information

MDPI

Article

Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms

Chukwutem Pinic Ufeli [1], Mian Usman Sattar [1], Raza Hasan [2,*] and Salman Mahmood [3]

[1] College of Science and Engineering, University of Derby, Kedleston Road, Derby DE22 1GB, UK; chitemufeli@gmail.com (C.P.U.); u.sattar@derby.ac.uk (M.U.S.)
[2] Department of Science and Engineering, Solent University, Southampton SO14 0YN, UK
[3] Department of Computer Science, Nazeer Hussain University, ST-2, Near Karimabad, Karachi 75950, Pakistan; salman.mahmood@nhu.edu.pk
* Correspondence: raza.hasan@solent.ac.uk

Abstract: In today's data-driven business landscape, effective customer segmentation is crucial for enhancing engagement, loyalty, and profitability. Traditional clustering methods often struggle with datasets containing both numerical and categorical variables, leading to suboptimal segmentation. This study addresses this limitation by introducing a novel application of Factor Analysis of Mixed Data (FAMD) for dimensionality reduction, integrated with K-means and Agglomerative Clustering for robust customer segmentation. While FAMD is not new in data analytics, its potential in customer segmentation has been underexplored. This research bridges that gap by demonstrating how FAMD can harmonize mixed data types, preserving structural relationships that conventional methods overlook. The proposed methodology was tested on a Kaggle-sourced retail dataset comprising 3900 customers, with preprocessing steps including correlation ratio filtering ($\eta \geq 0.03$), standardization, and encoding. FAMD reduced the feature space to three principal components, capturing 81.46% of the variance, which facilitated clearer segmentation. Comparative clustering analysis showed that Agglomerative Clustering (Silhouette Score: 0.52) outperformed K-means (0.51) at $k = 4$, revealing distinct customer segments such as seasonal shoppers and high spenders. Practical implications include the development of targeted marketing strategies, validated through heatmap visualizations and cluster profiling. This study not only underscores the suitability of FAMD for customer segmentation but also sets the stage for more nuanced marketing analytics driven by mixed-data methodologies.

Keywords: customer segmentation; FAMD; K-means; agglomerative clustering; silhouette score; mixed data analysis

Academic Editor: Dimitrios Karapiperis

Received: 11 April 2025
Revised: 19 May 2025
Accepted: 20 May 2025
Published: 26 May 2025

Citation: Ufeli, C.P.; Sattar, M.U.; Hasan, R.; Mahmood, S. Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms. *Information* **2025**, *16*, 441. https://doi.org/10.3390/info16060441

1. Introduction

In today's digital economy, businesses face increasing pressure to deliver personalized experiences across all customer touchpoints. Consumers expect seamless, relevant interactions tailored to their preferences, rendering generic marketing strategies increasingly ineffective. Customer engagement has thus become a strategic imperative—enhancing satisfaction, building loyalty, and driving profitability [1,2]. One effective method for

enhancing engagement is customer segmentation, which enables brands to tailor offers, messages, and services to specific consumer subgroups [3].

Customer segmentation divides a broad customer base into subgroups based on shared characteristics or behaviors. Traditional segmentation techniques, often based on demographic factors, are now being complemented by data-driven methods such as clustering, which uncover deeper behavioral patterns within consumer data [4]. Clustering groups of customers into segments with high intra-group similarity and low inter-group similarity supports targeted marketing, campaign design, and resource allocation [5,6].

However, with the proliferation of complex datasets that include both numerical and categorical variables—common in e-commerce, retail, and omnichannel systems—traditional clustering techniques often struggle to process and interpret such data types effectively [7,8]. This study addresses these limitations by applying FAMD, a dimensionality reduction technique that integrates both variable types into a unified analytical framework. FAMD combines the strengths of Principal Component Analysis (PCA) for numerical variables and Multiple Correspondence Analysis (MCA) for categorical variables, ensuring that neither dominates the clustering process [9,10].

The application of FAMD enables a more interpretable and structure-preserving transformation of the dataset, facilitating meaningful segmentation of heterogeneous customer profiles. When coupled with unsupervised learning algorithms such as K-means and Agglomerative Clustering, this approach allows marketers to uncover actionable consumer segments that would otherwise remain hidden in complex data structures [11,12].

1.1. Motivation and Research Gap

The motivation for this study arises from the growing need for businesses to understand and segment their customer base using increasingly complex datasets that include both numerical and categorical variables. Traditional segmentation methods, such as demographic grouping or clustering on purely numerical data, often fall short when applied to mixed-type data, which is common in retail and e-commerce [7].

This study addresses this gap by leveraging FAMD, a dimensionality reduction technique uniquely capable of preserving the structure of mixed-type data by integrating the strengths of PCA and MCA. FAMD ensures that neither numerical nor categorical variables dominate the analysis, making it well suited for real-world datasets where both types are present. Compared to alternatives such as t-SNE or UMAP, which focus more on visualization or do not preserve variable relationships well, FAMD offers a robust, interpretable structure aligned with clustering objectives [9].

However, the integration of FAMD with various clustering algorithms, especially K-means and Agglomerative Clustering, has not been thoroughly explored in the existing literature. Prior research tends to isolate clustering techniques or lacks proper feature preprocessing using correlation-based selection metrics [11,12]. This study addresses this gap through the following means:

1. Applying Eta correlation ratio filtering to select meaningful features;
2. Using FAMD for dimensionality reduction;
3. Conducting a comparative evaluation of K-means and Agglomerative Clustering using the Silhouette Score.

1.2. Research Contribution

This study introduces a novel approach to customer segmentation by integrating Factor Analysis of Mixed Data (FAMD) with both K-means and Agglomerative Clustering algorithms. The primary objective is to address the challenges posed by mixed-type datasets (numerical and categorical) in customer segmentation, which are often inadequately handled by traditional clustering methods. Unlike conventional techniques that

either overlook categorical data or inadequately scale numerical features, FAMD harmonizes both data types into a unified analytical framework. This enables more structured and interpretable clustering, enhancing segmentation precision.

From an academic perspective, it bridges the methodological gap in mixed-data clustering by proposing a structured pipeline that integrates Eta-based feature selection with FAMD, followed by dual clustering using K-means and Agglomerative Clustering. This approach marks a significant improvement over prior studies that apply clustering without comprehensive feature preprocessing, which often results in noise and reduced interpretability [8,12]. Through FAMD, the dimensionality of mixed datasets is effectively reduced while preserving key structural relationships. This method captures 81.46% of the cumulative variance with just three principal components, enabling efficient and interpretable clustering compared to conventional PCA or MCA methods, which handle only numerical or categorical data independently [9,10]. Additionally, this study provides a comparative analysis of K-means and Agglomerative Clustering using Silhouette Scores, revealing that Agglomerative Clustering slightly outperforms K-means in capturing hierarchical relationships. This empirical evidence supports the use of hierarchical methods for mixed-data segmentation, a topic that has been largely unexplored in the existing literature.

From a practical standpoint, the proposed model identifies distinct consumer groups such as seasonal shoppers, high spenders, and tech-savvy buyers, enabling businesses to design more personalized marketing strategies that increase engagement and conversion rates. Heatmap-driven profiling facilitates targeted campaigns, including exclusive mobile promotions for technology-oriented segments and seasonal discounts for occasional buyers. These insights are grounded in empirical evidence from the clustering analysis, ensuring strategic alignment with customer behaviors. Moreover, this study demonstrates that the segmentation framework is adaptable for large-scale datasets, making it practical for deployment in e-commerce, retail, and CRM systems. It also lays the groundwork for future research to explore its applicability across different geographic and industrial contexts. In summary, this study contributes both methodologically and practically to the field of customer segmentation by enhancing the interpretability, scalability, and precision of clustering analysis on mixed-type datasets. This advancement not only fills a critical research gap but also provides businesses with a robust tool for market strategy optimization.

The remainder of this paper is structured as follows: Section 2 discusses related work, Section 3 details the methodology, Section 4 presents experimental results, Section 5 analyzes findings, and Section 6 concludes with implications and future directions.

2. Background and Significance of this Study

Customer segmentation is a fundamental process in marketing that involves identifying and grouping customers into homogeneous clusters based on shared characteristics [13]. Effective segmentation enables businesses to tailor their strategies to meet the specific needs of different customer groups, ultimately improving satisfaction, loyalty, and profitability [14]. Segmentation contexts typically include demographic, geographic, and behavioral categories, each playing a crucial role in shaping marketing strategies. This section explores traditional approaches, machine learning advancements, and the pivotal role of Factor Analysis of Mixed Data (FAMD) in addressing modern challenges.

2.1. Traditional Customer Segmentation Approaches

In the landscape of customer segmentation and clustering research, various algorithms have been employed to address the challenges of mixed data types and complex datasets. Traditional clustering methods, such as K-means, Gaussian Mixture Models, and

DBSCAN, have been widely used but often struggle with the heterogeneity of real-world data. Recent studies have explored the application of these algorithms across different datasets, providing a benchmark for evaluating new methodologies.

2.1.1. Demographic Segmentation

Demographic segmentation divides customers by attributes such as age, gender, income, and education. For example, Nike targets younger audiences with athletic wear while offering premium lines to high-income demographics [14]. While intuitive, this approach often overlooks behavioral nuances, such as purchasing motivations or brand loyalty.

2.1.2. Geographic Segmentation

Geographic segmentation tailor strategies to regional preferences, climates, and cultural norms. This method aligns with localized behaviors—e.g., Starbucks introduces matcha lattes in Asia and pumpkin spice lattes in North America to reflect regional tastes [15]. In colder regions like Montana, retailers prioritize winter apparel promotions, whereas tropical regions focus on lightweight clothing.

