# Assignment_4

Meenakshi Vaidhiyanathan

2024-03-17

Installing the necessary packages using `install.packages()` function.

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

Loading the necessary packages using `library()` function.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
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)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

After importing the dataset, numerical variables are selected and is normalized.

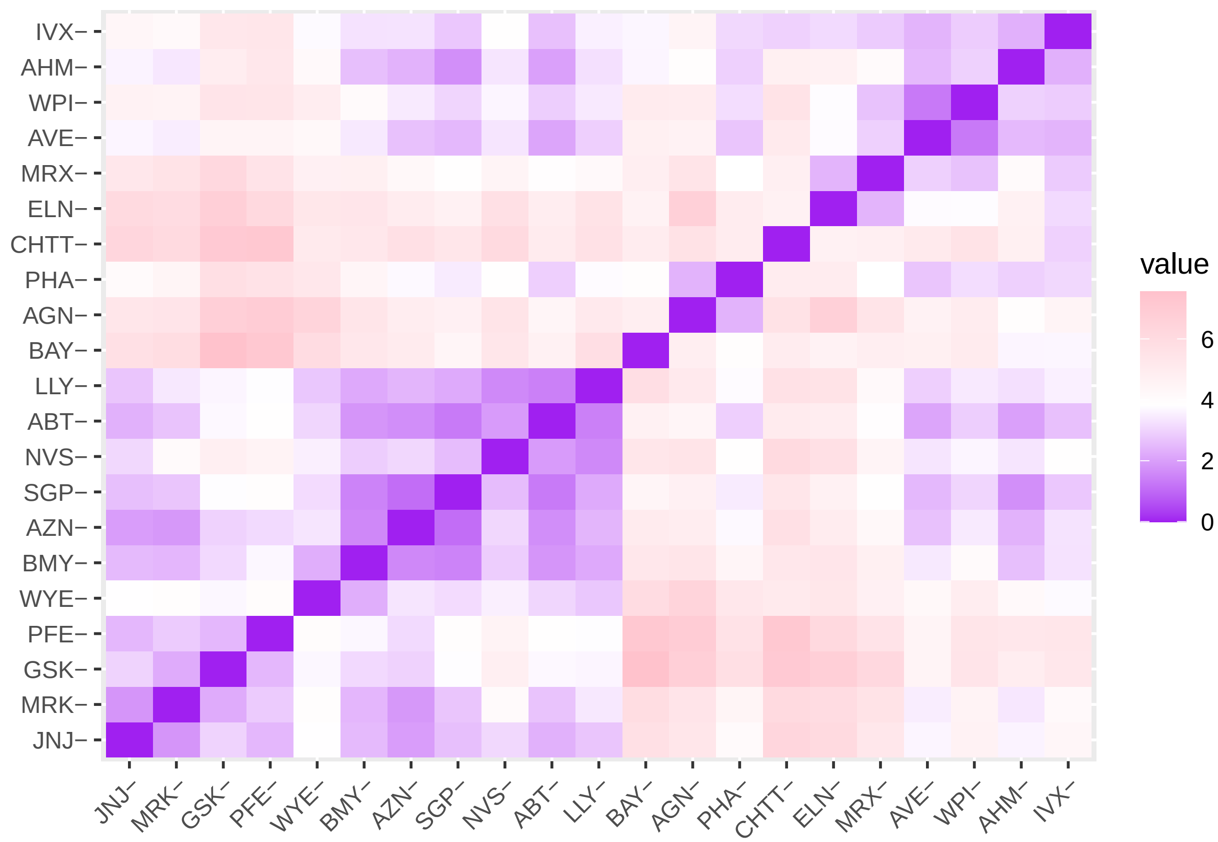
```r
library(readr)
Pharma <- read.csv("~/Downloads/Pharmaceuticals.csv")
rownames(Pharma)<- Pharma$Symbol
Pharmacy1 <- Pharma[,-c(1,2,12,13,14)]
Pharm_norm <- scale(Pharmacy1)
summary(Pharm_norm)
```

```
##    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
```

The functions `get_dist()` and `fviz_dist()` is used to calculate and visualize the distance matrix which visually depicts the similarity or dissimilarity of the different data points.

