515_Assignment4

2022-10-11

1. Being a Data Scientist for the Mall, I am tasked to improve customer experience and attract more customers using machine learning. The model I will be using is K Means Clustering which is an unsupervised learnig method where unlabled data are segmented into groups. The goal is to seperate customers into different groups and individually see what we can sell more or services that we can add or improve on like adding restaurants, restrooms, massage chair stations, etc based on the needs and characteristics of each group like income, age etc.

2,3,5) The dataset that I will be using is from Kaggle called Mall_Customer. The dataset link and code were obtained from https://data-flair.training/blogs/machine-learning-datasets/. The data was very clean and not much data cleaning or preparation was required. The only thing I did was changing the column name to Gender because for some reasons the data came in as 'Genre' for that column. There are also no missing values or NA which can be seen in the data describing code/ouput shown below. For real world data where the data is not this clean, I would go through each column to look for missing values as well as duplicate records to determine what to do with them. For missing values, I do not remove records with missing values right away, it depends if other variables or information are important and it also depends on how many missing values are there. Besides that, data formatting like str, int or float are also important and will be checked and changed for consistency and formatting requirements for certain pacakages. Having the wrong datatype will lead to record not being captured.

```
#Importing data set and saving it as df
url <- "D:/Documents/McDaniel/ANA515/Mall_Customers.csv"
df <- read_csv(url)
```

```
## Rows: 200 Columns: 5
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): CustomerID, Genre
## dbl (3): Age, Annual Income (k$), Spending Score (1-100)
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

```
df<- df %>%
  rename("Gender" = "Genre",
         "Annual_Income_k" = `Annual Income (k$)`,
         "Spending_Score" = "Spending Score (1-100)")
```

4. The mall customer dataframe consists of 5 variabels (Customer ID, Gender, Age, Annual Income and also Spending Score) and 200 rows, there are no NA's and the summary as well as the standard deviations are shown below.

```
names(df)
```

```
## [1] "CustomerID"      "Gender"          "Age"             "Annual_Income_k"
## [5] "Spending_Score"
```

```
ncol(df)
```

```
## [1] 5
```

```
nrow(df)
```

```
## [1] 200
```

```
sapply(df, anyNA)
```

```
##      CustomerID          Gender             Age Annual_Income_k  Spending_Score
##           FALSE           FALSE           FALSE           FALSE           FALSE
```

```
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

```
sd(df$Age)
```

```
## [1] 13.96901
```

```
summary(df$Annual_Income_k)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00
```

```
sd(df$Annual_Income_k)
```

```
## [1] 26.26472
```

```
summary(df$Spending_Score)
```

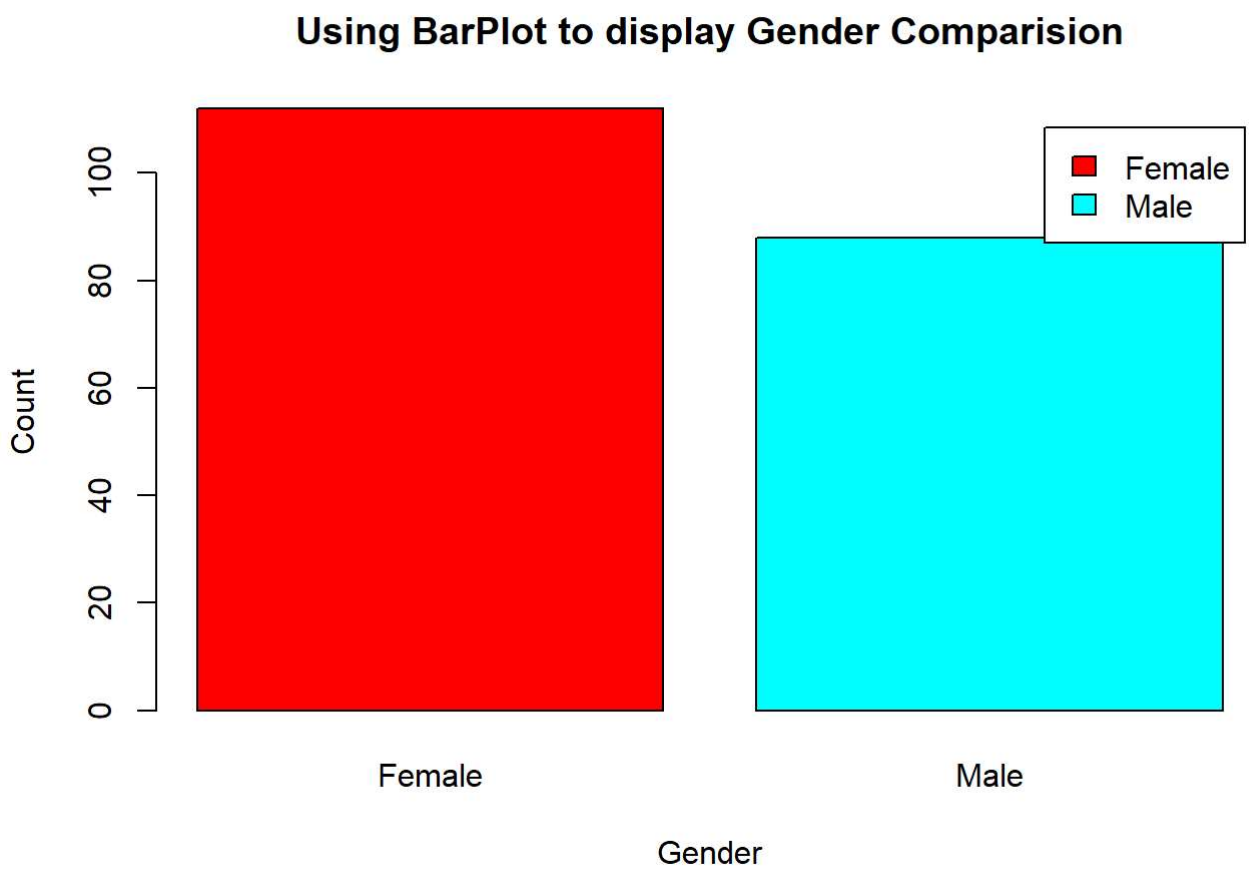
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.75   50.00   50.20   73.00   99.00
```

```
sd(df$Spending_Score)
```

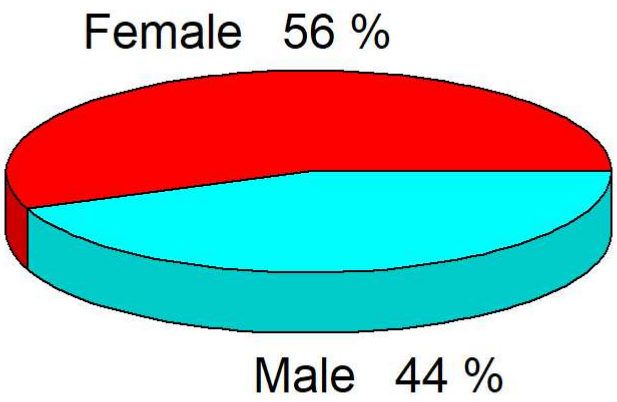
```
## [1] 25.82352
```

8. The data visualizations below are Exploratory Data Analysis that was performed to understand the data more. The first plot below is a bar plot that shows the comparison between male and female count. As we can see from the plot, there are more females than males in this dataset. The percentages are shown in the piechart where females are 56% of the dataset and males are 44%.The third diagram below shows the histogram of Customer's Ages and majority are in their early 30's which can also be seen in the box plot belowe where most customers are in between the age of 30-50. The second histogram describes the annual income and it shows that majority makes 70-80k a year which can also be seen from the density plot below the annual income histogram. A boxplot and histogram were also created for spending score and the charts show that most of the customer's have spending score of 40-60.

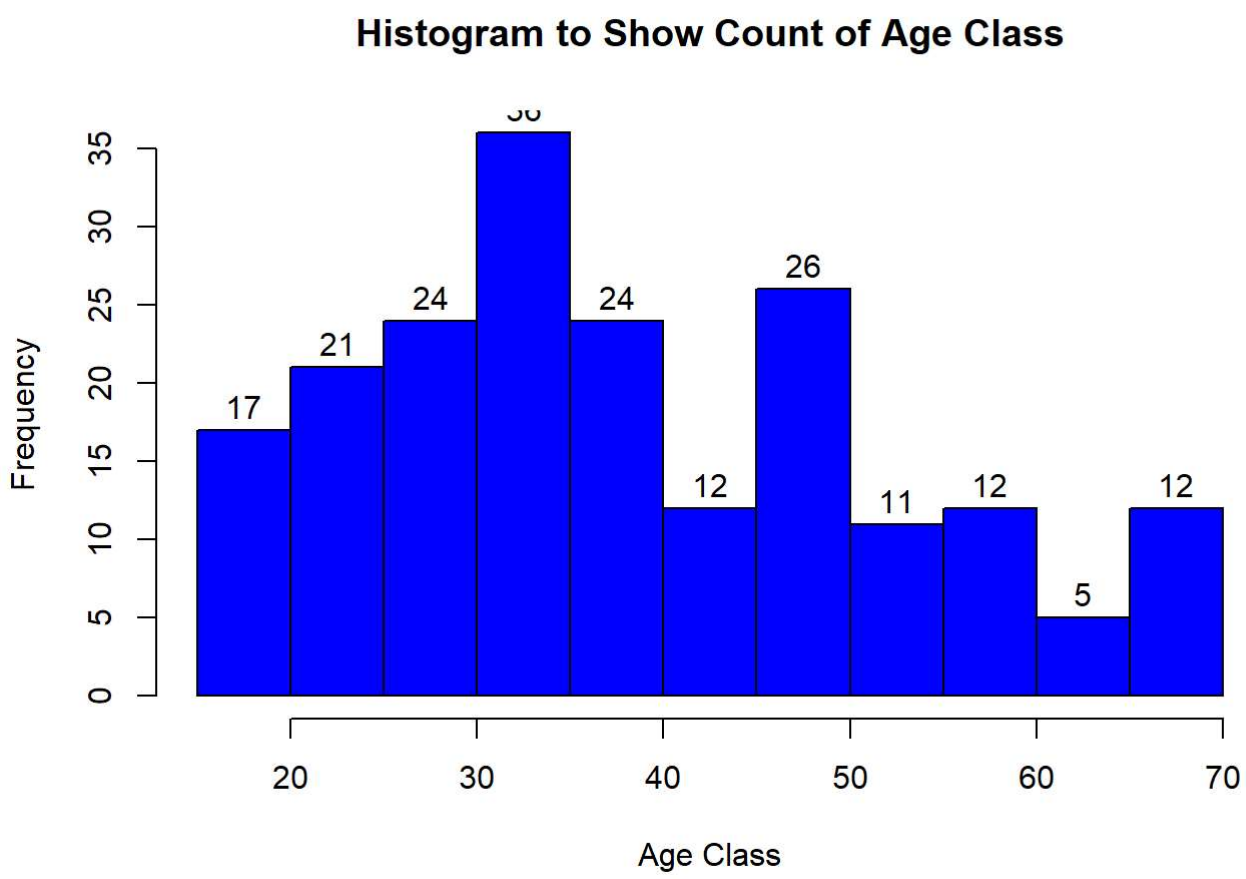
```
a=table(df$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```



