FML- ASSIGNMENT-4

LOKESH JETANGI

2023-03-17

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

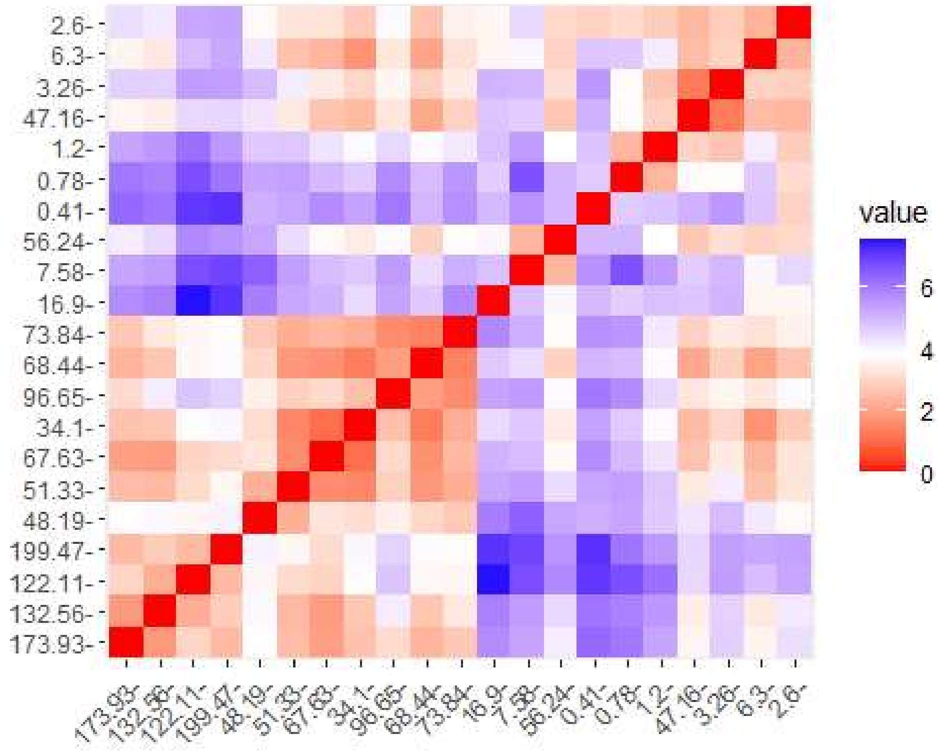
```
Pharmaceuticals<- read.csv("C:/Users/jetan/Downloads/Pharmaceuticals.csv")
head(Pharmaceuticals)

##   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
```

```
#install.packages("flexclust")
#install.packages("cluster")
#install.packages("tidyverse")
#install.packages("factoextra")
#install.packages("FactoMineR")
#install.packages("ggcorrplot")
#install.packages("tinytex")
#install.packages("NbClust")
library(tinytex)
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     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.1.8
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; 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)
library(NbClust)

# TASK-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.
```

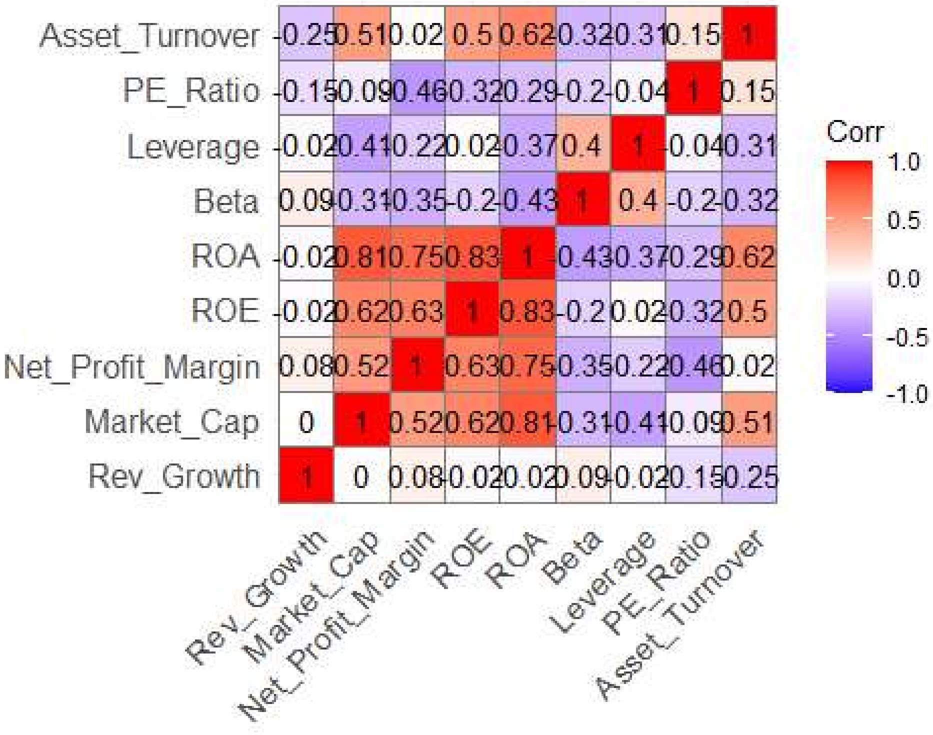
```
#Use only the numerical variables (1 to 9) to cluster
Pharmaceuticals1<-Pharmaceuticals[,c(3:11)]
row.names(Pharmaceuticals1)<-Pharmaceuticals1[,1]
view(Pharmaceuticals1)

#Using the Euclidean distance formula which is given by
```

