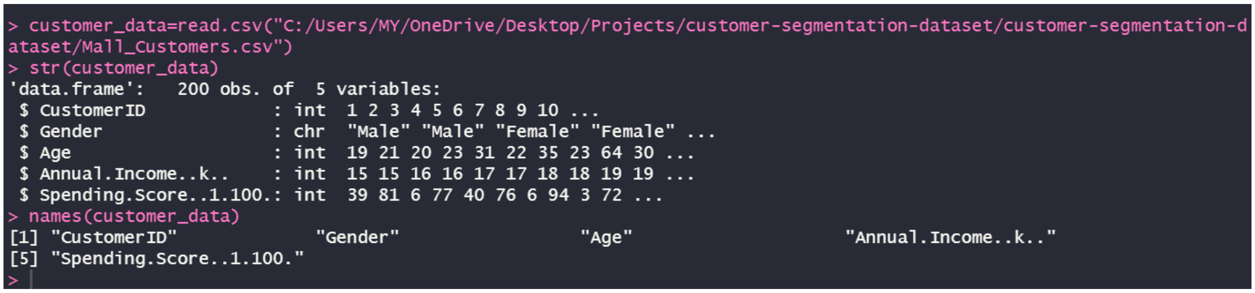
Customer Segmentation Using Machine Learning in R

Reading the DataSet

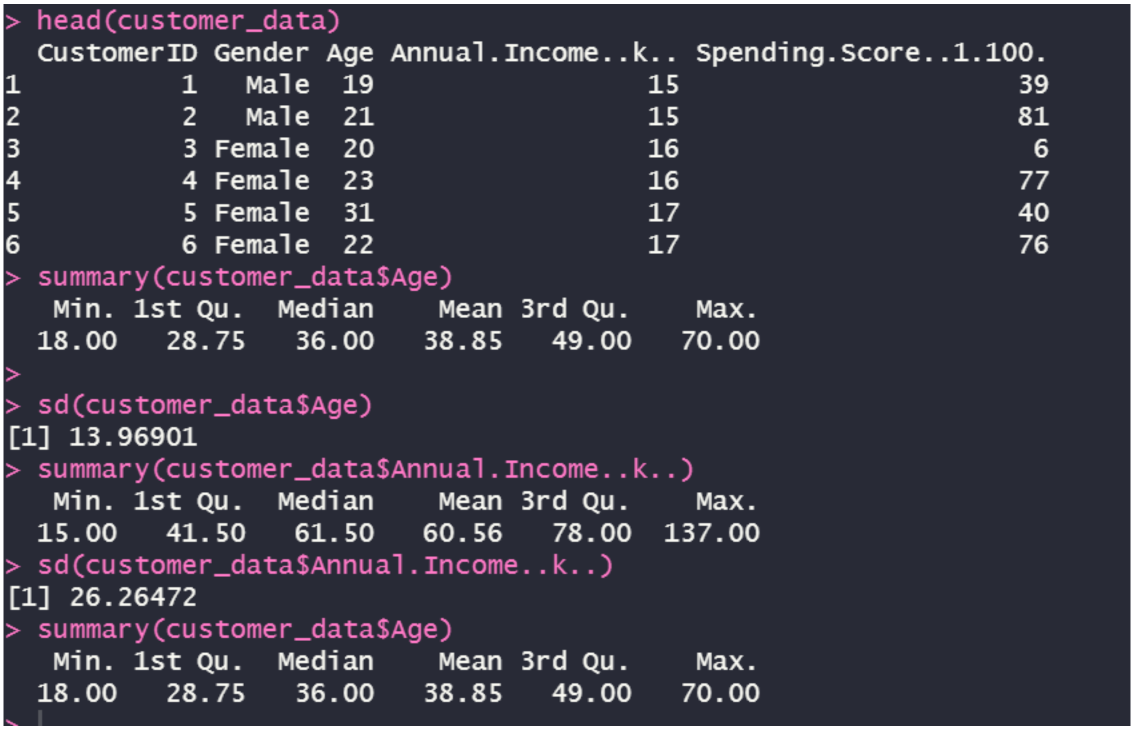
```
customer_data=read.csv("C:/Users/MY/OneDrive/Desktop/Projects/customer-segmentation-dataset/customer-segmentation-dataset/Mall_Customers.cs
str(customer_data)
names(customer_data)
```



```
> customer_data=read.csv("C:/Users/MY/OneDrive/Desktop/Projects/customer-segmentation-dataset/customer-segmentation-d
ataset/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                : 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"                    "Annual.Income..k.."
