SALES MANAGEMENT ANALYSIS WITH TEXT MINING METHODS

A THESIS SUBMITTED TO APPLIED DATA SCIENCE PROGRAM OF GRADUATE SCHOOL

OF

TED UNIVERSITY

BY

TUĞBA SOYER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

IN GRADUATE SCHOOL OF TED UNIVERSITY

FEBRUARY 2020

Approval of the Graduate School of TEDU

Associate Prof. Mehmet Rüştü Taner
Director of Graduate School

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Associate Prof. Semih Tümen
Director of Program

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Associate Prof. Semih Tümen

…………………………………….
Supervisor

Examining Committee Members

Associate Prof. İbrahim Ünalmış (TED University, Business Administration)

Dr. Güneş A. Aşık (TOBB University of Economics and Technology, Economics)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Tuğba Soyer

Signature:

ABSTRACT

SALES MANAGEMENT ANALYSIS with TEXT MINING METHODS

SOYER, Tuğba

MSc., Applied Data Science Program of Graduate School

Supervisor: Associate Prof. Semih TÜMEN

February 2020

The aim of this study is to examine offer contents of a major vehicle company using data mining techniques to categorize these offers based on difficulty level for the purpose of improving sales management process. The language of the offers is mainly in German. Categorizing the offers according to the difficulty level will allow the company to observe the changes in the customer requests over time and to make long term strategic sales management and marketing plans. A new method is developed inspired by the Support Vector Machine and Lexicon sensitivity analysis, and the offers are categorized with respect to their difficulty levels. At this stage, technical dictionaries are prepared with the support of the company's expert team. The study also constructs a large consolidated data set, which could serve as a basis for future sales management analysis of the offers. The data set used in this study contains 1,082,093 observations used in the vehicle sales process between 2006-2019. As a result of this study, unstructured and distributed offers data are converted into a data set ready for analysis. It is seen that the content of the special requests affects the difficulty level more than the number of requests included in the offer. It is understood how market demands have changed over time for the countries.

Keywords: Sales Management, Text Mining, Text Categorization, Lexicon Sentiment Analysis, Machine Learning, Automotive, German

ÖZ

METİN MADENCİLİK YÖNTEMLERİ İLE SATIŞ YÖNETİMİ ANALİZİ

Soyer, Tuğba

Yüksek Lisans, Uygulamalı Veri Bilimi Bölümü

Tez Yöneticisi: Doç. Dr. Semih Tümen

Şubat 2020

Bu çalışmanın amacı, satış yönetim sürecini iyileştirmek hedefiyle teklifleri zorluk seviyelerine göre kategorize etmek için veri madenciliği tekniklerini kullanarak büyük bir araç şirketinin teklif içeriğini incelemektir. Tekliflerin dili çoğunlukla Almancadır. Tekliflerin zorluk seviyesine göre sınıflandırılması, şirketin müşteri taleplerindeki değişiklikleri gözlemlemesine ve gelecek planları yaparken fikir edinmesine katkı sağlayacaktır. Destek Vektör Makinesi ve Lexicon duyarlılık analizinden esinlenerek yeni bir yöntem geliştirilmiş ve teklifler zorluk seviyelerine göre kategorize edilmiştir. Bu aşamada değerlendirme uzmanı grubun desteğiyle teknik içerikli özel sözlükler hazırlanmıştır. Ayrıca, tüm tekliflerin tek bir veri setinde birleştirilmesi ve içeriğin analiz edilmesi, satış yönetimi sürecindeki diğer olası çalışmaların temelini oluşturmuştur. Bu çalışmada kullanılan veri seti 2006-2019 yılları arasında araç satış sürecinde kullanılan 1,082,093 gözlemi içermektedir. Çalışma sonucunda, yapılandırılmamış ve dağınık teklif verileri analiz edilmeye hazır bir veri setine dönüştürülmüştür. Teklif içerisinde yer alan özel istek sayısı yerine isteklerin içeriğinin zorluk seviyesini daha çok etkilediği görülmektedir. Ülke bazında piyasa taleplerinin zaman içinde nasıl değiştiği anlaşılmaktadır.

Anahtar Kelimeler: Satış Yönetimi, Metin Madenciliği, Metin Kategorizasyonu, Lexicon Duygu Analizi, Makine Öğrenimi, Otomotiv, Almanca

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor Associate Prof. Semih Tümen for his support, practical suggestions and guidance throughout my thesis.

My department managers and colleagues always supported me throughout my studies. Specifically, I would also like to thank Dinç Bulut and Fatih Pilevne for their support in sharing the data set used in this study with me and for their support in the business process for the completion of the study.

Last and most precious, I want to thank my husband Burak Soyer for believing and supporting me throughout my all studies.

It would not have been possible to complete this thesis without your support and understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Motivation

Data science techniques are used in management analysis in many different contexts including, but are not limited to, profile analysis on social media, product marketing and communication, analyzing the period of artworks by doing emotion analysis, performance measurement of production companies, and development of sales management processes.

The data in social media channels such as Twitter, Instagram, where every second information flow is available, can be used for many different purposes. The study of categorizing both users and shared images by examining user profiles is an example of data analysis performed on social media channels (Hu et al., 2014). The findings suggest that the consumption of traditional news media has decreased while the number of online news users has increased (Stempel et al., 2000). In the study conducted by Gorodnichenko et al (2018), the text data of two major political events on Twitter are analyzed using the Sentiment analysis method and the questions of how the communication revolution affects the information flows between individuals and how they can affect these flows are examined. The rapid and widespread dissemination of information on social media affected the news media, making social media a powerful source of news. This has created the risk that social media, especially during high-impact events, may publish false drop and feint news. The study conducted by Karol Jan Borowiecki is another example of how wide the data analysis is. Using approximately 1,400[1] letters written by Wolfgang Amadeus Mozart, Ludwig van Beethoven and Franz Liszt throughout their lives, it was analyzed which emotions were revealed during the creation of the artworks. With the linguistic analysis method, how positive emotions (e.g. happiness) or negative emotions (e.g. grief) are reflected in the artworks is examined and the intensity of emotions during the creation of artworks were detected (Borowiecki, 2017).

[1] 1400 words are selected within the scope of the relevant study. It is not the total words in the artworks of 3 artists. See section IIIB in relevant study for a discussion of the selection.

Knowledge is the most valuable asset of a manufacturing enterprise, as it enables a business to differentiate itself from competitors and to compete efficiently and effectively to the best of its ability. Knowledge exists in all business functions, including purchasing, marketing, design, production, maintenance, and distribution; but knowledge can be notoriously difficult to identify, capture and manage (Harding et al., 2006). Moreover, over the years, the growth and meaninglessness of the data caused problems as well as being important. In order to carry out a retrospective study or to make decisions by making predictions for the future, there was a need to make the data meaningful and to obtain results through data analysis. Data processing and reporting are important for the firm to examine its own performance as well as to ensure its continuity in the market.

Data used in this study are gathered from a company operating in the automotive/vehicles sector. For products with high price performance such as passenger car, house etc., price offers should be prepared. The buyer requests a configuration with details of the product to be purchased and evaluates the quotation offered for the product. After the delivery of the offer, the product may be purchased or not. Even if the sale is not materialized, the offer process has to be experienced. Within the scope of this thesis, the changes in the quality and quantity of offers prepared for the vehicles are examined. The vehicles are divided into different categories. First of all, it is aimed to convert the offer contents into a meaningful data set in terms of customer specific requests [2]. Offer content has never been combined as a single data set before in the company and has never been analyzed using data mining methods. Converting the unstructured data into a data ready to analyze is very valuable as it will facilitate access to knowledge. Additionally, determining the difficulty levels of offers using the text mining method provides to observe how the market demands have changed over time.

Furthermore, the categorization of the offers supports the department that prepares the offer in the sense that it provides information on how the offers vary over years. The vehicle product is diverse in itself. Thus, each sub-product can be prepared by one or several responsible persons. After the categorization process is completed, the decentralization can be reviewed again. Such that the decentralization in the sub-product groups may vary depending on the categorization

[2] In addition to the serial features and options of a vehicle configuration, specific request can be demanded by customer. It can be added in the vehicle configuration as text format. In the thesis content, these requests are referred to as customer specific request which is called CSR.

of the offer by the difficulty level. For example, if offers prepared in one vehicle type become increasingly difficult over the years, and vice versa, the number of type-based employees need to be revised. Of course, analyzing the allocation of workload across marketing specialists is not the ultimate goal of this study, but in a department/section where 40% of the workload is offer preparation, the result is highly relevant for planning and execution. Furthermore, not every offer is guaranteed to be ordered but only has the “possibility”. The conversion of offers to orders means that the company earns money. When a firm makes future plans, it makes better decisions if it observes how a constraint that directly affects profit has changed over the years. Therefore, it is very important for the firm that the difficulty level of the offers is categorized and how it has changed over the years.

Finally, the details of the customer demands are examined for the company that has adopted the principle of customer focus. The offers are categorized and how the offers vary according to difficulty level are examined. The results of the study provide an opportunity for the sales management department to assess/develop themselves and make long-term strategic decisions.

1.2. Outline of the Thesis

In Section 2, literature review of the works related to text categorization and text mining in sales management is given. This section shows the value and contribution of the study carried out by examining the different studies in the literature.

The data set used in the thesis is introduced in Section 3. How the data set is made suitable for analysis is detailed. Information about the applications used in the thesis is given.

The Results section includes what categorization of offers and text mining results mean for sales management. For each vehicle group based on year and country, how data is transformed into knowledge is explained with examples.

The Conclusion and Future Works section contains inferences from the analysis results and how these inferences can be evaluated in the future.

2. LITERATURE REVIEW

In this section, data mining techniques in the literature used for both sales management and text categorization are examined. Data mining techniques is used in many areas of sales processes such as analysis of sales rates, marketing, Customer Relationship Management and customer categorization.

In the study conducted by Bahng and Kincade, the effect of weather on sales was investigated by using sales data of a retailer of branded women's business wear. As a result of this study, it was determined that more seasonal clothes were sold during the sales periods in which drastic temperature changes occurred (Bahng and Kincade, 2012).

Banks and insurance companies are sectors with a large customer database. They want to sell products such as loans and life insurance policies to their customers. A specific customer group in the database has already purchased the products. Customers who have not purchased the remaining product must be identified. At this stage, the process can be managed with the using of data mining (Ling and Li, 1998).

The marketing model has shifted from product-focused to customer-focused (Harding et al., 2006). Customer focus means more sales, more profit. The need to better understand customer behavior and the interest of many managers to focus on those customers who can return long-term profits has changed how marketers view the world (Gorodnichenko et al., 2018). The studies in this field are called to as Customer Relationship Management (CRM). CRM is also valid to the manufacturing sector as well as in the service sector.

Data mining for customer segmentation is applied by Morita et al. (2000) to determine which customers are likely to shift from one cellular company to another. A rule induction algorithm on the transformed data are used to build rules and afterwards the potential moves of the customers are predicted.

Hui and Jha (2000) used DBMiner to develop a decision-support system using a customer service database. Neural networks and case-based reasoning were used to mine the unstructured customer data to identify the machine faults.

With the increase in text information stored in electronic format, automatic text classification or text categorization became increasingly important (Kibriya et al., 2004). Automatic text classification/categorization is performed by use of machine learning models (Dalal et al, 2002). In text categorization studies, three of the most common applied machine learning algorithms are generally used. These are “Naive Bayes”, “Maximum Entropy Classification”, “Support Vector Machines” (Pang et al., 2002). Another method used for text categorization is Sentiment Analysis. Sentiment analysis is a research area that aims to specify subjective information such as attitude, emotion and opinion in a text document (Wlezien et al., 2017). There are two main methods used in sentiment analysis which are machine-learning based (supervised methods) and dictionary-based method (Lexicon-Based method). Machine learning methods used for text categorization are valid in sentiment analysis. Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN) are some examples for machine-learning based methods in sentiment analysis. Lexicon-based model consists specific dictionary for relevant dataset. In the study by Gräbner, D. et al. customer reviews of hotels in London are categorized using Lexicon analysis (Gräbner et al., 2012).

Although Naive Bayes method is called as its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization is in tendency to get good results in a surprisingly way (Lewis, 1998); Indeed, Domingos and Pazzani (1997) show that Naive Bayes is optimal for certain problem classes with highly dependent features.

Maximum Entropy Classification (abbreviated by “MaxEnt” or “ME”) is the other alternative and effective technique in a number of natural language processing applications (Berger et al., 1996). Nigam et al. (1999) reveal that this method rarely predominates the other method “Naive Bayes” at standard text classification.

Support Vector Machines (SVMs) are one of the types of learning machines, based on the “Structural Risk Minimization” principle from the computational learning theory. The idea of structural risk minimization is to attain a hypothesis *h* for which we may assure the lowest true error. The true error of *h* is the probability that *h* makes an error on an unseen and randomly selected test example Besides, it is important that SVMs are so effective at traditional text categorization and occasionally predominate the other method “Naive Bayes”, and these are the large-margin ones compared to the Naïve Bayes and Maximum Entropy Classification methods

(Joachims, 1998). In addition, there is a known fact for SVM method which is the algorithm performs the categorization by assigning the same values for each input. However, it is impossible that all inputs have the same value. SVM method is supported by feature weighting or feature selection to solve this problem and significant impact on the performance of this method (Wang, 2010).

3. DATA and METHODOLOGY

In this section, the information about the data and applied methodology is explained. The flowchart diagram of the methodology is shown in Figure 1.

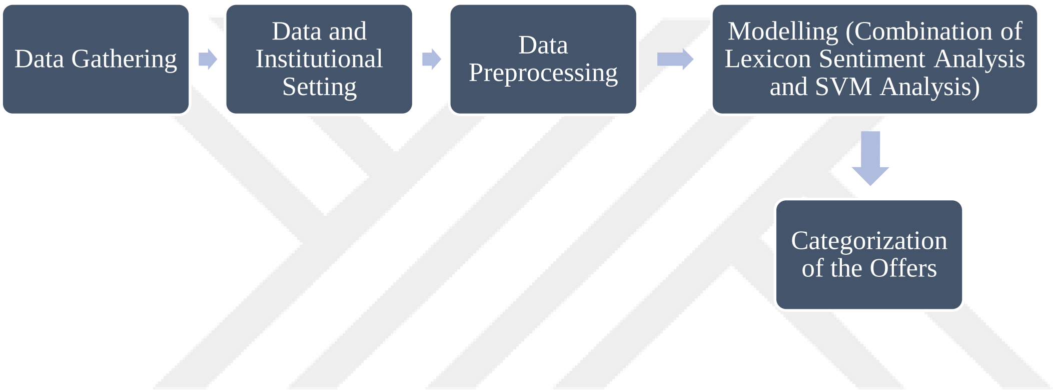

Data Gathering
Data and Institutional Setting
Data Preprocessing
Modelling (Combination of Lexicon Sentiment Analysis and SVM Analysis)
Categorization of the Offers

Figure 1 . Flowchart of Methodology

As shown in the flowchart diagram, firstly data is obtained and the content of the data to be analyzed should be examined in detail. Afterwards, a "data preprocessing" should be performed to obtain a data ready to be analyzed since there are characters in the data that affect the analysis processes and results. After the completion of the pre-processing process, Modelling (Combination of Lexicon Based Sentiment Analysis and SVM Analysis) process is examined. A new model[3] which is adapted to the data is created by taking a different perspective from Lexicon sentiment analysis and SVM text categorization studies in the literature. How to apply the created model is described. Finally, the categorization of offers, which is one of the desired results to be obtained within the scope of the thesis, is completed. While applying the extended model for categorization, various analyzes requested by the company are performed. These analyzes and their returns are explained in the Result and Future Work section.

