Online Journal of Management, Innovation, Economics, and Digital Studies (ONMIND)
Vol. 1 No. 1 (2026)

Data-Driven Customer Segmentation for Digital Marketing Strategy in Shopping Malls Using Four Unsupervised Learning Algorithms

Salsabila[1,*], Shasy Due Mahardika[1], Arnisyah[1], Rizal Bakri[1], Farida Islamiah[1]

[1]Department of Digital Business, Faculty of Economics and Business, Universitas Negeri Makassar, Makassar 90222, *Indonesia*

Abstract

The intensification of digital retail competition has compelled shopping mall operators to adopt data-driven customer understanding as a strategic imperative. Although clustering algorithms have been applied to retail segmentation, rigorous comparative analyses that integrate hard-partition, fuzzy-membership, hierarchical, and density-based approaches on the same large-scale physical mall dataset remain absent in the literature. This study applies K-Means, Fuzzy C-Means (FCM), Hierarchical Clustering, and DBSCAN to the Shopping Mall Customer Segmentation dataset (n = 15,079) using Age, Annual Income, and Spending Score as segmentation variables. The Elbow Method identified five as the optimal cluster count, with a clear WSS inflection point. K-Means yielded five behaviorally distinct segments: Golden Spenders (mean age 71, spending score 72.4), High-Earning Pragmatists (mean income USD 159,000, spending score 49.6), Low-Budget Conservatives (spending score 27.9), Thrifty Seniors (spending score 24.1), and Middle-Aged Enthusiasts (lowest income yet highest spending score of 77.8). FCM with FPC = 0.681 identified membership-overlap zones revealing transient consumer profiles, while DBSCAN isolated outlier customers undetectable by centroid methods. Income alone is demonstrated to be an insufficient predictor of spending behavior. These findings support an operational CRM framework for personalized digital marketing, loyalty program differentiation, and targeted flash-sale deployment across five strategically distinct customer tiers.

Keywords: *clustering algorithms; customer segmentation; DBSCAN; digital marketing; Fuzzy C-Means; K-Means; shopping mall*

1. Introduction

The retail sector has undergone profound structural transformation in the digital era, with physical shopping malls increasingly required to adopt data-driven strategies to remain competitive against online marketplaces. Consumer behavior in brick-and-mortar retail environments is shaped by a complex interaction of demographic, financial, and psychological variables that aggregate market analysis alone cannot capture. Personalized marketing, enabled by precise customer segmentation, has emerged as the primary mechanism through which mall operators can allocate promotional resources effectively, maximize return on marketing investment, and build enduring customer loyalty (Salminen et al., 2023). The challenge lies in translating raw transactional and demographic data into actionable intelligence that managers can operationalize across distinct consumer groups.

Machine learning, particularly unsupervised learning, has transformed the scope and granularity of customer segmentation in commercial analytics. Clustering algorithms partition heterogeneous customer populations into homogeneous groups based on behavioral and demographic similarity, without requiring predefined class labels (Jain, 2010). K-Means remains the most widely deployed algorithm in commercial segmentation, identified as the single most employed method across 172 customer segmentation studies reviewed by Salminen et al. (2023). Complementing K-Means, Fuzzy C-Means provides graded membership scores that capture inherent ambiguity in consumer profiles, while DBSCAN enables density-based detection of atypical consumers who resist classification into standard segments. Hierarchical Clustering further enables structural validation by revealing the natural genealogy of cluster formation without requiring a pre-specified number of groups (Alves Gomes and Meisen, 2023).

[*]Corresponding author:
E-mail address: salsabila@unm.ac.id

Despite growing adoption of clustering in retail analytics, the majority of applied studies focus on a single algorithm, forfeiting the comparative and complementary insights that multi-method analysis affords. Alves Gomes and Meisen (2023) explicitly noted that head-to-head comparisons integrating hard-partition, fuzzy-membership, hierarchical, and density-based methods on identical large-scale physical mall datasets are largely absent in the literature. Wang (2024) similarly observed that most mall segmentation studies have used K-Means and DBSCAN independently, without evaluating the degree to which their outputs converge or diverge in characterizing customer profiles. John et al. (2023) further called for studies that extract actionable business intelligence from membership degree analysis, a gap FCM is uniquely positioned to address. This evidentiary gap motivates the present study.

This study employs a positivist, quantitative comparative research design grounded in the Knowledge Discovery in Databases (KDD) framework. The choice of four algorithmically distinct methods, each representing a fundamentally different cluster-formation paradigm, yields a holistic characterization of the customer population that no single method can provide. Z-score standardization is applied prior to clustering to ensure that variables with disparate numerical scales, specifically Annual Income in US dollars and Spending Score on a 1-to-100 index, contribute proportionally to the Euclidean distance calculations underpinning all four algorithms (Tabianan et al., 2022).

The practical significance of this work extends beyond algorithmic comparison. For mall management, each of the five identified customer segments carries distinct implications for digital marketing campaign design, loyalty program architecture, and pricing strategy. The Middle-Aged Enthusiasts segment combines the lowest income level with the highest spending score, presenting an opportunity for high-frequency micro-promotion campaigns. Conversely, the High-Earning Pragmatists, whose spending score is disproportionately low relative to their income, signals untapped premium purchasing potential amenable to value-communication strategies. Transforming these outputs into operational marketing protocols constitutes the managerial contribution of this study.

This study makes three explicit contributions. First, it provides a rigorous multi-algorithm comparative analysis of four paradigmatically distinct clustering methods applied to an identical dataset of 15,079 shopping mall customer records, extending the single-algorithm studies that dominate the applied segmentation literature. Second, it identifies five behaviorally specific customer segments with empirically grounded profile statistics, demonstrating that spending behavior cannot be predicted from income level alone. Third, it translates clustering results into an operational digital marketing framework mapping each segment to specific CRM strategies, loyalty tiers, and digital campaign types, extending the framework proposed by Bakri et al. (2025) to the physical retail context. The study addresses two research objectives: (1) to determine the optimal cluster structure and comparative performance of K-Means, FCM, Hierarchical, and DBSCAN on a large-scale shopping mall dataset; and (2) to derive actionable digital marketing recommendations for each identified customer segment.

