FML Assignment 4

Gloria Stephen

2022-11-06

First CSV file and Required Packages are loaded

```
library(caret)
```

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

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

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

```
## 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 4.2.2
```

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

```
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- 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 4.2.2
```

```
library(readr)
Pharmaceuticals <- read.csv("C:/Users/idast/Downloads/Pharmaceuticals.csv")
view(Pharmaceuticals)
head(Pharmaceuticals)
```

```
##   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
```

```
str(Pharmaceuticals)
```

```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PL
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

```
summary(Pharmaceuticals)
```

```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio          ROE             ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median :0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
##  Min.   :-3.17   Min.   : 2.6      Length:21              Length:21
##  1st Qu.: 6.38   1st Qu.:11.2      Class :character       Class :character
##  Median : 9.37   Median :16.1      Mode  :character       Mode  :character
##  Mean   :13.37   Mean   :15.7
##  3rd Qu.:21.87   3rd Qu.:21.1
##  Max.   :34.21   Max.   :25.5
##    Exchange
##  Length:21
##  Class :character
##  Mode  :character
##
##
##
```

```
dim(Pharmaceuticals)
```

```
## [1] 21 14
```

```
colMeans(is.na(Pharmaceuticals))
```

```
##                Symbol                  Name            Market_Cap
##                     0                     0                     0
##                  Beta              PE_Ratio                   ROE
##                     0                     0                     0
##                   ROA        Asset_Turnover              Leverage
##                     0                     0                     0
##            Rev_Growth     Net_Profit_Margin Median_Recommendation
##                     0                     0                     0
##              Location              Exchange
##                     0                     0
```

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

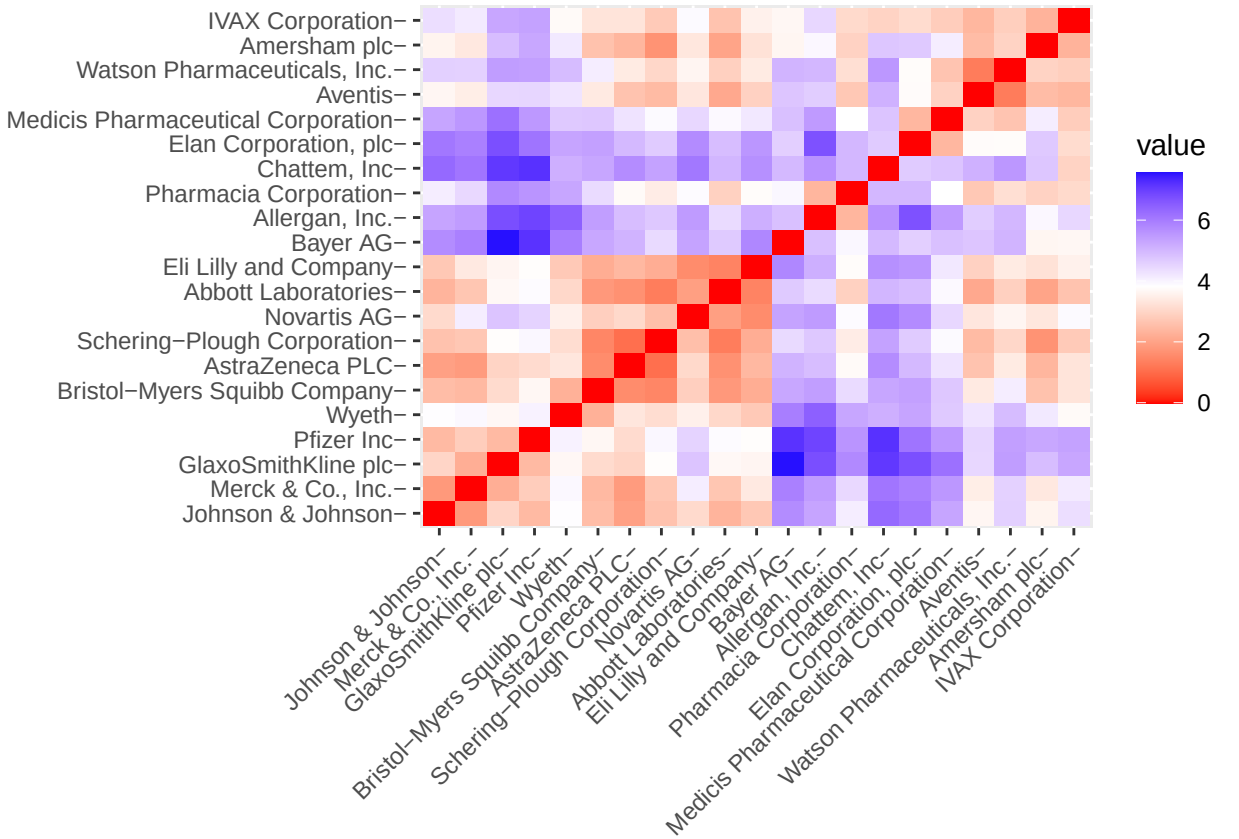
#1. Focusing on the numericals

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

Normalizing and Clustering the data by measuring and plotting

The default euclidean distance metric, which is scale-sensitive and requires data modification, is used.

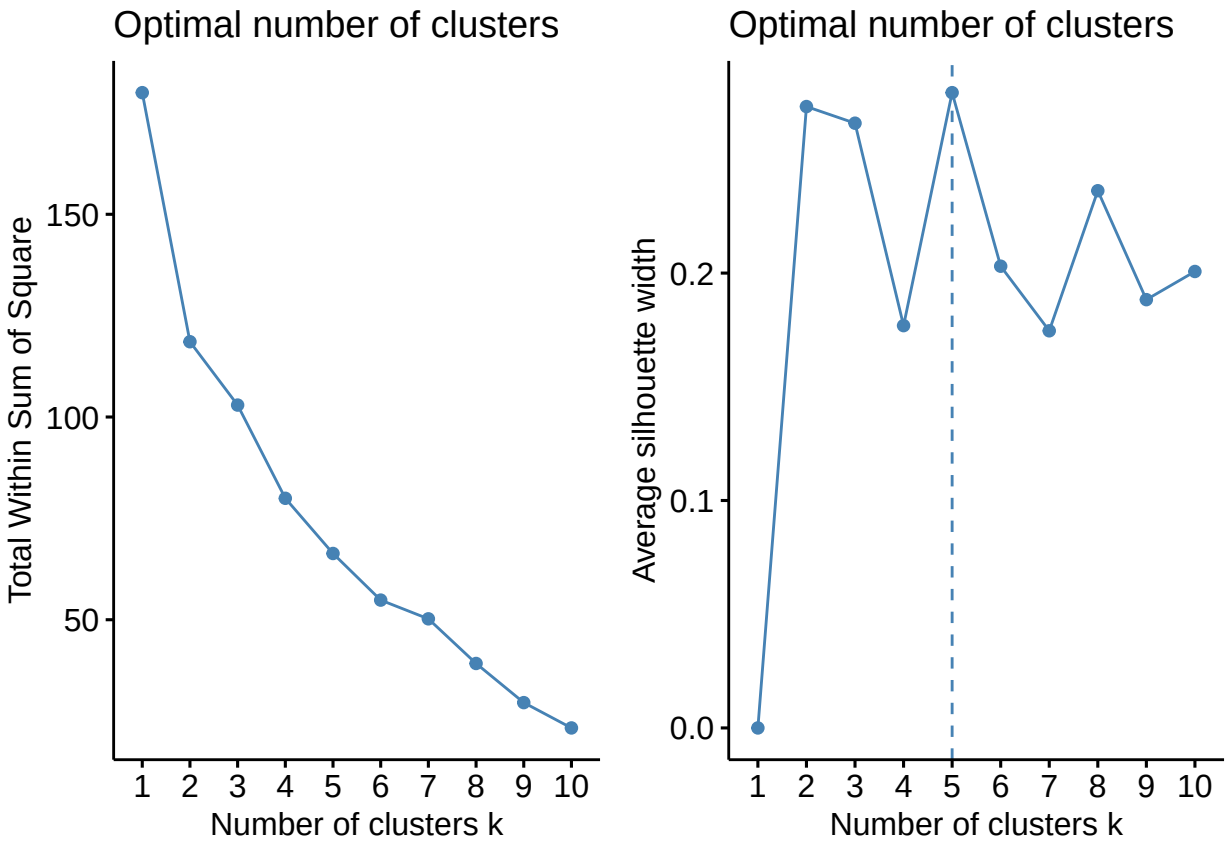
```
normalization.Pharmaceuticals.Que1 <- scale(Pharmaceuticals.Que1)
distance <- get_dist(normalization.Pharmaceuticals.Que1)
fviz_dist(distance)
```



The color intensity changes as distance increases in the graph.Since it represents the distance between two observations, the diagonal, as we would anticipate, has a value of zero.

##The optimal K value The Elbow chart and the Silhouette Method are two of the most efficient methods for determining the number of clusters for the k-means model when there are no external factors.The first illustration demonstrates that as more clusters are added, cluster heterogeneity decreases.The latter evaluates how similar an object is to its cluster in comparison to other clusters. Using the

