# MACHINE LEARNING

ISHIKA NISHA

11/7/2021

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

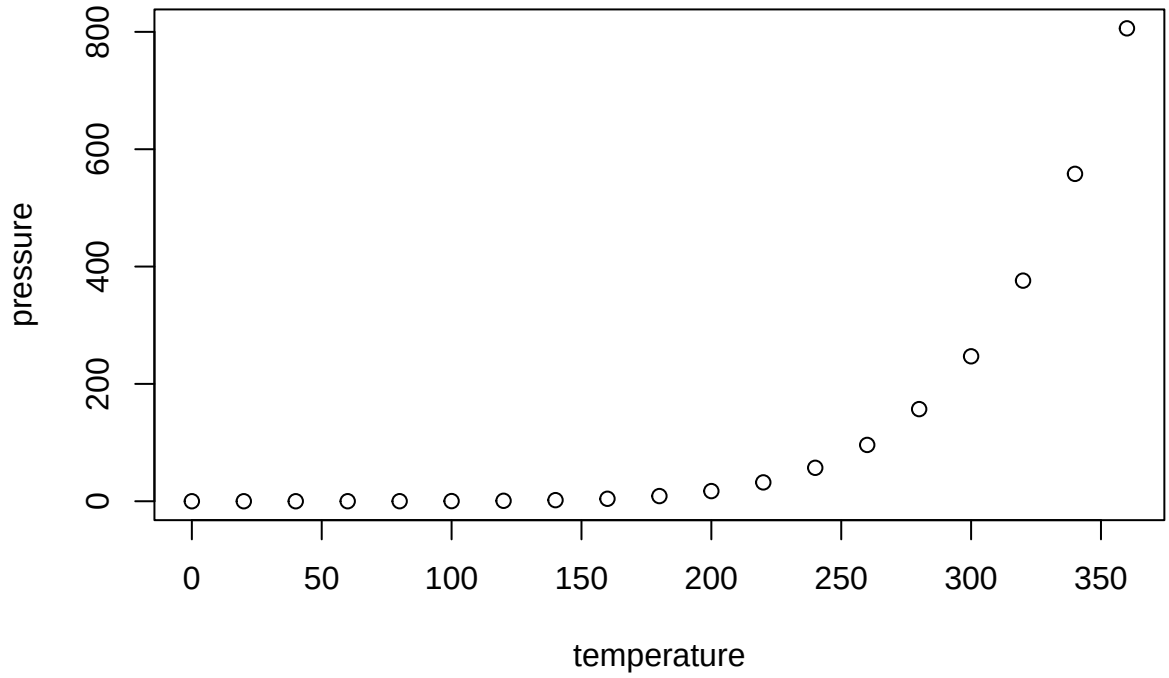
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

## Including Plots

You can also embed plots, for example:




**Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.**

title: "MACHINE LEARNING 4" author: "ISHIKA NISHA" date: "11/7/2021" output: html_document

#Importing the required packages

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

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

```
## Loading required package: lattice
```

```
library(factoextra)
```

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.5     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
set.seed(123)
```

#set working directory

```
getwd()
```

```
## [1] "/Users/ishika/Documents/assignment"
```

## Reading the Pharmaceuticals Csv file

```
Assign <- read.csv('Pharmaceuticals.csv')
```

```
summary(Assign)
```

```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio          ROE             ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median :0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100
```

```
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
##  Min.   :-3.17   Min.   : 2.6      Length:21             Length:21
##  1st Qu.: 6.38   1st Qu.:11.2      Class :character      Class :character
##  Median : 9.37   Median :16.1      Mode  :character      Mode  :character
##  Mean   :13.37   Mean   :15.7
##  3rd Qu.:21.87   3rd Qu.:21.1
##  Max.   :34.21   Max.   :25.5
##    Exchange
##  Length:21
##  Class :character
##  Mode  :character
##
##
##
```

#cheaking for null values

