FML ASSIGNMENT 4 - 811290653

GEETHIKA VULLI

2023-11-10

SUMMARY

For this project, I will perform a non-various leveled bunch examination using the k-implies grouping strategy. The aim is to segregate the data into uniform groups so that important information can be extracted. First, we should stack the first dataset and the required bundles. It has information from about 21 pharmaceutical companies.

Justification for Selecting Market Capitalization, Beta, PE Ratio, ROE, ROA, Leverage, Rev Growth, and Net Profit Margin The chosen variables are typical financial metrics that are used to assess and compare the performance of businesses. They include Market Cap, Beta, PE Ratio, ROE, ROA, Asset Turnover, Leverage, Rev Growth, and Net Profit Margin. All of these factors combined offer a thorough picture of the efficiency, profitability, and financial stability of a company.

1. Market Capitalization: Varies from 0.41 to 199.47. shows the pharmaceutical companies' total size and market value.
2. Beta: Indicates how sensitive a company's returns are to changes in the market and varies from 0.18 to 1.11.
3. PE Ratio: expresses the value of a company's stock in relation to its earnings. It can range from 3.6 to 82.5.
4. ROI : varies between 3.9 and 62.9. shows the efficiency with which a company uses shareholder equity to turn a profit.
5. ROA : 0.3 to 1.1. Evaluates the capacity of an organization to make money off of its assets.
6. Asset Turnover: Indicates how well a company uses its assets to produce income. It's from 0.5 to 1.1.
7. Leverage: shows the extent to which a business uses debt to finance its operations; Ranges from 0 to 3.51.
8. Rev_Growth: Shows the percentage change in revenue over a given time period and varies from -3.17 to 34.21.
9. Net Profit Margin: This variable shows the percentage of revenue that is converted to profit and ranges from 2.6 to 25.54.

Normalising the data: For every variable to contribute proportionately to the clustering process, the numerical variables must be normalized. Normalization helps prevent one variable from dominating the clustering based solely on its magnitude because these variables may have different units or scales. In contrast, Beta is a fraction between 0 and 1, whereas Market Cap is in the hundreds.

K-means is frequently used in exploratory data analysis to find patterns and groupings within the data, and K-means clustering can reveal information about the financial profiles of pharmaceutical companies, which is

why I've chosen it over DBSCAN. DBSCAN is useful for datasets with dense regions because it can identify groups of companies with comparable financial characteristics, supporting investment analysis or strategic decision-making. It is also simple to interpret.K-means necessitates a predetermined number of clusters (k). This might be advantageous in some circumstances because the user can choose how many clusters to create. DBSCAN and hierarchical clustering may not provide a clear-cut choice for the number of clusters.

Five groups are created from the dataset based on numerical variables. Financial ratios and performance metrics are taken into consideration when providing an interpretation of each cluster. Net profit margin, revenue growth, leverage, beta, ROA, and ROE are examples of cluster characteristics.

cluster 1 - The hold,moderate sell,moderate buy and moderate sell are in order of greatest to lowest. They are listed on the NYSE and originated from the US, the UK, and Switzerland,in which US is the highest.

cluster 2 - has a different Hold and Moderate Buy median,where the hold is greater than the moderate buy, a different count from the US and Germany, and a different country count, the firms are evenly divided among AMEX,NASDAQ and NYSE.

cluster 3 - is only listed on the NYSE, has equal Hold and Moderate Buy medians, and is evenly divided across the US and Canada

cluster 4 - is distributed throughout the United States and the United Kingdom, has the same hold and moderate buy medians, and is also listed on the NYSE.

cluster 5 - has equal moderate buy and moderate sell , they are also distributed in France,Ireland,US countries and listed under NYSE exchange.

We examine the correlations between variables 10 to 12 and clusters. Within each cluster, the frequency distribution of non-clustered variables is shown using bar plots.Using the bar graph the explanation and the appropriate names are provided below the graph.

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:

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:

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

#installing the libraries using install.packages() and calling the requried libraries

```
library(tidyverse)  # data manipulation
```

```
## -- 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 ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
```

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

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

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

```
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
library(dbscan)
```

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

```r
#importing the dataset and reading the dataset
dataset <- read.csv("C:\\Users\\geeth\\Downloads\\Pharmaceuticals.csv")
head(dataset)
```

```
##   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
```

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

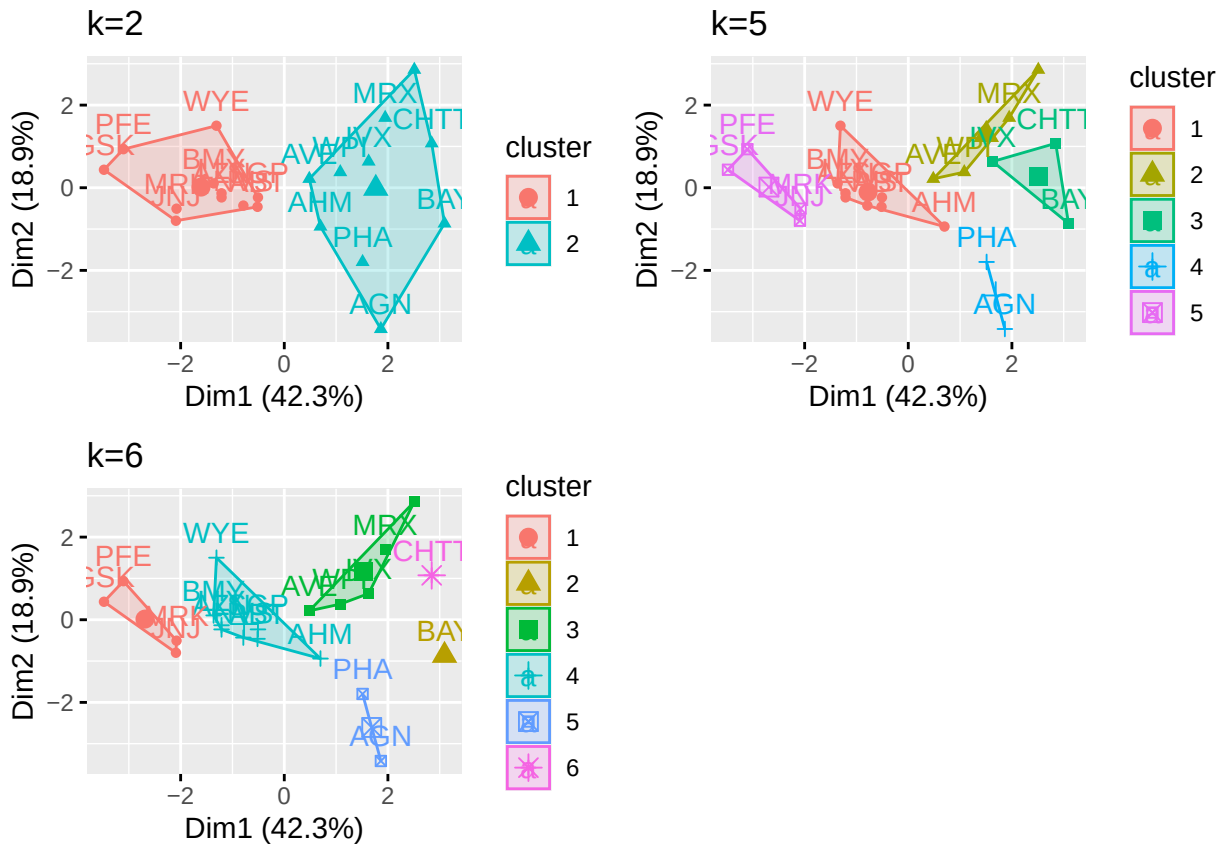
```r
#To remove any missing value that might be present in the data
P_data <- na.omit(dataset)
#Collecting numerical variables from column 1 to 9 to cluster 21 firms
row.names(P_data)<- P_data[,1]
Ph<- P_data[, 3:11]
head(Ph)
```

```
##     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
```

