MIS 64060: Assignment_4: k-Means for clustering

Eyob Tadele

10/26/2021

Project Objective

The purpose of this assignment is to use k-Means for clustering and analyse the pharmaceuticals dataset. An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file Pharmaceuticals.csv. Use cluster analysis to explore and analyze the given dataset.

(a) 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.

Importing a Pharmaceuticals.csv dataset into r, load relevant libraries, and printout stats about the data.

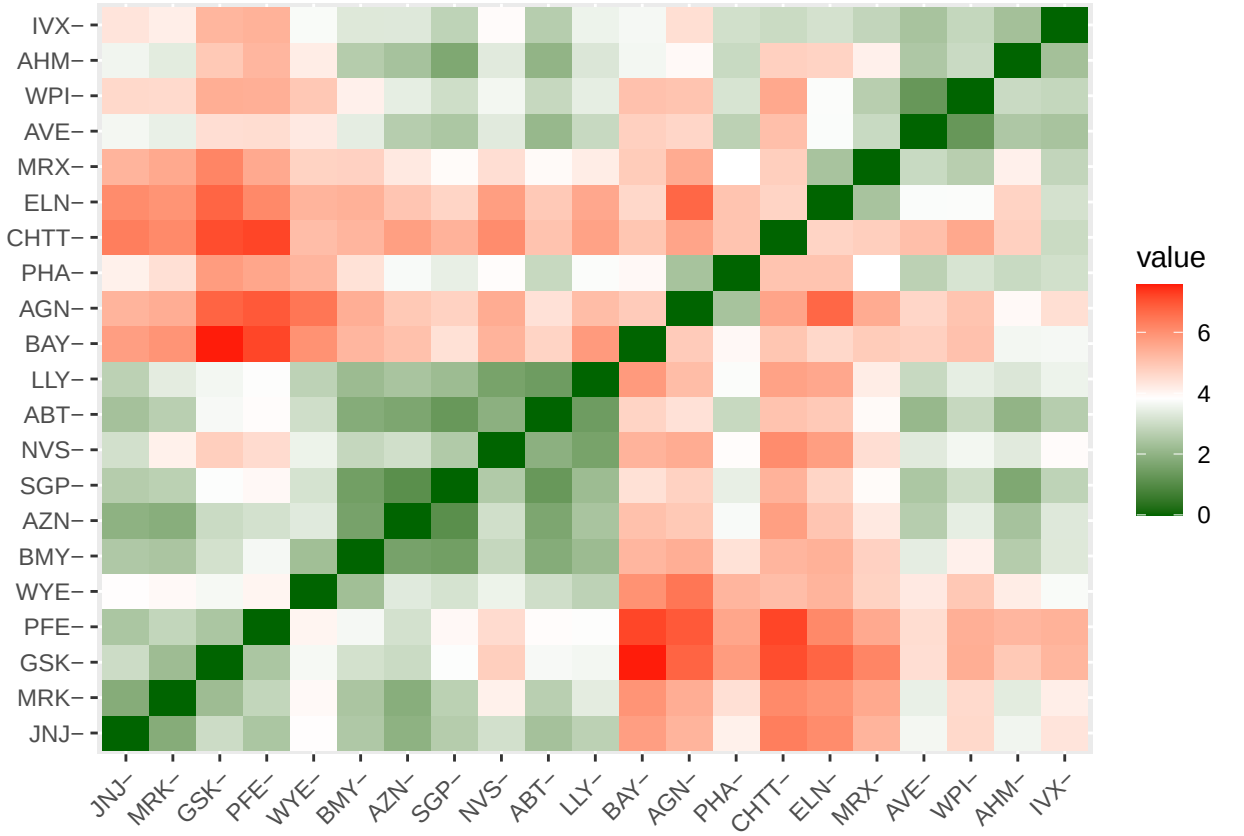
```
library(tidyverse)
library(factoextra)
library(flexclust)
library(cluster)
library(gridExtra)
set.seed(420)
pharmaData <- read.csv("Pharmaceuticals.csv")
rownames(pharmaData) <- pharmaData$Symbol  # setting the row names to company acronyms
pharma_df <- pharmaData[,-c(1,2,12,13,14)] # Selecting only the numerical variables of the dataset
ph_df <- scale(pharma_df)   # Scaling the dataset using z-score
summary(ph_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
```

