CUSTOMER SEGMENTATION WITH CLUSTERING METHODS IN THE RETAIL INDUSTRY

Hayriye ŞENTÜRK
Yıldız Teknik Üniversitesi, Türkiye
hayriyesenturk94@gmail.com
https://orcid.org/0000-0002-9523-8745

Ebru GEÇİCİ
Yıldız Teknik Üniversitesi, Türkiye
egecici@yildiz.edu.tr
https://orcid.org/0000-0002-7954-9578

Selçuk ALP
Yıldız Teknik Üniversitesi, Türkiye
alp@yildiz.edu.tr
https://orcid.org/0000-0002-6545-4287

Atıf	ŞENTÜRK, H.; GEÇİCİ,; ALP, S. (2024). CUSTOMER SEGMENTATION WITH CLUSTERING METHODS IN THE RETAIL INDUSTRY. *İstanbul Aydın Üniversitesi Sosyal Bilimler Dergisi,* 16(4),551-573.

ABSTRACT

Businesses that carry out marketing efforts moved away from product-oriented work, understood the importance of the customer, and shifted towards customer-centered practices. This situation has made customer-centered efforts more important and has caused businesses to focus more on activities related to customer relations. Today with tech development and increasing competition, company-customer relations become more important. Creating a customer profile is critical for businesses to recognize their customers and distinguish their most profitable customers. By understanding their customers' behavior, businesses can tailor their marketing and customer relationship management strategies. Thus, they can meet their customers' needs, increase their satisfaction and loyalty to their businesses and encourage them to shop with them again. Thus, this study aims to categorize customers based on RFM metrics and interpret the obtained clusters from a marketing perspective. At the segmentation phase, hierarchical and non-hierarchical clustering methods, namely k-means, AGNES, and DBSCAN, are used and the results are compared. First, data, which consist of the shopping information of 38975 customers who shopped from e-commerce in one year, are collected from a textile retail company in Istanbul. Then, the purchase amount

Geliş tarihi: 13.07.2024 – Kabul tarihi: 04.09.2024, DOI: 10.17932/IAU.IAUSBD.2021.021/iausbd_v16i4004
Araştırma Makalesi-Bu makale iThenticate programıyla kontrol edilmiştir.

spent by customers is additionally scored to reveal the most valuable customers. It is observed that better results are mined from the k-means algorithms. As a result, four different customer types are determined: loyal customer, potential customer, new customer, and lost customer types. In conclusion, profile-oriented marketing strategies are presented.

Keywords: *Customer Segmentation, Marketing, RFM Analysis, Unsupervised Learning, Clustering.*

PERAKENDE SEKTÖRÜNDE KÜMELEME YÖNTEMLERİ İLE MÜŞTERİ SEGMENTASYONU

ÖZ

Pazarlama çabaları yürüten işletmeler ürün odaklı çalışmalardan uzaklaşarak müşterinin önemini kavramış ve müşteri merkezli uygulamalara doğru kaymıştır. Bu durum müşteri merkezli çabaları daha önemli hale getirmiş ve işletmelerin müşteri ilişkileri ile ilgili faaliyetlere daha fazla yönelmelerine neden olmuştur. Günümüzde teknolojinin gelişmesi ve rekabetin artmasıyla birlikte şirket-müşteri ilişkileri daha da önem kazanmaktadır. İşletmelerin müşterilerini tanıması ve en kârlı müşterilerini ayırt edebilmesi için müşteri profili oluşturmak kritik öneme sahiptir. İşletmeler, müşterilerinin davranışlarını anlayarak pazarlama ve müşteri ilişkileri yönetimi stratejilerini uyarlayabilirler. Böylece müşterilerinin ihtiyaçlarını karşılayabilir, memnuniyetlerini ve işletmelerine olan bağlılıklarını artırabilir ve onları kendilerinden tekrar alışveriş yapmaya teşvik edebilir. Böylece bu çalışma, RFM metriklerine göre müşterileri kategorize etmeyi ve elde edilen kümeleri pazarlama perspektifinden yorumlamayı amaçlamaktadır. Segmentasyon aşamasında k-means, AGNES ve DBSCAN gibi hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri kullanılarak sonuçlar karşılaştırılmıştır. Öncelikle İstanbul'daki bir tekstil perakende firmasından bir yıl içinde e-ticaretten alışveriş yapan 38975 müşterinin alışveriş bilgilerinden oluşan veriler toplanmıştır. Daha sonra ek olarak müşterilerin harcadığı satın alma tutarı puanlanarak en değerli müşteriler ortaya çıkarılır. K-means algoritmalarından daha iyi sonuçlar elde edildiği görülmektedir. Sonuç olarak dört farklı müşteri tipi belirlendi: sadık müşteri, potansiyel müşteri, yeni müşteri ve kayıp müşteri tipleri. Sonuç olarak profil odaklı pazarlama stratejileri sunulmaktadır.

Anahtar Kelimeler: *Müşteri Segmentasyonu, Pazarlama, RFM Analizi, Gözetimsiz Öğrenme, Kümeleme.*

INTRODUCTION

The evolution of technology and market dynamics has significantly altered consumer shopping habits. E-commerce platforms have gained widespread popularity due to their convenience and reduced physical effort. As online supermarket shopping constitutes a substantial portion of the e-commerce market, providers must differentiate themselves by comprehensively understanding and addressing customer needs. While customer expectations and value perceptions vary in the context of online shopping services, a lack of customer profiling hinders businesses from delivering personalized experiences. This can lead to financial losses due to ineffective promotions and discounts. By accurately identifying customer segments and their specific needs, businesses can implement profitable sales strategies and cultivate strong customer loyalty. Consequently, effective customer segmentation enables companies to focus on the most lucrative market segments and develop targeted marketing strategies and business solutions. That is, businesses are trying to compete under challenging conditions such as economic, sociological, legal and inter-state tension. Businesses need to work harder to gain new customers and retain existing customers in challenging conditions and intense competition. It is important for businesses to be able to communicate with customers at the right time. While offers made at the right time are important to customers, the appearance of an offer when they do not need it can lead to unnecessary waste of time and complaints for customers. Today, most of the marketing approaches are starting to turn into customer-based marketing techniques to develop strong customer relations.

