kmeans

akhil yada

2022-11-06

```
#loading the required packages
library(flexclust)
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
library(cluster)
library(tidyverse)
```

```
## ── Attaching packages
## ─────────────────────────────────────── tidyverse 1.3.2 ──
```

```
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ─────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(factoextra)
```

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

```
pharma = read.csv("c:/Users/91773/Desktop/Pharmaceuticals.csv ")

# I am selecting columns from 3 to 11 and storing the data in variable
pharma1
pharma1 <- pharma[3:11]
# Using head function to display the first 6 rows of data
head(pharma1)
```

```
##   Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## 1      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## 2       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## 3       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## 4      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## 5      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## 6      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##   Net_Profit_Margin
## 1              16.1
## 2               5.5
## 3              11.2
## 4              18.0
## 5              12.9
## 6               2.6
```

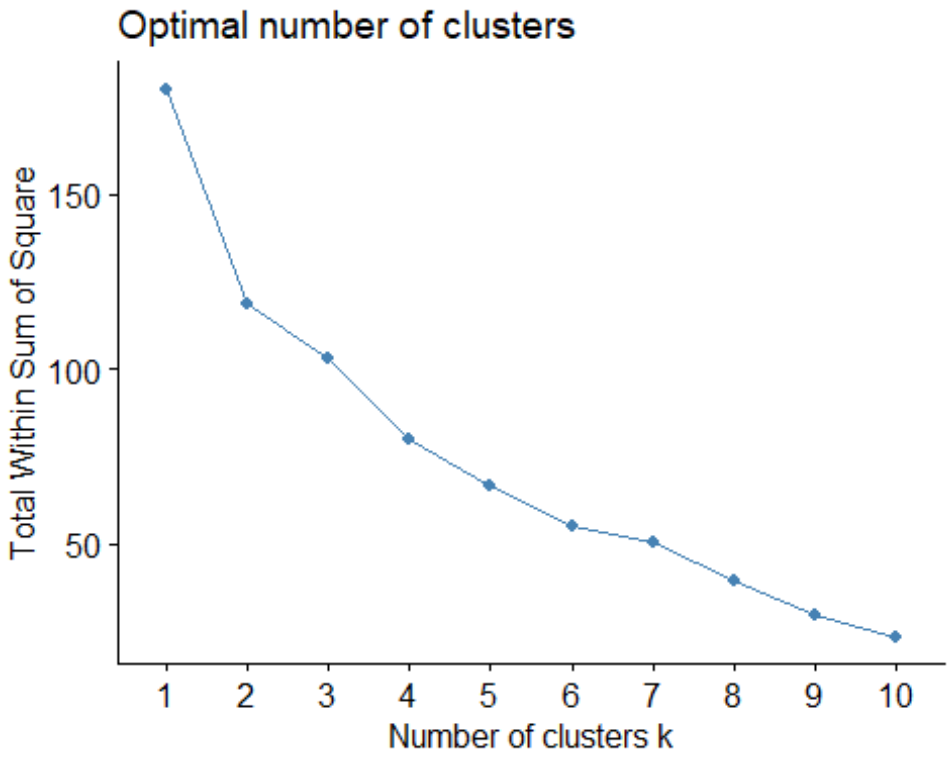
```
summary(pharma1)
```

```
##    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
```

```
# We will scale the data in pharma1 and record the scaled data in the pharma2
dataframe because the variables are measured with varying weights throughout
the rows.
pharma2 <- scale(pharma1)
colnames(is.na(pharma2))
```

