E-CUSTOMIZATION OF AN ONLINE RETAILER

by
SADAF ZARRIN

**Submitted to the Sabancı Graduate Business School
in partial fulfillment of
the requirements for the degree of Master of Science in Business
Analytics**

Sabancı University
December 2022

ABSTRACT

E-CUSTOMIZATION OF AN ONLINE RETAILER

SADAF ZARRIN

Business Analytics M.Sc. THESIS, December 2022

Thesis Advisor: Asst. Prof. Burak Gökgür
Thesis Co-Advisor: Assoc. Prof. Ayşe Kocabıyıkoğlu

Keywords: Online Retailer, Consumer Behavior Analysis, RFM Method, K-means Clustering, Association Rule Mining

Data science and machine learning algorithms enable companies to track consumer behavior from large datasets for different markets. In online retail platforms, these analyses can help to extract popular products and expose these to customers to increase the purchase probability. In this study, we used the data of an online Turkish retailer containing 841 customers and 1282 transactions to propose two frameworks to improve the recommendation system (RS) and website layout of the company to stimulate more purchases. In the first framework, we utilized RFM (Recency, Frequency, Monetary) analysis to evaluate people's purchasing behavior in the context of these three variables. Next, by using these features and the K-means clustering algorithm, we assigned customers to different clusters. The four segments thus built are Loyal Customers, High-potential Customers, Prosperous Customers, and Departed Customers. In the final step, relying on association rule mining, 10 products were extracted from the frequently bought itemsets of the first three clusters to offer to the customers by the recommendation system. In the second framework, we developed a mechanism that adopts the lift metric of association rule mining to change the layout of the company's website to increase purchase probability.

ÖZET

ÇEVRİMİÇİ BİR PERAKENDECİNİN E-ÖZELLEŞTİRİLMESİ

SADAF ZARRIN

İş Analitiği YÜKSEK LİSANS TEZİ, Aralık 2022

Tez Danışmanı: Asst. Prof. Burak Gökgür
İkinci Tez Danışmanı: Assoc. Prof. Ayşe Kocabıyıkoğlu

Anahtar Kelimeler: Çevrimiçi Perakendeci, Tüketici Davranışı Analizi, RFM Metodu, K-ortalama Kümeleme, Birliktelik Kuralı Madenciliği

Veri bilimi ve makine öğrenimi algoritmaları, şirketlerin farklı pazarlar için büyük veri kümelerinden tüketici davranışlarını izlemesine olanak tanır. Çevrimiçi perakende platformlarında, bu analizler popüler ürünleri ayıklamaya ve bunları müşterilere sunarak satın alma olasılığını artırmaya yardımcı olabilir. Bu çalışmada, 841 müşteri ve 1282 işlem içeren bir çevrimiçi Türk perakendecisinin verilerini kullanarak tavsiye sistemini (RS) ve şirketin web sitesi düzenini iyileştirerek daha fazla satın almayı teşvik edecek iki çerçeve önerdik. İlk olarak, insanların satın alma davranışlarını bu üç değişken bağlamında değerlendirmek için RFM (Yenilik, Sıklık, Para) analizini kullandık. Ardından, bu özellikleri ve K-means kümeleme algoritmasını kullanarak müşterileri farklı kümelere atadık. Bu şekilde oluşturulan dört bölüm, sadık müşteriler, yüksek potansiyel müşteriler, müreffeh müşteriler ve ayrılan müşterilerdir. Son adımda, birliktelik kuralı madenciliğine dayanarak, müşterilere sunmak için ilk üç kümenin her biri için sık satın alınan ürünlerden 10'ar ürün çıkarıldı. İkinci yöntemde, şirketin web sitesinin web düzeninde verimliliği artırmak için birliktelik kuralı madenciliğinin kaldırma metriğini benimseyen bir mekanizma geliştirdik.

ACKNOWLEDGEMENTS

I want to sincerely appreciate my family who has been a great source of support. I also express my gratitude to Burak Gökgür and Ayşe Kocabıyıkoğlu for their guidance throughout this study.

To my beloved family and friends

TABLE OF CONTENTS

LIST OF TABLES . . . . **x**

LIST OF FIGURES . . . . **xii**

1. INTRODUCTION **1**

2. LITERATURE REVIEW **5**

2.1. E-Commerce and Consumer Behavior Analysis 5

2.1.1. RFM 9

2.2. Machine Learning and Clustering 11

2.3. Data Mining and Market Basket Analysis 13

2.3.1. Association Rule Mining and Store Layout Design 14

2.3.2. Association Rule Mining and Recommendation Systems 15

2.4. Contribution of the Thesis 19

3. EMPIRICAL SETTING AND DATA **21**

3.1. Baskasindaarama 21

3.1.1. Purchase Process 22

3.1.2. Brands 22

3.1.3. Price and Promotions 23

3.1.4. Selling Strategy 23

3.2. Data 24

3.2.1. Daily Sales Data 24

3.2.2. Monthly Traffic Acquisition 25

3.2.3. Brand Category Lists 25

3.2.4. Page Visits 26

3.2.5. Demographic Data 26

3.3. Descriptive Analysis 27

3.4. Data Preprocessing 29

4. METHODOLOGY **30**

4.1. RFM Method 31
4.2. K-means Clustering Algorithm 32
4.2.1. Elbow Method 34
4.2.2. Silhouette Analysis 34
4.3. Association Rules 35
4.3.1. Apriori Algorithm 36
4.4. Describing the Model 39

5. ANALYSIS AND RESULTS 41
5.1. Recommendation System Strategies 41
5.1.1. Summary of the Segments 46
5.1.1.1. Prosperous Customers 53
5.1.1.2. High-potential Customers 54
5.1.1.3. Loyal Customers 55
5.1.1.4. Departed Customers 55
5.1.2. Frequent Rule Mining of the Clusters 57
5.2. Improved Web Layout Framework 61

6. DISCUSSION AND CONCLUSION 63
6.1. Theoretical Implications 64
6.2. Managerial Implications 65
6.3. Limitations and Future Research 65

BIBLIOGRAPHY 68

APPENDIX A 73

LIST OF TABLES

Table 2.1. Advantages and Disadvantages of Different Recommendation Systems 18

Table 5.1. Silhouette Index Values for Different Number of Clusters 45
Table 5.2. Size of Clusters 46
Table 5.3. Average of Standardized RFM Values of the Clusters 48
Table 5.4. Results of the K-S Test for the Recency Values of the Clusters . 49
Table 5.5. Results of the K-S Test for the Frequency Values of the Clusters 50
Table 5.6. Results of the K-S Test for the Monetary Values of the Clusters 50
Table 5.7. Kruskal-Wallis Test Results for the Recency of the Clusters 51
Table 5.8. Mann-Whitney U Test Results for the Recency of the Clusters . 51
Table 5.9. Kruskal-Wallis Test Results for the Frequency of the Clusters .. 52
Table 5.10. Mann-Whitney U Test Results for Frequency of the Clusters ... 52
Table 5.11. Kruskal-Wallis Test Results for the Monetary of the Clusters .. 52
Table 5.12. Mann-Whitney U Test Results for Monetary Values 53
Table 5.13. Frequent Rules for the Prosperous Customers' Cluster 58
Table 5.14. Frequent Rules for the High-potential Customers' Cluster 59
Table 5.15. Frequent Rules for the Loyal Customers' Cluster 60
Table 5.16. Initial Layout of the Webpage 61
Table 5.17. Optimized Layout of the Webpage 62
Table 5.18. Lift Values for Current and Optimized Website Layout 62

Table A.1. One Page of the Retailer's Daily Sales Data 74
Table A.2. One Page of the Retailer's Monthly Traffic Acquisition 75
Table A.3. One Page of the Retailer's Sold Brands Data 76
Table A.4. Daily Page Visits Reports 77
Table A.5. Customers Demographic Report 78
Table A.6. First Five Rows of the Main Data Used in the Thesis 79
Table A.7. RFM Dataframe 80
Table A.8. RFM Dataframe with Standardized Values 81
Table A.9. Initial Layout for the Healthy Meals Page 82

Table A.10.Improved Layout for the Healthy Meals Page 82
Table A.11.Initial Layout for the Legumes Page 82
Table A.12.Improved Layout for the Legumes Page 82

LIST OF FIGURES

Figure 3.1. Revenue of each Month in 2020 27
Figure 3.2. Overall Turnover in each Day of the Months 28
Figure 3.3. Overall Number of Purchases from each Category Class 28

Figure 4.1. Proposed Methodology 31

Figure 5.1. RFM Variables Histograms 42
Figure 5.2. RFM Variables Boxplots 43
Figure 5.3. Elbow Method Plot for the Optimal Number of Clusters 44
Figure 5.4. Silhouette Plots for K=2 and K=3 45
Figure 5.5. Silhouette Plots for K=4 and K=5 45
Figure 5.6. Silhouette Plot for K=6 46
Figure 5.7. Size of Clusters 46
Figure 5.8. Standardized RFM Values of the Clusters 48
Figure 5.9. Category Class Purchase Frequency for Prosperous Customers 54
Figure 5.10. Turnover Values for each Month in the Second Cluster 55
Figure 5.11. Turnover Values of each Month for the Departed Customers .. 56
Figure 5.12. 3D Scatter Plot of the Customers 56

Figure A.1. Snapshot of the Retailer's First Page 73

1. INTRODUCTION

Many technology-savvy consumers search for products and services on the internet before making a purchase. In this process, customization plays a highly significant role in adjusting a standard product or service to meet the individual customer's needs. The objective is to increase value for customers and therefore boost loyalty. Furthermore, the company and producers can benefit from increased profits (Montgomery and Smith, 2009). Internet-based customization applications have greatly advanced, as the internet provides a well-suited environment for interacting with information. One of the technologies and analytical tools that businesses have invested heavily in to guide customers in their purchases is recommendation systems. Personalized recommendations and filters help consumers find interesting items from a large pool of products based on their preferences. Long-tail phenomena effects are another reason recommendation systems are considered necessary. Physical markets are limited by shelf capacity to provide only the most popular items and in online markets and retailers, this phenomenon forces the markets to recommend special items to individual customers because it is impossible to present all available items to them (Leskovec et al., 2020). This issue should be considered not only in the recommendation of products but also in the order of placement of products on the website. The more products that have a higher possibility of buying at the same time are placed next to each other, the more they are exposed to the customers and can encourage them to buy.

Amazon is one of the famous companies for using recommendation systems and offering personalized items to customers (Smith and Linden, 2017). Twitter is also a social information network that uses recommendation systems to help users find their favorite tweets. Without using such systems, users can be easily overwhelmed by the huge number of tweets created each day and the great amount of information people want to convey to others (Kywe et al., 2019). Other famous websites like Reddit and Facebook also utilize recommendation systems to help their visitors to find easily what they want.

To implement recommendation systems or decide about the efficient website layout of the online markets, companies need to perform consumer behavior analysis based on previous data and regarding the customers' demand and behavior towards different products and categories. Without understanding the customers' previous behavior, we will be unable to extract their interests and implement a suitable method.

The primary focus of this research is to provide two methods with the help of consumer behavior analysis and data science algorithms to increase the revenue of online markets. Specifically, we focus on the data of an online retailer in Turkey named *Baskasindaarama.com*. This website is a marketplace that provides products from different categories like lifestyle, FMCG, and self-care. The data provides us with transactional information on 2020 sales. It has the purchasing data of 841 unique customers provided in 2126 rows. This website does not use any methods to place the products on the website. Each page is just offering different items of the same categories placed next to each other by their brands' names. Moreover, the retailer does not benefit from the advantages of a recommendation system implemented by modern data science algorithms. They just suggest new or seasonality products to all the visiting customers without considering the differences between them. The first proposed framework is an improved recommendation system and the second one is implementing a method to change the website layout to increase the likelihood of more purchases.

By implementing these two methods, we want to answer the following questions:

- How can we segment the customers considering their different purchasing behavior?

  To answer this question in the first framework, we use the 'date', 'turnover', and 'transaction' columns in the purchasing data of the website to implement the RFM, a consumer behavior analysis method, to derive recency, frequency, and monetary scores for each customer. The 'date' column gives us information about the exact date the customer purchased a specific item. Therefore, by considering this information and the last date in the data, we can find out about each customer's recency score. Frequency shows us the total number of purchases of each customer and by adding up the number of transactions for each customer, this score can be calculated. By using the sum of 'turnover' values for each customer, the amount of money that was spent on the website is specified for each person. In the next step, we segment the customers by utilizing these three scores in the K-means clustering which is an unsupervised machine learning algorithm.

- How can we build a customized recommendation system to increase the revenue of the retailer?

  To build the recommendation system, we use the results of the K-means clustering algorithm that builds the clusters of the customers. Members of each of these groups have similarities in recency, frequency, and monetary. By implementing the Apriori algorithm, which is an association rule mining algorithm, we can identify popular products in each cluster and use them in the recommendation system. This method exposes popular products to customers and increases sales and income.

- What are the most bought products and categories in each segment?

  By implementing the Apriori algorithm, we can find out about the frequent rules in each segment and extract the most bought items from these rules. Moreover, we can discover the categories of these popular items.

- How can we propose a customized website layout to enhance revenue?

  If we can find the items that are most bought together, we can put these items next to each other on the website and increase the purchase probability. By utilizing the 'lift' metric of the association rules and seeking to maximize the sum of two-way lift values via Excel, we extract these items and change the website layout.

The summary of the implemented methods to answer these questions is explained step by step in the following:

In the first framework, we identify purchasing patterns using the transaction data of the retailer and data mining techniques. First, we use RFM, a behavior analysis method, to find out about the three characteristics of the customers in the data. In the next step, we utilize the results to implement K-means clustering, an unsupervised machine learning algorithm, to cluster the customers. The final step is utilizing association rule mining to find the most purchased products together in each segment and use them in the recommendation system of the retailer.

The second framework is using association rule mining to extract the two-way lift values that show the strength of the rules of buying each of the two products together. With the help of these values and trying to maximize the sum of them, we can place the products with higher two-way lift values next to each other to increase the probability of buying more products.

This thesis is organized as follows: Section 2 provides background and the literature review on consumer behavior analysis, machine learning, and data mining algorithms, especially two utilized methods, K-means clustering, and association rule mining. In Section 3, we give brief information about the data we used in the thesis along with data preprocessing explanation. Section 4 discusses the two implemented methodologies. Section 5 presents the results of the applied methods and algorithms. Finally, Section 6 provides the concluding remarks, the contribution of the research, a summary of the thesis, and the research limitations.

2. LITERATURE REVIEW

In this chapter, we discuss the literature review by exploring the research that has been conducted under the headings of e-commerce and consumer behavior analysis, machine learning and clustering, and data mining and market basket analysis. These are three closely related areas and have significant implications for businesses, marketers, and researchers seeking to understand and predict consumer behavior. By reviewing the existing literature on these topics, we aim to provide an overview of the current state of knowledge and find appropriate tools to employ or identify gaps in the research that may be addressed in future studies. Moreover, we discuss our contribution to the literature in these areas at the end of the section, including the frameworks we took in our research to reach the new insights and findings we discovered. Overall, this literature review aims to highlight the importance of machine learning, data mining, and recommendation systems in consumer behavior analysis and the potential contributions to our understanding of consumers' behavior in general.

2.1 E-Commerce and Consumer Behavior Analysis

Before the advent of the internet, consumers could only browse and purchase products at retail brick-and-mortar stores. These stores have several issues, including a limitation in numbers and physical space for inventory which cause limited flexibility to adapt to changing consumer preferences or trends and high operating costs which leads to expensive pricing. Customers had more of the issues. They were having difficulty purchasing what they needed because of high prices, limitations in reaching due to inconvenient locations and limited operating hours, the lengthy process of purchasing their necessities, limited options, and crowded stores (Chan and Pollard, 2003).

Today, however, consumers are increasingly turning to online stores for their necessities. The benefits of online retailers for producers and brands include costs reduction in running and selling, operating 24/7/365, ease of scaling up, affordable marketing and advertising, no reach limitations and connecting with potential customers all around the globe, faster response to market demands and expanding their brands quickly, providing more products and the most important one; collecting customer data and discovering Insights. The advantages for the customer side include quick access to favored retailers without taking a lot of time, simplicity and comfort of purchase with no reach limitations, having a large number of choices with more detailed information, product, brand, and price comparison, lower prices and several payment options (Amazon, 2021; Malloy, 2019).

Considering the available trends and statistics, it is clear that the transition from brick-and-mortar methods of buying and selling to online methods has taken place and is continuing to rapidly grow. E-commerce is the new reality and is the common desire and behavior of sellers and customers. Statista estimates that worldwide online retail sales were close to 1,250 and 3 billion dollars in 2013 and 2018, respectively, and will reach 6.5 billion dollars in 2023 (Kywe et al., 2019). The rise of this level of transaction highlights the importance of using various marketing analysis methods to figure out customers' needs and behavior to enhance the offerings and services of online markets.

Customer behavior study is the process of collecting, analyzing, and interpreting data about the behavior of customers in a particular market or industry (Abrardi et al., 2022). By studying customer behavior, manufacturers, marketers, and sellers can figure out what factors, how, why, when, and where, influence their customers' purchase decisions (Kim and Kim, 2022). Also, It is a way for businesses to understand how customers make their purchasing decisions, what factors affect their decisions and behaviors, and how they respond to different marketing strategies and recommendations, such as new products/services, retailing and advertisement operations (Kardes et al., 2014). Utilizing the information gained through consumer behavior study, retailers can provide better shopping experiences, by optimizing store layout and minimizing operational costs which eventually leads to obtaining higher revenue and profitability for the retailer (Abrardi et al., 2022).

