CUSTOMER SEGMENTATION AND SHOPPING BEHAVIOR ANALYSIS WITH RFM AND K-MEANS

Tran Thi Mai[1], Nguyen Quynh Hoa[2]

Abstract

Customer segmentation plays a pivotal role in customer relationship management (CRM) and in optimizing marketing strategies. This study applies the RFM model (Recency, Frequency, Monetary) in combination with the K-Means algorithm to analyze shopping behavior from 9,994 retail transactions. The results reveal four distinct customer groups with clear differences in loyalty, transaction frequency, and spending value. Compared with previous studies that primarily focused on applying RFM–KMeans in e-commerce or banking (Chen et al., 2012; Rahman & Khan, 2021), this research offers two main contributions: (i) enhancing clustering quality through Box-Cox transformation and outlier treatment, which improved the Silhouette Score by nearly 44%; and (ii) extending behavioral analysis across multiple dimensions such as product categories, revenue–profit, customer lifecycle, and shipping methods. These findings provide practical implications for businesses to design personalized customer care strategies, optimize resource allocation, and sustain competitive advantage in the context of digital transformation.

Keywords: *RFM model; K-Means; customer segmentation; consumer behavior.*

PHÂN KHÚC KHÁCH HÀNG VÀ PHÂN TÍCH HÀNH VI TIÊU DÙNG DỰA TRÊN RFM VÀ K-MEANS

Tóm tắt

Phân khúc khách hàng đóng vai trò then chốt trong quản trị quan hệ khách hàng (CRM) và tối ưu hóa chiến lược marketing. Nghiên cứu này ứng dụng mô hình RFM (Recency, Frequency, Monetary) kết hợp với thuật toán K-Means để phân tích hành vi mua sắm từ dữ liệu bán lẻ 9.994 giao dịch. Kết quả cho thấy bốn nhóm khách hàng đặc trưng với sự khác biệt rõ rệt về mức độ trung thành, tần suất giao dịch và giá trị chi tiêu. So với các nghiên cứu trước chủ yếu dừng lại ở việc áp dụng RFM–KMeans trong thương mại điện tử hoặc ngân hàng (Chen và cộng sự, 2012; Rahman và Khan, 2021), nghiên cứu này có hai điểm mới: (i) cải thiện chất lượng phân cụm thông qua biến đổi Box-Cox và xử lý ngoại lai, giúp chỉ số Silhouette Score tăng gần 44%; (ii) mở rộng phân tích hành vi theo nhiều chiều cạnh như danh mục sản phẩm, doanh thu – lợi nhuận, vòng đời khách hàng và phương thức vận chuyển. Các phát hiện này cung cấp cơ sở thực tiễn để doanh nghiệp thiết kế chiến lược chăm sóc khách hàng cá nhân hóa, tối ưu phân bổ nguồn lực và duy trì lợi thế cạnh tranh trong bối cảnh chuyển đổi số

Từ khóa*:* *Mô hình RFM; K-Means; phân khúc khách hàng; hành vi tiêu dùng.*

JEL classification: C51, D11, D12.

DOI: 10.63767/TCKT.34.2025.104.117

1. Introduction

In the context of increasing competition and digital transformation, understanding customer behavior is crucial for businesses - especially in the retail sector, where transactional data is continuously recorded and highly detailed. The ability to leverage this data to analyze consumer behavior is not only an inevitable trend but also a strategic lever in Customer Relationship Management (CRM).

However, research gaps remain in three aspects: (i) many RFM–KMeans studies stop at describing the procedure without clearly demonstrating the improvement in clustering quality brought by preprocessing steps (e.g., skewness correction, outlier handling); (ii) there is a lack of multidimensional behavioral analysis after clustering (e.g., product categories, profit–revenue, customer lifecycle, delivery methods) to translate results into actionable recommendations;

and (iii) limited discussion exists on the transferability of the model to emerging markets such as Vietnam.

One of the most popular and effective tools for customer behavior analysis is the RFM model (Recency - Frequency - Monetary). This model evaluates three key aspects of the customer-business relationship: how recently a customer made a purchase (Recency), how often they make purchases (Frequency), and how much they spend (Monetary). Assessing customers based on these indicators provides a quantitative view of their loyalty, profit potential, and likelihood of engagement.

However, to classify customers more automatically and accurately, this study proposes combining the RFM model with the K-Means machine learning algorithm. This approach allows for clustering customers with similar behaviors, thereby enabling businesses to develop more effective personalized marketing strategies.

The objective of this study is to present a comprehensive customer segmentation process, which includes the following steps:

Data preprocessing

Data transformation using Box-Cox

Data normalization using StandardScaler

Determination of the optimal number of clusters using the Elbow method and Silhouette Score

Application of K-Means for clustering

The research team selected U.S. retail data for analysis because: (i) it is a publicly available dataset, rich in attributes and widely used as a benchmark, thereby ensuring the transparency of results; (ii) its structure (orders–customers–categories–shipping) is comparable to that of modern retailers in Vietnam, allowing the transfer of the analytical procedure with minimal contextual dependency; and (iii) given the limited data-sharing practices among many Vietnamese enterprises, employing a public benchmark dataset serves as a proof-of-concept step, enabling subsequent parameter tuning and outlier threshold adjustments when applied to internal data. This approach does not conflict with the urgency of the Vietnamese context; rather, it shortens the timeline for real-world implementation.

The study uses U.S. retail data to develop this process and derives insights that can be applied to Vietnamese businesses. The results of the analysis and visualization of customer segments offer deep insights, supporting strategic decision-making in marketing and customer management.

2. Theoretical framework

Customer segmentation is a key topic in marketing and has been extensively studied through quantitative models and data mining approaches. The RFM model is a widely used method for evaluating customer value and engagement (Fader, Hardie & Lee, 2005). Numerous studies, such as those by Christy et al. (2021) and Chen, Sain & Guo (2012), have demonstrated the effectiveness of combining RFM with the K-Means algorithm in segmenting customers in the retail and e-commerce sectors.

Recent research has expanded the application of the RFM model in customer segmentation by integrating it with various advanced clustering algorithms to improve accuracy and interpretability. For instance, John et al. (2024) used retail data from the UK to compare the performance of clustering algorithms such as K-Means, GMM, DBSCAN, and Agglomerative Clustering based on RFM features. Additionally, Vianna Filho et al. (2025) proposed a novel graph-based approach using the Max-K-Cut optimization algorithm to cluster RFM data with reduced computational complexity while maintaining segmentation quality.

Furthermore, studies conducted in Indonesia and Bangladesh (Jimmi Chitra & Heikal, 2024; Rahim & Khan, 2021) applied the RFM-KMeans approach to banking and retail customer data, emphasizing the enhanced personalization of CRM and increased customer lifetime value through behavioral clustering.

