FML_Assignment_4

Rupesh_Suragani

2023-11-12

#Directions of the problem

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

For each firm, the following variables are recorded:

1. Market capitalization (in billions of dollars)
2. Beta
3. Price/earnings ratio
4. Return on equity
5. Return on assets
6. Asset turnover
7. Leverage
8. Estimated revenue growth
9. Net profit margin
10. Median recommendation (across major brokerages)
11. Location of firm's headquarters
12. Stock exchange on which the firm is listed

Use cluster analysis to explore and analyze the given dataset as follows:

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.

---

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)

---

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

#Running all the libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(factoextra)
```

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

```
library(ISLR)
library(cluster)
```

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

```
library(dbscan)
```

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

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

```
library(fpc)
```

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

```
##
## Attaching package: 'fpc'
##
## The following object is masked from 'package:dbscan':
##
##     dbscan
```

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

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

```
#Loading the csv data
pharmacy <- read.csv("C:\\Users\\rupes\\OneDrive\\Desktop\\Kent State University\\FML\\Assignmen
t 4\\Pharmaceuticals.csv")
head(pharmacy)
```

```
##   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
```

```
pharmacy <- na.omit(pharmacy)

#Dimensions of the csv data
dim(pharmacy)
```

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

```
#Printing all the variable names
t(t(names(pharmacy)))
```

```
##       [,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"
```

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). From the given csv data (Pharmaceuticals.csv) the numerical variables are Market capitalization, Beta, Price/earnings ratio, Return on equity, Return on assets, Asset turnover, Leverage, Estimated revenue growth, and Net profit margin

Using Only the numerical variables (3 to 11 variables in the data) to cluster the 21 firms.

```
set.seed(123)

#Selecting the numerical variables 1 to 9

#In the Pharmacy data the numerical variables exist from column 3 to column 11.
row.names(pharmacy) <- pharmacy[,1]
num_pharma_data <- pharmacy[ , c(3 : 11)]
head(num_pharma_data)
```

```
##     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
```

```
#Dimensions
dim(num_pharma_data)
```

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

```
#Numeric variables
t(t(names(num_pharma_data)))
```

```
##       [,1]
##  [1,] "Market_Cap"
##  [2,] "Beta"
##  [3,] "PE_Ratio"
##  [4,] "ROE"
##  [5,] "ROA"
##  [6,] "Asset_Turnover"
##  [7,] "Leverage"
##  [8,] "Rev_Growth"
##  [9,] "Net_Profit_Margin"
```

#Normalizing the Numerical Pharma data of 21 firms using scale function

```
#Normalizing the numerical data (z-score)
Scale_pharma <- scale(num_pharma_data)
Scale_pharma
```

```
##      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
## ABT  -0.21209793 -0.52776752        0.06168225
## AGN   0.01828430 -0.38113909       -1.55366706
## AHM  -0.40408312 -0.57211809       -0.68503583
## AZN  -0.74965647  0.14744734        0.35122600
## AVE  -0.31449003  1.21638667       -0.42597037
## BAY  -0.74965647 -1.49714434       -1.99560225
## BMY  -0.02011273 -0.96584257        0.74744375
## CHTT  3.74279705 -0.63276071       -1.24888417
## ELN   0.61983791  1.88617085       -0.36501379
## LLY  -0.07130879 -0.64814764        1.17413980
## GSK  -0.31449003  0.76926048        0.82363947
## IVX   1.10620040  0.05603085       -0.71551412
## JNJ  -0.62166634 -0.36213170        0.33598685
## MRX   0.44065173  1.53860717        0.85411776
## MRK  -0.39128411  0.36014907       -0.24310064
## NVS  -0.67286239 -1.45369888        1.02174835
## PFE  -0.54487226  1.10143723        1.44844440
## PHA  -0.30169102  0.14744734       -1.27936246
## SGP  -0.74965647 -0.43544591        0.29026942
## WPI  -0.49367621  1.43089863       -0.09070919
## WYE   0.68383297 -1.17763919        1.49416183
## attr(,"scaled:center")
##       Market_Cap             Beta         PE_Ratio              ROE
##       57.6514286        0.5257143       25.4619048       25.7952381
##              ROA    Asset_Turnover         Leverage       Rev_Growth
##       10.5142857        0.7000000        0.5857143       13.3709524
## Net_Profit_Margin
##       15.6952381
## attr(,"scaled:scale")
```

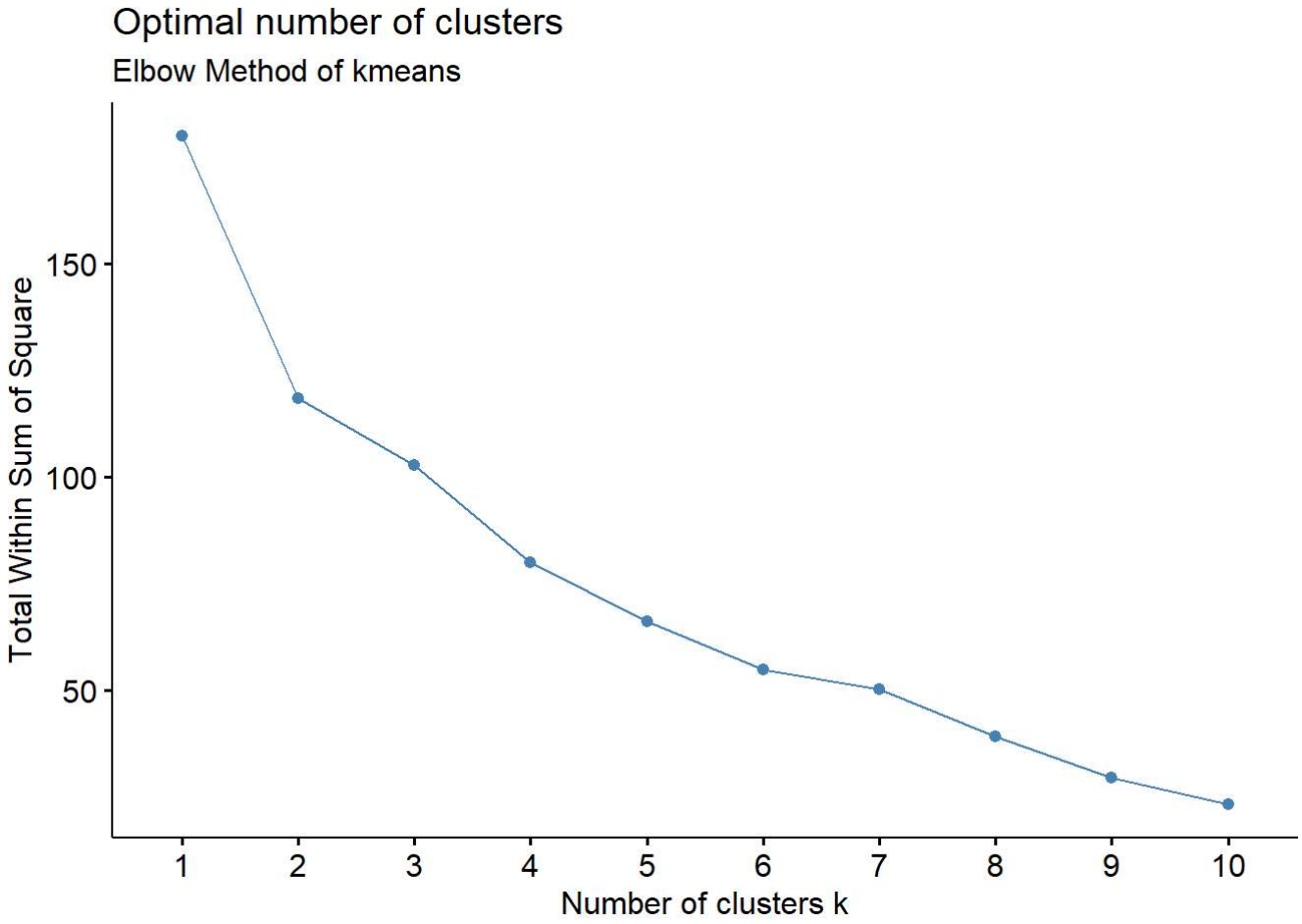
```
##        Market_Cap              Beta          PE_Ratio               ROE
##        58.6029595         0.2567406        16.3102568        15.0849752
##               ROA    Asset_Turnover          Leverage        Rev_Growth
##         5.3213988         0.2167948         0.7813103        11.0483351
## Net_Profit_Margin
##         6.5620482
```

#choosing different clustering algorithms

1. K_means Clustering:

One of the key decisions in cluster analysis is to determine the number of clusters and the optimal number of clusters will vary depending on the dataset. For getting out the optimal number of clusters to be formed, Elbow method can be implemented.

