ML-Assignment4

Harish Kunaparaju

2022-11-05

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

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

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

```
library(ISLR)
library(caret)
```

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

#Importing the dataset

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

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.

#Remove missing data and rescale variables for comparability before clustering data.

```
PS<- na.omit(Input) #gives the data after removing the missing values.
PS
```

```
##    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
```

#To cluster the 21 firms, just the quantitative variables (1-9) need be collected.

```
row.names(PS)<- PS[,1]
PS1<- PS[,3:11]
head(PS1)
```

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

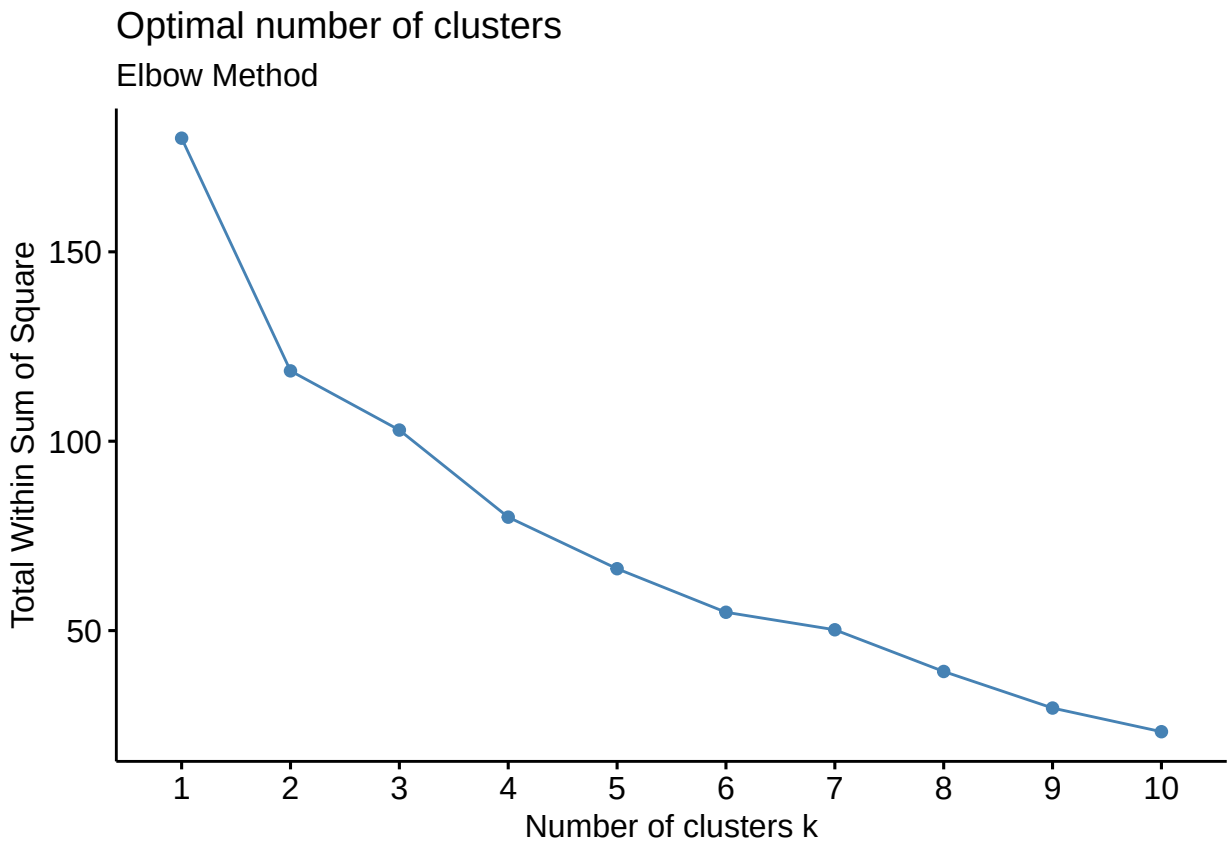
#Scale all the dataframe's quantitative variables

```
PS2<-scale(PS1)
head(PS2)
```

```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
##       Leverage Rev_Growth Net_Profit_Margin
## ABT -0.2120979 -0.5277675        0.06168225
## AGN  0.0182843 -0.3811391       -1.55366706
## AHM -0.4040831 -0.5721181       -0.68503583
## AZN -0.7496565  0.1474473        0.35122600
## AVE -0.3144900  1.2163867       -0.42597037
## BAY -0.7496565 -1.4971443       -1.99560225
```

