DS_Customer Segmentation

Input Data

```
cs_data = read.csv("./data/Mall_Customers.csv")
str(cs_data)
```

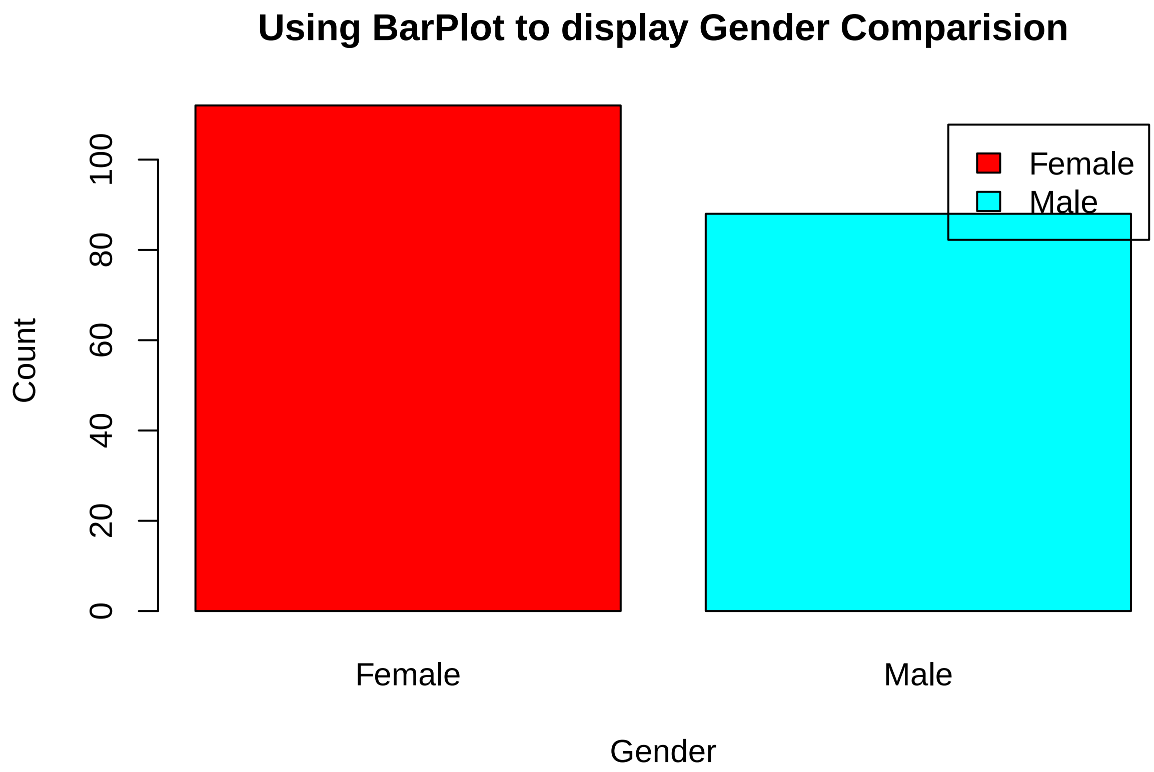
```
## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1 ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  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 ...
```

```
head(cs_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
```

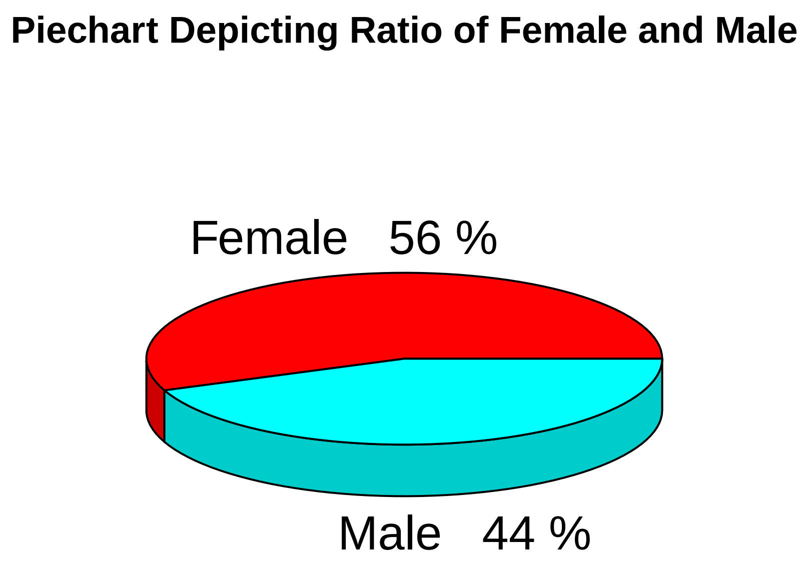
Customer Gender Visualization

```
### Barplot to show gender comparision
a = table(cs_data$Gender)
barplot(a, main = "Using BarPlot to display Gender Comparision",
        xlab = "Gender",
        ylab = "Count",
        col = rainbow(2),
        legend = rownames(a))
```



From the above barplot, we can observe that the number of females is higher than the males.

```
### Piechart to show the ratiio of gender distribution
library(plotrix)
pct = round(a/sum(a) * 100)
lbs = paste(c("Female", "Male"), " ", pct, "%", sep = " ")
pie3D(a, labels = lbs,
      main = "Piechart Depicting Ratio of Female and Male")
```



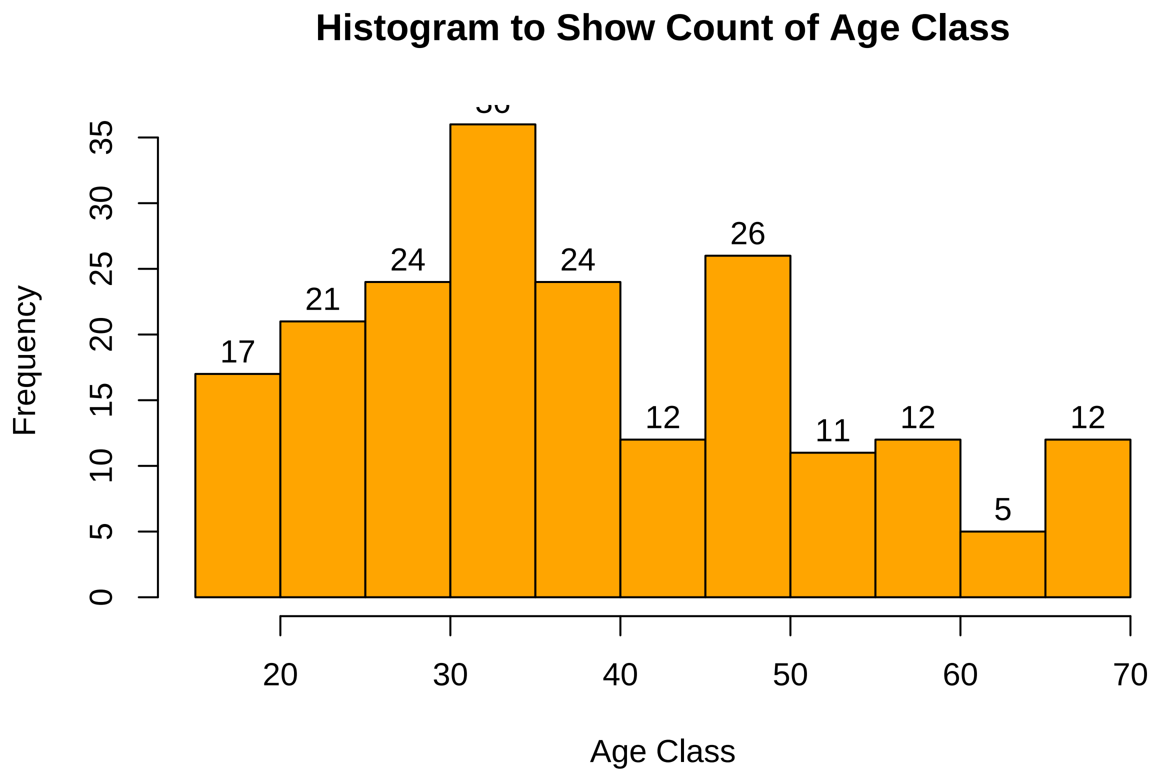
From the pie chart, we can conclude that the percentage of females is 56%, whereas the percentage of male in the customer dataset is 44%.

