Assignment 4

Ananth Kumar

27/10/2021

.libPaths("C:\Users\Ananth\OneDrive\Desktop\MSBA Kent\Fall 2021\Fundamentals of Machine Learning\Assignment\Ass 2")

```
library(factoextra) # clustering algorithms & visualization
library(ISLR)
library(tidyverse)  # data manipulation
library(caret)
library(flexclust)

set.seed(1234)

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

```
##   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
```

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.

```
#Considering only columns from 1 to 9 for 21 firms
Numeric<- KMC[,3:11] # numerical columns starts from 3 to 11 in the excel data.
head(Numeric) # descriptive data has been removed or we can say a data subset was done for the numerica
```

```
##   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
```

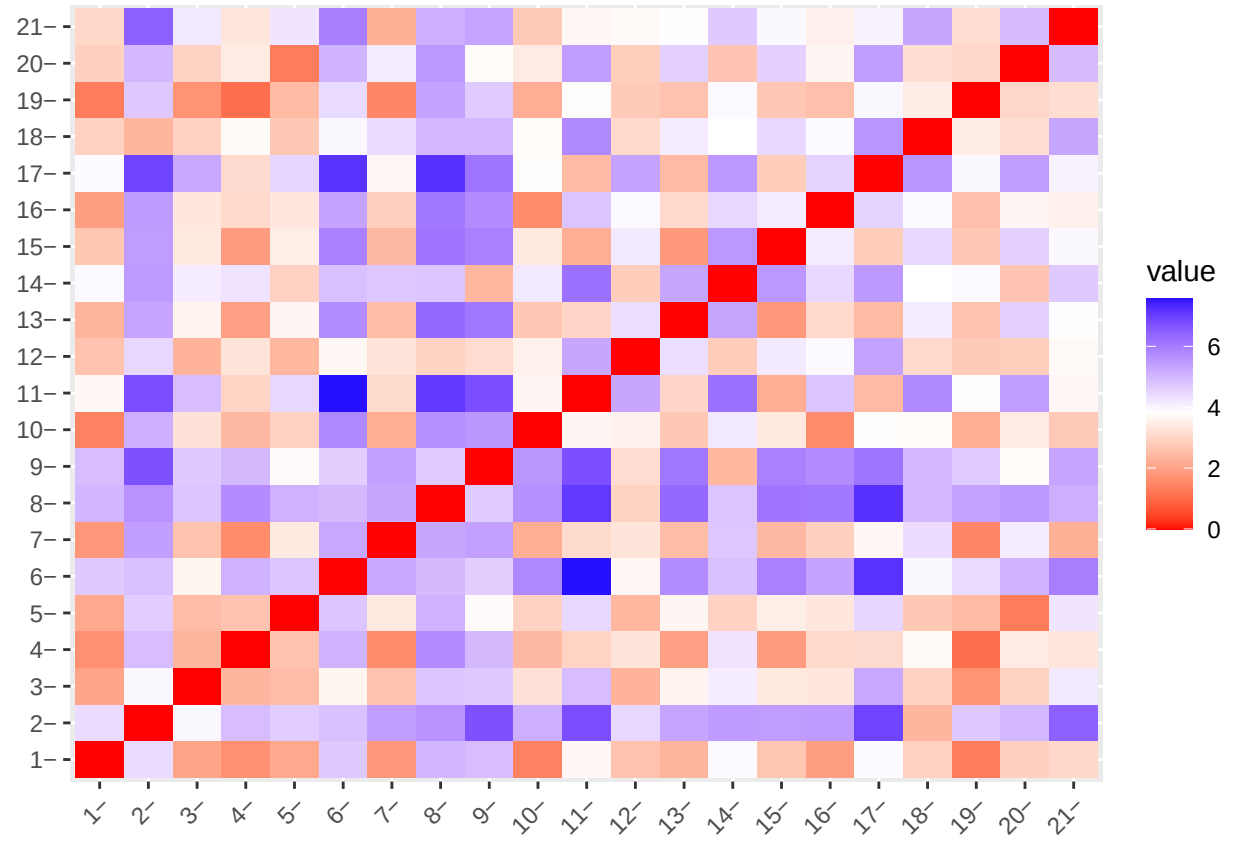
```r
library(factoextra) # clustering algorithms & visualization
library(flexclust)


#Normalizing the dataframe with range and scale method , We apply scale to represents a good measure of

