INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

PREDICTIVE ANALYTICS FOR CUSTOMER SEGMENTATION AND TARGETED MARKETING

Dr. J Josephine Virginia Sharmila
Assistant Professor, Department of Corporate Secretaryship,
Anna Adarsh College for Women (Autonomous)
Chennai 600040
Mail ID : jjosephinevirginia@annaadarsh.edu.in

Dr. G.M.Deivanayagi
Assistant Professor & Head, Department of Corporate Secretaryship
Anna Adarsh College For Women (Autonomous), Chennai 600 040
Mail ID: deivanayagigm@annaadarsh.edu.in

Dr Maya Elizabeth Augustine,
Assistant Professor, PG & Research Department of Commerce
Anna Adarsh College For Women (Autonomous) Chennai 600040
Mail ID mayaelizabeth@annaadarsh.edu.in

Abstract

Customer segmentation is an important component of modern marketing, as customers differ in their demographic characteristics, income levels, and spending behaviour. Traditional marketing approaches often rely on broad customer categories, which may fail to capture meaningful differences among customers. Machine learning and data analytics provide effective techniques for identifying customer groups and developing targeted marketing strategies.

This study presents a data-driven framework for customer segmentation and targeted marketing using K-means clustering and statistical analysis. The analysis uses the publicly available Mall Customer Segmentation dataset, comprising 200 customer records with information on age, gender, annual income, and spending score. After examining the dataset, annual income and spending score were standardized and used as input features for clustering. K-means models with different numbers of clusters were evaluated using inertia, the silhouette coefficient, the Davies–Bouldin index, and the Calinski–Harabasz score. A five-cluster solution was selected based on its overall clustering performance and interpretability, achieving a silhouette coefficient of 0.5547 and a Davies–Bouldin index of 0.5722. The identified segments represent customers with moderate income and spending, high income and high spending, lower income and high spending, high income and low spending, and lower income and low spending. Further statistical analysis indicated significant differences in age across the identified clusters, whereas the association between gender and cluster membership was not statistically significant. Based on these customer profiles, segment-specific marketing recommendations were developed to support

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

personalized promotions, customer engagement, and marketing resource allocation. The findings demonstrate the potential of unsupervised machine learning to identify interpretable customer segments. However, the effectiveness of the recommended marketing strategies requires further validation using actual transaction records and campaign-response data.

Keywords: *Predictive analytics, customer segmentation, targeted marketing, machine learning, K-means clustering, customer profiling, statistical analysis, data-driven marketing.*

1. Introduction

The rapid development of data analytics and machine learning has transformed the way organizations understand customer behaviour and develop marketing strategies. In a competitive business environment, understanding differences in customer characteristics, purchasing preferences, income levels, and spending patterns is essential for effective marketing decision-making. Traditional marketing strategies often treat customers as a relatively uniform group, which may result in less personalized promotional activities and inefficient allocation of marketing resources. Customer segmentation addresses this challenge by dividing customers into meaningful groups based on shared characteristics.

Machine learning techniques provide an effective means of identifying customer segments from available data. In particular, unsupervised learning algorithms can discover hidden patterns in customer attributes without requiring predefined segment labels. K-means clustering is widely used for this purpose because of its computational efficiency, straightforward implementation, and ability to generate interpretable customer groups. However, its performance depends on appropriate feature selection, data preprocessing, feature scaling, and the selection of a suitable number of clusters. Therefore, clustering results should be evaluated using multiple validation measures rather than relying solely on the number of groups identified.

Predictive analytics can support customer profiling by combining data analysis, statistical techniques, and machine learning to identify patterns that inform business decisions. In customer segmentation, demographic characteristics and spending-related variables can help organizations distinguish customers with different economic profiles and levels of spending engagement. Such information can support the development of differentiated marketing strategies, including personalized promotions, loyalty programmes, and customer engagement initiatives. Nevertheless, identifying customer groups alone does not establish whether a particular campaign will increase purchases or improve profitability. Actual marketing outcomes must be evaluated separately using transaction and campaign-response data.

