Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

E-Commerce Customer Segmentation Application Based on the K-Means Algorithm

Nehemia Dheadema Mareten[1], Jekoniah Nahum Pakage[2]*, Veronica Lois[3], Regina Arieskha[4], Muhammad Ifan Rifani Ihsan[5]

[1,2,3,4] Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
nehemiadm17@gmail.com[1], 15230325@bsi.ac.id[2], veronicalouise08@gmail.com[3], reginaarieskha01@gmail.com[4]*

Abstract

Ineffective e-commerce marketing serves as the background for this research, which aims to develop a customer segmentation application for targeted marketing. The K-Means Clustering method with RFM (Recency, Frequency, Monetary) analysis is applied to data from 178 customers. The research methodology includes data preprocessing, feature transformation, and the determination of the optimal K using the Elbow Method. The results indicate that K=3 is the optimal number of clusters. Three segments were successfully identified: 'Champions' (18.5%, 33 customers) with the highest Frequency/Monetary values, 'Active & Potential' (41%, 73 customers) with the lowest Recency (most recent), and 'At Risk' (40.5%, 72 customers) with the highest Recency (longest duration since last transaction). The study concludes that the developed Streamlit-based application successfully visualizes these segments interactively to support strategic decision-making in marketing.

Keywords: *Customer Segmentation; E-commerce; K-Means Algorithm; RFM Analysis; Streamlit*

1. Introduction

The rapid advancement of digital technology has significantly transformed the global commerce landscape, particularly with the proliferation of e-commerce platforms in Indonesia [4], [11]. These platforms not only offer convenience to consumers but also provide businesses with vast opportunities to reach broader markets [13]. Consequently, this shift influences consumer behavior, necessitating companies to adopt more flexible and adaptive business strategies to remain competitive [6].

In the current competitive landscape, understanding consumer behavior is paramount. E-commerce stakeholders face the substantial challenge of managing vast amounts of customer data characterized by diverse demographics and purchasing behaviors. Traditional mass marketing strategies have proven inefficient, as they fail to address the specific needs and preferences of distinct consumer groups [1].

To address these challenges, data-driven customer segmentation offers an effective solution. By categorizing customers into smaller, homogeneous groups based on their shopping habits, companies can tailor promotional strategies to be more personalized and targeted. A highly relevant approach involves utilizing the RFM (Recency, Frequency, Monetary) model processed with machine learning algorithms, specifically K-Means Clustering, to uncover hidden patterns within transaction data [5], [7], [8].

Previous research on customer segmentation using the RFM model and K-Means algorithm has been extensive. Various studies have demonstrated the effectiveness of this combination in identifying customer characteristics, distinguishing between loyal customers, potential customers, and those at risk of churn [2], [3]. However, the majority of existing studies focus on generating static analysis results presented in reports or simple visualizations. A research gap remains regarding the practical implementation of these analytical results into interactive tools accessible to non-technical decision-makers. The novelty of this research lies in the utilization of the Streamlit framework to develop a functional web application that integrates K-Means Clustering analysis with practical business requirements, bridging the gap between theoretical models and application [10].

Consequently, this study aims to design and develop an automated e-commerce customer segmentation application using the K-Means algorithm and Streamlit. The primary objective is to assist businesses in formulating effective, data-driven marketing strategies by

providing dynamic insights into customer segments. Furthermore, this research serves as an academic reference for the practical application of machine learning in developing modern business intelligence tools.

2. Research Methodology

This study employs a quantitative approach combined with system development methods. The research framework is designed systematically to ensure the accuracy of the segmentation results and the functionality of the developed application. The complete research flow is illustrated in Fig. 1.

