Assignment_4

Andrew Gutierrez

2022-10-30

First, I'll install the requisite packages.

```
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
```

```
library(ISLR)
library(flexclust)
```

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

Then, I'll read the Pharmaceuticals.csv file into a DataFrame in R:

```
pharma = read.csv("C:\\Users\\gutiera9\\Documents\\MSBA KSU\\Pharmaceuticals.csv",header=T,sep=",")

head(pharma)
```

```
##   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
```

1. 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.

First, I'll scale the numeric columns in the dataframe according to z-score:

```
set.seed(123)

pharma[,c(3:11)] <- scale(pharma[,c(3:11)] )
print(pharma)
```

```
##    Symbol                              Name Market_Cap        Beta    PE_Ratio
## 1     ABT               Abbott Laboratories  0.1840960 -0.80125356 -0.04671323
## 2     AGN                    Allergan, Inc. -0.8544181 -0.45070513  3.49706911
## 3     AHM                      Amersham plc -0.8762600 -0.25595600 -0.29195768
## 4     AZN                   AstraZeneca PLC  0.1702742 -0.02225704 -0.24290879
## 5     AVE                          Aventis -0.1790256 -0.80125356 -0.32874435
```

```
## 6     BAY                           Bayer AG -0.6953818  2.27578267  0.14948233
## 7     BMY       Bristol-Myers Squibb Company -0.1078688 -0.10015669 -0.70887325
## 8    CHTT                       Chattem, Inc -0.9767669  1.26308721  0.03299122
## 9     ELN              Elan Corporation, plc -0.9704532  2.15893320 -1.34037772
## 10    LLY              Eli Lilly and Company  0.2762415 -1.34655112  0.14948233
## 11    GSK                GlaxoSmithKline plc  1.0999201 -0.68440408 -0.45749769
## 12    IVX                   IVAX Corporation -0.9393967  0.48409069 -0.34100657
## 13    JNJ                  Johnson & Johnson  1.9841758 -0.25595600  0.18013789
## 14    MRX Medicis Pharmaceutical Corporation -0.9632863  0.87358895  0.19240011
## 15    MRK                  Merck & Co., Inc.  1.2782387 -0.25595600 -0.40231769
## 16    NVS                        Novartis AG  0.6654710 -1.30760129 -0.23677768
## 17    PFE                         Pfizer Inc  2.4199899  0.48409069 -0.11415545
## 18    PHA              Pharmacia Corporation -0.0240846 -0.48965495  1.90298017
## 19    SGP        Schering-Plough Corporation -0.4018812 -0.06120687 -0.40231769
## 20    WPI       Watson Pharmaceuticals, Inc. -0.9281345 -1.11285216 -0.43297324
## 21    WYE                             Wyeth -0.1614497  0.40619104 -0.75792214
##            ROE        ROA Asset_Turnover    Leverage  Rev_Growth
## 1   0.04009035  0.2416121      0.0000000 -0.21209793 -0.52776752
## 2  -0.85483986 -0.9422871      0.9225312  0.01828430 -0.38113909
## 3  -0.72225761 -0.5100700      0.9225312 -0.40408312 -0.57211809
## 4   0.10638147  0.9181259      0.9225312 -0.74965647  0.14744734
## 5  -0.26484883 -0.5664461     -0.4612656 -0.31449003  1.21638667
## 6  -1.45146000 -1.7127612     -0.4612656 -0.74965647 -1.49714434
## 7   0.59693581  0.8617498      0.9225312 -0.02011273 -0.96584257
## 8  -0.11237924 -1.1677918     -0.4612656  3.74279705 -0.63276071
## 9  -0.70899938 -1.0174553     -1.8450624  0.61983791  1.88617085
## 10  0.34502953  0.5610770     -0.4612656 -0.07130879 -0.64814764
## 11  2.45971647  1.8389364      1.3837968 -0.31449003  0.76926048
## 12 -0.29136529 -0.6979905     -0.4612656  1.10620040  0.05603085
## 13  0.18593083  1.0872544      0.9225312 -0.62166634 -0.36213170
## 14 -0.96753478 -0.9610792     -1.8450624  0.44065173  1.53860717
## 15  0.98142435  0.8429577      1.8450624 -0.39128411  0.36014907
## 16 -0.52338423  0.1288598     -0.9225312 -0.67286239 -1.45369888
## 17  1.31287998  1.6322239      0.4612656 -0.54487226  1.10143723
## 18 -0.81506519 -0.9047030     -0.4612656 -0.30169102  0.14744734
## 19 -0.21181593  0.5234929      0.4612656 -0.74965647 -0.43544591
## 20 -1.03382590 -0.6979905     -0.9225312 -0.49367621  1.43089863
## 21  1.92938746  0.5422849     -0.4612656  0.68383297 -1.17763919
##    Net_Profit_Margin Median_Recommendation    Location Exchange
```

