Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.14974
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Customer Segmentation Using RFM and K-Means Clustering to Support CRM in Retail Industry

Yohanni Syahra[1], Abdul Fadlil[2], Herman Yuliansyah[3]
[1]Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
[1]Doctoral Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]Department Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3]Department Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1)] yohannisyahra@umsu.ac.id, [2)] fadlil@mti.uad.ac.id, [3)] herman.yuliansyah@tif.uad.ac.id

Submitted :Jun 18, 2025 | **Accepted** :Jul 22, 2025 | **Published** :Jul 23, 2025

Abstract: In today's highly competitive retail landscape, businesses face increasing challenges in retaining customer loyalty and achieving sustainable growth. A common issue, particularly among small and medium-sized enterprises (SMEs), is the absence of a structured method for identifying and categorizing customers based on their value and behavior. This study addresses the challenge by implementing a data-driven customer segmentation approach using Recency, Frequency, and Monetary (RFM) analysis combined with the K-Means clustering algorithm. The research utilized 2,353 transaction records from 369 unique customers collected over three years from a local retail business. After preprocessing and normalizing the RFM values using Min-Max scaling, the Elbow Method was applied to determine the optimal number of clusters, resulting in four distinct customer segments. Cluster 3, labeled "Loyal Customers," consisted of customers with high purchase frequency and very high spending; Cluster 1 ("Potential Loyalists") included those with moderate activity; Cluster 0 represented "At-Risk Customers," and Cluster 2 comprised "One-Time Buyers." This segmentation framework supports the development of targeted Customer Relationship Management (CRM) strategies, such as loyalty programs and re-engagement campaigns. However, the approach also has limitations, including potential data bias due to the use of static transaction records and the challenge of interpreting clusters without qualitative customer feedback. Despite these constraints, the study demonstrates the practical utility of combining RFM analysis with clustering techniques to extract actionable insights in environments with limited technical infrastructure.

Keywords: Customer Segmentation, RFM Analysis, K-Means Clustering, Retail Industry, CRM.

INTRODUCTION

Digital economy, customer-centric strategies have become the cornerstone of business sustainability, especially in the highly competitive retail industry(Zhu, 2023). With the increasing availability of consumer data, businesses now have the opportunity to better understand customer behavior and adapt their services accordingly. However, transforming raw transaction data into meaningful insights remains a challenge for many retailers, particularly small- and medium-sized enterprises. The key to this transformation lies in customer segmentation, the practice of dividing a customer base into groups that exhibit similar characteristics, behaviors, or needs. A more refined segmentation process enables businesses to implement tailored marketing strategies, improve customer loyalty, and ultimately increase profitability.

Among the many models used for customer segmentation, the RFM (Recency, Frequency, Monetary) model stands out due to its simplicity and effectiveness. The RFM model evaluates customers based on how recently they made a purchase (Recency), how often they purchase (Frequency), and how much they spend (Monetary). Each of these dimensions provides valuable insight into a customer's engagement and value to the business. When combined, these metrics create a powerful profile that can be used to segment customers for targeted marketing

*name of corresponding author

efforts. However, segmenting customers manually based on RFM values can be both inefficient and subjective, especially when dealing with large datasets. This is where machine learning techniques, particularly clustering algorithms such as K-Means, offer significant advantages. K-Means clustering is a popular unsupervised learning algorithm that groups data into clusters based on similarity. When applied to RFM data, it can uncover natural groupings within the customer base that may not be immediately visible through traditional analysis.

Several previous studies have demonstrated the effectiveness of RFM and clustering techniques in various domains. For instance, K-Means clustering on RFM data to segment customers in the e-commerce sector, leading to improved targeting strategies(Ma, 2022). Similarly,found that combining RFM with clustering increased customer retention rates in the hospitality industry(Paul & Ramanan, 2019). These studies highlight the relevance and adaptability of the RFM-K-Means approach across industries. Despite its proven utility, few studies have applied this method specifically to smaller retail environments in developing regions, where CRM systems are often underutilized due to limited technical resources. In this context, our research seeks to fill the gap by applying the RFM model and K-Means clustering to real-world retail data from a small business environment. The study is motivated by the observation that many retail businesses collect transaction data but rarely use it for strategic decision-making. This gap presents a missed opportunity to enhance customer engagement, reduce churn, and optimize marketing efforts. We hypothesize that combining RFM analysis with K-Means clustering will enable the identification of distinct customer segments that can support more effective CRM practices. The primary problem addressed in this research is the lack of structured, data-driven customer segmentation in small retail businesses. Without segmentation, marketing strategies tend to be broad and ineffective, failing to meet the expectations of different customer types. By identifying behavioral patterns through RFM and clustering analysis(Fatahi & Rabiei, 2020), this research aims to transform basic transaction records into strategic insights that can directly support business growth. In addition to its practical contributions, this study also offers a methodological advancement by demonstrating how normalized RFM variables and unsupervised clustering can be effectively integrated and adapted for small-scale data contexts, contributing to the literature on lightweight customer segmentation models suitable for resource-constrained environments. The main objective of this study is to develop a practical and scalable customer segmentation model using RFM and K-Means clustering techniques. This model is designed to assist retail businesses in recognizing high-value customers, identifying inactive customers, and tailoring their marketing strategies based on real customer behavior.