[3] The new model can be called extended / combination of the Lexicon sentiment analysis and SVM text categorization. It is not a completely new model.

3.1. Data Gathering

Data must be obtained through the “Main System of the Offer” where the offers are submitted, via different report screens of the “Offer Details” and “Offer Evaluation Time Details” since there is no ready dataset for thesis study.

Two separate datasets are needed within the scope of the thesis. The first dataset contains the details of the offers and the second dataset contains the details of the evaluation process of the offers, the first dataset is named as “Offer Detail” and the second is called “Offer Time”.

Data can be acquired from the report screens in the main system by transferring to the spreadsheets. Maximum 20,000 rows of data are allowed at one time. So, the data have to be transferred to multiple spreadsheet files. Firstly, Offer Detail data was gathered. 82 separate files are transferred from report screen. By combining 82 different files, this dataset is created. Similarly, all the data required for the “Offer Time” dataset is transferred from report screens to 5 different files. By combining these files, the second dataset is created.

3.2. Data and Institutional Setting

The data used in this thesis belongs to one of the leader companies in the automotive sector. The company has two main product groups: bus and truck. The data includes information about the bus group. The company has bus manufacturing plants in 3 different countries. Within the scope of the study, vehicle groups produced by the central bus production factory with the most units and serving the premium market are examined which are High Premium, Premium Coach and Coach vehicles groups. Figure 2 shows the company's bus product tree.

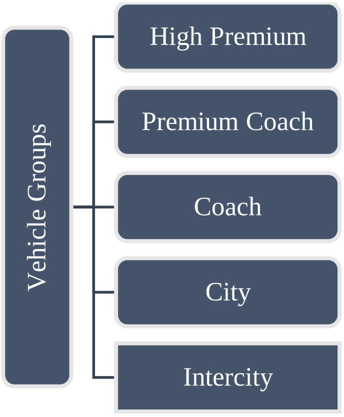


Figure 2. Vehicle Groups

Sales of vehicles are carried out through main sales companies and dealerships. The main sales companies and dealerships serve 36 countries including Turkey. During the sales process, a unique offer number for each customer needs to be configured. Vehicle configuration is created in a digital environment through a system used by seller. There are serial and option offered on the basis of vehicle type. The codes for the serial products are automatically selected in the configurator. The options are selected by the seller according to the customer's demands. For example, the seat type produced only from fabric is the serial for vehicles in the coach group, while the seat type produced from leather and fabric combination is available as an option. If the customer does not request the series and options, the requests are added to the configuration as “CSRs”. Since there are no ready codes in the configurator for them, requests are written to the configuration as Text. Continuing from the example of the seat, instead of the seat offered in the leather / fabric combination, a more luxurious and more comfortable seat is required. The feasibility of this request should be evaluated by the responsible persons in the production company and an offer should be submitted if this request is acceptable by production company.

After the evaluation is completed, the seller submits the offer to the end customer based on the costs of additional *CSRs* in the configuration. If the customer wants to buy the vehicle, the order is executed, and the vehicle is included in the production plan. Since order-based production is made, the production plan of a vehicle that has not been ordered cannot be made.

CSRs are the biggest feature that distinguishes the company from its competitors in the market. CSRs provide product flexibility to customers. So that a customer can buy the vehicle they dream of by means of flexibility.[4] Some customer specific requests can be implemented very easily, while others can be very difficult[5]. The difficulty of customer demands affects both the technical work and documentation process of the new request and the production process of the vehicle. In addition, an extra price is charged to the customer for each special request. Requests are carefully priced out by a responsible department in order to the price offer bears the technical work and production cost since they affect the company's profit rates. In fact, the company needs to observe how offers change over time while making a future plan.

[4] Requests are accepted if they are found appropriate based on international homologation rules and received approval for feasibility by the responsible departments in the manufacturing company.

[5] Technical feasibility and supply availability are the main criterions.

All in all, within the scope of this thesis, the contents of the offers up to the date of 26.08.2019 from the first offer registered in the system are examined. The offers are categorized according to difficulty levels, grouped by country and type, with using Text mining methods based on CSRs content. As a result of the categorization process, it is how the offers change over time.

As stated in the data gathering title, two different data sets were created within the scope of this study. In this section, the content of the datasets is detailed.

Dataset of the Offer Details

The offer details dataset contains the CSRs' texts and the details of the vehicle in a customer's configuration. As an example of how the order is received and processed: A bus transport company contacts the seller to add a new coach vehicle to its fleet. A new configuration is prepared for this customer. In this example, it is assumed that two new specific requests are demanded, apart from the series and options, which are curtain and seat fabric compatible with the customer's fleet. As mentioned before, new CSRs are added as text into the configuration. The configuration is transferred to the responsible person via a system for the offer evaluation. The configuration document and CSRs are displayed separately in this system. In this example, the offer is numbered dummy "12345". The responsible person evaluates the CSRs text and gives confirmation of the feasibility of the requests. Finally, he checks the overall configuration and, if the vehicle has all the appropriate conditions for the order, the system confirms "the vehicle is suitable for the order. Finally, the overall configuration is checked and, if the vehicle has the appropriate conditions for ordering, responsible person approves to order via the system. Details of the offer are automatically added to the database of the system. When a new offer is created, the next Number is automatically assigned as "Offer Number" which is "12345" in this example. CSRs in the new offer are recorded and added on following previous offer's CSRs. Finally, the system has the ability to transfer the database to spreadsheet files. The contents of the dataset are illustrated in the Table 1.[6]

[6] There is a total of 31 variables in this data set. All variables are not covered by the study. For example, pricing details. In order to obtain general information about this data set, sample table is created.

Offer Number	Vehicle Type	Customer Name	Unit	Country	CSR	Cost of the CSR
12345	High Premium	XYX Tour	5	Deutschland	Vorhangstoff in rot Farbe	……
12345	High Premium	XYX Tour	5	Deutschland	Fahrgastsitze schoepf in rot Farbe	……
12346	Coach	YXY Tour	3	Italien	Fahrzeug Lackierung Farbe in blau	……
12346	Coach	YXY Tour	3	Italien	Läufer azo in blau	……

Table 1. Contents of the Offer Details Dataset

The "Offer Details" data set consists of 1,082,093[7] observations and 31 variables. Offer number, customer name, offer creator name, the product group of the created offer, the CSR in the offer, price of each CSR, price calculation details of offers are a few examples among 31 variables. Distribution of the type of the variables in the dataset are that 11-numeric and 20-character variables

Summary Statistics of Offer Details Dataset

Vehicle groups, offer numbers, order numbers are summarized in this part.

Vehicle Groups

As previously mentioned, within the scope of this thesis study High Premium, Premium Coach and Coach vehicles' groups are examined.

Countries

As previously mentioned, a total of 36 countries, including turkey vehicle sales process is carried out. Countries are shown in Table 2.

Countries			
Germany	Belgium	Dubai	Bulgaria
France	Croatia	Slovenia	Greece
Italy	Czech Republic	Hungary	Romania

[7] 1,082,093 is that total number of CSR for all vehicle Groups.

Turkey	Estonia	Switzerland	Azerbaijan
Austria	Spain	Lithuania	Gabon
Poland	Norway	Luxembourg	Latvia
UK	Morocco	Bosnia	Mexico
Russia	Netherlands	Denmark	Israel
Sweden	Slovakia	Portugal	Iceland

Table 2. The countries which carried vehicle sales

<u>*Number of the Offer by the vehicle Type*</u>

Since 2006, when the first offer was prepared, 793 offers for high premium vehicles, 4731 offers for Premium Coach vehicles and 14314 offers for Coach vehicles were created. The number of offers for each vehicle group is shown in the Table 3.

Vehicle Groups	**Number of the Offer**
High Premium	793
Premium Coach	4731
Coach	14314

Table 3. The number of offers for each vehicle group

Dataset of the Offer Time

In addition to the content of the offer, there is a database in which the offer evaluation process is recorded. This data is obtained from a different report screen in the main system. The data set contains a total of 18 variables and 45,344 observations[8], including the offer number, vehicle group, customer name, the time when the offer evaluation was started to be evaluated, the time when the offer evaluation was completed, the evaluator and country of the offer. This data set and the offer numbers in the Offer Details data set are common. Offer number variable is common to the Offer Details and the Offer Time datasets. This variable is included in the offer

[8] There is a total of 45344 observations including vehicle groups outside the scope of the thesis.

details data set. Other variables in this dataset are not examined within the scope of this thesis. The content of the dataset is illustrated in the XG table.[9]

Offer Number	Vehicle Group	Customer Name	Evaluation Start Date	Evaluation End Date	Evaluator	Country
12345	*High Premium*	*XYX*	19.09.2016 15:24:41	22.09.2016 13:24:61	Burak	Deutschland
12346	*Coach*	*YXY*	11.02.2017 10:30:30	13.02.2017 10:30:30	Tuğba	Italy
12347	*Coach Premium*	*YXX*	12.02.2017 15:22:37	14.02.2017 15:22:37	Tuğba	Italy

Table 4. Table Contents of the Offer Time Dataset

3.3. Data Preprocessing

As a result of technological developments, the data used in the companies have shown various structurality. In fact, the data set can be categorized into structural, semi-structural and non-structural (Gandomi and Haider, 2015).

Structured data consists of clearly defined data types that make patterns easily searchable. So that, it is well organized and easily understandable by machine language. The structured data can be one-dimensional or multi-dimensional as well as a tabular or non-tabular format. For example, a table that lists sales amount by year.

Semi-structured data between the structured and unstructured data format covers continuity. In another words, it has not been well organized, but they are suitable to be organized and can be easily analyzed. Extensible Markup Language (XML) is an example of semi-structured data on the Web. The main reason that prevents it from being considered as structured data is XML documents containing user-defined data tags which make them machine-readable (Gandomi and Haider, 2015).

Unstructured data is not organized and under the influence of this situation, it cannot be easily processed and analyzed using conventional tools and methods. Image, video, mobile activity

[9] In order to obtain general information about Offer Time Dataset, sample table is created.

and social media activity are examples of unstructured data. In addition to given examples, text document is falls into the unstructured data category.

The fact that the non-structural data is in irregular format causes many factors that may negatively affect the results of the analysis and do not need to be analyzed. Data preprocessing should be used for standardizing the data and extracting unnecessary knowledge from unstructured text data. As a result of these preliminary processes, the efficiency of the analyzes is increased.

Text preprocessing methods vary by language and type of text document. For example, “@” is used frequently in the data has Emails while it is not in a document that holds sales data. In “Preprocessing Techniques for Text Mining articles” effective preprocessing steps are indicated as "Tokenization, Stop Word Removal, Stemming" (Kannan and Gurusamy, 2014). By means of these preprocesses, it is stated that the root words are defined, and the size of the text data are reduced (Vijayarani et al., 2015). It is not possible to apply these steps one to one for each data. Thus, these steps should be privatized based on the document to be analyzed.

In this section, data preprocessing steps used on Offer Data are explained. As shown in Figure 3, they consist of conversion of special letters, removing punctuations and spaces, tokenization and stemming.

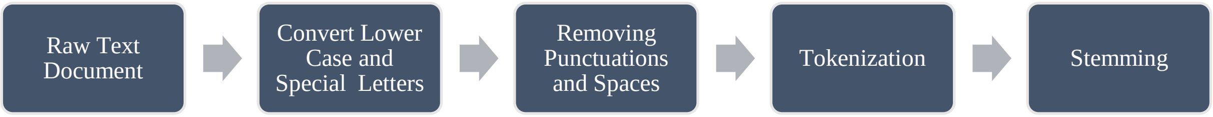

Raw Text Document
Convert Lower Case and Special Letters
Removing Punctuations and Spaces
Tokenization
Stemming

Figure 3. Text Preprocessing Steps

The following example of the Offer Text is used to show how each preprocessing step used in and affected the raw text document. This Offer Text is related to the customer's special request other than the serial or options offered for the curtain. Many different CSRs for curtain have been requested in different years. Without the year information, only the effect of the pre-processing steps on the request text are shown.

“ - Gardienenhaltebänder in Vorhangstoff ”

3.3.1. Convert Lower Case and Special Letters

One of the most common methods for text preprocessing step is conversion to lower case. Also, firstly lower-case step is applied. The letters ß and ä in the German alphabet can be written as "ss" and "ae". For example, the carpet can be written as "Läufer" or "Laeufer". The letters ß and ä are changed to "ss" and "ae" to prevent multiple versions of the same word. After conversion of lower case and the special letters preprocessing steps, example of the Offer Text seems as follow:

“ - gardienenhaltebaender in vorhangstoff ”

3.3.2. Removing Punctuations, Spaces and Stop words

Punctuation and spaces do not contain any information in the text analysis process. If there is a punctuation mark adjacent to the words, it may negatively affect the result. In addition, it makes the text document look heavier. In order to reduce the dimensionality of term space, punctuations, spaces and stop words should be removed (Vijayarani et al., 2015). After removing punctuation and spaces from the text document, given example of the Offer Text looks like as follows;

“ gardienenhaltebaender vorhangstoff ”

Thus, as shown in the example, "-" punctuation is deleted from the text document.

3.3.3. Tokenization

Tokenization is the process of creating the smallest meaningful units by dividing a flow of text into words, expressions, symbols, or elements such as tokens. Tokens may contain one or more words depending on the content of the analysis. Token lists become input for further processing, such as text mining. The language of the text to be tokenization determines the difficulty of the process. (Vijayarani et al., 2015). It is easier in languages where words are separated using white spaces. After this step, normally, text document consists of words list. But in this study, the text was not converted into a single word format. By using the Tokenization method, word lists for each vehicle groups are created to be used in the analysis phase. Because a request can have more than one interrelated word in it, which affects the difficulty level of the request.

3.3.4. Stemming

Another important step in text pre-processing is stemming. With stemming, words are converted to root forms. Depending on the structure of the language, the words may take different paragoge. Automatically eliminate these paragoge and use the root format of words. For example, “looked” is stemmed to look; “books” is stemmed to “book”.