This study uses the publicly available Mall Customer Segmentation dataset, which contains 200 customer records and five attributes: customer ID, gender, age, annual income, and spending score. Annual income and spending score are selected as the primary features for clustering, while age and gender are used to describe the resulting customer groups and examine their statistical characteristics. The dataset does not contain transaction dates,

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

purchase histories, or transaction-level revenue; therefore, Recency, Frequency, and Monetary (RFM) analysis cannot be directly performed using these data.

The proposed approach applies feature standardization and K-means clustering to identify distinct customer segments. Candidate clustering solutions are evaluated using inertia, the silhouette coefficient, the Davies–Bouldin index, and the Calinski–Harabasz score. Based on the comparative evaluation and interpretability of the results, a five-cluster solution is selected. The identified groups represent different combinations of annual income and spending score, providing a basis for developing segment-specific marketing recommendations. Statistical analysis is also conducted to examine age differences across clusters and the relationship between gender and cluster membership.

The main contribution of this study is the application of a reproducible, data-driven clustering framework to customer profiling and targeted marketing recommendations. By linking cluster characteristics with practical marketing actions, the study demonstrates how machine learning can support customer understanding and marketing planning. However, the proposed recommendations should be validated using actual purchase records and campaign-response data before drawing conclusions about customer retention, sales improvement, or marketing effectiveness.

2. Problem Statement

Organizations face challenges in understanding differences in customer income levels and spending behaviour, making it difficult to develop effective and personalized marketing strategies. Applying the same promotional approach to all customers may overlook important variations in spending patterns and result in inefficient allocation of marketing resources. Therefore, identifying meaningful customer groups based on their characteristics is essential for improving customer profiling and supporting targeted marketing decisions.

Traditional customer classification approaches may not adequately reveal underlying patterns among customers with different income and spending levels. Machine learning techniques, particularly K-means clustering, can help address this limitation by grouping customers according to similarities in selected attributes. However, the reliability and usefulness of the resulting segments depend on appropriate feature selection, data standardization, the choice of the number of clusters, and systematic evaluation of clustering quality.

In this study, the Mall Customer Segmentation dataset is used to identify customer groups based on annual income and spending score. The dataset contains 200 customer records with demographic information, annual income, and spending scores, but it does not include transaction histories, purchase dates, or actual revenue data. Consequently, transaction-based measures such as recency, purchase frequency, and monetary value cannot be calculated, and customer retention or campaign effectiveness cannot be directly measured.

The central problem addressed by this research is how to identify and evaluate meaningful customer segments using K-means clustering and statistical analysis, and how to translate the resulting customer profiles into practical, segment-specific marketing recommendations. The

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

study evaluates alternative clustering solutions using established validation metrics and examines differences in age and gender across the identified groups. The resulting framework aims to support data-driven customer profiling and marketing planning, while recognizing that the actual business impact of the recommended strategies requires validation using transaction and campaign-response data.

3. Research Objectives

The main objective of this study is to develop a data-driven framework for customer segmentation and targeted marketing recommendations using machine learning and statistical analysis. The specific objectives are:

1. Analysis of Customer Characteristics: To analyse customer characteristics using age, gender, annual income, and spending score from the Mall Customer Segmentation dataset.
2. Feature Preprocessing & Standardization: To preprocess and standardize the selected features, particularly annual income and spending score, to improve the reliability of distance-based clustering.
3. Identification of Customer Segments: To identify distinct customer segments using the K-means clustering algorithm based on annual income and spending score.
4. Multi-Metric Cluster Selection: To determine an appropriate number of clusters by comparing alternative clustering solutions using inertia, silhouette coefficient, Davies–Bouldin index, and Calinski–Harabasz score.
5. Profile Development & Interpretation: To develop and interpret customer profiles by examining the income and spending characteristics of each identified cluster.
6. Demographic & Statistical Testing: To examine demographic differences among customer segments through statistical analysis of age and the association between gender and cluster membership.
7. Targeted Marketing Formulation: To recommend segment-specific marketing strategies based on the characteristics of the identified customer groups.
8. Critical Assessment & Validation Pathways: To assess the usefulness and limitations of the proposed framework for customer profiling and marketing planning, while identifying the need for transaction and campaign-response data to validate actual business outcomes.