It is generally accepted that acquiring a new customer is about five times more costly than retaining an existing one, and ten times more costly to regain a dissatisfied customer. In addition, many studies have shown that a five-point increase in customer retention can increase profits by more than 25 percent (Marcus, 1998). Given the significant costs associated with acquiring new customers and the profitability of retaining existing ones, businesses must prioritize strategies that enhance customer loyalty. One effective approach is to stay attuned to evolving customer preferences and behaviors, which can shift rapidly in today's dynamic market environment. In short, customer needs are constantly changing. In this respect, it is important for businesses to be aware of these changes and always be able to respond to these changes. Customer segmentation can be used in this regard. Segmentation is one of fundamental strategy for managing marketing efforts towards customers (Zhang et al., 2009). Customer segmentation helps sellers identify better value propositions, allocate resources, identify and effectively track opportunities, anticipate problems and find solutions, and reflect on situations. The main purpose of customer profiling is to describe the behavioral and functional characteristics of customers (Adomavicius and Tuzhilin, 2001). In this way, the value of the customer for the company is demonstrated.

That is, it is important to present customer profiling obtained by customer segmentation to examine the dynamic nature of consumer preferences and the corresponding strategic adaptations undertaken by businesses. Thus, the aim of this study is to apply customer segmentation in a retail business and to present different strategies in each segment by using the results obtained. It can be used to help businesses identify the variables they need to focus more on and develop different marketing strategies for customers with different profiles. These strategies can increase profitability and help your business identify strengths and weaknesses in its overall business strategy. In the study, Recency – Frequency – Monetary (RFM) analysis, which is frequently used in customer clustering and lost customer research, are used to create customer segmentations. The results of RFM analysis and the results of k-means, agglomerative nesting (AGNES) clustering and density-based spatial clustering of applications with noise (DBSCAN) clustering algorithms, which are frequently used in the literature, are compared. In order to measure the clustering performance, the compatibility of the models proposed in the study with the Silhouette index is tested.

The rest of the paper is organized as follows: Section 2 provides literature review for usage of the RFM analysis. In section 3, we present a methodology which is used the clustering of the customer. Section 4, on the other hand, gives the results of the application of the methodology for each method. Then, we conclude and suggest about the extension of the study in section 5.

LITERATURE REVIEW

In this part of the study, studies on customer segmentation are presented. In this context, first, studies with RFM content, including the subject of this study, are presented. Then, studies including different methods integrated with RFM will be presented. Then, studies on the application areas of RFM are addressed.

In the literature, recently, there are many academic studies that use machine learning algorithms to identify customer segments, measure customer loyalty, win back customers, and predict the future. One of the studies is presented by Soeini and Fathalizade (2012). They propose an updated model for selecting target customers for direct marketing in their work in the insurance industry. A method is presented that extends the RFM model by including the time and cost since the customer made the first purchase. Dogan et al. (2018), on the other hand, propose two consumer segmentation models based on RFM analysis and k-means clustering in their study on a business operating in the retail sector in Turkey. In this study, loyalty cards for business customers are gradually defined according to cluster size and recommended values. Another study on the RFM model is carried out in the recycling sector by Erpolat Taşabat and Akca (2020). In the study, a metal product business determined the customers who could make the highest contribution to recycling according to the RFM scores they calculated.

In addition to the aforementioned studies on RFM, there are also studies in the literature in which RFM weights are introduced with different methodologies. In the study of Shih and Liu (2003), analytical hierarchy process (AHP) is applied to determine the weights of the parameters, considering that the weights of the RFM parameters may differ according to the products and sectors in the calculation of the customer lifetime value (CLV). Khajvand et al. (2011) discuss the issue of customer segmentation for a company that produces personal care products. In the study, CLV is addressed by applying RFM analysis and k-means clustering algorithm. Chuang and Shen (2008) use RFM analysis and k-means clustering algorithm in the customer value analysis they developed for a business operating in Taiwan. In their studies, firstly, importance weights of R, F, M variables are determined with AHP. Then, CLV is calculated using these weights and customers are analyzed with the obtained values.

Another study conducted with RFM includes models created by integrating different methodologies such as clustering algorithms. Zhang et al. (2015) propose the aggregation method as the RFM-C model by integrating RFM with a metric-based approach. Bhatia et al. (2022) use the data of shopping mall customers in their customer segmentation study, including variables such as age, spending score, monthly income, and gender. To determine the appropriate number of clusters, they apply the within cluster sum of squares method and used the obtained cluster number in the k-means algorithm. Solichin and Wibowo (2022) build the k-means clustering algorithm based on the RFM model and combined this model with the user event tracking parameter. They divide their customers into three categories, platinum, gold and silver, according to the model they used, and emphasized that customers could develop different strategies for each group. Ernawati et al. (2022) use the RFM and k-means algorithm to determine the needs of university students. In this study, they use geographic information system and k-means algorithm to add the potential of the regions as a variable and propose the RFM-D model by integrating it with the RFM model. They compare the performances obtained in the study, which they exemplified using Indonesian universities, with the customer lifetime value-based RFM method. According to the results obtained, it is emphasized that there should be target schools in regions with high potential. Using online sales data, Wu et al. (2020) divide customers into four different segments according to their shopping habits with RFM and k-means clustering methods. Wu et al (2021), on the other hand, calculate the indicators that improve RFM based on real purchase data, give weight to the indicators and finally classify the value of users using the k-means++ algorithm. According to their results, they state that the user classification based on the proposed RFM model is more accurate than the user classification based on the traditional RFM model, and the propose RFM model can more accurately define user value.