This study offers a practical solution by transforming raw retail transaction data into meaningful customer insights through the integration of RFM (Recency, Frequency, Monetary) analysis and K-Means clustering(Abbasimehr& Bahrini, 2022). By calculating RFM scores, the model captures key aspects of customer behavior, how recently and frequently they shop, and how much they spend. These scores are then processed using K-Means clustering to identify natural groupings among customers based on their purchasing patterns(Tabianan et al., 2022)(Nasyuha et al., 2022). The result is a clear segmentation of the customer base into actionable categories, such as loyal customers, frequent shoppers, and one-time buyers.The proposed segmentation framework serves as a valuable tool for small and medium-sized retail businesses aiming to enhance their Customer Relationship Management (CRM) strategies(Saha et al., 2021). Without requiring advanced infrastructure or technical expertise, this approach enables retailers to personalize their marketing efforts, design loyalty programs, and re-engage inactive customers more effectively. In the broader scope of digital transformation, this study emphasizes the role of data analytics in creating more adaptive, efficient, and customer-focused retail practices. The integration of RFM and clustering not only enriches the decision-making process but also supports long-term business growth through targeted and data-driven engagement strategies.

LITERATURE REVIEW

In recent years, the increasing availability of customer transaction data has driven the development of more sophisticated methods for customer segmentation. One of the most widely adopted approaches is the RFM (Recency, Frequency, Monetary) model, which evaluates customer behavior based on three key dimensions: the recency of the last transaction, the frequency of transactions over a period, and the monetary value spent. Originally introduced in direct marketing, RFM analysis has proven effective across multiple domains, especially in retail, due to its simplicity and strong correlation with customer lifetime value(Smaili & Hachimi, 2023).

Numerous studies have combined RFM with clustering techniques to generate more meaningful customer segments. K-Means clustering, in particular, has been favored for its computational efficiency and interpretability. For example, demonstrated how integrating RFM metrics with K-Means clustering enables businesses to uncover latent patterns in customer behavior(Christy et al., 2021). Their work suggested that the segmentation outcomes provided better targeting insights than relying on demographic segmentation alone. Highlighted the strength of unsupervised learning methods like K-Means in identifying homogeneous groups from high-dimensional behavioral data, reinforcing the method's role in customer profiling(Dol & Jawandhiya, 2023).

While many studies support the utility of RFM-K-Means integration, they often focus on large-scale organizations with access to advanced analytics infrastructure. For instance, used big data and parallel computing

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.14974 p-ISSN : 2541-044X

to optimize RFM segmentation in an e-commerce context(Li et al., 2021). Although the results showed improved precision in targeting campaigns, such implementations are rarely feasible for small retailers due to limited resources. This creates a practical gap in the literature: the need for lightweight, scalable segmentation approaches that smaller businesses can implement with minimal technical overhead.However, since RFM variables are measured in different scales and units, normalization is a critical preprocessing step to ensure fairness in clustering. The Min-Max normalization method is commonly used for this purpose, transforming each value to a range between 0 and 1. The formula is as follows:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X is the original RFM value, X_{min} and X_{max} are the minimum and maximum values of the feature respectively. This technique ensures that each RFM metric contributes equally to the clustering process, avoiding bias towards features with larger numeric ranges.

In another strand of research, scholars have explored enhancements to the traditional RFM model by incorporating variables such as customer loyalty, product preferences, and browsing behavior. These extensions, while informative, also introduce complexity and often require real-time data access, which may not be available in small retail settings (Chen & Lin, 2020). Therefore, a return to foundational RFM methods combined with interpretable clustering offers a compelling alternative, especially for businesses aiming to leverage offline sales records.Comparative studies have also evaluated different clustering algorithms, including hierarchical clustering and DBSCAN, for customer segmentation. However, many researchers agree that K-Means strikes a balance between performance and simplicity, making it suitable for routine segmentation tasks. For example, compared multiple clustering algorithms on retail datasets and found that K-Means offered the best trade-off in terms of accuracy and computational cost, particularly when RFM data was normalized beforehand(Rungruang et al., 2024).Despite this growing body of research, there remains a notable lack of studies focusing on the implementation of RFM and K-Means in small-scale, traditional retail environments, particularly in developing regions. Most literature assumes access to CRM systems or digital sales platforms, overlooking businesses that rely on basic transaction logs(Pynadath et al., 2023)(Sun et al., 2023). This study aims to bridge that gap by applying the RFM-K-Means framework using simple Excel-based transaction data, demonstrating how even basic tools can yield actionable customer insights(Heldt et al., 2021).

METHOD

To address the problem of unstructured customer data and the absence of targeted marketing strategies, this research proposes a data-driven solution using the RFM (Recency, Frequency, Monetary) model integrated with K-Means clustering(ASLANTAŞ et al., 2023). The method is applied to transactional data from a local retail business recorded over a span of three years. The data contains 2,353 transaction records from 369 unique customers and includes fields such as transaction date, item purchased, quantity, total spending, and customer identity.

