Assignment 4

K-Means for Clustering

Hollister Victor

BA 64060: Fundamentals of Machine Learning

June 15,2025

This was used due to QuartzBitmap\_Output error being received.

#Global Chunk setting for knitting   
knitr::opts\_chunk$set(dev = "png", error = TRUE)

Load Libraries  
#Load Required Libraries  
library(readr)   
library(dplyr)

##   
## 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(cluster)   
library(factoextra)

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

library(flexclust)

Load data set:  
#Load Data set   
pharma <- read\_csv("/Users/hollyvictor/Downloads/Pharmaceuticals.csv")

## Rows: 21 Columns: 14

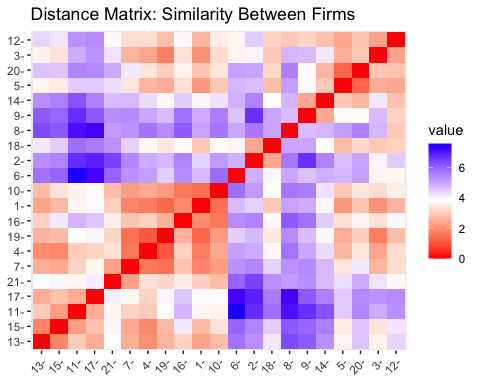
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Symbol, Name, Median\_Recommendation, Location, Exchange  
## dbl (9): Market\_Cap, Beta, PE\_Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev...  
##   
## ℹ Use `spec()` to retrieve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

Use only the numerical variables (1 to 9) to cluster the 21 firms

#Select Numeric Variables for the clustering   
num\_vars <- pharma %>%  
 select(Market\_Cap, Beta, PE\_Ratio, ROE, ROA,  
 Asset\_Turnover, Leverage, Rev\_Growth, Net\_Profit\_Margin)

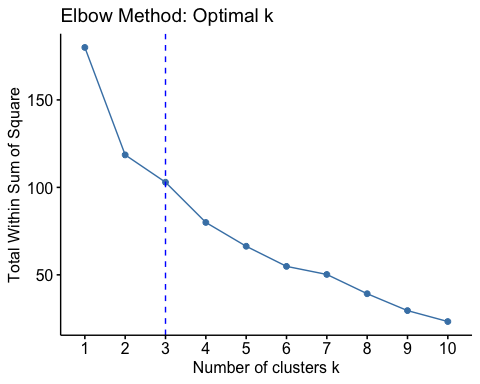
k-means is distance based. Features like market cap are very large and can dominate the data set. This ensures **equal weighting** in Euclidean distance calculations  
#scale numeric data  
num\_scaled <- scale(num\_vars)

Distance heatmap to show if there are clusters. The darker areas show similarities. The red diagonal should be disregarded as it is comparison against the same companies – so distance =0  
#visual the distance heat map   
fviz\_dist(dist(num\_scaled)) +  
 labs(title = "Distance Matrix: Similarity Between Firms")



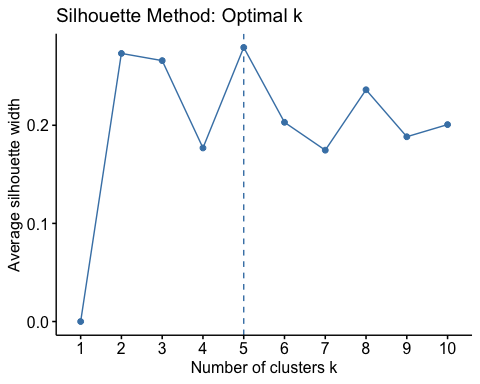
Elbow method:plots the points within cluster sum of squares (WSS) Elbow point suggests good k value.K=3 is optimal

#Determine optimal K   
fviz\_nbclust(num\_scaled, kmeans, method = "wss") +  
 geom\_vline(xintercept = 3, linetype = 2, color = "blue") +  
 labs(title = "Elbow Method: Optimal k")



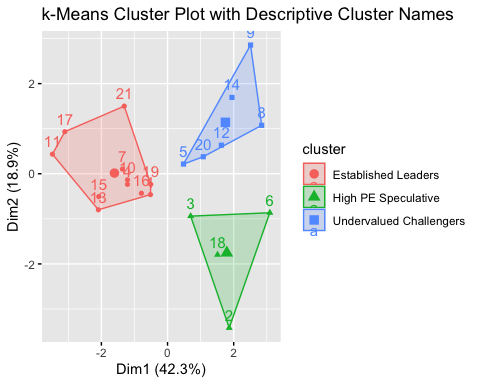
Silhouette method:measures how well a point lies within a cluster. The higher the score the better the clustering.In this case it suggests K=5

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



Using K=3 creates 3 clusters. Nstart=25 means it will try 25 random assignments and choose the best one.

