Assignment\_4

Hemani Panchmatiya

2025-10-14

Introduction: In this assignment we have applied K-Means clustering an unsupervised learning technique to segment 21 Pharmaceutical firms with 9 numerical financial variables. By grouping firms into cluster, we identify financial profile, assets market perception and gain insights. The analysis also explores median recommendation, location and exchange to provide investor market positioning. Each cluster is assigned a descriptive name reflecting the characteristic, interpretation and decision-making.

#Loading the packages

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(factoextra)

## Loading required package: ggplot2

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

library(cluster)  
library(ggplot2)  
library(flexclust)

#Import data set

Pharma <- read.csv("Pharmaceuticals.csv")

#View the sturcture

str(Pharma)

## 'data.frame': 21 obs. of 14 variables:  
## $ Symbol : chr "ABT" "AGN" "AHM" "AZN" ...  
## $ Name : chr "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PLC" ...  
## $ Market\_Cap : num 68.44 7.58 6.3 67.63 47.16 ...  
## $ Beta : num 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...  
## $ PE\_Ratio : num 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...  
## $ ROE : num 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...  
## $ ROA : num 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...  
## $ Asset\_Turnover : num 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...  
## $ Leverage : num 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...  
## $ Rev\_Growth : num 7.54 9.16 7.05 15 26.81 ...  
## $ Net\_Profit\_Margin : num 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...  
## $ Median\_Recommendation: chr "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...  
## $ Location : chr "US" "CANADA" "UK" "UK" ...  
## $ Exchange : chr "NYSE" "NYSE" "NYSE" "NYSE" ...

Question 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 numerical variables 1–9

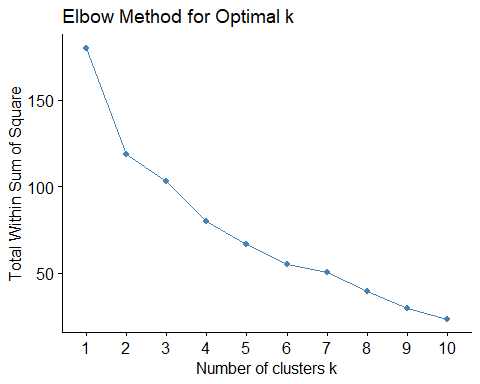
Pharma\_num <- Pharma[, 3:11]

#Standardizing/scale the data using Z- Score

Pharma\_scaled <- scale(Pharma\_num)

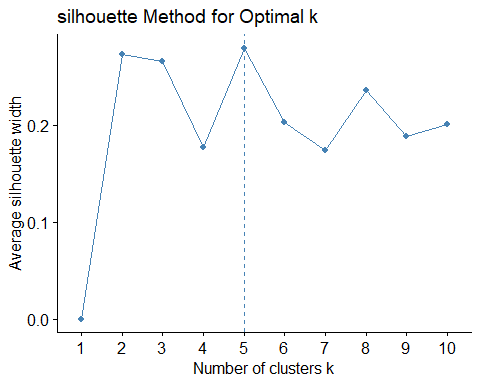
#Determing the optimal K #Elbow Method

fviz\_nbclust(Pharma\_scaled, kmeans, method = "wss") +  
 ggtitle("Elbow Method for Optimal k")



#Silhouette Method

fviz\_nbclust(Pharma\_scaled, kmeans, method = "silhouette") +  
 ggtitle("silhouette Method for Optimal k")



#Running K-means algorithm to cluster the dataset. We will choose an initial value of k = 5 identified through Silhouette Method

set.seed(789)  
k5 <- kmeans(Pharma\_scaled, centers = 5, nstart = 25) # k = 5, number of restarts = 25  
  
k5$centers # output the centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 2 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 3 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 4 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 5 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 1.36644699 -0.6912914 -1.320000179  
## 2 0.06308085 1.5180158 -0.006893899  
## 3 -0.27449312 -0.7041516 0.556954446  
## 4 -0.46807818 0.4671788 0.591242521  
## 5 -0.14170336 -0.1168459 -1.416514761

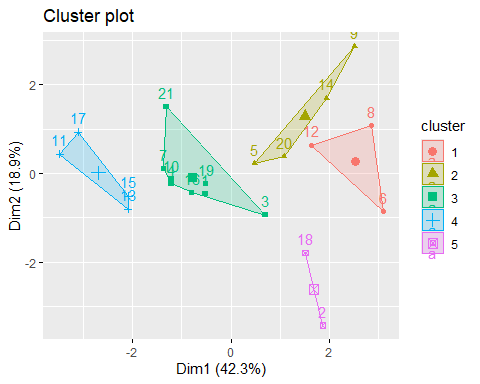
k5$size # Number of cars in each cluster

## [1] 3 4 8 4 2

#Identify the cluster of the 1-21st observation as an example  
k5$cluster[1:21]

## [1] 3 5 3 3 2 1 3 1 2 3 4 1 4 2 4 3 4 5 3 2 3

#Visualize the output  
  
fviz\_cluster(k5, data = Pharma\_scaled)



#Assign clusters to dataset

Pharma$clusters <- k5$cluster

#View the sturcture

str(Pharma)

## 'data.frame': 21 obs. of 15 variables:  
## $ Symbol : chr "ABT" "AGN" "AHM" "AZN" ...  
## $ Name : chr "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PLC" ...  
## $ Market\_Cap : num 68.44 7.58 6.3 67.63 47.16 ...  
## $ Beta : num 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...  
## $ PE\_Ratio : num 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...  
## $ ROE : num 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...  
## $ ROA : num 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...  
## $ Asset\_Turnover : num 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...  
## $ Leverage : num 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...  
## $ Rev\_Growth : num 7.54 9.16 7.05 15 26.81 ...  
## $ Net\_Profit\_Margin : num 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...  
## $ Median\_Recommendation: chr "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...  
## $ Location : chr "US" "CANADA" "UK" "UK" ...  
## $ Exchange : chr "NYSE" "NYSE" "NYSE" "NYSE" ...  
## $ clusters : int 3 5 3 3 2 1 3 1 2 3 ...

