KUVEMPU UNIVERSITY
Shankargatta,Shivamogga

Academic year 2023-2024

Department of Computer Science

"R-Programming Mini Project"

Project Report On:
"Customer Segmentation Classification"

Submitted By,

USN Number	Name
U06PE21S0093	Ajay B R
U06PE21S0009	Ananth J
U06PE21S0083	Chandan M K
U06PE21S0021	Hemanth Kumar S
U06PE21S0003	Nikhil S
U06PE21S0047	Suhas A N

Submitted To,

Ms Roopa D S
Head of the Department
Department of Computer Science
PESIAMS
Shivamogga

PES Institute of Advanced Management Studies,

NH-206,Sagar Road,Shivamogga

Table of Content

Acknowledgement:

I would like to thank all those who have contributed to the completion of the project and helped me with valuable suggestions for improvement. I am extremely grateful to our Prof. Roopa DS, Professor, Department of Computer Science, for providing us with the atmosphere for the creative work, guidance and encouragement.

Abstract:

The code reads customer data, explores demographics, and visualizes gender distribution. It analyzes age, annual income, and spending score, creating histograms and boxplots.K-means clustering is performed with silhouette analysis and gap statistics for optimal K-Results are visualized using scatter plots and PCA, providing insights into customer segments.

Introduction:

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of k-mean Clustering which is the essential algorithm for clustering unlabelled dataset.

Platform:

R STUDIO R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. One of the keys to R's explosive growth has been its densely populated collection of extension software libraries, known in R terminology as packages, supplied and maintained by R's extensive user community. Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. The best known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published. Text analysis in particular has become well established in R. There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations to advanced text modelling techniques such as fitting Latent Dirichlet Allocation models, R provides it all. One of the main advantages of performing text analysis in R is that it is often possible, and relatively easy, to switch between different packages or to combine them. Recent efforts among the R text analysis developers' community are designed to promote this interoperability to maximize flexibility and choice among users. As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features.

Project Specification:

- R Studio version 1.2.5033

Hardware Specifications:

- Microsoft® Windows® 7/8/10 (32- or 64-bit)
- 3 GB RAM minimum, 8 GB RAM recommended;
- 2 GB of available disk space minimum
- core processor of i3 minimum or above.

Dataset:

- Mall_Customers.csv

Packages Requried:

- plotrix
- purr
- cluster
- gridExtra
- grid
- nbClust
- factoextra
- ggplot2
- dplyr

About Project:

“Customer Data Analysis and Clustering”

1. Data Loading and Exploration: - Loads customer data from a CSV file and examines its structure using `str`, displays column names (`names`), and shows the first few rows (`head`).

2. Descriptive Statistics: - Computes and displays summary statistics and standard deviations for the 'Age', 'Annual Income', and 'Spending Score' columns.

3. Customer Gender Visualization:- Creates a bar plot and a 3D pie chart to visualize the distribution of customer genders.

4. Age Distribution Visualization: - Produces a histogram and boxplot to illustrate the distribution and summary statistics of customer ages.

5. Annual Income Analysis:- Generates a histogram, density plot, and boxplot for the analysis of customer annual income.

6. Spending Score Analysis:- Constructs a boxplot and histogram to analyze the distribution and summary statistics of customer spending scores.

7. K-means Clustering:- Applies the K-means clustering algorithm to cluster customers based on 'Age', 'Annual Income', and 'Spending Score' - Utilizes the elbow method to find an optimal number of clusters and evaluates cluster quality using the silhouette method and gap statistic.

8. Visualizing Clustering Results:- Uses principal component analysis (PCA) to visualize clustering results in a scatter plot, color-coding points by cluster.

Interpretation:

- The analysis suggests that customers can be grouped into clusters based on their age, annual income, and spending score.
- The optimal number of clusters is determined using the elbow method, silhouette method, and gap statistic.
- Visualization of clusters using PCA shows distinct segments based on the first two principal components.

Note:

Ensure the necessary R packages (`plotrix`, `purrr`, `cluster`, `gridExtra`, `NbClust`, `factoextra`, `ggplot2`) are installed.

(The code generates visualizations and statistical summaries to aid in understanding customer behavior and segmentation.)

