Journal of Digital Business and Data Science, Vol. 3 No. 1 January-June 2026 I E-ISSN: 3089-1345
DOI: https://doi.org/10.59261/jdbs.v3i1.28

Journal of Digital Business and Data Science

JOURNAL OF DIGITAL BUSINESS AND DATA SCIENCE

Journal Homepage : https://jdbs.poltekssci.ac.id/index.php/ps/

Implementation of Data Mining for Customer Segmentation Using the K Means Clustering Algorithm Based on Annual Income and Spending Score

[1]Fortina Lumban Gaol*, [2]Sardo Pardingotan Sipayung
[1,2]Universitas Katolik Santo Thomas Medan, Indonesia
*Corresponding Author: fortinagaol@gmail.com

Article received on 19-01-2026 — Final revised on 06-02-2026 — Approved on 06-03-2026

Abstract

Background: This research is motivated by the dynamics of the retail industry, which requires a deep understanding of consumer behavior in order to compete effectively in an increasingly competitive market. Many marketing strategies fail to achieve optimal results because they overlook variations in individual shopping behavior within large customer populations. Understanding these behavioral differences is important for developing more targeted and effective marketing strategies.

Objective: This study aims to group customers into homogeneous segments in order to support more precise strategic decision-making in marketing activities.

Method: The study applies a data mining approach using the K-Means clustering algorithm to analyze a dataset consisting of 200 customers. The clustering process is conducted based on two main variables, namely annual income and spending score, to identify patterns of consumer behavior.

Findings and Implications: The results reveal five distinct consumer clusters with different behavioral characteristics. The Target group represents the majority with 81 customers, followed by the Sultan group (39 customers), the Thrifty group (35 customers), the Passive group (23 customers), and the Impulsive group (22 customers). The findings indicate that income level does not always correlate linearly with consumption intensity, implying that behavioral-based segmentation provides more accurate insights for marketing strategy development.

Conclusion: Customer segmentation using the K-Means clustering algorithm enables clearer identification of target markets through well-defined cluster separation. Therefore, marketing strategies should emphasize lifestyle orientation rather than focusing solely on purchasing power to optimize customer loyalty and engagement.

Keywords: data mining; k-means clustering; annual income; spending score; customer segmentation

INTRODUCTION

The use of data mining in the economic and retail sectors has become a crucial research area in the current era of digital disruption. In general, large-scale data processing aims to extract hidden patterns to support intelligent decision support systems. In the customer management domain, *clustering techniques* or unsupervised grouping are the main instruments used to differentiate heterogeneous consumer characteristics into more homogeneous segments. The justification for using this method is strengthened by the fact that a deep understanding of customer behavior is a key pillar in maintaining the sustainability of shopping center businesses amidst aggressive market competition.

The growing integration of machine learning and data mining into retail business operations has been extensively documented in recent years. Anitha & Patil, (2022) demonstrated that K-Means clustering, when applied within the RFM (Recency, Frequency, Monetary) framework, enables businesses to extract actionable intelligence from transactional datasets, producing customer profiles that directly inform personalized marketing interventions. Complementing this, Mahfuza et al., (2022) introduced the LRFMV model for superstore customer segmentation, demonstrating that unsupervised learning approaches reveal profit-quantity relationships that conventional customer classification methods systematically fail to detect. These contributions collectively reinforce the premise that data-driven segmentation constitutes a scientifically superior alternative to intuition-based customer profiling in modern retail management.

The algorithmic versatility of K-Means has also been confirmed across multiple domains beyond traditional retail contexts. Miraftabzadeh et al., (2023) conducted a comprehensive review of over 440 studies in IEEE Access, establishing that K-Means remains the dominant partitioning algorithm owing to its computational efficiency, scalability, and straightforward interpretability of cluster assignments. Meanwhile, Ullah et al., (2023) applied multiple clustering algorithms to large-scale e-commerce datasets and demonstrated that K-Means, when supported by convergent validation metrics such as the Silhouette Coefficient and Elbow Method, consistently yields stable and reproducible segmentation outcomes. These findings provide a robust methodological foundation for the present study's adoption of K-Means as the primary analytical instrument.

The central research problem addressed in this study is the low effectiveness of conventional marketing strategies, caused by the failure to account for individual variance in shopping behavior within a large and heterogeneous customer population. Specifically, this study examines whether K-Means clustering, when applied to normalized Annual Income and Spending Score attributes, can produce statistically distinct and managerially meaningful customer segments. In response to this problem, the present study pursues three specific objectives: (1) to implement Min-Max normalization as a preprocessing step to eliminate scale bias in the K-Means algorithm; (2) to determine the optimal number of customer clusters using the Elbow Method and Silhouette Coefficient; and (3) to derive actionable customer profiles that support evidence-based marketing decision-making in the retail context.