```
#Graphical representation of optimal number of clusters using Elbow method
fviz_nbclust(Scale_pharma, kmeans, method = "wss") + labs(subtitle = "Elbow Method of kmeans")
```



From the above graph, the optimal number of clusters can be taken as 2 because in the graph of elbow method the elbow bent is seen at point 2. Hence K shall be taken as 2.

#Taking k =2 for calculating kmeans.

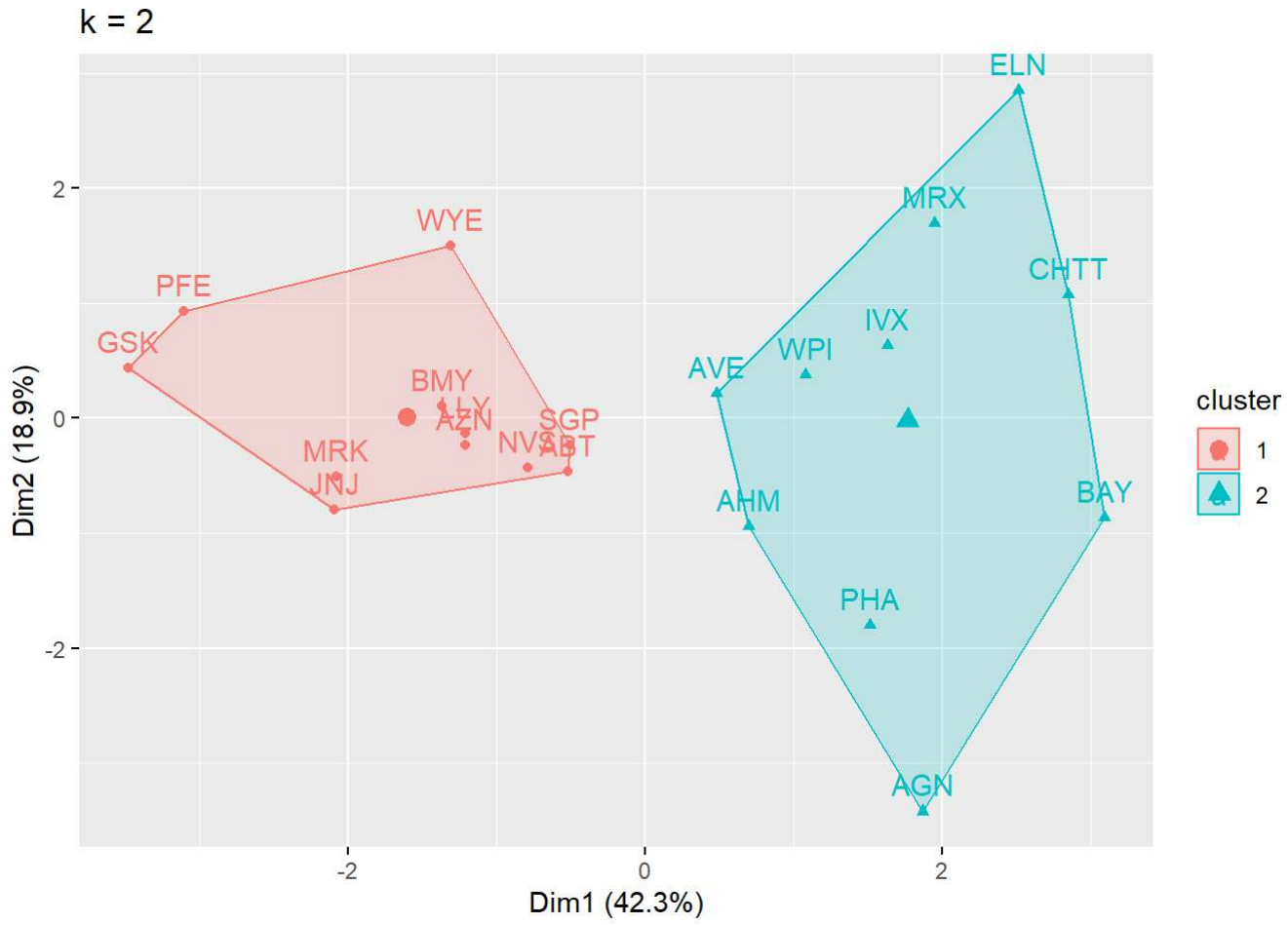
```
set.seed(159)
k = 2
elbow_cluster <- kmeans(Scale_pharma, centers = k, nstart = 21)
elbow_cluster
```