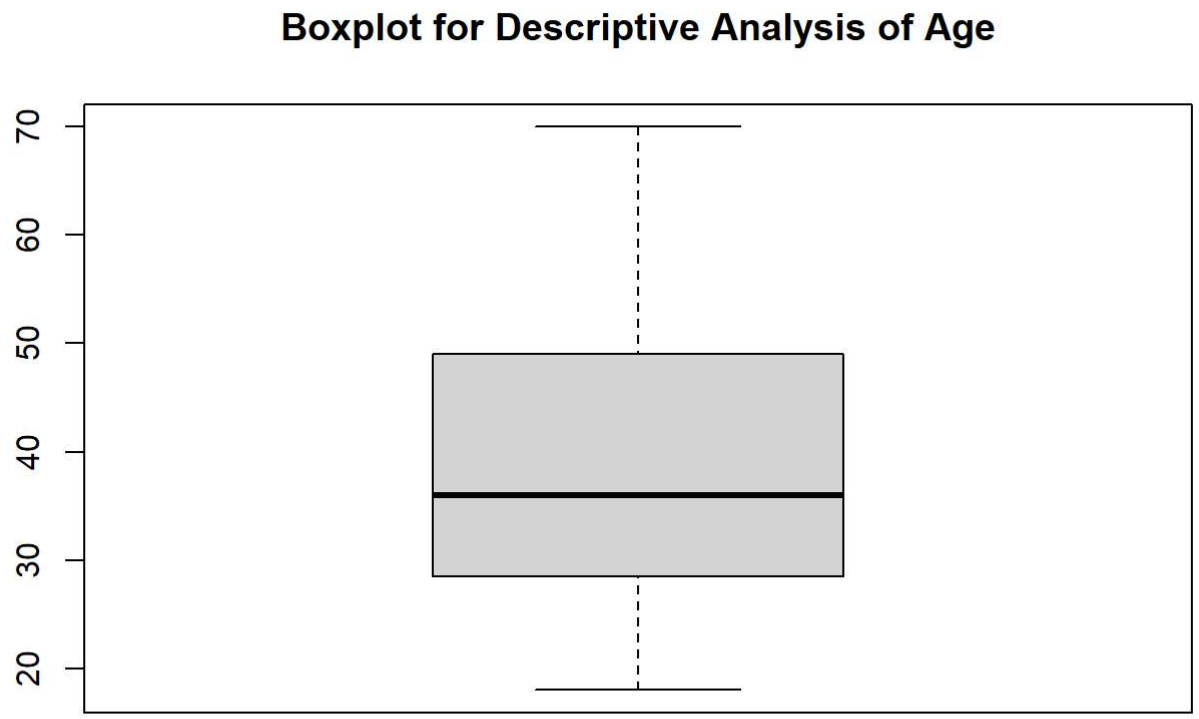

```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
   main="Pie Chart Depicting Ratio of Female and Male")
```




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



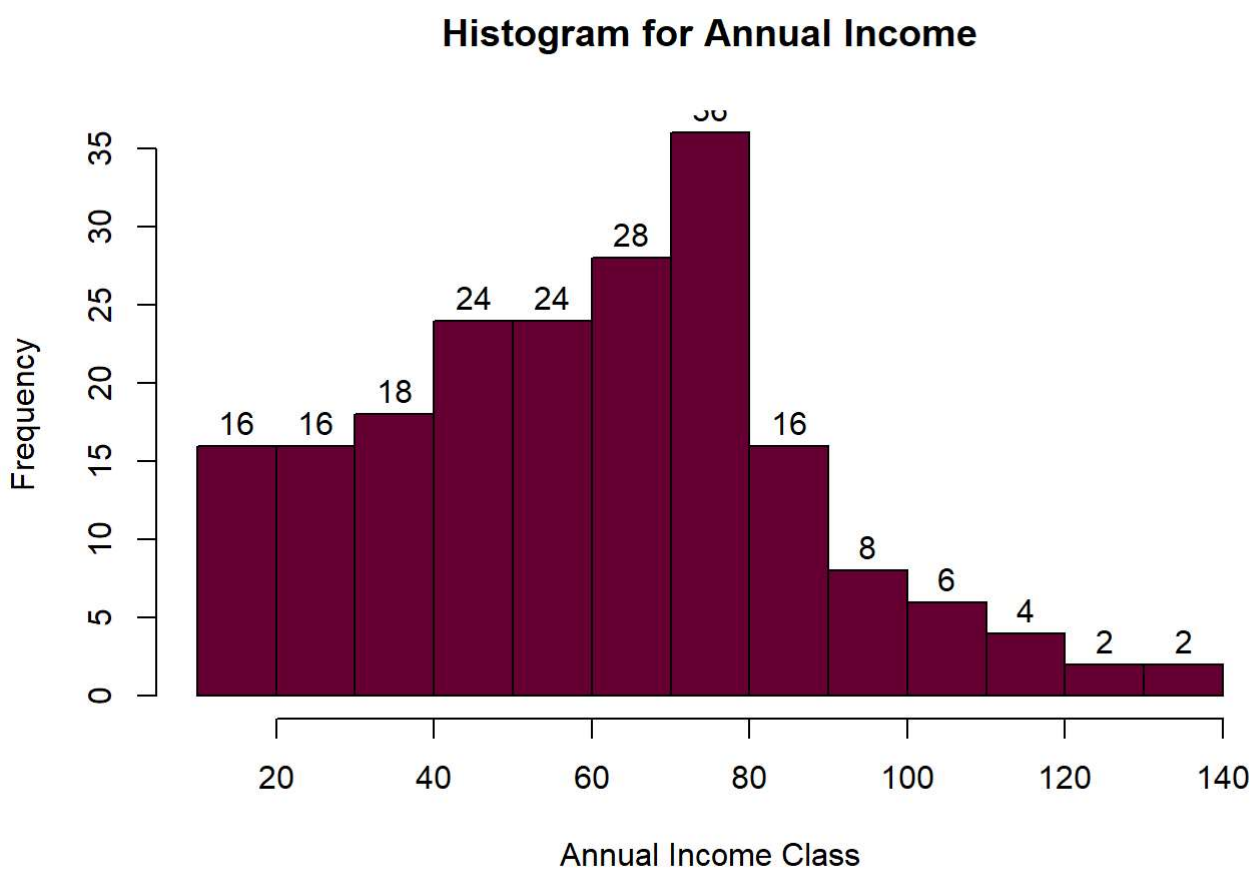

```
boxplot(df$Age,
       main="Boxplot for Descriptive Analysis of Age")
```



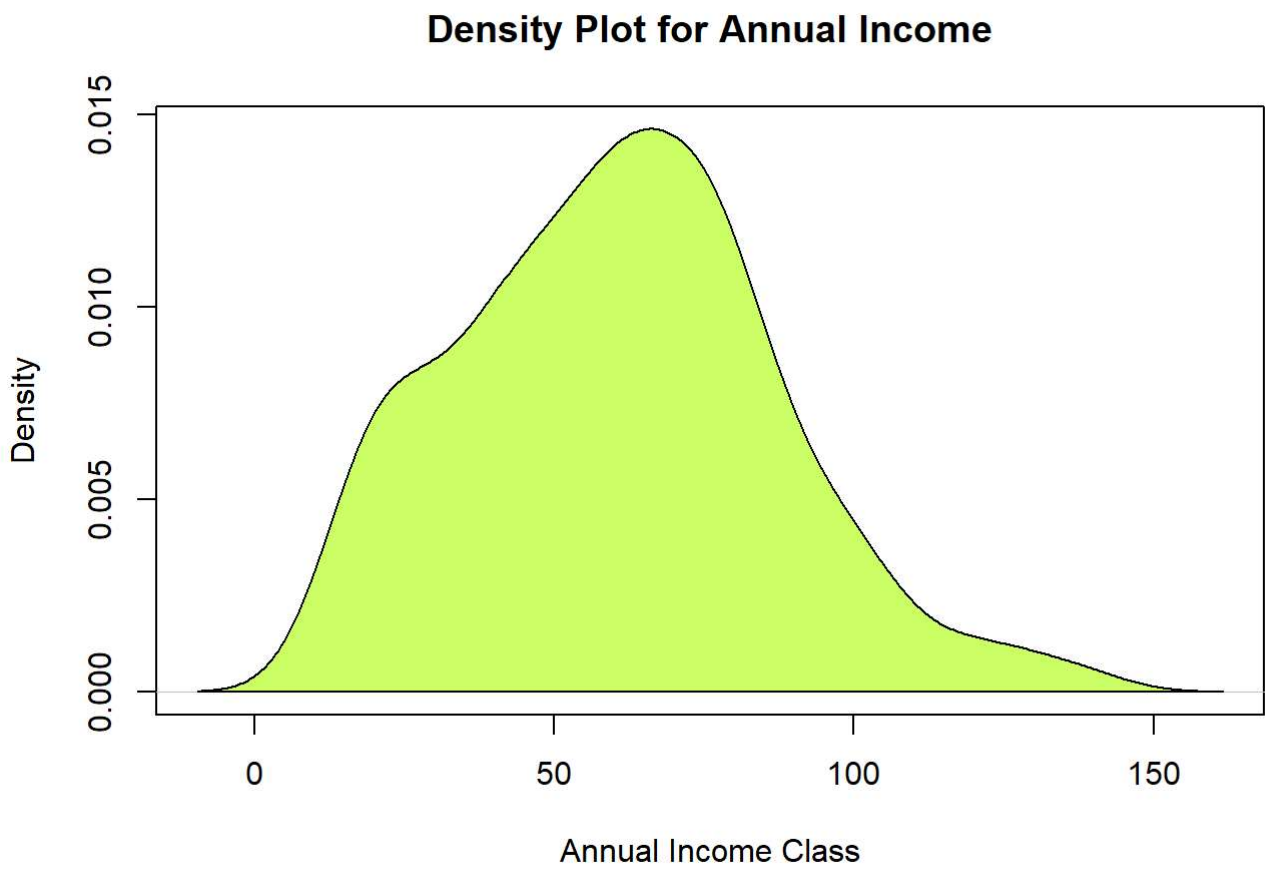

```
summary(df$Annual_Income_k)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00
```

