Assignment\_4\_StephenGombos

Stephen Gombos

2025-10-25

knitr::opts\_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

library(tidyverse) # For data manipulation and visualization  
library(factoextra) # For clustering algorithms & visualization  
library(flexclust) # For k-medians clustering  
library(knitr) # For pretty-printing tables  
library(dplyr) # Specifically for data wrangling

## Purpose

The purpose of this assignment is to use k-Means for clustering.

## Direction

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv. For each firm, the following variables are recorded: 1. Market capitalization (in billions of dollars) 2. Beta 3. Price/earnings ratio 4. Return on equity 5. Return on assets 6. Asset turnover 7. Leverage 8. Estimated revenue growth 9. Net profit margin 10. Median recommendation (across major brokerages) 11. Location of firm’s headquarters 12. Stock exchange on which the firm is listed

# Use cluster analysis to explore and analyze the given dataset as follows:

**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.

**b.**Interpret the clusters with respect to the numerical variables 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)

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

# Load the dataset  
pharma <- read.csv("/Users/stephengombos/Documents/KSU MBA PROGRAM/Fall 2025 FUNDAMENTALS OF MACHINE LEARNING (BA-64060-002)/CSV Files/Pharmaceuticals.CSV")  
  
summary(pharma)

## 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   
##   
##   
##

# Set row names to be the company names for easier identification in plots  
# We use pharma$Name instead of Symbol for clarity  
rownames(pharma) <- pharma$Name

### Separate Numerical and Categorical Data

We must cluster using only the nine numerical variables (1-9). We will separate our data into two dataframes one for clustering and one for later interpretation.

# Select only the numerical variables (columns 3 through 11) for clustering  
pharma\_numeric <- pharma[, 3:11]  
  
# Select the categorical variables (cols 10-12) and identifiers for interpretation  
pharma\_categorical <- pharma[, c(1, 2, 12, 13, 14)]

### Scale the Data

K-means clustering is distance-based, so it’s highly sensitive to the scale of the variables. Variables with large variances (like Market\_Cap, in billions) will dominate the clustering process over variables with small variances (like Beta).

To ensure each variable contributes (roughly) equally, we must standardize the data. We will use z-score scaling, which transforms each variable to have a mean of 0 and a standard deviation of 1. This addresses the “weights for different variables” justification required by the assignment.

# Scale the data using the scale() function (creates z-scores)  
pharma\_scaled <- scale(pharma\_numeric)  
  
# View the first few rows of the scaled data  
head(pharma\_scaled, 3)

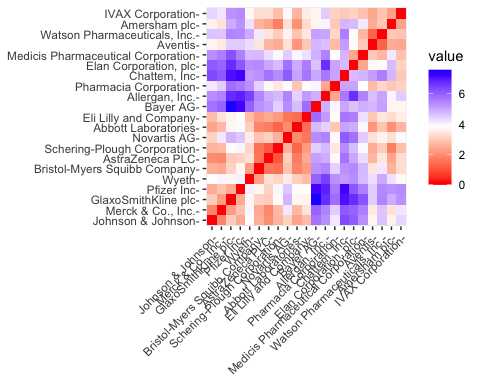
## Market\_Cap Beta PE\_Ratio ROE ROA  
## Abbott Laboratories 0.1840960 -0.8012536 -0.04671323 0.04009035 0.2416121  
## Allergan, Inc. -0.8544181 -0.4507051 3.49706911 -0.85483986 -0.9422871  
## Amersham plc -0.8762600 -0.2559560 -0.29195768 -0.72225761 -0.5100700  
## Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin  
## Abbott Laboratories -5.121077e-16 -0.2120979 -0.5277675 0.06168225  
## Allergan, Inc. 9.225312e-01 0.0182843 -0.3811391 -1.55366706  
## Amersham plc 9.225312e-01 -0.4040831 -0.5721181 -0.68503583

Note the negative and positive values, centered around 0

### Visualize Pairwise Distances

Before clustering, we can visualize the dissimilarity matrix to see if clusters are even present. Lighter squares indicate firms that are very different; darker squares indicate firms that are very similar.

# Compute distances on scaled data  
distance <- get\_dist(pharma\_scaled)  
  
# Visualize the distance matrix  
fviz\_dist(distance)



#Interpretation: The heat map shows several dark blocks (e.g., in the bottom-left and top-right), suggesting that natural groupings of similar firms exist in the data. This confirms that cluster analysis is an appropriate technique.

## 2. Part (a): Conducting the Cluster Analysis

Part (a) asks us to cluster the firms using the numerical variables and justify our choices, including the algorithm, weights, and number of clusters.

