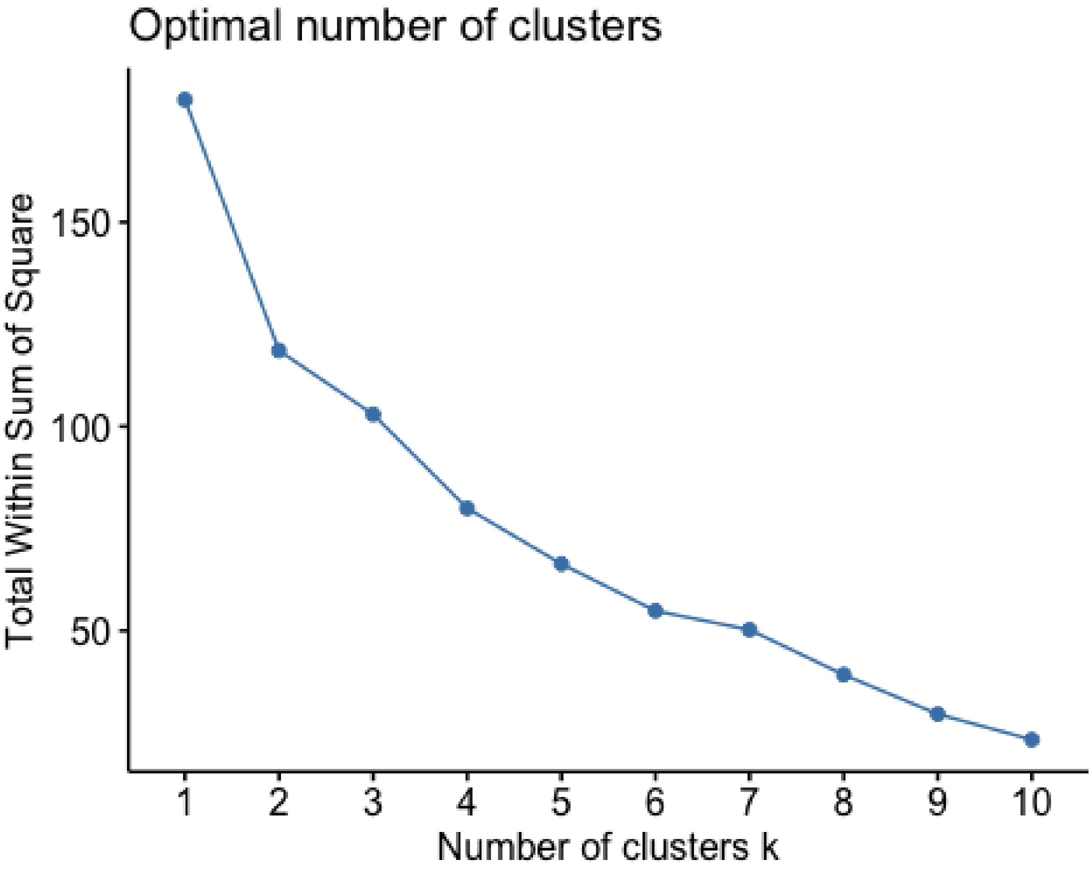
Assignment 4 - Fundamentals of Machine Learning