$$\text{distance} = \sqrt{(\mathrm{X2}-\mathrm{X1})^2+(\mathrm{Y2}-\mathrm{Y1})^2}$$

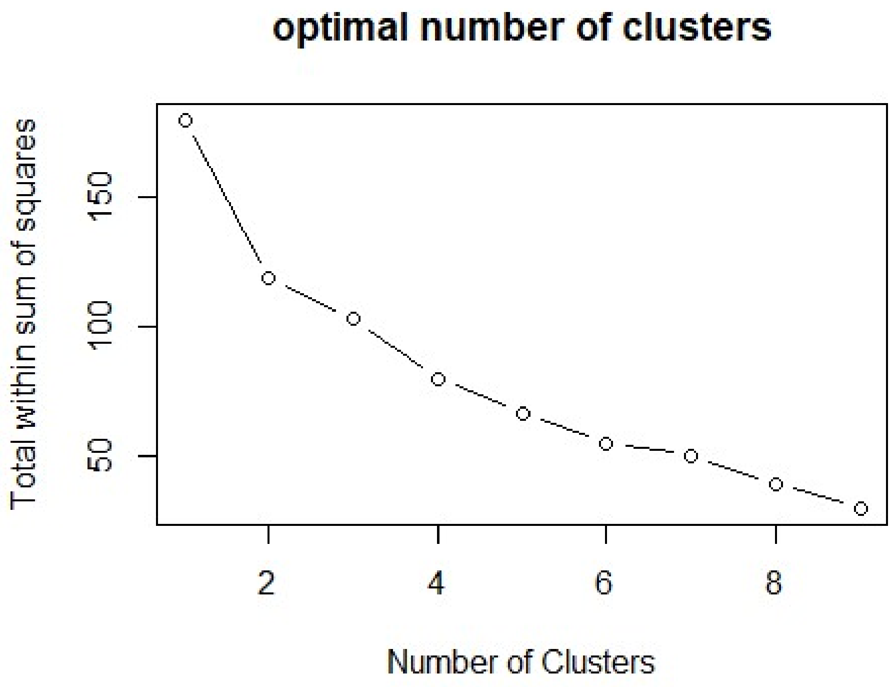
```
##Normalizing the data
Pharmaceuticals2<-scale(Pharmaceuticals1)
distance<-get_dist(Pharmaceuticals2)
fviz_dist(distance)
```



