Tejesh_Varma_Maddana_FML_Assignment_4

Tejesh Varma Maddana

2023-11-07

#Loading the required packages for analyzing the problem statement

```
rm(list = ls()) #cleaning the environment
library(readr)
library(cluster)
library(tidyr)
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ggplot2)
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.3     ✔ purrr     1.0.2
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors

library(pander)
library(caret)
library(knitr)
library(class)
library(reshape2)

##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths

library(kernlab)
```

```
##
## Attaching package: 'kernlab'
##
## The following object is masked from 'package:purrr':
##
##     cross
##
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```r
library(ggcorrplot)
library(dplyr)
library(e1071)
library(factoextra)
```

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

```r
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: modeltools
## Loading required package: stats4
##
## Attaching package: 'modeltools'
##
## The following object is masked from 'package:kernlab':
##
##     prior
##
##
## Attaching package: 'flexclust'
##
## The following object is masked from 'package:e1071':
##
##     bclust
##
## The following object is masked from 'package:kernlab':
##
##     kcca
```

```r
library(cowplot)
```

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

#Importing the data from the problem statement

```
P <- read.csv("~/Documents/KSU/Fundamentals of Machine Learning -
64060/Assignments/4. Assignment_4/Pharmaceuticals.csv")
head(P)

##   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
```

#Understand the bank data structure:- Display the structure of dataset by using the function "str()" so as to know about different data type, dimensions, and the elements in dataset.

```
str(P)

## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc."
"Amersham plc" "AstraZeneca PLC" ...
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08
0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6
27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1
```

```
31 ...
## $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5
...
## $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
## $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53
...
## $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
## $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3
23.4 ...
## $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy"
"Moderate Sell" ...
## $ Location             : chr  "US" "CANADA" "UK" "UK" ...
## $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...

#From the Structure we know that there are 21 obs. of  14 variables
```

#Using the "is.na()" function for checking the missing or not available values in the provided dataset.

```
colMeans(is.na(P))

##                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
```

#Since the result in all the columns indicated zero, it indicates there are no missing values in the provided dataset

Problem Statement - 1.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.

#Using the numerical variables from 1 to 9 to cluster the 21 firms.

```
P2 <- P[,c(1,3:11)]
```

#Assigning the rownames to the each firm

```
row.names(P2) <- P2[,1]
```

#Deleting the symbol column from the P2 dataframe

```
P2 <- P2[,-1]
head(P2)
```

```
##     Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## ABT      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## AGN       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## AHM       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## AZN      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## AVE      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## BAY      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##     Net_Profit_Margin
## ABT              16.1
## AGN               5.5
## AHM              11.2
## AZN              18.0
## AVE              12.9
## BAY               2.6
```

#Displaying the structure of dataframe P2

```
str(P2)

## 'data.frame':    21 obs. of  9 variables:
##  $ Market_Cap       : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta             : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08
0.18 ...
##  $ PE_Ratio         : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9
...
##  $ ROE              : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31
...
##  $ ROA              : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover   : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage         : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth       : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin: num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4
...
```

#Dropped the columns of 'Name', 'Median_Recommendation', 'Location', 'Exchange'

