Optimization of the K-Means Algorithm Using PCA Dimensionality Reduction for E-Commerce Customer Segmentation

Mahara Bengi[1,*], Syarifah Atika[2], Chichi Rizka Gunawan[3], Cicha Rizka Gunawan[4]
[1,3,4] Fakultas Sains dan Teknologi, Informatika, Universitas Samudra, Langsa, Indonesia
[2] Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Prima Indonesia, Medan, Indonesia
maharabengi@unsam.ac.id

Submitted: December 8, 2025; accepted: May 5, 2026

Abstract

The rapid growth of the e-commerce industry has generated increasingly large and complex customer datasets, creating opportunities for data-driven behavioral analysis. This study investigates customer segmentation using an unsupervised learning approach that integrates Principal Component Analysis (PCA) with the K-Means clustering algorithm. The dataset, obtained from a publicly available e-commerce customer dataset, consists of 350 records with multidimensional behavioral attributes. Exploratory analysis indicates that Total Spending, Number of Items Purchased, and Average Rating are dominant variables representing purchasing intensity. PCA reduced the dimensionality of the dataset while retaining 79.41% of the total variance, thereby improving structural efficiency without significant information loss. Clustering results reveal three well-defined customer segments representing high, moderate, and low engagement levels. Although the highest Silhouette Score was observed at k = 8, the improvement compared to k = 3 was marginal, and higher cluster configurations resulted in fragmented and less interpretable groups. Therefore, k = 3 was selected based on statistical stability and interpretability considerations. The findings demonstrate that PCA-enhanced K-Means clustering improves cluster clarity and structural coherence, providing an effective methodological framework for customer segmentation analysis. This study contributes to clustering optimization research by empirically evaluating the balance between dimensionality reduction, separation quality, and interpretability in e-commerce customer data.

Keywords*: E-commerce, PCA, K-Means, Customer Segmentation, Behavioral Analysis.*

1. Introduction

The rapid growth of the e-commerce industry has generated extremely large and complex volumes of customer data. Each customer exhibits diverse characteristics, ranging from age, gender, and satisfaction levels to purchasing patterns and total spending. This abundance of data holds substantial potential for understanding consumer behavior and designing more targeted marketing strategies. However, to fully leverage this potential, companies must classify customers into relevant groups through effective customer segmentation techniques. Customer segmentation itself is an essential marketing strategy that organizes customers based on specific characteristics, behaviors, or preferences. By understanding these distinct customer groups, companies can enhance retention, increase loyalty, and obtain long-term benefits through marketing initiatives that better align with customer needs [1], [2].

Clustering methods represent one approach capable of uncovering purchasing behavior patterns, enabling businesses to tailor marketing strategies such as loyalty programs or more specialized promotional activities [3], [4]. One of the most widely used techniques in customer segmentation is K-Means clustering, as it effectively forms groups based on similarities between individuals. However, K-Means has several inherent limitations. The algorithm is sensitive to differences in feature scales, vulnerable to multicollinearity, and tends to degrade in performance when applied to high-dimensional data [5], [6]. These issues may lead to suboptimal cluster results and visual interpretability challenges [7].

To address these limitations, Principal Component Analysis (PCA) is applied as a preprocessing stage prior to clustering. PCA maps high-dimensional data into a lower-dimensional space efficiently, thereby reducing computational complexity without eliminating essential information [8]. This method

decreases the number of features while retaining the majority of data variance, which can often reach approximately 80% using only a few principal components [9]. Previous empirical findings indicate that integrating PCA with K-Means yields more distinct and stable cluster structures, reflected in a Silhouette Score of 0.49—signifying improved cluster separation quality [9]. Moreover, PCA-based centroid initialization has been shown to provide more accurate cluster center representations, reducing the likelihood of suboptimal clustering configurations [8]. The combination of PCA and K-Means is expected to produce clusters that are more distinct, compact, and easier to interpret. This approach has also been demonstrated to improve the clarity of clustering outcomes in various studies, including research on human resources data showing that cluster structures become more prominent after PCA is applied prior to K-Means [10].