Visualization of Age Distribution

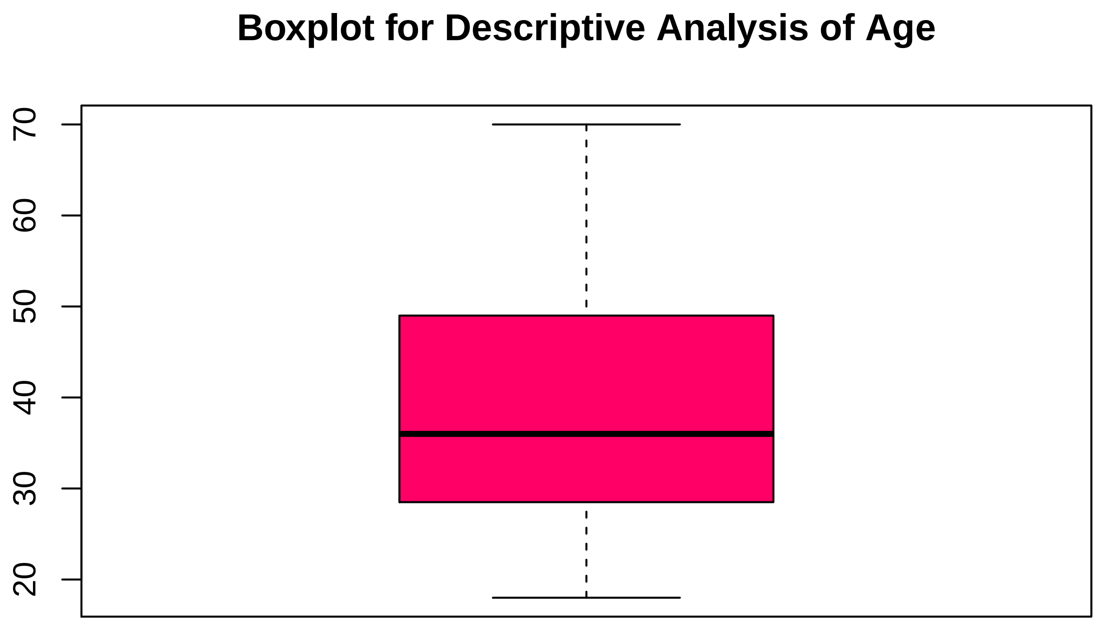
```
summary(cs_data$Age)
```

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

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




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



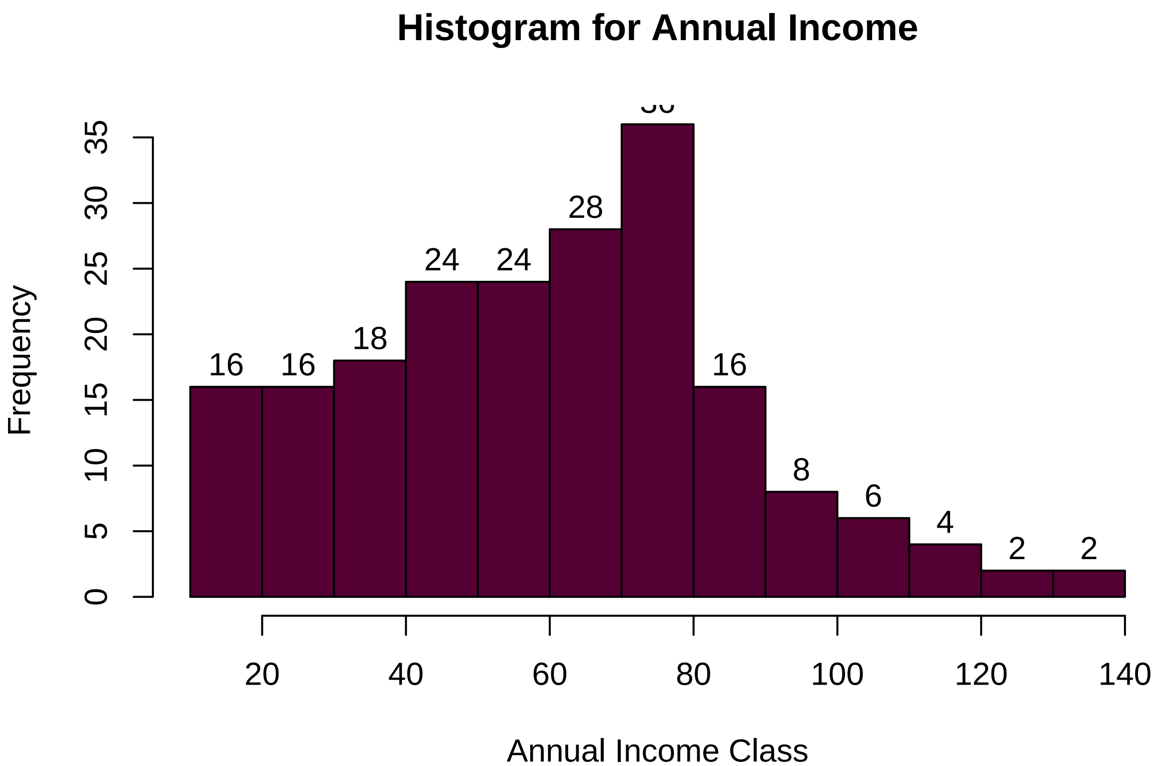
From the above two visualizations, we conclude that the maximum customer ages are between 30 and 35. The minimum age of customers is 18, whereas, the maximum age is 70.