Computing and visualizing the distance matrix using the functions get_dist() and fviz_dist() from the factoextra R package. This enables us to have visual understanding of the dis/similarity of the different data points.

```
set.seed(420)
distance <- get_dist(ph_df)
# displaying a dissimilarity and distance matrix
fviz_dist(distance, gradient = list(low = "dark green", mid = "white", high = "red"))
```




```
head(round(as.matrix(distance), 2),4) #displaying the first four rows rounded to 2 decimal places
```

```
##       ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK
## ABT 0.00 4.42 2.02 1.67 2.11 4.69 1.81 5.02 4.90 1.42 3.69 2.62 2.33 3.92 2.68
## AGN 4.42 0.00 3.95 4.91 4.64 4.85 5.42 5.61 6.70 5.14 6.75 4.47 5.32 5.48 5.44
## AHM 2.02 3.95 0.00 2.36 2.49 3.64 2.60 4.76 4.70 3.24 4.90 2.32 3.59 4.12 3.36
## AZN 1.67 4.91 2.36 0.00 2.63 5.07 1.57 5.72 4.97 2.41 2.96 3.28 1.96 4.27 1.86
```

```
##      NVS  PFE  PHA  SGP  WPI  WYE
## ABT 1.92 3.89 2.91 1.31 2.88 3.04
## AGN 5.47 6.91 2.37 4.73 5.01 6.45
## AHM 3.33 5.27 2.93 1.70 2.94 4.19
## AZN 3.06 3.11 3.72 1.08 3.41 3.32
```

As distance is an important factor in clustering, the distance matrix above shows the similarity or dissimilarity of each pair of observations based on thier distance (i.e. Green indicating similarity and red showing dissimilarity, in this specific example). The similarity can be used to decide which clusters should be combined or divided into another. Meaning, points with minimal distance value among them should be be in the same cluster.
Since it's required to have a K value in order to use the k-means algorithm, we can manually try different values of K to see how the dataset gets clustered. I've randomly selected k values of 2,3,4,5 and a restart value of 25(which is the default).

```
set.seed(420)
k2 <- kmeans(ph_df, centers = 2, nstart = 25) # k = 2, number of restarts = 25
k3 <- kmeans(ph_df, centers = 3, nstart = 25) # k = 3, number of restarts = 25
k4 <- kmeans(ph_df, centers = 4, nstart = 25) # k = 4, number of restarts = 25
k5 <- kmeans(ph_df, centers = 5, nstart = 25) # k = 5, number of restarts = 25

k2$size # Number of Pharmaceutical companies in each cluster, k=2
```

```
## [1] 10 11
```

```
k3$size # Number of Pharmaceutical companies in each cluster, k=3
```

```
## [1]  4  6 11
```

```
k4$size # Number of Pharmaceutical companies in each cluster, k=4
```

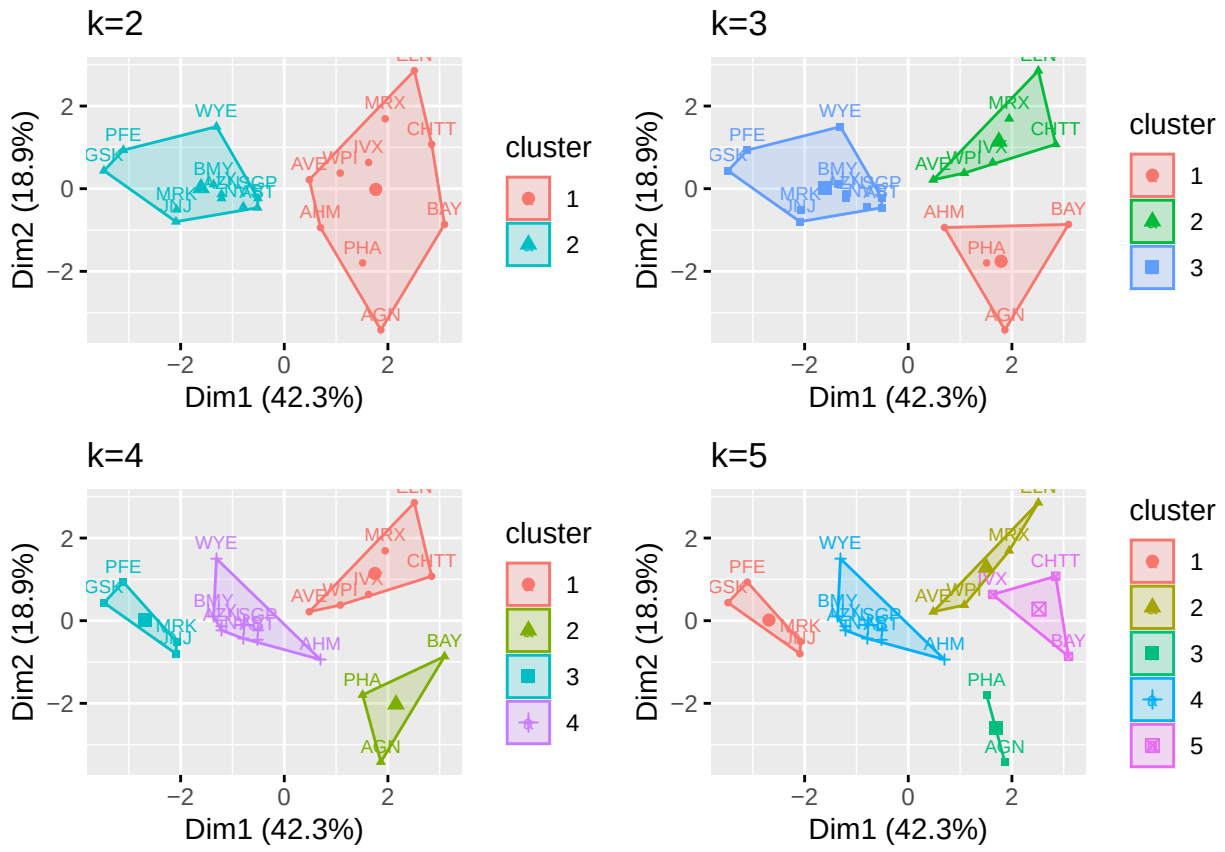
```
## [1] 6 3 4 8
```

```
k5$size # Number of Pharmaceutical companies in each cluster, k=5
```

