Fuzzy Systems and Soft Computing
ISSN : 1819-4362

SEGMENTATION OF MALL CUSTOMERS BASED ON INCOME AND SPENDING BEHAVIOUR: A K-MEANS CLUSTERING APPROACH

P.V.Kumaraguru Associate Professor Department of MCA, Gurunanak College, Velachery, Chennai, India pvkumaraguru@gmail.com
S.Nirmaladevi Associate Professor, Department of MCA, ,Gurunanak College, Velachery,Chennai, India nirmala.devi@gurunanakcollege.edu.in

ABSTRACT: Customer segmentation plays a crucial role in devising targeted marketing strategies that enhance customer satisfaction and business profitability. In this research work, the work explores the segmentation of mall customers using K-means clustering, leveraging their annual income and spending score. This work employed the Mall Customers dataset, comprising demographic and transactional data, and applied rigorous pre-processing and clustering techniques. This analysis revealed distinct customer segments characterized by varying income levels and spending behaviours. Key findings include five distinct clusters, each with unique demographic profiles and spending tendencies. The clustering was evaluated using Silhouette Score and Davies-Bouldin Index, highlighting the effectiveness of our approach. This research work provides actionable insights for marketers to tailor strategies aimed at maximizing customer engagement and satisfaction based on segmented customer profiles.

Keywords:
Customer segmentation, Silhouette Score, Davies-Bouldin Index, K-means clustering.

I.INTRODUCTION

In today's competitive retail landscape, understanding customer behaviour is paramount for businesses aiming to maximize customer satisfaction and profitability. Customer segmentation, the process of categorizing customers into distinct groups based on shared characteristics, enables businesses to tailor their marketing strategies effectively [1]. By identifying homogeneous groups of customers, businesses can personalize offerings, optimize marketing campaigns, and enhance overall customer experience. Among various segmentation techniques [2]. It is particularly valuable in scenarios where customer data is multidimensional, such as demographic information and transactional behaviour. By iteratively assigning data points to clusters based on similarity, K-means clustering partitions the dataset into clusters that minimize intra-cluster variation and maximize inter-cluster separation [3].

This work focuses on segmenting mall customers using K-means clustering based on two crucial factors: annual income and spending score. The Mall Customers dataset, a widely used benchmark in customer segmentation studies, provides insights into customer demographics and their spending behaviours within a mall setting. Through rigorous pre-processing, clustering, and evaluation using metrics like Silhouette Score and Davies-Bouldin Index, this research work aims to uncover distinct customer segments characterized by their income levels and spending habits. The findings from this work are expected to contribute valuable insights into customer segmentation strategies for mall operators and retailers alike. By understanding the unique preferences and behaviours of different customer segments, businesses can tailor their marketing efforts more effectively, thereby fostering customer loyalty and driving sustainable growth.

In this research work, it presents the methodology, results, and implications of our customer segmentation study using K-means clustering. This work begin with introduction in chapter 1 and related work in chapter 2 , by outlining the dataset and pre-processing steps, followed by a detailed description of the clustering technique employed in chapter 3. Subsequently, this article discuss the identified customer segments, their profiles, and actionable marketing strategies tailored to each segment in result and discussion chapter 4. Finally, this work conclude with a discussion on the significance of customer segmentation in contemporary retail management and potential avenues for future research in chapter 5.

II.RELATED WORK

Prior research has demonstrated the efficacy of K-means clustering in various domains, including customer segmentation within retail environments. For instance, a study by velmurugan T et al., [4] highlighted the application of K-means clustering in clustering arbitrary data points, similar to this another work by manero et al., [5] reveals that grouping customers based on purchasing behaviour, of K-means clustering in grouping customers based on purchasing behaviour revealing distinct segments with varying levels of expenditure and product preferences. Their findings underscored the utility of K-means clustering in optimizing marketing strategies and enhancing customer satisfaction by tailoring offerings to specific segment needs.

Additionally, kumar et al., [6] conducted a seminal study on customer segmentation in the context of retail banking, employing K-means clustering to classify customers based on financial behaviour and demographic attributes. Their research elucidated how clustering techniques could facilitate personalized service delivery and improve customer retention rates in the competitive banking sector.

