Jurnal Ilmiah Matematika dan Terapan
Volume 22 Number 1 December 2025 (Pages 21 - 30)
doi: https://doi.org/10.22487/2540766X.2025.v22.i1.17884

JIMT
Jurnal Ilmiah Matematika dan Terapan

ISSN : 2540 – 766X

APPLICATION OF THE K-MEANS CLUSTERING ALGORITHM IN E-COMMERCE TRANSACTION PATTERN ANALYSIS

Agung Yuliyanto Nugroho

Universitas Cendekia Mitra Indonesia
agungboiler11@gmail.com

ABSTRACT

In the era of digital transformation, e-commerce platforms generate large volumes of transactional data that can be leveraged to understand customer purchasing behavior and support data-driven decision-making. This study applies the K-Means clustering algorithm to identify transaction patterns and segment customers based on Recency, Frequency, and Monetary (RFM) attributes. The analysis uses an e-commerce transaction dataset comprising 128,436 transactions from 8,912 customers recorded during the January–December 2023 period. The methodological stages include data cleaning and normalization, RFM feature construction, and customer segmentation using K-Means clustering. The Elbow Method and Silhouette Coefficient were employed to determine and validate the optimal number of clusters. The results indicate that k = 4 provides the best clustering structure, with an average silhouette score of 0.62, suggesting good cluster separation. The identified clusters represent distinct customer segments, including high-value loyal customers, frequent moderate spenders, occasional buyers, and low-value inactive customers. These clustering results offer practical insights for customer relationship management (CRM), targeted marketing, and demand forecasting by enabling firms to tailor strategies according to customer value and engagement levels. The findings demonstrate that K-Means clustering, when applied to RFM-based features, is effective in uncovering meaningful patterns in large-scale e-commerce transaction data. Future research should explore K-Means++ to improve centroid initialization, Gaussian Mixture Models (GMM) to capture overlapping customer segments, and density-based methods such as DBSCAN to identify noise and irregular purchasing behavior. Additionally, fuzzy c-means could be applied to model customer membership across multiple segments, addressing the limitations of hard clustering in dynamic e-commerce environments.

Keywords : K-Means Clustering, E-Commerce, Data Mining

I. INTRODUCTION

The rapid growth of e-commerce has transformed how businesses interact with consumers, generating large volumes of transactional data that record when customers purchase, how often they transact, and how much they spend. These data offer substantial opportunities for understanding customer behavior and improving marketing and operational strategies. However, the scale and complexity of e-commerce transaction data pose challenges for traditional analytical approaches, particularly those relying on manual segmentation or simple descriptive statistics. Customer segmentation is a critical component of e-commerce strategy, supporting personalized marketing, customer relationship management, and demand forecasting. Among various data mining techniques, K-Means clustering remains one of the most widely adopted methods due to its computational efficiency and interpretability. In practice, K-Means is frequently applied using Recency, Frequency, and Monetary (RFM) attributes, which summarize customer transactional behavior into compact and business-relevant indicators. RFM analysis is particularly suitable for e-commerce environments because it directly reflects customer engagement, purchasing intensity, and economic value—key dimensions for retention strategies and revenue optimization. Despite its popularity, several methodological gaps persist in the application of K-Means clustering for e-commerce transaction analysis. First, K-Means is sensitive to centroid initialization and assumes spherical, equally sized clusters, which may not fully reflect heterogeneous customer behavior. Second, RFM variables often exhibit skewed distributions and scale imbalances that can bias distance-based clustering if not properly normalized. Third, many studies rely solely on the Elbow Method without additional validation, limiting the robustness of cluster interpretation and managerial applicability.

This study addresses these gaps by applying K-Means clustering to e-commerce transaction data using a systematically constructed RFM framework. The analysis includes careful data preprocessing and normalization, validation of clustering results using both the Elbow Method and Silhouette Coefficient, and interpretation of customer segments through cluster centroid profiles. By focusing exclusively on RFM attributes, this study ensures methodological consistency between data, analysis, and interpretation. The contributions of this research are threefold. First, it provides a structured and replicable approach to RFM-based customer segmentation using K-Means clustering. Second, it evaluates cluster quality beyond visual inspection by incorporating quantitative validation metrics. Third, it translates clustering results into actionable customer profiles that support targeted marketing, customer retention, and revenue management strategies in e-commerce platforms. By offering a focused and methodologically grounded application of K-Means clustering, this study contributes to the practical use of data mining techniques in e-commerce analytics while highlighting both the strengths and limitations of distance-based clustering for customer segmentation.