[5] "Spending.Score..1.100."
>
```

Get Data Insights

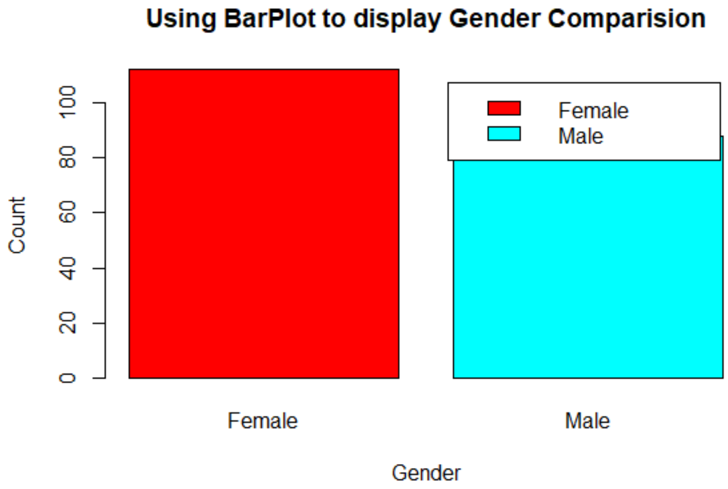
```
head(customer_data)
summary(customer_data$Age)

sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)
```

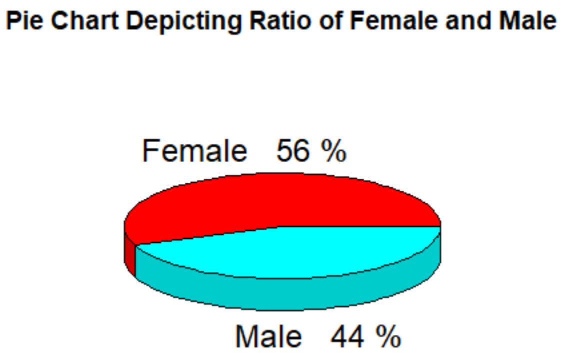


```
> head(customer_data)
  CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
1          1   Male  19                 15                     39
2          2   Male  21                 15                     81
3          3 Female  20                 16                      6
4          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] 13.96901
> 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 
>
```

Visualize the Gender Attribute

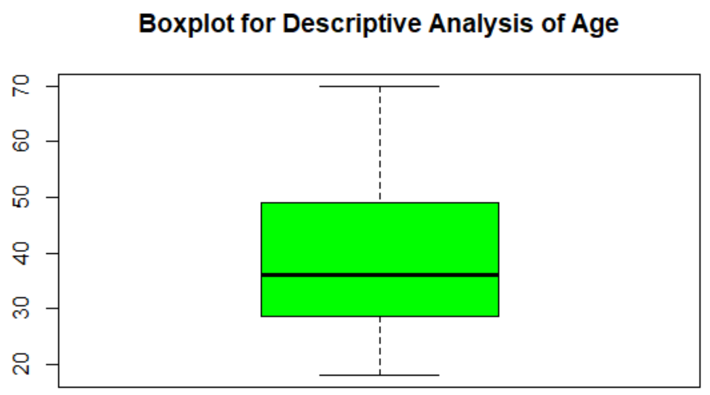
```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))

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

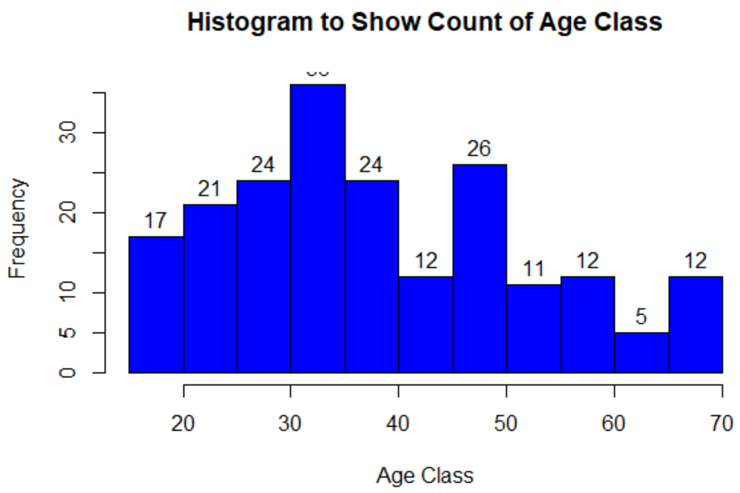




