Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

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-means clustering** which is the essential algorithm for clustering unlabeled dataset. Before ahead in this project, learn what actually customer segmentation is.

customer_data=read.csv("/home/dataflair/Mall_Customers.csv")

str(customer_data)

names(customer_data)

```
customer_data=read.csv("/home/dataflair/Mall_Customers.csv")
str(customer_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 ...
```

```
names(customer_data)
```

```
## [1] "CustomerID"             "Gender"
## [3] "Age"                    "Annual.Income..k.."
## [5] "Spending.Score..1.100."
```

head(customer_data)

summary(customer_data$Age)

```
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)

summary(customer_data$Annual.Income..k..)

sd(customer_data$Annual.Income..k..)

summary(customer_data$Age)

```
sd(customer_data$Age)
```

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

```
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] 26.26472
```

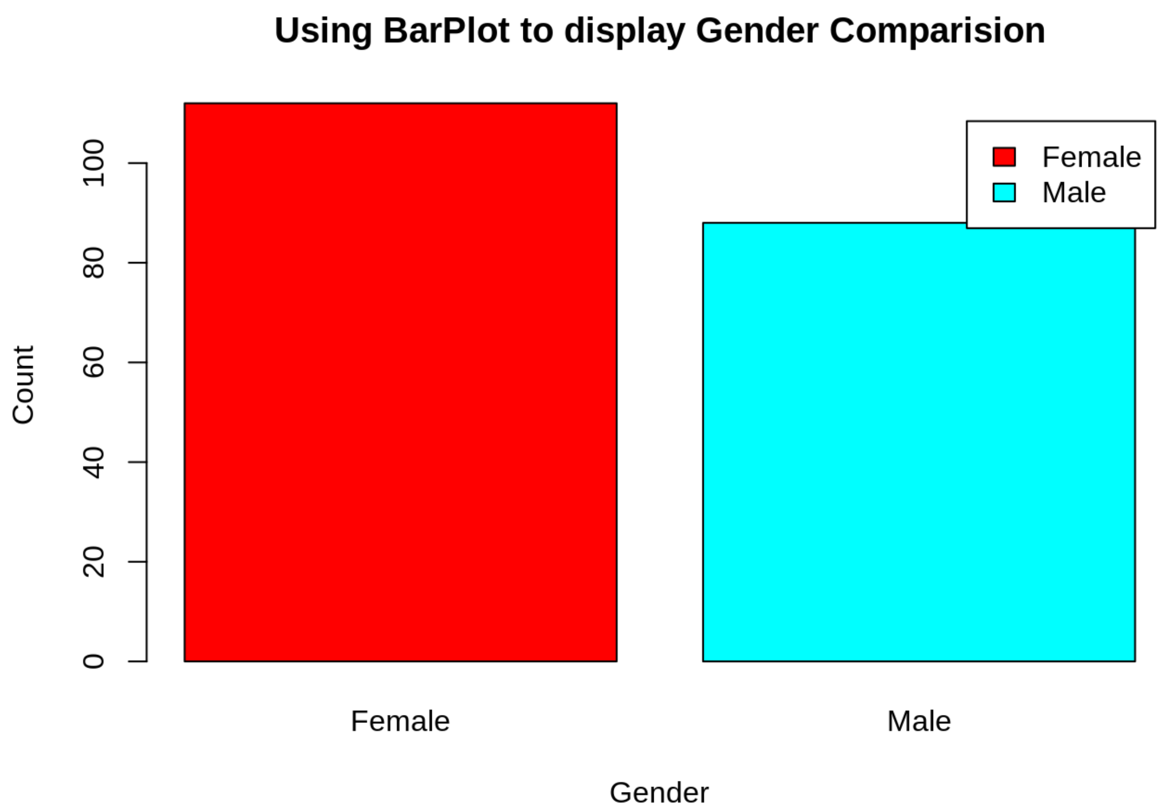
```
summary(customer_data$Age)
```

```
##    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.)
```

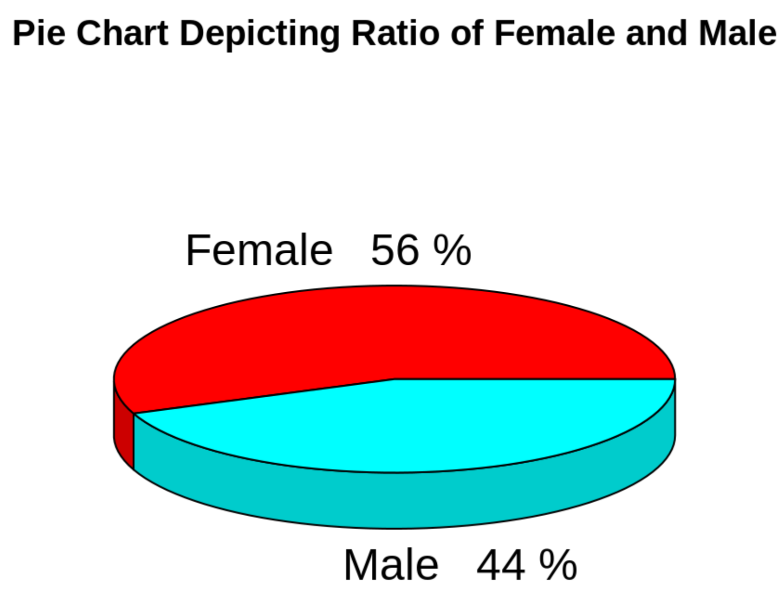
```
## [1] 25.82352
```

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



From the above barplot, we observe that the number of females is higher than the males. Now, let us visualize a pie chart to observe the ratio of male and female distribution.

