Assignment 4

Harish Kumar uddandi

R Markdown

```
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ISLR)
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.2.2

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

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.2

## ── Attaching packages
## ───────────────────────────────────────
## tidyverse 1.3.2 ──

## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ✔ purrr   0.3.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.2.2

## Loading required package: grid
## Loading required package: modeltools
## Loading required package: stats4

set.seed(64060)
getwd()

## [1] "C:/FALL/ML"
```

```
setwd("C:/FALL/ML")

KMC <- read.csv("Pharmaceuticals.csv")
head(KMC)

##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turn
over
## 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 Exc
hange
## 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
```

a. 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.

```
# Columns 1 - 9 for 21 firms
ColumnNums <- KMC [,3:11] # Considering column 3-11 i.e numerical variables
head(ColumnNums)

##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## 2       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 5      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 6      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##   Net_Profit_Margin
## 1                 16.1
## 2                  5.5
## 3                 11.2
## 4                 18.0
## 5                 12.9
## 6                  2.6
```

```
ColumnNums <- scale(ColumnNums)
 summary(ColumnNums)
```

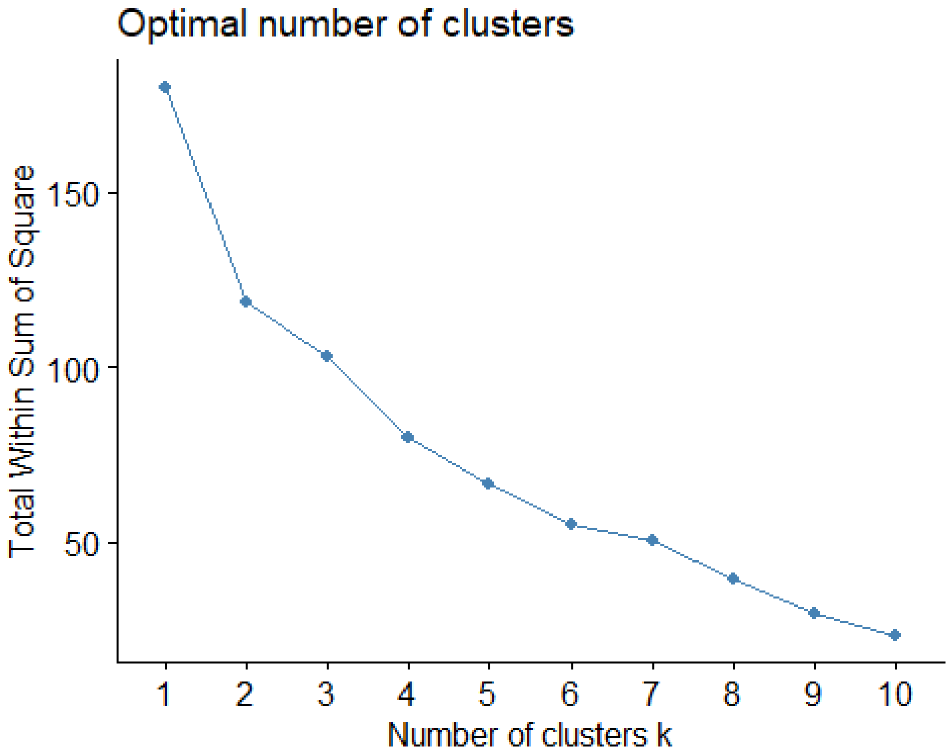
```
##    Market_Cap           Beta              PE_Ratio             ROE         
##  Min.   :-0.9768   Min.   :-1.3466   Min.   :-1.3404   Min.   :-1.4515  
##  1st Qu.:-0.8763   1st Qu.:-0.6844   1st Qu.:-0.4023   1st Qu.:-0.7223  
##  Median :-0.1614   Median :-0.2560   Median :-0.2429   Median :-0.2118  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.2762   3rd Qu.: 0.4841   3rd Qu.: 0.1495   3rd Qu.: 0.3450  
##  Max.   : 2.4200   Max.   : 2.2758   Max.   : 3.4971   Max.   : 2.4597  
##       ROA          Asset_Turnover      Leverage          Rev_Growth     
##  Min.   :-1.7128   Min.   :-1.8451   Min.   :-0.74966   Min.   :-1.4971  
##  1st Qu.:-0.9047   1st Qu.:-0.4613   1st Qu.:-0.54487   1st Qu.:-0.6328  
##  Median : 0.1289   Median :-0.4613   Median :-0.31449   Median :-0.3621  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000  
##  3rd Qu.: 0.8430   3rd Qu.: 0.9225   3rd Qu.: 0.01828   3rd Qu.: 0.7693  
##  Max.   : 1.8389   Max.   : 1.8451   Max.   : 3.74280   Max.   : 1.8862  
##  Net_Profit_Margin 
##  Min.   :-1.99560  
##  1st Qu.:-0.68504  
##  Median : 0.06168  
##  Mean   : 0.00000  
##  3rd Qu.: 0.82364  
##  Max.   : 1.49416
```

```