Numeric <- scale(Numeric)

distance_Numeric <-get_dist(Numeric, method = "euclidean", stand = FALSE)  # Euclidean distance is used


# we can see the distance between each observation
fviz_dist(
  distance_Numeric,
  order = FALSE,    # order is set to false so that the x axis and y axis values are sorted.
  show_labels = TRUE,
  lab_size = NULL,
  gradient = list(low = "red", mid = "white", high = "blue") # coloring based on the values of the obse
)
```

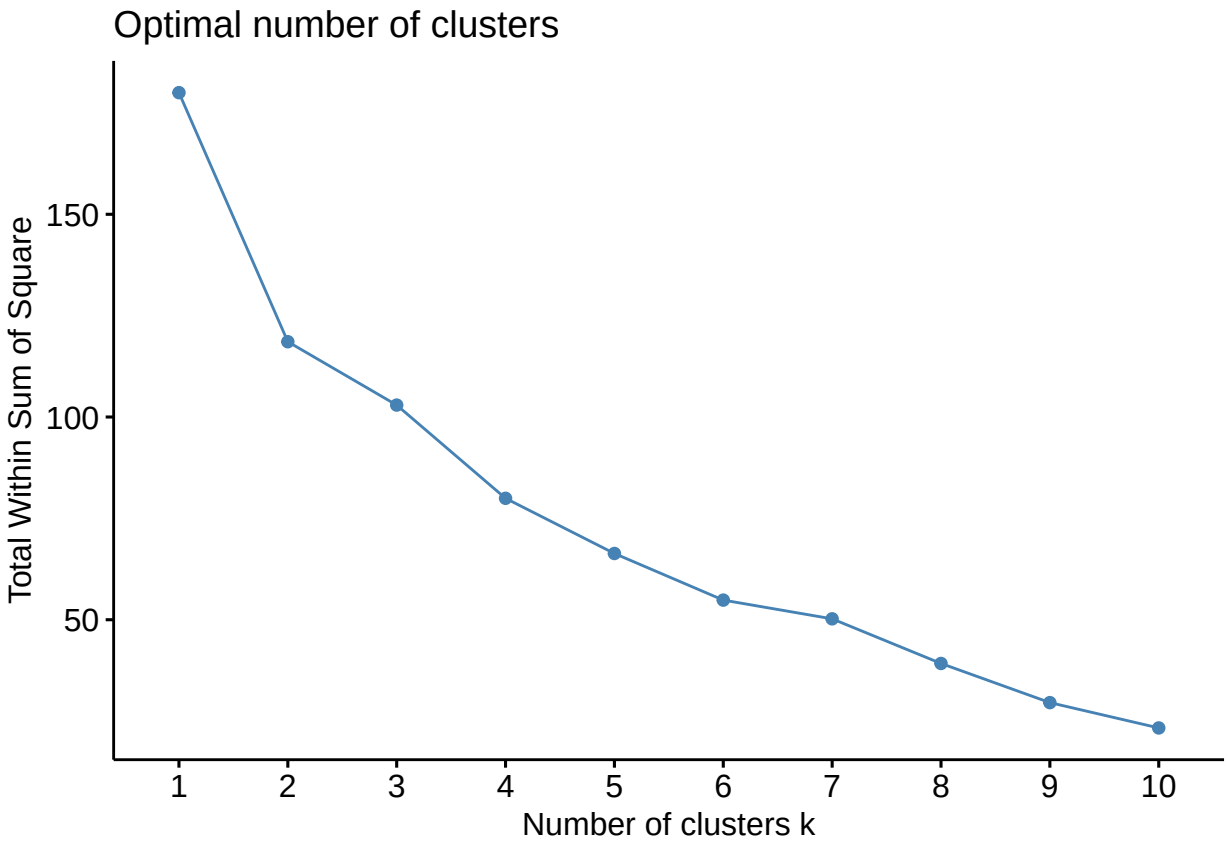


Using Within-Cluster-Sum of Squared Errors.

