# K_Means_Cluster

Joy Naik

## K-means Clustering Project in R

## Introduction

This project applies K-means clustering to a Mall Customers dataset using R to identify distinct clusters within the data. The project demonstrates the ability to preprocess data, apply K-means clustering, and interpret the results in R.

## Dataset

- **Source:** https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python/data

- **Description:** The dataset has total 5 columns, 3 numerical(Age, Income and Spending score), one categorical(gender) and the other is Customer ID.

## Project Workflow

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

   - The dataset didn't contain and missing or duplicate values.

   - We will be using the 3 numerical columns: Age, Income and Spending score as K-means is ran on numerical values.

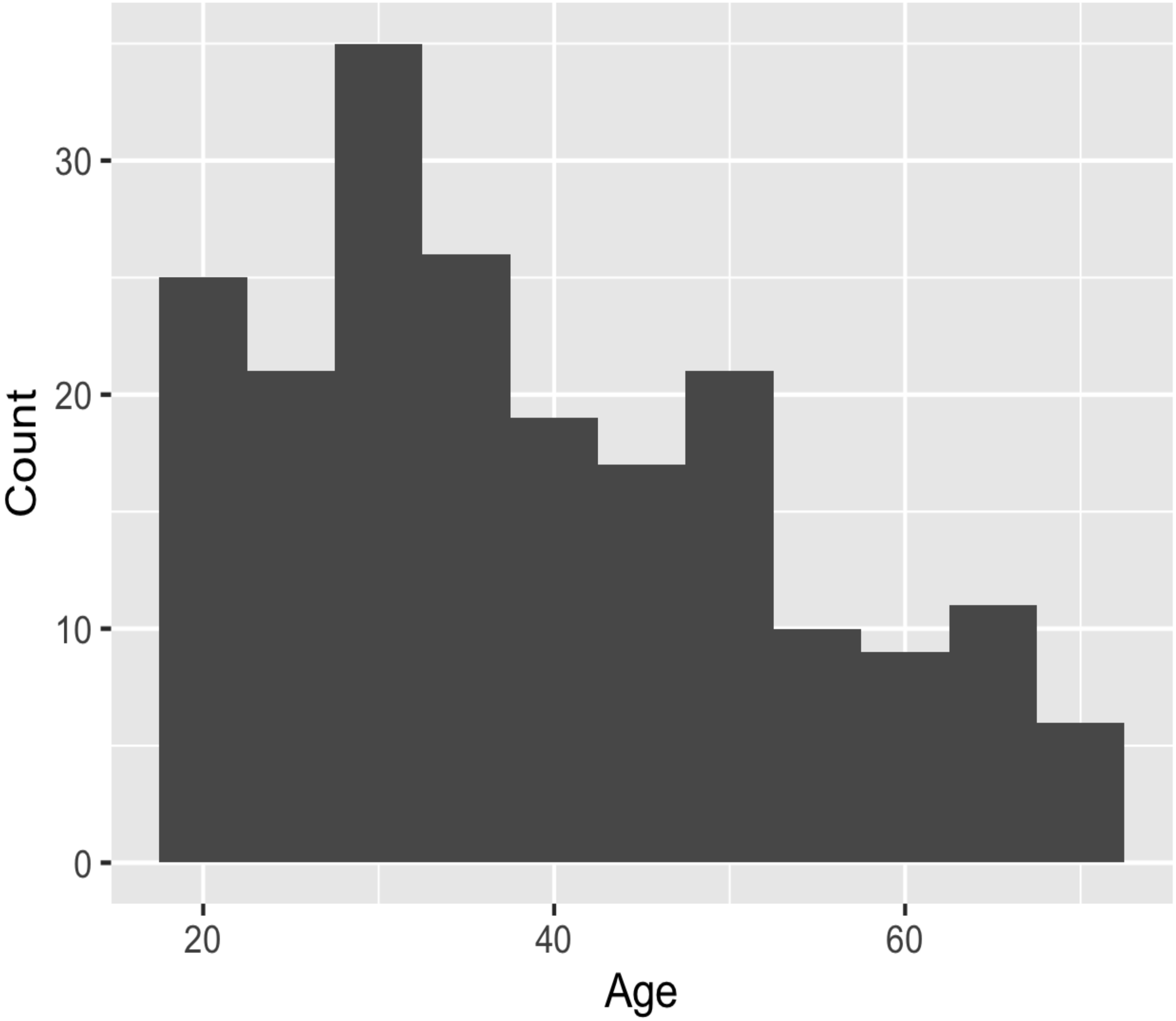
     ```
     quantdf <- mdf[c(-1,-2)]
     ```