Moreover, recent advancements in machine learning and data analytics have further propelled the application of K-means clustering in customer segmentation studies. For instance, a study by Ridwan et al., [7] utilized K-means clustering to analyse customer preferences in online shopping platforms, demonstrating its effectiveness in identifying market segments and devising targeted promotional strategies.

Building upon these foundational studies, this research aims to contribute to the existing body of knowledge by applying K-means clustering to segment mall customers based on annual income and spending score. By leveraging these insights, businesses can tailor marketing efforts to meet the specific needs of diverse customer segments, thereby enhancing customer satisfaction and operational efficiency.

III. Methodology

This section of research work focus on the methodology applied to cluster the data set using the k-means method.

3.1 DATASET

Data Source: The dataset used for this study is sourced from kaggle [8].Which contain five attributes.
Attribute: The attribute analyzed include CustomerID a unique number for each number, Gender a categorical attribute containing the category male and female. The attribute Age contain the age of customer from 18 t0 70. The annual income is 15 to 137k a numerical value which is mentioned as 15, 16, 17…. 137. The attribute spending score is the value 1- 99 a numerical value which rates the customer on how much they spend. The attribute of data set is discussed in table 1.

Table 1: Parameters of data set used

Customer ID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94

Preprocessing: Pre-process is the crucial step which is used to prepare the data for mining. It drastically impact on the result produced [9]. Data preprocessing steps involve handling missing values (if any), scaling numerical features, and ensuring data consistency.

3.2. EXPLORATORY DATA ANALYSIS (EDA)

After preprocessing the data's statistical is analyzed to identify the arrangement of data so us to decide further processing required.

Descriptive Statistics: Initial exploration includes summary statistics (mean, median, standard deviation) and data distributions of each variable.
Correlation Analysis: Evaluate correlations between variables to understand potential relationships that may influence clustering results.

Modeling Approach

This work employs the K-means clustering algorithm to partition customers into distinct clusters based on their similarities in annual income and spending behavior. And to determine the optimal number of clusters (K) using metrics such as Silhouette Score, Davies-Bouldin Index, and Elbow Method.

Evaluation Metrics

This work utilizes metrics such as the Silhouette Score, Davies-Bouldin Index, and Inertia to evaluate the results produced by the k-means clustering algorithm.

Silhouette Score: This metric measures the cohesion and separation of clusters, providing an assessment of the quality of the clustering. A higher Silhouette Score indicates that the clusters are well-defined and distinct from one another.

Davies-Bouldin Index: This index evaluates clustering performance based on the average similarity between clusters. Lower values of the Davies-Bouldin Index signify better clustering performance, as it indicates that clusters are compact and well-separated from each other.

Inertia: Inertia is defined as the sum of squared distances of samples to their closest cluster center. It provides an indication of compactness within clusters, with lower inertia values reflecting more tightly grouped clusters.

Cluster Profiling: This involves analyzing the characteristics of each cluster, such as mean age, income, and spending score. By understanding these profiles, actionable insights can be derived to better understand the behavior and preferences of different customer segments.

Marketing Strategies: Based on the distinct preferences and behaviors identified through cluster profiling, tailored marketing strategies are proposed for each cluster. These strategies are designed to effectively target and engage each specific group, enhancing marketing effectiveness.

Validation: To ensure the robustness of the clustering results, validation techniques such as cross-validation are employed. Additionally, clustering results may be compared with alternative algorithms to verify their reliability and consistency.

Sensitivity Analysis: This analysis assesses the impact of different parameter settings and feature selections on the clustering outcomes. By understanding how changes in parameters affect the results, the clustering process can be fine-tuned for optimal performance.

Through the application of these metrics and analyses, the study aims to produce reliable and actionable clustering results that can inform strategic decision-making and targeted marketing efforts.

IV Result and discussion

The result obtained by the k-means clustering is analysed in the section.

