# Customer Segmentation through KMeans Clustering: Insights into Income and Spending Patterns


## Abstract

This project utilizes KMeans clustering to analyze customer segmentation based on two key variables: Annual Income (k$) and Spending Score (1-
100). By applying this unsupervised machine learning technique, we aim to identify distinct c ustomer groups within the dataset to better understand consumer behavior and preferences

The dataset consists of annual incomes and spending scores for a sample population. Data p reprocessing steps include normalization using MinMaxScaler to ensure the features are on a comparable scale. KMeans clustering is then performed to partition the data into three clust ers.
centroids of these clusters provide insight into the average characteristics of each group.

Visualization techniques such as scatter plots with distinct markers and centroid highlights ar e employed to illustrate the clustering results effectively. Additionally, a detailed summary of the clusters, including mean values for income and spending scores, is provided to facilitate interpretation.

This clustering approach offers a powerful method for uncovering patterns and segments wit hinthe data, enabling targeted marketing strategies and better customer relationship manag ement.



## Keywords

K-Means Clustering, Customer Segmentation, Unsupervised Learning, Normalization, Cluster Centroids, Scatter Plot, Cluster Analysis, Min Max Scaler, Market Segmentation, Consumer Behavior.


## Introduction

In today's datadriven world, understanding customer behavior is paramount for businesse s to tailor their marketing strategies and enhance customer satisfaction. Clustering al gorithms, particularly KMeans,
provide a powerful means to segment customers based on specific attributes. This project fo cuses on applying KMeans clustering to segment customers using two key variables: Annual Income (k$) and Spending Score.By identifying distinct clusters within the dataset, we can u ncover patterns and insights that inform more targeted and effective marketing approaches.

The dataset comprises annual income and spending scores for a sample population, repre senting diverse customer profiles. We employ data preprocessing techniques such as normal ization to ensure features are on a comparable scale, which is critical for the accuracy of the clustering algorithm. The KMeans algorithm is then applied to partition the data into cluster s, with centroids representing the average characteristics of each cluster.

Visualization tools are utilized to plot the clusters and their centroids, providing a clear visual representation of the segmentation results. This visual insight, combined with statistical analysis, helps interpret the clusters and derive actionable insights.

Ultimately, this project demonstrates how KMeans clustering can be leveraged to enhance customer understanding and improve business strategies, showcasing the practical applications of machine learning in real-world scenarios.

## Objective

The objective of this project is to apply KMeans clustering to segment customers based on their annual income and spending score. By achieving this, the project aims to:

1. Identify Distinct Customer Segments: Group customers into clusters that exhibit similar behaviors and characteristics.
2. Enhance Customer Insights: Gain a deeper understanding of customer profiles, preferences, and spending patterns.
3. Inform Marketing Strategies: Provide actionable insights that can help tailor marketing campaigns, promotions, and product offerings to specific customer segments.
4. Improve Business Decision-Making: Support data-driven decisions that enhance customer satisfaction and business performance.
5. Visualize Clustering Results: Create clear visual representations of clusters and their centroids to facilitate easy interpretation and communication of insights.

## Data Source and Collection

Kaggle is a fantastic platform for finding open datasets and machine learning projects. Everyone can explore, analyze, and share quality data across various domains.

Link: Mall_Customers (kaggle.com)

## Methodology

This project employs a structured methodology to analyze and segment customers using K-Means clustering based on their annual income and spending score. The methodology comprises several key steps:

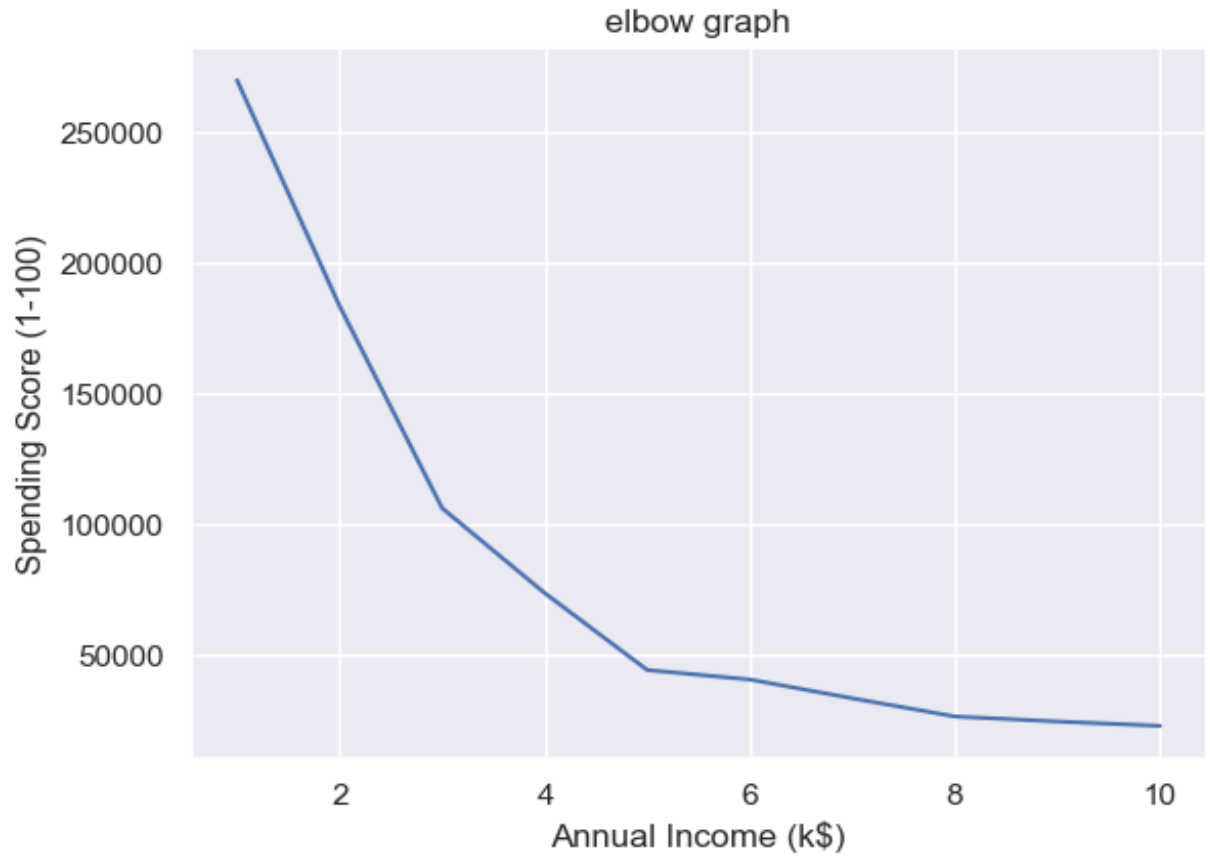
1. **Data Collection**:
   - Source: Kaggle dataset related to customer segmentation.
   - Attributes: Annual Income (k$) and Spending Score (1-100).
2. **Data Preparation**:
   - Load the dataset using Pandas.
   - Handle any missing values and clean the data.

- Normalize the data using MinMaxScaler to bring the features to a common sc ale.

3. **Exploratory Data Analysis (EDA)**:
    - Visualize the distribution of the attributes using scatter and box plots.
    - Analyze the relationship between annual income and spending score using scatter plots.
    - From boxplot we found one outlier as max value which is replaced by upper bound from inter quartile range.