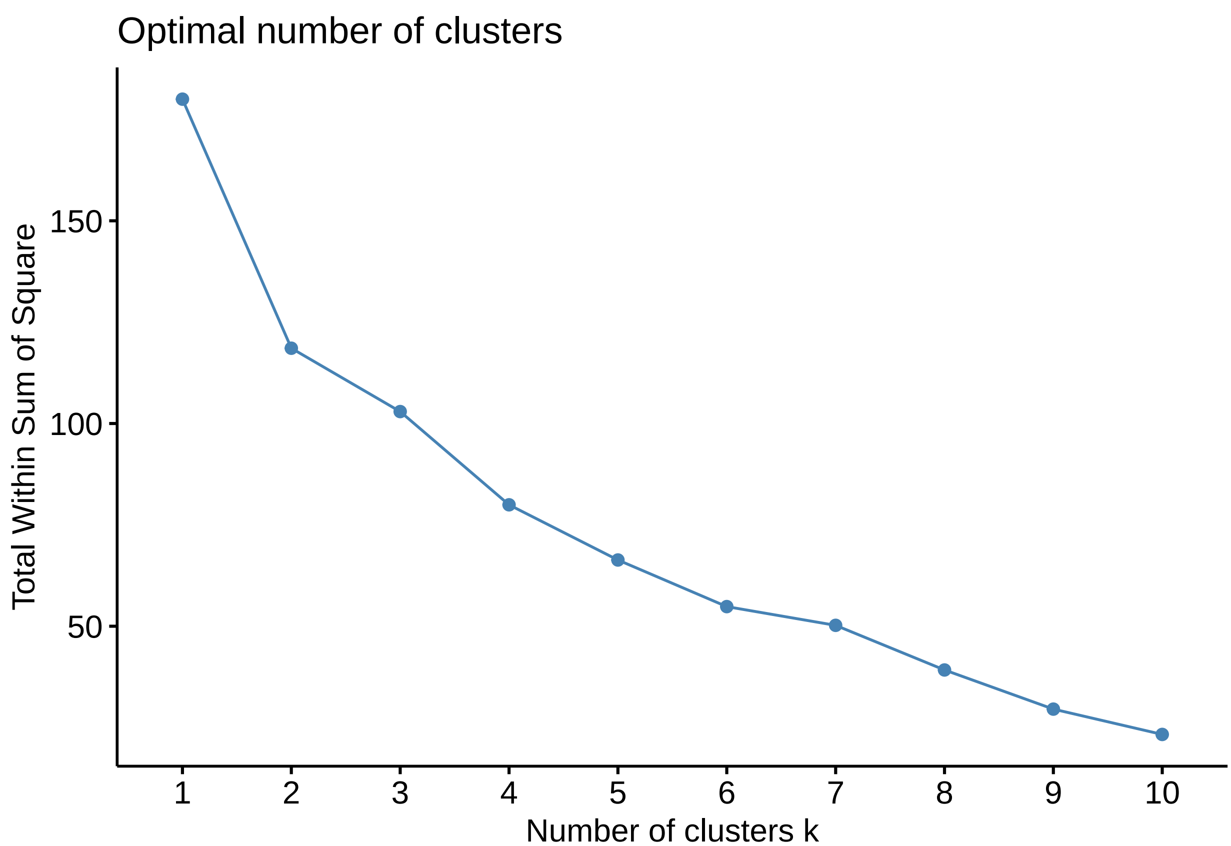
The parameter here with which each pair of observation is depicted with respect to clustering is distance. Pink color depicts dissimilarity while purple color shows similarity as seen below. Data points that have minimal distance between them belong to same cluster as similarity determines data points that can be combined into clusters.

```
set.seed(420)
dist_matrix <- get_dist(Pharm_norm)
fviz_dist(dist_matrix, gradient = list(low = " purple", mid = "white", high = "pink"))
```