While numerous studies have explored customer segmentation using K-Means, many rely either on purely transactional variables or on benchmark datasets without explicitly analyzing the impact of dimensionality reduction on cluster interpretability and stability. In this context, the dataset used in this study sourced from a publicly available e-commerce customer dataset provides a structured multidimensional representation of behavioral attributes, including spending intensity, purchase frequency, recency, discount usage, satisfaction level, and membership category. The scientific value of this dataset lies in its comprehensive behavioral structure, which enables a controlled evaluation of how variance-based feature compression influences clustering compactness, separability, and interpretability.

Therefore, rather than focusing solely on practical implementation, this research contributes methodologically by empirically examining the effectiveness of PCA-enhanced clustering within a multidimensional customer behavior framework. The integration of PCA and K-Means is expected to produce clusters that are more distinct, compact, and interpretable, thereby strengthening both analytical clarity and strategic relevance.

Accordingly, the objective of this research is to develop a customer segmentation model using PCA and K-Means and to evaluate the results using the Elbow Method to determine the optimal number of clusters, as well as the Silhouette Score to assess separation quality. By emphasizing methodological evaluation and reproducibility, this study aims to provide transparent and empirically grounded insights into clustering optimization for e-commerce customer segmentation.

2. Method

The research methodology in this study is visually illustrated through a flowchart that outlines all major stages, beginning from data collection and preprocessing to model evaluation, as presented in Figure 1.

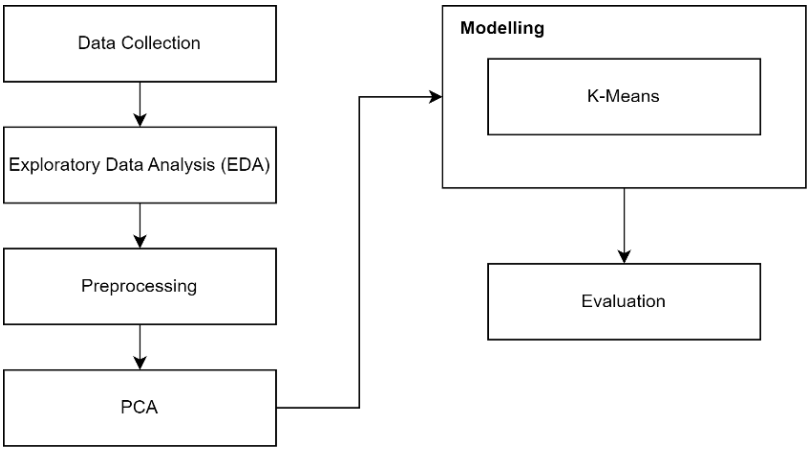

Fig. 1 Workflow of the Research Methodology

2.1. Data Collection

The dataset used in this study was obtained from a publicly available e-commerce customer dataset distributed through the Kaggle data science platform [11]. It consists of 350 customer records with 11 behavioral features, including user interaction indicators, purchase history, discount usage, and satisfaction levels. The dataset is publicly accessible and not affiliated with a specific commercial entity.

All records are anonymized and structured for analytical purposes. Additionally, the customer records are proportionally distributed across three membership categories: Gold, Silver, and Bronze.

2.2. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) stage was conducted to understand the fundamental structure and characteristics of the dataset prior to preprocessing and modeling. The analysis began by examining the distribution of numerical features through histogram visualizations to identify distribution patterns, value ranges, and potential outliers. This step provided an initial overview of variations and general tendencies in customer behavior.

Subsequently, relationships among numerical variables were evaluated using a correlation matrix, which was visualized through a heatmap. This step aimed to identify variables with strong or weak correlations, serving as an important consideration for feature selection and dimensionality reduction. Overall, EDA functions as a critical initial step for comprehensively understanding the dataset before applying further transformations such as PCA and clustering using K-Means.

