MIS-64060-001(A4)

Kiran Kour

2022-11-01

Installing required Packages

```
#install.packages("tidyverse")
#install.packages("factoextra")
#install.packages("flexclust")
#install.packages("cluster")
#install.packages("gridExtra")
#install.packages("ggplot2")
#install.packages("cowplot")
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

```
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
library(cluster)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ISLR)
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.2.2
```

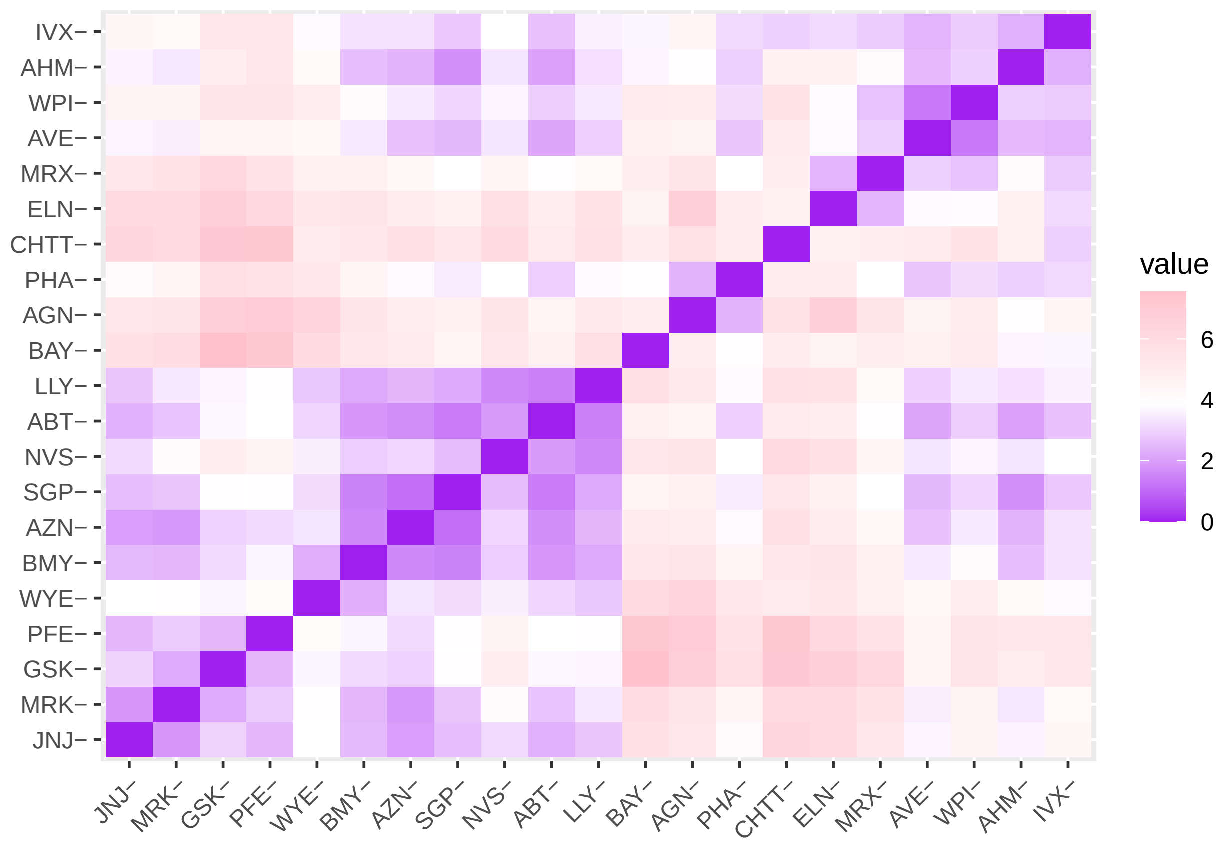
Importing the dataset, selecting the numericals variables and normalizing the dataset

```r
Pharma <- read.csv("Pharmaceuticals.csv")
rownames(Pharma)<- Pharma$Symbol
Pharma1 <- Pharma[,-c(1,2,12,13,14)]
Ph_norm <- scale(Pharma1)
summary(Ph_norm)
```