Customer segmentation analyses can be effectively employed in various sectors and one of them is retail industry. In a 2005 study on churn in the retail industry (Buckinx and van del Poel, 2005), the concept of partial churn is suggested. Considering the possibility that the loss in the retail sector may recur over time rather than continuously, this situation is expressed as partial loss. Using the data set of a supermarket chain in Europe, the researchers accept the customer churn time as three months. Shopping time, product category, payment method, customer demographics and complaints of 32,000 customers are used as variables. These variables are analyzed with RFM, logistic regression (LR), automatic relevance determination, neural networks, and random forests. Miguéis et al. (2012) use 7,200 customers' shopping information of a business to examine customer churn in the retail industry. In addition to the RFM value, the concept of variable memory, a method based on the "Markov Chain", is used to order products the customer purchases. The aim is to learn shopping behavior by analyzing the products' purchase orders. Tanaka et al. (2017) group the customers of a Japanese supermarket chain in order of importance using a hybrid method consisting of RFM analysis and LR. Sohrabi and Khanlari (2007) propose the measurement of CLV based on the RFM model. In the study, k-means clustering approach is used to determine CLV and customers are clustered according to RFM criteria and then propose a customer retention strategy. RFM analyses also have applications in sectors other than the retail sector. One of them is made by Chan (2008). The author develops a model that uses RFM analysis and genetic algorithm to segment the automobile sales market. According to the results obtained, special offers are determined for the most profitable customers. In his retail study, Ha (2007) divides the customers' 15-month period by using three-term periods and clusters them using RFM values for a total of five periods. In this study, RFM values are evaluated according to their average values and customer groups are determined according to these average values. Birant (2011) propose an integrated model in data mining in his study. The proposed model consists of five main parts. These sections can be listed as data preprocessing, RFM analysis, customer review process, segmentation, and forecasting.

This study, in contrast to the studies mentioned above, this study uses three different clustering algorithms to perform customer segmentation. In this study, three different clustering algorithms, namely hierarchical clustering (AGNES), partitioning clustering in non-hierarchical (k-means), density-based clustering in non-hierarchical clustering (DBSCAN), are used. This study seeks to segment customers based on RFM metrics and analyze the resulting clusters from a marketing perspective. The study employs hierarchical and non-hierarchical clustering techniques.

METHODOLOGY

In this study, RFM analysis, k-means clustering, AGNES clustering and DBSCAN clustering algorithms are used in customer segmentation analysis.

RFM Analysis

The RFM model was first developed by Hughes (1994) to select the most valuable customers from all customer data obtained in a given time period. It is a very useful model especially for achieving successful outcomes in customer relations, and it is widely used in the development of discount and promotion strategies in marketing areas and in the sales of loans and stocks of banks (Chang & Tsai, 2011; Hu et al., 2013).

In the analysis of RFM, the Recency (R) value represents the customer's last purchase, the Frequency (F) value represents the frequency of the customer's purchase, and the Monetary (M) represents the total purchase amount of the customer (McCarty & Patient, 2007). Based on these three attributes in the RFM analysis, a value between 1 and 5 is assigned to each symbol. Recency, Frequency and Monetary Scores are calculated by dividing them into 5 groups, each of which is 20%, and giving the highest score of 5 to the first 20%, 4 to the 2nd highest part, 3 to the 3rd part, 2 to the 4th part, and 1 point to the last 20% piece. RFM score scale is given in Table 1 (Cheng & Chen, 2009).

Table 1.
The score scale RFM

Score	Recency (%)	Frequency (%)	Monetary (%)
5	0-20	0-20	0-20
4	20-40	20-40	20-40
3	40-60	40-60	40-60
2	60-80	60-80	60-80
1	80-100	80-100	80-100

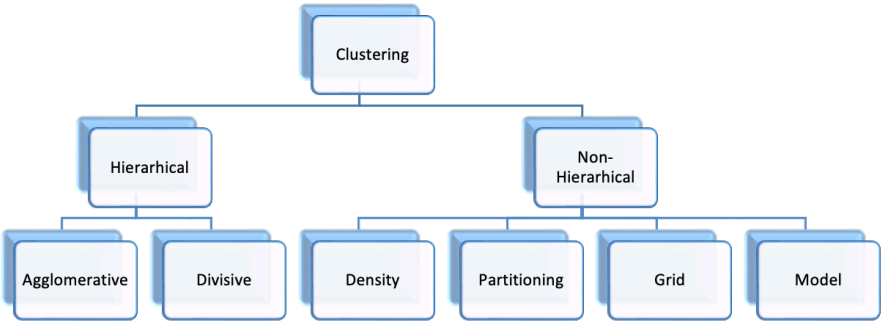
As a result, each customer is assigned a total score ranging from 111 to 555. Customers with 111 points are considered lost customers. Customers with a score of 555, on the other hand, are called loyal customers and are likely to bring the most profit for the business. In this study, customers in the retail sector will be divided into segments according to their purchasing values on the RFM scale by considering the purchasing habits of customers. Later, these segments will be used to integrate RFM results into clustering algorithms. Thus, strategies that should be retained or that will allow customer loyalty to be achieved can be developed.

Clustering Analysis

Cluster analysis is a method that allows to categorize the examined units into certain groups according to their similarities, to reveal the common features of the units and to develop a general definition about these clusters. Clustering results should show a high degree of homogeneity within clusters and a high degree of heterogeneity between clusters (Sharma, 1996).

