Assignment 4

Anuraag Vasal

11/6/2021

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(factoextra)

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

library(ggplot2)
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()

library(cowplot)
```

Importing Dataset

```
setwd("C:/Users/anura/Desktop/Machine Learning/Assignment 4")
Pharmaceuticals<- read.csv <- read_csv("Pharmaceuticals.csv")

## Rows: 21 Columns: 14

## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## chr (5): Symbol, Name, Median_Recommendation, Location, Exchange
## dbl (9): Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage,
Rev...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

# Summary
summary(Pharmaceuticals)

##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio          ROE            ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.
:0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st
Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median
:0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean
:0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd
Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.
:3.5100
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
##  Min.   :-3.17   Min.   : 2.6      Length:21             Length:21
##  1st Qu.: 6.38   1st Qu.:11.2      Class :character      Class :character
##  Median : 9.37   Median :16.1      Mode  :character      Mode  :character
##  Mean   :13.37   Mean   :15.7
##  3rd Qu.:21.87   3rd Qu.:21.1
##  Max.   :34.21   Max.   :25.5
##    Exchange
##  Length:21
##  Class :character
##  Mode  :character  
##                    
##                    
## 
```

#Data cleaning a)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.

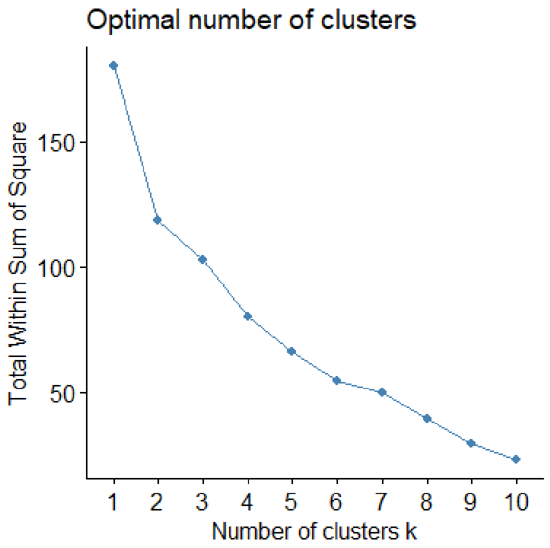
```r
# Checking NULL values in the dataset.
apply(Pharmaceuticals,2,function(x){any(is.na(x))})
```

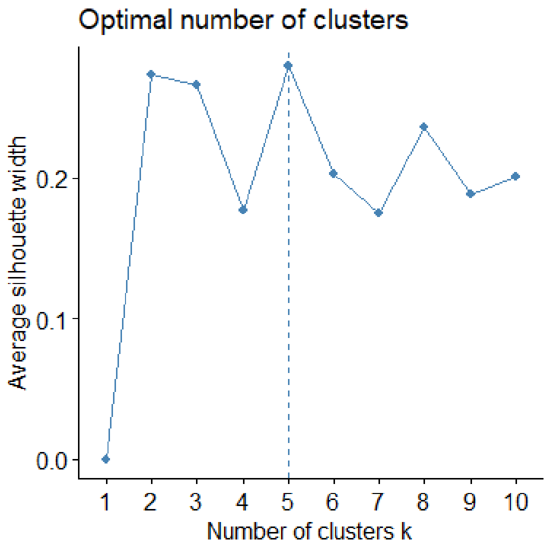
```
##                Symbol                  Name            Market_Cap 
##                 FALSE                 FALSE                 FALSE 
##                  Beta              PE_Ratio                   ROE 
##                 FALSE                 FALSE                 FALSE 
##                   ROA        Asset_Turnover              Leverage 
##                 FALSE                 FALSE                 FALSE 
##            Rev_Growth     Net_Profit_Margin Median_Recommendation 
##                 FALSE                 FALSE                 FALSE 
##              Location              Exchange 
##                 FALSE                 FALSE
```

```r
# Using only the numerical variables (1 to 9) for cluster analysis
Pharmaceuticals_1to9 <- Pharmaceuticals %>% select_if(is.numeric)
# Scaling the data frame (z-score).
set.seed(15)
scale_data <- as.data.frame(scale(Pharmaceuticals_1to9))
```

#Estimating the optimal number of clusters.

