---
title: "Assignment4_ClusterAnalysis"
output: html_document
date: '2022-03-18'
output:
html_document: default
pdf_document: default

```
rm(list=ls())

library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## 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)
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

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

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

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

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

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

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

```
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     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()
```

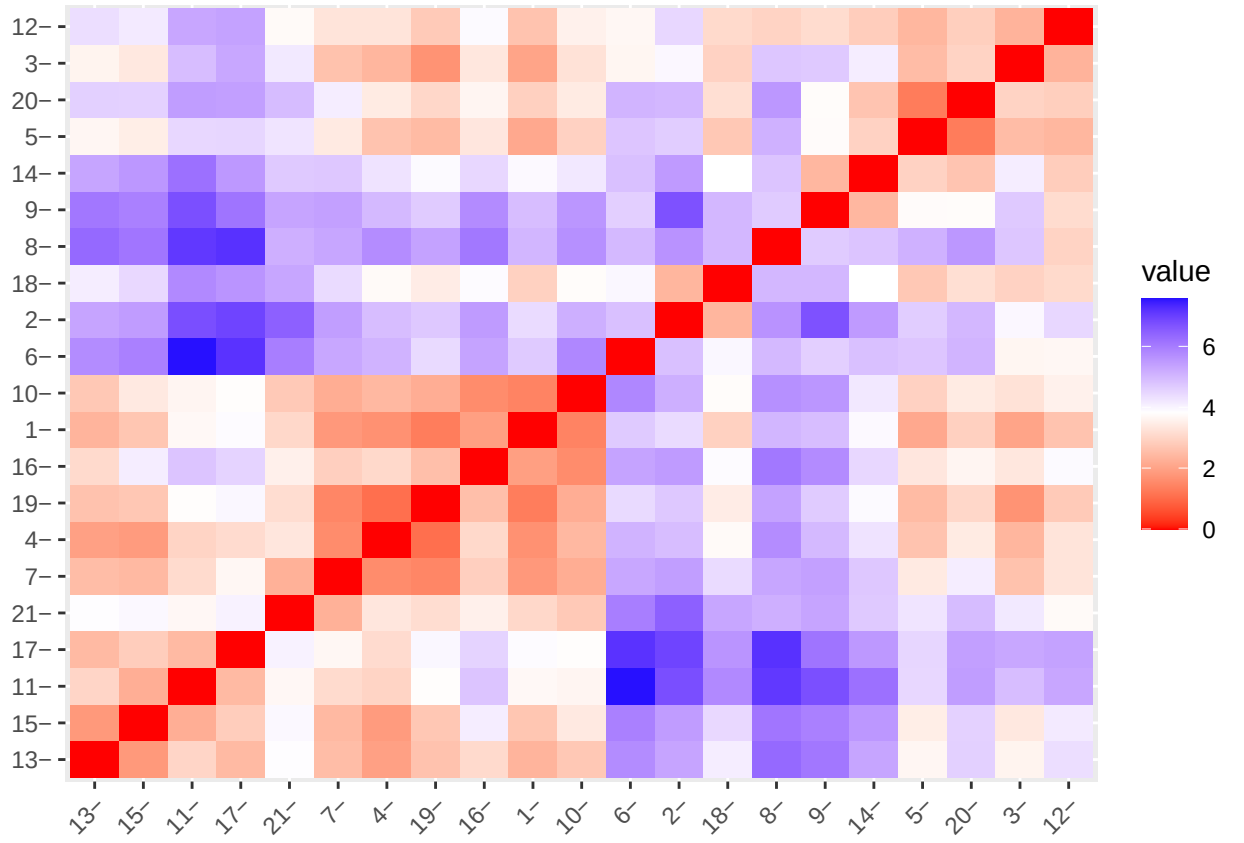
```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.1.3
```

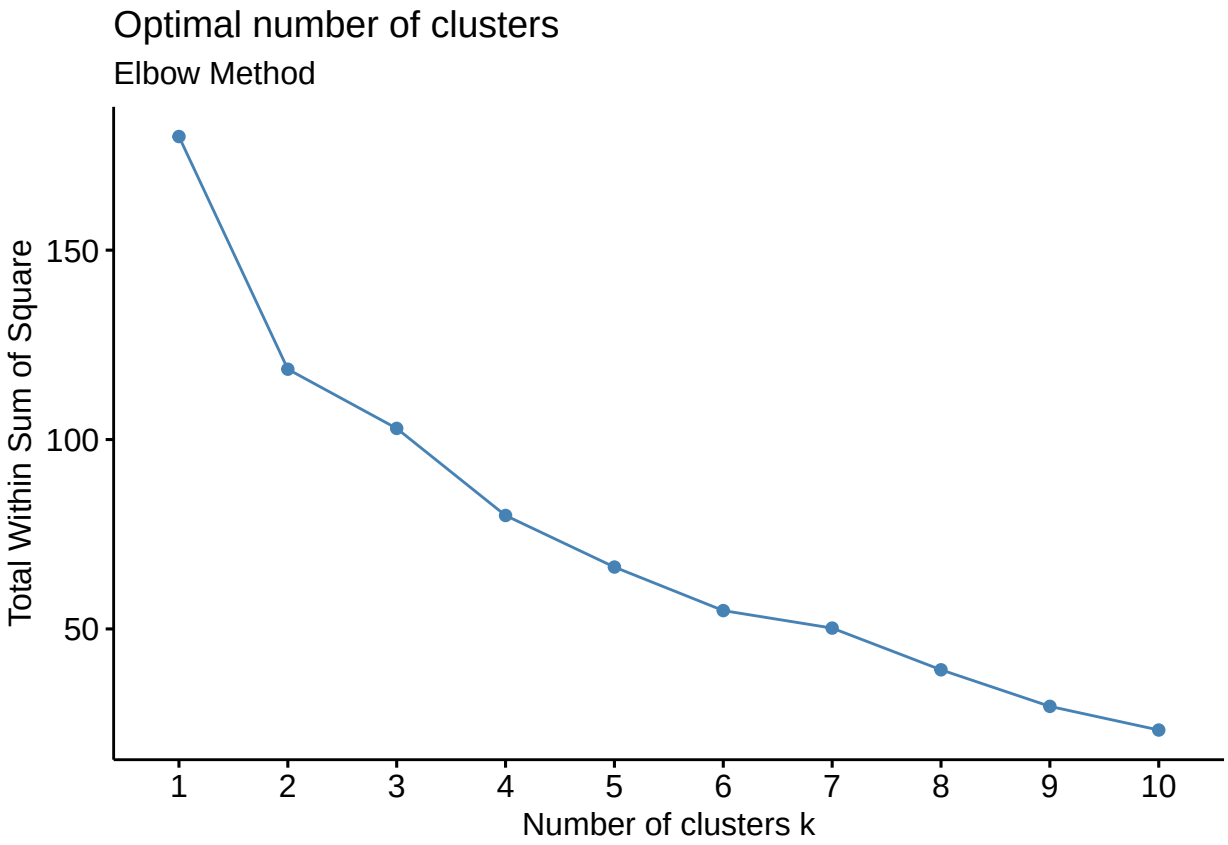
```
#read Pharmaceuticals.csv