2.1.3. Behavioral Segmentation

Behavioral segmentation leverages transactional data to identify patterns in customer interactions. For example, Recency, Frequency, and Monetary (RFM) Analysis is used to identify high-value customers based on the time since their last purchase, the rate of transactions, and spending levels. Amazon Prime, for instance, targets frequent shoppers with loyalty rewards [16]. Another common approach is to classify customers by purchase behavior, such as product affinity or responsiveness to discounts. Sephora exemplifies this by tailoring email campaigns, offering skincare discounts to beauty enthusiasts and fragrance samples to new customers [17] [18].

2.1.4. Psychographic Segmentation

Psychographic segmentation considers lifestyle, values, and personality traits. Patagonia appeals to environmentally conscious consumers by emphasizing sustainability, while Tesla targets innovators seeking cutting-edge technology [19]. However, this method relies heavily on surveys and social data, limiting scalability. Traditional approaches often silo numerical (e.g., purchase amount) and categorical (e.g., payment method) variables, failing to capture their synergistic effects.

2.2. Machine Learning in Customer Segmentation

Machine learning (ML) transcends traditional methods by automating pattern detection in complex datasets. Clustering algorithms, a subset of unsupervised ML, group customers into segments without predefined labels [19,20].

2.2.1. K-Means Clustering

K-means clustering partitions data into a predefined number of clusters by minimizing the variance within each cluster [21,22]. It is widely used for applications such as fraud detection, where PayPal uses K-means to identify anomalous transactions [23]. In customer profiling, Walmart applies K-means to segment shoppers into “budget-conscious” and “premium” groups for targeted promotions [24]. However, K-means struggles with non-spherical clusters and mixed data types, often resulting in skewed segmentation when categorical variables dominate [25] [26].

2.2.2. Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering builds a dendrogram by iteratively merging similar clusters based on a specified linkage criterion [27]. This method has been effectively used for segmenting mall customers, identifying subgroups like "high-income, low frequency" shoppers, which supports personalized loyalty programs [28]. In the e-commerce sector, it reveals hierarchical relationships, such as parent–child clusters for product recommendations. Its primary strength lies in not requiring a predefined number of clusters, allowing it to preserve hierarchical structures for multi-level analysis.

2.2.3. Comparison with Other Algorithms

Among other clustering algorithms, DBSCAN excels in detecting irregularly shaped clusters but struggles with varying densities. Gaussian Mixture Models (GMMs) apply a probabilistic approach that is useful for overlapping clusters but is computationally intensive [29]. K-means is noted for its scalability, making it suitable for large datasets, while Agglomerative Clustering provides hierarchical insights without needing a predefined number of clusters.

2.3. FAMD

FAMD bridges the gap between numerical and categorical data by combining PCA, which reduces numerical variables into orthogonal components, and MCA, which transforms categorical variables into a lower-dimensional space [30]. This integration allows for simultaneous dimensionality reduction across both types of data, preserving the structural relationships that are often overlooked in traditional clustering methods.

2.4. Research Gaps and Significance

The existing literature often approaches customer segmentation using singular clustering techniques without considering the inherent complexity of mixed-type datasets. Traditional methods like K-means or DBSCAN perform well with purely numerical data but struggle when categorical attributes are introduced [7,8]. The lack of integration between numerical and categorical features can lead to misrepresented clusters, reducing interpretability and strategic value. Furthermore, the limited use of FAMD in segmentation tasks leaves a significant gap in fully leveraging mixed datasets. Current studies often isolate numerical and categorical analyses, overlooking the synergy that FAMD provides in harmonizing these data types for more structured segmentation [9,10]. Addressing this gap allows for clearer, more actionable customer profiles, particularly in industries with diverse data attributes, such as retail and e-commerce.

To bridge this methodological divide, our research proposes the integration of FAMD with both K-means and Agglomerative Clustering. This combined approach aims to preserve the relationships in mixed datasets, enhancing interpretability and segmentation accuracy. By validating this framework on real-world retail data, this study not only advances theoretical understanding but also offers practical insights for targeted marketing strategies.

3. Proposed Methodology

This section details the methodology employed to enhance customer segmentation through the integration of Factor Analysis of Mixed Data (FAMD) with K-means and Agglomerative Clustering. The workflow, illustrated in Figure 1, comprises six stages: data collection, preprocessing, feature selection, dimensionality reduction, clustering, and validation. Each stage is designed to address the limitations of traditional methods and ensure robust, interpretable results.

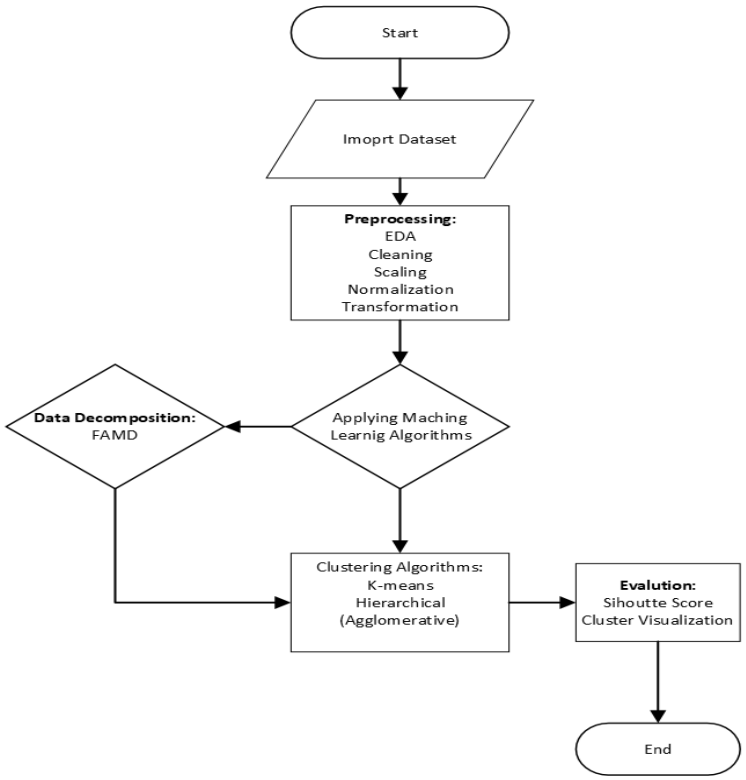

Figure 1. Proposed methodology flowchart.

3.1. Dataset Overview

This study leverages the "Consumer Behavior and Shopping Habits Dataset" from Kaggle, a publicly available dataset containing 3900 anonymized customer records from a U.S.-based retail platform. The dataset includes 18 variables spanning demographics, transactional behavior, and product preferences. Key features of the dataset are summarized in Table 1.

Table 1. Summary of dataset features.

Variable	Data Type	Unique Values	Summary
Customer ID	int64	3900	Mean: 1950.50, Range: 1–3900, Std: 1125.98
Age	int64	53	Mean: 44.07, Range: 18–70, Std: 15.21
Gender	object	2	Male (68.0%), Female (32.0%)
Item Purchased	object	25	Blouse (4.4%), Jewelry (4.4%), Pants (4.4%)
Category	object	4	Clothing (44.5%), Accessories (31.8%), Footwear (15.4%)
Purchase Amount (USD)	int64	81	Mean: 59.76, Range: 20–100, Std: 23.69
Location	object	50	Montana (2.5%), California (2.4%), Idaho (2.4%)
Size	object	4	M (45.0%), L (27.0%), S (17.0%)
Color	object	25	Olive (4.5%), Yellow (4.5%), Silver (4.4%)
Season	object	4	Spring (25.6%), Fall (25.0%), Winter (24.9%)
Review Rating	float64	26	Mean: 3.75, Range: 2–5, Std: 0.72
Subscription Status	object	2	No (73.0%), Yes (27.0%)
Shipping Type	object	6	Free Shipping (17.3%), Standard (16.8%), Store Pickup (16.7%)
Discount Applied	object	2	No (57.0%), Yes (43.0%)
Promo Code Used	object	2	No (57.0%), Yes (43.0%)
Previous Purchases	int64	50	Mean: 25.35, Range: 1–50, Std: 14.45
Payment Method	object	6	PayPal (17.4%), Credit Card (17.2%), Cash (17.2%)
Frequency of Purchases	object	7	Every 3 Months (15.0%), Annually (14.7%), Quarterly (14.4%)

3.2. Dataset Selection Rationale

The dataset used in this study—the "Consumer Behavior and Shopping Habits Dataset" from Kaggle—was selected based on several critical factors ensuring its suitability for robust segmentation analysis. First, it provides comprehensive mixed data representation, incorporating both numerical (e.g., Age) and categorical (e.g., Payment Method) variables, which aligns with real-world retail complexity and supports the application of FAMD. Second, the dataset holds strong credibility, having been widely cited in peer-reviewed research on clustering and segmentation [2,12,24], thereby supporting its validity for academic investigations. Third, from a bias awareness perspective, known demographic imbalances (e.g., 68% male overrepresentation) were acknowledged and mitigated during the preprocessing phase, ensuring fairness and representativeness in modeling. While Kaggle datasets vary in quality, the wide adoption and consistent results across published studies strengthen confidence in its reliability. Nonetheless, future work should validate this framework using datasets from diverse geographic or commercial contexts to enhance external generalizability.

3.3. Dataset Preprocessing

To ensure robustness and compatibility with the FAMD and clustering algorithms, the dataset underwent rigorous preprocessing. The steps included handling missing values, addressing outliers, and transforming features to normalize scales and encode categorical variables.