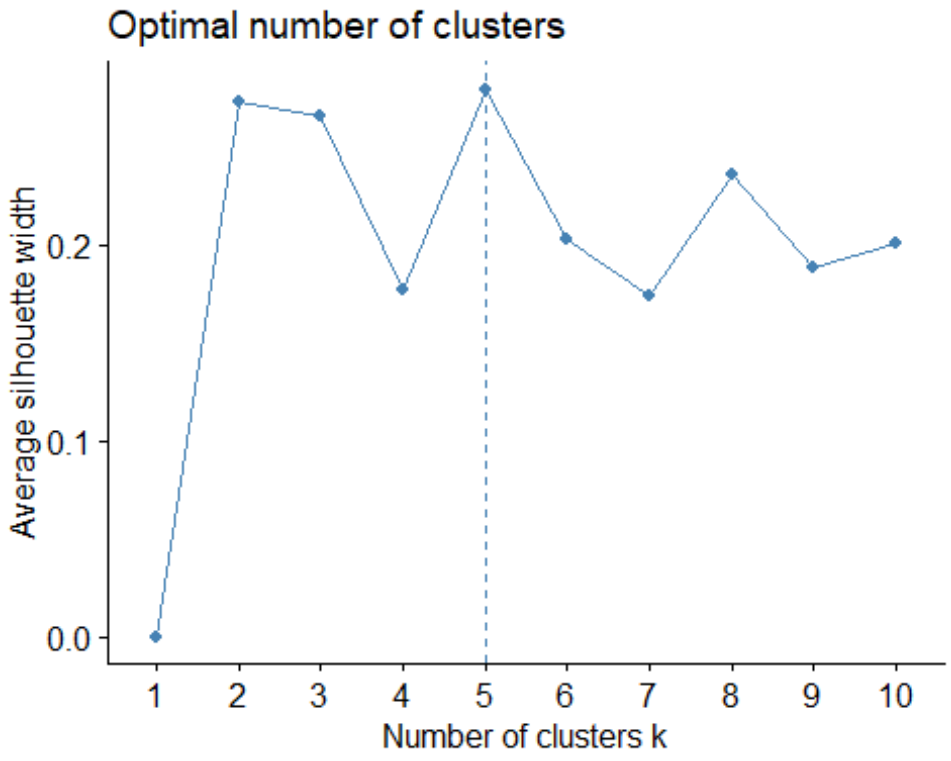
```
## [1] "Market_Cap"        "Beta"              "PE_Ratio"
## [4] "ROE"               "ROA"               "Asset_Turnover"
## [7] "Leverage"          "Rev_Growth"        "Net_Profit_Margin"
```

```
wss =fviz_nbclust(pharma2,kmeans,method = "wss")
wss
```



The optimal k thereby received using the wss method is k = 2 /newline

```
silhouette = fviz_nbclust(pharma2,kmeans,method = "silhouette")
silhouette
```



whereas by employing the silhouette method the optimal k received was k = 5. /newline

cluster formation using K-Means with k = 2 (WSS)

```
wss_kmeans <- kmeans(pharma2,centers = 2,nstart=25) # k = 2, number of
restarts = 25
wss_kmeans

## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##   Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
##
## Clustering vector:
##  [1] 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1 1 2 1 2 1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
```

```
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

after WSS Method i got 2 clusters of size 11 and 10. /newline *cluster formation using k-means with k=5 (silhouette)*

```
silhouette_kmeans <- kmeans(pharma2,centers = 5,nstart = 25) # k = 5, number of restarts = 25
silhouette_kmeans

## K-means clustering with 5 clusters of sizes 3, 2, 8, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 2 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 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  1.36644699 -0.6912914      -1.320000179
## 2 -0.14170336 -0.1168459      -1.416514761
## 3 -0.27449312 -0.7041516       0.556954446
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  [1] 3 2 3 3 5 1 3 1 5 3 4 1 4 5 4 3 4 2 3 5 3
##
## Within cluster sum of squares by cluster:
## [1] 15.595925  2.803505 21.879320  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

after silhouette method we got 5 clusters of size 4,2,3,8,4

plot for cluster WSS

```
fviz_cluster(wss_kmeans,pharma2)
```


```
fviz_cluster(silhouette_kmeans,pharma2)
```

for analysis binding the clusters result to the original data

```
# data formation for wss method
wss_clusters = wss_kmeans$cluster

silhouette_clusters = silhouette_kmeans$cluster
```

I noticed that the total sum of squares within the cluster for the Silhouette method is 62.35 (Sil k5$tot.withinss$), $which is smaller than 118.56 ($Elbk2$tot.withinss), the value I obtained for the Elbow method. This observation helped me choose the appropriate k value for this data set. I like to use the Silhouette approach for this assignment because the sum of the squares within the cluster is lower and results in homogeneous clusters. Therefore, 5 is the ideal value for k.

