Supporting Retail Sale by Using a Combination of RFM model and Clustering Agorithm

Thuy Nguyen Thi Thu[1]

[1] Informatics Department, Thuongmai University, Hanoi, Vietnam

DOI: https://doi.org/10.56293/IJMSSSR.2026.6108

IJMSSSR 2026
VOLUME 8
ISSUE 2 MARCH - APRIL **ISSN: 2582 – 0265**

Abstract: In the increasingly competitive retail sector, understanding customer behavior to formulate personalized marketing and sales strategies is imperative. This study proposes an automated customer segmentation approach based on the analysis of multidimensional transaction data from global shopping centers, specifically utilizing a comprehensive retail dataset from Istanbul spanning from 2021 to 2023. The core methodology integrates the RFM (Recency, Frequency, Monetary) analytical framework with an unsupervised machine learning clustering algorithm. The RFM model is applied to quantify individual buyer engagement by measuring the time elapsed since the last purchase (Recency), the total number of transactions (Frequency), and the total expenditure (Monetary). Subsequently, a clustering algorithm is employed to analyze these variables, thereby uncovering underlying behavioral patterns and grouping customers with similar characteristics into meaningful structures.
The results demonstrate the successful partitioning of the customer base into three strategic segments: Regular Customers (39.65%), Infrequent Customers (41.08%), and Potential Customers (19.26%). The quality and distinctness of these clusters are validated through robust evaluation metrics, achieving a Silhouette Score of 0.339, a Calinski-Harabasz Score of 68829.85, and a Davies-Bouldin Score of 1.11, indicating well-defined and adequately separated groupings. This practical implementation provides businesses with a powerful tool to comprehensively understand consumer profiles, thereby optimizing targeted outreach campaigns and enhancing overall customer retention strategies.

Keywords: RFM, Customers Segmentation, K-means, Clustering

1. Introduction

To maintain sustainability and enhance competitive advantage, enterprises must develop long-term strategies aimed at retaining loyal customers and identifying potential ones. Corporate clients typically possess diverse personal characteristics, behaviors, and consumption preferences. Consequently, applying a uniform marketing strategy to an entire customer base yields low efficiency and may overlook specific segments that could be highly profitable (Manero et al., 2018; Khong, 2021). Therefore, supporting personalization in corporate customer care is a pivotal activity, as it not only improves communication effectiveness but also mitigates the risk of customer churn.

With the exponential growth of transactional data, managing customer information creates significant pressure regarding the accurate classification of clients into distinct groups. It is crucial to analyze the needs of each customer segment to facilitate market segmentation and evaluate various target audiences within the system (Sitompul et al, 2019; Rachmawati et al., 2020). This approach enables the implementation of specialized marketing methods tailored to different clients, thereby enhancing customer loyalty and satisfaction.

The objective of this study is to perform customer data analysis and implement segmentation based on historical purchasing transactions. The proposed segmentation in this research utilizes the RFM model combined with the K-Means clustering algorithm. The objectives and contributions of this research are:

- Primary Objective: To utilize the RFM model to categorize the corporate customer base into three distinct groups: Potential, Regular, and Infrequent customers.
- Methodological Integration: To combine the RFM (Recency, Frequency, Monetary) method with Unsupervised Machine Learning (clustering) to uncover meaningful structures and analyze purchasing

behaviors. The quality of the clusters is validated by technical metrics, including the Silhouette Score (0.339), Calinski-Harabasz Index (68829.85), and Davies-Bouldin Index (1.11), demonstrating that the customer segments are well-separated and align closely with empirical data.

- Behavioral Analysis: To provide a detailed analysis of the behaviors of different customer groups. For instance, Cluster 0 (Regular Customers) represents those with moderate spending but high shopping frequency; Cluster 1 (Infrequent Customers) is characterized by few transactions, low spending, and long intervals since the last purchase; and Cluster 2 (Potential Customers) comprises clients with very high expenditure, substantial order values, and continuous engagement.

The rest of the sections are organized as follows. Section 2 includes the related works, Section 3 introduces the research framwork, Section 4 shows the experimental results and including the Discusion of Data Visualization, Conclusion is included in Section 5.