```
colSums(is.na(Assign))
```

```
##                Symbol                  Name            Market_Cap
##                     0                     0                     0
##                  Beta              PE_Ratio                   ROE
##                     0                     0                     0
##                   ROA        Asset_Turnover              Leverage
##                     0                     0                     0
##            Rev_Growth     Net_Profit_Margin Median_Recommendation
##                     0                     0                     0
##              Location              Exchange
##                     0                     0
```

#cheacking for numerical variables

```
head(Assign)
```

```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8            0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5            0.9
## 3    AHM        Amersham plc       6.30 0.46     20.7 14.9  7.8            0.9
## 4    AZN     AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4            0.9
## 5    AVE             Aventis      47.16 0.32     20.1 21.8  7.5            0.6
## 6    BAY           Bayer AG       16.90 1.11     27.9  3.9  1.4            0.6
##   Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location Exchange
## 1     0.42       7.54              16.1          Moderate Buy       US     NYSE
## 2     0.60       9.16               5.5          Moderate Buy   CANADA     NYSE
## 3     0.27       7.05              11.2            Strong Buy       UK     NYSE
## 4     0.00      15.00              18.0         Moderate Sell       UK     NYSE
## 5     0.34      26.81              12.9          Moderate Buy   FRANCE     NYSE
## 6     0.00      -3.17               2.6                 Hold  GERMANY     NYSE
```

#A Using only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

```
Assign_Numeric_Values <- Assign[,c(3:11)]#Columns upon which we want to cluster our datas.
Assign_Numeric_Values
```

```
##    Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1       68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## 2        7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## 3        6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4       67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 5       47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 6       16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
## 7       51.33 0.50     13.9 34.8 15.1            0.9     0.57       2.70
## 8        0.41 0.85     26.0 24.1  4.3            0.6     3.51       6.38
## 9        0.78 1.08      3.6 15.1  5.1            0.3     1.07      34.21
## 10      73.84 0.18     27.9 31.0 13.5            0.6     0.53       6.21
## 11     122.11 0.35     18.0 62.9 20.3            1.0     0.34      21.87
## 12       2.60 0.65     19.9 21.4  6.8            0.6     1.45      13.99
## 13     173.93 0.46     28.4 28.6 16.3            0.9     0.10       9.37
## 14       1.20 0.75     28.6 11.2  5.4            0.3     0.93      30.37
## 15     132.56 0.46     18.9 40.6 15.0            1.1     0.28      17.35
## 16      96.65 0.19     21.6 17.9 11.2            0.5     0.06      -2.69
## 17     199.47 0.65     23.6 45.6 19.2            0.8     0.16      25.54
## 18      56.24 0.40     56.5 13.5  5.7            0.6     0.35      15.00
## 19      34.10 0.51     18.9 22.6 13.3            0.8     0.00       8.56
## 20       3.26 0.24     18.4 10.2  6.8            0.5     0.20      29.18
## 21      48.19 0.63     13.1 54.9 13.4            0.6     1.12       0.36
##    Net_Profit_Margin
## 1               16.1
## 2                5.5
## 3               11.2
## 4               18.0
## 5               12.9
## 6                2.6
## 7               20.6
## 8                7.5
## 9               13.3
## 10              23.4
## 11              21.1
## 12              11.0
## 13              17.9
## 14              21.3
## 15              14.1
## 16              22.4
## 17              25.2
## 18               7.3
## 19              17.6
## 20              15.1
## 21              25.5
```

