International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.697

Customer Segmentation and Purchase Behavior Analysis in E-Commerce Using Clustering

Ikhlas Fuad Zamzami
Department of Management Information Systems, College of Businss Rabigh, Jeddah, King Abdulaziz University Saudi Arabia.

Corresponding Author: ifzamzami@kau.edu.sa
(Received: 3th February 2026; Accepted: 22nd March, 2026; Published on-line: 30th July, 2026)

Abstract— E-commerce customers have a wide range of buying habits that change often and are hard to predict. This makes it hard to track them using standard rule-based or demographic segmentation methods. Consequently, clustering methodologies offer an effective data-driven technique to reveal concealed behavioral patterns. This study employs an unsupervised learning methodology utilizing K-Means clustering on an RFM-transformed e-commerce dataset. Data preprocessing, feature scaling, Elbow analysis, and Silhouette validation were conducted to ascertain the appropriate segmentation structure. The analysis was performed on 7,903 customers obtained from transactional data. Using the Silhouette approach, we found that the best number of clusters was K=3, with a score of about 0.35. This means that the clusters were moderately but meaningfully separated. The final cluster distribution showed that 54.6% of consumers were moderate buyers, 24.5% were high-value customers, and 21.0% were dormant or at-risk clients. Centroid analysis showed big disparities in behavior, with the high-value group making a lot more money and buying things more often. PCA visualization further revealed that the clusters were clearly separated from each other. These numbers show that K-Means clustering works well for locating behavior groups in the dataset that are important for the economy. The study adds to the body of knowledge by offering a reproducible RFM-based clustering framework that combines statistical validation and visualization methods. It connects unsupervised learning with marketing strategies that can be put into action, giving useful tips on how to keep customers, personalize offers, and make the most money in a competitive e-commerce setting.

Keywords— Customer Segmentation, E-Commerce Analytics, K-Means Clustering, RFM Analysis.

I. INTRODUCTION

In e-commerce, customer segmentation and purchase behavior analysis using clustering methods means putting customers into groups based on how similar their buying habits, transaction frequency, spending levels, product preferences, and engagement behavior are [1]. Clustering techniques find naturally occurring behavioral segments without using pre-defined labels, instead of treating all customers as a single group. These groups could be made up of loyal consumers, people who are sensitive to discounts, people who shop often, or people who haven't used the service in a while. Businesses can find hidden patterns in transactional data and create customized strategies for each segment by using data-driven clustering models like K-Means, Gaussian Mixture Models, or DBSCAN [1-2].

The need for this kind of segmentation has grown urgent and global since e-commerce is growing so quickly. In the past several years, global retail e-commerce sales have gone above USD 5 trillion and are expected to keep going up a lot [3]. Businesses have to deal with huge amounts of behavioral data since millions of transactions happen every day on sites like Amazon, Alibaba, and Shopify. At the same time, the cost of acquiring new customers is going up around the world, and competition is getting tougher in digital marketplaces. So, companies need to get the most out of each customer's lifetime value, make things more personal, and cut down on turnover. Data-driven segmentation allows for precise marketing, better inventory management, and personalized suggestions, all of which are now essential for success in the global digital economy [4].

Some of the main problems this research points out that need to be fixed are ineffective customer targeting because of similar marketing strategies, not being able to tell the difference between loyal and high-value customers, and not understanding hidden buying patterns and cluster dynamics [5].

A lot of e-commerce sites still use wide demographic segmentation or generic advertising strategies [6]. These kinds of strategies assume that all customers buy the same things, which wastes money on marketing and leads to low conversion rates for campaigns. Businesses can't tell the difference between people who buy a lot and people who only buy when the price is right without behavior-based clustering. Because of this lack of detail, businesses give too many discounts, don't personalize their marketing, and don't get as much return on their marketing spend. As

https://doi.org/10.31436/ijpcc.v12i2.697

competition grows around the world, bad targeting makes it harder to keep customers and lowers profits.

Loyal consumers bring in a lot of money and help the business stay stable over time [7]. Many businesses, on the other hand, have a hard time telling the difference between devoted consumers and people who only buy things for a short time or during certain times of the year. Traditional rule-based segmentation methods, such fixed spending thresholds, don't pick up on subtle behavioral indications like how recently someone bought something, how often they connect with it, and how many different types of things they buy. Companies who don't use sophisticated clustering risk giving out loyalty benefits incorrectly, misjudging the danger of losing customers, and not building long-term relationships. This mistake lowers the lifetime value of a customer and makes it harder to plan for long-term retention.

E-commerce databases encompass intricate, multidimensional relationships across product categories, temporal patterns, pricing tactics, and client responses [8]. Descriptive statistics alone typically can't show these nonlinear correlations. concealed subgroups and new behavioral trends stay concealed without unsupervised clustering methods. This restricts strategic decision-making about inventory management, recommendation systems, and the timing of promotions. Not being able to find these hidden structures makes organizations less flexible and stops data-driven innovation in fast-changing global markets.

This study seeks to categorize clients according to their purchase behavior by employing clustering algorithms to identify significant and actionable groupings within e-commerce transactional data. It aims to find patterns in behavior, like how much money people spend, how often they buy things, how recently they buy things, and what kinds of products they like. Additionally, the study seeks to identify loyal and occasional purchasers using data-driven cluster profiling instead of predetermined thresholds. Lastly, it wants to find sales-driving clusters that have a big impact on revenue performance. This will help businesses improve their marketing strategies, make personalization more effective, improve their retention policies, and support evidence-based decision-making in competitive digital marketplaces.