```
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")
```

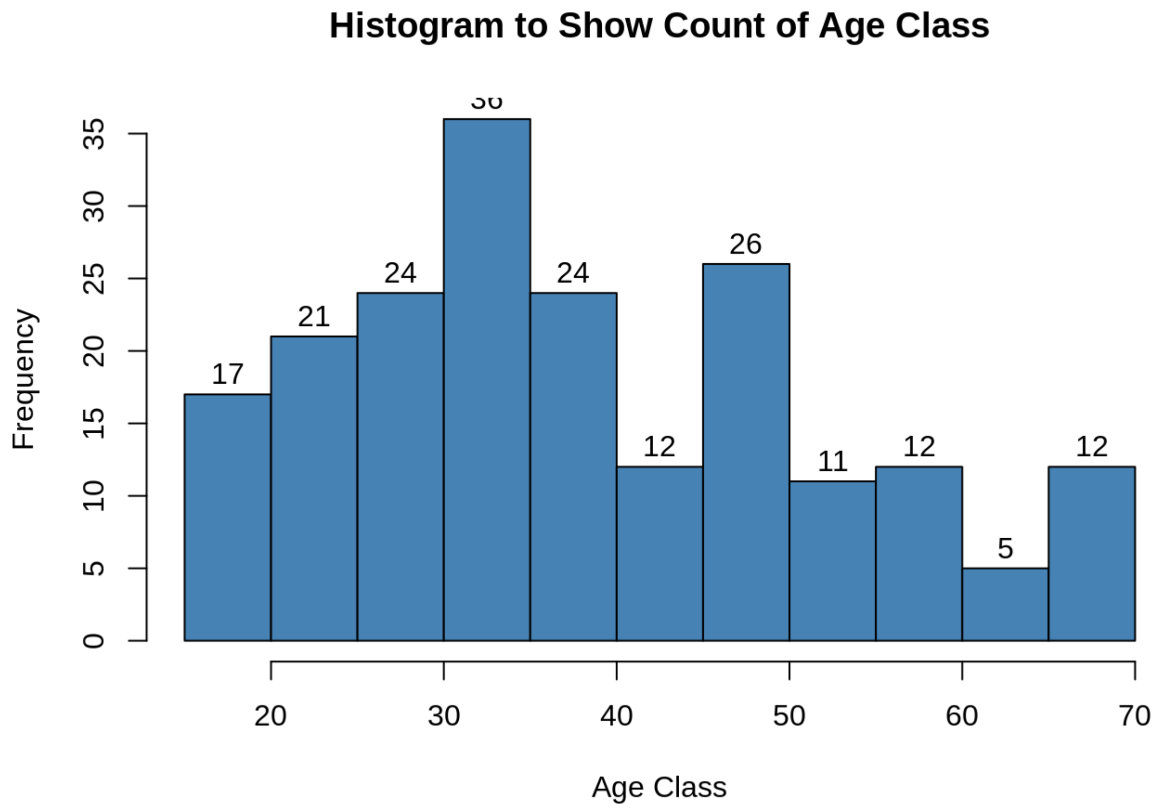


From the above graph, we conclude that the percentage of females is **56%**, whereas the percentage of male in the customer dataset is **44%**.

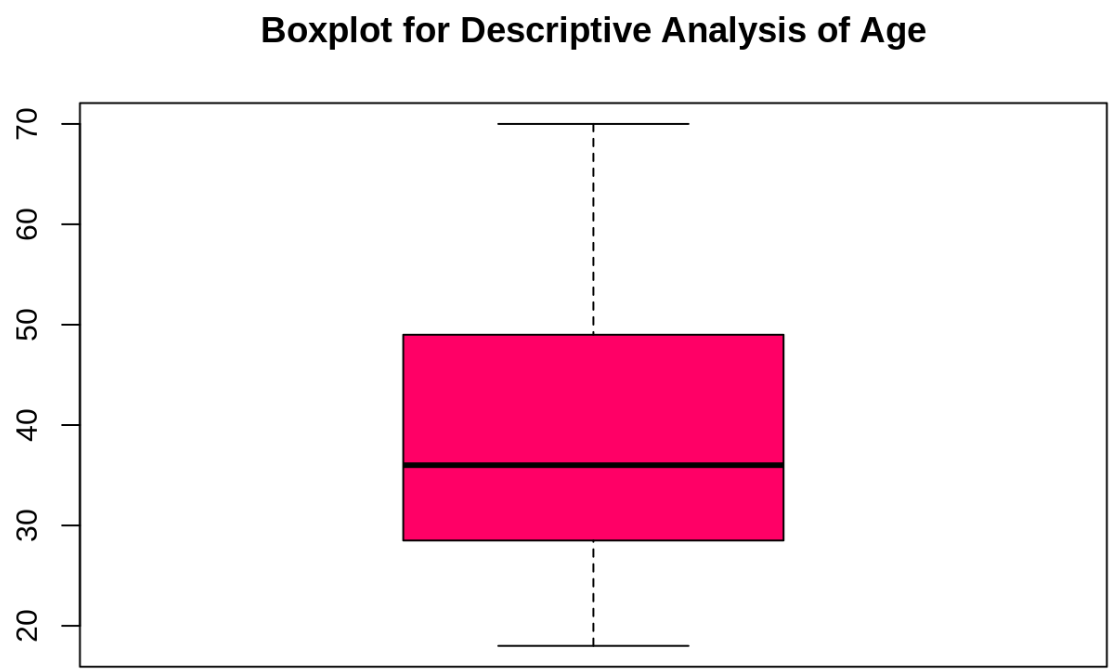
```
summary(customer_data$Age)
```

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

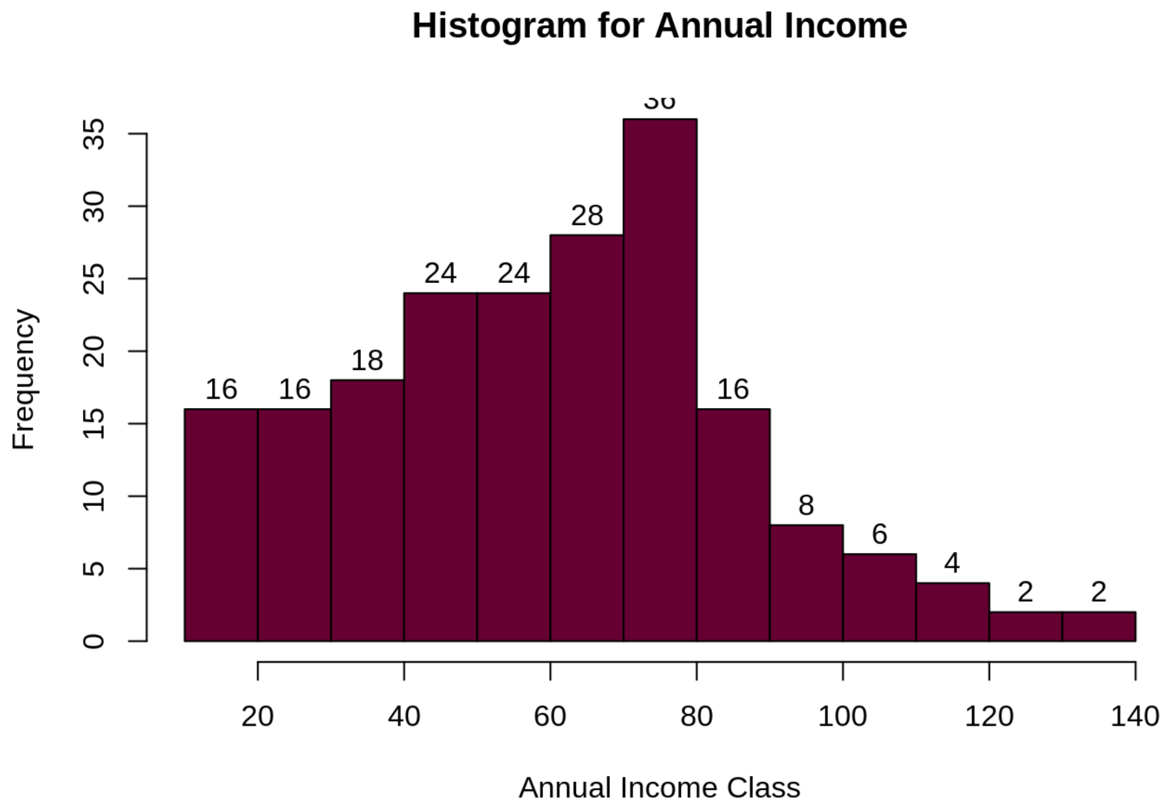
```
hist(customer_data$Age,
   col="blue",
   main="Histogram to Show Count of Age Class",
   xlab="Age Class",
   ylab="Frequency",
   labels=TRUE)
```



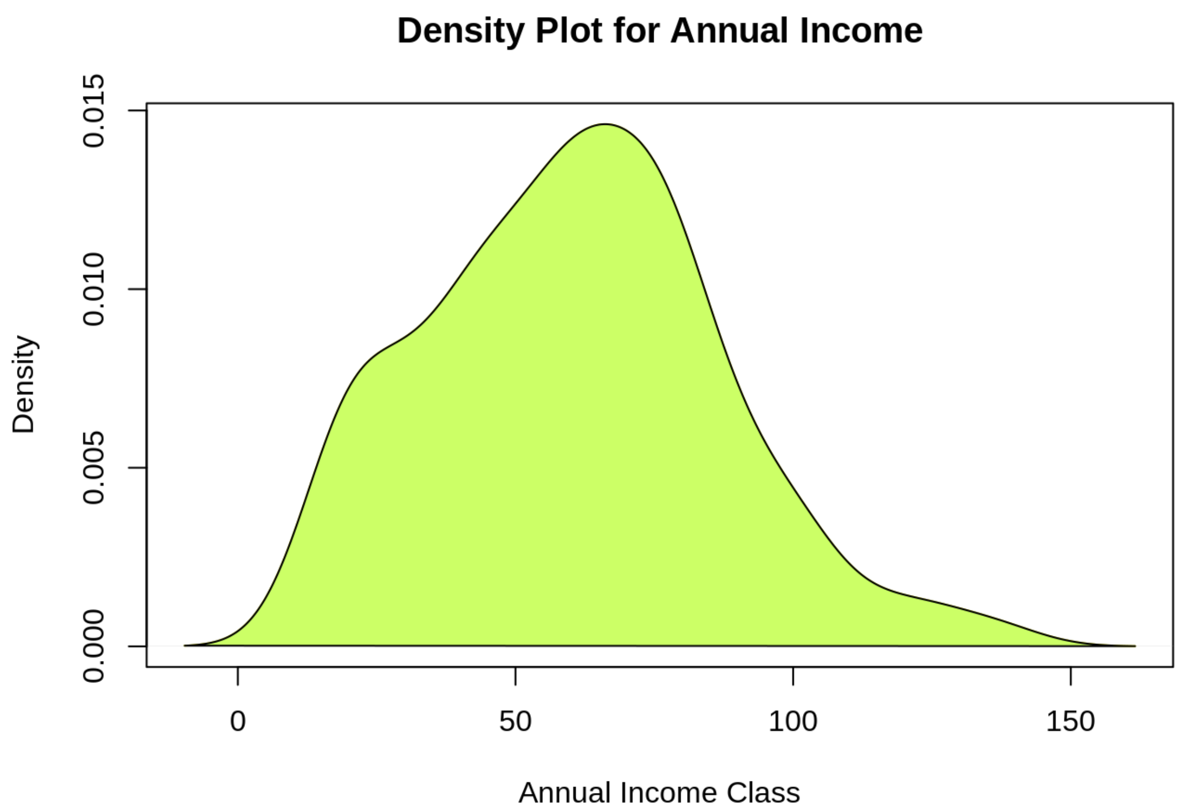

