Assignment\_4v1a

Bryan Mack

2025-06-14

Purpose: The purpose of this assignment is to use k-Means for clustering.

Directions: 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)
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

library(tidyverse) # data manipulation

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

#install.packages("factoextra") # if necessary  
library(factoextra) # clustering algorithms & visualization

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

library(ISLR)  
library(flexclust)  
set.seed(123)

#setwd("D:/R\_DATA/Assignment 4")  
getwd()

## [1] "D:/R\_DATA/Assignment 4"

#setwd() #used to set R working directory   
# i.e. D:/R\_DATA or (note / for windows OS)   
# setwd("D:\\\\R\_DATA\\\\")

#df<-Auto[,c(1,6)]  
#summary(df)  
#df<-Pharmaceuticals.csv  
#df <- read.csv("Pharmaceuticals.csv", stringsAsFactors = FALSE)  
getwd()

## [1] "D:/R\_DATA/Assignment 4"

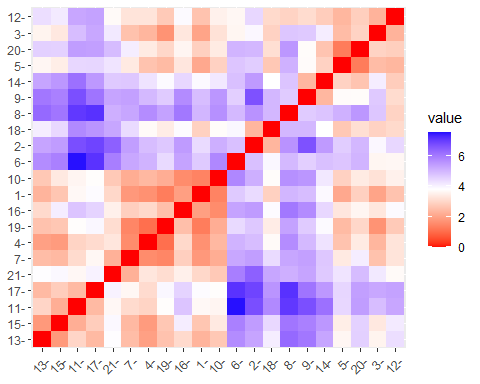
df <- read\_csv("Pharmaceuticals.csv", col\_select = c(Market\_Cap, Beta, PE\_Ratio, ROE, ROA,Asset\_Turnover, Leverage, Rev\_Growth, Net\_Profit\_Margin))

## Rows: 21 Columns: 9  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (9): Market\_Cap, Beta, PE\_Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#df<-Pharmaceuticals.csv  
summary(df)

## Market\_Cap Beta PE\_Ratio ROE   
## Min. : 0.41 Min. :0.1800 Min. : 3.60 Min. : 3.9   
## 1st Qu.: 6.30 1st Qu.:0.3500 1st Qu.:18.90 1st Qu.:14.9   
## Median : 48.19 Median :0.4600 Median :21.50 Median :22.6   
## Mean : 57.65 Mean :0.5257 Mean :25.46 Mean :25.8   
## 3rd Qu.: 73.84 3rd Qu.:0.6500 3rd Qu.:27.90 3rd Qu.:31.0   
## Max. :199.47 Max. :1.1100 Max. :82.50 Max. :62.9   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. : 1.40 Min. :0.3 Min. :0.0000 Min. :-3.17   
## 1st Qu.: 5.70 1st Qu.:0.6 1st Qu.:0.1600 1st Qu.: 6.38   
## Median :11.20 Median :0.6 Median :0.3400 Median : 9.37   
## Mean :10.51 Mean :0.7 Mean :0.5857 Mean :13.37   
## 3rd Qu.:15.00 3rd Qu.:0.9 3rd Qu.:0.6000 3rd Qu.:21.87   
## Max. :20.30 Max. :1.1 Max. :3.5100 Max. :34.21   
## Net\_Profit\_Margin  
## Min. : 2.6   
## 1st Qu.:11.2   
## Median :16.1   
## Mean :15.7   
## 3rd Qu.:21.1   
## Max. :25.5

df <- scale(df, center = TRUE, scale = TRUE)  
distance <- get\_dist(df)  
fviz\_dist(distance)



k4 <- kmeans(df, centers = 4, nstart = 25) # k = 4, number of restarts = 25  
  
# Visualize the output  
  
k4$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 1.69558112 -0.1780563 -0.1984582 1.2349879 1.3503431 1.153164e+00  
## 2 -0.03142211 -0.4360989 -0.3172485 0.1950459 0.4083915 1.729746e-01  
## 3 -0.82617719 0.4775991 -0.3696184 -0.5631589 -0.8514589 -9.994088e-01  
## 4 -0.52462814 0.4451409 1.8498439 -1.0404550 -1.1865838 1.480297e-16  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.4680782 0.4671788 0.5912425  
## 2 -0.2744931 -0.7041516 0.5569544  
## 3 0.8502201 0.9158889 -0.3319956  
## 4 -0.3443544 -0.5769454 -1.6095439

