Assignment 4

Melissa Dennis

2025-10-26

##Load Libraries  
  
library(factoextra)

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

## Loading required package: ggplot2

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

library(cluster)  
library(dplyr)

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

##   
## 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(skimr)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.2.1  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0  
## ✔ readr 2.1.4

## ── 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(kableExtra)

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

##   
## Attaching package: 'kableExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## group\_rows

library(caret)

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

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(e1071)

## Warning: package 'e1071' was built under R version 4.3.1

##   
## Attaching package: 'e1071'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## element

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.3.3

##   
## Attaching package: 'reshape2'  
##   
## The following object is masked from 'package:tidyr':  
##   
## smiths

library(reshape)

##   
## Attaching package: 'reshape'  
##   
## The following objects are masked from 'package:reshape2':  
##   
## colsplit, melt, recast  
##   
## The following object is masked from 'package:lubridate':  
##   
## stamp  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, smiths  
##   
## The following object is masked from 'package:dplyr':  
##   
## rename

library(crayon)

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

##   
## Attaching package: 'crayon'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## %+%

##Load data  
  
df <- read.csv("C:/Users/m\_den/Downloads/Pharmaceuticals.csv")

##Process data  
  
row.names(df) <- df[,1]  
df <- df[,-1]

##Prnit question  
  
cat("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.")

## 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.

##Check variables, then normalize data  
  
df.norm <- df %>%  
 select\_if(is.numeric) %>%  
 scale()

##Run kmeans algorithm  
  
km <- kmeans(df.norm, 9)

##Show cluster membership  
  
km$cluster

## ABT AGN AHM AZN AVE BAY BMY CHTT ELN LLY GSK IVX JNJ MRX MRK NVS   
## 2 8 2 2 6 7 2 1 6 5 9 6 3 6 3 5   
## PFE PHA SGP WPI WYE   
## 9 8 2 6 4

##Centroids  
  
km$centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.9767669 1.2630872 0.03299122 -0.11237924 -1.1677918 -0.4612656  
## 2 -0.2063280 -0.2481660 -0.33855413 -0.03813318 0.4069821 0.6457718  
## 3 1.6312072 -0.2559560 -0.11108990 0.58367759 0.9651061 1.3837968  
## 4 -0.1614497 0.4061910 -0.75792214 1.92938746 0.5422849 -0.4612656  
## 5 0.4708563 -1.3270762 -0.04364767 -0.08917735 0.3449684 -0.6918984  
## 6 -0.7960593 0.3205014 -0.45014035 -0.65331484 -0.7881923 -1.1070374  
## 7 -0.6953818 2.2757827 0.14948233 -1.45146000 -1.7127612 -0.4612656  
## 8 -0.4392513 -0.4701800 2.70002464 -0.83495252 -0.9234951 0.2306328  
## 9 1.7599550 -0.1001567 -0.28582657 1.88629822 1.7355802 0.9225312  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 3.7427970 -0.6327607102 -1.24888417  
## 2 -0.4271213 -0.4707453496 0.15311712  
## 3 -0.5064752 -0.0009913151 0.04644311  
## 4 0.6838330 -1.1776391857 1.49416183  
## 5 -0.3720856 -1.0509232612 1.09794407  
## 6 0.2717048 1.2256188344 -0.14861794  
## 7 -0.7496565 -1.4971443383 -1.99560225  
## 8 -0.1417034 -0.1168458746 -1.41651476  
## 9 -0.4296811 0.9353488559 1.13604194