```
WSS_1 <- fviz_nbclust(normalization.Pharmaceuticals.Que1, kmeans, method = "wss")
Silhouette <- fviz_nbclust(normalization.Pharmaceuticals.Que1, kmeans, method = "silhouette")
plot_grid(WSS_1, Silhouette)
```



According to the plotted charts, the elbow approach creates a line when k=2, whereas the silhouette method results in k=5. The k-means approach I'm using has k=5.

```
#using k-means with k=5 for making clusters
set.seed(101)
KMeans.Pharmaceuticals.Opt_1 <- kmeans(normalization.Pharmaceuticals.Que1, centers = 5, nstart = 50)
KMeans.Pharmaceuticals.Opt_1$centers
```

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

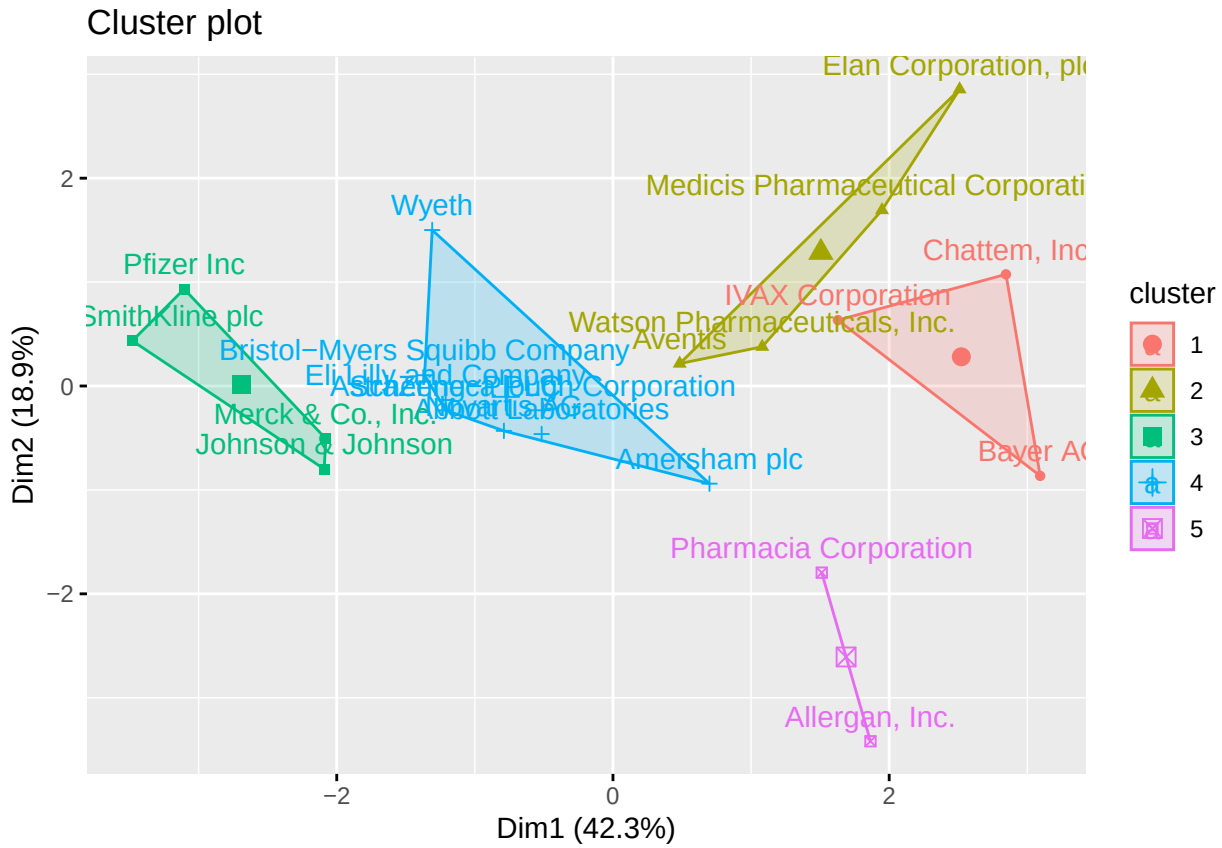
```
KMeans.Pharmaceuticals.Opt_1$size
```

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