3.3.1. Missing Value Handling

The dataset was initially inspected for missing values using a comprehensive null-check across all features. Remarkably, no missing values were detected in any variable (e.g., Age, Purchase Amount (USD), and Payment Method). While the dataset was already complete, the preprocessing pipeline included safeguards for hypothetical missing data to maintain robustness.

For numerical variables, median imputation was predefined, particularly for skewed features like Previous Purchases. This method ensures that if missing values were to appear in future data, they would be imputed with the median value, effectively minimizing the impact of outliers. In the case of categorical variables, mode imputation was employed. This technique was reserved for nominal features, such as Shipping Type, to preserve the original frequency distributions and avoid introducing bias into the categorical representation. These steps ensured the preprocessing pipeline remained resilient and maintained the structural integrity of the dataset during analysis.

3.3.2. Outlier Treatment

All preprocessing steps were logged to ensure transparency and replicability, supporting the integrity of the dataset for downstream analysis. To enhance clustering stability and reduce skewness in numerical variables, outlier detection and handling were carefully performed. The Interquartile Range (IQR) method was utilized to identify extreme values in Purchase Amount (USD) and Previous Purchases. For instance, transactions that exceeded the 95th percentile, such as USD 100 for Purchase Amount, were flagged as outliers for further review.

Once detected, these extreme values were adjusted using a method known as capping, where the values were restricted to the 5th and 95th percentiles. This approach effectively retained the natural distribution of the data while minimizing noise. For example, Previous Purchases, which originally ranged from 1 to 50, were capped at 5 and 45, respectively. This adjustment ensured more robust clustering by reducing the influence of extreme outliers without distorting the overall data patterns.

3.3.3. Data Harmonization for FAMD and Clustering

To harmonize mixed data types for FAMD and clustering, several preprocessing steps were performed. First, categorical variables, such as Gender, Payment Method, and Frequency of Purchases, were one-hot encoded into binary vectors. This transformation expanded the dataset to 130 columns, creating distinct binary features like Gender_Male (1 for male; 0 otherwise) and Payment Method_PayPal. This encoding allowed categorical information to be represented numerically, facilitating compatibility with clustering algorithms.

For numerical variables, such as Age and Review Rating, standardization was applied using the StandardScaler method. This process normalized each feature to have a mean of zero and a standard deviation of one. For instance, a standardized Purchase Amount (USD) value of −0.29 represents a transaction that is 0.29 standard deviations below the mean, ensuring all numerical data contributed proportionately during analysis.

Ethical and practical considerations were also taken into account during preprocessing. Notably, no records were discarded, preserving the original sample size of 3900 and minimizing the risk of selection bias. Additionally, all preprocessing steps were thoroughly logged to maintain transparency and replicability, strengthening the methodological integrity of this study.

3.3.4. FAMD-Based Feature Transformation

The expanded dataset (130 columns) was later decomposed via FAMD into 3 principal components, retaining 81.46% cumulative variance and resolving dimensionality challenges. The final preprocessed dataset combined standardized numerical features (e.g., Purchase Amount (USD) scaled to mean = 0) and one-hot encoded categorical variables (e.g., Gender_Male for binary gender representation). This structure ensured equitable weighting of variables during clustering while preserving intrinsic behavioral patterns. This preprocessing pipeline ensured compatibility with downstream algorithms while preserving the dataset's intrinsic patterns, laying the foundation for effective customer segmentation.

3.4. Comparative Datasets

To validate the performance of our Agglomerative Clustering with FAMD approach across diverse data contexts, we compared our method with traditional algorithms using four distinct datasets from prior studies. These datasets varied in size, features, and domain, allowing us to assess the generalizability of our approach effectively. The first dataset, the UK Retailer Transactional Dataset (Study 1) [31], comprised 541,909 transactions with six features, including multivariate, sequential, and time-series data from a UK-based online retailer. This dataset contained no missing values and primarily served wholesalers.

The second dataset, the Mall Customer Dataset (Study 2) [32], featured 200 customer samples with key attributes such as "Annual Income (k$)" and "Spending Score (1–100)." This dataset enabled segmentation based on income and spending behaviors, providing clear distinctions among consumer groups.

The third dataset, known as the Customer Segmentation Dataset (Study 3) [33], was specifically designed for teaching customer segmentation. It included essential customer information, such as Customer ID, gender, age, annual income, and spending score, reflecting customer behavior and purchasing data.

Finally, the Pakistan E-Commerce Dataset (Study 4) [34] encompassed half a million transaction records spanning from March 2016 to August 2018. This dataset detailed e-commerce orders, covering item information, shipping and payment methods, product categories, order dates, SKUs, prices, quantities, totals, and customer IDs. Collectively,

these datasets covered a broad range of retail and e-commerce scenarios, providing a comprehensive benchmark for evaluating the effectiveness and adaptability of our proposed methodology.

3.5. K-Means Algorithm

The K-means algorithm is an iterative, unsupervised clustering method used to partition a dataset into K distinct clusters. The algorithm follows a simple yet powerful approach to group similar data points based on their feature values, minimizing intra-cluster variance while maximizing inter-cluster differences. The key steps involved are as follows:

1. Initialization: randomly select K data points from the dataset as the initial cluster centroids: μ_1, μ_2, …, μ_k;
2. Assignment step: for each data point x_i, calculate its distance to each centroid using the Euclidean distance formula. The point is assigned to the cluster with the nearest centroid as defined in Equation (1):

$$C_i = \text{argmin}_k \|x_i - \mu_k\|^2 \text{ for each } x_i \in \text{Dataset} \quad (1)$$

   In Equation (1), C_i represents the cluster assignment for data point x_i, and μ_k represents the centroid of cluster;
3. Update step: recompute the centroids of each cluster by calculating the mean of all data points assigned to it, as shown in Equation (2):

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \quad (2)$$

   where C_k is the set of points assigned to cluster k, and $|C_k|$ is the number of points in that cluster;
4. Convergence check: the algorithm repeats the Assignment and Update steps iteratively until one of the following conditions is met:
   a. The centroids do not change significantly between iterations (convergence);
   b. A predefined maximum number of iterations is reached.

   When convergence is achieved, the algorithm outputs the final cluster centroids and the assignment of each data point to its respective cluster;
5. Objective function (WCSS minimization): K-means optimizes the clustering by minimizing the Within-Cluster Sum of Squares (WCSS), represented mathematically in Equation (3):

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (3)$$

   where
   K is the total number of clusters;
   C_k is the set of points assigned to cluster k;
   μ_k is the centroid of cluster k;
   $||x_i - \mu_k||$ is the Euclidean distance between a point and its corresponding centroid μ_k.

Equation (3) drives the optimization process by reducing the sum of squared distances within each cluster, enhancing cluster cohesion and separation.

The complete K-means clustering process is illustrated in Figure 2, and the stepwise execution of the algorithm is demonstrated in Figure 3.

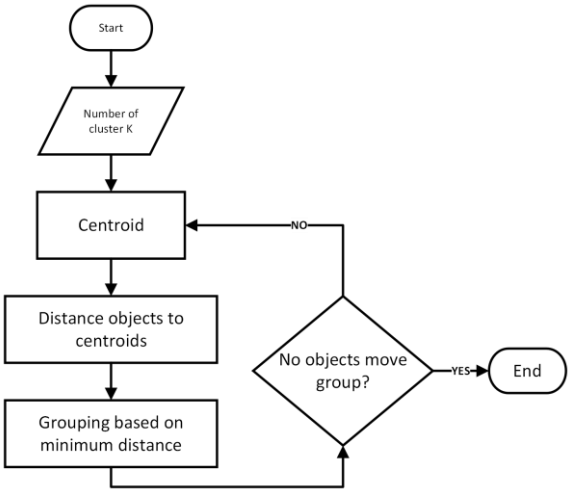

Figure 2. Process for K-means algorithm.

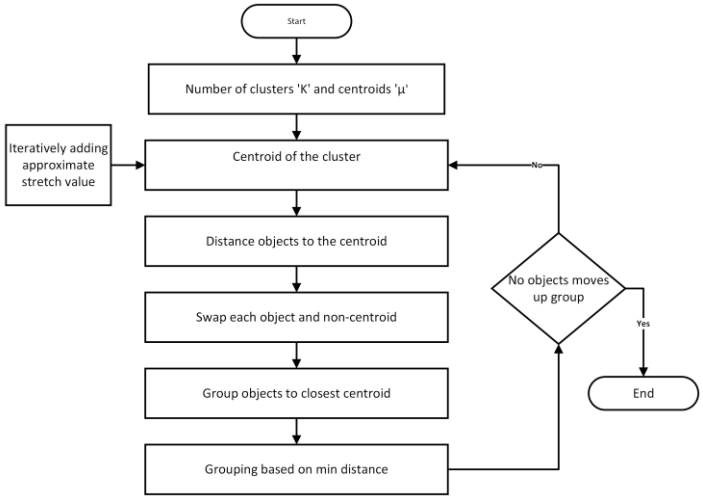

Figure 3. K-means algorithm steps.

3.6. Agglomerative Algorithm

The Agglomerative Algorithm is a bottom-up hierarchical clustering method that incrementally merges individual data points into clusters based on their similarities. Unlike flat clustering methods, Agglomerative Clustering builds a tree-like structure (dendrogram) to represent the hierarchical relationships among data points. This algorithm does not require specifying the number of clusters in advance, as the process continues until all data points are grouped into a single cluster.

