# Loyal Customer Segmentation Using RFM Model and K-Means Algorithm on E-Commerce Data

**Hersa Safitri[1], Haviluddin Haviluddin[1,*], and Anton Prafanto[1]**
[1]Department of Informatics, Faculty of Engineering, Universitas Mulawarman, Samarinda, East Kalimantan, Indonesia

**Corresponding author:** Haviluddin Haviluddin (e-mail: haviluddin@unmul.ac.id).

**ABSTRACT** Comprehensive customer segmentation in the e-commerce context is essential for supporting effective data-driven decision-making. This research aims to develop and validate a loyal customer segmentation model by integrating Recency, Frequency, and Monetary (RFM) analysis with the K-Means clustering algorithm to generate stable, well-separated, and managerially relevant clusters. The dataset analyzed comprises 392,732 transactions from 4,339 customers, obtained from a public e-commerce platform. The research workflow includes data preprocessing, RFM score computation, feature standardization, determination of the optimal number of clusters using the Elbow Method, and cluster evaluation using the Silhouette Coefficient and Davies–Bouldin Index (DBI). The experimental results indicate that a five-cluster configuration yields the best performance, achieving a Silhouette score of 0.6158 and a DBI of 0.7190. The stability of the five-cluster solution is confirmed over 30 random initializations (Silhouette = 0.6159 ± 0.0031; mean Adjusted Rand Index = 0.9892), and its practical value is demonstrated through a segment-level revenue contribution analysis and a comparison against six baseline methods. The Champions segment (comprising Top Champions and Champions) accounts for only 0.3% of the customer base yet contributes the highest total monetary value of US$190,808.54. In contrast, 94.5% of customers are classified as at risk or hibernating. These findings demonstrate that integrating RFM with K-Means, validated across multiple evaluation metrics, yields a reliable, measurable, and actionable customer segmentation framework. This approach effectively supports the development of customer retention strategies and the optimization of data-driven customer value.

**KEYWORDS** Customer Segmentation, Davies-Bouldin Index, K-Means, RFM

## I. INTRODUCTION

Customer retention is a central challenge in e-commerce: acquiring a new customer is considerably more expensive than retaining an existing one, while the rapid accumulation of transaction logs makes manual analysis of customer behavior infeasible [1]. Customer segmentation enables firms to allocate marketing resources according to customer value; however, manual segmentation based on ad-hoc thresholds is subjective, difficult to reproduce, and unable to capture the multidimensional nature of purchasing behavior. Clustering techniques have been applied across diverse domains [2], including image processing [3]. In e-commerce retail, the RFM model summarizes customer transactional behavior into three managerially meaningful dimensions and, when combined with clustering techniques, enables objective, data-driven identification of loyal customers, replacing subjective manual thresholds [4][5].

The K-means algorithm [6] is popular because it is quite efficient in handling large datasets and is capable of quickly partitioning data [7]. K-Means has been applied across a wide range of domains [8], and its scalability on large transaction datasets has been further enhanced through parallel optimizations [9], [10], [11], [12], [13], including data-driven decision-support applications [14], making the clustering results serve as a more representative basis for data-driven decision making. Nevertheless, applying K-Means to e-commerce customer data requires the number of clusters to be specified in advance and the resulting structure to be rigorously validated.

However, K-Means has limitations in determining the optimal number of clusters, is sensitive to centroid initialization, and is susceptible to outliers and noise, leading to biased segmentation [5], [15], [16], [17]. As a result, the validation cluster becomes crucial. The Silhouette Index is widely used to assess intra-cluster density and inter-cluster separation, while DBI serves as a complementary metric, with low values indicating more compact, clearly separated clusters [18], [19]. The development of silhouette

optimization has also been shown to improve accuracy on complex datasets, and DBI is effectively used in diverse contexts across domains [20], [21], [22], [23], [24], [25].

In a competitive business environment, customer segmentation becomes crucial to maintain retention and marketing efficiency [13]. RFM represents the customer value based on transaction behavior, and effectively differentiates intensity interactions and monetary contributions [20], [21], [23], [24], [26], [27], [28]. Integration of RFM and K-Means, with the determination of the cluster method, Elbow [29], as well as validation, Silhouette, and DBI, has been shown to improve segmentation quality by maintaining intra-cluster homogeneity and inter-cluster heterogeneity [26], [27], [30], [31].

Despite the popularity of RFM and K-Means integration, three specific gaps remain in prior studies. First, the number of clusters k is often determined using a single criterion, either the Elbow Method or the Silhouette Coefficient alone, which is prone to selection bias when the criteria disagree, as is common in highly skewed transaction data [4], [32]. Second, the sensitivity of K-Means to centroid initialization is rarely quantified [33], so the reported segmentations may not be reproducible. Third, validation is typically restricted to internal indices and does not demonstrate the business value of the resulting segments. To address these gaps, this research integrates RFM and K-Means with multi-metric validation (Elbow, Silhouette, and DBI), a multi-seed stability analysis, robustness checks using PCA [34] and the Local Outlier Factor (LOF) [35], and a business-oriented validation based on segment-level revenue contribution [13], [36], [37], [38]. This aims to identify the optimal customer loyalty segment structure and validate the data-driven retention support strategy. The main contributions of this research are threefold. First, we propose a multi-metric cluster validation framework that combines the Elbow Method, the Silhouette Coefficient [39], and the DBI [40], complemented by a 30-seed stability analysis, to reduce the subjectivity of selecting the number of clusters k in RFM-based segmentation. Second, we provide a business-interpretable five-segment customer typology (Top Champions, Champions, Need Attention, At Risk, and Lost/Hibernating) and validate it externally through revenue-contribution analysis, showing that 0.3% of customers generate 17.8% of total revenue. Third, we demonstrate the robustness of the proposed segmentation through PCA-space re-evaluation, LOF outlier screening, and a comparison against six baseline methods, including manual rule-based RFM quintile segmentation, in which K-Means++ achieves the best balance of internal validity and managerial actionability. This is structured as follows: the second section describes the methodology and experimental design; the third section presents the results and analysis; the fourth section discusses the implications of the findings; and the last section serves as a conclusion and a direction for further research.