Analysis of the Annual Income of the Customers

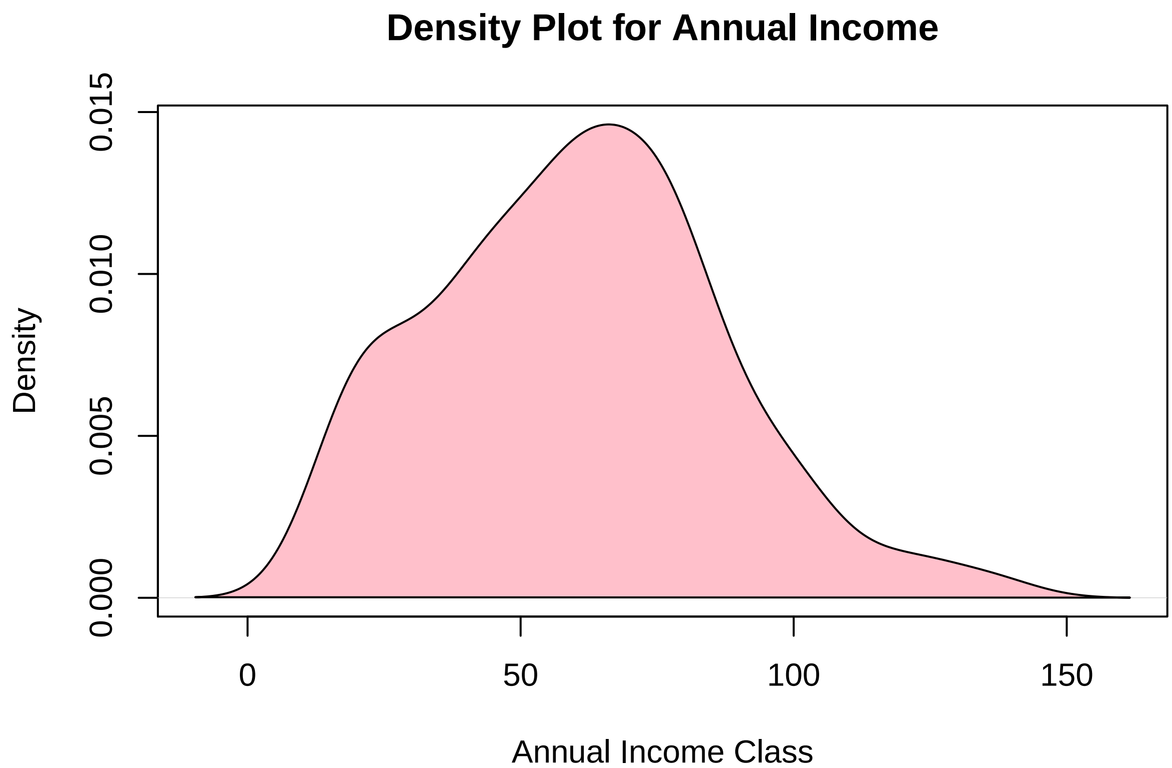
```r
summary(cs_data$Annual.Income..k..)
```

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

```r
hist(cs_data$Annual.Income..k..,
  col="#550033",
  main="Histogram for Annual Income",
  xlab="Annual Income Class",
  ylab="Frequency",
  labels=TRUE)
```




```r
plot(density(cs_data$Annual.Income..k..),
    col="pink",
    main="Density Plot for Annual Income",
    xlab="Annual Income Class",
    ylab="Density")
polygon(density(cs_data$Annual.Income..k..),
        col = "pink")
```



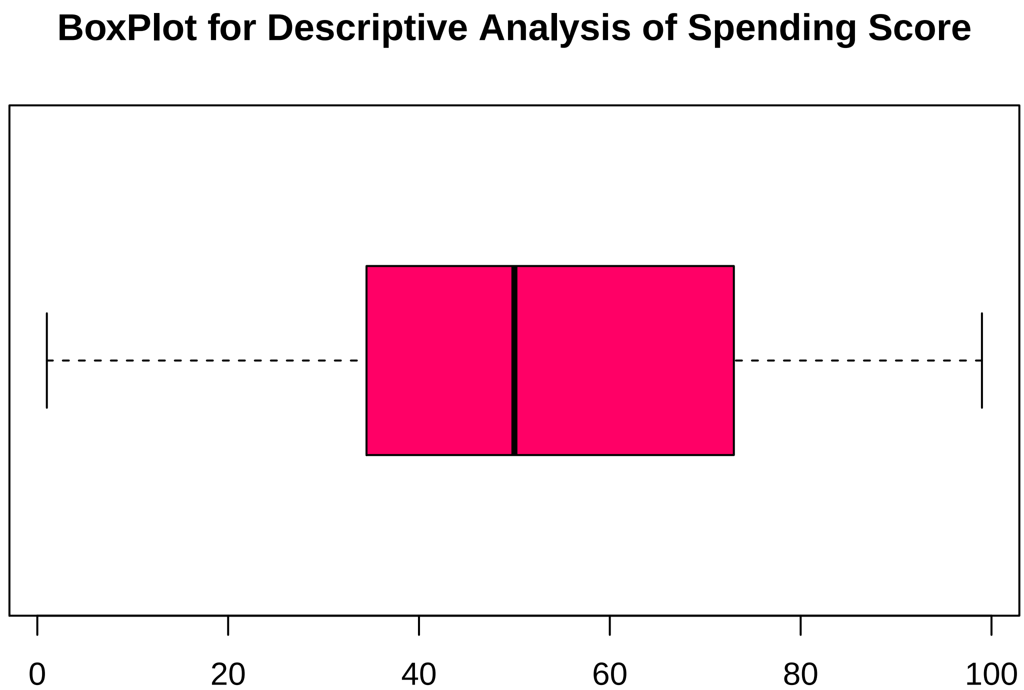
From the above descriptive analysis, we conclude that the minimum annual income of the customers is 15 and the maximum income is 137. People earning an average income of 70 have the highest frequency count in our histogram distribution. The average salary of all the customers is 60.56. In the Kernel Density Plot that we displayed above, we observe that the annual income has a normal distribution.

Analyzing Spending Score of the Customers

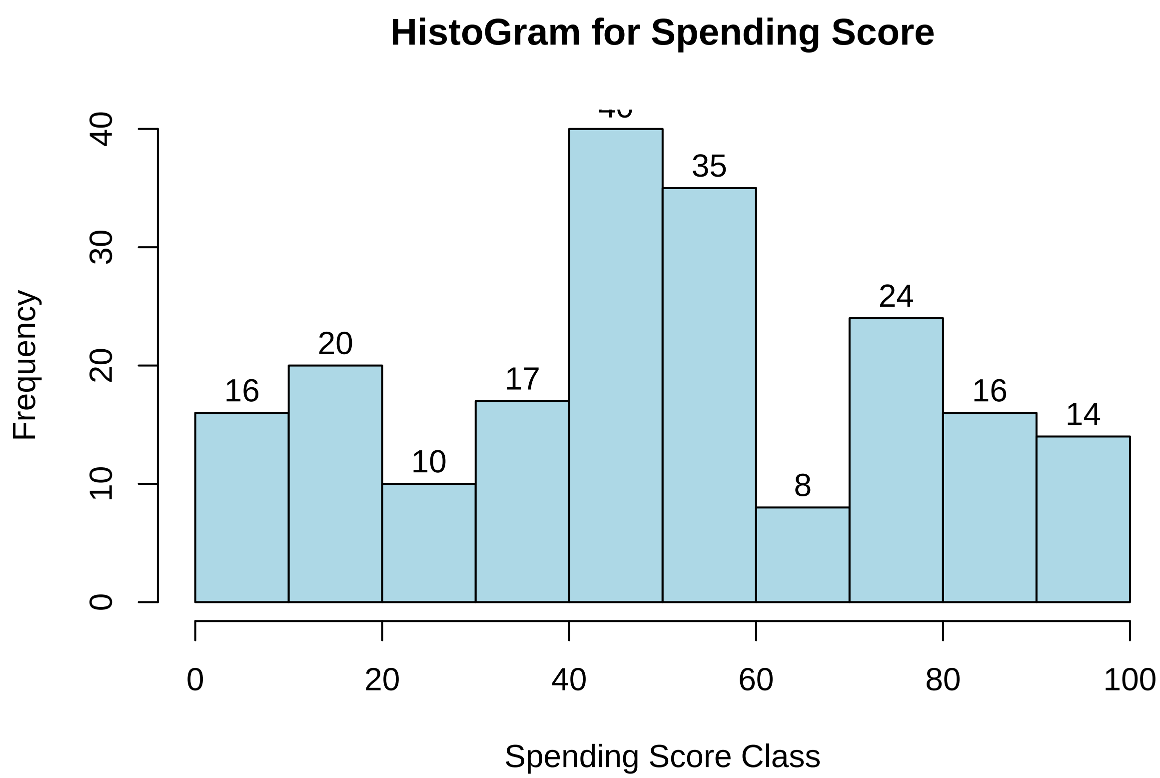
```
summary(cs_data$Spending.Score..1.100.)
```

