FML\_ASSIGNMENT\_4

Niharika Matsa

2023-11-12

**Loading the required libraries**

library(flexclust)

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(cluster)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2

## ── 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(FactoMineR)  
library(ggcorrplot)

1. 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.

data = read.csv("Pharmaceuticals.csv")  
head(data)

## 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

Pharmaceuticals = data[3:11]  
head(Pharmaceuticals)

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage Rev\_Growth  
## 1 68.44 0.32 24.7 26.4 11.8 0.7 0.42 7.54  
## 2 7.58 0.41 82.5 12.9 5.5 0.9 0.60 9.16  
## 3 6.30 0.46 20.7 14.9 7.8 0.9 0.27 7.05  
## 4 67.63 0.52 21.5 27.4 15.4 0.9 0.00 15.00  
## 5 47.16 0.32 20.1 21.8 7.5 0.6 0.34 26.81  
## 6 16.90 1.11 27.9 3.9 1.4 0.6 0.00 -3.17  
## Net\_Profit\_Margin  
## 1 16.1  
## 2 5.5  
## 3 11.2  
## 4 18.0  
## 5 12.9  
## 6 2.6

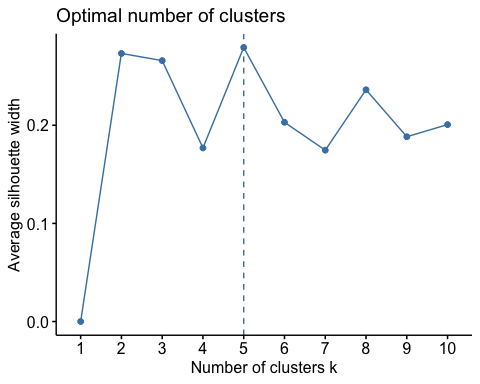
summary(Pharmaceuticals)

## 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

**Normalizing the data**

Pharma = scale(Pharmaceuticals)  
row.names(Pharma) = data[,1]  
distance = get\_dist(Pharma)  
correlation = cor(Pharma)

fviz\_nbclust(Pharma, kmeans, method = "silhouette")



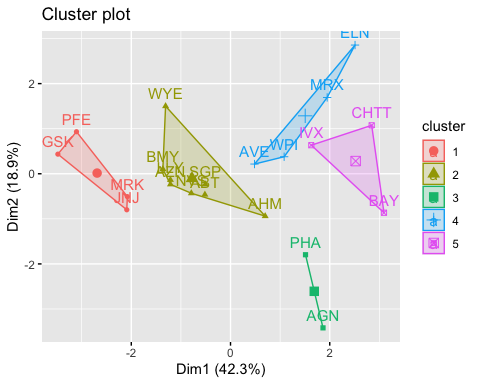
set.seed(69)  
k5 = kmeans(Pharma, centers = 5, nstart = 30)  
k5$size

## [1] 4 8 2 4 3

k5$centers

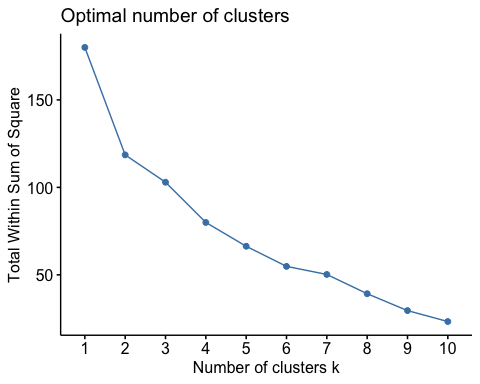
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 2 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 3 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 4 -0.76022489 0.2796041 -0.47742380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

fviz\_cluster(k5, data = Pharma)



**elbow**

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



**Manhattan**

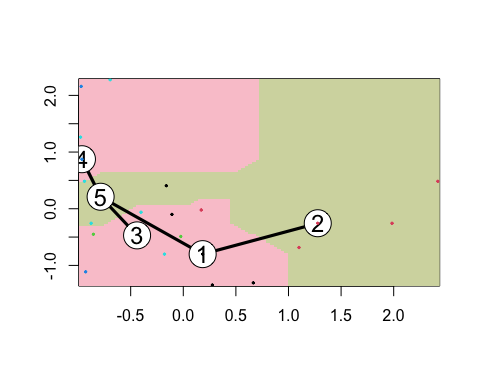
set.seed(50)  
k51 = kcca(Pharma, k = 5, kccaFamily("kmedians"))  
k51