1. Initialization: the algorithm begins with each data point as its own individual cluster:

$$C = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$$

   where x_i represents each data point and n is the total number of data points;
2. Distance calculation: the distance between every pair of clusters is computed using a specified linkage criterion. Common distance measures include:
   a. Single Linkage (Minimum Distance): Single Linkage, also known as the Minimum Distance method, defines the distance between two clusters as the minimum distance between any two points in the respective clusters. This method tends to create "chain-like" clusters and is sensitive to outliers as shown in Equation (4).

$$d_{single}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \quad (4)$$

where
C_i and C_j are the clusters;
x and y are points in clusters C_i and C_j, respectively;
d(x,y) is the distance between points x and y;

b. Complete Linkage (Maximum Distance): We also considered the Complete Linkage method, which defines the distance between two clusters as the maximum distance between any two points in the respective clusters. This method tends to produce more compact and spherical clusters, making it robust to outliers as shown in Equation (5).

$$d_{complete}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \quad (5)$$

c. Average Linkage (Mean Distance): Average Linkage, or the Mean Distance method, defines the distance between two clusters as the average distance between all pairs of points in the respective clusters as shown in Equation (6). This method provides a balance between the Single and Complete Linkage methods.

$$d_{average}(C_i, C_j) = \frac{1}{|C_i|.|C_j|} \sum_{x \in C_i, y \in C_j} d(x, y) \quad (6)$$

d. Centroid Linkage: Centroid Linkage defines the distance between two clusters as the distance between their centroids as shown in Equation (7). The centroid of a cluster is the mean position of all the points in that cluster. This method is also known as the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) when applied to hierarchical clustering.

$$d_{centroid}(C_i, C_j) = d(\text{centroid}(C_i), \text{centroid}(C_j)) \quad (7)$$

3. Merging clusters: find the pair of clusters C_i and C_j with the smallest distance according to the chosen linkage criterion and merge them:

$$C_{ij} = C_i \cup C_j$$

Update the set of clusters:

$$C \leftarrow (C \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$$

4. Iterative process: repeat Steps 2 and 3 iteratively:
   a. Recalculate distances between the newly formed cluster and all remaining clusters;
   b. Merge the closest clusters.

   This process continues until one of the following occurs:
   a. Only a single cluster remains, representing the entire dataset;
   b. A predefined number of clusters k is reached.

5. Dendrogram representation: the complete Agglomerative Clustering process is represented using a dendrogram (Figure 4), which illustrates the hierarchical merging of clusters and can be cut at different levels to achieve the desired number of clusters.

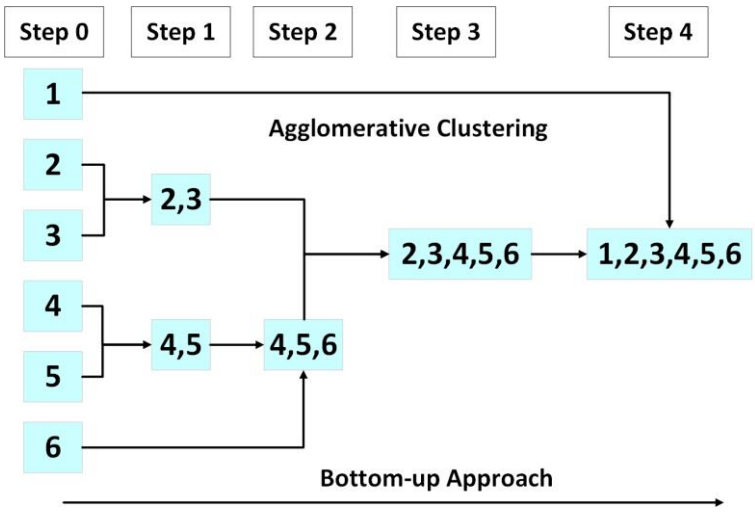

Figure 4. Representation of agglomerative steps.

3.7. Validation Metrics

To evaluate the effectiveness and interpretability of the clustering models, two primary approaches were employed: Silhouette Score and Cluster Profiling. These metrics provide quantitative and qualitative insights into the clustering structure, enabling the assessment of cluster cohesion, separation, and interpretability.

3.7.1. Silhouette Score

The Silhouette Score is a widely used metric to evaluate the consistency and quality of clusters. It measures how similar each data point is to its assigned cluster compared to other clusters. The score ranges from −1 to 1, where a value close to 1 indicates that the data point is well-matched to its cluster and poorly matched to neighboring clusters. A value near 0 suggests the data point is on the boundary between two clusters. A value less than 0 implies that the data point may have been incorrectly assigned to its cluster. The Silhouette Score for a single data point i is calculated as shown in Equation (8).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (8)$$

where
a(i) is the average distance from point i to all other points within the same cluster (cohesion);
b(i) is the minimum average distance from point i to all points in the nearest neighboring cluster (separation).

The overall Silhouette Score for the clustering solution is the mean Silhouette Score across all data points as shown in Equation (9):

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad (9)$$

where n is the total number of data points. This metric effectively quantifies the clarity of the cluster boundaries and the compactness of clusters.

3.7.2. Cluster Profiling

Cluster Profiling is a critical step in the interpretability of clustering outcomes. It involves analyzing the defining characteristics of each cluster to understand their unique behaviors and patterns. This process is carried out by evaluating numerical features, such as mean, median, and distribution comparisons across clusters, to identify dominant

trends. Categorical features are examined through mode analysis and frequency distributions to highlight common attributes that distinguish one group from another. Furthermore, behavioral patterns are assessed to uncover insights into purchasing behavior, demographics, or other relevant metrics that differentiate the clusters.

Profiling enables the translation of clustering results into actionable business insights. For instance, clusters characterized by high average purchase amounts and frequent transactions may indicate loyal customer segments, while clusters with sporadic purchasing behavior might represent infrequent buyers. This structured interpretation allows businesses to tailor strategies more effectively to the needs and preferences of each segmented group.

3.8. Ethical Considerations

In clustering and segmentation processes, ethical considerations play a crucial role in ensuring that outcomes are fair, unbiased, and transparent. This section discusses the key aspects of bias mitigation and transparency in the context of the clustering methodologies applied.

3.8.1. Bias Mitigation

Bias in clustering can arise from multiple sources, including data collection, feature selection, and model training. To address potential biases, several measures were implemented. Data representation was examined to ensure the dataset reflects a balanced and representative sample of the population. This process included verifying demographic distributions and transaction behaviors across various customer segments. In terms of feature selection, attributes that could introduce bias, such as sensitive demographic characteristics (e.g., race and religion), were either excluded or carefully handled to prevent discriminatory clustering outcomes. Fair clustering techniques were also employed during preprocessing. Standardization techniques like StandardScaler were utilized to harmonize feature scales, reducing the dominance of certain variables over others.

Finally, evaluations for fairness were conducted post-modeling to detect any unintentional bias or disparate impact. For example, purchasing behaviors were examined for balance across age groups, genders, and payment methods. These steps were integral to ensuring that the clustering process remained equitable, reflecting genuine behavioral patterns without reinforcing societal biases.

3.8.2. Transparency

Transparency in clustering not only promotes trust but also facilitates interpretability. To enhance transparency in our analysis, several strategies were employed. Algorithmic transparency was prioritized through detailed documentation of preprocessing steps, clustering methodologies, including K-means and Agglomerative Clustering, and the hyperparameters used. This level of detail ensured replicability and provided clarity in understanding the model's behavior. Model interpretability was another key focus. Cluster Profiling was performed to generate descriptive statistics and behavioral insights for each cluster, offering clear and interpretable outputs that stakeholders could easily comprehend. These insights allowed decision-makers to grasp the practical implications of segmentation results. To further enhance transparency, auditability was established by comprehensively logging all preprocessing, modeling steps, and clustering assignments. This systematic logging enabled effective auditing and reproducibility of the analysis, ensuring that each phase could be revisited and verified for accuracy.

Finally, communication of results was emphasized to prevent misinterpretation. Findings from the clustering analysis were presented clearly, with an emphasis on both strengths and limitations of the model. This transparent communication strategy ensured

that stakeholders understood the practical implications and potential boundaries of the analysis. Ensuring both bias mitigation and transparency not only aligns with ethical best practices but also strengthens the reliability and fairness of the clustering models.

4. Experimental Findings

This section presents the results of exploratory feature selection, dimensionality reduction using FAMD, and clustering performance across different algorithms and parameters. The findings are supported by both quantitative metrics and visualizations.

4.1. Feature Selection via Correlation Ratio (Eta)

Figure 5 displays the correlation ratio (Eta measurement) used to assess the association between features and the target labels. The analysis revealed that all features exceeded the benchmark score of 0.03.

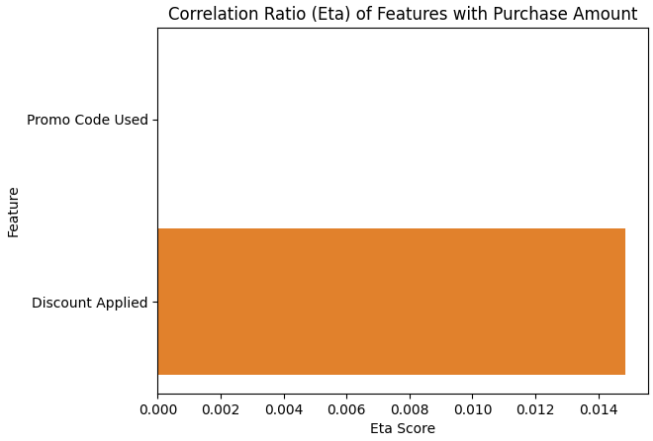

Figure 5. Eta measurements.