2.3. Preprocessing

The preprocessing stage began with data cleaning to ensure dataset quality prior to analysis. Missing values in numerical features such as Age, Total Spending, Items Purchased, Average Rating, and Days Since Last Purchase were handled using median imputation, while categorical features such as Gender, City, Membership Type, Discount Usage, and Satisfaction Level were filled using the mode to represent the most common category. A duplication check was also conducted, and no duplicate entries were found, ensuring the uniqueness of all observations.

Next, categorical variables were converted into numerical form using label encoding to enable model processing. The numerical features were then normalized using StandardScaler to standardize feature scales and improve the stability of the algorithm's computations. The outcome of this entire preprocessing stage was a clean, consistent, and normalized dataset ready for dimensionality reduction using PCA and clustering using K-Means.

2.4. Principal Component Analysis (PCA)

Dimensionality reduction using Principal Component Analysis (PCA) was conducted to simplify the number of features without eliminating essential information within the dataset. PCA computes principal components based on the highest variance of the normalized data, thereby capturing the structural patterns of the dataset more efficiently. The process begins with determining the explained variance ratio of each principal component to assess the proportion of information represented by each component. Based on these calculations, a set of components that retain the majority of the data variance is selected, reducing input complexity while preserving representational quality. The reduced-dimension dataset is then used as input for the clustering stage.

2.5. Modelling

The modeling stage employed the K-Means algorithm to generate customer segmentation based on the PCA-reduced dataset. Several values of k (number of clusters) were tested to observe the pattern of cluster formation and determine the most optimal configuration. The K-Means model was executed for k = 2 to k = 9 using random_state = 42 to ensure consistent centroid initialization across training runs. For each value of k, the algorithm performed its core operations, including centroid initialization, Euclidean distance calculation, assignment of data points to the nearest centroid, and centroid updates.

Upon fitting the model, the inertia value was recorded for each k. Inertia represents the within-cluster sum of squares (WSS), which serves as an indicator of cluster compactness. Additionally, the cluster labels produced for each k were used to compute the Silhouette Score, which reflects how well each data point fits within its assigned cluster. All results were then stored for subsequent evaluation.

2.6. Evaluation

Model evaluation was carried out using two primary approaches: the Elbow Method and the Silhouette Score, both of which were analyzed using the metrics generated during modeling. In the Elbow Method, inertia values were plotted against variations in the number of clusters to assess the compactness of each cluster configuration. The optimal number of clusters was identified at the "elbow point," where the decrease in inertia begins to slow, indicating that adding more clusters does not significantly enhance model quality.

Next, the Silhouette Score was used to evaluate the level of separation between clusters. Higher scores indicate well-defined clusters with strong internal cohesion. By jointly examining both evaluation plots, the optimal cluster number can be determined more objectively. The results of this evaluation serve as the basis for selecting the final K-Means model used in the subsequent customer segmentation analysis.

3. Result and Discussion

3.1. Dataset Analysis and EDA

The dataset used in this study consists of 350 observations with 11 variables representing e-commerce customer behavior. The exploratory stage was carried out to ensure the quality and completeness of the dataset prior to modeling. Data integrity checks confirmed that the dataset contains no missing values or duplicate entries, allowing it to be used directly without requiring additional imputation processes.

The distribution of all numerical features including Age, Total Spending, Items Purchased, Average Rating, Discount Usage, and Days Since Last Purchase shows variations within reasonable boundaries, reflecting the diversity of customer shopping behavior. The histograms in Figure 2 illustrate the distribution patterns of each variable, including dominant tendencies within moderate value ranges and the presence of several observations with relatively high values. These distribution patterns provide an initial understanding of the fundamental characteristics of the numerical data and serve as an important basis for ensuring dataset suitability prior to further analyses such as dimensionality reduction and clustering.

