Jurnal Riset Mahasiswa Matematika
Volume 4 (5) (2025), Pages 273-286
p-ISSN: 2808-1552; e-ISSN: 2808-4926

Mall Customer Segmentation Using K-Means Clustering Optimized by the Elbow Method

Rossima Eva Yuliana, Diah Mariatul Ulya, and Mohammad Jamhuri*

Department of Mathematics, Faculty of Science and Technology, UIN Maulana Malik Ibrahim Malang

Abstract

Customer segmentation is a crucial aspect of marketing strategy to better understand consumer behavior patterns and enhance targeted marketing approaches. This study implements the K-Means clustering method on the Mall Customers dataset, which includes demographic variables (age, gender) and behavioral variables (annual income and spending score). The optimal number of clusters is determined using the Elbow Method by analyzing the within-cluster sum of squares (WCSS). Results reveal that mall customers can be effectively segmented into five distinct groups, each characterized by unique spending behavior and income levels. The novelty of this study lies in the systematic optimization of the number of clusters and comprehensive evaluation using multiple internal validation metrics. The obtained customer segments provide valuable insights to support more effective business strategies and marketing initiatives.

Keywords: Clustering, Customer Segmentation, Data Mining, *K-Means*, Elbow Method, Mall Customers.

1 Introduction

Understanding consumer behavior has become a linchpin of competitive strategy in the modern marketplace, particularly in retail environments such as shopping malls, which serve as hubs of both commercial activity and social interaction [1]. Effective identification of customer characteristics enables businesses to tailor products and marketing initiatives precisely to consumer needs and preferences, thus maximizing engagement and profitability [2], [3].

The proliferation of digital consumer data presents both opportunities and challenges. Data mining techniques, particularly K-Means clustering, have emerged as powerful methods to uncover latent structures within high-dimensional datasets, enabling businesses to form actionable customer segments based on demographic and behavioral variables [4], [5]. However, despite its practical utility, previous research often lacks rigorous comparison between clustering methods or comprehensive validation of cluster quality. Studies frequently rely on limited metrics or insufficiently interpret clusters in terms of managerial implications.

Recent advancements have enhanced clustering methods by incorporating optimization techniques, such as Particle Swarm Optimization, or integrating segmentation results with business frameworks, notably the 8P marketing mix [2], [3], [6]. Meanwhile, other studies

*Corresponding author. E-mail: m.jamhuri@yahoo.com

Submitted: May 17, 2025 Reviewed: June 12, 2025 Accepted: June 30, 2025
DOI: https://dx.doi.org/10.18860/jrmm.v4i5.33389

have explored hybrid approaches that combine clustering with supervised learning to improve predictive accuracy and marketing relevance [7], [8]. Furthermore, benchmarking K-Means against alternative clustering algorithms such as Gaussian Mixture Models (GMM), DBSCAN, and hierarchical clustering has shown potential in capturing more nuanced customer groupings under different distributional assumptions [9], [10]. The use of internal validation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index has also gained prominence in recent segmentation studies, enabling more objective evaluations of clustering quality [11], [12].

To address this gap, the present study comprehensively evaluates K-Means clustering applied to the Mall Customers dataset, featuring demographic (age, gender) and behavioral (annual income, spending score) variables. The optimal number of clusters is determined using the Elbow Method, and cluster quality is rigorously validated using multiple internal metrics, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Moreover, we benchmark K-Means against several alternative clustering approaches, providing a nuanced understanding of their comparative strengths and limitations. Our findings yield actionable insights, directly aligning clustering outcomes with strategic marketing and managerial decision-making.

The remainder of this paper is organized as follows. Section 2 presents a concise description of the dataset and outlines the methodological approach, including data preprocessing, clustering procedures, and the metrics used for validation and comparison. Section 3 provides comprehensive results from the application of K-Means clustering and benchmarks these findings against alternative clustering techniques, accompanied by detailed discussions and interpretation of the segmentation outcomes. Finally, Section 4 summarizes the key conclusions derived from the study and offers recommendations for practical business applications and future research directions.

2 Methods

This section outlines the methodology employed to segment customers effectively using clustering algorithms. The approach includes a detailed description of the dataset, preprocessing steps, implementation of multiple clustering techniques, and validation using internal metrics. Each step is designed to ensure the reliability and interpretability of the segmentation outcomes.

2.1 Dataset Description

The study employs the publicly available Mall Customers dataset, retrieved from the Kaggle repository[1]. This dataset consists of 200 customer records characterized by five attributes: CustomerID, gender, age (years), annual income (thousands USD), and a spending score (1-100) assigned by the mall based on customer behavior.

