FML clustering assignment

2023-11-14

Summary;

Data Exploration

1. **Data Load and Summary**: Loaded the ‘Pharmaceuticals.csv’ dataset, observed its structure (21 observations, 14 variables), and identified factor variables representing Symbol, Name, Median_Recommendation, Location, and Exchange.
2. **Initial Visualization**: Displayed the structure of the dataset and visualized pairs of numerical variables to observe any potential clusters visually.

Cluster Analysis Preparation

3. **Data Preprocessing**: Removed factor variables and performed scaling to normalize the data for cluster analysis.
4. **Distance Matrix Calculation**: Computed the distance matrix to measure the dissimilarity between observations.

Hierarchical Agglomerative Clustering

5. **Complete Linkage Method**: Utilized hierarchical agglomerative clustering with complete linkage to create dendrograms and identified potential clusters based on visual inspection of the dendrogram.
6. **Average Linkage Method**: Conducted hierarchical agglomerative clustering with average linkage to explore alternative clustering structures.

Cluster Membership and Characterization

7. **Cluster Membership Analysis**: Determined the number of clusters and their membership using both methods, then characterized the clusters by aggregating numerical features and observing silhouette plots for cluster validation.
8. **K-means Clustering**: Performed K-means clustering with k=3 and evaluated cluster qualities using sum of squares and silhouette plots, comparing the results to hierarchical clustering.

Pattern Analysis and Interpretation

9. **Comparison Across Methods**: Analyzed the patterns within clusters concerning Median_Recommendation, Location, and Exchange across different clustering methodologies.

Insights for Equities Analyst

10. **Utilization for Equities Analysis**: Explored the implications of the clustering analysis for equities analysts studying the pharmaceutical industry, covering investment strategies, risk assessment, industry insights, investment allocation, strategy adaptation, and risk mitigation.

Summary and Cluster Insights

11. **Cluster Insights and Naming**:

- Identified three key clusters based on recurring characteristics observed across multiple methodologies.
- Provided insights into stability, global consistency, and investment diversity within these clusters.

12. **Cluster Naming and Rationale**:

- Named clusters based on dominant characteristics observed across methodologies, emphasizing stability, global presence, and investment diversity.

```
# Data Exploration
df <- read.csv('Pharmaceuticals.csv')
getwd()
```

```
## [1] "C:/Users/Lenovo/OneDrive/Desktop/FML"
```

```
summary(df)
```

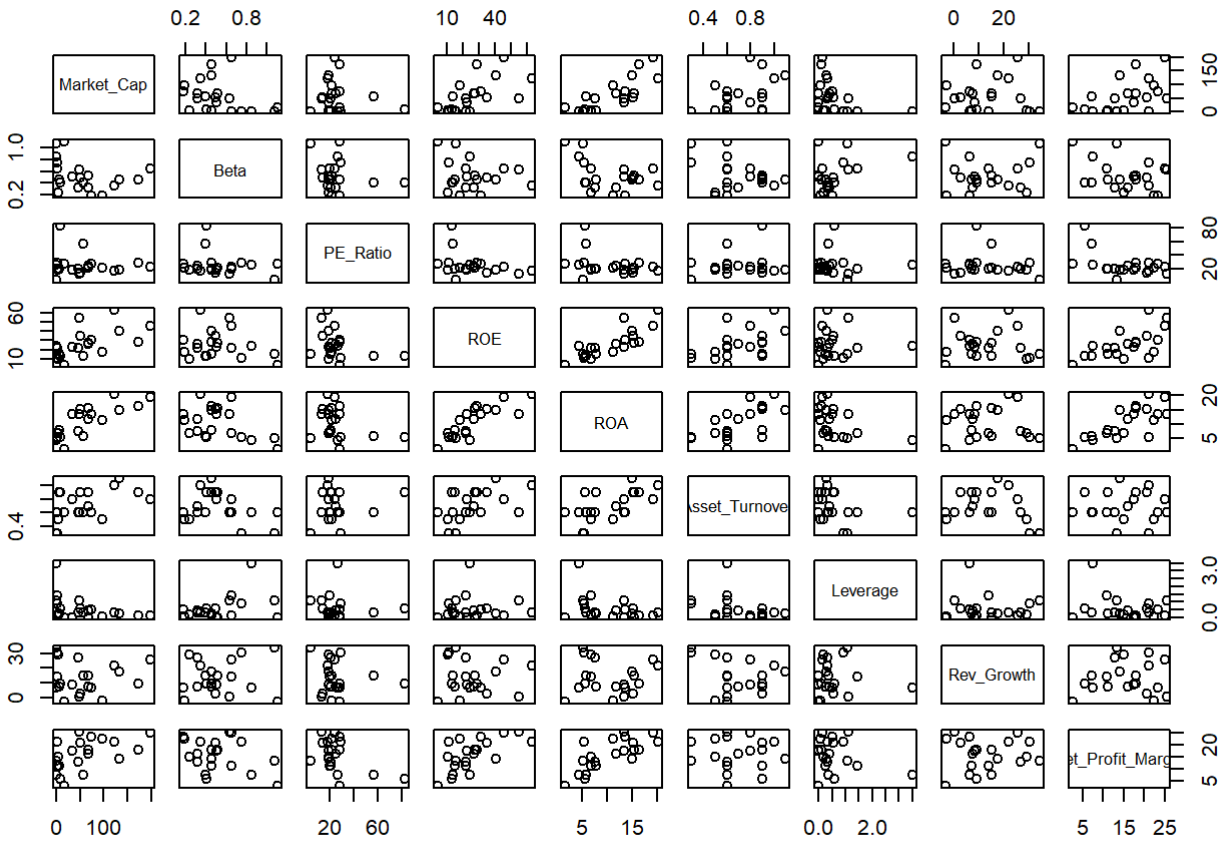
```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##   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

The dataset shows 21 observations and it includes 14 variables. The variable number 1,2,12,13 & 14 are factor variables representing Symbol, Name, Median_Recommendation, Location & Exchange. In order to start cluster analysis, we are going to remove the factor variables from our analysis, because cluster analysis is only ment for nonfactor variables or numerical variables.

