Customer Segmentation using Machine Learning in R

1. ***Import dataset:***

Code:

```
>Mall_Customers <- read_csv("Mall_Customers.csv",
+       col_types = cols(Age = col_integer(),
+       Annual.Income.k = col_integer(),
+       CustomerID = col_integer(), Gender = col_factor(levels = c()),
+       Spending.Score.1..100 = col_integer()))
```

#structure dataset

Code:

```
>str(Mall_Customers)
```

```
Console   Terminal   Jobs
C:/Users/Valued Customer/Desktop/Testing/customer-segmentation-dataset/
> str(Mall_Customers)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':
        200 obs. of  5 variables:
 $ CustomerID           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender               : Factor w/ 2 levels "Male","Femal
e": 1 1 2 2 2 2 2 2 1 2 ...
 $ Age                  : int  19 21 20 23 31 22 35 23 64 3
0 ...
 $ Annual.Income.k      : int  15 15 16 16 17 17 18 18 19 1
9 ...
 $ Spending.Score.1..100: int  39 81 6 77 40 76 6 94 3 72
 ...
 - attr(*, "spec")=
  .. cols(
  ..   CustomerID = col_integer(),
  ..   Gender = col_factor(levels = NULL, ordered = FALSE,
 include_na = FALSE),
  ..   Age = col_integer(),
  ..   Annual.Income.k = col_integer(),
  ..   Spending.Score.1..100 = col_integer()
  .. )
>
```

```
>names(Mall_Customers)
```

```
> names(Mall_Customers)
[1] "CustomerID"              "Gender"
[3] "Age"                     "Annual.Income.k"
[5] "Spending.Score.1..100"
>
```

```
>head(Mall_Customers)
```

```
> head(Mall_Customers)
# A tibble: 6 x 5
  CustomerID Gender   Age Annual.Income.k
       <int> <fct>  <int>           <int>
1          1 Male      19              15
2          2 Male      21              15
3          3 Female    20              16
4          4 Female    23              16
5          5 Female    31              17
6          6 Female    22              17
# … with 1 more variable: Spending.Score.1..100 <int>
```

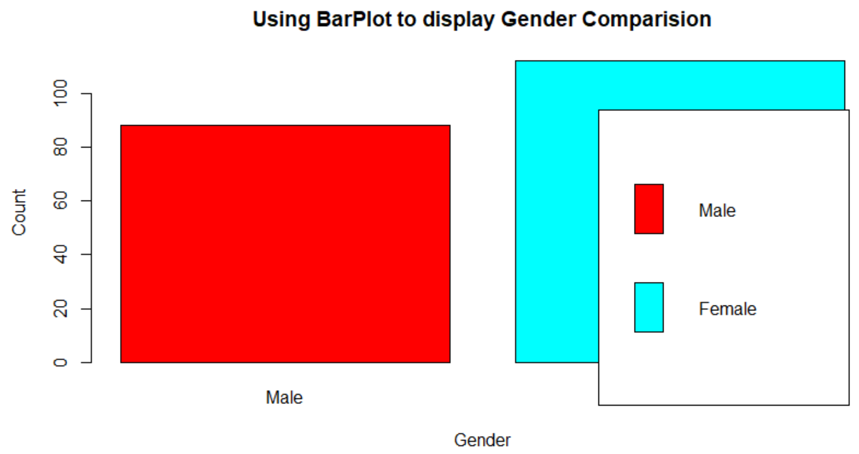
2. Analysis Dataset

2.1 Visualization Customer Gender Visualization

#Graphic

Code:

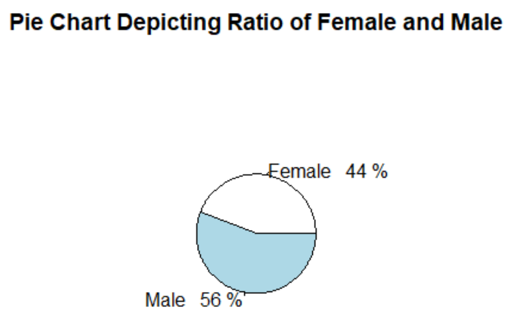
```
> a=table(Mall_Customers$Gender)
> barplot(a,main="Using BarPlot to display Gender Comparision",
+         ylab="Count",
+         xlab="Gender",
+         col=rainbow(2),
+         legend=rownames(a))
```



 # pie chart to observe the ratio of male and female distribution.

Code:

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



2.2 Visualization of Age Distribution

Code:

```
>summary(Mall_Customers$Age)
```

```
> summary(Mall_Customers$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   28.75   36.00   38.85   49.00   70.00 
> 
```

 # standard deviation

Code:

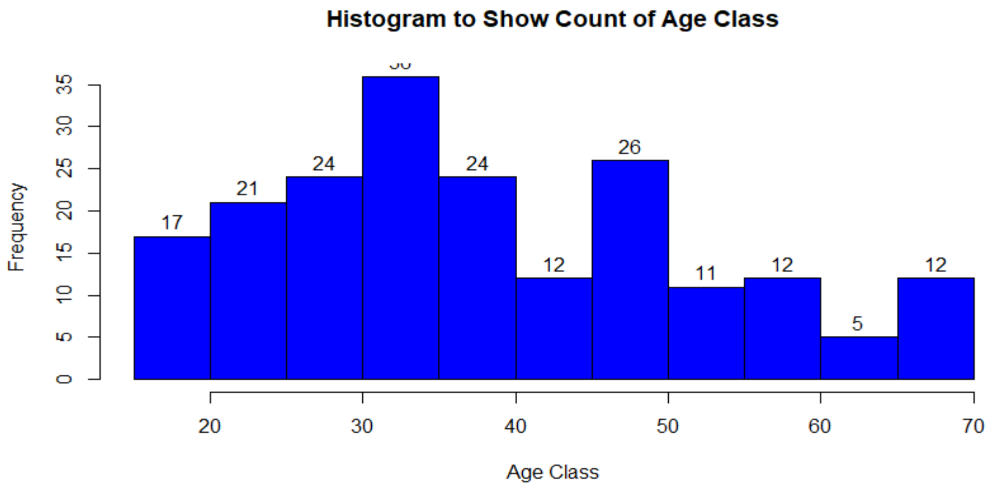
```
> sd(Mall_Customers$Age)
```

```
> sd(Mall_Customers$Age)
[1] 13.96901
```