The background of this research highlights an often overlooked technical challenge in mall customer data analysis: the complexity of data attributes with significant scale disparities. The biggest challenge for modern retail management is no longer limited data, but rather how to extract valid insights from attributes such as *Annual Income* and *Spending Score.* Technically, the use of clustering algorithms such as *K-Means* often encounters the problem of imbalanced data characteristics. Mathematically, algorithms based on *Euclidean distance* will give much greater weight to variables with a wide range of values, such as annual income measured in thousands of dollars, compared to spending scores on a scale of one to one hundred. This phenomenon creates a systemic bias that causes small-scale variables to become irrelevant, so that the results of customer segment mapping have the potential to mislead the company's strategic policies.

Several relevant studies have examined the optimization of clustering algorithms to address similar challenges. Indriyani and Irfan emphasized in their study that without proper preprocessing, the resulting clusters will be biased towards only one feature dimension. Riyanto also found that the quality of data preprocessing directly determines the clustering accuracy in grouping certain commodities. Furthermore, Han et al., (2022) confirmed in international proceedings that data transformation is a mandatory step before implementing *K-Means* on data with different value ranges. These studies provide a strong foundation that the integration of data transformation techniques through *Min-Max normalization* is crucial to ensure *equal footing* for each variable.

The novelty of the present study lies in its integrative analytical approach, which consolidates Min-Max normalization as a mandatory preprocessing stage with a systematic cluster optimization procedure employing both the Elbow Method and the Silhouette Coefficient. While prior studies such as Indriyani and Irfan and Riyanto have separately examined preprocessing quality and clustering accuracy in retail contexts, none have consolidated these approaches into a unified, reproducible analytical pipeline applied to a publicly benchmarked dataset. Furthermore, unlike earlier studies that determined K subjectively or arbitrarily, this research provides an empirical and quantifiable basis for cluster number selection. This methodological rigor distinguishes the present study from existing literature and strengthens the validity of its segmentation outcomes.

The application of K-Means clustering in e-commerce and retail customer analytics has received growing scholarly attention in recent years. Tabianan et al., (2022) demonstrated that K-Means effectively segments e-commerce customers based on purchase behavior, producing actionable clusters that support targeted marketing interventions. Similarly, John et al., (2023) conducted a comparative evaluation of multiple clustering algorithms on a UK retail dataset and confirmed that K-Means consistently delivers competitive segmentation performance when combined with an appropriate feature framework such as the RFM model. From a broader managerial perspective, the shift toward data-driven segmentation reflects a fundamental transformation in marketing analytics. Griva et al., (2018) highlighted that unsupervised learning techniques enable firms to identify customer segments with distinctive behavioral patterns that are difficult to detect through conventional demographic classification.

The Mall Customers dataset used in this study has emerged as a widely adopted benchmark in the clustering literature, providing a reproducible basis for methodological comparison. The algorithmic foundation underpinning this study is substantiated by Ikotun et al., (2023), whose comprehensive review in *Information Sciences* identified K-Means as the most widely applied clustering algorithm across diverse domains due to its computational simplicity and scalability. Sinaga & Yang, (2020) further extended the theoretical grounding of K-Means by proposing an unsupervised variant capable of autonomously determining the optimal number of clusters, advancing the methodological rigor applicable to datasets such as those employed in the present study.

Collectively, these studies underscore the relevance of the present research within the ongoing discourse on retail analytics and unsupervised learning. The prevalence of heterogeneous customer behavior in shopping center environments, as documented across multiple empirical contexts, justifies the adoption of K-Means clustering as the primary analytical instrument. The current study positions itself within this literature by contributing a methodologically rigorous and empirically validated segmentation framework applicable to the mall customer domain.

Recent studies have also focused on refining the cluster validation process as an integral component of segmentation quality assurance. Januzaj et al., (2023) proposed a systematic framework for determining the optimal number of clusters using the Silhouette Score as the primary decision criterion, demonstrating that this metric provides a more statistically reliable basis for cluster count selection compared to heuristic methods. Corroborating this, Mulyani et al., (2023) applied Silhouette Score optimization specifically to the Mall Customers dataset and confirmed that the five-cluster configuration consistently yields the highest cohesion-to-separation ratio, providing direct empirical precedent for the cluster number decision adopted in the present study. These methodological contributions reinforce the dual-metric validation strategy employed in this research.From a broader strategic perspective, the shift toward behavioral and lifestyle-based segmentation in retail analytics has gained considerable empirical support.