```
head(df)
```

```
##   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
```

```
pairs(df[3:11])
```




```
# Scatter plot
plot(df$Rev_Growth~ df$Net_Profit_Margin, df=df)
```

```
## Warning in plot.window(...): "df" is not a graphical parameter
```

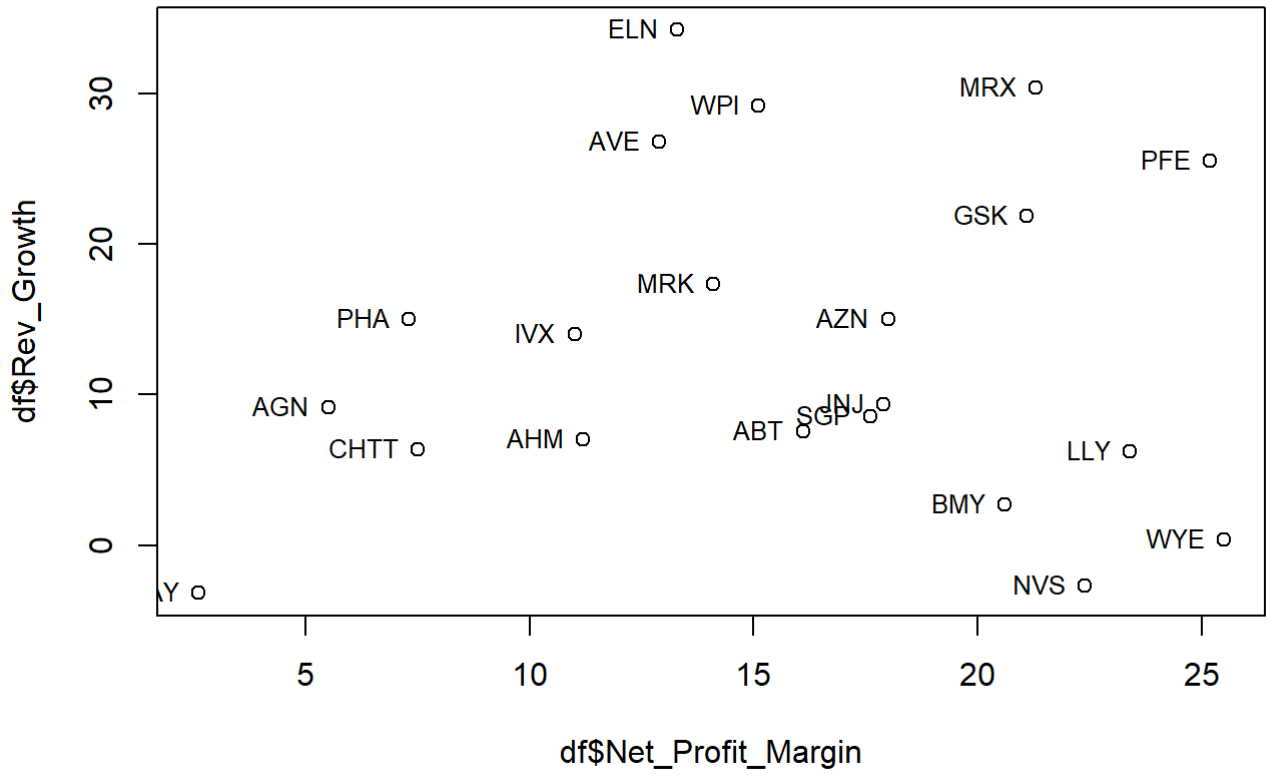
```
## Warning in plot.xy(xy, type, ...): "df" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "df" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "df" is not a
## graphical parameter
```

```
## Warning in box(...): "df" is not a graphical parameter
```

```
## Warning in title(...): "df" is not a graphical parameter
```

```
with(df,text(df$Rev_Growth ~ df$Net_Profit_Margin, labels=df

             $Symbol,pos=2, cex= 0.8))
```



From the depicted graph, focusing on the variables Rev Growth and Net Profit Margin, it's evident that four companies, represented by symbols in the right bottom corner, share characteristics suggesting the potential to form a cluster. These companies exhibit high net profit margins and low revenue growth. Conversely, in the right top corner, three companies, indicated by symbols, demonstrate both high net profit margins and robust revenue growth, hinting at the possibility of forming another distinct cluster. For the remaining companies, there is some overlap.Lets do cluster analysis .

```
#Scaling
z = df[, -c(1,2,12,13,14)]
df_scale = scale(z)
df_scale
```

```
##       Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
##  [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
##  [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
##  [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
##  [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
##  [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
##  [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
##  [7,] -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
##  [8,] -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
##  [9,] -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## [10,]  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## [11,]  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## [12,] -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## [13,]  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## [14,] -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## [15,]  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## [16,]  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## [17,]  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## [18,] -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## [19,] -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## [20,] -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## [21,] -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
##          Leverage  Rev_Growth Net_Profit_Margin
##  [1,] -0.21209793 -0.52776752        0.06168225
##  [2,]  0.01828430 -0.38113909       -1.55366706
##  [3,] -0.40408312 -0.57211809       -0.68503583
##  [4,] -0.74965647  0.14744734        0.35122600
##  [5,] -0.31449003  1.21638667       -0.42597037
##  [6,] -0.74965647 -1.49714434       -1.99560225
##  [7,] -0.02011273 -0.96584257        0.74744375
##  [8,]  3.74279705 -0.63276071       -1.24888417
##  [9,]  0.61983791  1.88617085       -0.36501379
## [10,] -0.07130879 -0.64814764        1.17413980
## [11,] -0.31449003  0.76926048        0.82363947
## [12,]  1.10620040  0.05603085       -0.71551412
## [13,] -0.62166634 -0.36213170        0.33598685
## [14,]  0.44065173  1.53860717        0.85411776
## [15,] -0.39128411  0.36014907       -0.24310064
## [16,] -0.67286239 -1.45369888        1.02174835
## [17,] -0.54487226  1.10143723        1.44844440
## [18,] -0.30169102  0.14744734       -1.27936246
## [19,] -0.74965647 -0.43544591        0.29026942
## [20,] -0.49367621  1.43089863       -0.09070919
## [21,]  0.68383297 -1.17763919        1.49416183
## attr(,"scaled:center")
##       Market_Cap             Beta         PE_Ratio              ROE
##       57.6514286        0.5257143       25.4619048       25.7952381
##              ROA    Asset_Turnover         Leverage       Rev_Growth
##       10.5142857        0.7000000        0.5857143       13.3709524
## Net_Profit_Margin
##       15.6952381
## attr(,"scaled:scale")
##       Market_Cap             Beta         PE_Ratio              ROE
##       58.6029595        0.2567406       16.3102568       15.0849752
##              ROA    Asset_Turnover         Leverage       Rev_Growth
```