```
boxplot(customer_data$Age,
    col="ff0066",
    main="Boxplot for Descriptive Analysis of Age")
```




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



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***.

summary(customer_data$Spending.Score..1.100.)

Min. 1st Qu. Median Mean 3rd Qu. Max.

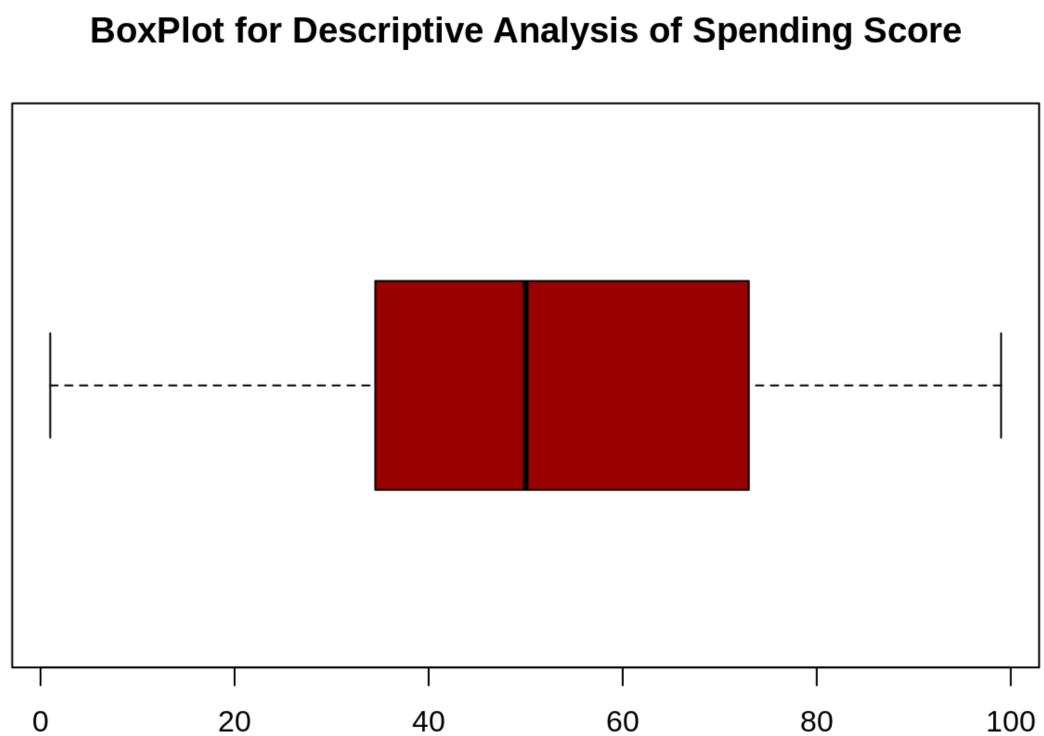
1.00 34.75 50.00 50.20 73.00 99.00

boxplot(customer_data$Spending.Score..1.100.,

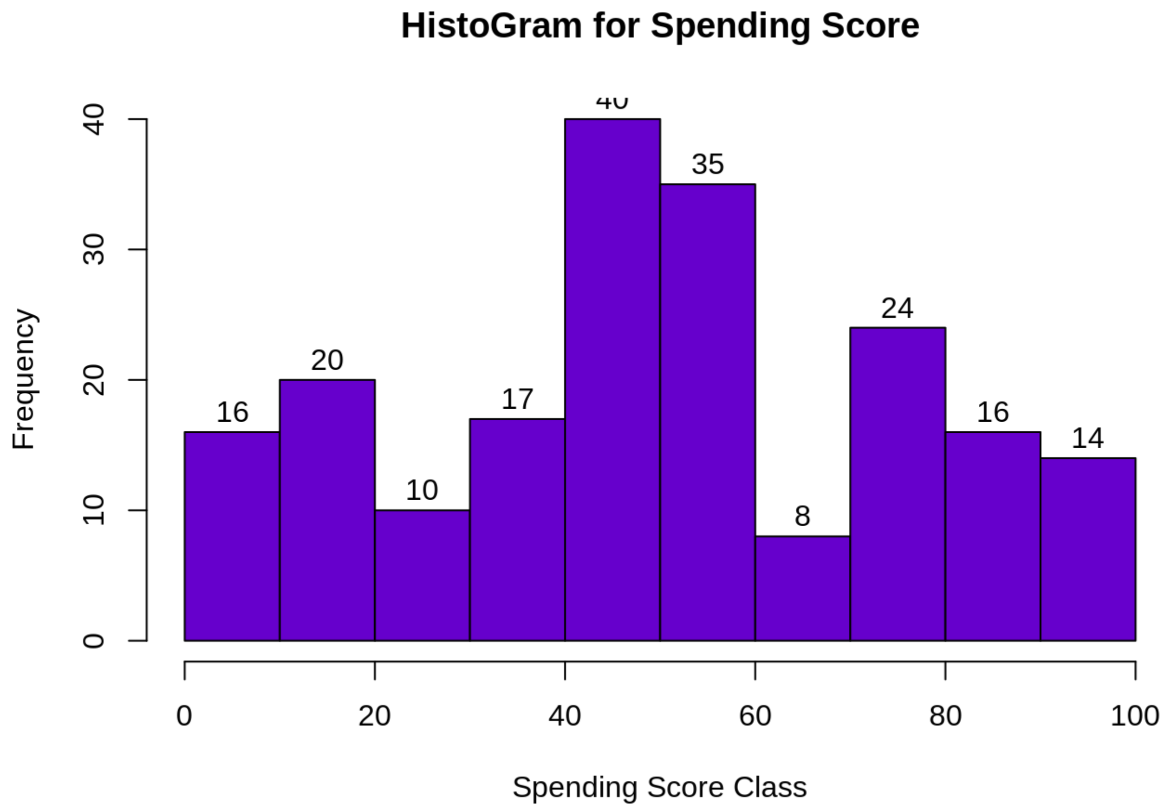
horizontal=TRUE,

col="#990000",

main="BoxPlot for Descriptive Analysis of Spending Score")



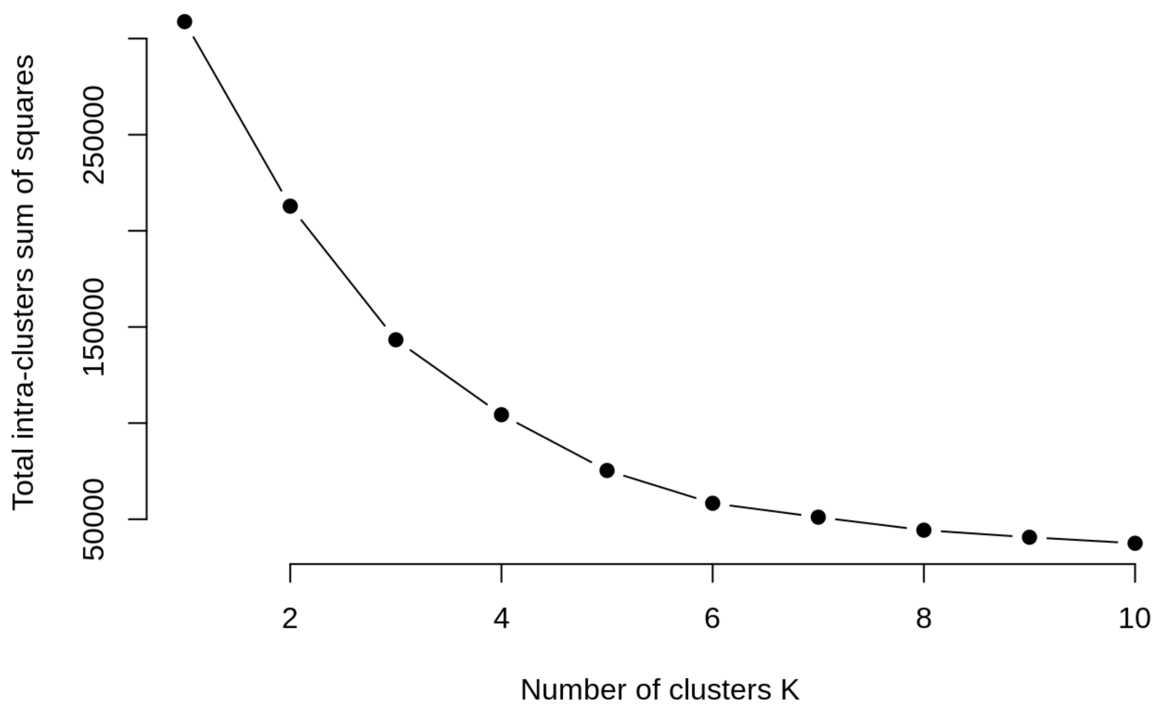

```
hist(customer_data$Spending.Score..1.100.,
    main="HistoGram for Spending Score",
    xlab="Spending Score Class",
    ylab="Frequency",
    col="#6600cc",
    labels=TRUE)
```



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.