As opposed to Manhattan distance, I am here utilizing Euclidean distance to measure the distance between the data points because it represents the absolute difference between the data points.

```
#b.interpretation
silhouette_clusters= as.data.frame(silhouette_clusters)
silhouette_2 = cbind(pharma2,silhouette_clusters)
cluster_mean = silhouette_2 %>% group_by(silhouette_clusters) %>%
summarise_all("mean")
cluster_mean

## # A tibble: 5 × 10
##   silhou…¹ Marke…²   Beta PE_Ra…³    ROE    ROA Asset…⁴ Lever…⁵ Rev_G…⁶
Net_Pr…⁷
##      <int>   <dbl>  <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1        1 -0.871   1.34  -0.0528 -0.618 -1.19   -0.461  1.37    -0.691 -
1.32
## 2        2 -0.439  -0.470  2.70   -0.835 -0.923   0.231 -0.142   -0.117 -
1.42
## 3        3 -0.0314 -0.436 -0.317   0.195  0.408   0.173 -0.274   -0.704
0.557
## 4        4  1.70   -0.178 -0.198   1.23   1.35    1.15  -0.468    0.467
0.591
## 5        5 -0.760   0.280 -0.477  -0.744 -0.811  -1.27   0.0631   1.52  -
0.00689
## # … with abbreviated variable names ¹silhouette_clusters, ²Market_Cap,
## #   ³PE_Ratio, ⁴Asset_Turnover, ⁵Leverage, ⁶Rev_Growth, ⁷Net_Profit_Margin
```

cluster 1 This cluster has lower leverage than other clusters, which means that its companies are less indebted than those in those other clusters. Among all the clusters, this one has the lowest revenue growth, but the companies in it have greater net profit margins.The companies in this cluster are performing better than Clusters 2,3, and 5, when the other factors are included. cluster 2 When compared to other clusters, this cluster's mean beta value is higher. This suggests that the companies in this cluster have more volatile stock prices. This cluster has the highest mean leverage, which suggests that these companies have a greater level of debt. Less Market Capital, ROA, Revenue Growth, and Net

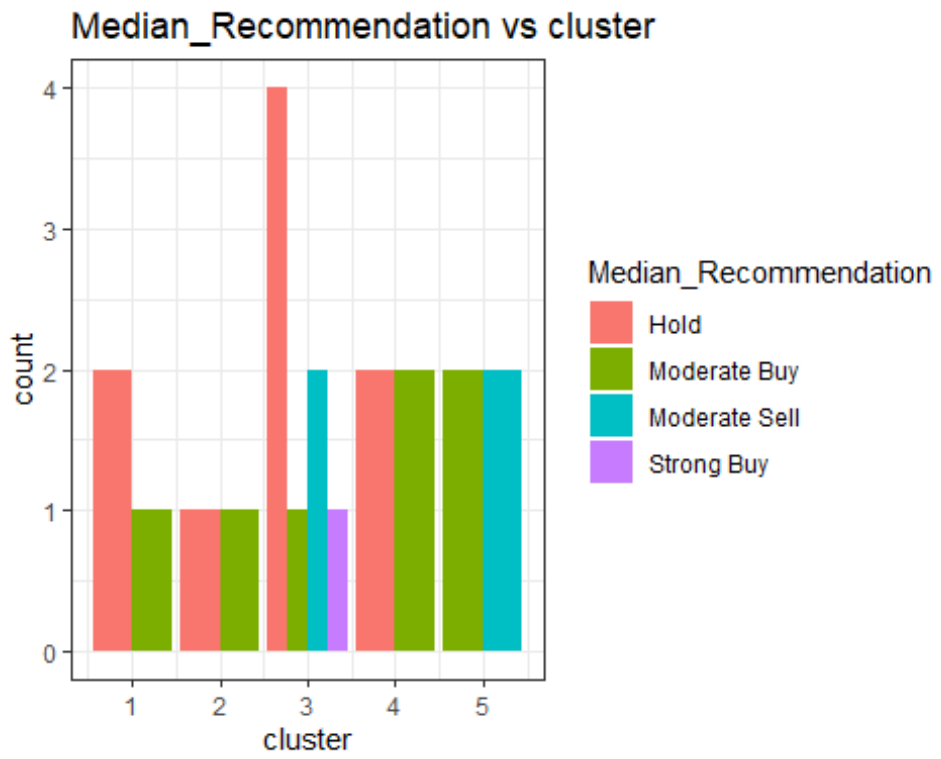
Profit Margin are displayed by the companies in this cluster. This suggests that these businesses must grow financially.

cluster 3 The net profit margin of the businesses in this cluster is the lowest. Additionally, this cluster has the lowest Return on Equity (ROE), indicating that the companies in this cluster struggle to turn equity investments into profits. Furthermore, this cluster has the highest Price-Earnings Ratio, a sign that the businesses are not making money. Even if these businesses' profits are falling, we can still see that their stocks are less volatile because this cluster has the lowest beta value.