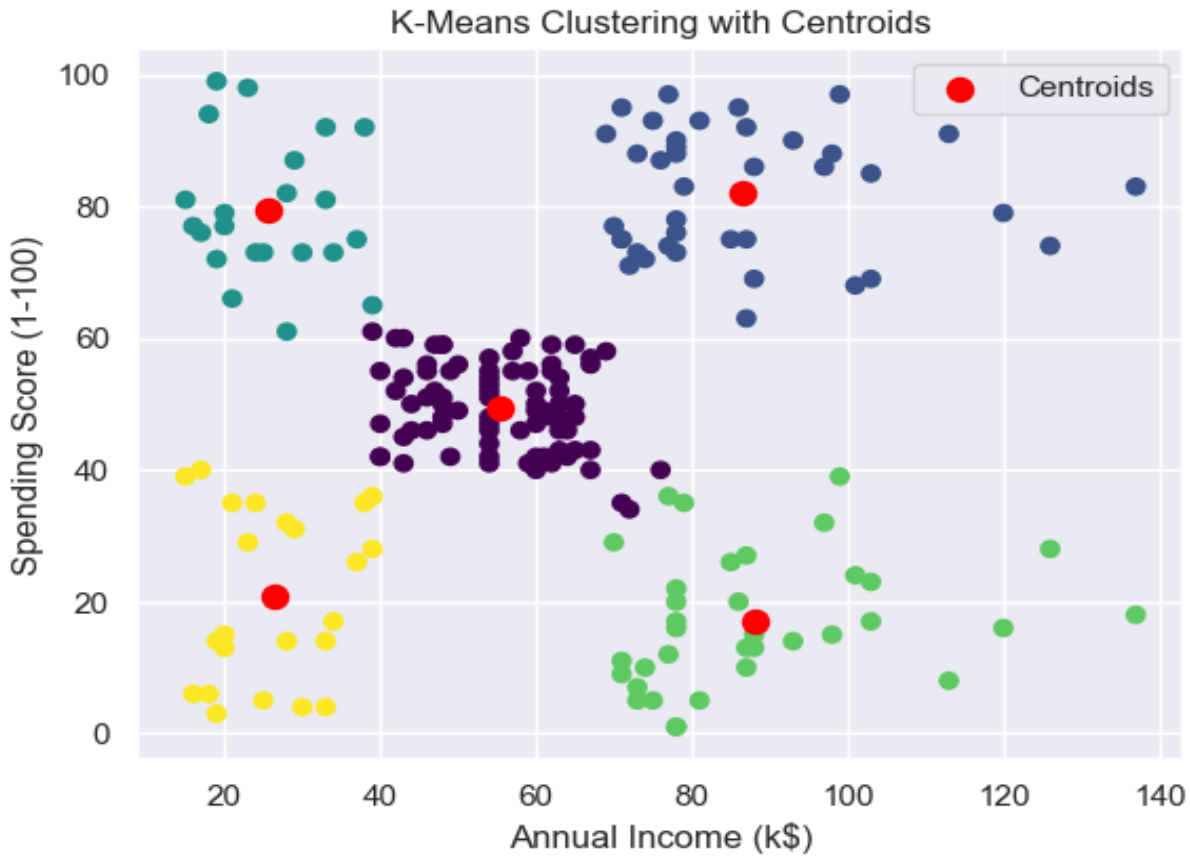
4. **K-Means Clustering**:
    - Select the number of clusters (k) using the Elbow Method.
    - Apply K-Means clustering to segment the data.
    - Calculate cluster centroids to understand the central point of each cluster.
5. **Visualization**:
    - Plot the clusters using scatter plots with distinct markers for each cluster.
    - Highlight cluster centroids on the plot for better visualization.
6. **Cluster Analysis**:
    - Summarize the characteristics of each cluster out of 5 by calculating the mean values of annual income and spending score.
    - Interpret the clusters to provide insights into customer segments.
7. **Evaluation**:
    - Assess the clustering results by examining the within-cluster sum of squares (WCSS).
    - Validate the clusters' relevance to business objectives and customer behavior.
8. **Reporting**:
    - Present the findings through visualizations and tables.
    - Provide actionable insights based on the clustering results.


Box Plot of Annual Income and Spending Score (1-100)
Values
140
120
100
80
60
40
20
0
Annual Income (k$)
Spending Score (1-100)


elbow graph
Spending Score (1-100)
250000
200000
150000
100000
50000
2
4
6
8
10
Annual Income (k$)

## Findings

This project involved the application of KMeans clustering to segment customers based on t heir annual income and spending score. The aim was to identify distinct five customer groups (suggested by Elbow curve) to provide insights into their behaviors and preferences.



| Customer ID | Genre | Age | Annual Income (k$) | Spending Score (1-100) | cluster |
|---|---|---|---|---|---|
| 1 | Male | 19 | 15 | 39 | 3 |
| 2 | Male | 21 | 15 | 81 | 2 |
| 3 | Female | 20 | 16 | 6 | 3 |
| 4 | Female | 23 | 16 | 77 | 2 |
| 5 | Female | 31 | 17 | 40 | 3 |
| … | … | … | … | … | … |
| 196 | Female | 35 | 103 | 79 | 0 |
| 197 | Female | 45 | 103 | 28 | 4 |
| 198 | Male | 32 | 103 | 74 | 0 |
| 199 | Male | 32 | 103 | 18 | 4 |
| 200 | Male | 30 | 103 | 83 | 0 |

After applying KMeans clustering with five clusters to the dataset containing CustomerID, Genre, Age, Annual Income (k$), and Spending Score (1-100), we observed the following characteristics for each cluster:

**Cluster 0: High Spenders**

- Spending Score Range: High
- Annual Income: $84,000
- Spending Score: 82
- Age: 33
- Description: Young professionals with high incomes and high spending, likely investing in luxury goods and experiences.

**Cluster 1: Low Spenders**

- Spending Score Range: Low
- Annual Income: $79,000

- Spending Score: 17
- Age: 40
- Description: Middle-aged individuals with high incomes but very cautious with their spending, focused on savings and investments.

**Cluster 2: Upper Middle Spenders**

- Spending Score Range: Upper Middle
- Annual Income: $48,000
- Spending Score: 56
- Age: 39
- Description: Balanced spenders with moderate incomes, exhibiting steady financial behavior.

**Cluster 3: Lower Middle Spenders**

- Spending Score Range: Lower Middle
- Annual Income: $26,000
- Spending Score: 21
- Age: 45
- Description: Older individuals with lower incomes and cautious spending behavior, reflecting careful budget management.