Description About Packages Used In Project:

1.plotrix Package:

- Used for 3D pie chart visualization (pie3D function).
- This package provides additional plotting functions beyond the base R plotting capabilities.

```
install.packages("plotrix")
library(plotrix)
pie3D(a, labels = lbs, main = "Pie Chart Depicting Ratio of Female and Male")
```

2. purrr Package:

- Used for functional programming and iteration.
- Specifically, it's used to apply the iss function to a sequence of cluster numbers in order to calculate intra-cluster sum of squares.

```
install.packages("purrr")
library(purrr)
iss_values <- map_dbl(k_values, iss)
```

3. cluster Package:

- Used for cluster analysis.
- Provides functions for k-means clustering (kmeans) and silhouette analysis (silhouette, clusGap).

```
install.packages("cluster")
library(cluster)
set.seed(123)
k2 <- kmeans(customer_data[,3:5], 2, iter.max = 100, nstart = 50, algorithm = "Lloyd")
```

4. gridExtra and grid Packages:

- Used for arranging and enhancing grid-based layouts of plots.
- Specifically, gridExtra is used for arranging multiple plots.

```
install.packages("gridExtra")
install.packages("grid")
library(gridExtra)
library(grid)
```

5. NbClust and factoextra Packages:

- Used for determining the optimal number of clusters.
- NbClust provides functions for computing indices of cluster validity.
- factoextra is used for visualizing the results of cluster analysis.

```
install.packages("NbClust")
install.packages("factoextra")
library(NbClust)
library(factoextra)
fviz_nbclust(customer_data[,3:5], kmeans, method = "silhouette")
```

6. ggplot2 Package:

- Used for creating sophisticated and customizable visualizations.
- Utilized for creating scatter plots to visualize clustering results using the first two principal components.

```
library(ggplot2)
ggplot(customer_data, aes(x = Annual.Income..k.., y = Spending.Score..1.100.)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```

Source Code:

```
customer_data=read.csv(choose.files())
str(customer_data)

names(customer_data)
head(customer_data)

summary(customer_data$Age)
sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)
sd(customer_data$Spending.Score..1.100.)


#Customer Gender Visualization

a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))


pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
install.packages("plotrix")
library(plotrix)
pie3D(a,labels=lbs,
      main="Pie Chart Depicting Ratio of Female and Male")


#Visualization of Age Distribution


summary(customer_data$Age)
```

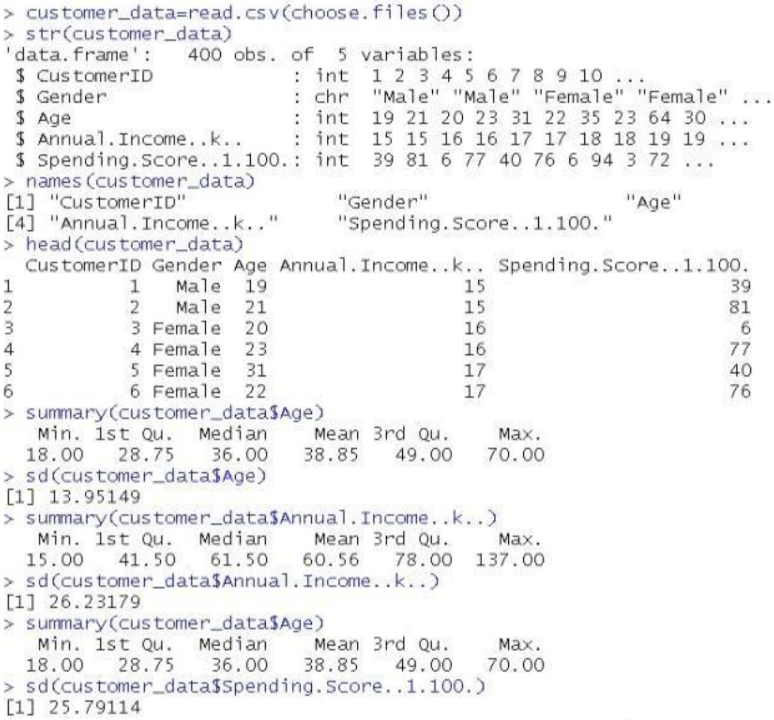
```
hist(customer_data$Age,
    col="blue",
    main="Histogram to Show Count of Age Class",
    xlab="Age Class",
    ylab="Frequency",
    labels=TRUE)

boxplot(customer_data$Age,
        col="#ff0066",
        main="Boxplot for Descriptive Analysis of Age")

