**CUSTOMER SEGMENTATION ANALYSIS OF MALL CUSTOMERS**

**A Data Science Project on Segmenting Customers Based on Income and Spending Behavior**

**By**

**Ahmed Talal**

**September 2025**

GitHub Repository: github.com/ahmaadtalal/CustomerSegmentation

**1. Introduction**

The purpose of this project is to perform **customer segmentation** using the Mall Customers dataset, aiming to uncover patterns in customer behavior and provide actionable insights for marketing and business strategy. Customer segmentation is a critical analytical approach that allows businesses to group customers with similar characteristics, enabling **targeted promotions, personalized engagement, and improved customer retention**.

**Project Goals:**

* **Identify distinct customer segments** based on **Annual Income** and **Spending Score**, allowing the business to recognize high-value and low-value customers.
* **Analyze demographic and behavioral patterns**, such as age and gender distribution across segments, to understand customer preferences and spending habits.
* **Visualize and interpret results** using plots and tables to provide clear, actionable insights for marketing strategies and business decision-making.

**Dataset Source:**

* Mall Customers dataset from Kaggle, consisting of 200 customers with attributes including **CustomerID, Gender, Age, Annual Income, and Spending Score**.

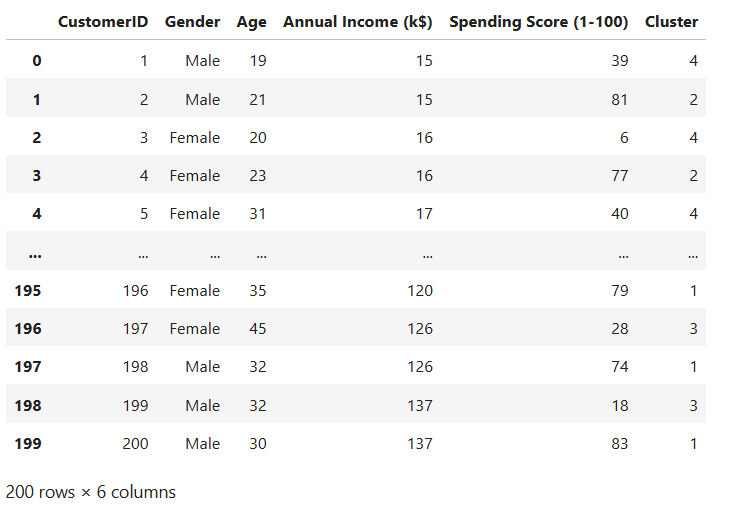
**Key Analytical Steps:**

* Data exploration and visualization to understand distributions and relationships.
* Feature scaling to prepare for distance-based clustering methods.
* Determination of the optimal number of clusters using **Elbow and Silhouette methods**.
* Application of **K-Means clustering** for segment identification.
* Profiling of each cluster and analysis of **average spending** to highlight the most profitable customer segments.

**2. Dataset Overview**

**Description:**

* The dataset contains 200 customers with 5 attributes: CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1–100).
* There are no missing values or duplicates.
* Gender distribution: 112 Female, 88 Male.
* Age range: 18–70 years.
* Annual Income range: 15k$–137k$.
* Spending Score range: 1–99.



A screenshot of a computer code

AI-generated content may be incorrect.

A table with numbers and numbers

AI-generated content may be incorrect.