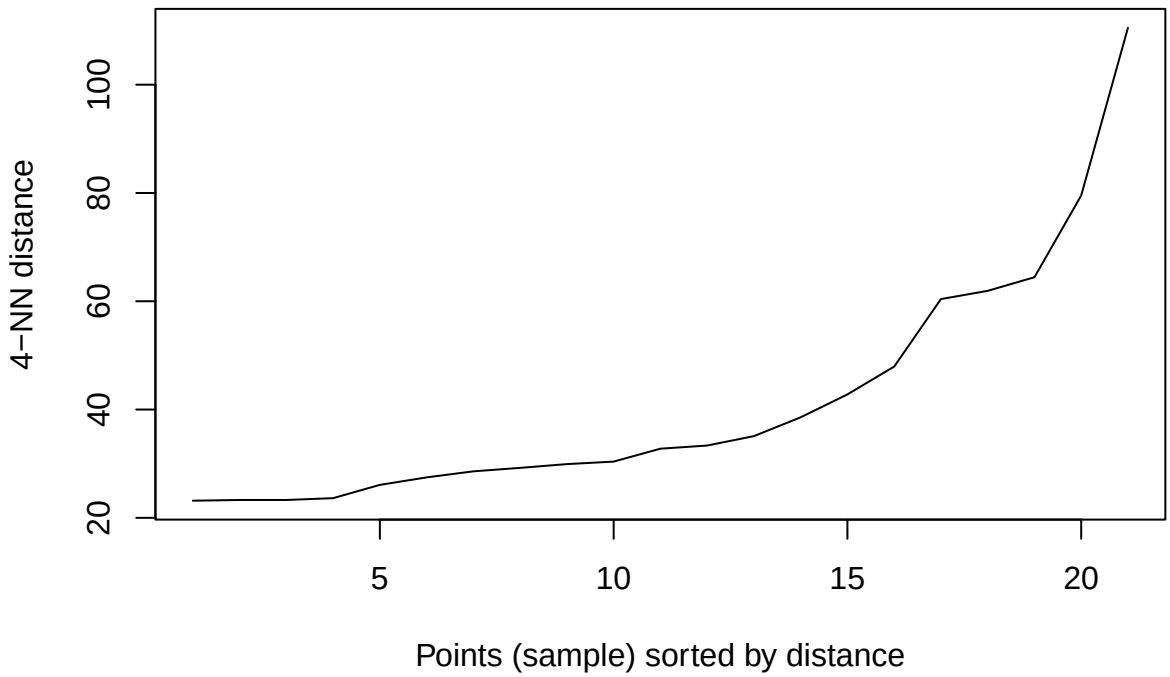
```r
#normalizing the data using Scale function
ph2<- scale(Ph)
head(ph2)
```

```
##      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
#Computing K-means clustering in R for different centers
#Using multiple values of K and examine the differences in results

km <- kmeans(ph2, centers = 2, nstart = 30)
km1<- kmeans(ph2, centers = 5, nstart = 30)
km2<- kmeans(ph2, centers = 6, nstart = 30)
Pl1<-fviz_cluster(km, data = ph2)+ggtitle("k=2")
pl2<-fviz_cluster(km1, data = ph2)+ggtitle("k=5")
pl3<-fviz_cluster(km2, data = ph2)+ggtitle("k=6")
grid.arrange(Pl1,pl2,pl3, nrow = 2)
```



```
#To get the best value of radius or eps.

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



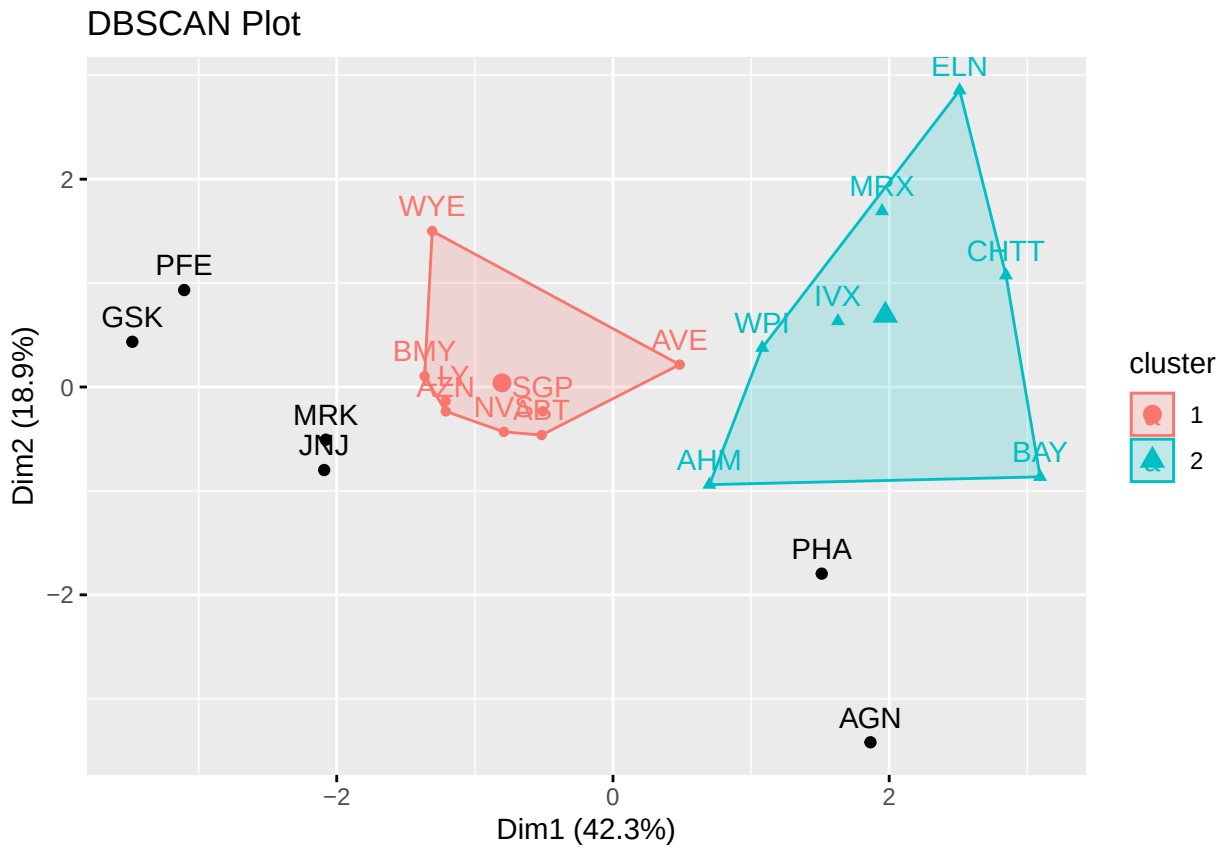

