

Assignment 4

Prerak Patel

3/18/2021

Pharmaceuticals Industry

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv. For each firm, the following variables are recorded:

Data Overview

```
str(Ph.data)
```

```
## spec_tbl_df [21 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ Symbol      : chr [1:21] "ABT" "AGN" "AHM" "AZN" ...
## $ Name        : chr [1:21] "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "A
straZeneca PLC" ...
## $ Market_Cap  : num [1:21] 68.44 7.58 6.3 67.63 47.16 ...
## $ Beta        : num [1:21] 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
## $ PE_Ratio    : num [1:21] 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
## $ ROE         : num [1:21] 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
## $ ROA         : num [1:21] 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
## $ Asset_Turnover : num [1:21] 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
## $ Leverage    : num [1:21] 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
## $ Rev_Growth  : num [1:21] 7.54 9.16 7.05 15 26.81 ...
## $ Net_Profit_Margin : num [1:21] 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
## $ Median_Recommendation: chr [1:21] "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sel
l" ...
## $ Location    : chr [1:21] "US" "CANADA" "UK" "UK" ...
## $ Exchange    : chr [1:21] "NYSE" "NYSE" "NYSE" "NYSE" ...
## - attr(*, "spec")=
## .. cols(
## ..   Symbol = col_character(),
## ..   Name = col_character(),
## ..   Market_Cap = col_double(),
## ..   Beta = col_double(),
## ..   PE_Ratio = col_double(),
## ..   ROE = col_double(),
## ..   ROA = col_double(),
## ..   Asset_Turnover = col_double(),
## ..   Leverage = col_double(),
## ..   Rev_Growth = col_double(),
## ..   Net_Profit_Margin = col_double(),
## ..   Median_Recommendation = col_character(),
## ..   Location = col_character(),
## ..   Exchange = col_character()
## .. )
```

Kmeans clustering is only done with variables having continuous data. Hence variables - 'symbol', 'Name', 'Median_Recommendation', 'Location', 'Exchange' will be dropped from further analytic steps

Data cleaning

```
summary(Ph.data)
```

```
##      Symbol      Name      Market_Cap      Beta
## Length:21      Length:21      Min.   : 0.41      Min.   :0.1800
## Class :character Class :character 1st Qu.: 6.30      1st Qu.:0.3500
## Mode  :character Mode  :character Median : 48.19      Median :0.4600
##                                     Mean  : 57.65      Mean   :0.5257
##                                     3rd Qu.: 73.84      3rd Qu.:0.6500
##                                     Max.   :199.47      Max.   :1.1100
##      PE_Ratio      ROE      ROA      Asset_Turnover      Leverage
## Min.   : 3.60      Min.   : 3.9      Min.   : 1.40      Min.   :0.3      Min.   :0.0000
## 1st Qu.:18.90      1st Qu.:14.9      1st Qu.: 5.70      1st Qu.:0.6      1st Qu.:0.1600
## Median :21.50      Median :22.6      Median :11.20      Median :0.6      Median :0.3400
## Mean   :25.46      Mean   :25.8      Mean   :10.51      Mean   :0.7      Mean   :0.5857
## 3rd Qu.:27.90      3rd Qu.:31.0      3rd Qu.:15.00      3rd Qu.:0.9      3rd Qu.:0.6000
## Max.   :82.50      Max.   :62.9      Max.   :20.30      Max.   :1.1      Max.   :3.5100
##      Rev_Growth      Net_Profit_Margin      Median_Recommendation      Location
## Min.   : -3.17      Min.   : 2.6      Length:21      Length:21
## 1st Qu.: 6.38      1st Qu.:11.2      Class :character      Class :character
## Median : 9.37      Median :16.1      Mode  :character      Mode  :character
## Mean   :13.37      Mean   :15.7
## 3rd Qu.:21.87      3rd Qu.:21.1
## Max.   :34.21      Max.   :25.5
##      Exchange
## Length:21
## Class :character
## Mode  :character
##
##
##
```