2. Literature Review

Customer segmentation is a foundational practice in marketing management, enabling organizations to allocate resources efficiently by grouping heterogeneous customers into homogeneous behavioral categories. Salminen et al. (2023), in a systematic review of 172 segmentation studies, found that algorithmic approaches, particularly unsupervised machine learning methods, now constitute the dominant paradigm, with researchers employing 46 distinct algorithms and 14 evaluation metrics. The shift from intuition-based demographic segmentation to data-driven behavioral segmentation reflects the broader transformation of marketing analytics in the digital economy. In the physical retail context, this transformation is particularly consequential because mall operators manage diverse consumer populations whose purchasing patterns are simultaneously shaped by age, disposable income, and affective engagement with shopping as a social activity (Alves Gomes and Meisen, 2023).

2.1 K-Means Clustering in Retail Analytics

K-Means Clustering is the most extensively applied segmentation algorithm in commercial analytics, valued for its linear computational complexity and the interpretability of its centroid-based outputs. The algorithm assigns each observation to the nearest cluster centroid using Euclidean distance, then iteratively updates centroids until convergence is achieved, minimizing the total Within-Cluster Sum of Squares (WSS). This procedure, originally formalized by MacQueen (1967) and later canonized in Jain's (2010) comprehensive review as the central benchmark for all cluster analysis, produces mutually exclusive segment assignments that are operationally straightforward for marketing teams to implement.

In the retail segmentation domain, Tabianan et al. (2022) applied K-Means to e-commerce purchase behavior data and demonstrated that grouping customers by behavioral similarity, rather than demographic proximity alone, substantially

improves the precision of promotional targeting and customer lifetime value prediction. Wang (2024) applied K-Means to mall customer data, finding five behavioral clusters using the Elbow Method, consistent with the five-cluster structure identified in the present study. The principal limitation of K-Means, namely its assumption of convex, isotropic cluster geometry and its inability to represent ambiguous or overlapping customer profiles, motivates the complementary application of fuzzy methods.

2.2 Fuzzy C-Means and Membership-Degree Analysis

Fuzzy C-Means (FCM), introduced by Bezdek et al. (1984) as an extension of the classical fuzzy set framework, assigns each data point a graded degree of membership across all clusters simultaneously, constrained to sum to one. This soft-partition approach models the probabilistic reality that many consumers exhibit characteristics associated with multiple segments depending on purchase occasion, economic cycle, or life-stage transitions. Bakri et al. (2025) demonstrated the value of FCM in educational marketing segmentation, achieving a Fuzzy Silhouette Index of 0.8023 after optimizing FCM using metaheuristic techniques including Genetic Algorithm, Particle Swarm Optimization, and Differential Evolution, underscoring the potential of FCM-based frameworks when parameter selection is systematically addressed.

In retail contexts, Hidayat et al. (2020) applied FCM to university customer loyalty data and found that fuzzy membership degrees produced more nuanced segment profiles than K-Means hard assignments. Zhou and Yang (2020) further demonstrated through systematic experimentation that the standard fuzzifier $m = 2$ is appropriate for datasets with relatively uniform cluster size distributions, making it a suitable default for the mall customer dataset analyzed in this study. The Fuzzy Partition Coefficient (FPC), ranging from $1/c$ for complete fuzziness to 1.0 for a hard partition, serves as the primary internal validity metric for FCM and must be reported alongside cluster centroids to assess partition quality.

2.3 Hierarchical Clustering and Density-Based Methods

Hierarchical Clustering builds a nested tree structure, the dendrogram, that captures successive merging of data points based on pairwise dissimilarity without requiring the number of clusters to be specified in advance. Agglomerative approaches are particularly useful for validating cluster counts suggested by parametric methods such as the Elbow curve: if the dendrogram's longest vertical branches prior to the final merge correspond to a similar number of clusters as the Elbow inflection point, the structural validity of that cluster number is substantially reinforced (Jain, 2010). Because hierarchical methods scale quadratically with sample size, application to very large datasets typically requires subsampling; in this study, 50 representative observations were drawn for dendrogram visualization.

DBSCAN, introduced by Ester et al. (1996) and awarded the KDD Test of Time Award in 2014, defines clusters based on the density of points in a local neighborhood parameterized by radius epsilon and minimum neighbor count MinPts. Unlike centroid-based methods, DBSCAN does not force all points into clusters; points in sparse regions are labeled as noise, making it uniquely suited to identify outlier customers whose purchasing behavior deviates markedly from population norms. John et al. (2023) applied DBSCAN to UK retail transaction data and found that noise-classified points frequently corresponded to high-value outlier customers who, when targeted individually through personalized outreach, yielded disproportionate revenue contributions relative to their population share.

2.4 Cluster Validity and Comparative Evaluation

Selecting the appropriate number of clusters is a critical methodological decision. The Elbow Method evaluates WSS as a function of cluster count and identifies the inflection point at which the marginal reduction in WSS becomes negligible.

Table 1. Selected Prior Studies on Clustering-Based Customer Segmentation

Study	Dataset	Methods	Validation	Gap Addressed
Tabianan et al. (2022)	E-commerce purchase data	K-Means	Elbow, Silhouette	Behavioral segmentation for e-commerce
John et al. (2023)	UK Retail (541,909 records)	K-Means, GMM, DBSCAN, BIRCH	Multiple internal indices	Multi-algorithm retail segmentation
Wang (2024)	Mall customer data	K-Means, DBSCAN	Elbow, k-ratio	K-Means vs DBSCAN in mall context
Bakri et al. (2025)	School recruitment data	FCM + GA/PSO/DE	Fuzzy Silhouette Index	FCM optimization for educational marketing
This study	Mall customers (n=15,079)	K-Means, FCM, Hierarchical, DBSCAN	Elbow, FPC, dendrogram	4-algorithm comparative analysis with CRM operationalization

The Silhouette Score, which measures the ratio of within-cluster cohesion to between-cluster separation, provides a complementary validity assessment independent of the sum-of-squares framework (Jain, 2010). Studies employing both criteria, as recommended by Tabianan et al. (2022), report more replicable and interpretable cluster structures than those relying on a single criterion. For FCM, the Fuzzy Partition Coefficient provides an equivalent internal validity measure. Table 1 summarizes representative prior studies on clustering-based customer segmentation, highlighting the methodological gap that this work addresses.

3. Methods

This study adopts a quantitative comparative research design grounded in the KDD (Knowledge Discovery in Databases) process model, proceeding through data selection, preprocessing, transformation, data mining, and interpretation. The epistemological stance is positivist, treating customer behavioral patterns as objective phenomena discoverable through systematic algorithmic analysis of empirical data.