However, previous studies have mainly focused on directly applying RFM–KMeans or comparing multiple clustering algorithms, without emphasizing improvements in clustering quality through data transformation and outlier treatment. Moreover, most of them stop at merely identifying customer groups, lacking in-depth post-clustering behavioral analysis. The distinct contributions of this study are: (i) applying Box-Cox transformation and outlier removal to enhance clustering accuracy, which significantly increases the Silhouette Score compared with the standard RFM–KMeans approach in prior research; and (ii) incorporating multidimensional analyses of customer behavior within each cluster to provide practical recommendations for Vietnamese enterprises.

K-Means Algorithm:

The K-Means clustering algorithm was introduced in 1957 by Lloyd and was officially published in 1982 (Lloyd, S. P., 1982). It is one of the most popular clustering methods and is based on data partitioning. The algorithm consists of the following steps:

Step 1: Initialization

The K-Means algorithm begins by randomly selecting k data points (clusters) from the dataset. k is the number of clusters to be formed and must be specified before running the algorithm.

Step 2: Assigning labels to each data point

After the initial k clusters are selected, the algorithm calculates the distance between each data point and the centers of the k clusters, assigning each point to the nearest cluster. The distance between two data points is typically calculated using the Euclidean distance, with the formula:

$$d(x,y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots + (y_n - x_n)^2}$$

Step 3: Updating cluster centroids

Once all data points have been assigned to clusters, the cluster centers are recalculated to improve the algorithm's performance. The new centroid of a cluster is determined by computing the average position of all data points within that cluster.

Step 4: Checking the stopping condition

The process of assigning data points and updating cluster centers is repeated until the centroids no longer change between iterations (or the change is below a specified threshold), or until the algorithm reaches the maximum number of iterations.

Among various clustering algorithms (DBSCAN, GMM, Agglomerative Clustering, etc.), K-Means was selected in this study for three main reasons:

Transparency and interpretability: The clustering results from K-Means are based on Euclidean distances in the RFM space, making them easier for businesses to understand and apply in customer management practices, whereas algorithms such as DBSCAN or GMM are often more difficult to interpret for non-technical users.

Computational efficiency: K-Means has low computational complexity, making it suitable for medium- and large-scale retail datasets. In contrast, other algorithms such as GMM or Agglomerative Clustering incur higher computational costs and are less scalable when the number of customers increases. Compatibility with preprocessing techniques: Since K-Means is sensitive to outliers and skewed data, this study leverages that characteristic to validate the effectiveness of Box-Cox transformation and outlier treatment. This also represents a key distinction and novel contribution compared with previous studies.

Based on the above theoretical foundations and literature review, this paper proposes a three-part research framework:

1. Systematize RFM-based customer behavioral characteristics.

2. Perform data preprocessing and transformation (e.g., Box-Cox transformation and standardization) to improve clustering quality.

3. Apply the K-Means algorithm and determine the optimal number of clusters using the Elbow Method and Silhouette Score, followed by result visualization.

This research framework is well-suited to addressing the customer segmentation problem based on purchasing behavior derived from transactional data.

3. Research methodology

In this study, the quantitative research method was implemented through several key steps. First, the collected data were processed and pre-processed to ensure completeness, accuracy, and suitability for analysis. Next, the RFM (Recency, Frequency, Monetary) model was applied to construct features that reflect customer behavior and value. Based on these features, the K-Means clustering algorithm was employed to identify groups of customers with similar characteristics. After clustering, the results were visualized using graphical tools to facilitate interpretation and practical insights. The entire analysis was conducted in Python 3.13.3 (64-bit), with the support of specialized libraries: pandas and numpy for data processing; scipy for Box-Cox transformation; scikit-learn for data standardization and clustering (StandardScaler, KMeans); and matplotlib and seaborn for data visualization.

4. Results and analysis

4.1 Data description and preprocessing

In this study, we use a well-known U.S. retail dataset that is publicly available on the website https://community.tableau.com/s/sample-superstore. The dataset consists of 9,994 transactions and 21 fields: 'RowID', 'OrderID', 'OrderDate', 'ShipDate', 'ShipMode', 'CustomerID', 'CustomerName', 'Segment', 'Country', 'City', 'State', 'Postal Code', 'Region', 'ProductID', 'Category', 'SubCategory', 'ProductName', 'Sales', 'Quantity', 'Discount', and 'Profit'. Four key fields were selected to build the RFM model: CustomerID, OrderID, OrderDate, and OrderValue, corresponding to an initial total of 793 customers. After data validation, 70 customers with missing OrderDate (NaN) information or identified as outliers were removed (as detailed in the Customer Segmentation section), resulting in 723 valid customers. The study further investigates the characteristics of customer groups using several data fields such as Category, Subcategory, Sales, and Profit.

Although the Superstore retail dataset has the advantages of being publicly available and rich in attributes, the use of this secondary data also has certain limitations. First, the dataset reflects the U.S. retail market, and thus consumer behavior patterns may differ from the Vietnamese context. Second, the data lack several important variables—such as transaction channels, online behaviors, and customer feedback, that could influence segmentation outcomes. Therefore, the results of this study should be regarded as a proof-of-concept, and in practical applications, Vietnamese enterprises are advised to adjust and calibrate the model based on their own internal data.

Calculate the RFM metrics for each customer as follows:

Recency (R): The number of days from the last purchase to the reference date (the last day in the dataset).

Frequency (F): The number of transactions (unique OrderIDs) during the entire study period.

Monetary (M): The total spending value (Sales) for each customer.

4.2 Preparing RFM data for clustering

To optimize the K-Means algorithm, RFM data was processed in two steps:

Box-Cox Transformation: Applied to skewed variables to bring the data closer to a normal distribution, reducing the influence of outliers.

Standardization using StandardScaler: Scales variables to a common range (mean 0, standard deviation 1), ensuring each metric has an equivalent impact during the clustering process.

Calculate and examine the distribution of the fields: Recency, Frequency, and Monetary, as illustrated by the following charts:

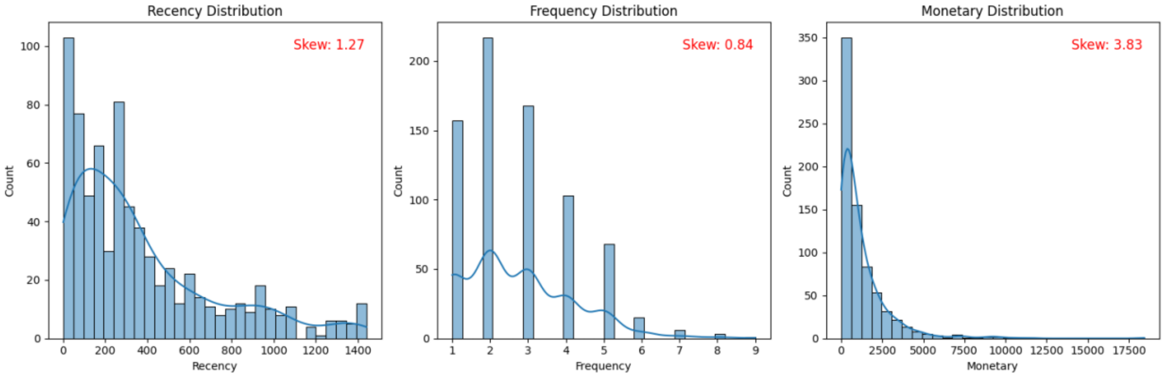

Figure 1: Distribution and Skewness of RFM Variables

Source: Author, via Python

Observations:

Initial analysis reveals that all three RFM metrics (Recency, Frequency, Monetary) exhibit a right-skewed distribution—a common characteristic in customer behavior data. This indicates that the majority of customers tend to have recent purchases, with average frequency and spending levels. However, some important customer groups also exist, such as:

High-value groups: Customers who shop frequently and spend significantly.