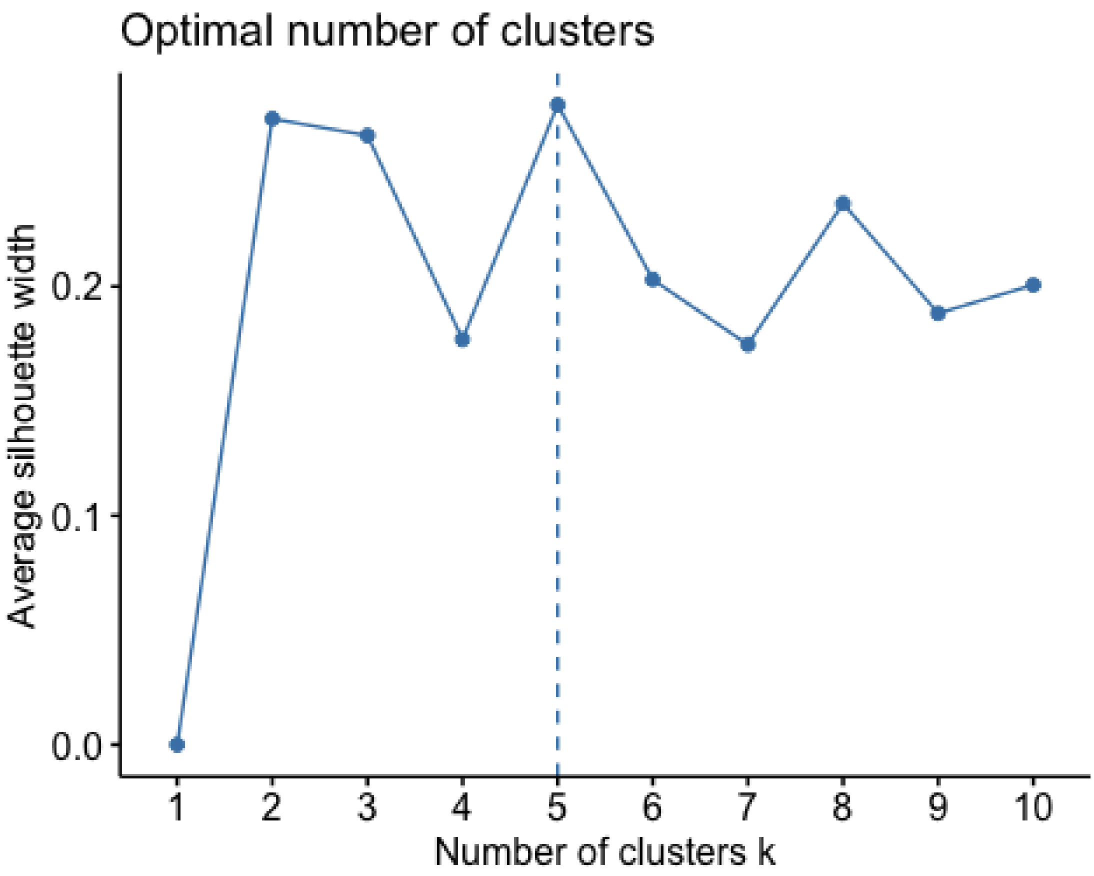
Julia Thacker

10/27/2021

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ISLR)
library(factoextra)

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

library(flexclust)

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

pharmaceuticals<-read.csv("Pharmaceuticals.csv")
set.seed(123)

pharmaceuticalsdf<-pharmaceuticals[,c(3:11)]
pharmaceuticalsdf<-scale(pharmaceuticalsdf)