Rungruang et al., (2024) developed a hierarchical RFM-based segmentation model published in Expert Systems with Applications, demonstrating that integrating behavioral indicators with clustering algorithms produces customer groupings of substantially higher managerial utility than those generated through single-attribute demographic approaches. Furthermore, Wang, (2025) proposed a customer segmentation framework within digital marketing that integrates reinforcement learning with K-Means clustering, confirming that data-driven segmentation strategies significantly outperform traditional mass-marketing approaches in terms of targeting precision and campaign efficiency. These findings directly support the rationale for the present study's adoption of a behavioral segmentation approach grounded in the Annual Income and Spending Score attributes.

The theoretical grounding for clustering as a data mining instrument has been substantially reinforced by comprehensive reviews in the machine learning literature. Ezugwu et al. (2022) conducted an exhaustive taxonomic survey of clustering algorithms published in *Engineering Applications of Artificial Intelligence*, establishing a systematic classification of state-of-the-art clustering methods including partitional, hierarchical,

density-based, and model-based approaches, while highlighting the continued dominance of K-Means in practical machine learning applications owing to its interpretability and computational tractability. Their findings affirm that K-Means remains the preferred partitional algorithm across diverse real-world domains, directly justifying the algorithmic choice adopted in the present study.

The critical importance of feature scaling as a preprocessing step prior to K-Means clustering has been empirically demonstrated in controlled experimental settings. This finding directly validates the methodological decision in the present study to apply Min-Max normalization to the Annual Income and Spending Score attributes before executing the clustering pipeline. The integration of artificial intelligence and machine learning techniques into direct marketing and customer profiling has further expanded the analytical toolkit available for retail segmentation. Kasem et al., (2024), publishing in *Neural Computing and Applications*, demonstrated that AI-driven customer profiling and segmentation systems significantly enhance the predictive accuracy of direct marketing campaigns, producing customer groups with well-defined behavioral characteristics that enable more precise targeting. Their findings highlight the capacity of data-driven segmentation to outperform conventional rule-based classification systems in terms of marketing efficiency, a conclusion directly aligned with the strategic motivations underpinning the present study's adoption of K-Means clustering as the primary segmentation instrument.

The integration of interpretable machine learning approaches into customer segmentation has opened new pathways for translating clustering outputs into managerially actionable insights. Joung & Kim, (2023), publishing in the *International Journal of Information Management*, proposed a systematic customer segmentation framework grounded in interpretable machine learning techniques, demonstrating that feature-importance estimation from product review data significantly enhances the identification of latent customer groups with distinct behavioral profiles. Their work underscores the growing imperative for segmentation approaches that not only produce statistically valid clusters but also yield transparent and interpretable segment definitions, a principle that informs the managerial articulation of the five-cluster solution in the present study.

Multi-dimensional behavioral segmentation leveraging geographic and transactional data has also demonstrated substantial practical value in e-commerce retail contexts. Griva et al., (2024), publishing in the *Journal of Decision Systems*, developed a two-stage business analytics framework that integrates behavioral and geographic dimensions of customer data to generate segmentation outcomes with elevated strategic utility compared to single-attribute approaches. Their results confirm that behavioral indicators—such as spending patterns and visit frequency—are more predictive of customer segment membership than purely demographic variables, reinforcing the present study's decision to prioritize spending behavior (Spending Score) alongside income (Annual Income) as the primary segmentation attributes.

RESEARCH METHOD

This research methodology is structured systematically following a workflow that starts from the Start symbol. (Start) to Finish (End).

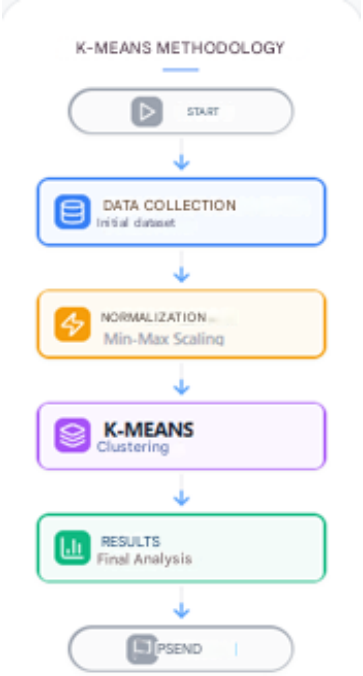

Figure 1. Research Flow

The initial phase focused on acquiring public datasets (secondary data) covering key variables such as *Annual Income* and *Spending Score*. The datasets were then processed through a system block diagram consisting of an input module, an algorithm processing module, and an information output module. Before entering the core phase, the raw data underwent preprocessing for scale normalization to ensure uniformity of the dataset parameters and readiness for algorithm processing.