```
# DBSCAN Algorithm at eps=30 and minpts =4
dbs <- dbscan::dbscan(Ph, eps = 30, minPts = 4)

# 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)
```

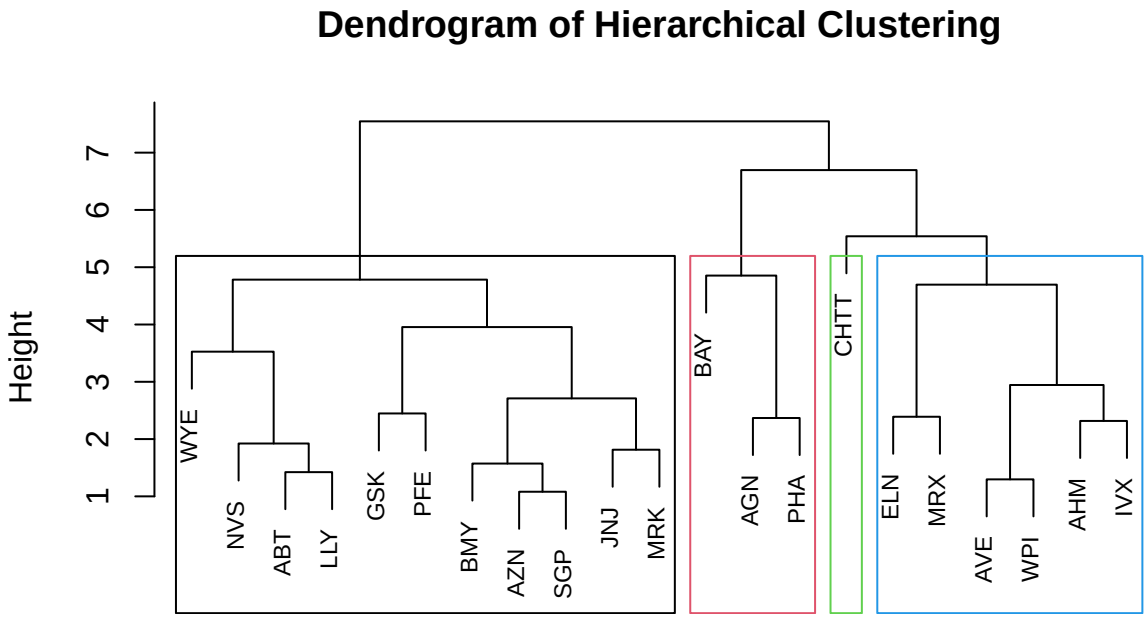
```
##  [1] 1 0 2 1 1 2 1 2 2 1 0 2 0 2 0 1 0 0 1 2 1
```

```
# Visualization of clusters
fviz_cluster(dbs, Ph) + ggtitle("DBSCAN Plot")
```



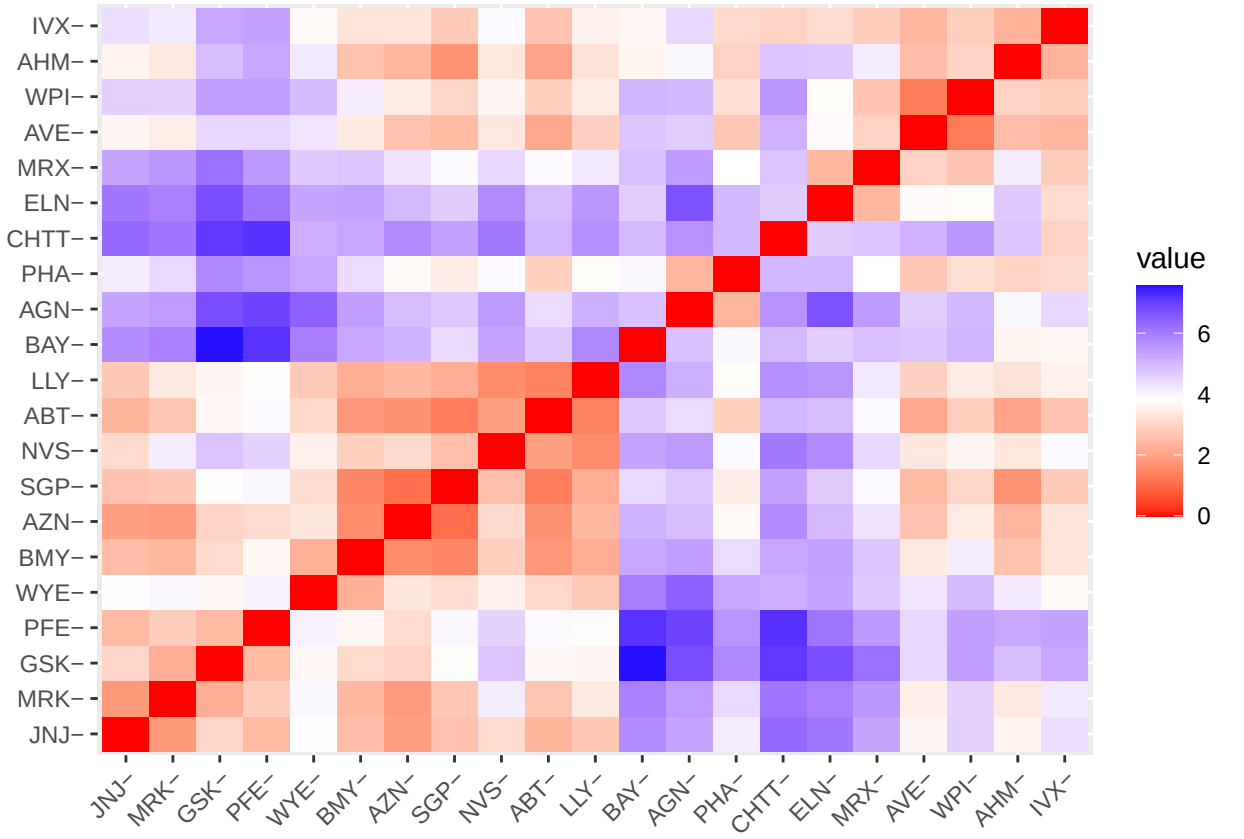
#I've chosen K-means over DBSCAN because it's frequently used in exploratory data analysis to find patterns and groupings in the data, and because K-means clustering can reveal information about the financial profiles of pharmaceutical companies. DBSCAN is useful for datasets with dense regions and can help with investment analysis and strategic decision-making by revealing groups of companies with comparable financial characteristics. It is also simple to interpret.K-means necessitates a predetermined number of clusters (k). This might be advantageous in some circumstances because the user can choose how many clusters to create. DBSCAN and hierarchical clustering may not provide a clear-cut choice for the number of clusters.

