Assignment\_4

2025-10-24

## Loading packages

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(class)  
library(gmodels)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.1 ✔ stringr 1.5.2  
## ✔ ggplot2 4.0.0 ✔ 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(flexclust)

## Importing the dataset

pharma <- read\_csv("./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.

## Task 1

#separating numeric data  
set.seed(135)  
pharma.num <- pharma[, 3:11]  
summary(pharma.num)

## Market\_Cap Beta PE\_Ratio ROE   
## Min. : 0.41 Min. :0.1800 Min. : 3.60 Min. : 3.9   
## 1st Qu.: 6.30 1st Qu.:0.3500 1st Qu.:18.90 1st Qu.:14.9   
## Median : 48.19 Median :0.4600 Median :21.50 Median :22.6   
## Mean : 57.65 Mean :0.5257 Mean :25.46 Mean :25.8   
## 3rd Qu.: 73.84 3rd Qu.:0.6500 3rd Qu.:27.90 3rd Qu.:31.0   
## Max. :199.47 Max. :1.1100 Max. :82.50 Max. :62.9   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. : 1.40 Min. :0.3 Min. :0.0000 Min. :-3.17   
## 1st Qu.: 5.70 1st Qu.:0.6 1st Qu.:0.1600 1st Qu.: 6.38   
## Median :11.20 Median :0.6 Median :0.3400 Median : 9.37   
## Mean :10.51 Mean :0.7 Mean :0.5857 Mean :13.37   
## 3rd Qu.:15.00 3rd Qu.:0.9 3rd Qu.:0.6000 3rd Qu.:21.87   
## Max. :20.30 Max. :1.1 Max. :3.5100 Max. :34.21   
## Net\_Profit\_Margin  
## Min. : 2.6   
## 1st Qu.:11.2   
## Median :16.1   
## Mean :15.7   
## 3rd Qu.:21.1   
## Max. :25.5

#This data is required normalization because the magnitude of Market\_Cap is too high compare to other variables, which will influence the whole result.  
  
#normalizing the data  
pharma.num.scaled <- scale(pharma.num)  
summary(pharma.num.scaled) #now all the data normalized and almost in the same scale

## Market\_Cap Beta PE\_Ratio ROE   
## Min. :-0.9768 Min. :-1.3466 Min. :-1.3404 Min. :-1.4515   
## 1st Qu.:-0.8763 1st Qu.:-0.6844 1st Qu.:-0.4023 1st Qu.:-0.7223   
## Median :-0.1614 Median :-0.2560 Median :-0.2429 Median :-0.2118   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.2762 3rd Qu.: 0.4841 3rd Qu.: 0.1495 3rd Qu.: 0.3450   
## Max. : 2.4200 Max. : 2.2758 Max. : 3.4971 Max. : 2.4597   
## ROA Asset\_Turnover Leverage Rev\_Growth   
## Min. :-1.7128 Min. :-1.8451 Min. :-0.74966 Min. :-1.4971   
## 1st Qu.:-0.9047 1st Qu.:-0.4613 1st Qu.:-0.54487 1st Qu.:-0.6328   
## Median : 0.1289 Median :-0.4613 Median :-0.31449 Median :-0.3621   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.8430 3rd Qu.: 0.9225 3rd Qu.: 0.01828 3rd Qu.: 0.7693   
## Max. : 1.8389 Max. : 1.8451 Max. : 3.74280 Max. : 1.8862   
## Net\_Profit\_Margin   
## Min. :-1.99560   
## 1st Qu.:-0.68504   
## Median : 0.06168   
## Mean : 0.00000   
## 3rd Qu.: 0.82364   
## Max. : 1.49416

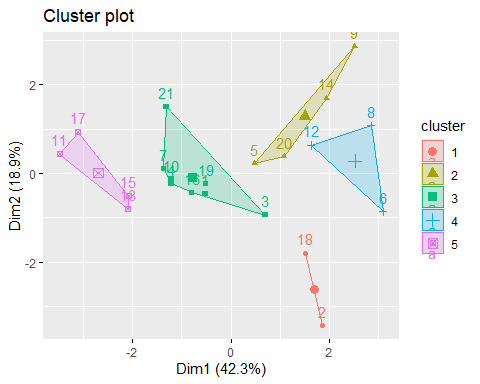
#I am going to use the "Elbow method" and "Average Silhouette method" to find the best value for k  
fviz\_nbclust(pharma.num.scaled, kmeans, method="wss") #Elbow method



#The elbow method suggests that the optimal number of clusters lies around 4, 5, or 6, because from 4 it started decreasing at a much smaller rate. In other words, k=4/5/6 provides the best value between bias and overfitting. However, the exact number of clusters remains somewhat ambiguous based on this method alone.  
  
fviz\_nbclust(pharma.num.scaled, kmeans, method="silhouette") #Silhouette method



#On the other hand, the silhouette method provides a clearer indication, identifying k = 5 as the optimal cluster number. Compared to the elbow method, the silhouette approach is generally more reliable, objective, and easier to interpret. It not only evaluates how cohesive (tight) the clusters are internally but also how well-separated they are from each other.  
#Therefore, I choose the k value 5  
  
