

Customer Segmentation using WRFM-Based Feature Representation with K-Means and PCA

Sumpunt Khongmark and Daranat Tansui

Department of Computer and Informatics for Management,
Faculty of Communication Sciences, Prince of Songkla University

Email¹: daranat.t@psu.ac.th

Date Received : 24 October 2025

Date Revised : 18 June 2026

Date Accepted : 19 June 2026

Date Published online : 26 June 2026

Abstract

This study investigates customer segmentation using a Weighted Recency, Frequency, and Monetary (WRFM)-based feature representation combined with Principal Component Analysis (PCA) and K-Means clustering. A dataset consisting of 286 transactions from 100 customers was transformed into RFM variables and normalized using Min–Max scaling to ensure comparability. Entropy-based weighting was then applied to construct WRFM features, providing a balanced representation of customer behavior. PCA was employed to reduce dimensionality and eliminate redundant information prior to clustering. Four clustering methods—K-Means, PCA-enhanced K-Means, Gaussian Mixture Model (GMM), and Hierarchical Clustering—were evaluated using Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. The results indicate that PCA-enhanced K-Means achieved the best performance (Silhouette = 0.473, Calinski–Harabasz = 144.817, Davies–Bouldin = 0.725), demonstrating improved cluster compactness and separation. The findings show that combining WRFM-based features with PCA effectively reduces bias and improves clustering quality. The segmentation results identify four distinct customer groups, including high-value loyal customers, potential customers, moderate customers, and at-risk customers. These insights support targeted marketing strategies such as retention programs, personalized promotions, and customer relationship management. This study does not propose a new

¹ Corresponding author

clustering model but focuses on empirical evaluation and applied validation of existing techniques. The approach is scalable and applicable to other domains involving multidimensional customer data.

Keywords: Customer Segmentation, Weighted RFM (WRFM), Entropy-Based Weighting, Principal Component Analysis, K-Means Clustering

1. Introduction

Customer segmentation is a fundamental process in strategic marketing, enabling organizations to understand purchasing behaviors and design targeted campaigns that address diverse customer needs. By identifying distinct customer groups, businesses can optimize resource allocation, enhance customer loyalty, and improve overall profitability. Among various approaches, the Recency, Frequency, and Monetary (RFM) model is widely adopted due to its simplicity and interpretability in capturing key aspects of customer behavior.

However, traditional RFM-based segmentation often suffers from limitations when applied to real-world datasets. Differences in scale among RFM variables may introduce bias, while conventional clustering algorithms may struggle to effectively separate customer groups in complex and high-dimensional data. These issues can lead to suboptimal segmentation and reduced effectiveness in supporting marketing decisions. To address these challenges, this study proposes an integrated framework that combines normalization, Weighted Recency-Frequency-Monetary (WRFM) modeling, Principal Component Analysis (PCA), and clustering techniques. Normalization ensures comparability across variables, while entropy-based weighting constructs the WRFM score to provide a balanced representation of customer importance. PCA is then applied to reduce dimensionality and eliminate redundant information, enhancing cluster separability.

K-Means clustering is employed due to its computational efficiency and scalability, and its performance is compared with other methods, including Gaussian Mixture Model (GMM) and Hierarchical Clustering. By integrating PCA and WRFM into the clustering process, the proposed framework aims to improve segmentation accuracy and interpretability.

Despite the growing adoption of PCA-enhanced clustering, comprehensive evaluations that integrate weighted RFM models with multiple clustering techniques remain limited. This

study addresses this gap by systematically comparing clustering performance and providing actionable customer insights based on the segments derived.

Objectives of the Study

1. To preprocess and transform customer transaction data into normalized RFM variables and construct a Weighted RFM (WRFM) model using entropy-based weighting.
2. To evaluate and compare the performance of clustering algorithms (K-Means, PCA-enhanced K-Means, Gaussian Mixture Model, and Hierarchical Clustering) using standard evaluation metrics.
3. To identify meaningful customer segments that provide actionable insights for marketing strategies, including high-value, potential, moderate, and at-risk customer groups.

Research Questions

RQ1: How does the integration of PCA and WRFM improve clustering performance compared to traditional K-Means in RFM-based customer segmentation?

RQ2: What customer insights can be derived from the identified clusters to support strategic marketing decisions, such as targeted promotions, retention, and customer relationship management?

2. Literature Review & Theoretical Background

2.1 Customer Segmentation and the RFM Model

Customer segmentation is a crucial process in strategic marketing, enabling businesses to recognize differences in purchasing behavior and design targeted campaigns. Proper analysis of customer groups not only boosts sales but also fosters long-term customer relationships. One of the most widely adopted frameworks for behavioral analysis is the RFM (Recency, Frequency, Monetary) model, reflecting purchase frequency, recency of the last transaction, and total spending (Wong et al., 2024). RFM has gained popularity in analytical marketing due to its ability to provide clear insights into customer purchasing behavior. Several studies confirm that RFM-based analysis allows businesses to plan promotional strategies more effectively, particularly when targeting loyal high-value customers or identifying those at risk of churn (John et al., 2023). Additionally, RFM provides a simple yet powerful foundation for integrating advanced machine learning and clustering methods. Its structured metrics make it

highly adaptable for diverse industries, ranging from e-commerce and banking to telecommunications. This flexibility allows organizations across different sectors to apply consistent customer evaluation criteria while tailoring segmentation strategies to their specific business contexts.

2.2 Clustering Approaches for RFM Analysis

Among clustering algorithms, K-Means is one of the most widely used methods due to its simplicity, scalability, and computational efficiency. However, it exhibits performance limitations when applied to high-dimensional or imbalanced datasets, often resulting in unclear or overlapping clusters (Syahra et al., 2025). To address these issues, dimensionality reduction techniques have been integrated prior to clustering. Principal Component Analysis (PCA) is frequently applied to reduce noise and complexity in the data while preserving the underlying distribution. Patibandla et al. (2025) and Dessureault and Massicotte (2021) demonstrated that combining PCA with K-Means enhances segmentation quality by simplifying the data structure and improving cluster separation. Recent studies have also emphasized that PCA reduces redundancy among correlated RFM features, allowing the clustering algorithm to focus on the most informative dimensions. Additionally, PCA facilitates better visualization of clusters in a two-dimensional space, enabling clearer interpretation of customer behavior patterns. These advantages make PCA-K-Means particularly suitable for marketing applications where interpretability and actionability are critical.

2.3 Recent Advances in RFM-Based Segmentation

Beyond PCA-enhanced K-Means, several innovative approaches have emerged in recent years. Wang et al. (2024) integrated RFM with multivariate time series clustering (LRFMS) to capture evolving customer behaviors more effectively. Vianna Filho et al. (2025) applied graph-based methods by formulating the problem as a max k-cut on graphs constructed from RFM dimensions, thereby improving both computational efficiency and segmentation accuracy. Madhiraju et al. (2024) combined RFM with advanced machine learning models such as XGBoost and CatBoost, demonstrating improved prediction accuracy and clearer classification of customer groups. Similarly, Anitha and Patil (2022) used RFM with K-Means to cluster retail customers, showing benefits in predicting purchasing behavior and supporting

business decision-making. In addition, Smaili and Hachimi (2023) proposed the RFM-D model by introducing a diversity parameter to improve segmentation accuracy.