Age Distribution

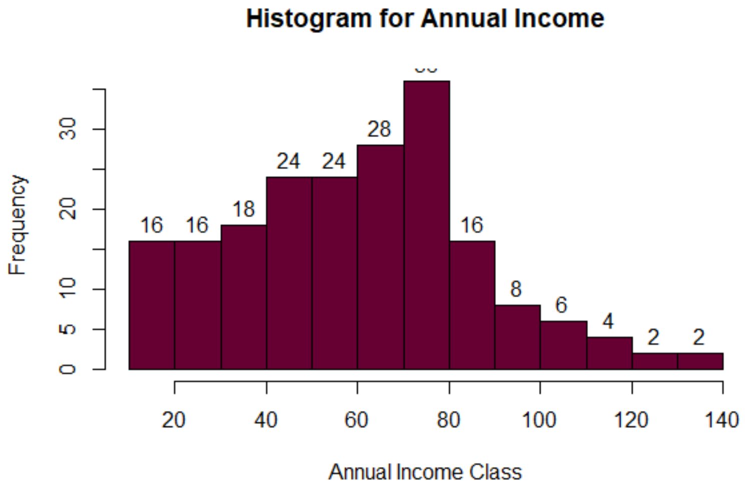
```
summary(customer_data$Age)
hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)
boxplot(customer_data$Age,
        col="green",
        main="Boxplot for Descriptive Analysis of Age")
```



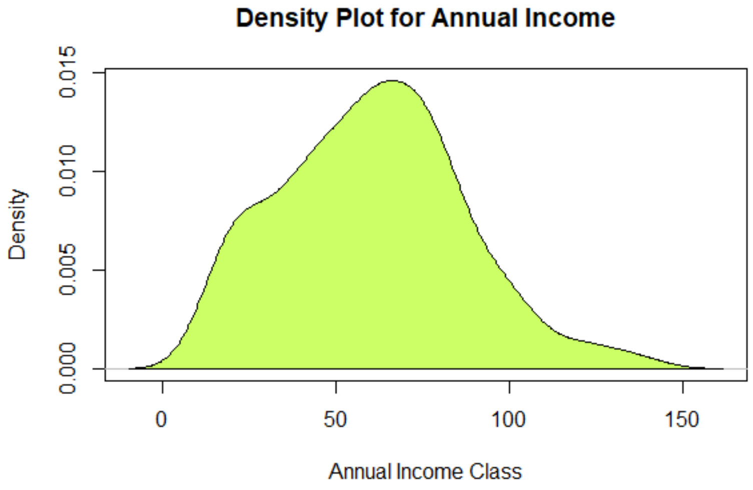


Analysis of Annual Income of Customers

```
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)
```

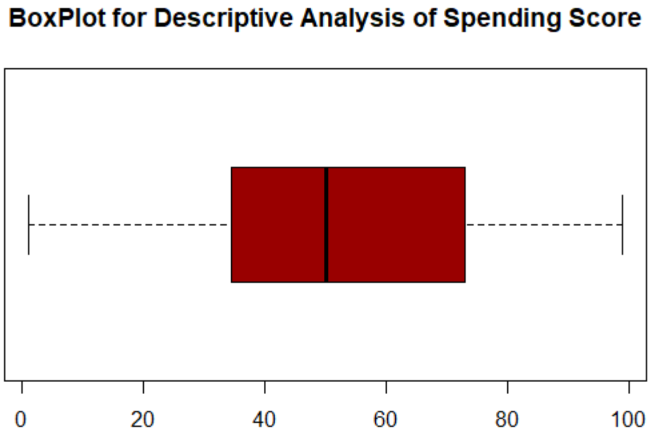


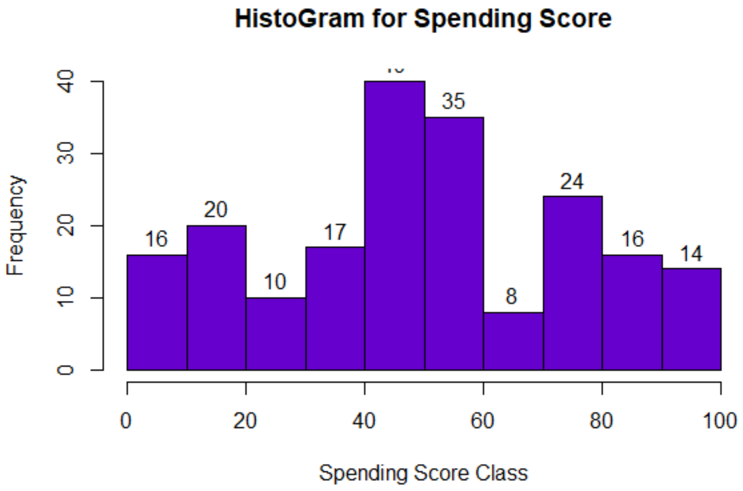
```
plot(density(customer_data$Annual.Income..k..),
     col="yellow",
     main="Density Plot for Annual Income",
     xlab="Annual Income Class",
     ylab="Density")
polygon(density(customer_data$Annual.Income..k..),
        col="#ccff66")
```



Expenditure of Customers

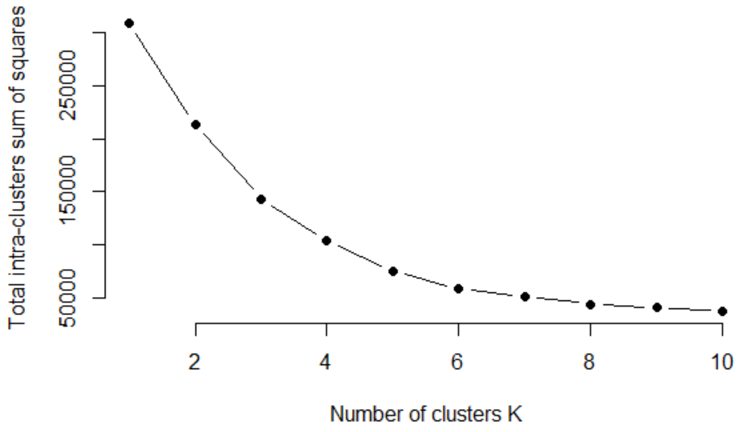
```
summary(customer_data$Spending.Score..1.100.)
#Min. 1st Qu. Median Mean 3rd Qu. Max.
