FML Assignment_4

Bhargav

2023-10-27

Summary

Questions - Answers

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. A. For clustering of this data we need to consider all the numeric variables from 1 to 9, because we need to cluster and suggest the companies depending on their equity, equity depends on financial factors such as profit, market value, price to earning ratio, Return on equity, Return on assets, leverage, etc. all these factors were defined in 1 to 9 variables and the weights for different variables were taken equally, as they all play equal role in defining the equity of the firm.
2. Market Capitalization: it shows the company's total size and market value.
3. Beta: it indicates how sensitive a company's returns are to changes in the market.
4. PE Ratio: it expresses the value of a company's stock in relation to its earnings.
5. ROE: it shows the efficiency with which a company uses shareholder equity to turn a profit.
6. ROA : it Evaluates the capacity of an organization to make money off of its assets.
7. Asset Turnover: it Indicates how well a company uses its assets to produce income.
8. Leverage:it shows the extent to which a business uses debt to finance its operations.
9. Rev_Growth:it Shows the percentage change in revenue over a given time period.
10. Net Profit Margin: This variable shows the percentage of revenue that is converted to profit.

I have tried clustering with all the 3 algorithms Kmeans, DBSCAN and Hierarchical clustering. out of which Kmeans clustering gave me the best result this particular data set without any outliers and good number of clusters. when I tried with DBSCAN clustering, 2 clusters are formed with 15 points and the remaining 6points are showed as outliers without considering into any of the clusters. And DBSCAN can be used for more denser data. So, DBSCAN cannot be considered as a good method to cluster for this dataset. When I tried with Hierarchical clustering, four clusters are formed with some points in each cluster 11 being the highest size of the one cluster and 1 being the lowest size, but when I tried with Kmeans with no.of clusters as 5, the clusters formed are relatively better with no.of points and the distances from the centers than from hierachcial clustering. So, I have considered Kmeans Algorithm for the clustering of the Dataset. And the no.of clusters taken for the Kmeans clustering is 5, as I have taken the optimal value of 5 from silhouette method, and I have done the clustering with no.of clusters as 2 which was shown by Elbow method. but the the clusters formed by taking the no.of points as 5 are better as the points are more closer to the centroids. The clutsers formed are: first cluster with size of 4 and with companies AVE, WPI, MRX, ELN second cluster with size 2 and with companies PHA, AGN third cluster with size 4 and with companies GSK, PFE, MRK, JNJ fourth cluster with size 3 and with companies IVX, CHTT, BAY fifth cluster with size 8 and with companies WYE, BMY, LLY, AZN, NVS, ABT, SGP, AHM

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)? (those not used in forming the clusters) A. The clusters formed with respect to numerical variables are: Cluster1 with companies AVE, WPI, MRX, ELN has high revenue growth and beta value. but have low asset turnover, return on equity and return on asset.And the market capitalization is also relatively low. based on these, it is possible that these companies are still growing and they are at early stage. These companies might be investing heavily in marketing and sales. However, the high revenue growth and beta value suggest that they are expected to improve their earnings more rapidly in the coming days. these companies are distinguished by their higher growth potential and low profitability.

Cluster2 with companies PHA, AGN has high Price or earnings ratio and asset turnover, but have low net profit margin, return on equity and return on asset. and the market capitalization is also relatively low. However, the high asset turnover and price or earnings ratios suggest that they are expected to improve their earnings more rapidly in the future, while having little net profit in the past. However, with its high price, investors get more risk.

Cluster3 with companies IVX, CHTT, BAY has high market capitalization, return on equity, Return on assets and Asset turnover. but they have lowest Beta and profit to return Ratio. Based on these features these companies are matured and well established companies. the low beta value suggests that their stock prices are more stable, so that it was less risky to invest. but the low profit return ratio shows that they are not so efficient in generating profits. these companies are distinguished by their maturity, stability, and profitability.

Cluster4 with companies WYE, BMY, LLY, AZN, NVS, ABT, SGP, AHM has high beta value and leverage. but have lowest net profit margin, market capitalization. And relatively low return on equity, return on asset, revenue growth. based on these features, we can say that these companies are riskier to invest than other companies as they have high beta value which means their stock price was unstable and high leverage means more debts. and there profit margin is also low. but, if the market was high they can earn more profits due to that high beta value. these companies are distinguished by higher risk and potential for higher returns.

Cluster5 with companies GSK, PFE, MRK, JNJ has highest net profit margin, asset turnover, return on equity, Return on assets. but have lowest Beta, profit to return Ratio, revenue growth. these features shows that these companies have high financial performance and low risk. the high net profit margins, asset turnovers, returns on equity, and returns on assets, indicates efficient operations and strong profitability. and lowest beta value and revenue growth shows the stock price was more stable and less revenue growth.these represents a group of mature and well-established companies with strong financial performance and low risk profiles. **** The pattern with respect to the clusters with variables 10 to 12: Cluster1, Recommended as Moderate Buy and Moderate Sell from Locations France, Ireland and US and listed under NYSE.

Cluster2, Recommended as Hold and Moderate Buy from Locations US and canada and listed under NYSE.

Cluster3, Recommended as Hold and Moderate Buy from Locations UK and US and listed under NYSE.

Cluster4, Recommended as Hold and Moderate Buy from Locations Germany and US and listed under AMEX, NASDAQ and NYSE.

Cluster5, Recommended Hold, Moderate Sell, Strong Buy & Moderate Buy from Locations Switzerland, UK and US and listed under NYSE.

3.Provide an appropriate name for each cluster using any or all of the variables in the dataset. A. Appropriate names: Cluster1: High Growth potential Cluster2: High Risk High Reward Cluster3: Stability and Profitability Cluster4: High Beta High Risk Cluster5: Low Risk High Profitability

---

Problem Statement

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv Download Pharmaceuticals.csv. For each firm, the following variables are recorded:
Market capitalization (in billions of dollars) Beta Price/earnings ratio Return on equity Return on assets Asset turnover Leverage Estimated revenue growth Net profit margin Median recommendation (across major brokerages) Location of firm's headquarters Stock exchange on which the firm is listed Use cluster analysis to explore and analyze the given dataset as follows: ******* # Data Import And Cleaning

Load the Required Libraries

```r
library(class)
library(caret)
```

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

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

```r
library(e1071)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
```

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

```r
library(ISLR)
library(factoextra)
```

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

```r
library(dbscan)
```

```
##
## Attaching package: 'dbscan'
##
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```
library(cluster)
library(klustR)
```

```
## Warning: package 'klustR' was built under R version 4.3.2
```

```
library(ggplot2)
library(dplyr)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.3.2
```

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

Import the data which was in CSV format

```
# import the data
pharma.data <- read.csv("C:/Users/BHARGAV/OneDrive/Desktop/FML Assignment/Assignment_4/Pharmaceuticals.c
dim(pharma.data)
```

```
## [1] 21 14
```

```
t(t(names(pharma.data)))# The 't' function creates a transpose of the dataframe
```

```
##       [,1]
##  [1,] "Symbol"
##  [2,] "Name"
##  [3,] "Market_Cap"
##  [4,] "Beta"
##  [5,] "PE_Ratio"
##  [6,] "ROE"
##  [7,] "ROA"
##  [8,] "Asset_Turnover"
##  [9,] "Leverage"
## [10,] "Rev_Growth"
## [11,] "Net_Profit_Margin"
## [12,] "Median_Recommendation"
## [13,] "Location"
## [14,] "Exchange"
```