Many algorithms have been proposed for cluster analysis. These algorithms can be grouped under two headings in the literature as hierarchical and non-hierarchical clustering algorithms. Hierarchical clustering algorithms group the two most similar objects into a cluster. Hierarchical clustering algorithms are divided into two as agglomerative and divisive approaches. Non-hierarchical clustering algorithms are algorithms that cluster data directly. Non-hierarchical algorithms change cluster centers until all points are at a minimum distance from their respective cluster centers. In agglomerative hierarchical clustering methods, each observation is first treated as a cluster and then clusters close to each other are combined to reduce the number of clusters at each step. In divisive hierarchical clustering methods, the processes start with considering all the observations in a single cluster and smaller clusters are created by separating the dissimilar observations in the cluster (Timm, 2002). Non-hierarchical clustering algorithms, on the other hand, are divided into four subcategories: density-based, partitioning, grid-based and model approaches (Wu and Chow, 2004) (Figure 1).

Figure 1.
The classification of the clustering algorithms



The common goal of both techniques is to maximize the inter-cluster differences and intra-cluster similarities. Many methods (Euclidean, square Euclidean, Minkowski, Manhattan, Pearson Correlation method, Chebychev, Mahalanobis etc.) have been developed to determine the distance between observations in order to determine the differences or similarities between the observations.

The main purpose of using clustering algorithms is to divide customers into clusters using customer segmentation data obtained with RFM. In this way, decision makers will be able to determine different marketing strategies for clusters containing customers with similar characteristics. For this purpose, in this study, k-means from partitioning-based clustering methods, AGNES from non-hierarchical clustering algorithms, DBSCAN from density methods are used for clustering, and Euclidean algorithm is used to measure distances between units.

K-means Clustering

The k-means algorithm is a partitioning-based clustering algorithm developed by J.B MacQueen in 1967. The *k* parameter in the k-means algorithm determines the number of clusters. Since *k* unique clusters are formed and the center of each cluster is the average of the values forming the cluster, the algorithm is called k-means. The general logic of the k-means algorithm is to divide a data set consisting of *n* observations into *k* clusters determined before the algorithm is executed. The aim is to ensure that the distance between the clusters formed at the end of the clustering process is maximum and the distance within the cluster is minimum (Jain, 2010). The k-means algorithm is designed as an iterative algorithm in which the clusters are constantly updated until the optimal solution is reached (Steinbach et al., 2000).

Agglomerative Nesting (AGNES)

AGNES (Gowda and Krishna, 1978) is a type of hierarchical clustering technique that groups data objects into a tree of clusters. It begins by creating clusters composed of single data objects, and then iteratively, using some distance metric merges such clusters into larger ones. This second step is repeated several times until a single cluster is obtained (Toshniwal et al., 2020). The hierarchical clustering method is a series of steps that reveal a tree-like structure by removing or adding an element from the clusters (Ketchen and Shook, 1996). Hierarchical clustering techniques combine units at different stages and take into account the distance or similarity of units in a data set to each other to determine the distance or similarity level at which items are placed (Wu and Chow, 2004). It is also known as AGNES, which stands for agglomerative hierarchical clustering algorithm. The way it works is from the bottom up. Here, each observation is considered as a single-element set. Then, at each step of the algorithm, each cluster is merged with the other cluster most similar to it to form a larger cluster. This process continues until all elements become one big set (root). The results are plotted as a dendrogram (tree diagram) (Johnson and Wichern, 2007).

Agglomerative clustering method firstly treats each observation as a cluster and initially there are as many clusters as the number of observations. The two closest observations then form a new cluster, and the process continues by decreasing the number of clusters at each step (Johnson and Wichern, 2007). By reducing the number of clusters by one, the iterated distance matrix is found. This process continues until it becomes one root.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN algorithm is introduced in 1996 and formed the basis of density-based clustering methods (Ester, 1996). DBSCAN is an algorithmic approach to data clustering that is based on density measures. The algorithm considers the density of objects when creating clusters. It is characterized by the inclusion of only those clusters that meet the pre-defined density threshold, with any data points falling

below this threshold being classified as noise. The DBSCAN method utilizes the minimum input point parameters and epsilon to identify clusters. The process of determining the requisite parameters is one of trial and error; consequently, the determination of the parameters must be tested on multiple occasions to obtain the requisite number of clusters (Brahmana et al., 2020). Clusters are defined by denser data objects. Low-density object clusters indicate outliers or noisy spots. DBSCAN is particularly useful for large databases and data sets containing noisy objects. It is also often used to identify clusters of various sizes and shapes.

Methods Used to Determine the Performance of Clustering Algorithms

To measure the performance of the clustering algorithms, some criteria called similarity index have been developed (Bolshakova and Azuaje, 2003). The most used of these techniques are Silhouette index, Calinski-Harabazs index, and Davies-Bouldin index (Maulik and Bandyopadhyay, 2002). The Calinski-Harabasz index evaluates cluster validity based on the cluster aggregate and cluster aggregate mean. The cluster with the highest index value gives the appropriate number of clusters (Caliński and Harabasz, 1974). The Silhouette index calculates cluster performance based on the two-way difference based on the distances between clusters and the distance within the cluster. Silhouette index takes values between -1 and 1 (Rousseeuw, 1987). The Davies-Bouldin index is calculated over the similarity between each cluster and all other clusters, and the highest value is assigned as cluster similarity. The smaller the index, the better the clustering result (Liu, Li, Xiong, Xuedong, Wu, 2010). In this study, Silhouette index is used to measure the performance of clustering algorithms.

