FML_Assignment4

Thanasit C.

2024-03-21

1. Summary

The dataset contains multiple financial ratios of 21 pharmaceutical stocks. The clustering process is used to differentiate stocks into 5 clusters, including 'Overpriced stocks', 'Start-up stocks', 'Cash cow stocks', 'Growth stocks', and 'Best stocks.' Each cluster has its own characteristics, details can be found in Section 5.
I performed both Euclidean and Manhattan K-means, but only the Euclidean calculation is used in all explanation sections because the results are much better compared to the Manhattan calculation. The optimal K is equal to 5, calculated using the 'WSS' and 'Silhouette' methods.

2. Library

```
library(dplyr)
library(tidyr)
library(factoextra)
library(flexclust)
library(caret)
```

3. Import Data

```
## 3.1. Set working directory
setwd("/Users/sieng/Documents/Study/MS.Business Analytics/SPRING
2024/Fundamental of Machine Learning/Assignment/Assignment 4")

## 3.2. Import csv file as dataframe format
maindf <- read.csv("Pharmaceuticals.csv") %>% as.data.frame()

## 3.3. Check data structure
str(maindf)

## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc."
"Amersham plc" "AstraZeneca PLC" ...
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08
0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6
27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1
31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5
```

```
...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53
...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3
23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy"
"Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

4. Data Manipulation

4.1 Handle missing value

```
# 1) Find N/A value
sumna <- sum(is.na(maindf))
print("Number of N/A values in data set")
sumna

colsumna <- colSums(is.na(maindf))
print("Number of N/A by column")
colsumna

## [1] "Number of N/A values in data set"
## [1] 0
## [1] "Number of N/A by column"
##                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
```

4.2 Reassign data attributes.

```
# 4.2 correcting data attributes
## 1).number()/integer() ###############

## 2).factor() ###############
maindf$Symbol <- factor(maindf$Symbol)
maindf$Name <- factor(maindf$Name)
maindf$Median_Recommendation <- factor(maindf$Median_Recommendation, levels =
c("Strong Buy", "Moderate Buy", "Hold", "Moderate Sell", "Strong Sell"))
```

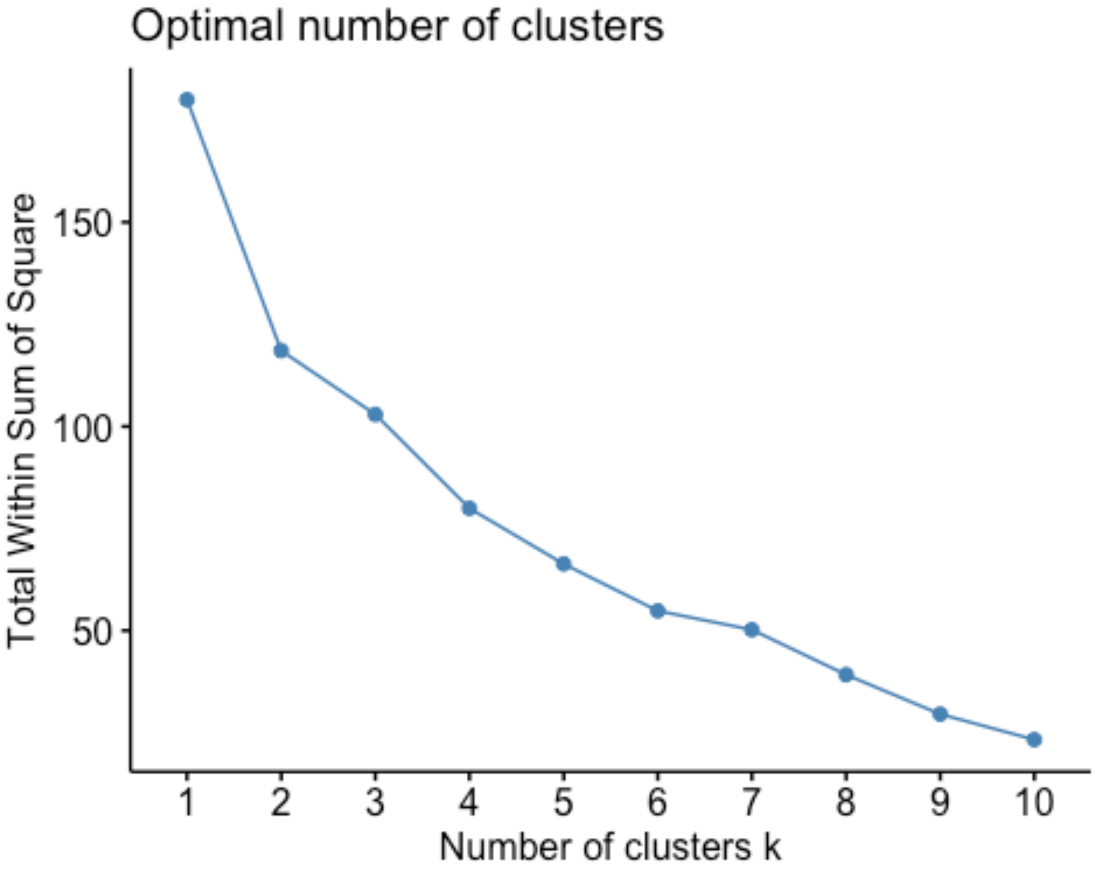
```
maindf$Location <- factor(maindf$Location)
maindf$Exchange <- factor(maindf$Exchange)