2. Related Works

In the context of digital transformation and intensifying competition, customer data has emerged as a strategic asset, enabling enterprises to optimize marketing activities, enhance user experience, and increase customer lifetime value. Analyzing customer sales data allows organizations to gain a profound understanding of purchasing behaviors, loyalty levels, and the economic value of distinct customer segments. Among various behavioral analysis methods, the RFM (Recency, Frequency, Monetary) model, integrated with clustering techniques, has proven to be an effective tool for customer segmentation and strategic decision support. According to Shirole et al. (2021), customer segmentation is based on discovering significant differences to categorize clients into target groups. This modeling allows organizations to target specific segments, facilitating more efficient resource allocation and maximizing cross-selling and up-selling opportunities. Furthermore, segmentation enhances customer service while boosting loyalty and retention rates.

The RFM model specifically evaluates customers across three dimensions: the time since the last purchase (Recency), the frequency of transactions (Frequency), and the total expenditure (Monetary). This methodology enables businesses to identify high-value groups, potential leads, and customers at risk of churn. Recent studies have confirmed that RFM provides a critical foundation for behavioral analysis based on historical transactional datasets and can be effectively integrated with unsupervised machine learning algorithms like K-means to improve segmentation accuracy. The K-Means algorithm is a prominent clustering technique utilized for cluster analysis. It offers several advantages, including superior computational speed, ease of implementation, dynamic performance on distributed data, and higher accuracy in results. A variation of K-means is the K-Medoids algorithm, which improves upon the former by selecting an actual "medoid" from the dataset as the center, thereby making it less sensitive to outliers (Aliyev et al., 2020).

Several studies employ K-means while determining the number of clusters using the Elbow method. For instance, Marisa et al. (2019) utilized this method with the Sum of Squared Errors (SSE) at k = 2 to define customer segments. Similarly, Herman et al. (2022) analyzed the financial efficiency of food retailers in Hungary and Romania using both K-Means and K-Medoids. Their results indicated that the choice of clustering method significantly impacts financial performance evaluation, with K-Medoids producing a more stable and balanced group distribution. Additionally, Shabrina et al. (2023) compared K-means, K-medoids, and fuzzy C-means through the analysis of school quality validation data. Their findings showed that K-Medoids achieved the highest efficiency (DBI = 0.8; SI = 0.46). In that study, the Davies-Bouldin Index (DBI) and Silhouette Coefficient were used to determine the optimal number of clusters, which was identified as K=3.

The selection between K-means and K-medoids depends on data characteristics, analytical objectives, accuracy requirements, and noise tolerance (Anikin et al., 2017). Generally, K-means is considered the optimal choice for large-scale datasets, such as sales transactions (Brahmana et al., 2020). In data processing, K-Means often yields better results than K-Medoids as it performs effectively in calculating mean distances within cluster centers. Previous research indicates that using K-Means or K-Medoids can produce significantly different results, suggesting that these algorithms do not always guarantee optimal performance as they are highly dependent on the specific dataset. Furthermore, K-means has a significantly lower computational complexity compared to K-medoids. Consequently, this study employs K-means along with the Davies-Bouldin Index (DBI), Calinski-

Harabasz Score (CH Score), and the Silhouette Coefficient to determine the optimal number of clusters.

3. Research Framework

The process of analyzing customer data using RFM models and clustering algorithms to support businesses is illustrated in Figure 1.