The root of the words could be used in the newly created dictionary for analysis, but this method was not used because German was a language in which more than one word was combined and new words were derived. For example, “seitenwandverkleidung”. This word was formed by combining the words “seiten” - “wand” - “verkleidung”.

3.4. Modelling: Combining Lexicon Sentiment Analysis and SVM Analysis

Text categorization varies according to the content and language of the data set examined. The data set examined within the scope of the thesis is the text document of the CSR of a product group in the automotive sector. Most of the words in the data set are technical word.

Sentiment analysis is a research area that aims to specify subjective information such as attitude, emotion and opinion in a text document (Wlezien et al., 2017). Two main methods are used for sentiment analysis which are Lexicon-based method (unsupervised approach) and Machine Learning based method (supervised approach) (Kolchyna et al., 2015). Machine Learning based studies are hinged on text categorization, while Lexicon-based model is hinged on pre-defined dictionary creation.

In the dictionary-based method, which is Lexicon based sentiment analysis, a polarity score is assigned to the unigram in the dictionary, then the overall polarity score of the text is calculated as the sum of the polarities of the unigram (Kolchyna et al., 2015).

Sentiment analysis is not performed in this study, but a new method is developed by utilizing the custom dictionary creation process in the Lexicon based method process.

Standard SVM uses all input features with the same importance when categorizing. However, it is a fact that the features of the inputs to be used in categorization are not identical. Two main methods are used to determine the differences in the features of the inputs. These are feature selection and feature weighting. The purpose of feature selection is to improve the generalization performance of the learning algorithm to eliminate irrelevant variables (Rakotomamonjy, 2003). Thus, a new subset is created using efficient variables for classification. On the other hand, the feature weighting attempts to estimate the relative importance of each feature and assign a weight corresponding to these estimates (Jin et al., 2005). Rather than making a binary decision on the relevance of feature, it is preferred to use a feature weight that has a continuous value, since it determines a more sensitive relevance (Wang, 2010).

In this study, the words to be used for offer categorization are assigned value with inspired by feature weighting method. 5-point Likert scale method is used to determine the scores to represent the values. It is impossible to give points to every word in the offer. Therefore, scoring is done by selecting certain words. There may be differences in CSR Texts according to the vehicle group. Therefore, it is necessary to examine the words on the basis of the vehicle group and dictionaries are created based on the selected words. Scores representing the difficulty levels are determined for the words in the dictionaries. In determining the score for words are conducted with a team who are experts in offer evaluation. After that, the words in the dictionaries are searched for each CSR Text line, respectively, and if detected, the difficulty level of the relevant word is added to the total score value of the CSR Text line. With this step, the total value of each word in the CSR Text content is calculated. There may be more than one special request within an offer. This means that an offer contains multiple lines of CSR text. In addition to scoring for CSR Text, the overall difficulty level of the offer needs to be determined. For this, average value of the difficulty levels of all CSR Texts is calculated. The steps of the applied method are shown in Figure 4.

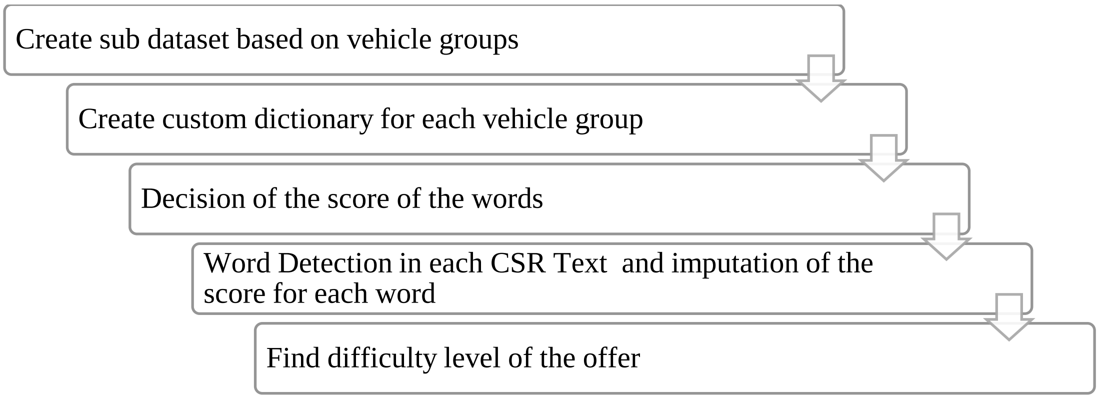

Create sub dataset based on vehicle groups
Create custom dictionary for each vehicle group
Decision of the score of the words
Word Detection in each CSR Text and imputation of the score for each word
Find difficulty level of the offer

Figure 4. The steps of the applied method

Create Sub Dataset Based on Vehicle Groups

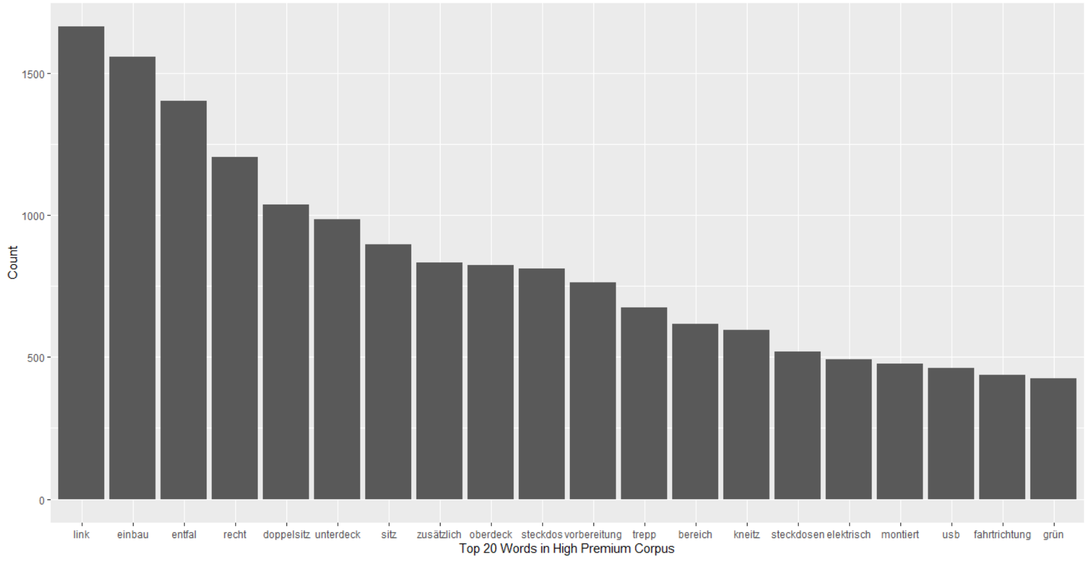
Separate data sets are created for the High Premium, Premium Coach and Coach vehicle groups. Analysis was not performed using the main data set because too much observation would affect the runtime of the detection loops and extend the analysis process. Also, Sub data sets are created for time saving. The Table 5 shows the number of variables and observations of each sub-data sets.

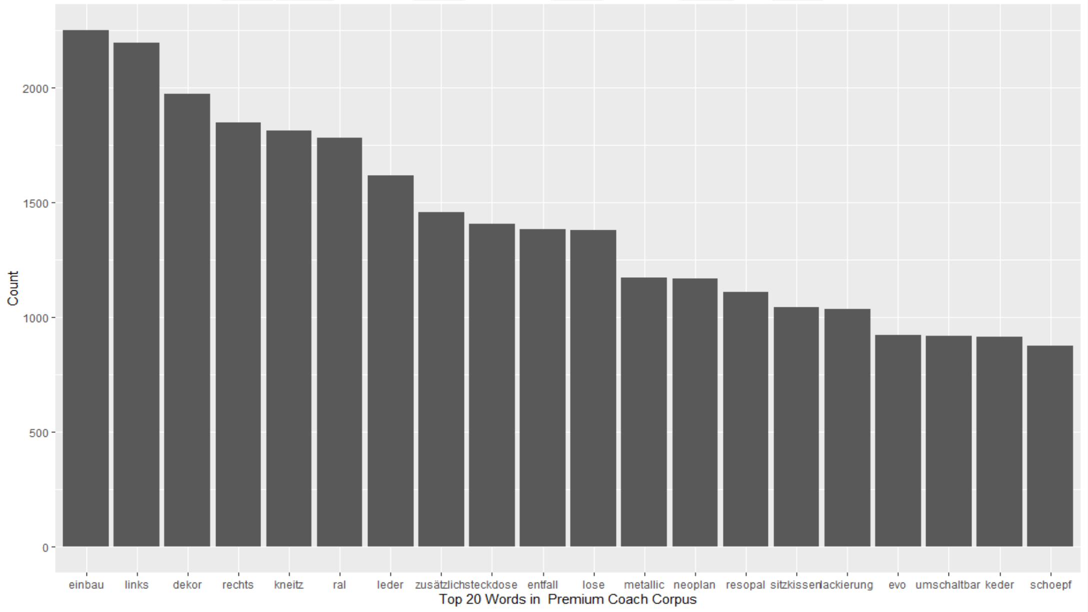
Vehicle Groups					
High Premium		**Premium Coach**		**Coach**	
Number of Observations	Number of Variables	Number of Observations	Number of Variables	Number of Observations	Number of Variables
9,464	31	34,151	31	144,519	31

Table 5. Content of sub-data sets

Create Custom Dictionary by Vehicle Groups

Since vehicle groups serve different markets, different specific request may be demanded in addition to common requests. Therefore, instead of using a single dictionary, different dictionaries are created for each vehicle group. When creating dictionaries, firstly, the number of times each word is used in the sub datasets is counted. Top 100 most used words by vehicle group are determined. These are in **Appendix A**. In addition, the first 20 most commonly used words for each vehicle group are shown in Figure 5. [10]





[10] Corpus text format was used to plot the Top20 words graph. Since the conversion of ”ä” and “ß” letters is done in data frame format, these are included in the graphic.

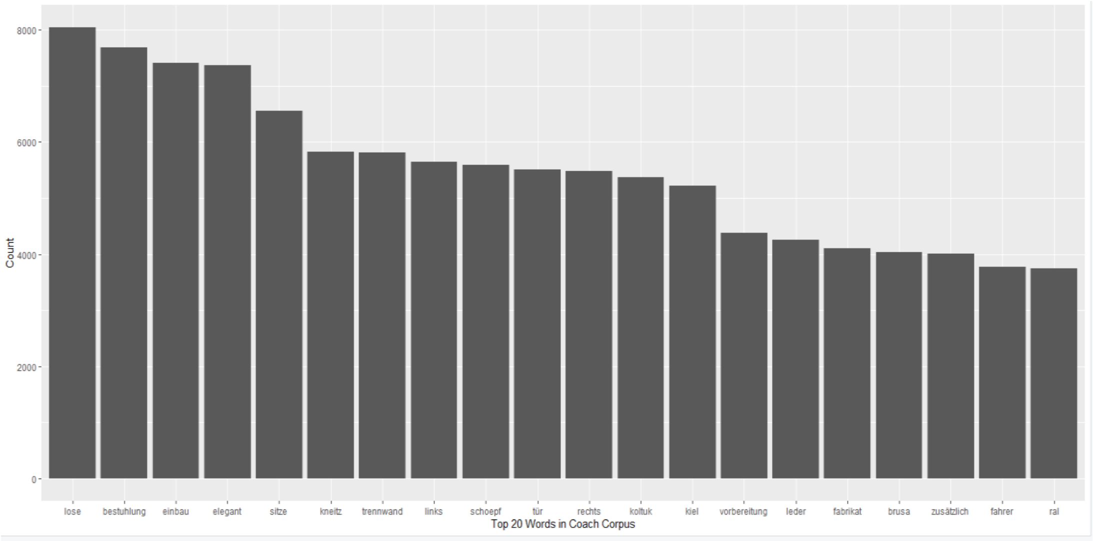


Figure 5. Top 20 most used words by vehicle groups

The words in the relevant lists are not used completely, because too many words appear in the lists, which can affect the difficulty level of the request if they are used in combination with a different word. For example, “einbau”. This word means " assemble". But it is not clear what the assembly is. Therefore, it is not appropriate to use this word directly. Dictionaries are created with the support of offer evaluation experts by taking advantage of Top100 Word Lists. The number of words in the newly created dictionaries and how many of the first 100 words are used in these dictionaries are shown in the Table 6. The dictionaries are created for each vehicle group are given in **Appendix B.**

	Vehicle Groups		
	High Premium	**Premium Coach**	**Coach**
Length of Dictionary	**153**	**163**	**183**
Percentage of the first 100 words included in the dictionary	**25**	**40**	**35**

Table 6. Content of Dictionaries

Decision of the words score

In order to categorize an offer, the difficulty level of the CSR texts must first be determined. Scores representing the difficulty level of words are added to the custom dictionaries created for each group of vehicles. 5-point Likert scale method was used to determine the difficulty levels. The scores and their meanings for the difficulty levels are shown in Table 7.

Score	Meaning
1	Very easy
2	Easy
3	Medium
4	Difficult
5	Very Difficult

Table 7. Likert Scoring for Difficulty Level

The difficulty levels of the words in the created dictionaries are shown in Appendix A. The distribution of the words used in the dictionaries according to the difficulty levels is shown in Figure 6.

