DATA MINING FOR CUSTOMER SEGMENTATION AND PROFILING:

A CASE STUDY FOR A FAST MOVING CONSUMER GOODS (FMCG) COMPANY

by

Gülçin Buruncuk

Submitted to
The Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements
for the degree of

Master of Arts
in
Management Information Systems

Bogaziçi University
2006

The thesis of Gülçin Buruncuk
is approved by

Assistant Prof. Bertan Badur (Committee Chairperson) ____________________

Assistant Prof. Hande Kımıloğlu ____________________

Assistant Prof. Ulaş Akküçük ____________________

June 2006

ABSTRACT

DATA MINING FOR CUSTOMER SEGMENTATION AND PROFILING: A CASE STUDY FOR A FAST MOVING CONSUMER GOODS (FMCG) COMPANY

Data mining is a process of extracting hidden information from large databases by analyzing data from different perspectives. Segmentation and profiling analyses are data mining applications used to detect valuable customers of companies. Determining discrete valuable customer segments allows companies to focus on these groups and reallocate their limited sources to serve them.

The aim of this study is to propose a base for the customer relationship management activities by using data mining tools and applications for a FMCG company. Customer master data and sales transactions of customers are converted to meaningful information that can be used for customer relationship management activities. Customer segments and city segments are constructed using the buying behavior data of customers as the input. Nonhierarchical clustering algorithm is used to implement the segmentation analyses. Profiles of customer and city segments are defined using the characteristics of customers included in these segments.

Results of the customer and city segmentation analyses are combined by developing a new reporting environment with OLAP functionalities. Meaningful information obtained at the end of the analyses will help company to develop effective customer relationship management activities focusing on the valuable customers and valuable cities which will result in increasing the long term profitability of the company.

ÖZET

MÜŞTERİ SEGMENTASYONU VE PROFİL ÇIKARILMASI İÇİN VERİ MADENCİLİĞİ UYGULAMASI: HIZLI TÜKETİM SEKTÖRÜNDE BİR ÖRNEK OLAY İNCELEMESİ

Veri madenciliği, büyük veri tabanlarında yer alan verinin farklı açılardan incelenerek sakladığı gizli bilgilerin ortaya çıkarılması sürecidir. Müşteri segmentasyonu ve profillerin çıkarılması, şirketlerin değerli müşterilerinin belirlenmesi amacıyla kullanılan veri madenciliği uygulamalarıdır. Diğer müşteri guruplarından faklı ancak kendi içinde benzerlik gösteren değerli müşteriler gurubu elde etmek, şirketlerin kısıtlı kaynaklarını bu gurup için kullanmasına olanak sağlar.

Bu çalışmanın amacı, veri madenciliği araçları ve uygulamalarını kullanarak hızlı tüketim sektöründe yer alan bir şirket için, müşteri ilişkileri yönetimi aktivitelerine temel olabilcek bir yapı geliştirmektir. Müşteri ana verisi ve satış işlemleri, müşteri ilşikileri yönetimi için kullanılabilcek anlamlı verilere dönüştürülmektedir. Müşteri ve il segmentleri müşterilerin alışveriş davranışlarına göre oluşturulmuştur. Segmentasyon modellemesi için hiyerarşik olmayan kümeleme yöntemleri kullanılmıştır. Müşteri ve il segmentlerinin profilleri kapsadıkları müşterilerin özellikleri kullanılarak çıkarılmıştır.

Müşteri ve il segmentasyonuna ait sonuçlar OLAP fonksiyonaliteleri kullanılarak oluşturulşan yeni bir raporlama ortamı ile birleştirilmiştir. Tüm analizlerin sonucunda elde edilen anlamlı bilgi, şirketin değerli müşterilere ve değerli illere odaklanan efektif müşteri ilişkileri yönetimi aktiviteleri oluşturmasına ve sonuç olarak uzun dönemde karlılığını arttırmasına hizmet edecektir.

ACKNOWLEDGMENT

Firstly I want to thank to my advisor Assistant Professor Bertan Badur for his teaching. I learned not only academic knowledge but also the attitude of doing systematic analyses. I owe many things to my advisor and I am proud of being his student.

Secondly, I appreciate my brother and parents for the wonderful family atmosphere they have created for me from the beginning of my life. With our close neighbors they always encouraged me when I had troubles. Without them it was impossible to accomplish anything in my life.

Grateful thanks to my fiancé Ozan Aksoy. In spite of being physically far away from me, he always supported me and made me trust myself. Existence of him always makes my life easier and more beautiful which enables me to achieve difficult tasks.

Many thanks to one of the most important things I acquired from my master education: my valuable friends Gonca, Çağla, Ergin and Ürün. Also thanks to my close friends: İpek, Hepşen, Şebnem, Duygu and Ebru, for making my life enjoyable. Anytime that I have spent with them reduced the difficulties of the thesis period.

Finally, I want to thank to my managers and colleagues for their understanding. They always relaxed me when I felt disheartened. It was a pleasure to share the same working environment with them.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

A newly developed business culture which shifts the focus from a product oriented view to a customer oriented view gave rise to a challenge for the traditional mass marketing process by a new approach called one-to-one marketing. Emergence of this new culture increased the importance of establishing close relationships with customers and the concept of Customer Relationship Management (CRM) became incredibly important.

CRM can be defined as a customer-centric business strategy which focuses on managing the selected customers and business interactions established with them in order to; maximize customer satisfaction, minimize customer service costs, and as a consequence, manage the customer life cycle more intelligently to optimize the long term value (Ragins, Greco, 2003; Tan et al., 2002; Bradshaw, Brash, 2001).

Increasing the long term profitability of company is one of the main goals of CRM activities. Strategies such as, acquiring new customers, increasing the value of customer, and retaining the valuable customers are used to achieve this goal. Yet, to come up with successful strategies, various analyses drawing on significant amount of data about customers and their buying behaviors are needed. This new approach; employing bigger datasets to obtain better results, requires searching massive data stores to derive valuable information, which is extremely difficult to do manually for many market researchers.

Therefore, as a result of the need to convert these large amounts of customer data into meaningful information, data mining became an important concept which can be used to develop a base for subsequent CRM strategies. Descriptive and predictive techniques of data mining are exploited by analyzing customer information from various different perspectives in order to discover hidden patterns in these datasets which, at the end, provides useful information to make important strategic decisions.

Nevertheless, data mining process is not very straightforward. In a competing environment, retaining the valuable customers instead of acquiring new ones is accepted as a more effective strategy to increase long term profits. However, deciding on which customers should be retained is an important issue. For every company, there is a wide bunch of customers including some non-valuable ones as well. Customer segmentation, partitioning the market into smaller groups, and profiling these groups by describing the customers according to their attributes, are important applications of data mining to be carried out to distinguish the valuable customers.

The logic behind data mining techniques includes partitioning the customers into smaller groups according to the similarities among them with respect to some predefined variables. In the literature, various methods are proposed to execute this partitioning. The standard approach proposed in the literature to decide on the base of partitioning is, using either the basic Customer Lifetime Value (CLV) or the components of the Recency-Frequency-Monetary (RFM) method which is used to determine the CLV. (Berger et al., 2002; Berger et al., 2003) There are also other researchers who propose to extend the standard method by including additional variables into analyses (Libabi et al., 2002; Hogan et al., 2002).

This thesis examines a company operating in Fast Moving Consumer Goods (FMCG) market. Competition in the FMCG market started to grow. There will be considerable number of competitors in the market in coming years. Many alternatives will be available for the customers and for these customers switching between competitors becomes easier. Thus, the sector that the company operates in is a suitable environment for CRM strategies: in order to increase their long term profitability, companies need to determine their valuable customers and develop CRM strategies to retain them. Because of some legal and practical obstacles, target customers of CRM activities are limited to business type customers rather than end customers. In this study, business type customers are referred as customers of the company for simplicity. The most important step of developing successful CRM strategies is analyzing the customer data with data mining applications and techniques.

To fulfill this step, this thesis implements segmentation and profiling analyses to determine the valuable customers of the case company. In addition to the customers, cities that the customers are nested in are also partitioned into small groups via other segmentation and profiling analyses. The input knowledge required to differentiate the customers and the cities are extracted from the master data and raw sales transactions of the customers by using descriptive and predictive data mining techniques such as clustering. Components of basic CLV determination method Recency, Frequency and Monetary are used with some additional extensions to partition the company's customers and cities. Additionally, a reporting base has been developed at the end of these analyses which can be used as a base for the CRM activities of the case company.

The results of the data mining procedures carried on in this study can be used to derive valuable CRM strategies for the case company. Smaller manageable customer and city groups obtained for the company via segmentation and profiling analyses will give the opportunity to describe the characteristics of the customers of the company both at the customer segments level and the city segments level.

CHAPTER 2

LITERATURE SURVEY

A newly developed business culture which is focusing on a customer oriented view has replaced the old model of product oriented view. With this market evolution, the traditional process of mass marketing is being challenged by a new approach of one-to-one marketing. The marketing goal of the traditional process was to reach more customers and expand the customer base. With the increasing costs of acquiring new customers, the marketing goal of new model became to conduct business with current customers. As a consequence of this, the marketing focus shifted away from the breadth of customer base to the depth of each customer's needs. (Rygielski et al., 2002) Evolution of this new model increases the importance of establishing close customer relationships and determining the valuable customers to continue to work with via segmentation.

In this chapter an overview of data mining concepts is presented with its objectives and corresponding application areas. Afterwards data mining applications for customer relationship management is examined. Methodology that will be followed in this study as well as detailed explanation about the steps of customer segmentation will be analyzed in the following chapters.

What is Data Mining

Data mining is the process of extracting hidden information such as data attributes trends or patterns from large databases by analyzing data from different perspectives and summarizing it into useful information. The extraction process is achieved usually by finding correlations or patterns among dozens of fields of large databases which are usually constructed as data warehouses. Data mining gains the attention of people as a result of the accumulation of large amounts of data in the databases and the increasing need to analyze and then convert them into meaningful information. In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for querying the necessary information and the ability to store large databases is critical to data mining. Table 1 summarizes the evolution from data collection to data mining and gives a general view about the need for data mining. (Thearling, 2004)

Table 1 Steps in the Evolution of Data Mining

Evolutionary Step	*Business Question*	*Enabling Technologies*	*Characteristics*
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Prospective, proactive information delivery

Data mining uses the historical accumulated data as a guide, when effective decisions are needed to predict the future. This is achieved by offering a rich capability for modeling historical data and then using this model to predict likely future outcomes. This ability to give advance information about the future is unique to data mining and makes business professionals have a new perspective of factors, which truly contribute to business success or failure.

The historical data passes through some data mining steps in order to be meaningful for the analyzers. Steps of data mining projects will be covered in the methodology part of this study with some extensions.

Usage Areas of Data Mining

Data mining is a broad technology that can potentially benefit any functional areas within a business where there is a major need or opportunity for improved performance and where data is available for analysis that can impact the performance improvement. Table 2 shows examples of business applications in various sectors and industries that can most benefit from data mining. (Lubel, 1998; Musaoğlu, 2003)

Table 2 Examples of Data Mining Business Applications in Various Sectors

Sector / Industry	*Application*
Marketing / Retailing	√ Market basket analysis √ Finding market segments √ Identifying loyal customers √ Predicting what type customers will respond to mailing √ Finding customer purchase behavior patterns √ Finding associations among customer characteristics √ Determine items for cross selling / up-selling √ Detecting seasonal differences in sales patterns √ Product placement √ Forecasting sales / demand / revenue
Banking / Finance	√ Predicting customers that are likely to change their credit cards √ Identifying loyal customers √ Identifying fraudulent behavior √ Detecting patterns of fraudulent credit card usage √ Credit Scoring √ Risk assessment of credit √ Determine credit card spending by customer groups √ Segmentation of customers √ Analysis of customer profitability √ Managing portfolios √ Forecasting price changes in foreign currency markets √ Distribution channel analysis
Telecommunications	√ Churn analysis
Internet	√ Text Mining √ Web marketing
Manufacturing	√ Inventory Control √ Equipment failure analysis √ Resource Management √ Process / quality control √ Capacity management

Sector / Industry	*Application*
Insurance / Healthcare	√ Identifying fraudulent behavior √ Predicting which customers will buy new products √ Medical treatment analysis
Transportation	√ Loading pattern analysis √ Distribution channel analysis

Data Mining Techniques

Data mining analyzes relationships and patterns between fields of large databases by using the information gained from the user queries in order to find useful information. These analyses are done by using different data mining functionalities. Data mining can be interpreted as an interdisciplinary field including database systems, statistics, machine learning, and visualization. Depending on the case in hand and data mining method being used, techniques from other disciplines may be applied during analysis. Data mining techniques can be classified into two categories: descriptive data mining techniques and predictive data mining techniques. (Han, Kamber, 2000)

Descriptive Data Mining Techniques: These techniques describe the dataset in a summarative manner and presents interesting general properties of the data.

Predictive Data Mining Techniques: These techniques analyze the data in order to construct one or a set of models with which they attempt to predict the future. The main data mining functionalities under these main classes are as follows: (Han, Kamber, 2000, Withrow, 2003)

- *Concept/Class Description*, *Characterization and Discrimination*: Concept description is the most basic form of descriptive data mining. It gives information about the properties of data in a summarative manner.

- *Association Analysis:* Analysis about discovering relationships among huge amounts of data. These analyses are useful especially in selective marketing, decision analysis and business management. A popular area of application is market basket analysis, which studies the buying habits of customers by searching for set of items that are frequently purchased together by a specified customer. In association rule mining analysis firstly frequent item sets that are satisfying minimum support threshold are found. Than by using these item sets strong association rules in the form of A → B are generated. These rules also satisfy a minimum confidence threshold. Only the rules that have threshold above minimum confidence threshold and minimum support threshold are generated.
- *Classification and Prediction:* Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. While classification predicts categorical labels (classes), prediction models continuous valued functions. An example for this model may be assigning a consumer to a particular sales cluster based on their income level. There are some algorithms that are used for these analyses.
- *Cluster Analysis:* It is the process of grouping a set of physical or abstract objects into classes of similar objects called clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis has wide applications including market or customer segmentation, pattern recognition, biological studies, spatial data

analysis and many others. An example for clustering may be the analysis of business consumers for unknown attribute groupings. To do this the algorithm should get the well defined consumer attributes for searching.

Each of these techniques is applied via some predefined data mining algorithms. Figure 1 illustrates the relation between data mining applications areas, data mining techniques and algorithms.

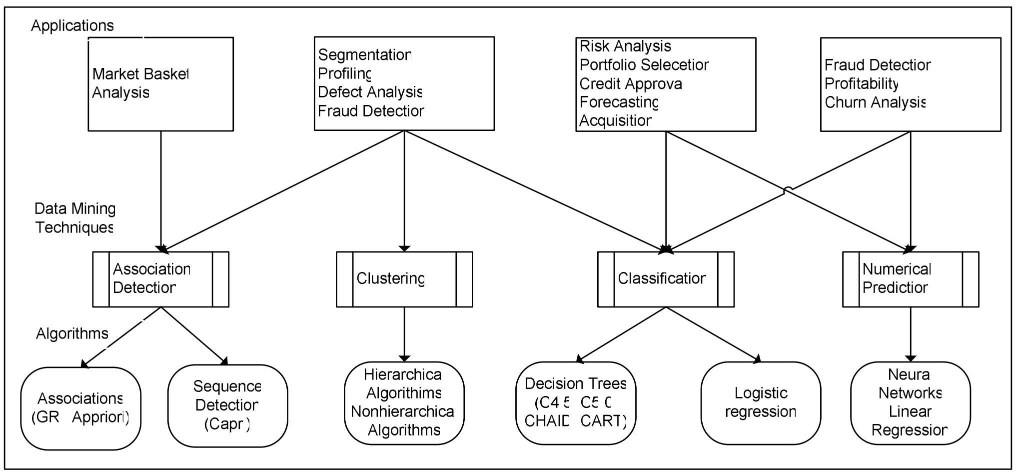


Figure 1 Data mining application areas, techniques and algorithms
Source Musaoğlu (2003)

Data Mining and Customer Relationship Management

Customer Relationship Management can be defined as a customer-centric business strategy focusing on managing selected customers and business interactions with them, in order to maximize customer satisfaction, minimize customer service costs and as a consequence optimize the long term value and manage the customer lifecycle intelligently.

The objectives of the CRM process can be summarized as shaping customers' perceptions of the organization and its products through identifying customers; creating customer knowledge; building committed customer relationships and;

gaining clearer insight and more intimate understanding of customers' buying behaviors and thus helping to build an effective competitive advantage (Ragins, Greco, 2003; Tan et al., 2002; Bradshaw, Brash, 2001). In order to achieve it goals CRM is redesigning functional activities and reengineering work processes with the support of intelligent application of CRM technologies. This combination of business processes and technology makes CRM neither a concept nor a technological term. Instead CRM is accepted as a business strategy that is being supported but not driven by the technology. (Tan et al., 2002)

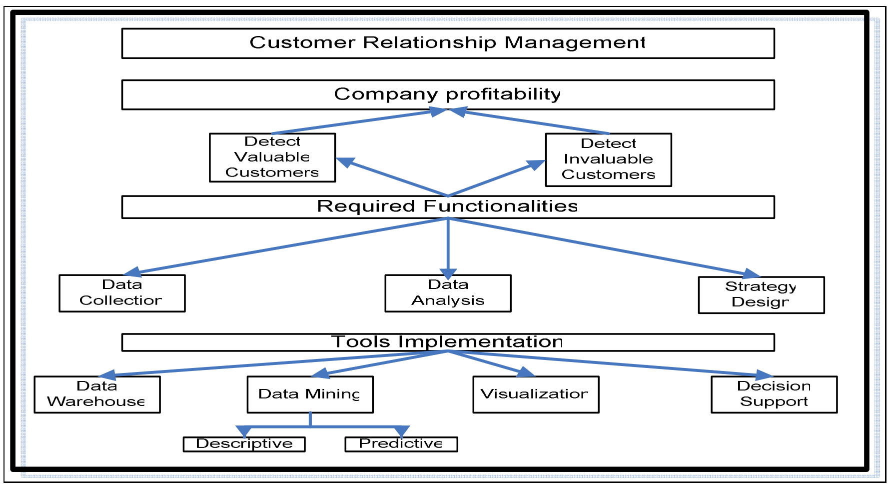


Figure 2 CRM overview
Source Lejeune (2001)

Several authors (Lejeune, 2001, Ryals, 2003, Eldestein, 2000) have been advancing the argument that increasing company profitability is one of the main goals of CRM activities. As it is shown in Figure 3, the goal can be achieved by detecting valuable and invaluable customers via segmentation and define strategies for them. The first task, identifying segments, requires collecting significant amount of data about customers and their buying behaviors. Although theory proposes to use more data for better results, analyzers cannot deal with massive data stores while searching valuable information.

Accumulation of large amounts of customer information and increasing need to convert them into meaningful information made data mining an important concept in developing a base for CRM strategies. By analyzing customer information and discovering hidden patterns in it data mining helps to understand past customer consumption behavior data in order to identify patterns for making strategic decisions (Rygielski et al., 2002).

Data mining techniques in CRM are used to identify additional products and services that should be offered to customers, to suggest the best time to make a cross sell or up sell offer, to develop strategies to increase customer value or to retain valuable customers based on their current usage patterns (Berry, Linoff, 2004; Ryals, 2003). Liu and Shih (2004) propose to use segmentation for product recommendation, when Lejeune (2001) defines segmentation as a way to detect the churn probability of customers and to define customer segments for cross-selling.

When the subject is to retain valuable customers, which is accepted as a more efficient strategy to increase long term profitability of company in a competing environment, the first issue is to identify market segments containing valuable or potential valuable customers and then armed with this information companies can target retention offers for predefined customer segments. One of the approaches being used in order to determine the valuable customers is Customer Lifetime Value (CLV) which aims to define valuable customers according to density and length of the relationship established between the company and the customer (Hwang et al., 2004; Buckinx, Poel, 2004). Generally, RFM (Recency, Frequency, and Monetary) method has been used to measure CLV (Berger, et al., 2002; 2001; Berger et al., 2003). RFM is one of the most powerful methods used for more than fifty years to predict customer behavior and assess the relationship between the enterprise and customers (Liu, Shih, 2004). Bult and Wansbeek (1995) defined the terms in turn as period since the last purchase, number of purchases made within a certain period and money spent during a certain period. However, according to Libai, et al. (2002) there are some limitations to the basic CLV determination approach such as not considering the short term switching behavior of customers and not offering comprehensive means for incorporating marketing mix variables and customer perceptions into the calculations. Additionally, Hogan et al. (2002) proposes an extended CLV model in order to overcome the deficiencies of RFM methodology. With the aim of avoiding the drawbacks resulted from limitations indicated above, instead of directly using CLV as the variable of the segmentation analysis components of CLV, Recency, Frequency and Monetary and other variables that are proposed by literature will be used in this study.

Customer Segmentation and Profiling

Customer segmentation has consequently been regarded as one of the most critical elements in achieving successful modern marketing and customer relationship management (Berson et al., 2000). Weinstein (2004) identifies customer segmentation as the process of partitioning markets into groups of potential customers with similar needs and/or characteristics who are likely to exhibit similar purchase behavior. Prospective activity of customer segmentation: customer profiling is the process of describing customers by their attributes, such as age, income and gender.

Segmentation offers to a company a way to know about the value of their customers. Knowing the profile of each customer, the company can treat the customer according to his/her individual needs in order to increase the lifetime value of customer. In the study by Kim et al., (2006) a case study has been analyzed with respect to customer segmentation and strategy development based on CLV. Results of the study show that refined strategies can be developed for the segments at the end of the segmentation process.

Wedel, Kamakura (1997) argues that selection of the segmentation variables is one of the critical issues of successful segmentation. Segmentation variables can be broadly classified into two groups; general variables which include customer demographics and lifestyles, and product specific variables which includes customer purchasing behaviors and intensions. According to Tsai, Chiu (2004) product specific variables should be included into segmentation analysis because of the inadequacy of general variables to determine purchasing patterns of customers. Several researches argue the potential variables for segmentation studies (Buckinx, Poel, 2004; Berger et al., 2002; Bayon et al., 2002). The variables that are proposed by literature to be used in segmentation are discussed below.

Segmentation Variables in Literature

- Length of Customer - Supplier Relationship

"Length of Customer – Supplier Relationship (LoR)" can be defined as the number of days passed from the first shopping of customer at the supplier. Variable shows how long the specified customer has been working with the company. Buckinx and Poel (2004) argue that the extent to which a customer is able to identify himself with a company is positively related to the period he is willing to continue this relationship. It is also mentioned that length of the relationship is positively associated to the perceived future stability of the relationship (Bayon et al., 2002).

- Frequency

"Frequency" can be defined as the number of purchases the customer made with representatives of the company from the beginning of its relationship with the company. Buckinx and Poel (2004) argue that the customer's frequency of purchases may be used to predict their future behavior because it is positively correlated to customer's expected future use (Buckinx, Poel, 2004). Two types of frequency are proposed by literature:

Frequency:

The variable frequency indicates the total number of orders given within four years by specified customers.

rFrequency:

"rFrequency" is the average purchase frequency of the customers. It is the ratio: frequency divided by LoR-1 as shown in Equation 1.

$$rFrequency = \frac{Frequency(cust_n)}{LoR(cust_n)} \tag{1}$$

"rFrequency" is used to equalize the chances of both new and old customers to be labeled as valuable with respect to their purchase frequency. Logically customers with longer LoR may have greater frequency values than the newer ones. Buckinx, Poel (2004) argue that by comparing frequency of each customer with his LoR a more comparable value is calculated to be used for the comparisons.

- Frequency Last Period:

"Frequency Last Period" shows how many times a customer has purchased goods from the company within the specified last period of analysis period. Buckinx and Poel (2004) argue that "Frequency Last Period" should be included in the analysis because of their power of predicting future purchase behavior of customers better than variables of overall period.

- Recency

"Recency" can be described as the number of days that passed between the last two transactions of the customer with the company within the observation period. Buckinx and Poel (2004) argue that the lower the value of "Recency", the higher the probability that a customer stays loyal. Different variations of the "Recency" variable are discussed in literature (Buckinx, Poel, 2004; Bayon et al., 2002).

Average Inter Purchase Time (IPT)

"Average Inter Purchase Time" reflects the "Recency" variable over the entire time period the customer has relation with the company. The formulation of the variable as is follows:

$$IPT = \frac{\sum (t-(t-1))}{TotalNumberofPurchases} \tag{2}$$

$$t: timeofthelastpurchase$$

$$t-1: timeofthepreviouspurchasebeforethelastone$$

- Monetary

"Monetary" variable can be defined as the total amount of spending that the customer made during its life time. "Monetary" value of each customer's past purchase behavior tends to be effective in predicting future purchase patterns and is used in the literature to determine future patterns (Buckinx, Poel, 2004). Variations of "Monetary" variable discussed in literature are as follows:

Monetary:

Total amount of spending that the customer made during its relationship with the company.

rMonetary:

"rMonetary" is the average spending of the customers. It is the ratio: monetary divided by LoR-1. Different from Monetary variable the length o the relationship of the customer with the supplier is taken into consideration in this variable (Buckinx, Poel, 2004).

“rMonetary” is calculated by dividing the monetary value of each customer to its length of relationship with the company as it is shown in Equation 3. The main reason to use this variable is to calculate comparable values for each customer with respect to monetary variable and avoid the wrong partitioning of customers into segments because of having incomparable figures.

$$rMonetary = \frac{Monetary(cust_n)}{LoR(cust_n)} \quad (3)$$

- rMajorTrip:

“rMajorTrip” indicates the proportion of transactions that includes a volume of purchase greater than the average volume of purchases done within analysis period. For example if one customer has purchased n times on average x liters per purchase, than “rMajorTrip” indicates how many of these n purchases exceeded the average x liters in terms of sales volume.

Steps to calculate this variable can be summarized as follows:

1.Calculation of average monetary value for the customer by using the formula:

$$AverageMonetary = \frac{Monetary}{TotalNumberofPurchases} \quad (4)$$

2.Calculating what percentage of the customer’s purchases are above the average monetary value by using the formula:

$$rMajorTrip = \left[\frac{(\forall Count(\forall(Monetary - AverageMonetary) > 0))}{TotalNumberofPurchases}\right] * 100 \quad (5)$$

- Time of the Day of Purchase– Timing of Shopping

“Timing of Shopping” is a variable that represents the average of each customer’s checkout time, in other words the time the specified customer left the shop. Buckinx and Poel (2004) argue that people do not shop all at the same time during the day or week. This difference may result from service quality differences among the several moments of the day such as shopping environment conditions or the attitudes of service personnel and has affect on the future buying patterns.

- Buying Behavior across Product Categories and Brand Purchase Behavior

This variable aims to catch the purchase pattern of a specified customer against special product categories. Buckinx and Poel (2004) argue that customer may start their relationship with the retailer by buying specific products. It is also claimed that the start of buying specific products or products from certain categories may be the indicator of a changing loyalty towards a company. On the other hand, if the customer is not pleased with the specific product or product from specific category even because of its price or quality, the probability of defection increases (Buckinx, Poel, 2004).

- Mode of Payment (MOP)

In state of shopping customers are offered several possible ways to pay their bill. The use of each of these modes of payment might be useful to classify customers into different segments (Buckinx, Poel, 2004).

- Customer Demographics

Several authors (Mittal and Kamakura, 2001; Vakratsas, 1998; Buckinx, Poel, 2004) have been advanced the argument that demographic characteristics of customers in some studies may be used to partition customers into different segments. Selection of the customer demographics is based on the general specialties of dataset that will be used for the analysis.

CHAPTER 3

METHODOLOGY AND PROBLEM DEFINITION

This chapter presents the methodology that will be followed for the customer segmentation and profiling analysis as well as the problem definition of the study. This section begins with the explanation of the methodology in detail with all steps that should be followed. The problem definition part explains the framework of the study. At last, business environment and general characteristics of the available data is discussed in the business environment section.

Methodology

Yen, Fang (2002) emphasizes the importance of using a predefined methodology for data mining and customer relationship management projects in order to avoid undesirable outcomes of learning process such as learning things that are not true and learning things that are true but not useful.

There are different predefined methodologies for both data mining and customer relationship management projects. Some of these methodologies are CRISP-DM and Two Crow Methodology (Edelstein, 2000; Crisp DM, 2000). In each methodology the life cycle of a project consists of different phases. It is common in all methodologies that the sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase which phase or which particular task of a phase, has to be performed next (Crisp DM, 2000).

With some needed extensions being made to Two Crow and CRISP_DM Methodologies, the phases of the methodology and the relationship between these phases that will be used in this study is shown in Figure 3. The outer circle in Figure 3 symbolizes the cyclical nature of data mining itself (CRISP DM, 2000).

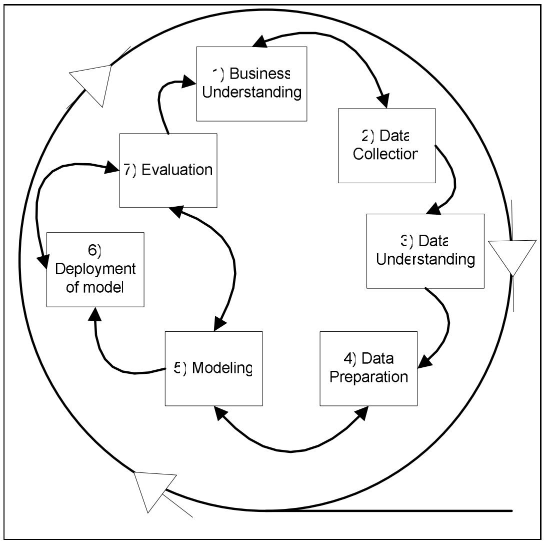


Figure 3 Steps of methodology used in this study

Steps of each phase are outlined in the following part:

1. Business Understanding

The initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition. Basic steps for this phase are as follows;

a. Determine data mining goals

b. Decide how the data mining would work to realize the objective to maximize customer satisfaction and minimize customer retention costs.

2. Data Collection:

The data collection phase aims to build database that contains the needed information for the analysis that will be done with data mining functionalities. This initial collection includes data loading from external resources for data understanding.

3. Data Understanding

The data understanding phase aims to explore data to understand the features of data in hand by analyzing the descriptive statistics, distribution of data etc. The phase contains activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. Basic steps for this phase are as follows;

a. Describing data: The step contains activities done to understand the general properties of data.

b. Exploring data: The step contains activities which can be addressed using querying, visualization and reporting. These include distribution of key attributes, relations between pairs or small groups of attributes or some simple statistical analyses. These analyses may address directly the data mining goals as well as contributing to data description and quality reports.

4. Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks include attribute selection, sample or data subset selection as well as data transformation and cleaning of data for modeling tools. Basic steps for this phase are as follows;

a. Data Selection: The step contains activities aims to decide on the data to be used for the analysis. Data Selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

b. Data Cleaning: The step contains activities to be achieved to raise data quality to the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling.

c. Data Construction: This task includes constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes.

5. Modeling

The Modeling phase covers all activities to build data mining model and explore alternative models to find the one that is most useful in solving the specified business problem with optimal results. During the activities of this phase because based on the needs of alternative models stepping back to the data preparation phase is often necessary. Basic steps for this phase are as follows;

a. Select Modeling Technique: This phase refers to selecting the specific modeling technique such as decision tree building with C4.5 or neural network generation with back propagation.

b. Build Model: This phase refers to running the modeling tool on the prepared dataset to create one or more models

c. Assess Model: In this phase alternative models are being assessed according to some predefined data mining success criteria and knowledge of the model builder. Different from the evaluation phase of the methodology this step only considers models whereas the evaluation phase also takes into account all other results.

6. Deployment of Model

This phase refers to running the modeling tool on the prepared dataset to create one or more models.

7. Evaluation

The evaluation phase aims to evaluate the model and review the steps executed to construct the model to determine whether it properly achieves the business objectives or not.

Problem Definition

The initial phase of data mining projects as mentioned in the methodology is business understanding. The phase focuses on understanding the project objectives and converting this knowledge into a data mining problem definition.

With the increasing number of competitors, the alternatives of the customers and the switching probability of a customer between the competitors have been increased in every type of market named as Business to Business (B2B) where both parties in the relation are business parties; Business to Customer (B2C) where the relation is established between a business party and; end customer and Business to Business to Customer (B2B2C) in which there is an intermediary business part between the producer company and end customer. When customers in B2C type market can change their suppliers easily without any switching cost, switching between alternative suppliers is a costly action for customers in B2B and B2B2C types markets especially if they are working on contractual basis. However, with the increasing competition in these types of markets, the markets have been fluctuating with large number of choices served to the customers. As a result of this, the probability that the customers may change their choices although it costs them big amounts has increased. The mentioned facts force the case company to be more careful about effective management of customer relationships in order to defense its market share against potential competitors and to increase its long term profitability.

Almost all firms have limited resources to serve their customers and managing customer relationships does not mean to satisfy every single customer's need. Indeed in order to protect its markets share company should use its limited resources in an effective manner by selecting the valuable customers and making efforts to keep them. Based on these facts case company decided to determine customer groups to which it should give priority in managing its relationship with. When defining the valuable customer groups, it is accepted that labeling the long life customers of the company as the profitable ones and use the limited resources to support the relationship with it may be unprofitable for the company. Instead in this study, all customers containing the short and long life ones will be treated equally and by using segmentation analysis with distinguishing variables profitable ones will be selected among them. Another way company prefers to manage relationships with the customers is to determine the valuable cities in which company has customers and develop special customer relationship activities for the ones in these cities. To put into action this alternative just like customers, cities in which company performs can be grouped as valuable and invaluable ones via segmentation analysis.

In this study, customers of the company will be segmented according to their buying behavior. Customer lifetime value components will be used with some extensions in order to define the segments which contain valuable customers. Additionally, cities in which the company performs will be segmented according to the buying behaviors of the customers located in each of them. For both segmentation analyses not only the variables available in the data warehouse of company but also the new derived ones will be used. Information gained from both segmentation analyses will be used to form a reporting environment which can be used as a base for developing CRM strategies.

In order to achieve the mentioned objectives of the study, data mining techniques will be used with the following goals:

- Preparation of a dataset with both existing variables that company already uses and new derived ones. Dataset will be used to partition the customers of company into small manageable groups for CRM activities.

- Preparation of a dataset, with derived variables that can be used to partition the cities in which company performs into small manageable groups for CRM activities. Variables for cities will be derived by using the ones of customers located in each of them.

- Segmentation analysis of company's customers
- Segmentation analysis of the cities in which company performs
- Profiling of the segments constructed for both customers and cities.
- Creation of a new reporting environment with information gained from segmentation analyses to develop customer relationship management strategies.

Business Environment Description

Case company is one of the companies that have activities in Fast Moving Consumer Goods (FMCG) sector with a significantly great market share compared to its competitors. FMCG is a classification that refers to wide range of frequently purchased consumer products including beverages, food products, cigarettes, toilet soaps, creams, toothpaste, shampoos and detergents(Wikipedia, 2006). Among these categories case company is focusing on beverages. When the situation of the market is analyzed it is obvious that there is not a serious competitor threat for the company right now. However, the market has started to fluctuate in recent years and it is expected that a number of competitors of the case company will increase in the coming years. In order to be ready for the possible competitor threats developing a base for CRM strategies became incredibly important for the case company.

The company works in B2B2C type market on a contractual base. There are two types of customers of the case company as listed below:

- *Business Type* : Distributors and Selling Points
- *Customer Type* : End Customers

Flow of orders and goods between the case company and its customers are visualized in Figure 4. As it is shown, the connection between the selling points and case company is obtained by the distributors. However, based on size of distributors orders of selling points are collected by the sales personnel of the case company or the sales personnel of the distributors. When the issue is transmitting the goods to the selling points, again distributors are in the intermediary position between the case company and selling points. Case company transmits the goods to the distributors and according to the needs of the selling points the goods are distributed to them by the specified distributors.

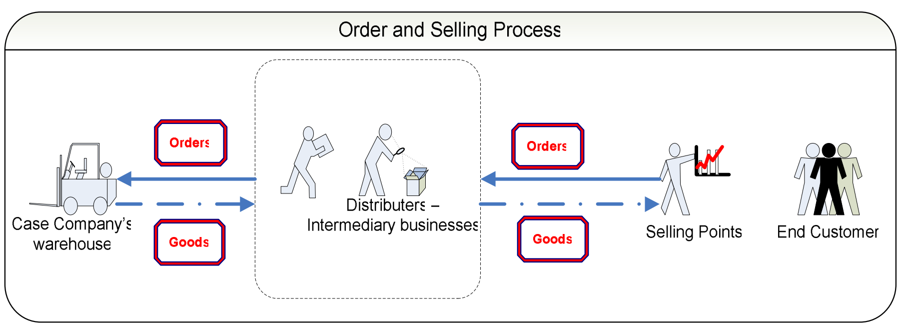


Figure 4 Order and selling process of case company

However, because of some governmental restrictions, the end customers cannot be direct targets of the CRM activities in the specified sector. Additional to this, the data warehouse of the case company does not contain data related to the sales transactions between the sales points and end customers. As a result of this, the target customers in this study will be business type customer of the company (distributors). Segmentation and profiling analyses will be made with the data related to this type of customers. For the sake of simplicity, business type of customers – distributors and selling points will be referenced as customers from now on in the study.

Transactions related to orders and finalized sales to the customers are recorded simultaneously to the Enterprise Resource Planning (ERP) System of the company. At the end of each day, the specified data is extracted to the data warehouse of the company. The transaction data includes details such as the product code that is being sold, related product brand code, volume code of the product, date the transaction took place as well as the selling point code which indicates to whom the products are sold. On the other hand, master data includes customer location related variables such as geographical region and city of the customer, as well as the position of the customer location. Additionally, customer master data includes variables related to the working style of customers such as period the customer works, the way it prefers to pay, brand categories it prefers to sell.

Data needed for the analyses is taken from the data warehouse of the case company. Real data of the company is sampled and nearly 60,000 customers are used for the analyses. However because of confidentiality data is not used in the original format and recoded. By using this data, variables that will be used for segmentation analysis are derived. Variables that will be derived are selected among the ones literature purposed according to the availability of data, characteristics of case company as well as characteristics of B2B market.

CHAPTER 4

DATA UNDERSTANDING AND PREPARATION

Dataset that will be used for segmentation analysis is constructed by using case company's customer master data and raw sales data over a three year period. Sales transactions for a randomly selected large sample of customers are extracted from company's data warehouse via a reporting tool that is being used to report sales activities of the company. Customer master data is directly used after some data preparation activities that will be discussed in detail in the following sections. On the other hand, sales data is transmitted into different variables that are defined according to the needs of the problem in hand under the light of literature survey.

Master data of company's customers were entered into database by a variety of employees, including sales representatives all over Turkey, customer services personnel and information systems personnel located at headquarters. As a natural effect of company's business flow, master data entry and maintenance have lower priority than other activities which are more customer directed. Based on this fact, some customer master data were incomplete and some data were clearly in error when there is no significant problem about the sales transactions. As a result, some data preparation activities are performed on customer master data whereas data construction activities are achieved to create the needed variables from sales transactions. With all effort made for different tasks of this phase, data understanding and preparation took the longest time and effort among other activities completed during the study.

Data Selection

Selection of appropriate data for the analysis according to data mining goals, quality and technical constraints include two main activities: selection of attributes (columns) as well as selection of records (rows) in a table.

In most applications, data selection phase is completed at the beginning of the Data Preparation processes. Different than general flow, data selection activities are divided into two main parts in this study. One of these parts, selection of attributes, is achieved before the data preparation activities started together with random selection of records from the data warehouse of case company to build a large sample. On the other hand, this large sample is decreased to a smaller one by reconsidering the data selection step. This second part is performed based on the results of the data cleaning process which is another step of data preparation.

At the end of the first part of data selection process, sales transactions over three year period for approximately 80000 customers have been collected and the attributes of these customers that will be used in the analysis are determined. This initial elimination to select the large sample is achieved on a random base.

To fulfill the aim of selecting the attributes that will be used in the analysis, master data for company's customers is analyzed. While selecting the attributes, some initial conditions discussed by Berry, Linoff (2004) are taken into consideration. Berry, Linoff (2004) argue that attributes for which almost all records have the same value as well as the ones that do not have value for most of the customers should not be included in the analysis because they are useless to distinguish between different rows. In addition, same sources indicate that categorical columns that take different value for almost every row do not have predictive value and should be discarded from the analyses. When the first two considerations are not valid for the dataset in hand, based on the third consideration Customer Name, Address, Telephone Number, Contact Person attributes are not included to analysis. Table 3 shows the general characteristics of the attributes selected. Detailed information related to each variable can be found in Appendix A.

Table 3 Data Dictionary of Categorical Variables

Field Name	*Description*	*Variable Type*	*Data Expression*	*Length*	Can Hold Null	Derived or Not
Sales Directorate	The directorate which the customer is bound to.	Categorical - Nominal	Integer	4	No	No
Customer Code	The unique number that is given from the system to each customer	Numeric - Discrete	Integer	7	No	No
City	City where the customer is located	Categorical - Nominal ‖ Numeric Discrete	Integer	2	Yes	No
Customer Type	defines the type of the customer determined according to the way they are using when selling the products of the company	Categorical - Nominal	Text	6	Yes	No
Working Period	Defines the working period of the customer	Categorical - Nominal	Text	8	Yes	No
Customer Group	Defines the group of the customer which is determined according to the physical and legal structure of their shops	Categorical - Nominal	Text	20	Yes	No
SES Group	Defines the socio economic status of the people who lives around the customer's location	Categorical - Nominal	Text	10	Yes	No
Region Description	Defines the region of the city that the customer has located	Categorical - Nominal	Text	10	Yes	No
Position Group	Defines the positioning of the places that the customer has located	Categorical - Nominal	Text	10	Yes	No
Customer Structure	Defines the group of customer which is defined according to the visual presentation of them.	Categorical - Nominal	Text	8	Yes	No
Visit Frequency	The characteristic shows visit ferquency of the firm for the specified customer.	Categorical - Nominal	Text	20	Yes	No
Customer Specialty	Defines the group of customers which is defined according to the products they are selling.	Categorical - Nominal	Text	20	Yes	No
Working Type	Defines the group of customers which is defined according to their payment method	Categorical - Nominal	Text	20	Yes	No

Data Cleaning

Data cleaning activities aim to raise data quality to the level required by the selected analyses techniques. Bearing this in mind, unreasonable entries for each variable are analyzed and cleaned, if appropriate. As the cleaning method, insertion of suitable defaults is used. These defaults are determined by taking other available attributes of specified customer as references.

Second part of the data selection phase, selection of rows is being done with the aim of acquiring reasonable records for the analysis. By keeping this aim in mind, after the data cleaning step finished, records with missing values for most of the variables that will be used for the analysis are removed from the sample. Unreasonable records such as those of customers who have non-zero amount of purchase but have never made any transactions are removed. At the end of this phase a dataset that contains 57,933 customers is remained to be used in the subsequent analyses.

Data Construction

Data construction phase includes constructive data preparation operations such as, the production of derived attributes, entering new records or transforming the values of existing attributes. Two available operations are performed in this analysis. Firstly, sales transactions of 57,933 selected customers are used to derive new variables in order to represent the essential facts that the dataset does not currently represent with the available attributes. As mentioned before, variables to be derived are determined according to the needs of the problem in hand availability of dataset among the ones proposed in the literature. Since the dataset was not useful variables that are proposed by literature such as; Mode of Payment, Usage of Promotions, Timing of Shopping and Risk are not used in this study. On the other hand in order to measure the variability of data standard deviations of Amount and Recency variables are derived during the data preparation phase. Additionally, in order to measure the differences between different years of analysis period some variables on year base are also derived. Microsoft Office application, Excel capabilities are used to derive these variables from the raw sales transactions of customers within three years. In addition, in this step variables that will be used to partition the cities into smaller groups are derived from these sales transactions, too. Table 4 shows the general information related to the variables derived in this phase. Detailed information about these variables can be found in Appendix A.

Secondly, as it is shown in Table 4, since the measurement scales of the variables are different and the modeling algorithm that will be used is not able to handle these different scales, values of the variables are transformed before the partitioning process start. Data is transformed into standard scores (z-scores) to eliminate the bias introduced by the different scales of different attributes used in the analyses. Formula to calculate standard score for a variable is shown in Equation 6.

$$z\_score = \frac{\mathrm{X} - \mu}{\sigma} \qquad (6)$$

Table 4 Data Dictionary for Continuous Variables

Field Name	*Aliases*	*Description*	*Variable Type*	*Data Expression*	*Measurement Scale*	*Can Hold Null*	Derived or Not	How to Calculate
Length of Customer - Supplier Relationship_1	LoR_1	Shows how long the specified customer is working with the company during the analysis period; four year.	Numeric - Continuous	Number	Days	No	Yes	(Last purchase date – First purchase date) within analysis period.
Length of Customer - Supplier Relationship_2	LoR_2	Shows how long the company is working with the specified customer. Different from the Length of relationship_1 variable it does not shows only the duration in the analysis period	Numeric - Continuous	Number	Days	No	Yes	(Last purchase date – Customer Opening Date)
Frequency		The number defines how many times the specified customer purchased from the firm during the analysis period	Numeric - Discrete	Number	Count	No	No	
rFrequency		Shows number of purchases customer made relative to the length of relationship (LoR_1)	Numeric - Continuous	Number	Proportion	No	Yes	(Frequency / Length of Relationship_1)
Frequency last one year		The number defines how many times did the distributor purchased from the firm during the last one year	Numeric - Discrete	Number	Count	No	No	
Recency		The number defines the duration passed between the last two purchases of customer from the firm.	Numeric - Continuous	Number	Days	No	Yes	(Date of the last Purchase – Date of the previous purchase before the last one) within the analysis period.

Field Name	*Aliases*	*Description*	*Variable Type*	*Data Expression*	*Measurement Scale*	*Can Hold Null*	Derived or Not	How to Calculate
Average Inter Purchase Time	IPT	The number defines the average of the periods passed between each purchases of the customer from the firm.	Numeric - Continuous	Number	Days	No	Yes	Calculate the average of the time pass between each two purchases of the distributor. (∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) within the analysis period.
Standard Deviation of Recency	StdDev Recency	Shows the standard deviation of the inter purchase time.	Numeric - Continuous	Number	Number	No	Yes	Calculate the standard deviation of the recency. StDev (∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) within the analysis period.
Coefficient Variation of Recency	CvRecency	Shows the ratio of StdRecency to Mean Recency	Numeric - Continuous	Number	Number	No	Yes	(StDev (∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) / Average (Date of the last Purchase – Date of the previous purchase before the last one)) within the analysis period.
Total Amount		Shows the total amount of products that the specified customer purchased from the company during the analysis period	Numeric - Continuous	Number	Liter	No	No	

Field Name	*Aliases*	*Description*	*Variable Type*	*Data Expression*	*Measurement Scale*	*Can Hold Null*	Derived or Not	How to Calculate
Amount		The number defines the average of the amounts the customer purchased from the firm during the specified period.	Numeric - Continuous	Number	Liter	No	Yes	(Total Amount / Frequency) within the analysis period
rTotal Amount		Shows total amount of products that the specified customer purchased from the company during the analysis period relative to the length of relationship (LoR_1).	Numeric - Continuous	Number	Proportion	No	Yes	(Total Amount / Length of Relationship_1) within the analysis period
rAmount		Shows average amount of products that the specified customer purchased from the company during the analysis period relative to the length of relationship (LoR_1).	Numeric - Continuous	Number	Proportion	No	Yes	((Total Amount / Frequency) / Length of Relationship_1) within the analysis period
Standard Deviation of Amount	StdevAmount	Shows the standard deviation of the average amount of products that the specified customer purchased from the company during the analysis period	Numeric - Continuous	Number	Number	No	Yes	(StDev (Total Amount / Frequency)) within the analysis period
rMajorTrip		Shows the percentage of the purchases of a customer which exceeds the average amount for the purchases that specified customer has done. The variable indicates the percentage of purchases that could be classified as a big shopping incidence.	Numeric - Continuous	Number	Percentage	No	Yes	(every (Count (every (Amount for specified order - Average Amount) > 0) / Total Number of Purchases) * 100)

Field Name	*Aliases*	*Description*	*Variable Type*	*Data Expression*	*Measurement Scale*	*Can Hold Null*	Derived or Not	How to Calculate
Frequency for 2002 / 2003 / 2004		The number defines how many times the specified customer purchased from the firm during the year at issue	Numeric - Discrete	Number	Count	No	No	
Average Inter Purchase Time for 2002 / 2003 / 2004		The number defines the average of the periods passed between each purchases of the customer from the firm during the year at issue.	Numeric - Continuous	Number	Days	No	Yes	Calculate the average of the time pass between each two purchases of the distributor. (∑ (Date of the Last Purchase in year at issue – Date of the previous purchase before the last one) / Total Number of Purchases) within the analysis period.
Total Amount for 2002 / 2003 / 2004		Shows the total amount of products that the specified customer purchased from the company during the year at issue	Numeric - Continuous	Number	Liter	No	No	
Amount for 2002 / 2003 / 2004		The number defines the average of the amounts the customer purchased from the firm during the year at issue	Numeric - Continuous	Number	Liter	No	Yes	(Total Amount / Frequency) within the year at issue
Average Sales_2 City		The number defines the average amount of products customers in the specified city purchased from the firm during the specified period.	Numeric - Continuous	Number	Liter	No	Yes	∑ (Amount_Customer where City_Customer= City at issue) / Count of customer in the city

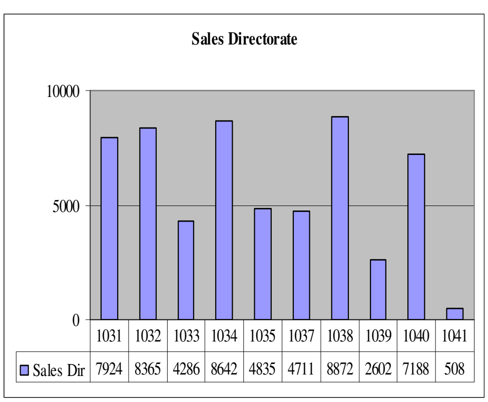
Field Name	*Aliases*	*Description*	*Variable Type*	*Data Expression*	*Measurement Scale*	*Can Hold Null*	Derived or Not	How to Calculate
Average IPT City		The number defines the average of the periods passed between each purchases of the customers in a specified city from the firm.	Numeric - Continuous	Number	Days	No	Yes	Avg (IPT_Customer where City_Customer = City at issue)
Count of Customers City		The number defines how many customers does the company have in the specified city.	Numeric - Discrete	Number	Count	No	No	Count(Customers where City_Customer = City at issue)
Average Frequency City		The number defines how many times the customers in the specified city purchased from the firm during the analysis period, on average	Numeric - Discrete	Number	Count	No	Yes	Avg (Frequency_Customer where City_Customer = City at issue)
Average Frequency Last Year City		The number defines how many times did the customers in specified city purchased from the firm during the last one year, on average	Numeric - Discrete	Number	Count	No	No	Avg (Frequency Last One Year_Customer where City_Customer = City at issue)
Average Recency City		The number defines the average duration passed between the last two purchases of customers in a specified city from the firm.	Numeric - Continuous	Number	Days	No	Yes	Avg (Recency_Customer where City_Customer = City at issue)
Sales per Customer City		The number defines per capita consumption of company's products for the specified city. Results of year 2000 population census, declared by government are used for calculation	Numeric - Continuous	Number	Liter	No	Yes	∑ (Total Amount_Customer where City_Customer= City at issue) / Population of the city

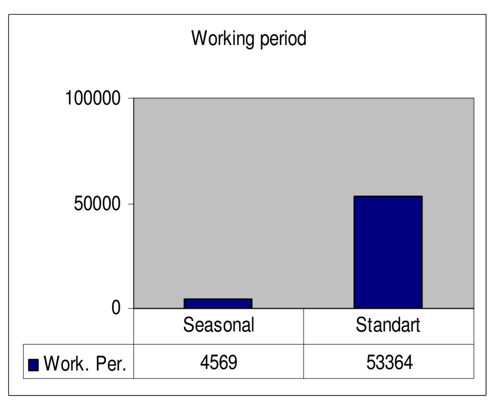
Data Examination

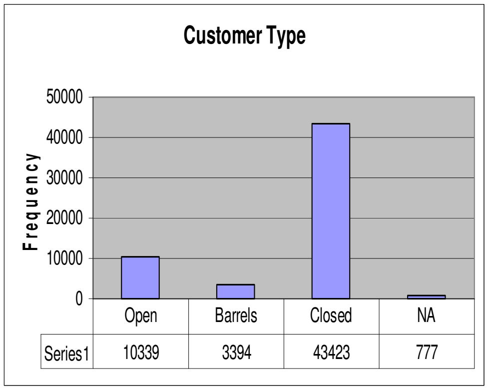
The step contains activities which can be addressed using visualization and reporting. In data examination step, two different analyses are applied to the dataset to understand the general characteristics of data that will be used in the analyses and get familiar with it. First analysis is done to understand general distributions of categorical variables in the dataset. Histograms and pie charts are created with categorical variables for 57,933 cases to analyze the general characteristics of data. Motivation to develop charts and the corresponding results will be discussed in following parts of this chapter. On the other hand, in order to deepen the understanding about the general characteristics of the derived attributes, functionalities of SPSS analysis tool is used. The descriptive statistics of these variables will also be analyzed in the following parts of this chapter.

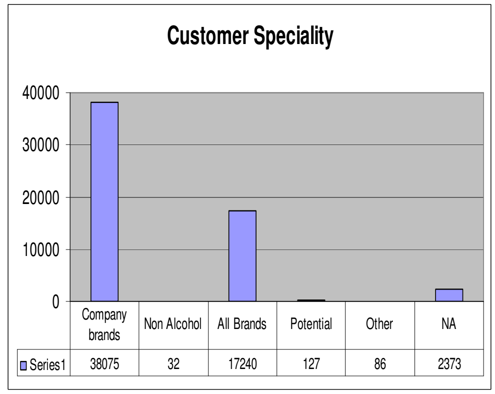
Data Examination for Categorical Variables

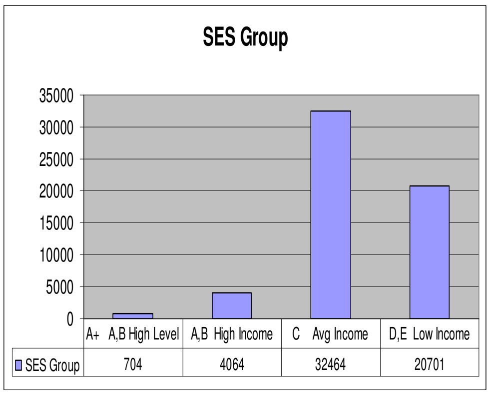
Histograms such as those in Figure 5, show how often each value or range of values occurs in the dataset used for the analyses. The vertical axis is the count of records, and the horizontal axis is the corresponding values in the column. The shape of histograms shows the distribution of values which are accepted as the same distribution as the original dataset. By analyzing these distributions, the most frequent values for each variable as well as the less common ones are determined.

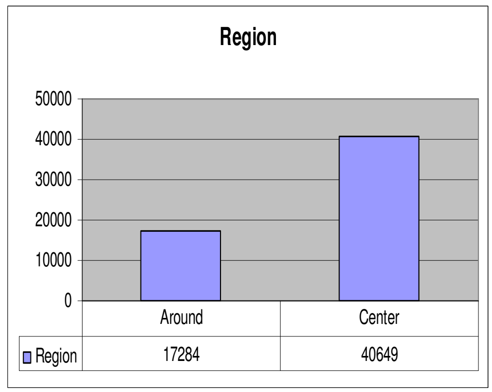


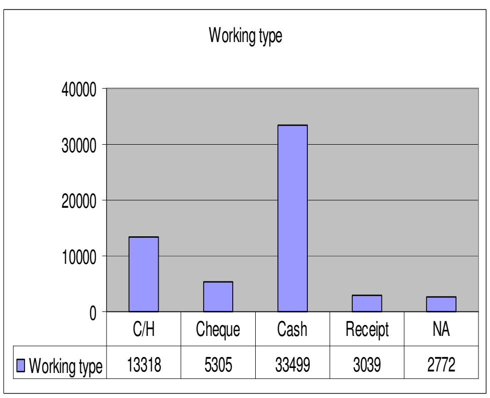


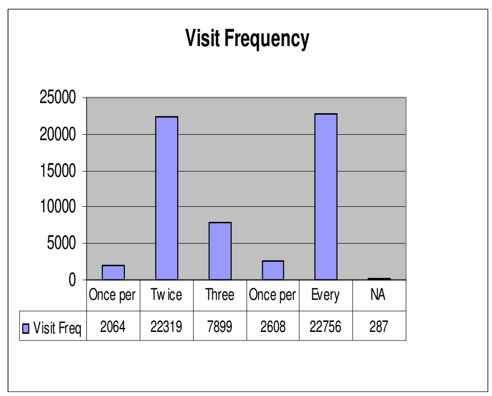

Customer Type
Frequency
50000
40000
30000
20000
10000
0
Open
Barrels
Closed
NA
Series1
10339
3394
43423
777

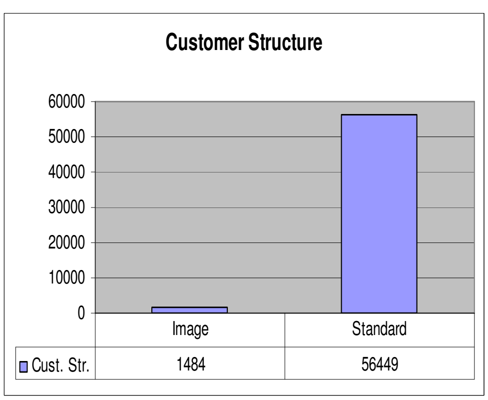

Customer Speciality
40000
30000
20000
10000
0
Company brands
Non Alcohol
All Brands
Potential
Other
NA
Series1
38075
32
17240
127
86
2373

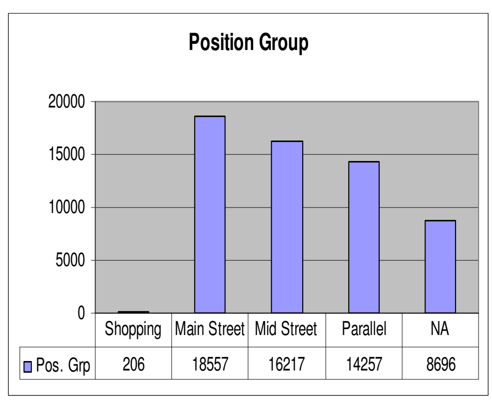

SES Group
35000
30000
25000
20000
15000
10000
5000
0
A+ A,B High Level
A,B High Income
C Avg Income
D,E Low Income
SES Group
704
4064
32464
20701


Region
50000
40000
30000
20000
10000
0
Around
Center
Region
17284
40649


Working type
40000
30000
20000
10000
0
C/H
Cheque
Cash
Receipt
NA
Working type
13318
5305
33499
3039
2772


Visit Frequency
25000
20000
15000
10000
5000
0
Once per
Twice
Three
Once per
Every
NA
Visit Freq
2064
22319
7899
2608
22756
287


Customer Structure
60000
50000
40000
30000
20000
10000
0
Image
Standard
Cust. Str.
1484
56449


Position Group
20000
15000
10000
5000
0
Shopping
Main Street
Mid Street
Parallel
NA
Pos. Grp
206
18557
16217
14257
8696

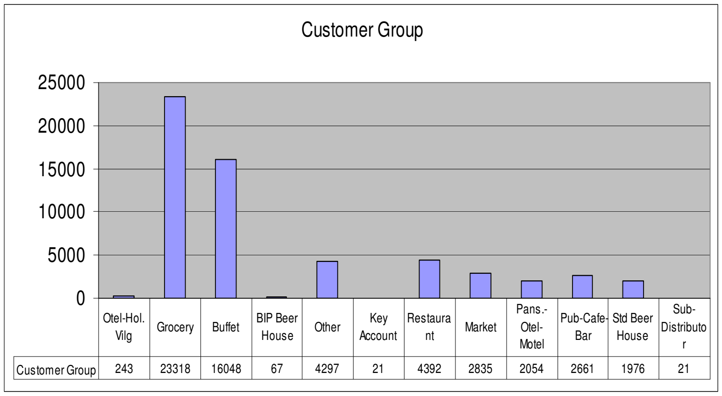


Figure 5 Frequency diagrams of categorical variables

Data Examination for Continuous Variables

To understand the general characteristics of continuous variables that will be used in analysis for both customers and cities, analysis capabilities of SPSS software is used. For 57933 cases at the customer level and 78 cases at the city level, descriptive statistics of the variables are computed and characteristics of these variables in terms of location and dispersion are analyzed.