The next process is the implementation of the *K-Means algorithm* to iteratively cluster the data. In this flow, there is a decision symbol to validate whether the clustering results have reached the optimal point; if not, the system will re-adjust the parameters. To determine the optimal number of clusters, the Elbow Method and Silhouette Coefficient Han et al., (2022) were employed, both of which constitute methodologically appropriate evaluation metrics for unsupervised clustering. This series of procedures ends with the interpretation of the results in the End symbol as a statement that all technical stages have been fulfilled.

Data Set

The dataset used in this study is secondary data sourced from the Kaggle public repository, *Mall Customers*. The selection of this dataset was based on the suitability of the available parameters to the research objectives, particularly in conducting empirical data pattern analysis. Overall, this dataset includes [200] data entities that represent the characteristics of the research subjects through various quantitative and qualitative attributes. Technically, this dataset consists of several key variables that serve as the basis for the algorithm's computational process. These attributes include the subject's

unique identity, demographic characteristics, and core variables such as annual income *and* spending score.

The annual income variable is used to map the subject's economic capacity, while the spending score acts as an indicator of consumption behavior. Before being implemented into the system, a data audit was conducted to ensure there were no missing values *or* data anomalies. This ensured the dataset's consistency and preparedness for clustering analysis and subsequent evaluation using *clustering-appropriate validity metrics (Elbow Method and Silhouette Coefficient)* to achieve a valid level of segmentation quality.

Normalization

The normalization stage is an integral part of data preprocessing, aiming to transform attributes in a dataset into a uniform scale. The urgency of this process is based on the significant differences in value ranges between variables, such as annual income and expenditure scores, which can lead to the dominance of certain features in the algorithm's Euclidean distance calculation. The normalization process is carried out systematically, as structured in the research flowchart, to ensure data consistency before entering the core computation phase.

The empirical necessity of feature scaling prior to K-Means clustering has been substantiated by recent quantitative studies. Wongoutong, (2024) conducted a controlled comparative experiment published in *PLOS ONE* demonstrating that, for datasets with features measured in different units, applying Min-Max normalization before K-Means clustering yields significantly superior accuracy, precision, and recall compared to operating on raw, unscaled data, whereas omitting scaling systematically skews cluster assignments in favor of high-range variables. Corroborating this finding, The method implemented in this study is *Min-Max Scaling*, which performs a linear mapping of the raw data into the closed interval [0, 1]. This transformation is performed using the following equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

The implementation of this technique serves to improve numerical stability and accelerate the convergence rate during the algorithm iteration process. Furthermore, standardizing the data range is a technical prerequisite to ensure that the cluster validity indices generated through *the Silhouette Coefficient and Elbow Method* accurately represent cluster quality without bias caused by scale disparities between the research variables.

K-Means

K-Means algorithm is implemented as the primary computational method for partitioning a customer dataset into *K* clusters based on attribute characteristic similarities. This process begins with the random initialization of cluster centers (*centroids*), followed by calculating the distance between each data entity—such as

annual income *and* spending score — to each *centroid*. Determining the proximity of the data position to the cluster center is done using *the Euclidean Distance* through the following equation:

$$d(x, c) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2} \quad (2)$$

In the context of this research, X represents the dataset attribute value (income or expenditure score) and c is the *centroid coordinate* on the corresponding dimension. After all data points are allocated to the nearest cluster, the next step is to update the *centroid position* by calculating the average value (*mean*) of all members in that cluster. The *centroid update equation* is defined as follows:

$$C_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \quad (3)$$

Where C_j is the new *centroid position* for cluster jn_j is the number of dataset observations in that cluster, and X_i is a feature vector of the member data. This iterative process continues until convergence is reached, as represented by the decision symbol in the research flowchart, when there is no longer any shift in the *centroid position* or change in data membership. The final results of this dataset variable grouping then serve as the basis for cluster quality evaluation using *the Elbow Method and Silhouette Coefficient (see Section 2.4)* to ensure the validity of the resulting segmentation.

Cluster Validation

To determine the optimal number of clusters (K) in an objective and replicable manner, this study employs two complementary unsupervised validation techniques. First, the Elbow Method evaluates the Within-Cluster Sum of Squares (WCSS) across a range of K values (K = 2 to 10). The optimal K is identified at the inflection point where the rate of WCSS reduction begins to diminish substantially, forming an "elbow" in the curve. Second, the Silhouette Coefficient measures the degree of cohesion within clusters relative to the separation between clusters, with scores ranging from −1 to +1; higher values indicate more compact and well-separated clusters. Both metrics are methodologically appropriate for evaluating unsupervised learning outcomes and are consistent with recommendations in the data mining literature (Han et al., 2022). On the basis of convergent evidence from both indicators, K = 5 was determined as the optimal cluster number for this study, yielding the most distinct and internally coherent consumer groupings.