```
## 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"
```

```
#Finding the Centroids
elbow_cluster$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
```

```
#Visualization of cluster
fviz_cluster(elbow_cluster, Scale_pharma) + ggtitle("k = 2")
```




```
#Finding which firm belongs to which cluster
elbow_cluster$cluster
```

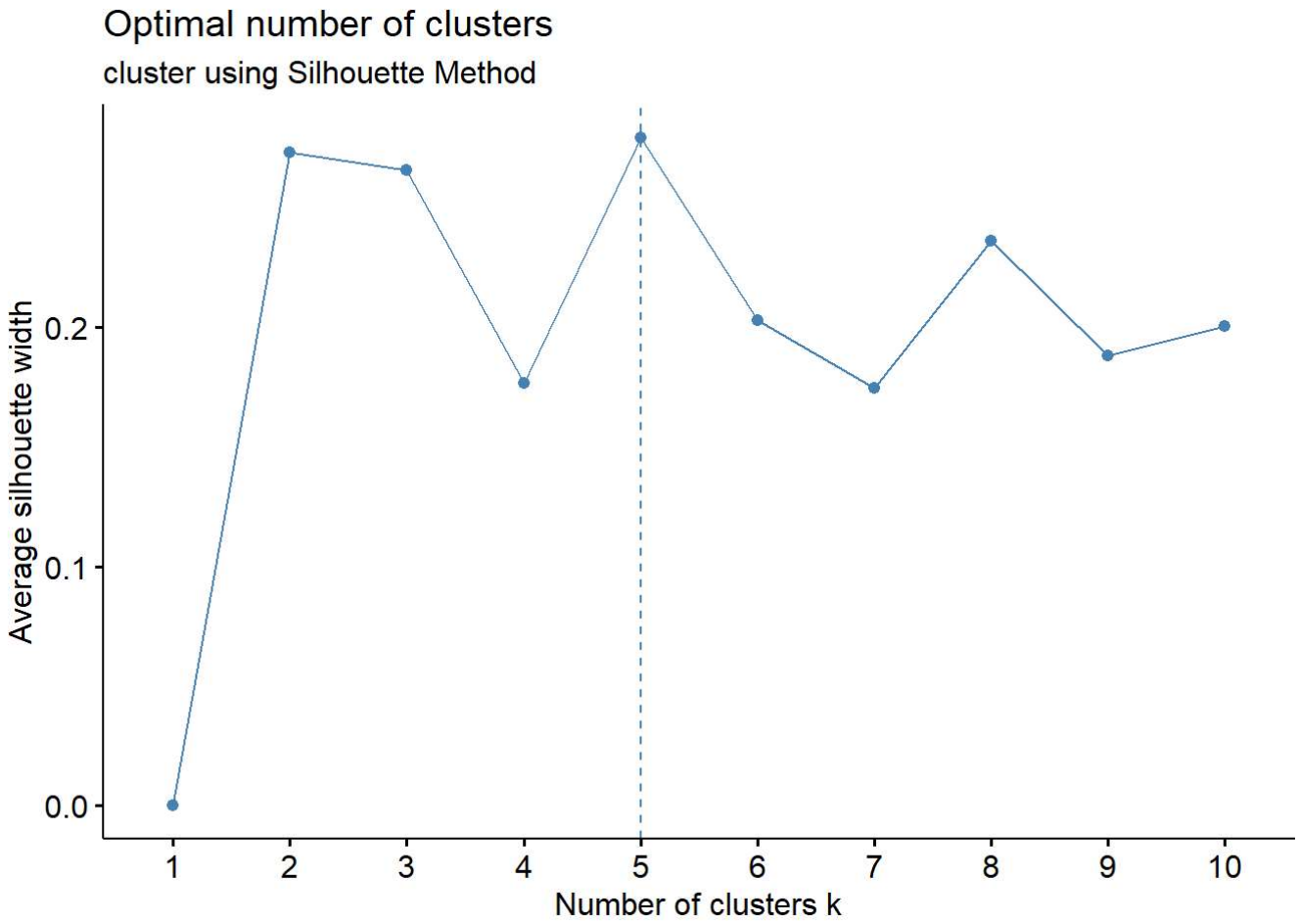
```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    2    2    1    2    2    1    2    2    1    1    2    1    2    1    1
##  PFE  PHA  SGP  WPI  WYE
##    1    2    1    2    1
```

```
#cluster size
elbow_cluster$size
```

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

Similar to Elbow method we can use Average silhouette method to find the best k(optimal number of clusters)

```
fviz_nbclust(Scale_pharma, kmeans, method = "silhouette") + labs(subtitle = "cluster using Silho
uette Method ")
```



On analyzing silhouette method the optimal number of clusters can be taken as 5.

```
set.seed(159)
k = 5
sil_cluster <- kmeans(Scale_pharma, centers = k, nstart = 21)
sil_cluster
```