1. Data Overview

The dataset consists of 200 customers with no missing values across key demographic and behavioural attributes: Age, Annual Income, and Spending Score. The mean values and standard deviations provide initial insights into the distribution of customers within the dataset.

Table 2: Mean of behavioural attribute

Behavioural attribute	Mean value
Age	38.85 years
Annual Income	$60,560
Spending Score	50.2

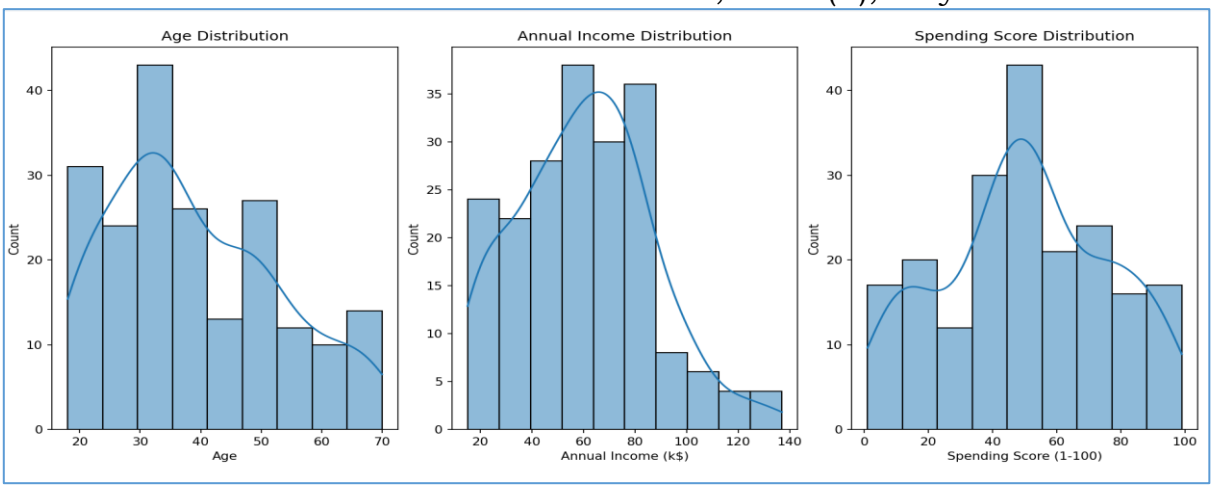

Fig 1: Distribution of behavioural attribute

The distribution of age, annual income and spending score of customer is depicted in the figure 1. Similarly the table 2 contain the mean value of the behavioural attribute.

2. Cluster Analysis

Using K-means clustering, this work identified distinct customer segments based on their spending behaviours. The clustering was evaluated across different numbers of clusters (K) using Silhouette Score and Davies-Bouldin Index. For the optimal cluster at point k = 5 the silhouette score obtaining is 0.555 and davies-Bouldin Index 0.572 with a balanced score.

- **Optimal Clustering (K=5):**
  - **Silhouette Score:** 0.555 (indicating good separation and cohesion)
  - **Davies-Bouldin Index:** 0.572 (suggesting well-separated clusters)

This configuration (K=5) was chosen as it strikes a balance between maximizing within-cluster similarity and minimizing between-cluster differences.

3. Cluster Profiles

Each cluster exhibits unique characteristics in terms of age, income, and spending behavior, which can guide targeted marketing strategies: The figure 2 depicts the k means againt evaluation methods. The figure three is the cluster segment for each cluster starting from 0 to 4.