- Variables at the customer level

Table 5 shows the statistical values of the variables at the customer level. In a first look to all of the variables at the customer level, the first issue to be taken into consideration is the dispersion of the variables. As shown in Table 5, except rMajorTrip, for each of the variables in hand, the mean is greater than the median and both are greater than the mode. This characteristic of the variables reveals that the dispersion of all of the variables is right skewed and values are cumulated around the first quartile. Together with the other characteristics of the distributions of the variables which will be mentioned in the following part, it is concluded that none of the variables has a normal distribution.

When the 3rd quartile (from which %75 of all of the values are smaller) is compared to the maximum value, it can be observed that the maximum values are sometimes hundreds times greater than the 3rd quartile values for all of variables. This observation can be interpreted as the existence of excessive outliers.

When Table 5 analyzed it is obvious that none of the variables except LoR_2 has a missing value. The reason behind this fact is; all of the variables in hand are "derived" variables. All are calculated from the raw sales data of the customers.

Table 5 Descriptive Statistics of the Variables at the Customer Level.

		LoR_1	LoR_2	Frequency	rFrequency	Frequency Last year	Recency	IPT	Average Purchase	Total Amount	rMajor Trip	StdDev Recency	StdDev Amount	rAmount	rTotal Amount	CV Recency
Valid		57933	36816	57933	57933	57933	57933	57933	57933	57933	57933	57933	57933	57933	57933	57933
Missing		0	21117	0	0	0	0	0	0	0	0	0	0	0	0	0
Mean		392.49	2049.47	66.42	0.1686	47.87	14.97	10.62	141.64	9982.62	36.75	9.4	117.5	.829	22.632	.888
Median		363	1459	51	0.1482	40	7	6.51	75.96	3708	37.21	5.153	49.02	.227	11.392	.777
Mode		0	1826	1	0	0	3	0	12	12	50	.000	.000	.000	.000	.000
Std. Deviation		252.17	2133.13	60.51	0.1203	40.509	31.652	17.36	277.83	24779.85	15.33	35.494	345.471	9.923	55.3515	5.0732
Range		1095	38275	785	2	335	827	615	27162.9	1146039	98.84	7206.563	53037.6	1503.960	5066.2017	112.124
Minimum		0	0	0	0	0	0	0	0	0	0	.000	.000	.000	.000	.00
Maximum		1095	38275	785	2	335	827	615	27162.9	1146039	98.84	7206.563	53037.6	1503.960	5066.2017	1112.1
Percentiles	25	225	572	22	0.0875	17	3	4.24	39.37	1224	12	2.841	19.831	.1070	4.745	.567882
	75	456	2701	93	0.2279	70	14	10.96	153.96	10123.04	60	9.7418	118.807	.5028	25.465	1.0375

In the following session, variables in Table 5 are analyzed more specifically to give more detailed information about the specific characteristic as well as other commonalities of the variables.

Inter Purchase Time:

When the results for Inter Purchase Time (IPT) variable are analyzed, it is obvious that with a standard deviation 1.7 times greater than its average IPT variable is highly dispersed. The distribution of the variable has the same specialties, discussed above for all variables.

By analyzing the frequency diagram, it is revealed that more than half of the IPT values are smaller than 1 week and nearly %70 of all cases are smaller than 10 days and %80 smaller than two weeks. This shows that the average frequencies of customers generally do not exceed two weeks. However, outliers greater than two weeks constitutes %20 of all cases.

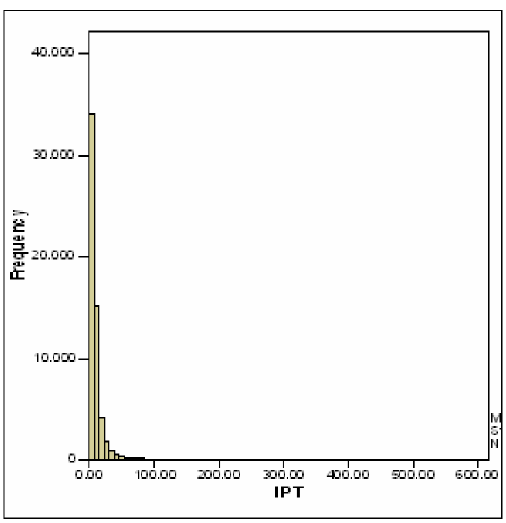


Figure 6 Frequency diagram of IPT variable

LOR:

As shown in Table 5, by having a standard deviation 1.6 times smaller than the mean of variable LoR's distribution is more dispersed than a normal distribution. However when the dispersion of the other variables analyzed it is clear that "LoR" can be accepted as one of the most "normal" variable with a relatively meaningful range value which is three times greater than its mean.

"LoR-2" variable has many missing values. The company recorded many passive customers who do not purchase at all. All these recorded but passive variables are regarded as missing. Variable seems unreliable because some cases takes meaningless values. For example the maximum value of "LoR-2" variable is more than 100 years which is impossible because the company is only thirty five years old. Therefore, this variable is discarded from the subsequent analysis.

Frequency diagram in Figure 7 reveals that "LoR-1" variable has again a right skewed distribution but now it is less skewed compared to the other variables. Another thing that can be observed from the diagram is that, data is mostly accumulated around the median of variable.

Although it has a mode of zero, which can be explained by the existence of single time purchasers, when we look at the histogram of "LoR" variable, it can be seen that the cluster that has a mid point of 300 that reaches the highest frequency level. This shows that the most frequently observed relationship age is one year for the analyzed cases.

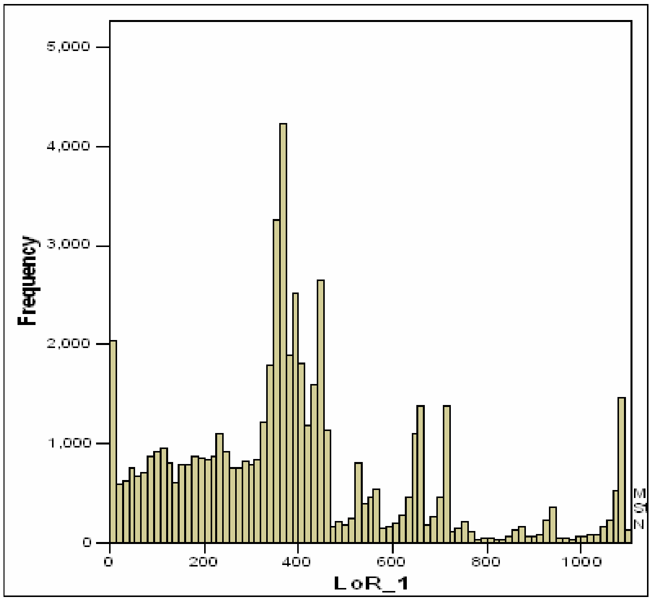


Figure 7 Frequency diagram of LoR_1 variable

<u>Frequency, Frequency Last Year:</u>

These two variables are closely related to each other therefore will be discussed together. Descriptive statistics of these variables draws a similar picture as the other variables. Same as the others, the variables are highly dispersed and right skewed. However, different from the previous variables, this time mean, median scores are closer to each other since data is accumulated around both first and second quartiles. Another important characteristic of the distribution of "Frequency" variable is the high number of cases having the value of "zero". But the mode of the variable is "one" which points that the biggest group of customers is the ones who purchased only once. On the other hand, "Frequency last year" has a mode of "zero" which indicates that last year the most observed purchase frequency is zero.

Frequency diagram of Frequency variable shows that the maximum value is nearly 8.4 times greater than the 3rd quartile, which also verifies that the variable is highly dispersed mainly because of the existence of the outliers. On the other hand, different from the frequency variable, Frequency Last Year has a maximum value which is nearly 5 times greater than the 3rd quartile which is a smaller value compared to the one for frequency. This shows that the dispersion of this variable is less dispersed than the frequency variable. But there is still, a significant amount of outliers for this variable.

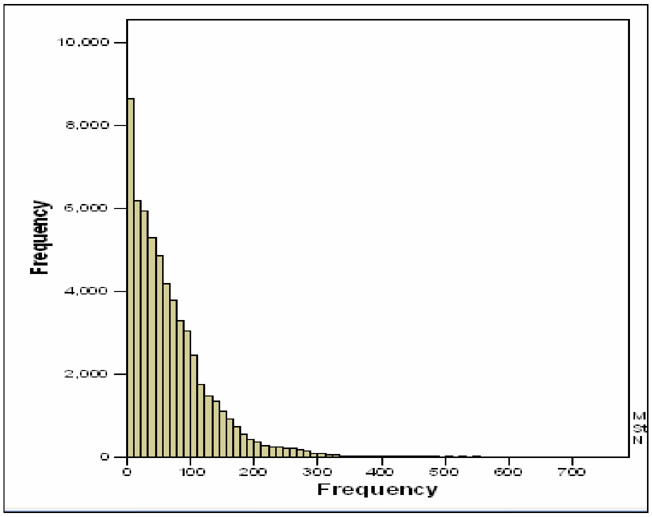


Figure 8 Frequency diagram of Frequency variable

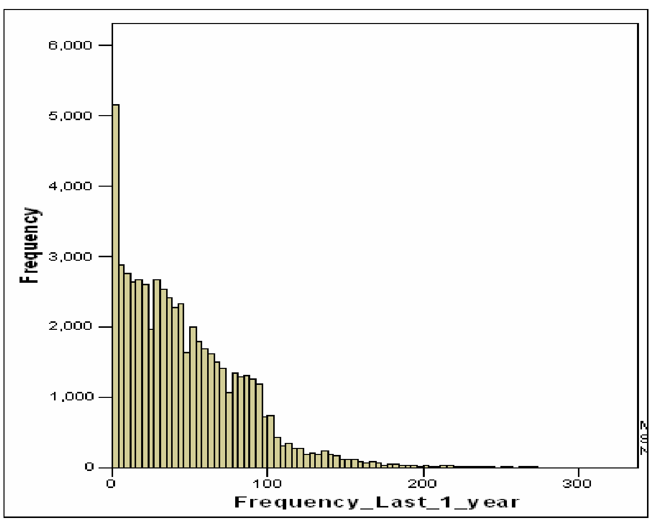


Figure 9 Frequency diagram of Frequency Last One Year Variable

rFrequency:

rFrequency indicates the purchase frequency of the customers relative to their length of relationship. Again the variable is right skewed. However, the standard deviation is considerably smaller than the mean (0.7 of it) which points a relatively less dispersed distribution. As it can be seen from the frequency diagram, most of the cases are accumulated around the first and second quartile.

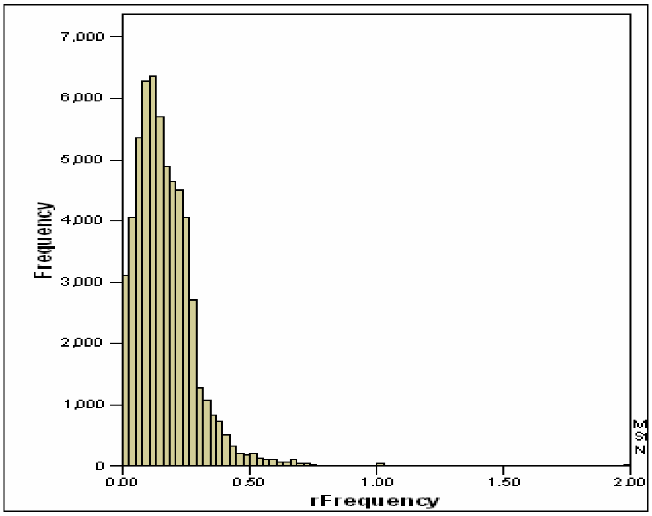


Figure 10 Frequency diagram of rFrequency variable

Total sales for four years, Average sales for four years:

These two variables are closely related to each other therefore will be discussed together. The distribution of these variables has the same specialties, for all variables discussed above. When the Frequency diagrams in Figure 11, Figure 12 analyzed it is noticed, that values are accumulated mostly around the first quartile of the variables.

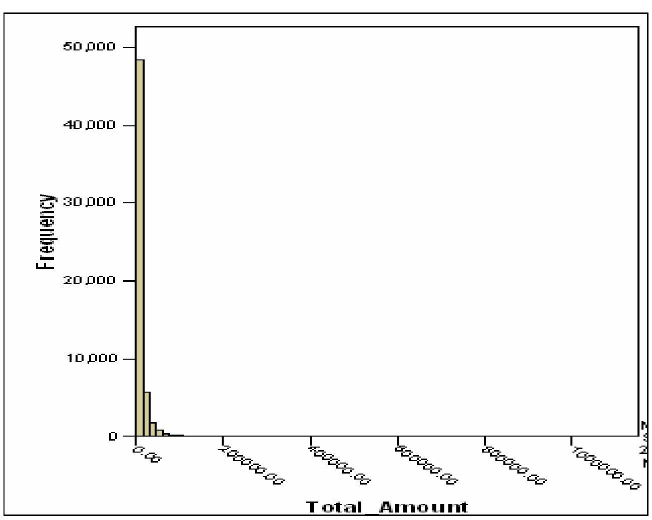


Figure 11 Frequency diagram of Total Amount variable

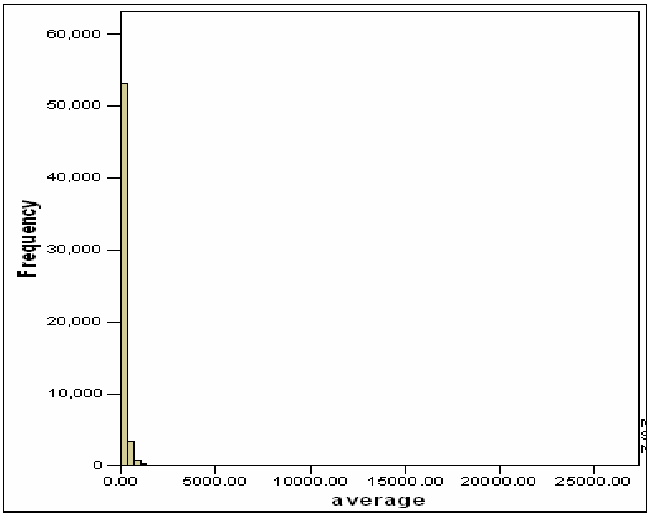


Figure 12 Frequency diagram of Average Sales variable

Recency

When the descriptive scores of the Recency variable are analyzed it is obvious that the dispersion of the variable is high because of the outliers again.

Another finding is that the value for the third quartile is just fourteen. This means that seventy five percent of the customers have at most two weeks between their last two purchases. This fact reconfirms that the mean reason of the dispersion is the existence of outliers. When the frequency diagram is analyzed it can be seen that the data are mostly accumulated around the second quartile which is at the same time the median value.

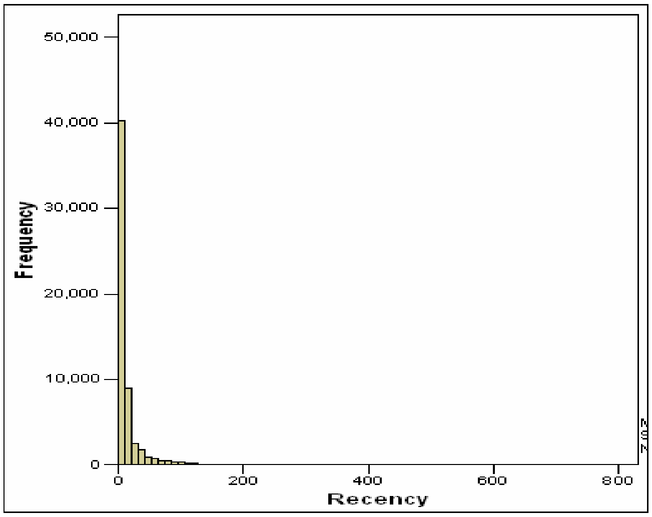


Figure 13 Frequency diagram of Recency variable

Standard Deviation Recency and CV Recency

These two variables are closely related to each other and therefore will be discussed together. Both of the variables are right skewed and highly cumulated in the value zero, i.e. the mode is zero. Therefore, these variables have the strange distribution that can be seen from the histograms below. The reason of observing high number of cases taking the value of zero depends on the distribution of the variable frequency. Since the mode of frequency is one, which means that there are many customers who purchased only once, then the standard deviations and therefore coefficient of variations of the durations between purchases for these customers are zero.

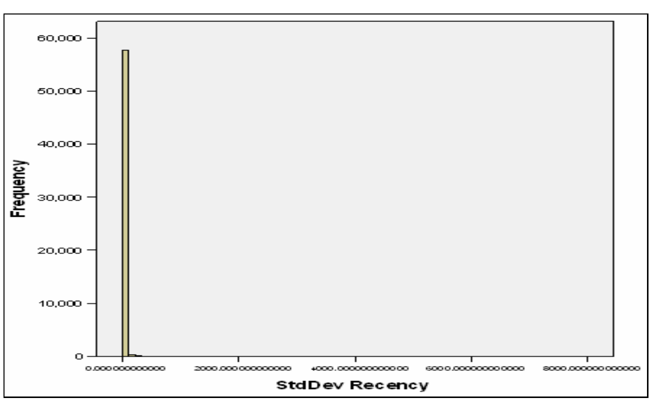


Figure 14 Frequency Diagram of Standard deviation Recency variable

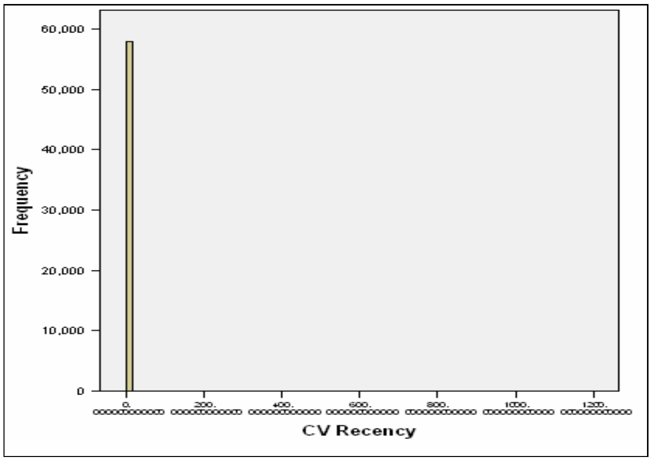


Figure 15 Frequency Diagram of Coefficient Variance of Recency variable

rTotal Amount, rAmount and Standard Deviation Amount

Analysis show that same as the variables “Standard Deviation Recency” and “CV Recency”, these three variables have right skewed distributions and modes of zero. The reason of high number of cases taking values around zero for these variables is again because of the stem variable “total amount”. The distribution of “total amount” was highly right skewed indicating that there are many customers who purchased low amounts.   The histogram of these variables can be seen in Figure 16, Figure 17 and Figure 18.

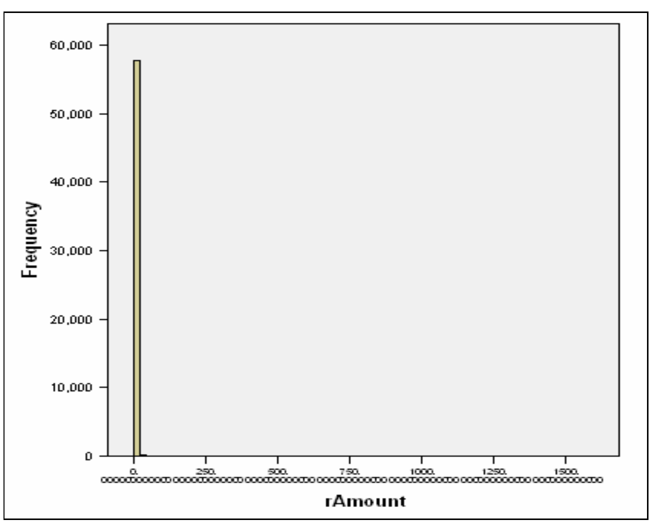


Figure 16 Frequency Diagram of rAmount variable

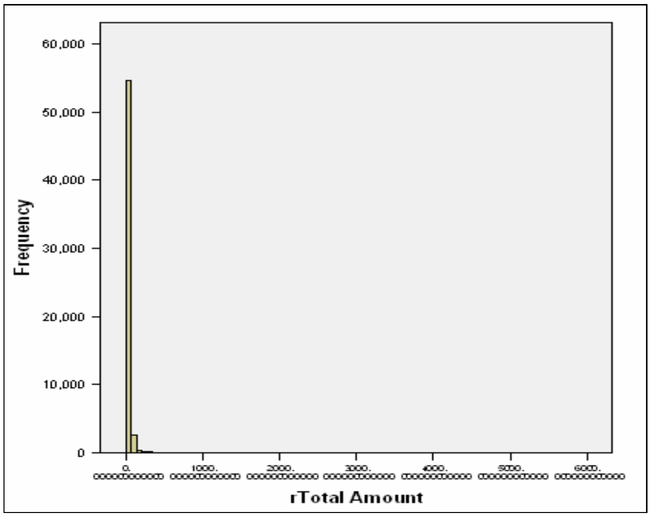


Figure 17 Frequency Diagram of rTotal Amount variable

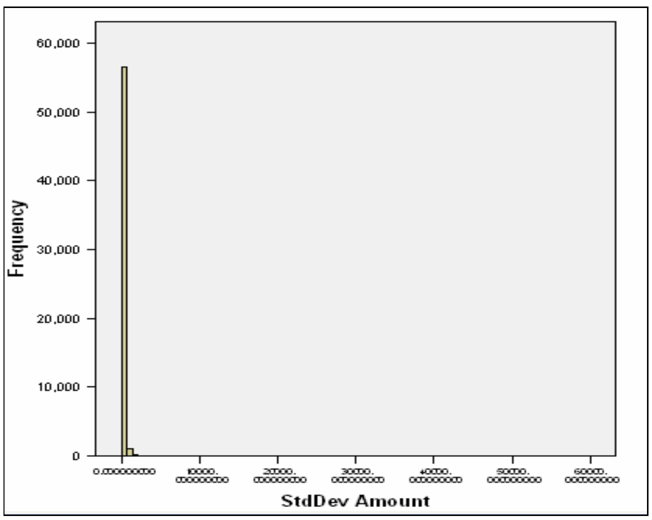


Figure 18 Frequency Diagram of Standard Deviation of Amount variable

<u>rMajorTrip</u>

Table 5 reveals the descriptive scores of this variable. When we compare the mean of this variable with the mean of total number of purchases namely frequency variable, the mean of “rMajorTrip” is slightly greater than the half of the mean of frequency variable. This shows that there are purchases with very few purchase volume which pulls the average sales volume down.

The median and mean of the variable is very close to each other and the standard deviation is 0.4 of the mean. These observations indicate that “rMajorTrip” has a distribution very close to normal distribution. This conclusion is also supported by the histogram of this variable presented in Figure 19: Except the first frequency category, cases that have “rMajorTrip” values less than one, the distribution of the variable is approximate to the normal distribution. The high frequency in first category represents customers who purchase very regularly the same volume of purchase in each transaction. Therefore, such cases have zero or very few number of purchases that exceeds the average volume of purchase.

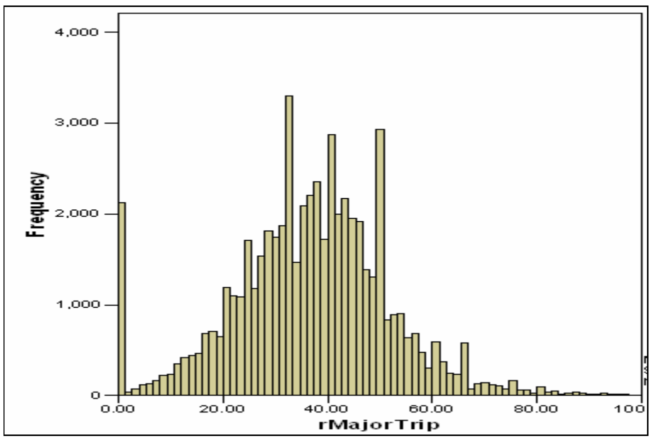


Figure 19 Frequency diagram of rMajorTrip variable

- Variables by year

In Table 6, the descriptive of the variables; frequency, total sales, average sales and IPT is given but broken down by years.

The number of purchases in 2003 decreased compared to 2002 but again in 2004 the number of purchases reached a higher value even than 2002. A similar pattern is observed for the total number of sales: a decrease in 2003 but an increase in 2004. However the total sales in 2004 could not exceed the sales volume in 2002. As a consequence of the trends in frequency and total sales variables, the average sales follow a monotonically decreasing pattern: each year, the average number of sales per purchase decreased.

With regards to "IPT", the same pattern as the total sales is observed: a decrease in 2003 and an increase in 2004, but 2004 values are on average smaller than 2002 values.

Together with the above observations, it can be concluded that a decrease in sales volume as well as the number of purchases is observed in year 2003. However, this decrease is recovered in 2004.

Table 6 Statistics for Variables by Year

	Frequency 2002	Frequency 2003	Frequency 2004	Total Amount 2002	Total Amount 2003	Total Amount 2004	Average Amount 2002	Average Amount 2003	Average Amount 2004
Valid	5720	32217	56307	5720	32217	56307	5720	32186	56301
Missing	52213	25716	1626	52213	25716	1626	52213	25747	1632
Mean	39.5	24.7	54.47	7729.15	4557.95	6982.9	203.2	158.45	136.19
Median	34	13	44	3679.2	1056	3147.5	113.1	78.71	70.32
Mode	1	1	1	24	24	24	24	24	12
Std. Deviation	32.227	27.791	46.903	13156.78	15008.57	14156	313.6	301.51	259.08
Range	260	267	687	361156.4	687905.1	572481	7082	15957	16192.24
Minimum	1	0	0	0	0	0	0	0	0
Maximum	261	267	687	361156.4	685357	572481	7082	15593	16192.24
Percentiles 25	4	21	1000.2	276	27.78	1080	56.64	38.65	50.0
Percentiles 75	38	77	9764.18	3828.48	46.15	7997.8	230.4	172.34	412.8

- Variables at the City Level

Table 7 shows the descriptive scores of the variables at the city level. For all of these variables, again, mean is greater than median and median is greater than mode. This shows the skewness of the distribution of these variables. All of the variables, except "Count of Customers City" and "Sales per Customer City", standard deviation is smaller than the mean. These two variables "Count of Customers City" and "Sales per Customer" are the two variables which have very high standard deviations and therefore they are highly dispersed. This issue can also be observed from the histogram of city level variables. Although all of them are right skewed, almost all of them approximate to the normal distribution but the variables "Count of Customers

City" and "Sales per Customer City" have a distribution far from normal; for both of the variables, cases are accumulated around the first and second percentiles.

Table 7 Descriptive Statistics of the Variables at the City Level.

	Count of Customers City	*Average Frequency City*	*Average Freq Last Year City*	*Average Recency City*	*Average Sales2 City*	*Average IPT City*	*Sales per Person City*
Valid	78	78	78	78	78	78	78
Missing	0	0	0	0	0	0	0
Mean	743.3	57.4	50.4	17.2	181.4	11.3	4.4
Median	156	54.3	48.2	15.6	149.6	10.7	2.4
Mode	1	1	1	0	69.0	0	0.0
Std. Deviation	1927.2	23.7	19.4	9.1	132.0	5.3	5.3
Range	13639	118.4	98.7	57.8	771.0	36.0	23.9
Minimum	1	1	1	0	69.0	0	0.0
Maximum	13640	119.4	99.7	57.8	840	36.0	23.9
Percentiles 25	42.6	38.8	12.6	306020.2	8.1	0.9	36
Percentiles 75	74.6	60.5	20.2	3255695.2	13.1	6.0	146.42

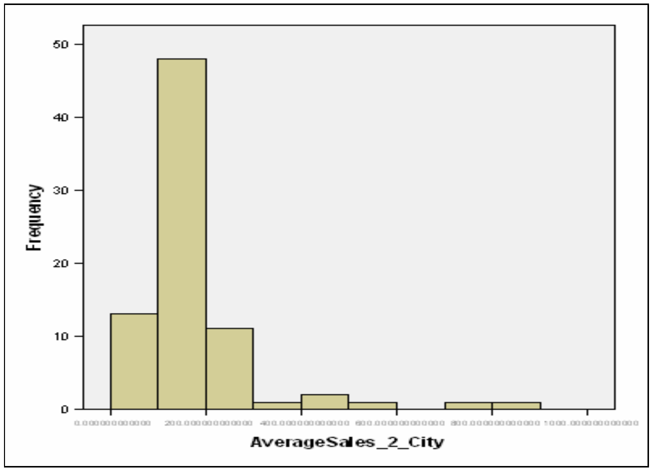


Figure 20 Frequency Diagram of Average Sales City Variable

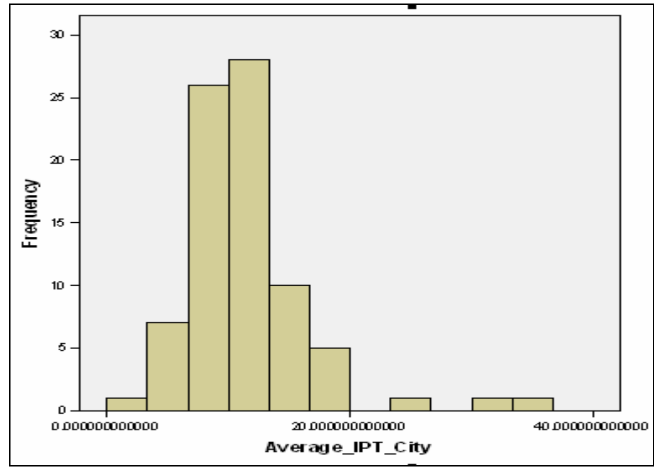


Figure 21 Frequency Diagram of Average IPT City Variable

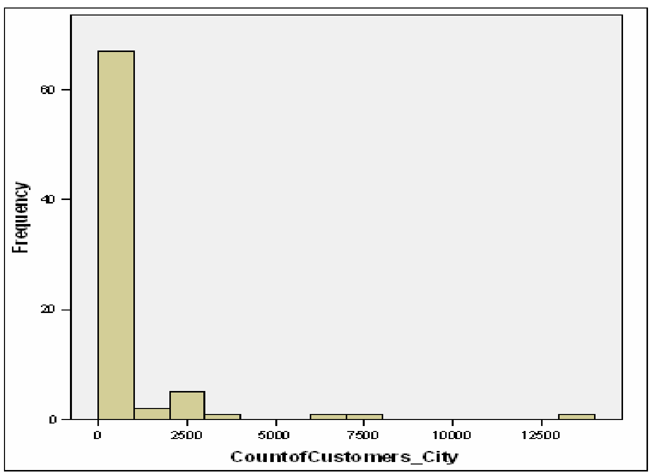


Figure 22 Frequency Diagram of Count of Customers City Variable

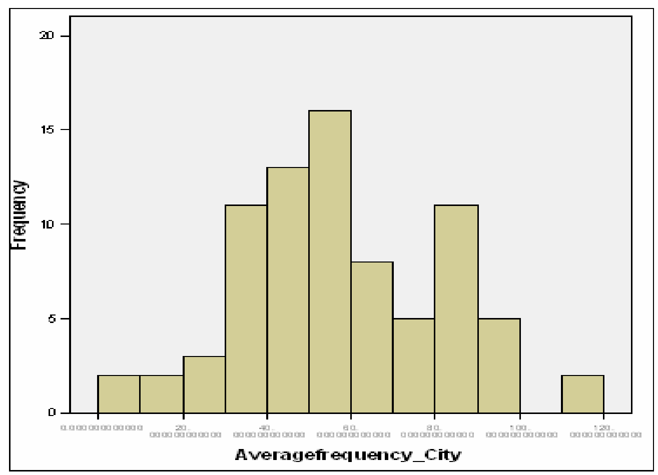


Figure 23 Frequency Diagram of Average Frequency City Variable

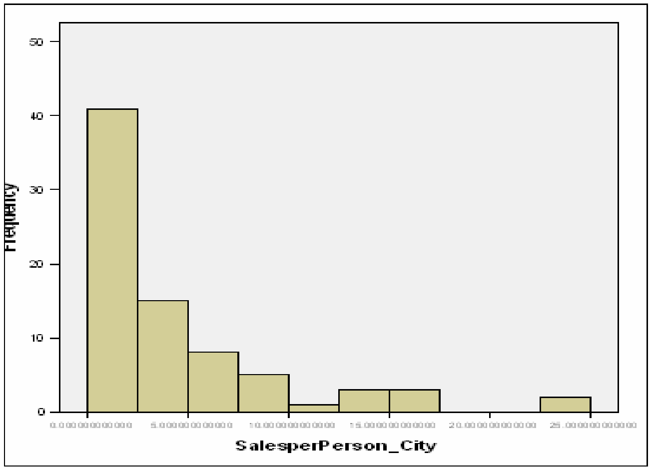


Figure 24 Frequency Diagram of Sales per Customer City Variable

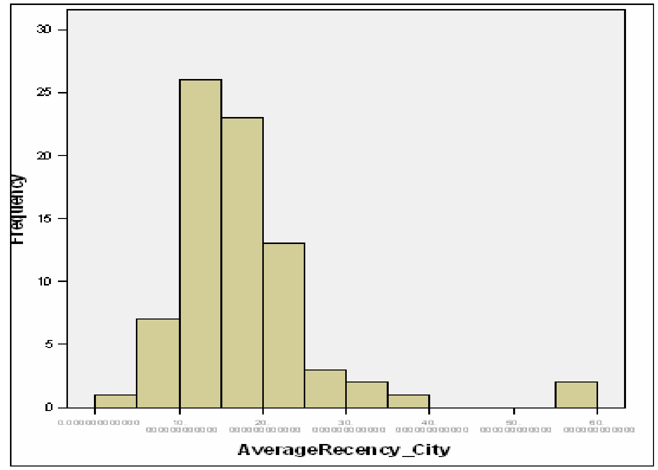


Figure 25 Frequency Diagram of Average Recency City Variable

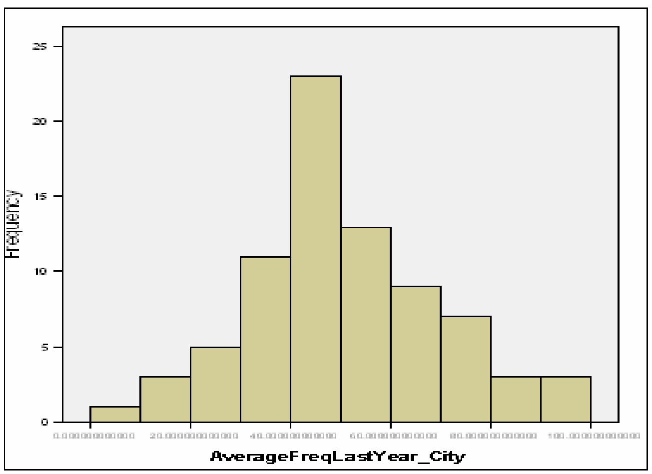


Figure 26 Frequency Diagram of Average Frequency Last Year City Variable

CHAPTER 5

FACTOR ANALYSES FOR VARIABLE SELECTION

Factor Analysis is an explorative statistical method used to define the underlying structure in a data matrix in order to reduce number of data in the original dataset or number of variables that define it. This method analyzes the structure of the interrelationships (correlations) among a large number of variables by defining a set of common underlying dimensions, known as factors (Hair et al., 1995).

Factor Analysis can be used either to reduce the number of data in the original dataset or number of variables in it. In both cases Factor Analysis aims to summarize the information of original dataset with minimum loss of information.

Types of Factor Analysis

R-Type Factor Analysis: Factor analysis type which aims to summarize the characteristics that define the dataset by identifying underlying dimensions.

Q-Type Factor Analysis: Factor analysis type which aims to summarize the individual respondents based on their characteristics. Cluster analysis generally preferred instead of Q-Type Factor Analysis because of its computational difficulties.

Steps of Factor Analysis

Factor analysis, in any application of it, is applied by following the steps shown in Figure 27.

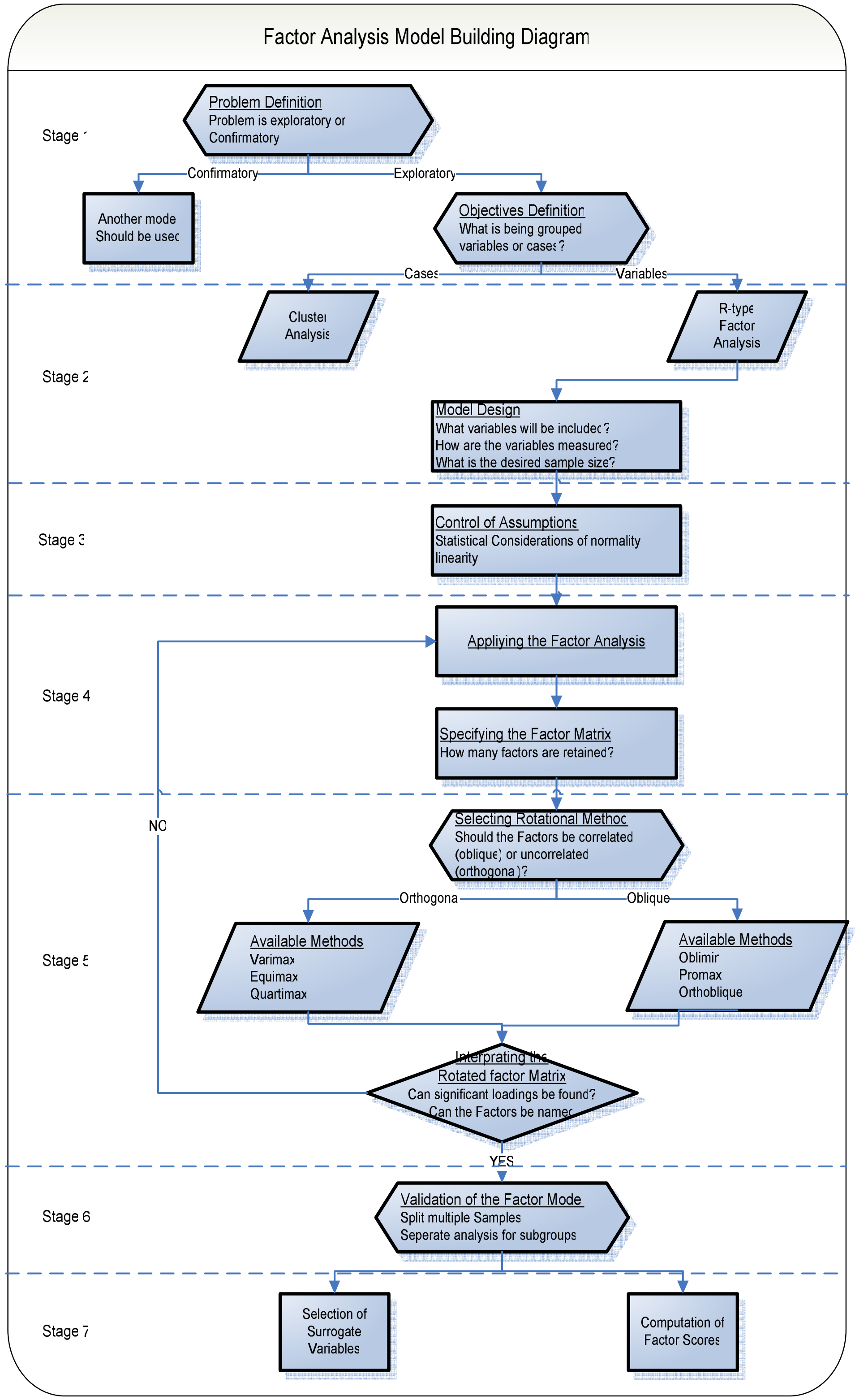


Figure 27 Factor Analysis Model Building Diagram

Stage One: Factor Analysis Problem Definition

Objectives of Factor Analysis

Hair et al. (1995) summarizes the objectives of factor analysis as follows:

1. Identify the structure of relationships among variables by examining correlations between them.

2. Identify representative variables from a much larger set of variables for use in subsequent multivariate analysis.

3. Create an entirely new set of variables, smaller in number, to partially or completely replace the original set of variables for subsequent techniques.

Stage Two: Factor Analysis Design

Data Characteristics for Factor Analysis

Hair et al. (1995) argues that the sample that will be used for the Factor Analysis should not be smaller than fifty cases and for better results it should be larger than one hundred cases. As a general rule the minimum sample size should be at least five times greater than the number of the variables. Appropriate variables for the factor analysis should not be categorical ones rather they should be at interval or ratio level. In addition, the measures of variables being analyzed should have the same scale. For example with a dataset that contains two variables in scales of days and amount factor analysis cannot be applied. In order to come up with comparable measures of the variables in the dataset, z-scores of these variables should be computed.

Stage Three: Controlling the Assumptions of Factor Analysis

Factor analysis is a data reduction technique that relies upon the fact that the variables are empirical indicators for some common underlying dimensions. Based on this fact basic assumption of the factor analysis is; variables in the analysis should be sufficiently correlated with each other. (Hair et al., 1995) In addition to this assumption, a dataset can be accepted as appropriate for factor analysis if the Bartlett Test of Sphericity and Kaiser Meyer Olkin correlation matrix measures catch the limits.

- <u>Bartlett Test of Sphericity</u>

A statistical test for the existence of correlations among variables. Ledakis (1999) argues that, Bartlett's test of sphericity tests the Null hypothesis, which states that variables in correlation matrix are not related. As the value of the test increases and the associated significance level decreases, the likelihood increases that the Null hypothesis can be rejected and the alternative hypothesis accepted (i.e., the variables that constitute the correlation matrix are related). In contrast, as the value of the test decreases and the associated significance level increases, the likelihood that the Null hypothesis is true increases and, in turn, the alternative hypothesis must be rejected. If the significance level of this test, which is calculated by statistical tool, is greater than 0.10 it means that the dataset is not suitable for the Factor Analysis.

- <u>Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO)</u>

Another statistical procedure for determining the suitability of the dataset for factor analysis. The KMO is an index for comparing the degree of the observed correlation coefficients to the degree of the partial correlation coefficients in the dataset. Partial correlation exists between two variables when the added effects of other variables on the correlation have been eliminated (Ledakis, 1999). KMO index can has value between zero and one. As the value of the index increases, the suitability of the dataset for factor analysis increases, too. Generally, this measure must be above 0.5, and values higher than 0.8 are preferred.

Stage Four: Applying the Factor Analysis and Specifying the Factor Matrix

Criteria for the Number of Factors to be Extracted

In factor analysis, optimum number of factors is determined by using some empirical guidelines rather than exact quantitative solutions. In most of the analysis one of the following criterions is used to decide the number of factors to extract.

- Latent Root Criterion:

Latent root criterion based on the fact that an underlying dimension of the dataset can be named as factor only if it should account for variance of at least a single variable. Since each single variable contributes a value of one to the eigenvalues, only factors whose Eigenvalues greater than one are considered significant. This criterion is accepted reliable if the number of variables is between twenty and fifty. Otherwise there is a tendency to extract too few or more factors.

- A Priori Criterion

A Priori Criterion is used when how many factors to extract have already known before executing the factor analysis.

- Percentage of Variance Extracted

Percentage of variance criterion aims to select the factors which explain at least a specified amount of variance that ensures these are significant factors for the analysis. Although there is not an absolute threshold adopted for all applications, in the natural sciences the factoring procedure usually should not be stopped until the extracted factors account for at least ninety five percent of the variance or until the last factor accounts for only a small portion. (Less than five percent) On the other hand, in the social sciences sixty percent of the total variance is accepted as a satisfactory solution. (Hair et al., 1995)

- Scree Test Criterion

The Scree test is used to identify the optimum number of factors that can be extracted before the amount of specific variance begins to dominate the common variance structure (Hair et al., 1995). The Scree Plot diagram shows the number of factors with their relative eigenvalues. In Scree test criterion the shape of the resulting curve is analyzed to determine the maximum number of factors for the analysis. This number is indicated by the first point the curve begins to flatten.

Stage Five: Interpretation of Factors

Three steps are followed to interpret the factors.

1. Analyzing the initial un-rotated factor matrix
2. Employing a rotational method
3. Interpreting the rotated factor matrix

- Analyzing the initial un-rotated factor matrix

Initial un-rotated matrix is analyzed to determine number of factors that will be extracted. However, in most cases factor loadings shown in un-rotated factor matrix do not provide adequate information to significantly distribute variables to the factors. Hair et al. (1995) defines factor loadings as the correlation between each variable and the factor, which shows the correspondence between them. The higher loadings make the related variable representative of the factor among all variables loaded on it. Un-rotated factor solutions extract the factors according to their importance. The first factor accounts for the largest amount of variance and subsequent ones accounts for smaller portions of it.

- Employing a Rotational Method

Since in most cases it is not possible to distribute the variables among factors with information in un-rotated factor loading matrix, factor rotational methods are employed to achieve adequate information for interpretations. When implementing the rotation of factors, the reference axes of the factors are turned about the origin until some other position has been reached. Since the un-rotated factor solutions extract factors in the order of their importance and gives the significant amount of variance to the first factor, subsequent factors are extracted based on the residual amount of variance. The ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern. (Hair et al., 1995)

There are two main types of rotation named as orthogonal factor rotations and oblique factor rotations.

- Orthogonal Factor Rotations:

In orthogonal factor rotations the axes are maintained at ninety degrees. The objective of this method is to maximize variable’s loading on a single factor or to make the number of high loadings as few as possible. Figure 28 demonstrates the orthogonal factor rotation.

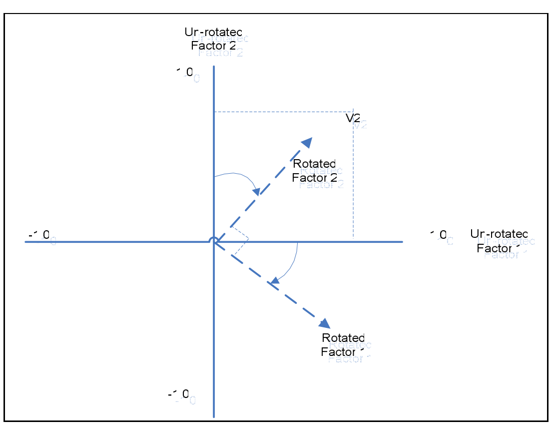


Figure 28 Orthogonal Factor Rotation

Three major orthogonal approaches have been developed: Quartimax, Varimax and Equimax.

- Oblique Factor Rotations:

When the axes are rotated without retaining the ninety degree angle between the reference axes the rotational procedure is called as oblique rotation. Figure 29 demonstrates the oblique factor rotations.

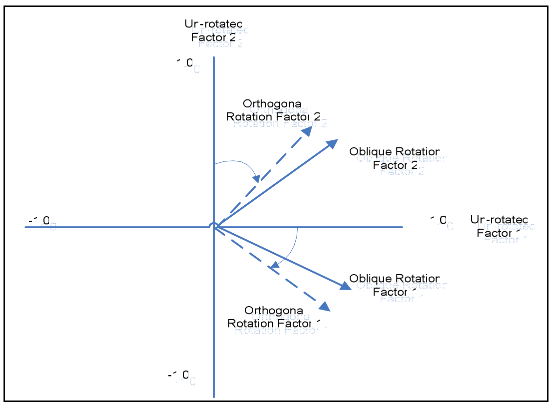


Figure 29 Oblique Factor Rotation

There is not an analytical reason to favor one rotational method over another. The choice should be made on the basis of the particular needs of the problem in hand as well as the ability of the statistical program that is being used.

- Interpreting the rotated factor matrix

Following steps should be followed when analyzing the factor matrix.

1. Examining the Factor Loading Matrix:

Hair et al. (1995) defines factor loadings as the correlation between each variable and the factor, which shows the correspondence between them. When analyzing the Factor Loading Matrix for each factor all variables should be examined to find the highest loading for that variable. When the highest loading (largest absolute factor loading) is found significance level for this variable should be tested. In order to consider the factor loadings as significant three different methods can be used:

▪ First method based on some practical information used by the analysts. According to this; factor loadings greater than ±.30 are considered to meet minimum level; loadings of ±.40 are considered more important; and if the loadings are ±.50 or greater, they are considered practically significant (Hair et al., 1995).

▪ Different from the first method second one determines the significance level according to the sample size of the analysis. Table 8 shows the minimum difference that should be between the highest factor loading (highest absolute value) and the second one in order to accept this as significant with the corresponding sample sizes (Source: Hair, et al., 1995).

Table 8 Guidelines for Identifying Significant Factor Loadings Based on Sample Size

Factor Loading	*Sample Size Needed for Significance*
0.30	350
0.35	250
0.40	200
0.45	150
0.50	120
0.55	100
0.60	85
0.65	70
0.70	60
0.75	50

▪ Third method focuses on the disadvantage of preceding methods such as not considering the number of variables being analyzed. According to this method as the number of variables being analyzed increases, the acceptable level for considering a loading as significant decreases.

2. Examining the Communalities Matrix:

After the rotated factor loading matrix is analyzed, variables that do not load on any factor should be determined. To achieve this additional to analyze the factor loading matrix, communalities matrix should also be analyzed. Hair, et al. (1995)

indicates that communalities for each variable represent the amount of variance accounted for by the factor solution for each variable. If the communality values for the variables do not meet acceptable levels of explanation, specified variables should be eliminated from the analysis dataset.

3. Naming the Factors

As the last step of the Factor Matrix interpretation, factors should be named according to the pattern of factor loadings.

Stage Six: Validation of Factor Analysis

Validation of factor analysis has two main parts:

▪ Assessing the degree of generalizability of the results to the population. Aim of this validation method is to confirm the results of the analysis by evaluating the consistency of results with the ones coming from small samples. Sampling may be achieved by splitting the original dataset or creating a separate one.

▪ Detection of effect of outliers. Aim of this validation is to assess the impact of outliers on the results of the analysis. In order to achieve this validation factor analysis should be applied with and without observations identified as outliers. If the ineffectiveness of the outliers' existence is justified, the results should have greater generalizability.

Stage Seven: Additional Uses of the Factor Analysis Results

Additional uses of factor analysis results include the computation of factor scores as well as selection of surrogate variables for subsequent analysis with other statistical techniques. According to the objective of the analysis in hand one of these can be used. Available additional use methods and objectives of these methods are as follows:

- Selecting Surrogate Variables

If the objective of the analysis is to identify appropriate variables for subsequent application with other statistical techniques, the factor matrix is examined and the variable with the highest factor loading is selected on each factor as a surrogate representative for that particular factor (Hair et al., 1995).

- Using Factor Scores

Factor scores are computed when the objective is to create a new and smaller set of variables to replace the original dataset. Hair et al. identifies factor scores as composite measures for each factor that contains the affect of each variable in it with respect to their loadings. As a result of this factor scores represent a composite of all variables loading on the factor, when surrogate variables represent only a single variable. However, a disadvantage of factor scores is that they are based on correlations with all the variables in the factor (Hair et al., 1995).

Factor Analysis to Define Variables of Customer Segmentation Analysis

Through the data collection and data preparation phases 27 characteristics of customers have been identified and calculated as variables that will be used to cluster them into smaller groups. Variables of the analysis are listed in Table 4.

As noted before in this study Recency-Frequency-Monetary method is used to determine the valuable customers with some extensions. As a result of this, these three variables are selected as the main variables of the analysis. In order to define the additional variables that will be used for analysis and their sequence, rather than adding them as separate variables, more general evaluative dimensions are used in the analysis. Factor Analysis is used in this study to identify the underlying evaluative dimensions of data.

Factor analysis is applied by following the steps of “Factor Analysis Model Building Procedure” shown in Figure 27. The steps of the model with corresponding results are summarized in Table 9. Detailed explanation regarding to the analysis steps can be found in the following sections.

Table 9 Summarized Results of Factor Analysis

Analysis Step		*Step Description*	*Expected Solution*	*Result for the Analysis*
1		Objectives Definition	Reduce the 27 variables to a smaller number	
2		Factor Analysis Method Selection	R-Type Factor Analysis	
3		Factor Analysis Assumptions Control		
	3.1	Control of correlations between variables	Shows the correlation level between the variables in the dataset. If there is not significant correlation between the variables this means that the dataset is not suitable for the Factor Analysis	74 of the 105 correlations are significant at the 0.01 level means that more than %70 of correlations is significant. The dataset is suitable for Factor Analysis
	3.2	Analyze of Kaiser Meyer Olkin Measure of Sampling	Adequacy value shows the total variance shaped by all variables. If the value is closer to 1 it shows that the data is suitable for Factor Analysis	Value for the case: 0.764. So the dataset is suitable for the Factor Analysis
	3.3	Test of Bartlett's Sphericity	Shows whether there is a correlation between the variables. If the significance level for dataset is greater than 0.10 it means the dataset is not suitable for the Factor Analysis	Value for the case: 0. So the dataset is suitable for the Factor Analysis
4		Factor Analysis Application		
	4.1	Number of factors Selection	Eigenvalues. When the eigenvalue is greater than 1 it means that the factor has contribution greater than the variable by itself.	Analysis returns 5 components with eigenvalues greater than 1. Total contribution for the solution is %76 which is an adequate value. Since the last factor contributes less than 5 % the factoring procedure has stopped there.
	4.1.1	Scree Plot	Scree plots diagram contains the information regarding the possible factors and their relative exploratory power as expressed by their eigenvalues.	Scree plot shows that first Five Factors have eigenvalues greater than 1.
5		Factors Interpretation		

Analysis Step		*Step Description*	*Expected Solution*	*Result for the Analysis*
	5.1	Analyze of Factor Loadings for un-rotated solution		Variables are not significantly loaded to the factors.
	5.2	Employing the rotation method	Achieve simpler and theoretically more meaningful factor solutions.	Orthogonal – Varimax Factor Rotational method is employed
	5.3	Analyzing the Communalities Matrix	Shows for every variable the contribution to the overall variance build by the model. Smaller values shows that the variable does not have so much contribution to the model. Ones that have Extraction values smaller than 0.50 will not be included in the mo	None of variables in the dataset has communality value less than 0.50 which certifies inclusion of all variables in the further analysis
	5.4	Interpretation of Rotated Factor Loading Matrix	For each variable factor loads will be analyzed and the greatest ones will be selected. There must be at least 0.10 difference between two factor loads in order to designate the variable to a factor.	
	5.5	Naming The Factors	Variables assigned to the Factors are analyzing and according to their characteristics names of the factors are given.	According to the common characteristics of the variables assigned to the factors, they are named as Amount, Recency, Frequency, LoR and Other
6		Validation of factor Analysis	Validation is achieved to assess the generalizability of the results to the population	Validation of factor analysis is achieved by splitting the original dataset into two samples and applying the same analysis to them
7		Surrogate Variables Selection	Identifying appropriate variables for subsequent application with other statistical techniques	Based on literature Recency, Frequency and Amount are selected as the base variables for the further statistical techniques. Factor Analysis approved this by calculating the highest factor loadings for these variables in Factor-1, facor-2 and Factor-3. Length of Relationship-1 and rMajorTrip are selected as surrogate variables from Factor-4 and Factor-5 by having the highest factor loadings in these factors.

Stage One: Factor Analysis Problem Definition

Objectives Definition

The objective of the Factor Analysis that performed in this study is to identify the structural relationships among variables with the aim of grouping large numbers of variables into a smaller number of homogenous sets and identifying representative variables for use in clustering analysis. If the 27 variables can be summarized in a smaller number of variables, then clustering analysis can be made in a more effective manner.

Stage Two: Factor Analysis Model Design

Selecting the Factor Analysis Method

In this analysis to define the underlying relationships between variables R-type Factor analysis will be used which focuses on summarizing the characteristics.

Data Characteristics for Factor Analysis

Regarding the adequacy of the sample size, in this analysis there are 57979 cases. This value is 2750 times greater than the number of variables. The specified ratio shows that the sample size is adequate for getting reasonable results from factor analysis. In addition to this, sample size provides and adequate basis for the calculation of the correlations between variables. None of the variables used in this analysis are categorical ones. In order to come up with comparable measures of the variables in the dataset, z-scores of these variables computed before the analysis is applied.

Stage Three: Control of Assumptions

Assessing the factorability of the correlation matrix

In order to identify statistically significant variables correlation analysis is applied to the dataset in hand. Results of the correlation analysis are shown in Table

10. Examination of the results shows that seventy four of the one hundred and five correlations are significant at the 0.01 level. The corresponding significance level for this value is seventy percent which provides an adequate basis for proceeding to the other controls for assumptions.

Table 10 Correlations Among Variables

Correlations

	LOR_1	LoR2	Frequency	Frequency last one year	rFrequency	Recency	IPT	Standard Deviation of Recency	CV Recency	Amount	Total Amount	Std Dev Amount	RAmount	RTotal Amount	RMajorTrip
LOR_1	1	.217**	.66**	.33**	0.01	-0.00	-.02**	03**	.02**	.1**	.34**	.12**	-.08**	.08**	0.00
LoR2	.22**	1	.16**	.09**	-.01*	01*	.02**	04**	.06**	.03**	.08**	.02**	-.02**	.04**	0.00
Frequency	.66**	.157**	1	.82**	.59**	-.19**	-.29**	-.11**	.01*	.03**	.41**	.07**	-.05**	.21**	.02**
Freq last 1 year	.33**	.092**	.82**	1	.72**	-.21**	-.33**	-.13**	0.01	-.02**	.26**	.01*	-.06**	.18**	.03**
rFrequency	0.01	-.011*	.59**	.72**	1	-.24**	-.37**	-.15**	0.00	-.04**	.17**	-.01**	.17**	.27**	.03**
Recency	-0.00	.012*	-.19**	-.21**	-.24**	1	.59**	2**	.02**	.01*	-.07**	.01*	-0.01	-.07**	0.00
IPT	-.03**	.019**	-.29**	-.33**	-.37**	59**	1	25**	0.00	.02**	-.1**	.02**	-0.00	-.10**	0.00
Standard Deviation of Recency	.03**	.041**	-.11**	-.13**	-.15**	2**	.25**	1	.92**	0.01	-.04**	.01**	-0.01	-.04**	-0.00
CV Recency	.02**	.065**	.01*	0.00	0.01	02**	0.00	92**	1	-0.00	0.00	0.00	-0.01	-0.00	0.00
Amount	.01**	.032**	.03**	-.02**	-.04**	01*	.02**	0.01	-0.00	1	.59**	.85**	.2**	.62**	0.00
Total Amount	.34**	.084**	.41**	.26**	.17**	-.07**	-.1**	.04**	0.00	.6**	1	.49**	.02**	.72**	0.01
Std Dev Amount	.12**	.025**	.06**	.01*	-.01**	01*	.02**	01**	0.00	.85**	.49**	1	.15**	.49**	-0.01
R Amount	-.07**	-.019**	-.05**	-.05**	.17**	-0.01	-0.00	0.01	-0.00	.19**	.02**	.15**	1	.43**	0.00
R Total Amount	.08**	.038**	.20**	.18**	.27**	-.07**	-.10**	.04**	-0.00	.62**	.72**	.49**	.43**	1	.01*
rMajorTrip	0.00	0.002	.02**	.03**	.03**	0.00	0.00	0.00	0.00	0.00	0.01	-0.01	0.00	.01*	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Other controls of the assumptions are Bartlett Test and Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO). The dataset for this analysis can be accepted as appropriate for Factor Analysis because it catches the limits for the following correlation matrix measures.