```
## 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"
```

```
#Finding the Centroids
sil_cluster$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
```

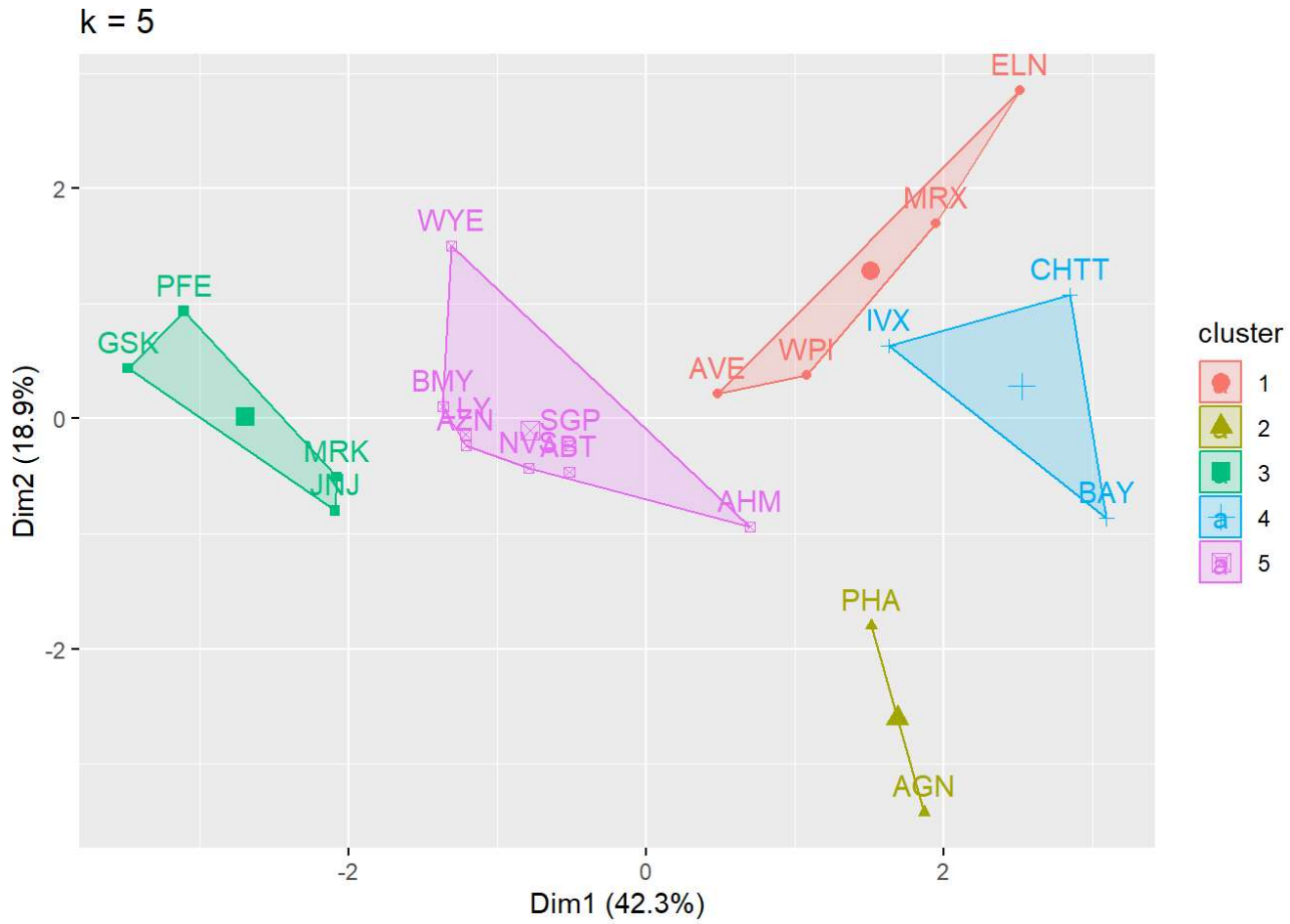
```
#Finding which firm belongs to which cluster
sil_cluster$cluster
```

```
##  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
```

```
#cluster size
sil_cluster$size
```

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

```
#Visualization of cluster
fviz_cluster(sil_cluster, Scale_pharma) + ggtitle("k = 5")
```



From the output of this kmeans clustering with k value of 5. we can see that 4 Firms comes under first cluster, 2 firms under second cluster, 3 companies under third cluster, 8 firms 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

#Kmeans cluster analysis for fitting the data with 5 clusters

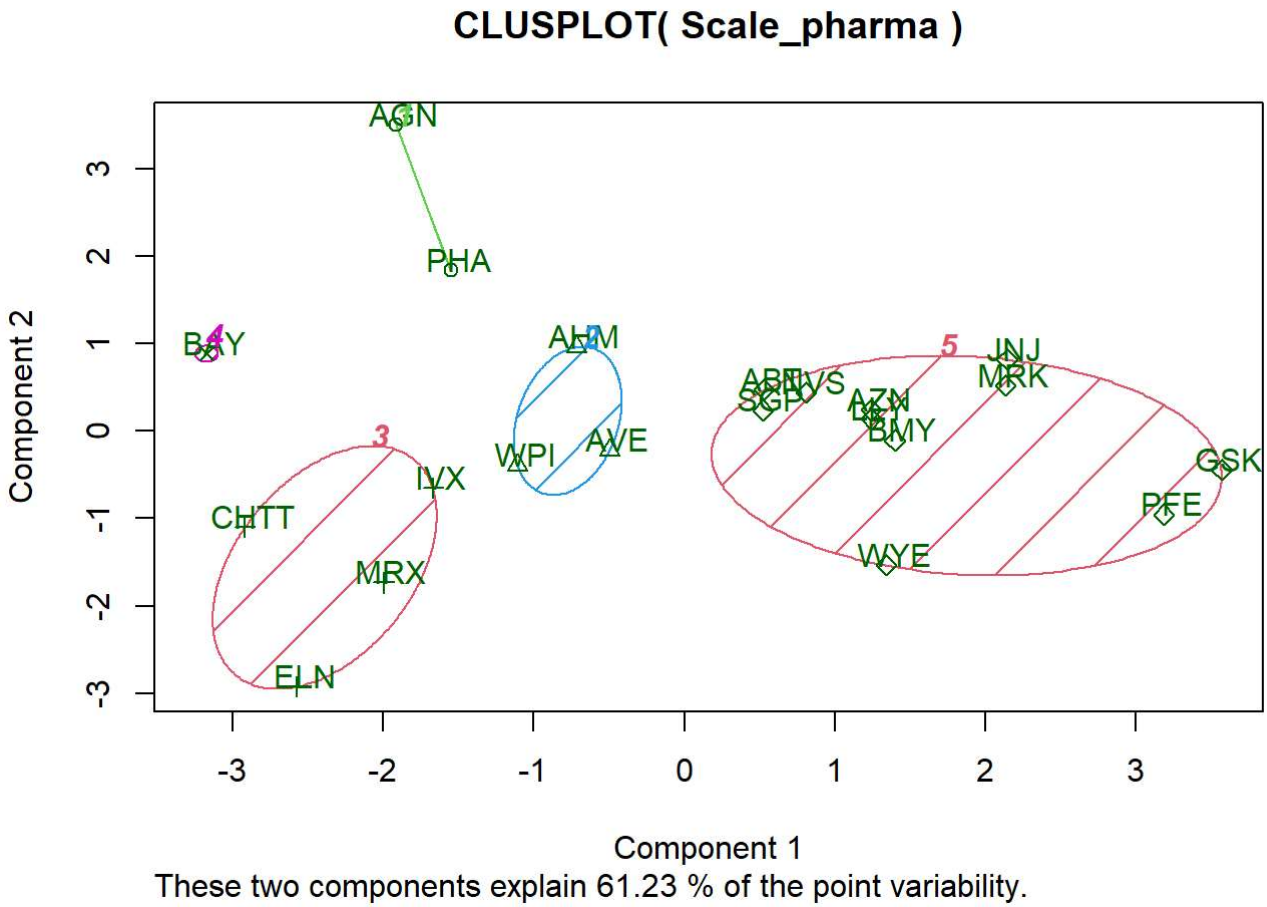
```
fit_data <- kmeans(Scale_pharma, 5)
```

#Calculating mean of all quantitative variables in each cluster