2.4 Extensions of the RFM Framework

While the traditional RFM model provides a simple and interpretable representation of customer behavior, it may suffer from scale imbalance, particularly when monetary values dominate due to larger numerical ranges. To address this limitation, normalization methods such as Min–Max scaling are applied to transform RFM variables into a comparable range, ensuring that each dimension contributes proportionally to the clustering process. In addition to normalization, weighting approaches such as entropy-based methods are used to determine feature importance based on data distribution, assigning higher weights to more informative variables. The resulting Weighted RFM (WRFM) model provides a more balanced and meaningful representation of customer behavior. Previous studies have demonstrated that integrating weighted RFM models with clustering algorithms significantly improves segmentation performance and interpretability. Previous studies have demonstrated improvements in RFM-based segmentation through advanced modeling approaches. Wang et al. (2024) proposed an extended RFM framework incorporating additional behavioral dimensions, while Tripathi et al. (2025) showed that integrating feature transformation techniques with clustering enhances segmentation performance. These findings highlight the importance of integrating normalization and weighting techniques into RFM-based segmentation. Therefore, the incorporation of normalization and WRFM serves as a critical enhancement to traditional RFM-based segmentation frameworks, enabling more accurate customer grouping and supporting more effective data-driven marketing strategies.

3. Research Methodology

This study aims to develop a customer segmentation framework based on RFM analysis, enhanced with entropy-based weighting, dimensionality reduction, and clustering techniques. The proposed methodology consists of four main phases: data preprocessing, RFM construction and scoring, feature transformation, and clustering, followed by a comprehensive evaluation phase. To achieve this aim, the study is guided by three specific objectives:

1. To preprocess and transform customer transaction data into RFM features.
2. To evaluate and compare the performance of clustering algorithms, including K-Means, PCA-K-Means, Gaussian Mixture Model (GMM), and Hierarchical Clustering, using standardized evaluation metrics.
3. To identify meaningful customer segments that provide actionable insights for marketing strategies, such as distinguishing high-value customers, moderate spenders, and at-risk groups. These objectives are operationalized through the research framework illustrated in Figure 1, which presents the overall process from data preprocessing to clustering and evaluation. Each phase is designed to build upon the previous step, ensuring a systematic and reproducible workflow for customer segmentation analysis.

3.1 Dataset Design and RFM Metric Calculation

This study utilizes a transactional dataset collected from a local construction company in southern Thailand. The transactions primarily involve construction-related products, including electrical supplies, tools, and hardware materials. Due to a confidentiality agreement with the company, detailed product-level attributes are not disclosed in this study. The data set consists of 286 purchase records from 100 unique customers, where each transaction is associated with three main attributes: Customer ID (unique identifier), Purchase Date (transaction date), and Purchase Amount (monetary value). Since each customer may have multiple purchase records, the dataset is well-suited for deriving behavioral metrics using the RFM model. Specifically, Recency represents the time since the most recent purchase, Frequency denotes the total number of transactions, and Monetary reflects the total spending of each customer (Aliyev et al., 2020; Wong et al., 2024).

Given the presence of multiple transactions per customer, the dataset enables the construction of aggregated behavioral features. These features form the basis of the RFM model, which is widely used for analyzing customer behavior (Wong et al., 2024) and subsequently employed for clustering analysis. The raw transaction data is preprocessed to ensure consistency by converting date formats, aggregating purchases per customer, and removing duplicates or missing values. After preprocessing, RFM values are computed for each customer and organized into a structured table. This RFM table subsequently serves as the primary input for the clustering and segmentation analysis.

This RFM table subsequently serves as the primary input for the clustering and segmentation analysis. The structured format of the RFM table ensures that all customers are represented consistently, facilitating direct comparison of purchasing behaviors across the dataset. Moreover, the aggregated features provide a standardized foundation for applying normalization, entropy-based weighting, and clustering algorithms in the subsequent stages of the analysis. This systematic approach also supports reproducibility, as the same preprocessing and aggregation procedures can be applied to transactional datasets from other industries. By maintaining a clear data preparation process, the study ensures that the resulting customer segments are both reliable and interpretable for practical marketing decision-making.

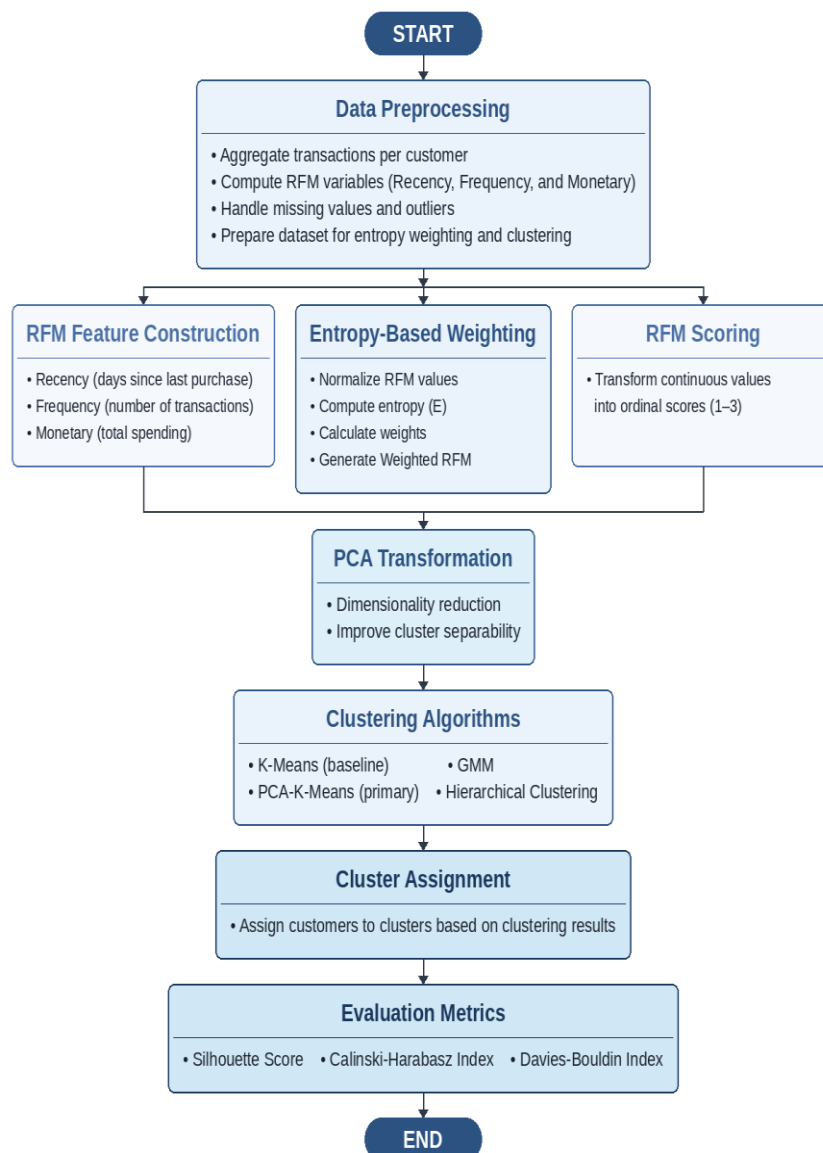


Figure 1 WRFM-Based Customer Segmentation Framework

3.1.1 Data Design

The dataset, titled `Customer_multiple_purchases.csv`, serves as the foundation for customer segmentation analysis. Each record corresponds to an individual transaction, and customers are uniquely identified using Customer ID.