```
summary(Assign_Numeric_Values)#summary of the extracted columns
```

```
##    Market_Cap          Beta            PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
```

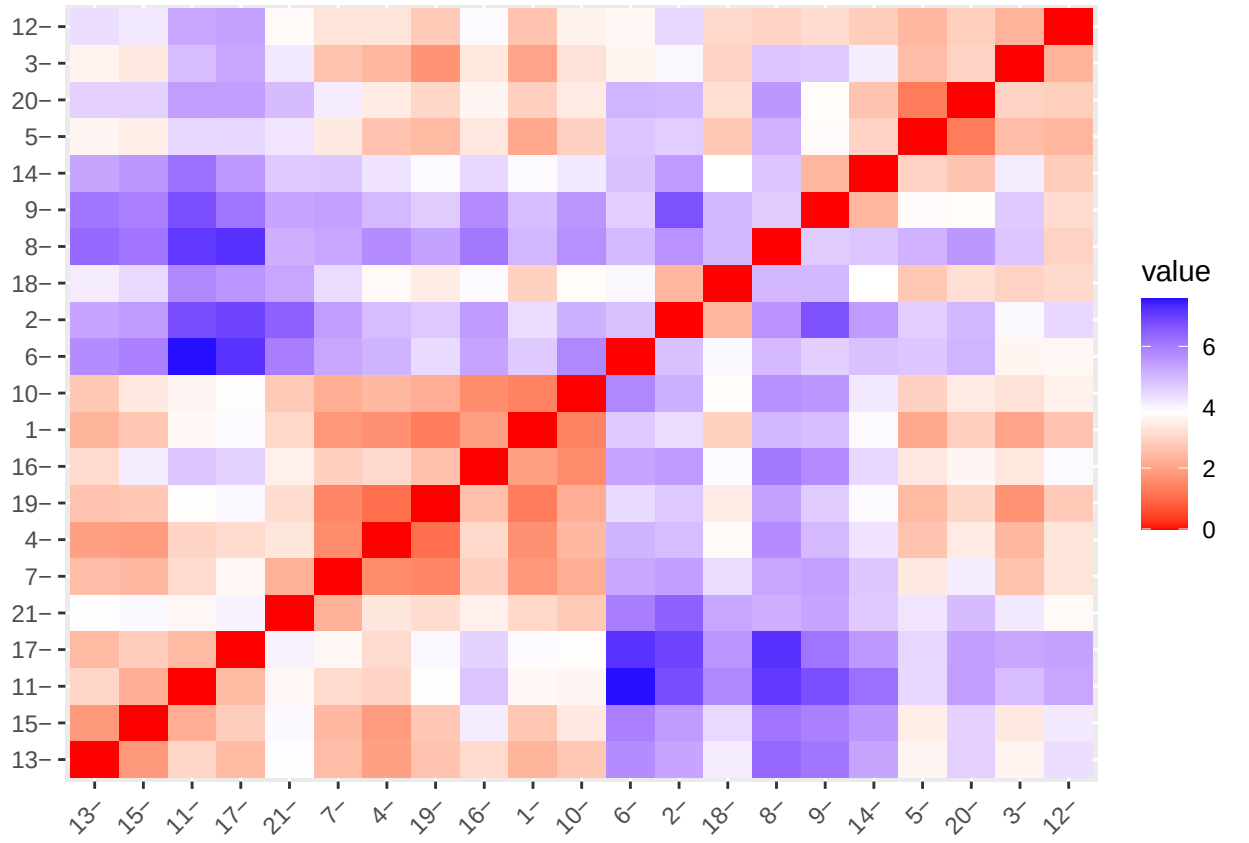
```
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

#scaling the numeric variables

```
Scale_Assign <-scale(Assign_Numeric_Values)
D_Assign <- get_dist(Scale_Assign)
```

