Assignment_4

Hannah Cronin

2022-10-30

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ dplyr   1.0.10
## ✔ tibble  3.1.8      ✔ stringr 1.4.1
## ✔ tidyr   1.2.1      ✔ forcats 0.5.2
## ✔ purrr   0.3.4
## ── 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/ve3WB
a
```

```
library(ISLR)
library(cluster)
Pharmaceuticals <- read_csv("/Users/hannahcronin/Desktop/GITHUB/64060_-HCRONIN-FML/Assig
nment_4/Pharmaceuticals.csv")
```

```
## Rows: 21 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (5): Symbol, Name, Median_Recommendation, Location, Exchange
## dbl (9): Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

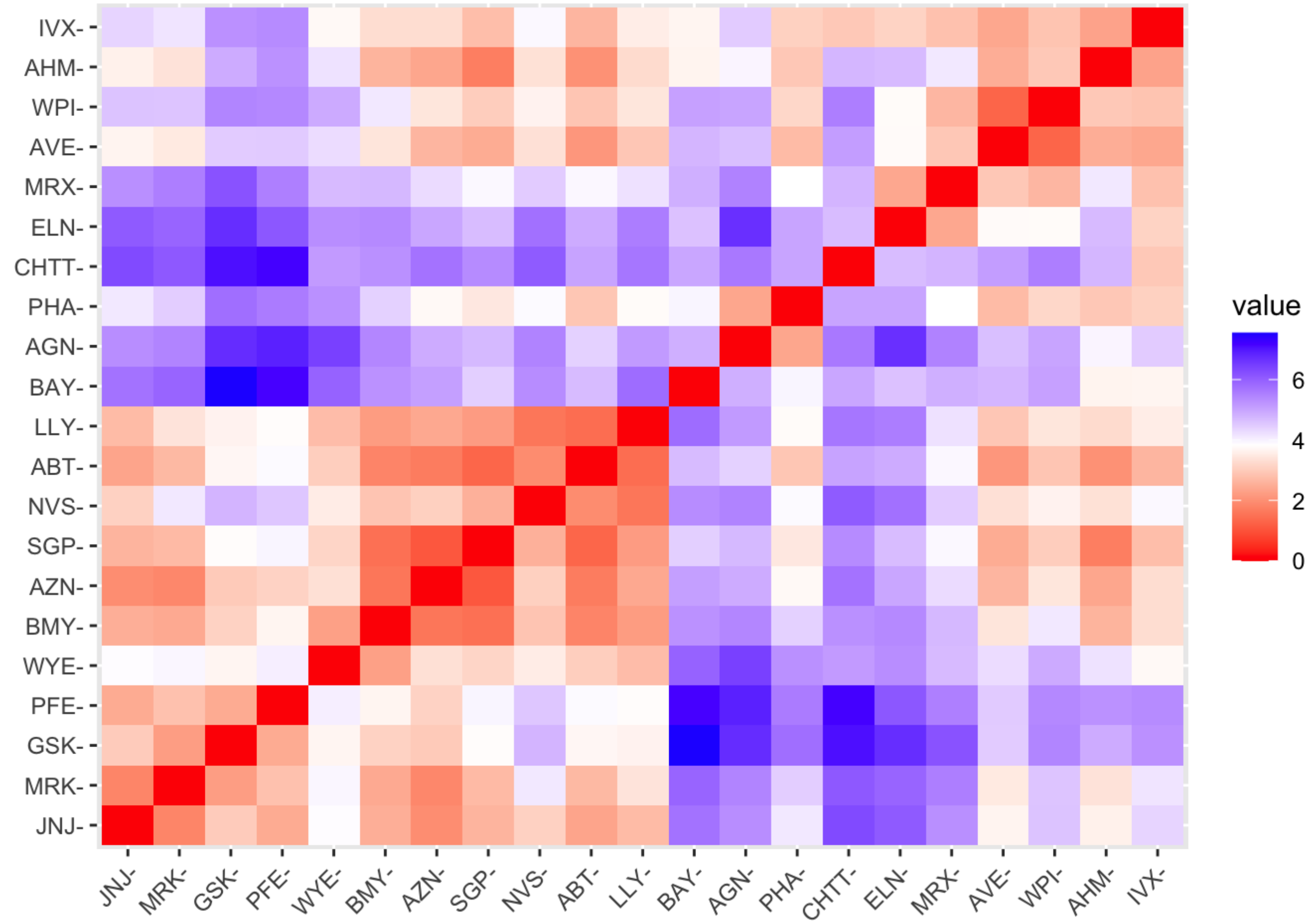
```
set.seed(123)
df = Pharmaceuticals[, c(3,4,5,6,7,8,9,10,11)]
rownames(df) <- c('ABT','AGN','AHM','AZN','AVE','BAY','BMY','CHTT','ELN','LLY','GSK','IV
X','JNJ','MRX','MRK','NVS','PFE','PHA','SGP','WPI','WYE')
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
colnames(df) <- c('Market Cap','Beta','PE_Ratio','ROE','ROA','Asset_Turnover','Leverage'
,'Rev_Growth','Net_Profit_Margin')
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
```

```
df = scale(df) #to normalize data
rownames(df) <- c('ABT','AGN','AHM','AZN','AVE','BAY','BMY','CHTT','ELN','LLY','GSK','IV
X','JNJ','MRX','MRK','NVS','PFE','PHA','SGP','WPI','WYE') #row names kept disappearing
colnames(df) <- c('Market Cap','Beta','PE_Ratio','ROE','ROA','Asset_Turnover','Leverage'
,'Rev_Growth','Net_Profit_Margin') #also to ensure column names stick
distance = get_dist(df)
fviz_dist(distance)
```

