INTERACTION OF DESCRIPTIVE AND PREDICTIVE ANALYTICS WITH PRODUCT NETWORKS: THE CASE OF SAM'S CLUB

by

BERNA ÜNVER

Submitted to
the Graduate School of Management
in partial fulfillment of
the requirements for the degree of
Master of Science

SABANCI UNIVERSITY

June 2019

INTERACTION OF DESCRIPTIVE AND PREDICTIVE ANALYTICS WITH PRODUCT NETWORKS: THE CASE OF SAM'S CLUB

APPROVED BY

Prof. Dr. Füsun Ülengin
(Thesis Supervisor) .....

Assoc. Prof. Dr. Abdullah Daşçı .....

Assoc. Prof. Dr. Özgür Kabak .....

DATE OF APPROVAL: 27.06.2019

ABSTRACT

INTERACTION OF DESCRIPTIVE AND PREDICTIVE ANALYTICS WITH PRODUCT NETWORKS: THE CASE OF SAM'S CLUB

BERNA ÜNVER

Business Analytics M.Sc. Thesis, June 2019

Thesis Supervisor: Prof. Dr. Füsun Ülengin

Keywords: two-stage clustering analysis, CLV, customer segmentation, product network analysis, HITS algorithm

Due to the fact that there are massive amounts of available data all around the world, big data analytics has become an extremely important phenomenon in many disciplines. As the data grow, the need for businesses to achieve more reliable and accurate data-driven management decisions and to create value with big data applications grows as well. That is the reason why big data analytics becomes a primary tech priority today.

In this thesis, initially we used a two-stage clustering algorithms in the customer segmentation setting. After the clustering stage, the customer lifetime value (CLV) of clusters were calculated based on the purchasing behaviors of the customers in order to reveal managerial insights and develop marketing strategies for each segment. At the second stage, we used HITS algorithm in product network analysis to achieve valuable insights from generated patterns, with the aim of discovering cross-selling effects, identifying recurring purchasing patterns, and trigger products within the networks. This is important for practitioners in real-life application in terms of emphasizing the relatively important transactions by ranking them with corresponding item sets.

From practical point of view, we foresee that our proposed methodology is adaptable and applicable to other similar businesses throughout the world, providing a road map for the potential applications.

ÖZET

ÜRÜN AĞLARININ BETİMLEYİCİ VE KESTİRİMSEL ANALİTİKLERLE ETKİLEŞİMİ: SAM'S CLUB VAKA ANALİZİ

BERNA ÜNVER

İş Analitiği Yüksek Lisans Tezi, Haziran 2019

Tez Danışmanı: Prof. Dr. Füsun Ülengin

Anahtar Kelimeler: iki aşamalı kümeleme analizi,müşteri yaşam süresi değeri, müşteri segmentasyonu, ürün ağı analizi, HITS algoritması

Günümüzde büyük miktarda kullanılabilir veri bulunması nedeniyle büyük veri analizi birçok disiplinde son derece önemli bir konu haline gelmiştir. Kullanılabilir veri miktarı büyüdükçe, işletmelerin daha güvenilir ve daha doğru veri odaklı yönetim kararları alma ve büyük veri uygulamalarıyla değer yaratma gereksinimi de artmaktadır. Büyük veri analizinin günümüzde birincil teknoloji önceliği haline gelmesinin nedeni budur.

Tez kapsamında öncelikle müşteri segmentasyonu bağlamında iki aşamalı kümeleme algoritması kullanılmıştır. Kümeleme aşamasından sonra, müşterilerin satın alma davranışlarına dayanarak yönetimsel içgörülerin ortaya çıkması ve her segment için pazarlama stratejilerinin geliştilmesi amacıyla kümelerin müşteri yaşam süresi değeri (CLV) hesaplanmıştır. İkinci aşamada, ürün ağı analizinde HITS algoritmasını ortaya çıkan örüntülerden değerli öngörüler edinmek, çapraz satış etkilerini keşfetmek, yinelenen satın alma alışkanlıklarını ve ürün ağlarında tetileyici ürünleri belirlemek amacıyla kullandık. Bu, gerçek hayattaki uygulayıcılar ve uygulamalar için göreceli olarak önemli işlemleri ilgili ürün setleriyle birlikte sıralayarak vurgulamak açısından önemlidir.

Pratik uygulamalar açısından, önerilen metodolojinin dünyadaki diğer benzer işletmeler için uyarlanabilir ve uygulanabilir olduğunu ve potensiyel uygulamalar için bir yol haritası oluşturacağı öngörülmektedir.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Füsun Ülengin, for her considerable encouragements, worthwhile guidance and insightful comments in order to complete this thesis. I feel honored for the opportunity to work under the supervision of her.

I would like to extend my heartfelt gratitude thanks to Prof. Dr. Ilker Topcu, for his outstanding and invaluable guidance, keen interest and tremendous cooperation.

I would like to express my sincere gratitude to Prof. Dr. Jennifer Shang for her warm and welcoming nature throughout our U.S. journey.

I would like to thank the following universities for their assistance with the collection of my data: University of Arkansas, and University of Pittsburg - Katz Business School.

I would like to dedicate this work to my family and my cousins: Özkan Ünver, Kevser Ünver, Büşra Ünver, Tunahan Ünver, Sema Ünver, Kübra Ünver, and İrem Özkan. Thank you for your love, support and faith in me. Special thanks also go to my close friends: Nur Beyza Aksoy, Aslı Ürem, Hatice Çakır, Zeynep Anlar, and Fethi Özkan. Finally, I would like to thank Arda Ağababaoğlu for his worthwhile motivation and encouragement.

I hope that I have made all of them a little proud! Thanks for the trusted bestowed on me.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ongoing forecasts indicate that revenue from big data and business analytics worldwide will reach 260 billion U.S. dollars in 2022, 233 billion U.S. dollars in 2021, and 208 billion U.S. dollars in 2020 [1]. This is an incredible global acceleration, leading to big structural and operational changes in business world. In conjunction with these forecasts, data-driven management has gained top priority for businesses to achieve more reliable and accurate management decisions and to create value with big data applications, especially in the last decade. According to a survey which conducted by Ascend2 and Research Partners in 2017, the most important data-driven objectives in marketing setting can be listed as basing more decisions on data analysis, acquiring more new customers, integrating data across platforms, enriching data quality and completeness, segmenting target markets, attributing sales revenue to marketing, and aligning marketing and sales teams [2].

Revolution of big data affects marketing researches and practices by exploring entirely new ways of understanding consumer behavior and formulating marketing strategies [3], [4]. Businesses aim to create consumer insights by gathering, storing, and analyzing big data related to the characteristics and behaviors of their customers in order to get competitive advantages for the future [5]. Big data analytics in the

marketing field focuses on better understanding the consumer behavior, effectively allocating the advertising budgets, improving the accuracy of the pricing strategies and demand forecasts and increasing customer satisfaction and loyalty.

This knowledge helps the businesses to develop more reliable and sustainable decision-making and strategic planning [6]. Strong customer relationships, lower management risks, improvements in operation efficiency, efficient marketing strategies and operation management are today more likely to be performed with the help of big data analytics application within the organizations [7]. Therefore, it seems that the tools, procedures and philosophies in the big data setting will continue to spread incredibly day by day and they will absolutely change long-standing management experiences and practices.

According to one of the most important review in management research area, conducted by Sheng et al. [7], there are three keystones for businesses to obtain value from big data and its potential applications. These keystones can be listed as value discovery, value creation, and value realization. Firstly, companies have made changes in their organizational alignment and IT structure via innovation and investment, and human resource management in order to discover value coming from big data for the last decade. Secondly, value creation has a significant role in strategic-decision making. Operation efficiency, marketing effectiveness, and cross-border decisions are all highly dependent on the performance of created value. Thirdly, value realization is measured by observing business development metrics such as financial performance, organizational success, and competition advantages. All these keystones are required to high level technology support with advanced techniques and applications.

In the context of this thesis, we focused on marketing segmentation and product network practices. Marketing segmentation is helpful for managers in order to target the appropriate marketing efforts to the most profitable and sustainable segments.

Businesses tend to spend time and effort in order to offer right products and services to the right customers in the big data revolution era. From the customer perspective, the past purchase and promotional-response history of the customers can help to retrieve information about micro-segmentation and prepare personalized promotions at least for similar customer segments. So, segmentation insights help companies to customize the marketing plans, identify the trends, plan the advertising campaigns, and deliver the relevant products to target customers [8]. Also, it helps to make proper marketing interventions for customers sharing similar preferences and purchasing patterns [9]. According to Bain & Company's "Management Tools & Trends 2018", marketing segmentation became one of the top ten executive management tools all over the world [10]. Therefore, promotion and price planning, category planning, reward programs for loyal customers, extension of core offers, right assortment planning, retention programs, and targeted communications planning can be more effective in marketing segmentation practices.

Product network analysis has commonly been used in order to gain valuable insights about customer purchasing behavior by identifying patterns with co-occurrences in transactional datasets. This type of analysis creates significant advantages for businesses to group products that co-occur in stores' layout design for the purpose of increasing in chance of cross-selling, driving recommendation engines, and targeting marketing campaigns with promotional coupons which includes related items they purchased frequently together. Moreover, product network analysis provides a solid base for category management domain in order to identify the products which are most likely to trigger cross-category sales, and to determine the most important products in terms of creating category loyalty.

A case study was conducted by analyzing the data set of Sam's Club, a division of Wal-Mart Stores, Inc. At the first stage, a novel approach based on two-stage clustering was applied in order to describe and predict purchasing behaviors of the

consumers. Initially, k-medoid clustering was used to group the individual customers. Subsequently, hierarchical clustering, was utilized in order to regroup those customers into distinct customer segments. After the clustering phase, the customer lifetime value (CLV) of the clusters were computed based on the purchasing behaviors of their members in order to reveal managerial insights and develop marketing strategies for each segment.

At the second stage, product networks were created for the top two of individual and business clusters. Due to the fact that the remaining clusters had relatively low customer lifetime value, we decided to create two general product networks by including all business clusters for business product network and all individual clusters for individual product network. Based on general product networks, one is for individual and the other is for business, we discovered valuable insights from generated patterns inside the networks. One of the most important aim of this thesis is to discover the cross-selling effects between items which are included in HITS model, and to find hubs in the transaction data set. Also, recurring purchasing patterns, complement, substitute and trigger products were identified within the network. We used HITS algorithm in order to perform product network analysis. The most important difference between HITS algorithm and classical association rule mining is that each transaction has different weights instead of equal weight assumption used in association rule mining [11]. This is important for practitioners in real-life application in terms of emphasizing the relatively important transactions by ranking them with corresponding item sets. The contributions of this thesis are reported in the following section.

1.1 Contributions of the Thesis

Contributions of the thesis can be summarized as follows:

- The most important contribution of this study from practical point of view is that the proposed methodology can be adapted and applied to other similar businesses throughout the world, providing a road map for potential applications.

- One of the most important contribution is the successive usage of these two-stage clustering algorithms, allowing the deeper understanding of each segment (a set of similar members). It can be reported that one of the main strong characteristics of this thesis is to create managerial insights for each segment based on the cluster characteristics and the CLV assessment metrics. These managerial insights are expected to help companies and marketing practitioners to develop effective and efficient marketing strategies.

- Another contribution is the usage of HITS algorithm in the product network analysis setting to achieve valuable insights from generated patterns, with the aim of discovering cross-selling effects, identifying recurring purchasing patterns, and trigger products within the networks. This is important for practitioners in real-life application in terms of emphasizing the relatively important transactions by ranking them with corresponding item sets.

1.2 Outline of The Thesis

- **Chapter 2** presents a detailed literature survey and background.

- **Chapter 3** provides the framework of the proposed methodology, including data collection, data derivation, data cleaning, descriptive analysis, two-stage clustering, and customer lifetime value estimation.

- **Chapter 4** highlights the product network analysis by using HITS algorithm.

- **Chapter 5** consists of conclusion and further suggestions.

- **Chapter A** includes Appendices.

1.3 Publications

- B. Ünver, F. Ülengin, and Y.I. Topcu (2019) "Assessing CLV scores of the Customer Segments Through a Weighted RFM Decision Model" The 25th International Conference on Multiple Criteria Decision Making (MCDM2019), June 16-21, Istanbul, TURKEY.

Chapter 2

Literature Survey and Background

2.1 Big Data in Marketing Analytics

When the term "big data" is searched in Google Scholar in the area of science, engineering and social science, many resources are encountered. There is no perfectly fitted threshold for the size and type of data, which can be accepted as big data [7].

Big data has a volume as expressed with petabytes, exabytes, or zettabytes. Although one of the hot topics for big data is related to its volume, the most important thing is that the ability to analyze vast and complex data sets [12]. Businesses focus basically on the features of big data, which are listed as velocity, volume, variety, and veracity. Volume represents the large size of data; velocity can be defined as speed or frequency of data generation; variety refers to various forms of data which ca be structured, semi-structures, and unstructured; and veracity is used to describe generated data accuracy [13].

In today's business world, companies spend a great effort in order to uncover hidden knowledge from big data. This knowledge can enable companies to develop more reliable and sustainable decision-making processes, as well as strategic planning phase

[14]. Data-driven management has gained top priority for businesses to achieve more reliable and accurate decisions and to create value with big data applications. Big data analytics has gained an incredible acceleration in business practices, combining massive data sets and advanced analytics techniques. Big data applications help companies to determine the competitors and customers' requirements in more reliable and accurate way. Moreover, businesses are more likely to reach as much information about customers' life as possible, because in todays' world, they are willing to respond efficiently to the customers' changing demands and expectations in a short time [12]. In today's world, strong customer relationships, lower management risks, improvements in operation efficiency, efficient marketing strategies and operation management are more likely to be performed with the help of big data analytics application within the organizations [7].

2.2 Segmentation

There is a certain fact that consumers are offered great variety of products and information never seen before. This situation causes an increase on consumers' diversified demands and expectations. Recommendation systems have gained a popularity in order to fulfill customers' demand and expectations. These systems are aimed to retain loyal customers and to attract new ones [15].

Customer segmentation was firstly developed by American marketing expert, Wendell R. Smith in the middle of 1950s [16]. Customer segmentation can be defined as classification of customers based on their value, demands, preference and other factors depending on business strategies, models and purposes. The main purpose of customer segmentation is to achieve distinct segments, which means that customers in the same groups should have certain similarities, on the other hand, customers in different groups have distinct characteristics [17]. Marketing segmentation is beneficial for companies to gain insights about current customers, as well as to determine potential customers for the company. It is an important fact that retention of customers is more important than spending effort to find new customers. Customization of marketing plans, identification of trends, planning of product development, planning of advertising campaigns, and delivery of relevant products can be supported with customer segmentation implementations [8].

2.2.1 Clustering in the context of Marketing Segmentation

Clustering is one of the most commonly used technique in the context of marketing segmentation [18], [19], [20], [21].

Murray et al. [18] concluded that historical transaction data create a valuable chance for analysts to achieve patterns which can be beneficial to predict consumer behavior. They proposed a marketing segmentation methodology based on customers' historical data by using dynamic time warping in the context of time-series clustering. It is important for practitioners to extract appropriate attributes from the data, because this data should be processed in order to reflect the customer behavior.

Griva et al. [19] proposed a clustering approach for customer visit segmentation using basket sales data. Using product categories, they classified customer visits by creating a product taxonomy with different levels from categories to items. Based on the results of proposed customer visit segmentation, the decisions on marketing campaigns for each distinct customer segment and on the redesign of a store's layout can be employed for product recommendation.

Tripathi et al. [20] proposed a hybrid solution with the combination of two separate clustering algorithms which are k-means and hierarchical for customer segmentation. It is reported that the usage of two clustering algorithms have outperformed compared to one clustering algorithm.

Huang et al. [21] conducted a case study in the context of analyzing retail customers' shopping patterns via three different clustering approaches. It is stated that based on clustering results, marketing strategies, cross- and up-selling opportunities can be revised in order to increase spending per visit, as well as customer loyalty.

RFM (Recency, Frequency, and Monetary) analysis, which is used to evaluate customers based on their past purchasing behaviors, is commonly used in the literature [8], [15], [17], [22], [23], [24], [25].

Christy et al. [8] proposed three different clustering algorithms based on RFM analysis in order to obtain distinct customer segments in the context of marketing segmentation.

Rodrigues and Ferreira [15] proposed a recommendation algorithm after applying customer segmentation and association rule mining in order to determine the best products for each target customer groups to recommend. Customer segmentation stage was performed using RFM variables to detect buying habits.

Wu and Lin [17] developed a customer segmentation model based on consumption level and consumption fluctuation for the purpose of optimizing marketing strategies according to different customer segments.

Chang and Tsai [22] developed a group RFM model to discover better customer consumption behavior. Based on the group RFM model, they clustered customers into different groups with respect to group RFM variables in order to measure customer loyalty and contribution. From management perspective, it can be used for planning of personalized purchasing and inventory management system.

Han et al. [23] proposed a clustering approach in order to design category strategies for each cluster. Category indices, which is used in category data clustering algorithm, were created by using average sales frequency, average sales volume, average sales revenue, average gross profit and average growth rate of each category. In this study, it was also applied an extended RFM model (Weighted RFM model) for clustering process. Finally, these two models were compared with each other.

Cheng and Chen [24] proposed a procedure using RFM attributes into clustering algorithm. The main objective is to cluster customer value in order to determine customer loyalty.

Tsai and Chiu [25] introduced a purchase-based segmentation methodology based on transactions history of customers in order to provide homogeneous marketing programs for each distinct segment. Also, they used RFM model in order to analyze the relative probability of each customer clusters after segmentation.

Figure 2.1 shows method(s) and tool(s), and attributes that are utilized in the corresponding articles.

			Attributes						
Authors	**Year**	**Method (s) and Tool (2)**	**Recency**	**Frequency**	**Monetary**	**Total Spending**	**Avg. Int. Time**	**Total # of Scanned Items**	**Avg # of Scanned Items**
Murray et al.	2017	Hierarchical Clustering, Marketing Segmentation		x		x		x	
Griva et al.	2018	Customer Visit Segmentation, Clustering	x	x	x		x		x
Tripathi et al.	2018	Customer Segmentation, Clustering			x	x			
Huang et al.	2009	Clustering		x	x	x			
Christy et al.	2018	Customer Segmentation, RFM Analysis	x	x	x				
Rodrigues and Ferreira	2016	RFM Analysis, Clustering	x	x	x				
Chang and Tsai	2011	RFM Analysis, Segmentation	x	x	x				x
Han et al.	2014	Marketing Segmentation, Clustering	x	x	x				x
Tsai and Chiu	2004	RFM Analysis, Marketing Segmentation, Clustering	x	x	x				
Cheng and Chen	2009	RFM Analysis, Clustering	x	x	x				
Wu and Lin	2005	RFM Analysis, Clustering	x	x	x				

FIGURE 2.1: Summary of Articles

All the articles summarized in segmentation section have different purposes and methodologies. This is important for us in order to perform data analysis in the context of this thesis. There are various limitations and further suggestions for these articles which we should point out. Some articles [8], [17], [20], [21], and [24] lacked adequate size of data to evaluate the proposed approach comprehensively. Instead of using a large volume data, they utilized sampling or filtering methods when they conducted their proposed methods. Other groups of articles [20], and [21] had limited number of attributes. The majority of articles except [19] only proposed purchase-based segmentation by including either products/product categories or customers in the transactional data. Another group of articles [21], and [23] took only short time periods into consideration.

The following section on Customer Lifetime Value includes the summarization of articles that proposed CLV segmentation in marketing setting.

2.3 Customer Lifetime Value (CLV)

Due to the fact that there is an important need to determine which customers are more profitable and loyal for companies in such a competitive business environment, CLV segmentation has evolved from year to year .

Customer-centric strategies, in other words customized marketing strategies have gained a great importance in the marketing area.

The continuous retention of customers, customer loyalty, new product and service developments and higher profits via customer analytics applications are popular research and implication areas in customer relationship management. There are four dimensions in customer relationship management: finding the customer identity, customers charm, retention of customers and customers growth [26].

Sheshasaayee and Logeshwari [26] combined RFM and LTV (Lifetime Value) model in order to perform segmentation, then to execute campaign planning and implementation based on the segmentation results. Another remarkable purpose in this study was to find target customers for developing efficient marketing strategies.

Tirenni et al. [27] proposed a value-based segmentation in order to determine customer lifetime value for each customer segment and to allocate efficiently marketing assets.

Ray and Mangaraj [28] developed a value-based customer segmentation utilizing a data mining method including AHP into it. They used AHP in order to define the importance (relative weight) of LRFM (Length, Recency, Frequency, and Monetary) in the calculation of customer lifetime value after applying a clustering approach for segmentation.

Liu and Shih [29] proposed a novel product recommendation system by using clustering approach in customer segmentation and AHP in the determination of the weights of recency, frequency, and monetary attributes which included in customer lifetime value calculation.

Hiziroglu and Sengul [30] proposed a comparative study by assessing two different customer lifetime value models within the scope of segmentation. They utilized RFM model to calculate customer lifetime value as one of the methods in this study.

Khajvand et al. [31] proposed a customer segmentation using RFM model and an extended version of RFM analysis method by adding an additional parameter, which is called count item, in order to estimate CLV values for each customer segment.

Hosseini et al. [32] proposed two RFM models to cluster customers, one includes non-weighed parameters, on the other hand, the other involves in weighted parameters. Then, they assessed CLV rankings.

Khajvand and Tarokh [33] utilized an adapted weighted RFM model to perform customer segmentation. They, they assessed CLV values of each segment based on six recent seasons.

Hosseini and Shaban [34] classified customers based on their values using RFM model and k-means clustering method. To evaluate customer values of segments, they aimed to achieve better results with analyzing the changes in customer value based on the time stamps.

Santoso and Erdaka [35] conducted two separate experiments in order to estimate customer lifetime value by developing several hypothesis in the context of research model. They developed their hypotheses with recency, monetary, and frequency attribute. In order to test hypotheses, they utilized multiple regression method by calculating customer life time value.

Figure 2.2 gives a summary of the articles in this section.

			Attributes		
Authors	**Year**	**Method (s) and Tool (2)**	**Recency**	**Frequency**	**Monetary**
Sheshasaayee and Logeshwari	2017	Customer Segmentation, CLV model, RFM Analysis, Clustering	x	x	x
Tirenni et al.	2007	CLV model	x	x	x
Ray and Mangaraj	2016	Clustering, Pairwise Comparison - AHP, CLV model	x	x	x
Liu and Shih	2005	Clustering, Pairwise Comparison - AHP, CLV model	x	x	x
Hiziroglu and Sengul	2012	CLV model	x	x	x
Khajvand et al.	2011	CLV model, RFM Analysis, Pairwise Comparison - AHP	x	x	x
Hosseini et al.	2010	RFM Analysis, Clustering, CLV model	x	x	x
Khajvand and Tarokh	2011	RFM Analysis, CLV model, Pairwise Comparison - AHP	x	x	x
Hosseini and Shaban	2015	RFM Analysis, CLV model, Clustering	x	x	x
Santoso and Erdaka	2015	RFM Analysis, CLV model	x	x	x

FIGURE 2.2: Summary of Articles

All articles under the customer lifetime value considered, CLV calculation with RFM model is widely used in segmentation setting. It is useful for marketing practitioners to determine customer loyalty, customer retention and customer churn rates. There are some limitations and future directions which we should indicate. Some articles [26], and [35] evaluated customers' past purchasing behaviors within a short time period such as 4-6 months. Another groups of articles [28], [29], and [30] had relatively small size of data. Some of the articles [26], [28], [29], [32], and [34] utilized both clustering and CLV model, but they used RFM attributes in both clustering and CLV model.

Our contribution in customer lifetime value setting is that we utilized three different methods to determine the weights of RFM attributes. It allows us to benchmark the results of CLV scores of our customer segments.

The next section, which refers to Hubs and Authorities, evaluates the articles which used HITS algorithm in product network setting.

2.4 Hubs and Authorities (HITS)

Hyperlink-Induced Topic Search, also known as Hubs and Authorities was firstly developed by Kleinberg (1999) [36] in order to rank pages in the contexts on the World Wide Web. The basic objective of the usage of HITS algorithm in this study was to detect hubs and authorities of the pages iteratively.

The main idea behind the usage of HITS in transaction data sets is that the weights of transactions, in other words hub scores, and the weights of items, which is authority scores, are in a mutually reinforcing relationships [11], [37], and [38].

Sun and Bai [11] utilized HITS algorithm in movie ranking data set used by NetFlix for the purpose of discovering the cross-selling effects between items by utilizing w-support and w-confidence as the rule selection thresholds.

Wang and Su [37] used HITS algorithm in order to rank items in the retail data set. There was an additional factor in the case study: individual profits of items. They used both real and synthetic data sets when searching appropriate associations among items by taking into consideration individual items' profits. One of the similar study with Wang and Su [37] belongs to Ramasamy and Lokeshkumar [38], they utilized HITS algorithm in a large dataset with only binary attributes to analyze cross-selling effects by taking into consideration the hub scores of transactions previously.

In this section, a detailed literature review is conducted in order to analyze the articles, which utilized at least one method that we used in the context of this thesis, by focusing on main objective(s), methodology, further suggestions, and limitations.

