# Clustering for Mall Customer data

The Mall customers data set (https://www.kaggle.com/datasets/kandij/mall-customers) contains some basic data about your customers like Customer ID, age, gender, annual income and spending score. In this project, we try to cluster the customers based on the basic data.

- Summary Statistics

| | Age | Annual_Income | Spending_Score |
|---|---|---|---|
| X | Min. :18.00 | Min. : 15.00 | Min. : 1.00 |
| X.1 | 1st Qu.:28.75 | 1st Qu.: 41.50 | 1st Qu.:34.75 |
| X.2 | Median :36.00 | Median : 61.50 | Median :50.00 |
| X.3 | Mean :38.85 | Mean : 60.56 | Mean :50.20 |
| X.4 | 3rd Qu.:49.00 | 3rd Qu.: 78.00 | 3rd Qu.:73.00 |
| X.5 | Max. :70.00 | Max. :137.00 | Max. :99.00 |

Table 1: Summary Statistics

- Matrix scatterplot and Correlation matrix

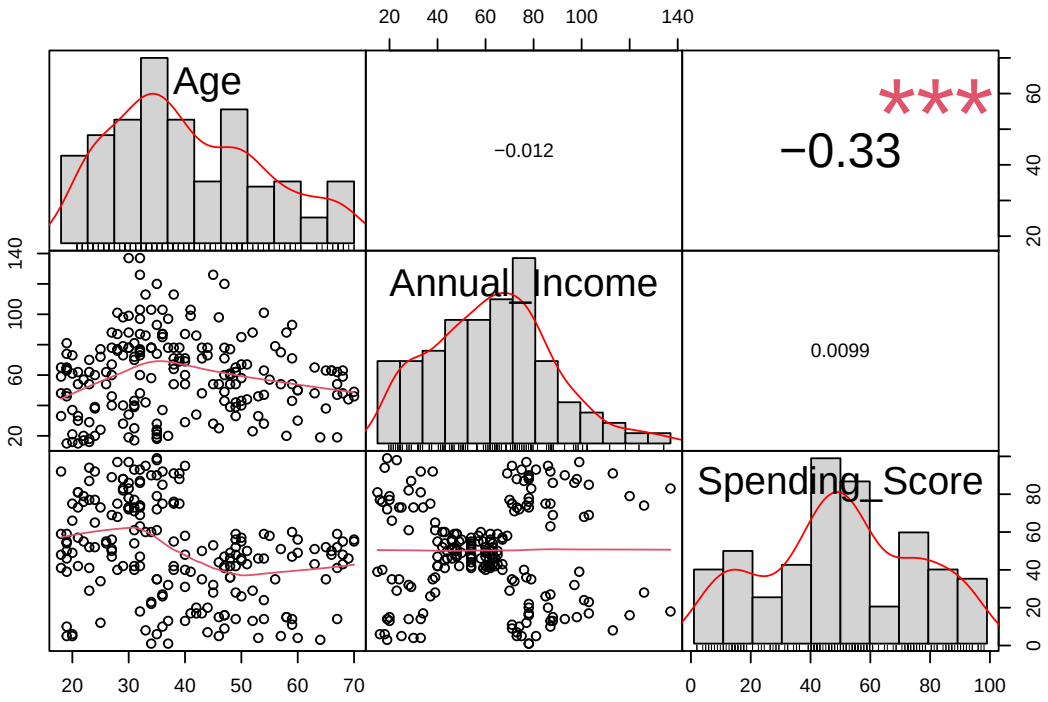


Figure 1: Matrix scatterplot

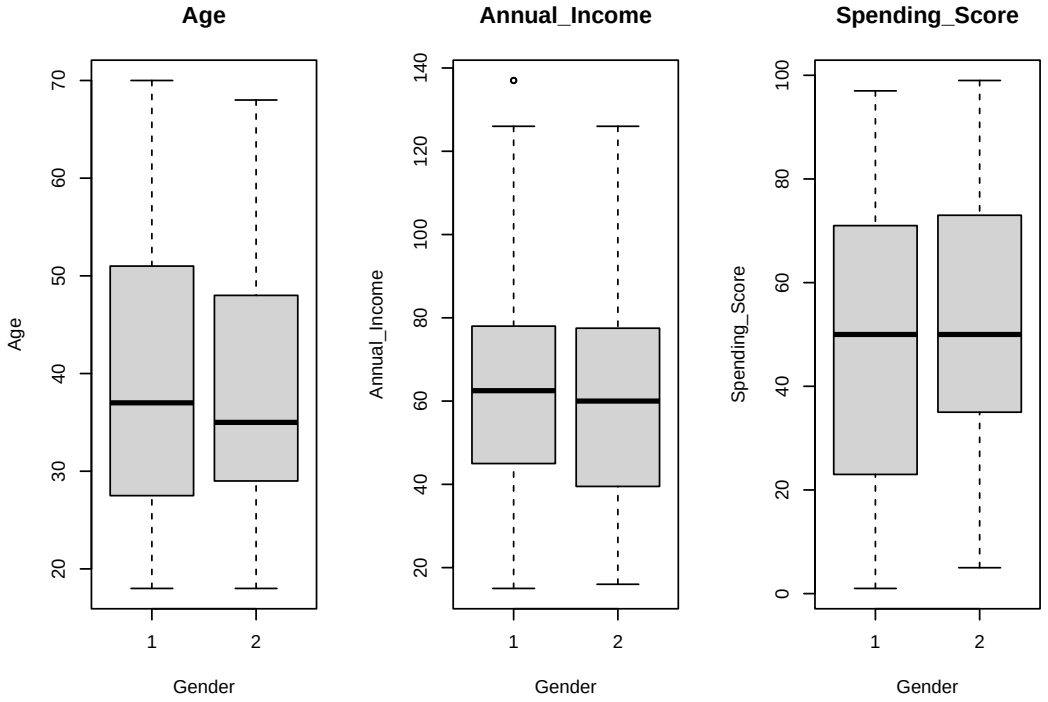


Figure 2: Box Plots

| | Age | Annual_Income | Spending_Score |
|---|---|---|---|
| Age | 1.00 | -0.01 | -0.33 |
| Annual_Income | -0.01 | 1.00 | 0.01 |