```
KMeans.Pharmaceuticals.Opt_1$withinss
```

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

```
fviz_cluster(KMeans.Pharmaceuticals.Opt_1, data = normalization.Pharmaceuticals.Que1)
```



Using the data, we can determine the five clusters based on how far off they are from the cores. While Cluster e.5 does not have a high Asset Turnover, Cluster n.2 has a high Beta. Market Capital is high for Cluste.4. We may also quantify the size of each cluster. Cluste.1 has the most companies, whilst Cluste just has two. 3. The within-cluster sum of squared distances reveals data dispersion: cluste.1 (21.9) is less homogenous than cluste.3 (2.8). The output of the algorithm reveals the five groups into which the data has been separated.

#2.Interpretation of clusters using numerical variables With only two clusters, we worried losing some of the properties of the data, so I decided to run the model again with only three clusters to better understand the cluster analysis.

```
#using k-means with k=3 for making clusters
set.seed(102)
KMeans.Pharmac_1 <- kmeans(normalization.Pharmaceuticals.Que1, centers = 3, nstart = 50)
KMeans.Pharmac_1$centers
```

```
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.8261772  0.4775991 -0.3696184 -0.5631589 -0.8514589     -0.9994088
## 2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
```

```
## 3 -0.6125361  0.2698666  1.3143935 -0.9609057 -1.0174553      0.2306328
##     Leverage Rev_Growth Net_Profit_Margin
## 1  0.8502201  0.9158889        -0.3319956
## 2 -0.3331068 -0.2902163         0.6823310
## 3 -0.3592866 -0.5757385        -1.3784169
```

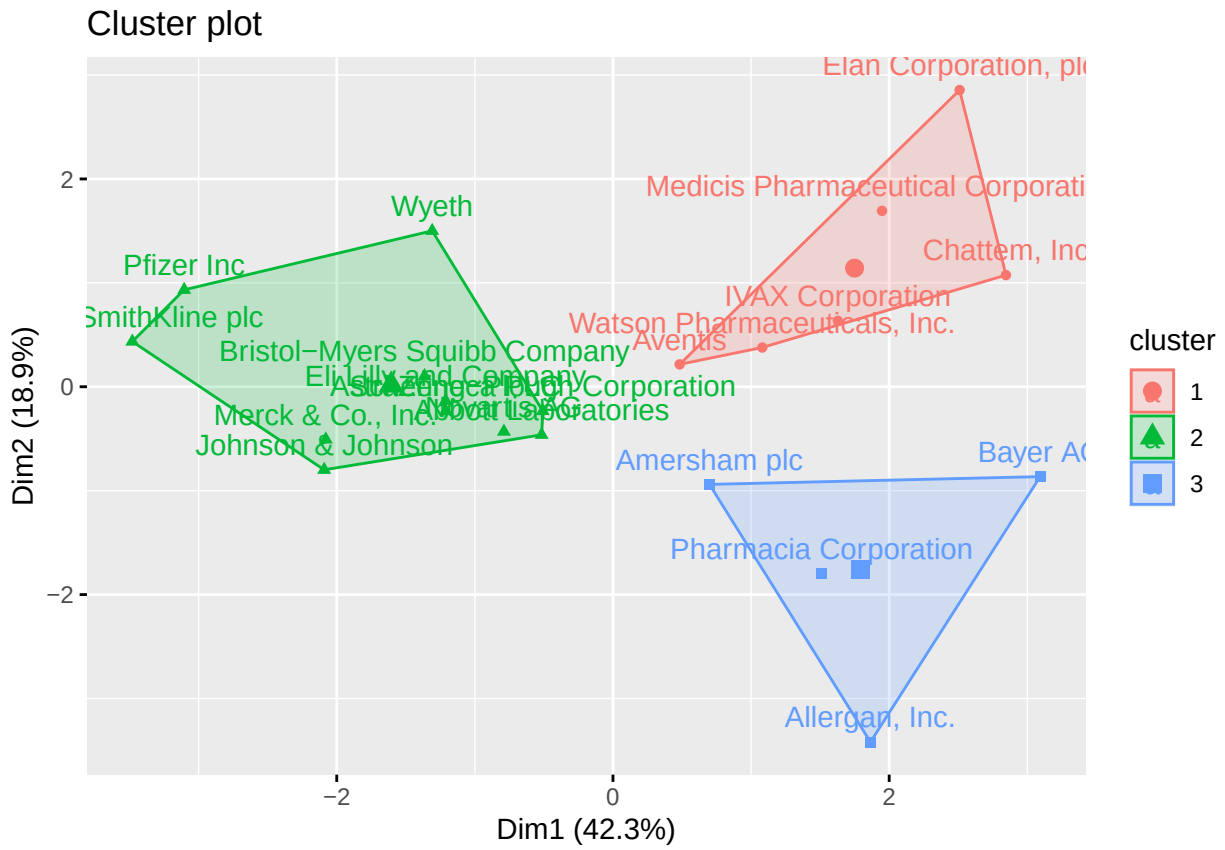
```
KMeans.Pharmac_1$size
```

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

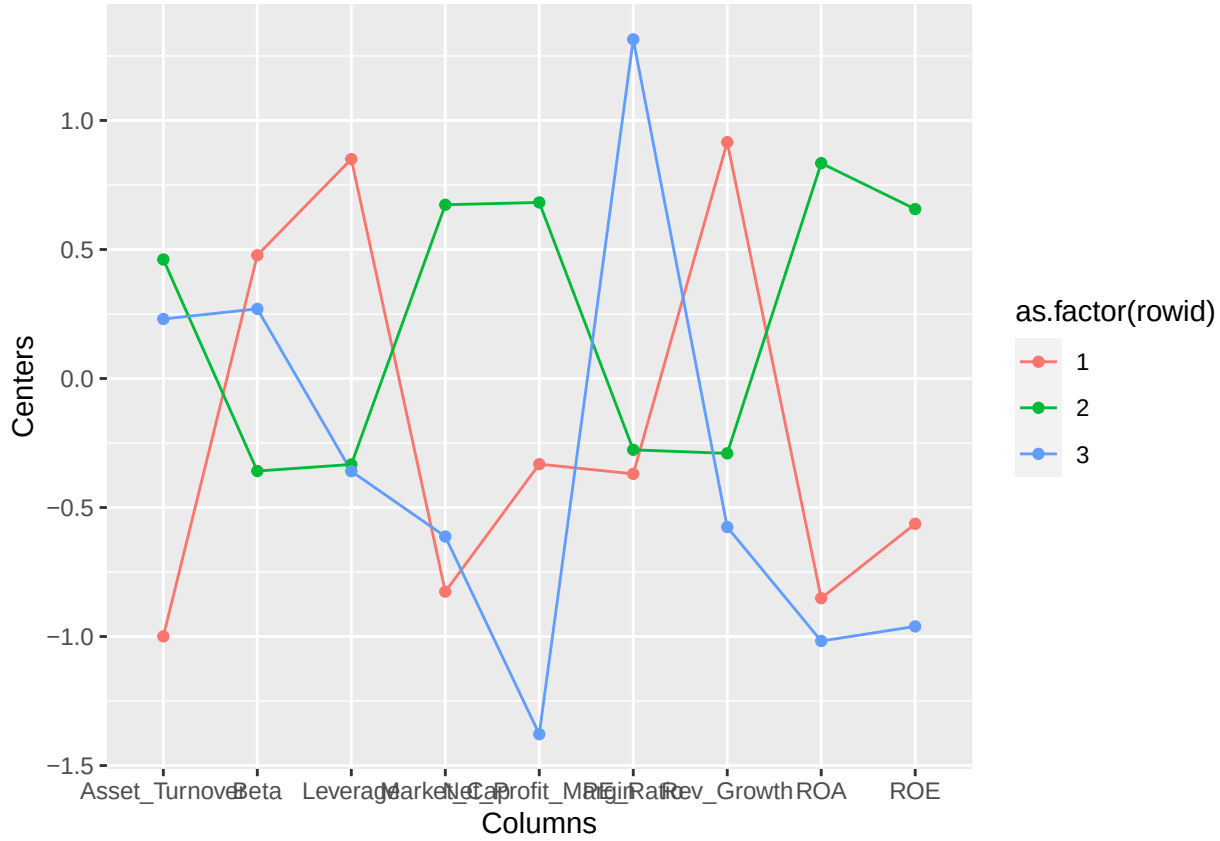
```
KMeans.Pharmac_1$withinss
```

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