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

Answer: The k-means clustering of the pharmaceutical firms resulted in five groups with distinct financial characteristics. Cluster 1 is characterized by a high price-to-earnings (PE) ratio relative to other variables such as Market Cap, Beta, ROE, ROA, Asset Turnover, Leverage, Revenue Growth, and Net Profit Margin, suggesting that these firms may be highly valued relative to earnings despite modest growth. Cluster 2 exhibits high revenue growth compared to other variables, indicating a focus on expansion, but these firms have low efficiency and profitability. Cluster 3 has above-average net profit margins, along with higher ROE and ROA, reflecting profitable and operationally efficient mid-sized companies with moderate growth. Cluster 4 is defined by very high leverage relative to other variables, representing small, high-risk firms with low profitability. Finally, Cluster 5 has high market capitalization, combined with strong profitability, moderate growth, and conservative leverage, representing large, mature, low-risk firms. Collectively, these clusters highlight the diversity of size, profitability, growth, efficiency, and financial risk among pharmaceutical firms.

#Cluster-wise mean of numerical variables

clusters\_summary <- Pharma %>%  
 group\_by(clusters) %>%  
 summarise(across(Market\_Cap:Net\_Profit\_Margin, mean, .names = "mean\_{.col}"))  
print(clusters\_summary)

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

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

# Median Recommendation distribution by cluster  
table(Pharma$cluster, Pharma$Median\_Recommendation)

##   
## Hold Moderate Buy Moderate Sell Strong Buy  
## 1 2 1 0 0  
## 2 0 2 2 0  
## 3 4 1 2 1  
## 4 2 2 0 0  
## 5 1 1 0 0

# Location distribution by cluster  
table(Pharma$cluster, Pharma$Location)

##   
## CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US  
## 1 0 0 1 0 0 0 2  
## 2 0 1 0 1 0 0 2  
## 3 0 0 0 0 1 2 5  
## 4 0 0 0 0 0 1 3  
## 5 1 0 0 0 0 0 1

# Exchange distribution by cluster  
table(Pharma$cluster, Pharma$Exchange)

##   
## AMEX NASDAQ NYSE  
## 1 1 1 1  
## 2 0 0 4  
## 3 0 0 8  
## 4 0 0 4  
## 5 0 0 2

#Insights: Cluster 3 shows the strongest market confidence under Median Recommendation, with most firms receiving “Buy” or “Strong Buy” ratings. These firms are primarily located in the US, UK, and Switzerland, and are listed on the NYSE, reflecting their established and reputable market presence. Clusters 1 and 4 indicate investor caution due to modest growth or high leverage, respectively. Cluster 1 includes firms based in the US and Germany and is distributed across AMEX, NASDAQ, and NYSE, suggesting diversity in firm size and maturity. Cluster 4 firms are primarily from the US and listed on the NYSE, representing small but high-risk firms. Cluster 2 exhibits mixed market recommendations, reflecting the volatility and uncertainty typical of rapidly growing firms. Its members are spread across France, Ireland, and the US, and are listed on the NYSE, aligning with their expansion-oriented nature. Cluster 5 also demonstrates strong market confidence, consisting of large, stable firms located in Canada and the US, all listed on the NYSE, highlighting their mature and low-risk profiles.

#Question D: Provide an appropriate name for each cluster using any or all of the variables in the data set. #Assign descriptive cluster names

cluster\_names <- c(  
 "High PE, Modest Growth Firms", # Cluster 1  
 "Fast-Growing Startups", # Cluster 2  
 "Profitable Mid-Size Firms", # Cluster 3  
 "High-Risk Leveraged Firms", # Cluster 4  
 "Large, Mature Market Leaders" # Cluster 5  
)  
Pharma$Cluster\_Name <- cluster\_names[as.numeric(Pharma$clusters)]  
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  
## clusters Cluster\_Name  
## 1 3 Profitable Mid-Size Firms  
## 2 5 Large, Mature Market Leaders  
## 3 3 Profitable Mid-Size Firms  
## 4 3 Profitable Mid-Size Firms  
## 5 2 Fast-Growing Startups  
## 6 1 High PE, Modest Growth Firms

#Insights: Cluster 1 – High PE, Modest Growth Firms: High price-to-earnings ratios indicate strong investor expectations despite limited growth.

Cluster 2 – Fast-Growing Start ups: Rapid revenue growth with lower efficiency and profitability, typical of early-stage firms.

Cluster 3 – Profitable Mid-Size Firms: Mid-sized firms with strong profitability, returns, and balanced growth.

Cluster 4 – High-Risk Leveraged Firms: Small firms with high debt and low profitability, indicating financial risk.

Cluster 5 – Large, Mature Market Leaders: Well-established, financially strong firms with high market capitalization and low leverage.

#Summary: The five clusters collectively highlight the structural diversity of the pharmaceutical industry. Firms range from high-risk, growth-oriented start ups to financially stable market leaders. This segmentation reflects differences in profitability, leverage, growth strategy, and risk, providing a clear understanding of competitive positioning within the sector.