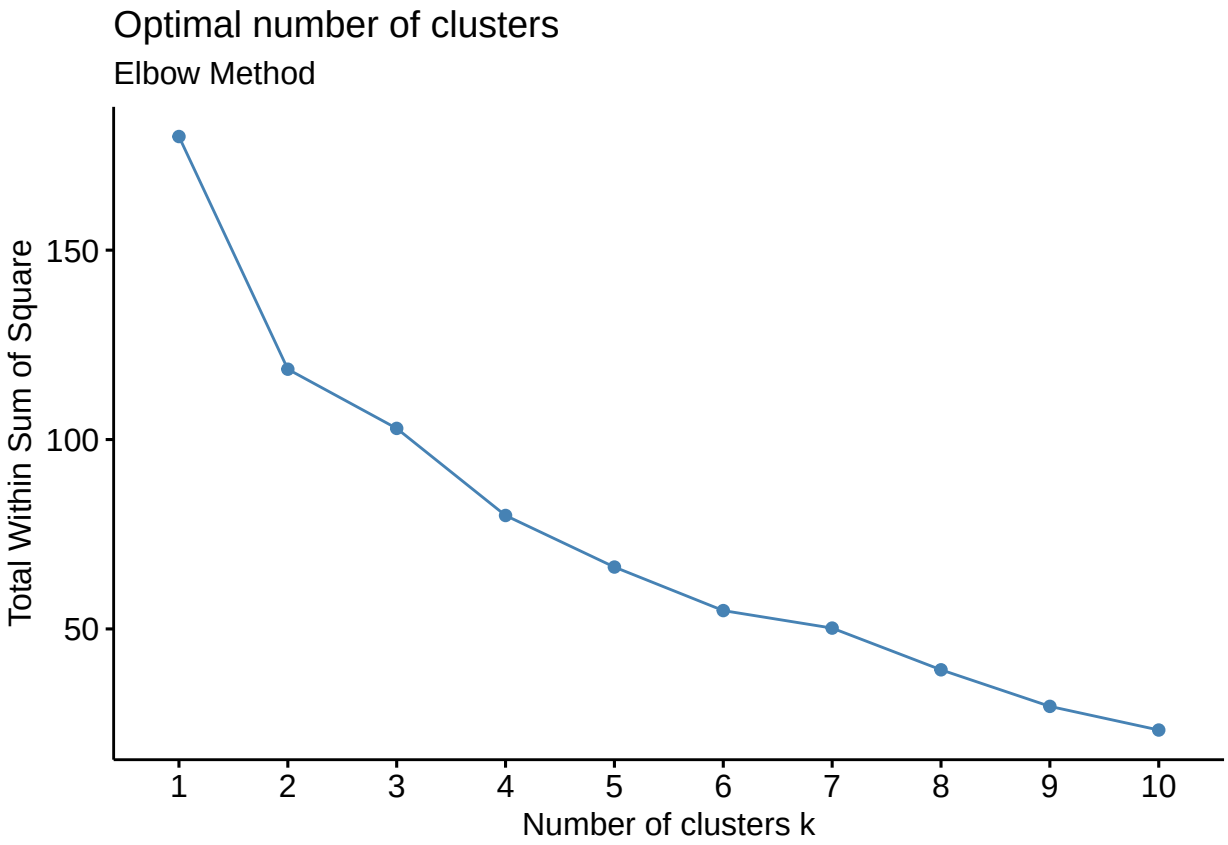
#To view and understand the distance matrix

```
fviz_dist(D_Assign)
```




