Assignment 4

Jayasri

11/07/2021

K-Means Algorithm (CLUSTERING)

The objective of this assignment is to use **K-Means for Clustering** the Pharmaceutical data

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) - Total market value of company's outstanding shares of stock.
2. **Beta** - Volatility of the security in regards to Market movement.
3. **Price/earnings ratio** - Higher Ratio could conclude either security is overvalued or expect to have future profits.
4. **Return on equity** - Ratio of Revenues, excluding all dividends, to the number of equities. Higher the ROE, better the company is performing.
5. **Return on assets** - an indicator of how profitable a company is relative to its total assets (Net Income / Total Assets).
6. **Asset turnover** - It means how efficiently the company is using it's assets to produce the asset, Higher the better.
7. **Leverage** - Ratio of Debt to Equity. Lesser the better.
8. **Estimated revenue growth** - Revenue growth is the increase (or decrease) in a company's sales from one period to the next.
9. **Net profit margin** - Ratio of net profit to the revenue. Higher the better.
10. **Median recommendation** (across major brokerages) -The mean or median recommendation that analysts make on a stock. The consensus recommendation is calculated simply by compiling recommendations and taking the average or median.
11. **Location of firm's headquarters** - Location of the firm.
12. **Stock exchange** on which the firm is listed

LOADING REQUIRED PACKAGES

```
# Loading the required packages and install packages if necessary
#install.packages("psych")
library(psych)
#install.packages("dplyr")
library(dplyr)
#install.packages("cowplot")
library(cowplot)
#install.packages("esquisse")
library(esquisse)
#install.packages("ggplot2")
library(ggplot2)
#install.packages("factoextra")
library(factoextra)
#install.packages("cluster")
library(cluster)
#install.packages("tidyverse")
library(tidyverse)
#install.packages("tibble")
library(tibble)
#install.packages("knitr")
library(knitr)
```

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.

Data Preparation

```
Pharmaceutical_data<- read.csv("Pharmaceuticals.csv")
# Extracted Numerical Variables only
df_Numerical <-Pharmaceutical_data[,c(3:11)]
row.names(df_Numerical) <- Pharmaceutical_data[,1]
# Pharamaceutical Data
kable(df_Numerical, caption= "Pharmaceutical Data")
```

Table 1: Pharmaceutical Data

	Market_Cap	Beta	PE_Ratio	ROE	ROA	Asset_Turnover	Leverage	Rev_Growth	Net_Profit_Margin
ABT	68.44	0.32	24.7	26.4	11.8	0.7	0.42	7.54	16.1
AGN	7.58	0.41	82.5	12.9	5.5	0.9	0.60	9.16	5.5
AHM	6.30	0.46	20.7	14.9	7.8	0.9	0.27	7.05	11.2
AZN	67.63	0.52	21.5	27.4	15.4	0.9	0.00	15.00	18.0
AVE	47.16	0.32	20.1	21.8	7.5	0.6	0.34	26.81	12.9
BAY	16.90	1.11	27.9	3.9	1.4	0.6	0.00	-3.17	2.6
BMY	51.33	0.50	13.9	34.8	15.1	0.9	0.57	2.70	20.6
CHTT	0.41	0.85	26.0	24.1	4.3	0.6	3.51	6.38	7.5
ELN	0.78	1.08	3.6	15.1	5.1	0.3	1.07	34.21	13.3
LLY	73.84	0.18	27.9	31.0	13.5	0.6	0.53	6.21	23.4
GSK	122.11	0.35	18.0	62.9	20.3	1.0	0.34	21.87	21.1
IVX	2.60	0.65	19.9	21.4	6.8	0.6	1.45	13.99	11.0
JNJ	173.93	0.46	28.4	28.6	16.3	0.9	0.10	9.37	17.9
MRX	1.20	0.75	28.6	11.2	5.4	0.3	0.93	30.37	21.3
MRK	132.56	0.46	18.9	40.6	15.0	1.1	0.28	17.35	14.1
NVS	96.65	0.19	21.6	17.9	11.2	0.5	0.06	-2.69	22.4
PFE	199.47	0.65	23.6	45.6	19.2	0.8	0.16	25.54	25.2
PHA	56.24	0.40	56.5	13.5	5.7	0.6	0.35	15.00	7.3
SGP	34.10	0.51	18.9	22.6	13.3	0.8	0.00	8.56	17.6
WPI	3.26	0.24	18.4	10.2	6.8	0.5	0.20	29.18	15.1
WYE	48.19	0.63	13.1	54.9	13.4	0.6	1.12	0.36	25.5

```
# Checking for NA Values
NA_Check<-colMeans(is.na(df_Numerical))
print(paste("There are no 'NA(Missing)' Values in the Dataset"))
```

[1] "There are no 'NA(Missing)' Values in the Dataset"

Data Normalization

```
# Data Normalization
df_Norm_Numerical <- scale(df_Numerical)
```

