Assignment 4

Brody Coyne

2022-10-31

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## 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
```

Importing the Pharmaceutical data and displaying the first few rows of that data

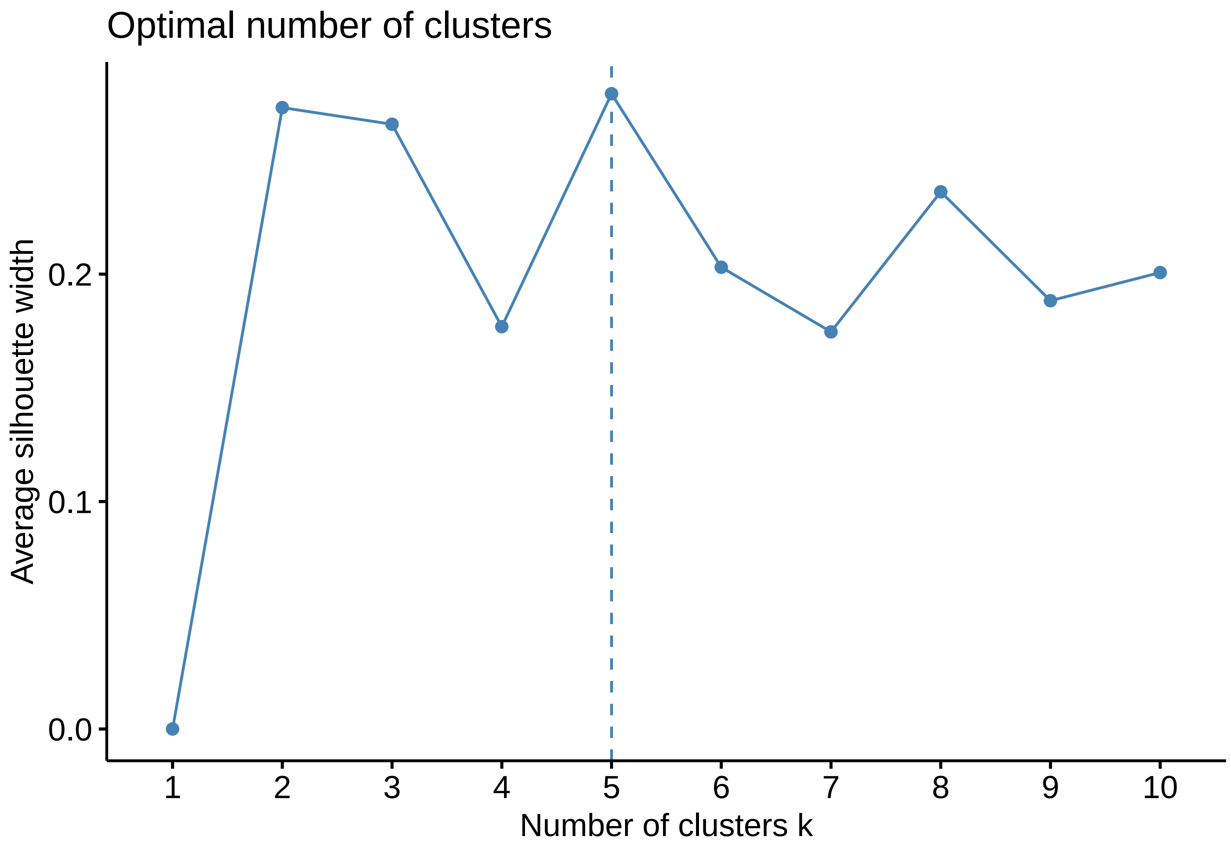
```
mydata<-read.csv("Pharmaceuticals.csv")
head(mydata)
```

```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8            0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5            0.9
## 3    AHM        Amersham plc       6.30 0.46     20.7 14.9  7.8            0.9
## 4    AZN    AstraZeneca PLC       67.63 0.52     21.5 27.4 15.4            0.9
## 5    AVE            Aventis      47.16 0.32     20.1 21.8  7.5            0.6
## 6    BAY            Bayer AG      16.90 1.11     27.9  3.9  1.4            0.6
##   Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location Exchange
## 1     0.42       7.54              16.1         Moderate Buy       US     NYSE
## 2     0.60       9.16               5.5         Moderate Buy   CANADA     NYSE
## 3     0.27       7.05              11.2           Strong Buy       UK     NYSE
## 4     0.00      15.00              18.0        Moderate Sell       UK     NYSE
## 5     0.34      26.81              12.9         Moderate Buy   FRANCE     NYSE
## 6     0.00      -3.17               2.6                 Hold  GERMANY     NYSE
```