## II. METHODS

### A. RESEARCH DESIGN

This research applies to a framework comprising several stages: data acquisition, data preprocessing, RFM feature extraction, K-Means clustering implementation, and cluster validation and interpretation. The research process begins with collecting customer transaction data, followed by data cleaning and normalization. Furthermore, RFM (Recency, Frequency, Monetary) values are calculated to represent customer characteristics. The next stage is to determine the optimal number of clusters, followed by implementing the K-Means algorithm. Finally, the clustering results are evaluated using internal validation metrics to assess cluster quality. The overall research flow is illustrated in Figure 1.

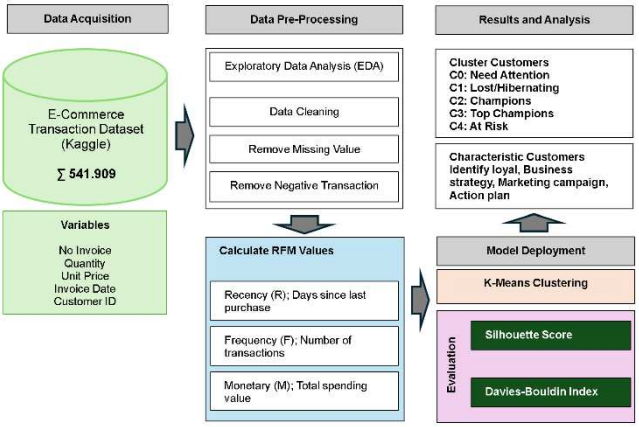


**Figure 1. Research Framework.**

### B. DATA DESCRIPTION

E-commerce transaction logs contain customer identities, transaction times, purchase frequency, and monetary values. The dataset was obtained from Kaggle and contains 541,909 online retail transaction records from a UK-based online retailer covering December 2010 to December 2011. Three filtering steps were applied sequentially (Table I). First, 135,080 records without a CustomerID (anonymous purchases) were removed, because customer-level RFM aggregation is impossible without an identifier, leaving 406,829 records. Second, 5,225 exact duplicate records were removed, leaving 401,604 records. Third, 8,872 records corresponding to cancelled invoices (InvoiceNo prefixed with "C") and non-positive quantities (returns) were removed, yielding a final dataset of 392,732 valid transactions from 4,339 unique customers. The three RFM features operate on very different scales: Recency ranges from 1 to 374 days (std = 100.01), Frequency from 1 to 210 invoices (std = 7.71), while Monetary ranges from US$0 to US$280,206.02 (std = 8,984.25) with extreme right skewness. Because K-Means assigns observations using Euclidean distance, the squared-difference term of the Monetary dimension would dominate the distance computation by several orders of magnitude

without scaling, effectively reducing the clustering to a one-dimensional partition of spending. Therefore, each feature was standardized using z-score normalization, $z = (x - \mu)/\sigma$, so that all three dimensions contribute equally to the distance metric and the resulting clusters reflect joint recency-frequency-monetary behavior rather than monetary value alone.

TABLE I.
DATA FILTERING STEPS

| Step | Operation | Rows removed | Rows remaining |
|---|---|---|---|
| 0 | Raw transactions (Dec 2010 to Dec 2011) | - | 541,909 |
| 1 | Remove records without CustomerID | 135,080 | 406,829 |
| 2 | Remove exact duplicate records | 5,225 | 401,604 |
| 3 | Remove cancelled invoices (C) and Quantity ≤ 0 | 8,872 | 392,732 |
| 4 | Final unique customers | | 4,339 |

### C. RECENCY, FREQUENCY, MONETARY (RFM)

The stage extraction feature converts the raw transaction data into three main variables.

- Recency (R): the number of days between the analysis reference date T, set to one day after the most recent transaction in the dataset (December 10, 2011), and the date of customer i's most recent transaction, $t_i^{last}$, so that $R_i \geq 1$ is expressed in days, as formalized in (1).

$$R_i = T - t_i^{last} \quad (1)$$

$$F_i = |I_i| \quad (2)$$

- Frequency (F): the number of distinct invoices (purchase occasions) of customer i during the observation period, formally $F_i = |I_i|$, where $I_i$ denotes the set of unique invoices of customer i; this counts orders rather than individual item lines, as formalized in (2).
- Monetary (M): the total spending of customer i during the observation period, computed as the sum over all the customer's transaction lines j of the purchased quantity $q_{ij}$ multiplied by the unit price $p_{ij}$ [4], as formalized in (3).

$$M_i = \Sigma\,(j \in T_i)\,(q_{ij} \times p_{ij}) \quad (3)$$

where $T_i$ is the set of transaction rows of customer i, $q_{ij}$ the quantity, and $p_{ij}$ the unit price of row j. RFM analysis is widely used in customer segmentation to identify high-value, loyal customers based on transactional behavior, including recency, frequency, and monetary value.

### D. K-MEANS ALGORITHM

The standard iterative algorithm, K-Means, is often referred to as Lloyd's Algorithm and is very popular in Machine Learning. This algorithm was introduced in 1957 by Stuart Lloyd, an engineer at Bell Labs. This was widely published in 1982. The working stages of K-Means consist of Stage 1, which involves determining the number of clusters (k). Stage 2 is initialization of the centroid. Stage 3 is data grouping to calculate the distance of each data point to all existing centroids. Stage 4 is updating the centroid (update step). Stage 5 is convergence (iteration); repeat stages 3 and 4. The process stops when the centroid no longer changes. Again, change position, centroid displacement is very small below the threshold/tolerance, and the specified maximum number of iterations has been reached.

