International Journal of New Political Economy 5(2): 115-151, 2024

The Evaluation Clustering Algorithm of Iran's Online Shopping Consumer Market

Khatereh Khorasani[1], Mansour Zaranezhad[2]*, Ghanbar Amirnezhad[3], Ali Kangarani Farahani[4]

ARTICLE INFO

Article history:

Date of submission: 11-01-2024

Date of revise: 13-07-2024

Date of acceptance: 31-07-2024

JEL Classification:

C52

J44

M31

Keywords:

Clustering

Data mining

RFM

E-commerce

K-Means

ABSTRACT

With the fast growth of e-commerce and the emerging new retail trend—online and offline integration—it is important to recognize the target market and satisfy customers with different needs by analyzing their online search behaviors. Accordingly, in this study, several internet companies in Iran were investigated. The companies were divided into 5 categories based on their product type: food, cosmetics and luxury goods, industrial goods and their accessories, sanitary goods, detergents, and clothing. Then the trading data of the companies in a certain period are analyzed. The data of this research includes customer transaction records from 2018 to 2019, after removing incomplete and missing data, this number has reached 349 records or the company. According to the inquiry from the Ministry of Mining Industry and Trade, there are 51,307 internet shopping and service sites and 36,200,000 internet buyers in the country. Clustering provides a good understanding of customer needs and helps identify potential customers. Dividing customers into sectors also increases the company's income. It is believed that retaining customers is more important than finding new customers. For example, companies can employ marketing strategies specific to a particular segment to retain customers. This study first performed RFM analysis on transaction data and then applied clustering using k-means. Then the results obtained from the methods were compared with each other.

1. Ph.D Candidate, Department of Management, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran.
2. Professor, Department of Management, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
3. Associate Professor, Department of Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.
4. Assistant Professor, Department of Management, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran
*** Corresponding Author Email Address**: M.zarran@scu.ac.ir
DOI: https://doi.org/10.48308/jep.5.2.115

1. Introduction

As there are many E-Commerce Systems which can provide the similar services, it is hard for the nodes to identify the appropriate and trustworthy service providers from these similar service alternatives. Applying a trust model to the study of service selection has been proved to be an effective approach. Trustworthiness of nodes is considered as a service quality metric which reflects the performance characteristics of the services. The approach helps a service requester to decide whether the service provided is trustworthy or not. Most of trust based service selection approaches assess the trustworthiness of service providers based on the direct service experiences or indirect recommendations. In general, if a node has not enough interaction experiences with service providers, then recommendations would be needed for decision making. These recommendations may include a number of untrustworthy or unfair users' feedbacks. Various recommendation filtering methods have been proposed in the area. One method is to use trust threshold to filter untrustworthy recommendations. However, it neglects the most critical service requirements of requester. This greatly impacts on the results of recommendation. Another method is to compute social similarity to filter untrustworthy recommendations. But it is impractical to build an accurate social network graph for all nodes in an IoT network where the number of nodes is too large, and many nodes often join in or leave the network due to their mobility (Wu and Liang, 2023). In fact, E-commerce systems have gained immense popularity and are now implemented in almost all areas of business. These platforms serve as online marketing and promotional channels to reach customers and promote products (Shih et al., 2024). In recent years, the rapid development of e-commerce has led to diverse customer needs. Identifying the various customer segments and establishing relationships with them has become an essential aspect of conducting business and ensuring future revenue streams (Laudien et al., 2024). Retaining valuable existing customers is equally important as attracting new

customers. A valuable approach to gaining deeper insights into customers is segmenting them into groups and analyzing the characteristics of each segment (Liu and Yu, 2023). Well-defined customer segmentation facilitates efficient allocation of marketing resources, enables targeted marketing efforts towards specific customer groups, and fosters long-term customer relationships (Ivens et al., 2024). As a manager, recognizing customer patterns and habits is crucial. The retail industry, in particular, often competes to increase their customer base and maximize profits (Sharma and Sadagopan, 2022). Leveraging data mining algorithms allows businesses to extract valuable knowledge from customer data, both demographic and behavioral, empowering marketing experts to design campaigns that align with customer interests (Wu and Liang, 2023). Data clustering, a fundamental component of data mining, plays a significant role in dividing data elements into distinct groups or clusters (Hashemi et al., 2023). Clustering aims to classify similar elements into the same cluster while grouping dissimilar elements into different clusters based on calculated similarities. As a key machine learning technique, clustering finds applications in various domains, including information granulation (Jendoubi et al., 2023). The main issue of this study is to utilize data mining techniques, specifically the K-Means clustering algorithm and Recency, Frequency, and Monetary (RFM) analysis, to cluster the online consumer market, identify customer needs, and characterize each segment more accurately. Behavioral data is used for clustering as it is readily available and evolves continuously with time and purchase history. RFM analysis, which evaluates customers based on their recency, frequency, and monetary aspects, is employed. A scoring method is developed to assess the scores of these three variables, and subsequently, RFM scores are consolidated to predict future patterns by analyzing present and past customer histories (Smaili and Hachimi, 2023). The calculated values of recency, frequency, and monetary are then used in the K-Means algorithm to cluster the customer base. Analyzing the behavior of each cluster helps identify the customer groups that contribute the most profits to the company (Cheung et

al., 2023). Thorough analysis of the clusters facilitates targeting the right customers and providing them with tailored advertisements and offers.

This study differentiates itself from previous literature on RFM and K-Means by mapping clustering across internet companies, regardless of their field of activity. While other studies focus on specific sectors (Rungruang et al., 2024; Gustriansyah et al., 2022), our study explores the use of big data. While various clustering methods have been used in research, only a limited number of articles have investigated the efficiency of these methods and their ability to predict customer behavior.

In this study, first the theoretical literature and research background will be examined, then the research methodologist will be introduced, and in the fourth part, the findings of the research will be evaluated, and finally, the final part of this study will be the discussion and conclusion.

2. Literature Review

As online shopping transcends geographical boundaries, international studies play a crucial role in providing perspectives and insights into the research methodologies applied. However, there has been limited research on online shopping in Asian countries, particularly the Middle East. While some Asian countries have started investigating this issue in the past decade, specific research on the type of consumers that internet companies in Iran should focus on is lacking. This study aims to fill that gap by conducting research on all types of internet companies in Iran (Shirmohammadi et al., 2024). Various customer segmentation techniques have been proposed in the literature, including lifetime period, demographic-based, propensity-based, and value-based segmentation (Agag et al., 2024). Valentini et al. (2024) suggest a three-dimensional approach to improving customer lifetime value (CLV), customer satisfaction, and understanding customer behavior. They emphasize that consumers are unique, each with their own needs, and segmentation helps identify and meet those demands effectively. Ding et al. (2024) concentrate on customer behavioral factors in their study. They

analyze user data using clustering algorithms to determine purchase behavior within the e-commerce system. The study establishes relationships between three clusters: event type, products, and categories. K-Means clustering is employed to process the collected data and segment customers. Papadimitriou and Tsoukala (2024) use the K-Means clustering algorithm to effectively segment customers who have purchased apparel items. They employ principal component analysis (PCA) for dimensionality reduction and focus on determining associations between customers and brand, product, and price based on their purchase habits. Amin et al. (2023) utilize customer telecommunication big data to target both important and potential churn customers. They employ the Recency, Frequency, Monetary (RFM) analysis technique to generate customer segments and design targeted marketing campaigns based on common characteristics within each segment. Their dataset comprises a combination of structured and unstructured data. Heriqbaldi et al. (2023) highlight the importance of segmentation strategies aligned with consumer needs and organizational marketing strategies in imperfect markets. Nazari Ghazvini et al. (2023) utilizes the RFM model and applies the K-Means algorithm to segment datasets. They validate various dataset clusters based on the Silhouette Coefficient, comparing the results with parameters like sales recency, frequency, and volume. Ilbeigipour et al. (2022) employ the K-Means and self-organizing map (SOM) algorithms to cluster customer characteristics using the RFM model for an insurance dataset. This helps identify customer needs, understand their characteristics, and tailor services accordingly.

