Project Template-3

Section 3

Introduction.

Customer Segmentation is one the most important modeling of unsupervised learning. Using clustering techniques, business can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, this project will make use of K-means clustering which is the essential algorithm for clustering unlabeled dataset.

Source of the dataset : https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

Purpose : This data set is created only for the learning purpose of the customer segmentation concepts , also known as market basket analysis.

Content

This is a sample data to mimics supermarket mall data which can be acquired through membership cards , It has some basic data about regular mall customers like :

- Customer ID
- Age
- Gender
- Annual Income and
- Spending score

The problem statement

Project targets to present customer segmentation analysis in form of visual representation to marketing team of a shopping mall to help them to plan effective marketing stretegies for prosprective customres. Based on available variables in dataset , project is trying to provide analysis on following data points :

- Price segmentation is basic and very common factor to determine prospective customers.
- How does demographic information like Gender/ Age can be used to identify prospective customer ?
- How does annual Income can be used to Identify customer buying pattern ?
- Finally spending score to determine prospective customer.

Approach (How you addressed this problem statement)

Project begins with loading dataset from .csv file followed by basic statistical analysis and data exploration. Finally, going through the input data to gain necessary insights about it.

Following are key analysis and their presentations :

1. Customer Gender Visualization

2. Visualization of Age Distribution
3. Analysis of the Annual Income of the Customers
4. Analyzing Spending Score of the Customers
5. Specifically, use of a clustering algorithm called **K-means clustering** by analyzing and visualizing the data , cluster assignment and finally determining optimal cluster using following three popular methods :

i. Elbow method
ii. Silhouette method
iii. Gap statistic

Analysis

```
library(plotrix)

library(purrr)
library(cluster)
library(gridExtra)
library(grid)

library(NbClust)
library(factoextra)
```

Important package libraries :

```
## Loading required package: ggplot2

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

Shopping mall dataset

```
setwd("~/MadR/Workspaces/dsc520")
customer_data=read.csv("~/MadR/Workspaces/dsc520/data/Mall_Customers.csv")

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)
```

```
##    CustomerID        Gender               Age        Annual.Income..k..
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00
##  Spending.Score..1.100.
##  Min.   : 1.00
##  1st Qu.:34.75
##  Median :50.00
##  Mean   :50.20
##  3rd Qu.:73.00
##  Max.   :99.00
```

```r
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 ...
```

Data Exploration

```r
'Age standard deviation'
```

```
## [1] "Age standard deviation"
```

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

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

```r
'Annual Income standard deviation'
```

```
## [1] "Annual Income standard deviation"
```

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

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

```r
'Spending standard deviation'
```

```
## [1] "Spending standard deviation"
```

```
sd(customer_data$Spending.Score..1.100.)
```

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

Implications

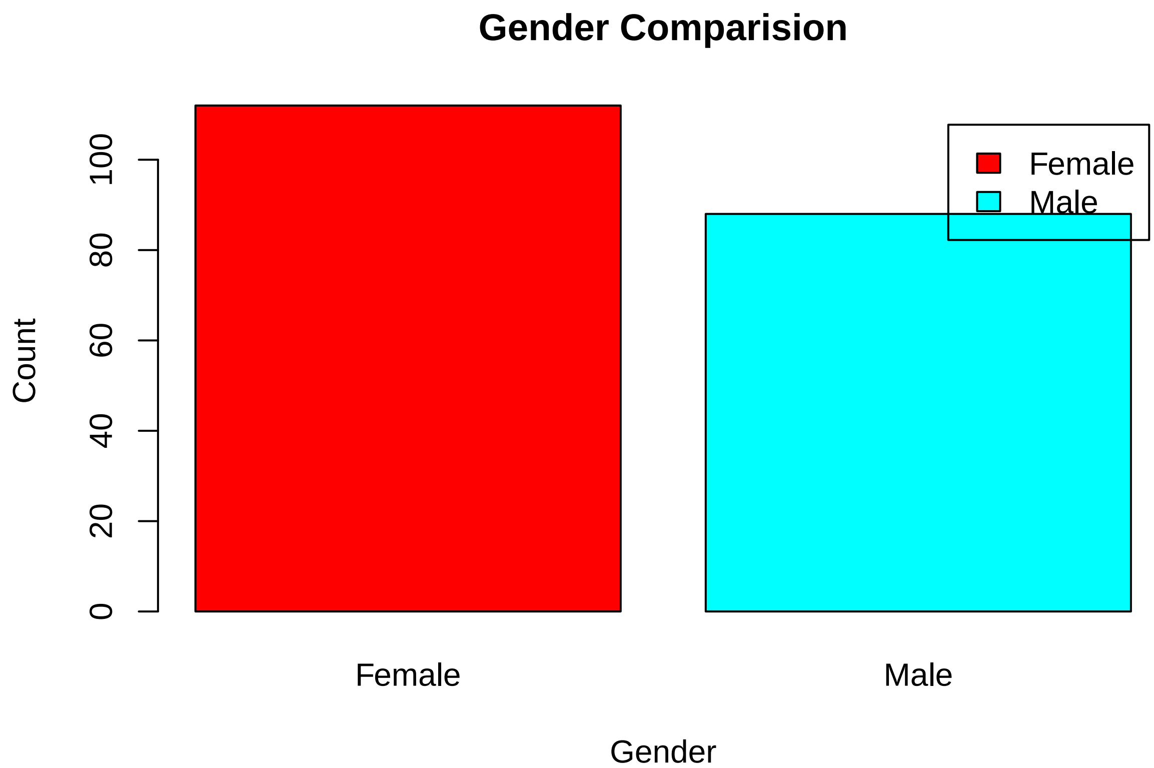
Analysis presented in the project helps to identify prospective customers based on their age group, Income range , gender and spending scores. This analysis was done on a smaller set of available customer data of a shopping mall. Same model can be replicated for any other marketing team (from other businesses) to build their stretegy to target prospective cudtomers.

However, this study was done on a limited sized data with only 5 variables.This is recommnded to collect more customer data to build a bigger dataset with few more demographic information along with shopping history of a customer for better insight.

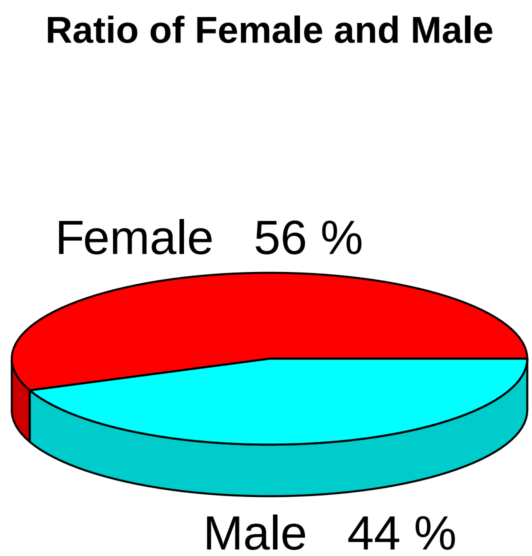
Visualization and Statistics

Customer Gender Visualization