Checking missing values

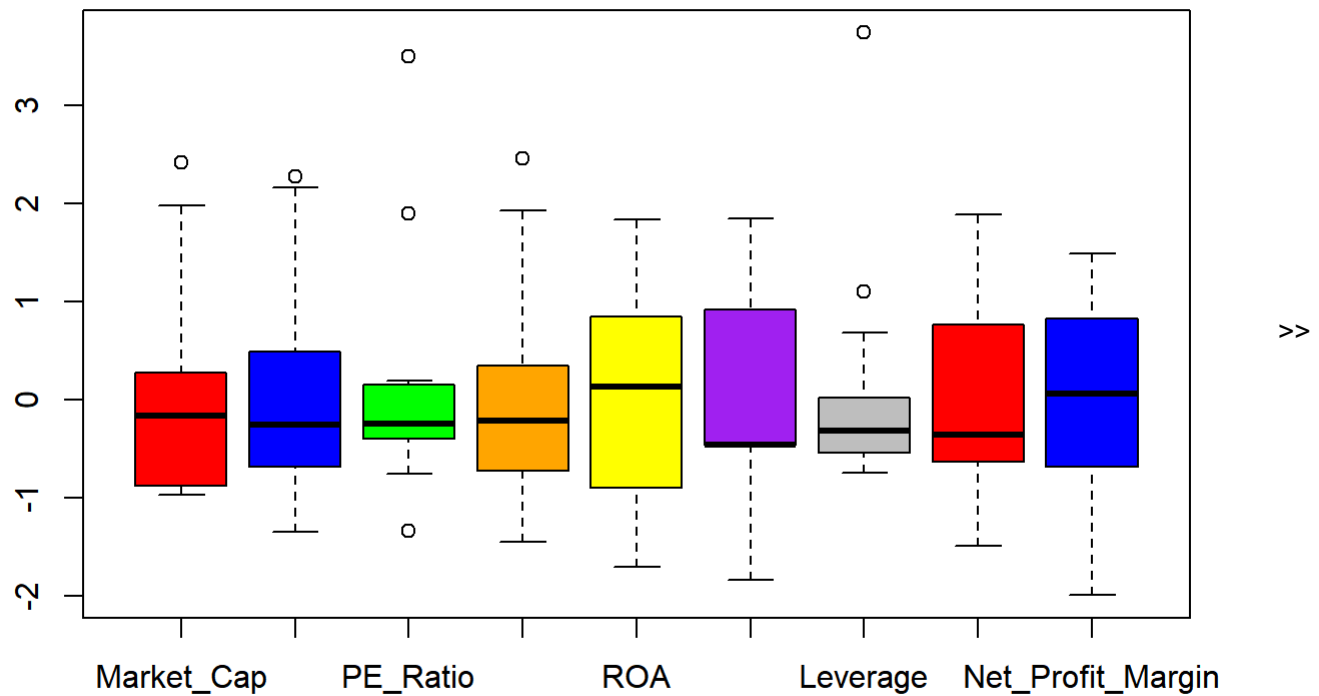
```
colSums(is.na(Ph.data))
```

```
##      Symbol      Name      Market_Cap
##      0      0      0
##      Beta      PE_Ratio      ROE
##      0      0      0
##      ROA      Asset_Turnover      Leverage
##      0      0      0
##      Rev_Growth      Net_Profit_Margin      Median_Recommendation
##      0      0      0
##      Location      Exchange
##      0      0
```

Analyzing outliers for every variable before normalizing the variable, Outliers should not be taken for granted. As in our problem extreme points of some of the variables may be the triggers of a sell off or buy of a particular stock, which if missed may lead to an unrecoverable opportunity cost.

```
#normalizing data to fit all variables in the same graph
# Scaling the data frame (z-score)
data <- data.frame(scale(Ph.data[,3:11]))

boxplot(data, col=c("red","blue","green", "Orange","yellow", "Purple", "grey" ))
```



There are 8 outlier points over 9 variables of the pharmaceutical data. While selecting the optimized K value for implementing K-means algorithm. We will need to remove these outliered points before evaluating the optimized k value.

```
library(factoextra)
```

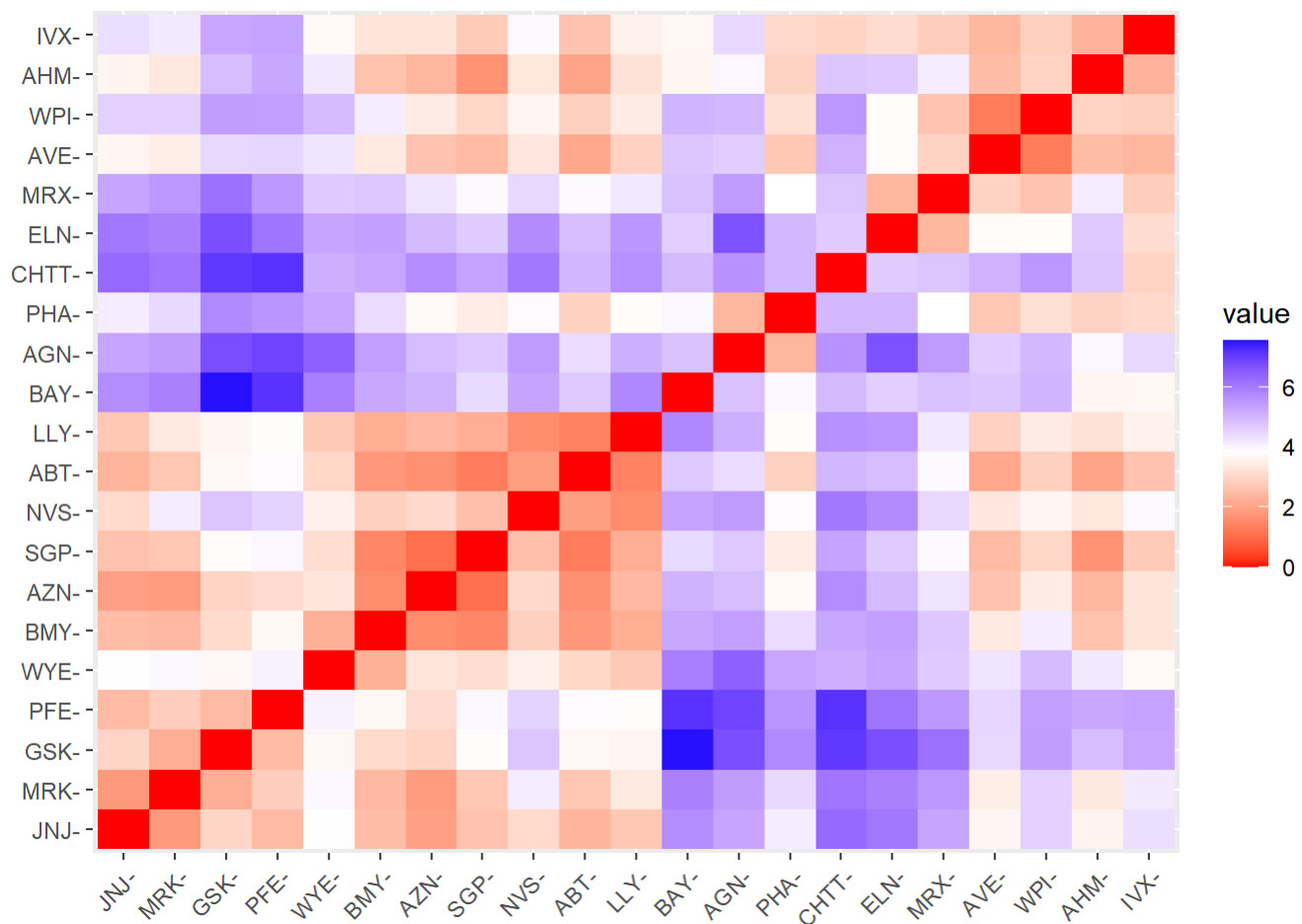
```
## Warning: package 'factoextra' was built under R version 4.0.4
```