str(maindf)

## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : Factor w/ 21 levels "ABT","AGN","AHM",..: 1 2 3
5 4 6 7 8 9 13 ...
##  $ Name                 : Factor w/ 21 levels "Abbott Laboratories",..: 1
2 3 4 5 6 7 8 9 10 ...
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08
0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6
27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1
31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5
...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53
...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3
23.4 ...
##  $ Median_Recommendation: Factor w/ 5 levels "Strong Buy","Moderate
Buy",..: 2 2 1 4 2 3 4 2 4 3 ...
##  $ Location             : Factor w/ 7 levels "CANADA","FRANCE",..: 7 1 6 6
2 3 7 7 4 7 ...
##  $ Exchange             : Factor w/ 3 levels "AMEX","NASDAQ",..: 3 3 3 3 3
3 3 2 3 3 ...

summary(maindf)

##      Symbol                    Name      Market_Cap          Beta
##  ABT    : 1   Abbott Laboratories: 1   Min.   :  0.41   Min.   :0.1800
##  AGN    : 1   Allergan, Inc.     : 1   1st Qu.:  6.30   1st Qu.:0.3500
##  AHM    : 1   Amersham plc       : 1   Median : 48.19   Median :0.4600
##  AVE    : 1   AstraZeneca PLC    : 1   Mean   : 57.65   Mean   :0.5257
##  AZN    : 1   Aventis            : 1   3rd Qu.: 73.84   3rd Qu.:0.6500
##  BAY    : 1   Bayer AG           : 1   Max.   :199.47   Max.   :1.1100
##  (Other):15   (Other)            :15
##     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      Strong Buy   :1         CANADA     : 1
##  1st Qu.: 6.38   1st Qu.:11.2      Moderate Buy :7         FRANCE     : 1
##  Median : 9.37   Median :16.1      Hold         :9         GERMANY    : 1
##  Mean   :13.37   Mean   :15.7      Moderate Sell:4         IRELAND    : 1
##  3rd Qu.:21.87   3rd Qu.:21.1      Strong Sell  :0         SWITZERLAND: 1
##  Max.   :34.21   Max.   :25.5                              UK         : 3
##                                                            US         :13
##    Exchange
##  AMEX   : 1
##  NASDAQ: 1
##  NYSE   :19
##
##
##
##
```

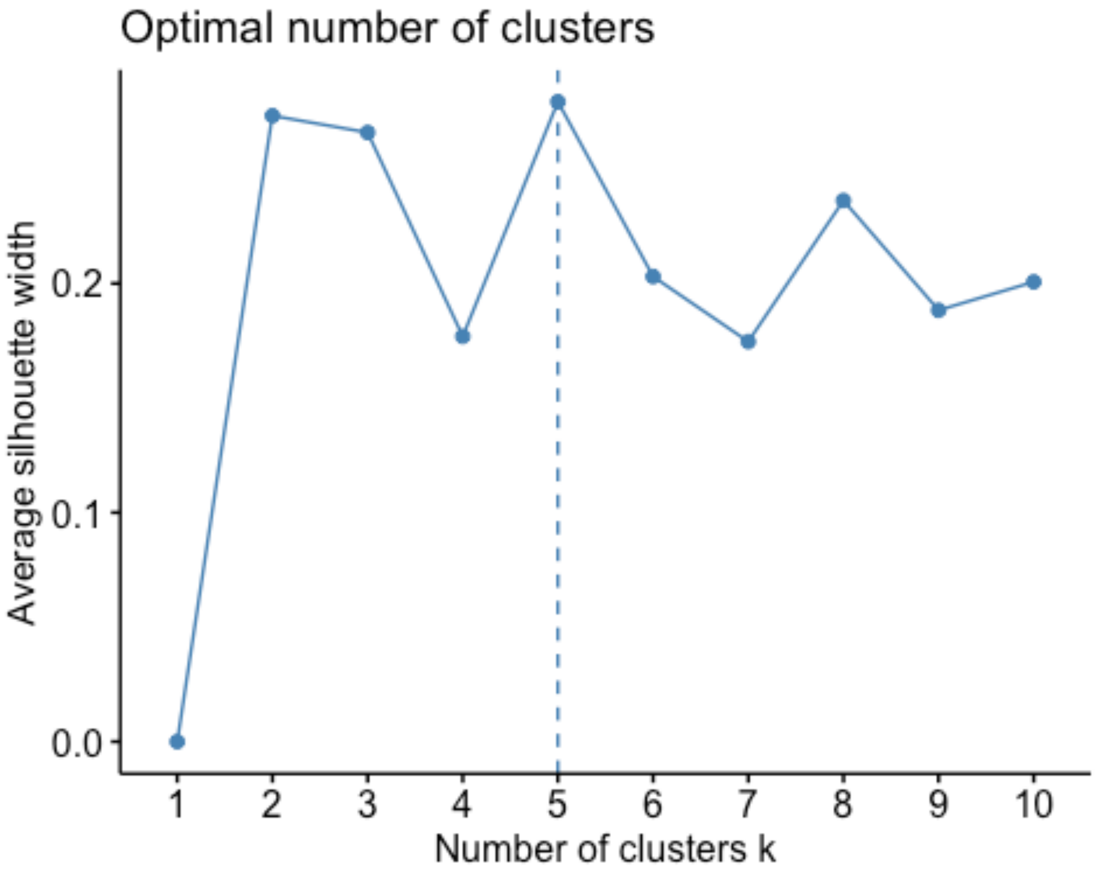
5. Question and Analysis

5.1 Question_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.
Answer_A; I performed clustering by using both Euclidean and Manhattan methods. The optimal K is equal to 5 based on ‘Silhouette’ method, unfortunately, ‘Wss’ method is not provide a clear result.

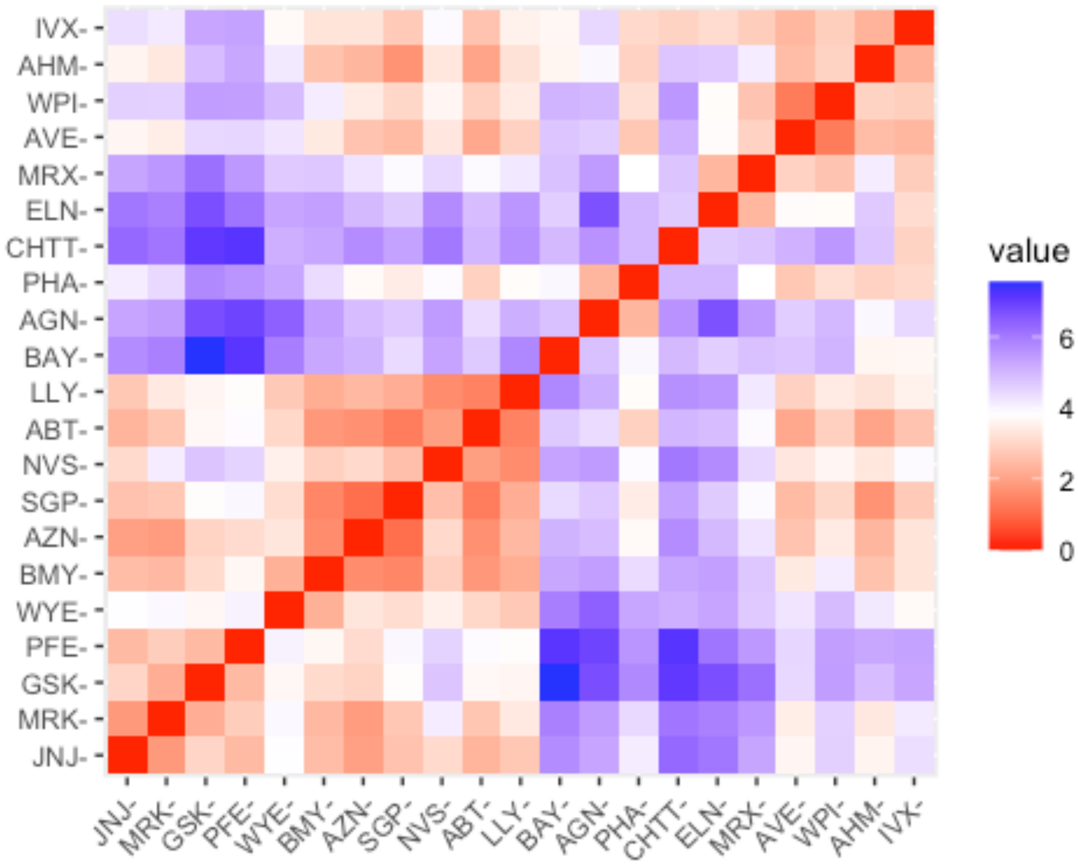
```r
## Answer_A
set.seed(22)
## 1) Data selection
QA_data <- maindf %>%
             tibble::column_to_rownames("Symbol") %>%
             select(2:10)