At-risk groups: Customers with high Recency scores (meaning a long time since their last interaction).

To enhance the effectiveness of K-Means clustering, it's necessary to adjust the distribution of the RFM variables to be closer to a normal (symmetrical) distribution. The goal is to reduce the skewness coefficient closer to zero, which improves the accuracy of distance calculations and increases the efficiency of clustering.

Three common transformation methods were examined:

Log transformation

Square root transformation

Box-Cox transformation

These methods are applied to adjust the shape of the distribution and reduce data skewness. The criterion for selecting the optimal method is its ability to bring the skewness coefficient closer to zero. Note that these methods are primarily applicable to positive data; in cases with negative or zero values, appropriate adjustments must be made before transformation.

For the Recency variable, we selected a skewness coefficient of -0.1. The results are illustrated in the chart shown in Figure 2.

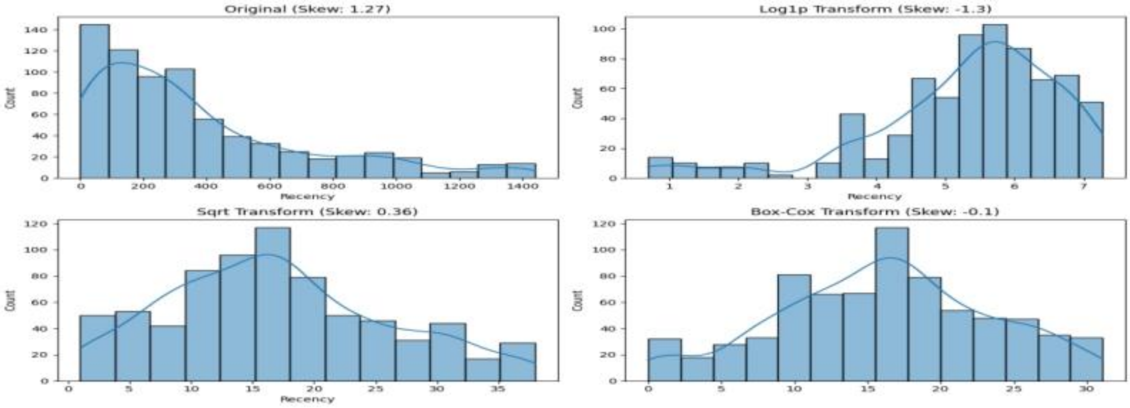

Figure 2: Skewness Reduction Techniques for Recency

Source: Author, via Python

For the Frequency variable, we selected a skewness coefficient of -0.03. The results are illustrated in the chart shown in Figure 3.

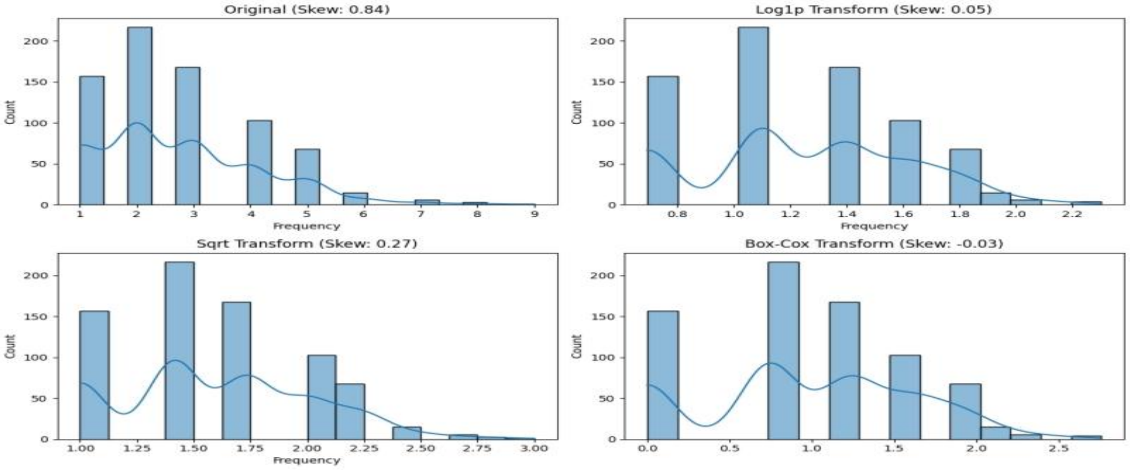

Figure 3: Skewness Reduction Techniques for Frequency

Source: Author, via Python

For the Monetary variable, we selected a skewness of -0.03. The results are illustrated in the chart shown in Figure 4.

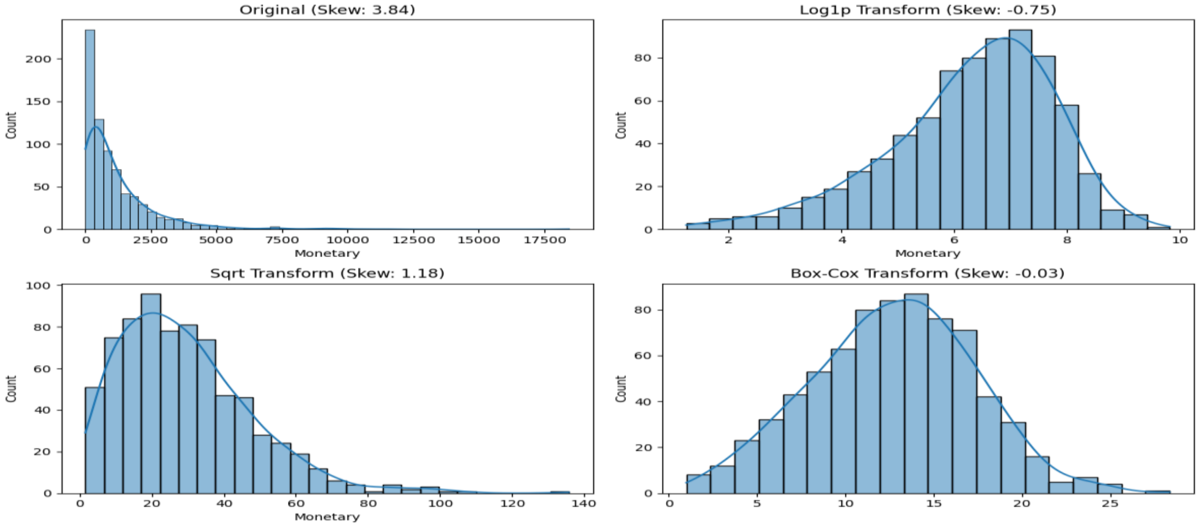

Figure 4: Skewness Reduction Techniques for Monetary

Source: Author, via Python

We standardized the three variables: Recency, Frequency, and Monetary using the StandardScaler class from Python's sklearn.preprocessing library. The purpose of data standardization is to rescale the data to ensure that all features contribute equally, allowing algorithms (especially those based on distance or gradient descent) to perform efficiently, quickly, and stably.

