Data Science Final Project Report

**Project Title : ** Supermarket customer data segmentation
Submitted By : Rubaba Binte Rahman

ID : ** 20-42182-1 , **Section - D

Project Description:

I took a dataset of 200 individuals from a local supermarket who have paid bills through membership cards. In the dataset there are 5 attributes such as – customer ID, gender, age, annual income and spending score. Spending score is something that assigned by the customers based on their defined parameters like customer behavior and purchasing data. To understand the customer's need like who can be easily converge [Target Customers]. We applied K-means clustering algorithm to extract information and gather insights for finding target customer .The result of the analysis can be given to marketing team and plan the strategy accordingly.

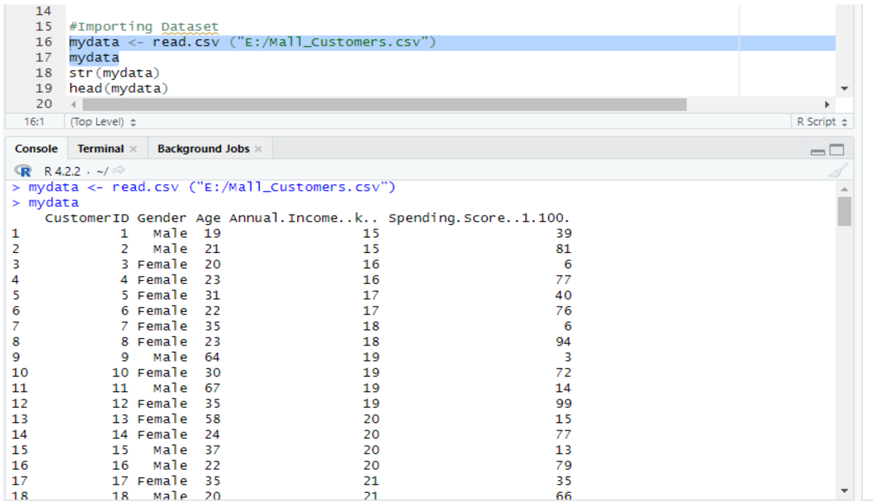
Dataset From Kaggle : https://www.kaggle.com/datasets/vjchoudhary7

1. Installing necessary packages

```
#Installing necessary packeages
install.packages("stats")
install.packages("dplyr")
install.packages("factoextra")
install.packages("ggfortify")
install.packages("NbClust")
library(stats)
library(dplyr)
library(factoextra)
library(ggfortify)
library(NbClust)
```

2. Imported Dataset and Summarized

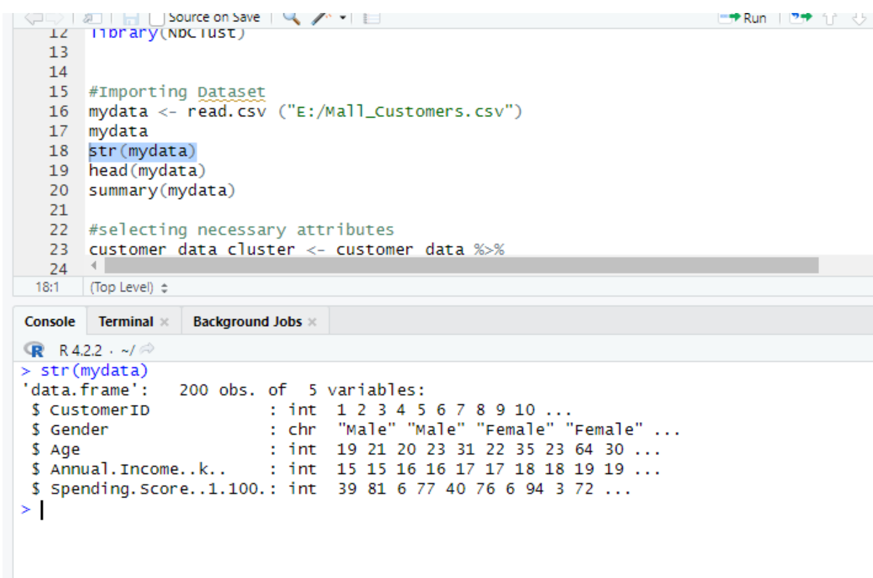
mydata <- read.csv ("E:/Mall_Customers.csv")
mydata



```
#Importing Dataset
mydata <- read.csv ("E:/Mall_Customers.csv")
mydata
str(mydata)
head(mydata)
```

```
> mydata <- read.csv ("E:/Mall_Customers.csv")
> mydata
   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
1           1   Male  19                 15                     39
2           2   Male  21                 15                     81
3           3 Female  20                 16                      6
4           4 Female  23                 16                     77
5           5 Female  31                 17                     40
6           6 Female  22                 17                     76
7           7 Female  35                 18                      6
8           8 Female  23                 18                     94
9           9   Male  64                 19                      3
10         10 Female  30                 19                     72
11         11   Male  67                 19                     14
12         12 Female  35                 19                     99
13         13 Female  58                 20                     15
14         14 Female  24                 20                     77
15         15   Male  37                 20                     13
16         16   Male  22                 20                     79
17         17 Female  35                 21                     35
18         18   Male  20                 21                     66
```