pharma_df <- read.csv('C:\\Users\\madhu\\Downloads\\Pharmaceuticals.csv')

colSums(is.na(pharma_df)) # verify null column sums
```

```
##                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
```

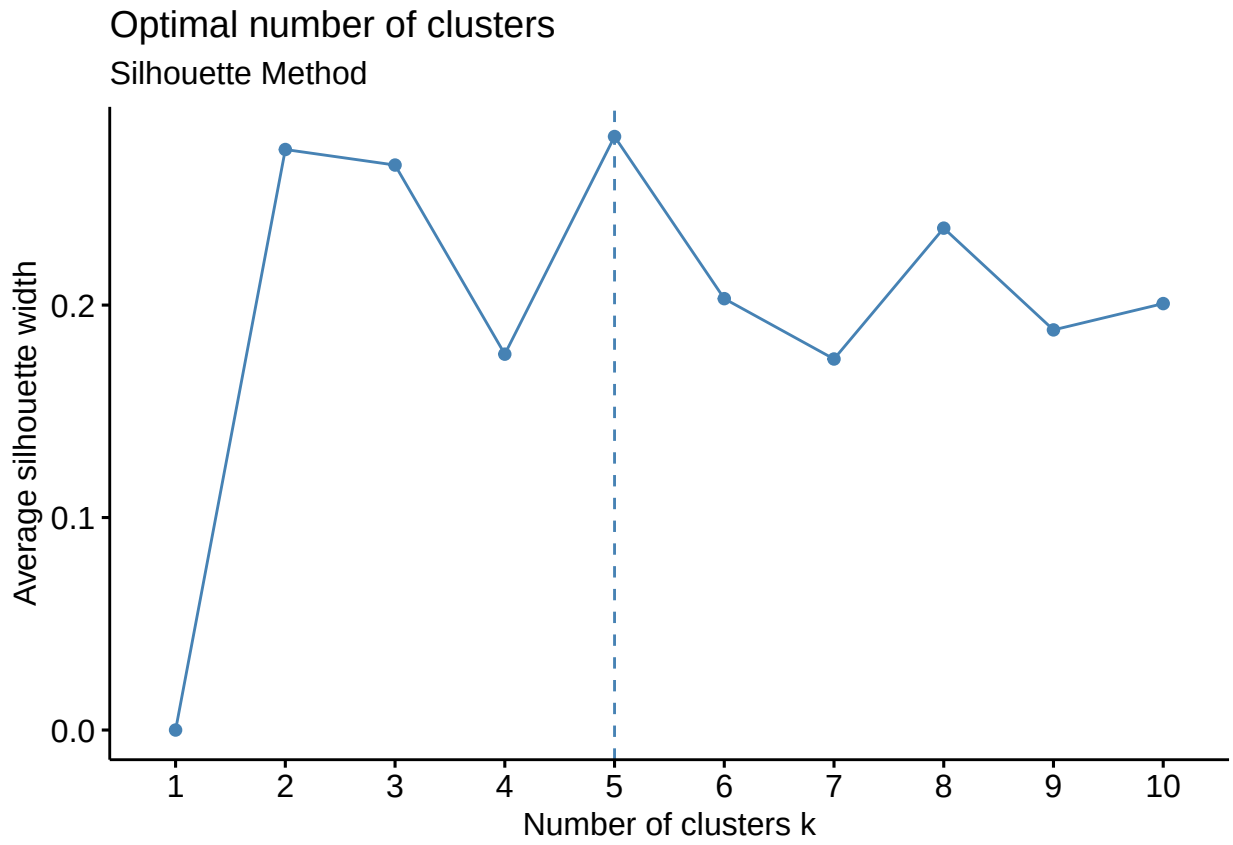
```
#Cluster analysis for pharma

#a. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made
#conducting the cluster analysis, such as weights for different variables, the specific clustering algo

pharma_df_numeric <- pharma_df[,c(3:11)] #numerical from 3 to 11

#scale quantitative variables in DF by z-score because normalization is very important in cluster analy

nor <- as.data.frame(scale(pharma_df_numeric))
distance <- get_dist(nor)
fviz_dist(distance) #visuvalize a distance matrix
```




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

fviz_nbclust(nor, FUNcluster = kmeans, method = "wss") + labs(subtitle = "Elbow Method")
```



