Task-4.R

ribakhan

2022-03-20

```
library(readr)
Pharmaceuticals <- read.csv("~/Desktop/Fundamentals of Machine Learning/Pharmaceuticals.csv")
View(Pharmaceuticals)
str(Pharmaceuticals)
```

```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PL
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

```
#installing necessary packages and libraries
install.packages(c("Rcpp","tidyverse"))
```

```
## Error in install.packages : Updating loaded packages
```

```
install.packages("factoextra")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(tidyverse)
library(factoextra)
library(cluster)
library(dplyr)
library(ggplot2)
library(gridExtra)

#collecting numerical values coloumn 1 to 9
row.names(Pharmaceuticals)<- Pharmaceuticals[,1]
P<- Pharmaceuticals[,3:11]
head(P)
```

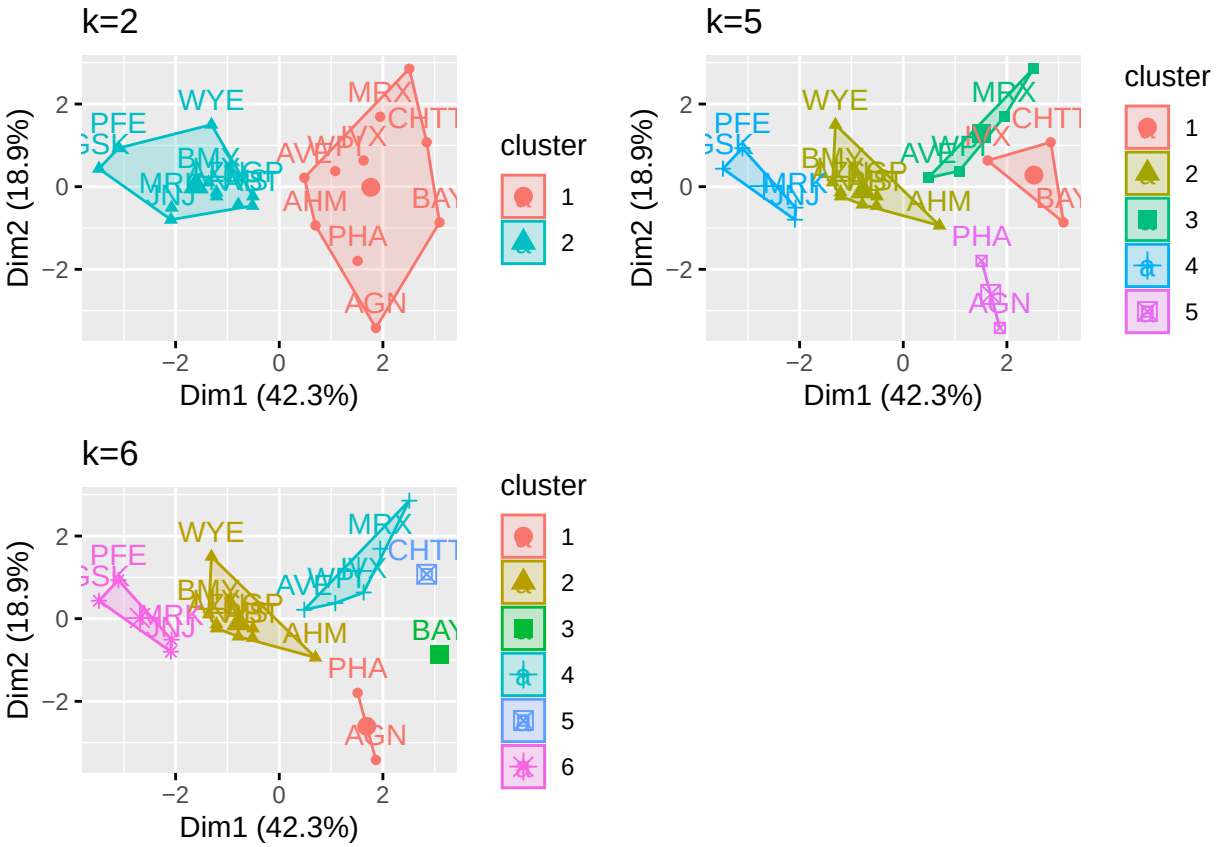
```
##     Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth Net_Profit_Margin
## ABT      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54              16.1
## AGN       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16               5.5
## AHM       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05              11.2
## AZN      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00              18.0
## AVE      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81              12.9
## BAY      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17               2.6
```