The dataset includes the following attributes:

- 1) Customer ID: Unique identifier for each customer
- 2) Purchase Date: Date of transaction
- 3) Purchase Amount: Monetary value of each purchase

3.1.2 Data Preprocessing

The raw transactional data was preprocessed to ensure data consistency and analytical readiness. First, all Purchase Date values were converted into a standard datetime format to enable accurate temporal calculations. Next, transactions were aggregated at the customer level, ensuring that each customer is represented by a single consolidated record.

Following aggregation, the RFM metrics were computed as follows:

- 1) Recency (R): Number of days since the most recent purchase
- 2) Frequency (F): Total number of transactions per customer
- 3) Monetary (M): Total spending per customer

In addition, missing values and duplicate records were identified and removed to improve data quality. The resulting RFM table serves as the primary input for subsequent stages, including entropy-based weighting, dimensionality reduction, and clustering, as illustrated in Figure 1.

Step 1: Preprocessing

The first step of the RFM analysis involves preparing the raw transaction data for further processing, as illustrated in Figure 1. The dataset contains three attributes: Customer ID, Purchase Date, and Purchase Amount.

Initially, all purchase dates are converted into a standard datetime format to ensure consistency. A reference date is defined as the most recent purchase date in the dataset, which serves as a baseline for calculating Recency.

Following this, customer-level aggregation is performed. For each customer, the most recent purchase date is identified, and Recency is calculated as the number of days between the reference date and the last purchase date. Frequency is determined by counting the total number of purchases made by each customer, while Monetary is obtained by summing all purchase amounts. The results are presented in an RFM table, which lists Recency, Frequency, and Monetary values for each customer. For example, customer CUST0001 exhibits a Recency of 0 days, a Frequency of 4, and a Monetary value of 11,001, indicating recent activity, frequent purchases, and high spending behavior.

This step transforms raw transactional data into structured behavioral features, forming the foundation for subsequent RFM scoring and clustering processes.

3.1.3 RFM Calculation

Following preprocessing, the constructed RFM variables (Recency, Frequency, and Monetary) are used to represent customer purchasing behavior (Aliyev et al., 2020).

- 1) **Recency:** Number of days since the customer's last purchase, measured relative to the reference date.
- 2) **Frequency:** Total number of purchase transactions recorded for each customer within the observation period.
- 3) **Monetary:** Sum of all purchase amounts linked to each customer.

The resulting RFM table provided a structured numerical representation of customer behavior, which was utilized in the subsequent clustering and segmentation process. These features capture key behavioral characteristics of customers and serve as the foundational input for subsequent transformation and clustering processes.

Step 2: RFM Scoring (Feature Transformation)

In this study, continuous RFM values obtained from the pre-processing stage were transformed into discrete ordinal scores using a threshold-based scoring scheme. This transformation ensures comparability among the three dimensions, enabling the data to be directly utilized in clustering algorithms. Specifically, Recency was inversely scored, such that customers with more recent purchases received higher scores. Frequency and Monetary were positively scored, with higher values corresponding to higher scores. Each dimension was

categorized into three levels: High (score = 3), Medium (score = 2), and Low (score = 1). The thresholds were defined as follows:

- 1) Recency: 0 days $\rightarrow$ High (3); 1–15 days $\rightarrow$ Medium (2); >15 days $\rightarrow$ Low (1)
- 2) Frequency: ≥ 4 purchases $\rightarrow$ High (3); 2–3 purchases $\rightarrow$ Medium (2); 1 purchase $\rightarrow$ Low (1)
- 3) Monetary: ≥ 8000 units $\rightarrow$ High (3); 4000–7999 units $\rightarrow$ Medium (2); <4000 units $\rightarrow$ Low (1)