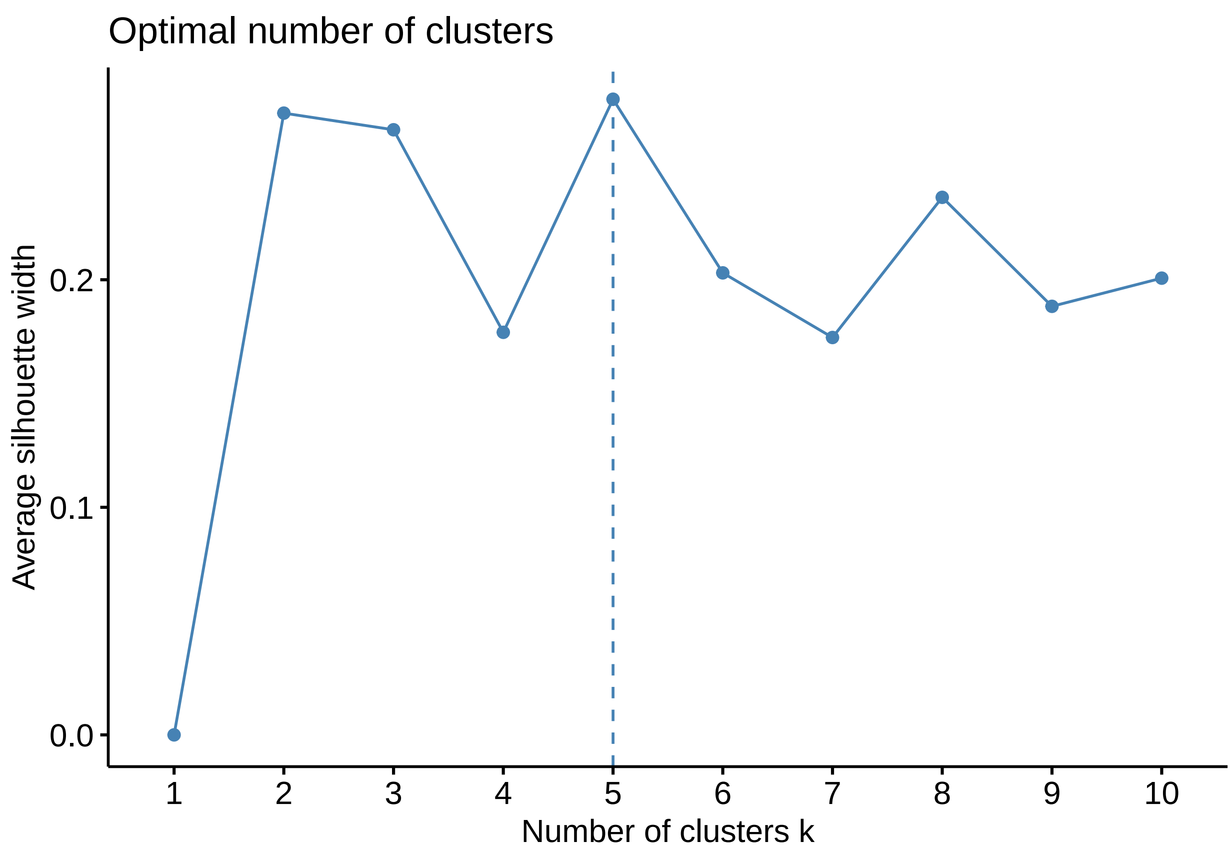



In order to find optimal k value, WSS and Silhouete methods are used as seen below. We find that the optimal k value is found to be 2 using WSS method while k value as 5 using silhouette method.

```
WSS_K_Value <- fviz_nbclust(Pharm_norm,kmeans,method="wss")
WSS_K_Value
```




```r
Silhouette_K_Value <- fviz_nbclust(Pharm_norm,kmeans,method="silhouette")
Silhouette_K_Value
```




Using k value as 2 from WSS method using `kmeans()` function. `Pharm_norm` is passed as an argument into `kmeans()` function.