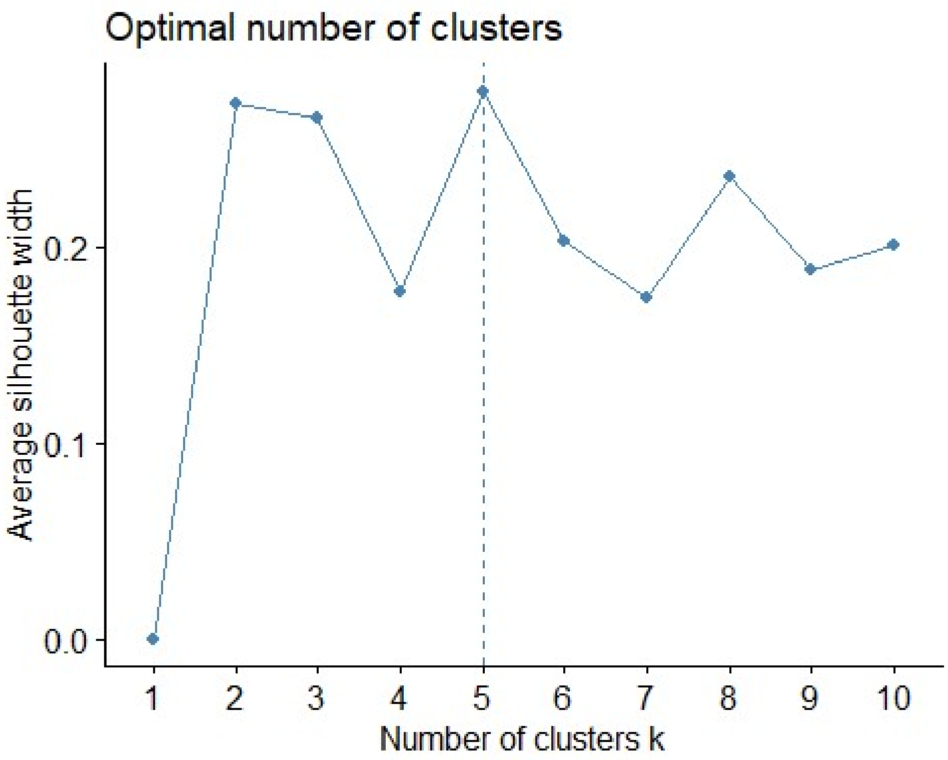
```
## plotting the graph
corelation<-cor(Pharmaceuticals2)
ggcorrplot(corelation, outline.color="grey50", lab=TRUE, hc.order = TRUE,
type = "full")
```



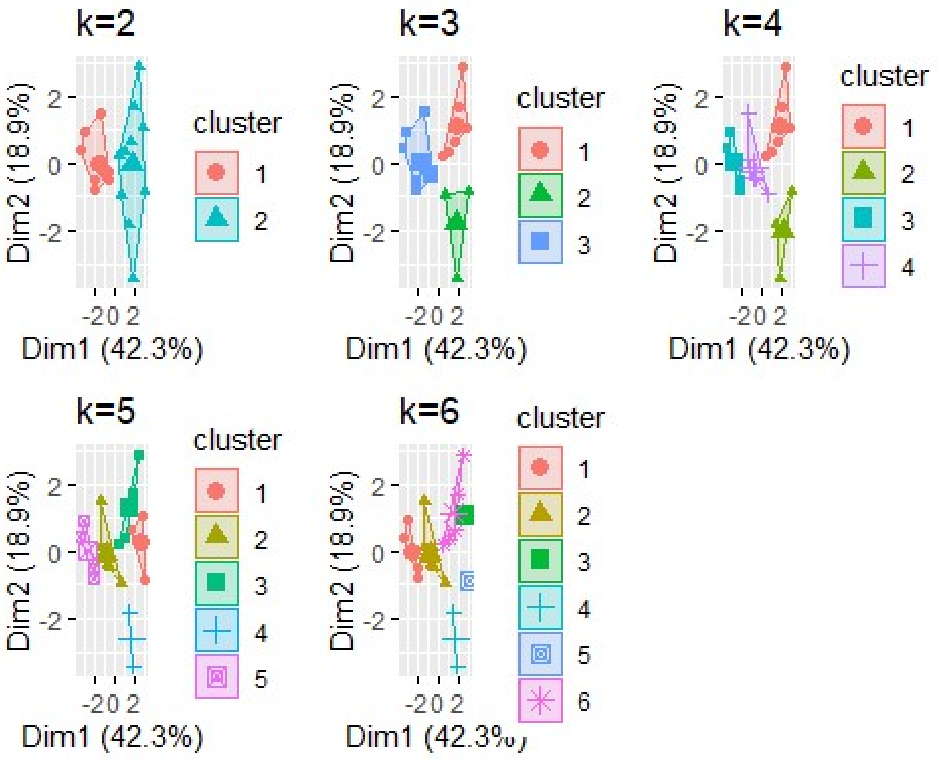
```
##There are two main methods to find the value of K or number of cluster
#Elbow chart and the Silhouette Method
#To determine the number of clusters to do the cluster analysis using Elbow
Method
set.seed(123)
wss<- vector()
for(i in 1:9 )wss[i]<- sum(kmeans(Pharmaceuticals2,i) $withinss)
plot(1:9, wss, type = "b", main =paste("optimal number of clusters"),
     xlab = "Number of Clusters",
     ylab = "Total within sum of squares")
```



```
#It is not clear from the Elbow method's above graph whether to use k=2 or 3,
4, or 5.
#Silhouette method for determining number of clusters

fviz_nbclust(Pharmaceuticals2, kmeans, method = "silhouette")
```



