Assignment\_4

2022-10-22

#Inputting and Setting up Data/Packages **A**

mydata <- read.csv("C:\\Users\\bucol\\Downloads\\Pharmaceuticals.csv")  
library(tidyverse) # data manipulation

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(factoextra) # clustering algorithms & visualization

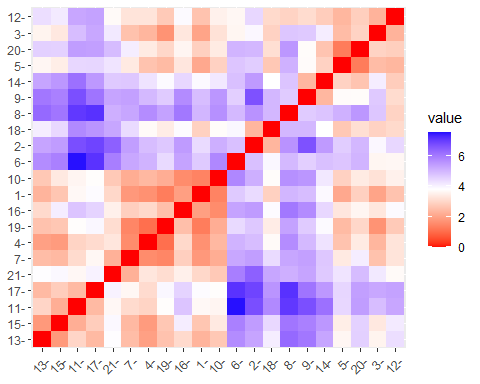
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(ISLR)  
set.seed(49)  
  
PharmData <- select(mydata, 3:11)  
summary(PharmData)

## 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

Now we need to normalize our data since Euclidean distance is very affected by scale. Then we will get our distances.

NormPharm <- scale(PharmData)  
distance <- get\_dist(NormPharm)  
fviz\_dist(distance)



fviz\_nbclust(NormPharm, kmeans, method = "wss")



#It's a bit challenging to read this graph to determine the elbow point, so we'll also check witht the silhouette method  
fviz\_nbclust(NormPharm, kmeans, method = "silhouette")



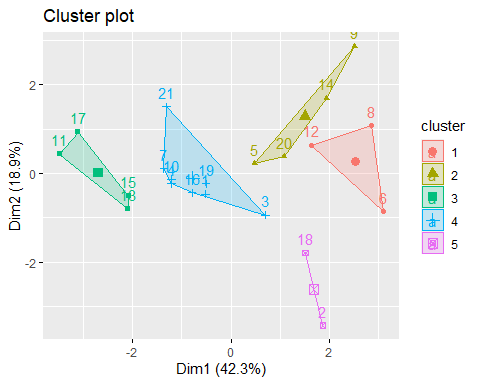
#From here, we know to choose 5 clusters.

#Now we cluster!

set.seed(49)  
clusterk5 <- kmeans(NormPharm, center = 5, nstart = 25)  
  
#Let's Visualize  
clusterk5$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 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 4 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 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 -0.46807818 0.4671788 0.591242521  
## 4 -0.27449312 -0.7041516 0.556954446  
## 5 -0.14170336 -0.1168459 -1.416514761

fviz\_cluster(clusterk5, data= NormPharm)



clusterk5$size

## [1] 3 4 4 8 2

**b** Looking at our cluster plot, we can see:

Cluster 1: rows 6, 8, 12 Cluster 2: rows 5, 9, 14, 20 Cluster 3: rows 11, 13, 15, 17 Cluster 4: rows 1, 3, 4, 7, 10, 16, 19, 21 Cluster 5: rows 2, 18

**c** Comparing this the variables not used in the cluster analysis, there aren’t very discernable patterns. -The countries seem pretty randomly distributed throughout the clusters.  
-The only two non NYSE companies are in cluster 1. -The reccommendations also seem relatively random.

**d** Naming Clusters

Cluster 1: Lowest Market Capitalization, Return on Equity, and Asset Turnover, but highest Beta and Return on Assets Cluster 2: Lowest Price/earnings ratio, but highest Revenue Growth Cluster 3: Lowest Leverage, but highest Return on equity, return on assets, asset turnover, net profit margin and market capitalization Cluster 4: Lowest revenue growth Cluster 5: Lowest beta, but highest Price/earnings ratio