

Assignment_4

Niharika

2024-03-16

Summary

Questions - Answers

- a. 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. 1. For the clustering analysis of this dataset, all numeric variables from 1 to 9 were considered, encompassing financial factors like profit, market value, price-to-earning ratio, return on equity, return on assets, leverage, etc. Equal weights were assigned to these variables, acknowledging their equal role in defining a firm's equity.

Exploring three clustering algorithms—Kmeans, DBSCAN, and Hierarchical clustering—revealed that Kmeans yielded the most favorable results. DBSCAN resulted in two clusters with 15 points, designating the remaining six points as outliers, rendering it unsuitable for this dataset. Hierarchical clustering, on the other hand, produced four clusters, with varying sizes, but Kmeans with five clusters stood out as the preferred choice.

The decision to use Kmeans with five clusters was influenced by both the silhouette method, which indicated an optimal value of 5, and the elbow method, which suggested 2 clusters. However, the clusters formed with five points demonstrated better cohesion with points closer to their centroids.

The resulting clusters are as follows:

1. Cluster 1 (Size: 4) - Companies: AVE, WPI, MRX, ELN

2. Cluster 2 (Size: 2) - Companies: PHA, AGN

3. Cluster 3 (Size: 4) - Companies: GSK, PFE, MRK, JNJ

4. Cluster 4 (Size: 3) - Companies: IVX, CHTT, BAY

5. Cluster 5 (Size: 8) - Companies: WYE, BMY, LLY, AZN, NVS, ABT, SGP, AHM

- b. Interpret the clusters with respect to the numerical variables used in forming the clusters. A. The clusters, based on numerical variables, reveal distinct characteristics for each group of companies:

Cluster 1 (AVE, WPI, MRX, ELN): Companies in this cluster exhibit high revenue growth and beta values but have low asset turnover, return on equity, return on assets, and market capitalization. This suggests that these companies are likely in their early stages of growth, heavily investing in marketing and sales. The high revenue growth and beta values indicate an expectation of improved earnings in the future, emphasizing their higher growth potential despite current low profitability.

Cluster 2 (PHA, AGN): This cluster includes companies with high price-to-earnings ratios and asset turnover but low net profit margin, return on equity, return on assets, and market capitalization. The high asset turnover and price-to-earnings ratios suggest an anticipated improvement in earnings, though with little past net profit. However, the higher risk is associated with the high stock price.

Cluster 3 (IVX, CHTT, BAY): Companies in this cluster have high market capitalization, return on equity, return on assets, and asset turnover, but they exhibit the lowest beta and profit-to-return ratio. These characteristics point to mature and well-established companies with stable stock prices and lower risk. However, their lower profit return ratio indicates a potential inefficiency in profit generation. This cluster is distinguished by maturity, stability, and overall profitability.

Cluster 4 (WYE, BMY, LLY, AZN, NVS, ABT, SGP, AHM): Companies in this cluster have high beta values and leverage but lower net profit margin, market capitalization, return on equity, return on assets, revenue growth, and profit return ratio. This suggests that these companies are riskier investments, given their unstable stock prices, higher debt levels, and lower profit margins. However, in a bullish market, they have the potential for higher returns.

Cluster 5 (GSK, PFE, MRK, JNJ): This cluster comprises companies with the highest net profit margin, asset turnover, return on equity, and return on assets. They also have the lowest beta, profit-to-return ratio, and revenue growth. These characteristics indicate high financial performance and low risk, portraying a group of mature and well-established companies with stable stock prices and strong profitability. ***** c. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters) 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.

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

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

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

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

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

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## — Attaching core tidyverse packages ————— tidyverse 2.0.0 —
## ✓ dplyr     1.1.4    ✓ readr     2.1.5
## ✓ forcats   1.0.0    ✓ stringr   1.5.1
## ✓ lubridate 1.9.3    ✓ tibble    3.2.1
## ✓ purrr     1.0.2    ✓ tidyrr    1.3.1
```

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

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

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

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

```
library(dbSCAN)
```

```
## Warning: package 'dbSCAN' was built under R version 4.3.3
```

```
##
## 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 4.3.3
```

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

```
## Warning: package 'gridExtra' was built under R version 4.3.3
```

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

Import the data which was in CSV format

```
# import the data
pharmaceuticals.data <- read.csv("C:\\\\Users\\\\nihar\\\\Downloads\\\\Pharmaceuticals.csv")
dim(pharmaceuticals.data)
```

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

```
t(t(names(pharmaceuticals.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

```
set.seed(159) # Important to ensure that we get the same sample if we rerun the code
# Remove the unwanted columns
row.names(pharmaceuticals.data) <- pharmaceuticals.data[,1]
clustering.data <- pharmaceuticals.data[,3:11]# 1 and 5 are the indexes for columns ID and ZIP
dim(clustering.data)
```

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

```
# Summary of the data
summary(clustering.data)
```