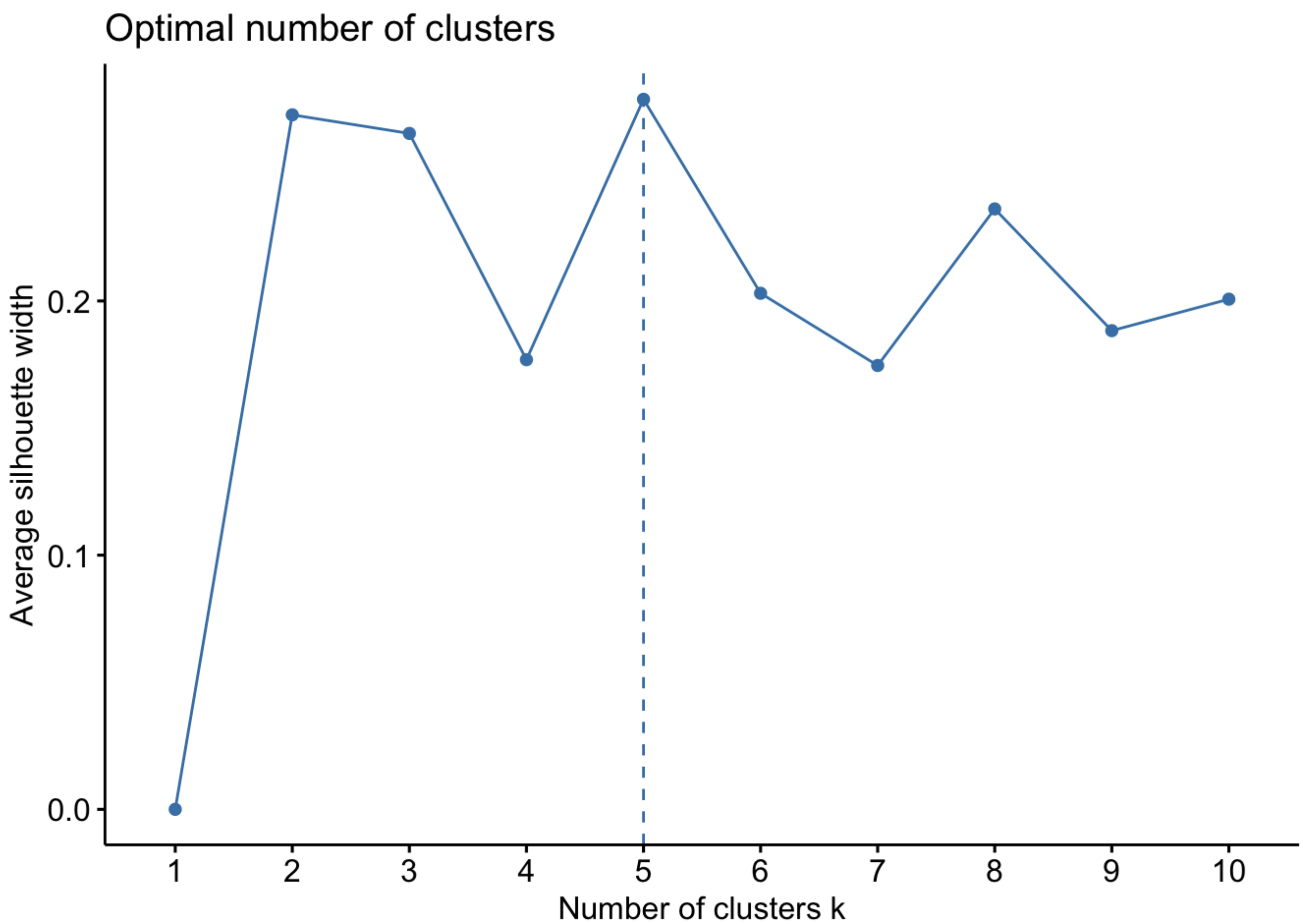



```
summary(df)
```

```
##    Market Cap          Beta             PE_Ratio            ROE
##  Min.   :-0.9768   Min.   :-1.3466   Min.   :-1.3404   Min.   :-1.4515
##  1st Qu.:-0.8763   1st Qu.:-0.6844   1st Qu.:-0.4023   1st Qu.:-0.7223
##  Median :-0.1614   Median :-0.2560   Median :-0.2429   Median :-0.2118
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.2762   3rd Qu.: 0.4841   3rd Qu.: 0.1495   3rd Qu.: 0.3450
##  Max.   : 2.4200   Max.   : 2.2758   Max.   : 3.4971   Max.   : 2.4597
##       ROA          Asset_Turnover       Leverage          Rev_Growth
##  Min.   :-1.7128   Min.   :-1.8451   Min.   :-0.74966   Min.   :-1.4971
##  1st Qu.:-0.9047   1st Qu.:-0.4613   1st Qu.:-0.54487   1st Qu.:-0.6328
##  Median : 0.1289   Median :-0.4613   Median :-0.31449   Median :-0.3621
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.8430   3rd Qu.: 0.9225   3rd Qu.: 0.01828   3rd Qu.: 0.7693
##  Max.   : 1.8389   Max.   : 1.8451   Max.   : 3.74280   Max.   : 1.8862
##  Net_Profit_Margin
##  Min.   :-1.99560
##  1st Qu.:-0.68504
##  Median : 0.06168
##  Mean   : 0.00000
##  3rd Qu.: 0.82364
##  Max.   : 1.49416
```