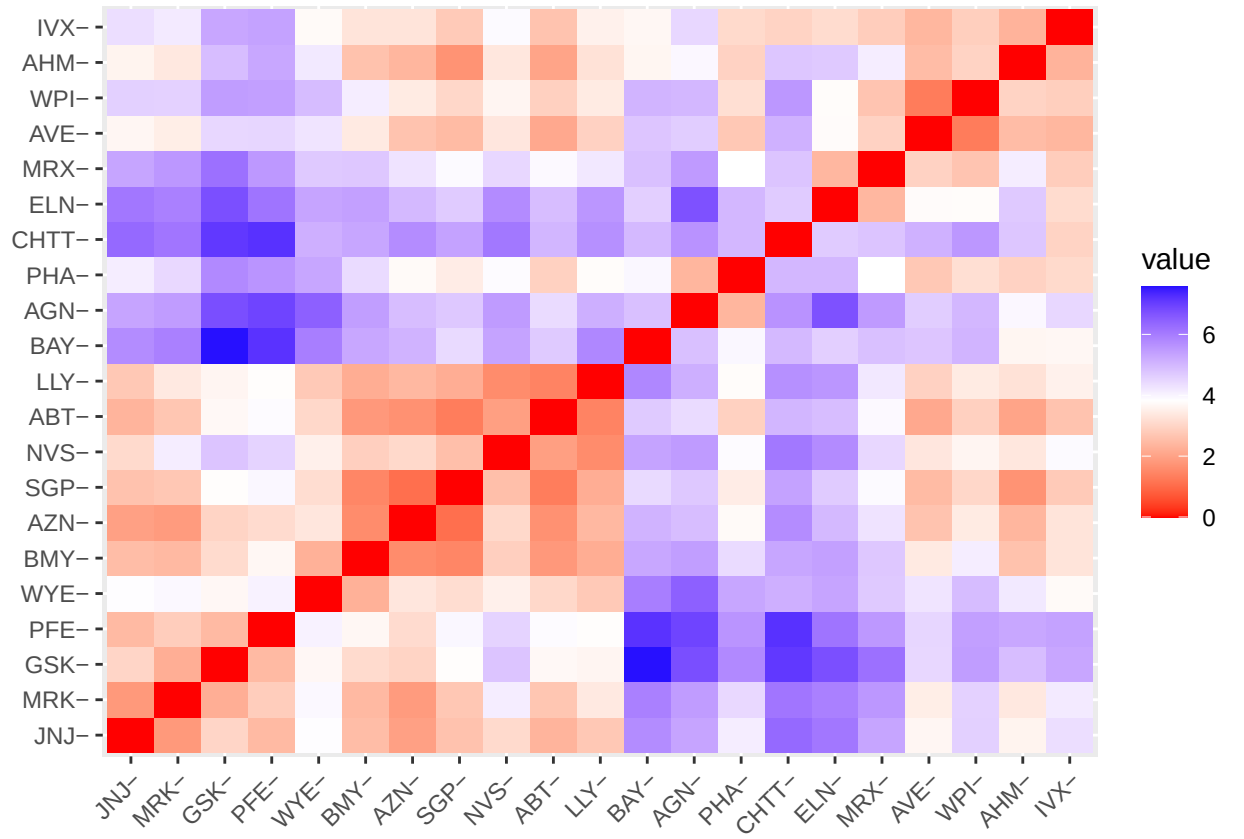
Distance Measure

For computing distance, we are going to use the **get_distance** function.It uses the Euclidean distance as default metric and **fviz_clust** is to visualize the distance matrix.

```
#Computing distance. Euclidean distance as a default metric
distance<- get_dist(df_Norm_Numerical)
head(distance)
```

```
## [1] 4.415575 2.018793 1.669541 2.111983 4.690231 1.805543
```

```
# Plotting the distance.
fviz_dist(distance)
```



Distance Matrix Plot

This plot shows the different intensity of color for different distances.As we can see , the diagonal has a value equal to zero because it indicates the distance of an observation from itself.The purple indicates that farthest distance between the point and red indicates the nearest distance.

Running the K-means Model

The main objective of the Clustering is: We group similar items into a cluster, with each cluster behind different from other clusters. That is, the variation among the items in a cluster should be small compared to the variation between the clusters.

- First we will select a random value for k and run the model. Here, I selected **k=2** as random number and run the k Model

```
set.seed(123)

# Running the K means model with Random number K=2
k2<- kmeans(df_Norm_Numerical,centers = 2,nstart = 25)

# To see the Results
#print(k2)

# To see the centers of the 2 clusters
k2$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
```

```
# To check the size of clusters
k2$size
```

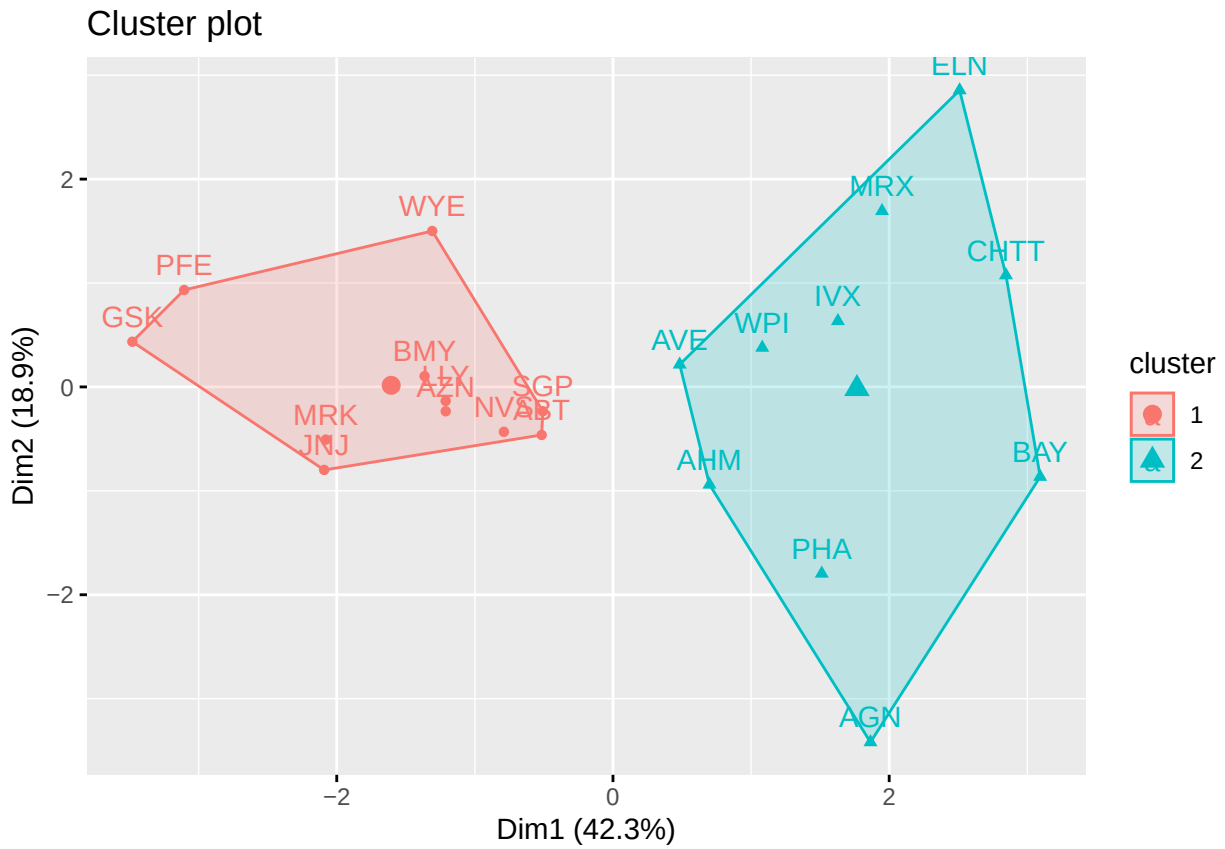
```
## [1] 11 10
```

```
# Identify the cluster of the 20th observation as Example
k2$cluster[20]
```

```
## WPI
##   2
```

Visualize When K=2

```
#Lets visualise the 2 clusters
fviz_cluster(k2,data = df_Norm_Numerical)
```



This plot gives us a detailed representation of the 2 clusters with the firms based on their Numerical variables. However, we chose the k value as random number without a substantial reason. So now we have to perform the data driven methods to chose our best k.