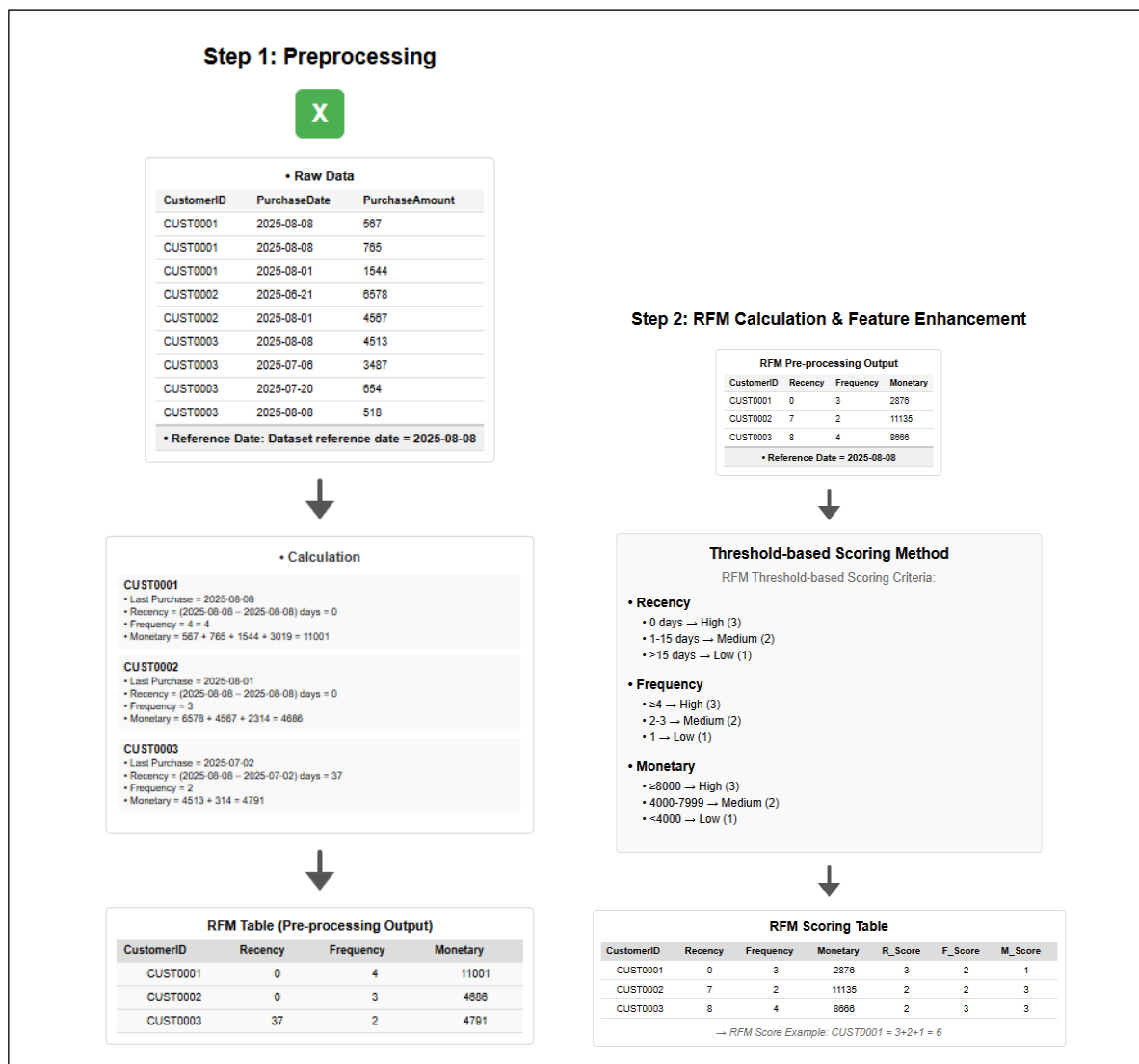


Figure 2 Example of RFM Scoring Table after transformation

This scoring procedure yields three integer values per customer, represented as a feature vector (R_score, F_score, M_score). For instance, customer CUST0001 with values (Recency = 0, Frequency = 3, Monetary = 2876) is represented as (3, 2, 1). Similarly, CUST0002 with (7, 2, 11135) becomes (2, 2, 3). These discrete vectors reduce scale differences across features and standardize behavioral patterns into a unified framework.

This transformation ensures that customers with similar behavioral patterns are grouped consistently, while differences in recency, frequency, and monetary value are clearly represented. By converting continuous variables into discrete scores, the study reduces extreme-value bias and produces a balanced, interpretable dataset for clustering. The resulting scores are summarized in Figure 2, which presents the RFM Scoring Table illustrating how continuous values are converted into ordinal scores for subsequent clustering.

To support further analysis, an optional composite score $RFM_sum = R_score + F_score + M_score$ can be computed for descriptive comparison, although the clustering stage utilizes the full 3-dimensional vector. The resulting RFM Scoring Table combines the original metrics (Recency, Frequency, Monetary) with their categorical scores (R_score, F_score, M_score), which are then used as input for the clustering algorithms in the next step. This feature is an engineering solution, as it not only ensures compatibility among features but also enhances the interpretability of customer groups. By explicitly defining scoring thresholds, the study produces segments that align with practical marketing criteria, enabling actionable insights such as identifying high-value customers, at-risk customers, and low-value customers.

3.2 Entropy-Based Weighting

In this study, an entropy-based weighting method is employed to enhance the traditional RFM model by assigning objective weights to each dimension—Recency (R), Frequency (F), and Monetary (M). Unlike conventional RFM approaches that treat all dimensions equally, this method determines the relative importance of each feature based on data variability. This approach allows more informative features to contribute more significantly to the analysis (Qu et al., 2022).

First, the RFM values obtained from the previous step were normalized to ensure comparability across dimensions. The normalized value for customer and variable is defined as: First, the RFM values obtained from the previous step were normalized to ensure

comparability across dimensions. The normalized value p_{ij} for customer i and variable j is defined as:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^n x_{ij}} \quad (1)$$

where x_{ij} represents the value of customer i in variable j , and n is the total number of customers.

Next, the entropy value for each variable was calculated to measure the degree of dispersion or uncertainty in the data:

A higher entropy value indicates that the data are more uniformly distributed, implying lower discriminative power. Conversely, a lower entropy value suggests greater variability and higher importance for clustering.

The entropy values were then transformed into weights using the following formula:

$$w_j = \frac{1-E_j}{\sum_{j=1}^m (1-E_j)} \quad (2)$$

where m is the number of variables (in this case, $m = 3$ corresponding to R, F, and M). The resulting weights (w_1, w_2, w_3) represent the relative importance of each RFM dimension, with the sum of all weights equal to 1.

Finally, the weighted RFM score for each customer was computed as:

$$WRFM_i = w_1 R_i + w_2 F_i + w_3 M_i \quad (3)$$

This weighted representation enhances the traditional RFM model by emphasizing more informative features, thereby improving the effectiveness of subsequent clustering. The resulting weighted RFM values were then used as input for dimensionality reduction and clustering algorithms in the next stage. The entropy weight method (EWM) is a widely used objective weighting technique that determines the importance of each variable based on the degree of dispersion in the dataset. A lower entropy value indicates higher variability and greater discriminative power, thus assigning a higher weight to that variable (Qu et al., 2022).

After computing the Weighted RFM scores for all 100 customers, each customer is represented by a weighted combination of Recency, Frequency, and Monetary values. This transformation enhances the discriminative power of the dataset by emphasizing more

informative features based on entropy-derived weights. However, the weighted RFM variables may still contain redundancy and potential correlation among features. Therefore, to further improve the structure of the data and enhance clustering performance, a dimensionality reduction technique is applied in the next step.

3.3 Feature Transformation using PCA

Principal Component Analysis (PCA) is employed to transform the normalized RFM variables (R_{norm} , F_{norm} , M_{norm}) into a new set of uncorrelated components (Dessureault & Massicotte, 2021). This step reduces dimensionality while preserving the most significant variance in the data.

To enhance clustering performance and reduce noise, PCA is applied after the normalization process. First, the normalized RFM variables are standardized using Z-score normalization, as shown in Equation (1), to ensure that all variables contribute equally to the analysis. Next, the covariance matrix is computed as shown in Equation (2), capturing the relationships among the standardized variables. Eigenvalues and eigenvectors are then obtained by solving Equation (3), where the eigenvectors represent the principal directions of maximum variance (Abdulhafedh, 2021), where the eigenvectors represent the principal directions of maximum variance. Finally, the principal components are derived using Equation (4), projecting the original data onto a new set of uncorrelated axes. These transformed features provide a compact and informative representation of customer behavior and are subsequently used as input for clustering algorithms.

This transformation improves cluster separability and enhances the overall performance of the clustering process. Prior to PCA, all features are standardized using the Z-score normalization:

$$Z = \frac{X - \mu}{\sigma} \quad (4)$$

where X is the original feature, μ is the mean, and σ is the standard deviation.

The covariance matrix is then computed as:

$$C = \frac{1}{n-1} Z^T Z \quad (5)$$

Eigenvalues and eigenvectors are obtained by solving:

$$Cv = \lambda v \quad (6)$$

Finally, the principal components are derived as:

$$PC = ZV \quad (7)$$

These components provide a reduced and uncorrelated representation of the original features, improving clustering performance.

The output of the RFM scoring process is a transformed dataset in which each customer is assigned discrete scores for Recency, Frequency, and Monetary (R, F, M), typically ranging from 1 to 3 (or 1 to 5 depending on the scoring scheme).

3.4 Customer Segmentation using Normalized Weighted RFM Model

This study employs the Recency–Frequency–Monetary (RFM) model to analyze customer purchasing behavior. RFM is widely used in customer analytics to evaluate customer value based on three dimensions: recency (R), frequency (F), and monetary value (M) (Wong et al., 2024).

However, since R, F, and M are measured on different scales, normalization is required to ensure comparability. In this study, Min–Max normalization is applied to transform all variables into a common range [0,1], as shown in Equation (1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

where X represents the original R, F, or M values.

After normalization, entropy weighting is used to determine the relative importance of each factor. The entropy method evaluates the variability of each attribute, assigning higher weights to more informative features, which enhances the effectiveness of customer segmentation (Aliyev et al., 2020).

