Assignemnt_4_Clustering

Sudarshan.Rayapati

2023-11-13

SUMMARY

Hear I am performing a non-various leveled bunch examination using the k-implies grouping strategy.Goal 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.

Reason for Using Market Capitalization, Beta, PE Ratio, ROE, ROA, Leverage, Revenue Growth, and Net Profit Margin The variables that were chosen are common financial measures that are used in business performance evaluating and comparing. Market Cap, Beta, PE Ratio, ROE, ROA, Turnover of Assets, Leverage, Rev Growth, and Net Profit Margin are a few of them. Once taken as a whole, these variables provide a full picture of a company's productivity, profitability, and stability.

1. Market Capitalization: This statistic, which ranges from 0.41 to 199.47, shows the total market value of a company's shares.
2. Beta: a number that ranges from 0.18 to 1.11 that shows how sensitive a company's stock price is to changes in the market overall.
3. PE Ratio: This number, which varies from 3.6 to 82.5, shows the multiple of a company's profits per share and aids in assessing the stock's relative value.
4. ROI: This number, which ranges from 3.9 to 62.9, measures how well a company uses shareholder equity to generate profit.
5. Return on Assets (ROA): This statistic, which ranges from 0.3 to 1.1, shows the profitable a company's assets are.
6. Asset Turnover: this number, which ranges from 0.5 to 1.1, shows the way a company uses its assets to generate revenue.
7. Leverage: a number between 0 and 3.51 that indicates how much a business depends on debt to fund its operations. Rev_Growth: this variable, which ranges from -3.17 to 34.21, indicates the percentage change in a company's revenue during a certain time period.
8. Net Profit Margin: This number, which ranges from 2.6 to 25.54, indicates the percentage of the profits of a company that turns into profit.

We examine the connection present between clusters and variables 10 to 12.Bar plots are applied to show the frequency distribution of non-clustered variables within each cluster.Below the graph, using the bar graph, are the necessary labels and an explanation.

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)
```

```
## -- 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   3.4.4     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 error
```

```
library(cluster)
library(factoextra)
```

```
## 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
```

```r
library(flexclust)
```

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

```r
library(dbscan)
```

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

#importing the dataset and reading the dataset

```r
data <- read.csv("/Users/sudarshan/Desktop/FML/dataset/Pharmaceuticals.csv")
head(data)
```

```
##   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
phramdata <- na.omit(data)
#Collecting numerical variables from column 1 to 9 to cluster 21 firms
row.names(phramdata)<- phramdata[,1]
Phram<- phramdata[, 3:11]
head(Phram)
```

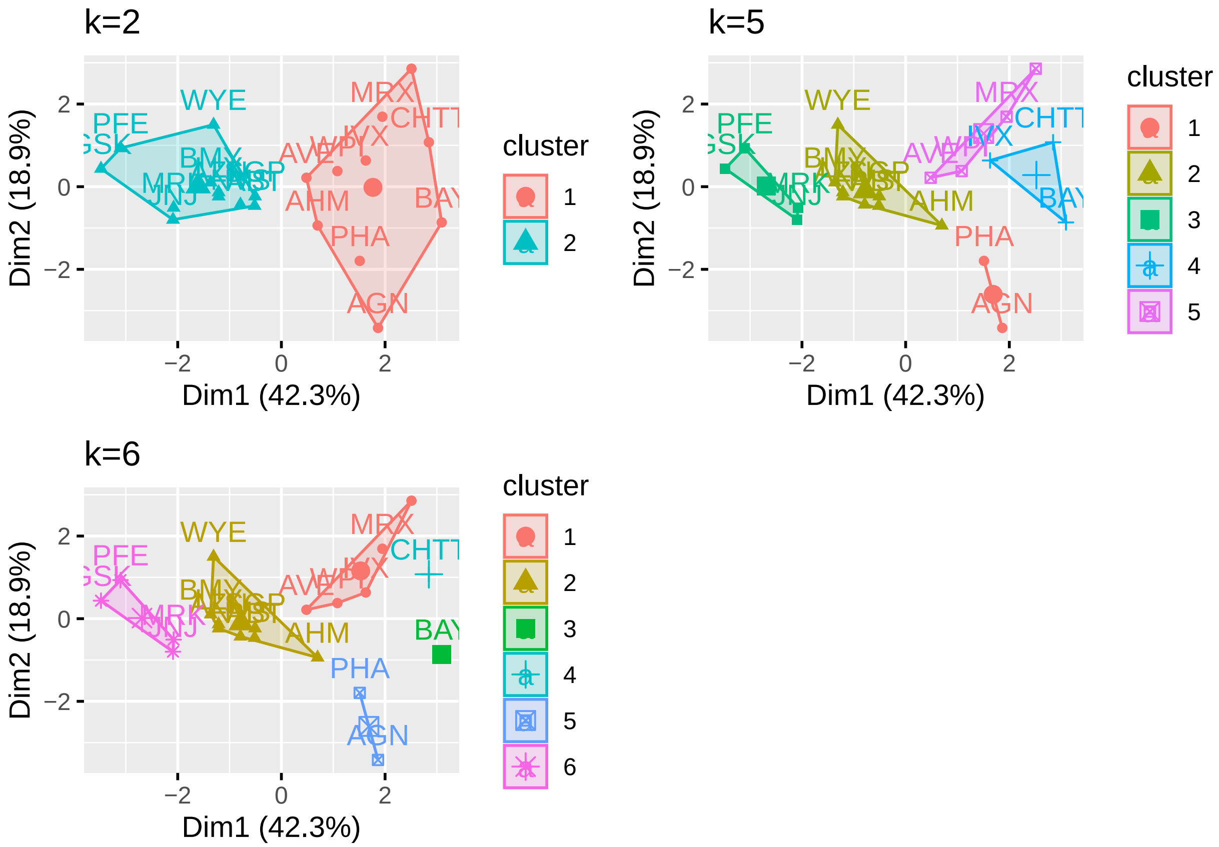
```
##     Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## ABT      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## AGN       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## AHM       6.30 0.46     20.7 14.9  7.8            0.9     0.27       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
phram2<- scale(Phram)
head(phram2)
```

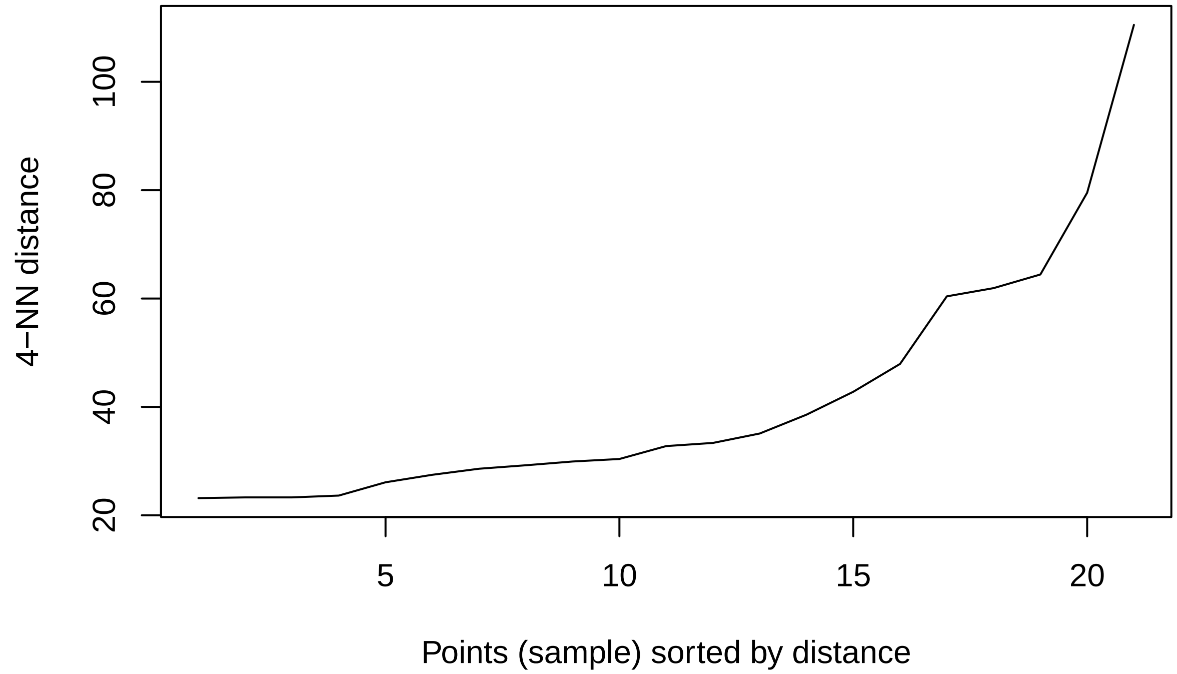
```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
##       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
```

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