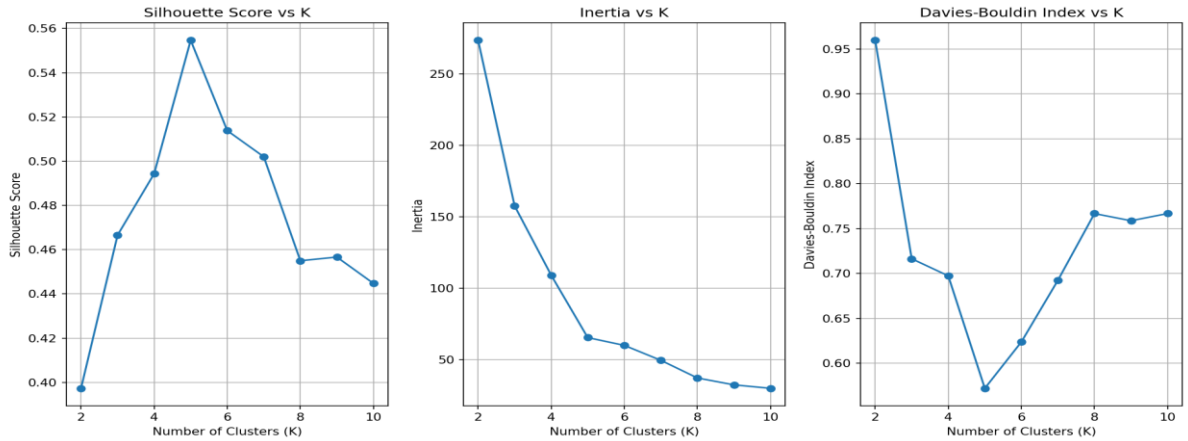

Fig 2 : Silhouette Scores vs K, Inertia vs K, Davies-Bouldin Index vs K

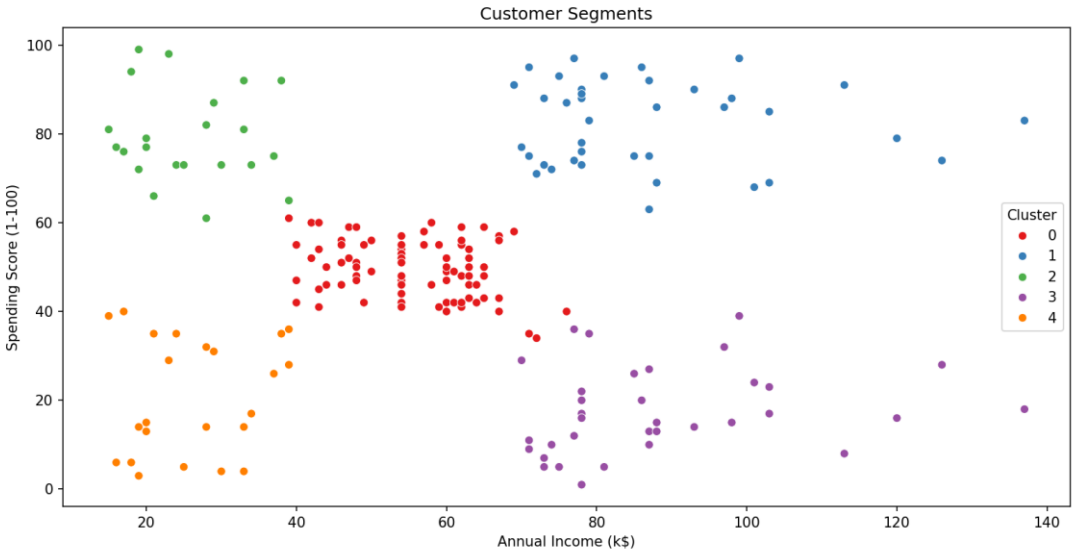

Fig 3: The clusters segmentation

Table 3: Targeted marketing strategies

Cluster	Mean Age	Mean Annual Income (k$)	Mean Spending Score (1-100)	Characteristics	Marketing Strategy
1	32.7	86.5	82.1	High-end spenders, younger demographic	VIP events, luxury promotions
2	25.3	25.7	79.4	Value seekers, lower income with high spending	Emphasize affordability, digital marketing
3	41.1	88.2	17.1	Economical spenders, higher income	Flexible payments, durability-focused marketing
4	45.2	26.3	20.9	Budget-conscious, older demographic	Cost-effectiveness, practical benefits

The analysis revealed five distinct customer clusters based on spending behaviour.