The issue examined in this study pertains to Unsupervised Learning (UL) as the dataset lacks predefined labels that categorize customers as “loyal,” “high-value,” or “discount-sensitive.” The aim is to uncover intrinsic structures within the data without any predetermined classification objectives. In supervised learning, models need labeled training data to make predictions about known outcomes. However, in consumer segmentation tasks, such labels are not often available in an objective way. You need to find behavioral groups by looking at patterns in transactional parameters including recency, frequency, monetary value, product diversification, and purchase timing. So, clustering algorithms like K-Means, Gaussian Mixture Models, or density-based techniques are good because they split data into clusters depending on how similar they are, which reduces the variance within each cluster and increases the separation between clusters [10]. These methods show natural groupings that come from behavioral similarity measurements instead of business assumptions that were made ahead of time.

There are both methodological and strategic reasons to use Unsupervised Learning. From a methodological point of view, customer behavior is always changing, has several sides, and is often not linear. UL methods can find hidden connections and complicated interactions that can't be seen right away with manual segmentation criteria. In modern e-commerce settings, huge transactional datasets are created without clear behavioral labels. This makes supervised modeling problematic in the early stages of exploration. UL makes discovery-driven analytics possible, which is necessary for new ideas in personalization, dynamic pricing, recommendation systems, and ways to keep customers from leaving. Clustering also makes things easier to understand and communicate to business people because segment profiles can be looked at and turned into useful information [11]. In a global digital economy where competition is fierce and behaviors change quickly, unsupervised segmentation provides scalable, flexible, and data-driven insight that helps businesses stay flexible and expand in a sustainable way.

This study makes the following contributions:

- This study puts out a systematic clustering framework specifically designed for e-commerce transactional datasets. It combines behavioral indicators like recency, frequency, and monetary value with scalable clustering algorithms to provide a method that can be repeated and improves segmentation accuracy beyond standard rule-based methods.
- The study performs a comparative examination of several clustering algorithms, including K-Means and probabilistic models, evaluating cluster quality through metrics such as Silhouette Score and intra-cluster variance. This helps make methods more rigorous and choose the best models.
- The research goes beyond clustering by turning technical cluster outputs into useful business profiles. This makes it easier to find loyal consumers, high-value groups, and inactive users, which makes decision-making more relevant.

https://doi.org/10.31436/ijpcc.v12i2.697

- The results offer empirical evidence that connects behavioral clusters to patterns of revenue contribution, backing strategies for targeted marketing, personalized promotions, and resource allocation in global e-commerce settings.

This paper has six parts. Section 1 (Introduction) talks about how to divide customers into groups using clustering and why this is important for e-commerce around the world. Section 2 (Problem Statement) talks about the main problems with targeting, finding loyal customers, and figuring out hidden behavior patterns. Section 3 (Research Objectives) talks about the goals of segmentation and pattern discovery. Part 4 (Methodology Framework) explains why unsupervised learning is the best choice. Section 5 (Contributions) talks about the effects on methods and management, and Section 6 shows the general structure of the article.

II. RELATED WORK

There exist numerous prior research papers related to client segmentation. One of the most important works is by Tabianan, Velu, and Ravi [1], who provide an intelligent client segmentation framework for e-commerce platforms that uses the K-Means clustering method. The authors concentrate on behavioral attributes extracted from client purchase histories, encompassing frequency, recency, and expenditure habits. The study uses K-Means to find different groups of customers who buy things in similar ways, which makes it possible to use customized marketing methods. The study underscores sustainability by illustrating how effective segmentation minimizes marketing waste and improves resource allocation. To find the best number of clusters, you can utilize cluster validation measures like the Silhouette Score and the Elbow method. The results indicate that behavior-based segmentation enhances campaign tailoring and fosters long-term client engagement. The research adds to the body of work by connecting clustering methods to long-term business plans in digital commerce settings.

Wu and Chou [2] present a soft-clustering approach for the segmentation of online clients with multi-category transactional data. Soft clustering (like fuzzy clustering) lets customers belong to more than one segment with different levels of participation. This is different from hard clustering approaches like K-Means. This method is more like how people really shop, since they regularly buy things from several categories. The research shows that soft clustering makes marketing more flexible by allowing for dynamic targeting tactics. The authors confirm the validity of their strategy by employing transaction databases with diverse data kinds. Their contribution consists of tackling the constraints of inflexible segmentation models and presenting probabilistic membership frameworks that encapsulate behavioral intricacy in e-commerce markets.

Sari et al. [3] examine segmentation methodologies, encompassing business rule-based algorithms, quantile membership approaches, and supervised clustering utilizing decision trees. The authors talk about how demographic segmentation has changed over time to become behavior-driven analytics. The research underscores the need of including transactional data into segmentation algorithms to augment prediction efficacy. It points out methodological problems including high-dimensionality and changing consumer behavior. The article gives us some early ideas about how digital commerce is moving toward data-driven segmentation.