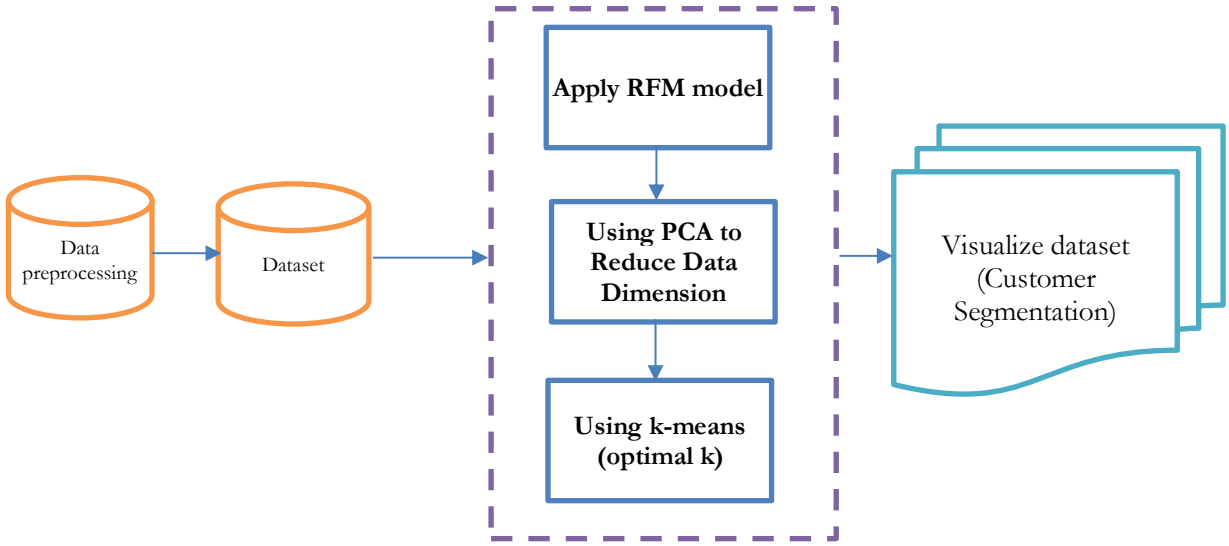

Figure 1: Customers' Segmentation Analysis Framework

Step 1: Data Preprocessing: the data is cleaned and prepared by handling missing values, removing duplicates, and standardizing formats to ensure data quality and analytical reliability.
Step 2: Dataset Construction: Data is organized into a structured dataset containing customer identifiers and transaction attributes required for further analysis.
Step 3: Apply the RFM Model: Compute Recency, Frequency, and Monetary metrics to quantify customer engagement, purchasing behavior, and economic value.
Step 4: Dimensionality Reduction Using PCA: Apply Principal Component Analysis to reduce feature dimensionality, minimize redundancy, and improve clustering efficiency while preserving key variance.
Step 5: Customer Clustering Using K-means (Optimal k): Perform K-means clustering to group customers into homogeneous segments, selecting the optimal number of clusters using validation metrics (Silhouette Score, Calinski Harabasz, and Davies Bouldin).
Step 6: Visualization of Customer Segmentation: Visualize clustered results using plots and charts to interpret segment characteristics and support data-driven decision-making

4. Experimental Results and Discussions

Data collection: Data is collected from Kaggle about 10 different shopping malls between 2021 and 2023 in Istanbul. Dataset includes 99456 sale transactions with essential information such as invoice numbers, customer IDs, age, gender, payment methods, product categories, quantity, price, order dates, and shopping mall locations. The attribute information as follows.

- invoice_no: Invoice number. Nominal. A combination of the letter 'I' and a 6-digit integer uniquely assigned to each operation.
- customer_id: Customer number. Nominal. A combination of the letter 'C' and a 6-digit integer uniquely assigned to each operation.
- gender: String variable of the customer's gender.
- age: Positive Integer variable of the customer's age.

- category: String variable of the category of the product purchased.
- quantity: The quantities of each product (item) per transaction. Numeric.
- price: Unit price. Numeric. Product price per unit in Turkish Liras (TL).
- payment_method: String variable of the payment method (cash, credit card or debit card) used for the transaction.
- invoice_date: Invoice date. The day when a transaction was generated.
- shopping_mall: String variable of the name of the shopping mall where the transaction was made.

We use Tableau to analyze data. For example, Figure 2 shows the alternative payment methods. The chart is a horizontal stacked bar chart that displays the quantity of items sold, broken down by payment method: Cash, Credit Card, and Debit Card. The Credit Card is the most used payment method, with the highest quantity of sales across almost all categories. Cash and Debit Card payments are also used but represent a smaller portion of the total sales volume

