Assignment 4

Lava Kumar

3/19/2022

First CSV file and Required Packages are loaded

In this project, I will use 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. Let's start by loading the required packages and the original dataset. It contains information about 21 pharmaceutical companies.

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

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3

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

```
library(dplyr)
```

```
##
## 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 4.1.3
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'forcats' was built under R version 4.1.3

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

```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.1.3
```

```
#Reading the dataset
library(readr)
Pharmaceut <- read.csv("C:/Users/lavak/Documents/R/Assignment4/Pharmaceuticals.csv")
view(Pharmaceut)
head(Pharmaceut)
str(Pharmaceut)
summary(Pharmaceut)
dim(Pharmaceut)
colMeans(is.na(Pharmaceut))

row.names(Pharmaceut) <- Pharmaceut[,2]
Pharmaceut <- Pharmaceut[,-2]
```

1)Using 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.

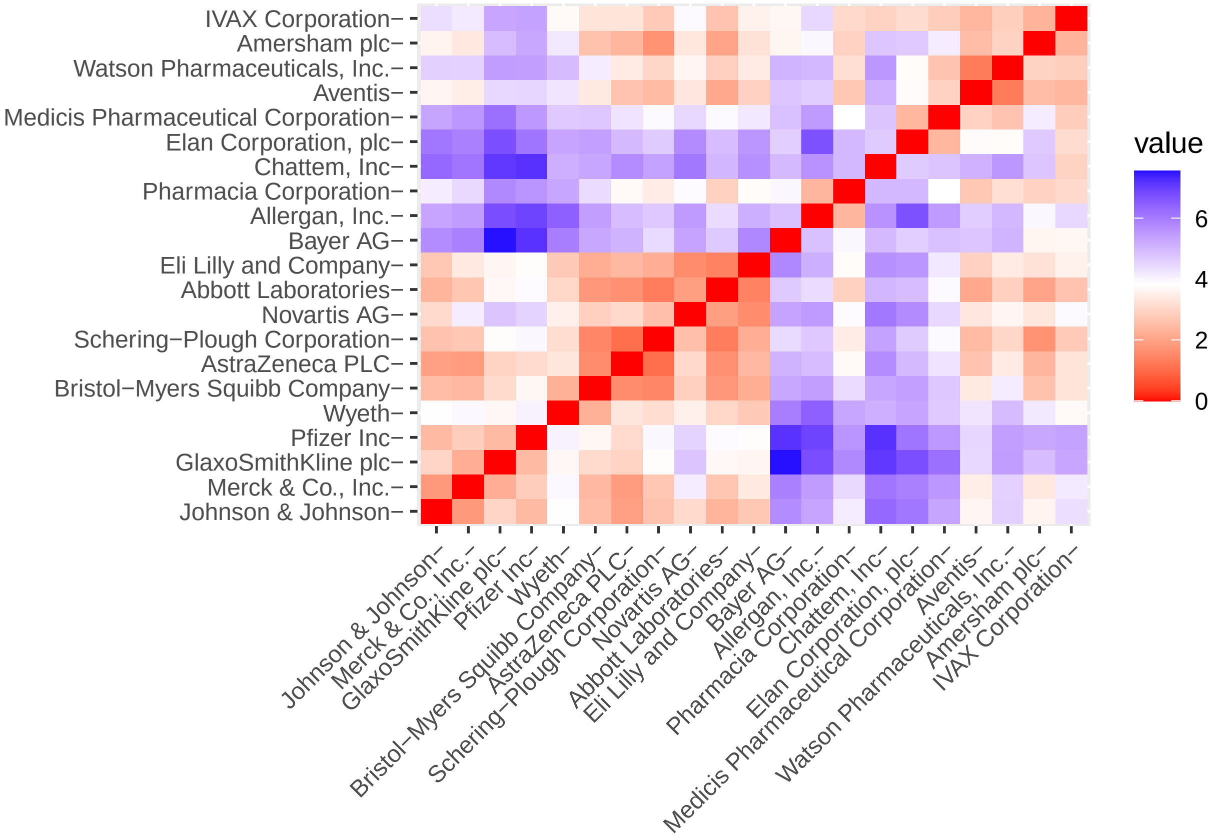
Focusing on a subset of the original dataset that only contains numerical variables for the first part of the assignment.

```
#with the exception of "Symbol" and the last 3 non-numerical variables
Pharmaceut.Que1 <- Pharmaceut[,-c(1,11:13)]
```

Normalizing and Clustering the data

I compute the distance between each observation in this part. Because the Euclidean distance metric is utilized by default and is scale sensitive, data must first be modified.