Chapter 3

Data Analysis

3.1 Data Collection

Sam's Club is a membership-based club, which provides goods and services for individual customers and business owners with different types and sizes. Both individual and business (industrial) customers have a membership card to shop at Sam's Club stores.

There are nine main departments at Sam's Clubs:

- grocery;
- office;
- pharmacy, health & beauty;
- jewelry, flowers & gifts;
- home and appliances;
- electronics & computers;

- apparel, shoes, sports &fitness;
- toys, games, books & entertainment;
- auto & tires

Sales at Sam's Club stores are unique resources for Sam's Club database. UA_SAMSCLUB_small database from the University of Arkansas Enterprise Systems Teradata source was used as a data source in this study. The database contains store visit information of seven stores from 7/31/2005 through 11/2/2006. There are more than 9 million transactions and 86 attributes in total, which are attached in Appendix A.1 and A.2.

The database involves six different tables:

- STORE_VISIT
- ITEM_SCAN
- MEMBER_INDEX
- ITEM_DESC
- STORE_INFORMATION
- SUB_CATEGORY_DESC

After several meetings with experts [1] to discuss the literature and the aim of the study, the attributes that can be used for this study were revealed.

[1] Assoc. Prof. Dr. Ron Freeze (Associate Director of Technology for Enterprise Systems, University of Arkansas), Dr. Michael Gibbs (Associate Director for Enterprise Systems, University of Arkansas), Assoc. Prof. Dr. Nitin Vasant Kale (Information Technology Program and Dept. of Industrial and Systems Engineering at University of Southern California), Prof. Dr. Jennifer Shang (Professor of Business Administration, Area Director for Business Analytics and Operations at University of Pittsburgh - Katz Business School) and Prof. Dr. Ilker Topcu (Istanbul Technical University - Department of Industrial Engineering)

There are 22 distinct attributes involving five tables as given at Table 3.1.

Table 3.1: Attributes extracted from UA_SAMSCLUB_small database

Table Name	**Attributes**
STORE_VISITS	visit_number
	store_number
	membership_number
	tender_type
	tender_amount
	total_visit_amount
	transaction_date
	transaction_time
	total_unique_item_count
	total_scan_count
ITEM_SCAN	visit_number
	store_number
	item_number
	item_quantity
	total_scan_amount
	transaction_date
	unit_cost_amount
	unit_retail_amount
MEMBER_INDEX	membership_number
	zip_code
STORE_INFORMATION	store_number
	store_name
	city
	state
	zip_code
ITEM_DESC	item_number
	category_number
	primary_description
	brand_name

As can be seen in Figure 3.1, the selected attributes constitute an entity relationship diagram.

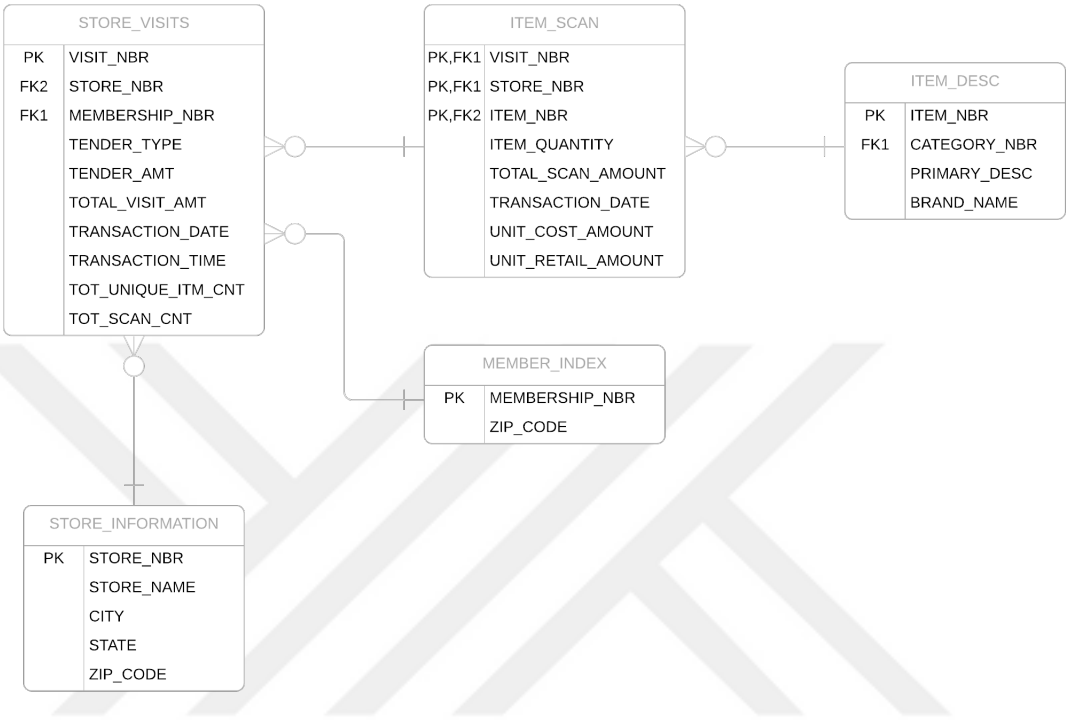


FIGURE 3.1: The Entity Relationship Diagram

3.1.1 Transaction Attributes

The descriptions and explanations of the selected attributes are given below:

- Visit Number (VISIT_NBR)

Visit number describes each different shopping trip with a nine-digit number. For example, if a member has five different shopping trips, she/he has five different visit numbers. There are 431,070 different visit numbers (i.e. shopping trips) in our transaction dataset.

- Transaction Date (TRANSACTION_DATE)

Transaction Date refers to the day of transaction with the date format. In our transaction dataset, the start date is July 31, 2005 and the end date is November 2, 2006.

- Transaction Time (TRANSACTION_TIME)

Transaction time defines the time of day that the transaction is started. Transaction time starts at 7:00 am and ends at 10:00 pm.

- Store Number (STORE_NBR)

Store number refers to store identification number, which means that each store has a unique store number. There are seven different stores, therefore, we have seven different store numbers in our transaction dataset as shown in Table 3.2.

TABLE 3.2: Store Information

Store Number	**Store Name**	**# of TRX**[a]
6	Extreme Retailers, ATLANTA, GA	180,931
7	Extreme Retailers, ATLANTA, GA	160,729
8	Extreme Retailers, AUGUSTA, GA	170,681
10	Extreme Retailers, BATON ROUGE, LA	768
59	Extreme Retailers, JACKSON, NY	328,155
66	Extreme Retailers, KANSAS CITY, MO	245,679
68	Extreme Retailers, KANSAS CITY, MO	144,285

[a] TRX : Transaction

- Store Name (STORE_NAME)

There are seven different stores in our transaction dataset. The store numbers, the names of store names and corresponding number of transactions are given at Table 3.2.

- Store City (STORE_CITY)

Store city is a location-based attribute and it provides an information indicating the city where the store is located.

- Store State (STORE_STATE)

Store state is another location-based attribute and it provides an information indicating the state where the store is located.

- Store Zip Code (ZIP_CODE)

Store zip code is another attribute which gives a location information about stores.

- Membership Number (MEMBERSHIP_NBR)

Each member has a unique membership number, which is assigned to the member upon joining the club. There are 91,876 different membership numbers in our transaction dataset. Therefore, we have 91,876 members.

- Member Zip Code (ZIP_CODE)

Member zip code is an attribute which gives a location information about members' residence.

- Tender Type (TENDER_TYPE)

Tender type defines the type of payment used in each visit. There are seven different tender types which can be listed as 0: Cash, 1: Check, 2: Gift Card, 3: Discover, 4: Direct Credit, 5: Business Credit, 6: Personal Credit.

We decided to focus on four tender types, namely cash, direct credit, business credit, and personal credit. The main reason of this decision was that we aim at revealing appropriate results and insights with a big data application. Therefore, we have a strong opinion that the choice of these four types of tenders would be suitable for being a benchmark study in terms of the applicability in retail sector in Turkey.

- Item Number (ITEM_NBR)

Item number refers to a unique number assigned to every different item for sale. There are totally 6981 different item numbers in our dataset.

- Item Quantity (ITEM_QUANTITY)

Item quantity helps to quantify of a unique item that is scanned during a transaction.

- Tender Amount (TENDER_AMT)

Tender amount describes the amount spent at the purchase. Occasionally, a member can use more than one tender type at a unique visit. In this case, there are two tender amount values for the member in the same visit.

- Total Unique Item Count (TOT_UNIQUE_ITM_CNT)

Total unique item count describes the number of unique items purchased per visit.

- Total Scanned Count (TOT_SCAN_CNT) $\Rightarrow$ Total Number Scanned (TOT_NBR_SCANNED)

Total scanned count refers to total number of scanned items per visit. It was necessary to change the name of this attribute to prevent the confusion between total scanned count and total scan amount. The new name for this attribute in the dataset is total number scanned (TOT_NBR_SCANNED).

- Total Visit Amount (TOT_VISIT_AMT) ⇒ Total Value per Visit (TOT_VALUE_PER_VISIT)

Total visit amount specifies the total monetary value of the entire visit. We needed to change the name of this attribute to prevent confusion between total visit amount and total scan amount. The new name for this attribute in the dataset is total value per visit (TOT_VALUE_PER_VISIT).

- Total Scan Amount (TOTAL_SCAN_AMOUNT)

Total scan amount refers to the total number of items scanned per visit number.

- Unit Cost Amount (UNIT_COST_AMOUNT)

Unit cost amount value was obtained by dividing cost by unit amount. This is a scrubbed value, which meant that costs and units were rounded to achieve an approximate unit cost amount.

- Unit Retail Amount (UNIT_RETAIL_AMOUNT)

Unit retail amount value was captured via dividing purchase price by unit amount. This was a scrubbed value, which meant that purchase prices and units were rounded to achieve an approximate unit retail amount.

- Category Number (CATEGORY_NBR)

Category number is a number assigned to each category of items. There are 61 category numbers in our dataset. Each category number has different items with different primary descriptions.

- Primary Description (PRIMARY_DESC)

This attribute helps to get informative description of items.There is just one category number for the items with the same primary descriptions.

- Brand Name (BRAND_NAME)

There is at least one brand name associated with the item in our dataset.

3.2 Data Derivation

There were 3 steps for the configuration of the data. These steps can be listed as below;

- Adjustment of data types

All extracted attributes from database pretended as numeric attribute. To handle with this problem, we made adjustments based on the types of attributes. For example, transaction date was converted to date format from numeric format.

- Derivation of transaction attributes

There were valuable attributes derived from existing ones. For example, category attribute was derived from category number and primary description attributes with grouping category numbers and corresponding primary descriptions.

- Derivation of customer attributes

Customer attributes were derived based on transaction dataset. We achieved customer and transaction datasets for business and individual members at the end of this step.

Table 3.3 and Table 3.4 exhibit the number of transactions and the number of members according to type of datasets and type of members.

Table 3.3: Information on Transaction Dataset

Member Type	**The Number of Transactions**
Individual Member	1,046,457
Business Member	66,952

Table 3.4: Information on Customer Dataset

Member Type	**The Number of Members**
Individual Member	47,013
Business Member	1,454

3.2.1 Derived Transaction Attributes

- Parts of Day (PartsOfDay)

Based on the transaction time, we derived parts of day attribute having three sections; namely, morning, afternoon, and evening. Morning defines the visits made before noon. Afternoon visits are defined as the visits between noon and 5:00 p.m. Evening, on the other hand, consists of the visits after 5:00 p.m.

- Interpurchase time (InterpurchaseTime)

Interpurchase time refers to the number of days between two consecutive shopping trips. For example, if a customer visits Sams Club five times, there will be four different interpurchase time values in the transaction dataset for that customer.

- Category (Category)

There are 61 different category numbers (CATEGORY_NBR) extracted from the database.

In descriptive analysis, category numbers are more likely to cause conflicts and difficulties. As can be seen in Table 3.5, we grouped these category numbers under categories based on departments of Sam's Club.

Table 3.5: Categories

#	**Category Name**
1	Apparel & Shoes
2	Auto
3	Beverages
4	Books & Entertainment
5	Bread & Bakery
6	Candy & Snacks
7	Canned, packaged foods
8	Cigarettes & Tobacco
9	Dumped Item
10	Electronics
11	Furniture & Mattresses
12	Health & Beauty
13	Home and Appliances
14	Household Essentials and Pets
15	Jewelry, Flowers & Gifts
16	Meat, Poultry, Seafood, Eggs & Diary
17	Membership
18	Office
19	Outdoor, Patio & Garden
20	Sports & Fitness
21	Toys & Games

3.2.2 Derived Customer Attributes

- Recency (Recency)

Recency reflects the number of days in-between the end of dataset period (November 2, 2006) and the last purchase of a customer.

- Frequency (Frequency)

Frequency represents the total number of shopping trips of each customer. For example, if a customer has five different visit numbers in dataset, then the frequency value of that customer will be five.

- Unique category count (Unique_Category_Count)

Customers purchase items in different categories during their own shopping history. Unique category count shows the number of different categories in which a customer is purchasing items. We derived unique category count attribute by analyzing and counting distinct categories for each member.

- Unique item count (Unique_Item_Count)

Unique item count, on the other hand, shows the number of distinct items purchased by a customer in her/his own shopping history. We derived unique item count attribute by counting item numbers for each member.

- Total spending (Total_Spending)

Total spending value per each visit of each customer is the value of Tender Amount (TENDER_AMT) attribute. Total spending is the summation of these values for each customer in her/his shopping history.

- Monetary (Monetary)

Monetary attribute shows the average spending amount per visit for each customer. It was derived by dividing total spending by frequency, in other words number of different visit numbers.

- Average of interpurchase time (Avg_InterpurchaseTime)

Average interpurchase time is an important indicator that shows the average number of days between two consecutive shopping trips of a customer in her/his shopping history. Average interpurchase time value is equal to sum of interpurchase time divided by number of purchase intervals (i.e. frequency-1).

- Total number of scanned items (Tot_Nbr_Scanned)

Total number of scanned items is the summation of total numbers of all scanned items for each customer at each shopping trip.

- Average number of scanned items (Avg_Nbr_Scanned)

Average number of scanned items is an important attribute which shows how many items on average a customer purchases per visit. It is equal to total number of scanned items divided by frequency.

- Standard deviation of spending (SD_TotalSpending)

Standard deviation of spending reflects the variation in the the customer spending in her/his shopping trips. For example, let's consider that there are two members with 3 shopping trips. One member spends $200 at all trips and the other member spends $100, $300, and $200, respectively. Although their total spending values are exactly

the same, they definitely have different characteristics in terms of spending behavior. Standard deviation of spending helps to analyze and explain these dissimilarities between members with respect to their total spending.

- Standard deviation of total number of scanned items (SD_TotalScanned)

Standard deviation of total number of scanned items represents the variation in the total scanned items during in her/his shopping trips. To derive standard deviation of total number of scanned items, we analyzed each separate visit for each customer. For example, let's consider two members with 4 shopping trips. A member purchases 20 items for each visit. The other member purchases 25,530,20 items, respectively. Although their total number of scanned items in total shopping history are exactly the same, it is certain that they have different characteristics in terms of shopping behavior. Standard deviation of total number of scanned items is helpful to analyze and explain these differences between members with respect to total number of scanned items.

3.3 Data Cleaning

3.3.1 Initial Dataset

As mentioned before the data at the Walton College Teradata platform has more than 9 million transactions and 86 attributes. The master data is limited by the absence or brevity of metadata with respect to some Teradata dataset attributes. This means that there are some transactions that have no or inaccurate data. For example, some transactions have missing primary descriptions and missing category numbers. We could not have a chance to make backtrack search to find and fill out these transactions with accurate descriptions. Therefore, to achieve more accurate and reliable predictive results we did not consider these transactions. After extracting the data from Teradata platform, we obtained an initial dataset that had 1,235,565 transactions (i.e. transactions with no missing descriptions for categorical variables and filtered transactions for numeric variables).

Two additional steps were conducted to clean the initial transaction dataset.

Firstly, we excluded the transactions with missing numeric values. Although many attributes were connected with each other via numeric values, we needed to use imputation methods as much as we can. If there was no chance to fill the value with the help of the other numeric attributes, these transactions were not taken into consideration to get better predictive performance.

Secondly, we excluded the transactions made between 10.00 pm and 7.00 am because the stores were actually closed at that period.

After these two eliminations, we came up with a transaction dataset having 1,231,228 transactions as seen in Table 3.6.

Table 3.6: Initial Cleaning of the Transaction Dataset

Elimination Steps	**# of TRX**[a]	**Decreased ratio**	**Total decreased ratio**
Initial Dataset	1,235,565		
Elimination Step 1	1,231,558	0.32%	**0.35%**
Elimination Step 2	1,231,228	0.03 %	

[a] TRX : Transaction

3.3.2 Individual and Business Transaction Datasets

Due to their different characteristics, after cleaning the initial dataset, we decided to group them into two; one for individual members (i.e. initial individual transaction dataset) and the other for business members (i.e. initial business transaction dataset). This separation was an important contribution of our research.
Among 1,231,228 transactions, 1,158,613 belonged to individual members and 72,615 of them belonged to business members. For both of these two groups of data, there were an additional data cleaning phase using five attributes, namely, tender amount, total unique item count, total number of scanned items, total value per visit, and total scan amount.
We determined the upper limits via 68–95–99.7 rule, also known as the empirical rule. This is a fact that approximately 99.7 % of the observations fall within three standard deviations of the mean.

After eliminating the transactions having values beyond these upper limits, we came up with an individual transaction dataset having 1,046,457 transactions and a business transaction dataset having 66,952 transactions as seen in Tables 3.7 and 3.8.

Table 3.7: Elimination Results for Individual Transaction Dataset

Eliminations	**# of TRX**[a]	**Decreased ratio**
Initial Dataset	1,158,613	
After Elimination Stage	1,046,457	**9.68%**

[a] TRX : Transaction

Table 3.8: Elimination Results for Business Transaction Dataset

Eliminations	**# of TRX**[a]	**Decreased ratio**
Initial Dataset	72,615	
After Elimination Stage	66,952	**7.8%**

[a] TRX : Transaction

3.3.3 Individual and Business Customer Datasets

Similarly to the transaction data set, the customer dataset was also separated into two: initial individual customer dataset and initial business customer dataset. There were 88,855 individual members and 2,142 business members in the corresponding initial customer datasets.

The most important cleaning in customer datasets was screening out members whose frequency values were equal to one. This means that these members made just one visit in the whole period. Therefore, interpurchase time of these members cannot be calculated and their monetary values (average total spending) become misleading. Therefore, we screened out the members who make just one visit. This further cleaning resulted with 51,836 individual members and 1,608 business members at the corresponding datasets.

Subsequently, we analyzed the distribution related to the total spending and the total number of scanned items, as well as the standard deviation of spending attributes. Similarly, we used the empirical rule in order to determine the upper limits for these variables and then eliminated members having values beyond these upper limits. This step has a significant role on getting representable customer datasets. All numeric variables in the customer datasets were connected to each other because they

were derived attributes from transaction datasets. Since the change in a variable would have an adverse effect on other variable(s), this step in data cleaning was essential to achieve representable member population.

As a result, we came up with an individual customer dataset having 47,013 members and a business customer dataset having 1,454 members as seen in Tables 3.9 and 3.10.

Table 3.9: Elimination Results for Individual Customer Dataset

Eliminations	**# of Members**	**Decreased ratio**
Members making more than one visit	51,836	
After Elimination Stage	47,013	**9.31%**

Table 3.10: Elimination Results for Business Customer Dataset

Eliminations	**# of Members**	**Decreased ratio**
Members making more than one visit	1,608	
After Elimination Stage	1,454	**9.58%**

3.4 Descriptive Analysis

Based on all the transactions extracted between the beginning and the end of dataset period (7/31/2005 - 11/2/2006), we made the visualization of findings in the context of descriptive analysis.

It is certain that descriptive insights have significant effect on deciding which attributes to be taken into account for clustering stage.

Descriptive analysis on the transaction and customer datasets are revealed in the following sections.

3.4.1 Transaction Datasets

This section aims to provide some critical visual evaluations for the transaction datasets of both individual and business members.

3.4.1.1 Individual Members

This section is designed to analyze individual transaction data set in a comprehensive way. The distribution of daily transactions per part of day, the distribution of visits and members per parts of day, and the distribution of transactions per day for top six categories are visualized as follows.

Figure 3.2 reveals the distribution of transactions per part of day for each day.

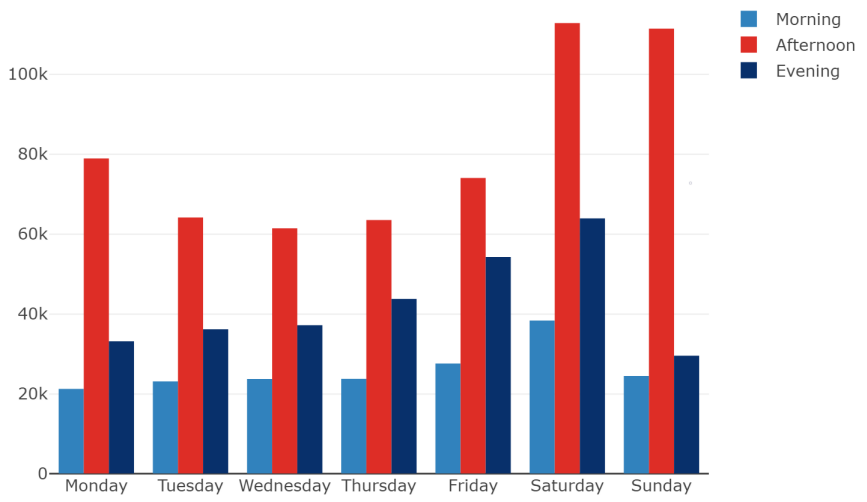
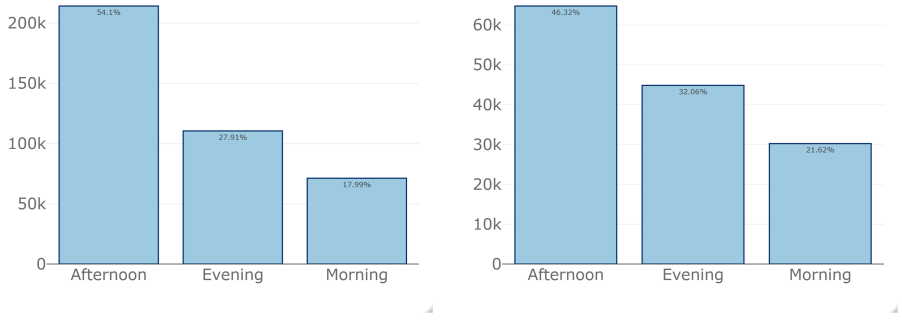


FIGURE 3.2: The distribution of Daily Transactions per Part of Day

Transactions made by individual members Saturday afternoons (112,841 transactions) and Sunday afternoons (111,451 transactions) are leading. Among the evenings, the busiest one is Saturday evenings (63,936 transactions), followed by Friday evenings (54,251 transactions).

Figure 3.3 exhibits the distribution of visits and members per parts of day for individual transaction data set.



(A) Distribution of Visits (B) Distribution of Members

FIGURE 3.3: Distribution of Visits and Members per Parts of Day

As can be seen in Figure 3.3, the majority of individual members visit stores in the afternoons. For the whole dataset period, in the afternoons, 64,765 members make

214,090 visits. In the evenings, 44,831 individual members make 110,434 visits. In the mornings, 30,239 members make 71,186 visits.

The distribution of transactions per day for the top six leading categories is given in Figure 3.4.

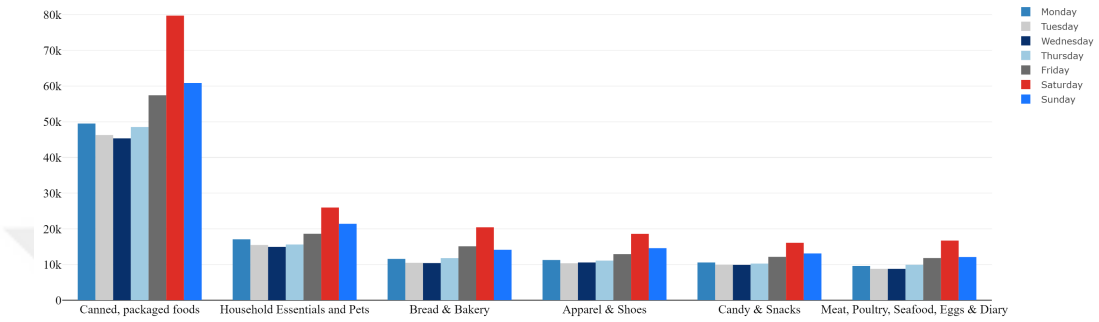


FIGURE 3.4: The distribution of Transactions per Day for top six Categories

Individual members, for buying products of "canned and packaged foods" category, make 79,725 transactions on Saturdays, 60,854 transactions on Sundays, and 57,402 on Fridays.

3.4.1.2 Business Members

This section is designed to analyze business transaction data set in a comprehensive way. The distribution of daily transactions per part of day, the distribution of visits and members per parts of day, and the distribution of transactions per day for top six categories are visualized as follows.

Figure 3.5 exhibits the distribution of transactions per part of day for each day.