Dropping the columns that were not required for clustering

```
# Remove the unwanted columns
row.names(pharma.data) <- pharma.data[,1]
clust.data <- pharma.data[ ,3:11]# 1 and 5 are the indexes for columns ID and ZIP
dim(clust.data)
```

```
## [1] 21  9
```

```r
# Summary of the data
summary(clust.data)
```

```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
##       ROA        Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

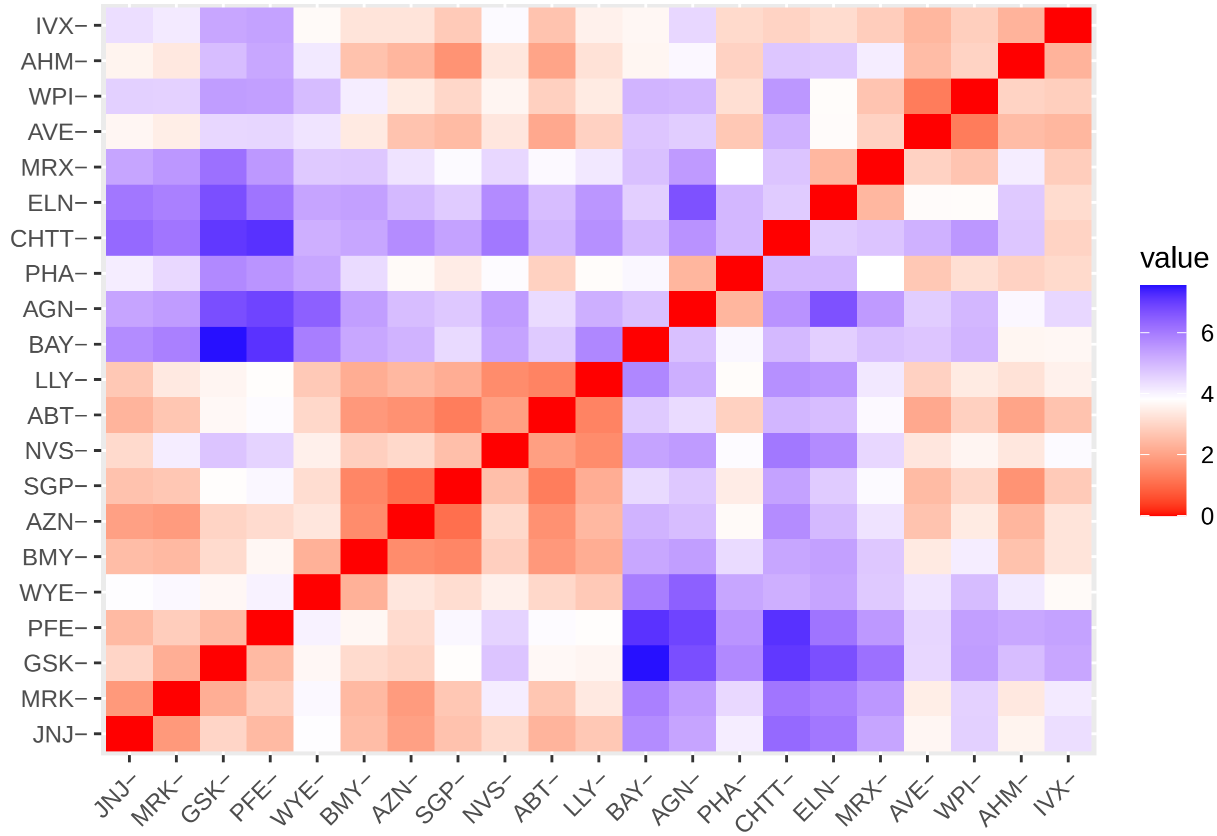
Scaling the data

```r
# scale the data using scale function
scaled.data <- scale(clust.data)
head(scaled.data)
```

```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
##       Leverage Rev_Growth Net_Profit_Margin
## ABT -0.2120979 -0.5277675        0.06168225
## AGN  0.0182843 -0.3811391       -1.55366706
## AHM -0.4040831 -0.5721181       -0.68503583
## AZN -0.7496565  0.1474473        0.35122600
## AVE -0.3144900  1.2163867       -0.42597037
## BAY -0.7496565 -1.4971443       -1.99560225
```