```
##         5.3213988         0.2167948         0.7813103        11.0483351
## Net_Profit_Margin
##         6.5620482
```

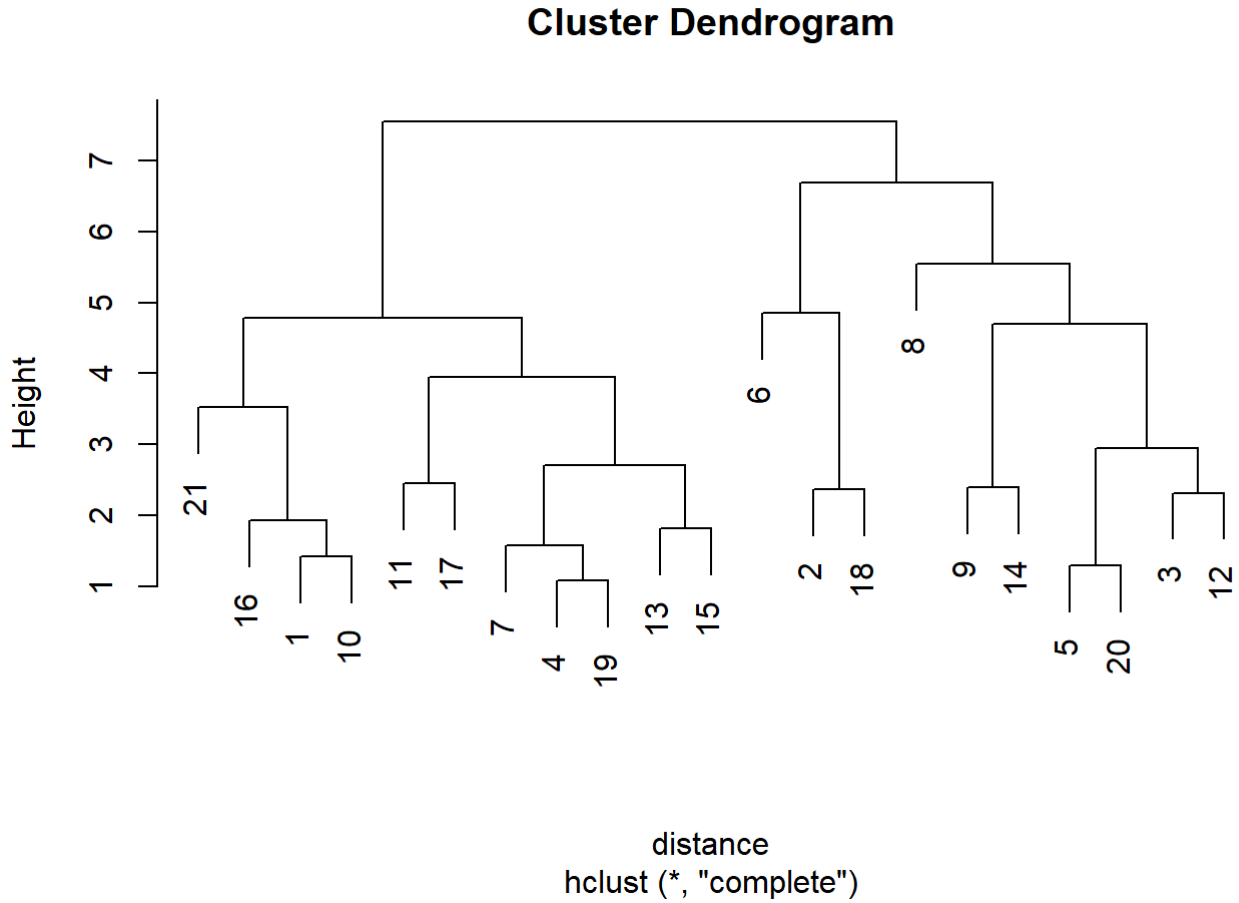
In order to do the cluster analysis, we need to normalize our dataset to be able to compare them to each other. After that we are going to calculate the distance matrix based on all the variables to see which companies have the lowest distance to each other which will enable them to form the same cluster and which companies have the highest distance which will put them in different clusters.