#visualizing the clusters  
pharma.k <- kmeans(pharma.num.scaled,centers=5,nstart=25)  
fviz\_cluster(pharma.k, data=pharma.num.scaled)



#From the graph, it is clear that these five clusters are well defined and separated

## Task 2

#adding the cluster to the original dataset to summarize and find relation among cluster and variables  
pharma$cluster <- pharma.k$cluster  
#creating tables with mean values of all variables for each cluster  
pharma.num.summary <- aggregate(pharma.num, by = list(Cluster = pharma$cluster), mean)  
pharma.num.summary

## Cluster Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 1 31.910000 0.40500 69.5000 13.20000 5.600000 0.7500  
## 2 2 13.100000 0.59750 17.6750 14.57500 6.200000 0.4250  
## 3 3 55.810000 0.41375 20.2875 28.73750 12.687500 0.7375  
## 4 4 6.636667 0.87000 24.6000 16.46667 4.166667 0.6000  
## 5 5 157.017500 0.48000 22.2250 44.42500 17.700000 0.9500  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 0.475000 12.080000 6.400000  
## 2 0.635000 30.142500 15.650000  
## 3 0.371250 5.591250 19.350000  
## 4 1.653333 5.733333 7.033333  
## 5 0.220000 18.532500 19.575000

#Cluster 1- medium market capital with moderate risk, revenue and low profit margin  
#Cluster 2- low market capital with high revenue, high debt and good profit margin  
#Cluster 3- large market capital with low risk, revenue but high profit margin  
#Cluster 4- very low market capital with high risk, debt and low revenue and profit margin  
#Cluster 5- largest market capital with low risk, debt and high revenue and highest profitability

## Task 3

#creating table to understand the distribution of these categorical variables within clusters  
pharma %>% group\_by(cluster,Median\_Recommendation) %>% summarise(RecomC=n())

## `summarise()` has grouped output by 'cluster'. You can override using the  
## `.groups` argument.

## # A tibble: 12 × 3  
## # Groups: cluster [5]  
## cluster Median\_Recommendation RecomC  
## <int> <chr> <int>  
## 1 1 Hold 1  
## 2 1 Moderate Buy 1  
## 3 2 Moderate Buy 2  
## 4 2 Moderate Sell 2  
## 5 3 Hold 4  
## 6 3 Moderate Buy 1  
## 7 3 Moderate Sell 2  
## 8 3 Strong Buy 1  
## 9 4 Hold 2  
## 10 4 Moderate Buy 1  
## 11 5 Hold 2  
## 12 5 Moderate Buy 2

pharma %>% group\_by(cluster,Location) %>% summarise(LocationC=n())

## `summarise()` has grouped output by 'cluster'. You can override using the  
## `.groups` argument.

## # A tibble: 12 × 3  
## # Groups: cluster [5]  
## cluster Location LocationC  
## <int> <chr> <int>  
## 1 1 CANADA 1  
## 2 1 US 1  
## 3 2 FRANCE 1  
## 4 2 IRELAND 1  
## 5 2 US 2  
## 6 3 SWITZERLAND 1  
## 7 3 UK 2  
## 8 3 US 5  
## 9 4 GERMANY 1  
## 10 4 US 2  
## 11 5 UK 1  
## 12 5 US 3

pharma %>% group\_by(cluster,Exchange) %>% summarise(ExchangeC=n())

## `summarise()` has grouped output by 'cluster'. You can override using the  
## `.groups` argument.

## # A tibble: 7 × 3  
## # Groups: cluster [5]  
## cluster Exchange ExchangeC  
## <int> <chr> <int>  
## 1 1 NYSE 2  
## 2 2 NYSE 4  
## 3 3 NYSE 8  
## 4 4 AMEX 1  
## 5 4 NASDAQ 1  
## 6 4 NYSE 1  
## 7 5 NYSE 4

#Cluster 1- Canada, US based and mostly NYSE  
#Cluster 2- moderate buy-sell and mostly NYSE  
#Cluster 3- Mix recommendation type, mostly US, NYSE  
#Cluster 4- Mix of exchange type  
#Cluster 5- Mostly US based and NYSE  
  
#From my perspective I didn't find any specific pattern among those variables and clusters. But I have some general observation that most companies are Us-based and exchange type is NYSE across the clusters. Most recommendation variations are in Cluster 3 and most exchange variations are in Cluster 4

## Task 4:

#Naming the clusters corresponding the results of the variables representing growth, profit and risk, found in task-2  
  
#Cluster 1- mediocre stable companies [moderate growth, moderate risk and revenue]  
#Cluster 2- fast growing emerging companies [companies with high risk and high growth with decent profit]  
#Cluster 3- profitable companies [low risk and low growth but highly profitable]  
#Cluster 4- risky companies [low growth and low profit with high risk]  
#Cluster 5- market dominating established companies [large growth and high profit with minimum risk]