Assignment K-means Clustering

Jyoti Phogat

2022-11-02

```
library(Hmisc) #Contents and Describe
```

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

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

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

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(leaps) #Variable selection
library(MASS)
library(NbClust)
```

```
pharmaceutical_data <- read.csv("Pharmaceuticals.csv", header=TRUE)
pharmaceutical_data
```

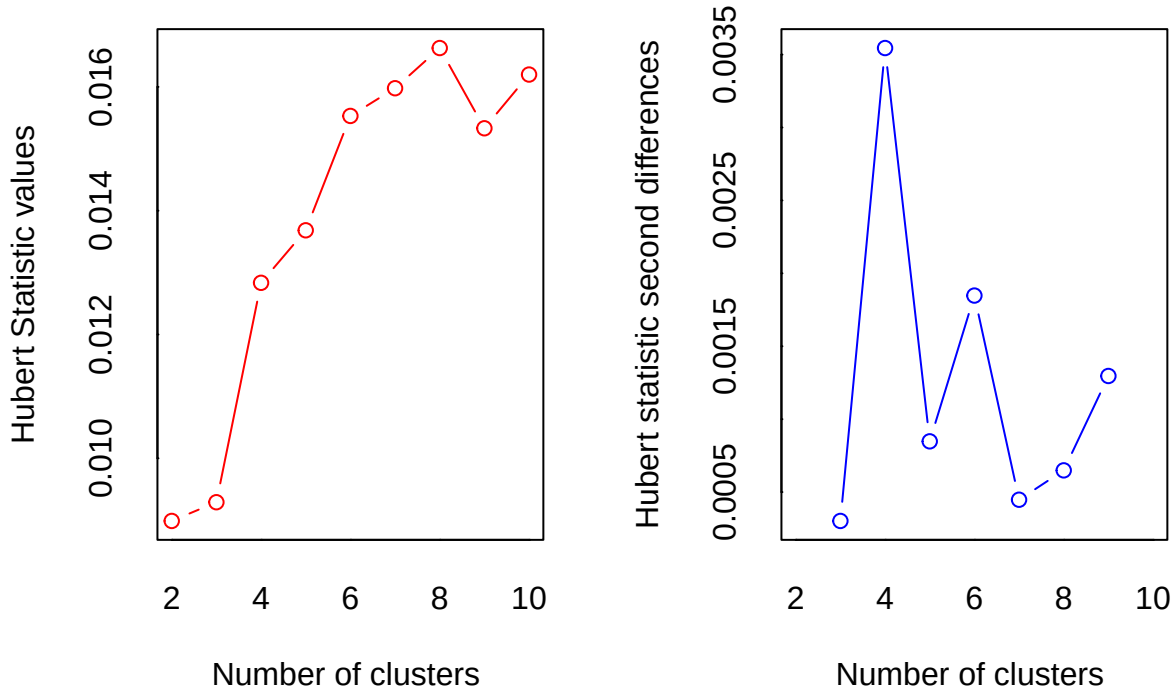
```
##    Symbol                                Name Market_Cap Beta PE_Ratio  ROE  ROA
## 1     ABT               Abbott Laboratories      68.44 0.32     24.7 26.4 11.8
## 2     AGN                    Allergan, Inc.       7.58 0.41     82.5 12.9  5.5
## 3     AHM                      Amersham plc       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
```

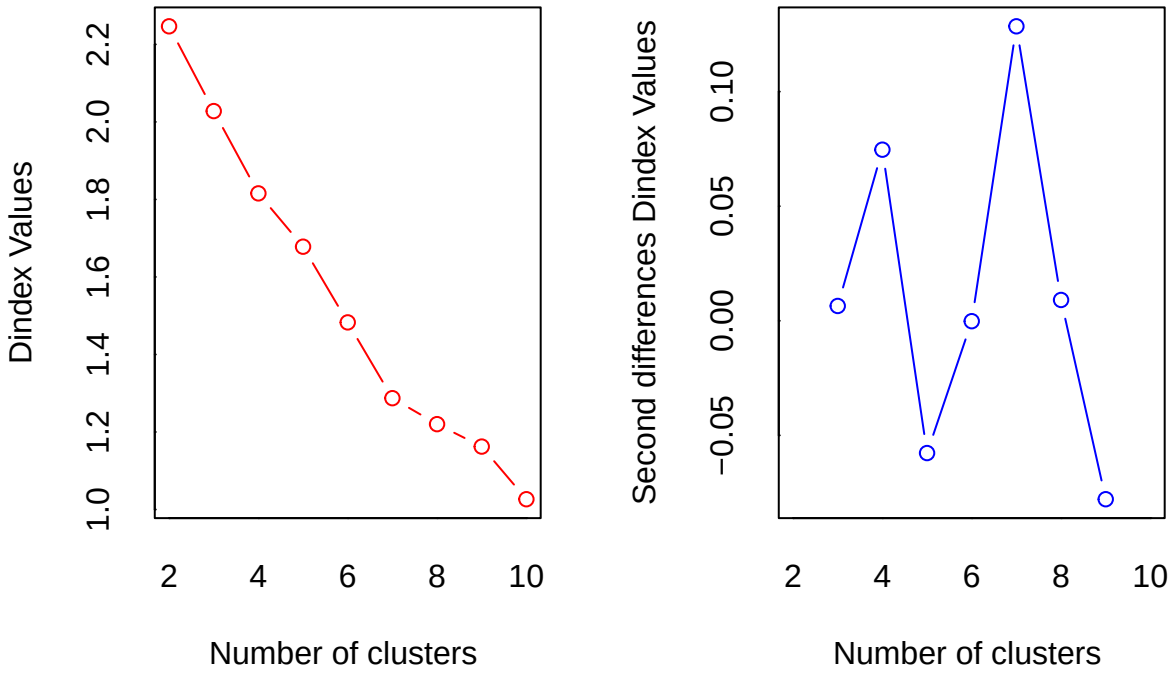
```
pc<-pharmaceutical_data[,2]
row.names(pharmaceutical_data) <- pharmaceutical_data[,2]
pharmaceutical_data <- pharmaceutical_data[, -c(1,2,12,13,14)]
pharmaceutical_data.norm <- sapply(pharmaceutical_data, scale)
set.seed(42)