#K-means K=3  
set.seed(123)  
km\_3 <- kmeans(num\_scaled, centers = 3, nstart = 25)  
pharma$Cluster3 <- as.factor(km\_3$cluster)  
  
#Add descriptive labels   
  
pharma <- pharma %>%  
 mutate(Cluster3\_Label = case\_when(  
 Cluster3 == 1 ~ "High PE Speculative",  
 Cluster3 == 2 ~ "Established Leaders",  
 Cluster3 == 3 ~ "Undervalued Challengers"  
 ))  
  
#Plot Cluster Results  
  
fviz\_cluster(list(data = num\_scaled, cluster = pharma$Cluster3\_Label)) +  
 labs(title = "k-Means Cluster Plot with Descriptive Cluster Names")



#Summarize clusters k=3  
summary\_3 <- pharma %>%  
 group\_by(Cluster3\_Label) %>%  
 summarise(across(Market\_Cap:Net\_Profit\_Margin, mean, .names = "mean\_{.col}"))  
print(summary\_3)

## # A tibble: 3 × 10  
## Cluster3\_Label mean\_Market\_Cap mean\_Beta mean\_PE\_Ratio mean\_ROE mean\_ROA  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Established Leaders 97.1 0.434 21.0 35.7 15.0   
## 2 High PE Speculative 21.8 0.595 46.9 11.3 5.1   
## 3 Undervalued Challen… 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>

#explore with K=5 this will test if finer segmentation will give better insight.  
set.seed(123)  
km\_5 <- kmeans(num\_scaled, centers = 5, nstart = 25)  
pharma$Cluster5 <- as.factor(km\_5$cluster)  
  
summary\_5 <- pharma %>%  
 group\_by(Cluster5) %>%  
 summarise(across(Market\_Cap:Net\_Profit\_Margin, mean, .names = "mean\_{.col}"))  
print(summary\_5)

## # A tibble: 5 × 10  
## Cluster5 mean\_Market\_Cap mean\_Beta mean\_PE\_Ratio mean\_ROE mean\_ROA  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 55.8 0.414 20.3 28.7 12.7   
## 2 2 6.64 0.87 24.6 16.5 4.17  
## 3 3 31.9 0.405 69.5 13.2 5.6   
## 4 4 157. 0.48 22.2 44.4 17.7   
## 5 5 13.1 0.598 17.7 14.6 6.2   
## # ℹ 4 more variables: mean\_Asset\_Turnover <dbl>, mean\_Leverage <dbl>,  
## # mean\_Rev\_Growth <dbl>, mean\_Net\_Profit\_Margin <dbl>

In summary:

Use **k = 3** for **executive-level segmentation** — it's clean, easy to understand, and still informative.Use **k = 5** if you want more **precision** or are making **targeted strategic decisions** (e.g., choosing between 2 small firms with slightly different ROEs and risk).For this assignment I moved forward using K=3

#Analyze Categorical relationships  
table(pharma$Cluster3, pharma$Median\_Recommendation)

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

table(pharma$Cluster3, pharma$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

table(pharma$Cluster3, pharma$Exchange)

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

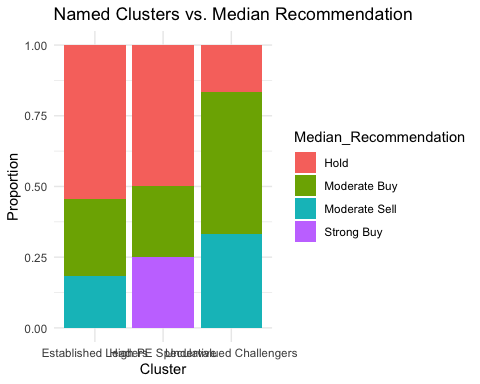
Analysis of Variables 10–12 (not used in clustering):

Median Recommendation(Analyst Rating) Location ( Country of Headquarters) Exchange (stock exchange NYSE NASDAQetc)

Established Leaders had more Hold Ratings. High PE Speculative had a mix of strong buy – investors expecting growth from these. Undervalued Challengers had mixed or less enthusiastic ratings. Broker sentiment aligns with the financial profiles of the clusters.

