Dorcas Munoko

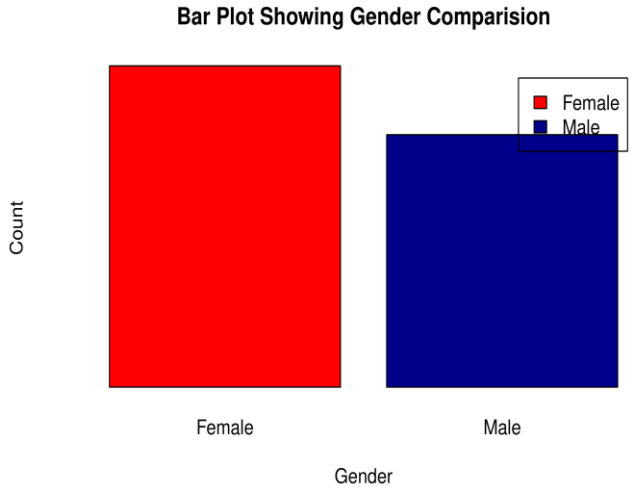
Customer Segmentation Analysis

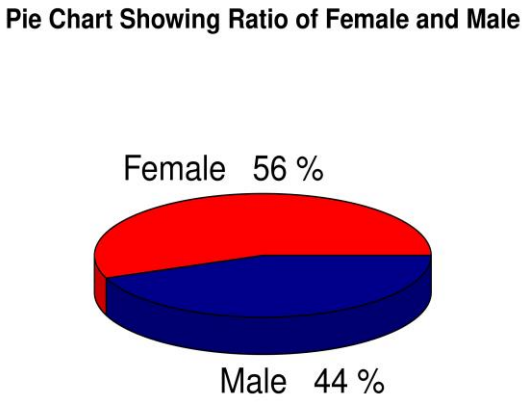
Introduction

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. The dataset contained 200 observations with main features including customer ID, gender, age, annual income, and spending score.