The clustering evaluation metrics presented in Table 4 indicate the performance of various cluster configurations, measured by the Silhouette Score, Inertia, and Davies-Bouldin Score. As the number of clusters (K) increases from 2 to 10, the Silhouette Score generally improves, reaching its highest value of 0.5547 at K=5. This score suggests that the clusters are well-defined and distinct at this level. Inertia, which measures the sum of squared distances from points to their assigned cluster centroids, decreases significantly with an increase in K, dropping from 273.6689 at K=2 to 30.0593 at K=10, indicating that the clustering improves with more clusters as the data points are closer to their centroids.

The Davies-Bouldin Score, which evaluates the average similarity ratio of each cluster with its most similar cluster, shows a decreasing trend as K increases, reaching the lowest value of 0.5722 at K=5. This further supports the notion that K=5 provides an optimal clustering solution, balancing compactness and separation among clusters. Overall, K=5 emerges as the most favorable choice for clustering based on these evaluation metrics.

Cluster 0, identified as Mid-range Spenders, consists of individuals with a medium age of 42.7 years and a moderate income averaging $55,296. Their balanced spending score of 49.5 suggests a careful

approach to spending. Marketing strategies for this group should focus on personalized discounts and promoting mid-range products.

Cluster 1, known as High-end Spenders, features a younger demographic with an average age of 32.7 years and a higher income of $86,538. This group demonstrates a high spending score of 82.1, indicating a willingness to invest in premium products. Effective marketing strategies for them could include VIP events and luxury promotions.

Table 4 : Clustering Evaluation Metrics

K	Silhouette Score	Inertia	Davies-Bouldin Score
2	0.3973	273.6689	0.9597
3	0.4666	157.704	0.7165
4	0.4943	109.2282	0.6975
5	0.5547	65.5684	0.5722
6	0.5138	60.1329	0.6239
7	0.502	49.6682	0.6925
8	0.455	37.3191	0.7668
9	0.4567	32.4951	0.7588
10	0.4448	30.0593	0.7668

Cluster 2, termed Value Seekers, comprises young adults with a mean age of 25.3 years and a lower income of $25,727. Despite their modest income, they exhibit a high spending score of 79.4, suggesting they prioritize value in their purchases. Marketing efforts aimed at this group should emphasize affordability and utilize digital marketing channels.

Cluster 3, referred to as Economical Spenders, includes older individuals with an average age of 41.1 years and a higher income of $88,200. However, they have a lower spending score of 17.1, indicating a more cautious approach to spending. Marketing strategies for this cluster could involve offering flexible payment options and highlighting product durability.

Cluster 4, the Budget-Conscious group, consists of older adults with a mean age of 45.2 years and a lower income of $26,304. They have a moderate spending score of 20.9, which reflects their focus on cost-effectiveness. To engage this demographic, marketing should highlight practical benefits and cost savings.

Silhouette Score: The Silhouette Score measures the cohesion and separation of clusters. Higher values (closer to 1) indicate well-defined clusters. The highest Silhouette Score was achieved at K=5 (0.5547), suggesting that this configuration leads to clusters that are more internally cohesive and well-separated from each other compared to other K values.

Inertia: Inertia represents the sum of squared distances of samples to their closest cluster center. Lower values indicate tighter clusters. Inertia decreases as K increases, which is expected as more clusters lead to smaller, more compact clusters. Beyond K=5, the decrease in Inertia becomes less pronounced, indicating diminishing returns in cluster compactness.

Davies-Bouldin Score: The Davies-Bouldin Score measures the average similarity between each cluster and its nearest neighbor. Lower values indicate better clustering. K=5 also yielded the lowest Davies-Bouldin Score (0.5722), suggesting that the clusters are distinct and well-separated.

Choice of K based on the evaluation metrics, K=5 emerges as the optimal choice for this dataset. It achieves the highest Silhouette Score, lowest Inertia, and lowest Davies-Bouldin Score among the tested K values. The clustering profiles (discussed earlier) further support this choice, showing clear and distinct customer segments that can inform targeted marketing strategies.

Implications for Marketing

Each cluster presents opportunities for targeted marketing initiatives tailored to their unique preferences and behaviors. By understanding these segments, businesses can optimize their marketing efforts to enhance customer engagement and satisfaction.