```
a=table(customer_data$Gender)
barplot(a,
        main = "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=" ")
pie3D(a,
      labels  =lbs,
      main="Ratio of Female and Male")
```



Customer Age Visualization

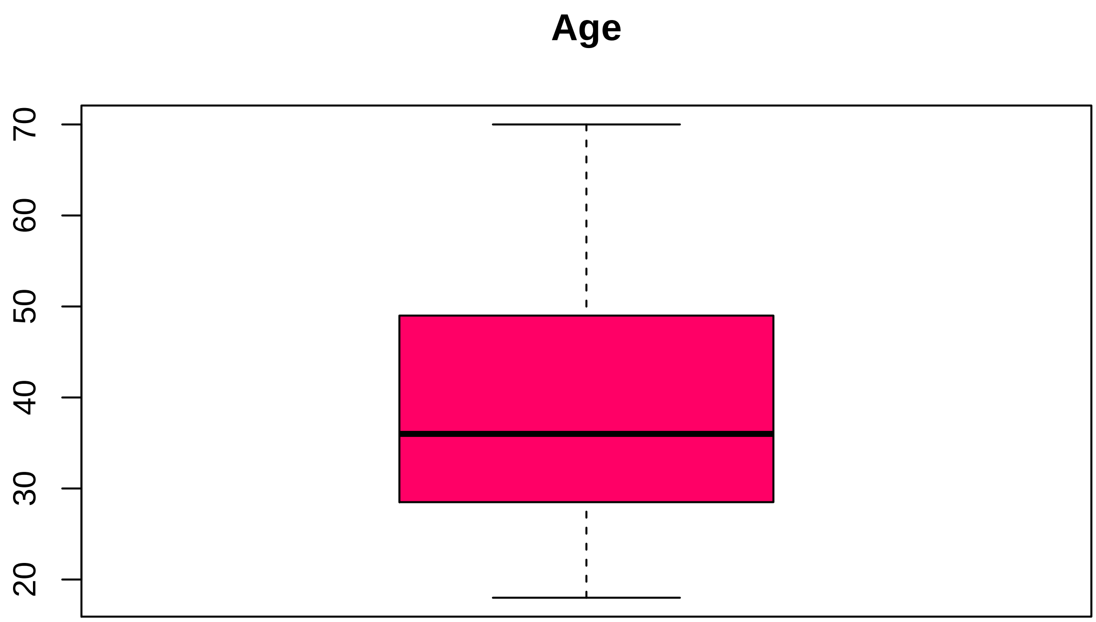
maximum customer ages are between 30 to 35 with minimum /maximum ages are 18 & 70 respectively

```
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="Age class",
     xlab="Age Class",
     ylab  ="Frequency",
     labels=TRUE)
```


```
boxplot(customer_data$Age,
        col="#ff0066",
        main="Age")
```



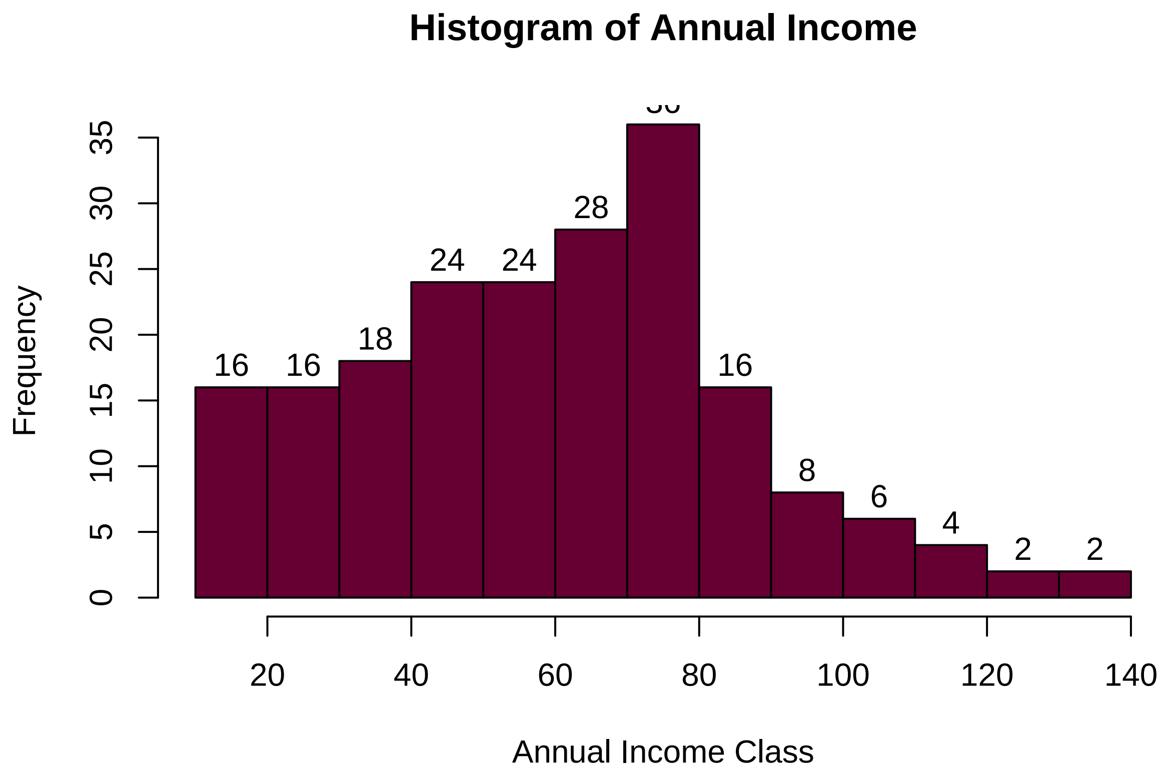
Analyzing Annual Income

Income range lies between 15 to 137K range with average income being 60.56

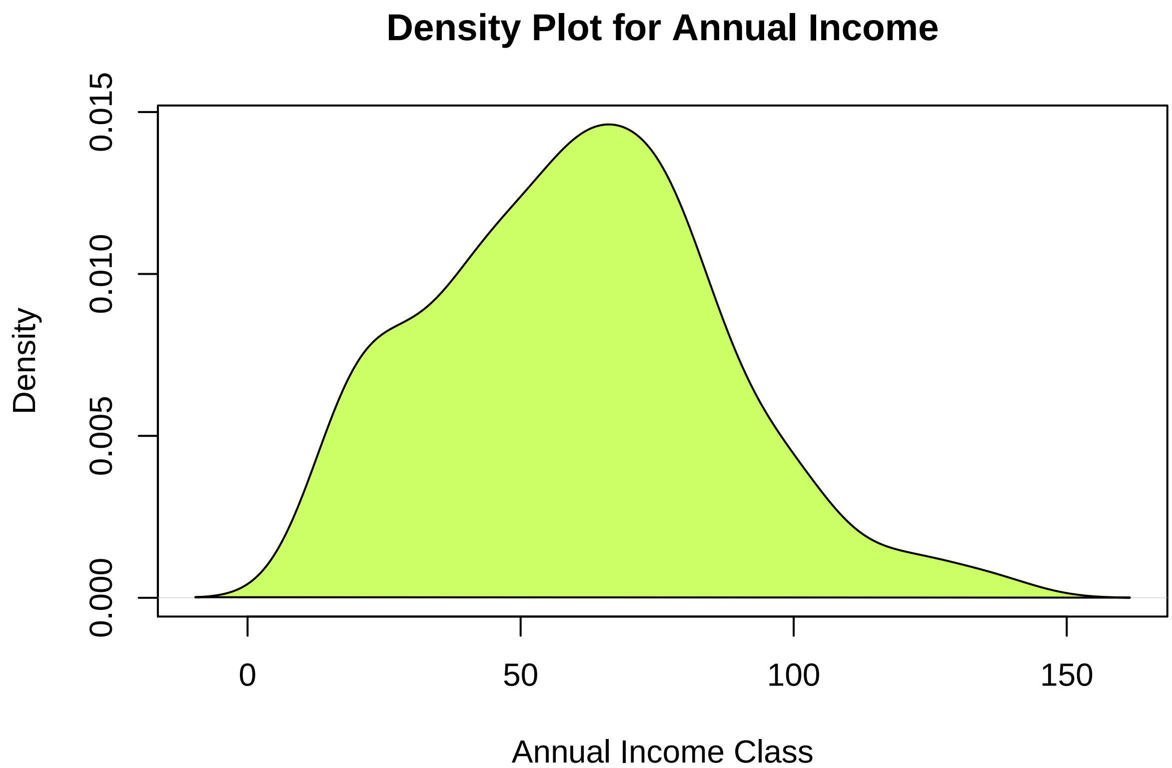
```
summary(customer_data$Annual.Income..k..)
```

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

