FML\_ASSIGNMENT\_4

Eswar dumpa

2024-03-16

# The necessary packages are loaded  
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

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

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

## 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

#install.packages("cowplot")  
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

#install.packages("flexclust")  
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

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

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

# It imports the "Pharmaceuticals" dataset from the specified file path  
Pharmacy <- read.csv("C:/Users/eshwa/Documents/Fundamentals of Machine Learning/ASSN 4/Pharmaceuticals.csv")

# The "Pharmacy" data set will be viewed  
view(Pharmacy)

# It displays first few rows of the "Pharmacy" 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

# It displays summary statistics for the "Pharmacy" 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   
##   
##   
##

#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.  
  
# Calculates the column wise mean of missing values in "Pharmacy" dataset  
colMeans(is.na(Pharmacy))

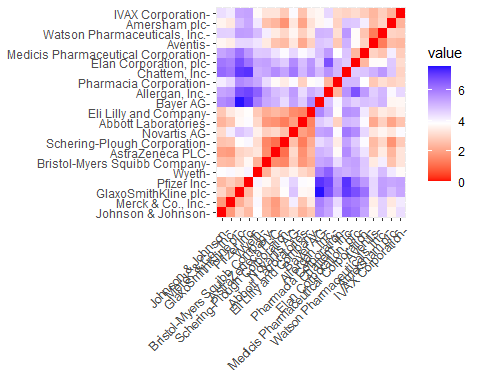
## Symbol Name Market\_Cap   
## 0 0 0   
## Beta PE\_Ratio ROE   
## 0 0 0   
## ROA Asset\_Turnover Leverage   
## 0 0 0   
## Rev\_Growth Net\_Profit\_Margin Median\_Recommendation   
## 0 0 0   
## Location Exchange   
## 0 0

# Sets row names of "Pharmacy" to the values in second column.  
row.names(Pharmacy) <- Pharmacy[,2]  
# Removes the second column from "Pharmacy" dataset  
Pharmacy <- Pharmacy[,-2]  
# Removes the first column and columns 11 to 13 from the updated "Pharmacy" dataset  
Pharmacy.1 <- Pharmacy[,-c(1,11:13)]

# Checks the dimensions of "Pharmacy" dataset  
dim(Pharmacy)

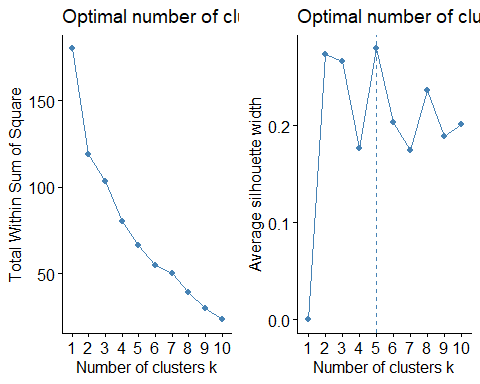
## [1] 21 13

# Standardizes columns of "Pharmacy.1" using the scale function  
norm.Pharmacy.1 <- scale(Pharmacy.1)  
# Calculates distance matrix based on the standardized data  
dist <- get\_dist(norm.Pharmacy.1)  
# Visualizes distance matrix using function  
fviz\_dist(dist)



# The chart illustrates how color intensity varies with distance traveled. The diagonal line that shows the separation between two observations has a value of zero, as would be expected.  
  
# For finding the best K Value: For a k-means model, the Elbow chart and the Silhouette Method are useful tools for determining the number of clusters, particularly in situations when outside influences are not significant. The Elbow graphic illustrates how overall cluster diversity declines as the number of clusters increases. In contrast, the Silhouette Method assesses an object's cluster alignment with other clusters in order to provide insight on the cohesiveness of the clusters.