```
fviz_cluster(KMeans.Pharmac_1, data = normalization.Pharmaceuticals.Que1)
```



This means that managing and identifying clusters during analysis is much easier. There are currently 4 data points in cluste. 6, 11, and 11 data items in cluste.3 respectively.

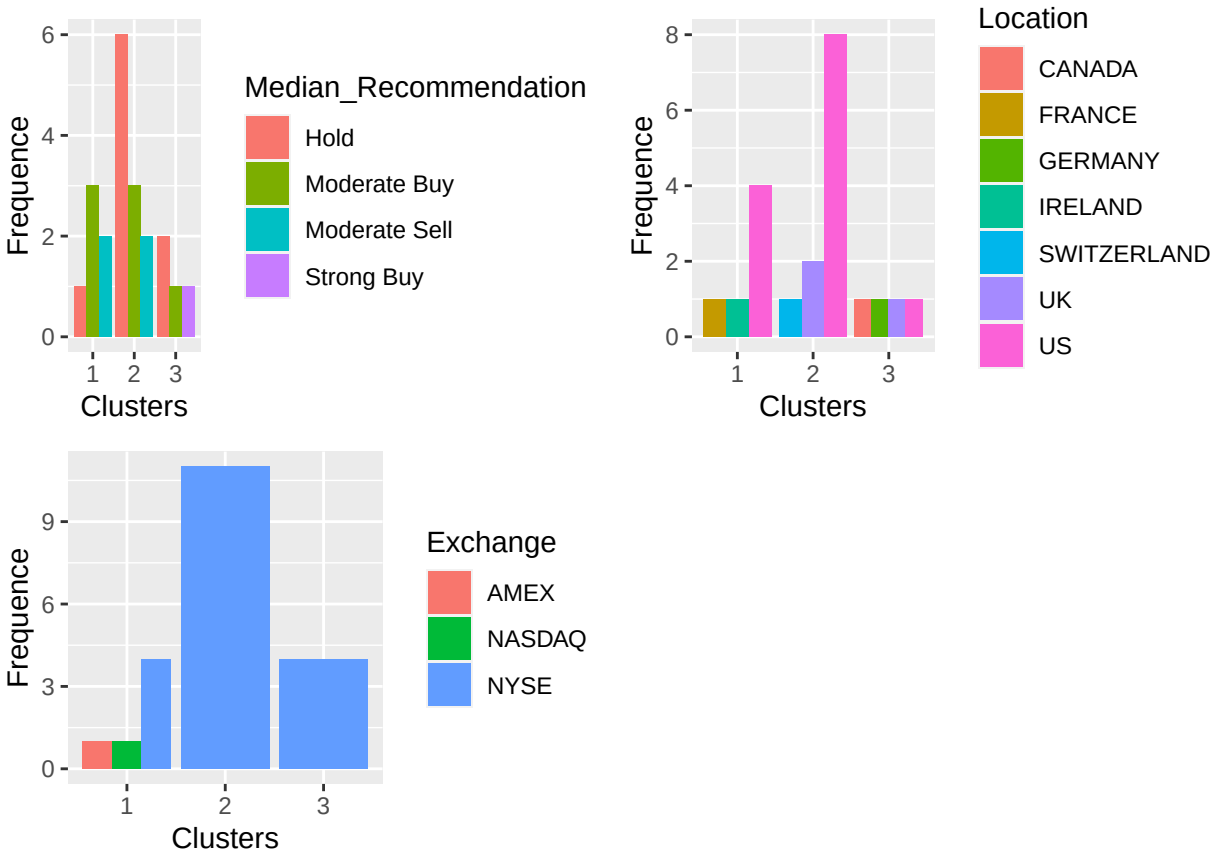


The second graph shows that businesses in cluste.1 have a low net profit margin and a high price to earnings ratio, while businesses in cluste.2 have a low asset turnover and return on asset (ROA), but a high leverage and expected revenue growth. With regard to any of the parameters we examined, Cluste.3 did not stand out.

#3. Pattern in clusters with respect to numerical variables The remaining three category factors to be considered are Stock Exchange, Location, and Median Recommendation. To visualize the distribution of businesses grouped by clusters and to identify any trends in the data, I choose to utilize bar charts.