```
##    Market_Cap           Beta            PE_Ratio            ROE
##  Min.   :-0.9768   Min.   :-1.3466   Min.   :-1.3404   Min.   :-1.4515
##  1st Qu.:-0.8763   1st Qu.:-0.6844   1st Qu.:-0.4023   1st Qu.:-0.7223
##  Median :-0.1614   Median :-0.2560   Median :-0.2429   Median :-0.2118
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.2762   3rd Qu.: 0.4841   3rd Qu.: 0.1495   3rd Qu.: 0.3450
##  Max.   : 2.4200   Max.   : 2.2758   Max.   : 3.4971   Max.   : 2.4597
##       ROA          Asset_Turnover       Leverage          Rev_Growth
##  Min.   :-1.7128   Min.   :-1.8451   Min.   :-0.74966   Min.   :-1.4971
##  1st Qu.:-0.9047   1st Qu.:-0.4613   1st Qu.:-0.54487   1st Qu.:-0.6328
##  Median : 0.1289   Median :-0.4613   Median :-0.31449   Median :-0.3621
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.8430   3rd Qu.: 0.9225   3rd Qu.: 0.01828   3rd Qu.: 0.7693
##  Max.   : 1.8389   Max.   : 1.8451   Max.   : 3.74280   Max.   : 1.8862
##  Net_Profit_Margin
##  Min.   :-1.99560
##  1st Qu.:-0.68504
##  Median : 0.06168
##  Mean   : 0.00000
##  3rd Qu.: 0.82364
##  Max.   : 1.49416
```

Computing and visualizing the distance matrix using the functions get_dist() and fviz_dist(). This enables us to have visual understanding of the dis/similarity of the different data points.

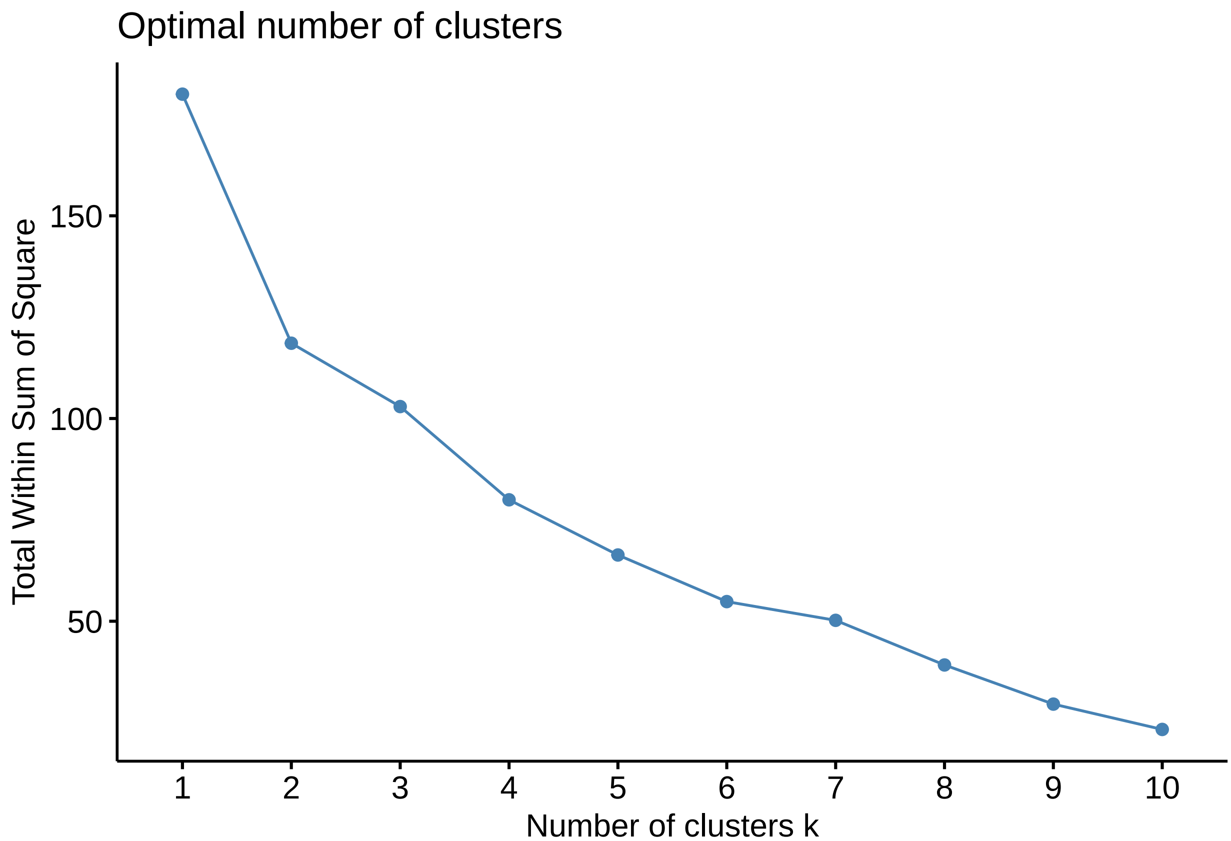
```r
set.seed(420)
distance <- get_dist(Ph_norm)

# displaying a dis/similarity and distance matrix
fviz_dist(distance, gradient = list(low = " purple", mid = "white", high = "pink"))
```



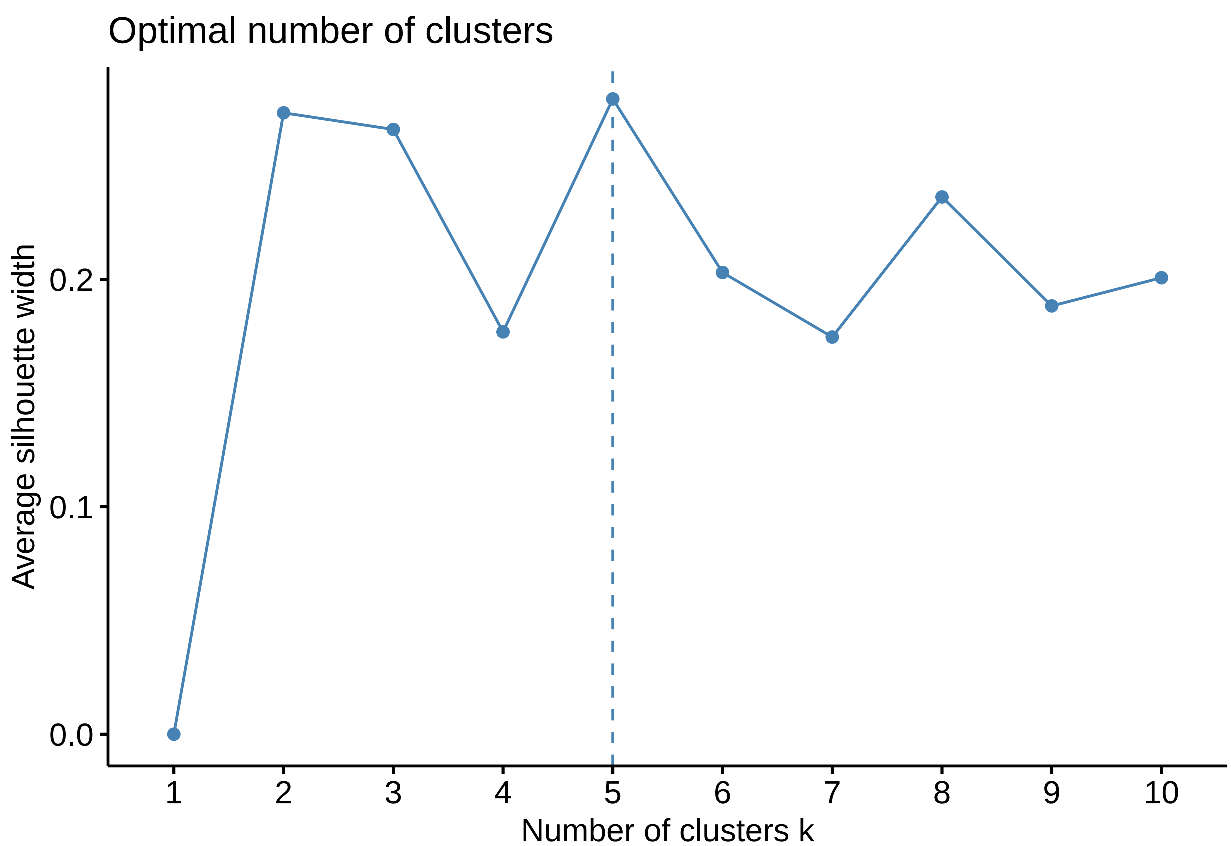
An essential factor in clustering is distance; the distance matrix above shows the similarity or dissimilarity of each pair of observations based on their distance (i.e., purple indicating similarity and pink showing dissimilarity in this specific example). The similarity can decide which clusters should be combined or divided into another. This means points with minimal distance value among them should be in the same cluster.

Using WSS and Silhouete method to find the optimal K value