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

```

v_name <- Ph.data[,1]
row.names(data) <- unlist(v_name) #Adding rownames from the original dataset as identifiers
distance <- get_dist(data,"euclidean")
fviz_dist(distance,
  order = TRUE,
  show_labels = TRUE,
  lab_size = NULL,
  gradient = list(low = "red", mid = "white", high = "blue"))

```



Determining k

Before determining k we need a dataframe containing data without the outliers

```

# Function to detect all outliers from the numerical variable data
an <- function(x){
  q1 <- quantile(data[,x],0.25)
  q3 <- quantile(data[,x],0.75)
  iqr <- q3 -q1
  lower <- q1-1.5*iqr
  upper <- q3+1.5*iqr
  data[x][(data[x]<lower) | (data[x]>upper), ]
}

dummy <- vector('list',length = length(data))
for(i in seq_along(data)){
  dummy[[i]] <- an(names(data)[i])
}
names(dummy) <- names(data)

temp_data <- data %>% filter(Market_Cap != dummy[[1]], Beta != dummy[[2]], !(PE_Ratio %in% dummy
[[3]]),
                          ROE != dummy[[4]], !(Leverage %in% dummy[[6]]))

```

List of all points from each variable resulting outliers are filtered out from the source data and saved into a temporary data; temp_data. Which is further used in the Elbow method, silhouette method and gap-static method to measure the optimized value of K

```

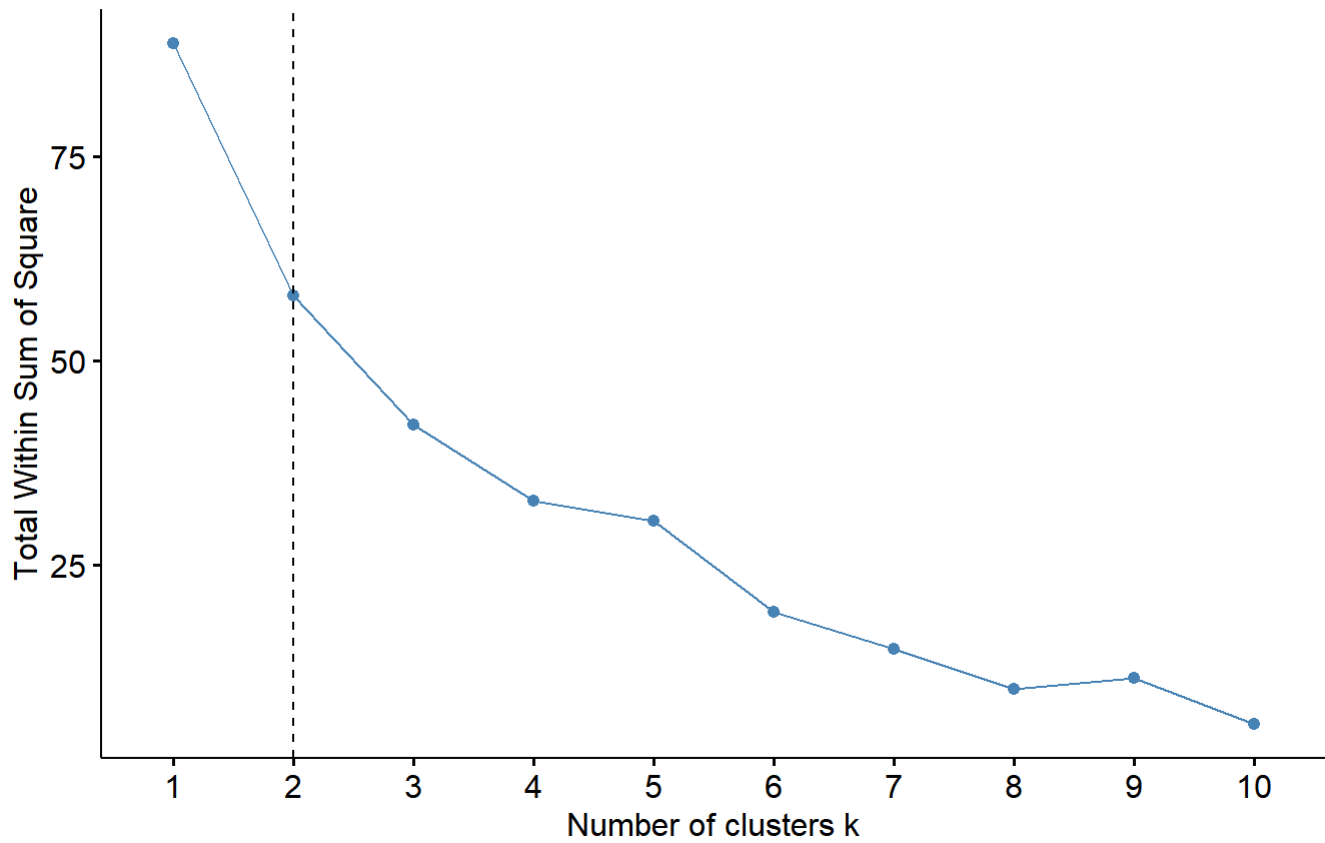
library(factoextra)

# Elbow method
fviz_nbclust(temp_data, kmeans, method = "wss") +
  geom_vline(xintercept = 2, linetype = 2)+
  labs(subtitle = "Elbow method")