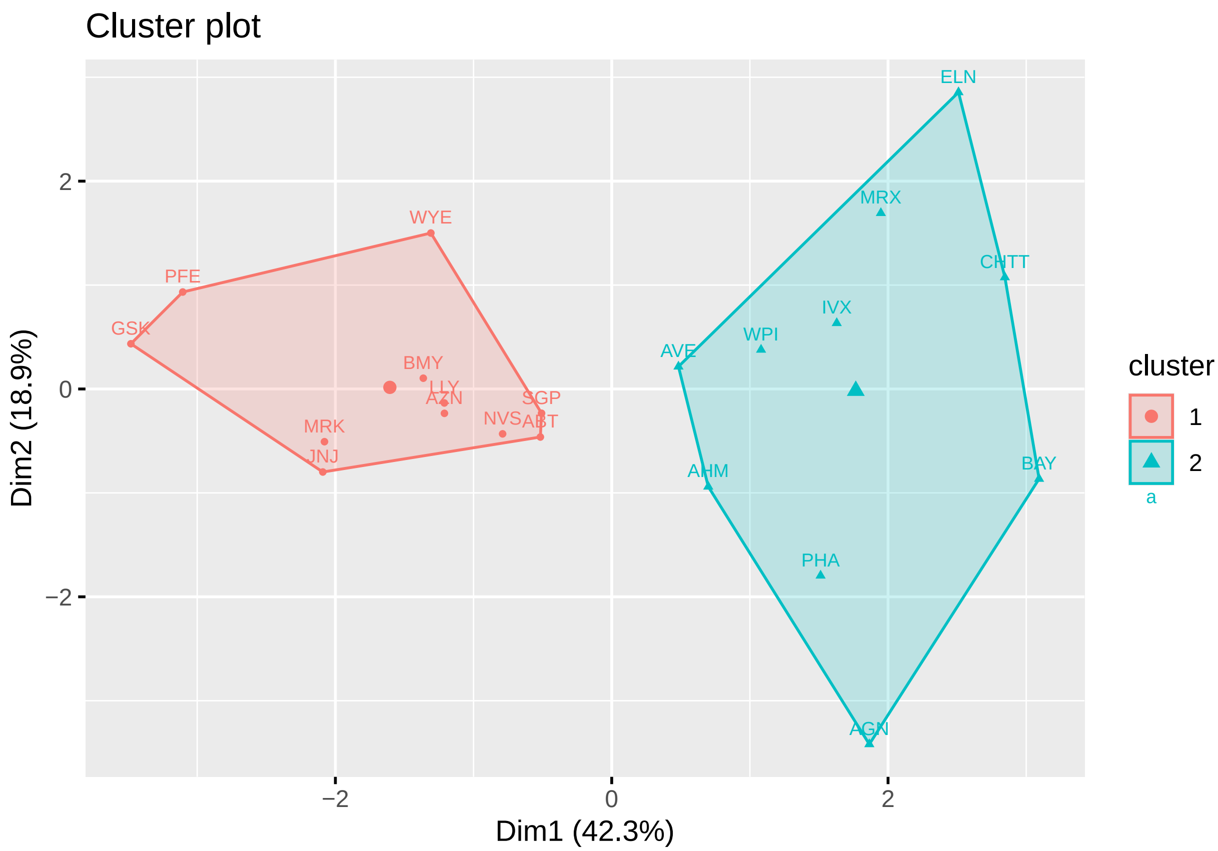
```
k_means_2<- kmeans(Pharm_norm, centers=2, nstart = 25)
k_means_2
```

```
## 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"
```

The two clusters are visualized using the `fviz_cluster()` function by passing `k_means_2` as an argument.

```
fviz_cluster(k_means_2, data = Pharm_norm, pointsize = 1, labelsize = 7)
```




***Running the kmeans with k=5 which we got by employing the Silhouette_K_Value method***
Using k value as 5 from Silhouette method to execute k means.

