MichaelBasta_Assignment_4

Michael Basta

2022-10-30

```
Pharma <- read.csv("/Users/michaelbasta/Documents/Fundmentals of Machine Learning /Module 6/Pharmaceuti
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

```
library(ISLR)
#install.packages("flexclust")
library(flexclust)
```

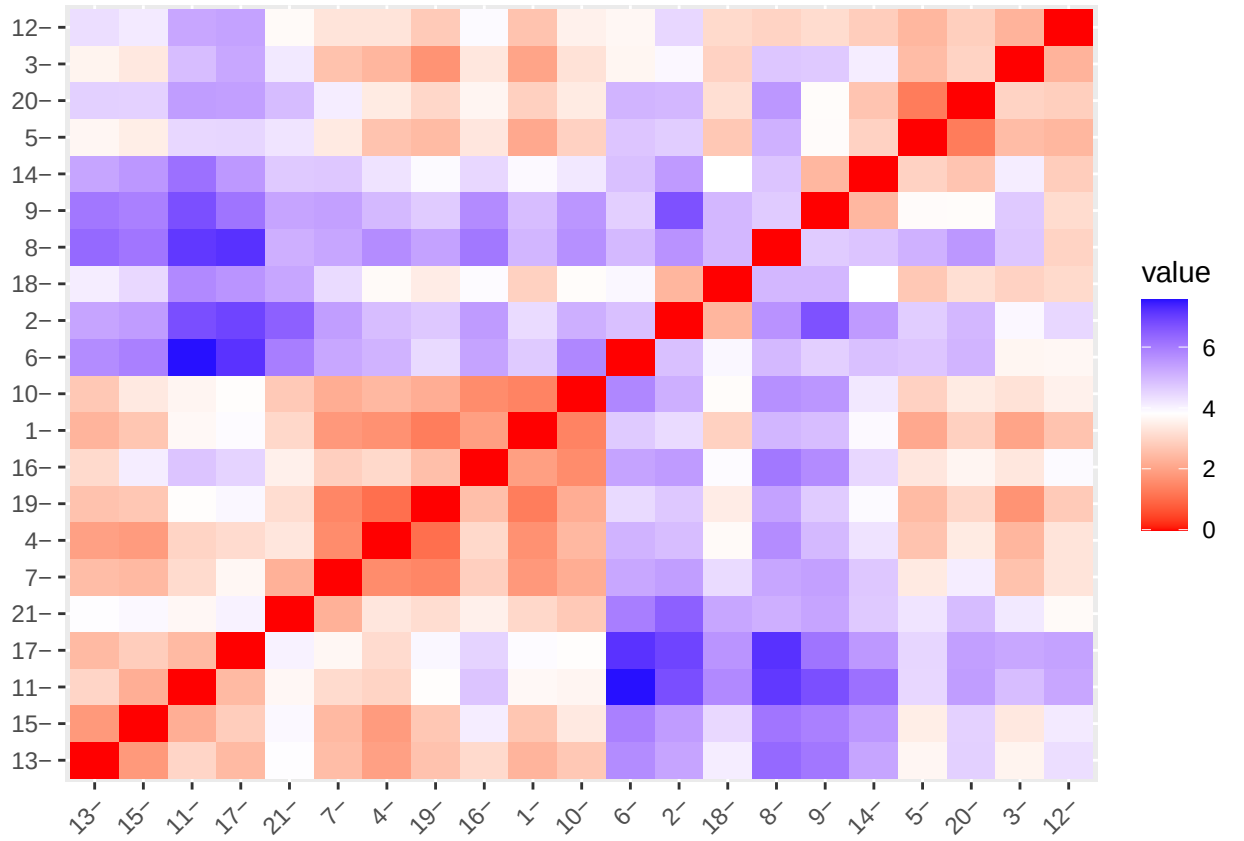
```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
set.seed(123)