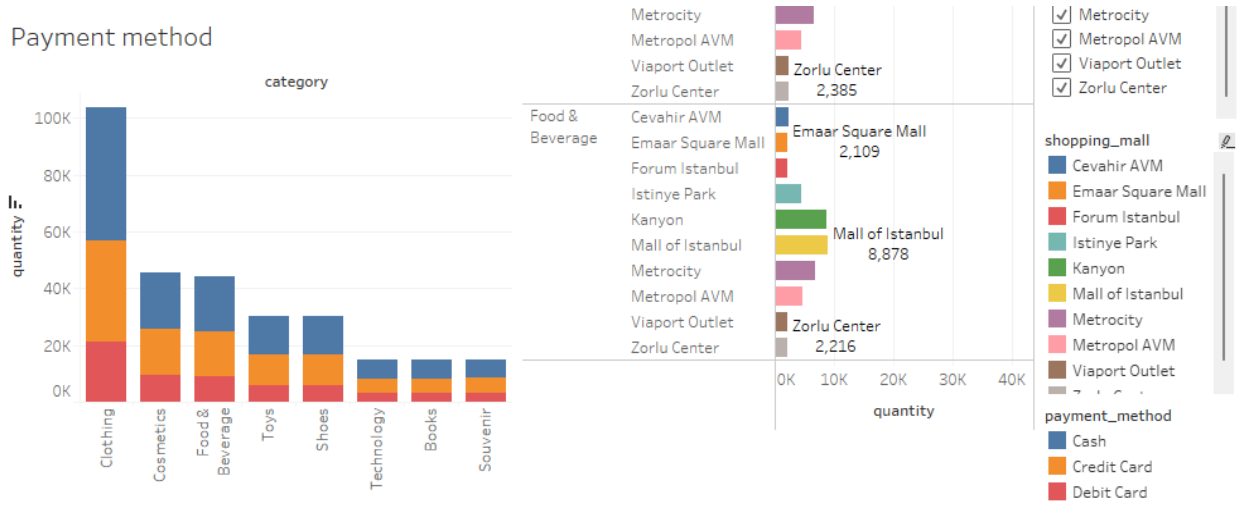

Figure 2: data analysis example

In a dataset comprising eight product categories, we identified the five categories with the highest sales volumes. Clothing ranked first, accounting for 40.8% of total sales. This predominance may be attributed to the age distribution of customers in the dataset (20–60 years), a demographic group that exhibits a high demand for apparel products. Cosmetics ranked second, representing 17.9% of total sales. This outcome can be explained by the gender composition of the dataset, in which female customers outnumber male customers, thereby contributing to higher cosmetic purchases. These findings are illustrated in Figure 3.

Figure 3: product categories data analysis

RFM model

RFM is a widely adopted analytical framework for evaluating customer value and facilitating customer segmentation. The acronym RFM represents three key behavioral dimensions:

- **Recency (R):** This dimension measures the time elapsed since a customer's most recent purchase. A lower recency value indicates more recent transactional activity, suggesting stronger customer engagement and a higher likelihood of responsiveness to marketing initiatives.
- **Frequency (F):** This dimension reflects how often a customer makes purchases within a specified time frame. Higher frequency values denote more frequent interactions with the business, which may indicate greater customer loyalty, satisfaction, or habitual purchasing behavior.
- **Monetary (M):** This dimension captures the total expenditure of a customer over a defined period. Customers with higher monetary values contribute more substantially to organizational revenue and are often associated with greater potential lifetime value.

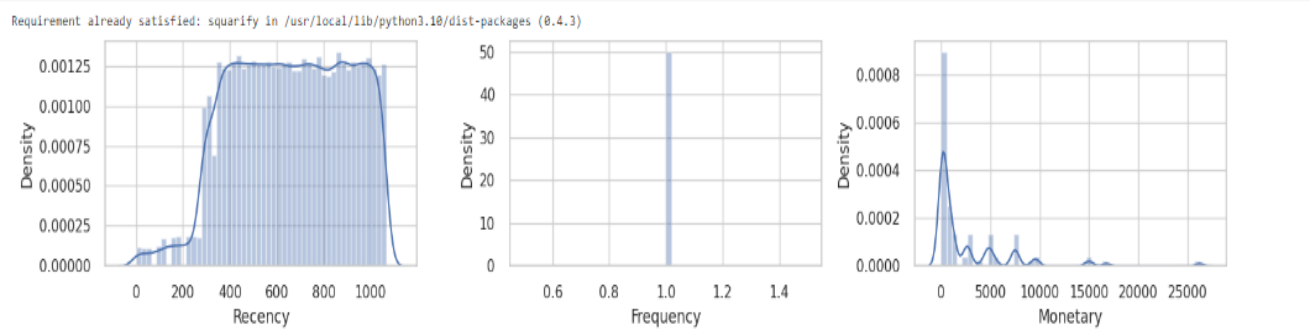

Figure 4: RFM visualization for dataset

In this paper, we will calculate the Recency, Frequency and Monetary indices of each customer based on customer_id and also show the distribution of each part to the customer. In our dataset, the Recency index is very evenly distributed, increasing from 3000-10000, showing that the customer's purchase time is quite close. For the Frequency distribution, there is only one column at number 1 showing that customers only buy once with each customer_id. As for the Monetary column, customers usually spend 500-3000 on their orders mainly, with higher spending also occurring but it gradually decreases, the higher it is, the less spending it does (see **Figure 4**).