```

Optimal number of clusters

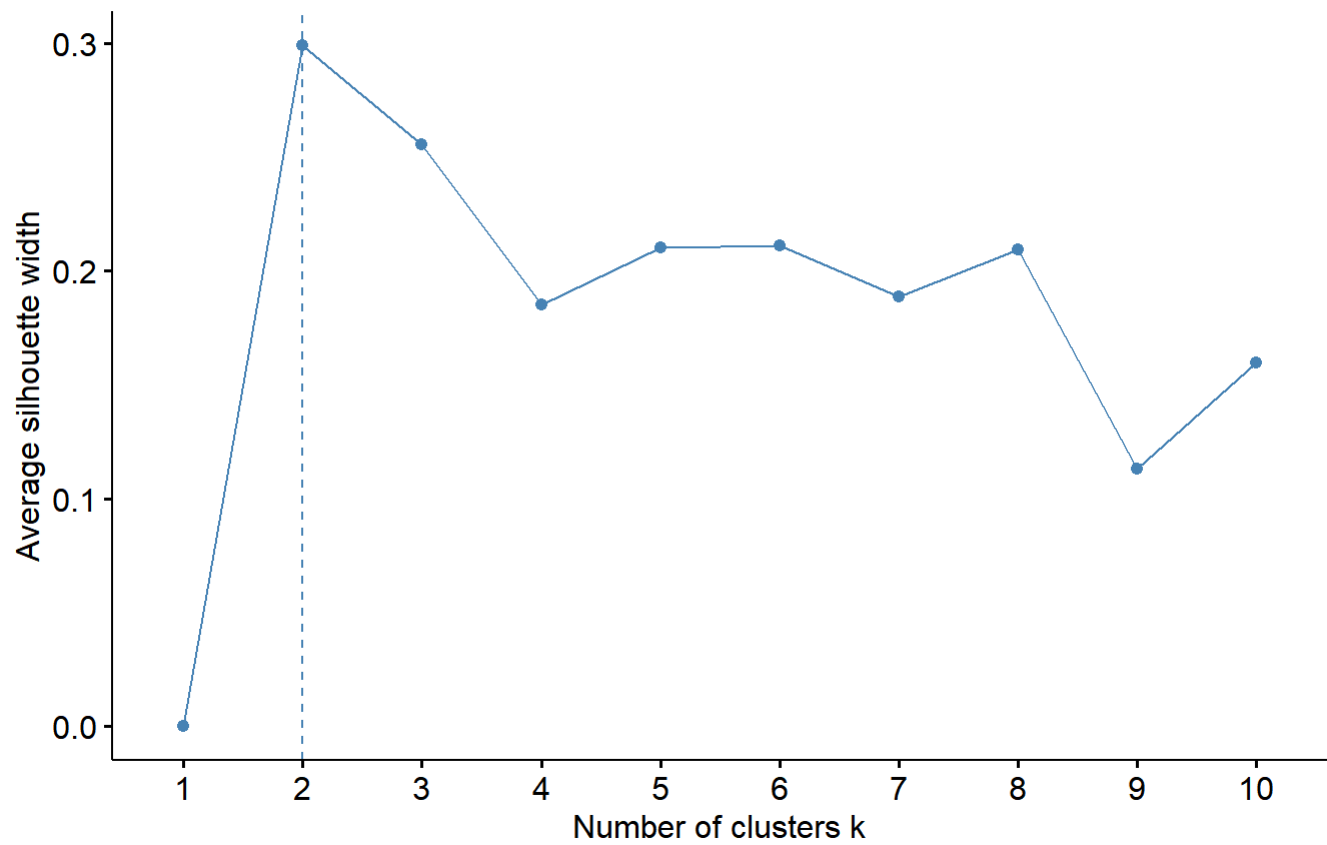
Elbow method



```
# Silhouette method  
fviz_nbclust(temp_data, kmeans, method = "silhouette")+  
  labs(subtitle = "Silhouette method")
```