Despite the popularity of online shopping, there is a lack of empirical studies focusing specifically on this topic in Iran. Zhang et al. (2024) aimed to identify key or strategic customers using the RFM model. They analyzed registered transactions from Refah Chain Store in Iran and determined the weight of each variable using the fuzzy analytic hierarchy process (AHP). Customers were then clustered using the K-means and Customers were then clustered using the K-means and two-step algorithms, and it was shown that

K-means is the best method according to the Silhouette index.They analyzed registered transactions from Refah Chain Store in Iran and determined the weight of each variable using the fuzzy Analytic Hierarchy Process (AHP). Customers were then clustered using the K-means and two-step algorithms, with K-means proving to be the superior method according to the Silhouette index.Smaili and Hachimi (2024) integrated three patterns, including RFM, fuzzy analytic hierarchy process (FAHP), and K-means, to rank customers in terms of credit. They identified transactions, refunds, and RFM variables as key factors affecting bank customer ratings. Chen et al. (2023) conducted a study to model villagers' intention to adopt e-marketing and performed rural provincial clustering. The research model incorporated the theory of planned behavior (TPB) and rural economy geography, resulting in the proposed geographic model of planned behavior (GeoTPB).

In summary, this study addresses the lack of research on online shopping in Iran by conducting a comprehensive analysis of all types of internet companies. It draws upon various customer segmentation techniques and clustering algorithms to gain insights into customer behavior and preferences. The findings of this study will contribute to the development of tailored marketing strategies and customer-centric approaches in the online business landscape.

3. Methodology

This section provides detailed information on the proposed objective, the algorithm used, and the experimental framework for achieving the desired outcomes of the study.

1.3. Selection of Clustering Algorithm

A cluster refers to a meaningful group of objects that share common characteristics. Clustering is often employed for customer segmentation and further analysis. The literature survey conducted by Cong et al. (2024) and Ma et al. (2024) supports this approach.In this section, clustering is performed in two scenarios, and the results are examined, compared, and

analyzed. The two approaches considered are clustering based on the RFM criterion and clustering using the K-means algorithm.

2.3. RFM Model

The RFM analysis is a widely utilized technique in database marketing for customer segmentation and identification (Smaili and Hachimi, 2023). This model distinguishes important customers from a large dataset based on three attributes: Recency, Frequency, and Monetary value. Recency refers to the time elapsed since the last purchase, Frequency represents the number of purchases within a specific time period, and Monetary denotes the value of the purchase made in the last period (Gustriansyah et al., 2022). To apply the RFM model, customer transaction data from the analysis period is used. The data is processed into the RFM model and subsequently normalized to ensure that the three values (R, F, and M) do not have significant gaps. The normalization method commonly employed is the min-max method as shown in equation 1:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where x_{norm} is the normalized value, x is the data value, x_{min} is the smallest value of the normalized attribute, and x_{max} is the largest value of the normalized attribute.The RFM model proves effective in predicting customers' future purchase behavior based on their past purchases. Firms can identify customers worth targeting by analyzing their past purchase behavior using the RFM model, which is widely applied in database marketing to develop effective marketing strategies. RFM models are often used to target specific customers for marketing programs, such as direct mail, in order to enhance response rates (Bueno et al., 2022). This indicates that RFM helps firms select which customers to target with promotional offers, leading to benefits such as increased response rates, reduced order costs, and greater profitability (Rungruang et al., 2024).

3.3. K-Means Algorithm

The goal of the K-means algorithm is to divide M points in N dimensions into K clusters to minimize the within-cluster sum of squares. The algorithm seeks a "local" optimum solution, ensuring that no movement of a point from one cluster to another will reduce the within-cluster sum of squares. The algorithm requires as input a matrix of M points in N dimensions and a matrix of K initial cluster centers in N dimensions. The number of points in cluster L is denoted by NC (L), and D (I, L) represents the Euclidean distance between point I and cluster L. The general procedure involves searching for a K-partition with the locally optimal within-cluster sum of squares by moving points from one cluster to another (Li et al., 2023).

One of the applications of K-means is customer segmentation. The K-Means clustering algorithm is a prototype-based partition clustering technique that finds the user-specified number of clusters represented by their centroids. K-Means is computationally faster and performs well on large datasets compared to other clustering methods. Another advantage of using K-Means is that it requires only one input parameter 'K' compared to other algorithms (Papadimitriou and Tsoukala, 2024).

Input: Customer Dataset containing 'n' instances and K: the number of clusters,

Output: Customer data partitioned into k clusters.

Algorithm: Initially, depending on the value of k, k random points are chosen as initial centroids, the distances of each data point from the chosen centroids are evaluated using the Euclidean distance, the distance values are compared, and the data point is assigned to the centroid with the shortest Euclidean distance value and repeat the previous steps. The process is stopped if the clusters obtained are the same as those in the previous step.

3.4. Davies- Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) is a method used to measure the validity of a cluster. It maximizes the inter-cluster distance while minimizing the

distance between points within a cluster. A higher inter-cluster distance indicates clear differences between clusters, while a lower intra-cluster distance indicates high similarity among objects within the cluster (Ros et al., 2023). Clustering results obtained from the proposed method are evaluated with the DBI method. This helps determine the correlation between the method of determining centroids based on the Sum of Squared Error and the enhancement of cluster quality based on the obtained DBI value. The steps of the proposed method for determining initial centroids in this paper are as follows (Hosen et al. 2023):

1. Determine and input the number of clusters, 2. Determine the number of centroids to be tested, 3. Choose centroids randomly, 4. Calculate the Sum of Squared Error (SSE) value for each centroid, 5. Select the centroid with the minimum SSE from all calculated centroids. Repeat from step 2 until the number of tested centroids is equal to the input number of centroids. The centroid with the minimum Sum of Square Error is used as the initial centroid in the clustering step, 6. Calculate the distance between data points and centroids using the Euclidean distance formula, 7. Classify data with the minimum distance from step 6, 8. Calculate and obtain new centroids based on the mean value from the membership of each cluster and 9. If the new centroid value is equal to the input centroid, the clustering is stopped; otherwise, go back to step 6.

3.5. Data analysis methods

It is an empirical research study that follows a descriptive research design. The data for this research has been collected from primary sources. The primary data includes customers' transaction records. The sample size of the study consisted of 349 Internet companies based in Iran. The sample elements in this research are individuals over the age of 18 who have made at least one online purchase. The data was collected over a period of one year, from 2022 to 2023. Convenience sampling was used as the sample procedure in this study. Two databases were utilized, which include personal profiles of customers and transaction data. The figure below illustrates these databases.

Fig 1. Research methodology (Design: Authors)

4. Findings

1.4. Measurement

In this study, the statistical population includes data related to various parameters such as gender, latency (R), frequency (F), and monetary value (M) of online sales. Age and education data were only available for some stores and were omitted in the clustering process.The data is organized in a cubic matrix with dimensions of 349 × 4 × 466. This matrix represents 349 companies and 4 available parameters (gender, R, F, and M). Each company has a maximum of 466 records. The summary of the general criteria of the statistical population is presented in the table below.

Table 1. General criteria of the statistical population

Maximum Purchase			Population			No. of Companies
Value (Rials)	Frequency	Latency	Men	Women	Total	
1791900000	45	3187	82457	80177	162634	349

Source: Research finding

According to the statistical population, a preliminary categorization can be made to gain an initial understanding. Based on the available parameters, the following five categories are identified among the companies: Foodstuffs, Cosmetics, Luxury accessories, Industrial supplies and their accessories and Sanitary ware, detergents and clothing. The initial categorization of companies into the aforementioned categories is presented in Table 2.4.

Table 2. Initial division of companies

Sanitary ware, detergents and clothing	Industrial supplies and their accessories	Luxury accessories	Cosmetics	Foodstuffs	
96	23	113	65	52	No. of Companies
3.8732	460.4225	18729000	96.6761	0	Average resolution criteria

Source: Research finding

4.2. Data clustering and analysis of results

The following table presents the approaches used in data clustering, which includes clustering based on the RFM criterion:

Table 3. Clustering approach

Clustering by criteria RFM and kmeans
For all customers of the whole company
For a company
Analysis of RFM and kmeans results

Source: Research finding

4.3. Clustering based on RFM and kmeans criteria

Clustering based on RFM criteria and kmeans for all customers of all companies. In this section, all customers and their information are combined into one dataset, and this dataset is then clustered using the RFM criteria and K-means algorithm. The total customer data is divided into 5 categories based on the R (Recency), F (Frequency), and M (Monetary value) indicators, resulting in 125 categories in a 3D space. The following tables present the overall results and comparison of criteria.