Clustering analysis

Using PCA to Reduce Data Dimension: To handle the dimensionality of the data, we selected the top three Principal Components (PCs), which are the features that capture the most variance within the dataset. By reducing the data to these top 3 PCs, the 3D scatter plot experiment allowed them to visually inspect the cohesion within each group and the overall quality of separation between the different customer clusters (see Figure 5).

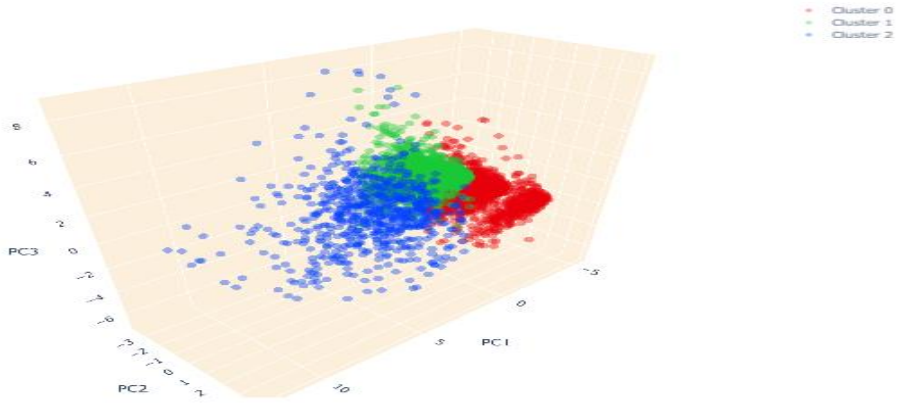

Figure 5: Data in 3D visualization

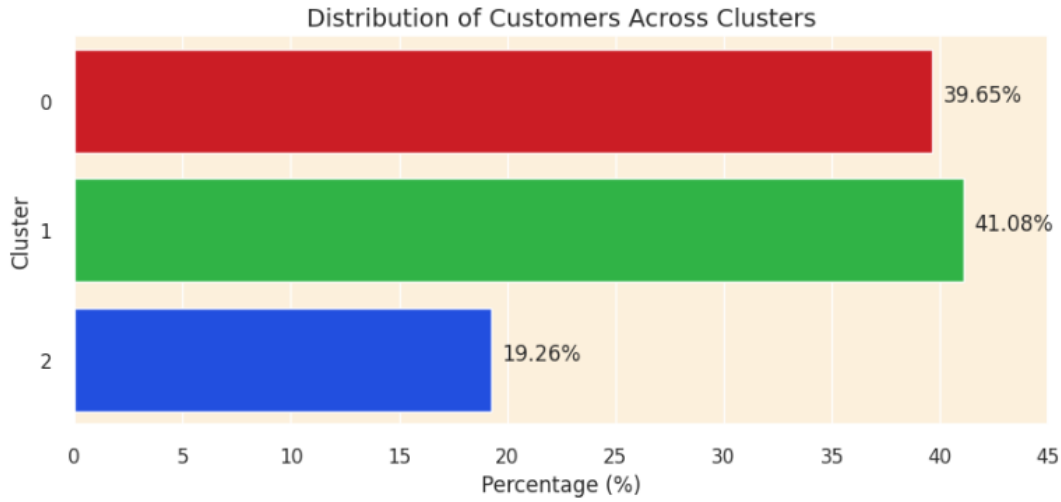

Figure 6: Percenatge of data in alternative clusters

Finding the Optimal Number of Clusters (K): Based on RFM model, we pre-determined that data would segment into three specific target groups: Potential customers, Regular customers, and Infrequent customers. To validate whether was an effective choice, we analyse the resulting clusters using three key evaluation metrics (see **Figure 7**).

Metric	Value
Number of Observations	94485
Silhouette Score	0.3391009697381492
Calinski Harabasz Score	68829.8578786589
Davies Bouldin Score	1.108416951736551

Figure 7: Evaluation metrics for optimal of K=3.