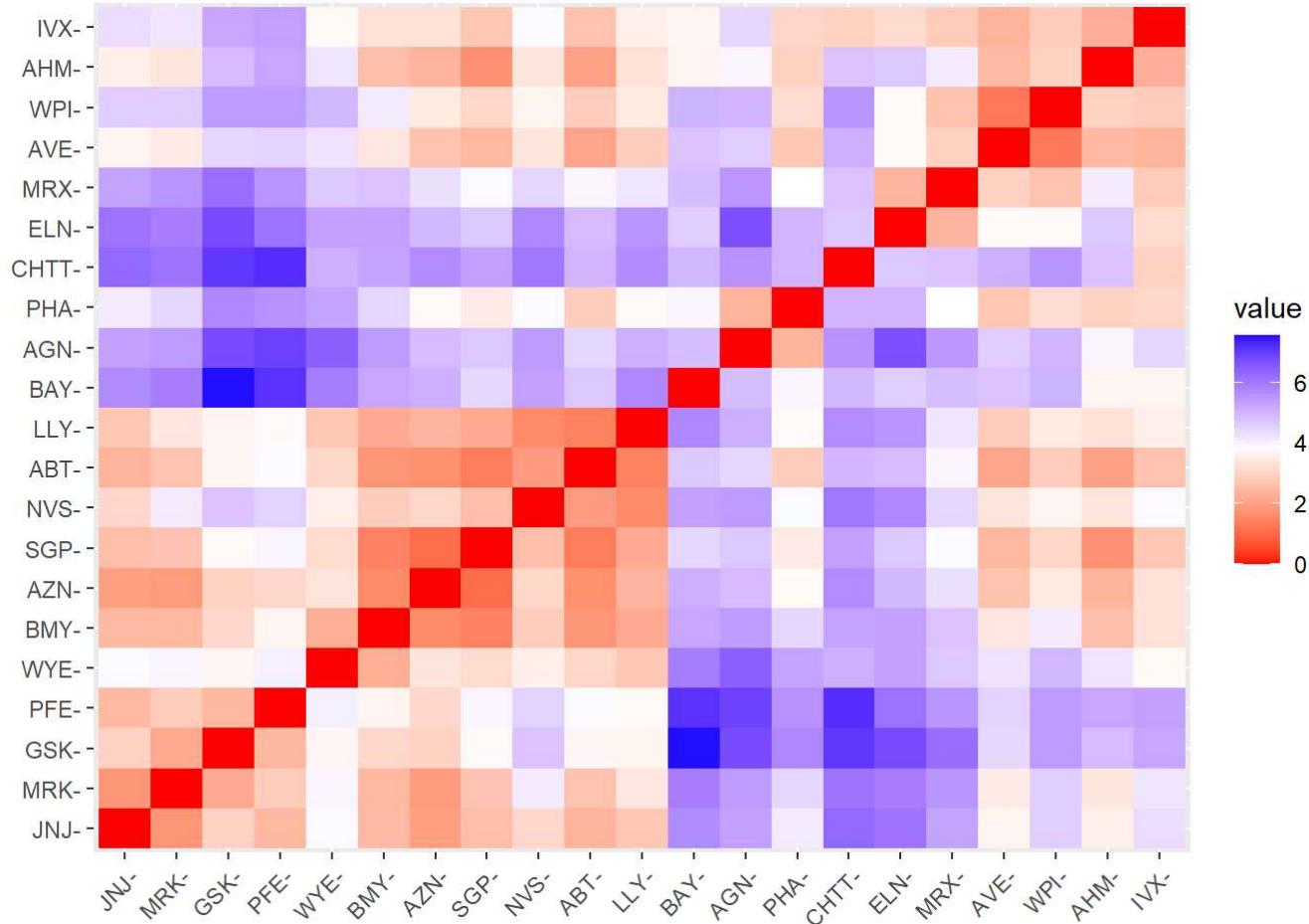
```
##      Market_Cap          Beta        PE_Ratio         ROE
## Min.   : 0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
## 1st Qu.: 6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
## Median :48.19   Median :0.4600   Median :21.50   Median :22.6
## Mean   :57.65   Mean   :0.5257   Mean   :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

```
# scale the data using scale function
scaling_data <- scale(clustering.data)
head(scaling_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
```