## 2) Normalization
QA_norm_process <- caret::preProcess(QA_data, method = c("center", "scale"))
QA_data_norm <- predict(QA_norm_process, QA_data)

## 3) Find optimal K
## Method = wss
fviz_nbclust(QA_data_norm, FUNcluster = kmeans, method = "wss")
```



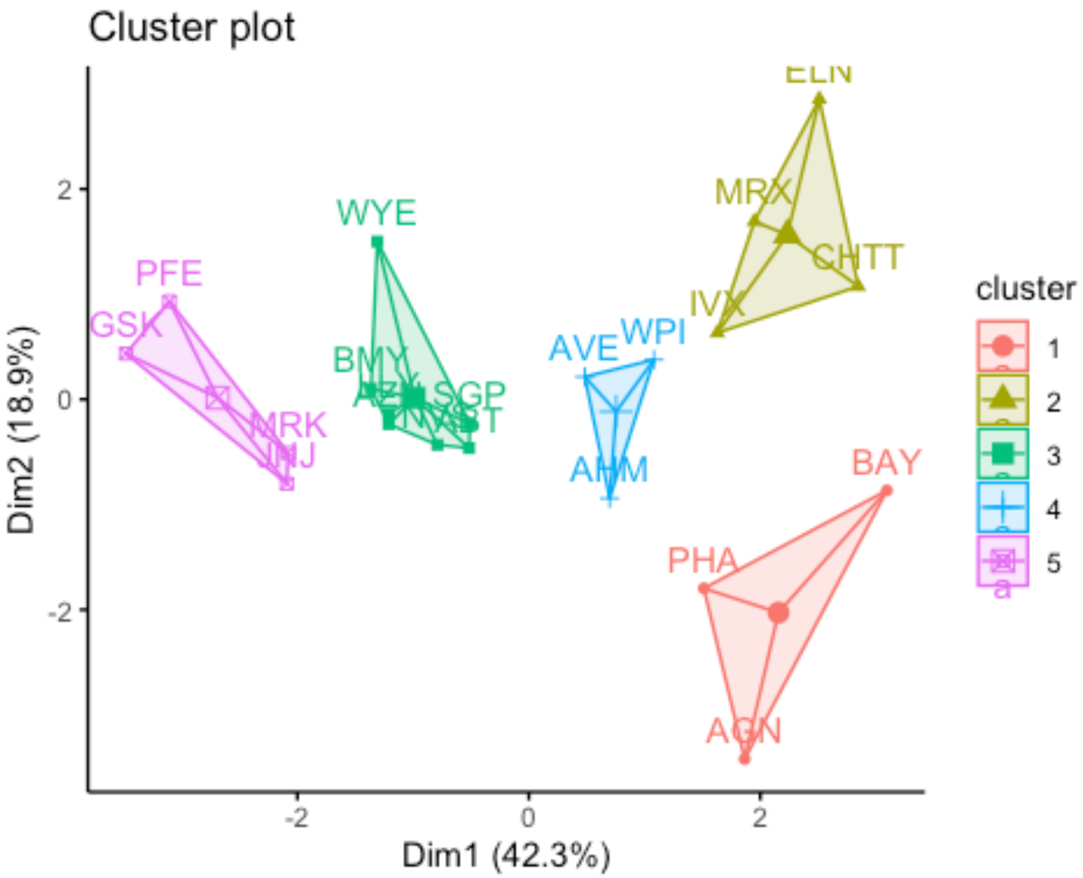
```
### Method = silhouette
fviz_nbclust(QA_data_norm, FUNcluster = kmeans, method = "silhouette")
```



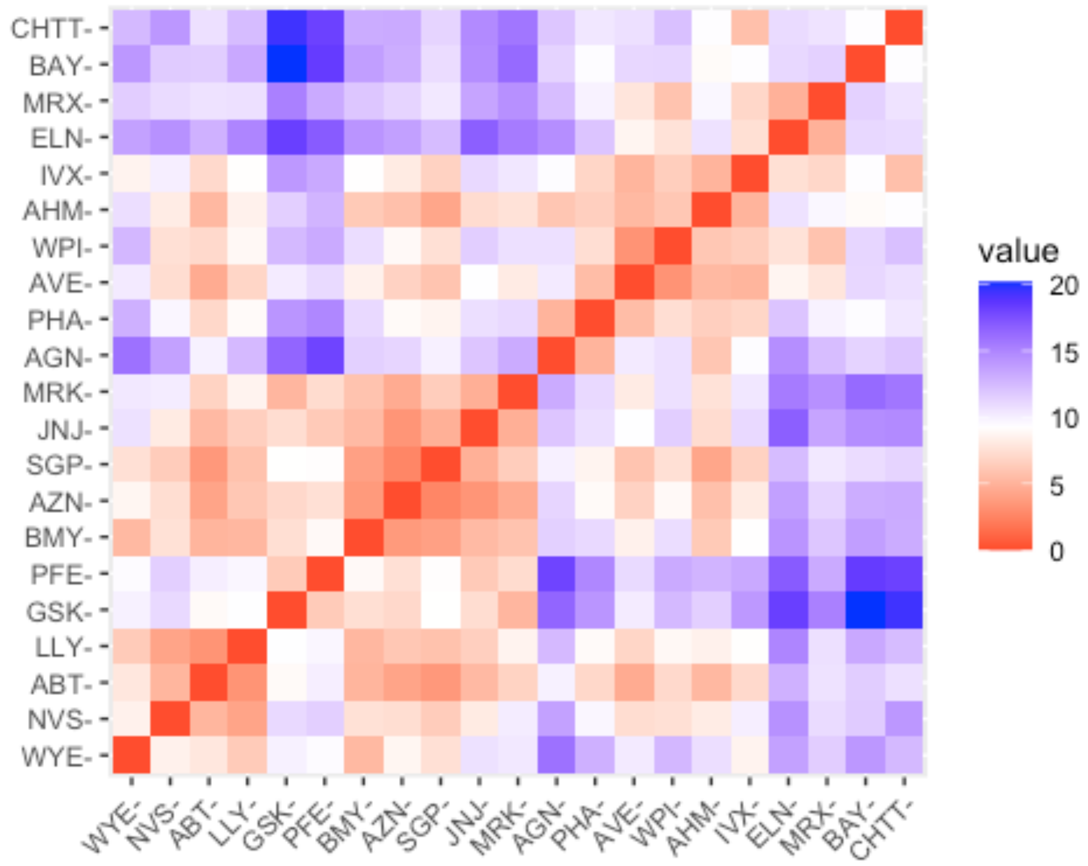
