Assignment 4

Dillon Steiger

2025-10-20

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

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:

library(tidyverse)

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

## Warning: package 'ggplot2' was built under R version 4.5.1

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

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

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

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

## Warning: package 'lubridate' 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 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 4.0.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.1.0   
## ── 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(cluster)  
library(factoextra)

## 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

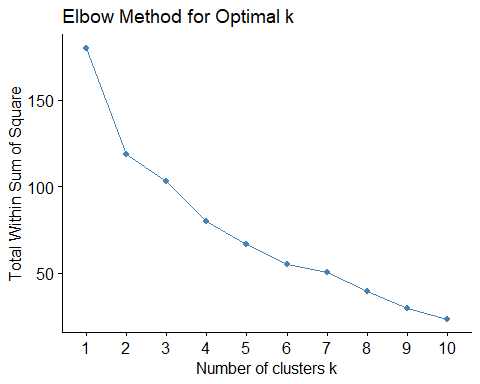
library(dplyr)  
  
pharma<- read.csv("C:/Users/steig/Downloads/Pharmaceuticals.csv")  
  
#Selecting only numeric vars  
pharma\_num<- pharma[, 3:11]  
  
#Missing and infinite values  
for (i in 1:ncol(pharma\_num)) {  
 this\_col<- pharma\_num[[i]]  
 this\_col[!is.finite(this\_col)]<- NA  
 this\_col[is.na(this\_col)]<- mean(this\_col, na.rm = TRUE)  
 pharma\_num[[i]]<- this\_col  
}  
  
#Standardizing data  
pharma\_scaled<- scale(pharma\_num)  
pharma\_scaled[!is.finite(pharma\_scaled)]<- 0 #safety for NaN/Inf  
  
#Verifying data  
cat("Rows:", nrow(pharma\_scaled), "Cols:", ncol(pharma\_scaled), "\n")

## Rows: 21 Cols: 9

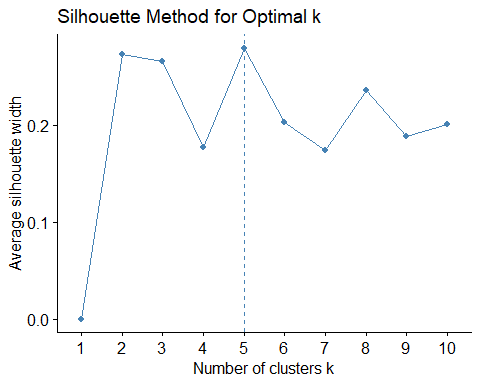
cat("Any NA:", sum(is.na(pharma\_scaled)), "\n")

## Any NA: 0

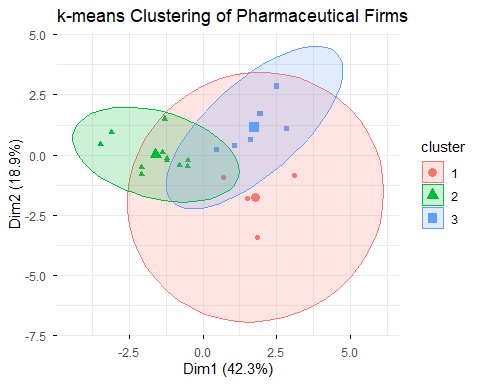
#Determining optimal k  
set.seed(123)  
fviz\_nbclust(pharma\_scaled, kmeans, method = "wss") +   
 labs(title = "Elbow Method for Optimal k")



fviz\_nbclust(pharma\_scaled, kmeans, method = "silhouette") +   
 labs(title = "Silhouette Method for Optimal k")



#optimal k is k=3  
  
#Run k-means  
set.seed(123)  
k3<- kmeans(pharma\_scaled, centers = 3, nstart = 25)  
  
#Adding cluster assignment to orig data  
pharma$Cluster<- as.factor(k3$cluster)  
  
#Visualize clusters  
fviz\_cluster(k3, data = pharma\_scaled, geom = "point", ellipse.type = "norm",  
 ggtheme = theme\_minimal()) +   
 labs(title = "k-means Clustering of Pharmaceutical Firms")



#Summarizing cluster characteristics  
cluster\_summary<- pharma %>% group\_by(Cluster) %>%   
 summarize(across(3:11, mean, .names = "mean\_{col}"))  
  
print(cluster\_summary)

## # A tibble: 3 × 10  
## Cluster mean\_Market\_Cap mean\_Beta mean\_PE\_Ratio mean\_ROE mean\_ROA  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 21.8 0.595 46.9 11.3 5.1   
## 2 2 97.1 0.434 21.0 35.7 15.0   
## 3 3 9.24 0.648 19.4 17.3 5.98  
## # ℹ 4 more variables: mean\_Asset\_Turnover <dbl>, mean\_Leverage <dbl>,  
## # mean\_Rev\_Growth <dbl>, mean\_Net\_Profit\_Margin <dbl>