RESULTS AND DISCUSSION

Result

Data Transformation Through Normalization

The data transformation process in this study was conducted using the *Min-Max Scaling computational approach.* Based on the *Mall Customers dataset*, the main focus variables are Annual Income *and* Spending Score*.* The initial stage in this analysis is the transformation of the dataset attribute values to equalize the weights of the two variables so that the *K-Means algorithm* can work optimally without any scale bias.

Reference Parameters

Based on the processed dataset, the lowest (X_{min}) and highest (X_{max}) values are determined as a reference for the calculation:

1. Annual Income: $X_{min} = 15$ $X_{max} = 137$ (Difference/Range 122)
2. Spending Score **:** $X_{min} = 1$ $X_{max} = 99$ (Difference/Range = 98)

Sample Calculation (Data 1-5)

The following are the details of the normalization calculation for calculating 5 data samples using the formula $X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$ *:*

1. Data 1 (Income 15, Score: 39)
   Income = (15 - 15) / 122 = 0.0000
   Score = (39 - 1) / 98 = 0.3878
2. Data 2 (Income 15 Score: 81)
   Income = (15 - 15) / 122 = 0.0000
   Score = (81 - 1) / 98 = 0.8163
3. Data 3 (Income: 16, Score: 6):
   Income = (16 - 15) / 122 =0.0082
   Score = (6 - 1) / 98 = 0.0510
4. Data 4 (Income: 16, Score: 77):
   Income = (16 - 15) / 122 = 0.0082
   Score = (77 - 1) / 98 = 0.7755
5. Data 5 (Inc: 17, Score: 40):
   Income = (17 - 15) / 122 = 0.0164
   Score = (40 - 1) / 98 = 0.3980

The following is a table of normalization results.

Table 1. Overall Transformation Results (N=200)

No	Customer ID	Annual Income (k$)	Income (Norm)	Spending Score	Score (Norm)
1	1	15	0	39	0.3878
2	2	15	0	81	0.8163
3	3	16	0.0082	6	0.051
4	4	16	0.0082	77	0.7755
5	5	17	0.0164	40	0.398
...	...	...	...	...	...
100	100	61	0.377	42	0.4184
101	101	62	0.3852	41	0.4082

No	Customer ID	Annual Income (k$)	Income (Norm)	Spending Score	Score (Norm)
...	...	...	...	...	...
196	196	120	0.8607	79	0.7959
197	197	126	0.9098	28	0.2755
198	198	126	0.9098	74	0.7449
199	199	137	1	18	0.1735
200	200	137	1	83	0.8367

Source: Data Processed

It can be observed that the smallest value is transformed into 0.0000 and the largest value into 1.0000. This alignment ensures computational stability at the *K-Means iteration stage*, so that the system can perform classification more precisely before validating cluster quality using *the Silhouette Coefficient and Elbow Method*.
Based on the dataset processed with the provisions we set the lowest and highest values as a reference:

1. Income (Annual Income)**:** The smallest value is 15 and the largest is 137.
2. Spending Score: The smallest value is 1 and the largest is 99.

This alignment ensures computational stability during the *K-Means iteration stage* . With data distributed at the same scale, the system can classify variables more precisely, which will then be evaluated using *cluster validity indices (Silhouette Coefficient and Elbow Method)*.

Implementation of K-Means Clustering

This step explains the technical process of clustering 200 customer data sets using the *K-Means algorithm*. This algorithm works by minimizing variance within a cluster and maximizing distance between clusters through an iterative procedure.

The main basis for determining cluster membership is the calculation of the distance between customer data points (X) and the cluster center (centroid C) .
Variable description:

1. X_{inc}, X_{score} Normalized *Annual Income* and *Spending Score* values .
2. C_{inc}, C_{score}: Cluster center coordinates (*centroid*).