4. Research Questions

- **RQ1**: How can K-means clustering be used to identify meaningful customer segments based on annual income and spending score?
- **RQ2**: What is the appropriate number of customer clusters based on clustering evaluation metrics, including inertia, silhouette coefficient, Davies–Bouldin index, and Calinski–Harabasz score?
- **RQ3**: What are the distinguishing characteristics of the identified customer segments in terms of annual income and spending score?

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

- **RQ4**: Are there statistically significant differences in customer age across the identified clusters?
- **RQ5**: Is there a statistically significant association between gender and customer cluster membership?
- **RQ6**: How can the identified customer profiles be translated into segment-specific marketing recommendations?
- **RQ7:** What are the limitations of using demographic, income, and spending-score data for customer segmentation, and what additional data are required to evaluate actual marketing effectiveness?

5. Literature Review

5.1 Customer Segmentation Using Data Analytics

Customer segmentation is an important marketing technique that divides customers into groups based on similarities in their characteristics, preferences, and behaviour. Data analytics and machine learning have expanded traditional segmentation approaches by enabling researchers to identify patterns that may not be evident through manual classification. Customer segmentation can help organizations understand differences among customer groups and develop more appropriate marketing strategies.

Salminen et al. (2023) conducted a systematic literature review of algorithmic customer segmentation covering 172 relevant articles. Their findings identified K-means as one of the most frequently used clustering algorithms and highlighted the importance of selecting suitable features, evaluating cluster quality, and interpreting the resulting customer segments. The review also emphasized that technical clustering metrics alone may not establish whether the identified segments are useful for practical marketing decisions. The present study follows this data-driven approach by analysing annual income and spending score to identify customer groups with distinct characteristics.

5.2 K-Means Clustering for Customer Segmentation

K-means is an unsupervised machine learning algorithm that groups observations into a predefined number of clusters based on their similarity. The algorithm assigns each observation to a cluster and iteratively updates cluster centroids to minimize within-cluster variation. Its computational simplicity and interpretability make it suitable for customer profiling.

Research on customer segmentation has frequently used K-means to identify groups with different customer characteristics. Salminen et al. (2023) found that K-means was the most frequently used algorithm among the methods examined in their systematic review. However, the performance of K-means depends on feature selection, data scaling, the number of clusters, and the distribution of observations. In the present study, annual income and spending score are standardized before clustering so that differences in numerical scale do not disproportionately influence the distance calculations. Candidate solutions with different cluster counts are then compared to determine an appropriate segmentation structure.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

5.3 Customer Segmentation Using Income and Spending Characteristics

Customer income and spending score can be used to distinguish groups with different economic profiles and spending tendencies. This approach is particularly relevant when the available dataset contains customer-level attributes rather than detailed purchase transactions.

Babu et al. (2024) investigated mall customer segmentation using Fuzzy C-Means clustering and a dataset of 200 customers. Their study identified five customer segments based on income, spending score, and related attributes. The research demonstrated the potential of clustering techniques and visual analytics to reveal distinct customer profiles that can inform marketing decisions. Unlike Fuzzy C-Means, which allows observations to have different degrees of membership in multiple clusters, conventional K-means assigns each customer to one cluster. The present study uses K-means to produce clearly defined customer groups and examines their income and spending characteristics. The results are interpreted as customer profiles rather than direct measurements of purchase frequency, loyalty, or profitability.