### E. CLUSTER VALIDATION AND EVALUATION

This research employs internal validation indices to ensure that the resulting segmentation is not arbitrary. The Silhouette score measures the balance between intra-cluster cohesion and inter-cluster separation. The silhouette score combines both cohesion and separation to evaluate clustering quality in a single metric. Meanwhile, the Davies–Bouldin Index (DBI) evaluates the ratio between within-cluster dispersion and the distance between clusters, where lower values indicate more compact and well-separated clusters. The evaluation is conducted across multiple values of k, with the optimal number of clusters determined based on the consistency of both metrics.

Furthermore, business validation is performed by analyzing the distribution of RFM values within each cluster to ensure clear heterogeneity between segments, such as distinguishing loyal, potential, and at-risk customers. This approach confirms that the resulting segmentation is not only mathematically valid but also meaningful and applicable for customer retention strategies.

### F. ANALYSIS, SEGMENTATION, AND RECOMMENDATION

The final stage focuses on analyzing the characteristics of each cluster formed. Each segment is interpreted based on the profile RFM and classified into categories such as loyal, potential, at-risk, or inactive customers. The results of this segmentation then become the basis for the formulation strategy, the business adjustment campaign, and a follow-up plan for decision-makers.

## III. RESULTS AND DISCUSSION

### A. PREPROCESSING AND RFM MODEL FORMATION

The dataset contains 392,732 valid transactions representing 4,339 unique customers. The RFM model was calculated with a reference date of December 10, 2011, and a summary of its descriptive statistics is presented in Table II. The high variation in Frequency (std = 7.71 against a mean of only 4.27 invoices) and Monetary (std = US$8,984.25 against

a mean of US$2,048.22) reflects strong heterogeneity and right-skewness of customer behaviour.

### B. DETERMINING THE OPTIMAL CLUSTER AMOUNT

Determining the optimal number of clusters is a key step in K-Means. RFM data is first standardized with StandardScaler, and then the number of clusters is determined through a multi-metric evaluation integrating the Elbow Method (inertia), Davies–Bouldin Index (DBI), and Silhouette Score to assess each k value (k = 2–10). The evaluation results are presented in Table III and Figure 2.

TABLE II.
STATISTICS DESCRIPTIVE VARIABLES RFM

| Statistics | Recency | Frequency | Monetary |
|---|---|---|---|
| Mean | 92.52 | 4.27 | 2,048.22 |
| Std | 100.01 | 7.71 | 8,984.25 |
| Min | 1 | 1 | 0.00 |
| Max | 374 | 210 | 280,206.02 |

TABLE III.
EVALUATION MULTI-METRIC FOR DETERMINATION OF CLUSTER OPTIMAL

| k | Elbow (inertia) | DBI | Silhouette |
|---|---|---|---|
| 2 | 9,016.15 | 0.7452 | 0.8959 |
| 3 | 5,442.38 | 0.7105 | 0.5944 |
| 4 | 4,097.64 | 0.7532 | 0.6161 |
| 5 | 3,122.37 | 0.7190 | 0.6158 |
| 6 | 2,473.54 | 0.6273 | 0.5983 |
| 7 | 2,023.40 | 0.6339 | 0.5162 |
| 8 | 1,716.12 | 0.6954 | 0.4768 |
| 9 | 1,446.94 | 0.7107 | 0.4777 |
| 10 | 1,284.02 | 0.6710 | 0.4779 |

Table III shows the existence of a trade-off between indicators, where the highest Silhouette Score appears at k = 2, the lowest (best) DBI value at k = 6, and the Elbow Method indicates k = 5 as the most stable point. Although k = 2 attains the highest Silhouette score (0.8959), inspection of the cluster sizes (Table IV) reveals a degenerate solution: 4,313 customers (99.4%) fall into a single cluster while only 26 customers form the second cluster, so the high score is an artifact of one massive, internally cohesive group rather than evidence of meaningful structure. Conversely, k = 6 minimizes the DBI (0.6273) but fragments the high-value tail into micro-clusters of only 2 and 4 customers (≤ 0.1% of the population), which are statistically unstable and cannot serve as actionable campaign targets, while its Silhouette score drops to 0.5983. The k = 5 solution offers the best compromise supported by four quantitative criteria: its Silhouette score (0.6158) is within 0.0003 of the highest non-degenerate value across all k, its DBI (0.7190) remains below 1, it coincides with the elbow point of the inertia curve, and its smallest cluster still contains six customers, large enough to be operationally targetable. Moreover, the resulting segments exhibit a monotonic business ordering, with mean Recency increasing as mean Monetary decreasing, confirming managerial interpretability, and the multi-seed stability analysis reported below shows that k = 5 yields the lowest variance of the Silhouette score among the candidate solutions.

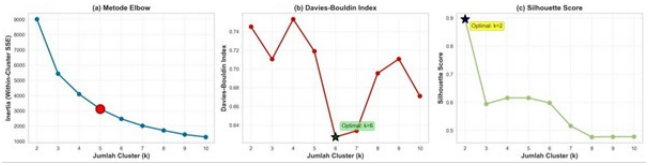


**Figure 2. Multi-Metric Visualization (Elbow, DBI, and Silhouette) Scores.**

Figure 2 presents three evaluation graphs in parallel. The Elbow (inertia) graph shows a slowing down of SSE at k = 5 (3,122.37) as the elbow point, the DBI graph reaches its minimum value at k = 6 (0.6273), while the Silhouette graph shows the score recorded a maximum value at k = 2 (0.8959) with a decreasing trend as k increased.

As additional validation, clustering was evaluated using Principal Component Analysis (PCA) space, which retains 95% of the variance in the RFM data and reduces the dimensionality from three to two. The evaluation silhouette in the PCA space shows that k = 4 is optimal, with a score of 0.6342. Outlier detection using Local Outlier Factor (LOF, n_neighbors = 20) identified 4,296 inliers (99.0%) and 43 outliers (1.0%), with Silhouette Score inlier of 0.6180. This confirms that the clustering results are robust to the presence of outliers.

Figure 3 shows the clustering results in two-dimensional PCA space, with each cluster represented by a different color, making it easier to assess the degree of separation between clusters in the reduced-dimensional space.