# Calculates Within Cluster Sum of Squares (WSS) for different numbers of clusters using k-means algorithm  
WSS <- fviz\_nbclust(norm.Pharmacy.1, kmeans, method = "wss")  
# Calculates Silhouette scores for different numbers of clusters using k-means algorithm  
Sil <- fviz\_nbclust(norm.Pharmacy.1, kmeans, method = "silhouette")  
# Displays plots of WSS and Silhouette scores  
plot\_grid(WSS, Sil)



# The charts indicate different optimal values for k, the Elbow Method suggests k=2, while the Silhouette Method produces k=5. Despite this, I have decided to use k=5 for k-means method in my analysis.

# Set the seed for reproducibility  
# Performs k-means clustering on normalized "Pharmacy.1" data with 5 centers   
# Displays the cluster centers obtained from k-means clustering  
set.seed(123)  
KMeans.Pharmacy.Opt <- kmeans(norm.Pharmacy.1, centers = 5, nstart = 50)  
KMeans.Pharmacy.Opt$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

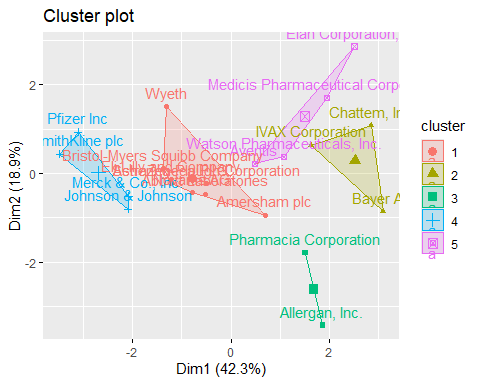
# Display size of each cluster  
KMeans.Pharmacy.Opt$size

## [1] 8 3 2 4 4

# Display within-cluster sum of squares  
KMeans.Pharmacy.Opt$withinss

## [1] 21.879320 15.595925 2.803505 9.284424 12.791257

# Visualize k-means clusters using a scatter plot  
fviz\_cluster(KMeans.Pharmacy.Opt, data = norm.Pharmacy.1)



# Based on the dataset's closeness to core points, we were able to identify five clusters. Cluster 2 is noteworthy for its high Beta, whilst Cluster 4 is notable for its high Market Capital.   
# Conversely, Cluster 5 exhibits a low asset turnover rate.When comparing the number of firms inside each cluster, Cluster 1 has the most, whilst Cluster 3 only has two.   
# The information about data dispersion can be obtained from the within-cluster sum of squared distances: Compared to Cluster 3 (2.8), Cluster 1 (21.9) is less homogeneous.The findings of the algorithm are visualized, enabling us to observe the many groups into which the data has been split.

#b. Interpret the clusters with respect to the numerical variables used in forming the clusters.  
  
# Set seed for reproducibility  
# Performs k-means clustering on the normalized "Pharmacy.1" data with 3 clusters  
# Displays cluster centers  
  
set.seed(123)  
KMeans.Pharmacy <- kmeans(norm.Pharmacy.1, centers = 3, nstart = 50)  
KMeans.Pharmacy$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

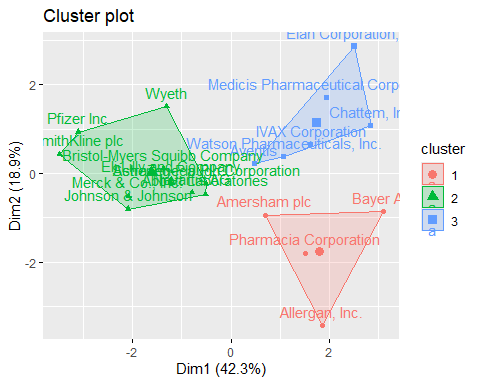
# Displays sizes of each cluster obtained from the k-means clustering.  
KMeans.Pharmacy$size