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

```
boxplot(cs_data$Spending.Score..1.100.,
   horizontal=TRUE,
   col="#ff0066",
   main="BoxPlot for Descriptive Analysis of Spending Score")
```




```
hist(cs_data$Spending.Score..1.100.,
    main="HistoGram for Spending Score",
    xlab="Spending Score Class",
    ylab="Frequency",
    col="lightblue",
    labels=TRUE)
```



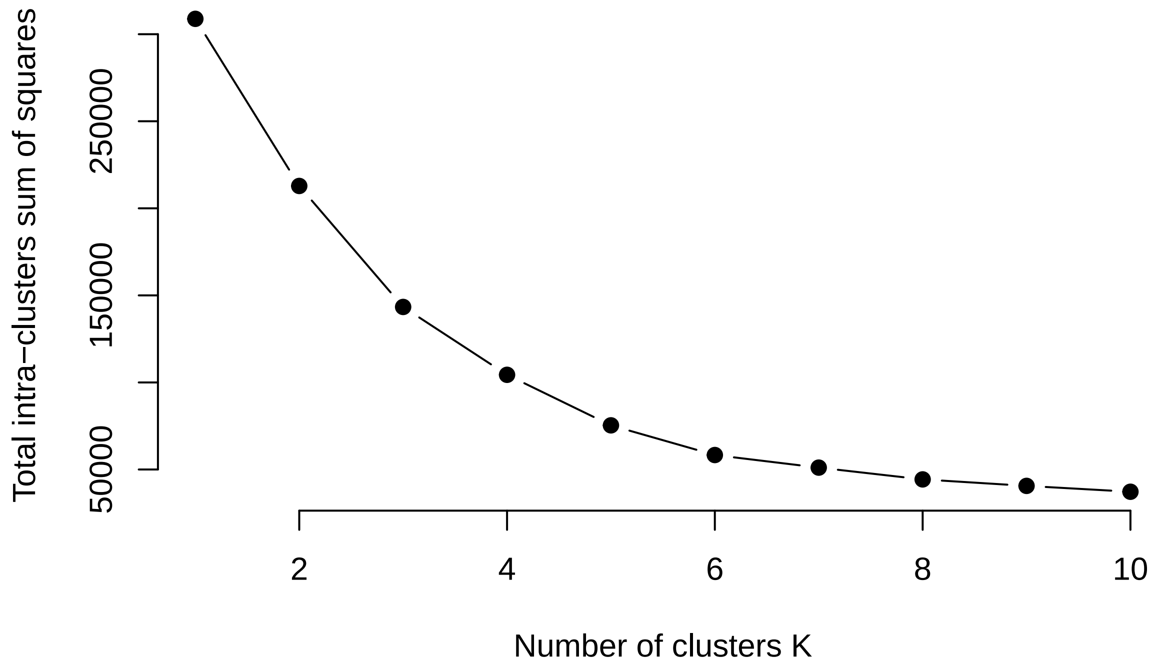
The minimum spending score is 1, maximum is 99 and the average is 50.20. We can see Descriptive Analysis of Spending Score is that Min is 1, Max is 99 and avg. is 50.20. From the histogram, we conclude that customers between class 40 and 50 have the highest spending score among all the classes.

K-means Algorithm

```
library(purrr)
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
set.seed(123)
# Function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(cs_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")
```



From the above graph, we conclude that 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

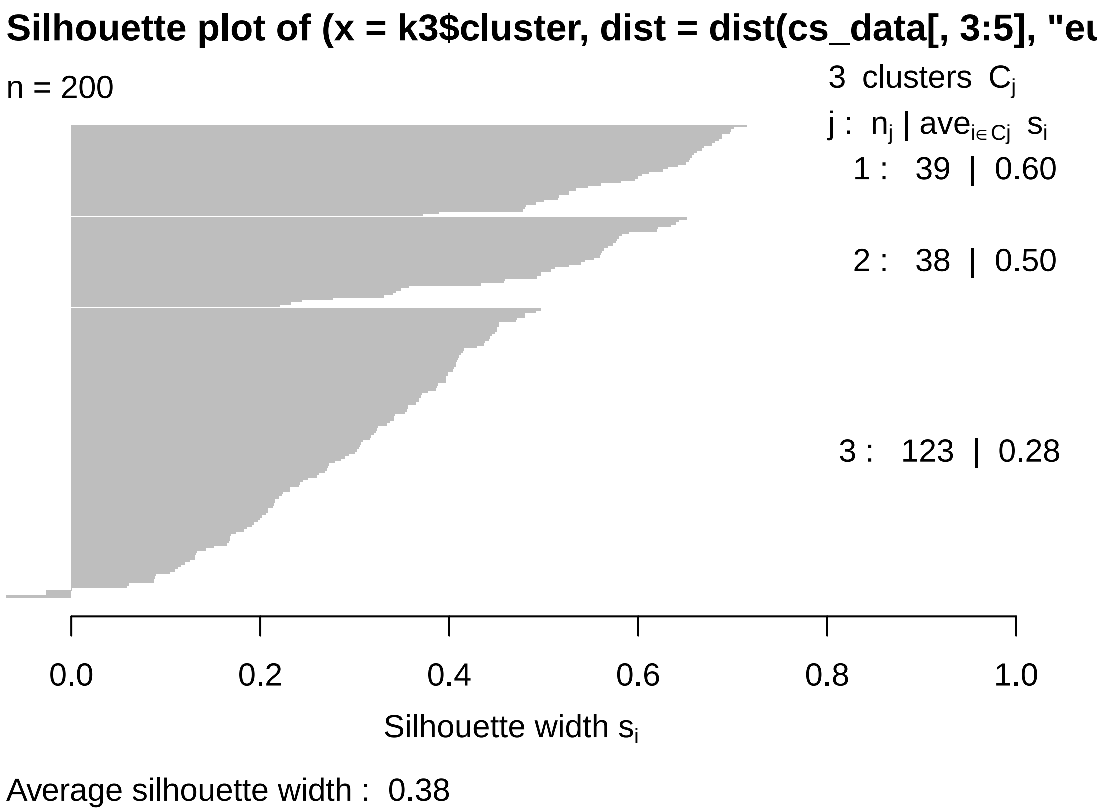
Average Silhouette Method

```r
library(cluster)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.2
```