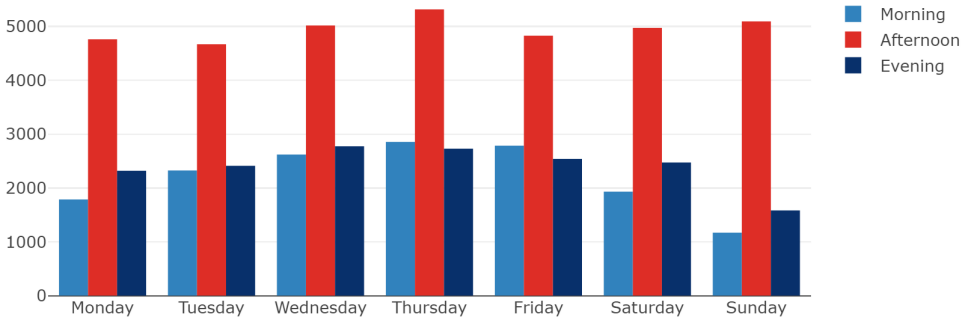
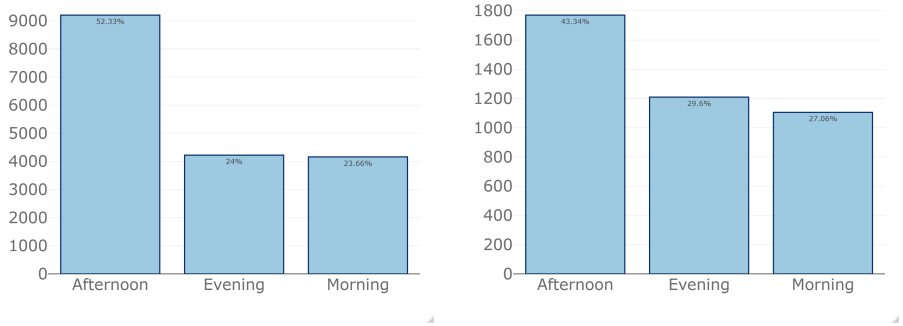


FIGURE 3.5: The distribution of Daily Transactions per Part of Day

Transactions made by business members Thursday afternoons (5,314 transactions) are leading, followed by Sunday afternoons (5,092 transactions) and Wednesday (5,051 transactions). The busiest mornings are Thursday mornings (2,854 transactions) while the busiest evening is Wednesday evenings (2,774 transactions).

Figure 3.6 exhibit the number of business members and visits per part of day for the whole dataset period.



(A) Distribution of Visits (B) Distribution of Members

FIGURE 3.6: Distribution of Visits and Members per Parts of Day

The majority of members visit in the afternoons; 1,770 business members make 9,202 visits. In the evenings, 1,209 business members make 4,221 visits. Finally, in the mornings, there are 1,105 members and 4,161 visits.

The distribution of transactions per day for the top six leading categories is given in Figure 3.7.

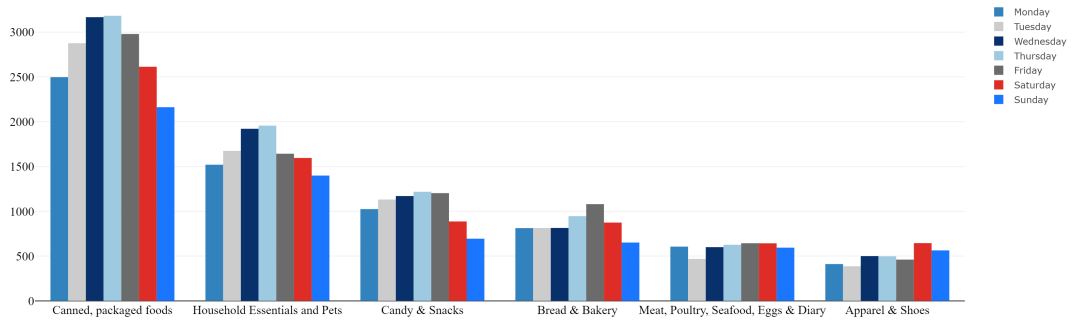


FIGURE 3.7: The distribution of Transactions per Day for top six Categories

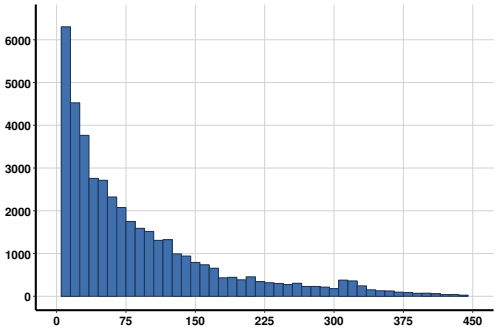
Business members make 3,181 transactions on Thursdays, 3,166 transactions on Wednesdays, and 2,978 on Fridays to buy "canned and packaged foods".

3.4.2 Customer Datasets

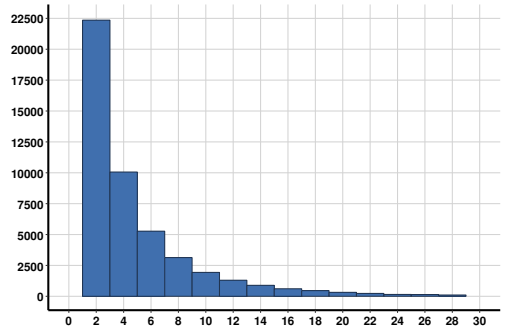
This section aims to evaluate distributions of recency, frequency, and monetary attributes, as well as correlation analysis for the purpose of deciding on attributes that are included in two-stage clustering.

3.4.2.1 Individual Members

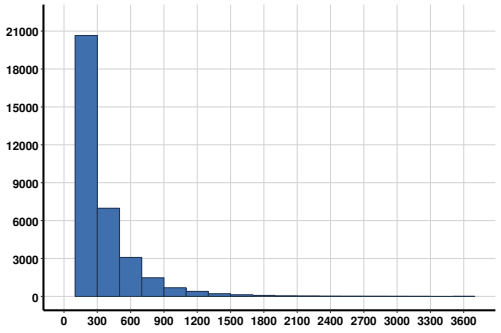
Figure 3.8 exhibits the distributions of recency, frequency, and monetary values of the individual members. The reason behind analyzing RFM attributes' distributions is that we utilized weighted RFM model in the customer lifetime value setting.



(A) Distribution of Recency



(B) Distribution of Frequency



(C) Distribution of Monetary

FIGURE 3.8: Distributions of RFM Attributes for Individual Members

Recency values of half (50.04%) of the individual members (23,526 members among 47,013 of them) are less than 50 days. 83.17% of the members have a recency value less than 150 days. On the other hand, frequency values of nearly half (47.55%) of the individual members (22,356 members) are less than 4 shopping trips. 80.17% of the members have a frequency value less than 8 trips. Last but not least, according to monetary values, 86.86% of the members (40,835 members) have an average spent which is less than $500 per visit while 97.39% of the members spent less than $1,000 per visit.

Figure 3.9 exhibits the correlation analysis results representing the mutual relationships among the attributes of the individual customer dataset.

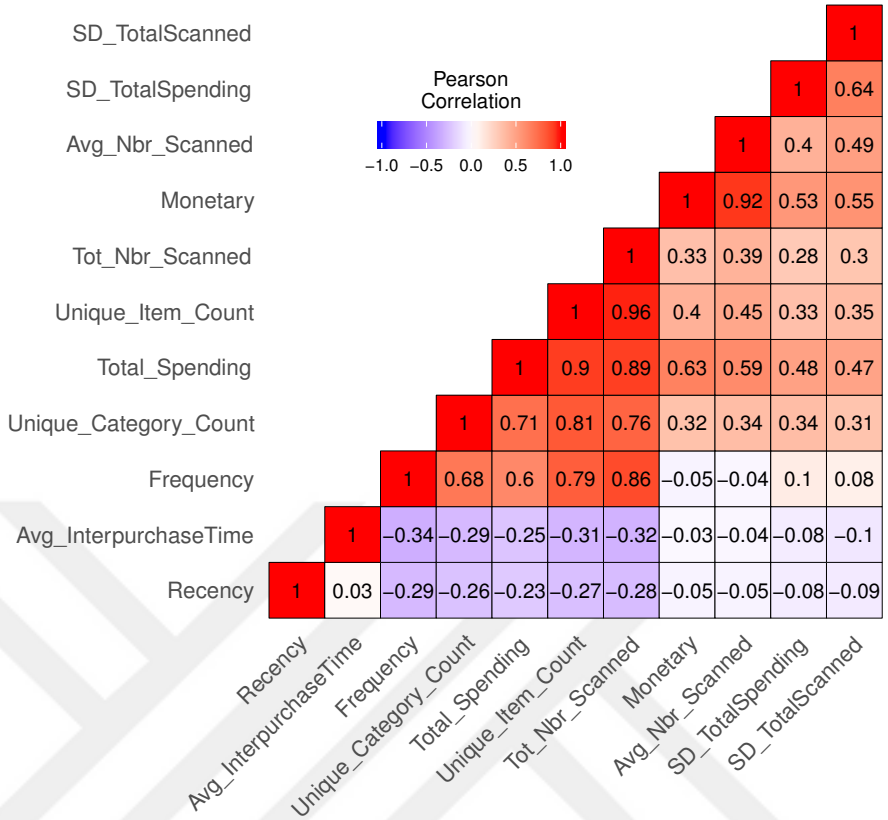


FIGURE 3.9: Correlation Matrix for Individual Customer Dataset Attributes

The important findings can be summarized as follows:

- There is a nearly perfect positive (uphill) relationship between "unique item count" and "total number of scanned items" (the correlation coefficient r is 0.96).

- There is a nearly perfect positive relationship between "monetary" and "average number of scanned items" (r = 0.92).

- There is a very strong positive relationship between "unique item count" and "total spending". (r = 0.9).

- There is a very strong positive relationship between “total number of scanned items” and “total spending” (r = 0.89).

- There is a strong positive relationship between “frequency” and “total number of scanned items” (r = 0.86).

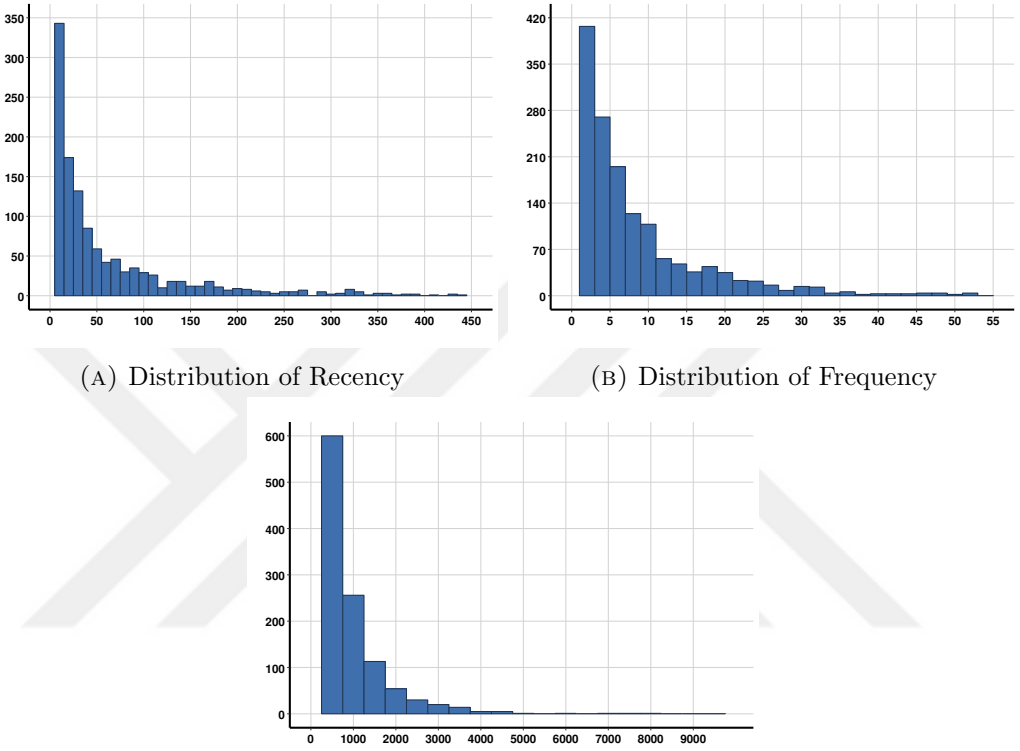
- There is a strong positive relationship between “unique category count” and “unique item count”. (r = 0.81).

- There is a strong positive relationship between “frequency” and “unique item count” (r = 0.79).

- There is a strong positive relationship between “unique category count” and “total number of scanned items” (r = 0.76).

3.4.2.2 Business Members

Figure 3.10 exhibits the recency, frequency, and monetary values of the business members.



(A) Distribution of Recency

(B) Distribution of Frequency

(C) Distribution of Monetary

FIGURE 3.10: Distributions of RFM Attributes for Business Members

Recency values of nearly half (52.54%) of the business members (i.e. 764 members among 1,454 of them) are less than 25 days. 83.84% of the members have a recency value less than 100 days. Frequency values of 37.55% of the individual members (i.e. 546 members) are less than 5 shopping trips. 81.16% of the members have a frequency value less than 15 trips. According to monetary values, 75.72% of the members (i.e. 1,101 members) have an average spent which is less than $1,000 per visit. 93.05% of the members, on the other hand, spent less than $2,000 per visit.

Figure 3.11 exhibits the correlation analysis results representing the mutual relationships among the attributes of the individual customer dataset.

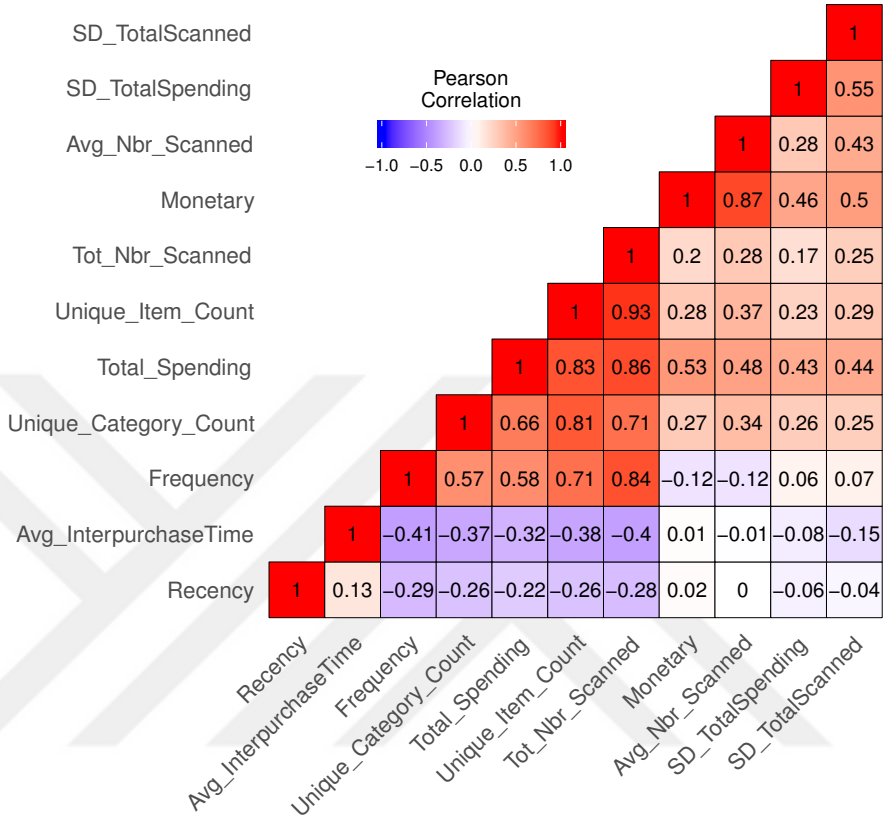


FIGURE 3.11: Correlation Matrix for Business Customer Dataset Attributes

The important results are reported as follows:

- There is a nearly perfect positive (uphill) relationship between "unique item count" and "total number of scanned items" (the correlation coefficient r is 0.93).

- There is a strong positive relationship between "monetary" and "average number of scanned items" (r = 0.87).

- There is a strong positive relationship between "total number of scanned items" and "total spending" (r = 0.86).

- There is a strong positive relationship between “frequency” and “total number of scanned items” (r = 0.84).

- There is a strong positive relationship between “unique item count” and “total spending” (r = 0.83).

- There is a strong positive relationship between “unique category count” and “unique item count” (r = 0.81).

- There is a strong positive relationship between “frequency” and “unique item count” (r = 0.71).

- There is a strong positive relationship between “unique category count” and “total number of scanned items” (r = 0.71).

Since there are strong positive relationships, we decided to use one attribute in each pair of strongly related attributes.

The next section provides the attributes which are included in two-stage clustering.

3.5 Predictive Analysis

3.5.1 Two-Stage Clustering

One of the aims of this study is to predict the purchasing behavior of the retail customers. For this purpose, we utilized cluster analysis to divide the individual and business members into distinct customer segments.

Based on the relationships among attributes (i.e. correlation analysis results given in Figures 3.9 and 3.11 and the experts' opinions, we selected the following attributes to be used in cluster analysis as seen in Table 3.11.

Table 3.11: Attributes included in clustering analysis

Attribute Name
Recency
Frequency
Total Spending
Monetary
Average Interpurchase Time
Standard Deviation of Spending
Standard Deviation of Total Scanned Items

Initially, *k-medoid* clustering method was utilized on the individual and business customer datasets where k value was specified as 15. Values of customers with respect to the above given attributes were standardized and Manhattan distance was used as the *k-medoid* distance metric.

After revealing 15 customer groups through *k-medoid* method, the average attribute values of customers in each group were computed. Subsequently, we formed a matrix where the rows represent the corresponding groups, the columns represent the selected attributes, and the entries represent the computed average values.

As a next step, we applied the *hierarchical clustering* algorithm to that matrix. Hence, we standardized the average values and used the Euclidean distance to obtain the dissimilarity matrix in *hierarchical clustering*. To construct the hierarchical model, the observations were clustered using Ward's method.

Based on the results of the *hierarchical clustering* analysis and experts' opinions, we created clusters using 15 groups coming from *k-medoid* method.

Finally, the summary statistics of each cluster were revealed, and the interpretations of statistical results were derived.

3.5.2 Unsupervised Learning

As aforementioned; in this research, we used a clustering method to group the individual and business customers and then another clustering method was used to regroup those groups of customers in order to get distinct customer segments.

The objective behind classical clustering methods is to create clusters from a set of observations by breaking the data to a certain number of groups in a way to maximize the similarities of observations in each cluster and maximize the dissimilarities of observations in different clusters.

3.5.2.1 *k-medoid* Clustering

Kaufman and Rousseeuw [39] proposed the *k-medoid* method, which is similar to the classical clustering methods. *k-medoid* divides the dataset of n observations into k clusters where the number k is specified apriori.
k-medoid method searches k representative observations (medoids) which can be defined as specific observations having the minimum average dissimilarity of all observations in their clusters.

Therefore, they can be regarded as the most centrally located observation in each cluster (i.e. minimize the distance between points assigned to the same cluster and a point specified as the medoid of the cluster). As the method minimizes the sum of pairwise dissimilarities, it provides more robust results than other classical clustering methods such as *k-means* etc. do, especially when the datasets have noise and outliers [40].

The main steps of *k-medoid* in the setting of algorithm "Partitioning around medoids (PAM)" clustering method are as follows [39]:

- For n observations $x_1, x_2, \ldots, x_n$, n(n-1)/2 dissimilarities between observations i and j (i.e. d(i, j)) are computed.

- By using a 0-1 Integer Programming model, which minimizes the total dissimilarity, the representative observations for each cluster are selected and each observation j are assigned to one of the selected representative observations.

Park and Jun [41] conducted a comparative study using *k-means* and *k-medoid* with both real and artificial data sets. In this study, it is stated that *k-medoid* outperforms *k-means*. There are three main advantages of *k-medoid*. First, *k-medoid* is based on the dissimilarities between pairs of objects, so it works well on the mixed data. Second, *k-medoid* algorithm determines representative objects as reference points, on the other hand, these points coming from *k-means* method may be unobservable. The final advantage is that *k-medoid* algorithm is less sensitive to outliers compared to *k-means* algorithm.

Velmurugan and Santhanam [42] developed a comparative study using *k-means* and *k-medoid* clustering algorithms for uniformly and normally distributed input data points. They reported that *k-means* algorithm is more efficient in smaller data sets, on the other hand, *k-medoid* algorithm outperforms in larger data sets.

Clustering results are subjective and dependent on implementation. There are several criteria to specify the quality of clustering results. First, the similarity measure for clustering method and its implementation has an effect on quality of clustering results. Second, the extent to which clustering algorithm is capable of exploration some or all hidden patterns, and finally, the definition of clusters and representation are important for clustering results evaluation [42].

One of the most important advantage for *k-medoid* clustering algorithm is that a medoid is the most centrally located object within a cluster as a reference point. In *k-means* clustering algorithm, the mean value of the objects within a cluster is used as a reference point. Therefore, *k-means* algorithm is more sensitive to the outliers. However, partitioning method in *k-medoid* can outperform because it aims to minimize the sum of dissimilarities between each object and corresponding reference point [42].

3.5.2.2 Hierarchical Clustering

Number of clusters is not known a priori in *hierarchical clustering* [43]. After utilizing the analysis, a tree-like visual representation of the observations called dendrogram is revealed. Dendrogram allows researcher to view at once the clustering of n observations obtained for each possible number of clusters, from 1 to n.

There are two approaches for *hierarchical clustering*: agglomerative (bottom-up) and divisive (top-down).

The steps of the bottom-up approach used in this research are as follows:

- It starts with assigning each observation to its own cluster.

- The closest two clusters are identified and then merged.

- If all observations are in a single cluster, then it stops, else the previous step is repeated.

- A dendrogram representing these iterative steps is revealed.

One of the criteria used in *hierarchical clustering* as well as in this research is Ward's method. For agglomerative *hierarchical clustering*, Ward [44] proposed to use an objective function of the error sum of squares that will to be minimized when selecting which pair of clusters should be merged at an iterative step.

Ward method is more complex compared to other methods such as single-linkage, complete-linkage and average linkage used in *hierarchical clustering*. However, it can be reported as more accurate method minimizing the variance between elements [45]. Ward method allows also minimize total within-cluster variance, in other words, maximize between-cluster variance.

3.6 The Clustering Results

3.6.1 Clustering Results for Individual Members

We came up with individual customer groups after utilizing *k-medoid* method on the individual customer dataset by standardizing values, using Manhattan distance metric, and specifying k value as 15.

Figure 3.12 exhibits the average values of individual customers in each group with respect to attributes.

Group	Total Members	Percentage	Average Recency	Average Frequency	Average Total Spending	Average Monetary	Average IPT	Average SD Total Spending	Average SD Total Scanned Items
1	3836	8.16%	110.39	2.63	230.69	93.65	27.42	10.09	1.28
2	2334	4.96%	27.09	15.94	2228.81	146.07	20.76	26.75	3.88
3	4402	9.36%	31.56	7.3	874.82	126.99	35.79	21.07	2.98
4	2904	6.18%	157.14	3.28	444.42	140.58	78.24	20.42	2.71
5	2810	5.98%	97.87	3.05	1691.64	571.74	55.4	72.08	11.7
6	5035	10.71%	30.63	3.32	315.2	102.15	59.65	14.01	1.8
7	3716	7.90%	75.16	3.09	899.65	302.71	77.72	47.1	6.94
8	4611	9.81%	75.25	3.24	655.01	214.4	27.89	30.35	4.71
9	3682	7.83%	48.13	7.2	1699.53	247.34	36.45	40.17	5.19
10	1865	3.97%	25.44	17.4	5064.63	306.79	19.28	43.4	6.3
11	1500	3.19%	43.6	8.29	5317.76	672.92	29.99	61.83	9.48
12	2704	5.75%	46.3	6.29	2656.08	450.34	34.2	51.79	7.66
13	1365	2.90%	68.23	3.1	3861.53	1296.52	63.65	56.54	9.08
14	3419	7.27%	71.14	2.36	322.54	137.84	160.01	17.88	2.37
15	2830	6.02%	312.77	2.9	462.91	164.26	40.75	23.13	2.79

FIGURE 3.12: Average Values of Individual Customers in each Group

Then, we revealed a *hierarchical clustering* dendrogram as shown in Figure 3.13 utilizing *hierarchical clustering* using Ward's method on the matrix given in Figure 3.12 by standardizing the average values and using Euclidean distance metric.

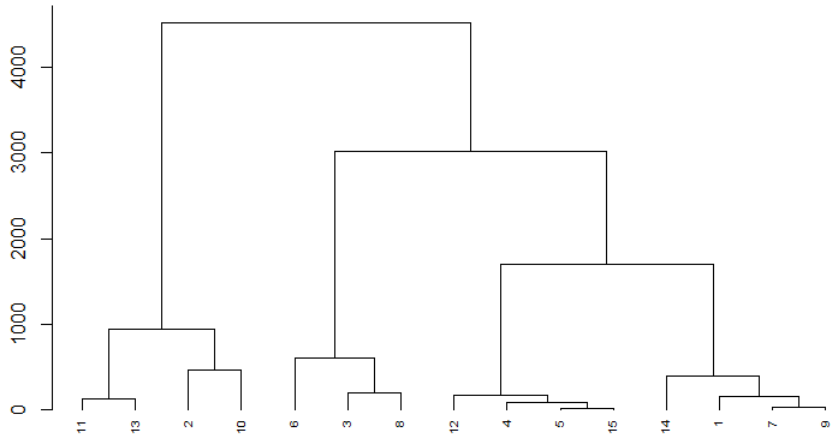


FIGURE 3.13: Hierarchical Clustering Dendrogram for Individual Customer Groups

Based on the clustering dendrogram and experts' opinion, we achieved eight clusters for individual customers as can be seen in Figure 3.14.

