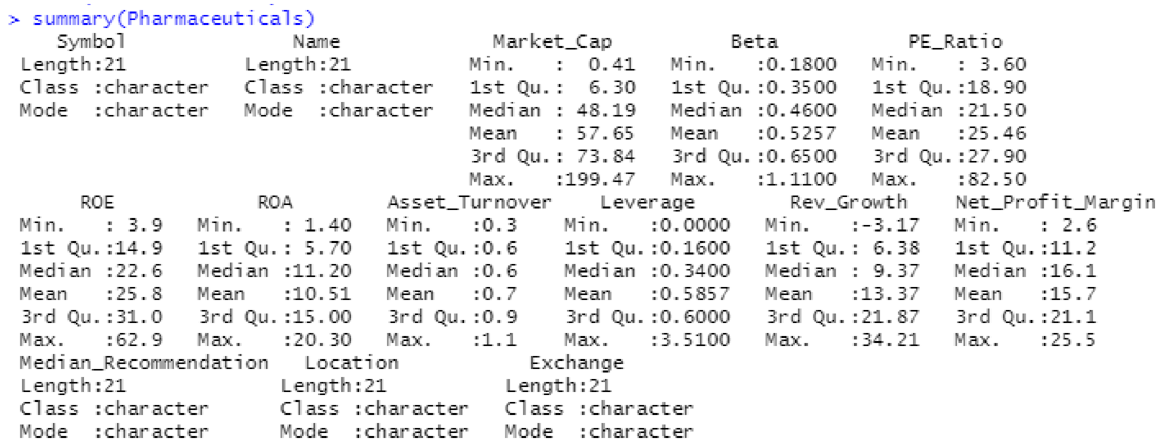
Sharon Kasturiarachi

MIS 64060 Fundamentals of Machine Learning

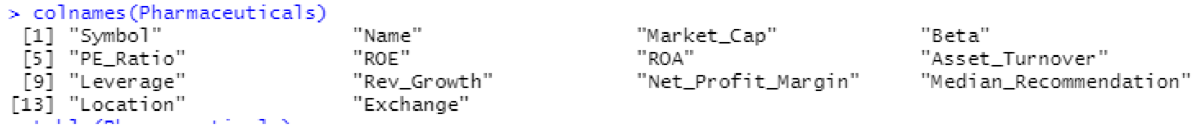
Dr. Rouzbeh Razavi

Assignment 4- Answers

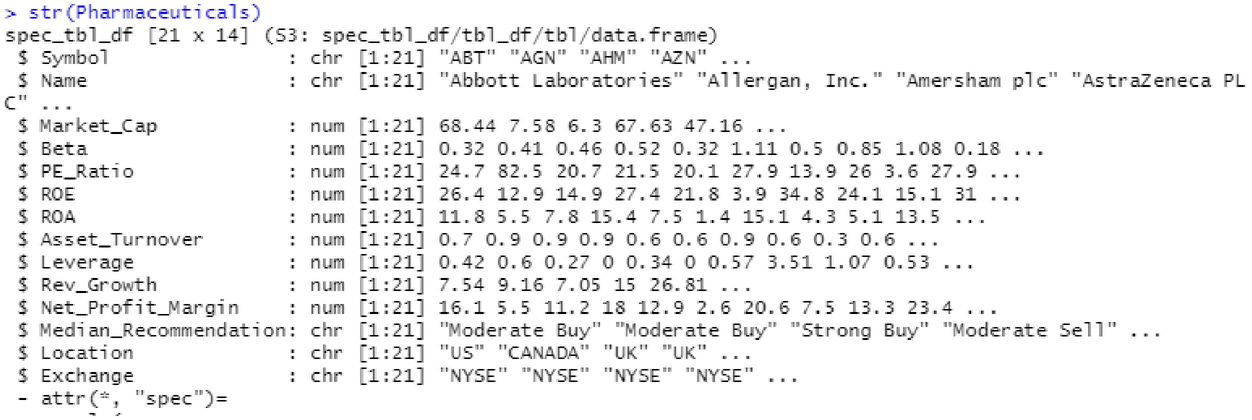
Kasturiarachi-Assignment 4



```
> summary(Pharmaceuticals)
    Symbol              Name             Market_Cap          Beta            PE_Ratio    
 Length:21          Length:21          Min.   :  0.41   Min.   :0.1800   Min.   : 3.60  
 Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500   1st Qu.:18.90  
 Mode  :character   Mode  :character   Median : 48.19   Median :0.4600   Median :21.50  
                                       Mean   : 57.65   Mean   :0.5257   Mean   :25.46  
                                       3rd Qu.: 73.84   3rd Qu.:0.6500   3rd Qu.:27.90  
                                       Max.   :199.47   Max.   :1.1100   Max.   :82.50  
      ROE             ROA        Asset_Turnover     Leverage        Rev_Growth     Net_Profit_Margin
 Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000   Min.   :-3.17   Min.   : 2.6     
 1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600   1st Qu.: 6.38   1st Qu.:11.2     
 Median :22.6   Median :11.20   Median :0.6    Median :0.3400   Median : 9.37   Median :16.1     
 Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857   Mean   :13.37   Mean   :15.7     
 3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000   3rd Qu.:21.87   3rd Qu.:21.1     
 Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100   Max.   :34.21   Max.   :25.5     
 Median_Recommendation   Location           Exchange        
 Length:21             Length:21          Length:21         
 Class :character      Class :character   Class :character  
 Mode  :character      Mode  :character   Mode  :character  
```