devAskNewPage(ask=TRUE)

nc <- NbClust(pharmaceutical_data.norm, min.nc=2,
```

```
                max.nc=10, method="kmeans")
```

```
## Warning in pf(beale, pp, df2): NaNs produced
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```



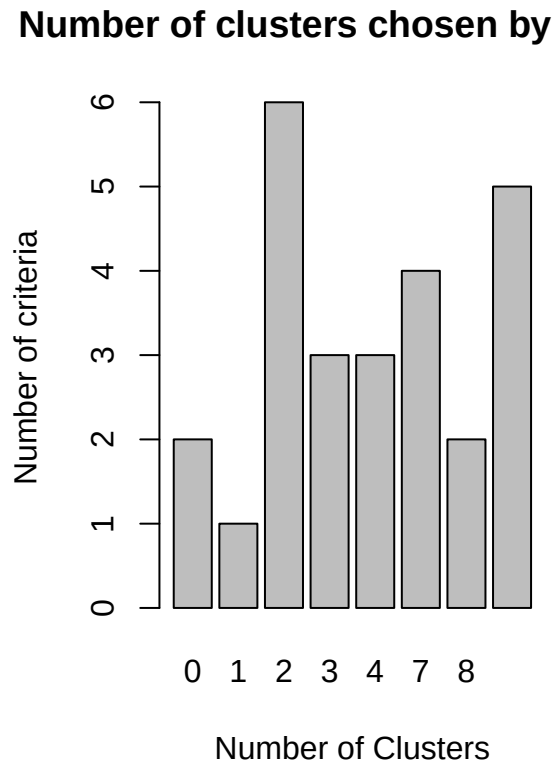

```
## *** : The D index is a graphical method of determining the number of clusters.
```

```
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *****************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 3 proposed 4 as the best number of clusters
## * 4 proposed 7 as the best number of clusters
## * 2 proposed 8 as the best number of clusters
## * 5 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *****************************************************************
```

```
table(nc$Best.n[1,])
```