TABLE IV.
CLUSTER-SIZE DIAGNOSTICS FOR CANDIDATE NUMBERS OF CLUSTERS

| k | Smallest cluster | Largest cluster (%) | Diagnosis |
|---|---|---|---|
| 2 | 26 (0.6%) | 99.4 | Degenerate |
| 5 | 6 (0.14%) | 70.0 | Balanced and actionable |
| 6 | 2 (0.05%) | 67.6 | Micro-clusters, unstable |

### C. CLUSTERING STABILITY ANALYSIS

Because K-Means is sensitive to centroid initialization, the clustering was repeated over 30 different random seeds (each with n_init = 10) for the candidate solutions k = 2, 5, and 6. Table V reports the mean and standard deviation of the Silhouette score and DBI, together with the Adjusted Rand Index (ARI) of each run against the reference partition (random_state = 42) [41]. The k = 5 solution is highly stable: the Silhouette score varies by only ±0.0031 and the DBI by ±0.0042 across seeds, while the partitions themselves are nearly identical (mean ARI = 0.9892; minimum 0.9248). In contrast, k = 2 exhibits the largest partition variability (minimum ARI = 0.6968), indicating that its apparent quality depends on initialization. These results confirm that the reported five-segment structure is reproducible and not an artifact of a particular initialization.

TABLE V.
CLUSTERING STABILITY OVER 30 RANDOM SEEDS

| k | Silhouette (mean ± std) | DBI (mean ± std) | ARI (mean) | ARI (min) |
|---|---|---|---|---|
| 2 | 0.9001 ± 0.0096 | 0.7344 ± 0.0242 | 0.9495 | 0.6968 |
| 5 | 0.6159 ± 0.0031 | 0.7159 ± 0.0042 | 0.9892 | 0.9248 |
| 6 | 0.5958 ± 0.0042 | 0.6381 ± 0.0213 | 0.9906 | 0.9619 |

### D. BASELINE COMPARISON

To position the proposed approach, six baselines were evaluated on the identical standardized RFM features with the same number of clusters (k = 5): K-Means with random initialization, Ward agglomerative clustering [42], BIRCH [43], a Gaussian Mixture Model (GMM) [44], MiniBatch K-Means [45], and a manual rule-based segmentation that assigns customers to five tiers by RFM quintile scores [46]. As shown in Table VI, K-Means++ achieves the best non-degenerate internal validity (Silhouette = 0.6158; DBI = 0.7190; Calinski–Harabasz = 3,433.6). BIRCH appears superior on Silhouette and DBI, but its solution is degenerate, placing 99.2% of customers in a single cluster, the same pathology observed for k = 2, which confirms that internal indices must be read jointly with the cluster-size distribution. The rule-based quintile segmentation performs worst (Silhouette = −0.0217), quantitatively confirming the weakness of manual threshold-based segmentation that motivates this research [32].

TABLE VI.
BASELINE COMPARISON AT K = 5

| Method | Silhouette | DBI | CH | Largest cluster (%) |
|---|---|---|---|---|
| K-Means++ (proposed) | 0.6158 | 0.7190 | 3,433.6 | 70.0 |
| K-Means (random init) | 0.5018 | 0.7351 | 2,796.3 | 62.1 |
| Agglomerative (Ward) | 0.5722 | 0.8101 | 3,199.2 | 64.6 |
| BIRCH | 0.8010* | 0.6148* | 926.4 | 99.2 (degenerate) |
| Gaussian Mixture | 0.0550 | 1.9182 | 787.0 | 32.9 |
| MiniBatch K-Means | 0.3358 | 0.8936 | 1,066.8 | 42.7 |
| Rule-based RFM quintile | −0.0217 | 1.8275 | 503.4 | 27.2 |

*The apparently superior BIRCH scores arise from a degenerate one-large-cluster solution and therefore do not indicate better segmentation quality.

These outcomes are directly explained by the distribution of the RFM variables. Monetary is extremely right-skewed (skewness = 19.34; mean US$2,048 vs. maximum US$280,206) and Frequency is similarly skewed (12.10), so the customer base consists of a very small high-value head and a long passive tail: the top 1% of customers account for 31.8% of total revenue and the top 20% for 74.7%, consistent with the Pareto principle. This skewness explains why the At Risk and Lost/Hibernating segments dominate the population (94.5% of customers) while contributing only 50.8% of revenue, and why effective retention strategy must protect the disproportionately valuable Champions segments while re-engaging the large At Risk segment that is still within the re-engagement window (mean Recency = 44 days).

### E. CUSTOMER RESULT SEGMENTATION

K-Means model with k = 5 (initialization K-Means ++, random_state = 42, n_init = 10) implemented on data RFM standardized, and produces five segments of customers, with the distribution presented in Table VII.

The Top Champions segment (Cluster 3) includes 6 customers (0.1%) with very low recency (7.67 days), high frequency (43 transactions), and the highest monetary value of US$190,808.54, representing the most valuable and premium customers overall. The Champions segment (Cluster 2) consists of 8 customers (0.2%) with very low recency (6.5 days), the highest frequency among all segments (120.62 orders), and a monetary value of US$55,099.49, reflecting the most active transaction intensity. Combined, these two segments account for only 0.3% of customers but contribute significantly to total revenue. The Need Attention Segment (Cluster 0) accounts for 5.2% of customers, with relatively low recency (15 days), moderate frequency (±21 transactions), and high monetary value (±US$12K), indicating great potential but requiring further management. Meanwhile, the At-Risk segment (Cluster 4), with 3,038 customers (70%), has a high recency (44.13 days), low frequency, and moderate monetary value, so it is at risk of churn but can still be re-engaged. Meanwhile, the Lost/Hibernating segment (Cluster 1), with 1,061 customers (24.5%), has very high recency (248.66 days) and very low frequency (1.55 transactions), indicating customers who are close to being lost and require an aggressive win-back strategy. Combined, these two segments account for 94.5% of customers. High recency (44–249 days), low frequency (1–4 transactions), and low to medium monetary value (US$476–US$1.3K), thus requiring intensive re-engagement efforts.