```
library(factoextra) # clustering algorithms & visualization
library(flexclust)

#elbow1 <- scale(Numeric)

fviz_nbclust(Numeric,kmeans,method="wss")
```

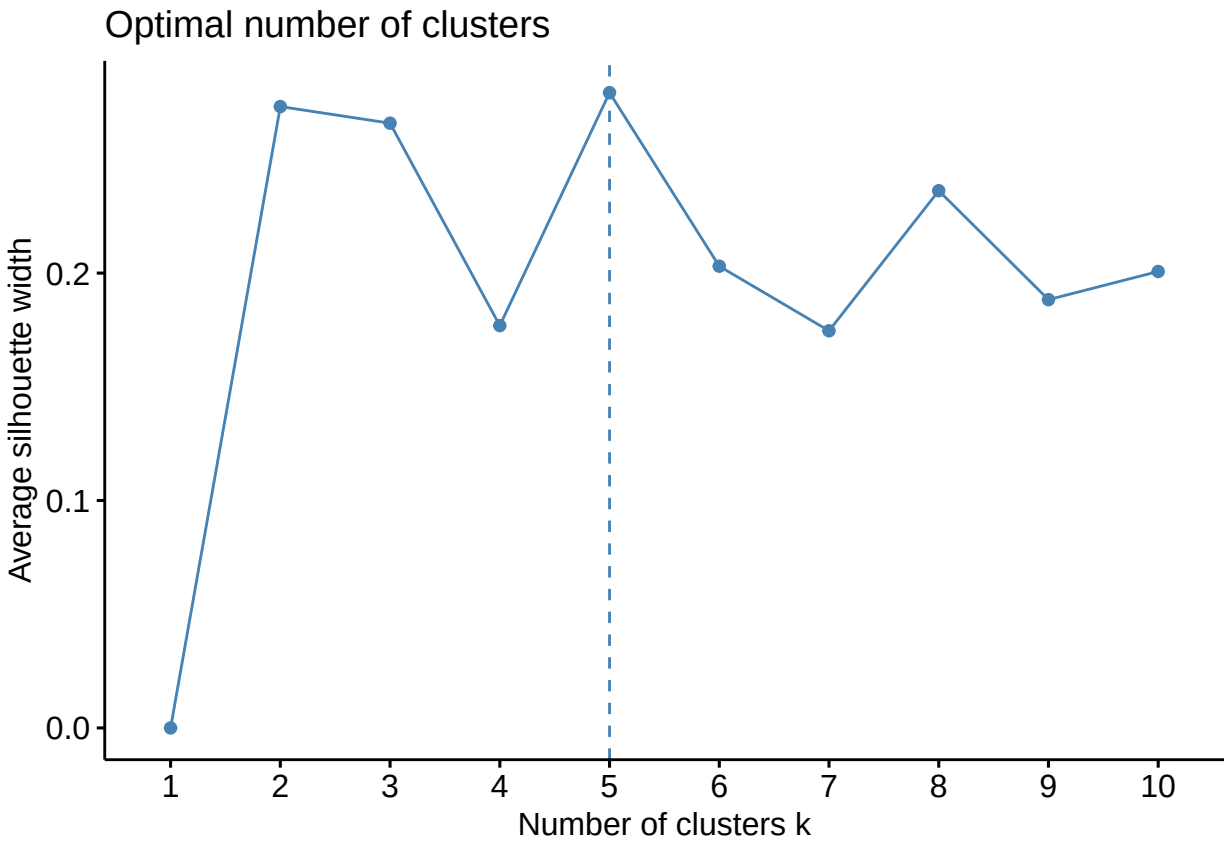


the plot looks like an arm with a clear elbow at k = 2, Howefver this choice ambigous we could either choose 2,3,4,5 and the graph is not sharp and clear.

```
library(flexclust)

# By using silhouette method, we can observe that

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



In the above graph generated by silhouette method, we can seea clear peak at k = 5 and this is clear and sharp and even highlited by the r studio. Hence considering silhouette method.

```
#Applying kmean

k5 <- kmeans(Numeric, centers = 5, nstart = 25) # k/centers  = 5, number of restarts = 25 , cluster mea
k5
```

```
## 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.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
k5$centers # Centers for each cluster for each and every columns
```

```
##    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
```

```
k5$size # Number of  observation in each cluster
```