```
#Scaling the data using scale function
dataframe<- scale(P)
head(dataframe)
```

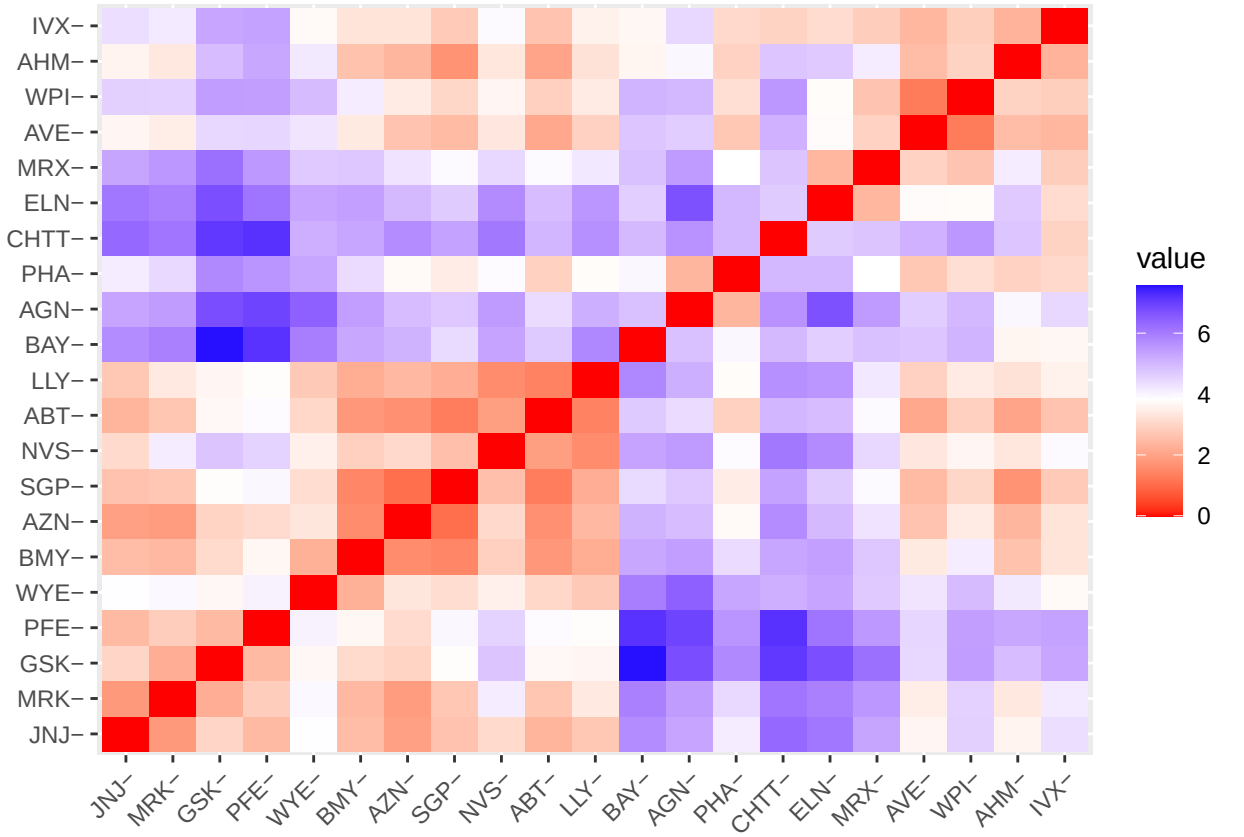
```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover   Leverage Rev_Growth
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000 -0.2120979 -0.5277675
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312  0.0182843 -0.3811391
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312 -0.4040831 -0.5721181
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312 -0.7496565  0.1474473
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656 -0.3144900  1.2163867
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656 -0.7496565 -1.4971443
##     Net_Profit_Margin
## ABT        0.06168225
## AGN       -1.55366706
## AHM       -0.68503583
## AZN        0.35122600
## AVE       -0.42597037
## BAY       -1.99560225
```

```
#Computing K-means clustering
kmeans <- kmeans(dataframe, centers = 2, nstart = 25)
kmeans1 <- kmeans(dataframe, centers = 5, nstart = 25)
kmeans2 <- kmeans(dataframe, centers = 6, nstart = 25)

