# Sbaig1_Assignment4

```
setwd("C:/Users/shari/OneDrive/Desktop/Business Analytics/Sem 1/Machine
Learning/ML_Assignment4")

Pharma <- read.csv("C:/Users/shari/OneDrive/Desktop/Business Analytics/Sem
1/Machine Learning/ML_Assignment4/Pharmaceuticals.csv")

library(factoextra)

library(ISLR)
library(tidyverse)

library(caret)

library(grid)
library(modeltools)

library(stats4)
library(lattice)
library(flexclust)

library(cluster)

set.seed(123)
head(Pharma)

##   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
```

**##Use cluster analysis to explore and analyze the given dataset as follows:**

**## A. Use 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.**

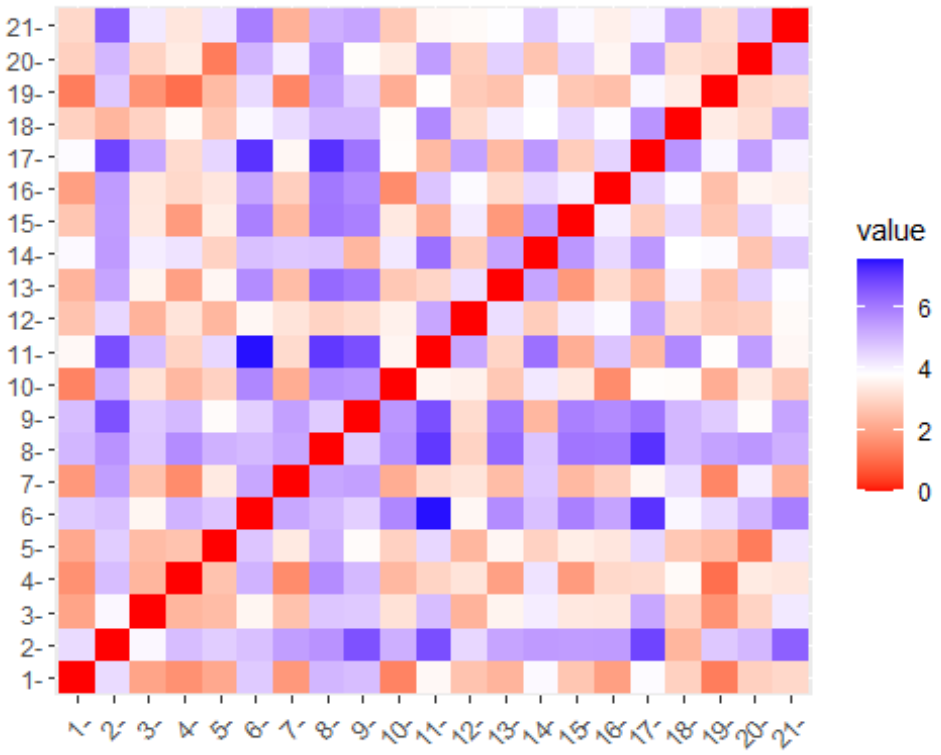
```
Numeric<- Pharma[,3:11]
head(Numeric)

##   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
##   Net_Profit_Margin
## 1              16.1
## 2               5.5
## 3              11.2
## 4              18.0
## 5              12.9
## 6               2.6
```

**Normalizing the data framewith Range and Scale method.**

```
Numeric <- scale(Numeric)
distance_Numeric <-get_dist(Numeric, method = "euclidean", stand = FALSE)
```