Finally, the normalized weighted RFM score is calculated as follows:

$$WRFM = w_R \cdot R_{norm} + w_F \cdot F_{norm} + w_M \cdot M_{norm} \quad (9)$$

where w_R, w_F, w_M are the entropy-based weights of Recency, Frequency, and Monetary, respectively.

Since the RFM variables (Recency, Frequency, and Monetary) are measured on different scales, normalization is required to ensure fair contribution from each feature. To demonstrate this effect, an example comparison before and after normalization is presented in Table 1.

Table 1 presents an example illustrating the effect of normalization on RFM variables

Scenario	Recency (R)	Frequency (F)	Monetary (M)	Interpretation
Before Normalization	20	4	11001	Dominated by Monetary (M)
After Normalization	0.2	0.5	0.8	Balanced contribution

Table 1 presents an illustrative example of the effect of normalization on RFM variables, high-magnitude variables such as Monetary (M) may dominate the analysis. Min–Max normalization scales all features to [0,1], ensuring balanced contributions and improving the reliability of subsequent weighting and clustering.

This step helps in visualizing customer groups and enhances the separation between clusters.

Step 3: Data Preparation and Feature Transformation

This step focuses on transforming transactional data into customer-level features using the RFM (Recency, Frequency, Monetary) model.

1. Recency Calculation (R)

Recency is calculated as the number of days between the customer's most recent transaction and the reference date (i.e., the latest transaction date in the dataset). This measure reflects how recently a customer has engaged with the business.

2. Frequency Calculation (F)

Frequency represents the total number of transactions made by each customer within the observation period.

This variable indicates the level of customer engagement.

3. Monetary Calculation (M)

Monetary value is defined as the total amount of money spent by each customer across all transactions.

This reflects the economic value contributed by the customer.

4. Aggregation of Customer-Level Data

All transactional records are aggregated to obtain a single R, F, and M value per customer.

This transformation converts transactional data into a structured dataset suitable for analytical modeling.

5. Feature Normalization

Since R, F, and M are measured on different scales, Min-Max normalization is applied to transform all variables into the range [0,1]

This approach enables a fair comparison between clustering methods under both normalized and non-normalized scenarios.

This step prevents high-magnitude variables, particularly Monetary, from dominating the model.

6. Preparation for Weighted RFM Model

The normalized RFM variables (R_{norm} , F_{norm} , M_{norm}) are used as input for the entropy-based weighting process and customer segmentation model described in.

3.5 Clustering Algorithms and Assignment

To identify meaningful customer groups, multiple clustering algorithms were applied to the RFM dataset, with and without feature transformation using Principal Component Analysis (PCA). Each customer is represented as a feature vector derived from the normalized RFM variables (R_norm, F_norm, M_norm) and the corresponding weighted RFM score (WRFM_norm), extending the traditional RFM scoring approach (Aliyev et al., 2020; Patibandla et al., 2025). Since R, F, and M are measured on different scales, Min–Max normalization is applied to ensure comparability and to prevent high-magnitude variables, particularly Monetary, from dominating the clustering process (Jain, 2010). This results in a more balanced and reliable segmentation. For dimensionality reduction and improved visualization, PCA is optionally applied to transform the feature space into a lower-dimensional representation while preserving the most significant variance in the data (Jolliffe, 2016). The clustering process incorporates multiple approaches, including centroid-based, probabilistic, and hierarchical methods, to enable comprehensive comparison and evaluation.

K-Means Clustering: K-Means is one of the most widely used clustering algorithms due to its simplicity and scalability. It partitions the dataset into k clusters by minimizing the within-cluster sum of squares (WCSS). In this study, the traditional K-Means algorithm is employed as the baseline method for performance comparison.

3.5.1 K-Means Original

The traditional K-Means algorithm was used as the baseline method due to its simplicity and computational efficiency. It partitions the dataset into k clusters by minimizing the within-cluster sum of squares. (WCSS) (MacQueen, 1967)

$$\operatorname{argmin}_C \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu\|^2 \quad (10)$$

Where μ_i is the centroid C_i

3.5.2 PCA-K-Means

To improve clustering quality, Principal Component Analysis (PCA) was applied prior to executing the K-Means algorithm. PCA reduces dimensionality and eliminates redundant noise, thereby enhancing cluster separation. Customers are first projected onto the most significant principal components before clustering is performed, resulting in improved interpretability and performance (Dessureault & Massicotte, 2021). PCA transforms the original data.

$$Z = XW \quad (11)$$

where W [$w_1, w_2 \dots w_m$] contains the eigenvectors associated with the largest eigenvalues of the covariance matrix.

Clustering is then performed on Z using the standard K-Means objective:

$$\operatorname{argmin}_c \sum_{i=1}^k \sum_{z \in c_i} \|z - \mu\|^2 \quad (12)$$

Where k = number of clusters, C_i = set of data points assigned to cluster i , Z = data point in the reduced feature space, μ = centroid of cluster

3.5.3 Gaussian Mixture Model (GMM)

Unlike K-Means, which assumes spherical clusters, the Gaussian Mixture Model (GMM) applies a probabilistic approach by modeling the data as a mixture of Gaussian distributions. Each customer is assigned to a cluster based on the probability of belonging to a Gaussian component. This method provides greater flexibility, enabling clusters of varying shapes and densities (John et al., 2023).

3.5.4 Hierarchical Clustering

Hierarchical clustering builds a hierarchy of clusters using agglomerative (bottom-up) or divisive (top-down) strategies. In this study, the agglomerative approach was applied, where customers are iteratively merged based on similarity until the desired number of clusters is reached. This method supports intuitive visualization and helps reveal the underlying structure of the data (Abdulhafedh, 2021). The probability density of data point x is:

$$p(x) = \sum_{k=1}^k \pi_k N_{(x/\mu_k, \Sigma k)} \quad (13)$$

Where π_k mixing coefficient of the k -th Gaussian $\sum_{k=1}^K \pi_k = 1$
 μ_k, Σ_k mean and covariance of the Gaussian component

$N(x/\mu_k, \Sigma_k)$ = Gaussian probability density function

Parameter estimation is performed through the Expectation-Maximization algorithm.

3.5.5 Evaluation Metrics

To evaluate the clustering performance in an unbiased manner, this study utilizes three well-established metrics: Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. These evaluation metrics are applied to clustering results derived from the normalized RFM features to ensure fair comparison across different algorithms.

(1) Silhouette Score

Silhouette Score evaluates the degree to which a data point belongs to its assigned cluster relative to other clusters. For each instance i , the Silhouette coefficient is computed as follows:

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (14)$$

where a_i represents the mean distance between i and all other points within the same cluster (intra-cluster cohesion), and b_i denotes the minimum average distance from i . The resulting score lies between -1 and 1, with higher values reflecting more distinct and well-separated clusters.

(2) Calinski–Harabasz Index

The Calinski–Harabasz Index, also known as the Variance Ratio Criterion, measures clustering quality by comparing the dispersion between clusters to the dispersion within clusters. It is calculated using the formula:

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \quad (15)$$

where SSB denotes the inter-cluster variance, SSW represents the intra-cluster variance, n is the number of observations, and k is the number of clusters. A higher CH score suggests that the clusters are well-separated and internally cohesive.