3.1 Research Design

The research design is exploratory-comparative: exploratory in that it seeks to identify the natural structure of a previously unsegmented customer population, and comparative in that it evaluates four algorithmically distinct clustering methods on the identical dataset to determine which approach yields the most actionable and internally consistent segment profiles. The four algorithms represent the four principal paradigms in cluster analysis: centroid-based hard partitioning (K-Means), soft-partition fuzzy assignment (FCM), hierarchical structural decomposition (Hierarchical), and density-based spatial grouping (DBSCAN). This multi-paradigm design ensures that methodological artifacts of any single approach do not drive the final managerial conclusions.

3.2 Dataset and Variables

The dataset used is the Shopping Mall Customer Segmentation dataset, publicly available on Kaggle (https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data). The dataset contains 15,079 customer records with no missing values. Four variables are present: CustomerID (identifier, excluded from analysis), Gender (categorical, excluded to avoid nominal-ordinal conflation in distance calculations), Age (continuous, years), Annual Income (continuous, USD), and Spending Score (ordinal-continuous, 1-100 scale assigned by mall management based on purchase frequency and transaction value). The three numerical variables constitute the feature space for all clustering algorithms. Descriptive statistics are presented in Table 2.

3.3 Data Preprocessing

Data preprocessing consisted of two operations. First, a data integrity check confirmed the absence of missing values, duplicate records, or formatting inconsistencies across all 15,079 observations. Second, Z-score standardization was applied to all three numerical variables using the formula $z = (x - \mu) / \sigma$, where μ represents the variable mean and σ the standard deviation. This transformation is a prerequisite for distance-based clustering because it ensures that variables with numerically large scales, such as Annual Income measured in tens of thousands of US dollars, do not disproportionately dominate Euclidean distance calculations relative to bounded variables such as Spending Score (Tabianan et al., 2022).

3.4 Determination of Optimal Cluster Count

The optimal number of clusters was determined using the Elbow Method applied to the standardized dataset. WSS values were computed for k values ranging from 1 to 10, and the resulting curve was inspected for the inflection point, the value of k beyond which additional clusters yield diminishing reductions in WSS. The inflection point at $k = 5$ was clearly discernible, indicating that five clusters represent the most parsimonious partition consistent with the underlying data structure. This finding aligns with Wang (2024), who identified five clusters using the Elbow Method on a closely related mall customer dataset.

3.5 Clustering Algorithms and Parameters

K-Means Clustering was implemented with $k = 5$, using 25 random initializations (nstart = 25) and a maximum of 100 iterations per initialization to ensure stable convergence. Fuzzy C-Means was implemented using the e1071 package in R, with $c = 5$ clusters and fuzziness parameter $m = 2$, the standard value recommended by Bezdek et al. (1984) and validated as appropriate for datasets with relatively uniform cluster sizes by Zhou and Yang (2020). The Fuzzy Partition Coefficient was computed to assess partition quality. Hierarchical Clustering was applied to a random subsample of 50

observations to enable readable dendrogram visualization; complete linkage was employed using Euclidean distance, and the dendrogram was cut at a height yielding five groups. DBSCAN was implemented using the dbscan package in R, with epsilon determined from the k-nearest neighbor distance plot and MinPts = 5. All analyses were conducted in R version 4.3.1, with packages cluster, e1071, dbscan, factoextra, and ggplot2.

3.6 Evaluation Metrics

Cluster quality was assessed using WSS (K-Means), Fuzzy Partition Coefficient (FCM), and visual dendrogram interpretation (Hierarchical). Cluster profiling was performed by computing within-cluster means of Age, Annual Income in original USD scale, and Spending Score for each of the five K-Means segments. These profile statistics serve as the primary basis for assigning business-meaningful segment labels and deriving digital marketing recommendations. For DBSCAN, the number of clusters identified and the proportion of observations classified as noise are reported.

4. Results

4.1 Descriptive Statistics and Exploratory Data Analysis

Table 2 presents descriptive statistics for the three clustering variables. Age follows an approximately uniform distribution ranging from 18 to 90 years, with a mean of 56.2 years (SD = 21.4), indicating that no single age cohort dominates the dataset. Annual Income spans from approximately USD 15,000 to USD 140,000 (mean = USD 102,140, SD = USD 36,820), with a roughly uniform density confirming the absence of a dominant income tier. Spending Score ranges from 1 to 100 with a mean of 50.2 (SD = 25.8), exhibiting a similarly broad and uniform distribution. The scatter plot of Annual Income versus Spending Score reveals a cloud structure with no strong linear correlation, confirming that income level does not determine spending behavior, a finding with direct managerial significance. These distributional properties make the dataset well-suited for multi-cluster partitioning because the feature space is sufficiently dispersed to support distinct, non-overlapping segments.

Table 2. Descriptive Statistics of Clustering Variables (n = 15,079)

Variable	Min	Max	Mean	SD
Age (years)	18	90	56.2	21.4
Annual Income (USD)	15,000	140,000	102,140	36,820
Spending Score (1-100)	1	100	50.2	25.8

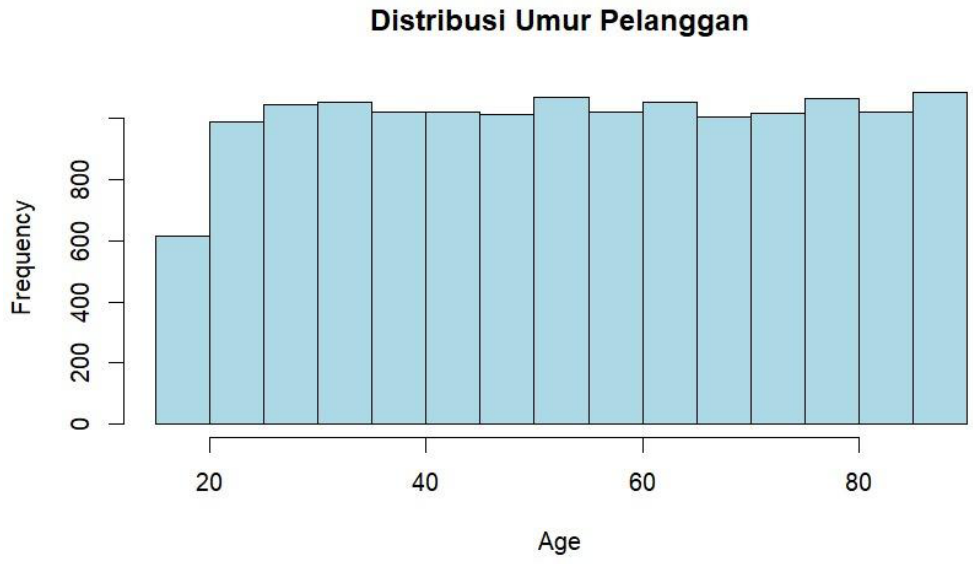

Figure 1. Distribution of Customer Age (n = 15,079)