5.4 Evaluation of Clustering Performance

Evaluating clustering quality is essential because unsupervised learning does not provide predefined class labels against which the results can be assessed. Multiple validation measures are therefore used to examine the compactness and separation of the resulting clusters.

Salminen et al. (2023) highlighted the diversity of evaluation metrics used in customer segmentation research and identified the need for more systematic assessment and interpretation of clustering results. The present study employs four evaluation measures: Inertia (measures within-cluster sum of squared distances), Silhouette coefficient (evaluates cohesion and separation), Davies–Bouldin index (measures similarity between clusters), and Calinski–Harabasz score (compares between-cluster to within-cluster dispersion). In the experimental analysis, the five-cluster solution achieved a silhouette coefficient of 0.5547 and a Davies–Bouldin index of 0.5722, supporting its selection.

5.5 Statistical Analysis and Interpretation of Customer Segments

Clustering identifies groups of similar observations, while statistical analysis can help examine whether customer characteristics differ across those groups. Descriptive statistics summarize the profiles of individual clusters, whereas inferential tests can investigate relationships involving additional demographic variables. In this study, a one-way analysis of variance is used to evaluate differences in age across the five clusters, while a chi-square test is used to examine the association between gender and cluster membership.

5.6 Research Gap

The reviewed literature establishes the importance of machine learning for customer segmentation and highlights the need for appropriate clustering evaluation and meaningful interpretation. However, several practical research gaps remain:

(1) Need for interpretable customer profiles supporting marketing decisions;
(2) Need for systematic cluster evaluation across multiple validation metrics;

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

(3) Need for demographic analysis alongside cluster assignments;
(4) Need to connect segmentation with actionable marketing recommendations;
(5) Need to acknowledge the limitations of customer-level attributes without transaction records;
(6) Need for empirical external validation. The present study explicitly addresses these gaps.

6. Proposed Methodology

6.1 Research Design

This study adopts a quantitative, data-driven research design to identify meaningful customer segments using machine learning and statistical analysis. The study uses the Mall Customer Segmentation dataset to examine relationships between annual income and spending score. The research workflow follows: Mall Customer Segmentation dataset → Data inspection and preprocessing → Feature selection → Feature standardization → K-means clustering → Cluster validation → Customer profiling → Statistical analysis → Targeted marketing recommendations.

6.2 Dataset Description

The study uses the publicly available Mall Customer Segmentation dataset from Kaggle, containing 200 customer records across five attributes describing customer demographics, annual income, and spending behaviour.

Variable	Description	Analytical Use
CustomerID	Unique customer identifier	Record identification; excluded from clustering
Genre	Customer gender (categorical)	Demographic profiling and chi-square analysis
Age	Customer age in years	Descriptive analysis and comparison across clusters
Annual Income (k$)	Annual income in thousands of dollars	Primary clustering feature
Spending Score (1–100)	Spending score assigned to customer	Primary clustering feature

6.3 Data Preprocessing & 6.4 Feature Standardization

Preprocessing involves dataset inspection, missing-value verification (0 missing values), feature selection (Annual Income & Spending Score), CustomerID exclusion, and feature standardization. Standardizing features ensures that differences in numerical scales do not disproportionately influence distance calculations:

$$Z = (X - \mu) / \sigma$$

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

where X is the original value, μ is the mean, σ is the standard deviation, and Z is the standardized score.