```
WSS <- fviz_nbclust(Ph_norm,kmeans,method="wss")
WSS
```




```
Silhouette <- fviz_nbclust(Ph_norm,kmeans,method="silhouette")
Silhouette
```



We got the optimal K=2 by employing the WSS method and K=5 by employing the Silhouette method.

Running the kmeans with k=2 which we got by employing the WSS method

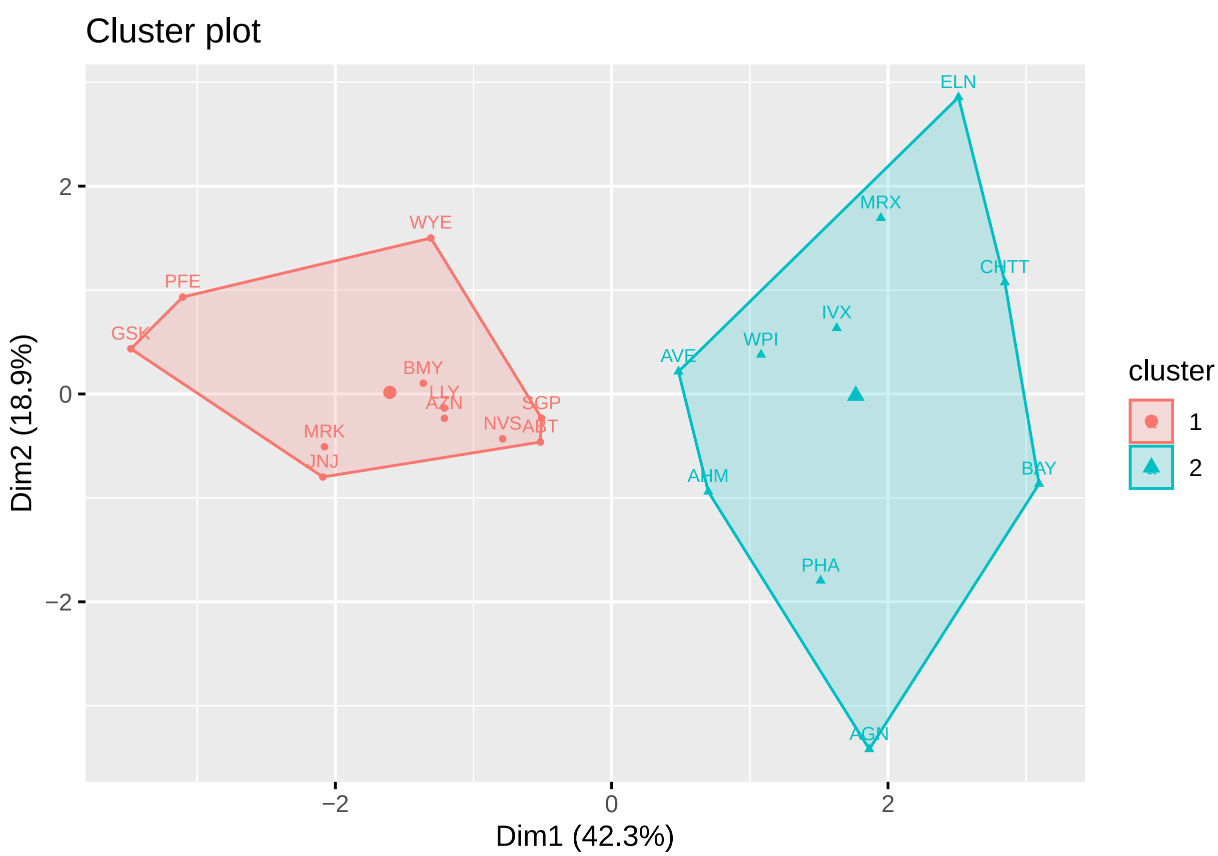
```
k2<- kmeans(Ph_norm, centers=2, nstart = 25)
k2
```