```r
kmeans_ss <- kmeans(phram2, centers = 2, nstart = 30)
kmeans_ss1<- kmeans(phram2, centers = 5, nstart = 30)
kmeans_ss2<- kmeans(phram2, centers = 6, nstart = 30)
Plott12<-fviz_cluster(kmeans_ss, data = phram2)+ggtitle("k=2")
plott22<-fviz_cluster(kmeans_ss1, data = phram2)+ggtitle("k=5")
plott33<-fviz_cluster(kmeans_ss2, data = phram2)+ggtitle("k=6")
grid.arrange(Plott12,plott22,plott33, nrow = 2)
```



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

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



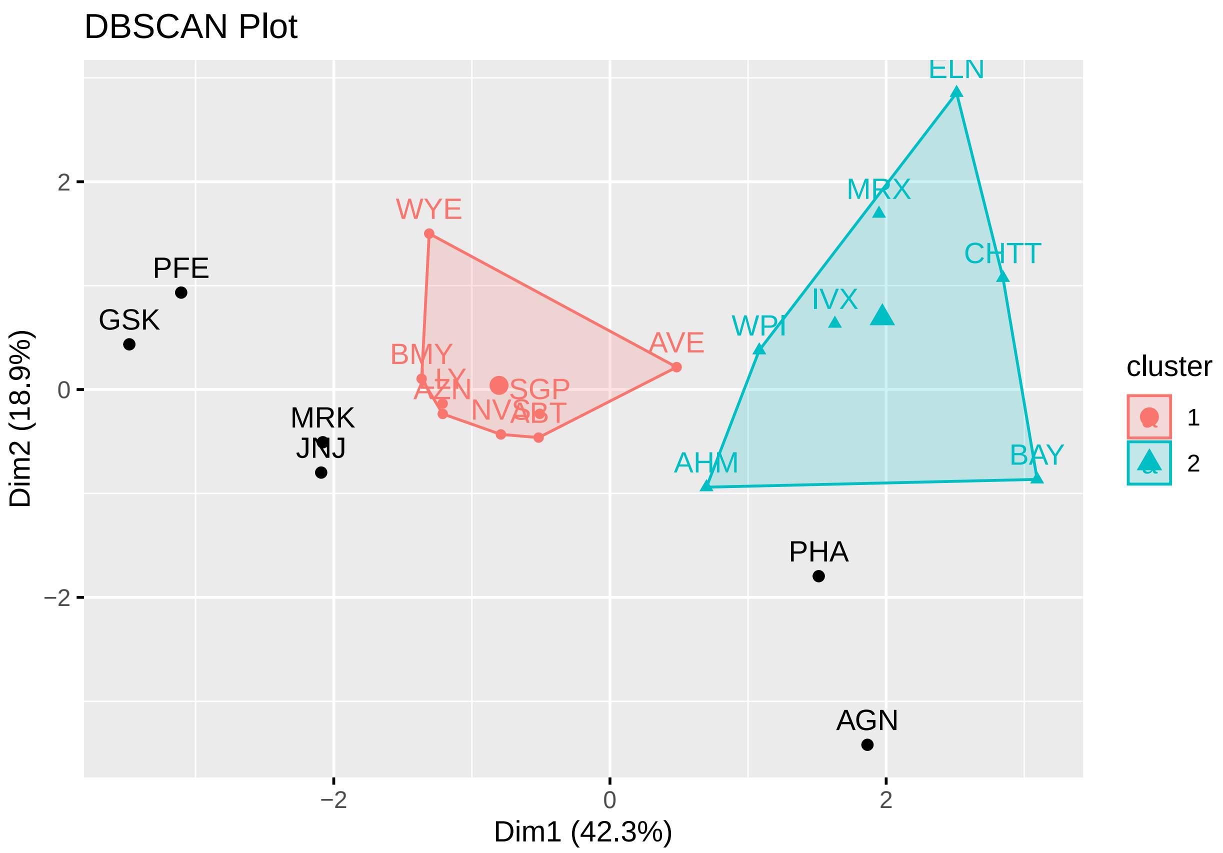
```
# DBSCAN Algorithm at eps=30 and minpts =4
db <- dbscan::dbscan(Phram, eps = 30, minPts = 4)