APPLICATION

Identifying customer profiles is an important part of the customer relationship management process. In this study, customer profiles are created for e-commerce customers of a textile company by using RFM and clustering algorithms. To make this analysis python programming language is used with the clustering algorithm packages. The data used in the study are taken from a textile company operating in the retail sector in Istanbul. Data covers 38975 purchases by customers from February 2021 to December 2021. In the study, first, data editing is carried out. For this purposes, analysis, arrangement, and extraction of missing data are carried out. Then the analysis phase is handled.

Recency-Frequency-Monetary Analysis

For RFM analysis, “Purchase Frequency”, “Last Purchase Date”, and “Customer's Total Spending Amount” information are calculated. Since the variables “Monetary", "Recency" and "Frequency" take different values, the data are normalized using the "min-max normalization method". An RFM score is calculated for each customer using the data obtained. Customers are segmented according to the resulting RFM scores. According to these scores, customers are defined as "platinum", "gold", "silver", "bronze", and "risky". Table 2 shows

the customer RFM values. In the RFM analysis, the RFM score is obtained by summing the R+F+M values.

Table 2
Customer RFM scores[1]

Customer ID	R	F	M	Recency Value	Frequency Value	Monetary Value	Score	RFM Score
12402	324	1	225.60	1	1	1	3	111
12403	50	1	427.70	4	1	2	7	412
12405	149	1	1710.39	2	1	4	7	214
12406	23	2	3415.30	4	2	4	10	424
12407	50	5	1708.12	4	3	4	11	434
12408	97	2	284.13	3	2	2	7	322
12409	295	1	311.55	2	1	2	5	212
12410	135	11	3907.50	3	5	5	13	355
12411	79	7	2994.02	3	4	5	12	345
12412	273	4	650.94	2	3	3	8	233

In Table 2, it is seen that the customer with ID 12402 in the first line made a purchase 324 days ago, made a purchase only once, and the total purchase amount was 225.60 TL. Likewise, it is determined that customer with ID 12407 also shopped 50 days ago and made 5 purchases a year and paid 1,708 TL for these purchases.

After the RFM score is obtained, the value added by each customer to the business can be evaluated more accurately. Based on the obtained R, F, M values, these values are averaged, and if all three values are above the average, they are called loyal customers. Customers with an RFM score of 555 (combining the top 5 for each metric), for example, are the most valuable customer segment of the business. Customers with high RFM scores, which are based on the obtained R, F, M values, rank higher, while customers with lower RFM scores rank lower. While the top 20% represent the opportunity areas of the business, these are the most important customers of the business. The bottom 20% represent areas that the company avoids, which are the business' worst customers. For example, if a customer with an RFM score of 214 is reviewed, these customers are considered risky customers. The Recency value is 2 when the customers are sorted from the largest to the smallest according to their most recent purchase date, and similarly, the frequency variable is given 1 point since it is in the bottom 20% when the average time between the purchases of the customers is ranked from the lowest to the highest. Customers in this group are customers with low currency and frequency values, that is, they do not come to the business very often and have not

[1] Note that, R = Day, F = Shopping frequency, M = Shopping cart amount in TL

come to the business recently, but these customers are also customers who shop at high amounts when they come. The RFM scores given in Table 2 were used during the clustering process. If the R and F parameters are above the average and the M parameter is below the average, they form the "potential customers" group. If the R parameter is above the average and the F and M parameters are below the average, they can be defined as "new customers" (Wei et al., 2013).

Based on this RFM score, customers are divided into "platinum", "gold", "silver", "bronze", and "risky" customer groups. This distinction is obtained by ordering the RFM scores of the customers based on the values of the customers' R, F, M parameters from largest to smallest. That is, customers are ranked from top to bottom, from largest to smallest, according to the RFM score they received. Afterwards, the obtained list is divided into five groups (platinum, gold, silver, bronze, and risky) using the created ranking. The customer segments formed when the e-commerce shopping customers in this study are segmented and the values of the customers in each segment are given in Table 3.

Table 3

Average values for RFM clusters

Customer Type	Recency	Frequency	Monetary	Number of Customer
Risky	270.29	1.14	247.03	10779
Bronze	231.46	2.20	648.79	10762
Silver	146.66	3.86	1251.48	6923
Gold	68.93	7.56	3075.24	9847
Platinum	16.56	21.77	11316.08	664

Determining the Number of Clusters for Clustering Algorithms

Silhouette index is widely used in clustering algorithms to determine the appropriate number of clusters. The relationship between the Silhouette Width index and the cluster structure is given in Table 4. An average Silhouette index above 0.5 indicates that it is admissible at an acceptable level, and a value above 0.7 indicates that it is very strong (Ng & Han, 1994).

Table 4

Relationship between average Silhouette width index and cluster structure (Adapted from Ng and Han (1994))

Average Silhouette Width	Cluster Structure
0.71-1	Strong
0.51-0.7	Successful
0.26-0.5	Weak
<0.25	Unsuccessful

In this study, k-means, AGNES, and DBSCAN clustering algorithms are used. The appropriate number of clusters for all three clustering algorithms is determined by the Silhouette index.

Table 5
Index Values of Clustering Algorithms

Methodologies	Number of Clusters	Silhouette Score
K-means	3	0.509129
	4	**0.531356**
AGNES	**3**	**0.442102**
	4	0.428409
DBSCAN	3	0.368886
	4	**0.407764**

According to the Silhouette index, k-means is the best clustering algorithm and the best number of clusters is 4. Below are the comparative results of each of the k-means, AGNES and DBSCAN algorithms with RFM.

Application of the Clustering Algorithms
Application of K-Means
For the k-means clustering method, after determining the optimal number of clusters as 4 with the Silhouette index, python programming language is used to implement the k-means method based on the customer's RFM score. The average RFM values of the clusters formed as a result of the application of this method are shown in Table 6.