#List of objects
str(mydata)



```
library(NbClust)

#Importing Dataset
mydata <- read.csv ("E:/Mall_Customers.csv")
mydata
str(mydata)
head(mydata)
summary(mydata)

#selecting necessary attributes
customer_data_cluster <- customer_data %>%
```

```
> str(mydata)
'data.frame':	200 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : chr  "Male" "Male" "Female" "Female" ...
 $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
> 
```

#list of objects in a dataframe
head(mydata)

```
mydata
str(mydata)
head(mydata)
summary(mydata)

#selecting necessary attributes
customer_data_cluster <- customer_data %>%
```

```
19:1   (Top Level)

Console   Terminal   Background Jobs

R 4.2.2 · ~/
> head(mydata)
  CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
1          1   Male  19                 15                     39
2          2   Male  21                 15                     81
3          3 Female  20                 16                      6
4          4 Female  23                 16                     77
5          5 Female  31                 17                     40
6          6 Female  22                 17                     76
>
```

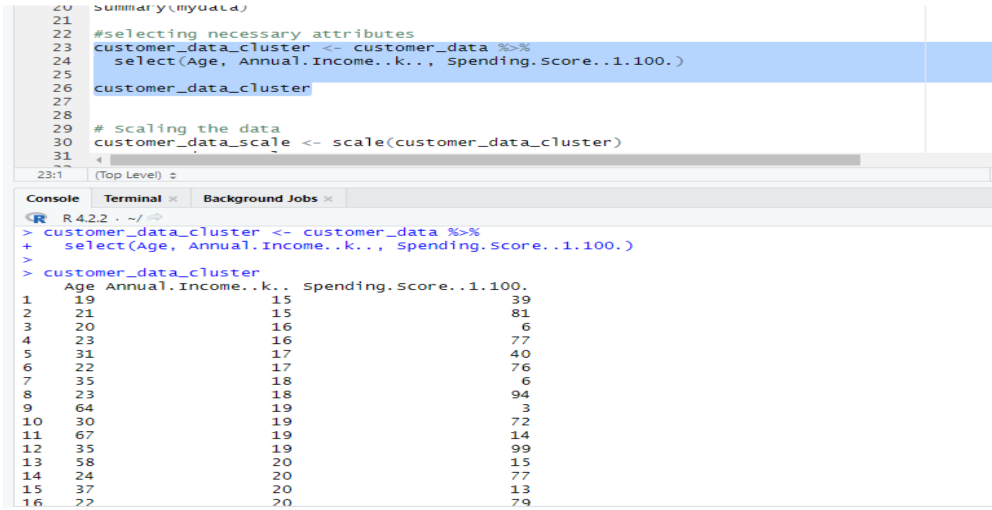
#Summary of attributes
summary(mydata)

```
str(mydata)
head(mydata)
summary(mydata)

#selecting necessary attributes
customer_data_cluster <- customer_data %>%
```

```
20:1   (Top Level)

Console   Terminal   Background Jobs

R 4.2.2 · ~/
> summary(mydata)
   CustomerID        Gender               Age        Annual.Income..k..  Spending.Score..1.100.
 Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00     Min.   : 1.00
 1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50     1st Qu.:34.75
 Median :100.50   Mode  :character   Median :36.00   Median : 61.50     Median :50.00
 Mean   :100.50                      Mean   :38.85   Mean   : 60.56     Mean   :50.20
 3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00     3rd Qu.:73.00
 Max.   :200.00                      Max.   :70.00   Max.   :137.00     Max.   :99.00
>
```

3. Selecting necessary attributes

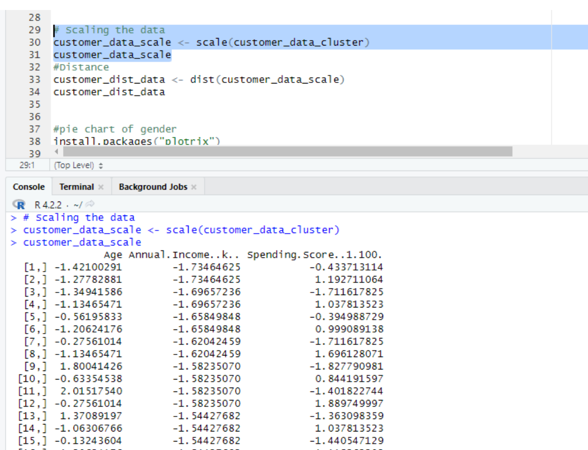
customer_data_cluster <- customer_data %>%
select(Age, Annual.Income..k.., Spending.Score..1.100.)
customer_data_cluster