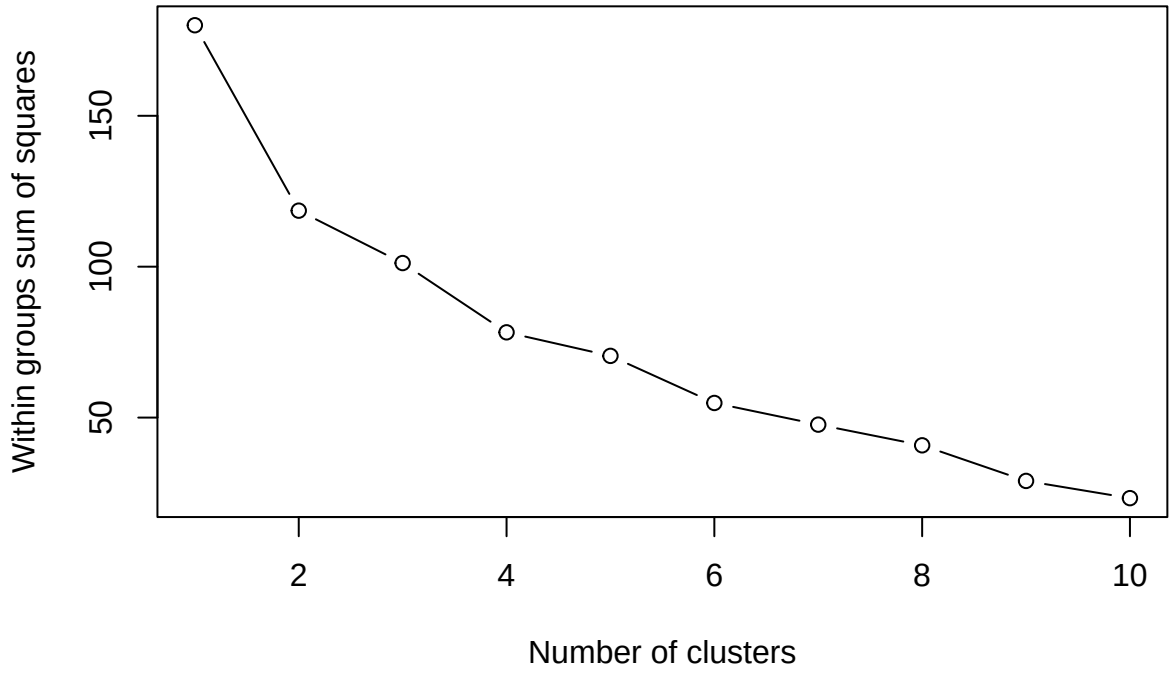
```
##
##  0  1  2  3  4  7  8 10
##  2  1  6  3  3  4  2  5
```

```
barplot(table(nc$Best.n[1,]), xlab="Number of Clusters", ylab="Number of criteria", main="Number of clus
```



#Use the wss plot to check the number of clusters

```
wssplot <- function(pharmaceutical_data.norm, nc=10, seed=42) {
  wss <- (nrow(pharmaceutical_data.norm)-1)*sum(apply(pharmaceutical_data.norm, 2, var))
  for (i in 2:nc) {
    set.seed(42)
    wss[i] <- sum(kmeans(pharmaceutical_data.norm, centers=i)$withinss)
  }
  plot(1:nc, wss, type="b", xlab="Number of clusters", ylab="Within groups sum of squares")
}
wssplot(pharmaceutical_data.norm,nc=10)
```



Perform k-means cluster analysis

```
fit.km <- kmeans(pharmaceutical_data.norm, 5, nstart=10)
fit.km$size
```

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

```
fit.km$centers
```

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

```
fit.km$withinss
```

```
## [1] 21.879320  2.803505 12.791257 15.595925  9.284424
```

calcualte cluster centroidsfit.km$centers

```
fit.km$centers
```

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

```
fit.km$withinss
```

```
## [1] 21.879320  2.803505 12.791257 15.595925  9.284424
```