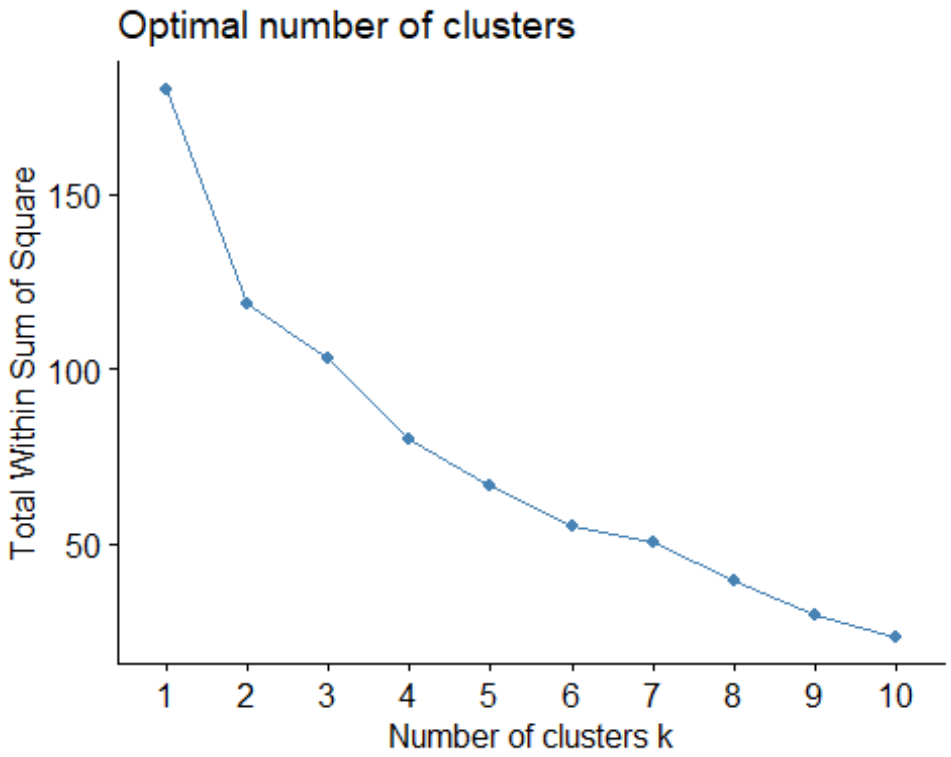
**In the below code we can see the distance between each observation and color code is provided depending on the observation values.**

```
fviz_dist(distance_Numeric, order = FALSE, show_labels = TRUE, lab_size = NULL, gradient =
list(low = "red", mid = "white", high = "blue"))
```




**Using the elbow method below to find the optimal k**

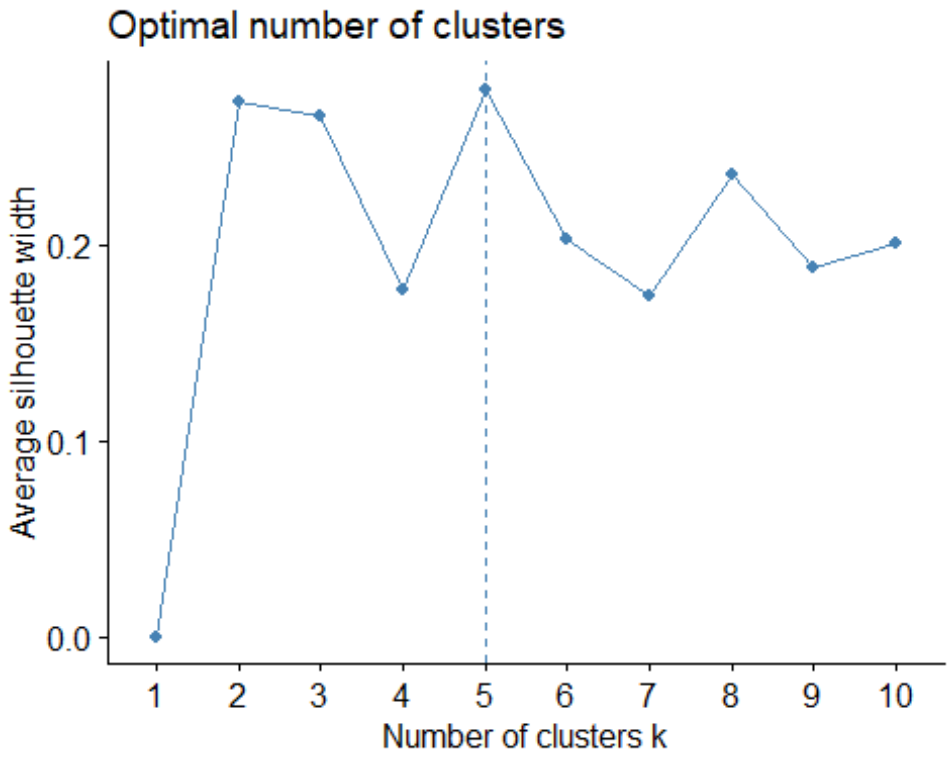
```
elbowpharma <- scale(Numeric)
fviz_nbclust(Numeric,kmeans,method="wss")
```