Cluster	Group	Total Members	Percentage	Average Recency	Average Frequency	Average Total Spending	Average Monetary	Average IPT	Average SD Total Spending	Average SD Total Scanned Items
1	11	2865	6.09%	43.6	8.29	5317.76	672.92	29.99	61.83	9.48
	13			68.23	3.1	3861.53	1296.52	63.65	56.54	9.08
2	2	4199	8.93%	27.09	15.94	2228.81	146.07	20.76	26.75	3.88
	10			25.44	17.4	5064.63	306.79	19.28	43.4	6.3
3	6	5035	10.71%	30.63	3.32	315.2	102.15	59.65	14.01	1.8
4	3	9013	19.17%	31.56	7.3	874.82	126.99	35.79	21.07	2.98
	8			75.25	3.24	655.01	214.4	27.89	30.35	4.71
5	12	2704	5.75%	46.3	6.29	2656.08	450.34	34.2	51.79	7.66
6	4	8544	18.17%	157.14	3.28	444.42	140.58	78.24	20.42	2.71
	5			97.87	3.05	1691.64	571.74	55.4	72.08	11.7
	15			312.77	2.9	462.91	164.26	40.75	23.13	2.79
7	14	3419	7.27%	71.14	2.36	322.54	137.84	160.01	17.88	2.37
8	1	11234	23.90%	110.39	2.63	230.69	93.65	27.42	10.09	1.28
	7			75.16	3.09	899.65	302.71	77.72	47.1	6.94
	9			48.13	7.2	1699.53	247.34	36.45	40.17	5.19

FIGURE 3.14: Assignment of Individual Customer Groups to Clusters

The average values of individual customers in each cluster with respect to attributes are given in Figure 3.15.

Cluster	Recency	Frequency	Total Spending	Monetary	Inter-purchase Time	Standard Deviation Spending	Standard Deviation Total Scanned
1	55.33	5.82	4623.95	970.03	46.03	59.31	9.29
2	26.36	16.59	3488.35	217.45	20.1	34.14	4.95
3	30.63	3.32	315.2	102.15	59.65	14.01	1.8
4	53.91	5.22	762.36	171.71	31.75	25.82	3.87
5	46.3	6.29	2656.08	450.34	34.2	51.79	7.66
6	189.2	3.08	860.74	290.22	58.31	38.31	5.7
7	71.14	2.36	322.54	137.84	160.01	17.88	2.37
8	78.33	4.28	933.39	213.18	47.02	32.19	4.43

FIGURE 3.15: Average Values of Individual Customers in each Cluster

Based on the average values of customers with respect to attributes in Figure 3.15, the following findings are revealed:

- Due to the fact that the lower the recency, the better the cluster is; Clusters 2 and 3 can be considered as the best clusters according to their average recency values. The number of days between the end of dataset period and the last purchase of the members in these clusters are 26.36 and 30.63 days on average, respectively. On the other hand, Cluster 6 is the worst cluster. The members in this cluster have not shopped for the last 189.2 days on average.

- Average frequency value of Cluster 2 is the highest during the dataset period, the members in Cluster 2 make much more shopping visits (16.59 times on average) than the members in other clusters. On the other hand, Clusters 3, 6 and 7 are the worst clusters based on the average frequency.

- Regarding to average total spending amounts, Clusters 1 is leading, followed by Clusters 2 and 5. During the data set period, the members in these clusters spend on average $4,623.95, $3,488.35, and $2,656.08, respectively. The worst average total spending amounts belong to Clusters 3 and 7, which are just spending $315.20 and $322.54, respectively.

- As the monetary attribute refers to the average spending amount per visit, we can say that the members in Cluster 1 spend as much as $970.03 per visit on average. This value makes Cluster 1 the leading cluster with a great difference. Clusters 3, 4, and 7 are the worst clusters with the lowest monetary values.

- Similar to the recency case, the lower the average interpurchase time, the better the cluster is. Hence, Cluster 2 becomes the best cluster followed by Clusters 4 and 5. The average time interval between two consecutive shopping visits of the members in Cluster 2 is just 20.1 days on average. On the other hand, Cluster 7 becomes the worst cluster with an average value of 160.01 days.

- When standard deviation of spending values attribute, i.e. the variation in the spending of customers during their shopping trips, is taken into consideration, we can conclude that the average variation in the spending of the members in Clusters 3 and 7 are low while the average variation in the spending of the members in Clusters 1 and 5 are high.

- According to the standard deviation of the total scanned items attribute, i.e. the variation in the total number of items purchased by customers in their shopping trips, we can conclude that the average variation in the number of items for Clusters 3 and 7 are low while the average variation in the number of items for Clusters 1 and 5 are high.

3.6.2 Clustering Results for Business Members

We came up with business customer groups after utilizing *k-medoid* method on the business customer dataset by standardizing values, using Manhattan distance metric, and specifying k value as 15.

Figure 3.16 exhibits the average values of business customers in each group with respect to attributes.

Group	Total Members	Percentage	Average Recency	Average Frequency	Average Total Spending	Average Monetary	Average IPT	Average SD Total Spending	Average SD Total Scanned Items
1	110	7.57%	27.63	4.08	841.04	224.69	57.22	23.66	2.67
2	73	5.02%	50.52	6.26	16063.99	2881.26	32.97	184.6	21.05
3	82	5.64%	85.85	2.93	1318.36	494.5	114.37	52.98	4.57
4	130	8.94%	28.55	7.92	6053	883.19	21.29	88.07	11.33
5	72	4.95%	39.04	6.6	4925.07	822.79	36.48	177.99	7.62
6	98	6.74%	20.46	14.1	18574.2	1395.66	22.57	115.33	14.03
7	86	5.91%	37.88	3.88	3002.34	888.6	77.68	84.26	7.51
8	55	3.78%	283.11	4.05	3527.75	984	32.62	93.23	10.92
9	118	8.12%	32.42	4.36	969.91	257.9	20.79	32.51	3.71
10	106	7.29%	44.32	5.74	5942.67	1189.53	43.68	126.35	19.36
11	102	7.02%	12.16	29.67	17600.33	628.1	12.81	91.97	10.7
12	98	6.74%	178.28	3.76	1232.68	358.57	38.68	40.51	4.35
13	98	6.74%	13.82	23.59	6469.57	292.34	16.76	59.05	6.65
14	126	8.67%	29.48	5.21	3193.5	673.59	37.35	56.4	7.95
15	100	6.88%	23.08	11.29	4045.56	366.74	26.08	52.05	5.94

FIGURE 3.16: Average Values of Business Customers in each Group

Then, we revealed a *hierarchical clustering* dendrogram, as shown in Figure 3.17, utilizing hierarchical clustering using Ward's method on the matrix given in Figure 3.16 by standardizing the average values and using Euclidean distance metric.

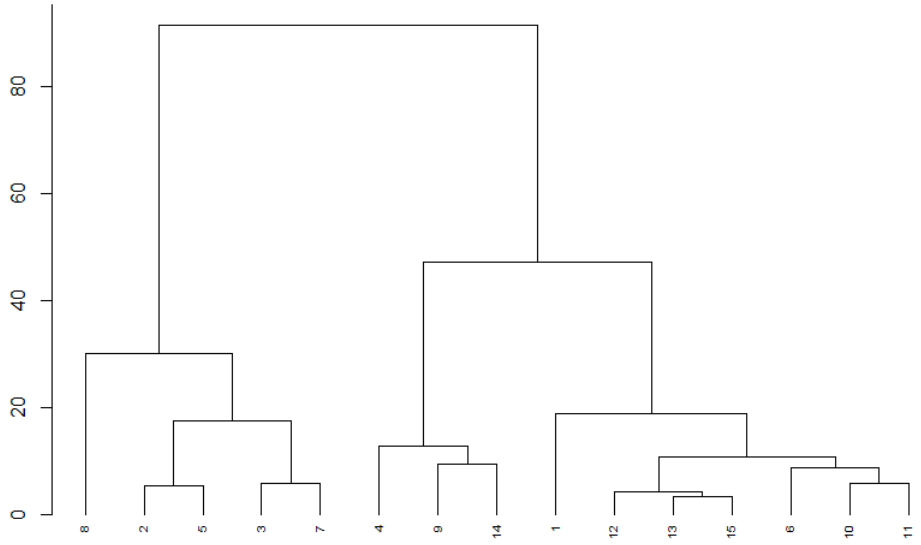


FIGURE 3.17: Hierarchical Clustering Dendrogram for Business Customer Groups

Based on the clustering dendrogram and the experts' opinion, we created six clusters for business customers as can be seen in Figure 3.18.

Cluster	Group	Total Members	Percentage	Average Recency	Average Frequency	Average Total Spending	Average Monetary	Average IPT	Average SD Total Spending	Average SD Total Scanned Items
1	6	306	21.05%	20.46	14.1	18574.2	1395.66	22.57	115.33	14.03
	10			44.32	5.74	5942.67	1189.53	43.68	126.35	19.36
	11			12.16	29.67	17600.33	628.1	12.81	91.97	10.7
2	12	296	20.36%	178.28	3.76	1232.68	358.57	38.68	40.51	4.35
	13			13.82	23.59	6469.57	292.34	16.76	59.05	6.65
	15			23.08	11.29	4045.56	366.74	26.08	52.05	5.94
3	1	110	7.57%	27.63	4.08	841.04	224.69	57.22	23.66	2.67
4	4	374	25.72%	28.55	7.92	6053	883.19	21.29	88.07	11.33
	9			32.42	4.36	969.91	257.9	20.79	32.51	3.71
	14			29.48	5.21	3193.5	673.59	37.35	56.4	7.95
5	3	223	15.34%	85.85	2.93	1318.36	494.5	114.37	52.98	4.57
	7			37.88	3.88	3002.34	888.6	77.68	84.26	7.51
	8			283.11	4.05	3527.75	984	32.62	93.23	10.92
6	2	145	9.97%	50.52	6.26	16063.99	2881.26	32.97	184.6	21.05
	5			39.04	6.6	4925.07	822.79	36.48	177.99	7.62

FIGURE 3.18: Assignment of Business Customer Groups to Clusters

The average values of business customers in each cluster with respect to attributes are given in Figure 3.19.

Cluster	Recency	Frequency	Total Spending	Monetary	Inter-purchase Time	Standard Deviation Spending	Standard Deviation Total Scanned
1	25.96	16.39	13873.95	1068.40	26.63	111.36	14.77
2	71.40	12.87	3916.81	339.40	27.17	50.55	5.65
3	27.63	4.08	841.04	224.69	57.22	23.66	2.67
4	30.09	5.88	3485.88	615.29	26.54	59.87	7.78
5	116.00	3.57	2512.70	767.21	80.06	74.97	7.27
6	44.82	6.43	10532.94	1859.13	34.71	181.32	14.38

FIGURE 3.19: Average Values of Individual Customers in each Cluster

Based on the average values of customers with respect to attributes in Figure 3.19, the following findings are revealed:

- As aforementioned, Clusters 1 and 3 can be considered as the best clusters according to their low recency values. The number of days between the end of dataset period and the last purchase of the members in these clusters are 25.96 and 27.63 days on average, respectively. On the other hand, Cluster 5 is the worst cluster. The members in this cluster have not shopped from Sam's Club for the last 116 days on average.

- According to the average frequency values, Clusters 1 and 2 are the best clusters. The members in these clusters make 16.39 and 12.87 shopping visits on

average during the dataset period. Cluster 5 is the worst cluster based on average frequency.

- In terms of the average total amount spent, Clusters 1 has the highest value, followed by Cluster 6. During the data set period, the members in these clusters spend $13,873.95 and $10,532.94 on average, respectively. The worst average total spending amount belongs to Cluster 3 with a spending of just $841.04.

- Taking into account monetary values, we can reveal that Cluster 6 is the best cluster, followed by Cluster 1. The members in these clusters spend $1,859.13 and 1,068.40 per visit on average, respectively. On the other hand, Cluster 2 and Cluster 3 are the worst clusters with the lowest monetary values.

- With respect to average interpurchase time; Clusters 1, 2, and 4 become the best clusters as the average time interval between two consecutive shopping visits of the members in these clusters are as low as approximately 27 days. On the hand, Cluster 5 is the worst cluster with an average value of 80.06 days.

- Taking into account standard deviation of spending values attribute, i.e. the variation in the spending of customers in their shopping trips, we can conclude that the average variation in the spending of the members in Clusters 3 is low while the average variation in the spending of the members in Cluster 6 is high.

- According to the standard deviation of total scanned items attribute, i.e. the variation in the total number of items purchased by customers in their shopping trips, we can conclude that the average variation in the number of items for

Cluster 3 is low while the average variation in the number of items for Clusters 1 and 6 is high.

3.7 Customer Lifetime Value (CLV)

After clustering individual and business members of Sam's Club into customer segments, we aimed to estimate customer lifetime value (CLV) of the clusters based on the purchasing behaviors of their members within a certain time period from 7/31/2005 through 11/2/2006.

In order to compute CLV value for each cluster specified at the predictive analytics stage, we used a weighted RFM (recency, frequency, monetary) model which is explained in detail below.

CLV can be defined as "the present value of the future cash flows attributed to the customer relationship" [46]. It is important to highlight that the CLV knowledge is important for companies to determine which customer segments are more profitable and loyal. Companies can use CLV as a metric for the assessment of different segments of customers to develop efficient and appropriate marketing and sales strategies from both financial and operational perspectives [47].

There are several studies which propose a marketing analysis method using recency, frequency, and monetary (RFM) variables to compute CLV value for each customer segment [28], [47]. Some of them used "weighted RFM model" by assessing importance of recency, frequency, and monetary variables [28], [29], [48], and [49].

3.7.1 The Weighted RFM Model

3.7.1.1 The Model based on Subjective Weights

To assess the relative weights of RFM variables, we assessed judgements of the experts by asking pairwise comparison questions, formed a pairwise comparison matrix using those judgements, and used the values of the eigenvector as relative weights extracted from that matrix. Assessing weights in this manner is called as using AHP in the literature [28],[29], [47], and [49].

Accordingly, we conducted a questionnaire survey. As can be seen in Figure 3.20, at the survey questionnaire, we asked questions in pairwise comparison manner using a nine-point scale suggested by Saaty [50] to assess judgments of the experts concerning the relative priorities (weights) of the variables.

With respect to goal "**calculating CLV for each cluster of members**" compare the following variable pairs:

1=Equal importance 3=Moderately more important 5=Strongly more important
7=Very strongly more important 9=Extremely more important

Recency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Frequency
Frequency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Monetary
Monetary	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Recency

FIGURE 3.20: Pairwise Comparison Questions

Six marketing professors[2] responded the questionnaire.

They chose any value between 2 and 9 at the left-hand side of 1 in the scale if they thought that the first variable is more important than the second one. On the other hand, if they thought that the second variable is more important than the first one,

[2]Prof. Dr. Şebnem Burnaz (ITU), Prof. Dr. Nimet Uray (Kadir Has Univ.) Prof. Dr. Banu Elmadağ Baş (ITU), Prof. Dr. Cenk Kocaş (Sabanci Univ.), Assoc. Prof. Dr. Elif Karaosmanoğlu (ITU), Asst. Prof. Dr. Kıvırcım Döğerlioğlu Demir (Sabancı Univ.)

they chose any value between 2 and 9 at the right-hand side of 1 in the scale. If they believed that both variables had exactly the same importance, then they picked 1 in the middle.

We computed the geometric means of all paired-comparison judgments of different respondents for each question in order to reveal the aggregated group judgments. For this purpose, we utilized inverse values if right-hand side number was selected and the number itself if it is at the left-hand side. Group judgments (i.e. geometric means), were then arranged in an aggregated pairwise comparison matrix. In the matrix, the value for an (i, j)-pair is in the range 1–9 if variable i is more important than variable j. The value is in the range 1–1/9 if variable j is more important than variable i. This matrix is a reciprocal matrix, in other words given the (i, j)-value, the corresponding (j, i)-value will be the inverse of the (i, j)-value.

The aggregated pairwise-comparison matrix is shown at Table 3.12.

Table 3.12: The Aggregated Pairwise Comparison Matrix

	R	**F**	**M**
R	1	0.3052	0.4569
F	3.2767	1	0.6788
M	2.1886	1.4731	1

The relative importance of each variable was computed at the next step. For this purpose, the eigenvector of the pairwise comparison matrix was extracted. As proposed by Saaty [50], the relative weights of the variables are the corresponding values at the eigenvector. The easiest way for the computation of eigenvector starts with the normalization of the pairwise comparison matrix (i.e. dividing each element by its column sum) so that each column adds to one. The arithmetic mean of the values of each row in the normalized matrix is an element of the eigenvector.

The calculated eigenvector (i.e. the relative weight vector) is given in Table 3.13.

TABLE 3.13: Relative Weight of RFM Variables

	$\mathbf{w}^{a}$
Recency	15.95 %
Frequency	39.49 %
Monetary	44.57 %

[a] w: weight

3.7.1.2 The Model based on Objective Weights

In this section, we utilized multiple regression analysis with forward stepwise selection in order to assess the relative objective weights of RFM variables.

Forward stepwise selection starts a model with no predictors and it iterates by adding predictors into the model one by one until all independent variables, in other words predictors, are included in the model. The most important principle in forward stepwise selection is that the predictor which contribute the highest additional improvement to the fit is included into the model [51].

To perform multiple regression analysis, initially, we splitted our overall customer data set, which comprised of individual and business members, into train and test data sets. Our dependent variable in multiple regression analysis was "Total Spending". After the forward stepwise selection was applied, we calculated relative importance for each of the predictors, which is known as sharply value in marketing area, within the model. After obtaining the relative importance of predictors, the relative weights of RFM variables were normalized.

The corresponding weights extracted from regression output are given in Table 3.14.

TABLE 3.14: Relative Weight of RFM Variables

	$\mathbf{w}^{a}$
Recency	5.3 %
Frequency	44.7 %
Monetary	50.0 %

[a] w: weight

3.7.1.3 The Aggregated Model

In this section, the subjective weights coming from the experts' judgments and objective weights revealed from regression analysis were aggregated by using the simple average method.

The aggregated relative weights for RFM variables are given in Table 3.15.

TABLE 3.15: Relative Weight of RFM Variables

	$\mathbf{w}^{a}$
Recency	10.63 %
Frequency	42.09 %
Monetary	47.28 %

[a] w: weight

For the following section, we created decision matrices with the averages of recency, frequency, and monetary attributes. Then, we utilized simple additive weighting to compute CLV scores with three different weights: subjective, objective and aggregated weights.

3.8 The Decision Matrices

In the next stage, we constructed two decision matrices, one for individual customers and the other for business customers. Matrices have clusters in the rows; recency, frequency, and monetary variables in the columns; and average RFM values of the customers in each cluster are the entries. The values of these RFM matrices were extracted from Figures 3.15 and 3.19.

Tables 3.16 and 3.17 exhibit the decision matrices for clusters of individual and business customers.

TABLE 3.16: The Average RFM Values of Individual Customers in each Cluster

Cluster	Recency	Frequency	Monetary
1	55.33	5.82	970.03
2	26.36	16.59	217.45
3	30.63	3.32	102.15
4	53.91	5.22	171.71
5	46.3	6.29	450.34
6	189.2	3.08	290.22
7	71.14	2.36	137.84
8	78.33	4.28	213.18

TABLE 3.17: The Average RFM Values of Business Customers in each Cluster

Cluster	Recency	Frequency	Monetary
1	25.96	16.39	1068.4
2	71.4	12.87	339.4
3	27.63	4.08	224.69
4	30.09	5.88	615.29
5	116.0	3.57	767.21
6	44.82	6.43	1859.13

3.9 Simple Additive Weighting

In order to compute the CLV scores of the clusters; obtaining a global (total) score by adding weighted contributions from each variable is frequently preferred in the literature [29], [32], and [33].

We used Simple Additive Weighting (SAW) method, which is a performance aggregation-oriented method, it is also called as Weighted Average or Weighted Sum method [52]. SAW uses linear normalization values based on Tchebycheff distance as contributions from variables.

A common numerical scaling system such as normalization is required to permit addition among attribute values. Normalization allows obtaining dimensionless units as well as comparable scales, which allow inter-variable and intra-variable comparisons, where the larger value becomes, the more preference it has.

$$V(a_i) = \sum_{j=1}^{n} w_j r_{ij} \tag{3.1}$$

where r_{ij} denotes normalized value of row element i with respect to column element j and n is the number of column elements, which are variables. The normalized value is defined as the ratio between individual distance and combined distance of a row element from the origin point.

Therefore, the normalized value of x_{ij} (performance value of row element i with respect to column element j) is given as [53]:

$$r_{ip}(p) = \frac{(x_{ij} - 0)}{(\sum_{k=1}^{m} [x_{kj} - 0]^p)^{\frac{1}{p}}} \tag{3.2}$$

where p indicates the distance type (that is to say, p=1 for Manhattan distance, p=2 for Euclidean distance, and p=∞ for Tchebycheff distance) and m is the number of row elements. The larger normalized value becomes, the more preference it has.

Therefore, if the column element is a cost attribute (i.e. lower value of the row element with respect to that column element is better), before using Equation 3.2, we should convert values at that column element by taking inverse values ($1/x_{ij}$), thus converting cost attribute to benefit attribute (i.e. higher value is better).

If the Tchebycheff distance is selected, Equation 3.2 becomes:

$$r_{ij} = \frac{x_{ij}}{max_k x_{kj}} \tag{3.3}$$

If the column element is a cost attribute then taking inverse value makes Equation 3.3 as follows:

$$r_{ij} = \frac{min_k x_{kj}}{x_{ij}} \tag{3.4}$$

In the decision matrices given in Tables 3.16 and 3.17, RFM variables have different units: days for recency, times for frequency, and dollars for monetary. Besides, managers prefer higher values of frequency and monetary and lower values of recency. Thus, normalization was needed for revealing global scores in our case. Finally, CLV scores of clusters were computed by multiplying global scores assessed through SAW by 100.

3.9.1 The Results based on Subjective Weights

3.9.1.1 Results for Individual Members

After utilizing SAW method on RFM matrix for individual members, we came up with the normalized matrix and the CLV scores of the clusters shown in Table 3.18.

Table 3.18: Normalized RFM Matrix and CLV Scores

Weight	**15.9 %**	**39.49 %**	**44.57 %**	
Cluster	**Recency**	**Frequency**	**Monetary**	**CLV**
1	0.476	0.351	1.000	66.02
2	1.000	1.000	0.224	65.43
5	0.569	0.379	0.464	44.74
4	0.489	0.315	0.177	28.11
3	0.861	0.200	0.105	26.32
8	0.337	0.258	0.220	25.35
6	0.139	0.186	0.299	22.89
7	0.371	0.142	0.142	17.86

As can be seen in Table 3.18, Clusters 1 and 2 have the highest CLV, 66.02 and 65.43, respectively. Cluster 5 with a CLV of 44.74 follows them. Other clusters (Clusters 4, 3, 8, 6, and 7) have low CLV scores less than 30.

3.9.1.2 Results for Business Members

Similarly, after using the SAW method on RFM matrix for business members, we came up with the normalized matrix and the CLV scores of the clusters shown in Table 3.19.

As can be seen in Table 3.19, Clusters 1 with a CLV of 81.05 is the leading cluster while Cluster 6 with a CLV of 69.29 follows it. Then, Clusters 2 and 4 take place with CLV scores of 44.94 and 42.67, respectively. Clusters 5 and 3 have the lowest CLV.

Table 3.19: Normalized RFM Matrix and CLV Scores

Weight	**15.9 %**	**39.49 %**	**44.57 %**	
Cluster	**Recency**	**Frequency**	**Monetary**	**CLV**
1	1.000	1.000	0.575	81.05
6	0.579	0.392	1.000	69.29
2	0.364	0.785	0.183	44.94
4	0.863	0.359	0.331	42.67
5	0.224	0.218	0.413	30.56
3	0.940	0.249	0.121	30.20

3.9.2 The Results based on Objective Weights

3.9.2.1 Results for Individual Members

When the objective weights were utilized in the calculation of CLV scores, the following CLV scores for individual clusters were obtained as shown in Table 3.20.

Table 3.20: Normalized RFM Matrix and CLV Scores

Weight	**5.3 %**	**44.7 %**	**50.0 %**	
Cluster	**Recency**	**Frequency**	**Monetary**	**CLV**
1	0.476	0.351	1.000	68.21
2	1.000	1.000	0.224	61.21
5	0.569	0.379	0.464	43.18
4	0.489	0.315	0.177	25.51
8	0.337	0.258	0.220	24.30
6	0.139	0.186	0.299	24.00
3	0.861	0.200	0.105	18.77
7	0.371	0.142	0.142	15.43

Based on the statistics in Table 3.20, Clusters 1 and 2 have the highest CLV, 68.21 and 61.21, respectively. Cluster 5 with a CLV of 43.18 follows them. Other clusters (Clusters 4, 8, 6, 3 and 7) have low CLV scores less than 30.

