Mall Customers Data Analysis Report

Assignment: Clustering and Fitting (30%); **Module:** Applied Data Science 1

Name: Deepak Raj Manickam

Student ID: 23070139; **Student Email ID:** dm24aav@herts.ac.uk

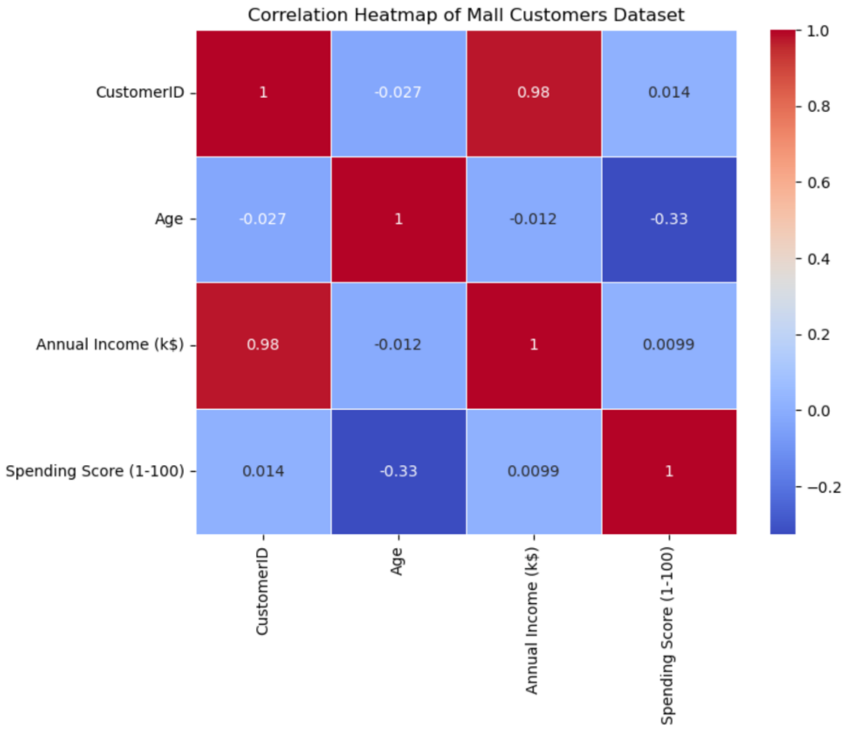
GitHub Repository: https://github.com/deepakraj-04/ADS1_CFA

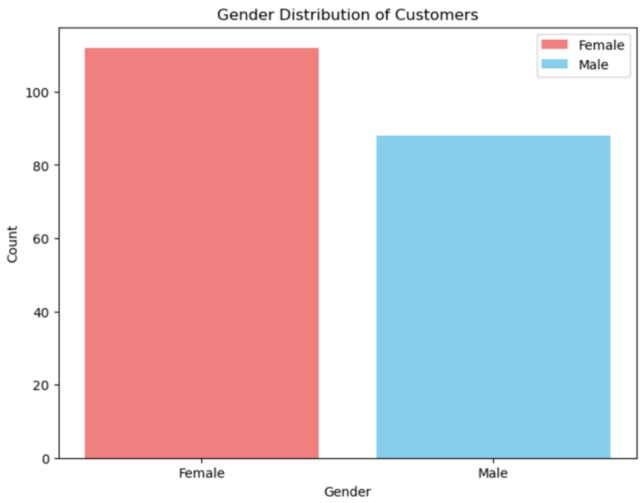
Introduction:

The mall customers dataset consists of 200 entries in which each representing a customer and includes the following columns such as CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100). The dataset is clean with no missing values or duplicate values in which we have a solid foundation for further analysis.

Plot 1: Correlation Heatmap of Mall Customers Dataset (Heatmap)

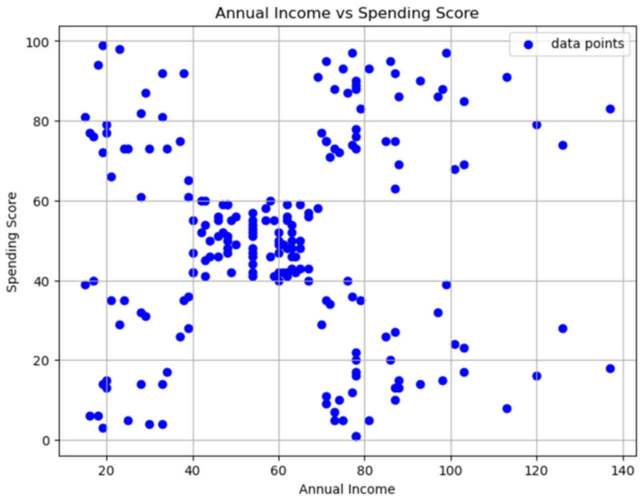
The heatmap gives us insights about the relationship between numerical values in the dataset. In the heatmap we can see a strong positive correlation exists between CustomerID and Annual Income (k$) (0.98) which indicates that more customer IDs are associated with higher annual incomes. There is a moderate negative correlation exists between Age and Spending Score (1-100) (-0.33) which indicates that old aged customers tend to spend less.