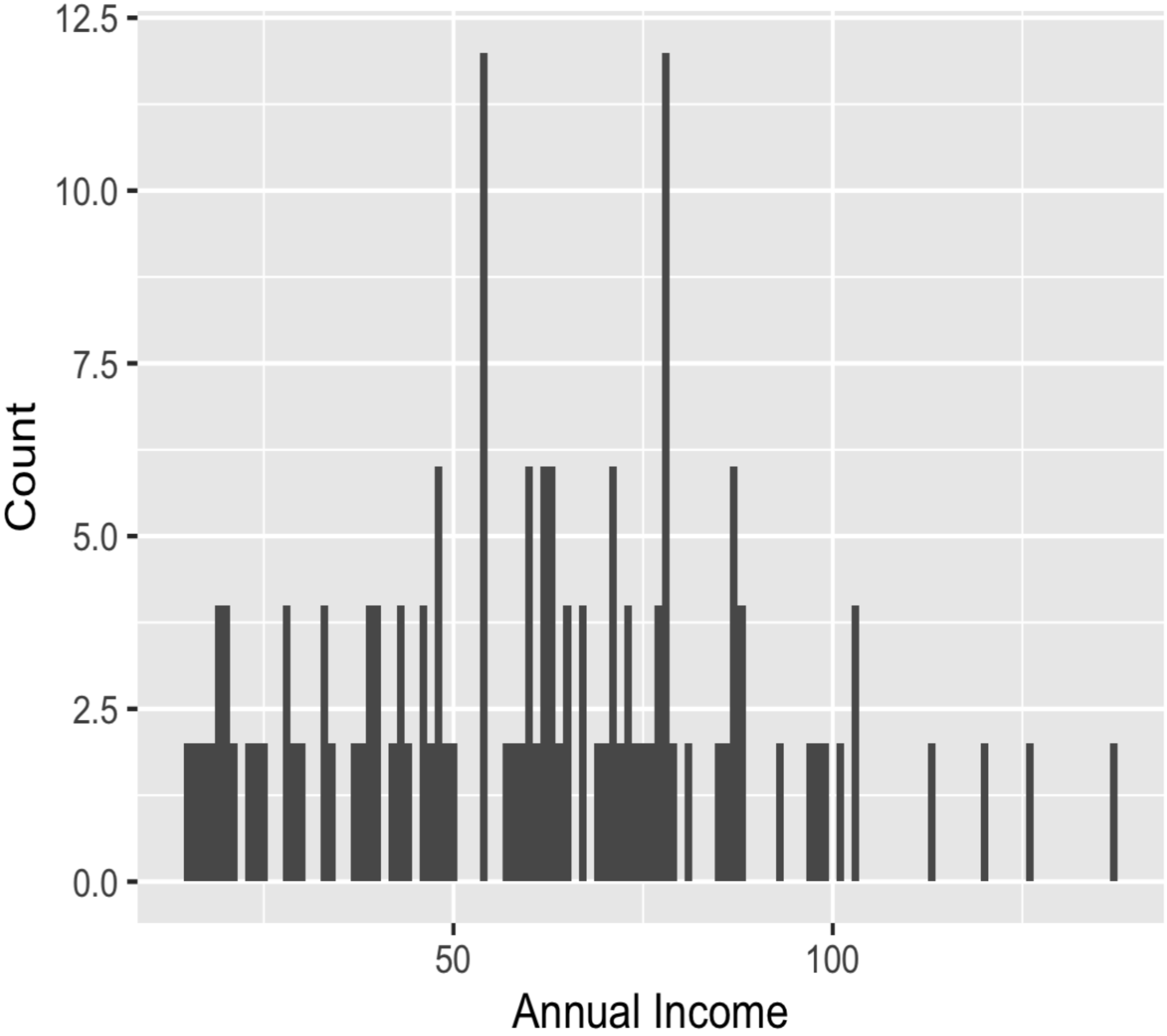
   - On checking the histograms, no major skewness is found in any of the features.