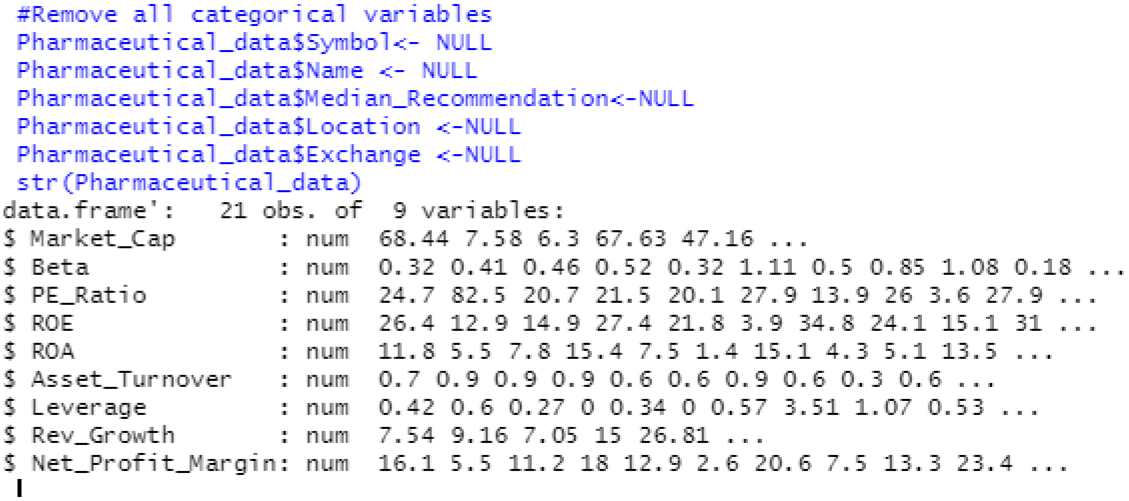


```
> colnames(Pharmaceuticals)
 [1] "Symbol"                "Name"                  "Market_Cap"            "Beta"                 
 [5] "PE_Ratio"              "ROE"                   "ROA"                   "Asset_Turnover"       
 [9] "Leverage"              "Rev_Growth"            "Net_Profit_Margin"     "Median_Recommendation"
[13] "Location"              "Exchange"             
```



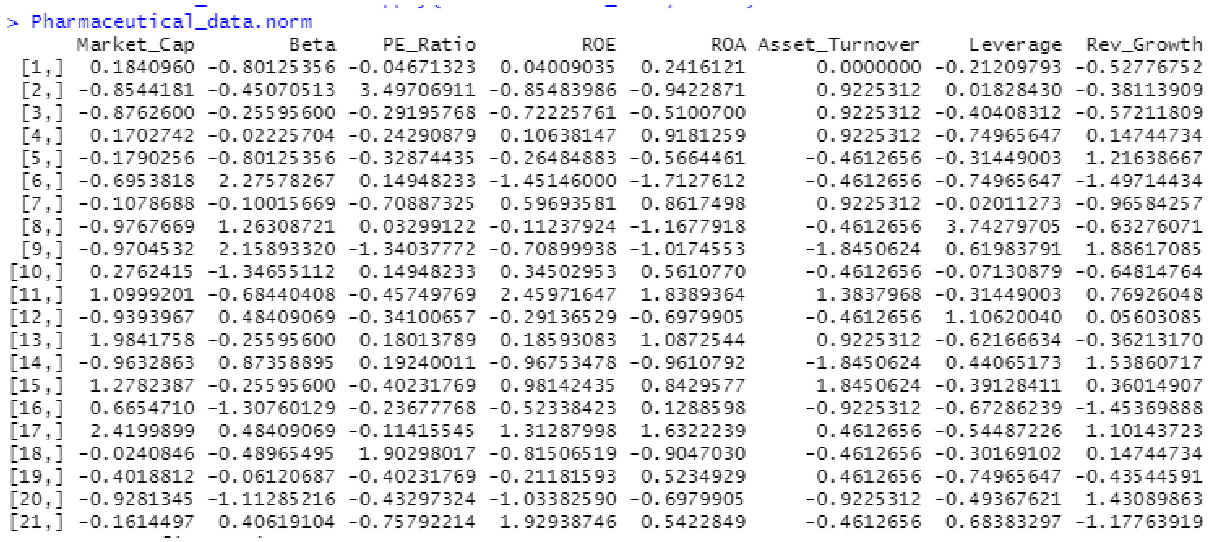
```
> str(Pharmaceuticals)
spec_tbl_df [21 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Symbol               : chr [1:21] "ABT" "AGN" "AHM" "AZN" ...
 $ Name                 : chr [1:21] "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PL
C" ...
 $ Market_Cap           : num [1:21] 68.44 7.58 6.3 67.63 47.16 ...
 $ Beta                 : num [1:21] 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
 $ PE_Ratio             : num [1:21] 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
 $ ROE                  : num [1:21] 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
 $ ROA                  : num [1:21] 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
 $ Asset_Turnover       : num [1:21] 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
 $ Leverage             : num [1:21] 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
 $ Rev_Growth           : num [1:21] 7.54 9.16 7.05 15 26.81 ...
 $ Net_Profit_Margin    : num [1:21] 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
 $ Median_Recommendation: chr [1:21] "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
 $ Location             : chr [1:21] "US" "CANADA" "UK" "UK" ...
 $ Exchange             : chr [1:21] "NYSE" "NYSE" "NYSE" "NYSE" ...
 - attr(*, "spec")=
```

After removing the variables that we do not want to include in the model

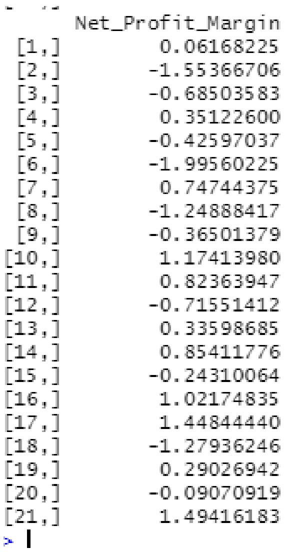


```
#Remove all categorical variables
Pharmaceutical_data$Symbol<- NULL
Pharmaceutical_data$Name <- NULL
Pharmaceutical_data$Median_Recommendation<-NULL
Pharmaceutical_data$Location <-NULL
Pharmaceutical_data$Exchange <-NULL
str(Pharmaceutical_data)
data.frame':   21 obs. of  9 variables:
$ Market_Cap       : num  68.44 7.58 6.3 67.63 47.16 ...
$ Beta             : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
$ PE_Ratio         : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
$ ROE              : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
$ ROA              : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
$ Asset_Turnover   : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
$ Leverage         : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
$ Rev_Growth       : num  7.54 9.16 7.05 15 26.81 ...
$ Net_Profit_Margin: num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
```