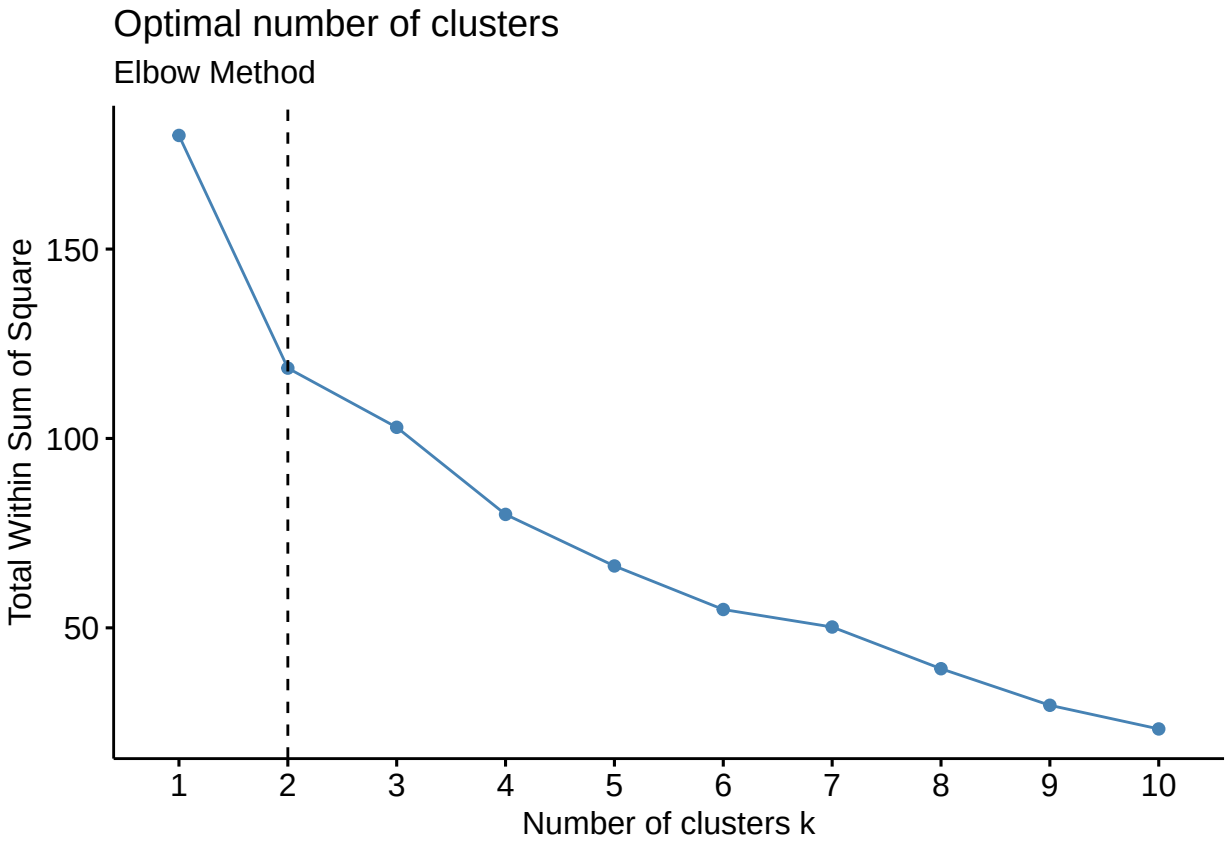
Choosing the best K

When there are no external considerations,the choice of K is done by using data driven methods. such as:

- **The Elbow Method**
- **the Average Silhouette Method**

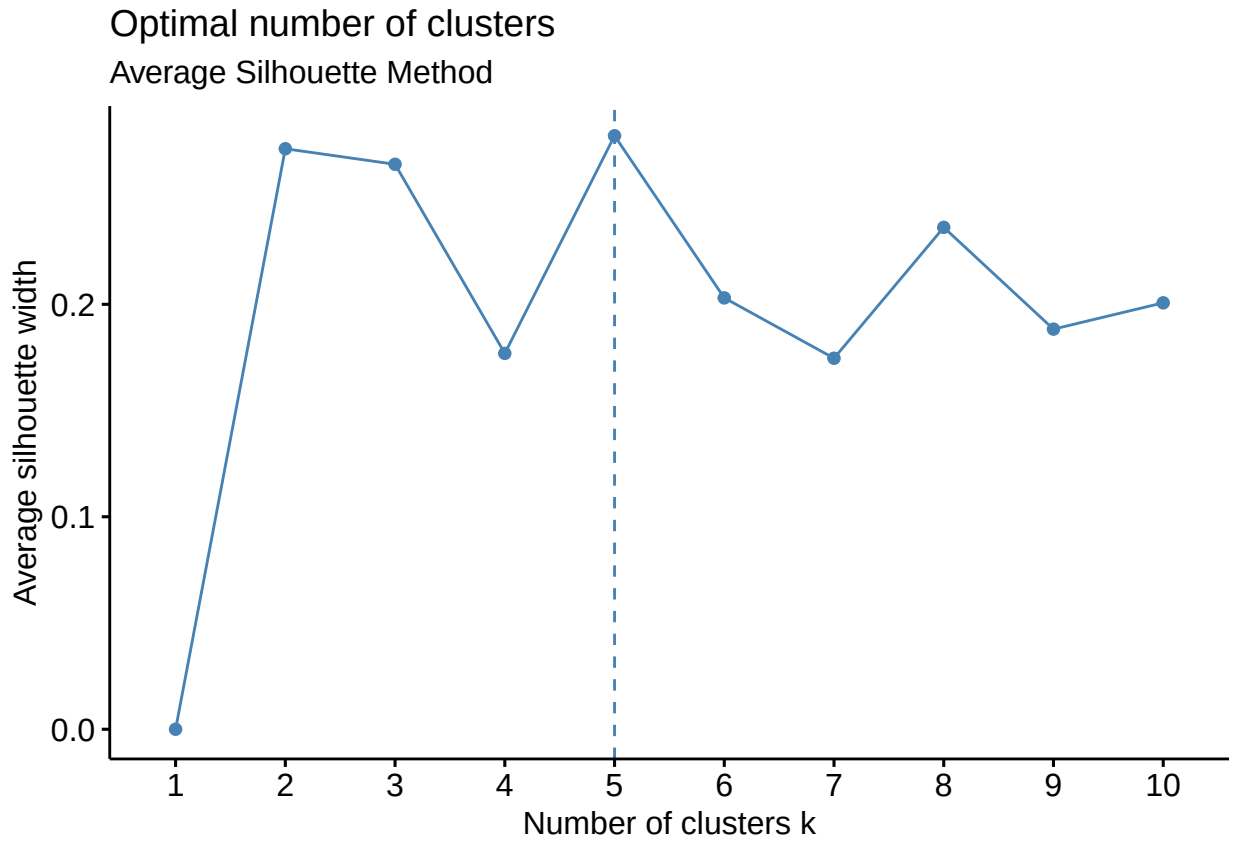
Elbow Method

```
# Elbow Method
fviz_nbclust(df_Norm_Numerical,kmeans,method = "wss") +
  geom_vline(xintercept = 2,linetype=2)+
  labs(subtitle = "Elbow Method")
```