#By using the scale function, Normalizing the data

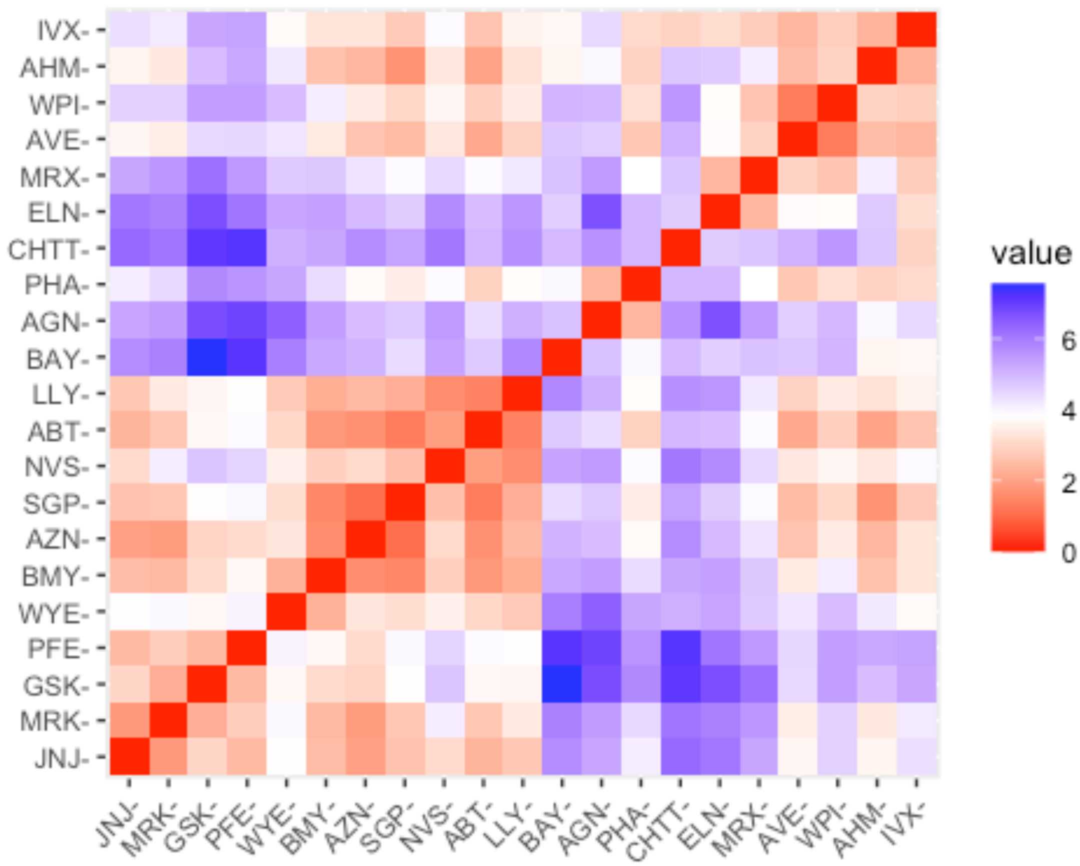
```
set.seed(44)
P_Norm <- scale(P2)
#normalizing the data by subtracting the mean of the data and dividing by the
standard deviation
pandoc.table(head(P_Norm),style="grid", split.tables = Inf)

##
##
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+
## |    | Market_Cap |   Beta   | PE_Ratio |   ROE   |   ROA   |
Asset_Turnover | Leverage | Rev_Growth | Net_Profit_Margin |
##
```

```
+=========+============+==========+==========+=========+=========+===========
=====+==========+============+===================+
## | **ABT** |   0.1841   | -0.8013  | -0.04671 | 0.04009 | 0.2416  |       0
| -0.2121  |  -0.5278   |      0.06168      |
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+
## | **AGN** |  -0.8544   | -0.4507  |  3.497   | -0.8548 | -0.9423 |
0.9225      | 0.01828  |  -0.3811   |      -1.554       |
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+
## | **AHM** |  -0.8763   |  -0.256  |  -0.292  | -0.7223 | -0.5101 |
0.9225      | -0.4041  |  -0.5721   |      -0.685       |
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+
## | **AZN** |   0.1703   | -0.02226 | -0.2429  | 0.1064  | 0.9181  |
0.9225      | -0.7497  |   0.1474   |      0.3512       |
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+
## | **AVE** |   -0.179   | -0.8013  | -0.3287  | -0.2648 | -0.5664 |    -
0.4613      | -0.3145  |   1.216    |      -0.426       |
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+
## | **BAY** |  -0.6954   |  2.276   |  0.1495  | -1.451  | -1.713  |    -
0.4613      | -0.7497  |   -1.497   |      -1.996       |
## +---------+------------+----------+----------+---------+---------+--------
--------+----------+------------+-------------------+

# Displaying the top 6 Observation from pharma_Norm
```

#Clustering the data by using euclidean distance and plotting the graph to interpret the results #By using the Euclidean distance formula

$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2)}$$

```
#Finding the distances between observations in the data by using the above
Euclidean distance formula
P_d <- get_dist(P_Norm)

#Considering the distance matrix P_d as its main argument, visualizing the
distances by using the 'fviz_dist()' function which displays the heat-map.

fviz_dist(P_d, order = TRUE, show_labels = TRUE)
```



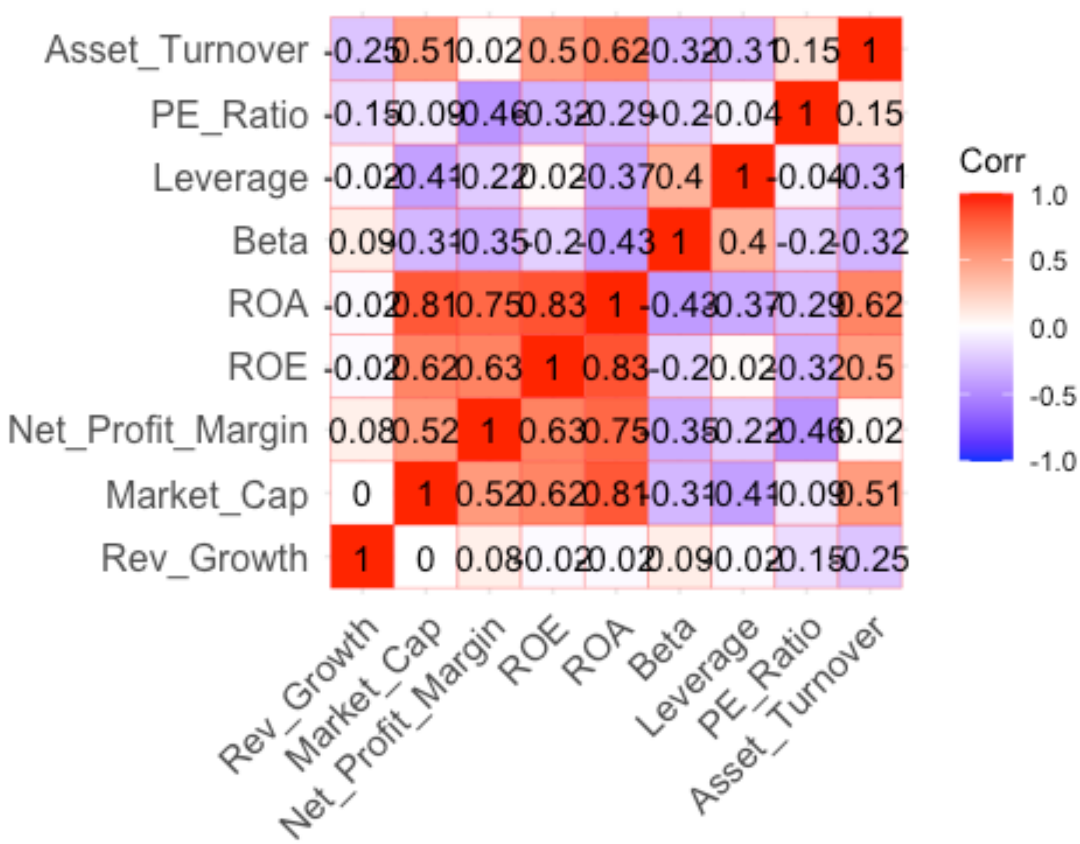

```
countries <- P[,c(1,2)]
unique(countries)

##    Symbol                               Name
## 1     ABT               Abbott Laboratories
## 2     AGN                    Allergan, Inc.
## 3     AHM                      Amersham plc
## 4     AZN                   AstraZeneca PLC
## 5     AVE                           Aventis
## 6     BAY                          Bayer AG
## 7     BMY        Bristol-Myers Squibb Company
## 8    CHTT                     Chattem, Inc
## 9     ELN             Elan Corporation, plc
## 10    LLY             Eli Lilly and Company
## 11    GSK               GlaxoSmithKline plc
## 12    IVX                  IVAX Corporation
## 13    JNJ                 Johnson & Johnson
## 14    MRX Medicis Pharmaceutical Corporation
## 15    MRK                 Merck & Co., Inc.
## 16    NVS                        Novartis AG
## 17    PFE                         Pfizer Inc
## 18    PHA             Pharmacia Corporation
## 19    SGP         Schering-Plough Corporation
```

```
## 20    WPI       Watson Pharmaceuticals, Inc.
## 21    WYE                               Wyeth
```