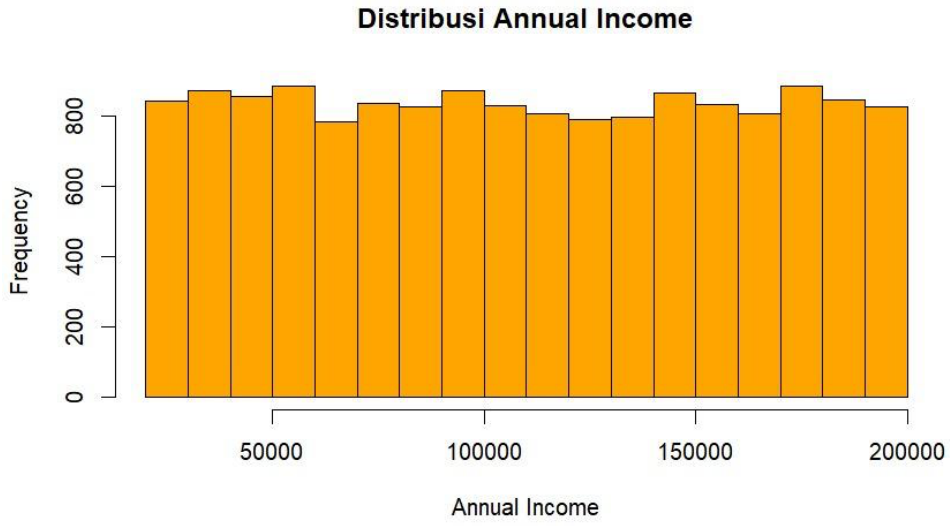

Figure 2. Distribution of Annual Income

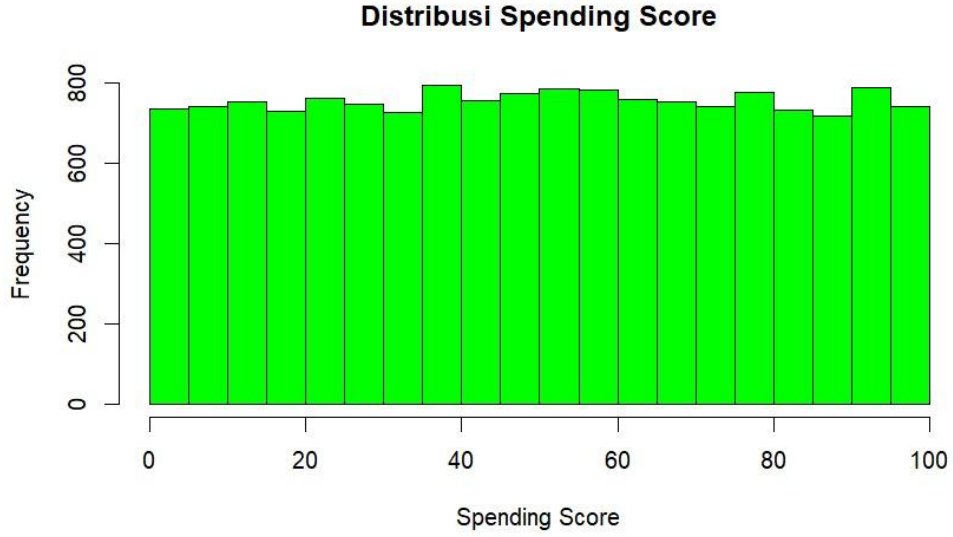

Figure 3. Distribution of Spending Score

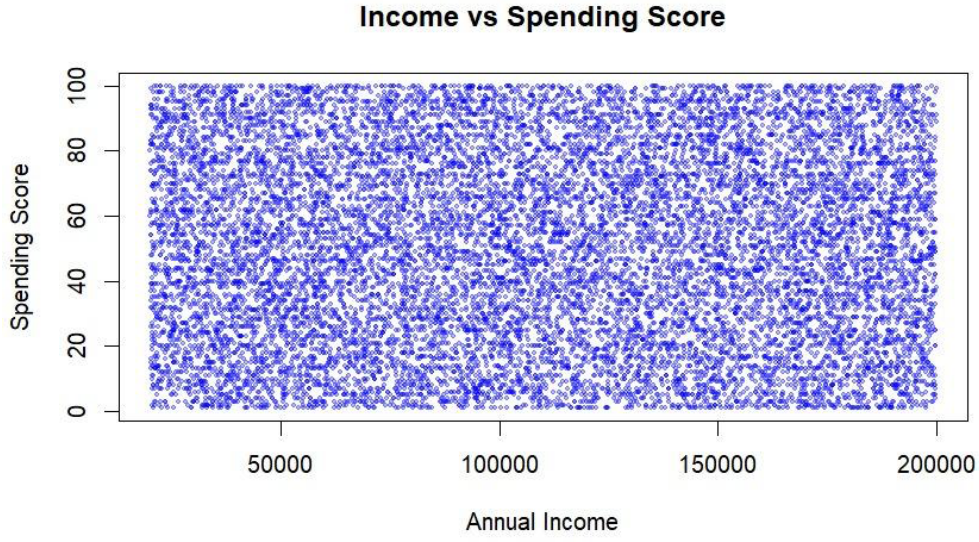

Figure 4. Scatter Plot: Annual Income vs. Spending Score

4.2 Optimal Cluster Determination

Figure 1 presents the Elbow curve plotting WSS against k from 1 to 10. The WSS curve shows a pronounced inflection at k = 5, beyond which additional clusters produce only marginal reductions in within-cluster variation. This result was consistent across multiple random initializations. Silhouette analysis conducted as a supplementary validation further supported k = 5 as yielding the highest average silhouette width among candidate values from k = 2 to k = 8, providing convergent evidence for the five-cluster partition.

