FML ASSIGNMENT - 4

Sai Harshitha Akula

2023-11-19

#install.packages("factoextra")  
#install.packages("cowplot")  
#install.packages("flexclust")  
#install.packages("cluster")  
#install.packages("NbClust")

#Loading required packages.  
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.2

## Loading required package: ggplot2

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

library(cowplot)

## Warning: package 'cowplot' was built under R version 4.3.2

library(caret)

## Warning: package 'caret' was built under R version 4.3.2

## Loading required package: lattice

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 4.3.2

## Warning: package 'tidyr' was built under R version 4.3.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.0  
## ✔ readr 2.1.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ purrr::lift() masks caret::lift()  
## ✖ lubridate::stamp() masks cowplot::stamp()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.3.2

## Loading required package: grid  
## Loading required package: modeltools  
## Loading required package: stats4

library(cluster)

## Warning: package 'cluster' was built under R version 4.3.2

library(NbClust)

#Importing the dataset.  
PharmaceuticalsData <- read.csv("C:/Users/saiha/OneDrive/Documents/R PROGRAMMING/Pharmaceuticals.csv")

head(PharmaceuticalsData)

## 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

colMeans(is.na(PharmaceuticalsData))

## 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(PharmaceuticalsData) <- PharmaceuticalsData[,2]  
PharmaceuticalsData <- PharmaceuticalsData[,-2]

#Summary of the dataset.  
summary(PharmaceuticalsData)

## Symbol Market\_Cap Beta PE\_Ratio   
## Length:21 Min. : 0.41 Min. :0.1800 Min. : 3.60   
## Class :character 1st Qu.: 6.30 1st Qu.:0.3500 1st Qu.:18.90   
## Mode :character Median : 48.19 Median :0.4600 Median :21.50   
## Mean : 57.65 Mean :0.5257 Mean :25.46   
## 3rd Qu.: 73.84 3rd Qu.:0.6500 3rd Qu.:27.90   
## Max. :199.47 Max. :1.1100 Max. :82.50   
## ROE ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. : 3.9 Min. : 1.40 Min. :0.3 Min. :0.0000 Min. :-3.17   
## 1st Qu.:14.9 1st Qu.: 5.70 1st Qu.:0.6 1st Qu.:0.1600 1st Qu.: 6.38   
## Median :22.6 Median :11.20 Median :0.6 Median :0.3400 Median : 9.37   
## Mean :25.8 Mean :10.51 Mean :0.7 Mean :0.5857 Mean :13.37   
## 3rd Qu.:31.0 3rd Qu.:15.00 3rd Qu.:0.9 3rd Qu.:0.6000 3rd Qu.:21.87   
## Max. :62.9 Max. :20.30 Max. :1.1 Max. :3.5100 Max. :34.21   
## Net\_Profit\_Margin Median\_Recommendation Location Exchange   
## Min. : 2.6 Length:21 Length:21 Length:21   
## 1st Qu.:11.2 Class :character Class :character Class :character   
## Median :16.1 Mode :character Mode :character Mode :character   
## Mean :15.7   
## 3rd Qu.:21.1   
## Max. :25.5

dim(PharmaceuticalsData)

## [1] 21 13

colMeans(is.na(PharmaceuticalsData))

## Symbol Market\_Cap Beta   
## 0 0 0   
## PE\_Ratio ROE ROA   
## 0 0 0   
## Asset\_Turnover Leverage Rev\_Growth   
## 0 0 0   
## Net\_Profit\_Margin Median\_Recommendation Location   
## 0 0 0   
## Exchange   
## 0

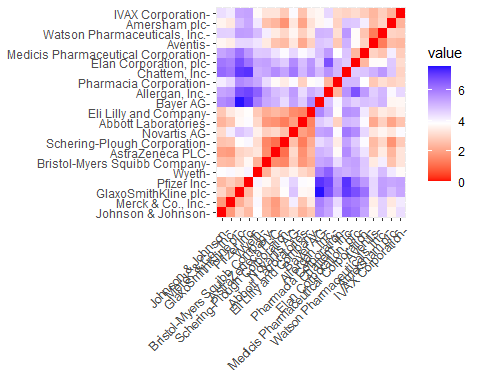
#a) Performing a cluster analysis involves making several decisions to ensure the process is meaningful and relevant to the underlying data structure. In the context of clustering 21 firms using only numerical variables (1 to 9).