```
## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    2    2    1    2    2    1    2    2    1    1    2    1    2    1    1
##  PFE  PHA  SGP  WPI  WYE
##    1    2    1    2    1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
##
```

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

Visualizing the Two Clusters

```
fviz_cluster(k2, data = Ph_norm, pointsize = 1, labelsize = 7)
```



Running the kmeans with k=5 which we got by employing the Silhouette method

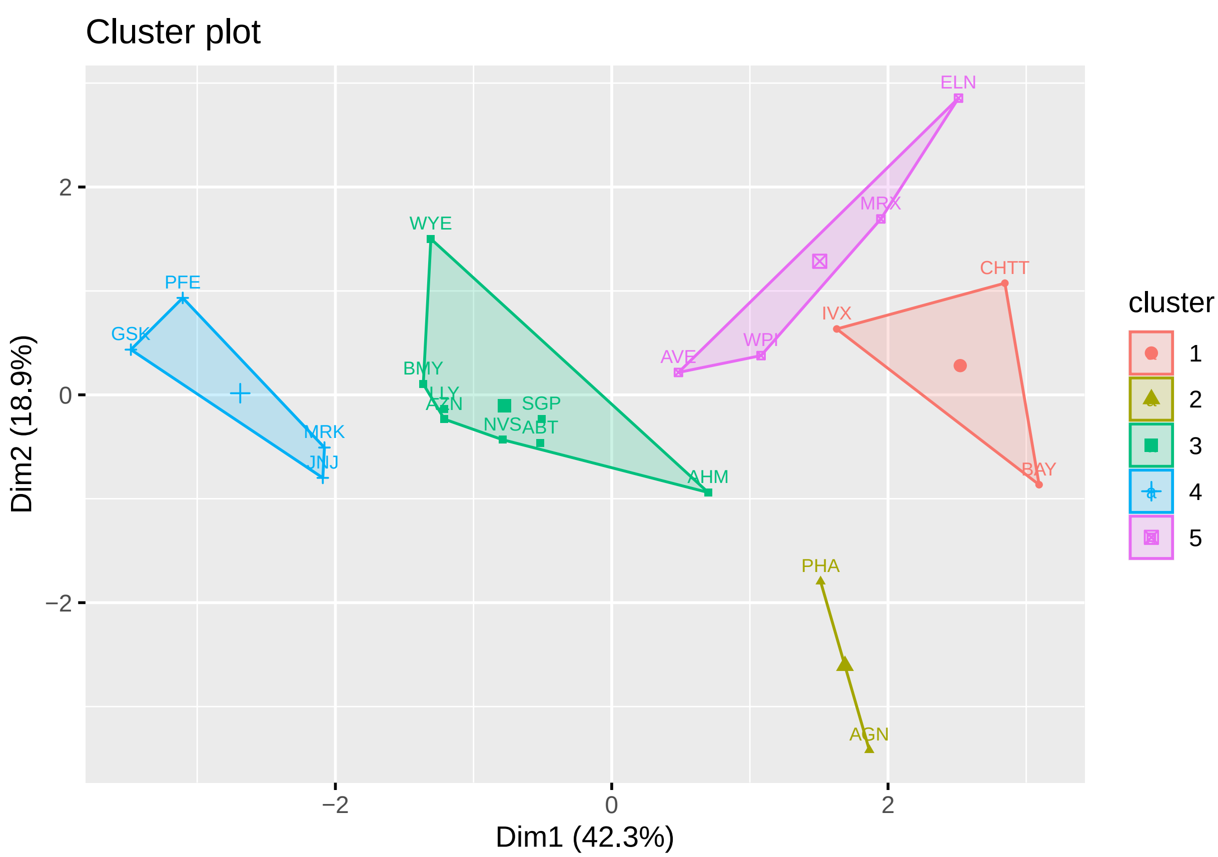
```
k5 <- kmeans(Ph_norm,centers=5,nstart=25)
k5
```

```
## K-means clustering with 5 clusters of sizes 3, 2, 8, 4, 4
##
## Cluster means:
##     Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 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  1.36644699 -0.6912914      -1.320000179
```

```
## 2 -0.14170336 -0.1168459      -1.416514761
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    3    2    3    3    5    1    3    1    5    3    4    1    4    5    4    3
##  PFE  PHA  SGP  WPI  WYE
##    4    2    3    5    3
##
## Within cluster sum of squares by cluster:
## [1] 15.595925  2.803505 21.879320  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Visualizing the Five clusters

```r
fviz_cluster(k5, data = Ph_norm, pointsize = 1, labelsize = 7)
```



B.) Interpreting the clusters we got from WSS and Silhouette with respect to the median of the numerical variables used in forming the clusters by using the original data.