```
#normalizing data
norm.Pharmaceut.Que1 <- scale(Pharmaceut.Que1)
#measuring and plotting distance
dist <- get_dist(norm.Pharmaceut.Que1)
fviz_dist(dist)
```

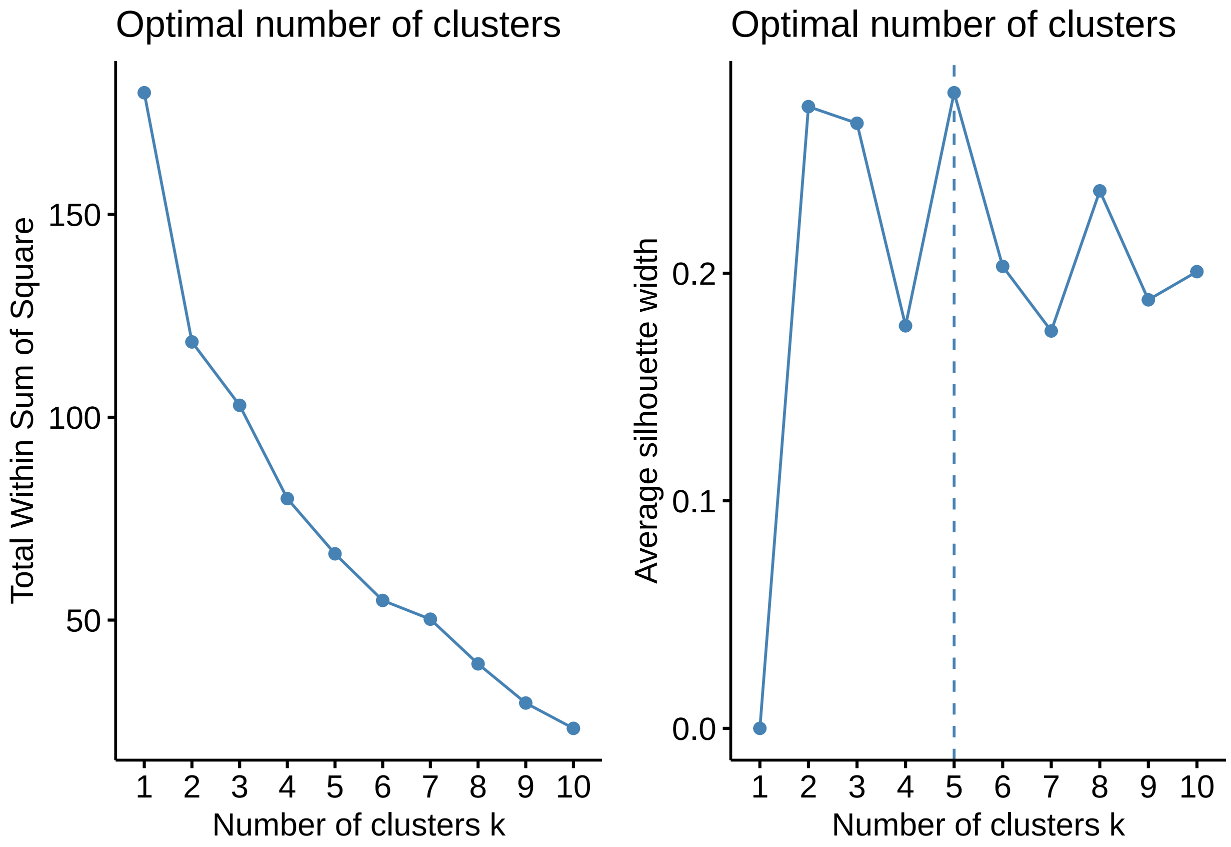


The graph depicts how intensity of color varies with distance. The diagonal, as we would predict, has a value of zero since it represents the distance between two observations.

##Finding the optimal K value

When there are no external factors, 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. The former shows how cluster heterogeneity decreases when more clusters are introduced. The latter compares an object's similarity to its cluster to the other clusters.

```r
#Using elbow chart and silhouette method
WSS <- fviz_nbclust(norm.Pharmaceut.Que1, kmeans, method = "wss")
Silho <- fviz_nbclust(norm.Pharmaceut.Que1, kmeans, method = "silhouette")
plot_grid(WSS, Silho)
```



The plotted charts show that in the elbow method line occurs when k=2, whereas the Silhouette Method produces k=5. I am using the k-means method with k=5.