Data Preprocessing

The raw data was first cleaned and standardized. Duplicate entries were removed, and transaction dates were converted into a consistent datetime format. Customer names were also normalized to prevent duplication due to inconsistent spelling or formatting.**Table 1** presents a sample of 100 cleaned transaction records, which include standardized transaction dates, normalized customer names, and the removal of duplicate entries. This subset of data serves as a representative overview of customer purchases and product distributions used in subsequent analysis.

Table. 1 Transaction Records

Date	Name of Product	QTY	TOTAL	Costumer
2022-01-03	Karak Kaliang	2	16000	Atur
2022-01-03	Rendang Telur	1	9000	Atur
2022-01-03	Dakak2 Beras	1	4000	Atur
2022-01-03	Rendang Telur	1	10000	Tempura
2022-01-04	Karak Kaliang	1	10000	Satria Akar Daya
2022-01-04	Rendang Telur	1	10000	Satria Akar Daya
2022-01-04	Sanjai Balado	1	10000	Satria Akar Daya
2022-01-04	Sanjai Lado Ijo	1	10000	Satria Akar Daya
2022-01-04	Dakak2 Beras	1	10000	Satria Akar Daya
2022-01-04	Sanjai Lado Ijo	1	10000	Sri Selamat
2022-01-04	Sanjai Balado	1	10000	Muslim
2022-01-04	Sanjai Lado Ijo	1	10000	Muslim

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.14974
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2022-01-04	Rendang Telur	1	10000	Muslim
2022-01-04	Rendang Telur	1	10000	Rahman
2022-01-05	Karak Kaliang	1	10000	Sumiati
2022-01-05	Sanjai Lado Ijo	1	10000	Sumiati
2022-01-07	Rendang Telur	1	10000	Budi
2022-01-07	Dakak2 Beras	2	20000	Halifah
2022-01-07	Karak Kaliang	1	10000	Halifah
2022-01-07	Sanjai Lado Ijo	1	10000	Halifah
2022-01-07	Rendang Telur	1	10000	Halifah
2022-01-07	Rendang Telur	1	10000	Halifah
2022-01-07	Karak Kaliang	3	30000	Indah
2022-01-07	Dakak2 Beras	1	10000	Indah
2022-01-07	Rendang Telur	1	10000	Indah
2022-01-07	Sanjai Lado Ijo	1	10000	Indah
2022-01-07	Rendang Telur	5	50000	Yoan
2022-01-07	Rendang Telur	1	10000	Yoan
2022-01-07	Ubi Cancang	1	8000	Yudi
2022-01-07	Sanjai Balado	2	16000	Yudi
2022-01-08	Sanjai Balado	1	10000	Dedy
2022-01-08	Rendang Telur	1	10000	Dedy
2022-01-08	Sanjai Lado Ijo	2	20000	Dedy
2022-01-08	Dakak2 Beras	1	10000	Dedy
2022-01-08	Sanjai Balado	1	10000	Dedy
2022-01-08	Rendang Telur	2	20000	Kaka
2022-01-08	Rendang Telur	2	20000	Upi
2022-01-08	dakak2 Ungu	1	10000	Upi
2022-01-11	Sanjai Lado Ijo	1	10000	Free Kaka
2022-01-11	Sanjai Balado	1	10000	Free Kaka
2022-01-11	Rendang Telur	1	10000	Free Kantin Da
2022-01-11	Rendang Telur	1	8000	Atur
2022-01-11	Rendang Telur	1	8000	Atur
2022-01-11	Dakak2 Beras	1	4000	Atur
2022-01-11	dakakungu	1	4000	Atur
2022-01-12	Sanjai Lado Ijo	2	20000	Astri
2022-01-12	Rendang Telur	14	126000	Rahman
2022-01-12	Karak Kaliang	16	128000	Rahman
2022-01-12	Sanjai Balado	10	80000	Rahman
2022-01-12	Sanjai Lado Ijo	4	32000	Rahman
2022-01-12	Dakak2 Beras	10	90000	Rahman
2022-01-12	Dakak2 Beras	5	20000	Anto
2022-01-12	sanjaibumbu	2	8000	Anto
2022-01-12	Ubi Cancang	6	24000	Anto
2022-01-17	Ubi Cancang	4	40000	Kiara Cell
2022-01-17	Rendang Telur	1	10000	Tempura
2022-01-17	dakak2 Ungu	1	10000	Tempura
2022-01-23	Sanjai Lado Ijo	1	10000	Anugerah
2022-01-23	Sanjai Balado	1	10000	Anugerah
2022-01-23	Karak Kaliang	1	10000	Anugerah
2022-01-23	Rendang Telur	1	10000	Anugerah
2022-01-23	Dakak2 Beras	1	10000	Anugerah
2022-01-23	dakakungu	1	10000	Anugerah
2022-01-23	Karak Kaliang	2	20000	Devi
2022-01-23	Sanjai Balado	2	20000	Devi
2022-01-23	Sanjai Lado Ijo	1	10000	Devi
2022-01-23	Sanjai Balado	1	10000	Devi
2022-01-23	Karak Kaliang	1	10000	Fahmi