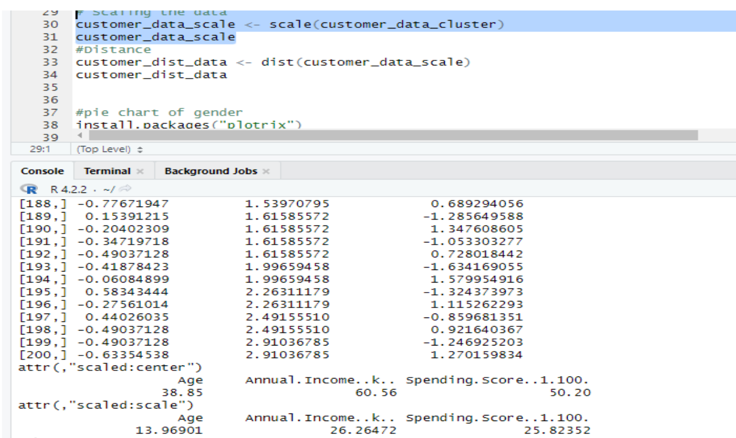


4. Scaling the data

customer_data_scale <- scale(customer_data_cluster)

customer_data_scale

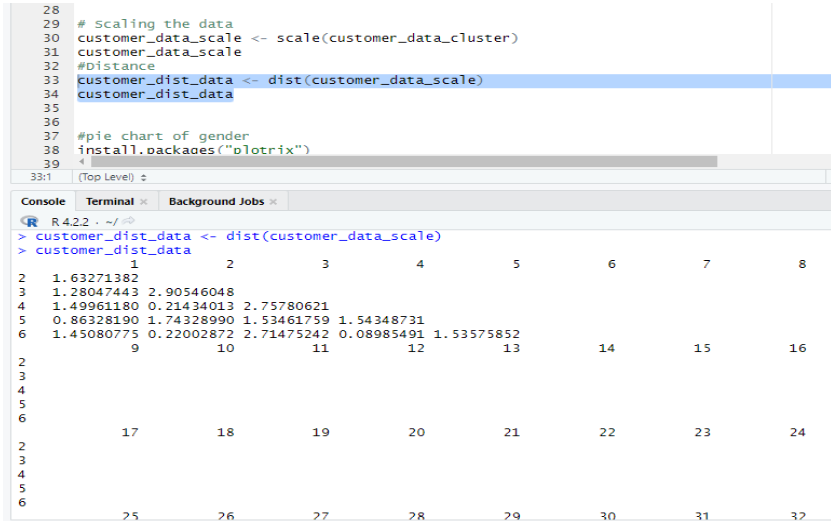




5. Distance Measure

customer_dist_data <- dist(customer_data_scale)

customer_dist_data



6. Pie chart of gender

install.packages("plotrix")

a= table(mydata$Gender)

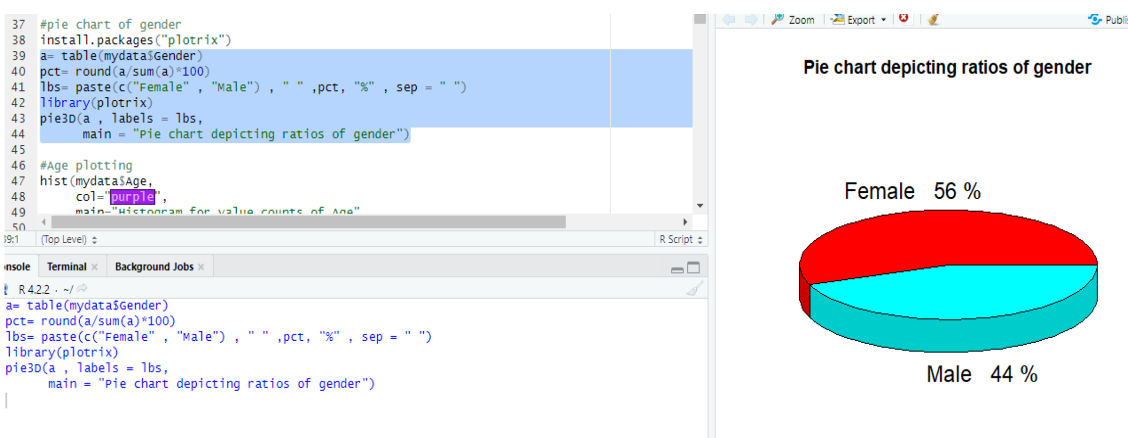
pct= round(a/sum(a)*100)

lbs= paste(c("Female" , "Male") , " " ,pct, "%" , sep = " ")

library(plotrix)

pie3D(a , labels = lbs,

main = "Pie chart depicting ratios of gender")