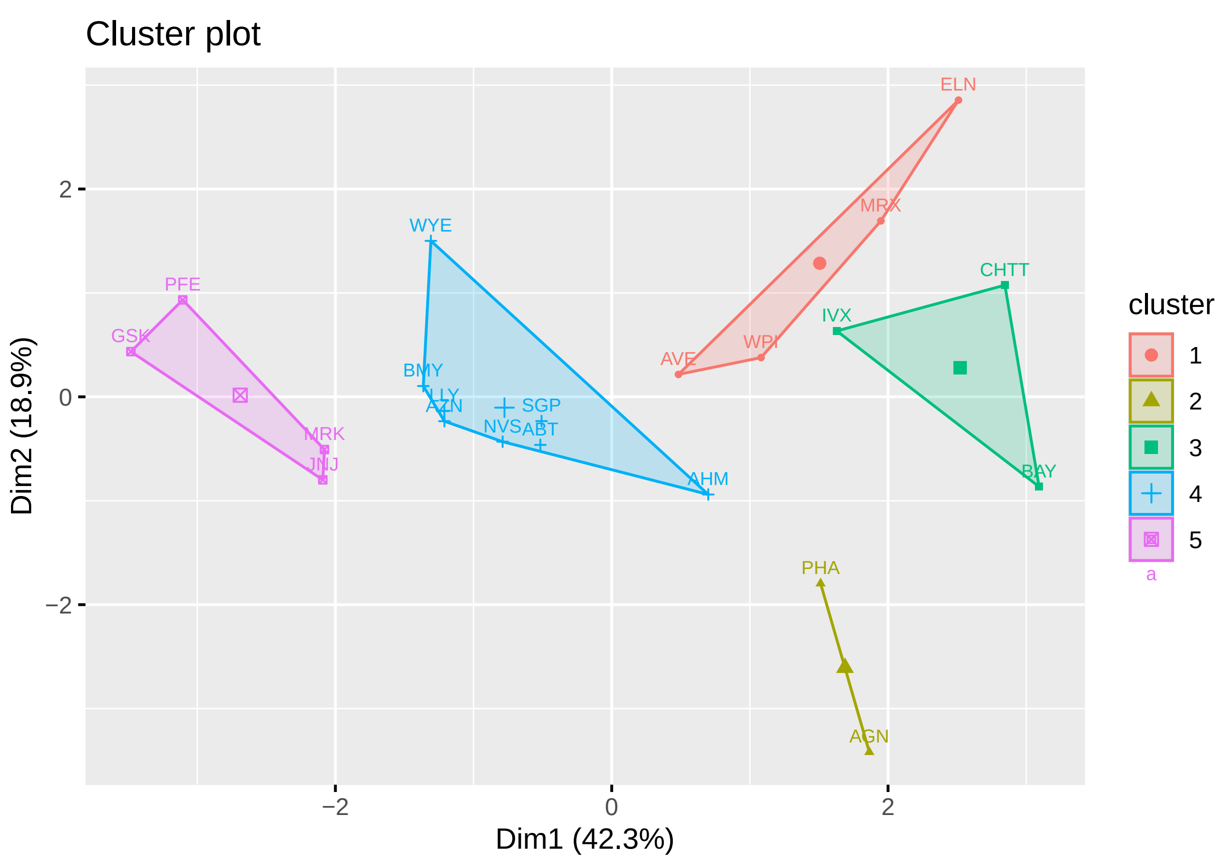
```
k_means_5 <- kmeans(Pharm_norm,centers=5,nstart=25)
k_means_5
```

```
## K-means clustering with 5 clusters of sizes 4, 2, 3, 8, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 3 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
```

```
## 1  0.06308085  1.5180158      -0.006893899
## 2 -0.14170336 -0.1168459      -1.416514761
## 3  1.36644699 -0.6912914      -1.320000179
## 4 -0.27449312 -0.7041516       0.556954446
## 5 -0.46807818  0.4671788       0.591242521
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    4    2    4    4    1    3    4    3    1    4    5    3    5    1    5    4
##  PFE  PHA  SGP  WPI  WYE
##    5    2    4    1    4
##
## Within cluster sum of squares by cluster:
## [1] 12.791257  2.803505 15.595925 21.879320  9.284424
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

The five clusters are visualised using `fviz_cluster()` function by passing `k_means_5`.

```
fviz_cluster(k_means_5, data = Pharm_norm, pointsize = 1, labelsize = 7)
```




***B.) Interpreting the clusters we got from WSS_K_Value and Silhouette_K_Value with respect to the median of the numerical variables used in forming the clusters by using the original***

***data.***

B) From WSS and Silhouette method, clusters with respect to median of numerical variables are interpreted.

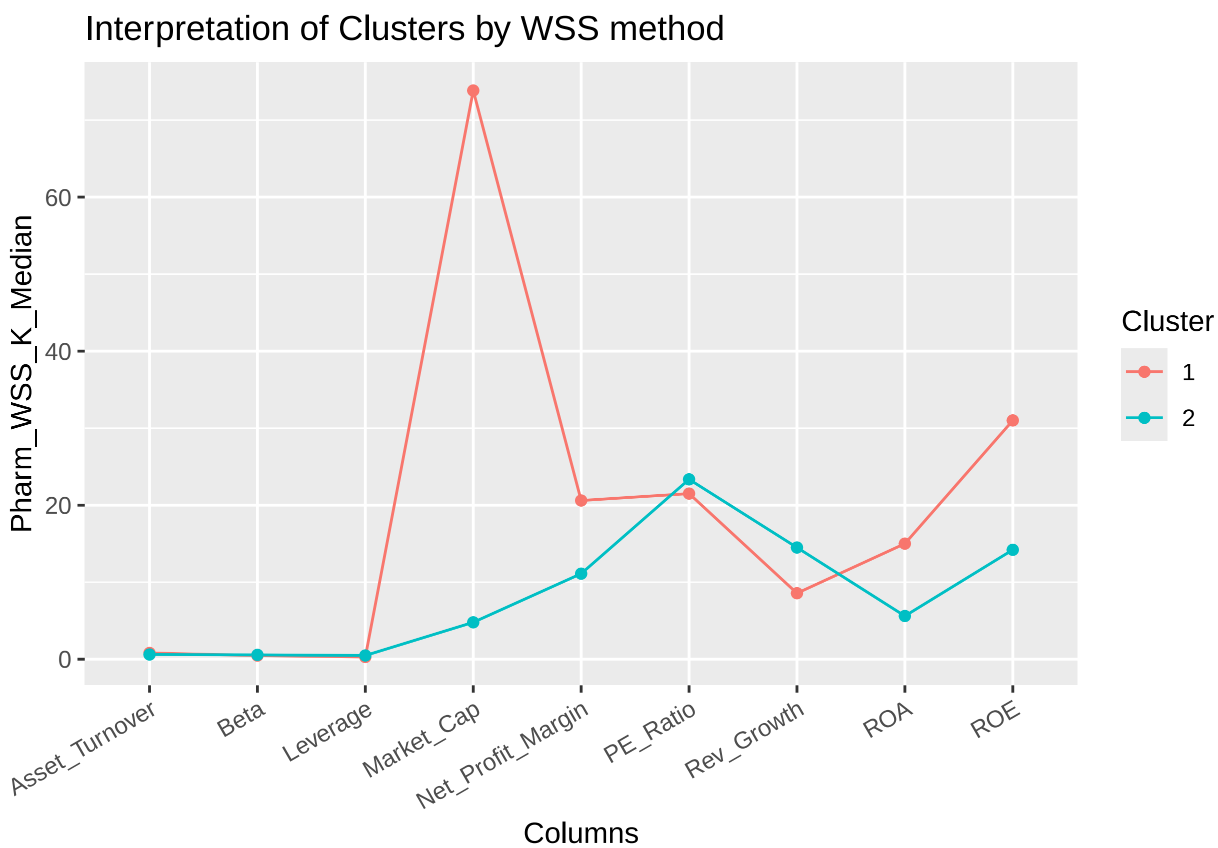
```
#Data Transformation for WSS method