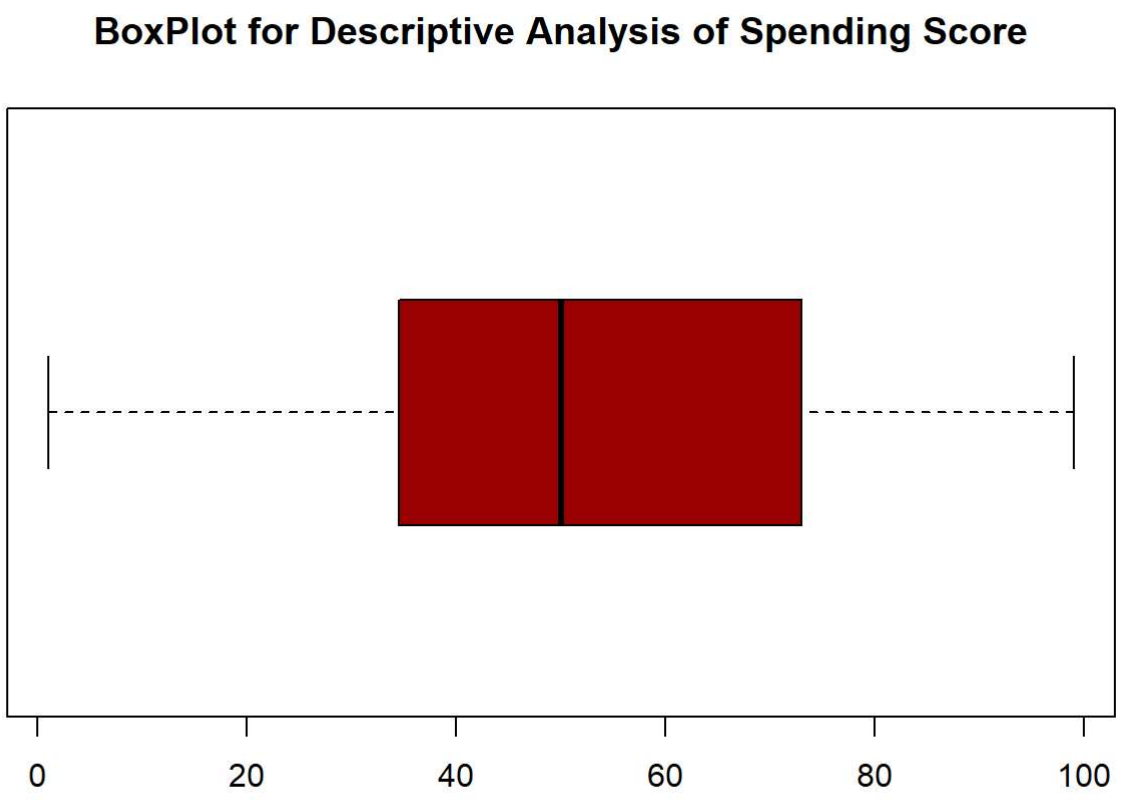
```
hist(df$Annual_Income_k
,
  col="#660033",
  main="Histogram for Annual Income",
  xlab="Annual Income Class",
  ylab="Frequency",
  labels=TRUE)
```



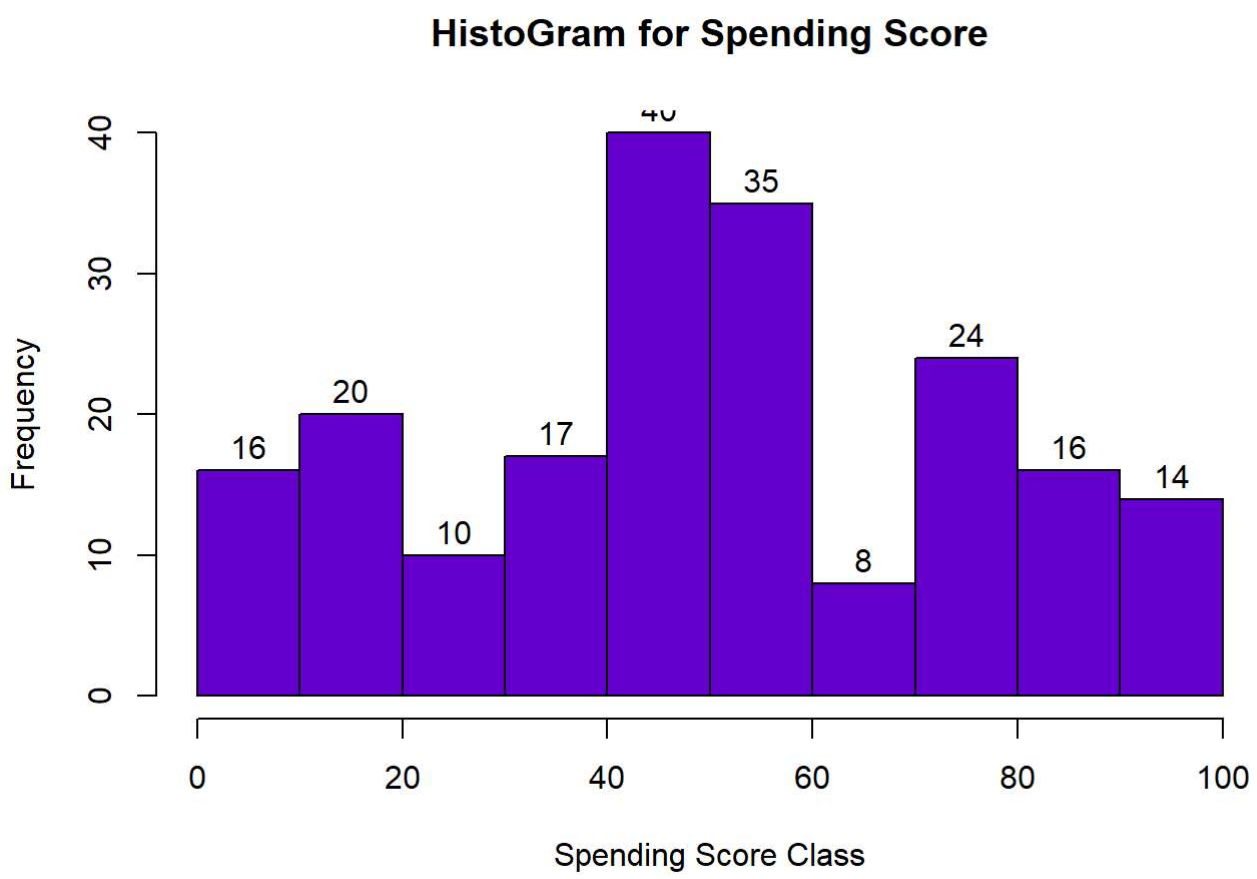

```
plot(density(df$Annual_Income_k),
    col="yellow",
    main="Density Plot for Annual Income",
    xlab="Annual Income Class",
    ylab="Density")
polygon(density(df$Annual_Income_k),
        col="#ccff66")
```




```
boxplot(df$Spending_Score,
   horizontal=TRUE,
   col="#990000",
   main="BoxPlot for Descriptive Analysis of Spending Score")
```



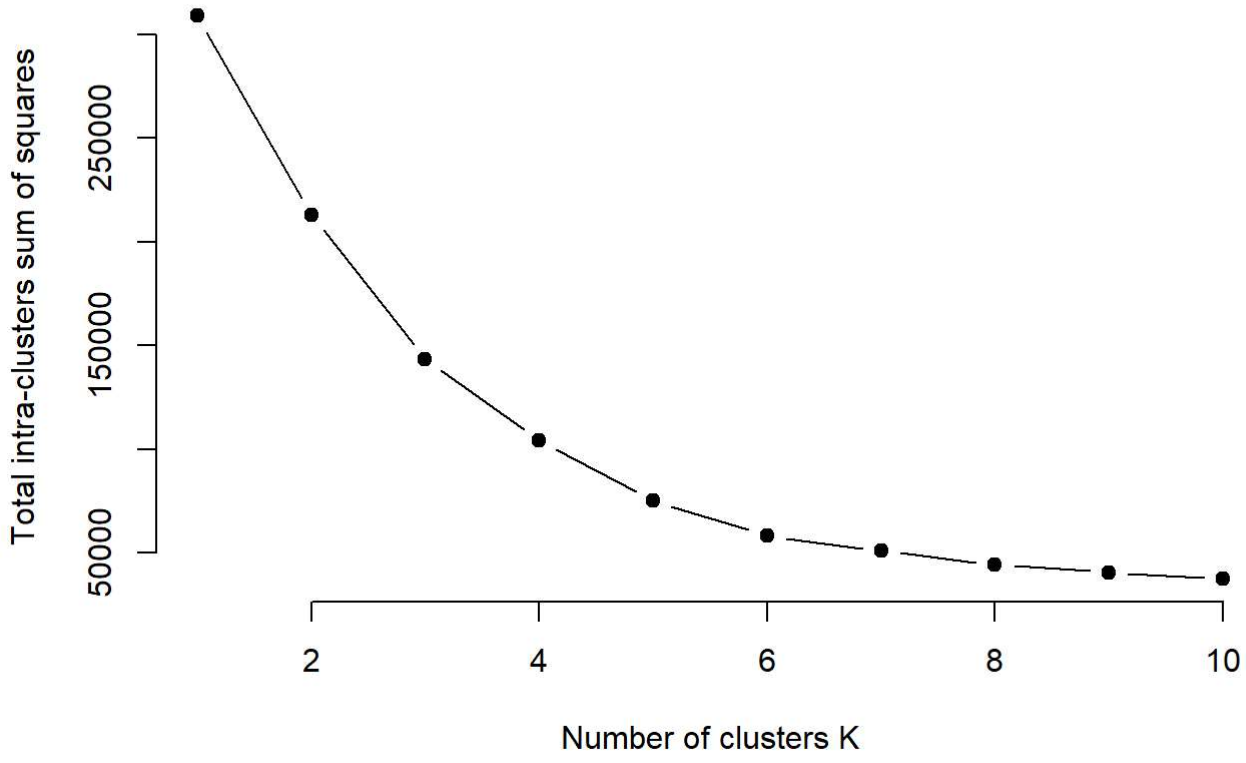

```
hist(df$Spending_Score,
    main="HistoGram for Spending Score",
    xlab="Spending Score Class",
    ylab="Frequency",
    col="#6600cc",
    labels=TRUE)
```