```
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)
```

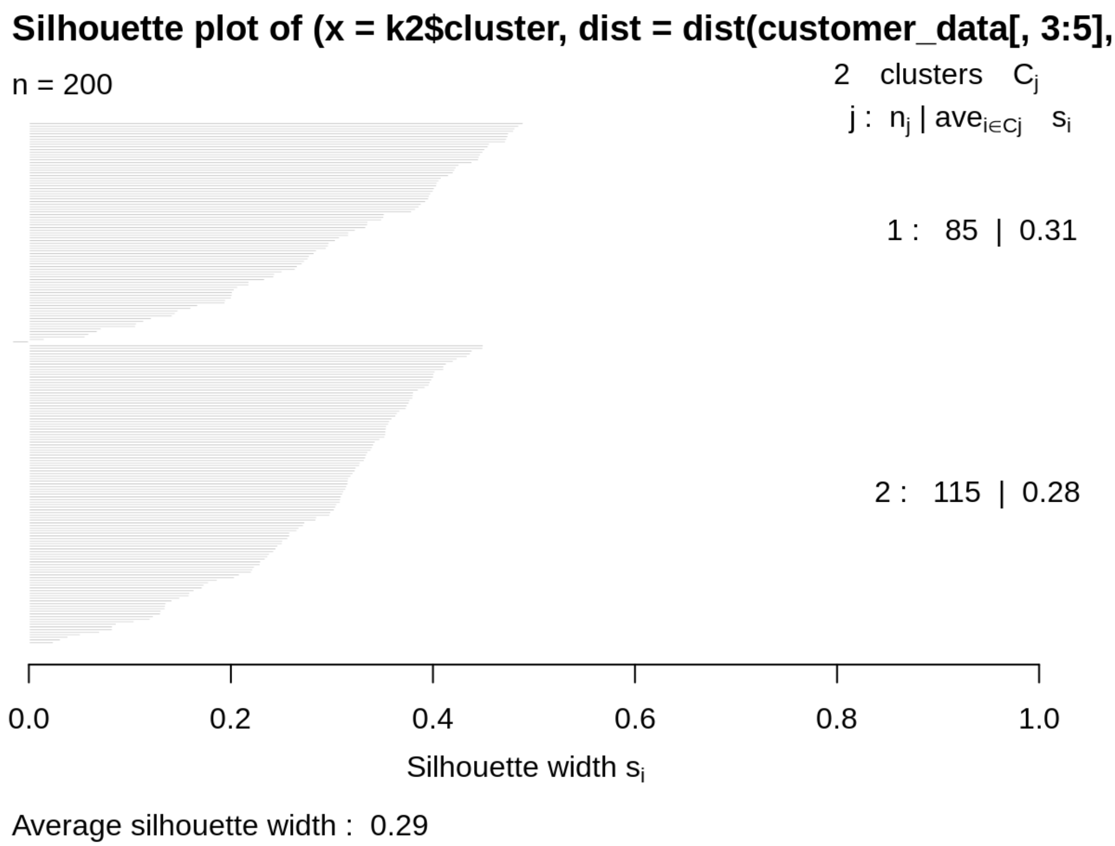
```
plot(k.values, iss_values,
   type="b", pch = 19, frame = FALSE,
   xlab="Number of clusters K",
   ylab="Total intra-clusters sum of squares")
```



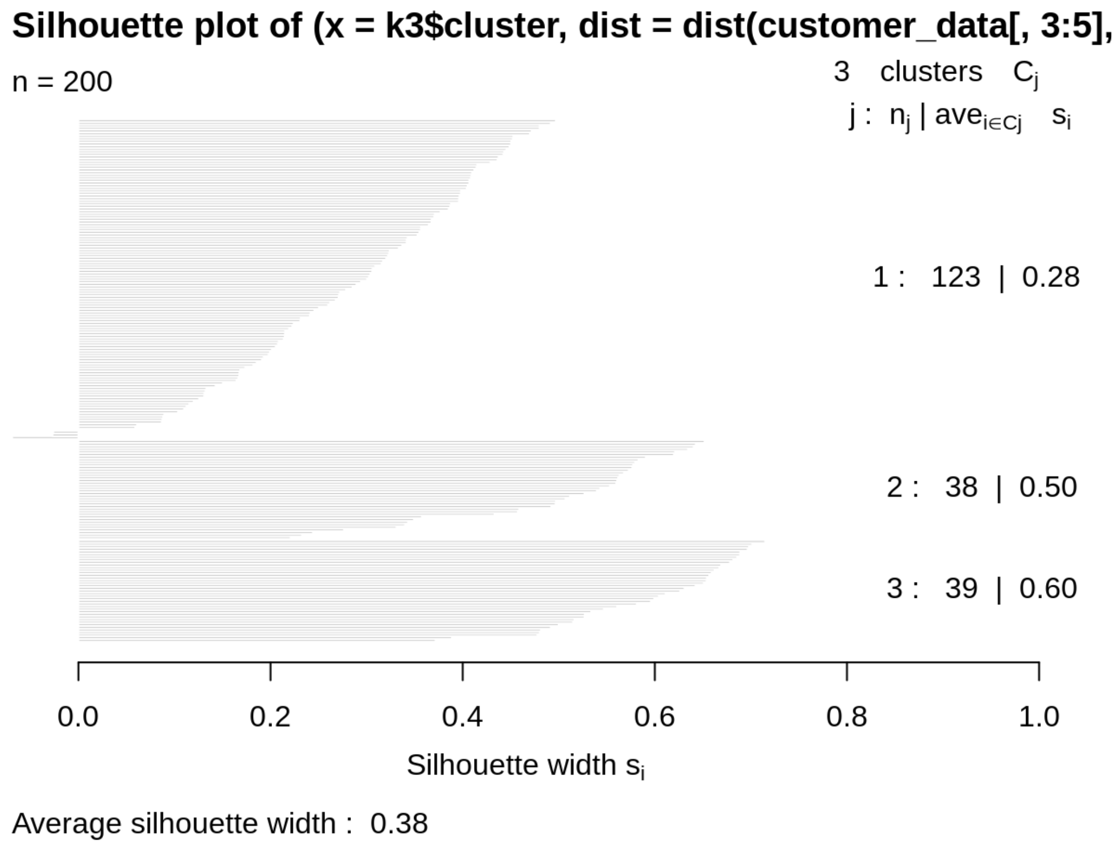
From the above graph, we conclude that 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

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

k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
```

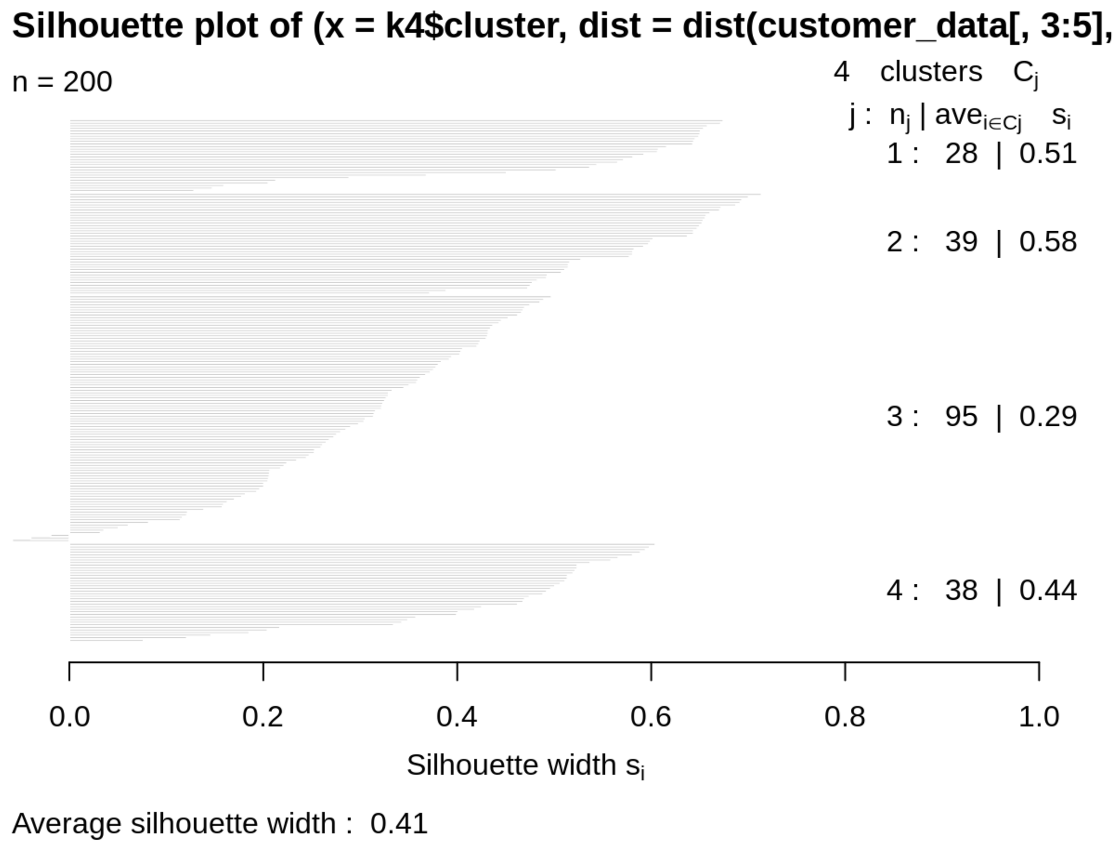
```
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))
```