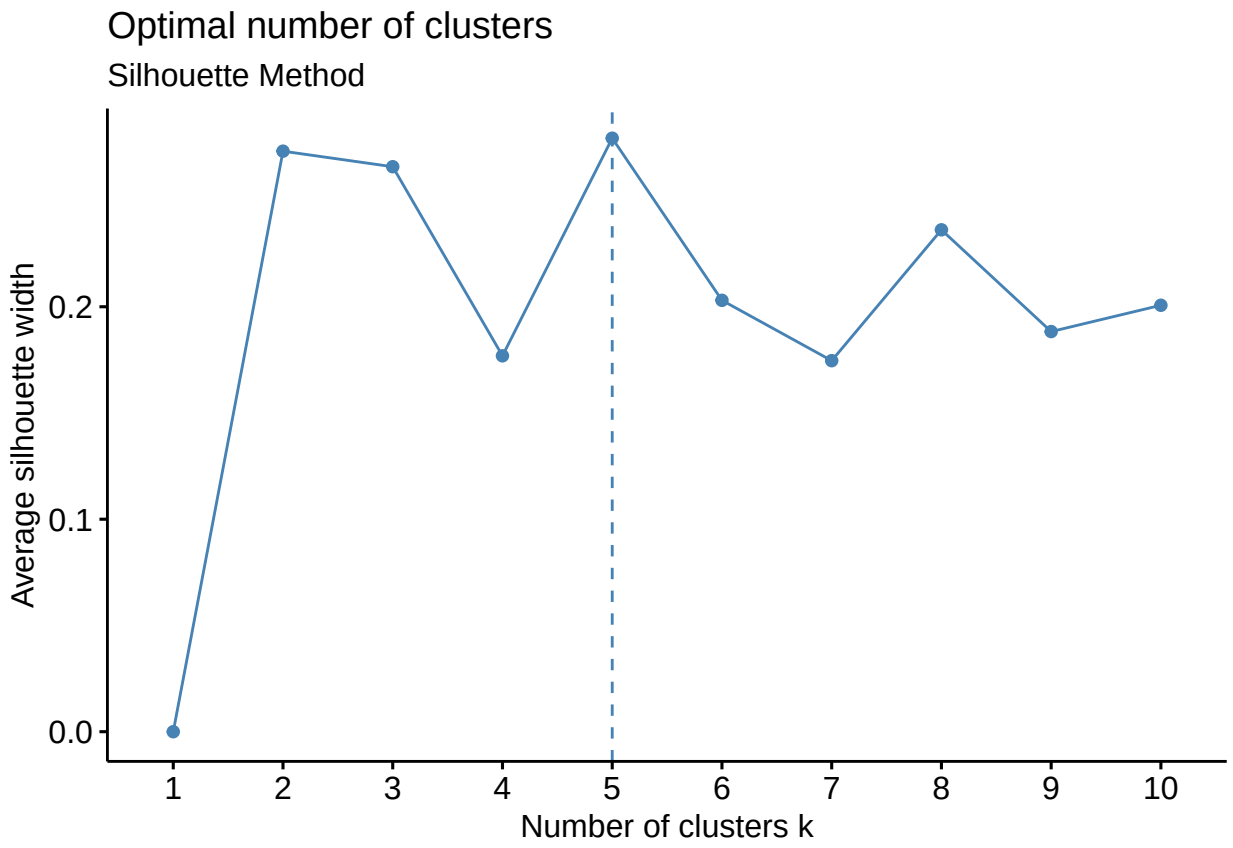
##Determining the no of clusters to do the cluster analysis using Elbow Method

```
fviz_nbclust(PS2, kmeans, method = "wss") + labs(subtitle = "Elbow Method")
```