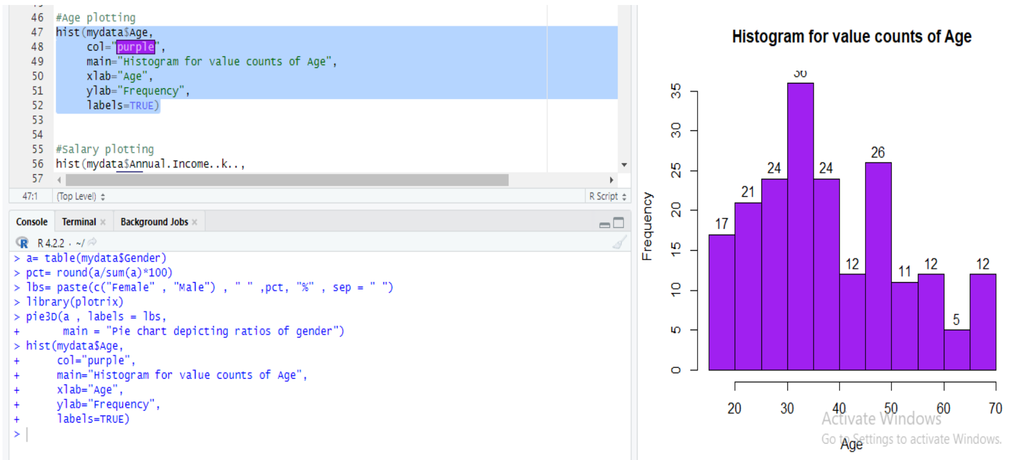
7. Age plotting

hist(mydata$Age,

col="purple",

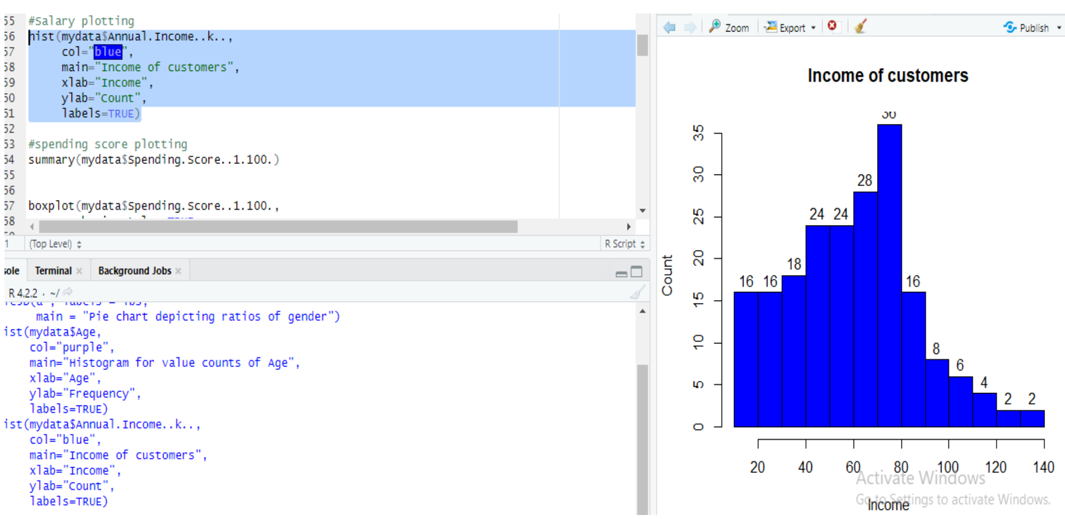
main="Histogram for value counts of Age",

xlab="Age",
ylab="Frequency",
labels=TRUE)