- Bartlett Test of Sphericity

As it is shown in Table 11, the significance level of this test is 0. If this value is greater than 0.10 it means that the dataset is not suitable for the Factor Analysis. Since it is smaller than 0.10 dataset is accepted as suitable for the factor analysis.

Table 11 Results for Bartlett Test of Sphericity and KMO Index

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.764499927
Bartlett's Test of Sphericity	Approx. Chi-Square	192133.2791
	Df	351
	Sig.	0

- Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO)

As it is shown in Table 11, the value for this statistic is 0.76. Since the value is greater than 0.5 the dataset is accepted as suitable for factor analysis.

Stage Four: Applying the Factor Analysis and Specifying the Factor Matrix

Factor analysis is applied by using SPSS for Windows statistical tool in this study. The characteristics selected for the applied Factor Analysis are summarized in Table 12.

Table 12 Characteristic of Applied Factor Analysis

Factor Analysis Characteristics	*Selected Characteristics for this Analysis*
Factor Extraction Method	Principal Components
Eigenvalues that will be extracted	Ones with value over 1 (One)
Rotation Method	Varimax
Missing Values	Not applicable in data

Selecting the Number of Components:

Factors that are representing the underlying dimensions in the original dataset are extracted by using the principal component analysis. In order to determine the number of Factors that will be used in the analysis Percentage of Variance Extracted, The Latent Root Criterion and Scree Test Criterion are employed. Table 13 shows the information regarding the twenty seven possible factors and their explanatory power as expressed by their eigenvalues and percentage of variance. Latent root criterion considers the factors as significant only if their eigenvalues are greater than one (Hair et al., 1995). According to this criterion, five factors are extracted from the dataset. On the other hand the Scree Test represented in Figure 30, which accepts the first point the curve begins flatten as the maximum number of factors to be extracted, does not support the result of the latent root criterion. The test indicates that three or four factors may be appropriate for this analysis. When the eigenvalues for the forth and fifth factors are examined, it is determined that they are greater than one which certifies their inclusion in the further analysis. Since the following factors have eigenvalues smaller than one, the factoring procedure is stopped at five factors.

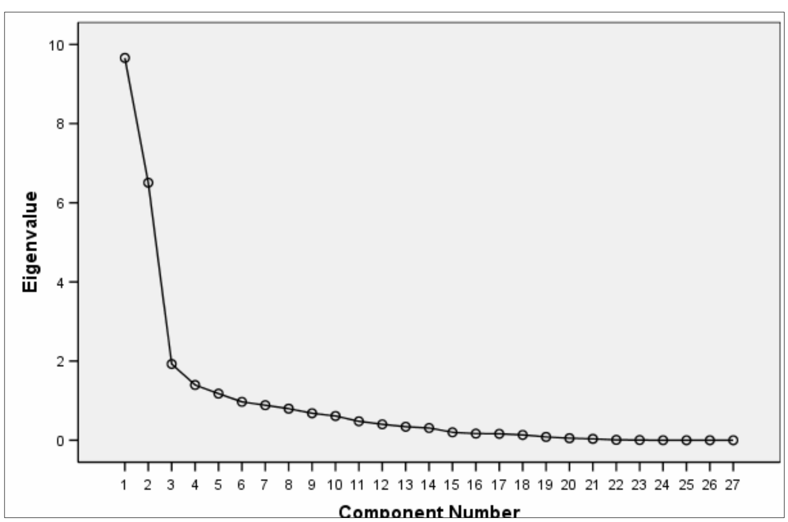


Figure 30 Scree Test Curve

In addition to controlling the eigenvalues, Percentage of Variance Extracted criterion can also be used to determine the number of factors. Based on the general acceptations, if the model explains the sixty percent of total variance, it is accepted as a satisfactory solution. According to this assumption five factors are extracted by SPSS which accounts for 76% of total variance. Since the latest factor accounts for only a small portion of total variance with 4.36%, factoring procedure stopped at the fifth factor. By combining the results of these three criteria five factors are extracted from the dataset for further analysis.

Table 13 Results for the Extraction of Component Factors

Total Variance Explained									
Component	*Eigenvalues*			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	9.658288331	35.77143826	35.77143826	9.658288331	35.77143826	35.77143826	9.095229117	33.68603377	33.68603377
2	6.506743606	24.09905039	59.87048866	6.506743606	24.09905039	59.87048866	5.414838924	20.05495898	53.74099274
3	1.926432045	7.134933501	67.00542216	1.926432045	7.134933501	67.00542216	3.144298086	11.64554847	65.38654121
4	1.396697623	5.172954158	72.17837632	1.396697623	5.172954158	72.17837632	1.780307391	6.593731079	71.98027229
5	1.177998635	4.362957909	76.54133422	1.177998635	4.362957909	76.54133422	1.231486722	4.561061933	76.54133422
6	0.970124674	3.593054349	80.13438857						
7	0.885864623	3.280980084	83.41536866						
8	0.797411913	2.953377455	86.36874611						
9	0.681774803	2.525091865	88.89383798						
10	0.613064623	2.270609715	91.16444769						
11	0.478427342	1.771953119	92.93640081						
12	0.402349041	1.490181633	94.42658244						
13	0.340865497	1.262464802	95.68904725						
14	0.310290875	1.149225464	96.83827271						
15	0.20019832	0.741475258	97.57974797						
16	0.168364538	0.623572362	98.20332033						
17	0.162970076	0.603592874	98.8069132						
18	0.134149449	0.496849812	99.30376302						
19	0.08396257	0.310972481	99.6147355						
20	0.051772827	0.19175121	99.80648671						
21	0.03553948	0.131627704	99.93811441						
22	0.009408201	0.034845189	99.9729596						
23	0.005765035	0.021351981	99.99431158						
24	0.000775239	0.002871254	99.99718283						
25	0.000724532	0.00268345	99.99986628						
26	3.18979E-05	0.00011814	99.99998443						
27	4.20522E-06	1.55749E-05	100						

Stage Five: Interpreting the Factors

Analyzing the initial un-rotated factor matrix

Table 14 shows the result of stage four, un-rotated component analysis factor matrix.

Table 14 Un-rotated Component Analysis Factor Matrix

Component Matrix						
	Component					
Variable \ *Factor*	*1*	*2*	*3*	*4*	*5*	*Difference between two highest loadings*
LOR_1	0.20968	-0.22565	0.59024*	-0.58391**	0.00815	0.00633
LoR2	0.12079	-0.07375	0.40577**	-0.45287*	0.22092	0.04710
Frequency	0.46225	-0.82344*	0.27950	0.12582	0.01573	0.36119
rFrequency	0.45042	-0.82492*	0.12480	0.26338	0.01834	0.37449
Freq last 1 yr	0.41705	-0.78548*	0.08205	0.28142	0.07030	0.36843
Recency	-0.18319	0.43382*	0.41319**	0.38332	0.04494	0.02063
IPT	-0.28373	0.56607*	0.43005	0.38037	0.23745	0.13602
StddevRec (IPT)	-0.31534	0.60180*	0.45790	0.32323	-0.09667	0.14389
CV Recency	-0.09000	0.21790	0.23023**	0.03017	-0.67283*	0.44260
Amount	0.86231*	0.47297**	-0.06096	-0.02869	0.03203	0.38934
Total Amount	0.96513*	0.14313**	0.04125	0.03546	0.00582	0.82199
Stddev Amount	0.82476*	0.40890**	0.00331	-0.05167	-0.14541	0.41586
rAmount	0.82720*	0.48379**	-0.12589	0.04725	0.04210	0.34341
rTotal Amount	0.95552*	0.16540**	-0.02464	0.08822	0.01487	0.79011
rMajorTrip	-0.04590	0.05450	-0.15605**	-0.06236	0.70586*	0.54981
2002 Frequency	0.36921	-0.58210*	0.54049**	-0.21532	-0.05377	0.04160
2002 Total Sales	0.82832*	0.08810	0.22398**	-0.11549	-0.04355	0.60433
2002 Average Sales	0.74941*	0.41907**	-0.03176	-0.00711	0.01189	0.33034
2002 IPT	-0.02200	0.16714	0.25680*	-0.10939	0.19395**	0.06285
2003 Frequency	0.42033**	-0.78140*	0.10825	0.26130	0.02500	0.36107
2003 Total Sales	0.92081*	0.14748**	-0.02425	0.09102	0.03046	0.77333
2003 Average Sales	0.81790*	0.46532**	-0.03496	-0.03002	0.03601	0.35258
2003 IPT	-0.31636	0.58197*	0.33513**	0.16638	0.16452	0.24684
2004 Frequency	0.41698**	-0.78609*	0.08154	0.28161	0.06913	0.36911
2004 Total Sales	0.89473*	0.14823**	-0.04108	0.08634	0.01729	0.74650
2004 Average Sales	0.76689*	0.44101**	-0.05681	-0.05330	-0.01440	0.32588
2004 IPT	-0.30382	0.55562*	0.31895**	0.06424	0.05454	0.23667
						Total
Sum of Squared Factor Loadings (Eigenvalues)	9.658288	6.506744	1.926432	1.926432	1.177999	20.66616
% of Variance	35.77144	24.09905	7.134934	5.172954	4.362958	76.54133

* Highest factor loading for the variable
** Second highest factor loading for the variable

Numeric values, in the upper left part of Table 14, represent the factor loadings of each variable on each of the factors. This part of the table is analyzed in order to determine the highest factor loadings (largest absolute factor loading) for each variable. In order to consider the factor loadings as significant in this analysis the method of determining the significance level according to the sample size is selected. Table 8 Guidelines for Identifying Significant Factor Loadings Based on Sample Size shows the minimum difference that should be between the highest factor loading (highest absolute value) and the second one in order to accept this as significant with the corresponding sample sizes. By analyzing this table with the sample size of the data, it is accepted that there should be at least 0.10 differences between the highest factor loading and the second one in order to accept the factor loading as significant. Difference between two highest loadings column of Table 14 shows that some variables do not significantly load to only one factor. The situation makes the interpretation of factors extremely difficult with un-rotated factor matrix solution.

Bottom of Table 14 shows some statistical values related to un-rotated component analysis factor matrix. Hair et al. (1995) explains that sum of squared factor loadings (Eigenvalues) indicates the relative importance of each factor in accordance with the variance of the set of variables being analyzed. As expected, un-rotated factor solution has extracted the factors according to their importance. As a result of this, when factor one accounts for the most variance following ones account for the less and less. Total Sum of Squared Factor Loadings 20.66 represents the total amount of variance extracted by the factor solution. Last row of Table 7 shows the percentage of variance extracted by each factor. Total percentage of variance shows the amount of the variance extracted by the factor solution. Hair et al. (1995) indicates that, if the variables of the analysis are all very different from one another, the index has lower values when if the variables fall into one or more highly redundant or related groups the index will approach to 100 percent. Total percentage of variance for the solution is calculated as 76.54 by the system. The value shows that 76.54 percent of the total variance is captured by the information the factor matrix contains. Since the index for the solution is high it can be confirmed that the variables in the analysis are highly related to one another.

Employing a rotational method

Orthogonal – Varimax Factor Rotational method is employed to achieve simpler and theoretically more meaningful factor solutions. The factor matrix is rotated to redistribute the variance calculated by the un-rotated factor solution from initial factor to following ones.

Interpreting the rotated factor matrix

1. Examining the Factor Loading Matrix:

Table 15 shows the Varimax rotated component analysis factor matrix.

Table 15 Rotated Component Analysis Factor Matrix

Rotated Component Matrix						
	Component					
Factor / Variable	1	2	3	4	5	Difference between two highest loadings
LOR_1	0.05521	0.15271**	-0.08059	0.85822*	0.12198	0.70550
LoR2	0.04865	0.02489	0.00391	0.64846*	-0.12280**	0.52566
Frequency	0.06856	0.92902*	-0.24504**	0.23891	0.03213	0.68398
rFrequency	0.06777	0.94347*	-0.26925**	0.03719	-0.00676	0.67421
Freq last 1 yr	0.05611	0.90069*	-0.25102**	-0.00401	-0.06614	0.64967
Recency	0.00281	-0.09529**	0.72211*	-0.06235	0.08380	0.62682
IPT	-0.03511	-0.21551**	0.85351*	-0.04486	-0.09366	0.63801
StddevRec (IPT)	-0.05082	-0.29255**	0.79946*	-0.04437	0.23626	0.50691
CV Recency	0.00014	-0.15805**	0.15504	-0.00700	0.71634*	0.55829
Amount	0.98156*	-0.08664**	0.00815	0.02241	-0.03621	0.89492
Total Amount	0.93508*	0.25800**	-0.07877	0.08812	0.00622	0.67707
StddevAmount	0.91778*	-0.05376	-0.02121	0.05735	0.14934	0.86043
rAmount	0.95899*	-0.09097**	0.02096	-0.07838	-0.06035	0.86801
rTotalAmount	0.93992*	0.24297**	-0.07856	0.00454	-0.01744	0.69695
rMajortrip	-0.01511	-0.08768**	0.04150	0.03927	-0.72137*	0.67987
2002 Frequency	0.06376	0.62948*	-0.13563	0.60854**	0.16587	0.02094
2002 Total Sales	0.77588*	0.23365	-0.05035	0.30156**	0.09600	0.54223

Rotated Component Matrix							
	Component						
Factor / Variable	1	2	3	4	5		Difference between two highest loadings
2002 Average Sales	0.85584*	-0.06804**	0.03080	0.01739	-0.01003		0.78780
2002 IPT	0.03402	-0.09723	0.25004*	0.05724**	-0.11937		0.19280
2003 Frequency	0.05918	0.89442*	-0.25243**	0.02066	-0.01602		0.64199
2003 Total Sales	0.90097*	0.24681**	-0.07549	0.00357	-0.03275		0.65416
2003 Average Sales	0.93677*	-0.08911**	0.03007	0.03739	-0.03353		0.84767
2003 IPT	-0.05860	-0.36909**	0.68021*	0.02916	-0.05139		0.31112
2004 Frequency	0.05582	0.90099*	-0.25175**	-0.00460	-0.06515		0.64925
2004 Total Sales	0.87830*	0.22920**	-0.08567	-0.00908	-0.02428		0.64910
2004 Average Sales	0.88096*	-0.10868**	-0.00755	0.02921	0.00858		0.77228
2004 IPT	-0.05950	-0.40286**	0.57923*	0.07897	0.04771		0.17636
							Total
Sum of Squared Factor Loadings (Eigenvalues)	9.095229	5.414839	3.144298	1.780307	1.231487		20.66616
% of Variance	33.68603377	20.05495898	11.64554847	6.593731079	4.561061933		76.54133422

* Highest factor loading for the variable
** Second highest factor loading for the variable

Bottom of Table 15 shows some statistical values related to rotated component analysis factor matrix. When the values in this part are compared with the ones in Table 14 it is shown that total amount of the variance extracted is same for both solutions, 76.54 % . However, by applying the Varimax rotation variance has been distributed from initial factors to following ones. As a result of this, percentage of variance extracted by each factor is different in rotated matrix as well as the factor loading pattern.

In the un-rotated factor solution all variables loaded significantly on the first, second and third factors and their loadings cannot be defined as significant. When difference between two highest loadings column of Table 15 analyzed it is shown that the variables are significantly loaded to factors by not having difference smaller than 0.10. The factor loadings also show that variables are distributed between factors and none of the variables loads significantly to more than one factor.

2. Examining the Communalities Matrix:

Table 16 shows the communalities for the factor matrix. Communalities matrix is being analyzed to eliminate the variables that do not load to any factors. Numeric values in the table show the amount of variance accounted for by the factor solution for each variable (Hair et al., 1995). In this analysis if 50 percent of the variance in a variable has not been extracted by the factor analysis, this variable is discarded from the dataset. As it is shown in table 9 none of the variables in the dataset has communality value less than 0.50 which certifies inclusion of all variables in the further analysis.

Table 16 Communalities

Communalities		
	Initial	*Extraction*
LoR_1	1	0.78428
LoR2	1	0.53858
Frequency	1	0.98593
rFrequency	1	0.96865
Freq last 1 yr	1	0.88178
Recency	1	0.54145
IPT	1	0.78694
Std Dev Recency	1	0.78509
CV Recency	1	0.56220
Amount	1	0.97285
Total Amount	1	0.95495
Std Dev Amount	1	0.87125
R Amount	1	0.93816
R Total Amount	1	0.94898
R Major Trip	1	0.53155
2002 Frequency	1	0.81654
2002 Total Sales	1	0.75928
2002 Average Sales	1	0.73844
2002 IPT	1	0.54395
2003 Frequency	1	0.86789
2003 Total Sales	1	0.87944
2003 Average Sales	1	0.88891
2003 IPT	1	0.60583
2004 Frequency	1	0.88255
2004 Total Sales	1	0.83195
2004 Average Sales	1	0.78888
2004 IPT	1	0.50985

3. Naming the Factors

Table 17 shows the factors with variables that have highest loading on them.

Table 17 Factors with Corresponding Variables

Component									
1-Amount		*2-Frequency*		*3-Recency*		*4-LoR*		*5-Other*	
Total Amount	0.9350	R Frequency	0.9434	IPT	0.8535	LOR_1	0.8582	R Major Trip	-0.7213
Std Dev Amount	0.9177	Frequency	0.9290	Std Dev Recency	0.7994	LoR_2	0.6484	CV Recency	0.7163
R Total Amount	0.9399	2004 Frequency	0.9009	Recency	0.7221				
R Amount	0.9589	Freq last one year	0.9006	2002 IPT	0.2300				
Amount	0.9815	2003 Frequency	0.8944	2003 IPT	0.6802				
2004 Total Sales	0.8783	2002 Frequency	0.6294	2004 IPT	0.5792				
2004 Average Sales	0.8809								
2003 Total Sales	0.9009								
2003 Average Sales	0.9367								
2002 Total Sales	0.7758								
2002 Average Sales	0.8558								

According to the analysis Factor-1 has eleven significant loadings when Factor-2 and Factor-3 have six and Factor-4 and Factor-5 have two ones. Factors are named according to the common specialties of the variables located in them. Variables related to the purchased amount of customers are located in Factor-1and based on this Factor-1 is named as Amount. Table 17 shows that all variables in Factor-1 have positive signs, which indicate that all of them are varying together. Factor-2 contains the variables related to Frequency; Factor-3 contains the ones related to Recency when Factor-4 is shaped by the Length of Relationship variables. For all these factors the variables they contain are of same sign, suggesting that these perceptions are quite similar among respondents. Different from preceding factors Factor-5 has two variables with different signs. Thus, rMajorTrip move opposite direction to the Coefficient Variance of Recency. Since these two variables do not have any common specialty, this factor is named as “Other”.

Stage Six: Validation of Factor Analysis

Validation of factor analysis applied to the original dataset has two main parts in this analysis:

- The first part of the validation is achieved by splitting the original dataset into two samples and applying the factor analysis with the same specifications to each of them. Analysis results for two sample datasets and the original dataset are compared to assess the generalizability of the results to the population. For sampling procedure, random sampling specialty of SPSS analysis tool is used. Factor analysis is applied to the samples again by following the steps of “Factor Analysis Model Building Procedure” shown in Figure 27. The steps of the model with corresponding results are summarized in Table 18.

Table 18 Summarized Results of Factor Analysis Validation

		Step Description	*Expected Solution*	*Result for the Analysis*
1		Objectives Definition	Assessing the generalizability of the results to the population by applying the Factor Analysis two samples created by splitting the original dataset into two parts.	
2		Factor Analysis Method Selection	R-Type Factor Analysis	
3		Factor Analysis Assumptions Control		
	3.1	Control of correlations between variables	Shows the correlation level between the variables in the dataset. If there is not significant correlation between the variables this means that the dataset is not suitable for the Factor Analysis	74 of the 105 correlations are significant at the 0.01 level means that more than %70 of correlations is significant. The dataset is suitable for Factor Analysis
	3.2	Analyze of Kaiser Meyer Olkin Measure of Sampling	Adequacy value shows the total variance shaped by all variables. If the value is closer to 1 it shows that the data is suitable for Factor Analysis	For both parts Value for the case: 0.846. So the dataset is suitable for the Factor Analysis
	3.3	Test of Bartlett's Sphericity	Shows whether there is a correlation between the variables. If the significance level for dataset is greater than 0.10 it means the dataset is not suitable for the Factor Analysis	For both parts Value for the case: 0. So the dataset is suitable for the Factor Analysis
4		Factor Analysis Application		
	4.1	Number of factors Selection	Eigenvalues. When the eigenvalue is greater than 1 it means that the factor has contribution greater than the variable by itself.	Both parts return 5 components with eigenvalues greater than 1. Total contribution for the first part is %76 when this value is %78 for the second part. Since the last factors contribute less than 5 % the factoring procedure has stopped there for both parts.
5		Factors Interpretation		
	5.1	Analyze of Factor Loadings for un-rotated		Variables are not significantly loaded to the factors.

		Step Description	*Expected Solution*	*Result for the Analysis*
		solution		
	5.2	Employing the rotation method	Achieve simpler and theoretically more meaningful factor solutions.	Orthogonal – Varimax Factor Rotational method is employed
	5.3	Analyzing the Communalities Matrix	Shows for every variable the contribution to the overall variance build by the model. Smaller values show that the variable does not have so much contribution to the model. Ones that have Extraction values smaller than 0.50 will not be included in the mo	None of variables in the dataset has communality value less than 0.50 which certifies inclusion of all variables in the further analysis
	5.4	Interpretation of Rotated Factor Loading Matrix	For each variable factor loads will be analyzed and the greatest ones will be selected. There must be at least 0.10 differences between two factor loads in order to designate the variable to a factor.	
	5.5	Naming The Factors	Variables assigned to the Factors are analyzed and according to their characteristics names of the factors are given.	According to the common characteristics of the variables assigned to the factors, they are named as Amount, Recency, Frequency, LoR and Other

Table 19 contains the Eigenvalues and Total Variances of the factors extracted for general dataset and factor models of two samples. The table shows that the results are comparable in terms of eigenvalues and total variances of factors.

Table 19 Total Variances Extracted Comparison

Total Variance Explained – Comparison						
Component	*Values for All Dataset*		*Values for Sample 1*		*Values for Sample 2*	
	Total	% of Variance	Total	% of Variance	Total	% of Variance
1	9.095229117	33.68603377	8.62353065	31.93900241	8.594586682	31.83180252
2	5.414838924	20.05495898	5.18438067	19.20140989	4.983527694	18.45750998
3	3.144298086	11.64554847	3.600952859	13.33686244	4.123151867	15.27093284
4	1.780307391	6.593731079	2.12537735	7.871767962	2.118064576	7.844683615
5	1.231486722	4.561061933	1.201496597	4.449987397	1.242982645	4.603639428

Table 20 shows the factors extracted for general dataset and two samples with the variables significantly loaded to them. The table shows that the variables loaded significantly to the factors are same with the general dataset. There are small differences between the factor loading values of variables on factors. However, this difference should be accepted because it is resulted from the fact that the sample sizes are different for general dataset and the samples. Table 20 clarifies that results are comparable for general dataset and two samples.

Based on these results it can be concluded that results are stable within the dataset that is used for the analysis.

Table 20 Factors with Corresponding Variables Comparison

Component																			
1-Amount				2-Frequency				3-Recency				4-LoR				5-Other			
	All	Samp1	Samp2		All	Samp1	Samp2		All	Samp1	Samp2		All	Samp1	Samp2		All	Samp1	Samp2
Total Amount	0.93	0.978	0.977	rFrequency	0.943	0.794	0.774	IPT	0.854	0.660	0.675	LoR_1	0.858	0.904	0.912	r Major Trip	-0.721	-0.831	-0.715
Std Dev Amount	0.92	0.848	0.837	Frequency	0.929	0.899	0.898	Std Dev Recency	0.799	0.837	0.862	LoR_2	0.648	0.608	0.582	CV Recency	0.716	0.899	0.084
r Total Amount	0.94	0.892	0.894	2004 Frequency	0.901	0.898	0.880	Recency	0.722	0.717	0.781								
r Amount	0.96	0.917	0.920	Frequency last one yea	0.901	0.816	0.805	2002 IPT	0.230	0.433	0.445								
Amount	0.98	0.850	0.862	2003 Frequency	0.894	0.899	0.884	2003 IPT	0.680	0.729	0.788								
2004 Total Sales	0.88	0.743	0.723	2002 Frequency	0.629	0.717	0.710	2004 IPT	0.579	0.721	0.757								
2004 Average Sales	0.88	0.885	0.874																
2003 Total Sales	0.90	0.869	0.866																
2003 Average Sales	0.94	0.938	0.946																
2002 Total Sales	0.78	0.849	0.850																
2002 Average Sales	0.86	0.914	0.920																

• Second part of the validation aims to assess the impact of outliers on the results of the analysis. This part of the validation is achieved by applying the factor analysis with same specifications to the dataset that does not contain any outliers. As noted before, variables that have z-scores smaller than minus three or greater than plus three are thought as outliers but not discarded from the dataset because they are accepted as important subset of all data set. For only determine the effect of these outliers another dataset without outliers is prepared to be used in the factor analysis. This part of validation justified the ineffectiveness of outliers, which accounts for almost two percent of data.

Factor analysis is applied to the dataset without outliers by following the steps of "Factor Analysis Model Building Procedure" shown in Figure 27. The steps of the model with corresponding results are summarized in Table 21.

Table 21 Summarized Results of Factor Analysis Validation with Dataset without Outliers

Analysis Step		*Step Description*	*Expected Solution*	*Result for the Analysis*
1		Objectives Definition	Reduce the 27 variables to a smaller number	
2		Factor Analysis Method Selection	R-Type Factor Analysis	
3		Factor Analysis Assumptions Control		
	3.1	Control of correlations between variables	Shows the correlation level between the variables in the dataset. If there is not significant correlation between the variables this means that the dataset is not suitable for the Factor Analysis	74 of the 105 correlations are significant at the 0.01 level means that more than %70 of correlations is significant. The dataset is suitable for Factor Analysis
	3.2	Analyze of Kaiser Meyer Olkin Measure of Sampling	Adequacy value shows the total variance shaped by all variables. If the value is closer to 1 it shows that the data is suitable for Factor Analysis	Value for the case: 0.843. So the dataset is suitable for the Factor Analysis
	3.3	Test of Bartlett's Sphericity	Shows whether there is a correlation between the variables. If the significance level for dataset is greater than 0.10 it means the dataset is not suitable for the Factor Analysis	Value for the case: 0. So the dataset is suitable for the Factor Analysis
4		Factor Analysis Application		
	4.1	Number of factors Selection	Eigenvalues. When the eigenvalue is greater than 1 it means that the factor has contribution greater than the variable by itself.	Analysis returns 5 components with eigenvalues greater than 1. Total contribution for the solution is %76 which is an adequate value. Since the last factor contributes less than 5 % the factoring procedure has stopped there.
	4.1.1	Scree Plot	A Scree plot contains the information regarding the possible factors and their relative explatory power as expressed by their eigenvalues.	Scree plot shows that first Five Factors have eigen values greater than 1.
5		Factors Interpretation		
	5.1	Analyze of Factor Loadings for unrotated solution		Variables are not significantly loaded to the factors.

Analysis Step		*Step Description*	*Expected Solution*	*Result for the Analysis*
	5.2	Employing the rotation method	Achieve simpler and theoretically more meaningful factor solutions.	Orthogonal – Varimax Factor Rotational method is employed
	5.3	Analyzing the Communalities Matrix	Shows for every variable the contribution to the overall variance build by the model. Smaller values show that the variable does not have so much contribution to the model. Ones that have Extraction values smaller than 0.50 will not be included in the mo	None of variables in the dataset has communality value less than 0.50 which certifies inclusion of all variables in the further analysis
	5.4	Interpretation of Rotated Factor Loading Matrix	For each variable factor loads will be analyzed and the greatest ones will be selected. There must be at least 0.10 differences between two factor loads in order to designate the variable to a factor.	
	5.5	Naming The Factors	Variables assigned to the Factors are analyzing and according to their characteristics names of the factors are given.	According to the common characteristics of the variables assigned to the factors, they are named as Amount, Recency, Frequency, LoR and Other
6		Validation of factor Analysis	Validation is achieved to assess the generalizability of the results to the population	Validation of factor analysis is achieved by splitting the original dataset into two samples and applying the same analysis to them. Results for the validation assure that results are stable within the dataset.
7		Surrogate Variables Selection	Identifying appropriate variables for subsequent application with other statistical techniques	Based on literature Recency, Frequency and Amount are selected as the base variables for the further statistical techniques. Length of Relationship-1 and rMajorTrip are selected as surrogate variables from Factor-4 and Factor-5 by having the highest factor loadings in these factors.

Table 22 contains the eigenvalues and total variances of the factors extracted for original dataset and for dataset without outliers. The table shows that the results are comparable in terms of eigenvalues and total variances of factors.

Table 22 Total Variances Extracted Comparison

Total Variance Explained – Comparison				
Component	*Values for Dataset with outliers*		*Values for Dataset without outliers*	
	Total	% of Variance	Total	% of Variance
1	9.095229117	33.68603377	8.432282	31.23067282
2	5.414838924	20.05495898	5.164841	19.12904076
3	3.144298086	11.64554847	3.770419	13.96451352
4	1.780307391	6.593731079	2.140759	7.928737331
5	1.231486722	4.561061933	1.257198	4.656289414

Table 23 shows the factors extracted for original dataset and for dataset without outliers with the variables significantly loaded to them. Table shows that the variables loaded significantly to the factors are same with the original dataset. Table 22 clarifies that results are comparable for original dataset and dataset without outliers.

These results justify the ineffectiveness of outliers, which accounts for almost two percent of data.

Table 23 Factors with Corresponding Variables Comparison

Component														
1-Amount			2-Frequency			3-Recency			4-LoR			5-Other		
	Original	Without Outliers		Original	Without Outliers		Original	Without Outliers		Original	Without Outliers		Original	Without Outliers
Total Amount	0.935	0.837	r Frequency	0.943	0.901	IPT	0.854	0.841	LoR_1	0.858	0.897	r Major Trip	-0.721	-0.724
Std Dev Amount	0.918	0.895	Frequency	0.929	0.801	Std Dev Recency	0.799	0.755	LoR_2	0.648	0.616	CV Recency	0.716	0.727
r Total Amount	0.94	0.820	2004 Frequency	0.901	0.893	Recency	0.722	0.635						
r Amount	0.959	0.882	Frequency last one year	0.901	0.891	2002 IPT	0.23	0.508						
Amount	0.982	0.974	2003 Frequency	0.894	0.820	2003 IPT	0.68	0.741						
2004 Total Sales	0.878	0.848	2002 Frequency	0.629	0.716	2004 IPT	0.579	0.719						
2004 Average Sales	0.881	0.910												
2003 Total Sales	0.901	0.860												
2003 Average Sales	0.937	0.939												
2002 Total Sales	0.776	0.728												
2002 Average Sales	0.856	0.876												

Stage Seven: Additional Uses of the Factor Analysis Results

Since the objective of the analysis is to identify appropriate variables for subsequent applications, as the additional use of factor analysis results selection of surrogate variables is selected.

- Selecting Surrogate Variables

If the objective of the analysis is to identify appropriate variables for subsequent application with other statistical techniques, the factor matrix is examined and the variable with the highest factor loading is selected on each factor as a surrogate representative for that particular factor. (Hair et al., 1995)

Table 17 shows the factors with the variables that have highest loadings on them. When selecting the surrogate variables the ones with the highest loadings should be preferred. However this general acceptation has not been followed because as the base of this analysis RFM methodology has been selected. Although Recency, Frequency and Total Amount variables does not have the highest loadings on the factors they are assigned these variables are selected as the surrogate variables. On the other hand, for factor four and factor five the variables with highest loadings are selected as surrogate variables, these are LoR_1 and r Major Trip in turn. Instead of all twenty seven variables these surrogate ones will be used in the further analysis.

Factor Analysis to Define Variables of City Segmentation Analysis

Through the data collection and data preparation phases, seven variables are calculated to define characteristics of cities in which company has activities with aim to use as characteristics of cities when partitioning them into smaller groups. Variables are listed in Table 5.

In order to define the variables that will be used to partition the cities into smaller groups by identifying the underlying evaluative dimensions of data, factor analysis is applied by following the steps of "Factor Analysis Model Building Procedure" shown in Figure 27. The steps of the model with corresponding results are summarized in Table 24. Detailed explanation regarding the analysis steps can be found in the following sections.

Table 24 Summarized Results of Factor Analysis

Analysis Step		*Step Description*	*Expected Solution*	Result for the Analysis
1		Objectives Definition	Reduce the *7 vari*ables to a smaller number	
2		Factor Analysis Method S*electi*on	R-Type Factor Analysis	
3		Facto*r* A*nalysi*s Assumptions Control		
	3.1	Analyze of Kaiser Meyer Olkin Measure of Sampling	Adequacy value shows the total variance shaped by all variables. If the value is closer to 1 it shows that the data is suitable for Factor Analysis	Value for the case: 0.58. So the dataset is su*itable* for the Factor Analysis
	3.2	Test of Bartlett's Sphericity	Shows whether there is a correlation between the variables. If the significance level for dataset is greater than 0.10 it means the dataset is not suitable for the Factor Analysis	Value for the case: 0. So the dataset is su*ita*ble for the Factor Analysis
4		*Factor* Analysis Application		
	4.1	Number of factors Selection	Eigenvalues. When the eigenvalue is greater than 1 it means that the factor has contribution greater than the variable by itself.	Analysis returns 3 components with eigenvalues greater than 1. Total contribution for the solution is *%83 which* is an adequate value.
	4.1.1	Scree Plot	Scree plots contain the information regarding the possible factors and their relative expletory power as expressed by their eigenvalues.	Scree plot shows that first Three Factors hav*e e*igen values greater than 1.
5		*F*actors Interpretation		
	5.1	Analyze of Factor Loadings for un-rotated solution		Variables are not signif*icantl*y loaded to the factors.
	5.2	Employing the rotation method	Achieve simpler and theoretically more meaningful factor solutions.	Orthogonal – Varimax Factor *Rotati*onal method is employed

Analysis Step		*Step Description*	*Expected Solut*ion	Result for the Analysis
	5.3	Analyzing the Communalities Matrix	Shows for every variable the contribution to the overall variance build by the model. Smaller values shows that the variable does not have so much contribution to the model. Ones that have Extraction values smaller than 0.50 will not be included in the mo	None of variables in the dataset has communality value less than 0.50 which certifies inclusion of all var*iables* in the further analysis
	5.4	Interpretation of Rotated Factor Loading Matrix	For each variable factor loads will be analyzed and the greatest ones will be selected. There must be at least 0.10 difference between two factor loads in order to designa*te the* variable to a factor.	
	5.5	Naming The Factors	Variables assigned to the Factors are analyzing and according to their characteristics names of the factors are given.	According to the common characteristics of the variables assigned to the factors, they are named as Amount*, R*ecency- Frequency and Other
6		Validation of factor Analysis	Validation is achieved to assess the generalizability of the results to the population	Validation of factor analysis is achieved by splitting the original dataset into two samples and apply*in*g the same analysis to them
7		Surrogate Variables Selection	Identifying appropriate variables for subsequent application with *othe*r statistical techniques	

Stage One: Factor Analysis Problem Definition

Objectives Definition

The objective of the Factor Analysis that performed in this study is to identify the structural relationships among variables with aim to group large number of variables into a smaller number of homogenous sets and identify representative variables for use in clustering analysis. If the seven variables can be represented in a smaller number of variables, then clustering analysis can be made in more effective manner.

Stage Two: Factor Analysis Model Design

Selecting the Factor Analysis Method

In this analysis to define the underlying relationships between variables R-type Factor analysis will be used which focuses on summarizing the characteristics.

Data Characteristics for Factor Analysis

Regarding the adequacy of sample size, in this analysis there are seventy eight cities. This value is not smaller than fifty cases and 10.1 times greater than the number of variables. This ratio shows that sample size is adequate for getting reasonable results from factor analysis. None of the seven variables are categorical ones and z-score of these variables are used in the analysis in order to come up with comparable measures.

Stage Three: Control of Assumptions

- Bartlett Test of Sphericity

As it is shown in Table 25, significance level of this test is 0, which shows that dataset can be accepted as a suitable one for the factor analysis.

Table 25 Results for Bartlett Test of Sphericity and KMO Index

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.578
Bartlett's Test of Sphericity	Approx. Chi-Square	508.152
	Df	28
	Sig.	0

- Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO)

As it is shown in Table 25, the value for this statistic is 0.58. Since the value is grater than 0.5 the dataset is accepted as suitable for factor analysis.

Stage Four: Applying the Factor Analysis and Specifying the Factor Matrix

Factor analysis is applied by using SPSS for Windows statistical tool in this study. The characteristics selected for the applied Factor Analysis are same with the ones for factor analysis of customer characteristics, summarized in Table 12.

Selecting the Number of Components:

Factors that are representing the underlying dimensions in the original dataset are extracted by using the principal component analysis. In order to determine the number of Factors that will be used in the analysis Percentage of Variance Extracted, The Latent Root Criterion and Scree Test Criterion are employed. Table 26 shows the information regarding the seven possible factors and their explanatory power as expressed by their eigenvalues and percentage of variance. According to latent root criterion, three factors are extracted from the dataset. Additional to this, Scree Test represented in Figure 31, supports the result of the latent root criterion by indicating two or three factors as appropriate number of factors. Since the following factors have eigenvalues smaller than one, the factoring procedure is stopped at three factors.

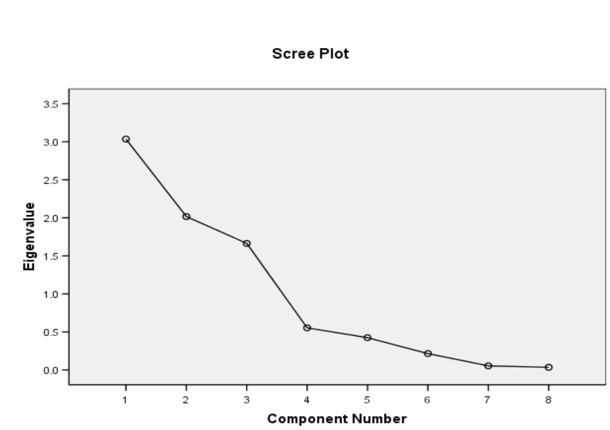


Figure 31 Scree Test Curve

According to general assumption of Percentage of Variance Extracted criterion three factors are extracted by SPSS which accounts for eighty three percent of total variance. By combining the results of these three criteria three factors are extracted from the dataset for further analysis.

Table 26 Results for the Extraction of Component Factors

Total Variance Explained									
Component	*Eigenvalues*			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	3.034926900	37.936586250	37.93658626	3.034926901	37.93658626	37.93658626	2.627529	32.84411541	32.84411541
2	2.016379265	25.20474082	63.14132707	2.016379265	25.20474082	63.14132707	2.282431	28.53038998	61.37450539
3	1.664179026	20.80223782	83.9435649	1.664179026	20.80223782	83.9435649	1.805525	22.5690595	83.9435649
4	0.553980192	6.924752398	90.8683173						
5	0.42471868	5.308983497	96.17730079						
6	0.216079267	2.700990838	98.87829163						
7	0.05451678	0.681459746	99.55975138						
8	0.03521989	0.440248622	100						

Stage Five: Interpreting the Factors

Analyzing the initial un-rotated factor matrix

Table 27 shows the result of stage four, un-rotated component analysis factor matrix.

Table 27 Un-rotated Component Analysis Factor Matrix

Component Matrix					
	Component				
Variable \ *Factor*	*1*	*2*	*3*		*Difference between two highest loadings*
Average Sales for City_2	-0.55921**	0.759905*	0.133817		0.200690015
Average IPT for City	-0.58992*	-0.58345**	0.097224		0.006461776
Count of Customers in the City	0.310027**	-0.10774	0.885969*		0.575942052
Average frequency for City	0.902828*	0.158525	-0.17266**		0.744302994
Average Frequency last year for City	0.853122*	0.207952	-0.3404**		0.51271744
Average Recency for City	-0.41279**	-0.67195*	-0.10958		0.259155163
Sales per Customer in the City	0.474347**	-0.10323	0.80573*		0.33138256
					Total
Sum of Squared Factor Loadings (Eigenvalues)	3.034926901	2.016379265	1.664179026		6.715485192
% of Variance	37.93658626	25.20474082	20.80223782		83.9435649

* Highest factor loading for the variable
** Second highest factor loading for the variable

Same analyses with factor analysis of customer characteristics are employed to analyze Factor Loading Matrix shown in Table 27. Differences between two highest loadings column of Table 27show that some variables do not significantly load to only one factor. The situation makes the interpretation of factors extremely difficult with un-rotated factor matrix solution.

Total Sum of Squared Factor Loadings 6.71 at bottom part of the Table 27, represents the total amount of variance extracted by the factor solution. Last row of Table 27 shows the percentage of variance extracted by each factor which is calculated as 83.94 by the system. The value shows that 83.94 percent of the total variance is represented by the information the factor matrix contains. Since the index for the solution is high it can be confirmed that the variables in the analysis are highly related to one another.

Employing a rotational method

Just like factor analysis of customer characteristics Orthogonal – Varimax Factor Rotational method is employed to achieve simpler and theoretically more meaningful factor solutions by redistributing the variance between factors.

Interpreting the rotated factor matrix

4.Examining the Factor Loading Matrix:

Table 28 shows the Varimax rotated component analysis factor matrix.

Table 28 Rotated Component Analysis Factor Matrix

Component Matrix					
	Component				
Factor / *Variable*	*1*	*2*	*3*		*Difference between two highest loadings*
Average Sales for City_2	-0.02157	0.94494*	-0.12127**		0.923370746
Average IPT for City	-0.82406*	-0.13053**	-0.04202		0.693530791
Count of Customers in the City	-0.00331	-0.02351**	0.944512*		0.920999857
Average frequency for City	0.836405*	-0.39814**	0.109321		0.438265999
Average Frequency last year for City	0.861067*	-0.37509**	-0.06932		0.485973733
Average Recency for City	-0.69415*	-0.34961**	-0.17277		0.344538207
Sales per Customer in the City	0.144783**	-0.12819	0.920581*		0.775798399
					Total
Sum of Squared Factor Loadings (Eigenvalues)	3.034926901	2.016379265	1.664179026		6.715485192
% of Variance	35.93	24.904741	23.092238		83.943565

* Highest factor loading for the variable
** Second highest factor loading for the variable

Comparisons between Table 27 and Table 28 show that, total amount of the variance extracted is same for both solutions, 83.94 percent. However, since by applying the Varimax rotation variance has been distributed from earlier factors to later ones, percentage of variance extracted by each factor is different in rotated matrix as well as the factor loading pattern.

When the difference between two highest loadings column of Table 27 is analyzed it is shown that the variables are significantly loaded to factors by not having difference smaller than 0.10. The factor loadings also show that variables are distributed between factors and none of the variables loads significantly more than one factor.

5.Examining the Communalities Matrix:

Table 29 shows the communalities for the factor matrix which will be used to eliminate the variables that do not load to any factors. As it is shown in Table 29 none of the variables in the dataset has communality value less than 0.50 which certifies inclusion of all variables in the further analysis.

Table 29 Communalities

Communalities		
	Initial	***Extraction***
Average Sales for City_2	1	0.78428
Average IPT for City	1	0.53858
Count of Customers in the City	1	0.98593
Average frequency for City	1	0.96865
Average Frequency last year for City	1	0.88178
Average Recency for City	1	0.54145
Sales per Customer in the City	1	0.78694

6.Naming the Factors

Table 30 shows the factors with variables that have highest loading on them.

Table 30 Factors with Corresponding Variables

Component					
1-Recency-Frequency		*2-Amount*		*3-Other*	
Average Frequency last year for City	0.861	Average Sales for City_2	0.944	Count of Customers in the City	0.8535
Average frequency for City	0.836			Sales per Customer in the City	0.7994
Average IPT for City	-0.824				
Average Recency for City	-0.694				

According to the analysis Factor-1 has four significant loadings when Factor-2 has one and Factor-3 has two ones. According to the common specialties of the variables loaded to each factor, factors are named as Recency-Frequency, Amount and Other. Different from other two factors, two of the variables loaded to Factor-1 have positive signs when other two have negative signs. Thus, Average Frequency and Average Frequency Last year variables move opposite direction to the Average IPT and Average Recency.

Stage Six: Validation of Factor Analysis

Validation of factor analysis applied to the original dataset is achieved by splitting the original dataset into two samples and applying the factor analysis with the same specifications to them. Sampling is done via Random sampling functionality of SPSS analysis tool. Analysis results for two sample datasets and the original dataset are compared to assess the generalizability of the results to the population. The steps of the analysis for validations with corresponding results are summarized in Table 31.

Table 31 Summarized Results of Factor Analysis Validation

Analysis Step		*Step Description*	*Expected Solution*	Result for the Analysis
1		Objectives Definition	Assessing the generalizability of the results to the population by applying the Factor Analysis two samples created by splitting the original dataset into two parts.	
2		Factor Analysis Method Selection	R-Type Factor Analysis	
3		Factor Analysis Assumptions Control		
	3.1	Analyze of Kaiser Meyer Olkin Measure of Sampling	Adequacy value shows the total variance shaped by all variables. If the value is closer to 1 it shows that the data is suitable for Factor Analysis	For part 1 Value for the case: 0.49 when for the part 2 it is 0.59. So the dataset is suitable for the Factor Analysis.
	3.2	Test of Bartlett's Sphericity	Shows whether there is a correlation between the variables. If the significance level for dataset is greater than 0.10 it means the dataset is not suitable for the Factor Analysis	For both parts Value for the case: 0. So the dataset is suitable for the Factor Analysis
4		*Factor* Analysis Application		
	4.1	Number of factors Selection	Eigenvalues. When the eigenvalue is greater than 1 it means that the factor has contribution greater than the variable by itself.	Both parts return 3 components with eigenvalues greater than 1. Total contribution for the first part is %84 when this value is %86 for the second part.
5		Factors Interpretation		
	5.1	Analyze of Factor Loadings for un-rotated solution		Variables are not significantly loaded to the factors.
	5.2	Employing the rotation method	Achieve simpler and theoretically more meaningful factor solutions.	Orthogonal – Varimax Factor Rotational method is employed

	5.3	Analyzing the Communalities Matrix	Shows for every variable the contribution to the overall variance build by the model. Smaller values shows that the variable does not have so much contribution to the model. Ones that have Extraction values smaller than 0.50 will not be included in the mo	None of variables in the dataset has communality value less than 0.50 which certifies inclusion of all var*iables* in the further analysis
	5.4	Interpretation of Rotated Factor Loading Matrix	For each variable factor loads will be analyzed and the greatest ones will be selected. There must be at least 0.10 differences between two factor loads in order to designa*te the* variable to a factor.	
	5.5	Naming The Factors	Variables assigned to the Factors are analyzed and according to their characteristics names of the factors are given.	According to the common characteristics of the variables assigned to the factors, they are named as Amount, Recency- Frequency and Other

Table 32 contains the Eigenvalues and Total Variances of the factors extracted for general dataset and factor models of two samples. The table shows that the results are comparable in terms of eigenvalues and total variances of factors.

Table 32 Total Variances Extracted Comparison

Total Variance Explained - Comparison						
Component	*Values for All Dataset*		*Values for Sample 1*		*Values for Sample 2*	
	Total	% of Variance	Total	% of Variance	Total	% of Variance
1	3.034926901	37.93658626	3.056003132	38.20003916	3.136927136	39.2115892
2	2.016379265	25.20474082	1.92105758	24.01321975	2.165220552	27.0652569
3	1.664179026	20.80223782	1.759784316	21.99730395	1.600347234	20.00434042

Table 33 shows the factors extracted for general dataset and two samples with the variables significantly loaded to them. As it is shown in Table 33, when variables loaded to the factors are same for all models there are some differences between the loadings resulted from the fact of having different sample sizes. Table 33 clarifies that results are comparable for general dataset and two samples.

Based on these results it can be concluded that results are stable within the dataset that is used for the analysis

Table 33 Factors with Corresponding Variables Comparison

Component											
1-Frequeny - Recency				2-Amount				3-Other			
	All	Samp1	Samp2		All	Samp1	Samp2		All	Samp1	Samp2
Average Frequency last year for City	0.861	0.94	0.78	Average Sales for City_2	0.944	0.92	0.94	Count of Customers in the City	0.853	0.95	0.95
Average frequency for City	0.836	0.86	0.79					Sales per Customer in the City	0.799	0.94	0.91
Average IPT for City	-0.824	-0.85	-0.83								
Average Recency for City	-0.694	-0.48	-0.78								

Stage Seven: Additional Uses of the Factor Analysis Results

Just like the factor analysis for customer characteristics, since the objective of the analysis is to identify appropriate variables for subsequent applications, as the additional use of factor analysis results selection of surrogate variables is selected.

- Selecting Surrogate Variables

With the objective of identifying appropriate variables for subsequent application with other statistical techniques, the factor matrix is examined and the variable with the highest factor loading is selected on each factor as a surrogate representative for that particular factor.

Table 30 shows the factors with the variables that have highest loadings on them. When selecting the surrogate variables the ones with the highest loadings are preferred. Results of the analysis show that; Average Frequency for City, Average Sales for City and Count of Customers are the surrogate variables of factors determined at the end of the analysis. Instead of all seven variables these surrogate ones will be used in the further analysis.

CHAPTER 6

CLUSTER ANALYSES FOR SEGMENTATION

Customer segmentation is the process of partitioning markets into groups of potential customers with similar needs and/or characteristics who are likely to exhibit similar purchase behavior (Weinstein, 2004). Cluster analysis is used to achieve objectives of this data mining application.

Cluster Analysis is an explorative statistical method to group objects based on the characteristics they process (Hair et al., 1995). Cluster analysis groups objects so that the degree of association is strong between members of the same cluster and weak between members of different clusters. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

In this chapter a brief summary of cluster analysis steps will be discussed. Afterwards alternative cluster analyses models that are build to segment company's customers and cities in which company performs will be analyzed. Following steps are followed in the application of cluster analysis.

Steps for Cluster Analysis

Stage One: Deriving Clusters and Assessing the Overall Fit

Selection of Clustering Algorithm

Clustering algorithms are the procedures used to partition the dataset into small groups by maximizing the differences between clusters relative to the difference within them as represented in Figure 32. Clustering algorithms can be analyzed under two main groups named as hierarchical and nonhierarchical cluster procedures.

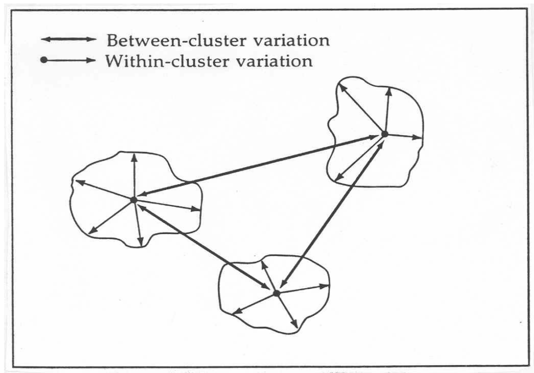


Figure 32 Cluster diagram showing between and within cluster variation.
Source: Hair et al. , 1995

- Hierarchical Clustering Procedures:

Hierarchical clustering procedures involves methods to partition the object into smaller groups by constructing a hierarchy of a tree like structure. Hierarchical clustering procedures follow one of two approaches: Agglomerative Methods and Divisive Methods. Agglomerative methods start with each observation as a cluster and with each step combine observations to form clusters until there is only one large cluster. Divisive methods begin with one large cluster and proceed to split into smaller clusters items that are most dissimilar (University of Illinois at Chicago, 2001). The cluster formation methods of four hierarchical clustering procedures are explained in Table 34.

Table 34 Clustering Methods

Clustering Procedure Name	*Clustering Method Name*	*Method of Forming Clusters*
Hierarchical Clustering		
	Single Linkage Analysis	An observation is joined to cluster if it has a certain level of similarity with at least one of the members of that cluster. Connections between clusters are based on links between single entities.
	Complete Linkage Analysis	An observation is joined to cluster if it has a certain level of similarity with all current members of that cluster.

Clustering Procedure Name	*Clustering Method Name*	*Method of Forming Clusters*
	Average Linkage Analysis	An observation is joined to cluster if it has a certain average level of similarity with all current members of that cluster.
	Ward's Method	Ward's method is designed to generate clusters in such a way as to minimize the within cluster variance.
Nonhierarchical Clustering		These methods begin with the partition of observations into a specified number of clusters. This partition may be on a random or nonrandom basis. Observations are then reassigned to clusters until some stopping criterion is reached. Methods differ in the nature of reassignment and stopping rules.
	K-means Analysis	Cases are reassigned by moving them to the cluster whose centroid is closest to that case. Reassignment continues until every case is assigned to the cluster with the nearest centroid. Such a procedure implicitly minimizes the variance within each cluster.
	Hill Climbing Methods	Cases are not reassigned to the cluster with the nearest centroid but are moved from one cluster to another if a particular statistical criterion is obtained. Reassignment continues until optimization occurs. The objective function to be optimized may be minimizing the within cluster variance, or obtaining the largest eigenvalue, etc.

Source: Punj, Stewart, 1983.

- <u>Nonhierarchical Clustering Procedures</u>:

Nonhierarchical clustering procedures involves methods to partition the object into smaller groups by assigning them to specified number of clusters. These methods begin with the partition of observations into a specified number of clusters on a random or non random basis. Observations are then reassigned to clusters until some stopping criterion is reached. There are three approaches to assign the objects to the clusters in non-hierarchical clustering procedures named as Sequential Threshold Method, Parallel Threshold Method and Optimizing Procedure. (Hair et al., 1995)

  - Sequential Threshold Method starts by selecting one cluster seed and includes all objects within a pre-specified distance. When all objects within the distance are included, a second cluster seed is selected and all objects within the pre-specified distance are included. Then a third seed is selected, and the process continues as before. When an object is clustered with a seed, it is no longer considered for other seeds.

- Parallel Threshold Method selects several cluster seeds simultaneously in the beginning and assigns objects within the threshold distance to the nearest seed. As the process evolves, threshold distances can be adjusted to include fewer or more objects in the clusters.
- Optimizing Procedure is similar to the other two except that it allows for reassignment of objects. If in the course of assigning objects, an object becomes closer to another cluster that is not the cluster it was originally assigned, then an optimizing procedure will switch the object to the more similar cluster.

The cluster formation methods of nonhierarchical clustering procedures are explained in Table 34.

- Selection between Hierarchical and Non-Hierarchical Procedures:

The clustering procedure that will be used in the analysis can be chosen according to the characteristics of the research problem in hand and evolving specialties of the available procedures and their fitness to the problem. The disadvantages of both procedures are summarized in Table 35.

Table 35 Disadvantages of Hierarchical and Nonhierarchical Clustering Procedures

Disadvantages of Hierarchical Methods	Disadvantages of Non-Hierarchical Methods
Undesirable early combinations created by the system may persist throughout the analysis and lead to artificial results.	Usage of methods depends on the ability of the researcher to select the seed points according to some practical, objective or theoretical basis.
Existence of outliers may affect the analysis results which force the analyzer to delete the cases from the dataset.	
Not suitable to analyze large samples	

Selection of Similarity Measurement

Partitioning process in cluster analysis is based on inter object similarity measurement. Hair et al. (1995) defines inter object similarity as a measurement of correspondence or resemblance between objects to be clustered. To partition the objects, firstly the characteristics defines the similarity are determined. Then, similarity measures are calculated for all pairs of objects with these characteristics.

The calculated similarity measures are used to compare the objects with themselves for purpose of grouping similar objects together into clusters.

Inter-object similarity, in cluster analysis, can be measured via two ways.

- Association Measurement: If dataset contains qualitative data association measures of similarity are used to compare objects.
- Distance Measurement: If the dataset contains quantitative data distance measures of similarity should be used. Distance measure of similarity represents similarity as the proximity of observations to one another across the pre-determined variables (Hair et al., 1995). There are several ways to calculate the distance between two objects. Selection of the measure is usually based on the needs of analysis. Two of distance measures will be explained below. One of them is Manhattan distance which calculates the distance by using the sum of absolute differences between variables. The formula to calculate Manhattan distance between two objects measured on two variables (X, Y) is shown in Equation 7.

$$D_M(x, y) = \sum_{x=i}^{n} |x_i - y_i| \quad (7)$$

On the other hand, Euclidian distance is accepted as the most commonly used distance measure. Euclidian distance between two objects is calculated as the length of the hypotenuse of a right triangle formed between them. The formula to calculate Euclidian distance is shown in Equation 8.

$$D_E(x, y) = \sqrt{\sum_{x=i}^{n} (x_i - y_i)^2} \quad (8)$$

Stage Two: Determining Number of Clusters

There is no generally accepted standard procedure for determining the number of clusters. The decision should be guided by practical judgment, common sense and interpretability of results. Hair et al. (1995) argues that inter-cluster distance can also be used as a guide for the cluster number selection. When using a criterion such as within cluster sum of squares, this can be plotted against the number of clusters in a diagram. And the changes in the criterion can be monitored to select the number of clusters. (University of Illinois at Chicago, 2001)

Stage Three: Validation of Clusters

Validation of the cluster analysis is achieved to assure that the cluster solution is representative of the general population. Hair et al. (1995) proposes three different validation methods. The one that will be used in the analysis can be choose according to the needs of the analysis and availability of dataset.

1. Applying cluster analysis with same specifications to different samples and comparing the cluster solutions to asses the correspondence of results. This method usually not used because of the unavailability of different samples for the analysis as well as the time constraints.
2. Splitting the original dataset into two samples and applying the cluster analysis with same specifications to these samples. In this method; results of the both parts are analyzed separately and compared to each other.
3. Analyzing the value of control variables among the clusters also called as criterion or predictive validity. In this method of validation

variables that are not included in the cluster analysis are selected as control variables. When selecting the control variables, the dispersion of it among the clusters should be predictable by referring to the variables that formed the clusters. If the variables have unmeaningful values among the clusters the clustering procedure should be repeated.

Stage Four: Interpretation of Clusters

The interpretation of the cluster steps aims to analyze the general structure of the derived clusters and give names to them that are describing their nature. Clusters' centroids can be used as a guide in the interpretation of clusters. If dataset is transformed before the partition process start, z-score values for the clusters' centroids should be converted to the original variables for the interpretation.