 # graphic

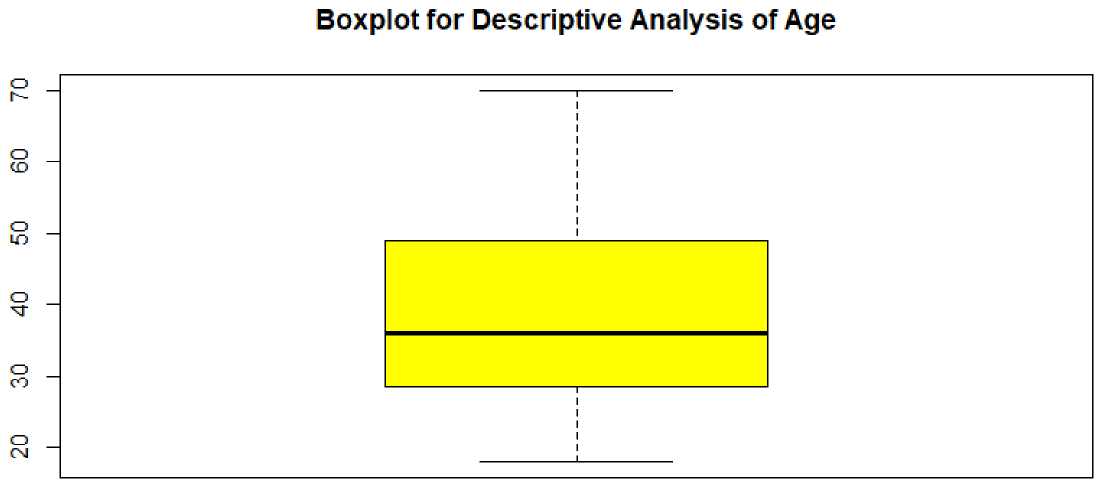
Code:

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



Code:

```
>boxplot(Mall_Customers$Age,
+         col="yellow",
+         main="Boxplot for Descriptive Analysis of Age")
```



According to the graphs, we can see the maximum customer ages are between 30 and 35. The minimum age of customers is 18, whereas, the maximum age is 70.

2.3Visualization Analysis of the Annual Income of the Customers

Code:

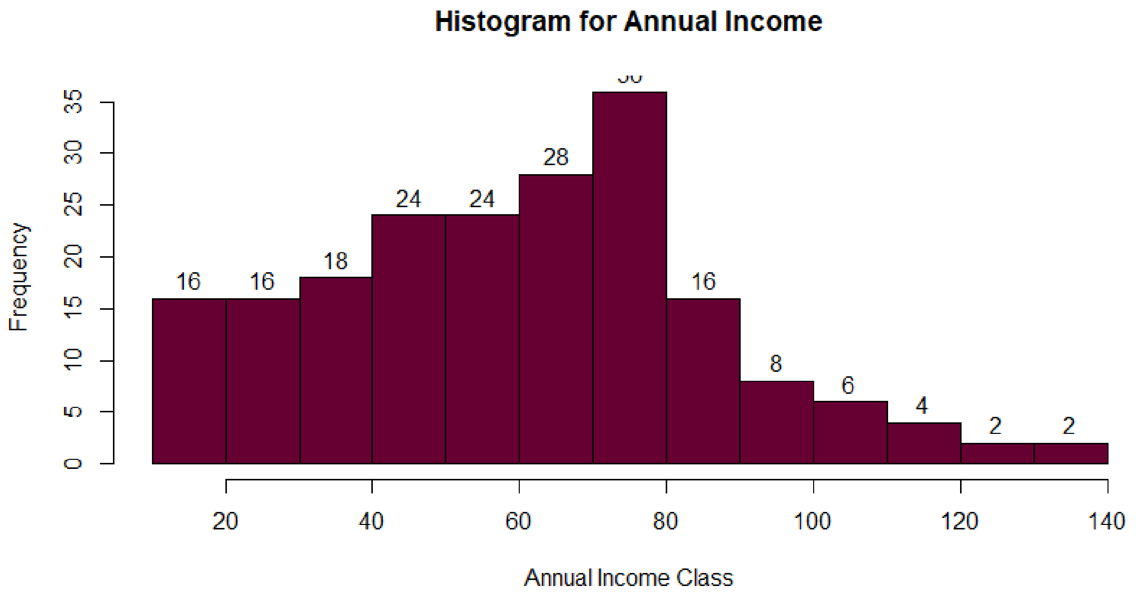
```
>summary(Mall_Customers$Annual.Income.k)
```

```
> summary(Mall_Customers$Annual.Income.k)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.00   41.50   61.50   60.56   78.00  137.00 
```

 # Graphic

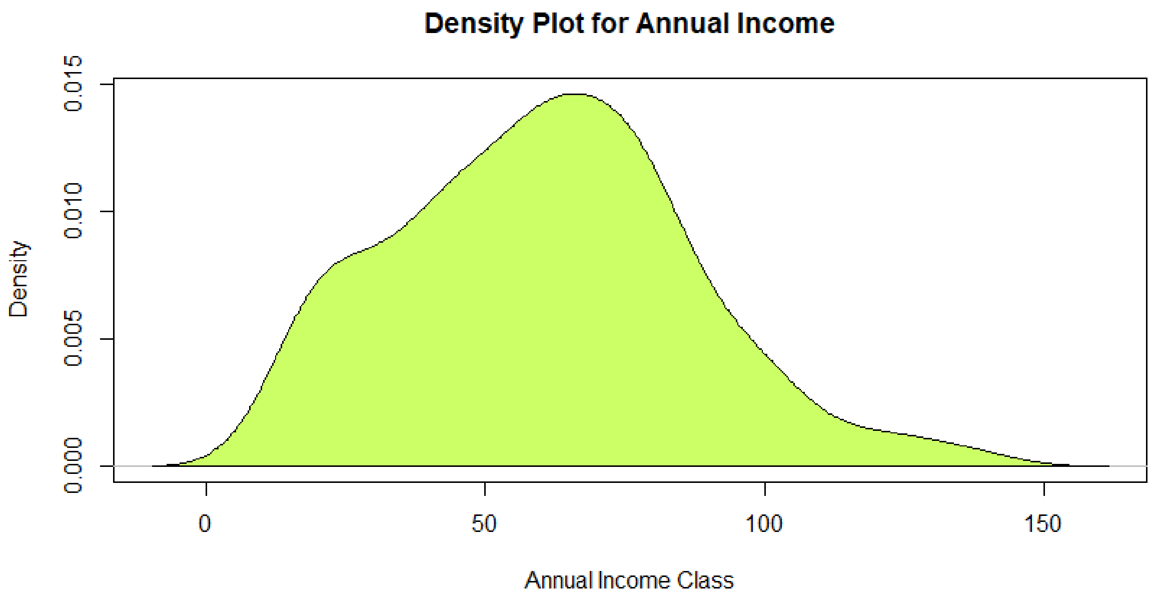
Code:

```
>hist(Mall_Customers$Annual.Income.k,
+       col="#660033",
+       main="Histogram for Annual Income",
+       xlab="Annual Income Class",
+       ylab="Frequency",
+       labels=TRUE)
```