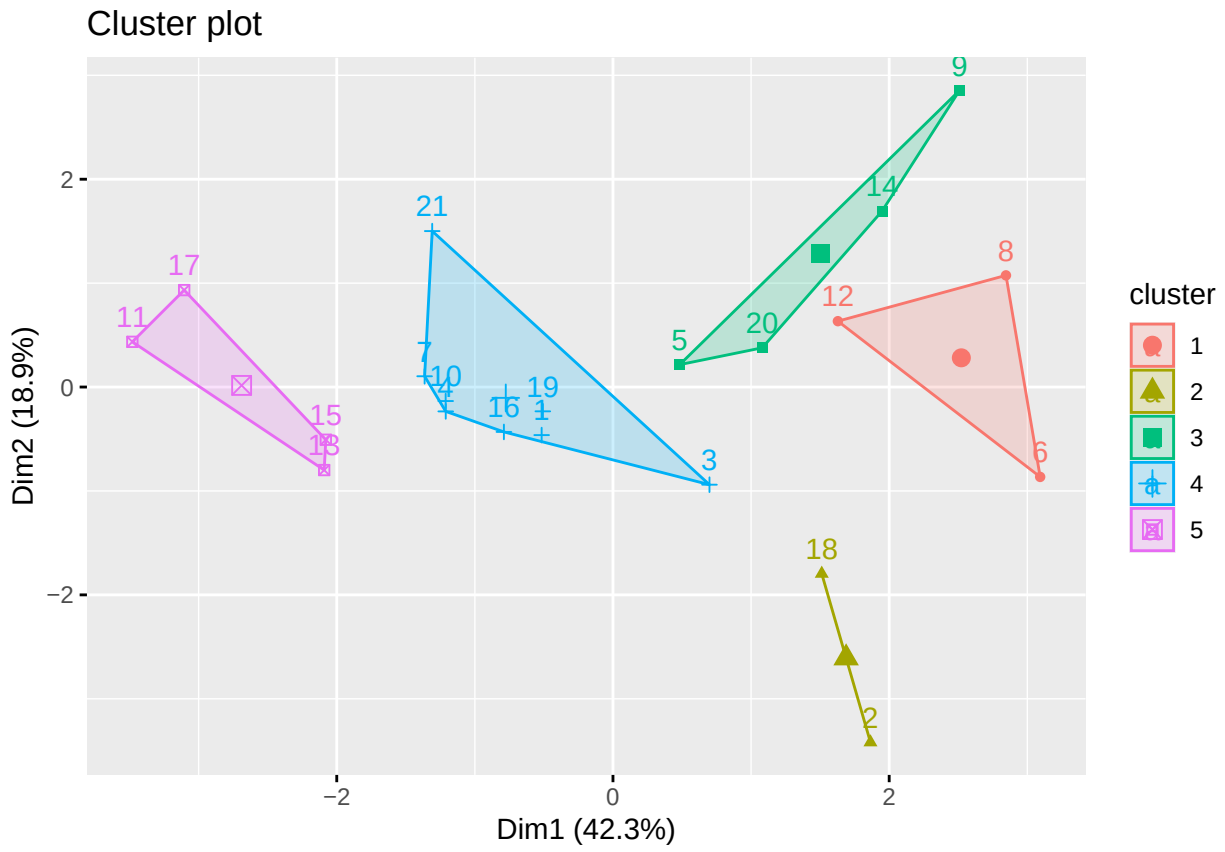

```
# Silhouette Method on scaled data to determine the number of clusters
fviz_nbclust(nor,FUNcluster = kmeans,method = "silhouette")+labs(subtitle="Silhouette Method")
```




```
## The plots reveal that 5 clusters are sufficient to capture the data variations

set.seed(64060)
k5 <- kmeans(nor, center = 5, nstart = 25) # where k = 5
k5$centers #centriods
```

```
##    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 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 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  1.36644699 -0.6912914      -1.320000179
## 2 -0.14170336 -0.1168459      -1.416514761
## 3  0.06308085  1.5180158      -0.006893899
## 4 -0.27449312 -0.7041516       0.556954446
## 5 -0.46807818  0.4671788       0.591242521
```

```
fviz_cluster(k5, data = nor) #cluster plot viz
```




```
k5$size
```