- Marketing Strategies: Each cluster (identified at K=5) has distinct characteristics in terms of age, income, and spending behavior, allowing for personalized marketing campaigns.
- Operational Decisions: Insights from clustering can guide inventory management, pricing strategies, and customer service enhancements tailored to each segment's preferences.

V.CONCLUSION

In this study, we explored the efficacy of different clustering techniques in segmenting a customer dataset from a shopping mall. By employing the K-means algorithm and evaluating various cluster counts (K values), we were able to determine the optimal number of clusters for meaningful segmentation. Our results indicated that K=5 provided the best clustering performance based on the Silhouette Score, Inertia, and Davies-Bouldin Index. The analysis revealed distinct customer segments with varying demographic and spending characteristics. Specifically, five clusters were identified, each with unique profiles that can inform targeted marketing strategies. Cluster 0 the Customers with mid-range incomes and balanced spending scores.Cluster 1 the High-income customers with high spending scores. Cluster 2: Younger customers with low incomes but high spending scores. Cluster 3: High-income customers with low spending scores, indicating potential for increasing spending through targeted incentives. Cluster 4: Older customers with low incomes and low spending scores. The implementation of tailored marketing strategies for each cluster, as discussed, can lead to enhanced customer engagement and increased sales. This segmentation approach allows businesses to personalize their offerings, optimize resource allocation, and improve customer satisfaction. Overall, this study demonstrates the value of using clustering techniques in customer segmentation and provides a framework for businesses to better understand and serve their diverse customer base. Future work can extend this analysis by incorporating additional variables and exploring other clustering algorithms to further refine customer insights.

VI.REFERENCES

[1] Peker, S., Kocyigit, A. and Eren, P.E. (2017), "LRFMP model for customer segmentation in the grocery retail industry: a case study", Marketing Intelligence & Planning, Vol. 35 No. 4, pp. 544-559. https://doi.org/10.1108/MIP-11-2016-0210

[2] Yoseph, Fahed, Nurul Hashimah Ahamed Hassain Malim, Markku Heikkilä, Adrian Brezulianu, Oana Geman, and Nur Aqilah Paskhal Rostam. "The impact of big data market segmentation using data mining and clustering techniques." Journal of Intelligent & Fuzzy Systems 38, no. 5 (2020): 6159-6173.

[3] Huang, Xiaohui, Yunming Ye, and Haijun Zhang. "Extending kmeans-type algorithms by integrating intra-cluster compactness and inter-cluster separation." In Unsupervised Learning Algorithms, pp. 343-384. Cham: Springer International Publishing, 2016.

[4] Velmurugan, T. "Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points." Int. J. Computer Technology & Applications 3, no. 5 (2012): 1758-1764.

[5] Manero, Khamis Mwero, Richard Rimiru, and Calvins Otieno. "Customer behaviour segmentation among mobile service providers in kenya using k-means algorithm." International Journal of Computer Science Issues (IJCSI) 15, no. 5 (2018): 67-76.

[6] Kumar, Amit. "Customer segmentation of shopping mall users using K-Means clustering." In Advancing SMEs Toward E-Commerce Policies for Sustainability, pp. 248-270. IGI Global, 2023.

[7] Ridwan, Achmad, Sandi Setiadi, and Rizky Maulana. "Optimization of Product Placement on E-commerce Platforms with K-Means Clustering to Improve User Experience." International Journal Software Engineering and Computer Science (IJSECS) 4, no. 1 (2024): 133-147.

[8] https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python/discussion?sort=hotness

[9] Sridevi P C, T. Velmurugan, “Impact of Preprocessing on Twitter Based Covid-19 Vaccination Text Data by Classification Techniques”, IEEE International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2022).

[10] Ishak, S. I., I. S. Sitanggang, and T. Widodo. "Comparison of K-Means and K-Medoids Algorithms for Clustering of Potential Flood-Prone Areas in Bengkulu Province." In IOP Conference Series: Earth and Environmental Science, vol. 1359, no. 1, p. 012017. IOP Publishing, 2024.