Assignment-4

Harika

2023-11-19

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

#install.packages("factoextra")  
library(factoextra)

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

## 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)

## ── 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()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

#install.packages("cowplot")  
library(cowplot)

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

##   
## Attaching package: 'cowplot'  
##   
## The following object is masked from 'package:lubridate':  
##   
## stamp

library(readr)  
#install.packages("flexclust")  
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

#install.packages("cluster")  
library(cluster)

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

#install.packages("NbClust")  
library(NbClust)

Pharmaceuticals <- read.csv("C:/Users/Harika/Downloads/Pharmaceuticals.csv")  
###to read the given dataset  
view(Pharmaceuticals)  
###to view the given dataset.  
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

###to call first few observations from the given dataset.  
str(Pharmaceuticals)

## 'data.frame': 21 obs. of 14 variables:  
## $ Symbol : chr "ABT" "AGN" "AHM" "AZN" ...  
## $ Name : chr "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PLC" ...  
## $ 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" ...

###to see the structure of the given dataset.  
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   
##   
##   
##

###to see the summary for the given dataset.  
dim(Pharmaceuticals)

## [1] 21 14

###to see how many rows and columns are there in the given dataset.  
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]

FIRST QUESTION:

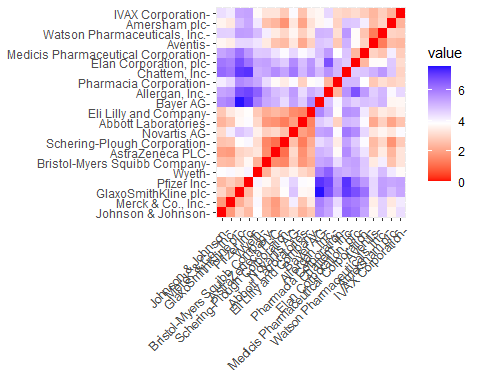
a.Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on

Pharmaceuticals1 <- Pharmaceuticals[,-c(1,11:13)]  
###with exception of "Symbol" and the last three non-numerical variables.

NORMALIZING AND CLUSTERING THE DATA

Here, I have calculated the separation between each observation and the data must be altered first because the Euclidean distance measure, which is scale sensitive used by default.

norm.Pharmaceuticals1 <- scale(Pharmaceuticals1)  
###the data is normalized.  
distance <- get\_dist(norm.Pharmaceuticals1)  
fviz\_dist(distance)



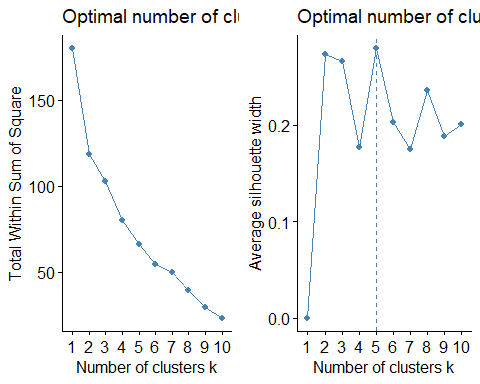
###to measure and plot distance for the given dataset.

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

#To find the Optimal K value

The Elbow chart and the Silhouette Method are two of the best methods for determining the number of clusters for the k-means model when there are no outside influences. The Elbow chart illustrates how adding more clusters causes a decrease in cluster heterogeneity, while the Silhouette Method assesses how closely related an object’s cluster is to those of other clusters.

WSS <- fviz\_nbclust(norm.Pharmaceuticals1, kmeans, method = "wss")  
Silhouette <- fviz\_nbclust(norm.Pharmaceuticals1, kmeans, method = "silhouette")  
plot\_grid(WSS, Silhouette)



###we used elbow chart and silhouette methods.

The charts above indicate that, according to the elbow method, the bend occurs when k=2, while the Silhouette method suggests k=5. I have chosen to use the k-means method with k=5.

###using k-means k=5 for making clusters  
set.seed(123)  
Kmeans.Pharamaceuticals.Optimalno <- kmeans(norm.Pharmaceuticals1, centers = 5, nstart = 50)  
Kmeans.Pharamaceuticals.Optimalno$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

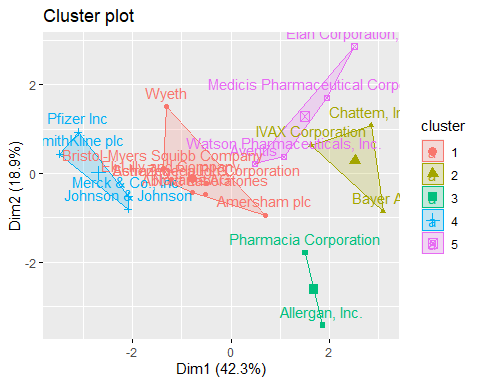
Kmeans.Pharamaceuticals.Optimalno$size

## [1] 8 3 2 4 4

Kmeans.Pharamaceuticals.Optimalno$withinss

## [1] 21.879320 15.595925 2.803505 9.284424 12.791257

fviz\_cluster(Kmeans.Pharamaceuticals.Optimalno, data = norm.Pharmaceuticals1)

 > Using the data, we can define five clusters based on their distance from the cores. Cluster 4 has a high Market Capital, whereas Cluster 2 has a high Beta, and Cluster 5 has a low Asset Turnover. We can also determine the size of each cluster. Cluster 1 has the most enterprises, whereas Cluster 3 has only two. The within-cluster sum of squared distances reveals information about data dispersion: Cluster 1 (21.9) is less homogeneous than Cluster 3 (2.8). By visualizing the algorithm’s output, we can observe the five groups into which the data has been grouped.