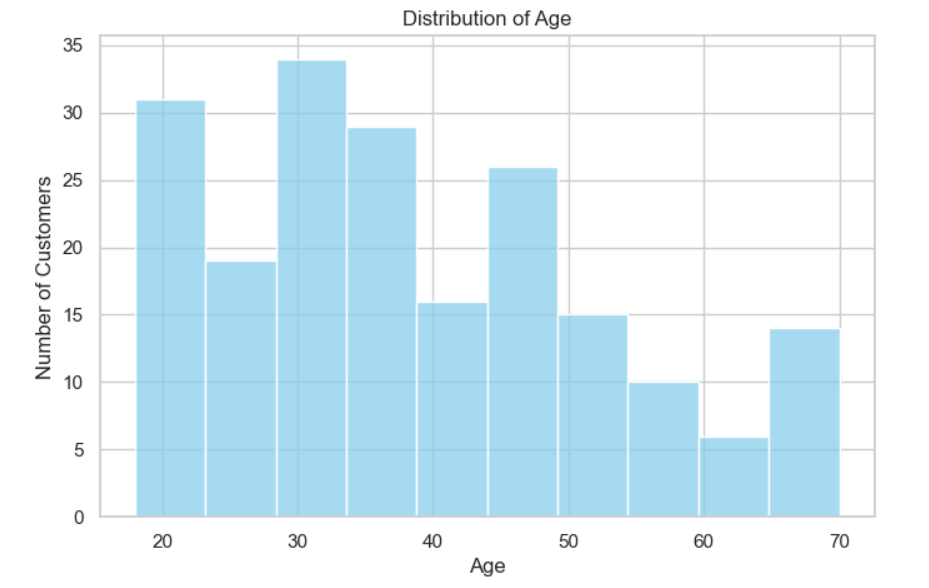
**Observations:**

* Gender distribution is slightly skewed toward females, but generally balanced.
* Income and spending scores exhibit wide variation, with some outliers.
* Provides sufficient diversity for meaningful customer segmentation analysis.

**3. Exploratory Data Analysis (EDA)**

**3.1 Univariate Analysis**

**Purpose:** Understand distribution of each variable.



A green graph with white grid

AI-generated content may be incorrect.

A graph with red bars

AI-generated content may be incorrect.

A graph showing a number of people in different colors

AI-generated content may be incorrect.

**Insights:**

* The majority of customers are between 30 and 50 years old, representing the core shopping demographic.
* Annual Income is mostly concentrated between 40k$ and 80k$, though there are a few high-income and low-income outliers.
* Spending Scores tend to cluster around 40–60, indicating moderate spending behavior for most customers.
* There is a slight predominance of female customers (112 vs. 88), but overall gender distribution is fairly balanced.
* These patterns suggest a diverse customer base, providing opportunities for segmentation based on income, spending, and demographic characteristics.

**3.2 Bivariate Analysis**

**Purpose:** Explore relationships between features.

A graph with green and orange dots

AI-generated content may be incorrect.

**A graph with red and blue dots

AI-generated content may be incorrect.**

**Insights:**

* Most customers are concentrated in the 40–60 spending score and 40k$–80k$ annual income range, forming the largest cluster.
* Gender overlaps across spending and income, indicating that both males and females are distributed similarly across segments.
* Younger customers sometimes exhibit high spending despite having lower incomes, highlighting potential high-value opportunities.
* A few outliers exist in both income and spending, representing niche customer behaviors that may require special attention.

**4. Feature Scaling**

**Purpose:** K-Means is distance-based; features need to be standardized.

**A screenshot of a screen

AI-generated content may be incorrect.**

**Insights:**

* After standardization, the features (Annual Income and Spending Score) have a mean close to 0 and a standard deviation of approximately 1, ensuring all features contribute equally to distance calculations.
* Negative values indicate customers who are below the dataset average for a feature, while positive values indicate customers who are above average.
* Scaling ensures that clustering is not biased by differing feature ranges and allows K-Means to form more accurate segments.

**5. Determining Optimal Number of Clusters**

**5.1 Elbow Method**

**A graph with a line

AI-generated content may be incorrect.**

**Insight:**

* The inertia (sum of squared distances to cluster centers) decreases as the number of clusters increases.
* The rate of decrease slows significantly after k=5, indicating that adding more clusters beyond 5 yields diminishing returns.
* This suggests that 5 clusters is an optimal choice for segmenting the customers.

**5.2 Silhouette Score**

**A graph with a line

AI-generated content may be incorrect.**

**Insight:**

* The silhouette score measures how similar each customer is to its own cluster compared to other clusters.
* The score peaks at k=5, confirming that 5 clusters provide the best separation and cohesion among the groups.
* Together with the Elbow method, this provides strong evidence to use 5 clusters for K-Means clustering.