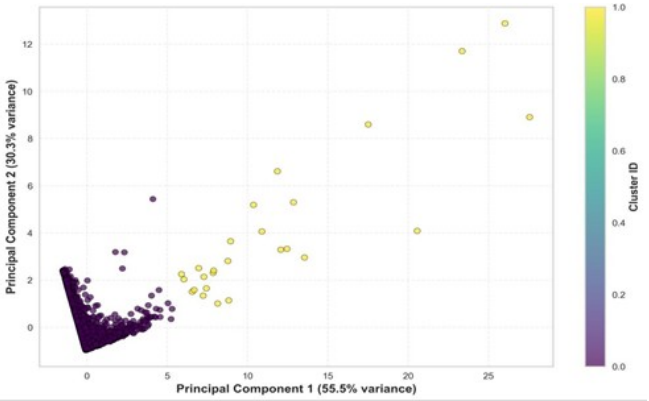


**Figure 3. Clustering Visualization in 2D PCA Projection.**

### F. CLUSTERING RESULT VALIDATION

Outlier detection using LOF (n_neighbors = 20) identified 43 customers (1.0%) out of a total of 4,339. The Silhouette

TABLE VII.
DISTRIBUTION AND CHARACTERISTICS CUSTOMER SEGMENT

| Cluster | Segment | Amount | % | Recency | Frequency | Monetary |
|---|---|---|---|---|---|---|
| 0 | Need Attention | 226 | 5.2 | 15.24 | 20.82 | 12,349.05 |
| 1 | Lost / Hibernating | 1,061 | 24.5 | 248.66 | 1.55 | 476.42 |
| 2 | Champions | 8 | 0.2 | 6.5 | 120.62 | 55,099.49 |
| 3 | Top Champions | 6 | 0.1 | 7.67 | 43 | 190,808.54 |
| 4 | At Risk | 3,038 | 70 | 44.13 | 3.61 | 1,318.37 |

Score on the inlier data (0.6180) is consistent with the overall score (0.6158), confirming the robustness of the clustering results. In addition, the PCA pipeline with a 95% variance threshold produces an optimal k = 4 cluster with a Silhouette Score of 0.6342, which is relatively consistent with the findings in the original RFM space (k = 5).

## G. CLUSTER PROFILE

Based on the RFM profile in Table VII, each cluster is interpreted and labeled accordingly. Behavior Cluster 3 (Top Champions) customers have the highest monetary value (US$190,808.54) and very low recency (7.67 days). Cluster 2 (Champions) has the highest transaction frequency (120.62 orders), very low recency (6.5 days), and a monetary value of US$55,099.49. The strategy for both focuses on exclusive loyalty, personalized service, special offers, and upselling and cross-selling, with more premium treatment for Top Champions.

Cluster 0 (Need Attention) includes 226 customers (5.2%) with low recency, relatively high monetary value, and moderate frequency, so they have the potential to become Champions if managed properly. Suggested approaches include personalized communications, frequency-increasing incentives, targeted offers, customer feedback, and reminder campaigns.

Cluster 4 (At Risk) is the largest segment, comprising 3,038 customers (70%), with a relatively high recency (44.13 days) and low frequency (3.61 transactions). Unlike Lost / Hibernating, this segment is still within the re-engagement range.

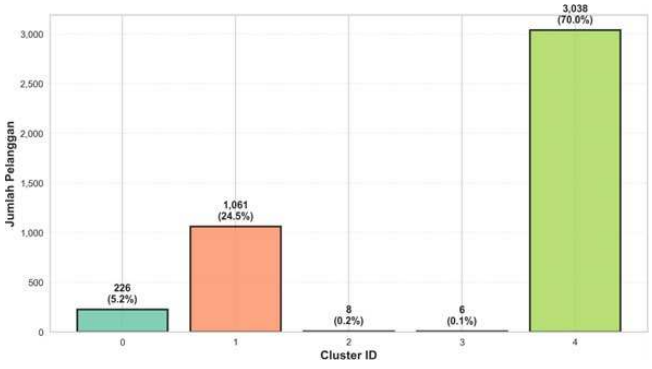


**Figure 4. Bar Chart of Customer Distribution per Cluster.**

Cluster 1 (Lost / Hibernating) consists of 1,061 customers (24.5%) with very high recency (248.66 days), very low frequency (1.55 transactions), and low monetary value (US$476.42), indicating customers who are almost lost or long-term inactive.

Figure 4 presents the distribution of customers as a bar chart, highlighting the size differences between clusters. Cluster 4 dominates, accounting for around 70% of the total customers.

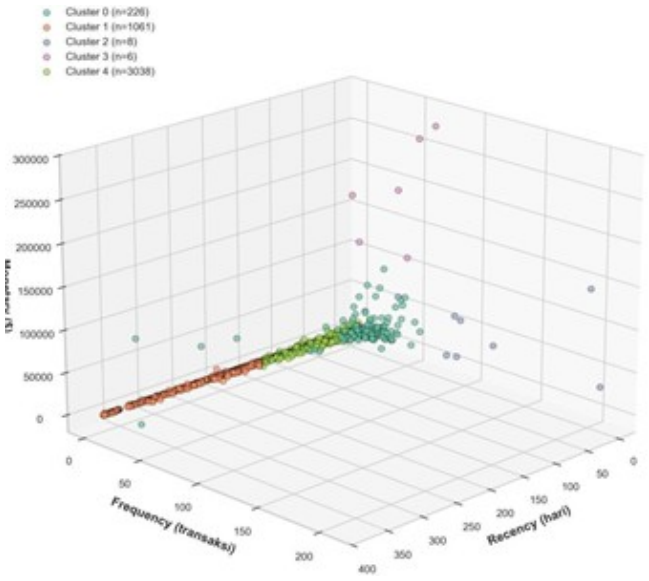


**Figure 5. 3D Visualization of RFM Segmentation Based on Clusters.**

Figure 5 displays a three-dimensional scatter plot visualization on the Recency, Frequency, and Monetary axes, with each cluster distinguished by color. This display clarifies the cluster's position in the RFM space and confirms the separation between the Champions segments and other segments.