 #The distance between each data point and the centroid is calculated using t
he Eucledian distance.
Distance_ColumnNums <-get_dist(ColumnNums, method = "euclidean", stand = FALS
E)
Distance_ColumnNums
```

```
##           1        2        3        4        5        6        7        8
## 2  4.415575                                                               
## 3  2.018793 3.945745                                                      
## 4  1.669541 4.909566 2.364249                                             
## 5  2.111983 4.642699 2.487172 2.632282                                    
## 6  4.690231 4.853901 3.636353 5.065563 4.764654                           
## 7  1.805543 5.419487 2.600986 1.572582 3.400602 5.273023                  
## 8  5.020726 5.612226 4.760341 5.719174 5.096246 4.969438 5.287400         
## 9  4.901141 6.695261 4.695844 4.974521 3.748778 4.608660 5.378092 4.675606
## 10 1.422680 5.140253 3.238353 2.405951 2.910766 5.804419 2.189107 5.657801
## 11 3.689906 6.747789 4.904614 2.957494 4.476690 7.546154 3.099023 7.080175
## 12 2.624729 4.470028 2.316548 3.282195 2.386850 3.658011 3.279927 2.951511
## 13 2.333874 5.317942 3.593764 1.958326 3.640773 5.724303 2.511309 6.310233
## 14 3.920297 5.479080 4.120549 4.269231 2.927258 4.848442 4.734766 4.786213
## 15 2.680733 5.443918 3.361981 1.859280 3.472410 5.918477 2.432281 6.101541
## 16 1.922731 5.468844 3.331743 3.056196 3.330879 5.331004 2.866126 6.063738
## 17 3.887235 6.906828 5.268858 3.109413 4.495242 7.163993 3.666674 7.180257
## 18 2.908982 2.367912 2.925627 3.715808 2.718441 3.955926 4.408645 5.000709
## 19 1.312599 4.725384 1.704709 1.080519 2.464855 4.426418 1.478433 5.346513
## 20 2.882610 5.007086 2.943946 3.414127 1.296549 5.055769 4.116074 5.540296
## 21 3.038549 6.446458 4.185594 3.324966 4.254562 5.954379 2.269808 5.127981
##           9       10       11       12       13       14       15       16
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10 5.554227
## 11 6.731204 3.631174
## 12 3.115283 3.537378 5.276601
## 13 6.070533 2.722434 2.988672 4.354581
## 14 2.389723 4.191466 6.187185 2.825394 5.306512
## 15 5.921987 3.380695 2.218040 4.164267 1.814184 5.532520
## 16 5.732322 1.577953 4.783039 3.899915 3.083678 4.478040 4.112418
## 17 6.123133 3.783136 2.447177 5.356598 2.447341 5.518379 2.831329 4.536250
## 18 5.007721 3.754900 5.773960 3.073579 4.112432 3.827019 4.448933 3.884035
## 19 4.665611 2.205815 3.780283 2.763476 2.604437 3.907501 2.710607 2.542763
## 20 3.756437 3.412378 5.437193 2.857109 4.591764 2.653341 4.569336 3.626404
## 21 5.312455 2.747839 3.670720 3.719962 3.858028 4.709401 3.935039 3.525940
##          17       18       19       20
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18 5.587119
## 19 3.955078 3.449579
## 20 5.403128 3.172178 3.026610
## 21 4.026095 5.286507 3.145472 4.922945
```

