Assignment 4 - 64060

Marella (Martin) Murphy

2023-10-29

This is my submission for Assignment 4, clustering financial data for Pharmaceutical companies. The progression of my analysis starts with exploring and reading the data, loading the raw data and then creating a new data frame to work with. Once I have the data frame in place with labeled rows and columns, I then decided to use 2 clusters based on running two separate models for choosing the optimal number of clusters.

read.csv("Pharmaceuticals.csv") #testing that I have the correct directory set and RStudio can read the desired data set.

## Symbol Name Market\_Cap Beta PE\_Ratio ROE ROA  
## 1 ABT Abbott Laboratories 68.44 0.32 24.7 26.4 11.8  
## 2 AGN Allergan, Inc. 7.58 0.41 82.5 12.9 5.5  
## 3 AHM Amersham plc 6.30 0.46 20.7 14.9 7.8  
## 4 AZN AstraZeneca PLC 67.63 0.52 21.5 27.4 15.4  
## 5 AVE Aventis 47.16 0.32 20.1 21.8 7.5  
## 6 BAY Bayer AG 16.90 1.11 27.9 3.9 1.4  
## 7 BMY Bristol-Myers Squibb Company 51.33 0.50 13.9 34.8 15.1  
## 8 CHTT Chattem, Inc 0.41 0.85 26.0 24.1 4.3  
## 9 ELN Elan Corporation, plc 0.78 1.08 3.6 15.1 5.1  
## 10 LLY Eli Lilly and Company 73.84 0.18 27.9 31.0 13.5  
## 11 GSK GlaxoSmithKline plc 122.11 0.35 18.0 62.9 20.3  
## 12 IVX IVAX Corporation 2.60 0.65 19.9 21.4 6.8  
## 13 JNJ Johnson & Johnson 173.93 0.46 28.4 28.6 16.3  
## 14 MRX Medicis Pharmaceutical Corporation 1.20 0.75 28.6 11.2 5.4  
## 15 MRK Merck & Co., Inc. 132.56 0.46 18.9 40.6 15.0  
## 16 NVS Novartis AG 96.65 0.19 21.6 17.9 11.2  
## 17 PFE Pfizer Inc 199.47 0.65 23.6 45.6 19.2  
## 18 PHA Pharmacia Corporation 56.24 0.40 56.5 13.5 5.7  
## 19 SGP Schering-Plough Corporation 34.10 0.51 18.9 22.6 13.3  
## 20 WPI Watson Pharmaceuticals, Inc. 3.26 0.24 18.4 10.2 6.8  
## 21 WYE Wyeth 48.19 0.63 13.1 54.9 13.4  
## Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin Median\_Recommendation  
## 1 0.7 0.42 7.54 16.1 Moderate Buy  
## 2 0.9 0.60 9.16 5.5 Moderate Buy  
## 3 0.9 0.27 7.05 11.2 Strong Buy  
## 4 0.9 0.00 15.00 18.0 Moderate Sell  
## 5 0.6 0.34 26.81 12.9 Moderate Buy  
## 6 0.6 0.00 -3.17 2.6 Hold  
## 7 0.9 0.57 2.70 20.6 Moderate Sell  
## 8 0.6 3.51 6.38 7.5 Moderate Buy  
## 9 0.3 1.07 34.21 13.3 Moderate Sell  
## 10 0.6 0.53 6.21 23.4 Hold  
## 11 1.0 0.34 21.87 21.1 Hold  
## 12 0.6 1.45 13.99 11.0 Hold  
## 13 0.9 0.10 9.37 17.9 Moderate Buy  
## 14 0.3 0.93 30.37 21.3 Moderate Buy  
## 15 1.1 0.28 17.35 14.1 Hold  
## 16 0.5 0.06 -2.69 22.4 Hold  
## 17 0.8 0.16 25.54 25.2 Moderate Buy  
## 18 0.6 0.35 15.00 7.3 Hold  
## 19 0.8 0.00 8.56 17.6 Hold  
## 20 0.5 0.20 29.18 15.1 Moderate Sell  
## 21 0.6 1.12 0.36 25.5 Hold  
## Location Exchange  
## 1 US NYSE  
## 2 CANADA NYSE  
## 3 UK NYSE  
## 4 UK NYSE  
## 5 FRANCE NYSE  
## 6 GERMANY NYSE  
## 7 US NYSE  
## 8 US NASDAQ  
## 9 IRELAND NYSE  
## 10 US NYSE  
## 11 UK NYSE  
## 12 US AMEX  
## 13 US NYSE  
## 14 US NYSE  
## 15 US NYSE  
## 16 SWITZERLAND NYSE  
## 17 US NYSE  
## 18 US NYSE  
## 19 US NYSE  
## 20 US NYSE  
## 21 US NYSE

