Name: Ammu Ambika Ramachandran
Student ID: 23006390
Github:

Mall Customer Segmentation Analysis Using Clustering and Fitting

Objective

The objective of this analysis was to perform customer segmentation using clustering techniques based on customer data, specifically focusing on the following features:**Age, Annual Income (k$),Spending Score (1-100)** The dataset use is downloaded from kaggle https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python which includes customer demographics and spending behavior. The goal was to apply K-Means clustering to divide customers into distinct groups and evaluate the performance of different clustering configurations using metrics such as silhouette scores and elbow scores. Also, I performed an analysis of the data and visualized my findings using Histograms, correlation matrix and box plot

Data Preprocessing and Exploratory Data Analysis (EDA)

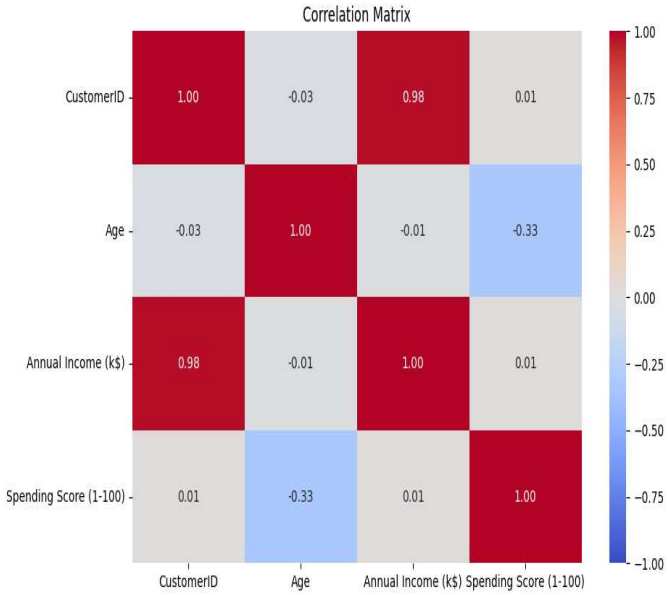
Data overview

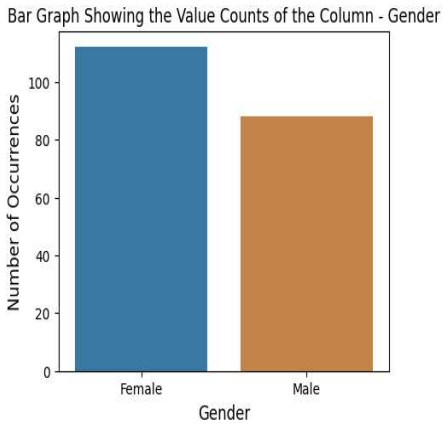
The initial dataset contains 200 rows and 5 columns. There are no missing values, as confirmed by the `df.info()` method. We observed the following basic data types:

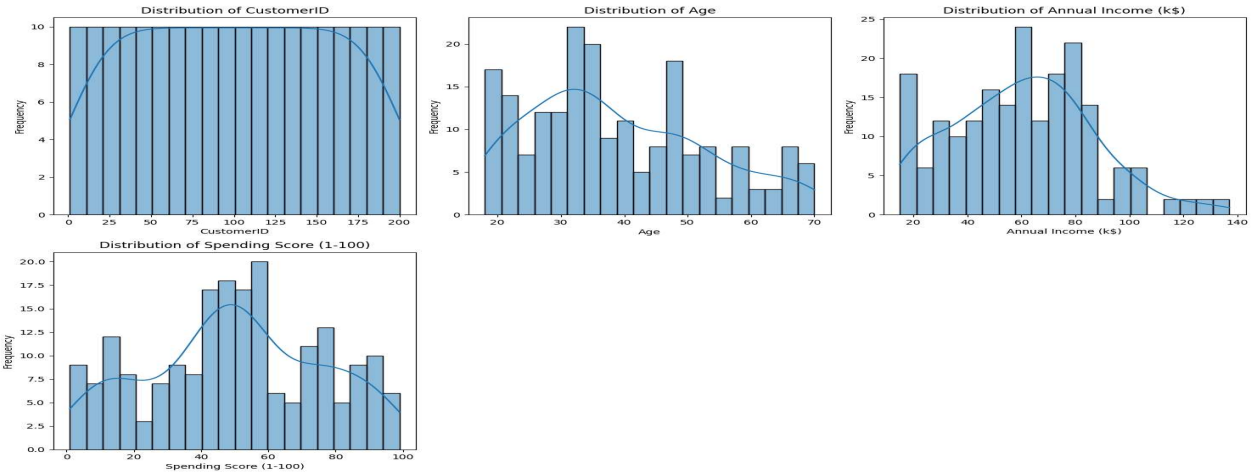
- Numerical: **Age**, **Annual Income (k$)**, **Spending Score (1-100)**
- Categorical: **Gender**
- The dataset contained a **Gender** column, which was converted into a categorical variable for use in the analysis. This feature was expected to provide valuable information forcustomer segmentation. In addition, basic statistical analyses were performed on numerical columns (Age, Annual Income, and Spending Score) to understand their distributions.