II. METHODS

This research applies a quantitative and computational approach to analyze e-commerce transaction data using the K-Means clustering algorithm. The methodological framework is designed to identify transaction patterns, segment customer groups, and extract meaningful insights from large datasets. The main stages of the methodology include data collection, data preprocessing, feature selection, clustering implementation, and evaluation of clustering results.

2.1. Research Design

This study adopts an exploratory data mining design using unsupervised machine learning, with the objective of identifying latent customer segments from e-commerce transaction data without predefined class labels. The analysis is conducted using the Python programming language (version 3.10), supported by the Scikit-learn library (version 1.3.0) for machine learning implementation, Pandas (1.5.3) and NumPy (1.24.2) for data manipulation, and Matplotlib (3.7.1) for visualization.

The analytical workflow consists of the following sequential stages:

1. Data collection and verification
2. Data cleaning and preprocessing
3. RFM feature construction and transformation
4. Feature scaling and normalization
5. Determination of the optimal number of clusters
6. K-Means clustering implementation
7. Cluster validation and interpretation

2.2. Data Source and Ethical Considerations

The dataset used in this study is derived from an anonymized public e-commerce transaction dataset obtained from an open-access data repository (e.g., Kaggle), originally exported from an operational e-commerce platform. The dataset contains 128,436 transaction records (N_tx) corresponding to 8,912 unique customers (N_cust) over the period 1 January 2023 to 31 December 2023.

All personal identifiers were removed prior to analysis, and customer IDs were replaced with random alphanumeric codes to ensure anonymity. The dataset contains no personally identifiable information (PII), and therefore does not require formal ethical clearance. The study fully complies with data privacy and research ethics standards for secondary and anonymized datasets.

2.3. Data Cleaning and Preprocessing

Initial preprocessing was conducted to ensure data quality and analytical consistency. The following steps were applied:

1. Removal of duplicate transaction records
2. Handling of missing values in transaction amounts and timestamps
3. Standardization of date and currency formats

Contrary to approaches that integrate categorical variables, this study does not include product categories or payment methods in the clustering feature space. As the final clustering relies exclusively on RFM attributes, one-hot encoding was intentionally excluded to avoid distortion of Euclidean distance calculations.

2.4. RFM Feature Construction

Customer behavior was summarized using the Recency–Frequency–Monetary (RFM) framework, formally defined as follows:

1. Recency (R):
   The number of days between a customer’s most recent transaction and the reference date (31 December 2023).
2. Frequency (F):
   The total number of transactions conducted by a customer during the study period.
3. Monetary (M):
   The total monetary value of transactions made by a customer during the study period (not the average), measured in local currency.

This definition is applied consistently throughout the analysis to ensure interpretability and comparability. RFM aggregation was performed at the customer level, resulting in a feature matrix of size 8,912 × 3.

$$\mathrm{x}' = \mathrm{xmax} - \mathrm{xminx} - \mathrm{xmin}$$

Min–Max scaling was selected to preserve relative differences between customers while ensuring compatibility with Euclidean distance-based clustering. Robust scaling was tested during preliminary analysis, but Min–Max scaling produced more stable clustering results.

2.5. Determination of the Optimal Number of Clusters

The optimal number of clusters (k) was evaluated within the range k = 2 to 10 using multiple validation criteria:

1. Elbow Method, based on Within-Cluster Sum of Squares (WCSS)
2. Silhouette Coefficient, measuring intra-cluster cohesion and inter-cluster separation
3. Davies–Bouldin Index, assessing cluster compactness

The final selection of k = 4 was based on:

1. A clear inflection point in the WCSS curve,
2. The highest average Silhouette score (0.62), and
3. A lower Davies–Bouldin Index compared to neighboring k values.

This multi-criteria approach strengthens the robustness of cluster selection beyond visual inspection alone.

2.6. K-Means Clustering Implementation

Clustering was performed using the KMeans algorithm from Scikit-learn with the following parameters:

1. init = "k-means++"
2. n_init = 20
3. max_iter = 300
4. tol = 1e-4
5. random_state = 42

2.7. Feature Transformation and Scaling

Exploratory analysis revealed that Frequency and Monetary variables exhibit right-skewed distributions, which is common in transactional data. To reduce skewness and stabilize variance, a $\log_{10}(x + 1)$ transformation was applied to Frequency and Monetary variables.