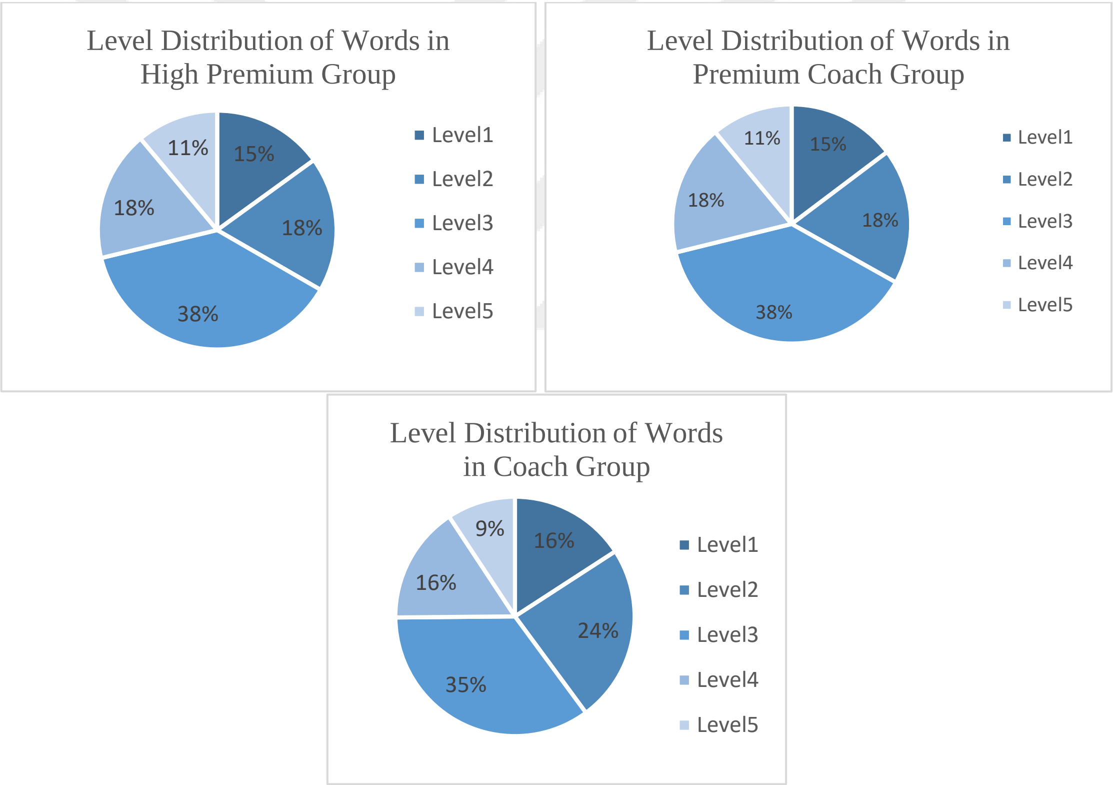
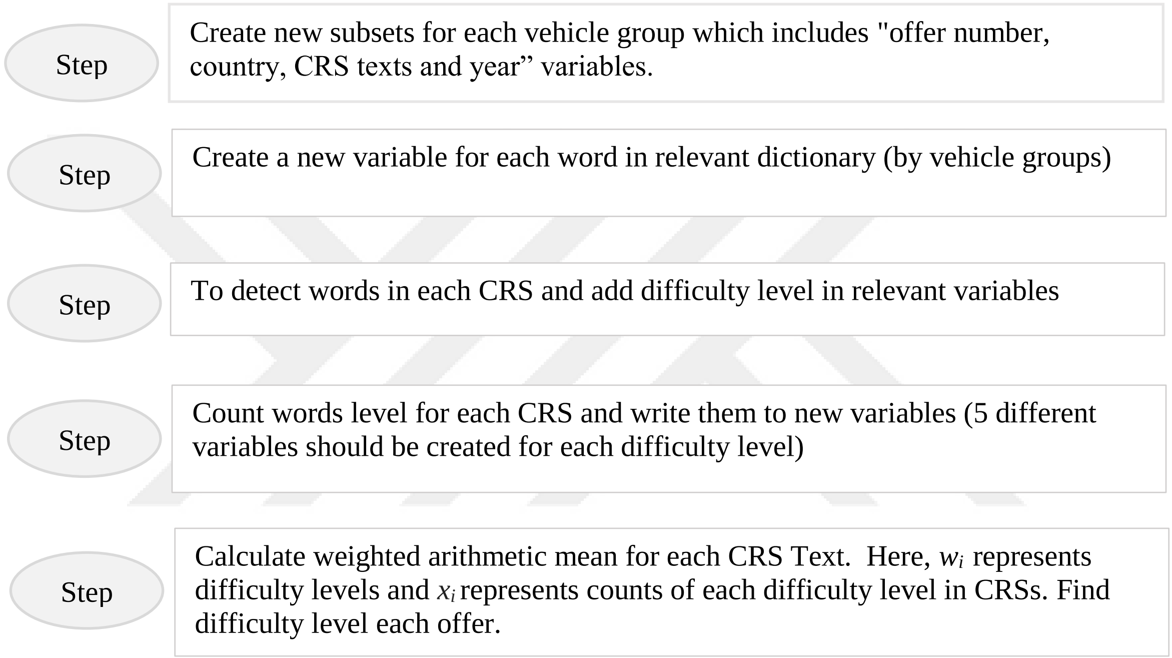


Figure 6. Level Distribution of Words by vehicle group

3.5. Categorization of the Offers

Three new data frames are created on the basis of the vehicle group by copying the offer number, country, CRS texts and year variables in the Offer Details dataset. A new variable is assigned to each word in the prepared dictionaries. Each offer contains multiple CRS texts. For each CRS text, the words in the dictionary are detected and the variable assigned to the word is added to the difficulty level score of the word. This allows searching for multiple words in a CRS text and calculating the total word difficulty value in the corresponding text. The weighted arithmetic mean is calculated for each CSR text using the difficulty level sum and counts of each difficulty level. Then, filtering is done based on offer number. The difficulty level of the offer is calculated by taking the average of the CSR texts whose weighted arithmetic mean is calculated for each offer. The steps applied to the categorization process are shown in Figure 7.



Step

Create new subsets for each vehicle group which includes "offer number, country, CRS texts and year" variables.

Step

Create a new variable for each word in relevant dictionary (by vehicle groups)

Step

To detect words in each CRS and add difficulty level in relevant variables

Step

Count words level for each CRS and write them to new variables (5 different variables should be created for each difficulty level)

Step

Calculate weighted arithmetic mean for each CRS Text. Here, w_i represents difficulty levels and x_i represents counts of each difficulty level in CRSs. Find difficulty level each offer.

Figure 7. The steps applied to the categorization process

Formula of the weighted arithmetic mean is that:

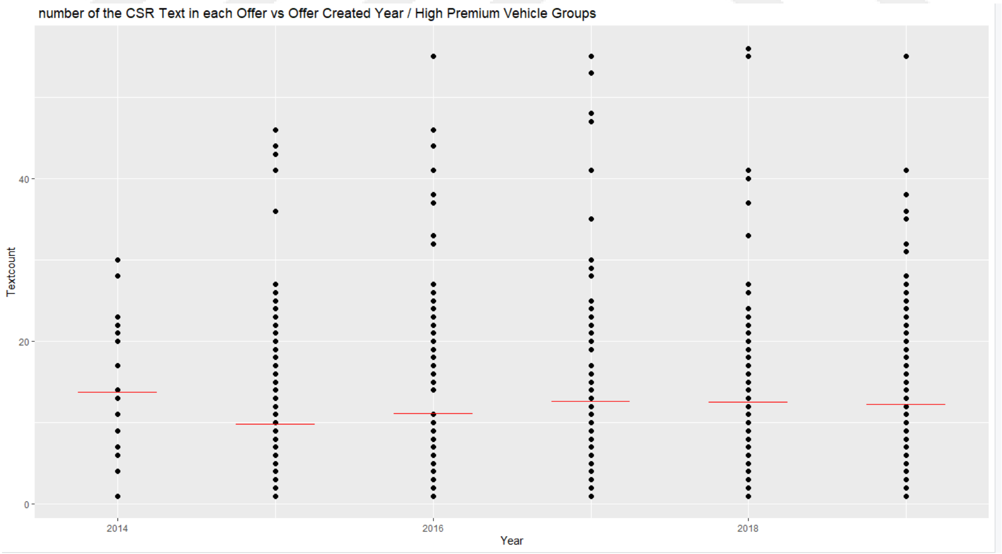
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i}$$

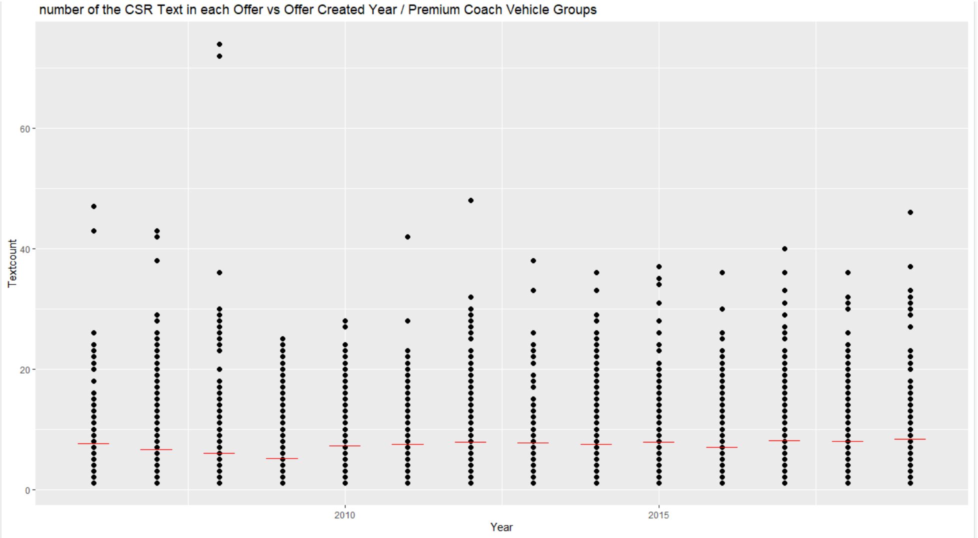
where $\bar{x}$ is weighted arithmetic mean, w_i (*non-negative weights)* represents difficulty levels and x_i *(data)* represents counts of each difficulty level in CRS text.

4. RESULTS

As a result of the study, meaningful information was obtained for sales management as well as categorization of customer-specific offers. The database that contains the data of the offers prepared since 2006 has never been examined before. Within the scope of this thesis, all offer contents are analyzed with text mining methods. As a result of this analysis, the distribution of the number of CSR Text by years, the variation of the total number of offers by years and total number of offers by country were obtained. In addition, categorization was performed by using country constraints in the vehicle group.

The first meaningful information obtained as a result of the analysis is the number of CSR text contained in the offers. Teams that evaluate offers by vehicle groups are different. The number of team members and workload distribution should be equal for each group of vehicles. The number of CSRs in the offer does not directly affect the workload because an offer with too many CSR text may fall into a very easy category. In this case, it can be said that an offer is created with easy content but a little longer evaluation period. The analysis results of the distribution of CSRs in three vehicle groups are shown in Figure 8.





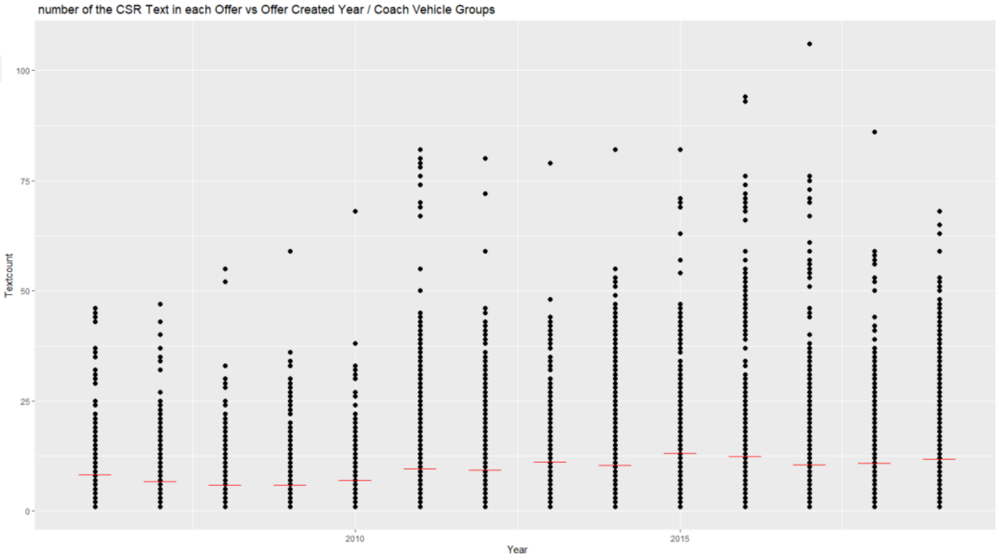


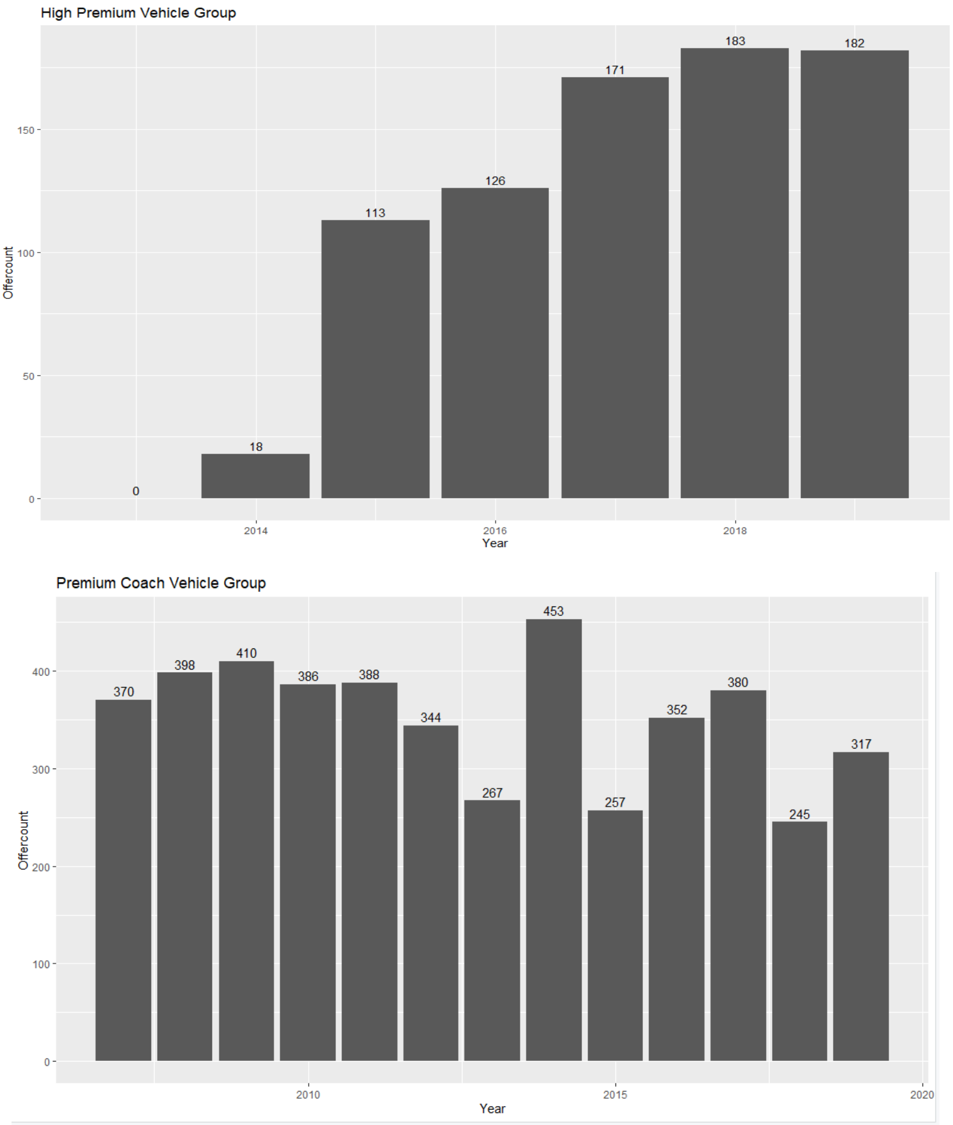
Figure 8. Distribution of CSR Text number by years for each vehicle group

According to the CSR text distribution graphs, the average CSR text of the offers in the High Premium vehicle group is higher than the other two vehicle groups. The first offer for the High Premium vehicle group is created in 2014. The highest CSR Text average belongs to 2014 while the lowest average in 2015. The offer with the highest CSR Texts is created in 2018.