##Using Silhouette method for determining no of clusters

```
fviz_nbclust(PS2, kmeans, method = "silhouette")+ labs(subtitle = "Silhouette Method")
```



The number of clusters is 5 in the above plots, which is sufficient to display the data variations.

```
set.seed(60000)
k5<- kmeans(PS2,centers=5,nstart = 25)
```

#Visualizing the output

```
k5$centers  #for centroids
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
##      Leverage Rev_Growth Net_Profit_Margin
## 1  1.36644699 -0.6912914      -1.320000179
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.46807818  0.4671788       0.591242521
## 5 -0.14170336 -0.1168459      -1.416514761
```

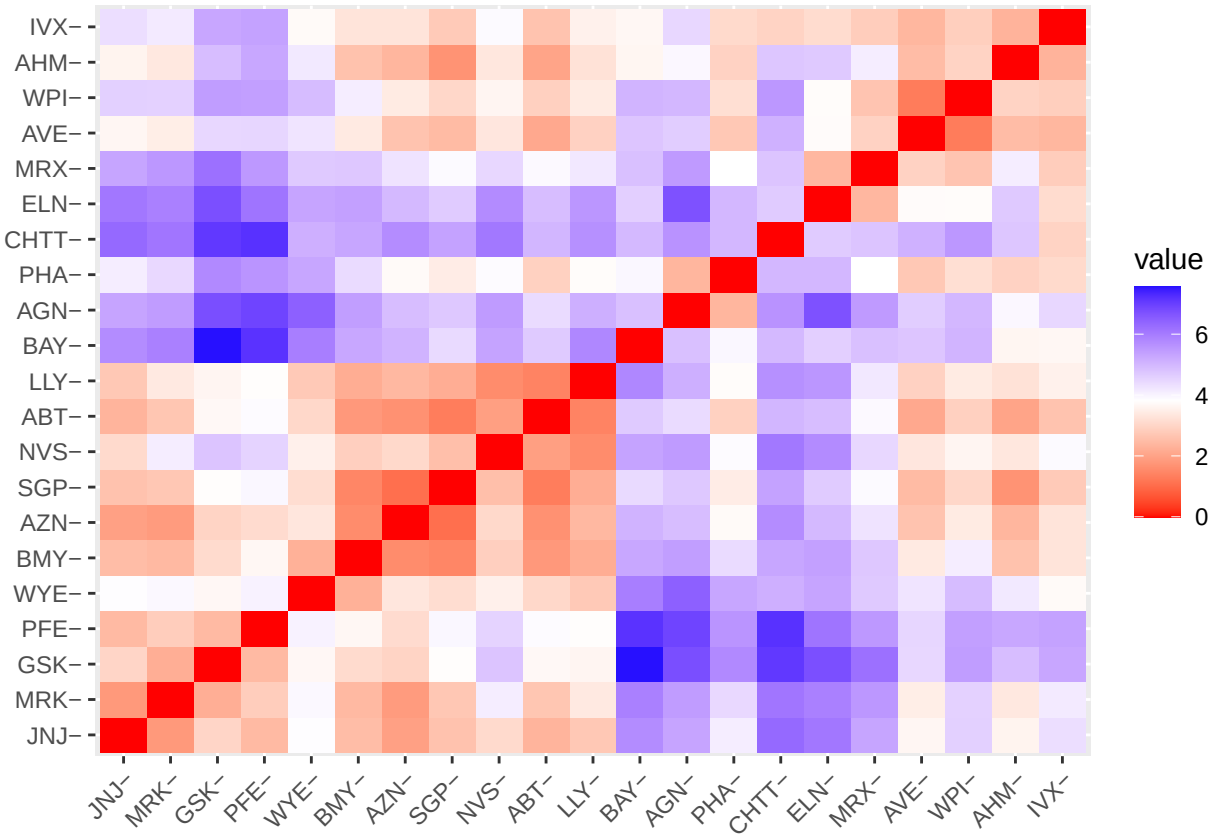
```
fviz_cluster(k5,data = PS2) # to Visualize the clusters
```


```
k5
```

```
## K-means clustering with 5 clusters of sizes 3, 4, 8, 4, 2
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
##      Leverage Rev_Growth Net_Profit_Margin
## 1  1.36644699 -0.6912914      -1.320000179
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.46807818  0.4671788       0.591242521
## 5 -0.14170336 -0.1168459      -1.416514761
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    3    5    3    3    2    1    3    1    2    3    4    1    4    2    4    3
##  PFE  PHA  SGP  WPI  WYE
##    4    5    3    2    3
##
## Within cluster sum of squares by cluster:
## [1] 15.595925 12.791257 21.879320  9.284424  2.803505
##  (between_SS / total_SS =  65.4 %)
```