In the internet era, analysis of customer behavior in e-commerce has become an effective and powerful analytical tool for businesses and sellers that are trying to better understand how online shoppers interact with an application or website, as well as predict consumer behavior since it significantly affects profitability due to the scale of the business (Alfian et al., 2019). Online shopping is on the rise, which allows companies to collect huge amounts of data on customer behavior, such as what products they purchase, how they interact with websites, applications and social media, and how they make their buying decisions. The data collected can be used to gain insights into customer preferences, predict future purchasing decisions and patterns, and tailor marketing strategies and recommendations to specific customer segments (Manis and Madhavaram, 2023).

Data is as much essential as the methods and tools to investigate customers' behavior. For customer behavior analysis in the traditional way, survey data, sales data, and CRM data were collected through questionnaire surveys, A/B testing, focus groups analysis, historical transaction databank, and customer relationship management systems as data gathering tools to be utilized, which were hard to reach, convert and analyze (Foxall, 2001). With the spread of e-commerce and the use of social media, and thanks to the emergence and development of big data technologies, accessing data and employing it has become much more convenient than before. In addition to traditional tools, new tools and methods such as customers' clickstreams, transactional histories, and shopping carts also help to make the data-gathering process more accurate and easier.

There are several approaches to gathering data and analyzing online customers' behaviors on various platforms, including:

Predictive analytics: This approach involves utilizing statistical and machine learning algorithms on historical data to identify and estimate the likelihood of future events and outcomes. A Predictive analytics approach can be employed to predict customer behavior in actions like churning or the likelihood of purchasing a specific product or at a specific time (Surendro et al., 2019).

Social media analytics: This approach involves studying customer behavior and their sentiment on social media platforms. A social media listening tool that includes tracking and analyzing large volumes of data can be used for this purpose. Moreover, posts can be read and analyzed manually (J. Zhao et al., 2021).

Web analytics: This approach involves studying data collected from a company's website, such as page views, conversion rates, time spent on the site, clickstreams, and suggestions to others. Businesses can use the information gained through this approach from data to understand how their customers interact with their website and what actions they take (Štimac et al., 2021).

Customer feedback and surveys: This approach involves collecting and analyzing customer feedback through methods such as surveys, customer reviews, and focus groups. Businesses can use this to understand how their products and services are perceived by their customers and how to improve them (Sari and Helena, 2021).

According to the available data, the questions, and the purpose of this research, predictive analytics and web analytics approaches will be used.

There are several methods and techniques to analyze customer behavior, which are also applicable to analyzing customer behavior in e-commerce settings. Some of these methods are mentioned below:

RFM (Recency, Frequency, Monetary) value analysis: This method analyzes how recently a customer made a purchase, how often makes purchases, and the total amount of money spends.

Customer segmentation: This method, based on characteristics such as demographics, interests, and behaviors, divides customers into different groups. This method helps businesses to target desired segments through suitable marketing and customer service efforts specific to each segment. Cohort analysis: In this method, customers are grouped into cohorts according to some common characteristic, such as the month they made their first purchase or the source of their first contact. This method helps businesses track the behavior of cohorts over time and compare them.

A/B testing: This method presents two different versions of a product or marketing action to two different groups of customers and compares the results. This method helps businesses understand which version is the most effective to influence consumer behavior.

Customer journey mapping: This method creates visual representations of the steps a customer takes in his or her interactions with a company website or application, from initial awareness to post-purchase evaluation. This method helps businesses realize the needs and motivations of customers at each stage of their journey. Heat map analysis: This method uses visual representations of websites and online stores to show where customers click and spend most of their time. This method helps businesses find out areas of the site that are most popular and actions that customers are taking on the website.

Cart analysis: This method analyzes the contents of abandoned shopping carts of customers to realize and explain the reasons that customers do not complete their purchases. This method helps businesses Identify any problems or obstacles that may prevent customers from purchasing (Solomon et al., 2017).

For this study, the RFM method will be used to analyze consumer behavior. This method and its characteristics are explained in the following subsection.

2.1.1 RFM

RFM analysis model is a useful behavior-based marketing analysis technique proposed by (Hughes, 1996). The model differentiates important customers according to their behavior from data by calculating customer scores for 3 attributes, namely; Recency, Frequency, and Monetary. The following is a detailed description of the RFM model attributes:

Recency of the last purchases (R): refers to the interval between the latest consuming behavior occurring and the last date in data. The shorter interval is better for recency.

Frequency of purchases (F): refers to the number of transactions over the course of a particular period, such as two times in a year, two times in a quarter, or two times in a month. The many the frequency means the bigger F.

Monetary value of purchases (M): refers to the amount of money used for consumption during a given period. The more the monetary means the bigger M.

RFM scores are very effective attributes to do behavioral multidimensional customer segmentation. According to the literature, R and F have a significant impact on the likelihood that customers will develop a new trade with businesses such that, as R and F grow, the more likely corresponding customers are to engage in a new transaction with a business. Further, it has been shown that the larger the M, the more likely the corresponding customers are to purchase products or services from enterprises again in the future (Cheng and Chen, 2009).

Although RFM is a useful and widely used method that helps to form customer segmentation and develop a marketing strategy for each segment. However, there are discussions about the importance of each of the scores. Hughes (2005) have considered the three attributes equal in terms of importance, therefore, their weights have to be identical. As a counterpoint, Stone (2008) claims that due to the characteristics of each industry and marketing objectives, the importance of the three attributes differs and therefore, the weights of the attributes are not equal. In this research, we considered the importance of all three features equal to each other.

RFM has been used in a great number of research because of its multiple advantages. Below we mentioned some of these:

- RFM is based on the historical data of the customers, which is easily accessible for the majority of businesses. Therefore, the analysis can be done fast and with minimal additional data collection.

- This method has shown to be a very successful technique for identifying valuable customers and focusing marketing efforts. Customers that score well on all three RFM characteristics are more likely to be receptive to marketing campaigns, according to research.

- RFM utilizes three simple factors to cluster the customers which makes it easy for the businesses to understand and analyze the findings.

- This analysis assists companies to segment the customers and target customers to implement successful marketing campaigns with higher response rates.

- This method can be used to segment clients in any industry, regardless of the type of product or service offered.

- To give a full picture of a company's customer base, RFM can be used in conjunction with other techniques like machine learning algorithms.

- Customers who haven't bought anything in a while or who have been spending less money can be identified using RFM which will enable businesses to take action to keep these clients.

- By examining the changes in the behavior of the customers who were the target of the campaign, RFM analysis can be used to determine the efficacy of various marketing campaigns.

- RFM can be integrated with other marketing tools such as CRM to help save time and improve the efficiency of different analyses.

Group-specific marketing, which is marketing targeted to a certain segment of customers, is much more needed and effective than the traditional mass marketing perspective. By reviewing the literature and noticing the advantages of RFM, we see that many researchers in the fields of manufacturing (Stormi et al., 2020), retail, and service (Das and Singh, 2023) frequently use the RFM method to aim this objective. RFM is cost-effective and easy to apply in quantifying customer behavior and is a transparent and easy to understand model. However, since it only considers existing customers, applying it solely can cause weak decisions about prospective customers and can lead to poor decisions because calculating the overall RFM score only gives the ranking of customers from best to worst (Jo-Ting et al., 2010). To overcome these limitations, machine learning clustering algorithms and data mining techniques can be employed in conjunction with RFM to successfully implement consumer behavior analysis in a large amount of data. In the following, we briefly talk about these mentioned methods.

2.2 Machine Learning and Clustering

In today's world, where a great deal of data is generated every moment, the importance of utilizing machine learning (ML) algorithms to analyze this volume of data is evident. There are different types of ML such as Supervised, Unsupervised, and Reinforcement algorithms which are employed for the aims of different tasks and applications, such as classification, regression, clustering, anomaly detection, control, and optimization according to the nature of the problem and dataset (Sarker, 2021).

With the digitalization and rapid growth in technology, the re-emergence and re-attention of data science, and the opportunity of implementing ML algorithms, behavioral science has evolved. Nowadays, machine learning is a powerful tool that can be used in consumer behavior analysis and making predictions about the future actions of customers such as purchasing patterns and churning. Through various algorithms, large sets of data can be analyzed to identify patterns and trends in consumer behavior which can be useful for group-specific and targeted marketing campaigns, as well as for identifying customers who may need extra attention to prevent them from becoming inactive or churning. Additionally, machine learning models can be used to create different types of personalized recommendation systems for customers based on their past behavior history. Overall, the use of machine learning in consumer behavior analysis can help companies to better understand their customers and make more informed business decisions (G Martín et al., 2021).

Clustering is a technique in unsupervised machine learning which involves partitioning a dataset into subsets by grouping similar objects or data points together into clusters based on a defined distance measurement between each subset. Machine learning clustering algorithms have many applications in pattern recognition, image analysis, and market analysis (Madhulatha, 2012). There are several machine learning clustering algorithms such as; K-means, hierarchical clustering (Agglomerative and Divisive), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM, Spectral Clustering, Affinity Propagation, Mean-shift Clustering, and so on Ray (2019).

Summary of some selected papers from the literature are as follows:

As an instance of the service sector using RFM, Ernawati et al. (2022) used the RFM model and spatial analysis as the basis for developing a geo-marketing strategy for a university in Indonesia. This framework would assist the university in selecting its promotional target market. In the study, data mining techniques and geographic information systems (GIS) were used to quantify the "value" of feeder schools and to examine enrollment patterns of the high-value feeder schools. Researchers found the distribution of feeder schools across regencies and cities and trends of enrolled students from the highest-value school segment, which can help university management choose target feeder schools more effectively. Decision-makers can use the findings to create a geo-marketing strategy for promotion and allocate resources accordingly.

To discover direct patterns and forecast future purchases of a retailer, Chattopadhyay et al. (2022), employed RFM and K-means algorithm. Next, six data mining models were applied to the patterns in order to predict profitable customers in each segment and whether each would purchase in the next six months. This method helped the researchers target the most consistently profitable customer groups to apply diversified marketing strategies.

In their paper, Christy et al. (2021) discuss performing RFM analysis on a business's data and using K-means and Fuzzy C-means algorithms to cluster customers. The authors proposed a new method of choosing the initial centroids in K-means and then compared the results of the two algorithms based on their execution time and cluster compactness. Results showed that the proposed K-means algorithm took less time and required fewer iterations than the fuzzy method.

Asllani and Halstead (2015) conducted a study to segment the customers using historical data from a company and the RFM methodology. The research then suggests and illustrates the use of a goal programming approach to decide the segments that should be targeted to maximize the profit for a hypothetical direct marketing campaign while considering various marketing priorities.

In their study, Essayem et al. (2022) used K-means clustering algorithm based on RFM to cluster the customer of a retail store. They acquired the data from the POS of the store. They decided to divide customers into 5 segments and then compared the results of each segment. According to their behavior, one segment was named potential loyal customers, while another showed possible customer churn.

Among ML clustering algorithms, K-means clustering has a great number of advantages such as scalability to handle large datasets and can be parallelized to run on multiple machines, general applicability, high-speed performance, efficiency, the guarantee of convergence, and ease of interpretation (Han et al., 2022). Many of the researchers in consumer behavior analysis have used the combination of RFM and K-means in their research. Considering the advantages of this algorithm and the literature, we realized that one of the most common and promising methods for clustering is the K-means algorithm and decided to use this in our thesis.

2.3 Data Mining and Market Basket Analysis

Data mining is a discipline growing in popularity and importance and is characterized by discovering patterns and knowledge from large amounts of data in databases that are useful for decision-making. Data mining can provide a significant competitive advantage to an organization by gaining meaningful insights into their operations to assist managers. Data mining process involves the use of various techniques, such as statistical analysis, machine learning techniques, and data visualization, to extract useful information from datasets (Bose and Mahapatra, 2001).

Market Basket Analysis (MBA) is a data mining technique used in retail to identify relationships, associations or co-occurrences between items in a customer's shopping basket (Ünvan, 2021). The goal of this analysis is to find interesting correlations, frequent patterns, associations, or causal structures that satisfy a minimum level of support and confidence between sets of items in data repositories such as transactional databases. It is widely used in a variety of areas such as inventory control, market analysis, and risk management (Kotsiantis and Kanellopoulos, 2006).

Association Rule Mining is an MBA method that identifies data item relationships and was first introduced by Agrawal et al. (1993) and now is one of the most important research techniques in data mining. It is a popular method to identify items that are frequently purchased together, also known as "affinity analysis" which can be used to create product bundles, and store layouts by determining the items that should be placed together in store displays or in a website, and to identify items that customers are likely to purchase together in the future for designing recommendation systems (Ünvan, 2021; X. Zhao and Keikhosrokiani, 2022).

In this study, we use association rule mining, an important data mining technique for the aim of market basket analysis for designing recommendation systems and store layouts in a segmented market.

2.3.1 Association Rule Mining and Store Layout Design

As stated above, one of the applications of association rules in market analysis is designing store layouts for retailers to increase revenue. In the following, some of the papers related to this subject are mentioned.

A study conducted by Cil (2012), talks about implementing association rule mining and multidimensional scaling technique to propose a new layout for Migros, a retailer in Turkey. This enables retailers to group products according to consumer purchasing patterns and helps them to save time and therefore, satisfy today's busy consumers. Because they can find what they need, customers are happier, and retailers' profit increases.

Association rule mining and optimization-based heuristics are used to propose a dynamic shelf relocation scheme for brick-and-mortar retailers in a working paper. This method provides them with a system for extracting and grouping profitable product pairs and optimizing the allocation of departments to store aisles. Depending on the nature of a retailer's target market, strategic rearrangement might produce more profit than a more traditional unchanged shelf space arrangement (Edirisinghe and Munson, n.d.).

Research conducted by Surjandari and Seruni (2005), used association rule mining and mix merchandising to find the strong rules between the products of a supermarket to design the layout of the market. They used WEKA software to extract the association rules.

Ozgormus and Smith (2020) proposed an analytical method using the dataset of Migros, a retailer in Turkey, to optimize the block layout of this retailer. They used the tabu search meta-heuristic algorithm on the market baskets to identify aspects impacting the sales ratio. In the end, considering desired adjacencies, they found some layout designs that helps to increase the revenue and suggested those to the retailer.

A study conducted by Halim et al. (2019), examined the data of an amusement arcade to solve the problem of using only certain types of games in the center by the customers while other games remained idle. By applying market basket analysis to the data, they proposed two layouts for the game center to increase the use of all devices by customers and increase revenue.

Association rule mining is a powerful technique that can be used to analyze customer purchasing patterns and identify relationships between different products. Using two-way lift, a measure of the strength of association between two items can optimize the store layout design to increase sales and improve the shopping experience for customers. Considering the research done so far, in which most of them have utilized this method for brick-and-mortar retailing, we will apply it to each product segment to suggest optimal layout design for the website which can identify unexpected relationships between products, leading to new product placement and cross-selling opportunities.

2.3.2 Association Rule Mining and Recommendation Systems

A recommendation system is a type of algorithm that uses data mining and machine learning techniques to analyze data on users' past behavior and preferences, usually associated with software tools or technologies, that make personalized recommendations and suggest items that are considered relevant to a particular user, typically when there is a great deal of information available and searching or selecting items is difficult (Almonte et al., 2022).

Recommendation systems are widely used in various applications such as e-commerce, social media, entertainment, and news to suggest songs, movies, or products to customers With the aim of helping users to discover new products, services, or content that they might be interested in. As a result, the implementation of a recommendation system leads to an increase in consumer satisfaction by providing personalized recommendations that align with their individual preferences, thus allowing customers to make more informed purchase decisions. Additionally, the use of recommendation systems can also benefit the marketer by increasing conversion rates and revenue through targeted marketing efforts (Deldjoo et al., 2020).

According to X. Zhao and Keikhosrokiani (2022), based on the type, recommendation systems can be categorized into four main groups:

- Collaborative Filtering

A collaborative-based recommendation system is a type of recommendation system that uses the past behavior, preferences, or activities of a group of similar users to make recommendations. It works by identifying the similarity between the users in their behavior such as their ratings or actions on the website of retailers. These systems can be very effective because of capturing the subtle patterns in users' behavior. However, if there are less data to work with and if there are not many users with similar preferences in the data, this system can not be much effective (Almonte et al., 2022).

Collaborative filtering systems can be implemented in two ways: Memory-Based Collaborative Filtering and Model-Based Collaborative filtering, the first one makes recommendations based on the memory of past interactions, while the second one, makes recommendations based on a model learned from past interactions (Behera and Nain, 2022).

- Content-Based Filtering

Content-based filtering is a type of recommendation system that recommends items that are similar to the items that a user has liked or consumed in the past. It uses the features of the items, such as their genre, director, actors, etc. in multimedia to make recommendations (Deldjoo et al., 2020). For online retailing, these systems typically use features such as item metadata (e.g. title, description, category) and customer behavior data (e.g. purchase history, click data) to make recommendations based on Item-item similarity or User-item similarity (Mehta et al., 2021).

Content-based filtering can be improved by combining it with collaborative filtering methods, this way it can take into account both the features of the items and the preferences of other users to make recommendations (Mehta et al., 2021).

- Hybrid Methods

Hybrid recommendation systems are a combination of different types of recommendation systems, such as content-based filtering and collaborative filtering. These systems utilize the strengths of each approach to make personalized recommendations. Hybrid methods have been shown to be more effective than pure collaborative or content-based methods in several scenarios and can improve the overall performance of the recommendation system (Mehta et al., 2021).