```r
# distance between each variable
distance <- get_dist(scaled.data)
```

```r
# Visualize the distance
fviz_dist(distance)
```



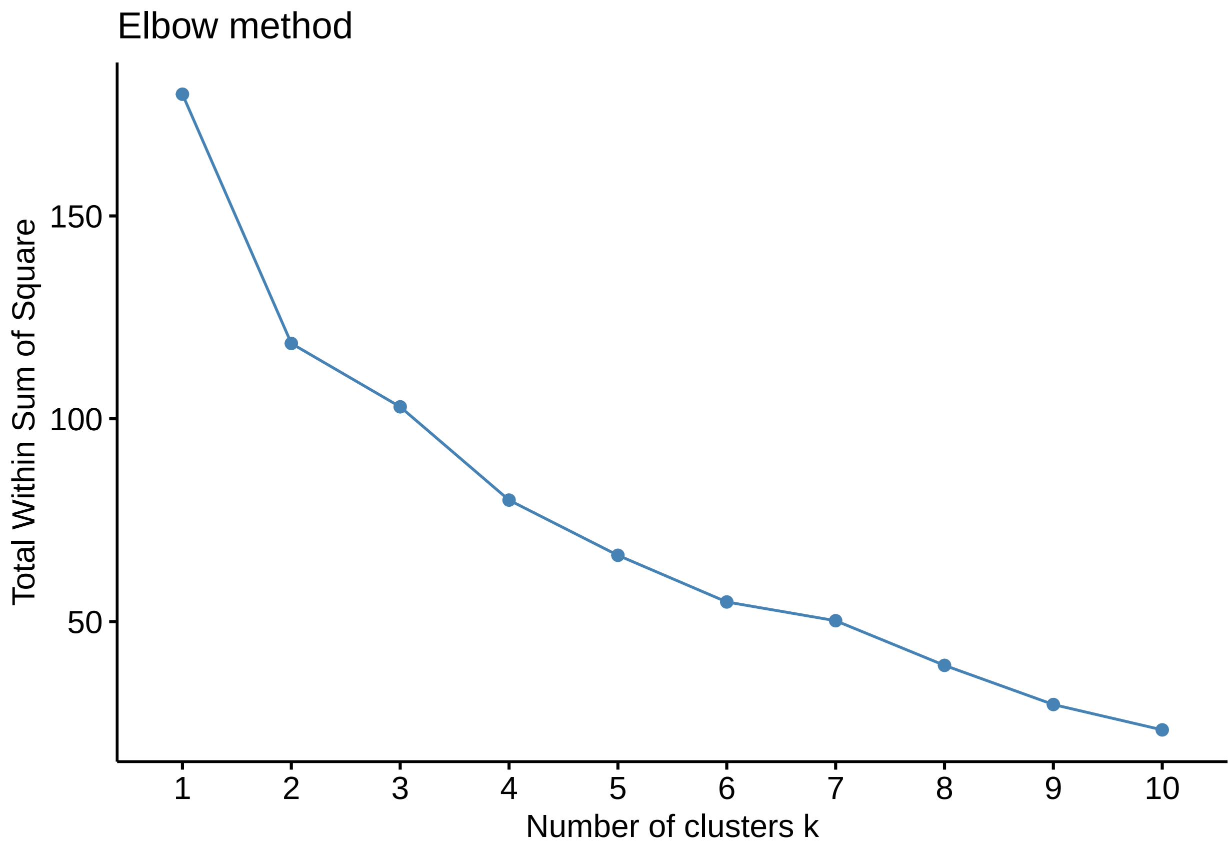
Questions

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.

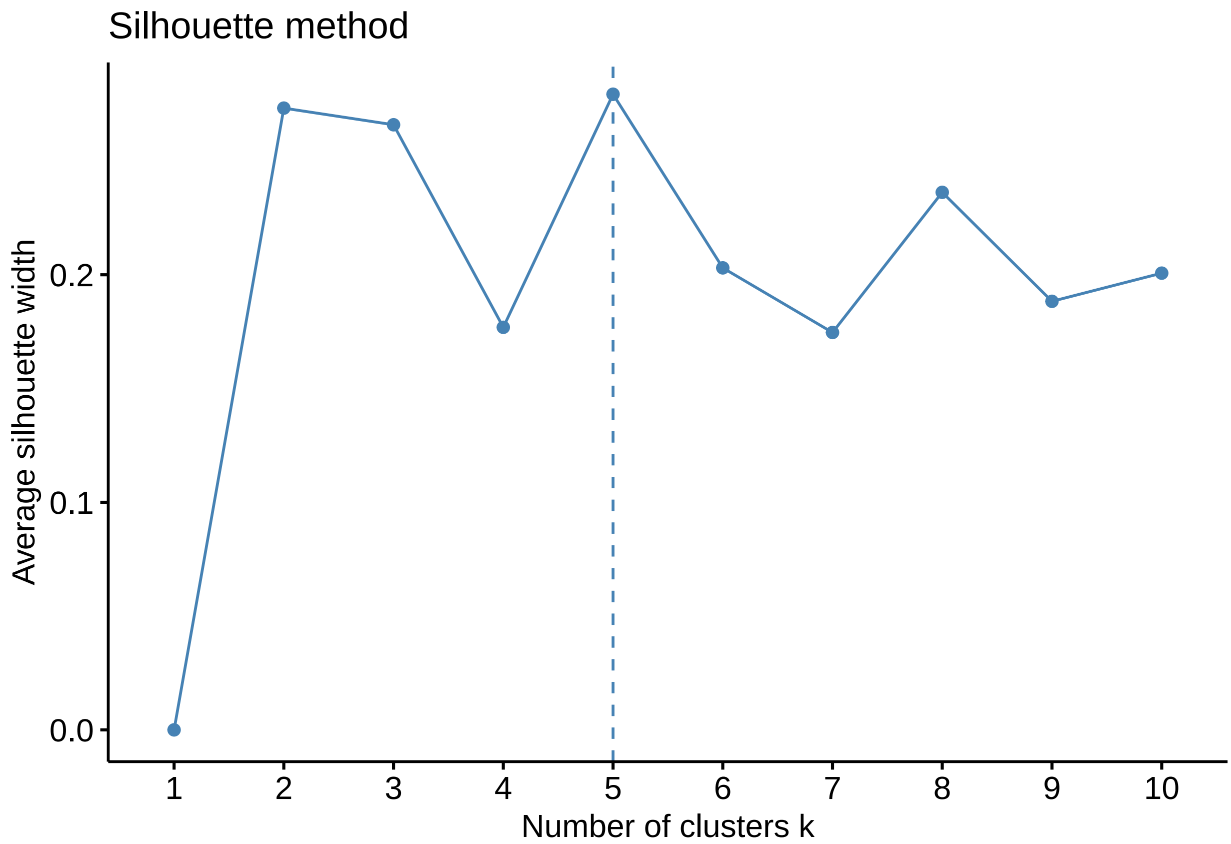
Kmeans Clustering

For getting the best value of K(no. of clusters) for kmeans

```r
# sum of squares method
fviz_nbclust(scaled.data, kmeans, method = "wss") + ggtitle("Elbow method")
```




```r
# silhouette method
fviz_nbclust(scaled.data, kmeans, method = "silhouette") + ggtitle("Silhouette method")
```



from the plot of WSS(Sum of squares) or elbow method, we can see that the curve was bent(as elbow) at point 2, so we have to consider the k value as 2. however it is still unclear due to less sharpness in the graphical representation.

```
# consider k=2
k <- 2
set.seed(159)
# kmeans algorithm
k_wss <- kmeans(scaled.data, centers = k, nstart=21)
k_wss
```

```
## 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"
```

```r
# To get the centroids of the clusters
cat("These are the centers of the clusters", "\n")
```

```
## These are the centers of the clusters
```

```r
k_wss$centers
```

```
##   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
```

```r
# Get the size of each cluster
cat("The Size of the each cluster is", "\n")
```

```
## The Size of the each cluster is
```

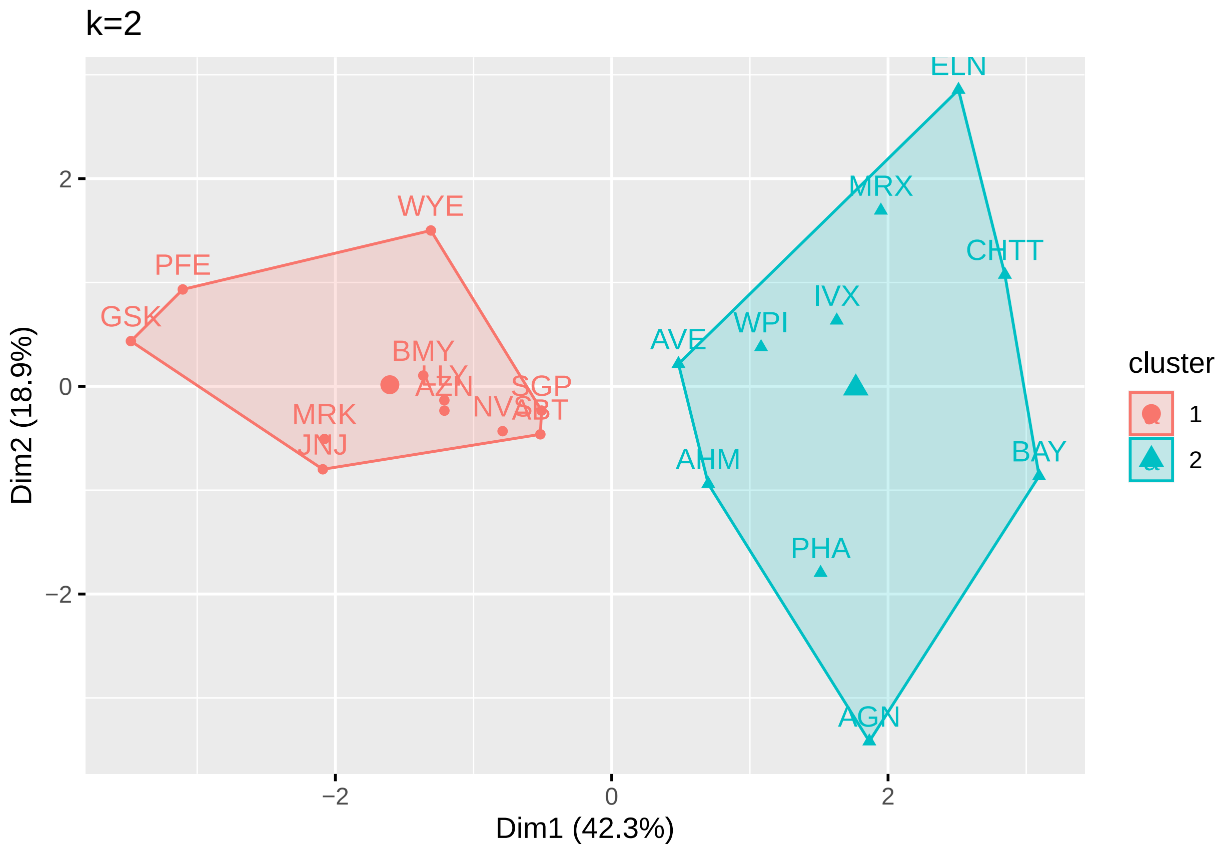
```r
k_wss$size
```