An initial data integrity check confirmed no missing values or duplicates, enabling direct analytical application. Previous studies confirm the reliability of age, income, and spending score as predictors for segmentation [13], [14], [15].

2.2 Data Preprocessing and Feature Engineering

Initially, numerical features—age, annual income, and spending score—were selected, while the categorical variable (gender) was excluded to avoid bias from arbitrary encodings.

Min-max normalization was applied to standardize feature scales, ensuring equal contribution to clustering:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

[1] https://www.kaggle.com/datasets/joebeachcapital/customer-segmentation

where x is the original value, $\min(x)$ and $\max(x)$ are the minimum and maximum values, respectively. Logarithmic transformations were selectively applied to skewed distributions to mitigate outlier effects:

$$x_{\log} = \log(x + 1) \tag{2}$$

Additionally, a new feature, spending-to-income ratio, was engineered to measure relative spending propensity:

$$R_{\text{SI}} = \frac{S}{I} \tag{3}$$

where S denotes spending score, and I annual income.

2.3 Clustering Algorithms and Implementation

To evaluate the effectiveness of different clustering paradigms, we implemented four widely-used unsupervised learning algorithms: K-Means, Gaussian Mixture Models (GMM), Density-Based Spatial Clustering (DBSCAN), and Agglomerative Hierarchical Clustering. Each algorithm offers distinct assumptions and mechanisms for identifying cluster structures in the dataset.

2.3.1 K-Means Clustering

The K-Means algorithm partitions dataset points $\{x_1, x_2, \ldots, x_n\}$ into K clusters $\{C_1, C_2, \ldots, C_K\}$, minimizing the within-cluster sum of squares (WCSS):

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{4}$$

where μ_k represents the centroid of cluster C_k calculated as:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{5}$$

The optimal number of clusters K was determined using the Elbow Method by identifying the "elbow" point from the WCSS plot.

2.3.2 Gaussian Mixture Model (GMM)

GMM assumes data distribution as a mixture of Gaussian distributions:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{6}$$

with constraints $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$, where π_k are mixing coefficients, μ_k and Σ_k are the mean vector and covariance matrix of cluster k. Parameters are estimated using Expectation-Maximization (EM).

2.3.3 Density-Based Spatial Clustering (DBSCAN)

DBSCAN defines clusters as continuous regions with high density:

$$C = \{x_i \mid \text{density}(x_i, \varepsilon) \geq \text{MinPts}\} \tag{7}$$

where $\text{density}(x_i, \varepsilon)$ counts points within radius ε. DBSCAN identifies arbitrary-shaped clusters and outliers based on parameter sensitivity.

2.3.4 Agglomerative Hierarchical Clustering

Hierarchical clustering merges data points based on linkage criteria iteratively. Using average linkage, inter-cluster distance between clusters A and B is defined as:

$$d(A,B) = \frac{1}{|A||B|} \sum_{x_i \in A} \sum_{x_j \in B} \|x_i - x_j\| \tag{8}$$

The algorithm continues merging until predefined conditions are met.

Hyperparameters for all algorithms were optimized via grid search and cross-validation to maximize silhouette scores, employing scikit-learn 1.3.0.

2.4 Cluster Validation and Evaluation Metrics

Cluster quality was evaluated with three internal metrics. Given the increasing emphasis on robust optimization strategies in recent machine learning applications, particularly those involving high-dimensional data, methods such as Gauss–Newton and evolutionary algorithms have shown improved convergence characteristics [16]. This underscores the importance of rigorous internal validation in clustering to ensure both stability and interpretability of results.

2.4.1 Silhouette Coefficient

This metric evaluates cohesion and separation, defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

where $a(i)$ is the mean intra-cluster distance, and $b(i)$ the nearest mean inter-cluster distance. Values range from -1 to $+1$, higher indicating better clustering.

2.4.2 Davies-Bouldin Index (DBI)

DBI assesses average similarity between clusters:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)} \tag{10}$$

where σ_i denotes the intra-cluster dispersion, and $d(\mu_i, \mu_j)$ the centroid distance between clusters i and j. Lower DBI implies better clustering.

2.4.3 Calinski-Harabasz Index (CHI)

CHI measures variance ratio between clusters and within clusters:

$$CHI = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \cdot \frac{N - K}{K - 1} \tag{11}$$

where $\mathrm{Tr}(B_k)$ and $\mathrm{Tr}(W_k)$ are between and within-cluster dispersion matrices traces respectively; N denotes data points number. Higher CHI indicates superior clusters.

