ML Assignment 4

Pavan Chaitanya

# Loading Libraries

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.2.2

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

library(dplyr)  
library(ggplot2)  
library(hrbrthemes)

## Warning: package 'hrbrthemes' was built under R version 4.2.2

## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.  
## Please use hrbrthemes::import\_roboto\_condensed() to install Roboto Condensed and  
## if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow

Task a:

set.seed(100)  
# Importing the dataset  
Pharmacy <- read.csv("C:/Users/Pavan Chaitanya/Downloads/Pharmaceuticals.csv")  
head(Pharmacy)

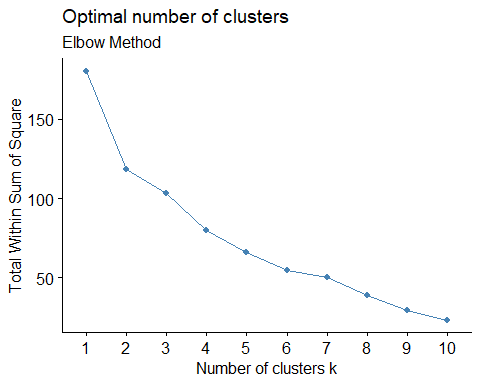
## Symbol Name Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 ABT Abbott Laboratories 68.44 0.32 24.7 26.4 11.8 0.7  
## 2 AGN Allergan, Inc. 7.58 0.41 82.5 12.9 5.5 0.9  
## 3 AHM Amersham plc 6.30 0.46 20.7 14.9 7.8 0.9  
## 4 AZN AstraZeneca PLC 67.63 0.52 21.5 27.4 15.4 0.9  
## 5 AVE 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

#Cleaning the data and checking for any null values in each of the column

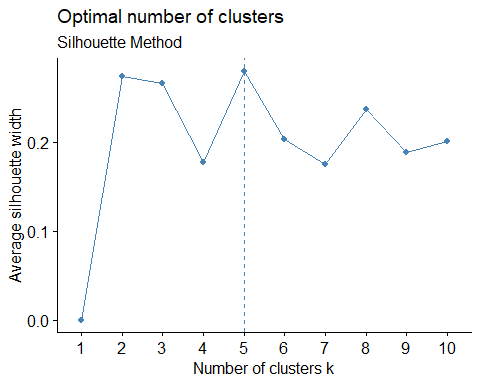
colSums(is.na(Pharmacy)) #returns the number of null values in each column

## Symbol Name Market\_Cap   
## 0 0 0   
## Beta PE\_Ratio ROE   
## 0 0 0   
## ROA Asset\_Turnover Leverage   
## 0 0 0   
## Rev\_Growth Net\_Profit\_Margin Median\_Recommendation   
## 0 0 0   
## Location Exchange   
## 0 0

# Selecting the numericals variables and normalizing the dataset.  
rownames(Pharmacy)<- Pharmacy$Symbol  
Pharmaceuticals <- Pharmacy[,c(3:11)]  
  
  
  
#Normalizing the numerical variables   
Normalized\_Pharmaceuticals = scale(Pharmaceuticals)  
  
# Elbow Method on scaled data to determine the value of k  
fviz\_nbclust(Normalized\_Pharmaceuticals,kmeans,method = "wss")+labs(subtitle="Elbow Method")



# Silhouette Method on scaled data to determine the number of clusters  
fviz\_nbclust(Normalized\_Pharmaceuticals,kmeans,method = "silhouette")+labs(subtitle="Silhouette Method")

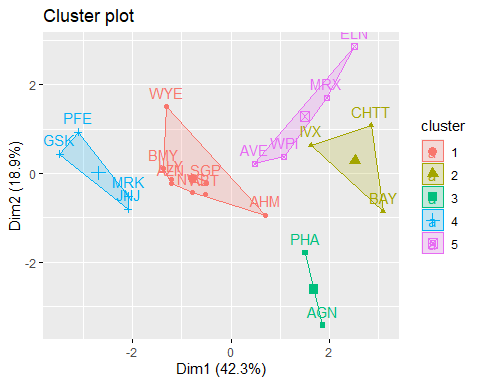


The distribution of data points on the scale is depending on the weight of each variable, and this has an impact on the clusters. The spacing between the data points will be affected as a result, and the clusters will follow.