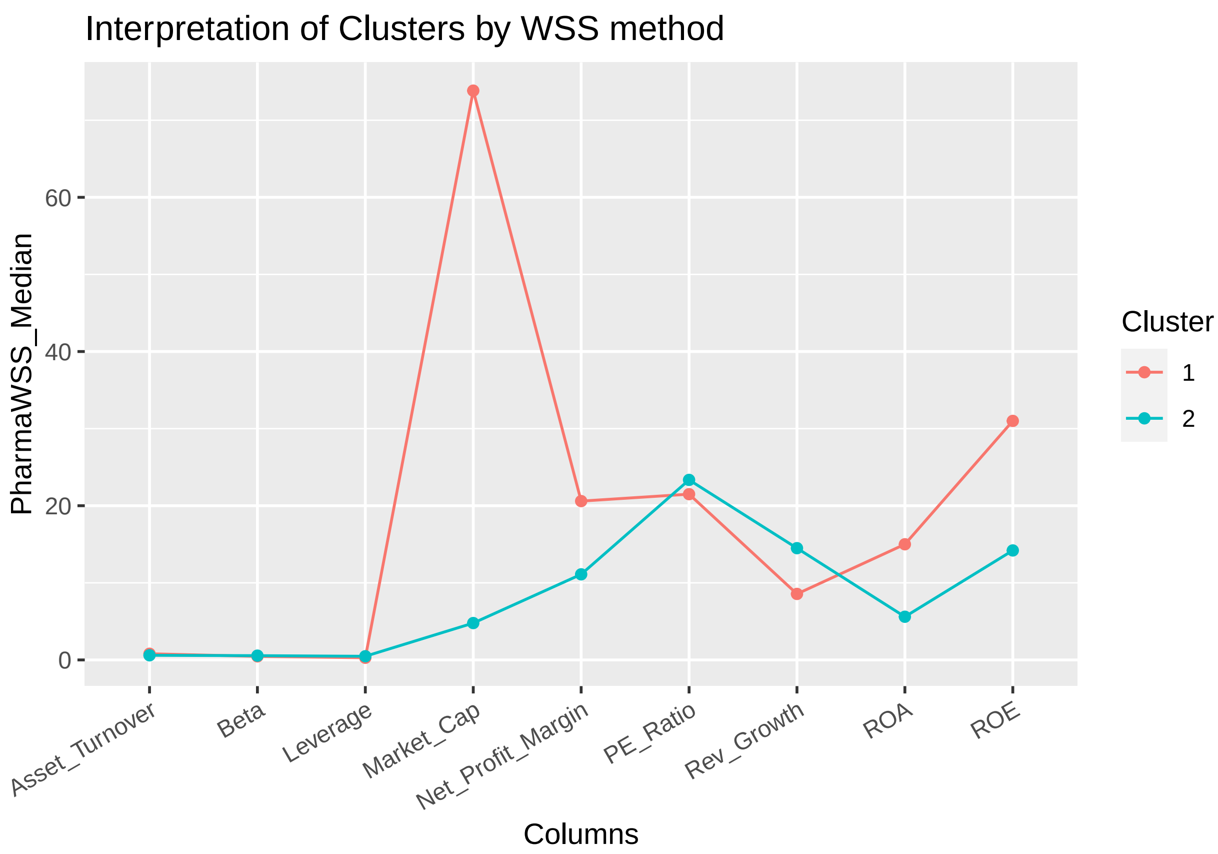
```
#Data Transformation for WSS method

Pharma2_WSS <- cbind(Pharma1, k2$cluster)

colnames(Pharma2_WSS) <- c("Market_Cap", "Beta", "PE_Ratio", "ROE","ROA","Asset_Turnover","Leverage","Re

Pharma2_WSS$Groups <- as.numeric(Pharma2_WSS$Groups)
```

```
PharmaWSS_Median<- aggregate(Pharma2_WSS,by=list(k2$cluster),FUN=median)
PharmaWSS_Median
```

```
##   Group.1 Market_Cap  Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage
## 1       1      73.84 0.460    21.50 31.0 15.0            0.8    0.280
## 2       2       4.78 0.555    23.35 14.2  5.6            0.6    0.475
##   Rev_Growth Net_Profit_Margin Groups
## 1      8.560              20.6      1
## 2     14.495              11.1      2
```

Visualizing the Interpretation between the Clusters formed by WSS method and the numerical variables

```
centers <- data.frame(PharmaWSS_Median[,-c(1,11)]) %>% rowid_to_column() %>%
gather('Columns', 'PharmaWSS_Median',-1)
ggplot(centers, aes(x = Columns, y = PharmaWSS_Median, color = as.factor(rowid))) +
geom_line(aes(group = as.factor(rowid))) + geom_point() +
labs(color = "Cluster", title = 'Interpretation of Clusters by WSS method') +
theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1))
```



Based on the above analysis, the formed clusters can be interpreted as follows;

- By seeing the ***WSS cluster 1*** it can be interpreted that it has bigger Market Capital with a value of 73.84 ,ROE with a value of 31.0, ROA with a value of 15.0 and Net profit margin with a value of 20.6 as compared to the ***WSS cluster 2***which has a market value of just 4.78, ROE value of 14.2, ROA value of 15.0 and Net profit margin of 11.1. It will be profitable to invest in the companies that are under cluster 1 because it has considerable high return on investment and in investing, companies with larger market capitalization are often safer investments as they represent more established companies with generally longer history in business.Also we can see that, the Beta value(Vulnerability to systematic risk) for ***WSS cluster 1*** is low with contrast to ***WSS cluster 2***, which ideally should be low which typically means that the stock is considered less risky.