**Cluster 4: High Income, Low Spend**

- Spending Score Range: Low
- Annual Income: $101,000
- Spending Score: 22
- Age: 41
- Description: Individuals with the highest incomes but conservative spending habits, likely focusing on financial security and investments.
- These insights help in understanding the diverse financial behaviors and preferences among different customer segments, providing a basis for targeted marketing strategies and personalized services.

## Conclusion

This project successfully demonstrated the application of KMeans clustering to segment customers based on their annual income and spending score. By clustering the data into five distinct groups, we gained valuable insights into different customer profiles and their financial behaviors.

Key findings highlighted distinct clusters:

1. **High Spenders (Cluster 0)**:
    - Young professionals with high incomes and high spending scores.
    - Likely to invest in luxury goods and lifestyle experiences.
2. **Low Spenders (Cluster 1)**:
    - Middle-aged individuals with relatively high incomes but very low spending scores.
    - Focused on savings and investments rather than expenditure.
3. **Upper Middle Spenders (Cluster 2)**:
    - Individuals with moderate incomes and balanced spending scores.
    - Represent a steady financial behavior and balanced lifestyle.
4. **Lower Middle Spenders (Cluster 3)**:
    - Older individuals with lower incomes and cautious spending behaviors.
    - Reflect careful budget management and conservative spending habits.
5. **High Income, Low Spenders (Cluster 4)**:
    - The highest income group with very conservative spending scores.
    - Likely focused on financial security and long-term investments.

These clusters provide actionable insights that can inform targeted marketing strategies, en hance customer satisfaction, and improve business decision-making. The use of K-Means clustering, combined with effective data visualization and statistical analysis, showcas es the practical benefits of machine learning in customer segmentation.

By leveraging this methodology, businesses can better understand their customer base, tailo r their offerings, and ultimately drive better business outcomes. This project underscores the value of datadriven approaches in uncovering patterns and making informed decisions.

## Appendix:

TABLE SHOWING THE DESCRIPTIVE STATISTICS OF THE VARIABLES

| *Statistic* | *Customer ID* | *Age* | *Annual Income (k$)* | *Spending Score (1-100)* |
|---|---|---|---|---|
| *count* | *200.00* | *200.00* | *200.00* | *200.00* |
| *mean* | *100.50* | *38.85* | *60.56* | *50.20* |
| *std* | *57.88* | *13.97* | *26.26* | *25.82* |
| *min* | *1.00* | *18.00* | *15.00* | *1.00* |
| *25%* | *50.75* | *28.75* | *41.50* | *34.75* |
| *50%* | *100.50* | *36.00* | *61.50* | *50.00* |
| *75%* | *150.25* | *49.00* | *78.00* | *73.00* |
| *max* | *200.00* | *70.00* | *137.00* | *99.00* |

Data Type:

| *index* | *Column* | *Non-Null Count* | *Data type* |
|---|---|---|---|
| *0* | *CustomerID* | *200 non-null* | *integer* |
| *1* | *Genre* | *200 non-null* | *object* |
| *2* | *Age* | *200 non-null* | *integer* |
| *3* | *Annual Income (k$)* | *200 non-null* | *float* |
| *4* | *Spending Score (1-100)* | *200 non-null* | *integer* |

## Literature Review

### 1. Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases

This review article discusses various customer segmentation methods, including KMeans clustering, and their application in personalized customer targeting in ecommerce. It highlights the importance of understanding customer behavior and preferences to enhance marketing strategies and improve customer satisfaction.

### 2. How Can Algorithms Help in Segmenting Users and Customers? A Systematic Review and Research Agenda for Algorithmic Customer Segmentation

This systematic review explores different algorithms used for customer segmentation, including KMeans clustering. It addresses key questions such as the choice of algorithm, the number of customer segments, and the evaluation of results. The review provides a comprehensive research agenda for future studies in this area.

### 3. Customer Segmentation Using Machine Learning: A Literature Review

This literature review focuses on the application of machine learning techniques, including K-Means clustering, for customer segmentation. It discusses the role of customer analytics in understanding customer behavior and improving customer retention. The review also identifies gaps in the current research and suggests areas for further study.

### 4. Customer Segmentation Analysis for Improving Sales Using Clustering

This research paper demonstrates the practical application of K-Means clustering for customer segmentation based on purchasing behavior. It shows how clustering can identify distinct customer groups and proposes strategies to target these segments effectively.

### 5. Mall Customer Segmentation Using K-Means Clustering

This study applies KMeans clustering to segment mall customers based on demographic data spending patterns, and mall visit frequency. It suggests optimizing the KMeans algorithm to improve customer segmentation and enhance marketing efforts