```
aggregate(Scale_pharma, by = list(fit_data$cluster), FUN = mean)
```

```
##   Group.1 Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1       1 -0.4392513 -0.4701800  2.7000246 -0.8349525 -0.9234951      0.2306328
## 2       2 -0.6611400 -0.7233539 -0.3512251 -0.6736441 -0.5915022     -0.1537552
## 3       3 -0.9624758  1.1949250 -0.3639982 -0.5200697 -0.9610792     -1.1531640
## 4       4 -0.6953818  2.2757827  0.1494823 -1.4514600 -1.7127612     -0.4612656
## 5       5  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.1417034 -0.1168459        -1.4165148
## 2 -0.4040831  0.6917224        -0.4005718
## 3  1.4773718  0.7120120        -0.3688236
## 4 -0.7496565 -1.4971443        -1.9956023
## 5 -0.3331068 -0.2902163         0.6823310
```

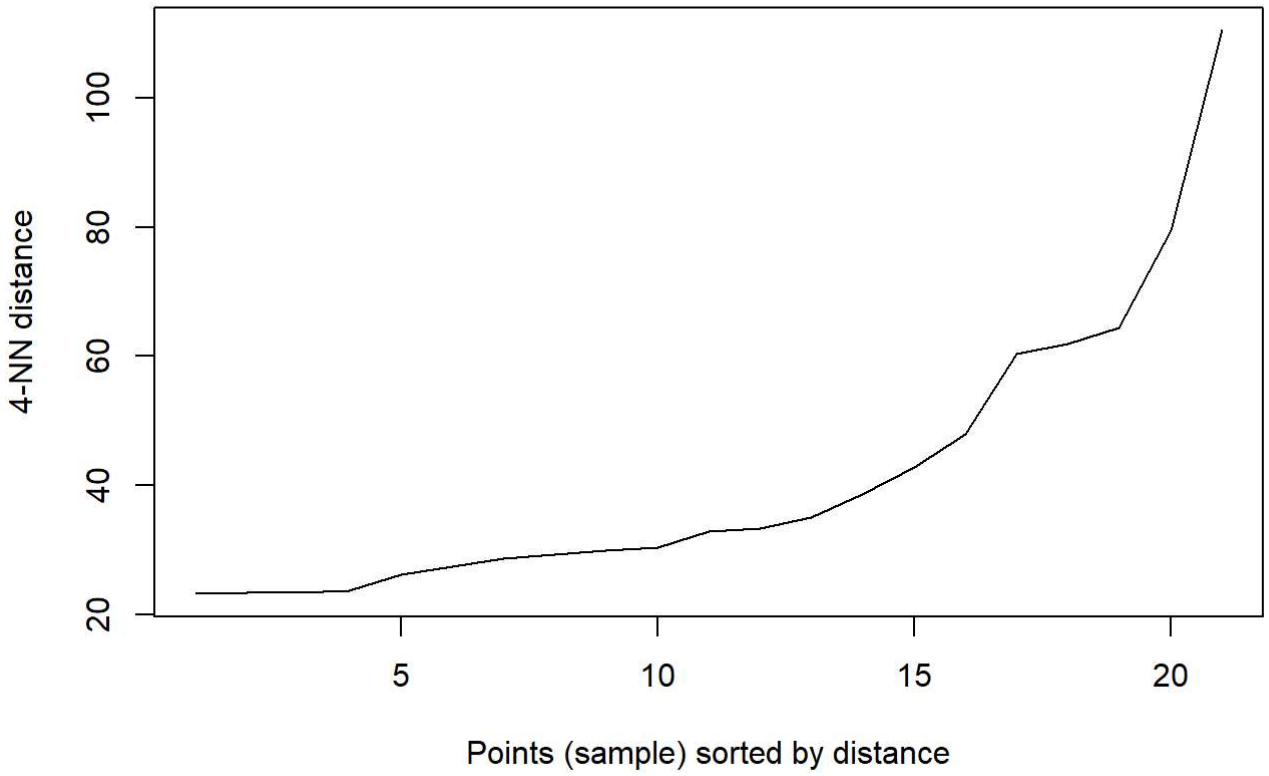
```
clusplot(Scale_pharma, fit_data$cluster, color = TRUE, shade = TRUE, labels = 2, lines = 0)
```



2.DBSCAN clustering

#Determining the Optimal ‘eps’ value