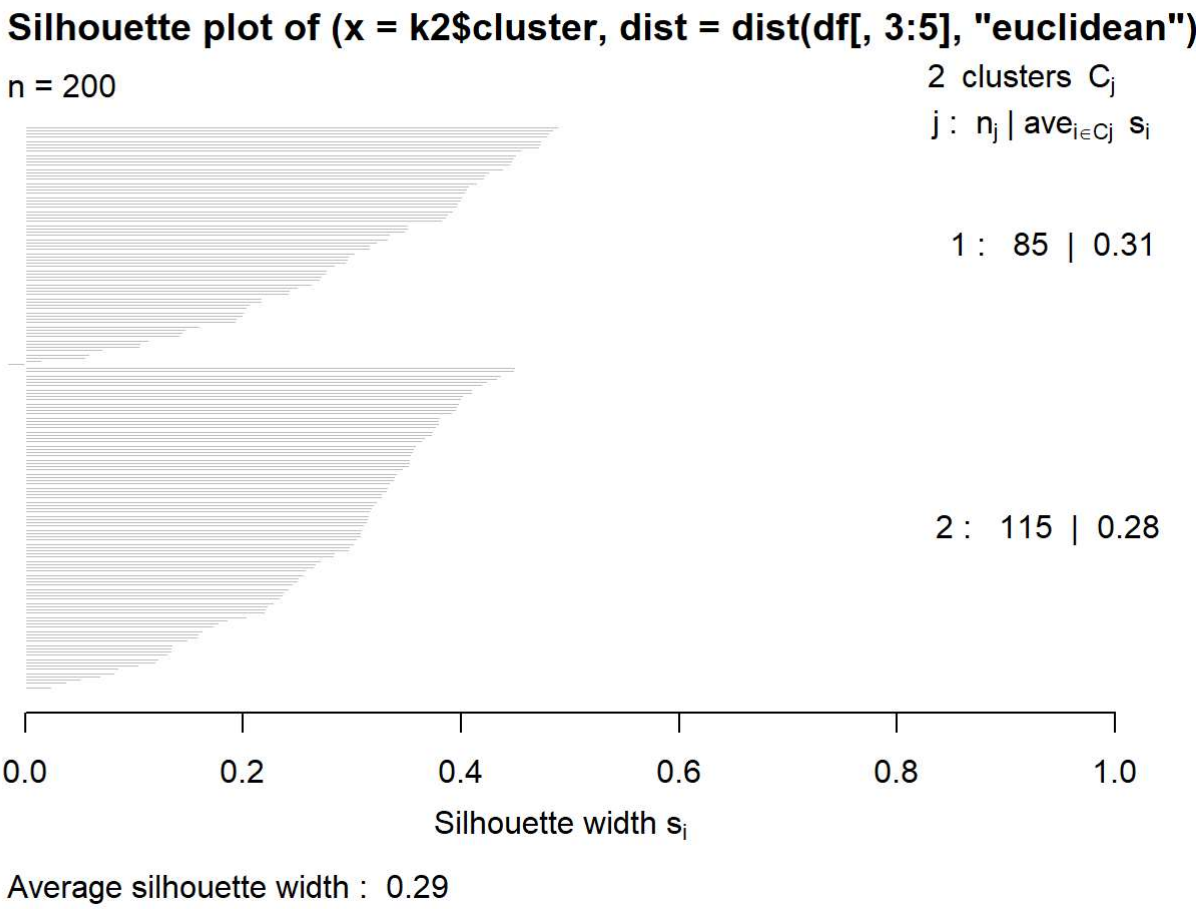
6,8) As mentioned above, the model used is K-Means Clusttering and the first step is to determine how many clusters we would be producing. Number of clusters will be known as 'k'. To determine the optimal number of clusters, there are 3 differnt methods that were used which were the Elbow method, Silhoutte Method and Gap Statistic. The first plot below is the elbow method, the optimal 'k' is where the line bends which is also called the knee and in this case its 4. The following 9 plots are called the Silhoutte plots which is the second method. The wider the average width the better it is, we plotted the charts for k=2 to k=10 and the widest width we got was 0.45 which is k=6, 6 clusters. Lastly, the gap statistic plot shows us that the optimal number of clusters is 6. Since 2 methods suggested 6 clusters, we would go with k=6.

```
library(purrr)
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(df[,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,
```

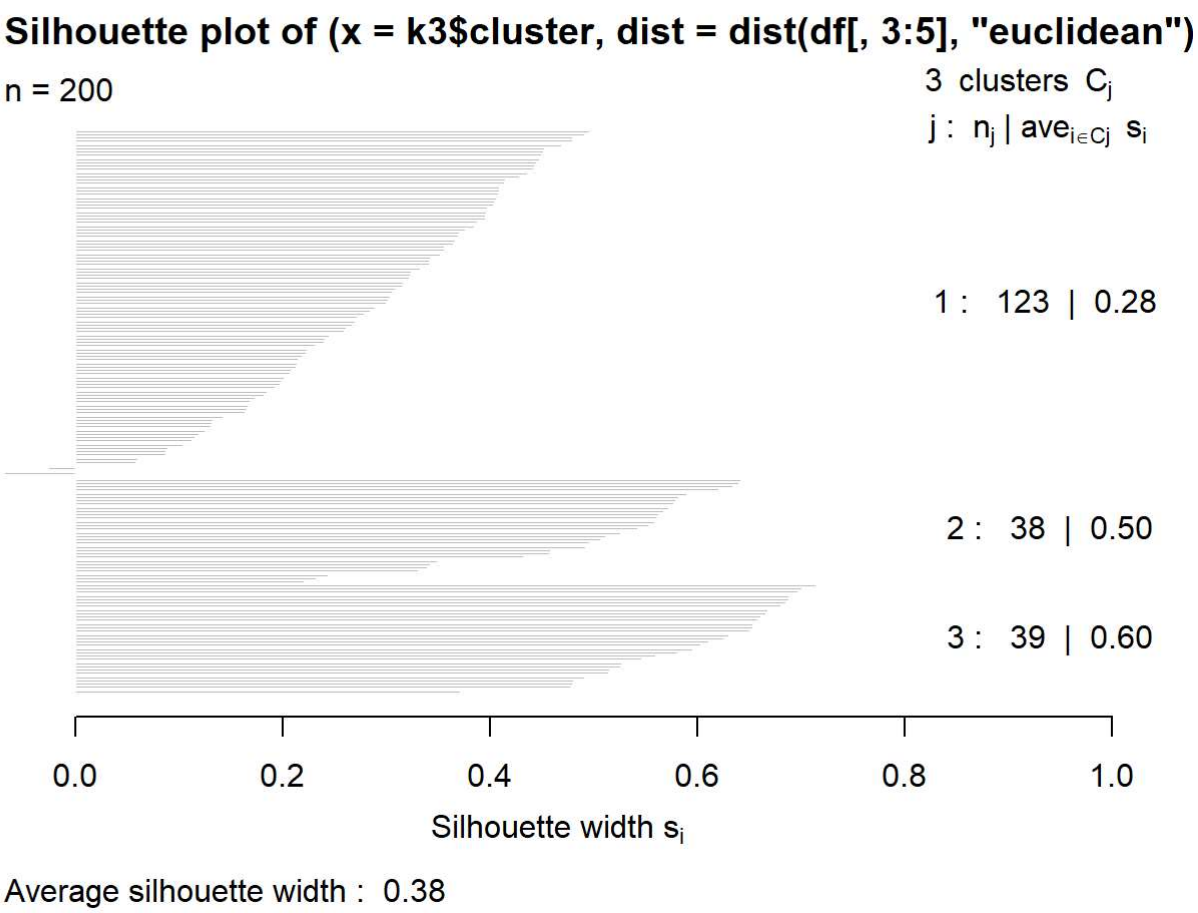
```
    xlab="Number of clusters K",
    ylab="Total intra-clusters sum of squares")
```



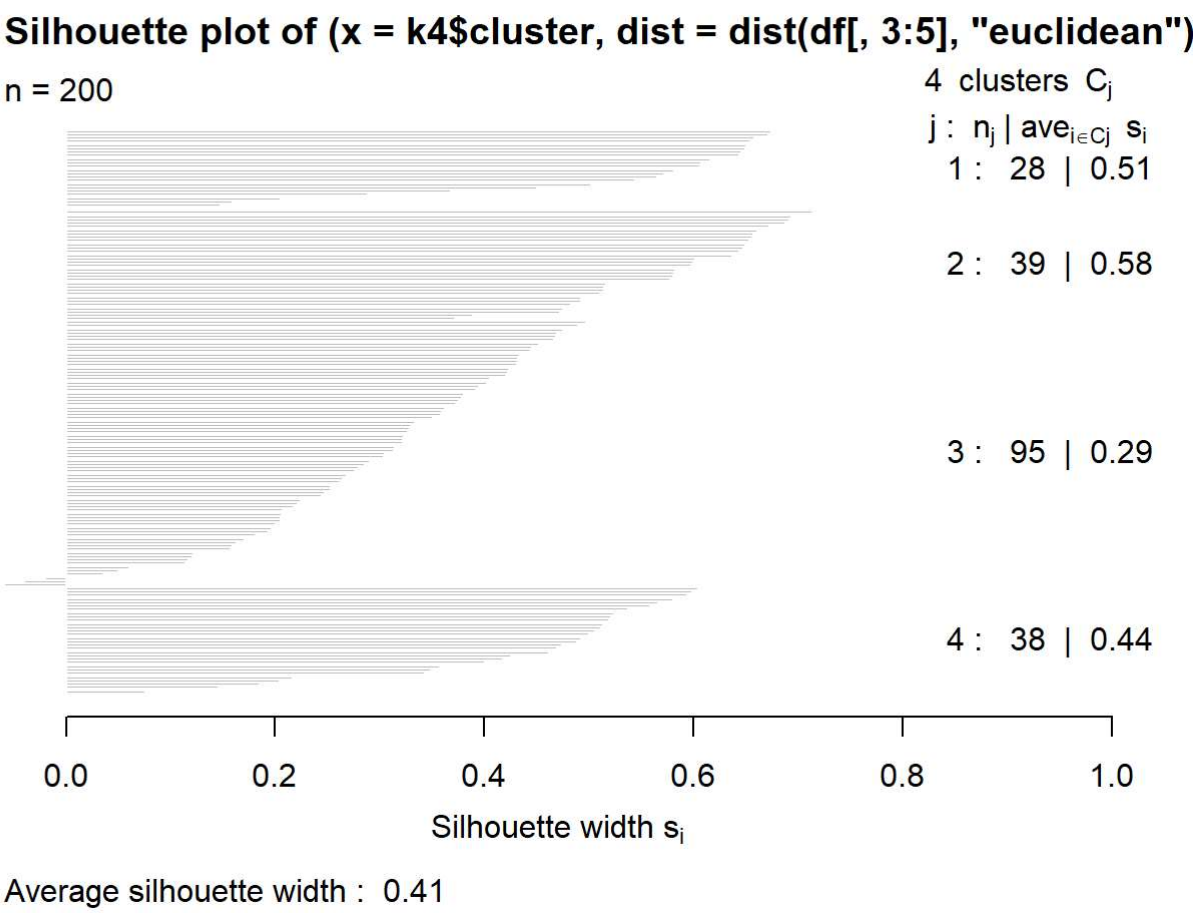

```
k2<-kmeans(df[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(df[,3:5],"euclidean")))
```