```
## [1] 11 10
```

```r
# To get which point belongs to which cluster
k_wss$cluster
```

```
##  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
```

```r
# Visualization of clusters
fviz_cluster(k_wss,data = scaled.data) + ggtitle("k=2")
```



from the output of this kmeans clustering with k value of 2. we can see that 11 companies comes under one cluster and the remaining 10 comes under another cluster, by taking all the numerical variables as these all are the financial measures are to be considered to know the equity, as equity depends on Market capital, net profit, return on assets, asset turnover, etc. And from the clusters we can see that some of the points like AGN, ELN, GSK, etc.. are far away from the centroids, which shows us that the number of clusters taken was not enough. ***** from the plot of silhouette method, we can see that the maximum average silhouette width is at point 5, so we have to consider the k value as 5.

```r
# consider k=5
k <- 5
set.seed(159)
# kmeans algorithm
k_sil <- kmeans(scaled.data, centers = k, nstart=20)
k_sil
```

```
## K-means clustering with 5 clusters of sizes 4, 2, 4, 3, 8
##
## 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  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 5 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
##      Leverage Rev_Growth Net_Profit_Margin
## 1  0.06308085  1.5180158      -0.006893899
## 2 -0.14170336 -0.1168459      -1.416514761
```

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

```r
# To get the centroids of the clusters
cat("These are the centers of the clusters", "\n")
```

```
## These are the centers of the clusters
```

```r
k_sil$centers
```

```
##    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  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 5 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
##      Leverage Rev_Growth Net_Profit_Margin
## 1  0.06308085  1.5180158      -0.006893899
## 2 -0.14170336 -0.1168459      -1.416514761
## 3 -0.46807818  0.4671788       0.591242521
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.27449312 -0.7041516       0.556954446
```

```r
# Get the size of each cluster
cat("The Size of the each cluster is",  "\n")
```

```
## The Size of the each cluster is
```

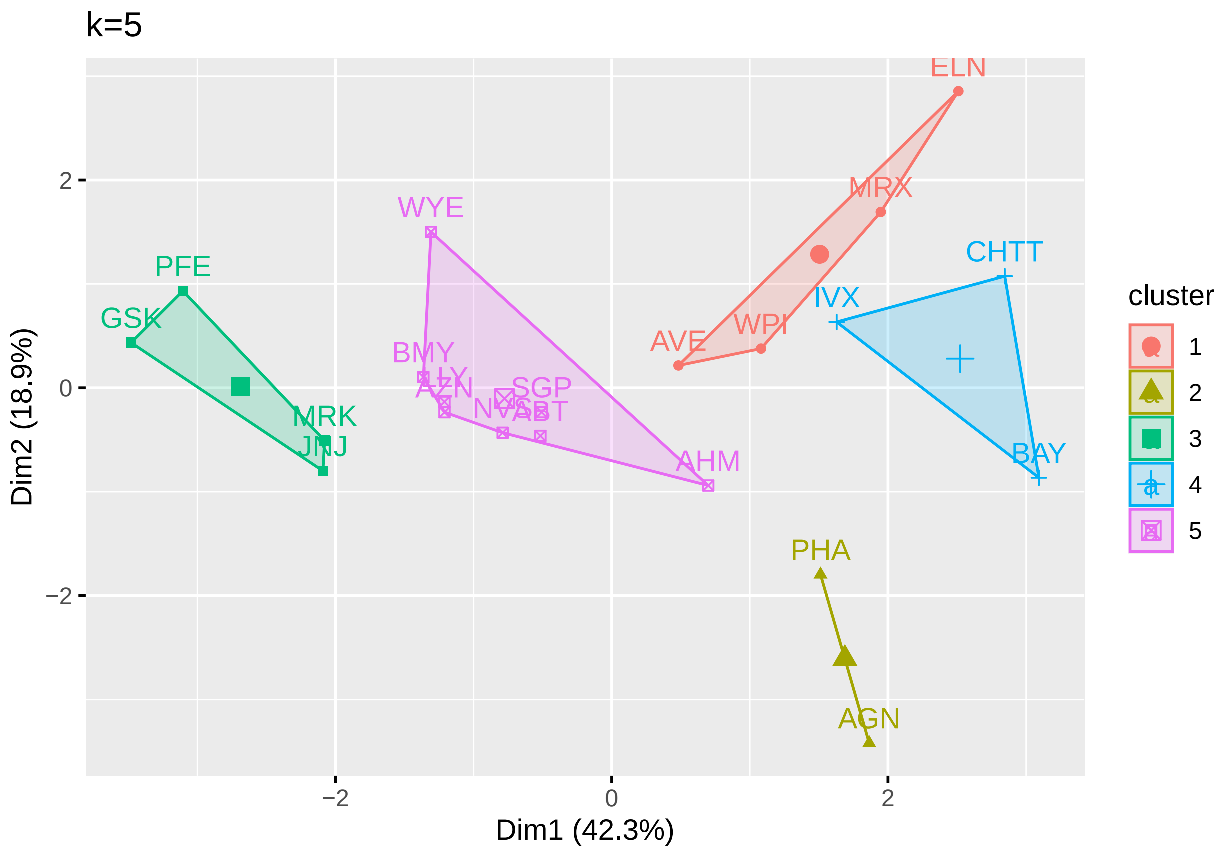
```r
k_sil$size
```

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

```r
# To get which point belongs to which cluster
k_sil$cluster
```

```
## ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##   5    2    5    5    1    4    5    4    1    5    3    4    3    1    3    5
## PFE  PHA  SGP  WPI  WYE
##   3    2    5    1    5
```