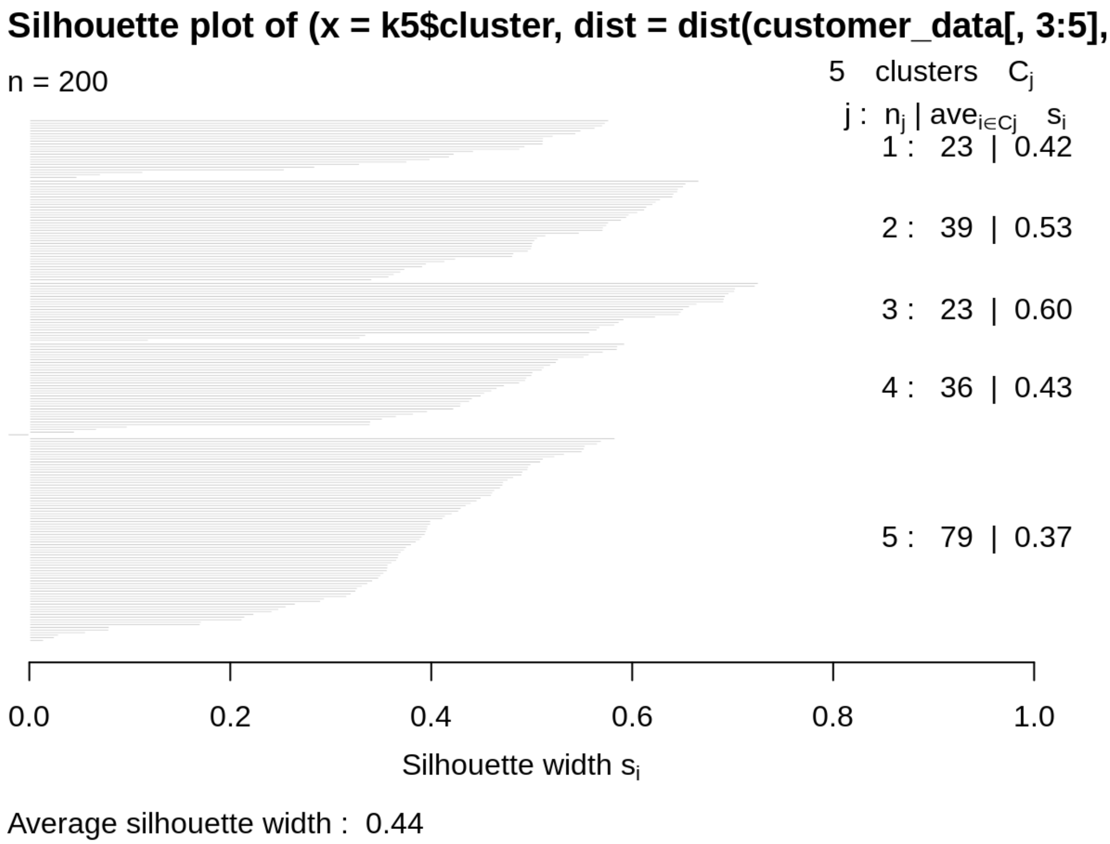

```
k3<-kmeans(customer_data[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")

s3<-plot(silhouette(k3$cluster,dist(customer_data[,3:5],"euclidean")))
```



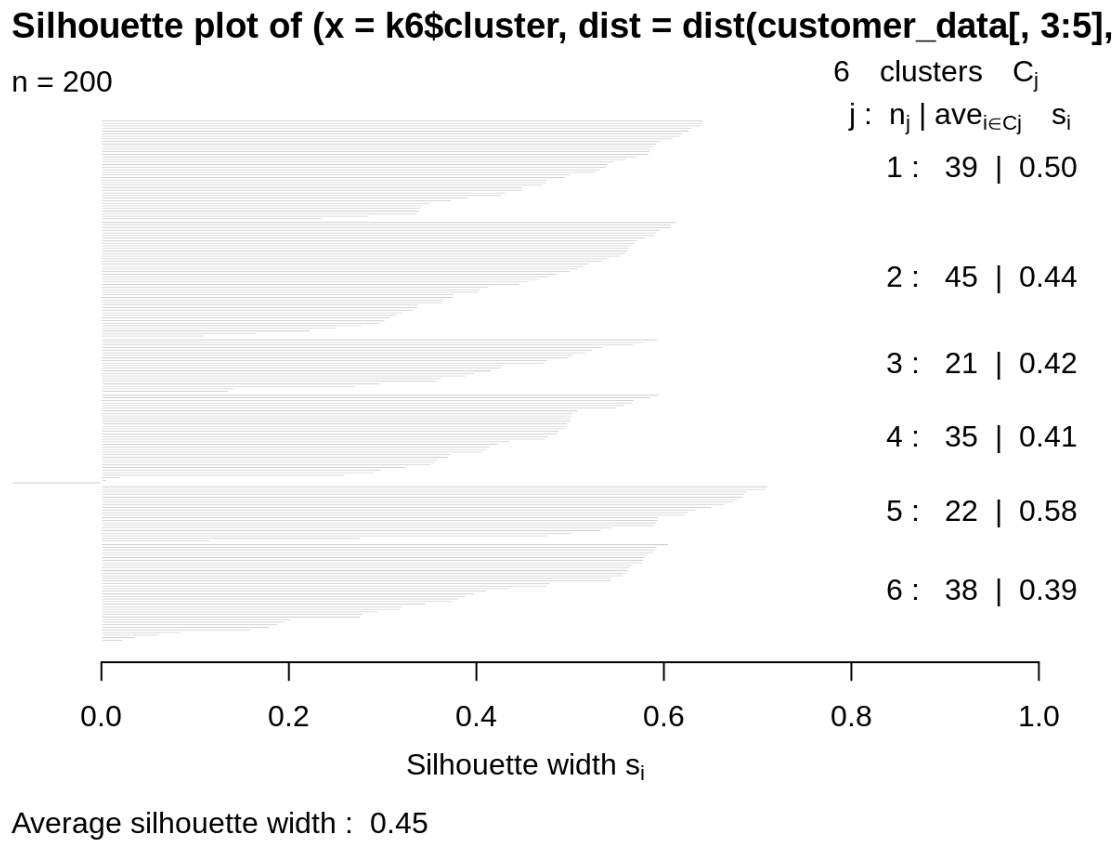

```
k4<-kmeans(customer_data[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")

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")))
```

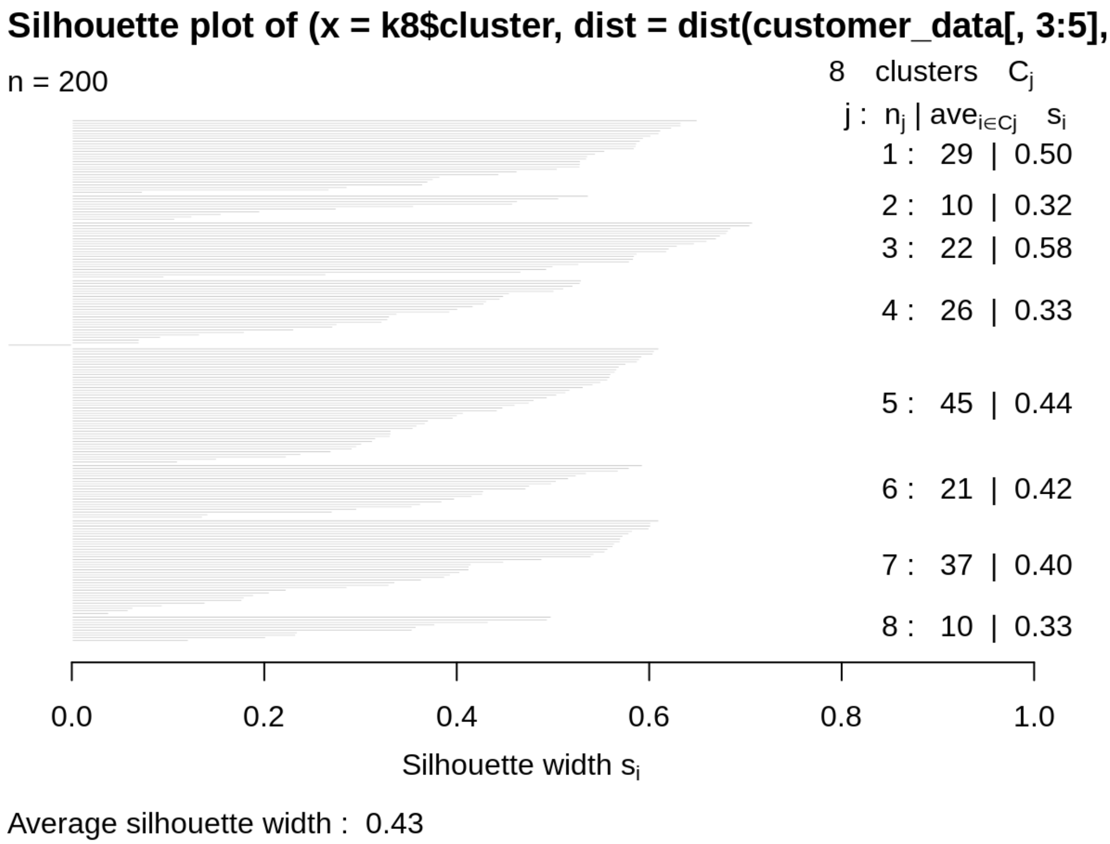



```
k6<-kmeans(customer_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")

s6<-plot(silhouette(k6$cluster,dist(customer_data[,3:5],"euclidean")))
```