Table 6
Values for the K-means method

Cluster Type	Recency	Frequency	Monetary	Number of Customer
Cluster 1	21.19	18.94	9795.97	7431
Cluster 2	164.29	5.11	1832.95	11859
Cluster 3	58.24	3.11	834.25	7276
Cluster 4	280.73	1.31	308.54	12409

When the values in Table 6 are examined, it can be seen that the most profitable customers are in the first cluster. It can be said that the average recency value is more recent than other clusters, the frequency level is much higher than the other clusters, and the average basket amount is by far higher. Therefore, the customers in this cluster are the most loyal customer base of the business. The second profitable customer segment of the business is the third cluster. Its monetary value is relatively low compared to cluster number two, but the relevance and frequency

of customers is significantly higher. Bringing the customers in this cluster into the loyal customer category is an important strategy for the business. The weakest link among the four clusters is the customer group that makes up the fourth cluster. Compared to other clusters, we can see that the value of each variable is lower than other clusters. This makes them the least profitable customer group for the business. The common customer numbers in the clusters created with RFM and k-means are given in Table 7.

Table 7

Distribution of customers in segments obtained with RFM and K-means

	Platinum	Gold	Silver	Bronze	Risky
Cluster 1	5413	2018	0	0	0
Cluster 2	0	4977	3831	3015	0
Cluster 3	0	2898	2475	1903	0
Cluster 4	0	0	0	4302	8107

Application of Agglomerative Nesting

The dendrogram is generally used to represent hierarchical cluster results. The vertical lines in the dendrogram represent distances, while the horizontal lines indicate clusters of junctions. The cluster intersection points on the scale show which clusters are formed and the distance between them. In order to calculate the coefficients in the hierarchical clustering method, the Complete method, in which the best result is obtained from seven methods (ward, average, single, complete, weighted, centroid, median), is used. According to the AGNES results, three customer segments are obtained. The average RFM values of the clusters formed as a result of the application of this method are shown in Table 8.

Table 8

Values for the AGNES method

Customer Type	Recency	Frequency	Monetary	Number of Customer
	Average	Average	Average	
Cluster 1	40.21	17.89	9294.31	8838
Cluster 2	173.05	3.84	1208.36	18451
Cluster 3	275.42	1.29	236.48	11679

The number of common customers in the clusters created with RFM and AGNES is given in Table 9.

Table 9
Distribution of customers in segments obtained with RFM and AGNES

	Platinum	Gold	Silver	Bronze	Risky
Cluster 3	0	0	80	2834	8765
Cluster 2	211	5245	5915	6498	582
Cluster 1	5545	3293	7	0	0

Customers in the first cluster obtained by hierarchical clustering are clustered by RFM as 5545 platinum, 3293 gold and 7 silver. The customers in the first cluster shopped an average of 40 days ago, and their annual frequency of visiting the business is 17, and they have spent an average of 9,294.314 TL until now. The second cluster consists of 211 platinum, 5245 gold, 5915 silver, 6498 bronze and 582 risky customer segments. It is found that this cluster shopped on average 173 days before the last business, visited the business approximately 4 times a year, and spent an average of 1,208,355 TL. The third cluster consists of customers segmented as 80 silver, 2834 bronze and 8765 risky. An average of 275 days have passed since the last shopping of the third cluster, and the frequency of these customers visiting the business is 1 and the average amount they spend is 236.484 TL.

Application of Density-Based Spatial Clustering of Applications with Noise
There are two clustering parameters in the DBSCAN algorithm: (i) Epsilon and (ii) MinPoints. In the first stage, after the MinPoints value, the Epsilon base value is determined according to the determined MinPoints base value.

The lowest and highest number of customers in the data set are determined. Multiples of the lowest number of shoppers and the highest number of shoppers are taken. This coefficient shows the maximum number of clusters that can be assigned in the problem. After the minPoints value is determined, trials are made using the DBSCAN algorithm to determine the epsilon value. The epsilon value starts at 1 and repeats until the epsilon values form a single set. In this study, four customer segments are obtained according to the results of DBSCAN Cluster Analysis. The number of common customers in the clusters created with RFM and DBSCAN is given in Table 10.

Table 10
Values for the DBSCAN clustering method

Customer Type	Recency	Frequency	Monetary	Number of Customer
Cluster 1	32.75	15.94	6556.97	1362
Cluster 2	74.29	3.11	2821.95	8261
Cluster 3	158.24	1.31	1834.25	18608
Cluster 4	280.73	1.06	308.54	10738

The number of common customers in the clusters created with RFM and DBSCAN is given in Table 11.

Table 11
Distribution of customers in segments obtained with RFM and DBSCAN

	Platinum	Gold	Silver	Bronze	Risky
Cluster 4	0	0	80	2608	8050
Cluster 3	0	4750	5915	6673	1268
Cluster 2	5136	3125		0	0
Cluster 1	1340	22	0	0	0

Customers in the first cluster, obtained by the DBSCAN clustering method, are clustered by RFM as 1340 platinum and 22 gold. The customers in the first cluster shop an average of 32 days ago, and their annual frequency of visiting the business is 27, and they have spent an average of 6,556,972 TL until now. The second cluster consists of 5136 platinum, 3125 gold, eight silver customer segments. It is found that this cluster shop an average of 74 days before the last business, visit the business 3 times a year on average, and spent an average of 2,821,947 TL. The third cluster consists of customers segmented as 4750 gold, 5915 silver, 6673 bronze and 1268 risky. An average of 158 days have passed since the last shopping of the third cluster, and the frequency of these customers visiting the business is 1 and the average amount they spend is 1,834,253 TL. Customers in the fourth cluster are grouped by RFM as 80 silver, 2608 bronze and 8050 risky. Customers in the fourth cluster shop an average of 280 days ago and spend an average of 308,537 TL, which is considerably lower than the other clusters.

