ML pharmaceuticals

Sai Supriya

2022-10-31

First CSV file and Required Packages are loaded

In this project, I'm using the k-means clustering technique to do a non-hierarchical cluster analysis. The goal is to divide the data into homogeneous clusters from which we may extract meaningful information.

###Importing the dataset

```
#packages are loaded
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.2
```

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

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

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

```
library(factoextra)
```

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

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.1     v stringr 1.4.0
## v readr   2.1.3     v forcats 0.5.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(cowplot)
#Reading the dataset
library(readr)
Pharmaceuticals <- read.csv("Downloads/Pharmaceuticals.csv")
view(Pharmaceuticals)
head(Pharmaceuticals)
str(Pharmaceuticals)
summary(Pharmaceuticals)
dim(Pharmaceuticals)
colMeans(is.na(Pharmaceuticals))
row.names(Pharmaceuticals) <- Pharmaceuticals[,2]
Pharmaceuticals <- Pharmaceuticals[,-2]
```

1)Used only the numerical variables (1 to 9) to cluster the 21 firms. Justifying 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,

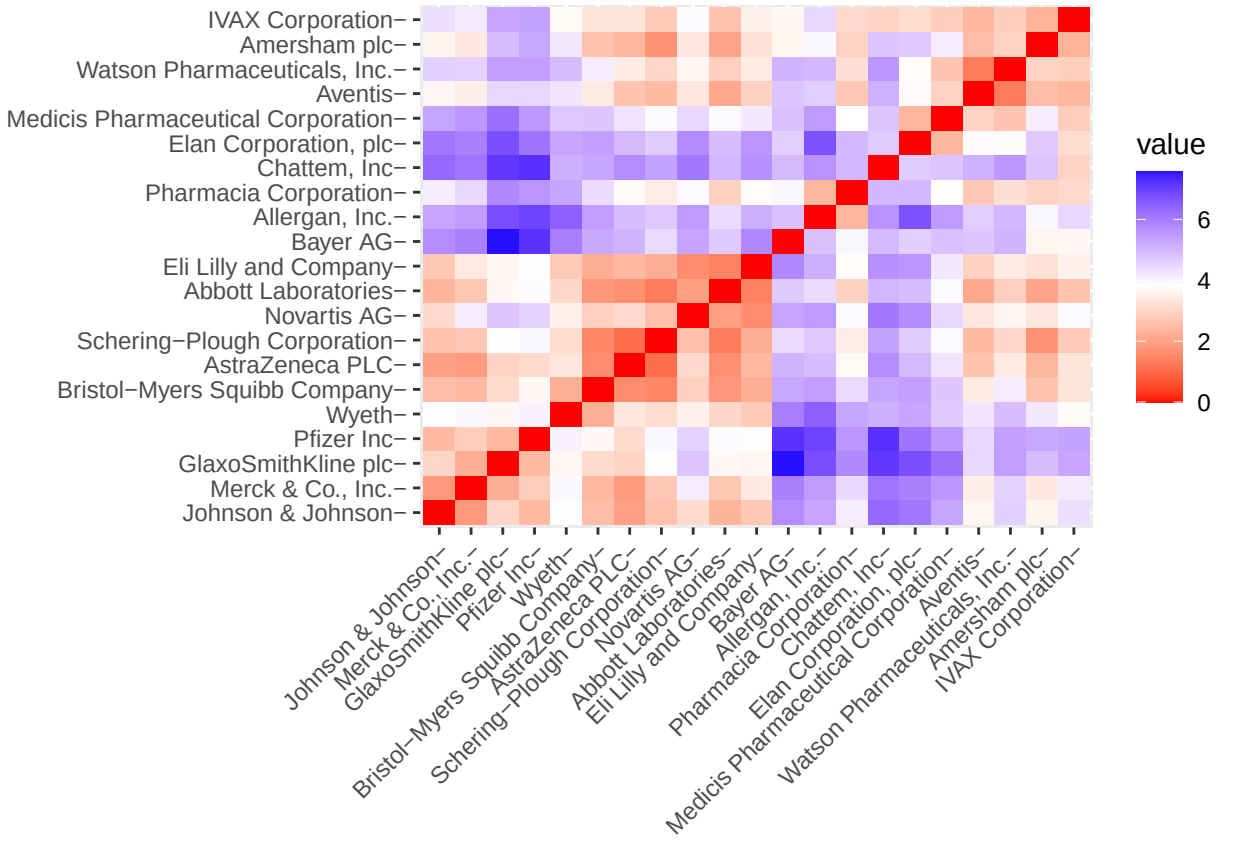
Focusing on a subset of the original dataset that only contains numerical variables till now.