```r
# Visualization of clusters
fviz_cluster(k_sil, scaled.data) + ggtitle("k=5")
```

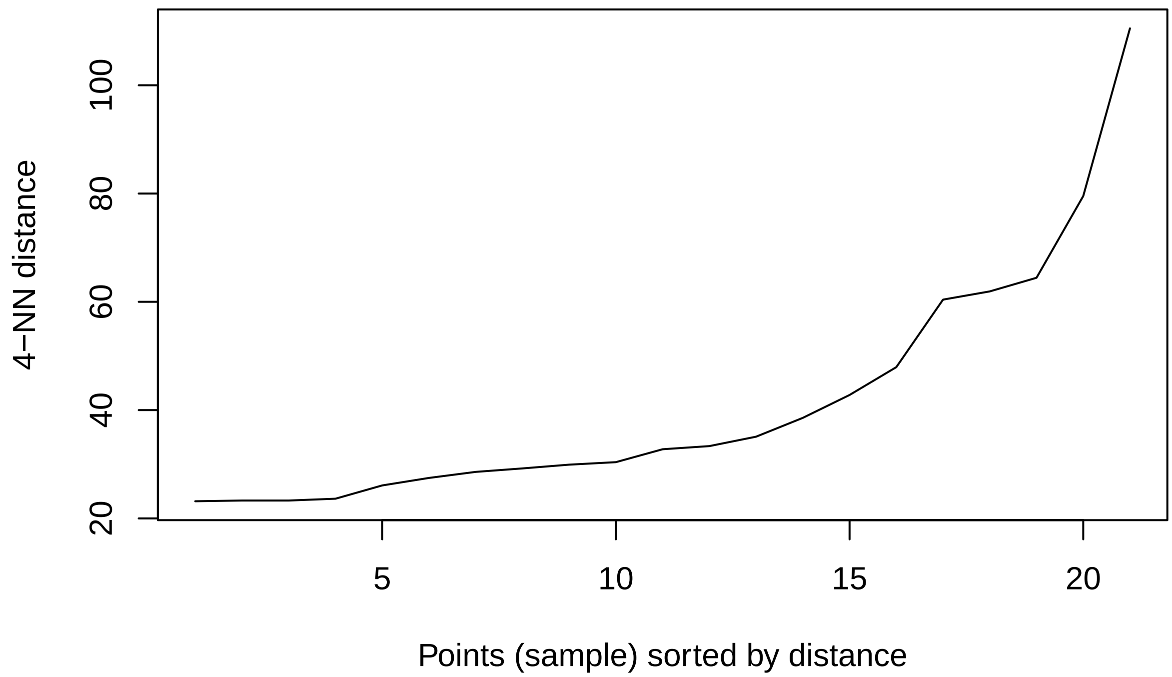


from the output of this kmeans clustering with k value of 5. we can see that 4 companies comes under first cluster, 2 companies under second cluster, 3 companies under third cluster, 8companies under fourth cluster and the remaining comes under fifth cluster, by taking all the numerical variables as these all are the financial measures are to be considered to know the equity, as equity depends on Market capital, net profit, return on assets, asset turnover, etc. And in this we can see the points are much nearer to the centroids. And this cluster might be the best. lets consider the remaining clusters

DBSCAN Clustering

To get the best value of radius or eps.

```r
# Graph to get the best value of radius at min points of 4.
dbscan::kNNdistplot(clust.data, k=4)
```



KNN-dist plot is used ro determine the optimal value of radius for DBSCAN clustering, we need to take the radius from where the curve was bent. From the above Plot, we can see that the curve was bent at distance between 20 and 40. so, consider the radius or EPS value as 30 at minimum points of 4.

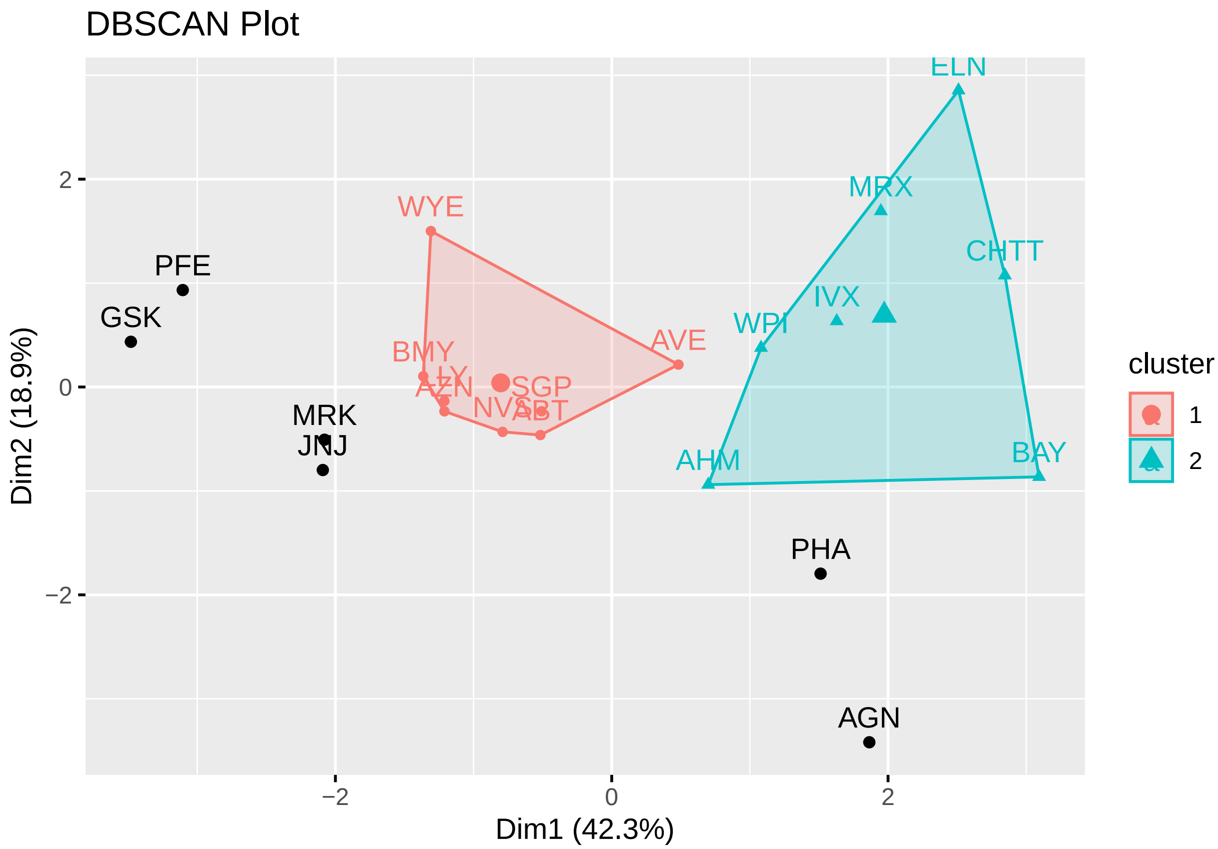
```
# DBSCAN Algorithm at eps=30 and minpts =4
dbs <- dbscan::dbscan(clust.data, eps = 30, minPts = 4)

# Output of the clusters
print(dbs)
```

```
## DBSCAN clustering for 21 objects.
## Parameters: eps = 30, minPts = 4
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 2 cluster(s) and 6 noise points.
##
## 0 1 2
## 6 8 7
##
## Available fields: cluster, eps, minPts, dist, borderPoints
```