The first offer for the Premium Coach group is created in 2006 and the offer with the highest CSR Text is created in 2008. The highest CSR Text average belongs to 2019 while the lowest average in 2009. The average CSR Text by years does not exceed 10.

The offer containing the highest CSR text is created in 2017 for the Coach vehicle group. The highest CSR Text average belongs to 2015. The deviation of the average CSR Text is not high after 2009. Finally, the average CSR text in all 3 vehicle groups is the lowest in 2019.

The total number of offers varies according to years. This is related to market needs and sales performance. The variation of the total number of offers for three vehicle groups by years is shown on Figure 9. In addition, the graphs showing the creating year of each offer are in Appendix C.



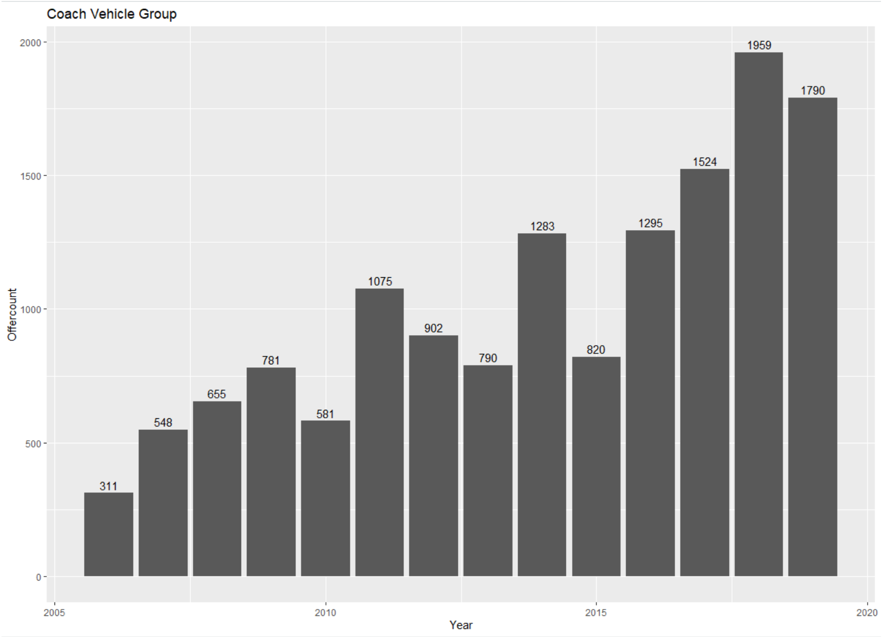
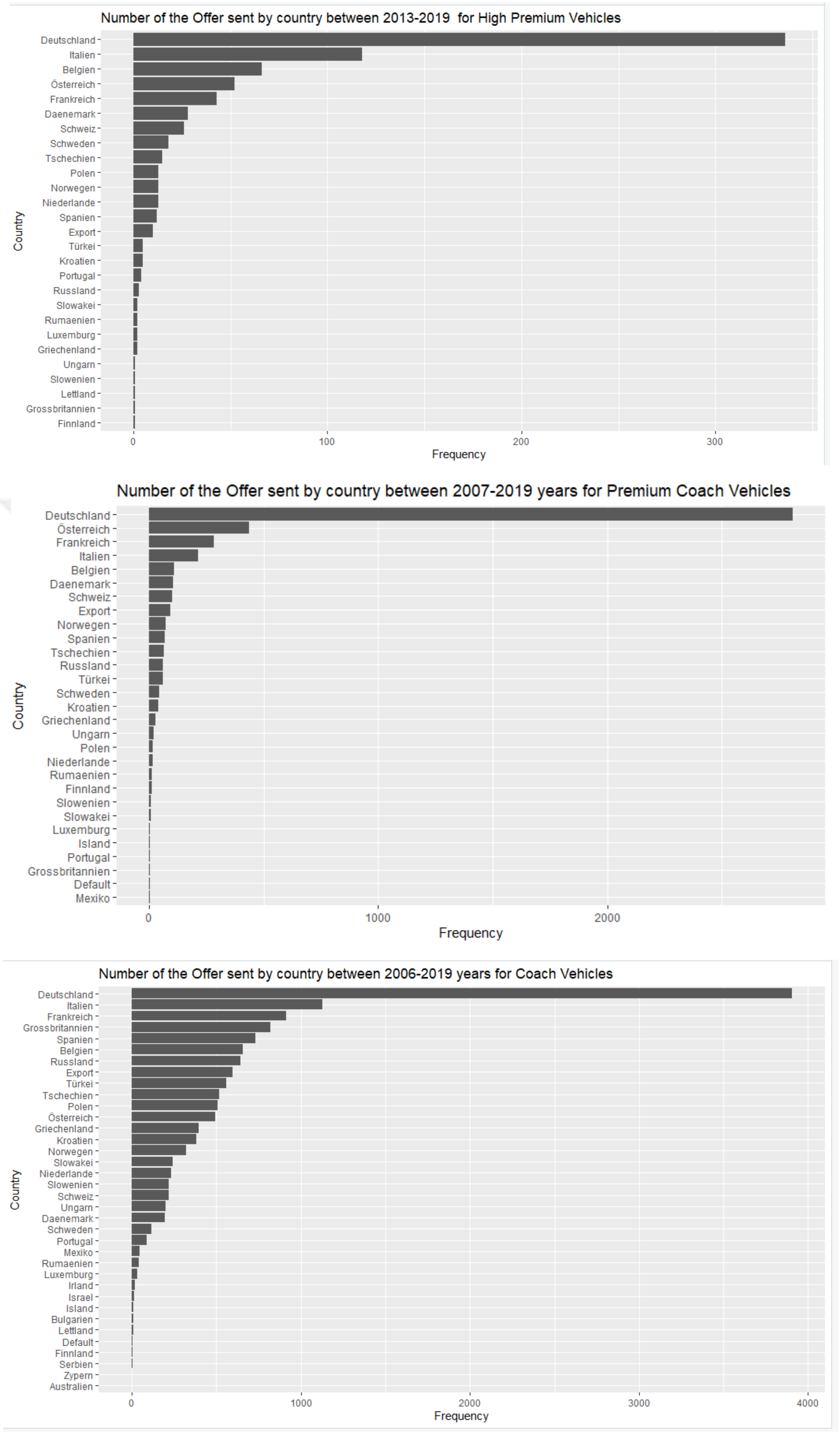


Figure 9. Number of the offers by years for each vehicle group

The first offer is created in 2014 for High Premium vehicle group. The total number of offers created in 2014 is too low compared to other years which is 18. After this year, the number of annual offers increases regularly. Since the data in 2019 is until August, the number of offers exceeds 2018. Furthermore, when comparing Figure 9 and Figure 8, the high number of offers does not mean that there is too much CSR Text. For example, Although the number of the minimum offer is created in 2014, the average number of CSR Text in this year is not the lowest.

The highest offer in the Premium Coach vehicle is created in 2014 while the minimum offer belongs to 2013. Although the number of offers increase regularly between 2015 and 2017, 36% decrease is observed in 2018 compared to the previous year.

The country of the offer is as important as content of the offers. The CSRs vary according to market requirements. Categorization analyze is firstly done based on countries. Before categorization results, how the total number of offers varies by countries is examined. The number of offers by countries for each vehicle group is shown in Figure 10.



Number of the Offer sent by country between 2013-2019 for High Premium Vehicles

Number of the Offer sent by country between 2007-2019 years for Premium Coach Vehicles

Number of the Offer sent by country between 2006-2019 years for Coach Vehicles

Figure 10. The total number of offers by country for each vehicle group

Since the country names in the data set remain in German, the country names shown in the Figure 10 are also in German. Additionally, countries named as Export are Dubai, Bulgaria, Gabon, Bosnia and Azerbaijan.

When the graphs are examined, the highest offer was created for Germany for the three vehicle groups. Italy ranks second for High Premium and Coach vehicles. The top 6 countries that make the highest offer for the High Premium and Premium Coach vehicles are the same. These countries are Belgium, Denmark, Germany, France, Italy and Austria. Countries for the Coach group are Belgium, Germany, France, England, Italy and Spain. Furthermore, Australia has once created an offer for the Coach vehicle but has never created an offer for other vehicle groups. Default country is assigned to the non-sales offer created by marketing department. This type of offer is not created for the High Premium vehicle group.

At this stage, offers are categorized based on countries. Instead of displaying all the countries on a graph, the graphs are created according to the top 6 countries with the most offers. Furthermore, graphs showing the categorization results for all countries are in Appendix D. The difficulty level of the offers varies between “1-5”. “0" difficulty level means categorization failed.

Categorization Results for High Premium Vehicle Group

The countries that request the most offers in the High Premium range are Belgium, Denmark, Germany, France, Italy and Austria. The offer categorization results for these countries are shown in Figure 11.

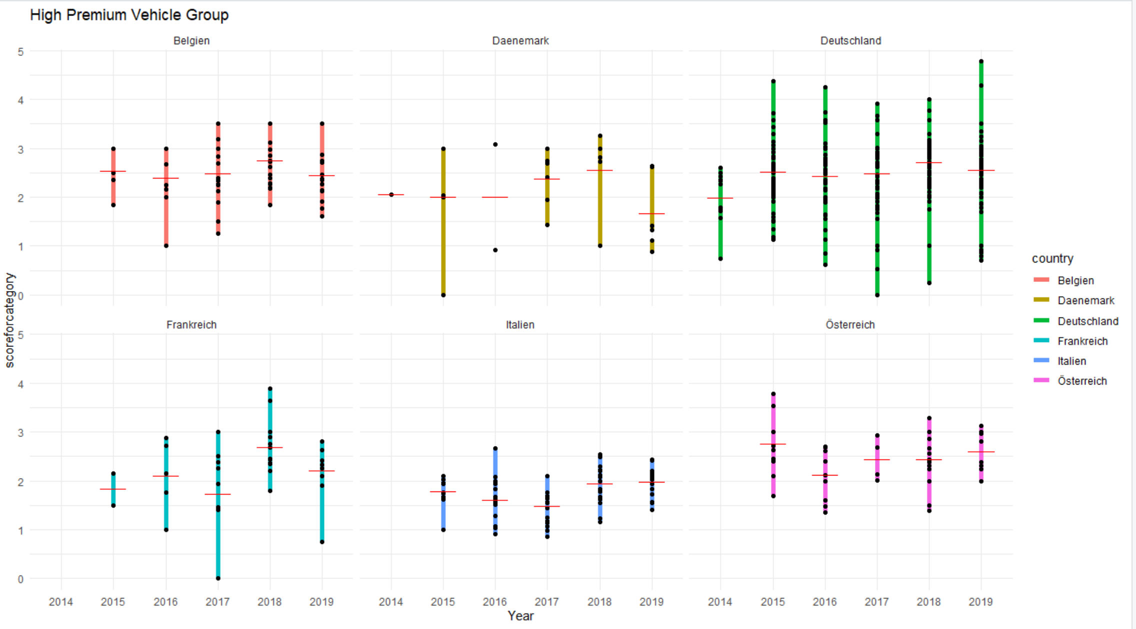


Figure 11. Offer categorization by country based in High Premium vehicles

The categorization results in this vehicle group belong to the years 2014-2019. The red line on the graph shows the average difficulty level. The average difficulty level of the offers created by Germany varies between 2-3. Within the top 6 countries, the overall average difficulty level of Italy is the lowest such that the average level of difficulty by years never exceed "2". On the other hand, Austria's offers have the highest levels of difficulty. According to the offer categorization result graph of all countries in Appendix D, only one offer has been created by Finland, England, Latvia, Russia, Slovenia and Hungary. Among these countries, Russian offers could not be categorized. The distribution of average difficulty levels by years for the High Premium vehicle group is shown in Figure 12.

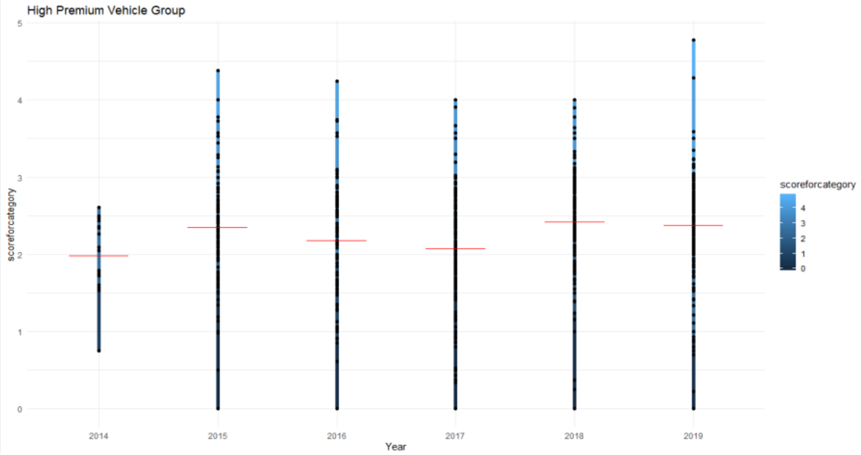


Figure 12. Offer categorization by year in High Premium vehicles

The average difficulty level for High Premium vehicles is the lowest in 2014. The average difficulty level is the highest in 2018. According to the graph, no offers are sent over the average difficulty level of 2.5 between 2014 and 2019 years.

Categorization Results for Premium Coach Vehicle Group

The countries that request the most offers for the Premium Coach vehicle group are the same as the High Premium vehicle group. The categorization results for these countries are shown in Figure 13.