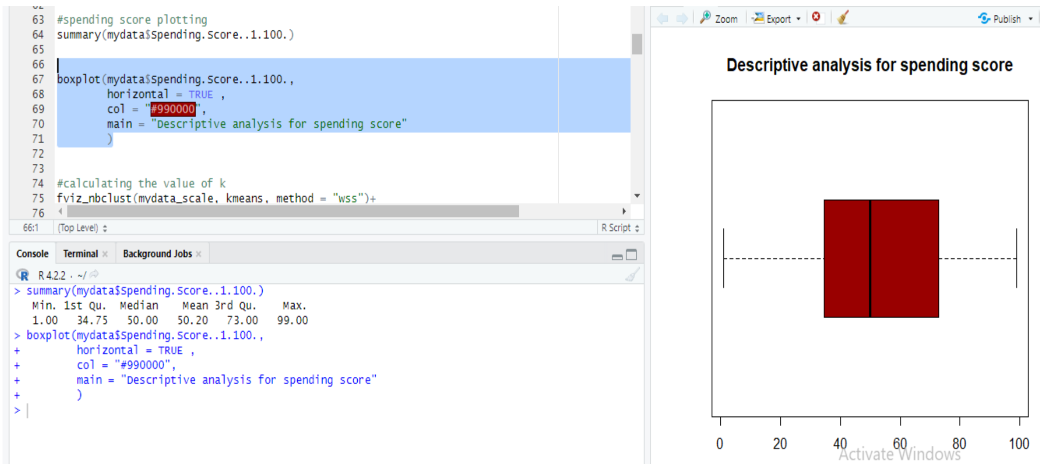
8. Salary plotting

hist(mydata$Annual.Income..k..,
col="blue",
main="Income of customers",
xlab="Income",
ylab="Count",
labels=TRUE)



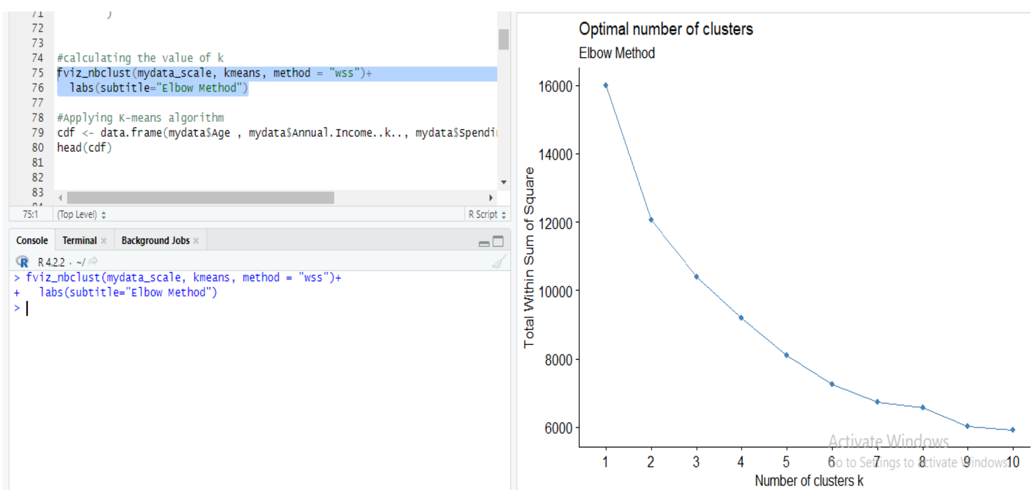
9. Spending score plotting

```
summary(mydata$Spending.Score..1.100.)
boxplot(mydata$Spending.Score..1.100.,
        horizontal = TRUE ,
        col = "#990000",
        main = "Descriptive analysis for spending score"
        )
```



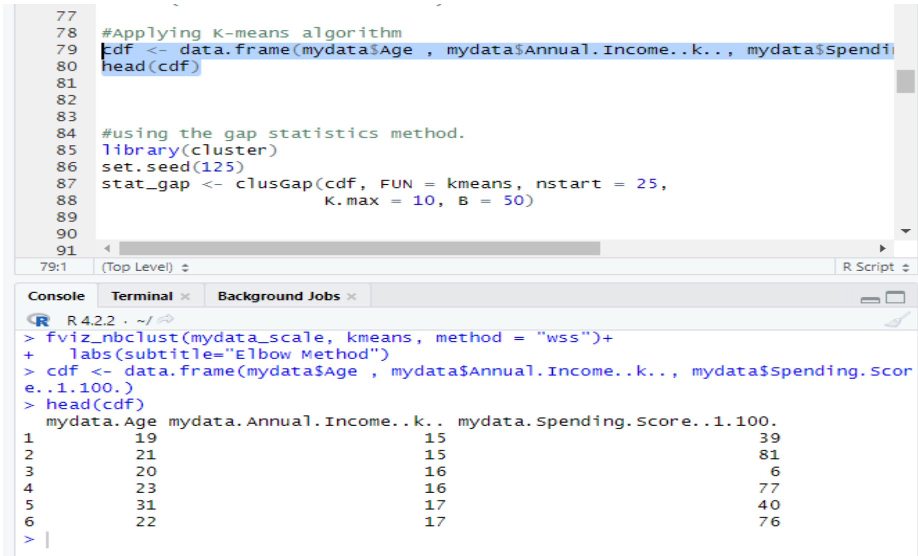
10. Calculating the value of k

```
fviz_nbclust(mydata_scale, kmeans, method = "wss")+
  labs(subtitle="Elbow Method")
```