Colour intensity varies with increasing and decreasing distance. The heat-map below shows the separation between two Pharma companies observations. The red diagonals have a value of zero, and the dark blue diagonals have a value of six, indicating their extreme separation from one another.

#Determine whether the variables selected for clustering have any correlation with one another.

```
corr<-cor(P_Norm)
ggcorrplot(corr,outline.color = "red",lab = TRUE,hc.order = TRUE,type =
"full")
```




```
#The Market capitalization (market_cap), profit margin, and Return on equity
(ROE) all have a significant positive correlation with return on assets
(ROA). Accordingly, it is expected that the values of Market_cap, Profit
Margin, and ROE will rise along with the value of ROA, and vice versa.
```

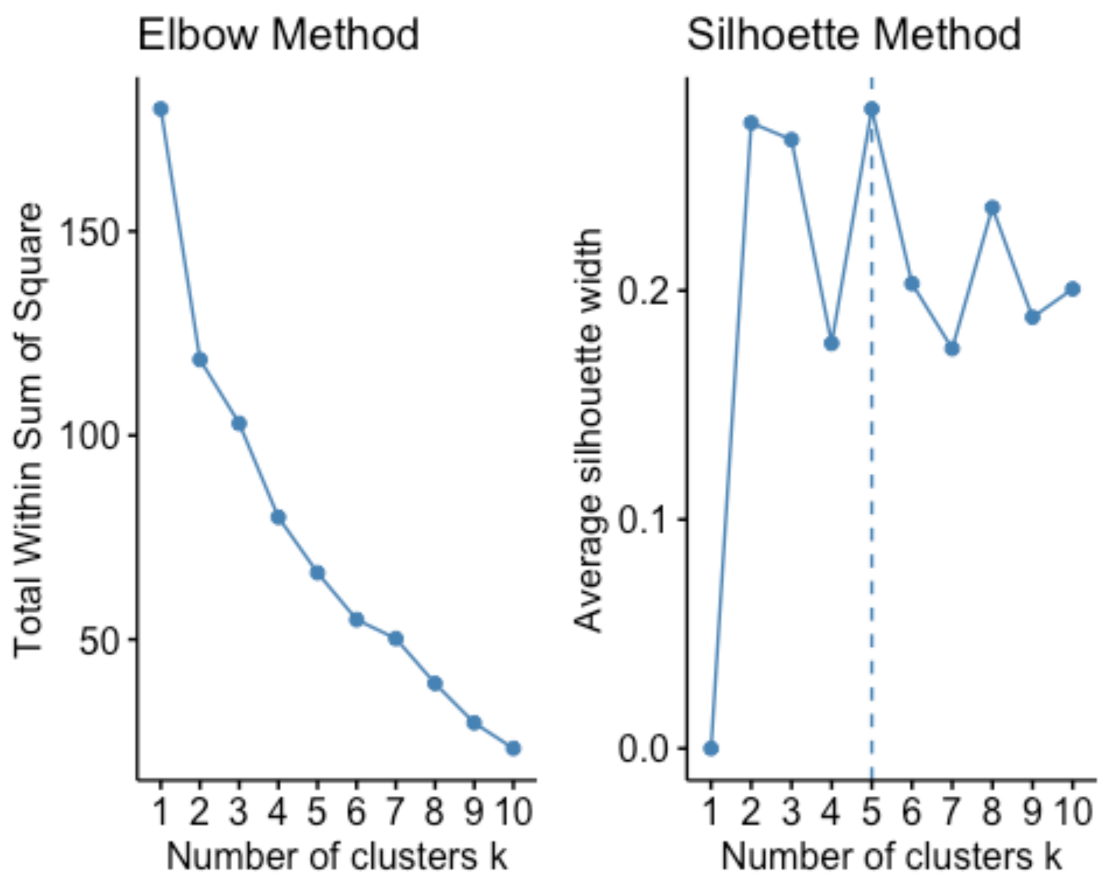
Problem Statement-2:- Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

#Finding the number of cluster's for grouping similar countries together.