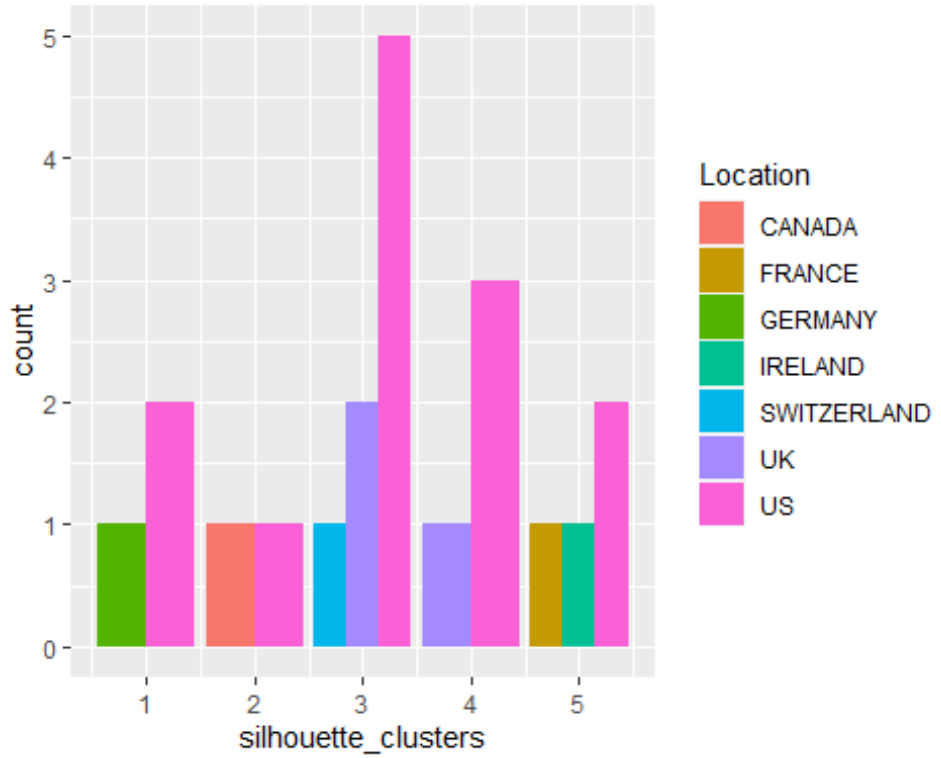
cluster 4 The businesses in this cluster have the highest market capitalization, net profit margin, return on assets (ROA), return on equity (ROE), and asset turnover. The companies in this cluster have the lowest mean leverage values, which indicates that they have less debt relative to shareholders' equity. As a result, when compared to other clusters, this cluster has the highest performing businesses.

cluster 5 High revenue growth among the businesses in this cluster is an indication that business development is going as planned. The companies should, ideally, use their assets to boost revenue, which raises the asset turnover ratio. The asset turnover ratio for this cluster is the lowest, nevertheless. The fact that this group of businesses has the lowest price-to-earnings ratio suggests that their earnings are higher.