Cluster Analysis to Segment Customers and Cities

Two different cluster analyses are applied to the datasets, prepared at the end of the data preparation step, with aim to obtain smaller manageable customer and city groups for customer relationship management projects and activities of company. For both analyses alternative models are built with surrogate variables determined at the end of factor analyses and partitioning of the cases is performed based on the similarity of objects for these surrogate variables. Selection between the alternative models is made according to the manageability of results with the help of basic objective of clustering procedure; minimizing the within cluster distance when maximizing the between clusters one.

Selection of Clustering Algorithm

K-means nonhierarchical clustering method is selected as the clustering algorithm of these analyses in consideration of the general characteristics of datasets that will be partitioned. On account of the sample size, hierarchical methods are not appropriate for these analyses because they are not suitable to analyze large samples. Additional to this, since existence of outliers has more effects on hierarchical methods compared to the nonhierarchical ones, outliers that are accepted as under sampling of actual groups make the use of nonhierarchical methods favorable instead of hierarchical ones. Among the available non-hierarchical clustering methods dependent on the capability of the analysis tool that is being used, k-means method is selected as the clustering algorithm. Being familiar to the algorithm of k-means method was another issue that has positive influence on this selection.

Selection of Similarity Measurement

Inter-object similarity, in cluster analyses of this study, is measured via distance type measurements because the datasets that will be partitioned contain quantitative data. Both Manhattan and Euclidian distances are calculated during the analyses in order to measure the inter object similarity. Distance measures that are calculated during the analyses are listed below:

- Sum of Squared Errors (SSE):

Han, Kamber (2001) accept the SSE as one of the most common measures used to evaluate the results of K-means clustering and describe SSE as total amount of variation that exists to be explained by the independent variables. This baseline is calculated by summing the squared differences between the actual variables and centroids of the clusters they are assigned.

- Total / Average Euclidian Distances of the Cases from Cluster Center:

Distance from the cluster center measures represent the distance between the cluster center and cases grouped in it, in total and on average. Specified measures indicate the wideness of the clusters. Smaller values for Average Euclidian Distance show that the cluster is a compact one when a bigger value shows that the cluster is a broad one. Average Euclidian Distances of the Cases from Cluster Center is called as Within Cluster Distance in the following sections.

- Average Within Cluster Distance:

This measure is calculated by dividing the Sum of SSE values for clusters with the total number of cases. The value is used to control whether the alternative solution is applicable according to the basic goal of clustering procedures, minimizing the within cluster distance when maximizing the between clusters one. This control will be achieved by comparing this value with Between Clusters Distance value.

- Between Cluster Manhattan / Euclidian Distances:

These measures specify the distance between the clusters by means of different measures and compared with Average within clusters distance in order to control the basic goal of clustering procedure as mentioned above.

- Total Manhattan / Euclidian Distance of the Cluster Center form the Center of the all Clusters:

These measures represent the distance between the cluster center and the center off all clusters by means of difference distance measures. When these measures have bigger values the probability that the cluster contains outliers increases.

- Total Manhattan / Euclidian Distance of the Cases form the Center of the all Dataset:

This measure shows the total distance between the cases and center of all dataset and indicates the general variance of the dataset.

Determining Number of Clusters

Inter-cluster distance is used as a guide for the cluster number selection, in these analyses. Euclidian distances from the cluster center they assigned for each case is summed to calculate sum of square, and this value is plotted against the number of clusters in a diagram in order to monitor the changes related to this comparison. It is obvious that as the number of clusters increases, within cluster sum of square decreases and approaches to zero. When selecting the number of clusters, the plotted diagram is analyzed in order to find the point after which the curve smoothes.

Validation of Clusters

Validation of the cluster analysis is achieved by splitting the original dataset into two samples and carrying out clustering on each half. Results of the two sets of clusters are compared to determine the degree to which similar clusters have been identified. Cluster analysis is first carried out on one half of the cases available for the analysis. At the end of this analysis centroids defining the clusters are obtained. Objects in the second sample are then assigned to one of the identified clusters on the basis of smallest Euclidian distance between the specified object and cluster centroid. Analysis results for two sample datasets and the original dataset are compared to assess the generalizability of the results to the population. For sampling procedure, random sampling specialty of SPSS analysis tool is used.

Interpretation of Clusters

As the last step of the cluster analysis, one of the alternative models is selected as the most useful one in partitioning the cases into small groups according to the manageability of results with the help of the basic objective of clustering procedure: minimizing the within cluster distance when maximizing the between clusters one. Clusters of the selected alternative model are interpreted in the following chapter.

Cluster Analysis to Segment Customers of Company

Three different alternative clustering models are built with surrogate variables determined at the end of factor analysis, in order to find the most useful one in partitioning the customers of company into small manageable groups for customer relationship management projects and activities. Alternative clustering models built in this analysis are as follows:

1. Clustering with “Recency”, “Frequency”, and “Total Amount” variables: This is the basic clustering model literature proposes by the name of RFM Model. Additionally, these three variables are selected as surrogate variables of first three factors determined at the end of the Factor Analysis.
2. Clustering with “Recency”, “Frequency”, “Total Amount” and “Length of Relationship” variables: In this second alternative model the fourth surrogate variable determined via factor analysis is included in the analysis.
3. Clustering with “Recency”, “Frequency”, “Total Amount”, “Length of Relationship” and “rMajorTrip” variables: Third alternative model is

build with all surrogate variables determined at the end of the Factor Analysis.

Alternative Model One:

Alternative Model One is constructed with three surrogate variables Recency, frequency and Total Amount. This is also the basic clustering model literature proposes by the name of RFM Model.

Determining Number of Clusters

Table 36 shows the SSEs for number of clusters two to twenty, and change occurs when the cluster number increased by one. In Figure 33 these SSEs are plotted against number of clusters. Based on information shown on Table 36 and Figure 33 ten is selected as number of clusters of this alternative model. Figure 33 realizes this selection by showing that the curve is becomes smoother after k equals to ten. Supporting this Table 36 shows that last maximum change occurs when number of clusters is increased to eleven from ten.

Table 36 Number of Clusters and Within Cluster Sum of Squares

Number of Clusters	*Within cluster sum of square*	*Change occurs w*hen number of cluster increases
2	149751.9	0.091
3	136091.6	0.254
4	101576.8	0.301
5	70974.77	0.102
6	63710.65	0.012
7	62962.37	0.063
8	58967.97	0.146
9	50333.3	0.079
10	46344.81	0.470
11	24578.74	0.086
12	22455.08	0.032
13	21743.01	0.047
14	20713.05	0.005
15	20616.49	0.047
16	19653.31	0.034
17	18986.51	0.051
18	18025.28	0.091
19	16385.75	0.042
20	15698.67	1

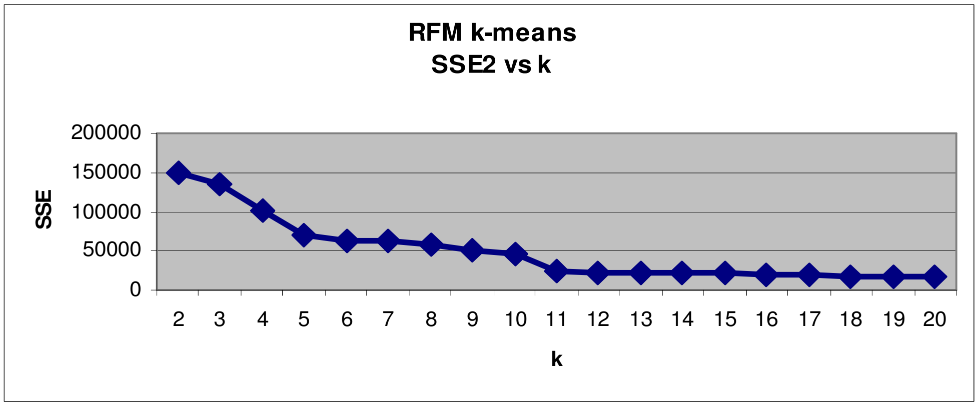


Figure 33 Number of Clusters versus Within Clusters Sum of Squares

Validation of Clusters

At the end of the partitioning process applied with three surrogate variables customers of the company are partitioned into eleven different clusters. Table 37 shows these clusters with corresponding final cluster centers, cases partitioned into them and within cluster distance produced by these cases. Between Euclidian Clusters Distance of this alternative is calculated as 4.82 which is greater than all within cluster distances except the one for cluster eight. Analysis made on this cluster shows that cluster can be interpreted as an outlier one. This information allows us to ignore the greater within cluster distance of this cluster. Average within cluster distance is calculated as 0.484 for this alternative model as it is shown in bottom of Table 37. It is smaller than the Between Cluster Distance of alternative model one; 4.82. Based on this information it can be concluded that cluster analysis applied with three surrogate variables achieved the main goal of cluster analysis, minimize the within cluster distance when maximizing the between clusters one.

Table 37 Results of Alternative Model One for Cluster Analysis of Customer Dataset

Cluster Number	*Frequency*	*Recency*	*Total Amount*	*Number of Cases*	*Total Euclidian Distance of the cases from Cluster Center*	*Within Cluster Distance*
1	0.128	-0.269	-0.093	17,439	6700.416	0.384
2	-0.487	4.327	-0.226	841	1121.655	1.334
3	-0.603	1.498	-0.262	4,122	2917.438	0.708
4	1.772	-0.265	4.152	651	1016.516	1.561
5	3.211	-0.351	0.868	1,773	1803.226	1.017
6	-0.730	-0.143	-0.306	23,323	8223.042	0.353
7	0.639	-0.242	1.499	2,221	1821.749	0.820
8	2.438	-0.299	30.759	20	111.996	5.600
9	1.283	-0.338	0.123	7,270	3581.869	0.493
10	1.863	-0.287	11.402	88	272.544	3.097
11	-0.646	11.029	-0.224	185	489.149	2.644
All Data Set				57933	28059.599	0.484

With an aim to validate whether the results of cluster analysis is representative of the general dataset or not, analysis are applied to two samples with the procedures described before. Table 38 shows the distance measures calculated during the analyses for general dataset and two samples. In Table 39 cluster centers calculated by the system for two samples are shown with corresponding information related to the clusters.

Analysis shown in Table 38 points out that the results of samples and general dataset are comparable in terms of distance measures. There are small differences between the calculated values for general dataset and two samples. However, these differences should be accepted because it is resulted from the fact that the sample sizes are different for general dataset and the samples.

Table 38 Distance Measures Comparison between General Dataset and Samples

Comparison Variable	*All dataset*	*Validation Sample_1*	*Validation Sample_2*
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	1.647	1.655	1.639
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.238	1.154	1.145
Between Clusters Manhattan Distance	6.979	6.959	7.179
Between Clusters Euclidian Distance	4.823	4.859	4.988
Average Euclidian Distances of the Cases from Cluster Center	0.484	0.484	0.483

Table 39 also confirms the consistency of the results as the cluster sizes are almost exact and the cluster centroids are very similar. Based on this information it can be concluded that results are stable within the dataset that is used for the analysis.

Table 39 Validation Results of Alternative Model One for Cluster Analysis of Customer Dataset

		Final Cluster Centers			Final Clusters Information		
	Cluster	Frequency	Recency	Total Amount	Number of Cases in the Cluster	Total Distance of the Cases from Cluster Center	Within Cluster Distance
Validation Sample One	1	0.128	-0.269	-0.093	1130	961.177	0.851
	2	-0.487	4.327	-0.226	2047	1442.523	0.705
	3	-0.603	1.498	-0.262	3690	1799.870	0.488
	4	1.772	-0.265	4.152	424	531.090	1.253
	5	3.211	-0.351	0.868	8	26.021	3.253
	6	-0.730	-0.143	-0.306	966	957.361	0.991
	7	0.639	-0.242	1.499	336	551.605	1.642
	8	2.438	-0.299	30.759	8733	3357.845	0.385
	9	1.283	-0.338	0.123	33	118.086	3.578
	10	1.863	-0.287	11.402	11314	3964.607	0.350
	11	-0.646	11.029	-0.224	94	221.395	2.355
Validation Sample Two	1	0.663	-0.246	1.552	1083	6700.416	0.384
	2	-0.600	1.487	-0.262	2045	1121.655	1.334
	3	1.252	-0.335	0.115	3808	2917.438	0.708
	4	-0.450	4.264	-0.215	425	1016.516	1.561
	5	2.856	-0.255	31.997	11	1803.226	1.017
	6	3.154	-0.341	0.852	873	8223.042	0.353
	7	1.820	-0.293	4.364	283	1821.749	0.820
	8	0.114	-0.270	-0.094	8838	111.996	5.600
	9	1.944	-0.303	11.759	41	3581.869	0.493
	10	-0.737	-0.145	-0.308	11650	272.544	3.097
	11	-0.709	11.142	-0.252	101	489.149	2.644

Alternative Model Two:

Second alternative model is built with four surrogate variables determined at the end of the factor analysis of customer data: “Recency”, “Frequency”, “Total Amount” and “Length of Relationship”

Determining Number of Clusters

In Table 40 the sum of squared values for number of clusters two to twenty is shown with related changes occurred when the number of cluster increased. These values are visualized in Figure 34. By examining the figures ten should be selected as number of clusters for Alternative Model Two because the curve in Figure 34 smoothes after k equals to ten and additional to this, in Table 40 last maximum change occurs when the number of clusters increaser to eleven from ten.

Table 40 Number of Clusters and Within Cluster Sum of Squares

Number of Clusters	*Within cluster sum of square*	*Change occurs w*hen number of cluster increases
2	164649.4	0.195
3	132573.7	0.169
4	110213.1	0.187
5	89579.27	0.118
6	79042.88	0.155
7	66766.89	0.118
8	58876.83	0.065
9	55074.89	0.066
10	51441.7	0.134
11	44550.13	0.050
12	42323.7	0.062
13	39685.05	0.080
14	36494.55	-0.002
15	36551	0.087
16	33385.03	0.041
17	32025.68	0.032
18	31011.86	0.003
19	30914.37	0.005
20	30765.33	1

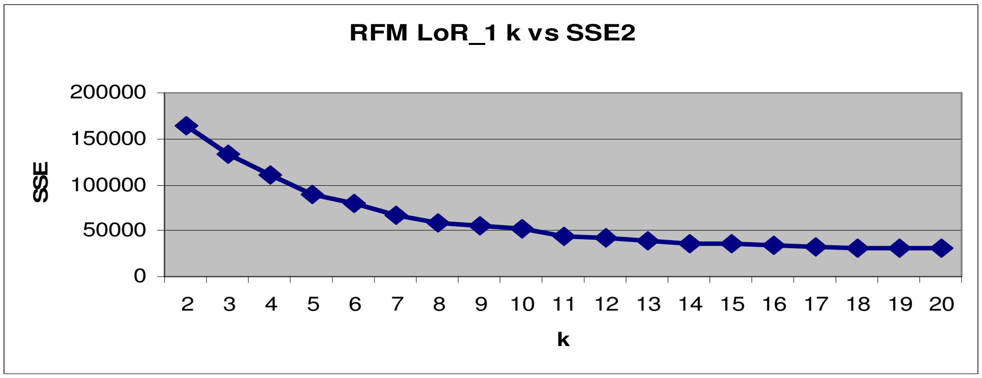


Figure 34 Number of Clusters versus Within Clusters Sum of Squares

Validation of Clusters

Table 41 shows eleven clusters with final cluster centers calculated by the system, cluster sizes and within cluster distance produced by the cases partitioned into them. At the end of the analysis Between Clusters Distance is calculated as 4.42 for Alternative Model Two. With this value Between Clusters Distance is greater than average within cluster distance, 0.706 as well as within cluster distances of all clusters except cluster eight which is the cluster formed by outliers. Based on this information it can be concluded that cluster analysis applied with four surrogate variables achieved the main goal of cluster analysis, minimize the within cluster distance when maximizing the between clusters one.

Table 41 Results of Alternative Model Two for Cluster Analysis of Customer Dataset

Cluster Number	*LoR_1*	*Frequency*	*Recency*	*Total Amount*	*Number of Cases*	*Total Euclidian Distance of the cases from Cluster Center*	*Within Cluster Distance*
1	-0.117	-0.293	0.358	1.201	861	1320.802	1.534
2	-0.489	-0.628	1.253	1.017	4073	3609.861	0.886
3	4.295	1.450	-0.359	-0.246	7709	6208.268	0.805
4	-0.229	-0.272	0.180	2.875	1240	1973.454	1.591
5	-0.117	-0.293	0.358	1.201	2145	2948.839	1.375
6	-0.489	-0.628	1.253	1.017	14337	6428.550	0.448
7	4.295	1.450	-0.359	-0.246	150	477.317	3.182
8	-0.229	-0.272	0.180	2.875	23	139.001	6.044
9	-0.117	-0.293	0.358	1.201	5647	5228.691	0.926
10	-0.489	-0.628	1.253	1.017	21564	12018.189	0.557
11	4.295	1.450	-0.359	-0.246	184	526.326	2.860
All Dataset					57933	40879.298	0.706

Validation of results' representativeness of general dataset is achieved by dividing general dataset into two samples and applying analyses to these two samples with the procedures described before. Distance measures calculated for general dataset as well as two samples are shown in Table 42. On the other hand Table 43 summarizes the results of analyses achieved with two validation samples.

Table 42 shows that results are comparable in terms of distance measures for general dataset and two samples. There are small differences for the calculated values which are resulted from the different objects included in the analysis and as a result of this should be accepted.

Table 42 Distance Measures Comparison between General Dataset and Samples

Comparison Variable	*All dataset*	*Validation Sample_1*	*Validation Sample_2*
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	2.376	2.386	2.376
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.465	1.469	1.460
Between Clusters Manhattan Distance	7.314	6.562	6.755
Between Clusters Euclidian Distance	4.424	3.827	4.016
Average Euclidian Distances of the Cases from Cluster Center	0.706	0.753	0.700

Table 43 also confirms the consistency of the results as the cluster sizes are almost exact and the cluster centroids are very similar. Based on information Table 43 includes it can be concluded that results are stable within the dataset that is used for the analysis.

Table 43 Validation Results of Alternative Model Two for Cluster Analysis of Customer Dataset

	Final Cluster Centers					Final Clusters Information		
	Cluster	LOR_1	Frequency	Recency	Total Amount	Number of Cases in the Cluster	Total Distance of the Cases from Cluster Center	Within Cluster Distance
Validation Sample One	1	2.382	2.854	-0.328	0.965	1092	1501.029	1.375
	2	0.516	-0.532	10.150	-0.201	100	267.053	2.671
	3	1.328	1.926	-0.299	24.473	14	94.088	6.721
	4	1.819	2.222	-0.196	7.033	113	306.826	2.715
	5	-0.330	-0.664	1.231	-0.278	2123	1705.581	0.803
	6	1.530	0.234	-0.094	0.030	2712	2494.112	0.920
	7	-0.127	-0.471	3.675	-0.219	607	874.352	1.440
	8	1.153	0.916	-0.232	2.536	696	1021.169	1.467
	9	0.352	1.250	-0.358	0.172	1806	3034.057	1.680
	10	-0.067	-0.162	-0.220	-0.147	10571	5813.090	0.550
	11	-1.101	-0.834	-0.203	-0.340	6941	3049.320	0.439
Validation Sample Two	1	2.400	2.798	-0.302	0.966	1001	1377.891	1.377
	2	0.712	-0.683	10.662	-0.256	115	349.676	3.041
	3	1.091	2.516	-0.265	26.528	14	142.585	10.185
	4	1.513	2.107	-0.326	7.662	111	345.736	3.115
	5	-0.310	-0.662	1.235	-0.286	2127	1716.520	0.807
	6	1.520	0.242	-0.094	0.028	2880	2646.789	0.919
	7	-0.090	-0.427	3.697	-0.212	590	875.153	1.483
	8	1.112	0.901	-0.244	2.499	660	954.293	1.446
	9	0.360	1.253	-0.359	0.159	3830	3021.294	0.789
	10	-0.067	-0.159	-0.225	-0.147	10688	5854.919	0.548
	11	-1.095	-0.832	-0.211	-0.339	7142	3127.726	0.438

Alternative Model Three:

Alternative Model Three is built with all surrogate variables determined at the end of the Factor Analysis.

Determining Number of Clusters

Table 44 shows the sum of square values for a number of clusters ranging two to twenty, and change that occurs when the cluster number increased by one. In Figure 35 these values are plotted against number of clusters. In Table 44, it is shown that when the number of clusters is increased to nine from the eight last maximum changes occurs. Additional to this, Figure 35 shows that after eight clusters the curve becomes smoother. Based on this information eight is selected as number of clusters that will be used for the rest of analysis.

Table 44 Number of Clusters and Within Cluster Sum of Squares

Number of Clusters	*Within cluster sum of square*	*Change occurs when* number of cluster increases
2	222402.405	0.144
3	190320.700	0.117
4	167980.017	0.171
5	139257.028	0.128
6	121495.692	0.095
7	109988.508	0.091
8	100003.957	0.068
9	93161.160	0.044
10	89055.383	0.037
11	85784.146	0.067
12	78737.325	0.040
13	75563.686	0.066
14	70542.222	0.030
15	68456.518	0.030
16	66388.962	0.021
17	64991.404	0.027
18	63219.265	0.065
19	59122.867	0.020
20	57945.922	1.000

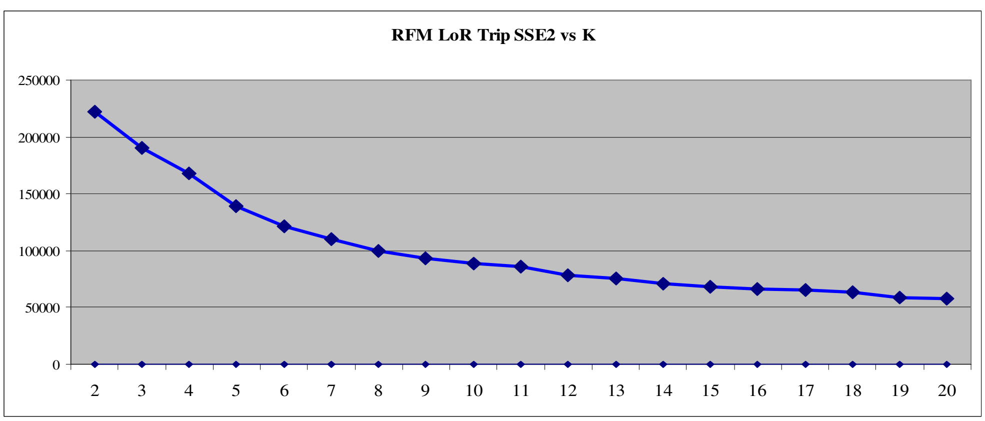


Figure 35 Number of Clusters versus Within Clusters Sum of Squares

Validation of Clusters

At the end of the partitioning process applied with five surrogate variables customers of the company are partitioned into eight different clusters. Table 45 shows these clusters with corresponding final cluster centers, cases partitioned into them and within cluster distance produced by these cases. Between Euclidian Clusters Distance of Alternative Model Three is calculated as 4.92 which is greater from all within cluster distances except the one for cluster three whose members are all outliers. Between Euclidian Cluster Distance value 4.92 is greater than Average within Clusters Distance of 1.139. Comparison result show that cluster analysis applied with five surrogate variables achieved the main goal of cluster analysis: to minimize the within cluster distance when maximizing the between clusters one.

Table 45 Results of Alternative Model Three for Cluster Analysis of Customer Dataset

Cluster Number	*LoR_1*	*Frequency*	*Recency*	*Total Amount*	*rMajorTrip*	*Number of Cases*	*Total Euclidian Distance of the cases from Cluster Center*	*Within Cluster Distance*
1	-0.140	-0.533	2.432	-0.238	0.074	2941	4677.124	1.590
2	-0.569	-0.581	-0.083	-0.254	0.924	15632	15443.096	0.988
3	1.223	2.214	-0.277	24.666	0.010	36	266.194	7.394
4	1.701	1.707	-0.274	0.496	-0.115	6464	10300.686	1.594
5	0.486	-0.544	9.451	-0.205	0.157	292	933.178	3.196
6	-0.782	-0.706	-0.134	-0.316	-1.158	11397	11660.652	1.023
7	0.265	0.281	-0.249	0.000	-0.037	20152	20020.444	0.993
8	1.660	2.045	-0.267	4.150	-0.011	1019	2682.378	2.632
All Data Set						57933	65983.752	1.139

In order to validate whether the results of cluster analysis is representative of the general dataset or not, analyses are applied to two samples with the procedures described before. Table 46 shows the distance measures values calculated by SPSS for general dataset as well as for the two samples. In Table 47 information related to the clusters constructed at the end of partitioning process is shown for two samples with corresponding final cluster centers.

Table 47 shows that the results are comparable in terms of distance measures. There are small differences for the calculated values. However, this difference should be accepted because it is resulted from the fact that the objects included in the analyses are for general dataset and the samples.

Table 46 Distance Measures Comparison between General Dataset and Samples

Comparison Variable	*All dataset*	*Validation Sample_1*	*Validation Sample_2*
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	3.140	2.464	2.443
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.771	1.474	1.465
Between Clusters Manhattan Distance	6.452	5.769	5.837
Between Clusters Euclidian Distance	4.094	3.835	3.873
Average Euclidian Distances of the Cases from Cluster Center	1.139	0.791	0.791

Table 47 also confirms the consistency of the results as the cluster sizes are almost exact and the cluster centroids are very similar. With this information it can be concluded that results are stable within the dataset that is used for the analysis.

Table 47 Validation Results of Alternative Model Three for Cluster Analysis of Customer Dataset

	Final Cluster Centers						Final Clusters Information		
	Cluster	LOR_2	Frequency	Recency	Total Amount	rMajor Trip	Number of Cases in the Cluster	Total Distance of the Cases from Cluster Center	Within Cluster Distance
Validation Sample One	1	1.678	2.010	-0.270	3.992	-0.017	538	1243.267	2.311
	2	0.069	-0.191	-0.153	-0.143	-0.009	11642	7671.343	0.659
	3	2.417	1.713	-0.209	0.552	-0.030	1833	2510.780	1.370
	4	0.480	1.089	-0.334	0.244	-0.020	5089	4699.671	0.923
	5	-1.057	-0.812	-0.152	-0.334	-0.030	7787	4009.909	0.515
	6	0.329	-0.550	8.706	-0.184	0.006	162	443.202	2.736
	7	-0.223	-0.575	2.218	-0.252	-0.009	1705	2064.134	1.211
	8	1.486	1.970	-0.294	22.892	-0.011	19	120.242	6.329
Validation Sample Two	1	1.570	1.997	-0.266	4.113	-0.023	495	1251.657	2.529
	2	0.073	-0.189	-0.158	-0.146	-0.010	11750	7717.319	0.657
	3	2.405	1.619	-0.192	0.541	-0.032	1804	2439.220	1.352
	4	0.497	1.078	-0.335	0.226	-0.019	5218	4753.911	0.911
	5	-1.052	-0.812	-0.160	-0.334	-0.032	7983	4090.791	0.512
	6	0.481	-0.536	9.084	-0.251	-0.008	180	547.378	3.041
	7	-0.194	-0.572	2.195	-0.253	-0.012	1704	2067.506	1.213
	8	1.080	2.143	-0.243	22.924	-0.014	24	208.185	8.674

Selection between Alternative Models:

Selection between the alternative models is made according to the manageability of results with the help of basic objective of clustering procedure: minimizing the within cluster distance when maximizing the between clusters one. Analysis made on the final cluster centers of three alternative models indicates that, Alternative Model Three with eight clusters is the most manageable one among others. The solution has detected the outliers better than other two alternative models and gave more manageable results. On the other hand, Table 13 shows the Average within Cluster Distance and Between Cluster Distance measures values in Euclidian base for three alternative models build and number of clusters determined for each model. As it is shown in Table 48, Alternative Model One has the smallest value in terms of Average within Cluster Distance measure, when Alternative Model Three has the greatest one. Although the aim of the clustering is to find the solution which ensures the minimum within cluster distance, given that the alternative models are built with different number of surrogate variables and resulted with different number of clusters, it is not logical to compare these values by themselves to select the most useful alternative model. When the Between Clusters distance measure values are analyzed, it is seen that Alternative Model Three again has the greatest value compared to other two models. With a bigger Between Cluster Distance Alternative Model Three ensures that cases in the clusters obtained at the end of the analysis are different from each other.

Table 48 Distance Measures for Alternative Models

Alternative Model	*Number of Clusters*	*Average Within Cluster Distance*	*Between Clusters Distance*
1	11	0.484	4.823
2	11	0.706	4.424
3	8	1.139	4.924

Additionally, Alternative Model Three resulted with the most manageable customer segments compared to the other two models. According to the results of these analyses Alternative Model Three is selected as the most useful model in partitioning the customers of company into smaller manageable groups.

Cluster Analysis to Segment Cities Company Performs

Three different alternative clustering models are built with surrogate variables determined at the end of factor analysis, in order to find the most useful one in partitioning the cities in which company performs into small manageable groups for customer relationship management projects and activities.

Alternative clustering models built in this analysis will be analyzed in the following sections.

Alternative Model One:

First alternative model is built with variables proposed by the literature as variables of basic model, RFM. Although "Frequency" and "Recency" not loaded to a factor with highest loadings, since RFM model is selected as the base model of this study, variables at issue are used in first alternative model with "Average Sales City" variable which is selected as surrogate variable of Second Factor.

Determining Number of Clusters

SSEs for different number of clusters from two to twenty are shown in Table 49 with the changes that occur when the number of clusters increased by one. In Figure 36 these values are plotted against number of clusters.

By analyzing Table 49 and Figure 36, number of clusters is selected as seven for Alternative Model One.

Table 49 Number of Clusters and Within Cluster Sum of Squares

Number of Clusters	*Within cluster sum of square*	*Change occurs w*hen number of cluster increases
2	166.323	0.348
3	108.460	0.343
4	71.279	0.125
5	62.362	0.184
6	50.882	0.213
7	40.056	0.079
8	36.906	0.110
9	32.836	0.185
10	26.755	0.000
11	26.759	0.097
12	24.152	0.033
13	23.351	0.192
14	18.864	0.148
15	16.076	-0.020
16	16.405	0.204
17	13.066	-0.022
18	13.355	0.199
19	10.697	0.109
20	9.536	1

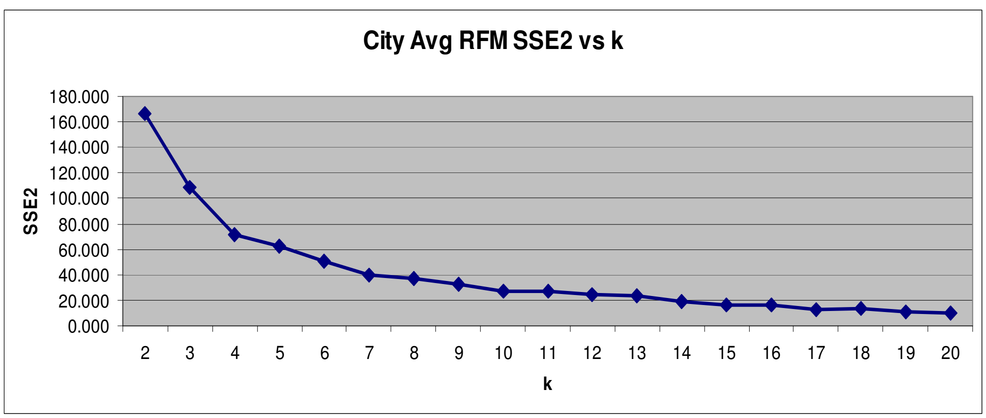


Figure 36 Number of Clusters versus Within Clusters Sum of Squares

Validation of Clusters

Table 50 shows seven clusters with corresponding final cluster centers, number of cases partitioned into them and within cluster distance produced by these cases. Between Clusters Distance measure is calculated as 1.76 for this Alternative Model One. Value at issue is greater than the within cluster distance of seven clusters as well as average within cluster distances shown in Table 50. This indicates that Alternative Model One achieves the basic goal of cluster analysis, minimizing the within cluster distance when maximizing the between clusters one.

Table 50 Results of Alternative Model One for Cluster Analysis of City Dataset

Cluster Number	*Average Frequency*	*Average Recency*	*Average Sales 2*	*Number of Cases*	*Total Euclidian Distance of the cases from Cluster Center*	*Within Cluster Distance*
1	-1.741	4.282	-0.103	2	0.632	0.316
2	-0.283	-0.296	0.957	11	9.198	0.836
3	-0.395	-0.009	-0.428	33	18.744	0.568
4	-2.250	-1.297	4.795	2	1.266	0.633
5	1.102	-0.602	-0.260	22	13.628	0.619
6	1.257	0.973	0.284	4	3.644	0.911
7	-1.284	1.736	-0.298	4	2.955	0.739
All Data Set				78	50.067	0.642

In Table 51 distance measures calculated during the analyses made with general city dataset and two samples are shown. On the other hand Table 52 shows the final cluster centers calculated by the system at the end of the partitioning processes of two samples and related clusters information.

Table 51 shows that general dataset and two samples do not have similar values with respect to different distance measures. In spite of this results of samples and general dataset are accepted as comparable in terms of distance measures because the sample sizes of the analyses are small and different for general dataset and the samples.

Table 51 Distance Measures Comparison between General Dataset and Samples

Comparison Variable	*All dataset*	*Validation Sample_1*	*Validation Sample_2*
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	2.017	1.845	2.192
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	2.962	2.311	3.610
Between Clusters Manhattan Distance	2.738	2.633	2.945
Between Clusters Euclidian Distance	3.679	3.213	4.475
Average Euclidian Distances of the Cases from Cluster Center	0.484	0.484	0.483

Table 52 shows that although numbers of cases partitioned into clusters are similar to each other for two samples, cluster centroids are not similar. With this information it can be conclude that results are not stable for general dataset and two samples.

Table 52 Validation Results of Alternative Model One for Cluster Analysis of City Dataset

		Final Cluster Centers			Final Clusters Information		
	Cluster	Average Frequency	Average Recency	Average Sales 2	Number of Cases in the Cluster	Total Distance of the Cases from Cluster Center	Within Cluster Distance
Validation Sample One	1	1.679	1.338	-0.540	1	0.000	0.000
	2	-0.299	-0.009	-0.300	21	13.486	0.642
	3	-0.822	-1.111	2.561	1	0.000	0.000
	4	-2.118	-0.709	4.600	1	0.000	0.000
	5	-1.114	1.691	-0.364	3	2.378	0.793
	6	1.325	1.393	1.726	1	0.000	0.000
	7	0.855	-0.555	-0.194	11	6.181	0.562
Validation Sample Two	1	0.971	0.720	0.121	1	1.149	1.149
	2	-0.453	-0.121	-0.123	20	15.247	0.762
	3	-0.443	0.436	2.045	1	1.675	1.675
	4	-2.382	-1.884	4.991	1	1.266	1.266
	5	-1.759	3.479	-0.102	3	6.129	2.043
	6	.	.	.	0	0.000	NA
	7	1.256	-0.536	-0.248	13	9.692	0.746

Alternative Model Two:

Alternative Model Two is built with "Average Frequency", "Average Recency", "Average Sales" and "Count of Customers" variables. "Count of Customer" variables is selected as surrogate variable of factor three at the end of the factor analysis.

Determining Number of Clusters

By analyzing Table 53 and Figure 37 twelve is selected as number of clusters for Alternative Model Two.

Table 53 Number of Clusters and Within Cluster Sum of Squares

Number of Clusters	*Within cluster sum of square*	*Change occurs w*hen number of cluster increases
2	98.819	0.337
3	65.487	0.110
4	58.271	0.373
5	36.563	0.084
6	33.483	0.259
7	24.800	0.133
8	21.501	0.252
9	16.073	0.244
10	12.145	0.186
11	9.884	0.004
12	9.845	0.344
13	6.455	0.218
14	5.051	0.090
15	4.597	0.082
16	4.220	0.101
17	3.795	0.165
18	3.171	0.172
19	2.625	0.181
20	2.151	1

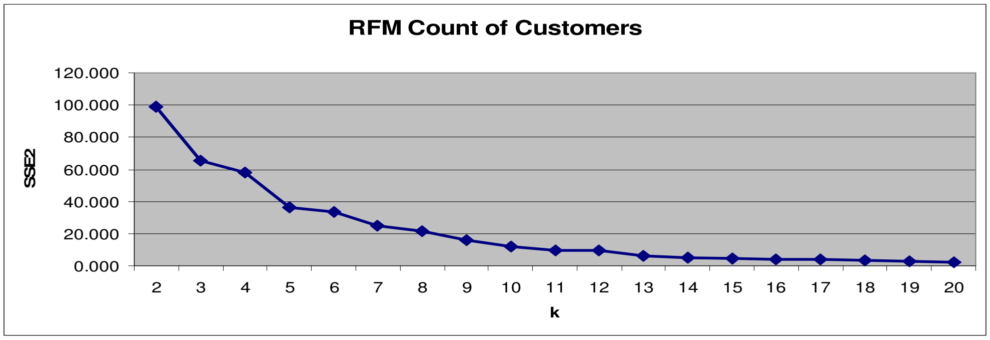


Figure 37 Number of Clusters versus Within Clusters Sum of Squares

Validation of Clusters

Twelve clusters with their final cluster centers number of cases partitioned into them and related total and within cluster distances are shown in Table 54. Between Clusters Euclidian Distance is calculated as 1.89 for Alternative Model Two which is greater than all twelve within clusters distances and average within clusters distances. With this information it can be concluded that Alternative Model Two achieves the basic goal of cluster analysis just like the previous ones.

Table 54 Results of Alternative Model Two for Cluster Analysis of City Dataset

Cluster Number	*Average Frequency*	*Average Recency*	*Average Sales 2*	*Count Of Customers*	*Number of Cases*	*Total Euclidian Distance of the cases from Cluster Center*	*Within Cluster Distance*
1	1.379	-0.141	0.030	6.692	1	0.000	0.000
2	1.234	0.833	-0.196	-0.308	3	1.626	0.542
3	1.325	1.393	1.726	-0.342	1	0.000	0.000
4	-0.443	-0.037	0.026	-0.305	28	20.9	0.746
5	-2.250	-1.297	4.795	-0.384	2	1.266	0.633
6	0.788	-0.167	0.084	1.164	2	0.764	0.382
7	-0.242	-0.445	-0.490	3.203	2	0.238	0.119
8	-1.215	1.524	-0.381	-0.350	5	3.991	0.798
9	-0.015	-0.204	-0.668	0.193	13	7.334	0.564
10	-1.741	4.282	-0.103	-0.384	2	0.632	0.316
11	1.201	-0.692	-0.264	-0.244	17	10.691	0.629
12	-0.632	-0.337	2.303	-0.354	2	1.676	0.838
All Data Set					78	49.117	0.630

Validation of results' representativeness of general dataset is achieved by dividing general dataset into two samples and applying analyses to these two samples with the procedures described before. Distance measures calculated for general dataset as well as two samples are shown in Table 55. On the other hand Table 56 summarizes the results of analyses achieved with two validation samples.

Figures in Table 55 show that results are more comparable in terms of distance measures for general dataset and two samples than the ones for Alternative Model One. The small differences for the calculated values result from using different objects for the analyses and as a result of this should be accepted.

Table 55 Distance Measures Comparison between General Dataset and Samples

Comparison Variable	*Value for All dataset*	*Value for Validation Sample_1*	*Value for Validation Sample_2*
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	2.511	2.569	2.444
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.565	1.579	1.565
Between Clusters Manhattan Distance	3.443	3.190	2.748
Between Clusters Euclidian Distance	1.892	1.827	1.104
Average Euclidian Distances of the Cases from Cluster Center	0.630	0.417	0.566

However, Table 56 shows that there are differences between the results of two samples in terms of cluster sizes and cluster centroids. Since the sample sizes are small when the number of clusters is selected same with the general dataset, some clusters are remain without any case partitioned into them. The fact causes difference between the cluster centroids. It can be said that with similar values for sample one and sample two in both Table 55 and Table 56 Alternative Model Two is more stable than Alternative Model One. However, analyses do not confirm the stability of results.

Table 56 Validation Results of Alternative Model Two for Cluster Analysis of City Dataset

	Final Cluster Centers					Final Clusters Information		
	Cluster	*Average Frequency*	*Average Recency*	*Average Sales 2*	*Count Of Customers*	Number of Cases in the Cluster	Total Distance of the Cases from Cluster Center	Within Cluster Distance
Validation Sample One	1	1.325	1.393	1.726	-0.342	1	0.000	0.000
	2	-0.726	0.508	-0.339	-0.283	11	7.468	0.679
	3	-0.822	-1.111	2.561	-0.328	1	0.000	0.000
	4	-2.118	-0.709	4.600	-0.384	1	0.000	0.000
	5	-0.064	-0.232	-0.710	0.556	4	1.742	0.436
	6	-0.418	2.433	-0.645	-0.385	1	0.000	0.000
	7	1.379	-0.141	0.030	6.692	1	0.000	0.000
	8	1.679	1.338	-0.540	-0.372	1	0.000	0.000
	9	-0.242	-0.445	-0.490	3.203	2	0.238	0.119
	10	1.238	-0.497	-0.458	-0.063	4	1.595	0.399
	11	0.049	-0.254	0.448	-0.308	5	1.777	0.355
	12	0.338	-0.576	-0.251	-0.276	7	3.449	0.493
Validation Sample Two	1	.	.	.	.	0	0.000	NA
	2	-0.788	0.157	-0.419	-0.245	7	2.863	0.409
	3	-0.443	0.436	2.045	-0.379	1	0.000	0.000
	4	-2.382	-1.884	4.991	-0.385	1	0.000	0.000
	5	0.339	-0.253	-0.305	0.936	4	2.901	0.725
	6	-1.759	3.479	-0.102	-0.358	3	3.311	1.104
	7	.	.	.	.	0	0.000	NA
	8	1.011	0.580	-0.024	-0.276	2	0.420	0.210
	9	.	.	.	.	0	0.000	NA
	10	1.463	-0.732	-0.266	-0.297	9	5.682	0.631
	11	-0.489	-0.382	0.655	-0.362	6	4.295	0.716
	12	0.023	-0.173	-0.483	-0.246	6	2.591	0.432

Alternative Model Three:

In the third alternative model second variable loaded to factor three is included in the analysis instead of "Count of Customers" variable and model is build with "Average Frequency", "Average Recency", "Average Sales" and "Sales per Customer" variables.

Determining Number of Clusters

By analyzing Table 57 and Figure 38 seven is selected as number of clusters for Alternative Model Three.

Table 57 Number of Clusters and Within Cluster Sum of Squares

Number of Clusters	*Within cluster sum of square*	*Change occurs* when number of cluster increases
2	3.094	0.280
3	2.227	0.262
4	1.642	0.280
5	1.183	0.116
6	1.045	0.240
7	0.795	0.043
8	0.760	0.072
9	0.706	0.110
10	0.628	0.114
11	0.556	0.086
12	0.508	0.036
13	0.490	0.121
14	0.431	0.054
15	0.407	0.053
16	0.386	0.073
17	0.358	0.047
18	0.341	0.042
19	0.327	1.000
20	3.094	0.280

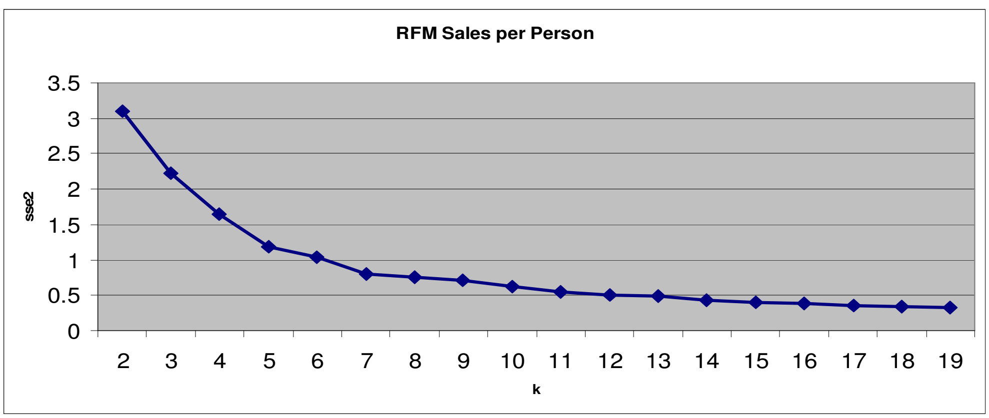


Figure 38 Number of Clusters versus Within Clusters Sum of Squares

Validation of Clusters

Table 58 shows these clusters with corresponding final cluster centers, cases partitioned into them and within cluster distance produced by these cases. Between Euclidian Clusters Distance of Alternative Model Three is calculated as 1.80 which is greater from all within cluster distances. Between Euclidian Cluster Distance is 4.92 for Alternative Model Three and greater than Average within Clusters Distance, 0.88. Based on this information it can be concluded that cluster analysis applied with four surrogate variables achieved the main goal of cluster analysis, minimize the within cluster distance when maximizing the between clusters one.

Table 58 Results of Alternative Model Three for Cluster Analysis of City Dataset

Cluster Number	*Average Frequency*	*Average Recency*	*Average Sales 2*	*Sales per Customer City*	*Number of Cases*	*Total Euclidian Distance of the cases from Cluster Center*	*Within Cluster Distance*
1	1.028	-0.588	-0.277	0.011	22	19.366	0.880
2	-0.434	-0.023	-0.171	-0.360	34	27.879	0.820
3	-1.774	-1.235	4.051	-0.689	3	3.933	1.311
4	-1.741	4.282	-0.103	-0.844	2	0.632	0.316
5	-1.284	1.736	-0.298	-0.669	4	3.059	0.765
6	0.883	0.972	0.838	-0.626	4	5.372	1.343
7	0.283	-0.221	-0.245	2.325	9	9.006	1.001
All Data Set					78	69.247	0.888

In order to validate whether the results of cluster analysis is representative of the general dataset or not, analysis are applied to two samples with the procedures described before. Table 59 shows the distance measures calculated for general dataset as well as for the two samples. In Table 60 information related to the clusters constructed at the end of partitioning process is shown for two samples with corresponding final cluster centers.

Table 59 shows that there are smaller differences between the values compared to other alternative models. Based on Table 60 it can be concluded that results are comparable in terms of distance measures. There are small differences for the calculated values. However, this difference should be accepted because it is resulted from the fact that the different objects are used for the analyses of general dataset and the samples.

Table 59 Distance Measures Comparison between General Dataset and Samples

Comparison Variable	*All dataset*	*Validation Sample_1*	*Validation Sample_2*
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	2.742	2.595	2.902
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.665	1.566	1.766
Between Clusters Manhattan Distance	3.072	3.038	3.328
Between Clusters Euclidian Distance	1.805	1.920	2.012
Average Euclidian Distances of the Cases from Cluster Center	0.888	0.755	1.031

Results in Table 60 also confirm the consistency of the results as the cluster sizes are almost exact and the cluster centroids are very similar. It is obvious that with similar values for sample one and sample two in both Table 59 and Table 60 Alternative Model Three is more stable than Alternative Model One and Two. Based on these results it can be concluded that results are stable within the dataset that is used for the analysis.

Table 60 Validation Results of Alternative Model Three for Cluster Analysis of City Dataset

	Final Cluster Centers					Final Clusters Information		
	Cluster	*Average Frequency*	*Average Recency*	*Average Sales 2*	*Sales per Customer City*	*Number of Cases in the Cluster*	*Total Distance of the Cases from Cluster Center*	*Within Cluster Distance*
Validation Sample One	1	-1.114	1.691	-0.364	-0.759	3	2.400	0.800
	2	0.721	-0.408	-0.284	0.028	14	12.532	0.895
	3	-0.822	-1.111	2.561	-0.438	1	0.000	0.000
	4	-2.118	-0.709	4.600	-0.786	1	0.000	0.000
	5	0.226	-0.280	-0.400	2.960	4	3.962	0.990
	6	1.325	1.393	1.726	-0.433	1	0.000	0.000
	7	-0.412	0.125	-0.226	-0.390	15	10.539	0.703
Validation Sample Two	1	-1.759	2.009	-0.102	-0.696	3	6.213	2.071
	2	1.103	-0.424	-0.239	0.012	14	15.772	1.127
	3	-0.443	0.436	2.045	-0.676	1	1.692	1.692
	4	-2.382	-1.884	4.991	-0.844	1	1.267	1.267
	5	0.420	-0.272	-0.326	1.965	4	4.892	1.223
	6	.	.	.	.	0	0.000	NA
	7	-0.555	-0.103	-0.057	-0.429	16	14.280	0.893

Selection between Alternative Models:

Selection between the alternative models is made according to the manageability of results with the help of basic objective of clustering procedure: minimizing the within cluster distance when maximizing the between clusters one. Different from the selection between alternative models for customer segmentation, in this part of study stability of the models based on the results of validation analysis also has effect on the selection of alternative model to partition the cities.

Analysis made on the final clusters produced by three alternative models indicates that, Alternative Model Three with seven clusters is the most manageable one among others. On another side, Table 61 shows the Average within Cluster Distance and Between Cluster Distance measures values in Euclidian base for three alternative models build and number of clusters determined for each model. Table 61 shows that Alternative Model Two has the smallest value in terms of Average within Cluster distance. On the other hand Alternative Model three has the greatest one. Although aim of the clustering is to find the solution which ensures the minimum within cluster distance, given that these two alternative models resulted with different number of clusters, it is not logical to compare these values by them selves to select the most useful alternative model. Since Alternative Model Two partitioned cases into greater number of clusters, there are of course less number of cases in each cluster which makes the within cluster distance smaller. When the Between Clusters distance measure values are analyzed, it is seen that Alternative Model Two has the greatest value compared to other two models. This difference again resulted from the different number of clusters selected for each alternative model.

Table 61 Distance Measures for Alternative Models

Alternative Model	*Number of Clusters*	*Average Within Cluster Distance*	*Between Clusters Distance*
1	7	0.642	1.76
2	12	0.630	1.89
3	7	0.88	1.80

Additional to this, Alternative Model three is the most stable model among others based on the results of validation analyses. Based on all this information Alternative Model Three is selected as the most useful one in partitioning the cities into smaller manageable groups for CRM activities of company.

CHAPTER 7

CLUSTER INTERPRETATIONS FOR PROFILING

In this chapter clusters obtained at the end of the cluster analyses, which are discussed in chapter six, are interpreted and profiles of them are defined using the characteristics of customers included in these clusters. Clusters are interpreted by analyzing the characteristics of clusters which are grouped under two main topics, namely, general characteristics and characteristics related to continuous variables. Additional to this in the interpretation of customer clusters another perspective: characteristics related to categorical variables are used.

- General characteristics

Under the general characteristics topic, the size and the wideness of the cluster as well as the possibility of being outliers for the cluster members are analyzed with reference to the statistics calculated for each cluster. The wideness of a cluster is evaluated by analyzing the variables indicating the distance of the cases from cluster center in the aggregate and on average. Clusters which have greater values on average are accepted as wider than the ones with smaller values. The possibility of being outliers for the cluster members is evaluated by analyzing the variables measuring the distance of the cluster center form the center of the all clusters. The distance between the cluster center and the overall center are calculated both on Euclidian and Manhattan bases. In addition, the Between Clusters Distance is computed, again on both Euclidian and Manhattan bases, and this statistics is also interpreted by comparing it to the distances between cluster centers and the overall center. Clusters which are far away from the overall center, compared to other clusters by using the between clusters distance as a benchmark distance, are accepted as the ones that have greater possibility of containing outlier cases.

- <u>Characteristics related to continuous variables</u>

The second topic of interpretation, characteristics related to continuous variables, includes analyses comparing the differences between clusters in terms of the considered continuous variables. In order to define the variables that will be used to interpret clusters one way Analysis of Variance (ANOVA) test applied to the dataset. Variables used for the clustering process and some control variables are analyzed in this analysis. ANOVA test produce the p values by comparing the between clusters variance and within cluster variances for each variable used in analysis. With ninety five percent of confidence level, if the p value is less than 0.05, it means that there is a significant difference between clusters with respect to the variable that is being analyzed. Otherwise, this variable should not be used to interpret the clusters. ANOVA test assumes that the variables analyzed have equal variances. If the opposite case occurs different methods of ANOVA test should be used to make the multiple comparisons.

For interpretation purposes, cluster centers are compared to the mean of the variable for the general dataset. In addition, the maximum and minimum of the cluster means are analyzed. Unique tables are prepared for each cluster to make the comparisons mentioned above. Left part of these tables contains the cluster center, rank of cluster for this variable among all clusters. Additionally, to make the comparison easier, the maximum and minimum values of the cluster centers as well as the mean of the variable for the general dataset are also included in these tables. In order to determine whether the specified cluster is significantly different form the other clusters with respect to each variable ANOVA tests is applied again with Tamhane's 2 method and clusters are compared one by one. P-values are shown in right part of tables.

- Characteristics related to categorical variables

Categorical variables that are listed in Table 3 are used by the company to define characteristics of their customers. Since the dataset being analyzed contains categorical variables in order to determine the ones that will be used to define the clusters produced at the end of the partitioning process the chi-square (c2), contingency test is used. A chi-square analysis is used to calculate the probability that a relationship found in a sample between two variables is due to chance (random sampling error). It does this by measuring the difference between the actual frequencies in each cell of a table and the frequencies one would expect to find if there were no relationship between the variables in the population from which the (simple random) sample has been drawn. If the actual counts are different from the expected counts the system calculated, p value of the test becomes smaller than 0.05 with ninety five percent confidence level. This means that these subgroups are significantly different from each other by means of specified variables. The larger these differences are, the less likely it is that they occurred by chance.

The contingency test has some restrictions on its use such as: when there are only two categories, no expected value may be smaller than five and when there are more than two categories, no more than 20% of the expected values may be smaller than five, and no expected value may be smaller than one.

Interpretation of Customer Clusters

Based on the information gained from segmentation analysis discussed in Chapter Six, customer base is partitioned into eight different segments. Table 62 shows the clusters with number of cases partitioned into them and corresponding percentage of their size compared to all dataset.

Table 62 Number of Cases in Customer Clusters

Cluster Number	*Number of Cases in Cluster*	*Percentage of Data in Cluster*
1	2941	5.08 %
2	15632	26.98 %
3	36	0.06 %
4	6464	11.16 %
5	292	0.50 %
6	11397	19.67 %
7	20152	34.79 %
8	1019	1.76 %
Total	57933	100 %

Cluster interpretation summaries can be found in Appendix B.

- Characteristics related to continuous variables

Table 63 shows the result of the homogeneity variances test which is applied to the dataset in order to analyze whether the variances of the variables are equal or not. Facts in table shows that variances of the variables that are being analyzed are not equal and special methods should be used to make the multiple comparisons.

Table 63 Test of Homogeneity Variances

Test of Homogeneity of Variances				
	Levene Statistic	df1	df2	Sig
LoR_1	1,213.4512	7	57,925	0.000
Frequency	2,172.5032	7	57,925	0.000
rFrequency	58.4277	7	57,925	0.000
Frequency last one year	882.5477	7	57,925	0.000
Recency	3,990.4165	7	57,925	0.000
IPT	4,693.7085	7	57,925	0.000
Amount	551.8920	7	57,925	0.000
Total Amount	5,145.4411	7	57,925	0.000
R Amount	57.2589	7	57,925	0.000
rMajorTrip	269.2200	7	57,925	0.000
R Total Amount	1,437.1562	7	57,925	0.000

Based on the results of Homogeneity variances test, Tamhane's T2 Multiple Comparison method that assumes variances of the variables are not equal, is used in the analysis to make comparisons. Table 64 shows the results of the ANOVA analysis applied to the dataset. Figures in Table 64 indicate that these variables show significant differences between the clusters and can be used to interpret them.

Table 64 Significance Testing of Variables

ANOVA Analysis of Variables						
		Sum of Squares	df	Mean Square	F	Sig.
LoR_1	Between Groups	35145.4016	7	5020.7717	12763.1248	0.0000
	Within Groups	22786.5984	57925	0.3934		
	Total	57932.0000	57932			
Frequency	Between Groups	36749.3724	7	5249.9103	14356.1536	0.0000
	Within Groups	21182.6276	57925	0.3657		
	Total	57932.0000	57932			
Frequency last year	Between Groups	20207.0275	7	2886.7182	4432.4261	0.0000
	Within Groups	37724.9725	57925	0.6513		
	Total	57932.0000	57932			
Recency	Between Groups	45608.9610	7	6515.5659	30626.7107	0.0000
	Within Groups	12323.0390	57925	0.2127		
	Total	57932.0000	57932			
IPT	Between Groups	14788.6586	7	2112.6655	2836.5014	0.0000
	Within Groups	43143.3414	57925	0.7448		
	Total	57932.0000	57932			
Total Amount	Between Groups	43361.3964	7	6194.4852	24625.9911	0.0000
	Within Groups	14570.6036	57925	0.2515		
	Total	57932.0000	57932			
rMajorTrip	Between Groups	28790.9121	7	4112.9874	8175.5629	0.0000
	Within Groups	29141.0879	57925	0.5031		
	Total	57932.0000	57932			
Amount	Between Groups	11569.9656	7	1652.8522	2065.0834	0.0000
	Within Groups	46362.0344	57925	0.8004		
	Total	57932.0000	57932			
rFrequency	Between Groups	7099.5000	7	1014.2143	1155.7244	0.0000
	Within Groups	50832.5000	57925	0.8776		
	Total	57932.0000	57932			
rAmonut	Between Groups	210.9812	7	30.1402	30.2467	0.0000
	Within Groups	57721.0188	57925	0.9965		
	Total	57932.0000	57932			
rTotal Amount	Between Groups	21589.2230	7	3084.1747	4911.7081	0.0000
	Within Groups	36372.4424	57925	0.6279		
	Total	57961.6654	57932			

Clusters' centroids will be used as a guide to interpret them. Since the dataset is transformed before the partition process start, z-scores for the clusters' centroids converted to original variables for interpretation. Table 65 shows final cluster centers determined by the system with the corresponding original values for the variables used in the segmentation. On the other hand, Table 66 shows the z-scores and original values for the other variables that will be used in interpretations.