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

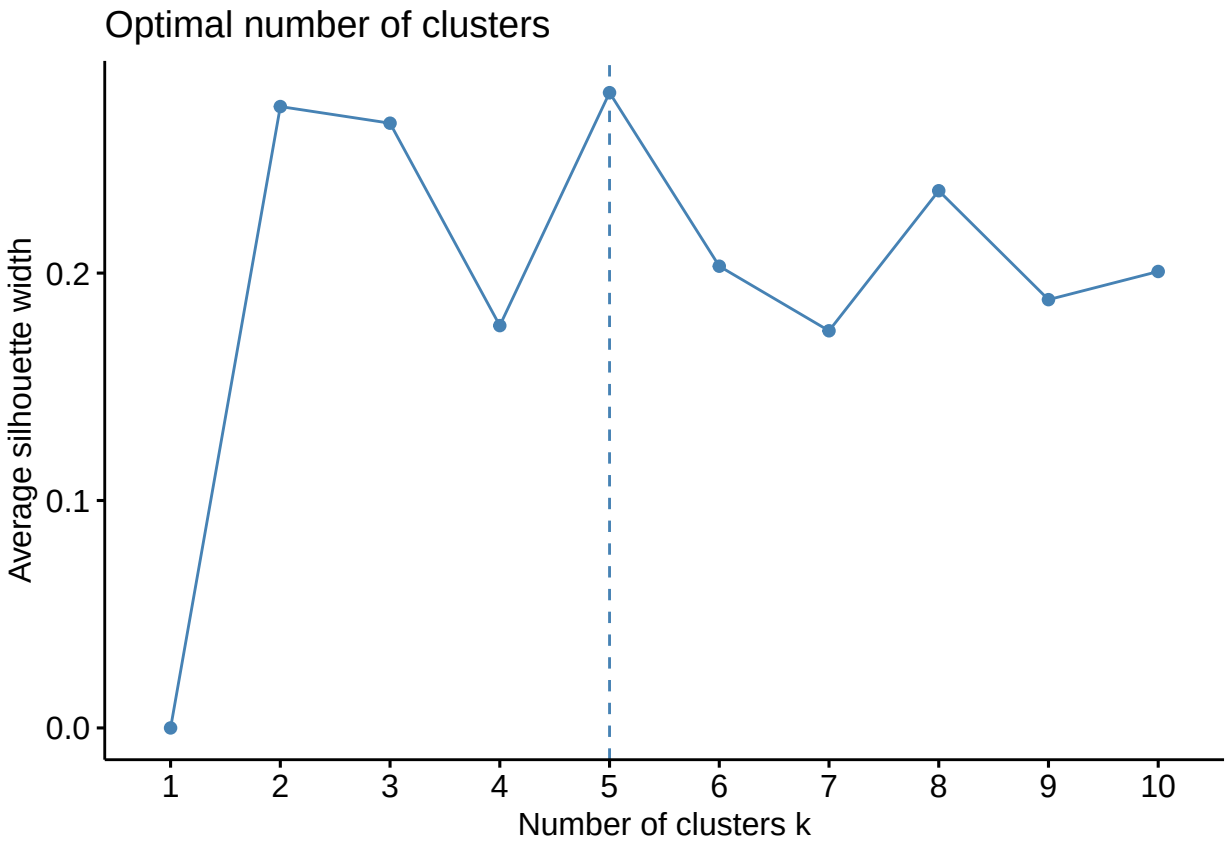
```
# assigning the visual outputs to objects that can be displayed in grid.arrange function
kv2 <- fviz_cluster(k2, data = ph_df, pointsize = 1, labelsize = 7) + ggtitle("k=2")
kv3 <- fviz_cluster(k3, data = ph_df, pointsize = 1, labelsize = 7) + ggtitle("k=3")
kv4 <- fviz_cluster(k4, data = ph_df, pointsize = 1, labelsize = 7) + ggtitle("k=4")
kv5 <- fviz_cluster(k5, data = ph_df, pointsize = 1, labelsize = 7) + ggtitle("k=5")

# arranging the clustering plots above into grids to make it easy for comparison and readability
grid.arrange(kv2,kv3,kv4,kv5)
```



