Assignment 4

Ameer Abdlrasul

2023-10-29

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 3.4.4 ✔ 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

pharmaceutical\_0data<-read.csv("~/Downloads/Pharmaceuticals.csv")  
pharmaceutical\_data<-na.omit(pharmaceutical\_0data)

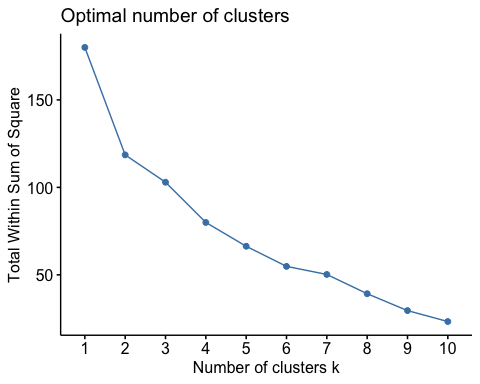
Using the numerical variables (1 to 9) to cluster the 21 firms.

row.names(pharmaceutical\_0data)<-pharmaceutical\_0data[,1]  
Clustering\_data<-pharmaceutical\_0data[,3:11]

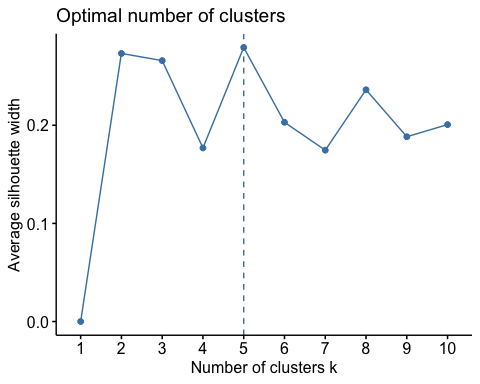
Scaling the data according to requirement

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

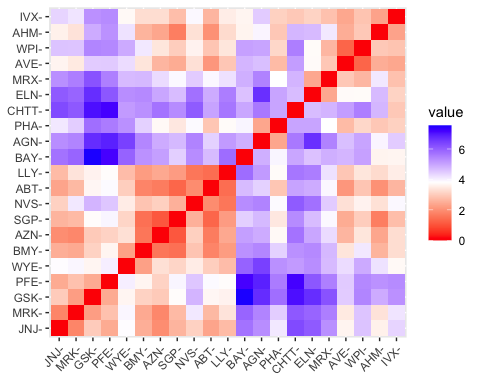
k\_wss<-fviz\_nbclust(ScaledData,kmeans,method="wss")  
k\_silhouette<-fviz\_nbclust(ScaledData,kmeans,method="silhouette")  
k\_wss



k\_silhouette



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

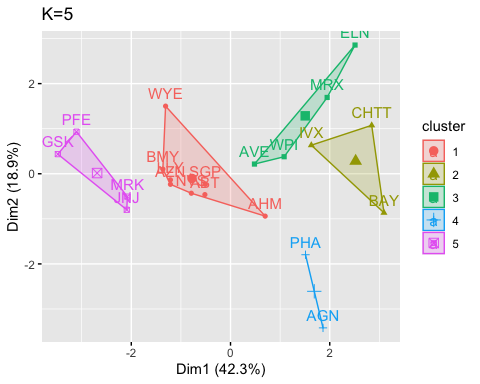


We will be choose silhouette k value of 5 because this will ensure that the sum of squares is low along with proper separation within the clusters.