Optimal number of clusters

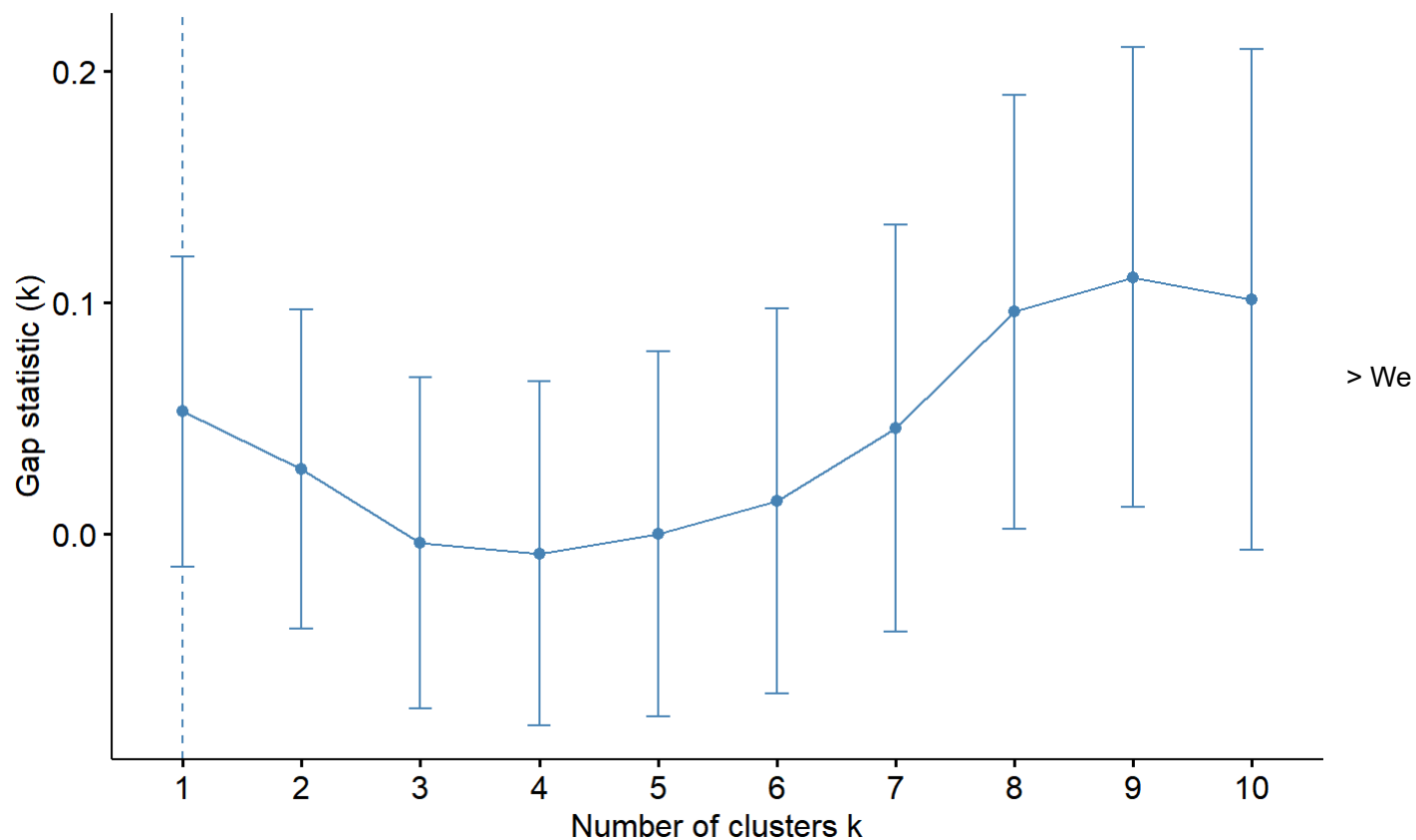
Silhouette method



```
set.seed(123)
fviz_nbclust(temp_data, kmeans, nstart = 25, method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```


Optimal number of clusters

Gap statistic method



can conclude that the values of K can be 2 or 1. We will consider k-value to be 2 and continue with generating clusters with kmeans modelling technique.

```
ph.cluster2 <- kmeans(data, 2, 25)
ph.cluster2
```

```
## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##   Market_Cap      Beta  PE_Ratio      ROE      ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159    0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575   -0.5073922
##   Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163    0.6823310
## 2  0.3664175  0.3192379   -0.7505641
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY  CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##   1   2   2   1   2   2   1   2   2   1   1   2   1   2   1   1
##  PFE  PHA  SGP  WPI  WYE
##   1   2   1   2   1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
## (between_SS / total_SS =  34.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# Visualize the output
ph.cluster2$centers      # output the centers
```