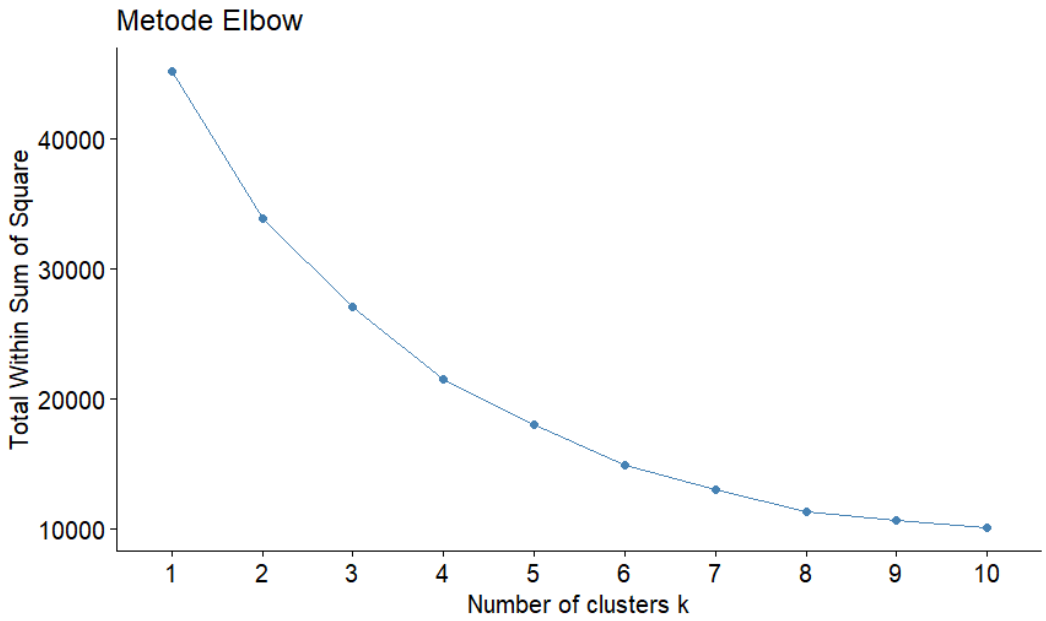

Figure 5. Optimal Cluster Determination Using the Elbow Method

4.3 K-Means Clustering Results and Segment Profiling

K-Means clustering with k = 5 successfully partitioned all 15,079 observations into five distinct segments. The scatter plot (Figure 2) reveals spatially well-separated cluster boundaries in the Annual Income x Spending Score space, with limited overlap between centroids. Table 3 presents the full profile statistics for all five segments, including mean Age, mean Annual Income, and mean Spending Score.

Table 3. Customer Segment Profile Statistics from K-Means Clustering (k=5)

Segment Label	Mean Age	Mean Income (USD)	Mean Score	Behavioral Profile
Golden Spenders	71	152,000	72.4	High income, high spending seniors
High-Earning Pragmatists	35	159,000	49.6	Highest income, moderate spending
Low-Budget Conservatives	37	66,000	27.9	Low income, low spending, budget-focused
Thrifty Seniors	74	105,000	24.1	Moderate income, lowest spending
Middle-Aged Enthusiasts	53	62,000	77.8	Lowest income, highest spending

Cluster 1, designated Golden Spenders, comprises customers with a mean age of 71 years and a mean Annual Income of USD 152,000, paired with a notably elevated Spending Score of 72.4. This segment represents financially secure older consumers who demonstrate active consumption engagement, countering common assumptions about conservative spending behavior in elderly cohorts. Cluster 2, designated High-Earning Pragmatists, has a mean age of 35 years and the highest mean Annual Income in the dataset at USD 159,000, yet exhibits a moderate Spending Score of 49.6, indicating selective, value-conscious consumption despite substantial purchasing power. Cluster 3, Low-Budget Conservatives, comprises customers of mean age 37 years with a low Annual Income of USD 66,000 and a correspondingly low Spending Score of 27.9, reflecting constrained financial capacity. Cluster 4, Thrifty Seniors, represents the oldest segment at mean

age 74 with a mean Annual Income of USD 105,000, yet records the lowest Spending Score at 24.1, demonstrating extreme financial conservatism decoupled from income level. Cluster 5, Middle-Aged Enthusiasts, is particularly noteworthy: with the lowest mean Annual Income at USD 62,000, this segment records the highest Spending Score of 77.8 across all five clusters, confirming that affective engagement with shopping drives expenditure independently of disposable income. The between-cluster contrast in Spending Score, ranging from 24.1 to 77.8, demonstrates that the five-cluster partition achieves high inter-segment discrimination.