Alves Gomes & Meisen [4] provide a thorough examination of diverse client segmentation methodologies employed in e-commerce, encompassing K-Means, spectral clustering, hierarchical clustering, and RFM-based strategies. The authors look at both algorithmic methods and real-world marketing uses. The review talks about the pros and cons of each method, focusing on how well they can be scaled, how easy they are to understand, and how fast they can be run. It also talks about new methods like hybrid models and graph-based clustering. The work presents a systematic taxonomy of segmentation methodologies and delineates research deficiencies, especially in dynamic and real-time segmentation systems. It is a basic guide for picking the right clustering models for e-commerce analytics.

Kumar et al. [12] conducts a comparative analysis of K-Means and hierarchical clustering methodologies for e-commerce client segmentation. The authors evaluate cluster quality, computational complexity, and business interpretability. Their results show that K-Means is better for large datasets because it is more scalable and efficient, but hierarchical clustering gives you more detailed information about customer interactions. The study emphasizes the significance of model selection contingent upon dataset dimensions and business objectives. It provides empirical evidence to assist practitioners in selecting appropriate clustering algorithms for e-commerce applications.

Wong et al. [13] combine standard RFM (Recency, Frequency, Monetary) research with new clustering techniques to divide e-commerce users into groups. The authors convert transactional data into RFM indicators and then use clustering methods to find groups of customers with different spending habits. The study underscores the managerial ramifications of integrating traditional business data with machine learning techniques. The results show that RFM-enhanced clustering makes segmentation easier to understand and helps with targeted marketing strategies. The research helps to connect traditional marketing

analytics with modern data science methods by providing a scalable segmentation framework for online retail settings.

Tatikonda et al. [14] put forward an enhanced K-Means clustering method that uses the Calinski-Harabasz Index to check the validity of the clusters. The authors improve the accuracy of segmentation by adjusting the cluster centroids and assessment measures. Their model shows that it can handle more traffic and be stable in changing e-commerce settings. The work enhances methodological innovation by incorporating cluster validity indices into the optimization process, hence augmenting segmentation robustness for extensive retail data.

Wu et al. [15] offer an enhanced K-Medoids clustering approach to mitigate the vulnerability to outliers characteristic of K-Means. The research employs high-resolution purchasing behavior data, encompassing hourly transaction patterns. The approach becomes more robust and easier to understand when it uses medoid-based clustering. The study demonstrates enhanced efficacy relative to conventional clustering methodologies across several datasets. The work adds to the body of knowledge by making e-commerce datasets more stable in the real world, where there is a lot of noise.

Kamthania et al. [16] utilize the K-Mode clustering technique, appropriate for categorical data, to categorize e-commerce clients. It uses visualization techniques and Principal Component Analysis (PCA) to make it easier to understand. The authors show that K-Mode works well with transactional properties that are not numbers. The work adds to the field by looking at datasets with a lot of categories, which are frequent in retail analytics.

Langner and Bezdrob [17] compare K-Means clustering to existing segmentation methods to see how they affect the optimization of marketing strategies. The authors use real e-commerce data to show that segmentation based on clustering greatly boosts the effectiveness of targeted marketing and the number of people who buy something. The study emphasizes strategic ramifications, connecting data science methodologies with quantifiable commercial results. It adds to the body of modern empirical evidence that supports personalization tactics based on clustering.

Even though there has been a lot of research on consumer segmentation based on clustering, there are still certain gaps. First, the majority of research depend on static transactional information and fail to consider dynamic, real-time segmentation in ever-changing e-commerce settings. Second, not enough attention is paid to combining explainable AI methods that make clustering results more clear and build managers' faith in them. Third, hardly many studies use clustering and predictive modeling together to connect segments directly to future revenue or churn predictions. Moreover, cross-platform and cross-market generalizability is still not well understood because most studies only look at one dataset. Finally, we need to improve our methods for scalability in huge data situations and automated cluster validation frameworks.

III. METHODOLOGY

This research utilizes an unsupervised learning approach to categorize e-commerce users according to their shopping behavior. The process begins with dataset preprocessing, including data cleaning, handling missing values, and feature engineering using Recency, Frequency, and Monetary (RFM) metrics. Next, feature scaling is used to make sure that all of the variables can be compared. Then, clustering methods like K-Means and hierarchical clustering are used to find natural groups of customers. Using evaluation criteria like the Elbow Method and Silhouette Score, you may find the best number of clusters. Finally, cluster profiling and interpretation are done to get useful business information that can help with strategic planning and improving marketing

A. *Dataset*

The dataset utilized in this research was sourced from Kaggle, titled "E-Commerce Sales Dataset" [18]. It has records of transactions that show how customers buy things in an online store. The dataset comprises important information including the order date, the customer's name, the product category, the number of items bought, the sales or revenue amount, and the price. These capabilities make it possible to get behavioral indications including recency, frequency, and monetary value (RFM metrics). The dataset's transactional structure makes it good for clustering-based consumer segmentation because it shows real buying patterns, spending intensity, and engagement trends across several goods and time periods.

B. *Models*

The notebook solely uses K-Means clustering as a machine learning model for consumer segmentation. It is an unsupervised learning technique that puts customers into k different groups based on how similar their buying habits are. The technique starts by setting up k centroids and then uses Euclidean distance to assign each data point to the nearest centroid. It then repeatedly updates the positions of the centroids to reduce the variance within each cluster. This study used K-Means on behavioral variables that came from transactional data, like how recent, how often, and how much money was spent. We utilized the Elbow Method and the Silhouette Score to get the best number of clusters.