Average Silhouette Method

```
# Average Silhouette Method
fviz_nbclust(df_Norm_Numerical,kmeans,method = "silhouette")+
labs(subtitle = "Average Silhouette Method")
```



From the two methods plotted, we can see that the Elbow Method gives us the optimal value of k is 2, While the Silhouette Method gives us k=5 as a result.

We have already ran the model with K=2 (which was the random number).Now we will run the k-means algorithm with 5 clusters and lets visualize it.

K-means Model using K=5

```
set.seed(123)

# Kmeans model using k=5
k5 <- kmeans(df_Norm_Numerical,centers = 5,nstart = 25)

# To see the centers of the 5 clusters
k5$centers
```

```
##    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
```

```
# To check the size of the 5 clusters
k5$size
```

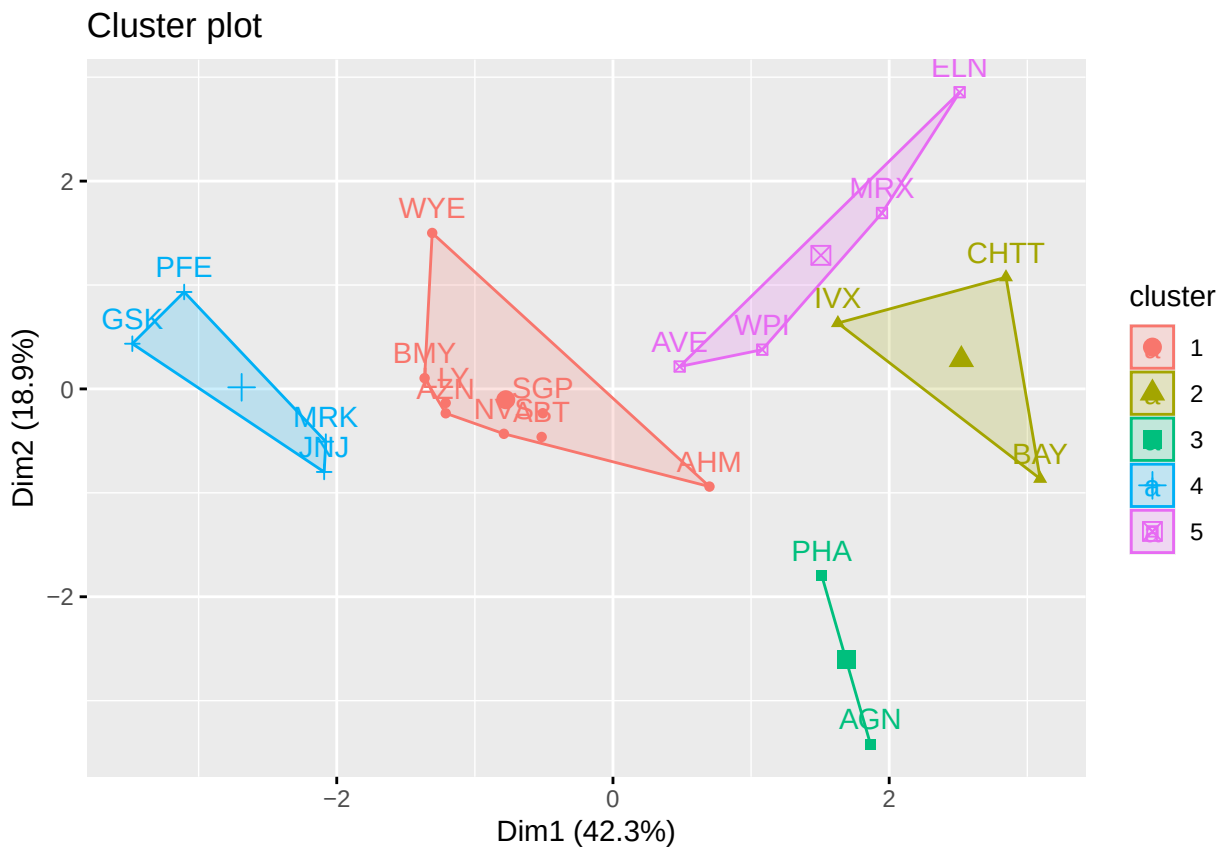
```
## [1] 8 3 2 4 4
```

```
#To identify to which cluster sizes belong
table(k5$cluster)
```

```
##
## 1 2 3 4 5
## 8 3 2 4 4
```

Visualize when K=5

```
# Lets visualize the 5 clusters
fviz_cluster(k5,data = df_Norm_Numerical)
```