- Association Rules

Association rule-based recommendation systems use the concept of association rules, which are a set of if-then statements that describe relationships between items in a dataset. These systems analyze transactional data to identify patterns of co-occurring items and use these patterns to make recommendations. Association rule-based recommendation systems are commonly used in the retail and e-commerce industry for market basket analysis (Ricci et al., 2015), due to their high explainability (Zhang et al., 2020). However, these systems are less commonly used in other areas such as music, news, or movie recommendations.

Table 2.1 represents the advantages and disadvantages of each group of recommendation systems (Mehta et al., 2021).

Considering the advantages and disadvantages of each type of recommendation system, the purpose of this research, and the available data, association rule-based recommendation systems were chosen.

Considering the strengths and weaknesses of each of the mentioned types, association rules-based recommendation system has been chosen for this research due to the following reasons:

1. Collaborative Filtering is sensitive to data sparsity and considering that the available data are sparse, the use of this method may not lead to reliable results.

2. Content-Based Filtering is sensitive to data quality, and highly relies on item features that the existing data do not have enough features about each item and leads to results with a lack of diversity in recommendations because only considers item features not user preferences.

3. Since Association Rules-based Recommendation Systems can uncover hidden patterns and relationships between items, recommends based on purchase history and co-occurring items, and are explainable and interpretable. This type of recommendation system can provide real-time recommendations with constant updating which is in alignment with the goals of the site and the essence of online retailing. Not only it can be used for market basket analysis but also it can be used for product placement and layout design optimization. Most important, It easily can be integrated with other methods to improve recommendation accuracy which is the ultimate objective of this research.

The Apriori algorithm and its variants are widely used in association rule-based recommendation systems (Varzaneh et al., 2018), which will also be used in this research. In the following some of the studies are mentioned:

In their research, Essayem et al. (2022), worked on user behavior analysis and utilized the sales data of a company for sales prediction and building recommendation models. First, they used the RFM analysis method with Random Forest and XGBoost machine learning algorithms for sales prediction. Their results showed the XGBoost has higher performance and accuracy than Random Forest. Next, Apriori, one of the association rule mining algorithms for basket analysis was used to recommend products for the customers.

Kumar and Balakrishnan (2019) worked on a recommendation system developed by using the Apriori algorithm to recommend different agricultural products including

Table 2.1 Advantages and Disadvantages of Different Recommendation Systems

Collaborative Filtering	**Advantages**	· Leverages crowd wisdom for accurate recommendations. · Handles the ’cold-start’ problem for new users or items with little or no ratings. · Handles sparse data for users or items with limited ratings. · Provides personalized recommendations without item features.
	Disadvantages	· Sensitive to data quality, relies on user ratings/actions · Sensitive to data sparsity, relies on user ratings/actions · Scalability issues with increasing number of users and items · Privacy concerns with sharing personal user preferences
Content-based Filtering	**Advantages**	· Accurate recommendations with item features · Handles ’cold-start’ problem for new users · Handles ’item cold-start’ problem for new items · Personalized recommendations based on user’s past preferences
	Disadvantages	· Sensitive to data quality, relies on item features · Scalability issues with increasing number of items · Suffers from ’information overload’ with many items to recommend · Lacks diversity in recommendations, only considers item features not user preferences
Association Rules	**Advantages**	· Handles large amounts of data and scales well to large datasets · Uncovers hidden patterns and relationships between items · Recommends based on purchase history and co-occurring items · Provides real-time recommendations with constant updating · Simple and easy to understand, relies on counting and association techniques · Can be used for market basket analysis and product placement optimization · Easily integrates with other methods to improve recommendation accuracy
	Disadvantages	· May not consider user preferences or ratings · Can be computationally expensive for large datasets · May not consider other factors such as time, location, or context
Hybrid Methods	**Advantages**	· More accurate recommendations by combining different approaches · Handles ’cold-start’ problem for new users or items with little or no ratings · Handles sparse data for users or items with limited ratings · Personalized recommendations by considering both item features and similar user preferences
	Disadvantages	· More complex to implement and require more resources · Sensitive to data quality, relies on user ratings/actions and item features · Scalability issues with increasing number of users and items · Sensitive to data sparsity, relies on user ratings/actions

vegetables and fruits to the customers of a website. For implementing this method, they used the ordered data of the past 8 months of the website.

Yan et al. (2022) used the Apriori algorithm to design a recommendation system to recommend personalized information for the users of a website. Another paper by using a Fuzzy system creates the association rules from customers' buying habits and then by using the Apriori algorithm extracts the frequently bought items for use in the recommendation system.

Research conducted by Obeidat et al. (2019) on a website that provides online courses to customers built two recommendation systems and compared the results. The first one was constructed by clustering the students based on their course selection and then implementing the Apriori algorithm to create the association rules and extract the most bought items. The second one was implementing the system without clustering the customers. Their results show that Clustered rules provide better coverage than individual rules.

The data of a book management system was utilized to build a recommendation system for personalized suggestions. They used an improved Apriori algorithm and defined a threshold for the support and confidence values to mine strong association rules. The results of their experiment show the positive impact of the implemented RS on recommending books (Zhou, 2020).

2.4 Contribution of the Thesis

Considering the conducted research mentioned above for each method and algorithm, the overall contribution of this study is adding new insights to the field of recommendation systems, customer behavior analysis, and website layout design by proposing two new frameworks:

1. Making association rules-based recommendations based on clusters of customers rather than general recommendations

The application of a combination of RFM scoring, K-means clustering, and the Apriori algorithm for association rule-based recommendation systems could improve the accuracy and relevance of the recommendations made to customers by taking into account their past behaviors while grouping them into similar segments. Additionally, the combination of these techniques could uncover new insights and patterns in customer behavior that were not previously known.

2. Association rules-based and category-wise layout design

The second proposed framework is inspired by the papers that used association rule mining to change the layout of different brick-and-mortar stores. By using the lift values extracted from the Apriori algorithm to obtain the measure of association between different products in each product category on the website, improve the web layout based on the previous behavior of customers.

The use of these two frameworks for the online retailer *'Baskasindaarama.com'*, which does not use any exact scientific method for recommendation system and website layout design, will have the following operational and applicational contribution: The first framework can contribute to their selling process and increase selling and revenue amount by personalized recommendations with the help of consumer behavior analysis. The second framework can contribute to the marketing process of the website by optimizing the layout of the website to better match the preferences of customers which can lead to improved conversion rate and cross-selling.

3. EMPIRICAL SETTING AND DATA

We discuss the retailer in detail in this section. In 3.1 part, we introduce the examined retailer in the thesis providing information about its purchase process, sold brands, price and promotions, and selling strategies. The 3.2 part and subsections are allocated to explain the daily sales data, monthly traffic acquisition data, brands categories, page visits, and demographic data. In the 3.3 part, we describe the descriptive statistics of the data. 3.4 part is dedicated to explaining the data preprocessing before implementing the methods on it.

3.1 Baskasindaarama

Baskasindaarama.com is an online retailer founded by Miss Esra Sarihan and Miss Melis Sarihan in 2016 in Izmir. They started their business with small initial capital and by selling a few brands on their website. The primary aim was selling high-quality products and offering high-quality delivery service and trying to build a friendly warm relationship with their customers. Another aim of their website was to promote and sell products of women farmers and female entrepreneurs. They specifically gave the opportunity for females who use domestic primary raw materials to produce their products. Most of these products are being sold under the groceries category of the website. A snapshot of the website is indicated in Figure A.1. In 2020, the retailer was working with over 100 brands under 6 to 9 categories on the website changing in number according to different situations like seasonality. They provide services all over Turkey. All these aims and efforts put them as one of the top 20 women entrepreneurs in the ‘Startup Turkey’ event with business owners and participants from 63 different countries. Moreover, they won the Bronze Stevie Award in 2019 which is one of the prestigious international awards for successful businesses. They won the prize again in 2020 to be the only retailer in Izmir to win the prize in two consecutive years. Caring for the high quality of the products, high-quality packaging, unique designs, and trending features are the characteristics that make this online retailer unique and different from others. In the following parts, we provided detailed information about the purchasing process, pricing strategies and promotions, and selling strategies of the company.

3.1.1 Purchase Process

This merchant operates as a drop shipping company and does not have its own inventory. As a result, we can claim that it serves as a hub for customer order delivery. The process of preparing the orders of the customers is as follows: First, they make a contract with the producers and brands and then upload the products to the website. If a customer places an order, the brands will produce the order or send the products from their specified warehouses to the retailer. In the next step, after quality control of the products, the retailer will put the products into special packages and send them to the customer with the invoice. For delivering the products to the customers, the retailer uses UPS services. Because trust along with high quality is important to the retailer. If something happens to the products the retailer will resend them without additional charges.

3.1.2 Brands

At the first steps of launching their business, baskasindaarama.com signed contracts with brands like Roli Doner and Bugatti which are perceived as luxurious brands. In the next steps, they started to work with high-quality and affordable products such as Neutrogena and Simple with more potential customers. This strategy helped the company to capture more portion of the market.

By implementing these strategies and making sure to have the trust of the customers, they had the chance to introduce and include some other international brands on their website and be sure to have the chance to sell them. In addition to all these strategies, by checking different brands' objectives before including them in their platform, they arrived at this point of working with lots of brands and selling over 500 products on the website.

3.1.3 Price and Promotions

To come up with the most appropriate low prices, the company is in constant contact with the brands to become aware of the finished product costs, limits, desired selling prices, and promotions and set the selling prices accordingly. The brands send new campaigns, updated prices, or promotions biweekly to the company. Therefore, the pricing policy of the retailer is mostly determined by the brands, and they can only decide about their own profit margins.

Moreover, the retailer checks the prices of other platforms for highly competitive products. For keeping the competitive advantage for perishable products such as baked goods, food, or dairy, they keep the prices close to the market prices and with a small profit margin.

3.1.4 Selling Strategy

The first and main part of the strategies of the retailer is selling high-quality products at affordable prices. The second step after catching the attention of the customers with this strategy is trying to offer more expensive, lesser-known products which allows the customer to choose from a more product variety with different properties.

Another strategy for increasing sales is providing a feature that enables customers to ask about the availability of their needed products that are currently out of stock. Along with helping to build a friendly relationship with the customers, this feature helps the retailer to become aware of the customer's needs and desires and therefore, determine the order quantity in advance.

One other strategy used by the company is utilizing a scoring system to reward customers for their specific activities. These include shopping for a certain amount or introducing the platform to a friend. By gaining a specific number of scores, discount codes will be provided to the customers for their future purchases.

The general strategy of the retailer is not to store perishable products. However, after the pandemic, they reconsidered this strategy and allocated space for these products in their stock. Promoting these products with a close expiry date for the customers considering their overall spending on the website is another selling strategy of the retailer. This method helps the customers to become aware of other products and motivates them to order more.

One of the efforts of the company in setting strategies until now was entering the eastern regions of Turkey. Not being familiar with e-businesses and social media and not using financial cards by women in these areas are the difficulties in catching the market. However, because of the low number of competitors penetrating this market would be very beneficial and profitable.

The selling strategy that we want to add to all these efforts is designing an improved recommendation system and changing the website layout to increase the purchase probability. We discuss the data that is used to support this thesis in the following part of this section.

3.2 Data

The main data used by the retailer to analyze and control their resources is generated via Google Analytics and are in different levels providing information on the retailer's performance and operations. The first and most important report is Daily Sales which shows the purchased products and customers and the amount of turnover for the transactions. This data is explained in the 3.2.1 part. The second report explained in 3.2.2 is the Monthly Traffic Acquisition of the website. This report shows the used sales channels by the retailer to gain website traffic and promote its products each month. The third report depicts the brands and products and main categories sold by the retailer. The next report shows the daily views of the webpages on the website. The following subsections are explanations of each of these reports.

3.2.1 Daily Sales Data

Daily sales data of the retailer is in 11 excel documents and each has the purchase information for one month from January to December. The file for March is not included since there was no transaction during that month. Each sheet in the report summarizes the daily transactions. Table A.1 represents one page of this report.

Each sheet of data contains 8 columns. The first column depicts the ID of the customers which are coded with a 'C' at the beginning of the customer number.

The second column is the URL representing the webpage of the bought item. The ‘Chart situation’ column represents two different statuses of the transactions which are ’completed’ or ‘pending cart’. The ‘number of ordered’ column is showing the number of items being purchased in the transaction. The ‘price’ column shows the price of one unit of the item. The ‘turnover’ column shows the total revenue from those items in the transactions. This value is equal to the number of orders multiplied by the price. The ‘profit margin’ column is varied from 10% to 50% and depicts the profit percentage of each item in a range. Some values of this column are 0 which is related to some special campaigns with 0 percent profit. The last column shows the ‘special campaigns’ of the website.

To be able to work with the data and run different models on it, we combined all the documents into one single sheet. We added the date column. Another column named transaction was added containing the date transactions statuses. We can say that this column shows the market baskets. By using a Python script and the URLs, we extracted other related information and added them to the data. These columns are the name of the items, the main category, brand, and category class of the items.

3.2.2 Monthly Traffic Acquisition

The channels that led users to the website are shown in the monthly traffic acquisition report for the first half of 2020. Each report sheet is devoted to a single month and contains rows that display the channels and two columns that define new and returning customers. Search engines like Google, direct access, and referrals like social media and blogs make up the channels for the first four months. Paid advertisements and displays such as YouTube are added to the channels for May and June. One page of the report is shown in Table A.2.

3.2.3 Brand Category Lists

The information about the brands and their assigned categories that they were working with in 2020 is provided in the brand list report and is represented in Table A.3. The report is like a flowchart showing the main categories at the top and the brands under them. The 7 categories of the report are jewelry, cosmetics, apparel, baby care, stationary, lifestyle, and groceries. There are some changes in the categories now. However, the data we are using consists of these 7 categories.

3.2.4 Page Visits

The Page Visit report indicates the daily visit to the product pages. At the time we received the data, only the first 5 months of 2020 were covered by this report. It contains the URL link of the products on the retailer's website and the number of visits. The overall number of visits, the number of users who left the website right away, and the average time spent on each page are also included. Table A.4 shows one page of this report.

3.2.5 Demographic Data

The customer demographics for the first half of 2020 are included in this report and one table of it related to the gender, age, and nationality of the customers is shown in Table A.5. This report provides details about the visitors' gender, age, nationality, interests, and the devices they use to access the website. The top 9 nations with the highest visit percentage are displayed in the "country of Origin" field. The 'age' category is divided into 6 groups, each starting at 18 and going up by 10, and is again displayed in percent. The percent utilization of the various device categories (mobile, desktop, and tablet) is also displayed. Additionally, the interest category, which is separated into 30 groups, shows the consumers' interests based on their searches. In the next part, we talk about descriptive analytics and the insights we received from the data.

3.3 Descriptive Analysis

The descriptive insights gleaned from the data are presented in this section. The data contains sales information for the full calendar year of 2020, from January 1 to December 31. It contains 2126 rows, and each row is related to a specific item of a transaction. There are also 18 columns in the data that each depict one unique piece of information about the transactions. The first five rows of the data are shown in Table A.6.

There are 1282 unique transactions in the data. The number of customers is 841. In total 4125 items from 599 products and from 97 categories in 7 category classes are sold. The mean for the ordered numbers is 1.94 with a standard deviation equal to 3.22. The data shows transactions in 252 days of 2020 from 59 different brands. The number of sold items in one month (March) is zero. As we discussed before, turnover is the price of each item multiplied by the number of orders and is very low in the first 5 months of the year. This can be the result of the start of the Covid-19 pandemic. The mean of turnover is 393.59 and the standard deviation is 1144.302 which is a high value. In the second half of the year, the revenue increases greatly. It is descending from the 6th month to the 11th month. However, there is a grand increase in December and that would be because of the Christmas campaign. Figure 3.1 below shows the total revenue for each month.

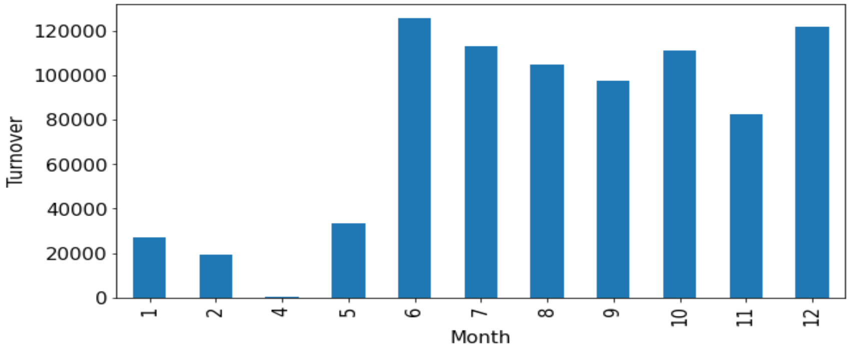

Figure 3.1 Revenue of each Month in 2020

The number of orders of an item in each transaction is between 1 and 100. Additionally, the dataset contains transactions containing up to six unique products. Moreover, there are 19 special campaigns in the data and most of them are in the first two months and the last six months of the year. One campaign related to Christmas lasted for one month and others lasted for no more than two or three days. The customer with the (C846) customer code has the highest value of turnover equal to

12589 TL in total. According to Figure 3.2 below, the total turnover value is higher in the first half of the months. We can conclude that most customers buy their needs on the first days of the months.