```
# distance between each variable
dis <- get_dist(scaling_data)
# Visualize the distance
fviz_dist(dis)
```



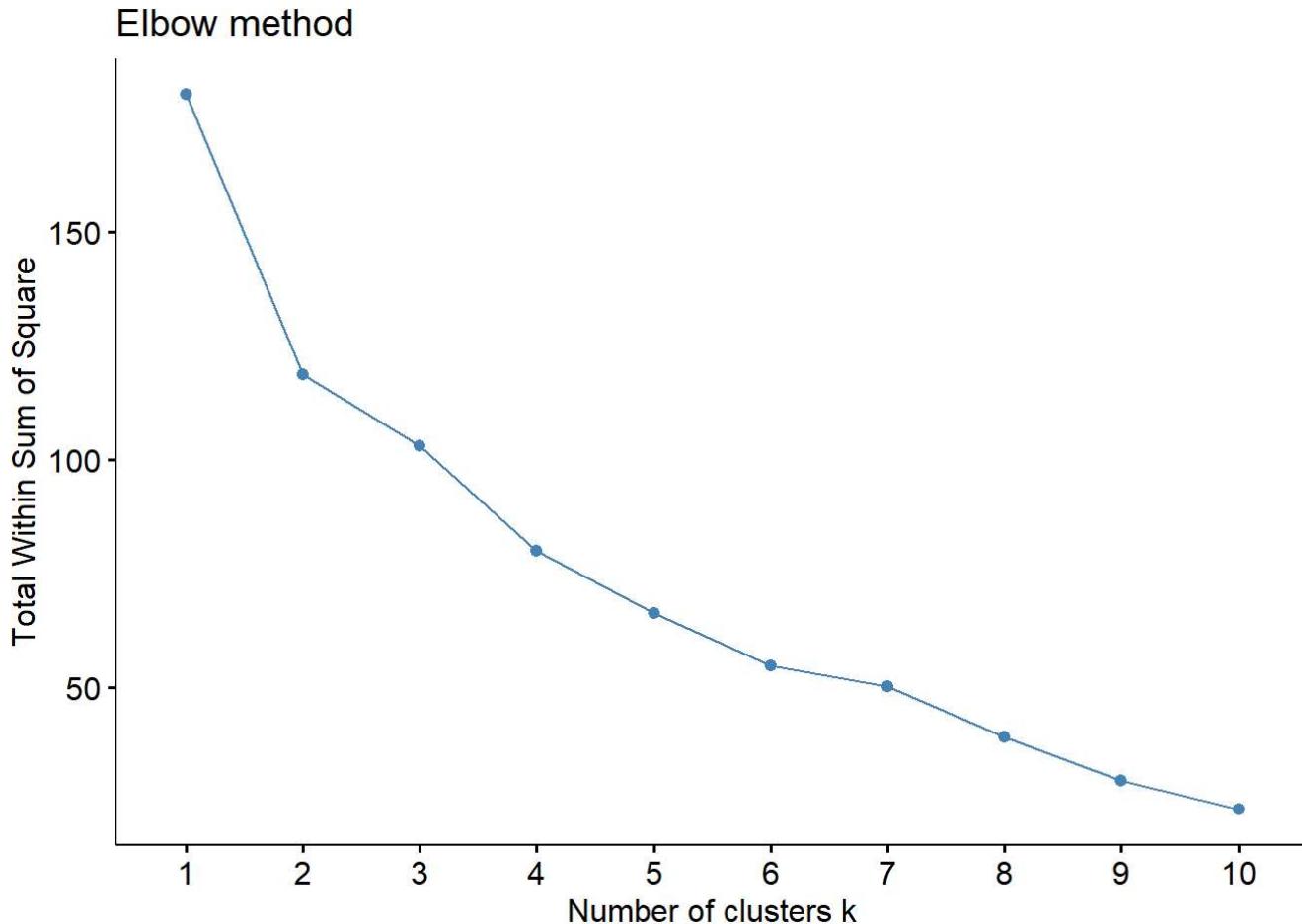
Questions

a. 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

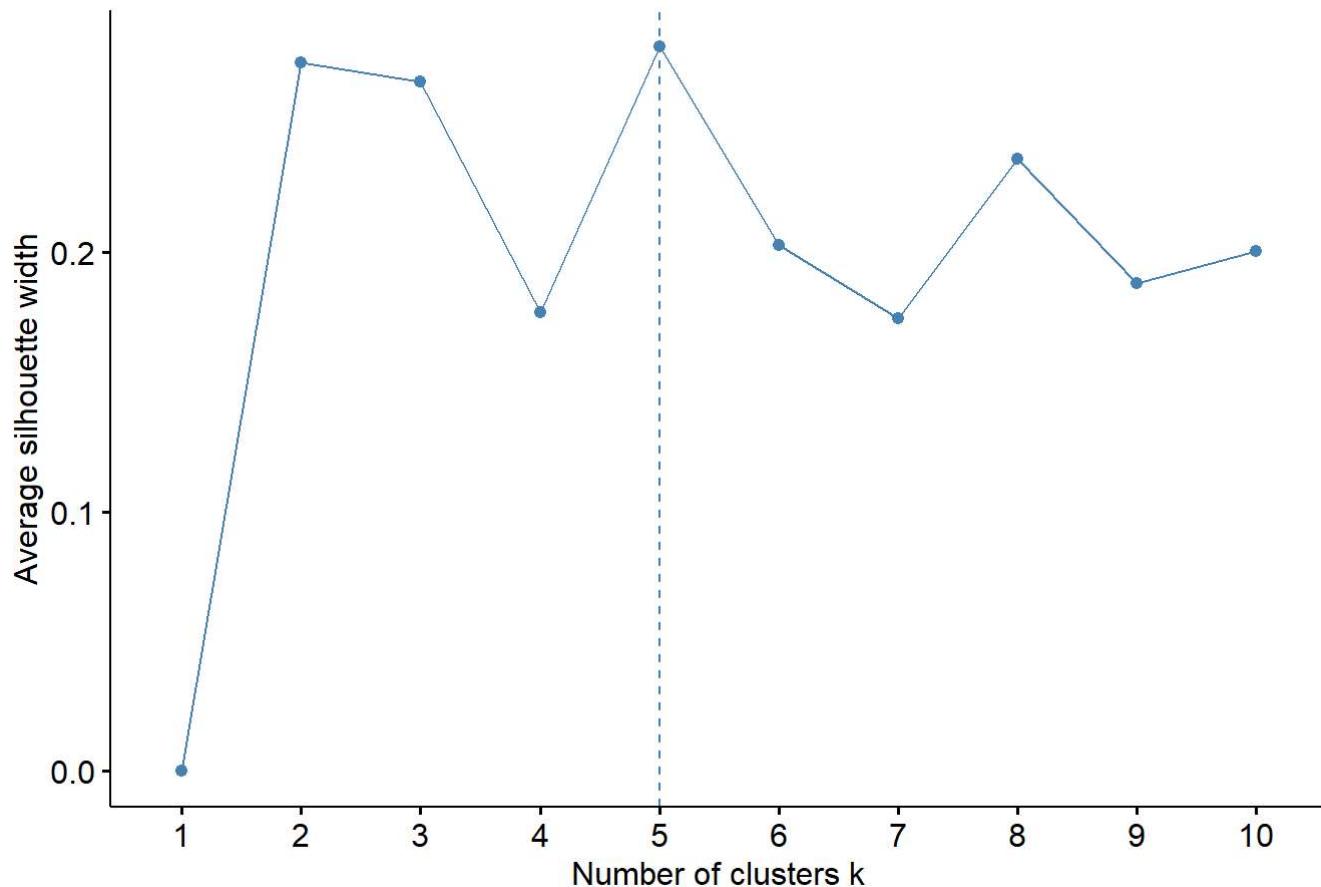
For getting the best value of K(no. of clusters) for kmeans