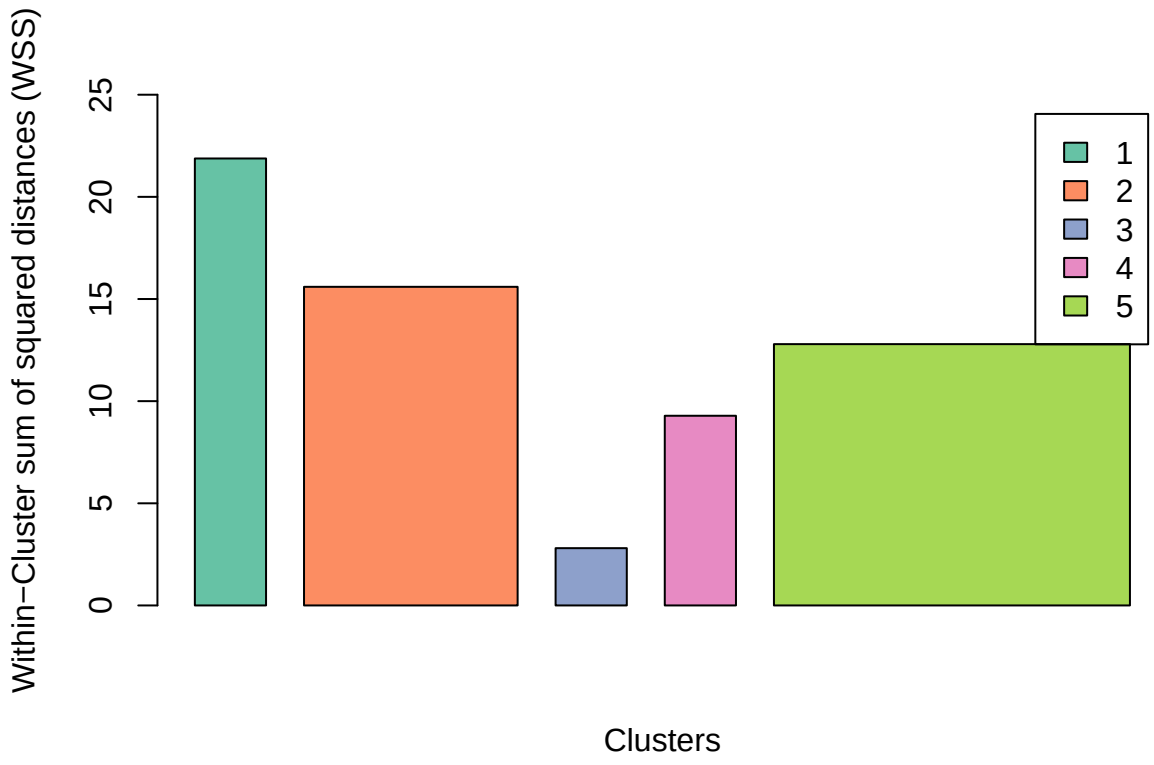


when we visualize the plot, We can see that 5 clusters with size of different characterizations:

- Cluster1 -(8 Firms) - Firms with Low Revenue Growth
- Cluster2 -(3 Firms) - High Leverage
- Cluster3 -(2 Firms) - High P/E Ratio
- Cluster4 -(4 Firms) - High Market Capitalization
- Cluster5 -(4 Firms) - Low Asset Turnover

Plot for Cluster and WithinSS

```
library(RColorBrewer)
coul <- brewer.pal(5, "Set2")
barplot(k5$withinss,k5$cluster,
        xlab = "Clusters",
        ylab = "Within-Cluster sum of squared distances (WSS)",
        ylim = c(0,25),
        col=coul,legend.text = c("1","2","3","4","5"))
```



This is the plot for clusters with their corresponding Within cluster sum of squared distances.

WithinSS is a measure of dispersion of the data within the cluster.When we take look at the above plot,It is clearly evident that Cluster1 is less homogeneous compared to the Cluster3. Moreover the records in the data set are very few to divide into 5 clusters and not able to find any significant trends with 5 clusters.The goal here isn't just to make clusters, but to make good, meaningful clusters.

For better Interpretation of cluster analysis, I decided to run the model with k=3. Since, with only 2 clusters, we might lose the main characteristics of the data. So lets run the model with k=3

Running the K-means Model with K=3 (For Better Interpretation)

```
#Run the K model with k=3
set.seed(123)
k3<- kmeans(df_Norm_Numerical,centers = 3,nstart = 25)

# To see the centers of the 5 clusters
k3$centers
```

```
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.6125361  0.2698666  1.3143935 -0.9609057 -1.0174553      0.2306328
## 2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 3 -0.8261772  0.4775991 -0.3696184 -0.5631589 -0.8514589     -0.9994088
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3592866 -0.5757385        -1.3784169
## 2 -0.3331068 -0.2902163         0.6823310
## 3  0.8502201  0.9158889        -0.3319956
```

```
# To check the size of clusters
k3$size
```

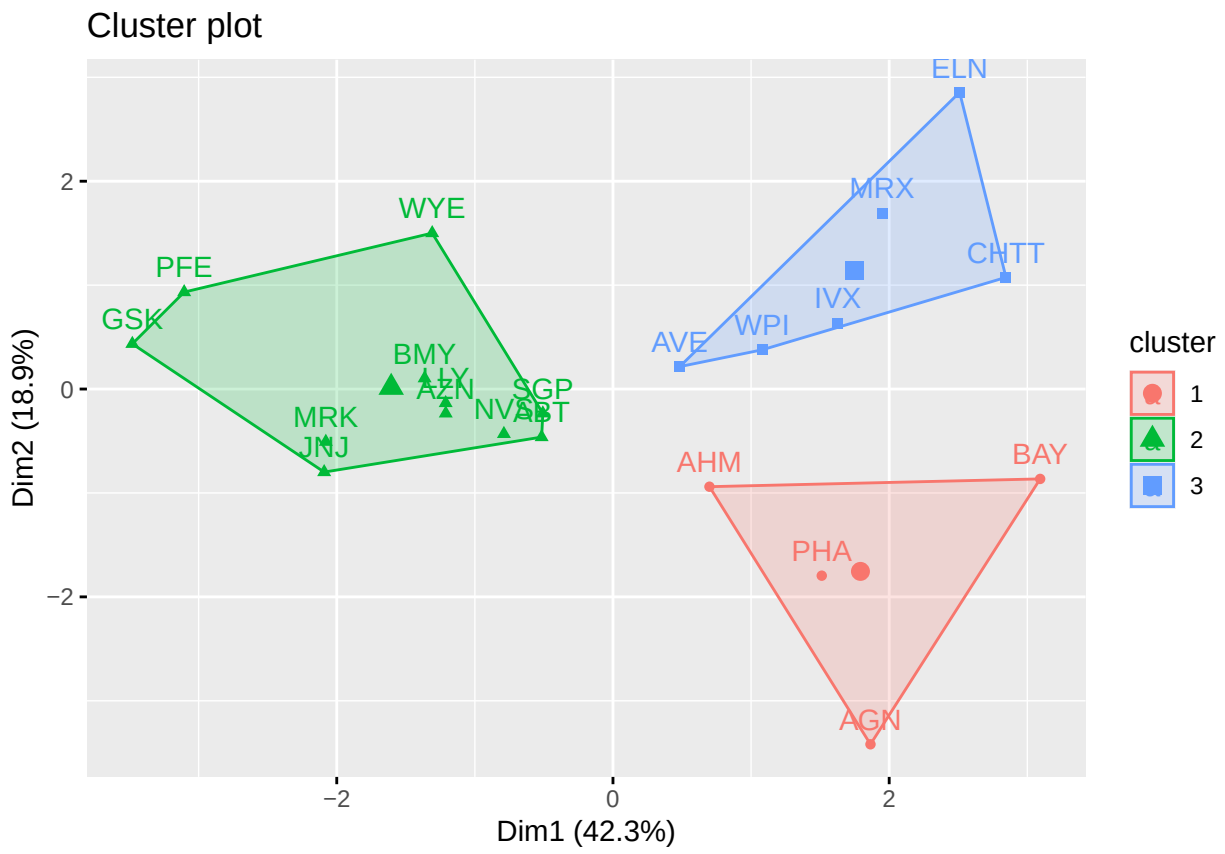
```
## [1]  4 11  6
```

```
#To identify to which cluster seizes belong
table(k3$cluster)
```

```
##
##  1  2  3
##  4 11  6
```