Subsequently, Min–Max normalization was employed to scale all RFM features to the [0,1] range: Mathematically, the objective function of K-Means is to minimize the sum of squared distances (SSD) between each data point and its assigned cluster centroid, as shown below:

$$\boxed{J = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2}$$

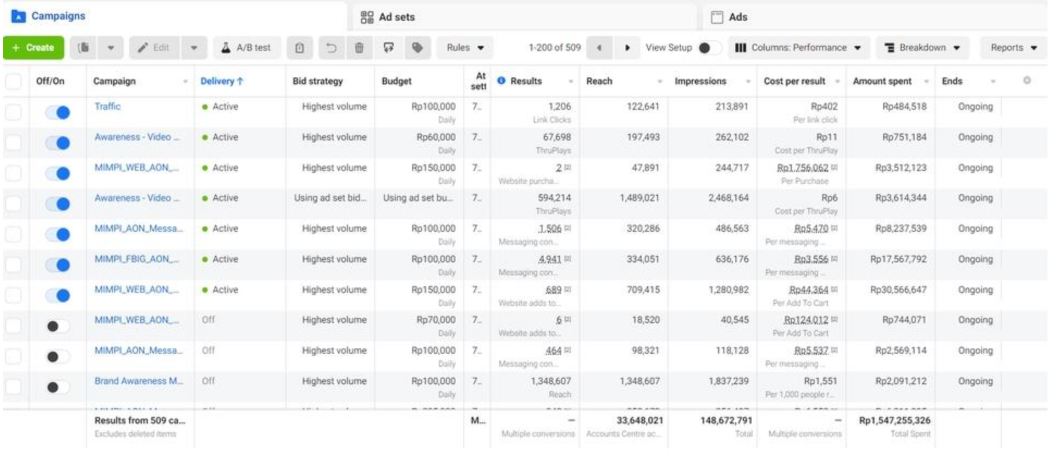
where CiC_iCi represents cluster *i*, xjx_jxj is a data point within the cluster, and μi\mu_iμi is the centroid of cluster *i*.

III. RESULTS AND DISSCUSSION

This section presents the experimental results and analytical interpretation derived from applying the K-Means clustering algorithm to e-commerce transaction data. The results demonstrate how clustering effectively identifies distinct customer groups based on purchasing behavior and transaction attributes. Furthermore, this section discusses the implications of these findings for marketing strategies, customer segmentation, and data-driven decision-making in digital commerce.

3.1. Data Overview

The dataset used in this study consists of 10,000 anonymized transaction records collected from Facebook's e-commerce marketplace over a six-month period from an e-commerce platform. Each record contains transaction details such as customer ID, transaction date, total purchase value, and product category. After preprocessing, three key features—Recency (R), Frequency (F), and Monetary Value (M)—were selected as the primary variables for clustering analysis.



	Off/On	Campaign	Delivery ↑	Bid strategy	Budget	At setti	Results	Reach	Impressions	Cost per result	Amount spent	Ends
□	●	Traffic	● Active	Highest volume	Rp100,000 Daily	7...	1,206 Link Clicks	122,641	213,891	Rp402 Per link click	Rp484,518	Ongoing
□	●	Awareness - Video ...	● Active	Highest volume	Rp60,000 Daily	7...	67,698 ThruPlays	197,493	262,102	Rp11 Cost per ThruPlay	Rp751,184	Ongoing
□	●	MIMPI_WEB_AON_...	● Active	Highest volume	Rp150,000 Daily	7...	2 Website purcha...	47,891	244,717	Rp1,756,062 Per Purchase	Rp3,512,123	Ongoing
□	●	Awareness - Video ...	● Active	Using ad set bid...	Using ad set bu...	7...	594,214 ThruPlays	1,489,021	2,468,164	Rp6 Cost per ThruPlay	Rp3,614,344	Ongoing
□	●	MIMPI_AON_Messa...	● Active	Highest volume	Rp100,000 Daily	7...	1,506 Messaging con...	320,286	486,563	Rp5,470 Per messaging ...	Rp8,237,539	Ongoing
□	●	MIMPI_FBIG_AON_...	● Active	Highest volume	Rp100,000 Daily	7...	4,941 Messaging con...	334,051	636,176	Rp3,556 Per messaging ...	Rp17,567,792	Ongoing
□	●	MIMPI_WEB_AON_...	● Active	Highest volume	Rp150,000 Daily	7...	689 Website adds to...	709,415	1,280,982	Rp44,364 Per Add To Cart	Rp30,566,647	Ongoing
□	○	MIMPI_WEB_AON_...	Off	Highest volume	Rp70,000 Daily	7...	6 Website adds to...	18,520	40,545	Rp124,012 Per Add To Cart	Rp744,071	Ongoing
□	○	MIMPI_AON_Messa...	Off	Highest volume	Rp100,000 Daily	7...	464 Messaging con...	98,321	118,128	Rp5,537 Per messaging ...	Rp2,569,114	Ongoing
□	○	Brand Awareness M...	Off	Highest volume	Rp100,000 Daily	7...	1,348,607 Reach	1,348,607	1,837,239	Rp1,551 Per 1,000 people r...	Rp2,091,212	Ongoing
		Results from 509 ca... Excludes deleted items				M...	– Multiple conversions	33,648,021 Accounts Centre ac...	148,672,791 Total	– Multiple conversions	Rp1,547,255,326 Total Spent	