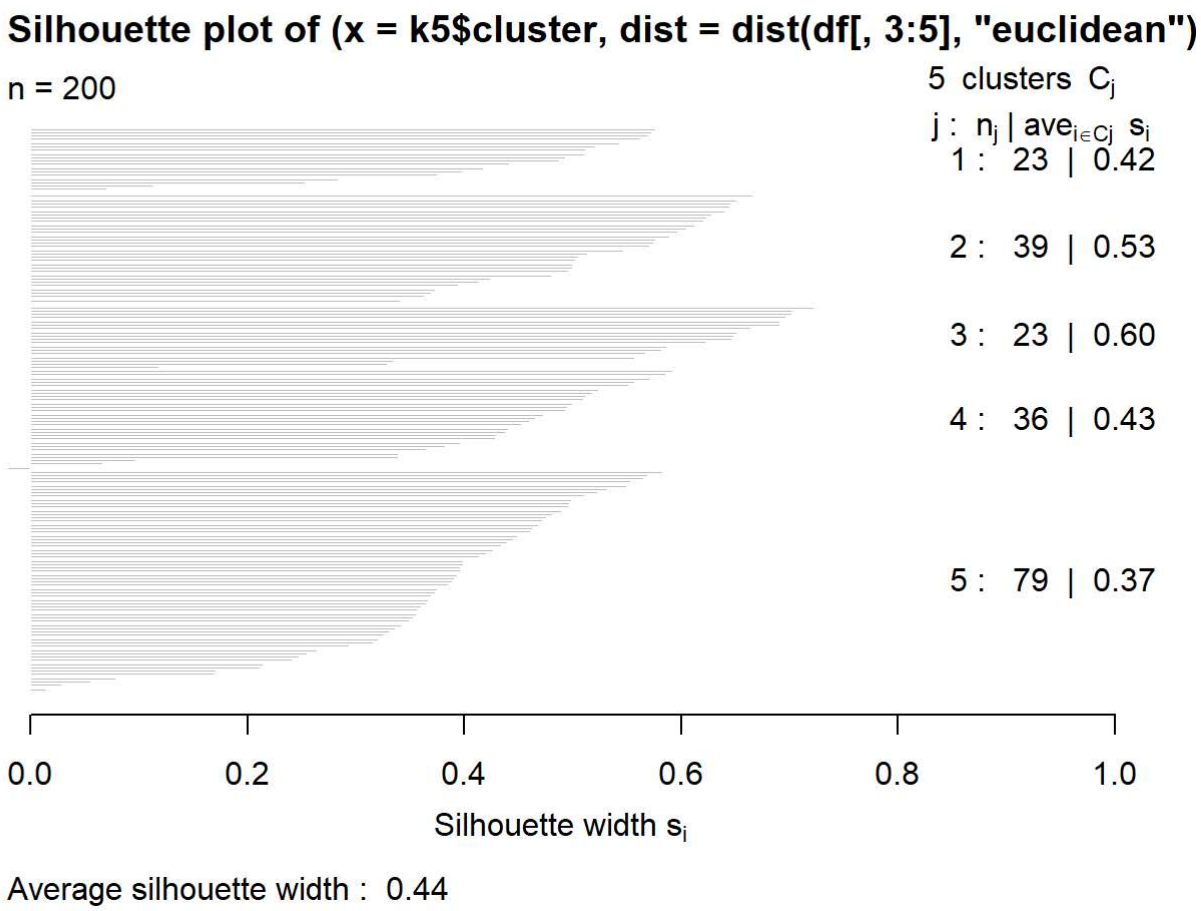

```
k3<-kmeans(df[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")
s3<-plot(silhouette(k3$cluster,dist(df[,3:5],"euclidean")))
```



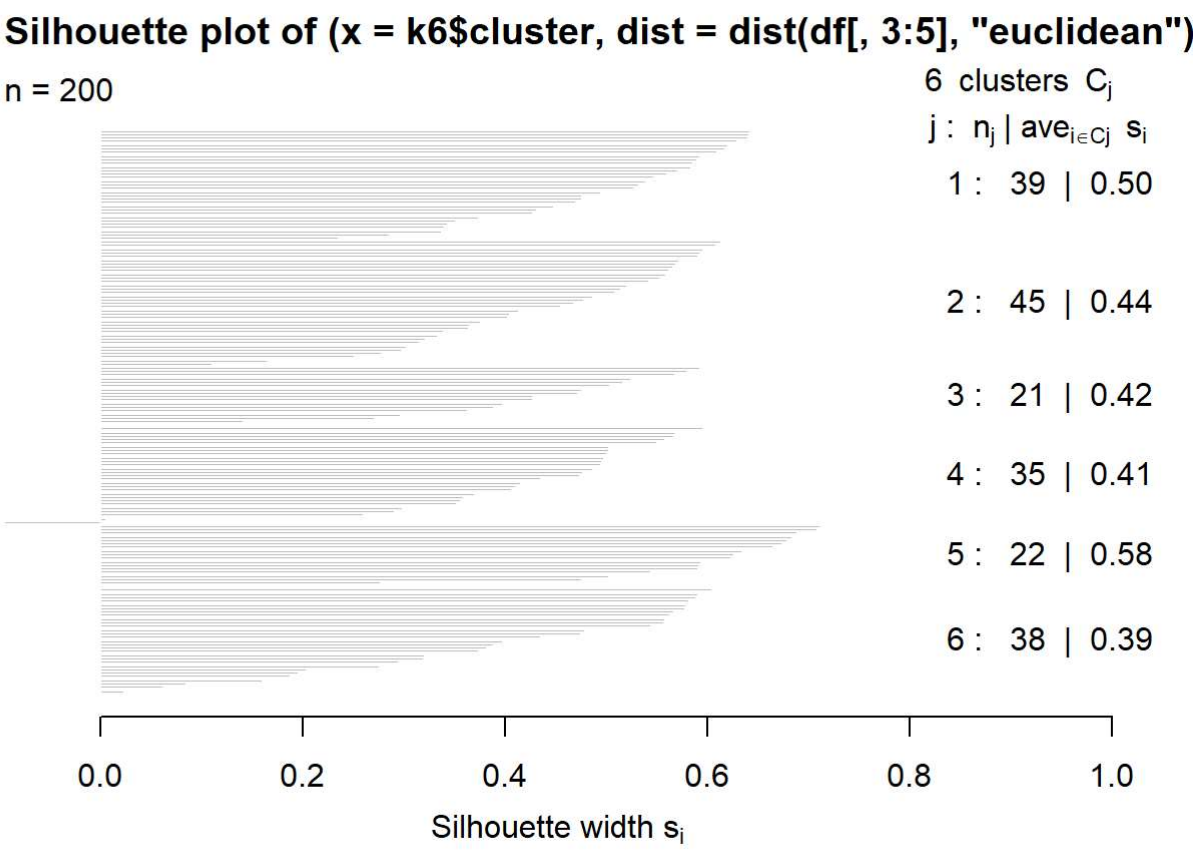

```
k4<-kmeans(df[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")
s4<-plot(silhouette(k4$cluster,dist(df[,3:5],"euclidean")))
```



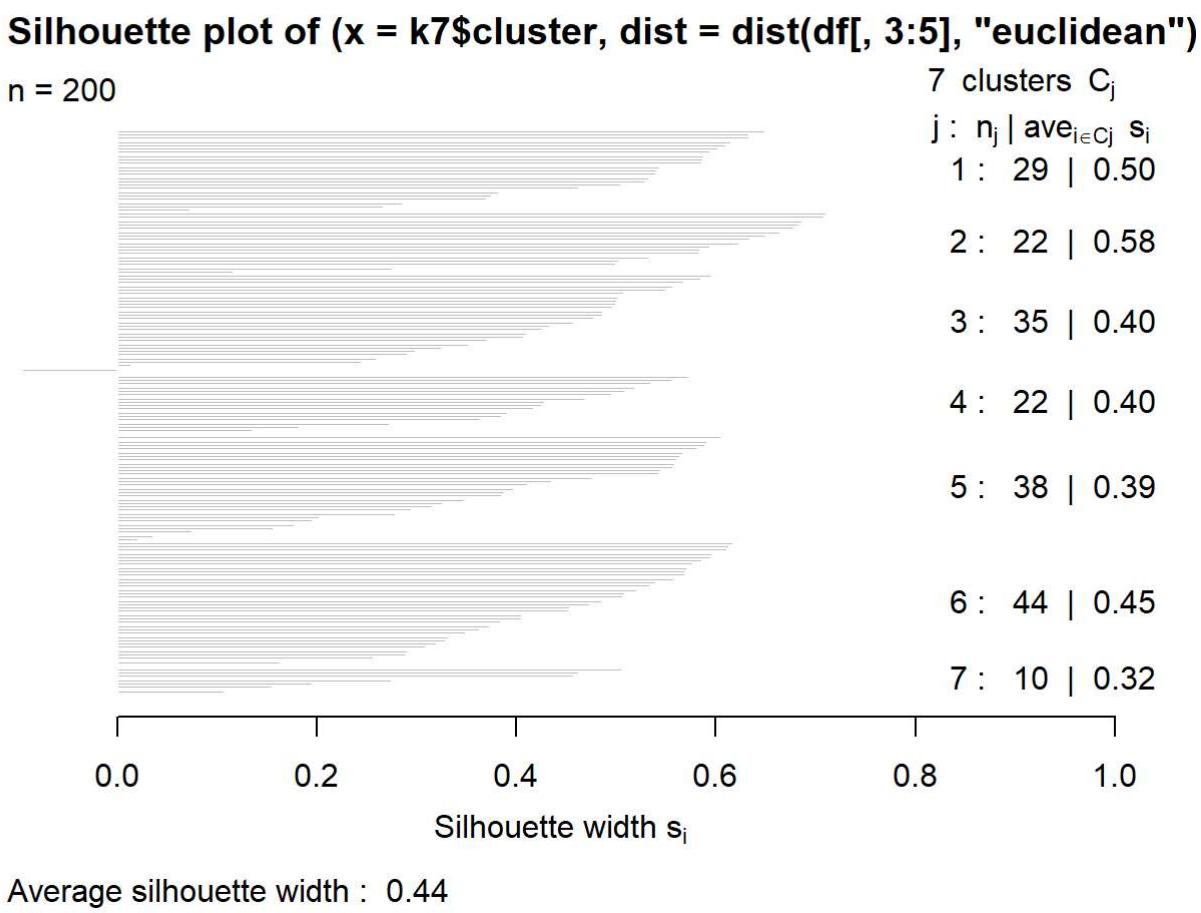

```
k5<-kmeans(df[,3:5],5,iter.max=100,nstart=50,algorithm="Lloyd")
s5<-plot(silhouette(k5$cluster,dist(df[,3:5],"euclidean")))
```



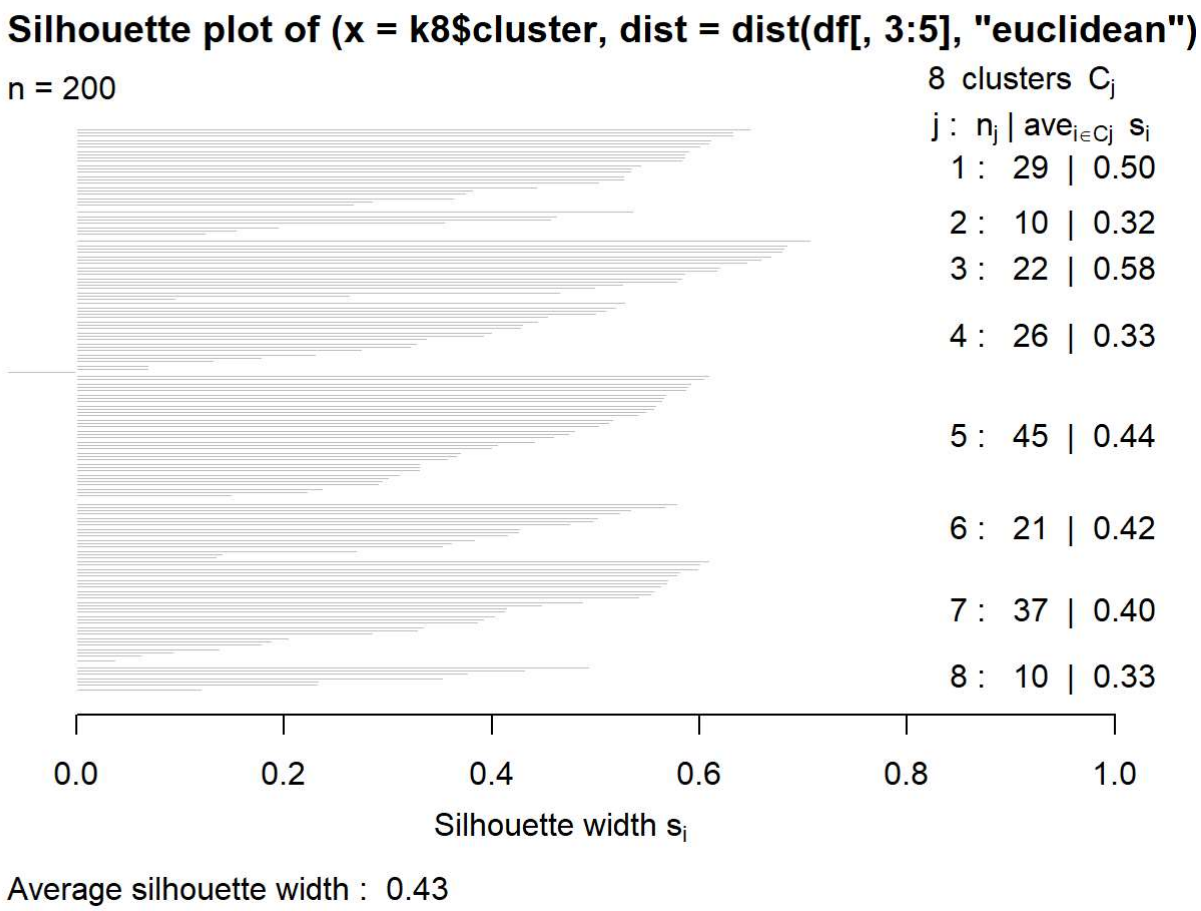

```
k6<-kmeans(df[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
s6<-plot(silhouette(k6$cluster,dist(df[,3:5],"euclidean")))
```