Table 4. Clustering along the R axis based on the RFM criterion

SSE	No. of women	No. of men	center	Cluster
0	0	32527	0	1
0	0	32527	0	2
0	0	32526	0	3
0	0	32527	0	4
35984	14560	17967	151.33	5

Source: Research finding

Table 5. Clustering along the F axis based on the RFM criterion

SSE	No. of women	No. of men	center	Cluster
81804	14555	17972	1.2971	1
0	0	32527	0	2
0	0	32526	0	3
0	0	32527	0	4
0	0	32527	0	5

Source: Research finding

Table 6. Clustering along the M axis based on the RFM criterion

SSE	No. of women	No. of men	center	Cluster
30266	14555	17972	4315900	1
0	0	32527	0	2
0	0	32526	0	3
0	0	32527	0	4
0	0	32527	0	5

Source: Research finding

Clustering based on Kmeans algorithm for all customers of all companies

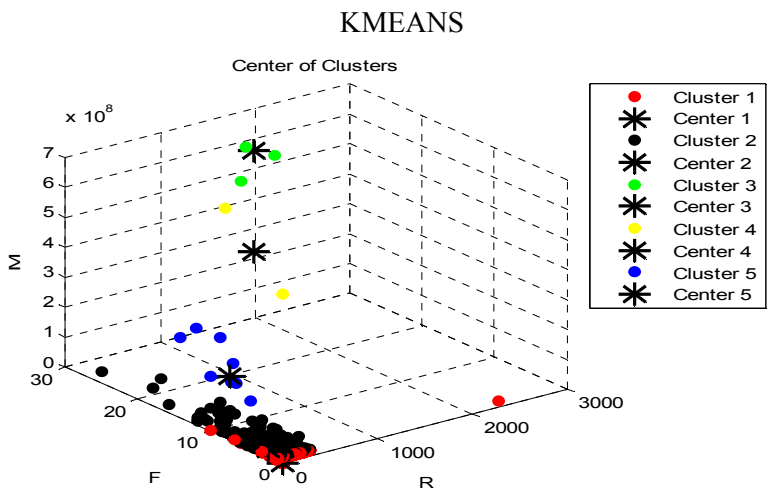


Fig 2. Clustering Kmeans algorith for all companies
Source: Research finding

Clustering based on RFM and kmeans criteria for for the category of companies. In this section, all the customers and information of a company, as examples of all 5 groups of food, cosmetics, luxury goods, detergents, and clothing, are analyzed separately by RFM and KMEANS methods, the results of which are presented in the following tables and figures.

Foodstuffs

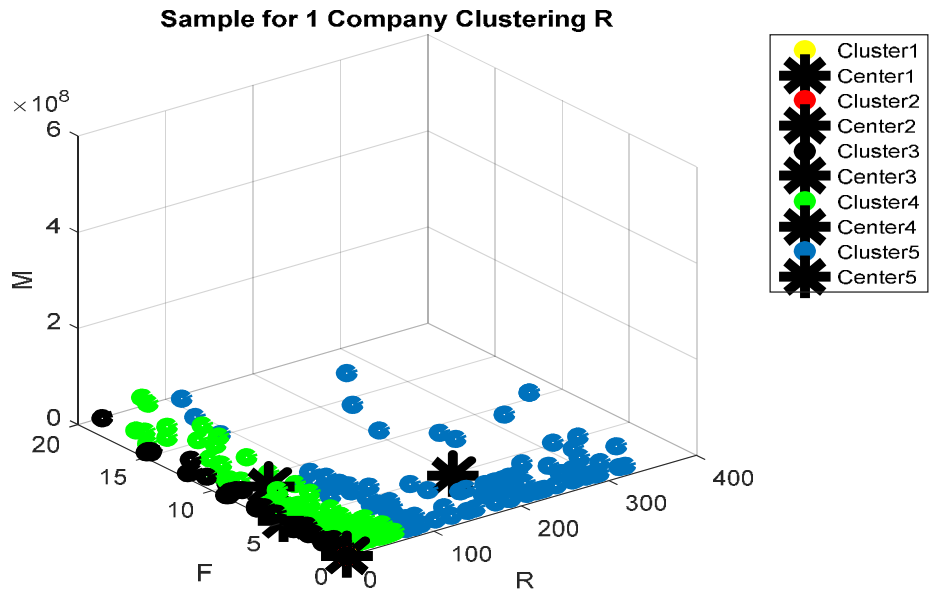


Fig 3. Clustering along the R axis based on the RFM criterion for foodstuffs

Source: Research finding

Table 7. Clustering along the R axis based on the RFM criterion for foodstuffs

Clustering along the R axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0	0	93	0	1
0	0	93	0	2
2283	27	67	4.0106	3
26896	44	49	40.892	4
3265	50	43	193.31	5

Source: Research finding

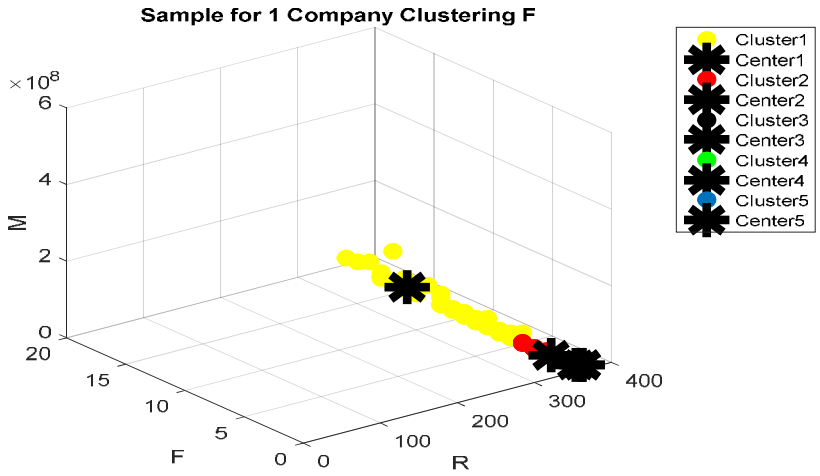


Fig 4. Clustering along the F axis based on the RFM criterion for foodstuffs
Source: Research finding

Table 8. Clustering along the F axis based on the RFM criterion for Foodstuffs

Clustering along the F axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
8736.3	47	46	14.744	1
118.67	47	46	2.6667	2
23.33	68	26	0.54225	3
0	93	0	0	4
0	93	4	0	5

Source: Research finding

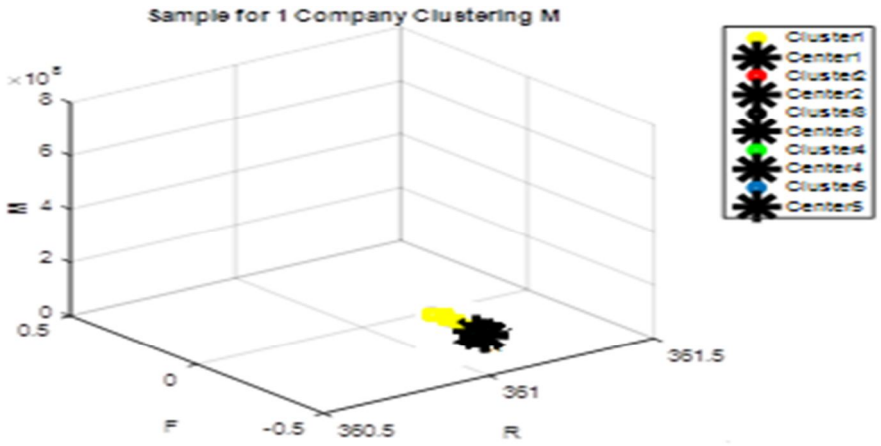


Fig 5. Clustering along the M axis based on the RFM criterion for foodstuffs
Source: Research finding