Normalizing the data for each variable to be treated equally by the distance measure

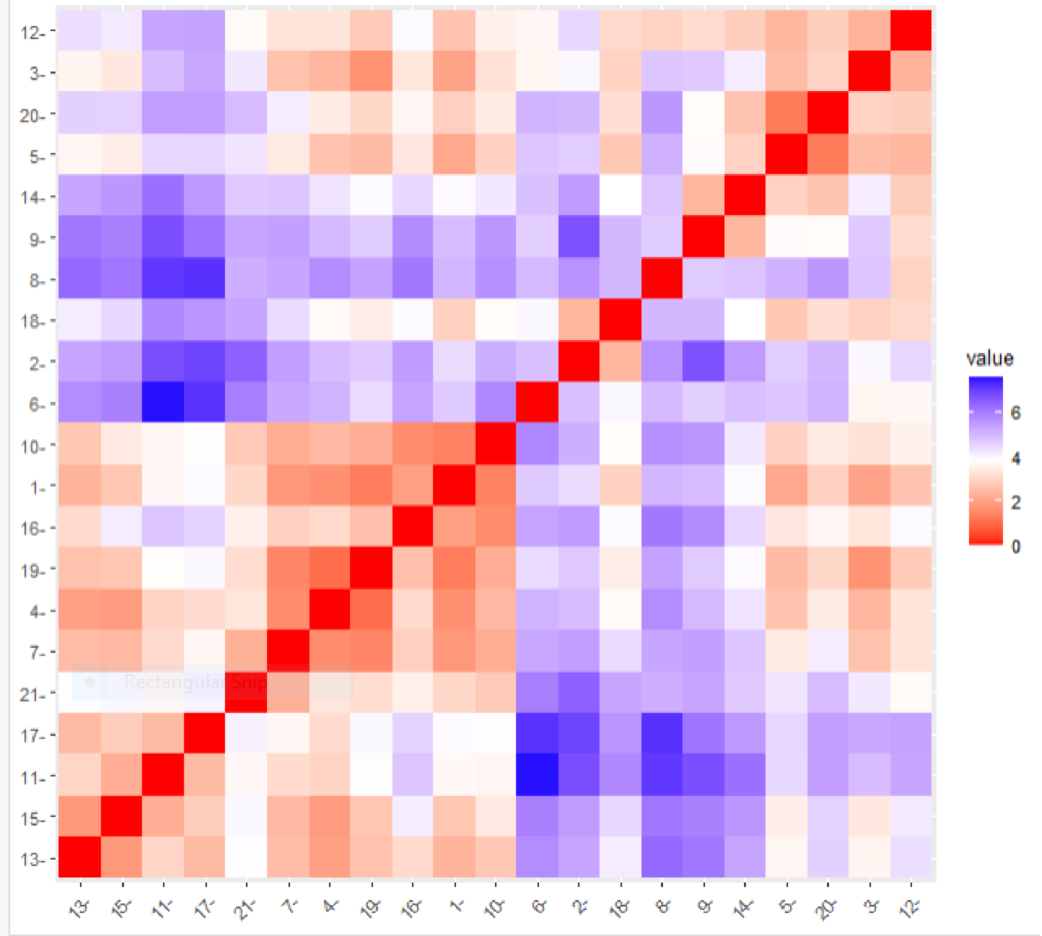


```
> Pharmaceutical_data.norm
       Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover    Leverage  Rev_Growth
 [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000 -0.21209793 -0.52776752
 [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312  0.01828430 -0.38113909
 [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312 -0.40408312 -0.57211809
 [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312 -0.74965647  0.14744734
 [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656 -0.31449003  1.21638667
 [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656 -0.74965647 -1.49714434
 [7,] -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312 -0.02011273 -0.96584257
 [8,] -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656  3.74279705 -0.63276071
 [9,] -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624  0.61983791  1.88617085
[10,]  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656 -0.07130879 -0.64814764
[11,]  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968 -0.31449003  0.76926048
[12,] -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656  1.10620040  0.05603085
[13,]  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312 -0.62166634 -0.36213170
[14,] -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624  0.44065173  1.53860717
[15,]  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624 -0.39128411  0.36014907
[16,]  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312 -0.67286239 -1.45369888
[17,]  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656 -0.54487226  1.10143723
[18,] -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656 -0.30169102  0.14744734
[19,] -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656 -0.74965647 -0.43544591
[20,] -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312 -0.49367621  1.43089863
[21,] -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656  0.68383297 -1.17763919
```