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
distance<- dist(PS2, method = "euclidean")
fviz_dist(distance)
```



#Using K-Means Cluster Analysis- to Fit the data with 5 clusters

```
fit<-kmeans(PS2,5)
```

#calculating the mean of all quantitative variables in each cluster

```
aggregate(PS2,by=list(fit$cluster),FUN=mean)
```

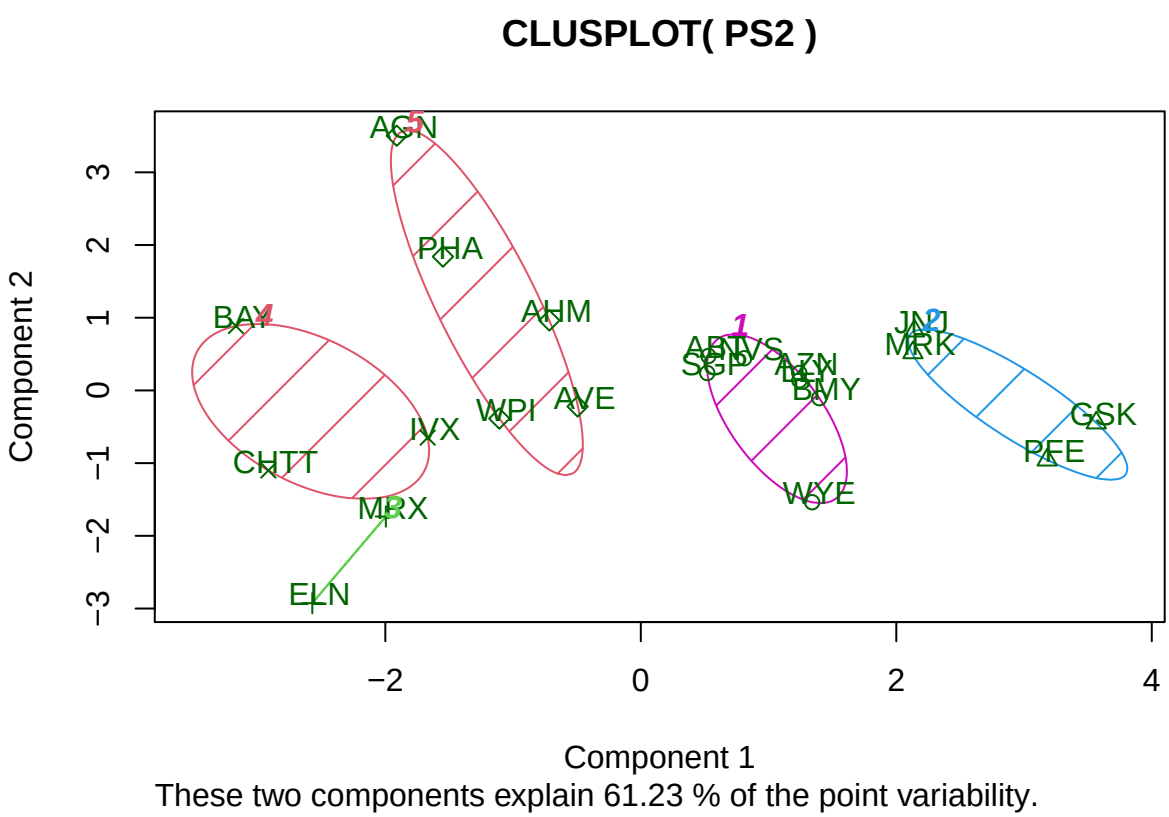
```
##   Group.1  Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1  0.08926902 -0.4618336 -0.32086149  0.3260892  0.5396003
## 2       2  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
## 3       3 -0.96686975  1.5162611 -0.57398880 -0.8382671 -0.9892673
## 4       4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478
## 5       5 -0.57238455 -0.6220844  0.86927480 -0.7381675 -0.7242993
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1   6.589509e-02 -0.2559803 -0.7230135         0.7343816
## 2   1.153164e+00 -0.4680782  0.4671788         0.5912425
## 3  -1.845062e+00  0.5302448  1.7123890         0.2445520
## 4  -4.612656e-01  1.3664470 -0.6912914        -1.3200002
## 5  -2.442491e-16 -0.2991312  0.3682951        -0.8069490
```