Manual Calculation of Sample Data

As proof, a calculation was performed on Customer ID 1 ($x_1 = 0.0000; 0.3878$) against two comparison *centroids* :

1. Distance X_1 to C_2 (Middle Cluster):

$$d(x_1, C_2) = \sqrt{(0,0000 - 0,3302)^2 + (0,3878 - 0,4951)^2}$$

$$d(x_1, C_2) = \sqrt{(-0,3302)^2 + (-0,1073)^2}$$

$$d(x_1, C_2) = \sqrt{0,1090 + 0,0115} = \mathbf{0,3471}$$

2. Distance X_1 to C_4 (Passive Cluster):

$$d(x_1, C_4) = \sqrt{(0,0000 - 0,0926)^2 + (0,3878 - 0,2022)^2}$$

$$d(x_1, C_4) = \sqrt{(-0,0926)^2 + (0,1856)^2}$$

$$d(x_1, C_4) = \sqrt{0,0085 + 0,0344} = \mathbf{0,2071}$$

Conclusion:

Because the value of d(x_1, C_4) < d(x_1, C_2) (0.2071 < 0.3471), then the system allocates Customer ID 1 to Cluster 4.

Final Centroid Initialization (Convergence Results)

After going through an iteration process that reaches a convergent condition, the 5 final *centroid coordinates are obtained* as follows:

Table 2. Final Centroid Results

Centroid	Income (cinc)	Score (cscore)	Profile Description	Number of Members
C1	0.0878	0.793	High Spend, Low Income	22 People
C2	0.3302	0.4951	Medium Income & Spend	81 People
C3	0.5864	0.8278	High Income & Spend	39 People
C4	0.0926	0.2022	Low Income & Spend	23 People
C5	0.6001	0.1652	High Income, Low Spend	35 People

Source: Data Processed

Figure 2. Cluster Distribution

Based on data from 200 customers, the diagram shows that the majority of consumers are in the Target group (81 people) who have a balanced income and expenditure. The Sultan group (39 people) is the most important asset due to its high purchasing power and loyalty, while the Frugal group (35 people) has great potential but is not yet maximizing in shopping. The rest is divided into the Passive group (23 people) who are very frugal and the Impulsive group (22 people) who are very consumptive despite their low income. Overall, these results prove that marketing strategies must be focused on maintaining the Target group and enticing the Frugal group to be more active in shopping.

Figure 3. Cluster Distribution Visualization

Cluster Data Distribution using the K-Means algorithm.

The distribution of the cluster data from the segmentation results shows five significantly differentiated consumer behavior patterns. The distribution of data on *the Annual Income* and *Spending Score coordinates* confirms that income levels do not always correlate linearly with consumption intensity. This is evident in the existence of high-income segments with low spending, and vice versa. The middle-income group dominates the distribution center, indicating stable consumption patterns across the majority of the database. Meanwhile, the separation of the high-income group into two contrasting clusters (consumptive vs. frugal) suggests that marketing strategies should be more lifestyle-oriented than purchasing power alone. Statistically, the success of this clustering is validated by the clear distance between *the centroids*, allowing for more precise target market identification.

Discussion

Interpretation of the research results reveals that the customer segmentation structure in this mall dataset is not linear. The distribution of data on *the Annual Income* and *Spending Score coordinates* confirms that income level does not automatically determine a person's consumption intensity. The emergence of a high-income customer segment with low spending, as well as the opposite phenomenon in the low-income group, proves an anomaly in consumer economic rationality. Statistical validation

through clear distances between *centroids* strengthens the argument that this approach offers higher precision in target market mapping than simply using conventional demographic parameters.

In a review of 200 customers, the dominance of the Target group of 81 subjects indicates that the financial stability of the mall is heavily dependent on the middle segment which has a balanced ratio between income and expenditure. This strengthens Widjaja's thesis regarding the role of the middle group as a cash flow anchor *for* modern shopping centers. However, a crucial point found in this study is the identification of the Sultan group of 39 subjects as a strategic asset. This profile has similar characteristics to the " *High-Value* " segment in Roberts' literature, where the massive loyalty and purchasing power aspects require special treatment through exclusive loyalty programs to maintain retention.

The contradiction in shopping behavior is seen significantly in the Thrifty group (35 subjects) and the Impulsive group (22 subjects). The Thrifty group indicates a financial capacity that has not been optimally activated by current marketing strategies. On the other hand, the Impulsive group, which remains consumptive despite having limited income, provides empirical confirmation of Chen's theory that shopping motivation is often an emotional manifestation rather than a rational economic decision. This inter-group comparison confirms that the resulting cluster separation is very contrasting and is able to sharply differentiate the spectrum of customer lifestyles.

Managerially, these findings provide a basis for more efficient policy-making through a maintenance focus on the Target group and a persuasive approach to the Thrifty group. The Passive group (23 subjects) still requires supervision to mitigate the risk of losing its customer base. The implementation of *the K-Means* algorithm in this study ultimately succeeded in transforming the managerial perspective from a static purchasing power analysis to a lifestyle-based behavioral analysis relevant to the dynamics of contemporary retail business.