Table 9. Clustering along the M axis based on the RFM criterion for foodstuffs

Clustering along the M axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
1.8594	47	46	1.2868	1
2.9823	57	41	1.3925	2
4.4252	63	31	2.2103	3
0	93	0	0	4
0	93	0	0	5

Source: Research finding

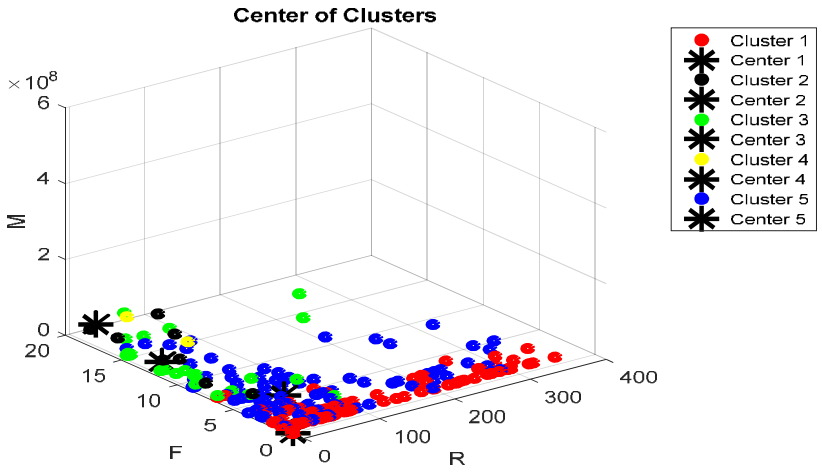
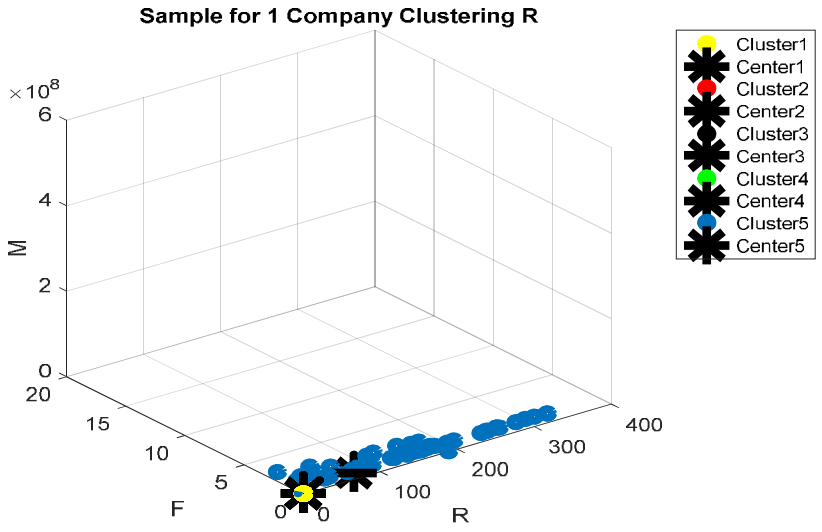


Fig 6. Clustering kmeans for foodstuffs, Source: Research finding



Cosmetic materials and accessories

Fig 7. Clustering along the R axis based on the RFM criterion Cosmetic materials and accessories, Source: Research finding

Table 10. Clustering along the R axis based on the RFM criterion Cosmetic materials and accessories

Clustering along the R axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0	47	93	0	1
0	0	93	0	2
0	0	94	0	3
0	0	93	0	4
9.3141	27	66	74.892	5

Source: Research finding

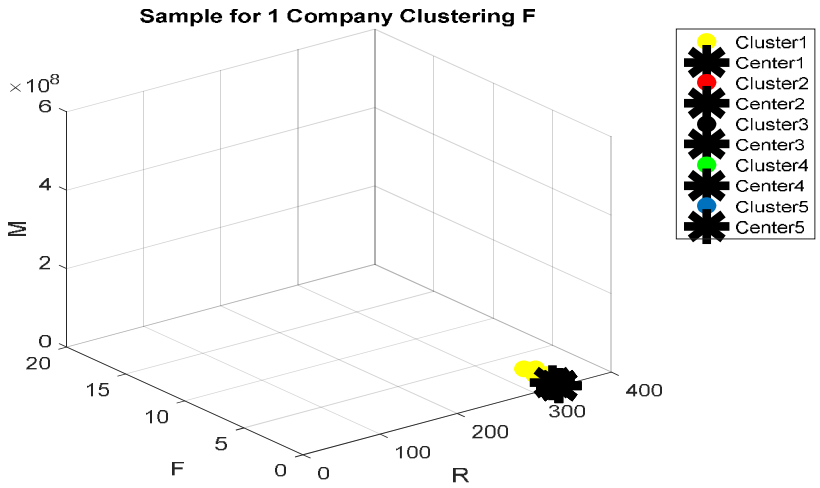


Fig 8. Clustering along the F axis based on the RFM criterion Cosmetic materials and accessories, Source: Research finding

Table 11. Clustering along the F axis based on the RFM criterion Cosmetic materials and accessories

Clustering along the f axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**Center**	**Cluster**
46.473	27	66	0.5914	1
0	0	93	0	2
0	0	94	0	3
0	3	93	0	4
0	0	92	0	5

Source: Research finding

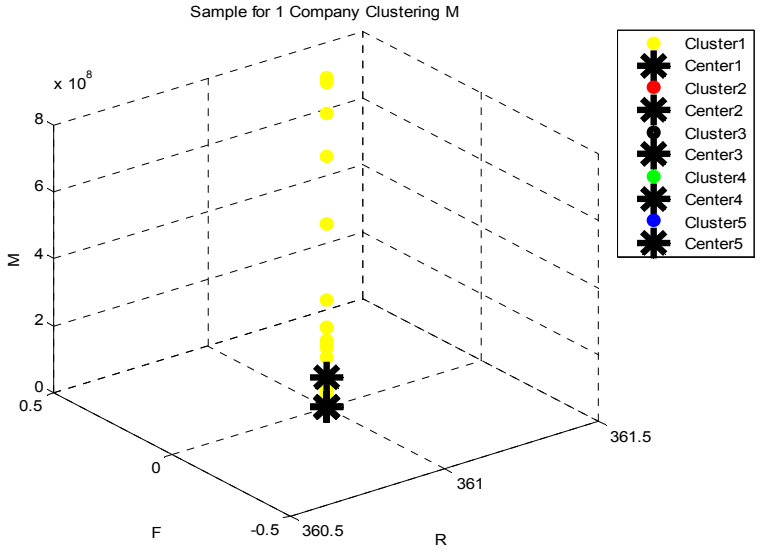


Fig 9. Clustering along the M axis based on the RFM criterion Cosmetic materials and accessories, Source: Research finding

Table 12. Clustering along the M axis based on the RFM criterion Cosmetic materials and accessories

Clustering along the M axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0	47	93	0	1
0	0	93	0	2
0	0	94	0	3
0	0	93	0	4
9.3141	27	66	74.892	5

Source: Research finding

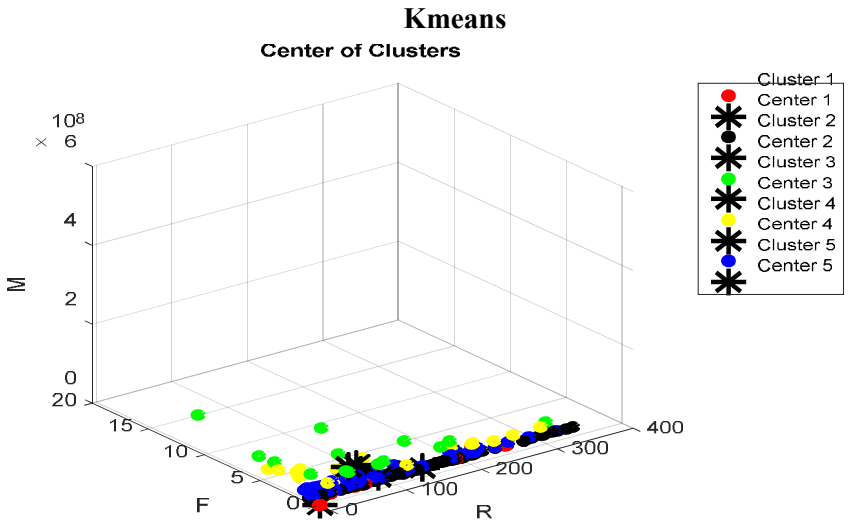


Fig 10. Clustering kmeans for Cosmetic materials and accessories
Source: Research finding