#Analysis of the Annual Income of the Customers

summary(customer_data$Annual.Income..k..)
hist(customer_data$Annual.Income..k..,
    col="#660033",
    main="Histogram for Annual Income",
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)

plot(density(customer_data$Annual.Income..k..),
    col="black",
    main="Density Plot for Annual Income",
    xlab="Annual Income Class",
    ylab="Density")
polygon(density(customer_data$Annual.Income..k..),
        col="#ccff66")

boxplot(customer_data$Spending.Score..1.100.,
        horizontal=TRUE,
        col="#990000",
        main="BoxPlot for Descriptive Analysis of Spending Score")

hist(customer_data$Spending.Score..1.100.,
    main="HistoGram for Spending Score",
    xlab="Spending Score Class",
    ylab="Frequency",
    col="#6600cc",
    labels=TRUE)

#K-means Algorithm
install.packages("purrr")
library(purrr)
```

```
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss
}

k.values <- 1:10


iss_values <- map_dbl(k.values, iss)

plot(k.values, iss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total intra-clusters sum of squares")

#Average Silhouette Method
install.packages("cluster")
install.packages("gridExtra")
install.packages("grid")

library(cluster)
library(gridExtra)
library(grid)


k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))

k3<-kmeans(customer_data[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")
s3<-plot(silhouette(k3$cluster,dist(customer_data[,3:5],"euclidean")))

k4<-kmeans(customer_data[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")
s4<-plot(silhouette(k4$cluster,dist(customer_data[,3:5],"euclidean")))

k5<-kmeans(customer_data[,3:5],5,iter.max=100,nstart=50,algorithm="Lloyd")
s5<-plot(silhouette(k5$cluster,dist(customer_data[,3:5],"euclidean")))

k6<-kmeans(customer_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
s6<-plot(silhouette(k6$cluster,dist(customer_data[,3:5],"euclidean")))

k7<-kmeans(customer_data[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")
s7<-plot(silhouette(k7$cluster,dist(customer_data[,3:5],"euclidean")))
```

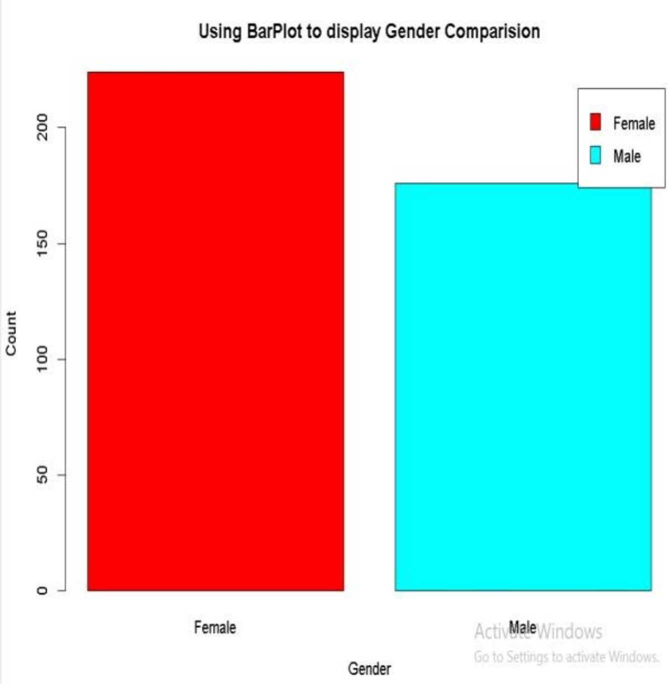
```
k8<-kmeans(customer_data[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")
s8<-plot(silhouette(k8$cluster,dist(customer_data[,3:5],"euclidean")))

k9<-kmeans(customer_data[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")
s9<-plot(silhouette(k9$cluster,dist(customer_data[,3:5],"euclidean")))

k10<-kmeans(customer_data[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")
s10<-plot(silhouette(k10$cluster,dist(customer_data[,3:5],"euclidean")))


install.packages("NbClust")
install.packages("factoextra")
library(NbClust)
library(factoextra)

fviz_nbclust(customer_data[,3:5], kmeans, method = "silhouette")