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.14974 p-ISSN : 2541-044X

2022-01-23	Sanjai Lado Ijo	1	10000	Fahmi
2022-01-23	Rendang Telur	1	10000	Kiki
2022-01-23	Karak Kaliang	2	20000	Boy
2022-01-23	Karak Kaliang	1	10000	Roni Yuni
2022-01-23	Dakak2 Beras	1	10000	Roni Yuni
2022-01-23	dakakungu	1	10000	Roni Yuni
2022-01-23	Sanjai Balado	1	10000	Roni Yuni
2022-01-23	Rendang Telur	1	10000	Roni Yuni
2022-01-23	Sanjai Lado Ijo	1	10000	Roni Yuni
2022-01-25	Karak Kaliang	15	120000	Upi
2022-01-25	Dakak2 Beras	5	45000	Upi
2022-01-25	dakak2 Ungu	5	45000	Upi
2022-01-25	Sanjai Lado Ijo	5	40000	Upi
2022-01-25	Sanjai Balado	5	40000	Upi
2022-01-26	Dakak2 Beras	2	20000	Ita
2022-01-26	Karak Kaliang	2	20000	Ita
2022-01-27	Sanjai Balado	3	24000	Kantin Tgd
2022-01-27	Karak Kaliang	1	8000	Kantin Tgd
2022-01-27	dakak2 Ungu	2	18000	Kantin Tgd
2022-01-27	Sanjai Lado Ijo	3	24000	Kantin Tgd
2022-01-27	Dakak2 Beras	1	10000	Kantin Tgd
2022-01-27	Karak Kaliang	1	10000	Suci
2022-01-27	Sanjai Lado Ijo	2	20000	Suci
2022-01-28	Bakso Ikan Nila	2	20000	Zul Diza Ponsel
2022-01-30	Sanjai Lado Ijo	1	10000	Astri
2022-01-30	Rendang Telur	1	10000	Kaka
2022-01-31	Karak Kaliang	4	32000	Kantin Darul Adib
2022-01-31	Dakak2 Beras	10	80000	Kantin Darul Adib
2022-01-31	dakak2 Ungu	10	80000	Kantin Darul Adib
2022-01-31	Sanjai Lado Ijo	10	80000	Kantin Darul Adib
2022-01-31	Sanjai Balado	6	48000	Kantin Darul Adib
2022-02-03	Sanjai Balado	2	20000	Aisyah

Table 1 presents a sample of cleaned and structured transaction records used in this study to perform customer segmentation analysis. The table includes five main columns:

1. Date – This column records the date of each transaction. All dates are formatted consistently (YYYY-MM-DD) to facilitate temporal analysis, particularly for calculating the *Recency* variable in the RFM model.
2. Name of Product – This column lists the names of the products purchased by customers. These include various items such as "Karak Kaliang," "Rendang Telur," and "Sanjai Balado," indicating the diversity of products sold in the retail business.
3. QTY (Quantity) – This column shows the number of units purchased for each product per transaction. It reflects the buying volume per product.
4. TOTAL – This field represents the total spending (in local currency) for each product line item per transaction. It serves as the basis for calculating the *Monetary* value in the RFM analysis.
5. Customer – This column identifies the customer associated with each transaction. Customer names were normalized during preprocessing to ensure consistency and remove duplication caused by spelling variations.

RFM Variable Calculation

RFM variables were computed to analyze purchasing behavior. Recency was determined by calculating the number of days between the most recent transaction of each customer and the latest transaction date in the dataset. Frequency was measured as the total number of transactions made, while Monetary referred to the sum of all spending recorded. The resulting RFM table contained three scores per customer, representing patterns in transaction timing, frequency, and overall value.RFM Formulas:

- **Recency (R)** = Date of latest transaction in the dataset – Date of customer's most recent transaction
- **Frequency (F)** = Total number of transactions made by the customer
- **Monetary (M)** = Sum of all transaction amounts (TOTAL) made by the customer

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.14974
e-ISSN : 2541-2019
p-ISSN : 2541-044X

To better understand customer purchasing behaviour, RFM (Recency, Frequency, and Monetary) analysis was performed on a sample of 100 cleaned transaction records. Recency was calculated by measuring the number of days between the customer's last purchase and the date of the last transaction in the dataset. Frequency represents how often customers make purchases, while Monetary records the total amount spent.Table 2. The resulting RFM highlights different customer profiles based on their activity. For example, customers with low Recency scores (e.g., 0–5 days) are considered recently active, while those with higher Frequency and Monetary values indicate strong engagement and high value. This analysis enables data-driven segmentation, facilitating strategies such as targeted marketing, retention campaigns, or loyalty programmes aligned with actual customer behaviour.