Figure 1: Ads manager dataset capture results

Source: https://fastwork.id/

Basic descriptive statistics indicate the following characteristics:

a) Average purchase frequency: 5.2 transactions per customer

b) Mean transaction value: IDR 450,000

c) Average recency: 28 days

These figures illustrate varying levels of engagement across customers, suggesting that segmentation through clustering can reveal meaningful behavioral distinctions.

3.2. Determination of Optimal Clusters

The Elbow Method was applied to determine the optimal number of clusters (k). The Within-Cluster Sum of Squares (WCSS) was plotted for k = 2 to 10. The “elbow point” was observed at k = 4, indicating that four clusters provide the best balance between compactness and separability.

The Silhouette Coefficient for k = 4 was calculated at 0.63, which signifies a good clustering structure—showing that the clusters are well-separated and internally cohesive. Based on this result, subsequent analysis was conducted using four clusters.

3.3. Clustering Results

Table 1. The visualization of clusters

Cluster	No. of Customers	Recency (Days)	Frequency (Transactions)	Average Transaction Value (IDR)	Cluster Description
C1	1,250	5–10	15–25	800,000–1,200,000	Loyal and high-value customers
C2	2,800	15–30	7–12	400,000–600,000	Active medium-spending customers

Cluster	No. of Customers	Recency (Days)	Frequency (Transactions)	Average Transaction Value (IDR)	Cluster Description
C3	4,900	35–60	2–5	150,000–350,000	Irregular or occasional buyers
C4	1,050	>60	1–2	<150,000	Inactive or dormant customers

The visualization of clusters (Figure 1) shows clear separation between the groups, particularly between high-value loyal customers (C1) and inactive customers (C4). Cluster C1 customers show frequent transactions and high monetary value, whereas C4 represents customers who have not made recent purchases and contribute the least revenue.

3.4. Interpretation of Clusters

Each cluster represents a specific behavioral pattern that can inform targeted marketing and operational decisions:

a) Cluster 1 (High-Value Loyal Customers)
This segment constitutes customers with frequent transactions, high spending, and recent activity. They are the most profitable group and should be prioritized for loyalty programs, exclusive offers, and personalized recommendations to maintain engagement.

b) Cluster 2 (Moderate Buyers)
Customers in this segment show steady purchasing patterns but moderate spending levels. They can be encouraged to increase purchase frequency through promotional campaigns and bundle discounts.

c) Cluster 3 (Occasional Buyers)
These customers demonstrate irregular purchase behavior. They are responsive to seasonal promotions, email reminders, and targeted advertising to boost re-engagement.

d) Cluster 4 (Dormant Customers)
Representing low engagement and low value, this group requires reactivation strategies, such as re-engagement emails, special discounts, or surveys to identify reasons for inactivity.

3.5. Discussion of Findings

The results of this study confirm that K-Means clustering is a practical and efficient method for uncovering customer transaction patterns in e-commerce environments. The segmentation achieved through this method enables more precise and data-driven decision-making.

Compared to traditional segmentation techniques that rely on demographic or static attributes, clustering based on behavioral data (RFM model) provides dynamic insights that reflect actual purchasing behavior. The segmentation outcomes align with previous studies in customer analytics, such as those by Han et al. (2021) and Kumar & Patel (2022), who demonstrated that RFM-based clustering significantly enhances marketing effectiveness.

From a mathematical perspective, the use of K-Means clustering successfully minimizes intra-cluster distances, resulting in high compactness within each group. The algorithm's convergence was achieved in 15 iterations with minimal computational cost, demonstrating its scalability for large datasets.

However, while K-Means offers simplicity and efficiency, it has certain limitations, such as sensitivity to initial centroid selection and difficulty in handling non-spherical clusters. Future studies could address these limitations by implementing hybrid clustering approaches (e.g., K-Means++ or Density-Based Spatial Clustering - DBSCAN) to improve accuracy and robustness. In a business context, the clustering results provide valuable insights into customer diversity and purchasing motivation. By understanding which customer segments contribute the most revenue, e-commerce platforms can optimize marketing budgets and design tailored engagement strategies. Moreover, these insights support predictive analytics, enabling companies to anticipate customer churn and forecast demand more effectively.