#Elbow and Silhouette methods are used to find the optimal number of clusters. #Elbow Method

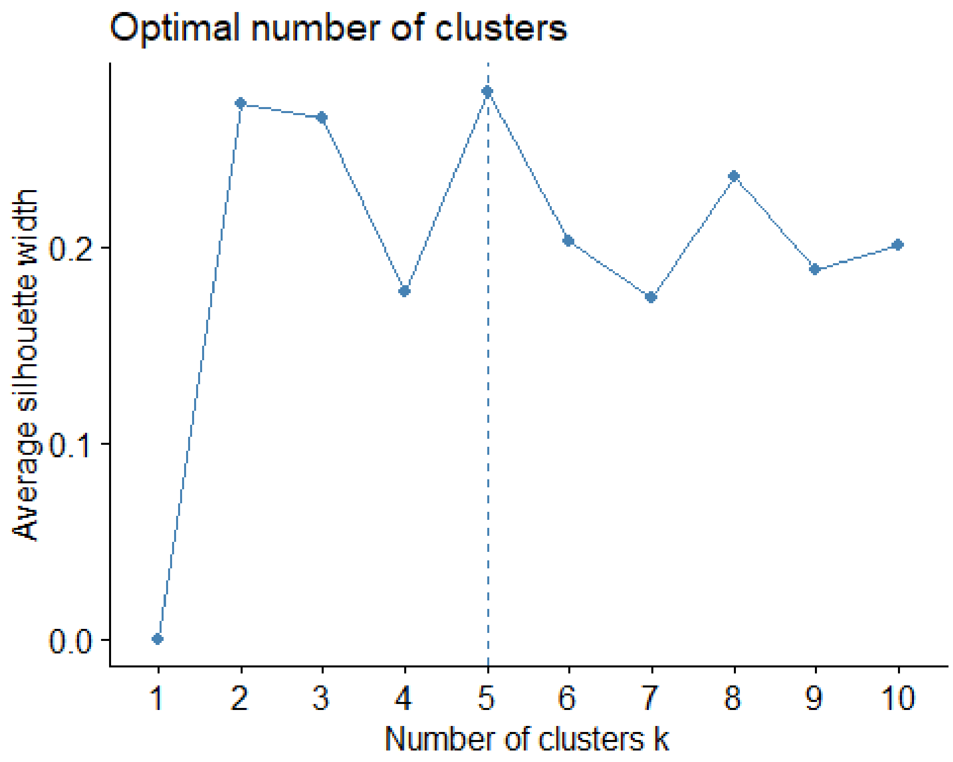
```
library(factoextra) # clustering algorithms & visualization
library(flexclust)
fviz_nbclust(ColumnNums,kmeans,method="wss")
```



#the plot shows a clear elbow is at k = 2. Also as the above graph is not clear as it did not show any sharp point at 2. We can use 3 or 4 or 5 as the 'K' value too.

#Silhouttes method

```
#Silhouttes method
fviz_nbclust(ColumnNums,kmeans,method="silhouette")
```



#The optimal clusters were determined as 2 using the elbow approach, but when we utilized the Silhouettes method, we obtained a value of 5. We will use the silhouettes approach in this case because the elbow method was unclear in identifying the optimal cluster. #We have determined how many clusters there are. We will now use the K-means method.

```
#Applying K-means Algorithm
KMean_chk <- kmeans(ColumnNums, centers = 5, nstart = 25) #Number of restarts
= 25
KMean_chk

## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 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 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
## 
```

```
## Clustering vector:
##  [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1
## 
## Within cluster sum of squares by cluster:
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withi
nss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

#Centers

```
KMean_chk$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 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
```

#Size