4.3. Customer clustering with K-Means

The customer segmentation process includes:

Determining the number of clusters (K): Using the Elbow Method based on the total within-cluster sum of squares (inertia).

Silhouette Score: Evaluating the separation between clusters.

Cluster labeling: Each customer is assigned to one of four clusters, reflecting similar behavioral characteristics.

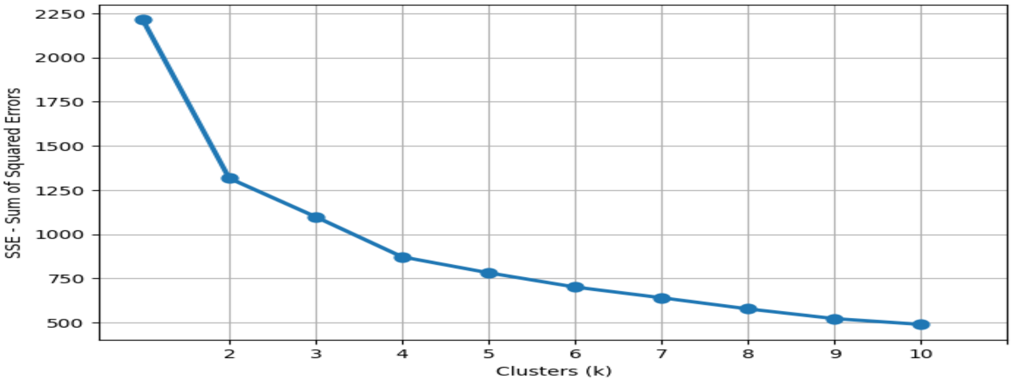

Figure 5: The Elbow Method

Source: Author, via Python

Observations from the Elbow Method Plot, observing the changes in the plot, we note:

From k = 1 to k = 2: The SSE (Sum of Squared Errors) decreases very sharply (from ~2200 to ~1300).

From k = 2 to k = 3: The SSE still shows a significant drop (from ~1300 to ~1100).

From k = 3 to k = 4: The SSE continues to decrease quite well (from ~1100 to below 900).

For k > 4: The curve begins to flatten out, and the reduction in SSE from k = 4 to k = 5, and subsequent points, is no longer as significant as before.

The point where the curve "bends" most distinctly is at k = 4. After this point, adding a new cluster doesn't yield substantial benefits in making the clusters much tighter. Therefore, based on the **Elbow Method**, the optimal number of clusters for partitioning this dataset is 4.

Re-evaluate the number of clusters with k = 4 using the Silhouette Score, as illustrated in the following chart:

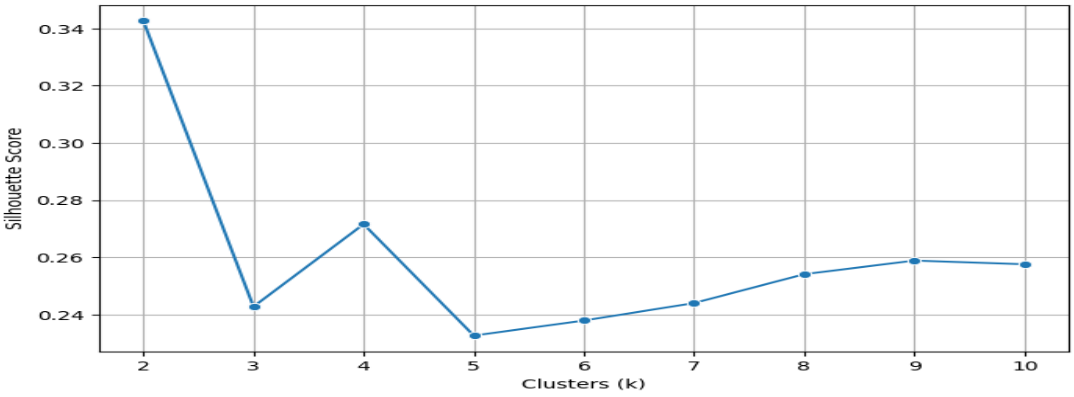

Figure 6: Silhouette Score by Number of Clusters

Source: Author, via Python

Observations:

The highest Silhouette Score occurs at k = 2 (~0.34): This indicates that dividing the data into two clusters yields the highest intra-cluster cohesion and the most distinct separation between clusters. However, having only two groups may be too limited for effective market segmentation.

A sharp drop in score at k = 3 and k = 5 (~0.24, ~0.235): As the number of clusters increases, the clustering quality decreases significantly, suggesting that the clusters begin to overlap or become less distinct. These are not suitable options.

From k = 6 to k = 10**:** The Silhouette Score increases slightly but does not exceed 0.26, indicating that models with more clusters do not significantly improve clustering quality.

k = 4 has the second-highest Silhouette Score (~0.27): This is a reasonable alternative for segmentation while maintaining a relatively good level of cluster separation.

After selecting the appropriate number of clusters for the dataset, the research team filtered customers belonging to each cluster and visualized them in 3D space using distinct colors. The results are illustrated in the chart in Figure 7.