```
# sum of squares method  
fviz_nbclust(scaling_data, kmeans, method = "wss") + ggtitle("Elbow method")
```



```
# silhouette method  
fviz_nbclust(scaling_data, kmeans, method = "silhouette") + ggtitle("Silhouette method")
```

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.wss1 <- kmeans(scaling_data, centers = k, nstart=21)
k.wss1
```

```

## 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     1     2     2     1     2     2     1     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"

```

```

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

```

```

## These are the centers of the clusters

```

```

k.wss1$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

```

```

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

```

```

## The Size of the each cluster is

```

```

k.wss1$size

```

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

```
# To get which point belongs to which cluster
k.wss1$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 2 1 1
## PFE PHA SGP WPI WYE
## 1 2 1 2 1
```

The results of the Kmeans clustering with a k value of 2 indicate that 11 companies form one cluster, while the remaining 10 are grouped into another cluster. Considering all numerical variables crucial for understanding equity, such as market capitalization, net profit, return on assets, and asset turnover, it becomes apparent that some data points, like AGN, ELN, GSK, etc., are notably distant from the centroids. This suggests that the chosen number of clusters may not be sufficient to accurately represent the underlying structure of the data. ***** 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.

```
# consider k=5
k <- 5
set.seed(159)
# kmeans algorithm
k.sil <- kmeans(scaling_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"

```

```

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

```

```

## These are the centers of the clusters

```

```

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
```

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

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

```
k.sil$size
```

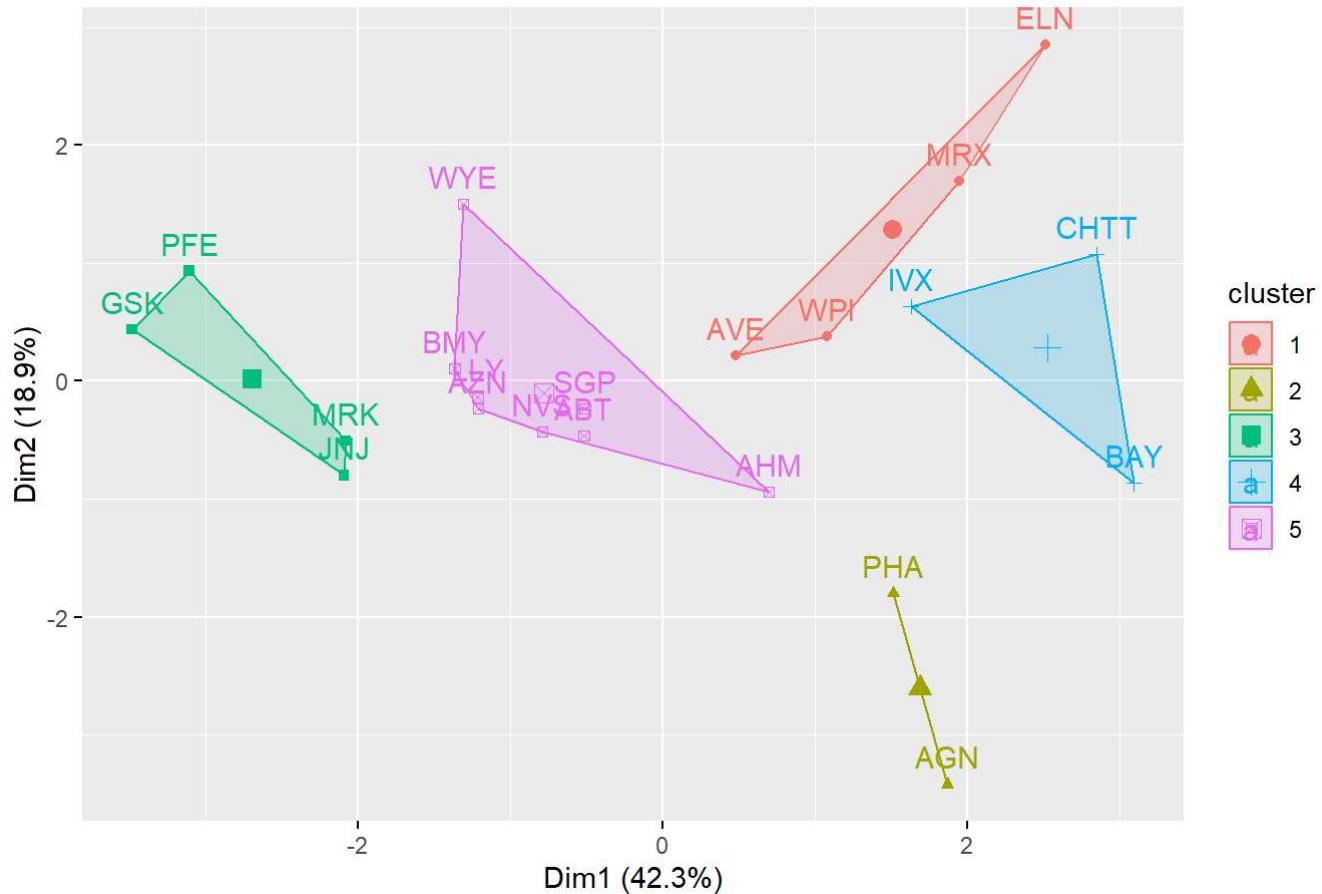
```
## [1] 4 2 4 3 8
```

```
# 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
```

```
# Visualization of clusters
fviz_cluster(k.sil, scaling_data) + ggtitle("k=5")
```