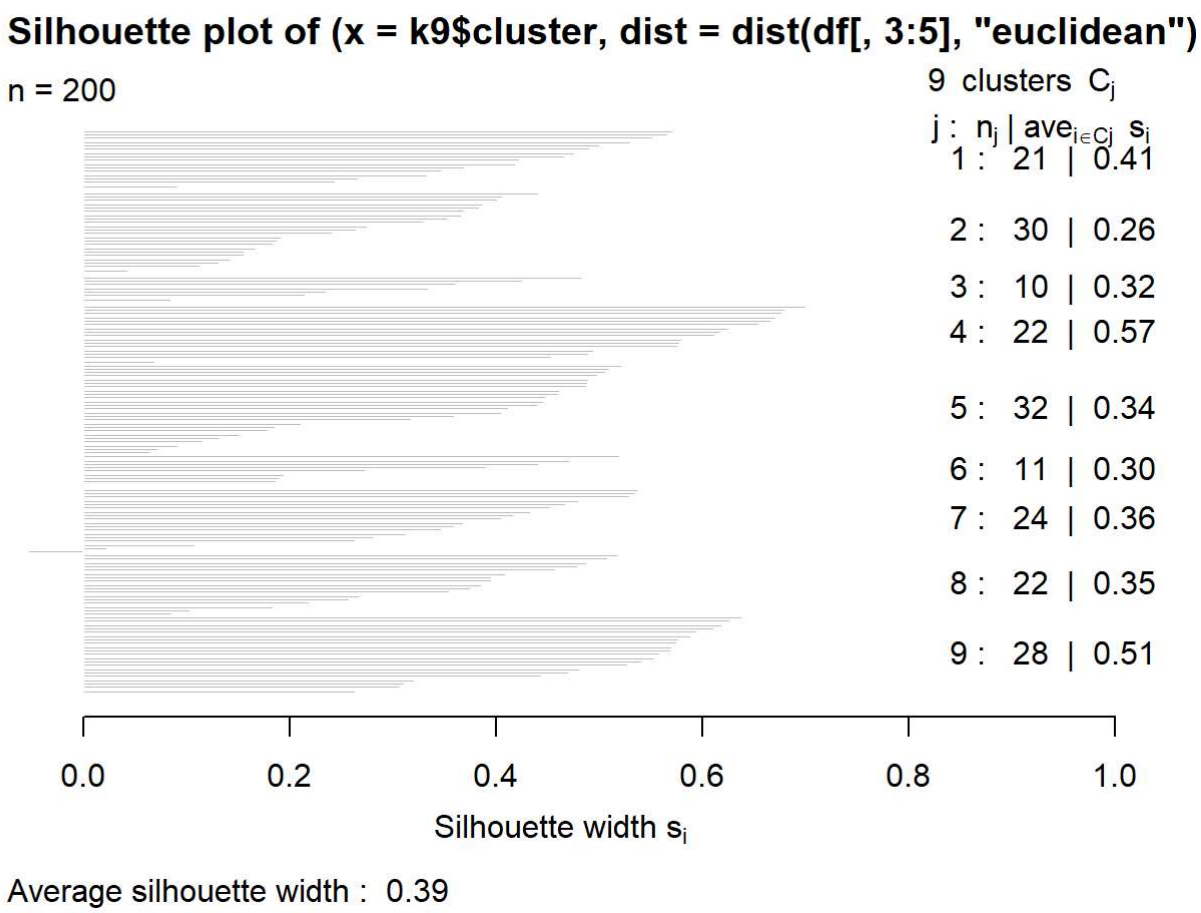

```
k7<-kmeans(df[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")
s7<-plot(silhouette(k7$cluster,dist(df[,3:5],"euclidean")))
```



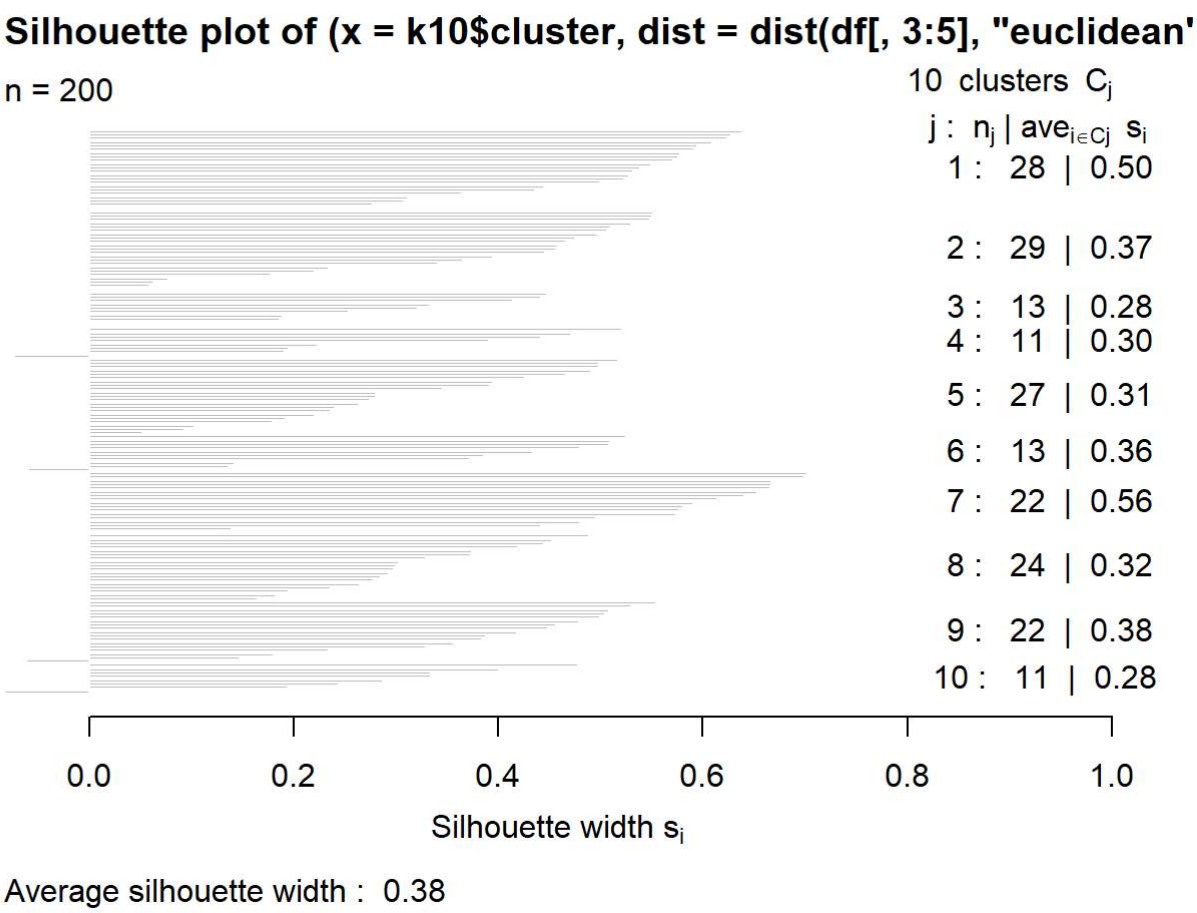

```
k8<-kmeans(df[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")
s8<-plot(silhouette(k8$cluster,dist(df[,3:5],"euclidean")))
```



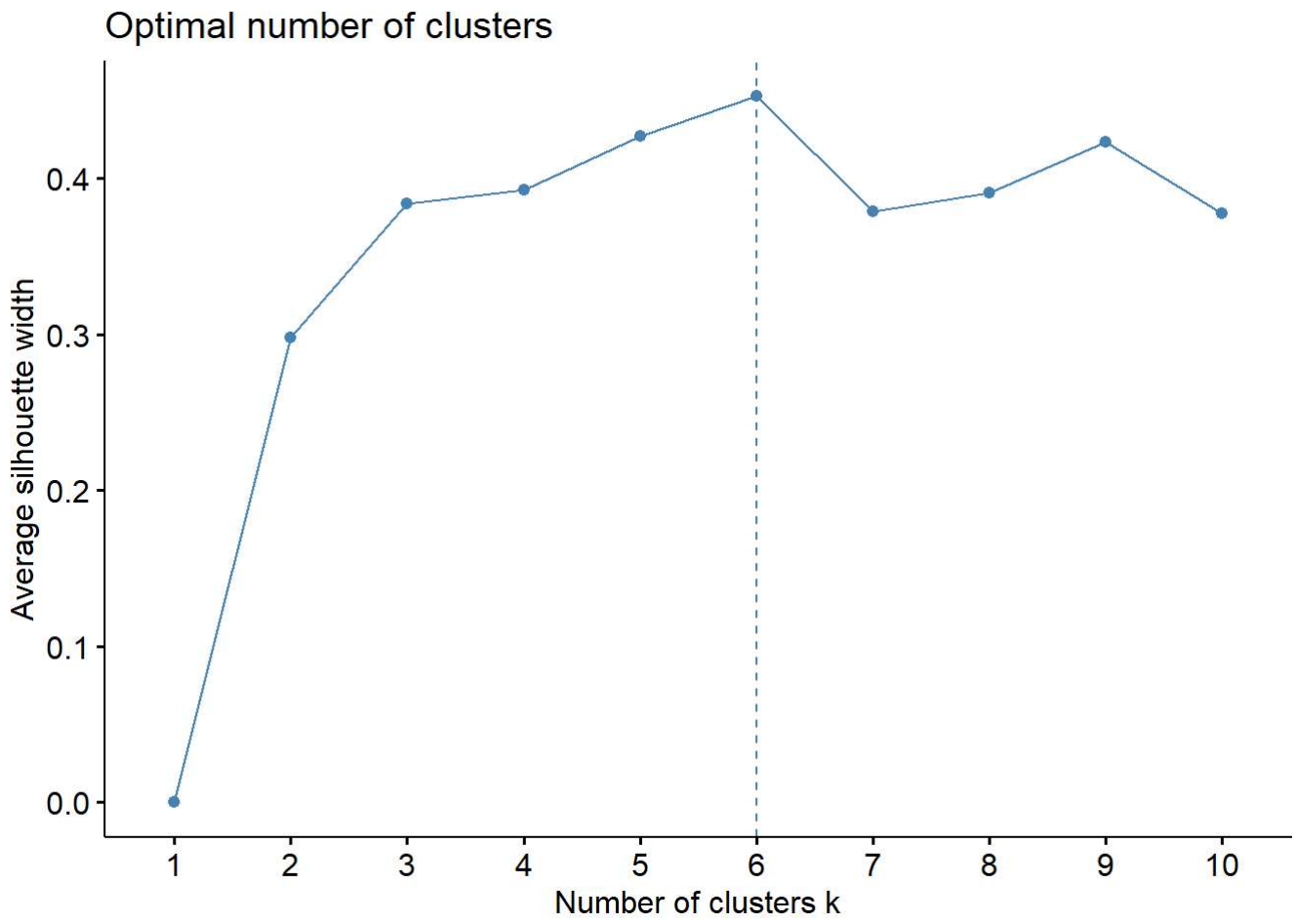

```
k9<-kmeans(df[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")
s9<-plot(silhouette(k9$cluster,dist(df[,3:5],"euclidean")))
```