WSS: At k = 2, the plot resembles an arm with a distinct elbow; however, because of the decision uncertainty, we might also choose 2,3,4,5, and the graph is not sharp and clear.

Silhouette We can clearly see a peak at k = 5 in the graph above that was produced using the silhouette method. So, take into account the silhouette strategy.

#Silhouette:  
Sil\_k5 = kmeans(Normalized\_Pharmaceuticals, centers=5,nstart=50)  
  
fviz\_cluster(Sil\_k5,data=Normalized\_Pharmaceuticals)



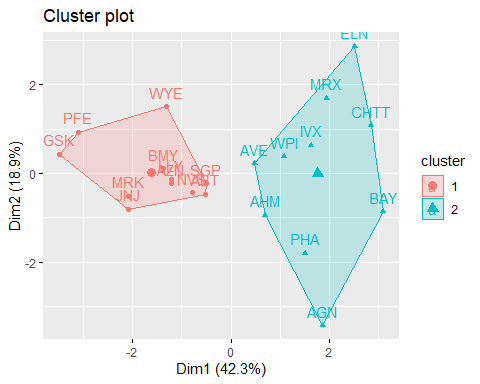
Silhouette\_group=Sil\_k5$cluster  
Sil\_k5$withinss

## [1] 21.879320 15.595925 2.803505 9.284424 12.791257

Sil\_k5$tot.withinss

## [1] 62.35443

#WSS  
Elb\_k2 = kmeans(Normalized\_Pharmaceuticals, centers=2,nstart=50)  
  
fviz\_cluster(Elb\_k2,data=Normalized\_Pharmaceuticals)



Elb\_k2$withinss

## [1] 43.30886 75.26049

Elb\_k2$tot.withinss

## [1] 118.5693

The total sum of squares within the cluster for Silhouette method is 62.35 which is smaller than WSS method 11.56.

The best value is k=5(Silhouette)

Task b:

Silhouette\_group = as.data.frame(Silhouette\_group)  
  
Sil\_Pharmaceuticals=cbind(Pharmaceuticals,Silhouette\_group)  
  
Cluster\_mean= Sil\_Pharmaceuticals %>% group\_by(Silhouette\_group) %>% summarise\_all("mean")  
Cluster\_mean

## # A tibble: 5 × 10  
## Silhouette…¹ Marke…² Beta PE\_Ra…³ ROE ROA Asset…⁴ Lever…⁵ Rev\_G…⁶ Net\_P…⁷  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 55.8 0.414 20.3 28.7 12.7 0.738 0.371 5.59 19.4   
## 2 2 6.64 0.87 24.6 16.5 4.17 0.6 1.65 5.73 7.03  
## 3 3 31.9 0.405 69.5 13.2 5.6 0.75 0.475 12.1 6.4   
## 4 4 157. 0.48 22.2 44.4 17.7 0.95 0.22 18.5 19.6   
## 5 5 13.1 0.598 17.7 14.6 6.2 0.425 0.635 30.1 15.6   
## # … with abbreviated variable names ¹​Silhouette\_group, ²​Market\_Cap, ³​PE\_Ratio,  
## # ⁴​Asset\_Turnover, ⁵​Leverage, ⁶​Rev\_Growth, ⁷​Net\_Profit\_Margin

Cluster 1

This cluster’s companies are less indebted than those in other clusters because it has lower leverage than those other clusters.

This cluster has the lowest revenue growth of all the groups, but the businesses in it have the highest net profit margins.

When the other factors are taken into account, this cluster’s businesses are performing better than Clusters 2, 3, and 5.