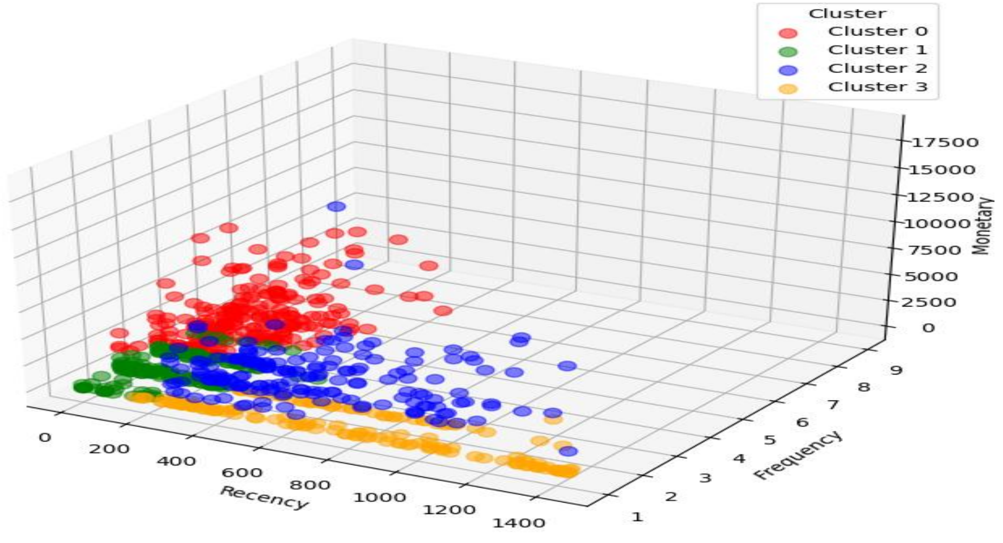

Figure 7: Customer Segmentation in 3D Space RFM

Source: Author, via Python

Following the initial clustering phase, the research team observed that several outliers from Cluster 2 had been incorrectly assigned to Cluster 0. Additionally, the overall Silhouette Score was relatively low, approximately 0.27. To address this issue, the team undertook a series of steps to handle the outliers and enhance the clustering performance, as outlined below:

Step 1: Extract all customers belonging to Cluster 2 for separate analysis.

Step 2: For each RFM metric (Recency, Frequency, Monetary), the Interquartile Range (IQR) method was employed to detect and eliminate outliers:

Compute the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile) for each variable.

Calculate the IQR as IQR = Q3 – Q1.

Remove any customers whose values fall:

below Q1 – 1.5×IQR

above Q3 + 1.5×IQR.

Step 3: After removing the outliers from Cluster 2, the remaining customers were merged with those from Clusters 0, 1, and 3.

Subsequently, the Silhouette Score — a widely used metric for assessing intra-cluster cohesion and inter-cluster separation — was re-evaluated. The results indicated a substantial improvement, with the score increasing from 0.27 to 0.3880, representing an approximate 44% enhancement. This improvement is both statistically and technically significant, as illustrated in Figure 8.

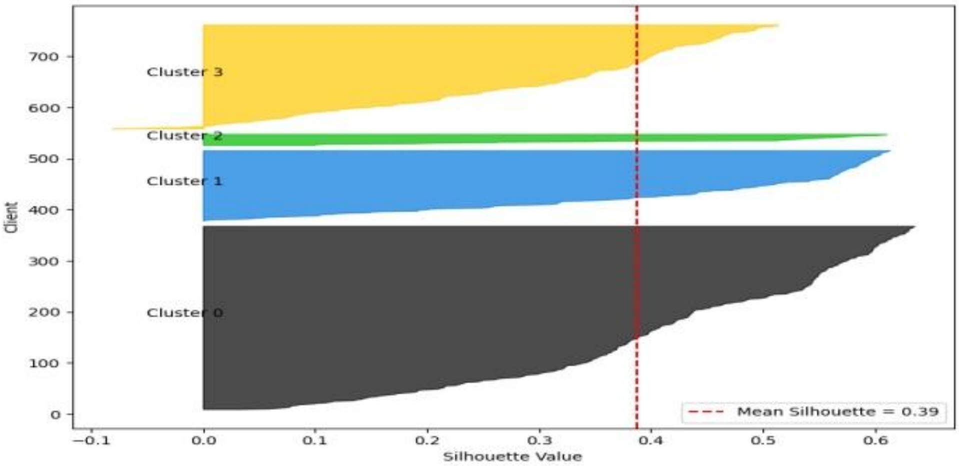

Figure 8: Silhouette chart after removing outliers

Source: Author, via Python

After handling the outliers, the research team conducted a second round of clustering, resulting in the following chart:

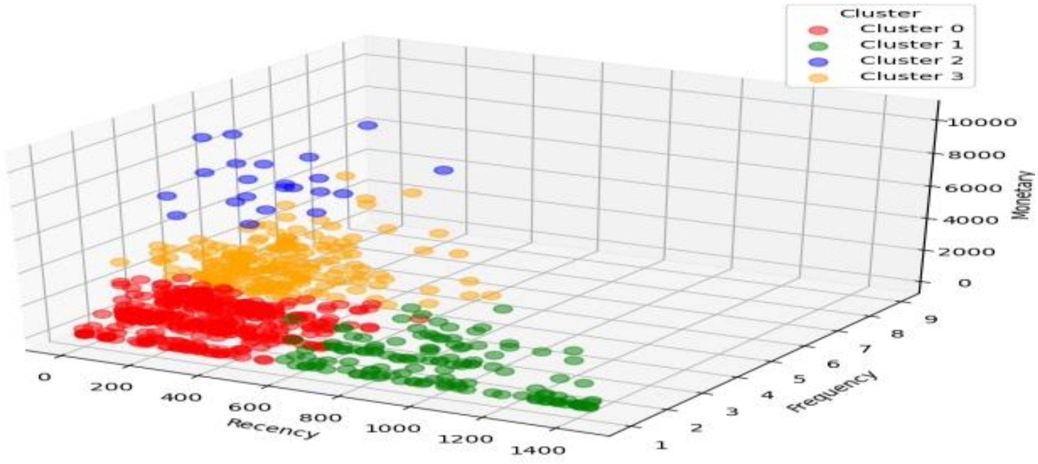

Figure 9: Customer Segmentation in 3D Space RFM -Cleaned

Source: Author, via Python

To examine statistically significant differences among clusters, ANOVA was conducted on the three R,F,M indicators. The results show that all three variables differ significantly across customer clusters (Recency: F = 608.240, p < 0.001; Frequency: F = 449.831, p < 0.001; Monetary: F = 469.492, p < 0.001). This confirms that the clusters are not only descriptively distinct but also statistically significant, thereby strengthening the reliability of the clustering model.

The Post-hoc Tukey HSD results further validate the robustness of the clustering. For Recency, Cluster 1 is significantly different from the other three clusters, indicating that it represents churned customers with the longest inactivity period. In contrast, the difference between Cluster 2 and Cluster 3 is not statistically significant, reflecting their similarity in recent interactions. For Frequency and Monetary, all clusters are significantly different (p < 0.001), demonstrating that transaction frequency and spending value are powerful discriminators. Thus, the customer clusters identified are not only descriptively distinct but also statistically validated, ensuring the robustness of the RFM–KMeans model. *(Source: Python)*

In addition, the basic descriptive statistics for the four clusters are presented in Table 1.