k=5

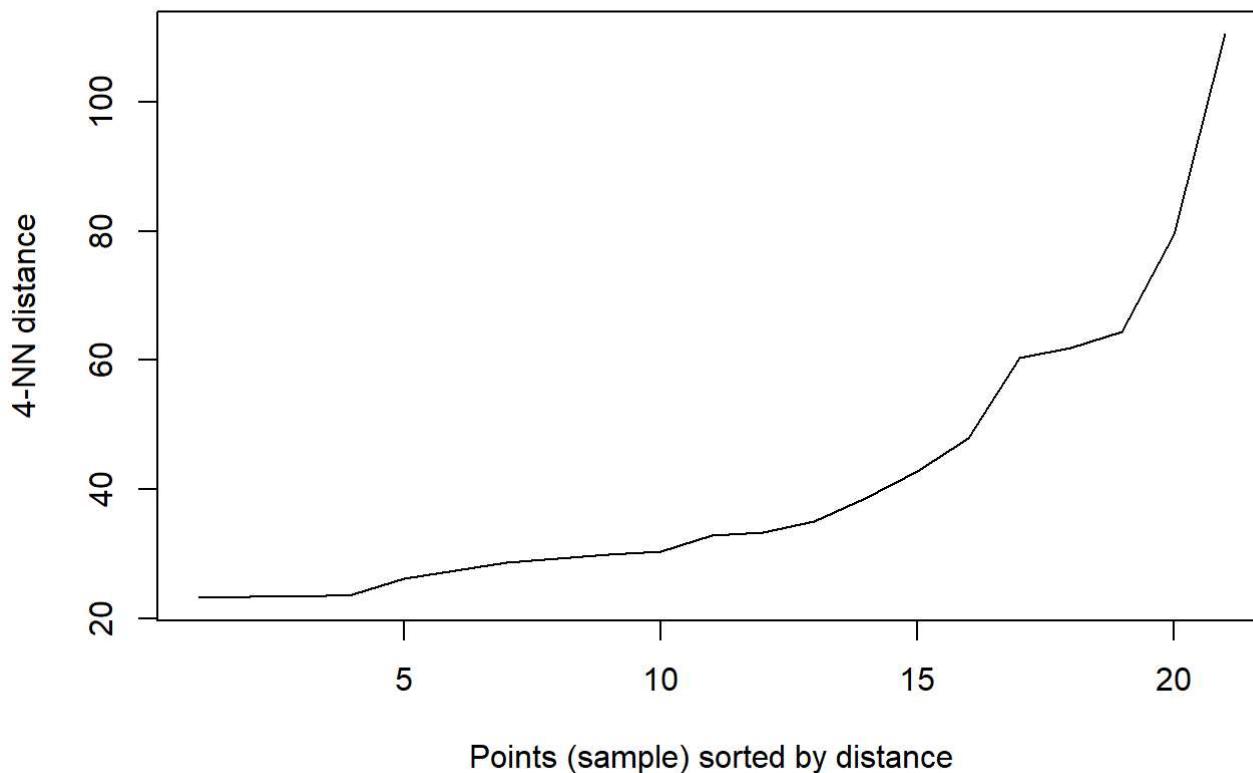


Examining the outcome of the Kmeans clustering with a k value of 5 reveals that the first cluster comprises 4 companies, the second cluster has 2 companies, the third cluster contains 3 companies, the fourth cluster includes 8 companies, and the remaining companies are assigned to the fifth cluster. All relevant numerical variables influencing equity, such as market capitalization, net profit, return on assets, and asset turnover, were considered. Notably, the data points appear much closer to the centroids in this clustering configuration, suggesting that this clustering arrangement is more favorable. Now, let's delve into the characteristics of the remaining clusters.

DBSCAN Clustering

To get the best value of radius or eps.

```
# Graph to get the best value of radius at min points of 4.
dbSCAN::kNNdistplot(clustering.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_1 <- dbSCAN::dbSCAN(clustering.data, eps = 30, minPts = 4)

# Output of the clusters
print(dbs_1)
```