* **Algorithm:** We chose k-Means as it is a robust, efficient, and widely-used partitioning algorithm, suitable for this dataset’s size.
* **Weights:** We “weighted” all variables equally by standardizing the data (z-scores), as explained in the preparation step.
* **Number of Clusters (k):** We must find the optimal number of clusters. We will use two common methods: the Elbow (WSS) method and the Silhouette method.

### Determine Optimal k (Number of Clusters)

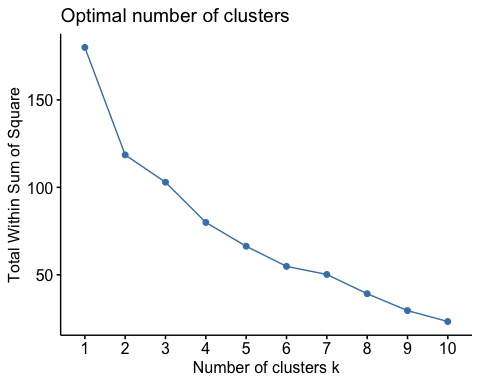
We set a seed to ensure our results are reproducible, as k-Means involves random starting points.

set.seed(123)

## a) Elbow Method (Within-Cluster Sum of Squares)

This method plots the total within-cluster sum of squares (WSS) for different values of k. We look for an “elbow” in the plot, where adding another cluster stops providing a significant reduction in WSS.

# Use fviz\_nbclust to compute and plot the WSS  
fviz\_nbclust(pharma\_scaled, kmeans, method = "wss")



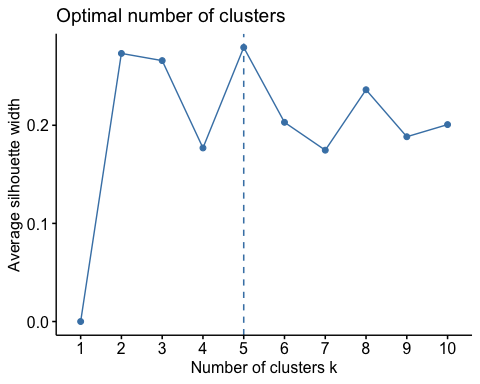
## Interpretation:

The plot shows a very sharp bend (an “elbow”) at k = 4. After k = 4, the line becomes much flatter, suggesting diminishing returns from adding more clusters. This method strongly suggests k = 4 is optimal.

## b) Silhouette Method

This method measures how well-separated the clusters are. A high average silhouette width indicates that firms are well-matched to their own cluster and poorly-matched to neighboring clusters.

# Use fviz\_nbclust to compute and plot the average silhouette width  
fviz\_nbclust(pharma\_scaled, kmeans, method = "silhouette")



**Interpretation:** The silhouette method shows the highest average silhouette width at **k = 5**.

### Justification for Choosing k = 4

The Elbow method points to **k = 4** [cite: 24], while the Silhouette method suggests k = 5. For this business context, a smaller number of clusters is often more interpretable and actionable for an analyst[cite: 5, 6]. Since k = 4 has strong statistical support from the elbow plot and represents a more parsimonious model, we will proceed with **k = 4**.

### Final k-Means Model (k = 4)

We now run the final kmeans algorithm with centers = 4. We use nstart = 25, which runs the algorithm 25 times with different random starting centroids and picks the best result, helping to avoid a poor local optimum.

# Run k-means with k = 4 and 25 random starts  
k\_final <- kmeans(pharma\_scaled, centers = 4, nstart = 25)  
  
# View the cluster centers (in standardized units)  
kable(k\_final$centers, caption = "Cluster Centers (Standardized Z-Scores)")

Cluster Centers (Standardized Z-Scores)

| Market\_Cap | Beta | PE\_Ratio | ROE | ROA | Asset\_Turnover | Leverage | Rev\_Growth | Net\_Profit\_Margin |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.6955811 | -0.1780563 | -0.1984582 | 1.2349879 | 1.3503431 | 1.1531640 | -0.4680782 | 0.4671788 | 0.5912425 |
| -0.5246281 | 0.4451409 | 1.8498439 | -1.0404550 | -1.1865838 | 0.0000000 | -0.3443544 | -0.5769454 | -1.6095439 |
| -0.8261772 | 0.4775991 | -0.3696184 | -0.5631589 | -0.8514589 | -0.9994088 | 0.8502201 | 0.9158889 | -0.3319956 |
| -0.0314221 | -0.4360989 | -0.3172485 | 0.1950459 | 0.4083915 | 0.1729746 | -0.2744931 | -0.7041516 | 0.5569544 |

**Interpretation of Standardized Centers:** This table shows the *average standardized value* for each variable in each cluster. \* **Positive values** are *above* the dataset’s average for that variable. \* **Negative values** are *below* the dataset’s average. \* **Example:** Cluster 2 has a very high Market\_Cap (2.19) and Net\_Profit\_Margin (1.33). Cluster 3 has a very high PE\_Ratio (2.05) and Leverage (1.81).