6.5 K-Means Clustering & 6.6 Cluster Evaluation

K-means partitions observations into K clusters by minimizing the within-cluster sum of squared distances J = Σ Σ ||x_i - μ_k||^2. Candidate solutions from K=2 to K=8 were evaluated using four core metrics:

Clusters (K)	Inertia	Silhouette Coeff.	Davies–Bouldin Index	Calinski–Harabasz Score
2	269.6910	0.3213	1.2670	95.67
3	157.7040	0.4666	0.7165	151.34
4	108.9213	0.4939	0.7096	174.60
5	65.5684	0.5547	0.5722	248.65
6	55.0573	0.5399	0.6546	243.09
7	44.8648	0.5281	0.7148	254.62
8	37.1481	0.4567	0.7579	267.91

6.7 Customer Segment Interpretation & Statistical Profiling

The five-cluster solution produced optimal silhouette (0.5547) and Davies–Bouldin (0.5722) scores. Each group was profiled and assigned descriptive labels:

Cluster	Customers	% Share	Mean Age	Mean Income (k$)	Mean Spending	Segment Interpretation
0	81	40.5%	42.72	55.30	49.52	Moderate-income, moderate-spending customers
1	39	19.5%	32.69	86.54	82.13	High-income, high-spending customers
2	22	11.0%	25.27	25.73	79.36	Lower-income, high-spending customers
3	35	17.5%	41.11	88.20	17.11	High-income, low-spending customers
4	23	11.5%	45.22	26.30	20.91	Lower-income, low-spending customers
Total	200	100%	—	—	—	—

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

Inferential Statistical Analysis: One-way ANOVA revealed significant differences in age across clusters ($F(4,195) = 12.37$, $p < 0.001$). Conversely, a Chi-square test of independence showed no significant association between gender and cluster membership ($\chi^2(4) = 2.23$, $p = 0.693$).

7. Data Analysis and Statistical Techniques

7.1 Descriptive Analysis

Descriptive metrics summarize the numerical attributes of the 200 retail customers:

Statistic	Age (years)	Annual Income (k$)	Spending Score (1–100)
Number of customers	200	200	200
Mean	38.85	60.56	50.20
Standard deviation	13.97	26.26	25.82
Minimum	18	15	1
25th percentile	28.75	41.50	34.75
Median	36.00	61.50	50.00
75th percentile	49.00	78.00	73.00
Maximum	70	137	99

7.2 Segment-Wise Analysis & Statistical Testing

Segment analysis highlights key disparities across groups, such as Cluster 1 (high income \$86.54k, high spending score 82.13) versus Cluster 3 (high income \$88.20k, low spending score 17.11), proving income alone is insufficient to predict spending.

Statistical Test	Test Statistic	Degrees of Freedom	p-value	Decision ($\alpha = 0.05$)
One-way ANOVA: Age across clusters	$F = 12.37$	4, 195	$p < 0.001$	Statistically Significant
Chi-square: Gender & cluster membership	$\chi^2 = 2.23$	4	$p = 0.693$	Not Significant

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

8. Results and Discussion

8.1 Dataset Summary & Clustering Evaluation

The evaluation across cluster solutions confirms that K=5 optimizes separation and cohesion. Evaluating inertia alongside silhouette coefficient and Davies–Bouldin index prevents under- or over-segmentation.

8.2 Clustering Evaluation

Table 5. Comparison of clustering performance

Number of clusters ((K))	Inertia	Silhouette score	Davies–Bouldin index	Calinski–Harabasz score
2	269.6910	0.3213	1.2670	95.67
3	157.7040	0.4666	0.7165	151.34
4	108.9213	0.4939	0.7096	174.60
5	**65.5684**	**0.5547**	**0.5722**	**248.65**
6	55.0573	0.5399	0.6546	243.09
7	44.8648	0.5281	0.7148	254.62
8	37.1481	0.4567	0.7579	267.91

Interpretation

The five-cluster solution achieved the highest silhouette score (0.5547) and the lowest Davies–Bouldin index (0.5722) among the tested alternatives. These results suggest that (K=5) provides the strongest balance of cluster cohesion and separation according to these two measures.

The Calinski–Harabasz score was slightly higher at (K=7) than at (K=5). Therefore, the selection of five clusters is based on the combined evidence of multiple validation metrics and the interpretability of the resulting profiles, rather than on every metric identifying the same optimum.

