FML\_Asiignment\_4\_Clustering

2023-11-11

library(ggplot2)  
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(flexclust)

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

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(cluster)  
library(tidyverse)

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

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

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

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.0

## ── 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

library(gridExtra)

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

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

#Importing the Dataset  
  
pharma<-read.csv("Pharmaceuticals.csv")  
summary(pharma)

## 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   
##   
##   
##

#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.

#Removing all the null values from the datasets and selecting the monetary variables  
colSums(is.na(pharma))

## 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 <- pharma[,2]  
Pharma\_data<- pharma[, 3:11]  
head(Pharma\_data)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## 1 68.44 0.32 24.7 26.4 11.8 0.7 0.42 7.54  
## 2 7.58 0.41 82.5 12.9 5.5 0.9 0.60 9.16  
## 3 6.30 0.46 20.7 14.9 7.8 0.9 0.27 7.05  
## 4 67.63 0.52 21.5 27.4 15.4 0.9 0.00 15.00  
## 5 47.16 0.32 20.1 21.8 7.5 0.6 0.34 26.81  
## 6 16.90 1.11 27.9 3.9 1.4 0.6 0.00 -3.17  
## Net\_Profit\_Margin  
## 1 16.1  
## 2 5.5  
## 3 11.2  
## 4 18.0  
## 5 12.9  
## 6 2.6

# Scaling and Normalisation of dataset.  
class(Pharma\_data)

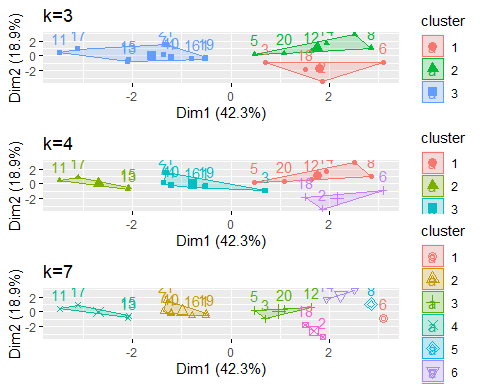
## [1] "data.frame"

Pharma\_scale<- scale(Pharma\_data)  
# Calculate means of the scaled columns  
head(Pharma\_scale)

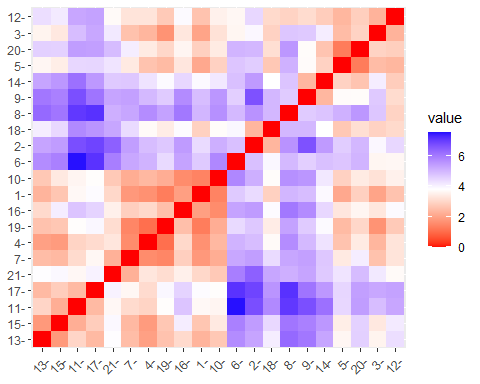
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## [1,] 0.1840960 -0.80125356 -0.04671323 0.04009035 0.2416121 0.0000000  
## [2,] -0.8544181 -0.45070513 3.49706911 -0.85483986 -0.9422871 0.9225312  
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700 0.9225312  
## [4,] 0.1702742 -0.02225704 -0.24290879 0.10638147 0.9181259 0.9225312  
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461 -0.4612656  
## [6,] -0.6953818 2.27578267 0.14948233 -1.45146000 -1.7127612 -0.4612656  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## [1,] -0.2120979 -0.5277675 0.06168225  
## [2,] 0.0182843 -0.3811391 -1.55366706  
## [3,] -0.4040831 -0.5721181 -0.68503583  
## [4,] -0.7496565 0.1474473 0.35122600  
## [5,] -0.3144900 1.2163867 -0.42597037  
## [6,] -0.7496565 -1.4971443 -1.99560225

Pharma\_scale <- as.data.frame(scale(Pharma\_data))