```
KMean_chk$size
```

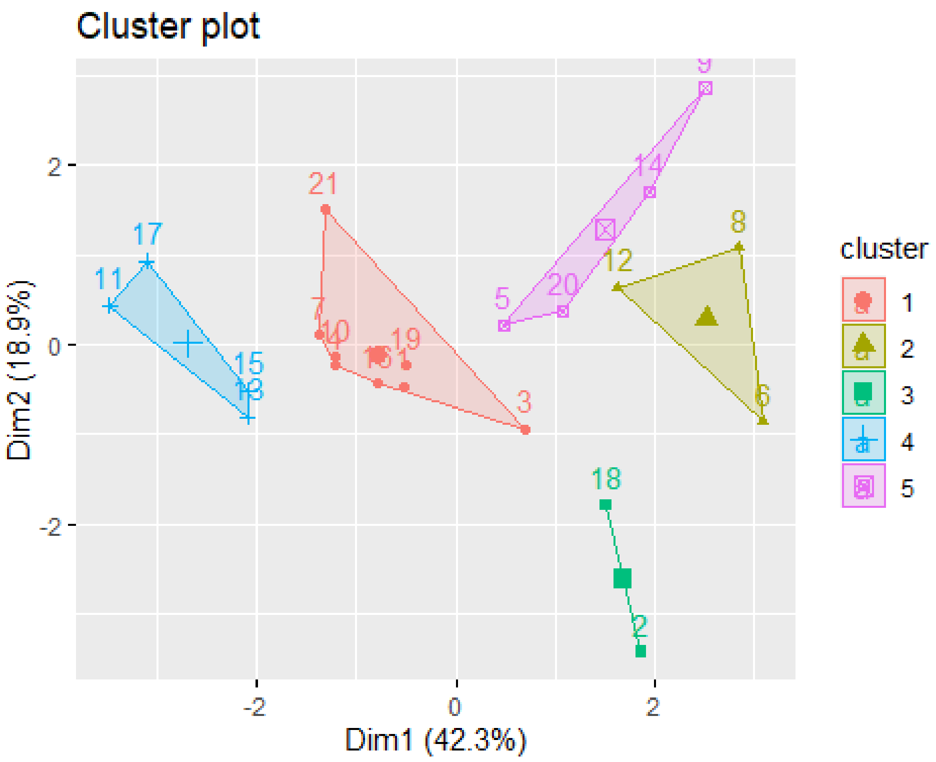
```
## [1] 8 3 2 4 4
```

#Cluster

```
KMean_chk$cluster[c(1:21)]
```

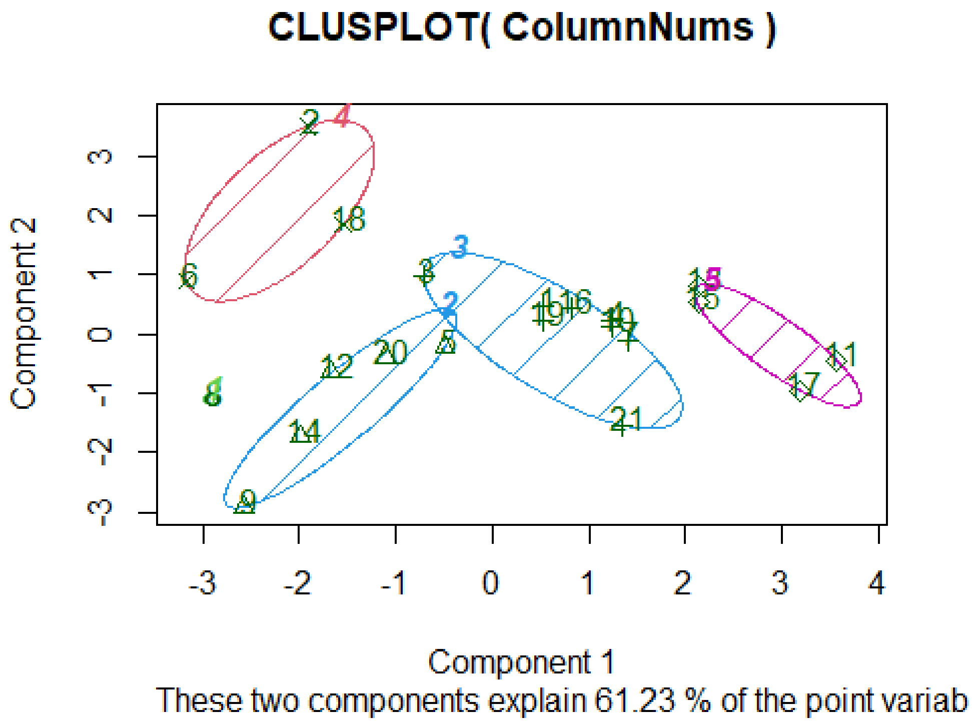
```
##  [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1
```