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

```
#K-Means Cluster Analysis - Fit the data with 5 clusters

data_fit <- kmeans(nor, 5)
aggregate(nor, by = list(data_fit$cluster), FUN = mean)
```

```
##   Group.1  Market_Cap       Beta   PE_Ratio        ROE        ROA
## 1       1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431
## 2       2 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022
## 3       3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792
## 4       4 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838
## 5       5  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1   1.153164e+00 -0.4680782  0.4671788         0.5912425
## 2  -1.537552e-01 -0.4040831  0.6917224        -0.4005718
## 3  -1.153164e+00  1.4773718  0.7120120        -0.3688236
## 4   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   6.589509e-02 -0.2559803 -0.7230135         0.7343816
```

```
norm <- as.data.frame(nor, data_fit$cluster)
norm
```

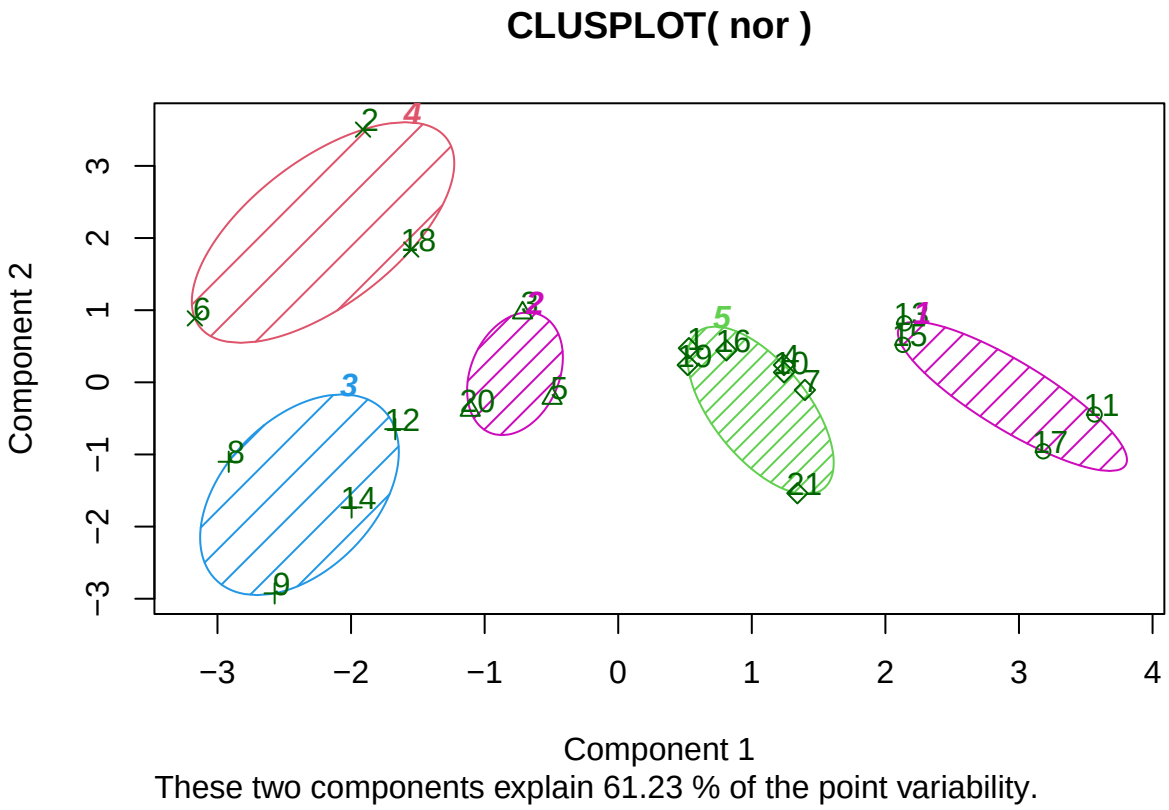
```
##   Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 5  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## 4 -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## 2 -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## 5  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## 2 -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## 4 -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
## 5 -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
## 3 -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
## 3 -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## 5  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## 1  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## 3 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## 1  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## 3 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## 1  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## 5  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## 1  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## 4 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## 5 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## 2 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## 5 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
##      Leverage  Rev_Growth Net_Profit_Margin
## 5 -0.21209793 -0.52776752        0.06168225
## 4  0.01828430 -0.38113909       -1.55366706
## 2 -0.40408312 -0.57211809       -0.68503583
## 5 -0.74965647  0.14744734        0.35122600
## 2 -0.31449003  1.21638667       -0.42597037
## 4 -0.74965647 -1.49714434       -1.99560225
## 5 -0.02011273 -0.96584257        0.74744375
## 3  3.74279705 -0.63276071       -1.24888417
## 3  0.61983791  1.88617085       -0.36501379
## 5 -0.07130879 -0.64814764        1.17413980
## 1 -0.31449003  0.76926048        0.82363947
## 3  1.10620040  0.05603085       -0.71551412
## 1 -0.62166634 -0.36213170        0.33598685
## 3  0.44065173  1.53860717        0.85411776
## 1 -0.39128411  0.36014907       -0.24310064
## 5 -0.67286239 -1.45369888        1.02174835
## 1 -0.54487226  1.10143723        1.44844440
## 4 -0.30169102  0.14744734       -1.27936246
## 5 -0.74965647 -0.43544591        0.29026942
## 2 -0.49367621  1.43089863       -0.09070919
## 5  0.68383297 -1.17763919        1.49416183
```

