BA 64060 Assignment 4

Kyle Breeden

2025-10-26

## R Markdown

#rm(list = ls())  
#install.packages("tidyverse") # if necessary  
library(tidyverse) # data manipulation

## Warning: package 'tidyverse' was built under R version 4.5.1

## Warning: package 'readr' was built under R version 4.5.1

## Warning: package 'forcats' was built under R version 4.5.1

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.1 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.3.0  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── 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

#install.packages("factoextra") # if necessary  
library(factoextra) # clustering algorithms & visualization

## Warning: package 'factoextra' was built under R version 4.5.1

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

#install.packages("ISLR") # if necessary  
library(ISLR)  
#install.packages("readr") # if necessary  
set.seed(123)  
library(readr)  
assign4\_dataset <- read\_csv("C:/Users/kylej/OneDrive/Documents/BA 64060/Assignment 4/Pharmaceuticals.CSV")

## Rows: 21 Columns: 14  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Symbol, Name, Median\_Recommendation, Location, Exchange  
## dbl (9): Market\_Cap, Beta, PE\_Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(assign4\_dataset)

## 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   
##   
##   
##

#Question a  
##use descriptive statistics above to identify numeric columns.  
assign4\_numeric <- assign4\_dataset[, 3:11]  
##check for missing values  
colSums(is.na(assign4\_numeric))

## Market\_Cap Beta PE\_Ratio ROE   
## 0 0 0 0   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## 0 0 0 0   
## Net\_Profit\_Margin   
## 0

##Generate correlation matrix to assess relationships within numerical variables  
cor(assign4\_numeric)

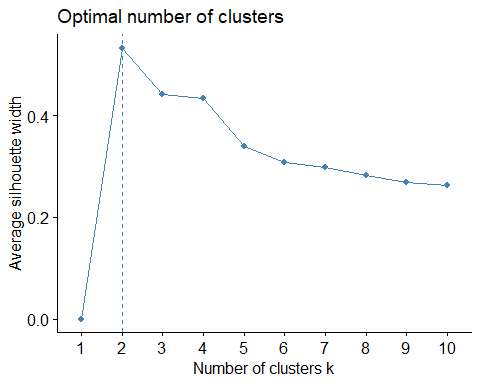
## Market\_Cap Beta PE\_Ratio ROE ROA  
## Market\_Cap 1.000000000 -0.31250762 -0.08798317 0.61952576 0.80908852  
## Beta -0.312507620 1.00000000 -0.19716312 -0.20273345 -0.42583638  
## PE\_Ratio -0.087983169 -0.19716312 1.00000000 -0.32205434 -0.29207790  
## ROE 0.619525759 -0.20273345 -0.32205434 1.00000000 0.83168600  
## ROA 0.809088517 -0.42583638 -0.29207790 0.83168600 1.00000000  
## Asset\_Turnover 0.507917513 -0.32069694 0.14974635 0.49612507 0.61977107  
## Leverage -0.408937481 0.40116206 -0.03985770 0.01560562 -0.36535802  
## Rev\_Growth 0.003788982 0.08807135 -0.15499183 -0.01905389 -0.02118403  
## Net\_Profit\_Margin 0.516711077 -0.34546582 -0.46240116 0.63395830 0.74875756  
## Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin  
## Market\_Cap 0.50791751 -0.40893748 0.003788982 0.51671108  
## Beta -0.32069694 0.40116206 0.088071348 -0.34546582  
## PE\_Ratio 0.14974635 -0.03985770 -0.154991834 -0.46240116  
## ROE 0.49612507 0.01560562 -0.019053892 0.63395830  
## ROA 0.61977107 -0.36535802 -0.021184032 0.74875756  
## Asset\_Turnover 1.00000000 -0.30817546 -0.253024565 0.01862763  
## Leverage -0.30817546 1.00000000 -0.021881004 -0.22135214  
## Rev\_Growth -0.25302457 -0.02188100 1.000000000 0.08478937  
## Net\_Profit\_Margin 0.01862763 -0.22135214 0.084789374 1.00000000

##ROE, ROA, and Net\_Profit\_Margin are strongly correlated but the rest are moderatly or weakly coorelated. I will use Euclidean distance since most variables are moderatley or weakly correlated and scale the vartiables using z-score.