#In this code chunk I mirrored the code from the class Github example to have all the correct libraries. I then added two more libraries trying to get the Plot labels correct.  
  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.1

## Warning: package 'ggplot2' was built under R version 4.3.1

## Warning: package 'tibble' was built under R version 4.3.1

## Warning: package 'tidyr' was built under R version 4.3.1

## Warning: package 'readr' was built under R version 4.3.1

## Warning: package 'purrr' was built under R version 4.3.1

## Warning: package 'dplyr' was built under R version 4.3.1

## Warning: package 'stringr' was built under R version 4.3.1

## Warning: package 'forcats' was built under R version 4.3.1

## Warning: package 'lubridate' was built under R version 4.3.1

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── 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(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.1

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

library(ISLR)

## Warning: package 'ISLR' was built under R version 4.3.1

library(dplyr)  
library(ggplot2)  
set.seed(123)  
  
summary("Pharmaceuticals.csv")

## Length Class Mode   
## 1 character character

Pharma.df<- read.csv("Pharmaceuticals.csv", header = TRUE , )  
summary(Pharma.df)

## 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   
##   
##   
##

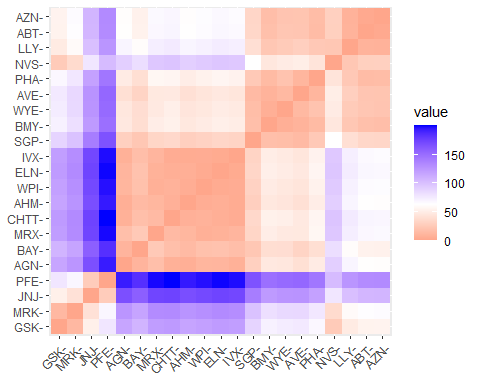
df <- Pharma.df[,c(3,11)]   
row.names(df) <- Pharma.df$Symbol   
  
#I like symbol better as row name, I tried name first and it was too lengthy and looked messy.  
  
#I am choosing to focus on three measurements for the data analysis, Market capitalization and Net Profit Margin.  
  
summary(df)

## Market\_Cap Net\_Profit\_Margin  
## Min. : 0.41 Min. : 2.6   
## 1st Qu.: 6.30 1st Qu.:11.2   
## Median : 48.19 Median :16.1   
## Mean : 57.65 Mean :15.7   
## 3rd Qu.: 73.84 3rd Qu.:21.1   
## Max. :199.47 Max. :25.5

df < - scale(df)

## Market\_Cap Net\_Profit\_Margin  
## ABT FALSE FALSE  
## AGN FALSE FALSE  
## AHM FALSE FALSE  
## AZN FALSE FALSE  
## AVE FALSE FALSE  
## BAY FALSE FALSE  
## BMY FALSE FALSE  
## CHTT TRUE FALSE  
## ELN TRUE FALSE  
## LLY FALSE FALSE  
## GSK FALSE FALSE  
## IVX FALSE FALSE  
## JNJ FALSE FALSE  
## MRX FALSE FALSE  
## MRK FALSE FALSE  
## NVS FALSE FALSE  
## PFE FALSE FALSE  
## PHA FALSE FALSE  
## SGP FALSE FALSE  
## WPI FALSE FALSE  
## WYE FALSE FALSE

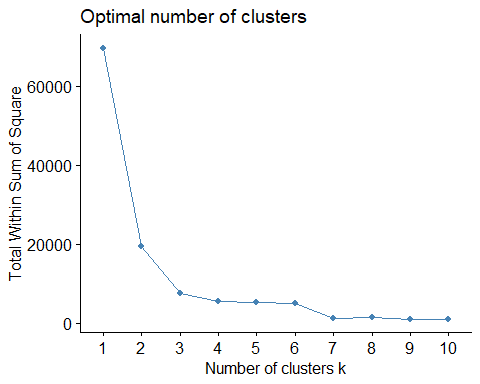
distance <- get\_dist(df)  
fviz\_dist(distance)



df < - scale(df)

