FML\_Assignment\_4

Ravi Gadde

2023-03-18

Loading the data set and the Libraries

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.3.3

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(cluster)  
library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'tidyr' was built under R version 4.3.2

## Warning: package 'readr' was built under R version 4.3.2

## Warning: package 'dplyr' was built under R version 4.3.2

## Warning: package 'stringr' was built under R version 4.3.2

## Warning: package 'lubridate' was built under R version 4.3.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ 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)

## Warning: package 'factoextra' was built under R version 4.3.3

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

library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.3.3

library(tinytex)

## Warning: package 'tinytex' was built under R version 4.3.2

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.3.3

P\_Data<-read.csv("C:/Users/Dell/Downloads/Pharmaceuticals.csv")  
P\_Data<-na.omit(P\_Data)

*TASK 1* The 21 firms are grouped using the numerical variables (1–9).

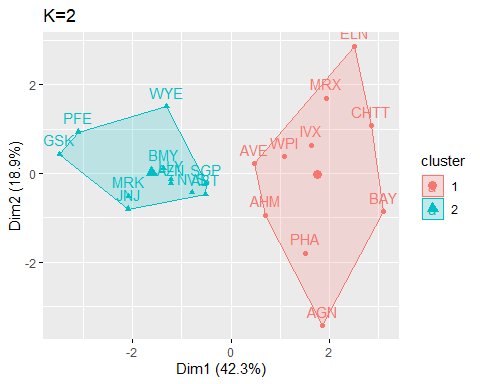
row.names(P\_Data)<-P\_Data[,1]  
Clustering\_dataset<-P\_Data[,3:11]

data scalability

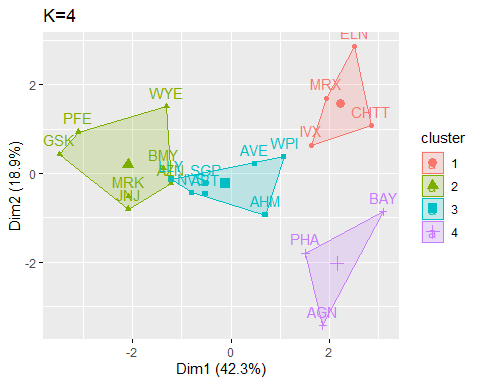
set.seed(143)  
Scaled\_data<-scale(Clustering\_dataset)

Kmeans computation using random K values

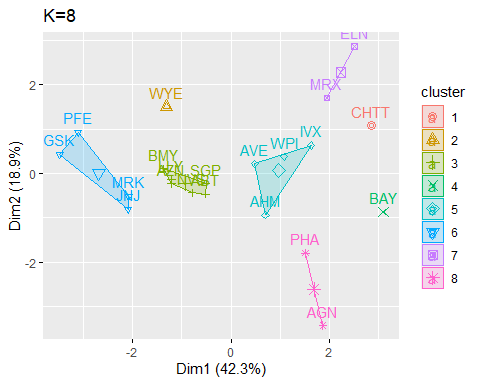
set.seed(143)  
kmeans\_2\_centers<-kmeans(Scaled\_data,centers = 2, nstart = 15)  
kmeans\_4\_centers<-kmeans(Scaled\_data,centers = 4, nstart = 15)  
kmeans\_8\_centers<-kmeans(Scaled\_data,centers = 8, nstart = 15)  
plot\_kmeans\_2\_centers<-fviz\_cluster(kmeans\_2\_centers,data = Scaled\_data) + ggtitle("K=2")  
plot\_kmeans\_4\_centers<-fviz\_cluster(kmeans\_4\_centers,data = Scaled\_data) + ggtitle("K=4")  
plot\_kmeans\_8\_centers<-fviz\_cluster(kmeans\_8\_centers,data = Scaled\_data) + ggtitle("K=8")  
plot\_kmeans\_2\_centers



plot\_kmeans\_4\_centers



plot\_kmeans\_8\_centers



Finding the optimal K appropriate for clustering using WSS and Silhouette

wss<-fviz\_nbclust(Scaled\_data,kmeans,method="wss")  
silhouette<-fviz\_nbclust(Scaled\_data,kmeans,method="silhouette")  
wss



silhouette



distance<-dist(Scaled\_data,metho='euclidean')  
fviz\_dist(distance)

 k is 2 from WSS and 5 from silhouette. The number 5 ensures that the sum of squires inside each cluster is minimal and that there is considerable spacing between them.

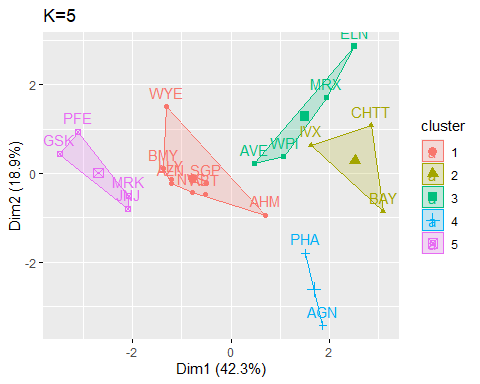
*TASK 2*

Using Kmeans to find an appropriate k

set.seed(143)  
kmeans\_5\_centers<-kmeans(Scaled\_data,centers = 5, nstart = 10)  
kmeans\_5\_centers