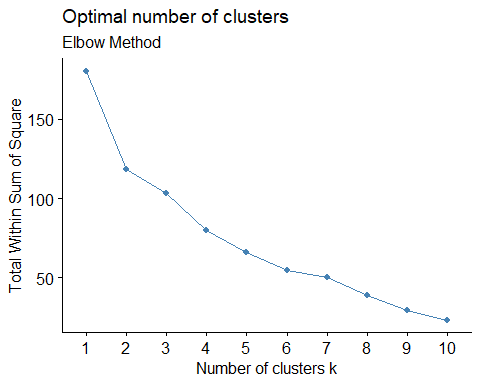
# Calculate K-means clustering for various centers, use a variety of K values, and compare the results.  
kmeans\_1normalization <- kmeans(Pharma\_scale, centers = 3, nstart = 30)  
kmeans\_2normalization<- kmeans(Pharma\_scale, centers = 4, nstart = 30)  
kmeans\_3normalization<- kmeans(Pharma\_scale, centers = 7, nstart = 30)  
Plot\_1results<-fviz\_cluster(kmeans\_1normalization, data = Pharma\_scale)+ggtitle("k=3")  
Plot\_2results<-fviz\_cluster(kmeans\_2normalization, data = Pharma\_scale)+ggtitle("k=4")  
Plot\_3results<-fviz\_cluster(kmeans\_3normalization, data = Pharma\_scale)+ggtitle("k=7")  
grid.arrange(Plot\_1results,Plot\_2results,Plot\_3results, nrow = 3)



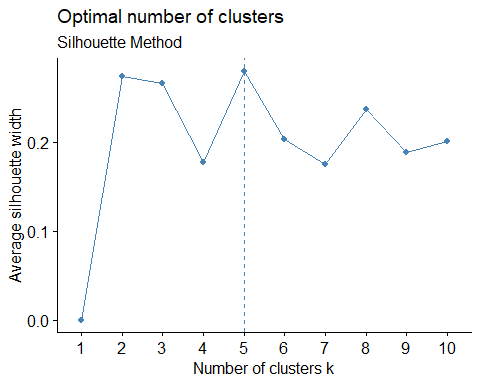
distance<- dist(Pharma\_scale, method = "euclidean")  
fviz\_dist(distance)



# Estimating the number of clusters  
# Elbow Method is used in scaling the data to determine the value of k  
fviz\_nbclust(Pharma\_scale, FUNcluster = kmeans, method = "wss") + labs(subtitle = "Elbow Method")



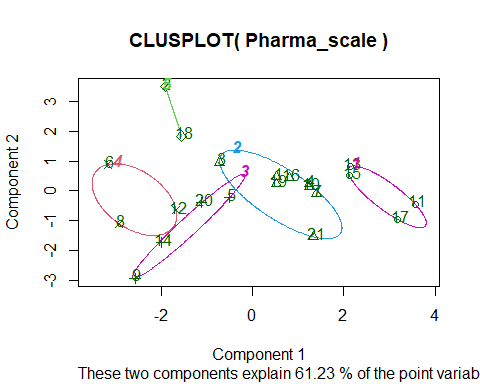
# Silhouette Method is used in scaling the data to determine the number of clusters  
fviz\_nbclust(Pharma\_scale,FUNcluster = kmeans,method = "silhouette")+labs(subtitle="Silhouette Method")



# Final analysis and Extracting results using 5 clusters and Visualize the results  
set.seed(200)  
final\_Clusters<- kmeans(Pharma\_scale, 5, nstart = 20)  
print(final\_Clusters)

## K-means clustering with 5 clusters of sizes 4, 8, 4, 3, 2  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 2 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 3 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 4 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 5 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.46807818 0.4671788 0.591242521  
## 2 -0.27449312 -0.7041516 0.556954446  
## 3 0.06308085 1.5180158 -0.006893899  
## 4 1.36644699 -0.6912914 -1.320000179  
## 5 -0.14170336 -0.1168459 -1.416514761  
##   
## Clustering vector:  
## [1] 2 5 2 2 3 4 2 4 3 2 1 4 1 3 1 2 1 5 2 3 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 9.284424 21.879320 12.791257 15.595925 2.803505  
## (between\_SS / total\_SS = 65.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

clusplot(Pharma\_scale,final\_Clusters$cluster, color = TRUE, labels = 2,lines = 0)



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

PC\_Cluster <- pharma[,c(12,13,14)]%>% mutate(clusters = final\_Clusters$cluster)%>% arrange(clusters, .desc = TRUE)  
PC\_Cluster