```r
wss1 <- fviz_nbclust(scale_data,FUNcluster = kmeans,method = "wss")
sill1 <- fviz_nbclust(scale_data,FUNcluster = kmeans,method = "silhouette")
plot_grid(wss1, sill1)
```





From Elbow method best K is 2 and From Silhouette Method k is 5.

Finding optimal number of clusters.

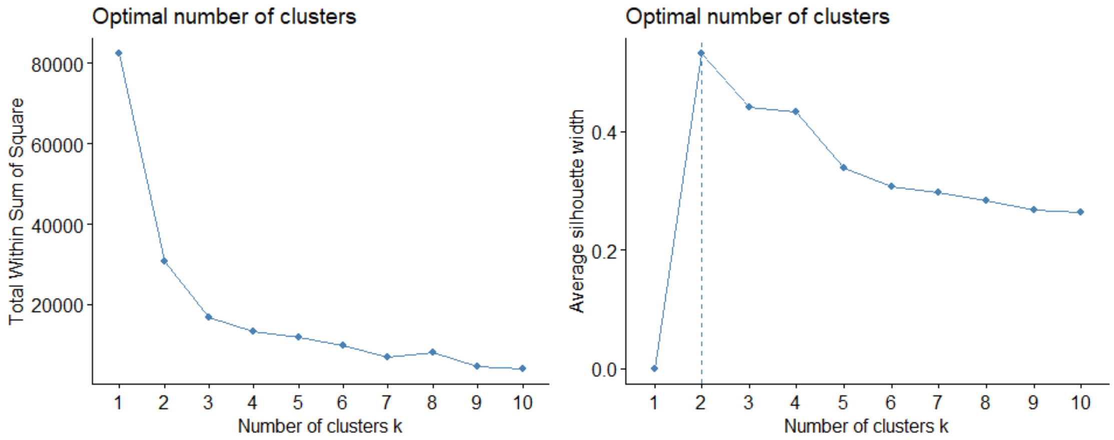
Finding IQR

```
#Finding outliers
outlier_dectecion <- function(x,df = Pharmaceuticals_1to9)
{
  q1 = quantile(df[[x]],0.25) #25th Percentile
  q3 = quantile(df[[x]],0.75) #75th Percentile
  IQR = q3 - q1
  upper_bound = q3 + 1.5 * IQR
  lower_bound = q1 - 1.5 * IQR
  df[(df[x]<lower_bound) | (df[x]>upper_bound),x]
}
out <- vector('list', length(names(Pharmaceuticals_1to9)))
for (i in seq_along(Pharmaceuticals_1to9)){
  x1 <- outlier_dectecion(names(Pharmaceuticals_1to9)[i])
  out[[i]] <- x1
}
names(out) <- names(Pharmaceuticals_1to9)
AfterHandling_outliers <- Pharmaceuticals_1to9 %>%
  filter(Market_Cap != out[[1]], Beta != out[[2]],
         !(PE_Ratio %in% out[[3]]), ROE != out[[4]], !(Leverage %in%
out[[7]]))
```

Estimating the optimal number of clusters

Elbow Method and Silhouette Method

```
wss2 <- fviz_nbclust(AfterHandling_outliers,FUNcluster = kmeans,method =
"wss")
sil2 <- fviz_nbclust(AfterHandling_outliers,FUNcluster = kmeans,method =
"silhouette")
plot_grid(wss2, sil2)
```



#For Model building, Considering scaled data without omitting Outliers K = 2

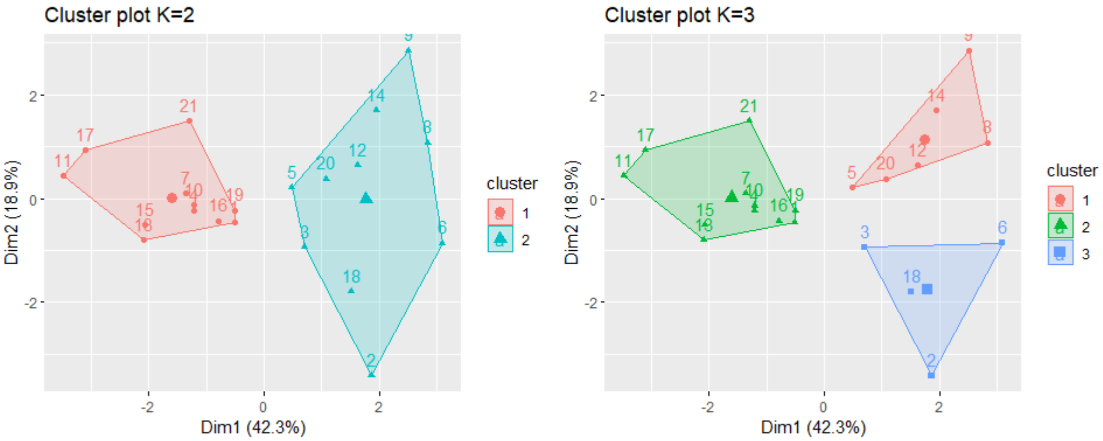
```
model_K2 <- kmeans(scale_data, centers = 2, nstart = 25)
model_K2