## 1.00 34.75 50.00 50.20 73.00 99.00
boxplot(customer_data$Spending.Score..1.100.,
        horizontal=TRUE,
        col="#990000",
        main="BoxPlot for Descriptive Analysis of Spending Score")
hist(customer_data$Spending.Score..1.100.,
     main="HistoGram for Spending Score",
     xlab="Spending Score Class",
     ylab="Frequency",
     col="#6600cc",
     labels=TRUE)
```





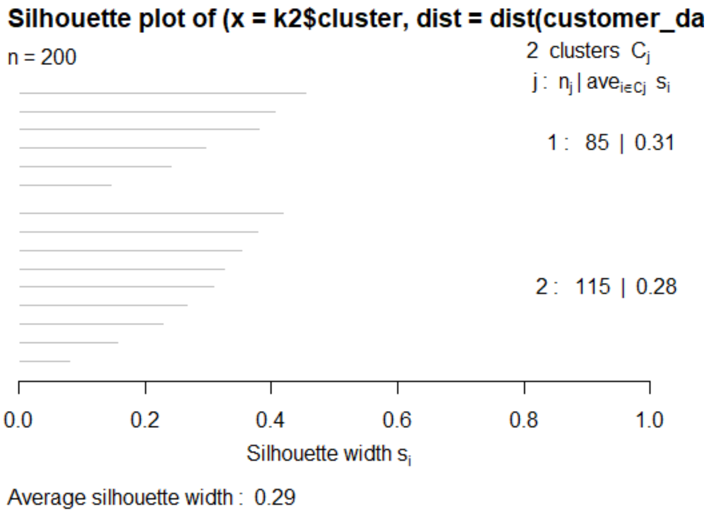
Applying K-means Clustering

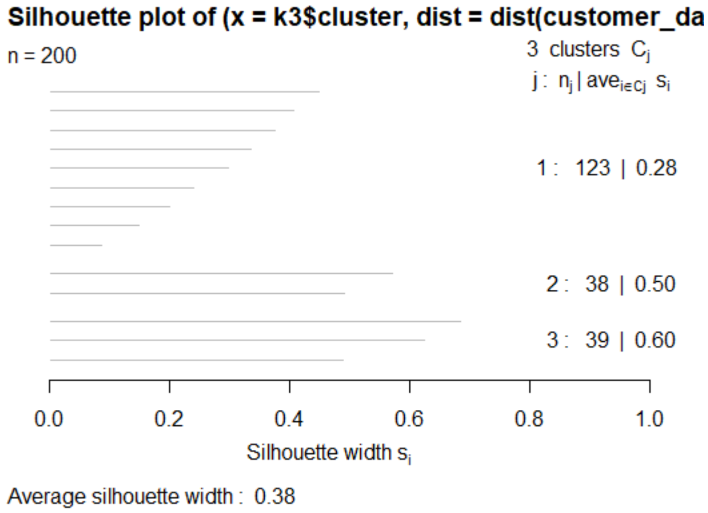
```
library(purrr)
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss
}
k.values <- 1:10
iss_values <- map_dbl(k.values, iss)
plot(k.values, iss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total intra-clusters sum of squares")
```

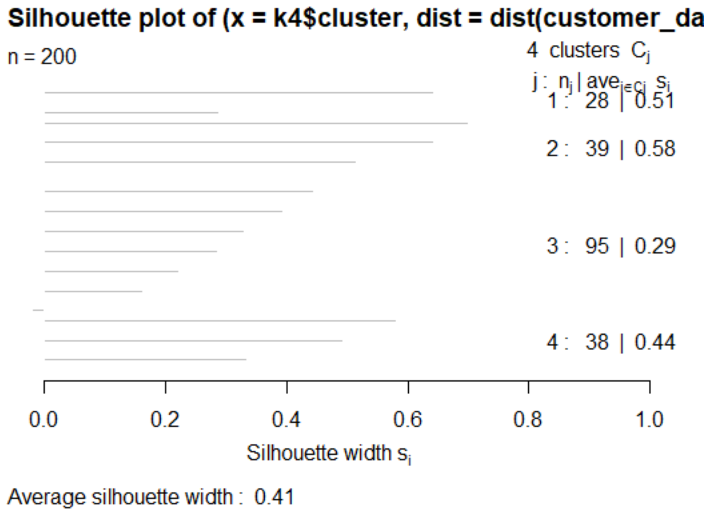


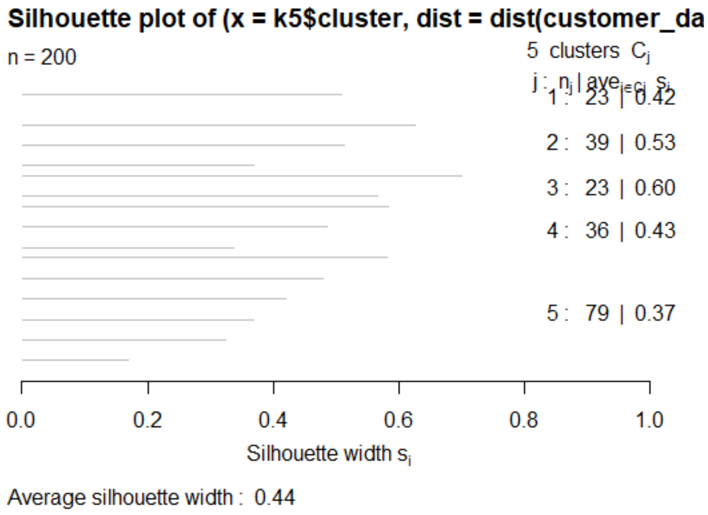
Applying Average sillhoutte method

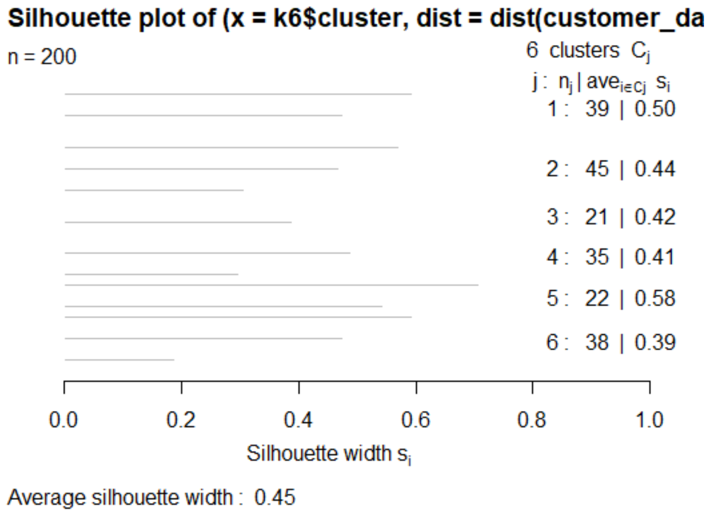