```
#with the exception of "Symbol" and the last 3 non-numerical variables
Pharmaceuticals.Q1 <- Pharmaceuticals[,-c(1,11:13)]
```

Normalizing and Clustering the data

Here I calculated the distance between each observation because, the Euclidean distance metric is utilized by default and is scale sensitive, data must first be normalised before calculating the distance.

```
#Normalizing data
norm.Pharmaceuticals.Q1<- scale(Pharmaceuticals.Q1)
#Measuring and plotting distance between the observations
distance <- get_dist(norm.Pharmaceuticals.Q1)
fviz_dist(distance)
```



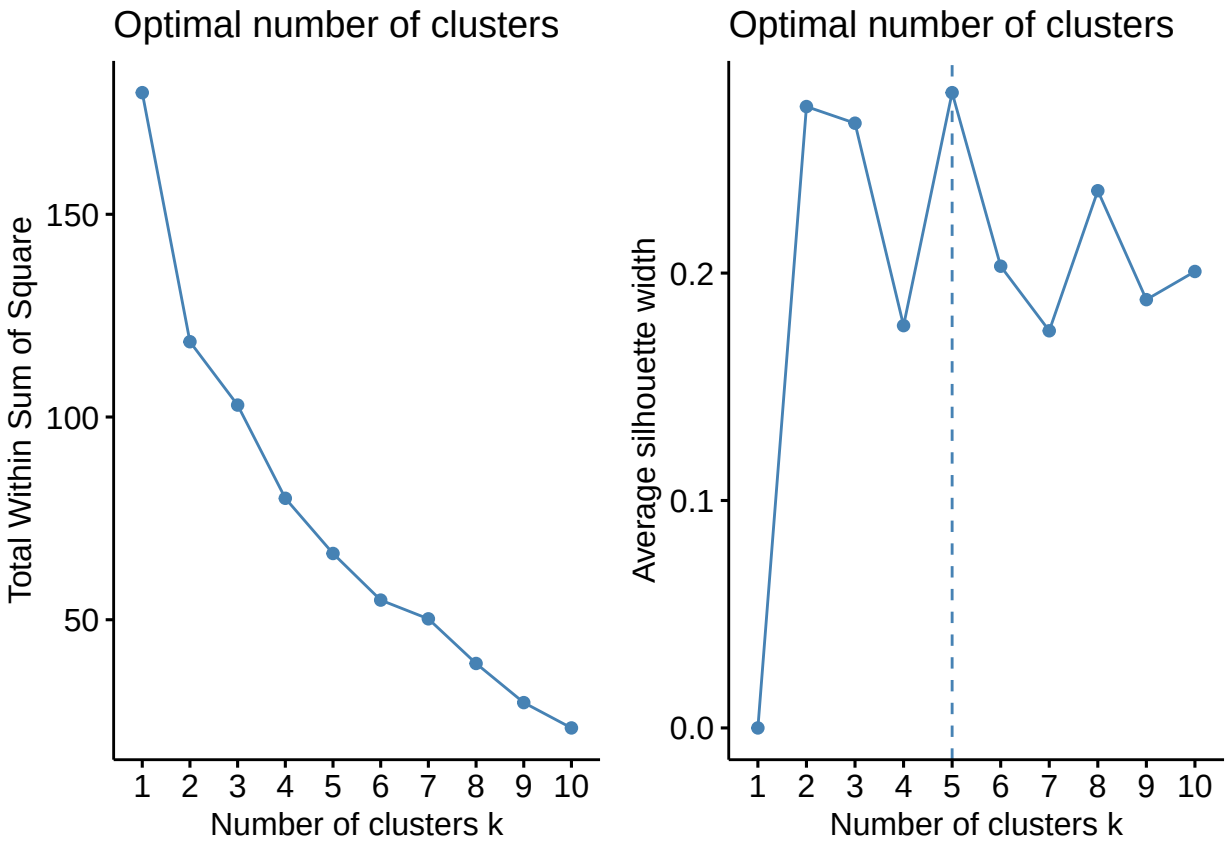
The graph depicts the intensity variation of color with distance. The diagonal represents that we had predicted which is in red, has a value of zero since it represents the distance between two observations in the graph.

##Finding the optimal K value

The Elbow chart and the Silhouette Method are two of the most effective approaches for calculating the number of clusters for the k-means model when you don't have external factors influencing it.