```
hist(customer_data$Annual.Income..k,
     col="#660033",
     main="Histogram of 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")
```



Analyzing Spending Score

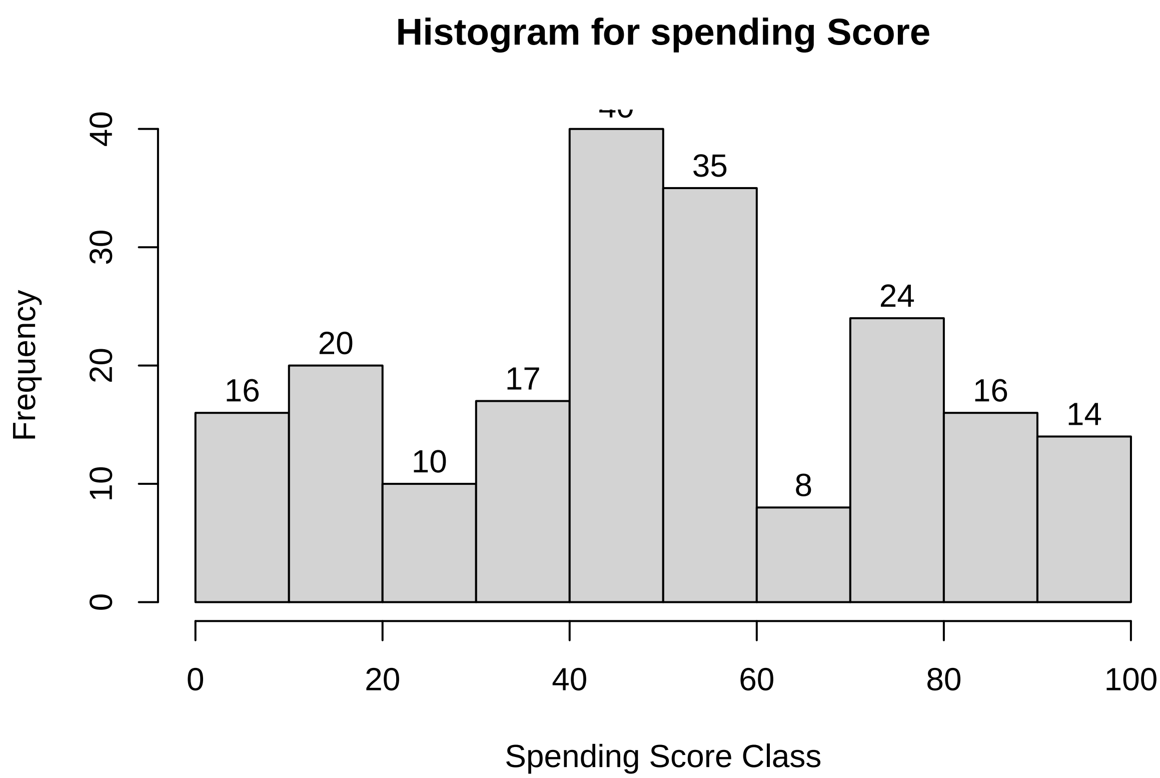
Spending score range is between 1 to 99 with average of 50.20. Histogram indicates customer between 40 and 50 have the highest spending score

```
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
```

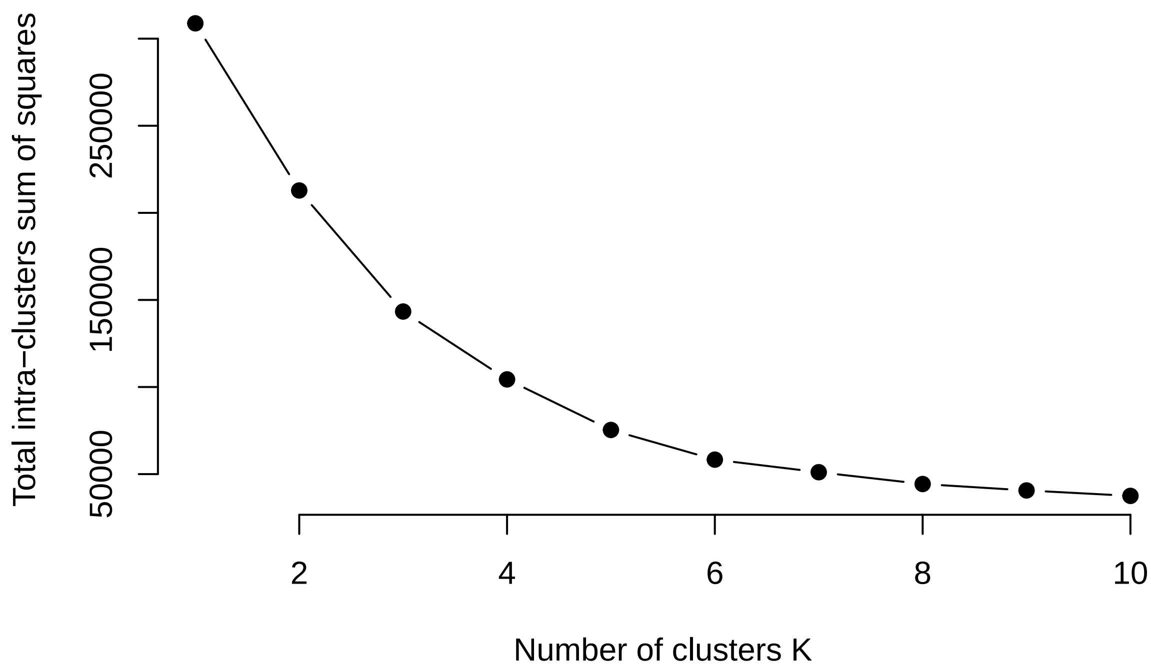
```
boxplot(customer_data$Spending.Score..1.100.,
        horizontal = TRUE,
        col="#990000",
        main="Boxplot of Spending Score")
```