#There are two main methods to find the value of K or number of cluster: Elbow chart and the Silhouette Method

#Determining the best value for k using an elbow chart Methods

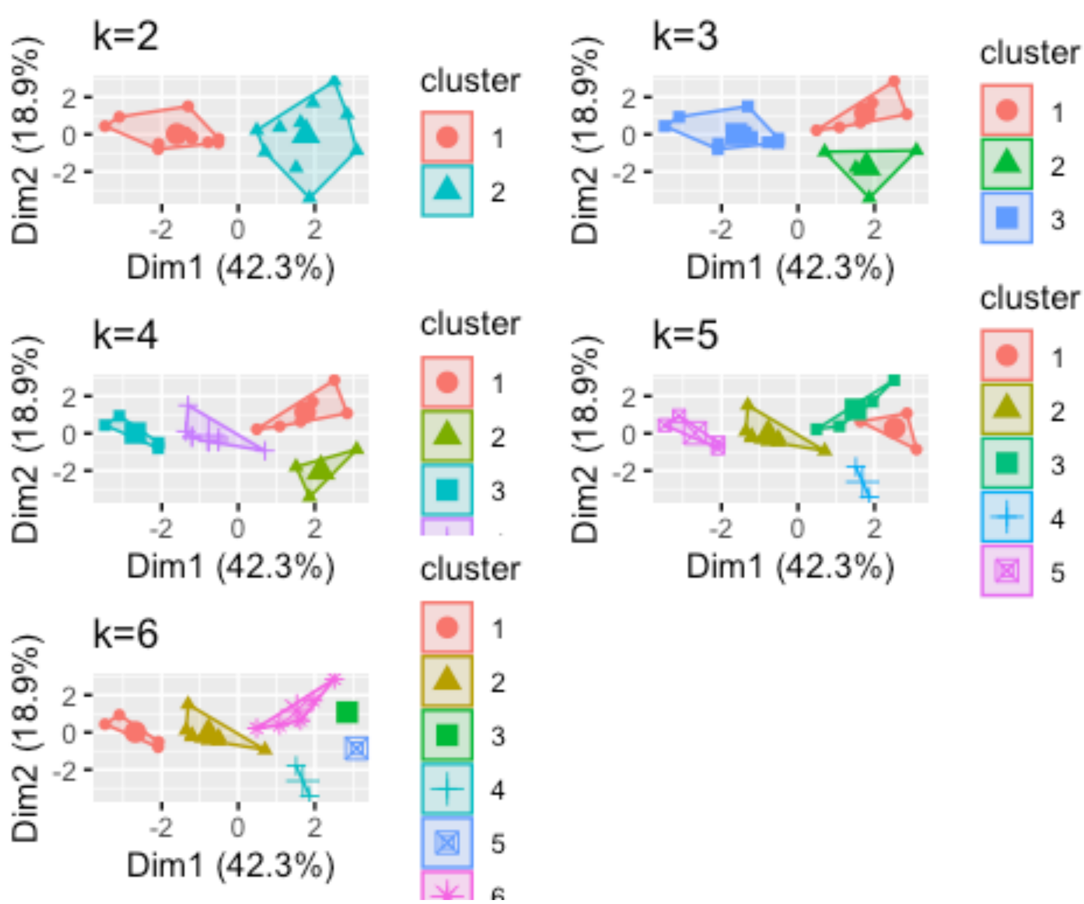
```
Elbow_method <- fviz_nbclust(P_Norm, kmeans, method = "wss")+ggtitle("Elbow
Method")

#Determining the best value for k using the Silhouette Method

Silhouette_method <- fviz_nbclust(P_Norm, kmeans, method =
"silhouette")+ggtitle("Silhoette Method")
plot_grid(Elbow_method, Silhouette_method, nrow = 1)
```



#As the silhouette approach indicates k=5, and the elbow method indicates k = 2 or 6, we are attempting to determine the ideal value of k. will examine every number between 2 and 6 and considering number of restarts = 25

```
k_2<-kmeans(P_Norm,centers =2,nstart=25)
k_3<-kmeans(P_Norm,centers =3,nstart=25)
k_4<-kmeans(P_Norm,centers =4,nstart=25)
k_5<-kmeans(P_Norm,centers =5,nstart=25)
```

```
k_6<-kmeans(P_Norm,centers =6,nstart=25)
p_1<-fviz_cluster(k_2,geom = "point", data=P_Norm)+ggtitle("k=2")
p_2<-fviz_cluster(k_3,geom = "point", data=P_Norm)+ggtitle("k=3")
p_3<-fviz_cluster(k_4,geom = "point", data=P_Norm)+ggtitle("k=4")
p_4<-fviz_cluster(k_5,geom = "point", data=P_Norm)+ggtitle("k=5")
p_5<-fviz_cluster(k_6,geom = "point", data=P_Norm)+ggtitle("k=6")
library(gridExtra)

##
## Attaching package: 'gridExtra'

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

grid.arrange(p_1,p_2,p_3,p_4,p_5)#The value 5 has no overlap and also
creating 5 different clusters
```