```
# Calculate distance matrix
distance = dist(df_scale)
print(distance, digits = 3)
```

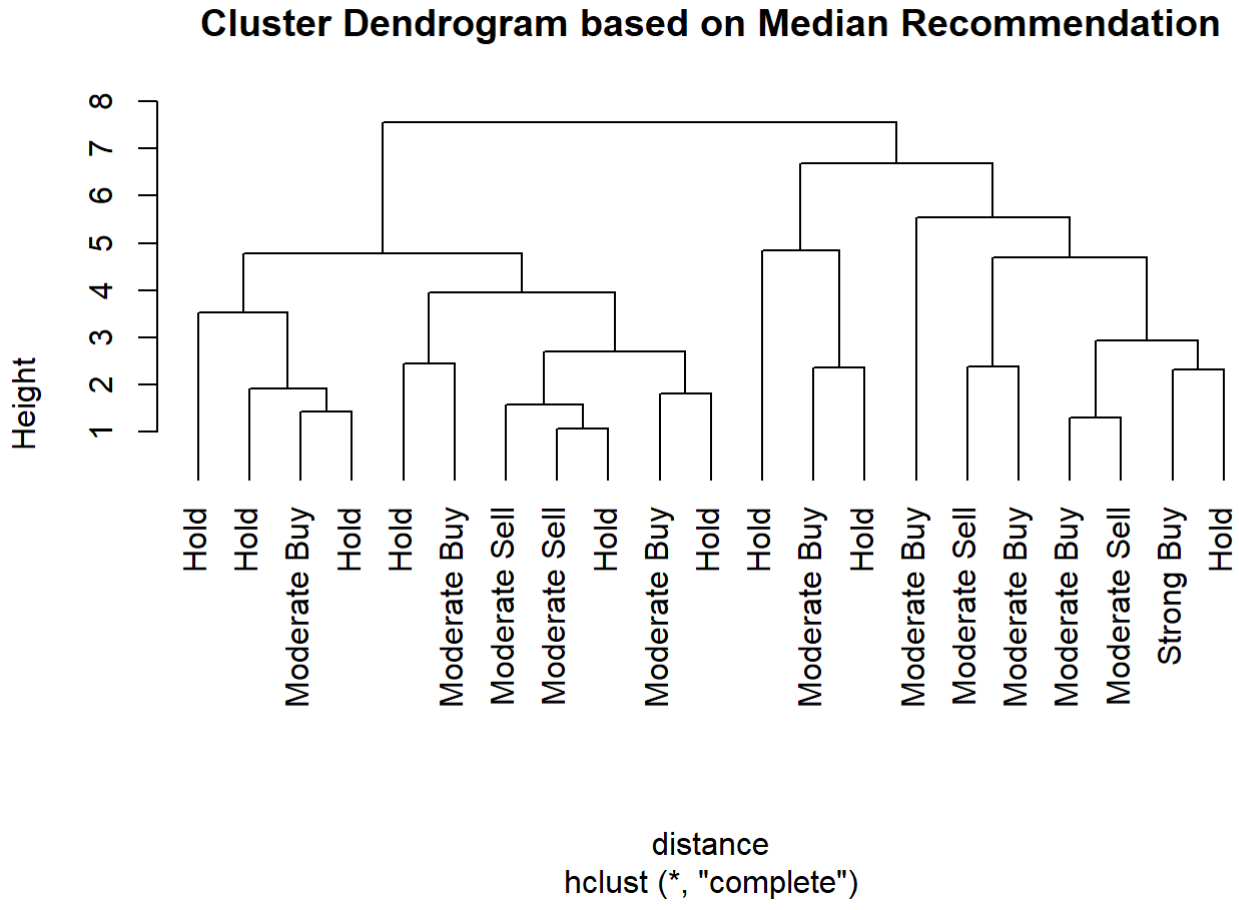
```
##       1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 2  4.42
## 3  2.02 3.95
## 4  1.67 4.91 2.36
## 5  2.11 4.64 2.49 2.63
## 6  4.69 4.85 3.64 5.07 4.76
## 7  1.81 5.42 2.60 1.57 3.40 5.27
## 8  5.02 5.61 4.76 5.72 5.10 4.97 5.29
## 9  4.90 6.70 4.70 4.97 3.75 4.61 5.38 4.68
## 10 1.42 5.14 3.24 2.41 2.91 5.80 2.19 5.66 5.55
## 11 3.69 6.75 4.90 2.96 4.48 7.55 3.10 7.08 6.73 3.63
## 12 2.62 4.47 2.32 3.28 2.39 3.66 3.28 2.95 3.12 3.54 5.28
## 13 2.33 5.32 3.59 1.96 3.64 5.72 2.51 6.31 6.07 2.72 2.99 4.35
## 14 3.92 5.48 4.12 4.27 2.93 4.85 4.73 4.79 2.39 4.19 6.19 2.83 5.31
## 15 2.68 5.44 3.36 1.86 3.47 5.92 2.43 6.10 5.92 3.38 2.22 4.16 1.81 5.53
## 16 1.92 5.47 3.33 3.06 3.33 5.33 2.87 6.06 5.73 1.58 4.78 3.90 3.08 4.48 4.11
## 17 3.89 6.91 5.27 3.11 4.50 7.16 3.67 7.18 6.12 3.78 2.45 5.36 2.45 5.52 2.83
## 18 2.91 2.37 2.93 3.72 2.72 3.96 4.41 5.00 5.01 3.75 5.77 3.07 4.11 3.83 4.45
## 19 1.31 4.73 1.70 1.08 2.46 4.43 1.48 5.35 4.67 2.21 3.78 2.76 2.60 3.91 2.71
## 20 2.88 5.01 2.94 3.41 1.30 5.06 4.12 5.54 3.76 3.41 5.44 2.86 4.59 2.65 4.57
## 21 3.04 6.45 4.19 3.32 4.25 5.95 2.27 5.13 5.31 2.75 3.67 3.72 3.86 4.71 3.94
##      16   17   18   19   20
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17 4.54
## 18 3.88 5.59
## 19 2.54 3.96 3.45
## 20 3.63 5.40 3.17 3.03
## 21 3.53 4.03 5.29 3.15 4.92
```

Now that we are ready to perform cluster analysis, we are going to use Hierarchical agglomerative clustering method with Complete Linkage as the first method.

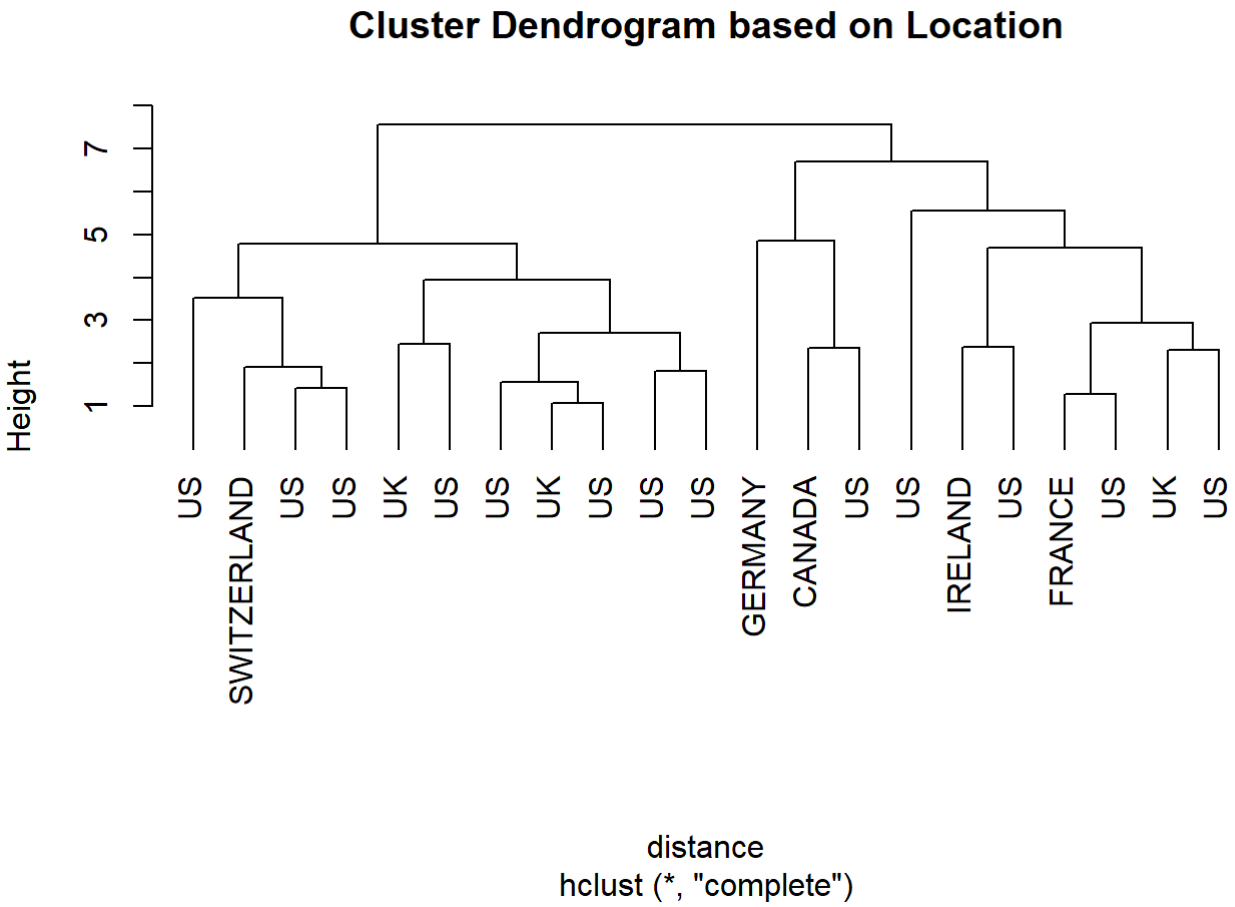
```
# Hierarchical agglomerative clustering
#Cluster Dendogram with Complete Linkage