```r
hist(customer_data$Spending.Score..1.100.,
     main = "Histogram for spending Score",
     xlab = "Spending Score Class",
     ylab = "Frequency",
     labels = TRUE)
```

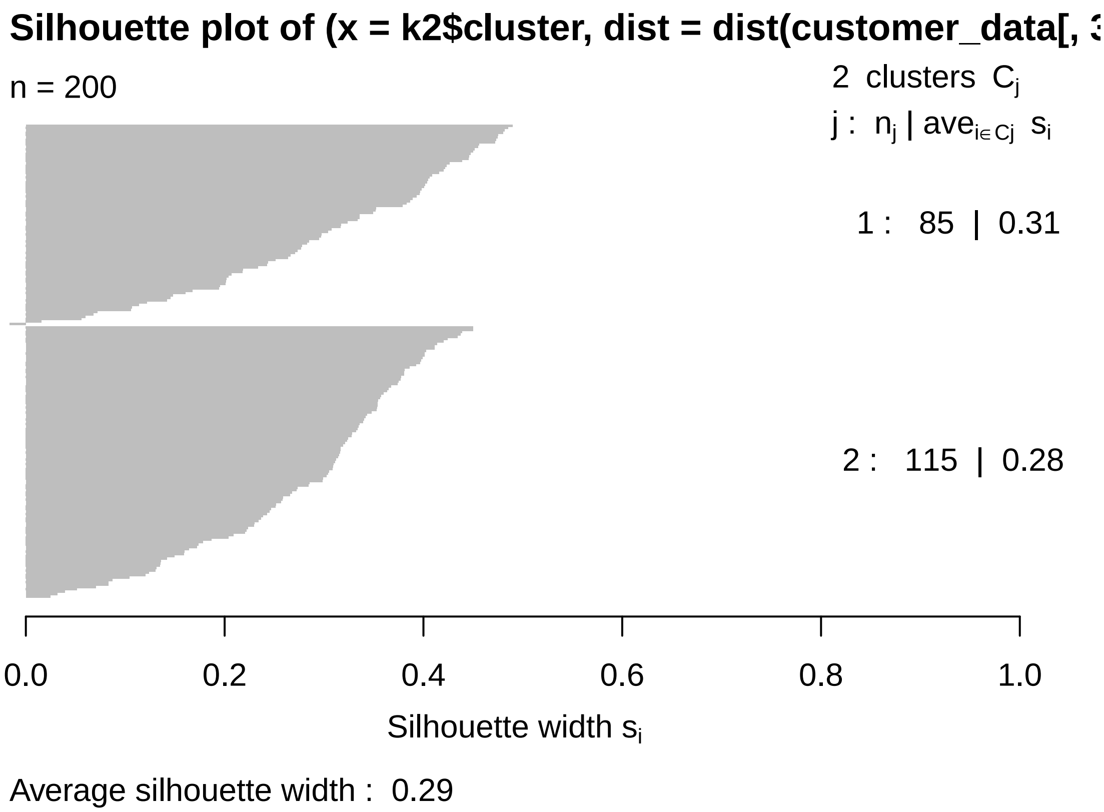


K-means Algorithm

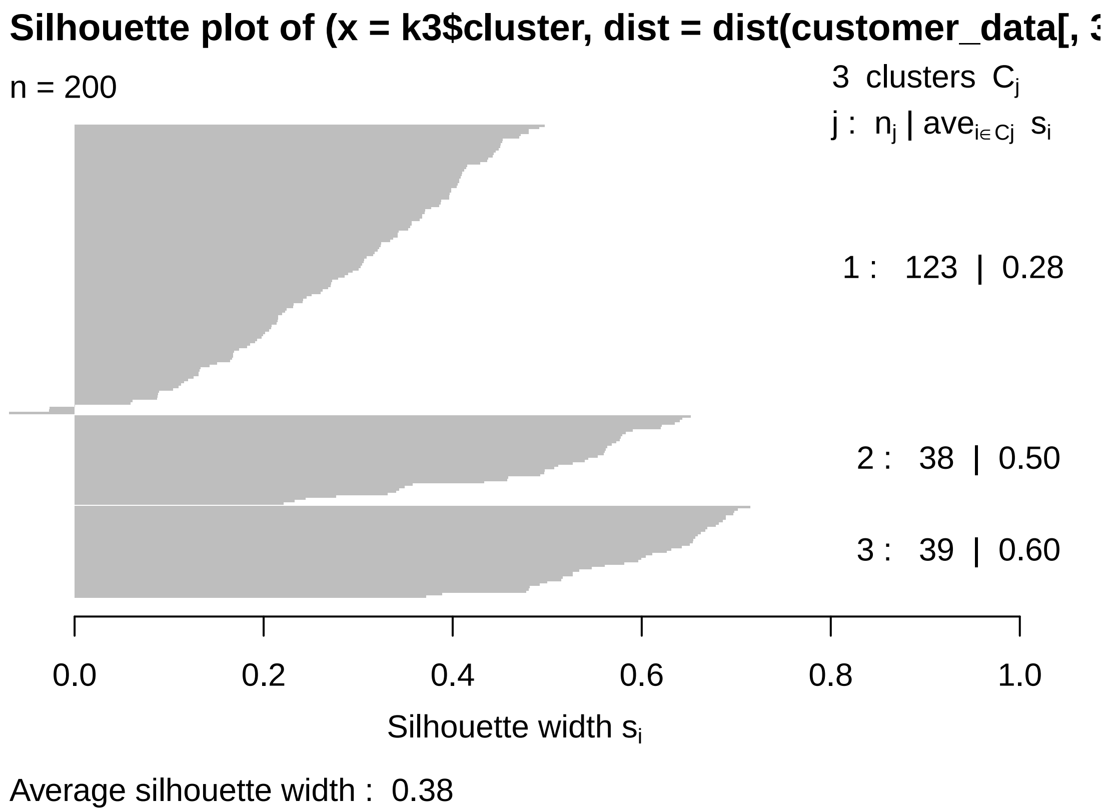
```
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
  kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss
}

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")
```



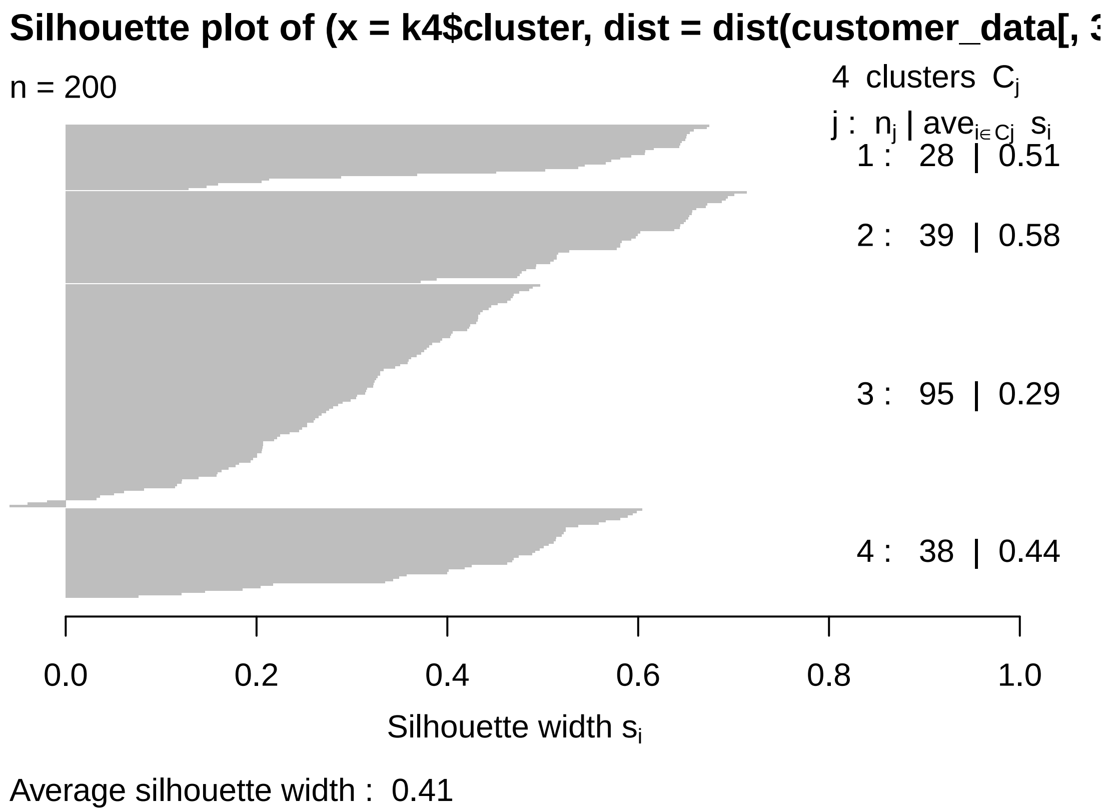