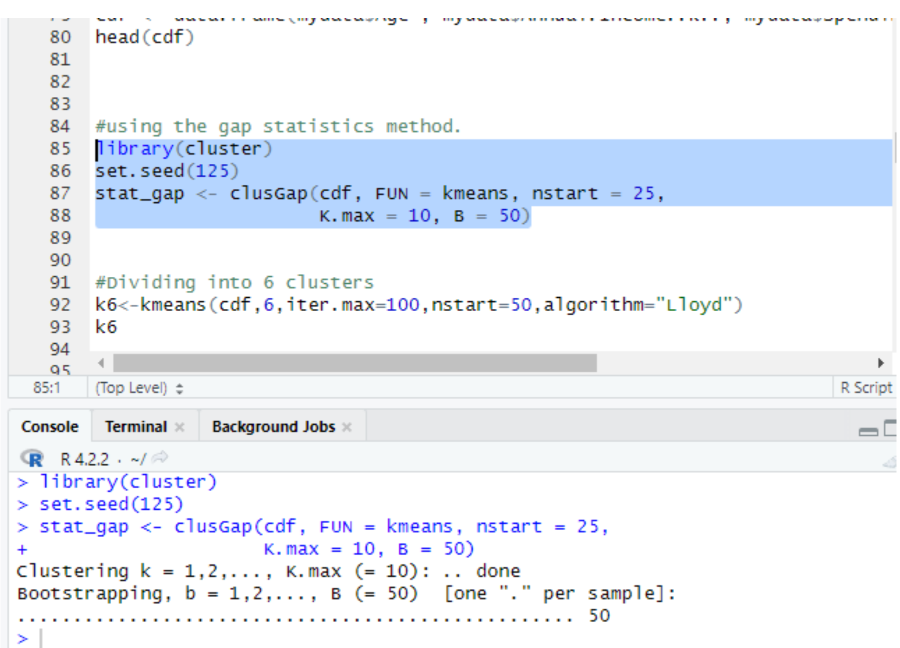
11.Applying K-means algorithm

cdf<-data.frame(mydata$Age,mydata$Annual.Income..k..,
mydata$Spending.Score..1.100.)
head(cdf)



12.Using the gap statistics method.

library(cluster)
set.seed(125)
stat_gap <- clusGap(cdf, FUN = kmeans, nstart = 25,
K.max = 10, B = 50)



13.Dividing into 6 clusters

k6<-kmeans(cdf,6,iter.max=100,nstart=50,algorithm="Lloyd")