hc.c <- hclust(distance)
plot(hc.c)
plot(hc.c,labels=df$Company,main='Cluster Dendrogram')
```



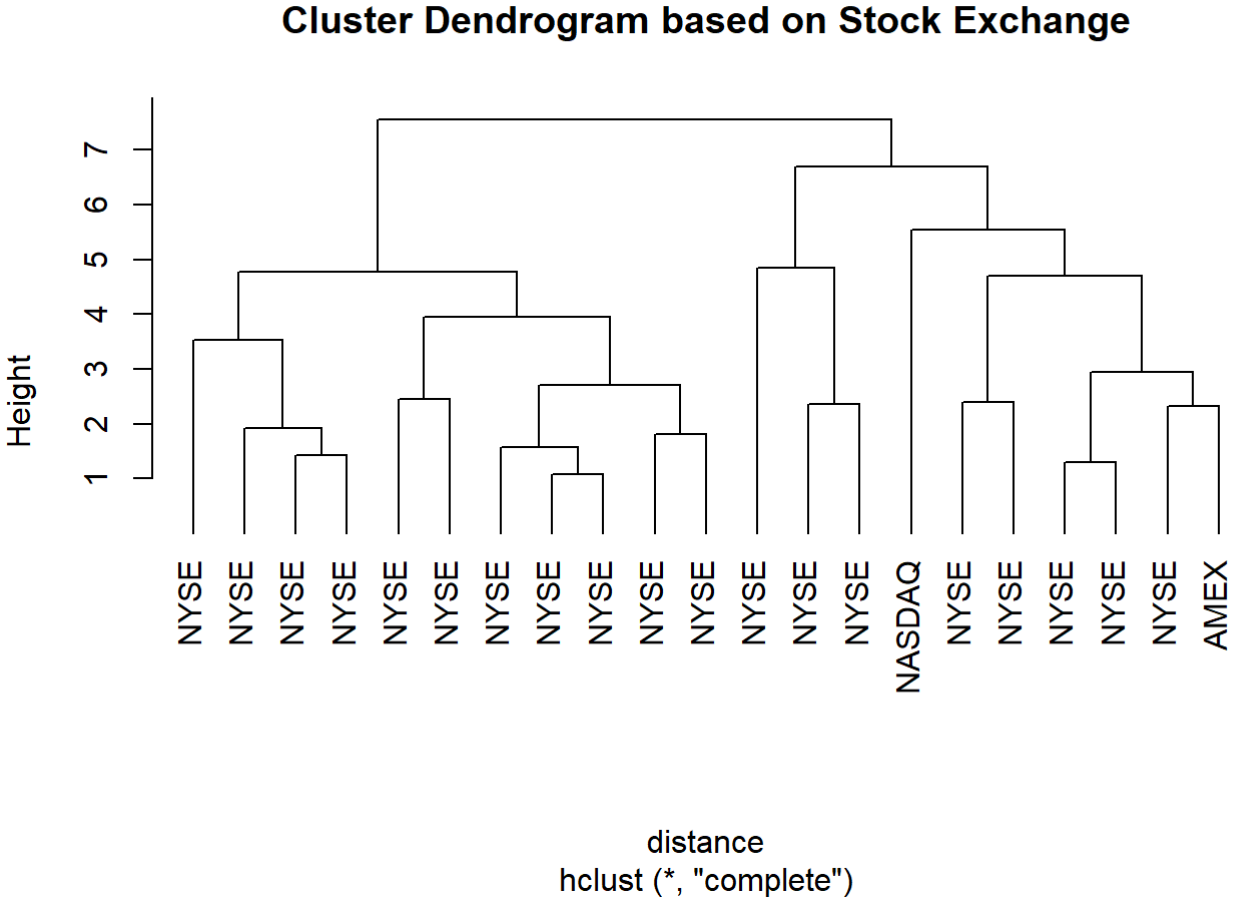

```
plot(hc.c,hang=-1, labels = df$Median_Recommendation, main='Cluster Dendrogram based on Media
n Recommendation')
```