Table 65 Final Cluster Centers in z-values and Original Values for Segmentation Variables

Final Cluster Centers									
		Cluster							
		1	2	3	4	5	6	7	8
LoR_1	z-value	-0.140	-0.569	1.223	1.701	0.486	-0.782	0.265	1.660
	Original value	357.310	249.039	701.000	821.524	515.099	195.276	459.331	811.083
Frequency	z-value	-0.533	-0.581	2.214	1.707	-0.544	-0.707	0.281	2.045
	Original value	34.166	31.265	200.361	169.691	33.521	23.671	83.418	190.171
Recency	z-value	2.432	-0.083	-0.277	-0.275	9.451	-0.134	-0.249	-0.267
	Original value	91.964	12.336	6.194	6.285	314.106	10.731	7.079	6.508
Total Amount	z-value	-0.238	-0.254	24.666	0.496	-0.205	-0.316	0.000	4.150
	Original value	4085.8	3696.3	621207.8	22270.9	4914.5	2163.5	9981.7	112817.1
rMajorTrip	z-value	0.074	0.924	0.010	-0.116	0.157	-1.159	-0.037	-0.011
	Original value	37.887	50.921	36.902	34.976	39.152	18.982	36.174	36.573

Table 66 Final Cluster Centers in z-values and Original Values for Control Variables

Final Cluster Centers									
		Cluster							
		1	2	3	4	5	6	7	8
rFrequency	z-value	-0.629	-0.163	1.228	0.562	-0.852	-0.387	0.236	0.609
	Original value	0.093	0.149	0.316	0.236	0.066	0.122	0.197	0.242
Frequency last year	z-value	-0.495	-0.475	0.983	0.919	-0.761	-0.650	0.475	0.959
	Original value	27.813	28.614	87.694	85.078	17.031	21.533	67.101	86.730
IPT	z-value	1.225	0.062	-0.397	-0.274	5.505	0.010	-0.208	-0.309
	Original value	31.893	11.696	3.731	5.861	106.180	10.801	7.010	5.256
Amount	z-value	0.019	-0.054	11.883	0.011	-0.021	-0.135	-0.034	2.468
	Original value	146.86	126.64	3443.06	144.75	135.82	104.17	132.21	827.39
rAmount	z-value	-0.024	0.080	0.693	-0.065	-0.054	0.021	-0.053	0.064
	Original value	0.589	1.619	7.710	0.187	0.292	1.035	0.306	1.463
rTotal Amount	z-value	-0.210	-0.095	19.464	0.126	-0.249	-0.199	0.012	2.641
	Original value	10.98	17.3	1099.7	29.588	8.882	11.628	23.286	168.7

- <u>Characteristics related to categorical variables</u>

Eight clusters that are produced at the end of the partitioning process are tested with chi-square test with aim to define their difference from each other. When the test is applied to all dataset, since some clusters have few cases, restrictions of the chi square analysis is not followed. More than 20% of the cases had counts smaller than five. Literature argues that when the expected counts are smaller than five small groups can either combined or discarded from the dataset. As a result of this by discarding the clusters with few cases, cluster three and cluster five contingency test is applied again. The result of contingency test shows that clusters are different with respect to these categorical variables by having significance values smaller than 0.05 and all variables can be used to define the clusters. Table 67 shows the result of the contingency test.

Table 67 Result of Contingency Tests

Test with Six Clusters - 57605 Cases			
Variable	*Chi-Square Value*	*df*	*p-value*
Sales Directorate	14061.353	45	0.000
Customer Type	3288.927	15	0.000
Working Period	2988.105	5	0.000
Position Group	4440.743	55	0.000
SES Group	948.866	20	0.000
Region	1031.705	5	0.000
Position Group	1090.009	20	0.000
Customer Structure	97.361	5	0.000
Visit Frequency	1603.383	25	0.000
Customer Specialty	776.989	25	0.000
Working Type	2470.940	20	0.000

In order to control whether a specified cluster is different from the rest of dataset with respect to categorical variables, contingency tests are applied. Results of the contingency tests in Table 67 show that with respect to “Sales Directorate” and “Customer Type” variables all clusters are different from the rest of dataset. On the other hand with respect to other categorical variables some clusters are not different from the rest of dataset. Detailed analysis will be done regarding each variable when clusters are being interpreted one by one.

Table 68 Contingency Test Results of Comparison between Cluster and Rest of Data

Variable	1- All	2- All	3- All	4- All	5- All	6- All	7- All	8- All
Sales Directorate	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Customer type	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Working Period	0.294	0.000	0.049	0.000	0.049	0.000	0.000	0.000
Customer Group	0.000	0.000	0.160	0.000	0.000	0.000	0.000	0.000
SES Group	0.001	0.000	0.105	0.000	0.160	0.000	0.000	0.000
Region	0.153	0.000	0.014	0.000	0.000	0.000	0.000	0.000
Position Group	0.315	0.000	0.018	0.000	0.000	0.401	0.000	0.000
Customer Structure	0.139	0.456	0.330	0.431	0.847	0.147	0.052	0.000
Visit Frequency	0.326	0.000	0.241	0.000	0.000	0.000	0.000	0.000
Customer Specialty	0.000	0.000	0.156	0.000	0.058	0.000	0.000	0.000
Working Type	0.000	0.000	0.000	0.000	0.068	0.000	0.000	0.000

In order to decide whether a category of categorical variable is one of the main features of cluster or not, percentage distribution of specified category is compared with the proportion of this category among all dataset. Figures constructed for each cluster with the categorical variables show the percentage distribution of variables for the cluster and for the all dataset. For each categorical variable categories with greater percentage proportion compared to the general dataset are selected as main features of clusters.

- <u>Interpretation and Profiling Sequence for Customer Clusters</u>

Profiling sequence of clusters is determined via a customized ranking in this study. Rank of each cluster is determined by the rank of its cluster center's among all clusters with respect to variables used in interpretation. The rank of cluster center is represented with a figure between one and eight where one represents the cluster center with biggest value. Different from other variables only for "IPT" and "Recency" variables one represents the cluster center with smallest value. The sum of ranks assigned to the cluster center with respect to variables used in interpretation represents the interpretation sequence of the clusters. Clusters with smaller sum of ranks will be interpreted before the others. Bottom part of Table 69 shows the average ranking for clusters and their interpretation sequence.

Table 69 Final Customer Cluster Center Ranks

Final Cluster Center Ranks								
	Cluster							
	1	2	3	4	5	6	7	8
LoR_1	6	7	3	1	4	8	5	2
Frequency	5	7	1	3	6	8	4	2
Frequency last year	6	5	1	3	8	7	4	2
Recency	7	6	1	2	8	5	4	3
IPT	7	6	1	3	8	5	4	2
Total Amount	6	7	1	3	5	8	4	2
rMajorTrip	3	1	4	7	2	8	6	5
Amount	3	7	1	4	5	8	6	2
rFrequency	7	5	1	3	8	6	4	2
rAmount	5	2	1	8	7	4	6	3
rTotal Amount	7	5	1	3	8	6	4	2
Total Ranking Point for Clusters	62	58	16	40	69	73	51	27
Interpretation Sequence for Clusters	*6*	*5*	*1*	*3*	*7*	*8*	*4*	*2*

Cluster Three

- General Characteristics

Table 70 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 70 General Characteristics of Cluster Three

Subject	V*alue*	*Status Among all Clusters*
Number of Cases in the Cluster	36	8th highest
Total Euclidian Distance of the cases from the Cluster Center	266.1942909	8th highest
Average Euclidian Distance from the Cluster Center	7.394285858	1st highest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	24.48882711	1st highest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	21.25971943	1st highest

Table 70 shows that thirty six cases are grouped under cluster three at the end of the partitioning process. The value represents 0.06 % of the general dataset, which is relatively small compared to the other clusters.

Distance form the Cluster Center measures represent the total and average distance between the cases grouped in the cluster and the cluster center. Since the cluster contains only thirty six cases, total Euclidian distance for this cluster is the smallest among all clusters. However when the average Euclidian distance is considered, cluster three has the highest value among all clusters. Based on this information, it can be concluded that most cases in this cluster are not close to their center and the cluster is a wide one compared to other ones.

The last two columns of the table represent the distance between the cluster center and the center of all clusters with respect to different distance measures. Since this cluster has the highest values for these measures, i.e. it is the farthest cluster from the center of all clusters; it indicates that this cluster contains outliers. Between

Cluster Manhattan Distance is 6.45 and Between Cluster Euclidian Distance is 4.09. Both values are smaller than the ones for this cluster which support the idea that this cluster contains the cases that are outliers.

- Characteristics Related to Continuous Variables

Table 71 contains information needed to make interpretations related to continuous variables defining the cluster.

"LoR" shows that customers in this cluster are working with the company nearly for two years on average. When we compare this cluster with the others, it is concluded that cluster three has a relatively high "LoR" which make it to lie in the third rank among all. Despite its ranking, p-values for this variable shows that cluster three is not significantly different from cluster four and cluster eight with respect to "LoR". These clusters are in second and first rank among all clusters with respect to this variable. Based on this information it is obvious that customers in this cluster are not the oldest ones but still they can be interpreted old customers compared to the others.

Table 71 Cluster Three Cluster center Values and Significance Values between the Means of Clusters

Cluster 3 - *Stars*						*Significance Values between Clusters*						
						Cluster 3-1	*Cluster 3-2*	*Cluster 3-4*	*Cluster 3-5*	*Cluster 3-6*	Cluster 3-7	Cluster 3-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Rank between the clusters*	p value	p value	P value	p value	p value	p value	p value
LoR_1	701.0000	392.4930	821.5241	195.2760	3	0.000	0.000	0.133	0.002	0.000	0.000	0.263
Frequency	200.3611	66.4161	200.3611	23.6706	1	0.000	0.000	0.572	0.000	0.000	0.000	1.000
Frequency last one year	87.6944	47.8689	87.6944	17.0308	1	0.000	0.000	1.000	0.000	0.000	0.464	1.000
Recency	6.1944	14.9735	314.1062	6.1944	1	0.000	0.000	1.000	0.000	0.021	1.000	1.000
IPT	3.7310	10.6223	106.1797	3.7310	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	621207.7897	9982.6234	621207.78	2163.4934	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rMajorTrip	36.9024	36.7469	50.9214	18.9822	4	1.000	0.000	0.996	0.998	0.000	1.000	1.000
Amount	3443.0654	141.6421	3443.0654	104.1787	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rFrequency	0.3164	0.1686	0.3164	0.0661	1	0.000	0.000	0.072	0.000	0.000	0.001	0.133
rAmount	7.7095	0.8296	7.7095	0.1874	1	0.033	0.126	0.019	0.022	0.060	0.022	0.104
rTotal Amount	1099.7329	22.6327	1099.7329	8.8824	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Cluster three is in the first rank among all clusters with respect to "Frequency" variable. The "Frequency" value greater than 200 for this cluster shows that customers in this cluster purchased more than the ones in other clusters within the observation period. Supporting this, "Frequency Last Year" variable also ranks in the first among all clusters. Also related to "Frequency", "rFrequency" variable has a value greater than all of the other clusters. Based on the value of this variable it is clear that customers in this cluster bought frequently from the company compared to their long "LoR" within the observation period. Although it is in the first rank for three variables mentioned above, p-values computed by ANOVA points out that, cluster three is not significantly different from cluster four and cluster eight with respect to "Frequency", "Frequency Last Year" and "rFrequency" variables. To sum up, customers in this cluster have the highest values for the frequency variables but this issue by itself does not make the cluster different from the other ones.

Both "Total Amount" and "Amount" variables show that the customers in this cluster are the ones who purchased the biggest amount from the company on total and on average. Cluster three has a "Total Amount" value which is five times greater than the nearest cluster. The same holds for the "Amount" variable, that is, cluster three is more than four times greater than the nearest one. These findings indicate that customers in this cluster buy significantly greater amounts than the customers in other clusters. This information is also supported by the ANOVA results; cluster three has significantly greater values for these two variables compared to the others.

Cluster three ranks in the first among all clusters with respect to the variables related to the time passed between purchases of customers: "IPT" and "Recency". When the needed comparisons are done for the "IPT" variable, results point out that the cluster has the smallest “IPT” compared to the other clusters. This very small “IPT” shows that customers in this cluster bought products from the company nearly in every four days on average. With respect to “IPT” variable the difference between this cluster and the others is found to be statistically significant by the ANOVA. Moreover, the value of the “Recency” variable for this cluster shows that on average there are 6 days between the last two purchases of the customers. Briefly, it can be concluded that customers in cluster three buy products from the company very frequently and this makes the cluster significantly different from the other ones.

The cluster has “rMajorTrip” which is very close to the general average of the dataset. With the value of 36.9 %, “rMajorTrip” variable indicates that these customers bought products of company in a systematic manner.

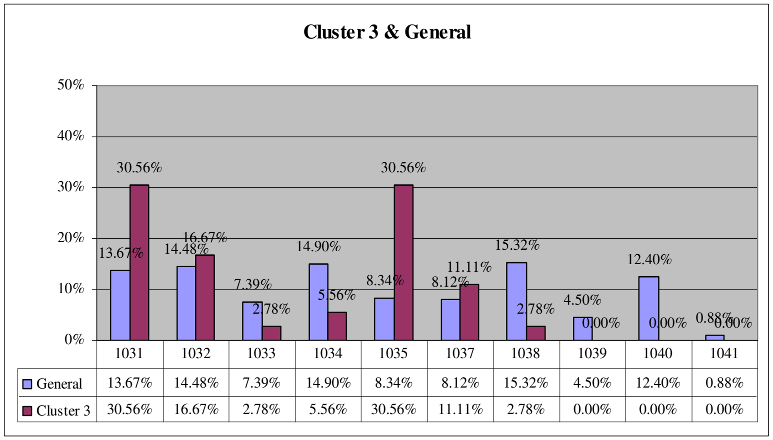
In short, based on the information gathered from the variables it can be concluded that, cluster three contains the most valuable customers of the company with greatest “Total Amount”, “rAmount”, “Amount” and “IPT” variables which are significantly different from other clusters. With its all characteristics the cluster can be named as Star Customers.

- Characteristics Related to Categorical Variables

Analyzing the figures from 39 to 49 characteristics of cluster three related to categorical variables are determined. Table 72 summarizes the main features of cluster three with respect to categorical variables.

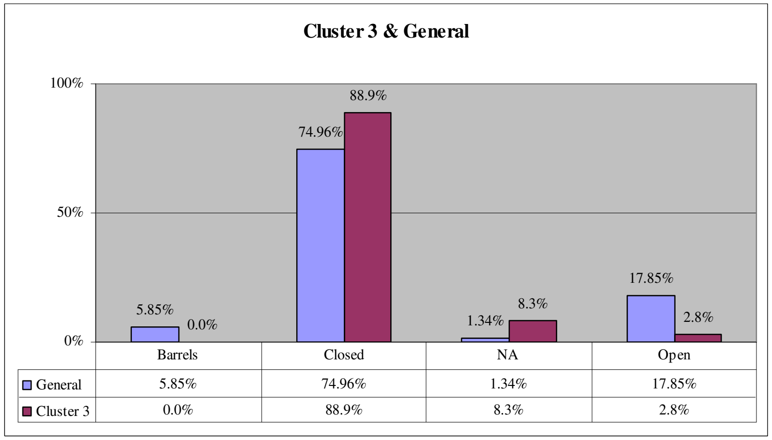
Table 72 Categorical Variables Analysis for Cluster Three

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1031, 1032, 1035 and 1037
Customer Type	Closed, NA
Working Period	Standard
Customer Group	Does not characterize cluster based on Contingency test results (Table 68)
SES Group	Does not characterize cluster based on Contingency test results (Table 68)
Region	Center
Position Group	Shopping Center, Mid Street and NA
Customer Structure	Company Brands
Visit Frequency	Does not characterize cluster based on Contingency test results (Table 68)
Customer Specialty	Company Brands, NA
Working Type	Cash, Cheque, NA



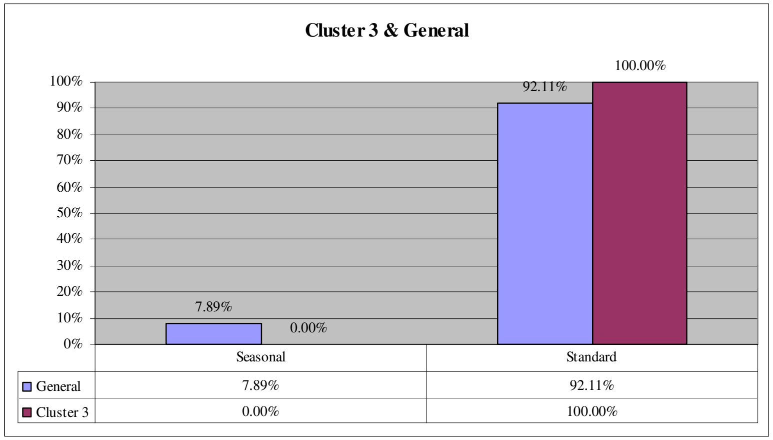
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 3	30.56%	16.67%	2.78%	5.56%	30.56%	11.11%	2.78%	0.00%	0.00%	0.00%

Figure 39 Sales directorate cluster three general comparison



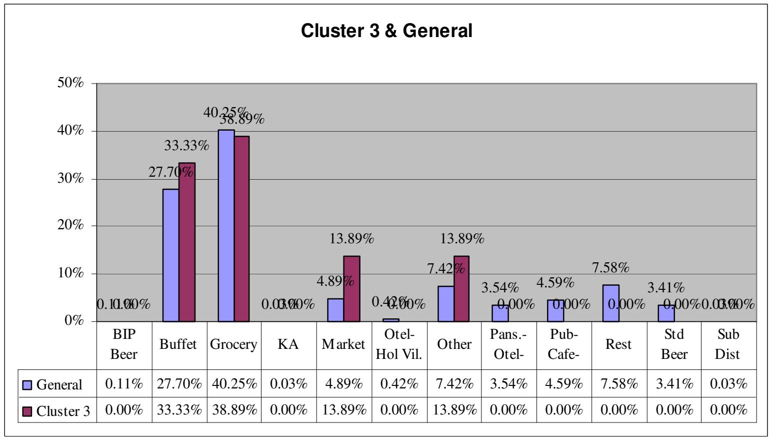
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 3	0.0%	88.9%	8.3%	2.8%

Figure 40 Customer type cluster three general comparison



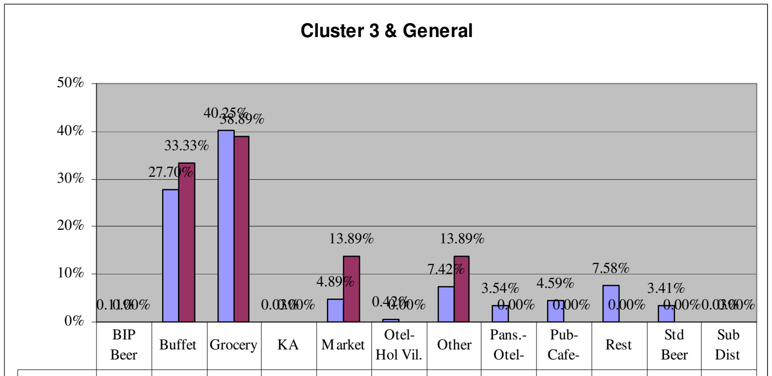
	Seasonal	Standard
General	7.89%	92.11%
Cluster 3	0.00%	100.00%

Figure 41 Working period cluster three general comparison



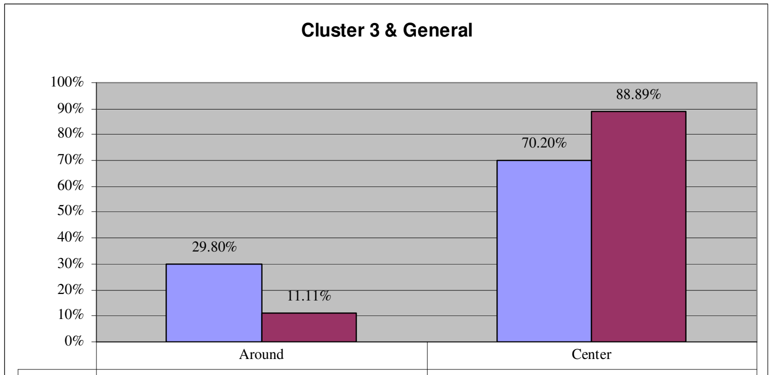
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 3	0.00%	33.33%	38.89%	0.00%	13.89%	0.00%	13.89%	0.00%	0.00%	0.00%	0.00%	0.00%

Figure 42 Customer group cluster three general comparison



	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 3	0.00%	33.33%	38.89%	0.00%	13.89%	0.00%	13.89%	0.00%	0.00%	0.00%	0.00%	0.00%

Figure 43 Region cluster three general comparison



	Around	Center
General	29.80%	70.20%
Cluster 3	11.11%	88.89%

Figure 44 SES group cluster three general comparison

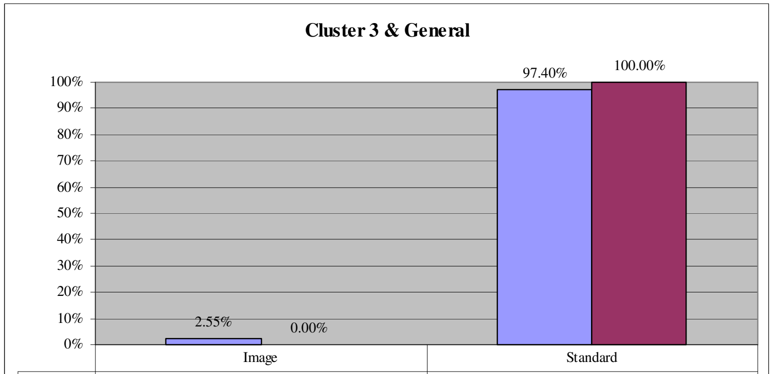
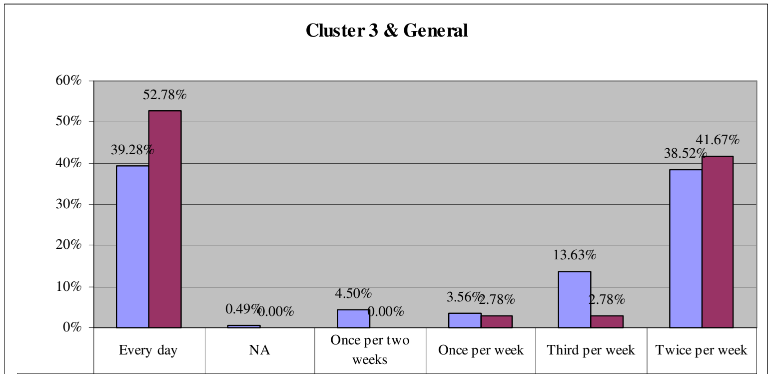


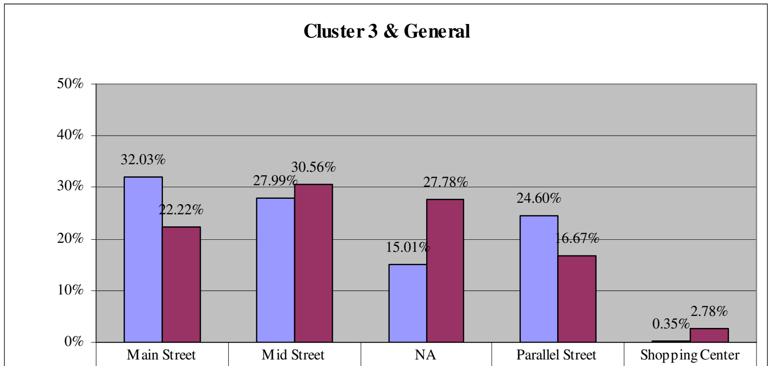
	Image	Standard
General	2.55%	97.40%
Cluster 3	0.00%	100.00%

Figure 45 Visit frequency cluster three general comparison



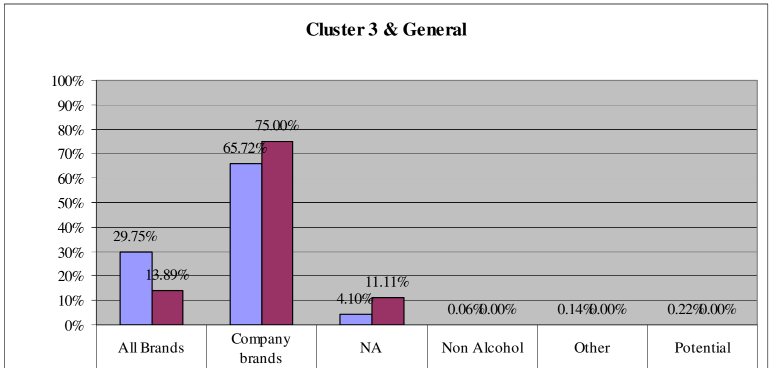
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 3	52.78%	0.00%	0.00%	2.78%	2.78%	41.67%

Figure 46 Customer structure cluster three general comparison



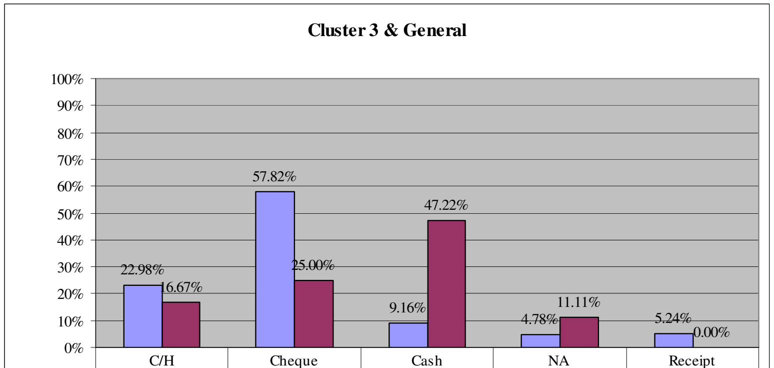
	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 3	22.22%	30.56%	27.78%	16.67%	2.78%

Figure 47 Customer specialty cluster three general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 3	13.89%	75.00%	11.11%	0.00%	0.00%	0.00%

Figure 48 Position group cluster three general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 3	16.67%	25.00%	47.22%	11.11%	0.00%

Figure 49 Working type cluster three general comparison

Cluster Eight

- General Characteristics

Table 73 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 73 General Characteristics of Cluster Eight

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	1019	6th biggest
Total Euclidian Distance of the cases from Cluster Center	2682.378269	6th biggest
Average Euclidian Distance from Cluster Center	2.632363366	3rd biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	3.77268996	8th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	2.311554324	8th biggest

There are 1019 customers in this cluster which accounts for 1.76% of all customers of the company. This is a relatively small percentage compared to other clusters.

Cluster eight has the third biggest average Euclidian distance value among all clusters which shows that the cluster is considerably wide compared to most of the other clusters.

Total Manhattan and Euclidian distance measures of this cluster have the smallest values among all clusters. These are approximately two times smaller than between cluster Manhattan and Euclidian distances. Based on this information it is obvious that this cluster is the closest cluster to the center of all clusters. In other words, there are no outliers in this cluster.

- Characteristics Related to Continuous Variables

Table 74 contains information needed to interpret the cluster with continuous variables representing it.

Customers in cluster eight have been working with the company for almost 2.2 years. It is the second longest length of relationship value among all clusters. On the other hand, ANOVA reveals that cluster eight is not significantly different from cluster three and cluster four with respect to “LoR”. However, despite the statistically insignificant differences, since cluster three has the longest value for this variable, customers in cluster eight can also be evaluated as long time customers.

The variable “Frequency” shows that customers in this cluster on average bought 190 times from the company within the observation period. This is the second greatest value among the cluster centers. However the difference between the cluster with the highest “Frequency” and this cluster which is only 5% is not statistically significant as it is shown in the right part of Table 11. This cluster can also be accepted as a cluster containing customers who buy frequently. Also the cluster has the second highest “Frequency Last Year”. However there is only 2% difference between the leading cluster and this cluster, which is again not a statistically significant. “rFrequency” of this cluster also shows that relative to their length of relationship customers in this cluster buy frequently from the company.

Table 74 Cluster Eight Cluster center Values and Significance Values between the Means of Clusters

Cluster *1 - Valuable Customers*						*Significance Values between Clusters*						
						Cluster 8-1	*Cluster 8-2*	*Cluster 8-3*	*Cluster 8-4*	*Cluster 8-5*	Cluster 8-6	Cluster 8-7
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Rank between the clusters*	p value	p value	p value	p value	p value	p value	P value
LoR_1	811.0834	392.4930	821.5241	195.2760	2	0.000	0.000	0.263	0.999	0.000	0.000	0.000
Frequency	190.1708	66.4161	200.3611	23.6706	2	0.000	0.000	1.000	1.000	0.000	0.000	0.000
Frequency last one year	86.7301	47.8689	87.6944	17.0308	2	0.000	0.000	1.000	1.000	0.000	0.000	0.000
Recency	6.5083	14.9735	314.1062	6.1944	3	0.000	0.000	1.000	1.000	0.000	0.000	0.967
IPT	5.2556	10.6223	106.1797	3.7310	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	112817.075	9982.623	621207.789	2163.493	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rMajorTrip	36.5726	36.7469	50.9214	18.9822	5	0.096	0.000	1.000	0.001	0.459	0.000	1.000
Amount	827.3993	141.6421	3443.0654	104.1787	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rFrequency	0.2419	0.1686	0.3164	0.0661	2	0.000	0.000	0.133	0.995	0.000	0.000	0.000
rAmount	1.4625	0.8296	7.7095	0.1874	3	0.000	1.000	0.104	0.000	0.000	0.027	0.000
rTotal Amount	168.7599	22.6327	1099.7329	8.8824	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000

When the “IPT” and “Recency” variables are analyzed, it is observed that the customers in this cluster have bought from the company nearly every 6 days and there is 6.5 days on average between their last two purchases. The cluster has the second smallest “IPT” and the third smallest “Recency” which shows that customers in this cluster have bought frequently and their last two purchases are closer to each other.

With respect to “Total Amount” and “Amount” variables, the cluster has the second greatest values among all clusters. However these figures are five times smaller than the largest ones so that there is a great difference between the first cluster and this cluster. On the other hand when this cluster is compared to the one following it, namely the third cluster, this cluster is nearly five times greater than the third one. Therefore customers in this cluster are buying higher amounts compared to other clusters except cluster three.

In addition, ANOVA results summarized in Table 74 supports that purchasing amount variables of this cluster are significantly different than cluster three which means that customers in this cluster do not purchase as much as the ones in cluster three. “rAmount” and “rTotal Amount” variables also have higher for this cluster. However, it is obvious that there is a great difference between the values for this cluster and cluster three because of the difference between the “Amount” and “Total Amount” of the two clusters.

“rMajorTrip” variable for the cluster is closer to the general average of the dataset. It shows that just like cluster three, the customers in this cluster bought products of company in a systematic manner.

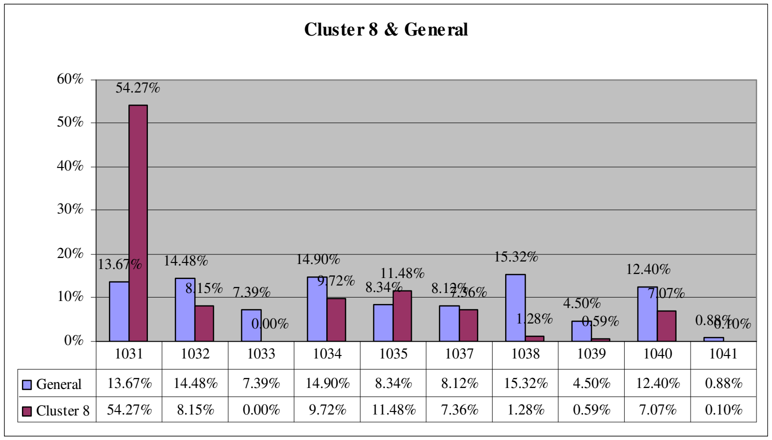
On the basis of information gained from analyzing variables separately, the cluster can be interpreted as a cluster containing valuable customers. Although the cluster has similar values with cluster three (Star Customers), the two clusters differ with respect to purchasing amount variables. In general the cluster contains customers who are valuable for the company with high “LoR”, “Frequency” values. But they are not buying as much as the customers in cluster three.

- Characteristics Related to Categorical Variables

Analyzing the figures from 50 to 60 characteristics of cluster eight related to categorical variables are determined. Table 75 summarizes the main features of cluster eight with respect to categorical variables.

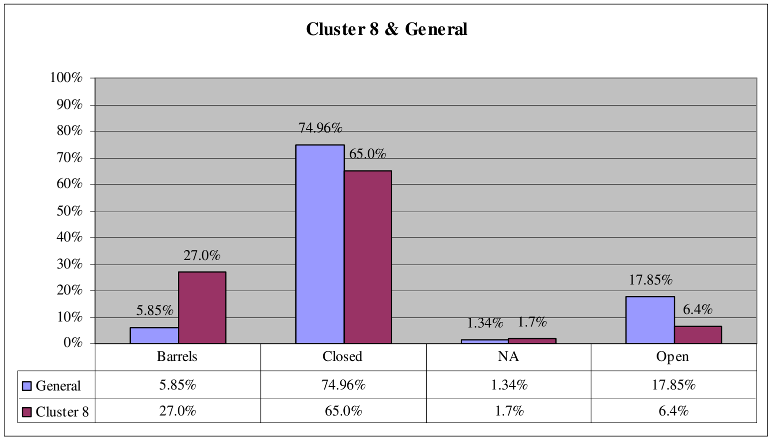
Table 75 Categorical Variables Analysis for Cluster Eight

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1031, 1035
Customer Type	Barrels, NA
Working Period	Standard
Customer Group	Otel, Holiday Village, Project Beer House, Buffet, Key account, Market, Pub Cafe Bar, Standard Beer House and Subordinate Distributor
SES Group	A, B-High Income, A+, A, B-High Level and D,E-Low Income
Region	Center
Position Group	Shopping Center, Main Street, “Parallel Street and NA
Customer Structure	Company Brands
Visit Frequency	Every day and “Once per week
Customer Specialty	Company Brands, Non Alcohol
Working Type	Cheque



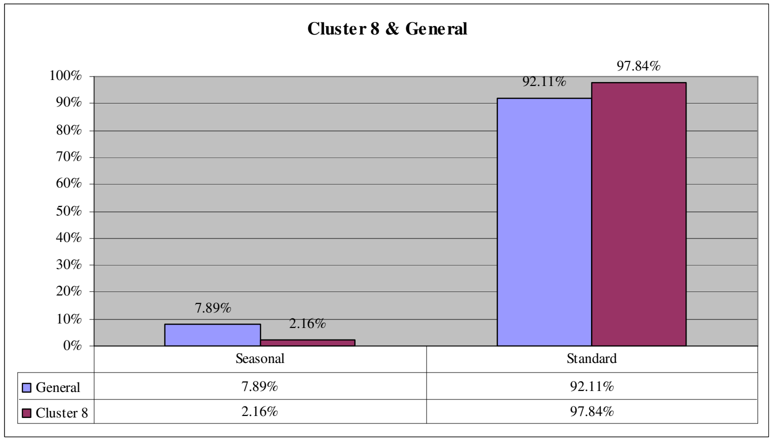
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 8	54.27%	8.15%	0.00%	9.72%	11.48%	7.36%	1.28%	0.59%	7.07%	0.10%

Figure 50 Sales directorate cluster eight general comparison



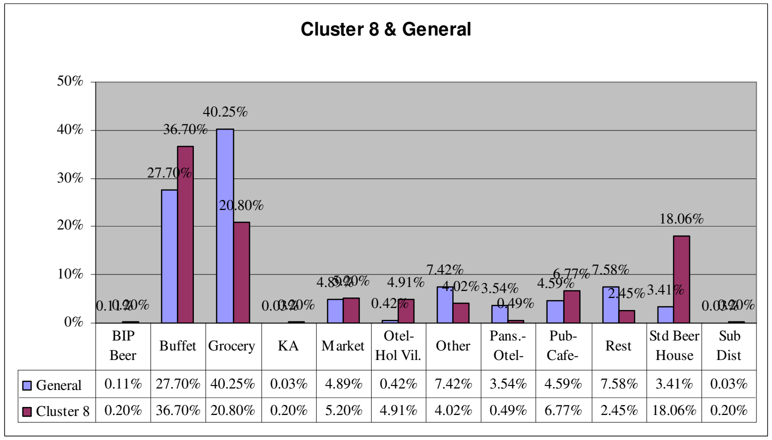
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 8	27.0%	65.0%	1.7%	6.4%

Figure 51 Customer type cluster eight general comparison



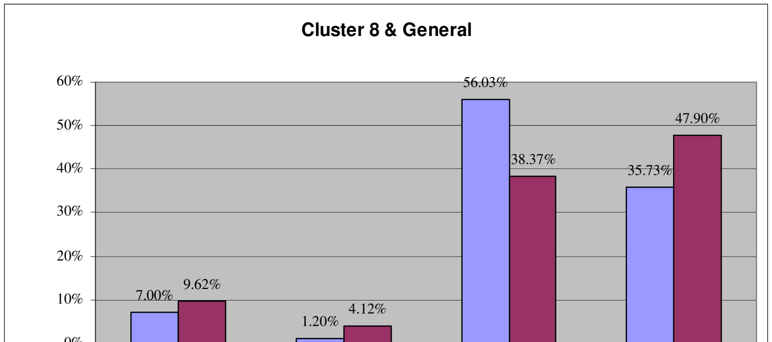
	Seasonal	Standard
General	7.89%	92.11%
Cluster 8	2.16%	97.84%

Figure 52 Working period cluster eight general comparison



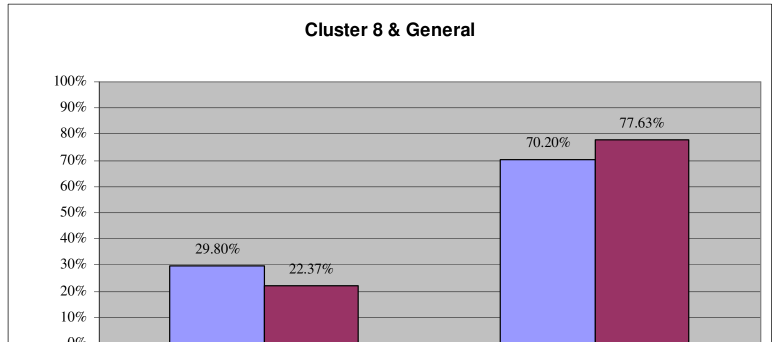
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer House	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 8	0.20%	36.70%	20.80%	0.20%	5.20%	4.91%	4.02%	0.49%	6.77%	2.45%	18.06%	0.20%

Figure 53 Customer group cluster eight general comparison



	A,B High Income	A+ A,B High Level	C Average Income	D,E low Income
General	7.00%	1.20%	56.03%	35.73%
Cluster 8	9.62%	4.12%	38.37%	47.90%

Figure 54 SES group cluster eight general comparison



	Around	Center
General	29.80%	70.20%
Cluster 8	22.37%	77.63%

Figure 55 Region cluster eight general comparison

	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 8	37.88%	13.94%	19.04%	27.97%	1.18%

Figure 56 Position group cluster eight general comparison

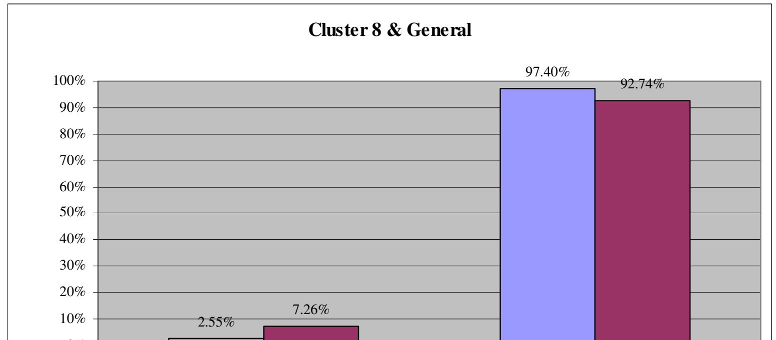
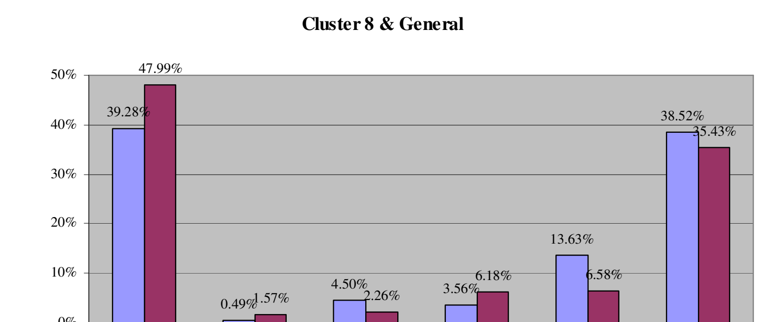


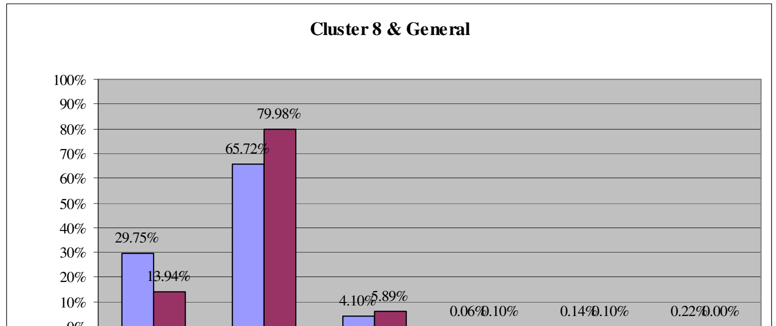
	Image	Standard
General	2.55%	97.40%
Cluster 8	7.26%	92.74%

Figure 57 Customer structure cluster eight general comparison



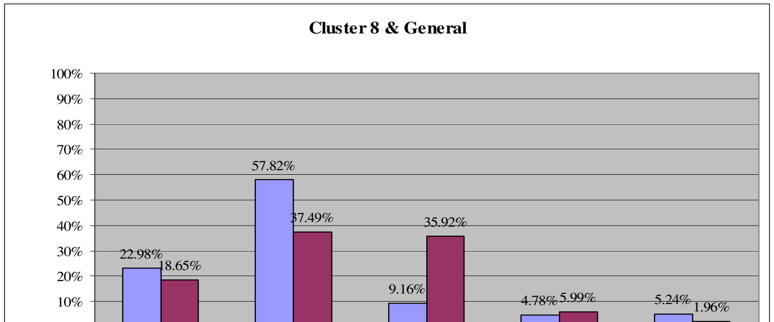
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 8	47.99%	1.57%	2.26%	6.18%	6.58%	35.43%

Figure 58 Visit frequency cluster eight general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 8	13.94%	79.98%	5.89%	0.10%	0.10%	0.00%

Figure 59 Customer specialty cluster eight general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 8	18.65%	37.49%	35.92%	5.99%	1.96%

Figure 60 Working type cluster eight general comparison

Cluster Four

- General Characteristics

Table 76 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 76 General Characteristics of Cluster Four

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	6464	4th biggest
Total Euclidian Distance of the cases from Cluster Center	10300.686	4th biggest
Average Euclidian Distance from Cluster Center	1.5935467	4th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	5.9581304	6th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	3.6482855	7th biggest

There are 6464 customers in this cluster which accounts for 11.16% of the general sample. The cluster has a moderate number of cases compared to other clusters.

Average Euclidian distance of the cases from cluster center represents the wideness of the cluster. Having the fourth biggest average distance, this cluster is wider than other four clusters but narrower than the three remaining ones.

Total Manhattan and Total Euclidian distances represent the distance between the center of this cluster and center of all clusters. Total distances computed for this cluster show that cases in this cluster are close to the center of all clusters because it has the sixth and seventh highest distances. Therefore, cases in this cluster can be accepted as close to the center and it is obvious that there is not a possibility of being an outlier for the members of this cluster.

- Characteristics Related to Continuous Variables

Table 77 contains information needed to make interpretations related to continuous variables representing the cluster center.

Table 13 shows that customers in this cluster have the greatest length of relationship compared to other clusters. By combining this information with the cluster size, it can be concluded that %11.16 of the customers constitute the group that has the longest relationship with the company.

This cluster is in the third rank with respect to "Frequency" and "Frequency last Year" variables. However, p-values computed by ANOVA show that this cluster is not significantly different from the first and second clusters with respect to these two variables. As a result, being in third rank does not mean that customers with long life time did not buy frequently from the company. Customers in this cluster also buy frequently from the company just like the ones in cluster three and cluster eight.

"IPT" variable shows that customers in this cluster buy products from the firm every 5.8 days on average. This is quite smaller than the mean of the dataset and lies in the third rank among all clusters. There is a significant difference between this cluster and cluster three and cluster eight in terms of the "IPT" variable. Based on this information it can be concluded that customers in this cluster buy products from the company within smaller intervals compared to the other parts of the dataset but not as frequently as the ones in cluster three and cluster eight. In addition, for the "Recency" variable there is not a significant difference between this cluster and cluster three and cluster eight, and the cluster has a smaller "Recency" compared to the overall the mean of the general dataset. By having a smaller “Recency” for the cluster center it, is obvious that there is a short time period between the last two purchases of the customers in this cluster just like the ones in cluster three and cluster eight.

This cluster is in the 7th position among all clusters with respect to “rMajorTrip”. This shows that customers in this cluster are buying from the company in consistent amounts. But sometimes (not so occasionally) they are buying in high amounts.

Table 77 Cluster Four Cluster Center Values and Significance Values between the Means of Clusters

Cluster 4 - Frequ*eny Buyers / Consistent*						*Significance Values between Clusters*						
						Cluster 4-1	*Cluster 4-2*	*Cluster 4-3*	*Cluster 4-5*	*Cluster 4-6*	Cluster 4-7	Cluster 4-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	p value	p value
LoR_1	821.5241	392.4930	821.5241	195.2760	1	0.000	0.000	0.133	0.000	0.000	0.000	0.999
Frequency	169.6914	66.4161	200.3611	23.6706	3	0.017	0.000	0.572	0.000	0.000	0.000	1.000
Frequency last one year	85.0781	47.8689	87.6944	17.0308	3	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Recency	6.2851	14.9735	314.1062	6.1944	2	0.000	0.000	1.000	0.000	0.000	0.000	1.000
IPT	5.8605	10.6223	106.1797	3.7310	3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	22270.8575	9982.6234	621207.7897	2163.4934	3	0.000	0.010	0.000	0.000	0.002	0.000	0.000
rMajorTrip	34.9758	36.7469	50.9214	18.9822	7	0.000	0.010	0.996	0.003	0.000	0.000	0.001
Amount	144.7526	141.6421	3443.0654	104.1787	4	1.000	0.000	0.000	1.000	0.000	0.000	0.000
rFrequency	0.2362	0.1686	0.3164	0.0661	3	0.000	0.000	0.072	0.000	0.000	0.000	0.995
rAmount	0.1874	0.8296	7.7095	0.1874	8	0.000	0.000	0.019	0.000	0.000	0.000	0.000
rTotal Amount	29.5876	22.6327	1099.7329	8.8824	3	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The cluster is in the third ranking with respect to the “Total Amount” variable. However there is a big difference between this cluster and the preceding one. The “Total Amount” of this cluster is five times smaller than of the preceding cluster. Although the “Total Amount” of this cluster is greater than the mean of the dataset, the cluster has the closest value for the “Amount” to the mean of dataset. Since the “Amount” is calculated by dividing the “Total Amount” by “LoR”, with a relatively small “Total Amount” and the greatest “LoR”, the cluster has a small “Amount” value. These findings show that the customers in this cluster buy frequently from the company but not in high amounts.

Different than all other variables, “rAmount” for this cluster is the smallest value among all clusters. This shows that customers in this cluster buys significantly small amounts compared to their length of relationship.

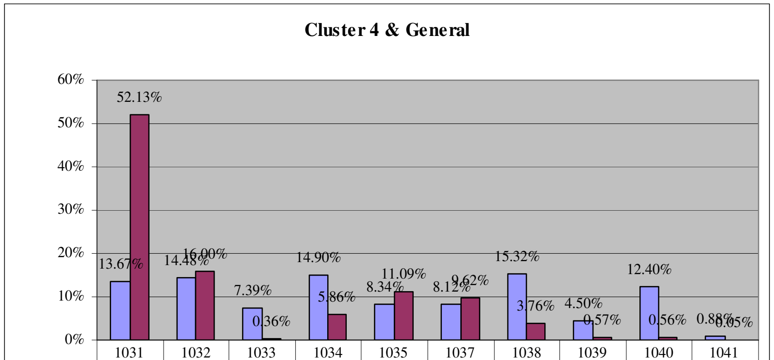
The points discussed above show that customers in this cluster have the longest relationships with company and they are buying in a frequent manner. However they purchase in smaller amounts. As a result, this cluster has smaller “Amount” and “rAmount”. In a word, it can be concluded that compared to their longer relationships, customers in this cluster did not buy in high amounts from the company. Therefore customers in this cluster are labeled as "Frequent Buyers".

- Characteristics Related to Categorical Variables

Analyzing the figures from 61 to 71 characteristics of cluster four related to categorical variables are determined. Table 78 summarizes the main features of cluster four with respect to categorical variables.

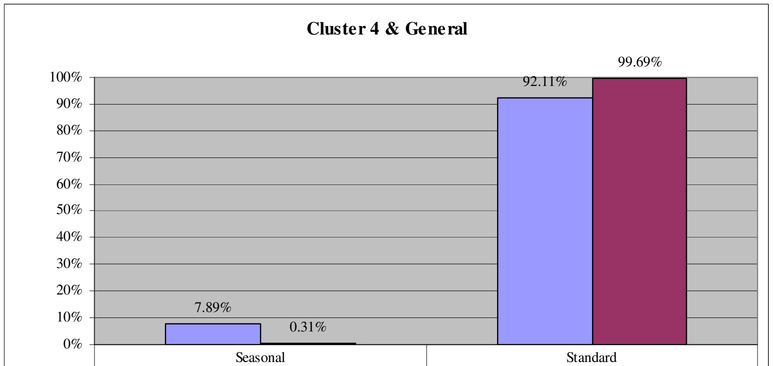
Table 78 Categorical Variables Analysis for Cluster Four

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1031, 1032, 1035 and 1037
Customer Type	Closed, Barrels
Working Period	Standard
Customer Group	Buffet, Standard Beer House
SES Group	A,B-High Income, A+, A,B-High Income, D,E-Low Income
Region	Center
Position Group	Shopping Center, NA
Customer Structure	Does not characterize cluster based on Contingency test results (Table 68)
Visit Frequency	Every day, Once per week
Customer Specialty	NA
Working Type	Cheque, Cash, NA



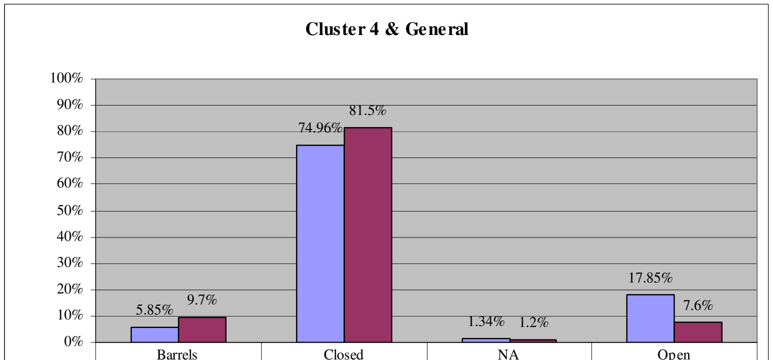
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 4	52.13%	16.00%	0.36%	5.86%	11.09%	9.62%	3.76%	0.57%	0.56%	0.05%

Figure 61 Customer type cluster four general comparison



	Seasonal	Standard
General	7.89%	92.11%
Cluster 4	0.31%	99.69%

Figure 63 Working period cluster four general comparison



	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 4	9.7%	81.5%	1.2%	7.6%

Figure 62 Sales directorate cluster four general comparison

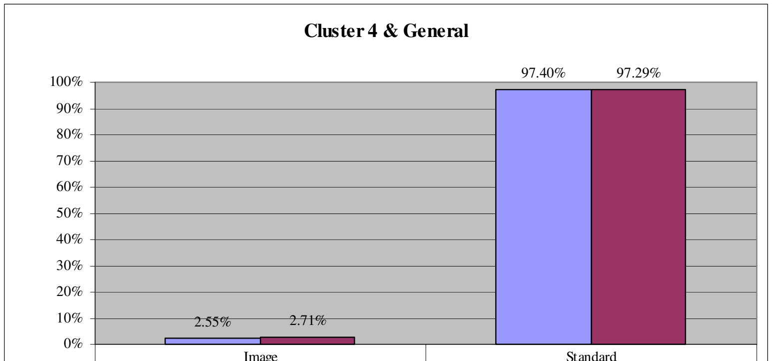
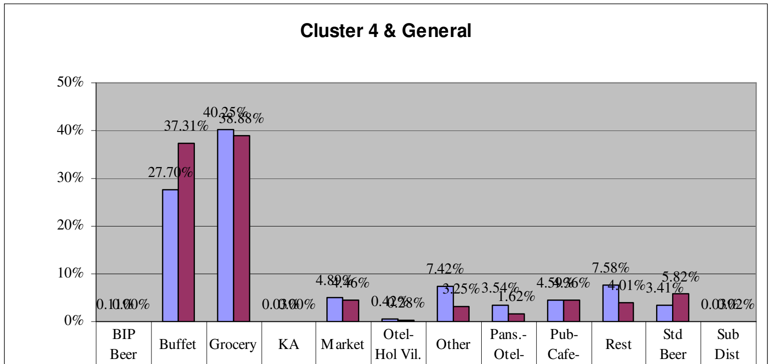


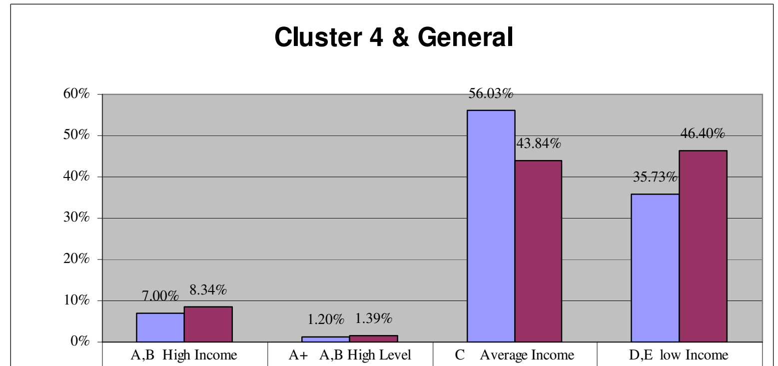
	Image	Standard
General	2.55%	97.40%
Cluster 4	2.71%	97.29%

Figure 64 Customer structure cluster four general comparison



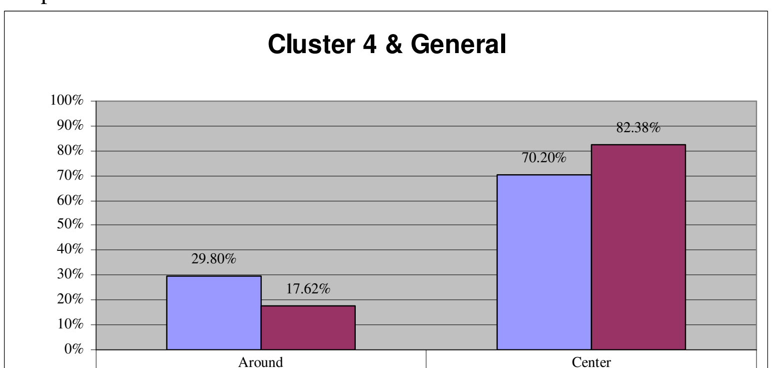
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 4	0.00%	37.31%	38.88%	0.00%	4.46%	0.28%	3.25%	1.62%	4.36%	4.01%	5.82%	0.02%

Figure 65 Customer group cluster four general comparison



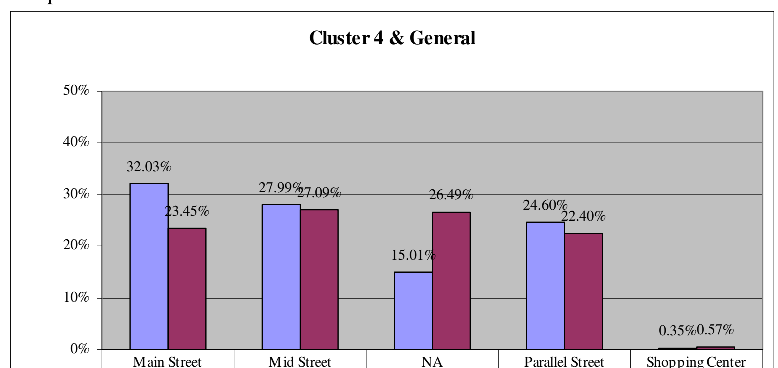
	A,B High Income	A+ A,B High Level	C Average Income	D,E low Income
General	7.00%	1.20%	56.03%	35.73%
Cluster 4	8.34%	1.39%	43.84%	46.40%

Figure 67 SES group cluster four general comparison



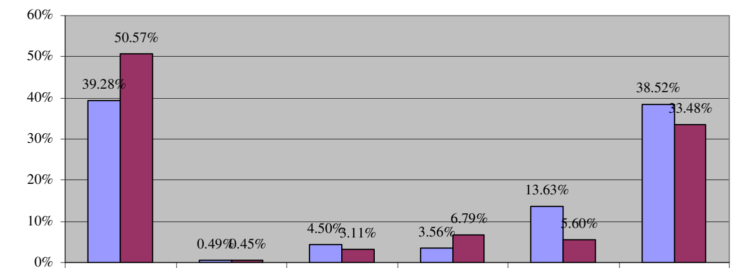
	Around	Center
General	29.80%	70.20%
Cluster 4	17.62%	82.38%

Figure 66 Region cluster four general comparison



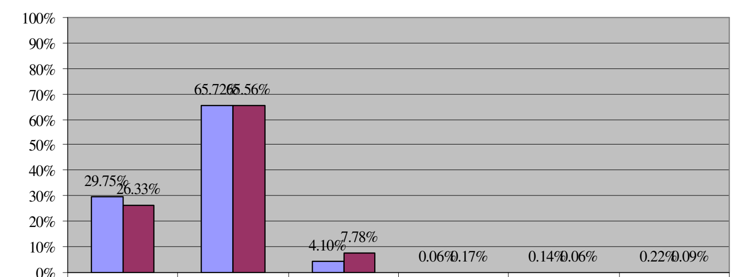
	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 4	23.45%	27.09%	26.49%	22.40%	0.57%

Figure 68 Position Group cluster four general comparison



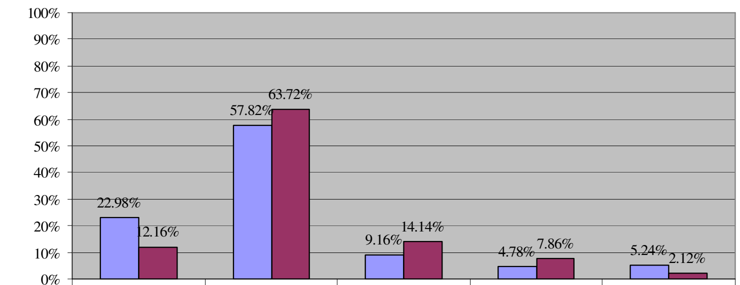
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 4	50.57%	0.45%	3.11%	6.79%	5.60%	33.48%

Figure 69 Visit frequency cluster four general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 4	26.33%	65.56%	7.78%	0.17%	0.06%	0.09%

Figure 70 Customer Specialty cluster four general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 4	12.16%	63.72%	14.14%	7.86%	2.12%

Figure 71 Working Type cluster four general comparison

Cluster Seven

- General Characteristics

Table 79 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 79 General Characteristics of Cluster Seven

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	20152	1st biggest
Total Euclidian Distance of the cases from Cluster Center	20020.444	1st biggest
Average Euclidian Distance from Cluster Center	0.9934718	7th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	5.3336817	7th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	3.8773688	6th biggest

There are 20152 customers in this cluster which accounts for 34.79% of the general dataset. This is the largest percentage compared to other clusters and naturally represents the biggest group of all clusters. In other words, most of the customers of the company are grouped under this cluster at the end of the partitioning process.

Average Euclidian distance measure of this cluster is the seventh biggest one among all clusters. The distance measure of this cluster is almost the same as the smallest one. As a result, it is concluded that cluster seven is one of the narrowest clusters in this analysis.

Total Manhattan and Total Euclidian distances represent the distance between the center of this cluster and center of all clusters. Values for this cluster show that cases in this cluster are close to the center of all clusters because the cluster has the seventh and sixth biggest distances. Distance measures of this cluster are approximately the same as of the preceding cluster: cluster four. Consequently, cases in this cluster can be accepted as close as the ones in cluster four and it is obvious that there is not a possibility of being an outlier for the customers in this cluster.