```
dbscan::kNNdistplot(num_pharma_data, k = 4)
```

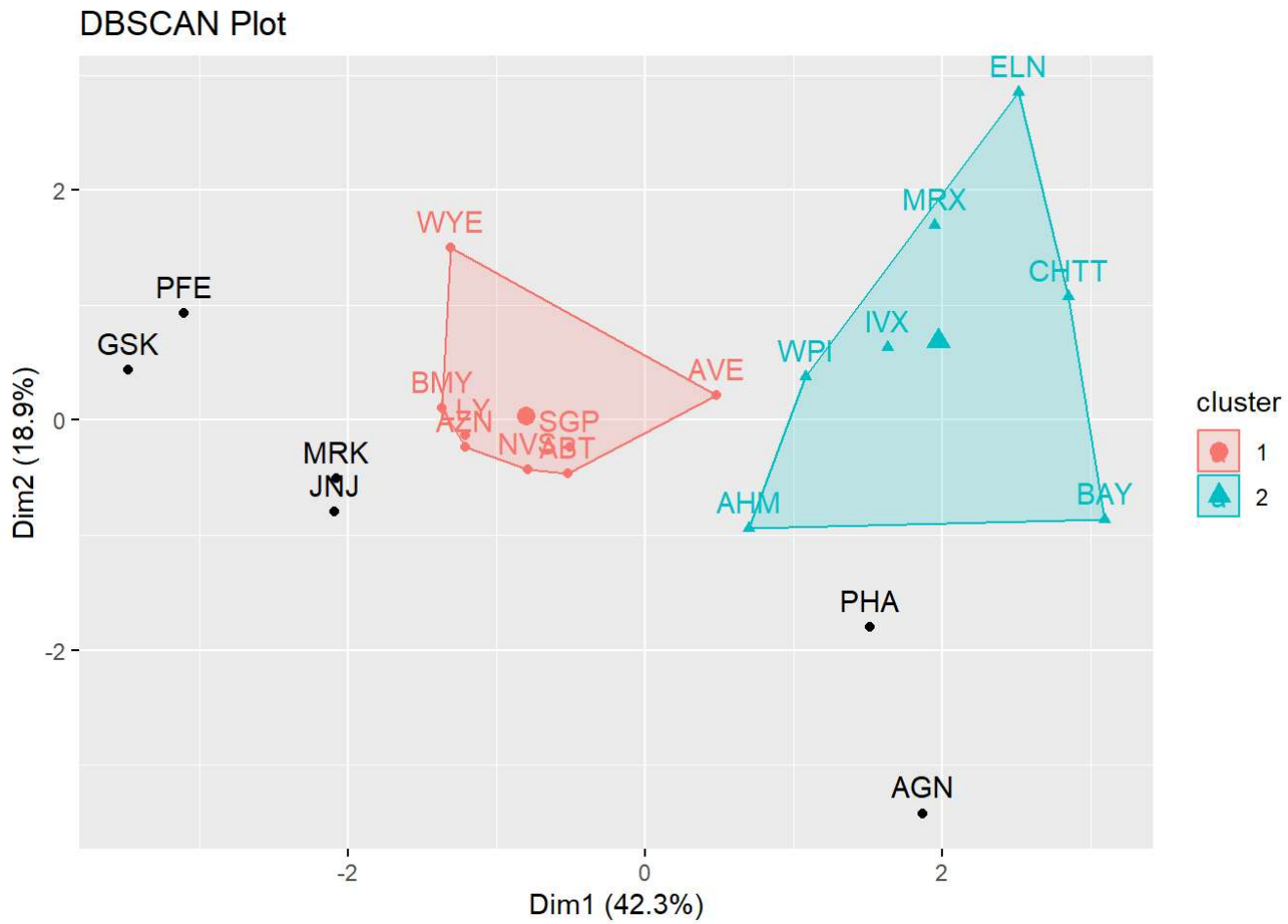


KNN-dist plot is used to determine the optimal value of radius for DBSCAN clustering, we need to take the radius from where the curve was bent. From the above Plot, we can see that the curve was bent at distance between 20 and 40. so, consider the radius or EPS value as 30 at minimum points of 4.

```
dbcluster <- dbscan::dbscan(num_pharma_data, eps = 30, minPts = 4)
dbcluster
```

```
## 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
```

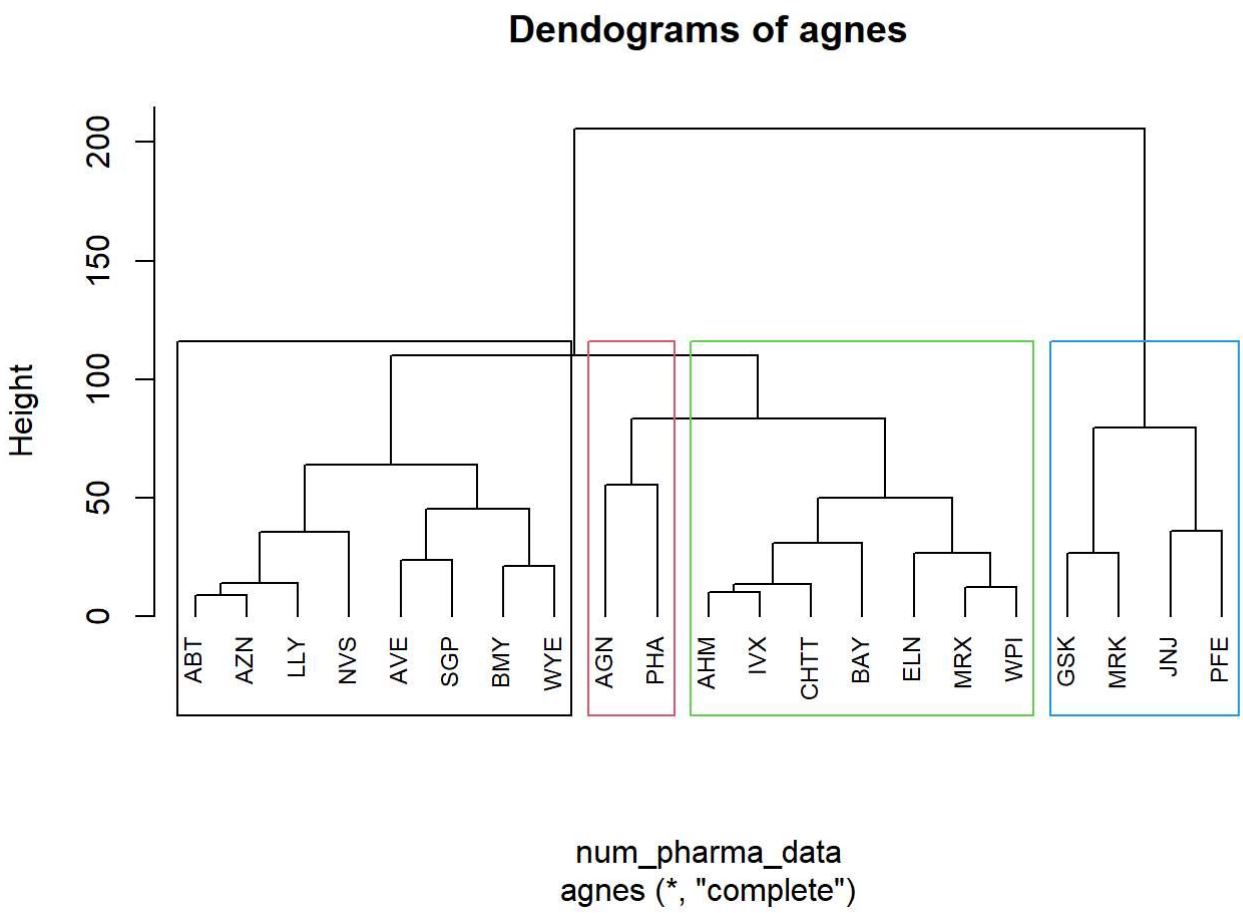
```
# Visualization of clusters
fviz_cluster(dbcluster, num_pharma_data) + ggtitle("DBSCAN Plot")
```



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.

3. Hierarchical Clustering

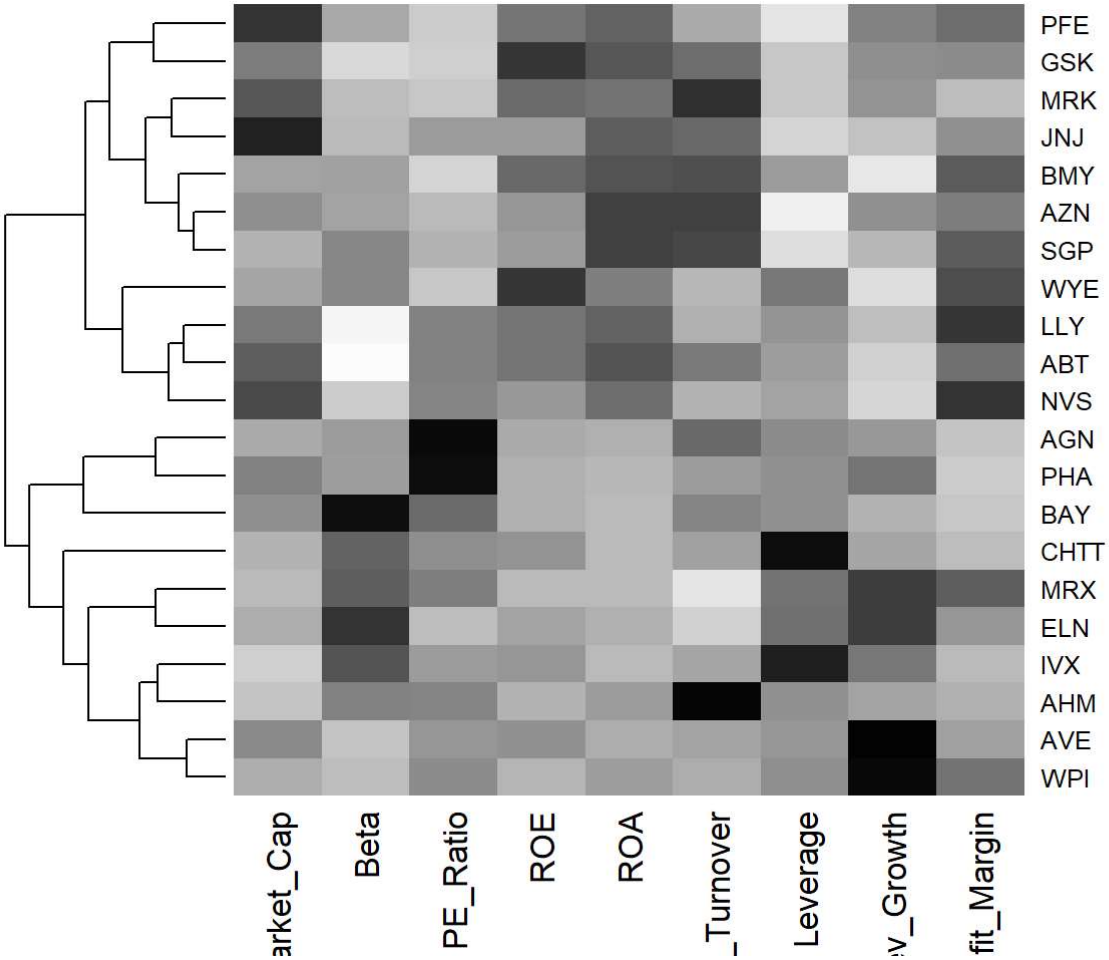
```
#Dissimilarity matrix
d <- dist(num_pharma_data, method = "euclidean")