Inertia decreases as the number of clusters increases, which is expected for K-means. It should be interpreted using the elbow curve and in conjunction with the other metrics.

The selected solution provides five interpretable customer profiles for exploratory marketing analysis. Further stability testing across random initializations and independent samples would strengthen confidence in the segmentation.

8.3 Customer Segmentation Findings

The clustering analysis identified five customer groups with different income and spending-score characteristics.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

Table 6. Summary of the five identified segments

Segment	Number of customers	Percentage	Main characteristic
Middle-income, moderate-spending	81	40.5%	Largest segment with moderate spending score
High-income, high-spending	39	19.5%	High income and high spending score
Lower-income, high-spending	22	11.0%	Lower income and high spending score
High-income, low-spending	35	17.5%	High income and low spending score
Lower-income, low-spending	23	11.5%	Lower income and low spending score

Interpretation

The largest segment is the middle-income, moderate-spending group, which represents 40.5% of the dataset. This group may provide opportunities for personalized recommendations and broad customer-engagement initiatives.

The high-income, high-spending group represents 19.5% of customers and may be considered for loyalty-focused strategies. In contrast, the high-income, low-spending group represents 17.5% of customers, suggesting that income by itself is insufficient to identify high spenders. This group may warrant additional research into product preferences and shopping motivations.

The lower-income, high-spending segment demonstrates that a relatively high spending score is not restricted to customers with high annual income. The lower-income, low-spending segment may require further customer research before marketing resources are allocated to promotional incentives.

These interpretations provide a basis for forming marketing hypotheses, but the spending score is not actual expenditure and cannot be used to calculate revenue contribution

8.4 Targeted Marketing Implications

The empirical customer profiles translate directly into actionable targeted strategies:

Customer Segment	Suggested Marketing Strategy	Performance Metric for Future Testing
Middle-income, moderate-spending	Personalized product recommendations & category offers	Incremental purchase rate & engagement
High-income, high-spending	Loyalty rewards, premium services & exclusive access	Repeat purchases & incremental profit

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

Lower-income, high-spending	Relevant value bundles & loyalty benefits	Retention & net profit after promo costs
High-income, low-spending	Product discovery, tailored recommendations & feedback	Incremental conversion vs control group
Lower-income, low-spending	Low-cost engagement experiments & preference research	Incremental response & cost-effectiveness

9. Major Findings

- **1. Five Distinct Segments**: K-means clustering successfully partitioned mall customers into five highly interpretable groups based on annual income and spending score.
- **2. Optimal Mathematical Fit:** The five-cluster solution achieved the highest silhouette coefficient (0.5547) and lowest Davies–Bouldin index (0.5722).
- **3. Age Disparities:** Significant age differences exist across clusters ($F = 12.37$, $p < 0.001$), with younger averages in high-spending groups.
- **4. Gender Independence:** Gender showed no statistically significant relationship with cluster membership ($\chi^2 = 2.23$, $p = 0.693$).
- **5. Actionable Strategies:** Clear marketing pathways were established for each segment, distinguishing high-value retention targets from conversion-opportunity groups.

10. Recommendations

Based on the customer segments identified through K-means clustering, the following recommendations are proposed for practical application and future empirical testing.

1. **Develop segment-specific marketing strategies:** Design differentiated marketing approaches for the five customer segments based on their annual income and spending scores rather than applying a uniform strategy to all customers.
2. **Prioritize high-income, high-spending customers:** Consider personalized product recommendations, premium offers, and loyalty-program incentives for this segment. Their actual retention potential and response to such initiatives should be evaluated using transaction and campaign data.
3. **Encourage spending among lower-income, high-spending customers:** Test affordable product bundles, value-oriented promotions, and suitable discounts to support engagement while considering customers' spending preferences.
4. **Explore opportunities among high-income, low-spending customers:** Use preference-based recommendations, product discovery campaigns, and carefully designed promotional offers to investigate whether their spending can be increased.
5. **Design suitable offers for moderate-income, moderate-spending customers:** Test relevant product recommendations and targeted offers that reflect this segment's observed income and spending characteristics.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