```r
## 4) K-means: Euclidean
### Distance Matrix
QA_data_dist_euc <- dist(QA_data_norm, method = "euclidean")
#as.matrix(QQA_data_dist_euc)
fviz_dist(QA_data_dist_euc)
```



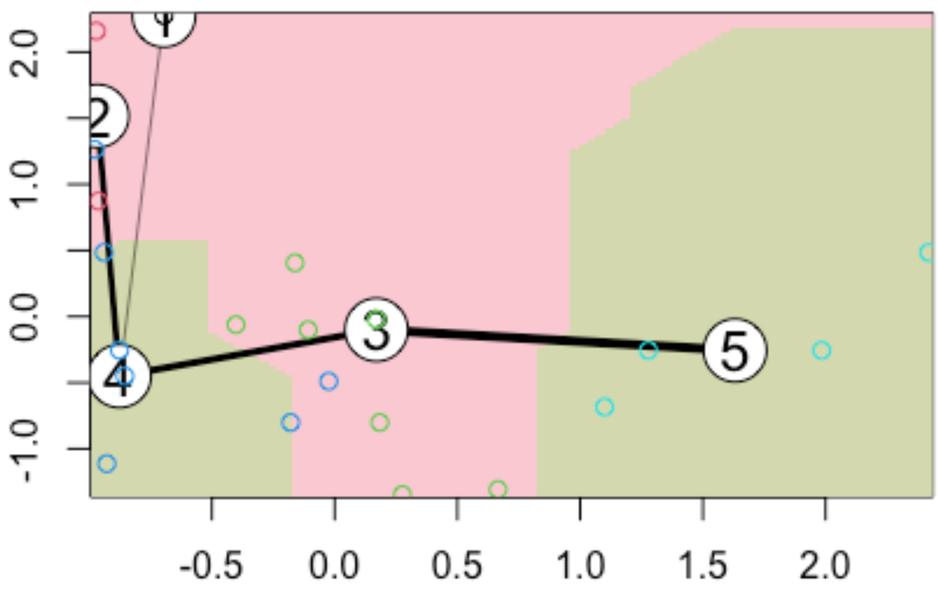
```
## Optimal K = 5
set.seed(22)
QA_Kmean_Euc_opt <- kmeans(QA_data_norm, centers = 5)
fviz_cluster(QA_Kmean_Euc_opt, data = QA_data_norm, ggtheme =
theme_classic(), star.plot = TRUE)
```