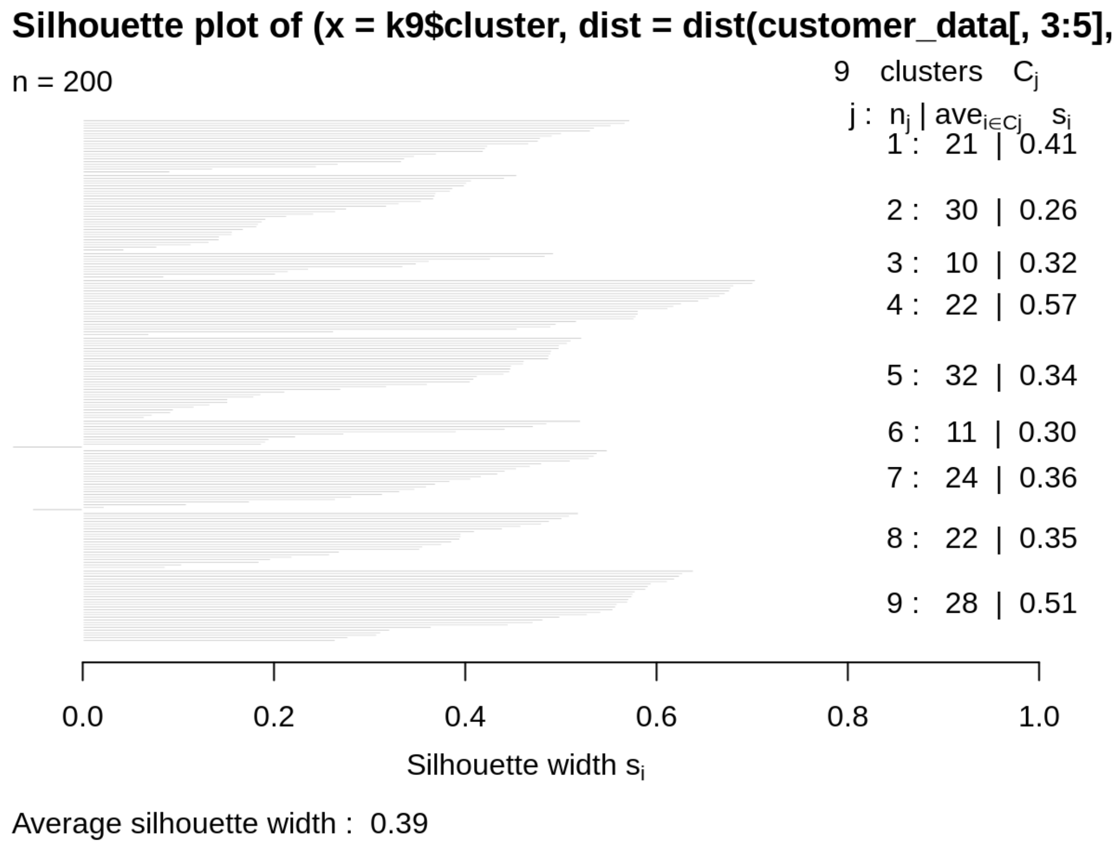



```
k7<-kmeans(customer_data[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")

s7<-plot(silhouette(k7$cluster,dist(customer_data[,3:5],"euclidean")))
```

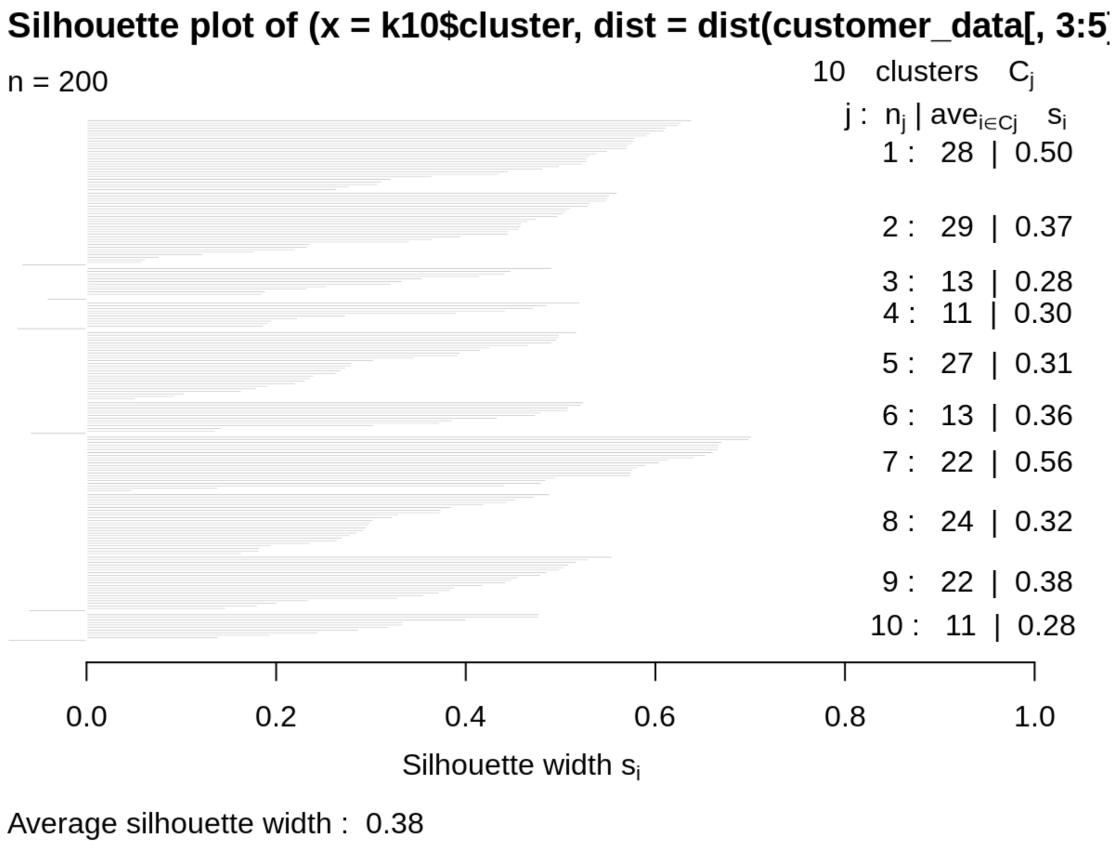

```
k8<-kmeans(customer_data[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")

s8<-plot(silhouette(k8$cluster,dist(customer_data[,3:5],"euclidean")))
```



