ML-Assignment 4

2022-11-04

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ✔ purrr 0.3.5   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ purrr::lift() masks caret::lift()

library(factoextra)

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

library(esquisse)  
set.seed(123)

getwd()

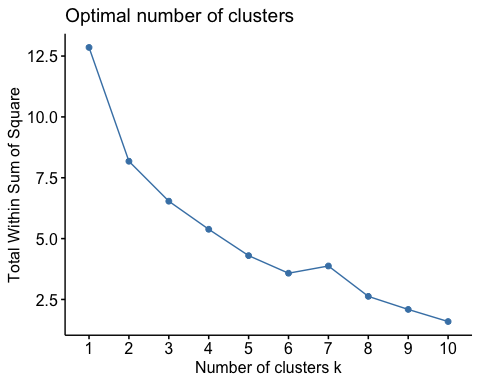
## [1] "/Users/thupiliabhinav/Desktop/ML/ML- Assignment 4"

setwd("/Users/thupiliabhinav/Desktop/ML/ML- Assignment 4")  
pharma <- read.csv("Pharmaceuticals.csv")

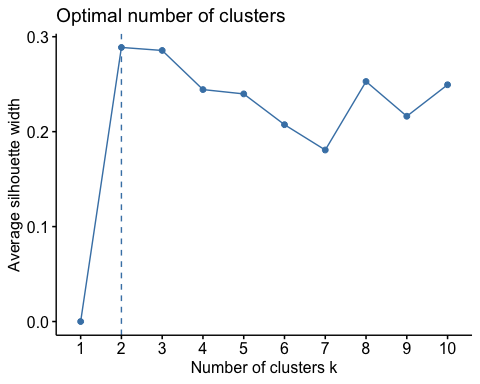
**#a. Using only the numerical variables (1 to 9) to cluster the 21.**

norm\_mean<- pharma %>% select('Market\_Cap', 'Beta', 'PE\_Ratio', 'ROE', 'ROA', 'Asset\_Turnover', 'Leverage', 'Rev\_Growth', 'Net\_Profit\_Margin')  
  
**#Scaling the Data.**  
norm\_train <- preProcess(norm\_mean, method = "range")  
norm\_predict<-predict(norm\_train, norm\_mean)

fviz\_nbclust(norm\_predict, kmeans, method = "wss")



fviz\_nbclust(norm\_predict, kmeans, method = "silhouette")



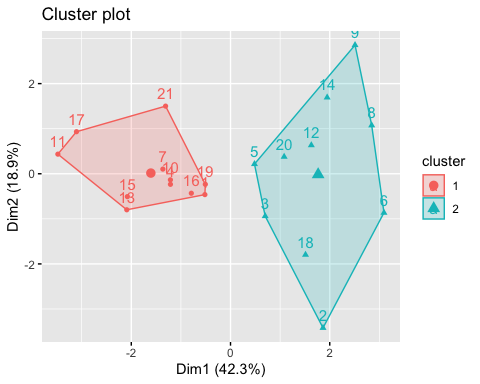
**#From above we calculate kmeans optimal being k=2:**

k\_means\_2 <- kmeans(norm\_predict, centers = 2, nstart = 25)  
k\_means\_2$centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## 1 0.48580145 0.2727273 0.2199562 0.5389831 0.7171717 0.6250 0.09272209  
## 2 0.06949161 0.4806452 0.3399240 0.1864407 0.2238095 0.3625 0.24843305  
## Rev\_Growth Net\_Profit\_Margin  
## 1 0.3567294 0.7673680  
## 2 0.5368646 0.3567686

**#Graphical representation of kmeans using cluster:**

fviz\_cluster(k\_means\_2, data= norm\_mean)



**#Grouping of clusters with original data:**

k\_cluster<- k\_means\_2$cluster  
group\_k <- cbind(pharma,k\_cluster)

#Calculating mean for both clusters:

aggregate(group\_k[,-c(1,2,12:14)],by=list(group\_k$k\_cluster),FUN="mean")

## Group.1 Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## 1 1 97.11364 0.4336364 20.95455 35.7 14.95455 0.80 0.3254545  
## 2 2 14.24300 0.6270000 30.42000 14.9 5.63000 0.59 0.8720000  
## Rev\_Growth Net\_Profit\_Margin k\_cluster  
## 1 10.16455 20.17273 1  
## 2 16.89800 10.77000 2

**#b).** Interpret the clusters with respect to the numerical variables used in forming the clusters.

# From above we can observe through clustering by “WSS” and “Silhouette” optimal K is 2 #Cluster 1- has companies with High-Market\_Cap, PE\_Ratio, ROE, ROA, and Net\_Profit\_Margin.

#Cluster 2- has companies with Low- Market\_Cap, PE\_Ratio, ROE, ROA and Net\_Profit\_Margin.

**#Grouping of clusters with Original Data:**

groupk2<-cbind(group\_k, pharma$Location,pharma$Exchange)

**#c).**Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? #With respect to numerical values to columns (10 to 12) are as follows:

#Analysis under column 10.Mediation\_recommendation: #Mediation\_recommendation under cluster 1 consists-

1. hold recommendations-6 , b) buy recommendations-3,

c) sell recommendations - 2

##Mediation\_recommendation under cluster 2 consists-

a) buy recommendations-5 , b) hold recommendations-3, c) sell recommendations - 2

#Analysis under columns 11.pharmaExchange: #Majority of pharma locations in cluster-1 and cluster-2 are US- based and for pharma exchange the majority is NYSE for both the clusters.

#**d).**Naming for each cluster using any or all of the variables in the dataset: #Cluster 1- With majority mediation recommendations being held, this cluster is named “HOLD CLUSTER”.

#Cluster 2- With majority mediation recommendations being bought, this cluster is named “BUY CLUSTER”.