Assignment-4

Archana Gadicharla

2024-03-17

#Loading the required packages  
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.3

## 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 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.1  
## ✔ readr 2.1.5

## ── 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

library(cowplot)

## Warning: package 'cowplot' was built under R version 4.3.3

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

library(readr)  
library(flexclust)

## Warning: package 'flexclust' was built under R version 4.3.3

## Loading required package: grid

## Loading required package: modeltools

## Loading required package: stats4

library(cluster)  
library(NbClust)

#Data importing   
pharmacy <- read.csv("C:\\Users\\archa\\Downloads\\Pharmaceuticals.csv")  
###to read the given dataset  
#View(pharmacy)  
###to view the given dataset.  
head(pharmacy)

## 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(pharmacy)

## '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" ...

#Checking the structure of the given dataset

summary(pharmacy)

## 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.

colMeans(is.na(pharmacy))

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

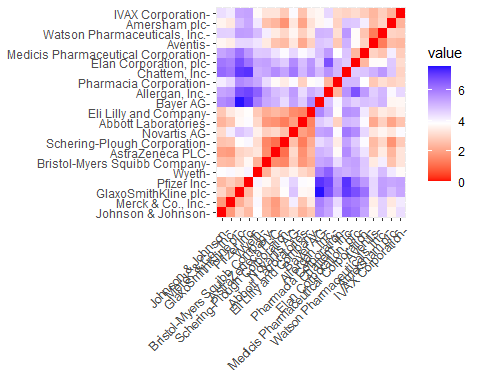
summary(pharmacy)

## 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

QUESTION 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

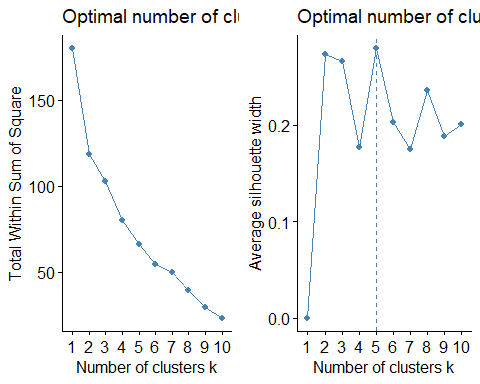
pharmacy1 <- pharmacy[,-c(1,11:13)]  
###with exception of "Symbol" and the last three non-numerical variables

#Normalizing the data  
norm.pharmacy1 <- scale(pharmacy1)  
###the data is normalized.  
distance <- get\_dist(norm.pharmacy1)  
fviz\_dist(distance)

 The graph visually represents color intensity variation corresponding to distance. As expected, the diagonal shows a value of zero, indicating the distance between two observations.

In terms of determining the optimal K value, both the Elbow chart and the Silhouette Method are effective techniques for discerning the number of clusters in a k-means model, particularly when external factors are absent. The Elbow chart illustrates how increasing the number of clusters leads to a reduction in cluster heterogeneity, while the Silhouette Method evaluates the proximity of an object’s cluster to others.

#Using silhouette method to find optimal k  
WSS <- fviz\_nbclust(norm.pharmacy1, kmeans, method = "wss")  
Silhouette <- fviz\_nbclust(norm.pharmacy1, kmeans, method = "silhouette")  
plot\_grid(WSS, Silhouette)

 Based on the elbow chart, the inflection point occurs at k=2, indicating a potential number of clusters. However, the Silhouette method suggests k=5 as a suitable choice. For this analysis, I have chosen to utilize the k-means method with k=5, as it aligns with the Silhouette method and provides a more nuanced understanding of the data’s clustering structure.

###using k-means k=5 for making clusters  
set.seed(123)  
Kmeans.pharmacy <- kmeans(norm.pharmacy1, centers = 5, nstart = 50)  
Kmeans.pharmacy$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

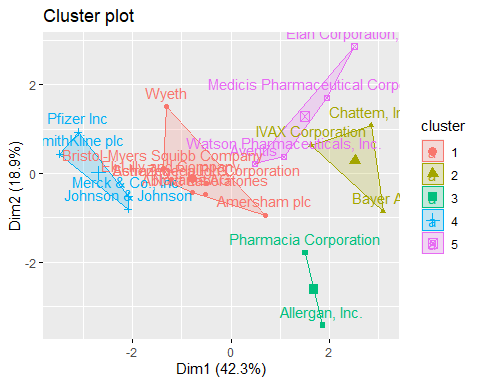
Kmeans.pharmacy$size

## [1] 8 3 2 4 4

Kmeans.pharmacy$withinss

## [1] 21.879320 15.595925 2.803505 9.284424 12.791257

fviz\_cluster(Kmeans.pharmacy, data = norm.pharmacy1)

 Based on the dataset, we can classify five clusters based on their proximity to the centroids. Cluster 4 is characterized by a high Market Capitalization, while Cluster 2 is notable for its high Beta, and Cluster 5 exhibits a low Asset Turnover. Furthermore, we can evaluate the size of each cluster, with Cluster 1 containing the highest number of enterprises, whereas Cluster 3 comprises only two. The within-cluster sum of squared distances provides insights into the dispersion of data: Cluster 1 (21.9) shows less homogeneity compared to Cluster 3 (2.8). By visualizing the output of the algorithm, we can observe the division of data into the five distinct groups.