## Median\_Recommendation Location Exchange clusters  
## 1 Hold UK NYSE 1  
## 2 Moderate Buy US NYSE 1  
## 3 Hold US NYSE 1  
## 4 Moderate Buy US NYSE 1  
## 5 Moderate Buy US NYSE 2  
## 6 Strong Buy UK NYSE 2  
## 7 Moderate Sell UK NYSE 2  
## 8 Moderate Sell US NYSE 2  
## 9 Hold US NYSE 2  
## 10 Hold SWITZERLAND NYSE 2  
## 11 Hold US NYSE 2  
## 12 Hold US NYSE 2  
## 13 Moderate Buy FRANCE NYSE 3  
## 14 Moderate Sell IRELAND NYSE 3  
## 15 Moderate Buy US NYSE 3  
## 16 Moderate Sell US NYSE 3  
## 17 Hold GERMANY NYSE 4  
## 18 Moderate Buy US NASDAQ 4  
## 19 Hold US AMEX 4  
## 20 Moderate Buy CANADA NYSE 5  
## 21 Hold US NYSE 5

#Cluster 1 consists of 1,3,4,7,10,16,19, and 21 (highest leverage, highest rev growth, lowest market cap, lowest beta, and lowest PE ratio).

#Cluster 2: 6, 8, 12 (lowest Net Profit Margin, greatest Beta and largest Rev Growth)

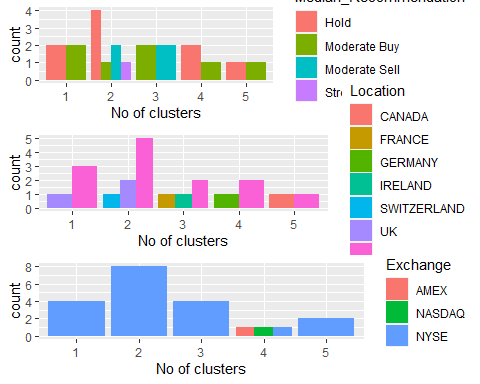
#Cluster 3: 5, 9, 14, 20, (highest ROE, lowest ROA, lowest Net Profit Margin, lowest PE Ratio, strongest Rev Growth)

#Cluster 4-2, 18 (highest PE ratio, lowest asset turnover, and lowest beta)

#Cluster 5: 11, 13, 15, 17, (lowest beta/PE ratio and highest market cap, ROE, ROA, and asset turnover ratio)

#Task3 #In terms of the numerical, are there any clusters that exhibit a pattern. (10 to 12) variables? (those n not utilized in the cluster formation).

plot1\_normalization<-ggplot(PC\_Cluster, mapping = aes(factor(clusters), fill=Median\_Recommendation))+geom\_bar(position = 'dodge')+labs(x ='No of clusters')  
plot2\_normalization<- ggplot(PC\_Cluster, mapping = aes(factor(clusters),fill = Location))+geom\_bar(position = 'dodge')+labs(x ='No of clusters')  
plot3\_normalization<- ggplot(PC\_Cluster, mapping = aes(factor(clusters),fill = Exchange))+geom\_bar(position = 'dodge')+labs(x ='No of clusters')  
grid.arrange(plot1\_normalization, plot2\_normalization, plot3\_normalization)



# According to the graph

#Cluster 1: The Hold median is the highest in this cluster, which also includes unique Hold, Moderate Buy, Moderate Sell, and Strong Buy medians. They trade on the NYSE and are based in the US, the UK, and Switzerland.

#Cluster 2: The number of firms on AMEX, NASDAQ, and NYSE is equal, but there is a distinct count difference between the US and Germany, along with a clear Hold and Moderate Buy median.

#Cluster 3: traded on the New York Stock Exchange, features unique counts for France, Ireland, and the United States, and equally mild buy and sell order medians.

#Cluster 4 is dispersed over the US, UK, and listed in; it has the same hold and moderate purchase medians.

#Cluster 5: #exclusively listed on the NYSE, with median values of Hold and Moderate Buy, equally dispersed throughout the US and Canada.

#The clusters exhibit a certain pattern with regard to the media recommendation variable: #Cluster 1 and Cluster 2 have hold recommendations. # The buy recommendations for Clusters 3, 4, and 5 are moderate.

# (d) Give each cluster a suitable name using any or all of the dataset’s variables.

#Cluster 1 :- HIGH HOLD CLUSTER  
#Cluster 2 :- HOLD CLUSTER  
#Cluster 3 :- BUY-SELL CLUSTER  
#Cluster 4 :- HOLD-BUY CLUSTER  
#Cluster 5 :- HOLD-BUY CLUSTER