Visualize when (K=3)

```
fviz_cluster(k3,data = df_Norm_Numerical)
```



When we visualize the plot, We can see that 3 clusters with sizes of different characterizations:

- **Cluster1:** 4 Firms (AHM,PHA,BAY,AGN)
- **Cluster2:** 11 Firms (PFE,GSK,MRK,JNJ,BMY,SGP,ABT,WYE,AZN,LLY,NVS)
- **Cluster3:** 6 firms (AVE,WPI,IVX,CHTT,MRX,ELN)

Weightage for different variables:

- To consider weightage for the variables, we have to check for the Standard Deviation or Variance. If the sd/variance of any variables is high, which implies that those variables has higher weightage.
- In our case, P/E Ratio, Market_Capital, ROE, ROA, Net_Profit_Margin and Asset_Turnover has high variance , so we can consider these variables to have higher weightage among all the variables.

Also, we can assign weightage to each variables using PCA (Principal Component Analysis)

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

```
#Adding a cluster column for all the firms

Pharma_data_Cluster<- cbind(Pharmaceutical_data,cluster=as.factor(k3$cluster))
Pharma_Clust<-Pharma_data_Cluster[order(Pharma_data_Cluster$cluster),]
```

With respect to Numeric Variables, we have combined the data with their corresponding clusters.

Symbol	Name	Market_Cap	Beta	PE_Ratio	ROE	ROA	Asset_Tur	Leverage	Rev_Grow	Net_Profit	Median_R	Location	Exchange	Cluster
AGN	Allergan, Ir	7.58	0.41	82.5	12.9	5.5	0.9	0.6	9.16	5.5	Moderate	CANADA	NYSE	1
AHM	Amersham	6.3	0.46	20.7	14.9	7.8	0.9	0.27	7.05	11.2	Strong Buy	UK	NYSE	1
BAY	Bayer AG	16.9	1.11	27.9	3.9	1.4	0.6	0	-3.17	2.6	Hold	GERMANY	NYSE	1
PHA	Pharmacia	56.24	0.4	56.5	13.5	5.7	0.6	0.35	15	7.3	Hold	US	NYSE	1
ABT	Abbott Lat	68.44	0.32	24.7	26.4	11.8	0.7	0.42	7.54	16.1	Moderate	US	NYSE	2
AZN	AstraZene	67.63	0.52	21.5	27.4	15.4	0.9	0	15	18	Moderate	UK	NYSE	2
BMY	Bristol-My	51.33	0.5	13.9	34.8	15.1	0.9	0.57	2.7	20.6	Moderate	US	NYSE	2
LLY	Eli Lilly and	73.84	0.18	27.9	31	13.5	0.6	0.53	6.21	23.4	Hold	US	NYSE	2
GSK	GlaxoSmitl	122.11	0.35	18	62.9	20.3	1	0.34	21.87	21.1	Hold	UK	NYSE	2
JNJ	Johnson &	173.93	0.46	28.4	28.6	16.3	0.9	0.1	9.37	17.9	Moderate	US	NYSE	2
MRK	Merck & C	132.56	0.46	18.9	40.6	15	1.1	0.28	17.35	14.1	Hold	US	NYSE	2
NVS	Novartis A	96.65	0.19	21.6	17.9	11.2	0.5	0.06	-2.69	22.4	Hold	SWITZERL	NYSE	2
PFE	Pfizer Inc	199.47	0.65	23.6	45.6	19.2	0.8	0.16	25.54	25.2	Moderate	US	NYSE	2
SGP	Schering-P	34.1	0.51	18.9	22.6	13.3	0.8	0	8.56	17.6	Hold	US	NYSE	2
WYE	Wyeth	48.19	0.63	13.1	54.9	13.4	0.6	1.12	0.36	25.5	Hold	US	NYSE	2
AVE	Aventis	47.16	0.32	20.1	21.8	7.5	0.6	0.34	26.81	12.9	Moderate	FRANCE	NYSE	3
CHTT	Chattem, I	0.41	0.85	26	24.1	4.3	0.6	3.51	6.38	7.5	Moderate	US	NASDAQ	3
ELN	Elan Corpc	0.78	1.08	3.6	15.1	5.1	0.3	1.07	34.21	13.3	Moderate	IRELAND	NYSE	3
IVX	IVAX Corp	2.6	0.65	19.9	21.4	6.8	0.6	1.45	13.99	11	Hold	US	AMEX	3
MRX	Medicis Ph	1.2	0.75	28.6	11.2	5.4	0.3	0.93	30.37	21.3	Moderate	US	NYSE	3
WPI	Watson Ph	3.26	0.24	18.4	10.2	6.8	0.5	0.2	29.18	15.1	Moderate	US	NYSE	3

```
# Plotting a graph for the average of all the numerical variables by clusters

Average_of_Variables<- data.frame(k3$centers) %>% rowid_to_column()
colnames(Average_of_Variables) <- c("RowID","MarketCap","Beta","P/E","ROE","ROA",
                                    "Asset_T/O",
                                    "Lev","Rev","Profit")
colnames(Average_of_Variables)
```

```
##  [1] "RowID"     "MarketCap" "Beta"      "P/E"       "ROE"       "ROA"
##  [7] "Asset_T/O" "Lev"       "Rev"       "Profit"
```