#Since value of K = 5 is making more sense will create 5 clusters for our analysis

```
P_Kmeans <- kmeans(P_Norm, centers = 5, nstart = 25)
pandoc.table(P_Kmeans$centers,style="grid", split.tables = Inf)
```

```
##
##
## +------------+---------+----------+---------+---------+---------------+--
--------+-----------+------------------+
## | Market_Cap |  Beta   | PE_Ratio |   ROE   |   ROA   | Asset_Turnover |
Leverage | Rev_Growth | Net_Profit_Margin |
##
+============+=========+==========+=========+=========+===============+=====
=====+===========+==================+
## |  -0.03142  | -0.4361 | -0.3172  |  0.195  | 0.4084  |     0.173     | -
0.2745  |  -0.7042   |       0.557       |
## +------------+---------+----------+---------+---------+---------------+--
--------+-----------+------------------+
## |  -0.4393   | -0.4702 |   2.7    | -0.835  | -0.9235 |     0.2306    | -
0.1417  |  -0.1168   |      -1.417       |
## +------------+---------+----------+---------+---------+---------------+--
--------+-----------+------------------+
## |  -0.7602   | 0.2796  | -0.4774  | -0.7438 | -0.8107 |     -1.268    |
0.06308  |   1.518    |     -0.006894     |
## +------------+---------+----------+---------+---------+---------------+--
--------+-----------+------------------+
## |  -0.8705   |  1.341  | -0.05284 | -0.6184 | -1.193  |    -0.4613    |
1.366   |  -0.6913   |       -1.32       |
## +------------+---------+----------+---------+---------+---------------+--
--------+-----------+------------------+
## |   1.696    | -0.1781 | -0.1985  |  1.235  |  1.35   |     1.153     | -
0.4681  |   0.4672   |      0.5912       |
## +------------+---------+----------+---------+---------+---------------+--
--------+-----------+------------------+
```

```
P_Kmeans$size
```

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

```
P_Kmeans$withinss
```

```
## [1] 21.879320  2.803505 12.791257 15.595925  9.284424
```

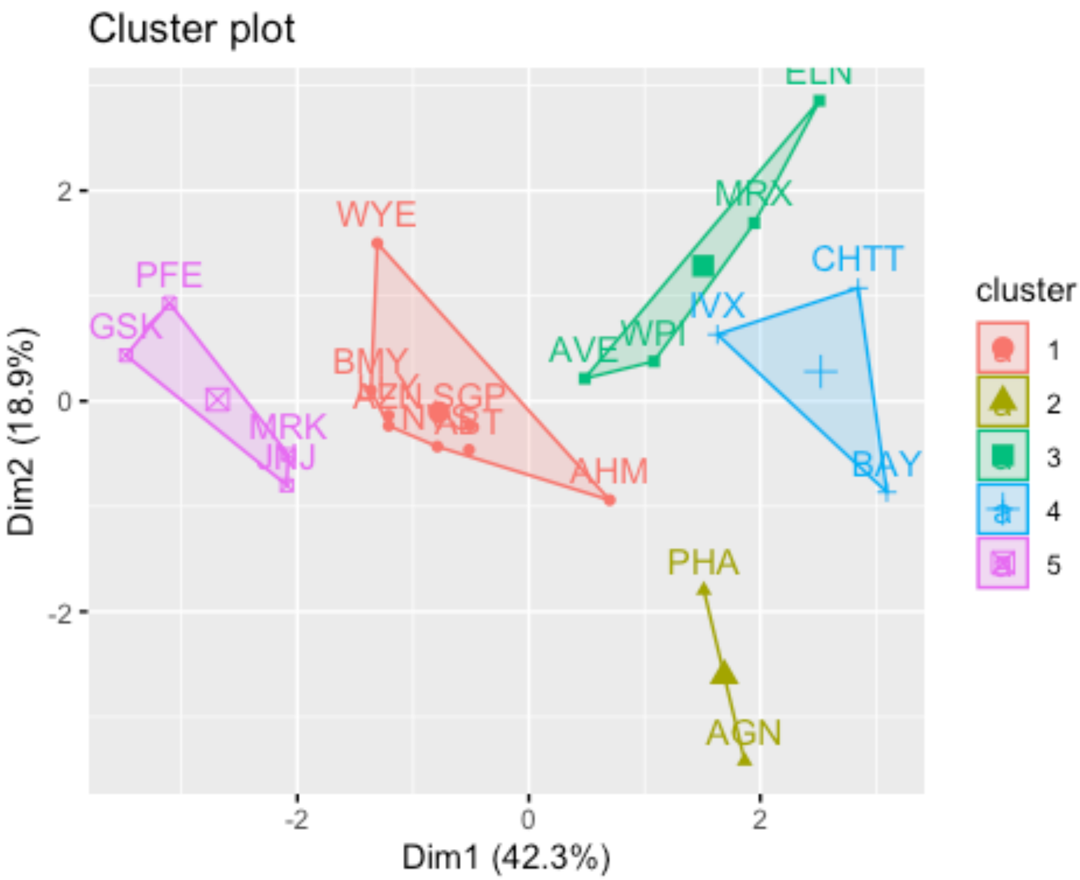
```
P_Kmeans$cluster[16]
```

```
## NVS
##   1
```

```
paste("The 16th Observation is country NVS and belongs to cluster",
P_Kmeans$cluster[16])
```

```
## [1] "The 16th Observation is country NVS and belongs to cluster 1"
```

```
fviz_cluster(P_Kmeans, data = P_Norm)
```