```
#using k-means with k=5 for making clusters
set.seed(123)
KMe.Pharmaceut.Opt <- kmeans(norm.Pharmaceut.Que1, centers = 5, nstart = 50)
KMe.Pharmaceut.Opt$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
```

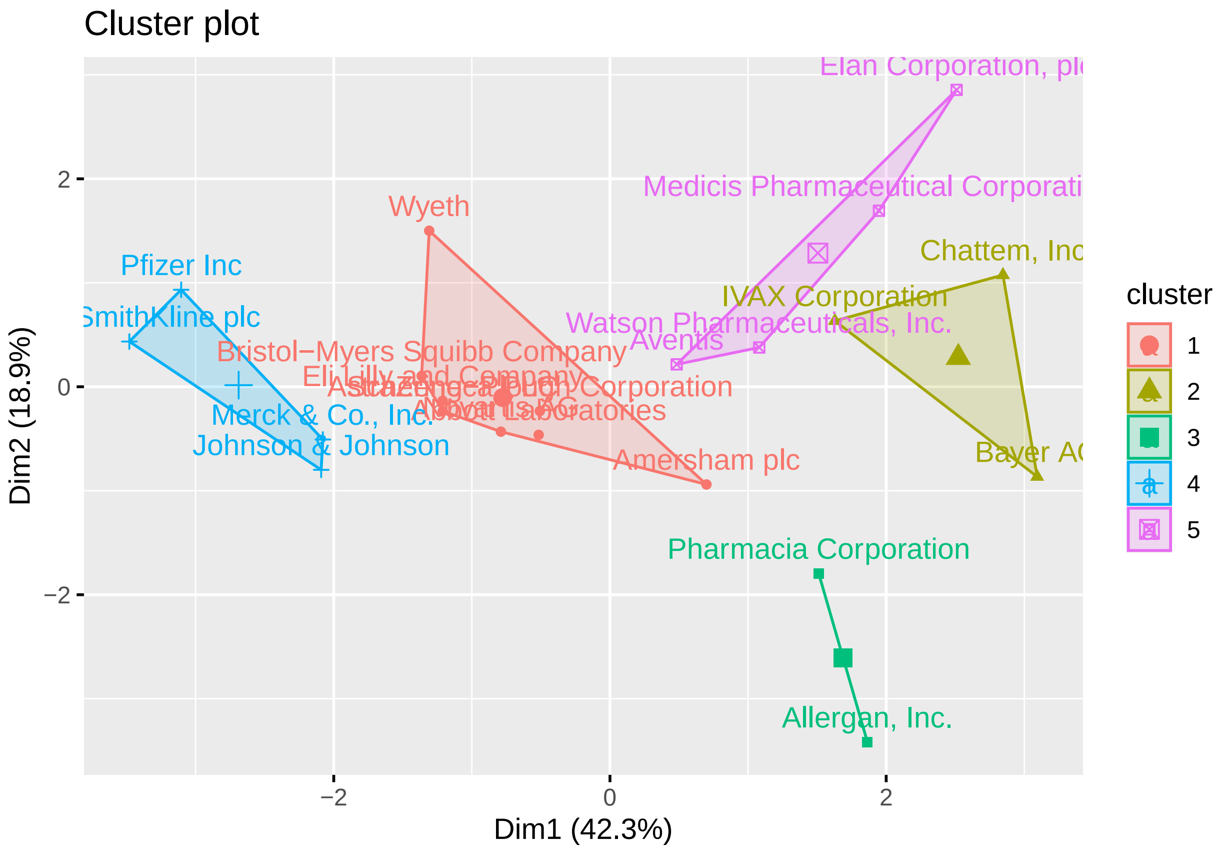
```
KMe.Pharmaceut.Opt$size
```

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

```
KMe.Pharmaceut.Opt$withinss
```

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

```
fviz_cluster(KMe.Pharmaceut.Opt, data = norm.Pharmaceut.Que1)
```



Using the data, we may define the five clusters depending on their distance from the cores. Cluste.4 has a high Market Capital, whereas Cluster n.2 has a high Beta and Cluste.5 does have a low Asset Turnover. We can also find out how big each cluster is. Cluste.1 has the most enterprises, whereas Cluste.3 has only two. The within-cluster sum of squared distances reveal information regarding data dispersion: cluste.1 (21.9) is less homogenous than cluste.3 (2.8). By visualizing the algorithm's output, we can observe 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. I choose to run the model again with only three clusters to acquire a better grasp of the cluster analysis, because with only two clusters, we feared losing some of the data's features.