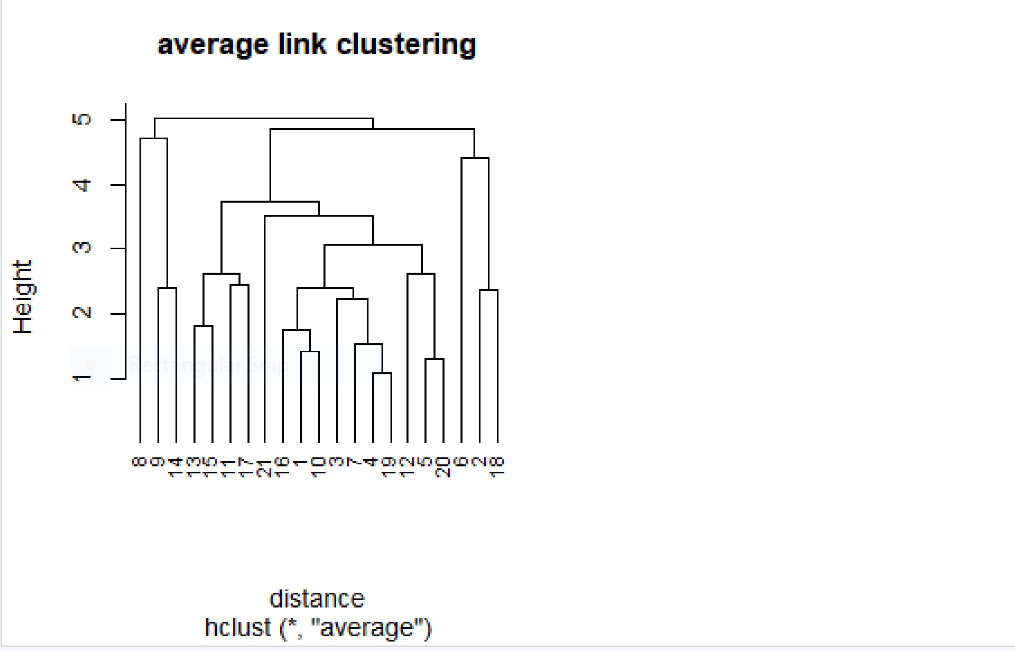
```
      Net_Profit_Margin
 [1,]        0.06168225
 [2,]       -1.55366706
 [3,]       -0.68503583
 [4,]        0.35122600
 [5,]       -0.42597037
 [6,]       -1.99560225
 [7,]        0.74744375
 [8,]       -1.24888417
 [9,]       -0.36501379
[10,]        1.17413980
[11,]        0.82363947
[12,]       -0.71551412
[13,]        0.33598685
[14,]        0.85411776
[15,]       -0.24310064
[16,]        1.02174835
[17,]        1.44844440
[18,]       -1.27936246
[19,]        0.29026942
[20,]       -0.09070919
[21,]        1.49416183
>
```

Let’s look at the distance between observations and using the Euclidean distance to find the similarity between the observations

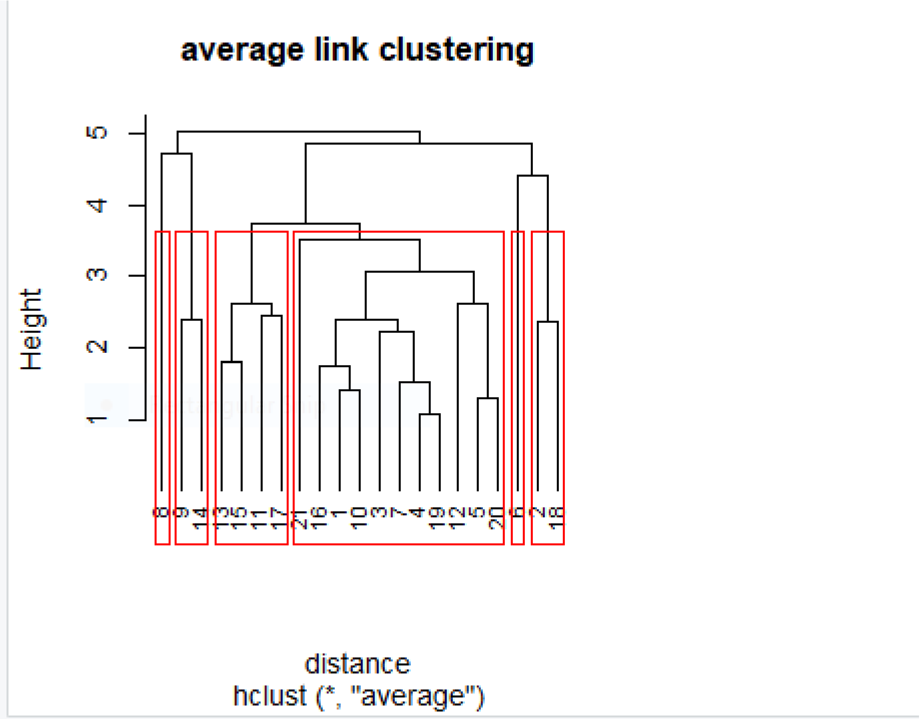


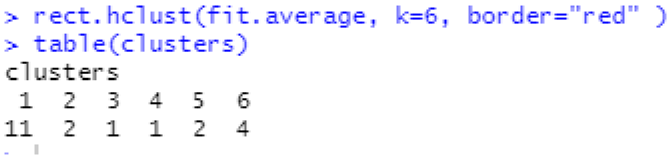
Using the Hierarchical clustering to identify the relationship between individual data points and clusters.

The average linkage helps to find the average distance between two clusters



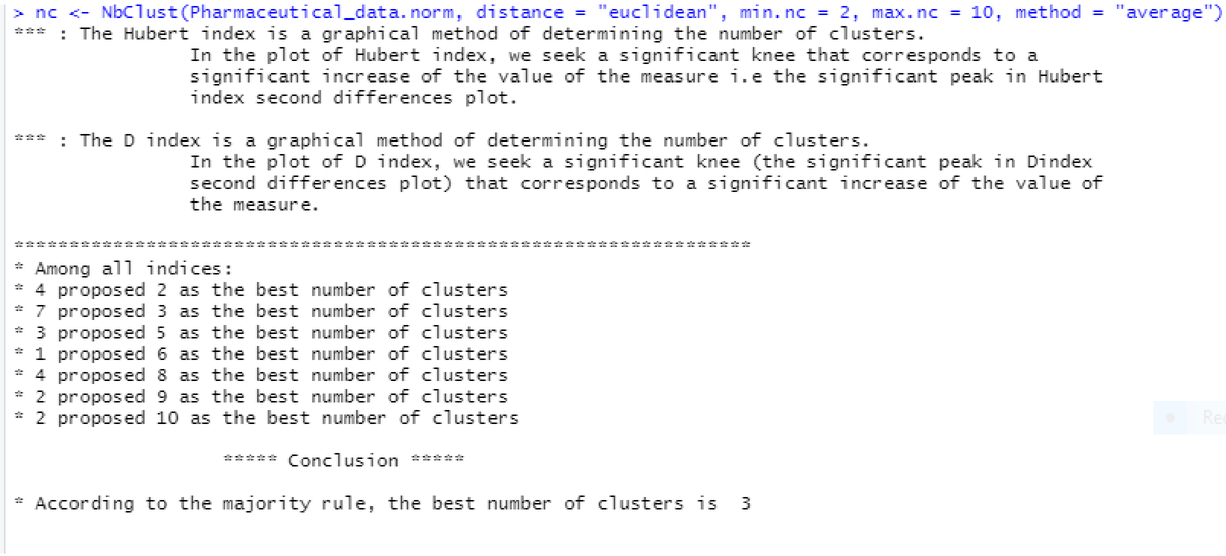
Using the cutree function to cut the dendogram into 6 clusters. Here the dendogram shows the relationship between individual data points and clusters, where the height is the distance between clusters.