```
#Using elbow chart and silhouette method for calculating the kmeans

WSS <- fviz_nbclust(norm.Pharmaceuticals.Q1, kmeans, method = "wss")
Silhoutte <- fviz_nbclust(norm.Pharmaceuticals.Q1, kmeans, method = "silhouette")
plot_grid(WSS, Silhoutte)
```



The plotted charts produces k=2, k=5 for the methods when used as Elbow and Silhoutte respectively. I am using the k-means method with k=5.

```
#using k-means with k=5 for making clusters
set.seed(646)
KMeans.Pharmaceuticals.Opt <- kmeans(norm.Pharmaceuticals.Q1, centers = 5, nstart = 25)
KMeans.Pharmaceuticals.Opt$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 2  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1  1.36644699 -0.6912914      -1.320000179
## 2 -0.46807818  0.4671788       0.591242521
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.27449312 -0.7041516       0.556954446
## 5  0.06308085  1.5180158      -0.006893899
```

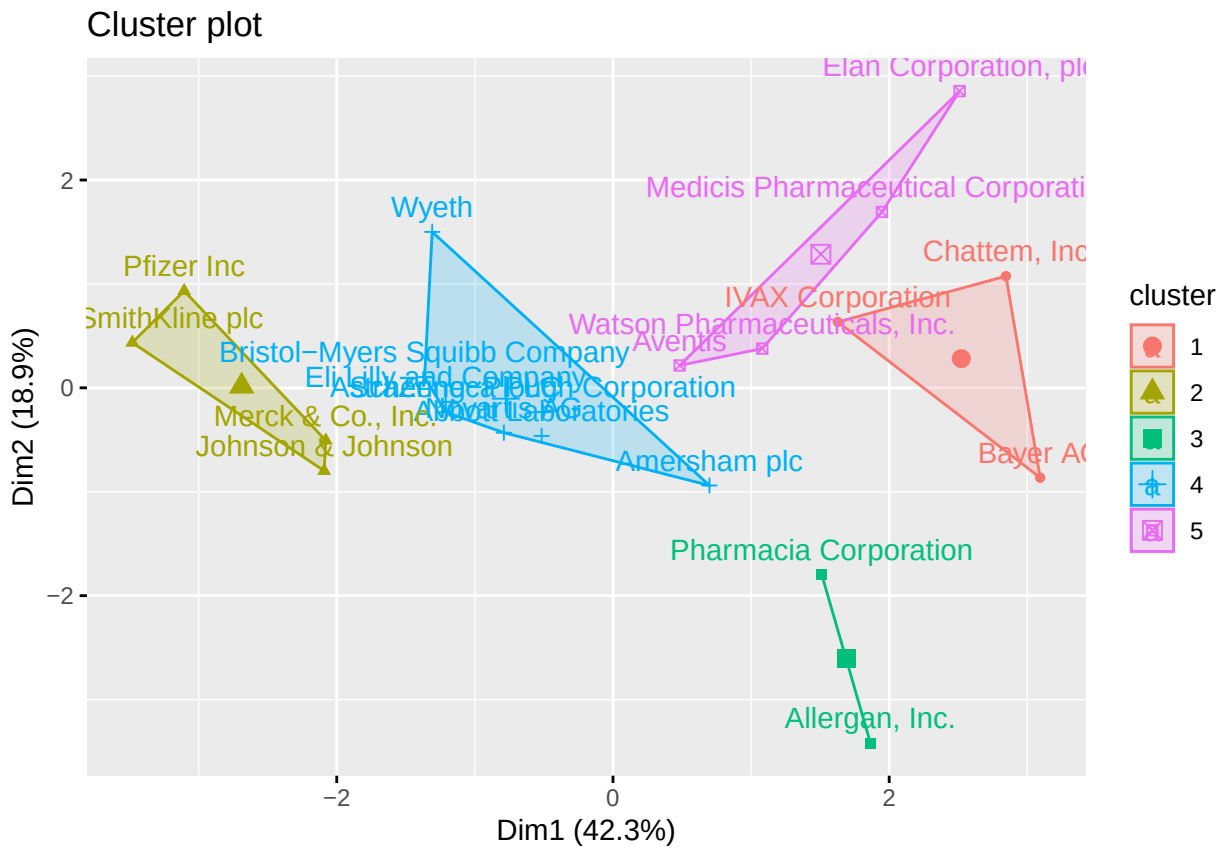
```
KMeans.Pharmaceuticals.Opt$size
```

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