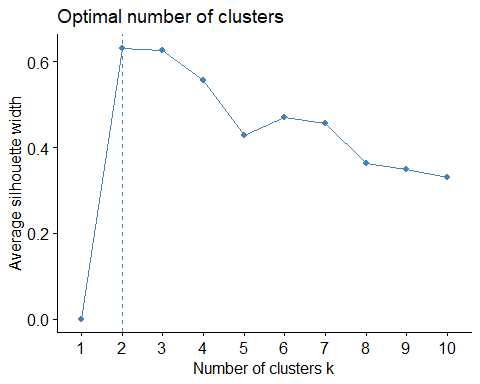
## Market\_Cap Net\_Profit\_Margin  
## ABT FALSE FALSE  
## AGN FALSE FALSE  
## AHM FALSE FALSE  
## AZN FALSE FALSE  
## AVE FALSE FALSE  
## BAY FALSE FALSE  
## BMY FALSE FALSE  
## CHTT TRUE FALSE  
## ELN TRUE FALSE  
## LLY FALSE FALSE  
## GSK FALSE FALSE  
## IVX FALSE FALSE  
## JNJ FALSE FALSE  
## MRX FALSE FALSE  
## MRK FALSE FALSE  
## NVS FALSE FALSE  
## PFE FALSE FALSE  
## PHA FALSE FALSE  
## SGP FALSE FALSE  
## WPI FALSE FALSE  
## WYE FALSE FALSE

fviz\_nbclust(df, kmeans, method = "wss")



#I see here that the optimal number of clusters to use is either two or three, I am not certain so I'm going to look at the silhouette chart next.

fviz\_nbclust(df, kmeans, method = "silhouette")



#This confirms that two clusters are optimal.

k2 <- kmeans(df, centers = 2, nstart = 25) # k = 2, number of restarts = 25  
  
k2$centers #The two groups are distinguished by size of Market Cap, cluster 1 has an average market cap of $144B, and cluster 2 has an average market cap of only $30B.

## Market\_Cap Net\_Profit\_Margin  
## 1 30.3725 14.30625  
## 2 144.9440 20.14000

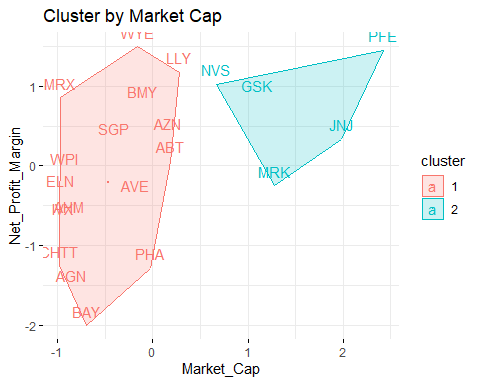
k2$size # number of companies in each cluster

## [1] 16 5

k2$cluster #list of which cluster each company was included in. I would like to learn to flip this table so it's all on one line!

## ABT AGN AHM AZN AVE BAY BMY CHTT ELN LLY GSK IVX JNJ MRX MRK NVS   
## 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 2   
## PFE PHA SGP WPI WYE   
## 2 1 1 1 1

# visualize the output. This visual makes it very clear that the 5 largest companies in terms of Market capitalization all have a net profit margin greater than $15B - I am assuming billions, I couldn't find $s presented in. This is where we begin to see a pattern with respect to variables 10-12. The top 5 market cap companies are all traded on the NYSE, and none are ranked sell.  
fviz\_cluster(k2, data = df, main = "Cluster by Market Cap",  
 geom = "text", ggtheme = theme\_minimal())



library(flexclust)

## Warning: package 'flexclust' was built under R version 4.3.1

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Warning: package 'modeltools' was built under R version 4.3.1

## Loading required package: stats4

set.seed(123)  
# Next I am going to try using the Manhattan distance  
k2 = kcca(df, k=2, kccaFamily("kmedians"))  
k2

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = df, k = 2, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2   
## 5 16

# Next I am going to use the predict function.  
clusters\_index <- predict(k2)  
dist(k2@centers)

## 1  
## 2 107.2821

image(k2)  
points(df, col=clusters\_index, pch=24, cex=0.6)