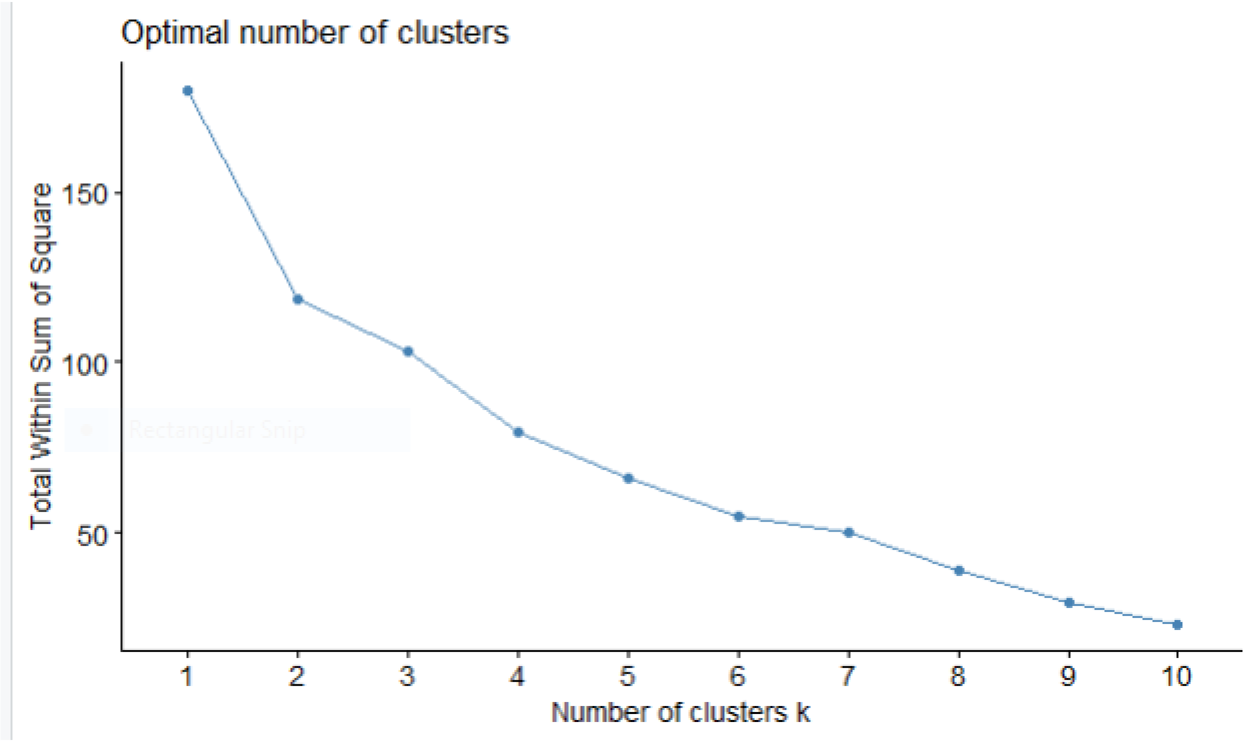
```
> rect.hclust(fit.average, k=6, border="red" )
> table(clusters)
clusters
 1  2  3  4  5  6 
11  2  1  1  2  4 
```

Now to determine the optimal number of clusters using the k-means algorithm.