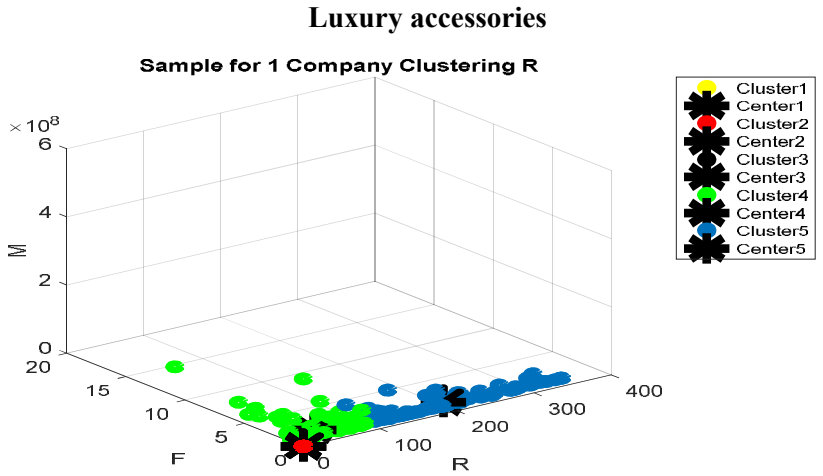


Fig 11. Clustering along the R axis based on the RFM criterion Luxury accessories
Source: Research finding

Table 13. Clustering along the R axis based on the RFM criterion Luxury accessories

Clustering along the R axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**Center**	**Cluster**
0	0	93	0	1
0	0	93	0	2
1.0189	0	94	0	3
0	0	93	35.634	4
4.0101	0	93	199.25	5

Source: Research finding

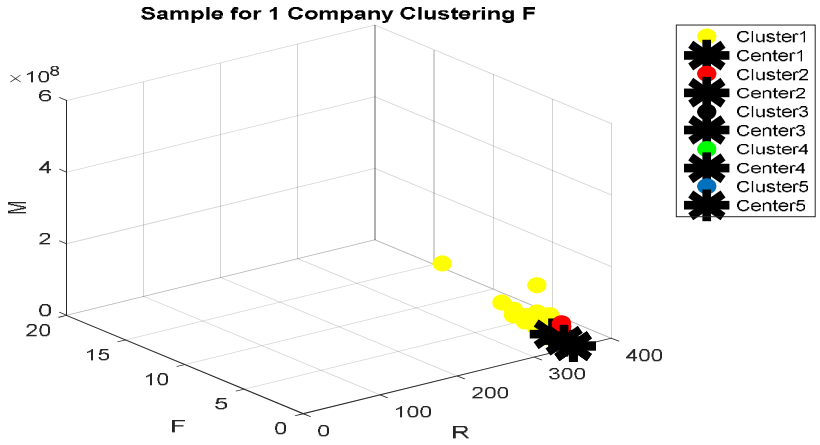


Fig 12. Clustering along the F axis based on the RFM criterion Luxury accessories
Source: Research finding

Table 14. Clustering along the F axis based on the RFM criterion Luxury accessories

Clustering along the F axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
194.26	1	93	1.7242	1
18.28	0	93	0.73118	2
0	0	94	0	3
0	0	93	0	4
0	2	92	74.892	5

Source: Research finding

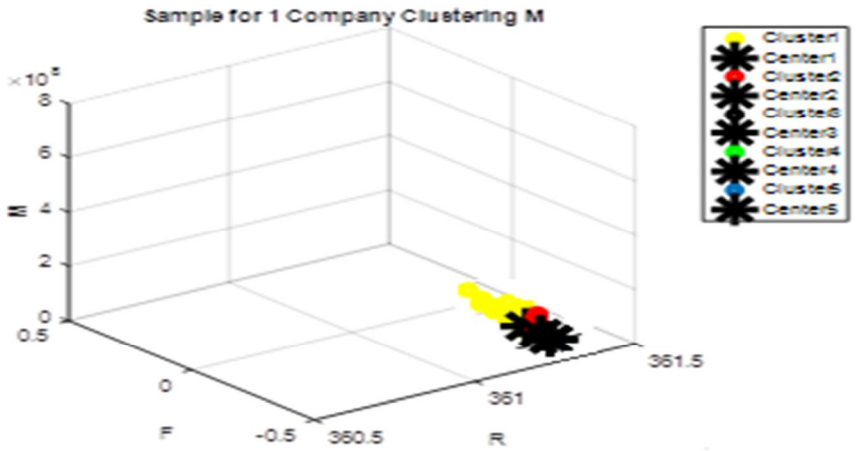


Fig 13. Clustering along the M axis based on the RFM criterion Luxury accessories
Source: Research finding

Table15. Clustering along the M axis based on the RFM criterion Luxury accessories

Clustering along the M axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
3.533	47	93	1.8546	1
2.9338	0	93	2.494	2
0	0	94	0	3
0	0	93	0	4
0	27	93	0	5

Source: Research finding

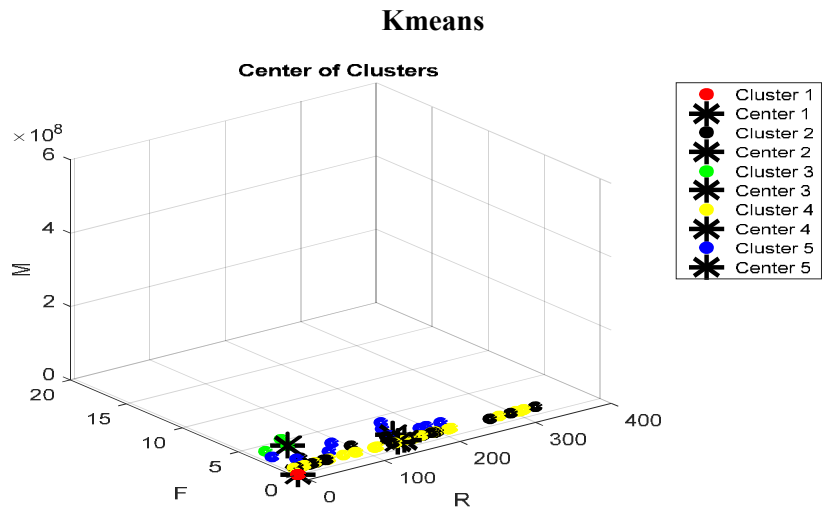


Fig 14. Clustering kmeans for Luxury accessories, Source: Research finding

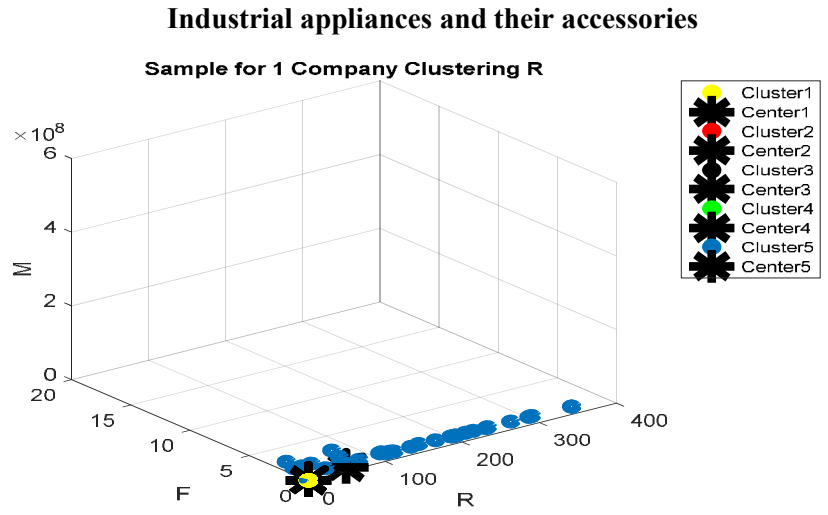


Fig 15. Clustering along the R axis based on the RFM criterion Industrial appliances and their accessories, Source: Research finding