## H. BUSINESS-ORIENTED VALIDATION

Beyond internal indices, the practical value of the segmentation was assessed through a revenue-contribution analysis (Table VIII), which evaluates the segments on an outcome dimension, total segment revenue share, rather than on the clustering criteria themselves. The five segments separate customer value sharply and monotonically: the two Champions segments jointly contain 0.32% of customers yet generate 17.84% of total revenue (US$1.59M), with average per-customer spending 93.2× (Top Champions) and 26.9× (Champions) the population mean; the Need Attention segment (5.2% of customers) contributes a further 31.4% of

revenue; and the two passive segments, although covering 94.5% of customers, contribute the remaining 50.8%. The segmentation also reproduces the empirical Pareto structure of the data, in which the top 20% of customers account for 74.7% of revenue, without this information being supplied to the algorithm. As an additional external reference, the manual rule-based RFM quintile segmentation in Table VI exhibits poor internal validity (Silhouette = −0.0217), demonstrating that the proposed data-driven segmentation captures value structure that threshold-based manual segmentation cannot.

TABLE VIII.
SEGMENT-LEVEL REVENUE CONTRIBUTION

| Cluster | Segment | Customers (%) | Revenue share (%) | Avg. Monetary vs mean |
|---|---|---|---|---|
| 3 | Top Champions | 0.14 | 12.88 | 93.2× |
| 2 | Champions | 0.18 | 4.96 | 26.9× |
| 0 | Need Attention | 5.21 | 31.40 | 6.0× |
| 4 | At Risk | 70.02 | 45.07 | 0.64× |
| 1 | Lost / Hibernating | 24.45 | 5.69 | 0.23× |

### I. INTERPRETATION OF CUSTOMER LOYALTY SEGMENTATION RESULT

The segmentation results show that loyal customers (Champions) account for only 0.3% of the population. However, they have a market-monetary average that is 15–40 times higher than that of other segments, confirming the highly unequal distribution of customer value in line with the Pareto principle. Need Attention Segment (5.2%) has strong potential to develop into Champions. Due to low recency and high monetary value, although the frequency still needs to be increased, it is a priority for loyalty programs. Meanwhile, most customers (94.5%) are in the At Risk and Lost/Hibernating segments, reflecting serious retention challenges, especially for Cluster 1, with a recency of 249 days and a frequency of 1.55, indicating customers are nearly lost and require a win-back strategy.

### J. MARKETING STRATEGY IMPLICATIONS

The Champions segment requires VIP treatment through early access products, appreciation programs, and upselling premium products. Need Attention Segment requires personalized offers, incentives to increase transaction frequency, and loyalty programs. Meanwhile, the At Risk / Hibernating segment needs to focus on campaign re-engagement, such as special discounts, win-back emails, and customer surveys, to identify the causes of decreased activity. Although they account for only 0.3% of customers, retention Champions is crucial because of their large revenue contribution. In parallel, converting Need Attention into Champions has the potential to significantly increase customers' lifetime value.

### K. VALIDATION METHODOLOGY

Multi-metric approaches (Elbow, DBI, and Silhouette) ensure a more objective selection of k. Different results show DBI is optimal at k = 6 and Silhouette at k = 2, which align with business interpretability; thus, k = 5 is selected based on cluster-size diagnostics (Table IV), the 30-seed stability analysis (Table V), and managerial interpretability, rather than as a mere compromise. K-Means was applied to the original RFM space without PCA to keep the interpretation directly relevant to marketing strategy, although PCA yields a slightly higher Silhouette score (0.6342 vs 0.6158), RFM segmentation is more communicative for stakeholders and more applicable in campaign management.

### L. RESEARCH CONTRIBUTION

This research demonstrates that integrating multiple validation indices yields a more robust and objective k selection than using a single metric alone. RFM segmentation effectively identifies high-value customers in an e-commerce context with a highly skewed value distribution, while generating immediately actionable segments for targeted retention and acquisition strategies.

Findings also confirm that different recommendations k from DBI and Silhouette can be aligned via the Elbow Method to obtain a stronger consensus. Practically, this research offers an applicable framework for e-commerce practitioners, with clear segmentation (Champions, Need Attention, At Risk, Lost/Hibernating) and specific, implementable strategy recommendations.

Besides that, this research emphasizes the importance of balancing technical sophistication and business relevance, where the RFM approach is more communicative for stakeholders and easier to integrate into existing marketing systems.

### M. LIMITATIONS AND FUTURE RESEARCH

This research has several limitations. The resulting segmentation is static because it is based on historical data, so real-world implementation requires dynamic segmentation updates. The RFM model relies on only three key metrics, so it doesn't capture product preferences, channels, or customers' seasonal patterns. Furthermore, segment characteristics may differ across e-commerce verticals. The business-oriented validation in this research is based on observed revenue concentration; direct causal validation of segment-targeted interventions, such as A/B testing of retention campaigns per segment or longitudinal tracking of customer lifetime value, remains future work, as it requires experimental data beyond the public transaction log.

In the future, research can be developed through the application of dynamic segmentation, expansion model RFM with dimensions addition (for example, duration, engagement, or loyalty), as well as integration with model predictive for projecting churn and customer lifetime validation based on

A/B testing and multi-channel analysis also has the potential to strengthen findings by linking segmentation directly to actual business outcomes.