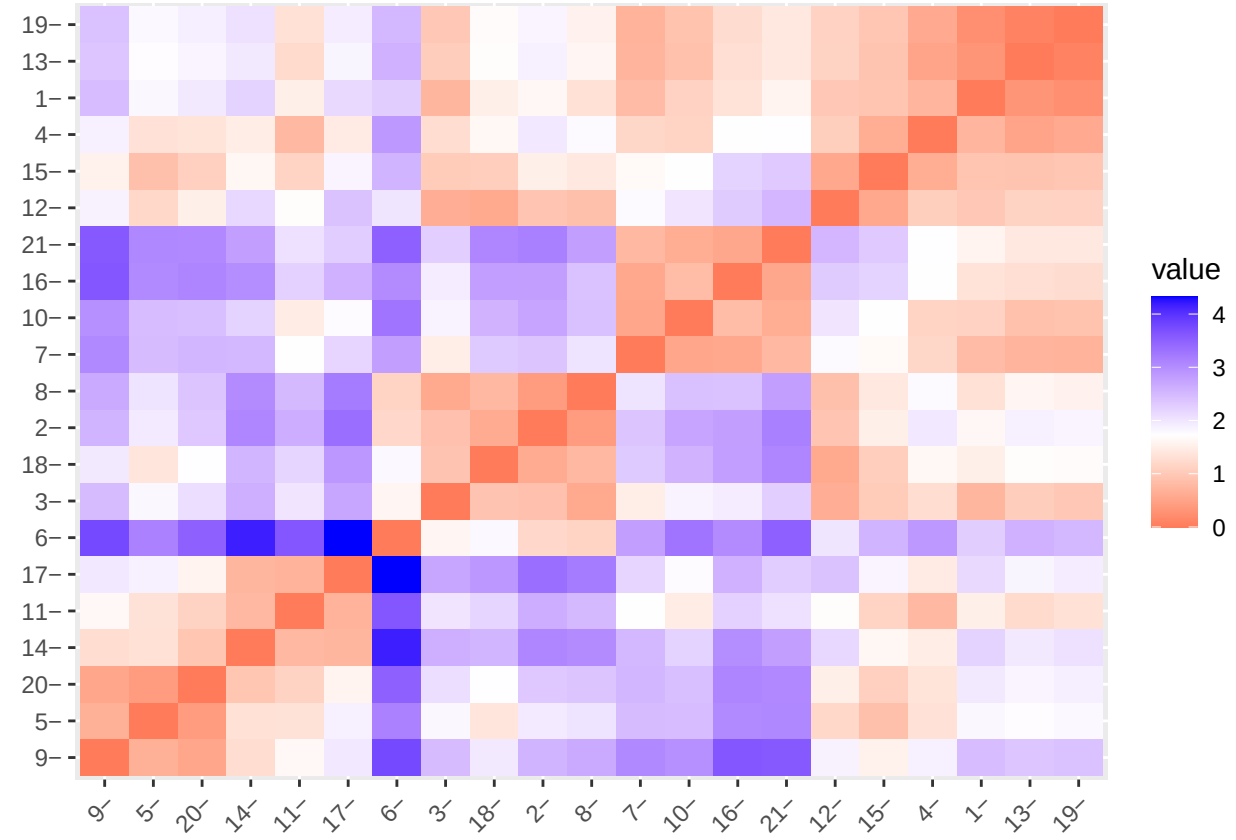
```
## 1         0.06168225          Moderate Buy          US     NYSE
## 2        -1.55366706          Moderate Buy      CANADA     NYSE
## 3        -0.68503583            Strong Buy          UK     NYSE
## 4         0.35122600         Moderate Sell          UK     NYSE
## 5        -0.42597037          Moderate Buy      FRANCE     NYSE
## 6        -1.99560225                 Hold     GERMANY     NYSE
## 7         0.74744375         Moderate Sell          US     NYSE
## 8        -1.24888417          Moderate Buy          US   NASDAQ
## 9        -0.36501379         Moderate Sell     IRELAND     NYSE
## 10        1.17413980                 Hold          US     NYSE
## 11        0.82363947                 Hold          UK     NYSE
## 12       -0.71551412                 Hold          US     AMEX
## 13        0.33598685          Moderate Buy          US     NYSE
## 14        0.85411776          Moderate Buy          US     NYSE
## 15       -0.24310064                 Hold          US     NYSE
## 16        1.02174835                 Hold SWITZERLAND     NYSE
## 17        1.44844440          Moderate Buy          US     NYSE
## 18       -1.27936246                 Hold          US     NYSE
## 19        0.29026942                 Hold          US     NYSE
## 20       -0.09070919         Moderate Sell          US     NYSE
## 21        1.49416183                 Hold          US     NYSE
```