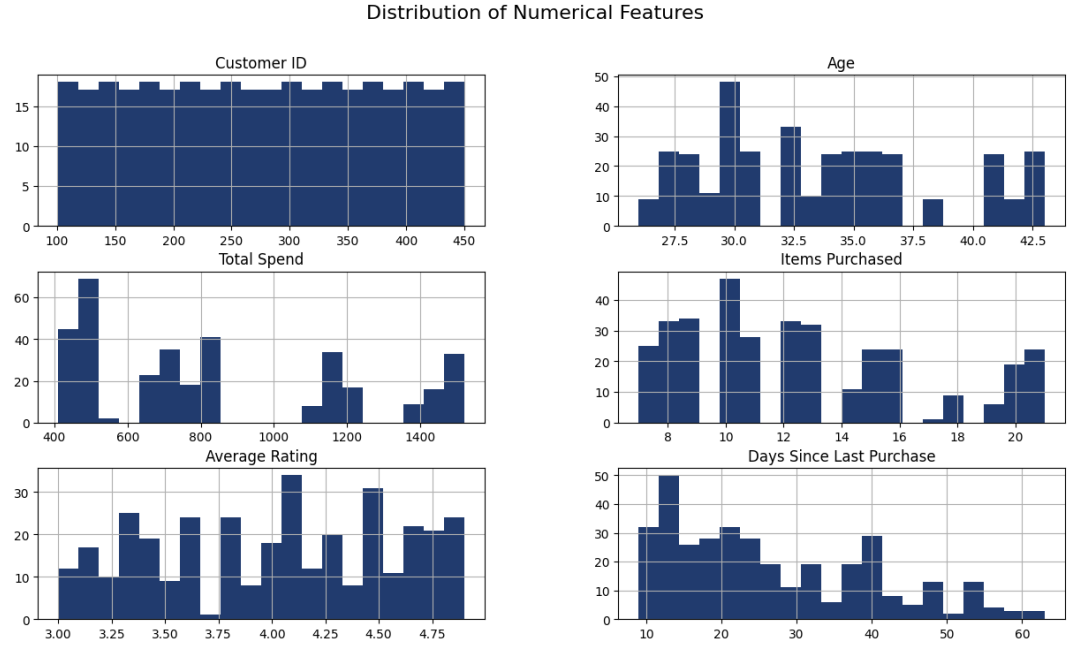

Fig. 2 Histogram of Numerical Feature Distributions

Correlation analysis among numerical variables was conducted using a correlation matrix, as shown in Figure 3. The visualization reveals strong correlations among Total Spending, Items Purchased, and Average Rating, indicating that these three variables play a major role in describing customer purchasing intensity and quality. Conversely, variables such as Customer ID and Age show weaker relationships with other variables, suggesting minimal contribution to the modeling process.