# Estimating the number of clusters # Using the Elbow Method on scaled data to determine the value of k

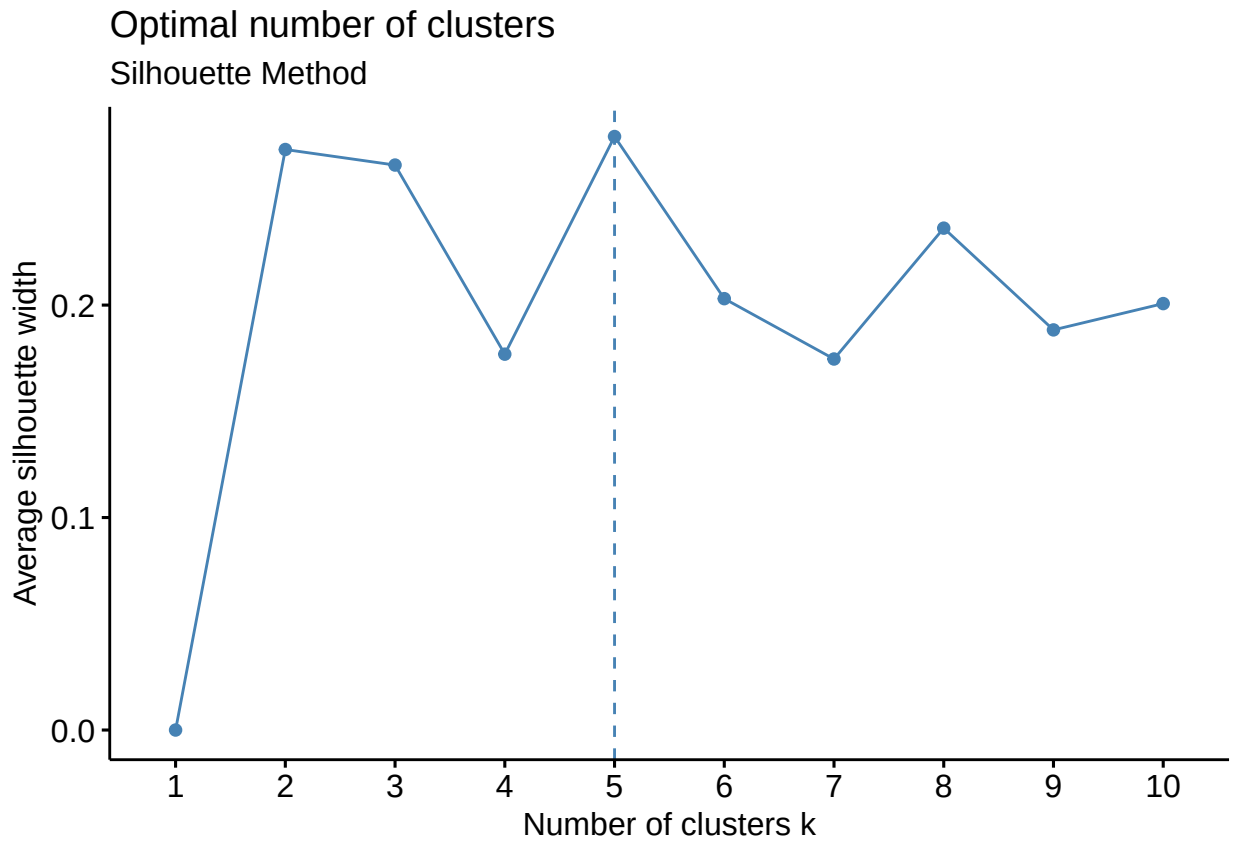
```
fviz_nbclust(Scale_Assign,FUNcluster = kmeans,method = "wss")+labs(subtitle="Elbow Method")
```




According to the Elbow Method k=2

## Silhouette Method on scaled data to determine the number of clusters, Measures of Simirality and ranges

```
fviz_nbclust(Scale_Assign,FUNcluster = kmeans,method = "silhouette")+labs(subtitle="Silhouette Method")
```




The plots reveal that 5 clusters are sufficient.

```
set.seed(10)
Kmeans_D <- kmeans(Scale_Assign,centers=5,nstart=25) #k=5
Kmeans_D$centers   #Centroids
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 2 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 3 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 5 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.46807818  0.4671788       0.591242521
## 2 -0.27449312 -0.7041516       0.556954446
## 3  0.06308085  1.5180158      -0.006893899
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.14170336 -0.1168459      -1.416514761
```

#Size of each cluster

```
Kmeans_D$size
```

```
## [1] 4 8 4 3 2
```

#Finding out the cluster of 8th observation in the dataset, we can similarly find the different observations of the dataset.

```
Kmeans_D$cluster[8]
```

```
## [1] 4
```

#Vizualizing the clusters

```
fviz_cluster(Kmeans_D,data=Scale_Assign)
```


```
#K-Means Cluster Analysis - Fit the data with 5 clusters
Kmeans_N <- kmeans(Scale_Assign, 5)
aggregate(Scale_Assign,by=list(Kmeans_N$cluster),FUN=mean)
```

```
##   Group.1 Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1  0.6733825 -0.3586419 -0.27635122  0.6565978  0.8344159
## 2       2 -0.9767669  1.2630872  0.03299122 -0.1123792 -1.1677918
## 3       3 -0.5246281  0.4451409  1.84984387 -1.0404550 -1.1865838
## 4       4 -0.7307042 -0.4214928 -0.34867046 -0.5780744 -0.6181243
## 5       5 -0.9668697  1.5162611 -0.57398880 -0.8382671 -0.9892673
##   Asset_Turnover    Leverage Rev_Growth Net_Profit_Margin
## 1   4.612656e-01 -0.33310678 -0.2902163         0.6823310
## 2  -4.612656e-01  3.74279705 -0.6327607        -1.2488842
```