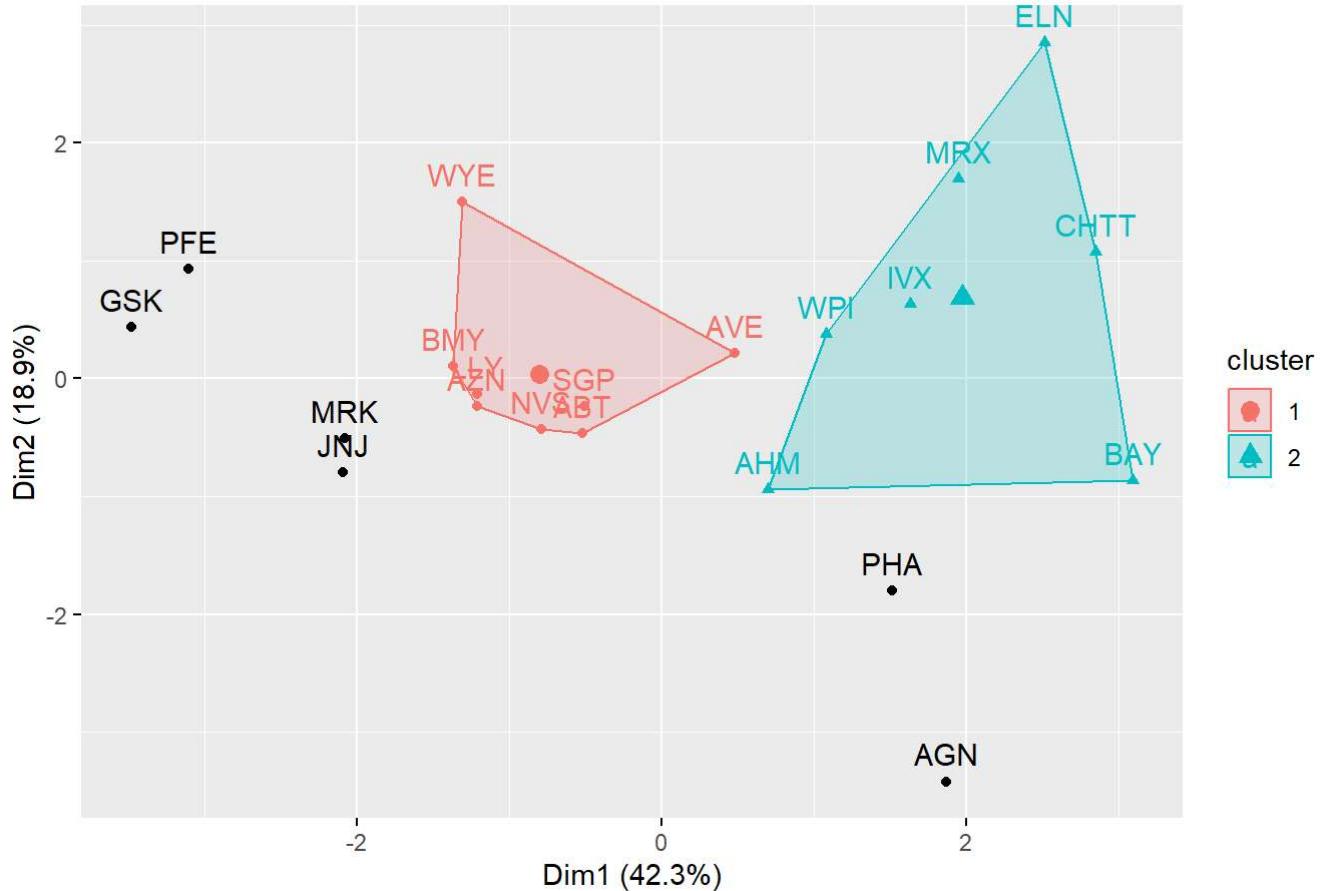
```
## DBSCAN clustering for 21 objects.
## Parameters: eps = 30, minPts = 4
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 2 cluster(s) and 6 noise points.
##
## 0 1 2
## 6 8 7
##
## Available fields: cluster, eps, minPts, dist, borderPoints
```

```
# To get which point belongs to which cluster
print(dbs_1$cluster)
```

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

```
# Visualization of clusters
fviz_cluster(dbs_1, clustering.data) + ggtitle("DBSCAN Plot")
```

DBSCAN Plot



Analyzing the results and visual representation of the DBSCAN clustering with a radius of 30 and a minimum of 4 points, it's evident that two clusters have been formed. The first cluster consists of 8 points, the second cluster has 7 points, and there are six remaining points identified as outliers. The presence of a notable number of outliers, as observed in the plot, indicates that this clustering process may not be considered effective. A desirable cluster should ideally have a minimal number of outliers.

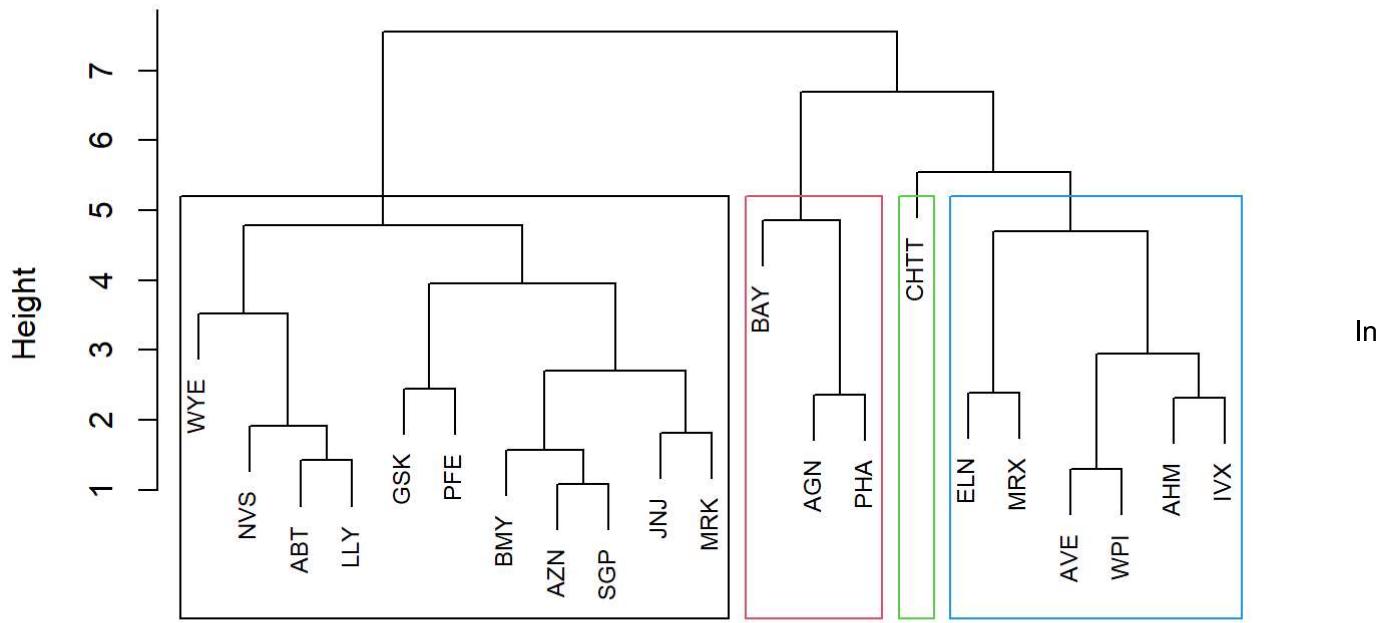
Hierarchical Clustering

```
# Get the euclidean distance for the data
d1_1 <- dist(scaling_data, method = "euclidean")

# Hierarchical Clustering
hc1_1 <- hclust(d1_1, method = "complete")

# Visualize the output Dendrogram at height=5
plot(hc1_1, cex = 0.75, main = "Dendrogram of Hierarchical Clustering")
rect.hclust(hc1_1, h=5, border = 1:4)
```