Correlation heat map: The correlation between numerical features is low to moderate

- Age vs Annual Income(k$):0.09(weak positive correation)
- Age vs Speding Score(1-100):_0.04(very weak negative correlation).





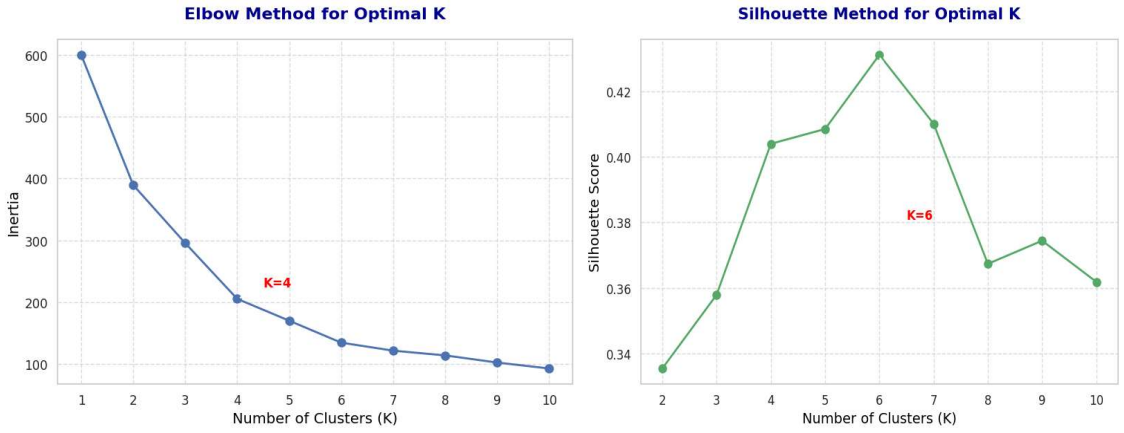


4. Statistical Summary: The statistical summary of the dataset revealed diversity in the customer population in terms of age, income, and spending scores. This diversity provided a solid foundation for segmentation and indicated that meaningful groupings could be made based on these features.

Clustering with K-Means:

After preprocessing the data, K-Means clustering was applied to group customers into segments based on their behavior and demographics.

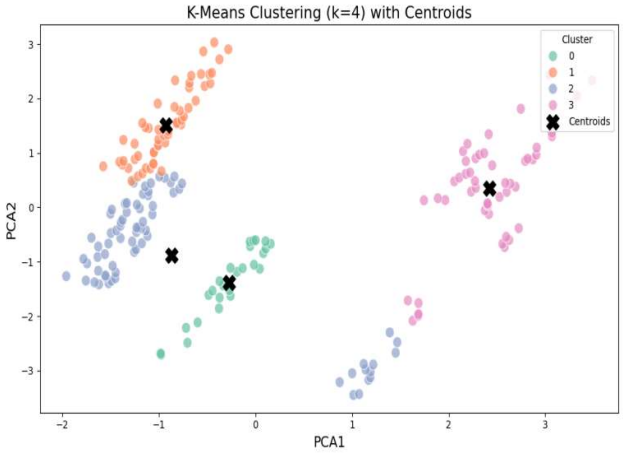
Elbow Method and silhouette Method: The clustering process began with the **Elbow Method** to determine the optimal number of clusters. By plotting inertia (the sum of squared distances between points and their cluster centroids) against the number of clusters, the "elbow" appeared at **4 clusters**. This suggested that 4 clusters might be the ideal choice.Then, silhouette Method to verify the number of clusters this was applied. The analysis provided a higher silhouette score for 6 clusters suggesting that splitting the data into clusters would result in better separation between the segments