## IV. CONCLUSION

This research successfully applies segmentation, RFM-based customer loyalty, and K-Means to 4,339 e-commerce customers. Multi-metric evaluation using Elbow, Silhouette, and Davies–Bouldin Index identified five optimal segments, with Silhouette Score 0.6158 and DBI 0.7190. The five-cluster solution is stable across 30 random initializations (mean ARI = 0.9892), outperforms six baseline methods in non-degenerate internal validity, and is externally supported by a segment-level revenue contribution analysis. The results identified the Top Champions and Champions segments at 0.3% with the highest loyalty level, the Need Attention segment at 5.2%, the At-Risk segment at 70%, and the Lost/Hibernating segment at 24.5%. Each segment has a distinct RFM profile and requires a specific marketing strategy, so this integrated approach enables segmentation, which is applicable for increasing customer retention and customer lifetime value in e-commerce. The application of metaheuristic algorithms to obtain accurate results requires further research.

## AUTHOR CONTRIBUTION

**Hersa Safitri:** Conceptualization, Methodf ology, software, Data Cuartions, Formal Analysis, Investigation, Visualization, Writing, Original Draft Preparation,
**Haviluddin Haviluddin:** Conceptualization, Methodology, Validation, Supervison, Project Administration, Writing, Review & Editing
**Anton Prafanto:** Review & Validation

## COPYRIGHT



## REFERENCES

[1] F. F. Reichheld and W. E. Sasser, "Zero defections: Quality comes to services," *Harvard Bus. Rev.*, vol. 68, no. 5, pp. 105-111, 1990.
[2] H. Yue, H. Zhang, and Y. Dai, "Application of PSO-integrated K-means algorithm in resident digital portrait classification," *PLOS One*, vol. 20, no. 8, p. e0329123, Aug. 2025, doi: 10.1371/journal.pone.0329123.
[3] P. Pickerling, H. Armanto, and S. K. Bastari, "Multilevel Image Thresholding Memanfaatkan Firefly Algorithm, Improved Bat Algorithm, dan Symbiotic Organisms Search," *J. Intell. Syst. Comput.*, vol. 1, no. 1, pp. 1-8, Aug. 2019, doi: 10.52985/insyst.v1i1.24.
[4] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Mark. Cust. Strategy Manag.*, vol. 19, no. 3, pp. 197-208, 2012, doi: 10.1057/dbm.2012.17.
[5] S. I. Murpratiwi, I. G. Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *J. Pendidik. Teknol. Dan Kejuru.*, vol. 18, no. 2, p. 152, Sep. 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.
[6] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Berkeley, CA, USA, 1967, pp. 281-297.
[7] H. Li and W. Zhu, "Artistic Color Matching Technology Based on Silhouette Coefficient and Visual Perception," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 6, 2024, doi: 10.14569/IJACSA.2024.0150642.
[8] L. Yang et al., "Assessing nursing students' palliative care training needs and profiles: a cross-sectional study using K-means clustering," *BMC Nurs.*, vol. 24, no. 1, p. 1235, Oct. 2025, doi: 10.1186/s12912-025-03797-0.
[9] S. M. Hadi, A. H. Alsaeedi, M. I. Dohan, R. R. Nuiaa, S. Manickam, and A. S. D. Alfoudi, "Dynamic Evolving Cauchy Possibilistic Clustering Based on the Self-Similarity Principle (DECS) for Enhancing Intrusion Detection System," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 5, pp. 252-260, Oct. 2022, doi: 10.22266/ijies2022.1031.23.
[10] R. Nuiaa, S. Manickam, A. H. Alsaeedi, and D. E. J. Al-Shammary, "Evolving Dynamic Fuzzy Clustering (EDFC) to Enhance DRDoS_DNS Attacks Detection Mechnism," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 1, Feb. 2022, doi: 10.22266/ijies2022.0228.46.
[11] N. N. Cahayana, U. S. Aesyi, and K. Kharisma, "Model Analisis Sentimen pada Ulasan Pengguna Mobile Banking Menggunakan Kombinasi K-Means dan Naive Bayes," *Angkasa J. Ilm. Bid. Teknol.*, vol. 17, no. 1, p. 44, May 2025, doi: 10.28989/angkasa.v17i1.2500.
[12] J. Jin, "Student behavior patterns in vocational education big data based on clustering algorithm," *Discov. Artif. Intell.*, vol. 5, no. 1, p. 197, Aug. 2025, doi: 10.1007/s44163-025-00433-3.
[13] S. Zhang, S. Chen, X. Yu, and S. Mei, "Research on collaborative filtering algorithm based on improved K-means algorithm for user attribute rating and co-rating," *Sci. Rep.*, vol. 15, no. 1, p. 19600, Jun. 2025, doi: 10.1038/s41598-025-96705-0.
[14] Aidil, J. P. Sugiono, E. I. Setiawan, and A. S. Putra, "Pembentukan Aturan Fuzzy Untuk Pemberian Rekomendasi Penerima Bantuan Keluarga Berumah Tidak Layak Huni Menggunakan K-means Clustering," *J. Intell. Syst. Comput.*, vol. 4, no. 2, pp. 85-92, Oct. 2022, doi: 10.52985/insyst.v4i2.216.
[15] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *J. Appl. Inform. Comput.*, vol. 7, no. 1, pp. 89-97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.
[16] L. Yuan and Y. Chen, "Automatic modulation classification method using fixed K-Means algorithm for feature clustering processing," *PLOS One*, vol. 20, no. 10, p. e0333098, Oct. 2025, doi: 10.1371/journal.pone.0333098.
[17] L. E. Ekemeyong Awong and T. Zielinska, "Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification," *Sensors*, vol. 23, no. 18, p. 7925, Sep. 2023, doi: 10.3390/s23187925.
[18] S. Ramadhani, D. Azzahra, and T. Z, "Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis," *Digit. Zone J. Teknol. Inf. Dan Komun.*, vol. 13, no. 1, pp. 24-33, May 2022, doi: 10.31849/digitalzone.v13i1.9292.
[19] R. Ying, T. Taniguchi, K. Nabeta, K. Ishigami, N. Kikuchi, and S. Okamoto, "Clustering First-order reversal curve diagram using the Gaussian mixture model and the Davies-Bouldin index," *Jpn. J. Appl. Phys.*, vol. 63, no. 11, p. 11SP04, Nov. 2024, doi: 10.35848/1347-4065/ad892d.
[20] M. S. Abidin, Y. Kustiyahningsih, E. Rahmanita, B. D. Satoto, and M. I. Firmansyah, "Application of the DBSCAN Algorithm in MSME Clustering using the Silhouette Coefficient Method," *J. Sist. Cerdas*, vol. 7, no. 3, pp. 260-270, Dec. 2024, doi: 10.37396/jsc.v7i3.472.
[21] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal

evaluation of two convex clusters," *PeerJ Comput. Sci.*, vol. 11, p. e3309, Nov. 2025, doi: 10.7717/peerj-cs.3309.