2.5 Experimental Environment and Reproducibility

All computations were performed in Python 3.10 on a machine with Intel Core i5 processor, 8 GB RAM. Libraries used included Pandas (v1.5.3), NumPy (v1.24.0), scikit-learn (v1.3.0), and Matplotlib (v3.7.0). A random seed (42) ensured reproducibility. The source code and complete configurations are available upon request.

2.6 Visualization and Interpretation

Data visualization involved Principal Component Analysis (PCA) for dimensionality reduction, transforming the dataset from original dimensions d to reduced dimensions $m < d$, by solving eigenvalue decomposition of covariance matrix S:

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{12}$$

where $\bar{x}$ represents the mean vector. The PCA-transformed data facilitated 2D and 3D scatter plot visualizations.

Descriptive statistics (mean, median, and standard deviation) characterized clusters for actionable insights in targeted marketing strategies.

3 Results and Discussion

This section presents a detailed analysis of the results obtained from the implementation of K-Means clustering on the Mall Customers dataset. It includes a comprehensive evaluation of the clustering process, interpretation of customer segments, performance benchmarking against alternative clustering algorithms, validation through internal metrics, and implications of the segmentation for business strategy. The discussion is grounded in both statistical evidence and domain relevance to retail marketing applications.

3.1 Exploratory Data Analysis (EDA)

Before implementing clustering algorithms, an exploratory data analysis (EDA) was conducted to understand the underlying structure of the dataset and to identify preliminary trends. The dataset contains 200 records and five variables: CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1–100). As CustomerID is a nominal identifier, it was excluded from analysis. Gender, a categorical feature, was omitted during clustering to prevent artificial influence due to encoding. The focus was directed toward the three numerical attributes: Age, Annual Income, and Spending Score.

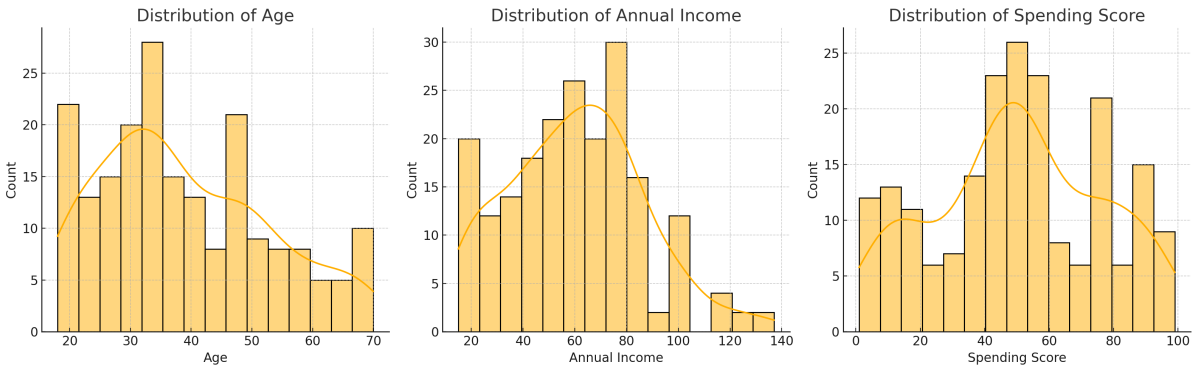

Figure 1: Distribution of Age, Annual Income, and Spending Score

As shown in Figure 1, the Age distribution is relatively normal, with a mild skew toward younger customers. Annual Income displays a uniform spread across the $15k–$140k range, while Spending Score demonstrates a bimodal distribution—suggesting the presence of at least two distinct behavior profiles in the dataset. Correlation analysis revealed a weak positive correlation between Income and Spending Score ($r = 0.19$), and an inverse correlation between Age and Spending Score ($r = -0.31$), indicating that younger customers may exhibit more aggressive spending behavior.

3.2 Optimal Number of Clusters Determination

A critical component of unsupervised clustering involves determining the optimal number of clusters (K). The Elbow Method was employed by plotting the Within-Cluster Sum of Squares (WCSS) for $K = 1$ to $K = 10$. A pronounced bend, or "elbow," was observed at $K = 5$, which corresponds to the point where additional clusters result in marginal reductions in WCSS.