```
##   Market_Cap      Beta  PE_Ratio      ROE      ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159    0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575   -0.5073922
##   Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163    0.6823310
## 2  0.3664175  0.3192379   -0.7505641
```

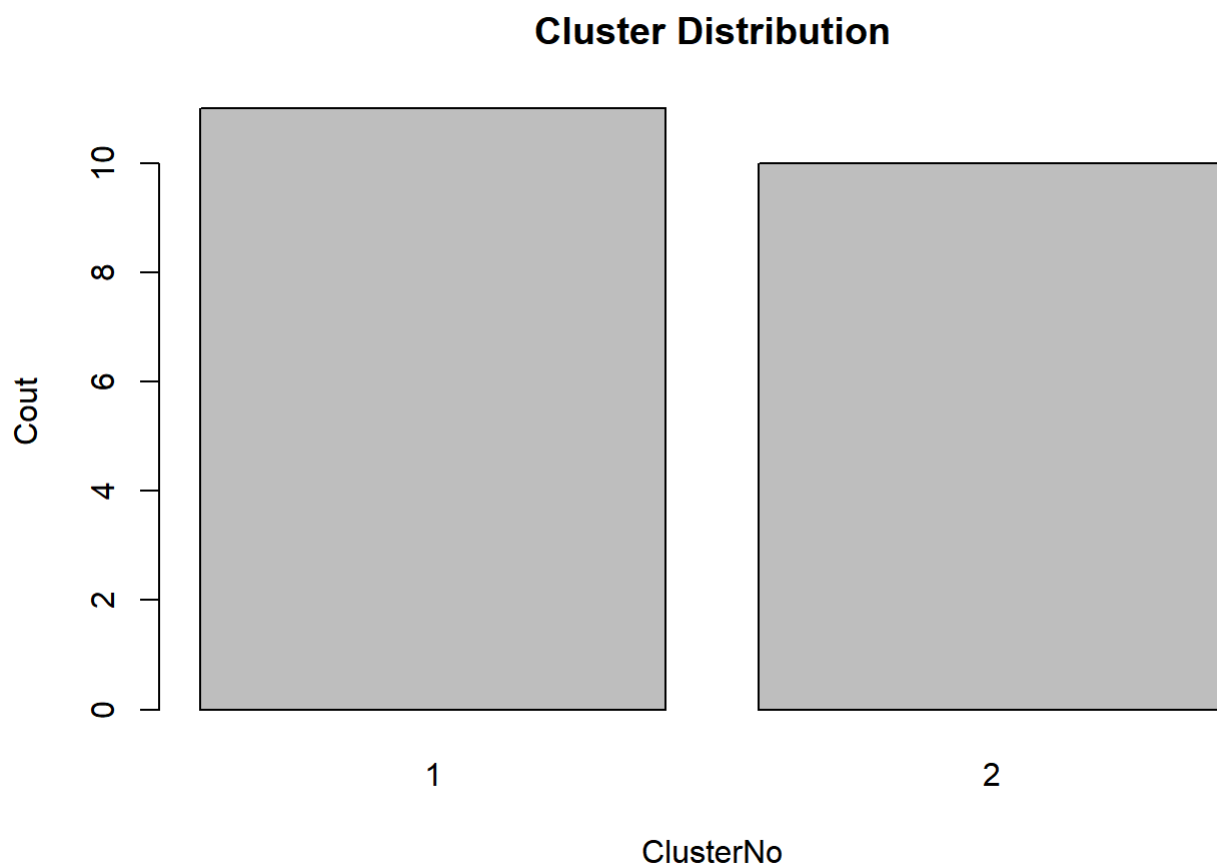
```
ph.cluster2$size      # Number of companies in each cluster
```

```
## [1] 11 10
```

```
ph.cluster2$cluster    # Identify the cluster of all observation
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY  CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##   1   2   2   1   2   2   1   2   2   1   1   2   1   2   1   1
##  PFE  PHA  SGP  WPI  WYE
##   1   2   1   2   1
```