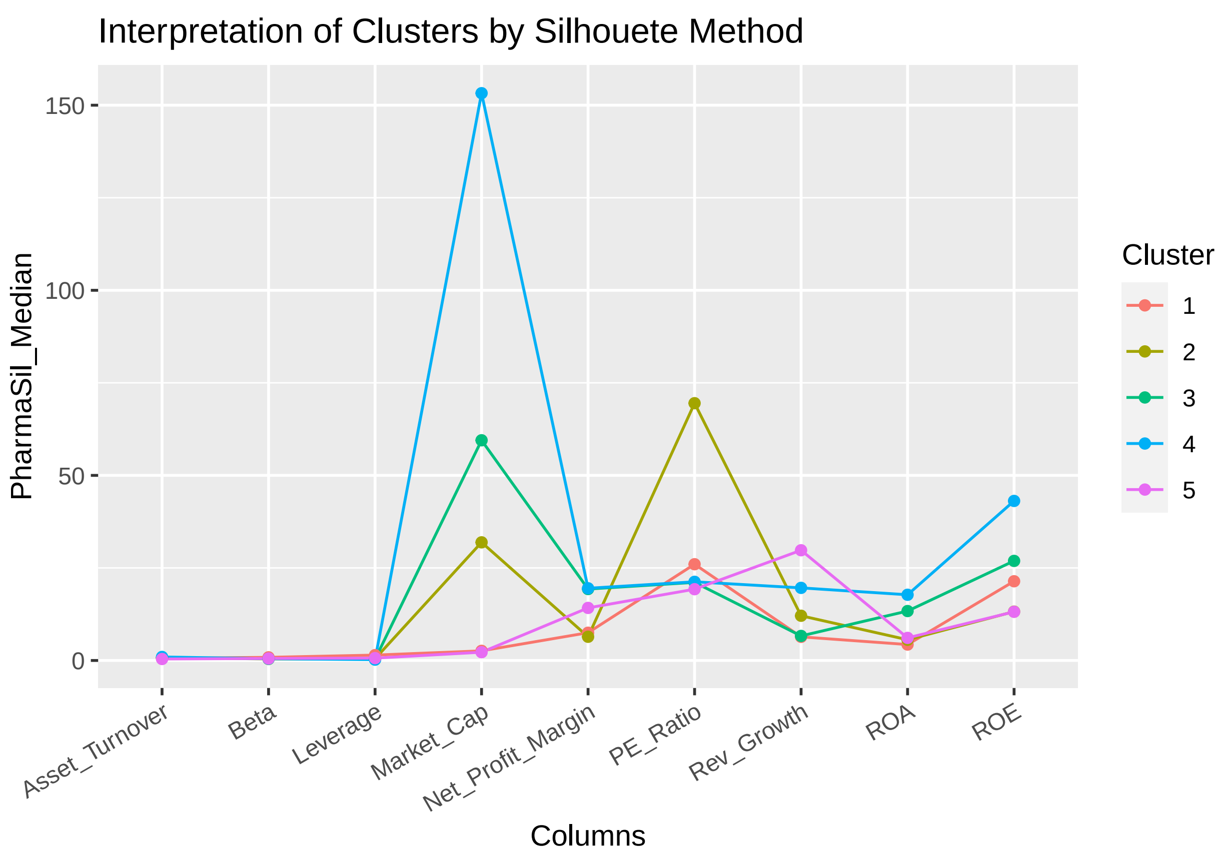
```
# Data Transformation for Silhouette Method

Pharma2_Sil <- cbind(Pharma1,k5$cluster)

colnames(Pharma2_Sil) <- c("Market_Cap", "Beta", "PE_Ratio", "ROE","ROA","Asset_Turnover","Leverage","R

Pharma2_Sil$Groups <- as.numeric(Pharma2_Sil$Groups)
```

```
PharmaSil_Median<- aggregate(Pharma2_Sil,by=list(k5$cluster),FUN=median)
PharmaSil_Median
```

```
##   Group.1 Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
## 1       1      2.600 0.850    26.00 21.40  4.30           0.60    1.450
## 2       2     31.910 0.405    69.50 13.20  5.60           0.75    0.475
```

```
## 3       3     59.480 0.480    21.10 26.90 13.35           0.75     0.345
## 4       4    153.245 0.460    21.25 43.10 17.75           0.95     0.220
## 5       5      2.230 0.535    19.25 13.15  6.10           0.40     0.635
##   Rev_Growth Net_Profit_Margin Groups
## 1      6.380               7.5      1
## 2     12.080               6.4      2
## 3      6.630              19.3      3
## 4     19.610              19.5      4
## 5     29.775              14.2      5
```

```
centers <- data.frame(PharmaSil_Median[,-c(1,11)]) %>% rowid_to_column() %>%
gather('Columns', 'PharmaSil_Median',-1)
ggplot(centers, aes(x = Columns, y = PharmaSil_Median, color = as.factor(rowid))) +
geom_line(aes(group = as.factor(rowid))) + geom_point() +
labs(color = "Cluster", title = 'Interpretation of Clusters by Silhouete Method') +
theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1))
```



Based on the above analysis, the formed clusters can be interpreted as follows;

• The companies in ***Silhouette Cluster 1*** have high Beta (i.e. vulnerable to market changes) and Leverage (making it bad,considering its Profit_Margin, ROA, and Rev_Growth are low).They have moderate PE Ratio but have less than moderate Asset Turnover, Market Cap, Revenue Growth and ROE.

• The first thing that stands out in ***Silhouette Cluster 2*** is its higher PE_Ratio, suggesting the stock's price is high relative to the earnings and possibly overpriced. Also the Net Profit Margin and ROE appears to be the lowest among the clusters.

- The companies in ***Silhouette Cluster 3*** have high Net Profit Margin as compared to the other clusters. They have over moderate values in Market Capital,ROE,ROA and Revenue Growth and less than moderate in Beta, Leverage and PE Ratio.
- ***Silhouette Cluster 4*** has a bigger Market Cap, ROE, ROA, Asset Turnover, and Net Profit Margin; also has a lesser Beta(vulnerability to systematic risk), PE Ratio(growth in the future), and Leverage. This might suggest a cluster of well established big pharma companies.
- ***Silhouette Cluster 5*** appears to have the highest Rev_Growth but relatively unremarkable in the other factors,including low Market Cap and Asset turnover.