Table 1: *Basic statistical description table for 4 clusters*

	mean	**std**	**min**	**max**	**median**	**count**
Cluster	**Recency**					
0	246.14	166.85	1	648	250	359
1	960.48	247.08	586	1440	918.5	138
2	161.50	156.27	4	615	123.5	22
3	184.95	163.10	1	854	154.5	204
Cluster	**Frequency**					
0	2.11	0.75	1	4	2	359
1	1.57	0.68	1	4	1	138
2	4.64	1.00	3	7	5	22
3	4.40	1.07	2	9	4	204
Cluster	**Monetary**					
0	622.22	616.77	2.81	3588.10	450.14	359
1	509.29	634.60	2.48	2824.23	238.66	138
2	6715.31	1899.00	4459.12	10402.53	6538.53	22
3	1779.01	996.06	85.39	4240.49	1730.34	204

Source: Author, via Python

The customer segmentation results based on the K-Means algorithm and RFM (Recency, Frequency, Monetary) metrics reveal clear differentiation in customer purchasing behavior. The model divides the customer base into four clusters (Cluster 0 to Cluster 3), reflecting variations in engagement levels, transaction frequency, and monetary value.

Based on the 3D clustering visualization (Figure 9) and the descriptive statistics table (Table 1), the following observations can be made:

Cluster 2: This is the VIP loyal customer group, representing the smallest proportion (22 customers), but with the highest average spending (Monetary mean) of $6,715. They also have high purchase frequency (average 4.63 times) and very recent transactions (average recency of 161 days). This group delivers the most value to the business and should be prioritized with special care strategies, exclusive offers, and personalized experiences.

Cluster 3: This group includes stable, loyal customers and accounts for the second-largest share (204 customers). They exhibit good spending levels ($1,779), frequent purchases (4.4 times on average), and fairly recent transaction history (185 days). This segment is highly nurturable and could potentially be upgraded to VIP status through loyalty programs, recurring offers, or membership packages.

Cluster 0: This is the average-value customer group and comprises the largest portion (359 customers). They have made purchases relatively recently (average recency of 246 days) but display low to moderate frequency and spending. This segment is suitable for upselling campaigns or light promotions to encourage repeat purchases and increase customer lifetime value.

Cluster 1: This group shows the longest time since the last purchase (average 960 days), along with low frequency and low monetary value—indicating potential churn or inactivity. Strategies for this group should focus on reactivation through final-chance offers, behavior surveys, or removal from ineffective marketing campaigns.

Additionally, the Silhouette Score reached 0.388, indicating a reasonably good clustering quality for real-world data, especially after outlier removal. This suggests that the model has formed relatively well-separated clusters with strong practical applicability. The proportion of customers by cluster is illustrated in the following chart:

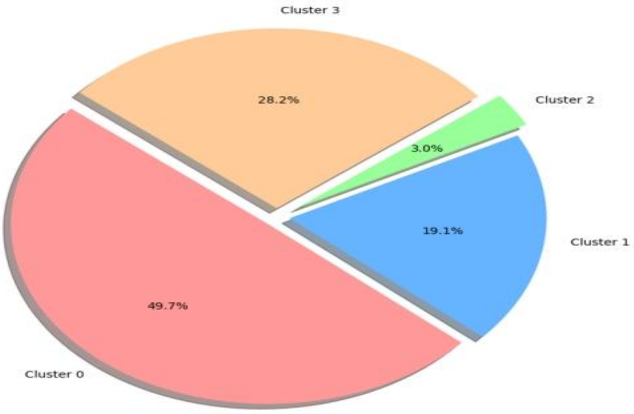

Figure 10: Customer Segmentation in 3D Space RFM -Cleaned

Source: Author, via Python

After the customer clusters were identified using the RFM model combined with the K-Means algorithm, the next step in the analysis process is to delve deeper into the distinctive consumption characteristics of each group. This aims to gain a better understanding of the behavior, value, and potential of different customer segments.

An analysis of the distribution of transaction volume by product category (Category) and customer cluster is conducted, with the results illustrated in the chart shown in Figure 11.

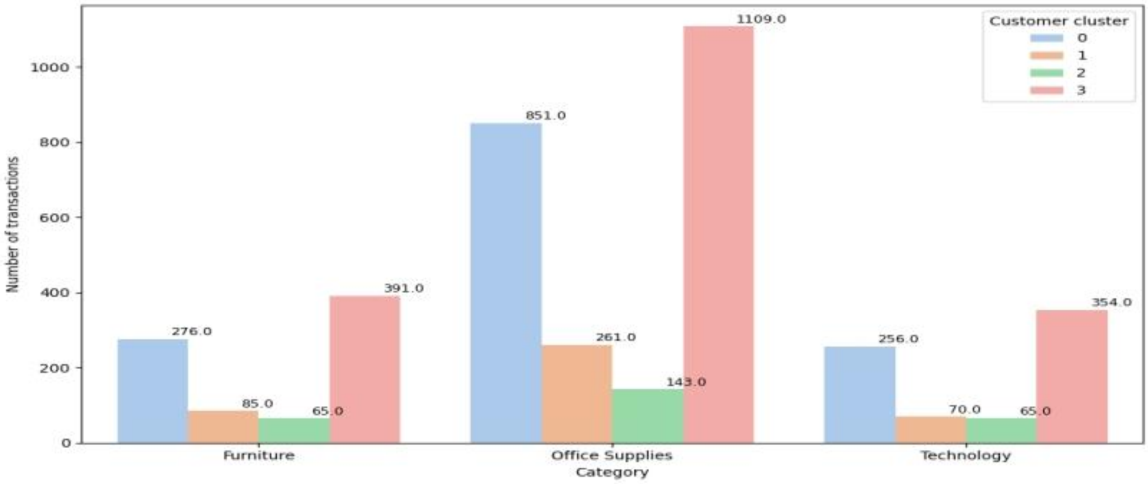

Figure 11 : Number of transactions by Category and customer cluster

Source: Author, via Python

The chart illustrates the number of transactions across the three main product categories (Category: Furniture, Office Supplies, Technology), segmented by the four customer clusters (Cluster 0–3), highlighting significant differences in consumption behavior among the groups.

Cluster 3 has the highest transaction volume, particularly in the *Office Supplies* category, indicating that this group consists of frequent buyers with large-scale and stable demand for office-related products. Additionally, this cluster also shows high transaction levels in both *Furniture* and *Technology*, demonstrating diversity in their purchasing behavior.

In contrast, Cluster 2—identified as the VIP group based on high Monetary value—has the lowest number of transactions across all three categories. This suggests that their spending is concentrated on high-value products rather than a large number of smaller purchases. Such behavior is typically associated with corporate clients or high-budget customers who make selective purchases.

Clusters 0 and 1 display average purchasing behavior, with a focus on *Office Supplies* and fewer transactions in *Furniture* and *Technology*. This reflects a general customer group with basic, stable needs—appropriate for broad-based marketing campaigns.

An analysis of transaction volume for detailed product types (SubCategory) across clusters was also conducted, with the results shown in the chart in Figure 12.