#Hierarchical clustering using complete Linkage
hc_complete <- agnes(num_pharma_data, method = "complete")

#plot the obtained dendogram
pltree(hc_complete, cex = 0.75, hang = -1, main = "Dendograms of agnes")
rect.hclust(hc_complete, k = 4, border = 1:4)
```



In hierarchical clustering, 4 clusters are formed. From the above dendrogram we can say that first cluster with size 8 second cluster with size 2 third cluster with size 7 fourth cluster with size 4

But here the hierarchical clustering is not that suggestible because 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(Scale_pharma), 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
cluster1 <- pharmacy[,c(2:11)]  %>%
                  mutate(cluster = sil_cluster$cluster) %>% arrange(cluster, ascending = T)

# dataset with clusters
cluster1
```

```
##                                   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
```

```
cluster1[,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
```

calculating the mean of all numerical variables in each cluster

```
# calculate the mean of all numerical variables
aggregate(Scale_pharma, by=list(sil_cluster$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.

```
# add the clusters to the scaled data
scale_pharma1 <- data.frame(Scale_pharma, sil_cluster$cluster)
scale_pharma1
```

```
##      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 sil_cluster.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 the firms 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 firms 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 firms 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 firms 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 firms 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 firms 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 firms are riskier to invest than other firms 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 firms are distinguished by higher risk and potential for higher returns.

Cluster5 with firms 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 to 12)

```
# Add the clusters to the data
data1 <- pharmacy[12:14] %>% mutate(Clusters = sil_cluster$cluster)
data1
```

```
##      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
```

Based on mean values:

```
filter(data1, data1$Clusters==1)
```

```
##     Median_Recommendation Location Exchange Clusters
## AVE           Moderate Buy   FRANCE     NYSE        1
## ELN          Moderate Sell  IRELAND     NYSE        1
## MRX           Moderate Buy       US     NYSE        1
## WPI          Moderate Sell       US     NYSE        1
```

Cluster 1 - AVE, ELN, MRX, and WPI comprise Cluster 1. The highest metrics in this cluster are Market_cap, ROA, ROE, and Asset_Turnover; the lowest are Beta and PE_Ratio.

```
filter(data1, data1$Clusters==2)
```

```
##     Median_Recommendation Location Exchange Clusters
## AGN           Moderate Buy   CANADA     NYSE        2
## PHA                   Hold       US     NYSE        2
```

Cluster 2 - AGN, PHA make up Cluster 2 has the lowest PE Ratio, Asset Turnover, and the highest Rev_Growth.

```
filter(data1, data1$Clusters==3)
```

```
##     Median_Recommendation Location Exchange Clusters
## GSK                  Hold       UK     NYSE        3
## JNJ          Moderate Buy       US     NYSE        3
## MRK                  Hold       US     NYSE        3
## PFE          Moderate Buy       US     NYSE        3
```