(3) Davies–Bouldin Index

The Davies–Bouldin Index (DBI) quantifies clustering quality by computing the average dissimilarity between each cluster and the one most like it. Similarity is determined by the ratio of intra-cluster dispersion to inter-cluster distance. A lower DBI value indicates better-defined and more distinct clusters.

$$DB = \frac{1}{K} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \quad (16)$$

where σ_i denotes the mean distance between the data points in cluster i and its centroid c_i , and $d(c_i, c_j)$ represents the distance between the centroids of clusters i and j . A lower Davies–Bouldin Index indicates more effective and well-separated clustering.

4. Results and Analysis

This section presents the experimental results of customer segmentation using the RFM framework integrated with multiple clustering techniques. For clarity and structured discussion, the findings are organized into three main parts: (i) evaluation of clustering performance across different algorithms, (ii) interpretation of cluster characteristics derived from the segmentation results, and (iii) discussion of business implications and their relevance to strategic marketing.

4.1 Descriptive Statistics of the Dataset

The results summarized in Tables 2 and 3 provide an overview of the dataset and clustering performance. Among the evaluated methods, the PCA-K-Means approach achieved the best overall performance, with a Silhouette Score of 0.473, a Calinski-Harabasz Index of 144.82, and a Davies-Bouldin Index of 0.725. These metrics collectively indicate that

PCA-K-Means generated compact and well-separated clusters, outperforming the other clustering techniques considered in this study.

Compared to traditional K-Means and Hierarchical Clustering, the PCA-K-Means model demonstrated greater stability across multiple evaluation metrics, highlighting its robustness in handling RFM-based data. Furthermore, the relatively low Davies-Bouldin Index suggests that the clusters exhibit strong internal cohesion while maintaining clear separation from one another. This characteristic enhances the interpretability of the segmentation results, making them more suitable for practical applications in marketing analysis and customer relationship management.

Table 2 Summary of the dataset

Metric	Value
Number of Customers	100
Number of Transactions	286
Average Frequency	2.85
Average Monetary	7384.71

Table 3 Sample of RFM Table (R, F, M, Scores)

CustomerID	Recency	Frequency	Monetary
CUST0001	20	4	11001
CUST0002	20	3	4686
CUST0003	57	4	9596
CUST0004	14	4	6760
CUST0005	130	1	3656

4.2 Feature Engineering Using Entropy-Based Weighting

This section presents the features of engineering process applied to the RFM variables. Initially, Min–Max normalization was employed to standardize the RFM features, as shown in Table 3. Subsequently, an entropy-based weighting method was applied to determine the relative importance of each feature (Table 5). The resulting weights were then used to

compute a weighted RFM score (WRFM), as presented in Table 6. This approach improves customer behavior representation by emphasizing more informative features.

Table 4 Normalized RFM Values (Min–Max Scaling)

CustomerID	R_norm	F_norm	M_norm
CUST0001	0.111732	0.75	0.525756
CUST0002	0.111732	0.5	0.217421
CUST0003	0.318436	0.75	0.457155
CUST0004	0.078212	0.75	0.318686
CUST0005	0.726257	0	0.167131

Table 5 Entropy-Based Feature Weights

Feature	Entropy	Divergence	Weight
R_norm	0.935013	0.064987	0.325723
F_norm	0.910167	0.089833	0.450258
M_norm	0.955305	0.044695	0.224019

Table 6 Weighted Normalized RFM Features

CustomerID	R_norm	F_norm	M_norm	WRFM_norm
CUST0001	0.111732	0.75	0.525756	0.491866
CUST0002	0.111732	0.5	0.217421	0.310229
CUST0003	0.318436	0.75	0.457155	0.543827
CUST0004	0.078212	0.75	0.318686	0.434561
CUST0005	0.726257	0	0.167131	0.273999

This section presents the features of engineering process applied to the RFM variables. As shown in Table 4, Min–Max normalization was performed to standardize the RFM features into a comparable scale.

Table 5 presents the entropy-based weighting results, indicating that the Frequency feature has the highest importance (weight = 0.4503), followed by Recency (0.3257) and

Monetary (0.2240). This suggests that customer purchase frequency plays a more significant role in distinguishing customer segments in this dataset.

Using these weights, a weighted RFM score (WRFM) was computed, as shown in Table 6. The WRFM values provide a more informative representation of customer behavior, which improves the effectiveness of the subsequent clustering process.

4.3 Principal Component Analysis (PCA) Results

To improve clustering performance, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the RFM dataset and eliminate redundancy among features. PCA transforms correlated variables into a set of orthogonal (uncorrelated) principal components while preserving most of the original data variance. This transformation enhances computational efficiency and facilitates more effective visualization and interpretation of customer segments.

As shown in Table 7, the first principal component (PC1) explains 73.44% of the total variance, followed by PC2 with 21.57%, and PC3 with 4.99%. Together, the first two components account for approximately 95.01% of the total variance, indicating that most of the information in the original dataset can be effectively represented in a reduced two-dimensional space.

This result demonstrates that PCA successfully compresses the RFM features while retaining essential information, thereby improving the efficiency and quality of the clustering process. Consequently, the reduced feature space is suitable for subsequent clustering analysis and visualization.

Table 7 PCA Explained Variance

Principal Component	Explained Variance Ratio	Explained Variance (%)
PC1	0.734382	73.438178
PC2	0.215689	21.568909
PC3	0.049929	4.992912

Figure 3 presents the distribution of customers in a two-dimensional space obtained through Principal Component Analysis (PCA). The original RFM variables are transformed into principal components, where PC1 and PC2 capture most of the data variance, enabling an

effective low-dimensional representation. The visualization reveals clear separation among customer groups, indicating that customers with similar purchasing behaviors are positioned close to one another, while those with distinct behaviors are located further apart. This demonstrates that PCA successfully preserves the underlying structure of the data. Furthermore, the application of K-Means clustering ($k = 4$) identifies four distinct customer segments, as indicated by different colors in the figure. The clusters exhibit minimal overlap and well-defined boundaries, suggesting strong segmentation performance.

Overall, the results confirm that combining PCA with K-Means clustering provides an effective approach for visualizing and identifying meaningful customer segments, supporting data-driven decision-making in customer relationship management. The clear spatial separation among clusters indicates that the underlying customer behavior patterns are well captured, supporting the robustness and reliability of the segmentation model.