```r
k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))
```



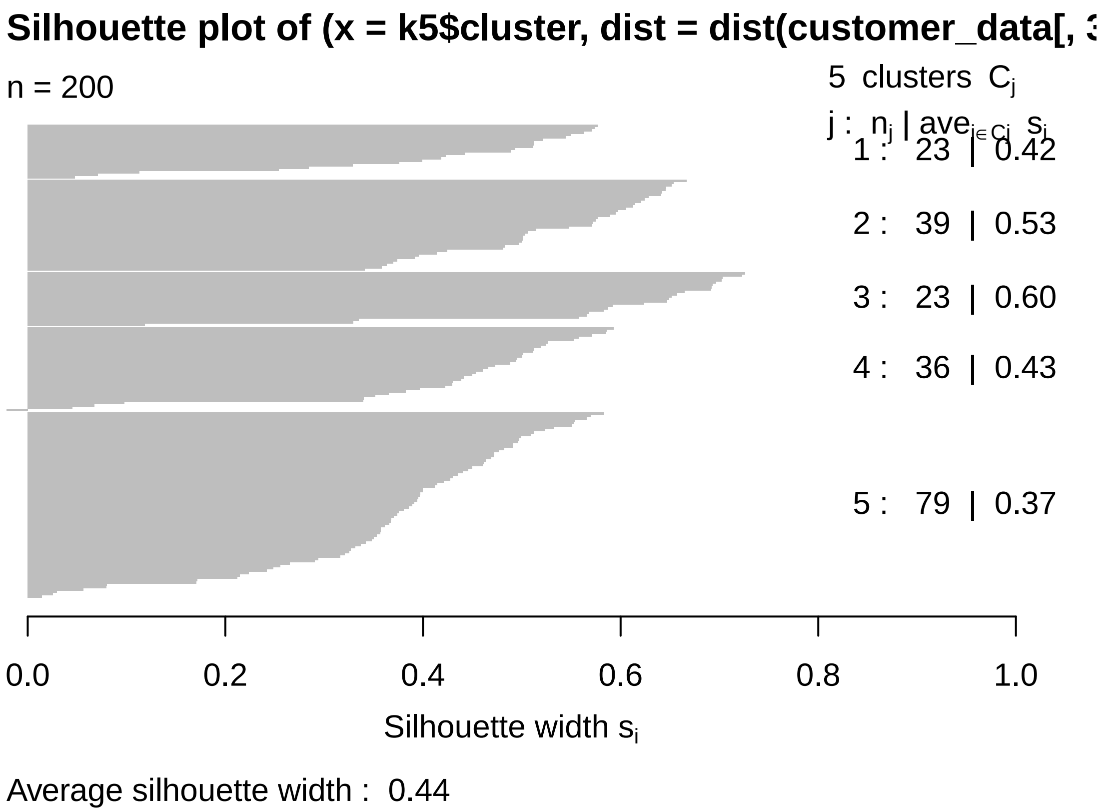

```
k3<-kmeans(customer_data[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")
s3<-plot(silhouette(k3$cluster,dist(customer_data[,3:5],"euclidean")))
```



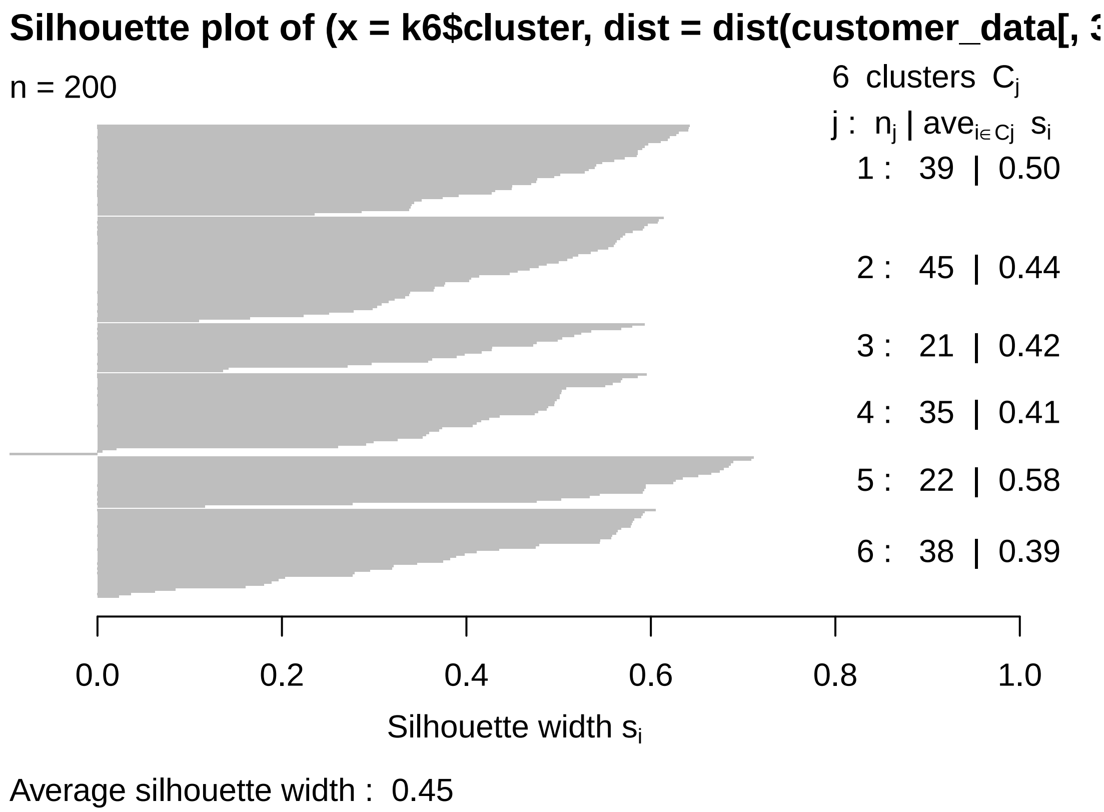

```
k4<-kmeans(customer_data[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")
s4<-plot(silhouette(k4$cluster,dist(customer_data[,3:5],"euclidean")))
```




```
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")))
```

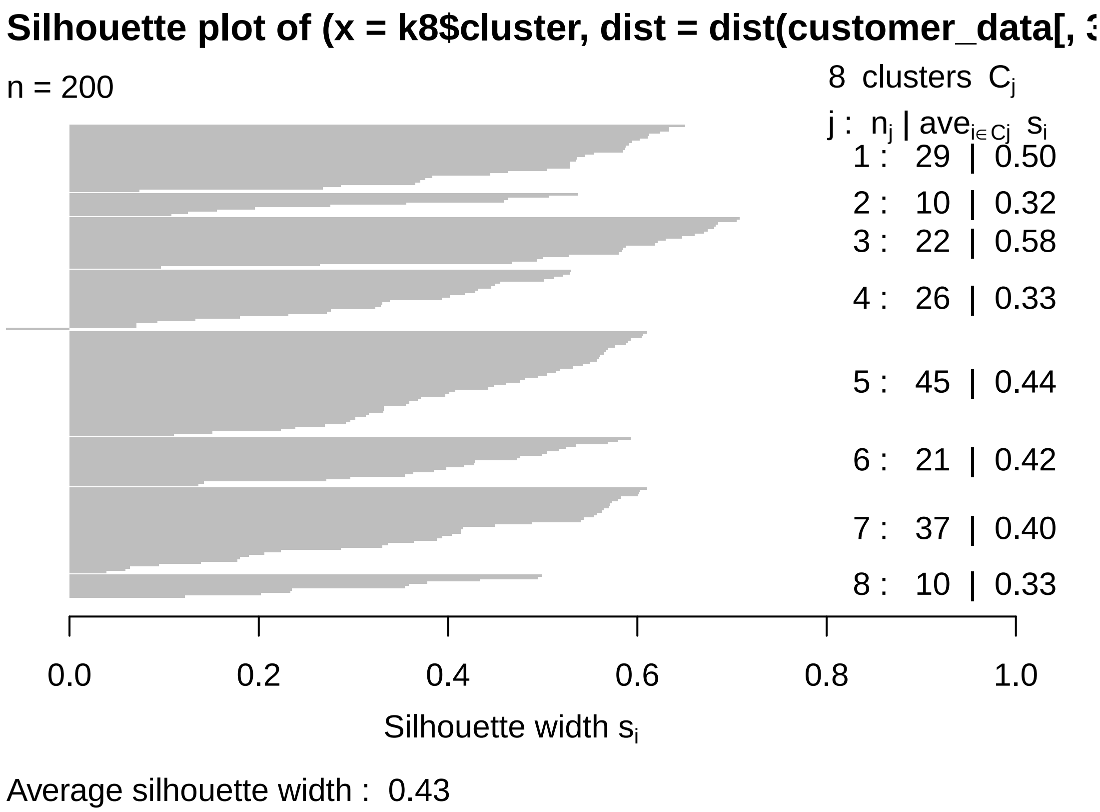