| Spending_Score | -0.33 | 0.01 | 1.00 |

Table 2: Correlation matrix

- As the correlation between most of the variables are very low, therefore, rather than going with the correlation-based distance, we can use metric-based distance for clustering.
- Also I would suggest standardizing the variables as they are in different scales and some have very high ranges.
- There is not much variation of the variables when the Gender is condisered.

**Hierarchical Clustering with Complete linkage**

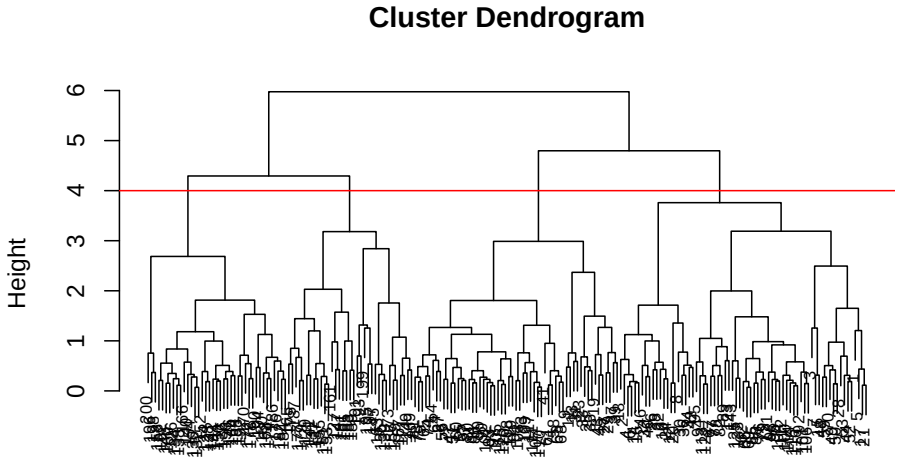


Figure 3: Hierarchical Clustering with Complete linkage

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 69 | 57 | 74 |

Table 3: Number of observations within each cluster (Hierarchical Clustering)

| | Age | Annual_Income | Spending_Score | clust |
|---|---|---|---|---|
| Cluster 1 | 27.275 | 42.783 | 56.377 | 1.000 |
| Cluster 2 | 55.333 | 47.316 | 41.088 | 2.000 |
| Cluster 3 | 36.946 | 87.338 | 51.459 | 3.000 |

Table 4: Cluster means of the variables (Hierarchical Clustering)

