Assignment 4 Fundamentals of ML

Ankith Dasu

R Markdown

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ISLR)
library(factoextra)
```

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

```
library(tidyverse) # manipulation of data
```

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

```
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     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(flexclust)
```

```
## Loading required package: grid
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
set.seed(64060)

setwd("/Users/ankithdasu/Desktop/Spring 2022/Fundamentals of Machine Learning/Assignment 4")
getwd()
```

```
## [1] "/Users/ankithdasu/Desktop/Spring 2022/Fundamentals of Machine Learning/Assignment 4"
```

```
KMC1 <- read.csv("Pharmaceuticals.csv")
KMC1
```

```
##    Symbol                                Name Market_Cap Beta PE_Ratio  ROE  ROA
## 1     ABT               Abbott Laboratories      68.44 0.32     24.7 26.4 11.8
## 2     AGN                    Allergan, Inc.       7.58 0.41     82.5 12.9  5.5
## 3     AHM                      Amersham plc       6.30 0.46     20.7 14.9  7.8
## 4     AZN                   AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4
## 5     AVE                          Aventis      47.16 0.32     20.1 21.8  7.5
## 6     BAY                         Bayer AG      16.90 1.11     27.9  3.9  1.4
## 7     BMY      Bristol-Myers Squibb Company      51.33 0.50     13.9 34.8 15.1
## 8    CHTT                      Chattem, Inc       0.41 0.85     26.0 24.1  4.3
## 9     ELN            Elan Corporation, plc       0.78 1.08      3.6 15.1  5.1
## 10    LLY            Eli Lilly and Company      73.84 0.18     27.9 31.0 13.5
## 11    GSK              GlaxoSmithKline plc     122.11 0.35     18.0 62.9 20.3
## 12    IVX                 IVAX Corporation       2.60 0.65     19.9 21.4  6.8
## 13    JNJ                Johnson & Johnson     173.93 0.46     28.4 28.6 16.3
## 14    MRX Medicis Pharmaceutical Corporation       1.20 0.75     28.6 11.2  5.4
## 15    MRK                Merck & Co., Inc.     132.56 0.46     18.9 40.6 15.0
## 16    NVS                      Novartis AG      96.65 0.19     21.6 17.9 11.2
## 17    PFE                       Pfizer Inc     199.47 0.65     23.6 45.6 19.2
## 18    PHA            Pharmacia Corporation      56.24 0.40     56.5 13.5  5.7
## 19    SGP      Schering-Plough Corporation      34.10 0.51     18.9 22.6 13.3
## 20    WPI     Watson Pharmaceuticals, Inc.       3.26 0.24     18.4 10.2  6.8
## 21    WYE                           Wyeth      48.19 0.63     13.1 54.9 13.4
##    Asset_Turnover Leverage Rev_Growth Net_Profit_Margin Median_Recommendation
## 1             0.7     0.42       7.54              16.1          Moderate Buy
## 2             0.9     0.60       9.16               5.5          Moderate Buy
## 3             0.9     0.27       7.05              11.2            Strong Buy
## 4             0.9     0.00      15.00              18.0         Moderate Sell
## 5             0.6     0.34      26.81              12.9          Moderate Buy
## 6             0.6     0.00      -3.17               2.6                 Hold
## 7             0.9     0.57       2.70              20.6         Moderate Sell
## 8             0.6     3.51       6.38               7.5          Moderate Buy
## 9             0.3     1.07      34.21              13.3         Moderate Sell
## 10            0.6     0.53       6.21              23.4                 Hold
## 11            1.0     0.34      21.87              21.1                 Hold
## 12            0.6     1.45      13.99              11.0                 Hold
## 13            0.9     0.10       9.37              17.9          Moderate Buy
## 14            0.3     0.93      30.37              21.3          Moderate Buy
## 15            1.1     0.28      17.35              14.1                 Hold
## 16            0.5     0.06      -2.69              22.4                 Hold
## 17            0.8     0.16      25.54              25.2          Moderate Buy
## 18            0.6     0.35      15.00               7.3                 Hold
## 19            0.8     0.00       8.56              17.6                 Hold
## 20            0.5     0.20      29.18              15.1         Moderate Sell
## 21            0.6     1.12       0.36              25.5                 Hold
##       Location Exchange
## 1           US     NYSE
## 2       CANADA     NYSE
## 3           UK     NYSE
## 4           UK     NYSE
## 5       FRANCE     NYSE
```