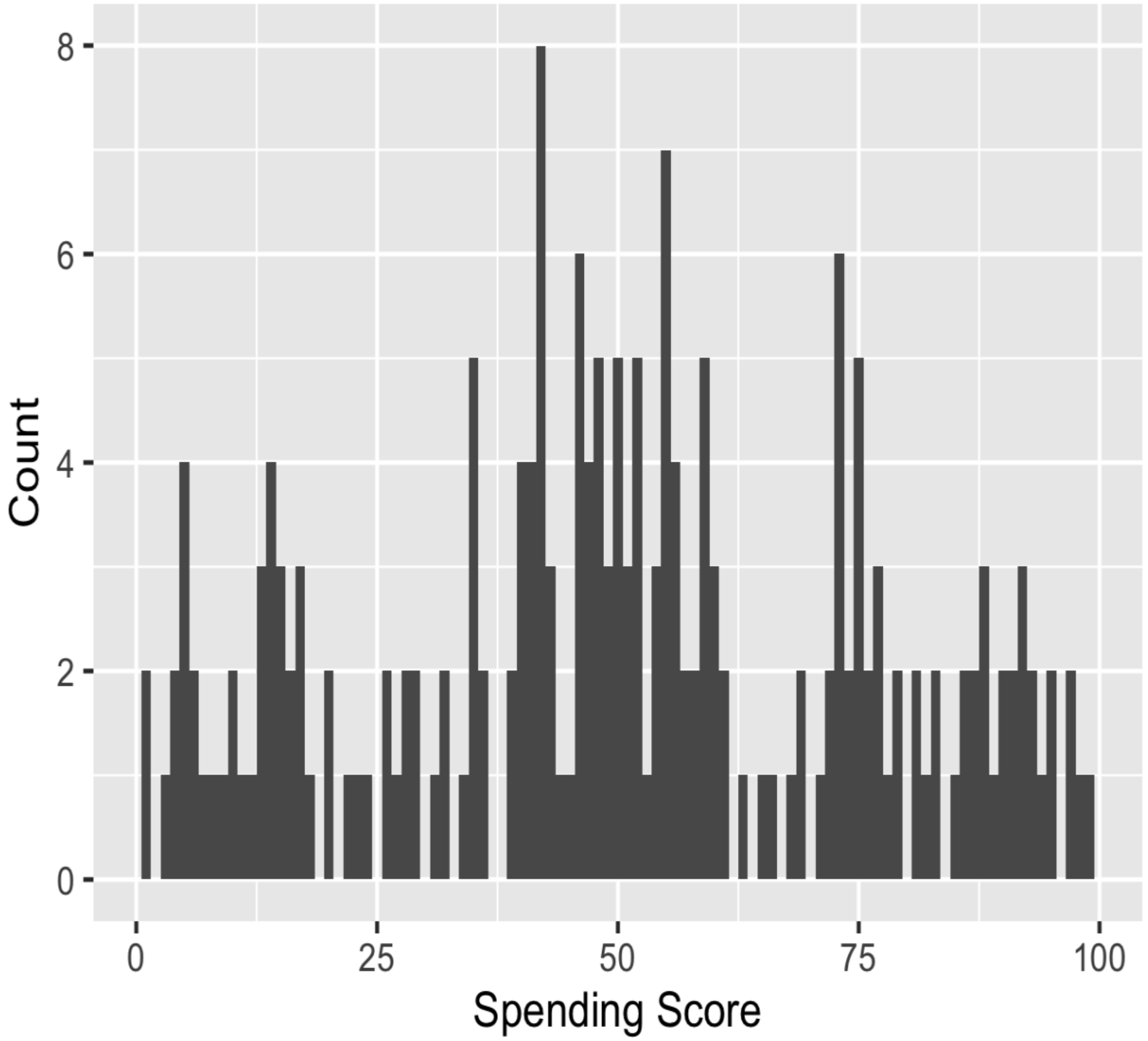
     ```
     ggplot(data = mdf) +
     geom_histogram(mapping = aes(x= quantdf$Age),
     binwidth = 5) +
     labs(x = "Age", y = "Count")