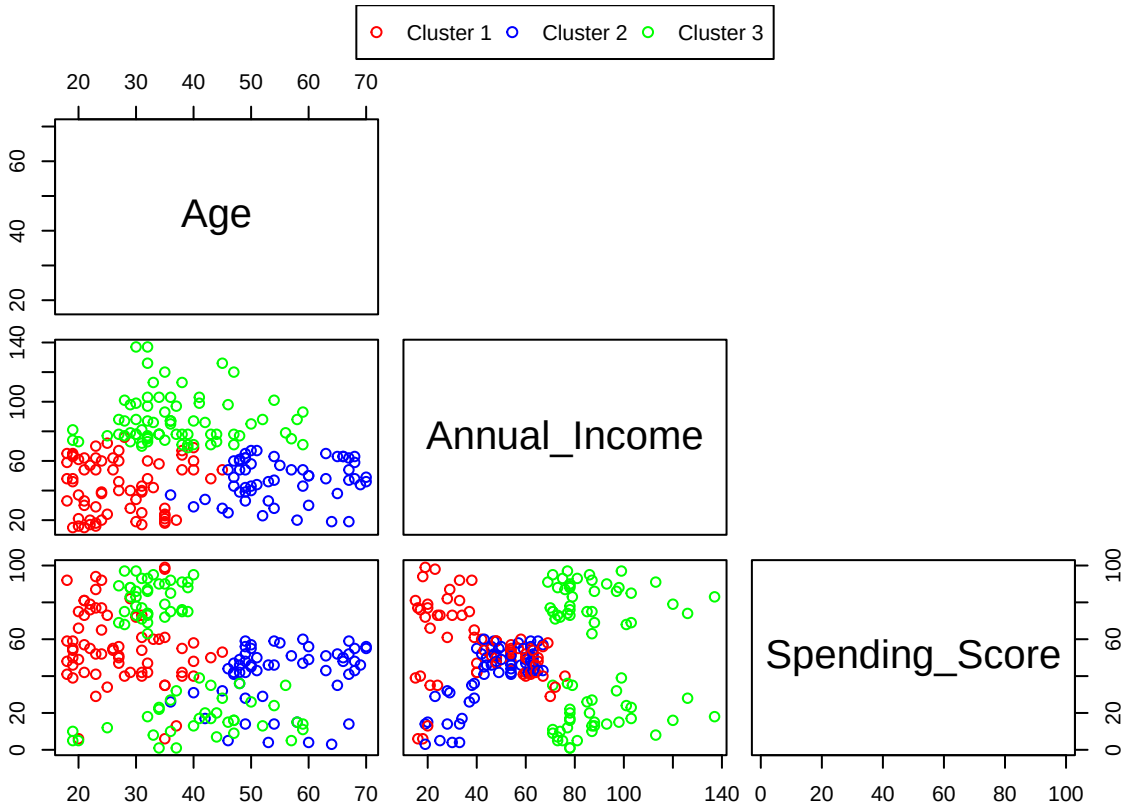


Figure 4: Matrix scatter plots

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 69 | 57 | 39 | 35 |

Table 5: Number of observations within each cluster (Hierarchical Clustering)

| | Age | Annual_Income | Spending_Score | clust |
|---|---|---|---|---|
| Cluster 1 | 27.275 | 42.783 | 56.377 | 1.000 |
| Cluster 2 | 55.333 | 47.316 | 41.088 | 2.000 |
| Cluster 3 | 32.692 | 86.538 | 82.128 | 3.000 |
| Cluster 4 | 41.686 | 88.229 | 17.286 | 4.000 |

Table 6: Cluster means of the variables (Hierarchical Clustering)