```
## 5) K-means: Manhattan
### Distance Matrix
QA_data_dist_man <- dist(QA_data_norm, method = "manhattan")
#as.matrix(QA_data_dist_man)
fviz_dist(QA_data_dist_man)
```



```
## Optimal K = 5
set.seed(22)
QA_Kmean_Man_opt <- kcca(QA_data_norm, k = 5, family =
kccaFamily("kmedians"))
Man_cluster_index <- predict(QA_Kmean_Man_opt)
image(QA_Kmean_Man_opt)
points(QA_data_norm, col = Man_cluster_index)
```



5.2 Question_B: Interpret the clusters with respect to the numerical variables used in forming the clusters.

```
QB_data <- QA_data %>%
             mutate(EucPrediction = QA_Kmean_Euc_opt$cluster) %>%
             mutate(ManPrediction = Man_cluster_index) %>%
             arrange(EucPrediction)
QB_data
```

```
##      Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## AGN        7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## BAY       16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
## PHA       56.24 0.40     56.5 13.5  5.7            0.6     0.35      15.00
## CHTT       0.41 0.85     26.0 24.1  4.3            0.6     3.51       6.38
## ELN        0.78 1.08      3.6 15.1  5.1            0.3     1.07      34.21
## IVX        2.60 0.65     19.9 21.4  6.8            0.6     1.45      13.99
## MRX        1.20 0.75     28.6 11.2  5.4            0.3     0.93      30.37
## ABT       68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## AZN       67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## BMY       51.33 0.50     13.9 34.8 15.1            0.9     0.57       2.70
## LLY       73.84 0.18     27.9 31.0 13.5            0.6     0.53       6.21
## NVS       96.65 0.19     21.6 17.9 11.2            0.5     0.06      -2.69
## SGP       34.10 0.51     18.9 22.6 13.3            0.8     0.00       8.56
## WYE       48.19 0.63     13.1 54.9 13.4            0.6     1.12       0.36
## AHM        6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## AVE       47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## WPI        3.26 0.24     18.4 10.2  6.8            0.5     0.20      29.18
## GSK      122.11 0.35     18.0 62.9 20.3            1.0     0.34      21.87
## JNJ      173.93 0.46     28.4 28.6 16.3            0.9     0.10       9.37
## MRK      132.56 0.46     18.9 40.6 15.0            1.1     0.28      17.35
## PFE      199.47 0.65     23.6 45.6 19.2            0.8     0.16      25.54
##      Net_Profit_Margin EucPrediction ManPrediction
## AGN                5.5             1             4
## BAY                2.6             1             1
## PHA                7.3             1             4
## CHTT               7.5             2             4
## ELN               13.3             2             2
## IVX               11.0             2             4
## MRX               21.3             2             2
## ABT               16.1             3             3
## AZN               18.0             3             3
## BMY               20.6             3             3
## LLY               23.4             3             3
## NVS               22.4             3             3
## SGP               17.6             3             3
## WYE               25.5             3             3
## AHM               11.2             4             4
## AVE               12.9             4             4
## WPI               15.1             4             4
## GSK               21.1             5             5
## JNJ               17.9             5             5
## MRK               14.1             5             5
## PFE               25.2             5             5

## [1] "Answer_B"

## [1] "In Question_A, I performed both Euclidean and Manhattan K-means, but
the results of Euclidean are much better, as shown in the table above. Below,
all explanations are based on the Euclidean methodology. Since the optimal K
equals 5, there are 5 clusters."

## [1] "The first cluster (1) includes AGN, BAY, and PHA, considered as the
highest-risk companies due to their high PE ratio, low ROE and ROA, and low
profit margin. Investing in this group is not a good choice since it is
overpriced (PE ratio is too high) and has low profitability."

## [1] "The second cluster, 2, comprises high-risk companies including CHITT,
ELN, IVX, and MRX. This group has a very small market capitalization,
relatively high beta, high leverage ratio, high revenue growth, and a decent
net profit margin. Companies in this group are small companies in an
expansion phase with a high leverage level but still have a great chance for
growth."
```

```
## [1] "The third group, 3, comprises moderate-risk companies including ABT,
AZN, BMY, LLY, NVS, SGP, and WYE. These are mid-cap companies with moderate
beta, PE ratio, good ROE, ROA, and asset turnover. The net profit margin is
high; however, revenue growth is relatively low. This group of companies is
characterized as cash cows in a mature phase with low potential for growth
but capable of generating a lot of cash."