6. **Use appropriate strategies for lower-income, low-spending customers:** Consider accessible pricing and value-focused product options without assuming that low spending indicates customer inactivity or a lack of interest.
7. **Monitor cluster stability:** Reassess the clustering results when updated customer data become available. Compare cluster profiles over time before making substantial changes to marketing strategies.
8. **Evaluate marketing effectiveness using additional data:** Collect transaction histories and campaign-response data to evaluate measures such as conversion rate, repeat purchases, incremental revenue, customer retention, and profit. These outcomes cannot be established from the current dataset alone.
9. **Extend the analytical framework:** Future studies could incorporate purchase frequency, recency, transaction value, product preferences, and campaign-response records. These additional variables would support more comprehensive segmentation and enable the empirical evaluation of marketing strategies.
10. **Protect customer privacy:** Apply appropriate data-minimization, access-control, security, and customer-consent practices when collecting and processing additional customer information.

Overall, the recommendations provide a basis for developing segment-specific marketing strategies. However, they should be treated as **proposed strategies rather than proven improvements in marketing performance**, because the current analysis identifies customer profiles but does not measure campaign outcomes.

11. Conclusion

This study applies a data-driven customer segmentation framework using the Mall Customer Segmentation dataset and K-means clustering. The framework includes data preprocessing, exploratory data analysis, feature standardization, cluster selection and evaluation, customer profiling, and the development of segment-specific marketing recommendations. Annual income and spending score were used as the primary clustering variables, while age and gender were examined to understand additional characteristics of the resulting customer groups.

The analysis identified five distinct customer segments representing different combinations of annual income and spending behaviour. The five-cluster solution achieved a silhouette score of 0.5547 and a Davies–Bouldin index of 0.5722, indicating a useful degree of cluster separation under the selected evaluation criteria. The resulting profiles included moderate-income, moderate-spending customers; high-income, high-spending customers; lower-income, high-spending customers; high-income, low-spending customers; and lower-income, low-spending customers. The statistical analysis also identified significant age differences across clusters, whereas the association between gender and cluster membership was not statistically significant.

These findings demonstrate how unsupervised machine learning can help organizations identify interpretable customer groups and develop differentiated marketing strategies. For example, businesses may investigate premium offers for high-income, high-spending customers and value-oriented promotions for lower-income, high-spending customers.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

However, these strategies are recommendations for future testing rather than demonstrated improvements in customer retention, sales, or profitability.

The study is limited by the dataset's size and the availability of only demographic, income, and spending-score attributes. It does not contain transaction histories, purchase frequency, recency, monetary transaction values, or marketing campaign outcomes. Consequently, the framework cannot perform RFM analysis or establish the actual effectiveness of targeted campaigns. Future research should incorporate transaction-level and campaign-response data, evaluate cluster stability, and use appropriate experimental methods to measure changes in conversion, customer retention, incremental revenue, and profit.

In conclusion, the study provides an empirical foundation for applying K-means clustering to customer segmentation and developing data-informed marketing recommendations. Extending the analysis with richer behavioural data and campaign evaluations would help determine whether these recommendations produce measurable business benefits.

12. Limitations

The study has several limitations that should be considered when interpreting its findings.