CONCLUSION

As technology develops and market conditions change, the way customers shop also changes. Many customers have started to use e-commerce platforms widely and frequently due to the advantage of time and the fact that it does not require physical effort. In other words, customers' shopping habits are changing, and businesses need to develop sales strategies to keep up with these habits. Moreover, since online supermarket purchasing is an important part of online shopping, online supermarket service providers must ensure a reason for them to be chosen by customers. This can only be achieved by fully understanding customer needs and offering value propositions that fit those needs. When customers use online shopping services, on the other hand, their needs and expectations are not the same in terms of the added value they create for the business. If a business does not create customer profiles, it cannot fully know its customers and offer them personalized service. Since the customer profile is not known, the business may make a loss while waiting for profit because of special discounts and customer advantages. Once the similarly behaving customer groups are identified and

the needs and wishes of these customer groups are well understood, businesses can implement profitable sales strategies and create healthy loyalty programs. Thus, the success of online supermarket brands requires adapting and adopting marketing channels, product portfolios, product diversification methods, delivery speeds, mobile applications, website design, pricing policies, and campaigns according to target consumer groups. In this context, customer segmentation is an easy way for these businesses to organize and manage customer relationships. Therefore, if businesses can identify customer segments correctly, they can choose the more profitable one than others and develop marketing strategies and business solutions targeting specific customer groups. This study seeks to categorize customers based on RFM criteria and interpret the identified clusters within a marketing context of retail industry.

In this study, RFM analysis is incorporated into clustering techniques. Based on customers' purchasing habits, customers are segmented according to their purchase value on the RFM scale. The customer segmentation process helps tailor the marketing, service and sales plan to the needs of different customer groups, increasing brand loyalty and customer satisfaction. In the customer segmentation phase, hierarchical and non-hierarchical clustering methods are used, and the results are compared. In the second stage, clustering analyzes are carried out using the data obtained from the RFM analysis. It is observed that the machine learning algorithms used perform better with the k-means clustering method. Then, potential customers who can participate in the campaign and services are determined. Offering benefits that maximize profit in the sales area, increase customer satisfaction, and reduce unnecessary costs; can provide maximum profit with minimum time and effort. According to the results of the study, the segments that spend the most are loyal customers and high-potential customers. The business needs to retain these customers and conduct research to ensure the continuity of its relationship with the business.

The positions of the four clusters created in the cluster analysis in the customer relations and value matrices, the characteristics attributed to each cluster, and the suggested marketing practices according to these characteristics are explained below. Among the clusters formed by the k-means algorithm:

- The first cluster with the highest RFM score is the most profitable cluster. If customer relationship management is implemented, long-term relationships will be able to be established in this cluster with the highest income. Efforts should be made to ensure that these clusters are not lost by offering special discount promotions and extra discounts with each purchase. Various discounts can be made in the membership system by introducing a special membership system for these people. These discounts should not be continuous and should not coincide with general discount times. Making good wishes from time to time (especially on special days) and sending messages that make them feel

important will make them feel valuable and this will increase the loyalty of the customer.

- Customers in the second cluster have low R scores, medium F scores, and high M scores. These customers are the customers who have not come to the business recently, who have a low frequency of coming to the business, but who shop in large sums when they come. Once these customers are identified, promotions should be offered based on their preferences and encouraged to visit the business more often. Such customers can be converted into loyal and more profitable customers for the company through cross-selling and effective cost management.
- The people in the third cluster make frequent purchases from the business, but they do not earn us the desired income. There are 7958 people in this segment. On average, they shopped 24 days ago. Their shopping frequency is three times a year. The shopping invoices of these people can be analyzed with algorithms such as apriori, so they can buy more products. The business may offer special discounts to encourage these customers to buy together, with announcements such as "50% off two and three items", according to the purchases made by their customers.
- There are 13464 people in the fourth cluster with the lowest RFM score and the highest number of observations. An average of 290 days of shopping was made and the frequency of shopping was found to be once a year. To increase and maintain the frequency of visits of this group called lost customers, it can announce seasonal discounts and recommend cheap products by simply sending an e-mail. Special incentives may be offered just for this customer segment, such as earning double points, to keep customers coming back to your business. Offers such as discounted prices or free, one-time shipping for a limited time can help mobilize these customers.

The results of this study show that e-commerce customers in clusters 1 and 2 should be the target segment. As these customers find the timely delivery and ease of use of mobile applications very important to them when using online shopping services, brands need to pay more attention to these issues to increase their market share. The promised delivery schedule greatly influences the e-commerce service provider's impression on these customers. Keeping the merchandise service "correct" is a challenge for both corporate logistics operations and customers, leading to disappointment and loss of trust in the brand. It would therefore make sense to recommend that these businesses prioritize their logistics systems and delivery obligations.

This study is expected to contribute to the literature in two different ways:

- The first is to show the applicability of RFM and clustering algorithms in the retail sector, and

- The second is to reveal the behavioral differences between clusters because of this analysis.
- We can list the limitations of the study and the topics to be addressed in future studies as follows:
- The most important limitation of the study is the use of only one year data. In future studies, the changes in customer groups can be observed by analyzing the data of consecutive years.
- Moreover, it is not enough to simply segment customers into certain segments. Answering the question of which products are preferred by the determined segments and which brands these products belong to is an important issue for retailers, especially when deciding on product variety. Therefore, RFM and apriori analysis can be used together to provide effective customer, product and brand matching in future studies.
- The concept of data mining, which can be explained as the process of extracting meaningful information from data stacks, has gained importance and its importance is increasing day by day. By using different data mining techniques, the scope of the study can be expanded by going down to details such as the customer's education, age, and rate of return to campaigns.
- Alternative methods can be tried to create clusters using algorithms different from the clustering algorithm.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2001). Using data mining methods to build customer profiles. *Computer*, *34*(2), 74-82.