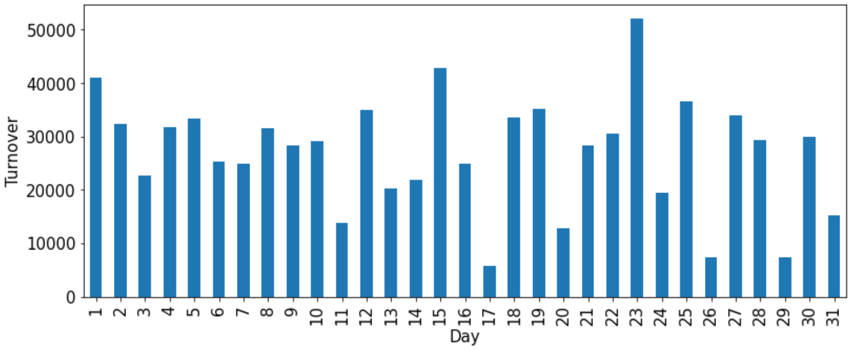

Figure 3.2 Overall Turnover in each Day of the Months

Figure 3.3 indicates the overall purchased items from 7 categories in the data. Most of the bought items are from the Grocery, Cosmetics, and Lifestyle category classes. Additionally, the greatest number of bought items are from Barbeque, Flour, and Sause categories.

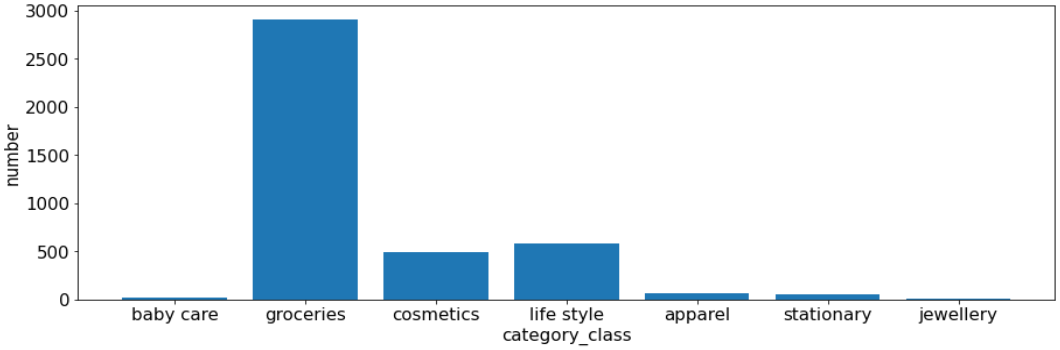

Figure 3.3 Overall Number of Purchases from each Category Class

3.4 Data Preprocessing

Data preprocessing is the most important part of a data science problem and most of the workload is related to this part. However, as we mentioned in the 'Daily Sales Data' subsection, since we created the data ourselves by using Python scripts to extract the information from the web pages, there was not much work to be done at this step. Some of the web pages did not have data related to the categories. Therefore, there were some null values here. First, we checked for the null values in the data and we filled them with the information we acquired from the website. Second, we checked for duplicate data and there were no such columns. In the final step, we checked the data types in Python. The 'date' column was defined as an object in the data and this was an obstacle to using it as date and time values. Therefore, because it plays a crucial role in our methods, we changed the column to a 'datetime' variable. In the next section, we discuss in detail the methodology that we used to implement our frameworks.

4. METHODOLOGY

In this section, we explain the methods and algorithms that we used to implement our frameworks on the data. We know that this retailer has two ways to select the products for recommending them to the customers. The first one is named the Fixed method which means that they recommend some new but unvarying items to all the customers. The second method is the Season-based method which refers to recommending products related to the season they are in. Although this mechanism introduces new and most demanded products; however, they don't consider the customers' preferences and heterogeneity. Developing a mechanism that provides customized recommendations according to the similarities between the customers would result in their satisfaction and revenue increase. For the successful implementation of a mechanism to help us in this process, we used RFM, the K-means algorithm, and association rule mining respectively.

Moreover, we provide another framework by using the lift values of the association rules mining to rearrange the website layout of the retailer to increase the probability of selling more products to the customers. Figure 4.1 is an illustration of these two proposed frameworks. We explain the utilized models and algorithms in the following.

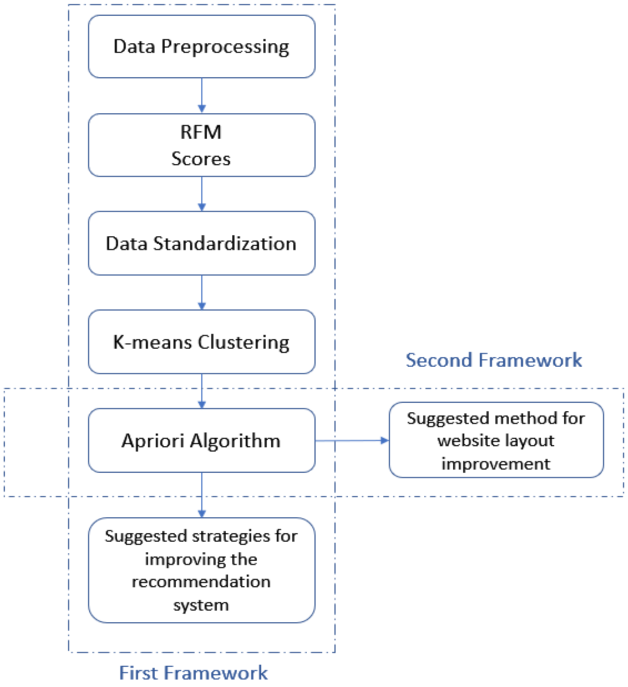

Figure 4.1 Proposed Methodology

4.1 RFM Method

RFM (Recency, Frequency, Monetary) is a behavior-based model and is used to assess customer behavior and generate predictions based on that behavior in the database (Hughes, 1996). This model has been extensively applied in many sectors and successful businesses, especially in direct marketing (Jo-Ting et al., 2010). Decision-makers can develop effective marketing strategies by adopting the RFM model. The definition of RFM according to Jo-Ting et al. (2010) is:

Recency: the amount of time since the customer's last purchase up to the last date in the data.

Frequency: indicates the number of purchases made during a given time frame.

Monetary: signifies the amount of money spent within this time frame.

For valued customers, the frequency and monetary scores are high, and the recency score is low. With the help of these three given scores, researchers can classify the customers and build their CRM strategies. RFM is simple and easy to use, yet it is powerful in identifying the target customers to increase the revenue of the companies. It is important to keep in mind that RFM analysis is based on past purchase behavior and therefore, it doesn't take into account future buying potential or changes in customer behavior. However, it is still a valuable tool for understanding customer behavior and making data-driven decisions (Christy et al., 2021).

We utilize this method in our research because of its successful results in a great number of research. Moreover, the major variables in our data are related to date, number of purchases, and the revenue of the company, which are very useful for implementing this method. For calculating the recency in our research, we counted the days from the last purchase of each customer to the last date in the data and recorded it as the recency score of each customer. To obtain the frequency score, we counted the number of transactions related to each customer. For the monetary score, we added up the turnover values for each customer in the 'turnover' data column. The new scores were stored in a new dataframe.

4.2 K-means Clustering Algorithm

K-means is a fast iterative and point-based unsupervised machine learning clustering method and was first developed by MacQueen (1967). K-means clustering algorithm aims to cluster the data points of a dataset to K number of clusters. It starts initially with arbitrary cluster centers and continues assigning all points to the closest center. The centroid of each cluster is then recomputed as the mean of all the points in the cluster. It repeats the assignment process until no change happens between two iterations. The algorithm tries to maximize the intra-cluster distance and minimize the distance between the points in each cluster. Most widely, the Euclidean distance method is used for calculating the distance in the algorithm. Equation 4.1 is the Euclidean distance formula where p and q are two points in Euclidean space and q_i and p_i are the Euclidean vectors starting from the origin of the space.

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{4.1}$$

To clarify the logic of the algorithm and understand its flow of computation, the pseudocode of the K-means clustering algorithm is indicated in the following where D is the list of data points (Nazeer and Sebastian, 2009):

```
input:
    $D = \{d_1, d_2, \ldots, d_n\}$
    $K$= number of clusters
output:
    $K$ clusters containing different points of the dataset
pseudocode:
Step 1. specify the number of clusters ($K$)
Step 2. assign the centroids randomly
Step 3. do {
            assign each point to the nearest cluster
            recalculate the mean value of the clusters and update the centroids
            }
termination point. while {
            there is no change in the points of each cluster
            }
```

K-means is sensitive to the initial placement of centroids. Therefore, it is a good practice to run the algorithm multiple times with different initial centroids to get the best result. Moreover, it is crucial to select the appropriate number of clusters before running the algorithm, otherwise, it may not perform well. In our study, we apply this algorithm to the recency, frequency, and monetary scores of the customers extracted from the RFM method. It would help us to generate the clusters for the customers according to their shopping behavior. Additionally, two methods are used in our study to find the best number of clusters which are described in the following.

4.2.1 Elbow Method

Elbow method is a way of helping the researchers to decide the number of optimal clusters in a cluster analysis. It calculates the total Within Clusters Sum of Squares (WCSS), which is the squared distance between the cluster center and each point in the data for a given range as the number of segments. When the dramatic decrease in total WCSS happens, the method chooses that as the optimal number of clusters. Increasing the number of segments after this value would not result in better modeling. By plotting WCSS against the number of clusters, we can see the elbow in the optimum value (Bholowalia and Kumar, 2014).

4.2.2 Silhouette Analysis

Silhouette analysis method is a way of representing how well the results of a clustering algorithm are. Silhouette index measures the distance of neighboring clusters and represents how well the data points are separated by measuring the similarity of each point to its assigned cluster and the other ones. The result is a value between -1 and +1 and the more the value is close to 1, it shows the point is assigned to a good cluster and is far from its neighboring clusters. If the result is close to -1, it indicates that the data point is more like other clusters and has been assigned to the wrong cluster. Equation 4.2 is used for computing the Silhouette index. a is referred to mean intra-cluster distance and b is the mean nearest-cluster distance. The more the value of the Silhouette index, the better the data points are clustered. Moreover, the Silhouette width for each cluster in the Silhouette plot is an indicator of the size of the clusters (Rousseeuw, 1987). Both the Silhouette index and the Silhouette plot can help the researchers to find the optimal cluster number.

$$SI = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{max(a_i, b_i)} \tag{4.2}$$

4.3 Association Rules

As we discussed in the literature review, association rule mining is one of the most important data mining techniques with an iterative approach for market basket analysis which seeks to uncover frequent patterns or associations among a transaction database. Suppose D is a database with a number of transactions T_s and I is defined as a set of m distinct items $I = I_1, I_2, ..., I_m$. T is a transaction in the dataset containing a set of items from I such that $T \subseteq I$.
$X \Rightarrow Y$ is an association rule where X and Y are two item sets such that $X, Y \subset I$ and $X \cap Y = \varnothing$. This rule is interpreted as 'X implies Y'. X is labeled as antecedent and Y is named consequent (Kotsiantis and Kanellopoulos, 2006).

Support, confidence, and, lift are three measures to understand the strength and nature of the association rules (Berry and Linoff, 2004). Support of an association rule refers to the number of occurrences or frequency of a particular item or combination of items in a dataset. It is usually expressed as a percentage or a fraction of the total number of transactions in the dataset. Low support values indicate that items or combinations of items are rare, while high support values indicate that items or combinations are common. The support value is utilized as a threshold to determine which items or combinations of items to include in the association rules.

$$support(X \Rightarrow Y) = (X \cup Y)/Total\ number\ of\ transactions(T) \tag{4.3}$$

Confidence in an association rule refers to the likelihood that a particular rule or association is true. In other words, it is the strength of the relationship between an antecedent (a preceding event or condition) and a consequent (a subsequent event or condition). Typically, confidence is expressed as a percentage or a fraction and is calculated as the number of times the antecedent and consequent occur together, divided by the number of times the antecedent occurs. In general, high confidence values indicate that the rule is likely to be true and that the antecedents and consequents strongly correlate, while low confidence values indicate less likelihood of the rule being true. Confidence is used to prune the set of association rules generated by the algorithm, by removing rules that have low confidence values.

$$confidence(X \Rightarrow Y) = (X \cup Y)/Total\ number\ of\ transactions\ contains\ X \tag{4.4}$$

For instance, if the support value of an item in a retailer's dataset is 0.2%, it means that 0.2 percent of transactions in the dataset contain purchasing that item. Moreover, if the confidence of the association rule $X \Rightarrow Y$ is 50%, it means that 50% of the transactions containing X also contain Y.

Lift of an association rule is defined as the observed support value divided by the expected support if X and Y were independent. An association rule is considered strong if lift>1 and the higher value of lift implies a stronger connection between items. A value less than 1 indicates that the antecedent and consequent are negatively associated, and the lower the value, the stronger the negative association. A value of 1 indicates that the antecedent and consequent are independent. Lift can be used to identify rules that have a stronger association than would be expected by chance, and to help prioritize which rules to examine further (Tseng and Lin, 2007).

$$lift(X \Rightarrow Y) = support(X \cup Y)/support(X).support(Y) \tag{4.5}$$

The equation can also be written as below(Sagin and Ayvaz, 2018):

$$lift(X \Rightarrow Y) = confidence(X \Rightarrow Y)/support(Y) \tag{4.6}$$

Since the number of association rules is sometimes very large, minimum Support, Confidence, and Lift values are defined before implementing the algorithm. In the following, we talk about the Apriori algorithm, a basic algorithm for implementing association rule mining.

4.3.1 Apriori Algorithm

Apriori is the first classical algorithm introduced for mining the most repeated patterns in association analysis. Researchers have found this algorithm very interesting since it was first introduced. Its impact on association rules has been significant, as well as on the advancement of data mining.

The algorithm's fundamental principle is to iteratively combine current frequent itemsets to create new candidate itemsets, then prune any candidates that do not satisfy the minimal support criterion. To find frequent itemsets, the algorithm employs a recursive approach. First, a threshold for minimum support and confidence is defined. Finding all itemsets with support and confidence larger than or equal to the minimum threshold is the next step of the algorithm. These itemsets are called frequent itemsets. Next, the algorithm creates new candidate itemsets by merging previously discovered frequent itemsets. The new candidates are then tested to see if they meet the minimum support threshold. If so, they are included in the group of frequent itemsets, and the process is then repeated. If not, they are thrown away. Until no more frequent itemsets can be formed, this process is repeated (Zeng and Jia, 2022).

The steps of the Apriori algorithm can be summarized as the following:

- Establish a minimum support threshold, then search the database for all itemsets that satisfy it. These are the first frequent itemsets generated by the algorithm.
- By merging current frequent itemsets, create new candidate itemsets. Then, evaluate these candidates to see if they fulfill the required level of support.
- Repeat the previous step until no new frequent itemsets can be generated. The final output is the set of all frequent itemsets.

Below the pseudocode of the simple Apriori algorithm is indicated (Tang et al., 2013):

input:

```
    T: the set of transactions
    min_support: minimum support threshold
```

output:

```
    set of frequent itemsets
```

pseudocode:

```
step 1. unique_items = find_unique_items(T)
step 2. item_combinations = create_item_combinations(unique_items)
step 3. frequent_itemsets = find_frequent_itemsets(item_combinations, T, min_support)
return frequent_itemsets
```

Pseudocode of find_unique_items function used in the algorithm:

```
find_unique_items($T$):
    unique_items = set()
    for $t$ in $T$:
        for item in $t$:
            unique_items.add(item)
return unique_items
```

Pseudocode of create_items_combination function:

```
create_item_combinations(items):
    item_combinations = [ ]
    for $i$ in range(1, len(items)):
        item_combinations.extend(combinations(items, i))
    return item_combinations
```

Pseudocode of find_frequent_itemsets function:

```
find_frequent_itemsets(item_combinations, $T$, min_support):
    frequent_itemsets = {}
    for itemset in item_combinations:
        support = calculate_support(itemset, $T$)
        if support >= min_support:
            frequent_itemsets[itemset] = support
    return frequent_itemsets
```

Pseudocode of calculate_support function:

```
calculate_support(itemset, T):
    support = 0
    for $t$ in $T$:
        if set(itemset).issubset(set($t$)):
```

```
        support += 1
    return support
```

Since the Apriori algorithm should scan the whole database to generate the frequent itemset, if the dataset and the dimension for the candidate item are large, too many itemsets will be generated, which would decrease the algorithm's efficiency. That is the reason for making many improvements to the algorithm. We used Apriori to find the popular items in each cluster of customers. Since our clusters are not very large and the number of items is limited, we use the basic Apriori algorithm for mining the frequent rules.

4.4 Describing the Model

We utilize the described methods for implementing two frameworks illustrated in Figure 4.1. The first framework is generating a recommendation system by using consumer behavior analysis. We use one-year selling data of the online marketplace 'Baskasindaarama.com'. First, we run some preprocessing on the data which is described in section 3. Next, the Recency, Frequency, and Monetary scores of the RFM method are calculated for every 841 customers in the data. The output of the model is a table with 3 columns; each indicating these 3 scores for the customers represented in each row. Since there are some outliers in two columns, we standardize the data to guarantee the quality of the results. In the fourth step, we use these variables as the three dimensions of the K-means clustering algorithm. This method is used to place the most similar customers according to their behavior in the same segments. This algorithm benefits some properties like the ease of implementation and independency of initial values. Elbow method and Silhouette analysis are two utilized methods to determine the optimal number of clusters. These steps result in understanding the number of customers in each cluster, their specific characteristics, their transactions information, the most bought items in each cluster, and the popular categories of the products among them. In the next step, Apriori, an association rule mining algorithm, is utilized to find out about frequent items that were purchased together in the segments. With the help of these extracted features, the website can customize the recommendations for each customer. This method would help the business to increase its revenue. The results of this analysis are presented in chapter 5. The code for this framework was created using Google Collaboratory and the Python 3.8.16 libraries. Specifically, to create the models and use special functions the Pandas, Matplotlib, ScikitLearn, and Mlxtend libraries are used.