     ggplot(data = mdf) +
     geom_histogram(mapping = aes(x= quantdf$`Annual
     Income (k$)`), binwidth = 1) +
      labs(x = "Annual Income", y = "Count")

     ggplot(data = mdf) +
     geom_histogram(mapping = aes(x= quantdf$`Spending
     Score (1-100)`), binwidth = 1) +
     abs(x = "Spending Score", y = "Count")
     ```

- We will standardize or normalize the data to have same units and store it in a new data frame.

```
quantdfn <- scale(quantdf)
```


Count
30
20
10
0
20
40
60
Age


12.5
10.0
7.5
5.0
2.5
0.0
Count
50
100
Annual Income


Count
0
2
4
6
8
0
25
50
75
100
Spending Score

## 2. K-means Clustering

- First, we will create a custom function for within sum of square(wss) which returns the WSS for each iteration.

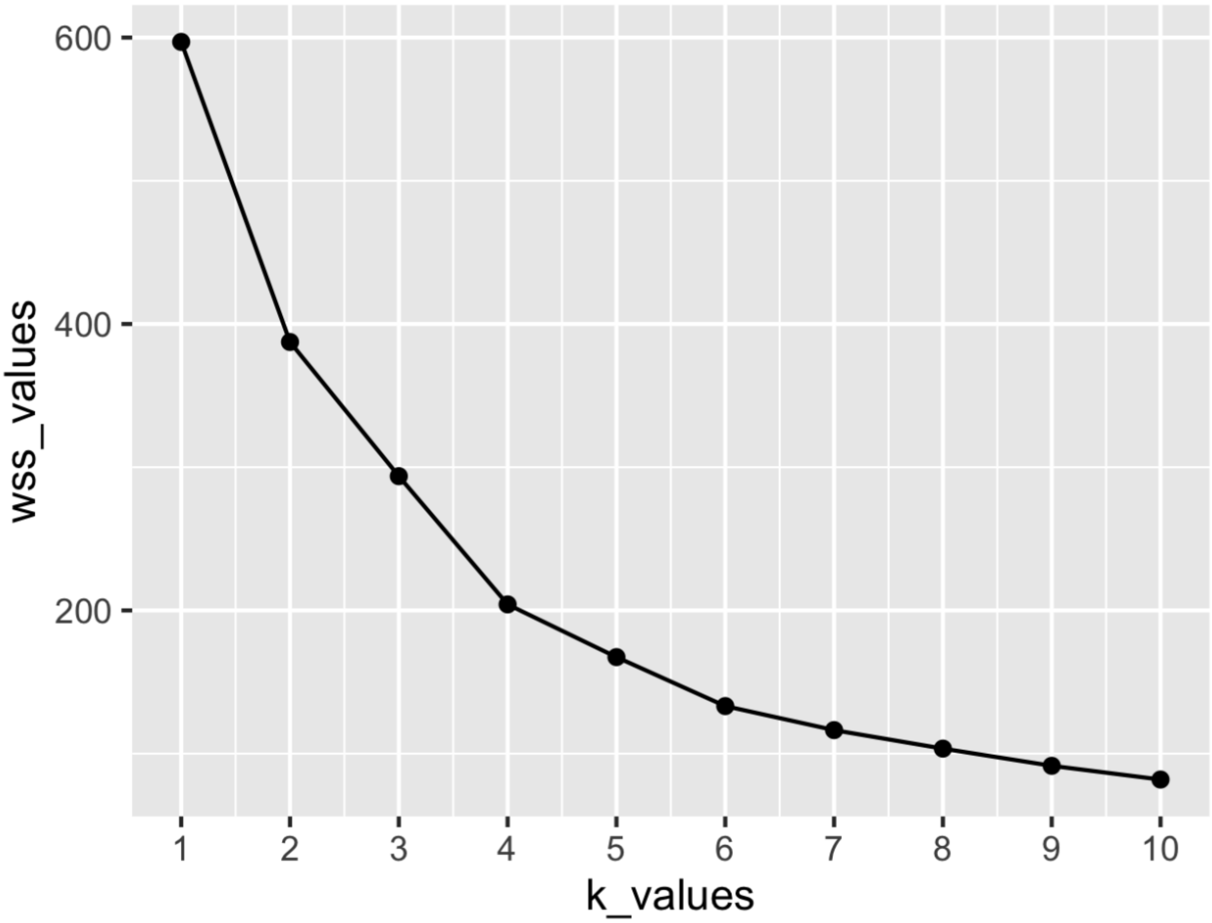
```
wss <- function(k){

kmeans(quantdfn, k, nstart= 10) } $tot.withinss
```