1. **Limited dataset size:** The analysis uses 200 customer records from the Mall Customer Segmentation dataset. The relatively small sample may limit the generalizability of the identified segments to other customer populations.
2. **Limited feature availability:** The dataset contains age, gender, annual income, and spending score. It does not include transaction histories, purchase frequency, recency, monetary transaction values, browsing behaviour, or campaign-response information. Therefore, RFM analysis and detailed behavioural profiling cannot be performed.
3. **Limited generalizability:** The dataset represents a specific retail customer sample. The resulting clusters may not represent customers from other retailers, industries, geographic regions, or time periods.
4. **Dependence on K-means assumptions:** K-means uses distance-based grouping and can be sensitive to feature scaling, initial centroid selection, outliers, and the shape of the underlying data. Different clustering algorithms or parameter settings may produce different customer segments.
5. **Dependence on selected features and cluster count:** The segmentation is based primarily on standardized annual income and spending score. The choice of five clusters is supported by the evaluation metrics and interpretability, but alternative feature combinations or clustering methods may produce different results.
6. **Limited evidence of demographic relationships:** Although the analysis identifies statistically significant age differences across clusters, the association between gender and cluster membership is not statistically significant. These results should be interpreted within the limitations of the dataset and the exploratory analysis.
7. **No direct evaluation of marketing effectiveness:** The dataset does not contain campaign records, customer responses, sales outcomes, or profit margins. Consequently, the study cannot establish whether the proposed marketing recommendations improve customer retention, conversion rates, revenue, or profitability.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

Overall, the findings provide useful exploratory customer profiles, but further validation with larger and more comprehensive datasets is required before applying the framework broadly.

13. Future Work

Future research can extend the present study in the following directions.

1. **Compare clustering algorithms:** Evaluate K-means against hierarchical clustering, Gaussian mixture models, and density-based methods such as DBSCAN to determine whether alternative approaches produce more meaningful and stable customer segments.
2. **Incorporate additional customer attributes:** Extend the analysis with relevant demographic, behavioural, and preference-related variables, where available, to develop richer customer profiles.
3. **Integrate transaction-level data:** Include purchase dates, transaction frequency, and monetary transaction values to support RFM analysis and more comprehensive behavioural segmentation.
4. **Develop predictive models:** Apply supervised learning to predict purchase propensity, customer churn, or customer lifetime value when suitable historical data and outcome labels are available.
5. **Examine temporal cluster stability:** Use customer data collected across multiple periods to determine whether segment membership and customer characteristics remain stable or change over time.
6. **Evaluate targeted marketing strategies:** Conduct controlled experiments or appropriate comparative studies to assess whether segment-specific campaigns produce measurable improvements in conversion, repeat purchases, incremental revenue, and profit.
7. **Validate the framework across datasets:** Apply the methodology to datasets from multiple retailers, industries, and geographic markets to assess its robustness and generalizability.
8. **Improve customer privacy and data governance:** Establish appropriate data-minimization, security, consent, and access-control practices when integrating more detailed customer information.

These future research directions can extend the current income-and-spending-based segmentation framework into a more comprehensive predictive marketing system. Such extensions should be empirically evaluated to determine whether the resulting customer profiles lead to measurable business improvements.

References

[1.] 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*, 527–570. https://doi.org/10.1007/s10257-023-00640-4

[2.] Babu, T., Ebin, P. M., & Nair, R. R. (2024). Data-driven insights: Mall customer segmentation through fuzzy C-means clustering. In *Proceedings of the 2024*

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/001

International Conference on Business and Information Technologies (BITCON). IEEE. https://doi.org/10.1109/BITCON63716.2024.10984994

[3.] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

[4.] 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*, 4995–5005. https://doi.org/10.1007/s00521-023-09339-6

[5.] Kumar, N. (2025). Intelligent customer segmentation: Unveiling consumer patterns with machine learning. *Journal of Umm Al-Qura University for Engineering and Architecture*, *16*, 774–783. https://doi.org/10.1007/s43995-025-00180-7

[6.] Pham, V. H., McIver, T., Robertson, K., & Thaichon, P. (2025). Effective customer segmentation using the recency, frequency, and monetary framework, combined with density-based clustering. *Journal of Strategic Marketing*. https://doi.org/10.1080/0965254X.2025.2571534

[7.] 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

[8.] Vjchoudhary7. (n.d.). *Mall customer segmentation data* [Data set]. Kaggle. https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python