I used the silhouette method to find the best number of clusters.

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



Running K-Means

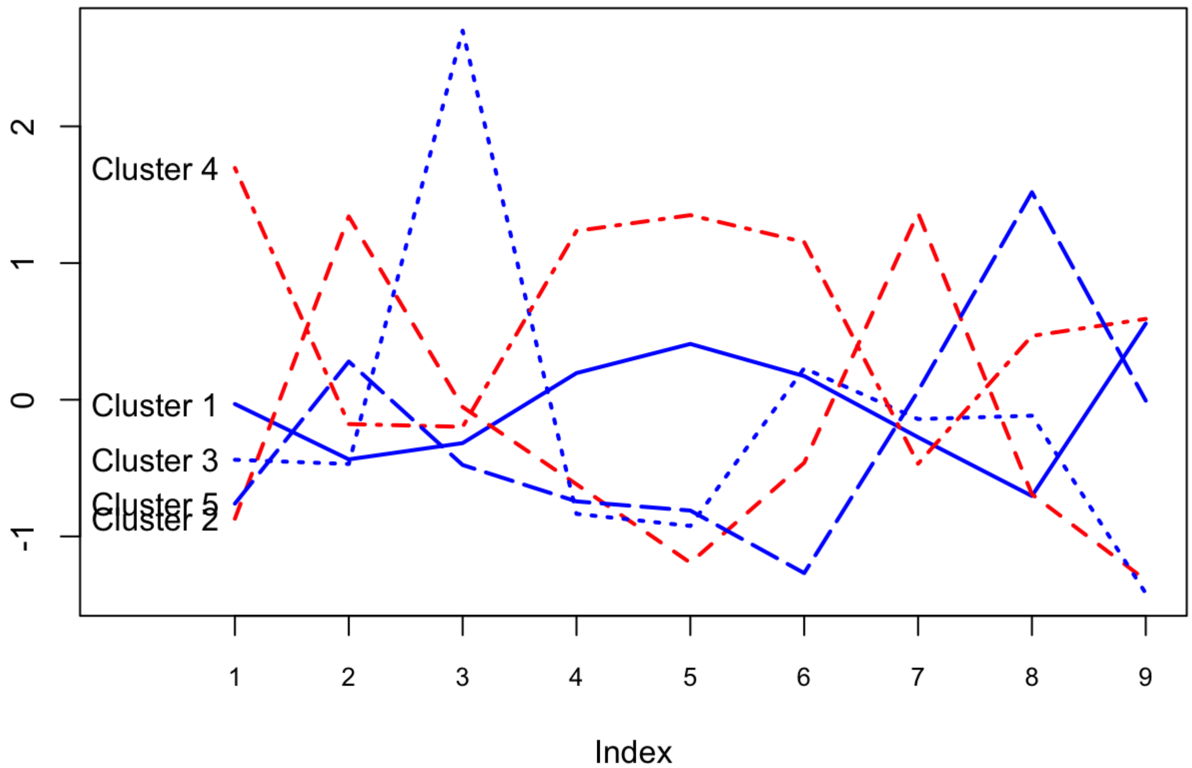
```
km = kmeans(df, 5, nstart = 25)
km$cluster
```

```
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    1    3    1    1    5    2    1    2    5    1    4    2    4    5    4    1
##  PFE  PHA  SGP  WPI  WYE
##    4    3    1    5    1
```