For this exercise, I've chosen to cluster the pharmaceutical companies by two variables - Revenue Growth, and Net Profit Margin. I believe these two variables in particular are ones that investors would be very interested in grouping these companies by, as they provide reasonable measures of how well each company is doing financially.

Since neither of these fields have any extreme outliers, the Euclidean distance measure should suffice for calculating the distance between observations. These two fields are not necessarily correlated with each other either. A company might be experiencing high revenue growth for example, but if their expenses are also high, then their net profit margin would comparatively be lower. So high revenue growth does not necessarily beget a high profit margin. Since Euclidean distance ignores relationships between variables, the lack of correlation between these two figures should work just fine for this particular model.

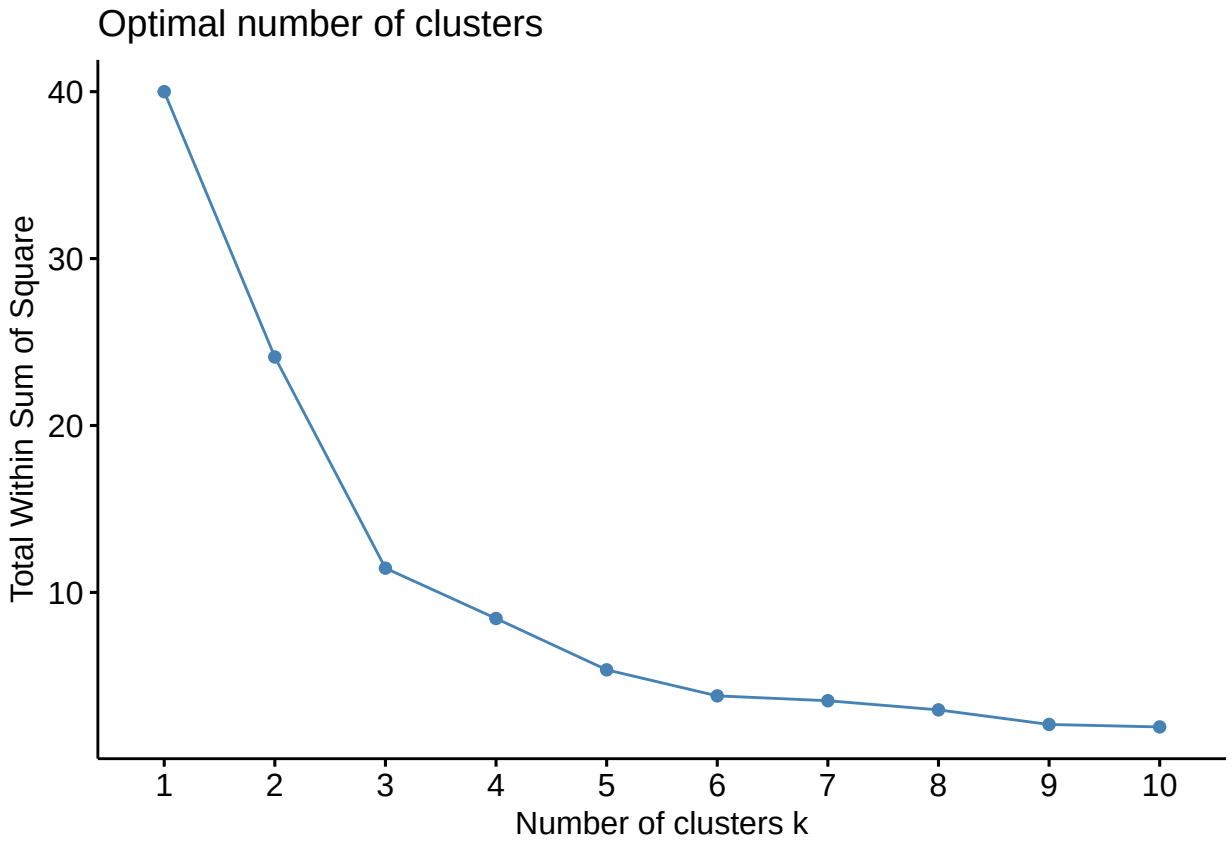
I'll now use the get_dist() function to display the Euclidean distances of each data point.

```
distance <- get_dist(pharma[,c(10,11)],method="euclidean")
fviz_dist(distance)
```