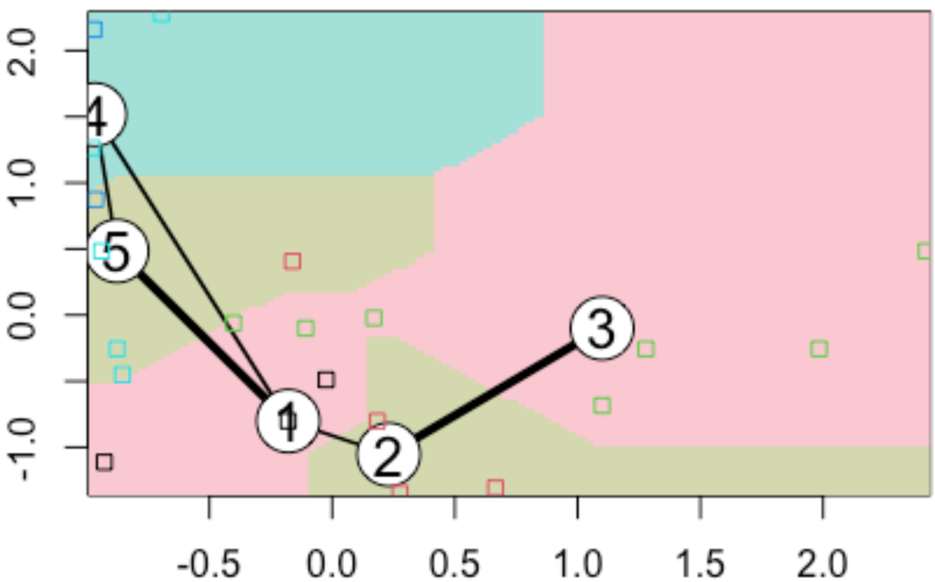



```
#By analyzing the findings, there are five clusters based on the entire data. While "cluster 2" has just two countries, "cluster 3" has the greatest number of firms.Moreover, Cluster 5 exhibits strong ROA and ROE asset turnover, while Cluster 2 displays a high PE ratio.Data dispersion is revealed by the cluster sum of square distance, cluster 2 (2.8) is less homogeneous than cluster 1 (21.9).
```

#Additionally, Kcca is being used to obtain the clusters rather than Kmeans because Kmeans uses the mean and KCCA utilises the KMedian.

```
#using k-means with k=3 for making clusters
set.seed(180)
P_KCCA_3 <- kcca(P_Norm, k = 5, kccaFamily("kmedians"))

## Found more than one class "kcca" in cache; using the first, from namespace 'kernlab'

## Also defined by 'flexclust'

## Found more than one class "kcca" in cache; using the first, from namespace 'kernlab'

## Also defined by 'flexclust'
```

```
P_KCCA_3

## kcca object of family 'kmedians'
##
## call:
## kcca(x = P_Norm, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 3 4 7 2 5

#Apply predict function

clusters_index <- predict(P_KCCA_3)
dist(P_KCCA_3@centers)

##          1        2        3        4
## 2 3.058723
## 3 3.233089 2.436130
## 4 3.075904 4.602752 4.887507
## 5 2.514770 3.447921 3.866226 3.329264

image(P_KCCA_3)
points(P_Norm, col = clusters_index, pch = 22, cex = 1)
```




```
#K-means clustering and KCCA have been used to solve the problem. While K-means clustering divides a set of data points into K clusters, KCCA is used to determine the correlation between two sets of variables. Furthermore, K-means clustering is an unsupervised learning technique that doesn't require any labelled data, whereas KCCA is a supervised learning method that does.
```