Question 2: 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.pharmacy1, 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

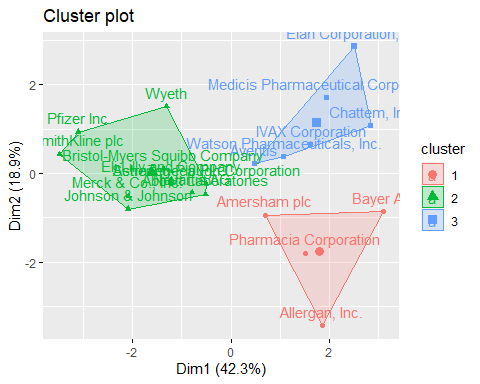
Kmeans.Pharmaceuticals$size

## [1] 4 11 6

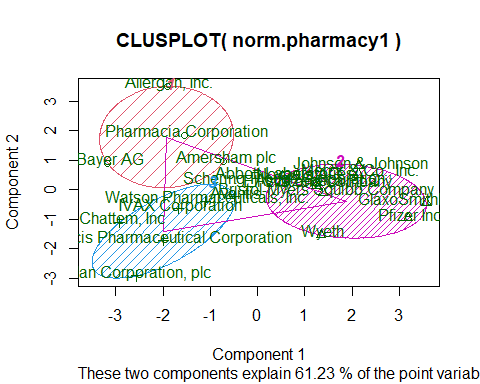
Kmeans.Pharmaceuticals$withinss

## [1] 20.54199 43.30886 32.14336

fviz\_cluster(Kmeans.Pharmaceuticals, data = norm.pharmacy1)

 This simplifies the process of identifying and managing the clusters within the analysis. Currently, there are 4 observations in cluster 1, 11 observations in cluster 2, and 6 observations in cluster 3.

library(cluster)  
clusplot(norm.pharmacy1,Kmeans.Pharmaceuticals$cluster,color = TRUE,shade =TRUE, labels=2,lines= TRUE)

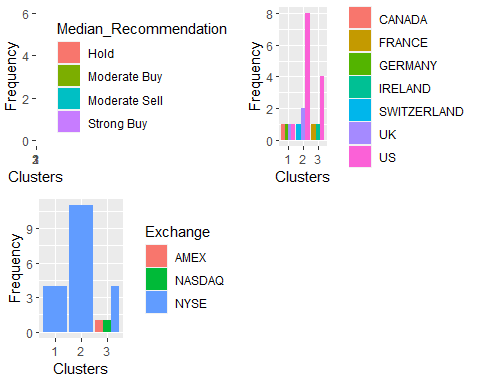
 From the second visualization, it’s clear that companies in cluster 1 display a juxtaposition of low Net Profit Margin and high Price/Earnings ratio. On the other hand, companies in cluster 2 exhibit low Asset Turnover and Return on Asset (ROA), coupled with high Leverage and Estimated Revenue Growth. However, cluster 3 does not reveal any discernible distinguishing features across the parameters analyzed.

Question 3: Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

Utilizing bar charts to explore trends based on the three categorical variables - Median Recommendation, Location, and Stock Exchange - can offer valuable insights into the distribution of firms across clusters, potentially revealing patterns within the dataset.

#The pharmacy data is partitioned for the last 3 variables  
pharmacy3 <- pharmacy %>% select(c(11,12,13)) %>%  
 mutate(Cluster = Kmeans.Pharmaceuticals$cluster)

Median\_Rec <- ggplot(pharmacy3, mapping = aes(factor(Cluster), fill=Median\_Recommendation)) +  
geom\_bar(position = 'dodge') +  
labs(x='Clusters', y='Frequency')  
Location <- ggplot(pharmacy3, mapping = aes(factor(Cluster), fill=Location)) +  
geom\_bar(position = 'dodge') +  
labs(x='Clusters', y='Frequency')  
Exchange <- ggplot(pharmacy3, mapping = aes(factor(Cluster), fill=Exchange)) +  
geom\_bar(position = 'dodge') +  
labs(x='Clusters', y='Frequency')  
plot\_grid(Median\_Rec,Location,Exchange)

 Inference: The provided graph illustrates a distinct pattern among the clusters. Cluster 3 primarily consists of companies headquartered in the United States, all of which have received a ‘hold’ recommendation for their shares. Moreover, these companies are exclusively listed on the New York Stock Exchange. In cluster 2, there is a tendency towards ‘Moderate Buy’ shares, with only two companies listed on alternative exchanges or indexes such as AMEX and NASDAQ. Cluster 1 presents a diverse composition, featuring four firms from four different countries, all of which have their stocks traded on the NYSE.

Question 4: Provide an appropriate name for each cluster using any or all of the variables in the dataset Inference: In this analysis, we synthesize all the data from the dataset to delineate three discernible groups within the cohort of 21 pharmaceutical companies.

Cluster 1 is identified as ‘overvalued international firms’ based on various factors: their global footprint, listing on the NYSE, low Net Profit Margin, and elevated Price/Earnings ratio. These companies have a presence across multiple continents and raise capital on the prestigious NYSE. However, despite their lofty market valuations, their current earnings fail to justify such high levels. To avoid a potential decline in their stock prices, these companies must focus on investment and improving their earnings to align with investors’ expectations.

Cluster 2 is categorized as a ‘growing and leveraged firm’ for specific characteristics: ‘Moderate buy’ ratings, low asset turnover and Return on Assets (ROA), high leverage, and expected revenue growth. Despite their current modest profitability and substantial debt load, these companies enjoy favorable investor sentiment, likely due to the anticipation of future growth prospects.

Cluster 3 is identified as an ‘Established US firm’ because of its United States-based operations, listing on the NYSE, and ‘Hold’ ratings received from analysts.