```
## 6      GERMANY     NYSE
## 7           US     NYSE
## 8           US   NASDAQ
## 9      IRELAND     NYSE
## 10          US     NYSE
## 11          UK     NYSE
## 12          US     AMEX
## 13          US     NYSE
## 14          US     NYSE
## 15          US     NYSE
## 16 SWITZERLAND     NYSE
## 17          US     NYSE
## 18          US     NYSE
## 19          US     NYSE
## 20          US     NYSE
## 21          US     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

Column_Nums <- KMC1 [,3:11] # Considering column 3-11 i.e quantitative variables
head(Column_Nums)
```

```
##   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
```

```
Column_Nums <- scale(Column_Nums)
summary(Column_Nums)
```

```
##    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
```

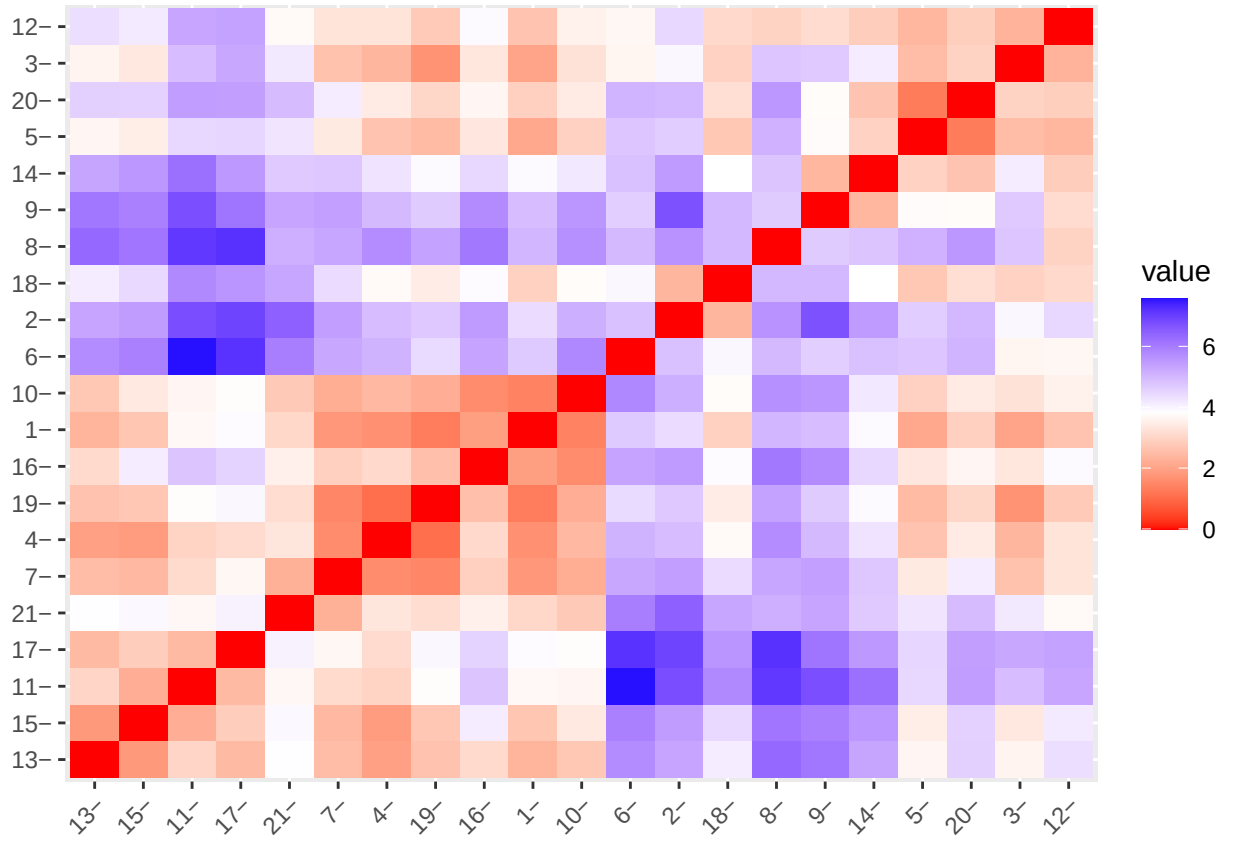
```
##  Min.   :-1.7128   Min.   :-1.8451   Min.   :-0.74966   Min.   :-1.4971
##  1st Qu.:-0.9047   1st Qu.:-0.4613   1st Qu.:-0.54487   1st Qu.:-0.6328
##  Median : 0.1289   Median :-0.4613   Median :-0.31449   Median :-0.3621
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  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
```