Pharma_2_WSS_K <- cbind(Pharmacy1, k_means_2$cluster)

colnames(Pharma_2_WSS_K) <- c("Market_Cap", "Beta", "PE_Ratio", "ROE","ROA","Asset_Turnover","Leverage"

Pharma_2_WSS_K$Groups <- as.numeric(Pharma_2_WSS_K$Groups)
```

```
Pharm_WSS_K_Median<- aggregate(Pharma_2_WSS_K,by=list(k_means_2$cluster),FUN=median)
Pharm_WSS_K_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
```

Clusters from WSS method and numerical variables are visualized.

```
centers_clust <- data.frame(Pharm_WSS_K_Median[,-c(1,11)]) %>% rowid_to_column() %>%
gather('Columns', 'Pharm_WSS_K_Median',-1)
ggplot(centers_clust, aes(x = Columns, y = Pharm_WSS_K_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;*** From above analysis, the clusters are interpreted as:

From, WSS_K_Value cluster 1: It has higher Market capital of 73.84, ROE is 31.0, ROA is 15.0 and Net Profit margin is 20.6 as compared to WSS_K_Value cluster 2 whose market value is 4.78, ROE of 14.2, ROA of 15.0 and Net profit margin of 11.1. Cluster 1 investment is profitable as it as greater history in business depicted by well established companies as they are more safer given that they have large market capitalization.

Vulnerability to systemic risk depicted by Beta value for WSS_K_Value cluster 1 is less as compared to WSS_K_Value cluster 2. WSS_K_Value cluster 2 should have been low as in a lesser risker stock value.

Transformation of data using Silhouette method.

```
Ph_2_Sil <- cbind(Pharmacy1,k_means_5$cluster)
colnames(Ph_2_Sil) <- c("Market_Cap", "Beta", "PE_Ratio", "ROE","ROA","Asset_Turnover","Leverage","Rev_
Ph_2_Sil$Groups <- as.numeric(Ph_2_Sil$Groups)
```

`aggregate()` function is used below and is calculated with respect to median.