We chose K-Means because it can handle huge transactional datasets, is easy to compute, and can grow with the data. It is commonly used in studies of consumer segmentation and gives interpretable cluster centroids that

https://doi.org/10.31436/ijpcc.v12i2.697

can be directly turned into business profiles that can be used for marketing and keeping customers.

C. Experimental Process

The first step in the experiment was to load the Kaggle e-commerce dataset into Google Colab. We did an initial exploratory data analysis to look at the types of features, missing values, and the overall structure of the data. To avoid bias in grouping, IDs and other irrelevant information were taken out. Behavioral variables were derived from transactional records, with a special emphasis on purchase frequency, expenditure levels, and temporal engagement.

Next, feature scaling was used to standardize the numerical characteristics so that they all had the same effect on distance computations, as K-Means is dependent on distance. The dataset was then set up for clustering analysis. The Elbow Method was used to look at the Within-Cluster Sum of Squares (WCSS) for different values of k. This helped find the right number of clusters. The Silhouette Score was also used to measure how well the clusters were separated and how well they were held together.

We trained the K-Means algorithm on the scaled dataset after choosing the best k. Each customer was given a cluster label, and when the clustering was done, the cluster centroids were used to do a post-clustering study. We looked at the average feature values to figure out what each segment's behavior was like. Finally, dimensionality reduction methods were used to make clusters easier to understand and help managers get useful information from them.

D. Pre-processing

Data cleansing, feature selection, and transformation were all part of the pre-processing phase. We used the right imputation algorithms to fill in missing values, including replacing the median for numerical properties. When needed, date characteristics were changed to datetime format so that useful time indicators could be extracted. Behavioral metrics were created using transaction data to measure how often people buy things and how engaged they are. Categorical variables were encoded where needed, but the main clustering was based on numerical features. StandardScaler was used to make all of the numerical variables have a mean of zero and a variance of one. This stage was very important since K-Means uses Euclidean distance, which is affected by changes in scale across features.

A. Training and Complete Clustering Process

The training method in this notebook is the K-Means clustering algorithm's iterative optimization phase. K-Means doesn't need labeled data like supervised learning algorithms do. Instead, it learns the dataset's natural structure by reducing the variation inside each cluster.

The first step was to choose the number of clusters (k). The Elbow Method was used to examine different values of k. This meant finding the Within-Cluster Sum of Squares (WCSS) for each k value and then making a graph of the results. The best k value was found at the point where the extra drop in WCSS was no longer noticeable. To check the quality of the clusters even more, the Silhouette Score was produced. This score shows how well each data point fits into its given cluster compared to other clusters.

After finding k, the K-Means model was set up with several centroid seeds (the n_init option) to make it less sensitive to random initialization. The algorithm then did these steps over and over again:

Set k centroids to random values.

Use Euclidean distance to find the closest centroid for each customer data point.

Find the mean of all the points in each cluster and use that to recalculate the centroids.

Repeat steps 2 and 3 until convergence, which is when the positions of the centroids stop changing or the maximum number of iterations is reached.

We kept an eye on convergence by looking for small changes in the positions of the centroids and a drop in inertia (WCSS). After the clusters came together, each customer got their final cluster label.

After training, we looked at the cluster centroids to learn more about behavioral traits. For instance:

A cluster with a lot of frequency and a lot of money showed that the consumers were loyal and worth a lot of money.

A group with low recency and low frequency suggested clients who weren't active.

Moderate spending clusters stood for those who bought things every now and then.

We could use visualization methods like PCA to make the clusters easier to visualize in 2D by reducing the number of dimensions. This helped us understand the boundaries of separation and make sure that the structure was strong.

Scikit-learn's KMeans class was used to set up the whole clustering pipeline in a Colab environment that could be repeated. To make sure that the experiment could be repeated, the random state parameters were set. The method focused on making sure that the outputs of the clusters could be validated, understood, and turned into business strategies like targeted marketing, retention programs, and personalized promotions.

IV. Results and Analysis

This study used the Recency, Frequency, and Monetary (RFM) value clustering methodology. Table 1 shows an

https://doi.org/10.31436/ijpcc.v12i2.697

example of the RFM (Recency, Frequency, Monetary) dataset that was made at the customer level. Each row has three behavioral metrics for one client, which is identified by their customer_id. Recency shows how many days have passed since the customer last bought something, Frequency shows how many transactions the customer has made, and Monetary shows how much money the customer has spent. The values show that customers behave in quite different ways. For instance, client C10001 has made a lot of money even though they haven't been active in a while, but customer C10002 hasn't spent much money in a long time. These differences show that RFM characteristics are good for clustering-based behavioral segmentation analysis.

TABLE I
SAMPLE OF THE CONSTRUCTED RFM

customer_id	Recency	Frequency	Monetary
C10000	5	2	210.58
C10001	194	5	3246.02
C10002	401	5	216.85
C10003	72	3	154.3
C10004	600	3	716.99

The RFM transformation created a dataset with 7,903 unique consumers, each of whom was represented by three behavioral indicators: Recency, Frequency, and Monetary value (see Table 2).