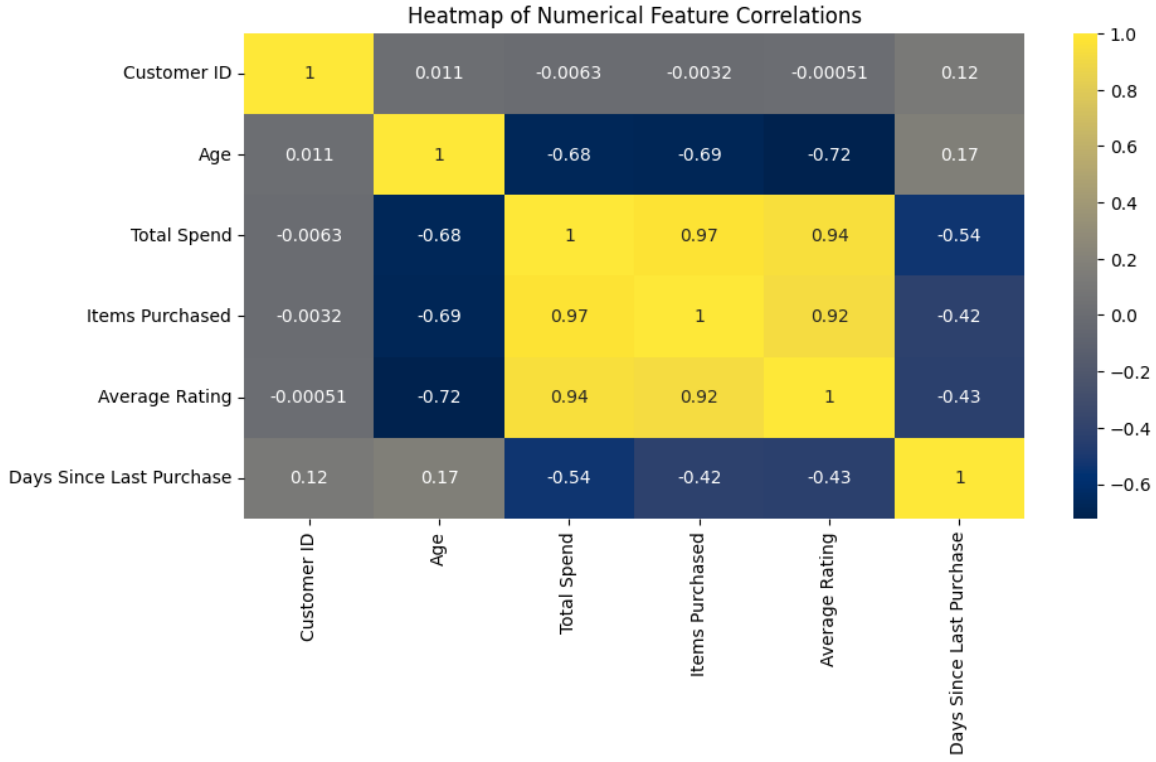

Fig. 3 Correlation Heatmap of Numerical Features

The insights obtained from the exploratory analysis clarify variable patterns and relationships, forming a foundation for identifying the most informative features for PCA and for constructing an appropriate clustering approach using K-Means.

3.2. Effectiveness of PCA in Dimensionality Reduction

The dimensionality reduction process using Principal Component Analysis (PCA) aims to reduce feature complexity without losing important information. The PCA results show that two principal components (PC1 and PC2) are able to explain 79.41% of the total variance, with PC1 contributing 61.87% and PC2 contributing 17.54%. This information is summarized in Table 1.

The first principal component (PC1) demonstrates the highest contribution from Total Spending, Items Purchased, and Average Rating, suggesting that PC1 represents the intensity and quality of customer purchasing behavior. Meanwhile, the second principal component (PC2) is strongly associated with Age and Days Since Last Purchase, representing behavioral tendencies influenced by demographic factors and transaction frequency.

Table 1.Explained Variance Ratio of PCA Components

Principal Component	Variance Ratio	Cumulative Variance
PC1	0.6187	0.6187
PC2	0.1754	0.7941

The distribution of data in the two-dimensional PCA space is shown in Figure 4, which displays natural separation patterns across customer groups. This separation provides strong evidence that the data possesses a cluster structure suitable for further exploration using the K-Means algorithm.