```
k6<-kmeans(customer_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
s6<-plot(silhouette(k6$cluster,dist(customer_data[,3:5],"euclidean")))
```

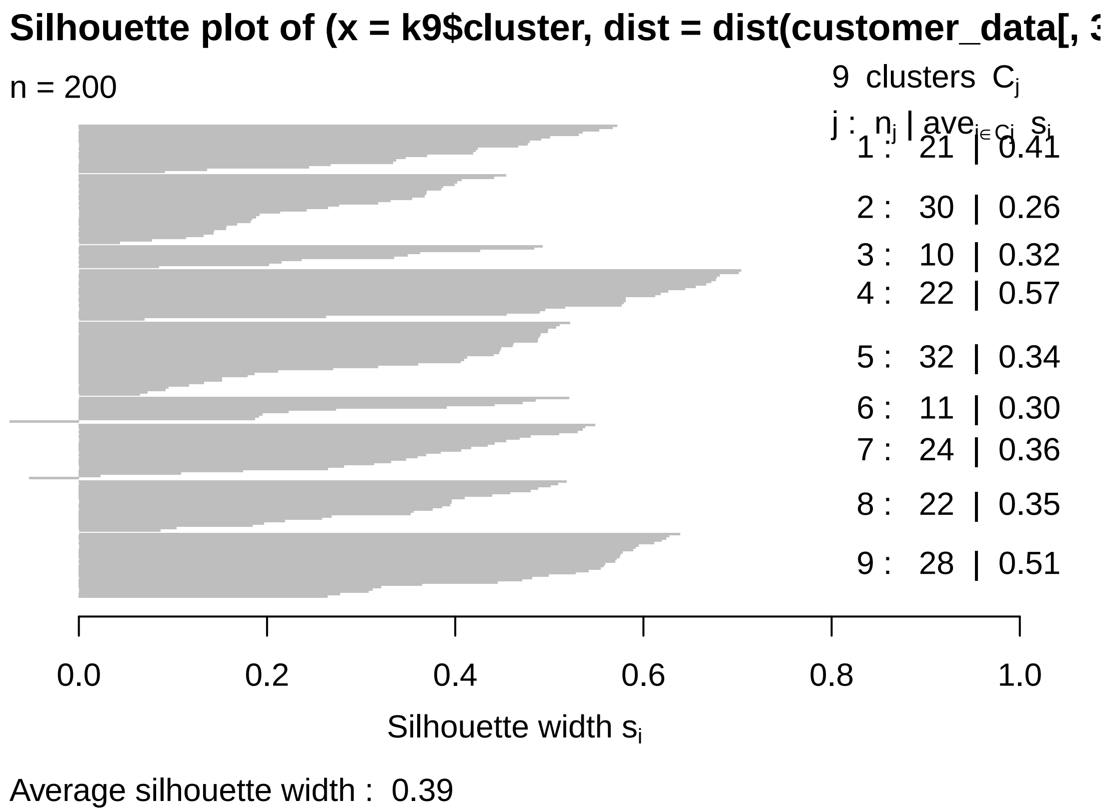



```
k7<-kmeans(customer_data[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")
s7<-plot(silhouette(k7$cluster,dist(customer_data[,3:5],"euclidean")))
```

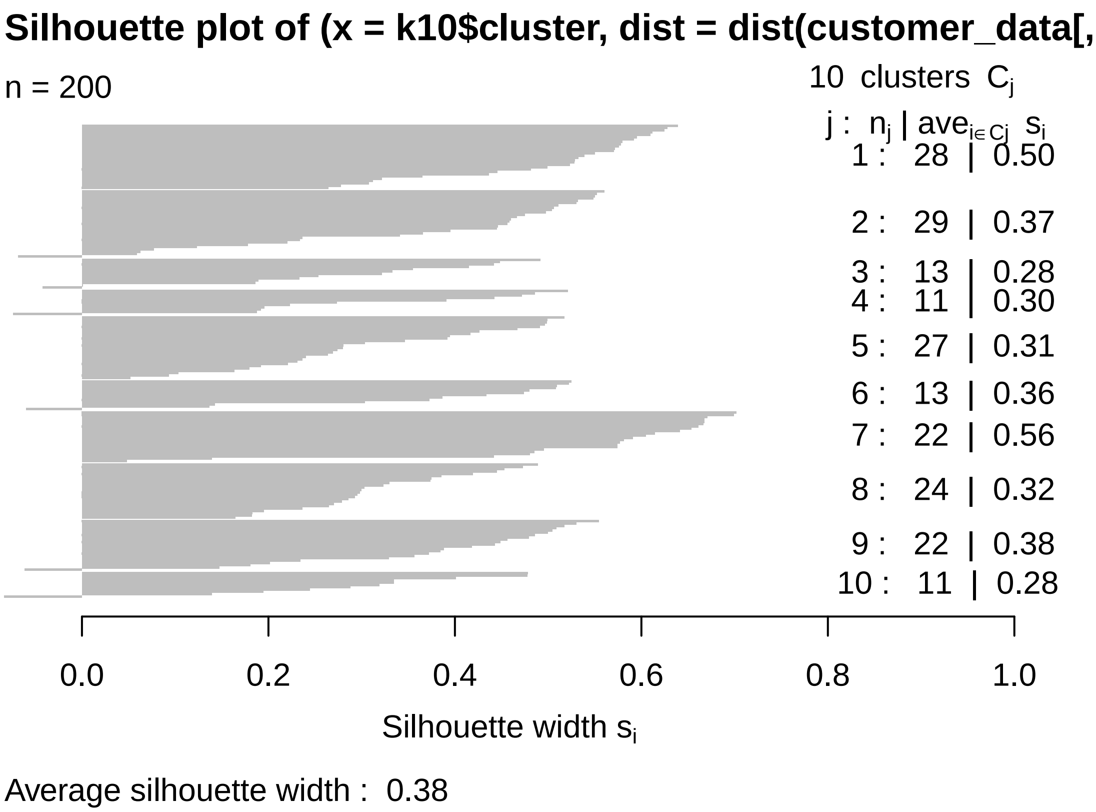

```
k8<-kmeans(customer_data[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")
s8<-plot(silhouette(k8$cluster,dist(customer_data[,3:5],"euclidean")))
```



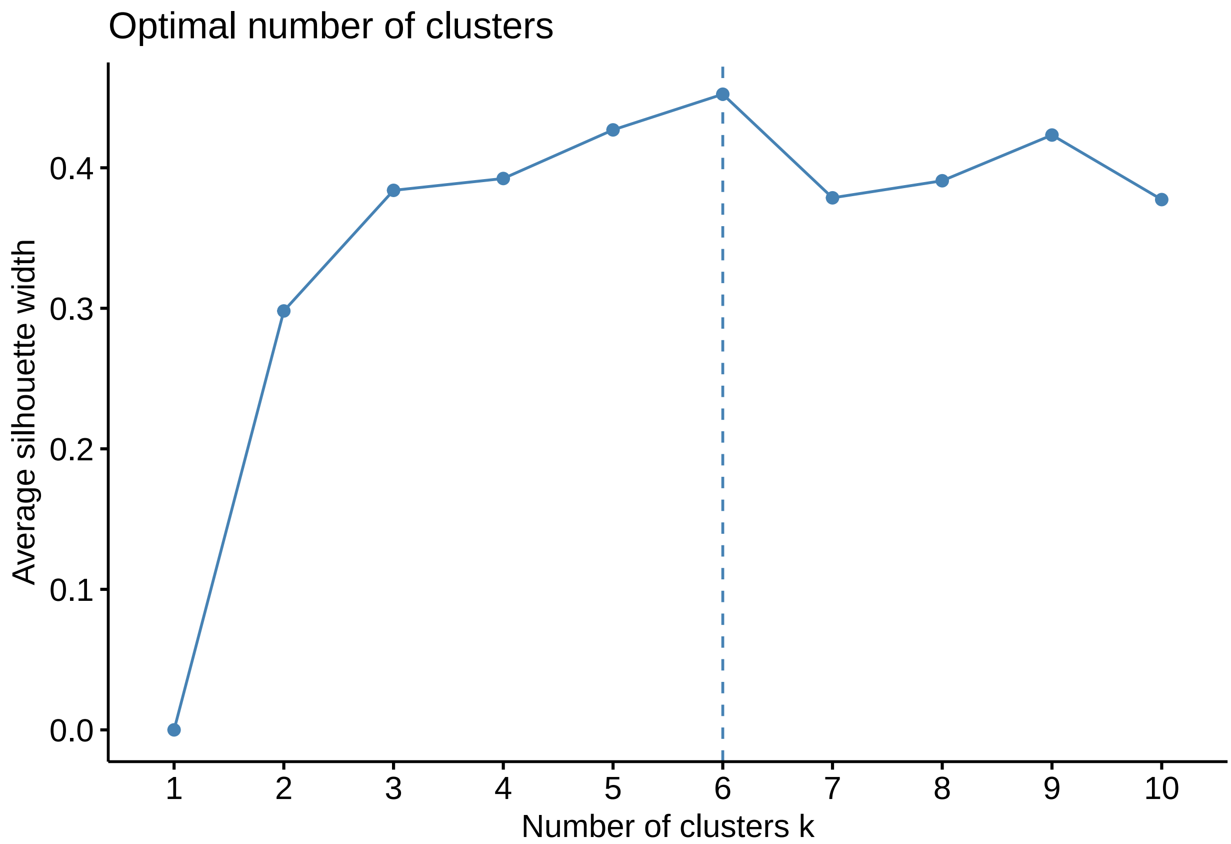