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

```
## 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(db$cluster)
```

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

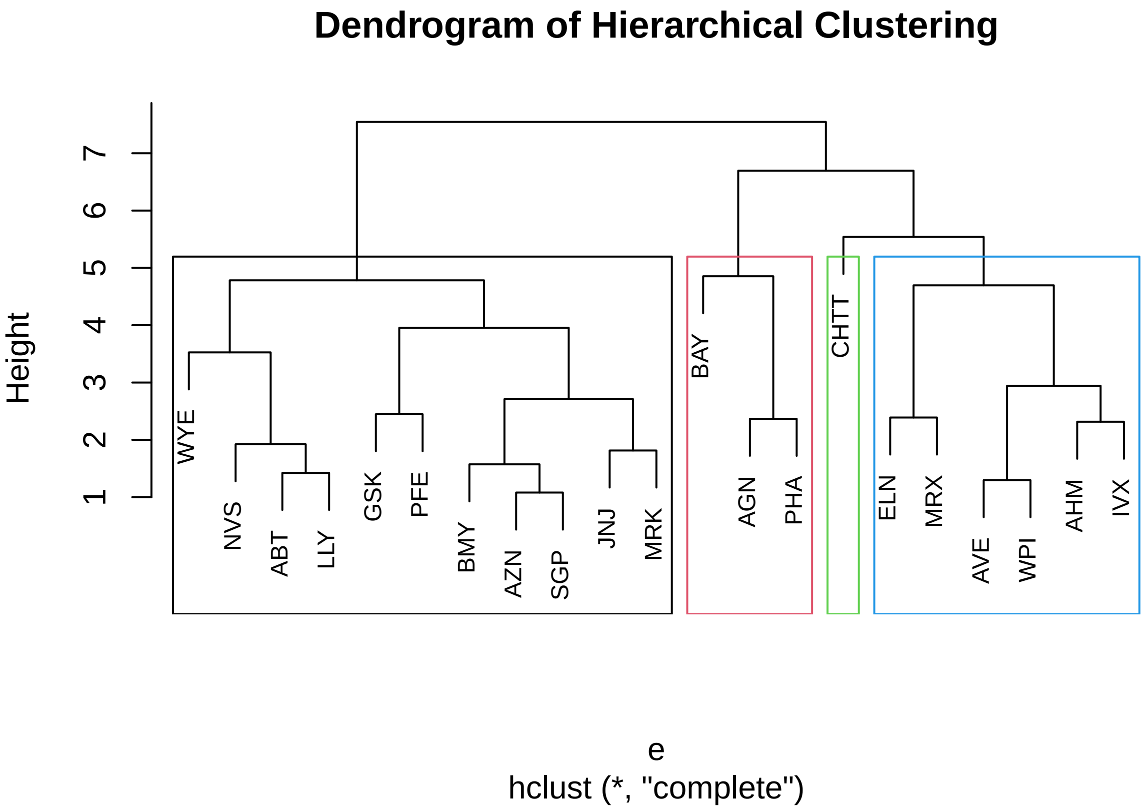
```
# Visualization of clusters
fviz_cluster(db, Phram) + ggtitle("DBSCAN Plot")
```



#K-means is widely used in exploratory data analysis to find patterns and groupings in the data, and I chose it over DBSCAN because K-means clustering provides details about the financial profiles of pharmaceutical companies. DBSCAN may be useful in showing groups of companies having comparable financial characteristics, and these can aid in strategic decision-making and investment analysis for datasets with large regions. It is also simple to figure out. The K-means algorithm needs a certain amount of clusters, k. Because the user can decide how many clusters to build, this might be useful in some situations. The amount of clusters may not be simply identified via DBSCAN or hierarchical clustering.