```
# Hierarchical Clustering

# Get the euclidean distance for the data
ed <- dist(ph2, method = "euclidean")

# Hierarchical Clustering
h <- hclust(ed, method = "complete")

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



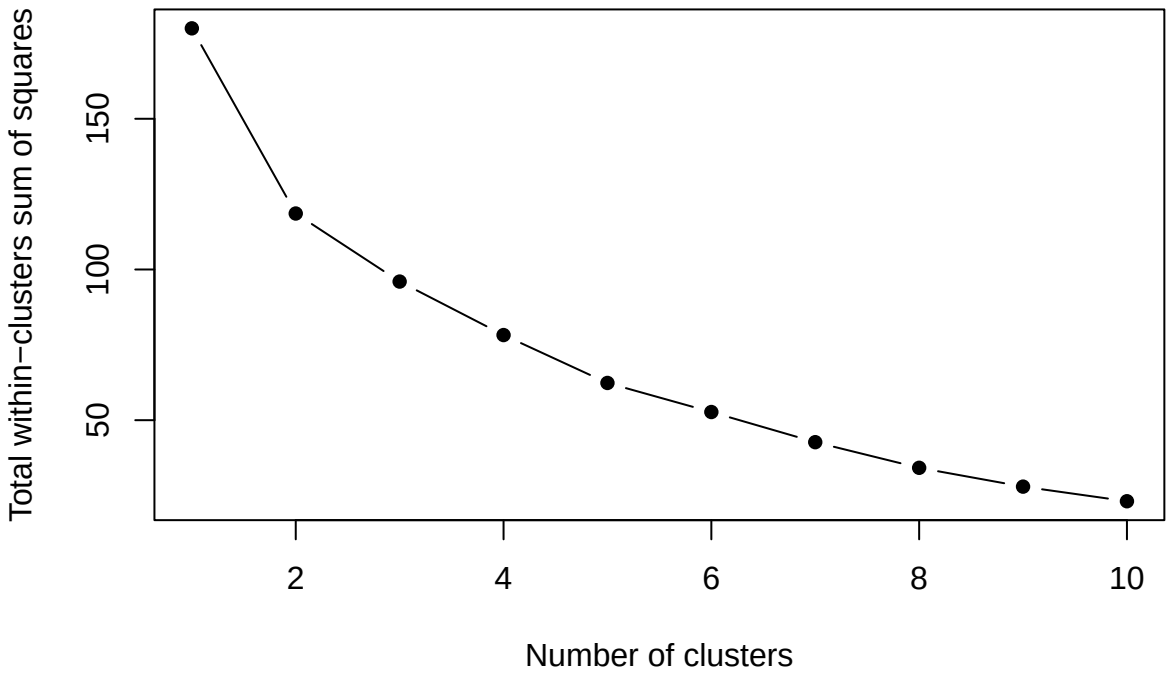
ed
hclust (*, "complete")

```
#Determining optimal clusters using Elbow method
dis <-  dist(ph2, method = "euclidean")# for calculating
#distance matrix between rows of a data matrix.
fviz_dist(dis)# Visualizing a distance matrix
```



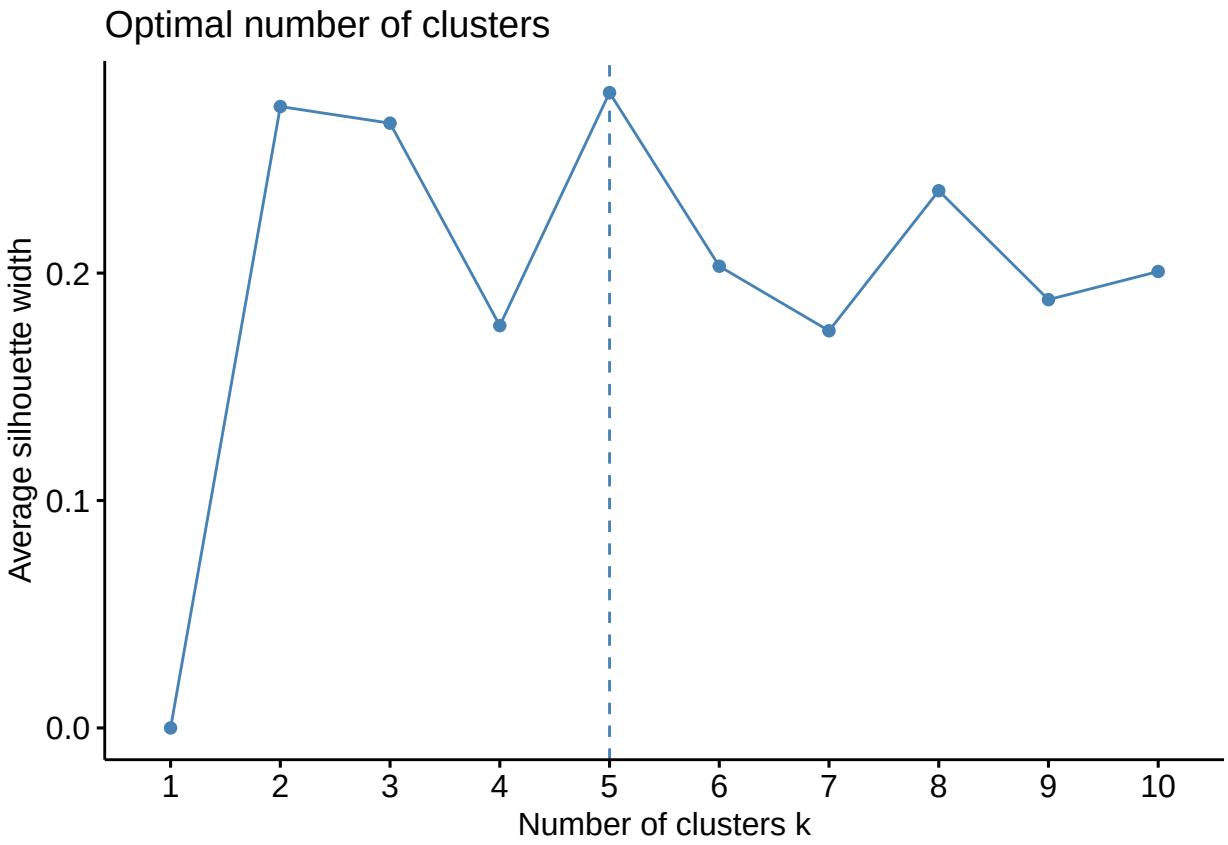
#For each k, calculate the total within-cluster sum of square (wss) tot.withinss is total within-cluster sum of squares Compute and plot wss for k = 1 to k = 10 extract wss for 2-15 clusters The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters k =5.

```
set.seed(123)
wss<- function(k){
kmeans(ph2, k, nstart =10)$tot.withinss
}
k.values<- 1:10
wss_cluster<- map_dbl(k.values, wss)
plot(k.values, wss_cluster,
     type="b", pch = 16, frame = TRUE,
     xlab="Number of clusters",
     ylab="Total within-clusters sum of squares")
```