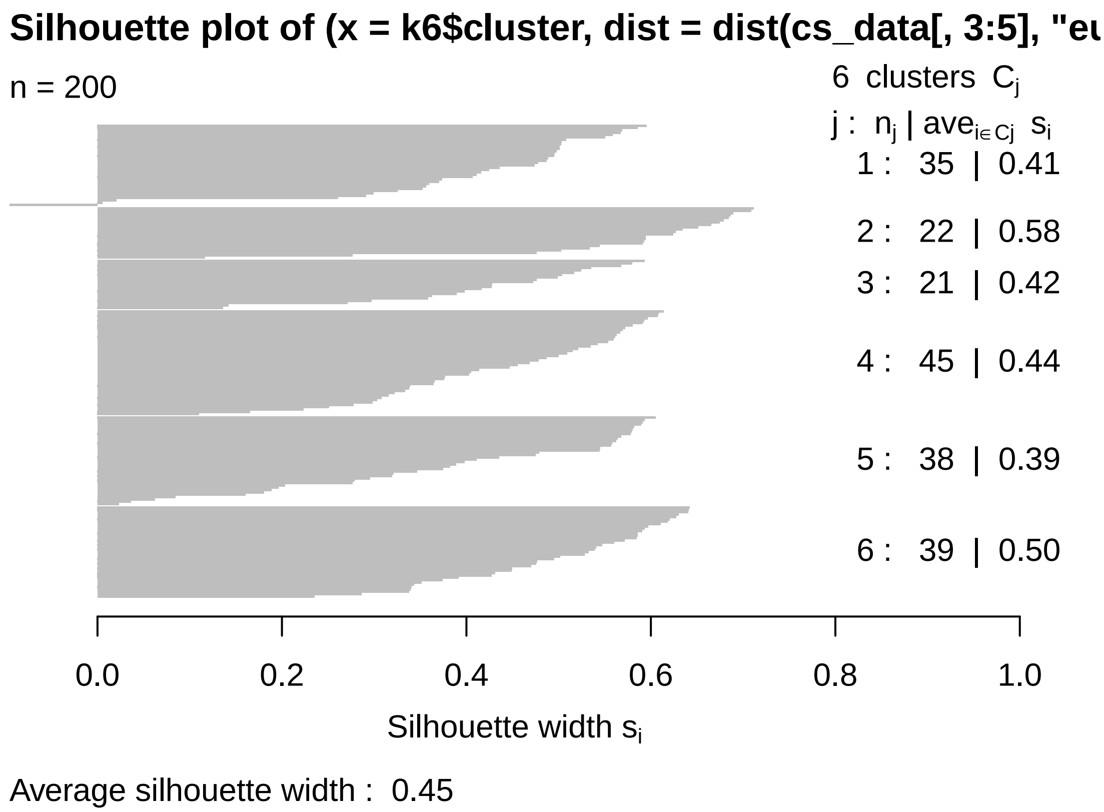
```r
library(grid)

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


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

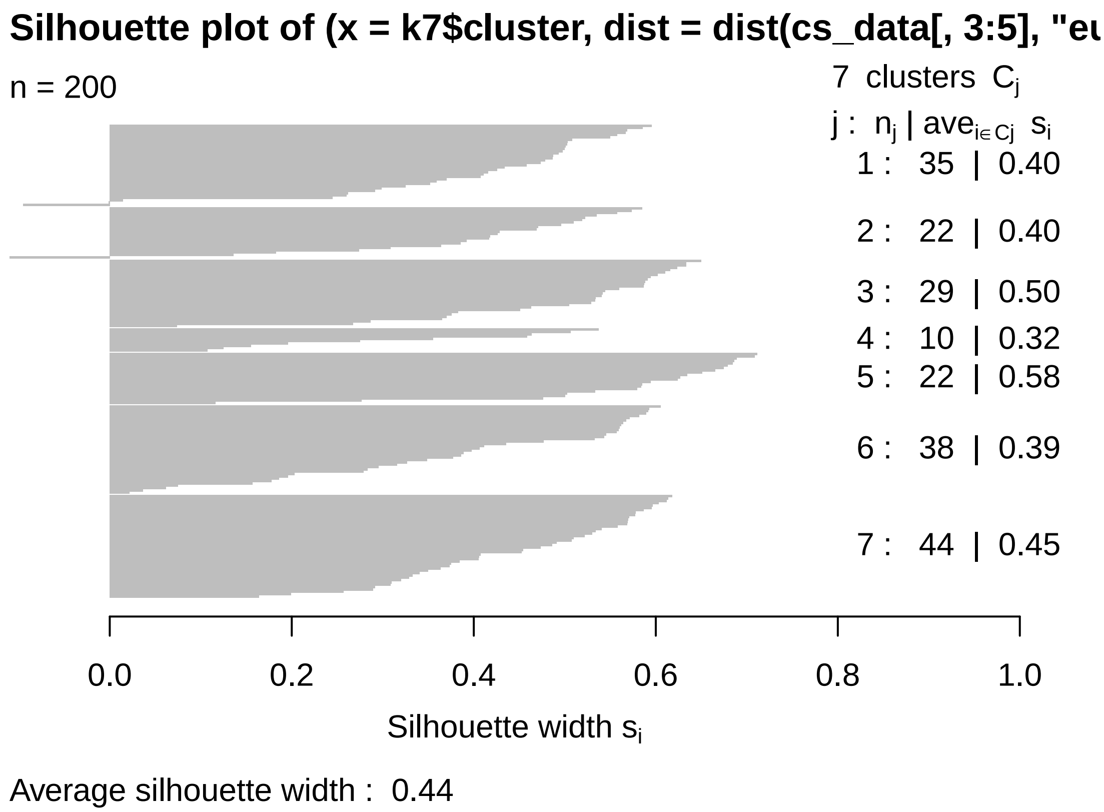



```
k4<-kmeans(cs_data[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")
s4<-plot(silhouette(k4$cluster,dist(cs_data[,3:5],"euclidean")))
```

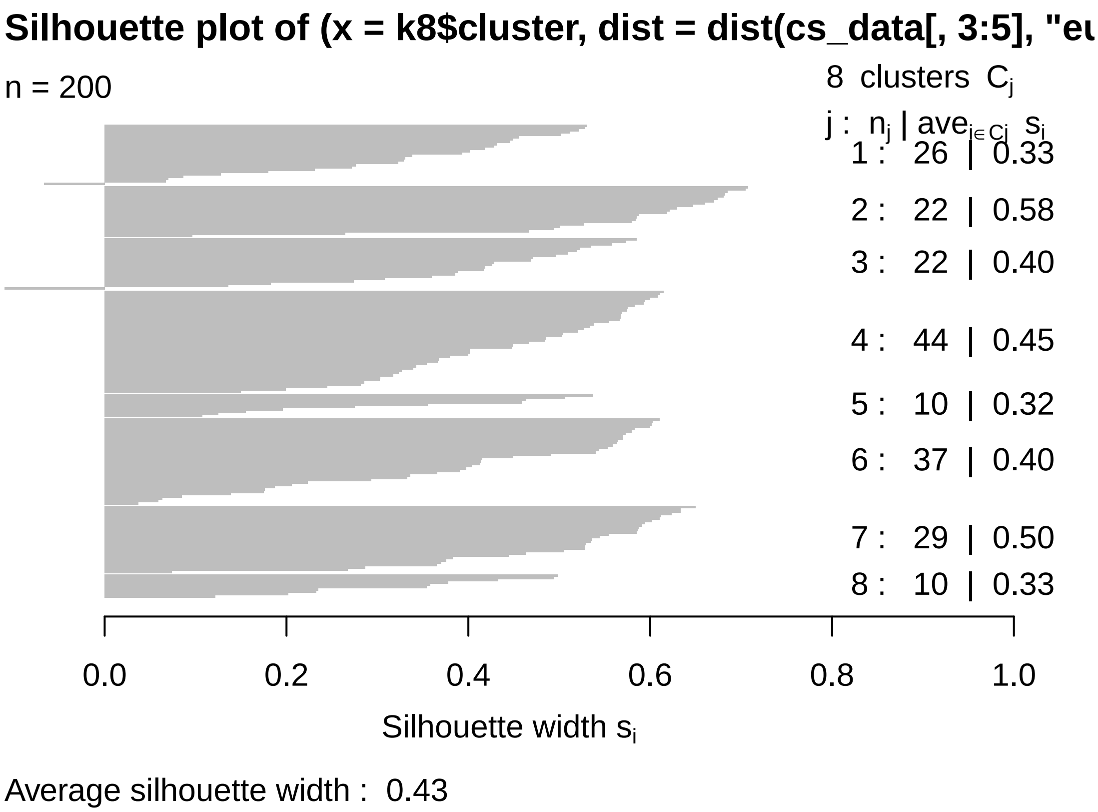