## [1] 4 11 6

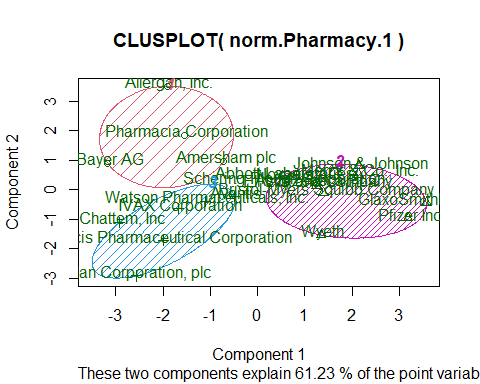
# Displays within-cluster sum of squares for each cluster  
KMeans.Pharmacy$withinss

## [1] 20.54199 43.30886 32.14336

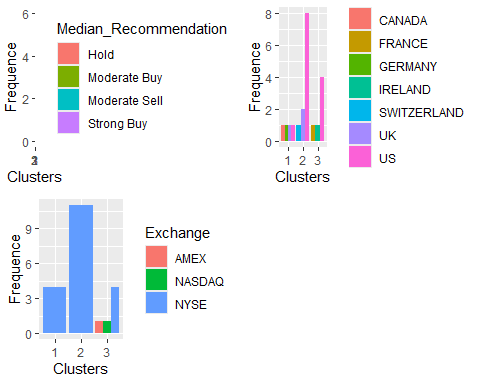
# Visualize k-means clusters using a scatter plot  
fviz\_cluster(KMeans.Pharmacy, data = norm.Pharmacy.1)



clusplot(norm.Pharmacy.1,KMeans.Pharmacy$cluster,color = TRUE,shade =TRUE, labels=2,lines=0)



#c. Is there a pattern in clusters with respect to numerical variables (10 to 12)?  
  
# Bar charts were my choice for examining trends in the data for the final three categorical variables: stock exchange, location, and median recommendation. These graphs give a clearer picture of the distribution of enterprises among various clusters, facilitating a better comprehension of data trends.  
  
Pharmacy.2 <- Pharmacy%>% select(c(11,12,13)) %>%   
 mutate(Cluster = KMeans.Pharmacy$cluster)  
Med\_Recom <- ggplot(Pharmacy.2, mapping = aes(factor(Cluster), fill=Median\_Recommendation)) +  
 geom\_bar(position = 'dodge') +  
 labs(x='Clusters', y='Frequence')  
Loc <- ggplot(Pharmacy.2, mapping = aes(factor(Cluster), fill=Location)) +  
 geom\_bar(position = 'dodge') +   
 labs(x='Clusters', y='Frequence')  
Ex <- ggplot(Pharmacy.2, mapping = aes(factor(Cluster), fill=Exchange)) +  
 geom\_bar(position = 'dodge') +   
 labs(x='Clusters', y='Frequence')  
plot\_grid(Med\_Recom, Loc, Ex)



# The majority of the companies in cluster 3 are clearly American, and all of them advise holding their shares, according to the chart.Only on the New York Stock Exchange are they traded. We have chosen stocks for cluster 2 with a "Moderate Buy" recommendation; just two businesses (AMEX and NASDAQ) are from separate exchanges. Cluster 1 shows that even though the four companies' stocks are all traded on the NYSE, they are all from different nations.

#d. Provide an appropriate name for each cluster using any or all of the variables in the dataset.  
  
#1) Cluster 1): Global Giants: These businesses are regarded as "overvalued international firms" due to their extensive global reach, NYSE listing, low net profit margins, and high price/earnings ratios.Their existing earnings do not adequately justify their high market value. They must make investments and boost profitability to satisfy investor expectations if they want to keep their stock prices high.  
#2) Cluster 2: Opportunities for Growth: Because of their "Moderate buy" evaluations, high leverage, poor ROA, low asset turnover, and projected revenue growth, this group is referred to as "growing and leveraged firms".They are highly valued by investors despite their current lack of profitability and large debt load because they perceive promise for future growth.  
#3) Cluster 3 - Stable US Companies: Because they are US-based, NYSE-listed, and have a "Hold" rating, the companies in this cluster are classified as "mature US firms".Compared to the other clusters, they are regarded as stable and mature, suggesting a more cautious attitude to investing.