The second framework considers three pages of the company's website, each related to one category of the products, and next, the current two-way lift values of all the products on that page are extracted. Next, the sum of lifts is calculated to find the total lift. Moving forward, the Excel Solver is used to maximize this value by changing the place of the products. This framework is implemented in Microsoft Excel by utilizing VLOOKUP, INDEX functions, Solver, and What-If Analysis.

The next section is dedicated to representing and analyzing the results of the frameworks and acquiring managerial insights.

5. ANALYSIS AND RESULTS

In this chapter, we discuss in detail the employed frameworks and the outcome of methods used in different steps on the data. In addition, we describe the constructed customer segments and the features of each. Lastly, we present the results of each framework.

5.1 Recommendation System Strategies

The first framework is implementing a methodology to improve the retailer's recommendation system. After making some changes to the data in the preprocessing part, RFM was the first method that was used to help in customer segmentation. The output of RFM is three scores for each customer. The Recency score aims to find out if a customer recently had a purchase from the website. Therefore, we calculated the number of days between the last day in the data, which was 2020/12/31, and the date of the customer's last purchase. To find the Frequency score, we counted the number of transactions of each consumer. Lastly, we calculated the Monetary score by summing up the turnover values of each customer's transactions in the data. We entered the results for 841 customers into a Python dataframe to use in the later processes. The results for the first five customers in the data are shown in Table A.7.

The recency variable's mean is 141.94 and the standard deviation is around 100. The minimum is 0 and the maximum is 365, showing that at least one customer purchased only once on the first day of the year and for some customers, the last purchase was on the last day of the year.

For the frequency variable, the mean is 1.5 and the standard deviation is 0.92 which means the number of purchases for most of the customers is only once. The minimum is 1 and the maximum is 6 showing the highest number of transactions for a customer in the data.

The mean for the monetary variable is 994 with 10 for the minimum and 12589 for the maximum and the standard deviation is 1890 showing the large variation for the feature.

We plotted the three scores and the results can be seen in Figure 5.1. The plots show that for most of the customers, the recency is between 0 to 200 days. More than 500 customers purchase only once from the website, and for a high portion of customers, the frequency is between 0 and 1000 Turkish Liras.

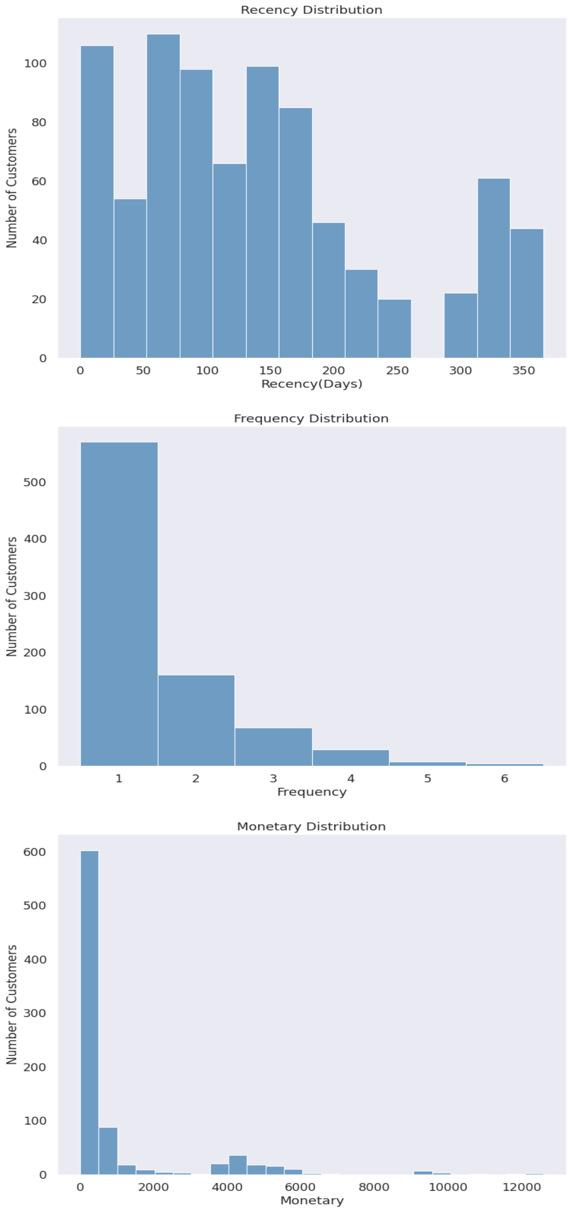

Figure 5.1 RFM Variables Histograms

Figure 5.2 shows the boxplots of the variables. We can acquire some information from these plots as well. The minimum value for the recency is 0 and the maximum is 365. There are no outliers in the plot and the interquartile is between 60 and 190 and its range is 130. For the frequency variable, the minimum and the first quartile are the same which shows a significant portion of the data is equal to 1. Moreover, there are some outliers in the variable. For the monetary, there are a great number of outliers and most of the data is less than 2000.

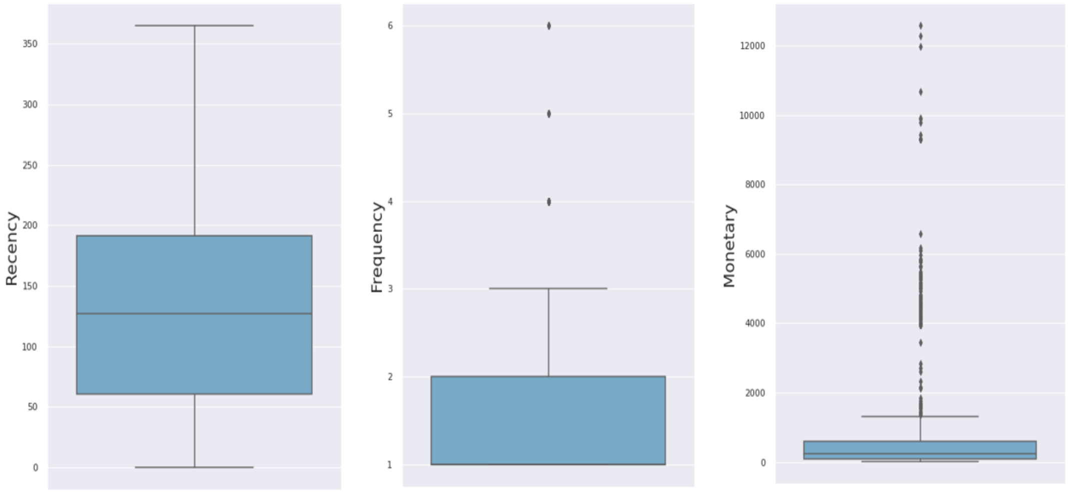

Figure 5.2 RFM Variables Boxplots

The natures of the obtained RFM scores are different and their range varies a great deal. Moreover, there are many outliers in frequency and monetary scores. Therefore, standardization of the data is a good way to acquire reliable results in the following moves.

Data standardization which is changing the data points of a variable in a way that the mean becomes 0 and the standard deviation changes to 1, is a crucial operation that has a great impact on the algorithms' performance before implementing them on the data. Variables must be standardized to have an equal variance to prevent clusters from being dominated by variables with the highest levels of variation. Standardization is also very common in marketing research because of the different nature of the variables in this area. For instance, in combining an income variable with the age variable for clustering without standardization, the income variable likely dominates the other because of its higher range. Therefore, to become sure of the equal contribution of the variables, we need to use standardization (Su et al., 2009). To obtain the z-score of each data point, Webber et al. (2008) propose 5.1 Equation where μ is the mean and σ is the standard deviation of the data points.

$$z_i = \frac{x_i - \bar{\mu}}{\sigma} \tag{5.1}$$

We implemented the z-score standardization on our data and added the results of this step as three new columns to the previous dataframe for RFM values. The first rows of this dataframe are shown in Table A.8. We can use the standardized values to implement the clustering algorithm.

Two methods were utilized to decide on the final number of clusters before implementing the K-means clustering algorithm. Figure 5.3 represents the results for the Elbow method. As discussed in section 4.2.1, the elbow method is calculating the total within clusters sum of squares for different numbers of Ks and where the considerable decrease happens, we can choose it as the best number of clusters. The distortion score in the plot indicates the sum of the squared distance from the points of each cluster to the center of the cluster. The fit time exhibits the amount of time to train the algorithm for that number of clusters. According to this plot, the best number of clusters for the K-means algorithm is 4 which resembles the elbow.

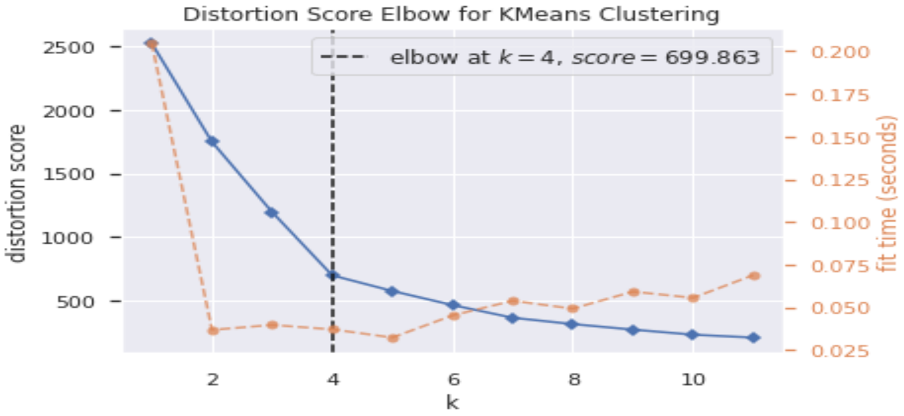

Figure 5.3 Elbow Method Plot for the Optimal Number of Clusters

The second utilized method is Silhouette analysis. Silhouette index was calculated for different Ks between 2 and 6. We mentioned that it shows how well the data points are clustered by measuring their similarity to the assigned cluster. The higher index represents better clustering. The results are represented in Table 5.1. According to this table, K=4 has the highest score and the value of the optimal number of clusters is the same as with the Elbow method.

Table 5.1 Silhouette Index Values for Different Number of Clusters

K-value	Silhouette index
2	0.44106
3	0.41323
4	0.51338
5	0.47441
6	0.487912

Added to the silhouette index, we also based our selection on the Silhouette plots analysis. In Figure 5.4 we present the results for K=2 and K=3. We see a huge fluctuation in the size of the clusters. Moreover, we notice some data points with negative Silhouette scores in clusters 0 and 1 which infers that there are certain customers that are wrongly clustered. The plots of K=4 and K=5 are shown in Figure 5.5. The variation in these clusters is not much and there are no negative Silhouette scores. However, because the dataset is small and the number of customers is low, it is better to keep the number of clusters small. In the plot for K=6 in Figure 5.6, we notice the Silhouette coefficients are relatively low and there are two very thin clusters. Finally, by considering the Elbow method, Silhouette indexes and plots as well as the business itself, we decided to continue with 4 clusters.

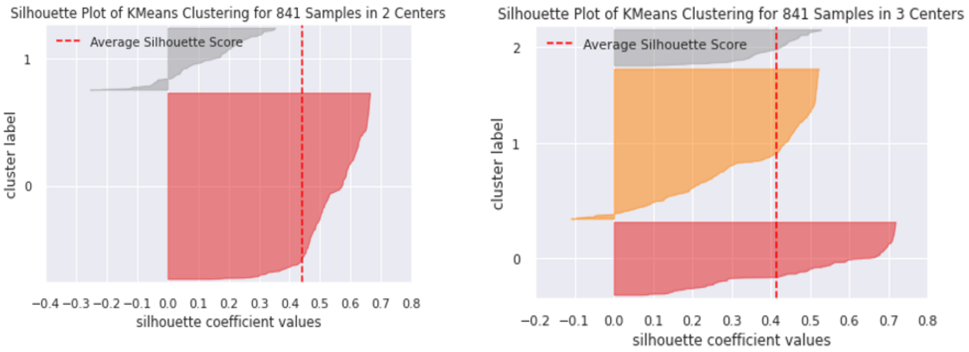

Figure 5.4 Silhouette Plots for K=2 and K=3

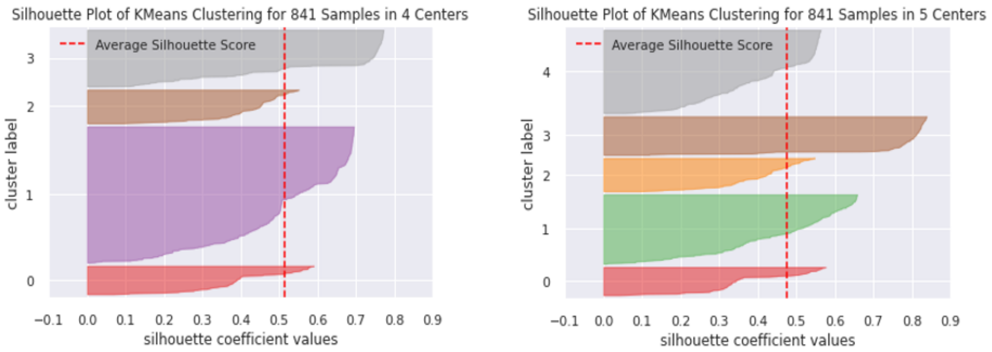

Figure 5.5 Silhouette Plots for K=4 and K=5

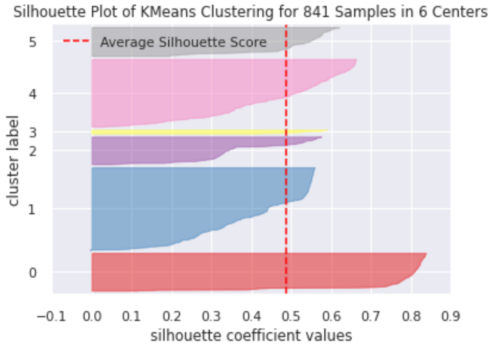

Figure 5.6 Silhouette Plot for K=6

The next part is dedicated to giving details about the 4 clusters created by the K-means clustering algorithm and their comparison in terms of different characteristics.

5.1.1 Summary of the Segments

We implemented the K-means clustering algorithm with K=4 number of clusters and the output of the algorithm is represented in Table 5.2. The smallest cluster has 94 customers and the size of the largest one is 449. Figure 5.7 also represents the number of customers in each segment.

Table 5.2 Size of Clusters

Cluster	Size
1	111
2	449
3	94
4	187

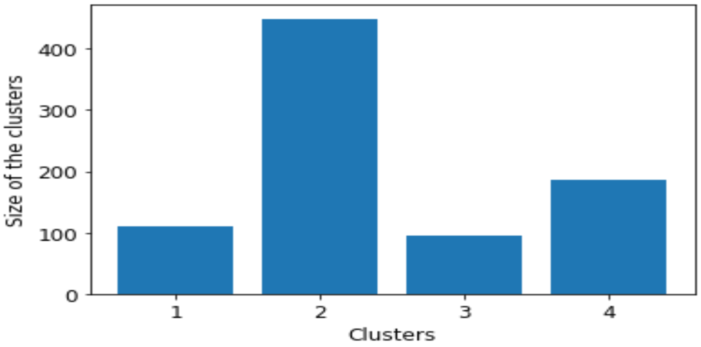

Figure 5.7 Size of Clusters

The good selection of initial centroids in K-means clustering can have a significant impact on the final clustering result. In some studies, researchers have used various methods to implement the K-means clustering algorithm with different initial centroids to obtain better results. In some other studies, like the ones we mentioned in the literature review, choosing centroids randomly is a common approach, as it is easy to implement and often results in good solutions. However, randomly choosing centroids can also lead to poor solutions since it is possible to converge to a local optimum rather than the global optimum, particularly if the data has a non-uniform distribution or if the number of clusters is not well-defined. One way of preventing the algorithm to trap in the local optimum in this situation is to run it several times to increase the chance of finding the good initialization and keep the results with the lowest sum of squared errors (Fränti and Sieranoja, 2019).

One alternative to randomly choosing centroids is to use a technique called "k-means++" which can produce better initial centroids. The k-means++ algorithm selects the initial centroids such that they are more likely to be distant from each other, which can help ensure that the final clusters are more distinct (Arthur and Vassilvitskii, 2006).

Another alternative is to use some other clustering algorithm as a preprocessing step to initialize the centroids. For instance, using hierarchical clustering to obtain a dendrogram and then cutting it at a certain level to get the number of clusters and then using the resulting clusters as initial centroids for K-means (Kanungo et al., 2002).

In general, the best approach will depend on the specific characteristics of the data and the goals of the analysis. In this research, we have employed the approach of running the algorithm with random initial centroids but for several times which means while implementing the K-means algorithm on the results of RFM on our dataset, we did not use any special methods to select the initial centroids of the clusters and algorithm chose them randomly. However, as a result of running the algorithm for several times to check the results, we noticed that the results were the same in all the replications and decided to continue with these results. One of the possible reasons for obtaining the same outcome in all the runs might be the sparsity of the data and the inherent difference of the data of each cluster with other clusters which every time leads to the same clusters that are at maximum distance from each other.

In the next step, to find out more about the characteristics of the segments, we analyzed the values of RFM scores for each cluster according to the snake plot shown in Figure 5.8 which gives a summary of the attributes. The X-axis shows the standardized RFM metrics and Y-axis represents the mean value of each metric for the 4 created clusters. The average values of the scores for the clusters are also shown in Table 5.3.