# a.

df <- Pharma[,3:11]
summary(df)
```

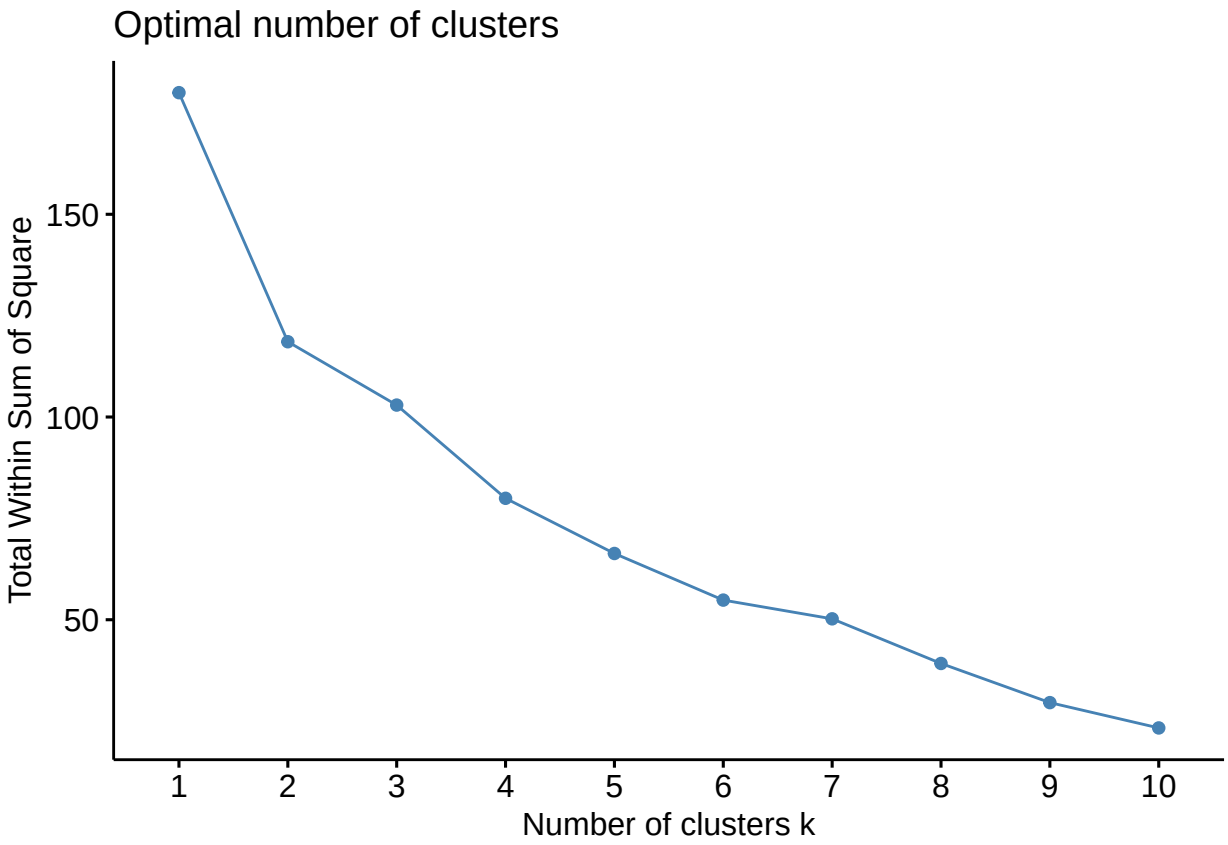
```
##    Market_Cap          Beta           PE_Ratio          ROE
##  Min.   :  0.41   Min.   :0.1800   Min.   : 3.60   Min.   : 3.9
##  1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90   1st Qu.:14.9
##  Median : 48.19   Median :0.4600   Median :21.50   Median :22.6
##  Mean   : 57.65   Mean   :0.5257   Mean   :25.46   Mean   :25.8
##  3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90   3rd Qu.:31.0
##  Max.   :199.47   Max.   :1.1100   Max.   :82.50   Max.   :62.9
```