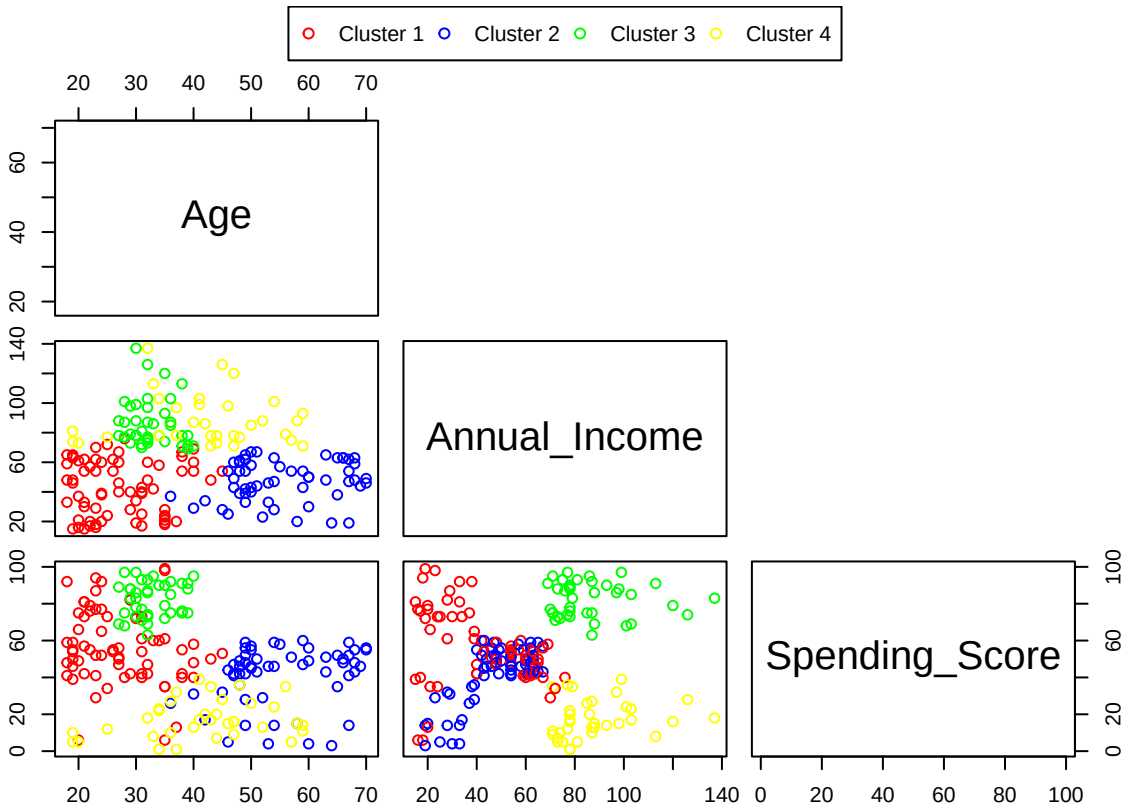


Figure 5: Matrix scatter plots

**K-means clustering with k = 3**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 91 | 41 | 68 |

Table 7: Number of observations within each cluster (k=3)

| | Age | Annual_Income | Spending_Score |
|---|---|---|---|
| Cluster 1 | 51.275 | 61.802 | 34.209 |
| Cluster 2 | 32.854 | 87.341 | 79.976 |
| Cluster 3 | 25.838 | 42.750 | 53.647 |

Table 8: Cluster means of the variables for k-means clustering (k=3)