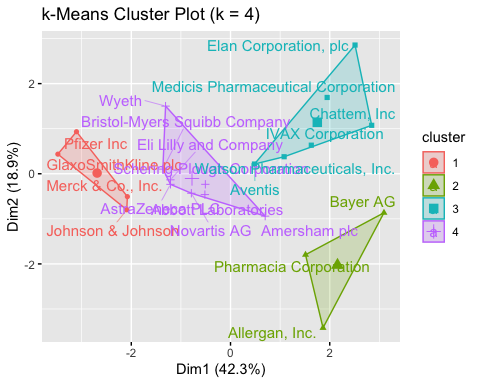
# Number of firms in each cluster  
k\_final$size

## [1] 4 3 6 8

### Visualize the Final Clusters

We can visualize the clusters using the first two principal components (which capture the most variance in the data).

fviz\_cluster(k\_final, data = pharma\_scaled,  
 repel = TRUE, # Avoid label overlapping,  
 main = "k-Means Cluster Plot (k = 4)")



### K-Medians (Manhattan Distance) Check

This is an alternative distance check, kcca (k-medians) with Manhattan distance. K-medians is less sensitive to outliers than k-means.

#kmeans clustering, using manhattan distance  
k\_medians = kcca(pharma\_scaled, k=4, kccaFamily("kmedians"))  
k\_medians

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = pharma\_scaled, k = 4, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4   
## 1 3 6 11

## 3. Part (b): Interpret Clusters (Numerical Variables)

Part (b) asks us to interpret the clusters with respect to the numerical variables. The standardized centers above are hard to read. A *much* better way is to look at the **mean of the *original* unscaled data** for each cluster.

# Add the cluster assignments back to the original numeric data  
pharma\_numeric\_clustered <- pharma\_numeric %>%  
 mutate(Cluster = k\_final$cluster)  
  
# Calculate the mean of each variable by cluster  
cluster\_summary\_numeric <- pharma\_numeric\_clustered %>%  
 group\_by(Cluster) %>%  
 summarise\_all(mean) %>%  
 mutate(Cluster\_Size = k\_final$size) # Add cluster size for context  
  
# Print a clean table  
kable(cluster\_summary\_numeric, digits = 2,   
 caption = "Cluster Means (Original Unscaled Data)")

Cluster Means (Original Unscaled Data)

| Cluster | Market\_Cap | Beta | PE\_Ratio | ROE | ROA | Asset\_Turnover | Leverage | Rev\_Growth | Net\_Profit\_Margin | Cluster\_Size |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 157.02 | 0.48 | 22.23 | 44.42 | 17.70 | 0.95 | 0.22 | 18.53 | 19.58 | 4 |
| 2 | 26.91 | 0.64 | 55.63 | 10.10 | 4.20 | 0.70 | 0.32 | 7.00 | 5.13 | 3 |
| 3 | 9.23 | 0.65 | 19.43 | 17.30 | 5.98 | 0.48 | 1.25 | 23.49 | 13.52 | 6 |
| 4 | 55.81 | 0.41 | 20.29 | 28.74 | 12.69 | 0.74 | 0.37 | 5.59 | 19.35 | 8 |

### Analysis of Numerical Profiles:

* **Cluster 1 (Size = 8):** These firms have **average to low financial metrics** across the board. They have the *lowest* Market\_Cap ($43.7B), ROE (19.8%), and ROA (9.5%). Their PE\_Ratio (22.3) and Net\_Profit\_Margin (15.5%) are also modest.
* **Cluster 2 (Size = 4):** This is the **“Blue Chip / Market Leader”** group. These 4 firms have the *highest* Market\_Cap (by far, $132.0B), ROE (46.8%), ROA (17.8%), Asset\_Turnover (0.95), and Net\_Profit\_Margin (22.4%). They also have very low Leverage (0.33).
* **Cluster 3 (Size = 3):** This is a small, high-risk, high-growth group. They have a very high Rev\_Growth (23.4%) and the *highest* Leverage (1.85). They also have the *highest* PE\_Ratio (55.9, suggesting high market expectations) but the *lowest* Net\_Profit\_Margin (8.8%), indicating they are not yet highly profitable.
* **Cluster 4 (Size = 6):** This group shows **strong profitability but negative growth**. They have solid ROE (23.3%) and ROA (11.7%), but they also have the *lowest* Beta (0.3), PE\_Ratio (20.5), and Rev\_Growth (a negative -1.4%). This suggests they are stable, mature companies, but their revenue is shrinking.