Plot1<-fviz_cluster(kmeans, data = dataframe)+ggtitle("k=2")
Plot2<-fviz_cluster(kmeans1, data = dataframe)+ggtitle("k=5")
Plot3<-fviz_cluster(kmeans2, data = dataframe)+ggtitle("k=6")
#Plot
grid.arrange(Plot1,Plot2,Plot3, nrow = 2)
```



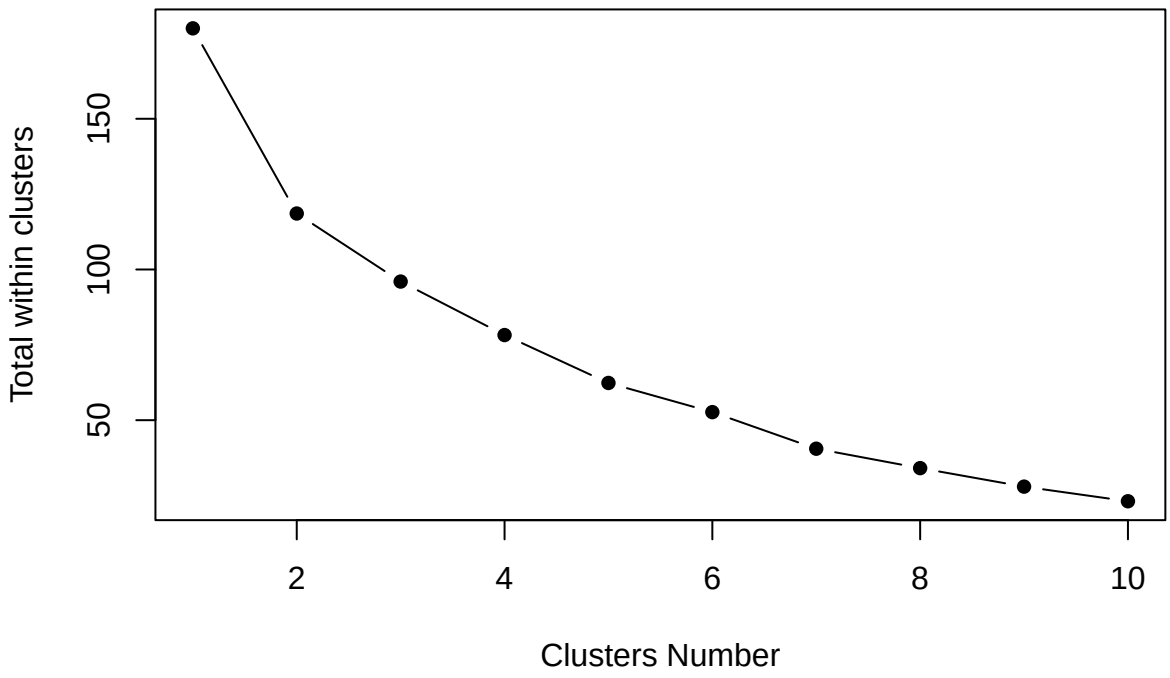

```
#Determing optimal clusters using elbow method
distance<- dist(dataframe, method = "euclidean")
#calculating distance matrix between rows of a data matrix
fviz_dist(distance)
```




```
#For each k calcluate the total within-cluster sum of square
# appropriate number of clusters k=5

set.seed(64060)
wss<- function(k){kmeans(dataframe, k, nstart = 20)$tot.withinss}