#In our analysis, we narrow our focus to a subset of the complete dataset, specifically emphasizing numerical variables. This strategic decision allows us to hone in on quantitative aspects and exclude non-numeric features, streamlining the clustering process

#Excluding the variable "Symbol" and the final three categorical variables in the dataset.  
PharmaceuticalsData1 <- PharmaceuticalsData[,-c(1,11:13)]

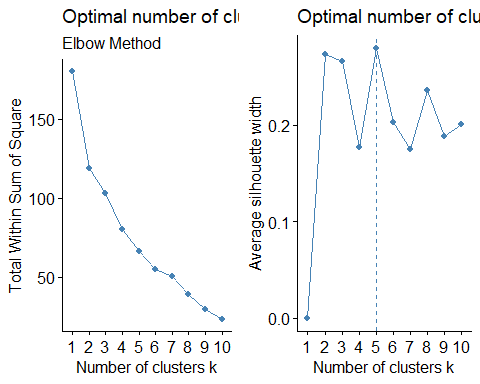
#In this step, the dissimilarity between each observation is computed. To ensure accurate results, the data needs to be adjusted because the default Euclidean distance measure, which is sensitive to scale, is utilized.

#Normalising the data  
norm.PharmaceuticalsData1 <- scale(PharmaceuticalsData1)  
  
#Measuring and plotting the distance  
dist <- get\_dist(norm.PharmaceuticalsData1)  
fviz\_dist(dist)



#The graph shows that as we move along the diagonal, the color becomes less intense, reaching zero at the center because it represents the distance between two observations.  
  
#The Elbow chart and the Silhouette Method are helpful tools for figuring out how many clusters to use in a k-means model when no external factors are influencing the decision. The Elbow chart shows how increasing the number of clusters reduces the differences within each cluster. On the other hand, the Silhouette Method evaluates how well an item's cluster aligns with the clusters of other items, helping to determine the optimal number of clusters for the data.

Pharma\_WSS <- fviz\_nbclust(norm.PharmaceuticalsData1, kmeans, method = "wss") + labs(subtitle = "Elbow Method")  
Pharma\_Silho <- fviz\_nbclust(norm.PharmaceuticalsData1, kmeans, method = "silhouette")  
plot\_grid(Pharma\_WSS, Pharma\_Silho) + labs(subtitle = "Silhouette Method")



#The elbow method suggests that the optimal number of clusters, k, is 2, based on the point where the line starts to bend. However, the Silhouette Method indicates that k should be 5. Despite these differences, I have decided to use the k-means method with k set to 5 for my analysis.

#Using k-means method with k=5.  
set.seed(123)  
K\_Means.PharmaceuticalsData.optimal <- kmeans(norm.PharmaceuticalsData1, centers = 5, nstart = 50)  
K\_Means.PharmaceuticalsData.optimal$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

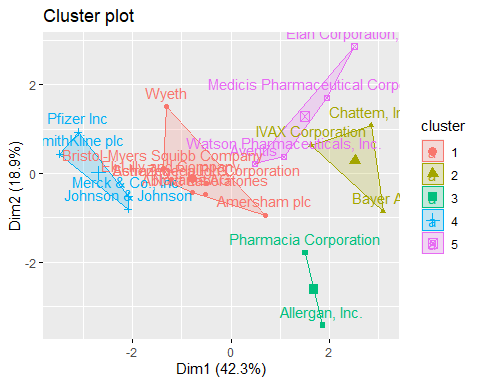
K\_Means.PharmaceuticalsData.optimal$size

## [1] 8 3 2 4 4

K\_Means.PharmaceuticalsData.optimal$withinss

## [1] 21.879320 15.595925 2.803505 9.284424 12.791257

fviz\_cluster(K\_Means.PharmaceuticalsData.optimal, data = norm.PharmaceuticalsData1)



#Using the data, we can categorize the firms into five clusters based on their distance from the central tendencies. Cluster 4 stands out for its high Market Capital, while Cluster 2 is notable for a high Beta, and Cluster 5 exhibits a low Asset Turnover. Additionally, it's worth noting that Cluster 1 encompasses the most enterprises, while Cluster 3 consists of only two. Examining the within-cluster sum of squared distances provides insights into the dispersion of data; for instance, Cluster 1 (21.9) is less internally consistent than Cluster 3 (2.8). Visualizing the algorithm's output allows us to easily observe the distinct groups formed by the data.

#b).Interpreting the clusters with respect to the numerical variables used in forming the clusters.

#I decided to rerun the model with only three clusters to gain a more comprehensive understanding of the cluster analysis. Having just two clusters raised concerns about potentially overlooking important features in the data.