Code:

```
>plot(density(Mall_Customers$Annual.Income.k),
+      col="green",
+      main="Density Plot for Annual Income",
+      xlab="Annual Income Class",
+      ylab="Density")
> polygon(density(Mall_Customers$Annual.Income.k),
+         col="#ccff66")
```



The graphics shows that the minimum annual income of the customers is 15 and the maximum 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.

2.4 Visualization Analyzing Spending Score of the Customers

Code:

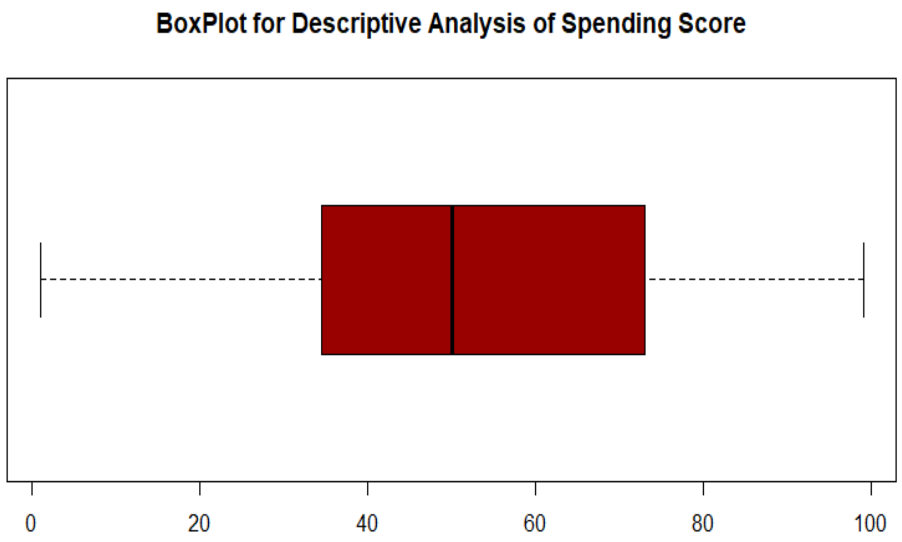
```
>summary(Mall_Customers$Spending.Score.1..100)
```

```
> summary(Mall_Customers$Spending.Score.1..100)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   34.75   50.00   50.20   73.00   99.00 
```

 # Graphic

Code:

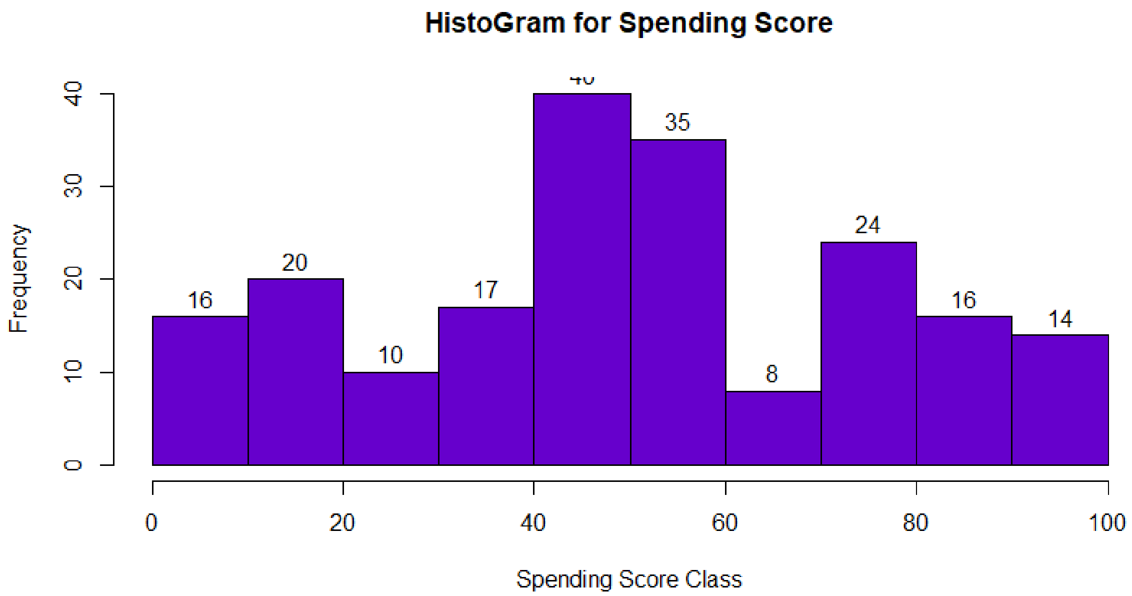
```
> boxplot(Mall_Customers$Spending.Score.1..100,
+         horizontal=TRUE,
+         col="#990000",
+         main="BoxPlot for Descriptive Analysis of Spending Score")
```



Code:

```
> hist(Mall_Customers$Spending.Score.1..100,
+      main="HistoGram for Spending Score",
+      xlab="Spending Score Class",
+      ylab="Frequency",
```

```
+       col="#6600cc",
+       labels=TRUE)
```



We can see from the graphs that the minimum spending score is 1, maximum is 99 and the average is 50.20. From the histogram, we conclude that customers between class 40 and 50 have the highest spending score among all the classes.