Table 2. RFM Analysis

Customer	Recency	Frequency	Monetary
Aisyah	0	1	20000
Anugerah	11	6	60000
Astri	4	2	30000
Satria Akar Daya	30	5	50000
Zul Diza Ponsel	6	1	20000
Sri Selamat	30	1	10000
Budi	27	1	10000
Ita	8	2	40000
Dedy	26	5	60000
Devi	11	4	60000
Fahmi	11	2	20000
Free Kaka	23	2	20000
Free Kantin Da	23	1	10000
Halifah	27	5	60000
Indah	27	4	60000
Kak Yoan	27	2	60000
Kaka	4	2	30000
Kantin Darul Adib	3	5	320000
Kantin Tgd	7	5	84000
Kawan Nek Upi	26	2	30000
Kiara Cell	17	1	40000
Kiki	11	1	10000
Muslim	30	3	30000
Upi	9	5	290000
Atur	23	7	53000
Boy	11	1	20000
Rahman	22	6	466000
Roni Yuni	11	6	60000
Suci	7	2	30000
Sumiati	29	2	20000
Tempura	17	3	30000
Anto	22	3	52000
Yudi	27	2	24000

Normalization

Before proceeding to customer segmentation, it is essential to normalize the RFM values to ensure that each metric contributes equally to the analysis. Since Recency, Frequency, and Monetary values are measured on different scales, normalization transforms these variables into a common range, typically between 0 and 1. This step prevents any single metric from disproportionately influencing the results and allows for a fair comparison among customers. Normalized RFM scores are particularly important when applying clustering algorithms or assigning weighted scores in subsequent segmentation strategies.The table 3. below presents the normalized RFM scores for each customer. Normalization was performed using Min-Max scaling to ensure that Recency, Frequency, and Monetary metrics contribute equally to subsequent segmentation analysis. Each value now ranges between 0 and 1.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.14974
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3. normalized RFM scores for each customer

Customer	Recency	Frequency	Monetary
Aisyah	0.000000	0.000000	0.021930
Anugerah	0.366667	0.833333	0.109649
Astri	0.133333	0.166667	0.043860
Satria Akar Daya	1.000000	0.666667	0.087719
Zul Diza Ponsel	0.200000	0.000000	0.021930
Sri Selamat	1.000000	0.000000	0.000000
Budi	0.900000	0.000000	0.000000
Buk Ita	0.266667	0.166667	0.065789
Dedy	0.866667	0.666667	0.109649
Devi	0.366667	0.500000	0.109649
Fahmi	0.366667	0.166667	0.021930
Free Kaka	0.766667	0.166667	0.021930
Free Kantin Da	0.766667	0.000000	0.000000
Halifah	0.900000	0.666667	0.109649
Indah	0.900000	0.500000	0.109649
Yoan	0.900000	0.166667	0.109649
Kaka	0.133333	0.166667	0.043860
Kantin Darul Adib	0.100000	0.666667	0.679825
Kantin Tgd	0.233333	0.666667	0.162281
Kawan Nek Upi	0.866667	0.166667	0.043860
Kiara Cell	0.566667	0.000000	0.065789
Kiki	0.366667	0.000000	0.000000
Muslim	1.000000	0.333333	0.043860
Upi	0.300000	0.666667	0.614035
Atur	0.766667	1.000000	0.094298
Boy	0.366667	0.000000	0.021930
Rahman	0.733333	0.833333	1.000000
Roni Yuni	0.366667	0.833333	0.109649
Suci	0.233333	0.166667	0.043860
Sumiati	0.966667	0.166667	0.021930
Tempura	0.566667	0.333333	0.043860
Wak Anto	0.733333	0.333333	0.092105
Yudi	0.900000	0.166667	0.030702

K-Means Clustering

K-Means Clustering is an unsupervised machine learning algorithm used to partition a dataset into K distinct, non-overlapping clusters based on feature similarity. In the context of customer segmentation, this technique groups customers with similar purchasing behaviors, as represented by their normalized RFM scores. The primary goal is to minimize intra-cluster variation while maximizing inter-cluster differences. K-Means is particularly effective for RFM segmentation, as it objectively assigns customers to behavioral profiles based on distance metrics(Smaili & Hachimi, 2023)(Rungruang et al., 2024).The algorithm operates in iterative steps: it begins by randomly selecting K centroids, assigns each data point to the nearest centroid, recalculates centroids as the mean of the assigned points, and repeats this process until cluster assignments stabilize.