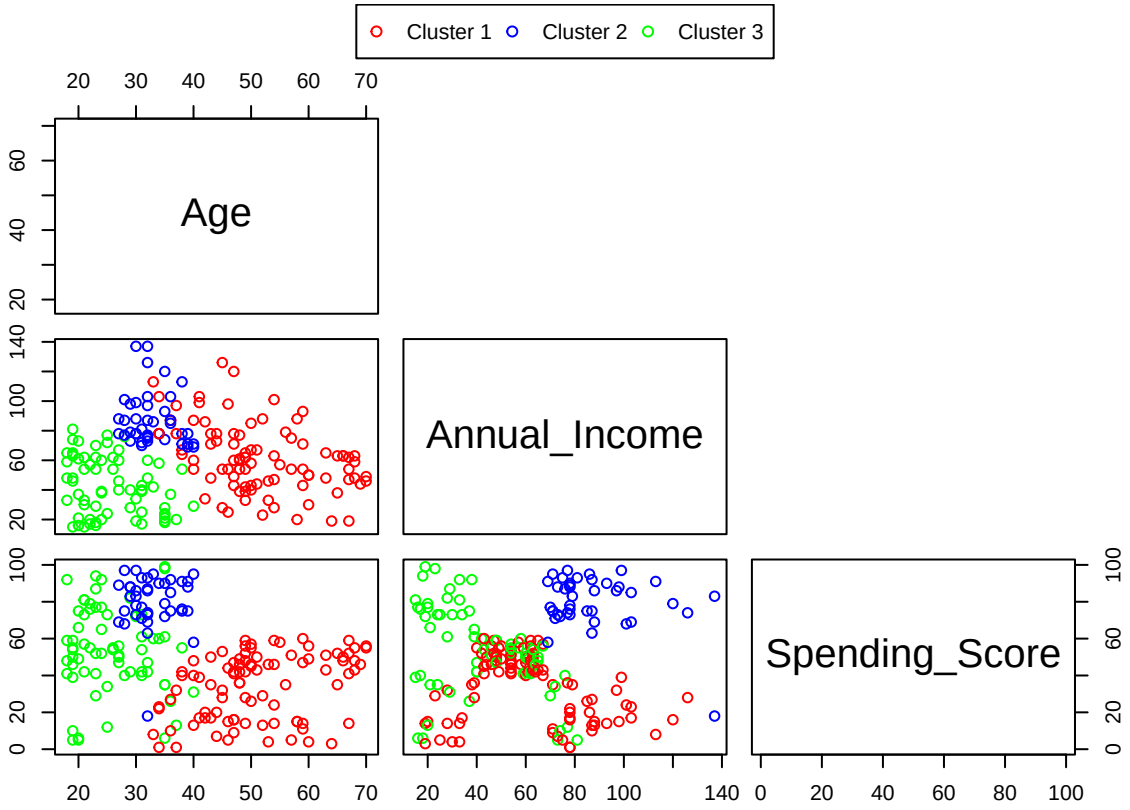


Figure 6: Matrix scatter plots

- Age is a good variable to cluster the data into 3 clusters.

**K-means clustering with k = 4**

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 65 | 38 | 57 | 40 |

Table 9: Number of observations within each cluster (k=4)

| | Age | Annual_Income | Spending_Score |
|---|---|---|---|
| Cluster 1 | 53.985 | 47.708 | 39.969 |
| Cluster 2 | 39.368 | 86.500 | 19.579 |
| Cluster 3 | 25.439 | 40.000 | 60.298 |
| Cluster 4 | 32.875 | 86.100 | 81.525 |

Table 10: Cluster means of the variables for k-means clustering (k=4)