```
library(cluster)
library(gridExtra)
library(grid)
k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))

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")
```

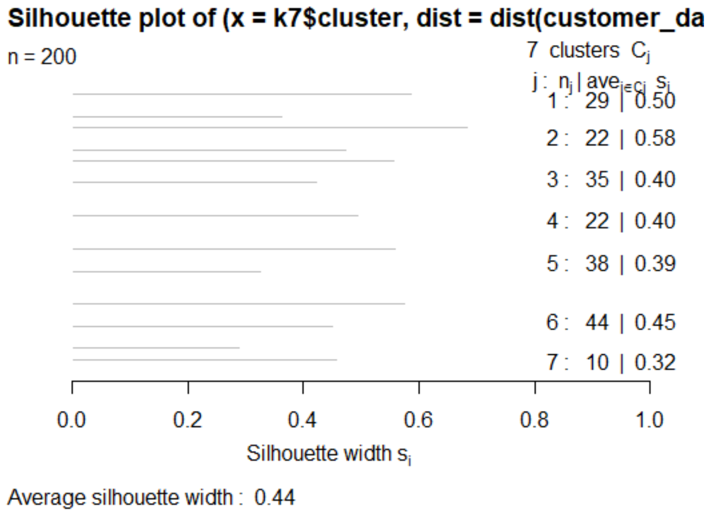

Silhouette plot of (x = k2$cluster, dist = dist(customer_da
n = 200
2 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 85 | 0.31
2 : 115 | 0.28
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.29

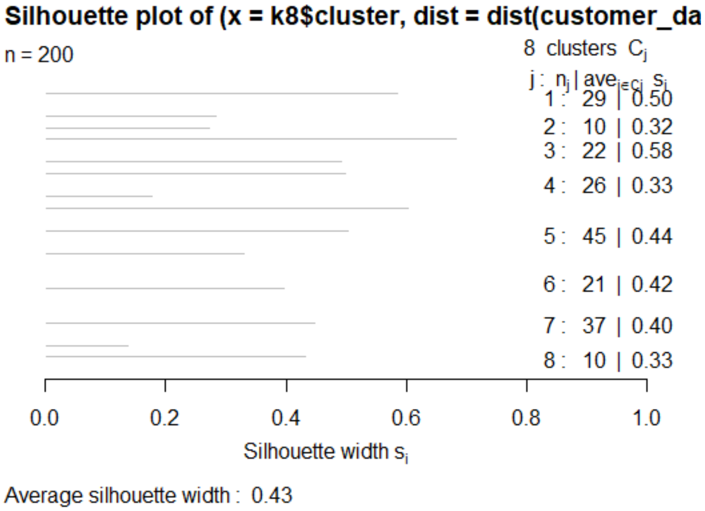

Silhouette plot of (x = k3$cluster, dist = dist(customer_da
n = 200
3 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 123 | 0.28
2 : 38 | 0.50
3 : 39 | 0.60
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.38

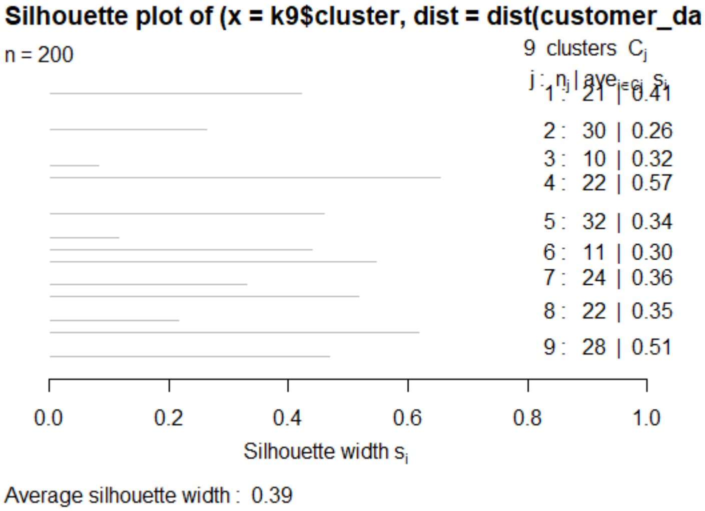

Silhouette plot of (x = k4$cluster, dist = dist(customer_da
n = 200
4 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 28 | 0.51
2 : 39 | 0.58