3.9.2.2 Results for Business Members

When the objective weights were utilized in the calculation of CLV scores, the following CLV scores for business clusters were obtained as shown in Table 3.21.

Table 3.21: Normalized RFM Matrix and CLV Scores

Weight	**5.3 %**	**44.7 %**	**50.0 %**	
Cluster	**Recency**	**Frequency**	**Monetary**	**CLV**
1	1.000	1.000	0.575	78.73
6	0.579	0.392	1.000	70.61
2	0.364	0.785	0.183	46.15
4	0.863	0.359	0.331	37.16
5	0.224	0.218	0.413	31.56
3	0.940	0.249	0.121	22.15

Based on the statistics given in Table 3.21, Clusters 1 with a CLV of 78.73 is the leading cluster while Cluster 6 with a CLV of 70.61 follows it. Then, Clusters 2 and 4 take place with CLV scores of 46.15 and 37.16, respectively. Clusters 5 and 3 have the lowest CLV with the values of 31.56 and 22.15 respectively.

3.9.3 The Results based on Aggregated Weights

3.9.3.1 Results for Individual Members

CLV scores of individual members with aggregated weights are given in Table 3.22.

Table 3.22: Normalized RFM Matrix and CLV Scores

Weight	**10.63 %**	**42.09 %**	**47.28 %**	
Cluster	**Recency**	**Frequency**	**Monetary**	**CLV**
1	0.476	0.351	1.000	67.11
2	1.000	1.000	0.224	63.32
5	0.569	0.379	0.464	43.96
4	0.489	0.315	0.177	26.81
8	0.337	0.258	0.220	24.83
6	0.139	0.186	0.299	23.44
3	0.861	0.200	0.105	22.55
7	0.371	0.142	0.142	16.64

Based on the aggregated model results for individual members as seen in Table 3.22, Cluster 1 is the leading cluster with CLV of 67.11, followed by Cluster 2 63.32. Cluster 5 is ranked as third with CLV of 43.96. The remaining clusters, which are Cluster 4, 8, 6, 3, and 7 have CLV values under 30.

3.9.3.2 Results for Business Members

Table 3.23 exhibits CLV scores of business members with aggregated weights.

Table 3.23: Normalized RFM Matrix and CLV Scores

Weight	**10.63 %**	**42.09 %**	**47.28 %**	
Cluster	**Recency**	**Frequency**	**Monetary**	**CLV**
1	1.000	1.000	0.575	79.89
6	0.579	0.392	1.000	69.95
2	0.364	0.785	0.183	45.55
4	0.863	0.359	0.331	39.92
5	0.224	0.218	0.413	31.06
3	0.940	0.249	0.121	26.18

As can be sen in Table 3.23, based on the aggregated model results for business members, Cluster 1 is the leading cluster with CLV of 79.89, followed by Cluster 6

69.95. Cluster 2 is ranked as third with CLV of 45.55, followed by Cluster 3 with CLV of 39.92. Cluster 5 has a CLV of 31.06 as the fifth ranked cluster. The lowest CLV value in business clusters belongs to Cluster 3 with CLV of 26.18.

3.9.4 Comparison of CLV Scores

Table 3.24 shows the comparison of CLV scores for individual clusters based on subjective, objective and aggregated weights.

TABLE 3.24: Comparison Table of CLV for Individual Members

Cluster No	Subjective	Objective	Aggregated
1	66.02	68.21	67.11
2	65.43	61.21	63.32
3	26.32	18.77	43.96
4	28.11	25.51	26.81
5	44.74	43.18	24.83
6	22.89	24.00	23.44
7	17.86	15.43	22.55
8	25.35	24.3	16.64

The rankings of Cluster 1 and 2 are the same in these three scenarios. One of the remarkable changes happens for Cluster 3, although it has relatively low CLV scores in both subjective and objective weights' cases, it is ranked as third valuable cluster by using aggregated weights. Similar situation appears for Cluster 5. Cluster 5 has relatively high CLV scores based on subjective and objective weights' cases, but its CLV score and corresponding ranking among all clusters decreases dramatically when we utilize aggregated weights. There is a relatively increase in the CLV score of Cluster 7 when aggregated weights are used. On the other hand, there is a relatively decrease in CLV score of Cluster 8 in case of aggregated weights' usage. There is almost no change in Cluster 4 CLV score.

Table 3.25 shows the comparison of CLV scores for business clusters based on subjective, objective and aggregated weights.

TABLE 3.25: Comparison Table of CLV for Business Members

Cluster No	Subjective	Objective	Aggregated
1	81.05	78.73	79.89
6	69.29	70.61	69.95
2	44.94	46.15	45.55
4	42.67	37.16	39.92
5	30.56	31.56	31.06
3	30.20	22.15	26.18

Cluster 1 is the leading cluster in all three cases. There are no changes in CLV scores and corresponding ranking in Cluster 2, 5 and 6. When we evaluate overall ranking, there is no change in ranking of business clusters in terms of CLV scores.

We compared our subjective approach by utilizing objective and aggregated weights when we calculated CLV scores of clusters. We can say that the top two clusters in both individual members and business members do not change in ranking. This is important for us, since we are planning to deal with top two clusters in both individuals and business members in product network stage.

Chapter 4

Product Network Analysis using Hubs and Authorities (HITS)

Basically, a network can be defined as a set of elements interconnected with each other [54]. Graphs are one of the most common techniques to visualize networks. There are set of elements, which refer to vertices, with the pairs of them connected by links, which refer to edges, inside the network [55]. It is important to note that graphs and networks form structural models, which can be accepted a convenient form of networks for us to analyze how various schemes act together. In the product network setting, a vertex refers to a product and an edge refers to a relationship between products [54].

In this section, we aimed to create appropriate product networks for both the top two individual and business clusters. Due to the fact that the remaining clusters have relatively low customer lifetime value, we decided to create two general product networks by including all business clusters for business product network and all individual clusters for individual product network. Based on the general product networks, one developed for individual and the other for business, we obtained valuable insights from *generated patterns* inside the networks.

One of the most important aim for us is to discover the *cross-selling* effects between items which are included in HITS model, and to find hubs in the transaction data set. Also, we target to identify recurring purchasing patterns, complement, substitute and trigger products within the networks.

We used the HITS algorithm in order to perform product network analysis. The most important difference between HITS algorithm and classical association rule mining is that each transaction has different weights instead of equal weight assumption used in association rule mining [11]. This is important for practitioners in real-life application in terms of emphasizing the relatively important transactions by ranking them with corresponding item sets. To the best our knowledge, HITS has rarely been applied in the domain of transactional data in retail sector.

4.1 Flow Diagram for HITS Algorithm

This section is designed to propose a flow diagram in order to explain the consecutive steps in the context of HITS algorithm. Figure 4.1 shows the flow diagram for the implementation of HITS algorithm in our datasets. In order to perform network analysis, we used R software, since it has all in-built functionalities. The following libraries from R were utilized in our analysis: reshape2, dplyr, igraph, arules, arulesViz, colorspace, and htmlwidgets. Firstly, we defined some basic terminologies for HITS algorithm as also seen in Figure 4.1.

Adjacency matrix provides a better representation of a network. The adjacency matrix A of a single graph with elements A_{ij} can be defined as follows [56].

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Transaction Object is a converted version of **Transaction Matrix** for the network analysis. **Hub and Authority** concepts are explained in the following section.

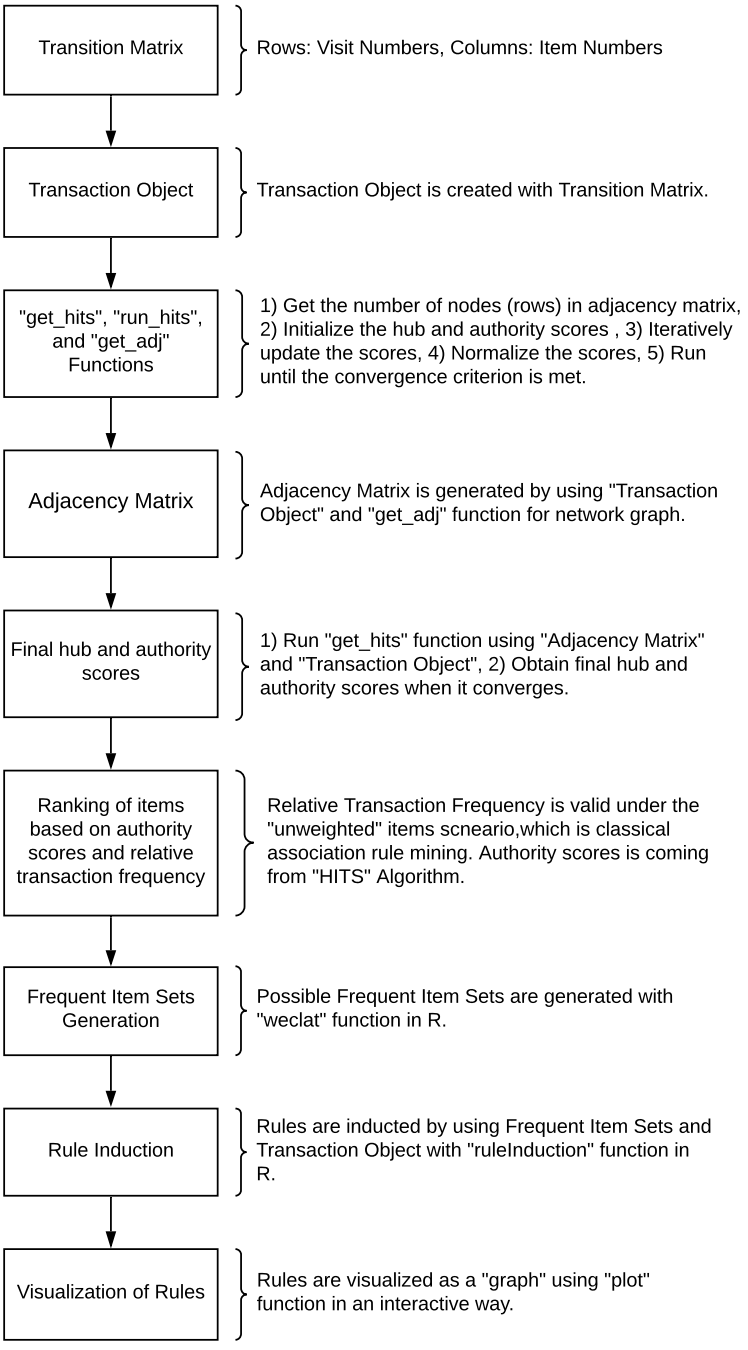


FIGURE 4.1: Flow Diagram for HITS Algorithm

4.2 Basic Principles of Hubs and Authorities (HITS)

Kumar and Sengottuvelan [38] proposed to use weighted association rule mining with HITS to determine rules, in other words the item sets, regarding good transactions, which are referred as hubs and to reveal some infrequent rules or item sets with *cross-selling* effects. The reinforcing relationship between transactions and items is similar to the relationship between hubs and authorities in the HITS model [11], [36], [37], and [38]. This means that every transaction seems to be a link/hub and items belong to transactions as an authority with many links/hubs.

The most important basic principle in HITS algorithm is that items that belong in relatively more transactions have relatively higher weight or importance, in other words, authority score. Similarly, transactions that comprise of many items have relatively higher weight or importance, in other words, hub score. To sum up, a transaction can be accepted as a good transaction with higher hub score, if it has relatively more items. Moreover, an item can be defined as a good item with higher authority score, if it is included in many transactions.

4.2.1 Ranking of Transactions with HITS

Transaction datasets can be expressed as a bipartite graph without the loss of information. Basically, there are some notations as given below. Figure 4.2 shows a typical representation of transaction database indicating transaction ID and corresponding item or item sets, and the bipartite graph.

Bipartite network, also called a *two-node network*, has two kinds of vertices. One of them refers to original vertices and other one represent groups to which they belong [56]. In the context of our study, the first vertex is the items, and the second one is transactions which they appear.

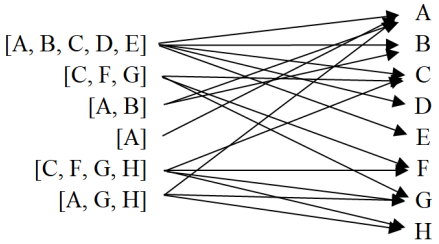
- D = $T_1, T_2, ..., T_m$: Transaction List

- I = $i_1, i_2, ..., i_n$: Item Set

- D is equal to the bipartite graph G =(D,I,E), where E = $(T, i) : i \in T, T \in D, i \in I$

TID	Transaction
100	[A, B, C, D, E]
200	[C, F, G]
300	[A, B]
400	[A]
500	[C, F, G, H]
600	[A, G, H]

(A) Database



(B) Bipartite graph

FIGURE 4.2: The bipartite graph representation of a database. (a) Database (b) Bipartite graph

As mentioned before, the reinforcing relationship between transactions and items is similar to the relationship between hubs and authorities in the HITS model [11], [36], [37], and [38]. Based on this similarity, transactions can be accepted as pure hubs and items as pure authorities within the context of HITS algorithm. In order to calculate hub scores of transactions and authority scores of items, we use the following Equations 4.2 and 4.3.

$$auth(i) = \sum_{T:i \in T} hub(T) \tag{4.2}$$

$$hub(T) = \sum_{i:i \in T} auth(i) \tag{4.3}$$

HITS algorithm calculates different hub and authority scores within each iteration. When HITS model ultimately converges, hub scores, in other words, hub weights of all transactions are gathered. Based on the hub weights of transactions, we can define the transactions whether they include high-value items or not. This assumption

means that any transaction with few items have a chance to become a good hub weights if most of the items within the transaction are top ranked depending on authority scores, which reflects the significance of an item [11], and [38]. On the contrary, any transaction with many common items may have low hub weight.

4.2.2 W-support and W-confidence

W-support is accepted as the generalization of support by taking the transactions' weights into consideration. Since the transaction weights are different with each other, a frequent item set may not be as important as it appears [38]. We can express the w-support of an item set X as given 4.4.

$$wsupp(X) = \frac{\sum_{T:X \subset T \wedge \in D} hub(T)}{\sum_{T:T \in D} hub(T)} \tag{4.4}$$

where hub(T) is the hub weight of transaction T.

We use Equation 4.5 and 4.6 in order to define w-support and w-confidence for association rules in whole transaction data set.

$$wsupp(X \Rightarrow Y) = wsupp(X \cup Y) \tag{4.5}$$

$$wconf(X \Rightarrow Y) = \frac{wsupp(X \cup Y)}{wsupp(X)} \tag{4.6}$$

According to Equation 4.6, w-confidence can be referred as the ratio of hub weights coming from X together with Y to the total hub weights coming from X. According to Equation 4.5 and 4.6, w-support reflects how significantly X and Y exist together; w-confidence reflects how strong the rule is.

The support and confidence values in Figures 4.3, 4.5, 4.7, 4.9, 4.11, and 4.13 are all w-support and w-confidence values in the context of HITS algorithm.

4.3 Product Networks, Rules, and Measures

4.3.1 General Product Network for Individual Members

We took all transactions of individual members into consideration when we created general individual product network. As reported in Table 3.3 and Table 3.4, there are 1,046,457 transactions and 47,013 individual members.

There are 219,084 different visit numbers and 5,412 different item numbers, which refers to unique numbers representing different items.

Figure 4.3 shows the general product network rules for individual members.

The majority of items in network rules belongs to Vegetables & Fruits. The detailed evaluations for general product network are given at the end of product network structure.

Rule Number	Rules	Support	Confidence	Lift
1	{Cigarettes & Tobacco $5.00 OFF} => {MARLBORO}	0.0012	0.676	144.308
2	{Lighter Fluid} => {Charcoal Briquettes}	0.0009	0.562	83.724
3	{Ladies Cotton Tshirt} => {Ladies Cotton Fleece}	0.0012	0.531	214.686
4	{Charcoal Starter} => {Charcoal Briquettes}	0.0011	0.659	98.219
5	{Charcoal Starter,Bananas} => {Charcoal Briquettes}	0.0001	0.850	126.681
6	{MARLBORO} => {Cigarettes & Tobacco $5.00 OFF}	0.0029	0.610	86.378
7	{Bananas,MARLBORO} => {Cigarettes & Tobacco $5.00 OFF}	0.0001	0.579	82.043
8	{On the Border Salsa,Bananas} => {Tortilla Chip}	0.0001	0.586	129.988
9	{Kibbles & Chucks,Navel Oranges} => {Bananas}	0.0001	0.563	9.857
10	{Red Delicious Apple,Honey Maid Graham} => {Bananas}	0.0001	0.667	11.683
11	{OTIS SPUNKMEYER Bread,Red Delicious Apple} => {Bananas}	0.0001	0.563	9.857
12	{Rotisserie Chicken,Kaiser Rolls} => {Bananas}	0.0001	0.625	10.952
13	{Red Delicious Apple,Animal Crackers} => {Bananas}	0.0001	0.500	8.762
14	{Peaches,VALENCIA Oranges} => {Bananas}	0.0001	0.667	11.683
15	{Assorted Muffins,Navel Oranges} => {Bananas}	0.0001	0.533	9.346
16	{Grapefruit,Dried Pitted Prunes} => {Bananas}	0.0001	0.571	10.014
17	{Rotisserie Chicken,Stoneground Wheat} => {Bananas}	0.0001	0.500	8.762
18	{Navel Oranges,Stoneground Wheat} => {Bananas}	0.0001	0.533	9.346
19	{Rotisserie Chicken,Peaches} => {Bananas}	0.0001	0.643	11.265
20	{Red Delicious Apple,Nectarines} => {Bananas}	0.0001	0.542	9.492
21	{Rotisserie Chicken,Red Delicious Apple} => {Bananas}	0.0001	0.519	9.086

FIGURE 4.3: General Product Network Rules for Individual Members

Figure 4.4 shows the general product network structure for individual members.

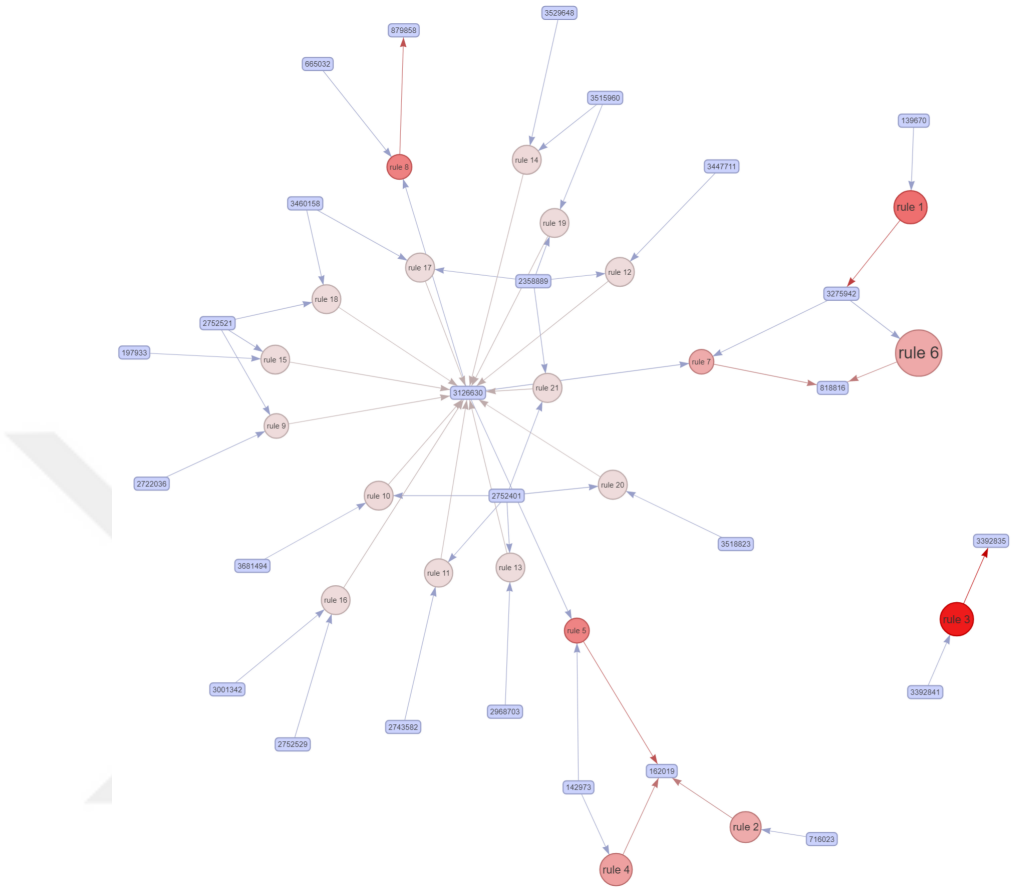


FIGURE 4.4: General Product Network for Individual Members

When we analyze the network structure, we can say that Bananas is the most authoritative item, followed by Navel Oranges, Red Delicious Apple and Rotisserie Chicken. Bananas mostly exists in right hand side, in other words as a consequent, in the network structure, on the other hand the other items with high authority scores exist in left hand side, in other words as a antecedent. This means that the purchase probability of Bananas, which is represented as weighted confidence, is dependent on different item or item sets that exists as antecedent in Figure 4.3.

One of the most remarkable findings from rules in Figure 4.3 is that people buying Charcoal Starter and Bananas are almost certain to buy Charcoal Briquettes (8.5 times out of 10). Also, people buying Charcoal Starter is more likely to buy Charcoal Briquettes (approximately 6.6 times out of 10). This is not surprising if you find Charcoal Briquettes next to Charcoal Starter on the shop shelf, but the inference from these rules which we can guess is that individual members are more likely to purchase the item sets which reflects their daily life habits and activities. For example, Rule 19: Rotisserie Chicken, Peaches ⇒ Bananas, is likely to be appeared in a typical family shopping card, because these item sets reflect the daily essentials. Another interesting rule is Rule 8: On the Border Salsa, Bananas ⇒ Tortilla Chip. We can infer that individual members tend to purchase some items which can be seem like complement items in overall.

Overall, we can propose appropriate marketing actions about promotions and cross-selling opportunities by combining mostly items whose categories are Vegetables and Fruit; Meat, Poultry, Seafood, Eggs & Diary; and Outdoor, Patio & Garden, as well as, organizing shelfs and catalog layouts.

Table 4.1 compare the scores of items which are included in the product network as seen Figure 4.4.

Item Score column gives the relative transaction frequency under the assumption that all transactions have equal weights as in classical association rule mining. *New Item Score* column refers to authority scores coming from HITS algorithm.

Although there are some exceptions for some items such as lighter fluid, ladies cotton fleece, assorted muffins etc., most of items have gone further up the order compared to original transaction frequency score.

TABLE 4.1: Comparison of Item Scores

Item Name	Item Score	New Item Score
Bananas	0.0571	0.7417
Navel Oranges	0.0128	0.0491
Red Delicious Apple	0.0100	0.0450
Rotisserie Chicken	0.0208	0.0369
Grapefruit	0.0038	0.0120
Nectarines	0.0027	0.0107
Peaches	0.0024	0.0098
Stoneground Wheat	0.0022	0.0098
Dried Pitted Prunes	0.0051	0.0094
Assorted Muffins	0.0089	0.0079
VALENCIA Oranges	0.0018	0.0079
Tortilla Chip	0.0045	0.0070
Animal Crackers	0.0053	0.0067
Kaiser Rolls	0.0031	0.0067
OTIS SPUNKMEYER Bread	0.0049	0.0064
Charcoal Briquettes	0.0067	0.0061
Honey Maid Graham	0.0041	0.0058
Kibbles & Chucks	0.0032	0.0036
Cigarettes & Tobacco $5.00 OFF	0.0071	0.0030
On the Border Salsa	0.0022	0.0026
MARLBORO	0.0047	0.0020
Charcoal Starter	0.0016	0.0018
Ladies Cotton Fleece	0.0025	0.0012
Ladies Cotton Tshirt	0.0022	0.0012
Lighter Fluid	0.0016	0.0009

The highest authority score belongs to Bananas, followed by Navel Oranges and Red Delicious Apple, and there are many rules which surround these authoritative items in Figure 4.4.

4.3.2 Individual Cluster 1 Product Network

We took only transactions of individual cluster 1 members into consideration for the individual cluster 1 product network.

As reported in Figure 3.14, there are 2,865 individual members. The number of transactions for these members is 74,650. There are 14,492 different visit numbers and 3,588 different item numbers.

Figure 4.5 shows the product network rules for individual cluster 1 members.