set.seed(125)
stat_gap <- clusGap(customer_data[,3:5], FUN = kmeans, nstart = 25,
                    K.max = 10, B = 50)
fviz_gap_stat(stat_gap)

k6<-kmeans(customer_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
k6


#Visualizing the Clustering Results using the First Two Principle Components

pcclust=prcomp(customer_data[,3:5],scale=FALSE) #principal component analysis
summary(pcclust)

pcclust$rotation[,1:2]




set.seed(1)
ggplot(customer_data, aes(x =Annual.Income..k.., y = 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 Mall Customers", subtitle = "Using K-means Clustering")
```

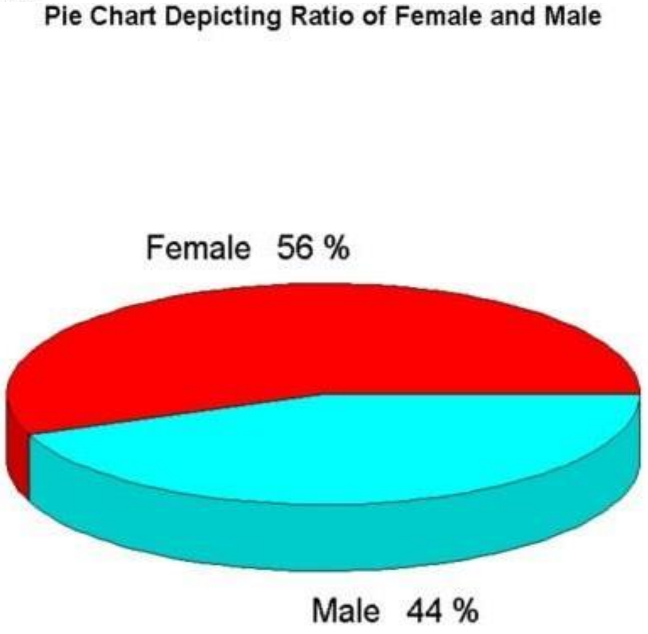
```
ggplot(customer_data, aes(x =Spending.Score..1.100., y =Age)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name=" ",
                      breaks=c("1", "2", "3", "4", "5","6"),
                      labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster 6")) +
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")



kCols=function(vec){cols=rainbow (length (unique (vec)))
return (cols[as.numeric(as.factor(vec))])}