3 : 95 | 0.29
4 : 38 | 0.44
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.41

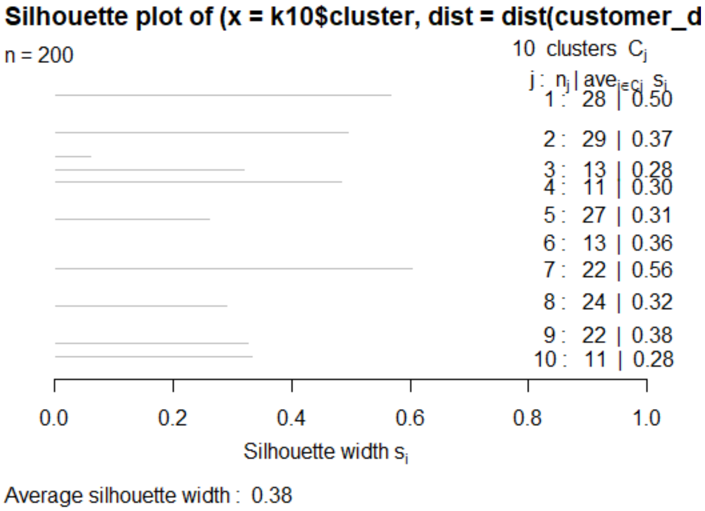

Silhouette plot of (x = k5$cluster, dist = dist(customer_da
n = 200
5 clusters C_j
j : n_j | ave_{i∈Cj} s_i
1 : 23 | 0.42
2 : 39 | 0.53
3 : 23 | 0.60
4 : 36 | 0.43
5 : 79 | 0.37
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.44


Silhouette plot of (x = k6$cluster, dist = dist(customer_da
n = 200
6 clusters C_j
j : n_j | ave_{i∈Cj} s_i
1 : 39 | 0.50
2 : 45 | 0.44
3 : 21 | 0.42
4 : 35 | 0.41
5 : 22 | 0.58
6 : 38 | 0.39
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.45


Silhouette plot of (x = k7$cluster, dist = dist(customer_da
n = 200
7 clusters C_j
j : n_j | ave_{i∈Cj} s_i
1 : 29 | 0.50
2 : 22 | 0.58
3 : 35 | 0.40
4 : 22 | 0.40
5 : 38 | 0.39
6 : 44 | 0.45
7 : 10 | 0.32
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.44


Silhouette plot of (x = k8$cluster, dist = dist(customer_da
n = 200
8 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 29 | 0.50
2 : 10 | 0.32
3 : 22 | 0.58
4 : 26 | 0.33
5 : 45 | 0.44
6 : 21 | 0.42
7 : 37 | 0.40
8 : 10 | 0.33
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.43


Silhouette plot of (x = k9$cluster, dist = dist(customer_da
n = 200
9 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 21 | 0.41
2 : 30 | 0.26
3 : 10 | 0.32
4 : 22 | 0.57
5 : 32 | 0.34
6 : 11 | 0.30
7 : 24 | 0.36
8 : 22 | 0.35
9 : 28 | 0.51
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.39


Silhouette plot of (x = k10$cluster, dist = dist(customer_d
n = 200
10 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 28 | 0.50
2 : 29 | 0.37
3 : 13 | 0.28
4 : 11 | 0.30
5 : 27 | 0.31
6 : 13 | 0.36
7 : 22 | 0.56
8 : 24 | 0.32
9 : 22 | 0.38
10 : 11 | 0.28
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.38