C.) Is there a pattern in the clusters with respect to the Categorical variables? (those not used in forming the clusters)

```
#Data Transformation for WSS method

Pharma3_WSS <- cbind(Pharma[,c(12,13,14)],k2$cluster)
colnames(Pharma3_WSS) <- c("Median_Recommendation", "Location", "Exchange", "Groups")
Pharma3_WSS$Groups <- as.numeric(Pharma3_WSS$Groups)

list(Pharma3_WSS)
```

```
## [[1]]
##      Median_Recommendation    Location Exchange Groups
## ABT           Moderate Buy          US     NYSE      1
## AGN           Moderate Buy      CANADA     NYSE      2
## AHM             Strong Buy          UK     NYSE      2
## AZN          Moderate Sell          UK     NYSE      1
## AVE           Moderate Buy      FRANCE     NYSE      2
## BAY                   Hold     GERMANY     NYSE      2
## BMY          Moderate Sell          US     NYSE      1
## CHTT          Moderate Buy          US   NASDAQ      2
## ELN          Moderate Sell     IRELAND     NYSE      2
## LLY                   Hold          US     NYSE      1
## GSK                   Hold          UK     NYSE      1
## IVX                   Hold          US     AMEX      2
## JNJ           Moderate Buy          US     NYSE      1
## MRX           Moderate Buy          US     NYSE      2
## MRK                   Hold          US     NYSE      1
## NVS                   Hold SWITZERLAND     NYSE      1
## PFE           Moderate Buy          US     NYSE      1
## PHA                   Hold          US     NYSE      2
## SGP                   Hold          US     NYSE      1
## WPI          Moderate Sell          US     NYSE      2
## WYE                   Hold          US     NYSE      1
```

Plotting Median Recommendation v/s WSS Clusters

```
ggplot(Pharma3_WSS, aes(fill = Median_Recommendation, x = as.factor(Groups))) +
geom_bar(position = 'stack') + labs(x="Cluster", y="Companies",
title = "Median Recommendation v/s WSS Clusters")
```

Through the above visualization we can interpret that:

- ***WSS Cluster 1*** has mixed recommendations with Hold recommendations being the highest it has moderate sell and buy recommendations as well, this can be because of it's high probability of profit gain due to the high value of Market Capital(73.84), ROE(31.0),ROA(15.0) and a huge Net profit margin(20.6) as compared to the ***WSS Cluster 2***. ***WSS Cluster 1*** companies have the potential to grow in the future and have profitable business on the basis of the values of different profit measuring parameters.

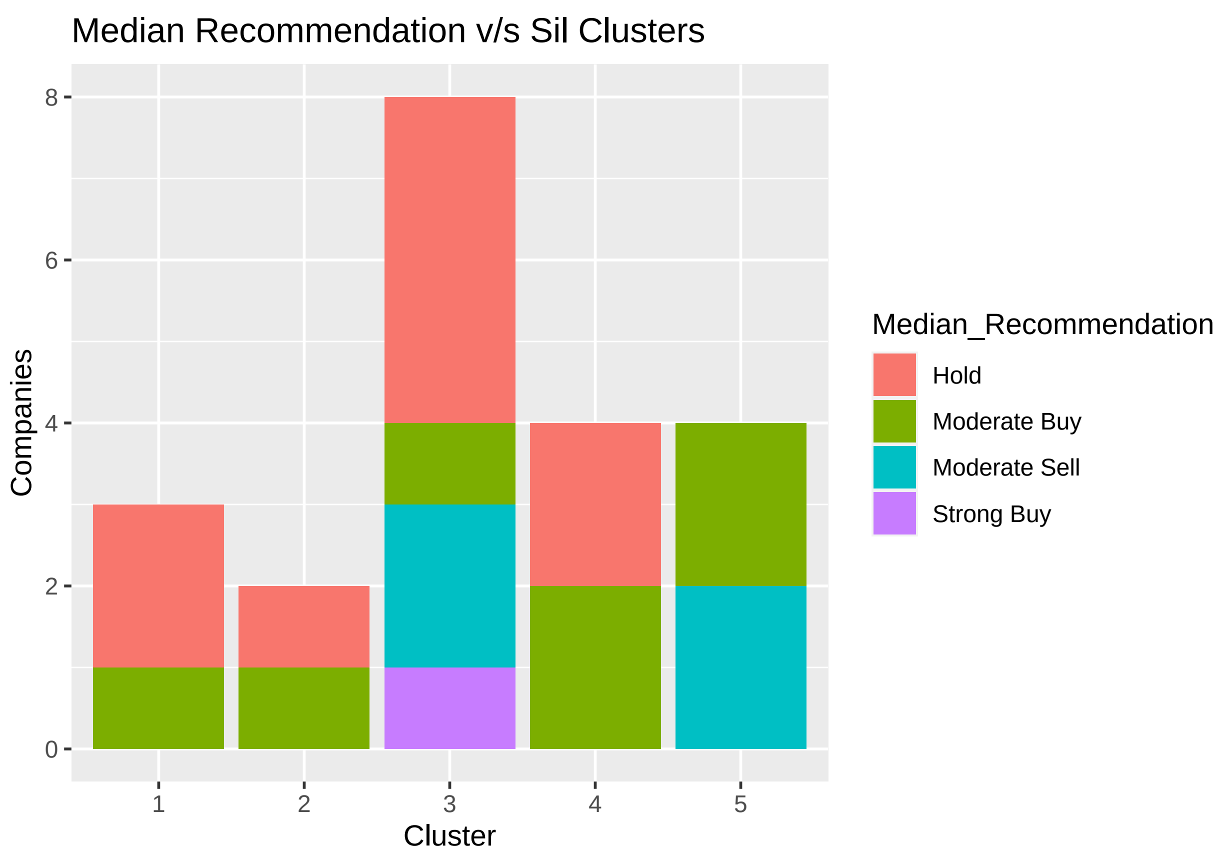
```
#Data Transformation for Silhouette method

Pharma3_Sil <- cbind(Pharma[,c(12,13,14)],k5$cluster)
colnames(Pharma3_Sil) <- c("Median_Recommendation", "Location", "Exchange", "Groups")
Pharma3_Sil$Groups <- as.numeric(Pharma3_Sil$Groups)

list(Pharma3_Sil)
```

```
## [[1]]
##      Median_Recommendation    Location Exchange Groups
## ABT           Moderate Buy          US     NYSE      3
## AGN           Moderate Buy      CANADA     NYSE      2
## AHM             Strong Buy          UK     NYSE      3
## AZN          Moderate Sell          UK     NYSE      3
## AVE           Moderate Buy      FRANCE     NYSE      5
## BAY                   Hold     GERMANY     NYSE      1
## BMY          Moderate Sell          US     NYSE      3
## CHTT          Moderate Buy          US   NASDAQ      1
## ELN          Moderate Sell     IRELAND     NYSE      5
```

```
## LLY                 Hold          US     NYSE      3
## GSK                 Hold          UK     NYSE      4
## IVX                 Hold          US     AMEX      1
## JNJ         Moderate Buy          US     NYSE      4
## MRX         Moderate Buy          US     NYSE      5
## MRK                 Hold          US     NYSE      4
## NVS                 Hold SWITZERLAND     NYSE      3
## PFE         Moderate Buy          US     NYSE      4
## PHA                 Hold          US     NYSE      2
## SGP                 Hold          US     NYSE      3
## WPI        Moderate Sell          US     NYSE      5
## WYE                 Hold          US     NYSE      3
```