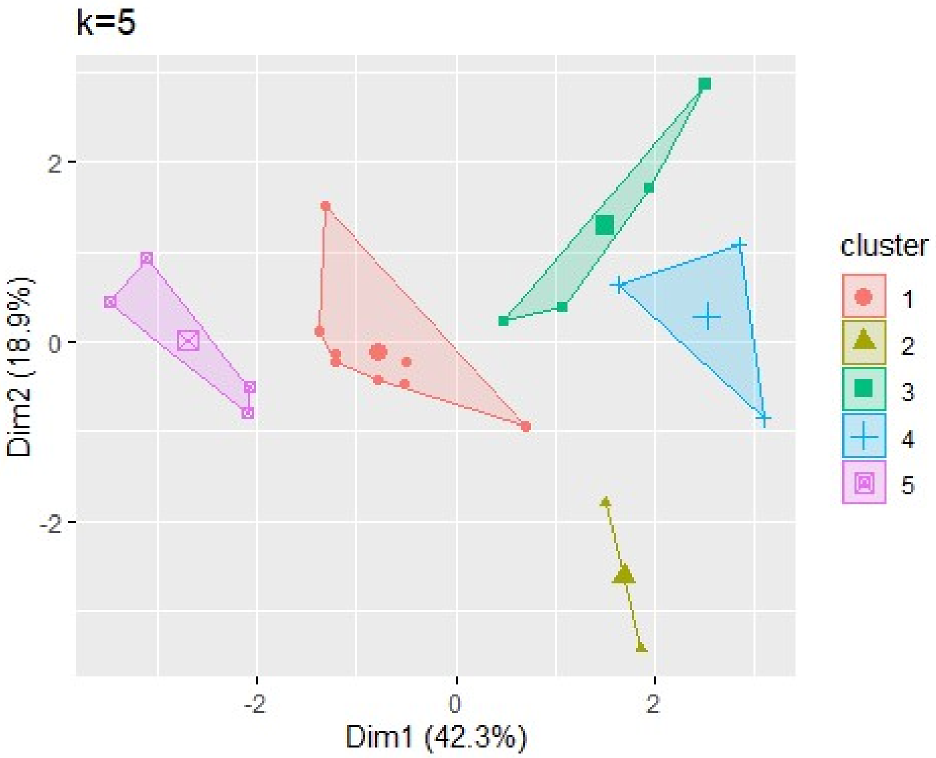
```
k2<-kmeans(Pharmaceuticals2,centers =2,nstart=25)
k3<-kmeans(Pharmaceuticals2,centers =3,nstart=25)
k4<-kmeans(Pharmaceuticals2,centers =4,nstart=25)
k5<-kmeans(Pharmaceuticals2,centers =5,nstart=25)
k6<-kmeans(Pharmaceuticals2,centers =6,nstart=25)
p1<-fviz_cluster(k2,geom = "point", data=Pharmaceuticals2)+ggtitle("k=2")
p2<-fviz_cluster(k3,geom = "point", data=Pharmaceuticals2)+ggtitle("k=3")
p3<-fviz_cluster(k4,geom = "point", data=Pharmaceuticals2)+ggtitle("k=4")
p4<-fviz_cluster(k5,geom = "point", data=Pharmaceuticals2)+ggtitle("k=5")
p5<-fviz_cluster(k6,geom = "point", data=Pharmaceuticals2)+ggtitle("k=6")

library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

grid.arrange(p1,p2,p3,p4,p5,nrow=2)
```



```
# from the above observation value of k=5 is making more sense
K5<-kmeans(Pharmaceuticals2,centers = 5, nstart = 25)
K5