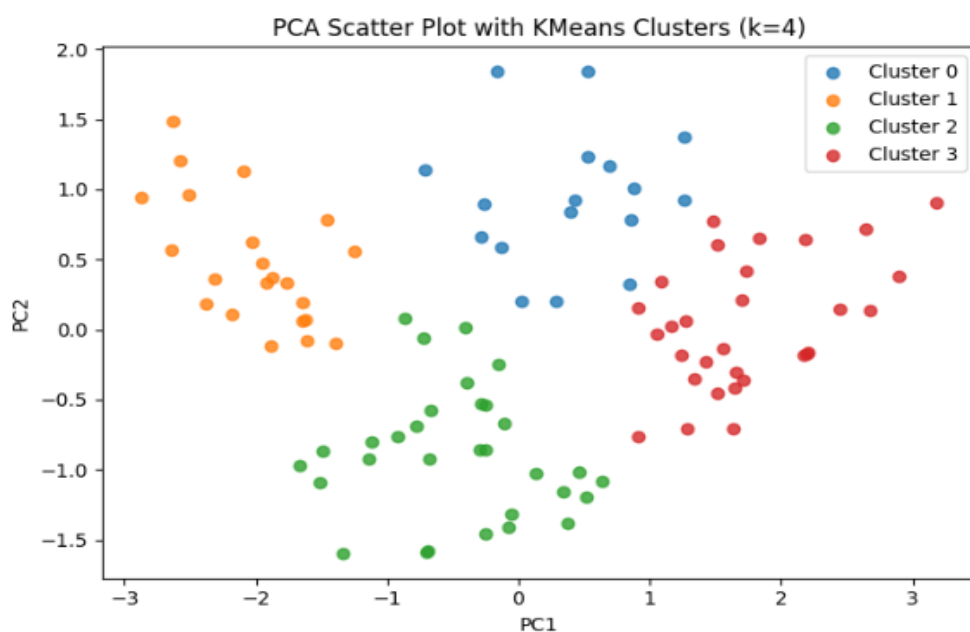


Figure 3 PCA Scatter Plot with K- Means Clusters ($k = 4$)

4.4 Clustering Results

Table 8 presents a comparative analysis of clustering performance using three evaluation metrics—Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index—across four clustering methods: K-Means, PCA- K-Means, Gaussian Mixture Model (GMM), and Hierarchical Clustering.

Table 8 Clustering Performance Comparison

Algorithm	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
K-Means (Original)	0.4137	125.5247	0.8733
PCA-K-Means	0.4733	144.8167	0.7253
GMM	0.3716	92.0015	0.9000
Hierarchical Clustering	0.4712	141.6544	0.7114

The results in Table 8 indicate that PCA-enhanced K-Means achieved the best clustering performance among all evaluated methods. Specifically, it obtained the highest Silhouette score (0.473), the highest Calinski–Harabasz index (144.817), and the lowest Davies–Bouldin index (0.725), demonstrating superior cluster quality compared to other algorithms. In contrast, traditional K-Means (Original) showed moderate performance and serves as a baseline for evaluating the improvements gained through PCA-based clustering.

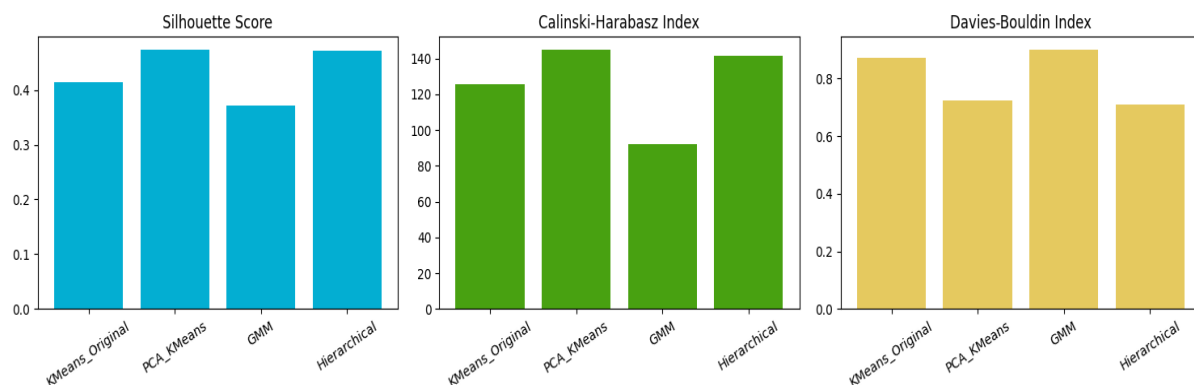
**Figure 4** Bar chart comparison of evaluation metrics.

Figure 4 presents a comparative analysis of four clustering algorithms—PCA-enhanced K-Means, K-Means (Original), Gaussian Mixture Model (GMM), and Hierarchical Clustering—using three evaluation metrics: Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index (Patibandla et al., 2025). The results consistently confirm that PCA-enhanced K-Means outperforms the other methods across all evaluation metrics.

4.5 Cluster Profiles and Interpretation

This section analyzes customer clusters using key indicators, including Recency, Frequency, and Monetary Value. Profiles help reveal behavioral patterns and support strategic segmentation. The cluster profiles are summarized in Table 9.

Table 9 Cluster profiles (mean Recency, Frequency, Monetary, Basket Size, Day Type)

Cluster	Recency	Frequency	Monetary
0	24	3.33	8,457
1	13	5	9,409
2	53.25	3.25	8,822.25
3	54	3	6,336

The profiling results reveal four distinct customer groups

1. Cluster 0 (New but Valuable Customers) Customers in this group exhibit relatively low recency (24 days since last purchase) and moderate purchase frequency, with a high monetary value (8,457). They represent recently engaged customers who demonstrate strong spending potential.
2. Cluster 1 (High-value Loyal Customers) This cluster shows the lowest recency (13 days) and the highest purchase frequency (5), coupled with the highest monetary value (9,409). These customers are highly engaged and generate substantial revenue, making them critical for retention strategies.
3. Cluster 2 (Dormant but High Spenders) With higher recency (53.25 days), this cluster indicates longer inactivity, though frequency and monetary values remain moderately high (3.25 and 8,822). These customers are at risk of churning but still contribute significantly to spending.
4. Cluster 3 (At-risk/Low-value Customers): Customers here have the longest recency (54 days), the lowest frequency (3), and the lowest monetary value (6,336). They represent at-risk customers who require reactivation campaigns or personalized incentives to re-engage

4.6 Summary of Findings

This section presents the findings of customer segmentation experiments using the proposed framework, which integrates Principal Component Analysis (PCA), Weighted Recency-Frequency-Monetary (WRFM), and K-Means clustering. To ensure clarity and comprehensiveness, the results are discussed in three main parts: (i) evaluation of clustering performance across multiple algorithms, (ii) interpretation of cluster characteristics, and (iii) business implications for strategic marketing.

4.6.1 Evaluation of Clustering Performance

The first part of the analysis evaluates the performance of four clustering algorithms: K-Means, PCA-enhanced K-Means, Gaussian Mixture Model (GMM), and Hierarchical Clustering. Three standard metrics—Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index—were employed to assess clustering quality in terms of compactness, separation, and overall validity. Among the tested methods, PCA-enhanced K-Means consistently delivered the best performance, achieving the highest Silhouette and Calinski–Harabasz values, along with the lowest Davies–Bouldin index. These results confirm that the integration of PCA improves clustering effectiveness by reducing noise and enhancing the separability of customer groups. In addition, the incorporation of WRFM further strengthens the model by providing a balanced and normalized representation of customer behavior. This directly addresses **Research Question 1 (RQ1)** by demonstrating that the combined PCA–WRFM–K-Means framework outperforms traditional clustering approaches.