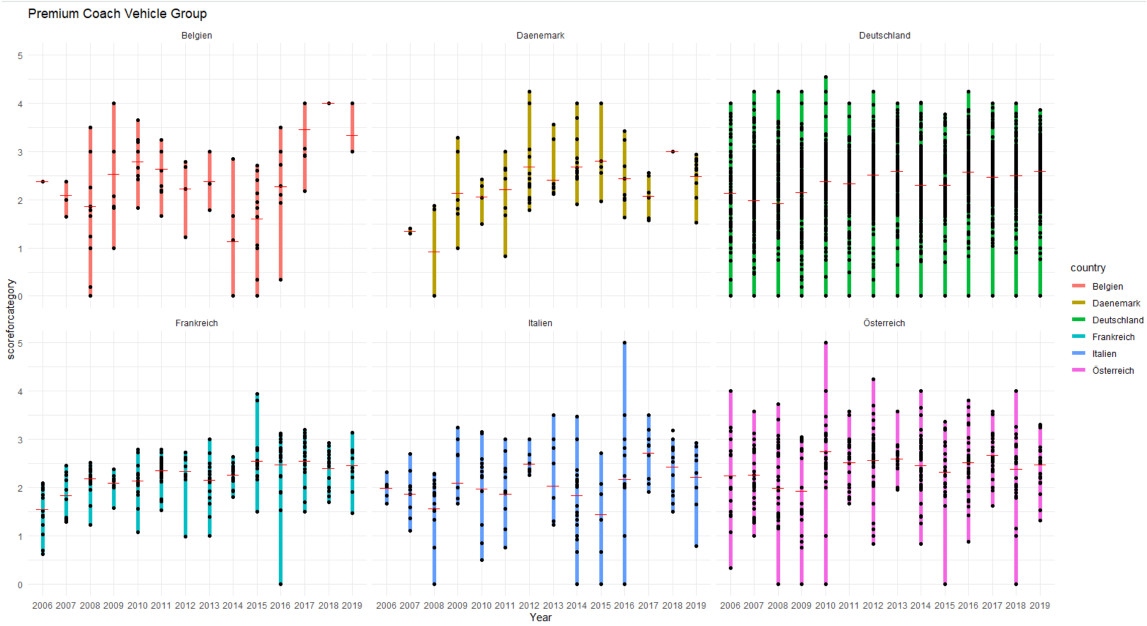


Figure 13. Offer categorization by country based in Premium Coach vehicles

The categorization results in this vehicle group belong to the years 2006-2019. The red line on the graph shows the average difficulty level. The average difficulty level of the offers created by Germany varies between 2-3. Within the top 6 countries, as in the High Premium vehicle, the country that creates the easiest offer is Italy and the most difficult offers is requested by Austria. Only one offer was created by both Belgium and Denmark in 2018, and these are the most difficult offers created by the relevant countries between 2006 and 2019.

According to the offer categorization result graph of all countries in Appendix D, only one offer has been created by Portugal in 2011. Two dummy offers were requested in 2009 and 2014 for this vehicle group. Several offers have been created for the Mexican and UK markets, but these markets are not active. Thus, only two offers were created by UK in 2009 and no other offers have been requested for this market in the last 10 years. Similarly, three offers were created for Mexico and the last one belongs to 2014. Also, there is no activity in this market for the past 5 years.

Besides the country-based categorization results, average categorization scores by years for the Premium Coach vehicle group are shown in Figure 14.

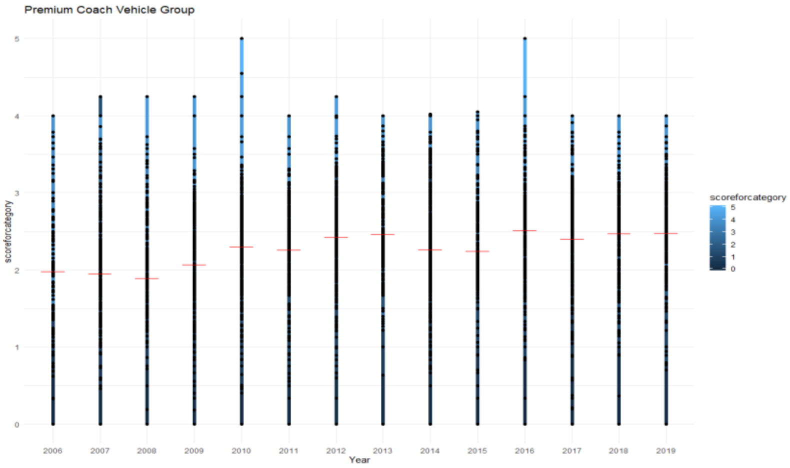


Figure 14.Offer categorization by year in Premium Coach vehicles

The average difficulty level of the offers created in 2016 is the highest compared to other years. In contrast, the lowest average is in 2008. There is no steady increase or decrease trend over the years. An offer was created at the most difficult level in 2010 and 2016.

Categorization Results for Coach Vehicle Group

The countries that create the most offers for the Coach vehicle group are Belgium, Germany, France, the UK, Italy and Spain. The categorization results of these countries' offers are shown in Figure 15.

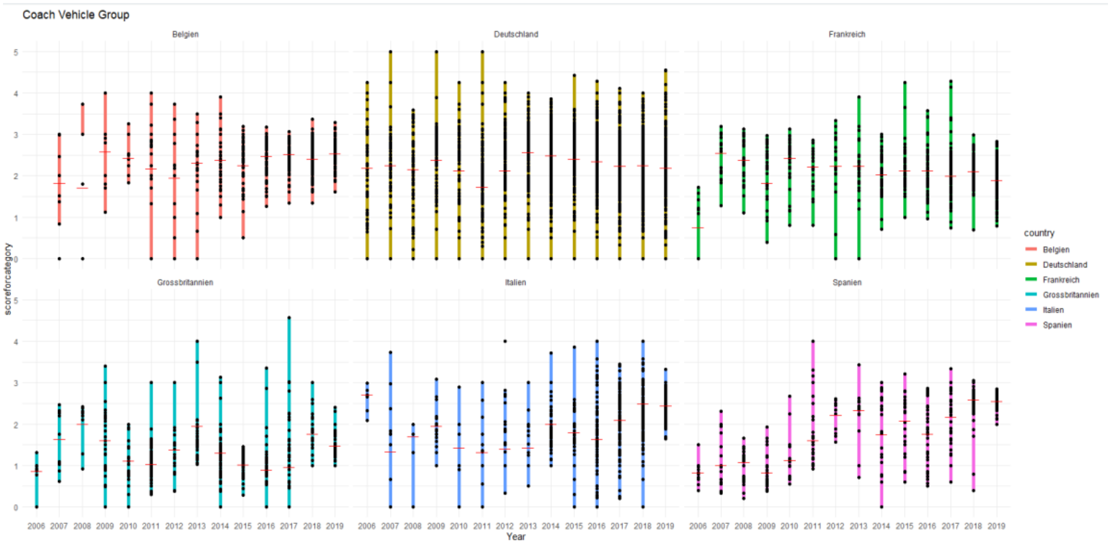


Figure 15. Offer categorization by country based in Coach vehicle group

The categorization results in this vehicle group belong to the years 2006-2019. The red line on the graph shows the average difficulty level. The average difficulty level of the offers created by Germany varies between 2-3. The overall average of the offers created by Italy and the UK is the lowest compared to the other top 6 countries. This means that the offer created by them are easy.

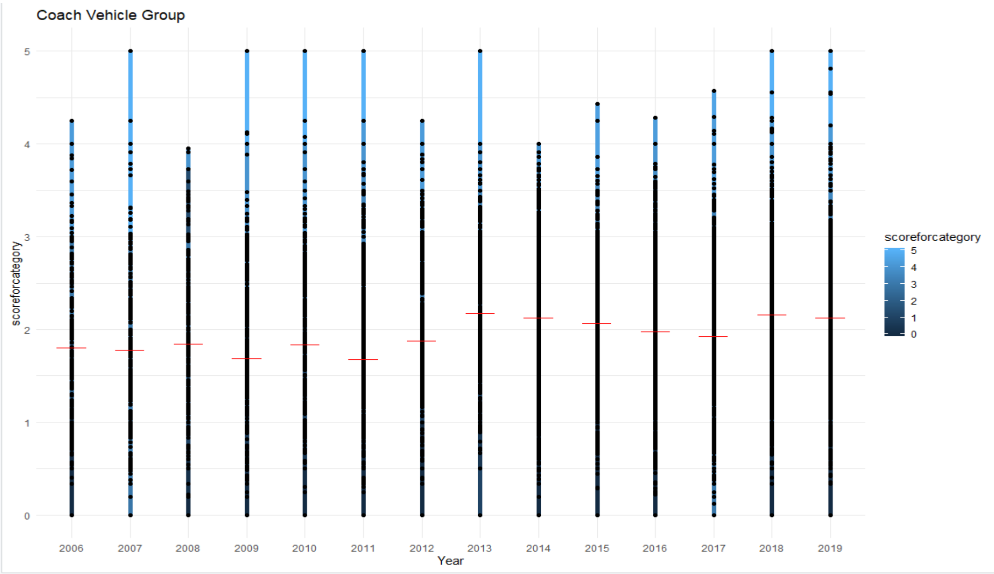


Figure 16. Offer categorization by year in Coach vehicles

According to the overall difficulty level average between 2006 and 2019, the difficulty level of the Coach vehicle is the easiest among the other two vehicle groups.

5. CONCLUSION & FUTURE WORK

In this study, the effect of the offer contents on sales management processes is examined by combining the offer documents presented to the customers in a single data set during the vehicle sales process. Country and year-based analyzes were conducted to analyze how market demands vary. Moreover, offers were categorized according to difficulty levels with a new method inspired by machine-learning and Lexicon sentiment analysis. However, there were offers that could not be categorized for all vehicle groups. As a result of this study, categorization could not be made in 18 of 793 offers in High Premium vehicle group, 144 of 4731 offers in Premium Coach vehicle group and 403 of 14314 offers in Coach vehicle group. The main reason for this is that no word detection in the CSR text within the offer. If the specific dictionaries used in categorization process is developed, 100 % categorization is provided. Since this study is the result of combining and developing two different literature studies, we think that this is the first study in this field. Additionally, it is seen that the content of the special requests affects the difficulty level more than the number of requests included in the offer.

The results of the categorization have been shared with the sales managers in the relevant countries and it has been reported how the market has changed over the years and these changes should be taken into consideration for future investments.

Another outcome of the study is that there are countries that are constantly creating difficult and simple offers on the basis of vehicle groups which are Austria and Italy. When allocating tasks in the offer evaluation department, it is necessary to look at the country from which more than one offer is sent to the responsible person during the process.

Finally, the future step in this study will be to estimate the difficulty level distribution of offer to be prepared in the coming years based on the categorized data of previous years' offers. In addition to the offer categorization, order categorization will be done. Order categorization will ensure that the difficulty of a vehicle to start production is known in advance and precautions for the possible challenges/problems encountered during the production process.

REFERENCES

Bahng, Y., & Kincade, D. H. (2012). The relationship between temperature and sales: Sales data analysis of a retailer of branded women's business wear. *International Journal of Retail & Distribution Management*, *40*(6), 410-426.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1), 39-71.

Borowiecki, K. J. (2017). How are you, my dearest mozart? well-being and creativity of three famous composers based on their letters. *Review of Economics and Statistics*, *99*(4), 591-605.

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. *International Journal of Computer Applications*, *28*(2), 37-40.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29(2-3), 103-130.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Gorodnichenko, Y., Pham, T., & Talavera, O. (2018). *Social media, sentiment and public opinions: Evidence from# Brexit and# USElection* (No. w24631). National Bureau of Economic Research.

Gottfried, J. and Shearer, E., 2016. News Use Across Social Media Platforms 2016. [Online] Pew Research Center's Journalism Project. (Available at: http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/ [Accessed 29 April 2018].)

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012, January). Classification of customer reviews based on sentiment analysis. In ENTER (pp. 460-470).

Harding, J. A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: a review.

Hu, Y., Farnham, S., & Talamadupula, K. (2015, April). Predicting user engagement on twitter with real-world events. In *Ninth International AAAI Conference on Web and Social Media*.

Hui, S. C., and Jha, G., 2000, "Data Mining for Customer Service Support," Inf. Manage., 38, pp. 1–13.

Jin, B., & Zhang, Y. Q. (2005, June). Support vector machines with evolutionary feature weights optimization for biomedical data classification. In NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society (pp. 177-180). IEEE.

Joachims, T. (1998). Making large-scale SVM learning practical (No. 1998, 28). Technical report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Kannan, S., & Gurusamy, V. (2014, October). Preprocessing techniques for text mining. In Conference Paper. India.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004, December). Multinomial naive bayes for text categorization revisited. In *Australasian Joint Conference on Artificial Intelligence* (pp. 488-499). Springer, Berlin, Heidelberg.

Lewis, D. D. (1998, April). Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

Ling, C. X., & Li, C. (1998, August). Data mining for direct marketing: Problems and solutions. In *Kdd* (Vol. 98, pp. 73-79).

Morita, T., Sato, Y., Ayukawa, E., and Maeda, A., 2000, "Customer Relationship Management Through Data Mining," Informs-Korms 2000, Seoul, pp. 1956–1963.

Murakami, D., Peters, G. W., Yamagata, Y., & Matsui, T. (2016). Participatory sensing data tweets for micro-urban real-time resiliency monitoring and risk management. IEEE Access, 4, 347-372.

Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering* (Vol. 1, No. 1, pp. 61-67).

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of machine learning research, 3*(Mar), 1357-1370.

Stempel, G.H., Hargrove, T. and Bernt, J.P., 2000. Relation of growth of use of the Internet to changes in media use from 1995 to 1999. *Journalism & Mass Communication Quarterly, 77*(1), pp.71-79.

Valle, M. A., Varas, S., & Ruz, G. A. (2012). Job performance prediction in a call center using a naive Bayes classifier. *Expert Systems with Applications*, *39*(11), 9939-9945.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Wang, T. (2010, May). i. In *2010 International Conference on Intelligent Computation Technology and Automation* (Vol. 2, pp. 518-521). IEEE.

Wlezien, C., Soroka, S., & Stecula, D. (2017). A Cross-National Analysis of the Causes and Consequences of Economic News. *Social Science Quarterly, 98(3),* 733-744.

Xing, H. J., Ha, M. H., Hu, B. G., & Tian, D. Z. (2009). Linear feature-weighted support vector machine. *Fuzzy Information and Engineering*, *1*(3), 289-305.