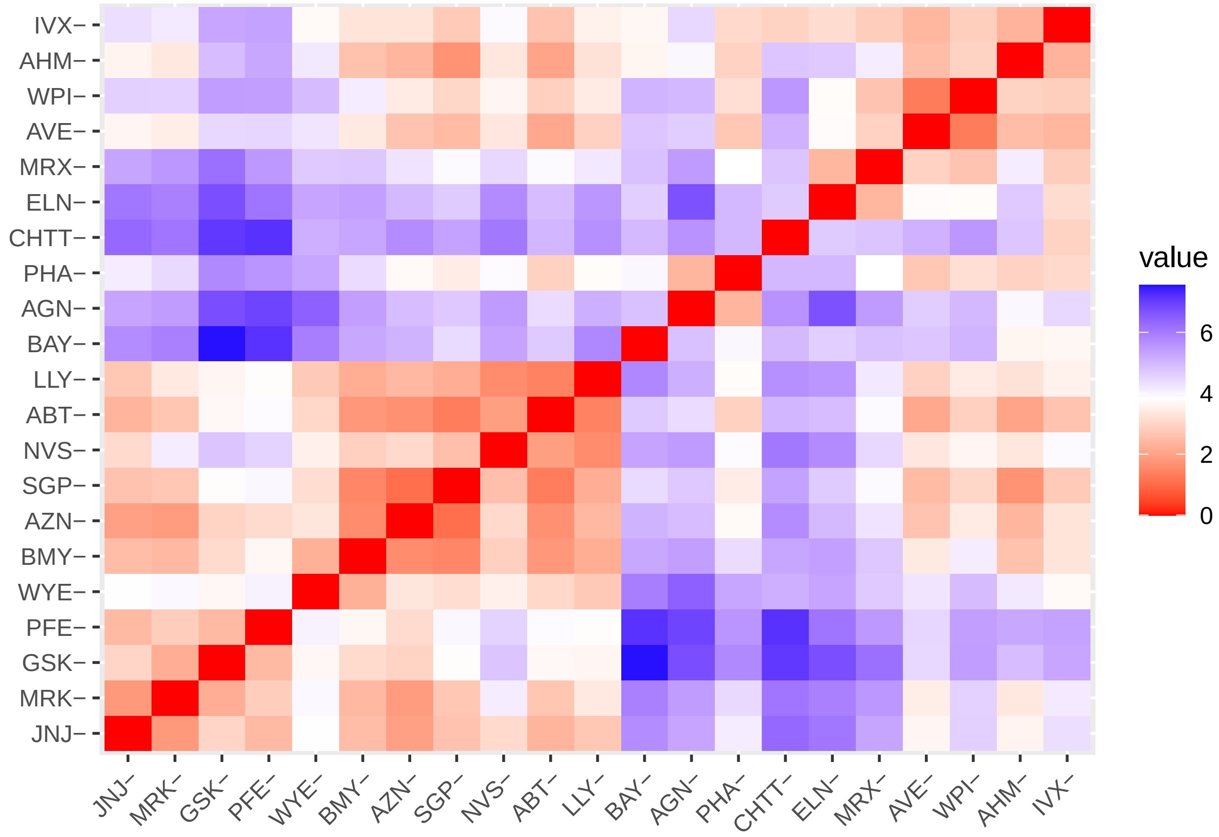
Hierarchical Clustering

```
# Geting euclidean distance for the data
e <- dist(phram2, method = "euclidean")

# Hierarchical Clustering
hhh <- hclust(e, method = "complete")

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



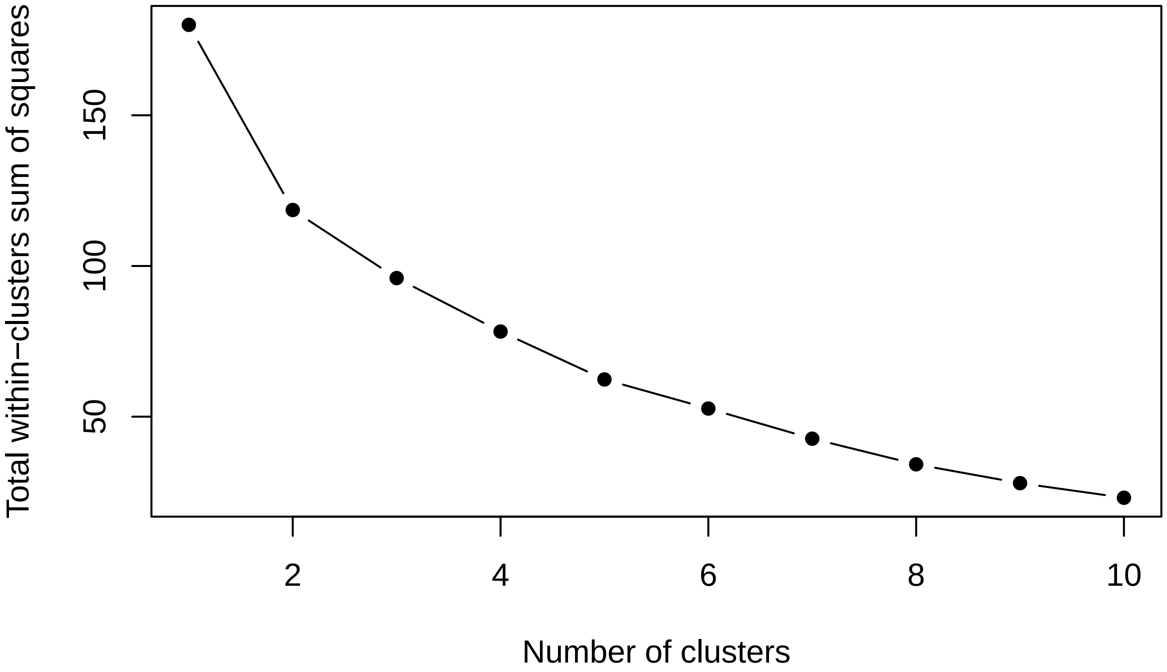
#Determining optimal clusters using Elbow method