**6. K-Means Clustering**

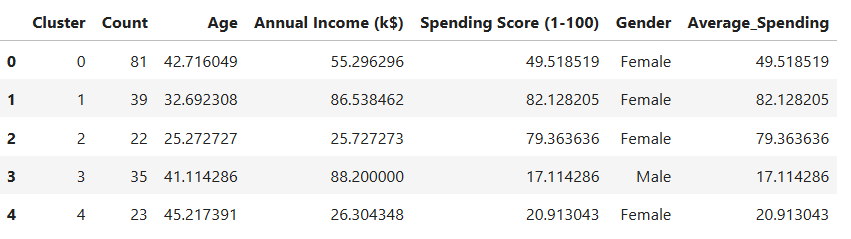
****

**Insights:**

* K-Means clustering resulted in a clear separation into 5 distinct clusters, consistent with the optimal cluster analysis.
* The scatterplot of scaled features demonstrates the technical separation achieved by the algorithm, showing how each customer is grouped based on standardized income and spending.
* The scatterplot with original (unscaled) features provides an intuitive view of customer segments in real-world units, making it easier to interpret and communicate business insights.
* These visualizations confirm that clustering effectively distinguishes high-value, medium-value, and low-value customer segments.

**7. Cluster Profiling**

**Cluster Characteristics:**

****

**Insights:**

* Clusters 1 & 2 represent high-value customers, with higher annual income and/or spending scores, making them key targets for premium marketing and loyalty programs.
* Cluster 0 corresponds to the average customer base, with moderate income and spending behavior, representing the core segment of regular shoppers.
* Clusters 3 & 4 are low-value segments, with lower spending scores and/or income, highlighting opportunities for engagement strategies to increase spending or maintain retention.
* These distinctions allow for tailored marketing strategies and resource prioritization based on customer value.

**7.1 Average Spending per Cluster**

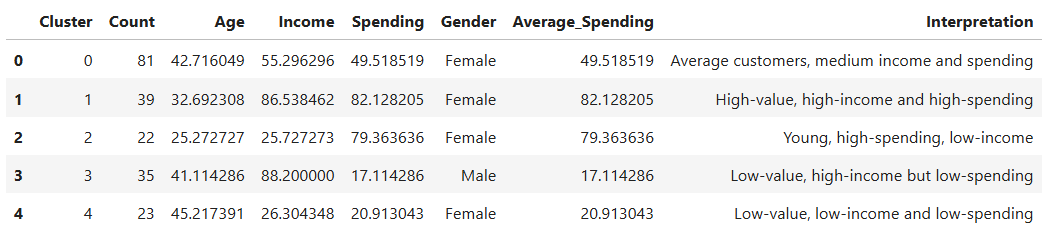
**A graph of a graph showing different colored bars

AI-generated content may be incorrect.**

**Insights**

* Cluster 1 has the highest average spending (~82), identifying it as the most profitable segment.
* Cluster 2 consists of younger customers with high spending (~79), representing a valuable growth opportunity.
* Cluster 0 shows moderate spending (~50), forming the core average customer group.
* Clusters 3 & 4 have low average spending (~17–21), indicating low-value segments that may require engagement strategies to increase their spending.
* This analysis highlights which segments contribute most to revenue and informs targeted marketing strategies for each cluster.

**8. Summary of Clusters**

****

**Key Insights**

* High-value segments: Clusters 1 & 2, characterized by high spending and/or income, represent the most profitable customers and should be the primary focus for premium marketing and retention strategies.
* Average-value segment: Cluster 0, with moderate spending and income, forms the core customer base and can be nurtured through regular promotions and engagement campaigns.
* Low-value segments: Clusters 3 & 4, with low spending and/or income, may require targeted campaigns or loyalty initiatives to encourage higher engagement and spending.
* Marketing strategy implications: Businesses can prioritize high-spenders, tailor personalized offers for medium-value customers, and design special programs for low-spenders to improve overall revenue and customer satisfaction.

**9. Conclusion**

* The analysis identified five distinct customer clusters, each with unique characteristics in terms of income, spending behavior, and demographics.
* The findings provide actionable insights for marketing, promotions, and customer engagement strategies, allowing the business to focus on high-value segments while addressing opportunities within low-value groups.
* Visualizations, including scatterplots, bar plots, and cluster summary tables, effectively illustrate the segmentation and support the analytical insights.
* Average spending analysis highlights the most profitable customer segments, enabling targeted campaigns to maximize revenue and enhance customer satisfaction.
* Overall, this segmentation provides a data-driven foundation for strategic decision-making in customer relationship management and marketing planning.