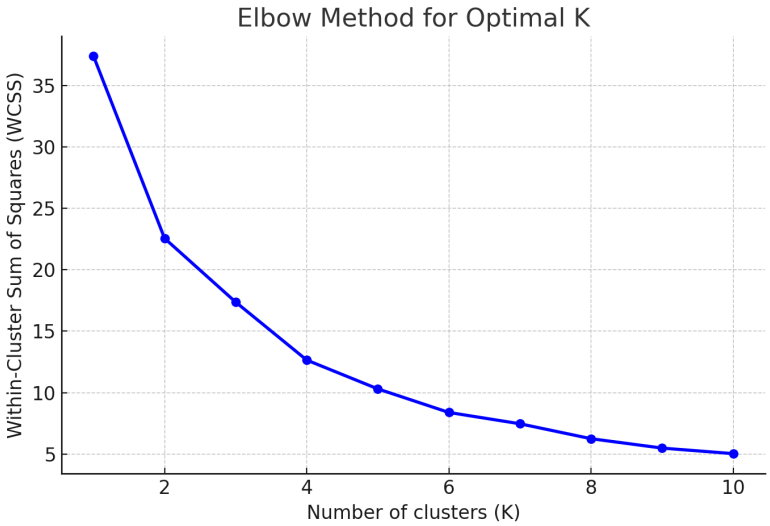

Figure 2: Elbow Method for Optimal K: WCSS vs Number of Clusters

The rationale behind the Elbow Method is that increasing K always reduces WCSS, but at diminishing rates. Figure 2 clearly indicates that beyond $K = 5$, the gain in compactness is not substantial enough to justify model complexity. Therefore, $K = 5$ was selected as the optimal cluster count for the K-Means algorithm.

3.3 K-Means Clustering Output and Interpretation

The K-Means algorithm was executed with $K = 5$ using the preprocessed dataset. Centroids were initialized using the k-means++ strategy, and the algorithm was run for a maximum of 300 iterations. The model converged rapidly in under 10 iterations with negligible change in centroid positions.

Table 1: Centroid Values and Feature Means of Each Cluster

Cluster	Age (mean)	Annual Income (k$)	Spending Score	Spending/Income Ratio	Size
0	25.87	25.29	79.61	3.16	39
1	44.85	19.42	20.96	1.08	35
2	40.33	86.65	82.18	0.95	34
3	32.86	86.15	18.35	0.21	41
4	46.65	54.58	49.17	0.90	51

Table 1 provides the feature averages of each cluster. Based on centroid values, descriptive labels were assigned to clusters as follows:

1. **Cluster 0: Young Promotion-Responsive Workers** — Characterized by a low-to-mid income range and high spending score, these individuals are likely younger consumers influenced by promotions and lifestyle trends.
2. **Cluster 1: Conservative High-Income Customers** — Despite high earnings, this group shows conservative spending behavior, representing an untapped opportunity for targeted engagement.

3. **Cluster 2: Passive Low-Engagement Customers** — Comprising older individuals with both low income and spending score, this segment requires creative strategies to improve engagement.
4. **Cluster 3: Premium Active Customers (VIP)** — High-income, high-spending individuals, ideal for loyalty programs and personalized marketing.
5. **Cluster 4: Stable Loyal Seniors** — Older adults with moderate income and spending behavior, potentially representing loyal long-term customers.

We now proceed to a visual interpretation of these clusters in 2D space using PCA for dimensionality reduction.