#(b)Interpret the clusters with respect to the numerical variables used in forming the clusters

```
cluster 1 - Row 8, 6, 12
cluster 2 - Row 2, 18
cluster 3 - Row 5, 9, 14, 20
cluster 4 - Row 3, 4, 7, 10, 16, 19, 21
cluster 5 - Row 11, 13, 15, 17

By the output of function::  aggregate(nor, by = list(data_fit$cluster), FUN = mean), we can observe th
```

```
cluster 1 has highest Market_Cap, highest ROE, highest ROA, lowest Leverage and lowest Beta
cluster 2 has lowest Beta, lowest PE_Ratio
cluster 3 has lowest Market_Cap, highest Beta, highest Leverage, highest Rev_Growth, lowest PE_ratio
cluster 4 has highest PE_Ratio, lowest ROE, lowest ROA, lowest Net_Profit_Margin
cluster 5 has highest Asset_Turnover, lowest Revenue growth, highest Net_Profit_Margin
```

```
#cluster plot

clusplot(nor, data_fit$cluster, color = TRUE, shade =TRUE, labels = 2, lines = 0)
```




```
#(c)Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

Moderate buy, hold, strong buy recommendations

Cluster 1 has highest ROE, highest ROA, highest Market_Cap but Rev_Growth is not indicated to moderate s

Cluster 2 has lowest Beta, lowest Asset_Turnover so hold Recommendation

Cluster 3 has highest Beta, highest Leverage, highest Rev_Growth is strong to buy Recommendation

Cluster 4 has highest PE_Ratio, lowest ROE, ROA, Net_Profit_Margin is to hold buy Recommendation

Cluster 5 has highest Asset_Turnover, highest Net_Profit_Margin, lowest revenue growth is risky but to h
```

```
Cluster 5 and Cluster 3 moderate to buy Recommendation
Cluster 1,4 is hold Recommendation

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

  Cluster 1 - highest Market_Cap,highest Leverage,highest Rev_Growth, lowest Leverage and Beta cluster -

  Cluster 2 - lowest Rev_Growth,lowest PE_Ratio cluster - on hold

  Cluster 3 - lowest PE_Ratio,lowest_ROE,lowest ROA, highest Leverage, highest Rev_growth, lowest Net_Pr

  Cluster 4 - highest PE_Ratio, lowest ROA,lowest Asset_Turnover, lowest Net_Profit_Margin cluster - - h

  Cluster 5 - highest Asset_Turnover, Net_Profit_Margin, lowest Rev_Growth cluster - strong buy
```