```
##       ROA         Asset_Turnover    Leverage        Rev_Growth
##  Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17
##  1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38
##  Median :11.20   Median :0.6    Median :0.3400   Median : 9.37
##  Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37
##  3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87
##  Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21
##  Net_Profit_Margin
##  Min.   : 2.6
##  1st Qu.:11.2
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5
```

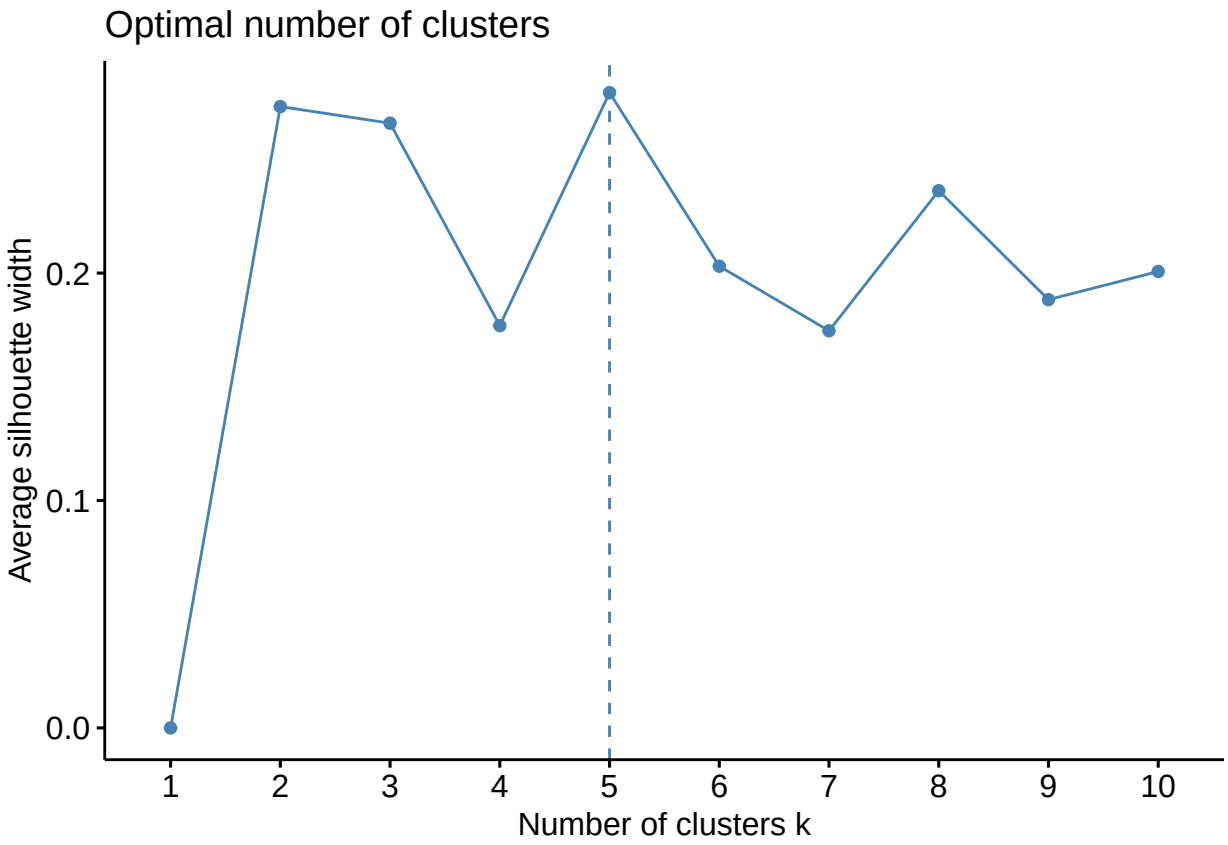
```
# It doesn't seem that there's a direct correlation between variables so Euclidean distance should be s
# Also since it is scale dependant we had to scale it before applying
df <- scale(df)
distance <- get_dist(df)
fviz_dist(distance)
```



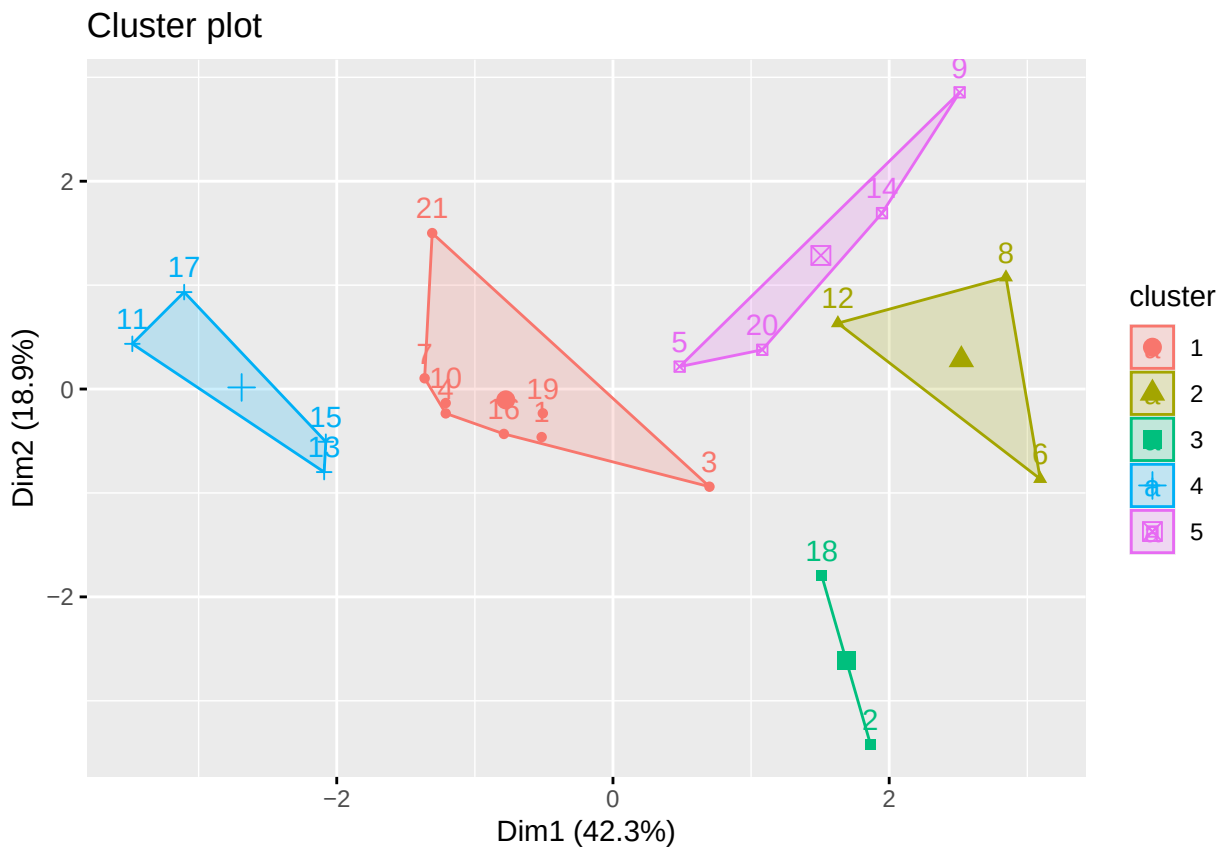