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

## 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<-fviz\_cluster(kmeans\_5,data = ScaledData) + ggtitle("K=5")  
plot\_kmeans\_5

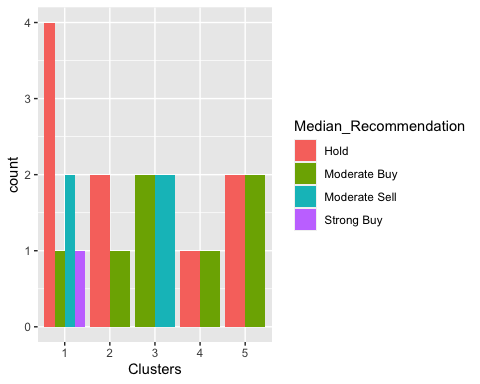


Clustering\_data\_1<-Clustering\_data%>%  
 mutate(Cluster\_no=kmeans\_5$cluster)%>%  
 group\_by(Cluster\_no)%>%summarise\_all('mean')  
Clustering\_data\_1

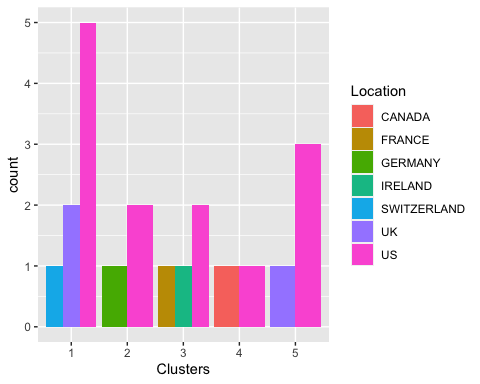
## # A tibble: 5 × 10  
## Cluster\_no Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover Leverage  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 55.8 0.414 20.3 28.7 12.7 0.738 0.371  
## 2 2 6.64 0.87 24.6 16.5 4.17 0.6 1.65   
## 3 3 13.1 0.598 17.7 14.6 6.2 0.425 0.635  
## 4 4 31.9 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>

Companies are grouped into following clusters: 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 From the above clusters 1.Cluster\_1 Companies with moderate returns on equity and investment 2.Cluster\_2 Companies with Poor returns on Assets(ROA), Return on Equity (ROE), Low market Capitalization, and weak Asset turnover which implies that these companies are Highly Risky 3.Cluster\_3 Companies Similar to those in cluster 2 but with slightly lower levels of risk 4.Cluster\_4 Companies with very high price to earnings (P/E) ratios but poor ROA and ROE, making them more risky than those in cluster 2 5.Cluster\_5 Companies with high market capitalization, ROE and ROA

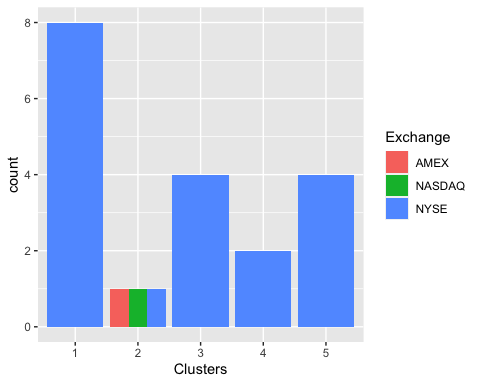
Clustering\_data\_2<- pharmaceutical\_data[,12:14] %>% mutate(Clusters=kmeans\_5$cluster)  
ggplot(Clustering\_data\_2, mapping = aes(factor(Clusters), fill =Median\_Recommendation))+geom\_bar(position='dodge')+labs(x ='Clusters')



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



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



By examining the data, a distinct pattern becomes apparent concerning the connection between the clusters and the ‘Median Recommendation’ variable. Cluster 2 predominantly indicates recommendations spanning from ‘hold’ to ‘moderate buy,’ while Cluster 3 tends to favor recommendations from ‘moderate buy’ to ’moderate sell.’Upon closer inspection of the geographical distribution, it becomes evident that a substantial percentage of pharmaceutical companies are headquartered in the United States, and there is no conspicuous spatial arrangement. Nevertheless, there is no apparent association between clusters and the stock exchange, aside from the observation that the majority of these companies are publicly listed on the New York Stock Exchange (NYSE).

Naming clusters:

**Cluster 1 = Investment Cluster.**

**Cluster 2 = Tight Cluster.**

**Cluster 3 = Low Profit Cluster.**

**Cluster 4 = Risky Cluster.**

**Cluster 5 = Large Purchase Cluster.**