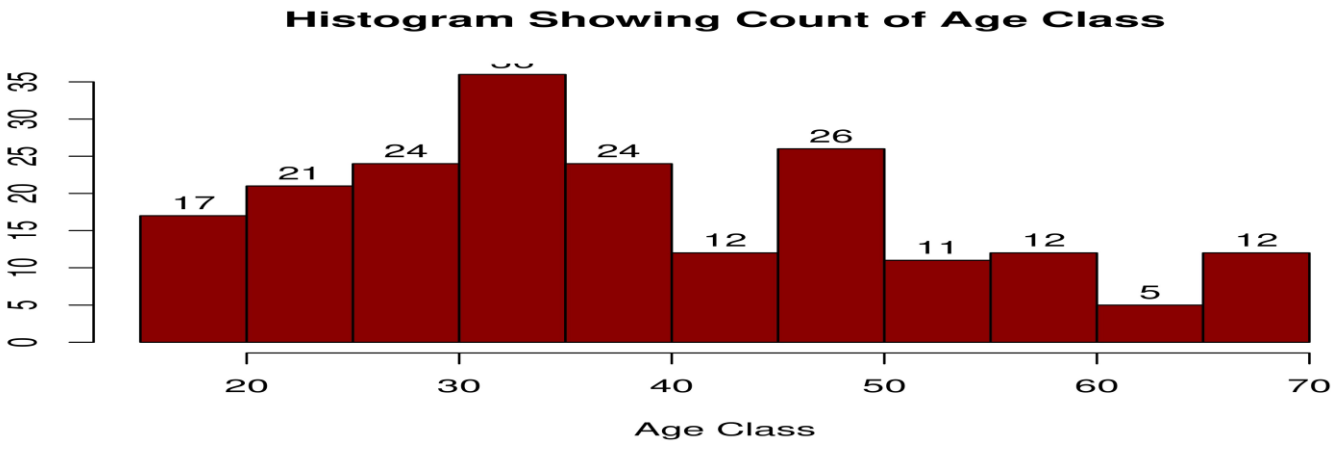
Description of Features

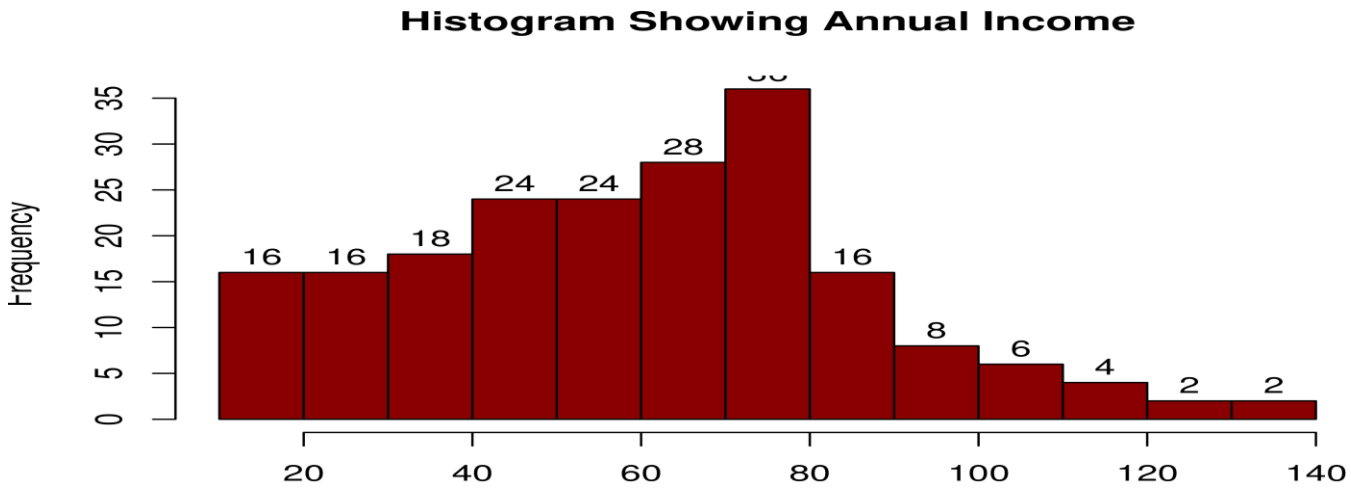
Using bar plot to display gender comparison, it was observed that the number of females was higher as compared to males. Females accounted for 56% of customers in the dataset while male accounted for 44%. The maximum customer ages were between 30 and 35. The minimum age of customers was 18, whereas the maximum age was 70.