Plotting Median Recommendation v/s Silhouette Clusters

```
ggplot(Pharma3_Sil, aes(fill = Median_Recommendation, x = as.factor(Groups)) +
geom_bar(position = 'stack') + labs(x="Cluster", y="Companies",
title = "Median Recommendation v/s Sil Clusters")
```



The pattern that can be interpreted from the Median recommendations with respect to Silhouette Clusters are:

Companies in ***Sil Cluster 1*** are recommended a Hold or Moderate Buy, this can be because of the high BETA value and the leverage value.***Sil Cluster 1***companies has a beta value of 0.850 which means they are highly volatile as compared to other companies.Because of this reason they must be put on hold by measuring the volatility and high risk degree. ***Sil Cluster 2*** are considered overpriced and buying is not ideal. However, one of the recommendations is for a Moderate Buy, which doesn't make sense here. ***Sil***

Cluster 3 has mixed recommendations of Moderate buy/sell and hold. It is found to be second profit earning cluster in future because of decent Market capital value, ROE, ROA and Net profit margin. It has decent Beta and leverage value which does not indicate much of volatility and risk degree in investment. The pattern of median recommendation in ***Sil Cluster 4*** is shockingly surprising. Even though it has the highest values of Market capital, ROE,ROA,Asset turnover, Revenue growth and considerably less value of beta, leverage and PE ratio it is still considered to be moderate buy or hold. It is plausibly the highest revenue generating cluster with a huge scope of earning great profits still it has recommendations of hold. In ***Sil Cluster 5*** it has recommendations of Moderate buy and Moderate sell which dosen't makes sense because there are companies in this cluster which has high beta value and leverage as compared to other companies which will not drive investors to invest or buy shares in this cluster.

Plotting Locations v/s Silhouette Clusters

```
ggplot(Pharma3_Sil, aes(fill = Location, x = as.factor(Groups))) +
geom_bar(position = 'stack') + labs(x="Cluster", y="Location",
title = "Location v/s Sil Clusters")
```

The pattern observed by the above visualization is that all the clusters have companies that are US based. Companies in ***Sil Cluster 3*** which in comparison to other clusters is doing well and has majortity of its companies originating in US. Secondly, the best cluster observed in Silhouette method i.e., ***Sil Cluster 4*** also has majority of its companies US based. This can be conclude that companies which are better performing are established in the US.

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

- Sil Cluster 1- ***'Poorly Performing Pharma'***, with low performance across all the featuresand very high BETA and Leverage value.

- Sil Cluster 2- ***'Overpriced Pharma'***, with high PE ratio.
- Sil Cluster 3: ***'Currently Profitable Pharma'*** with good Net_Profit_Margin, but lowest Revenue Growth.
- Sil Cluster 4: ***'Big Pharma'***, with high Market Capital, ROE, ROA, Asset Turnover, and Net profit margin.
- Sil Cluster 5: ***'Future Potential Pharma'***, with highest Rev_Growth.

Conclusion:

The size and value of a company (Market Capital Value) can inform the level of risk you might expect when investing in its stock, as well as how much your investment might return over time. The ROA figure gives investors an idea of how effective the company is in converting the money it invests into net income. The higher the ROA number, the better, because the company is able to earn more money with a smaller investment. Put simply, a higher ROA means more asset efficiency.Additionally, when we talk about ROE- The higher the ROE, the better a company is at converting its equity financing into profits. ***'Big Pharma'*** cluster formed through Silhouette method has all this characteristics and values. Therefore, ***'Big Pharma'*** cluster would generate higher amount of profits and will be very beneficial for the investors to invest in ***Big Pharma*** companies .

I have considered ***'Big Pharma'*** cluster from Silhouette method more optimal than WSS Clusters because if we compare the median values of variables in these clusters ***'Big Pharma'*** cluster have values which are higher than the clusters formed by WSS method. It shows that individuals will most likely be investing in this cluster as it will be profitable and less riskier for them in the future.