## K-means clustering with 2 clusters of sizes 11, 10
## 
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
## 
## Clustering vector:
##  [1] 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1 1 2 1 2 1
## 
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

K = 3

```
model_K3 <- kmeans(scale_data, centers = 3, nstart = 25)
model_K3

## K-means clustering with 3 clusters of sizes 6, 11, 4
## 
```

```
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.8261772  0.4775991 -0.3696184 -0.5631589 -0.8514589     -0.9994088
## 2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 3 -0.6125361  0.2698666  1.3143935 -0.9609057 -1.0174553      0.2306328
##     Leverage Rev_Growth Net_Profit_Margin
## 1  0.8502201  0.9158889        -0.3319956
## 2 -0.3331068 -0.2902163         0.6823310
## 3 -0.3592866 -0.5757385        -1.3784169
##
## Clustering vector:
##  [1] 2 3 3 2 1 3 2 1 1 2 2 1 2 1 2 2 2 3 2 1 2
##
## Within cluster sum of squares by cluster:
## [1] 32.14336 43.30886 20.54199
##  (between_SS / total_SS =  46.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

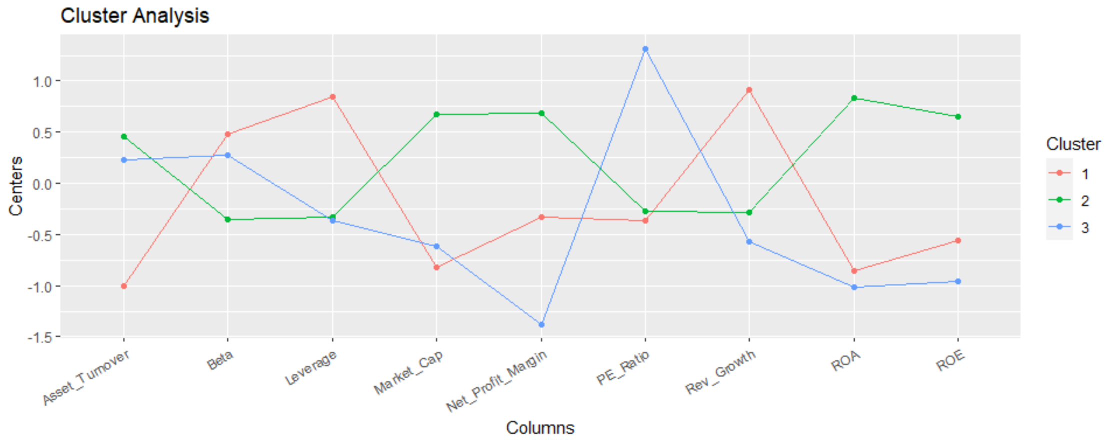
```
K_2 <- fviz_cluster(model_K2,data = scale_data, main = 'Cluster plot K=2')
K_3 <- fviz_cluster(model_K3,data = scale_data, main = 'Cluster plot K=3')
plot_grid(K_2, K_3)
```



#Considering all the observations, 3 looks like an optimal k.

#Q2: Cluster analysis