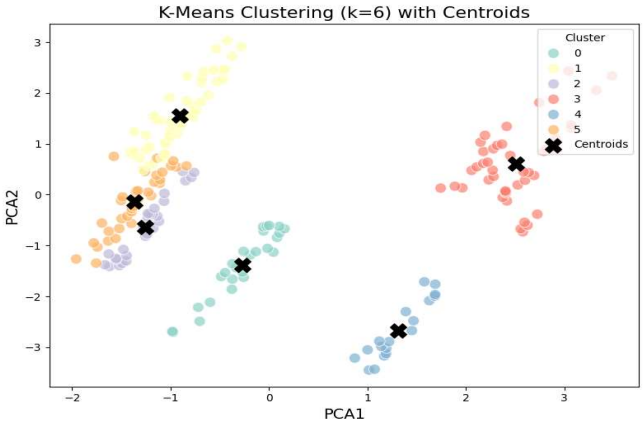


4. Cluster Evaluation

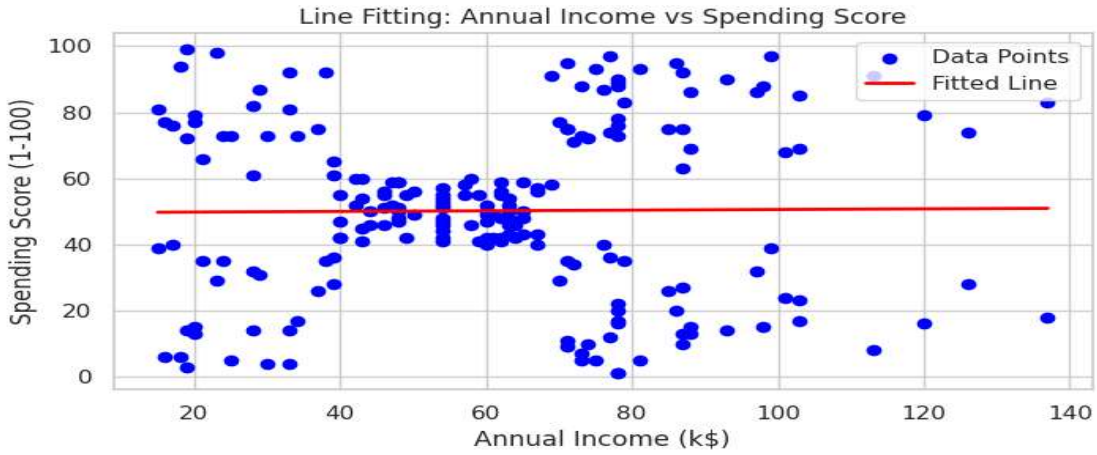
We analyzed the clusters formed by the K-Means algorithm to understand the customer segments better.**Cluster Summary (K=4),Cluster 0**: Older customers with lower spending scores — likely less engaged with promotions or shopping activities.,**Cluster 1**: Younger customers with high spending scores — active shoppers who engage with mall promotions and discounts,**Cluster 2**: Middle-aged customers with moderate income and spending — a balanced group likely to be price-sensitive.,**Cluster 3**: Older customers with higher spending scores — potentially retirees or those with more disposable income.**Cluster Summary (K=6).**When 6 clusters were used, the segmentation became more granular, revealing more distinct distinctions, particularly in terms of spending behavior and income levels. The additional clusters allowed for a more detailed understanding of customer behavior.

5. Visualization: The clusters were visualized using **scatter plots** and which clearly showed the separation between the different clusters. For further clarity, **Principal Component Analysis (PCA)** was applied for dimensionality reduction, and the clusters were even more distinct, with clear centroids marking the center of each group.





Line Fitting



Conclusion and Insights

Clustering Insights: The clustering analysis revealed distinct customer segments with differing characteristics:

- **Younger customers (Cluster 1)** with high spending scores are more likely to be frequent shoppers who respond well to promotions and discounts.
- **Older customers (Clusters 0 and 3)**, with lower or moderate spending scores, may benefit more from loyalty programs or retention strategies aimed at increasing engagement.
- **Middle-aged customers (Cluster 2)** with moderate income and spending scores could be targeted with more value-based or price-sensitive offerings.