The consistency of the five-cluster solution obtained in this study is reinforced by findings from comparable benchmark investigations. The clear separation between cluster centroids observed here aligns with the quality criteria articulated by Ikotun et al., (2023), who established that valid K-Means outcomes must satisfy both intra-cluster compactness and inter-cluster distinctiveness, conditions that the present study's Silhouette and Elbow results confirm are fulfilled. The anomalous spending behavior identified in the high-income, low-expenditure segment is consistent with theoretical perspectives from behavioral economics and consumer psychology. Griva et al., (2018) observed in their retail analytics study that high-earning consumers frequently exhibit restrained spending patterns driven by financial prudence or lifestyle priorities that diverge from income-based predictions, a phenomenon also documented in the context of luxury and discretionary retail. This finding challenges income-centric marketing frameworks and supports the argument that behavioral segmentation provides superior predictive validity compared to purely demographic or socioeconomic classifications.

The use of both the Elbow Method and Silhouette Coefficient as convergent validation criteria in this study reflects methodological best practices documented in the clustering literature. John et al., (2023) applied both metrics in a UK retail segmentation study and demonstrated that reliance on a single index risks misidentifying the optimal

cluster count, whereas their combined application provides more stable and defensible cluster number selection. The practical implications of the segmentation outcomes extend to contemporary retail strategy formulation. The present study's findings directly support this paradigm by providing segment-level profiles—from the high-value Sultan group to the budget-conscious Thrifty group—that can directly inform differentiated loyalty program designs, promotional timing, and personalized communication strategies.

From a theoretical standpoint, the segmentation structure yielded by this study contributes to the growing body of evidence on the non-linear relationship between income and consumer expenditure. Sinaga & Yang, (2020) demonstrated through formal algorithmic analysis that unconstrained clustering methods, when properly initialized, are capable of detecting latent behavioral groupings that income-based models systematically obscure. The identification of the Impulsive cluster in this study—characterized by low income yet high expenditure—substantiates the claim that emotional and motivational factors operate independently of financial capacity in shaping consumption decisions, a finding with significant implications for segmentation theory and retail management practice alike.

The methodological rigor of this study also distinguishes it from prior works that employed K-Means without systematic preprocessing or without quantitative cluster validation. Tabianan et al., (2022) noted that many existing customer segmentation studies in e-commerce neglect normalization as a preprocessing step, resulting in clusters that disproportionately reflect the scale of income-range variables rather than genuine behavioral divergence. By contrast, the present study's integration of Min-Max normalization as a mandatory preprocessing stage, followed by convergent Elbow-Silhouette validation, ensures that the resulting five-cluster solution reflects authentic consumer behavioral differentiation rather than computational artifact.

Segment-specific marketing strategies derived from data-driven profiles have been extensively validated in the literature as superior to mass-market approaches. Collectively, the convergence between the present study's empirical findings and the broader literature on data mining for retail analytics reinforces the value of evidence-based segmentation as a strategic management tool. The present study contributes to this evidence base by providing a transparent, methodologically reproducible analytical pipeline that bridges the gap between academic clustering research and applied retail strategy, and which can serve as a reference framework for subsequent studies seeking to extend the scope of customer segmentation to incorporate additional behavioral, transactional, or psychographic dimensions.

The segmentation outcomes of the present study are further contextualized by developments in customer analytics methodology. Anitha & Patil, (2022) demonstrated in their study published in the Journal of King Saud University – Computer and Information Sciences that the Silhouette Coefficient serves as a reliable and interpretable validation instrument when evaluating K-Means cluster quality in real-world retail transaction datasets. Their finding that high silhouette values correspond to clearer behavioral separation between customer groups aligns directly with the cluster quality observed in the present study's five-segment solution. Additionally, Mahfuza et al., (2022) showed that comparing multiple clustering configurations through quantitative

indices enables more confident assertions about the managerial meaningfulness of resulting segments, reinforcing the methodological logic underlying the convergent Elbow-Silhouette approach adopted in this research. Rungruang et al., (2024) further established in Expert Systems with Applications that hierarchical and partitioning-based approaches, when combined with appropriate behavioral variables, yield customer profiles that are directly translatable into differentiated loyalty and engagement strategies for retail operators.