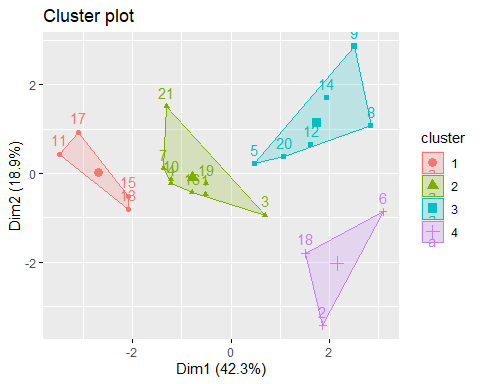
k4$size # Number of cars in each cluster

## [1] 4 8 6 3

k4$cluster[10] # Identify the cluster of the 120th observation as an example

## [1] 2

fviz\_cluster(k4, data = df) # Visualize the output



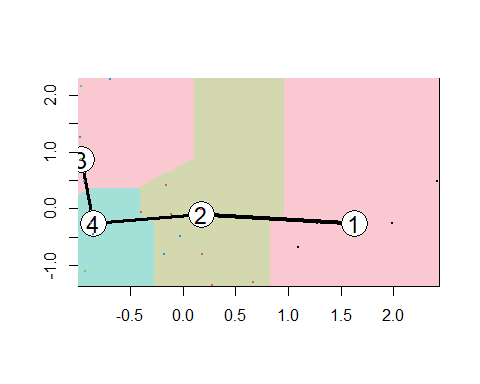
#install.packages("flexclust")  
#library(flexclust)  
set.seed(123)  
#kmeans clustering, using manhattan distance  
k4 = kcca(df, k=4, kccaFamily("kmedians"))  
k4

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = df, k = 4, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4   
## 4 7 3 7

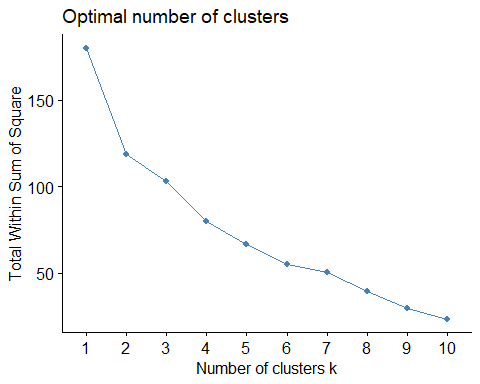
#Apply the predict() function  
clusters\_index <- predict(k4)  
dist(k4@centers)

## 1 2 3  
## 2 2.608581   
## 3 5.395288 3.872647   
## 4 4.664586 2.864999 3.010141

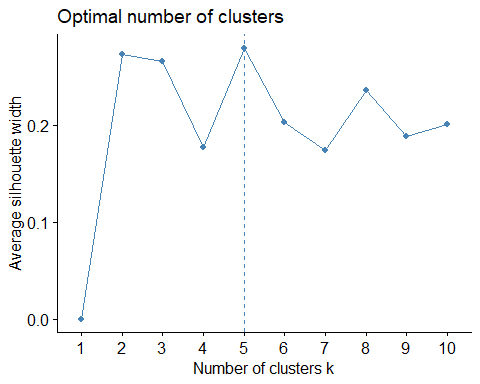
image(k4)  
points(df, col=clusters\_index, pch=19, cex=0.3)