Rule Number	Rules	Support	Confidence	Lift
1	{Lighter Fluid} => {Charcoal Briquettes}	0.0014	0.625	56.965
2	{Hamburger Buns} => {Hot Dog Buns}	0.0023	0.635	122.625
3	{MARLBORO} => {Cigarettes & Tobacco $5.00 OFF}	0.0039	0.602	73.331
4	{Fresh Cut Green} => {Whole Kernel Corn}	0.0014	0.559	36.811
5	{Ladies Cotton Tshirt} => {Ladies Cotton Fleece}	0.0031	0.638	108.721
6	{Ladies Cotton Fleece} => {Ladies Cotton Tshirt}	0.0031	0.518	108.721
7	{Hamburger Buns} => {Hot Dog Buns}	0.0024	0.507	98.055
8	{Stoneground Wheat} => {Bananas}	0.0025	0.556	8.629
9	{Tomato Paste,Twin Pack Ketchup} => {Tomato Sauce}	0.0007	0.769	42.548
10	{Creole Seasoning,Twin Pack Ketchup} => {FRENCH'S Mustard}	0.0006	0.571	34.649
11	{FRENCH'S Mustard,Creole Seasoning} => {Twin Pack Ketchup}	0.0006	0.727	18.922
12	{Asparagus Spears,Cut Green Beans} => {Whole Kernel Corn}	0.0009	0.619	40.778
13	{Asparagus Spears,Whole Kernel Corn} => {Cut Green Beans}	0.0009	0.619	39.872
14	{Cheer Powder Ultra,Dawn Dish Detergent} => {Bountry Paper Towels}	0.0006	0.500	6.849
15	{Diced Tomatoes,Whole Kernel Corn} => {Cut Green Beans}	0.0009	0.545	35.132
16	{Kosher Petite Pickle,Cut Green Beans} => {Whole Kernel Corn}	0.0007	0.600	39.524
17	{Kosher Petite Pickle,Whole Kernel Corn} => {Cut Green Beans}	0.0007	0.643	41.406
18	{GLAD Trash Bag,TIDE Liquid Ultra} => {Bountry Paper Towels}	0.0005	0.538	7.376
19	{PLEDGE Lemon,GLAD Trash Bag} => {Bountry Paper Towels}	0.0007	0.529	7.252
20	{FRENCH'S Mustard,Mushroom Pieces} => {Twin Pack Ketchup}	0.0005	0.700	18.213
21	{Whole Kernel Corn,Dawn Dish Detergent} => {Cut Green Beans}	0.0005	0.500	32.204
22	{Cut Green Beans,Bananas} => {Whole Kernel Corn}	0.0007	0.556	36.596
23	{Whole Kernel Corn,Bananas} => {Cut Green Beans}	0.0007	0.588	37.888
24	{Cut Green Beans,Bountry Paper Towels} => {Whole Kernel Corn}	0.001	0.500	32.936
25	{Whole Kernel Corn,Bountry Paper Towels} => {Cut Green Beans}	0.001	0.609	39.205
26	{Whole Kernel Corn,Dawn Dish Detergent} => {Bountry Paper Towels}	0.0005	0.500	6.849
27	{FRENCH'S Mustard,Miracle Whip} => {Twin Pack Ketchup}	0.0007	0.529	13.774
28	{FRENCH'S Mustard,Dawn Dish Detergent} => {Twin Pack Ketchup}	0.0005	0.500	13.009
29	{Dawn Dish Detergent,Ritz Crackers} => {Bountry Paper Towels}	0.0006	0.571	7.827

FIGURE 4.5: Individual Cluster 1 Product Network Rules

We can say that individual members in cluster 1 are more likely to purchase items in canned, packaged foods and household essentials frequently together.

Figure 4.6 shows the product network structure for individual cluster 1 members.

According to network structure in Figure 4.6, Bountry Paper Towels, Dawn Dish Detergent, Twin Pack Ketchup are relatively high connected items with various

rules. Moreover, Cut Green Beans, Whole Kernel Corn, and FRENCH'S mustard are located at the central of network with relatively high authority scores in Table 4.2.

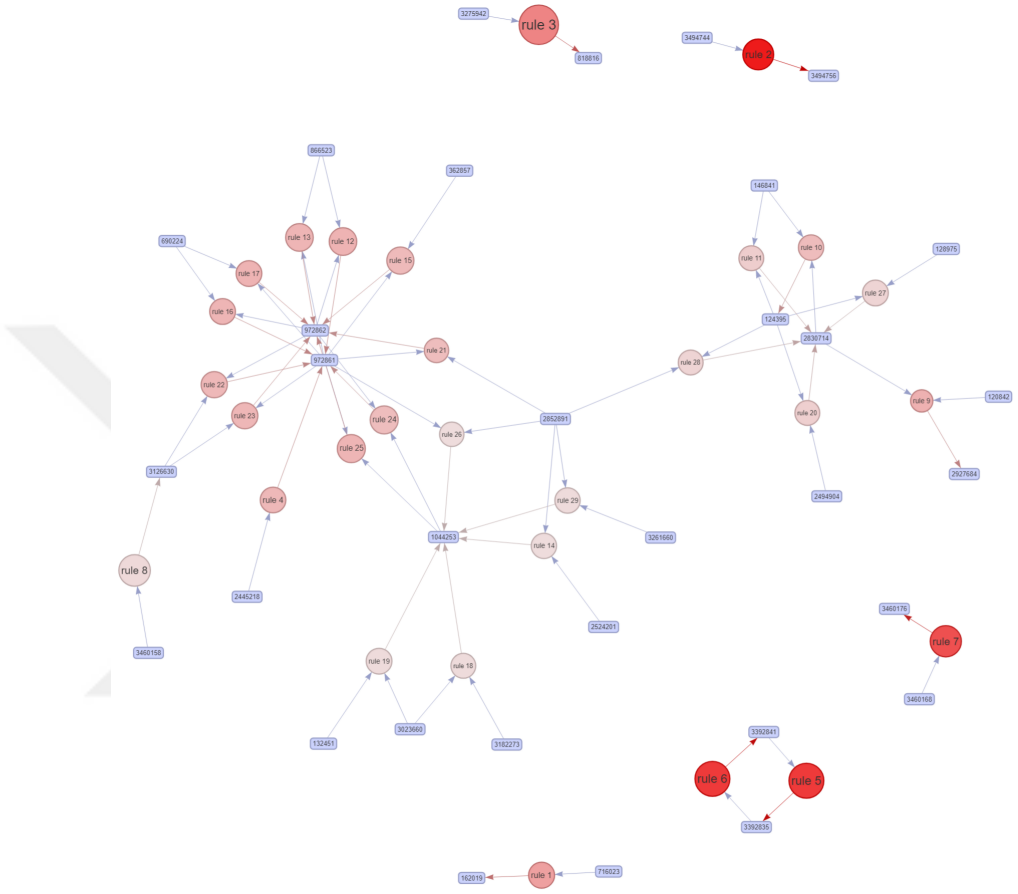


FIGURE 4.6: Individual Cluster 1 Product Network

The highest weighted confidence score belongs to Rule 9, which contains Tomato Paste and Twin Pack Ketchup as antecedents and Tomato Sauce as a consequent. Based on weighted confidence and corresponding lift value, it is not surprising if you find these items on the closer shelfs in the store. On the other hand, we can conclude that these items form good candidate to be substitute products, which are deliberately bought together much more often by sheer chance. The similar evaluation can be valid for Rule 20, which consists of FRENCH'S Mustard and Mushroom Pieces as antecedents, and Twin Pack Ketchup as consequent.

One of the interesting findings from rules in Figure 4.5 is that Whole Kernel Corn, Cut Green Bean, Twin Pack Ketchup and Bountry Paper Towels exist in different rules both as antecedent and as consequent with relative better lift rules.

In general, we can conclude that the members in individual cluster 1 are more likely to purchase complement items in canned, packaged foods and household essential items. Based on this information, potential marketing actions can be the combination of top items, in other words items that have relatively higher authority scores, in Household Essentials & Pets and Canned, packaged foods and Vegetables & Fruits by focusing on regular needs especially in Household Essentials, and complement items in Canned, packaged foods.

Bountry Paper Towels, Bananas, Twin Pack Ketchup, Dawn Dish Detergent and GLAD Trash Bag have the highest five authority scores, respectively, as can be seen in the New Item Score column in Table 4.2.

Table 4.2: Comparison of Item Scores

Item Name	**Item Score**	**New Item Score**
Bountry Paper Towels	0.0730	0.2619
Bananas	0.0644	0.1582
Twin Pack Ketchup	0.0384	0.1045
Dawn Dish Detergent	0.0281	0.0703
GLAD Trash Bag	0.0275	0.0643
Miracle Whip	0.0177	0.0400
Tomato Sauce	0.0181	0.0391
Ritz Crackers	0.0173	0.0386
FRENCH'S Mustard	0.0165	0.0372
Cut Green Beans	0.0155	0.0372
Whole Kernel Corn	0.0152	0.0352
Mushroom Pieces	0.0158	0.0312
PLEDGE Lemon	0.0121	0.0305
TIDE Liquid Ultra	0.0155	0.0299
Kosher Petite Pickle	0.0160	0.0288
Diced Tomatoes	0.0141	0.0276
Cheer Powder Ultra	0.0123	0.0240
Asparagus Spears	0.0130	0.0237
Charcoal Briquettes	0.0110	0.0141
Creole Seasoning	0.0050	0.0093
Tomato Paste	0.0049	0.0090
Stoneground Wheat	0.0043	0.0078
Cigarettes & Tobacco $5.00 OFF	0.0082	0.0069
Hot Dog Buns	0.0052	0.0062
Hot Dog Buns	0.0052	0.0055
Ladies Cotton Fleece	0.0059	0.0054
Hamburger Buns	0.0046	0.0052
Ladies Cotton Tshirt	0.0048	0.0051
Fresh Cut Green	0.0023	0.0048
MARLBORO	0.0064	0.0041
Hamburger Buns	0.0036	0.0035
Lighter Fluid	0.0022	0.0034

4.3.3 Individual Cluster 2 Product Network

We took only transactions of individual cluster 2 members into consideration for the individual cluster 2 product network.

As reported in Figure 3.14, there are 4,199 individual members. Corresponding number of transactions for these members is 156,365. There are 61,233 different visit numbers and 4,429 different item numbers.

Figure 4.7 exhibits product network rules for individual cluster 2.

The majority of items and item sets inside individual cluster 2 product network as seen in Figure 4.8 comprise Canned, packaged foods and products in Vegetables and Fruit category.

Rule Number	Rules	Support	Confidence	Lift
1	{MARLBORO} => {Cigarettes & Tobacco $5.00 OFF}	0.0039	0.696	76.503
2	{On the Border Salsa,Bananas} => {Tortilla Chip}	0.0001	0.667	158.225
3	{Peppered Gravy Mix,Bananas} => {Buttermilk Biscuit Milk}	0.0001	0.750	128.641
4	{AUSTRALIAN Navel Oranges} => {Bananas}	0.0002	0.524	6.716
5	{Charcoal Starter} => {Charcoal Briquettes}	0.0015	0.650	96.317
6	{Charcoal Starter,Bananas} => {Charcoal Briquettes}	0.0002	0.800	118.611
7	{Bananas,LYSOL Disinfectant} => {Bountry Paper Towels}	0.0001	0.545	17.163
8	{Navel Oranges,Electrasol Tabs} => {Bananas}	0.0001	0.625	8.013
9	{Kibbles & Chucks,Navel Oranges} => {Bananas}	0.0002	0.889	11.396
10	{Extra Fine Granulatd,Whole Almonds} => {Bananas}	0.0001	0.500	6.410
11	{Red Delicious Apple,Globe Grapes} => {Bananas}	0.0001	0.500	6.410
12	{Old Fashion White,Bag of Bagels} => {Bananas}	0.0002	0.583	7.479
13	{Thin Spaghetti,WINDFRESH 160 Load} => {Bananas}	0.0001	0.500	6.410
14	{Rotisserie Chicken,Kaiser Rolls} => {Bananas}	0.0002	0.700	8.975
15	{Granny Smith Apple,Navel Oranges} => {Bananas}	0.0002	0.688	8.814
16	{Rotisserie Chicken,Stoneground Wheat} => {Bananas}	0.0002	0.500	6.410
17	{Navel Oranges,Dawn Dish Detergent} => {Bananas}	0.0001	0.625	8.013
18	{Peaches,VALENCIA Oranges} => {Bananas}	0.0001	0.857	10.989
19	{Rotisserie Chicken,Peaches} => {Bananas}	0.0001	0.714	9.158
20	{Navel Oranges,Bag of Bagels} => {Bananas}	0.0001	0.556	7.123
21	{Cocktail Croissant,Navel Oranges} => {Bananas}	0.0002	0.500	6.410
22	{Iron Kids Bread,Kosher Petite Pickle} => {Bananas}	0.0001	0.500	6.410
23	{Kosher Petite Pickle,Rotisserie Chicken} => {Bananas}	0.0001	0.625	8.013
24	{Classic Roast Coffee,Navel Oranges} => {Bananas}	0.0002	0.563	7.212

FIGURE 4.7: Individual Cluster 2 Product Network Rules

Figure 4.8 exhibits the product network structure for individual cluster 2.

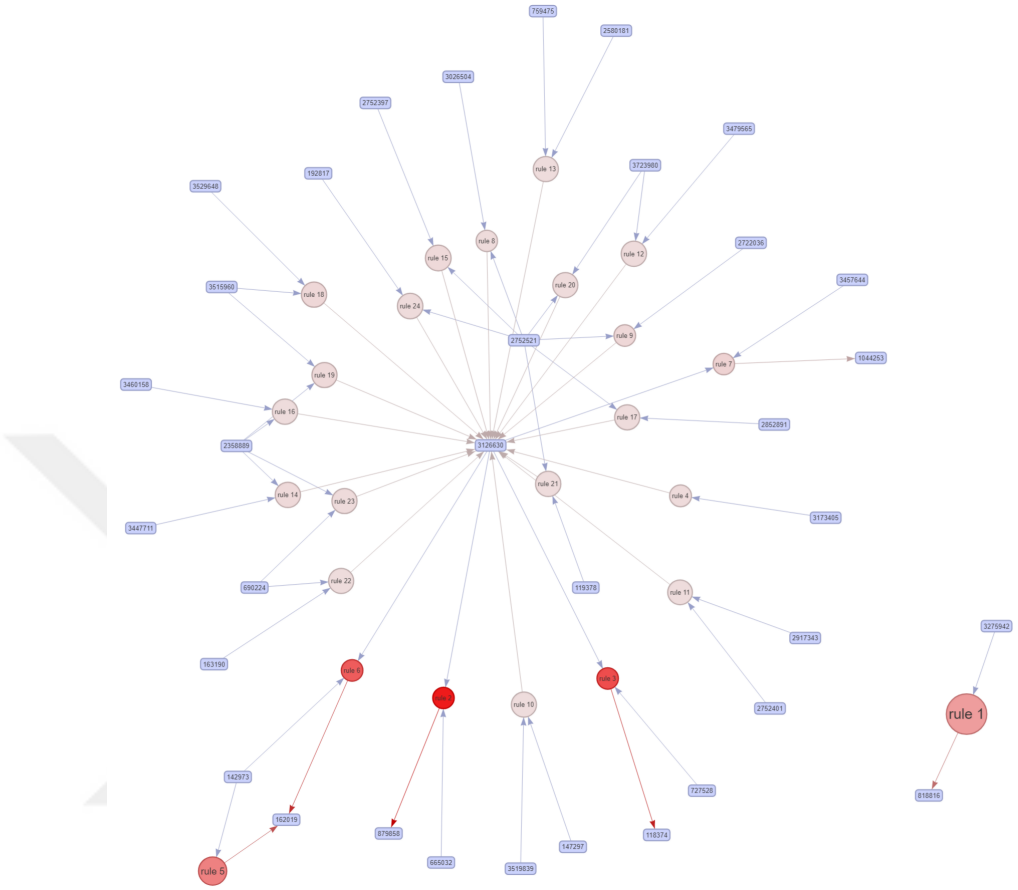


FIGURE 4.8: Individual Cluster 2 Product Network

We can easily observe that Bananas is the most authoritative item in Figure 4.8. The corresponding authority score of Bananas can be shown in Table 4.3, having the highest authority score and existing in many rules as a consequent. This is an important information in terms of customers who are more likely to buy other item or item sets before the Bananas.

The combination of item sets in Vegetables & Fruits and in Canned, packaged goods are mostly common in individual cluster 2 product network rules. For example, based on Rule 9, we can say that if a customer Kibbles & Chucks and Navel Oranges,

he/she is almost certain to buy (Confidence = 0.889, almost 9 times out of 10) Bananas.

For example, as aforementioned, items in Vegetables and Fruits category are mostly authoritative. Therefore, there are many transactions that indicate customers are more likely to buy these products together. Possible up-selling marketing actions can include offering additional similar product options, or other versions of current products, i.e organic, to fulfill better their needs. Especially, purchasing organic product options may be boosted for U.S. citizens, since there has been a growing perception to deal with obesity. The main assumption is that organic options are more profitable for Sam's Club.

Overall, we can infer that members in individual cluster 2 tend to buy products which are consumed in a short time period such as Vegetables & Fruits. For this segment, possible marketing actions should include mostly items in Vegetables & Fruits category. We can clearly say that up-selling opportunities can be highly possible for customers in this segment with higher-end products in terms of profitability for businesses.

The first item, Bananas, is the leading item with highest authority score in Table 4.3. Navel Oranges, Red Delicious Apple, and Bountry Paper Towels can be ranked as second, third, and fourth, respectively, when we consider their authority scores, as given in Table 4.3.

According to our network in Figure 4.8, we can see Bananas as the most authoritative item with many items and rules around it. Similarly, second and third highest authority scores belong to items in Vegetables & Fruits category.

Table 4.3: Comparison of Item Scores

Item Name	Item Score	New Item Score
Bananas	0.0780	0.8477
Navel Oranges	0.0169	0.0603
Red Delicious Apple	0.0130	0.0484
Bountry Paper Towels	0.0318	0.0453
Rotisserie Chicken	0.0214	0.0335
Extra Fine Granulated	0.0156	0.0259
Classic Roast Coffee	0.0149	0.0256
Iron Kids Bread	0.0100	0.0195
WINDFRESH 160 Load	0.0126	0.0152
Kosher Petite Pickle	0.0084	0.0134
Cocktail Croissant	0.0106	0.0131
VALENCIA Oranges	0.0026	0.0109
Bag of Bagels	0.0027	0.0106
Peaches	0.0027	0.0105
Bananas	0.0074	0.0104
Dawn Dish Detergent	0.0096	0.0101
Stoneground Wheat	0.0032	0.0094
Granny Smith Apple	0.0024	0.0087
Buttermilk Biscuit Milk	0.0058	0.0080
Kaiser Rolls	0.0041	0.0076
Thin Spaghetti	0.0034	0.0075
Old Fashion White	0.0020	0.0073
Globe Grapes	0.0028	0.0071
Whole Almonds	0.0053	0.0068
Tortilla Chip	0.0042	0.0064
Kibbles & Chucks	0.0036	0.0053
Charcoal Briquettes	0.0067	0.0050
Electrasol Tabs	0.0045	0.0044
LYSOL Disinfectant	0.0043	0.0033
Charcoal Starter	0.0022	0.0023
AUSTRALIAN Navel Oranges	0.0003	0.0020
Peppered Gravy Mix	0.0019	0.0019
On the Border Salsa	0.0013	0.0019
MARLBORO	0.0056	0.0019
Cigarettes & Tobacco $5.00 OFF	0.0091	0.0017

4.3.4 General Product Network for Business Members

We took all transactions of business members into consideration when we created general business product network.

As reported in Table 3.3 and Table 3.4, there are 66,952 transactions and 1,454 business members. There are 13,203 different visit numbers and 3,432 different item numbers, which refers to unique numbers representing different items.

Figure 4.9 shows general product network rules for business members.

Rule Number	Rules	Support	Confidence	Lift
1	{Dorito Nacho Cheese} => {Cheetos Crunchy}	0.0107	0.532	30.021
2	{Cheetos Crunchy} => {Dorito Nacho Cheese}	0.0107	0.603	30.021
3	{Chex Mix,Animal Crackers} => {Bountry Paper Towels}	0.0006	0.538	6.155
4	{FRENCH'S Mustard,Bountry Paper Towels} => {Twin Pack Ketchup}	0.0007	0.529	36.405
5	{DOW Bathroom Cleaner,Windex Combo} => {Bountry Paper Towels}	0.0006	0.636	7.274
6	{Extra Fine Granulated,GLAD Trash Bag 13Gal} => {Bountry Paper Towels}	0.0006	0.700	8.002
7	{BOUNCE Singles 160CT,LYSOL Toilet Essential} => {DOW Bathroom Cleaner}	0.0006	0.500	44.605
8	{DAWN Dish Detergent,DOW Bathroom Cleaner} => {Bountry Paper Towels}	0.0005	0.600	6.859
9	{PLEDGE Lemon,DOW Bathroom Cleaner} => {Bountry Paper Towels}	0.0009	0.524	5.988
10	{BOUNCE Singles 160CT,LYSOL Toilet Essential} => {Bountry Paper Towels}	0.0006	0.500	5.716
11	{PLEDGE Lemon,Special Roast Coffee} => {Bountry Paper Towels}	0.0005	0.750	8.573
12	{Special Roast Coffee,LYSOL Disinfectant} => {Bountry Paper Towels}	0.0006	0.583	6.668
13	{Pure Cane Sugar} => {Non-Dairy Creamer}	0.0127	0.612	23.895
14	{LYSOL Disinfectant,Pure Cane Sugar} => {Non-Dairy Creamer}	0.0008	0.769	30.048
15	{LYSOL Disinfectant,Non-Dairy Creamer} => {Pure Cane Sugar}	0.0008	0.667	32.242
16	{Classic Roast Coffee,Pure Cane Sugar} => {Non-Dairy Creamer}	0.0024	0.633	24.713
17	{Classic Roast Coffee,Non-Dairy Creamer} => {Pure Cane Sugar}	0.0024	0.500	24.181
18	{LYSOL Disinfectant,Pure Cane Sugar} => {Bountry Paper Towels}	0.0006	0.538	6.155
19	{Classic Roast Coffee,LYSOL Disinfectant} => {Bountry Paper Towels}	0.0005	0.545	6.235

FIGURE 4.9: General Product Network Rules for Business Members

Figure 4.10 demonstrates a general product network for business members.

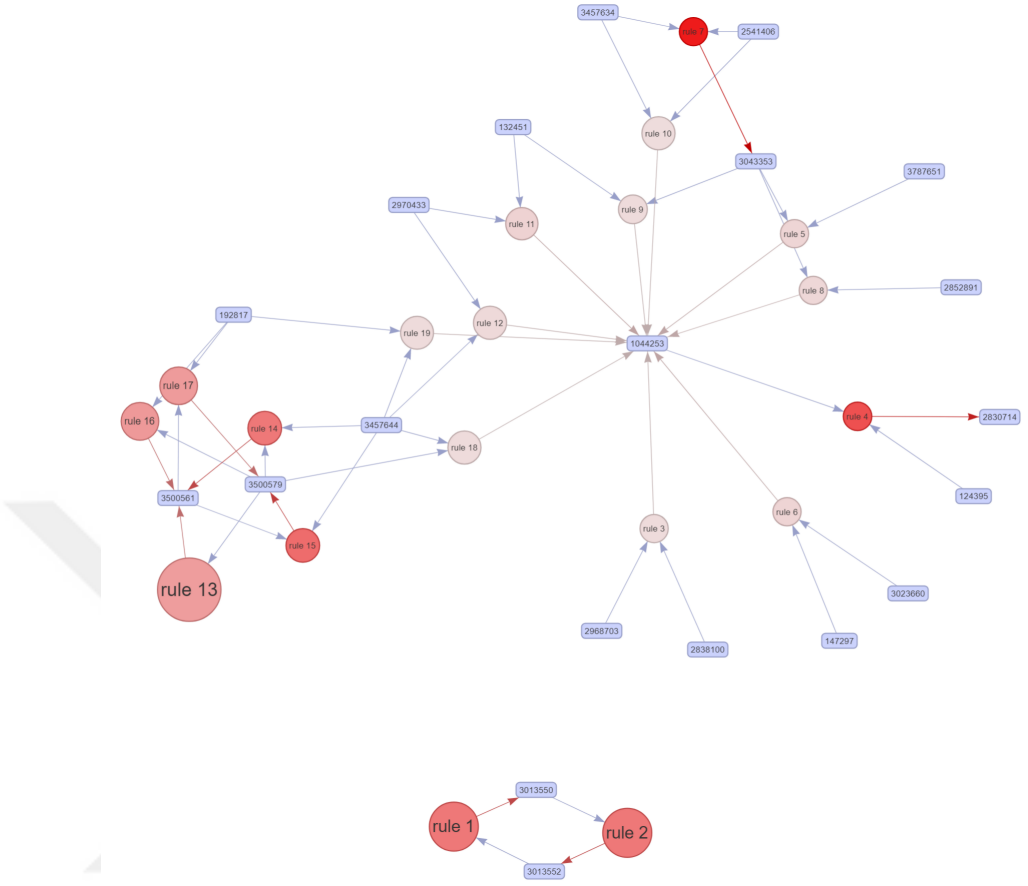


FIGURE 4.10: General Product Network for Business Members

According to the general product network for business members, we can infer that majority of items in the network belong to either Household Essentials and pets, or Canned, packaged foods category.

Bountry Paper Towels is the most frequently bought item with the highest authority score in Table 4.4, as also seen in Figure 4.10 with many connected items and rules.

One of the most interesting findings from this network is that pure cane sugar, non-dairy creamer and roast coffee can be accepted as complement products. The rules which include these items have relatively higher lift values by indicating that these items are deliberately bought together much more often to the business customers.