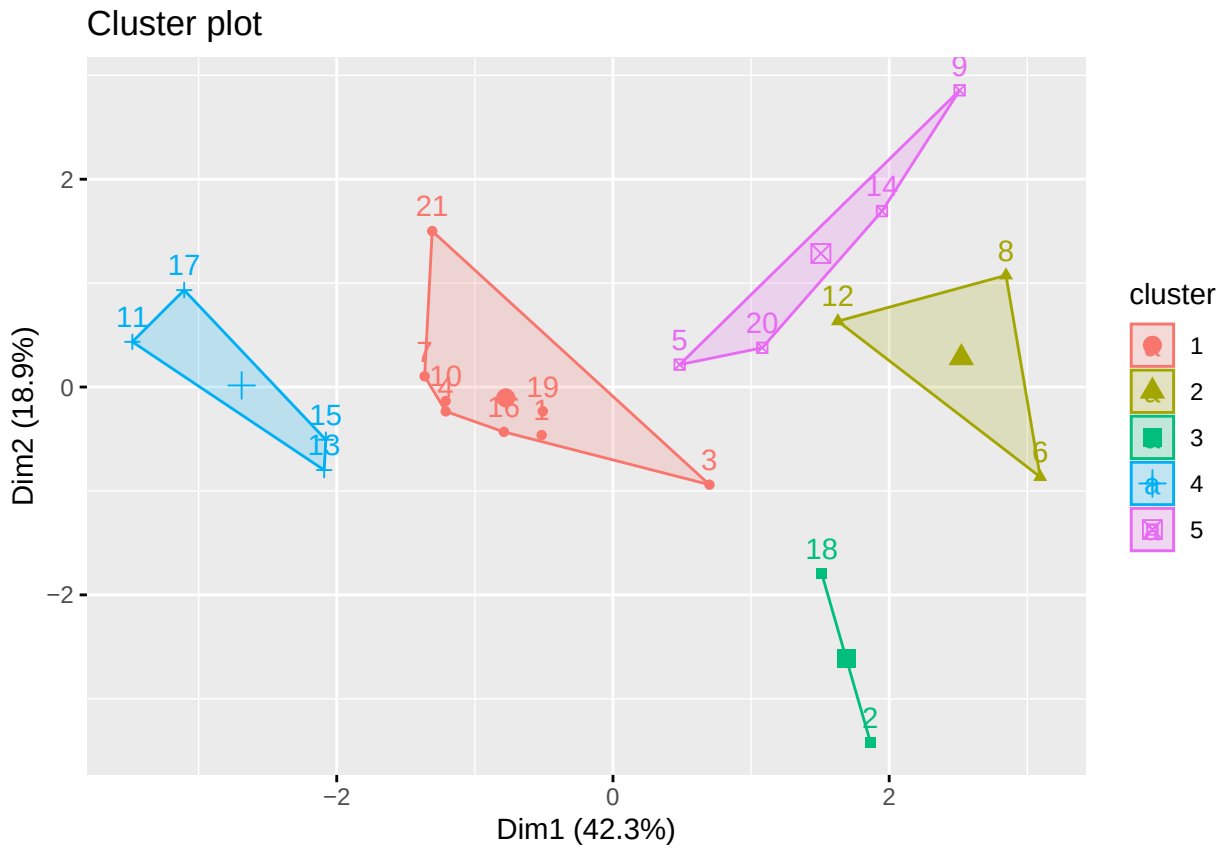
```
## [1] 8 3 2 4 4
```

```
k5$cluster[c(21,20,19)] # 19,20,21 observations and their respective cluster lables. 21th observation h
```

```
## [1] 1 5 1
```