```
 #Eucledian Distance is used to calculate the distance for all data points to centroid
Distance_Column_Nums <-get_dist(Column_Nums, method = "euclidean", stand = FALSE)
Distance_Column_Nums
```

```
##           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
```

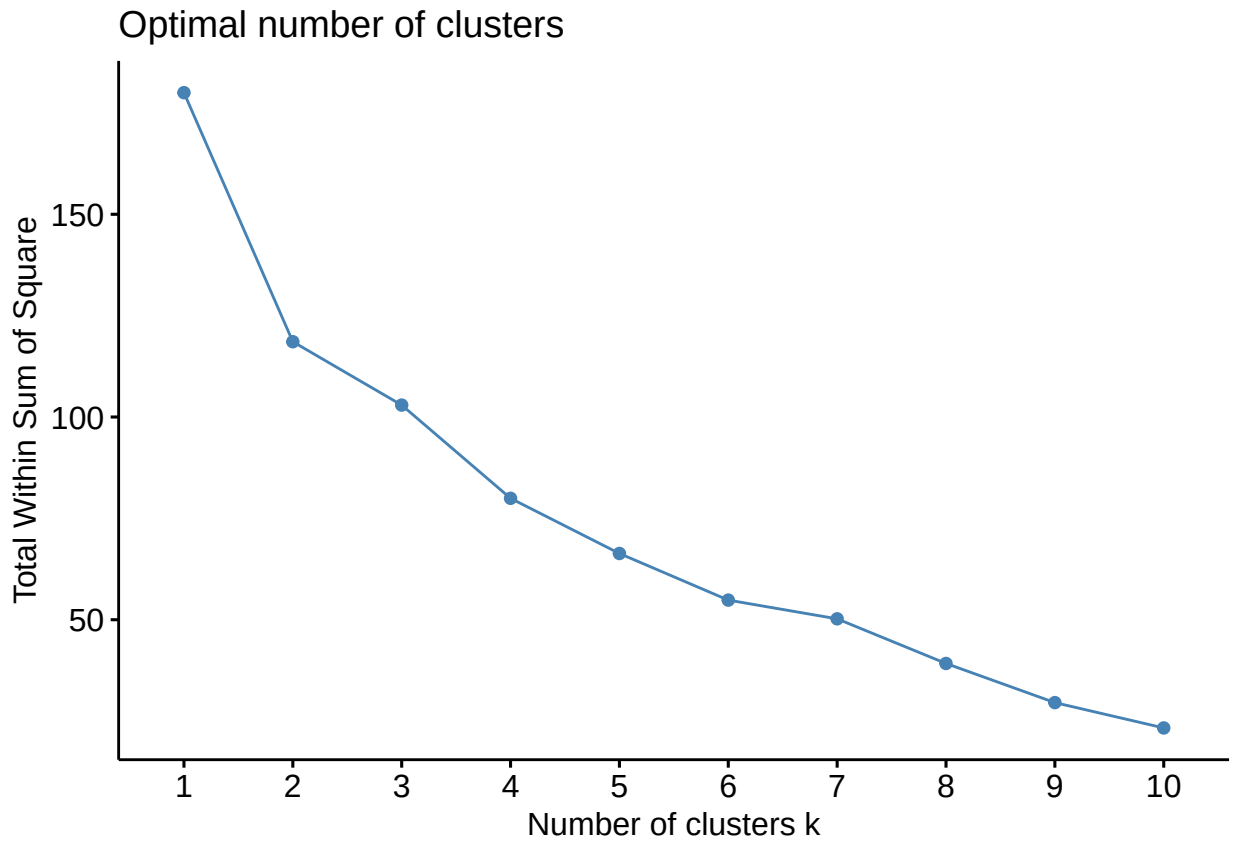
```
## 16 5.732322 1.577953 4.783039 3.899915 3.083678 4.478040 4.112418
## 17 6.123133 3.783136 2.447177 5.356598 2.447341 5.518379 2.831329 4.536250
## 18 5.007721 3.754900 5.773960 3.073579 4.112432 3.827019 4.448933 3.884035
## 19 4.665611 2.205815 3.780283 2.763476 2.604437 3.907501 2.710607 2.542763
## 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
```

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

fviz_dist(dist.obj = Distance_Column_Nums, order = TRUE, show_labels = TRUE)
```