```
plot(hc.c,hang=-1, labels = df$Location, main='Cluster Dendrogram based on Location')
```



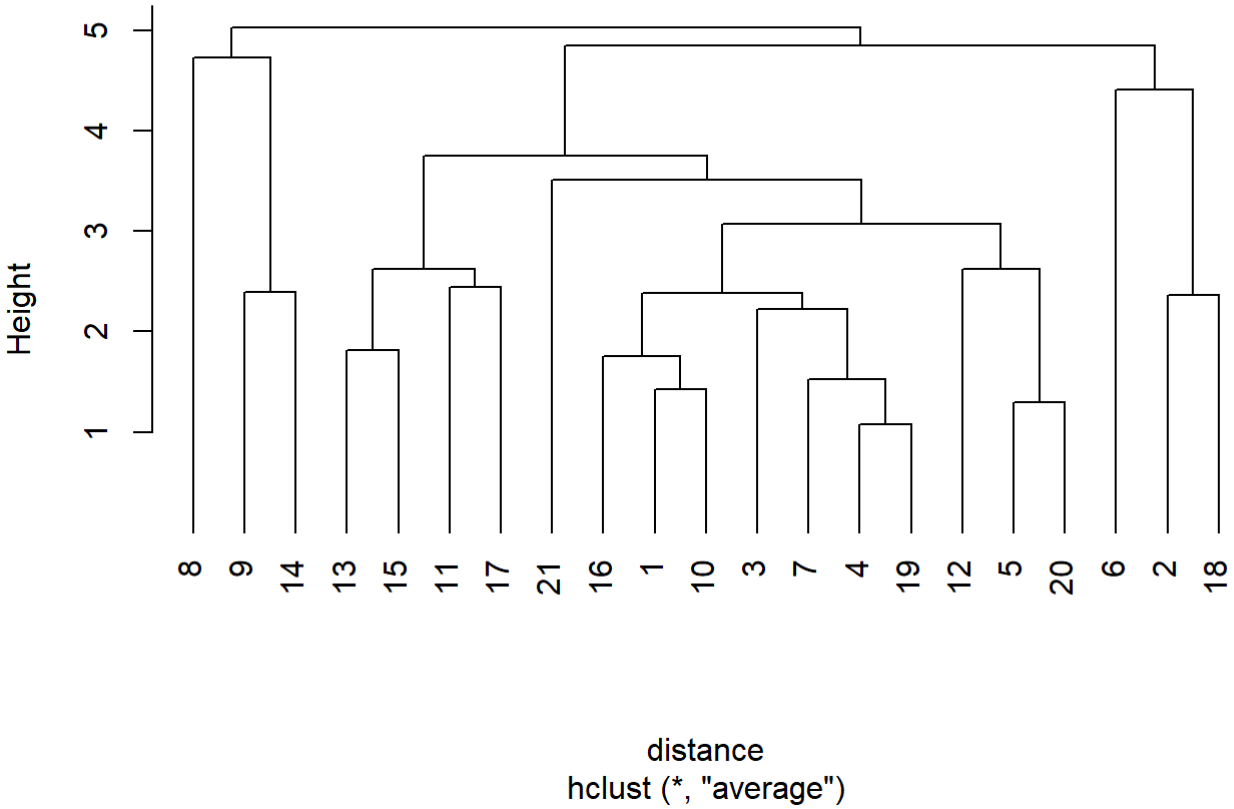

```
plot(hc.c,hang=-1, labels = df$Exchange, main='Cluster Dendrogram based on Stock Exchange')
```




```
#In this graph we can say that companies 2 and 18  have formed as 1 cluster at first, then 2, 18,6 have made a bigger cluster and as we go up, more companies will be formed in one cluster and as we can see at the height of 7, we can divide the whole data in 2 big clusters. Based o n this graph, we can suggest maybe choosing 3 clusters at the height 6 would be good for this dataset.
```

As the second method, we are going to use Hierarchical agglomerative clustering method with Average Linkage. As we can see in the graph, companies number 4 & 19 are again formed firstly to one cluster and then with company 7 they are making a bigger cluster and at the height 5, we have the minimum number of clusters which is 3.

```
# Hierarchical agglomerative clustering using "average" linkage
#Cluster Dendogram with Average Linkage
hc_a<-hclust(distance,method="average")
plot(hc_a,hang=-1)
```



Now lets make a cluster membership table to compare these two methods. If we set the number of clusters with each of the methods to 3, then we would get the following table:

```
# Cluster membership
member = cutree(hc.c,3)
table(member)
```

```
## member
##  1  2  3
## 11  3  7
```

```
member.c <- cutree(hc.c,3)
member.a <- cutree(hc_a,3)
table(member.c, member.a)
```