K-means clustering with 5 clusters of sizes 8, 4, 4, 2, 3

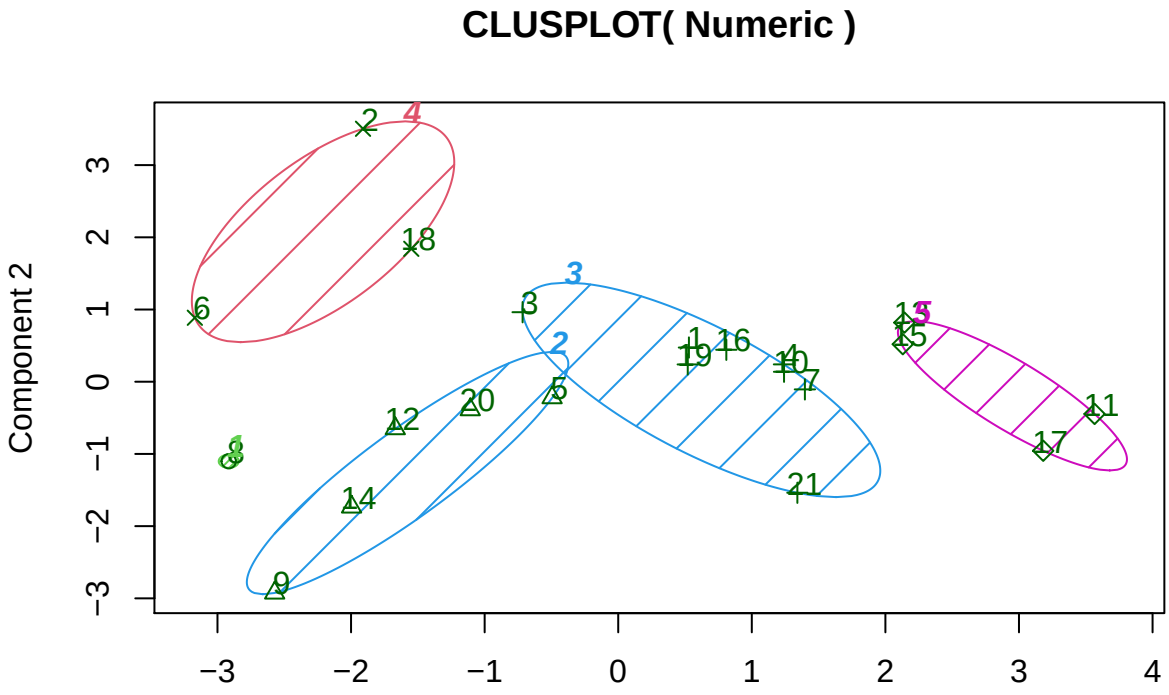
```
fviz_cluster(k5, data = Numeric)
```



From the above cluster graph we can see that there are 5 clusters in respective color and shapes. The Shapes in the center of the each cluster is the centriod or center point. The center is determined after 25 restarts which we have given in kmeans, I have tried to decrease the no of restarts < 25 and there is a discrepancy in centers and > 25 its the same as 25 that means that we have reached the final center points and no further centroid can be considered unless new data is added.

```
library(cluster)
fit <- kmeans(Numeric,5)

clusplot(Numeric, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0) # we can see the row numbers
```



(b)

Cluster_1(BLUE) - Row 3,19,1,16,21,7,10,1

Cluster_2(GREEN) - Row 2,18

Cluster_3(RED) - Row 12,6,18

Cluster_4(PINK FAR RIGHT) - Row 12,15,11,17

Cluster_5(PINK) - Row 9,14,20,5

```
#Below command gives the mean value of all quantitative variables for each cluster.

aggregate(Numeric,by=list(fit$cluster),FUN=mean) # Mean of clusters where selected numerical rows are u
```

```
##   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
```

```
Numeric1 <- data.frame(Numeric, fit$cluster)
```

Cluster_1 = has Highest Rev_growth and low leverage and low beta

Cluster_2 = has Highest PE ratio, Lowest ROE, Lowest ROA, Lowest Asset Turnover, Lowest Net Profit Margin

Cluster_3 = has Highest Market Cap, Highest ROE, Highest ROA, Highest Asset Turnover

Cluster_4 = has Highest Net Profit Margin, Lowest Beta, Lowest PE Ratio, Lowest Rev growth.

Cluster_5 = has Highest Beta, Highest Leverage, Highest Rev growth and Lowest Market Cap.

(c)

There is a pattern in the cluster with respect to the average recommended variable. Cluster 3, which has the highest market capitalization, highest ROE, highest ROA, and highest asset turnover, has no median sales recommendations. Cluster 3 mainly has purchase recommendations with strong purchase recommendations. Cluster 2 with the highest P / E, lowest ROE, lowest ROA, lowest asset turnover, and lowest net return usually has pending recommendations. Cluster 4, with the highest net margin, lowest beta, lowest PE ratio, and lowest Rev growth, is most often recommended to be put on hold.

(d)

We can name various clusters based on their dependence on the quantitative variables.

Cluster_1 - Lowest Leverage cluster and Highest Rev_growth.

Cluster_2 - High PE ratio, Low ROE, Low ROA, Low Asset Turnover and Negative Net Profit Margin Cluster

Cluster_3 - High Market Cap, ROE, ROA, Asset Turnover cluster

Cluster_4 - High Net Profit Margin, High Low Beta and Negative Rev growth cluster

Cluster_5 - High Beta, Negative Leverage, Low Rev growth and Low Market Cap cluster