```
clusters_centers <- data.frame(model_K3$centers) %>%
  rowid_to_column() %>%
  gather('Columns', 'Centers', -1)
ggplot(clusters_centers, aes(x = Columns, y = Centers, color =
as.factor(rowid))) +
  geom_line(aes(group = as.factor(rowid))) + geom_point() +
  labs(color = "Cluster", title = 'Cluster Analysis') +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1))
```



From the above graph we can infer that all cluster patterns are different;

1)Red: Companies have good Asset turnover and beta, But it's leverage, market cap, Net profit margin, Revenue Growth, ROA and ROE are low but it has good PE Ratio.

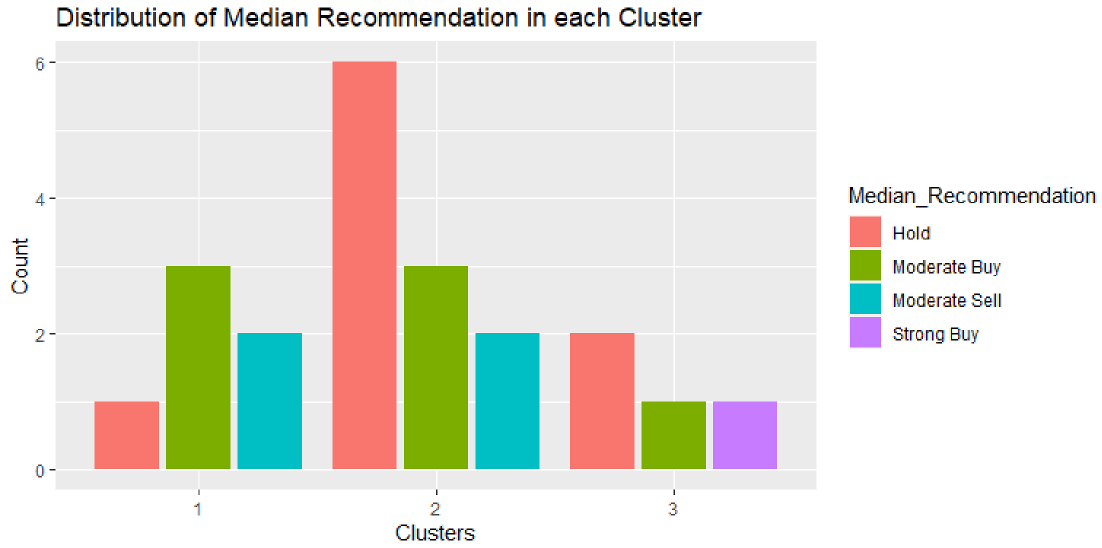
2)Green: Companies have low asset value(Asset turnover, ROA, ROE), But good revenue growth, beta and leverage.

3)Blue: Companies have good Asset value(Asset turnover, ROA, ROE) and market cap, But lacks in Beta, Leverage, PE Ratio and Revenue Growth.

#Q3: Applying above Cluster patterns

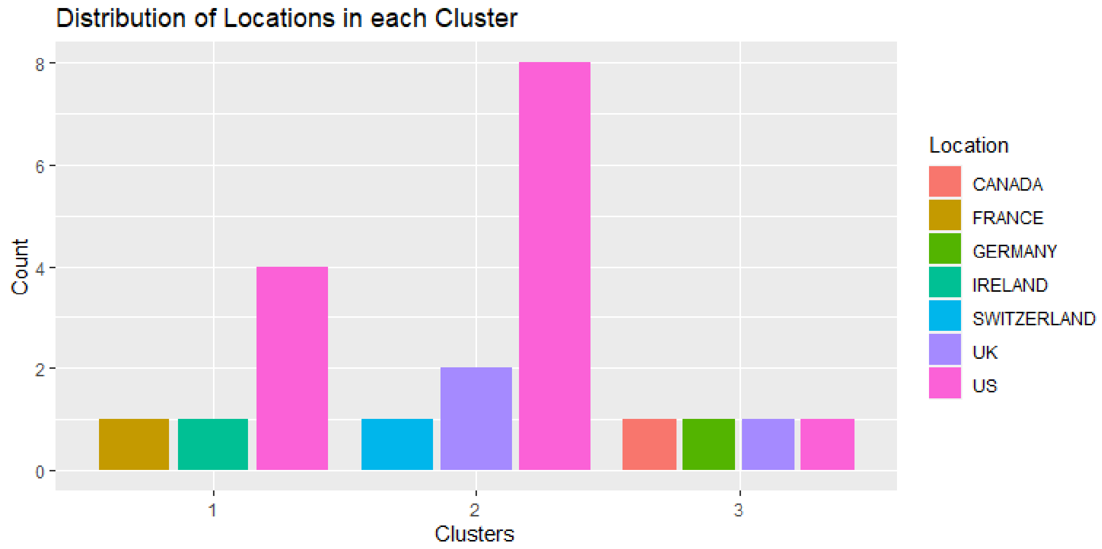
```
(Pharmaceuticals_10to12 <-  Pharmaceuticals %>%
select(c("Median_Recommendation","Location","Exchange")) %>%
  mutate(cluster_pattern = model_K3$cluster) %>%
arrange(desc(cluster_pattern)))