Cluster 3 - GSK, JNJ, MRK, and PFE make up Cluster 3; it has the lowest Market Cap, ROE, ROA, Leverage, Rev Growth, and Net Profit Margin, and the highest Beta and Leverage.

```
filter(data1, data1$Clusters==4)
```

```
##      Median_Recommendation Location Exchange Clusters
## BAY                  Hold  GERMANY     NYSE        4
## CHTT          Moderate Buy       US   NASDAQ        4
## IVX                  Hold       US     AMEX        4
```

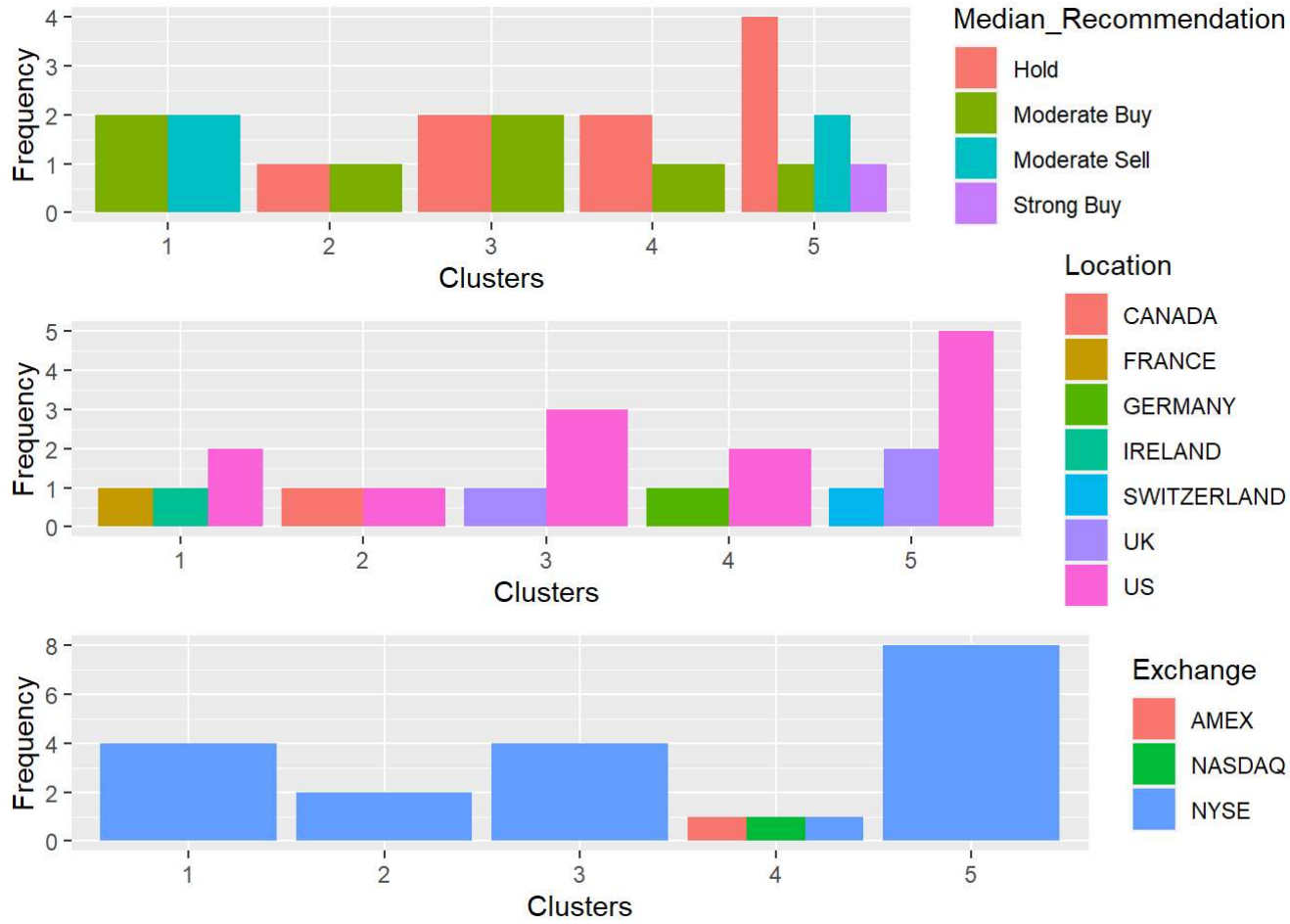
Cluster 4 - BAY, CHTT, and IVX make up Cluster 4, which has the lowest leverage and asset turnover ratios and the highest PE ratio.

```
filter(data1, data1$Clusters==5)
```

```
##     Median_Recommendation    Location Exchange Clusters
## ABT          Moderate Buy          US     NYSE        5
## AHM             Strong Buy         UK     NYSE        5
## AZN         Moderate Sell          UK     NYSE        5
## BMY         Moderate Sell          US     NYSE        5
## LLY                  Hold          US     NYSE        5
## NVS                  Hold SWITZERLAND     NYSE        5
## SGP                  Hold          US     NYSE        5
## WYE                  Hold          US     NYSE        5
```

Cluster 5: ABT, AHM, AZN, BMY, NVS, SGP, LLY, WYE ~ Cluster 5 has the lowest leverage, beta, and the highest Net Profit Margin.

```
# Plot the data with Median_Recommendation
recommendation <- ggplot(data1, mapping = aes(factor(Clusters), fill =Median_Recommendation)) +
geom_bar(position='dodge') + labs(x ='Clusters',y = 'Frequency')

# Plot the data with location
location <- ggplot(data1, mapping = aes(factor(Clusters), fill = Location)) + geom_bar(position
= 'dodge') + labs(x='Clusters',y = 'Frequency')

# Plot the data with Exchange
exchange <- ggplot(data1, mapping = aes(factor(Clusters), fill = Exchange)) + geom_bar(position
= 'dodge') + labs(x='Clusters',y = 'Frequency')

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



Cluster1, Recommended as Moderate Buy and Moderate Sell from Locations France, Ireland and US and was 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.

Appropriate names for the cluster

Cluster 1: Top Buying (High growth potential cluster)

Cluster 2: Significant Risk (High risK High reward cluster)

Cluster 3: Attempt it (Stability and profitablility cluster)

Cluster 4: Very Dangerous or Runaway (High risk and high beta cluster)

Cluster 5: A Perfect Asset (Low risk and high profitability cluster)