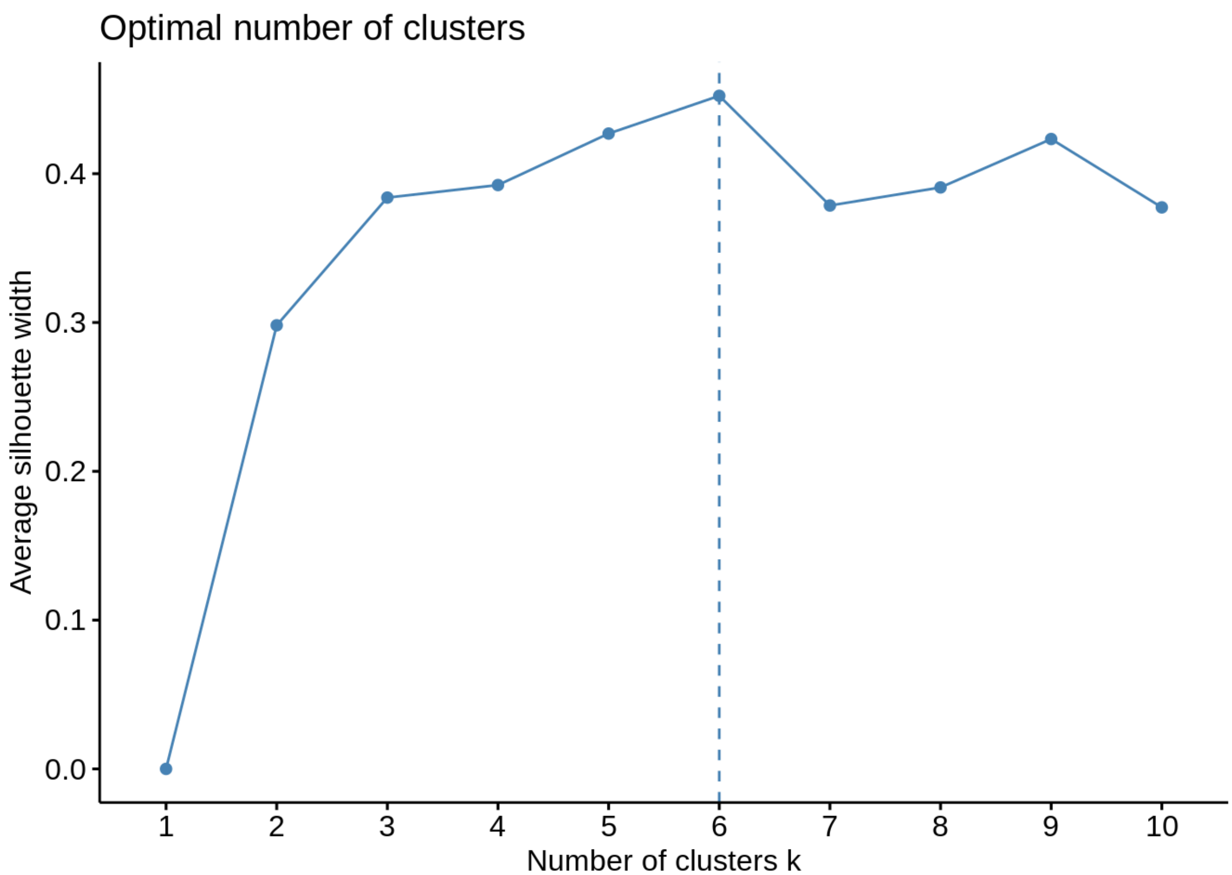

```
k9<-kmeans(customer_data[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")

s9<-plot(silhouette(k9$cluster,dist(customer_data[,3:5],"euclidean")))
```



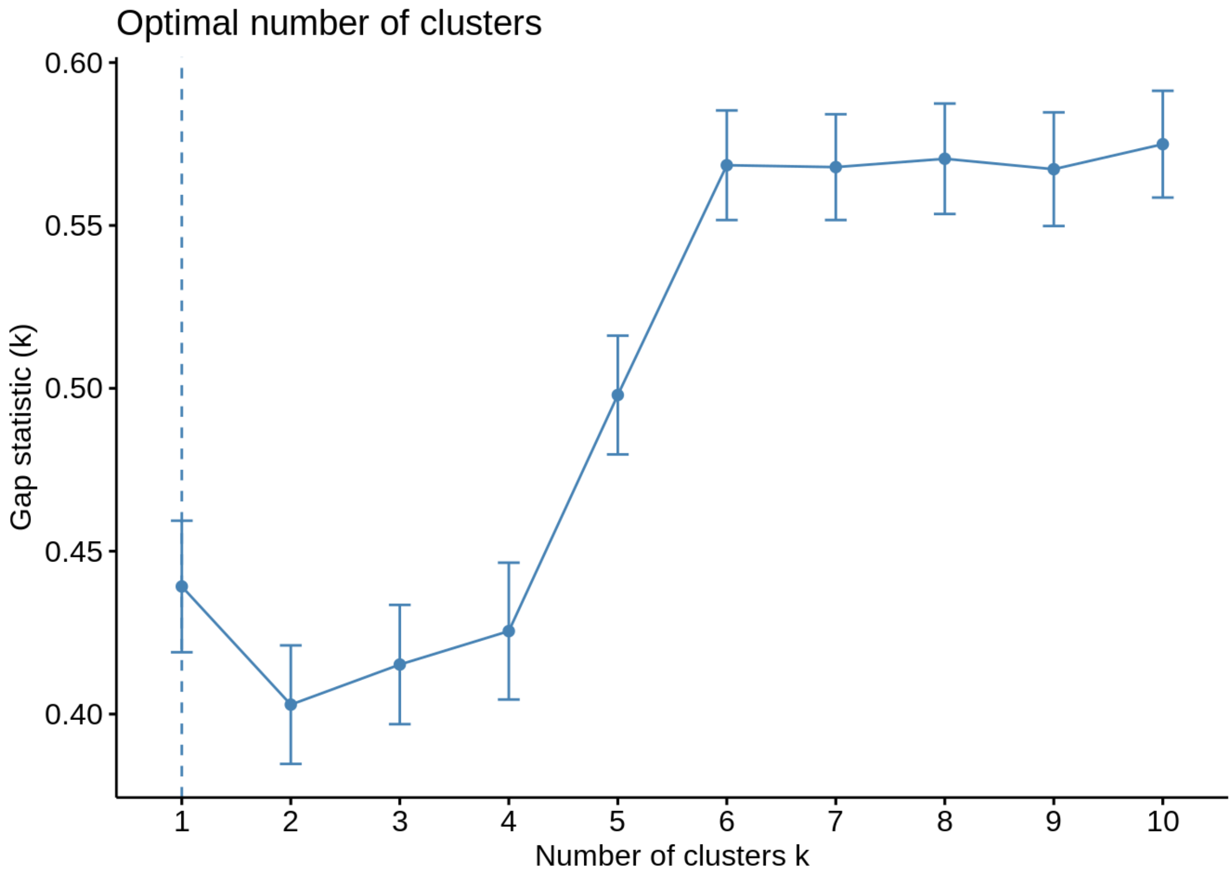

```
k10<-kmeans(customer_data[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")

s10<-plot(silhouette(k10$cluster,dist(customer_data[,3:5],"euclidean")))
```




```
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

```
## K-means clustering with 6 clusters of sizes 45, 22, 21, 38, 35, 39
##
## Cluster means:
##        Age Annual.Income..k.. Spending.Score..1.100.
## 1 56.15556           53.37778               49.08889
## 2 25.27273           25.72727               79.36364
## 3 44.14286           25.14286               19.52381
## 4 27.00000           56.65789               49.13158
## 5 41.68571           88.22857               17.28571
## 6 32.69231           86.53846               82.12821
##
## Clustering vector:
##   [1] 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3
##  [36] 2 3 2 3 2 1 2 1 4 3 2 1 4 4 4 1 4 4 1 1 1 1 1 4 1 1 4 1 1 1 4 1 1 4 4
##  [71] 1 1 1 1 1 4 1 4 4 1 1 4 1 1 4 1 1 4 4 1 1 4 1 4 4 4 1 4 1 4 4 1 1 4 1
```

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

summary(pcclust)

pcclust$rotation[,1:2]

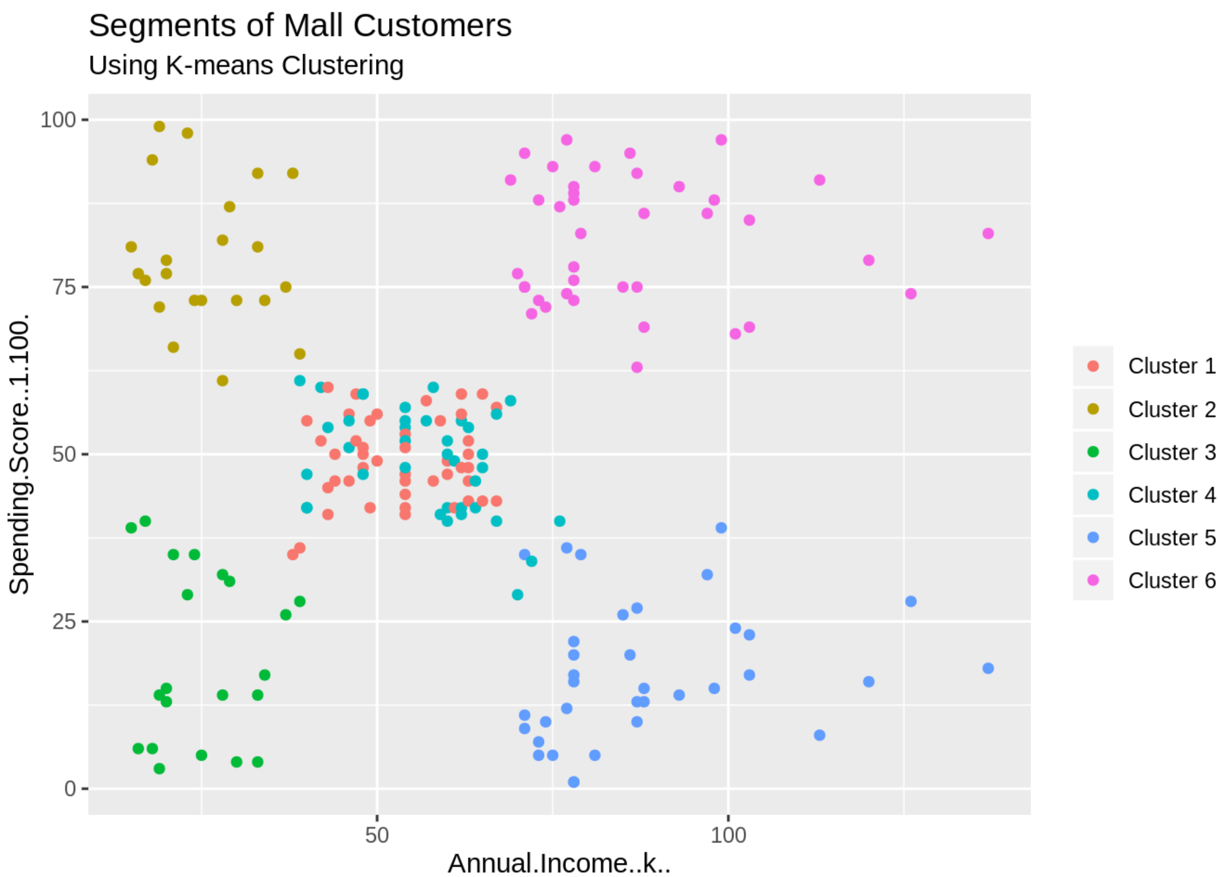
```
pcclust=prcomp(customer_data[,3:5],scale=FALSE)  #principal component analysis
summary(pcclust)
```