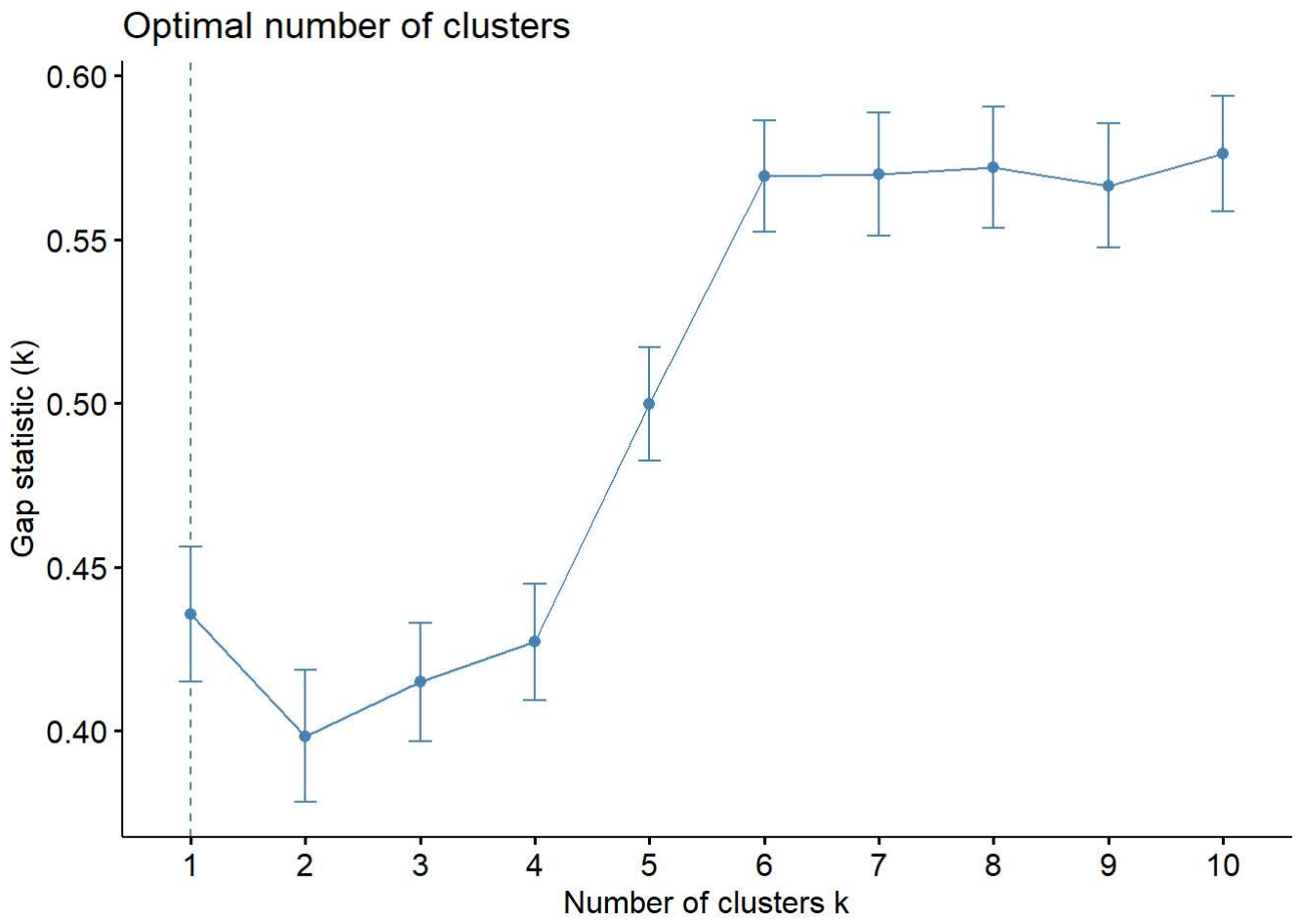

```
k10<-kmeans(df[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")
s10<-plot(silhouette(k10$cluster,dist(df[,3:5],"euclidean")))
```




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




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




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

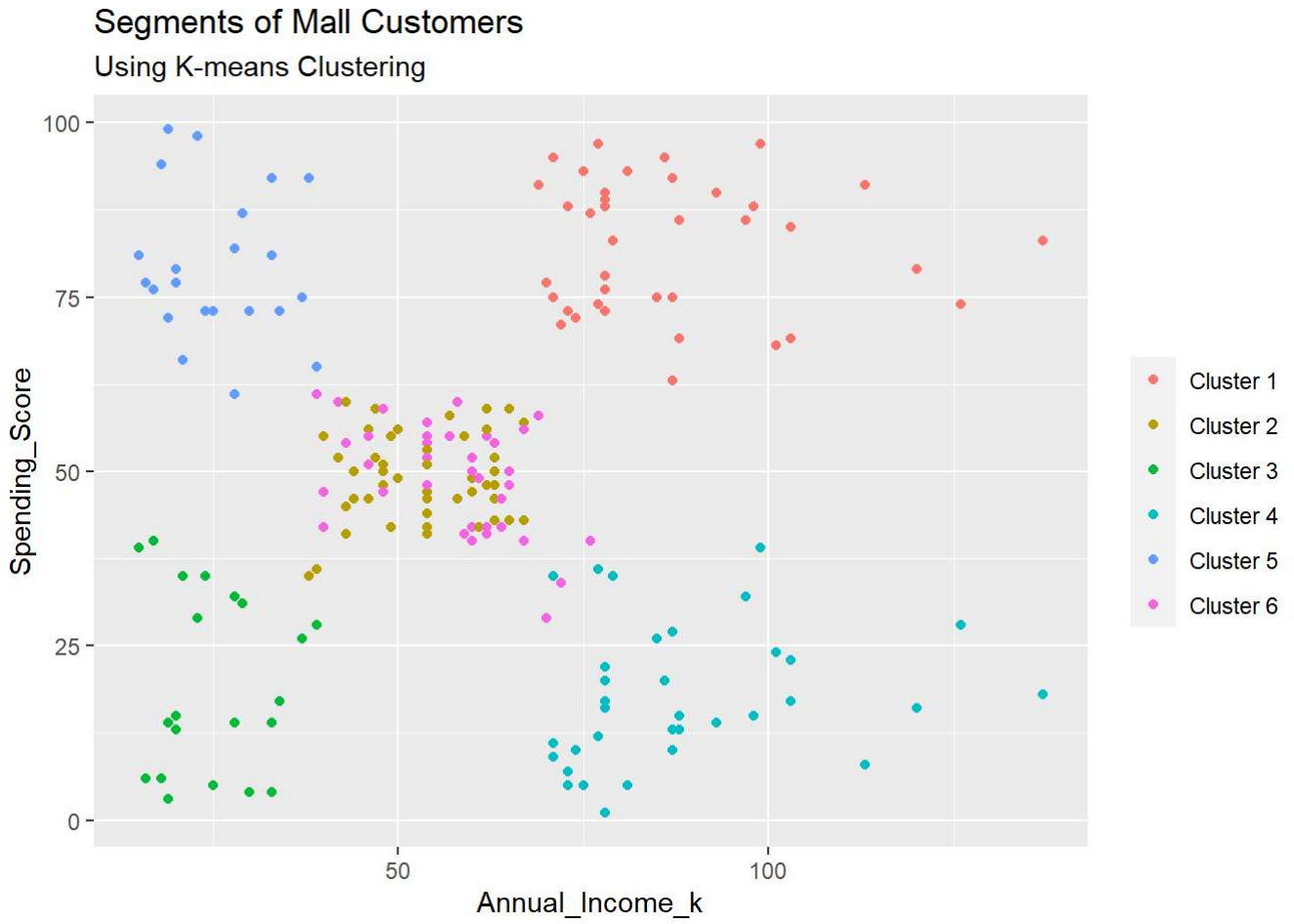
```
## Importance of components:
##                           PC1     PC2     PC3
## Standard deviation     26.4625 26.1597 12.9317
## Proportion of Variance  0.4512  0.4410  0.1078
## Cumulative Proportion   0.4512  0.8922  1.0000
```

```
pcclust$rotation[,1:2]
```

```
##                        PC1        PC2
## Age              0.1889742 -0.1309652
## Annual_Income_k -0.5886410 -0.8083757
## Spending_Score  -0.7859965  0.5739136
```