```
k6<-kmeans(cs_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
s6<-plot(silhouette(k6$cluster,dist(cs_data[,3:5],"euclidean")))
```




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

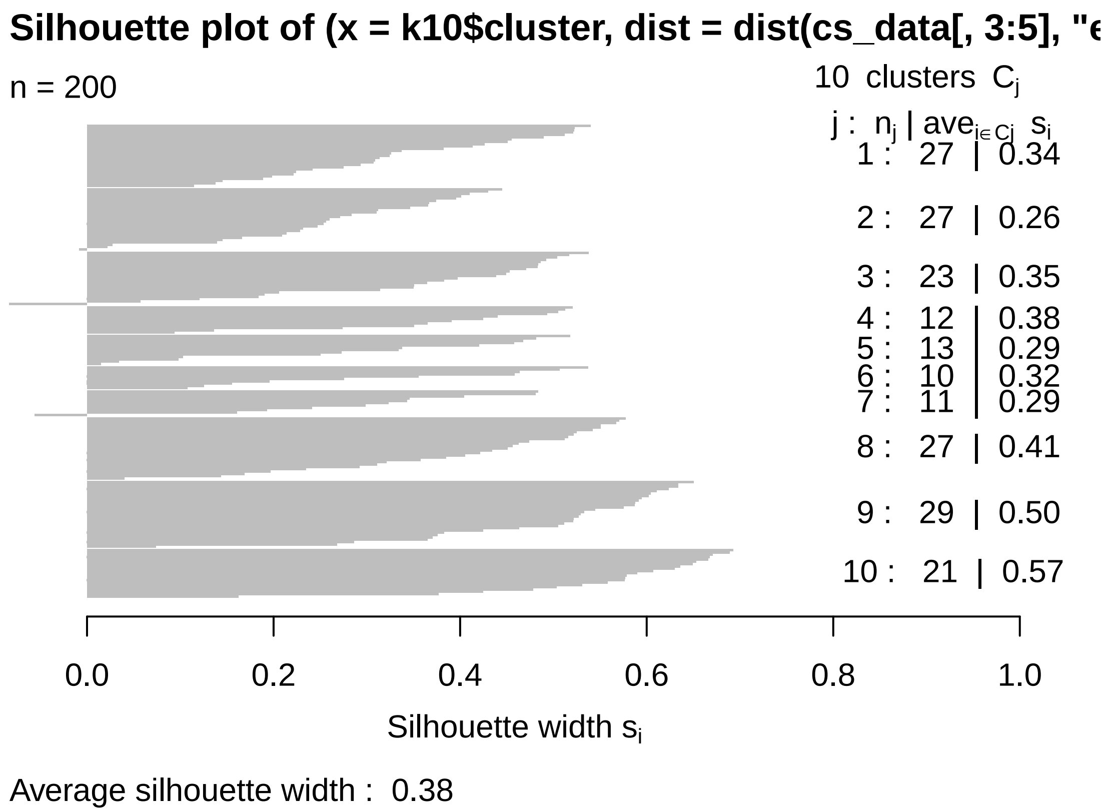



```
k8<-kmeans(cs_data[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")
s8<-plot(silhouette(k8$cluster,dist(cs_data[,3:5],"euclidean")))
```




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


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




```r
library(NbClust)
```

```
## Warning: package 'NbClust' was built under R version 3.5.2
```

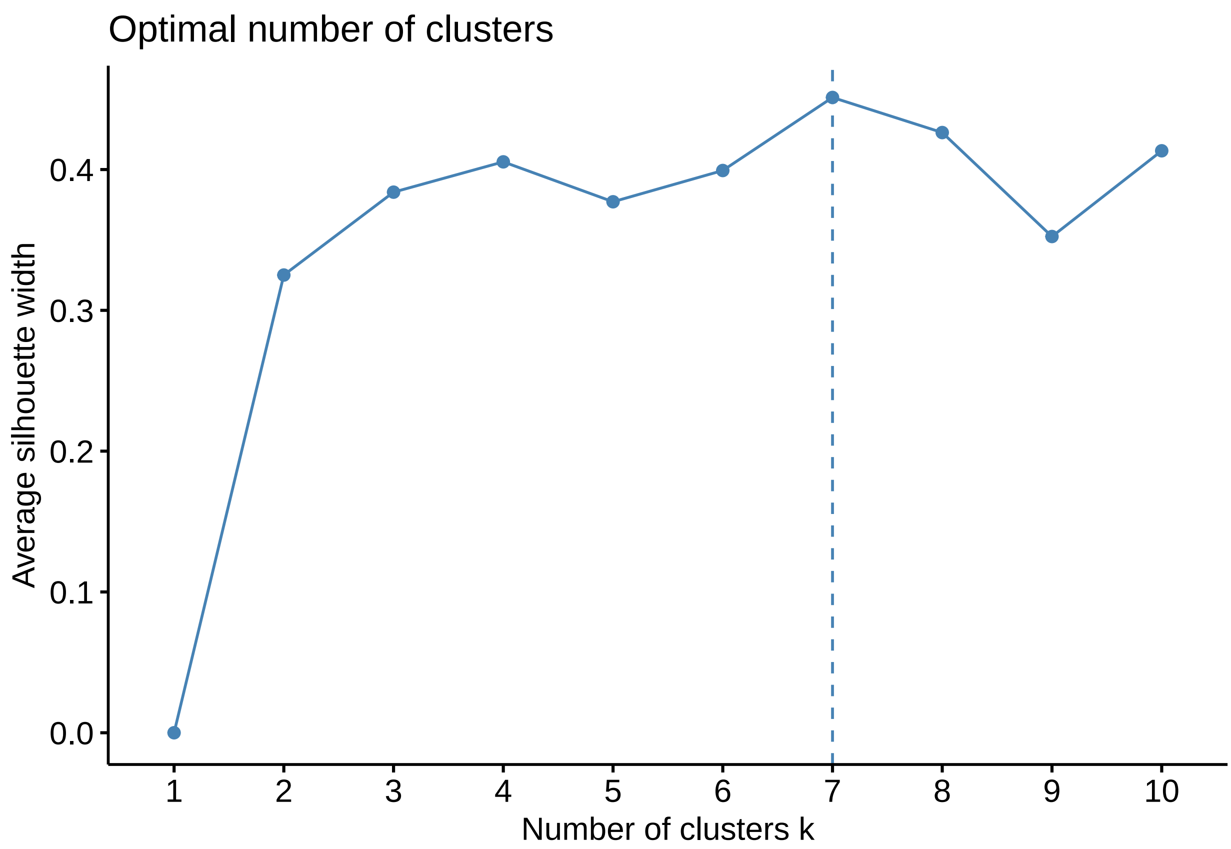
```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.5.3
```