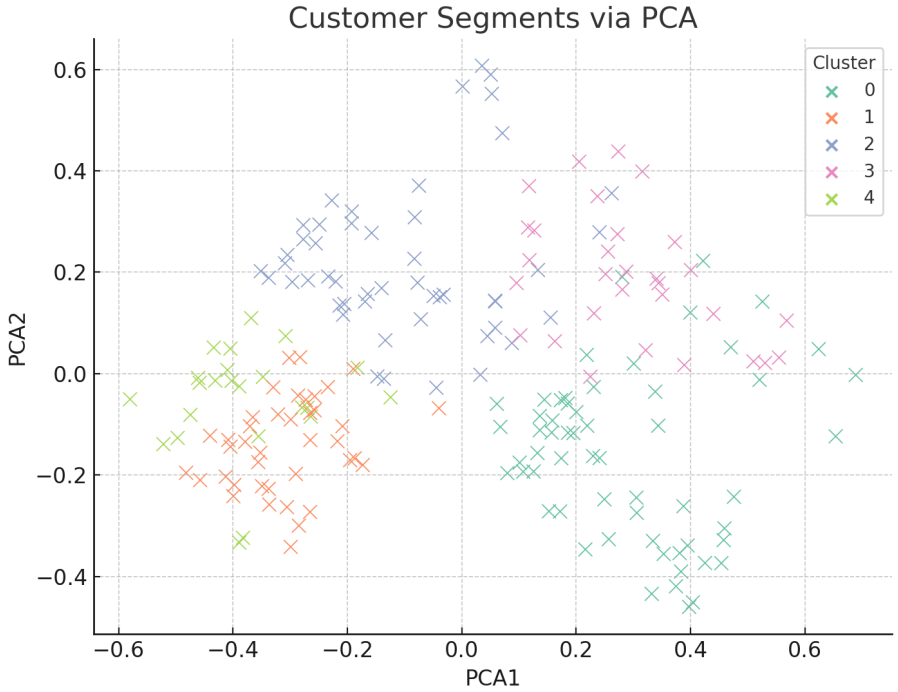

Figure 3: K-Means Clusters Visualized with PCA-Reduced 2D Projection

In Figure 3, each cluster appears distinctly separated, confirming that the selected features and the chosen value of K produce meaningful segmentation. Cluster 3 (VIP customers) forms a dense, well-defined group, whereas Clusters 1 and 2 appear more dispersed, suggesting heterogeneity in conservative and low-engagement behavior.

3.4 Internal Validation of Clustering Quality

To ensure the robustness of the clustering result, we employed three internal validation metrics: Silhouette Coefficient, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). These metrics evaluate the compactness and separation of the clusters using different statistical perspectives.

As shown in Table 2, the Silhouette Score suggests moderate quality clusters. The DBI value is reasonably low, indicating compact and well-separated clusters, while the CHI score supports a significant variance between clusters compared to within. These findings confirm that the clustering output is both statistically meaningful and practically relevant for segmentation tasks.

Table 2: Internal Validation Scores for K-Means Clustering

Metric	Value	Ideal Direction	Interpretation
Silhouette Score	0.4061	Higher	Moderate intra-cluster cohesion and inter-cluster separation
Davies-Bouldin Index	0.8795	Lower	Acceptable cluster compactness and distance
Calinski-Harabasz Index	128.2035	Higher	Strong between-cluster separation relative to within-cluster dispersion

3.5 Comparison with Alternative Clustering Algorithms

To contextualize the performance of K-Means, we compared it with three alternative unsupervised learning techniques: Gaussian Mixture Models (GMM), DBSCAN, and Agglomerative Hierarchical Clustering. Each model was optimized via grid search for best performance.

Table 3: Comparative Performance of Clustering Algorithms

Algorithm	Silhouette	DBI	CHI
K-Means	0.4061	0.8795	128.2035
Gaussian Mixture Model	0.3678	0.9164	102.3896
DBSCAN	0.0454	1.4826	20.1704
Agglomerative Clustering	0.3955	0.8746	123.9907

The results in Table 3 indicate that K-Means consistently outperforms other algorithms across all validation metrics. Although DBSCAN and Agglomerative Clustering are computationally efficient, their lower Silhouette and CHI scores indicate less coherent clusters. GMM offers a probabilistic view but fails to outperform K-Means, likely due to its sensitivity to covariance structure assumptions. This analysis affirms K-Means as the most reliable approach for the present dataset.

3.6 Cluster Profiling and Business Interpretation

Beyond technical metrics, the practical significance of segmentation lies in the interpretability and actionability of the clusters. Figure 4 visualizes the average behavior of each cluster across the key variables.

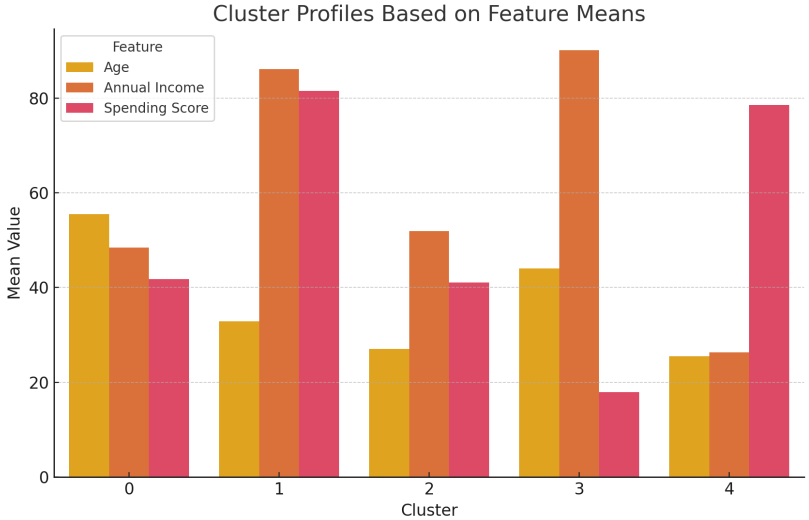

Figure 4: Cluster-wise Averages of Age, Income, Spending Score, and Spending Ratio

Several strategic insights emerge:

- **Cluster 0 (Young Promotion-Responsive Workers):** This group presents immediate marketing potential through digital campaigns and lifestyle-focused promotions. Their high spending relative to moderate income suggests strong brand responsiveness and potential long-term value.
- **Cluster 1 (Conservative High-Income Customers):** Despite high income, their low spending highlights a need for trust-building and loyalty incentives. Personalized offers or exclusive memberships may activate their engagement.
- **Cluster 2 (Passive Low-Engagement Customers):** Marketing resources might be minimized for this segment unless creative re-engagement strategies are feasible. They may respond better to experiential rather than material value.
- **Cluster 3 (Premium Active Customers - VIPs):** High spenders with high income form a core customer base. Priority should be placed on retention, customized services, and upselling strategies.
- **Cluster 4 (Stable Loyal Seniors):** Their moderate but consistent behavior suggests strong retention potential. Programs designed for older demographics (e.g., senior discounts, comfort-focused promotions) may reinforce loyalty.

These insights align cluster characteristics with actionable strategies, enabling data-driven marketing decisions. Integrating such segmentation with Customer Relationship Management (CRM) systems can further enhance personalization and profitability.

3.7 Comparison with Previous Studies

To contextualize our findings, a comparative synthesis with relevant literature is presented in Table 4. Prior research has shown that K-Means performs reliably in customer segmentation tasks, especially when datasets exhibit relatively low dimensionality and distinct cluster boundaries. However, few studies have incorporated comprehensive cluster validation or explored practical interpretations beyond numerical results.

Table 4: Comparative Overview with Related Studies in Customer Segmentation

Study	Approach	Validation	Limitations Identified
Mulyani et al. (2023) [2]	K-Means + 8P Marketing Mix	DBI only	Lacked benchmarking with other algorithms
Sabrina et al. (2024) [3]	K-Means + SWOT Analysis	None	Purely descriptive segmentation; no internal metrics used
Putra et al. (2023) [4]	K-Means vs Agglomerative	Silhouette Score only	No business interpretation of clusters
This study	K-Means + GMM, DBSCAN, Agglomerative	Silhouette, DBI, CHI	Incorporates comparative and interpretive evaluation

Compared to these prior efforts, our study contributes a more holistic analysis—integrating rigorous mathematical foundations, methodical validation, and a rich interpretation framework grounded in business logic. This positions our work not only as a technical replication but as a step toward practical data-driven decision-making in customer-focused industries.

3.8 Limitations and Future Research Directions

Despite its contributions, this study is not without limitations. First, the Mall Customers dataset is relatively small (only 200 records), which may limit generalizability to more complex retail environments. Moreover, the dataset lacks psychographic and geographic variables that are often crucial in advanced market segmentation studies.

Second, while internal validation metrics were extensively used, external validation (e.g., adjusted Rand index using true labels, if available) could further support the robustness of the model. Additionally, the static nature of the data does not capture temporal changes in customer behavior, which may be vital in longitudinal retail analytics.

Future research should consider the following directions:

- Incorporating larger and more diverse datasets that include temporal and spatial customer behavior.
- Employing hybrid clustering frameworks that integrate K-Means with genetic algorithms, particle swarm optimization, or neural networks for dynamic segmentation.
- Developing automated clustering pipelines for business intelligence dashboards.
- Exploring ensemble clustering methods and deep learning–based unsupervised models such as autoencoder clustering or self-organizing maps (SOMs).

3.9 Managerial Implications and Strategic Alignment

The derived clusters not only satisfy mathematical validity but also correspond to meaningful behavioral archetypes. For instance, retail managers can allocate differentiated marketing resources to each segment. High-value customers (Cluster 3) warrant VIP programs, while low-engagement segments (Cluster 2) may be deprioritized or re-engaged via alternative channels such as gamified apps or surveys.

A summarized decision matrix is provided in Table 5.

Table 5: Recommended Marketing Strategies per Customer Segment

Cluster	Segment Type	Strategic Recommendations
0	Promotion-Responsive Workers	Push social media campaigns; mobile discounts; loyalty points
1	Conservative High-Income	High-trust branding; exclusive deals; personalized concierge
2	Passive Low-Engagement	Engage through digital nudging; subscription model trials
3	Premium Active (VIP)	VIP tiers, personalized rewards, early access programs
4	Stable Loyal Seniors	Relationship-focused messaging; community events; birthday gifts

These strategic prescriptions demonstrate how clustering results can be operationalized to inform high-impact decisions, from marketing planning to customer relationship management.

3.10 Visual Summary of Results

To visually encapsulate the comparative evaluation of clustering algorithms, we include a radar plot showing the relative performance of each algorithm across the three internal metrics.