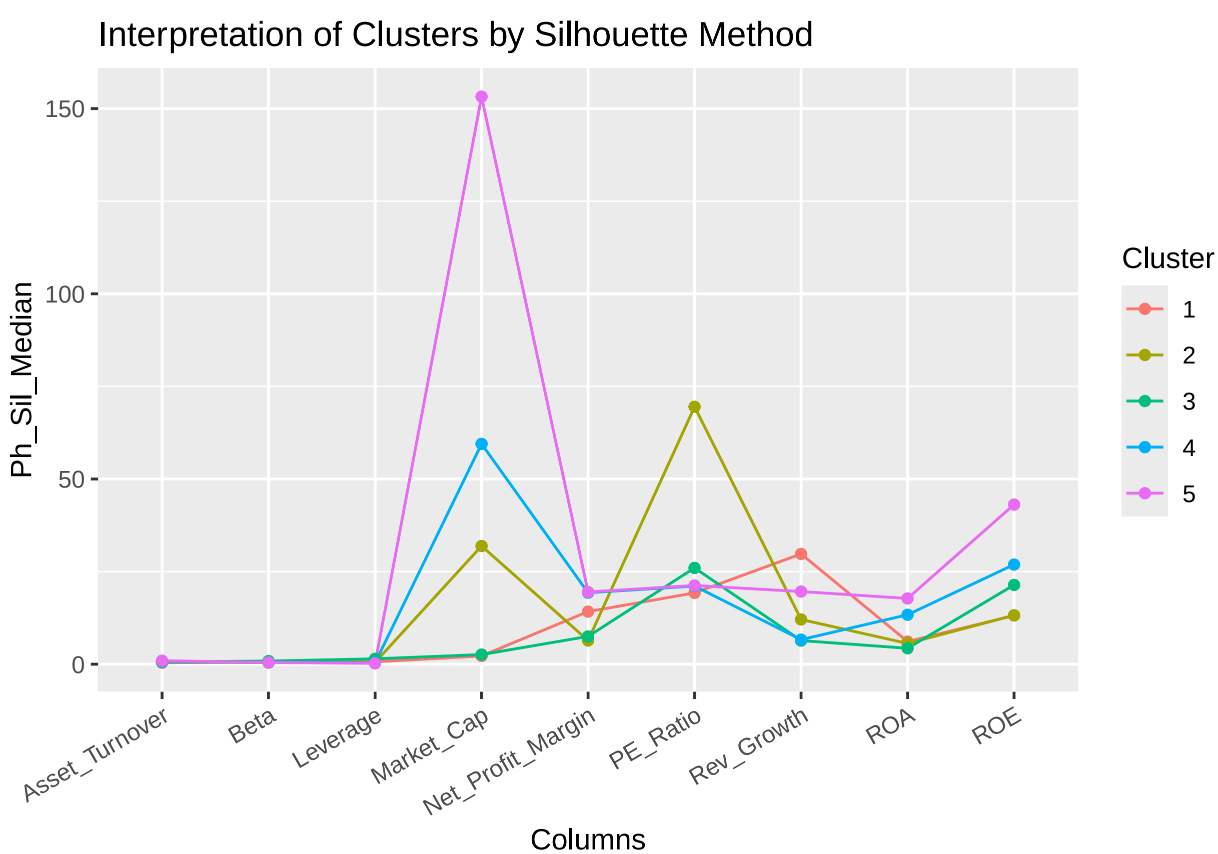
```
Ph_Sil_Median<- aggregate(Ph_2_Sil,by=list(k_means_5$cluster),FUN=median)
Ph_Sil_Median
```

```
##   Group.1 Market_Cap  Beta PE_Ratio   ROE  ROA Asset_Turnover Leverage
## 1       1      2.230 0.535    19.25 13.15 6.10           0.40    0.635
## 2       2     31.910 0.405    69.50 13.20 5.60           0.75    0.475
## 3       3      2.600 0.850    26.00 21.40 4.30           0.60    1.450
```

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

`ggplot()` function is used to depict clusters using Silhouette method. Silhouette_K_Value Cluster 1: has high beta and leverage. It's Profit margin, ROA, and Rev_Growth are columns depict low value. Asset Turnover, Market Cap, Revenue Growth and ROE columns have less than moderate values while PE ratio is moderate. Silhouette_K_Value Cluster 2: has high PE ratio. Stock price is high with respect to earnings. Lowest values of Profit Margin and ROE are recorded. Silhouette_K_Value Cluster 3: have high net profit margin in comparison to others while Market Capital,ROE,ROA and Revenue Growth have moderate values. Beta, Leverage and PE Ratio record less than moderate. Silhouette_K_Value Cluster 4: has less Beta, PE ratio and Leverage. Market Cap, ROE, ROA, Asset Turnover, Net Profit Margin have higher values. This says that the clusters represent that they are well establishes companies. Silhouette_K_Value Cluster 5: has highest revenue growth but the rest being low.