```
Pharmaceuticals.Que_3 <-  Pharmaceuticals %>% select(c(11,12,13)) %>%
mutate(Cluster = KMeans.Pharmac_1$cluster)
```

```
Median_Recom <- ggplot(Pharmaceuticals.Que_3, mapping = aes(factor(Cluster), fill=Median_Recommendation
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Location_0 <- ggplot(Pharmaceuticals.Que_3, mapping = aes(factor(Cluster), fill=Location)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Exchange_0 <- ggplot(Pharmaceuticals.Que_3, mapping = aes(factor(Cluster), fill=Exchange)) +
  geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
plot_grid(Median_Recom, Location_0, Exchange_0)
```



The graph makes it clear that most of the businesses in cluste.3 are American-based and all have a spread advice to keep their stock. The New York Stock Exchange is where they are all exchanged. We choose "Moderate Buy" shares for cluste.2 and only take into account two businesses whose equities are traded on other exchanges or indexes (AMEX and NASDAQ). The four businesses are located in four separate nations, as shown by Cluste.1, and their stocks are listed on the NYSE.

#4. Naming for each cluster using the variables in the dataset.

Hence, using the entire dataset of information, we can separate the list of 21 pharmaceutical businesses into three unique categories.

1) Cluster 1-Due to the following characteristics: international location, NYSE trading, low Net Profit Margin, and a high Price/Earnings ratio, Cluster_1 is referred to as "overvalued foreign enterprises." These companies operate across several continents and raise funds on the biggest stock exchange in the world (NYSE). Both of them are valued highly on the financial market, which is not supported by their current earnings levels. They must invest and boost earnings to satisfy investors if they do not want their stock price to plummet.

2) Cluster 2-Due to the following traits, Cluster_2 is labeled as a "growing and leveraged firm": "Moderate buy" assessments, low asset turnover and ROA, high leverage, and anticipated revenue growth. Investors who are ready to wait for future development tend to esteem them highly despite their poor profitability and significant debt.

3)Cluster_3- Due to its US location, NYSE listing, and "Hold" ratings, Cluster_3 is considered a "mature US corporation."