- Characteristics Related to Continuous Variables

Table 80 demonstrates the information needed to interpret the continuous variables.

Customers in this cluster have a relationship with the company for more than one year. This is greater than the mean of the dataset but in the fifth position among all clusters.

The cluster has the closest “Frequency” and “Frequency Last Year” values to the centers of all dataset and lies in the fourth ranking among all clusters. There is a significantly large difference between this cluster and the ones for the preceding cluster in terms of these two variables. This cluster has nearly two times smaller values than the preceding one. ANOVA shows that cluster seven is significantly different from all other clusters with respect to “Frequency” and “Frequency Last Year” variables.

Table 80 Cluster Seven Cluster Center Values and Significance Values between the Means of Clusters

Cluster *7 - Average Customers*						*Significance Values between Clusters*						
						Cluster 7-1	*Cluster 7-2*	*Cluster 7-3*	*Cluster 7-4*	*Cluster 7-5*	Cluster 7-6	Cluster 7-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	p value	p value
LoR_1	459.3312	392.4930	821.5241	195.2760	5	0.000	0.000	0.000	0.000	0.001	0.000	0.000
Frequency	83.4177	66.4161	200.3611	23.6706	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Frequency last one year	67.1012	47.8689	87.6944	17.0308	4	0.000	0.000	0.464	0.000	0.000	0.000	0.000
Recency	7.0794	14.9735	314.1062	6.1944	4	0.000	0.000	1.000	0.000	0.000	0.000	0.967
IPT	7.0103	10.6223	106.1797	3.7310	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	9981.6748	9982.6234	621207.7897	2163.4934	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rMajorTrip	36.1739	36.7469	50.9214	18.9822	6	0.000	0.000	1.000	0.000	0.139	0.000	1.000
Amount	132.2166	141.6421	3443.0654	104.1787	6	0.994	0.181	0.000	0.000	1.000	0.000	0.000
rFrequency	0.1970	0.1686	0.3164	0.0661	4	0.000	0.000	0.001	0.000	0.000	0.000	0.000
rAmount	0.3059	0.8296	7.7095	0.1874	6	0.000	0.000	0.022	0.000	1.000	0.000	0.000
rTotal Amount	23.2860	22.6327	1099.7329	8.8824	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Cluster seven is in the fourth rank when “IPT” and “Recency” variables are considered. Customers in this cluster buy products each week from the company. The “Recency” of the cluster also supports the information gathered from the “IPT” variable, i.e. there are seven days between the last two purchases of the customers.

An important observation for this cluster is it has the closest values for the “Total Amount”, “rTotal Amount” and “Amount” variables compared to the center of all dataset.

The cluster is in the 6^{th} order for the “rMajorTrip” variable. This shows that customers in this cluster are buying from the company in consistent amounts. But sometimes (not so occasionally) they are buying for bigger amounts.

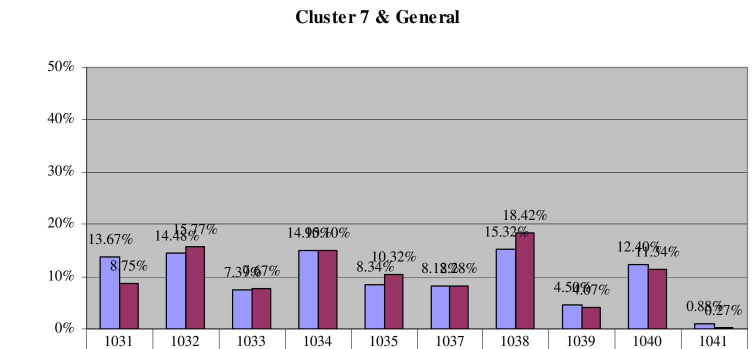
Briefly, customers in this cluster have values closest to the mean of the general dataset for most of the variables. Based on this information customers in this cluster are labeled as "Average Customers".

- Characteristics Related to Categorical Variables

Analyzing the figures from 72 to 82 characteristics of cluster seven related to categorical variables are determined. Table 81 summarizes the main features of cluster seven with respect to categorical variables.

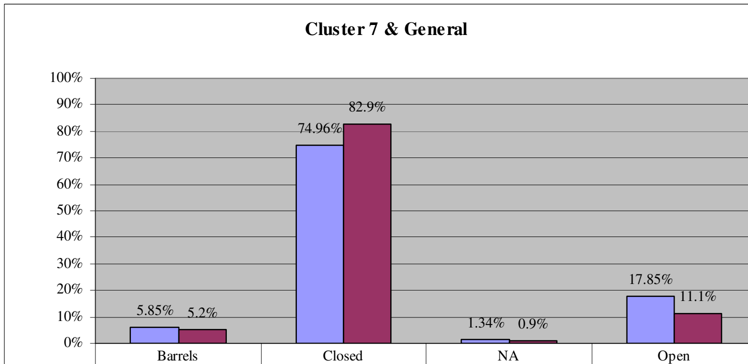
Table 81 Categorical Variables Analysis for Cluster Seven

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1032, 1033, 1034, 1035, 1037 and 1038
Customer Type	Closed
Working Period	Standard
Customer Group	Grocery and Buffet
SES Group	D,E-Low.Income
Region	Center
Position Group	Main Street
Customer Structure	Does not characterize cluster based on Contingency test results (Table 68)
Visit Frequency	One per two weeks, twice per week or three per week.
Customer Specialty	All brands
Working Type	CH, Cheque and Receipt.



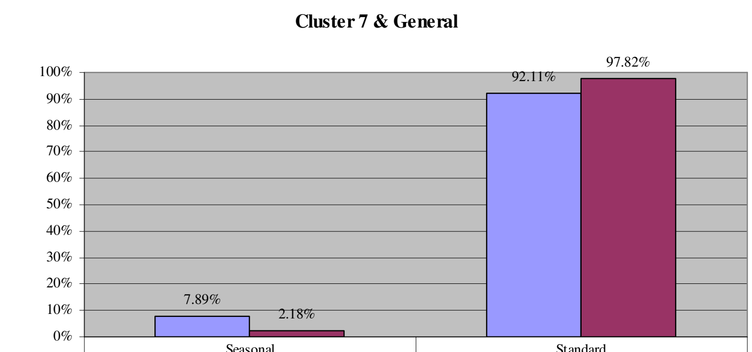
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 7	8.75%	15.77%	7.67%	15.10%	10.32%	8.28%	18.42%	4.07%	11.34%	0.27%

Figure 72 Sales directorate cluster seven general comparison



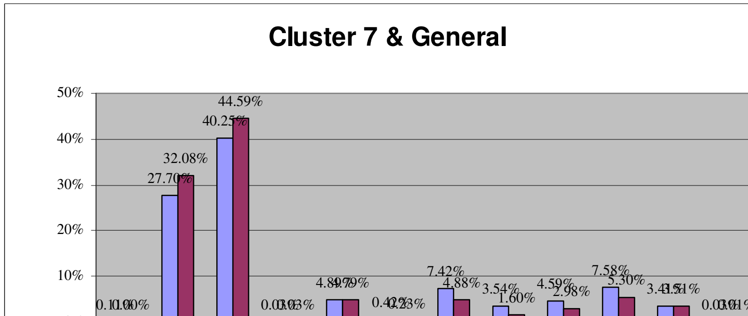
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 7	5.2%	82.9%	0.9%	11.1%

Figure 73 Customer type cluster seven general comparison



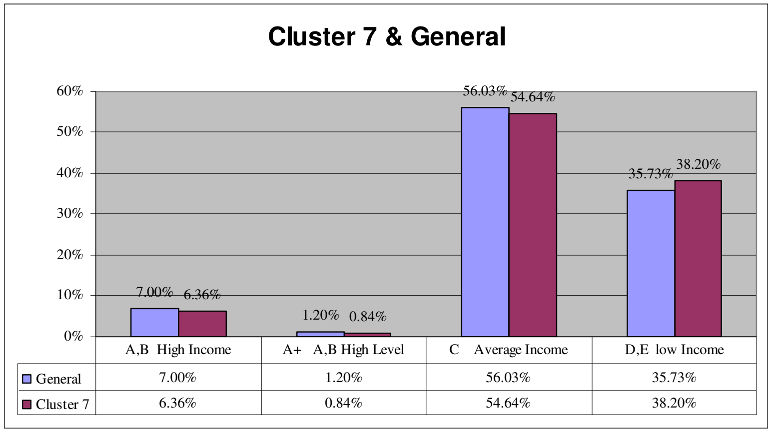
	Seasonal	Standard
General	7.89%	92.11%
Cluster 7	2.18%	97.82%

Figure 74 Working period cluster seven general comparison



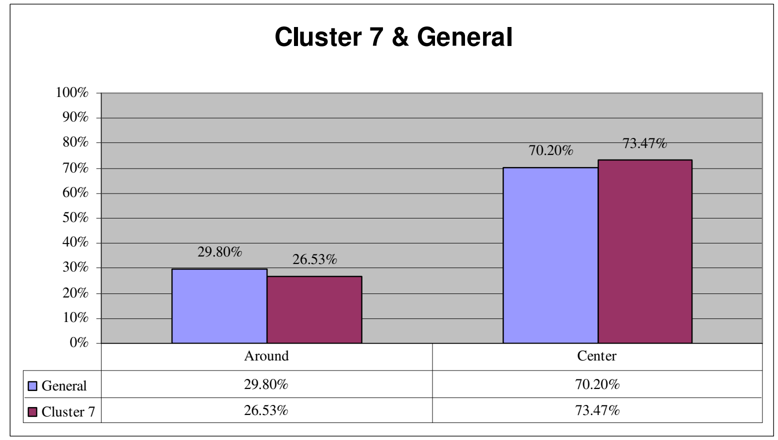
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer House	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 7	0.00%	32.08%	44.59%	0.03%	4.79%	0.23%	4.88%	1.60%	2.98%	5.30%	3.51%	0.01%

Figure 75 Customer group cluster seven general comparison



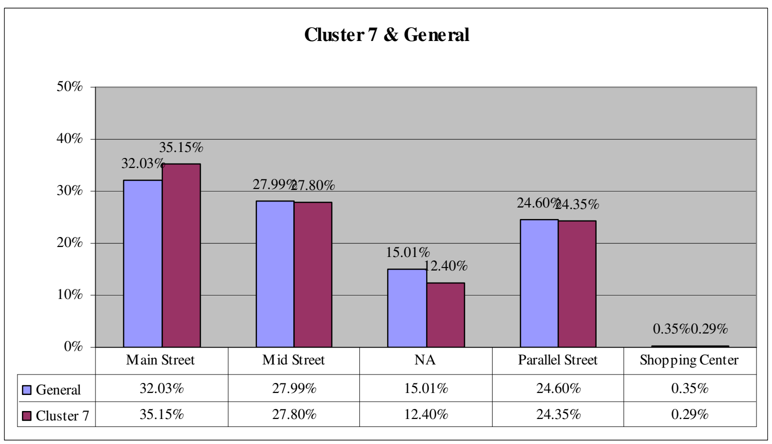
	A,B High Income	A+ A,B High Level	C Average Income	D,E low Income
General	7.00%	1.20%	56.03%	35.73%
Cluster 7	6.36%	0.84%	54.64%	38.20%

Figure 76 SES group cluster seven general comparison



	Around	Center
General	29.80%	70.20%
Cluster 7	26.53%	73.47%

Figure 77 Region cluster seven general comparison



	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 7	35.15%	27.80%	12.40%	24.35%	0.29%

Figure 78 Position group cluster seven general comparison

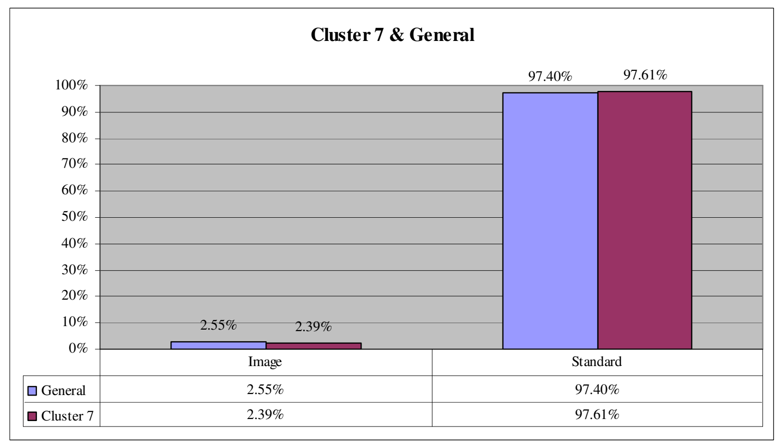


	Image	Standard
General	2.55%	97.40%
Cluster 7	2.39%	97.61%

Figure 79 Customer structure cluster seven general comparison

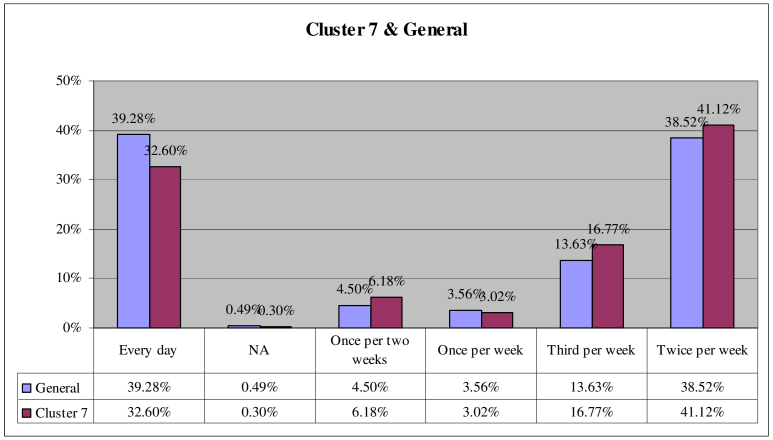


Figure 80 Visit frequency cluster seven general comparison

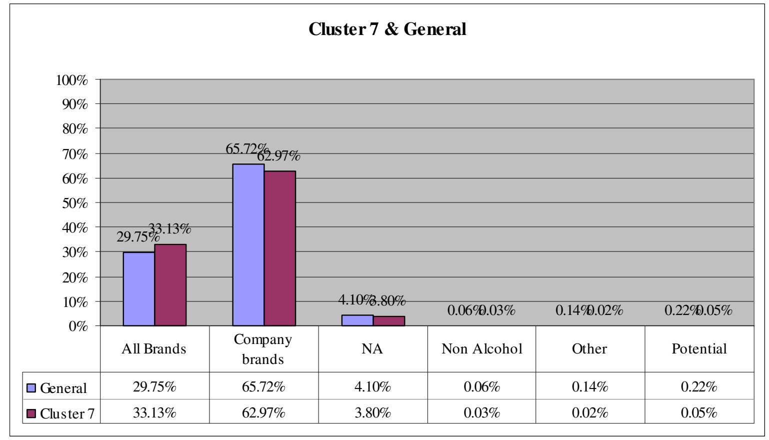


Figure 81 Customer structure cluster seven general comparison

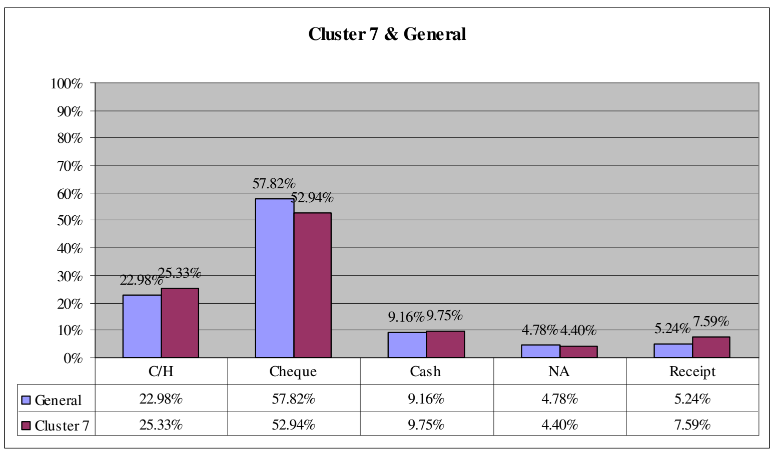


Figure 82 Working type cluster seven general comparison

Cluster Two

- General Characteristics

Table 82 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 82 General Characteristics of Cluster Two

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	15632	2nd biggest
Total Euclidian Distance of the cases from Cluster Center	15443.096	2nd biggest
Average Euclidian Distance from Cluster Center	0.9879155	8th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	7.2094903	4th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	4.2876392	4th biggest

There are 15632 customers in this cluster which accounts for 26.98% of the all dataset. This cluster is the second biggest cluster.

Cluster two has the smallest Average Euclidian distance which is the narrowest value compared to all other clusters.

This cluster is in the fourth position in terms of the measures indicating the distance between the cluster center and center of all clusters. Total Manhattan and Total Euclidian distances are approximately two times greater of the between clusters Manhattan and Euclidian distances. Consequently, cases in this cluster can be evaluated as ones that are not so far away from the center of all cluster centers and they cannot be evaluated as outliers.

- Characteristics Related to Continuous Variables

Table 83 contains information needed to make interpretations related to continuous variables.

Customers in this cluster have been working with the company for less than one year. Length of relationship for this cluster is 1.5 times smaller than the mean of the general dataset but 1.2 times greater than the cluster with the smallest length. In addition, the results of ANOVA certify that this cluster has significantly higher “LoR” compared to the cluster with the smallest “LoR” (Cluster Six). As a result, it is obvious that customers in this cluster are not the ones with shortest “LoRs” but compared to other clusters they have shorter relationship with the company.

“Frequency” and “Frequency Last Year” figures for this cluster center are smaller than the general mean of the data and closer to the minimum of the general dataset. Based on these facts, it can be concluded that customers in this cluster do not buy from the company frequently. However since the length of relationship for these customers are shorter than the preceding clusters, before concluding that these customers are not frequently buying from the company, the value of the “rFrequency” value should be analyzed. Value for this variable is closer to the one for Cluster 7 which contains “Average Customers”. This may mean that these customers have the potential to be the average customers in future.

Table 83 Cluster Two Cluster center Values and Significance Values between the Means of Clusters

Cluster 2 - *Potential Valuable Customers*						*Significance Values between Clusters*						
						Cluster 2-1	*Cluster 2-3*	*Cluster 2-4*	*Cluster 2-5*	*Cluster 2-6*	Cluster 2-7	Cluster 2-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	P value	p value
LoR_1	249.039	392.4930	821.5241	195.2760	7	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Frequency	31.264	66.4161	200.3611	23.6706	7	0.001	0.000	0.000	1.000	0.000	0.000	0.000
Frequency last one year	28.614	47.8689	87.6944	17.0308	5	0.999	0.000	0.000	0.006	0.000	0.000	0.000
Recency	12.335	14.9735	314.1062	6.1944	6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IPT	11.696	10.6223	106.1797	3.7310	6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	3696.31	9982.623	621207.78	2163.49	7	0.174	0.000	0.000	0.975	0.000	0.000	0.000
rMajorTrip	50.921	36.7469	50.9214	18.9822	1	0.000	0.010	0.000	0.000	0.000	0.000	0.000
Amount	126.642	141.642	3443.065	104.178	7	0.816	0.000	0.000	1.000	0.000	0.181	0.000
rFrequency	0.1489	0.1686	0.3164	0.0661	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rAmount	1.6187	0.8296	7.7095	0.1874	2	0.000	0.126	0.000	0.000	0.012	0.000	1.000
rTotal Amount	17.379	22.6327	1099.7329	8.8824	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The cluster center is in the 6th rank when “IPT” and “Recency” variables are considered. “IPT” of the cluster shows that customers in this cluster on average buy products from the company in every eleven days which higher than the mean of the general dataset. Also the “Recency” variable with a value of 12 supports the results derived from the “IPT” variable. It can be concluded that customers in this cluster buy the products not in very short intervals. However, since the “Amount” of the cluster centroid is not so small relative to its “LoR”, it is interpreted that these customers are not buying so frequently but buying in large amounts for each of their purchases. This idea is also supported by the “rMajorTrip” of the cluster reaches its maximum for this cluster.

“Total Amount” and “Amount” are relatively small compared to other clusters. “Total Amount” for this cluster is 1.7 times and “Amount” of this cluster centroid is 1.2 times greater than the smallest one. Thus, it can be concluded that these are non-valuable customers for the company. However the value of “rAmount” may change this interpretation. “rAmount” of the cluster is in the second position among all clusters. The first cluster was the one that is labeled as stars. Also ANOVA reveals that there is not a significant difference between this cluster and cluster three (stars) and cluster eight (valuable customers). The value for this cluster is almost the same as the one of cluster eight.

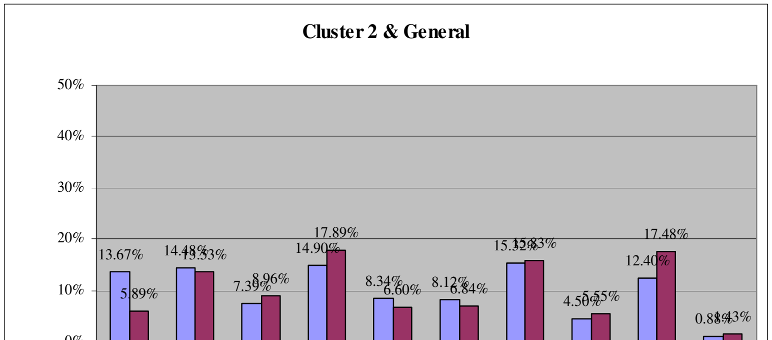
Although the variables for this cluster are not very high, since the “LoR” of the customers in this cluster are relatively smaller than the other clusters and “rAmount” variable is comparable to the valuable clusters, it is named as Potential Valuable Customers. Similarities with cluster eight also support the idea to name this cluster as Potential Valuable Customers.

- Characteristics Related to Categorical Variables

Analyzing the figures from 83 to 93 characteristics of cluster two related to categorical variables are determined. Table 84 summarizes the main features of cluster two with respect to categorical variables.

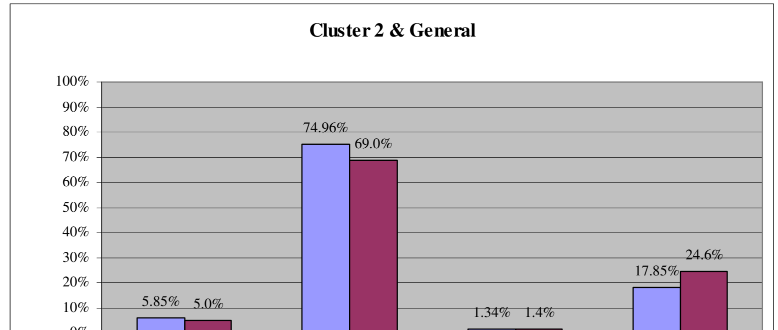
Table 84 Categorical Variables Analysis for Cluster Two

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1040, 1038, 1034 and 1032
Customer Type	Open
Working Period	Seasonal
Customer Group	Otel Holiday Village, Restaurant, Pension Otel Motel, Pub cafe bar, Subordinate distributor
SES Group	A,B-High Income, A+, A,B-High Income, C-Average Income
Region	Around
Position Group	Mid Street and Parallel Street
Customer Structure	Does not characterize cluster based on Contingency test results (Table 68)
Visit Frequency	Every day, Once per week
Customer Specialty	Company brands and Other
Working Type	CH or Cash



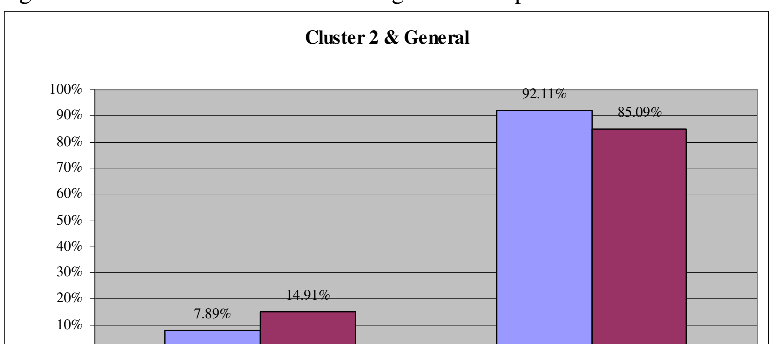
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 2	5.89%	13.53%	8.96%	17.89%	6.60%	6.84%	15.83%	5.55%	17.48%	1.43%

Figure 83 Sales directorate cluster two general comparison



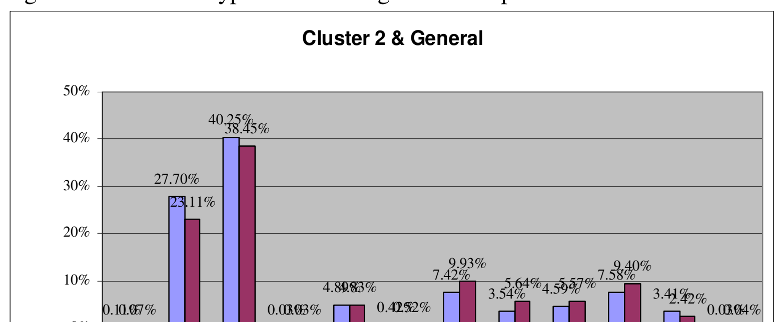
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 2	5.0%	69.0%	1.4%	24.6%

Figure 84 Customer type cluster two general comparison



	Seasonal	Standard
General	7.89%	92.11%
Cluster 2	14.91%	85.09%

Figure 85 Working period cluster two general comparison



	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 2	0.07%	23.11%	38.45%	0.03%	4.83%	0.52%	9.93%	5.64%	5.57%	9.40%	2.42%	0.04%

Figure 86 Customer group cluster two general comparison

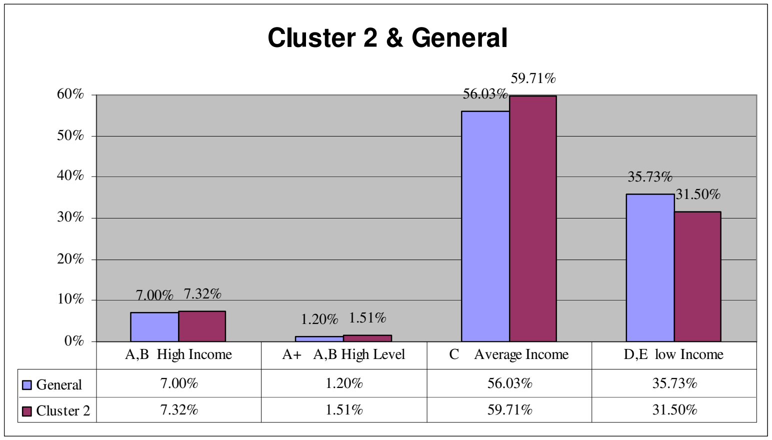


Figure 87 SES Group cluster two general comparison

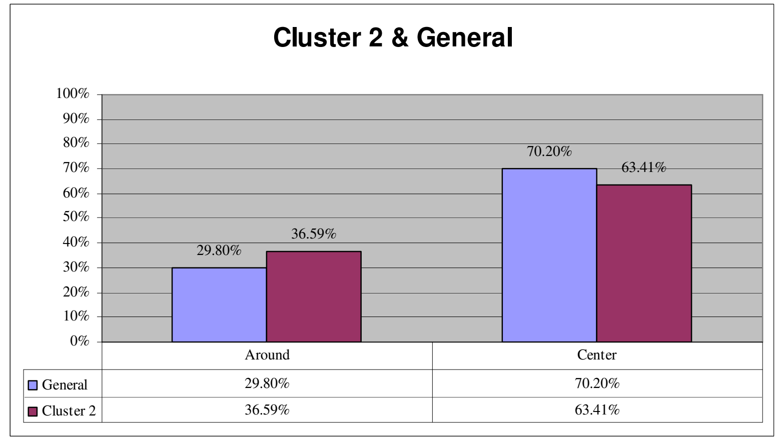


Figure 88 Region cluster two general comparison

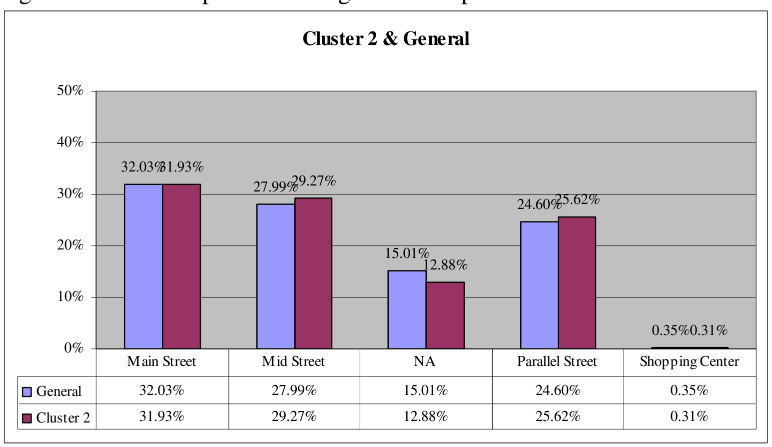


Figure 89 Position group cluster two general comparison

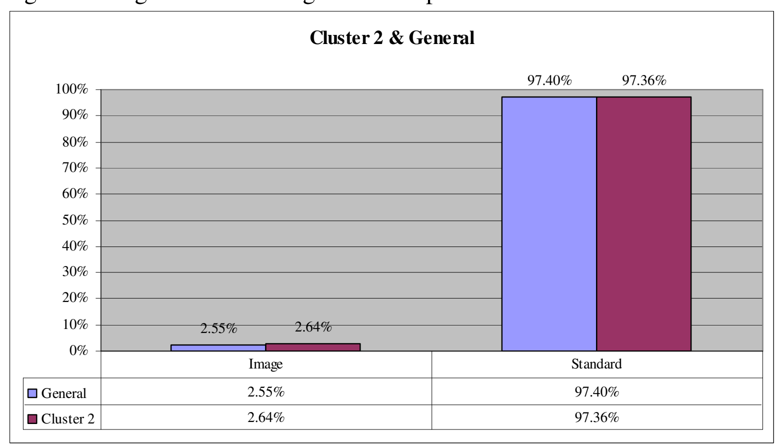
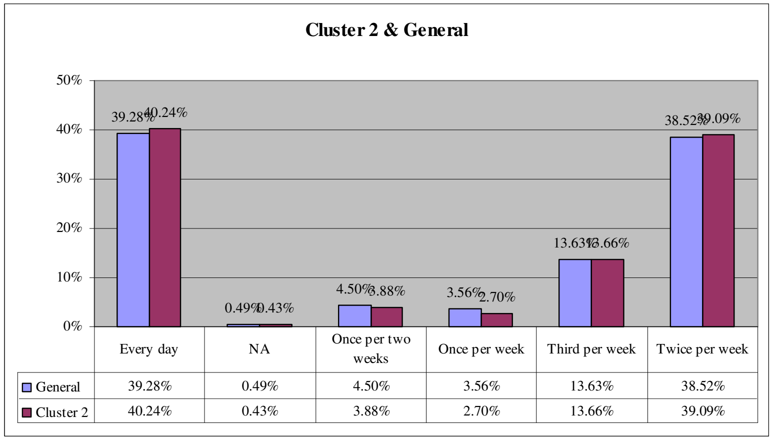
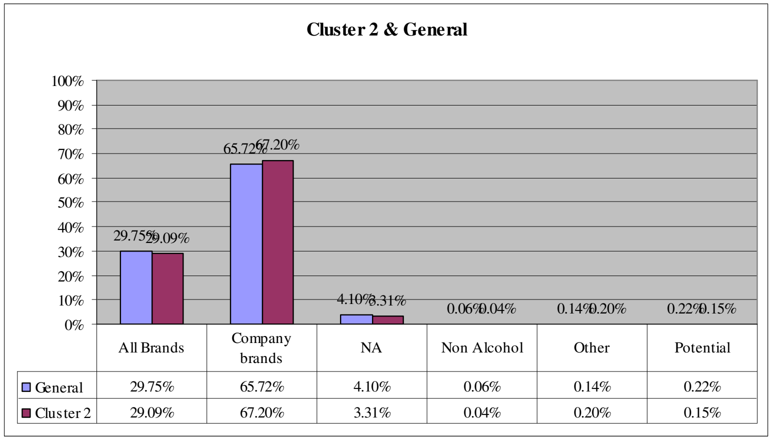


Figure 90 Customer structure cluster two general comparison



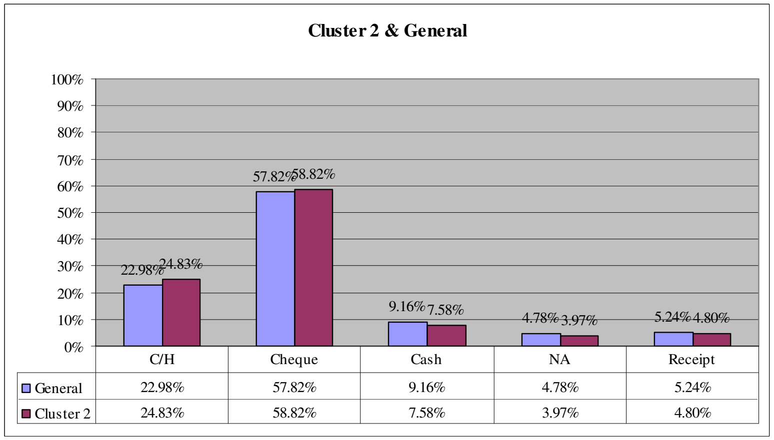
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 2	40.24%	0.43%	3.88%	2.70%	13.66%	39.09%

Figure 91 Visit frequency cluster two general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 2	29.09%	67.20%	3.31%	0.04%	0.20%	0.15%

Figure 92 Customer specialty cluster two general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 2	24.83%	58.82%	7.58%	3.97%	4.80%

Figure 93 Working type cluster two general comparison

Cluster One

- General Characteristics

Table 85 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 85 General Characteristics of Cluster One

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	2941	5th biggest
Total Euclidian Distance of the cases from Cluster Center	4677.1238	5th biggest
Average Euclidian Distance from Cluster Center	1.5903175	5th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	6.6158418	5th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	4.1124366	5th biggest

There are 2941 customers in this cluster that accounts for 5.08% of the general dataset. The size of the cluster may be interpreted as an average one compared to the other clusters in the dataset.

Average Euclidian distance of this cluster is almost the same as cluster four which makes the cluster have an average wideness level among all clusters.

Total Manhattan and Total Euclidian distances of the cluster are in the fifth rank among all clusters. These values are approximately the same as the between cluster Manhattan and Euclidian distances. Based on this information, it is concluded that cases in this cluster are not far away from the center of all clusters and are not outliers.

- Characteristics Related to Continuous Variables

Table 86 contains information needed to make interpretations related to continuous variables.

Customers in this cluster have been working with the company for almost one year which is the closest value to the average of the dataset.

“Frequency” and “Frequency Last Year” variables show that customers in this cluster did not purchase frequently from the company. Although this cluster is closer to Cluster two named as Potential Valuables in terms of these two variables, since the “LoR” of this cluster is greater than of cluster two, the interpretation is different. “rFrequency” variable supports this difference in interpretation. “rFrequency” of this cluster is in the 7th rank among all clusters. This shows that customers in this cluster buy for fewer times relative to their “LoR”.

“IPT” of this cluster center shows that, customers in this cluster buy rarely from the company. Also “Recency” variable is very high for this cluster compared to the others. From this information, it can be concluded that customers in this cluster, although they are working with the company for almost one year did not purchase frequently.

Table 86 Cluster One Cluster Center Values and Significance Values between the Means of Clusters

Cluster *1 - Potential Invaluable*						*Significance Values between Clusters*						
						Cluster 1-2	*Cluster 1-3*	*Cluster 1-4*	*Cluster 1-5*	*Cluster 1-6*	Cluster 1-7	Cluster 1-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Rank between the clusters*	p value	p value	p value	p value	p value	p value	p value
LoR_1	357.3098	392.4930	821.5241	195.2760	6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Frequency	34.1656	66.4161	200.3611	23.6706	5	0.001	0.000	0.000	1.000	0.000	0.000	0.000
Frequency last one year	27.8133	47.8689	87.6944	17.0308	6	0.999	0.000	0.000	0.018	0.000	0.000	0.000
Recency	91.9640	14.9735	314.1062	6.1944	7	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IPT	31.8931	10.6223	106.1797	3.7310	7	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	4085.8690	9982.6234	621207.7897	2163.4934	6	0.174	0.010	0.000	1.000	0.000	0.000	0.000
rMajorTrip	37.8871	36.7469	50.9214	18.9822	3	0.000	1.000	0.000	1.000	0.000	0.000	0.000
Amount	146.8597	141.6421	3443.0654	104.1787	3	0.816	0.000	1.000	1.000	0.002	0.994	0.000
rFrequency	0.0929	0.1686	0.3164	0.0661	7	0.000	0.000	0.000	0.029	0.000	0.000	0.000
rAmount	0.5887	0.8296	7.7095	0.1874	5	0.000	0.033	0.000	0.000	0.002	0.000	0.000
rTotal Amount	10.9883	22.6327	1099.7329	8.8824	7	0.000	0.000	0.000	0.998	0.998	0.000	0.000

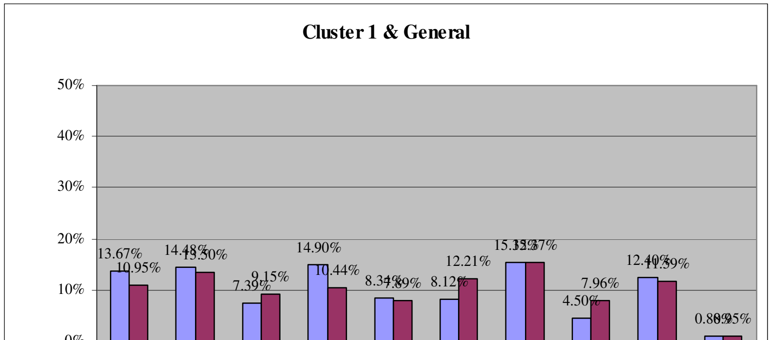
The “Total Amount” of this cluster is in the 6th rank and is almost 2.5 times smaller than the mean for the whole dataset. However, the “Amount” of this cluster is in the third rank and closer to the mean of the whole dataset. “rMajorTrip” of the cluster shows that customers in the cluster bought in a systematic manner. However, when the “rAmount” and “rTotal Amount” variables are analyzed it is observed that these are relatively small compared to other clusters and both are in 7th rank. With all these information it is concluded that customers in this clusters are the ones who do not purchase frequently and who purchase in smaller amounts. Under this circumstances customer is labeled as Potential Invaluable.

- Characteristics Related to Categorical Variables

Analyzing the figures from 94 to 104 characteristics of cluster one related to categorical variables are determined. Table 87 summarizes the main features of cluster one with respect to categorical variables.

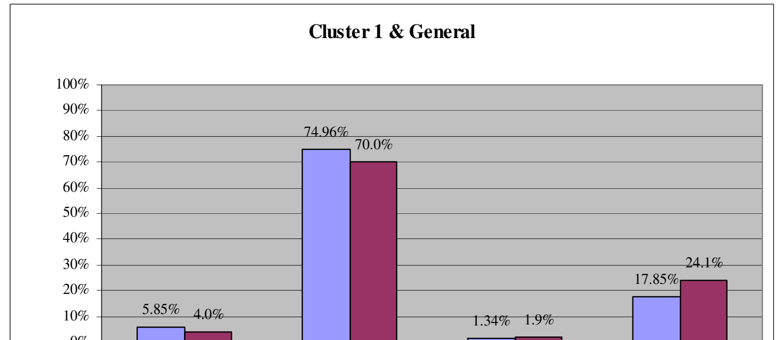
Table 87 Categorical Variables Analysis for Cluster One

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1033, 1037, 1038, 1039 and 1041
Customer Type	Open
Working Period	Does not characterize cluster based on Contingency test results (Table 68)
Customer Group	Otel Holiday village, restaurant, Market, Pension Otel Motel, Pub cafe bar, Subordinate distributor and Other
SES Group	A,B-High Income, A+, A,B-High Income, C-Average Income
Region	Does not characterize cluster based on Contingency test results (Table 68)
Position Group	Does not characterize cluster based on Contingency test results (Table 68)
Customer Structure	Does not characterize cluster based on Contingency test results (Table 68)
Visit Frequency	Does not characterize cluster based on Contingency test results (Table 68)
Customer Specialty	Company Brands, Other, NA
Working Type	CH, NA



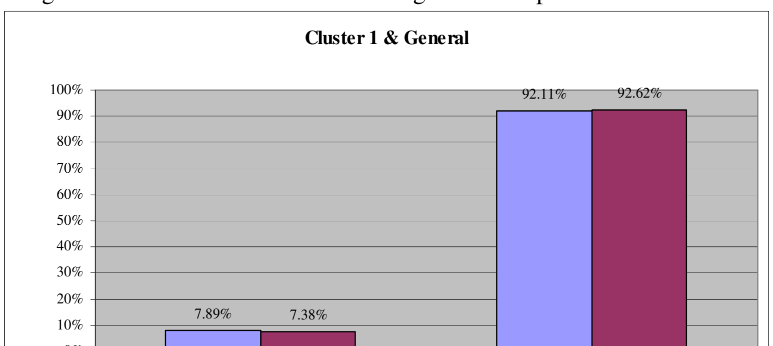
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 1	10.95%	13.50%	9.15%	10.44%	7.89%	12.21%	15.37%	7.96%	11.59%	0.95%

Figure 94 Sales directorate cluster one general comparison



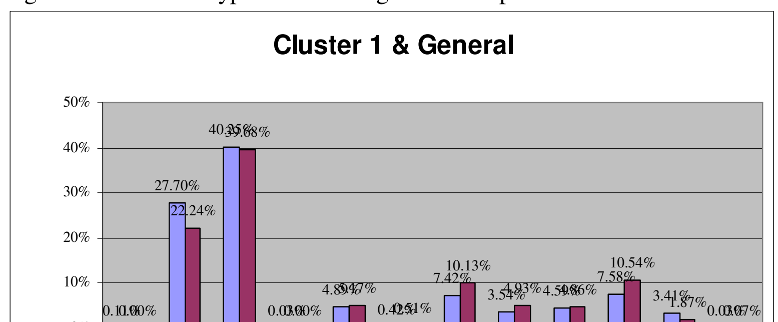
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 1	4.0%	70.0%	1.9%	24.1%

Figure 95 Customer type cluster one general comparison



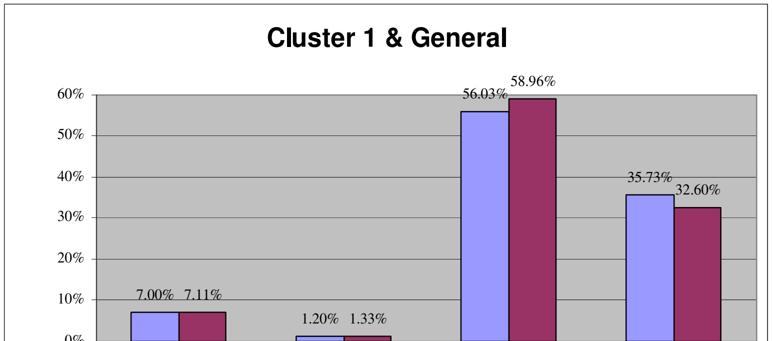
	Seasonal	Standard
General	7.89%	92.11%
Cluster 1	7.38%	92.62%

Figure 96 Working period cluster one general comparison



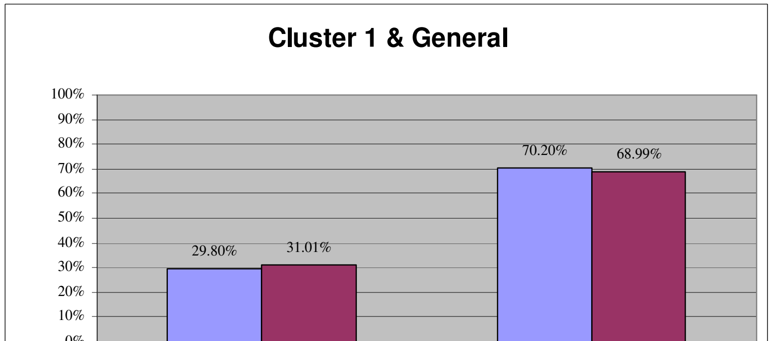
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 1	0.00%	22.24%	39.68%	0.00%	5.17%	0.51%	10.13%	4.93%	4.86%	10.54%	1.87%	0.07%

Figure 97 Customer group cluster one general comparison



	A,B High Income	A+ A,B High Level	C Average Income	D,E low Income
General	7.00%	1.20%	56.03%	35.73%
Cluster 1	7.11%	1.33%	58.96%	32.60%

Figure 98 SES status cluster one general comparison



	Around	Center
General	29.80%	70.20%
Cluster 1	31.01%	68.99%

Figure 99 Region cluster one general comparison

	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 1	31.21%	27.81%	16.18%	24.55%	0.24%

Figure 100 Position group cluster one general comparison

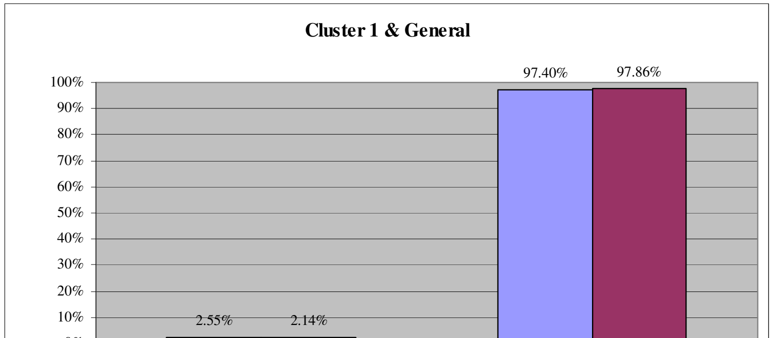
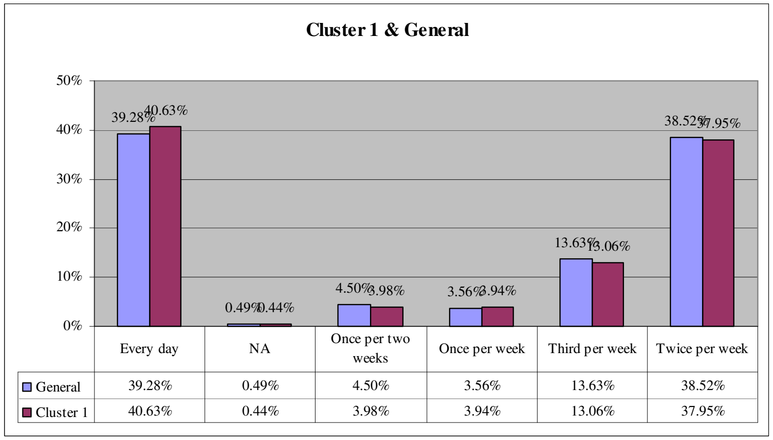


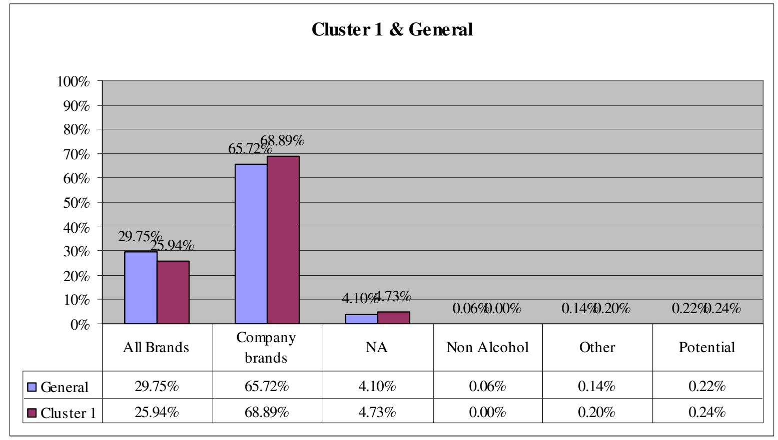
	Image	Standard
General	2.55%	97.40%
Cluster 1	2.14%	97.86%

Figure 101Customer structure cluster one general comparison



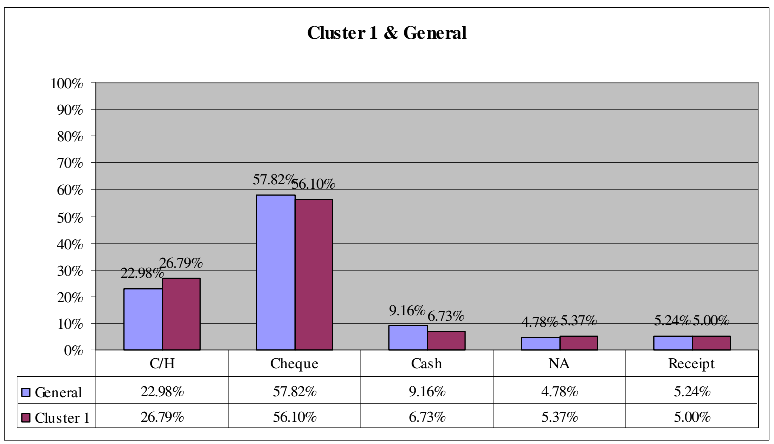
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 1	40.63%	0.44%	3.98%	3.94%	13.06%	37.95%

Figure 102 Visit frequency cluster one general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 1	25.94%	68.89%	4.73%	0.00%	0.20%	0.24%

Figure 103 Customer specialty cluster one general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 1	26.79%	56.10%	6.73%	5.37%	5.00%

Figure 104 Working type cluster one general comparison

Cluster Five

- General Characteristics

Table 88 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 88 General Characteristics of Cluster Five

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	292	7th biggest
Total Euclidian Distance of the cases from Cluster Center	933.17789	7th biggest
Average Euclidian Distance from Cluster Center	3.1958147	2nd biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	13.073502	2nd biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	9.0068771	2nd biggest

There are 292 customers in this cluster which accounts for 0.50% of the general dataset. This is a relatively small size compared to the other clusters. Cluster five is the seventh cluster among all clusters in terms of cluster size.

Normally Total Euclidian distance of the cases from the cluster center is not so high. However, Average Euclidian distance of the cluster is the second highest one among all clusters which shows that cluster five is a wide one compared to other clusters but it is not as wide as cluster three.

The distance between the cluster center and center of all clusters is represented by the Total Manhattan and Total Euclidian Distances. In terms of these variables, this cluster is in the 2nd ranking and has total distance values approximately two times greater than between clusters Manhattan and Euclidian distances. On the other hand these are quite smaller compared to cluster three. This information reveals that this cluster also contains cases that can be evaluated as outliers but not far away from the center as far as the ones in Cluster three.

- Characteristics Related to Continuous Variables

Table 89 contains summary of the information needed to interpret the continuous variables.

Customers in this cluster have been working with the company for more than 1.5 years. This value is 1.5 times smaller than the maximum, but 2.7 times greater than the minimum of this variable among all clusters. This cluster is in 4th rank among all clusters in terms of its length of relationship.

Cluster has the sixth smallest “Frequency” and eighth smallest “Frequency Last Year” measures. Therefore when the “Frequency” of the cluster is compared to its “Frequency Last Year”, its can be seen that the buying frequency of the customers in this cluster decreased in the last year of the observation. Since it has a high “LoR”, as a consequence “rFrequency” variable is lower which indicates that customers in this cluster did not buy frequently from the company.

Table 89 Cluster Five Cluster center Values and Significance Values between the Means of Clusters

Cluster 5 - Invaluable						*Significance Values between Clusters*						
						Cluster 5-1	*Cluster 5-2*	*Cluster 5-3*	*Cluster 5-4*	*Cluster 5-6*	Cluster 5-7	Cluster 5-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	p value	p value
LoR_1	515.0993	392.4930	821.5241	195.2760	4	0.000	0.000	0.002	0.000	0.000	0.001	0.000
Frequency	33.5205	66.4161	200.3611	23.6706	6	1.000	1.000	0.000	0.000	0.183	0.000	0.000
Frequency last one year	17.0308	47.8689	87.6944	17.0308	8	0.018	0.006	0.000	0.000	0.987	0.000	0.000
Recency	314.1062	14.9735	314.1062	6.1944	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IPT	106.1797	10.6223	106.1797	3.7310	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	4914.5549	9982.6234	621207.7897	2163.4934	5	1.000	0.975	0.000	0.000	0.016	0.000	0.000
rMajorTrip	39.1520	36.7469	50.9214	18.9822	2	1.000	0.000	0.998	0.003	0.000	0.139	0.459
Amount	135.8268	141.6421	3443.0654	104.1787	5	1.000	1.000	0.000	1.000	0.030	1.000	0.000
rFrequency	0.0661	0.1686	0.3164	0.0661	8	0.029	0.000	0.000	0.000	0.000	0.000	0.000
rAmount	0.2917	0.8296	7.7095	0.1874	7	0.000	0.000	0.022	0.000	0.000	1.000	0.000
rTotal Amount	8.8824	22.6327	1099.7329	8.8824	8	0.998	0.000	0.000	0.000	0.936	0.000	0.000

The “Recency” and “IPT” of this cluster are the maximum values of these two variables which mean that customers in this cluster did not buy frequently and there is almost one year between the last two purchases of them. When the ANOVA results are analyzed, it is observed that this cluster is not significantly different from the other clusters with respect to variables related to purchasing amount. Based on this information it can be said that, although the cluster center does not have the smallest value regarding to purchasing amount variables, with its high “LoR”, “IPT” and “Recency”, this cluster differs from other clusters.

“Total Amount” and “Amount” of this cluster are in the fifth rank among all clusters. These may be accepted as moderate values; however, “rAmount” and “rTotal Amount” of this cluster are in the seventh and eighth position. Since the “rAmount” and “rTotal Amount” variables are calculated by using the “LoR” variable, the values of these variables are in smaller rankings compared to “Amount” and “Total Amount” variables.

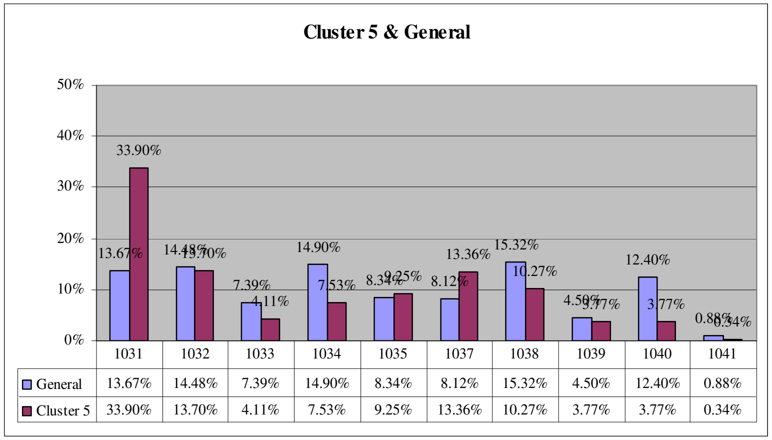
As a result of the above analyses, it is concluded that this cluster contains the customers who lost value in recent years. Higher “Total Amount” and “Amount” figures for this cluster center show that these customers bought significant amounts from the company. However, since the cluster center has the lowest “Frequency Last Year” it is obvious that customers did not purchase frequently in recent years. Based on this information customers in this cluster are labeled as “Invaluable Customers”.

- Characteristics Related to Categorical Variables

Analyzing the figures from 105 to 115 characteristics of cluster five related to categorical variables are determined. Table 90 summarizes the main features of cluster five with respect to categorical variables.