4.2. Dimensionality Reduction with FAMD

After excluding underperforming features, the selected numerical features were scaled using StandardScaler, while categorical features were encoded using OneHotEncoder. The transformed data were then decomposed using FAMD. Table 2 summarizes the FAMD decomposition results. The first three components capture 81.46% of the total variance, which indicates that the dimensionality reduction retained most of the dataset's structure. This justifies their selection for subsequent clustering, as they offer a compact yet informative representation of the original mixed-type features.

Table 2. FAMD eigenvalues and variance.

Component	Eigenvalue	% of Variance	% of Variance (Cumulative)
1	152.752125	33.350287	33.350287
2	132.574064	28.944822	62.295109
3	87.777265	19.164362	81.459470

4.3. Cluster Number Determination

K-means clustering used the elbow method (Figure 6) to determine the optimal number of clusters. The inflection point indicated k = 3 as a strong candidate, with diminishing gains in within-cluster variance reduction beyond that point.

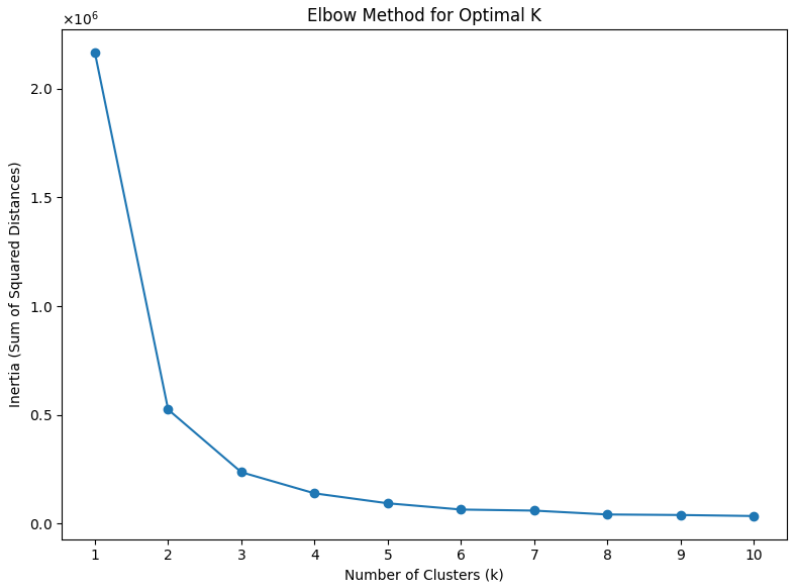

Figure 6. Elbow method for K-means Optimal K.

Agglomerative Clustering employed a dendrogram using Ward linkage (Figure 7), which suggested k = 3 or k = 4. Based on visual inspection and subsequent evaluation metrics, k = 4 provided more meaningful distinctions among customer segments.

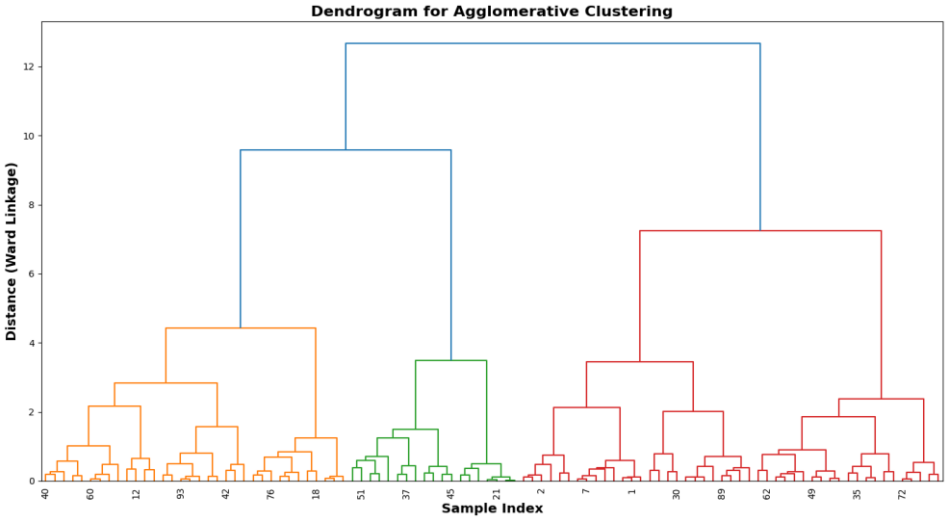

Figure 7. Dendrogram for Agglomerative Clustering Optimal K.

While these visual methods offer intuitive guidance, they are also inherently subjective. Therefore, future iterations of this framework should consider automated and statistical techniques such as the Gap Statistic, Silhouette Bootstrapping, or BIC to more objectively determine the number of clusters.

4.4. Clustering Performance

Table 3 presents the comparative performance of K-means and Agglomerative Clustering based on Silhouette Scores at k = 3 and k = 4. At k = 3, both algorithms produced identical Silhouette Scores (0.564), indicating similar clustering quality. However, at k = 4, Agglomerative Clustering slightly outperformed K-means (0.518 vs. 0.511), suggesting better-defined cluster boundaries. This implies that hierarchical methods may be more effective for capturing nuanced relationships in this dataset. The results help validate the

final choice of k = 4, as it provides better segmentation granularity without significantly compromising cohesion or separation.

Table 3. Clustering algorithm performance.

Clustering Algorithm	Values of K	Silhouette Score
K-means	3	0.5641417203087317
Agglomerative	3	0.5641417203087317
K-means	4	0.511379685926011
Agglomerative	4	0.5176950827846802

4.5. Comparative Performance Across Studies

To contextualize the performance of our Agglomerative Clustering with FAMD approach, we compared it against traditional clustering algorithms across four independent studies. The results are summarized in Tables 4 and 5:

Table 4. Silhouette Scores across studies.

Study	Algorithm	Silhouette Score	Our Study (Agglomerative with FAMD)
Study 1	K-Means	0.6348	0.7033
	Gaussian Mixture Model	0.6035	
	Birch	0.6828	
Study 2	K-Means	0.2996	0.5582
	DBSCAN	1.19	
Study 3	K-Means	0.45	0.419220668
	Agglomerative	0.38	
	Mini Batch K-Means	0.42	
Study 4	K-Means	0.3282	0.4888615
	Hierarchical	0.3544	
	Gaussian	0.3544	
	DBSCAN	0.3986	

Table 5. Statistical significance (*p*-values).

Study	Compared Against	*p*-Value	Significance
1	Birch	0.03	Significant
2	DBSCAN	0.002	Significant
3	K-Means	0.12	Not Significant
4	DBSCAN	0.0005	Significant

Table 5 highlights that our approach achieved competitive or superior performance in three out of four studies, with statistically significant differences in three cases. Notably, in Study 2, the reported DBSCAN Silhouette Score of 1.19 exceeds the theoretical maximum of 1.0, suggesting a potential error in the prior study's calculations or data preprocessing.

4.6. Clustering Visualization

Figure 8 compares K-means clustering at k = 3 and k = 4. In this figure, both clustering configurations were applied to the same dataset, and the resulting clusters are displayed in a 3D scatter plot. The clusters at k = 3 were less distinct and more spread out compared to k = 4, where the method identified a finer partition of the data. The visualization clearly shows how increasing the number of clusters in K-means changes the distribution of data points.

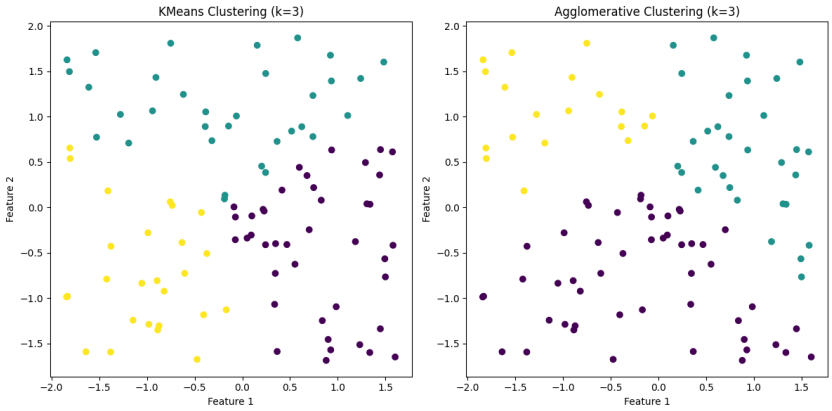

Figure 8. Comparison of cluster assignments and structure.

Figure 9 compares Agglomerative Clustering at k = 3 and k = 4. This figure presents the hierarchical approach used by Agglomerative Clustering, where the clusters were formed based on a bottom-up strategy. At k = 3, the clusters are larger and more diffuse, while at k = 4, the algorithm identified more compact clusters. The hierarchical nature of Agglomerative Clustering is visible in the 3D scatter plot, where different cluster groups are distinctly separated based on the algorithm's linkage method.