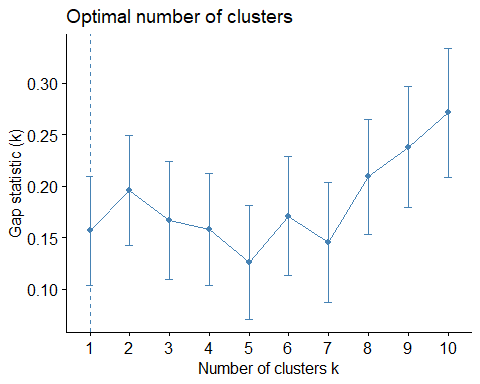
library(tidyverse) # data manipulation  
library(factoextra) # clustering & visualization  
library(ISLR)  
set.seed(123)  
  
df <- scale(df)  
fviz\_nbclust(df, kmeans, method = "wss")



fviz\_nbclust(df, kmeans, method = "silhouette")



fviz\_nbclust(df, kmeans, method = "gap\_stat")



Cluster in K5

set.seed(123)  
#kmeans clustering, using manhattan distance  
k5 = kcca(df, k=5, kccaFamily("kmedians"))  
k5

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = df, k = 5, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4 5   
## 3 4 3 7 4

k5 <- kmeans(df, centers = 5, nstart = 25) # k = 4, number of restarts = 25  
  
# Visualize the output  
  
k5$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 2 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 3 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 4 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 5 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.14170336 -0.1168459 -1.416514761  
## 2 1.36644699 -0.6912914 -1.320000179  
## 3 0.06308085 1.5180158 -0.006893899  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.46807818 0.4671788 0.591242521

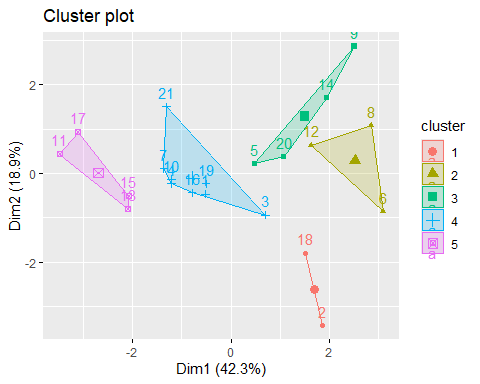
k5$size # Number of cars in each cluster

## [1] 2 3 4 8 4

k5$cluster[5] # Identify the cluster of the 120th observation as an example

## [1] 3

fviz\_cluster(k5, data = df) # Visualize the output



set.seed(123)  
#kmeans clustering, using manhattan distance  
k3 = kcca(df, k=3, kccaFamily("kmedians"))  
k3

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = df, k = 3, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3   
## 6 9 6

k3 <- kmeans(df, centers = 3, nstart = 25) # k = 4, number of restarts = 25  
  
# Visualize the output  
  
k3$centers # output the 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

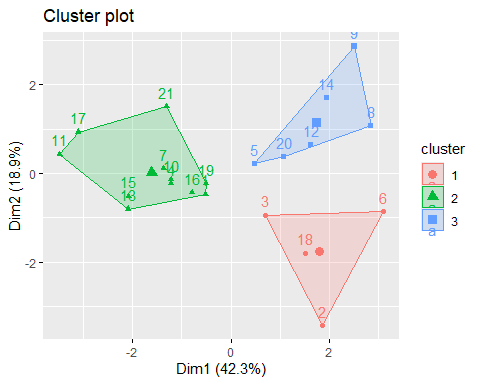
k3$size # Number of cars in each cluster

## [1] 4 11 6

k3$cluster[3] # Identify the cluster of the 120th observation as an example

## [1] 1

fviz\_cluster(k3, data = df) # Visualize the output



k3$centers # output the 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

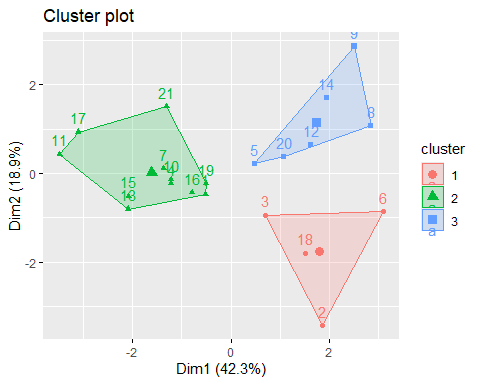
k3$betweenss # Number of cars in each cluster