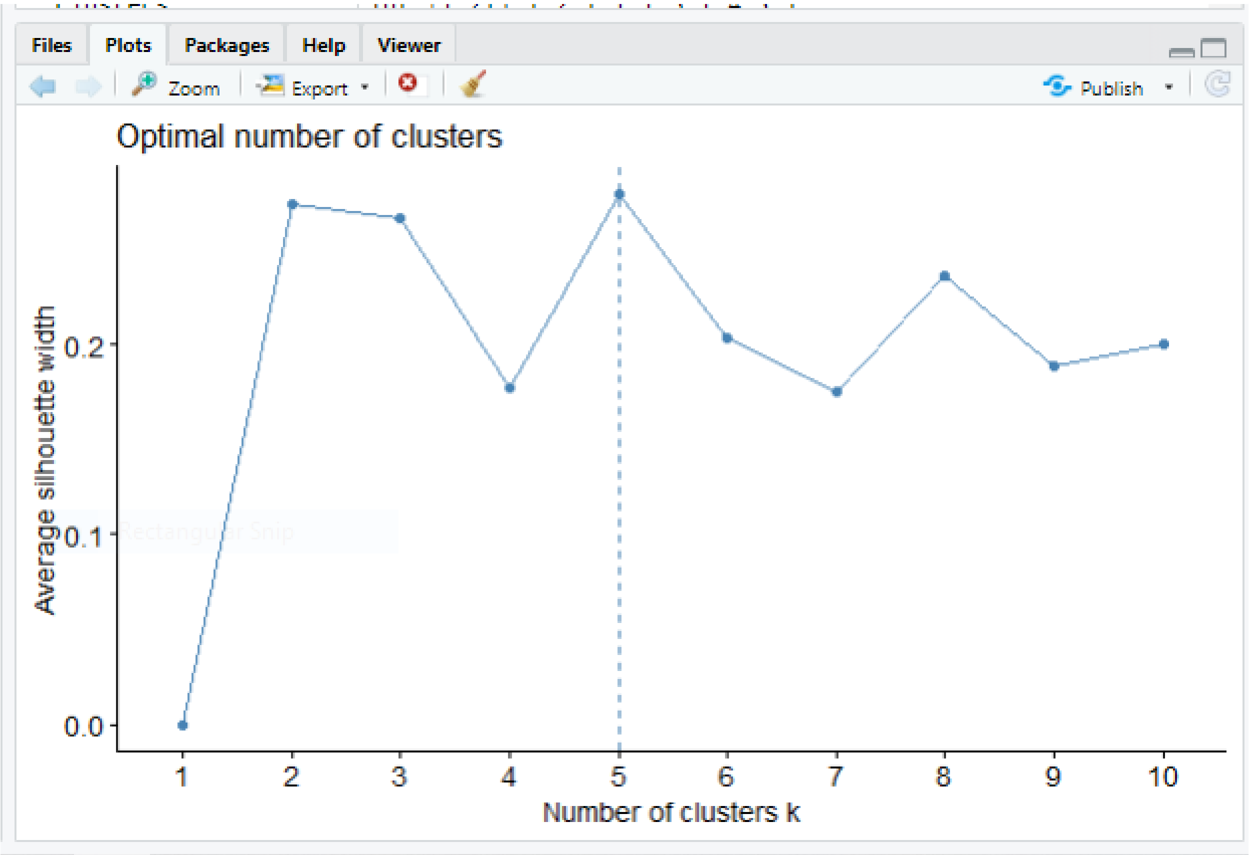


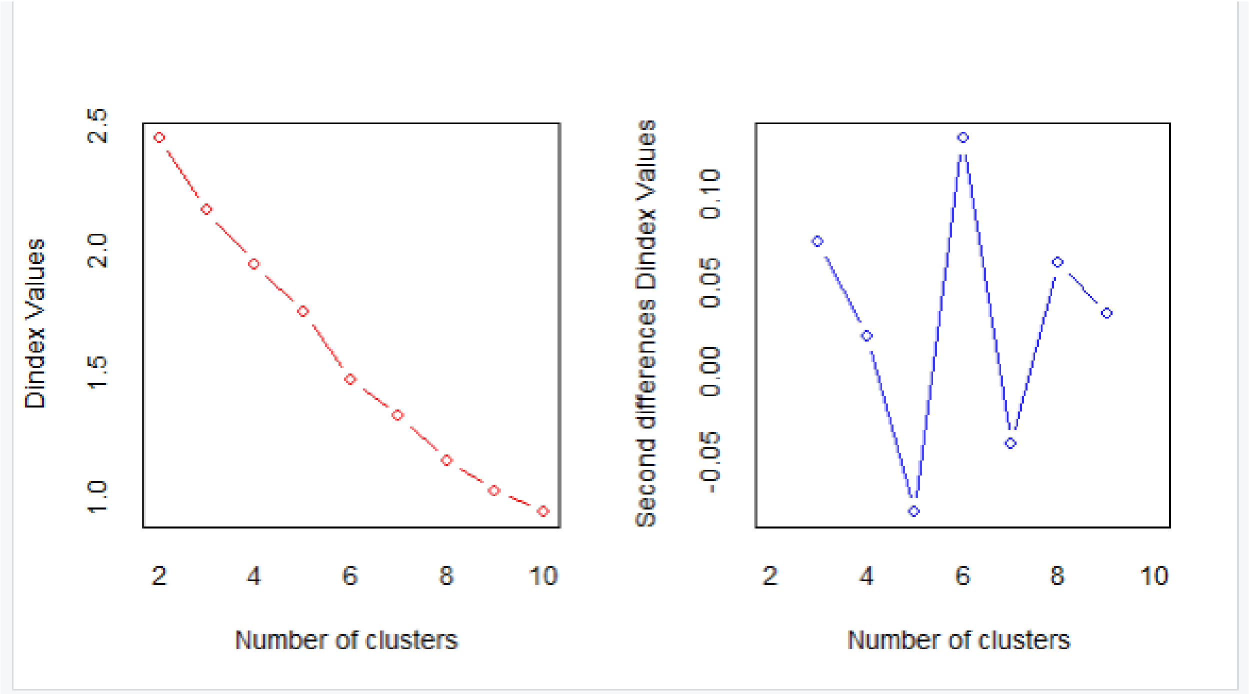
```
> nc <- NbClust(Pharmaceutical_data.norm, distance = "euclidean", min.nc = 2, max.nc = 10, method = "average")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a 
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot. 
 
*** : The D index is a graphical method of determining the number of clusters. 
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure. 
 
******************************************************************* 
* Among all indices:                                                
* 4 proposed 2 as the best number of clusters 
* 7 proposed 3 as the best number of clusters 
* 3 proposed 5 as the best number of clusters 
* 1 proposed 6 as the best number of clusters 
* 4 proposed 8 as the best number of clusters 
* 2 proposed 9 as the best number of clusters 
* 2 proposed 10 as the best number of clusters 

                   ***** Conclusion *****                            
 
* According to the majority rule, the best number of clusters is  3 
```