##Scale the data frame (z-score)   
assign4\_scaled <- scale(assign4\_numeric)  
##Use an elbow chart to determine k  
fviz\_nbclust(assign4\_scaled, kmeans, method = "wss")

 ##The elbow point at k = 2 provides the best value for k becasue the total within-cluster sum of squares decreases sharply up to this point but at a much slower rate beyond it.

##Use silhouette method to determine the number of clusters  
fviz\_nbclust(assign4\_numeric, kmeans, method = "silhouette")

 ##The chart above shows that the optimal number of clusters is k = 2, because this value achieves the largest silhouette score among all tested cluster counts.

##Run k-Means where K = 2  
set.seed(123)  
k2 <- kmeans(assign4\_scaled, centers = 2, nstart = 25)  
  
  
##Visualize the output  
k2$centers #output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 0.6733825 -0.3586419 -0.2763512 0.6565978 0.8344159 0.4612656  
## 2 -0.7407208 0.3945061 0.3039863 -0.7222576 -0.9178575 -0.5073922  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.3331068 -0.2902163 0.6823310  
## 2 0.3664175 0.3192379 -0.7505641

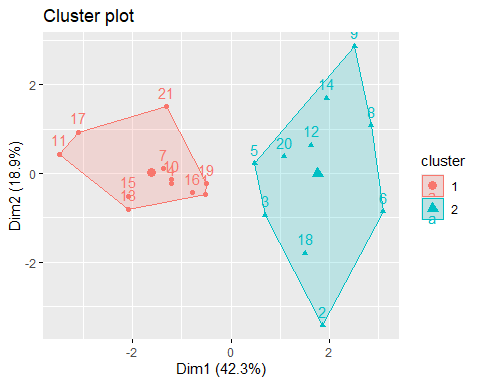
k2$size #Number of firms in each cluster

## [1] 11 10

k2$cluster[15] #Identify the cluster of the 15th observation as an example

## [1] 1

fviz\_cluster(k2, data = assign4\_scaled,) #Visualize the output



#Add cluster labels into original data set  
assign4\_clusters <- data.frame(assign4\_dataset, Cluster = k2$cluster)  
#I tried to add the symbol as the label & re-generate the cluster plot, but could not get it to work  
#Question b  
##Use the means of each variable to compare numerical variables in each cluster  
library(dplyr)  
cluster\_summary <- assign4\_clusters %>% group\_by(Cluster) %>% summarise(across(Market\_Cap:Net\_Profit\_Margin, mean))

##Cluster 1 represents the pharma firms with higher market cap, ROE, ROA, and Net Profit Margin and lower leverage, beta, and revenue growth. Clauster 2 represents the pharma fims with higher beta and revenue growth and lower market cap, ROE, ROA, and Net Profit Margin. The pharma firms in cluster 1 are large, profitable, and stable (lower beta, lower but steady revenue growth, and lower leverage). The pharma firms in cluster 2 are smaller growing companies with more risk.

#Question c  
##generate tables to see distributions   
table(assign4\_clusters$Cluster, assign4\_clusters$Median\_Recommendation)

##   
## Hold Moderate Buy Moderate Sell Strong Buy  
## 1 6 3 2 0  
## 2 3 4 2 1

table(assign4\_clusters$Cluster, assign4\_clusters$Location)

##   
## CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US  
## 1 0 0 0 0 1 2 8  
## 2 1 1 1 1 0 1 5

table(assign4\_clusters$Cluster, assign4\_clusters$Exchange)

##   
## AMEX NASDAQ NYSE  
## 1 0 0 11  
## 2 1 1 8

##Yes, patterns among the non-clustering variables support the financial distinctions identified earlier. Cluster 1 firms are predominantly U.S.-based, NYSE-listed, and viewed as stable performers. Cluster 2 firms are more geographically diverse, include listings on growth-oriented exchanges, and receive more favorable analyst recommendations, indicating higher perceived growth potential.

#Question d  
#Cluster 1: Established U.S. Pharma Firms  
#Cluster 2: Emerging Pharma Firms  
k2\_named <- k2  
k2\_named$cluster <- factor(k2$cluster,  
 levels = c(1, 2),  
 labels = c("Established U.S. Pharma Firms",  
 "Emerging Pharma Firms"))  
  
fviz\_cluster(k2\_named,   
 data = assign4\_scaled,  
 ggtheme = theme\_minimal(),  
 main = "K-Means Clusters of Pharmaceutical Firms (k = 2)")