For each data point in the dataset, calculate the **Euclidean distance** between the point and each centroid. Assign the data point to the cluster with the closest centroid(Yulisasih et al., 2024):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (2)$$

Where:
$d(x,y)$: the distance between data point x and centroid y
yi:donates the value of the i-th attribute value of centroid y,
xi: represents the i-th attribute value of the data point x

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.14974 p-ISSN : 2541-044X

i: refers to the index of the attribute,
n: indicates the total number of the attribute for each data point.
The WCSS is calculated to measure the compactness of the clusters. For each cluster, compute the sum of squared distances between each data point and its cluster centroid(Kuraria et al., 2018),(Cui, 2020):

$$WCSS = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \quad (4)$$

Where:

- K: Total number of clusters.
- k:An index that refers to the current cluster, ranging from 1 to K.
- S_k: The set of data points assigned to cluster k.
- p: Number of features or dimensions in the dataset.
- j: An index that refers to the current feature or dimension, ranging from 1 to p.
- x_{ij}: The j-th feature value of the i-th data point.
- $\bar{x}_{kj}$: The mean value of the j-th feature in cluster k (i.e., the centroid's value).
- $(x_{ij} - \bar{x}_{kj})^2$: The squared Euclidean distance between the j-th feature of data point x_i and the corresponding feature of the centroid of its cluster.

This step essentially aggregates the squared Euclidean distances within each cluster, aiming to capture how tightly the data points are grouped.Repeat Steps 2–4 until convergence is reached. This typically occurs when the cluster assignments no longer change or when the decrease in WCSS between iterations is below a defined threshold.

Cluster Interpretation

The clustering process in this study was conducted using the K-Means algorithm implemented in Python through the scikit-learn library, which is widely adopted in data science and machine learning applications for its robustness and ease of use. The normalized RFM dataset served as the input for clustering, where each row represents a unique customer characterized by three features: Recency, Frequency, and Monetary value.

Prior to clustering, the optimal number of clusters (k) was determined using the Elbow Method, which plots the Within-Cluster Sum of Squares (WCSS) against different values of k. The point at which the rate of decrease sharply shifts (the "elbow") was identified at k=4, suggesting that four clusters offer a good balance between under- and over-segmentation.

The clustering was carried out using the following Python libraries and steps:

- Libraries used: pandas, numpy, matplotlib, seaborn, and sklearn.cluster.KMeans
- Normalization: Performed using MinMaxScaler from sklearn.preprocessing
- K-Means parameters:
  - n_clusters=4 (number of clusters)
  - init='k-means++' (initialization method)
  - n_init=10 (number of times the algorithm will be run with different centroid seeds)
  - max_iter=300 (maximum number of iterations per run)
  - random_state=42 (for reproducibility)

After fitting the model to the normalized RFM data, each customer was assigned a cluster label (0–3). These labels were then used to group customers and calculate average RFM scores per cluster. The results were visualized using 2D scatter plots of RFM components, colored by cluster, to interpret the behavioral patterns of each group.Each resulting cluster was profiled based on its average RFM values. The clusters were labeled according to customer value, such as:

- **Cluster 0**: High Frequency, High Monetary – "Loyal Customers"
- **Cluster 1**: Low Frequency, High Recency – "At-Risk Customers"
- **Cluster 2**: Low Monetary, Low Frequency – "One-time Buyers"
- **Cluster 3**:High Frequency, Very High Monetary – *"Loyal Customers"*

These profiles were then linked to actionable CRM strategies, such as offering loyalty rewards to frequent spenders or re-engagement campaigns for dormant customers.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.14974
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

Following the normalization of the RFM values, the next step is to apply K-Means clustering to segment customers based on their purchasing behavior. The purpose of K-Means is to partition the customers into distinct groups such that those within the same cluster exhibit similar Recency, Frequency, and Monetary characteristics. The algorithm aims to minimize the Within-Cluster Sum of Squares (WCSS), which represents the total squared distance between each customer and the centroid of the cluster to which they are assigned. By reducing this variance, K-Means ensures that customers within a cluster are as similar as possible, while maximizing the differences between clusters. This segmentation is instrumental in identifying meaningful customer profiles that can be used to tailor marketing strategies and improve customer relationship management.

K-Means

K-Means clustering was applied to the normalized RFM dataset to identify distinct customer segments based on their purchasing behavior. The algorithm successfully grouped customers into four clusters by minimizing the Within-Cluster Sum of Squares (WCSS), ensuring that individuals within the same cluster share similar levels of recency, frequency, and monetary value. Each customer was assigned to a cluster, labeled from 0 to 3, which represents their behavioral profile. These clusters serve as the foundation for understanding customer loyalty, activity, and value, and can be used to develop targeted marketing strategies and personalized engagement initiatives.

Table 4. Result K-Means

Customer	Recency	Frequency	Monetary	Cluster
Aisyah	0	1	20000	2
Anugerah	11	6	60000	1
Astri	4	2	30000	2
Satria Akar Daya	30	5	50000	1
Zul Diza Ponsel	6	1	20000	2
Sri Selamat	30	1	10000	0
Budi	27	1	10000	0
Ita	8	2	40000	2
Dedy	26	5	60000	1
Devi	11	4	60000	1
Fahmi	11	2	20000	2
Free Kaka	23	2	20000	0
Free Kantin Da	23	1	10000	0
Halifah	27	5	60000	1
Indah	27	4	60000	0
Yoan	27	2	60000	0
Kaka	4	2	30000	2
Kantin Darul Adib	3	5	320000	3
Kantin Tgd	7	5	84000	1
Kawan Nek Upi	26	2	30000	0
Kiara Cell	17	1	40000	2
Kiki	11	1	10000	2
Muslim	30	3	30000	0
Upi	9	5	290000	3
Atur	23	7	53000	1
Boy	11	1	20000	2
Rahman	22	6	466000	3
Roni Yuni	11	6	60000	1
Suci	7	2	30000	2
Sumiati	29	2	20000	0
Tempura	17	3	30000	0
Anto	22	3	52000	0
Yudi	27	2	24000	0