```
ds <-  dist(phram2, method = "euclidean")
fviz_dist(ds)
```



#Total within-cluster sum of squares (wss) for each k is tot.withinss, which is the total within-cluster sum of squares. Measure and plot the wss for k = 1 to k = 10 and extract a wss for clusters 2–15. It is commonly accepted that a location of a bend, or knee, in the plot shows that k = 5 is the correct number of clusters.

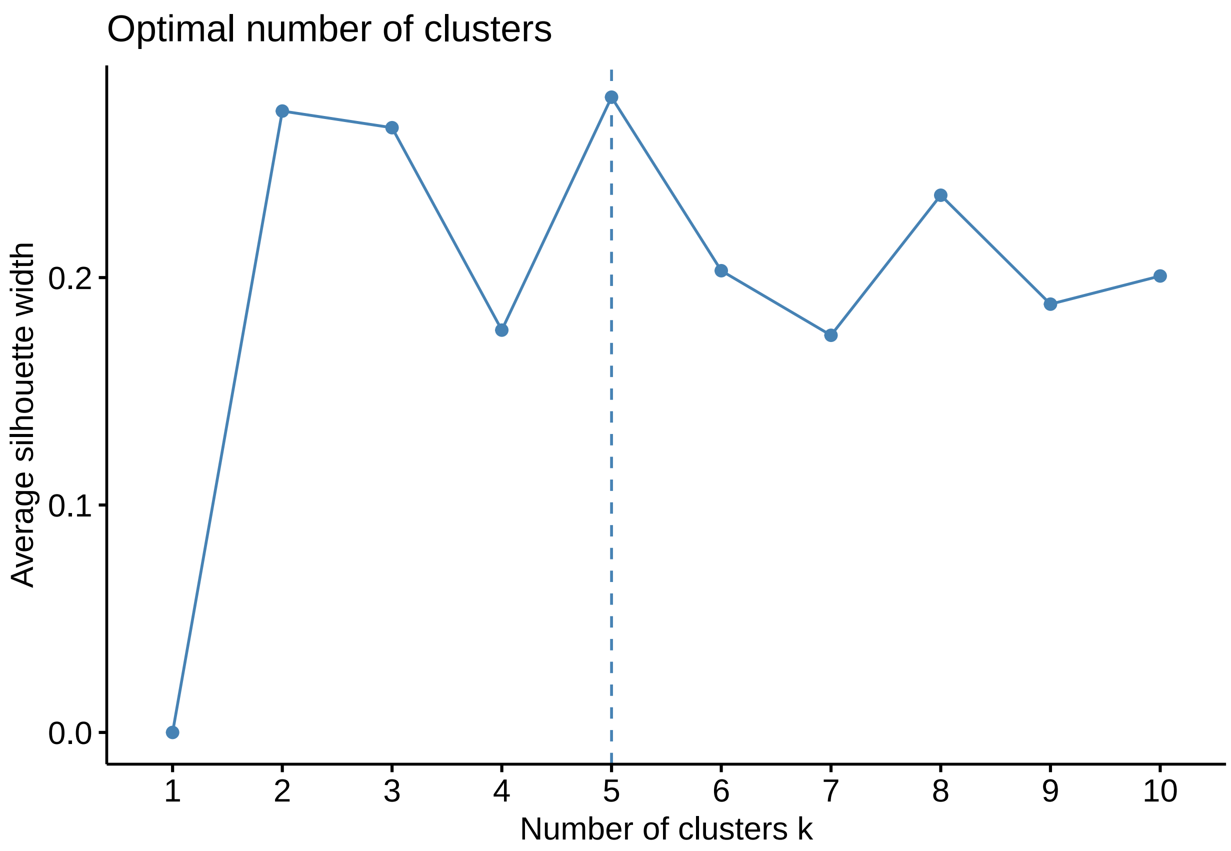
```
set.seed(123)
wss<- function(k){
kmeans(phram2, k, nstart =10)$tot.withinss
}
k.values<- 1:10
wss_cluster<- map_dbl(k.values, wss)
plot(k.values, wss_cluster,
     type="b", pch = 16, frame = TRUE,
     xlab="Number of clusters",
     ylab="Total within-clusters sum of squares")
```



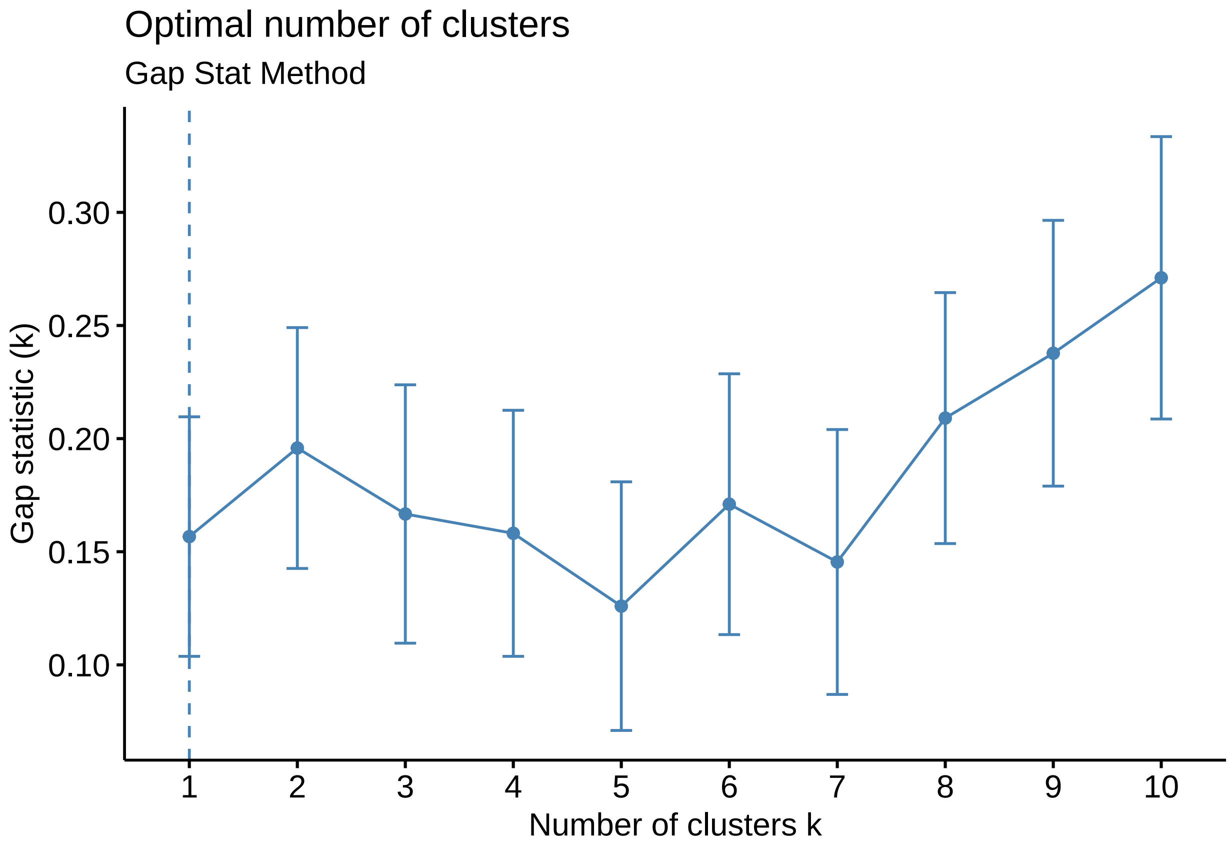
#The elbow at number two is visible in the graph above, but because the graphical representation is less sharp, it is still confusing.

#Using Silhouette & Gap stat methods