Dendrogram of Hierarchical Clustering

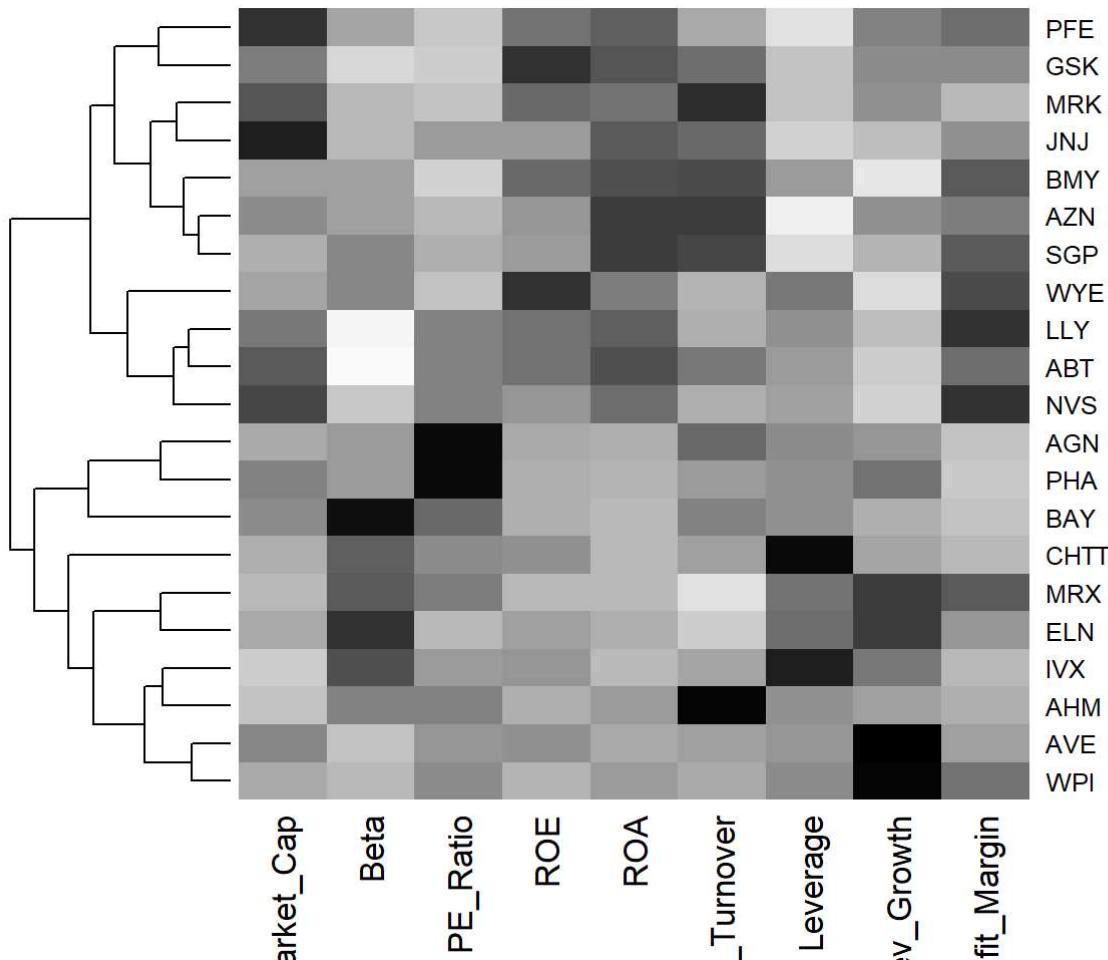


d1_1
hclust (*, "complete")

hierarchical clustering, a height value of 5 was chosen, where clusters are formed based on the distances between merged clusters compared to the remaining heights. At this height, the dendrogram indicates the creation of four clusters with the following sizes: 1. The first cluster has 11 data points. 2. The second cluster consists of 3 data points. 3. The third cluster comprises a single data point. 4. The fourth cluster includes 6 data points.

Observing the distribution of points across clusters, it's evident that one cluster has a substantial number of points, while another has a considerably smaller count. This uneven distribution suggests that this clustering approach may not be optimal for effectively grouping all the companies.

```
heatmap(as.matrix(scaling_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.

b. Interpret the clusters with respect to the numerical variables used in forming the clusters.

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

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

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

```
clustering.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

```
# calculate the mean of all numerical variables
aggregate(scaling_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.

```
# add the clusters to the scaled data
scaled.data1 <- data.frame(scaling_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
<hr/>						
## 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.