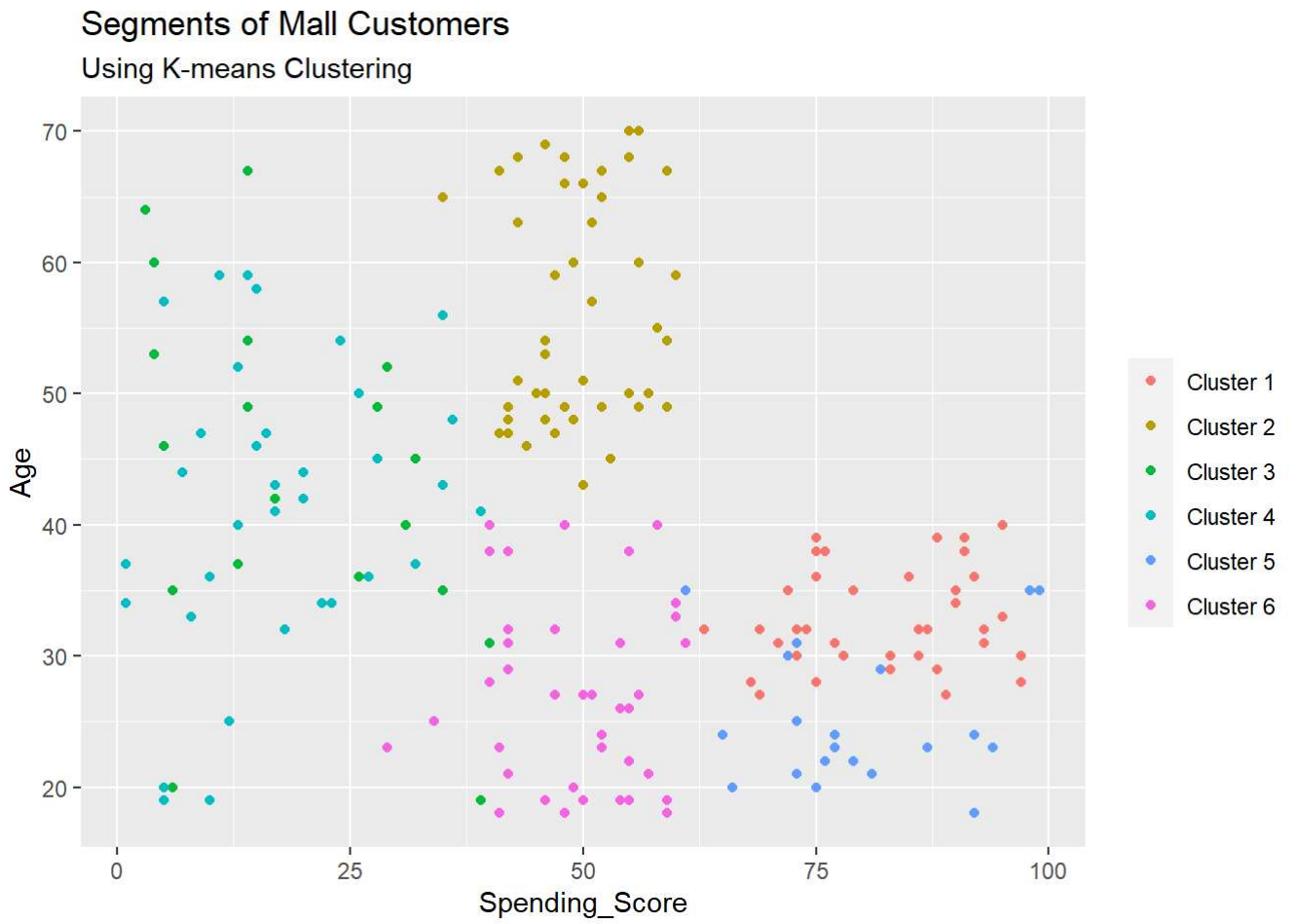
7,8) Below is the first output of the model: - Cluster 1: Low Income, High Spending - Cluster 2: High Income, High Spending - Cluster 3 & 4: Medium Income, Medium Spending - Cluster 5: Low Income, Low Spending - Cluster 6: High Income, Low Spending

```
set.seed(1)
ggplot(df, aes(x =Annual_Income_k, y = Spending_Score)) +
  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","Clus
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```



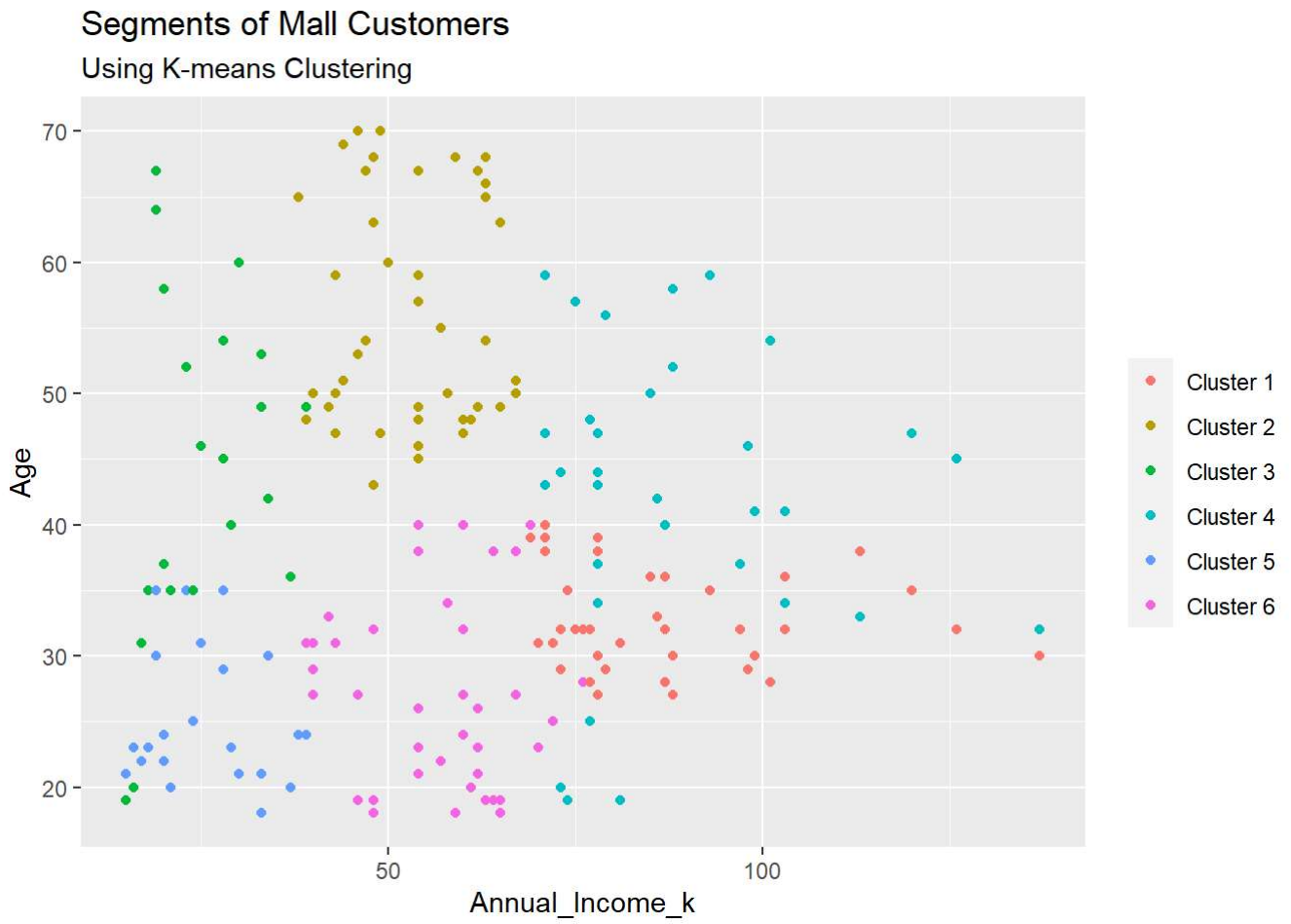
7,8) Below is the second output of of model. - Cluster 1 & 2: Young age, high spending
- Cluster 3: Older age, medium spending - Cluster 4: Younger age, Medium Spending - Cluster 5 & 6: Low Spending

```
ggplot(df, aes(x =Spending_Score, 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
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```



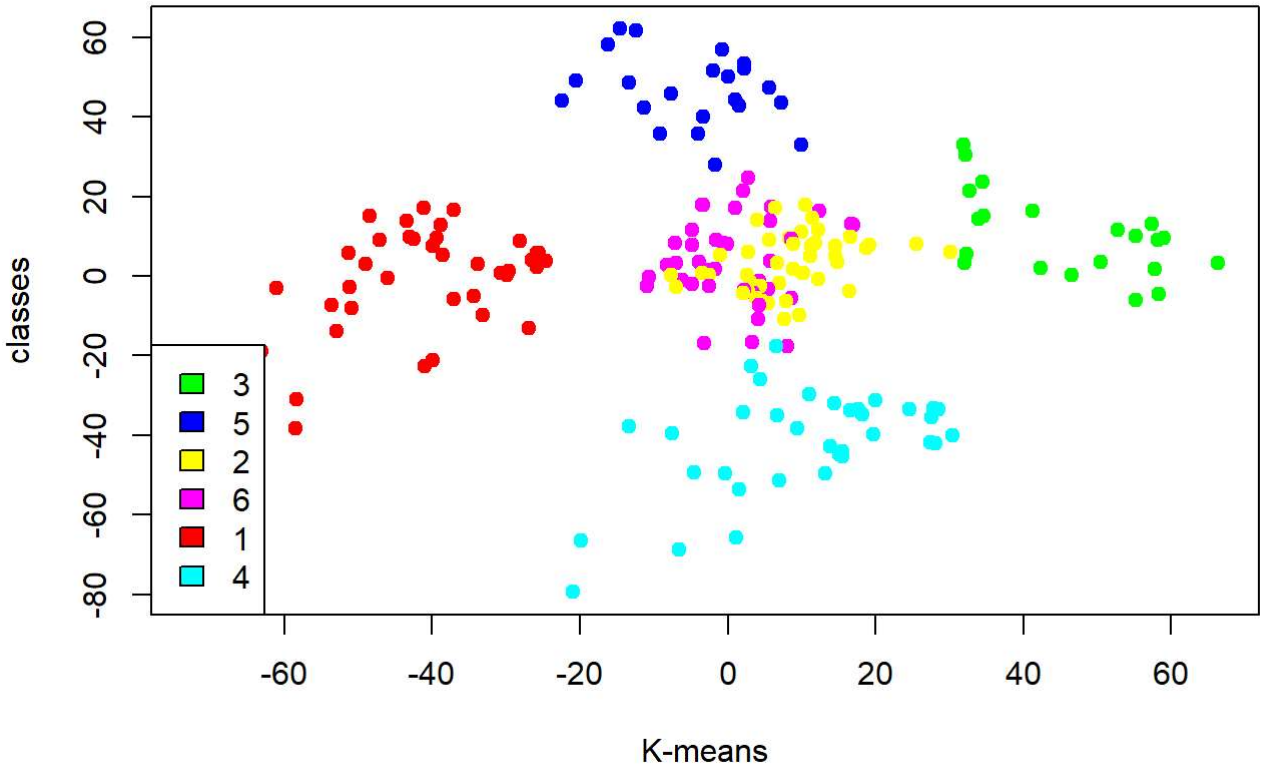
7,8) Below is the second output of of model. - Cluster 1: Young Age, Low Income - Cluster 2: Young age, high annual income - Cluster 3: Old age, medium income - Cluster 4: Young age, medium income - Cluster 5: Low income - Cluster 6: High Annual Income

```
ggplot(df, aes(x =Annual_Income_k, 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
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")
```



7,8) Below is the second output of of model. PCA1 is the first axis (x) and PCA2 is the second (y). - Cluster 3 & 4: Medium PCA1 and PCA2 score. - Cluster 2: Medium PCA2 score and Low PCA1 score - Cluster 1: Medium PCA1 score, high PCA2 score - Cluster 5: High PCA1 score, medium PCA2 score - Cluster 6: Medium PCA1 score, Low PCA2 score

```
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)))
```



Conclusion: We will be making improvements in our malls based on the 6 group of customers that we came up with.