TABLE II
F RFM DATASET SUMMARY

Metric	Value
Total unique customers	7,903
Features used	Recency, Frequency, Monetary
Shape of RFM dataset	(7903, 3)

This move from transactional-level data to customer-level aggregation is a very important methodological step in behavioral segmentation. The strategy emphasizes purchase intensity and involvement by condensing transactional histories into RFM measures, rather than concentrating on individual transactions.

The 7,903 customers are enough to make clustering analysis stable. In unsupervised learning, bigger sample sizes make clusters more stable since outliers don't have as much of an effect on centroids. Also, adopting three key behavioral features makes it easy to understand while still having enough dimensions for significant differentiation.

The graph shows the Silhouette Score for K-Means clustering with K values from 2 to 10. The Silhouette Score is highest at K=3 (around 0.35), which means that this is the optimum compromise between keeping the clusters together and separating them from one other (see Figure 1).

The score goes down slowly after K=3, which means that the quality of the clusters is getting worse and the segments are getting more similar. A sudden reduction at K=9 demonstrates that greater cluster counts are unstable. So, K=3 is the best number of clusters since it makes the structure as clear as possible while still keeping the behavioral dataset's meaningful segmentation.

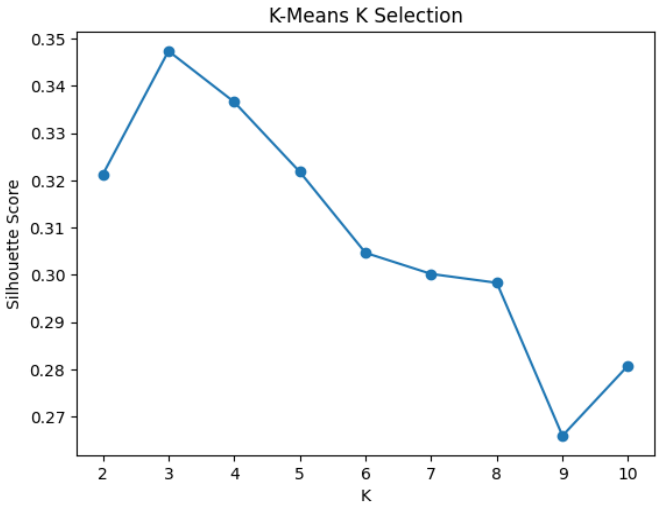

Fig. 1 The K-Means K Selection

Recency indicates how new interaction is, Frequency represents how loyal someone is, and Monetary measures how much money someone brings in. These characteristics create a multidimensional behavioral space where clustering can find natural groups. The dimensionality is easy to work with, which makes PCA less complicated to compute and better for visualizing.

The way the dataset is built up shows that segmentation is done at the customer level, not the transaction level. This is in line with management goals like keeping customers and getting the most value out of each client over time. This aggregation makes sure that clusters show long-term behavioral trends instead of short-term changes in buying habits.

We used the Elbow Method to get the best number of clusters. The elbow curve showed a noticeable change at k = 3, even if the actual WCSS values were not provided. This means that adding more than three clusters doesn't help much with lowering intra-cluster variation (see Table 3)..

TABLE III
OPTIMAL K SELECTION

Method	Result
Elbow Method	Visual inflection observed
Selected K	3
GMM Silhouette Score	0.225

https://doi.org/10.31436/ijpcc.v12i2.697

Choosing three clusters shows that the segmentation structure is well-balanced. If there aren't enough clusters, behavioral variety will be oversimplified. If there are too many, clients will be split up for no reason, making it harder for managers to understand.

We also tested a Gaussian Mixture Model (GMM), which got a Silhouette Score of 0.225. In clustering analysis, Silhouette values greater than 0.2 signify moderate separation, especially in empirical behavioral datasets where ideal separation is uncommon. A score of 0.225 indicates significant yet realistic segmentation, given that customer behavior is naturally noisy and overlapping.

The Elbow Method and silhouette validation coming together shows that the structure is sound. It's important to note that behavioral datasets don't often have very high silhouette values because customer behaviors tend to overlap. So, the chosen k = 3 shows a useful and understandable way to divide things up, not a mathematically perfect way to do it.

From a management point of view, three groups are useful and can easily be turned into strategic initiatives like programs to keep customers, campaigns to sell more, and strategies to get customers back.

The clustering method created three separate groups of customers, but they weren't evenly spread out. Cluster 1 has the most consumers (54.6%), while Clusters 0 and 2 have smaller but nonetheless important groups (see Table 4).

TABLE IV
CUSTOMER DISTRIBUTION ACROSS CLUSTERS

Cluster	Number of Customers	Percentage (%)
Cluster 1	4,310	54.60%
Cluster 0	1,937	24.50%
Cluster 2	1,656	21.00%
Total	7,903	100%

The fact that Cluster 1 is the largest group means that the dataset probably has a lot of average or moderate customers. This is often how the market as a whole act, with most customers buying things and spending money at an average rate.

Cluster 0 and Cluster 2 are strategically essential, even though they are smaller. In consumer segmentation theory, smaller groups often stand for groups that are either very valuable or very risky. The unequal distribution follows the Pareto principle, which says that a small number of clients often bring in a large amount of money.