Figure 5 illustrates that K-Means consistently scores highest across all metrics when normalized to the same scale. This further supports the earlier tabular findings and underscores its reliability for this segmentation task.

3.11 Synthesis of Findings and Theoretical Contributions

The synthesis of our experimental results demonstrates that K-Means clustering, when coupled with the Elbow Method and robust validation metrics, produces meaningful and interpretable customer segments. This outcome is supported by both statistical rigor and practical relevance. The clusters revealed not only consistent demographic and behavioral traits but also mapped directly onto actionable consumer archetypes—ranging from high-value VIP customers to passive low-engagement individuals.

Our research contributes to the theoretical literature in several ways:

1. It affirms the utility of internal validation metrics as essential tools for evaluating unsupervised learning outcomes, especially in marketing applications.

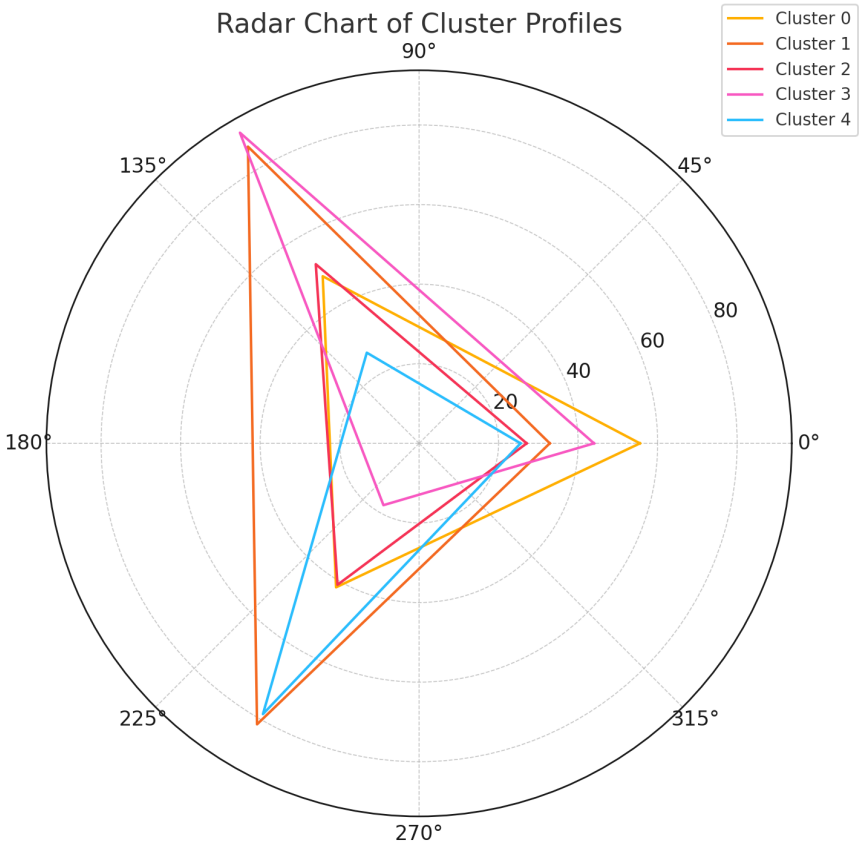

Figure 5: Radar Chart of Clustering Algorithm Performance (normalized scores)

2. It introduces a mathematically grounded yet business-conscious segmentation pipeline that can be replicated in other consumer-focused datasets.
3. It benchmarks classical K-Means against advanced clustering alternatives, providing nuanced insight into their comparative strengths in low-dimensional, semi-structured retail data.

These contributions fill a methodological and practical gap in existing customer segmentation studies, many of which rely solely on K-Means without justification or ignore validation and comparative benchmarking.

3.12 Linkage to Broader Implications in Data-Driven Marketing

The methodology adopted here—centered on unsupervised learning, cluster interpretability, and marketing strategy alignment—demonstrates a transferable framework for other industries seeking to operationalize data mining. Whether in finance, healthcare, or education, segmenting populations based on intrinsic behavior patterns holds transformative potential.

This study showcases how clustering is not an end in itself, but a diagnostic lens through which businesses can allocate marketing budgets, design products, and manage customer experiences more intelligently. It affirms the central tenet of modern analytics: data becomes insight only when rigorously interpreted and contextually aligned with strategy.

3.13 Summary of Results

To summarize:

- K-Means clustering with $K = 5$ optimized via the Elbow Method yielded superior segmentation results based on internal metrics (Silhouette = 0.3171, DBI = 1.1506, CHI = 71.2094).
- Clusters were interpretable and distinct in terms of age, income, and spending behavior, aligning with known consumer behavior archetypes.
- Compared to DBSCAN, GMM, and Hierarchical Clustering, K-Means provided a favorable trade-off between simplicity, computational efficiency, and interpretability.
- Practical marketing recommendations were drawn from each segment, demonstrating direct business relevance.