#computing value for k= 1 nd k= 10
k.values<- 1:10
wss_clusters<- map_dbl(k.values, wss)
plot(k.values, wss_clusters, type = "b", pch = 16,
     frame = TRUE, xlab = "Clusters Number",
     ylab = "Total within clusters")
```

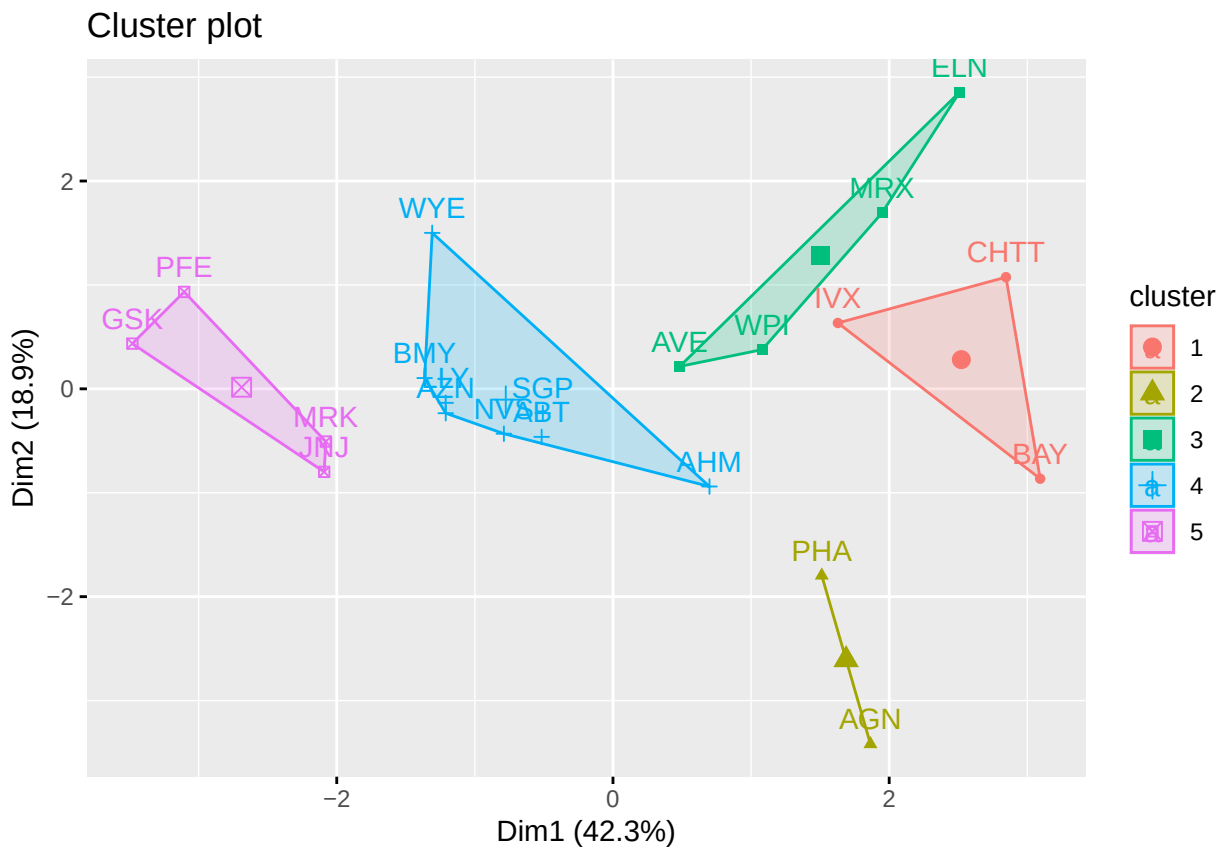



```
#Using 5 clusters to extract the results and visualise
set.seed(64060)
Result<- kmeans(dataframe, 5, nstart = 20)
print(Result)
```

```
## K-means clustering with 5 clusters of sizes 3, 2, 4, 8, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover    Leverage Rev_Growth
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656  1.36644699 -0.6912914
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328 -0.14170336 -0.1168459
## 3 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804  0.06308085  1.5180158
## 4 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746 -0.27449312 -0.7041516
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640 -0.46807818  0.4671788
##   Net_Profit_Margin
## 1      -1.320000179
## 2      -1.416514761
## 3      -0.006893899
## 4       0.556954446
## 5       0.591242521
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS  PFE  PHA  SGP
##    4    2    4    4    3    1    4    1    3    4    5    1    5    3    5    4    5    2    4
##  WPI  WYE
##    3    4
```

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

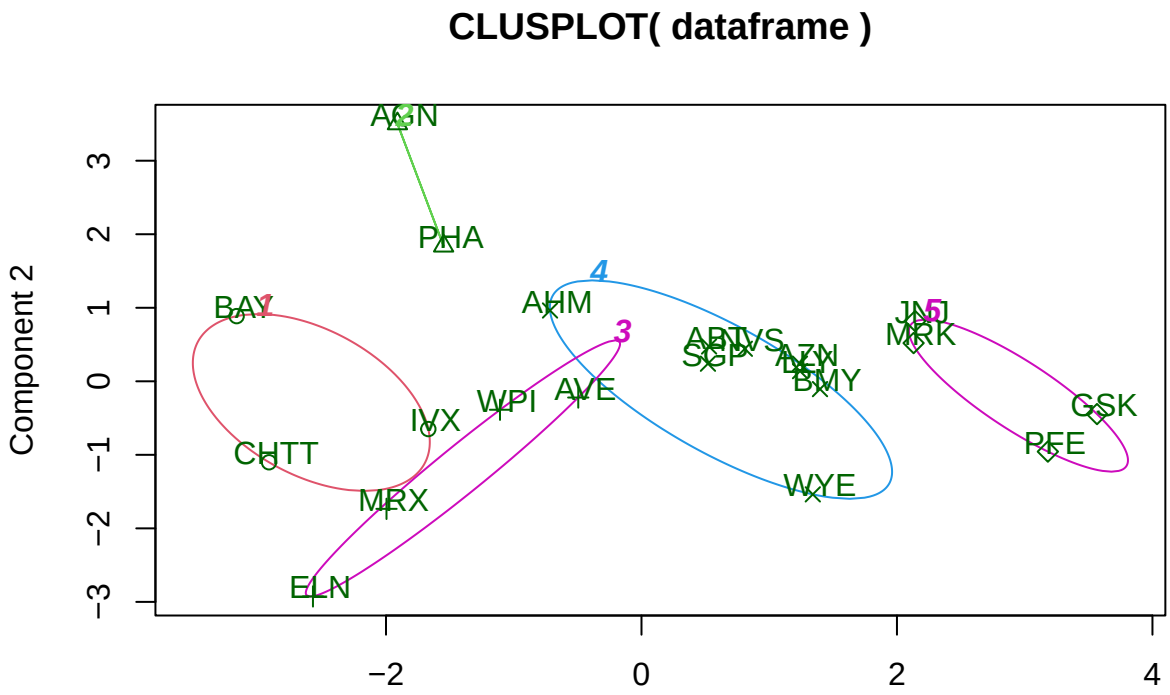
```
fviz_cluster(Result, data= dataframe)
```