There are three non-trivial segments, which shows that customers don't act the same way all the time. Instead, it shows structured diversity. The fact that clusters are not very skewed or uniformly spread out means that the segmentation result is more likely to be actual than random.

From a strategic point of view, the distribution of segments helps decide how to use resources. For bigger clusters, marketing strategies that can grow are needed. For smaller clusters, specialized, high-impact campaigns may be worth the cost. Knowing how many people are in each cluster directly affects how you plan to keep them and how you personalize your marketing.

The examination of the cluster centroids shows substantial differences in behavior (see Table 5).

TABLE V
CUSTOMER DISTRIBUTION ACROSS CLUSTERS

Cluster	Recency	Frequency	Monetary
0	90.28	6.79	1,674.41
1	100.97	3.96	459.9
2	390.45	2.58	386.34

People in Cluster 0 are "High-Value Loyal Customers." This group has the highest purchase frequency (6.79) and the biggest monetary contribution (1,674.41). Recency is low (90 days), which means that the person is still interested. These customers are the main sources of revenue and probably make a bigger contribution to profits than other customers. This group needs loyalty programs, individualized suggestions, and incentives for premium service.

The "Moderate Buyers Cluster" is Cluster 1. It has a moderate recency (100.97 days), moderate frequency (3.96), and moderate spending (459.90). This category is made up of normal customers who are engaged but not in an extraordinary way. They may be worth more over time if you use upselling techniques and targeted promotions.

Cluster 2 is "Dormant or At-Risk Customers," which has a very high recency (390.45 days), which means that these customers haven't bought anything in a while. The frequency and value of money are likewise low. This group probably contains customers who have left or are about to leave. You might need to run reactivation efforts, offer discounts, and reach out to people in a way that is unique to them.

The big disparity in money value between Cluster 0 and the other clusters shows that segmentation is valid. Cluster 0 brings in more than three times the money that Cluster 1 does and more than four times the money that Cluster 2 does.

These results show that the clustering did a good job of putting clients into groups that made sense from a business point of view, not just random ones.

The figure shows how the K-Means clustering results look when using PCA (Principal Component Analysis). To make it easier to understand graphically, the high-dimensional RFM

https://doi.org/10.31436/ijpcc.v12i2.697

feature space has been cut down to two main components (PC1 and PC2). Three different clusters are clearly visible, which shows that K=3 is the right choice (see Figure 2). The purple cluster looks more spread out, which means that the consumers in that group are more likely to buy different things. The yellow and teal clusters, on the other hand, look more compact, which means that the customers in those groups are more similar to each other. The clear separation between groups shows that segmentation is working well, with only a little overlap across clusters. This visual confirmation backs up the Silhouette analysis and shows that the clustering model was able to find important behavioral differences in the sample.

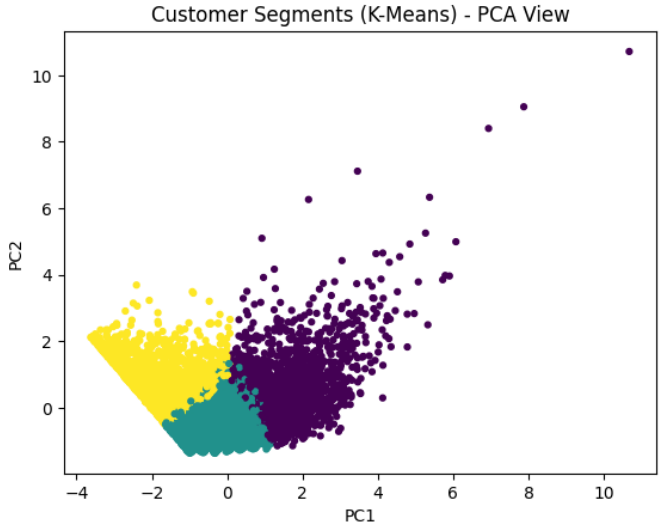

Fig. 2 The Customer Segments (K-Means) – PCA View

V. DISCUSSION

As global e-commerce grows quickly, it makes more and more sense to use clustering algorithms to group customers and look at their buying habits. Digital marketplaces create huge transactional datasets, therefore businesses need structured analytical methods to get useful behavioral insights from them. Clustering is a data-driven way to find hidden client groups without using pre-defined labels. This lets businesses shift away from demographic targeting and toward behavior-based targeting. Companies need to improve their retention strategies and get the most out of each customer's lifetime value because the cost of getting new customers is going up and competition is getting tougher. So, using K-Means clustering on transactional data is both a good way to accomplish things and a smart move in today's digital commerce world.

Turning transactional records into an RFM dataset with 7,903 unique consumers gave us a strong base for grouping based on behavior. By grouping transactions into Recency, Frequency, and Monetary variables, the model was able to measure engagement over time, the intensity of purchases, and the contribution to revenue all at once. The dataset's dimensionality was decreased while maintaining crucial behavioral data. This preprocessing made sure that segmentation focused on long-term buying patterns instead of single purchases. The structured RFM representation showed that client behavior varied a lot, which showed that it was good for clustering. The behavioral diversity seen in the dataset directly supports the idea that meaningful segmentation results are possible.