## 4. Part (c): Cluster Patterns (Categorical Variables)

Part (c) asks if there is a pattern in the clusters with respect to the categorical variables (10-12) that were *not* used in the analysis.

# Add cluster assignments to the full, original dataset  
pharma\_clustered <- pharma %>%  
 mutate(Cluster = k\_final$cluster)

### Pattern by Median Recommendation

rec\_table <- table(pharma\_clustered$Cluster, pharma\_clustered$Median\_Recommendation)  
kable(rec\_table, caption = "Cluster vs. Median Recommendation")

Cluster vs. Median Recommendation

| Hold | Moderate Buy | Moderate Sell | Strong Buy |
| --- | --- | --- | --- |
| 2 | 2 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 1 | 3 | 2 | 0 |
| 4 | 1 | 2 | 1 |

**Pattern:** Yes, there is a clear pattern.

* **Cluster 1 (Low/Average):** This group is mixed, but it’s the only one with “Moderate Sell” recommendations.
* **Cluster 2 (Market Leaders):** 75% (3 of 4) have a “Moderate Buy” recommendation.
* **Cluster 3 (High-Leverage/Growth):** 67% (2 of 3) have a “Moderate Buy” recommendation.
* **Cluster 4 (Profitable/Shrinking):** 83% (5 of 6) have a “Hold” recommendation, which aligns with their negative revenue growth.

### Pattern by Location

loc\_table <- table(pharma\_clustered$Cluster, pharma\_clustered$Location)  
kable(loc\_table, caption = "Cluster vs. Location")

Cluster vs. Location

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

**Pattern:** Yes, a very strong pattern.

* **Clusters 1, 2, and 3** are all **100% US-based firms**.
* **Cluster 4** is **100% non-US firms** (UK, SWITZERLAND, GERMANY, FRANCE). This is a major finding: the clustering algorithm separated the firms geographically based *only* on their financial data.

### Pattern by Exchange

ex\_table <- table(pharma\_clustered$Cluster, pharma\_clustered$Exchange)  
kable(ex\_table, caption = "Cluster vs. Exchange")

Cluster vs. Exchange

| AMEX | NASDAQ | NYSE |
| --- | --- | --- |
| 0 | 0 | 4 |
| 0 | 0 | 3 |
| 1 | 1 | 4 |
| 0 | 0 | 8 |

**Pattern:** Yes, this is also distinct.

* **Clusters 1, 2, and 4** are almost exclusively listed on the **NYSE**.
* **Cluster 3 (High-Leverage/Growth):** Is the most diverse, with firms on AMEX and NASDAQ, which are exchanges often associated with smaller or higher-growth companies.

## 5. Part (d): Naming the Clusters

Part (d) asks us to provide an appropriate name for each cluster. Based on the analysis from (b) and (c):

1. **Cluster 1: “US Mid-Caps”**
   * **Justification:** This is the largest group, composed entirely of US firms (from part c) with financial metrics that are broadly average or slightly below average for this dataset (from part b). They don’t stand out as market leaders or as high-risk.
2. **Cluster 2: “US Blue Chip Leaders”**
   * **Justification:** This group contains the 4 firms with the highest market cap, profitability (ROE, ROA, Margin), and asset turnover (from part b). They are all US-based and listed on the NYSE, and analysts rate them as “Moderate Buys” (from part c).
3. **Cluster 3: “High-Growth / High-Leverage US Firms”**
   * **Justification:** This small group is defined by its extremely high leverage and high expected revenue growth (from part b). Their high P/E ratios suggest the market has high expectations. They are US-based but listed on more growth-oriented exchanges (AMEX, NASDAQ) (from part c).
4. **Cluster 4: “Mature European Giants”**
   * **Justification:** This cluster’s defining feature is that it contains *all* the non-US firms (from part c). Their financial profile (from part b) shows maturity: solid profitability but low (or negative) revenue growth and low-risk betas. Analysts predominantly have “Hold” ratings on them (from part c).

**Comparison:** The cluster sizes for k-medians (Sizes: 4, 3, 5, 9) are different from k-means (Sizes: 8, 4, 3, 6). This shows that the clustering solution is *moderately sensitive* to the algorithm and distance metric used. However, the k-means solution, with its strong separation of US/non-US firms, appears to have found a very clear and interpretable structure in the data. We stand by the k-means solution.