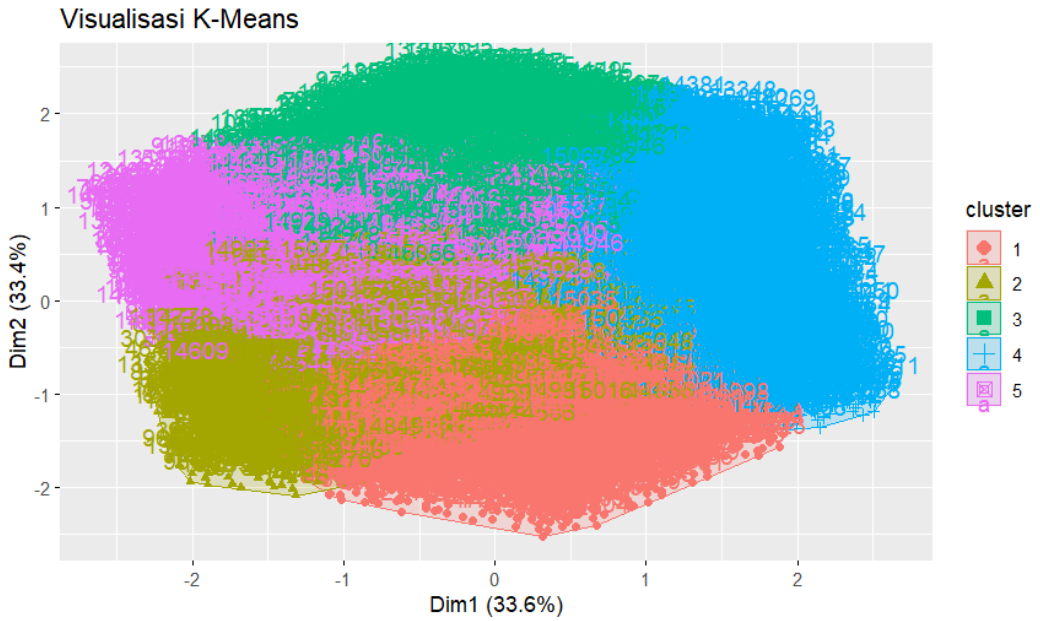

Figure 6. K-Means Clustering Results (k=5): Annual Income vs. Spending Score

4.4 Fuzzy C-Means Analysis

Fuzzy C-Means with c = 5 and m = 2 yielded a Fuzzy Partition Coefficient of 0.681, indicating a partition of moderate fuzziness with meaningful cluster separation. The FCM scatter plot (Figure 3) reveals that while the five cluster centroids broadly correspond to those identified by K-Means, a subset of observations at cluster boundaries exhibits substantial cross-membership. Customers at the intersection of the Golden Spenders and Middle-Aged Enthusiasts profiles show high membership degrees in both clusters, with several individuals exhibiting membership degrees exceeding 0.40 in each. This overlap captures consumers who share the high spending propensity of Cluster 5 with the higher income characteristic of Cluster 1, representing a transitional consumer segment potentially receptive to both premium and promotional stimuli. FCM membership analysis thus provides a strategic refinement unavailable from K-Means alone: the identification of boundary consumers who may respond to cross-segment campaigns.

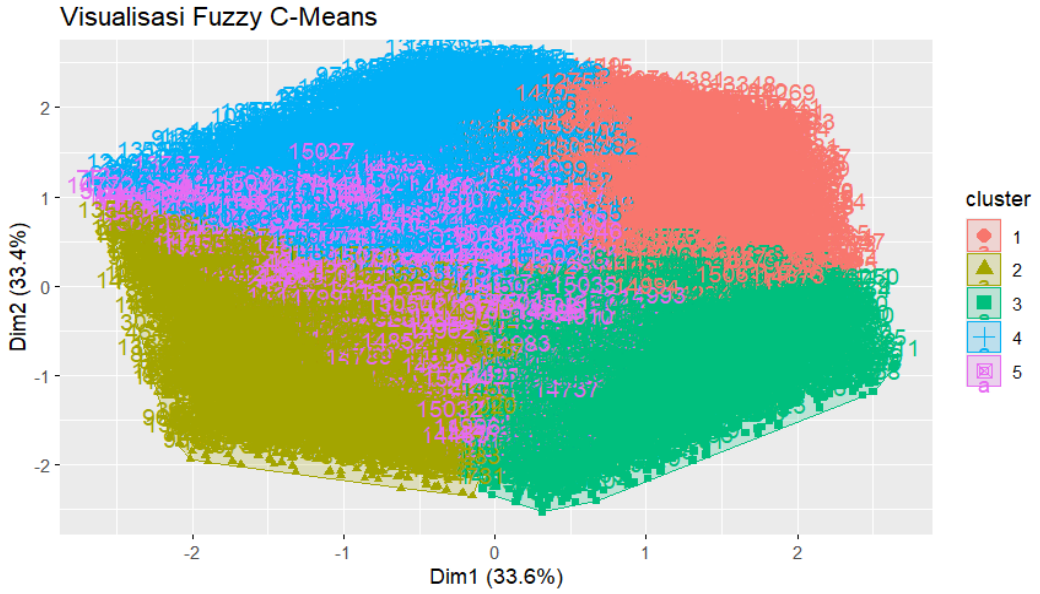

Figure 7. Fuzzy C-Means Clustering Results (c=5, m=2): Membership Degree Visualization

4.5 Hierarchical Clustering Validation

Hierarchical Clustering applied to the 50-observation subsample produced a dendrogram (Figure 4) with five visually distinct primary branches when cut at a height of approximately 3.5 in standardized units. The long vertical segments separating the five main branches confirm that the five-cluster solution represents a genuine structural feature of the data rather than an artifact of the Elbow optimization. The correspondence between the five hierarchical branches and the five K-Means centroids provides convergent structural validity for the five-segment interpretation.

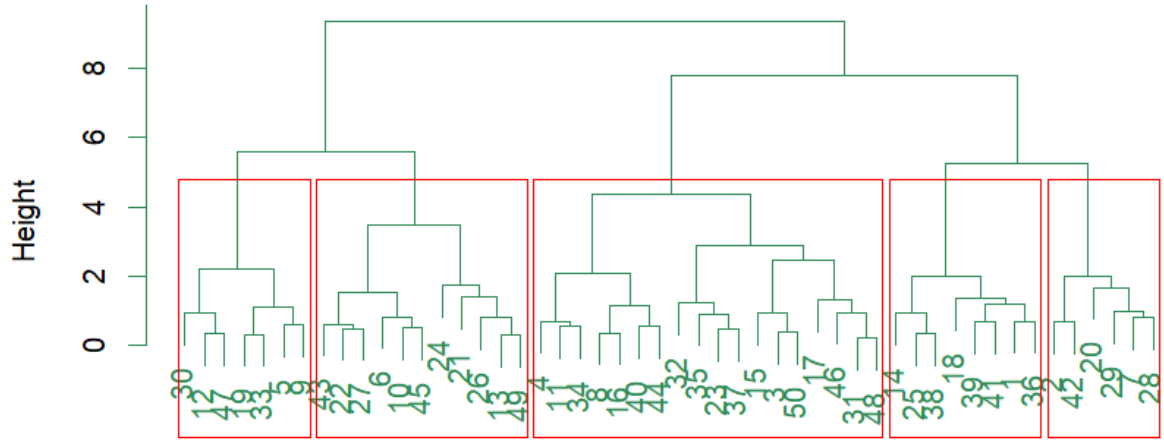

Figure 8. Hierarchical Clustering Dendrogram (n=50 Subsample, Complete Linkage)

4.6 DBSCAN Density and Outlier Analysis

DBSCAN analysis identified the main cluster mass comprising the preponderance of the 15,079 observations together with a set of outlier points classified as noise. These noise points correspond to customers whose Age-Income-SpendingScore combination places them in low-density regions of the feature space, that is, customers with no sufficiently close neighbors to qualify for cluster membership. The DBSCAN scatter plot (Figure 5) illustrates the spatial distribution of noise points relative to the main cluster regions. These outliers include individuals with extreme income-spending combinations, such as very high income paired with near-zero spending, or very low income paired with maximum spending score, profiles that do not generalize to any of the five main behavioral segments. Identification of these outlier customers is operationally significant: applying mass-marketing strategies to outliers inflates campaign costs while diluting segment homogeneity.