3. K-means Algorithm

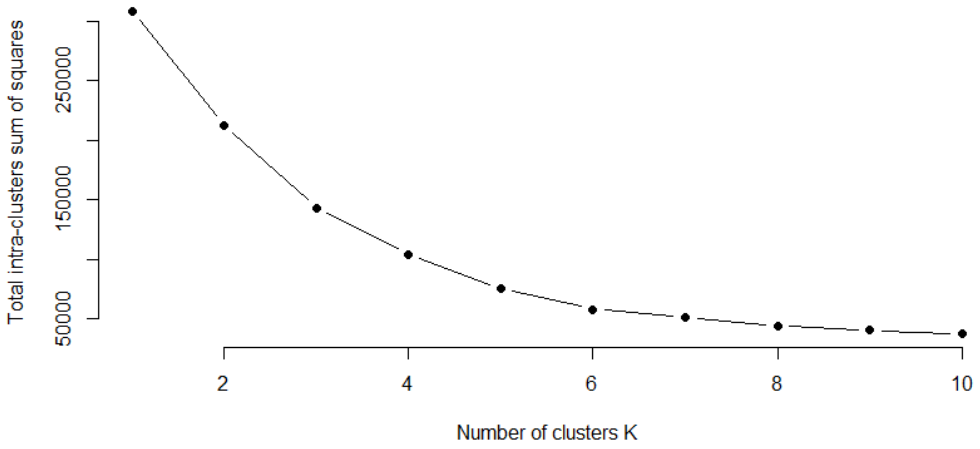
3.1 Elbow Method

Code:

```
> library(purrr)
> set.seed(123)
```

 # function to calculate total intra-cluster sum of square

```
> iss <- function(k) {
+     kmeans(Mall_Customers[,3:5],k,iter.max=100,nstart=100,algorithm="Llo
yd" )$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")
```



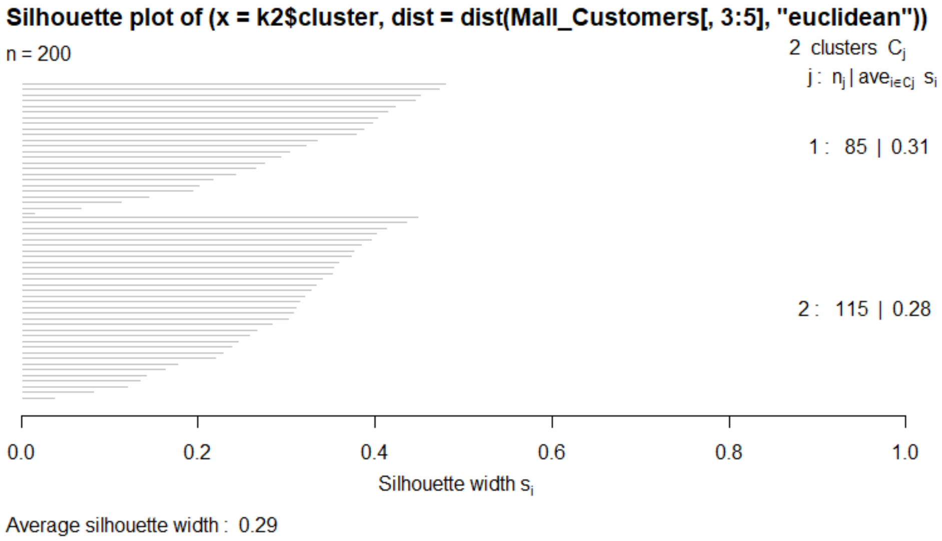
According to the graph, 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

3.2 Average Silhouette Method

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

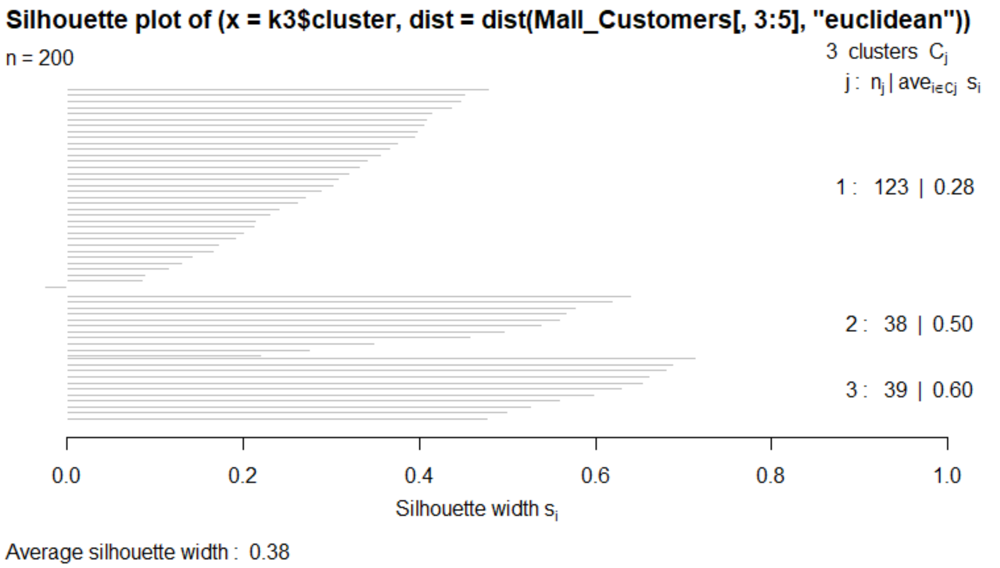
Code:

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



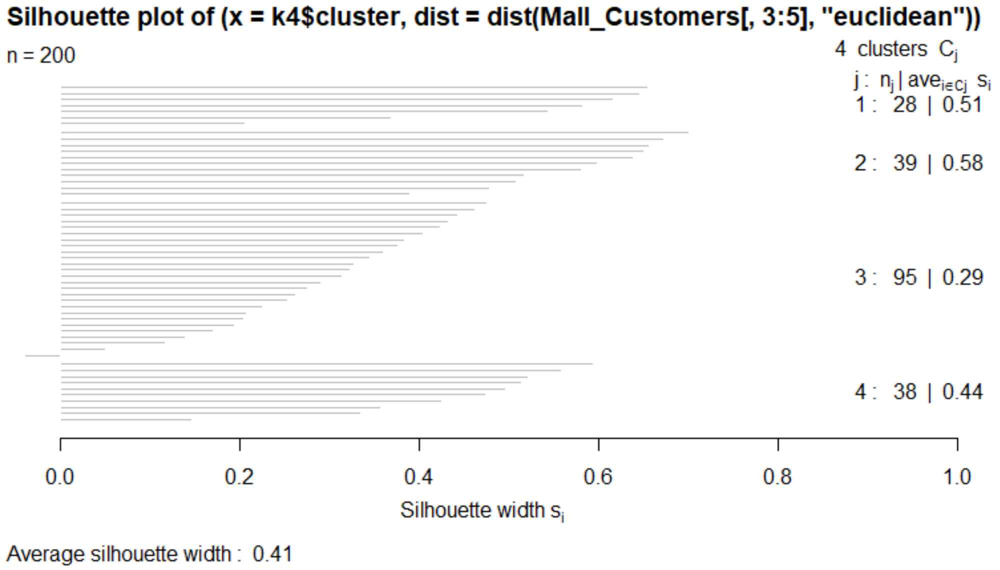
Code:

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



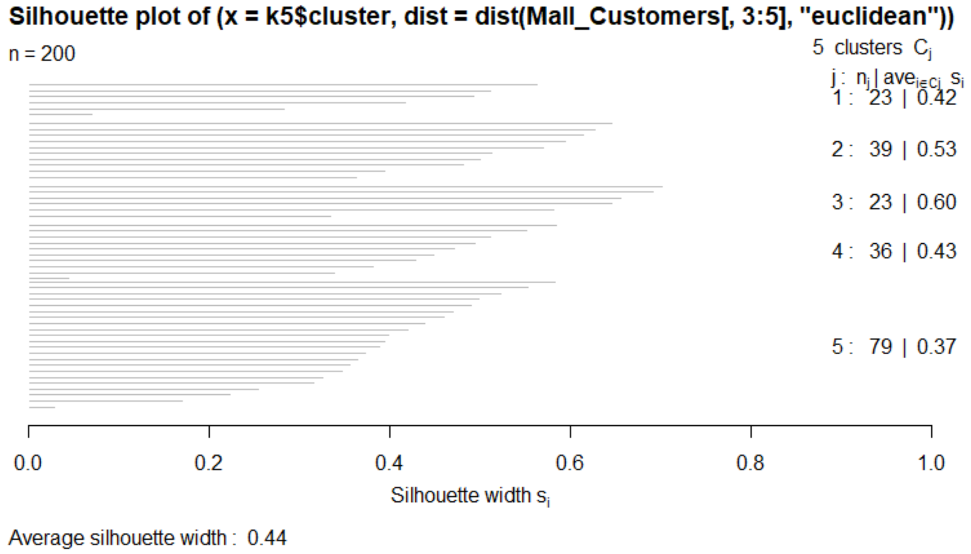
Code:

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



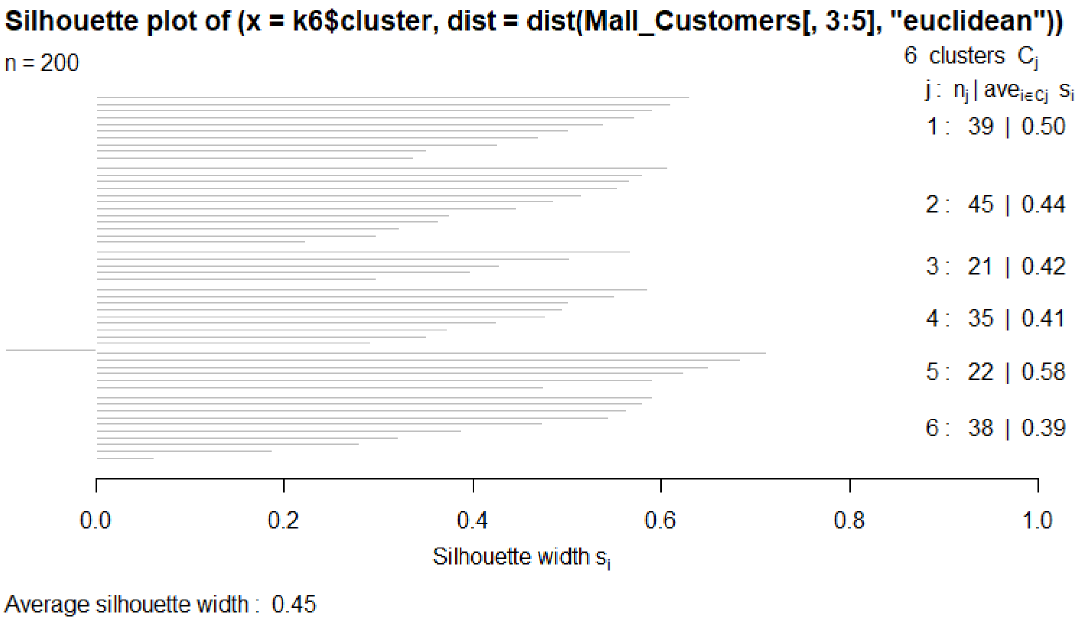
Code:

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



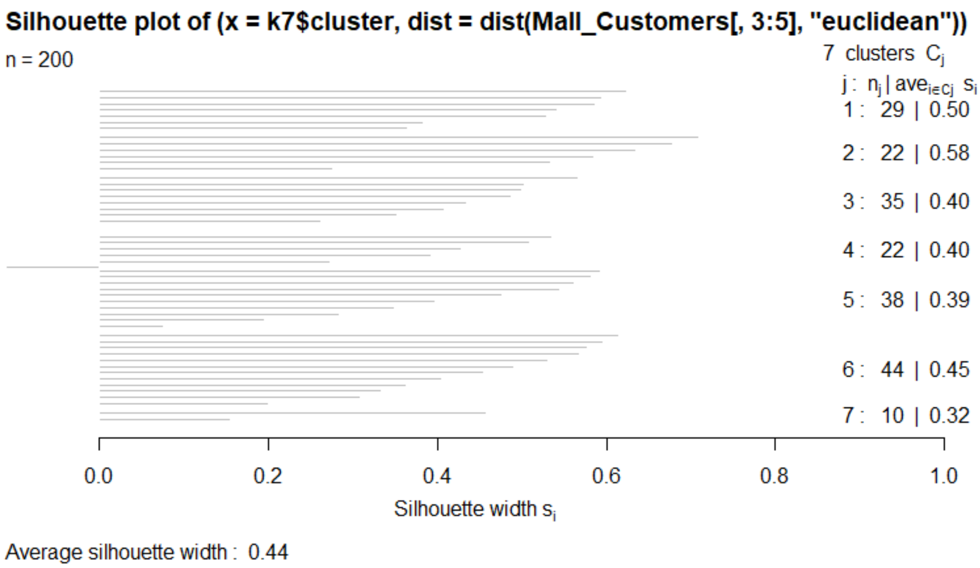
Code:

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



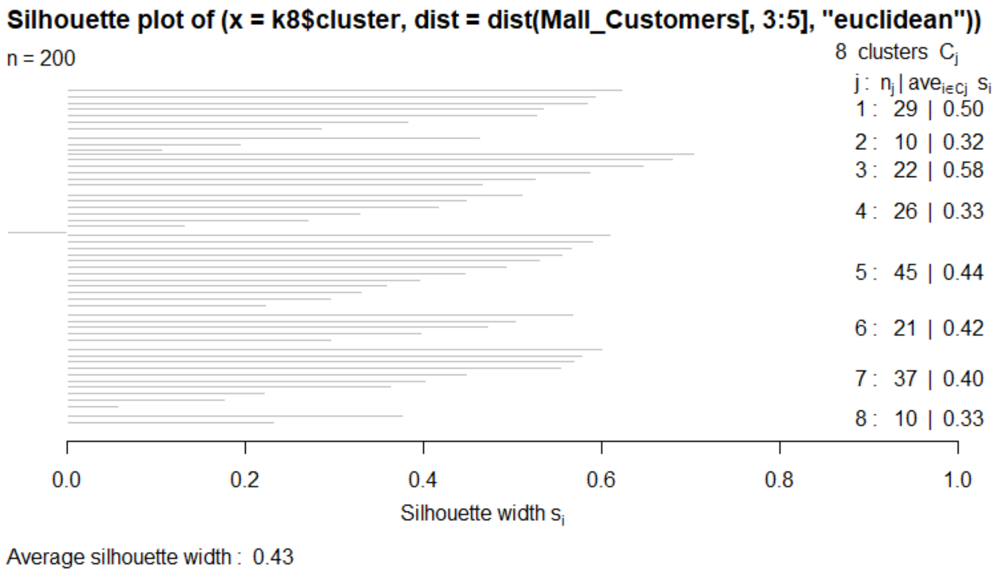
Code:

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



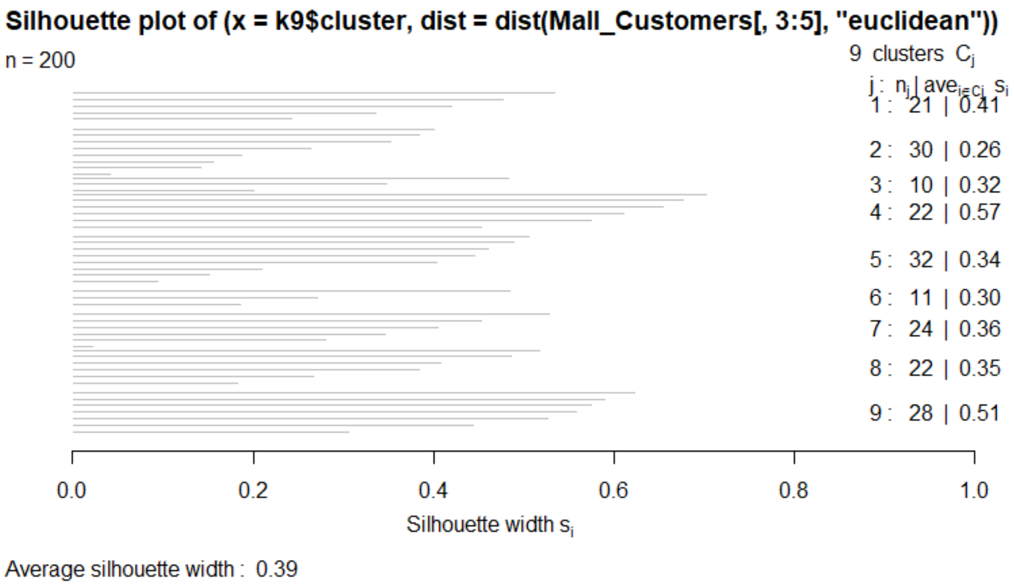
Code:

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



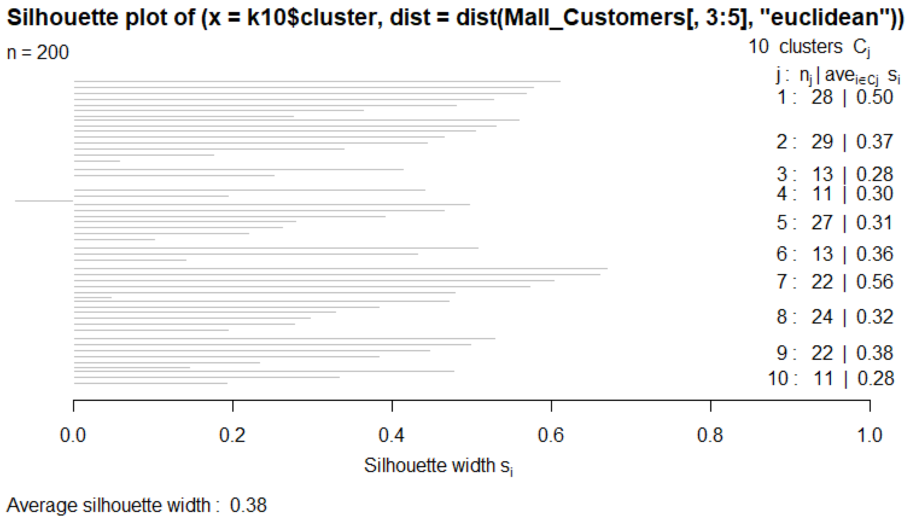
Code:

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



Code:

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

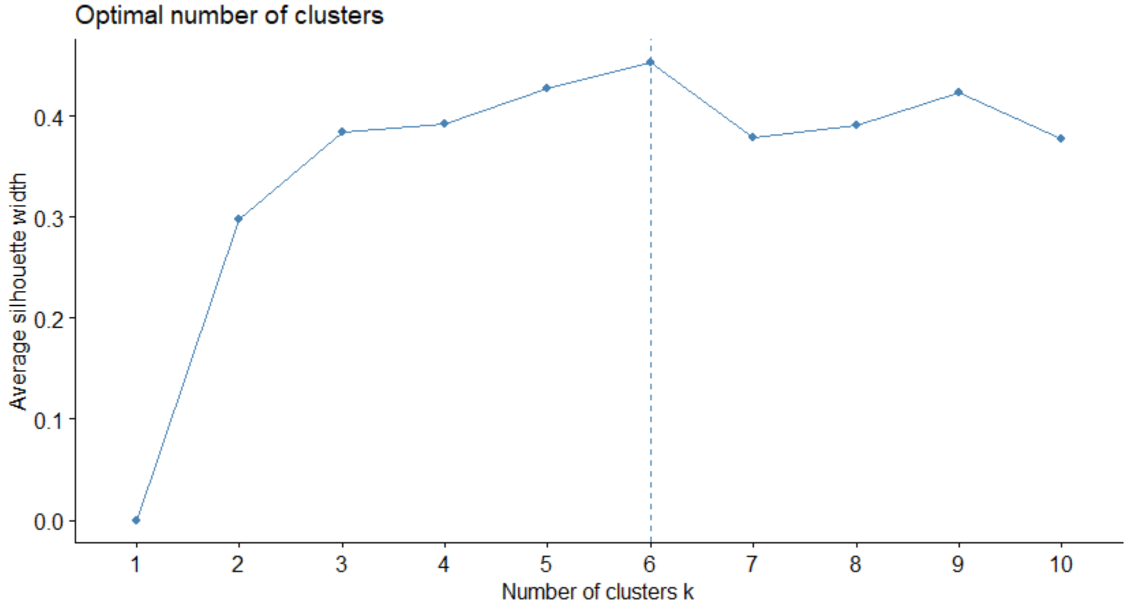


fviz_nbclust() function to determine and visualize the optimal number of clusters

```
> library(NbClust)
> library(factoextra)
```

Code:

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

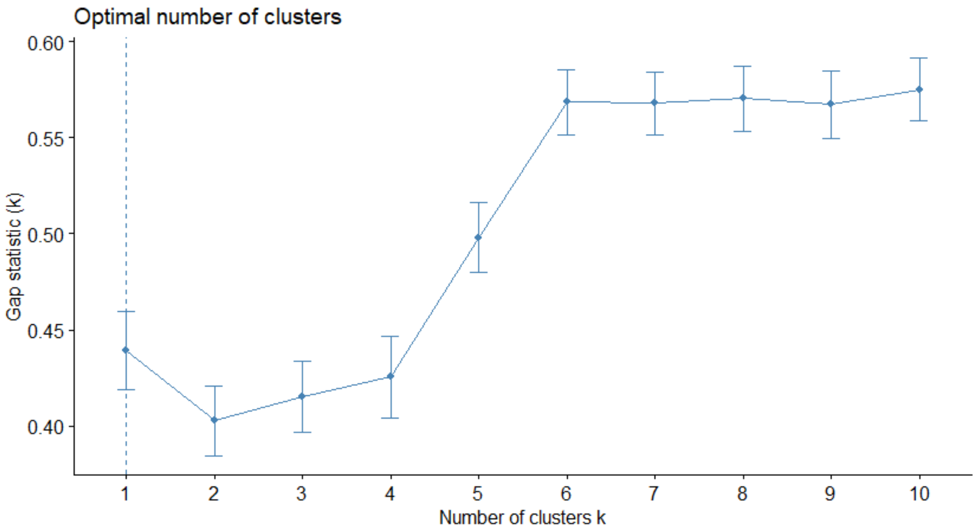


3.3 Gap Statistic Method

```
> library(cluster)
```

Code:

```
> set.seed(123)
> gap_stat <- clusGap(Mall_Customers[,3:5], FUN = kmeans, nstart = 25,
+                     K.max = 10, B = 50)
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 50)  [one "." per sample]:
.................................................. 50
> fviz_gap_stat(gap_stat)
```