## [1] "The fourth group, 4, comprises mid-cap companies with a strong
financial standing in every aspect, including AHM, APE, and WPI. They exhibit
moderate PE ratios and beta, good ROE, asset turnover, and leverage levels.
Although their ROA is somewhat low, they demonstrate high revenue growth and
net profit margins."

## [1] "Lastly, group 5 includes GSK, JNJ, MRK, and PFE. This group comprises
large-cap companies with excellent financial statuses. Investing in this
group is recommended; the stock prices are at fair value (moderate PE ratio),
and there is high potential for growth in the future."
```

5.3 Question_C: Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters).

```
QC_data <- maindf %>%
             tibble::column_to_rownames("Symbol") %>%
             select(-1) %>%
             mutate(EucPrediction = QA_Kmean_Euc_opt$cluster) %>%
             arrange(EucPrediction)

table(QC_data$Median_Recommendation, QC_data$EucPrediction)

##
##                   1 2 3 4 5
##   Strong Buy      0 0 0 1 0
##   Moderate Buy    1 2 1 1 2
##   Hold            2 1 4 0 2
##   Moderate Sell   0 1 2 1 0
##   Strong Sell     0 0 0 0 0

table(QC_data$Location, QC_data$EucPrediction)

##
##                 1 2 3 4 5
##   CANADA        1 0 0 0 0
##   FRANCE        0 0 0 1 0
##   GERMANY       1 0 0 0 0
##   IRELAND       0 1 0 0 0
##   SWITZERLAND   0 0 1 0 0
##   UK            0 0 1 1 1
##   US            1 3 5 1 3

table(QC_data$Exchange, QC_data$EucPrediction)
```

```
##
##          1 2 3 4 5
##   AMEX   0 1 0 0 0
##   NASDAQ 0 1 0 0 0
##   NYSE   3 2 7 3 4
```

```
## [1] "Answer_C"
```

```
## [1] "I created a tabular representation of the data across three
categories and cluster predictions by K-means. The Recommendation and
Prediction table might be somewhat challenging to understand due to the
spread of the data; however, we can observe that most of the stocks are
categorized as 'Hold' or 'Moderate Buy.' Out of 21 stocks, 13 are from the US
and are clustered in groups 2, 3, and 5. Stocks from other countries have
fewer representations, some with only one stock, making interpretation
challenging. Lastly, the majority of the stocks are listed on the NYSE, with
only one stock listed on the AMEX and NASDAQ."
```

```
## [1] "The first three tables are difficult to understand, so grouping the
data will aid in interpretation. Since the 'Exchange' variable doesn't
provide significant information, I decided to drop it. Then, I grouped
recommendations from 5 levels to 3 levels, including 'Buy,' 'Hold,' and
'Sell.' Lastly, regarding the country category, I grouped 'CANADA' with 'US'
as 'US' and grouped the other countries as 'EURO'."
```

```
QC_data2 <- QC_data %>%
              mutate(Median_Recommendation = gsub("Moderate ", "",
Median_Recommendation),
                     Median_Recommendation = gsub("Strong ", "",
Median_Recommendation)) %>%
              mutate(Location = gsub("CANADA", "US", Location),
                     Location = gsub("FRANCE", "EURO", Location),
                     Location = gsub("GERMANY", "EURO", Location),
                     Location = gsub("IRELAND", "EURO", Location),
                     Location = gsub("SWITZERLAND", "EURO", Location),
                     Location = gsub("UK", "EURO", Location))

table(QC_data2$Median_Recommendation, QC_data2$EucPrediction)
```

```
##
##        1 2 3 4 5
##   Buy  1 2 1 2 2
##   Hold 2 1 4 0 2
##   Sell 0 1 2 1 0
```

```
table(QC_data2$Location, QC_data2$EucPrediction)
```

```
##
##        1 2 3 4 5
##   EURO 1 1 2 2 1
##   US   2 3 5 1 3
```

```
table(QC_data2$Location, QC_data2$Median_Recommendation)

## 
##        Buy Hold Sell
##   EURO   2    3    2
##   US     6    6    2

## [1] "New three tables that I created are much easier to understand."

## [1] "First, the table between 'Recommendation' and 'Clustered Prediction.'
There are 'Buy' recommendations in every cluster. Half of the list is 'Hold'
recommended. Lastly, there are a few 'Sell' recommended stocks, which are in
clusters 2, 3, and 4."