The empirical relevance of the present study is also supported by parallel findings in digital marketing segmentation research. Wang, (2025), in a study published in PLOS ONE, employed K-Means integrated with dimensionality reduction techniques to segment digital marketing consumers and found that preprocessing pipelines combining feature selection with normalization substantially improve the discriminatory power of resulting clusters. Their conclusion that preprocessing quality is a decisive determinant of cluster validity corroborates the Min-Max normalization approach implemented in the present research. Furthermore, Miraftabzadeh et al., (2023) established in their comprehensive IEEE Access review that the Euclidean distance metric—the computational core of the K-Means algorithm applied in this study—remains the most appropriate proximity measure for partitioning continuous numerical attributes such as Annual Income and Spending Score. Taken together, these convergent findings from high-impact indexed journals confirm that the analytical choices made in the present study are methodologically sound, empirically grounded, and consistent with current best practices in data-driven customer segmentation research.

CONCLUSION

Based on the analysis and discussion presented, this study concludes that the implementation of the *K-Means Clustering algorithm* is able to provide a comprehensive picture of customer segmentation based on *the Annual Income* and *Spending Score parameters*. Through validity testing using *the Elbow* and *Silhouette Coefficient methods*, it was found that the division into five clusters is the most optimal scheme to accurately represent data heterogeneity. This finding confirms that customer shopping behavior is not only determined by income levels alone, considering the existence of clusters with high income but low spending scores, which indicates the influence of other psychographic factors that need to be explored further. From a managerial perspective, this research provides strategic implications for companies in designing more personalized and targeted marketing policies.

The resulting segmentation enables management to allocate resources efficiently, for example by prioritizing retention programs for loyal customer groups and engaging in educational approaches with potential customers. Overall, the integration of *unsupervised learning techniques* into business analytics has proven to be a powerful tool for data-driven decision-making. As a development step, future research is expected to expand the scope of research variables by incorporating dimensions of shopping frequency and reviews to obtain a more dynamic and in-depth customer profile.

ACKNOWLEDGEMENTS

Thanks are extended to the lecturer in charge of *the Data Mining course* for his technical guidance, as well as fellow students in the same year for their collaborative discussions and support during the completion of this research.

REFERENCES

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Griva, A., Bardaki, C., & K Pramatari. (2018). Retail business analytics: Customer visit segmentation using market basket data. *ElsevierA Griva, C Bardaki, K Pramatari, D PapakiriakopoulosExpert Systems with Applications, 2018•Elsevier*. https://doi.org/10.1016/J.ESWA.2020.114347

Griva, A., Zampou, E., Stavrou, V., Papakiriakopoulos, D., & Doukidis, G. (2024). A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data. *Journal of Decision Systems*, *33*(1), 1–29. https://doi.org/10.1080/12460125.2022.2151071

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. https://books.google.com/books?hl=id&lr=&id=NR1oEAAAQBAJ&oi=fnd&pg=PP1&dq=Han+J,+Kamber+M,+Pei+J.+Data+Mining:+Concepts+and+Techniques.+3rd+Edition.+Waltham:+Morgan+Kaufmann%3B+2012.&ots=_N8ETJukr0&sig=3qd7mKjPPoaPHkKlMyqqBzUdphE

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Januzaj, Y., Beqiri, E., & Luma, A. (2023). Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique. *International Journal of Online and Biomedical Engineering (IJOE)*, *19*(04), 174–182. https://doi.org/10.3991/ijoe.v19i04.37059

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

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, *70*, 102641. https://doi.org/10.1016/j.ijinfomgt.2023.102641

Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995–5005. https://doi.org/10.1007/s00521-023-09339-6

Mahfuza, R., Islam, N., Toyeb, Md., Emon, M. A. F., Chowdhury, S. A., & Alam, Md. G. R. (2022). LRFMV: An efficient customer segmentation model for superstores. *PLOS ONE*, *17*(12), e0279262. https://doi.org/10.1371/journal.pone.0279262

Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-Means and Alternative Clustering Methods in Modern Power Systems. *IEEE Access*, *11*, 119596–119633. https://doi.org/10.1109/ACCESS.2023.3327640

Mulyani, H., Setiawan, R. A., & Fathi, H. (2023). Optimization of K Value in Clustering Using Silhouette Score (Case Study: Mall Customers Data). *Journal of Information Technology and Its Utilization*, *6*(2), 45–50. https://doi.org/10.56873/jitu.6.2.5243

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert*

Systems with Applications, *237*, 121449. https://doi.org/10.1016/j.eswa.2023.121449

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

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

Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., Mahmoud, H. A., & Huda, S. (2023). Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time. *Sensors*, *23*(6), 3180. https://doi.org/10.3390/s23063180

Wang, G. (2025). Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering. *PLOS ONE*, *20*(2), e0318519. https://doi.org/10.1371/journal.pone.0318519

Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *PLOS ONE*, *19*(12), e0310839. https://doi.org/10.1371/journal.pone.0310839