## kcca object of family 'kmedians'   
##   
## call:  
## kcca(x = Pharma, k = 5, family = kccaFamily("kmedians"))  
##   
## cluster sizes:  
##   
## 1 2 3 4 5   
## 5 5 2 3 6

clusters\_index = predict(k51)  
dist(k51@centers)

## 1 2 3 4  
## 2 2.558034   
## 3 4.451230 4.795056   
## 4 4.222539 4.954336 4.589219   
## 5 2.645989 3.581581 3.351236 2.857647

image(k51)  
points(Pharma, col = clusters\_index, pch = 19, cex = 0.3)

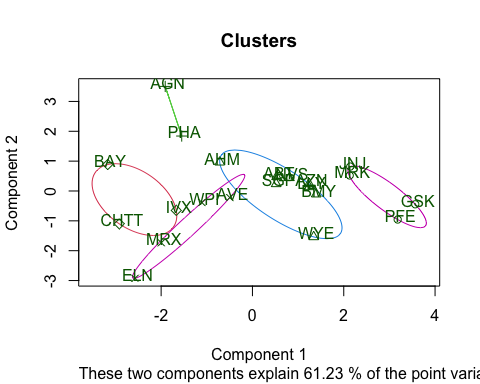
 2. Interpret the clusters with respect to the numerical variables used in forming the clusters. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

Pharmaceuticals %>% mutate(Cluster = k5$cluster) %>% group\_by(Cluster) %>% summarise\_all("mean")

## # A tibble: 5 × 10  
## Cluster Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 157. 0.48 22.2 44.4 17.7 0.95 0.22   
## 2 2 55.8 0.414 20.3 28.7 12.7 0.738 0.371  
## 3 3 31.9 0.405 69.5 13.2 5.6 0.75 0.475  
## 4 4 13.1 0.598 17.7 14.6 6.2 0.425 0.635  
## 5 5 6.64 0.87 24.6 16.5 4.17 0.6 1.65   
## # ℹ 2 more variables: Rev\_Growth <dbl>, Net\_Profit\_Margin <dbl>

**Interpretation:**

clusplot(Pharma,k5$cluster, main = "Clusters", color = TRUE, labels = 3, lines = 0)



**Below is the Cluster naming based on the companies:**

Cluster 1: ELN, MRX, WPI and AVE

Cluster 2: AGN and PHA

Cluster 3: AHM,WYE,BMY,AZN, LLY, ABT, NVS and SGP

Cluster 4: BAY, CHTT and IVX

Cluster 5: JNJ, MRK, PFE and GSK

**Interpretation**

**Cluster 1 - Best**

Cluster stands out with the best Net Profit Margin, the lowest PE ratio, and rapid sales growth. This cluster is considered a strong candidate for purchase or holding as a reserve.

**Cluster 2 - Substantial Risk**

Cluster 2 is characterized by a notably high PE ratio, signaling potential overvaluation. Investors should approach this cluster with caution due to the elevated valuation.

**Cluster 3 - Pursue**

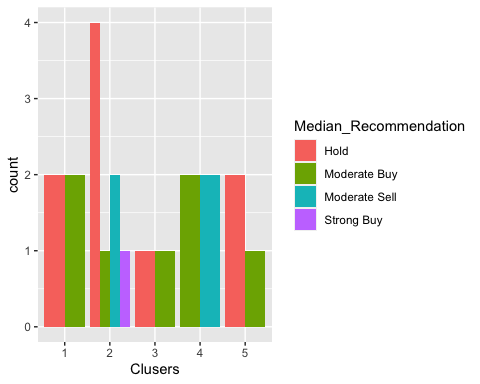
This Cluster represents a moderate-risk category. While not as extreme as some other clusters, careful consideration is still advised for entities in this group.

**Cluster 4 - Deadly, Despite Excellent PE Ratio**

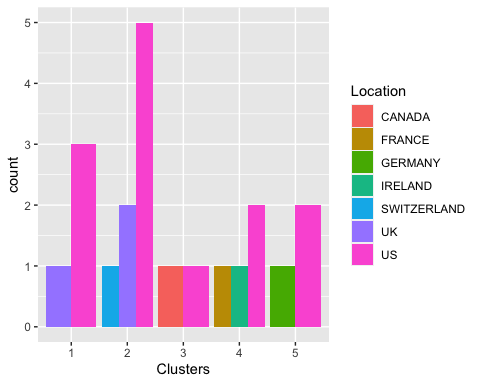
Despite having an excellent PE ratio, Cluster 4 is marked by exceptionally high risk, driven by elevated leverage, poor Net Profit Margin, and very low revenue growth. Ownership of entities in this cluster is considered highly risky.