digCluster<-k6$cluster; dignm<-as.character(digCluster); # K-means clusters

plot(pcclust$x[,1:2], col =kCols(digCluster),pch =19,xlab ="K-means",ylab="classes")
legend("bottomleft",unique(dignm),fill=unique(kCols(digCluster)))
```

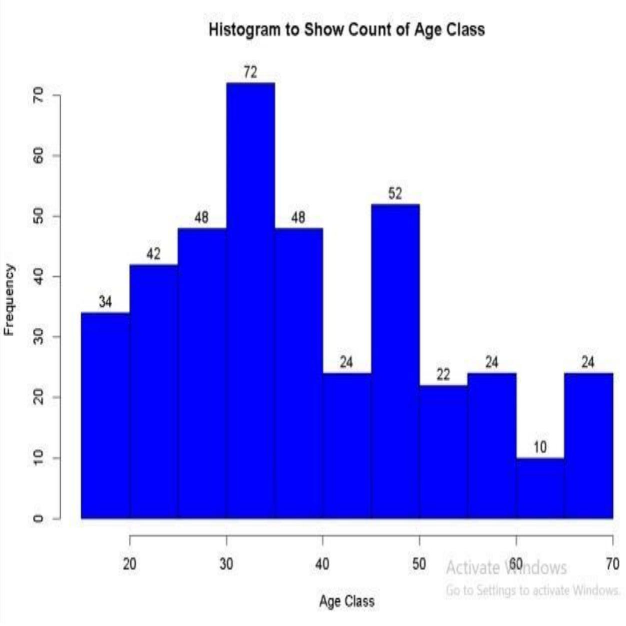
Project SnapShot:

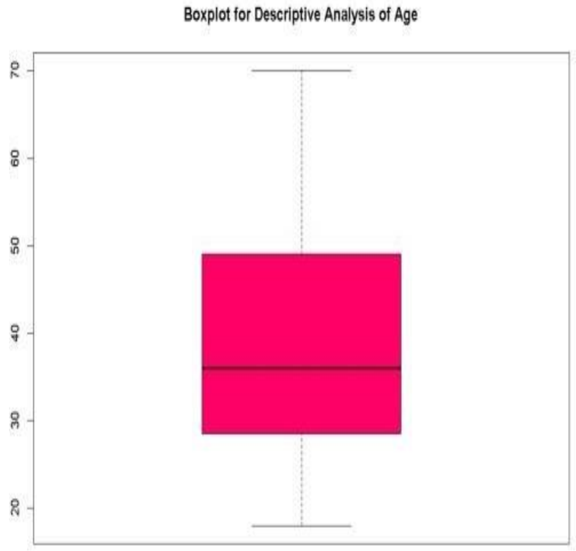


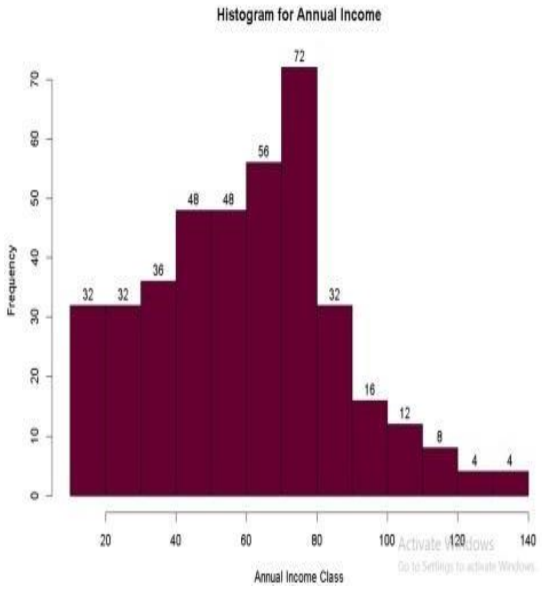
```
> customer_data=read.csv(choose.files())
> str(customer_data)
'data.frame':   400 obs. of  5 variables:
 $ CustomerID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                 : chr  "Male" "Male" "Female" "Female" ...
 $ Age                    : int  19 21 20 23 31 22 35 23 64 30 ...
 $ 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 ...
> names(customer_data)
[1] "CustomerID"             "Gender"                 "Age"
[4] "Annual.Income..k.."     "Spending.Score..1.100."
> head(customer_data)
  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(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   28.75   36.00   38.85   49.00   70.00
> sd(customer_data$Age)
[1] 13.95149
> summary(customer_data$Annual.Income..k..)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   41.50   61.50   60.56   78.00  137.00
> sd(customer_data$Annual.Income..k..)
[1] 26.23179
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   28.75   36.00   38.85   49.00   70.00
> sd(customer_data$Spending.Score..1.100.)
[1] 25.79114