```
## Loading required package: ggplot2
```

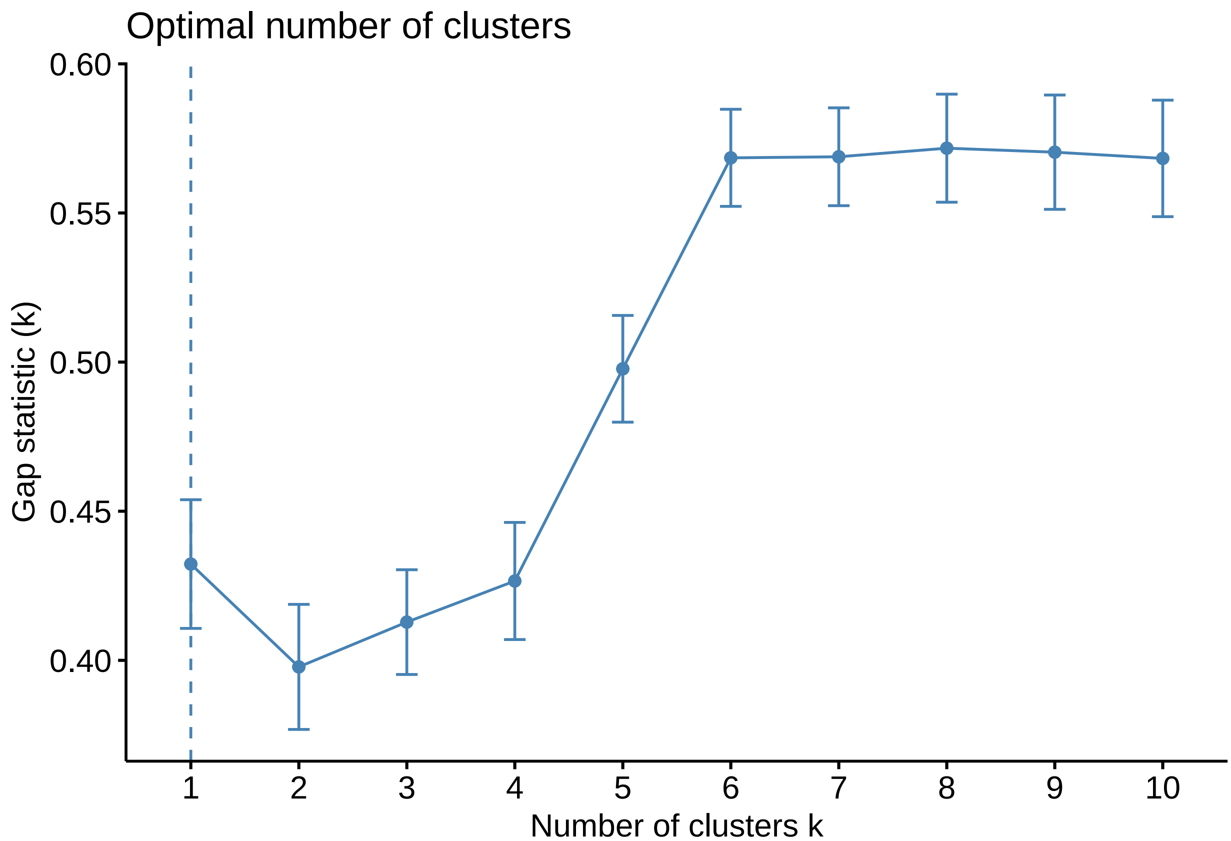
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

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



Gap Statistic Method

```
#Compute gap statistic
set.seed(123)
stat_gap <- clusGap(cs_data[,3:5], FUN = kmeans, nstart = 25,
            K.max = 10, B = 50)
fviz_gap_stat(stat_gap)
```




```r
#Take k = 6 as the optimal cluster
k6<-kmeans(cs_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
k6
```

```
## K-means clustering with 6 clusters of sizes 38, 45, 22, 21, 39, 35
##
## Cluster means:
##        Age Annual.Income..k.. Spending.Score..1.100.
## 1 27.00000           56.65789               49.13158
## 2 56.15556           53.37778               49.08889
## 3 25.27273           25.72727               79.36364
## 4 44.14286           25.14286               19.52381
## 5 32.69231           86.53846               82.12821
## 6 41.68571           88.22857               17.28571
##
## Clustering vector:
##   [1] 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
##  [36] 3 4 3 4 3 2 3 2 1 4 3 2 1 1 1 2 1 1 2 2 2 2 2 1 2 2 1 2 2 2 1 2 2 1 1
##  [71] 2 2 2 2 2 1 2 1 1 2 2 1 2 2 1 2 2 1 1 2 2 1 2 1 1 1 2 1 2 1 1 2 2 1 2
## [106] 1 2 2 2 2 2 1 1 1 1 1 2 2 2 2 1 1 1 5 1 5 6 5 6 5 6 5 1 5 6 5 6 5 6 5
## [141] 6 5 1 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6
## [176] 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5
##
## Within cluster sum of squares by cluster:
## [1]  7742.895  8062.133  4099.818  7732.381 13972.359 16690.857
##  (between_SS / total_SS =  81.1 %)
```

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Visualizing the Clustering Results using the First Two Principle Components

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

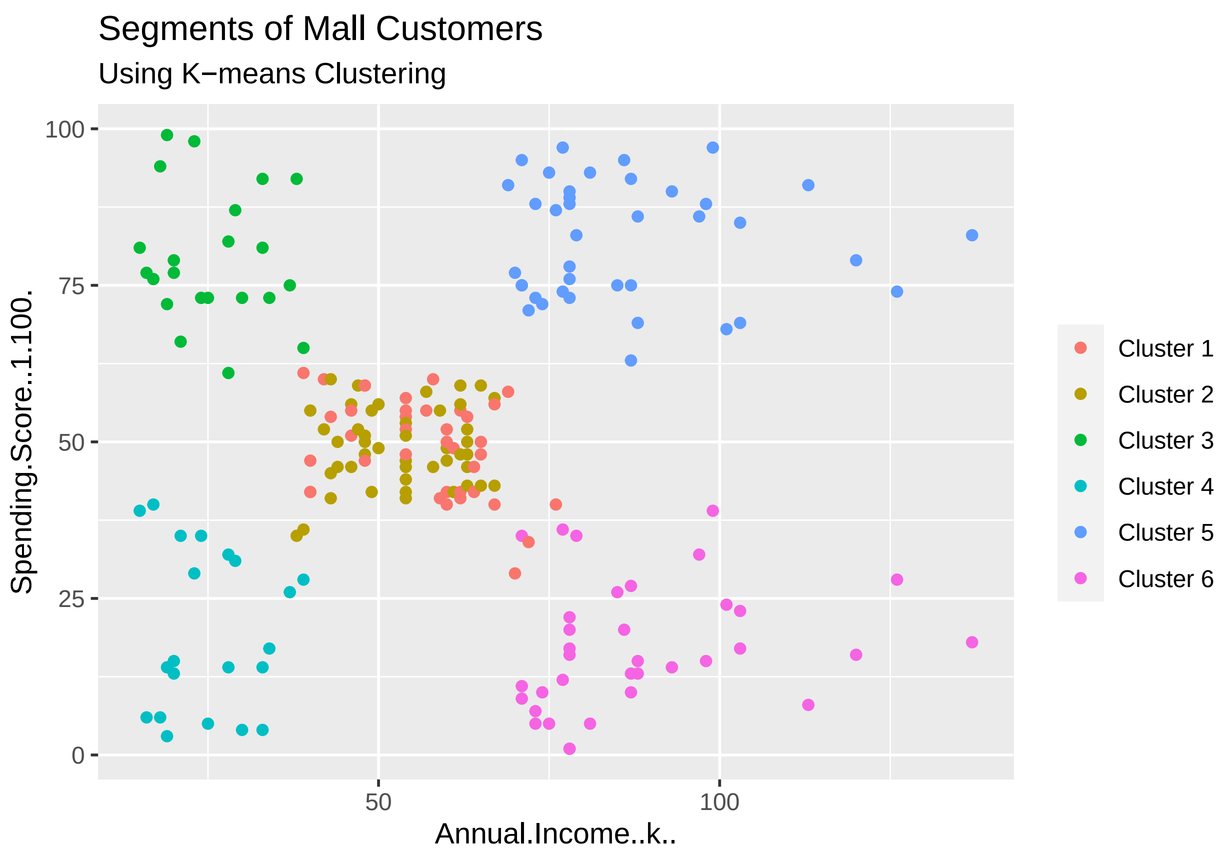
```
## Importance of components:
##                            PC1     PC2     PC3
## Standard deviation     26.4625 26.1597 12.9317
## Proportion of Variance  0.4512  0.4410  0.1078
## Cumulative Proportion   0.4512  0.8922  1.0000
```

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