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

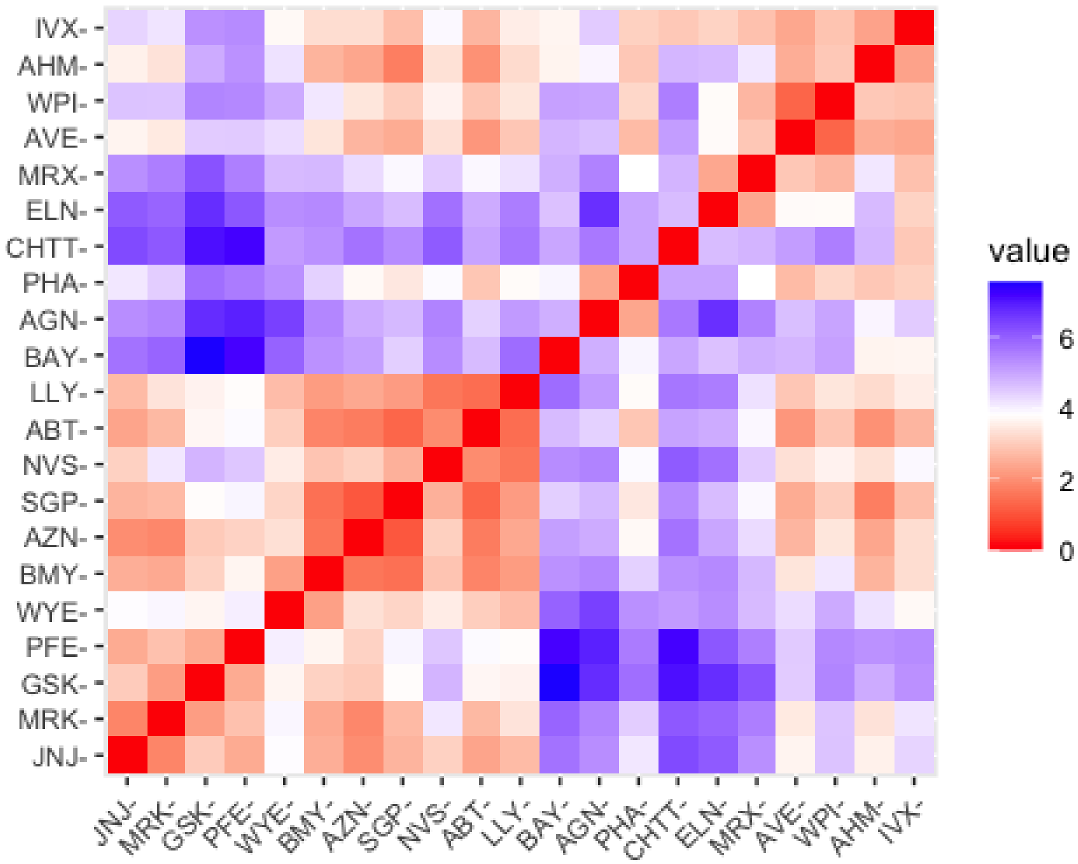


```
fviz_nbclust(pharmaceuticalsdf,kmeans,method="silhouette")
```



Filtered the data to only use the numerical variables, variables 1-9. Normalized the data using scale() and visualized the optimal number of clusters. Selected k=5 because it is the knee point using the wss method and the returned value using the silhouette method.

```
rownames(pharmaceuticalsdf)<-
c("ABT","AGN","AHM","AZN","AVE","BAY","BMY","CHTT","ELN","LLY","GSK","IVX","J
NJ","MRX","MRK","NVS","PFE","PHA","SGP","WPI","WYE")
distance<-get_dist(pharmaceuticalsdf)
fviz_dist(distance)
```



This plot shows correlations between some of the firms which could mean that they will form a cluster. One example would be LLY, ABT, NVS, SGP, AZN, BMY, and WYE, which are all showing a low value.

```
k5<-kmeans(pharmaceuticalsdf,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
```

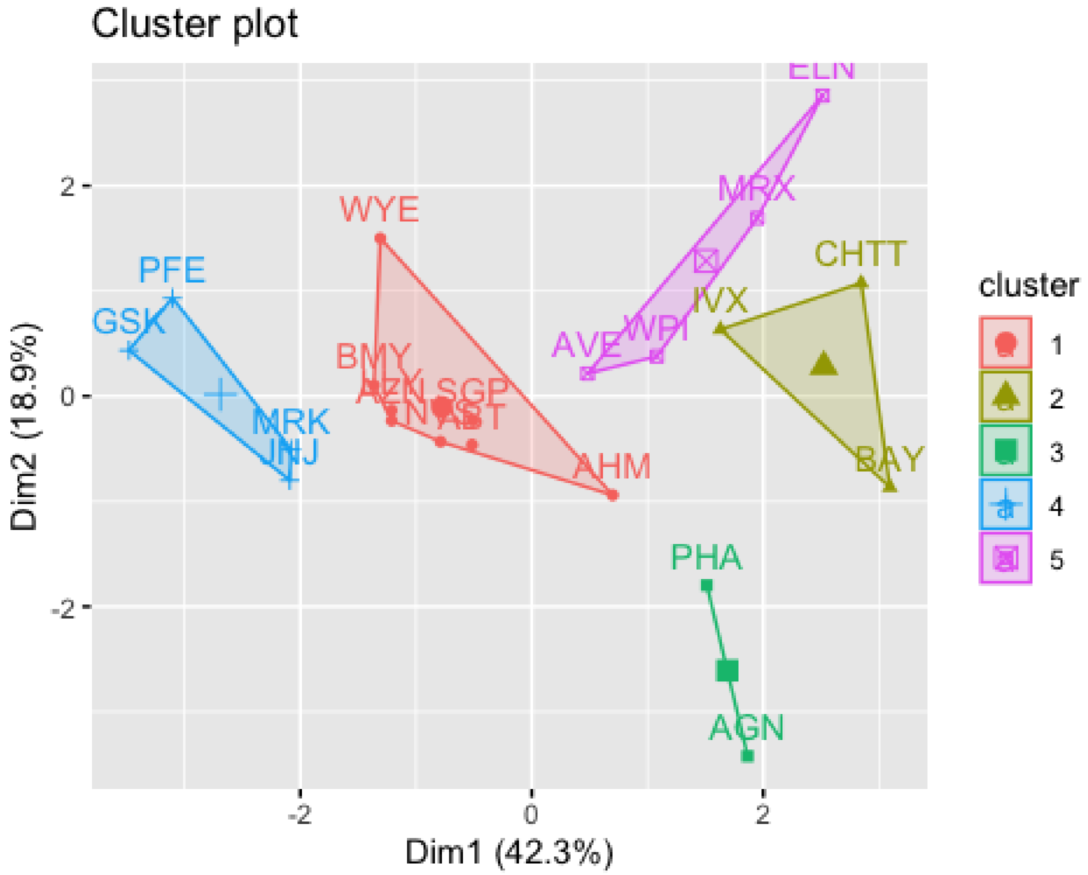
```
## [1] 8 3 2 4 4
```

```
k5$cluster[7]
```

```
## BMY
##   1
```