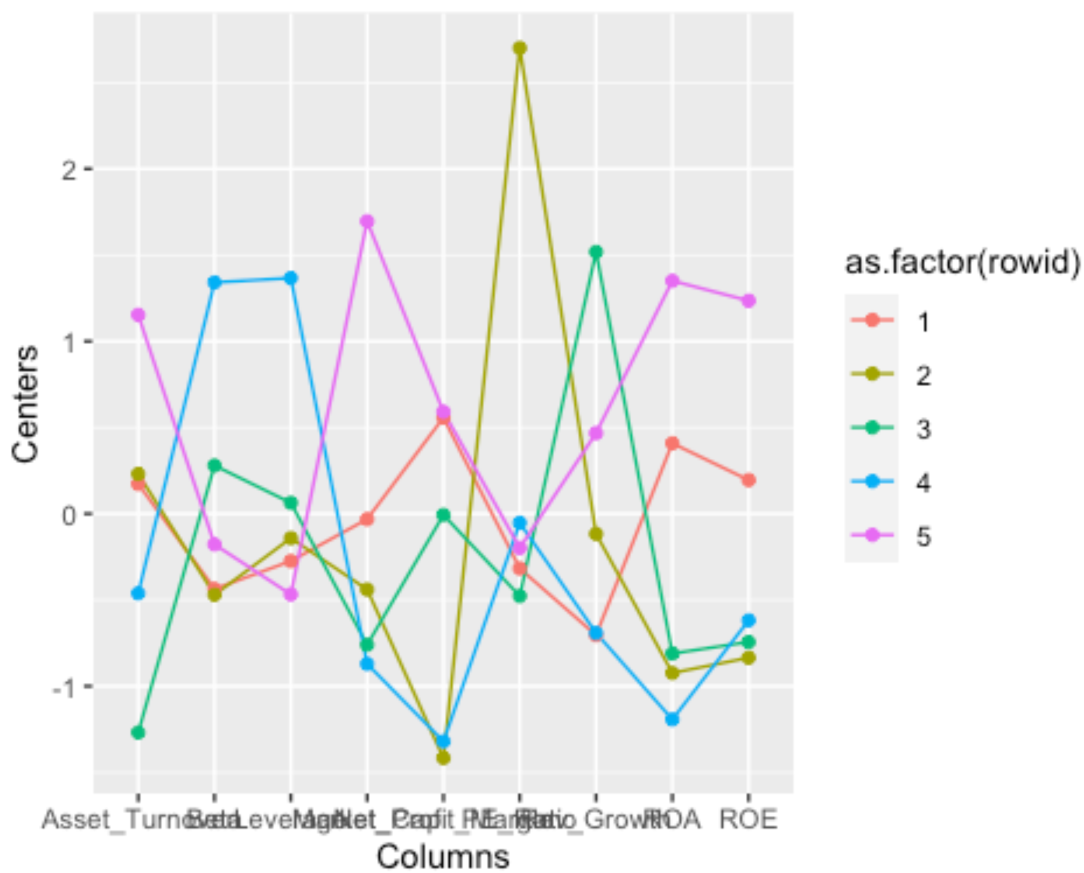
#Will Continue with cluster created by Kmeans since its more accurate for unsupervised learning method

```
#Plot of data grouped in clusters
Centroid_1 <- data.frame(P_Kmeans$centers) %>% rowid_to_column() %>%
gather('Columns', 'Centers', -1)
print(Centroid_1)
```

```
##    rowid       Columns      Centers
## 1      1    Market_Cap -0.031422109
## 2      2    Market_Cap -0.439251341
## 3      3    Market_Cap -0.760224892
## 4      4    Market_Cap -0.870515113
## 5      5    Market_Cap  1.695581115
## 6      1          Beta -0.436098941
## 7      2          Beta -0.470180039
## 8      3          Beta  0.279604106
## 9      4          Beta  1.340986857
```

```
## 10     5              Beta -0.178056346
## 11     1          PE_Ratio -0.317248516
## 12     2          PE_Ratio  2.700024643
## 13     3          PE_Ratio -0.477423799
## 14     4          PE_Ratio -0.052844340
## 15     5          PE_Ratio -0.198458234
## 16     1               ROE  0.195045857
## 17     2               ROE -0.834952524
## 18     3               ROE -0.743802224
## 19     4               ROE -0.618401510
## 20     5               ROE  1.234987906
## 21     1               ROA  0.408391543
## 22     2               ROA -0.923495091
## 23     3               ROA -0.810742783
## 24     4               ROA -1.192847826
## 25     5               ROA  1.350343113
## 26     1    Asset_Turnover  0.172974602
## 27     2    Asset_Turnover  0.230632802
## 28     3    Asset_Turnover -1.268480411
## 29     4    Asset_Turnover -0.461265604
## 30     5    Asset_Turnover  1.153164010
## 31     1          Leverage -0.274493115
## 32     2          Leverage -0.141703357
## 33     3          Leverage  0.063080849
## 34     4          Leverage  1.366446992
## 35     5          Leverage -0.468078185
## 36     1        Rev_Growth -0.704151557
## 37     2        Rev_Growth -0.116845875
## 38     3        Rev_Growth  1.518015830
## 39     4        Rev_Growth -0.691291399
## 40     5        Rev_Growth  0.467178770
## 41     1 Net_Profit_Margin  0.556954446
## 42     2 Net_Profit_Margin -1.416514761
## 43     3 Net_Profit_Margin -0.006893899
## 44     4 Net_Profit_Margin -1.320000179
## 45     5 Net_Profit_Margin  0.591242521
```

```r
ggplot(Centroid_1, aes(x = Columns, y = Centers, color = as.factor(rowid))) +
geom_line(aes(group = as.factor(rowid))) + geom_point()
```




```
#The graph demonstrates that companies in cluster.1 have a high price to
earnings ratio and a low net profit margin, whereas companies in cluster 3
have a high leverage ratio, a low return on asset (ROA), and a low asset
turnover rate. However, Cluster 2 did not stand out in relation to any of the
factors we looked at.
```

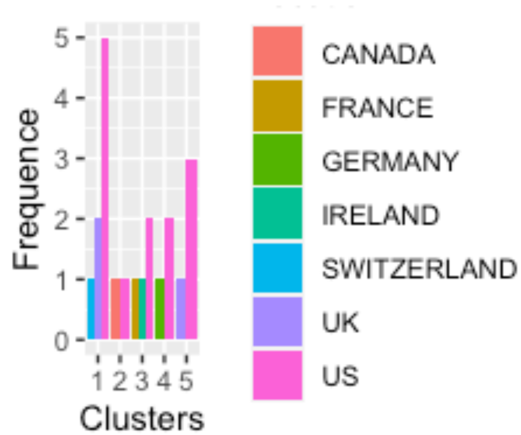
#Checking if there is any pattern in the clusters with respect to the numerical variables (10 to 12)?

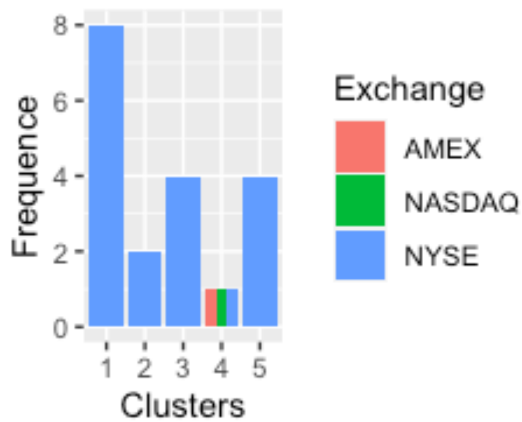
```
P_Pattern <-  P %>% select(c(12,13,14)) %>% mutate(Cluster =
P_Kmeans$cluster)
print(P_Pattern) #The remaining three category to be considered are Stock
Exchange, Location, and Median Recommendation.

##    Median_Recommendation    Location Exchange Cluster
## 1           Moderate Buy          US     NYSE       1
## 2           Moderate Buy      CANADA     NYSE       2
## 3             Strong Buy          UK     NYSE       1
## 4          Moderate Sell          UK     NYSE       1
## 5           Moderate Buy      FRANCE     NYSE       3
## 6                   Hold     GERMANY     NYSE       4
## 7          Moderate Sell          US     NYSE       1
## 8           Moderate Buy          US   NASDAQ       4
## 9          Moderate Sell     IRELAND     NYSE       3
```

```
## 10                  Hold          US      NYSE        1
## 11                  Hold          UK      NYSE        5
## 12                  Hold          US      AMEX        4
## 13          Moderate Buy          US      NYSE        5
## 14          Moderate Buy          US      NYSE        3
## 15                  Hold          US      NYSE        5
## 16                  Hold SWITZERLAND      NYSE        1
## 17          Moderate Buy          US      NYSE        5
## 18                  Hold          US      NYSE        2
## 19                  Hold          US      NYSE        1
## 20         Moderate Sell          US      NYSE        3
## 21                  Hold          US      NYSE        1
```