```
fviz_nbclust(phram2,kmeans,method="silhouette")
```



```
fviz_nbclust(phram2, kmeans, method = "gap_stat") + labs(subtitle = "Gap Stat Method")
```

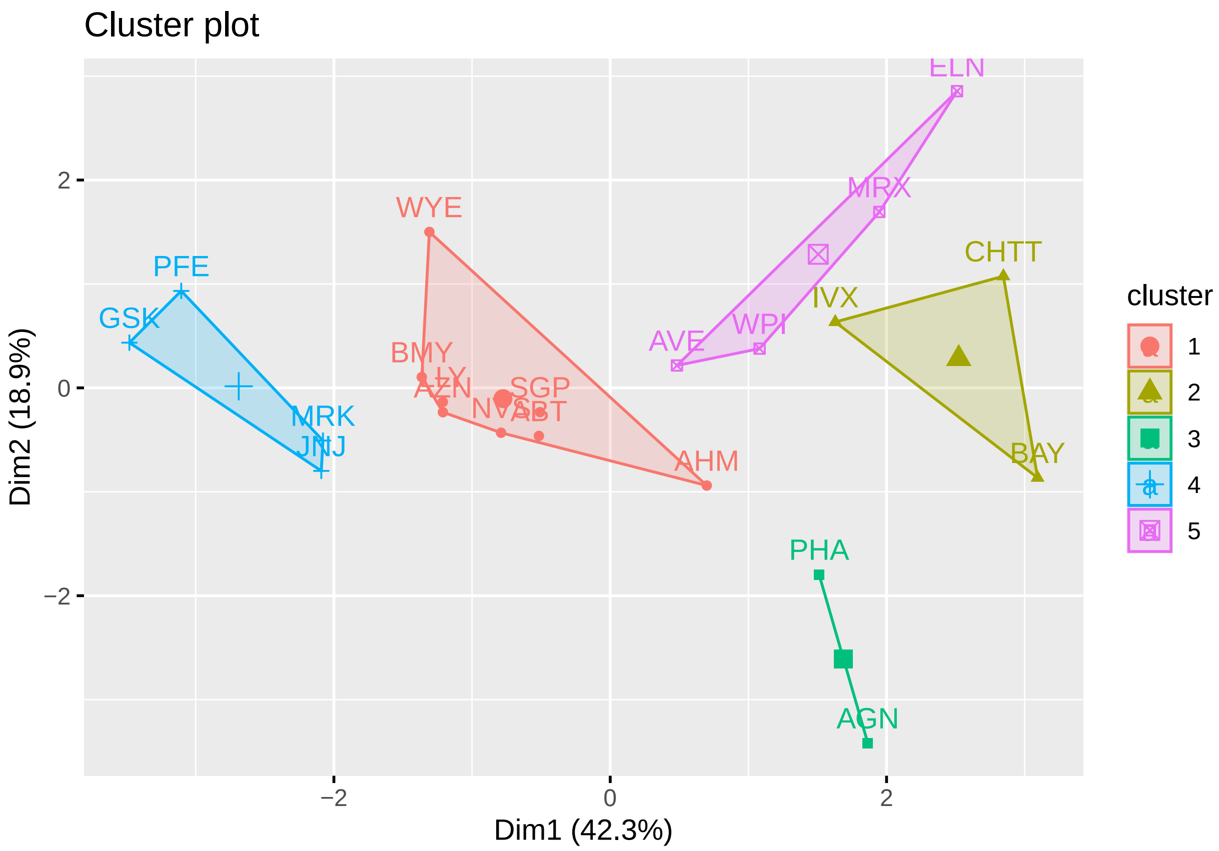


#The Silhouette approach will be used because of its clear picture of K=5. #Complete analysis, result extraction utilizing five clusters, and result visualization

```r
set.seed(123)
flan<- kmeans(phram2, 5, nstart = 25)
print(flan)
```

```
## 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(flan, data = phram2)
```



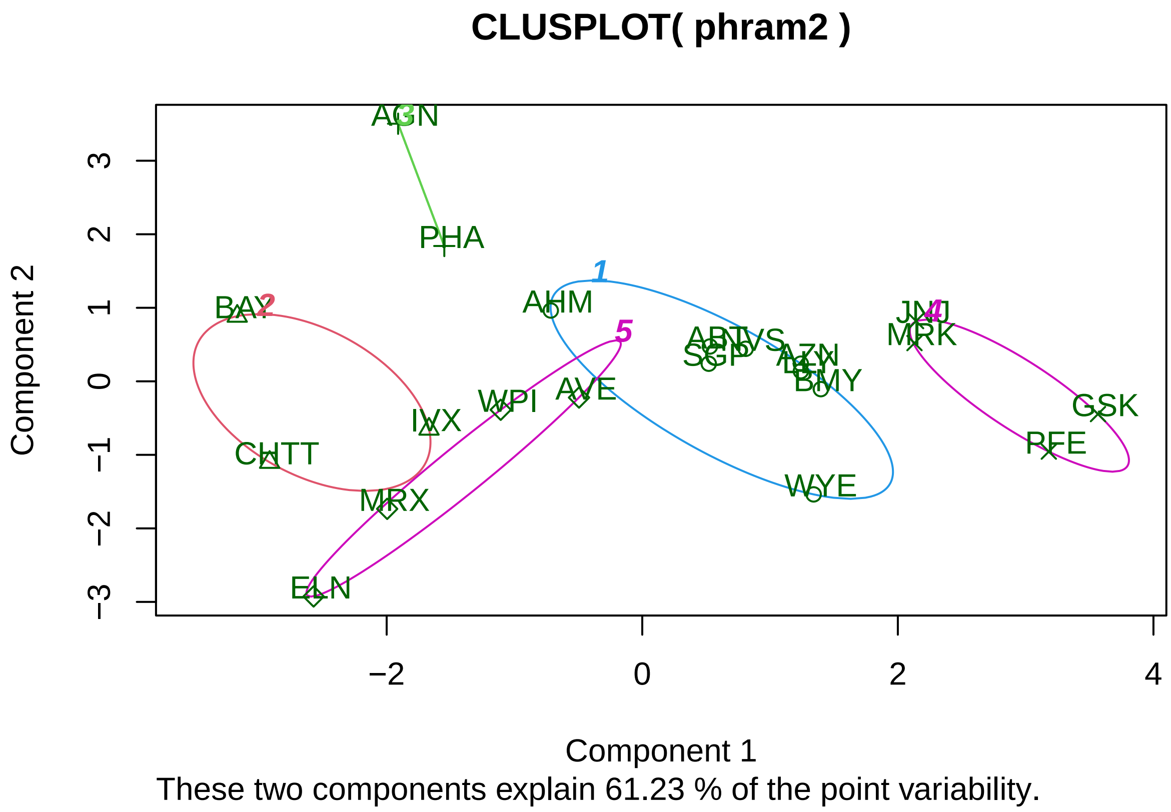
## 2.Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

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