 # Check that k = 6 as our optimal cluster

Code:

```
> k6<-kmeans(Mall_Customers[,3:5],6,iter.max=100,nstart=50,algorithm="Lloy
d")
> k6
```

```
Console   Terminal   Jobs
C:/Users/Valued Customer/Desktop/Testing/customer-segmentation-dataset/
> k6<-kmeans(Mall_Customers[,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
 [27] 3 2 3 2 3 2 3 2 3 2 3 2 3 2 1 2 1 4 3 2 1 4 4 4 1 4
 [53] 4 1 1 1 1 1 4 1 1 4 1 1 1 4 1 1 4 4 1 1 1 1 1 4 1 4
 [79] 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
[105] 1 4 1 1 1 1 1 4 4 4 4 1 1 1 1 4 4 4 6 4 6 5 6 5 6
[131] 5 6 4 6 5 6 5 6 5 6 5 6 4 6 5 6 5 6 5 6 5 6 5 6 5 6
[157] 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
```

```
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
 [27] 3 2 3 2 3 2 3 2 3 2 3 2 3 2 1 2 1 4 3 2 1 4 4 4 1 4
 [53] 4 1 1 1 1 1 4 1 1 4 1 1 1 4 1 1 4 4 1 1 1 1 1 4 1 4
 [79] 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
[105] 1 4 1 1 1 1 1 4 4 4 4 4 1 1 1 1 4 4 4 6 4 6 5 6 5 6
[131] 5 6 4 6 5 6 5 6 5 6 5 6 4 6 5 6 5 6 5 6 5 6 5 6 5 6
[157] 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
[183] 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6

Within cluster sum of squares by cluster:
[1]  8062.133  4099.818  7732.381  7742.895 16690.857
[6] 13972.359
 (between_SS / total_SS =  81.1 %)

Available components:

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

4. ***Visualizing the Clustering Results using the First Two Principle Components***

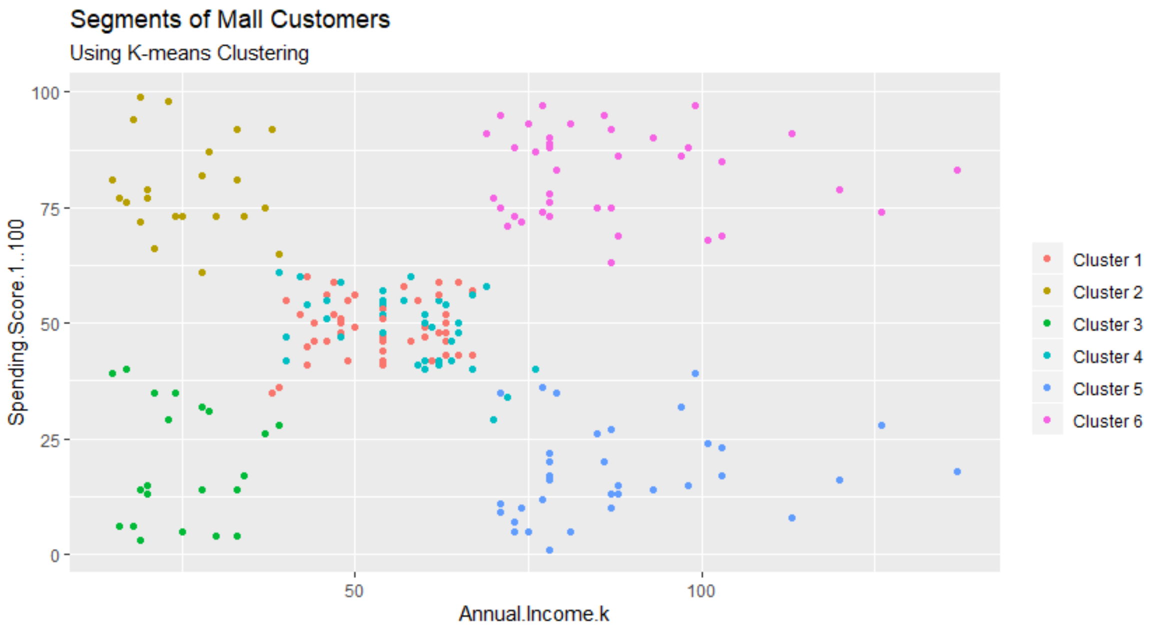
Code:

```
> pcclust=prcomp(Mall_Customers[,3:5],scale=FALSE)
> summary(pcclust)
```

```
> pcclust=prcomp(Mall_Customers[,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
>
> 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
>
```

Code:

```
> set.seed(1)
> ggplot(Mall_Customers, 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 Clus
tering")
```



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

Cluster 6 and 4 – These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.

Cluster 1 – 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 2 – This cluster denotes a high annual income and low yearly spend.

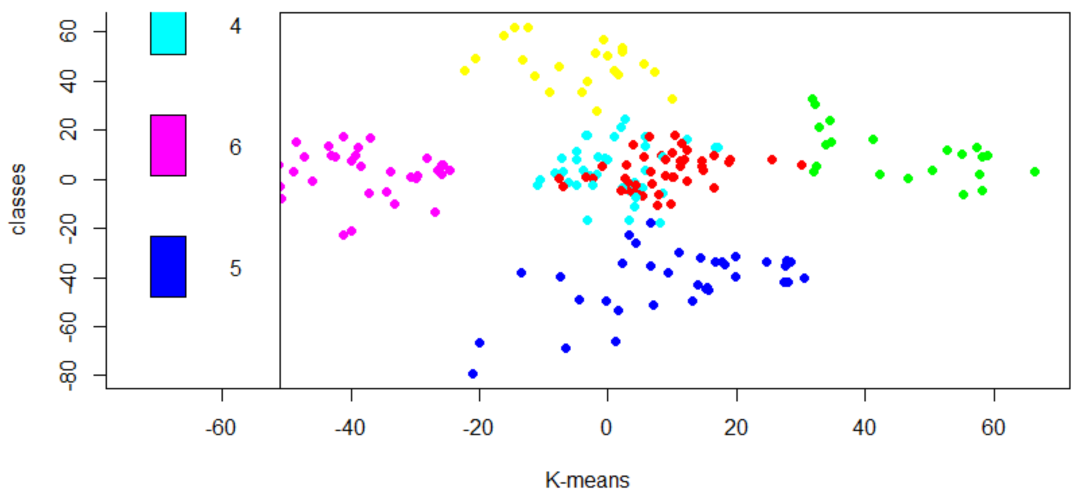
Cluster 5 – This cluster represents a low annual income but its high yearly expenditure.

Code:

```
> ggplot(Mall_Customers, 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 Clus
tering")
```

Code:

```
> kCols=function(vec){cols=rainbow (length (unique (vec)))
+ return (cols[as.numeric(as.factor(vec))])}
>
> digCluster<-k6$cluster; dignm<-as.character(digCluster);
>
> plot(pcclust$x[,1:2], col =kCols(digCluster),pch =19,xlab ="K-means",yla
b="classes")
> legend("bottomleft",unique(dignm),fill=unique(kCols(digCluster)))
```



Cluster 4 and 1 – These two clusters consist of customers with medium PCA1 and medium PCA2 score.

Cluster 6 – This cluster represents customers having a high PCA2 and a low PCA1.

Cluster 5 – In this cluster, there are customers with a medium PCA1 and a low PCA2 score.

Cluster 3 – This cluster comprises of customers with a high PCA1 income and a high PCA2.

Cluster 2 – This comprises of customers with a high PCA2 and a medium annual spend of income.

.