Table 16. Clustering along the R axis based on the RFM criterion Industrial appliances and their accessories

Clustering along the R axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0	1	93	0	1
0	2	94	0	2
0	0	93	0	3
0	0	93	0	4
8.4595	5	88	56.355	5

Source: Research finding

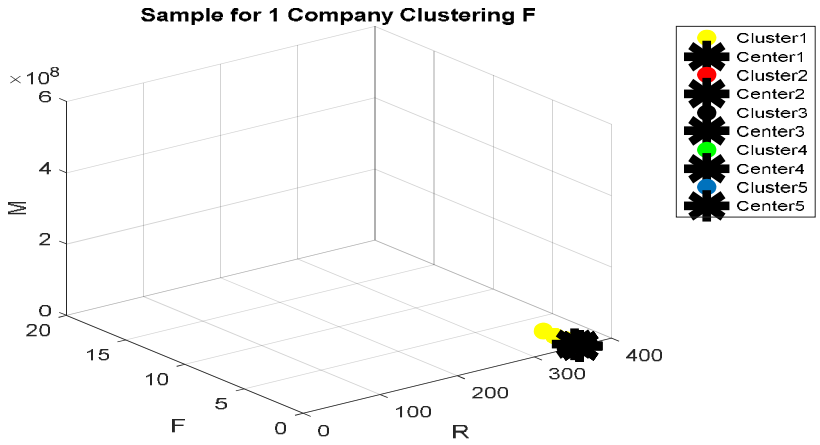


Fig 16. Clustering along the F axis based on the RFM criterion Industrial appliances and their accessories, Source: Research finding

Table 17. Clustering along the F axis based on the RFM criterion Industrial appliances and their accessories

Clustering along the F axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**The center of the cluster**	**Cluster**
42.925	5	88	0.44860	1
0	0	93	0	2
0	20	94	0	3
0	0	93	0	4
0	0	94	74.892	5

Source: Research finding

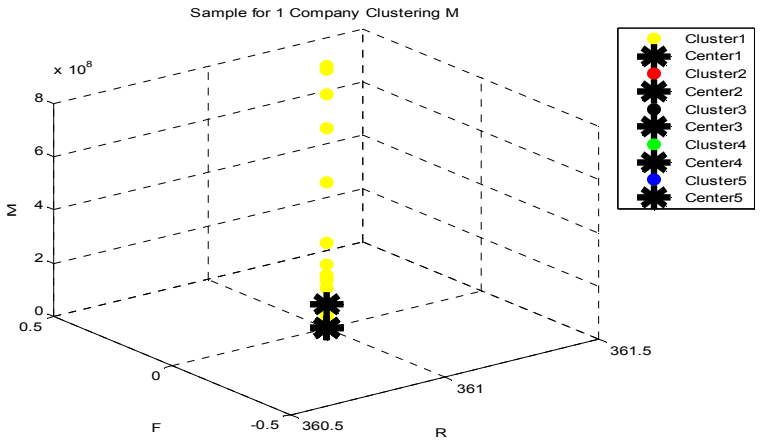


Fig 17. Clustering along the M axis based on the RFM criterion Industrial appliances and their accessories, Source: Research finding

Table18. Clustering along the M axis based on the RFM criterion Industrial appliances and their accessories

Clustering along the M axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0.42925	5	88	0.4486	1
0	0	93	0	2
0	0	94	0	3
0	0	93	0	4
0	2	96	74.892	5

Source: Research finding

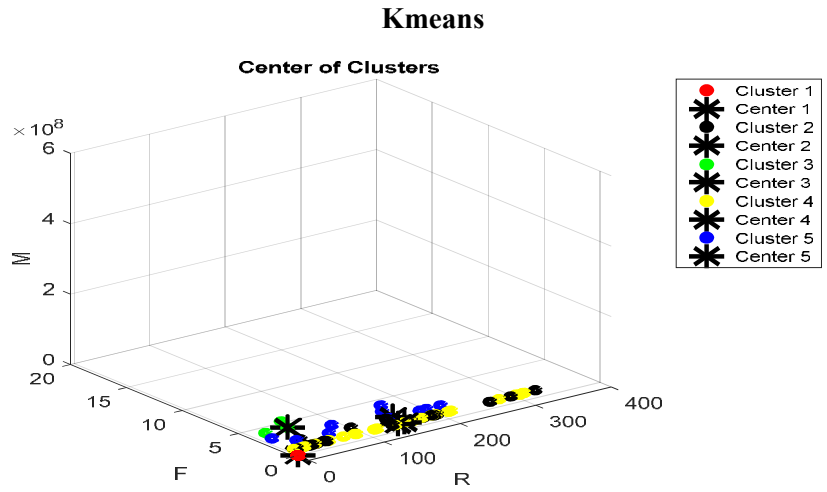


Fig 18. Clustering kmeans for Industrial appliances and their accessories
Source: Research finding

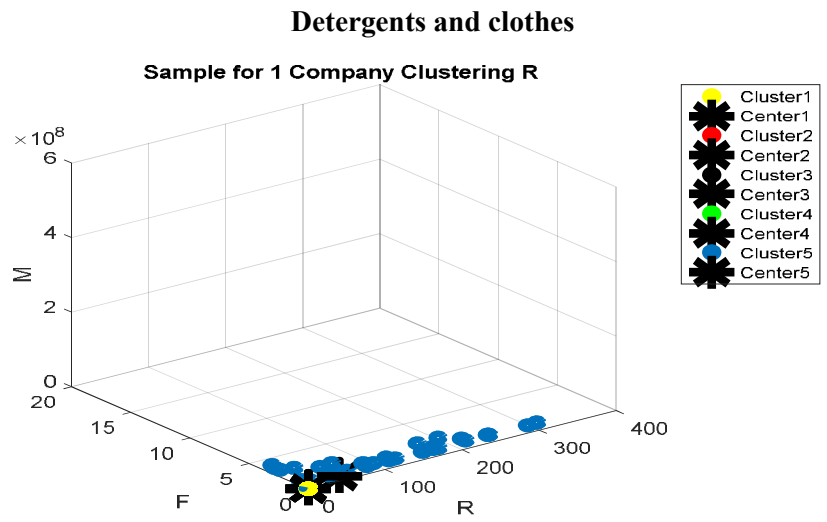


Fig 19. Clustering along the R axis based on the RFM criterion Detergents and clothes
Source: Research finding

Table 20. Clustering along the R axis based on the RFM criterion Detergents and clothes

Clustering along the R axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0	3	93	0	1
0	0	93	0	2
0	0	94	0	3
0	0	93	0	4
6.7557	11	82	49.634	5

Source: Research finding

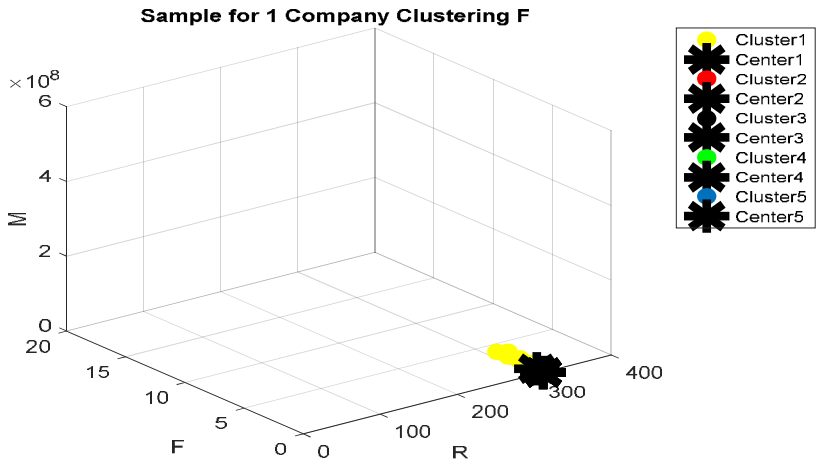


Fig 20. Clustering along the F axis based on the RFM criterion Detergents and clothes

Source: Research finding

Table 21. Clustering along the F axis based on the RFM criterion Detergents and clothes