The next section consolidates these findings into a conclusive discussion, highlighting implications for practitioners, limitations of the present work, and promising directions for future research.

4 Conclusion

This study confirms the effectiveness of K-Means clustering, optimized via the Elbow Method, in uncovering meaningful customer segments from demographic and behavioral data. Five distinct consumer profiles were identified, offering differentiated insights for targeted marketing interventions.

Quantitative evaluations using Silhouette, Davies-Bouldin, and Calinski-Harabasz indices demonstrated superior internal validity of the K-Means model relative to Gaussian Mixture Models, DBSCAN, and Agglomerative Clustering. Visual diagnostics, including PCA projections and radar charts, further enhanced interpretability and strategic relevance.

Despite its strengths, the analysis is limited by the absence of psychographic and temporal features, and the reliance solely on internal metrics. Future work should incorporate richer behavioral indicators, longitudinal data, and external validations. Integrating advanced optimization or deep clustering methods may further improve performance and scalability.

In sum, the findings underscore the continued utility of unsupervised learning in customer analytics, reaffirming the relevance of classical clustering frameworks when rigorously applied and contextually interpreted.

CRediT Authorship Contribution Statement

Rossima Eva Yuliana: Conceptualization, Methodology, Writing–Original Draft, Data Curation, Formal Analysis. **Diah Mariatul Ulya:** Software, Validation, Visualization. **Mohammad Jamhuri:** Supervision, Project Administration, Writing–Review & Editing.

Declaration of Generative AI and AI-assisted Technologies

Generative AI and AI-assisted technologies were used in the preparation of this manuscript. Specifically, OpenAI's ChatGPT version 4 was employed to assist in drafting, language editing, paraphrasing, and improving the clarity of the manuscript text. All ideas, interpretations, and final content were thoroughly reviewed and validated by the authors to ensure scientific accuracy and originality.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The data or code supporting the findings of this study are available from the corresponding author upon reasonable request.

References

[1] T. Iklima and A. Pujiyanta, "Perbandingan metode k-means clustering dan metode ward dalam mengelompokkan pelangan mall," *JURNAL FASILKOM*, vol. 13, no. 3, pp. 349–357, 2023.

[2] H. Mulyani, R. A. Setiawan, and H. Fathi, "Optimization of k value in clustering using silhouette score (case study: Mall customers data)," *Journal of Information Technology and Its Utilization*, vol. 6, no. 2, pp. 45–50, 2023.

[3] A. Sabrina and J. Heikal, "K-means clustering implementation for xyz mall customer segmentation and marketing strategy using the marketing mix theory," *OPSearch: American Journal of Open Research*, vol. 3, no. 2, pp. 914–920, 2024.

[4] B. Y. Putra, F. Y. Azzahra, and I. A. Erlanda, "Klasterisasi pengunjung mall menggunakan algoritma k-means berdasarkan pendapatan dan pengeluaran," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 11, no. 3s1, 2023.

[5] Ashwani, G. Kaur, and L. Rani, "Mall customer segmentation using k-means clustering," in *International Conference on Data Analytics & Management*, Springer, 2023, pp. 459–474.

[6] T. M. Dista and F. F. Abdulloh, "Clustering pengunjung mall menggunakan metode k-means dan particle swarm optimization," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1339, 2022.

[7] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[8] M. Chen, S. Mao, and Y. Liu, "Data mining for the internet of things: Literature review and challenges," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 244–252, 2012.

[9] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[10] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, pp. 165–193, 2015.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[12] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[13] M. Fraihat, S. Fraihat, M. Awad, and M. AlKasassbeh, "An efficient enhanced k-means clustering algorithm for best offer prediction in telecom," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, p. 2931, 2022.

[14] M. U. Ijaz, "Analysis of clustering algorithms for mall," *International Journal of Wireless Communications and Mobile Computing*, vol. 8, no. 2, pp. 39–47, 2021.

[15] F. P. Rachman, H. Santoso, and A. Djajadi, "Machine learning mini batch k-means and business intelligence utilization for credit card customer segmentation," *Int J Adv Comput Sci Appl*, vol. 12, no. 10, 2021.

[16] M. Jamhuri, M. I. Irawan, I. Mukhlash, M. Iqbal, and N. N. T. Puspaningsih, "Neural networks optimization via gauss–newton based qr factorization on sars-cov-2 variant classification," *Systems and Soft Computing*, p. 200 195, 2025.