APPENDICES

Appendix A

Counted Top 100 Words for High Premium Vehicle Group

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
link	1663	230	382	ral	246	hoch	187
einbau	1556	pkt	374	sitzreih	245	fahrzeug	185
entfal	1402	sicherheitsgurt	374	einbauort	230	code	184
recht	1204	sitzen	364	sitzrückenlehn	228	spannungsversorgung	181
doppelsitz	1037	lose	352	kofferraum	225	fabr	178
unterdeck	985	tisch	340	vorhangstoff	222	gesamt	177
sitz	896	anschluss	334	schalter	217	fms	174
zusätzlich	833	monitor	333	analog	216	flyerhalt	171
oberdeck	825	montieren	326	tür	215	zugänglich	170
steckdos	812	vorn	314	110	208	abmessungen	170
vorbereitung	762	wurstkoch	309	230v	206	umschaltbar	169
trepp	674	resop	299	nachträglichen	206	bordküch	169
bereich	617	werksseitig	289	ausführung	202	leicht	169
kneitz	594	kabel	287	fenstersitz	202	steuergerät	168
steckdosen	520	toilett	274	fensterplatz	201	220	167
elektrisch	492	ferngrau	273	seitenwand	199	sitzkissen	165
montiert	477	isero	273	stecker	196	rollstuhlplatz	159
usb	460	66656674	270	manda	194	versorgungskabel	159
fahrtrichtung	437	bmc	265	beleuchtung	191	anhang	158
grün	426	kaffeemaschin	264	gepäckablag	191	vorzusehen	156
fahrerplatz	417	tief	260	sieh	190	frischwassertank	155
heckbank	409	reih	259	abdeckung	190	staufach	152
möglich	402	erst	253	gerät	189	valencia	151
schoepf	395	fahrgastsitz	251	schnittstell	189	glitter	151
dekor	382	fabrikat	246	breit	187	seit	150

Counted Top 100 Words for Premium Coach Vehicle Group

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
einbau	2250	manda	789	garnfarbe	593	fahrgastsitze	476
links	2195	fahrerplatz	783	unten	591	9016	472
dekor	1971	glitter	777	neo	589	inkl	470
rechts	1848	bezug	761	vegal	587	crystal	468
kneitz	1812	tür	746	montiert	582	kühlschrank	463
ral	1779	silver	732	ausführung	575	fahrgastsitzseitenböden	463
leder	1615	fahrer	727	oben	566	silber	461
zusätzlich	1458	schwarz	713	met	566	bereich	453
steckdose	1404	esw	703	sitzreihe	560	ksw	450
entfall	1382	ingleston	688	steckdosen	554	elektrische	445
lose	1380	falte	677	siehe	554	toilettenbox	443
metallic	1170	beilegen	670	115160	550	slate	443
neoplan	1168	doppelsitz	670	seitenwangen	536	brett	441
resopal	1107	joker	669	perforiert	536	armaturenbrett	441
sitzkissen	1044	heckbank	668	muirhead	532	fahrgastsitzbezugsstoff	439
lackierung	1035	stoff	662	eingestickt	529	bmc	432
evo	922	kofferraum	654	230v	526	seitenwandmaterial	430
umschaltbar	917	analog	651	fahrzeug	514	vorhangstoff	429
keder	914	eswend	633	fabrikat	512	seite	423
schoepf	876	logo	625	comfort	510	polsterart	420
vorbereitung	875	bestuhlung	621	gardinenhalteband	510	jedoch	417
schalter	872	küche	609	farbe	504	nachträglichen	412
toilette	866	zusätzliche	606	9010	492	anhang	410
toilettentür	850	seitenwand	604	verbauen	485	hinten	406
sitze	812	montieren	600	vorne	476	möglich	405

Counted Top 100 Words for Coach Vehicle Group

Word	Frequency	Word	Frequency	Words	Frequency	Words	Frequency
lose	8047	mitlieferung	3411	code	2283	kswend	1811
bestuhlung	7684	laeufer	3381	aschenbecher	2280	2200	1810
einbau	7411	fahrerplatz	3364	ausführung	2235	eingebaut	1743
elegant	7366	200	3202	kabel	2206	fahrerliege	1730
sitze	6564	azo	3056	montiert	2205	haltegriff	1654
kneitz	5833	entfall	3002	beim	2119	slim	1653
trennwand	5814	farbe	2939	doppelsitz	2108	lackierung	1634
links	5642	muirhead	2924	3210	2103	vip	1633
schoepf	5594	metallic	2922	bug	2086	evo	1599
tür	5515	schalter	2894	monte	2073	sitzreihen	1585
rechts	5479	valencia	2825	sitzen	2073	hostes	1545
koltuk	5374	fuer	2815	heckbank	2060	manda	1543
kiel	5222	toilette	2793	schwarz	1987	alkoholmessgeräts	1526
vorbereitung	4383	seat	2750	tisch	1930	typ	1511
leder	4250	monitor	2709	ksw	1911	vorne	1498
fabrikat	4100	esw	2704	yolcu	1908	fahrgastsitze	1471
brusa	4033	without	2654	ingleston	1903	innen	1468
zusätzlich	4006	kofferraum	2550	kumasi	1878	steuergeraet	1463
fahrer	3767	avance	2543	2400	1868	edilecektir	1451
ral	3747	eswend	2496	door	1862	seitenscheiben	1437
steckdose	3667	seats	2487	bsaeule	1854	kopflaetzchen	1414
sirius	3626	oben	2473	seite	1849	feuerlöscher	1414
elektrische	3485	olmayacaktir	2443	sofor	1841	sitzkissen	1414
fahrzeug	3459	anthrazit	2396	keder	1830	ayirma	1412
olacaktir	3429	pos	2286	noppenbelag	1814	beilegen	1396

Appendix B

Dictionaries

Dictionary for High Premium Vehicle Group

Word	Difficulty Level	Word	Difficulty Level
coc	1	mikrowelle	3
kofferraumzuladung	1	kühlschrank	3
zielland	1	resopal	3
sprache	1	resop	3
dokumentation	1	lüfterquirle	1
bmc	5	lufterquirle	1
kmh	1	fza	5
hypoidantriebsachse	1	fahrtzielanzeig	5
haltestellenbremse	1	steuergeraet	5
stahlfelge	1	icomera	5
alufelge	1	mobitec	5
radzierblenden	1	hanover	5
radzierblendenlackierung	1	lawo	5
lackierung	2	adblue	4
lackiert	2	batterien	4
schloss	2	starterbatterien	4
schlüss	2	spannungswandler	4
treppe 1 rechts	3	phönix	4
scheibe	3	phonix	4
ral	2	phoenix	4
notspiegel	1	abc	4
kamera	3	led	5
poliger	2	diebstahl	4
polig	2	warmanlage	4
seitenwand	3	tachograph	4
seitenwandverkleidung	3	fms	4
verkleidung	3	wlan	4
mirax	3	wlan router	4
schoepf	3	dab	4
schöpf	3	smartlink	4
valencia	3	reisebegleitmikrofon	1
manda	3	mikrofon	1
gepaeckablagenunterseit	3	steckdose	3
muirhead	3	steckdosen	3
melito	3	230v	3
kneitz	3	usb	3
fahrerplatzbeleuchtung	3	doppelusb	3

beleuchtung	3	kombi steckdose	3
blau	2	tomtom	3
grün	2	lose	3
bodenbelag	3	attentionguard	1
gerflor	3	efficientcruise	1
tarabus	3	tresor	3
gaya wood	3	sitzanordnung	2
laeufer	2	bestuhlung	2
azo 200	1	punktgurten	2
azo	1	flixbus	2
teppich	1	markenemblem	3
abfallbehaelter	2	stoff	3
vorhangstoff	2	leder	3
gardienen	2	klapptisch	5
gardiene	2	clubtisch	5
gardienenhalteband	2	dos a dos	5
gardienenhaelteband	2	vis a vis	5
gardinenhalteband	2	zeitungsnetze	3
halteband	2	armlehne	3
haelteband	2	sitze	3
anise	2	sitz	3
borgstena	2	sitzkissen	1
staubox	4	doppelsitz	3
wc	4	sitzanordnung	3
cc	4	fahrersitz	3
toilette	4	sitzreihen	3
toillette	4	sitzreihe	3
haendetrockn	3	slim	3
umschaltbar	3	fussstützen	3
frischwassertank	2	rollstuhlplatz	5
faekalientankvolumen	3	sliding	5
dekor	4	slider	5
kühlschrank	4	wechselpodest	5
stehküche	4	fahrgastsitze	3
bordküche	4	sitzrückenlehne	3
kaffeemaschine	4	fahrgastsitzseitenböden	3
küche	4	fahrgastsitzrückseite	3
faekalientank	3	lederkopflaetzchen	3
wassertank	2	staufach	2
wurstkocher	3		

Dictionary for Premium Coach Vehicle Group

Word	Difficulty Level	Word	Difficulty Level
coc	1	lufterquirle	1
kofferraumzuladung	1	fza	5
zielland	1	fahrtzielanzeig	5
sprache	1	steuergeraet	5
dokumentation	1	icomera	5
bmc	5	mobitec	5
kmh	1	hanover	5
hypoidantriebsachse	1	lawo	5
haltestellenbremse	1	adblue	4
stahlfelge	1	batterien	4
alufelge	1	starterbatterien	4
radzierblenden	1	spannungswandler	4
radzierblendenlackierung	1	phönix	4
lackierung	2	phonix	4
lackiert	2	phoenix	4
schloss	2	abc	4
schlüss	2	led	5
treppe 1 rechts	3	diebstahl	4
scheibe	3	warmanlage	4
ral	2	tachograph	4
notspiegel	1	fms	4
kamera	3	wlan	4
poliger	2	wlan router	4
polig	2	dab	4
seitenwand	3	smartlink	4
seitenwandmaterial	3	reisebegleitmikrofon	1
verkleidung	3	mikrofon	1
mirax	3	steckdose	3
schoepf	3	steckdosen	3
schöpf	3	230v	3
valencia	3	usb	3
manda	3	doppelusb	3
gepaeckablagenunterseit	3	kombi steckdose	3
muirhead	3	tomtom	3
melito	3	lose	3
kneitz	3	attentionguard	1
fahrerplatzbeleuchtung	3	efficientcruise	1
beleuchtung	3	tresor	3
blau	2	sitzanordnung	2
grün	2	bestuhlung	2

bodenbelag	3	punktgurten	2
gerflor	3	markenemblem	3
tarabus	3	stoff	3
gaya wood	3	leder	3
laeufer	2	klapptisch	5
azo 200	1	clubtisch	5
azo	1	dos a dos	5
teppich	1	vis a vis	5
abfallbehaelter	2	zeitungsnetze	3
vorhangstoff	2	armlehne	3
gardienen	2	sitze	3
gardiene	2	sitz	3
gardienenhalteband	2	sitzkissen	1
gardienenhaelteband	2	doppelsitz	3
gardinenhalteband	2	sitzanordnung	3
halteband	2	fahrersitz	3
haelteband	2	sitzreihen	3
anise	2	sitzreihe	3
borgstena	2	slim	3
staubox	4	fussstützen	3
wc	4	rollstuhlplatz	5
cc	4	sliding	5
toilette	4	slider	5
toillette	4	wechselpodest	5
haendetrockn	3	fahrgastsitze	3
umschaltbar	3	sitzrückenlehne	3
frischwassertank	2	fahrgastsitzseitenböden	3
faekalientankvolumen	3	fahrgastsitzrückseite	3
dekor	4	lederkopflaetzchen	3
kühlschrank	4	staufach	2
stehküche	4	toilettentür	4
bordküche	4	glitter	3
kaffeemaschine	4	crystal	3
küche	4	toilettenbox	4
faekalientank	3	sicherheitsgurt	2
wassertank	2	flyerhalt	2
wurstkocher	3	9010	1
mikrowelle	3	ingleston	3
kühlschrank	3	garnfarbe	3
resopal	3	joker	2
resop	3	eingestickt	5
lüfterquirle	1		

Dictionary for Coach Vehicle Group

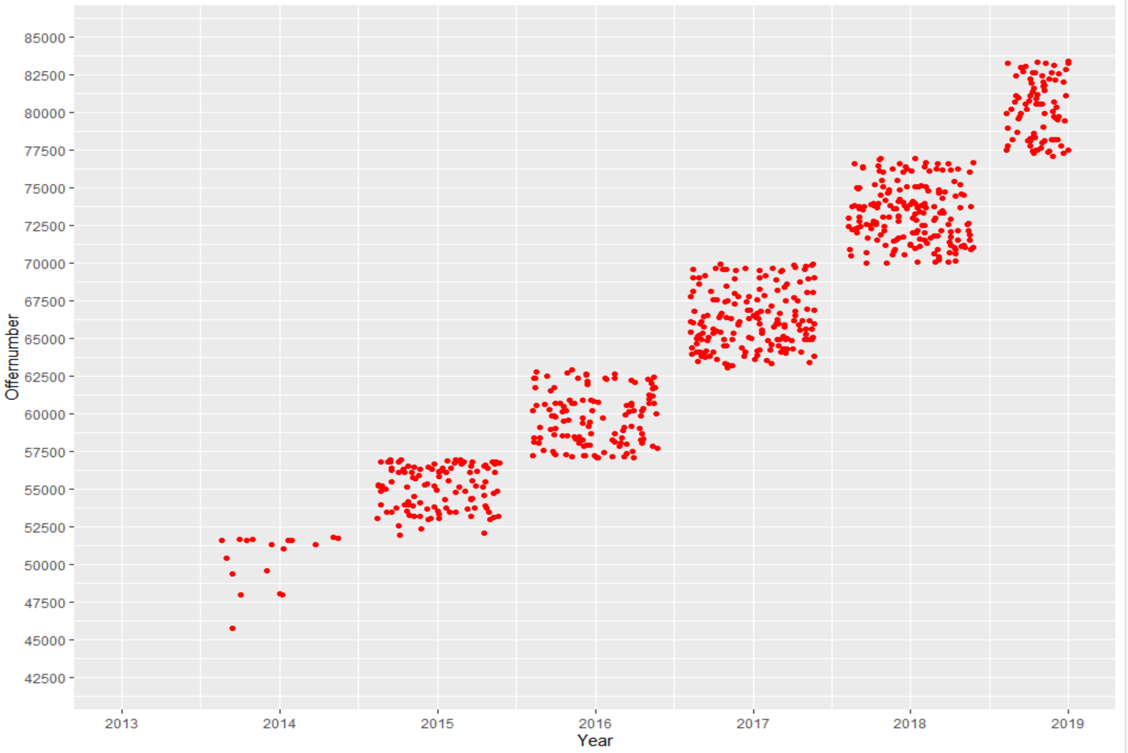
Word	Difficulty Level	Word	Difficulty Level
coc	1	flixbus	2
kmh	2	markenemblem	2
kofferraumzuladung	1	epengle	3
zielland	1	stoff	3
sprache	1	leder	3
dokumentation	1	klapptisch	5
Bmc	5	clubtisch	5
hypoidantriebsachse	1	dos a dos	4
haltestellenbremse	1	vis a vis	4
stahlfelge	1	zeitungsnetze	3
alufelge	1	armlehne	3
radzierblenden	1	sitze	3
radzierblendenlackierung	1	elegant	3
lackierung	2	doppelsitz	3
schloss	2	fussstützen	3
schlüss	2	rollstuhlplatz	5
scheibe	3	sliding	5
Ral	2	slider	5
notspiegel	1	wechselpodest	5
kamera	3	mittelarmlehnen	1
poliger	2	rollstuhlplaetze	5
seitenwand	3	schliessung	3
seitenwandverkleidung	3	bugklappe	3
verkleidung	3	gl6wbx	3
mirax	3	perforiert	1
schoepf	3	rangierlicht	2
schöpf	3	ladewechselrichter	5
manda	3	ısolierungen	3
melito	3	eraglonass	1
fahrerplatzbeleuchtung	3	sticker	1
beleuchtung	3	winter package	3
Blau	2	extinguishing	2
Grün	2	vga	2
bodenbelag	3	sitzkissen	1
gerflor	3	aschenbecher	2
tarabus	3	schulbusssymbol	2
gaya wood	2	fenwick	2
laeufer	2	nothammer	1
Azo	1	kneitz	3
teppich	1	koltuk	3
abfallbehaelter	2	toilette	4
vorhang	2	steckdosen	3

gardienen	2	vorhangstoff	2
gardiene	2	faekalientank	3
gardienenhalteband	2	kaffeemaschine	4
gardienenhaelteband	2	gepaeckablagenunterseit	3
halteband	2	steuergeraet	5
haelteband	2	abfallbehaelter	2
Anise	2	wassertank	2
borgstena	2	umschaltbar	3
staubox	4	bordküche	4
Wc	4	fahrtzielanzeig	5
Cc	4	tachograph	4
toilet	3	bestuhlung	3
toillette	4	lackiert	2
dekor	4	gardinenhalteband	2
kaffeemaschine	3	kühlschrank	4
küche	4	gaya wood	3
resopal	3	lüfterquirle	1
lüfterquirle	2	spannungswandler	4
lufterquirle	1	polig	2
Fza	5	7polsteckdose	2
mobitec	5	valencia	3
hanover	5	muirhead	3
Lawo	5	azo 200	1
adblue	4	haendetrockn	3
batterien	4	frischwassertank	2
starterbatterien	4	faekalientankvolumen	3
spannungswandler	4	icomera	5
phönix	4	wlan router	4
phonix	4	doppelusb	3
phoenix	4	kombisteckdose	3
Abc	4	tomtom	3
Led	5	mikrofonsteckdose	2
diebstahl	4	13polige	2
warmanlage	4	feuerlöscher	2
tachograph	2	alkoholmessgeraets	2
Fms	4	noppenbelag	1
Wlan	4	keder	1
Dab	4	sirius	3
smartlink	4	kiel	3
reisebegleitmikrofon	1	brusa	3
mikrofon	1	avance	3
steckdose	3	heckbank	3
230v	3	sitzen	3
Usb	3	sitzreihen	3
Lose	3	evo	3
attentionguard	1	seat	3