```
KMeans.Pharmaceuticals.Opt$withinss
```

```
## [1] 15.595925  9.284424  2.803505 21.879320 12.791257
```

```
fviz_cluster(KMeans.Pharmaceuticals.Opt, data = norm.Pharmaceuticals.Q1)
```



###Results:

From the given data, we defined the five clusters depending on their distance from the cores. ##1 Cluster.4 has a high Market Capital ##2 Cluster n.2 has a high Beta and ##3 Cluster.5 does have a low Asset Turnover. ##4) Cluster.1 has the most enterprises, whereas Cluste.3 has only two.

##The within-cluster sum of squared distances reveal information regarding data dispersion: homogeneity cluster.1 (21.9) < cluster.3 (2.8). In the end visualizing the algorithm's output, we have observed the five groups into which the data has been grouped.

##2)Interpreting the clusters with respect to the numerical variables used in forming the clusters. Now Im running the model with 5 clusters as a basis to obtain btter results as using only 2 might lose the features of data.

```
#using k-means algorithm with k=3 for making clusters
set.seed(643)
KMeans.Pharmaceuticals <- kmeans(norm.Pharmaceuticals.Q1, centers = 5, nstart = 25)
KMeans.Pharmaceuticals$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
```

```
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 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 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.46807818  0.4671788       0.591242521
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.14170336 -0.1168459      -1.416514761
```

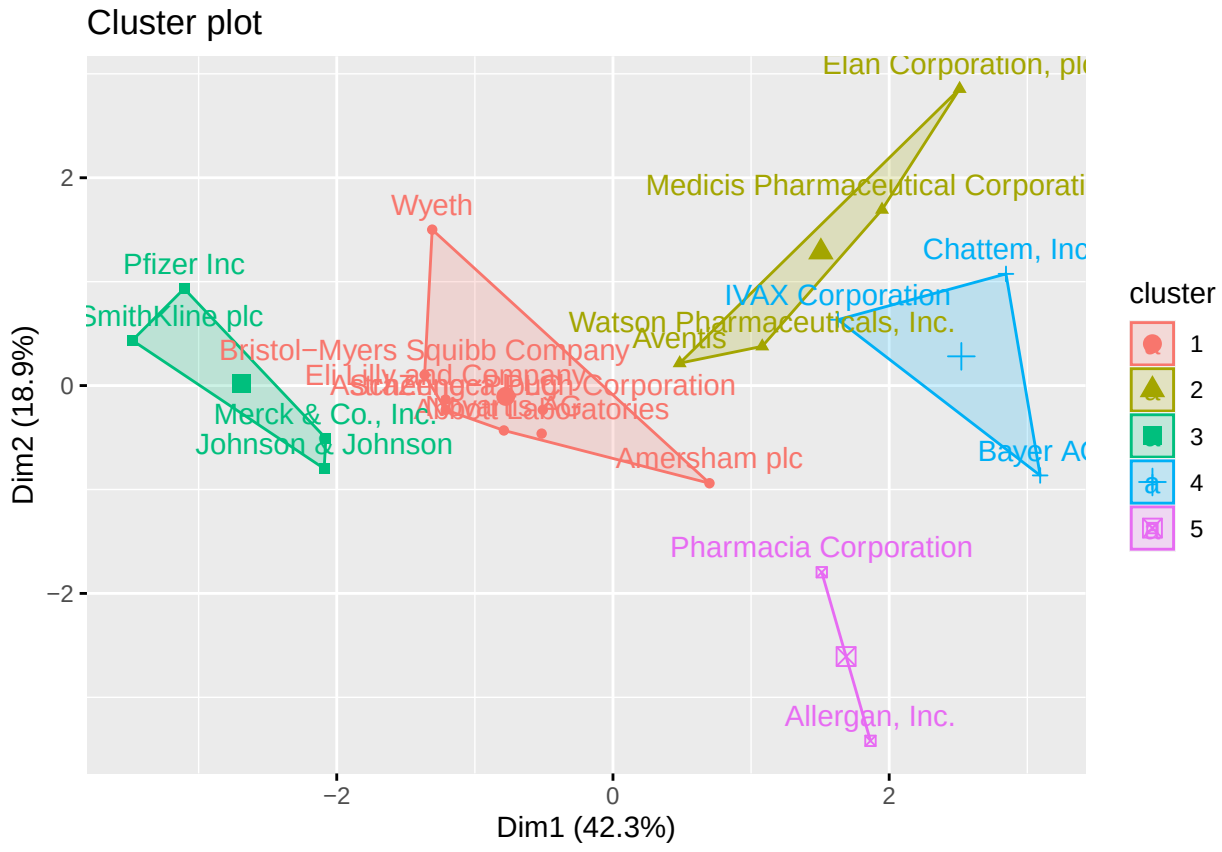
```
KMeans.Pharmaceuticals$size
```

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

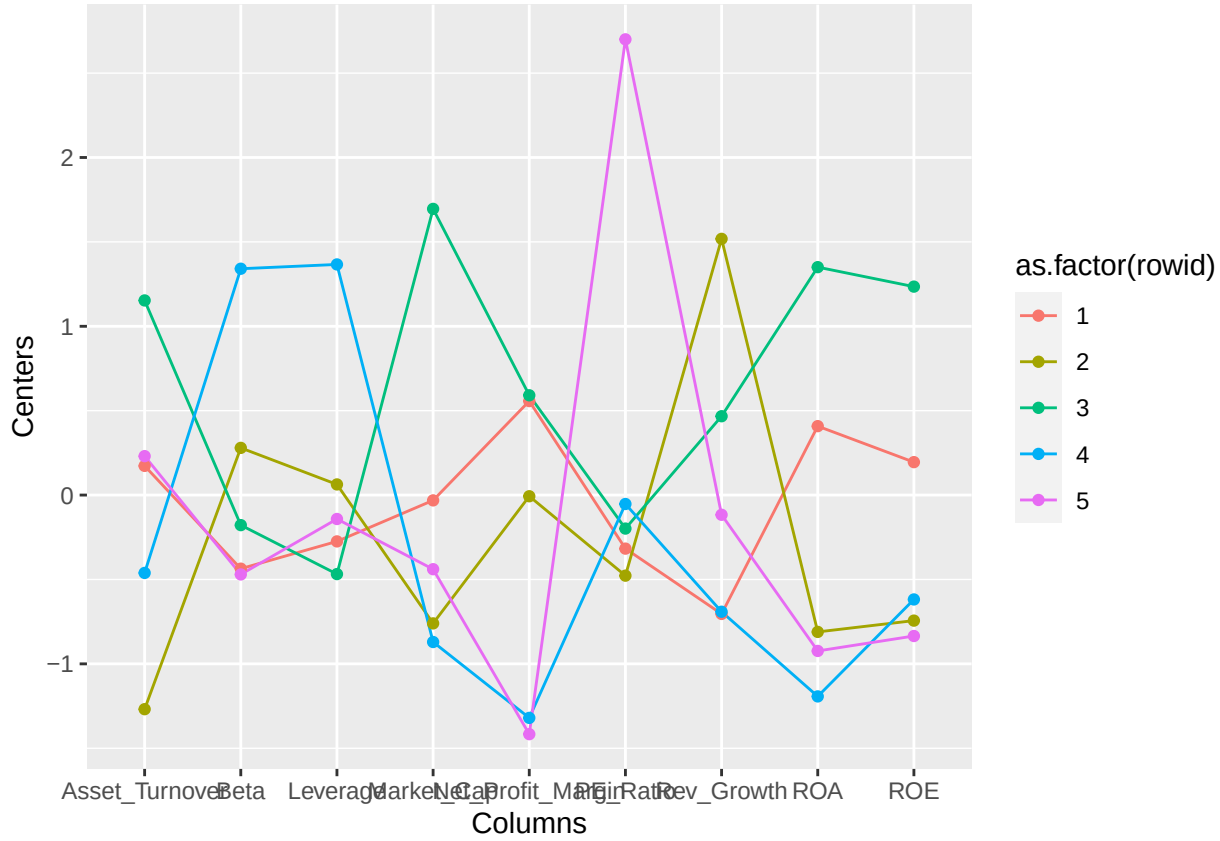
```
KMeans.Pharmaceuticals$withinss
```