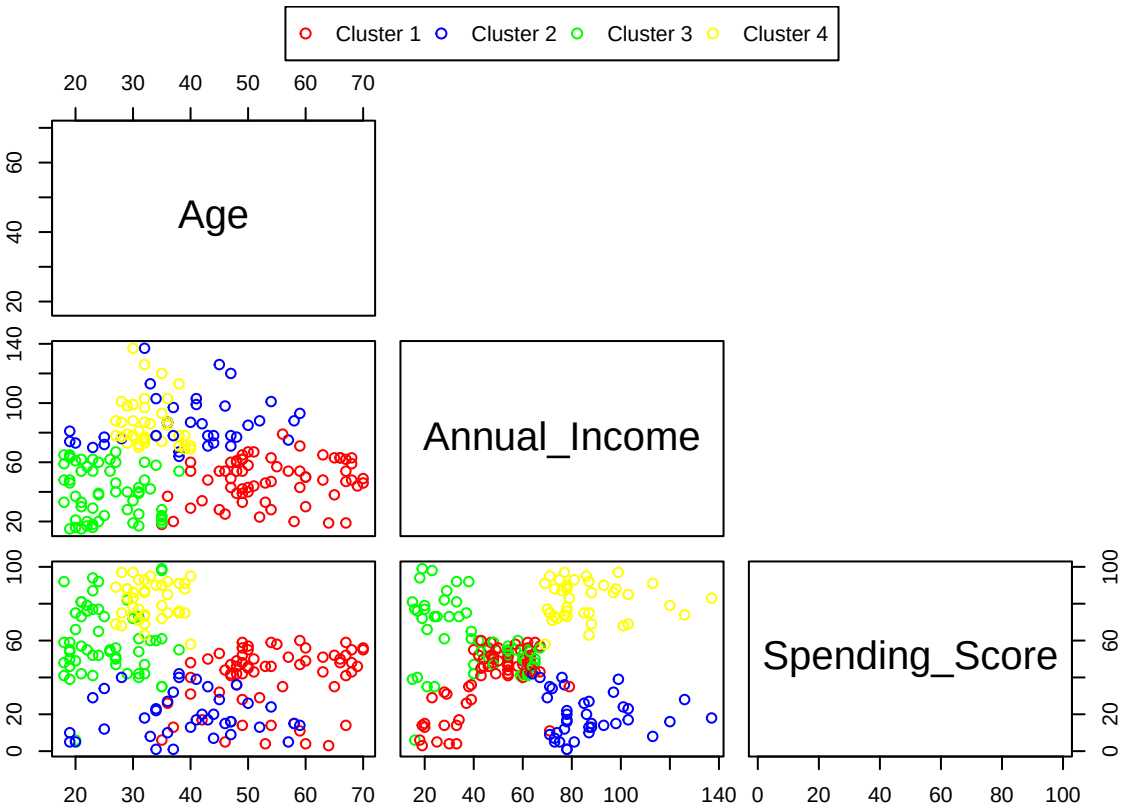


Figure 7: Matrix scatter plots

**K-means clustering with k = 5**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 20 | 39 | 54 | 47 | 40 |

Table 11: Number of observations within each cluster (k=4)

| | Age | Annual_Income | Spending_Score |
|---|---|---|---|
| Cluster 1 | 46.250 | 26.750 | 18.350 |
| Cluster 2 | 39.872 | 86.103 | 19.359 |
| Cluster 3 | 25.185 | 41.093 | 62.241 |
| Cluster 4 | 55.638 | 54.383 | 48.851 |
| Cluster 5 | 32.875 | 86.100 | 81.525 |

Table 12: Cluster means of the variables for k-means clustering (k=5)

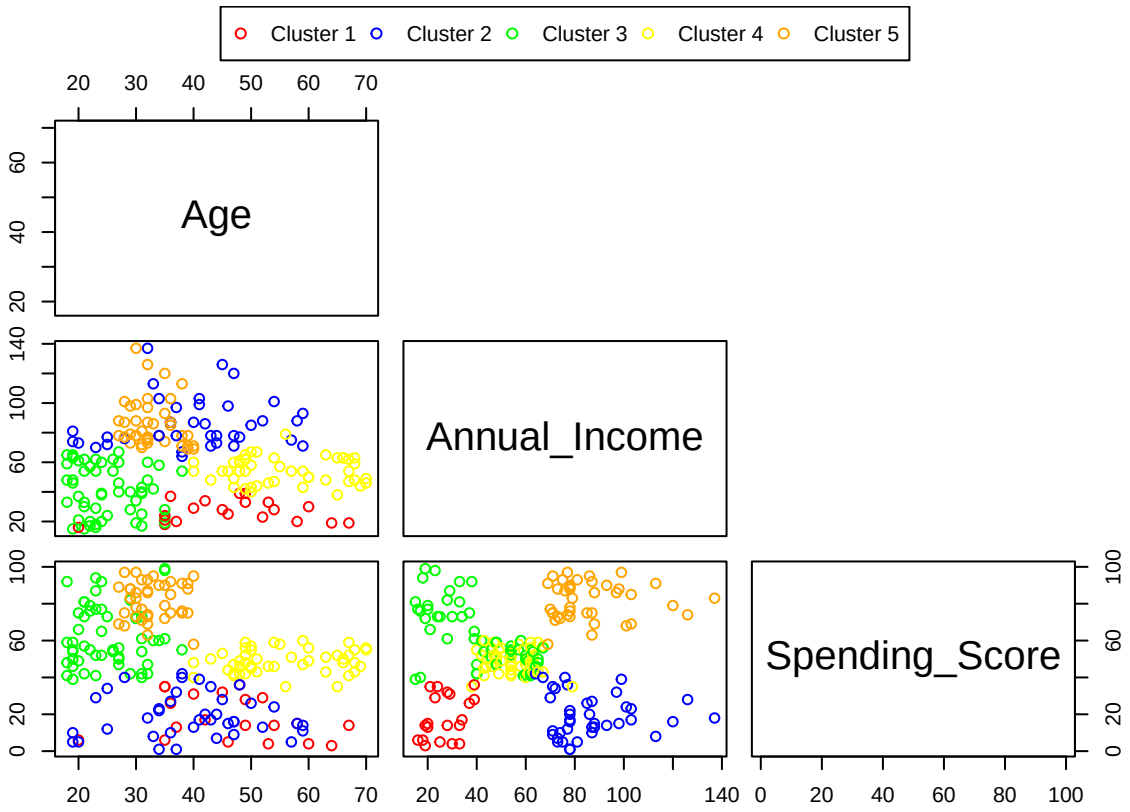


Figure 8: Matrix scatter plots

**K-means clustering with k = 6**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 24 | 45 | 39 | 33 | 21 | 38 |

Table 13: Number of observations within each cluster (k=4)