Clustering along the F axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**center**	**Cluster**
0	11	82	52.686	1
0	0	93	0	2
0	0	94	0	3
0	0	83	0	4
961.186	20	93	0	5

Source: Research finding

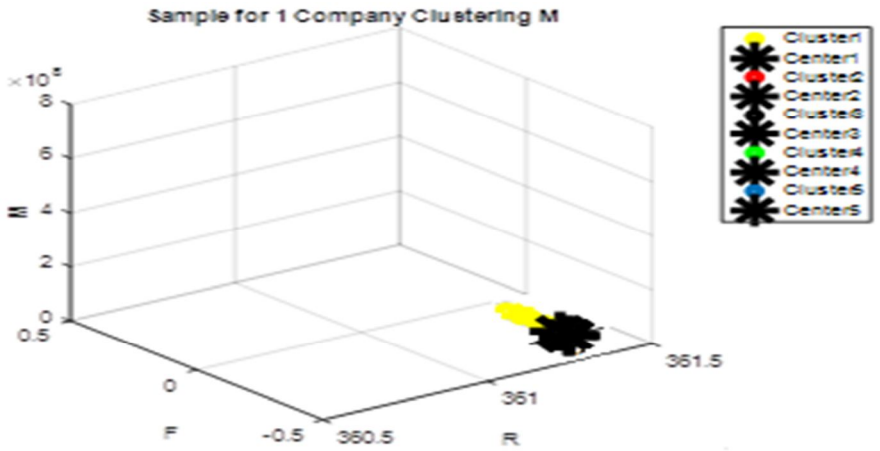


Fig 21. Clustering along the M axis based on the RFM criterion Detergents and clothes (Source: Research finding)

Table 22. Clustering along the M axis based on the RFM criterion Detergents and clothes

Clustering along the M axis based on the RFM criterion for Company				
SSE	**No. of women**	**No. of men**	**Center**	**Cluster**
61.183	11	82	52.688	1
0	0	93	0	2
0	0	94	0	3
0	3	93	0	4
0	0	66	0	5

Source: Research finding

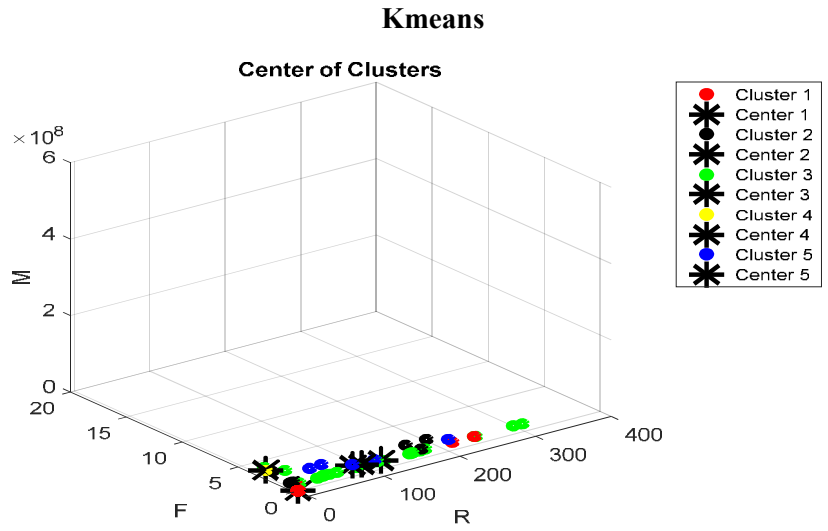


Fig 22. Clustering kmeans for Detergents and clothes (Source: Research finding)

5. Discussion and conclusion

E-commerce (electronic commerce) is the activity of electronically buying or selling products on online services or over the Internet. E-commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. E-commerce is the largest sector of the electronics industry and is in turn driven by the technological advances of the semiconductor industry. The term was coined and first employed by Robert Jacobson, Principal Consultant to the California State Assembly's Utilities & Commerce Committee, in the title and text of California's Electronic Commerce Act, carried by the late Committee Chairwoman Gwen Moore (D-L.A.) and enacted in 1984. E-commerce typically uses the web for at least a part of a transaction's life cycle although it may also use other technologies such as e-mail. Typical e-commerce transactions include the purchase of products (such as books from Amazon) or services (such as music downloads in the form of

digital distribution such as the iTunes Store). There are three areas of e-commerce: online retailing, electronic markets, and online auctions. The findings of the article were in line with the results of the research Zhang et al. (2024), Agag et al., (2024), Papadimitriou and Tsoukala (2024), Chen et al. (2023), Amin et al. (2023), Heriqbaldi et al. (2023), Nazari Ghazvini et al. (2023), Ilbeigipour et al. (2022) but it did not have much alignment with the research findings of Smaili and Hachimi (2024), Valentini et al. (2024), Ding et al. (2024). E-commerce is supported by electronic business. The existence value of e-commerce is to allow consumers to shop online and pay online through the Internet, saving the time and space of customers and enterprises, greatly improving transaction efficiency, especially for busy office workers, and also saving a lot of valuable time. E-commerce businesses may also employ some or all of the following:

- Online shopping for retail sales direct to consumers via web sites and mobile apps, conversational commerce via live chat, chatbots, and voice assistants;
- Providing or participating in online marketplaces, which process third-party business-to-consumer (B2C) or consumer-to-consumer (C2C) sales;
- Business-to-business (B2B) buying and selling;
- Gathering and using demographic data through web contacts and social media;
- B2B electronic data interchange;
- Marketing to prospective and established customers by e-mail or fax (for example, with newsletters);
- Engaging in pretail for launching new products and services;
- Online financial exchanges for currency exchanges or trading purposes.

Based on the analysis and results obtained from the clustering methods, the following conclusions can be drawn:

1. K-means Method:
   - The K-means method has achieved the most optimal distribution of cluster centers and has shown effective performance in dividing the data into different categories compared to other methods.

- Based on the Sum of Squared Error (SSE) criterion, the K-means method outperformed other methods when considering a group of companies or the entire dataset.
- The K-means method also achieved the best results based on the Davies-Bouldin Index (DBI), which is an important clustering evaluation index, for both data sets.

2. RFM Method:

- The RFM method, when applied to clustering a company's customers, resulted in only one cluster for all companies. As a result, it is not suitable for separating the customers into the desired five clusters.
- However, when analyzing customers of a specific company using the RFM method, valuable insights can still be gained. Most customers were placed in groups related to food and cosmetics, indicating their preferences and purchase behavior.
- Recommendations for the company based on RFM clustering include considering solutions such as providing discounts on occasions related to women, increasing the variety of products in the food and cosmetics categories, and offering special discounts to customers with high monetary value purchases.

In conclusion, the K-means method has shown superior performance in terms of proper distribution of cluster centers and overall data analysis compared to other methods. However, the RFM method can still provide valuable insights when analyzing customers of a specific company. By considering the clustering results and customer preferences, companies can tailor their strategies to encourage customers to buy more. Strategies such as providing discounts, increasing product variety, and targeting specific customer segments can be effective in improving sales performance. In the context of K-means clustering, all five expected categories were successfully formed, making it a suitable method for clustering and analyzing people's online shopping behavior. Surprisingly, contrary to initial assumptions, the analysis revealed that the most valuable customers were predominantly men.