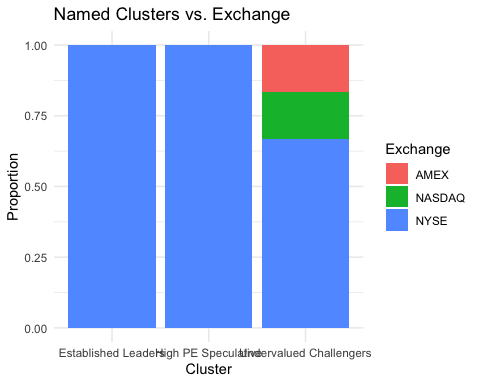
Regarding location established leaders tended to be mainly US based. This aligns with real-world patterns — many of the **largest and most mature pharmaceutical firms** areU.S.-based and listed on major U.S. exchanges like NYSE. There were more international firms in the undervalued challengers which suggests **rising firms in European markets** or **smaller U.S. players** with solid fundamentals but lower visibility or valuation.

#Plot Median Recommendation by cluster  
ggplot(pharma, aes(x = Cluster3\_Label, fill = Median\_Recommendation)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Named Clusters vs. Median Recommendation",   
 y = "Proportion", x = "Cluster") +  
 theme\_minimal()



Established leaders and PE Speculative: - NYSE and the Undervalued Challengers found on less traditional or smaller cap exchanges. This pattern aligns.

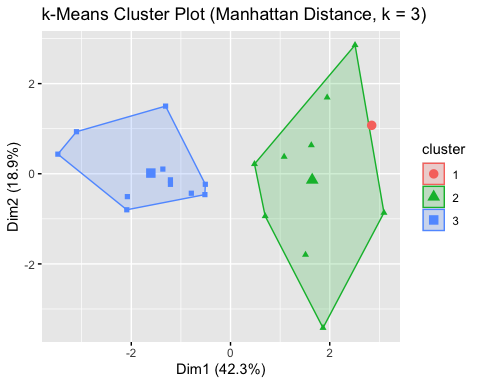
#Plot Exchange by Cluster  
ggplot(pharma, aes(x = Cluster3\_Label, fill = Exchange)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Named Clusters vs. Exchange",   
 y = "Proportion", x = "Cluster") +  
 theme\_minimal()



#additional exploration : using Manhattan Distance   
  
num\_matrix <- as.matrix(num\_scaled)  
km\_manhattan <- kcca(num\_matrix, k = 3, family = kccaFamily("kmeans", dist = "manhattan"))  
pharma$Cluster\_Manhattan <- as.factor(predict(km\_manhattan))  
  
#compare clusters  
table(pharma$Cluster3\_Label, pharma$Cluster\_Manhattan)

##   
## 1 2 3  
## Established Leaders 0 0 11  
## High PE Speculative 0 4 0  
## Undervalued Challengers 1 5 0

#plot Manhattan clustering   
print(  
 fviz\_cluster(  
 list(data = num\_scaled, cluster = predict(km\_manhattan)),  
 geom = "point"  
 ) +  
 labs(title = "k-Means Cluster Plot (Manhattan Distance, k = 3)")  
)



The robustness of the clustering results was assessed using k-means with Manhattan distance in place of the default Euclidean metric. This was appropriate because Manhattan distance, which measures the absolute differences across dimensions, might be a better choice when the clusters are not spherical or when the dataset contains outliers. The results showed strong consistency for two of the original clusters: all "Established Leaders" mapped cleanly to one cluster, and all "High PE Speculative" firms grouped together under a different cluster. This reinforces the strength and separation of those two segments. However, the "Undervalued Challengers" showed greater sensitivity—splitting us between two clusters under the Manhattan-based method. This indicates that this group may be more diverse in financial characteristics or that its boundaries are less clearly defined compared to the other two clusters.