**Looking at the above graph we can see that there is an elbow at 2, however it is still unclear due to less sharpness in the graphical representation.**

**Using the Silhouette method below**

```
fviz_nbclust(Numeric,kmeans,method="silhouette")
```




**We will use the Silhouette method becuase of the clear representation of K=5.**

```
k <- kmeans(Numeric, centers = 5, nstart = 25)
k

## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap      Beta    PE_Ratio       ROE       ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915     0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478    -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951     0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431     1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428    -1.2684804
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
```

```
##  [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1
##
## Within cluster sum of squares by cluster:
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"     "centers"     "totss"       "withinss"    "tot.withinss"
## [6] "betweenss"   "size"        "iter"        "ifault"
```

**Finding below cluster center for all rows and colomns**

```
k$centers
```

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

**Number of observation in each cluster**

```
k$size
```

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

**Observation and their respective cluster label.**

```
k$cluster[c(21,20,19)]
```

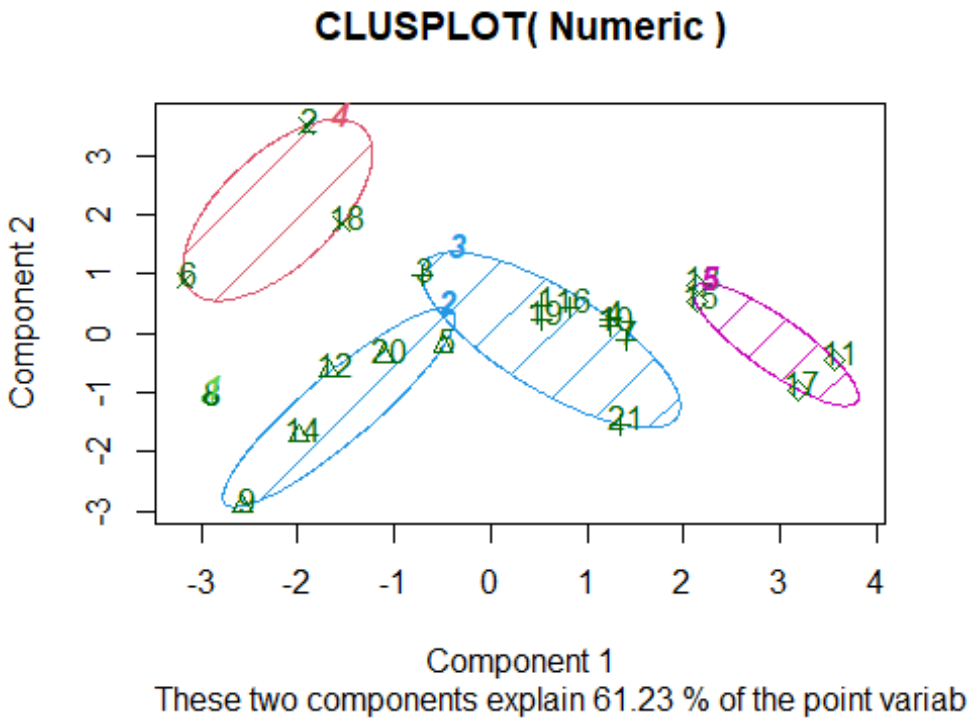
```
## [1] 1 5 1
```

**Applying Kmeans clustering with 5 clusters of the size 4,2,4,8,4.**

```
fviz_cluster(k, data=Numeric)
```


**On the cluster graph above, we can see that there are 5 clusters, each with its own color and shape. The center of the cluster is the centriod or the center point. We have reached the final center points after 25 restarts as there is no change until and unless the new data is added.**

```
Fi <- kmeans(Numeric,5)
clusplot(Numeric, Fi$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```




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

```
aggregate(Numeric,by=list(Fi$cluster),FUN=mean)

##   Group.1  Market_Cap      Beta    PE_Ratio       ROE       ROA
## 1       1 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918
## 2       2 -0.79605926  0.3205014 -0.45014035 -0.6533148 -0.7881923
## 3       3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 4       4 -0.52462814  0.4451409  1.84984387 -1.0404550 -1.1865838
## 5       5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1  -4.612656e-01  3.7427970 -0.6327607       -1.2488842
## 2  -1.107037e+00  0.2717048  1.2256188       -0.1486179
## 3   1.729746e-01 -0.2744931 -0.7041516        0.5569544
## 4   1.480297e-16 -0.3443544 -0.5769454       -1.6095439
## 5   1.153164e+00 -0.4680782  0.4671788        0.5912425