| | Age | Annual_Income | Spending_Score |
|---|---|---|---|
| Cluster 1 | 25.250 | 25.833 | 76.917 |
| Cluster 2 | 56.333 | 54.267 | 49.067 |
| Cluster 3 | 32.692 | 86.538 | 82.128 |
| Cluster 4 | 41.939 | 88.939 | 16.970 |
| Cluster 5 | 45.524 | 26.286 | 19.381 |
| Cluster 6 | 26.684 | 57.579 | 47.789 |

Table 14: Cluster means of the variables for k-means clustering (k=6)


Figure 9: Matrix scatter plots

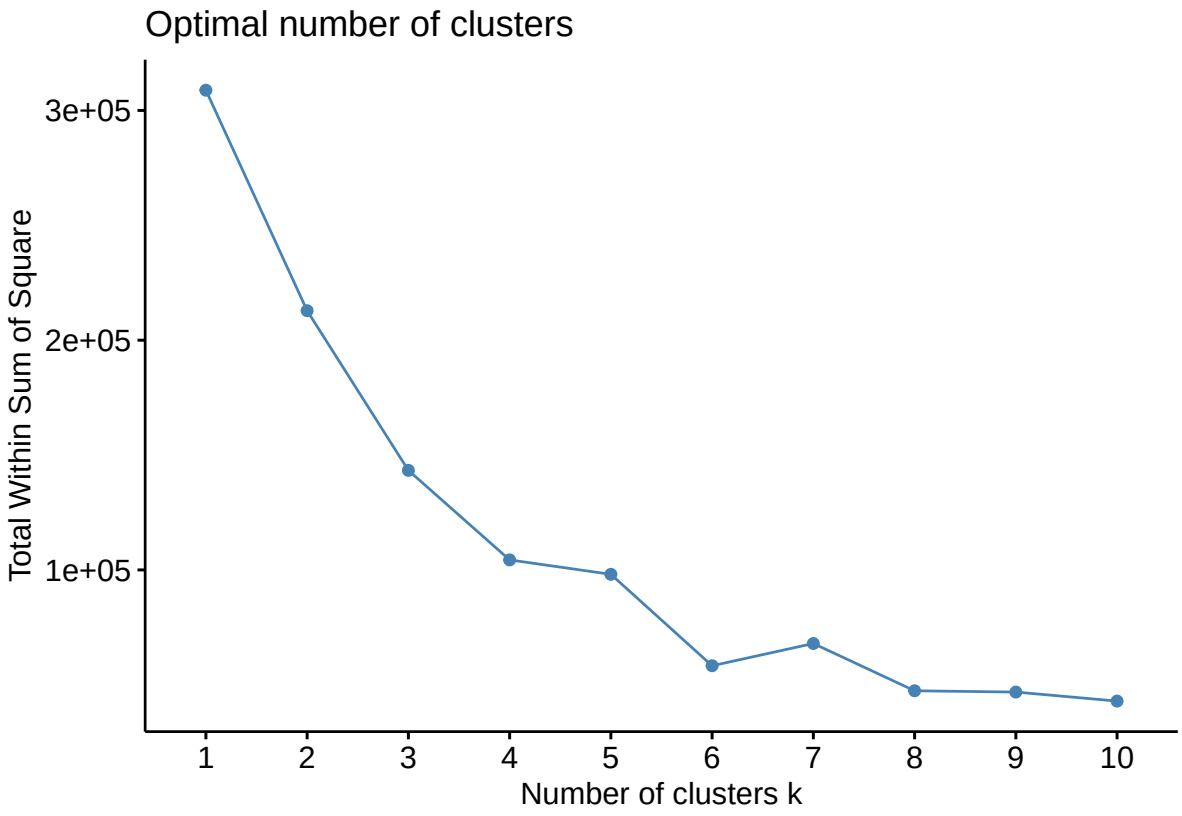


Figure 10: Plot of number of clusters vs total within cluster sum of squares