Assignment_4

Matt Kostoff

2022-10-30

Cluster analysis

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

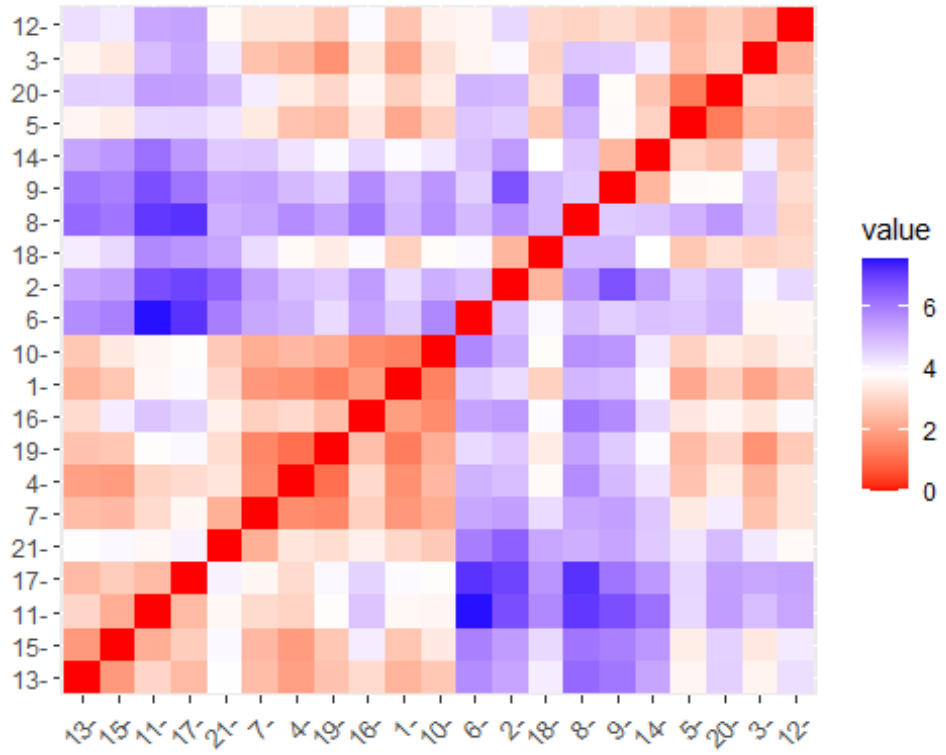
```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
library(ISLR)
set.seed(123)
Pharmaceuticals<-read.csv("Pharmaceuticals.csv")
Pharm.df<-Pharmaceuticals[,c(3,4,5,6,7,8,9,10,11)]
summary(Pharm.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
```

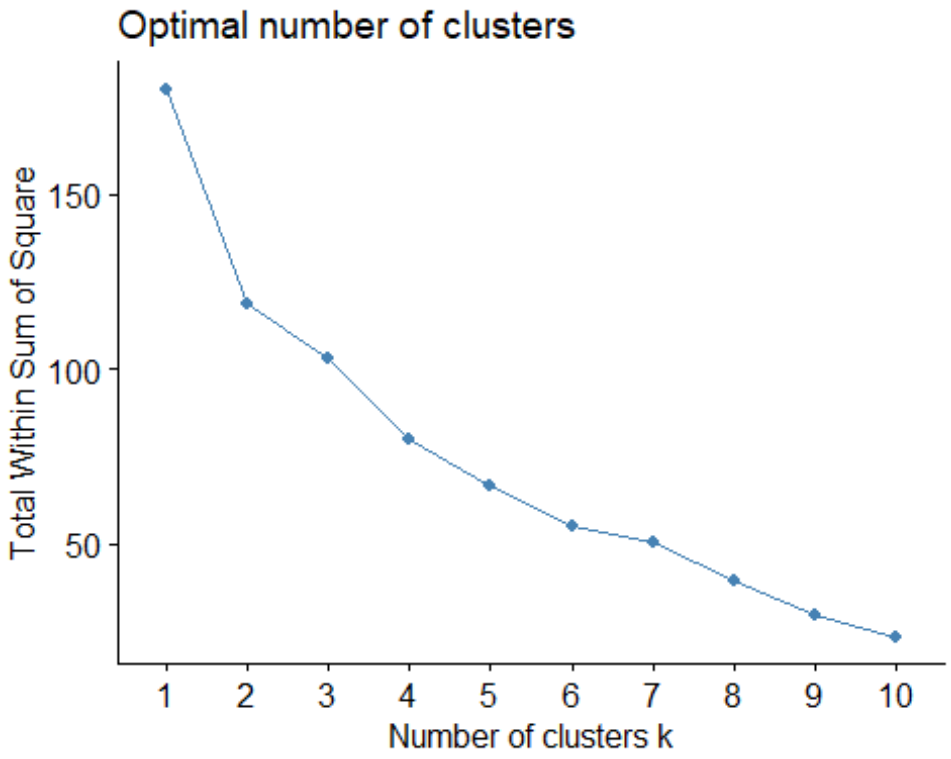
```
##  Median :16.1
##  Mean   :15.7
##  3rd Qu.:21.1
##  Max.   :25.5

# scaling the data frame
view(Pharm.df)
Pharm.df<-scale(Pharm.df)
distance<-get_dist(Pharm.df)
fviz_dist(distance)
```