## K-means clustering with 5 clusters of sizes 8, 3, 4, 2, 4  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 2 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 3 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## 4 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 5 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.27449312 -0.7041516 0.556954446  
## 2 1.36644699 -0.6912914 -1.320000179  
## 3 0.06308085 1.5180158 -0.006893899  
## 4 -0.14170336 -0.1168459 -1.416514761  
## 5 -0.46807818 0.4671788 0.591242521  
##   
## Clustering vector:  
## ABT AGN AHM AZN AVE BAY BMY CHTT ELN LLY GSK IVX JNJ MRX MRK NVS   
## 1 4 1 1 3 2 1 2 3 1 5 2 5 3 5 1   
## PFE PHA SGP WPI WYE   
## 5 4 1 3 1   
##   
## Within cluster sum of squares by cluster:  
## [1] 21.879320 15.595925 12.791257 2.803505 9.284424  
## (between\_SS / total\_SS = 65.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

plot\_kmeans\_5\_centers<-fviz\_cluster(kmeans\_5\_centers,data = Scaled\_data) + ggtitle("K=5")  
plot\_kmeans\_5\_centers



Clustering\_dataset\_1<-Clustering\_dataset%>% mutate(Cluster\_no=kmeans\_5\_centers$cluster)%>% group\_by(Cluster\_no)%>%summarise\_all('mean')  
Clustering\_dataset\_1

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

Following clusters have been created for companies:

Cluster\_1= ABT,AHM,AZN,BMY,LLY,NVS,SGP,WYE

Cluster\_2= BAY,CHTT,IVX

Cluster\_3=AVE,ELN,MRX,WPI

Cluster\_4=AGN,PHA

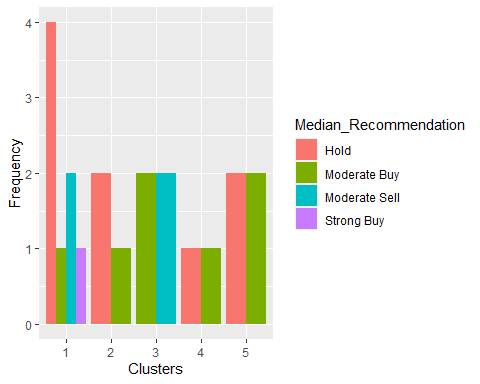
Cluster\_5=GSK,JNJ,MRK,PFE

This can be inferred from the clusters that were generated.

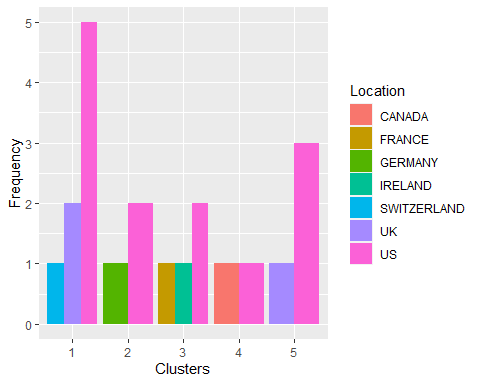
1. Cluster 1 contains a collection of businesses with a modest return on equity and return on investment.
2. Cluster 2 Companies have extremely low ROA, ROE, market capitalization, and asset turnover. This means that these businesses are exceedingly dangerous.
3. Similar to cluster 2, Cluster 3 features group corporations, but with slightly lower risk.
4. Companies in cluster 4 are more risky than those in cluster 2 because they have very good PE ratios but weak ROA and ROE.
5. Companies in Cluster 5 have excellent ROE, ROA, and market capitalization.

*TASK 3*

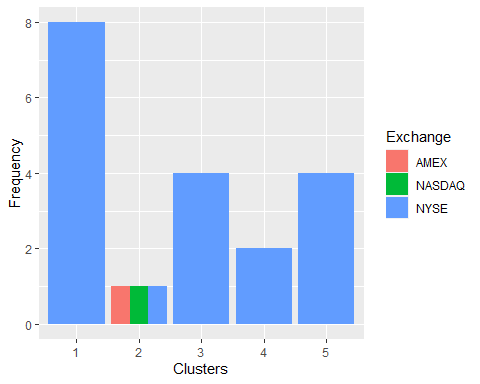
#Is there a pattern in the clusters with respect to the numerical   
#variables (10 to 12)? (those \n #not used in forming the clusters)  
Clustering\_datase\_2<- P\_Data[,12:14] %>% mutate(Clusters=kmeans\_5\_centers$cluster)  
ggplot(Clustering\_datase\_2, mapping = aes(factor(Clusters), fill =Median\_Recommendation))+geom\_bar(position='dodge')+labs(x ='Clusters',y='Frequency')



ggplot(Clustering\_datase\_2, mapping = aes(factor(Clusters),fill = Location))+geom\_bar(position = 'dodge')+labs(x ='Clusters',y='Frequency')



ggplot(Clustering\_datase\_2, mapping = aes(factor(Clusters),fill = Exchange))+geom\_bar(position = 'dodge')+labs(x ='Clusters',y='Frequency')



Clusters and the variable Median Recommendation exhibit a pattern, as can be observed. Similar to what the second cluster shows between moderate buy and hold, the third cluster recommends between moderate purchase and moderate sell. The majority of pharmaceutical businesses are based in the US, as can be seen from the location graph, although there isn’t much of a pattern there. With the exception of the bulk of companies being listed on NYSE, there is no discernible relationship between clusters and exchanges.

*TASK 4* - Naming clusters:

[It is done based on the net Market capitalization(size) and Return on Assets(money)]

Cluster 1: Large-Thousands Cluster 2: Extra Small-Penny Cluster 3: Small- Dollars Cluster 4: Medium-Hundreds Cluster 5: Extra Large-Millions