```
# To get which point belongs to which cluster
print(dbs$cluster)
```

```
##  [1] 1 0 2 1 1 2 1 2 2 1 0 2 0 2 0 1 0 0 1 2 1
```

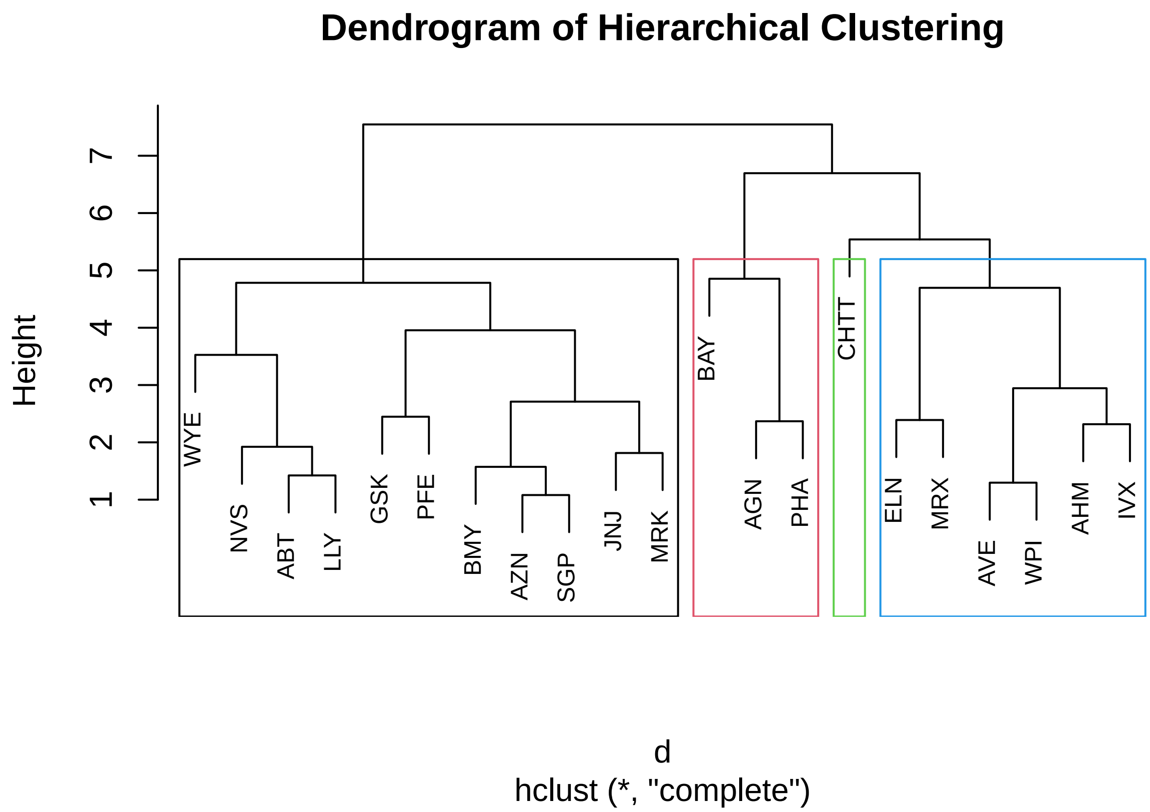
```r
# Visualization of clusters
fviz_cluster(dbs, clust.data) + ggtitle("DBSCAN Plot")
```



From the output and Plot of the DBSCAN clustering with the radius of 30 and minimum points of 4, we can see that 2 clusters are formed, one cluster with 8 points and the second cluster with 7 points and remaining six points as outliers. we can see the outliers from the plot. a good cluster should have minimum number of outliers, so we can say that this was not a good clustering process.

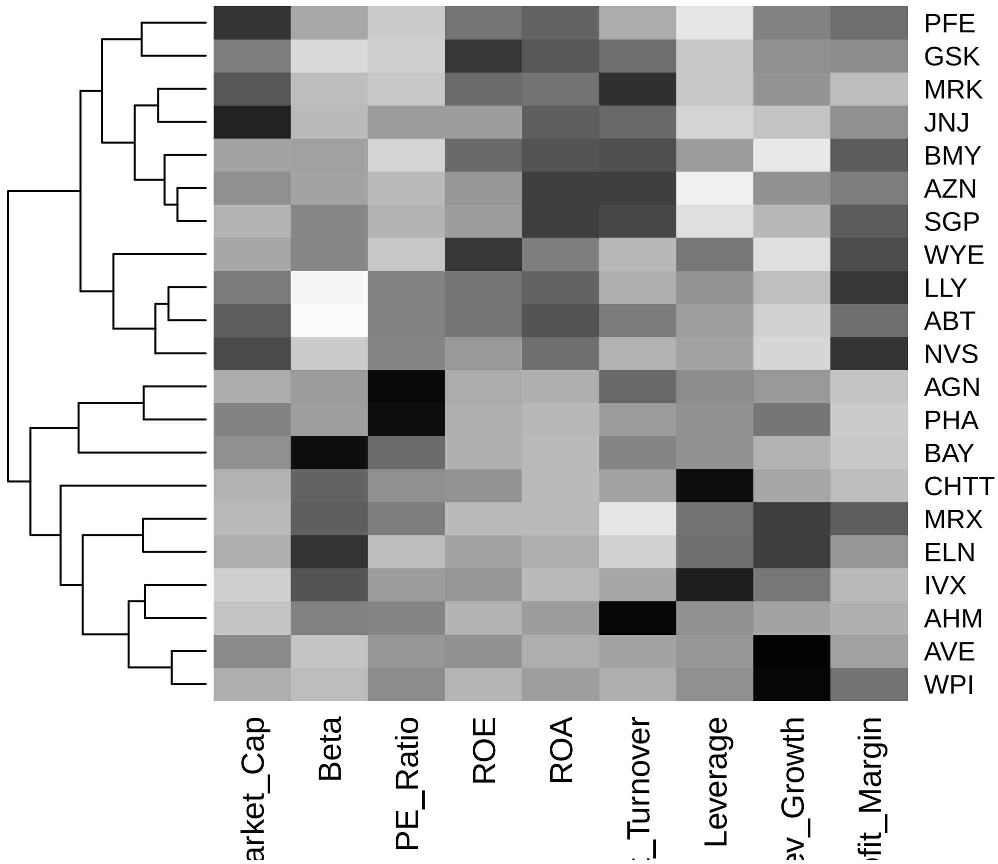
Hierarchical Clustering

```r
# Get the euclidean distance for the data
d <- dist(scaled.data, method = "euclidean")

# Hierarchical Clustering
hc <- hclust(d, method = "complete")

# Visualize the output Dendrogram at height=5
plot(hc, cex = 0.75, main = "Dendrogram of Hierarchical Clustering")
rect.hclust(hc, h=5, border = 1:4)
```



in hierarchical clustering, we have considered the height h=5. because at h=5 the clusters are formed correspond to the distance between the merged clusters compared to remaining heights. at this height 4 clusters are formed. from the dendrogram we can say that first cluster with size 11 second cluster with size 3 third cluster with size 1 fourth cluster with size 6 but here in this clustering, one cluster have many points and the other have too less, so this might not be a good one to do clustering of all the companies.

```
heatmap(as.matrix(scaled.data), Colv = NA, hclustfun = hclust,
        col=rev(paste("gray",1:99,sep="")))
```



Out of all these clusters I have found that Kmeans clustering with no.of clusters as 5 produce better clusters. **************

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)?