## K-means clustering with 5 clusters of sizes 8, 2, 4, 3, 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 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 3 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 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 -0.14170336 -0.1168459      -1.416514761
## 3  0.06308085  1.5180158      -0.006893899
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.46807818  0.4671788       0.591242521
##
## Clustering vector:
##  68.44   7.58    6.3  67.63  47.16   16.9  51.33   0.41   0.78  73.84
122.11
##      1      2      1      1      3      4      1      4      3      1
5
##    2.6 173.93    1.2 132.56  96.65 199.47  56.24   34.1   3.26  48.19
##      4      5      3      5      1      5      2      1      3      1
```

```
##
## Within cluster sum of squares by cluster:
## [1] 21.879320  2.803505 12.791257 15.595925  9.284424
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

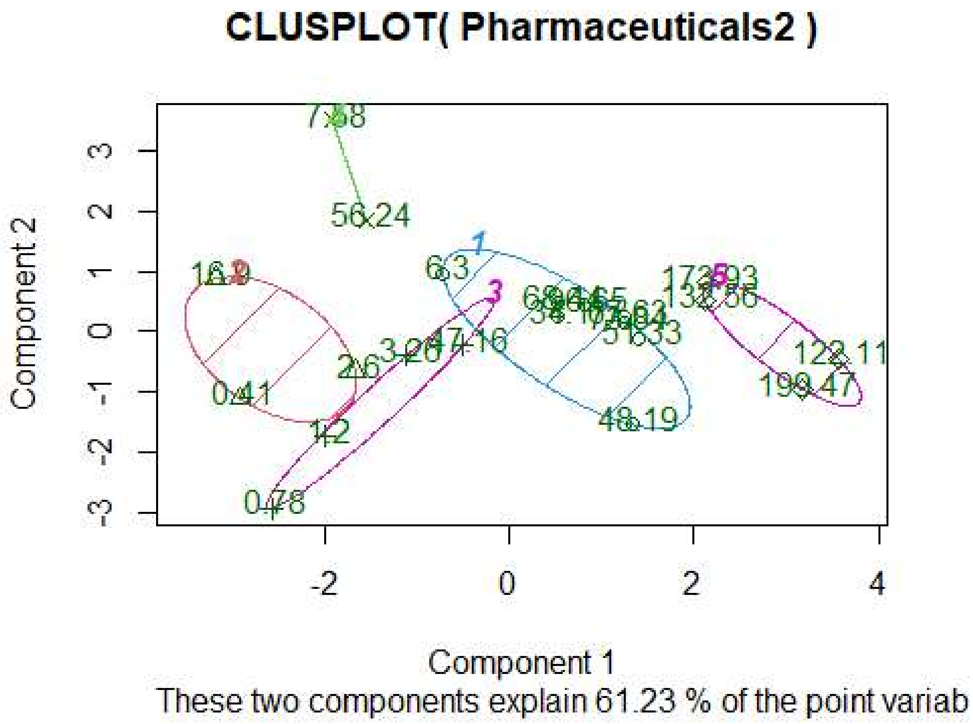
```
p5<-fviz_cluster(K5, geom ="point", data =Pharmaceuticals2 )+ggtitle("k=5")#
to Visualize the clusters
p5
```