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

```
# Add the clusters to the data
data_pattern1 <- pharmaceuticals.data[12:14] %>% mutate(Clusters = k.sil$cluster)
data_pattern1
```

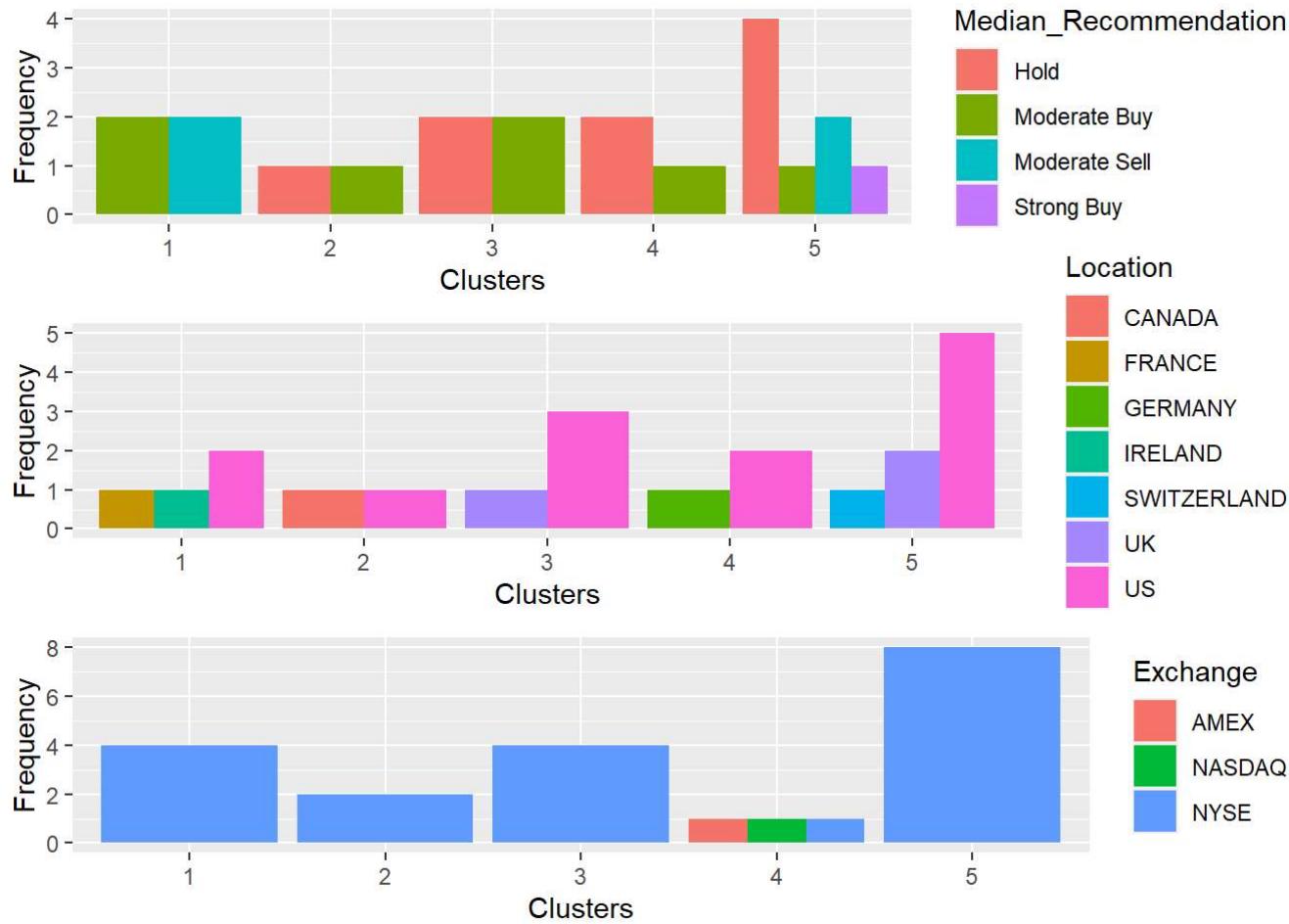
	Median_Recommendation	Location	Exchange	Clusters
## ABT	Moderate Buy	US	NYSE	5
## AGN	Moderate Buy	CANADA	NYSE	2
## AHM	Strong Buy	UK	NYSE	5
## AZN	Moderate Sell	UK	NYSE	5
## AVE	Moderate Buy	FRANCE	NYSE	1
## BAY	Hold	GERMANY	NYSE	4
## BMY	Moderate Sell	US	NYSE	5
## CHTT	Moderate Buy	US	NASDAQ	4
## ELN	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_pattern1, mapping = aes(factor(Clusters), fill = Median_Recommendation)) + geom_bar(position='dodge') + labs(x = 'Clusters',y = 'Frequency')

# Plot the data with Location
location <- ggplot(data_pattern1, mapping = aes(factor(Clusters), fill = Location)) + geom_bar(position = 'dodge') + labs(x='Clusters',y = 'Frequency')

# Plot the data with Exchange
exchange <- ggplot(data_pattern1, mapping = aes(factor(Clusters), fill = Exchange)) + geom_bar(position = 'dodge') + labs(x='Clusters',y = 'Frequency')

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



Cluster 1 is suggested for a combination of Hold and Moderate Buy recommendations, primarily located in North America (US/Canada) and listed on the NYSE.

Cluster 2 is recommended for Hold and Moderate Sell for the majority of its companies. These companies are situated in the US, Switzerland, and the UK, and are listed on the NYSE.

Cluster 3, with a recommendation of Moderate Buy and Moderate Sell, is located in France, Ireland, and the US, and is listed on the NYSE.

Cluster 4 is advised for Hold and Moderate Buy recommendations, primarily located in the UK and the US, and listed on the NYSE.

Cluster 5 is recommended for Hold and Moderate Buy, with companies located in AMEX, Germany, and the US, and listed on both NASDAQ and NYSE.

To summarize: - Cluster 1: Moderate Buy and Moderate Sell in France, Ireland, and the US, listed on NYSE.

- Cluster 2: Hold and Moderate Sell in the US, Switzerland, and the UK, listed on NYSE.
- Cluster 3: Hold and Moderate Buy in the UK and the US, listed on NYSE.