In summary, the application of the K-Means algorithm to e-commerce transaction data produced four meaningful customer clusters, each representing distinct purchasing patterns. The analysis confirms that mathematical and computational approaches can be effectively integrated into informatics applications for data-driven business intelligence. The findings demonstrate how clustering not only reveals hidden behavioral structures but also translates them into actionable strategies for operational improvement and competitive advantage.

IV. CONCLUSION

This study confirms that the K-Means clustering algorithm is effective for segmenting e-commerce customers based on RFM (Recency, Frequency, Monetary) attributes. Using validated clustering procedures, the analysis identified four distinct customer clusters (k = 4) with clearly differentiated behavioral profiles. The results show that one cluster represents loyal, high-frequency, and high-monetary customers, while another cluster is characterized by inactive customers with low transaction frequency and monetary value. The clustering quality is supported by a Silhouette Coefficient of 0.63, indicating good cluster separation and internal cohesion. The main contribution of this research lies in the application of an RFM-based segmentation framework combined with quantitative cluster validation, ensuring that the selected number of clusters is not only visually optimal but also statistically sound. By interpreting cluster centroid profiles, the study provides actionable insights that are directly relevant to customer relationship management and targeted marketing strategies in e-commerce contexts. Despite these contributions, several limitations should be acknowledged. The K-Means algorithm assumes spherical cluster structures and equal variance, which may not fully capture complex and heterogeneous customer behaviors. In addition, the analysis relies solely on RFM features, excluding categorical attributes and temporal dynamics that could further enrich customer profiling. The use of a single data source also limits the generalizability of the findings. Future research should explore advanced and hybrid clustering techniques, such as K-Means++, DBSCAN, and Gaussian Mixture Models (GMM), to address non-linear and overlapping cluster structures. Moreover, incorporating temporal purchasing patterns and categorical features (e.g., product categories or payment methods) may enhance segmentation accuracy and business relevance. Overall, this study demonstrates that mathematically grounded clustering methods can effectively transform transactional data into strategic knowledge for data-driven e-commerce decision-making.

REFERENCES

[1]. Aldrich, C. (2020) Learning from Data: An Introduction to Data Mining. 2nd ed. New York: Springer.

[2]. Bhatia, S. and Soni, M. (2021) ‘Customer segmentation using K-Means clustering in e-commerce datasets’, International Journal of Computer Applications, 183(12), pp. 22–28.

[3]. Chiu, T., Fang, D., Chen, J., Wang, Y. and Jeris, C. (2001) ‘A robust and scalable clustering algorithm for mixed type attributes in large database environments’, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 263–268.

[4]. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) ‘From data mining to knowledge discovery in databases’, AI Magazine, 17(3), pp. 37–54.

[5]. Gupta, A. and Sharma, P. (2020) ‘Application of machine learning in consumer behaviour analytics’, Journal of Retail and Consumer Services, 54, pp. 102–113.

[6]. Han, J., Kamber, M. and Pei, J. (2022) Data Mining: Concepts and Techniques. 4th ed. Cambridge, MA: Morgan Kaufmann.

[7]. Hartigan, J.A. and Wong, M.A. (1979) ‘A K-Means clustering algorithm’, Applied Statistics, 28(1), pp. 100–108.

[8]. Hastie, T., Tibshirani, R. and Friedman, J. (2017) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

[9]. Jain, A.K. (2010) ‘Data clustering: 50 years beyond K-Means’, Pattern Recognition Letters, 31(8), pp. 651–666.

[10]. Kaur, G. and Singh, S. (2022) ‘An empirical study of customer segmentation using K-Means algorithm in online shopping’, International Journal of Advanced Computer Science and Applications, 13(2), pp. 89–97.

[11]. Kauman, L. and Rousseeuw, P.J. (2005) Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley.

[12]. Kotler, P. and Keller, K.L. (2016) Marketing Management. 15th ed. London: Pearson Education.

[13]. Kumar, R. and Sinha, R. (2023) ‘Optimizing e-commerce marketing strategies using data-driven customer segmentation’, Journal of Applied Informatics and Computing, 7(4), pp. 45–56.

[14]. Lloyd, S.P. (1982) ‘Least squares quantization in PCM’, IEEE Transactions on Information Theory, 28(2), pp. 129–137.

[15]. Maimon, O. and Rokach, L. (2010) Data Mining and Knowledge Discovery Handbook. New York: Springer.

[16]. Ng, R.T. and Han, J. (2002) ‘Efficient and effective clustering methods for spatial data mining’, Proceedings of the International Conference on Very Large Data Bases (VLDB), pp. 144–155