```
##Applying K-means
#Visualizing the output
#centroids
k5<-kmeans(Pharmaceuticals2, centers = 5, nstart = 25)
k5$centers
```

```
##     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
```

```
#To view the cluster plot
clusplot(Pharmaceuticals2,k5$cluster,color = TRUE,shade = TRUE, labels  =2, 
lines = 0)
```



```
#TASK-2
#Interpret the clusters with respect to the numerical variables used in 
forming the clusters.

aggregate(Pharmaceuticals2,by=list(k5$cluster),FUN=mean)
```

```
##   Group.1  Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 2       2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478
## 3       3 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428
## 4       4 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951
## 5       5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
##   Asset_Turnover    Leverage Rev_Growth Net_Profit_Margin
## 1      0.1729746 -0.27449312 -0.7041516       0.556954446
## 2     -0.4612656  1.36644699 -0.6912914      -1.320000179
## 3     -1.2684804  0.06308085  1.5180158      -0.006893899
## 4      0.2306328 -0.14170336 -0.1168459      -1.416514761
## 5      1.1531640 -0.46807818  0.4671788       0.591242521
```

```
jl_pharmacy<-data.frame(Pharmaceuticals2,k5$cluster)
head(jl_pharmacy)
```

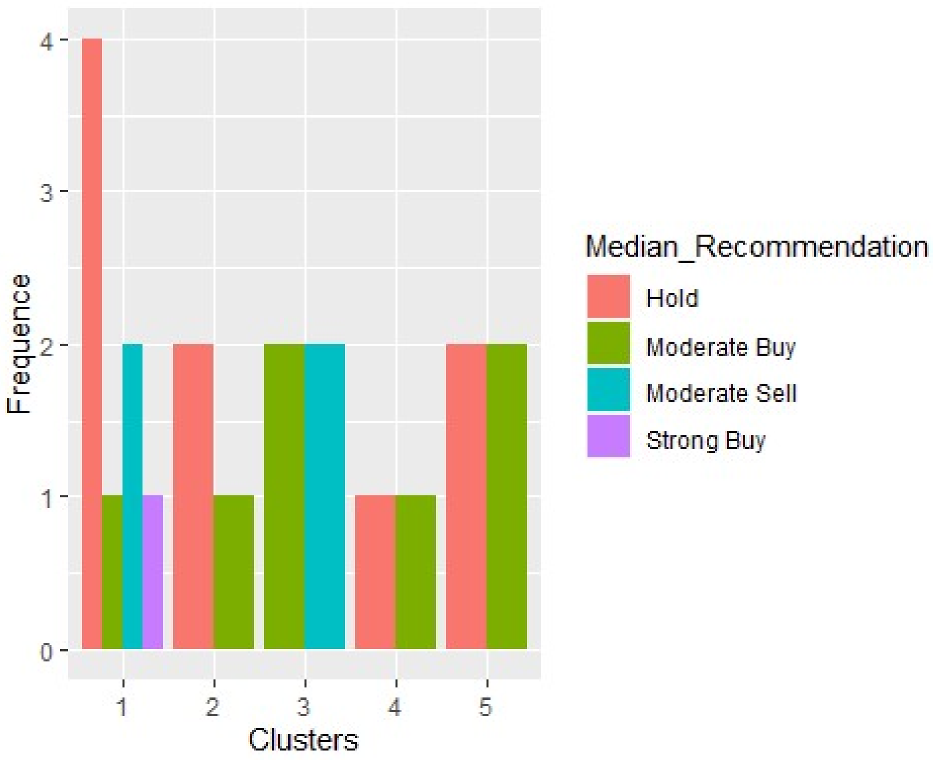
```
##        Market_Cap        Beta    PE_Ratio         ROE         ROA
Asset_Turnover
## 68.44  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121
0.0000000
## 7.58  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871
0.9225312
## 6.3   -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700
0.9225312
## 67.63  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259
0.9225312
## 47.16 -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -
0.4612656
## 16.9  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -
0.4612656
##          Leverage Rev_Growth Net_Profit_Margin k5.cluster
## 68.44 -0.2120979 -0.5277675        0.06168225          1
## 7.58   0.0182843 -0.3811391       -1.55366706          4
## 6.3   -0.4040831 -0.5721181       -0.68503583          1
## 67.63 -0.7496565  0.1474473        0.35122600          1
## 47.16 -0.3144900  1.2163867       -0.42597037          3
## 16.9  -0.7496565 -1.4971443       -1.99560225          2
```

```
#cluster-1: ABT,AHM,AZN,BMY,LLY,NVS,SGP,WYE
#This cluster features a low PE ratio and modearte leverage,net profit margin
and they have a moderate asset turnover, ROE and ROA that are near to value
zero

#cluster-2: BAY,CHTT,IVX
#This cluster has the highest beta and least net profit margine and they also
have low asset turnover and negative value of ROE,ROA.

#cluster-3: AVE,ELN,MRX,WPI
# This cluster has high Rev_Growth,Market_Cap is similar to all the clusters
they also have least Asset_Turnover and negative ROA,ROE.