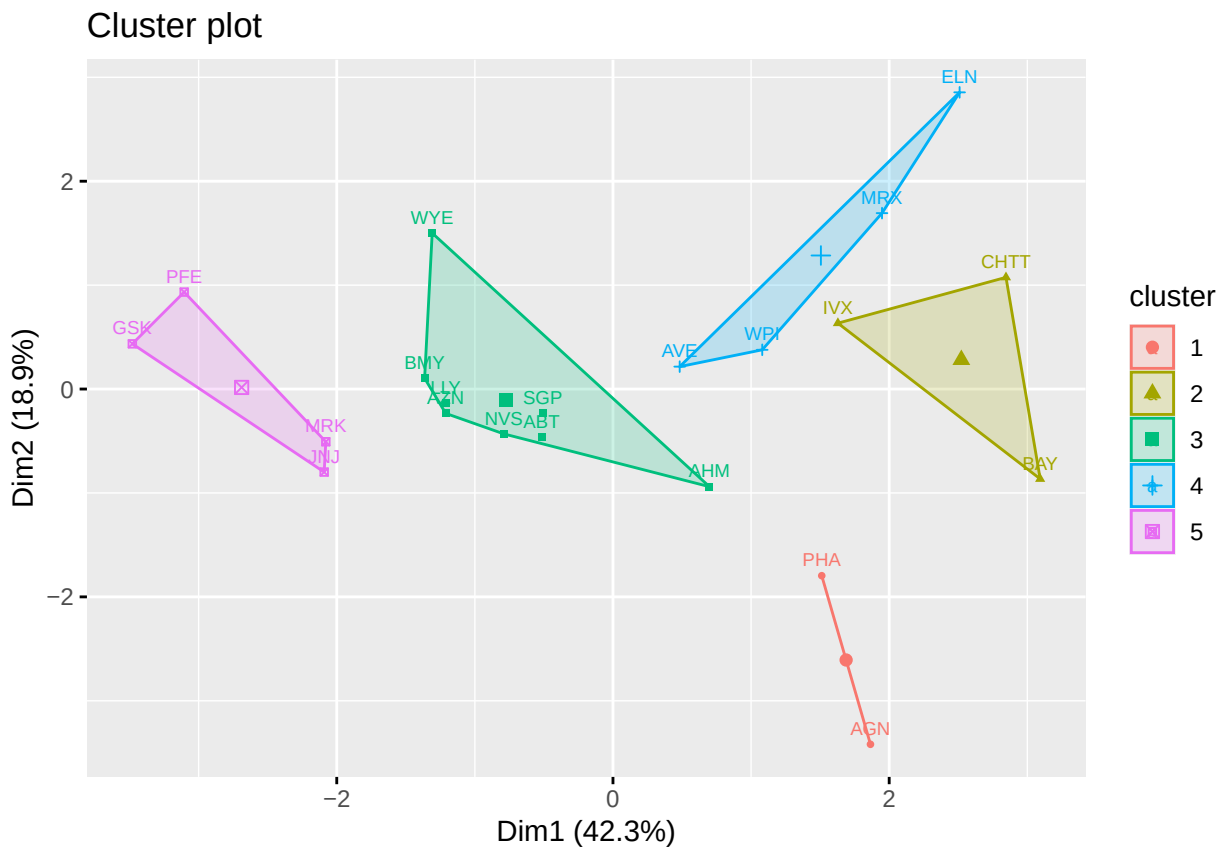
We can see from the above cluster visualizations that some clusters may include data points (i.e the pharmaceutical companies, in this case)that may not necessarily be similar in that cluster. In order to help in identifying the optimal number of clusters, we can apply three methods. Namely, Elbow method, Silhouette method, or Gap statistic. For this case, I've chosen to use the Silhouette method.

```
set.seed(420)
# applying the Silhouette method using the fviz_nbclust function
fviz_nbclust(ph_df, kmeans, method = "silhouette")
```