```
fviz_cluster(KMean_chk, data = ColumnNums)
```



From the above, 5 clusters have been identified. The symbols/shapes in each cluster are ‘centroids’ of that specific cluster.No other centroid can be considered until new data is added, due to the criteria of Nstart value 25 and higher.

```
library(cluster)
Cluster_Plot <- kmeans(ColumnNums,5)
clusplot(ColumnNums, Cluster_Plot$cluster, color=TRUE, shade=TRUE, labels=2,
lines=0)
```



(b) Interpret the clusters with respect to the numerical variables used in forming the clusters.

#In Excel, rows begin with 2. The rows have therefore been discussed starting with row one for our convenience. (Row 2 in this case)

First Cluster_Red = Rows are 2, 6, 18

Second Cluster_Green = Rows are 1,4,7,10,16,19,21

Third Cluster_Blue = Rows are 8,9,12,14

Fourth Cluster_Pink = Rows are 3,5,20

Fifth Cluster_Pink(last) = Rows are 11,13,15,17

We calculate the mean of all the numerical variables.

```
aggregate(ColumnNums,by=list(Cluster_Plot$cluster),FUN=mean)

##   Group.1  Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918
## 2       2 -0.79605926  0.3205014 -0.45014035 -0.6533148 -0.7881923
## 3       3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 4       4 -0.52462814  0.4451409  1.84984387 -1.0404550 -1.1865838
## 5       5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1  -4.612656e-01  3.7427970 -0.6327607        -1.2488842
## 2  -1.107037e+00  0.2717048  1.2256188        -0.1486179
## 3   1.729746e-01 -0.2744931 -0.7041516         0.5569544
## 4   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   1.153164e+00 -0.4680782  0.4671788         0.5912425
```

```
ColumnNums1 <- data.frame(ColumnNums, Cluster_Plot$cluster)
```

First Cluster = has Highest PE_Ratio and lowest Net_Profit_Margin, ROA

Second Cluster = has Highest Net_Profit_Margin and Lowest Rev_Growth, Beta

Third Cluster = has Highest Leverage, Beta and Lowest ROA

Fourth Cluster = has Highest Rev_Growth and Lowest Beta, ROE Market_Cap

Fifth Cluster = has Highest Market_Cap, ROA, ROE and Lowest Leverage

(c) Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

In First Cluster,There is a high PE Ratio and a low Net Profit Margin and ROA. The Median Recommendation for this cluster is “Moderate Buy” for all the points.

In Second Cluster, Low Rev Growth, Beta and high Net Profit Margin are present. The Median Recommendation is usually advised to be set on “Hold” for the majority of the points for this cluster.

In Third Cluster,High Leverage, Beta, and Low ROA are present. The Median Recommendation for this cluster primarily supports a Moderate Buy.

In Fourth Cluster ,High Rev Growth and Lowest Beta, together with ROE Market Cap are present . The Median suggestion indicates equal Strong Buy, Moderate Buy, and Moderate Sell recommendations for this cluster.

In Fifth Cluster, High Market Cap, Lowest Leverage, and High ROA and ROE present. Both Hold and Moderate Buy recommendations are included in the Median Recommendation for this cluster.

(d) Provide an appropriate name for each cluster using any or all of the variables in the dataset.

First Cluster- Low Net_Profit_Margin and ROA cluster or Moderate Buy Cluster

Second Cluster- Low Rev_Growth, Beta cluster or Hold Cluster

Third Cluster- High Leverage, Beta cluster or ’Moderate Cluster

Fourth Cluster- High Rev_Growth and Lowest Beta, ROE Market_Cap Cluster

Fifth Cluster- High Market_Cap, ROA, ROE and Lowest Leverage Cluster