```
set.seed(123)
```

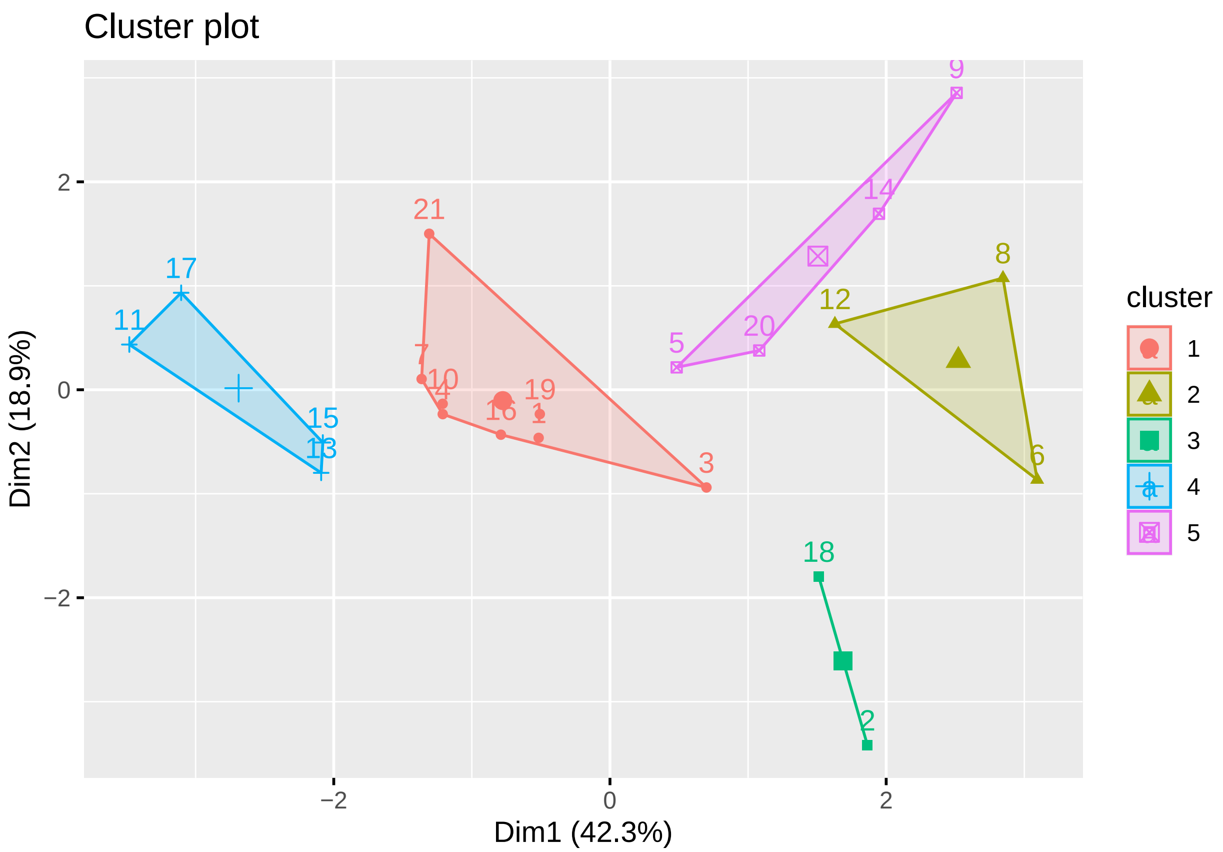
I took out the 9 numerical values from the Pharmaceutical data and then I calculated the optimal number of clusters. I decided to use the silhouette method because it gives an exact number for the optimal number of clusters, which in this case is 5. I also scaled the data so that all of the data would be normalized.

```r
cluster<-mydata[3:11]
cluster<-scale(cluster)
fviz_nbclust(cluster, kmeans, method = "silhouette")
```



I then used kmeans to create 5 clusters based of all of the numerical data. Due to the fact that it is a relatively small data set, I thought that using 25 for the nstart value would be high enough.

```r
k5<-kmeans(cluster, centers = 5, nstart = 25)
fviz_cluster(k5, data = cluster)
```



Then I checked the centers of the clusters as well as the size of the clusters.

```
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
```

The first cluster is the largest with 8 firms in it. I noticed that they are all traded on the NYSE and all either from the US or UK except for one from Switzerland.

The second cluster has firms that are pretty different. They are all traded on different exchanges but they are similar in the fact that there are two holds and one moderately buy.

The third cluster only has 2 firms and I would say that the categorical variables are pretty different. The only similarities are that they are both traded on the NYSE.

The fourth cluster likely has the most similar firms with most of them being US based and traded on the NYSE.

The fifth cluster is relatively different based on the firms with only half of the firms sharing the fact that they are US based and traded on the NYSE.

Based on only the 3 non-numerical variables that I looked at to compare the clusters, there seemed to not be many similarities. However, those three variables don't provide much insight into the company as a whole which means I would have to look into the firms on a deeper level to really test the effectiveness of the clustering.