```
# creating a table with clusters
clustered.data1 <- pharma.data[,c(2:11)]  %>%
                  mutate(cluster=k_sil$cluster) %>% arrange(cluster, ascending = T)
# dataset with clusters
clustered.data1
```

```
##                                   Name Market_Cap Beta PE_Ratio  ROE  ROA
## AVE                            Aventis      47.16 0.32     20.1 21.8  7.5
## ELN                Elan Corporation, plc     0.78 1.08      3.6 15.1  5.1
## MRX  Medicis Pharmaceutical Corporation     1.20 0.75     28.6 11.2  5.4
## WPI          Watson Pharmaceuticals, Inc.    3.26 0.24     18.4 10.2  6.8
## AGN                      Allergan, Inc.     7.58 0.41     82.5 12.9  5.5
## PHA                Pharmacia Corporation    56.24 0.40     56.5 13.5  5.7
## GSK                  GlaxoSmithKline plc   122.11 0.35     18.0 62.9 20.3
## JNJ                    Johnson & Johnson   173.93 0.46     28.4 28.6 16.3
## MRK                     Merck & Co., Inc.  132.56 0.46     18.9 40.6 15.0
## PFE                          Pfizer Inc   199.47 0.65     23.6 45.6 19.2
## BAY                            Bayer AG    16.90 1.11     27.9  3.9  1.4
## CHTT                       Chattem, Inc     0.41 0.85     26.0 24.1  4.3
## IVX                    IVAX Corporation     2.60 0.65     19.9 21.4  6.8
```

```
## ABT                Abbott Laboratories      68.44 0.32     24.7 26.4 11.8
## AHM                       Amersham plc       6.30 0.46     20.7 14.9  7.8
## AZN                    AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4
## BMY         Bristol-Myers Squibb Company      51.33 0.50     13.9 34.8 15.1
## LLY              Eli Lilly and Company      73.84 0.18     27.9 31.0 13.5
## NVS                        Novartis AG      96.65 0.19     21.6 17.9 11.2
## SGP         Schering-Plough Corporation      34.10 0.51     18.9 22.6 13.3
## WYE                              Wyeth      48.19 0.63     13.1 54.9 13.4
##      Asset_Turnover Leverage Rev_Growth Net_Profit_Margin cluster
## AVE             0.6     0.34      26.81              12.9       1
## ELN             0.3     1.07      34.21              13.3       1
## MRX             0.3     0.93      30.37              21.3       1
## WPI             0.5     0.20      29.18              15.1       1
## AGN             0.9     0.60       9.16               5.5       2
## PHA             0.6     0.35      15.00               7.3       2
## GSK             1.0     0.34      21.87              21.1       3
## JNJ             0.9     0.10       9.37              17.9       3
## MRK             1.1     0.28      17.35              14.1       3
## PFE             0.8     0.16      25.54              25.2       3
## BAY             0.6     0.00      -3.17               2.6       4
## CHTT            0.6     3.51       6.38               7.5       4
## IVX             0.6     1.45      13.99              11.0       4
## ABT             0.7     0.42       7.54              16.1       5
## AHM             0.9     0.27       7.05              11.2       5
## AZN             0.9     0.00      15.00              18.0       5
## BMY             0.9     0.57       2.70              20.6       5
## LLY             0.6     0.53       6.21              23.4       5
## NVS             0.5     0.06      -2.69              22.4       5
## SGP             0.8     0.00       8.56              17.6       5
## WYE             0.6     1.12       0.36              25.5       5
```

```r
cat("below are the list of firms with their corresponding clusters")
```

```
## below are the list of firms with their corresponding clusters
```

```r
clustered.data1[,c(1,11)]
```

```
##                                     Name cluster
## AVE                              Aventis       1
## ELN                 Elan Corporation, plc       1
## MRX  Medicis Pharmaceutical Corporation       1
## WPI         Watson Pharmaceuticals, Inc.       1
## AGN                        Allergan, Inc.       2
## PHA                 Pharmacia Corporation       2
## GSK                   GlaxoSmithKline plc       3
## JNJ                     Johnson & Johnson       3
## MRK                     Merck & Co., Inc.       3
## PFE                            Pfizer Inc       3
## BAY                              Bayer AG       4
## CHTT                         Chattem, Inc       4
## IVX                       IVAX Corporation       4
## ABT                   Abbott Laboratories       5
## AHM                          Amersham plc       5
```

```
## AZN                     AstraZeneca PLC       5
## BMY        Bristol-Myers Squibb Company       5
## LLY               Eli Lilly and Company       5
## NVS                          Novartis AG       5
## SGP         Schering-Plough Corporation       5
## WYE                                Wyeth       5
```

calculate the mean of all numerical variables in each cluster

```r
# calculate the mean of all numerical variables
aggregate(scaled.data, by=list(k_sil$cluster), FUN=mean)
```

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

Adding the cluster to normalised data.

```r
# add the clusters to the scaled data
scaled.data1 <- data.frame(scaled.data, k_sil$cluster)
scaled.data1
```

```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## AGN  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## AHM  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## AZN   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## AVE  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## BAY  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
## BMY  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
## CHTT -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
## ELN  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## LLY   0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## GSK   1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## IVX  -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## JNJ   1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## MRX  -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## MRK   1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## NVS   0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## PFE   2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## PHA  -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## SGP  -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## WPI  -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## WYE  -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
```

