Assignment 4

Rob DiVincenzo 10-29-2023 BA 64060

This is the submission for Assignment 4.

#Read data  
DF=read.csv("./Pharmaceuticals.csv") # Read the Pharmaceuticals csv file  
#Load libraries  
library(ISLR)  
library(tidyverse) # data manipulation

## ── 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 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ 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) # clustering algorithms & visualization

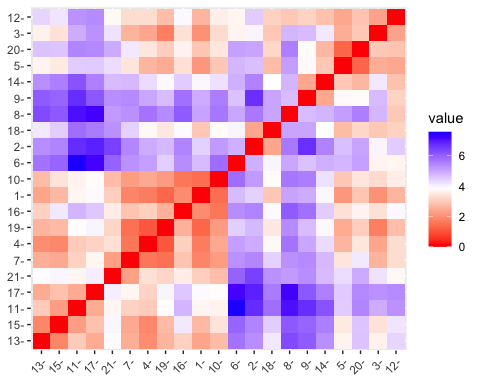
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(flexclust)

## Loading required package: grid  
## Loading required package: lattice  
## Loading required package: modeltools  
## Loading required package: stats4

set.seed(123)

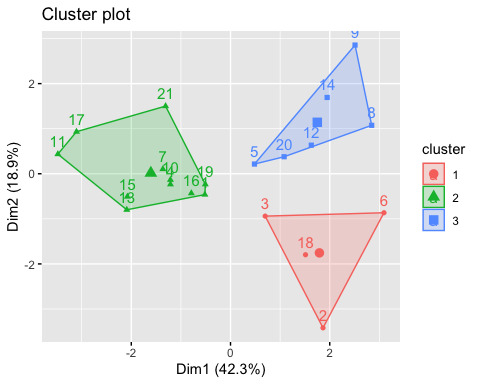
#part a) Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on  
  
#Select only the 9 numerical variables  
DF <- select(DF, Market\_Cap, Beta, PE\_Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev\_Growth, Net\_Profit\_Margin)  
  
# Scaling the data frame (z-score)   
DF <- scale(DF)  
distance <- get\_dist(DF)  
fviz\_dist(distance)

 The graph shows the distance between pharmaceutical companies. We will now cluster them with k-means algorithm starting with k=3 to analyze in 3 different clusters. Three clusters in our analysis makes sense from an equities analysis point of view because perhaps these companies could be analyzed as a natural categorization (small / mid / large caps, value / growth / quality stock, or perhaps underperformer / benchmark / overperformer).

#run kmeans to cluster the data  
k3 <- kmeans(DF, centers = 3, nstart = 25) # k = 3, number of restarts = 25  
# Visualize the output  
k3$size # Number of companies in each cluster

## [1] 4 11 6

fviz\_cluster(k3, data = DF) # Visualize the output



The output shows three distinct clusters.

Cluster 1 contains 4 companies, Cluster 2 contains 11 companies, and Cluster 3 contains 6. Since Cluster 2 contains over 50% of the sample, we’ll try k=4 for a better distribution of clusters.

#run kmeans to cluster the data  
k4 <- kmeans(DF, centers = 4, nstart = 25) # k = 3, number of restarts = 25  
# Visualize the output  
k4$size # Number of companies in each cluster

## [1] 8 4 3 6

fviz\_cluster(k4, data = DF) # Visualize the output



Cluster 1 now contains 8, Cluster 2 contains 4, Cluster 3 contains 3, and Cluster 4 contains 6 companies. The clusters are more balanced, with no cluster containing more than 50% and the largest cluster is not 3x larger than the smallest number (which is the case in k=5).

# b) Interpret the clusters with respect to the numerical variables used in forming the clusters.  
k4$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.03142211 -0.4360989 -0.3172485 0.1950459 0.4083915 1.729746e-01  
## 2 1.69558112 -0.1780563 -0.1984582 1.2349879 1.3503431 1.153164e+00  
## 3 -0.52462814 0.4451409 1.8498439 -1.0404550 -1.1865838 1.480297e-16  
## 4 -0.82617719 0.4775991 -0.3696184 -0.5631589 -0.8514589 -9.994088e-01  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.2744931 -0.7041516 0.5569544  
## 2 -0.4680782 0.4671788 0.5912425  
## 3 -0.3443544 -0.5769454 -1.6095439  
## 4 0.8502201 0.9158889 -0.3319956

Interpreting the data, we can indeed categorize many of these clusters into the categories we identified as potential above.

Cluster 1 generally has market caps which are neither high or low looking at the centers output, and no company is over 100B market cap or under 6.3B. They have the lowest Beta among the categories and among the highest net profit. Their PE ratio and revenue growth are both low. These stocks are less risky, but have limited upside.

Cluster 2 contains the largest market caps among the companies. They also have the highest profit margin paired with high revenue growth. While these companies are more risky (Beta) and more expensive (PE\_Ratio) than Cluster 1, they are growing faster and are likely industry leaders with more potential upside than Cluster 1.

Cluster 3 contains a cluster of risky and generally underperforming companies. These companies have the lowest net profit margins, are high beta (risk), smaller than average market cap, and have the highest price to earnings ratio. They’re the most expensive cluster according to PE ratios. They are the worst performers according to net profit margin, and are growing well under average according to revenue growth. This cluster seems popular by price, but has the highest potential for downside from current market price. The equity analyst should strongly consider avoiding these stocks on the long side. They should also consider them as a sell relative to the sample.

Cluster 4 contains a cluster of the smallest caps. They’re also the fastest growing companies of the entire set. While they’re still earning a negative net margin, it’s not nearly as bad as Cluster 3 and the growth potential is promising. These are the riskiest with the highest beta and highest leverage of the group, but their low P/E ratio is attractive given their high revenue growth. Aggressive boom or bust bet with the highest upside and also highest risk of bust.

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

There is indeed a correlation to the unused columns. Specifically I looked at median recommendation. Cluster 1 most likely falls into the “Hold” category, consistent with lower risk, value stocks. Value stocks typically don’t grow or shrink quickly, but they provide stable dividends. That is why many investors hold them in long term. Cluster 2 is likely to be a buy or a hold, consistent with high-growth large-cap industry leaders. They are only categorized as buy (50%) or hold (50%). Cluster 3 surprisingly also contained only buy or hold recommendations (and the only STRONG buy recommendation). These stocks could be popular or well-hyped given their PE ratio relative to their high beta, and low revenue growth and net profit margin. Perhaps our analysis is inaccurate and the analysts see or know more in these companies than the data supplies. However if our analysis is correct, the analysts recommended poor stocks for investors. It could simply be that the price is going up, so analysts are recommending them. Similar to a “momentum stock.” The last cluster, Cluster 4, accurately captures the high risk / high reward of this cluster. Cluster 4 contains the most recommended buys and ties for the most recommended sells. Only 1 stock in Cluster 4 is a recommended hold, reinforcing the analysis that these are more risky stocks than, say, Cluster 1.

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

Cluster 1 - Mid cap value, dividend stocks. Conservative and low growth. Generally worth holding in a mature portfolio with low risk tolerance.

Cluster 2 - Large cap growth, blue chips. Industry leaders, buy and hold. Generally worth holding in any portfolio.

Cluster 3 - Mid cap blend, momentum stocks. Overpriced for the risk according to the data, but high price and high interest for other reasons unknown to the data. Generally worth having positions in a short term trading portfolio (both long and short).

Cluster 4 - Small cap growth. Boom or bust potential. Worth holding in small amounts or in an immature portfolio with higher risk tolerance.