```
## [1] 21.879320 12.791257  9.284424 15.595925  2.803505
```

```
fviz_cluster(KMeans.Pharmaceuticals, data = norm.Pharmaceuticals.Q1)
```



We have identified the grouping and management of the clusters in the analysis. We can observe 4 data points in cluster.1, 11 data points in cluster.2, and 6 data points in cluster.3.



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

By looking at the mean values of all quantitative variables in each cluster.

##Cluster 1 - K(has highest Market_cap,ROA,ROE,Asset_Turnover and lowest is Beta,PE_Ratio.)

##Cluster 2 - (has highest Rev_Growth and lowest PE_Ratio, Asset_Turnover.)

##Cluster 3 - has highest Beta, Leverage and lowest Market_Cap, ROE, ROA, Leverage, Rev_Growth, Net_Profit_Margin.

##Cluster 4 - has highest PE_Ratio and lowest Leverage, Asset_Turnover. ##Cluster 5 - has highest Net_Profit_Margin and lowest leverage,Beta.

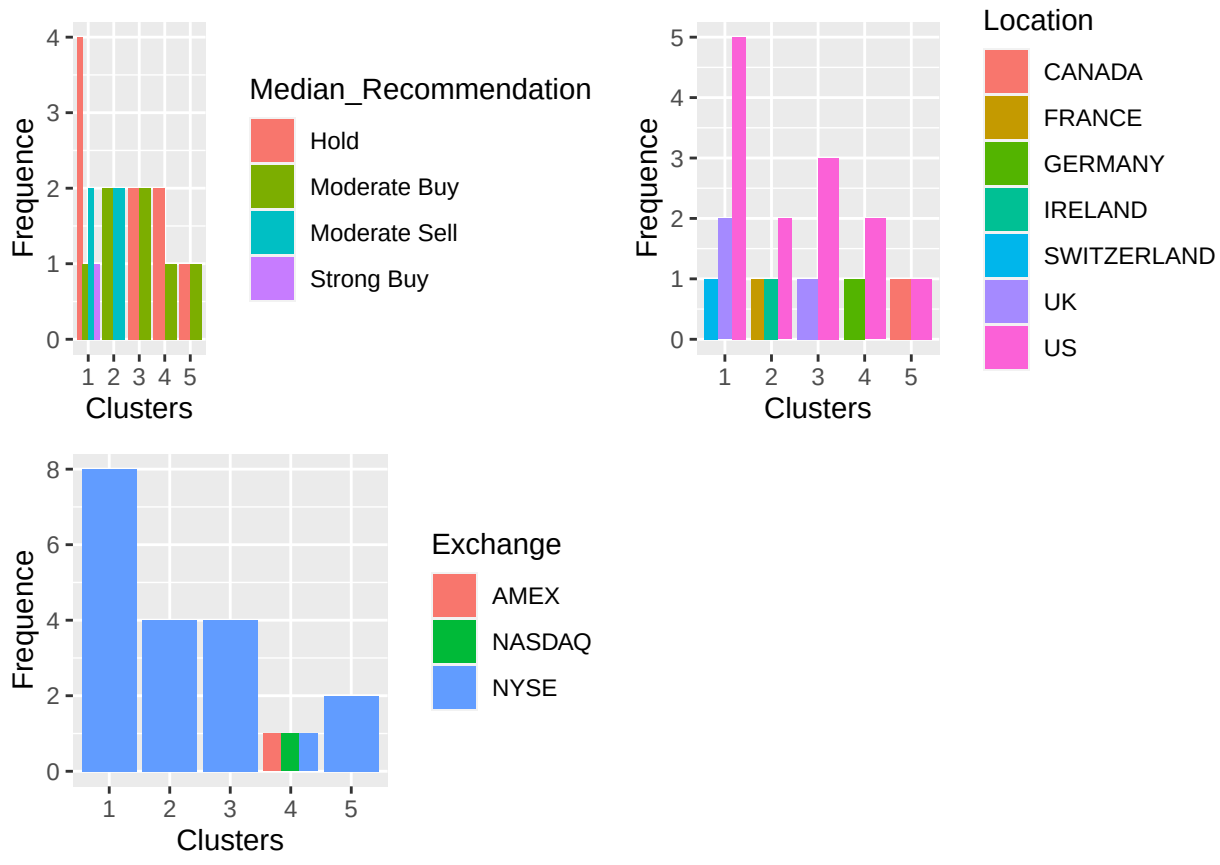
Task C)Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

Consider the last three categorical variables: Median Recommendation, Location, and Stock Exchange. In order to check for any trends in the data, I choose to utilize bar charts to graphically display the distribution of firms grouped by clusters.

```
#data set partitioning for last 3 variables
Pharmaceuticals.Q3 <-  Pharmaceuticals %>% select(c(11,12,13)) %>%
    mutate(Cluster = KMeans.Pharmaceuticals$cluster)
```