```
Phram%>%
  mutate(Cluster = flan$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(phram2,flan$cluster, color = TRUE, labels = 2,lines = 0)
```



Cluster 1- AHM,SGP,WYE,BMY,AZN, ABT, NVS, LLY - This group has the lowest rate of revenue growth and the largest net profit margin. These companies aren't highly capitalized and have low revenue growth. Because of their large number of successful merchandise, they have the largest net profit margin and return on equity. Therefore, they don't have to use up all of their resource. Since they are unable to borrow money from the capital market, these companies had low leverage.

Cluster 2 - BAY, CHTT, IVX - This cluster has high leverage and high beta, but low ROA, net profit margin, and sales growth. These businesses are examples of creative startups in the sector. They are little in terms of market capitalization, and their names are not as well-known as those of well-known brands. Their revenue growth is modest and their net profit margins are low due to their recent establishment, lack of experience, and lack of profitable products that can produce cash flow. They have a low ROA and a high degree of leverage since they heavily rely on R&D. Since they have a high beta and are investing in the future, their price will increase in a rising market.

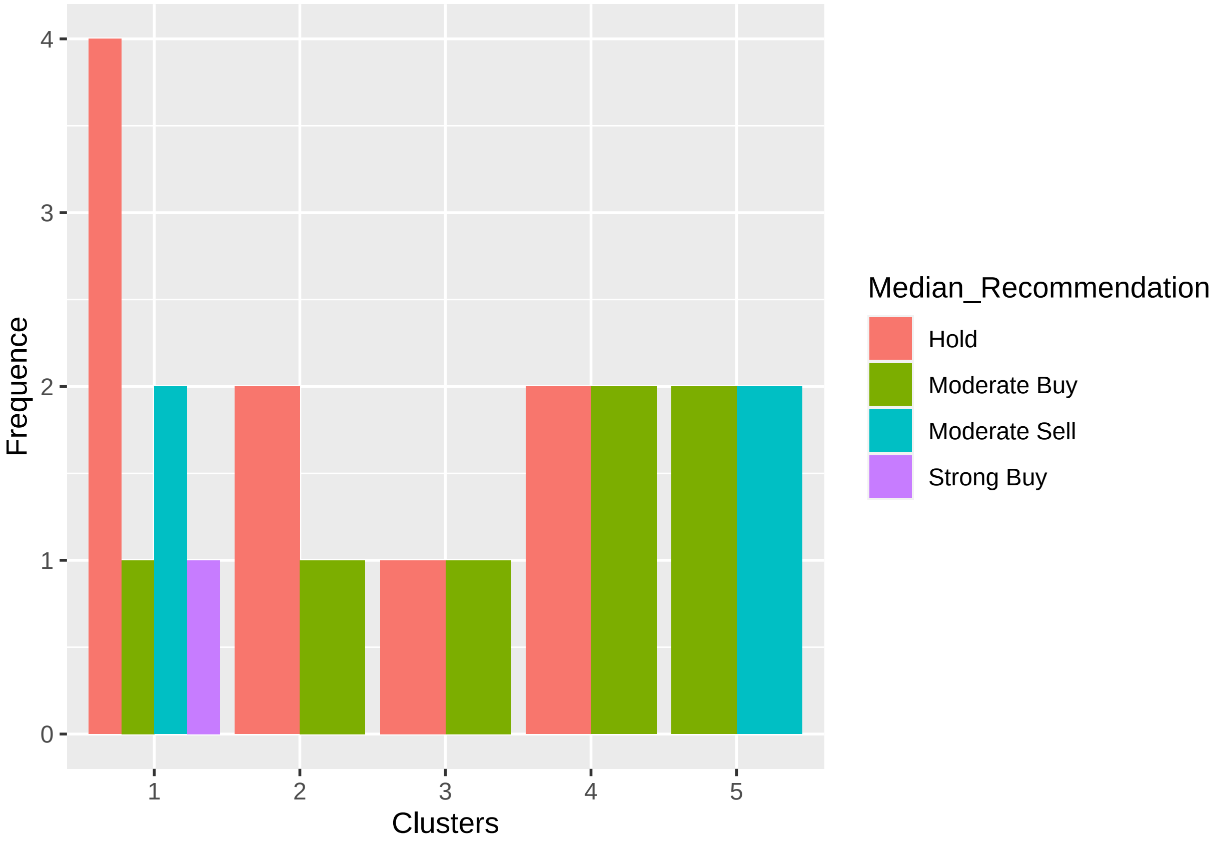
Cluster 3 - AGN, PHA - There are just two businesses in this cluster: AGN and PHA. Its net profit margin, low ROA, lowest beta, and highest P/E ratio are all present.Consequently, many businesses had modest net profits in the past but great expectations for the future. The market values them highly since they may invest a significant amount of money in D&I in cutting-edge technologies. Nevertheless, investors bear greater risk due to its high price.

Cluster 4 - JNJ, MRK, PFE,GSK -With respect to market capitalization, ROE and ROA, net profit margin, asset turnover, and leverage, this group has the greatest values. These companies represent the industry leaders with their greatest market capitalizations and significant positions. The remarkable capital usage of these organizations is demonstrated by their low leverage values, high ROE, ROA, and asset