#check if any cluster is a striking outlier

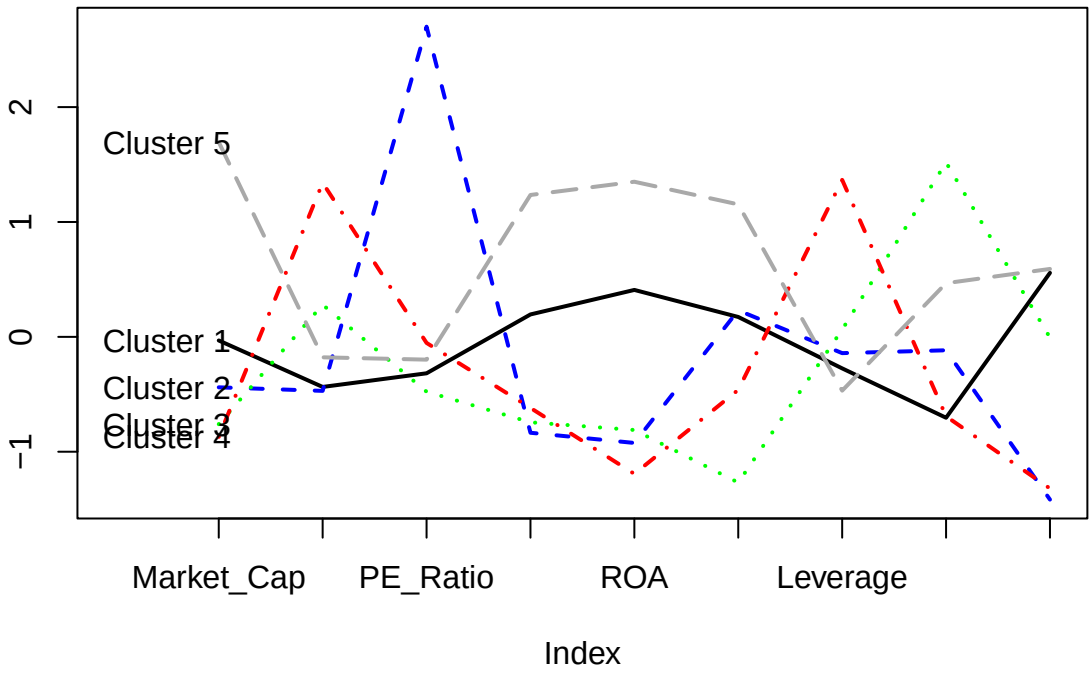
```
dist(fit.km$centers)
```

```
##          1        2        3        4
## 2 4.045579
## 3 3.299161 4.210877
## 4 3.711570 3.775790 3.230532
## 5 2.720924 5.275301 4.744753 5.457397
```

#Profile plot of Centroids

plot an empty scatter plot

```
plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(fit.km$centers), max(fit.km$centers)), xlim =
# label x-axes
axis(1, at = c(1:9), labels = names(pharmaceutical_data))
# plot centroids
for(i in c(1:5))lines(fit.km$centers[i,], lty = i, lwd = 2,
                      col = ifelse(i %in% c(1),"black",
                                   (ifelse(i %in% c(2),"blue",
                                         (ifelse(i %in% c(3),"green",
                                               (ifelse(i %in% c(4),"red","dark grey")))))))

text(x = 0.5, y = fit.km$centers[, 1], labels = paste("Cluster", c(1:5)))
```



Cluster 5 can be called as 'Big doing great' as it has the highest market cap with high Asset turnover, low Beta(risk) and a profit margin greater than all other clusters.

Cluster 3 can be called 'Recovering fast' as it has a low market cap, high Beta(risk), lowest asset turnover, good profit and highest revenue growth.

Cluster 4 can be called as 'High risk no recovery' as it has the highest Beta, low market cap, low ROE, low ROA, least revenue growth and least net profit margin.

Cluster 2 can be called as 'Stable going good' as it has the lease Beta(risk), high ROE, high ROA, good revenue growth and high Net Profit Margin.

Cluster 1 can be called as 'Stable best buy' as it has the least Beta just like Cluster 1 , has an average asset turnover, average revenue growth and has the highest PE Ratio among all the clusters which is the factor used to select which stocks to buy. PE Ratio is the ratio of current Stock market price to the earning per share.