From **Figure 7**, the detail can be explained as follows.

- **Silhouette Score (0.339):** This indicated a fair degree of separation between the clusters, though the score suggests there may still be slight overlaps.
- **Calinski Harabasz Score (68829.85):** A considerably high score, which verified that the three clusters were well-defined and represented substantial, underlying structures in the data.
- **Davies Bouldin Score (1.11):** A reasonably low score, suggesting decent separation and a moderate level of similarity between each cluster and its nearest neighbour.

Therefore, it can be seen that the 3-cluster structure was successful because the customer distribution was highly balanced and significant (approximately 40%, 41%, and 19%), proving the algorithm had identified meaningful groupings rather than just isolating noise or outliers.

Customer Segmentation Analysis: The customer segmentation experiment primarily relied on combining the RFM (Recency, Frequency, Monetary) method with an unsupervised learning clustering algorithm, notably K-means. The clustering experiment successfully categorised the dataset into three distinct behavioural profiles (**Figure 6**).

- **Cluster 0: Regular Customers (39.65%):** These customers make highly frequent purchases and have a relatively high average transaction value. They tend to reside in Istanbul, and their moderate spending shows an increasing trend over time.

- **Cluster 1: Infrequent Customers (41.08%):** Making up the largest segment, this group has low spending and a low number of transactions. They exhibit a very high number of days since their last purchase, indicating they have not interacted with the businesses for a long time.
- **Cluster 2: Potential Customers (19.26%):** These customers are the biggest spenders, boasting very high total spending and exceptionally high average order values. They have the lowest average time between purchases, shop frequently (often early in the day), and exhibit highly volatile but predictable monthly spending patterns

For further analysis, we visual dataset with 3 clusters to more understand about the customers purchase and behaviour (see in Figure 8).

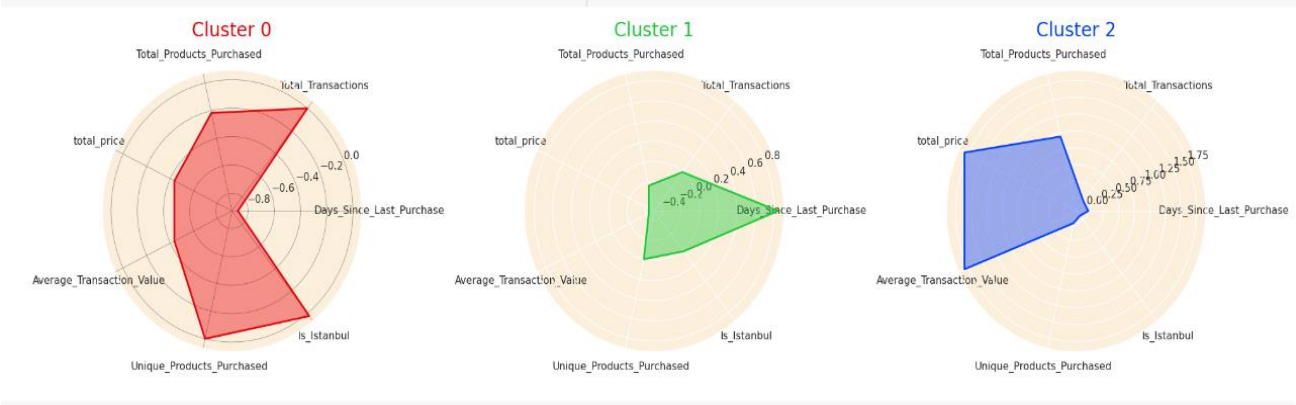

Figure 8: Customers' segmentation analysis

Cluster 0 (Red Chart): Regular Customers

- Customers in this cluster spend moderately and their total purchases are very frequent, as indicated by the Total_Transactions and unique_Products_Purchases metrics. They tend to spend moderately, showing that their spending is increasing over time. These customers shop frequently, as indicated by the High Hours value, and they primarily reside in Isband. Their average transaction value is relatively high, meaning that when they shop, they tend to buy items of moderate value.

Cluster 1 (Green Chart): Infrequent Customers

- Customers in this group tend to spend less, with a low number of frequent transactions but much less total spending and product purchases. They do not tend to shop for goods on the market, as indicated by a very high Days_Since_Last_Purchase value, indicating that the customer group has not been shopping for a long time. Their spending trends are on the lower side and they have a low monthly spend variance. The average transaction value is lower, suggesting that when they shop, they tend to spend less per transaction.