centers=5 is K=5 and nstart=25 is the number of restarts. The sizes of the 5 clusters are 8, 3, 2, 4, and 4 Also found which cluster firm 7, BMY belongs in and got a result of cluster 1.

```
fviz_cluster(k5,data=pharmaceuticalsdf)
```



Visualized the five clusters.

```
pharmaceuticals2<-cbind(pharmaceuticals,"Cluster_Number"=k5$cluster)
```

Added a column with the cluster number into the original data set.

Cluster 1 has the highest number of firms. Based on the variables of the firms in this cluster, this appears to be where firms with average performance will fall. They have the lowest average revenue growth among the 21 firms in this data set, and most of the other variables fall in the middle of the range of values. The Median Recommendation is Hold for half of these firms, meaning that it is not currently in a good position to be bought or sold. All of the firms in Cluster 1 are traded on the NYSE and more than half are located in the US.

Cluster 2 only contains stocks of 3 of the 21 firms. These three stocks all have the same Asset Turnover of 0.6. they also each have relatively low Net Profit Margin, but they differ in other variables such as Market Capitalization and ROE. Bayer AG has a higher Market Cap and low ROE while the other two firms have the opposite. In the cluster plot, this is shown with Bayer being the bottom point of the cluster. Bayer AG is also the only stock with the location of Germany, while the other two are located in the US. This is also the only cluster that has stocks from three different Exchanges.

Cluster 3 is the smallest cluster with only two firms. These two stocks each have relatively low ROE and ROA compared to the other 19 firms in the data. They also have the highest PE Ratio's out of all of the firms. This likely means that the stocks are overvalued, or they are expected to grow in the future. Allegran, Inc. has the highest PE Ratio at 82.5 and was given a median recommendation of Moderate Buy, possibly because of expected growth.

Cluster 4 includes stocks for four of the 21 firms and they are primarily located in the US except for one located in the UK. They are all traded on the NYSE and two have a Median Recommendation of Hold and two have a Median Recommendation of Moderate Buy. These four stocks have the highest Market Capitalization's of the 21 firms in the dataset. This likely means that these are the largest companies on the list. They also have relatively high asset turnover which means that the firms are performing well.

Cluster 5 also has four firms, but it is more diverse with one firm located in France, one in Ireland and two in the US. These firms were all given a Median Recommendation of either Moderate Buy or Moderate sell. These firms have the highest Revenue Growth, but fairly low Market Capitalizations. This could mean that these are smaller companies that are growing quickly.

Cluster 1 = Typical Companies Cluster 2 = 0.6 Asset Turnover Cluster 3 = High PE Ratios Cluster 4 = Large Companies Cluster 5 = Small companies growing quickly