```
k9<-kmeans(customer_data[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")
s9<-plot(silhouette(k9$cluster,dist(customer_data[,3:5],"euclidean")))
```



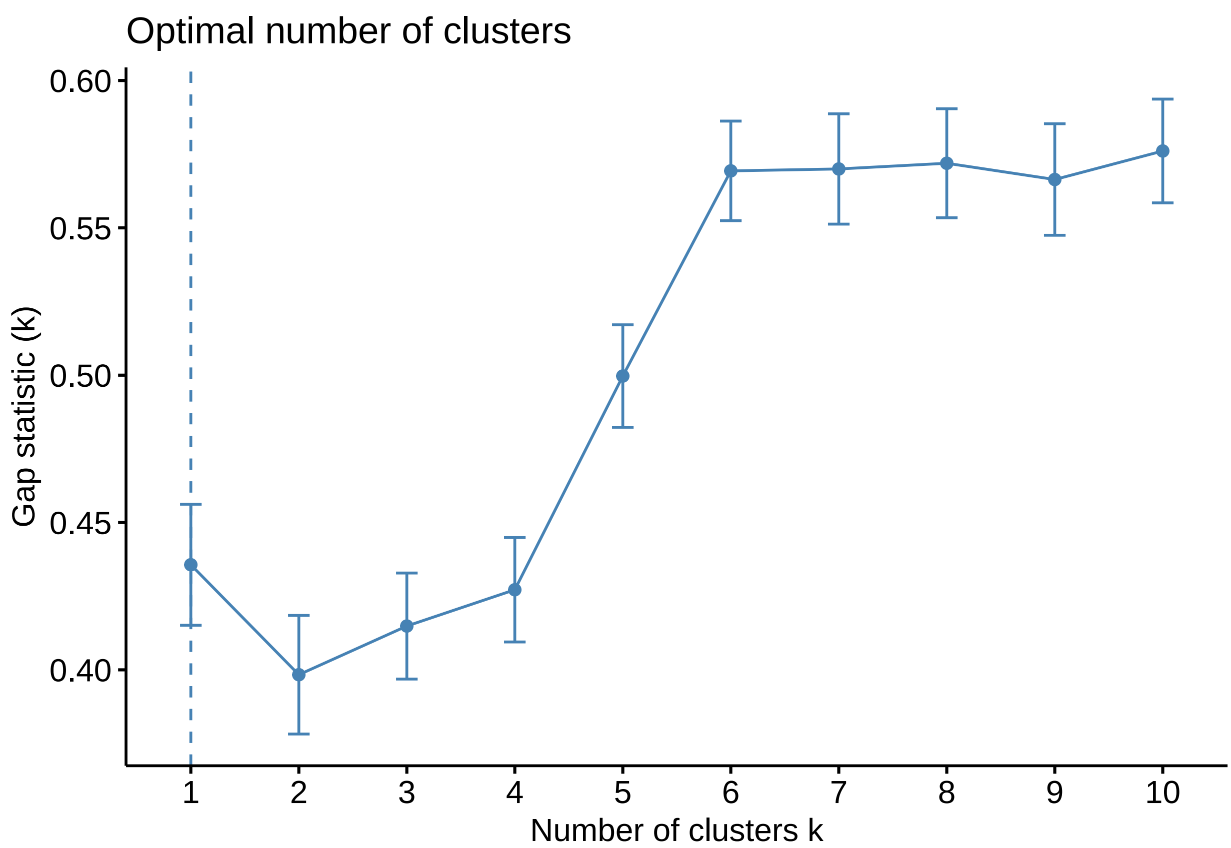

```
k10<-kmeans(customer_data[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")
s10<-plot(silhouette(k10$cluster,dist(customer_data[,3:5],"euclidean")))
```




```
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 2 3
##  [38] 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 1 1 1 1
##  [75] 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 4 1 1 1 1 1
## [112] 4 4 4 4 4 1 1 1 1 4 4 4 6 4 6 5 6 5 6 5 6 4 6 5 6 5 6 5 6 5 6 4 6 5 6 5 6
## [149] 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 5 6 5 6 5
## [186] 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 13972.359
##  (between_SS / total_SS =  81.1 %)
##
```

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

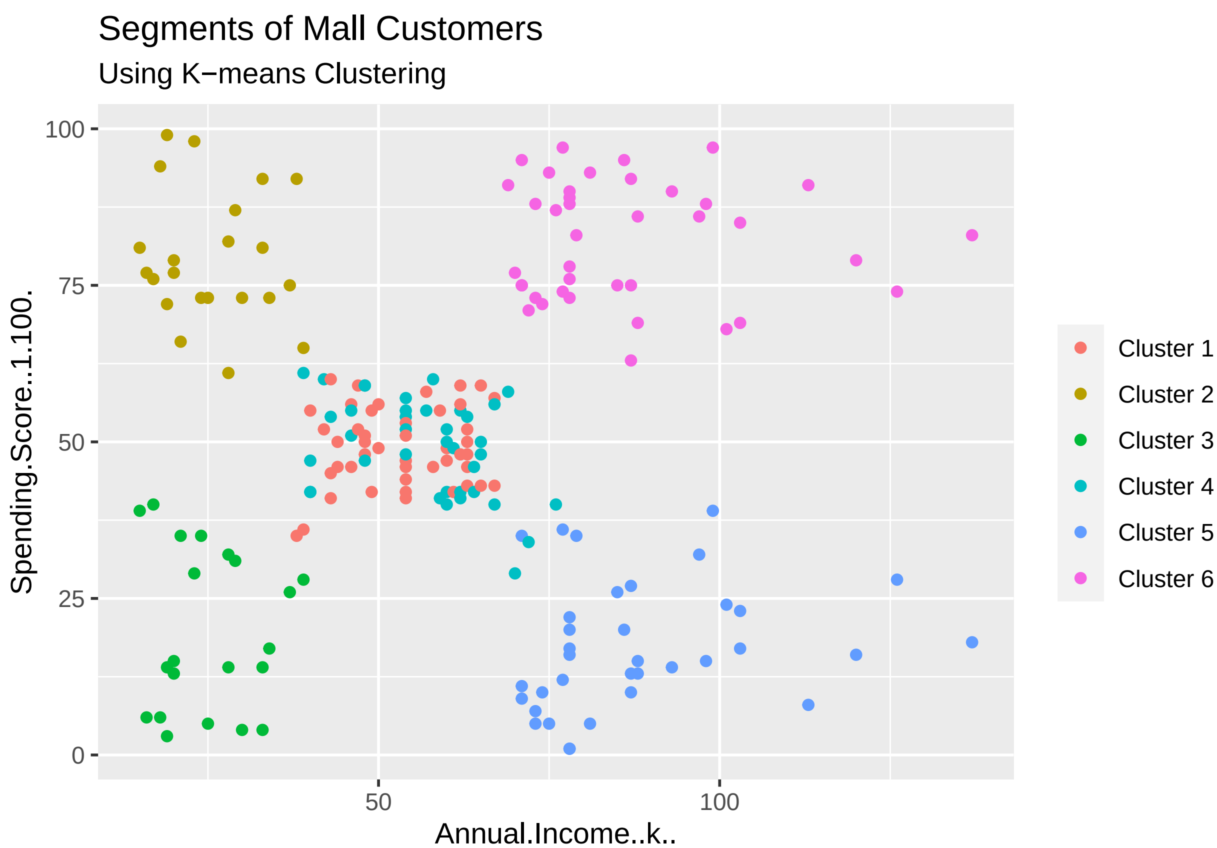
```r
pcclust=prcomp(customer_data[,3:5],scale=FALSE) #principal component analysis
summary(pcclust)
```