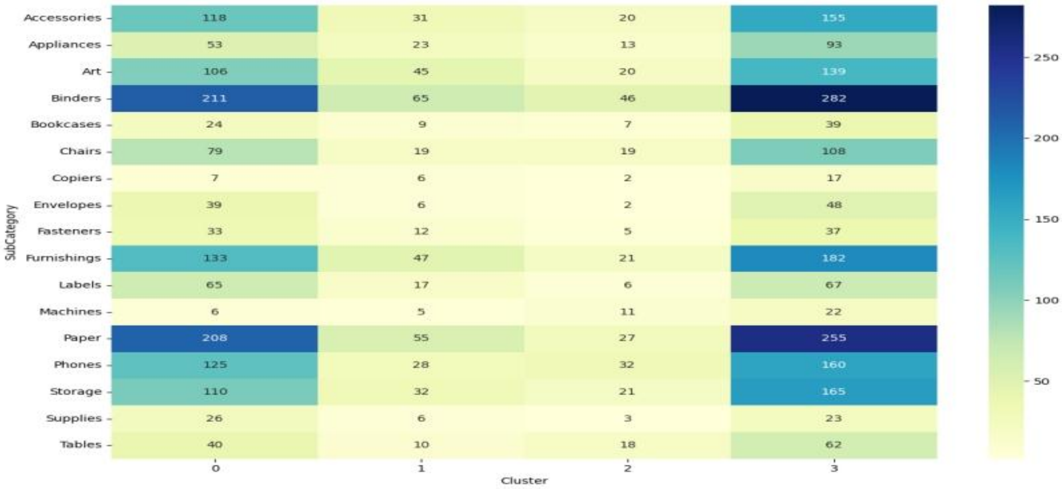

Figure 12: Number of transactions of SubCategory by customer cluster (Heatmap)

Source: Author, via Python

The analysis of transaction volume by product subcategory (Figure 12) reveals clear distinctions in purchasing behavior among customer groups. Specifically, customers in Clusters 0 and 3 show significantly higher total transaction volumes compared to Cluster 1, and especially Cluster 2.

Cluster 0, which contains the largest number of customers (359), contributes a high number of transactions despite having a relatively low average monetary value. This is reflected in high activity across various subcategories such as *Binders*, *Paper*, *Phones*, and *Furnishings*.

Cluster 3 (204 customers) represents loyal customers with high purchase frequency and moderate order value. This is evident through the prominent number of transactions in subcategories like *Binders*, *Paper*, *Accessories*, and *Storage*.

Notably, Cluster 2, while identified as the VIP group with the highest average spending per customer according to the RFM analysis, has a much lower total number of transactions due to its small size (only 22 customers).

Cluster 1 is characterized by the lowest purchase frequency and order value, with transaction volumes distributed thinly and consistently low across all product subcategories.

These findings show that combining transaction volume analysis with RFM-based behavioral insights provides a more comprehensive view of each customer segment, thereby enhancing the effectiveness of targeted marketing and customer care strategies.

An analysis of profit and revenue by customer cluster was also conducted, with the results illustrated in the chart shown in Figure 13.

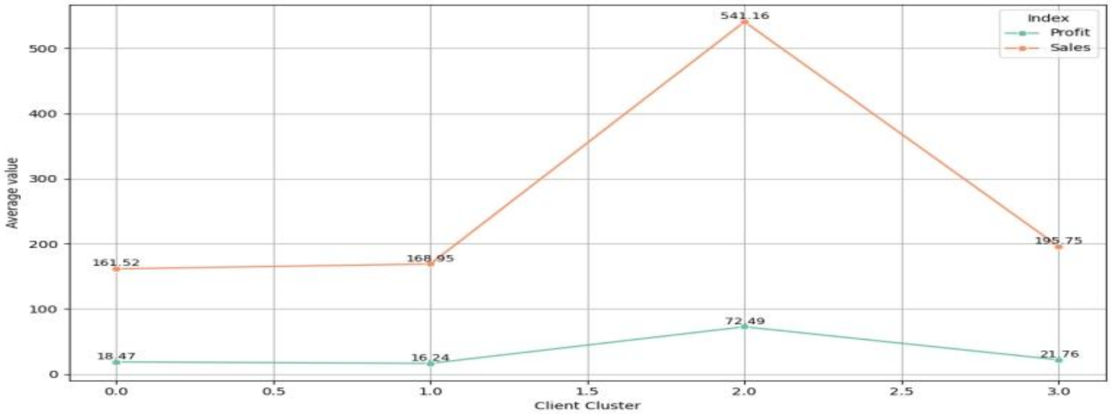

Figure 13: Compare average Profit and Sales by customer cluster

Source: Author, via Python

The results show that Cluster 2 stands out with the highest average sales and profit among all groups (Sales = 541.16; Profit = 72.49), despite having the smallest customer base. This reinforces the previous conclusion that Cluster 2 represents VIP customers who place high-value orders and generate significant revenue for the business, even if their purchase frequency is not particularly high.

In contrast, Clusters 0 and 1 have lower average sales and profits (Sales ≈ 160–170; Profit ≈ 16–18). These are likely general or new customers, with smaller order values and limited spending capacity.

Cluster 3 shows higher average sales than Clusters 0 and 1 (Sales = 195.75), but the average profit is not significantly higher (Profit = 21.76). This suggests that this group exhibits stable purchasing behavior, with a high number of orders but moderate value per transaction, aligning with the previously identified characteristics of loyal, high-frequency customers.

An analysis of customer lifecycle by cluster was also conducted, with the results illustrated in Figure 14.

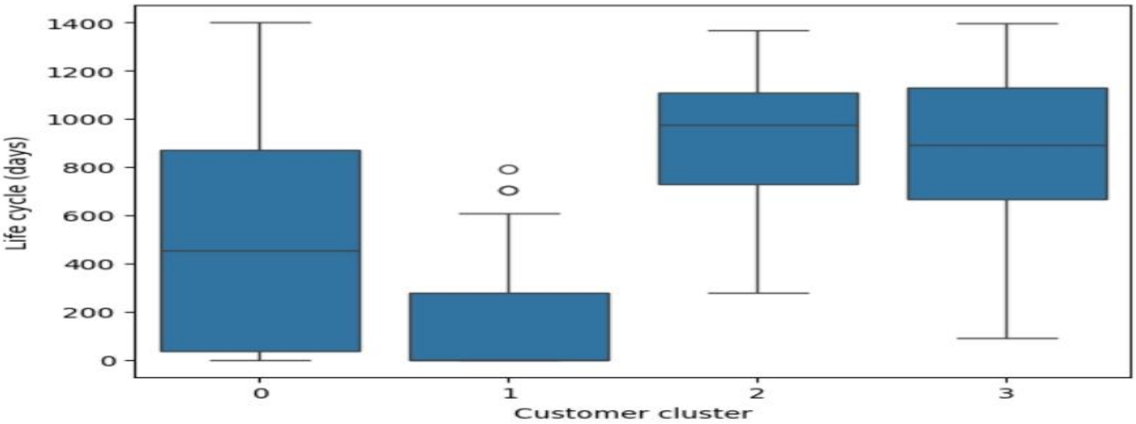

Figure 14: Customer Lifetime by customer cluster

Source: Author, via Python

The results show that:

Clusters 2 and 3 have the longest customer lifecycles (median over 800 days), indicating that these are long-term loyal customers who hold high value for the business. The company should consider implementing loyalty programs, VIP offers, or personalized strategies to maintain long-term relationships. Additionally, cross-selling and up-selling campaigns can help increase the average transaction value.