For example, in the first cluster, there were 319 male customers who were primarily online food buyers. To improve the sales situation and attract more customers, the company should consider increasing diversity in the food sector, offering special discounts, and providing free delivery services. In the weighted RFM clustering, the majority of customers belonged to the first category, which comprised food buyers, mainly men. Therefore, the same strategies mentioned earlier should be considered to target this customer segment. In K-means clustering, customers were distributed into five clusters, with the majority falling into the second category: buyers of cosmetics, predominantly women. In light of this, the company should consider implementing policies such as providing discounts on occasions related to women and expanding the variety of products in the food and cosmetics categories. The research objectives have been achieved, and the company can now develop a model and strategy for increasing sales based on the gender composition and type of purchases made by customers. By tailoring the customer purchase model and devising persuasive strategies to encourage more purchases, the company can capitalize on the differences in customer behavior and their distribution across clusters. In the RFM method, when clustering a company's customers, the majority of customers belong to clusters 4 and 5 along the R axis, indicating high purchase frequency by men. Furthermore, when clustering customers based on the F and M axes, most customers are placed in groups 1 and 2, representing food and cosmetics, respectively, and indicating the least purchase delay. To summarize, based on the clustering results, the company can develop a comprehensive model and strategy to drive sales growth. Taking into account the gender composition, customer behavior, and the RFM status of their purchases, the company can target specific customer segments and present tailored strategies to persuade and encourage them to make more purchases. Therefore, for Company 1, the clustering analysis based on the RFM method indicates that most customers are placed in groups 1 and 2, representing food and cosmetics. Since these customers make the most

purchases and do so regularly, the company should focus on strategies to encourage them to buy more. This can include providing discounts on occasions related to women, increasing the variety of products in the food and cosmetics categories, and implementing targeted marketing campaigns in these areas. Additionally, the company can offer special discounts to customers with the highest monetary value of purchases (M) to further incentivize their loyalty and increase sales.

In the case of K-means clustering, all five expected categories are formed, making it a more suitable method for analyzing people's online shopping behavior. Contrary to initial assumptions, the analysis reveals that the most valuable customers are predominantly men. For example, in the first cluster, there are 319 male customers who are online food buyers. To improve the company's sales situation and attract more customers, strategies such as increasing diversity in the food sector, offering special discounts, and providing free delivery services can be implemented. In the weighted RFM clustering, the majority of customers belong to the first category, which comprises food buyers and men. Therefore, the same strategies mentioned earlier should be considered. In the K-means clustering results, the majority of customers belong to the second category, which represents buyers of cosmetics and predominantly women. Thus, policies like providing discounts on occasions related to women and expanding the variety of products in the food and cosmetics categories should be considered by the company. By achieving the research objectives, the company can now establish a model and strategy to drive more sales. This can be done by considering the gender composition and type of purchases made by customers. The customer purchase model and strategies to persuade and encourage them to make more purchases can be tailored based on whether they are male or female and their RFM status. Furthermore, the analysis provides insights into customer behavior and their distribution across clusters. This information can help the company understand how customers make purchases and guide decision-making processes.

Funding

This study received no financial support from any organization.

Authors' contributions

All authors had contribution in preparing this paper.

Conflicts of interest

The authors declare no conflict of interest

References

Agag, G., Shehawy, Y., Almoraish, A., Eid, R., Lababdi, H., Labben, T., & Abdo, S. (2024). Understanding the relationship between marketing analytics, customer agility, and customer satisfaction: A longitudinal perspective. Journal of Retailing and Consumer Services, (77), 21-34.

Amin, A., Adnan, A., & Anwar, S. (2023). An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes. Journal of Applied Soft Computing, (137), 691-714.

Bueno, I., Velasco, J., Carrasco, R., & Herrera-Viedma, E. (2022). A geospatial model of RFM analysis: An application to tourism in the Iberian Peninsula. Journal of Procedia Computer Science, (214), 1047-1062.

Chen, Y., Liu, L., Zheng, D., & Li, B. (2023). Estimating travellers' value when purchasing auxiliary services in the airline industry based on the RFM model. Journal of Retailing and Consumer Services, (74), 957-974.

Cheung, C., Lee, W., Wang, M., & To, S. (2023). A multi-perspective knowledge-based system for customer service management. Journal of Expert Systems with Applications, 24(4), 457-470.

Cong, L., Ding, H., Xie, N., & Wie, X. (2024). Space delay-tolerant network routing algorithm based on node clustering and social attributes. Journal of Ad Hoc Networks, (155), 649-667.

Ding, K., Gong, X., Huang, T., & Choo, W. (2024). Recommend or not: A comparative analysis of customer reviews to uncover factors influencing explicit online recommendation behavior in peer-to-peer accommodation. Journal of European Research on Management and Business Economics, 30(1), 105-12.

Gustriansyah, R., Ermatita, E., & Palupi Rini, D. (2022). An approach for sales forecasting. Journal of Expert Systems with Applications, (207), 161-179.

Hashemi, E., Gholian-Jouybari, F., & Hajiaghaei-Keshteli, M. (2023). A fuzzy C-means algorithm for optimizing data clustering. Journal of Expert Systems with Applications, (227), 39-55. [In Persian]

Heriqbaldi, U., Jayadi, A., Erlando, A., Samudro, B., Widodo, W., & Esquivias, M. (2023). Survey data on organizational resources and capabilities, export marketing strategy, export competitiveness, and firm performance in exporting firms in Indonesia. Journal of Data in Brief, (48), 443-460.

Hosen, A., Hoshen Moz, S., Kabir, S., Galib, S., & Adnan, N. (2023). Enhancing Thyroid Patient Dietary Management with an Optimized Recommender System based on PSO and K-means. Journal of Procedia Computer Science, (230), 688-697.

Ilbeigipour, S., Albadvi, A., & Noughabi, E. (2022). Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making. Journal of Informatics in Medicine Unlocked, (32), 1-19.

Ivens, B., Kasper-Brauer, K., Leischnig, A., & Thornton, S. (2024). Implementing customer relationship management successfully: A configurational perspective. Journal of Technological Forecasting and Social Change, (199), 41-58.

Jendoubi, S., Baelde, A., & Tran, T. (2023). SSFuzzyART: A Semi-Supervised Fuzzy ART through seeding initialization and a clustered data generation algorithm to deeply study clustering solutions. Journal of Array, (19), 611-627.

Liu, S., & Yu, Z. (2023). Modeling and efficiency analysis of blockchain agriculture products E-commerce cold chain traceability system based on Petri net. Journal of Heliyon, 9(11), 217-231.Laudien, S., Reuter, U., Garcia, F., & Botella-Carrubi, D. (2024). Digital advancement and its effect on business model design: Qualitative-empirical insights. Journal of Technological Forecasting and Social Change, (200), 278-295.

Ma, F., Wang, C., Huang, J., Zhong, Q., & Zhang, T. (2024). Key Grids based Batch-Incremental CLIQUE Clustering Algorithm Considering Cluster Structure Changes. Journal of Information Sciences, (128), 1-17.

Nazari Ghazvini, S., Vakil Alroaia, Y., & Baharon, R. (2023). The Study of Electroencephalography in Neuromarketing Research, Consumer Behavior and Performance Method: A Systematic Analysis. Journal of System Management, 9(4), 185-204.

Papadimitriou, A., & Tsoukala, V. (2024). Evaluating and enhancing the performance of the K-Means clustering algorithm for annual coastal bed evolution applications. Journal of Oceanologia, (164), 1-16.

Ros, F., Raid, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. Journal of Neurocomputing, (528), 1-22.

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. Journal of Expert Systems with Applications, (237), 94-111.

Sharma, S., & Sadagopan, P. (2022). Influence of conditional holoentropy-based feature selection on automatic recommendation system in E-commerce sector. Journal of King Saud University - Computer and Information Sciences, 34(8), 5564-5577.

Shin, I., Silalahi, A., & Eunike, I. (2024). Engaging audiences in real-time: The nexus of socio-technical systems and trust transfer in live streaming e-commerce. Journal of Computers in Human Behavior Reports, (13), 71-86.

Shirmohammadi, Y., Abiyaran, P., & Peters, M. (2024). Virtual reality (VR) alongside Social Media Marketing Activities (SMMAs) as a solution for Management Information Systems (MIS). Journal of System Management, 10(1), 133-154.

Smaili, M., & Hachimi, H. (2023). New RFM-D classification model for improving customer analysis and response prediction. Ain Shams Engineering Journal, 14(12), 1-23.

Wu, X., & Liang, J. (2023). Study on trust evaluation and service selection for Service-Oriented E-Commerce systems in IoT environments. Egyptian Informatics Journal, 24(2), 257-263.

Valentini, T., Roederer, C., & Casteran, H. (2024). From Redesign to revenue: Measuring the effects of servicescape remodeling on customer lifetime value. Journal of Retailing and Consumer Services, (77), 66-82.

Zhang, H., Li, J., Zhang, J., & Dong, Y. (2024). Speeding up k-means clustering in high dimensions by pruning unnecessary distance computations. Journal of Knowledge-Based Systems, (284), 180-197.