```
## 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
```

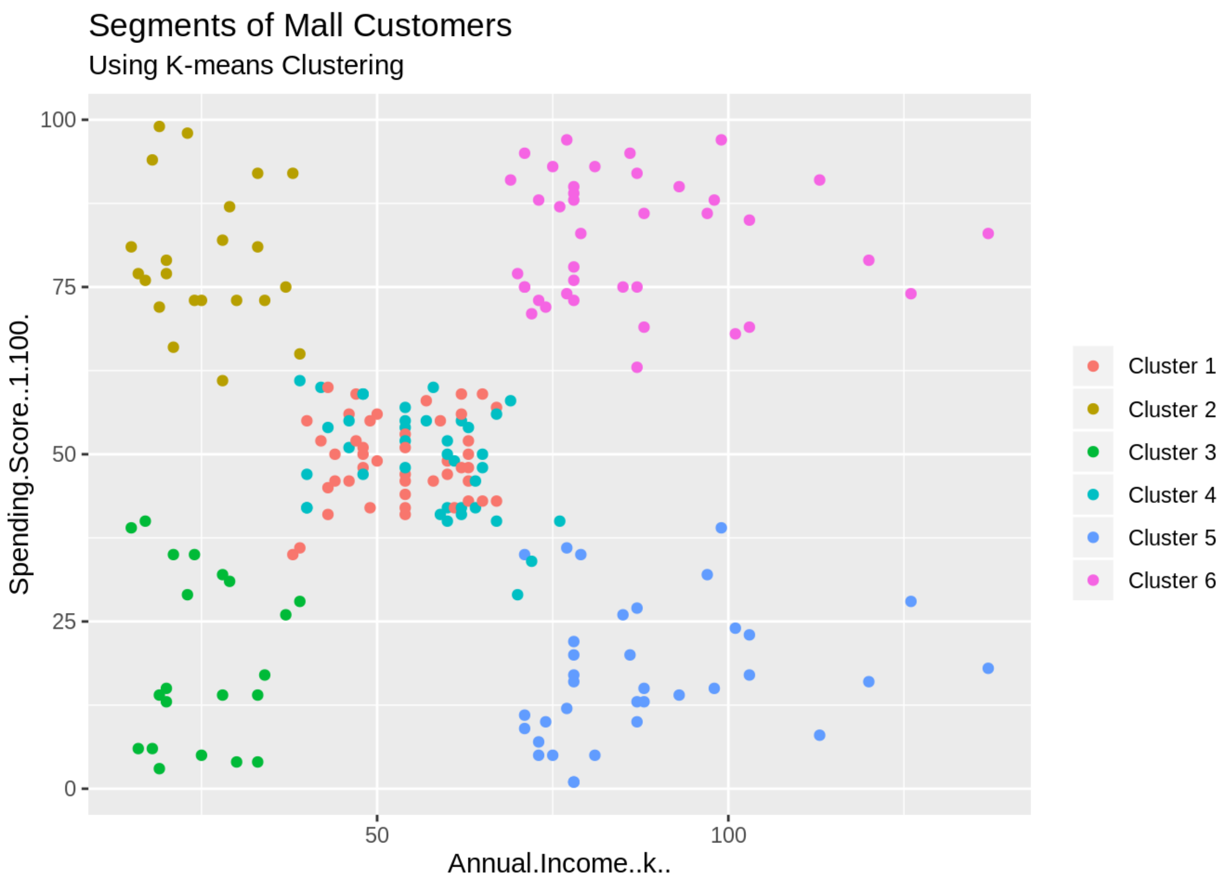
```
pcclust$rotation[,1:2]
```

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

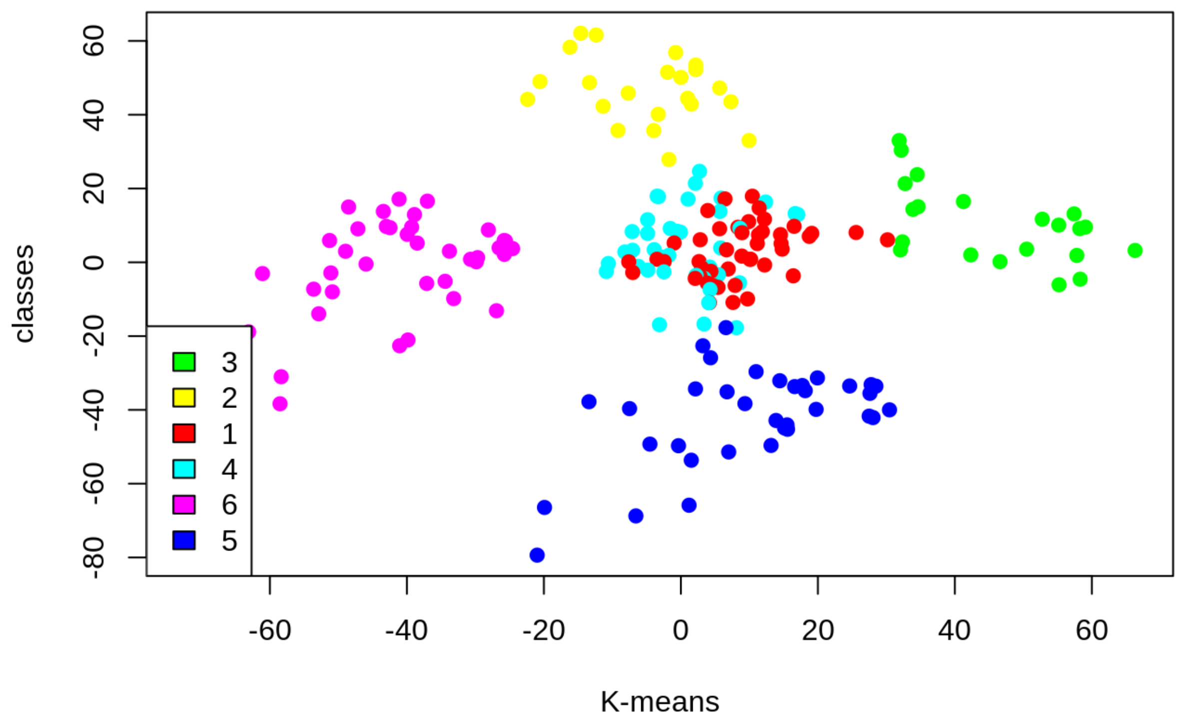
```
set.seed(1)

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)))
```



In this data science project, the customer segmentation model was described using a class of machine learning known as unsupervised learning. Specifically, clustering algorithm called K-means clustering. Analyzed and visualized the data.