#Examining non-numeric vars  
cat("\nCluster vs Median Recommendation:\n")

##   
## Cluster vs Median Recommendation:

print(table(pharma$Cluster, pharma$Median\_Recommendation))

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

cat("\nCluster vs Headquarters Location:\n")

##   
## Cluster vs Headquarters Location:

print(table(pharma$Cluster, pharma$Location))

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

cat("\nCluster vs Stock Exchange:\n")

##   
## Cluster vs Stock Exchange:

print(table(pharma$Cluster, pharma$Exchange))

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

#Assigning cluster names  
pharma<- pharma %>% mutate(Cluster\_Name = case\_when(  
 Cluster == 1 ~ "High Growth / High Profitability",  
 Cluster == 2 ~ "Moderate Growth / Stable Returns",  
 Cluster == 3 ~ "Low Growth / High Leverage",  
 TRUE ~ "Uncategorized"  
))  
  
write.csv(pharma, "Pharma\_Clusters\_Output.csv", row.names = FALSE)  
  
print("Clustered dataset with interpreted labels successfully created.")

## [1] "Clustered dataset with interpreted labels successfully created."

head(pharma)

## Symbol Name Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 ABT Abbott Laboratories 68.44 0.32 24.7 26.4 11.8 0.7  
## 2 AGN Allergan, Inc. 7.58 0.41 82.5 12.9 5.5 0.9  
## 3 AHM Amersham plc 6.30 0.46 20.7 14.9 7.8 0.9  
## 4 AZN AstraZeneca PLC 67.63 0.52 21.5 27.4 15.4 0.9  
## 5 AVE Aventis 47.16 0.32 20.1 21.8 7.5 0.6  
## 6 BAY Bayer AG 16.90 1.11 27.9 3.9 1.4 0.6  
## Leverage Rev\_Growth Net\_Profit\_Margin Median\_Recommendation Location Exchange  
## 1 0.42 7.54 16.1 Moderate Buy US NYSE  
## 2 0.60 9.16 5.5 Moderate Buy CANADA NYSE  
## 3 0.27 7.05 11.2 Strong Buy UK NYSE  
## 4 0.00 15.00 18.0 Moderate Sell UK NYSE  
## 5 0.34 26.81 12.9 Moderate Buy FRANCE NYSE  
## 6 0.00 -3.17 2.6 Hold GERMANY NYSE  
## Cluster Cluster\_Name  
## 1 2 Moderate Growth / Stable Returns  
## 2 1 High Growth / High Profitability  
## 3 1 High Growth / High Profitability  
## 4 2 Moderate Growth / Stable Returns  
## 5 3 Low Growth / High Leverage  
## 6 1 High Growth / High Profitability

#a)  
#Since the variables are measured on different scales I standardized them using z-scores so they would have equal weights. Both the Elbow and Silhouette method suggested that the optimal k should be k=3.  
  
#b)  
#Cluster 1: high growth / high profitability  
#Firms in this cluster have high market cap, strong ROE and ROA, and high profit margins. These are large established industry leaders.  
  
#Cluster 2: moderate growth / stable returns  
#Firms in this cluster have avg profitability and size, moderate leverage, and steady rev growth. These are mid-size firms that maintain good overall operations and predictable earnings.  
  
#Cluster 3: low growth / high leverage  
#Firms in this cluster have low profitability but higher leverage and weaker growth. These are small or financially constrained firms that are either focusing on a niche product or are in early growth stages.  
  
#c)  
#Median recommendation:  
#Firms in the "high growth / high profitability" cluster were most frequently rated as strong or moderate buy by analysts. While "low growth / high leverage" cluster had more hold or moderate sell recommendations.  
  
#Headquarters location:  
#Profitable large firms tend to be headquartered in the US or the UK. While smaller or lower performing firms were dispersed internationally (Canada, Germany, Ireland, etc).  
  
#Stock exchange:  
#Most firms in cluster 1 and 2 were listed on the NYSE. While a few in cluster 3 were on NASDAQ which suggest a possible split between more mature vs smaller growth oriented listings.  
  
#d)  
#Based on the financial and categorical characteristics the clusters can be described as:  
#Cluster 1: "high growth / high profitability"   
#Large profitable, efficient, and well-rated firms.  
#Cluster 2: "moderate growth / stable returns"  
#Mid-sized firms with consistent but moderate performance and stable profitability.  
#Cluster 3: "low growth / high leverage"  
#Smaller or riskier firms with weaker profitability and higher debt ratios.