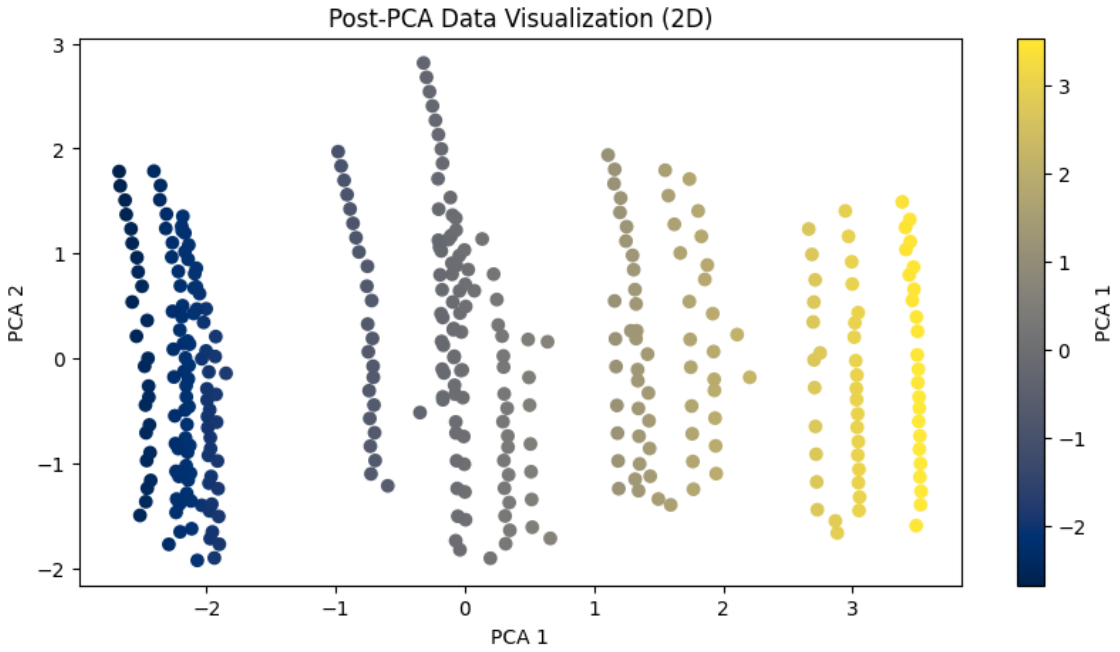

Fig. 4 Visualization of PCA-Transformed Data

In addition to reducing dimensionality and improving computational efficiency, PCA also helps minimize issues related to multicollinearity. The transformed dataset provides a cleaner and more manageable representation for the clustering stage.

3.3. Customer Cluster Formation Using K-Means

Applying the K-Means algorithm to the PCA-transformed data produced three main clusters that are stable and easily interpretable. The selection of the number of clusters was based on experiments with multiple k values, considering cluster separability, interpretability, and quantitative evaluations from the Elbow Method and Silhouette Score.

The distribution of cluster members is presented in Table 1, where clusters 0 and 1 contain the largest number of customers, while cluster 2 is smaller yet provides important insights into customers with distinct purchasing behaviors.

Table 2. Distribution of Cluster Members

Cluster	Number of Members
0	116
1	105
2	129

The average values of each feature across clusters, as presented in Table 2, indicate distinct characteristics:

- Cluster 1: the most active group, showing the highest Total Spending, Items Purchased, and Average Rating. Their short purchase intervals suggest highly loyal and high-intensity customers.
- Cluster 0: a moderate group with balanced spending and activity levels.
- Cluster 2: the least active group with the lowest purchasing intensity and the longest purchase intervals, representing customers who rarely transact or show low engagement.

Table 3. Average Numerical Feature Values per Cluster

Cluster	Number of Data Points	Average Age	Average Total Spending	Average Items Purchased	Average Rating	Days Since Last Purchase
0	116	39.36	473.39	8.49	3.33	31.61
1	105	29.77	1329.22	17.94	4.70	16.62
2	129	31.53	786.07	11.95	4.09	30.19

These findings confirm that the K-Means algorithm effectively identifies customer segments with significantly different behavioral patterns that are highly relevant for business use.

3.4. Model Evaluation

Model effectiveness was evaluated using the Elbow Method and Silhouette Score, as shown in Figure 5.