#cluster-4: AGN,PHA
#  This cluster has highest PE_Ratio and least net proft margine,
ROA,Market_cap

#cluster-5: GSK,JNJ,MRK,PFE
#This cluster has the highest ROE,ROA AND market_cap is high compare to other
clusters and they have low leverage.
```

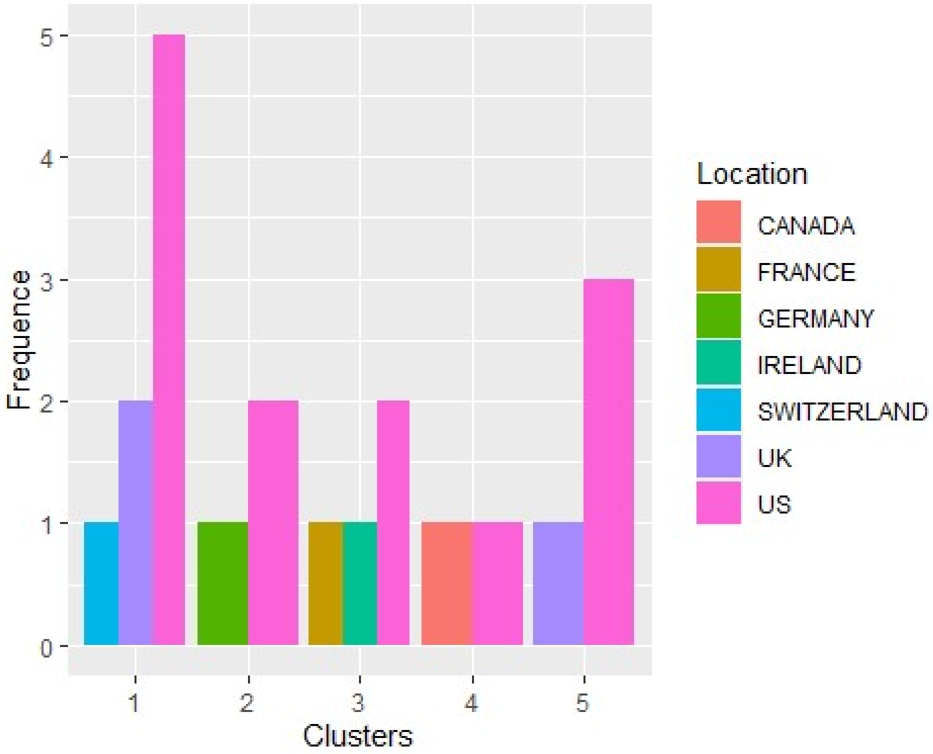
```
#TASK-3
# Is there a pattern in the clusters with respect to the numerical variables
(10 to 12)? (those not used in forming the clusters)
pattern <- Pharmaceuticals %>% select(c(12,13,14)) %>% mutate(Cluster =
k5$cluster)
print(pattern)
```