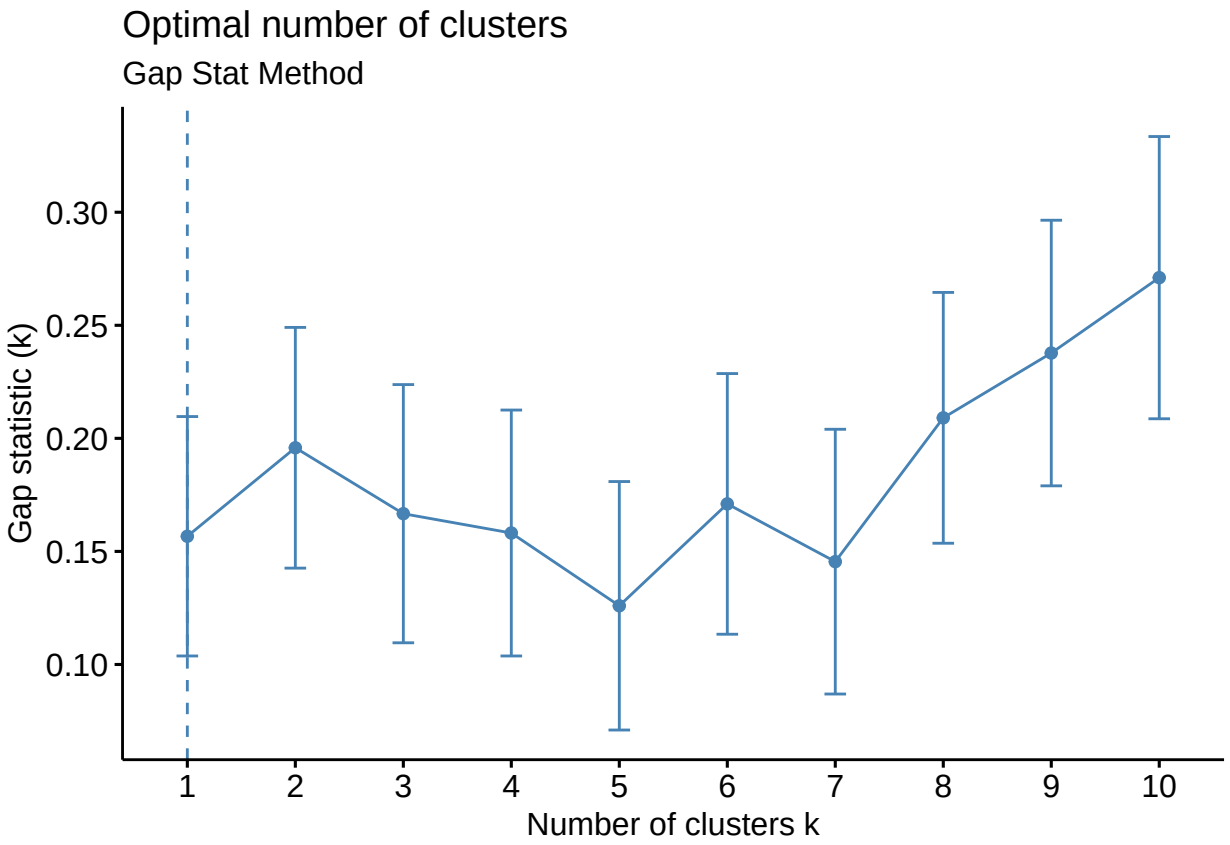


#Looking at the above graph we can see that there is an elbow at 2, however it is still unclear due to less sharpness in the graphical representation.

```
#Using the Silhouette method below
fviz_nbclust(ph2,kmeans,method="silhouette")
```




```
fviz_nbclust(ph2, kmeans, method = "gap_stat") + labs(subtitle = "Gap Stat Method")
```