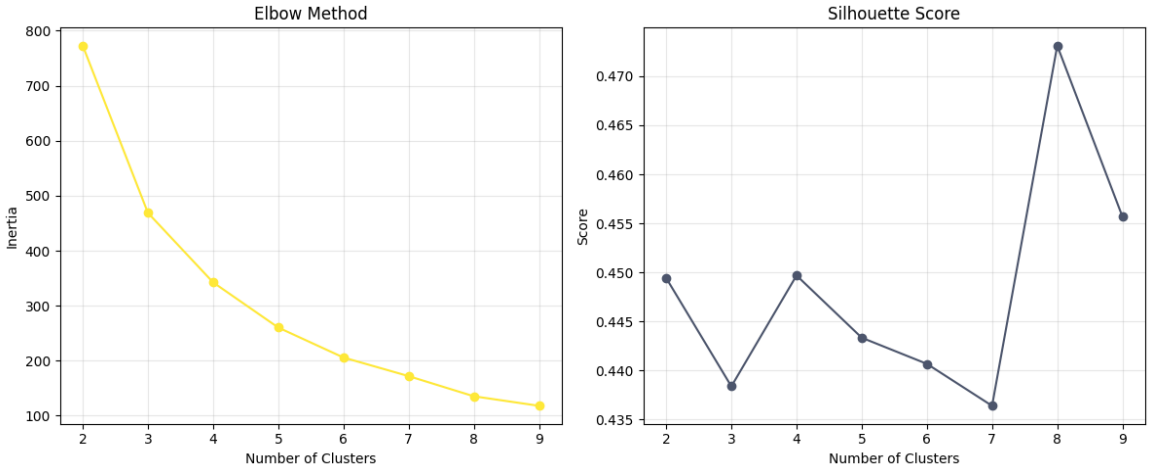

Fig. 5 Model Evaluation

The Elbow plot demonstrates a pronounced decrease in inertia between k = 2 and k = 4, indicating substantial improvement in within-cluster compactness within this interval. Beyond k = 4, the rate of inertia reduction diminishes considerably, suggesting that additional clusters provide limited incremental structural benefit. This pattern indicates the presence of an elbow region around k = 3–4.

The Silhouette Score analysis for k = 2 to k = 9 shows that the highest value is obtained at k = 8 (approximately 0.472), while the score at k = 3 is approximately 0.449. Although k = 8 yields the maximum numerical value, the improvement relative to k = 3 is marginal ($\Delta \approx 0.023$), indicating that the increase in separation quality is not substantial.

A closer structural examination reveals that the configuration at k = 8 produces greater fragmentation, with several clusters containing relatively small subsets of observations. This suggests over-segmentation, where minor variations in behavioral attributes are separated into additional clusters without generating meaningful differentiation. While such granularity may slightly improve the Silhouette metric, it reduces structural coherence and practical interpretability.

In contrast, the k = 3 configuration results in balanced cluster distributions and clear differentiation across key behavioral dimensions, including spending intensity, transaction frequency, and recency. The clusters remain sufficiently separated in the PCA-transformed space while maintaining interpretative clarity.

Therefore, considering statistical evidence, structural stability, and interpretability, k = 3 is selected as the optimal number of clusters.

3.5. Business Implications of Customer Segmentation

The segmentation results obtained from PCA and K-Means provide strategic insights that can support business decision-making. Based on the characteristics of each cluster:

- High-value cluster (Cluster 1**):** suitable targets for loyalty programs, exclusive recommendations, and premium benefits to enhance retention.
- Moderate cluster (Cluster 0**):** candidates for personalized promotions, discount coupons, or upselling campaigns to increase purchasing activity.
- Low-engagement cluster (Cluster 2): ideal for reactivation strategies such as reminder emails, bundling offers, or retention campaigns to reduce churn rates.

Thus, customer segmentation based on data analysis enables e-commerce companies to design more effective, targeted, and evidence-based marketing strategies.

4. Conclusion

The analysis conducted using Principal Component Analysis (PCA) and the K-Means algorithm demonstrates that the e-commerce customer dataset is of high quality and free from missing values, making it suitable for advanced analytical processes. The exploratory analysis reveals that Total Spending, Items Purchased, and Average Rating are the most dominant variables in representing overall customer behavior.