```
cls <- data.frame(ph.cluster2$cluster)
clsdf <- setDT(cls, keep.rownames = TRUE)[,]
colnames(clsdf) <- c("rn", "clusteN")
barplot(table(clsdf$clusteN), main="Cluster Distribution", xlab="ClusterNo", ylab="Cout")
```



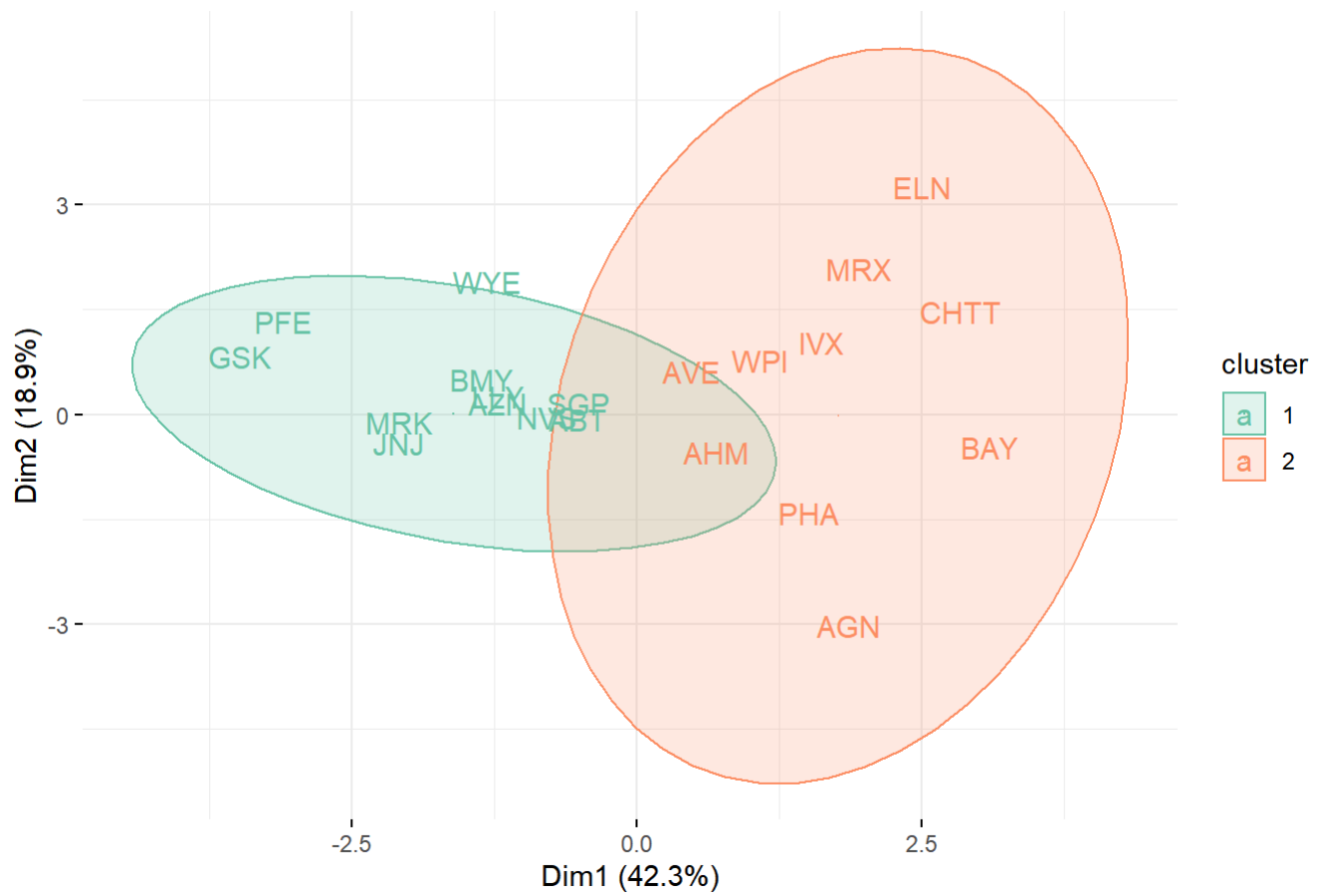
```
library(factoextra)
```

```
ph.cluster2$cluster
```

```
## ABT AGN AHM AZN AVE BAY BMY CHTT ELN LLY GSK IVX JNJ MRX MRK NVS
## 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1
## PFE PHA SGP WPI WYE
## 1 2 1 2 1
```

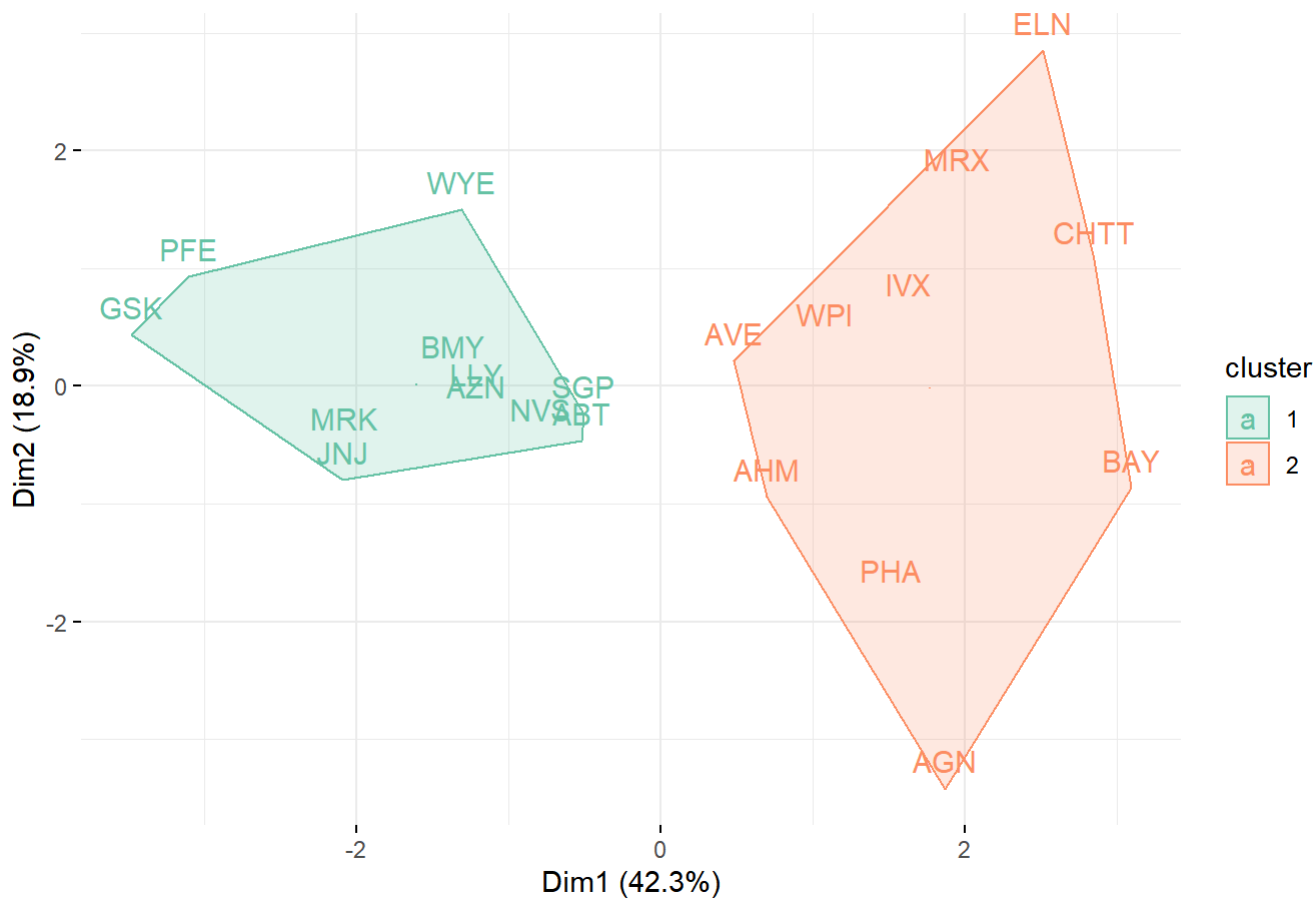
```
fviz_cluster(ph.cluster2, data, ellipse.type = "norm", geom = "text" ,
              palette = "Set2", ggtheme = theme_minimal())
```

Cluster plot