I decided to use KMeans/Euclidean distance because these financial ratios/statistics are not inherently correlated. A few may share some common demoninators, however they represent/pull other data from such different areas (ex: different financial statements) that I chose not to use the Manhattan distance metric.

Cluster 1: ABT, AHM, AZN, BMY, LLY, NVS, SGP, WYE Cluster 2: BAY, CHTT, IVX, Cluster 3: AGN, WPI Cluster 4: GSK, JNJ, MRK, PFE Cluster 5: AVE, ELN, MRX, WPI

```
plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(km$centers), max(km$centers)), xlim = c(0, 9))
axis(1, at = c(1:9), labels = names(df), cex.axis=.8)
for (i in c(1:5))
lines(km$centers[i,], lty = i, lwd = 2, col = ifelse(i %in% c(1, 3, 5),
                                                      "blue", "red"))
text(x = 0.3, y = km$centers[, 1], labels = paste("Cluster", c(1:5)))
```



Descriptions of each cluster: Cluster 1 = Low Beta, Low Rev_Growth, High Net_Profit_Margin (No extremes) Cluster 2 = Low Market_Cap, High Beta, Low ROA, High leverage Cluster 3 = High PE_Ratio, Low Net_Profit_Margin Cluster 4 = High Market_Cap, High ROE, High ROA, High Asset_Turnover, High Net_Profit_Margin Cluster 5 = Low PE_Ratio, Low Asset_Turnover

```
km$centers #numerical descriptions of each cluster
```

```
##    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
```

```
km$withinss #numerical descriptions of each cluster
```

```
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
```

```
km$size #numerical descriptions of each cluster
```