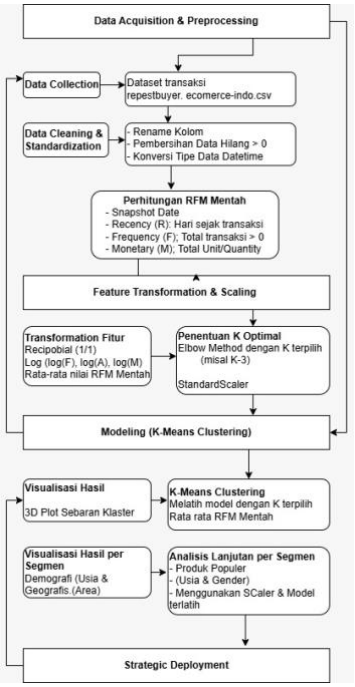

Fig. 1: flow diagram

2.1. Data acquisition

The primary data source for this study is a transaction dataset from an Indonesian e-commerce platform (repeatbuyer_ecommerce_indo.csv). The dataset comprises 1,000 transaction records gathered over a two-month period from June 1, 2022, to July 31, 2022. It involves 178 unique customers and includes key attributes such as Cust_ID, Trx_Date (transaction date), Product_Name, and Amount.

2.2. Data Preprocessing and RFM Calculation

Before analysis, the raw data undergoes a preprocessing stage. This involves converting the Trx_Date column into a datetime object and checking for missing values or inconsistencies. Following this, the data is transformed into an RFM (Recency, Frequency, Monetary) model to quantify customer behavior [8], [10]. The metrics are calculated as follows:

1) Recency (R)
   The number of days between the customer's last transaction and the snapshot date (August 1, 2022). A lower value indicates a more active customer [4].
2) Frequency (F)
   The total number of transactions made by a customer during the analysis period.
3) Monetary (M)
   The total value (sum of Amount) spent by the customer.

To handle the variations in data scale, the RFM values are normalized using the StandardScaler. This step is crucial for the K-Means algorithm, as it ensures that no single metric dominates the clustering process due to scale differences.

2.3. Clustering Algorithm (K-Means)

Customer segmentation is performed using the K-Means Clustering algorithm. K-Means partitions the dataset into (K) distinct, non-overlapping subgroups (clusters) where each data point belongs to the cluster with the nearest mean [12], [6]. The algorithm operates iteratively to minimize the Within-Cluster Sum of Squares (WCSS). Other studies have also successfully applied this algorithm for various evaluation and segmentation purposes [13].

To determine the optimal number of clusters (K), this study utilizes the Elbow Method. The algorithm is run for a range of (K) values (from 1 to 10), and the WCSS for each (K) is plotted. The optimal (K) is identified at the "elbow" point where the rate of decrease in WCSS slows down significantly.

2.4. Application Development

The final stage involves developing a web-based dashboard using Streamlit, an open-source Python framework designed for machine learning and data science projects. The application is built to:

1) Load the pre-trained K-Means model and segmented data.
2) Visualize the clusters using interactive 2D/3D scatter plots.
3) Display the statistical profile (mean RFM values) of each segment.
4) Provide filtering capabilities to allow stakeholders to analyze specific customer groups dynamically.

3. Result and Discussion

3.1. Data Overview

The experimental dataset comprises 1,000 transaction records involving 178 unique customers. The descriptive analysis of the raw data reveals that the transaction distribution is highly skewed, which justifies the need for data transformation before clustering. Most customers have a relatively low transaction frequency, while a small group of "power users" drives a significant portion of the total monetary value.

All title and author details must be in single-column format and must be centred.
Only the first word in a title must be capital and other word should be in small case. Author details must not show any professional title (e.g. Managing Director), any academic title (e.g. Dr.) or any membership of any professional organization (e.g. Senior Member IEEE).

To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith).
Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia). Email address is compulsory for the corresponding author.

3.2. Determination of Optimal Clusters

The first step in the K-Means analysis is determining the optimal number of clusters (K) using the Elbow Method. The algorithm was executed for (K=1) to (K=10), and the inertia (Within-Cluster Sum of Squares) was calculated for each iteration.

As shown in Fig. 2, the "elbow" point—where the reduction in inertia begins to diminish significantly is observed at (K=3). Therefore, three clusters were selected as the optimal number for segmenting the customer base.