In conclusion, we can promote the items either in Household Essentials and pets, or Canned, packaged foods by determining the most frequently bought ones and combining them into combo offers. It is also possible to get advantage of cross-selling effects from business members, due to the fact that common items which are purchased by business members are basically in two categories as mentioned above, since these items are more likely to reinforce value of each other by occurring together. Table 4.4 exhibits item scores for items in general business product network.

Table 4.4: Comparison of Item Scores

Item Name	**Item Score**	**New Item Score**
Bountry Paper Towels	0.0875	0.6689
Classic Roast Coffee	0.0311	0.0799
LYSOL Disinfectant	0.0220	0.0718
GLAD Trash Bag 13Gal	0.0192	0.0630
Non-Dairy Creamer	0.0256	0.0615
Pure Cane Sugar	0.0207	0.0550
PLEDGE Lemon	0.0167	0.0520
Special Roast Coffee	0.0225	0.0511
DAWN Dish Detergent	0.0205	0.0501
LYSOL Toilet Essential	0.0159	0.0472
BOUNCE Singles 160CT	0.0209	0.0469
DOW Bathroom Cleaner	0.0112	0.0359
Extra Fine Granulated	0.0189	0.0351
Windex Combo	0.0111	0.0294
Twin Pack Ketchup	0.0145	0.0261
FRENCH'S Mustard	0.0131	0.0209
Chex Mix	0.0127	0.0204
Dorito Nacho Cheese	0.0201	0.0131
Cheetos Crunchy	0.0177	0.0124

The highest authority score belongs to Bountry Paper Towels, followed by Classic Roast Coffee, LYSOL Disinfectant, and GLAD Trash Bag.

4.3.5 Business Cluster 1 Product Network

We took only transactions of business cluster 1 members into consideration for the business cluster 1 product network.

As reported in Figure 3.18, there are 306 business members. Corresponding number of transactions for these members is 20,229. There are 5,016 different visit numbers and 2,505 different item numbers.

Figure 4.11 demonstrates product network rules for business cluster 1.

One of the most remarkable insight from business cluster 1 product network is that Household Essential items and Canned, packaged foods exist extensively in product network rules.

Rule Number	Rules	Support	Confidence	Lift
1	{RSVP Fine Ballpoint} => {Bountry Paper Towels}	0.0008	0.571	5.838
2	{TIDE Liquid MTN Spring} => {Bountry Paper Towels}	0.0008	0.500	5.108
3	{GLADE Plug-In } => {Bountry Paper Towels}	0.0010	0.714	7.297
4	{Whole Kernel Corn} => {Cut Green Beans}	0.0022	0.647	79.162
5	{HEINZ Vinegar} => {Electrasol Tabs}	0.0012	0.667	72.696
6	{HEINZ Vinegar,Electrasol Tabs} => {Bountry Paper Towels}	0.0012	1.000	10.216
7	{Bountry Paper Towels,HEINZ Vinegar} => {Electrasol Tabs}	0.0012	0.857	93.466
8	{Bountry Paper Towels,HEINZ Vinegar} => {Trashbag 13GAL White}	0.0008	0.571	33.721
9	{Trashbag 13GAL White,HEINZ Vinegar} => {Bountry Paper Towels}	0.0008	1.000	10.216
10	{HEINZ Vinegar} => {Bountry Paper Towels}	0.0014	0.778	7.946
11	{Iron Kids Bread,Bountry Paper Towels} => {Bananas}	0.0008	0.800	23.886
12	{Dorito Nacho Cheese} => {Cheetos Cruncy}	0.0152	0.521	21.402
13	{Cheetos Cruncy} => {Dorito Nacho Cheese}	0.0152	0.623	21.402
14	{Chex Mix,Animal Crackers} => {Bountry Paper Towels}	0.0008	0.571	5.838
15	{Bountry Paper Towels,Animal Crackers} => {Chex Mix}	0.0008	0.571	48.581
16	{Bountry Paper Towels,Chex Mix} => {Animal Crackers}	0.0008	0.571	60.985
17	{Trashbag 55GAL Black,Trashbag 13GAL White} => {Bountry Paper Towels}	0.0010	0.625	6.385
18	{Bountry Paper Towels,Navel Oranges} => {Bananas}	0.0008	0.571	17.061
19	{Regular Hot Cocoa,Bountry Paper Towels} => {Special Roast Coffee}	0.0008	0.500	20.727
20	{Bountry Paper Towels,Red Delicious Apple} => {Bananas}	0.0008	0.571	17.061
21	{Highspeed 84 Brite,HEWLETT PACKARD Black Cartridge} => {Bountry Paper Towels}	0.0008	0.571	5.838
22	{FRENCH'S Mustard,Bountry Paper Towels} => {Twin Pack Ketchup}	0.0010	0.500	27.560
23	{Trashbag 13GAL White,SCOTCH Tape Refill} => {Bountry Paper Towels}	0.0008	0.667	6.811
24	{DOW Bathroom Cleaner,WINDEX Combo} => {Bountry Paper Towels}	0.0008	0.800	8.173
25	{PLEDGE Lemon,WINDEX Combo} => {Bountry Paper Towels}	0.0008	0.667	6.811
26	{PLEDGE Lemon,DOW Bathroom Cleaner} => {Bountry Paper Towels}	0.0008	0.800	8.173
27	{Dawn Dish Detergent,LYSOL Toilet 3 PK} => {Bountry Paper Towels}	0.0008	0.500	5.108
28	{Pure Cane Sugar} => {Non-Dairy Creamer}	0.0126	0.624	25.437
29	{Non-Dairy Creamer} => {Pure Cane Sugar}	0.0126	0.512	25.437
30	{LYSOL Disinfectant,Pure Cane Sugar} => {Non-Dairy Creamer}	0.0008	1.000	40.780
31	{LYSOL Disinfectant,Non-Dairy Creamer} => {Pure Cane Sugar}	0.0008	1.000	49.663
32	{Classic Roast Coffee,Pure Cane Sugar} => {Non-Dairy Creamer}	0.003	0.625	25.488
33	{Classic Roast Coffee,Non-Dairy Creamer} => {Pure Cane Sugar}	0.003	0.536	26.605
34	{Dawn Dish Detergent,LYSOL Disinfectant} => {Bountry Paper Towels}	0.0012	0.600	6.130

FIGURE 4.11: Business Cluster 1 Product Network Rules

Figure 4.12 shows the product network for Business Cluster 1.

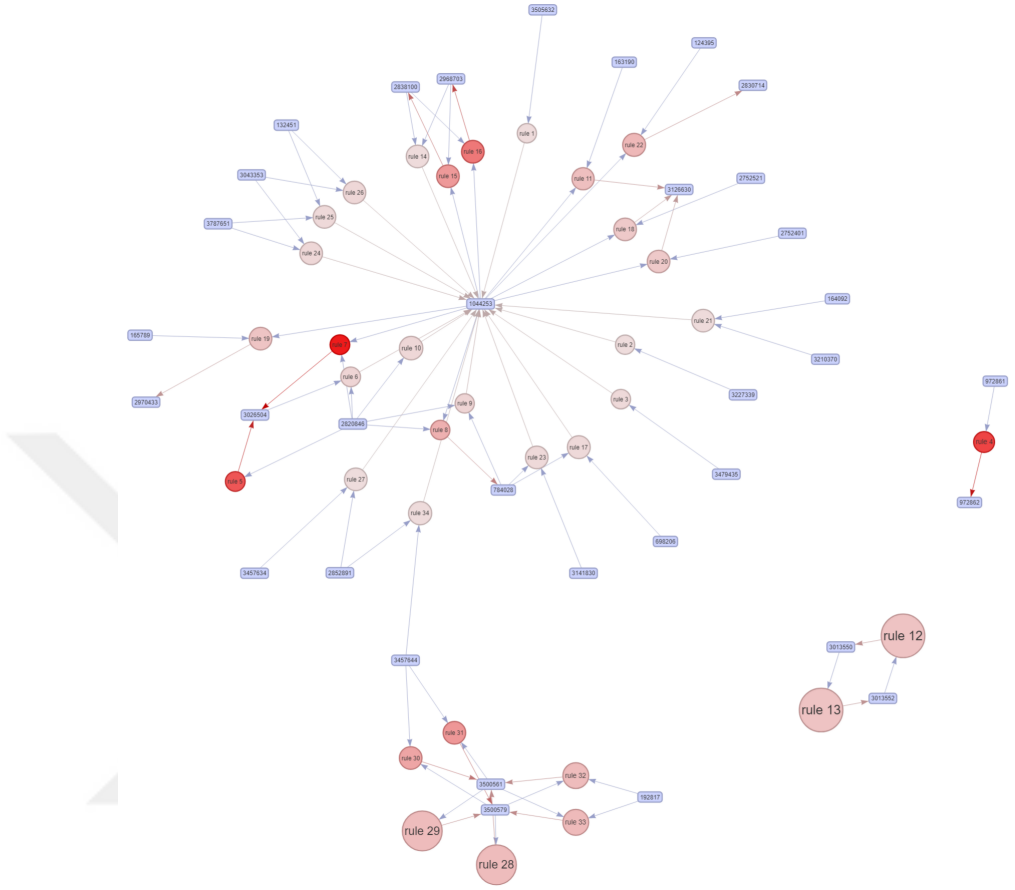


FIGURE 4.12: Business Cluster 1 Product Network

Bountry Paper Towels exists both as an antecedent and as a consequent in business cluster 1 product network. This indicates that there are different relationships with Bountry Paper Towels and other items which reinforce the purchase of each other by occurring together.

Generally, we can infer that on the right hand side of rules, in other words, antecedents consists of items from different categories such as Household Essentials and Pets and Canned, packaged foods. Items in Household Essentials and Pets are as common as items in Canned, packaged foods according to business cluster 1 product network.

There are some items from Office and Electronics category such as HEWLETT PACKARD Black Cartridge and SCOTCH Tape Refill. There are also some products from Vegetables & Fruit Category.

It is important to highlight that there are some rules including possible complement products such as Rule 28, 29, 30, 31, 32, and 33. Pure Cane Sugar, Non-Dairy Creamer, and Classic Roast Coffee can be potentially candidate to be complement items which reinforce value of each other by occurring frequently together. Based on this information, it can be possible to promote these items much more often to business members in cluster 1.

To sum up, we can propose appropriate marketing actions by focusing on Household Essential items and complement items, especially in Canned, packaged foods in order to plan appropriate discount programs. It is also possible to utilize markup complements in the context of marketing actions for this segment.

Table 4.6 exhibits item scores for items in general business product network.

Table 4.5: Comparison of Item Scores

Item Name	Item Score	New Item Score
Bountry Paper Towels	0.0979	0.6349
Classic Roast Coffee	0.0377	0.1009
LYSOL Disinfectant	0.0213	0.0741
Dawn Dish Detergent	0.0233	0.0668
Highspeed 84 Brite	0.0257	0.0604
Special Roast Coffee	0.0241	0.0561
Bananas	0.0335	0.0560
Non-Dairy Creamer	0.0245	0.0549
PLEDGE Lemon	0.0183	0.0549
Pure Cane Sugar	0.0201	0.0486
Trashbag 13GAL White	0.0169	0.0437
LYSOL Toilet 3 PK	0.0148	0.0390
Twin Pack Ketchup	0.0181	0.0309
DOW Bathroom Cleaner	0.0094	0.0306
WINDEX Combo	0.0106	0.0295
Electrasol Tabs	0.0092	0.0286
SCOTCH Tape Refill	0.0132	0.0263
FRENCH'S Mustard	0.0154	0.0257
HEWLETT PACKARD Black Cartridge	0.0110	0.0218
Red Delicious Apple	0.0128	0.0207
Chex Mix	0.0118	0.0189
Regular Hot Cocoa	0.0094	0.0182
Dorito Nacho Cheese	0.0291	0.0181
Navel Oranges	0.0098	0.0178
Trashbag 55GAL Black	0.0056	0.0177
Animal Crackers	0.0094	0.0171
Cheetos Cruncy	0.0243	0.0167
Cut Green Beans	0.0082	0.0140
Iron Kids Bread	0.0058	0.0111
HEINZ Vinegar	0.0018	0.0104
Whole Kernel Corn	0.0034	0.0066
GLADE Plug-In	0.0014	0.0065
TIDE Liquid MTN Spring	0.0016	0.0061
RSVP Fine Ballpoint	0.0014	0.0059

The highest authority score belongs to Bountry Paper Towels, followed by Classic Roast Coffee, LYSOL Disinfectant, Dawn Dish Detergent and High-speed 84 Brite.

4.3.6 Business Cluster 6 Product Network

We took only transactions of business cluster 6 members into consideration for the business cluster 6 product network. As reported in Figure 3.18, there are 145 business members. Corresponding number of transactions for these members is 4,624. There are 932 different visit numbers and 1,441 different item numbers.

Figure 4.13 exhibits product network rules for business cluster 6.

We can state that Household Essential items dominate the network structure with many rules.

Rule Number	Rules	Support	Confidence	Lift
1	{MEMBERS MARK Bath} => {Bountry Paper Towels}	0.0033	0.600	5.886
2	{Muenster Cheese} => {Bountry Paper Towels}	0.0043	0.800	7.848
3	{Ultra Wisk} => {Bountry Paper Towels}	0.0043	0.667	6.540
4	{FOLGERS Regular OCS} => {Bountry Paper Towels}	0.0043	0.500	4.905
5	{Electrasol Tabs} => {Bountry Paper Towels}	0.0043	0.800	7.848
6	{Raisin Bran Crunch} => {Bountry Paper Towels}	0.0033	0.600	5.886
7	{Mozz Proscuito Roll} => {Bountry Paper Towels}	0.0043	0.800	7.848
8	{JET DRY} => {BOUNCE Singles 160CT}	0.0043	0.571	16.643
9	{LYSOL Deodorizing} => {Bountry Paper Towels}	0.0065	0.545	5.351
10	{Extra Fine Granulatd} => {Bountry Paper Towels}	0.0054	0.500	4.905
11	{Twin Pack Ketchup} => {FRENCH'S Mustard}	0.0140	0.565	25.085
12	{FRENCH'S Mustard} => {Twin Pack Ketchup}	0.0140	0.619	25.085
13	{DOW Bathroom Cleaner,CASCADE Gel} => {Bountry Paper Towels}	0.0033	0.750	7.358
14	{Bountry Paper Towels,CASCADE Gel} => {DOW Bathroom Cleaner}	0.0033	0.500	24.526
15	{LYSOL Toilet 3 PK,WINDEX Combo} => {Bountry Paper Towels}	0.0033	0.600	5.886
16	{Bountry Paper Towels,LYSOL Toilet 3 PK} => {WINDEX Combo}	0.0033	0.500	27.412
17	{DOW Bathroom Cleaner,WINDEX Combo} => {Bountry Paper Towels}	0.0033	0.750	7.358
18	{DOW Bathroom Cleaner,LYSOL Toilet 3 PK} => {Bountry Paper Towels}	0.0033	0.600	5.886
19	{Bountry Paper Towels,LYSOL Toilet 3 PK} => {DOW Bathroom Cleaner}	0.0033	0.500	24.526
20	{GLAD Trash Bag,LYSOL Toilet 3 PK} => {BOUNCE Singles 160CT}	0.0033	1.000	29.125
21	{BOUNCE Singles 160CT,LYSOL Toilet 3 PK} => {GLAD Trash Bag}	0.0033	1.000	29.125
22	{COMET 4PK Household Essential,DOW Bathroom Cleaner} => {Bountry Paper Towels}	0.0033	0.750	7.358
23	{COMET 4PK Household Essential,Bountry Paper Towels} => {DOW Bathroom Cleaner}	0.0033	0.500	24.526
24	{COMET 4PK Household Essential,BOUNCE Singles 160CT} => {Bountry Paper Towels}	0.0033	0.600	5.886
25	{COMET 4PK Household Essential,Bountry Paper Towels} => {BOUNCE Singles 160CT}	0.0033	0.500	14.563
26	{GLAD Trash Bag,DOW Bathroom Cleaner} => {BOUNCE Singles 160CT}	0.0033	0.600	17.475
27	{BOUNCE Singles 160CT,DOW Bathroom Cleaner} => {GLAD Trash Bag}	0.0033	0.600	17.475
28	{GLAD Trash Bag,DOW Bathroom Cleaner} => {Bountry Paper Towels}	0.0033	0.600	5.886
29	{BOUNCE Singles 160CT,GLAD Trash Bag} => {Bountry Paper Towels}	0.0043	0.571	5.606
30	{Pure Cane Sugar} => {Non-Dairy Creamer}	0.0301	0.609	11.819
31	{Non-Dairy Creamer} => {Pure Cane Sugar}	0.0301	0.583	11.819

FIGURE 4.13: Business Cluster 6 Product Network Rules

Figure 4.14 exhibits the product network for business members in cluster 6.

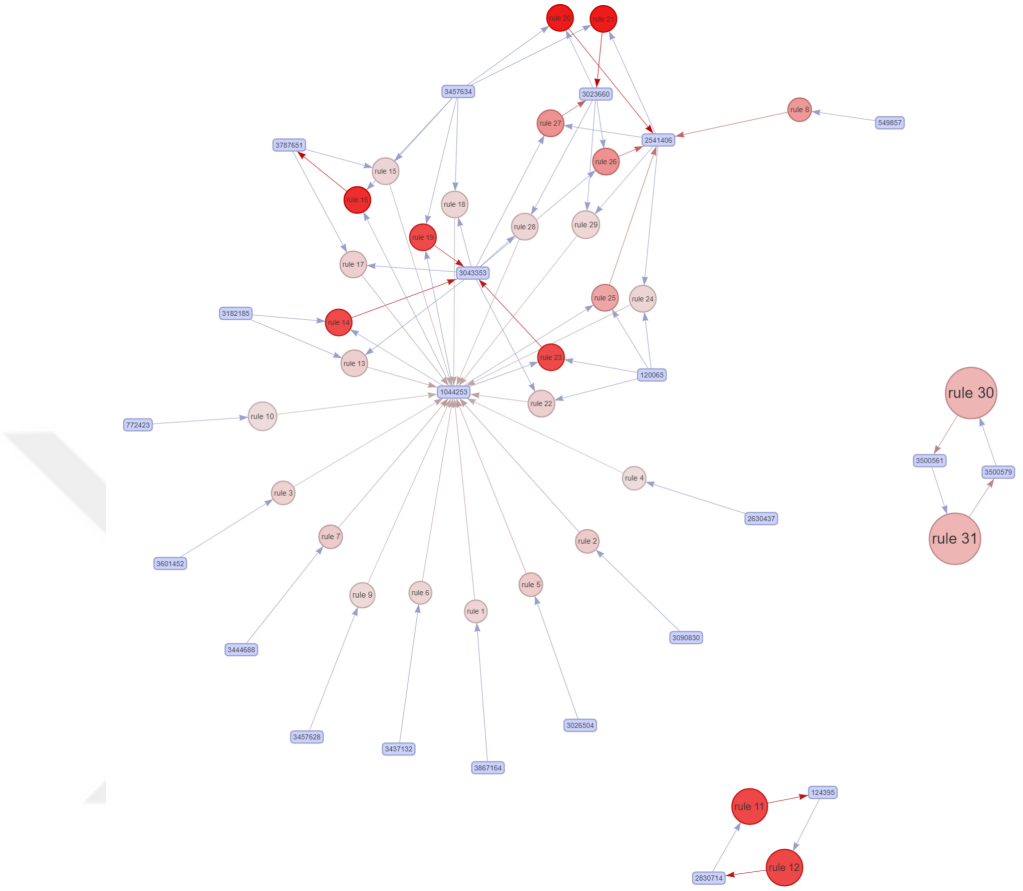


FIGURE 4.14: Business Cluster 6 Product Network

We see in Figure 4.13, there are many Household Essential items both as an antecedent and as a consequent.

As seen in Figure 4.14, Bountry Paper Towels and DOW Bathroom Cleaner are connected with many rules and items. The difference between these two items is that Bountry Paper Towels exists mostly as a consequent in the network, on the other hand, DOW Bathroom Cleaner are more likely to be an antecedent in this network. This means that DOW Bathroom Cleaner reinforce other items by occuring together, however, Bountry Paper Towers is affected by the purchase of other items within the rule.

For example, if a customer purchase DOW Bathroom Cleaner and CASCADE Gel, he/she is mostly likely to buy Bountry Paper Towels (Rule 13, confidence=0.75, 7.5 times out of 10). Similarly, if a customer purchase COMET 4PK Household Essential and DOW Bathroom Cleaner, he/she is mostly likely to buy Bountry Paper Towels (Rule 13, confidence=0.75, 7.5 times out of 10).

The most remarkable finding from this network structure is that shelf planning and catalog layout can be redesigned based on the frequently bought items in Household Essentials category. Also, it might be helpful in designing combo offers for the top selling products in Household Essentials. It is also possible to develop up-selling opportunities by using Household Essential items for the purpose of boosting sales and increasing profit.

Table 4.6 shows the item scores for items in business cluster 6 product network.

TABLE 4.6: Comparison of Item Scores

Item Name	**Item Score**	**New Item Score**
Bountry Paper Towels	0.1019	0.4908
GLAD Trash Bag	0.0343	0.1458
Pure Cane Sugar	0.0494	0.1427
Non-Dairy Creamer	0.0515	0.1366
BOUNCE Singles 160CT	0.0343	0.1262
DOW Bathroom Cleaner	0.0204	0.0833
COMET 4PK Household Essential	0.0204	0.0634
LYSOL Toilet 3 PK	0.0204	0.0625
WINDEX Combo	0.0182	0.0608
CASCADE Gel	0.0204	0.0587
FRENCH'S Mustard	0.0225	0.0571
Twin Pack Ketchup	0.0247	0.0529
Extra Fine Granulated	0.0107	0.0436
LYSOL Deodorizing	0.0118	0.0433
JET DRY	0.0075	0.0305
Mozz Proscuito Roll	0.0054	0.0288
Raisin Bran Crunch	0.0054	0.0251
Electrasol Tabs	0.0054	0.0250
FOLGERS Regular OCS	0.0086	0.0246
Ultra Wisk	0.0064	0.0219
Muenster Cheese	0.0054	0.0218
MEMBERS MARK Bath	0.0054	0.0206

According to Table 4.6, we can say that Bountry Paper Towels has the highest authority score, followed by GLAD Trash Bag, Pure Cane Sugar, Non-Dairy Creamer, BOUNCE Singles 160 CT,and DOW Bathroom Cleaner, respectively.

4.4 Marketing Implications of the Results

Product networks which we developed using HITS algorithm can be helpful in determining the recurring purchasing patterns, the complement and substitute products, and the trigger products which have high authority scores. Moreover, marketing practitioners can develop appropriate marketing strategies in inventory planning, designing of combo effects, planning of discount programs and promotions (i.e discount one product or more, mark up complements), shelf planning, catalog layout, recommending of products, and cross-selling, especially in higher margin products.

Our network outputs are expected to help businesses for deciding what types of items to put on sale, which items to be combined for the coupons, how to place merchandise on shelves based on products which are bought frequently together.

We can define the following characteristics of the individual and business member clusters as well as managerial insights for each, by utilizing products network outputs and two-stage clustering results.

4.4.1 Individual Clusters Marketing Implications

Cluster 1 is the leading cluster according to CLV score. Members in Cluster 1 made shopping occasionally. They made their last shopping a short time ago. They were spending very high amount of money both per each visit and in total. There were a moderate number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were high.

This cluster can be regarded as an "affluent segment". Due to being heavy spenders, they are more likely to be responsive towards cross-selling and up-selling opportunities. By showing special attention to members in this segment, the segment may

be turned into star segment. It would be beneficial to build one-to-one relationship in order to attract them to the store more frequently.

According to individual cluster 1 product network outputs, the individual cluster 1 members are more likely to purchase complements, i.e. on the border salsa and tortilla chip, which are deliberately bought together much more often by sheer chance, and household essential items. Potential marketing actions can be combination of top-selling items in household essentials, and canned, packaged foods and vegetables & fruit categories for the purpose of boosting sales and increasing profit.

Cluster 2 is ranked as second according to CLV score. Members in Cluster 2 shopped very frequently. Thus; although they spent moderately low amount of money per each visit, their total spending was high on average. Their last shopping happened very short time ago. The number of days between two consecutive visits was very low.

This cluster can be evaluated as "attached to brand segment". The focus for this segment would be to find ways to increase their average amount of spending per each visit. Since they have visited store frequently and recently, they are more likely to be early adopters for new products and offers. Moreover, up-selling marketing strategies may be beneficial to increase the value of the segment.

According to individual cluster 2 product network, the combination of item sets in vegetables & fruits, and in canned, packaged foods are mostly common. Possible up-selling marketing actions can include offering additional similar product options, or other versions of current products i.e. organic fruits and vegetables, to fulfill better their needs. Especially, purchasing of organic products may be boosted for U.S. citizens, since there has been a growing concern to deal with obesity. We assume that organic options are more profitable for Sam's Club. Combining complement items such as charcoal starters and charcoal briquettes can be advantageous to boost sales and revenue.

As aforementioned, the remaining clusters of individuals (Cluster 3, 4, 5, 6, 7, and 8) had relatively lower customer lifetime values, so we created a general individual product network for them. The following evaluations are based on this network output, and clustering outputs for each segment.

Cluster 3 consisted of rarely shopping customers. Their last shopping happened very short time ago. They were spending low amount of money per each visit and their total spending was low, as well. There was moderate number of days between their two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were low.