```
#using k-means algorithm with k=3 for making clusters
set.seed(123)
KMe.Pharmaceut <- kmeans(norm.Pharmaceut.Que1, centers = 3, nstart = 50)
KMe.Pharmaceut$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
```

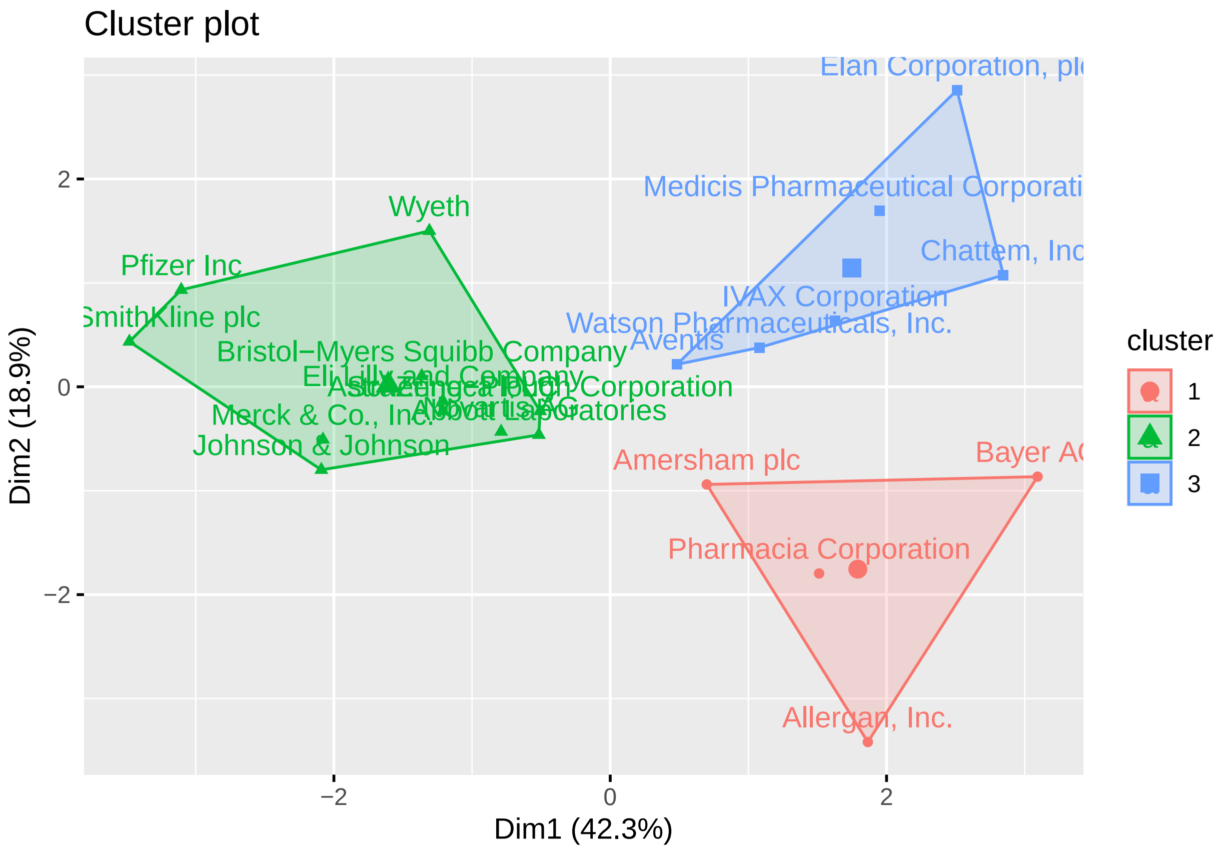
```
KMe.Pharmaceut$size
```

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

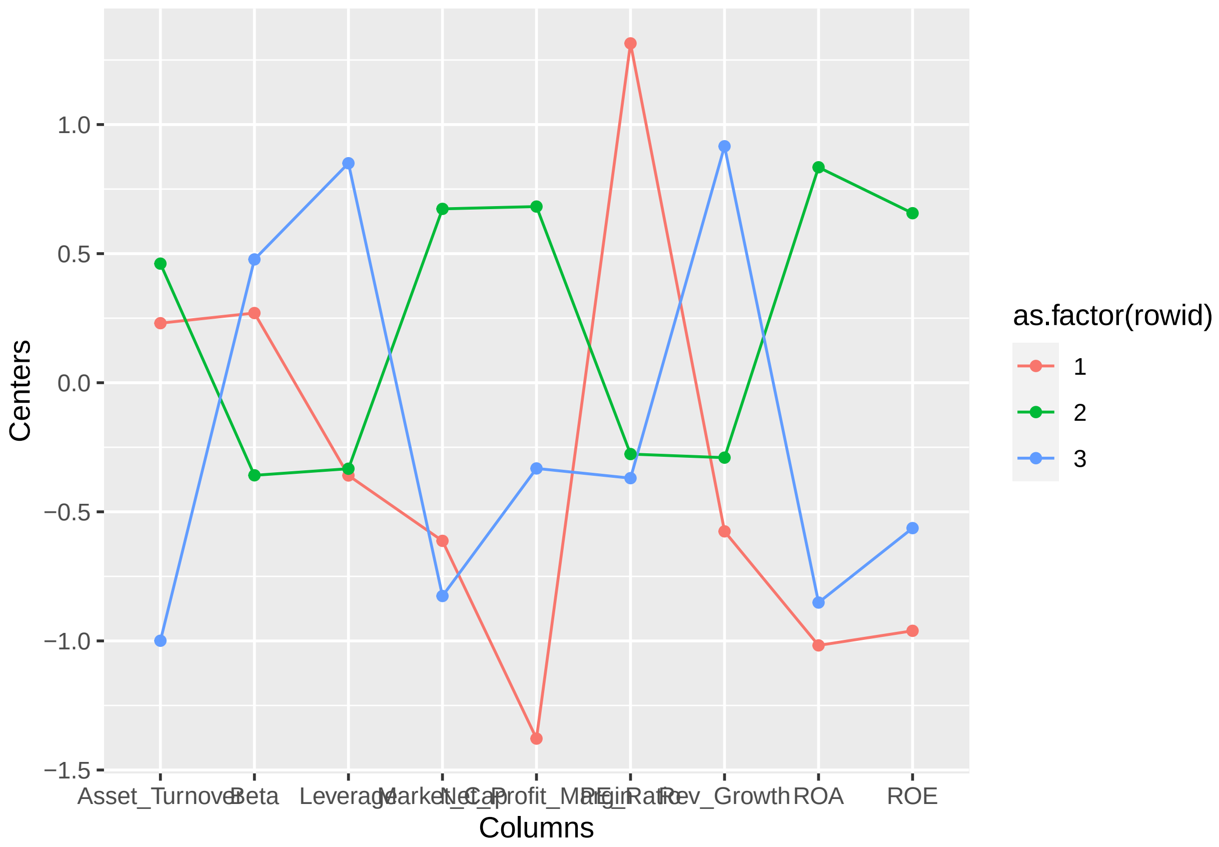
```
KMe.Pharmaceut$withinss
```