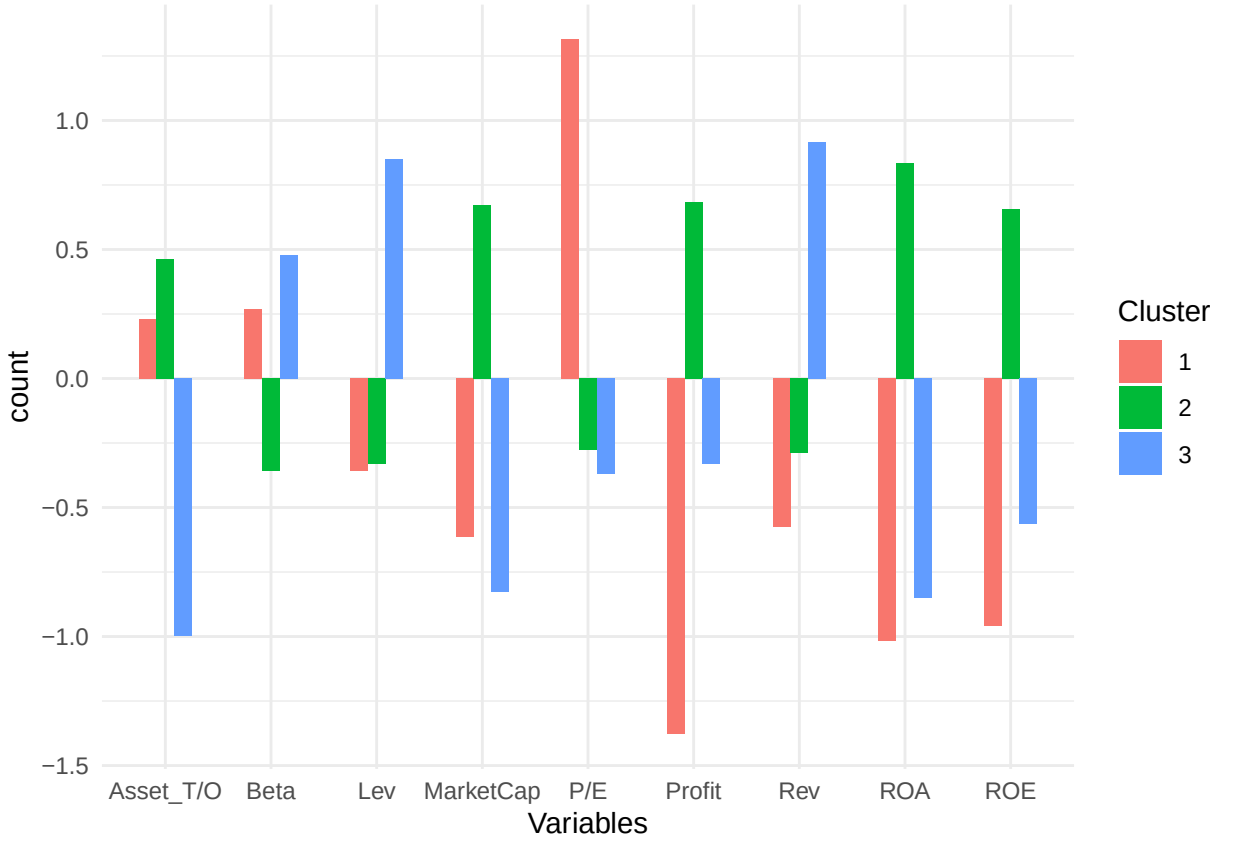
```
df<-Average_of_Variables %>%
  pivot_longer(MarketCap:Profit,names_to = "Variables",values_to = "Average") %>%
  arrange(Variables)

df$rowid<-as.factor(df$RowID)


#esquisser(df)

library(ggplot2)

ggplot(df) +
 aes(x = Variables, fill = rowid, group = rowid, weight = Average) +
 geom_bar(width=0.5,position = "dodge") +
 scale_fill_hue(direction = 1) +
 labs(fill = "Cluster") +
 theme_minimal()
```



MarketCap	Beta	P/E	ROE	ROA	Asset_T/O	Lev	Rev	Profit
Market_Capital	Beta	Price/Earnings Ratio	Return on Equity	Return on Assets	Asset_Turnover	Leverage	Rev_Growth	Net_Profit_Margin

The above plot helps us to differentiate by cluster wise with their averages of the numerical variables.

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

Let us now consider the 3 categorical variables: Median Recommendation, Location and Stock Exchange.

Median Recommendation: (across major brokerages) -The mean or median recommendation that analysts make on a stock. The consensus recommendation is calculated simply by compiling recommendations and taking the average or median.

Location: Location of the firm.

Exchange: The firm in which securities are bought and sold.

In order to look for possible trends within data, I decided to leverage bar charts to graphically visualize the distribution of the firms grouped by clusters.