Cluster 2 (Blue Chart): Potential Customers

- Customers in this group are big spenders with very high total spending and a very high average value of total purchase orders.
- They engage in frequent transactions but also have high frequency and cancellation rates.
- These customers have a very low average time between purchases, and they tend to shop early in the day.

Their monthly spending is highly volatile, suggesting their spending patterns may be more predictable than other groups

5. Conclusion

This study demonstrates the effectiveness of integrating the RFM analytical framework with unsupervised clustering techniques to achieve meaningful and actionable customer segmentation in the retail sector. By leveraging multidimensional transactional data from Istanbul shopping centers (2021–2023), the proposed

automated approach successfully identified three strategically significant customer segments—Regular, Infrequent, and Potential customers—each exhibiting distinct behavioral patterns. The evaluation metrics, including a Silhouette Score of 0.339, a Calinski–Harabasz Score of 68,829.85, and a Davies–Bouldin Score of 1.11, confirm the robustness, cohesion, and separation of the resulting clusters. These findings indicate that the model provides a reliable representation of customer heterogeneity, enabling retailers to better understand purchasing behaviors and segment-specific needs. From a managerial perspective, the proposed framework offers a data-driven foundation for designing personalized marketing initiatives, improving customer engagement, and enhancing retention strategies. By identifying high-value and at-risk customer groups, businesses can allocate resources more efficiently and implement targeted interventions to maximize customer lifetime value.

Future research may extend this work by incorporating additional behavioral or demographic variables, exploring alternative clustering algorithms, or applying the framework to real-time data environments to support dynamic customer segmentation.

References

1. Shirole, R., Salokhe, L., Jadhav, S. (2021). Customer Segmentation using RFM Model and K-Means Clustering. Int. J. Sci. Res. Sci. Technol. Vol. 8, pp. 591–597
2. Aliyev, M., Ahmadov, E., Gadirli, H., Mammadova, A., and Alasgarov, E. (2020). Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning, Available: http://arxiv.org/abs/2008.08662
3. Sitompul, B. J. D., Sitompul, O. S., and Sihombing, P. (2019). Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm, International Conference on Computing and Applied Informatics, vol. 1235, no. 1, pp. 1-6. DOI 10.1088/1742-6596/1235/1/012015.
4. Manero, K. M., Rimiru, R., and Otieno, C. (2018). Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm, International Journal of Computer Science, vol. 15, no. 5, pp. 67–76.
5. Khong, D. W. K. (2021). Rents: How Marketing Causes Inequality by Gerrit De Geest. Asian Journal of Law and Policy, 1(1), 83–86. https://doi.org/10.33093/ajlp.2021.5
6. Rachmawati, I.K. (2020). Collaboration Technology Acceptance Model, Subjective Norms and Personal Innovations on Buying Interest Online. Int. J. Innov. Sci. Res. Technol. 5, pp. 115–122.
7. Shabrina, A. N., Afdal, M., and Monalisa, S. (2023). Comparison Of K-Means, K-Medoids, and Fuzzy C-Means Algorithms for Clustering Drug User's Addiction Levels. J. Sist. Cerdas, vol. 6, no. 2, pp. 113–122, 2023, DOI: https://doi.org/10.37396/jsc.v6i2.313
8. Marisa, V, Ahmad, S. S. S., Yusof, Z. I. M., and Aziz, T. M. A. (2019). Segmentation Model of Customer Lifetime Value in Small an Medium Enterprise (SMEs) using K-Means Clustering and LRFM Model, International Journal of Integrated Engineering, vol. 11, pp. 169 -180.
9. Herman, E., Zsido, K. E., and Fenyves, V. (2022). Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation. Appl. Sci.2022, 12(16), 7985; https://doi.org/10.3390/app12167985.
10. Brahmana, R. W. S., Mohammed, F. A., Chairuang, K. (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. Lontar Komputer. Vol 11 No 1 (2020). DOI:10.24843/lkjiti.2020.v11. i01.p04
11. Anikin, I. V. (2017). Privacy Preserving DBSCAN Clustering Algorithm for Vertically Partitioned Data in Distributed Systems, International Siberian Conference on Control and Communications, vol. 10, pp.1-4, 2017.