```
##    Median_Recommendation    Location Exchange Cluster
## 1           Moderate Buy          US     NYSE       1
## 2           Moderate Buy      CANADA     NYSE       4
## 3             Strong Buy          UK     NYSE       1
## 4          Moderate Sell          UK     NYSE       1
## 5           Moderate Buy      FRANCE     NYSE       3
## 6                   Hold     GERMANY     NYSE       2
## 7          Moderate Sell          US     NYSE       1
## 8           Moderate Buy          US   NASDAQ       2
## 9          Moderate Sell     IRELAND     NYSE       3
## 10                  Hold          US     NYSE       1
## 11                  Hold          UK     NYSE       5
## 12                  Hold          US     AMEX       2
## 13          Moderate Buy          US     NYSE       5
## 14          Moderate Buy          US     NYSE       3
## 15                  Hold          US     NYSE       5
## 16                  Hold SWITZERLAND     NYSE       1
## 17          Moderate Buy          US     NYSE       5
## 18                  Hold          US     NYSE       4
## 19                  Hold          US     NYSE       1
## 20         Moderate Sell          US     NYSE       3
## 21                  Hold          US     NYSE       1
```

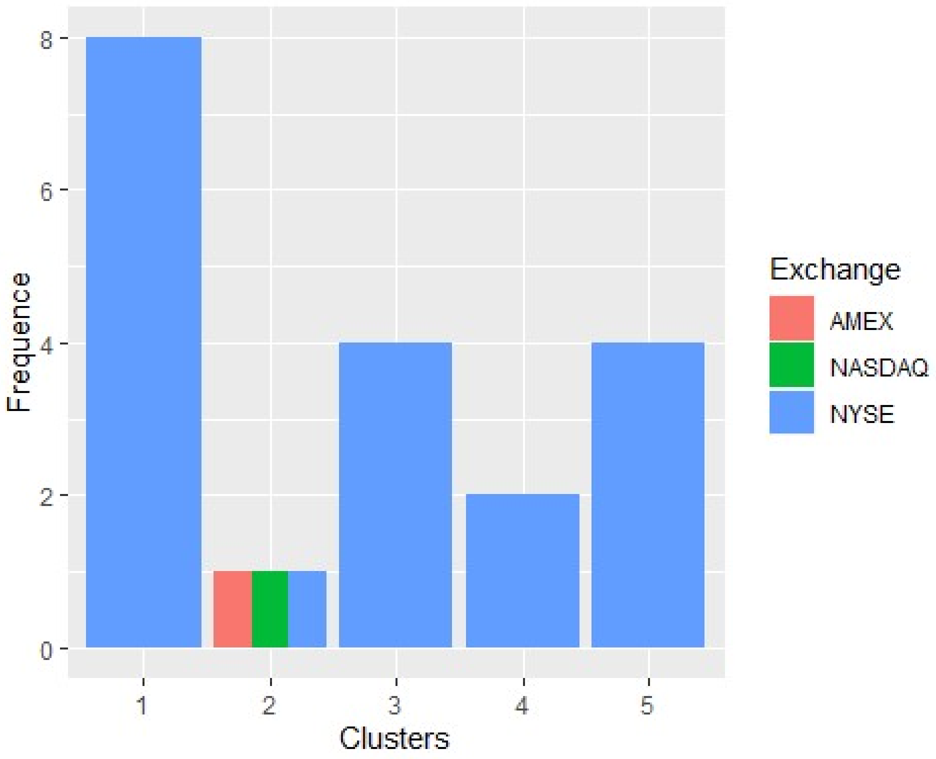
```
Median_Recommenation <- ggplot(pattern, mapping = aes(factor(Cluster),
fill=Median_Recommendation)) + geom_bar(position = 'dodge') +
  labs(x='Clusters', y='Frequence')
Median_Recommenation
```



```
Location <- ggplot(pattern, mapping = aes(factor(Cluster), fill=Location)) +
geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
Location
```



```
Exchange<- ggplot(pattern, mapping = aes(factor(Cluster), fill=Exchange)) +
geom_bar(position = 'dodge') + labs(x='Clusters', y='Frequence')
Exchange
```



```
#cluster-1
#Among the rating options available in cluster 1 are hold, moderate buy,
moderate sell, and strong buy. The median rating for Hold is the highest in
the pattern, indicating that investors usually view the companies in this
cluster as stable and having a moderate potential for growth. and low-risk
investments

#cluster-2
#The median values for hold and moderate buy are different. Compare to
moderate buy their is a  hight frequency rate for the hold and  they have
growth potential but may also have some level of risk

#cluster-3
#Similar median values for moderate buy and sell behavior are found in
Cluster 1, however it has a different count from the other clusters and they
may have some growth potential

#cluster-4
#In comparison to Cluster 4, Cluster 5's median values for hold and moderate
buy behavior are the same, indicating that this occurs the least frequently
```

```
#cluster-5
#The median values for hold and moderate buy behavior are identical in
cluster5 and they may have some potential for growth but also some level of
risk.

#TASK-4
#Provide an appropriate name for each cluster using any or all of the
variables in the dataset

#Cluster 1: Hold-Focused Cluster (due to the high median rating for Hold)

#Cluster 2: Hold-Preferred Cluster (due to the higher frequency rate for Hold
compared to moderate buy)

#Cluster 3: : Moderate Buy/Sell Cluster with Different Count (due to similar
median values for moderate buy and sell behavior but a different count
compared to other clusters)

#Cluster 4: Rare Buy/Hold Cluster (due to the least frequent occurrence of
the same median values for Buy and Hold behaviors, but with similar behavior
to Cluster 4)

#Cluster 5:Balanced Buy/Hold Cluster (due to the identical median values for
Buy and Hold behaviors)
```