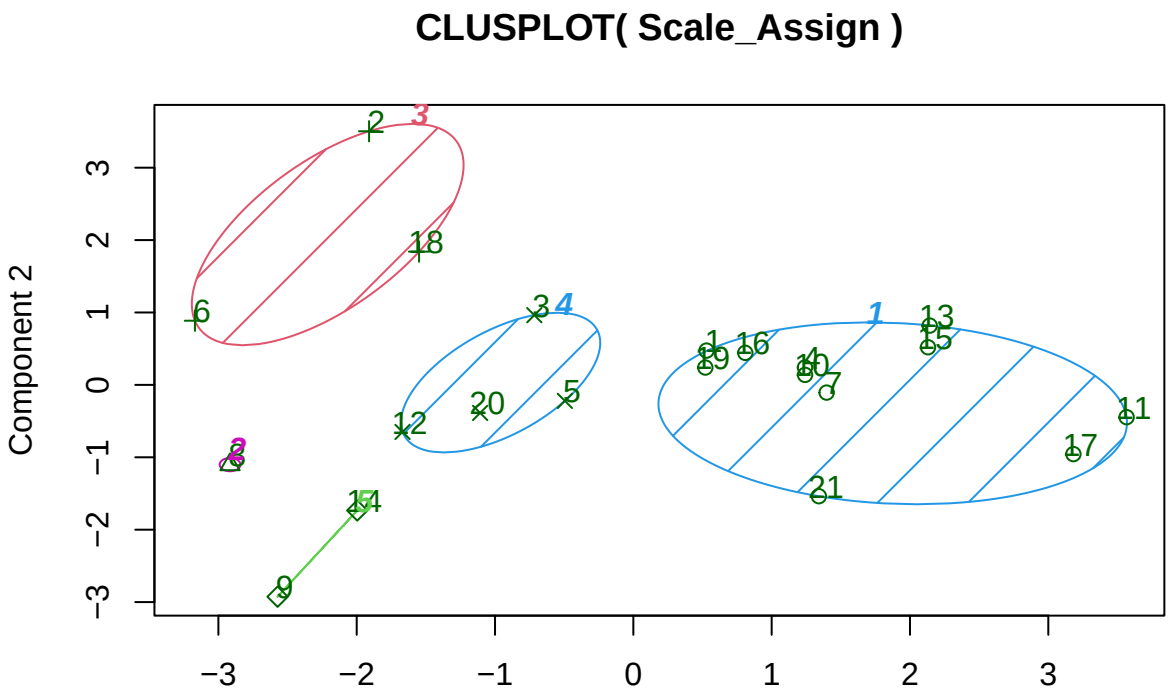
```
## 3  -3.330669e-16 -0.34435439 -0.5769454        -1.6095439
## 4  -2.306328e-01 -0.02651224  0.5327995        -0.4793074
## 5  -1.845062e+00  0.53024482  1.7123890         0.2445520
```

```
New_Assign <- as.data.frame(Scale_Assign, Kmeans_D$cluster)
New_Assign
```

```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## X2    0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## X5   -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## X2.1 -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## X2.2  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## X3   -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## X4   -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
## X2.3 -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498   9.225312e-01
## X4.1 -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918  -4.612656e-01
## X3.1 -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553  -1.845062e+00
## X2.4  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770  -4.612656e-01
## X1    1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364   1.383797e+00
## X4.2 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905  -4.612656e-01
## X1.1  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544   9.225312e-01
## X3.2 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792  -1.845062e+00
## X1.2  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577   1.845062e+00
## X2.5  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598  -9.225312e-01
## X1.3  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239   4.612656e-01
## X5.1 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030  -4.612656e-01
## X2.6 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929   4.612656e-01
## X3.3 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905  -9.225312e-01
## X2.7 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849  -4.612656e-01
##         Leverage  Rev_Growth Net_Profit_Margin
## X2   -0.21209793 -0.52776752        0.06168225
## X5    0.01828430 -0.38113909       -1.55366706
## X2.1 -0.40408312 -0.57211809       -0.68503583
## X2.2 -0.74965647  0.14744734        0.35122600
## X3   -0.31449003  1.21638667       -0.42597037
## X4   -0.74965647 -1.49714434       -1.99560225
## X2.3 -0.02011273 -0.96584257        0.74744375
## X4.1  3.74279705 -0.63276071       -1.24888417
## X3.1  0.61983791  1.88617085       -0.36501379
## X2.4 -0.07130879 -0.64814764        1.17413980
## X1   -0.31449003  0.76926048        0.82363947
## X4.2  1.10620040  0.05603085       -0.71551412
## X1.1 -0.62166634 -0.36213170        0.33598685
## X3.2  0.44065173  1.53860717        0.85411776
## X1.2 -0.39128411  0.36014907       -0.24310064
## X2.5 -0.67286239 -1.45369888        1.02174835
## X1.3 -0.54487226  1.10143723        1.44844440
## X5.1 -0.30169102  0.14744734       -1.27936246
## X2.6 -0.74965647 -0.43544591        0.29026942
## X3.3 -0.49367621  1.43089863       -0.09070919
## X2.7  0.68383297 -1.17763919        1.49416183
```