## # A tibble: 21 x 4
##    Median_Recommendation Location Exchange cluster_pattern
##    <chr>                 <chr>    <chr>              <int>
##  1 Moderate Buy          CANADA   NYSE                   3
##  2 Strong Buy            UK       NYSE                   3
##  3 Hold                  GERMANY  NYSE                   3
##  4 Hold                  US       NYSE                   3
##  5 Moderate Buy          US       NYSE                   2
##  6 Moderate Sell         UK       NYSE                   2
##  7 Moderate Sell         US       NYSE                   2
##  8 Hold                  US       NYSE                   2
##  9 Hold                  UK       NYSE                   2
## 10 Moderate Buy          US       NYSE                   2
## # ... with 11 more rows

ggplot(Pharmaceuticals_10to12, aes(fill = Median_Recommendation,
x = as.factor(cluster_pattern))) +
```

```r
  geom_bar(position = 'dodge2') +
labs(x="Clusters", y="Count",
     title = "Distribution of Median Recommendation in each Cluster")
```

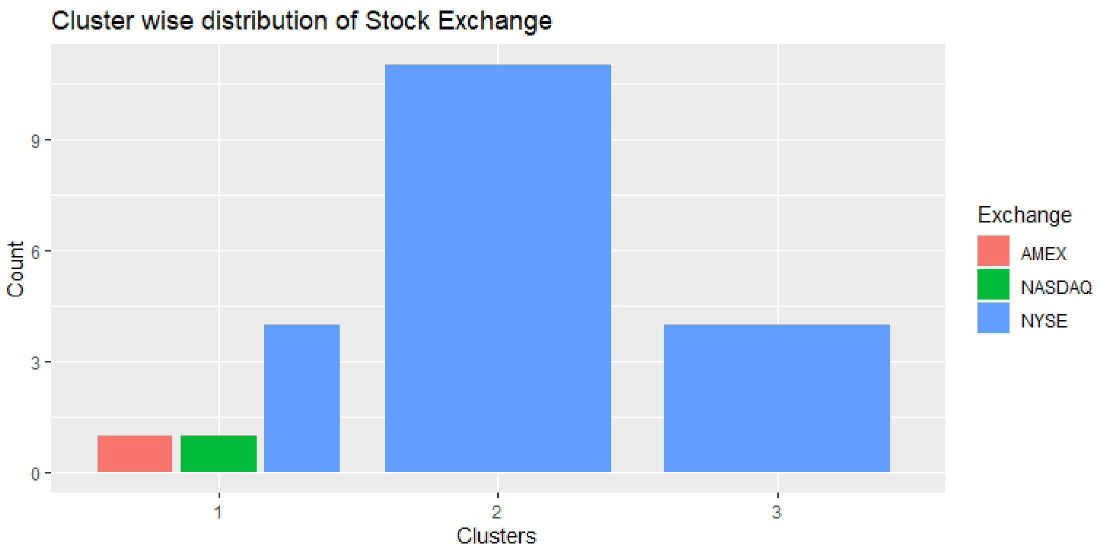


From the above graph we can infer that Cluster1 has moderate buy and sell ratio option which is unique from other clusters and Cluster2 has High Hold and sell ratio Cluster3 has good hold option.

```r
ggplot(Pharmaceuticals_10to12, aes(fill = Location,
x = as.factor(cluster_pattern))) +
  geom_bar(position = 'dodge2') +
labs(x="Clusters", y="Count",
     title = "Distribution of Locations in each Cluster")
```



```r
ggplot(Pharmaceuticals_10to12, aes(fill = Exchange,
x = as.factor(cluster_pattern))) +
  geom_bar(position = 'dodge2') +
labs(x="Clusters", y="Count",
     title = "Cluster wise distribution of Stock Exchange")
```



#Q4: Providing an appropriate name for each cluster.

#Small Cap: High PE Ratio.

#Mid Cap: Fast growing with less Market capital and assets.

#Large Cap: High Assets and Market capital.