Bhatia, T. K., Gupta, S., & Sharma, A. (2022, October). Analysis of Customer Segmentation Model through K-Means Clustering. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-6). IEEE.

Birant, D. (2011). Data Mining Using RFM Analysis, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978- 953-307-154-1, InTech.

Brahmana, R. S., Mohammed, F. A., & Chairuang, K. (2020). Customer segmentation based on RFM model using K-means, K-medoids, and DBSCAN methods. *Lontar Komput. J. Ilm. Teknol. Inf*, *11*(1), 32.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European journal of operational research*, *164*(1), 252-268.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert systems with applications*, *34*(4), 2754-2762.

Chang, H. C., & Tsai, H. P. (2011). Group RFM analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications*, *38*(12), 14499-14513.

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.

Chuang, H. M., & Shen, C. C. (2008, July). A study on the applications of data mining techniques to enhance customer lifetime value—based on the department store industry. In *2008 International Conference on Machine Learning and Cybernetics* (Vol. 1, pp. 168-173). IEEE.

Doğan, O., Ayçin, E., & Bulut, Z. (2018). Customer segmentation by using RFM model and clustering methods: a case study in retail industry. International Journal of Contemporary *Economics and Administrative Sciences*, 8(1), pp 1- 19.

Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2022). Target market determination for information distribution and student recruitment using an extended RFM model with spatial analysis. *Journal of Distribution Science*, *20*(6), 1-10.

Erpolat Taşabat, S., & Akca, E. (2020). Recyclıng Project Wıth RFM Analysıs In

Industrıal Materıal Sector. *Sigma: Journal of Engineering & Natural Sciences*, *38*(4), 1681-1692.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Gowda, K. C., & Krishna, G. J. P. R. (1978). Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, *10*(2), 105-112.

Ha, S. H. (2007). Applying knowledge engineering techniques to customer analysis in the service industry. *Advanced Engineering Informatics*, *21*(3), 293-301.

Hu, Y. H., Huang, T. C. K., & Kao, Y. H. (2013). Knowledge discovery of weighted RFM sequential patterns from customer sequence databases. *Journal of systems and software*, *86*(3), 779-788.

Hughes, A. M. (1994). Strategic database marketing: the masterplan for starting and managing a profitable. *Customer-based Marketing Program, Irwin Professional*.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis. 6th. *New Jersey, US: Pearson Prentice Hall*.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, *17*(6), 441-458.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57-63.

Ma, E. W., & Chow, T. W. (2004). A new shifting grid clustering algorithm. *Pattern recognition*, *37*(3), 503-514.

MacQueen, J. B. (1967). 5th Berkeley symposium on mathematical statistics and probability. *Berkeley, CA*.

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of consumer marketing*, *15*(5), 494-504.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(12), 1650-1654.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of business research*, *60*(6), 656-662.

Miguéis, V. L., Van den Poel, D., Camanho, A. S., & Falcão e Cunha, J. (2012). Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification*, *6*, 337-353.

Ng, R. T., & Han, J. (1994, September). E cient and E ective clustering methods for spatial data mining. In *Proceedings of VLDB* (pp. 144-155).

Sharma, S. (1996). Applied Multivariate Techniques. John Wiley&Sons. *Inc, New York*.

Shih, Y. Y., & Liu, C. Y. (2003). A method for customer lifetime value ranking—Combining the analytic hierarchy process and clustering analysis. *Journal of Database Marketing & Customer Strategy Management*, *11*, 159-172.

Soeini, R. A., & Fathalizade, E. (2012). Customer segmentation based on modified RFM model in the insurance industry. In *IACSIT Hong Kong Conferences* (pp. 101-104).

Sohrabi, B. & Khanlar, A. (2007). Customer Lifetime Value (CLV) Measurement Based on RFM Model. *Iranian Accounting & Auditing Review*, *14*(47), 7-20.

Solichin, A., & Wibowo, G. (2022, October). Customer Segmentation Based on Recency Frequency Monetary (RFM) and User Event Tracking (UET) Using K-Means Algorithm. In *2022 IEEE 8th Information Technology International Seminar (ITIS)* (pp. 257-262). IEEE.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.

Tanaka, T., Hamaguchi, T., Saigo, T., & Tsuda, K. (2017). Classifying and understanding prospective customers via heterogeneity of supermarket stores. *Procedia computer science*, *112*, 956-964.

Timm, N. H. (Ed.). (2002). *Applied multivariate analysis*. New York, NY: Springer New York.

Toshniwal, D., Chaturvedi, N., Parida, M., Garg, A., Choudhary, C., & Choudhary, Y. (2020). Application of clustering algorithms for spatio-temporal analysis of urban traffic data. *Transportation Research Procedia*, *48*, 1046-1059.

Wei, J. T., Lee, M. C., Chen, H. K., & Wu, H. H. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, *40*(18), 7513-7518.

Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., & Xu, G. (2020). An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Mathematical Problems in Engineering*, *2020*, 1-7.

Wu, J., Shi, L., Yang, L., XiaxiaNiu, Li, Y., XiaodongCui, ... & Zhang, Y. (2021). User value identification based on improved RFM model and k-means++ algorithm for complex data analysis. *Wireless Communications and Mobile Computing*, *2021*, 1-8.

Wu, S., & Chow, T. W. (2004). Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, *37*(2), 175-188.

Zhang, X., Feng, G., & Hui, H. (2009, June). Customer-churn research based on customer segmentation. In 2009 International Conference on Electronic Commerce and Business Intelligence (pp. 443-446). IEEE.

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. Marketing Science, 34(2), 195-208.