```
##         Leverage  Rev_Growth Net_Profit_Margin k_sil.cluster
## ABT  -0.21209793 -0.52776752        0.06168225             5
## AGN   0.01828430 -0.38113909       -1.55366706             2
## AHM  -0.40408312 -0.57211809       -0.68503583             5
## AZN  -0.74965647  0.14744734        0.35122600             5
## AVE  -0.31449003  1.21638667       -0.42597037             1
## BAY  -0.74965647 -1.49714434       -1.99560225             4
## BMY  -0.02011273 -0.96584257        0.74744375             5
## CHTT  3.74279705 -0.63276071       -1.24888417             4
## ELN   0.61983791  1.88617085       -0.36501379             1
## LLY  -0.07130879 -0.64814764        1.17413980             5
## GSK  -0.31449003  0.76926048        0.82363947             3
## IVX   1.10620040  0.05603085       -0.71551412             4
## JNJ  -0.62166634 -0.36213170        0.33598685             3
## MRX   0.44065173  1.53860717        0.85411776             1
## MRK  -0.39128411  0.36014907       -0.24310064             3
## NVS  -0.67286239 -1.45369888        1.02174835             5
## PFE  -0.54487226  1.10143723        1.44844440             3
## PHA  -0.30169102  0.14744734       -1.27936246             2
## SGP  -0.74965647 -0.43544591        0.29026942             5
## WPI  -0.49367621  1.43089863       -0.09070919             1
## WYE   0.68383297 -1.17763919        1.49416183             5
```

by comparing the mean values of all the numerical variables from the clusters Cluster1 with companies AVE, WPI, MRX, ELN has high revenue growth and beta value. but have low asset turnover, return on equity and return on asset.And the market capitalization is also relatively low. based on these, it is possible that these companies are still growing and they are at early stage. These companies might be investing heavily in marketing and sales. However, the high revenue growth and beta value suggest that they are expected to improve their earnings more rapidly in the coming days. these companies are distinguished by their higher growth potential and low profitability.

Cluster2 with companies PHA, AGN has high Price or earnings ratio and asset turnover, but have low net profit margin, return on equity and return on asset. and the market capitalization is also relatively low. However, the high asset turnover and price or earnings ratios suggest that they are expected to improve their earnings more rapidly in the future, while having little net profit in the past. However, with its high price, investors get more risk.

Cluster3 with companies IVX, CHTT, BAY has high market capitalization, return on equity, Return on assets and Asset turnover. but they have lowest Beta and profit to return Ratio. Based on these features these companies are matured and well established companies. the low beta value suggests that their stock prices are more stable, so that it was less risky to invest. but the low profit return ratio shows that they are not so efficient in generating profits. these companies are distinguished by their maturity, stability, and profitability.

Cluster4 with companies WYE, BMY, LLY, AZN, NVS, ABT, SGP, AHM has high beta value and leverage. but have lowest net profit margin, market capitalization. And relatively low return on equity, return on asset, revenue growth. based on these features, we can say that these companies are riskier to invest than other companies as they have high beta value which means their stock price was unstable and high leverage means more debts. and there profit margin is also low. but, if the market was high they can earn more profits due to that high beta value. these companies are distinguished by higher risk and potential for higher returns.

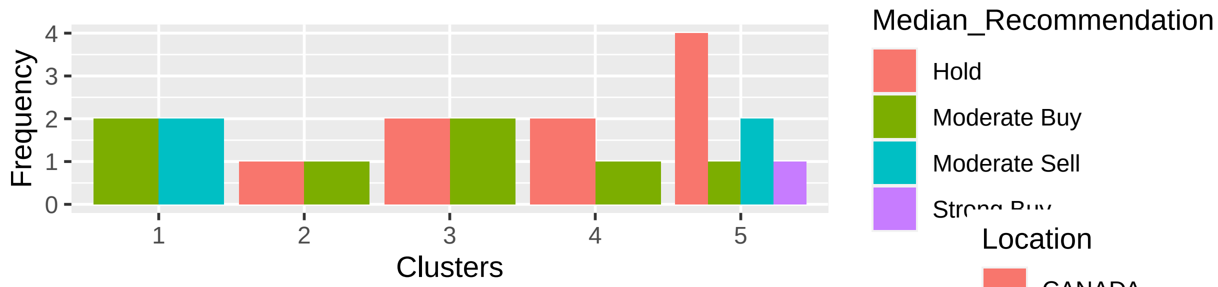
Cluster5 with companies GSK, PFE, MRK, JNJ has highest net profit margin, asset turnover, return on equity, Return on assets. but have lowest Beta, profit to return Ratio, revenue growth. these features shows that these companies have high financial performance and low risk. the high net profit margins, asset turnovers, returns on equity, and returns on assets, indicates efficient operations and strong profitability. and lowest beta value and revenue growth shows the stock price was more stable and less revenue growth.these represents a group of mature and well-established companies with strong financial performance and low risk profiles.

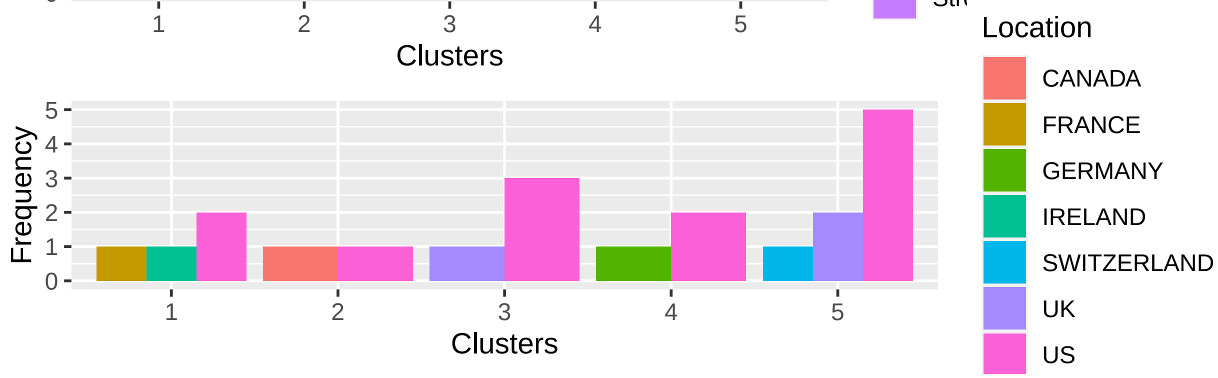
Is there a pattern in the clusters with respect to the numerical variables (10 to12)

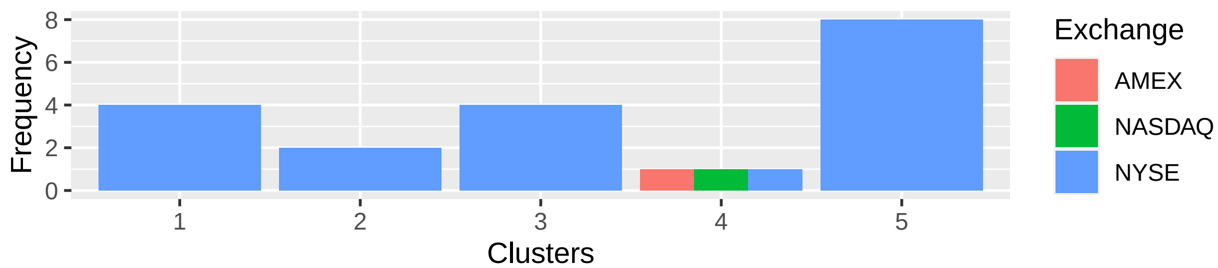
```
# Add the clusters to the data
data_pattern <- pharma.data[12:14] %>% mutate(Clusters = k_sil$cluster)
data_pattern
```

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

```
# Plot the data with Median_Recommendation
recommendation <- ggplot(data_pattern, mapping = aes(factor(Clusters), fill =Median_Recommendation)) + g

# Plot the data with location
location <- ggplot(data_pattern, mapping = aes(factor(Clusters), fill = Location)) + geom_bar(position =

# Plot the data with Exchange
exchange <- ggplot(data_pattern, mapping = aes(factor(Clusters), fill = Exchange)) + geom_bar(position =

grid.arrange(recommendation, location, exchange)
```







Cluster1, Recommended as Moderate Buy and Moderate Sell from Locations France, Ireland and US and listed under NYSE.

Cluster2, Recommended as Hold and Moderate Buy from Locations US and canada and listed under NYSE.

Cluster3, Recommended as Hold and Moderate Buy from Locations UK and US and listed under NYSE.

Cluster4, Recommended as Hold and Moderate Buy from Locations Germany and US and listed under AMEX, NASDAQ and NYSE.

Cluster5, Recommended Hold, Moderate Sell, Strong Buy & Moderate Buy from Locations Switzerland, UK and US and listed under NYSE. *************** 3.Provide an appropriate name for each cluster using any or all of the variables in the dataset. A. Appropriate names for the clusters: Cluster1: High Growth potential Cluster2: High Risk High Reward Cluster3: Stability and Profitability Cluster4: High Risk High Beta Cluster5: Low Risk High Profitability