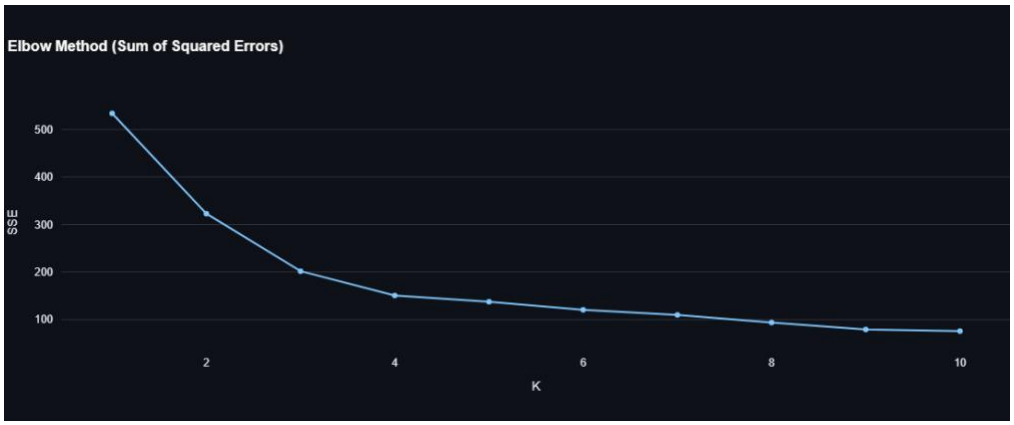

Fig 2: Elbow Method

3.3. Cluster Profiling and Interpretation

The K-Means algorithm successfully partitioned the 178 customers into three distinct segments. To interpret these segments, the centroid values of the standardized data were inverse-transformed back to their original RFM scales. The statistical profile of each cluster is summarized in Table I.

Table 1: Cluster Mean Values and Characteristics

Cluster	Recency (Days)	Frequency (Trans)	Monetary (Items)	Customer Count	Label
0	11.03	4.27	6.67	73 (41%)	Active & Potential
1	21.03	12.88	27.5	33 (18.5%)	Champions
2	43.57	5.90	5.90	72 (40.5%)	At Risk

Based on the RFM characteristics in Table I, the segments are analyzed as follows:

1) Cluster 0 (Active & Potential)

This is the largest segment, comprising 41% of the total customers. They have the lowest Recency (avg. 11 days), indicating they have transacted very recently. However, their Frequency and Monetary values are moderate. This group represents active users who have the potential to become loyal customers if nurtured correctly.

2) **Cluster 1 (Champions)**
Although this is the smallest group (18.5%), it is the most valuable. These customers exhibit the highest Frequency (avg. 12.8 transactions) and Monetary value (avg. 27.5 items). They are the loyal "Champions" of the platform.

3) **Cluster 2 (At Risk)**
This segment makes up 40.5% of the customer base. They have the highest Recency (avg. 43 days), meaning it has been a long time since their last purchase. Their low Frequency and Monetary scores suggest they are at high risk of churning (leaving the platform).

3.4. Implementation in Streamlit

The results of this segmentation were deployed into a web-based dashboard using the Streamlit framework. This application serves as a decision support tool for marketing teams.

Fig. 3 illustrates the application interface, showing the interactive 3D scatter plot of the clusters. Users can filter data by cluster labels to view specific customer lists. For instance, the marketing manager can quickly extract the list of "At Risk" customers to send reactivation emails or generate a list of "Champions" for a special reward program.

Fig 3: User Interface of the Customer Segmentation Application developed with Streamlit

The visualization clearly separates the three clusters, confirming that the K-Means algorithm with $K=3$ successfully differentiates customers based on their purchasing behavior.

4. Conclusion

This study successfully addresses the challenge of managing diverse e-commerce customer data by developing an automated segmentation application based on RFM analysis and K-Means Clustering. The application, built using the Streamlit framework, effectively segments customers and visualizes the results to support data-driven strategic decision-making.

The experimental results indicate that the Elbow Method determined $K=3$ as the optimal number of clusters. Three distinct customer segments were identified: 'Active & Potential' (characterized by low Recency), 'Champions' (high Frequency and Monetary values), and 'At Risk' (high Recency). The primary contribution of this research is bridging the gap between static data analysis and practical business implementation through an interactive and user-friendly web application.