#Using k-means with k=3.  
set.seed(123)  
K\_Means.PharmaceuticalsData <- kmeans(norm.PharmaceuticalsData1, centers = 3, nstart = 50)  
K\_Means.PharmaceuticalsData$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

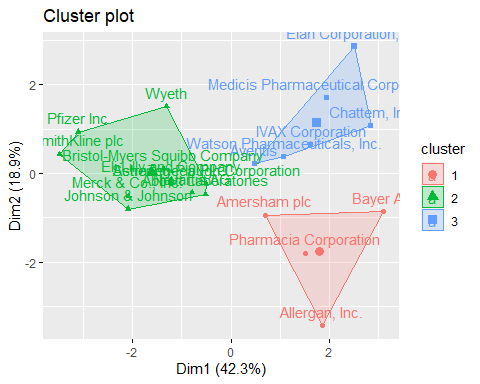
K\_Means.PharmaceuticalsData$size

## [1] 4 11 6

K\_Means.PharmaceuticalsData$withinss

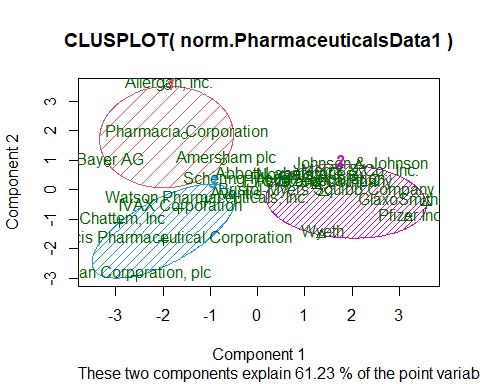
## [1] 20.54199 43.30886 32.14336

fviz\_cluster(K\_Means.PharmaceuticalsData, data = norm.PharmaceuticalsData1)



#The analysis has led to the identification and categorization of clusters. Specifically, there are four data points in cluster 1, eleven data points in cluster 2, and six data points in cluster 3.

#To view the cluster plot-  
clusplot(norm.PharmaceuticalsData1,K\_Means.PharmaceuticalsData$cluster,color = TRUE,shade =TRUE, labels = 2,lines = 0)

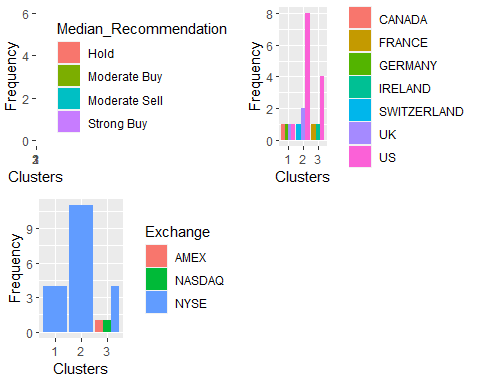


#c). Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

#To examine trends in the data, I opt to use bar charts to visually represent how firms are distributed across clusters based on the last three categorical variables: Median Recommendation, Location, and Stock Exchange

PharmaceuticalsData2 <- PharmaceuticalsData %>% select(c(11,12,13)) %>%   
 mutate(Cluster = K\_Means.PharmaceuticalsData$cluster)

Median\_Rec <- ggplot(PharmaceuticalsData2, mapping = aes(factor(Cluster), fill=Median\_Recommendation)) +  
 geom\_bar(position = 'dodge') +  
 labs(x='Clusters', y='Frequency')  
Location <- ggplot(PharmaceuticalsData2, mapping = aes(factor(Cluster), fill=Location)) +  
 geom\_bar(position = 'dodge') +   
 labs(x='Clusters', y='Frequency')  
Exchange <- ggplot(PharmaceuticalsData2, mapping = aes(factor(Cluster), fill=Exchange)) +  
 geom\_bar(position = 'dodge') +   
 labs(x='Clusters', y='Frequency')  
plot\_grid(Median\_Rec, Location, Exchange)



#The chart clearly shows that most companies in cluster 3 are from the United States and all of them have a "Hold" recommendation for their shares. Additionally, these companies are exclusively traded on the New York Stock Exchange. In cluster 2, we've selected stocks with a "Moderate Buy" recommendation, and there are only two companies listed on exchanges other than the NYSE, such as AMEX and NASDAQ. Cluster 1 consists of four firms located in four different countries, and interestingly, their stocks are all traded on the NYSE.

#d). Assigning meaningful names to each cluster based on the characteristics of the firms can be accomplished by considering the distinctive features captured by the numerical variables. The labels should reflect the common traits shared by the firms within each cluster, making it easier to interpret and communicate the essence of each group.

#Ans).Cluster 1: These companies are termed as “overvalued international firms” because they operate globally, are listed on the NYSE, have low Net Profit Margins, and high Price/Earnings ratios. Despite their high market valuations, their current earnings may not justify such high stock prices. To sustain their stock value, they need to invest and increase earnings to meet investor expectations.

Cluster 2: This group is identified as a “growing and leveraged firm.” They have “Moderate buy” evaluations, low asset turnover and Return on Assets (ROA), high leverage, and are expected to experience revenue growth. Although currently not very profitable and carrying significant debt, investors see potential in them and are willing to wait for future growth.

Cluster 3: These companies are characterized as “mature US firms” because they are based in the United States, listed on the NYSE, and have received “Hold” ratings. Their status suggests a stable and mature phase of development in the business.