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

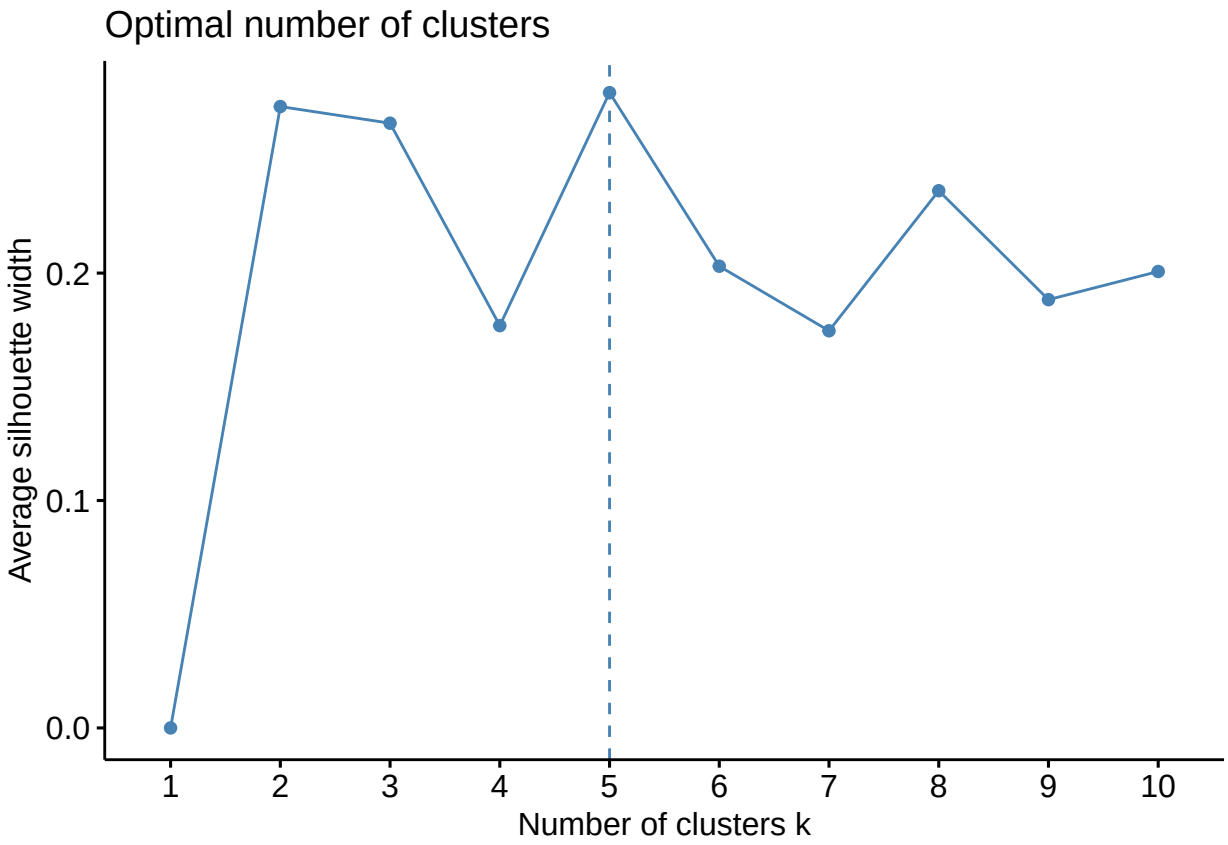
```
library(factoextra) # clustering algorithms & visualization
library(flexclust)


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



#in the plot 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(Column_Nums,kmeans,method="silhouette")
```



#As observed in elbow method, the optimal clusters identified as 2, but when we have used Silhouttes method, we got the value as 5. As the elbow method was not clear in determining the optimal cluster, we shall use Silhouttes method here #We have identified the number of clusters. Now we shall apply K-means algorithm

```
#Applying K-means Algorithm

KMeans4 <- kmeans(Column_Nums, centers = 5, nstart = 25) #Number of restarts = 25
KMeans4
```

```
## 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"
```