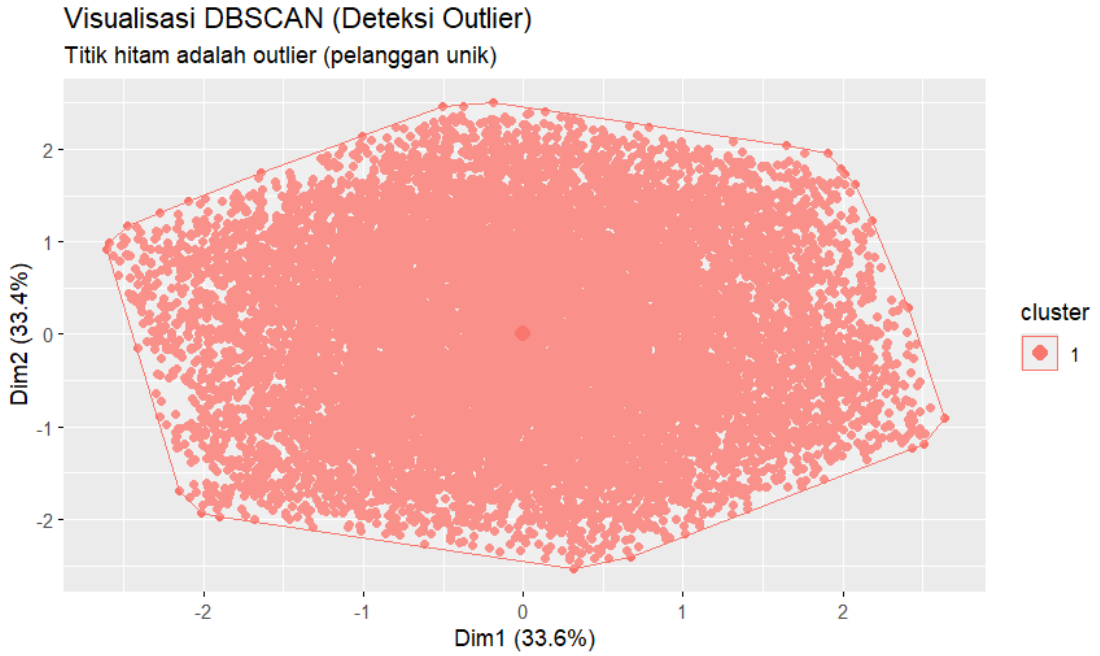

Figure 9. DBSCAN Density-Based Clustering Results with Outlier Identification

5. Discussion

5.1 Methodological Interpretation

The convergence of K-Means and Hierarchical Clustering on a five-cluster solution, corroborated by the Elbow inflection and dendrogram branch structure, provides strong multi-evidence support for a five-segment customer typology in this dataset. This finding replicates Wang (2024), who obtained an identical five-cluster structure on a related mall customer dataset, and aligns with the five-segment model reported by Tabianan et al. (2022) in e-commerce behavioral segmentation. The critical empirical anomaly, the extreme dissociation between income and spending behavior in Clusters 4 and 5, is independently confirmed by FCM, which allocates the lowest cross-cluster membership scores precisely to these two clusters, indicating their behavioral profiles are sharply defined and internally cohesive. The FPC of 0.681 compares favorably with the value of 0.62 reported by Hidayat et al. (2020) for five-cluster loyalty segmentation, suggesting that the five-segment partition is well-supported by the fuzzy structure of this dataset. DBSCAN adds a dimension absent from the other three methods by quantifying the extent to which the five-cluster model captures the full population: the noise points represent customers who cannot be assigned to any of the five segments with confidence, and whose marketing treatment should be individualized rather than segment-based.

5.2 Business Implications Per Segment

The Golden Spenders (Cluster 1) represent a high-value retention priority. With mean Annual Income of USD 152,000 and Spending Score of 72.4, this segment responds to premium experiential offerings: curated personal shopping services, exclusive membership tiers, and invitation-only preview events. Digital marketing for this segment should be delivered through personalized email and mobile push notifications rather than mass-broadcast channels, given the demographic preference for direct, respectful communication. The High-Earning Pragmatists (Cluster 2) represent a conversion opportunity of substantial commercial potential. Their elevated income (USD 159,000 mean) combined with a moderate spending score (49.6) suggests a rational, value-oriented decision-making style that is resistant to impulse-promotion but responsive to transparent value articulation, product quality demonstrations, and premium-tier loyalty benefits.

The Low-Budget Conservatives (Cluster 3) are best served through high-perceived-value promotions, bundled discounts, and cost-sensitive communication that does not create aspirational misalignment between the product offering and the consumer's financial self-concept. Digital channels with lower engagement barriers, such as push notifications and SMS campaigns, are cost-effective for this segment. The Thrifty Seniors (Cluster 4), despite above-average income of USD 105,000, exhibit the lowest spending score at 24.1. This profile suggests that financial conservatism is value-driven rather than resource-constrained, and that conversion requires strategies emphasizing utility, necessity, and trusted brand associations rather than novelty or social status. The Middle-Aged Enthusiasts (Cluster 5) represent the highest-priority segment for short-cycle digital marketing campaigns. Their paradoxical combination of the lowest mean income (USD 62,000) and the highest spending score (77.8) identifies them as intrinsically motivated, experiential consumers whose purchase decisions are driven by hedonic value. Flash sales of 24 hours or fewer, limited-edition product releases, and gamified loyalty programs are highly suited to this segment's responsiveness to urgency and exclusivity.

This finding extends the framework of Bakri et al. (2025), who demonstrated that FCM-derived membership degrees in educational marketing can be used to design adaptive, personalized outreach strategies. In the retail context, the FCM cross-membership data for Cluster 5 consumers who also exhibit partial Cluster 1 membership can be operationalized as a premium upgrade pathway: presenting aspirational premium products to high-spending middle-income consumers who already demonstrate elevated luxury orientation.