This cluster can be interpreted as a "promising segment". These customers may be informed regularly to increase their brand awareness and free trials may be offered to attract these members.

According to general individual product network outputs, individual members are more likely to purchase the item sets which reflects their daily habits and activities and daily essentials. Due to being promising segment, it would be advantageous to spend effort for attracting them with general promotions with top-selling items, i.e. bananas, navel oranges, and rotisserie chicken, in top-selling categories in product network structure.

Cluster 4 consisted of members who shop occasionally. They made their last shopping a short time ago. They were spending low amount of money per each visit and their total spending was moderately low. They had low number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were low.

This cluster can be evaluated as a "new comers segment". It would be beneficial to encourage them with loyalty programs and similar actions in order to maintain close relationship with the members.

Based on general individual product network, new comers segments can be kept in countenance by offering top item sets which are frequently bought together in the category of vegetables & fruits, meat, poultry, seafood, & eggs, and outdoor, patio & garden.

Cluster 5 consisted of occasionally shopping customers. They made their last shopping a short time ago. They spent high amount of money per each visit and their total spending was high. They had low number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were high.

This cluster can be regarded as "potential to evolve as star segment". To turn this segment into a star segment, the members may be encouraged to visit the store frequently by offering them cross-selling opportunities based on their past shopping behaviors. Moreover, limited time offers may be useful to attract this segment.

Item sets which include relatively authoritative items would be advantageous in terms of up-selling and cross-selling opportunities in order to attract these members for the purpose of evolving as star segment. This is important for Sam's Club, since if the appropriate cross-selling and up-selling are offered with the help of knowledge revealed from general individual network structure, this segment will have potentially evolve as star segment with a higher CLV score and better shopping characteristics.

Members in Cluster 6 shopped rarely. Their last shopping happened a while ago. They spent at an average amount of money per each visit and their total spending was moderately low. They had moderate number of days between two consecutive visits.

This cluster are more likely to be “low motion customers segment”. Because their last shopping happened a long time ago, they may be reactivated via general offers and reminding information in order to create brand awareness.

Due to the fact that this segment can be described as low motion customer segment, general promotions with top-selling item sets can be offered to reactivate them.

Cluster 7 had the lowest CLV score. Members in Cluster 7 shopped rarely. Their last shopping happened a short time ago. They spent low amount of money per each visit and their total spending was low. They had extremely high number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were low.

This cluster’s members can be accepted as “lost members”. Information about general offers can be useful to attract this segment.

Since the cluster 7 members can be referred to as ”lost members”, similarly with Cluster 6 case, they can be encouraged with mostly general top-selling item sets.

Cluster 8 shopped occasionally. Their last shopping happened a while ago. They spent moderately low amount of money both per each visit and in total. They had moderate number of days between two consecutive visits.

The members in this cluster act as “low value customers segment” in the concept of customer segmentation. Management should be careful about spending the correct amount of time, cost and effort to acquire these customers, to offer service them.

In order to increase the value of Cluster 8, referred to low value customer segment, it can be beneficial to offer general promotions with top-selling products in the category of vegetables & fruits, meat, poultry, seafood, & eggs, and outdoor, and patio & garden.

4.4.2 Business Clusters Marketing Implications

Cluster 1 is the leading cluster. Its members shopped very frequently. Their last shopping happened very short time ago. They spent moderately high amount of money per each visit and their total spending was high. They had low number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were high.

We can call this cluster "stars segment". The customers in this cluster have good values from all point of views. They should be solicited frequently by rewards and by building one-to-one relationships. In addition, they can become early adopters for new products and offers by promoting the brand.

Both individual items in the category of Household Essentials and complement items, i.e. pure-cane sugar, non-dairy creamer, and roast coffee, are more likely to be promoted to business 1 cluster members. They are also accepted stars segments based on their shopping characteristics revealed from two-stage clustering, so markup complements, up-selling promotions, and cross-selling promotions are helpful to attract these members for the purpose of boosting the profitability from this segment.

As aforementioned, the Cluster 2, 3, 4, and 5 have relatively lower customer lifetime values, so we created a general business product network for them. The following evaluations are based on the general product network outputs, and two-stage clustering outputs for each segment.

Cluster 2 shopped long time ago but before that date they shopped frequently. They spent low amount of money per each visit and their total spending was moderately low. They had low number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were low.

This cluster can be called “need-attention segment”. By offering discounts, its members may be reattracted to the stores. Management may encourage them to participate to special lottery and gaming opportunities to motivate them for spending more money per each shopping visit.

Special lottery or gaming opportunities can be implemented on top-selling items revealed from rules in general business network to attract this segment for the purpose of increasing their average spending amount in each visit. Using top-selling products such as Bountry Paper Towel, LYSOL Disinfectant, Dawn Dish Detergent etc., cross-selling marketing actions can be developed.

Cluster 3 had one of the lowest CLV scores. Its members in Cluster 3 shopped very rarely. Their last shopping happened very short time ago. They spent very low amount of money both per each visit and in total. They had high number of days between two consecutive visits. The variation in the total spending of the members and the variation in the total number of items purchased by them were low.

We can call this cluster “potential new and low value segment”. Monetary and total spending values of these members were very low. Although they shopped rarely, their average recency value was good. Therefore; they may be potentially new customers. Its members may be informed by creating better brand awareness. Additionally, free trials may be useful to attract these customers.

Especially, top-selling items in top-selling categories can be advantageous to create brand awareness for this segment. Moreover, due to being potential new and low value segments, their shopping habits are carefully investigated in order to fulfill their needs in a better way.

Cluster members shopped rarely but they had low number of days between two consecutive visits. Their last shopping happened short time ago. They spent moderately low amount of money per each visit and their total spending was moderately

low as well. The variation in the spending of the members and the variation in the total number of items purchased by them were low.

We can call this cluster "new comer segment". The members in this cluster may be encouraged by offering loyalty programs to them. Loyalty programs may be helpful to maintain relationship between the retailer and the customers, to increase total number of shopping trips of customers, and to increase their average spending.

Due to the fact that this segment has relatively low number of days between two consecutive visits, they can be a good candidate for cross-selling and up-selling opportunities, especially in frequently bought together item sets in general business product network.

Cluster 5 had one of the lowest CLV scores. Members in Cluster 5 shopped very rarely. Their last shopping happened very long time ago. They spent an average amount of money per each visit but their total spending was low. They had very high number of days between two consecutive visits. The variation in the total number of items purchased by them was low.

This cluster can be interpreted as "question mark or problem-child segment". The customers in this cluster have bad values from all point of views except spending an average amount of money per each visit. Retailer may reach these customers via reactivation campaigns in order to strengthen the relationship with these members and to increase their common awareness about the brand. Top-selling brands in general business network structure, especially in household essentials and canned, packaged goods are better candidates to reactivate these members.

Cluster 6 is the second-leading segment in terms of CLV. Its members shopped rarely. Their last shopping happened a while ago. They spent high amount of money per each visit and their total spending was moderately high. They had moderate number of days between two consecutive visits. The variation in the total spending of the

members and the variation in the total number of items purchased by them were high.

This cluster can be evaluated as "potential star segment". From marketing perspective, members in potential star segment may be encouraged with cross-selling opportunities and offering related products based on their past shopping behaviors. This may lead these customers to shop frequently by incurring high activation.

Since this segment is accepted as potential star segment, they tend to be better responsive candidate for the top-selling item sets in the business 6 product network. Moreover, up-selling opportunities with relatively authoritative items, such as Bountry Paper Towel, GLAD Trash Bag, Pure Cane Sugar, Non-Dairy Creamer, BOUNCE Singles 160CT etc., in the network structure are advantageous for the purpose of boosting sales and increasing profit from this segment.

Chapter 5

Conclusion and Further Suggestions

Due to the fact that there are massive amounts of available data all around the world, big data analytics has become a very important topic. As the data grow, the need for businesses to achieve more reliable and accurate data-driven management decisions and to create value with big data applications increases. That is the reason why big data analytics becomes a primary tech priority today. Most of the businesses tend to be able to gain competitive advantages by utilizing big data applications in all of their business areas. Big data revolution enables practitioners to discover various methods to gain insights in consumer behavior. Based on creative and result-oriented big data applications, companies have a chance to come up with appropriate marketing strategies such as planning of advertising budgets and price strategies, forecasting demand, determining customer satisfaction levels, increasing customer satisfaction, and planning product categories.

Initially, we conducted a detailed literature survey as seen in Chapter 2. We summarized all the articles by pointing out their utilized methods, tools, attributes in proposed methodology, limitations and further suggestions.

Subsequently, a data analysis was conducted in Chapter 3 which included in comprehensive applications in the context of SAM'S CLUB case study. We began with data collection step in data analysis part. Data collection comprised information on Sam's Club, UA_SAMSCLUB_SMALL database, corresponding attributes which were given in Appendix A . Then, selected attributes for the research with an entity relationship diagram in Figure 3.1 were presented. Descriptions and explanations for selected transaction attributes which were extracted directly from Sam's Club database were also given. The aim of the data derivation was to configure the data by adjusting the data types, deriving new transaction attributes and finally deriving customer attributes for customer tables. In order to derive customer attributes, we conducted a literature review in consumer behavior field. Then, derived transaction attributes and derived customer attributes were provided with their descriptions. At the data cleaning step, after mentioning master data issues to show the requirement of cleaning phase for performing the descriptive analysis on the datasets, the cleaning steps were summarized for both transaction datasets and customer datasets.

In the context of descriptive analysis, we analyzed the distribution of daily transactions per part of day, the distribution of visits and members per parts of day, and the distribution of transactions per day for top six categories are visualized, as well as evaluating distributions of recency, frequency, and monetary attributes, and finally, correlation analysis for the purpose of deciding on attributes that were included in two-stage clustering.

Predictive analysis performed by utilizing two-stage clustering with k-medoid and hierarchical cluster analyses to form customer segments for individual and business members. We aimed at achieving specific and supportive results for customer segments by interpreting statistical results of cluster analysis in order to evaluate them. The attributes included in cluster analysis were selected based on correlation analysis and experts' opinions. For creating groups of business and individual customers, k-medoid algorithm was used. Then, hierarchical clustering was used for clustering

groups of customers in order to achieve distinct customer segments. The successive usage of these two algorithms is an important contribution, allowing the deeper understanding of each member and each cluster (a set of similar members).

After two-stage clustering phase, we performed a customer lifetime value (CLV) estimation to compute CLV scores of each customer segment and to rank them based on CLV scores within a certain time period from 7/31/2005 through 11/2/2006. To compute CLV scores of each segment, we utilized the weighted RFM model with three different weighting methods: subjective, objective, and aggregated. This enabled us to compare the CLV results based on different weighting methods. We targeted to reveal managerial insights and develop marketing strategies for each segment. These insights are expected to improve the business processes of the companies and their performances, create new markets and opportunities as well as ensure sustainable competitive advantages.

After all data analyses steps, we used HITS algorithm in product network analysis to achieve valuable insights from generated patterns by aiming at discovering cross-selling effects, identifying recurring purchasing patterns, and complement, trigger and substitute products within the networks. This is important for practitioners in real-life applications in terms of emphasizing the relatively important transactions by ranking them with corresponding item sets.

From practical point of view, we foresee that our proposed methodology is adaptable and applicable to other similar businesses throughout the world by providing a road map for the potential applications.

There are some limitations and further suggestions for this thesis. Due to the fact that we utilized a data set which covers the transactions that happened approximately 14 years ago, the recommended marketing strategies may differ from the ones that will be based on the current data set(s). One of the strong reason behind this foresight is that we expect that there will be much more items, and brands in

the current state. Therefore, our proposed methodology in the data analysis and the product network setting can be applied and reevaluated with the current data sets if they are accessible.

Secondly, a comprehensive scenario analysis can be developed by evaluating the rules revealed from product network analysis. For example, it would be important for the business to analyze the profits of top item sets of the rules in product network structures according to the offered promotions for each segment, if and only if some additional data, i.e. actual, or forecasted marketing promotion costs, are available. Finally, rules extracted from HITS algorithm can be helpful for us to discover communities of products for further analysis.

Appendix A

Sam's Club Metadata

Attribute	Description	Values
ACTIVITY_CD	Activity Code	Y, N
BRAND_NAME	Name of the brand associated with the item	Null, name of brand
BUS_CR_TYP_STAT_CD	Business Credit Type Status Code	0-10
CARD_HOLDER_NBR	Card holder within an account	1-99
CATEGORY_NBR	Number assigned to a category of items	Null, 0-99
CMPLMNTRY_CARD_CNT	Number of extra cards given to an account	0-4
COLOR_DESC	Color description of an item	White, Almond, etc
CREATE_DATE	Date the item was created	Date
EFFECTIVE_DATE	Date the item began to be sold	Date
ELITE_STAT_CODE		0-4
EXPIRATION_DATE	Expiration date of an item	Date
FINELINE	Combination of category_nbr & sub_category_nbr	4 digit number
ISSUING_CLUB_NBR	The club that the member originally joined	1-150
ITEM_NBR	The number assigned to every different item for sale	Unique number (PK)
ITEM_QUANTITY	The quantity of a unique item that is scanned	
JOIN_DATE	Date the member joined the club	Date
LAST_RENEWAL_DATE	Last date that the member renewed their membership	Date
MEMBER_CODE		1,A,D,E,G,V,W,X,Y
MEMBER_STATUS_CD		A,D,E,T
MEMBER_TYPE		1,A,E,G,V,W,X
MEMBERSHIP_NBR	The number assigned to the member upon joining the club	
MFG_NBR	Number representing a manufacturer	
OBSOLETE_DATE	The date an item is no longer sold	Date
OPERATOR_NBR		
PRIMARY_DESC	The description of an item	Teal X-Large etc
QUALIFY_ORG_CODE		Null, 015-3001
REFUND_CODE	Code to indicate a return transaction	0 = Not Return, 1= Return
REGISTER_NBR	The register identification number where the transaction took place	1-85
RENEWAL_DATE	Date a membership should be renewed	Date
SALES_TAX_AMT	Tax charged for total visit	
SECONDARY_CARD_CNT	Number of cards other than primary card assigned to the membership	
SECONDARY_DESC	Additional description of an item	Sweatshirt, gift set etc
SIC	Standard Industry Classification code	783700, 443700 etc

FIGURE A.1: Sam's Club Metadata

SIZE_DESC	Text description of the size of the item, including clothing and non-clothing items	15CUFT, LARGE, etc
STATUS_CHG_DATE	The date an item last changed its status code	Date
STATUS_CODE	Whether an item is active or deactive	A = Active, D = Deactive
STORE_NAME	The name of the store	
STORE_NBR	Store identification number	1-150
SUB_CATEGORY_NBR	The number assigned to a sub_category of items	
TAX_COLLECT_CODE	Purchase taxable or not	0,1
TENDER_AMT	The amount tendered for the purchase	
TENDER_TYPE	Type of payment used	0 - Cash 1 - Check 2 - Gift Card 3 - Discover 4 - Direct Credit 5 - Business Credit 6 - Personal Credit
TOT_SCAN_CNT	Total number of scanned items per transaction	
TOT_UNIQUE_ITM_CNT	The number of unique items purchased per transaction	0-84
TOT_UNIT_COST	The cost of the item (scrubbed)	
TOTAL_SCAN_AMOUNT	The total number of items scanned per visit number	
TOTAL_VISIT_AMT	The total value of the entire transaction	
TRANSACTION_DATE	Date of the transaction	
TRANSACTION_TIME	The time of day that the transaction started	
UNIT_COST_AMOUNT	Cost/Unit (scrubbed)	
UNIT_RETAIL_AMOUNT	Purchase Price/Unit (scrubbed)	
VENDOR_NBR	The number of the vendor that supplies the item	
VISIT_NBR	Every time a member goes to the register and has their membership card scanned, this number is then created	9 digit #
VNPK_CUBIC_FT	How many cubic feet does a vendor pack take up	
VNPK_QTY	The quantity of items in a vendor pack	
ZIP_CODE	The zip-code of the store	
ZIP_CODE	The zip-code of the member	

FIGURE A.2: Sam's Club Metadata (continued)

Bibliography

[1] S. R. Department, "Revenue from big data and business analytics worldwide from 2015 to 2022 (in billion u.s. dollars)." https://www.statista.com/statistics/551501/worldwide-big-data-business-analytics-revenue/.

[2] "Most important data-driven marketing objectives." https://www.marketingcharts.com/charts/ascend2-most-important-data-driven-marketing-objectives-jul2017/.

[3] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," *Journal of Business Research*, vol. 69, no. 2, pp. 897–904, 2016.

[4] S. C Matz and O. Netzer, "Using big data as a window into consumers' psychology," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 7–12, 12 2017.

[5] E. Aktas and Y. Meng, "An exploration of big data practices in retail sector," *Logistics*, vol. 1, p. 12, 12 2017.

[6] L.-L. L. Chiang and C.-S. Yang, "Does country-of-origin brand personality generate retail customer lifetime value? A Big Data analytics approach," *Technological Forecasting and Social Change*, vol. 130, no. C, pp. 177–187, 2018.

[7] J. Sheng, J. Amankwah-Amoah, and X. Wang, "A multidisciplinary perspective of big data in management research," *International Journal of Production Economics*, vol. 191, pp. 97 – 112, 2017.

[8] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "Rfm ranking – an effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences*, 2018.

[9] E. T. Bradlow, M. Gangwar, P. Kopalle, and S. Voleti, "The Role of Big Data and Predictive Analytics in Retailing," *Journal of Retailing*, vol. 93, no. 1, pp. 79–95, 2017.

[10] D. Rigby, "Management tools and techniques: A survey," *California Management Review*, vol. 43, pp. 139–160, 12 2001.

[11] K. Sun and F. Bai, "Mining weighted association rules without preassigned weights," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 489–495, April 2008.

[12] H. v. S. Bernard Marr, *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. 01 2015.

[13] R. T. Bedeley, T. Ghoshal, L. S. Iyer, and J. Bhadury, "Business analytics and organizational value chains: A relational mapping," *Journal of Computer Information Systems*, vol. 58, no. 2, pp. 151–161, 2018.

[14] R. H. Chiang, V. Grover, T.-P. Liang, and D. Z. G. Editors, "Special issue: Strategic value of big data and business analytics," *Journal of Management Information Systems*, vol. 35, no. 2, pp. 383–387, 2018.

[15] F. Rodrigues and B. Ferreira, "Product recommendation based on shared customer's behaviour," *Procedia Computer Science*, vol. 100, pp. 136 – 146, 2016. International Conference on ENTERprise Information Systems/International

Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016.

[16] W. R. Smith, "Product differentiation and market segmentation as alternative marketing strategies," *Journal of Marketing*, vol. 21, no. 1, pp. 3–8, 1956.

[17] J. Wu and Z. Lin, "Research on customer segmentation model by clustering," in *Proceedings of the 7th International Conference on Electronic Commerce*, ICEC '05, (New York, NY, USA), pp. 316–318, ACM, 2005.

[18] P. W. Murray, B. Agard, and M. A. Barajas, "Market segmentation through data mining: A method to extract behaviors from a noisy data set," *Computers and Industrial Engineering*, vol. 109, pp. 233 – 252, 2017.

[19] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, "Retail business analytics: Customer visit segmentation using market basket data," *Expert Systems with Applications*, vol. 100, pp. 1 – 16, 2018.

[20] S. Tripathi, A. Bhardwaj, and P. E, "Approaches to clustering in customer segmentation," *International Journal of Engineering and Technology*, vol. 7, p. 802, 07 2018.

[21] S. Huang, E. Chang, and H. Wu, "A case study of applying data mining techniques in an outfitter's customer value analysis," *Expert Systems with Applications*, vol. 36, pp. 5909–5915, 4 2009.

[22] H.-C. Chang and H.-P. Tsai, "Group rfm analysis as a novel framework to discover better customer consumption behavior," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14499 – 14513, 2011.

[23] S. Han, Y. Ye, X. Fu, and Z. Chen, "Category role aided market segmentation approach to convenience store chain category management," *Decision Support Systems*, vol. 57, pp. 296 – 308, 2014.

[24] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via rfm model and rs theory," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 4176 – 4184, 2009.

[25] C.-Y. Tsai and C.-C. Chiu, "A purchase-based market segmentation methodology," *Expert Systems with Applications*, vol. 27, no. 2, pp. 265 – 276, 2004.

[26] A. Sheshasaayee and L. Logeshwari, "An efficiency analysis on the tpa clustering methods for intelligent customer segmentation," in *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 784–788, Feb 2017.

[27] G. Tirenni, C. Kaiser, and A. Herrmann, "Applying decision trees for value-based customer relations management: Predicting airline customers' future values," *Journal of Database Marketing & Customer Strategy Management*, vol. 14, pp. 130–142, Jan 2007.

[28] M. N. Ray, "Ahp based data mining for customer segmentation based on customer lifetime value," 2016.

[29] D.-R. Liu and Y.-Y. Shih, "Integrating ahp and data mining for product recommendation based on customer lifetime value," *Information and Management*, vol. 42, no. 3, pp. 387 – 400, 2005.

[30] A. Hiziroglu and S. Sengul, "Investigating two customer lifetime value models from segmentation perspective," *Procedia - Social and Behavioral Sciences*, vol. 62, pp. 766 – 774, 2012. World Conference on Business, Economics and Management (BEM-2012), May 4–6 2012, Antalya, Turkey.

[31] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study," *Procedia Computer Science*, vol. 3, pp. 57 – 63, 2011. World Conference on Information Technology.

[32] M. Seyed hosseini, A. Maleki, and M. Gholamian, "Cluster analysis using data mining approach to develop crm methodology to assess the customer loyalty," *Expert Syst. Appl.*, vol. 37, pp. 5259–5264, 07 2010.

[33] M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted rfm model in retail banking context," *Procedia Computer Science*, vol. 3, pp. 1327 – 1332, 2011. World Conference on Information Technology.

[34] M. Hosseini and M. Shabani, "New approach to customer segmentation based on changes in customer value," *Journal of Marketing Analytics*, vol. 3, 09 2015.

[35] A. Santoso and A. Erdaka, "Customer loyalty in collaborative consumption model: Empirical study of crm for product-service system-based e-commerce in indonesia," *Procedia Computer Science*, vol. 72, pp. 543–551, 12 2015.

[36] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of The ACM - JACM*, vol. 46, 01 1999.

[37] K. Wang and M.-Y. T. Su, "Item selection by "hub-authority" profit ranking," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), pp. 652–657, ACM, 2002.

[38] L. Ramasamy, "A framework for mining weighted association rule using hits progress," *ciit internation journal of Data Mining Knowledge Engineering*, vol. 12, pp. 183–187, 08 2010.

[39] L. Kaufman, P. Hopke, and P. Rousseeuw, "Using a parallel computer system for statistical resampling methods," *Computational Statistics*, vol. 2, pp. 129–141, 01 1988.

[40] D. Yu, G. Liu, M. Guo, and X. Liu, "An improved k-medoids algorithm based on step increasing and optimizing medoids," *Expert Syst. Appl.*, vol. 92, pp. 464–473, Feb. 2018.

[41] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336 – 3341, 2009.

[42] V. T and S. T, "Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points," *Journal of Computer Science*, vol. 6, 06 2010.

[43] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014.

[44] J. H. W. Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[45] D. Esztergár-Kiss and B. Caesar, "Definition of user groups applying ward's method," *Transportation Research Procedia*, vol. 22, pp. 25–34, 12 2017.

[46] P. Farris, *Marketing Metrics: 50+ Metrics Every Executive Should Master.* Wharton School Pub., 2006.

[47] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study," vol. 3, 01 2010.

[48] B. Stone and R. Jacobs, *Successful Direct Marketing Methods.* McGraw-Hill, 2001.

[49] M. Seyed hosseini, A. Maleki, and M. Gholamian, "Cluster analysis using data mining approach to develop crm methodology to assess the customer loyalty," *Expert Syst. Appl.*, vol. 37, pp. 5259–5264, 07 2010.

[50] T. Saaty, "Multicriteria decision marking : the analytic hierarchy process : planning, priority setting, resource allocation / thomas l. saaty," *SERBIULA (sistema Librum 2.0)*, 05 2019.

[51] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning – with Applications in R*, vol. 103 of *Springer Texts in Statistics*. New York: Springer, 2013.

[52] P. Koele, "Multiple attribute decision making: An introduction, sage university paper series on quantitative applications in the social sciences, 07-104, k. paul yoon and ching-lai hwang, thousand oaks, ca: Sage, 1995, 73 pp., isbn 0-8039-5486-7.," *Journal of Behavioral Decision Making*, vol. 10, no. 2, pp. 151–151, 1997.

[53] I. KWANGSUN YOON Senior Member and G. KIM, "Multiple attribute decision analysis with imprecise information," *IIE Transactions*, vol. 21, no. 1, pp. 21–26, 1989.

[54] N. Verma, *Market Basket Analysis with Network of Products*. Master's thesis, Ca Foscari, University of Venice, Venice, 2017.

[55] I. F. Videla-Cavieres and S. A. Ríos, "Extending market basket analysis with graph mining techniques: A real case," *Expert Syst. Appl.*, vol. 41, pp. 1928–1936, 2014.

[56] M. E J Newman, *Networks: An Introduction*. 01 2010.