```
fviz_nbclust(df, kmeans, method="wss")
```




```
fviz_nbclust(df, kmeans, method="silhouette")
```




```
# From the earlier step it shows that 5 clusters is the best choice
k4 <- kmeans(df, centers=5, nstart=25)
fviz_cluster(k4, data = df)
```



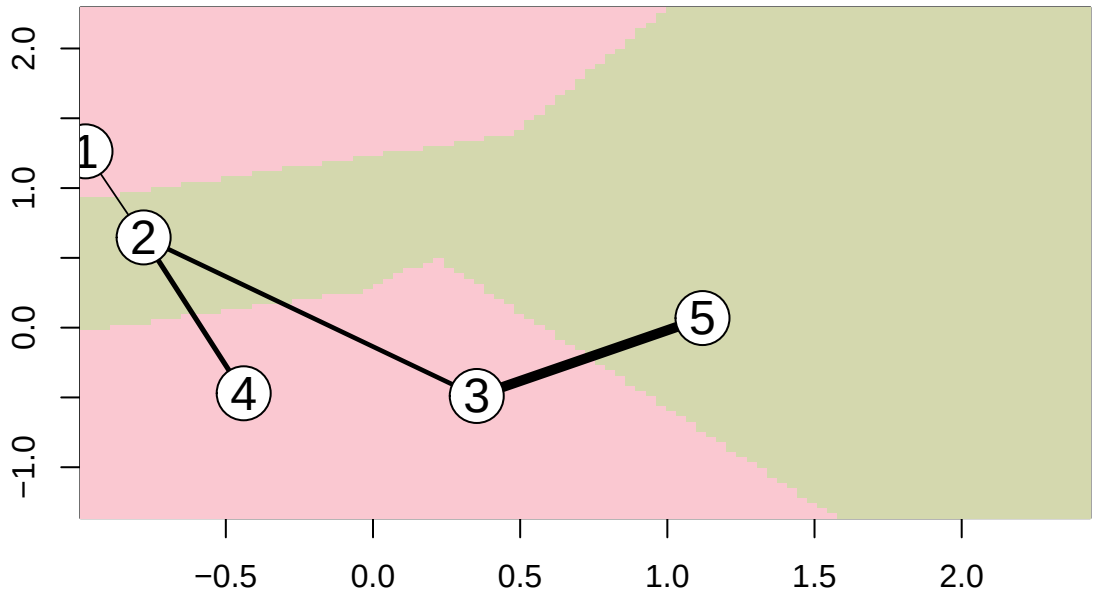

```
k4$size
```

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

```
k4$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
```

```
k4 = kcca(df, k=5, kccaFamily("kmeans"))
clusters_index = predict(k4)
image(k4)
```




```
# b.
## Looking at the data from the centroids of the clusters
## Cluster 1: (Market_cap, Beta, PE_Ratio, Leverage, Rev_Growth) are Lower than average (negative value
## Cluster 2: (Market_cap, PE_Ratio, ROE, ROA, Asset_Turnover, Rev_Growth, Net_Profit_Margin) are Lower
## Cluster 3:(Market_cap, Beta, ROE, ROA, Leverage, Rev_Growth) are Lower than average (negative values
## Cluster 4: (Beta, PE_Ratio, Leverage) are Lower than average (negative values) and (Market_Cap, ROE,
## Cluster 5:(Market_Cap, PE_Ratio, ROE, ROA, Asset_Turnover, Net_Profit_Margin) are Lower than average

# c.
## I don't see a pattern in the clusters with respect to the non-numerical variables

# d.
## Cluster 1: High returns and turnover
## Cluster 2: High leverage
## Cluster 3: High Price/earnings ratio
## Cluster 4: High Market Capitalization and returns
## Cluster 5: High Revenew Growth
```