5.3 Operational Framework for Digital Marketing

The five-segment model translates directly into a tiered CRM architecture. In a Customer Relationship Management database, each customer record can be assigned a primary cluster label derived from K-Means along with a secondary FCM membership vector. The primary label determines the default campaign assignment, while the secondary vector enables dynamic re-assignment when a customer's membership degree in an adjacent cluster exceeds 0.35. This threshold-based mechanism allows the marketing system to respond to life-stage transitions before those transitions are fully reflected in demographic data. DBSCAN-labeled outlier customers should be routed to a dedicated high-touch channel, such as a customer success representative, rather than automated campaigns. The update cycle for cluster re-assignment should be set to quarterly, consistent with the temporal resolution of purchase behavior data in typical mall loyalty program databases.

5.4 Limitations and Future Research Directions

Several limitations constrain the generalizability of these findings. First, the dataset is restricted to three variables; the absence of behavioral variables such as visit frequency, dwell time, and multichannel activity limits the depth of segment profiling. Second, the dataset represents a cross-sectional snapshot, precluding longitudinal analysis of segment migration patterns. Third, the analysis is conducted on a single mall dataset, and the five-cluster structure may not generalize to malls with different size, demographic composition, or geographic context. Fourth, the Hierarchical Clustering was applied to a 50-observation subsample for visualization purposes; while standard for large datasets, the dendrogram may not fully represent the hierarchical structure of all 15,079 records. Future research should incorporate time-series transaction data to enable walk-forward segmentation analysis, include additional behavioral features such as digital interaction metrics and loyalty app activity, and validate the five-cluster framework across multiple mall datasets to assess cross-institutional generalizability. The integration of FCM membership degree vectors with predictive churn models, as proposed by Bakri et al. (2025) in the educational context, presents a particularly promising direction for extending the present framework.

6. Conclusion

This study addressed two research objectives through a multi-algorithm clustering analysis of 15,079 shopping mall customer records. With respect to the first objective, the Elbow Method and hierarchical dendrogram analysis converged on five as the optimal cluster count, and the comparative evaluation demonstrated that K-Means, FCM, Hierarchical Clustering, and DBSCAN yield mutually reinforcing evidence for a five-segment customer typology defined by the Age-Income-SpendingScore feature space. K-Means produced the most interpretable and operationally actionable segment profiles; FCM added membership-degree granularity identifying transitional consumers at cluster boundaries; Hierarchical Clustering provided structural validation without parameter pre-specification; and DBSCAN isolated outlier customers whose individualized behavioral profiles resist generalization to any of the five main segments.

With respect to the second research objective, five operationally distinct customer segments were identified: Golden Spenders (mean age 71, income USD 152,000, spending score 72.4), High-Earning Pragmatists (age 35, income USD 159,000, spending score 49.6), Low-Budget Conservatives (age 37, income USD 66,000, spending score 27.9), Thrifty Seniors (age 74, income USD 105,000, spending score 24.1), and Middle-Aged Enthusiasts (age 53, income USD 62,000, spending score 77.8). The income-spending dissociation between Clusters 4 and 5 constitutes the central empirical finding of this study: the highest income paired with the lowest spending, and the lowest income paired with the highest spending, together demonstrate that income is an insufficient proxy for behavioral purchasing potential.

The three contributions of this study collectively advance the literature on clustering-based retail analytics. Mall operators are recommended to implement the tiered CRM architecture described in Section 5.3, with quarterly segment re-assignment based on FCM membership thresholds. Future research should prioritize longitudinal segmentation analysis that captures customer migration patterns across the five behavioral tiers, and should validate the proposed operational framework in controlled field experiments where segment-targeted campaigns can be evaluated against matched-control baselines. The integration of density-based outlier identification into routine CRM operations represents a particularly immediate practical priority, given that DBSCAN noise-classified customers represent a commercially underserved population whose individualized treatment potential has been systematically neglected in mass-marketing approaches.

Acknowledgements: The authors thank the Department of Digital Business, Universitas Negeri Makassar, for supporting student research activities in the Machine Learning course.

Funding Statement: This research received no specific funding from any external source.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Contribution: Salsabila: Conceptualization, supervision, writing (review and editing), final approval. Shasy Due Mahardika: Data analysis, methodology, writing (original draft). Arnisyah: Data curation, visualization, writing (original draft). Rizal Bakri: Methodology, critical revision, supervision, final approval. Farida Islamiah: Literature review, writing (review and editing), final approval.

References

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 21(3), 527-570. https://doi.org/10.1007/s10257-023-00640-4

Bakri, R., Sobirov, B., Astuti, N. P., Ahmar, A. S., & Singh, P. K. (2025). A new framework for dynamic educational marketing segmentation in student recruitment: Optimizing Fuzzy C-Means with metaheuristic techniques. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 9(3), 659-669. https://doi.org/10.29207/resti.v9i3.6515

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203. https://doi.org/10.1016/0098-3004(84)90020-7

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231). AAAI Press.

Hidayat, S., Rismayati, R., Tajuddin, M., & Merawati, N. L. P. (2020). Segmentation of university customers loyalty based on RFM analysis using fuzzy c-means clustering. Jurnal Teknologi dan Sistem Komputer, 8(2), 133-139. https://doi.org/10.14710/jtsiskom.8.2.2020.133-139

Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. Pattern Recognition Letters, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. Analytics, 2(4), 809-823. https://doi.org/10.3390/analytics2040042

Kaggle. (2020). Shopping Mall Customer Segmentation Dataset. https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson Education.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281-297). University of California Press.

Salminen, J., Mustak, M., Sufyan, M., & Jansen, B. J. (2023). How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. Journal of Marketing Analytics, 11(4), 677-692. https://doi.org/10.1057/s41270-023-00235-5

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243

Wang, Y. (2024). Research on the mall customers segmentation based on K-means and DBSCAN. In X. Li, C. Yuan, & J. Kent (Eds.), Proceedings of the 7th International Conference on Economic Management and Green Development (ICEMGD 2023). Applied Economics and Policy Studies. Springer. https://doi.org/10.1007/978-981-97-0523-8_129

Zhou, K., & Yang, S. (2020). Effect of cluster size distribution on clustering: A comparative study of k-means and fuzzy c-means clustering. Pattern Analysis and Applications, 23(1), 455-466. https://doi.org/10.1007/s10044-019-00783-6