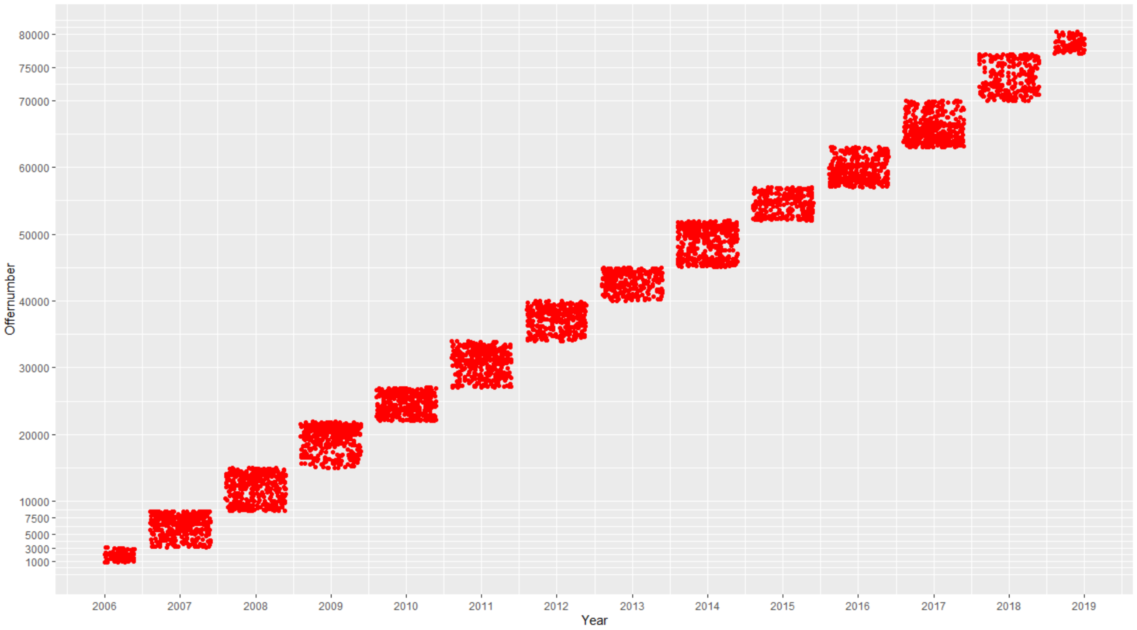
efficientcruise	1	seats	3
tresor	3	monitor	2
sitzanordnung	3	kopflaetzchen	3
punktgurten	2		

Appendix C

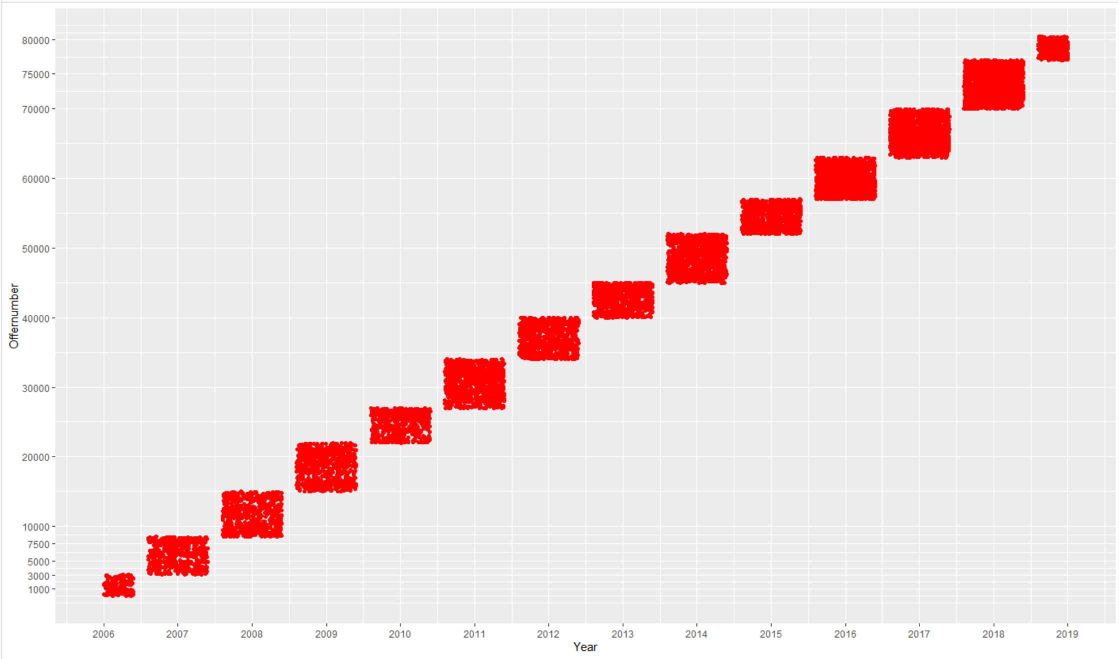
Distribution of the Offers by Years



High Premium Vehicle Group



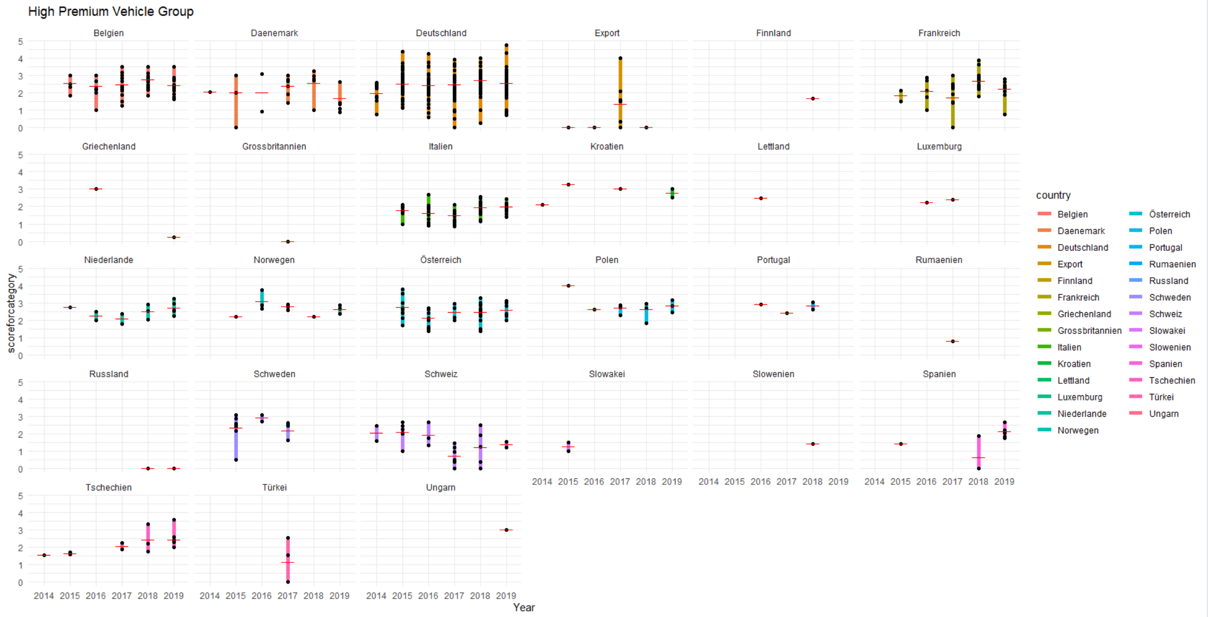
Premium Coach Vehicle Group

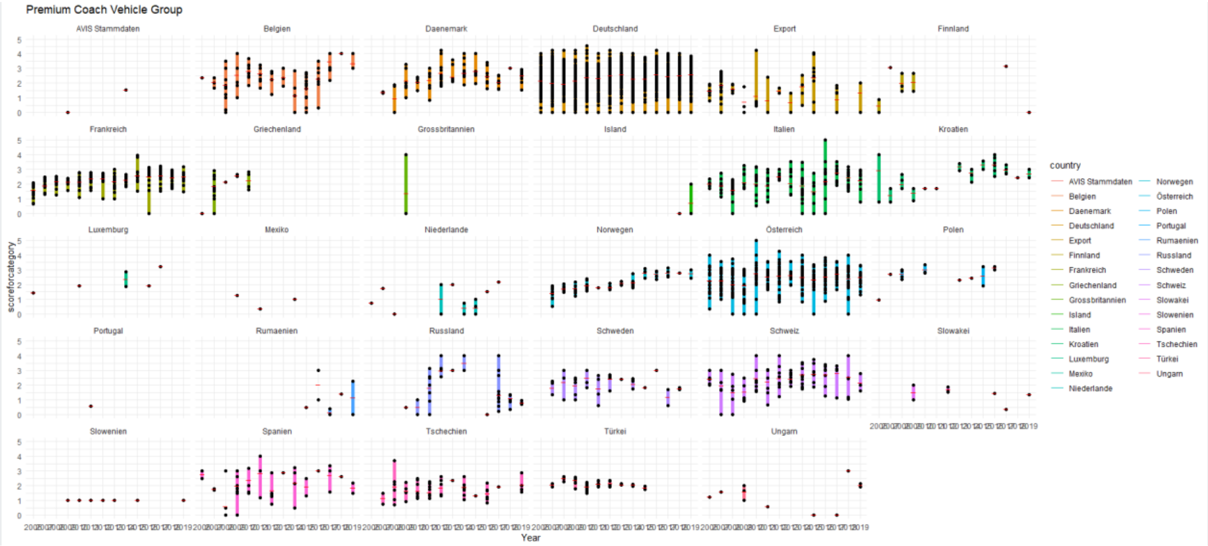


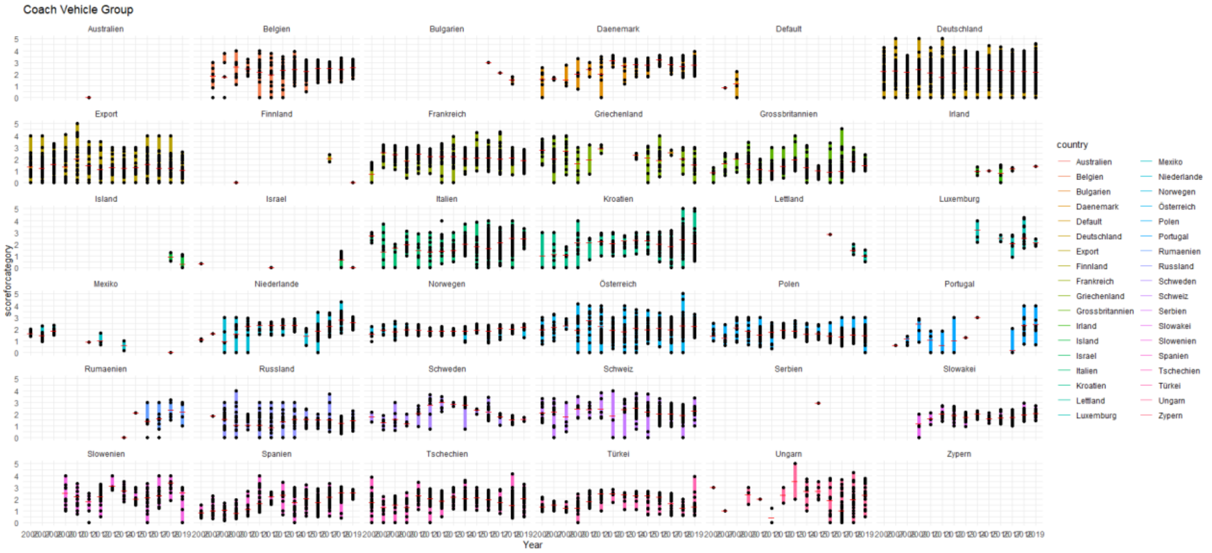
Coach Vehicle Group

Appendix D

Categorization Results for All Countries






Coach Vehicle Group
Australien
Belgien
Bulgarien
Daenemark
Default
Deutschland
Export
Finnland
Frankreich
Griechenland
Grossbritannien
Irland
Island
Israel
Italien
Kroatien
Lettland
Luxemburg
Mexiko
Niederlande
Norwegen
Österreich
Polen
Portugal
Rumaenien
Russland
Schweden
Schweiz
Serbien
Slowakei
Slowenien
Spanien
Tschechien
Türkei
Ungarn
Zypern
scoreforcategory
0
1
2
3
4
5
Year
country
Australien
Belgien
Bulgarien
Daenemark
Default
Deutschland
Export
Finnland
Frankreich
Griechenland
Grossbritannien
Irland
Island
Israel
Italien
Kroatien
Lettland
Luxemburg
Mexiko
Niederlande
Norwegen
Österreich
Polen
Portugal
Rumaenien
Russland
Schweden
Schweiz
Serbien
Slowakei
Slowenien
Spanien
Tschechien
Türkei
Ungarn
Zypern