4.6.2 Interpretation of Cluster Characteristics

The second part of the analysis focuses on interpreting the clusters derived from the PCA-enhanced K-Means model. Based on the updated WRFM-based profiling, customers are segmented into four distinct groups, each exhibiting unique behavioral characteristics. High-value loyal customers demonstrate recent purchasing activity, high frequency, and high monetary contribution, supported by high WRFM scores. Potential customers show balanced engagement and strong growth potential. Moderate customers maintain stable but average engagement levels, while at-risk customers exhibit long inactivity, low purchasing frequency, and low monetary value, reflected in lower WRFM scores. The integration of normalized RFM variables and WRFM scoring enables a more accurate and unbiased interpretation compared to traditional raw RFM analysis.

4.6.3 Business Implications

The final part discusses the practical implications of the segmentation results. Identifying high-value customers enables businesses to design retention strategies such as loyalty programs and personalized rewards. Potential and moderate customers can be targeted through upselling and cross-selling campaigns to enhance their lifetime value. Meanwhile, at-risk customers require reactivation strategies, including targeted promotions and personalized incentives. The proposed PCA-WRFM-K-Means framework provides actionable insights that support data-driven decision-making in customer relationship management. This directly addresses **Research Question 2 (RQ2)** by demonstrating how customer segmentation results can be translated into effective marketing strategies. Furthermore, the framework is scalable and adaptable to other domains involving complex customer data.

5. Discussion

The comparative evaluation confirms that PCA-enhanced K-Means consistently outperforms other clustering methods across all evaluation metrics. It achieves the highest Silhouette Score (0.473), the highest Calinski–Harabasz Index (144.817), and the lowest Davies–Bouldin Index (0.725), indicating compact, well-separated, and stable clusters. Compared to traditional K-Means, the improvement can be attributed to the dimensionality reduction capability of PCA, which reduces redundancy and enhances cluster separability. Moreover, the integration of WRFM further improves clustering quality by mitigating scale bias and ensuring balanced feature contributions. In contrast, the Gaussian Mixture Model (GMM) shows weaker performance, suggesting limitations in capturing clearly separated clusters in this dataset. Similarly, Hierarchical Clustering demonstrates lower effectiveness due to scalability issues and less balanced cluster formation.

These findings are consistent with recent studies. Wang (2025) reported that integrating PCA with K-Means improves segmentation accuracy in digital marketing. Syahra et al. (2025) confirmed the effectiveness of RFM-based clustering in customer relationship management, while Tripathi et al. (2025) emphasized the role of PCA in enhancing cluster stability. Zheng et al. (2024) further highlighted improvements in RFM-based analysis through optimized computational techniques. In summary, the integration of PCA, WRFM, and K-Means provides a robust, scalable, and effective framework for customer segmentation.

5.1 Limitations and Future Work

This study has several limitations that should be acknowledged.

First, PCA was applied only in combination with K-Means, while GMM and Hierarchical Clustering were performed on the original RFM features. This design choice was made because K-Means relies on Euclidean distance and is particularly sensitive to correlated and high-dimensional features, making it the most suitable candidate for PCA-based enhancement. However, this selective application means that the observed performance differences may partially reflect the effect of dimensionality reduction rather than algorithmic superiority alone. Future studies should consider applying PCA consistently across all clustering methods for a more controlled comparison.

Second, the clustering experiments were conducted using a single initialization setting (`random_state = 42`, `n_init = 50`). Although `n_init = 50` mitigates sensitivity to initial centroids, running experiments across multiple random seeds and reporting averaged results would further strengthen the robustness of the findings.

Third, the dataset was collected from a single construction company in southern Thailand, consisting of 286 transactions from 100 customers. The relatively small sample size and domain-specific context may limit the generalizability of the results. Future research should validate the proposed framework by using larger and more diverse datasets from different industries.

Finally, although the dataset involves construction-related products such as electrical supplies, tools, and hardware, detailed product-level attributes were not incorporated into the segmentation model due to confidentiality constraints. Incorporating product categories into future segmentation frameworks may enhance the depth and applicability of the analysis.

References

- Abdulhafedh, A. (2021). Incorporating K-means, hierarchical clustering and PCA in customer segmentation. *Journal of City and Development*, 3(1), 12–30.
- Aliyev, M., Ahmadov, E., Gadirlı, H., Mammadova, A., & Alasgarov, E. (2020). *Segmenting bank customers via RFM model and unsupervised machine learning*. arXiv. <https://arxiv.org/abs/2008.08662>

- Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), 1785–1792.
- Dessureault, J.-S., & Massicotte, D. (2021). *Feature selection or extraction decision process for clustering using PCA and FRSD*. arXiv. <https://arxiv.org/abs/2111.10492>
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.
- John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics*, 2(4), 809–823. <https://doi.org/10.3390/analytics2040042>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 374(2065), Article 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
- Madhiraju, B., Reddy, S., & Sasikala, G. (2024). Customer segmentation using RFM analysis. *EPRA International Journal of Economic and Business Review*, 12(7), 15-22.
- Patibandla, K. K., Daruvuri, R., & Mannem, P. (2025). Enhancing online retail insights: K-Means clustering and PCA for customer segmentation. In *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 388–393). IEEE. <https://ieeexplore.ieee.org/document/11011448>
- Qu, W., Li, J., Song, W., Li, X., Zhao, Y., Dong, H., Wang, Y., Zhao, Q., & Qi, Y. (2022). Entropy-weight-method-based integrated models for short-term intersection traffic flow prediction. *Entropy*, 24(7), Article 849. <https://doi.org/10.3390/e24070849>
- Smaili, M. Y., & Hachimi, H. (2023). New RFM-D classification model for improving customer analysis and response prediction. *Ain Shams Engineering Journal*, 14(12), Article 102254. <https://doi.org/10.1016/j.asej.2023.102254>
- Syahra, Y., Fadlil, A., & Yuliansyah, H. (2025). Customer segmentation using RFM and K-Means clustering to support CRM in retail industry. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 9(3), 1120-1131. <https://doi.org/10.33395/sinkron.v9i3.14907>

- Tripathi, S., Bachmann, N., Brunner, M., Tuezuen, A., Thienemann, A.-K., Pöchtrager, S., & Jodlbauer, H. (2025). Evaluation of clustering with PCA for market segmentation: A study using simulated and surrogate data. *Procedia Computer Science*, 253, 2063-2075. <https://doi.org/10.1016/j.procs.2025.01.267>
- Vianna Filho, A. L. C., de Lima, L., & Kleina, M. (2025). *A graph-based approach to customer segmentation using the RFM model*. arXiv. <https://arxiv.org/abs/2505.08136>
- Wang, G. (2025). Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering. *PLOS ONE*, 20(2), e0318519. <https://doi.org/10.1371/journal.pone.0318519>
- Wang, S., Sun, L., & Yu, Y. (2024). A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. *Scientific Reports*, 14, 17491.
- Wong, C.-G., Tong, G.-K., & Haw, S.-C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy*, 12(3), 97–125. <https://search.informit.org/doi/10.3316/informit.T2024102400007590180439661>
- Zheng, Y., Gan, W., Chen, Z., Zhou, P., & Fournier-Viger, P. (2024). *SeqRFM: Fast RFM analysis in sequence data*. arXiv. <https://arxiv.org/abs/2411.05317>