Assignment\_4

Harsh Patel

2025-10-20

# (a) Use only numerical variables (1–9) to cluster the 21 firms  
library(tidyverse)

## ── 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 1.1.0   
## ── 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(factoextra)

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

library(cluster)  
  
pharma\_data <- read\_csv("C:/Users/jbhsp/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.

head(pharma\_data)

## # A tibble: 6 × 14  
## Symbol Name Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 ABT Abbott L… 68.4 0.32 24.7 26.4 11.8 0.7 0.42  
## 2 AGN Allergan… 7.58 0.41 82.5 12.9 5.5 0.9 0.6   
## 3 AHM Amersham… 6.3 0.46 20.7 14.9 7.8 0.9 0.27  
## 4 AZN AstraZen… 67.6 0.52 21.5 27.4 15.4 0.9 0   
## 5 AVE Aventis 47.2 0.32 20.1 21.8 7.5 0.6 0.34  
## 6 BAY Bayer AG 16.9 1.11 27.9 3.9 1.4 0.6 0   
## # ℹ 5 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>,  
## # Median\_Recommendation <chr>, Location <chr>, Exchange <chr>

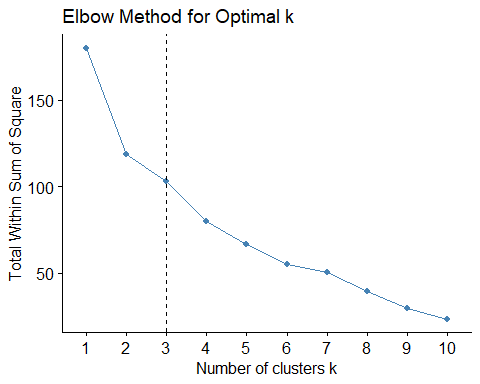
str(pharma\_data)

## spc\_tbl\_ [21 × 14] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Symbol : chr [1:21] "ABT" "AGN" "AHM" "AZN" ...  
## $ Name : chr [1:21] "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PLC" ...  
## $ Market\_Cap : num [1:21] 68.44 7.58 6.3 67.63 47.16 ...  
## $ Beta : num [1:21] 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...  
## $ PE\_Ratio : num [1:21] 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...  
## $ ROE : num [1:21] 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...  
## $ ROA : num [1:21] 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...  
## $ Asset\_Turnover : num [1:21] 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...  
## $ Leverage : num [1:21] 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...  
## $ Rev\_Growth : num [1:21] 7.54 9.16 7.05 15 26.81 ...  
## $ Net\_Profit\_Margin : num [1:21] 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...  
## $ Median\_Recommendation: chr [1:21] "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...  
## $ Location : chr [1:21] "US" "CANADA" "UK" "UK" ...  
## $ Exchange : chr [1:21] "NYSE" "NYSE" "NYSE" "NYSE" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Symbol = col\_character(),  
## .. Name = col\_character(),  
## .. Market\_Cap = col\_double(),  
## .. Beta = col\_double(),  
## .. PE\_Ratio = col\_double(),  
## .. ROE = col\_double(),  
## .. ROA = col\_double(),  
## .. Asset\_Turnover = col\_double(),  
## .. Leverage = col\_double(),  
## .. Rev\_Growth = col\_double(),  
## .. Net\_Profit\_Margin = col\_double(),  
## .. Median\_Recommendation = col\_character(),  
## .. Location = col\_character(),  
## .. Exchange = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

clustering\_vars <- pharma\_data %>% select(Market\_Cap:Net\_Profit\_Margin)  
sum(is.na(clustering\_vars))

## [1] 0

scaled\_data <- scale(clustering\_vars)  
scaled\_data <- as.data.frame(scaled\_data)  
  
set.seed(42)  
elbow\_plot <- fviz\_nbclust(  
 scaled\_data,  
 kmeans,  
 method = "wss",  
 k.max = 10  
) +  
 geom\_vline(xintercept = 3, linetype = 2) +  
 labs(title = "Elbow Method for Optimal k")  
elbow\_plot