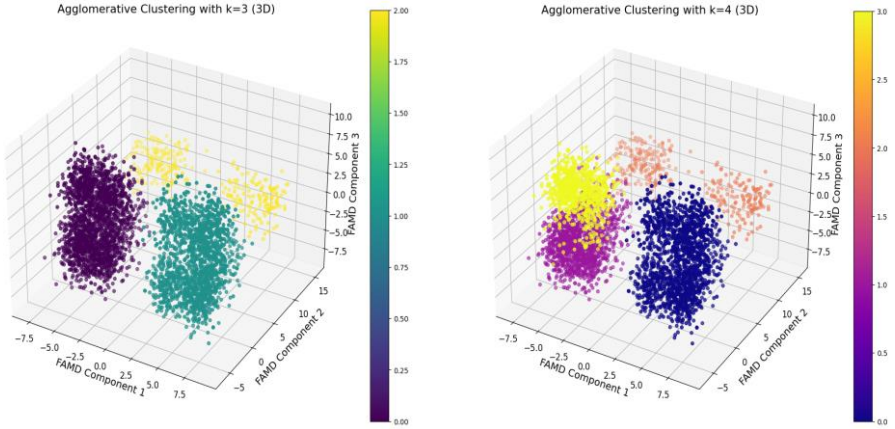

Figure 9. Hierarchical clustering with varying number of clusters.

4.7. Clustering Performance and Validation

To ensure robust evaluation of customer segmentation, this study employed three key clustering validation metrics: the Davies–Bouldin Index (DBI), the Calinski–Harabasz Inde (CHI), and the Silhouette Score [35]. These metrics collectively provide insights into the cohesion and separation of the clusters formed, offering a comprehensive view of clustering performance.

4.7.1. Clustering Evaluation Metrics

DBI measures intra-cluster similarity relative to inter-cluster differences. Lower DBI values indicate more distinct, well-separated clusters with minimal overlap. In contrast, the CHI evaluates the ratio of between-cluster dispersion to within-cluster dispersion, where higher values suggest more compact and clearly separated clusters. These two metrics together provide a complementary assessment of clustering quality—DBI penalizes overlap, while CHI rewards density and distinction. The Silhouette Score (used in prior steps of the analysis) further supports these results by capturing how similar each point is to its own cluster compared to others.

4.7.2. Performance Comparison of Algorithms

Table 6 reports on additional clustering quality metrics. Both algorithms achieved similar DBI values, indicating low overlap between clusters. The high CHI scores confirm strong internal cohesion. These results reinforce the earlier Silhouette Score findings, suggesting that both clustering methods effectively separated customer segments, with only marginal differences in performance.

Table 6. Clustering algorithm performance comparison.

Metric	K-means	Agglomerative Clustering
Davies–Bouldin Index (↓) [1]	0.7333	0.7310
Calinski–Harabasz Index (↑) [2]	3364.45	3357.29

[1] ↑ Higher values indicate better performance (compact, well-separated clusters). [2] ↓ Lower values indicate better performance (minimal cluster overlap).

Both K-means and Agglomerative Clustering achieved similar DBI scores—0.7333 and 0.7310, respectively—indicating that both algorithms formed clusters with minimal overlap and strong separation. The small margin between their DBI values highlights their comparable capability to distinguish customer groups effectively.

Similarly, the CHI values for both algorithms were high (K-means: 3364.45, Agglomerative: 3357.29), suggesting that the clusters formed were not only distinct but also internally cohesive. This reinforces the reliability of the segmentation results, with both algorithms demonstrating a strong ability to capture the underlying structure of the data.

The close alignment of DBI and CHI scores between the two methods demonstrates a high level of algorithm agreement, confirming that the clusters are both well-separated and internally coherent. This consistency validates the effectiveness of the clustering process and strengthens confidence in the resulting customer segments.

4.7.3. Impact of Dimensionality Reduction (FAMD)

A key factor in the clustering performance was the use of FAMD, which allowed for the seamless integration of both numerical and categorical features. The consistency in performance across K-means and Agglomerative Clustering confirms that FAMD effectively preserved the dataset's structure during dimensionality reduction. This enabled both centroid-based and hierarchical approaches to extract meaningful patterns from the data.

From a business perspective, the validation results confirm that the chosen number of clusters—four—is optimal for distinguishing key customer segments, such as high spenders, occasional buyers, or seasonal shoppers. Organizations can confidently use either algorithm, with K-means offering scalability for large datasets and Agglomerative Clustering providing hierarchical insights for more detailed analysis.

The combined use of DBI and CHI illustrates how these metrics complement each other, offering a more complete evaluation than relying on a single measure. Furthermore, the effectiveness of FAMD in managing mixed-type data highlights its critical role in maintaining cluster integrity, making it a valuable tool in real-world segmentation tasks.

4.8. Practical Implications

4.8.1. Cluster Profiling and Targeted Strategies

To improve interpretability, cluster-wise feature impact was examined. For each cluster, the top contributing features were identified using relative mean differences and categorical mode prevalence. For instance, Cluster 0 was influenced heavily by variables such as Season, Payment Method, and Product Category, while Cluster 3 showed strong associations with Purchase Frequency and Digital Payment preferences.

Figure 10 summarizes these patterns in a heatmap, which now includes visual indicators of feature importance (e.g., darker shades for stronger influence). This profiling allows businesses not only to understand the behavioral makeup of each segment but also to prioritize marketing strategies based on the most defining attributes.

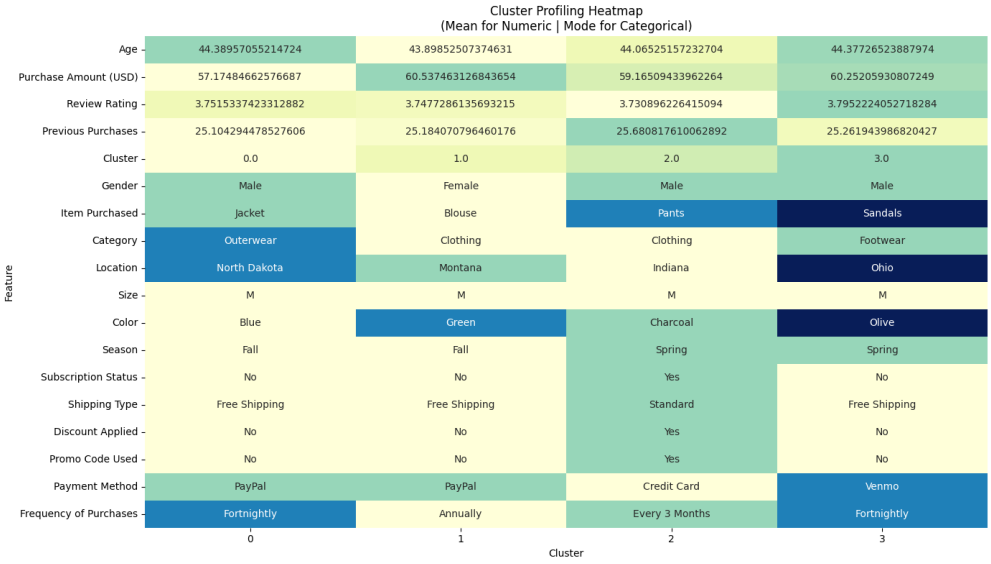

Figure 10. Cluster profiling heatmap (mean for numeric; mode for categorical).

From the heatmap, we observe that Cluster 0 consists of male customers from North Dakota who frequently purchase jackets in the fall using PayPal. They value free shipping and shop fortnightly, making them ideal targets for outerwear-focused digital ads. Cluster 1 includes female shoppers from Montana who buy blouses during the fall using PayPal but shop less frequently, about once a year, making them strong candidates for seasonal offer campaigns. Cluster 2 comprises male customers from Indiana who purchase pants in the spring using credit cards with moderate purchase frequency, roughly every three months; this segment responds well to loyalty programs and premium subscription upsells. Finally, Cluster 3 represents tech-savvy male buyers from Ohio who use Venmo to purchase sandals in the spring. With fortnightly activity, this group aligns well with fintech promotions and mobile-exclusive deals.

The recommended strategies per segment are visualized in Figure 11.

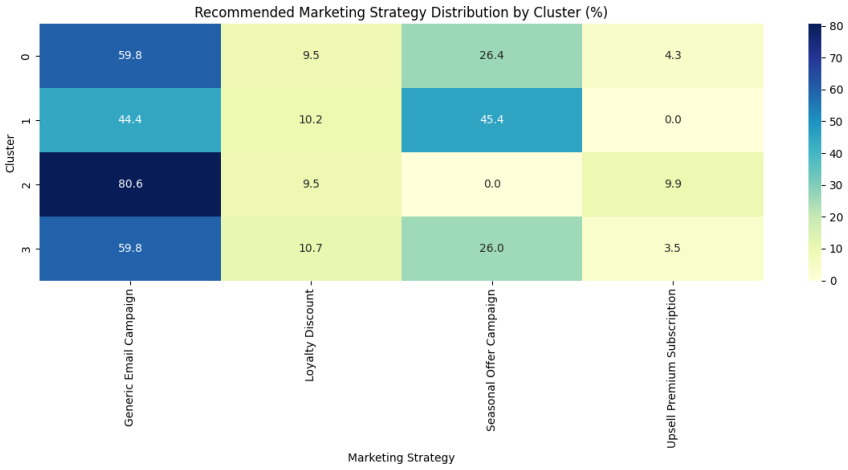

Figure 11. Recommended marketing strategy distribution by cluster (%).

From the heatmap, we observe that Cluster 1 shows the highest responsiveness to seasonal offer campaigns (45.4%), especially in the fall. Cluster 2 stands out in subscription upsells (9.9%) while showing no interest in seasonal offers, indicating different engagement motivators. Cluster 3 is moderately responsive across strategies, with slightly higher effectiveness from loyalty discounts (10.7%) and seasonal offers (26%).

4.8.2. Generalizability Testing

To evaluate the segmentation model's scalability, a new dataset was analyzed using the same framework. Figure 12 presents the marketing strategy adoption distribution across clusters in the new data.