```
fviz_cluster(ph.cluster2, data,  
              palette = "Set2", ggtheme = theme_minimal(), geom = "text" )
```

Cluster plot



head(Ph.data)

Sym...	Name	Market_Cap	B...	PE_Ratio	R...	R...	Asset_Turnover	Leverage
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ABT	Abbott Laboratories	68.44	0.32	24.7	26.4	11.8	0.7	0.42
AGN	Allergan, Inc.	7.58	0.41	82.5	12.9	5.5	0.9	0.60
AHM	Amersham plc	6.30	0.46	20.7	14.9	7.8	0.9	0.27
AZN	AstraZeneca PLC	67.63	0.52	21.5	27.4	15.4	0.9	0.00
AVE	Aventis	47.16	0.32	20.1	21.8	7.5	0.6	0.34
BAY	Bayer AG	16.90	1.11	27.9	3.9	1.4	0.6	0.00

6 rows | 1-10 of 14 columns

```
datadf <- setDT(Ph.data, keep.rownames = TRUE)[  
#cl.data <- datadf %>% merge(datadf, clsdf, by="rn", all = TRUE)  
cl.data <- cbind(datadf, clsdf)  
result <- cl.data[, -c(1,16)]  
result
```

Sym... <chr>	Name <chr>	Market_Cap <dbl>	B... <dbl>	PE_Ratio <dbl>	R... <dbl>	R... <dbl>	Asset_T
ABT	Abbott Laboratories	68.44	0.32	24.7	26.4	11.8	
AGN	Allergan, Inc.	7.58	0.41	82.5	12.9	5.5	
AHM	Amersham plc	6.30	0.46	20.7	14.9	7.8	
AZN	AstraZeneca PLC	67.63	0.52	21.5	27.4	15.4	
AVE	Aventis	47.16	0.32	20.1	21.8	7.5	
BAY	Bayer AG	16.90	1.11	27.9	3.9	1.4	
BMY	Bristol-Myers Squibb Company	51.33	0.50	13.9	34.8	15.1	
CHTT	Chattem, Inc	0.41	0.85	26.0	24.1	4.3	
ELN	Elan Corporation, plc	0.78	1.08	3.6	15.1	5.1	
LLY	Eli Lilly and Company	73.84	0.18	27.9	31.0	13.5	
1-10 of 21 rows 1-8 of 15 columns				Previous	1	2	3 Next

```
#writing the result file.
write_csv2(result,file="result.csv" )
```

```
#library(GGally)

#ggparcoord(result2,
#  columns = 1:9, groupColumn = 10
#  )
```

```
knitr::include_graphics('Assignment4_Dashboard_NonNumeric_vs_Cluster_relation.png')
```

Non-Numeric Vs Cluster N

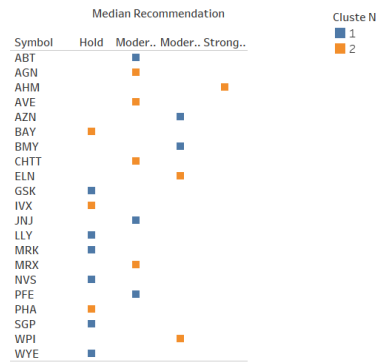
Cluster 1:

- Companies trading only in the NYSE.
- Companies with High Market Cap.

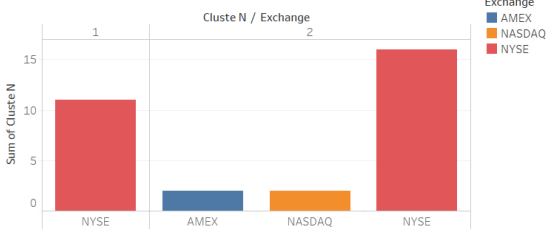
Cluster 2:

- Companies trading in AMEX, NASDAQ and NYSE; all of them
- Each country's company exists in cluster 2.
- Consists a company with strong buy recommendation.
- Companies with lower market cap.

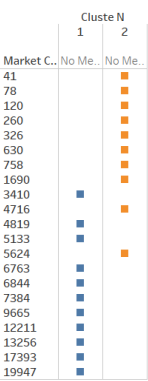
Symbol/Recommendation



Cluster N/ Exchange



Cluster N Vs Market Cap



Cluster N/ Country