k6

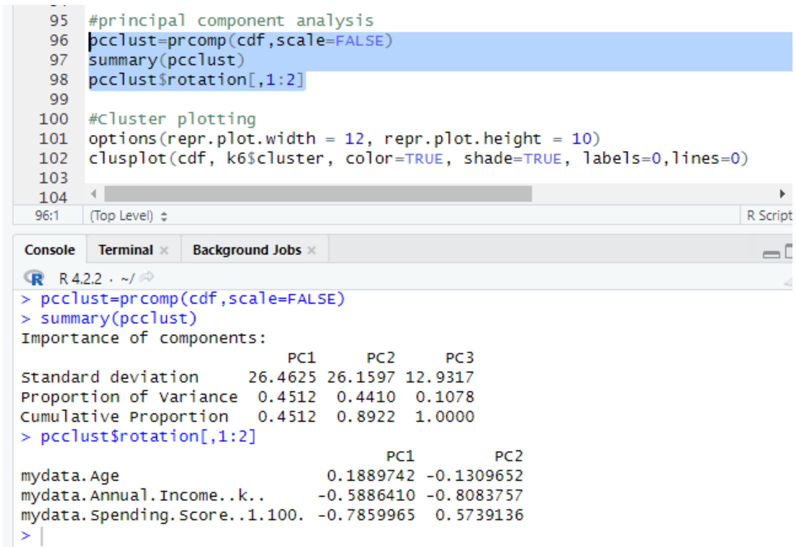
```
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 50)  [one "." per sample]:
.................................................. 50
> k6<-kmeans(cdf,6,iter.max=100,nstart=50,algorithm="Lloyd")
> k6
K-means clustering with 6 clusters of sizes 35, 22, 38, 44, 22, 39

Cluster means:
  mydata.Age mydata.Annual.Income..k.. mydata.Spending.Score..1.100.
1   41.68571                  88.22857                      17.28571
2   44.31818                  25.77273                      20.27273
3   27.00000                  56.65789                      49.13158
4   56.34091                  53.70455                      49.38636
5   25.27273                  25.72727                      79.36364
6   32.69231                  86.53846                      82.12821

Clustering vector:
  [1] 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2
 [36] 5 2 5 2 5 4 5 2 3 2 5 4 3 3 3 4 3 3 4 4 4 4 4 3 4 4 3 4 4 4 3 4 4 3 3
 [71] 4 4 4 4 4 3 4 3 3 4 4 3 4 4 3 4 4 3 3 4 4 3 4 3 3 3 4 3 4 3 3 4 4 3 4
[106] 3 4 4 4 4 4 3 3 3 3 3 4 4 4 4 3 3 3 6 3 6 1 6 1 6 1 6 3 6 1 6 1 6 1 6
[141] 1 6 3 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1
[176] 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6

Within cluster sum of squares by cluster:
[1] 16690.857  8189.000  7742.895  7607.477  4099.818 13972.359
 (between_SS / total_SS =  81.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

14.principal component analysis

pcclust=prcomp(cdf,scale=FALSE)

summary(pcclust)

pcclust$rotation[,1:2]



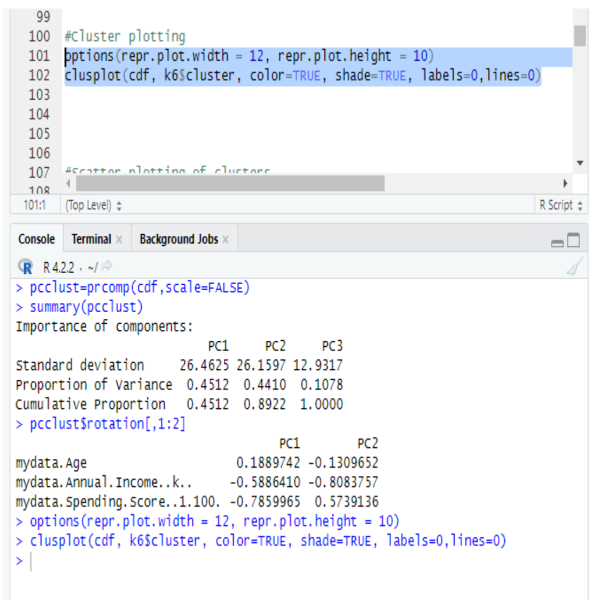
```
#principal component analysis
pcclust=prcomp(cdf,scale=FALSE)
summary(pcclust)
pcclust$rotation[,1:2]

#Cluster plotting
options(repr.plot.width = 12, repr.plot.height = 10)
clusplot(cdf, k6$cluster, color=TRUE, shade=TRUE, labels=0,lines=0)
```