#Centers

```
KMeans4$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

```
KMeans4$size
```

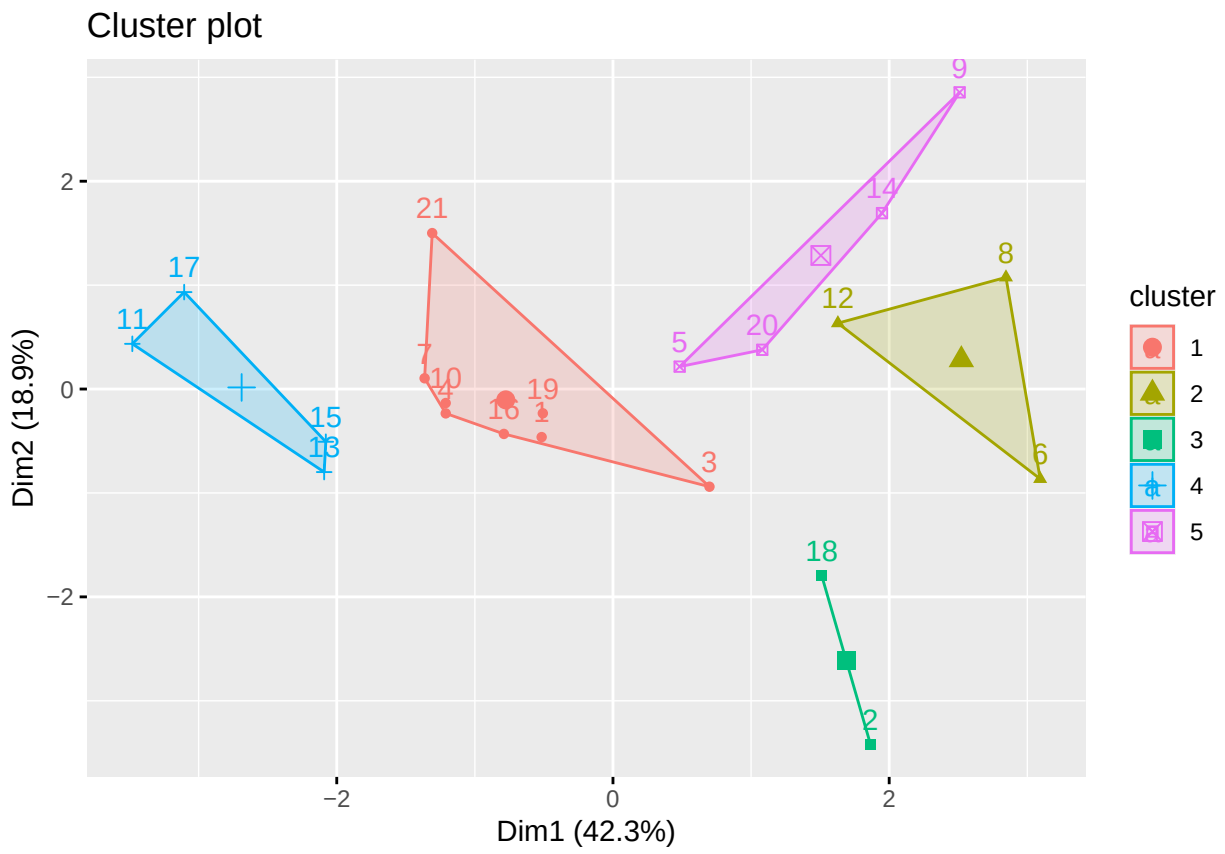
```
## [1] 8 3 2 4 4
```

#Cluster