#Using the bar charts to visualise the distribution of firms organised by clusters and to spot any data trends.

```
Median_Recom <- ggplot(P_Pattern, mapping = aes(factor(Cluster),
fill=Median_Recommendation)) +
  geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')

Location_0 <- ggplot(P_Pattern, mapping = aes(factor(Cluster),
fill=Location)) + geom_bar(position = 'dodge') + labs(x='Clusters',
y='Frequence')

Exchange_0 <- ggplot(P_Pattern, mapping = aes(factor(Cluster),
fill=Exchange)) +
geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
plot_grid(Median_Recom, Location_0, Exchange_0)
```





#According to the clustering analysis, the firms in each cluster share comparable attributes with regard to their Exchange, Location, and Median Recommendation.

Cluster-1:- The majority of the companies in cluster 1 are US-based firms that are listed on the New York Stock Exchange. Their stock has a broad recommendation to hold, indicating that they are dependable and generally low-risk investments.

Cluster -2:- The companies in cluster 2 are a combination of US and Canadian firms listed on the NYSE, are recommended to be bought or held because they have the potential for growth but may also carry some risk.

Cluster-3:- The companies in cluster 3 are listed on the NYSE and come from different places, have a modest buy or sell recommendation, indicating that there may be room for growth.

Cluster-4:- Companies in cluster 4 are based in the USA and Germany and are listed on stock exchanges other than NYSE (AMEX and NASDAQ), are recommended for a hold or modest purchase.

Cluster-5:- Companies in cluster 5 includes companies from the UK and the USA, have partially hold and buy recommendations for their NYSE-listed stocks, suggesting that they may have some growth potential but also considerable risk.

problem statement -3 :- Provide an appropriate name for each cluster using any or all of the variables in the dataset.

21 pharmaceutical firms can be divided into 5 groups based on the characteristics of the clusters and the detailed analysis as done.

Cluster 1: "Stable - efficient companies":- Businesses with normal levels for all financial parameters are thought to be running effectively and efficiently in their sector and against competitors. Additionally, American-based businesses that are listed on the New York shares Exchange dominate it. These businesses have a spread advise to hold onto their shares, implying that they are reliable and reasonably low-risk investments. Cluster 1 is characterized by high market capital, high ROE, high ROA, and high asset turnover.

Cluster 2: "Overpriced - Risky companies":- Despite the company's relatively low net profit margin, the market is valuing its shares at a premium to its present earnings due to its high price-to-earnings (PE) ratio and low net profit margin. It indicates that, despite the company's low profit margin relative to revenue, investors are prepared to pay a premium for each dollar of earnings the company makes.These businesses carry some risk since their stock price can drop in the future if they are unable to live up to the expectations of the market.

Cluster 3: "Growth oriented- Low risky companies":- A business that exhibits strong revenue growth along with low asset turnover may be a sign of substantial growth potential that isn't being realized at this time due to inefficient operations. Investors ought to take into account the industry and competitive environment of the business in addition to its capacity to maintain rapid revenue growth in the long run. It's also critical to assess the profitability of the business, since even with strong sales growth, profits may not increase if the company is not making the most use of its resources.Additionally, these are the companies from different regions that are listed on the New York Stock Exchange (NYSE), and their moderate buy or sell recommendation implies that they might have room for growth. Finally, Cluster 3 is characterized by similar beta values, high price/earnings ratio, and low ROE ROA, net profit margin.

Cluster 4- "Debt-ridden - very risky companies":- High leverage and low ROA and net profit margin may be signs that a company is borrowing a lot of money to fund its operations while producing insufficient profits or returns on assets. Investors may find this to be a worrying indication because it could be difficult for the business to pay off its debt and eventually get into financial difficulties.Additionally, they are recommended for holds or moderate buys on stock exchange marketplaces other than the New York Stock Exchange (AMEX and NASDAQ). Finally, Cluster 4 is characterized by below average ROE, ROA, and asset turnover with high estimated revenue growth.

Cluster 5- "Established - profitable companies":- Large, well-established businesses with a good financial position and a substantial market presence are usually those with a high market capitalization. A corporation with a high market capitalization has many outstanding shares and a high stock price, which contributes to a high value.Additionally, they have a buy and partially hold rating on the NYSE-listed equities they own.

In a simple, i can state that

#Cluster 1: Hold cluster -They have decent numbers.

#Cluster 2: Moderate Buy (or) Hold cluster.

#Cluster 3: Buy or Sell Cluster

#Cluster 4: Buy Cluster - It has good stability.

#Cluster 5: High Hold cluster