SECOND QUESTION:

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

###using k-means k=3 for making clusters  
set.seed(123)  
Kmeans.Pharmaceuticals <- kmeans(norm.Pharmaceuticals1, centers = 3, nstart = 50)  
Kmeans.Pharmaceuticals$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

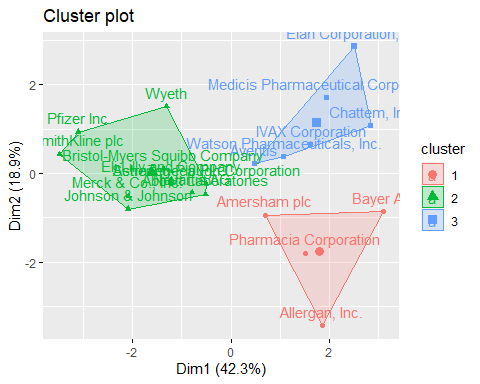
Kmeans.Pharmaceuticals$size

## [1] 4 11 6

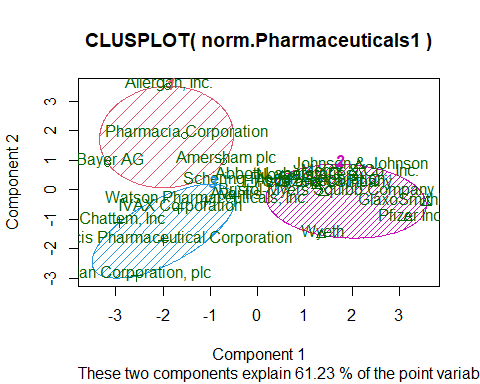
Kmeans.Pharmaceuticals$withinss

## [1] 20.54199 43.30886 32.14336

fviz\_cluster(Kmeans.Pharmaceuticals, data = norm.Pharmaceuticals1)

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

clusplot(norm.Pharmaceuticals1,Kmeans.Pharmaceuticals$cluster,color = TRUE,shade =TRUE, labels=2,lines=0)



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

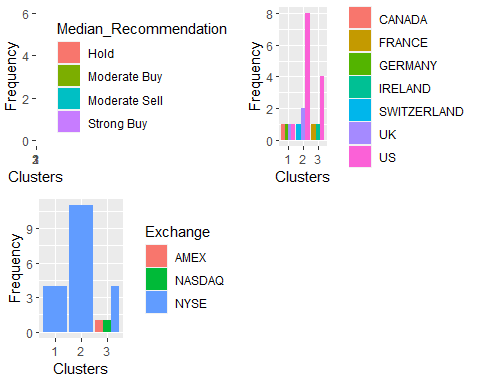
THIRD QUESTION:

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

By, considering the three last categorical variables ie., Median\_Recommendation, Location and Stock Exchange. To check for any trends in the data, I like to use bar charts for graphical representation of the distribution of firms which are grouped by clusters.

###dataset is partitioned for last 3 variables.  
Pharmaceuticals3 <- Pharmaceuticals %>% select(c(11,12,13)) %>%  
 mutate(Cluster = Kmeans.Pharmaceuticals$cluster)

Median\_Rec <- ggplot(Pharmaceuticals3, mapping = aes(factor(Cluster), fill=Median\_Recommendation)) +  
 geom\_bar(position = 'dodge') +  
 labs(x='Clusters', y='Frequency')  
Location <- ggplot(Pharmaceuticals3, mapping = aes(factor(Cluster), fill=Location)) +  
 geom\_bar(position = 'dodge') +   
 labs(x='Clusters', y='Frequency')  
Exchange <- ggplot(Pharmaceuticals3, mapping = aes(factor(Cluster), fill=Exchange)) +  
 geom\_bar(position = 'dodge') +   
 labs(x='Clusters', y='Frequency')  
plot\_grid(Median\_Rec,Location,Exchange)

 > The graph plainly illustrates that the majority of the companies in cluster 3 are based in the United States, and all of them have a ‘hold’ recommendation for their shares. They are all traded on the New York Stock Exchange. In cluster 2, we choose ‘Moderate Buy’ shares, including 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.

FOURTH QUESTION:

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

Here, we can compile all the given data from the dataset and identify the three distinct groups among the list of 21 pharmaceutical companies.

Cluster 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 conduct business on multiple continents while raising capital on the world’s largest stock exchange (NYSE). They have high financial market valuations that are not supported by their present earnings levels. To prevent their stock prices from collapsing, they must invest and increase earnings to meet investors’ expectations.

Cluster 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 substantial debt, they appear to be highly valued by investors willing to wait for future growth.

Cluster 3 qualifies as a ‘mature US firm’ since it is US-based, listed on the NYSE, and has ‘Hold’ ratings.