Num1 <- data.frame(Numeric, Fi$cluster)
Num1

##    Market_Cap       Beta    PE_Ratio       ROE       ROA Asset_Turnover
## 1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## 2  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
```

```
## 3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700     0.9225312
## 4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259     0.9225312
## 5  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461    -0.4612656
## 6  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612    -0.4612656
## 7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498     0.9225312
## 8  -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918    -0.4612656
## 9  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553    -1.8450624
## 10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770    -0.4612656
## 11  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364     1.3837968
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905    -0.4612656
## 13  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544     0.9225312
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792    -1.8450624
## 15  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577     1.8450624
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598    -0.9225312
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239     0.4612656
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030    -0.4612656
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929     0.4612656
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905    -0.9225312
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849    -0.4612656
##      Leverage  Rev_Growth Net_Profit_Margin Fi.cluster
## 1  -0.21209793 -0.52776752       0.06168225         3
## 2   0.01828430 -0.38113909      -1.55366706         4
## 3  -0.40408312 -0.57211809      -0.68503583         3
## 4  -0.74965647  0.14744734       0.35122600         3
## 5  -0.31449003  1.21638667      -0.42597037         2
## 6  -0.74965647 -1.49714434      -1.99560225         4
## 7  -0.02011273 -0.96584257       0.74744375         3
## 8   3.74279705 -0.63276071      -1.24888417         1
## 9   0.61983791  1.88617085      -0.36501379         2
## 10 -0.07130879 -0.64814764       1.17413980         3
## 11 -0.31449003  0.76926048       0.82363947         5
## 12  1.10620040  0.05603085      -0.71551412         2
## 13 -0.62166634 -0.36213170       0.33598685         5
## 14  0.44065173  1.53860717       0.85411776         2
## 15 -0.39128411  0.36014907      -0.24310064         5
## 16 -0.67286239 -1.45369888       1.02174835         3
## 17 -0.54487226  1.10143723       1.44844440         5
## 18 -0.30169102  0.14744734      -1.27936246         4
## 19 -0.74965647 -0.43544591       0.29026942         3
## 20 -0.49367621  1.43089863      -0.09070919         2
## 21  0.68383297 -1.17763919       1.49416183         3
```

**Cluster_1 = has Highest Rev_growth and low leverage and low beta**

**Cluster_2 = has Highest PE ratio, Lowest ROE, Lowest ROA, Lowest Asset Turnover, Lowest Net ProfitMargin**

**Cluster_3 = has Highest Market Cap, Highest ROE, Highest ROA, Highest Asset Turnover.**

**Cluster_4 = has Highest Net Profit Margin, Lowest Beta, Lowest PE Ratio, Lowest Rev growth.**

**Cluster_5 = has Highest Beta, Highest Leverage, Highest Rev growth and Lowest Market Cap.**

**C. Is there a pattern in the clusters with respect to the numerical variables ?**

**1. Based on the average recommended variable, there is a pattern in the cluster.**

**2. Despite having the highest market capitalization, highest ROE, highest ROA, and highest asset turnover, cluster.**

**3. doesn't have a median sales recommendation. 3. Instead, cluster 3 has strong purchase recommendations.**

**4. Most of the time, Cluster 2 with the lowest P / E, ROE, ROA, asset turnover, and net return has pending recommendations.**

**5. Cluster 4, which has the highest net margin, the lowest beta, the lowest PE ratio, and the lowest revenue growth, is most often recommended for hold.**

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

**Cluster_1 - Lowest Leverage cluster and Highest Rev_growth.**

**Cluster_2 - High PE ratio, Low ROE, Low ROA, Low Asset Turnover and Negative Net Profit Margin Cluster**

**Cluster_3 - High Market Cap, ROE, ROA, Asset Turnover cluster**

**Cluster_4 - High Net Profit Margin, High Low Beta and Negative Rev growth cluster**

**Cluster_5 - High Beta, Negative Leverage, Low Rev growth and Low Market Cap cluster**