Elbow method

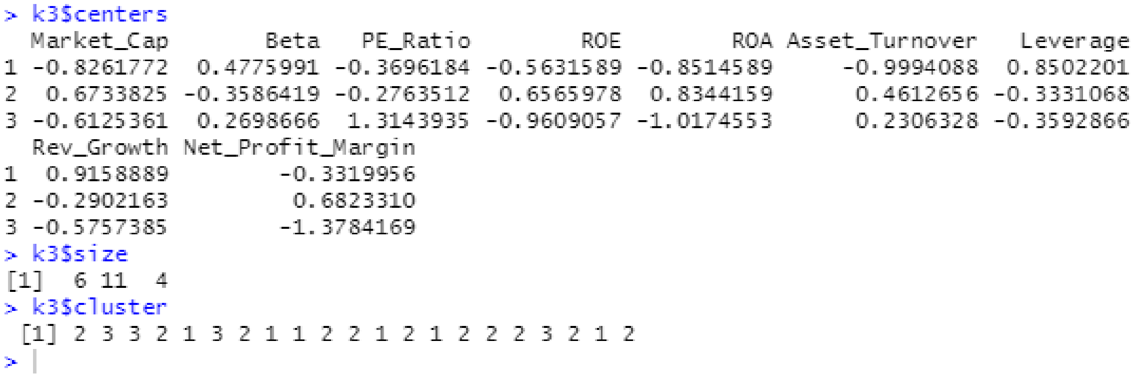


Average Silhouette method- shows best number of clusters is 5

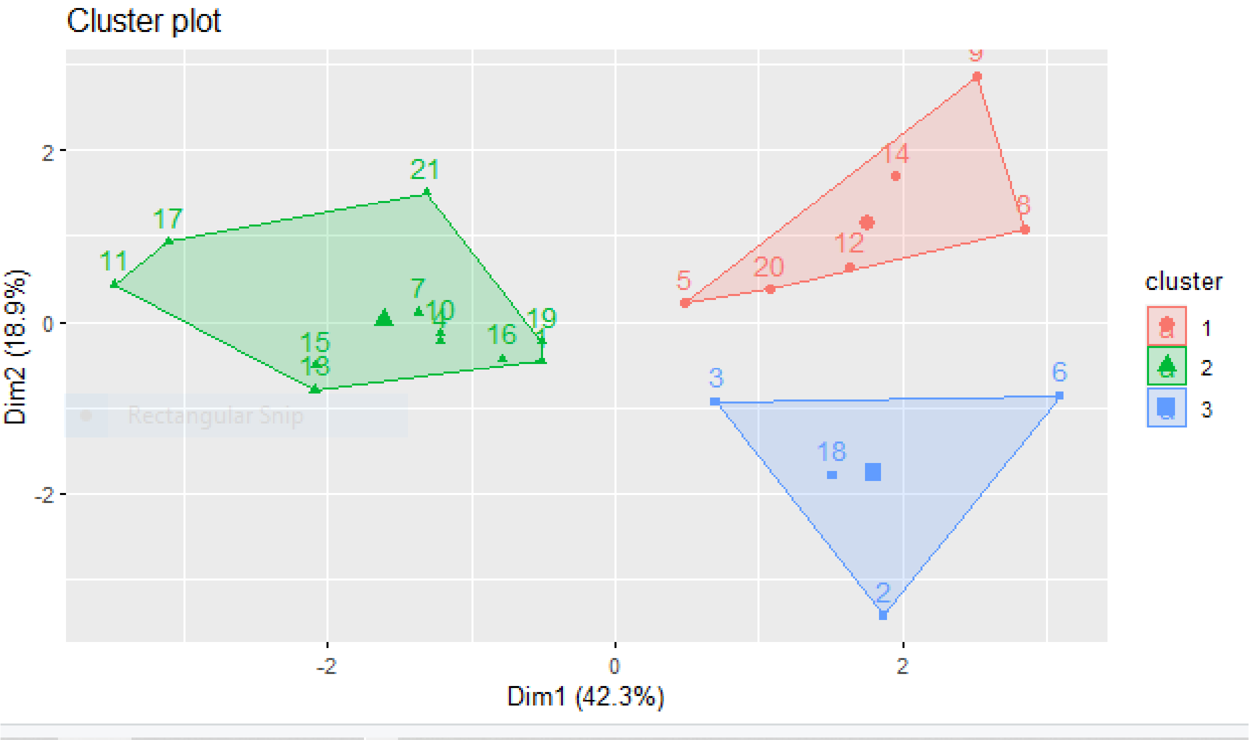




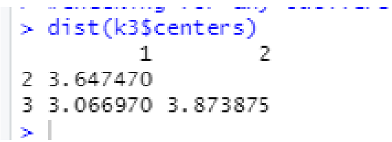
So, I will use 3 clusters to perform the k-means where k=3 and 25 restarts to perform the cluster analysis



```
> k3$centers
  Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover   Leverage
1 -0.8261772  0.4775991 -0.3696184 -0.5631589 -0.8514589     -0.9994088  0.8502201
2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656 -0.3331068
3 -0.6125361  0.2698666  1.3143935 -0.9609057 -1.0174553      0.2306328 -0.3592866
  Rev_Growth Net_Profit_Margin
1  0.9158889        -0.3319956
2 -0.2902163         0.6823310
3 -0.5757385        -1.3784169
> k3$size
[1]  6 11  4
> k3$cluster
 [1] 2 3 3 2 1 3 2 1 1 2 2 1 2 1 2 2 2 3 2 1 2
> |
```



Checking for any outliers



```
> dist(k3$centers)
         1        2
2 3.647470
3 3.066970 3.873875
> 
```

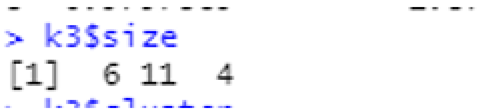
(b)

Now that we have the pharmaceutical companies belonging to one of the 3 clusters. We need to dive into each cluster to analyze the characteristics and variables.