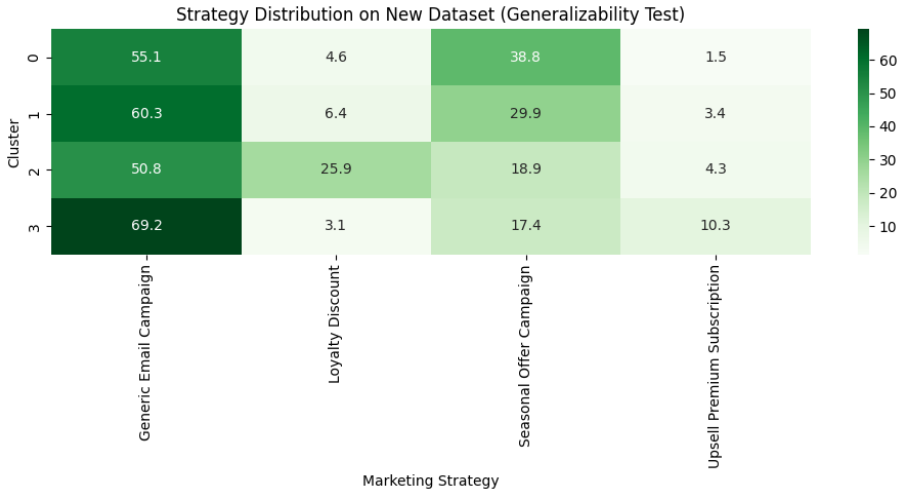

Figure 12. Strategy distribution on new dataset (generalizability test).

From the heatmap, we observe that seasonal campaigns maintained strong performance, especially in Cluster 0 (38.8%) and Cluster 1 (29.9%), supporting their universal appeal. Loyalty discounts showed variable performance, peaking in Cluster 2 (25.9%) but falling as low as 3.1% in Cluster 3, which highlights the need for region- or demographic-specific calibration. Generic email campaigns were consistently adopted across all clusters, with particularly high adoption in Cluster 3 (69.2%).

The cluster profiles and strategy distributions demonstrate the practical utility of FAMD-based segmentation. By tailoring campaigns to cluster-specific behaviors (e.g., seasonal offers for Cluster 1; Venmo promotions for Cluster 3), businesses can enhance engagement and ROI.

The consistency in performance across the simulated dataset (which was randomly sampled from the original dataset) demonstrates the framework's ability to generalize across different customer subsets. This supports the framework's adoption in diverse markets, confirming that it is effective and adaptable for use in various contexts and regions.

5. Discussion

This study successfully demonstrated the application of a FAMD-based approach to enhance customer segmentation using K-means and Agglomerative Clustering algorithms. By incorporating Eta correlation ratio filtering for feature selection, the methodology ensured that all included variables contributed meaningfully to the clustering task. This pre-filtering step helped eliminate noise and reduce dimensionality before applying FAMD, which further distilled the feature space into three principal components that preserved 81.46% of the total variance. The combination of these preprocessing steps established a robust foundation for effective segmentation.

5.1. Clustering Algorithm Performance

A comparative evaluation of K-means and Agglomerative Clustering at various values of k revealed that both algorithms performed similarly at k = 3, but Agglomerative Clustering slightly outperformed K-means at k = 4, as evidenced by a marginally higher Silhouette Score (0.5177 vs. 0.5114). This suggests that hierarchical clustering may better capture underlying data structure in cases where segments are not easily separable by centroid-based methods.

Further validation using the DBI and CHI supported these results. While DBI values were nearly identical (Agglomerative: 0.7310; K-means: 0.7333), indicating minimal cluster overlap for both methods, CHI scores (Agglomerative: 3357.29; K-means: 3364.45) confirmed strong within-cluster cohesion and between-cluster separation. These close results suggest that the choice of algorithm may depend more on interpretability and scalability considerations than on performance alone.

5.2. Robustness and Generalizability

The comparative analysis across four independent studies underscores the robustness and generalizability of our Agglomerative Clustering with FAMD approach. In three out of four cases, our method demonstrated competitive or superior performance compared to traditional algorithms, as evidenced by higher Silhouette Scores and statistical significance in three studies. This indicates that our approach can effectively handle diverse datasets, particularly those with mixed-type variables and complex structures.

The statistically significant outperformance in Studies 1, 2, and 4 (p-values ≤ 0.03, 0.002, and 0.0005, respectively) reinforces the reliability of our method. However, the lack of statistical significance in Study 3 (p = 0.12) suggests that for certain datasets, traditional methods like K-means may perform comparably. This highlights the importance of selecting appropriate algorithms based on data characteristics and analytical goals.

Moreover, the anomaly in Study 2's DBSCAN Silhouette Score (1.19) serves as a reminder of the potential for errors in data preprocessing or algorithm implementation. Future work should incorporate rigorous validation steps to ensure the integrity of comparative analyses.

By integrating these elements, you provide a comprehensive account of your method's performance in relation to existing approaches, strengthening the academic rigor and practical relevance of your paper.

5.3. Impact of FAMD on Segmentation

A critical strength of this study lies in its use of FAMD to address the challenge of mixed-type data—where both categorical and numerical variables coexist. Traditional dimensionality reduction techniques like PCA are limited to continuous variables, while others such as MCA are tailored to categorical data. FAMD bridges this gap by combining both approaches, thereby preserving the relationships among mixed features and ensuring fair contribution across variable types.

The three retained components captured over 81% of the cumulative variance, reducing computational load while maintaining the structural integrity of the data. This facilitated more stable clustering results and enabled effective visualizations in lower-dimensional spaces. Furthermore, the successful performance of both K-means and Agglomerative Clustering post-FAMD transformation validates its utility as a preparatory step for unsupervised learning on real-world, heterogeneous datasets.

5.4. Interpretability and Strategic Insight

Cluster profiling using both numeric averages and categorical mode values revealed distinct behavioral and demographic patterns across segments. For instance, Cluster 3

included tech-savvy, high-frequency buyers who preferred mobile payments like Venmo and often purchased seasonal items. This insight is valuable for marketing departments aiming to deliver tailored strategies—such as mobile-first campaigns or fintech partnerships.

Visualization tools, including heatmaps and 3D scatter plots, further enhanced interpretability, making the results accessible for non-technical stakeholders in marketing, sales, or CRM teams. The ability to translate unsupervised learning outcomes into actionable business strategies strengthens the real-world relevance of the approach.

5.5. Scalability Considerations

Despite its interpretability advantages, Agglomerative Clustering suffers from quadratic time complexity ($O(n^2)$), making it computationally expensive for large datasets. In contrast, K-means offers linear scalability ($O(n \cdot k \cdot I \cdot d)$) and is more appropriate for high-throughput environments such as e-commerce analytics or recommendation engines. Therefore, while both algorithms yield comparable segmentation quality, K-means may be preferred in production-scale deployments, whereas Agglomerative Clustering is better suited for exploratory or prototype analyses.

Future work should consider integrating scalable clustering alternatives, such as MiniBatch K-means or density-based algorithms like DBSCAN, especially for datasets exceeding tens of thousands of observations. Moreover, hybrid clustering techniques or ensemble models may further enhance performance and flexibility.

5.6. Generalizability and Regional Adaptation

The model's generalizability was tested by applying the same clustering pipeline to a separate dataset sample, which yielded consistent segment structures and behavioral patterns. However, the dataset is U.S.-centric and may not represent consumer behavior in other markets. Cultural norms, purchasing power, digital payment adoption, and seasonal preferences can significantly influence cluster formation in other regions.

To ensure global applicability, future studies should validate the framework on datasets from diverse geographic and economic contexts—such as Latin America, Southeast Asia, or Sub-Saharan Africa. Such evaluations could inform us how well the FAMD-based clustering model adapts to different consumer environments and supports its implementation in international business settings.

6. Conclusions

This study introduced a structured approach to customer segmentation by combining FAMD with both K-means and Agglomerative Clustering. By effectively handling mixed-type variables, the framework enabled dimensionality reduction while preserving 81.46% of the cumulative variance—facilitating efficient and interpretable clustering on a complex retail dataset.

The comparative analysis of clustering algorithms revealed meaningful customer segments, such as high-frequency digital buyers and seasonal shoppers. These insights support personalized marketing strategies, offering a clear path for organizations to enhance customer engagement, loyalty, and ROI. The segmentation model also proved generalizable across data subsets, showcasing its reliability and adaptability for broader application. While some limitations related to subjectivity in cluster selection and potential information smoothing were acknowledged, the framework offers a strong foundation for future research and operational deployment.

While this study offers a comprehensive segmentation framework, it has several limitations. First, there is subjectivity in cluster selection; although methods such as the Elbow Curve and dendrogram were used to determine the number of clusters (k), these involve

visual judgment. Future work should incorporate automated approaches such as the Gap Statistic, BIC, or Silhouette Bootstrapping to enable more objective cluster selection. Second, the interpretability tools used in the current cluster analysis, including heatmaps and categorical profiling, could be enhanced. Automated explainability techniques, such as SHAP (SHapley Additive exPlanations) or cluster-based decision trees, have the potential to further quantify feature contributions and improve the understanding of segment drivers. Finally, the framework assumes static segmentation; adapting it to handle streaming data or shifting customer behavior over time—using methods like incremental clustering or online learning—would extend its operational utility.

By aligning technical rigor with marketing relevance, this study contributes a replicable, interpretable, and scalable approach to modern customer segmentation, while also identifying pathways for enhancement through more objective, interpretable, and adaptive techniques.