- Next, we will have k values in the range 1:10 to determine the number of clusters.

```
k_values <- 1:10
```

- We will perform wss for each value of k and plot an elbow plot.

```
wss_values <- map_dbl(k_values, wss)

elbowdf <- data.frame(k_values, wss_values)


ggplot(elbowdf, mapping = aes(x = kvalues,

y = wss_values)) +

geom_line() +

geom_point()
```



- From the elbow plot, the optimal number of clusters seems to be 4, we will be using 4 clusters to segment our customers.
- Finally, we will run kmeans for 4 clusters with nstart of 1000 to iterate 1000 different initial points.

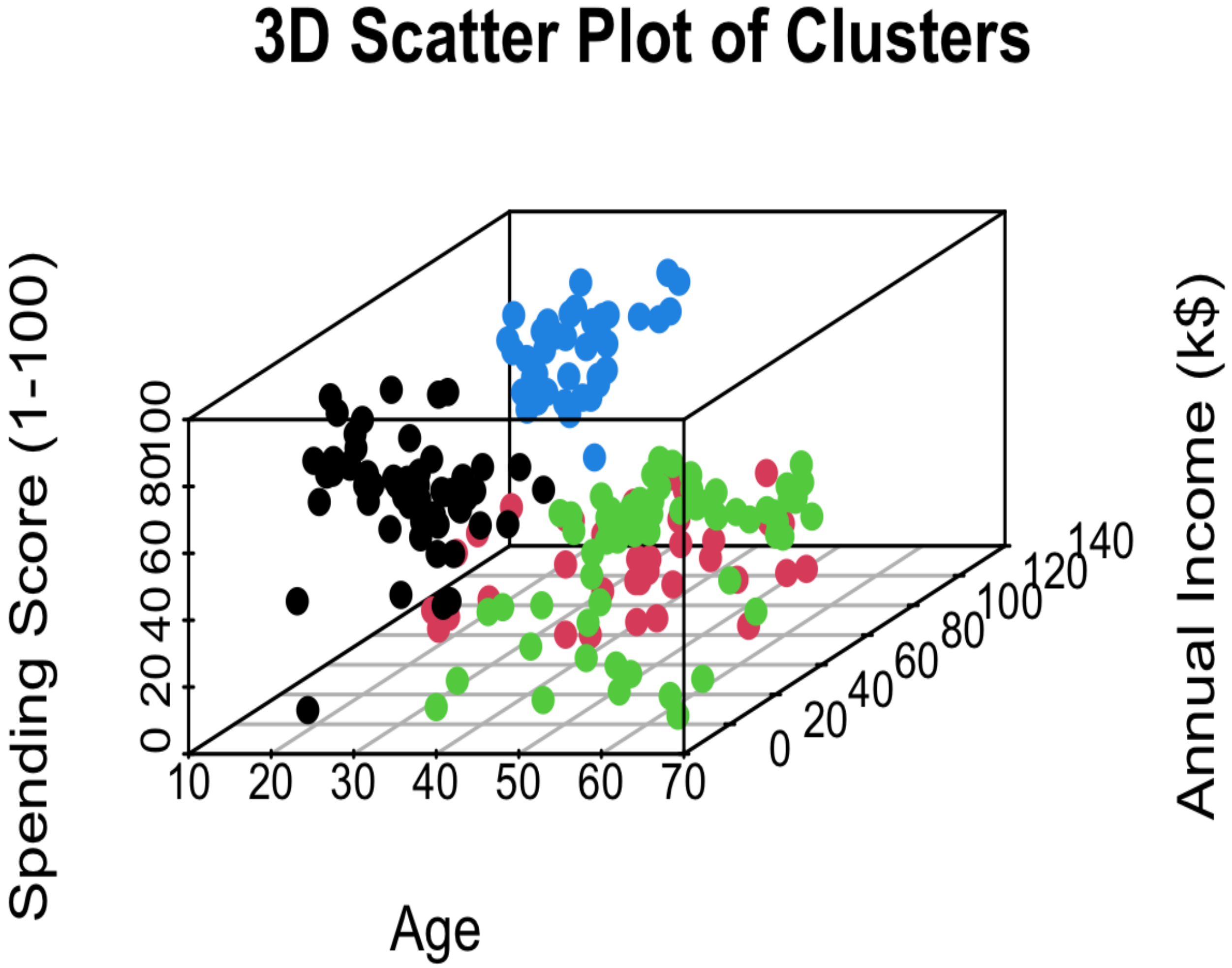
  ```
  k4 <- kmeans(quantdfn, 4, nstart = 1000)
  ```