Determining K

```
# "Elbow" Method
fviz_nbclust(Pharm.df, kmeans, method = "wss")
```




```r
# Average Silhouette Method
fviz_nbclust(Pharm.df, kmeans, method = "silhouette")
```

```
# optimal K = 5
```

K-Means

```
k5<-kmeans(Pharm.df, centers = 5, nstart = 25)
k5$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
```

```
k5$size
```

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

```
fviz_cluster(k5, data = Pharm.df)
```

Other Distances

```
library(flexclust)
```

```
## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

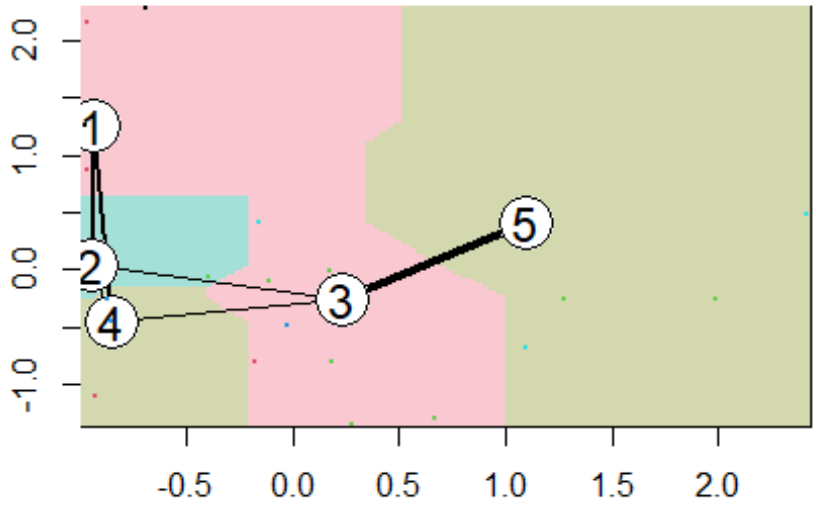
```
k5 = kcca(Pharm.df, k=5, kccaFamily("kmedians"))
k5
```

```
## kcca object of family 'kmedians'
##
## call:
## kcca(x = Pharm.df, k = 5, family = kccaFamily("kmedians"))
##
## cluster sizes:
##
## 1 2 3 4 5
## 3 4 8 3 3
```

```
clusters_index <- predict(k5)
dist(k5@centers)
```

```
##          1        2        3        4
## 2 3.091226
## 3 3.731864 3.691708
## 4 3.278099 3.937881 3.471411
## 5 5.478701 4.991624 2.857072 5.750727
```

```
image(k5)
points(Pharm.df, col=clusters_index, pch=19, cex=0.3)
```



Using numerical variables (1-9), there is no clear pattern in the clusters with respect to variables 10-12