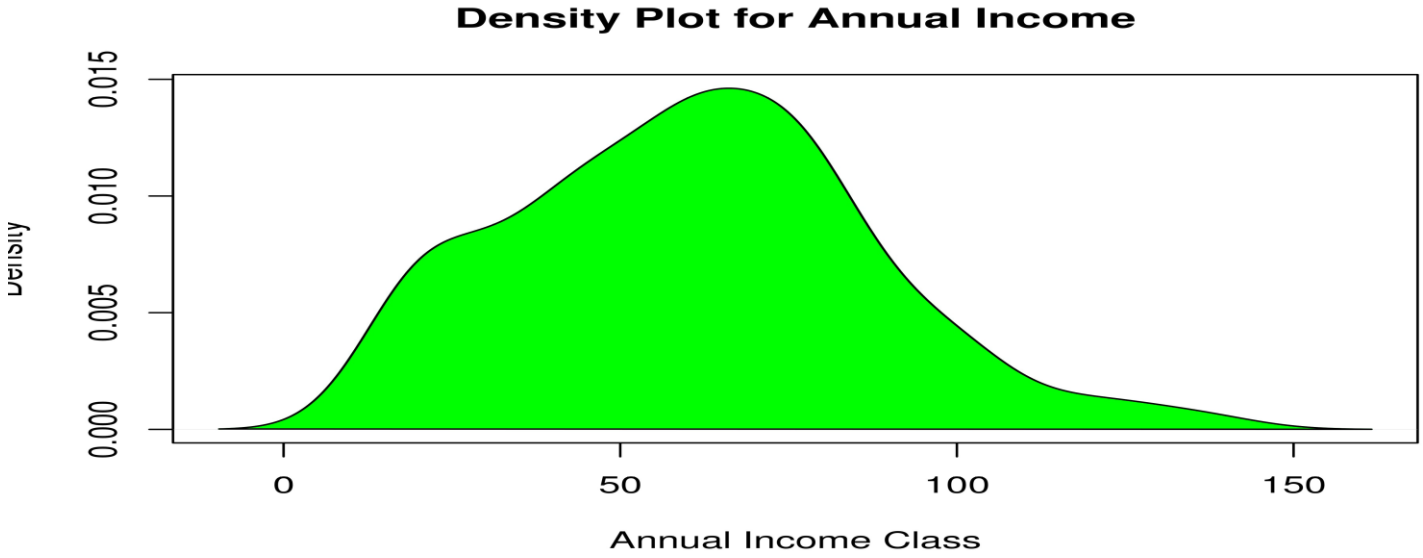




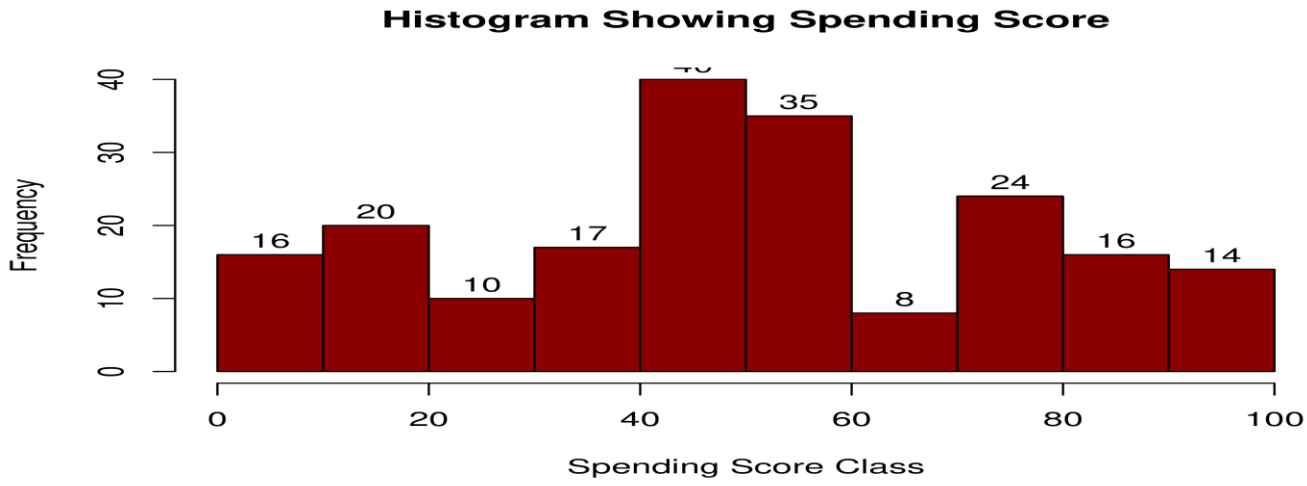
Further analysis was carried out to determine the income of the customers. It was observed that the minimum annual income of the customers was 15 and the maximum income was 137. People earning an average income of 70 had the highest frequency count. The average salary of all the customers was 60.56. The Kernel Density Plot displayed indicates that the annual income has a normal distribution.







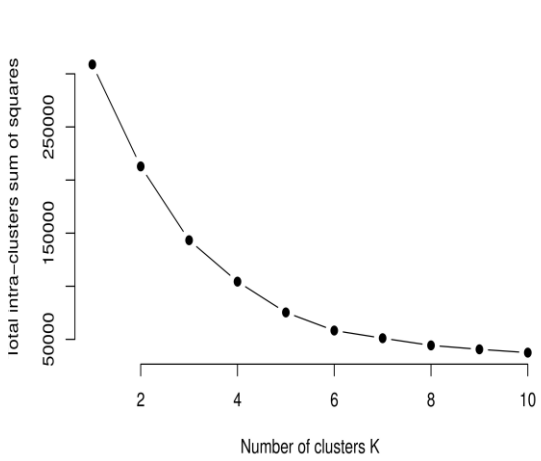
Spending score was also analyzed. It was observed that the minimum was 1, Max was 99 and average was 50.20. From the histogram, we conclude that customers between class 40 and 50 had the highest spending score among all the classes.