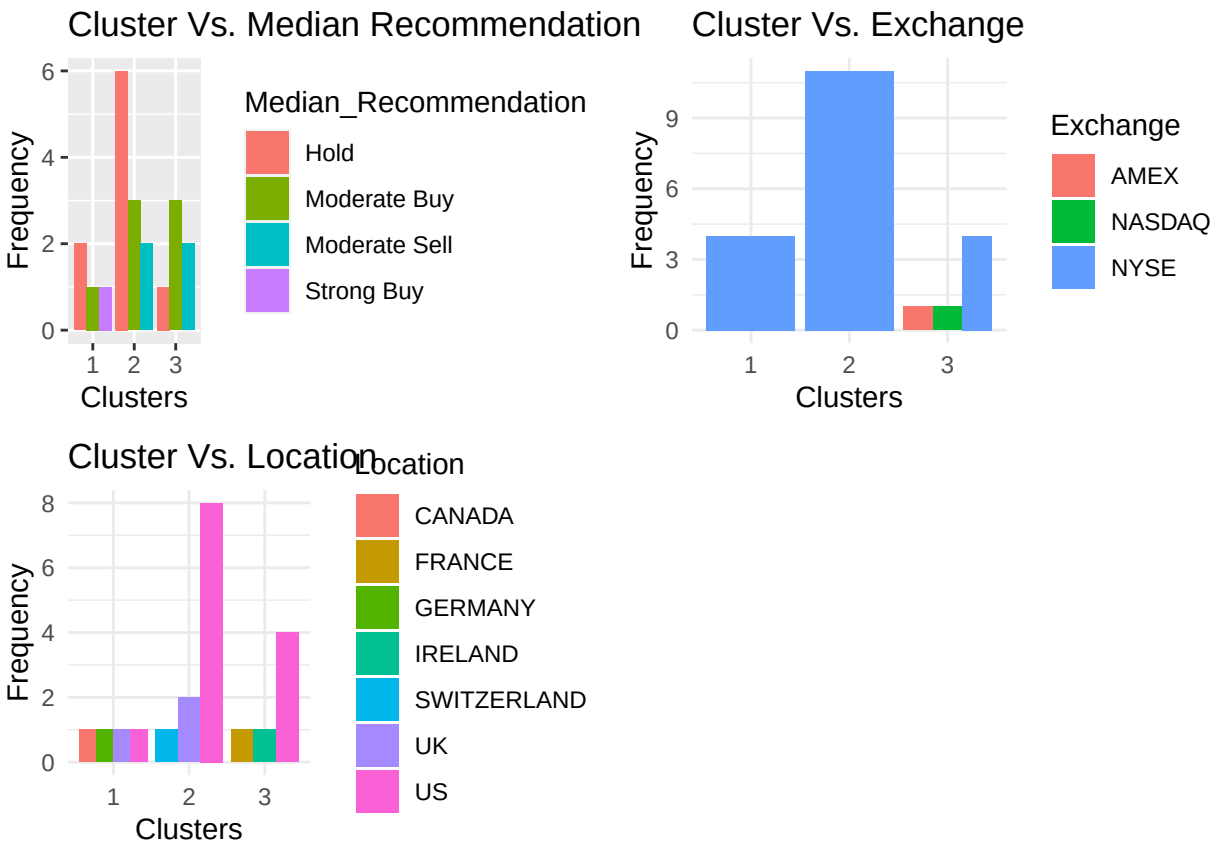
```
Median_Rec <- ggplot(Pharma_Clust) +
 aes(x = cluster, fill = Median_Recommendation) +
  geom_bar(position = "dodge") +
 scale_fill_hue(direction = 1) +
 labs(x = "Clusters", y = "Frequency", title = "Cluster Vs. Median Recommendation")

Exchange <-ggplot(Pharma_Clust) +
```

```
 aes(x = cluster, fill = Exchange) +
 geom_bar(position = "dodge") +
 scale_fill_hue(direction = 1) +
labs(x = "Clusters", y = "Frequency", title = "Cluster Vs. Exchange") +
 theme_minimal()

Location <-ggplot(Pharma_Clust) +
 aes(x = cluster, fill = Location) +
 geom_bar(position = "dodge") +
 scale_fill_hue(direction = 1) +
  labs(x = "Clusters", y = "Frequency", title = "Cluster Vs. Location") +
 theme_minimal()

plot_grid(Median_Rec,Exchange,Location)
```



I do not find any possible trends with Location and Exchange variables as they spread in all the clusters which doesn't play a major role in identifying the patterns between the clusters. Whereas Median Recommendation shows a slight pattern on the clusters.

- Cluster 1: Hold, Moderate Buy and Strong buy which indicates that firms may outperform in the future.
- Cluster 2 & Cluster 3 : Has a mix of Hold,Moderate Buy and moderate sell which indicates that average risk and profit.

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

CLUSTER 1 : **Overvalued International Firms**

- High PE Ratio either indicates that the stock is overvalued or analysts predict that the company will perform well in the future(Long term).
- Good Assert Turnover represents that that company is handling their assets efficiently to generate rev while ROE looks moderate.
- High beta indicates that companies are volatile to the Market movement.
- Lowest leverage among the clusters.

CLUSTER 2 : **High Capital Profitable Firms** (Mature Firms) - The companies with high number of stocks and / or high share prices.

- High Cap indicates that company has high total market price (Share price * total outstanding stocks).
- High ROE & ROA - Indicates that Companies are performing well and producing profits
- High Asset turnover depicts that company is efficiently handling all of their assets to generate the revenue
- Companies has a high Net Profit Margin which is the ration of Profit to the revenue. For every 100$, Company produces a profit of 20.17
- Low leverage indicates that company owns more asset than the debt which is less risk.

CLUSTER 3 : **High Leveraged and UnderValued Firms** - Company runs on the borrowed capital which could be risky. High Leverage indicates higher Debt than the Equity / asset the company owns.

- Undervalued (Low P/E)
- Low beta indicates that company is less volatile to Market movements
- High leverage means that company is running on high borrowed capital, high risk
- Higher revenue growth

In a nutshell,

Cluster 1 can be labeled as **OVERVALUED INTERNATIONAL FIRMS** due to high Price/Earnings ratio and low Net Profit Margin . Buying overvalued firms can be risky, as they might drop closer to their intrinsic value at any time, especially over the short term. Over the long term, the intrinsic value of healthy and growing companies will grow. But it's still possible to simply pay too much for a stock. so these companies need to meet investor's expectations by investing and increasing the profits and maintaining their stock price if they do not want their stock price to decrease and these firms should be able to perform Business Continuity Plan as these firms are volatile to the market movements due to high Beta.

Cluster 2 can be represented as **HIGH CAPITAL PROFITABLE FIRMS (MATURE FIRMS)** due to High Profit and market capitalization.These firms also has high ROE,ROA and Asset turnover which indicates that these firms are performing well and handling all their assets to generate a good revenue. Buying these firms is of less risk as these are well stable, runs on less debt and matured enough to handle the fluctuations in market.

Cluster 3 can be characterized as **HIGH LEVERAGED AND UNDERVALUED FIRMS** due to Low Price/Earnings ratio and high leverage. An undervalued stock is one with a market price that is significantly lower than its real or 'fair' value (market value $<$ fair value).But Undervalued stocks are expected to go higher and these firms are estimated to generate higher revenue growth. This seems to be a good option for few investors. However, these firms should concentrate on gaining investor confidence, financial health of the company and asset turnover.