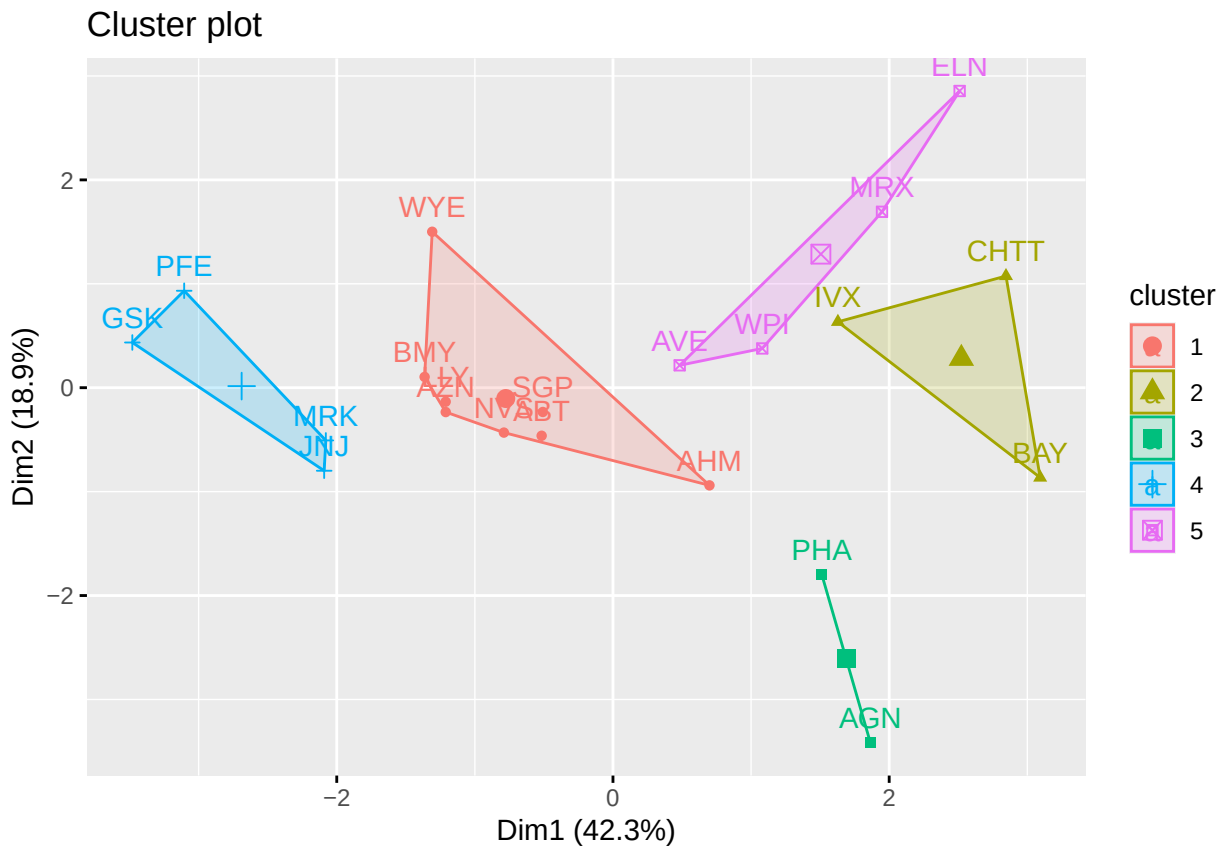
#We will use the Silhouette method becuase of the clear representation of K=5 #Final analysis and Extracting results using 5 clusters and Visualize the results

```
set.seed(123)
fl<- kmeans(ph2, 5, nstart = 25)
print(fl)
```

```
## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    3    1    1    5    2    1    2    5    1    4    2    4    5    4    1
```

```
##  PFE  PHA  SGP  WPI  WYE
##    4    3    1    5    1
##
## Within cluster sum of squares by cluster:
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(fl, data = ph2)
```



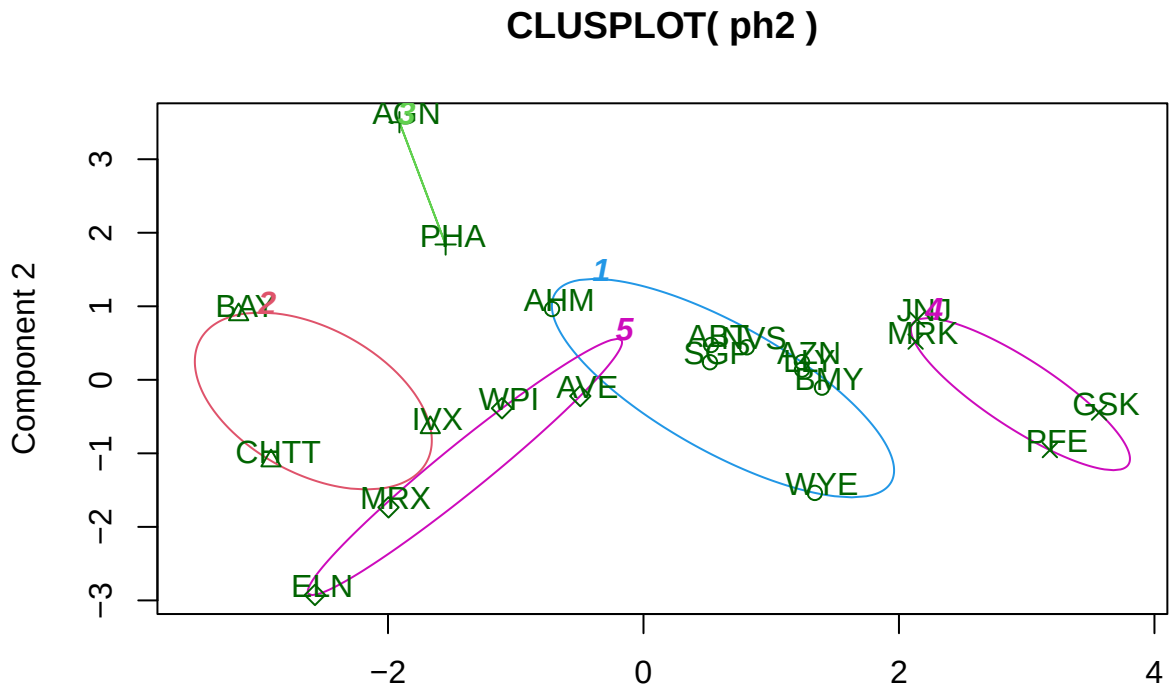
2.Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

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

```
#Interpreting the clusters with respect to the numerical variables used in forming the clusters
Ph%>%
  mutate(Cluster = fl$cluster) %>%
  group_by(Cluster)%>% summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 2       2       6.64 0.87      24.6  16.5  4.17          0.6      1.65
## 3       3      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 4       4     157.   0.48      22.2  44.4 17.7           0.95     0.22
## 5       5      13.1  0.598     17.7  14.6  6.2           0.425    0.635
## # i 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>
```

```
clusplot(ph2,fl$cluster, color = TRUE, labels = 2,lines = 0)
```



Cluster 1- AHM,SGP,WYE,BMY,AZN, ABT, NVS, LLY - This group has a high net profit margin and the lowest revenue growth. These companies have relatively low leverage and low revenue growth. They have the highest net profit margin and high return on equity, which means they have many products that bring them high profit . So they need not to consume much of their assets. These companies don't need to borrow money from the capital market, which makes their leverage low.

Cluster 2 - BAY, CHTT, IVX - This cluster has high beta and high leverage, but low ROA, revenue growth, and net profit margin. These companies represent innovative startups in the industry. They are relatively small from the market capitalization point of view, and their name is not known by people compared with those well-known brands. They have poor net profit margins and slow revenue growth since they are young, unproven businesses without lucrative products that will generate cash flow. As they highly rely on R&D, they have a high level of leverage and a low ROA. They are investing in the future, which indicates a high beta, so their price will rise in a rising market.