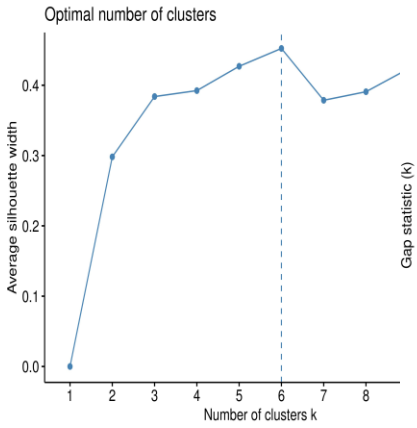


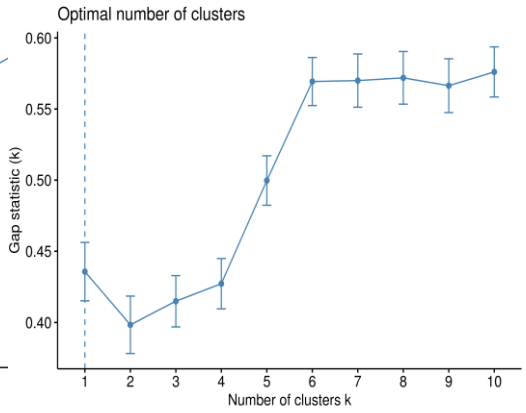
Dorcas Munoko

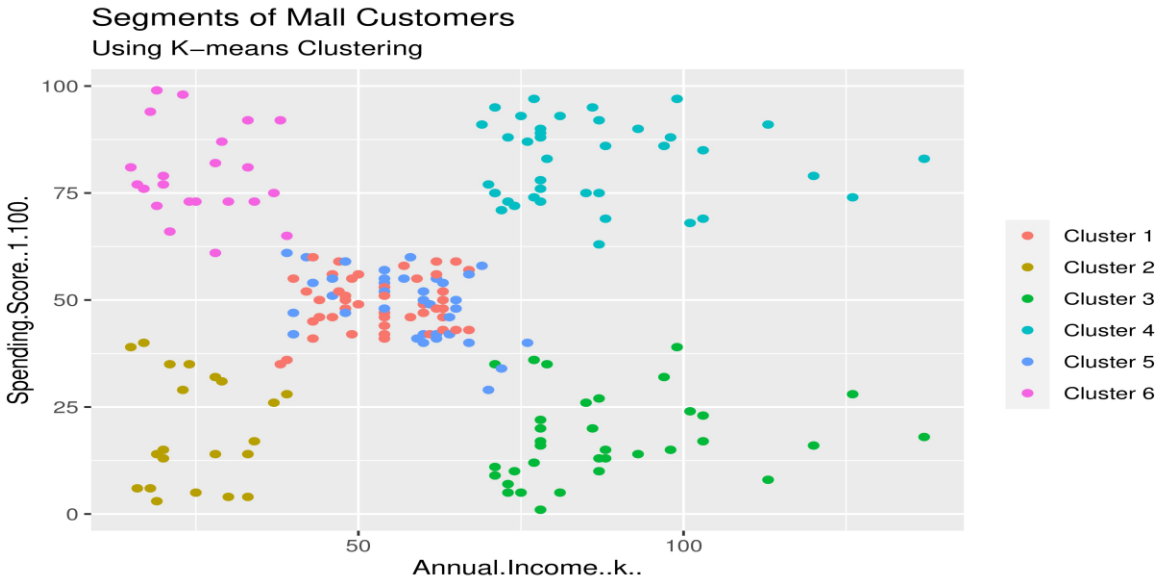
Models, Results, and Conclusion

K-means clustering algorithm was used. The elbow, silhouette, and gap statistic methods were used to determine the optimal clusters. With the elbow method, it was observed that 4 was the appropriate number of clusters since it appeared at the bend in the elbow plot. The average silhouette method determines how well within the cluster is the data object. A high average silhouette means that we have a good clustering. Finally, with the gap statistic method, it was observed that 6 was the appropriate number of clusters.