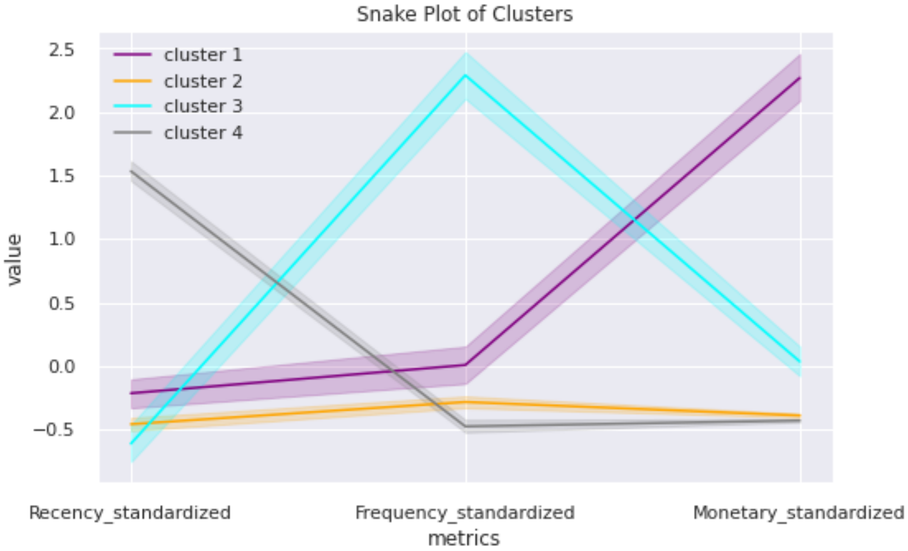

Figure 5.8 Standardized RFM Values of the Clusters

Table 5.3 Average of Standardized RFM Values of the Clusters

Cluster	Recency_standardized	Frequency_standardized	Monetary_standardized
1	-0.214	0.007	2.268
2	-0.458	-0.283	-0.389
3	-0.611	2.290	0.036
4	1.534	-0.475	-0.429

To support the claims about the differences between the recency, frequency, and monetary values of the clusters, we need to know if they are statistically different from each other or not. In the first step to determine the appropriate statistical test, we assessed the normality distribution of each value using the Kormogorov-Smirnov statistical test.

The Kolmogorov-Smirnov test (K-S test) is a nonparametric statistical test that compares a sample cumulative distribution function (CDF) to a reference probability distribution in order to test whether two samples have similar distributions. For testing the normality of a dataset, it compares the sample cumulative distribution function to the theoretical cumulative distribution function of the normal distribution. To determine whether to reject the null hypothesis which is the sample comes from a normal distribution, the test statistic (D) is calculated as the maximum difference between the two cumulative distribution functions, and then compared to a critical value from the reference table. Unless the test statistic (D) exceeds the critical value, we fail to reject the null hypothesis, indicating that there is insufficient evidence to suggest that it is not normal. In other words, the K-S test for normality compares the sample data to the normal distribution and measures how well the sample data fit the normal distribution. We can also examine the hypotheses by checking the p-values of the test. If the p-value is less than the significant threshold, we have enough evidence to reject the null hypothesis that the data comes from a normal distribution. Otherwise, we fail to reject H_0 (Wasserman, 2006).

We define the hypotheses for the recency values of the groups as below:

$H_0 =$ The distribution of the recency values of the cluster comes from a normal distribution.

$H_1 =$ The distribution of the recency values of the cluster deviates significantly from a normal distribution.

We tested these hypotheses using the Kolmogorov-Smirnov function of the Python Scipy library at the significance level $\alpha = 0.05$. If the p-value for the test is less than α, we reject the null hypothesis.

Table 5.4 indicates that the p-value of the recency values for the four clusters are less than 5% which shows they are not coming from a normal distribution.

Table 5.4 Results of the K-S Test for the Recency Values of the Clusters

Cluster	Statistic	P-value
1	0.247	$1.713e^{-6}$
2	0.324	$1.058e^{-42}$
3	0.375	$2.150e^{-12}$
4	0.715	$3.304e^{-97}$

We repeated the test for the frequency values of the 4 clusters to check if they have a normal distribution and the results are shown in Table 5.5. According to the results, the P-value for all the clusters is less than 5%. Therefore, we have enough evidence to reject the null hypothesis and conclude the values are not coming from a normal distribution.

Table 5.5 Results of the K-S Test for the Frequency Values of the Clusters

Cluster	Statistic	P-value
1	0.309	$6.015e^{-10}$
2	0.452	$2.349e^{-84}$
3	0.945	$7.296e^{-119}$
4	0.640	$4.718e^{-75}$

The test is repeated for the monetary values and the results are indicated in Table 5.6. Since all the p-values are less than our significance level, we conclude the monetary values for none of the clusters are coming from a normal distribution.

Table 5.6 Results of the K-S Test for the Monetary Values of the Clusters

Cluster	Statistic	P-value
1	0.931	$5.405e^{-130}$
2	0.495	$1.756e^{-102}$
3	0.330	$1.195e^{-9}$
4	0.553	$2.440e^{-54}$

From the results of the three tables, we know that our recency, frequency, and monetary values for none of the clusters in the dataset are normally distributed. Therefore, to check if they are statistically different, we should use a nonparametric statistical test.

The Kruskal-Wallis is a nonparametric statistical test and is used to assess if three or more independent groups are statistically different. This test is an extension of the Mann-Whitney U test which examines the difference between two samples that may have non-equal sizes. It is the nonparametric version of the ANOVA test (Kruskal and Wallis, 1952). The hypotheses of the Kruskal-Wallis test for our problem can be written below:

H_0: The population means of the 4 clusters are equal.

H_1: At least one of the groups has a mean statistically different from other groups.

First, we tested if the recency values of the 4 clusters are significantly different from each other at the significance level $\alpha = 0.05$. Table 5.7 indicates the result of the Kruskal-Wallis test on Python for the recency values. Since the p-value of the test is less than our significance level, therefore, we have sufficient evidence to reject the null hypothesis and conclude at least one of the groups is significantly different from the others.

Table 5.7 Kruskal-Wallis Test Results for the Recency of the Clusters

Variables	Statistic	P-value
recency values of the clusters	446.154	$2.220e^{-96}$

The next step is to pair-wisely check whether the distribution of R, F, and M values among the 4 clusters are significantly different from each other or not. In order to check this, we used the Mann-Whitney U test. The Mann-Whitney U test also known as the Wilcoxon rank sum test is a nonparametric test to check if the difference between two groups that have no specific distribution is significant. The null hypothesis of the test is that the population medians of the two groups are equal, and it is tested against the alternative hypothesis that the population medians are not equal. This can also be used as an approximate test for difference in means if the two datasets are not normal and the sample sizes are not too large (Hollander et al., 2013; Mann and Whitney, 1947). This test is like a t-test which is a parametric test examining the two groups are from a single population but under the normality assumption. We can write our null and alternative hypotheses as below:

H_0: The population means of the two groups are equal.

H_1: The population means of the two groups are significantly different from each other.

We used this test to test the difference between the recency values of each two cluster. The results of the tests are indicated in Table 5.8. We notice all the p-values are less than the significance level $\alpha = 0.05$ and therefore, we reject the null hypotheses for all the tests and conclude the mean of the recency values of each two groups are significantly different from each other.

Table 5.8 Mann-Whitney U Test Results for the Recency of the Clusters

Clusters	Statistic	P-value
1-2	31446	$1.906e^{-5}$
1-3	7194	$3.006e^{-6}$
1-4	111	$3.048e^{-46}$
2-3	25580	0.001
2-4	46.500	$8.350e^{-88}$
3-4	386.500	$4.679e^{-39}$

We repeated the Kruskal-Wallis test for examining if the frequency values of all four clusters are equal or if at least one of them is significantly different from the others. The result of the test is designated in Table 5.9. Since the p-value is less than 5%, we conclude the mean values are not equal for the clusters.

Table 5.9 Kruskal-Wallis Test Results for the Frequency of the Clusters

Variables	Statistic	P-value
frequency of the Clusters	391.187	$1.794e^{-84}$

The next step is to check the frequency values of which clusters are significantly different from each other. Same to checking the recency of the clusters, we applied the Mann-Whitney U test on every two clusters separately. Table 5.10 represents the results of the six tests. We see that all the p-values are less than 5% and the clusters are significantly different in the mean of frequency values.

Table 5.10 Mann-Whitney U Test Results for Frequency of the Clusters

Clusters	Statistic	P-value
1-2	29299	0.000
1-3	364	$1.061e^{-32}$
1-4	49753.500	$1.619e^{-7}$
2-3	10070	$1.061e^{-32}$
2-4	17526	$6.291e^{-55}$
3-4	13862	$2.881e^{-12}$

The Kruskal-Wallis test was used for the third time to check if the monetary values of the 4 clusters obtained from the K-means clustering algorithm are all equal or if at least one of them is statistically different from others. According to the result of the test represented in Table 5.11, the p-value is less than the significance level and the clusters are not equal in terms of mean and median.

Table 5.11 Kruskal-Wallis Test Results for the Monetary of the Clusters

Variables	Statistic	P-value
monetary of the Clusters	419.594	$1.260e^{-90}$

Mann-Whitney U test was applied for checking if the monetary values for each of the two clusters are statistically different from each other. P-values of these tests represented in Table 5.12, show the mean of monetary values of all the clusters are significantly different from each other.

Table 5.12 Mann-Whitney U Test Results for Monetary Values

Clusters	Statistic	P-value
1-2	49839	$6.460e-60$
1-3	10026	$6.037e-30$
1-4	20757	$3.184e-47$
2-3	4201	$2.457e-34$
2-4	48838.500	0.001
3-4	20757	$3.184e-47$

After implementing the statistical tests on the results of the K-means clustering algorithm, we conclude that the recency, frequency, and monetary values of all four clusters are statistically different from each other. With these results, we can comment on the different attributes of each cluster and their differences. The four subsections in the following are dedicated to describing the distinct characteristics of the four clusters.

5.1.1.1 Prosperous Customers

Cluster 1 with 111 customers is the third cluster in terms of population. It has the third lowest recency and frequency values. However, the highest monetary value with a huge difference belongs to this segment. The lowest number of ordered items in this cluster is 1 and the highest one is 8. Moreover, the highest price for the ordered items is 9300 Turkish Liras. The highest turnover for a transaction is 10640 TL. The total sold items for this cluster is 324 from 39 categories and the ‘Barbecue’ category has the highest number of purchases. The most bought item is the ’Tashoven Pro 75 Stone Oven’. There is no specific pattern for purchases on different days of the month for this segment. The plot for the number of purchases from each of the 7 category classes is represented in Figure 5.9 and we can see most of the bought items are from the ’Life style’ category class. Because of the very high monetary value of the segment, we named the group as *‘Prosperous Customers’*.

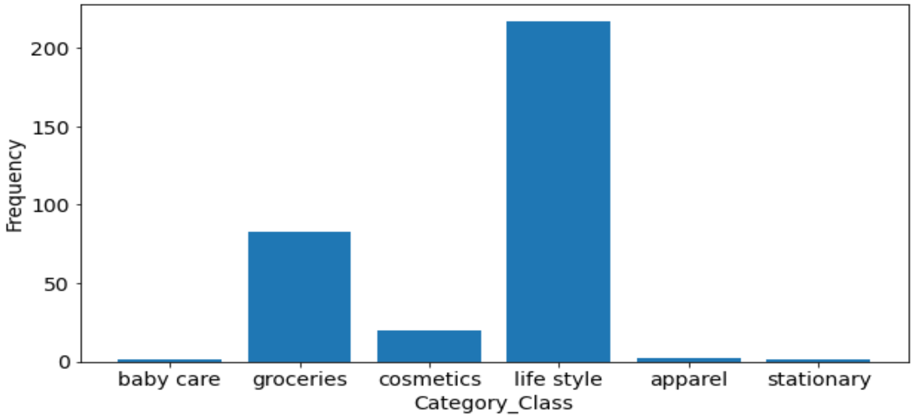

Figure 5.9 Category Class Purchase Frequency for Prosperous Customers

5.1.1.2 High-potential Customers

Cluster 2 has the highest number of customers equal to 449. It has the second lowest scores for all the attributes. 933 different items were bought from 89 categories with the lowest price of 3 TL and the highest price of 2140 TL. The mean of turnover value for this segment is 125 with a standard deviation equal to 205. Most of the purchases are from the ‘Flour’ category and the most bought items with ’Yayla 3 Pieces Meal Set 2’ with 37 purchases. A large number of purchases by this group were made in the first half of the month, which can be an indication that the customers of this group depend on their salaries received at the beginning of each month. The plot of the turnover for each month is shown in Figure 5.10. Since most of their purchases are related to the recent days in the data and they have reasonable values for frequency and monetary, we named the cluster as *‘High-potential Customers’*.

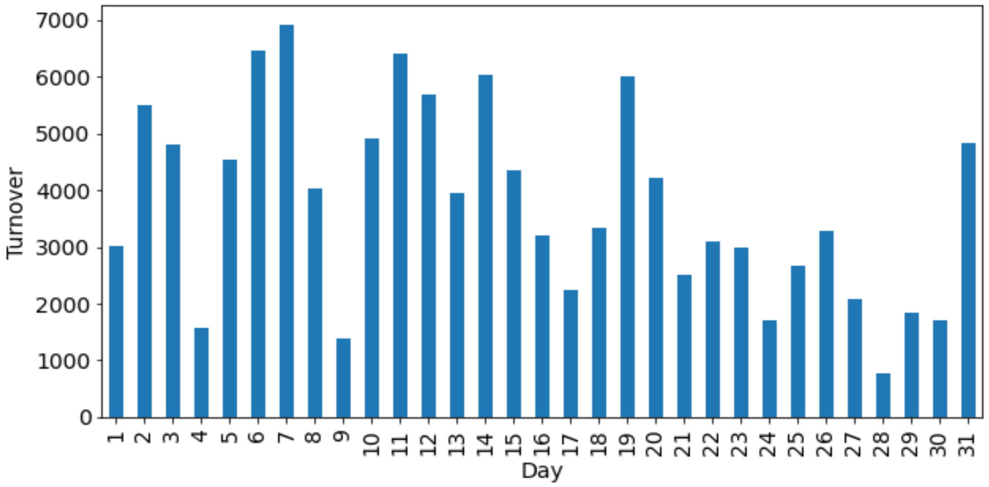

Figure 5.10 Turnover Values for each Month in the Second Cluster

5.1.1.3 Loyal Customers

According to figure 5.8, although cluster 3 has the least number of customers, it has the highest frequency value with a great difference from other segments. Moreover, they have the least recency value, indicating that they chose this marketplace as their recent trusted website. They also have the second higher monetary value. Because of these characteristics, we called the customers of this cluster as *'Loyal Customers'*. The number of customers in this segment is equal to 94. In total 1158 items were bought in 341 unique transactions. The maximum turnover of a transaction in this segment is 4200. Most of the transactions have been made in the first 10 days of the months. 'Tashoven Baking Stone' is the most bought item and most purchases have been made from the 'Flour' category.

5.1.1.4 Departed Customers

The fourth cluster has the lowest frequency and monetary values with the highest recency value which shows that they visited the website a long time ago and did not come back. We labeled this cluster as *'Departed Customers'*. They bought 747 items in total from 61 different categories. They bought only in the first 6 months of the year, and there are no purchases for the second half of the year. Figure 5.11

is the plot of turnover value for the first 6 months.

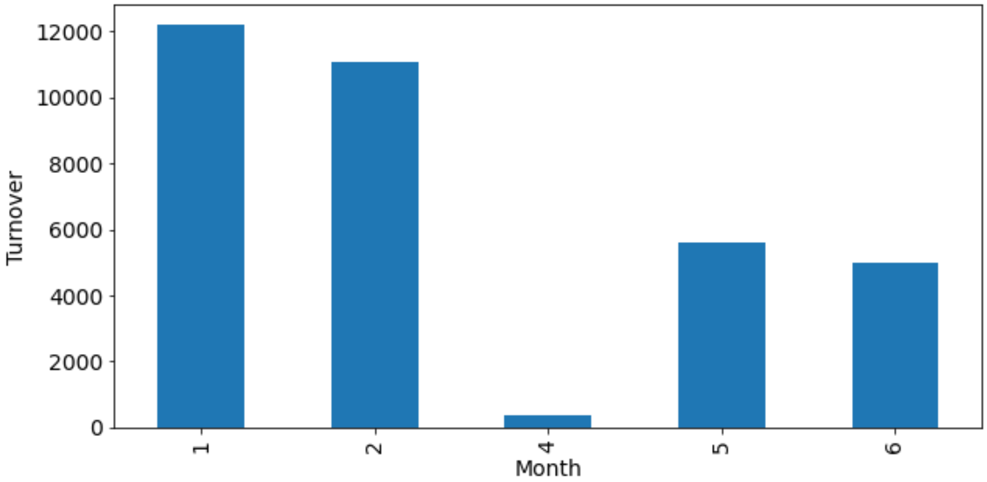

Figure 5.11 Turnover Values of each Month for the Departed Customers

We drew the 3D scatter plot of the customers according to their scores and the result is shown in figure 5.12. The *Prosperous Customers* are indicated with purple color, and we can see their higher monetary value in comparison to others. Their recency and frequency values are relatively low. The orange color is indicating the *High-potential Customers* with relatively small values for all three variables. *Loyal customers* can be seen with cyan dots with the highest frequency and low recency and the *Departed Customers* are shown in gray with the highest recency scores with a great distance to other dots.