turnover. From each dollar invested in these companies, they stand to benefit the most. They must have a small number of top-selling, market-dominating products in addition to established products that demand little in the way of capital or assets from the businesses but bring in big sums of money and have healthy net profit margins.

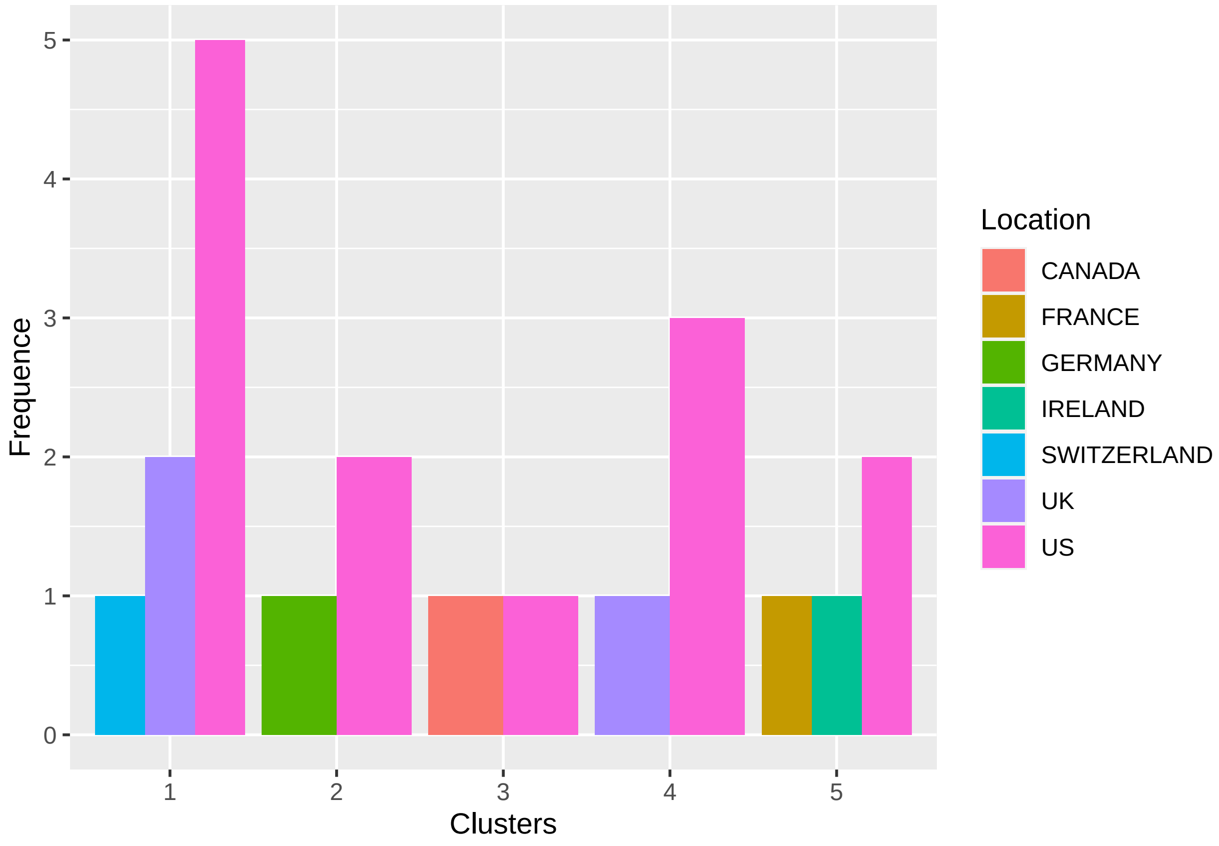
Cluster 5 - WPI, MRX,ELN,AVE - This cluster has low market capitalization, low P/E, low turnover rate, high beta, and strong revenue growth. Low ROE, ROA, and turnover rates indicate that these conventional small-sized enterprises may not have very strong capital usage capabilities. However, we can infer that the organizations are being led in the correct direction by either external market developments or internal reformation given the robust rate of revenue growth. The lowest P/E further suggests that their share price is still modest.

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

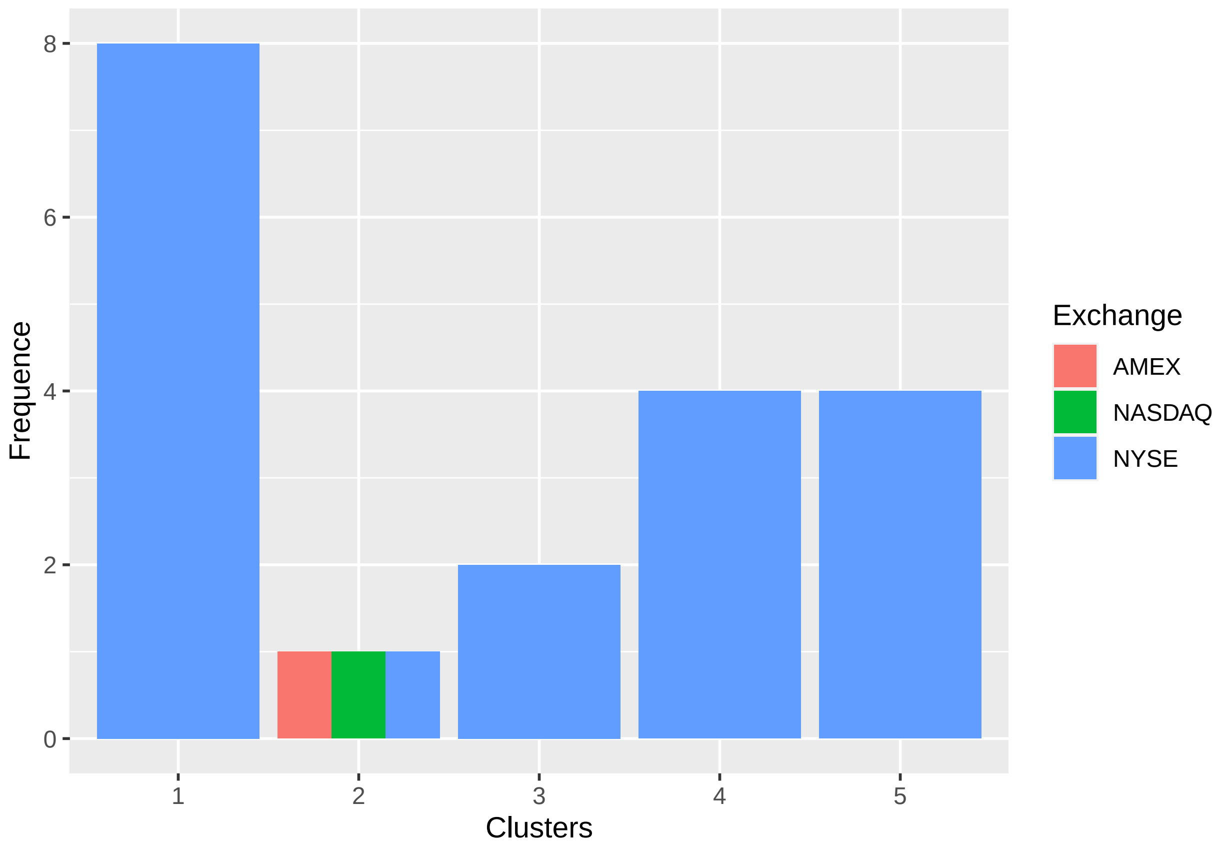
```
p <- data[12:14] %>% mutate(Clusters=flan$cluster)
ggplot(p, mapping = aes(factor(Clusters), fill =Median_Recommendation))+geom_bar(position='dodge')+labs
```



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



```
ggplot(p, 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 it has 3 locations US,UK and Switzerland and had only one exchange the NYSE

Cluster 2:

Median Recommendation: Cluster 2 has a strong hold rating and a low buy rating it was in 2 locations us gets more than germany and had (AMEX, NASDAQ, and NYSE)

Cluster 3:

Median Recommendation: Cluster 3 has a low hold and a low buy, according to the median it was in 2 locations US and canada and had only one exchnage nyse

Cluster 4:

Median Recommendation: Cluster 4 has a high hold and a high buy, according to the it was in 2 locations US and UK with one stock exchange in NYSE

Cluster 5:

Median Recommendation: Cluster 5 has a moderate buy and moderate sell and prominent in US and has one stock exchnage in NYSE

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

cluster 1 - Higher hold cluster cluster 2 - Hold cluster cluster 3 - Lowest cluster cluster 4 - Buy hold cluster cluster 5 - Buy sell cluster