```
KMeans4$cluster[c(1:21)]
```

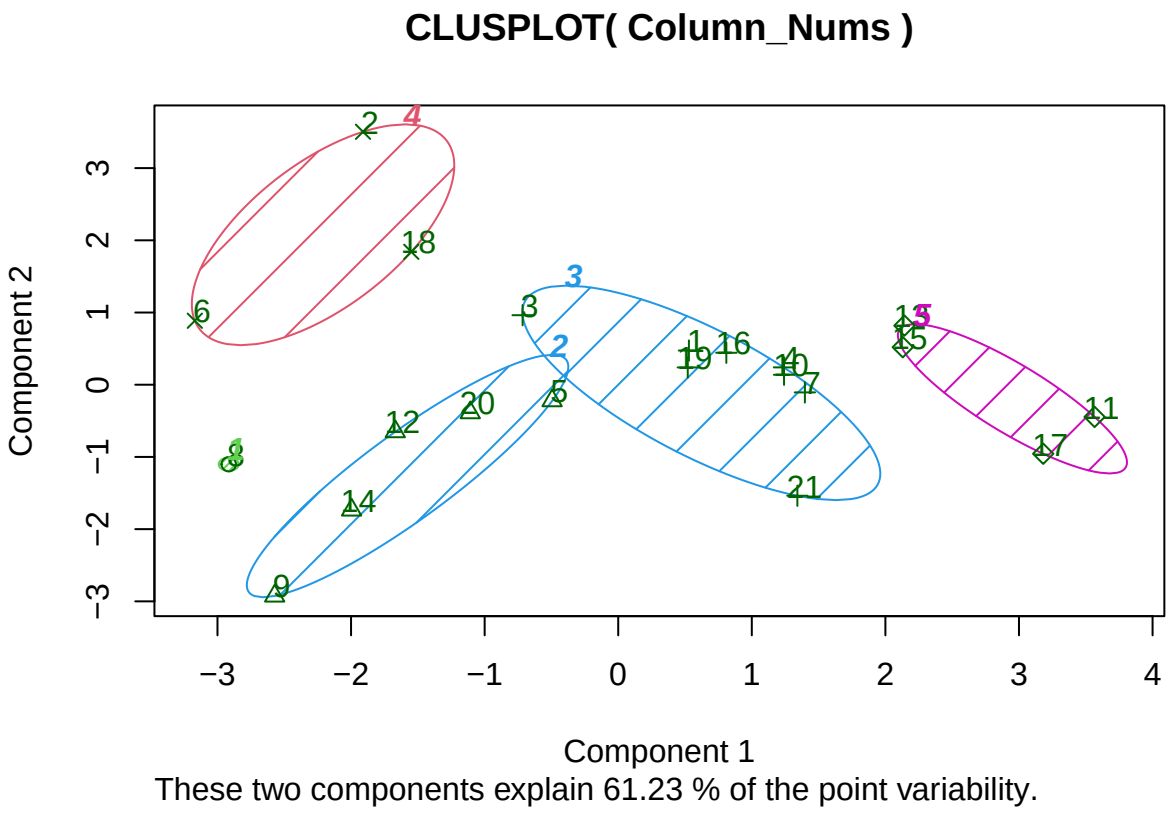
```
##  [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1
```

```
fviz_cluster(KMeans4, data = Column_Nums)
```



5 clusters have been noticed from the above. The symbols/shapes in each cluster are 'centroids' of that specific cluster. Nstart value 25 and above is defined as no other centroid can be taken into consideration until new data is being added.

```
library(cluster)

Clus_Plot <- kmeans(Column_Nums,5)

clusplot(Column_Nums, Clus_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.

#Rows in the excel start from 2. So for our convenience, the rows have been explained starting from 1.(Here 1= 2nd row) 1st Cluster_Red = Rows are 2, 6, 18 2nd Cluster_Green = Rows are 1,4,7,10,16,19,21 3rd Cluster_Blue = Rows are 8,9,12,14 4th Cluster_Pink = Rows are 3,5,20 5th Cluster_Pink(last) = Rows are 11,13,15,17

We find the mean of all the quantitative variables

```
aggregate(Column_Nums,by=list(Clus_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
```

```
Column_Nums1 <- data.frame(Column_Nums, Clus_Plot$cluster)
```

1st Cluster = has Highest PE_Ratio and lowest Net_Profit_Margin, ROA 2nd Cluster = has Highest Net_Profit_Margin and Lowest Rev_Growth, Beta 3rd Cluster = has Highest Leverage, Beta and Lowest ROA 4th Cluster = has Highest Rev_Growth and Lowest Beta, ROE Market_Cap 5th 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 1st Cluster, there is high PE_Ratio and lowest Net_Profit_Margin, ROA. For this cluster, the Median Recommendation is "Moderate Buy" for all the points. In 2nd Cluster, there is high Net_Profit_Margin and low Rev_Growth, Beta. For this cluster, the Median Recommendation is often recommended to be put on "Hold" for majority of the points. In 3rd Cluster, there is high Leverage, Beta and there is low ROA. For this cluster, the Median Recommendation suggests Moderate Buy mostly.
In 4th Cluster, there is high Rev_Growth and Lowest Beta, ROE Market_Cap. For this cluster, the Median recommendation suggests equal Strong Buy, Moderate Buy and Moderate Sell recommendations In 5th Cluster, there is high Market_Cap, ROA, ROE and Lowest Leverage. For this cluster, the Median Recommendation has both Hold and Moderate Buy recommendations.

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

1st Cluster- Low Net_Profit_Margin and ROA cluster or Moderate Buy Cluster

2nd Cluster- Low Rev_Growth, Beta cluster or Hold Cluster

3rd Cluster- High Leverage, Beta cluster or Moderate Cluster

4th Cluster- High Rev_Growth and Lowest Beta, ROE Market_Cap Cluster

5th Cluster- High Market_Cap, ROA, ROE and Lowest Leverage Cluster