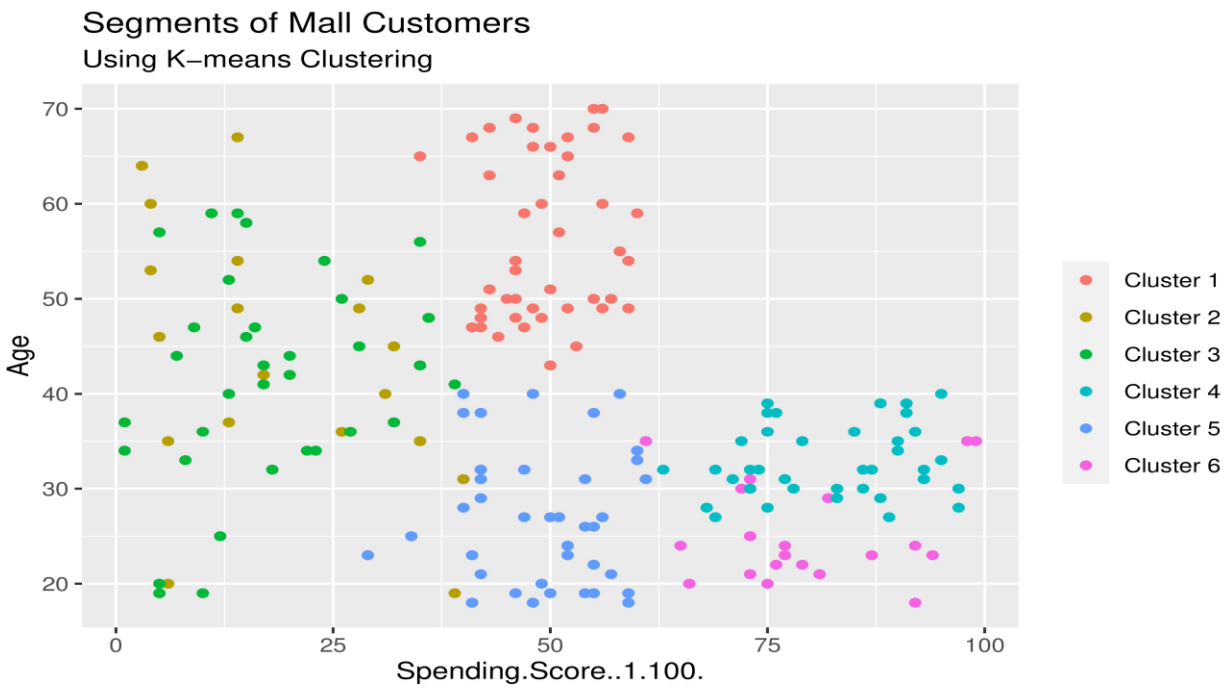
Cluster 6 and 4 – Represent customers with the medium income salary as well as the medium annual spend of salary.

Cluster 1 – Represents customers having a high annual income as well as a high annual spend.

Cluster 3 – Represents customers with low annual income as well as low yearly spend of income.

Cluster 2 – Represents a high annual income and low yearly spend.

Cluster 5 – Represents a low annual income but its high yearly expenditure.



Cluster 4 and 1 – These two clusters consist of customers with medium PCA1 and medium PCA2 score.

Cluster 6 – Represents customers having a high PCA2 and a low PCA1.

Cluster 5 – Represents customers with a medium PCA1 and a low PCA2 score.

Cluster 3 – Represents customers with a high PCA1 income and a high PCA2.

Cluster 2 – Represents customers with a high PCA2 and a medium annual spend of income.