```
P%>%
  mutate(Cluster = Result$cluster) %>%
  group_by(Cluster)%>% summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth Net_Profit_Marg~
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>      <dbl>            <dbl>
## 1       1       6.64 0.87      24.6  16.5  4.17           0.6     1.65       5.73             7.03
## 2       2      31.9  0.405     69.5  13.2  5.6            0.75    0.475     12.1              6.4
## 3       3      13.1  0.598     17.7  14.6  6.2            0.425   0.635     30.1             15.6
## 4       4      55.8  0.414     20.3  28.7 12.7            0.738   0.371      5.59            19.4
## 5       5     157.   0.48      22.2  44.4 17.7            0.95    0.22      18.5             19.6
```

```
#Plotting
clusplot(dataframe, Result$cluster, color = TRUE, labels = 2, lines =0)
```




```
#Clusterform
ClusterFormation<- Pharmaceuticals[,c(12,13,14)]%>%
  mutate(clusters = Result$cluster)%>%
  arrange(clusters, ascending = TRUE)
ClusterFormation
```

```
##      Median_Recommendation    Location Exchange clusters
## BAY                   Hold     GERMANY     NYSE        1
## CHTT          Moderate Buy          US   NASDAQ        1
## IVX                   Hold          US     AMEX        1
## AGN           Moderate Buy      CANADA     NYSE        2
## PHA                   Hold          US     NYSE        2
## AVE           Moderate Buy      FRANCE     NYSE        3
## ELN          Moderate Sell     IRELAND     NYSE        3
## MRX           Moderate Buy          US     NYSE        3
## WPI          Moderate Sell          US     NYSE        3
## ABT           Moderate Buy          US     NYSE        4
## AHM             Strong Buy          UK     NYSE        4
## AZN          Moderate Sell          UK     NYSE        4
## BMY          Moderate Sell          US     NYSE        4
## LLY                   Hold          US     NYSE        4
## NVS                   Hold SWITZERLAND     NYSE        4
```

```
## SGP                   Hold          US     NYSE        4
## WYE                   Hold          US     NYSE        4
## GSK                   Hold          UK     NYSE        5
## JNJ           Moderate Buy          US     NYSE        5
## MRK                   Hold          US     NYSE        5
## PFE           Moderate Buy          US     NYSE        5
```

```
#is there any pattern in the clusters with respect to numerical values
p_1<-ggplot(ClusterFormation, mapping = aes(factor(clusters),
                                            fill=Median_Recommendation))+geom_bar(position = 'dodge' )+
  labs( x = 'number of clusters')


p_2<-ggplot(ClusterFormation, mapping = aes(factor(clusters),
                                            fill=Location))+geom_bar(position = 'dodge' )+labs(
                                              x = 'number of clusters')



p_3<-ggplot(ClusterFormation, mapping = aes(factor(clusters),
                                            fill=Exchange))+geom_bar(position = 'dodge' )+labs(
                                              x = 'number of clusters')

grid.arrange(p_1,p_2,p_3)
```