```

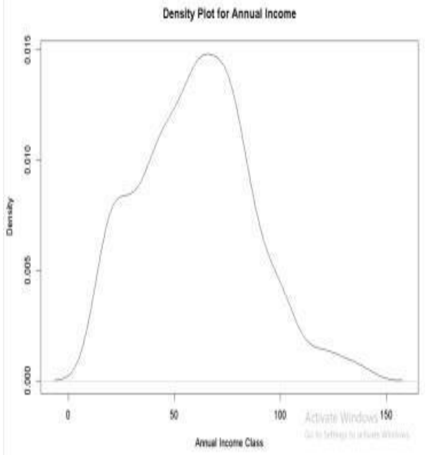


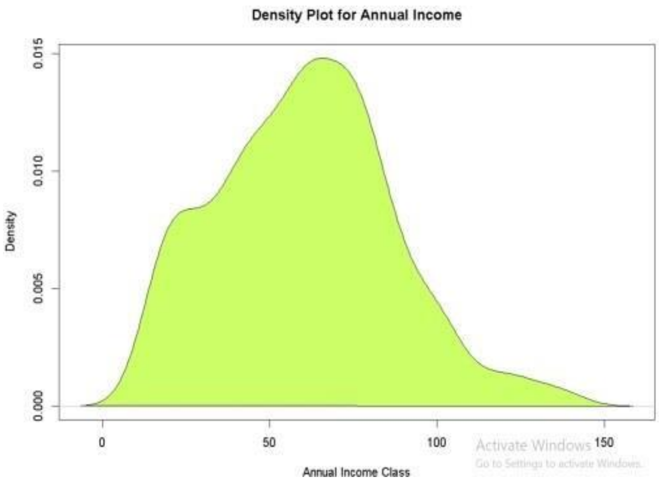


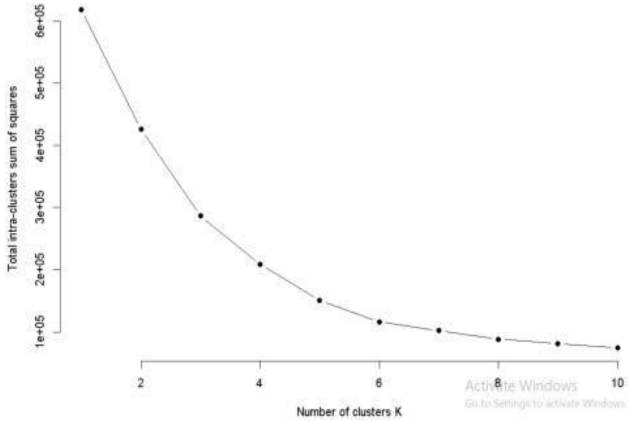


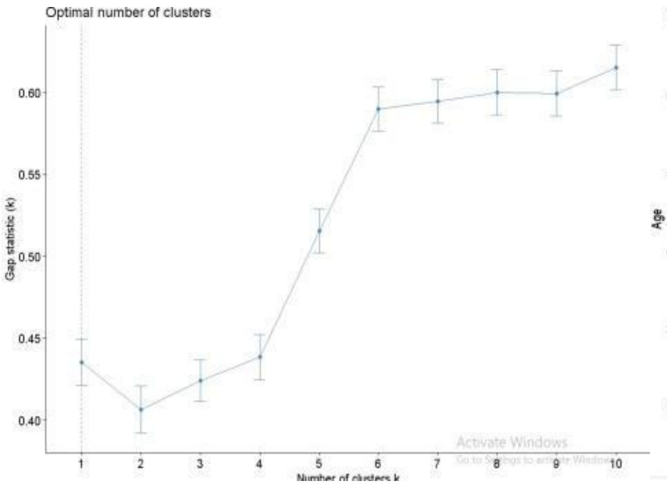

Boxplot for Descriptive Analysis of Age
20
30
40
50
60
70

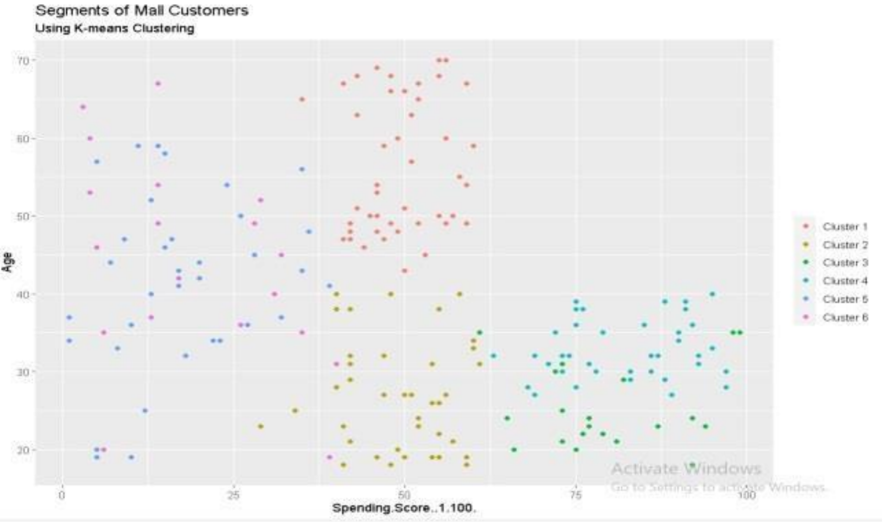

Histogram for Annual Income
Frequency
Annual Income Class
32
32
36
48
48
56
72
32
16
12
8
4
4
20
40
60
80
100
120
140

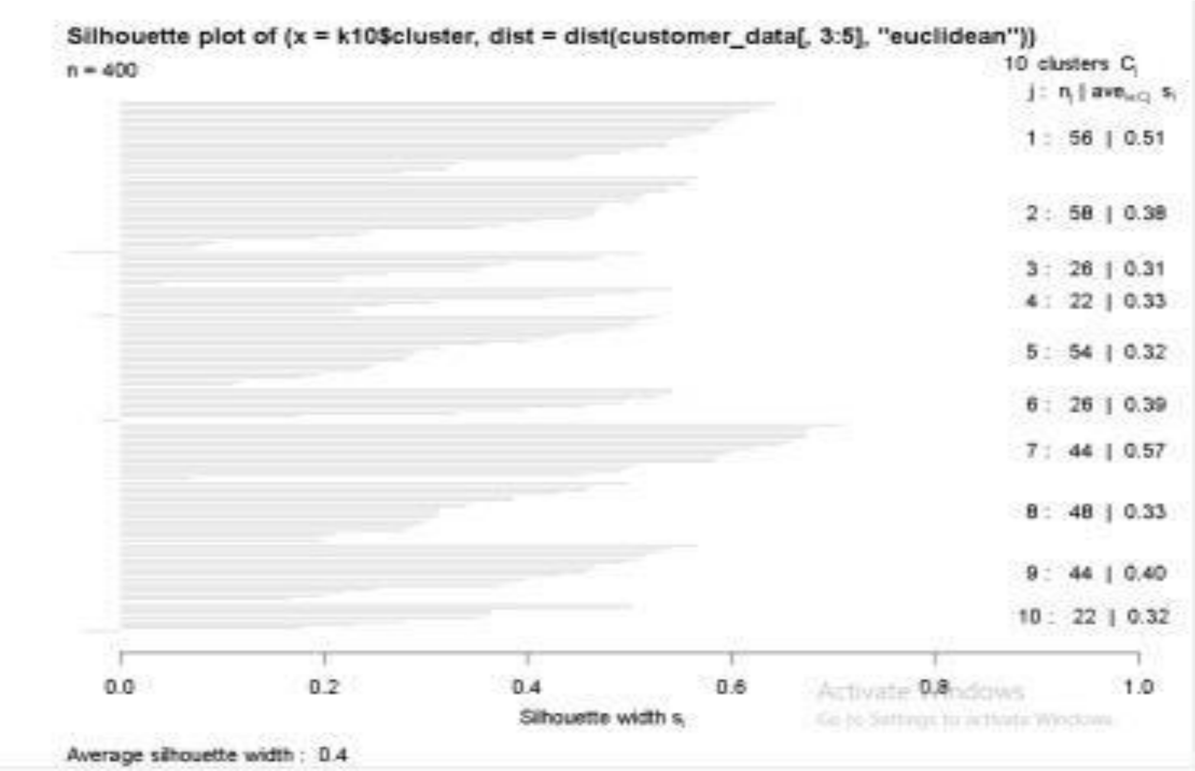

Density Plot for Annual Income
Density
Annual Income Class
0.000
0.005
0.010
0.015
0
50
100
150

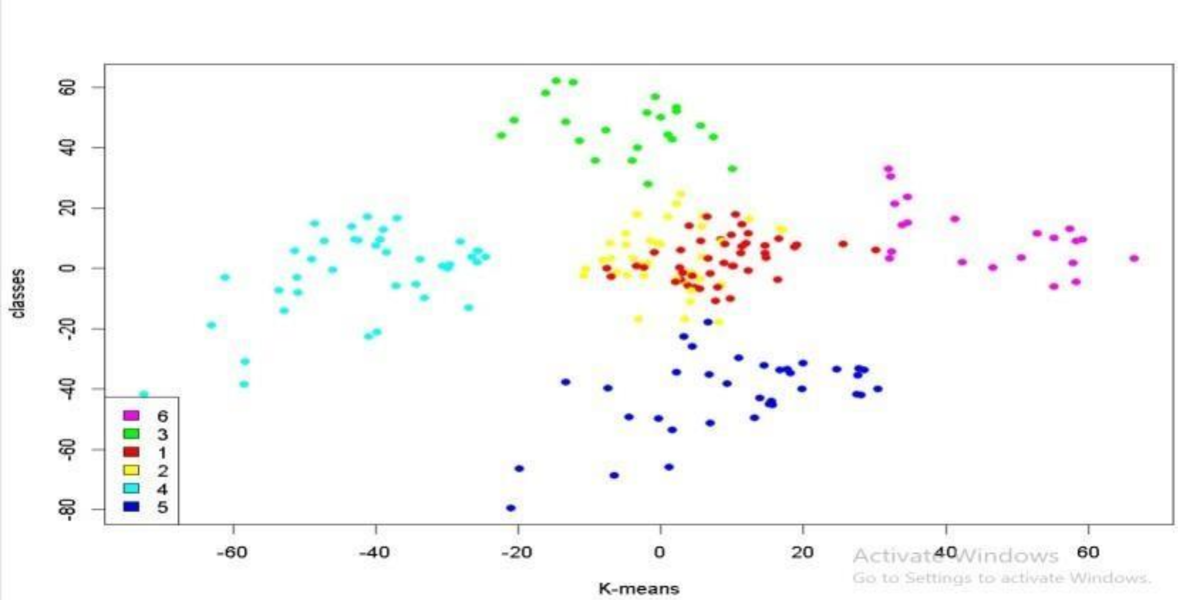

Density Plot for Annual Income
Density
Annual Income Class
0.000
0.005
0.010
0.015
0
50
100
150


Total intra-clusters sum of squares
Number of clusters K
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
2
4
6
8
10


Optimal number of clusters
Gap statistic (k)
Number of clusters k
0.40
0.45
0.50
0.55
0.60
1
2
3
4
5
6
7
8
9
10


Segments of Mall Customers
Using K-means Clustering
Age
Spending.Score..1.100.
20
30
40
50
60
70
0
25
50
75
100
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6


Silhouette plot of (x = k10$cluster, dist = dist(customer_data[, 3:5], "euclidean"))
n = 400
10 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 56 | 0.51
2 : 58 | 0.38
3 : 26 | 0.31
4 : 22 | 0.33
5 : 54 | 0.32
6 : 26 | 0.39
7 : 44 | 0.57
8 : 48 | 0.33
9 : 44 | 0.40
10 : 22 | 0.32
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.4