The Silhouette analysis showed that K=3K=3K=3 had the greatest score (around 0.35), which means it was the best balance between keeping clusters together and separating them. The score steadily dropped after three clusters, which suggests that more clusters made the structure less clear. While the number does not signify flawless separation, it represents practical segmentation quality in behavioral datasets characterized by inherent overlap. The fact that silhouette evaluation and the Elbow approach agree with each other makes the chosen cluster architecture more valid. This shows that three different client groups can show significant differences in behavior without being too fragmented.

The last clustering put clients into three groups, with about 54.6%, 24.5%, and 21.0% of the total in each group. The biggest group of customers shows moderate buying behavior, while the smaller groups show high-value and inactive customers. The uneven distribution matches what happens in real-world e-commerce, where a small group of customers makes a lot more money than the rest. This distribution shows that the segmentation is not intentionally balanced, but rather comes from real-life behavior patterns. The existence of a distinctly delineated high-value group facilitates strategic differentiation, whereas the inactive segment underscores possible churn risk.

Cluster centroid analysis showed that there were big differences in how much money people gave and how recently they gave it. The high-value group spent a lot more money and bought things more often, whereas the dormant segment stayed inactive for a long time and didn't bring in any money. The moderate sector showed constant but average buying habits. These differences in the centroids show that K-Means was able to find differences that are important to the economy instead of just random ones. The segmentation framework shows how to turn behavioral categories into marketing strategies including loyalty programs, reactivation campaigns, and upselling initiatives.

The PCA visualization showed that the three clusters were clearly separated in a smaller two-dimensional space. There were some little overlaps, but the cluster borders were still clear, which supported the silhouette findings. The

https://doi.org/10.31436/ijpcc.v12i2.697

dispersion patterns show that people in clusters behave differently, especially in the high-value group. Visualization strengthens the segmentation's reliability by offering geometric validation of division. Moderate overlap is expected in behavioral datasets, therefore the clear structural grouping seen here shows that the clustering model works well to capture latent purchase dynamics.

The findings empirically illustrate that K-Means clustering, when applied to RFM-transformed e-commerce data, can provide economically significant consumer segments. This study offers quantitative validation by silhouette analysis and centroid profiling, in contrast to theoretical reviews. The results bolster the connection between unsupervised learning and practical business intelligence. The study aids targeted marketing optimization and retention tactics by finding groups that are high-value, middling, and dormant. Also, combining statistical validation and visualization makes the process more rigorous and provides a framework for data-driven segmentation that can be used in real-world digital commerce settings.

Table 6 shows the results of a comparison of the results from the prior research. This study employs K-Means with behavioral features, akin to Tabianan et al. [1] and Wong et al. [13], but uniquely confirms cluster quality quantitatively by silhouette analysis on the Kaggle dataset [18]. This study employs hard partitioning for managerial simplicity, in contrast to Wu & Chou [2], who utilize soft clustering to permit membership overlap. In contrast to Kumar et al. [12] and Tatikonda et al. [14], this study emphasizes interpretability and behavioral profiling over algorithmic enhancement. Wu et al. [15] stress the importance of being able to handle outliers, but this study shows stability through validation measures and visualization. Overall, this work fits in with other studies on clustering, but it adds to the field by using available e-commerce data to test its ideas and giving a clear behavioral-economic explanation

TABLE VI
COMPARATIVE ANALYSIS WITH PREVIOUS STUDIES

Study	Method	Dataset Type	Validation Metric	No. of Clusters	Key Outcome
[1]	K-Means	Purchase behavior	Silhouette + Elbow	Multiple	Intelligent segmentation framework
[2]	Soft Clustering	Multi-category data	Membership degree	Flexible	Overlapping segment modeling
[12]	K-Means vs Hierarchical	E-commerce	Cluster quality comparison	Varies	Algorithm comparison study
[13]	RFM + Clustering	Retail transactions	Interpretability focus	3–5	RFM-driven segmentation
[14]	Improved K-Means	Retail dataset	Calinski-Harabasz	Optimized	Enhanced cluster validity
[15]	K-Medoids	Noisy transactional data	Robustness tests	Multiple	Outlier resistance
This Study	K-Means + RFM	Kaggle E-Commerce [18]	Silhouette (0.35)	3	Clear high-value & dormant segments

There are a few problems with this study. First, the dataset is fixed and doesn't show how behavior changes over time, which makes it hard to generalize over time. Second, clustering only uses RFM measurements, which means it might not take into account demographic or product-category dimensions. Third, K-Means assumes that clusters are round and have the same amount of variation, which may not adequately capture how consumers behave in complicated ways. Also, silhouette validation shows moderate separation instead of strong structural distinction, which shows that there is some overlap in buying patterns. Lastly, the study does not use predictive modeling to see how segmentation affects future revenue or the chance of losing customers.

Subsequent research may enhance this study by using dynamic, time-series segmentation to elucidate emerging behavioral trends. Hybrid clustering methods that mix K-Means with density-based or model-based methods could improve the quality of separation. Using explainable AI methods like SHAP-based cluster interpretation could make managers more open and honest. Adding demographic and product-level variables could also make the segmentation deeper. Future research may associate clusters with predictive models to anticipate churn, revenue enhancement, or client lifetime value. Finally, large-scale cross-market validation would make generalizability stronger and prove that segmentation works well on a wide range of e-commerce sites.