4. **Visualization**

   We can bind the cluster ID column to the dataset used for kmeans and store it a new data frame quantdfk4

   - We will use 3D plot to visualise the clusters.

```
scatterplot3d(quantdf$Age, quantdf $Annual
Income (k$), quantdf$`Spending Score (1-100)`,
color = quantdfk4$clusterID,
pch = 16, main = “3D Scatter Plot of Clusters”,
xlab = “Age”,
ylab = “Annual Income (k$)”,
zlab = “Spending Score (1-100)”)
```



## 5. **Results and Interpretation**

We can use cluster.stats to examine the characteristics of clusters.

```
cluster.stats(dist(quantdfn, method=“euclidean”),k4$cluster)
```

The ratio of average between clusters and average within cluster is around 2, which shows that the centroids of the clusters are apart from each other while the within cluster distance is less, hence the clusters are distinct.

We will summarize the clusters based on the mean of the clusters and compare it with the mean of our original data.

```
quantdfk4 %>%
group_by(clusterID) %>%
summarise_all(mean)
```

```
summarise_all(quantdf, mean)
```

## 6. Conclusion:

Average of the data

| Age | Income(k) | Spending Score |
|---|---|---|
| 38.8 | 60.6 | 50.2 |

Average of the clusters

| Cluster ID | Age | Income(k) | Spending Score |
|---|---|---|---|
| 1 | 25.4 | 40 | 60.3 |
| 2 | 39.4 | 86.5 | 19.6 |
| 3 | 54.0 | 47.7 | 40.0 |
| 4 | 32.9 | 86.1 | 81.5 |

**Cluster 1:** Young individuals with low income and above average spending score: These are young individuals who just started their careers, making less income, however they have good spending score indicating that they are willing to buy products inspite of low earnings. Trendy and attrative items can be marketed to such individuals.

**Cluster 2:** Individuals with average age, high income and low spending score: These are mid-aged individuals who are well settled in their careers but are reluctant to buy products or planning for the future or kids. Products related to investments or future planning can be marketed to them as they don't wish to spend on unnecessary items.

**Cluster 3:** Mid to older age individuals with below average income and little below average spending score: These individuals are approaching retirement and not making much income but has around average spending score. Products related to healthcare, retirement planning and children college programs can be marketed to them as they are willing to buy but little conservative.

**Cluster 4:** Below average age, very high income and very high spending score: These individuals have below average age but seems to be doing well with their careers, they earn high income and are also willing to spend more. Luxury and expensive items can be marketed to such individuals.

## Requirements

- R packages used in the project are `dplyr`, `ggplot2`, `cluster`, fpc, scatterplot3d etc.