Next, we have to determine the optimal value of k. We'll do this using an elbow chart.

```
fviz_nbclust(pharma[,c(10,11)], kmeans, method = "wss")
```



This elbow chart clearly shows that the optimal value of k (meaning the value that corresponds to the least amount of difference between items in each cluster) is 3. So, three clusters it is for my model.

Now that I have my ideal value of k, it's time to run our clustering model. For this I'll use the kmeans method, as it works particularly well with the Euclidean distance measure that I've chosen to use earlier (kmeans has the added benefits of being relatively easy to implement, and of producing generally tighter clusters than other clustering models).

```
k3 <- kmeans(pharma[,c(10,11)], centers = 3, nstart = 25)

print("Here are the centers of each cluster: ")
```

```
## [1] "Here are the centers of each cluster: "
```

```r
print(k3$centers)
```

```
##   Rev_Growth Net_Profit_Margin
## 1 -0.4799473        -1.2463443
## 2 -0.6779033         0.6845823
## 3  1.1861300         0.2859154
```

```r
print("Here are the number of companies in each cluster: ")
```

```
## [1] "Here are the number of companies in each cluster: "
```

```r
print(k3$size)
```

```
## [1] 6 8 7
```

```r
print("And here is the visual: ")
```

```
## [1] "And here is the visual: "
```

```r
fviz_cluster(k3, data = pharma[,c(10,11)])
```

Before moving on to the remaining assignment prompts, I'll merge the cluster output from my model with the original Pharmaceuticals dataframe - allowing me to easily see which companies are allocated to which clusters.

```
pharma <- cbind(pharma,data.frame(k3$cluster))
print(pharma[,c(2,12:15)])
```

```
##                                 Name Median_Recommendation    Location
## 1                 Abbott Laboratories          Moderate Buy          US
## 2                      Allergan, Inc.          Moderate Buy      CANADA
## 3                        Amersham plc            Strong Buy          UK
## 4                     AstraZeneca PLC         Moderate Sell          UK
```

```
## 5                           Aventis          Moderate Buy      FRANCE
## 6                         Bayer AG                 Hold     GERMANY
## 7       Bristol-Myers Squibb Company         Moderate Sell          US
## 8                     Chattem, Inc          Moderate Buy          US
## 9              Elan Corporation, plc         Moderate Sell     IRELAND
## 10             Eli Lilly and Company                 Hold          US
## 11               GlaxoSmithKline plc                 Hold          UK
## 12                  IVAX Corporation                 Hold          US
## 13                  Johnson & Johnson          Moderate Buy          US
## 14 Medicis Pharmaceutical Corporation          Moderate Buy          US
## 15                 Merck & Co., Inc.                 Hold          US
## 16                       Novartis AG                 Hold SWITZERLAND
## 17                        Pfizer Inc          Moderate Buy          US
## 18             Pharmacia Corporation                 Hold          US
## 19       Schering-Plough Corporation                 Hold          US
## 20      Watson Pharmaceuticals, Inc.         Moderate Sell          US
## 21                             Wyeth                 Hold          US
##    Exchange k3.cluster
## 1      NYSE          2
## 2      NYSE          1
## 3      NYSE          1
## 4      NYSE          2
## 5      NYSE          3
## 6      NYSE          1
## 7      NYSE          2
## 8    NASDAQ          1
## 9      NYSE          3
## 10     NYSE          2
## 11     NYSE          3
## 12     AMEX          1
## 13     NYSE          2
## 14     NYSE          3
## 15     NYSE          3
## 16     NYSE          2
## 17     NYSE          3
## 18     NYSE          1
## 19     NYSE          2
## 20     NYSE          3
## 21     NYSE          2
```

Note that my responses to the remaining assignment prompts B through D are actually contained in the 'A.Gutierrez Assignment 4 Responses' TXT file that is also included in my GitHub folder for this assignment.