## [1] 84.0058

k3$cluster[120] # Identify the cluster of the 120th observation as an example

## [1] NA

fviz\_cluster(k3, data = df) # Visualize the output



Tried to change weights of numeric variables.

df1 <- read\_csv("Pharmaceuticals.csv", col\_select = c(Market\_Cap, Beta, PE\_Ratio, ROE, ROA,Asset\_Turnover, Leverage, Rev\_Growth, Net\_Profit\_Margin))

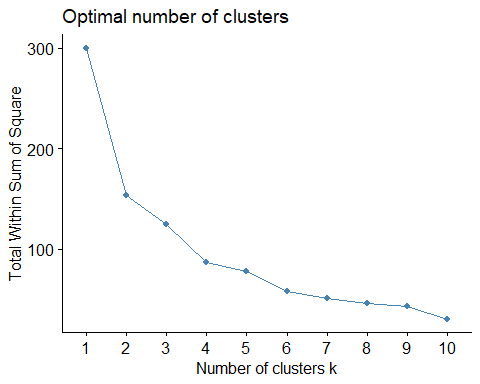
## Rows: 21 Columns: 9  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (9): Market\_Cap, Beta, PE\_Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

df1\_scaled <- scale(df1, center = TRUE, scale = TRUE)  
df1\_scaled <- as.data.frame(scale(df1))

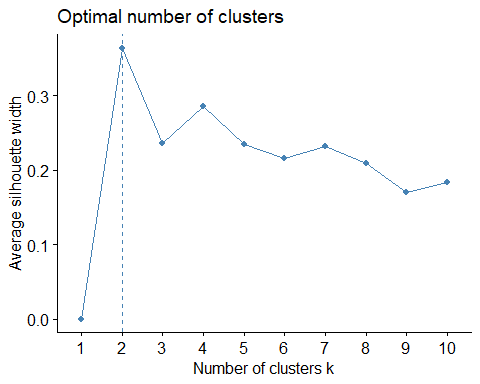
weights <- c(  
 Market\_Cap = 2.0, # Emphasize company size  
 Beta = 0.5, # Lower weight for risk  
 PE\_Ratio = 0.5, # Lower weight for valuation  
 ROE = 2.0, # Emphasize profitability  
 ROA = 1.0, # Moderate weight for efficiency  
 Asset\_Turnover = 0.5, # Lower weight for operational metric  
 Leverage = 0.5, # Lower weight for debt  
 Rev\_Growth = 1.0, # Moderate weight for growth  
 Net\_Profit\_Margin = 2.0 # Emphasize profitability  
)

df1\_weighted <- df1\_scaled  
for (col in names(weights)) {  
 df1\_weighted[[col]] <- df1\_scaled[[col]] \* weights[col]  
}

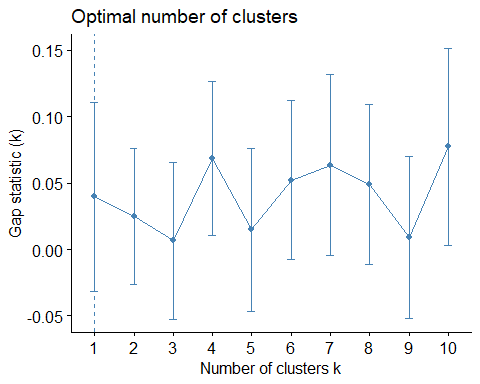
fviz\_nbclust(df1\_weighted, kmeans, method = "wss")



fviz\_nbclust(df1\_weighted, kmeans, method = "silhouette")



fviz\_nbclust(df1\_weighted, kmeans, method = "gap\_stat")



set.seed(123)  
#kmeans clustering, using manhattan distance  
k2\_weighted = kcca(df1\_weighted, k=2, kccaFamily("kmedians"))  
k2\_weighted

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = df1\_weighted, k = 2, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2   
## 4 17

k2\_weighted <- kmeans(df1\_weighted, centers = 2, nstart = 25) # k = 2, number of restarts = 25  
  
# Visualize the output  
  
k2\_weighted$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -1.419834 0.1765388 0.1198884 -1.351708 -0.7868256 -0.2096662  
## 2 1.561818 -0.1941927 -0.1318773 1.486878 0.8655082 0.2306328  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 0.1324781 0.2506303 -1.311886  
## 2 -0.1457259 -0.2756933 1.443074