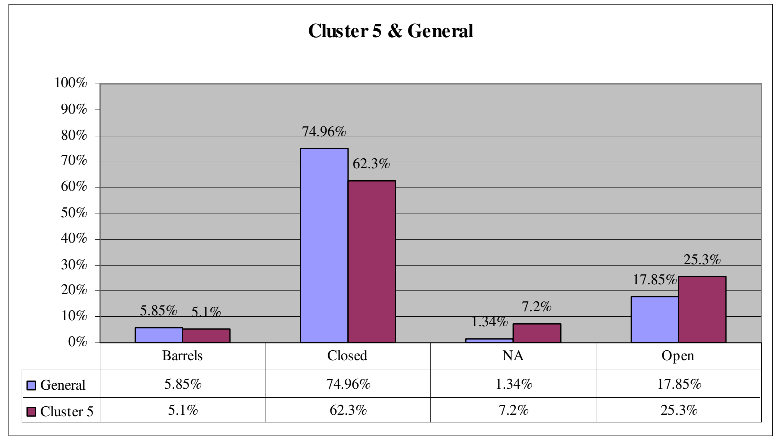
Table 90 Categorical Variables Analysis for Cluster Five

Categorical Variable	*Main Features for Cluster*
Sales Directorate	1032, 1035 and 1037
Customer Type	Open and NA
Working Period	Standard
Customer Group	Restaurant, Market , Pension Otel Motel and Other
SES Group	Does not characterize cluster based on Contingency test results (Table 68)
Region	Center
Position Group	NA
Customer Structure	Does not characterize cluster based on Contingency test results (Table 68)
Visit Frequency	Every Day, Once per week, Once per two weeks and NA
Customer Specialty	Does not characterize cluster based on Contingency test results (Table 68)
Working Type	Does not characterize cluster based on Contingency test results (Table 68)



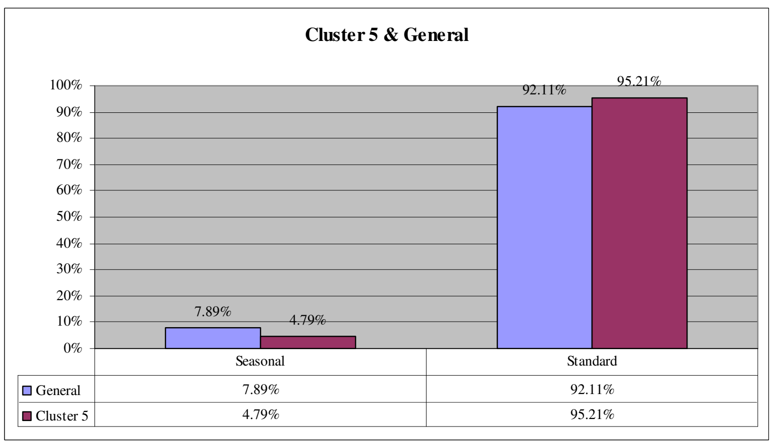
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 5	33.90%	13.70%	4.11%	7.53%	9.25%	13.36%	10.27%	3.77%	3.77%	0.34%

Figure 105 Sales directorate cluster one general comparison



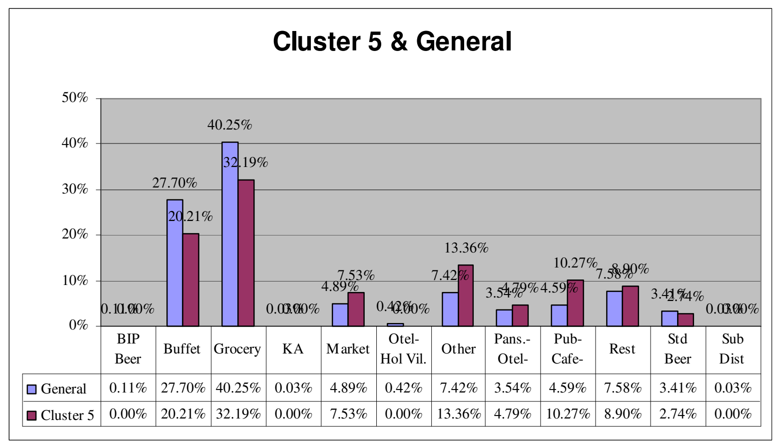
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 5	5.1%	62.3%	7.2%	25.3%

Figure 106 Customer type cluster one general comparison



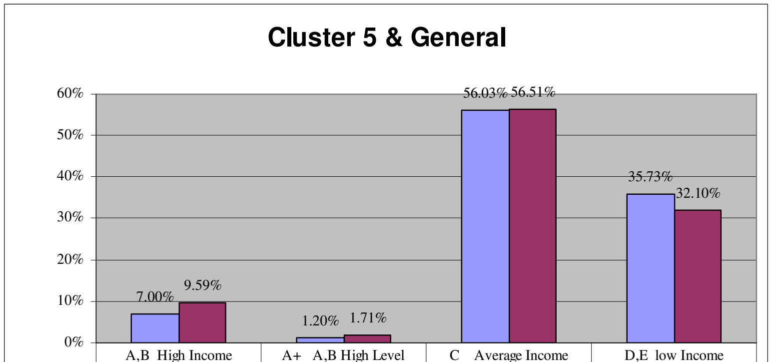
	Seasonal	Standard
General	7.89%	92.11%
Cluster 5	4.79%	95.21%

Figure 107 Working period cluster one general comparison



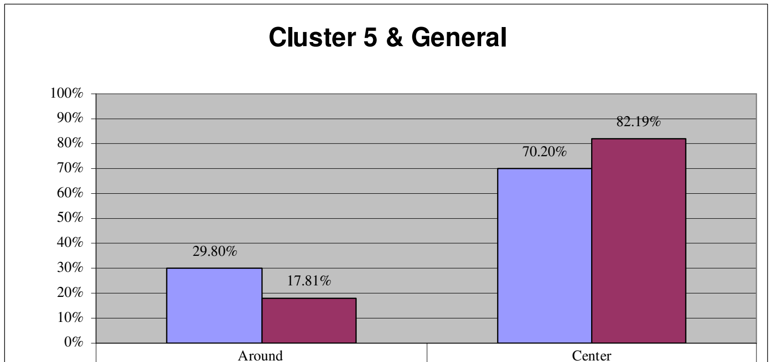
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 5	0.00%	20.21%	32.19%	0.00%	7.53%	0.00%	13.36%	4.79%	10.27%	8.90%	2.74%	0.00%

Figure 108 Customer group cluster one general comparison



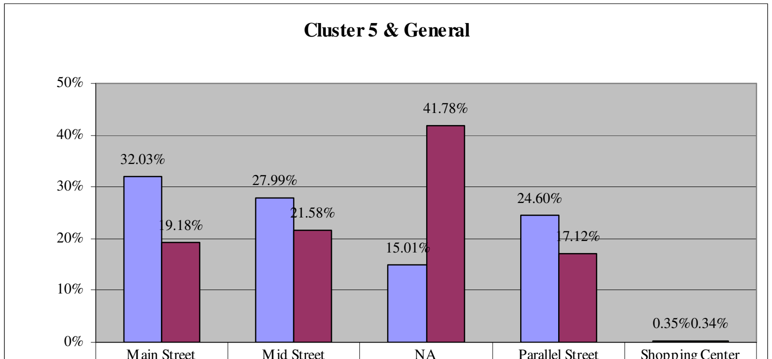
	A,B High Income	A+ A,B High Level	C Average Income	D,E low Income
General	7.00%	1.20%	56.03%	35.73%
Cluster 5	9.59%	1.71%	56.51%	32.10%

Figure 109 SES group cluster one general comparison



	Around	Center
General	29.80%	70.20%
Cluster 5	17.81%	82.19%

Figure 110 Region cluster one general comparison



	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 5	19.18%	21.58%	41.78%	17.12%	0.34%

Figure 111 Position Group cluster one general comparison

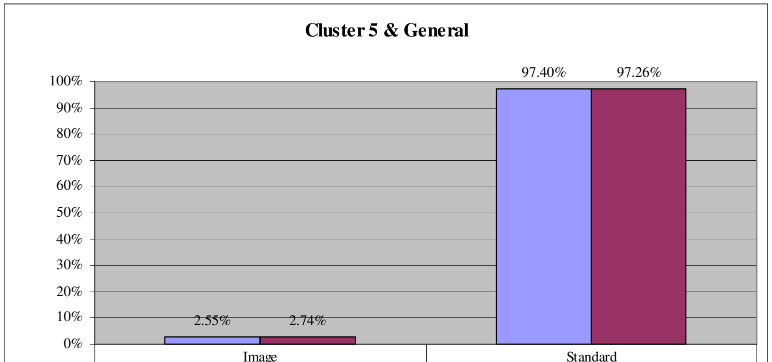
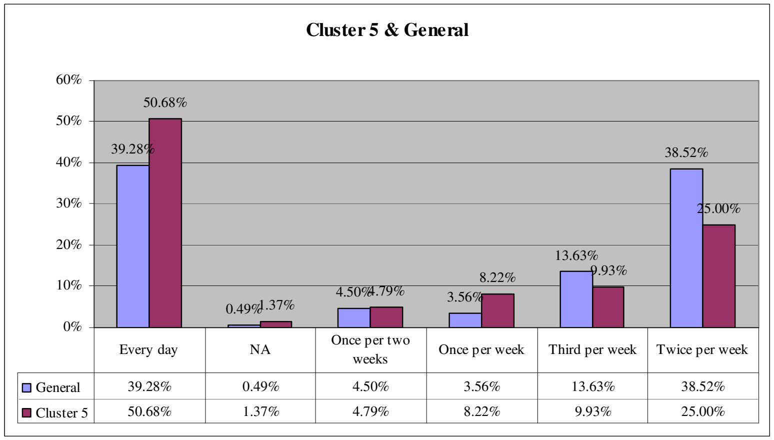


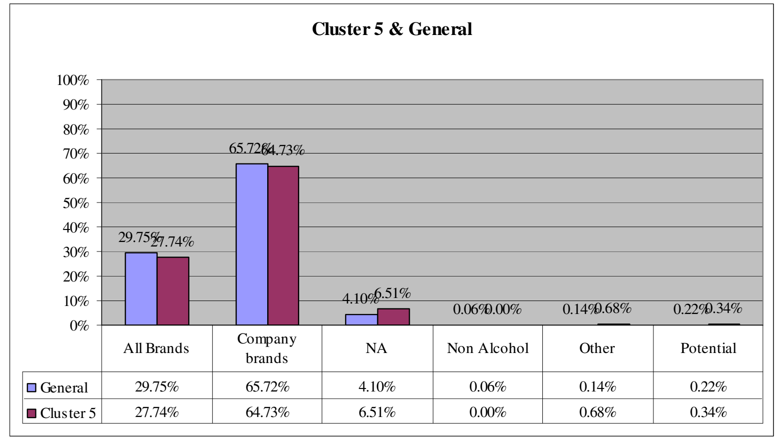
	Image	Standard
General	2.55%	97.40%
Cluster 5	2.74%	97.26%

Figure 112 Customer structure cluster one general comparison



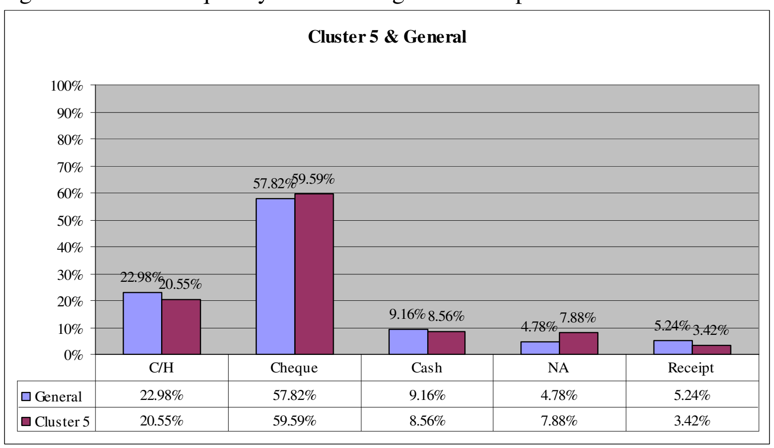
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 5	50.68%	1.37%	4.79%	8.22%	9.93%	25.00%

Figure 113 Visit frequency cluster one general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 5	27.74%	64.73%	6.51%	0.00%	0.68%	0.34%

Figure 114 Customer specialty cluster one general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 5	20.55%	59.59%	8.56%	7.88%	3.42%

Figure 115 Working type cluster one general comparison

Cluster Six

- General Characteristics

Table 91 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 91 General Characteristics of Cluster Six

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	11397	3rd biggest
Total Euclidian Distance of the cases from Cluster Center	11660.652	3rd biggest
Average Euclidian Distance from Cluster Center	1.0231334	6th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	7.6424142	3rd biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	4.4374894	3rd biggest

There are 11397 customers in this cluster which accounts for 19.67% of all dataset. This is one of the biggest clusters partitioned by the system.

This cluster is in the sixth rank among all clusters with respect to the average Euclidian distance from the cluster center which shows that the cluster is a narrow one compared to the other five clusters with higher values.

This cluster is in the 6^{th} position when the total Manhattan and Euclidian distances are considered. These measures are approximately two times greater than between clusters Manhattan and Euclidian distances. Thus, these findings indicate that the cases that are far away from the center but still not outliers grouped under cluster six.

- Characteristics Related to Continuous Variables

Table 92 contains information needed to make interpretations related to continuous variables.

“LoR” of this cluster is the smallest one among all clusters. This means that customers with the shortest length of relationships are grouped in this cluster. “LoR” for this cluster is 195 days, which is approximately half of a year.

“Frequency” of this cluster is the lowest among all clusters. On the other hand frequency for last year variable is not the lowest one. Since the “LoR” of these customers is not so high, having such a low “Frequency” is not surprising. Under these circumstances, when the “rFrequency” is analyzed, it is observed that “rFrequency” of this cluster is higher than of cluster one and cluster five which are labeled as Invaluable and Potential Invaluable respectively. In addition, this cluster is very close to cluster two which is labeled as Potential Valuable Customers in terms of this variable. Therefore, it is concluded that these customers are not the ones that have the lowest frequency but they should be interpreted as the ones that may have higher frequency values in future.

Customers in this cluster buy products from the firm approximately once in ten days. “Recency” of the cluster also supports this information with a value of 10.7. These values are closer to the mean of the general variables.

Table 92 Cluster Six Cluster Center Values and Significance Values between the Means of Clusters

Cluster 6 - *Potential Customers*						*Significance Values between Clusters*						
						Cluster 6-1	*Cluster 6-2*	*Cluster 6-3*	*Cluster 6-4*	*Cluster 6-5*	Cluster 6-7	Cluster 6-8
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	P value	p value	p value	p value	p value	p value
LoR_1	195.2760	392.4930	821.5241	195.2760	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Frequency	23.6706	66.4161	200.3611	23.6706	8	0.000	0.000	0.000	0.000	0.183	0.000	0.000
Frequency last one year	21.5331	47.8689	87.6944	17.0308	7	0.000	0.000	0.000	0.000	0.987	0.000	0.000
Recency	10.7312	14.9735	314.1062	6.1944	5	0.000	0.000	0.021	0.000	0.000	0.000	0.000
IPT	10.8010	10.6223	106.1797	3.7310	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total Amount	2163.4934	9982.6234	621207.7897	2163.4934	8	0.000	0.010	0.000	0.000	0.016	0.000	0.000
rMajorTrip	18.9822	36.7469	50.9214	18.9822	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Amount	104.1787	141.6421	3443.0654	104.1787	8	0.002	0.000	0.000	0.000	0.030	0.000	0.000
rFrequency	0.1221	0.1686	0.3164	0.0661	6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rAmount	1.0354	0.8296	7.7095	0.1874	4	0.002	0.012	0.060	0.000	0.000	0.000	0.000
rTotal Amount	11.6276	22.6327	1099.7329	8.8824	6	0.998	0.000	0.000	0.000	0.936	0.000	0.000

“rMajorTrip” of this cluster is the lowest among all. This shows that customers in this cluster are buying consistently from the company and their purchasing amounts do not fluctuate.

Both “Total Amount” and “Amount” are the lowest for this cluster. However, just like “Frequency”, since the “LoR” is lower for this cluster, this result is not a surprising issue. When the “rAmount” value is analyzed it is observed that this cluster is closer to cluster two which is labeled as Potential Valuable Customers. Also when the differences between the clusters are analyzed by ANOVA, it is examined that cluster six has a similar pattern to cluster three, stars.

As a consequence of the issues discussed above, it is concluded that, in spite of purchasing in lower amounts, customers in cluster six can be accepted as “Potential Customers”, just like the ones in cluster two.

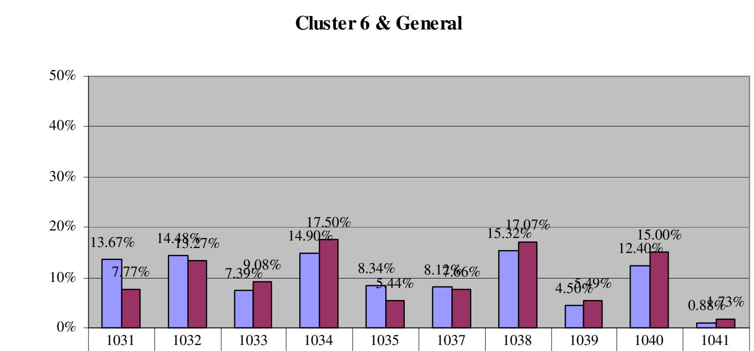
- Characteristics Related to Categorical Variables

Analyzing the figures from 116 to 126 characteristics of cluster six related to categorical variables are determined. Table 93 summarizes the main features of cluster six with respect to categorical variables.

Table 93 Categorical Variables Analysis for Cluster Six

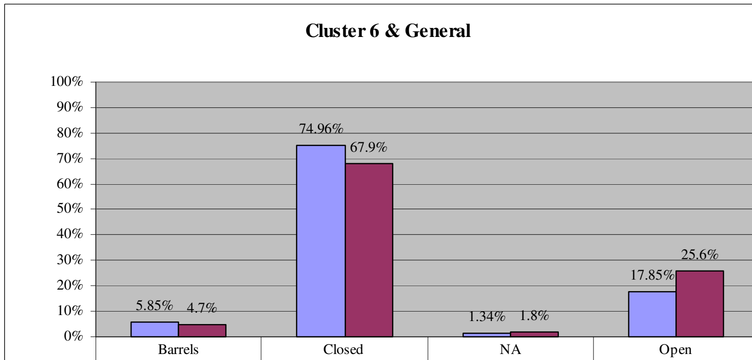
Categorical Variable	*Main Features for Cluster*
Sales Directorate	1033, 1034, 1038, 1039, 1040 and 1041
Customer Type	Open and NA
Working Period	Seasonal
Customer Group	Project Beer House, Other, Key account, Restaurant, Market , Pension Otel Motel, Pub cafe Bar, Subordinate Distributor and Other
SES Group	C-Average Income
Region	Center
Position Group	Does not characterize cluster based on Contingency test results (Table 68)
Customer Structure	Does not characterize cluster based on Contingency test results (Table A)
Visit Frequency	Every day, NA

Categorical Variable	*Main Features for Cluster*
Customer Specialty	Company brands and Non Alcohol, Other, Potential Customers
Working Type	Cash



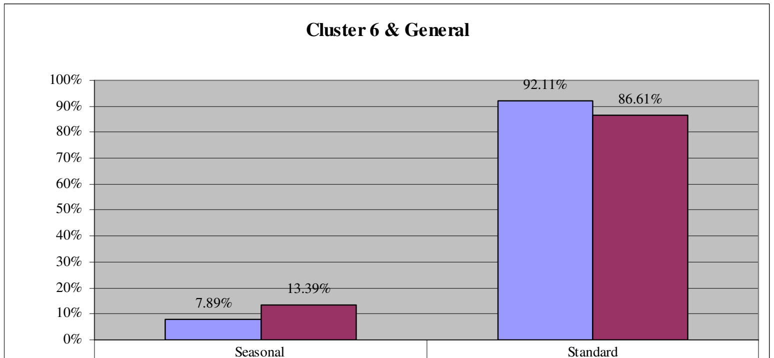
	1031	1032	1033	1034	1035	1037	1038	1039	1040	1041
General	13.67%	14.48%	7.39%	14.90%	8.34%	8.12%	15.32%	4.50%	12.40%	0.88%
Cluster 6	7.77%	13.27%	9.08%	17.50%	5.44%	7.66%	17.07%	5.49%	15.00%	1.73%

Figure 116 Sales directorate cluster one general comparison



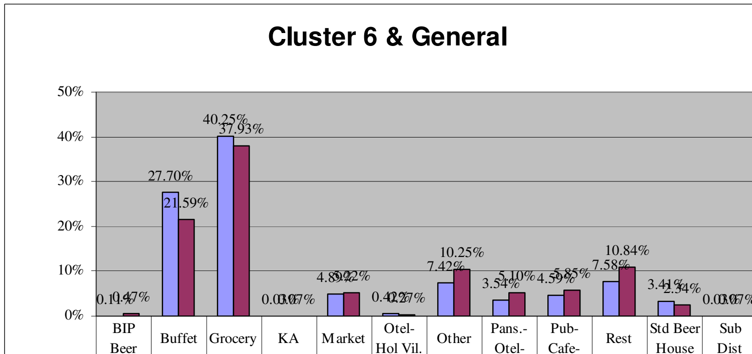
	Barrels	Closed	NA	Open
General	5.85%	74.96%	1.34%	17.85%
Cluster 6	4.7%	67.9%	1.8%	25.6%

Figure 117 Customer type cluster one general comparison



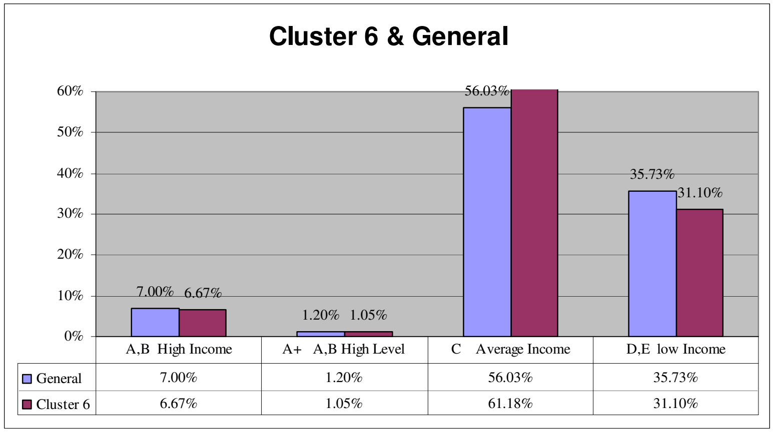
	Seasonal	Standard
General	7.89%	92.11%
Cluster 6	13.39%	86.61%

Figure 118 Working period cluster one general comparison



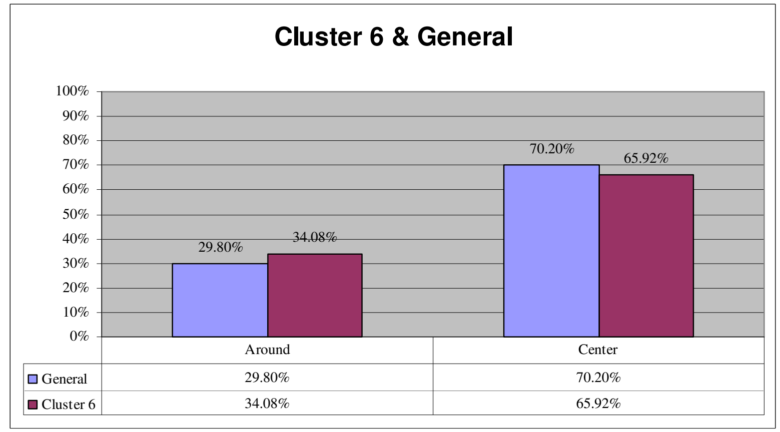
	BIP Beer	Buffet	Grocery	KA	Market	Otel-Hol Vil.	Other	Pans.-Otel-	Pub-Cafe-	Rest	Std Beer House	Sub Dist
General	0.11%	27.70%	40.25%	0.03%	4.89%	0.42%	7.42%	3.54%	4.59%	7.58%	3.41%	0.03%
Cluster 6	0.47%	21.59%	37.93%	0.07%	5.22%	0.27%	10.25%	5.10%	5.85%	10.84%	2.34%	0.07%

Figure 119 Customer group cluster one general comparison



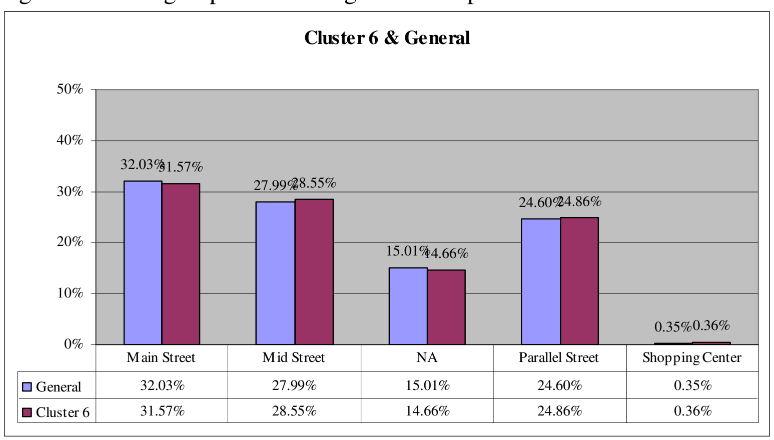
	A,B High Income	A+ A,B High Level	C Average Income	D,E low Income
General	7.00%	1.20%	56.03%	35.73%
Cluster 6	6.67%	1.05%	61.18%	31.10%

Figure 120 SES group cluster one general comparison



	Around	Center
General	29.80%	70.20%
Cluster 6	34.08%	65.92%

Figure 121 Region cluster one general comparison



	Main Street	Mid Street	NA	Parallel Street	Shopping Center
General	32.03%	27.99%	15.01%	24.60%	0.35%
Cluster 6	31.57%	28.55%	14.66%	24.86%	0.36%

Figure 122 Position group cluster one general comparison

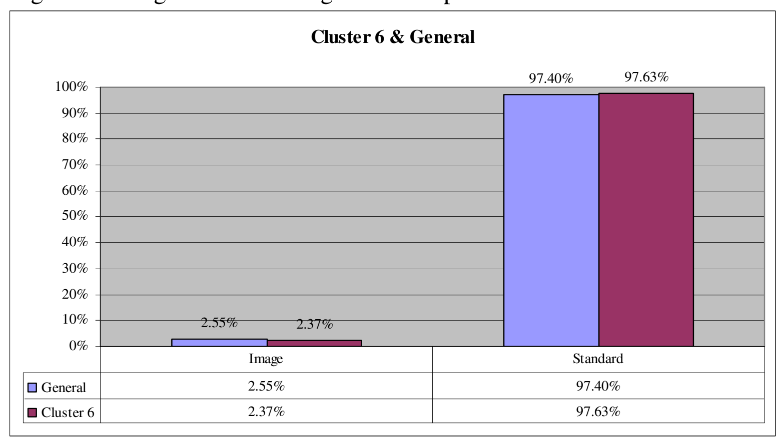
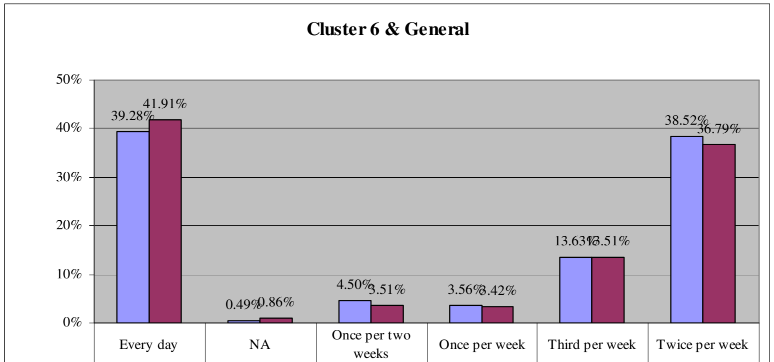


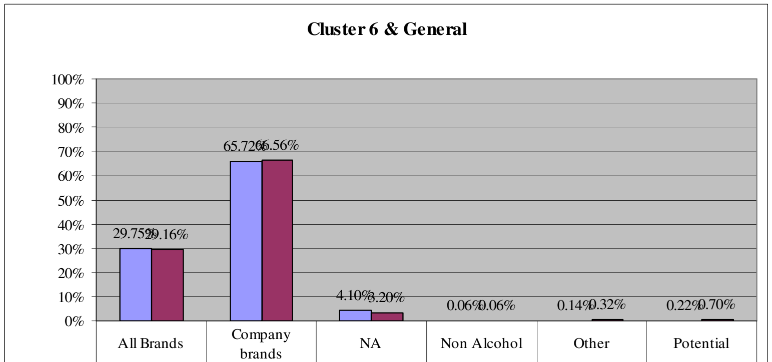
	Image	Standard
General	2.55%	97.40%
Cluster 6	2.37%	97.63%

Figure 123 Customer structure cluster one general comparison



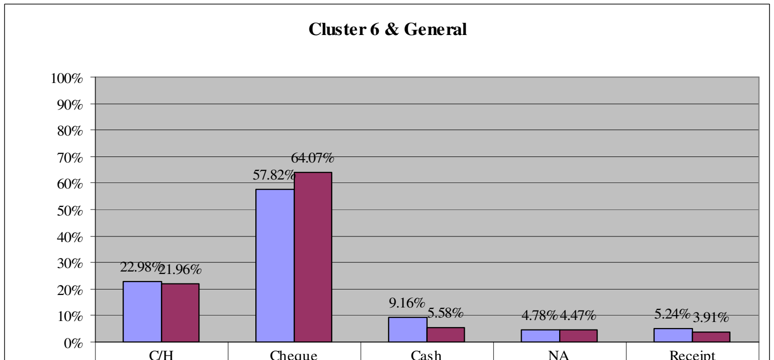
	Every day	NA	Once per two weeks	Once per week	Third per week	Twice per week
General	39.28%	0.49%	4.50%	3.56%	13.63%	38.52%
Cluster 6	41.91%	0.86%	3.51%	3.42%	13.51%	36.79%

Figure 124 Visit frequency cluster one general comparison



	All Brands	Company brands	NA	Non Alcohol	Other	Potential
General	29.75%	65.72%	4.10%	0.06%	0.14%	0.22%
Cluster 6	29.16%	66.56%	3.20%	0.06%	0.32%	0.70%

Figure 125 Customer specialty cluster one general comparison



	C/H	Cheque	Cash	NA	Receipt
General	22.98%	57.82%	9.16%	4.78%	5.24%
Cluster 6	21.96%	64.07%	5.58%	4.47%	3.91%

Figure 126 Working Type cluster one general comparison

Interpretation of City Clusters

At the end of the cluster analysis applied to the dataset, cities are grouped under seven clusters based on information gained from variables determined via factor analysis. Table 94 shows the clusters with the number of cases partitioned into them and the corresponding percentage of their size compared to all dataset.

Table 94Cluster Information for City Clusters

Cluster Number	*Number of Cases in Cluster*	Percentage *of* Data in *Cl*uster
1	22	28.21
2	34	43.59
3	3	3.85
4	2	2.56
5	4	5.13
6	4	5.13
7	9	11.54
Total	78	100

Cluster interpretation summaries can be found in Appendix B.

- Characteristics related to continuous variables

Table 95 shows the results of the ANOVA analysis applied to the dataset. Figures in Table 95 indicate that these variables show significant differences between the clusters and can be used to interpret them.

Table 95 Significance Testing of Variables

ANOVA City					
Variable	Sum of Squares	df	Mean Square	F	Sig.
Average Sales for City_2	55.624	6	9.271	30.792	0.000
Average IPT for City	35.279	6	5.880	10.006	0.000
Average frequency for City	55.585	6	9.264	30.716	0.000
Count of Customers in the City	32.216	6	5.369	8.512	0.000
Average Frequency last year for City	49.845	6	8.308	21.721	0.000
Average Recency for City	65.135	6	10.856	64.964	0.000
Sales per Customer in the City	59.242	6	9.874	39.478	0.000

Clusters' centroids will be used as a guide to interpret them. Since the dataset is transformed before the partition process start, z-scores for the clusters' centroids converted to original variables for interpretation. Table 96 shows final cluster centers determined by the system with the corresponding original values for the variables used in the segmentation. On the other hand, Table 97 shows the z-scores and original values for the other variables that will be used in interpretations.

Table 96 Final Cluster Centers in z-values and Original Values for Segmentation Variables

Final Cluster Centers								
		Cluster						
		1	2	3	4	5	6	7
Average Sales for City_2	z-value	-0.277	-0.171	4.051	-0.103	-0.298	0.838	-0.245
	Original value	144.916	158.842	715.954	167.882	142.089	291.990	149.059
Average frequency for City	z-value	1.028	-0.434	-1.774	-1.741	-1.284	0.883	0.283
	Original value	81.754	47.132	15.396	16.167	26.998	78.326	64.111
Average Recency for City	z-value	-0.588	-0.023	-1.235	4.282	1.736	0.972	-0.221
	Original value	11.861	17.030	5.941	56.417	33.126	26.133	15.219
Sales per Customer in the City	z-value	0.011	-0.360	-0.689	-0.844	-0.669	-0.626	2.325
	Original value	4.500	2.552	0.822	0.009	0.927	1.156	16.654

Table 97 Final Cluster Centers in z-values and Original Values for Control Variables

Final Cluster Centers								
		Cluster						
		1	2	3	4	5	6	7
Average IPT for City	z-value	-0.658	0.226	-0.560	1.058	2.243	-0.379	-0.122
	Original value	7.859	12.522	8.377	16.912	23.159	9.329	10.688
Count of Customers in the City	z-value	-0.150	-0.233	-0.366	-0.384	-0.358	-0.352	1.769
	Original value	455.045	294.059	38.667	3.500	52.500	65.250	4152.889
Average Frequency last year for City	z-value	1.062	-0.384	-1.401	-1.695	-1.235	0.748	-0.084
	Original value	71.007	42.992	23.294	17.583	26.499	64.930	48.798

▪ <u>Interpretation and Profiling Sequence for Customer Clusters</u>

Clusters are interpreted in their original order.

<u>Cluster One</u>

- General Characteristics

Table 98 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 98 General Characteristics of Cluster One

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	22	2nd biggest
Total Euclidian Distance of the cases from Cluster Center	19.365819	2nd biggest
Average Euclidian Distance from Cluster Center	0.8802645	4th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	3.7046759	4th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	2.119596	4th biggest

With a value of twenty two, cluster one is the second greatest cluster constructed at the end of the partitioning process. This value accounts for the 28.2% of all dataset which is a significantly high percentage compared to other clusters.

Average Euclidian distance of this cluster is in the fourth rank among all clusters which makes the cluster have an average level of wideness among all clusters.

The cluster is in the fourth rank when Total Manhattan and Total Euclidian distances are considered. Based on this information, it can be concluded that cases in this cluster are not far away from the center of all clusters and are not outliers.

- Characteristics Related to Continuous Variables

Table 99 contains information to interpret the continuous variables representing the cluster center.

Figures in Table 99 show that cities with greatest “frequency” values and shortest “IPT” values are partitioned into cluster one. Cities in this cluster are in the second rank in terms of “Count of Customer” located in them. In addition, the cluster is in the second rank when “Sales per Customer” variable is considered which shows that the consumption is high in the cities in this cluster. In order to figure out whether cities in this cluster are valuable, a variable “Total Sales”, is calculated by multiplying the “Average Sales” with “Count of Customers”. Although cities in this cluster are in the second rank in terms of “Total Sales” because of having high “Customer Count”, they have smaller values for “Average Sales” variable compared to other clusters. ANOVA show that, different from Cluster 7 which contains the most valuable cities, cities in this cluster have the highest “Average Frequency”. The reason for this difference lies in the big disparity between the numbers of customers in the two clusters. Based on information shown in Table 99, it can be concluded that cluster one contains cities whose customers purchase in high amounts from the company but not as high as the ones in cluster seven. Therefore, this cluster is named as Most Valuables.

Table 99 Cluster One Cluster Center Values and Significance Values between the Means of Clusters

Cluster 1 - Most Valuables						*Significance Values between Clusters*					
						Cluster 1-2	*Cluster 1-3*	*Cluster 1-4*	*Cluster 1-5*	Cluster 1-6	Cluster 1-7
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Rank between the clusters*	p value	p value	p value	p value	p value	p value
Average Sales for City	144.92	181.41	715.95	142.09	6	1.000	0.452	1.000	1.000	0.981	1.000
Average IPT for City	7.86	11.33	23.16	7.86	1	0.000	1.000	1.000	0.860	1.000	0.253
Count of Customers in the City	455.05	743.32	4152.89	3.50	2	1.000	0.129	0.060	0.156	0.180	0.519
Average frequency for City	81.75	57.41	81.75	15.40	1	0.000	0.368	0.315	0.034	1.000	0.161
Average Frequency last year for City	71.01	50.43	71.01	17.58	1	0.000	0.773	0.147	0.029	1.000	0.000
Average Recency for City	11.86	17.24	56.42	5.94	2	0.000	0.990	0.080	0.042	0.084	0.055
Population for City (2000)	725002.00	852692.71	2506369.22	162718.50	4	1.000	1.000	1.000	0.007	1.000	0.922
Sales per Customer in the City	4.50	4.44	16.65	0.01	2	0.189	0.117	0.000	0.018	0.003	0.000
Total sales	3587600.79		49045634.80	8468.40	2	0.425	0.515	0.006	0.009	0.151	0.888
Total frequency	33140.00		289202.78	42.00	2	0.686	0.027	0.018	0.023	0.091	0.809

Cluster Two

- General Characteristics

Table 100 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 100 General Characteristics of Cluster Two

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	34	1st biggest
Total Euclidian Distance of the cases from Cluster Center	27.878568	1st biggest
Average Euclidian Distance from Cluster Center	0.8199579	5th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	1.6777833	7th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.0453678	7th biggest

Thirty four cities are partitioned into cluster two which accounts for 43.6%, nearly half, of the all dataset. Thus, cluster two is the most crowded one among all clusters.

Cluster two has the fifth biggest average Euclidian distance which indicates that cases in the cluster are not very far away from each other and this cluster can be accepted as a relatively narrow one.

Cluster two has the smallest distance from the center of all clusters. Based on this information it can be concluded that cities in cluster two are the closest cities to the center of all clusters.

- Characteristics Related to Continuous Variables

Table 101 contains information needed to make interpretations related to continuous variables representing the cluster center.

Cluster two is in the third rank with respect to “Count of Customer”, “Sales per Customer”, “Total Amount” and “Total Frequency” variables. This shows that cities in this cluster include customers who purchase from the company frequently and in relatively higher amounts. ANOVA results show that the difference between cluster one (Most Valuables) and this cluster results from the time between purchases and frequency related variables; customers located at cities partitioned into cluster two are purchasing from the company not as frequently as the ones in cluster one. As a result, because it has many similarities with cluster one and has higher values from the preceding clusters, this cluster is named as Valuables.

Table 101 Cluster Two Cluster Center Values and Significance Values between the Means of Clusters

Cluster 2 – Valuables						*Significance Values between Clusters*					
						Cluster 2-1	*Cluster 2-3*	*Cluster 2-4*	*Cluster 2-5*	Cluster 2-6	Cluster 2-7
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	p value
Average Sales for City	158.84	181.41	715.95	142.09	4	1.000	0.460	1.000	1.000	0.992	1.000
Average IPT for City	12.52	11.33	23.16	7.86	5	0.000	1.000	1.000	0.986	0.986	0.870
Count of Customers in the City	294.06	743.32	4152.89	3.50	3	1.000	0.108	0.012	0.136	0.156	0.458
Average frequency for City	47.13	57.41	81.75	15.40	4	0.000	0.905	0.888	0.740	0.754	0.159
Average Frequency last year for City	42.99	50.43	71.01	17.58	4	0.000	0.999	0.816	0.753	0.870	0.589
Average Recency for City	17.03	17.24	56.42	5.94	4	0.000	0.762	0.115	0.108	0.331	0.803
Population for City (2000)	625550.82	852692.71	2506369.22	162718.50	5	1.000	1.000	1.000	0.000	1.000	0.916
Sales per Customer in the City	2.55	4.44	16.65	0.01	3	0.189	0.882	0.000	0.639	0.418	0.000
Total sales	1521858.5		49045634.8	8468.40	3	0.425	1.000	0.000	0.001	0.998	0.852
Total frequency	14391.47		289202.78	42.00	3	0.686	0.063	0.014	0.027	0.706	0.726

Cluster Three

- General Characteristics

Table 102 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 102 General Characteristics of Cluster Three

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	3	6th biggest
Total Euclidian Distance of the cases from Cluster Center	3.9329953	5th biggest
Average Euclidian Distance from Cluster Center	1.3109984	2nd biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	7.3537882	1st biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	4.2640695	1st biggest

Three cities are partitioned to this cluster that makes this cluster one of the smallest clusters.

Cluster three has the second highest Average Euclidian distance value, which shows that most cases in this cluster are not close to their center and the cluster is a wide one compared to the others.

With respect to Total Manhattan and Total Euclidian Distances, this cluster is in the first rank. It is the farthest cluster from the center of all clusters. This information indicates that this cluster can be evaluated as a cluster contains outliers.

- Characteristics Related to Continuous Variables

Table 103 contains information needed to make interpretations related to continuous variables representing the cluster center.

Cluster three is in the fifth position among all clusters with respect to “Sales per Customer”, “Total Sales” and “Total Frequency” variables and in the sixth position for the “Count of Customers” variable. Cities in this cluster do not include customers who purchase in high amounts from the company. However, when this cluster is compared to others, it is obvious that there are worse ones. Based on this information, cluster is named as Fit Class.

Table 103 Cluster Three Cluster Center Values and Significance Values between the Means of Clusters

Cluster 3 - Fit Class						*Significance Values between Clusters*					
	Value of Cluster Center	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Rank between the clusters*	*Cluster 3-1*	*Cluster 3-2*	*Cluster 3-4*	*Cluster 3-5*	Cluster 3-6	Cluster 3-7
Variables						p value	p value	p value	p value	p value	p value
Average Sales for City	715.95	181.41	715.95	142.09	1	0.452	0.460	0.457	0.378	0.437	0.410
Average IPT for City	8.38	11.33	23.16	7.86	2	1.000	1.000	1.000	0.934	1.000	1.000
Count of Customers in the City	38.67	743.32	4152.89	3.50	6	0.129	0.108	1.000	1.000	1.000	0.372
Average frequency for City	15.40	57.41	81.75	15.40	7	0.368	0.905	1.000	1.000	0.223	0.515
Average Frequency last year for City	23.29	50.43	71.01	17.58	6	0.773	0.999	1.000	1.000	0.790	0.992
Average Recency for City	5.94	17.24	56.42	5.94	1	0.990	0.762	0.026	0.043	0.143	0.869
Population for City (2000)	424171.33	852692.71	2506369.22	162718.50	6	1.000	1.000	1.000	1.000	1.000	0.860
Sales per Customer in the City	0.82	4.44	16.65	0.01	5	0.117	0.882	1.000	1.000	1.000	0.000
Total sales	767757.89		49045634.80	8468.40	5	0.515	1.000	1.000	1.000	1.000	0.838
Total frequency	1413.67		289202.78	42.00	5	0.027	0.063	1.000	1.000	0.992	0.666

Cluster Four

- General Characteristics

Table 104 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 104 General Characteristics of Cluster Four

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	2	7th biggest
Total Euclidian Distance of the cases from Cluster Center	0.6319068	7th biggest
Average Euclidian Distance from Cluster Center	0.3159534	7th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	6.252274	2nd biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	3.9306492	2nd biggest

Cluster four contains only two cities which is approximately 2.5% of all dataset. Thus, cluster four is the smallest cluster.

Since there are only two cases in the cluster, it has the smallest value with respect to total distance from the cluster center. In addition, cluster has the lowest average distance. This information indicates that this is the narrowest cluster among all.

Total Manhattan and Total Euclidian distances for this cluster show that cases in this cluster are not very close to the center of all clusters by having the second biggest value among all clusters. Therefore, cases in this cluster can be accepted as outliers and not close to the center of all clusters.

- Characteristics Related to Continuous Variables

Table 105 contains information needed to make interpretations related to continuous variables representing the cluster center.

Cities with lowest “Sales per Customer”, “Total Amount” and “Total Frequency” measures are partitioned into cluster four. ANOVA results show that cluster four is significantly different from cluster seven, cluster one and cluster two which include valuable cities, with respect to “Sales per Customer” variable. Having the minimum value for “Sales per Customer” variable, cluster four contains the cities in which products of the company are not consumed. In addition, cluster four contains the cities with fewer number of customers compared to the other clusters. Thus, although the cluster has the lowest “Total Sales” and “Average Sales per Customer”, the cluster is in the first position in terms of “Average Sales”. With all these information this cluster is named as Most Invaluable.

Table 105 Cluster Four Cluster Center Values and Significance Values between the Means of Clusters

Cluster 4 - Most Invaluable						*Significance Values between Clusters*					
						Cluster 4-1	*Cluster 4-2*	*Cluster 4-3*	*Cluster 4-5*	Cluster 4-6	Cluster 4-7
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	p value
Average Sales for City	167.88	181.41	715.95	142.09	3	1.000	1.000	0.457	1.000	0.996	1.000
Average IPT for City	16.91	11.33	23.16	7.86	6	1.000	1.000	1.000	1.000	1.000	1.000
Count of Customers in the City	3.50	743.32	4152.89	3.50	7	0.060	0.012	1.000	0.998	0.926	0.361
Average frequency for City	16.17	57.41	81.75	15.40	6	0.315	0.888	1.000	1.000	0.149	0.202
Average Frequency last year for City	17.58	50.43	71.01	17.58	7	0.147	0.816	1.000	0.999	0.187	0.630
Average Recency for City	56.42	17.24	56.42	5.94	7	0.080	0.115	0.026	0.026	0.007	0.105
Population for City (2000)	875693.00	852692.71	2506369.22	162718.50	2	1.000	1.000	1.000	1.000	1.000	0.989
Sales per Customer in the City	0.01	4.44	16.65	0.01	7	0.000	0.000	1.000	0.988	0.711	0.000
Total sales	8468.40		49045634.80	8468.40	7	0.006	0.000	1.000	0.988	0.906	0.822
Total frequency	42.00		289202.78	42.00	7	0.018	0.014	1.000	0.987	0.903	0.660

Cluster Five

- General Characteristics

Table 106 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 106 General Characteristics of Cluster Five

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	4	5th biggest
Total Euclidian Distance of the cases from Cluster Center	3.0591581	6th biggest
Average Euclidian Distance from Cluster Center	0.7647895	6th biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	3.2703681	2nd biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.6717354	2nd biggest

There are four cities in cluster five which is approximately 5.5% of all dataset which makes the cluster one of the smallest clusters.

Cluster five is has the sixth highest distance with to the cluster center variables which shows that cases in this cluster are not close to each other and the cluster is one of the widest ones.

The distance between the cluster center and center of all clusters is represented by the Total Manhattan and Total Euclidian Distances. This cluster is in the second rank among all clusters for these distance measures. This information shows that this cluster also contains cases that can be evaluated as outliers.

- Characteristics Related to Continuous Variables

Table 107 contains information needed to make interpretations related to continuous variables representing the cluster center.

Cluster five contains invaluable cities whose “Sales per Customer” values are very small as its “Total Sales”. Cities in this cluster do not include many customers. There are fifty two customers on average in the cities partitioned into this cluster. Moreover, cities with the lowest population are assigned to this cluster. Despite the cluster has low “Count of Customers” and population variables which are the denominators in the calculation of average values, cluster center has the smallest “Average Sales” and the second smallest “Sales per Customer” scores. This verifies that cities in this cluster include customers who purchase very small amounts from the company. Because of its similarities with Most Invaluable cluster, this cluster is labeled as Invaluable.

Table 107 Cluster Five Cluster Center Values and Significance Values between the Means of Clusters

Cluster 5 - Invaluable						*Significance Values between Clusters*					
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	*Cluster 5-1*	*Cluster 5-2*	*Cluster 5-3*	*Cluster 5-4*	Cluster 5-6	Cluster 5-7
						p value	p value	p value	p value	p value	p value
Average Sales for City	142.09	181.41	715.95	142.09	7	1.000	1.000	0.378	1.000	0.978	1.000
Average IPT for City	23.16	11.33	23.16	7.86	7	0.860	0.986	0.934	1.000	0.909	0.954
Count of Customers in the City	52.50	743.32	4152.89	3.50	5	0.156	0.136	1.000	0.998	1.000	0.376
Average frequency for City	27.00	57.41	81.75	15.40	5	0.034	0.740	1.000	1.000	0.202	0.100
Average Frequency last year for City	26.50	50.43	71.01	17.58	5	0.029	0.753	1.000	0.999	0.282	0.430
Average Recency for City	33.13	17.24	56.42	5.94	6	0.042	0.108	0.043	0.026	0.823	0.077
Population for City (2000)	162718.50	852692.71	2506369.22	162718.50	7	0.007	0.000	1.000	1.000	0.869	0.702
Sales per Customer in the City	0.93	4.44	16.65	0.01	6	0.018	0.639	1.000	0.988	1.000	0.000
Total sales	140165.88		49045634.80	8468.40	6	0.009	0.001	1.000	0.988	0.963	0.825
Total frequency	910.50		289202.78	42.00	6	0.023	0.027	1.000	0.987	0.965	0.664

Cluster Six

- General Characteristics

Table 108 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 108 General Characteristics of Cluster Six

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	4	4th biggest
Total Euclidian Distance of the cases from Cluster Center	5.3721552	4th biggest
Average Euclidian Distance from Cluster Center	1.3430388	1st biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	2.3855098	5th biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	1.4658811	5th biggest

Just like cluster five there are four cities in this cluster which is approximately 5.5% of all dataset. Thus, cluster six can also be accepted as a small cluster.

By having the biggest value with respect to average Euclidian distance from the cluster center, it is obvious that cluster six is the widest one among all clusters.

Cluster six has the fifth highest Total Manhattan Distance and Total Euclidian Distance measures. This information shows that cases in this cluster are not close to the center of all clusters and can be regarded as outliers.

- Characteristics Related to Continuous Variables

Table 109 contains information needed to make interpretations related to continuous variables representing the cluster center.

The cluster is in the fourth position with respect to “Sales per Customer”, “Total Sales”, and “Count of Customers” variables.

When the frequency related variables are analyzed, it is observed that the cluster is in the second rank. This may be seen as a high value however, since the “Average Frequency” is calculated using the “Total Frequency” of the city and “Count of Customers”; this does not mean that customers in these cities are buying more frequently than other ones however, this high frequency figure is observed because the cluster has fewer customers than many other clusters. When the Total Frequency variable is analyzed, the cluster lies in the fourth rank. All these findings indicate that this cluster can be named as Averages.

Table 109 Cluster Six Cluster Center Values and Significance Values between the Means of Clusters

Cluster 6 - Averages						*Significance Values between Clusters*					
						Cluster 6-1	*Cluster 6-2*	*Cluster 6-3*	*Cluster 6-4*	Cluster 6-5	Cluster 6-7
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	p value	p value	p value	p value	p value	p value
Average Sales for City	291.99	181.41	715.95	142.09	2	0.981	0.992	0.437	0.996	0.978	0.984
Average IPT for City	9.33	11.33	23.16	7.86	3	1.000	0.986	1.000	1.000	0.909	1.000
Count of Customers in the City	65.25	743.32	4152.89	3.50	4	0.180	0.156	1.000	0.926	1.000	0.380
Average frequency for City	78.33	57.41	81.75	15.40	2	1.000	0.754	0.223	0.149	0.202	0.999
Average Frequency last year for City	64.93	50.43	71.01	17.58	2	1.000	0.870	0.790	0.187	0.282	0.981
Average Recency for City	26.13	17.24	56.42	5.94	5	0.084	0.331	0.143	0.007	0.823	0.201
Population for City (2000)	764790.50	852692.71	2506369.22	162718.50	3	1.000	1.000	1.000	1.000	0.869	0.959
Sales per Customer in the City	1.16	4.44	16.65	0.01	4	0.003	0.418	1.000	0.711	1.000	0.000
Total sales	911951.35		49045634.80	8468.40	4	0.151	0.998	1.000	0.906	0.963	0.840
Total frequency	5429.00		289202.78	42.00	4	0.091	0.706	0.992	0.903	0.965	0.685

Cluster Seven

- General Characteristics

Table 110 shows distance measures calculated for the cluster as well as the order of these measures among all clusters.

Table 110 General Characteristics of Cluster Seven

Subject	*Value*	*Status Among all Clusters*
Number of Cases in the Cluster	9	3rd biggest
Total Euclidian Distance of the cases from Cluster Center	9.0059068	3rd biggest
Average Euclidian Distance from Cluster Center	1.0006563	3rd biggest
Total Manhattan Distance of the Cluster Center form the Center of the all Clusters	4.8748834	3rd biggest
Total Euclidian Distance of the Cluster Center form the Center of the all Clusters	2.8235641	3rd biggest

By having nine cities, cluster seven is the third biggest cluster. The size of cluster accounts for 11.53% of all dataset. However, when it is compared to the first and second biggest cluster this cluster is a big one but not as big as cluster one and cluster two.

Having the third highest average Euclidian distance, this cluster is wider than other four clusters but narrower than two clusters.

Variables indicating the distance between the cluster center and center of all clusters are in the fourth position among all clusters. Therefore, cases in this cluster can be evaluated as ones that are not so far away from the center of all clusters.

- Characteristics Related to Continuous Variables

Table 111 contains information needed to make interpretations related to continuous variables representing the cluster center.

Most crowded cities are partitioned into cluster seven by the system. In addition, cities in this cluster have the greatest “Count of Customers”. As a result despite the cluster is in the first rank with respect to “Sales per Customer”, the “Average Sales” and “Average Frequency” are not so high compared to other clusters. With an aim to see whether cities in this cluster are valuable in terms of “Total Sales”, this variable is calculated by multiplying “Average Sales” by the “Count of Customer”. ANOVA results show that at the city level, cluster seven has significantly higher “Sales per Customer” value compared to all other clusters. Analyzing the information shown in Table 111, it is concluded that cities in cluster seven are the most valuable ones for the company and named as Stars.

Table 111 Cluster Seven Cluster Center Values and Significance Values between the Means of Clusters

Cluster 7 - Stars						*Significance Values between Clusters*					
Variables	*Value of Cluster Center*	*Mean For General Dataset*	*Max Value of Cluster Centers*	*Min Value of Cluster Centers*	*Range between the clusters*	*Cluster 7-1*	*Cluster 7-2*	*Cluster 7-3*	*Cluster 7-4*	Cluster 7-5	Cluster 7-6
						p value	p value	p value	p value	p value	p value
Avg Sales for City	149.06	181.41	715.95	142.09	5	1.000	1.000	0.410	1.000	1.000	0.984
Average IPT for City	10.69	11.33	23.16	7.86	4	0.253	0.870	1.000	1.000	0.954	1.000
Count of Customers in the City	4152.89	743.32	4152.89	3.50	1	0.519	0.458	0.372	0.361	0.376	0.380
Average frequency for City	64.11	57.41	81.75	15.40	3	0.161	0.159	0.515	0.202	0.100	0.999
Average Frequency last year for City	48.80	50.43	71.01	17.58	3	0.000	0.589	0.992	0.630	0.430	0.981
Average recency for City	15.22	17.24	56.42	5.94	3	0.055	0.803	0.869	0.105	0.077	0.201
Population for City (2000)	2506369.22	852692.71	2506369.22	162718.50	1	0.922	0.916	0.860	0.989	0.702	0.959
Sales per Customer in the City	16.65	4.44	16.65	0.01	1	0.000	0.000	0.000	0.000	0.000	0.000
Total sales	49045634.80		49045634.80	8468.40	1	0.888	0.852	0.838	0.822	0.825	0.840
Total frequency	289202.78		289202.78	42.00	1	0.809	0.726	0.666	0.660	0.664	0.685

CHAPTER 8

REPORTING ENVIRONMENT DEVELOPMENT

OLAP Technology for Data Mining

On-line Analytical Processing (OLAP) tools are used for the interactive analysis of multidimensional data of varied granularities which facilitates effective data mining. Furthermore many other data mining techniques such as classification, prediction and clustering can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction (Han, Kamber, 2001).

OLAP tools view data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. When the word cube is heard it is commonly thought as a three-dimensional structure; however the cubes used in data mining analysis are constructed by n-dimensions. The data is stored as dimensions and facts in the cube instead of the rows and columns in relational data model. Facts or measures are numeric or factual data that represent a specific business activity (Samtani et al., 1998). Facts can be defined as the measures the analyzer tries to see the effects of dimensions on. Some examples of facts may be the total sales of a retailer company as well as the number of transactions that are made in a market. Dimensions on the other hand, are the perspectives or entities with respect to which an organization wants to keep records. Each dimension in a multidimensional model is constituted by a set of attributes. The attributes correspond to the columns in traditional databases. These attributes are selected by the user when building the cube according to a hierarchy. This hierarchical manner allows users to make detailed analyzes at different hierarchy levels. An illustration of a cube and hierarchical summarization of its dimensions are shown in Figure 127.

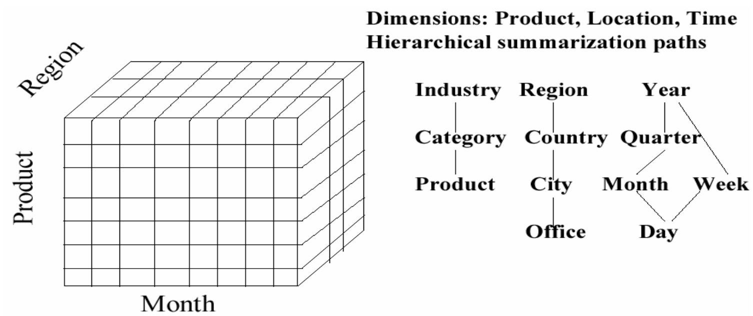


Figure 127 Multi Dimensional data

There are some common operations that are used to gain detailed information from the cubes in an effective and efficient manner. Some of these operations (Han, Kamber, 2001) are:

- *Roll up:* The roll up operation performs aggregation on a data cube either by climbing up a concept hierarchy for a dimension or by dimension reduction
- *Drill down:* It is the reverse of roll up. It navigates from less detailed data to more detailed data.
- *Slice and dice:* This operation performs a selection on one dimension of a given cube.

Multidimensional models can exist in form of star schema, snowflake schema and fact constellation schema.

- *Star Schema:* In this type, the data warehouse (database) contains a fact table and a set of dimension tables. In this type for every dimension only one dimension table is stored in the database. The fact table includes foreign keys that correspond to the primary keys of each of the dimension tables. Moreover the facts or measures are placed in the fact table. An example star schema is shown in Figure 128 (Dayal, Chaudhuri, 1997).

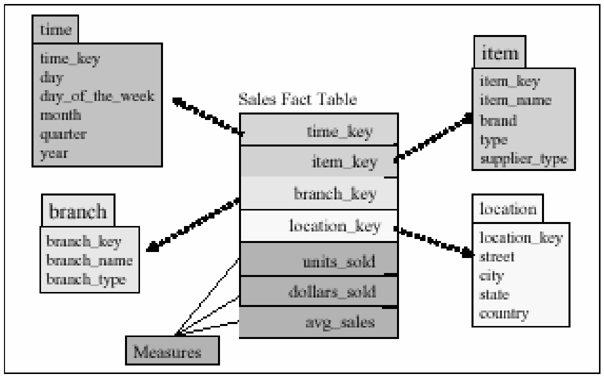


Figure 128 Star schema of a data warehouse.

- *Snowflake schema:* The snowflake schema is a type of star schema, a model in which some dimension tables are normalized by further splitting the data into additional tables. An example snowflake schema is shown in Figure 129 (Han, Kamber, 2001).

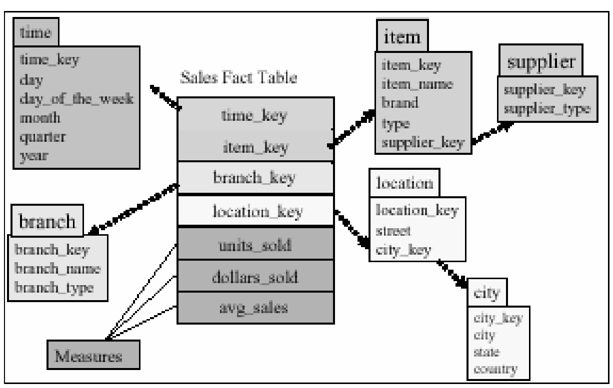


Figure 129 Snowflake schema of a data warehouse.

- *Fact constellation schema:* This schema is used for sophisticated applications that require multiple fact tables. The fact tables share the dimension tables. An example of fact constellation schema is shown in Figure 130.

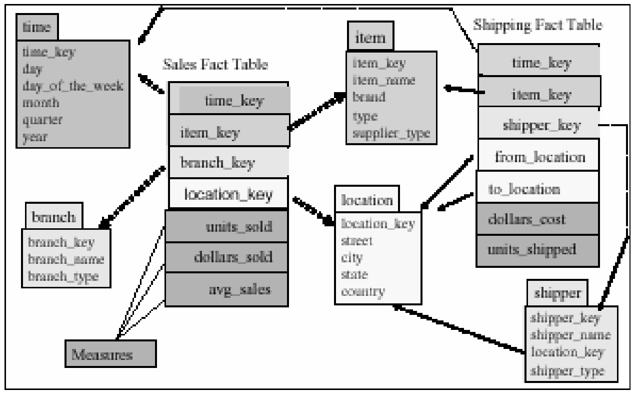


Figure 130 Fact constellation schema of a data warehouse.

Cube Design for Reporting Environment

In the results of data mining technique that is applied in this study, clustering is integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. Daily sales transaction data and customer master data for the customers used in segmentation and profiling analyses are modeled as a data cube with dimensions and facts. Multidimensional model is designed in form of snowflake schema which contains a fact table and a set of dimension tables.

With the analyses in this study effects of dimensions on the "Sales Amount" figure is being analyzed. As a result of this "Sales Amount" is placed in the fact table as the fact of the analyses. "Sales Amount" is analyzed with respect to product that is sold, customer who bought the product, as well as the time the product is sold. Based on these needs, dimensions of the cube are defined as Customer, Product and Time. These dimensions are split into additional tables to reduce redundancies such as volume of product, brand of product, sales region and cities in which the customer lives, as well as the customer and city segments the specified customer is assigned to at the end of the segmentation analyses achieved in this study. Designed snowflake schema with its dimensions is illustrated in Figure 131.