#Visualization Of ClustPlot

```
library(cluster)
clusplot(Scale_Assign ,Kmeans_N$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```




#B. Interpret the clusters with respect to the numerical variables used in forming the clusters.

Cluster 1 - Row 11,17,13,15

Cluster 2 - Row 21,7,10,4,16,19,3,1

Cluster 3 - Row 5,20,14,9

Cluster 4 - Row 12,8,6

Cluster 5 - Row 18,2

As above mention with the help of the following output =aggregate(Scale_Assign,by=list(Kmeans_N$cluster),FUN=mean) We can observe the followings

Cluster 1 has highest Market_Cap,highest ROE,highest ROA,highest Asset_Turnover,highest Net_Profit_Margin.

Cluster 2 has highest Leverage,lowest Market_Cap, lowest Rev_Growth.

Cluster 3 has highest PE_Ratio,lowest_ROE,lowest ROA,lowest Net_Profit_Margin,lowest Leverage.

Cluster 4 has lowest Beta.

Cluster 5 has highest Beta,highest Rev_Growth,lowest PE_Ratio,Lowest Asset_Turnover.

#C Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

Cluster 1 has highest Market_Cap,highest ROE,highest ROA,highest Asset_Turnover,highest Net_Profit_Margin with most of the cases of Median_Recommendation of Hold, Country US and exchange NYSE

Cluster 2 has highest Leverage with most of the cases of Median_Recommendation of Hold,Country Us and Exchange NYSE

Cluster 3 has highest PE_Ratio with most of Median_Recommendation of equal Moderate buy and Moderate sell,and Exchange NYSE

Cluster 4 has lowest Beta with most of the Median_Recommendation of Hold, and Country US

Cluster 5 has highest Beta,highest Rev_Growth with most of the Median_Recommendation of equal Hold and Moderate Buy, and Exchange NYSE

So we can Conclude( in terms of Median_Recommendation,Country,Exchange) Cluster 1,2,4 has most of the Median_Recommendation of Hold, Country US,Exchange NYSE(Only in Cluster 1 and 2) Cluster 3 has equal Median_Recommendation of Moderate Buy and Moderate Sell, Exchange NYSE Cluster 5 has equal Median_Recommendation of Hold and Moderate Buy, Exchange NYSE

#D Provide an appropriate name for each cluster using any or all of the variables in the dataset.

Cluster 1 Stellar (has highest Market_Cap,highest ROE,highest ROA,highest Asset_Turnover,highest Net_Profit_Margin.)

Cluster 2 Low (has highest Leverage,lowest Market_Cap, lowest Rev_Growth.)

Cluster 3 Least (has highest PE_Ratio,lowest_ROE,lowest ROA,lowest Net_Profit_Margin,lowest Leverage.)

Cluster 4 Medium (has only lowest Beta.)

Cluster 5 Strong (has highest Beta,highest Rev_Growth,lowest PE_Ratio,Lowest Asset_Turnover.)

I have name the clusters taking into consideration of only the numeric columns The orders are as follows 1.Stellar 2.Strong 3.Medium 4.Low 5.Least