Cluster 3 - AGN, PHA - This cluster contains only 2 companies: AGN and PHA. It has the highest P/E ratio, lowest beta, low ROA, and net profit margin.As a result, these companies have high hopes for the

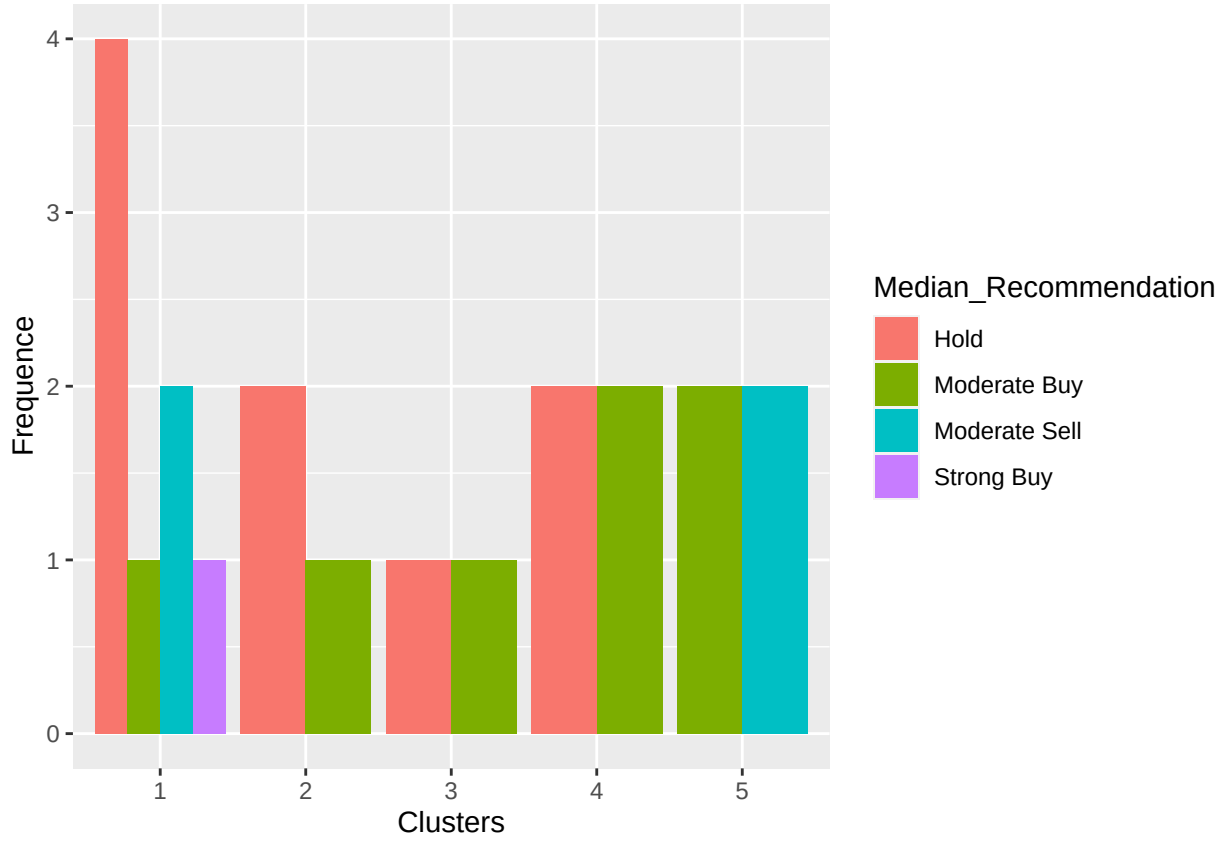
future while having little net profit in the past. Since they might spend a lot of money on D&I in cutting-edge technologies, the market values them highly. However, with its high price, investors get more risk.

Cluster 4 - JNJ, MRK, PFE,GSK -This group has the highest market capitalization, high ROE and ROA, high net profit margin, high asset turnover, and low leverage. With the largest market capitalizations and the most prominent positions in the industry, these companies stand for the leaders in this sector. These companies use capital exceptionally effectively, as seen by their high ROE, ROA, asset turnover, and lowest leverage values. They stand to gain the most from every dollar invested in these businesses. They must have a few best-selling and dominant products in the market, as well as mature products that require little capital or asset investment from the companies but generate large revenue and strong net profit margins—Pfizer is one example of this.

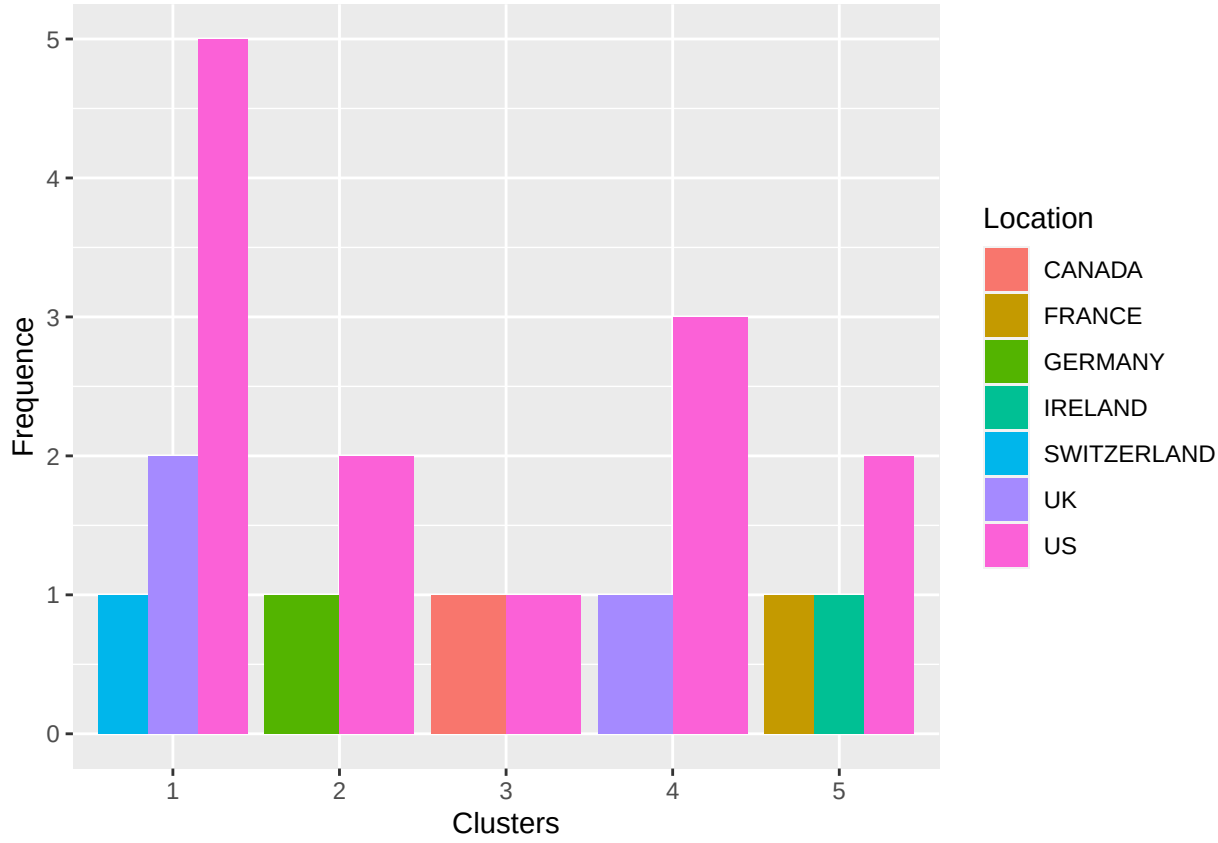
Cluster 5 - WPI, MRX,ELN,AVE - This cluster has high revenue growth, high Beta and low market capitalization, low P/E, and low turnover rate. These traditional, small-sized businesses have low ROE, ROA, and turnover rates, which suggests that they don't have particularly strong capital utilization skills. Nonetheless, given the strong rate of revenue growth, we can assume that the companies are being guided in the right direction by either internal reformation or external market changes. Additionally, we can infer that their share price is still cheap based on the lowest P/E.

Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

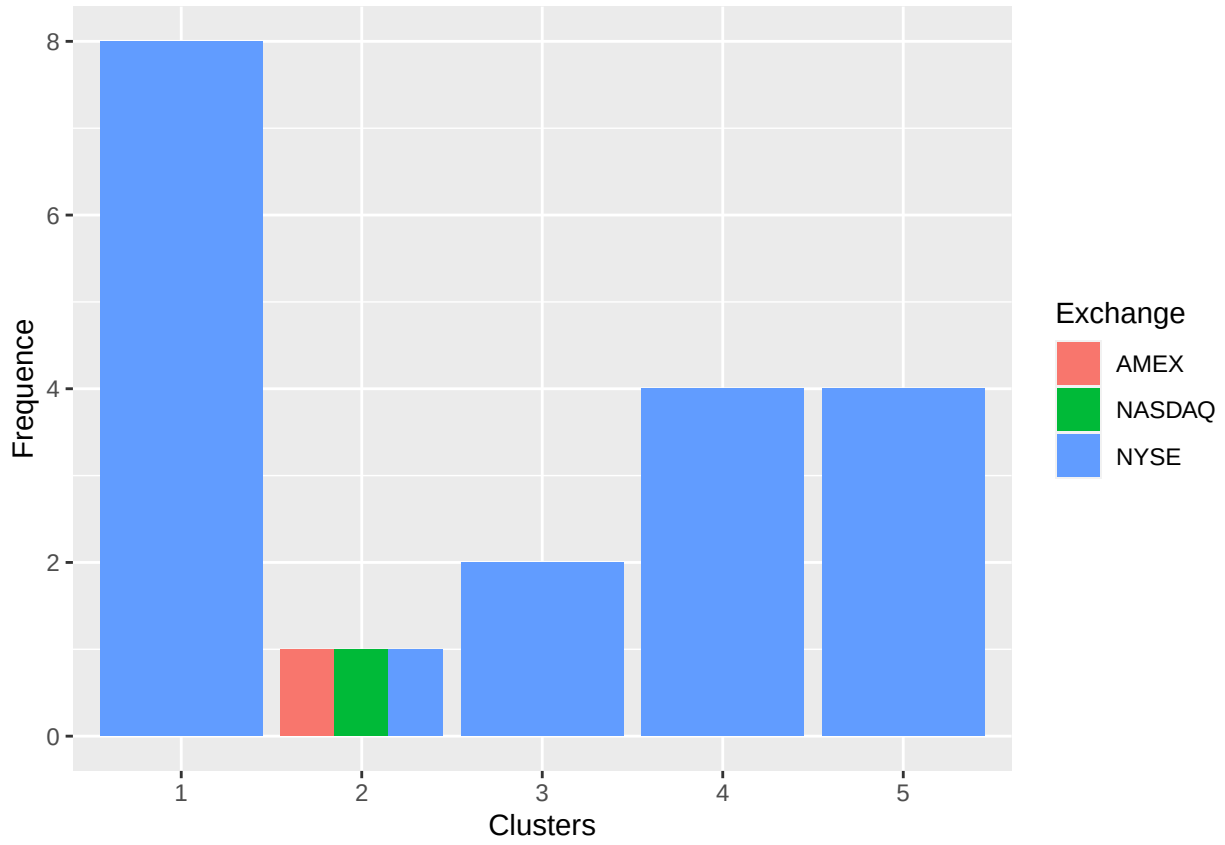
```
pattern_clue <- dataset[12:14] %>% mutate(Clusters=fl$cluster)
ggplot(pattern_clue, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='d
```



```
ggplot(pattern_clue, mapping = aes(factor(Clusters),fill = Location))+
  geom_bar(position = 'dodge')+labs(x ='Clusters',y = 'Frequence')
```




```
ggplot(pattern_clue, mapping = aes(factor(Clusters),fill = Exchange))+geom_bar(position = 'dodge')+
  labs(x ='Clusters',y = 'Frequence')
```



Cluster 1:

Median Recommendation: Cluster 1 is a very strong hold.

Location: Cluster 1 has three locations, with the United States outnumbering the United Kingdom and Switzerland.

Exchange: Cluster 1 has only one exchange, the NYSE, which has a large number of participants.

Cluster 2:

Median Recommendation: Cluster 2 has a strong hold rating and a low buy rating.

Location: Cluster 2 has two locations where the US ranks higher than Germany.

Exchange: Cluster 2 is home to three exchanges (AMEX, NASDAQ, and NYSE), which are all evenly distributed.

Cluster 3:

Median Recommendation: Cluster 3 has a low hold and a low buy, according to the median recommendation.

Location: Cluster 3 has only two locations (the United States and Canada) that are evenly distributed.

Exchange: Cluster 3 only has one exchange, which is the NYSE.

Cluster 4:

Median Recommendation: Cluster 4 has a high hold and a high buy, according to the median recommendation.

Location: Cluster 4 has two locations, with the US outnumbering the UK by a large margin.

Exchange: Cluster 4 only has one exchange, which is NYSE.

Cluster 5:

Median Recommendation: Cluster 5 has a moderate buy and moderate sell recommendation.

Location: Cluster 5 has three locations, the most prominent of which is the United States.

Exchange: Cluster 5 only has one exchange, which is the NYSE.

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

APPROPRIATE NAME :

cluster 1 - Elevated hold cluster

cluster 2 - Hold cluster

cluster 3 - Cheapest cluster

cluster 4 - Purchase hold cluster

cluster 5 - Purchase sell cluster