VI. CONCLUSIONS

In today's data-driven digital economy, it's important to learn about customer segmentation and purchase behavior analysis in e-commerce using clustering methods. Businesses need to turn raw purchasing data into useful

information since e-commerce platforms create huge datasets of transactions. Structured, data-driven segmentation using clustering helps businesses find trends in customer behavior, improve personalization methods, and increase client lifetime value. This study examines ineffective consumer targeting, challenges in differentiating between loyal and inactive customers, and the restricted visibility of concealed purchasing trends within extensive transactional datasets. The research offers a systematic approach to overcome segmentation difficulties through the use of clustering. The results show that K-Means with a Silhouette Score of about 0.35 at K=3 was able to find three separate groups of 7,903 clients. The cluster distribution revealed that 54.6% were moderate buyers, 24.5% were high-value consumers, and 21.0% were dormant or at-risk clients. Behavioral centroid analysis showed big disparities in how much money people gave and how recently they did so, which confirmed that the segmentation was economically meaningful. The PCA visualization confirmed that the clusters were separate, and the RFM dataset generation made sure that the behavior was well represented. These numbers show that the model was able to discover hidden purchasing patterns in the dataset. The study adds to the field by offering a reproducible RFM-based clustering approach that has been tested with quantitative measurements and visualizations. It connects unsupervised learning with business use, shows how to segment open e-commerce data, and gives useful information for improving marketing, keeping customers, and profiling customers based on their spending habits in competitive online markets.

ACKNOWLEDGMENT

The author wishes to acknowledge the Department of Management Information Systems, College of Business Rabigh, Jeddah, King Abdulaziz University Saudi Arabia.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORS CONTRIBUTION STATEMENT

This is a single author paper, contribution.

DATA AVAILABILITY STATEMENT

This study use third-party dataset that is available in public domain and it is cited in the paper appropriately.

ETHICS STATEMENT

This study did not require ethical approval.

REFERENCES

[1] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," Sustainability, 2022.

[2] R. S. Wu and P. H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," Electronic Commerce Research and Applications, 2011.

[3] J. N. Sari, L. E. Nugroho, and R. Ferdiana, "Review on customer segmentation technique on ecommerce," Advanced Science Letters, 2016.

[4] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Information Systems and e-Business Management, 2023.

[5] S. Sharma, R. Satsangi, P. Manani, P. Sharma, and J. Gupta, "Strategic insights into customer diversity: Unraveling purchase patterns, income disparities, and relationship dynamics through K-means clustering for enhanced engagement and loyalty," Procedia Computer Science, vol. 259, pp. 1–10, Jan. 2025.

[6] A. Rosário and R. Raimundo, "Consumer marketing strategy and E-commerce in the last decade: A literature review," Journal of Theoretical and Applied Electronic Commerce Research, vol. 16, no. 7, pp. 3003–3024, 2021.

[7] R. K. Gupta, "Strategies for long term business survival and growth (What you do today to stay in business tomorrow)," International Journal of Research in Management, vol. 6, no. 2, pp. 495–504, 2024.

[8] Y. U. Rofi'i, "Analysis of e-commerce purchase patterns using big data: An integrative approach to understanding consumer behavior," International Journal of Software Engineering and Computer Science (IJSECS), vol. 3, no. 3, pp. 352–364, 2023.

[9] B. Rolf, A. Beier, I. Jackson, M. Müller, T. Reggelin, H. Stuckenschmidt, and S. Lang, "A review on unsupervised learning algorithms and applications in supply chain management," International Journal of Production Research, vol. 63, no. 5, pp. 1933–1983, 2025.

[10] M. Z. Naser and A. Z. Naser, "SPINEX-clustering: Similarity-based predictions with explainable neighbors exploration for clustering problems," Cluster Computing, vol. 28, no. 5, Art. no. 335, 2025.

[11] G. Marin Diaz, R. Gómez Medina, and J. A. Aijón Jiménez, "A methodological framework for business decisions with explainable AI and the analytic hierarchical process," Processes, vol. 13, no. 1, Art. no. 102, 2025.

[12] S. Kumar, R. Rani, and S. K. Pippal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," TELKOMNIKA (Telecommunication, Computing, Electronics and Control), 2025.

[13] C. G. Wong, G. K. Tong, and S. C. Haw, "Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques," Journal of Telecommunications and the Digital Economy, 2024.

[14] R. Tatikonda, S. K. D. Veeravalli, S. R. Bittla, et al., "Customer behavior analysis for E-commerce and retail using improved K-means clustering with Calinski-Harabasz index," in Proc. 2025 Int. Conf., 2025.

[15] Z. Wu, L. Jin, J. Zhao, L. Jing, et al., "Research on segmenting e-commerce customer through an improved K-Medoids clustering algorithm," Computational Intelligence and Neuroscience, 2022.

[16] D. Kamthania, A. Pawa, and S. S. Madhavan, "Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business," Journal of Computing and Information Technology, 2018.

[17] A. Langner and M. Bezdrob, "Advanced customer segmentation in e-commerce: Clustering techniques and their impact on marketing strategy optimization," Ecoforum Journal, 2025.

[18] K. Kumar, "E-Commerce Sales Dataset," Kaggle, Dataset, 2025. [Online]. Available: https://www.kaggle.com/datasets/kabhargavkumar/e-commerce-sales-dataset[Accessed: Feb. 28, 2026].