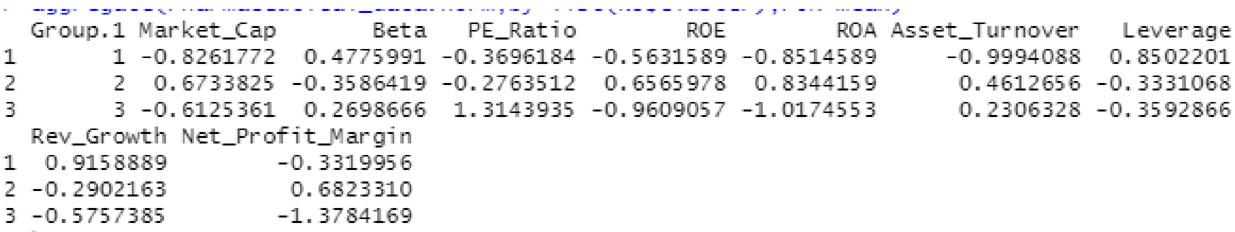
Cluster 1-rows 5, 8, 9, 12, 14, 20

Cluster 2-rows 4,7,10,11, 15, 16,17,18, 19,21

Cluster 3-rows 2, 3, 6, 18



```
> k3$size
[1]  6 11  4
```



```
  Group.1 Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover   Leverage
1       1 -0.8261772  0.4775991 -0.3696184 -0.5631589 -0.8514589     -0.9994088  0.8502201
2       2  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656 -0.3331068
3       3 -0.6125361  0.2698666  1.3143935 -0.9609057 -1.0174553      0.2306328 -0.3592866
  Rev_Growth Net_Profit_Margin
1  0.9158889        -0.3319956
2 -0.2902163         0.6823310
3 -0.5757385        -1.3784169
```

This indicates that there are 6 pharmaceuticals in cluster 1, 11 pharmaceuticals in cluster 2 and 4 pharmaceuticals in cluster 3

Cluster 1- is characterized by companies that high revenue growth, low ROA and ROE, lowest Market cap and Asset Turnover, highly leveraged, high beta more volatile

Cluster 2- companies that have high net profit margin, high ROE and ROA and highest market Cap

Cluster 3- companies with lowest revenue growth and profit margin, high PE ratio, low ROE, and ROA

The Hierarchical cluster algorithm shows Cluster 1 and Cluster 2 have similar pattern, consisting of companies that are highly profitable and low risk investment. On the other hand, Cluster 3 is comprised of not-for-profit companies with low leverage, and less debt. Moreover, Cluster 1 includes not-for-profit companies with high volatility due to high beta and considering revenue growth and high level of leverage like Cluster 3. The stock price is undervalued.

C) Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

Yes, there is a pattern in the clusters with respect to variable Median Recommendation categorized into Moderate Buy, Moderate Sell, Strong Buy, and hold.

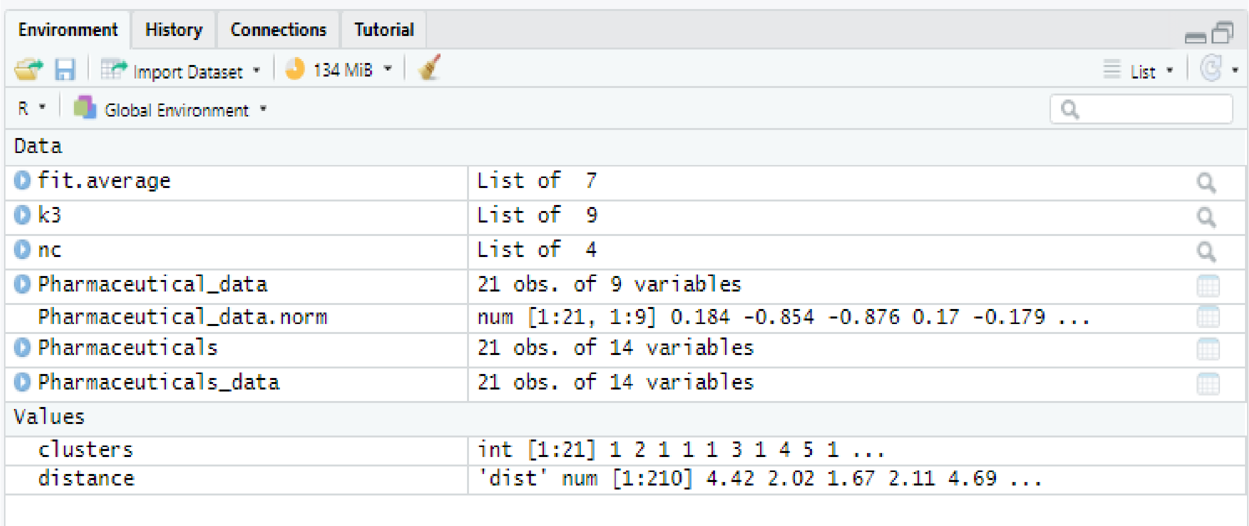
When comparing the clusters, Cluster 2 has the highest Market Cap, highest ROE and ROA, and highest Asset Turnover, yet does not indicate a moderate sell. On the other hand, Cluster 3 with high PE ratio, lowest revenue growth and profit margin, lowest ROE, ROA, and lowest asset turnover has a strong buy recommendation as the stock price is undervalued. Cluster 1 high beta, leverage, revenue growth and low market cap on hold recommendation.

(d) Naming the clusters according to variables

Cluster 1- highest revenue growth, highest leverage, highest beta, low ROA and ROE, lowest Market cap and Asset Turnover-Risky yet high revenue

Cluster 2 -High ROA, high ROE, high net profit and least risk- Moderate Buy

Cluster 3 - lowest revenue growth and profit margin, high PE ratio, low ROE, and ROA- Strong Buy



```
Data
fit.average                   List of  7
k3                            List of  9
nc                            List of  4
Pharmaceutical_data           21 obs. of 9 variables
Pharmaceutical_data.norm      num [1:21, 1:9] 0.184 -0.854 -0.876 0.17 -0.179 ...
Pharmaceuticals               21 obs. of 14 variables
Pharmaceuticals_data          21 obs. of 14 variables
Values
clusters                      int [1:21] 1 2 1 1 1 3 1 4 5 1 ...
distance                      'dist' num [1:210] 4.42 2.02 1.67 2.11 4.69 ...
```