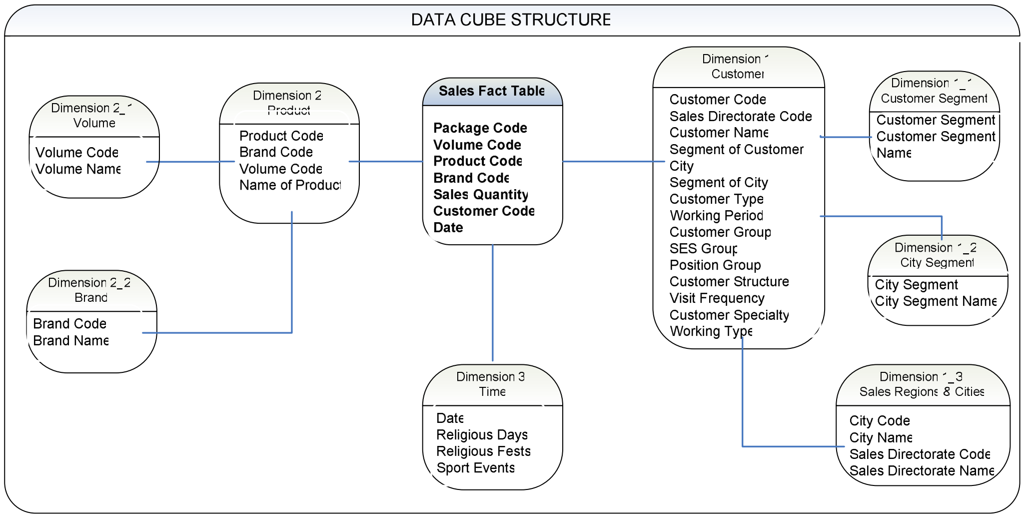


Figure 131 Illustration of snowflake schema

As shown in Figure 131 each dimension in the model is constituted by a set of attributes. Attributes included in the analyses are listed and described in Table 112.

Table 112 Attributes Included in the Analyses

Dimension	*Table*	*Field*	*Description*
	Sales Fact	Product Code	Specifies the product
		Package Code	Specifies the package of the product
		Volume Code	Specifies the volume of the product
		Brand Code	Specifies the brand of the product
		Customer Code	Specifies the customer who bought specified product of company in the sales transaction at issue.
		Sales Quantity	Specifies how much the product is sold in the specified sales transaction
		Date	Specifies the date on which the sales transaction is executed.
Customer	Customer	Categorical Variables selected for the analysis	For explanations please refer to Table 3.
	Customer Segment	Customer Segment	Specifies the customer segment the customer is assigned at the end of the segmentation analysis
		Customer Segment Name	Specifies the name of the customer segment at issue.
	City Segment	City Segment	Specifies the city segment the city in which customer located is assigned at the end of the segmentation analysis
		City Segment Name	Specifies the name of the city segment at issue.
	Sales Regions and Cities	City Code	Specifies the city in which customer is located

Dimension	*Table*	*Field*	*Description*
		City Name	Specifies the name of the city
		Sales Directorate Code	Specifies the sales directorate the city in which customer located is exists in.
		Sales Directorate Name	Specifies the name of the sales directorate.
Product	Prodcut	Product Code	Specifies the product
		Product Name	Specifies the name of the product
		Volume Code	Specifies the volume of the product
		Brand Code	Specifies the brand of the product
	Volume	Volume Code	Specifies the volume of the product
		Volume Name	Specifies the name of the volume
	Brand	Brand Code	Specifies the brand of the product
		Brand Name	Specifies the name of the brand
Time	Time	Date	Specifies the date on which the sales transaction is executed.
		Religous days	Shows whether the date at issue is religious or not.
		Religious fests	Shows whether the date at issue is religious fest or not.
		Sport events	Shows whether there is a sportive activity on the day or not.

With the aim of making different analyses at different hierarchy levels these attributes are selected according to a hierarchy if their structure is applicable for it. The hierarchical summarizations of the dimensions are shown in Table 113.

Table 113 Conceptual Hierarchies of Dimensions

Conceptual Hierarchies of Dimensions					
Product	*Time*	*Customer_1*	*Customer_2*	*Customer_3*	*Customer_4*
Brand	Year	Sales Directorate	Segment of City	Segment of Customer	Customer Type
Volume	Quarter	City	City	Customer Name	Customer Group
Product	Month	City name	Region		Customer name
	Day		Position Group		
			Customer Name		

“Time” and “Customer” dimensions have some attributes which cannot be a part of conceptual hierarchies but should be used as dimension during the analysis. Virtual dimensions are created with these attributes in order to be able to use them in the analysis. Virtual dimensions of the analysis are listed in Table 114.

Table 114 Virtual Dimensions

Virtual Dimensions	
Time	*Customer*
Religious Days	Customer Type
Religious Fests	Working Period
Sports Events	Customer Group
	SES Group
	Position Group
	Customer Structure
	Visit Frequency
	Customer Specialty
	Working Type

Reports for Creating Base for CRM Activities

In this section some sample reports are examined for creating a base for CRM activities. All reports are developed on the base of information coming from the analyses made for comparisons of sales trends for years 2002, 2003 and 2004. This analysis is achieved with the OLAP cube designed for this study. The effects of values of dimensions are analyzed with these reports. Information obtained with these reports can be used for CRM activities of the case company. Additionally, some other reports can be developed with the OLAP cube described in the preceding section.

Report One

In order to create a base for the CRM activities general characteristics of the "Star Customers" in "Star City Segment" who are purchasing "Extra Brand" are analyzed by the help of newly created reporting environment. These specifications are selected for the report because when the general sales report obtained from OLAP cube for all customers and all brands is analyzed it is realized that for the customers with these specifications the purchase amount of "Extra Brand" is decreased. Figure 132 shows the general Sales Report prepared from the OLAP cube. “Report One” is developed with the aim of finding some information that can be useful to investigate the reasons of this decrease.

Dimensions used in “General Sales Comparison Report” are listed in Table 115 with the operations used to get detailed information from the cube in an effective and efficient manner.

Table 115 Dimensions and Operations of General Sales Report

Dimension	*Operation*	*Detail Level*
Customer Segment Name	Drill down	Customer Segment Name
City Segment Name	Drill down	City Segment name
Brand	Slice	Extra, Brand_3, Normal
Time	Drill Down and Slice	Quarter - Quarter 3

Dimensions used in “Report One: Sales comparison for ‘Closed’ type customers who are selling ‘Extra’ are listed in Table 116 with the operations used to get detailed information from the cube in an effective and efficient manner.

Table 116 Dimensions and Operations of Report One

Dimension	*Operation*	*Detail Level*
Customer Segment Name	Slice	Stars
City Segment Name	Slice	Stars
Customer Type	Slice	Closed
Customer Group	Drill Down	
Customer Specialty	Drill Down	
Position Group	Drill Down	
Time	Drill Down and Slice	Quarter 3.

Figure 133 shows the report created with these specifications. The last column in the report shows the percentage of change between year 2003 and 2004 with respect to “Brand Extra” sales. The positive figures show that there is a decrease in the sales of “Brand Extra” in year 2004 compared to year 2003.

GENERAL SALES COMPARISON REPORT										
		Brand_3			Extra			Normal		
		2002	2003	2004	2002	2003	2004	2002	2003	2004
City Segment Name	Customer Segment Name	Quarter 3	Quarter 3	Quarter 3	Quarter 3	Quarter 3	Quarter 3	Quarter 3	Quarter 3	Quarter 3
Most Valuable Cities	Potential Valuable Customers	0	820.2	2220	0	3712	6848	0	87992.08	290742
	Stars	0	450	759	0	1980	2100	0	95310.6	96122
	Valuable Customers	0	481.2	285	0	1006	1526	0	66962.52	148346.4
Most Valuable Cities Total		0	1751.4	3264	0	6698	10474	0	250265.2	535210.4
Stars	Potential Valuable Customers	2052	15329.57	75956.8	3240	38796	240494	205850	1530597	7928980
	Stars	17736	44662.2	19328	6408	215270	102348	661256	6512945	2789277
	Valuable Customers	170521.2	222499.6	131312	52464.16	392192	359138	4888337	15881546	16249409
Stars Total		190309.2	282491.4	226596.8	62112.16	646258	701980	5755443	23925088	26967667
Valuable Cities	Potential Valuable Customers	0	0	96	0	12	210	0	312	32370
Valuable Cities Total		0	0	96	0	12	210	0	312	32370

Figure 132 General sales comparison report

REPORT ONE: SALES COMPARISON FOR "CLOSED" TYPE CUSTOMERS and EXTRA BRAND									
Sales Quantity						Brand Name	Quarter		
						Extra			
						2002	2003	2004	
City Segment Name	Customer Segment Name	Customer Type	Customer Group	Customer Specialty	Position Group	Quarter 3	Quarter 3	Quarter 3	Decrease Percentage
Stars	Stars	Closed	Buffet			60	24	156	
				All Brands	Mid Street	0	8716	1020	88.30
				Company Brands		0	18540	17472	5.76
					Main Street	0	22092	8844	59.97
					Mid Street	0	51714	45736	11.56
					Parallel Street	0	19182	0	100.00
			Buffet Total			60	120268	73228	39.11
			Grocery			0	24	24	0.00
				All Brands	Mid Street	1800	25504	828	96.75
					Parallel Street	3708	360	0	100.00
				Company Brands		0	720	1260	-75.00
					Main Street	0	33462	10284	69.27
					Mid Street	0	25428	5232	79.42
					Parallel Street	0	2220	3360	-51.35
			Grocery Total			5508	87718	20988	76.07
			Market	Company Brands	Mid Street	0	5028	3792	24.58
					Parallel Street	0	1056	1680	-59.09
					Shopping Center	0	0	960	
			Market Total			0	6084	6432	-5.72
Stars Total						5568	214070	100648	52.98

Figure 133 Report one: Sales comparison for "Closed" type customers who are selling "Extra"

Report shows that when we compare 2004 to 2003 the biggest decline is observed for grocery. For grocery customers located in almost all position groups purchase Amount of Extra Brand decrease with some expectations. It is interesting that grocery type customers located in the parallel street are lost if they are also working with other companies. On the other hand if they are working only with the case company the purchase amount of the customer is increased. Additional to this, the purchase amount of grocery customers located in mid street also decreases if they are also working with competitors more than the ones who are working only with the case company. The results show that For Brand Extra compared to the same period of the previous year, competitor threat effected the sales for grocery type customers who are selling products closed more than other ones. The information gained from the new reporting environment can be used for CRM activities targeted customers with these specifications.

Report Two

In order to create a base for the CRM activities effect of religious days on the sales amount of “Closed” type customers for “Brand 4” products is analyzed with respect to all city and customer segments by the help of a newly created reporting environment. These specifications are selected for the report by analyzing the effect of religious days on all types of customers. When the report shown in Figure 134 is analyzed it is realized that “Closed” type customers are the ones who are mostly effected by the religious days. Report shows that products of “Mixed” and “Brand 4” brands are the most affected ones among all brands. However since the “Mixed” brand is not a commonly produced one it is discarded from the analysis and “Brand 4” is used for the rest of the analysis. “Report Two” is developed with the aim of finding the City and Customer segments which are affected by the religious days mostly for "Closed" type customers who are selling "Brand 4".

Dimensions used in "Affect of Religious days- General Sales Comparison Report" are listed in Table 117 with the operations used to get detailed information from the cube in an effective and efficient manner.

Table 117 Dimensions and Operations of Affect of Religious Days-General Sales Comparison Report

Dimension	*Operation*	*Detail Level*
Customer Type	Drill down	
Brand	Drill down	Brand Name
Religious Days	Drill Down	

Dimensions used in "Report Two" are listed in Table 118 with the operations used to gain detailed information from the cube in an effective and efficient manner.

Table 118 Dimensions and Operations of Report Two

Dimension	*Operation*	*Detail Level*
Customer Segment Name	Drill Down	Customer Segment Name
City Segment Name	Drill Down	City Segment Name
Customer Type	Slice	Closed
Brand	Slice	Brand 4
Religious Days	Drill Down	

Figure 135 shows the report created with these specifications. The last column in the report shows the percentage of change between religious and non religious days with respect to "Brand 4" sales in "Closed" type customers. The positive figures show that there is a decrease in the sales of "Brand 4" in religious days compared non-religious ones.

Affect of Religious Days - General Sales Comparison Report					
Customer Type	*Brand Name*	*Non-Religious Day*	*Religious Day*	*Grand Total*	*Percentage of Change*
Barrels	Brand_1	992353.28	49492.08	1041845.36	95.01 %
	Brand_2	332090.08	16165.68	348255.76	95.1 %
	Brand_3	6278	430.2	6708.2	93.1 %
	Brand_4	1210	0	1210	100 %
	Dark	165692.27	9849.6	175541.87	94.05 %
	Extra	46725.52	1296	48021.52	97.22 %
	Light	196748.16	11111.04	207859.2	94.35 %
	Non-Alcohol	12801.12	3614.16	16415.28	71.76 %
	Normal	92537779.5	3920589.75	96458369.25	95.76 %
Barrels Total		94291677.93	4012548.51	98304226.44	95.74 %
Closed	Brand_1	7839263.13	264192	8103455.13	96.62 %
	Brand_2	851328.75	27409.68	878738.43	96.78 %
	Brand_3	7377350.69	260269.2	7637619.89	96.47 %
	Brand_4	295288	5980	301268	97.97 %
	Dark	1347714.12	54857.76	1402571.88	95.92 %
	Extra	17312865.52	840444	18153309.52	95.14 %
	Light	656502.24	25755.72	682257.96	96.07 %
	Mixed	36	0	36	100 %
	Non-Alcohol	154466.64	36836.16	191302.8	76.15 %
	Normal	351336793.3	12013454.13	363350247.4	96.58 %
Closed Total		387171608.4	13529198.65	400700807	96.50 %
Open	Brand_1	547102.02	25123.44	572225.46	95.40 %
	Brand_2	285192.56	10559.52	295752.08	96.29 %
	Brand_3	35062.58	1750.8	36813.38	95.00 %
	Brand_4	1308	30	1338	97.70 %
	Dark	79053.5	5057.52	84111.02	93.60 %
	Extra	86282	5162	91444	94.01 %
	Less Alcohol	15.84	0	15.84	100 %
	Light	143568.23	7378.31	150946.54	94.86 %
	Non-Alcohol	10152.72	3745.92	13898.64	63.10 %
	Normal	30506557.28	1248569.27	31755126.55	95.90 %
Open Total		31694294.73	1307376.78	33001671.51	95.87 %

Figure 134 Affect of religious days – General sales comparison report

REPORT TWO: AFFECT OF RELIGIOUS DAYS IN CITY AND CUSTOMER SEGMENTS DETAIL				
City Segment Name	*Customer Segment name*	*Non-Religious Days*	*Religious Days*	*Percentage of Change*
Fit Class	Average Customers	84	0	100%
	Potential Customers	48	0	100%
Fit Class Total		132	0	100%
Most Valuable Cities	Average Customers	11114	540	95.14%
	Frequently Buyers	504	0	100%
	Potential Customers	438	0	100%
	Potential Invaluable Customers	132	0	100%
	Potential Valuable Customers	1472	0	100%
	Valuable Customers	624	0	100%
Most Valuable Cities Total		14284	540	96.22%
Stars	Average Customers	106050	1092	98.97%
	Frequently Buyers	102082	2204	97.84%
	Potential Customers	15358	238	98.45%
	Potential Invaluable Customers	3534	36	98.98%
	Potential Valuable Customers	23066	424	98.16%
	Stars	2760	0	100%
	Valuable Customers	27050	1446	94.65%
Stars Total		279900	5440	98.06%
Valuable Cities	Average Customers	708	0	100%
	Frequently Buyers	12	0	100%
	Potential Customers	60	0	100%
	Potential Invaluable Customers	24	0	100%
	Potential Valuable Customers	168	0	100%
Valuable Cities Total		972	0	100%

Figure 135 Report Two: Affect of religious days in city and customer segments detail

Report shows that when we compare religious days to non- religious ones for "Closed" type customers who are selling "Brand 4", it is realized that in every city segment and customer segment there is a big decline in the sales of products. The percentage of decrease is almost 100% for all city and customer segments. The information gained from this report makes clear that religious days affect the sales amount of "Brand 4".

Report Three

In order to create a base for the CRM activities, characteristics of customers whose buying pattern is different for winter and summer periods of years 2003 and 2004 are analyzed. Different from the pervious reports as a part of "Report Three" another report has been created for more specific analysis on the customer base with the help of newly created reporting environment. "Report Three" shows the buying patterns of customers in all city segments for quarter one and quarter three periods of year 2003 and year 2004. Figure 136 represents the "Sales Comparison between Summer and Winter Periods" report. The last column in the report shows the percentage of change between winter and summer periods. The positive figures show that there is a decrease in the sales for summer period compared to winter one.

Dimensions used in "Sales Comparison between Summer and Winter Periods Report" are listed in Table 119 with the operations used to gain detailed information from the cube in an effective and efficient manner.

Table 119 Dimensions and Operations of Affect of Religious Days-General Sales Comparison Report

Dimension	*Operation*	*Detail Level*
Customer Type	Drill down	
Working Period	Drill down	
City Segment Name	Drill Down	City Segment Name
Time	Slice	2003-2004, Quarter One and Quarter Three

SALES COMPARISON BETWEEN WINTER AND SUMMER PERIODS								
Customer Type	Working Period	City Segment Name	Quarter 1	Quarter 3	Percentage of Decrease	Quarter 1	Quarter 3	Percentage of Decrease
Barrels	Seasonal	Most Valuable Cities	0	0		650	0	100.00
		Stars	2947.28	40108.48	-1260.86	13176.4	99743.36	-656.98
		Valuable Cities	0	0		36	654	-1716.67
	Standard	Most Valuable Cities	0	208228.88		295625	369151.32	-24.87
		Stars	7545249.04	11062446.42	-46.61	10838764.34	12909924.3	-19.11
		Valuable Cities	0	0		432	2888	-568.52
Barrels Total			7548196.32	11310783.78	-49.85	11148683.74	13382360.98	-20.04
Closed	Seasonal	Most Valuable Cities	0	708		24607.2	63751.8	-159.08
		Stars	10439.76	51450.67	-392.83	65925.96	204050.28	-209.51
		Valuable Cities	0	0		1230	2335.44	-89.87
	Standard	Most Valuable Cities	0	1039389.76		1154678.8	1585715.48	-37.33
		Stars	21568794.24	52304838.84	-142.50	43303374.22	66090636.3	-52.62
		Valuable Cities	0	21643.56		123952.28	154071.88	-24.30
Closed Total			21579234	53418030.83	-147.54	44673768.46	68100561.18	-52.44
Open	Seasonal	Most Valuable Cities	0	0		700	3792	-441.71
		Stars	46250.2	110968.64	-139.93	75054.44	156856.6	-108.99
		Valuable Cities	0	0		5677.2	14414.88	-153.91
	Standard	Most Valuable Cities	0	216154.16		239087.2	297783.23	-24.55
		Stars	2132169.08	3320675.88	-55.74	3418391.13	4782240.7	-39.90
		Valuable Cities	0	5896		25758.56	23934.8	7.08
Open Total			2178419.28	3653694.68	-67.72	3764668.53	5279022.21	-40.23
Grand Total			31305849.6	68382509.29	-118.43	59587120.73	86761944.37	-45.61

Figure 136 Sales Comparison between summer and winter periods

"Sales Comparison between Summer and Winter Periods Report" shows that in every city segment, the greatest changes of sales amounts between the winter and summer periods have occurred for the customers who are working only in "Seasonal" periods. By analyzing the report it is not surprising that the greatest change occurs for the customers in the "Stars" city segment which contains cities like Antalya and Izmir. Surprisingly in the "Most Valuable Cities" segment "Barrels" type of customers who are working seasonally are completely lost in the summer of 2004.

Another report has been created for more detailed analysis of these lost customers, based on the result of previous report. Figure 136 represents this report named as "Detailed Analysis of Sales Decrease". In reality it is not possible to satisfy all customers' specific needs with the limited sources of company; however with this report it is shown that if the company needs to understand the specific reasons that create the abnormality in the sales patterns, the newly developed environment let them to do it by showing details like the ones in this report. Dimensions used in "Detailed Analysis of Sales Decrease Report" are listed in Table 120 with the operations used to get detailed information from the cube in an effective and efficient manner.

Table 120 Dimensions and Operations of Affect of Religious Days-General Sales Comparison Report

Dimension	*Operation*	*Detail Level*
Customer Segment Name	Drill Down	Customer Name
City segment Name	Slice / Drill Down	Most Valuable Cities/ City Name
Time	Drill Down	Quarter

DETAILED ANALYSIS OF DECREASE IN SALES						
City Segment Name	*City Name*	*Customer Segment Name*	*Customer Name*	*Quarter 1*	*Quarter 2*	*Quarter 4*
Most Valuable Cities	Adana	Potential Valuable Customers	A.E.O Garden Restaurant	650	100	200
		Potential Valuable Customers Total		650	100	200
	Adana Total			650	100	200
Most Valuable Cities Total				650	100	200
Grand Total				650	100	200

Figure 137 Detailed analysis of decrease in sales

Based on the information gained from "Detailed Analysis of Decrease in Sales Report", illustrated in Figure 137, it is clear that he customer which seems lost in the previous report is a potentially valuable one and working in the seasonal period. But interestingly in the summer period it does not buy any products from the company although it eventually starts to buy some amounts. By analyzing this report the company may develop CRM strategies to make this customer a valuable one.

CHAPTER 9

CONCLUSION

Increasing competition in FMCG sector forces the companies to be careful about customer relationships to maintain their market share against potential competitors and increase their long term profitability. The present study aims to create a base for possible CRM activities of an FMCG company that performs in a B2B2C type market. Two segmentation and profiling analyses are employed to partition the company's customers and the cities that the customers are embedded in into small manageable groups for future CRM activities. Additionally, a reporting base has been developed using the information gained from these analyses with an intention to constitute a base for possible CRM activities of the case company.

The methodology used in this study is a combination of Two Crow and Crisp DM methodologies explained in Chapter 3. At the end of the data preparation phase, two different datasets are constructed using the variables proposed in the literature for segmentation and profiling analyses of the customers and cities.

In the modeling phase, clustering technique of data mining is employed with a nonhierarchical clustering technique: k-means. Clustering technique requires determining the variables that will be used to partition the objects into small groups beforehand. In order to determine which variables will be used in segmentation analyses among the ones prepared at the end of the data preparation step, factor analysis is applied to the datasets. At the end of the factor analysis, twenty seven variables in the customer dataset are loaded on five factors and "Recency", "Frequency", "Total Amount", "LoR" and "rMajorTrip" are selected as surrogate variables for customer segmentation analysis. In city dataset, seven variables are loaded to three factors and “Average Recency for City”, “Average Frequency for City”, “Average Sales for City”, “Count of Customers” and “Sales per Customer” are selected as surrogate variables for city segmentation analysis.

At the end of the segmentation analysis considering the buying behavior of company’s customers, eight different customer segments are constructed. In addition to customer segments, cities in which the company performs are partitioned into seven different segments. The results of validation analyses exploited to verify the results of segmentation analyses showed that segments are composed of customers and cities which manifest similar buying behavior. Analyses show that although “Recency”, “Frequency” and “Monetary” variables are enough to adequately partition the customers and cities into smaller groups, including other surrogate variables such as “LoR” and “Sales per Customer” makes the profiling process of segments easier and helps to create more manageable segments for CRM activities.

Segments are profiled using three different perspectives: general characteristics of segments, characteristics related to continuous variables and characteristics related to categorical variables. On the other hand, for city segments only the first two of the above perspectives are used. Results of profiling analyses show that variables related to purchasing amount, namely “Total Amount”, “Amount”, “rAmount”, “rTotal Amount” are the ones that generally differentiate the customer segments from each other. In addition to these variables, “Sales per Customer” variable has a powerful distinguishing effect for the city segments.

In the last part of the study, the results obtained from segmentation and profiling analyses are integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. A data cube is created from the daily transactions data of customers including product, customer and time information. The customer and city segments obtained at the end of the segmentation and profiling analyses are also included in the data cube. Three different scenarios are created by using the analysis functionalities of this cube. The information gained from these reports can be used as a base for CRM activities or new reports can be created using this reporting environment.

All of the analyses in this study are done manually after obtaining the clustering results from SPSS. Future work might be in developing an automatic cluster detection tool that will be able to perform the analyses achieved manually in this study such as validation of cluster analysis. In a similar vein, in the profiling part, all of the comparisons between the clusters are performed manually. Other future work might be in developing a segment profiling tool that will be able to compare the characteristics of the produced clusters and summarize the characteristics that define each of them.

In summary, increasing competition forces all companies to improve their relationship with their customers in order to increase their long term profit. This requires detecting valuable customers and retaining them instead of acquiring new ones. Once the valuable customers are determined and armed with this information, companies can target retention offers for predefined customer segments. Data mining functionalities such as clustering and profiling can be used to detect these valuable customers. If the results of these functionalities are combined with a reporting environment that allows multiple level analyses, an effective base for CRM studies can be developed.

REFERENCES

Arnold, S.J. (1979). A test for clusters. *JMR, Journal of Marketing Research (pre-1986)*, 545.

Athanassopoulos, A.D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. Journal of Business Research, 191.

Bauer, H.H. & Hammerschmidt, M. & Braehler, M. (2003). The customer lifetime value concept and its contribution to corporate valuation. Yearbook of Marketing and Consumer Research.

Bayon, G. & Gutsche, J. & Bauer, H. (2002). Customer equity marketing: Touching the intangible. European Management Journal, 213.

Berger, B.D. & Weinberg, B. & Hanna, R.C. (2003). Customer lifetime value determination and strategic implications for a cruise-ship company. Journal of Database
Marketing & Customer Strategy Management, 40.

Berger, P.D. & Bechwati, N.N. (2001). The allocation of promotion budget to maximize customer equity. Omega 50 29, 49.

Berger, P.D.& Bolton, R.N. & Bowman, D. & Briggs, E. et al. (2002). Marketing actions and the value of customer assets: A framework for customer customer asset management. Journal of Service Research : JSR, 39.

Berry, M. J. A. & Linoff, G. S. (2004). Data mining techniques for marketing, sales, and custoer relationship management. Indiana: Wiley Publishing.

Berson, A.& Smith, S. & Thearling, K. (2000). Building data mining applications for CRM. New York: McGraw-Hill.

Bloemer, J.M.M. et al. (2003). Comparing complete and partial classification for identifying customers at risk. Intern. J. of Research in Marketing, 117.

Bolton, R.N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. Marketing Science (1986-1998) ABI/INFORM Global, 45.

Bradshaw, D. & Brash, C. (2001). Managing Customer relationships in the e-business world: How to personalise computer relationships for increased profitability. International Journal of Retail & Distribution Management, 520.

Buckinx, W. & Van den Poel, D. (2004). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research.

Bult, J.R. & Wansbeek, T.J. (1995). Optimal selection for direct mail. Marketing Science, 378

Changchien, S.W. & Lu,T.C.(2001). Mining association rules procedure to support on-line recommendation by customers and products fragmentation. Expert Systems with Applications, 325.

Chen, J. & Ching, R.K.H. (2004). An empirical study of the relationship of IT intensity and organizational absorptive capacity on CRM performance. Journal of Global Information Management, 1.

Chen, J-S. & Ching, R.KH. & Lin, Y-S. (2004). Extended study of the k-means algorithm for data clustering and its applications. Journal of the Operational Research Society, 976.

Chiang, D.A. et al. (2003). Goal-oriented sequential pattern for network banking churn analysis. Expert Systems with Applications, 293.

Crie, D. (2004). Loyalty-generating products and the new marketing paradigm. Journal of Targeting, Measurement and Analysis for Marketing, 242.

Crosby, L.A. & Johnson, S.L. (2002). CRM and management. Marketing Management, 10.

Drew, J.H. & Mani, D.R. & Betz, A.L. & Datta, P. (2001). Targeting customers with statistical and data-mining techniques. Journal of Service Research : JSR, 205.

Edelstein, H. (2000). Building profitable customer relationships with data mining. U.S.A.:SPSS

Ehret, M. (2004). Managing the trade-off between relationships and value networks.Towards a value-based approach of customer relationship management in business-to-business markets. Industrial Marketing Management, 465.

Fox, T. & Stead, S. (2001). Customer relationship management: Delivering the benefits. Customer Relationship Management (UK) Ltd.

Friesner et al. (2004). Identifying Latent Outcome Measures in Inpatient Physical Therapy. 2004 Midwest Business Economics Association Conference Proceedings.

Geppert, K. (2002). Customer churn management: Retaining high-margin customers with customer relationship management techniques. KPMG.

Greenberg, P. (2001). CRM at the speed of light. Capturing and keeping customers in internet realtime. Osborne: McGraw-Hill.

Ha, S. H. et al. (2002). Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. Computers & Industrial Engineering, 801.

Hair, J. F. et al. (1995). Multivariate data analysis with readings. New Jersey: Prentice Hall.

Hamerly, G. & Elkan, C. Learning the k in k-means. Department of Computer Science and Engineering University of California, San Diego.

Han, J. & Kamber, M. (2001). Data mining: Concepts and techniques. San Diego: Academic Press.

Harrison-Walker, L.J. & Neeley, S.E. (2004). Customer relationship building on the internet in b2b marketing: a proposed typology. Journal of Marketing Theory and Practice, 19.

He, Z. et al. (2004). Mining class outliers: concepts, algorithms and applications in CRM. Expert Systems with Applications, 681.

Hogan, J.E. & Lemon, K.N. & Libai, B. (2003). What is the true value of a lost customer? Journal of Service Research : JSR, 196.

Hogan, J.E. et al. (2002). Linking customer assets to financial performance. Journal of Service Research : JSR, 26.

Hsieh, N.C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Systems with Applications, 623.

Hwang, H. & Jung, T. & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications, 181.

Jonkera, J.J. & Piersmab, N. & Van den Poel, D. (2004). Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert Systems with Applications, 159.

Kim, S.Y. et al. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications, 101.

Klastorin, T.D. (1983). Assessing cluster analysis results. JMR, Journal of Marketing Research, 92.

Krieger, A.M. & Green, P.E. (1996). Modifying cluster-based segments to enhance agreement with an exogenous response variable. JMR, Journal of Marketing Research, 351.

Ledakis, G. (1999). Factor Analytic Models of the Mattis Dementia Rating Scale in Dementia of the Alzheimer's Type and Vascular Dementia Patients. Doctoral Dissertation, Drexel University. http://schatz.sju.edu/multivar/factor.html

Lee, G. & Morrison, A.M. & O'Leary, J.T. (2006).The economic value portfolio matrix: A target market selection tool for destination marketing organizations. Tourism Management, 576.

Lejeune, M.A.P.M. (2001). Measuring the impact of data mining on churn management. Internet Research: Academic Research Library, 375.

Liao, S.H. & Chen, Y.J. (2004). Mining customer knowledge for electronic catalog marketing. Expert Systems with Applications.

Libai, B. & Narayandas, B. & Humby, C.(2002).Toward and individual customer profitability model: A segment-based approach. Journal of Service Research : JSR, 69.

Liu, D.R. & Shih, Y.Y. (2004). Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management.

Liu, D.R. & Shih, Y.Y. (2004). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. The Journal of Systems and Software.

Lockshin, L.S. & Spawton, A.L. & Macintosh, G. (1997). Using product, brand and purchasing involvement for retail segmentation. Journal of Retailing and Consumer Services, 171.

Mittal, V. & Pankaj & Tsiros, M. (1999). Attribute-level performance satisfaction, and behavioral intentions over time: A consumption-system approach. Journal of Marketing, 88.

Morwitz, V.G. & Schmittlein, D. (1992). Using segmentation to improve sales forecasts based on purchase intent: Which "Intenders" actually buy? JMR, Journal of Marketing Research, 391.

Musaoğlu, C. (2003). Customer acquisition and retention modeling in consumer finance sector using data mining. Thesis Study. Istanbul: Boğazici Press.

Palmer, R.A. & Millier, P. (2004). Segmentation: Identification, intuition, and implementation. Industrial Marketing Management, 779.

Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research, 483.

Pritscher, L. & Feyen, H. Data mining and strategic marketing in the airline industry. Atraxis AG, Swissair Group, Data Mining and Analysis, CKCB.

Punj, G. & Stewart, D.W. (1983). Cluster analysis in marketing research: Review and suggestions for application. JMR, Journal of Marketing Research, 134.

Ragins, E.J. & Greco, A. J. (2003). Customer relationship management and E-business: More than a software solution. Review of Business, 25.

Reinartz, W.J. & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. Journal of Marketing, 17.

Rowley, J.E. (2002). Reflections on customer knowledge management in e-business.Qualitative Market Research: An International Journal, 268.

Ryals, L. (2002). Are your customers worth more than money? Journal of Retailing and Consumer Services, 241.

Ryals, L. (2003). Creating profitable customers through the magic of data mining. Journal of Targeting, Measurement and Analysis for Marketing, 343.

Rygielski, C. & Wang, J-C. & Yen, D.C. (2002). Data mining techniques for customer relationship management. Technology in Society, 483.

Sam, I. (1994). Using lifetime value analysis for selecting new customers. Credit World, 37.

Shaw, M. (2001). Editorial: Lifetime values and valuing customers - who are you kidding?Journal of Targeting, Measurement and Analysis for Marketing, 101.

SPSS® Base 11.5 User's Guide

Srivastava, J. et al. (2002). A Case for Analytical Customer Relationship Management. M.-S. Chen, P.S. Yu, and B. Liu (Eds.): PAKDD, 14.

Stahl, H.K. & Matzler, K. & Hinterhuber, H.H. (2003). Linking customer lifetime value with shareholder value. Industrial Marketing Management, 267.

Stone, M. et al. (2003). The quality of customer information management in customer life cycle management. Journal of Database Management, 240.

Suha, E.H. & Nohb, K.C. & Suhc, C.K. (1999). Customer list segmentation using the combined response model. Expert Systems with Applications, 89.

Swift, R. S. (2001). Accelerating customer relationships using CRM and relationship techniques. Upper Saddle River: Prentice Hall.

Tan, X. & Yen, D.C. & Fang, X. (2002). Internet integrated customer relationship management: A key success factor for companies in the e-commerce arena. The Journal of Computer Information Systems, 77.

Thearling, K. (2000). An Introduction to Data Mining : Discovering hidden value in your data warehouse. White Paper, URL: http://www.thearling.com/text/dmwhite/dmwite.htm.

Tsai, C.Y. & Chiu, C.C. (2004). A purchase-based market segmentation methodology. Expert Systems with Applications, 265.

Tsai, C.Y. & Chiu, C.C. (2004). A purchase-based market segmentation methodology. Expert Systems with Applications, 265.

Wedel, S.& Kamakura, W. (1997). Market segmentation: Conceptual and methodological foundations. Boston: Kluwer.

Weinstein, A. (2004). Handbook of market segmentation: strategic targeting for business and technology firms. Binghamton, NY: Haworth Press

Wikipedia the free encyclopedia (2006), http://en.wikipedia.org/wiki/Fast_Moving_Consumer_Goods

Withrow, S., Data warehousing and mining basics (2003)
http://www.builder.com/5100-6388-1045046.htm

Yim, C.K. & Kannan, P.K. (1999). Consumer behavioral loyalty: A segmentation model and analysis. Journal of Business Research, 75.

APPENDICES

APPENDIX A

(Data Dictionaries)

Categorical Variables

Variable: Müdürlük -Sales Directorate		
Short Description: Expresses the directorate each customer is bound to.		
Variable Type: Categoric – Nominal		
Data Expression : Numeric		
Can hold null:	*Yes:*	*No:* x
Length : 4		
If Nominal		*If Ordinal*
Value	*Meaning*	
1031	İstanbul Sales Directorate	
1033	Doğu Marmara Sales Directorate	
1034	Trakya Sales Directorate	
1035	Orta Anadolu Sales Directorate	
1037	Güney Marmara Sales Directorate	
1038	Güney Sales Directorate	
1039	Karadeniz Sales Directorate	
1040	Akdeniz Sales Directorate	
1041	Güney Ege Sales Directorate	
If derived		
Calculation		

Variable: Nokta Kodu - Distributor Code		
Short Description: The unique number that is given from the system to each customer.		
Variable Type: Numeric – Discrete		
Data Expression : Numeric		
Can hold null:	*Yes:*	*No:* x
Length : 7		
If Discrete		*If Continous*
Value	*Meaning*	
1……	Traditional Sales Point	
2……	Distributor (Modern Sales Point)	
If derived		
Calculation		

<table>
<tr><td colspan="4">Variable: Nokta Türü - Customer Type</td></tr>
<tr><td colspan="4">Short Description:
The type of the customer which is determined according to the way the customers use when selling the products of the company.</td></tr>
<tr><td colspan="4">Variable Type:
Categoric – Nominal</td></tr>
<tr><td colspan="4">Data Expression : Text</td></tr>
<tr><td colspan="2">Can hold null:</td><td>Yes: x</td><td>No:</td></tr>
<tr><td colspan="4">Length : 6</td></tr>
<tr><td colspan="2">If Nominal</td><td colspan="2">If Ordinal</td></tr>
<tr><td>Value</td><td>Meaning</td><td colspan="2"></td></tr>
<tr><td>Kapalı – Closed</td><td>Customers who sell the products with original packets without any service.</td><td colspan="2"></td></tr>
<tr><td>Açık – Open</td><td>Customers who serve the products in their places with or without original packets.</td><td colspan="2"></td></tr>
<tr><td>Fıçı - Barrels</td><td>Customers who sell draught beer in their places.</td><td colspan="2"></td></tr>
<tr><td colspan="4"></td></tr>
<tr><td colspan="2">If derived</td><td colspan="2"></td></tr>
<tr><td colspan="2">Calculation</td><td colspan="2"></td></tr>
</table>

<table>
<tr><td colspan="4">Variable: Çalışma Dönemi – Working Period</td></tr>
<tr><td colspan="4">Short Description:
Defines the working period of the customer.</td></tr>
<tr><td colspan="4">Variable Type:
Categoric – Nominal</td></tr>
<tr><td colspan="4">Data Expression : Text</td></tr>
<tr><td colspan="2">Can hold null:</td><td>Yes: x</td><td>No:</td></tr>
<tr><td colspan="4">Length : 8</td></tr>
<tr><td colspan="2">If Nominal</td><td colspan="2">If Ordinal</td></tr>
<tr><td>Value</td><td>Meaning</td><td colspan="2"></td></tr>
<tr><td>Standart - Standard</td><td>Customers who works for full year.</td><td colspan="2"></td></tr>
<tr><td>Sezonluk - Seasonal</td><td>Customers who works for some periods – seasons of year.</td><td colspan="2"></td></tr>
<tr><td colspan="4"></td></tr>
<tr><td colspan="2">If derived</td><td colspan="2"></td></tr>
<tr><td colspan="2">Calculation</td><td colspan="2"></td></tr>
</table>

Variable: Müşteri Gurubu – Customer group		
Short Description: The group of customer determined according to the physical and legal structure of their shops.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 20		

If Nominal		*If Ordinal*
Value	*Meaning*	
Bakkal - Grocery	Small grocery stores mostly located around neighborhood.	
Market	Bigger grocery stores.	
Bufe - Buffet	Buffets	
Birahane – Beer House	Beer houses	
BIP Birahane	Specal kind of beer houses that are placed in a special campaign.	
Standard Birahane	Beer Houses	
KeyAccount	Market chains	
Lokanta - Restaurant	Restaurants	
Pansion-Otel-Motel	Guest houses, Motels and bigger Hotels.	
5Yıldızlı Otel – Tatil Köyü	Five Star Hotels and Holiday Villages.	
Pub-Café-Bar	Pubs- cafes and Bars.	
Tali Bayi – Subordinate Distributor	Special kinds of distributors who serve other distributors.	
Diğer - Other	Other Type of Customers.	
If derived		
Calculation		

Variable: SES Gurubu – Social and Economical Status Group		
Short Description: Defines the socio economic status of the people who lives around the customer's location.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 10		
If Nominal		*If Ordinal*
Value	*Meaning*	
A+	Top Level of Income	
A,B	High Level of Income	
C	Middle Level of Income	
D,E	Low Level of Income	
If derived		
Calculation		

Variable: Bölge Tanımı – Region Description		
Short Description: Defines the region of the city where the customer has located.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 10		
If Nominal		*If Ordinal*
Value	*Meaning*	
Merkez – Center	Customers located at the center of the city or midtown area.	
Çevre – Around	Customers located around the city center.	
If derived		
Calculation		

Variable: Konum Gurubu – Position Group		
Short Description: Defines the positioning of the places that the customer has located.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 10		
If Nominal		*If Ordinal*
Value	*Meaning*	
Alışveriş Merkezi – Shopping Center	The customer's location position is at shopping centers.	
Ana Arter – Main Street	The customer's location position is on the main traffic arteries.	
Ara Sokak – Mid Street	The customer's location position is at cross streets.	
Paralel Arter – Parallel Street	The customer's location position is on the parallel traffic arteries.	
Null	The position of the customer location is not defined.	
If derived		
Calculation		

Variable: Çalışma Şekli – Working Type		
Short Description: Defines the group of customers defined according to their way of payment.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 20		
If Nominal		*If Ordinal*
Value	*Meaning*	
C/H	Special Bank Account – Customers pay via bank transfer.	
Çek – Cheque	Cheque - Customer pays by cheque.	
Peşin – Cash	Cash –Customer pays by cash at the same time he collects the	

	goods.	
Senet – Receipt	Receipt - Customer pays by bill of exchange.	
Null	The payment type of the customer is not defined.	
If derived		
Calculation		

Variable: Nokta Yapısı – Customer Structure		
Short Description: Defines the group of customer which is defined according to their visual presentation.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 8		
If Nominal		*If Ordinal*
Value	*Meaning*	
Standart – Standard	Standard – Customers whose locations are not decorated with special visual materials. These places may contain some POP materials but not with special ones.	
Imaj - Image	Image – Customer's location has been decorated with some special visual materials of the company. These customers are the ones who are located at critical parts of the city and they are the ones who have high turnovers.	
If derived		
Calculation		

Variable: Ziyaret Frekansı – Visit Freuency.		
Short Description: The characteristic shows visit frequency of the firm for the specified customer.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 20		
If Nominal		*If Ordinal*
Value	*Meaning*	
Haftada Bir – Once per week	The customer is being visited once per week.	
Haftada İki – Twice per week	The customer is being visited twice per week.	
Haftada Üç – Three per week	The customer is being visited three times per week.	
Her Gün – Every day	The customer is being visited every day.	
İki Haftada Bir – Once per two weeks	The customer is being visited bi weekly.	
If derived		
Calculation		

Variable: Nokta Özellik – Distributor's speciality		
Short Description: Defines the group of customers defined according to the products they are selling.		
Variable Type: Categoric – Nominal		
Data Expression : Text		
Can hold null:	*Yes:* x	*No:*
Length : 20		
If Nominal		*If Ordinal*
Value	*Meaning*	
Şirket Markaları – Company Brands	Company Brands – The customer sells only the products of the company.	
Alkolsuz - Non-Alcohol	Only one Special Product – The customer sells only a special brand of the company launched last year.	
Tüm Markalar – All	All Brands – The customer	

Brands	sells both the company's products and competitor's products.	
Potansiyel – Potential	Potential Customer – The customer does not sell the company's products but may be a potential customer.	
Diğer - Other	Other – The customer's specialty is different from the groups defined above.	
Null	The specialty of the customer is not defined.	
If derived		
Calculation		

Continuous Variables at Customer Level

Variable: Length of Relationship_1			
Short Description: Shows how long the specified customer is working with the company during the analysis period: four year.			
Variable Type: Numeric – Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 1095 days.	
If derived			
Calculation			
(Last purchase date – First purchase date) within analysis period.			

Variable: Length of Relationship_2		
Short Description: Shows how long the company is working with the specified customer. Different from the length of relationship_1 variable, it does not show only the duration in the analysis period.		
Variable Type: Numeric – Continuous		
Data Expression : Number		
Can hold null:	*Yes:*	*No:* x

Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 16434 days.	
If derived			
Calculation			
(Last purchase date – Customer Opening Date)			

Variable: Frequency			
Short Description: The number defines how many times the specified customer purchased from the firm during the analysis period.			
Variable Type: Numeric – Continuous			
Data Expression : Number			
Can hold null:		*Yes:* x	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 785 times.	
If derived			
Calculation			

Variable: rFrequency			
Short Description: Shows number of purchases customer made relative to the length of relationship (LoR_1).			
Variable Type: Numeric –Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 2.	
If derived			
Calculation			
(Frequency / Length of Relationship_1)			

Variable: Frequency Last Year			
Short Description: Shows how many times specified customer purchased goods from the company during the last year of the analysis period.			
Variable Type: Numeric –Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 335 times.	
If derived			
Calculation			

Variable: Recency			
Short Description: Shows the number of days that passed between the last two transactions of the customer with the company within the observation period.			
Variable Type: Numeric –Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 827.	
If derived			
Calculation			
(Date of the last Purchase – Date of the previous purchase before the last one) within the analysis period.			

Variable: The Average Inter Purchase Times			
Short Description: Shows the average time passed between each two purchases of the customer from the company during the analysis period. The variable reflects the Recency variable over the entire time period the customer has relation with the company.			
Variable Type: Numeric –Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 118.	
If derived			
Calculation			
(∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) within the analysis period.			

Variable: Standard Deviation of Recency			
Short Description: Shows the standard deviation of the inter purchase time.			
Variable Type: Numeric –Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 143.	
If derived			
Calculation			
StDev (∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) within the analysis period.			

Variable: Coefficient of Variation of Recency			
Short Description: Shows the ratio of StdRecency to Mean Recency.			
Variable Type: Numeric -Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 346.	
If derived			
Calculation			
(StDev (∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) / Average (Date of the last Purchase – Date of the previous purchase before the last one)) within the analysis period.			

Variable: Total Amount			
Short Description: Shows the total amount of products that the specified customer purchased from the company during the analysis period.			
Variable Type: Numeric -Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 90000 liter.	
If derived			
Calculation			

Variable: Amount				
Short Description: Shows the average amount of products that the specified customer purchased from the company during the analysis period.				
Variable Type: Numeric –Continuous				
Data Expression : Number				
Can hold null:			*Yes:*	*No:* x
Length :				
If Discrete		*If Continuous*		
Value	*Meaning*	*Range of the value*		
		The variable can take a value between 0 and 2020 liter.		
If derived				
Calculation				
(Total Amount / Frequency) within the analysis period				

Variable: Standard Deviation of Amount				
Short Description: Shows the standard deviation of the average amount of products that the specified customer purchased from the company during the analysis period.				
Variable Type: Numeric –Continuous				
Data Expression : Number				
Can hold null:			*Yes:*	*No:* x
Length :				
If Discrete		*If Continuous*		
Value	*Meaning*	*Range of the value*		
		The variable can take a value between 0 and 1615 liter.		
If derived				
Calculation				
(StDev (Total Amount / Frequency)) within the analysis period				

Variable: rTotal Amount				
Short Description: Shows total amount of products that the specified customer purchased from the company during the analysis period relative to the length of relationship (LoR_1).				
Variable Type: Numeric -Continuous				
Data Expression : Number				
Can hold null:		*Yes:*		*No:* x
Length :				
If Discrete		*If Continuous*		
Value	*Meaning*	*Range of the value*		
		The variable can take a value between 0 and 225 liter.		
If derived				
Calculation				
((Total Amount / Frequency) / Length of Relationship_1) within the analysis period				

Variable: rAmount				
Short Description: Shows average amount of products that the specified customer purchased from the company during the analysis period relative to the length of relationship (LoR_1).				
Variable Type: Numeric -Continuous				
Data Expression : Number				
Can hold null:		*Yes:*		*No:* x
Length :				
If Discrete		*If Continuous*		
Value	*Meaning*	*Range of the value*		
		The variable can take a value between 0 and 50 liter.		
If derived				
Calculation				
((Total Amount / Frequency) / Length of Relationship_1) within the analysis period				

Variable: rMajorTrip			
Short Description: Shows the percentage of the purchases of a customer which exceeds the average amount for the purchases that specified customer has done. The variable indicates the percentage of purchases that could be classified as a big shopping incidence.			
Variable Type: Numeric –Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		The variable can take a value between 0 and 100 percentage.	
If derived			
Calculation			
((∀ Count (∀ (Amount for specified order – Average Amount) > 0) / Total Number of Purchases) * 100)			

Variable: Frequency for years			
Short Description: The number defines how many times the specified customer purchased from the firm during the specified year, 2002, 2003, 2004.			
Variable Type: Numeric – Continuous			
Data Expression : Number			
Can hold null:		*Yes:* x	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		2002 → The variable can take a value between 1 and 265 times. 2003 → The variable can take a value between 0 and 267 times. 2004 → The variable can take a value between 0 and 687 times.	
If derived			
Calculation			

Variable: Total Amount for years				
Short Description: Shows the total amount of products that the specified customer purchased from the company during the specified year.				
Variable Type: Numeric -Continuous				
Data Expression : Number				
Can hold null:			*Yes:*	*No:* x
Length :				
If Discrete		*If Continuous*		
Value	*Meaning*	*Range of the value*		
		2002 → The variable can take a value between 12 and 154.000 liter. 2003 → The variable can take a value between 7 and 16.000 liter. 2004 → The variable can take a value between 0 and 85.000 liter.		
If derived				
Calculation				

Variable: Amount for years				
Short Description: Shows the average amount of products that the specified customer purchased from the company during the specified year.				
Variable Type: Numeric -Continuous				
Data Expression : Number				
Can hold null:			*Yes:*	*No:* x
Length :				
If Discrete		*If Continuous*		
Value	*Meaning*	*Range of the value*		
		2002 → The variable can take a value between 0 and 2440 liter. 2003 → The variable can take a value between 8 and 4000 liter. 2004 → The variable can take a value between 0 and 2020 liter.		

If derived	
Calculation	
(Total Amount / Frequency) within the analysis period	

Variable: The Average Inter Purchase Times for years			
Short Description: Shows the average time passed between each two purchases of the customer from the company during the specified year. The variable reflects the Recency variable over the entire time period the customer has relation with the company.			
Variable Type: Numeric -Continuous			
Data Expression : Number			
Can hold null:		*Yes:*	*No:* x
Length :			
If Discrete		*If Continuous*	
Value	*Meaning*	*Range of the value*	
		2002 → The variable can take a value between 0 and 2440. 2003 → The variable can take a value between 0 and 303. 2004 → The variable can take a value between 0 and 214.	
If derived			
Calculation			
(∑ (Date of the Last Purchase – Date of the previous purchase before the last one) / Total Number of Purchases) within the analysis period.			

APPENDIX B

(Summary Cluster Interpretations for Profiling)

CLUSTER THREE - STARS	
General Characteristics	√ Smallest cluster with 36 customers √ Outlier – important subgroup of data set √ Wide cluster
Characteristics Related to Continuous Variables	√ Have long relationship with company √ Buys for the greatest amounts on Total and Average √ Greatest Total Amount relatively LoR √ Smallest time between purchases
Characteristics Related to Categorical Variables	√ Sales Directorates: 1031, 1032, 1035, 1037 √ Customer Type: Closed, NA √ Working period: Standard √ Region: Center √ Position Group: Shopping Center, Mid Street, NA √ Customer Specialty: Company Brands √ Working Type: Cash, Cheque

CLUSTER EIGHT – VALUABLE CUSTOMERS	
General Characteristics	√ Relatively small cluster with 1019 customers (1.76%) √ Wide Cluster √ Not outliers
Characteristics Related to Continuous Variables	√ High LoR √ Buys frequently and there is short time between each two purchases √ Buying for greater amounts from other clusters except Cluster Three: Stars √ Buying pattern of the clusters is not a fluctuating one (rMajorTrip)
Characteristics Related to Categorical Variables	√ Sales Directorate: 1031, 1035 √ Customer Type: Barrels, NA √ Working Period: Standard √ Customer Group: Otel, Holiday Village, Beer House, Pub café Bar √ SES Group: A, B-High Income, A+, A, B-High Level and D,E-Low Income √ Region: Center √ Position Group: Shopping Center, Main Street, Parallel street and NA √ Customer structure: Image √ Visit frequency: Every Day, once per week √ Customer specialty: Company Brands, Non-Alcohol √ Working Type: Cash, NA

CLUSTER FOUR – FREQUENT BUYERS	
General Characteristics	√ Cluster with average size - 6464 customers (11.16%) √ Average wideness √ Not outliers
Characteristics Related to Continuous Variables	√ Longest LoR √ Buy frequently but not as frequently as Stars and Valuables √ Not buy for big amounts in each transaction √ Customers buy significantly small amounts compared

	to their LoR. √ Buying patterns of customers is smooth but sometimes they buy for bigger amounts.
Characteristics Related to Categorical Variables	√ Sales Directorate: 1031, 1032, 1035 and 1037 √ Customer Type: Closed, Barrels √ Working Period: Standard √ Customer Group: Buffet, Standard Beer House √ SES Group: A,B-High Income, A+, A,B-High Income, D,E-Low Income √ Region: Center √ Position Group: Shopping Center, NA √ Customer Structure: Does not characterize cluster based on Contingency test results (Table A) √ Visit Frequency: Every day, Once per week √ Customer Specialty: NA √ Working Type: Cheque, Cash, NA

CLUSTER SEVEN – AVERAGE CUSTOMERS	
General Characteristics	√ Biggest Cluster - 20152 customers (34.79%) √ Narrowest Cluster √ Not outliers
Characteristics Related to Continuous Variables	√ Have average LoR √ Buys for average amounts with average frequency √ Time between purchases is seven days √ Buying patterns of customers is smooth
Characteristics Related to Categorical Variables	√ Sales Directorate: 1032, 1033, 1034, 1035, 1037 and 1038 √ Customer Type: Closed √ Working Period: Standard √ Customer Group: Grocery and Buffet √ SES Group: D,E-Low.Income √ Region: Center √ Position Group: Main Street √ Customer Structure: Does not characterize cluster based on Contingency test results (Table 68) √ Visit Frequency: One per two weeks, twice per week or three per week. √ Customer Specialty: All brands √ Working Type: CH, Cheque and Receipt.

CLUSTER TWO – POTENTIAL VALUABLE CUSTOMERS	
General Characteristics	√ Second biggest cluster - 15632 customers (26.98%) √ Narrowest Cluster √ Not outliers
Characteristics Related to Continuous Variables	√ Have shorter relationship with company √ Time between purchases is longer than 10 days. √ Does not buy frequently but relatively to their LoR they same with average customers √ Does not buy frequently but buys for big amounts. √ Buys big amounts relatively to their length of relationship
Characteristics Related to Categorical Variables	√ Sales Directorate: 1040, 1038, 1034 and 1032 √ Customer Type: Open √ Working Period: Seasonal

	√ √ Customer Group: Otel Holiday Village, Restaurant, Pension Otel Motel, Pub cafe bar, Subordinate distributor √ SES Group: A,B-High Income, A+, A,B-High Income, C-Average Income √ Region: Around √ Position Group: Mid Street and Parallel Street √ Customer Structure: Does not characterize cluster based on Contingency test results (Table 68) √ Visit Frequency: Every day, Once per week √ Customer Specialty: Company Brands, Other √ Working Type: CH, Cash

CLUSTER ONE – POTENTIAL INVALUABLE CUSTOMERS	
General Characteristics	√ Average cluster size. 2941 customers. (5.08%) √ Average wideness √ Not outliers
Characteristics Related to Continuous Variables	√ Have a relationship with company for one year. Near to the average of all data set. √ Do not buy frequently relatively to their LoR √ Do not buy for bigger amounts
Characteristics Related to Categorical Variables	√ Customer Type: Open √ Customers located at: 1033, 1037, 1038, 1039, 1041 Sales Directorates √ Customer Group: Otel, Restaurant, Pub √ SES: High Income, High Level, Average √ Region: Center √ Customer Specialty: Company Brands, Other √ Working Type: C/H

CLUSTER FIVE – INVALUABLE CUSTOMERS	
General Characteristics	√ Small sized cluster with 292 customers (0.05%) √ Wide cluster but not as wide as Cluster 3. √ Outliers but not as far as the ones in Cluster 3.
Characteristics Related to Continuous Variables	√ Have long relationship with company more than one and half year. √ Do not buy frequently relatively to their LoR √ Time between purchases is the greatest one. √ Buying frequency decreased in last year √ Bought for significant amounts but the volume they bought decreased in the last year.
Characteristics Related to Categorical Variables	√ Sales Directorate:1032, 1035 and 1037 √ Customer Type: Open and NA √ Working Period: Standard √ Customer Group: Restaurant, Market , Pension Otel Motel and Other √ Region: Center √ Position Group: NA √ Visit Frequency: Every Day, Once per week, Once per two weeks and NA

CLUSTER SIX – POTENTIAL CUSTOMERS	
General Characteristics	√ Big sized cluster with 11.397 customers (19.67%) √ Narrow cluster but not as wide as Cluster 3. √ Not outliers but away from the general center.
Characteristics Related to Continuous Variables	√ Shortest length of relationship √ Buy frequently in the last year of observation period. √ Time between purchases is more than 10 days. √ Purchasing pattern does not fluctuate. √ Has similar patterns with Cluster Three: Stars
Characteristics Related to Categorical Variables	√ Sales Directorate: 1033, 1039, 1040 √ Customer Type: Open √ Customer Group: Beer House, Restaurant, Pension, Bar √ SES Status: C Average Income √ Region: Around √ Visit Frequency: Every day √ Customer Specialty: Company Brands, Non-Alcohol √ Working Type: Cash

City Clusters

CLUSTER ONE – MOST VALUABLE CITIES	
General Characteristics	√ Second biggest cluster with 22 cities (28.21%) √ Average wideness √ Not outliers
Characteristics Related to Continuous Variables	√ Greatest frequency √ Shortest IPT √ Second greatest count of company customers located in √ Second greatest Sales per Customer figure √ Consumption of company products is high √ Customers buy significant amounts

CLUSTER TWO – VALUABLE CITIES	
General Characteristics	√ Biggest cluster with 34 cities (43.59%) √ Narrowest cluster. √ Not outliers
Characteristics Related to Continuous Variables	√ Customers in these cities buy frequently for relatively big amounts. √ Cities are not so crowded √ Have an average Sales per Customer figure √ Have similarities with cities in Cluster One.

CLUSTER THREE – FIT CLASS CITIES	
General Characteristics	√ Small sized cluster with 3 cities (3.85%) √ Wide cluster. √ Outliers.
Characteristics Related to Continuous Variables	√ Customers in the cities of this cluster do not buy high amounts. √ There are not so many customers of company in these cities √ Small Sales per Person figure √ Cities in this cluster are not so crowded.

CLUSTER FOUR – MOST INVALUABLE CITIES	
General Characteristics	√ Smallest cluster with 2 cities (2.56%) √ Narrowest cluster √ Outliers
Characteristics Related to Continuous Variables	√ Second most crowded cities √ Least count of company customers are located in. √ Smallest Sales per Customer figure. √ Smallest Total Amount and Total Frequency

CLUSTER FIVE – INVALUABLE CITIES	
General Characteristics	√ Small sized cluster with 4 cities (5.13%) √ Wide cluster. √ Outliers.
Characteristics Related to Continuous Variables	√ Least crowded cities √ Second smallest Sales per Customer figure √ Second smallest Total amount figure

CLUSTER SIX – AVERAGE CITIES	
General Characteristics	√ Small sized cluster with 4 cities (5.13%) √ Widest cluster. √ Outliers
Characteristics Related to Continuous Variables	√ Do not buy frequently √ Have smaller Sales per Customer and Total Amount figures

CLUSTER SEVEN – STAR CITIES	
General Characteristics	√ Third biggest cluster with 9 customers (11.54%) √ Average wideness √ Not outliers
Characteristics Related to Continuous Variables	√ Most crowded cities √ Greatest count of company customers are located in. √ Greatest Sales per Customer figure.