It can be inferred from the above graph that the optimal value of k=5. As a result, we can use 5 as the optimal k to cluster and analyze the pharmaceuticals dataset.

```
set.seed(420)
k5 <- kmeans(ph_df, centers = 5, nstart = 25)
fviz_cluster(k5, data = ph_df, pointsize = 1, labelsize = 7)
```



If there is a need to reduce the dimension of the features or assign different weights to the features, we can observe the variance within each of the 21 dimensions. Those features/variables with higher variance can be assigned a higher weight. Alternatively, if the variance is very small, we can consider dropping those features with low variance. We can also apply Principal Component Analysis to reduce the dimension.

(b) Interpret the clusters with respect to the numerical variables used in forming the clusters.

From the cluster mean values above, we can see that:

```
print(k5)
```

```
## K-means clustering with 5 clusters of sizes 2, 3, 8, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 4 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.14170336 -0.1168459      -1.416514761
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.27449312 -0.7041516       0.556954446
## 4  0.06308085  1.5180158      -0.006893899
```

```
## 5 -0.46807818  0.4671788       0.591242521
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK  NVS
##    3    1    3    3    4    2    3    2    4    3    5    2    5    4    5    3
##  PFE  PHA  SGP  WPI  WYE
##    5    1    3    4    3
##
## Within cluster sum of squares by cluster:
## [1]  2.803505 15.595925 21.879320 12.791257  9.284424
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

- Cluster 1 has a bigger Market Cap, ROE, ROA, Asset Turnover, and Net Profit Margin; also has a lesser Beta(vulnerability to systematic risk), PE Ratio(growth in the future), and Leverage. This might suggest a cluster of well established big pharma companies.
- The first thing that stands out in Cluster 2 is its higher PE_Ratio, suggesting the stock's price is high relative to the earnings and possibly overpriced. Also the Net Profit Margin and ROE appears to be the lowest among the clusters.
- Cluster 3 has the lowest Revenue Growth but a better current Net_Profit Margin. The Revenue Growth and PE Ratio are also lower suggesting bleak growth potential.
- Cluster 4 appears to have the highest Rev_Growth but relatively unremarkable in the other factors, including low Market Cap.
- Cluster 5 has the highest Beta(i.e. vulnerable to market changes) and highest Leverage(making it bad, considering its Profit_Margin, ROA, and Rev_Growth are low). This cluster appears to be performing poorly across all the features.

(c) Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

- As far as variable 12 (Exchange) is concerned, almost 90% of the dataset belongs in NYSE and this doesn't help in identifying a pattern.
- The Median_Recommendations somewhat align to the Clusters, but not completely. Companies in Cluster 1 are recommended a Hold or Moderate Buy, as it's performing well in most categories. Cluster 2 are considered overpriced and buying is not ideal. However, one of the recommendations is for a Moderate Buy, which doesn't make sense here. Recommendations to 'Hold' for Cluster 5 only makes sense, if we don't want to get rid of it at a low price. Clusters 3 and 4 aren't clear as far as recommendations are concerned. It has a mix of Hold,Moderate Sell/Buy.
- In terms of location, 67% are based in the US or Canada. The rest are in Europe. This also doesn't align clearly with the clusters. Generally, it looks like the non-selected variables(10-12) do not appear to show a clear pattern in the clusters. Feature 10(Median_Recommendation) slightly follows a pattern with the clusters, but only to a limited degree.

(d) Provide an appropriate name for each cluster using any or all of the variables in the dataset.

- Cluster 1: 'Big Pharma', with high market cap, ROE, ROA, Asset Turnover, and NPM.
- Cluster 2: 'Overpriced Pharma', with high PE ratio.
- Cluster 3: 'Currently Profitable Pharma' with good Net_Profit_Margin, but lowest Revenue Growth.
- Cluster 4: 'Future Potential Pharma', with highest Rev_Growth.
- Cluster 5: 'Poorly Performing Pharma', with low performance across all the features.