```
## [1] 20.54199 43.30886 32.14336
```

```
fviz_cluster(KMe.Pharmaceut, data = norm.Pharmaceut.Que1)
```



This facilitates the identification and management of the clusters in the analysis. We now have 4 data points in cluste.1, 11 data points in cluste.2, and 6 data points in cluste.3.



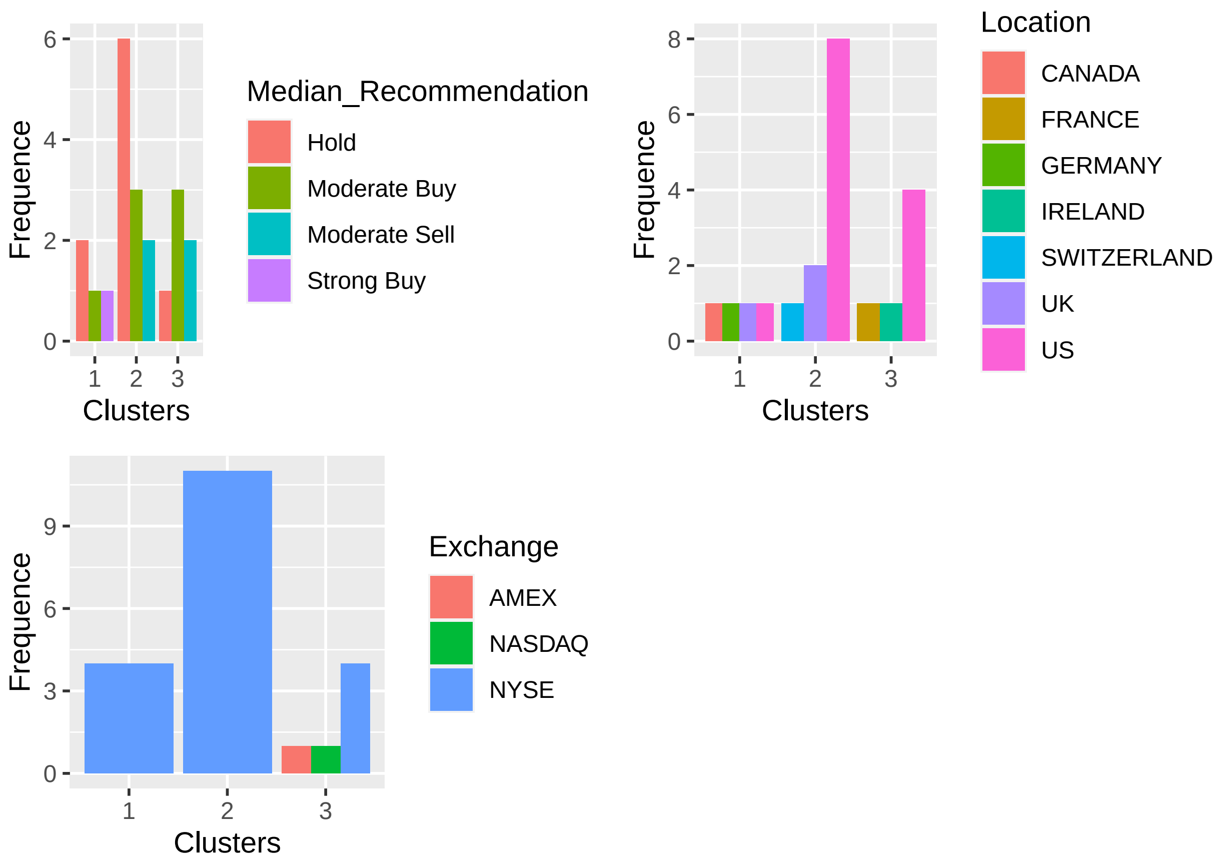
According to the second graphic, Companies in cluste.1 have a low Net Profit Margin and a high Price/Earnings ratio, whereas companies in cluste.2 have a low Asset Turnover and Return on Asset (ROA) but a high Leverage and Estimated Revenue Growth. Cluste.3 did not stand out in any of the parameters we looked at.

3)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
Pharmaceut.Que3 <-  Pharmaceut %>% select(c(11,12,13)) %>%
    mutate(Cluster = KMe.Pharmaceut$cluster)
```

```
#cluster plots
Med_Recom <- ggplot(Pharmaceut.Que3, mapping = aes(factor(Cluster), fill=Median_Recommendation)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Locat <- ggplot(Pharmaceut.Que3, mapping = aes(factor(Cluster), fill=Location)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Excha <- ggplot(Pharmaceut.Que3, mapping = aes(factor(Cluster), fill=Exchange)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
plot_grid(Med_Recom, Locat, Excha)
```



The graph plainly illustrates that the majority of the companies in cluste.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 cluste.2, we choose “Moderate Buy” shares, and we include just two companies whose stocks are listed on other exchanges or indexes (AMEX and NASDAQ). Cluste.1 shows that the four firms are located in four different countries, and their stocks are traded on the NYSE.

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

Thus, we can compile all of the data from the dataset and identify 3 distinct groups from the list of 21 pharmaceutical companies.

1)Cluste.1 is defined as “overvalued international firms” due to the following factors: international location, NYSE trading, low Net Profit Margin, and a high Price/Earnings ratio. These firms do business on many continents while raising capital on the world’s largest stock exchange (NYSE). They both have high financial market valuations that are not supported by their present earnings levels. If they do not want their stock price to collapse, they must invest and increase earnings to meet investors’ expectations.

2)Cluste.2 is categorized as a “growing and leveraged firm” because of the following characteristics: “Moderate buy” evaluations, low asset turnover and ROA, high leverage, and predicted revenue growth. Despite their current poor profitability and huge debt, they appear to be highly valued by investors willing to wait for future growth.

3)Cluste.3 qualifies as a “mature US firm” ,since it is US-based, listed on the NYSE, and has “Hold” ratings.