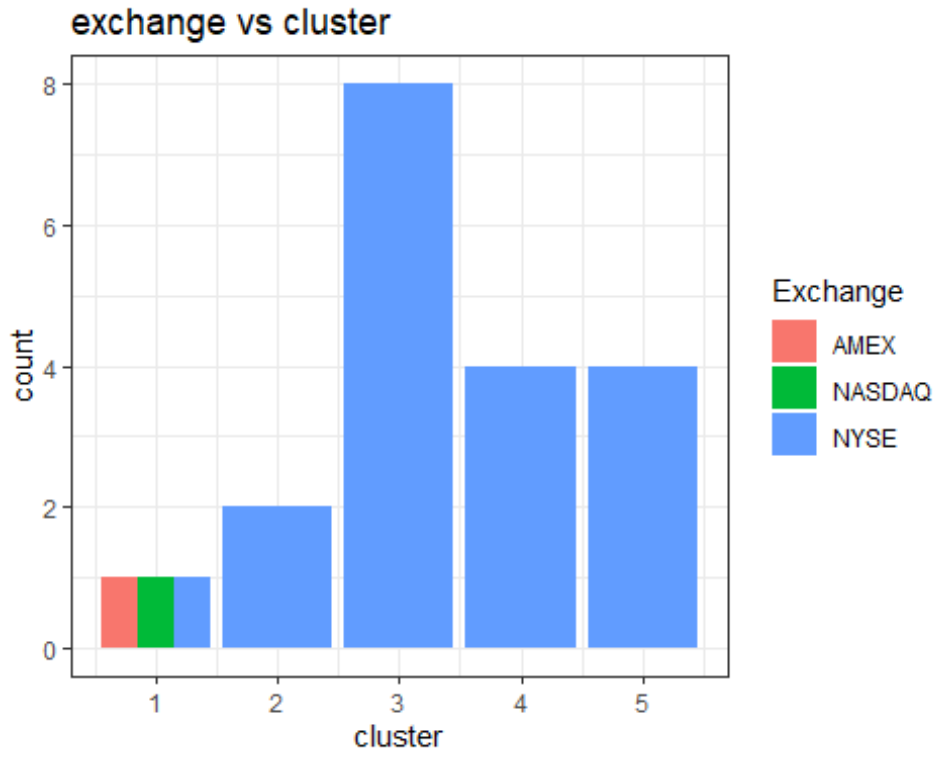
```
library(hrbrthemes)

## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use
these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto
Condensed and

##       if Arial Narrow is not on your system, please see
https://bit.ly/arialnarrow

pharma_6 = pharma[12:14]
pharma_7 = cbind(pharma_6,silhouette_clusters)

ggplot(pharma_7,aes(x=silhouette_clusters,fill =
Median_Recommendation))+geom_bar(position = "dodge") + labs(
  title = "Median_Recommendation vs cluster",
  x= "cluster"
) +
  theme_bw()
```




```
ggplot(pharma_7,aes(x=silhouette_clusters,fill = Location))+
geom_bar(position = "dodge")
```



```
labs(
  title = "location vs cluster",
  x = "Cluster"
) +
  theme_bw()

## NULL

ggplot(pharma_7,aes(x=silhouette_clusters, fill = Exchange))+
  geom_bar(position = "dodge")+
  labs(
    title = "exchange vs cluster",
    x= "cluster"
  )+
  theme_bw()
```



I don't notice any distinct patterns with regard to the category variables because these aren't taken into account while creating the cluster, even though a pattern is an identifiable sequence. I can see, though, that the plots do lend themselves to certain observations.

1. Looking at the median recommendation plot, I can see that Cluster 1 only has one "Strong Buy" recommendation and has a lot of "Hold" recommendations. All of the clusters have a distribution of moderate buy.
2. I can see that all of the clusters have US-based enterprises from the Location vs. Cluster Plot. However, different places can be found throughout all clusters.

3. It can be seen from the Exchange plot that 19 out of 21 companies are listed on the NYSE. There is no pattern provided by this variable.

D) Provide an appropriate name for each silhouette cluster

1. cluster 1 is “Poorly Performing Pharma” and has low performance across all features and extremely high BETA and Leverage values.

2. cluster 2 “Overpriced Pharma”, with a high PE ratio

3. cluster 3 “Currently Profitable Pharma,” has the lowest revenue growth but a solid net profit margin.

4.” Big Pharma” is in Cluster 4, and it has high market capitalization, ROE, ROA, asset turnover, and net profit margin.

5. The Sil Cluster 5 with the highest Rev Growth is “Future Potential Pharma.”

conclusion:

Safety, income, and capital growth are the three characteristics that define any investment. Each investor must choose the right combination of these three elements. Investments are always constrained by their “profit to loss ratio,” and any given investor would like to maximize profits with little to no loss. The cluster “Exorbitant Viability with Slighter Risk” in this data set exhibits all of these features. I believe that this can be the best cluster to choose for an investment given that there is a lower possibility of risk and better earnings, based on the study and interpretation done. Note: Choosing a cluster from the silhouette approach was made possible by the method’s improved ability to define the domain. can be used by anyone to make an effective choice regarding their investing options.