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

```
#cluster plots
Median_Recommendation <- ggplot(Pharmaceuticals.Q3, mapping = aes(factor(Cluster), fill=Median_Recommen
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Location <- ggplot(Pharmaceuticals.Q3, mapping = aes(factor(Cluster), fill=Location)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Stock_Exchange <- ggplot(Pharmaceuticals.Q3, mapping = aes(factor(Cluster), fill=Exchange)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
plot_grid(Median_Recommendation, Location,Stock_Exchange)
```



The graph plainly illustrates that the majority of the companies in cluster 3 are based in the United States, and they all have a spread recommendation to hold their shares. They are all traded on the New York Stock Exchange.

In cluster.2, we choose "Moderate Buy" shares, and we include just two companies whose stocks are listed on other exchanges or indexes (AMEX and NASDAQ).

Cluster.1 shows that the four firms are located in four different countries, and their stocks are traded on the NYSE.

In Conclusion

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

Cluster-1 - Moderate Buy (or) Hold cluster.

Cluster-2 - Low PE_Ratio, Asset_Turnover cluster (or) Hold cluster.

Cluster-3 - High Beta, Leverage cluster (or) Buy Cluster.

Cluster-4 - High PE_Ratio cluster (or) High Hold cluster.

Cluster-5 - High Net_Profit_Margin cluster (or) High Hold cluster.