## [1] "The second table shows the relationship between 'Country' and
'Recommendation'. Most 'US' stocks are clustered in groups 2, 3, and 5. On
the other hand, 'EURO' stocks are spread equally across all 5 clusters."

## [1] "The last table shows the relationship between 'Country' and
'Recommendation'. It seems that 'US' stocks perform well since most of them
are recommended to 'Buy' or 'Hold'. Conversely, the proportion of 'Sell'
recommendations for 'EURO' stocks is a bit high, at nearly 30%."
```

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

```
QD_data <- QC_data2
QD_data

##      Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## AGN        7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## BAY       16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
## PHA       56.24 0.40     56.5 13.5  5.7            0.6     0.35      15.00
## CHTT       0.41 0.85     26.0 24.1  4.3            0.6     3.51       6.38
## ELN        0.78 1.08      3.6 15.1  5.1            0.3     1.07      34.21
## IVX        2.60 0.65     19.9 21.4  6.8            0.6     1.45      13.99
## MRX        1.20 0.75     28.6 11.2  5.4            0.3     0.93      30.37
## ABT       68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## AZN       67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## BMY       51.33 0.50     13.9 34.8 15.1            0.9     0.57       2.70
## LLY       73.84 0.18     27.9 31.0 13.5            0.6     0.53       6.21
## NVS       96.65 0.19     21.6 17.9 11.2            0.5     0.06      -2.69
## SGP       34.10 0.51     18.9 22.6 13.3            0.8     0.00       8.56
## WYE       48.19 0.63     13.1 54.9 13.4            0.6     1.12       0.36
## AHM        6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## AVE       47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## WPI        3.26 0.24     18.4 10.2  6.8            0.5     0.20      29.18
## GSK      122.11 0.35     18.0 62.9 20.3            1.0     0.34      21.87
## JNJ      173.93 0.46     28.4 28.6 16.3            0.9     0.10       9.37
## MRK      132.56 0.46     18.9 40.6 15.0            1.1     0.28      17.35
## PFE      199.47 0.65     23.6 45.6 19.2            0.8     0.16      25.54
##      Net_Profit_Margin Median_Recommendation Location Exchange
EucPrediction
## AGN                5.5                   Buy       US     NYSE
1
## BAY                2.6                  Hold     EURO     NYSE
1
## PHA                7.3                  Hold       US     NYSE
1
## CHTT               7.5                   Buy       US   NASDAQ
2
## ELN               13.3                  Sell     EURO     NYSE
2
## IVX               11.0                  Hold       US     AMEX
2
## MRX               21.3                   Buy       US     NYSE
2
## ABT               16.1                   Buy       US     NYSE
3
## AZN               18.0                  Sell     EURO     NYSE
3
## BMY               20.6                  Sell       US     NYSE
3
## LLY               23.4                  Hold       US     NYSE
3
## NVS               22.4                  Hold     EURO     NYSE
3
## SGP               17.6                  Hold       US     NYSE
3
## WYE               25.5                  Hold       US     NYSE
3
## AHM               11.2                   Buy     EURO     NYSE
4
## AVE               12.9                   Buy     EURO     NYSE
4
## WPI               15.1                  Sell       US     NYSE
4
## GSK               21.1                  Hold     EURO     NYSE
5
## JNJ               17.9                   Buy       US     NYSE
5
## MRK               14.1                  Hold       US     NYSE
5
## PFE               25.2                   Buy       US     NYSE
5

## [1] "Answer_D"

## [1] "As the results of clustering process, there are 5 clusters. I've
explained some characteristics of each cluster in the Question_B. Below is an
appropriate name for each cluster."
```

```
## [1] "Cluster 1; 'Overvalued stocks'.  The main characteristic is an
extremely high P/E ratio and low net profit margin. Investors should avoid
this group of stocks."

## [1] "Cluster 2, 'Start-up stocks'. These are small-cap stocks in an
expansion phase, characterized by a high leverage ratio and high growth
potential but low return on assets. Investing in this group of stocks
requires a very careful understanding of the business."

## [1] "Cluster 3, 'Cash cow stocks'. This cluster comprises mid-to-large-cap
stocks with a good profit margin but low potential growth. You can expect a
high dividend yield from this cluster."

## [1] "Cluster 4, 'Growth stocks'. This cluster is a small-to-mid-cap stocks
trade at fair value with a very high growth potential. However, return on
investment or net profit margin seems to be a bit low. Investment in this
group of stock require a consistency in market updates,"

## [1] "Cluster 5, 'Best stocks'. Big-cap stocks trade at fair value, with
low risk (low beta and low leverage), high potential future growth, and a
high profitability ratio. Both dividends and capital growth can be expected;
there are no stocks better than these."
```