```
## 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
```

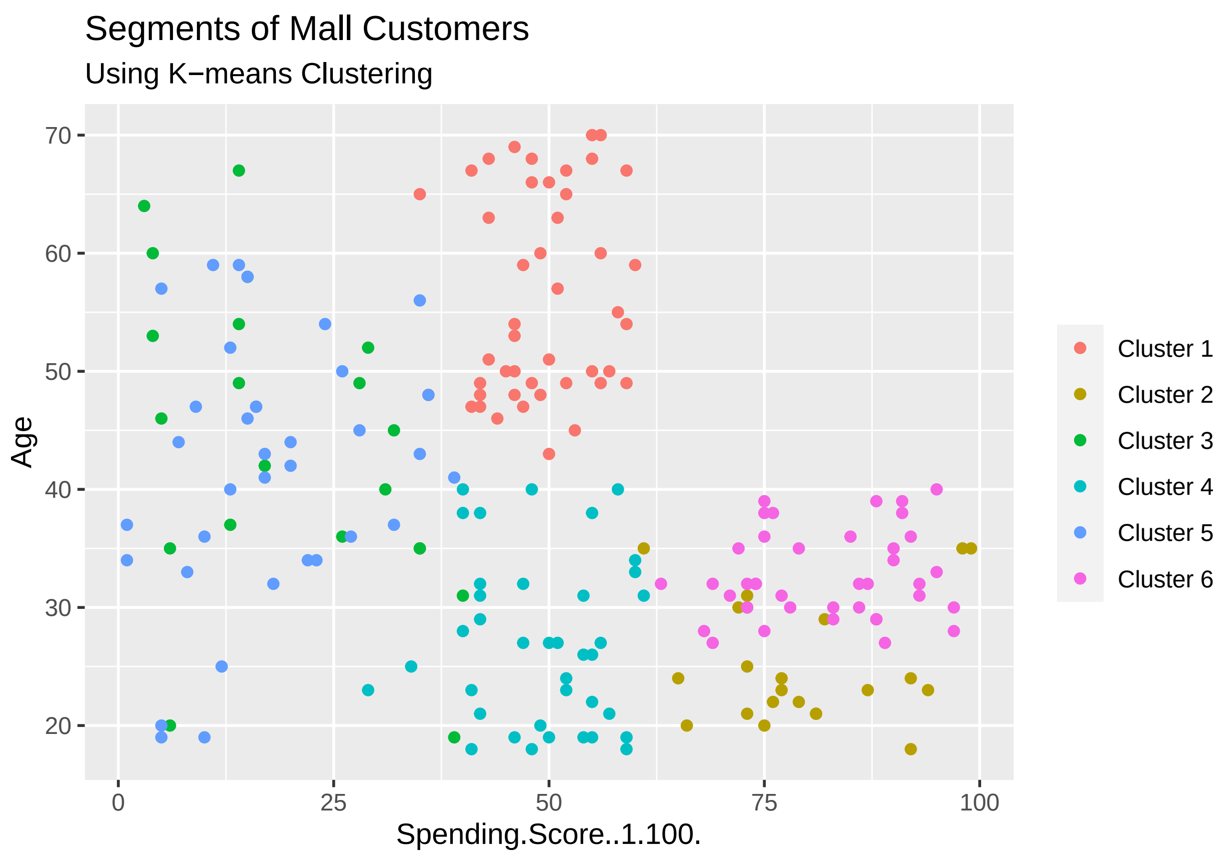
```r
pcclust$rotation[,1:2]
```

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

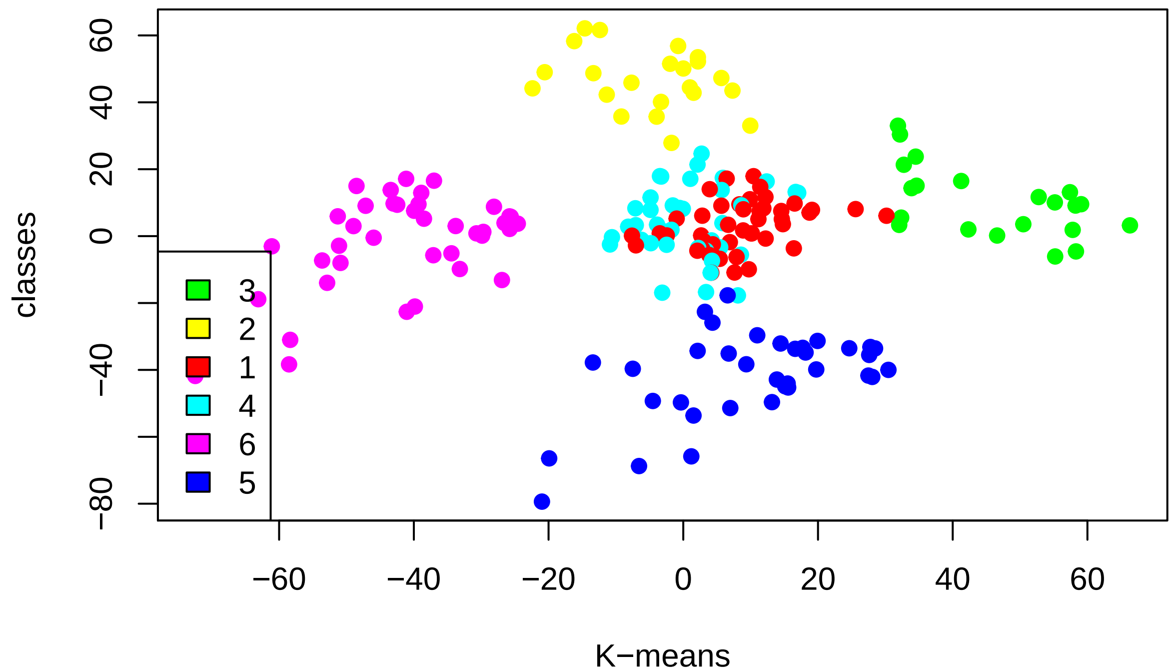
```r
set.seed(1)
ggplot(customer_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name=" ",
              breaks=c("1", "2", "3", "4", "5","6"),
              labels=c("Cluster 1", "Cluster 2", "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
  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)))
```



Limitations

In the process of project execution , I came across following limitations :

1. Customer tend to behave differently, and think differently, at different times or occasions. For example, dietary habits and preferences vary by occasion: Friday night diner is different from lunch during the week. Existing dataset doesn't have any variable with any behavarial attribute of the customer.

2. Shopping score can be derrived from shopping history in conjuction with shopping event but existing dataset feeds shopping score directly without any dependent variable being captured.

3. Data size to get more accurate reading and predictions.

Concluding Remarks

Project is developed using unsupervised ML technique **(KMeans Clustering Algorithm)** in the simplest form. Project mimics supermarket shopping mall data which can be acquired through membership cards , Dataset provides basic data about customers like Customer ID, age, gender, annual income and spending score.

K-Menas clustering helps to segment group of customer on the basis of their age, gender, income and shopping score along with various visual presentations for marketing team to identify prospective customers to plan their strategy.

Project scope is limited to shopping mall but this approach can be reused for other domains by factoring variables of respective datasets to conduct unsupervised ML for respective marketing team.