```
##                               PC1        PC2
## Age                     0.1889742 -0.1309652
## Annual.Income..k..     -0.5886410 -0.8083757
## Spending.Score..1.100. -0.7859965  0.5739136
```

Visuliase the clusters

```r
set.seed(1)
ggplot(cs_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")
```



From the above visualization, we observe that there is a distribution of 6 clusters as follows:

Cluster 1 and 4 - These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.

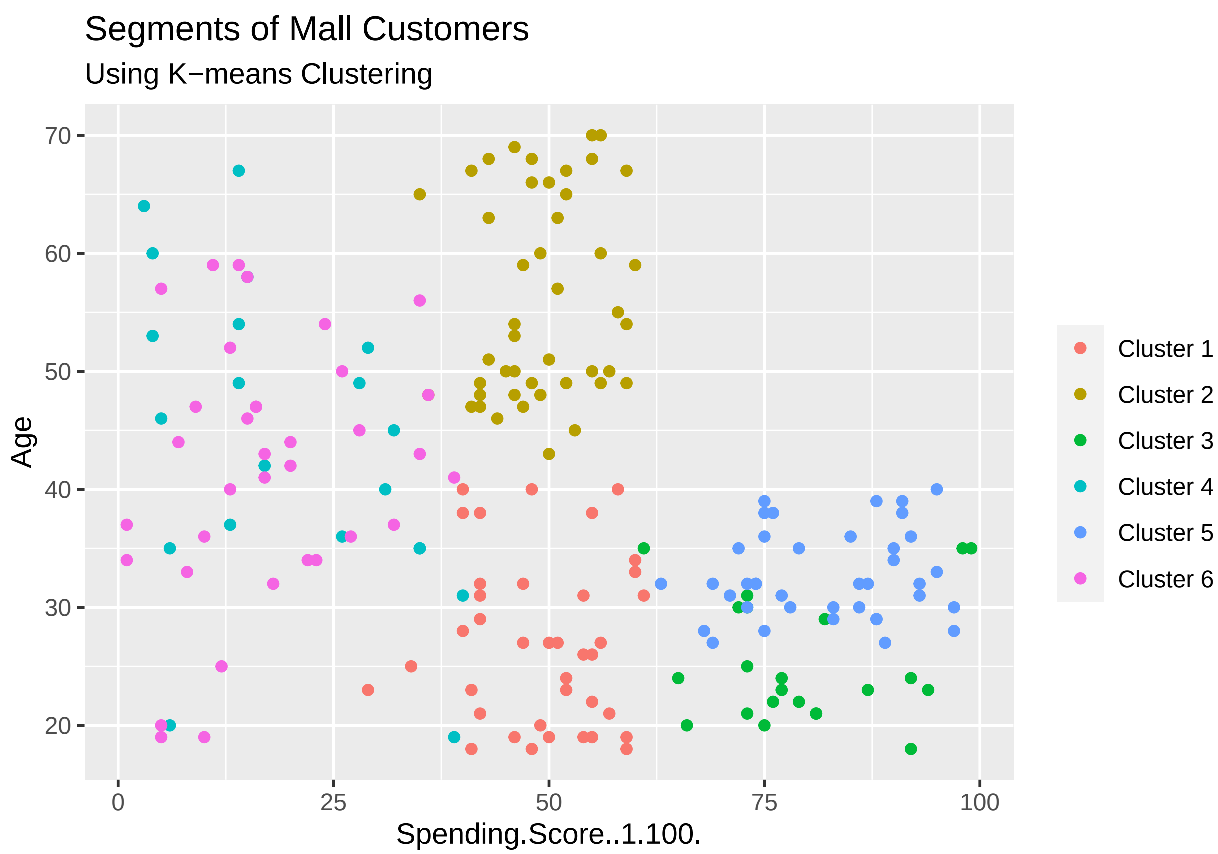
Cluster 2 - This cluster represents the customer_data having a high annual income as well as a high annual spend.

Cluster 3 - This cluster denotes the customer_data with low annual income as well as low yearly spend of income.

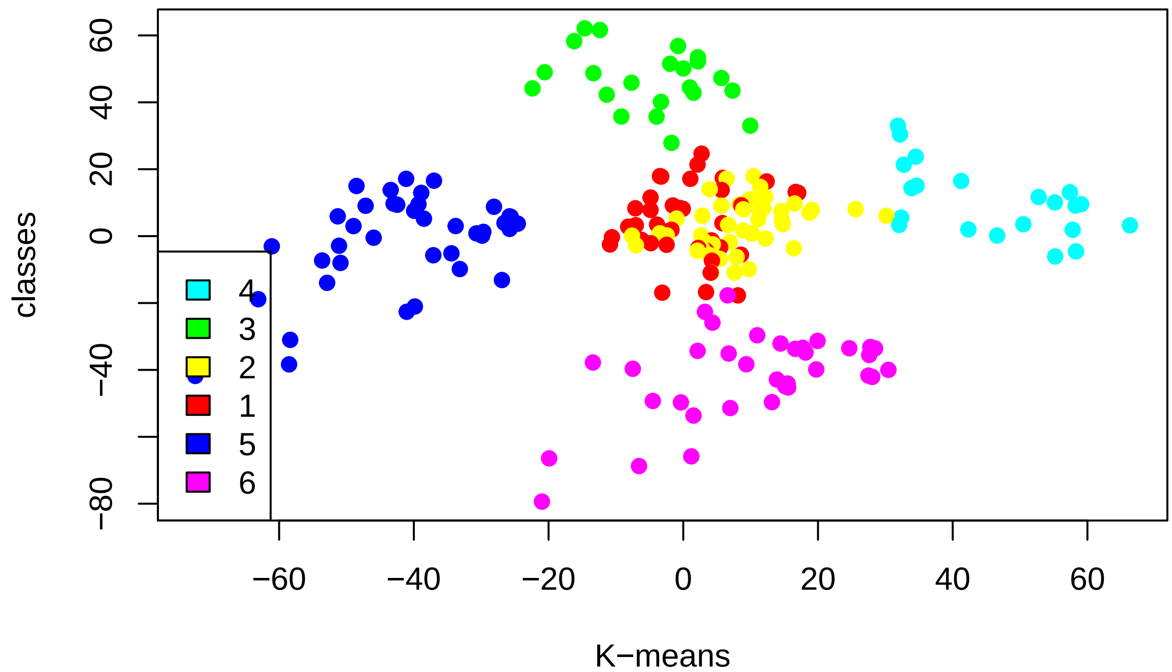
Cluster 5 - This cluster denotes a high annual income and low yearly spend.

Cluster 6 - This cluster represents a low annual income but its high yearly expenditure.

```
ggplot(cs_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
  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)))
```



Cluster 1 and 2 - These two clusters consist of customers with medium PCA1 and medium PCA2 score.

Cluster 5 - This cluster represents customers having a high PCA2 and a low PCA1.

Cluster 6 - In this cluster, there are customers with a medium PCA1 and a low PCA2 score.

Cluster 4 - This cluster comprises of customers with a high PCA1 income and a high PCA2.

Cluster 3 - This comprises of customers with a high PCA2 and a medium annual spend of income.

With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of customers, companies can release products and services that target customers based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.

Summary

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 analyzed and visualized the data and then proceeded to implement our algorithm. Hope you enjoyed this customer segmentation project of machine learning using R.