Cluster 2

This cluster has a greater mean beta value than other clusters. This shows that the stock prices of the companies in this cluster are more erratic. This cluster has the highest mean leverage, indicating that the debt levels of these businesses are higher. The companies in this cluster have less Market Capital, ROA, Revenue Growth, and Net Profit Margin. This indicates that these companies need to develop financially.

Cluster 3

The businesses in this cluster have the lowest net profit margins. Furthermore, this cluster has the lowest Return on Equity (ROE), a sign that the businesses in it have a hard time turning equity investments into profits. Additionally, this cluster has the highest Price-Earnings Ratio, which indicates that the companies are not profitable. Because this cluster has the lowest beta value, even though these companies’ profits are declining, we can still see that their stocks are less volatile.

Cluster 4

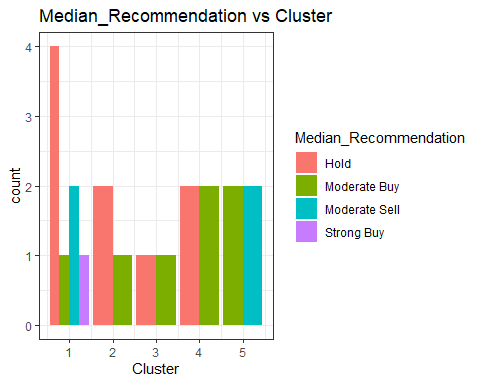
The market capitalization, net profit margin, return on assets (ROA), return on equity (ROE), and asset turnover of the companies in this cluster are all at their highest levels. The businesses in this cluster have the lowest mean leverage values, which means that their debt to shareholders’ equity ratios are lower. As a result, this cluster has the highest performing firms when compared to other clusters.

Cluster 5

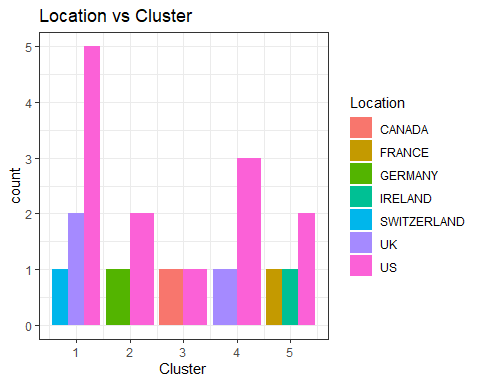
High revenue growth among the businesses in this cluster is an indication that business development is going as planned. The companies should, ideally, use their assets to boost revenue, which raises the asset turnover ratio. The asset turnover ratio for this cluster is the lowest, nevertheless. The fact that this group of businesses has the lowest price-to-earnings ratio suggests that their earnings are higher.

Task c:

Pharma\_categorical= Pharmacy[,12:14]  
  
Cluster\_Pharma\_categorial = cbind(Pharma\_categorical,Silhouette\_group)  
  
ggplot(Cluster\_Pharma\_categorial, aes(x = Silhouette\_group, fill = Median\_Recommendation)) +  
 geom\_bar(position = "dodge") + labs(title = "Median\_Recommendation vs Cluster", x = "Cluster") + theme\_bw()



# Looking at the median recommendation plot, I can see that Cluster 1 only has one "Strong Buy" recommendation and has a lot of "Hold" recommendations. All of the clusters have a distribution of moderate buy.  
  
  
ggplot(Cluster\_Pharma\_categorial, aes(x = Silhouette\_group, fill = Location)) +  
 geom\_bar(position = "dodge") + labs(title = "Location vs Cluster", x = "Cluster") +theme\_bw()



# I can see that all of the clusters have US-based enterprises from the Location vs. Cluster Plot. However, different places can be found throughout all clusters.

Task d :

Cluster names:

Cluster 1 - Enlarging Companies

Cluster 2 - Massive debt Companies

Cluster 3 - Little-profit Companies

Cluster 4 - Most efficient Companies

Cluster 5 - Increased Income Companies