Cluster 0 exhibits the greatest variability in customer lifecycles, with some very long-term customers, but most having shorter durations (median ≈ 450 days). It is advisable to further analyze behavioral patterns that may signal a transition toward loyalty. Periodic promotional campaigns or discounts for the next purchase can help increase purchase frequency.

Cluster 1 has the shortest lifecycle (median ≈ 150 days), suggesting that this group mainly consists of new customers or one-time/occasional buyers. It is recommended to implement onboarding campaigns to help new customers better understand the products/services. Early reminders or offers can encourage repeat purchases and reduce churn risk.

An analysis of order volumes by shipping method (Ship Mode) and customer cluster was also conducted, with the results illustrated in Figure 15.

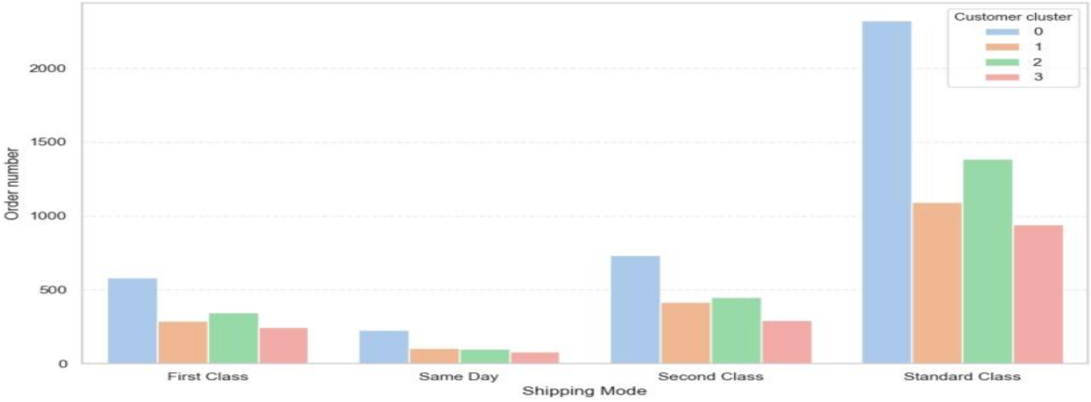

Figure 15: Number of orders by ShipMode and customer cluster

Source: Author, via Python

The chart reveals clear differences in shipping method preferences across customer clusters:

Cluster 0 has the highest number of orders across all shipping methods, especially Standard Class with over 2,200 orders. This reflects a customer group that places frequent, high-volume orders but has lower urgency for fast delivery.

Clusters 2 and 3 tend to favor Standard Class and Second Class shipping more than other methods. This suggests that these customers prioritize a balance between cost and delivery time.

In contrast, cluster 1 has the fewest total orders compared to the other groups, with a more even distribution across shipping methods. This may indicate a group of trial or irregular buyers without consistent purchasing habits.

Notably, same day delivery has the lowest usage across all clusters, implying that the high cost of this service is not a priority for most customers.

These findings provide valuable insights for businesses to adjust their operational strategies and design targeted shipping policies that align with the preferences of each customer segment.

5. Conclusion

The customer segmentation analysis using the RFM model combined with the K-Means algorithm identified four customer clusters with distinct consumption behaviors. These clusters clearly reflect the differentiation between high-value and loyal customers versus low-engagement groups, thereby providing valuable insights for marketing strategies and customer relationship management.

This study not only illustrates the application process of RFM and K-Means in customer segmentation but also complements it with statistical validation and a discussion on model reliability. This strengthens the validity and generalizability of the approach, contributing to the theoretical foundation of integrating the RFM model with clustering algorithms in quantitative marketing research.

One limitation of the study lies in the use of secondary, publicly available data (the Superstore dataset), which means the findings are exploratory and may not fully capture the specific characteristics of the Vietnamese market. Future research should apply the RFM model (or extend it to RFMX by incorporating features such as product type, transaction channel, and CRM feedback) to real-world datasets from Vietnamese enterprises. At the same time, combining statistical testing with model stability assessments (e.g., cross-validation, Silhouette, Davies–Bouldin) will further enhance the reliability and applicability of the results. This is particularly important as Vietnamese businesses accelerate digital transformation and require customer analytics tools that are both rigorous and practical.

REFERENCES

Vianna Filho, A. L. C., de Lima, L., & Kleina, M. (2025). A graph-based approach to customer segmentation using the RFM model. *Optimization and Control*https://doi.org/10.48550/arXiv.2505.08136.

Chen, D., Sain, S. L., & Guo, K (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3), 197–208. https://doi.org/10.1057/dbm.2012.17.

Christy, A. J., Umamakeswari, L., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 634–640. https://doi.org/10.1016/j.jksuci.2018.09.00.

Chitra, J., & Heikal, J. (2024). Customer segmentation using the K-Means Clustering algorithm in Foreign Banks in Indonesia. *Indonesia Accounting Research Journa*l, 11(4), 230–241. Retrieved from https://journals.iarn.or.id/index.php/Accounting/article/view/289.

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics*, 2(4), 809–823. https://doi.org/10.3390/analytics2040042.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.

Rahman, M. A., & Khan, M. M. R. (2021). Customer segmentation using RFM model and K-means clustering in the banking sector of Bangladesh. *Journal of Retailing and Consumer Services*, 60, 102130. https://doi.org/10.1016/j.jretconser.2020.102130.

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Researc*h, 42(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415.

Wang, S., Sun, L., & Yu, Y. (2024). A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. *Scientific Reports*, 14, 13421. https://doi.org/10.1038/s41598-024-57266-y.

Brahmana, R. K., Wardani, N. A., & Dewi, I. K. (2020). A comparative study of K-means, K-medoids and DBSCAN clustering algorithms on RFM model. *Lontar Kompute*r, 11(2), 87–95. https://doi.org/10.24843/LKJITI.2020.v11.i02.p06.

Sitorus, E. R., Lubis, A. R., & Sihombing, H. (2021). Customer segmentation using RFM and K-means in bottled water industry. *Jurnal Serambi Engineering*, 6(3), 1215–1224. https://doi.org/10.32672/jse.v6i3.3684.

Thông tin tác giả:
1. Trần Thị Mai
- Đơn vị công tác: Trường Đại học Kinh tế và Quản trị kinh doanh
- Địa chỉ email: *tranthimai879@gmail.com*
2. Nguyễn Quỳnh Hoa
- Đơn vị công tác: Trường Đại học Kinh tế và Quản trị kinh doanh

Ngày nhận bài: 7/8/2025
Ngày nhận bản sửa: 20/8/2025
Ngày duyệt đăng: 26/8/2025