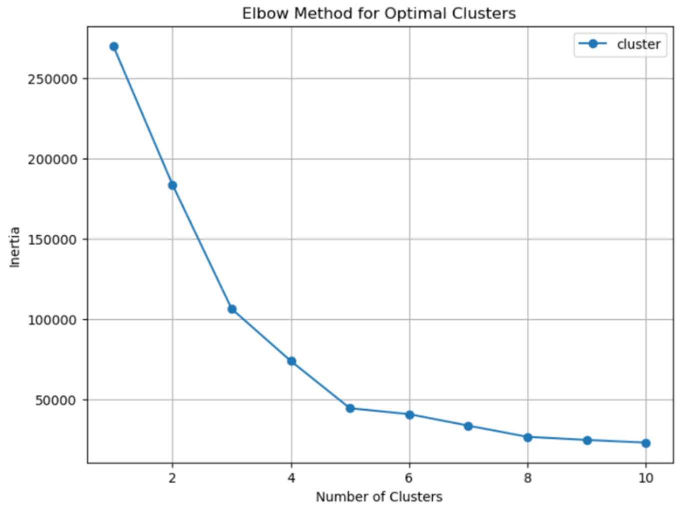
Plot 2: Gender Distribution of Customers (Bar Chart)

The bar chart displays the distribution of customers based on gender. In the bar chart we can significantly see that the number of female customers is more than the number of male customers.



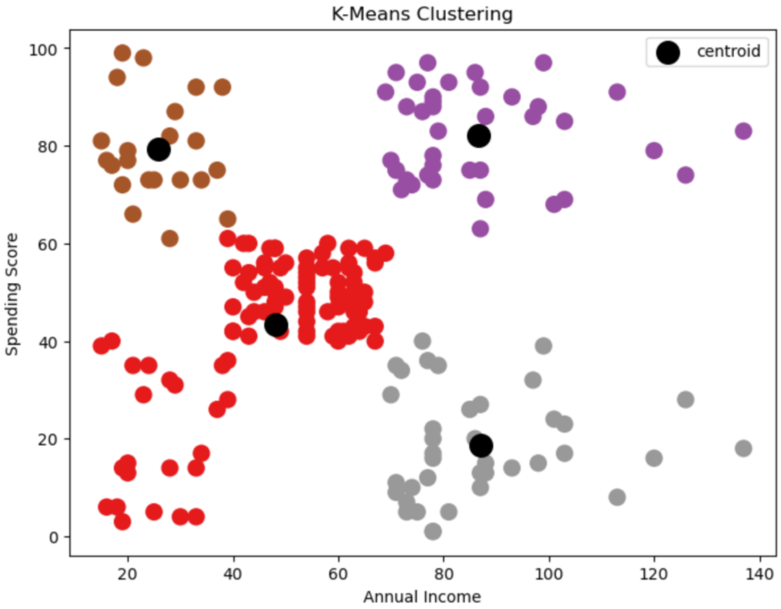
Plot 3: Annual Income vs Spending Score (Scatter Plot)

The Scatter Plot displays the relationship between Annual Income and Spending Score. In the scatter plot the points are scattered across the plot which indicates a weak linear correlation. However there are some noticeable clusters of points in the plot which suggests us that there is a group of customers similar annual income and similar spending patterns.



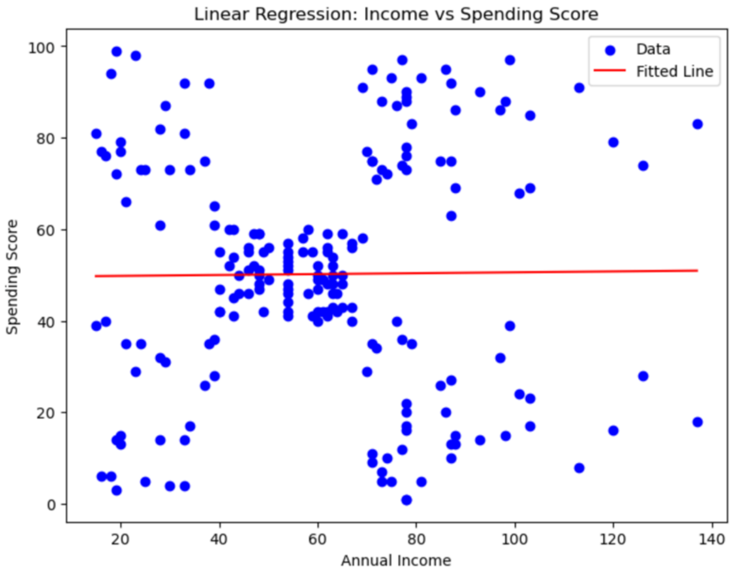
Plot 4: Elbow method for optimal clusters (Elbow plot)

The elbow plot helps us to choose the best number of clusters of the dataset. The elbow plot shows how the inertia changes as we increase the number of clusters. Generally the inertia decreases when the number of clusters gets increased. In the plot the elbow point appears to be in around 4 to 5 clusters which suggests us that using 4 or 5 clusters would be a great option for this dataset.



K-means Clustering (Clustering):

The K-means clustering plot displays the result of applying the K-Means clustering algorithm to the dataset with Annual Income and Spending Score columns. The data points are divided into four clusters represented by different color. The black dots represents centroids. In the plot cluster 1 (red) displays customers with low annual income and moderate spending scores, cluster 2 (purple) displays customers with high annual income and high spending scores, cluster 3 (brown) displays customers with high annual income and low spending scores and cluster 4 (gray) displays customers with moderate to low annual income and low spending scores.



Linear Regression: Income vs Spending Score (Fitting):

The plot displays the results of fitting a linear regression model to the dataset with relationship between annual income and spending score. The blue dots represent the data points while the red line represents the regression line.

Conclusion:

This report displays customer behavior and customer segmentation using clustering and linear regression. These insights can be useful to businesses in which they can develop better marketing strategies and understanding customer behaviors.