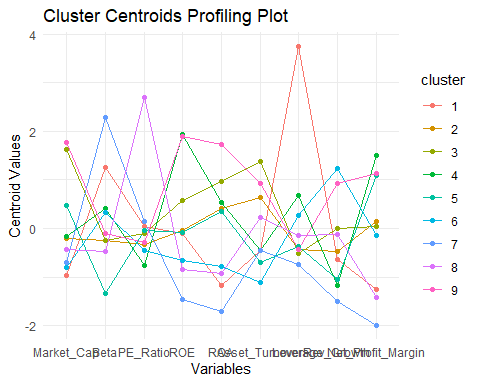
##Within-cluster sum of squares  
  
km$withinss

## [1] 0.000000 6.586586 1.645632 0.000000 1.244968 16.542597 0.000000  
## [8] 2.803505 2.994337

##Cluster size  
  
km$size

## [1] 1 5 2 1 2 5 1 2 2

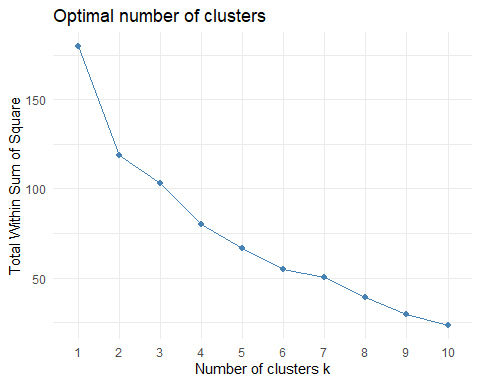
##Cluster centroids plot  
  
centers <- as.data.frame(km$centers)  
centers$cluster <- rownames(centers)  
centers.melt <- melt(centers, id.vars = "cluster")  
ggplot(centers.melt, aes(x = variable, y = value, group = cluster, color = cluster)) +  
 geom\_line() +  
 geom\_point() +  
 labs(title = "Cluster Centroids Profiling Plot", x = "Variables", y = "Centroid Values") +  
 theme\_minimal()



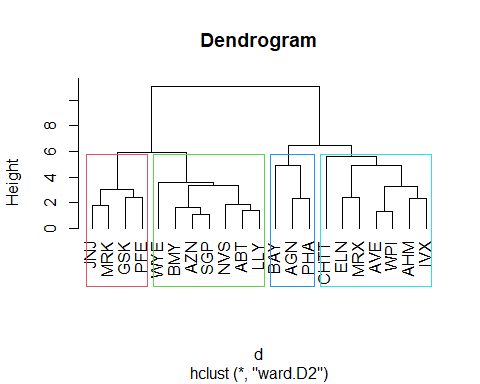
##Euclidean distance between clusters  
  
dist(km$centers)

## 1 2 3 4 5 6 7 8  
## 2 5.109456   
## 3 6.140122 2.215742   
## 4 5.127981 3.041934 3.789676   
## 5 5.810966 2.176635 3.141679 3.060875   
## 6 4.326767 3.003315 4.392667 4.243415 3.662034   
## 7 4.969438 4.509825 5.751103 5.954379 5.516608 4.239064   
## 8 5.181731 3.818951 4.629351 5.774966 4.399390 4.012153 4.266509   
## 9 7.024622 3.518438 2.155100 3.653032 3.952024 5.089393 7.255098 6.045571

##K-means clustering using kmeans function  
set.seed(123)  
wss <- numeric(9)  
for (i in 1:9) {  
 df.kmeans <- kmeans(df.norm, centers = i, nstart = 5)  
 wss[i] <- df.kmeans$tot.withinss  
}  
  
##Plot euclidean distance between clusters  
  
fviz\_nbclust(df.norm, kmeans, method = "wss") + theme\_minimal()



##Hierarchical clustering, using dendrogram to find the best k  
  
d <- dist(df.norm, method = "euclidean")  
hc <- hclust(d, method = "ward.D2")  
plot(hc, main = "Dendrogram", hang = -1)  
  
##highlight the clusters  
  
rect.hclust(hc, k = 4, border = 2:5)



##Print question  
  
cat("B. Interpret the clusters with respect to the numerical variables used in forming the clusters.")

## B. Interpret the clusters with respect to the numerical variables used in forming the clusters.

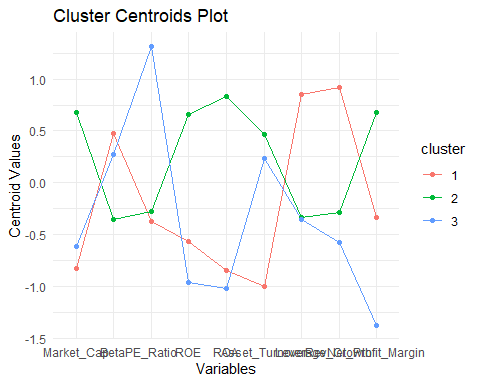
##Centroids of clusters  
  
km$centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.9767669 1.2630872 0.03299122 -0.11237924 -1.1677918 -0.4612656  
## 2 -0.2063280 -0.2481660 -0.33855413 -0.03813318 0.4069821 0.6457718  
## 3 1.6312072 -0.2559560 -0.11108990 0.58367759 0.9651061 1.3837968  
## 4 -0.1614497 0.4061910 -0.75792214 1.92938746 0.5422849 -0.4612656  
## 5 0.4708563 -1.3270762 -0.04364767 -0.08917735 0.3449684 -0.6918984  
## 6 -0.7960593 0.3205014 -0.45014035 -0.65331484 -0.7881923 -1.1070374  
## 7 -0.6953818 2.2757827 0.14948233 -1.45146000 -1.7127612 -0.4612656  
## 8 -0.4392513 -0.4701800 2.70002464 -0.83495252 -0.9234951 0.2306328  
## 9 1.7599550 -0.1001567 -0.28582657 1.88629822 1.7355802 0.9225312  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 3.7427970 -0.6327607102 -1.24888417  
## 2 -0.4271213 -0.4707453496 0.15311712  
## 3 -0.5064752 -0.0009913151 0.04644311  
## 4 0.6838330 -1.1776391857 1.49416183  
## 5 -0.3720856 -1.0509232612 1.09794407  
## 6 0.2717048 1.2256188344 -0.14861794  
## 7 -0.7496565 -1.4971443383 -1.99560225  
## 8 -0.1417034 -0.1168458746 -1.41651476  
## 9 -0.4296811 0.9353488559 1.13604194