**Cluster 5 - Fortune Overall Metrics** This Cluster showcases robust market capitalization, ROI, ROA, asset turnover, and Net Profit Margin. With a moderately valued PE ratio, entities in this cluster are deemed favorable for purchase and retention. The substantial revenue growth of 18.5% adds to the attractiveness of this cluster.

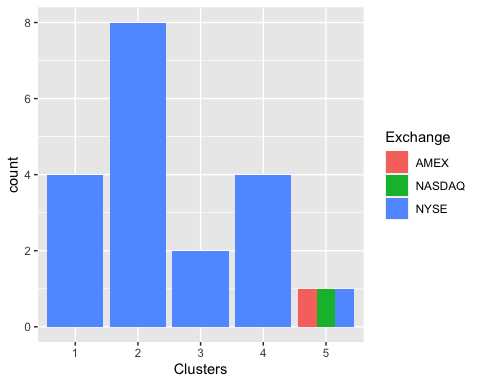
Pharmaceuticals1 = data[12:14] %>% mutate(Clusters = k5$cluster)  
ggplot(Pharmaceuticals1,mapping=aes(factor(Clusters),fill=Median\_Recommendation))+geom\_bar(position = 'dodge')+labs(x='Clusers')



ggplot(Pharmaceuticals1, mapping = aes(factor(Clusters),fill = Location))+geom\_bar(position = 'dodge')+labs(x = 'Clusters')



ggplot(Pharmaceuticals1, mapping = aes(factor(Clusters), fill = Exchange))+geom\_bar(position = 'dodge')+labs(x = 'Clusters')



A pattern can be observed in the median suggestions.

The most of the clusters/companies are listed on the NYSE and are based in the United States, but other than that, there doesn’t appear to be any discernible pattern among the clusters, locations, or exchanges.

**Cluster Interpretation according to variables:**

**Cluster 1**

**Median Suggestion** An average buy and sell suggestion is given for Cluster 1.

**Location** There are three places in Cluster 1, the most notable being the United States.

**Exchange** NYSE is the only one cluster in exchange.

**Cluster 2**

**Median Suggestion** Cluster 2 has a low hold and a low purchase.

**Location** The United States and Canada are the only two locations in Cluster 2, and they are dispersed equally.

**Exchange** NYSE is the only one cluster in exchange.

**Cluster 3**

**Median Suggestion** Cluster 3 has an extremely strong hold.

**Location** Cluster 3 has three locations, and is dominated by the United States, followed by the United Kingdom and Switzerland.

**Exchange** There is only one exchange in Cluster 3, the NYSE, and it has a big user base.

**Cluster 4**

**Median Suggestion** With a low buy rating, cluster 4 is rated as strongly held.

**Location** The US is ranked higher than Germany in two locations in Cluster 4.

**Exchange** Three equally distributed exchanges (AMEX, NASDAQ and NYSE) are located in Cluster 4.

**Cluster 5**

**Median Suggestion** A high buy and high hold rating are assigned to Cluster 5, based on the median recommendation.

**Location** There are two locations for Cluster 5, with a significant majority of the United States and the United Kingdom.

**Exchange** NYSE is the only one cluster in exchange.

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

**To name the clusters i have considered all the numerical variables below is the interpretations:**

**Cluster 1: High Profitability & Growth Leaders** This cluster excels in Net Profit Margin, has the lowest PE ratio, and experiences rapid sales growth. It is named for its emphasis on profitability and growth potential.

**Cluster 2: High Beta, Elevated PE Warning** Characterized by a notably high Beta and a warning for an elevated PE ratio, Cluster 2 is named for its emphasis on market sensitivity and the cautionary signal regarding valuation.

**Cluster 3: Moderate Risk, Balanced Metrics** Representing a moderate-risk category, Cluster 3 is named for its balance across various metrics. It avoids extremes and may offer a balanced risk-return profile.

**Cluster 4: High Risk, Low Profitability** Despite a strong PE ratio, Cluster 4 carries high risk due to elevated leverage, poor Net Profit Margin, and low revenue growth. It is named for its high-risk nature and lower profitability.

**Cluster 5: Robust Metrics & Growth Potential** Cluster 5 is named for its robust market capitalization, strong Return on Equity (ROE), Return on Assets (ROA), and growth potential indicated by substantial revenue growth. It represents entities with solid fundamentals and growth prospects.