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



C) WSS method used for data transformation.

```
Pharma_3_WSS_K <- cbind(Pharma[,c(12,13,14)],k_means_2$cluster)
colnames(Pharma_3_WSS_K) <- c("Median_Recommendation", "Location", "Exchange", "Groups")
Pharma_3_WSS_K$Groups <- as.numeric(Pharma_3_WSS_K$Groups)

list(Pharma_3_WSS_K)
```

```
## [[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
```

Using `ggplot()` to visualize Median recommendation v/s Clusters. WSS_K_Value Cluster 1: has highest hold recommendations. Buy and sell is moderate. Probability of profit gain is high because its Market Capital is 73.84, ROE is 31.0, ROA is 15.0 and a high Net profit margin of 20.6 as compared to the WSS_K_Value Cluster 2 while WSS_K_Value Cluster 1 has more potential to grow later.

```
ggplot(Pharma_3_WSS_K, aes(fill = Median_Recommendation, x = as.factor(Groups))) +
geom_bar(position = 'stack') + labs(x="Cluster", y="Companies",
title = "Median Recommendation v/s WSS Clusters")
```


Silhouette method data transformation.

```
Pharma_3_Silhouette <- cbind(Pharma[,c(12,13,14)],k_means_5$cluster)
colnames(Pharma_3_Silhouette) <- c("Median_Recommendation", "Location", "Exchange", "Groups")
Pharma_3_Silhouette$Groups <- as.numeric(Pharma_3_Silhouette$Groups)

list(Pharma_3_Silhouette)
```

```
## [[1]]
##      Median_Recommendation    Location Exchange Groups
## ABT           Moderate Buy          US     NYSE      4
## AGN           Moderate Buy      CANADA     NYSE      2
## AHM             Strong Buy          UK     NYSE      4
## AZN          Moderate Sell          UK     NYSE      4
## AVE           Moderate Buy      FRANCE     NYSE      1
## BAY                   Hold     GERMANY     NYSE      3
## BMY          Moderate Sell          US     NYSE      4
## CHTT          Moderate Buy          US   NASDAQ      3
## ELN          Moderate Sell     IRELAND     NYSE      1
## LLY                   Hold          US     NYSE      4
## GSK                   Hold          UK     NYSE      5
## IVX                   Hold          US     AMEX      3
## JNJ           Moderate Buy          US     NYSE      5
## MRX           Moderate Buy          US     NYSE      1
## MRK                   Hold          US     NYSE      5
## NVS                   Hold SWITZERLAND     NYSE      4
```

```
## PFE          Moderate Buy           US      NYSE       5
## PHA                  Hold           US      NYSE       2
## SGP                  Hold           US      NYSE       4
## WPI         Moderate Sell           US      NYSE       1
## WYE                  Hold           US      NYSE       4
```

Using `ggplot()` function to plot Median Recommendation v/s Silhouette Clusters. Silhouette Cluster 1 has high Beta of 0.850 and hence high volatility in comparision to others and high leverage value and hence provide Hold or moderate buy. Hence hold suggestion due to high risk. Silhouette Cluster 2 are expensive and not ideal to purchase. Silhouette Cluster 3 has mixed recommendations of Moderate buy or sell and hold. It has good market capital,ROE, ROA and Net profit margin. Hence can be considered as second profittable cluster.
Silhouette Cluster 4 has high Market capital, ROE,ROA,Asset turnover, Rev_Growth but less beta, leverage and PE ratio. It still suggests to be moderate buy or hold. Silhouette Cluster 5 suggests data points with high beta and leverage in comparison to others.

```
ggplot(Pharma_3_Silhouette, aes(fill = Median_Recommendation, x = as.factor(Groups)) +
geom_bar(position = 'stack') + labs(x="Cluster", y="Companies",
title = "Median Recommendation v/s Silhouette Clusters")
```


Using `ggplot()` function to visualize Location v/s Silhouette Clusters. This suggests that most companies are US based. cluster 3 is performing well while both cluster 3 and 4 are US based companies.

An appropriate name for each cluster using any or all of the variables in the dataset: Silhouette Cluster 1: " Pharmacy companies with poor performance" depicted by high Beta and leverage values with low performance. Silhouette Cluster 2: " Highly priced Companies" depicted by high PE Ratio. Silhouette

Cluster 3: " At present profitable companies" as it has low revenue growth and good net growth rate. Silhouette Cluster 4: " Large pharma companies" as it has high Market Capital, ROE, ROA, Asset Turnover including Net profit margin. Silhouette Cluster 5: " Future prospective pharmaceutical companies" as it has highest reveue growth.

```
ggplot(Pharma_3_Silhouette, aes(fill = Location, x = as.factor(Groups))) +
geom_bar(position = 'stack') + labs(x="Cluster", y="Location",
title = "Location v/s Silhouette Clusters")
```


Stock investment are influenced by ROA as in, higher ROA, lesser the investment with more earnings or profits. Median values of Silhouette method is higher than depicted in WSS method and hence Silhouette method is considered with respect to "Large pharma companies" and hence is more profitable and less risky.