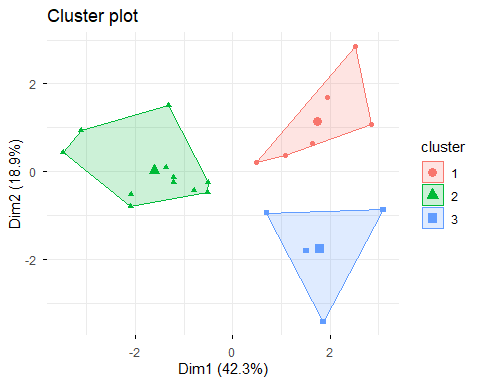
##K-means clustering, using k = 3  
  
df.k3 <- kmeans(df.norm, centers = 3, nstart = 5)  
df.k3

## K-means clustering with 3 clusters of sizes 6, 11, 4  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.8261772 0.4775991 -0.3696184 -0.5631589 -0.8514589 -0.9994088  
## 2 0.6733825 -0.3586419 -0.2763512 0.6565978 0.8344159 0.4612656  
## 3 -0.6125361 0.2698666 1.3143935 -0.9609057 -1.0174553 0.2306328  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 0.8502201 0.9158889 -0.3319956  
## 2 -0.3331068 -0.2902163 0.6823310  
## 3 -0.3592866 -0.5757385 -1.3784169  
##   
## Clustering vector:  
## ABT AGN AHM AZN AVE BAY BMY CHTT ELN LLY GSK IVX JNJ MRX MRK NVS   
## 2 3 3 2 1 3 2 1 1 2 2 1 2 1 2 2   
## PFE PHA SGP WPI WYE   
## 2 3 2 1 2   
##   
## Within cluster sum of squares by cluster:  
## [1] 32.14336 43.30886 20.54199  
## (between\_SS / total\_SS = 46.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

##Plot the clustered centroids values  
  
centers <- as.data.frame(df.k3$centers)  
centers$cluster <- rownames(centers)  
centers.melt <- melt(centers, id.vars = "cluster")  
ggplot(centers.melt, aes(x = variable, y = value, group = cluster, color = cluster)) +  
 geom\_line() +  
 geom\_point() +  
 labs(title = "Cluster Centroids Plot", x = "Variables", y = "Centroid Values") +  
 theme\_minimal()



##Cluster centroids plot  
##Visualize the clusters with symbol variable as label of the nodes  
  
fviz\_cluster(df.k3, data = df.norm, geom = "point", ellipse.type = "convex", ellipse = TRUE, label = "symbol", ggtheme = theme\_minimal())



##Print interpretation  
  
cat("Cluster 1 is defined by its extreme volatility and debt. They have the highest Leverage (0.85), meaning they carry a lot of debt, and very fast Rev Growth (0.92). But they are also the worst performers financially: they have very low market size (Market\_Cap: -0.83), extremely poor asset utilization (Asset\_Turnover: -1.00), and poor returns (ROE: -0.56, ROA: -0.85). They seem to be younger, aggressive companies focused on expansion at any cost, using high debt and sacrificing all profitability. Firms like AVE, CHTT, ELN, IVX, MRX, and WPI land here. \n \nCluster 2 is the group of solid, reliable winners. They have the highest profitability and efficiency metrics, like high ROE (0.66), high ROA (0.83), and the highest Net Profit Margin (0.68). They are the largest companies (Market\_Cap: 0.67) and are efficient with their assets (Asset\_Turnover: 0.46). They also have the lowest risk (Beta: -0.36) and low debt. These are large, mature, well-managed firms with excellent returns. This category includes JNJ, PFE, MRK, and a majority of the sample. \n \nCluster 3 is a concerning group that is highly unprofitable but highly valued by investors. They have a massive PE\_Ratio (1.31), meaning their stock price is high relative to their earnings. The reason is they are highly unprofitable, showing the lowest ROA (-1.02) and the lowest Net Profit Margin (-1.38). Their low returns and low market size suggest they are struggling companies that may be overvalued, potentially based on old growth expectations or speculation. AGN, AHM, BAY, and PHA are in this group.")