Activate Windows
Go to Settings to activate Windows.


60
40
20
0
-20
-40
-60
-80
classes
6
3
1
2
4
5
-60
-40
-20
0
20
40
60
K-means
Activate Windows
Go to Settings to activate Windows.

Result And Discussion:

1. Exploratory Data Analysis (EDA)

1.1 Descriptive Statistics

Age Distribution:
The summary statistics and visualizations reveal the age distribution of mall customers.
The histogram and boxplot provide insights into the spread and central tendency of customer ages.

Annual Income:
Summary statistics and visualizations, such as histograms and density plots, are presented to understand the distribution of annual income among customers.

Spending Score:
Descriptive statistics and visualizations, including boxplots and histograms, offer insights into the distribution of spending scores.

1.2 Gender Comparison
A bar plot and 3D pie chart are utilized to visually compare the gender distribution among customers.

2. K-means Clustering

2.1 Determining Optimal Number of Clusters
The total intra-cluster sum of squares is calculated for different values of k using the Elbow Method.
The Average Silhouette Method and Gap Statistic further assist in determining the optimal number of clusters.

2.2 K-means Clustering
K-means clustering is performed using the optimal number of clusters obtained from the analysis.
Silhouette plots for different cluster numbers provide insights into the quality of clustering.

2.3 Visualizing Clustering Results
Principal component analysis (PCA) is applied to visualize the clustering results in a two-dimensional space.
Scatter plots using the first two principal components show the segmentation of mall customers into distinct clusters.

3. Summary and Interpretation

Interpretation of Clusters:
Discuss the characteristics of each cluster in terms of age, annual income, and spending score.
Identify patterns or trends that emerge within each cluster.
Implications for Business:

Discuss potential business implications based on the identified customer segments.
Consider marketing strategies, personalized services, or targeted promotions for each cluster.
Limitations and Future Work:

Address any limitations in the analysis, such as assumptions made or data constraints.
Suggest potential areas for future research or improvements in the clustering approach.
This structure provides a comprehensive overview of the results obtained from both the exploratory data analysis and the K-means clustering. Feel free to customize and expand on each section based on the specific insights and findings from your dataset.

Conclusion:

In this data science project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analysed and visualized the data and then proceeded to implement our algorithm. Hope you enjoyed this customer segmentation project of machine learning using R