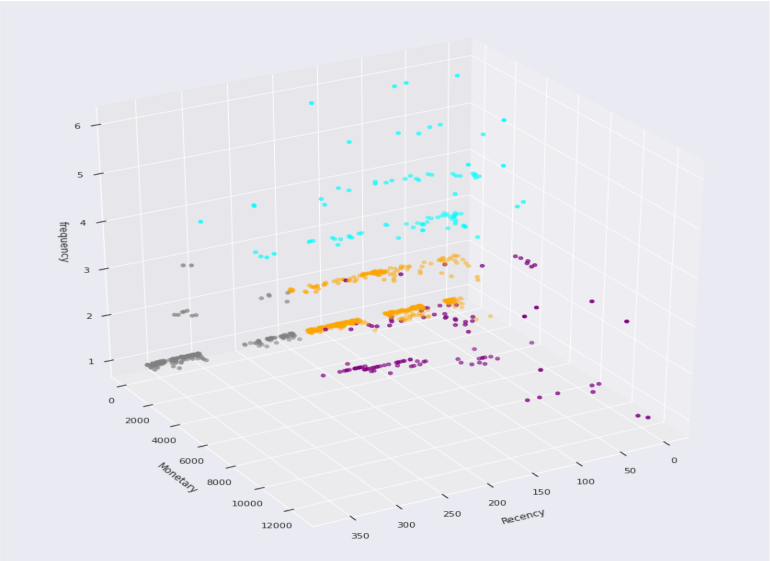

Figure 5.12 3D Scatter Plot of the Customers

5.1.2 Frequent Rule Mining of the Clusters

The next step after creating the clusters was discovering the most bought products and the frequent itemsets in each segment and utilizing them to improve the recommendation system of the website. Since the last purchase for all the customers of the *Departed* cluster was more than 6 months ago, we decided not to implement the Apriori algorithm on this segment. Because the products are changing more often, the history of the customers in this cluster will not be helpful. The best policy for this group is the website's current policy of recommending new and season-based products. For the other 3 segments, we utilized the Apriori algorithm to discover the association rules. First, we put together the items from the same transactions as one basket in the data. Next, after some attempts with different values for minimum support, we set it equal to 0.01 to exclude non-frequent itemsets.

The policy of the website is to recommend 10 new or season-based items at the bottom of a product that the visitor is checking. The number of rules generated for each cluster was relatively low because of the small number of transactions. Therefore, we included rules containing only one item with a high support value as well to recommend to the customers. These items are equal to the most repeated products in the transactions. The results for the *'Prosperous Customers'* are shown in Table 5.13. The 10 recommended products extracted by the algorithm are: Tashoven Pizza Shovel Large, Tashoven Pro 75 Stone Oven, Tashoven Protection Case, Tashoven barbecue, Tashoven Pizza Board, Tashoven Pro 100, Tashoven Baking Stone, Gourmezz Caramelized Onions, Tashoven Pro100 Protection Case, Buckwheat (Grechka) Flour.

The frequent itemsets of the *'High-potential Customers'* are shown in Table 5.14. The 10 results for this segment are: Gourmezz Roasted Hot Pepper Sauce, Gourmezz Spicy and Balsamic Tomato Sauce, Professional Shaving Bowl, Gourmezz Roasted Hot Pepper Sauce - Very Hot, Razor blade, Gourmezz Caramelized Onions, Gourmezz Pickled Red Onions, Yayla 3 Pieces Meal Set 2, Yayla 3 Pieces Meal Set 1, Tashoven Baking Stone.

The results for the third cluster, *'Loyal Customers'* according to Table 5.15 are as follows: Gourmezz Roasted Hot Pepper Sauce - Normal Hot, Gourmezz Spicy and Balsamic Tomato Sauce, Gourmezz Pickled Red Onions, Gourmezz Roasted Hot Pepper Sauce - Very Hot, Fast Acting Sebum Roll On For Acne Prone Skin, Clay Mask with Bentonite and Activated Charcoal for Acne Prone Skin, Tea Tree Oil Daily Balm For Acne Prone Skin, Rosehip Fruit Anti-Blemish Natural Spray Tonic, Rosehip Anti-Blemish Eye Contour Serum, Organic Wheat Flour.

Table 5.13 Frequent Rules for the Prosperous Customers' Cluster

Antecedents	**Consequents**	**Support**	**Confidence**	**Lift**
Tashoven Pizza Shovel Large	Tashoven Pro 75 Stone Oven	0.041	0.778	1.296
Tashoven Protection Case	Tashoven Pro 75 Stone Oven	0.200	0.919	1.532
Tashoven Pro 75 Stone Oven, Tashoven barbecue	Tashoven Protection Case	0.094	0.640	2.941
Tashoven barbecue	Tashoven Protection Case	0.100	0.586	2.693
Tashoven Pizza Board, Tashoven Pro 75 Stone Oven	Tashoven Protection Case	0.029	0.833	3.829
Tashoven Pro 100	-	0.050	-	-
Tashoven Baking Stone	-	0.050	-	-
Gourmezz Caramelized Onions-212 gr	-	0.030	-	-
Tashoven Pro100 Protection Case	-	0.030	-	-
Buckwheat (Grechka) Flour	-	0.030	-	-

Table 5.14 Frequent Rules for the High-potential Customers' Cluster

Antecedent	Consequents	Support	Confidence	Lift
Gourmezz Roasted Hot Pepper Sauce, Gourmezz Spicy and Balsamic Tomato Sauce,	Gourmezz Roasted Hot Pepper Sauce - Very Hot	0.020	0.875	21.375
Professional Shaving Bowl	Razor blade	0.011	1	85.500
Gourmezz Caramelized Onions-212 gr, Gourmezz Spicy and Balsamic Tomato Sauce	Gourmezz Roasted Hot Pepper Sauce - Very Hot	0.011	0.800	19.542
Gourmezz Pickled Red Onions	Gourmezz Roasted Hot Pepper Sauce - Very Hot	0.011	0.666	16.285
Yayla 3 Piece Meal Set 2	-	0.090	-	-
Yayla 3 Piece Meal Set 1	-	0.066	-	-
Tashoven Baking Stone				

Table 5.15 Frequent Rules for the Loyal Customers' Cluster

antecedents	consequents	support	confidence	lift
Gourmezz Roasted Hot Pepper Sauce - Normal	Gourmezz Spicy and Balsamic Tomato Sauce	0.039	0.800	10.150
Gourmezz Pickled Red Onions	Gourmezz Spicy and Balsamic Tomato Sauce	0.039	0.800	10.150
Gourmezz Roasted Hot Pepper Sauce - Very Hot	Gourmezz Spicy and Balsamic Tomato Sauce	0.029	0.666	8.458
Fast Acting Sebum Roll On For Acne Prone Skin	Clay Mask with Bentonite and Activated Charcoal for Acne Prone Skin	0.014	1	67.666
Tea Tree Oil Daily Balm For Acne Prone Skin	Clay Mask with Bentonite and Activated Charcoal for Acne Prone Skin	0.014	1	67.666
Rosehip Fruit Anti-Blemish Natural Spray Tonic	Rosehip Anti-Blemish Eye Contour Serum	0.014	1	67.666
Organic Wheat Flour	-	0.020	-	-

5.2 Improved Web Layout Framework

In the second framework, we used the association rules lift metric to make changes to the website layout. We explained in section 4 that lift is a measure of the strength of a rule and a lift value greater than 1 indicates a positive correlation between items and a greater lift is indicating a stronger rule. Therefore, by placing items with high lift values next to each other, we can increase the probability of buying the items together and increase the revenue.

This website has different pages each representing the products of different categories. Currently, the products are assigned randomly to different places on each page and the products of each specific brand are next to each other. We chose 3 pages that have the highest number of products in the data. If a product was not found in the data, we substituted it with another similar product. First, we calculated the lift values of each two products on each page, called two-way lifts, by using the What-If Analysis in Microsoft Excel. Next, considering the current layouts, we added the lift values for adjacent products by utilizing the Index function and next calculated the sum of all the results. Table 5.16 indicates the 21 products each represented by numbers from 1 to 21 in one of the webpages named 'peanut and hazelnut butter'. The initial result for the lift values of this page is equal to 89.91. In the next step, by using Solver, we tried to maximize the current lift value by reassigning the products to new places. After 20 minutes of running the defined model, the maximized value was equal to 264.42. The final improved layout is also shown in Table 5.17.

Table 5.16 Initial Layout of the Webpage

	1	**2**	**3**	**4**
A	10	12	14	8
B	4	2	6	5
C	3	11	7	9
D	18	13	19	20
E	1	21	16	15
F	17			

Table 5.17 Optimized Layout of the Webpage

	1	**2**	**3**	**4**
A	1	17	8	9
B	14	13	15	16
C	6	21	18	4
D	12	7	11	5
E	20	19	3	10
F	2			

This method was implemented on two other pages as well. The first one is the 'Healthy meals' page with 13 products. The initial layout of the webpage is shown in Table A.9 and the final result is shown in Table A.10. Moreover, Table A.11 indicates the initial layout of the 'Legume' page with 11 products, and Table A.12 is the improved layout.

The results for lift values of these three pages are represented in Table 5.18. We can conclude that by implementing this simple method on each page, we would be able to increase the probability of buying more products and increase revenue.

Table 5.18 Lift Values for Current and Optimized Website Layout

Current Lift	**Lift for Optimized layout**	**Percentage Change**
138.286	174.428	26%
35.500	55.000	54%
89.910	264.427	194%

6. DISCUSSION AND CONCLUSION

Consumer behavior analysis is essential in all types of markets to develop strategies in different segments like supply chain and CRM. This analysis is also important for online retailers to implement special policies on their websites to increase their income. In this study, two frameworks were developed to help increase the selling probability and the revenue of an online retailer in Turkey by using consumer behavior analysis, machine learning, and data mining algorithms.

The approach of the first framework was to generate a recommendation system for the website by considering the variety of behaviors of customers in previous purchases. We extracted the three behavioral features of recency, frequency, and monetary of the 841 customers with the RFM method and utilized these scores in the K-means clustering algorithm. The output of the algorithm is such that 13% are in the *Prosperous Customers* group, 53% are named *High-potential Customers* and 11% are in the *Loyal Customers* segment. The 22% of customers who did not buy anything during the last 6 months of the dataset and had the highest recency scores, were considered as *Departed Customers*. For the Departed group, the best recommendation strategy remains the website's current strategy, exposing them to products that are either new or related to the season. Since the products in today's market change very quickly, their purchasing history will not be useful. To utilize the results of the customer segmentation for the other 3 groups, we implemented the Apriori algorithm to find the frequent rules and extracted 10 products out of these rules with a high probability of purchasing to display to the customers.

Most of the products extracted from the frequent rules for the *Prosperous Customers* are from the expensive products on the website. Moreover, a great number of extracted products for *High-potential Customers* are low-price products. Since the recency and frequency scores of this group are low and customers are relatively new to the website, it seems they start with low-price products in their first purchases to become familiar with the website and see if they can trust it for their future and more expensive purchases. For *Loyal Customers*, we notice some products are from the BA market and related to food and kitchen and some are skin care products.

One reason for these products being in this group might be that the customers have bought the skin care products once and then were assured of the authenticity of the products and repeated their purchases. Another reason might be the advertisement of the website on social media or by the favorite and trusted influencers of the customers which made them trust the website and replicate their purchases.

The second framework sought to find an improved layout for the pages of the website by utilizing the lift metric of association rule mining. We demonstrated that modifying the current layout by attempting to maximize the total lift values by placing the products with greater two-way lifts near each other can increase the likelihood of selling more to customers. The increase of the lift values for the 3 improved layouts for the webpages were 26%, 54%, and 194%.

6.1 Theoretical Implications

From a theoretical point of view, the potential benefits and advantages of employing RFM, customer clustering with K-means, and association rule-based recommendation systems all together can be summarized as improved segmentation, personalized recommendations, accurate targeting, improved understanding of customer behavior, and cost-effective multi-perspective analysis.

While RFM analysis can be used to segment customers based on their past buying behavior, and customer clustering with K-means can be used to group similar customers together based on their characteristics or demographics, utilizing these methods together can provide a more detailed and accurate picture of a company's customers, which can be used to target marketing campaigns and tailor communication to specific groups of customers. Utilizing association rules-based recommendation systems in conjunction with RFM and K-means clustering can improve personalized product recommendations to customers based on their previous purchase behavior in groups with similar customers.

Integrated models lead to cost-effective systems with improved automation and multi-perspective analysis. This would be the case for the integrated framework proposed in this research.

6.2 Managerial Implications

By employing association rules, companies can identify patterns in customer behavior, which can be used to predict future behavior and make recommendations accordingly.

By using multiple methods such as RFM and K-means, companies can gain a better understanding of customer behavior and preferences, which can be used to make more informed business decisions. Furthermore, by combining the information obtained from RFM and K-means, companies can better target the most valuable customers with personalized offers and promotions.

Companies, marketers, and online retailers can utilize association rules based recommendation systems on each cluster of customers extracted through K-means clustering based on RFM methods. This framework can help marketers and online retailers to improve personalized recommendations and as a result, increase customer engagement and sales by providing them with products they are more likely to be interested in.

By identifying the most valuable customers, companies can optimize their resources and marketing efforts, reducing costs and increasing the return on investment. Moreover, by identifying customers at risk of defection using RFM analysis and targeting them with personalized offers and recommendations, companies can improve customer retention and increase long-term revenue.

By integrating all the methods into one system, companies can automate the process of segmentation, clustering, and recommendations, which would save time and improve efficiency. Moreover, utilizing multiple methods allows for analyzing customer data from multiple perspectives, which can provide a more complete understanding of the customers.

6.3 Limitations and Future Research

The data of the customers' previous transactions is a great help in implementing different strategies, and in our case, in developing two functional methodologies.

However, there were some limitations while implementing the project. First, the data relates to the 12 months of 2020, which was the very beginning of the Covid-19 pandemic, and similar to many businesses worldwide the mentioned retailer was also affected. This caused a reduction in sales and therefore, a loss of valuable information.

Second, we know that the performance of recommendation systems is based on previous transactions and that bigger datasets result in more reliable outcomes. However, our dataset was very small with a low number of transactions and customers. Therefore, to extract more reliable results, using the data from more years would be useful. Additionally, with more data, we would be able to analyze the behavior from different approaches like considering the seasonality.

Another important issue was the limitation in the number of variables. Unfortunately, we did not have access to variables such as age, gender, or the location of customers to conduct further analyses or analyses with different approaches.

One of the limitations of the RFM method is that it does not consider the product breadth, which means the variety of products that a customer has bought and it just considers the frequency of purchases in total. To overcome this problem, perhaps the combination of RFM with other methods can be used to obtain more accurate results.

Another problem was that Apriori is an inferential algorithm rather than a predictive one. Therefore, without implementing the outcome of the methodology on the website, we are unable to examine the ratio of the impact on the revenue. Moreover, Apriori is an expensive and time-consuming algorithm. However, because our dataset was small, the effect of these issues was not very tangible.

Future research on this topic might focus on using other methods in the related area to boost sales and, consequently, customer loyalty. For instance, utilizing a bigger dataset helps to acquire more reliable results.

Additionally, other clustering methods like Hierarchical clustering or DBSCAN clustering could be implemented to compare the results with our findings.

We implemented the K-means clustering by letting the algorithm start with random initial centroids. This research can be repeated by utilizing some methods to select the initial centroids and obtain better results.

Other recommendation system methods like collaborative-based filtering or content-based filtering could be implemented with larger datasets and with a greater number of variables to provide further insights for the retailer.

Apriori algorithm can be implemented based on the product categories to find out which categories are bought mostly together in different clusters.

We hope that the results of our study provide avenues for future research in the area and also help marketers choose the appropriate course of action to increase sales.

BIBLIOGRAPHY

Abrardi, L., Cambini, C., & Rondi, L. (2022). Artificial intelligence, firms and consumer behavior: A survey. *Journal of Economic Surveys*, *36*(4), 969–991.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 acm sigmod international conference on management of data* (pp. 207–216).

Alfian, G., Ijaz, M. F., Syafrudin, M., Syaekhoni, M. A., Fitriyani, N. L., & Rhee, J. (2019). Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores. *Asia Pacific Journal of Marketing and Logistics.*

Almonte, L., Guerra, E., Cantador, I., & De Lara, J. (2022). Recommender systems in model-driven engineering. *Software and Systems Modeling*, *21*(1), 249–280.

Amazon, T. S. B. (2021). *A complete guide on the advantages of ecommerce to business.* Retrieved 2022-10-30, from `https://sell.amazon.in/seller-blog/advantages-of-ecommerce`

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding* (Tech. Rep.). Stanford.

Asllani, A., & Halstead, D. (2015). A multi-objective optimization approach using the rfm model in direct marketing. *Academy of Marketing Studies Journal*, *19*(2), 65.

Behera, G., & Nain, N. (2022). Trade-off between memory and model-based collaborative filtering recommender system. In *Proceedings of the international conference on paradigms of communication, computing and data sciences* (pp. 137–146).

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley & Sons.

Bholowalia, P., & Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, *105*(9).

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & management*, *39*(3), 211–225.

Chan, P. S., & Pollard, D. (2003). Succeeding in the dotcom economy: Challenges for brick & mortar companies. *International Journal of Management*, *20*(1), 11.

Chattopadhyay, M., Mitra, S. K., & Charan, P. (2022). Elucidating strategic patterns from target customers using multi-stage rfm analysis. *Journal of Global Scholars of Marketing Science*, 1–31.

Cheng, C.-H., & Chen, Y.-S. (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, *36*(3), 4176–4184.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). Rfm ranking–an effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, *33*(10), 1251–1257.

Cil, I. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling. *Expert Systems with Applications*, *39*(10), 8611–8625.

Das, P., & Singh, V. (2023). Knowing your customers using customer segmentation. In *Computational methods and data engineering* (pp. 437–451). Springer.

Deldjoo, Y., Schedl, M., Cremonesi, P., & Pasi, G. (2020). Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)*, *53*(5), 1–38.