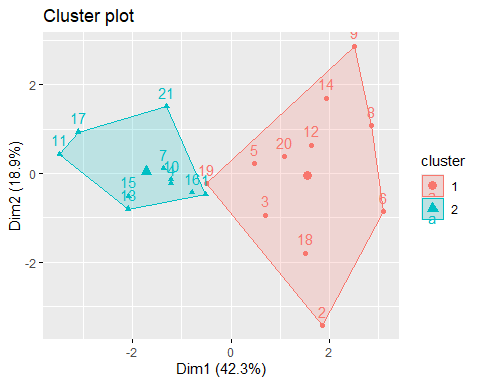
k2\_weighted$size # Number of cars in each cluster

## [1] 11 10

k2\_weighted$cluster[120] # Identify the cluster of the 120th observation as an example

## [1] NA

fviz\_cluster(k2\_weighted, data = df1\_weighted) # Visualize the output



set.seed(123)  
#kmeans clustering, using manhattan distance  
k3\_weighted = kcca(df1\_weighted, k=3, kccaFamily("kmedians"))  
k3\_weighted

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = df1\_weighted, k = 3, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3   
## 4 14 3

k3\_weighted <- kmeans(df1\_weighted, centers = 3, nstart = 25) # k = 4, number of restarts = 25  
  
# Visualize the output  
  
k3\_weighted$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 3.391162 -0.08902817 -0.09922912 2.4699758 1.3503431 0.57658201  
## 2 0.178538 -0.23091682 -0.16043075 0.6521784 0.5396003 0.03294754  
## 3 -1.481442 0.19725305 0.15199317 -1.4445152 -0.9178575 -0.25369608  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.2340391 0.4671788 1.182485  
## 2 -0.1279901 -0.7230135 1.468763  
## 3 0.1832087 0.3192379 -1.501128

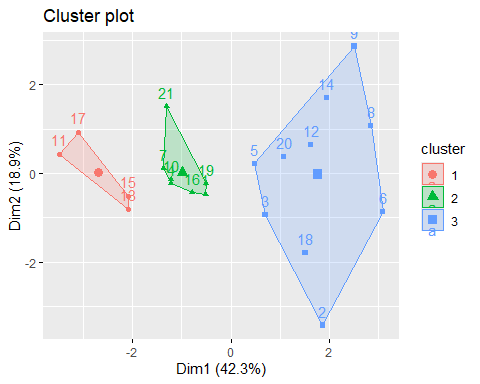
k3\_weighted$size # Number of cars in each cluster

## [1] 4 7 10

k3\_weighted$cluster[120] # Identify the cluster of the 120th observation as an example

## [1] NA

fviz\_cluster(k3\_weighted, data = df1\_weighted) # Visualize the output



Use cluster analysis to explore and analyze the given dataset as follows:

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.

I tried several different combinations in clustering including modifying the weights of the data. After I modified the weights I was shown in “fviz\_nbclust(df1\_weighted, kmeans, method =”silhouette”)” the optimal number of clusters was 2 however even after I think I weighted the data to increase the size of the market cap (Market\_Cap = 2.0, # Emphasize company size) relative to other factors I still found exceptions in the clustering. Using the weighted 3 cluster I see in the cluster 1 (assuming the number is the row in the data) 11 - Eli Lilly and Company 17 - Novartis AG 15 - Medicis Pharmaceutical Corporation (market cap only 1.2) 18 - Pfizer Inc

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

Based on the output I am getting I do not think that all of the numerical data should be used for the analysis and that a more specific goal should be defined before the variables are selected for analysis.

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

I was most interested in Median\_Recommendation however I was not able to see a clear relation between the clusters and the non-numeric data (Median\_Recommendation, Location, Exchange)

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

Using weighting I was trying to keep all of the variables and expected to see the following 3 categories:

Cluster 1: Large, highly profitable firms. Cluster 2: Mid-sized, growth-oriented firms. Cluster 3: Smaller or high-risk firms.