Cluster interpretation

Cluster interpretation was conducted to uncover meaningful behavioral patterns within each customer segment. By analyzing the average Recency, Frequency, and Monetary values of each cluster, distinct profiles

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.14974
e-ISSN : 2541-2019
p-ISSN : 2541-044X

emerged that reflect different levels of customer engagement and value. This step is critical in translating raw clustering results into actionable business insights. For instance, clusters with high frequency and monetary values are typically indicative of loyal customers who contribute significantly to revenue, whereas clusters with low frequency and high recency may represent at-risk or inactive customers. These interpretations enable organizations to tailor customer relationship management (CRM) strategies more effectively, rewarding loyalty, encouraging repeat purchases, or reactivating dormant customers. The table below summarizes the average RFM characteristics of each cluster, forming the basis for these strategic insights.

Table 5. Cluster Interpretation

Cluster	Recency	Frequency	Monetary	Interpretation
1	18.25	5.375	60875	Potential Loyalists
3	11.3334	5.333	358666.666	Loyal Customers
0	25.67	2.1665	29666.666	At-Risk Customers
2	7.9	1.5	26000	One-Time Buyers

These profiles provide actionable insights:

- Cluster 3 includes the most profitable and active customers, suitable for premium retention and loyalty strategies.
- Cluster 1 represents customers with consistent purchases but lower value, indicating growth potential.
- Cluster 0 contains inactive or declining customers who require reactivation.
- Cluster 2 captures casual buyers, possibly influenced by promotions.

These behavioral insights allow businesses to tailor their Customer Relationship Management (CRM) strategies more effectively.

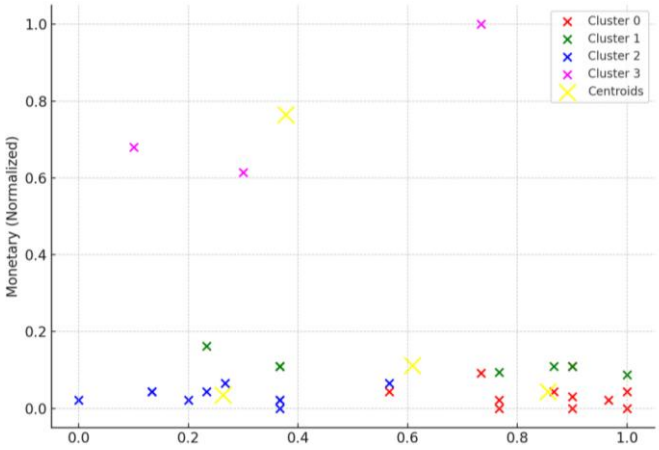

Figure 1. Cluster of Costumer

DISCUSSIONS

The results of the RFM analysis and K-Means clustering revealed distinct customer segments, each exhibiting unique purchasing behaviors. By normalizing the RFM variables and applying the K-Means algorithm, the dataset was successfully partitioned into four interpretable clusters. These clusters ranged from high-value, loyal customers to low-engagement, one-time buyers. Such segmentation highlights not only the diversity in customer engagement levels but also provides valuable insights for strategic business decision-making.One of the significant findings is the identification of Cluster 3 as comprising the most loyal and valuable customers, with high purchase frequency and the highest monetary contribution. These customers are crucial for sustaining long-term profitability and should be prioritized for loyalty programs and personalized services. On the other hand, Cluster 2 was identified as containing one-time buyers with low spending and infrequent interactions. This group may represent customers acquired through promotions or casual purchases and could benefit from re-engagement campaigns.

Furthermore, the application of K-Means demonstrated the practical advantages of unsupervised learning in customer behaviormodeling. Unlike traditional segmentation methods based on intuition or demographic profiles, clustering based on normalized transactional data allows for a more objective and data-driven approach. The results also confirm the importance of normalization, as it ensured that no single RFM dimension disproportionately influenced the clustering outcome.Comparatively, had other clustering methods such as Hierarchical Clustering or DBSCAN been applied, the structure and number of clusters might differ, especially in handling outliers or irregular purchasing patterns. However, K-Means was chosen for its efficiency and interpretability in segmenting medium-sized datasets with clear behavioral trends.