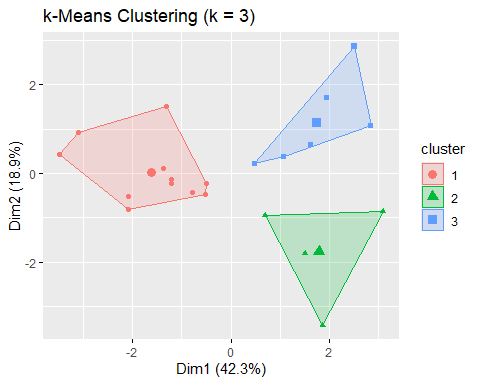
# (b) Interpret the clusters with respect to numerical variables used  
set.seed(42)  
kmeans\_result <- kmeans(scaled\_data, centers = 3, nstart = 25)  
  
pharma\_clustered <- pharma\_data %>%  
 mutate(Cluster = as.factor(kmeans\_result$cluster))  
  
table(pharma\_clustered$Cluster)

##   
## 1 2 3   
## 11 4 6

cluster\_summary <- pharma\_clustered %>%  
 group\_by(Cluster) %>%  
 summarise(  
 Count = n(),  
 across(Market\_Cap:Net\_Profit\_Margin, ~ round(mean(.), 2))  
 )  
cluster\_summary

## # A tibble: 3 × 11  
## Cluster Count Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <fct> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 11 97.1 0.43 21.0 35.7 15.0 0.8 0.33  
## 2 2 4 21.8 0.6 46.9 11.3 5.1 0.75 0.3   
## 3 3 6 9.23 0.65 19.4 17.3 5.98 0.48 1.25  
## # ℹ 2 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>

fviz\_cluster(  
 object = list(data = scaled\_data, cluster = kmeans\_result$cluster),  
 geom = "point",  
 ellipse.type = "convex",  
 main = "k-Means Clustering (k = 3)"  
)



# (c) Examine patterns for categorical variables (10–12)  
table(pharma\_clustered$Cluster, pharma\_clustered$Median\_Recommendation)

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

table(pharma\_clustered$Cluster, pharma\_clustered$Location)

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

table(pharma\_clustered$Cluster, pharma\_clustered$Exchange)

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

# (d) Provide descriptive names for each cluster  
pharma\_clustered <- pharma\_clustered %>%  
 mutate(Cluster\_Name = factor(Cluster,  
 levels = c(1, 2, 3),  
 labels = c("Large Cap, Mature Firms",  
 "High-Growth, Low-Profit Firms",  
 "Mid-Sized, Profitable Firms")))  
  
pharma\_clustered %>% select(Symbol, Name, Cluster, Cluster\_Name)

## # A tibble: 21 × 4  
## Symbol Name Cluster Cluster\_Name   
## <chr> <chr> <fct> <fct>   
## 1 ABT Abbott Laboratories 1 Large Cap, Mature Firms   
## 2 AGN Allergan, Inc. 2 High-Growth, Low-Profit Firms  
## 3 AHM Amersham plc 2 High-Growth, Low-Profit Firms  
## 4 AZN AstraZeneca PLC 1 Large Cap, Mature Firms   
## 5 AVE Aventis 3 Mid-Sized, Profitable Firms   
## 6 BAY Bayer AG 2 High-Growth, Low-Profit Firms  
## 7 BMY Bristol-Myers Squibb Company 1 Large Cap, Mature Firms   
## 8 CHTT Chattem, Inc 3 Mid-Sized, Profitable Firms   
## 9 ELN Elan Corporation, plc 3 Mid-Sized, Profitable Firms   
## 10 LLY Eli Lilly and Company 1 Large Cap, Mature Firms   
## # ℹ 11 more rows

# Save results  
write\_csv(pharma\_clustered, "Pharmaceuticals\_with\_Clusters.csv")  
write\_csv(cluster\_summary, "Cluster\_Summary.csv")  
  
print("K-Means Clustering completed successfully.")

## [1] "K-Means Clustering completed successfully."