## Cluster 1 is defined by its extreme volatility and debt. They have the highest Leverage (0.85), meaning they carry a lot of debt, and very fast Rev Growth (0.92). But they are also the worst performers financially: they have very low market size (Market\_Cap: -0.83), extremely poor asset utilization (Asset\_Turnover: -1.00), and poor returns (ROE: -0.56, ROA: -0.85). They seem to be younger, aggressive companies focused on expansion at any cost, using high debt and sacrificing all profitability. Firms like AVE, CHTT, ELN, IVX, MRX, and WPI land here.   
##   
## Cluster 2 is the group of solid, reliable winners. They have the highest profitability and efficiency metrics, like high ROE (0.66), high ROA (0.83), and the highest Net Profit Margin (0.68). They are the largest companies (Market\_Cap: 0.67) and are efficient with their assets (Asset\_Turnover: 0.46). They also have the lowest risk (Beta: -0.36) and low debt. These are large, mature, well-managed firms with excellent returns. This category includes JNJ, PFE, MRK, and a majority of the sample.   
##   
## Cluster 3 is a concerning group that is highly unprofitable but highly valued by investors. They have a massive PE\_Ratio (1.31), meaning their stock price is high relative to their earnings. The reason is they are highly unprofitable, showing the lowest ROA (-1.02) and the lowest Net Profit Margin (-1.38). Their low returns and low market size suggest they are struggling companies that may be overvalued, potentially based on old growth expectations or speculation. AGN, AHM, BAY, and PHA are in this group.

##Print questions  
  
cat("C. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in  
forming the clusters)")

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

##Add cluster variable to Pharmaceuticals data  
  
df$cluster <- df.k3$cluster  
  
##Table of cluster variable by location, stock exchange, and median recommendation  
  
table(df.k3$cluster, df$Location)

##   
## CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US  
## 1 0 1 0 1 0 0 4  
## 2 0 0 0 0 1 2 8  
## 3 1 0 1 0 0 1 1

table(df.k3$cluster, df$Exchange)

##   
## AMEX NASDAQ NYSE  
## 1 1 1 4  
## 2 0 0 11  
## 3 0 0 4

table(df.k3$cluster, df$Median\_Recommendation)

##   
## Hold Moderate Buy Moderate Sell Strong Buy  
## 1 1 3 2 0  
## 2 6 3 2 0  
## 3 2 1 0 1

##Print interpretation  
  
cat("The clusters show clear differences based on where the companies are from, where they trade, and what analysts think. For Cluster 2, they're mostly from stable places like the US and France, trade heavily on the NYSE, and get a mix of Hold and Moderate Buy ratings, suggesting they're solid but maybe fully priced. Cluster 1 are a US and Canadian mix, trade only on the NYSE, and analysts are cautious, giving them Hold or Moderate Sell ratings because of their high-risk profile. Cluster 3 are mostly from developed markets like the US and UK, trade mainly on the NYSE, and yet, despite struggling financially, they get ratings as high as Strong Buy. This strange mix of bad financials and good ratings helps explain why they have a super high PE ratio, investors or analysts might be betting big on a turnaround that hasn't happened yet.")

## The clusters show clear differences based on where the companies are from, where they trade, and what analysts think. For Cluster 2, they're mostly from stable places like the US and France, trade heavily on the NYSE, and get a mix of Hold and Moderate Buy ratings, suggesting they're solid but maybe fully priced. Cluster 1 are a US and Canadian mix, trade only on the NYSE, and analysts are cautious, giving them Hold or Moderate Sell ratings because of their high-risk profile. Cluster 3 are mostly from developed markets like the US and UK, trade mainly on the NYSE, and yet, despite struggling financially, they get ratings as high as Strong Buy. This strange mix of bad financials and good ratings helps explain why they have a super high PE ratio, investors or analysts might be betting big on a turnaround that hasn't happened yet.

##Print question  
  
cat("D. Provide an appropriate name for each cluster using any or all of the variables in the dataset")

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

##Print the names for each cluster  
  
cat(bold("Cluster 1:"),"Volatile Firms\n",bold("Cluster 2:"),"Reliable Firms","\n",bold("Cluster 3:"),"Overvalued Firms")

## Cluster 1: Volatile Firms  
## Cluster 2: Reliable Firms   
## Cluster 3: Overvalued Firms