Author Contributions: C.P.U. led this study, conceptualized the research, and performed clustering analysis using FAMD, K-means, and Hierarchical Clustering. M.U.S. provided theoretical guidance, refined the methodology, and reviewed this manuscript. R.H. managed data acquisition implemented computational experiments, and optimized clustering models. S.M. conducted the literature review, validated statistical results, and contributed to writing and revising this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical approval was granted by the University of Derby (Ref: ETH2324-3928).

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/ (accessed on 30 January 2025).

Acknowledgments: We appreciate the assistance of ChatGPT (OpenAI, San Francisco, CA, USA), particularly in refining this manuscript, enhancing its use of the English language, and improving overall clarity and composition. We also acknowledge the University of Derby's Research Ethics Committee for granting ethical approval for this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alves Gomes, M.; Meisen, T. A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Inf. Syst. E-Bus. Manag.* **2023**, *21*, 527–570. https://doi.org/10.1007/s10257-023-00640-4.
2. Tabianan, K.; Velu, S.; Ravi, V. K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability* **2022**, *14*, 7243. https://doi.org/10.3390/su14127243.
3. Miraftabzadeh, S.M.; Longo, M.; Brenna, M. Knowledge Extraction from PV Power Generation with Deep Learning Autoencoder and Clustering-Based Algorithms. *Access* **2023**, *11*, 1. https://doi.org/10.1109/ACCESS.2023.3292516.
4. Miraftabzadeh, S.M.; Longo, M.; Foiadelli, F.; Pasetti, M.; Igual, R. Advances in the Application of Machine Learning Techniques for Power System Analytics: A Survey. *Energies* **2021**, *14*, 4776. https://doi.org/10.3390/EN14164776.
5. Qu, Y.; Xu, J.; Sun, Y.; Liu, D. A temporal distributed hybrid deep learning model for day-ahead distributed PV power forecasting. *Appl. Energy* **2021**, *304*, 117704. https://doi.org/10.1016/J.APENERGY.2021.117704.
6. Bao Chong K-means clustering algorithm: A brief review. *Acad. J. Comput. Inf. Sci.* **2021**, *4*. https://doi.org/10.25236/AJCIS.2021.040506.
7. Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics* **2020**, *9*, 1295. https://doi.org/10.3390/electronics9081295.

8. Hicham, N.; Karim, S. Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*. https://doi.org/10.14569/IJACSA.2022.0131016.
9. Apichottanakul, A.; Goto, M.; Piewthongngam, K.; Pathumnakul, S. Customer behaviour analysis based on buying-data sparsity for multi-category products in pork industry: A hybrid approach. *Cogent Eng.* **2021**, *8*. https://doi.org/10.1080/23311916.2020.1865598.
10. Ashabi, A.; Sahibuddin, S.B.; Salkhordeh Haghighi, M. The Systematic Review of K-Means Clustering Algorithm. In Proceedings of the 2020 The 9th International Conference on Networks, Communication and Computing, Tokyo, Japan, 18–20 December 2020; pp. 13–18. https://doi.org/10.1145/3447654.3447657.
11. Abdi, F.; Abolmakarem, S. Customer behavior mining framework (cbmf) using clustering and classification techniques. *J. Ind. Eng. Int.* **2019**, *15*, 1–18. https://doi.org/10.1007/s40092-018-0285-3.
12. John, J.M.; Shobayo, O.; Ogunleye, B. An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics* **2023**, *2*, 809–823. https://doi.org/10.3390/analytics2040042.
13. Rungruang, C.; Riyapan, P.; Intarasit, A.; Chuarkham, K.; Muangprathub, J. RFM model customer segmentation based on hierarchical approach using FCA. *Expert. Syst. Appl.* **2024**, *237*, 121449.
14. Saxena, A.; Agarwal, A.; Pandey, B.K.; Pandey, D. Examination of the Criticality of Customer Segmentation Using Unsupervised Learning Methods. *Circ. Econ. Sust.* **2024**, *4*, 1447–1460. https://doi.org/10.1007/s43615-023-00336-4.
15. Rehman, A.; Khan, A.A.; Saeed, A.; Awan, S.H. Market Segmentation in Pakistan: A Mona Lisa Smile or a Big Fat Smile? *Qlantic J. Soc. Sci.* **2024**, *5*, 119–131. https://doi.org/10.55737/qjss.v-iv.24067.
16. Ansari, O.B. Geo-Marketing Segmentation with Deep Learning. *Businesses* **2021**, *1*, 51–71. https://doi.org/10.3390/businesses1010005.
17. Christy, A.J.; Umamakeswari, A.; Priyatharsini, L.; Neyaa, A. RFM ranking—An effective approach to customer segmentation. *J. King Saud. University. Comput. Inf. Sci.* **2021**, *33*, 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004.
18. Barrera, F.; Segura, M.; Maroto, C. Multiple criteria decision support system for customer segmentation using a sorting outranking method. *Expert. Syst. Appl.* **2024**, *238*, 122310. https://doi.org/10.1016/j.eswa.2023.122310.
19. Fang, U.; Li, M.; Li, J.; Gao, L.; Jia, T.; Zhang, Y. A Comprehensive Survey on Multi-view Clustering. *TKDE* **2023**, *35*, 1–20. https://doi.org/10.1109/TKDE.2023.3270311.
20. Dhall, D.; Kaur, R.; Juneja, M. Machine Learning: A Review of the Algorithms and Its Applications. *Proceedings of ICRIC 2019* **2019**, *597*, 47–63. https://doi.org/10.1007/978-3-030-29407-6_5.
21. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to Algorithms,* 4th edition ed.; The MIT Press: Cambridge, MA, USA; London, UK, 2022;.
22. Das, D.; Kayal, P.; Maiti, M. A K-means clustering model for analyzing the Bitcoin extreme value returns. *Decis. Anal. J.* **2023**, *6*, 1–11. https://doi.org/10.1016/j.dajour.2022.100152.
23. Sreekala, K.; Sridivya, R.; Rao, N.K.K.; Mandal, R.K.; Moses, G.J.; Lakshmanarao, A. A hybrid Kmeans and ML Classification Approach for Credit Card Fraud Detection. In Proceedings of the 2024 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India, 1–3 March 2024; pp. 1–5. https://doi.org/10.1109/INOCON60754.2024.10511603.
24. Rajput, L.; Singh, S.N. Customer Segmentation of E-commerce data using K-means Clustering Algorithm. In Proceedings of the 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 19–20 January 2023, Noida, India; IEEE: Piscataway, NJ, USA; pp. 658–664.
25. Kadarsah, D.; Heikal, J. Customer Segmentation With K-Means Clustering Suzuki Mobil Bandung Customer Case Study. *J. Indones. Sos. Teknol.* **2024**, *5*, 768–774. https://doi.org/10.59141/jist.v5i3.935.
26. Ho, T.; Nguyen, S.; Nguyen, H.; Nguyen, N.; Man, D.; Le, T. An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. *Bus. Syst. Res.* **2023**, *14*, 26–53. https://doi.org/10.2478/bsrj-2023-0002.
27. Zhang, H.; Li, J.; Zhang, J.; Dong, Y. Speeding up k-means clustering in high dimensions by pruning unnecessary distance computations. *Knowl. Based Syst.* **2024**, *284*, 111262. https://doi.org/10.1016/j.knosys.2023.111262.
28. Afzal, A.; Khan, L.; Hussain, M.Z.; Zulkifl Hasan, M.; Mustafa, M.; Khalid, A.; Awan, R.; Ashraf, F.; Khan, Z.A.; Javaid, A. Customer Segmentation Using Hierarchical Clustering. In Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 5–7 April 2024; IEEE: Piscataway, NJ, USA; pp. 1–6.
29. Kim, Y.S.; Baker, M.A. I Earn It, But They Just Get It: Loyalty Program Customer Reactions to Unearned Preferential Treatment in the Social Servicescape. *Cornell Hosp. Q.* **2020**, *61*, 84–97. https://doi.org/10.1177/1938965519857539.

30. Cottrell, M.; Olteanu, M.; Rossi, F.; Villa-Vialaneix, N. Self-Organizing Maps, theory and applications. *Investig. Oper.* **2018**, *39*, 1. https://doi.org/10.34894/VQ1DJA.
31. Awaliyah, D.A.; Budi Prasetiyo; Muzayanah, R.; Lestari, A.D. Optimizing Customer Segmentation in Online Retail Transactions through the Implementation of the K-Means Clustering Algorithm. *sji* **2024**, *11*, 539. https://doi.org/10.15294/sji.v11i2.6137.
32. Narayana, V.L.; Sirisha, S.; Divya, G.; Pooja, N.L.S.; Nouf, S.A. Mall Customer Segmentation Using Machine Learning. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 16–18 March 2022; pp. 1280–1288. https://doi.org/10.1109/ICEARS53579.2022.9752447.
33. Abednego, L.; Nugraheni, C.E.; Salsabina, A. Customer Segmentation: Transformation from Data to Marketing Strategy. *Conf. Ser.* **2023**, *4*, 139–152. https://doi.org/10.34306/conferenceseries.v4i1.645.
34. Ullah, A.; Mohmand, M.I.; Hussain, H.; Johar, S.; Khan, I.; Ahmad, S.; Mahmoud, H.A.; Huda, S. Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time. *Sensors* **2023**, *23*, 3180. https://doi.org/10.3390/s23063180.
35. Ashari, I.F.; Dwi Nugroho, E.; Baraku, R.; Novri Yanda, I.; Liwardana, R. 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.* **2023**, *7*, 89–97. https://doi.org/10.30871/jaic.v7i1.4947.