```
PS3<-data.frame(PS2,fit$cluster)
head(PS3)
```

```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
##       Leverage Rev_Growth Net_Profit_Margin fit.cluster
## ABT -0.2120979 -0.5277675        0.06168225           1
## AGN  0.0182843 -0.3811391       -1.55366706           5
## AHM -0.4040831 -0.5721181       -0.68503583           5
## AZN -0.7496565  0.1474473        0.35122600           1
## AVE -0.3144900  1.2163867       -0.42597037           5
## BAY -0.7496565 -1.4971443       -1.99560225           4
```

#view of the cluster plot

```
library(cluster)
clusplot(PS2,fit$cluster,color = TRUE,shade = TRUE,labels = 2,lines = 0)
```



These two components explain 61.23 % of the point variability.

#b.Interpret the clusters with respect to the numerical variables used in forming the clusters. By looking at the mean values of all quantitative variables in each cluster.

Cluster 1 - has highest Market_cap,ROA,ROE,Asset_Turnover and lowest is Beta,PE_Ratio.

Cluster 2 - has highest Rev_Growth and lowest PE_Ratio, Asset_Turnover.

Cluster 3 - has highest Beta, Leverage and lowest Market_Cap, ROE, ROA, Leverage, Rev_Growth,

Net_Profit_Margin.

Cluster 4 - has highest PE_Ratio and lowest Leverage, Asset_Turnover.

Cluster 5 - has highest Net_Profit_Margin and lowest leverage,Beta.

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

With respect to the Media recommendation variable, there is a pattern in the clusters.

Cluster 1 with highest Market_Cap, highest ROE, highest ROA, highest Asset_Turnover has equal Hold and Moderate Buy Recommendation.

Cluster 2 with lowest PE_Ratio and lowest Asset_Turnover has Hold Recommendation.

Cluster-3 with highest Beta, highest Leverage has mostly Moderate Buy Recommendation.

Cluster 4 with highest PE_Ratio has Hold Recommendation.

Cluster 5 with highest Net_Profit_Margin has mostly Hold Recommendation.

In terms of variables in clusters (10 to 12).

Clusters 1,3 has mostly Moderate Buy Recommendation.

Clusters 1,2,4,5 has Hold Recommendation.

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

Cluster-1 - Hold cluster.

Cluster-2 - Hold cluster.

Cluster-3 - Buy Cluster.

Cluster-4 - High Hold cluster.

Cluster-5 - High Hold cluster.