Edirisinghe, G., & Munson, C. L. (n.d.). Strategic rearrangement of retail shelf space allocations: Using data insights to encourage impulse buying. *Available at SSRN 4087605*.

Ernawati, S. S. K. B., Fauziah Kasmin, H., & Purwanugraha, A. (2022). Geomarketing promotional target selection using modified rfm with spatial and temporal analysis: A case study. *Journal of System and Management Sciences*, *12*(3), 156–180.

Essayem, W., Bachtiar, F. A., & Priharsari, D. (2022). Customer clustering based on rfm features using k-means algorithm. In *2022 ieee international conference on cybernetics and computational intelligence (cyberneticscom)* (pp. 23–27).

Foxall, G. R. (2001). Foundations of consumer behaviour analysis. *Marketing theory*, *1*(2), 165–199.

Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, *93*, 95–112.

G Martín, A., Fernández-Isabel, A., Martín de Diego, I., & Beltrán, M. (2021). A survey for user behavior analysis based on machine learning techniques: current models and applications. *Applied Intelligence*, *51*(8), 6029–6055.

Halim, S., Octavia, T., & Alianto, C. (2019). Designing facility layout of an amusement arcade using market basket analysis. *Procedia Computer Science*, *161*, 623–629.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons.

Hughes, A. M. (1996). Boosting response with rfm. *Marketing Tools*, *3*(3), 4–10.

Hughes, A. M. (2005). *Strategic database marketing*. McGraw-Hill Pub. Co.

Jo-Ting, W., Shih-Yen, L., & Hsin-Hung, W. (2010). A review of the application of rfm model. *African Journal of Business Management*, *4*(19), 4199–4206.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881–892.

Kardes, F., Cronley, M., & Cline, T. (2014). *Consumer behavior*. Cengage Learning.

Kim, Y.-J., & Kim, H.-S. (2022). The impact of hotel customer experience on customer satisfaction through online reviews. *Sustainability*, *14*(2), 848.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71–82.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Kumar, S., & Balakrishnan, K. (2019). Development of a model recommender system for agriculture using apriori algorithm. In *Cognitive informatics and* *soft computing* (pp. 153–163). Springer.

Kywe, S. M., Lim, E.-P., & Zhu, F. (2019). A survey of recommender systems in twitter. In *International conference on social informatics* (pp. 420–433).

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets.* Cambridge university press.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th berkeley symp. math. statist. probability.*

Malloy. (2019). *Benefits of e-commerce for customers and businesses.* Retrieved 2022-10-30, from `https://www.cloudtalk.io/blog/benefits-of-e-commerce-for-customers-and-businesses/`

Manis, K., & Madhavaram, S. (2023). Ai-enabled marketing capabilities and the hierarchy of capabilities: Conceptualization, proposition development, and research avenues. *Journal of Business Research*, *157*, 113485.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Mehta, P., Dongare, O., Tekale, R., Umare, H., & Wanve, R. (2021). A survey on hybrid recommendation systems.

Montgomery, A. L., & Smith, M. D. (2009). Prospects for personalization on the internet. *Journal of Interactive Marketing*, *23*(2), 130–137.

Nazeer, K. A., & Sebastian, M. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceedings of the world congress on engineering* (Vol. 1, pp. 1–3).

Obeidat, R., Duwairi, R., & Al-Aiad, A. (2019). A collaborative recommendation system for online courses recommendations. In *2019 international conference on deep learning and machine learning in emerging applications (deep-ml)* (pp. 49–54).

Ozgormus, E., & Smith, A. E. (2020). A data-driven approach to grocery store block layout. *Computers & Industrial Engineering*, *139*, 105562.

Ray, S. (2019). A quick review of machine learning algorithms. In *2019 international conference on machine learning, big data, cloud and parallel computing (comitcon)* (pp. 35–39).

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook* (pp. 1–34). Springer.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Sagin, A. N., & Ayvaz, B. (2018). Determination of association rules with market basket analysis: application in the retail sector. *Southeast Europe Journal of Soft Computing*, *7*(1).

Sari, Y., & Helena, Y. (2021). Consumer behavior analysis on shopee's e-commerce purchase decisions during the covid-19 pandemic. *International Journal Administration Business & Organization*, *2*(2), 33–47.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1–21.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at amazon.com. *IEEE internet computing*, *21*(3), 12–18.

Solomon, M. R., White, K., Dahl, D. W., Zaichkowsky, J. L., & Polegato, R. (2017).

Consumer behavior: Buying, having, and being (Vol. 12). Pearson Boston, MA.

Štimac, H., Bilandžić, K., et al. (2021). How web shops impact consumer behavior? *Tehnički glasnik*, *15*(3), 350–356.

Stone, B. (2008). Successful direct marketing methods: interactive, database, and customer-based marketing for digital age.

Stormi, K., Lindholm, A., Laine, T., & Korhonen, T. (2020). Rfm customer analysis for product-oriented services and service business development: an interventionist case study of two machinery manufacturers. *Journal of Management and Governance*, *24*(3), 623–653.

Su, C., Zhan, J., & Sakurai, K. (2009). Importance of data standardization in privacy-preserving k-means clustering. In *International conference on database systems for advanced applications* (pp. 276–286).

Surendro, K., et al. (2019). Predictive analytics for predicting customer behavior. In *2019 international conference of artificial intelligence and information technology (icaiit)* (pp. 230–233).

Surjandari, I., & Seruni, A. C. (2005). Design of product placement layout in retail shop using market basket analysis. *Makara Journal of Technology*, *9*(2), 1.

Tang, J.-Y., Chuang, L.-Y., Hsi, E., Lin, Y.-D., Yang, C.-H., & Chang, H.-W. (2013). Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the apriori algorithm. *BioMed Research International*, *2013*.

Tseng, M.-C., & Lin, W.-Y. (2007). Efficient mining of generalized association rules with non-uniform minimum support. *Data & Knowledge Engineering*, *62*(1), 41–64.

Ünvan, Y. A. (2021). Market basket analysis with association rules. *Communications in Statistics-Theory and Methods*, *50*(7), 1615–1628.

Varzaneh, H. H., Neysiani, B. S., Ziafat, H., & Soltani, N. (2018). Recommendation systems based on association rule mining for a target object by evolutionary algorithms. *Emerging Science Journal*, *2*(2), 100–107.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Webber, W., Moffat, A., & Zobel, J. (2008). Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval* (pp. 51–58).

Yan, S.-R., Pirooznia, S., Heidari, A., Navimipour, N. J., & Unal, M. (2022). Implementation of a product-recommender system in an iot-based smart shopping using fuzzy logic and apriori algorithm. *IEEE Transactions on Engineering Management*.

Zeng, J., & Jia, B. (2022). Live multiattribute data mining and penalty decision-making in basketball games based on the apriori algorithm. *Applied Bionics and Biomechanics*, *2022*.

Zhang, Y., Chen, X., et al. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, *14*(1), 1–101.

Zhao, J., Xue, F., Khan, S., & Khatib, S. F. (2021). Consumer behaviour analysis for business development. *Aggression and Violent Behavior*, 101591.

Zhao, X., & Keikhosrokiani, P. (2022). Sales prediction and product recommendation model through user behavior analytics. *Computers, Materials, & Continua*, 3855–3874.

Zhou, Y. (2020). Design and implementation of book recommendation management system based on improved apriori algorithm. *Intelligent Information Management*, *12*(3), 75–87.

APPENDIX A

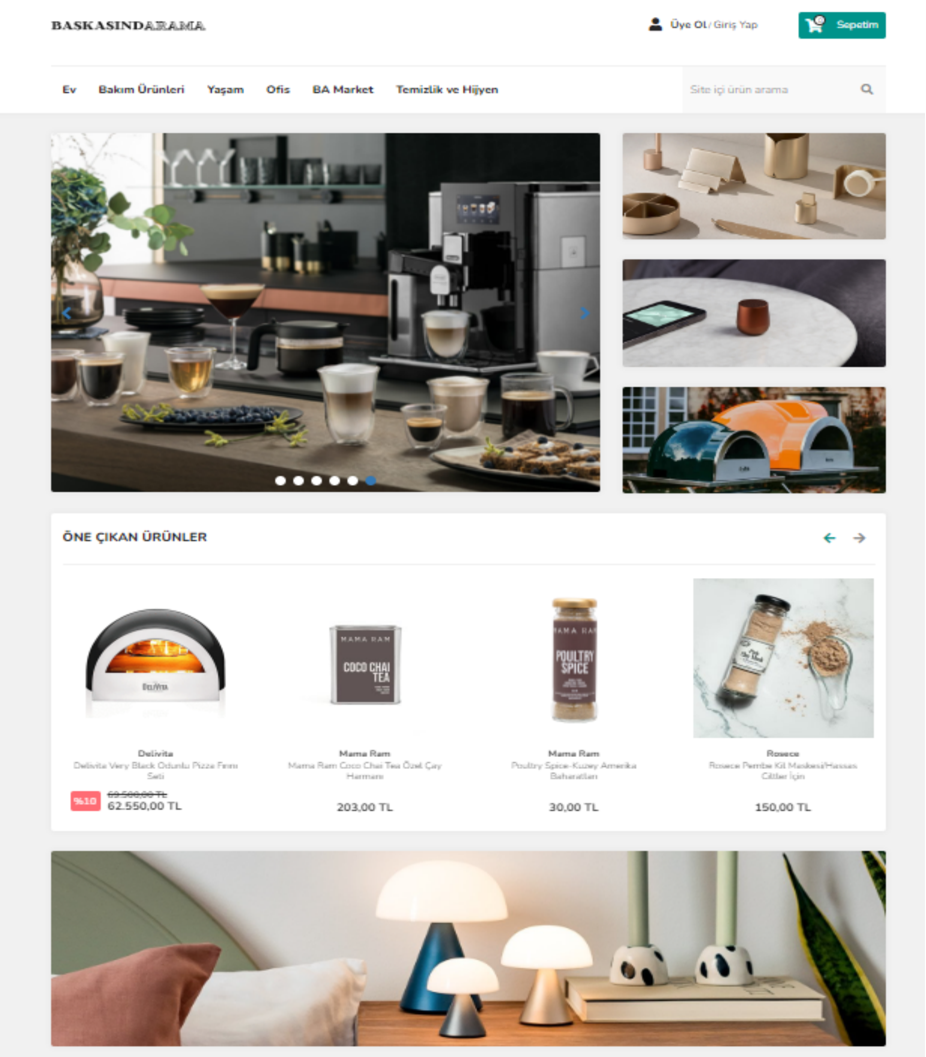

Figure A.1 Snapshot of the Retailer’s First Page

Table A.1 One Page of the Retailer's Daily Sales Data

			New Visitor	Returning Visitor			
			540	63			
Customer ID	Page Viewed	Chart Situation	Number of Order	Product Price	Turnover	Profit Margin	Special Campaigns
C62	http://www.baskasindaarama.com/urun/biyik-sekillendirici-wax	completed	1	59	59	20%-30%	-
C62	http://www.baskasindaarama.com/urun/vintage-biyik-fircasi-ve-taragi	completed	1	42	42	20%-30%	-
C63	http://www.baskasindaarama.com/urun/rosece-vucut-losyonu-yasemin-ve-sandal-agaci	completed	1	68.75	68.75	20%-30%	-
C64	http://www.baskasindaarama.com/urun/fabooks-perfect-is-boring-spiral-bloknot	completed	1	19	19	30%-40%	-
C64	http://www.baskasindaarama.com/urun/fabooks/you-are-super-duper-amazing-talented-spiralli-bloknot	completed	1	15	15	30%-40%	-
C64	http://www.baskasindaarama.com/urun/fabooks-i-love-london-mini-defter	completed	1	6.9	6.9	30%-40%	-
C65	http://www.baskasindaarama.com/urun/flip-alarm-saat	completed	1	265	265	20%-30%	-

Table A.2 One Page of the Retailer's Monthly Traffic Acquisition

	New Visitors	**Returning Visitors**
	17456	**5259**
Organic Search	61.40%	61.90%
Direct	27.20%	23.80%
Social	9.30%	7.20%
Referral	2.10%	7.10%

Table A.3 One Page of the Retailer's Sold Brands Data

Baskasindaarama.com						
Apparel	**Jewellery**	**Babycare**	**Cosmetics**	**Life style**	**Stationary**	**BA Market**
Aquella beachwear	Soul2Seven	Happy Folks	Bade Natural	Markaev	Fabooks	Yerlim Ciftlik
The Black Ears		Italtrike	Real Techniques	Sodalife	Lexon	Saf Nutrition
		My Konjac	Bold and Goodly	Tashoven		Melez Tea
		Melissa and Doug	Braun	Le Nouveau		Humm Organic
		Bade Natural	Eda Taspinar	Bestway		Gourmet Ladies
		Trim	Neutrogena	Bold and Goodly		Yayla
		Trunki	Glide'n Style			Guzel Gida

Table A.4 Daily Page Visits Reports

Total Number of Pages Viewed	Rate of Immediate Exit	Average Duration on One Page
5138	75.26%	0:02:20

Page Link	View Number
https://www.baskasindaarama.com/kategori/defter-ajanda	316
https://www.baskasindaarama.com/urun/market-alisveris-listem	745
https://www.baskasindaarama.com/kategori/ba-market?marka=gourmezz	526
https://www.baskasindaarama.com/kategori/kadin-bakim-urunleri	726
https://www.baskasindaarama.com/yeni-urunler	896
https://www.baskasindaarama.com/populer-urunler	755
https://www.baskasindaarama.com/kategori/kadin-bakim-urunleri?marka=bade-natural	354
https://www.baskasindaarama.com/kategori/makyaj-temizleme-urunleri	455
https://www.baskasindaarama.com/kategori/saglik-ve-hijyen-urunleri	365

Table A.5 Customers Demographic Report

1.Age	
Age	**Percentage(%)**
18-24	10.56%
25-34	52.11%
35-44	28.17%
45-54	9.15%
55-64	0.00%
65+	0.00%

2. Gender	
Female	61.80%
Male	38.20%

3. Location	
Turkey	82.09%
USA	7.93%
Germany	2.06%
Argentina	1.91%
Canada	1.76%
United Kingdom	0.73%
Austria	0.29%
China	0.29%
Cyprus	0.29%

Table A.6 First Five Rows of the Main Data Used in the Thesis

Row	Transaction	Customer	Customer_Code	Item	Category	Category_Class	Brand	Day	Month	Date	Situation	Number_Ordered	Price	Turnover	Profit_Margin	Special_Campaign	Url
0	20200101C9	C9	9	colombian filter coffee 10 pieces	coffee	BA MARKET	Filtr Café	1	1	2020-01-01	ORDER COMPLETED	1	35	35	%21-%30		https://www.baskasindarama.com/urun/colombian-filtre-kahve-10lu
1	20200101C8	C8	8	bade natural intense nourishing serum	skin care serums	cosmetics	Bade Natural	1	1	2020-01-01	ORDER COMPLETED	2	74,95	149,9	%31-%40	%50 DISCOUNT CAMPAIGN	https://www.baskasindarama.com/urun/bade-natural-yogun-besleyici-serum
2	20200101C7	C7	7	bade natural eyebrow and eyelash nourishing serum	skin care serums	cosmetics	Bade Natural	1	1	2020-01-01	ORDER COMPLETED	6	49,95	299,7	%31-%40	%50 DISCOUNT CAMPAIGN	https://www.baskasindarama.com/urun/bade-natural-kas-ve-kirpik-besleyici-serum
3	20200101C6	C6	6	bade natural eyebrow and eyelash nourishing serum	skin care serums	cosmetics	Bade Natural	1	1	2020-01-01	ORDER COMPLETED	2	49,95	99,9	%31-%40	%50 DISCOUNT CAMPAIGN	https://www.baskasindarama.com/urun/bade-natural-kas-ve-kirpik-besleyici-serum
4	20200101C5	C5	5	bade natural eyebrow and eyelash nourishing serum	skin care serums	cosmetics	Bade Natural	1	1	2020-01-01	ORDER COMPLETED	5	49,95	249,75	%31-%40	%50 DISCOUNT CAMPAIGN	https://www.baskasindarama.com/urun/bade-natural-kas-ve-kirpik-besleyici-serum

Table A.7 RFM Dataframe

Row	**Customer**	**Recency**	**Frequency**	**Monetary**
0	C1	138	5	510.2
1	C10	334	2	214.0
2	C100	331	2	284.9
3	C101	322	2	121.9
4	C102	71	2	488.9

Table A.8 RFM Dataframe with Standardized Values

Row	**Customer**	**Recency**	**Frequency**	**Monetary**	**Recency__Standardized**	**Frequency__Standardized**	**Monetary__Standardized**
0	C1	138	5	510.2	-0.039332	3.765232	-0.256417
1	C10	334	2	214.0	1.912878	0.515256	-0.413088
2	C100	331	2	284.9	1.882998	0.515256	-0.375586
3	C101	322	2	121.9	1.793355	0.515256	-0.461803
4	C102	71	2	488.9	-0.706669	0.515256	-0.267683

Table A.9 Initial Layout for the Healthy Meals Page

	1	**2**	**3**	**4**
A	2	7	10	11
B	3	4	9	13
C	5	1	12	6
D	8			

Table A.10 Improved Layout for the Healthy Meals Page

	1	**2**	**3**	**4**
A	1	8	6	10
B	12	13	3	5
C	7	9	4	2
D	11			

Table A.11 Initial Layout for the Legumes Page

	1	**2**	**3**	**4**
A	3	7	1	8
B	5	2	4	6
C	9	10	11	

Table A.12 Improved Layout for the Legumes Page

	1	**2**	**3**	**4**
A	2	7	3	5
B	1	4	8	6
C	9	10	11	