The application of PCA plays a crucial role in simplifying data dimensionality, retaining 79.41% of the total variance, which indicates that most relevant information from the original dataset is preserved. This contributes to producing models that are more stable, efficient, and interpretable. Furthermore, the K-Means algorithm successfully groups customers into three well-defined clusters: high activity, moderate activity, and low activity. Validation using the Elbow Method and Silhouette Score confirms that k = 3 is the optimal number of clusters, as it produces good separation and sufficient internal coherence. This segmentation offers strategic value for businesses, particularly in developing loyalty programs, designing more personalized promotions, and planning targeted reactivation strategies.

For future work, it is recommended to incorporate more diverse behavioral features and compare results with alternative clustering methods such as DBSCAN or GMM. Periodic cluster updates are also necessary to ensure alignment with evolving customer behavior.

Reference

[1] Purushottam Perapu, "Customer Segmentation Using K-Means Clustering for Personalized Marketing Campaigns," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 3, pp. 810–815, May 2025, doi: 10.32628/CSEIT25113344.

[2] I. Shah, "Customer Segmentation," *Int J Res Appl Sci Eng Technol*, vol. 12, no. 1, pp. 1586–1591, Jan. 2024, doi: 10.22214/ijraset.2024.58144.

[3] Refri Martiansah, Siti Monalisa, Fitriani Muttakin, and Mona Fronita, "Customer Segmentation Analysis Through RFM-D Model and K-Means Algorithm," *Jurnal Sistem Cerdas*, vol. 8, no. 1, pp. 1–11, Apr. 2025, doi: 10.37396/jsc.v8i1.504.

[4] V. V. Darma Oktavian, R. Ramadhan, and D. R. Fadhilla, "Segmentasi Pelanggan Berbasis RFM dengan Algoritma K-Means pada Data Transaksi Online Retail," *Jurnal Riset Informatika dan Teknologi Informasi*, vol. 2, no. 3, pp. 236–243, Jul. 2025, doi: 10.58776/jriti.v2i3.156.

[5] T. Garg and A. Malik, "Survey on Various Enhanced K-Means Algorithms," 2014. [Online]. Available: www.ijarcce.com

[6] B. Chong, "K-means clustering algorithm: a brief review," *Academic Journal of Computing & Information Science*, vol. 4, no. 5, 2021, doi: 10.25236/AJCIS.2021.040506.

[7] A. Chadha and S. Kumar, "An improved K-Means clustering algorithm: A step forward for removal of dependency on K," in *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*, IEEE, Feb. 2014, pp. 136–140. doi: 10.1109/ICROIT.2014.6798312.

[8] C. Zhang, J. Ou, W. He, H. Huang, G. Cheng, and Y. Gu, "Optimisation Research on K-Means Clustering Algorithm Based on Principal Component Analysis and Percentile Improvement," in

2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA), IEEE, Nov. 2024, pp. 148–153. doi: 10.1109/ICAICA63239.2024.10823007.

[9] A. Jauhari, I. O. Suzanti, D. R. Anamisa, and F. T. Admojo, "PCA-counseled k-means and k-medoids with dimension reduction for improved in determining optimal aid clustering," *Jurnal Ilmiah Kursor*, vol. 13, no. 1, pp. 46–55, Jul. 2025, doi: 10.21107/kursor.v13i1.460.

[10] S. A. Mousavian Anaraki, A. Haeri, and F. Moslehi, "A hybrid reciprocal model of PCA and K-means with an innovative approach of considering sub-datasets for the improvement of K-means initialization and step-by-step labeling to create clusters with high interpretability," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 1387–1402, Aug. 2021, doi: 10.1007/s10044-021-00977-x.

[11] Y. Abdelghfar, "E-commerce Customer Behavior," Kaggle Notebook, 2024. [Online]. Available: https://www.kaggle.com/code/youssefabdelghfar/e-commerce-customer-behavior