```
> pcclust=prcomp(cdf,scale=FALSE)
> summary(pcclust)
Importance of components:
                           PC1     PC2     PC3
Standard deviation     26.4625 26.1597 12.9317
Proportion of Variance  0.4512  0.4410  0.1078
Cumulative Proportion   0.4512  0.8922  1.0000
> pcclust$rotation[,1:2]
                                  PC1        PC2
mydata.Age                  0.1889742 -0.1309652
mydata.Annual.Income..k..  -0.5886410 -0.8083757
mydata.Spending.Score..1.100. -0.7859965  0.5739136
>
```

15.Cluster plotting

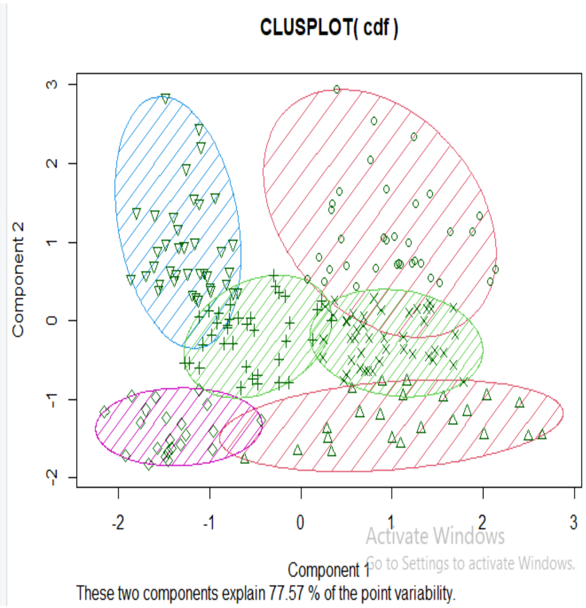
options(repr.plot.width = 12, repr.plot.height = 10)

clusplot(cdf, k6$cluster, color=TRUE, shade=TRUE, labels=0,lines=0)



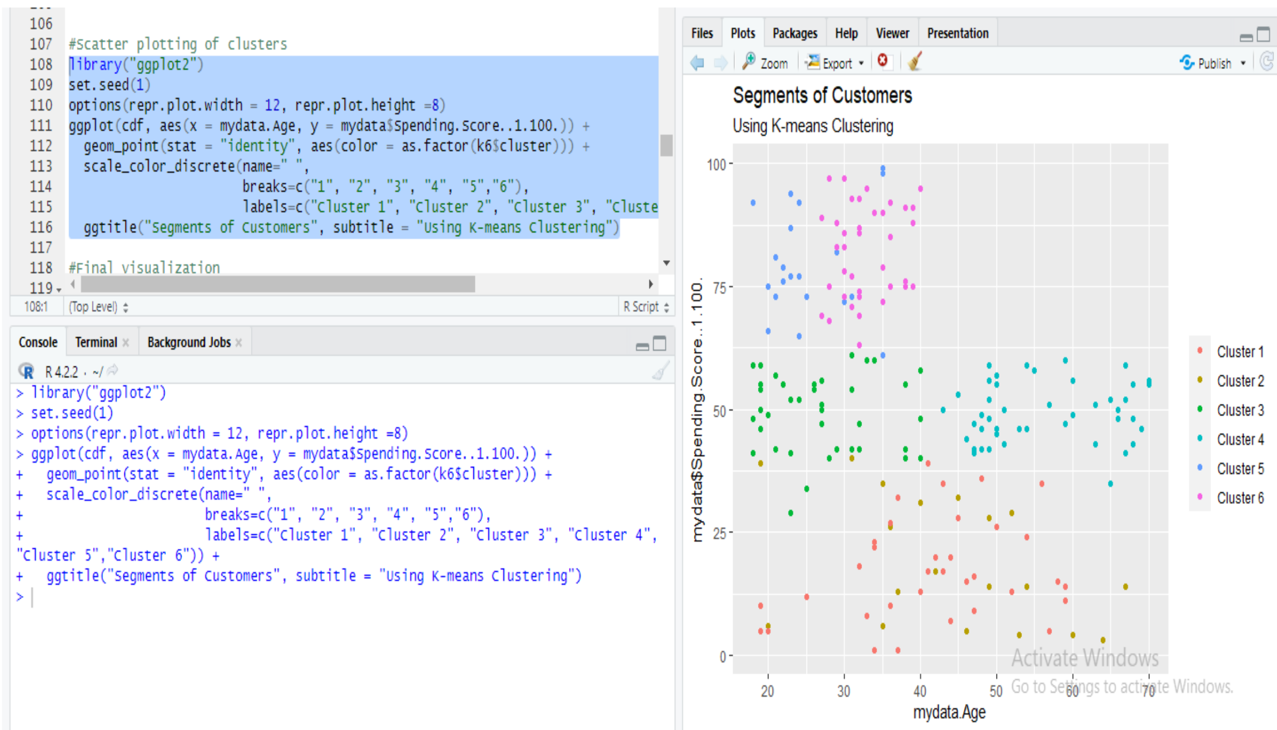
```
#Cluster plotting
options(repr.plot.width = 12, repr.plot.height = 10)
clusplot(cdf, k6$cluster, color=TRUE, shade=TRUE, labels=0,lines=0)
```

```
> pcclust=prcomp(cdf,scale=FALSE)
> summary(pcclust)
Importance of components:
                           PC1     PC2     PC3
Standard deviation     26.4625 26.1597 12.9317
Proportion of Variance  0.4512  0.4410  0.1078
Cumulative Proportion   0.4512  0.8922  1.0000
> pcclust$rotation[,1:2]
                                  PC1        PC2
mydata.Age                  0.1889742 -0.1309652
mydata.Annual.Income..k..  -0.5886410 -0.8083757
mydata.Spending.Score..1.100. -0.7859965  0.5739136
> options(repr.plot.width = 12, repr.plot.height = 10)
> clusplot(cdf, k6$cluster, color=TRUE, shade=TRUE, labels=0,lines=0)
>
```



16.Scatter plotting of clusters

```
library("ggplot2")
set.seed(1)
options(repr.plot.width = 12, repr.plot.height =8)
ggplot(cdf, aes(x = mydata.Age, y = mydata$Spending.Score..1.100.)) +
 geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
 scale_color_discrete(name=" ",
                    breaks=c("1", "2", "3", "4", "5","6"),
                    labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster
6")) +
 ggtitle("Segments of Customers", subtitle = "Using K-means Clustering")
```



17.Final visualization

```
kcols = function(vec){ cols = rainbow(length(unique(vec)))
return(cols[as.numeric(as.factor(vec))])
}
digCluster<- k6$cluster; dignm <- as.character(digCluster); #k-means cluster

plot(pcclust$x[,1:2], col = kcols(digCluster) , pch = 19 , xlab = "K-Means" , ylab = "Classes")

legend("bottomleft" , unique(dignm) , fill = unique(kcols(digCluster)))
```