For future research, it is recommended to extend the analysis period to at least 12 months to capture seasonal shopping patterns. Additionally, incorporating actual monetary values in currency, rather than unit counts, would provide more accurate revenue-based insights for the business.

References

[1] F. Nur Wahidah, "The Role of Social Media for E-Commerce in Indonesia.pdf," *J. Econ. Bus. Horiz.*, vol. 3, pp. 90–95, 2024.
[2] C. W. Prasetyandari, "E-Commerce as Indonesia's Economic Development Effort," *Indones. J. Econ. Manag.*, vol. 3, no. 1, pp. 70–78, 2022, doi: 10.35313/ijem.v3i1.4456.
[3] A. K. Fikri, A. U. Hamdani, and P. Hayati, "PENERAPAN E-COMMERCE BERBASIS CONTENT MANAGEMENT SYSTEM DENGAN METODE WATERFALL UNTUK PROMOSI," vol. 06, no. 01, 2025.
[4] Yoesoep Edhie Rachmad, "Social Media Marketing Mediated Changes In Consumer Behavior From E-Commerce To Social Commerce," *Int. J.*

Econ. Manag. Res., vol. 1, no. 3, pp. 227–242, 2022, doi: 10.55606/ijemr.v1i3.152.

[5] D. Mubarok *et al.*, "BIG DATA ANALYTICS DAN MACHINE LEARNING UNTUK MEMPREDIKSI PERILAKU KONSUMEN DI E-COMMERCE," 2025. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

[6] C. H. Ardana, A. A. A. A. A. Khoyum, and M. Faisal, "Segmentasi Pelanggan Penjualan Online Menggunakan Metode K-means Clustering," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 1, pp. 1–9, 2024, doi: 10.14421/jiska.2024.9.1.1-9.

[7] N. Hendrastuty, "Penerapan Data Mining MenggunakanAlgoritma K-Means ClusteringDalam Evaluasi HasilPembelajaran Siswa," *J. Ilm. Inform. Dan Ilmu Komput.*, vol. 3, no. 1, pp. 46–56, 2024, [Online]. Available: https://doi.org/10.58602/jima-ilkom.v3i1.26

[8] F. Amalia Maresti, G. Mustika Anugraheni, R. A. Hargiyanto, and K. Mustaqim, "Penerapan Exploratory Data Analysis (Eda) Dan Analisis Recency, Frequency, and Monetary (Rfm) Untuk Segmentasi Pelanggan E-Commerce," *Competitive*, vol. 19, no. 1, pp. 14–25, 2024, doi: 10.36618/competitive.v19i1.4059.

[9] R. M. Fauzan and G. Alfian, "Segmentasi Pelanggan E-Commerce Menggunakan Fitur Recency, Frequency, Monetary (RFM) dan Algoritma Klasterisasi K-Means," 2024.

[10] R. Artikel, R. C. Octavianus, H. Toba, and B. R. Suteja, "Segmentasi dan Pembentukan Model Regresi Nasabah Berbasis Analisis Recency , Frequency dan Monetary Segmentation and Formation of Customer Regression Model Based on Recency , Frequency and Monetary Analysis," vol. 8, pp. 474–484, 2022.

[11] T. K. Amarya, R. Firliana, and A. Ristyawan, "Aplikasi Deteksi Dini Penyakit Stroke Menggunakan Stramlit," *INOTEK, Vol. 9*, vol. 9, pp. 453–462, 2025.

[12] D. K. Dan and M. Rfm, "SEGMENTASI PELANGGAN MAJALAH PADA SITUS WEB E-COMMERCE SEGMENTATION OF MAGAZINE SUBSCRIBERS ON E-COMMERCE WEBSITE USING K-MEANS ++ AND RFM METHOD," vol. 11, no. 6, pp. 1243–1252, 2024, doi: 10.25126/jtiik.2024118208.

[13] D. Agustina, "FITUR SOCIAL COMMERCE DALAM WEBSITE E-COMMERCE DI INDONESIA," *J. Inform. Mulawarman*, vol. 12, no. 1, 2017.