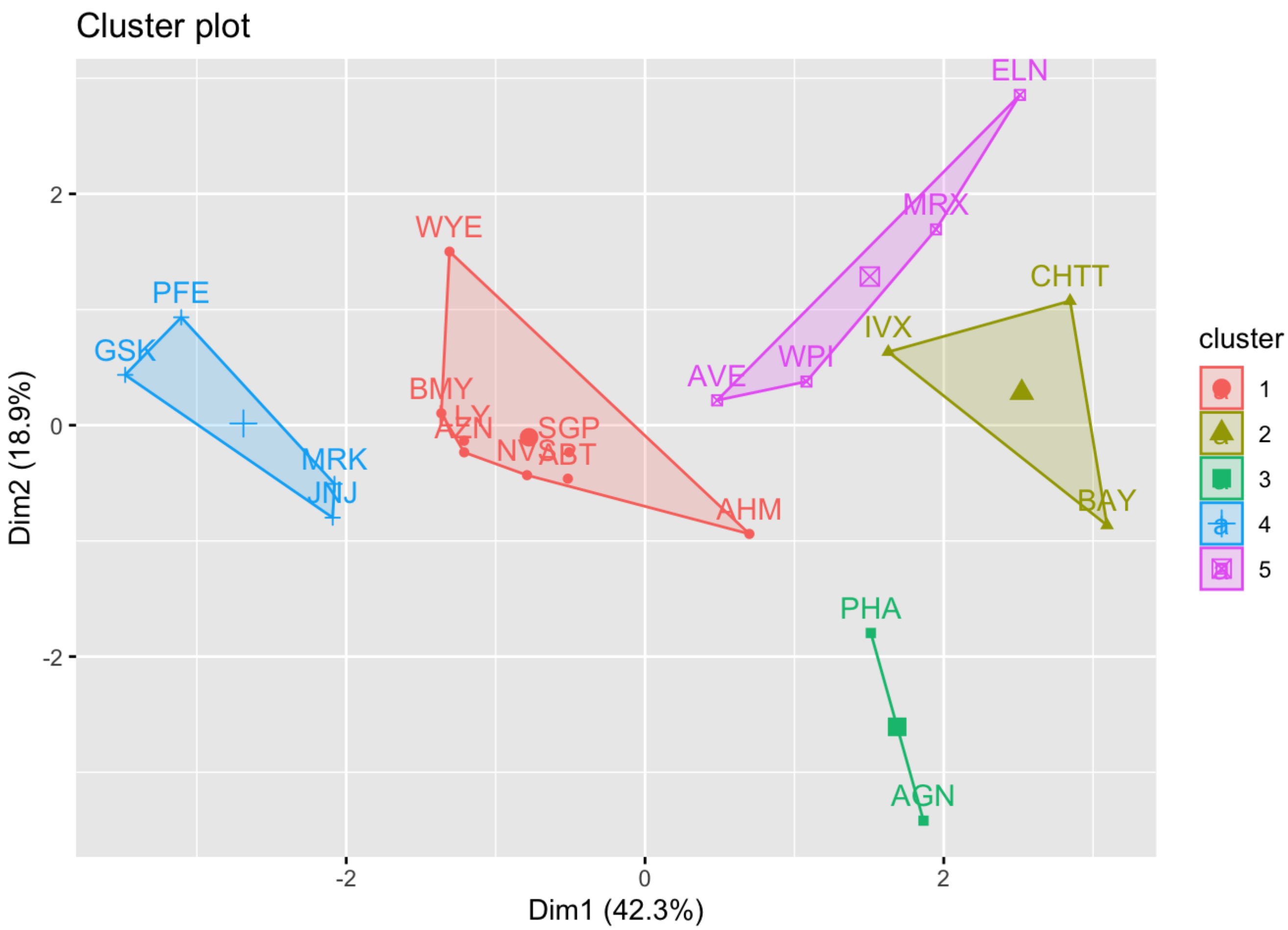
```
## [1] 8 3 2 4 4
```

Clusters 3,4 are the most homogenous- also amongst the smallest clusters. The least homogenous is Cluster 1, coincidentally also the largest cluster.

```
dist(km$centers)
```

```
##          1        2        3        4
## 2 3.711570
## 3 4.045579 3.775790
## 4 2.720924 5.457397 5.275301
## 5 3.299161 3.230532 4.210877 4.744753
```

```
fviz_cluster(km, data = df) # Visualize the output
```



Patterns among the clusters (categorical variables): Cluster 1: All traded on the NYSE, 5/8 of these are US companies Cluster 2: 2/3 of these are holds, 2/3 are US companies Cluster 3: Both listed on the NYSE Cluster 4: 2 are moderately buy, 2 are hold, 3/4 are US companies, all traded on NYSE Cluster 5: 2 Moderately sell, 2 moderately buy, all traded on NYSE

When it comes to the categorical variables, there's definite trends in the dataset related to trading environment and location- however since US and NYSE dominate both categories, I don't think these similiarities are related to the clustering. From the clusters that my model generated, there were no overhwhelming similarities when it came to the median_recommmendation for each company.

Names for each cluster: Cluster1: The_Middle_No_Extremes Cluster2: Low_Market_Cap_ROA_High_Beta_Leverage Cluster3: High_PE_Ratio_Low_Net_Profit_Margin Cluster4: High_Market_Cap_ROE_ROA_Asset_Turnover_Net_Profit_Margin Cluster5: Low_PE_Ratio_Asset_Turnover