[22] T. A. Mulyadi and D. Purnomo, "Optimasi Pelayanan Kapal Penumpang melalui Clustering Penumpang dengan Metode Silhouette Coefficient," *Edumatic J. Pendidik. Inform.*, vol. 7, no. 2, pp. 217-226, Dec. 2023, doi: 10.29408/edumatic.v7i2.21067.

[23] N. T. M. Sagala and A. A. S. Gunawan, "Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 1-10, Feb. 2022, doi: 10.21512/comtech.v13i1.7270.

[24] I. T. Utami, F. Suryaningrum, and D. Ispriyanti, "K-Means Cluster Count Optimization with Silhouette Index Validation and Davies Bouldin Index (Case Study: Coverage of Pregnant Women, Childbirth, and Postpartum Health Services In Indonesia In 2020)," *Barekeng J. Ilmu Mat. Dan Terap.*, vol. 17, no. 2, pp. 0707-0716, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0707-0716.

[25] P. Bombina, D. Tally, Z. B. Abrams, and K. R. Coombes "SillyPutty Improved clustering by optimizing the silhouette width," *PLOS ONE*, vol. 19, no. 6, p. e0300358, Jun. 2024, doi: 10.1371/journal.pone.0300358.

[26] S. E. A. Buananta, M. A. Ahmad, J. Mahmood, and P. Paradise, "Identification of Evaluation Results in E-Banking Services Transaction for Product Recommendation using the BIRCH and Davies Bouldin Index Method," *J. INFOTEL*, vol. 16, no. 2, pp. 427-440, May 2024, doi: 10.20895/infotel.v16i2.1116.

[27] A. Idrus, N. Tarihoran, U. Supriatna, A. Tohir, S. Suwarni, and R. Rahim, "Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm," *TEM J.*, pp. 1871-1876, Nov. 2022, doi: 10.18421/TEM114-55.

[28] H. Hairani, D. Susilowati, I. Puji Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK J. Manaj. Tek. Inform. Dan Rekayasa Komput.*, vol. 21, no. 2, pp. 275-282, Mar. 2022, doi: 10.30812/matrik.v21i2.1542.

[29] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267-276, 1953, doi: 10.1007/BF02289263.

[30] Y. Asriningtias and J. Aryanto, "K-Means Algorithm with Davies Bouldin Criteria for Clustering Provinces in Indonesia Based on Number of Events and Impacts of Natural Disasters," *Int. J. Eng. Technol. Nat. Sci.*, vol. 4, no. 1, pp. 75-80, Jul. 2022, doi: 10.46923/ijets.v4i1.147.

[31] P. Prihandoko, D. Jollyta, G. Gusrianty, M. Siddik, and J. Johan, "Cluster Validity for Optimizing Classification Model: Davies Bouldin Index - Random Forest Algorithm," *MATRIK J. Manaj. Tek. Inform. Dan Rekayasa Komput.*, vol. 24, no. 1, pp. 61-72, Nov. 2024, doi: 10.30812/matrik.v24i1.4043.

[32] S. Acar, F. Köroğlu, B. Duyuler, T. Kaya, and T. Özcan, "Customer segmentation using RFM model and clustering methods in online retail industry," in *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation (INFUS 2021)*, Lecture Notes in Networks and Systems, vol. 307. Cham, Switzerland: Springer, 2022, pp. 69-77, doi: 10.1007/978-3-030-85626-7_9.

[33] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA '07)*, Philadelphia, PA, USA, 2007, pp. 1027-1035.

[34] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, art. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[35] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. 2000 ACM SIGMOD Int. Conf. Management of Data*, Dallas, TX, USA, 2000, pp. 93-104, doi: 10.1145/342009.335388.

[36] T. Akbar, G. M. Tinungki, and S. Siswanto, "Performance Comparison of K-Medoids and Density Based Spatial Clustering of Application with Noise Using Silhouette Coefficient Test," *BAREKENG J. Ilmu Mat. Dan Terap.*, vol. 17, no. 3, pp. 1605-1616, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1605-1616.

[37] L. Guozhang, R. Alfred, R. Pailus, X. Fengchang, and H. Haviluddin, "Scalable and Efficient Student Behavior Prediction using Parallelized Clustering and AHP-weighted KNN," *J. ICT Res. Appl.*, vol. 19, no. 2, pp. 142-165, Dec. 2025, doi: 10.5614/itbj.ict.res.appl.2025.19.2.3.

[38] A. N. Fauzan, F. A. Wandi, A. Z. N. Aiman, and M. Wati, "Pengelompokan Minat Akademik Siswa SMA Negeri 1 Loa Janan Menggunakan Metode Clustering K-Means," 2025.

[39] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53-65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[40] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224-227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[41] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193-218, 1985, doi: 10.1007/BF01908075.

[42] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 236-244, 1963, doi: 10.1080/01621459.1963.10500845.

[43] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, no. 2, pp. 103-114, 1996, doi: 10.1145/233269.233324.

[44] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Boston, MA, USA: Springer, 2009, pp. 659-663, doi: 10.1007/978-0-387-73003-5_196.

[45] D. Sculley, "Web-scale k-means clustering," in *Proc. 19th Int. Conf. World Wide Web (WWW '10)*, Raleigh, NC, USA, 2010, pp. 1177-1178, doi: 10.1145/1772690.1772862.

[46] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.*, vol. 3, no. 1, pp. 1-27, 1974, doi: 10.1080/03610927408827101.