*name of corresponding author

CONCLUSION

This study successfully demonstrates the integration of RFM analysis and K-Means clustering as a practical approach to customer segmentation in the retail sector. Utilizing transactional data from 369 customers, the analysis identified four distinct customer segments: Loyal Customers, Potential Loyalists, At-Risk Customers, and One-Time Buyers. The application of Min-Max normalization and the Elbow Method contributed to accurate and meaningful clustering outcomes. The findings reveal that Loyal Customers (Cluster 3) contribute the highest revenue and should be prioritized with reward-based strategies. Potential Loyalists (Cluster 1) require nurturing and engagement to promote long-term retention. At-Risk Customers (Cluster 0), who exhibit low purchasing frequency and declining engagement, should be targeted with reactivation campaigns. Meanwhile, One-Time Buyers (Cluster 2) demonstrate minimal interaction and may respond better to introductory promotions aimed at encouraging repeat purchases. The consistency and clarity of the segmentation results, driven by unsupervised learning techniques, support more precise targeting, enhanced customer retention, and more efficient marketing strategies. Despite the effectiveness of this approach, a notable limitation lies in the reliance on static RFM variables, which may not fully capture evolving customer behavior. Future research should explore the incorporation of dynamic behavioral modeling and alternative clustering algorithms to improve segmentation accuracy. Nevertheless, the proposed framework presents a scalable and cost-effective solution, particularly beneficial for small retailers with limited access to advanced CRM systems.

.

REFERENCES

Abbasimehr, H., & Bahrini, A. (2022). An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. *Expert Systems with Applications*, *192*, 116373. https://doi.org/10.1016/j.eswa.2021.116373

ASLANTAŞ, G., GENÇGÜL, M., RUMELLİ, M., ÖZSARAÇ, M., & BAKIRLI, G. (2023). Customer Segmentation Using K-Means Clustering Algorithm and RFM Model. *Deu Muhendislik Fakultesi Fen ve Muhendislik*, *25*(74), 491–503. https://doi.org/10.21205/deufmd.2023257418

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Accounting, Auditing and Finance*, *1*(1), 5–8. https://doi.org/10.23977/accaf.2020.010102

Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining—A survey. *Engineering Applications of Artificial Intelligence*, *122*, 106071.

Fatahi, S., & Rabiei, M. (2020). *Users clustering Based on Search Behavior Analysis Using the LRFM Model (Case Study: Iran Scientific Information Database (Ganj)) Somayeh*. *36*(2), 419–442. https://doi.org/10.35050/JIPM010.2020.006

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, *127*(March), 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

Kuraria, A., Jharbade, N., & Soni, M. (2018). Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining. *International Journal of Scientific Research in Science, Engineering and Technology*, *6*(1), 190–195. https://doi.org/10.32628/ijsrset21841122

Li, M., Wang, Q., Shen, Y., & Zhu, T. Y. (2021). Customer relationship management analysis of outpatients in a Chinese infectious disease hospital using drug-proportion recency-frequency-monetary model. *International Journal of Medical Informatics*, *147*(July 2020). https://doi.org/10.1016/j.ijmedinf.2020.104373

Ma, J. (2022). *E-commerce Customer Segmentation Based on RFM Model* (pp. 926–931). https://doi.org/10.1007/978-981-16-8052-6_118

Nasyuha, A. H., Zulham, Z., & Rusydi, I. (2022). Implementation of K-means algorithm in data analysis. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *20*(2), 307. https://doi.org/10.12928/telkomnika.v20i2.21986

Paul, L., & Ramanan, T. R. (2019). An RFM and CLV analysis for customer retention and customer relationship management of a logistics firm. *International Journal of Applied Management Science*, *11*(4), 333. https://doi.org/10.1504/IJAMS.2019.103713

Pynadath, M. F., Rofin, T. M., & Thomas, S. (2023). Evolution of customer relationship management to data mining-based customer relationship management: a scientometric analysis. *Quality & Quantity*, *57*(4), 3241–3272. https://doi.org/10.1007/s11135-022-01500-y

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449. https://doi.org/10.1016/j.eswa.2023.121449

*name of corresponding author

Saha, L., Tripathy, H. K., Nayak, S. R., Bhoi, A. K., & Barsocchi, P. (2021). Amalgamation of customer relationship management and data analytics in different business sectors—a systematic literature review. *Sustainability (Switzerland)*, *13*(9). https://doi.org/10.3390/su13095279

Smaili, M. Y., & Hachimi, H. (2023). New RFM-D classification model for improving customer analysis and response prediction. *Ain Shams Engineering Journal*, *14*(12), 102254. https://doi.org/10.1016/j.asej.2023.102254

Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, *9*(2), e13384. https://doi.org/10.1016/j.heliyon.2023.e13384

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability (Switzerland)*, *14*(12), 1–15. https://doi.org/10.3390/su14127243

Yulisasih, B. N., Herman, H., Sunardi, S., & Yuliansyah, H. (2024). Evaluation of K-Means Clustering Using Silhouette Score Method on Customer Segmentation. *ILKOM Jurnal Ilmiah*, *16*(3), 330–342. https://doi.org/10.33096/ilkom.v16i3.2325.330-342

Zhu, N. (2023). Research on Customer Relationship Segmentation of Apparel Retail Industry through Data Mining. *HighTech and Innovation Journal*, *4*(2), 309–314. https://doi.org/10.28991/HIJ-2023-04-02-05

*name of corresponding author