```
##          member.a
## member.c  1  2  3
##        1 11  0  0
##        2  0  3  0
##        3  4  0  3
```

As we can see from the table, with the complete linkage method, 11 companies belong to cluster 1, 3 companies belong to cluster 2 & 7 companies belong to cluster 3.

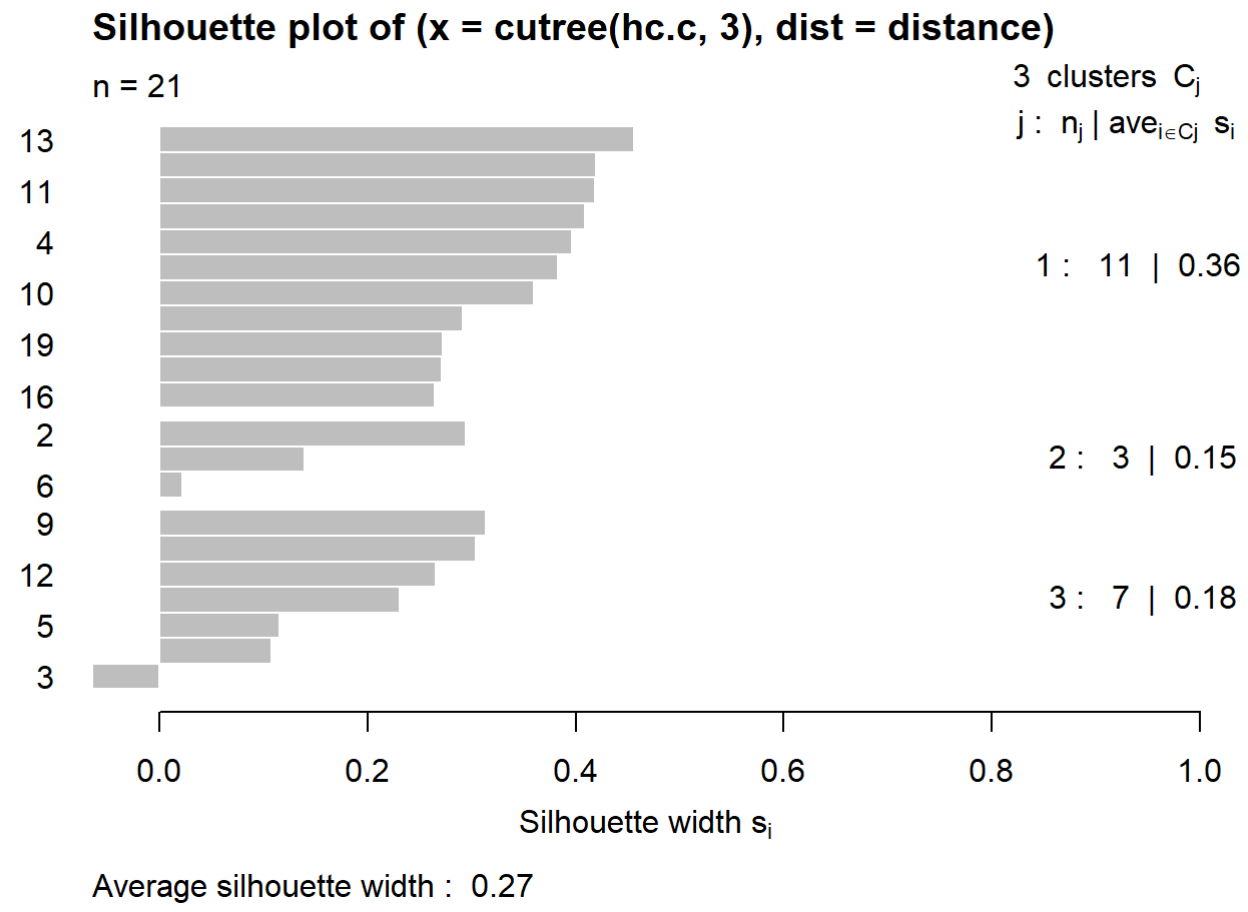
```
# Characterizing clusters
aggregate(df_scale,list(member),mean)
```

```
##   Group.1 Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1       1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159   4.612656e-01
## 2       2 -0.5246281  0.4451409  1.8498439 -1.0404550 -1.1865838   1.480297e-16
## 3       3 -0.8333319  0.3728055 -0.3585240 -0.5858873 -0.8026890  -7.248459e-01
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163          0.682331
## 2 -0.3443544 -0.5769454         -1.609544
## 3  0.6710340  0.7033165         -0.382430
```

```
aggregate(df[, -c(1,2,12,13,14)],list(member),mean)
```

```
##   Group.1 Market_Cap      Beta PE_Ratio      ROE       ROA Asset_Turnover
## 1       1  97.113636 0.4336364 20.95455 35.70000 14.954545      0.8000000
## 2       2  26.906667 0.6400000 55.63333 10.10000  4.200000      0.7000000
## 3       3   8.815714 0.6214286 19.61429 16.95714  6.242857      0.5428571
##    Leverage Rev_Growth Net_Profit_Margin
## 1 0.3254545  10.164545         20.172727
## 2 0.3166667   6.996667          5.133333
## 3 1.1100000  21.141429         13.185714
```

```
# Silhouette Plot
library(cluster)
plot(silhouette(cutree(hc.c,3), distance))
```



Cluster 1: Si values are very high. This means that the members within Cluster 1 are quite similar to each other, indicating a successful formation of this cluster. Cluster 2: Si values are average. While most members are well-matched to their own cluster, one member has a low Si value, suggesting it might be less suited to its

assigned cluster. Cluster 3: Si values are generally good, but there's a negative Si value. This negative value indicates an outlier member (like Company 3) that might not fit well within Cluster 3. Company 3 can be an outlier.

Now as the last clustering method, we are going to use K-means Clustering with setting the number of clusters to 3.

```
# K-means clustering

set.seed(123)
# when k=3
kc<-kmeans(df_scale,3)
kc
```

```
## K-means clustering with 3 clusters of sizes 7, 9, 5
##
## Cluster means:
##   Market_Cap        Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.9547543 -0.06120687 -0.3576482  1.0818081  1.1033619      0.8566361
## 2 -0.2375550 -0.73633718  0.4233386 -0.4489909 -0.2407172     -0.1025035
## 3 -0.9090570  1.41109654 -0.2613021 -0.7063477 -1.1114156     -1.0147843
##     Leverage  Rev_Growth Net_Profit_Margin
## 1 -0.2797499 -0.01818848         0.7082574
## 2 -0.3557313 -0.13595383        -0.1652117
## 3  1.0319661  0.27018076        -0.6941793
##
## Clustering vector:
##  [1] 2 2 2 1 2 3 1 3 3 2 1 3 1 3 1 2 1 2 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 25.26414 42.25037 31.94053
##  (between_SS / total_SS =  44.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Sum of Squares in Each Cluster:

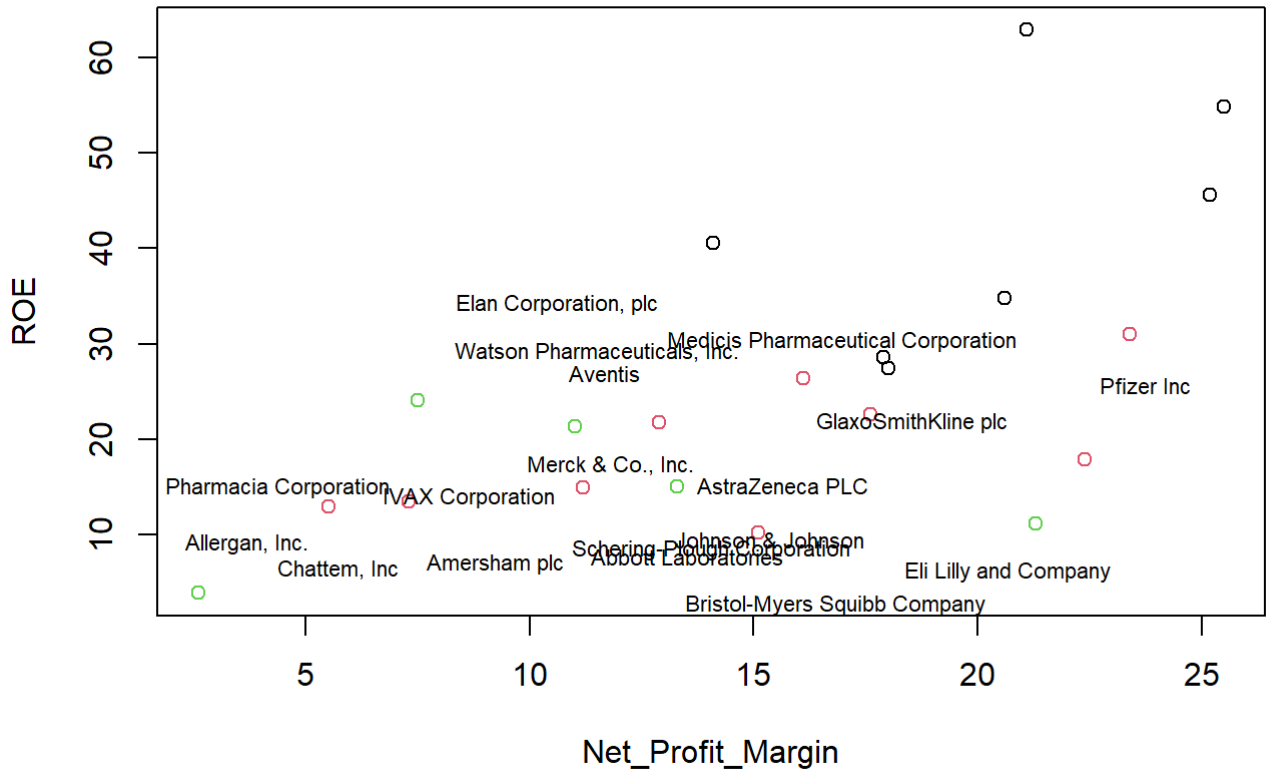
For Cluster 1, the sum of squares is 32.14. For Cluster 2, the sum of squares is 43.30. For Cluster 3, the sum of squares is 20.54. Rationale: Smaller sums of squares within each cluster are desirable. It means the data points within a cluster are closer to each other. So, lower values indicate more cohesive and tightly packed clusters. Sum of Squares Among All Clusters:

The sum of squares among all clusters is 46.7%. Rationale: Here, we are looking for a higher value. A higher percentage suggests that the clusters are distinct and separate from each other. We want the clusters to be as different as possible. Decision on Number of Clusters (k=3):

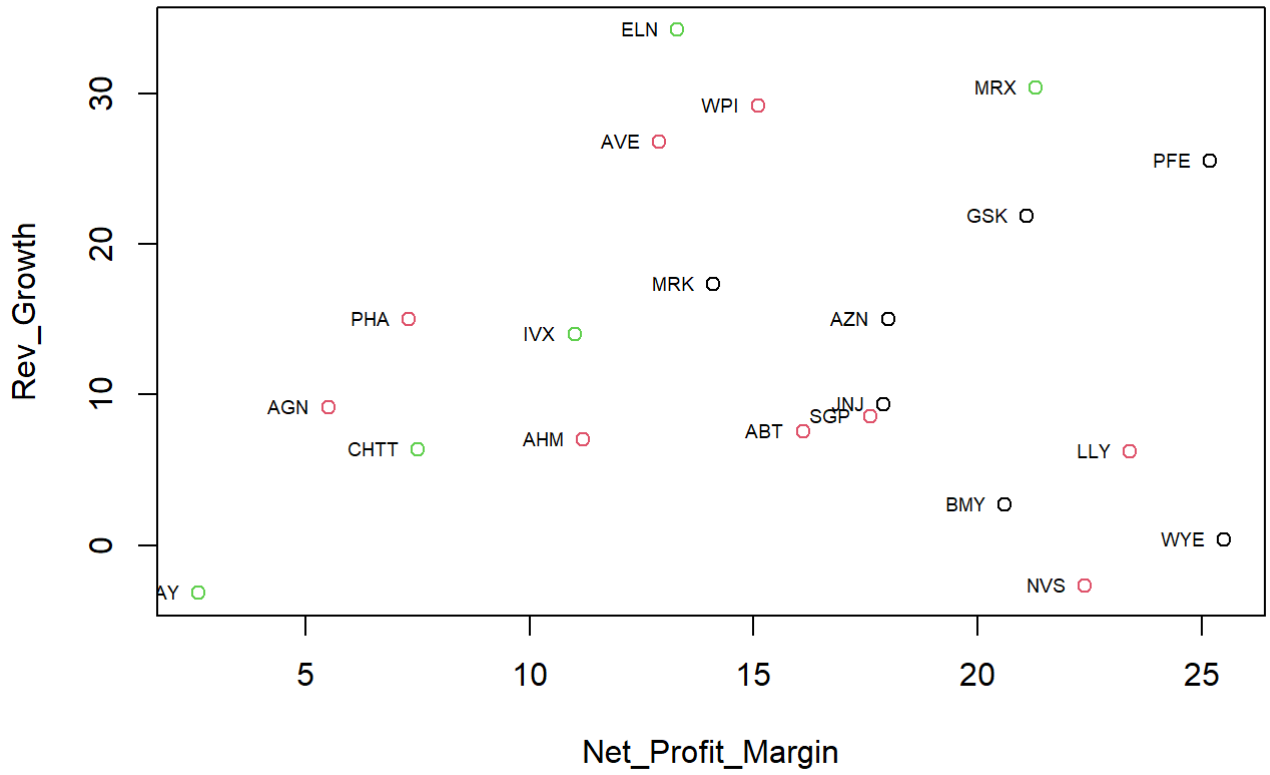
Considering the dataset has 21 observations and analyzing the plots that show the improvement in sum of squares based on the number of clusters and the dendrograms: Decision: Keeping the number of clusters as 3 seems appropriate for further analysis. Rationale: The chosen number of clusters balances the cohesion within each cluster (lower sum of squares) and the separation among clusters (higher percentage of sum of squares among all clusters).

the goal is to find a balance where clusters are tight internally (low within-cluster sum of squares) and distinct from each other (high among-cluster sum of squares). The choice of 3 clusters strikes a reasonable balance for this dataset based on the analysis of the sum of squares.

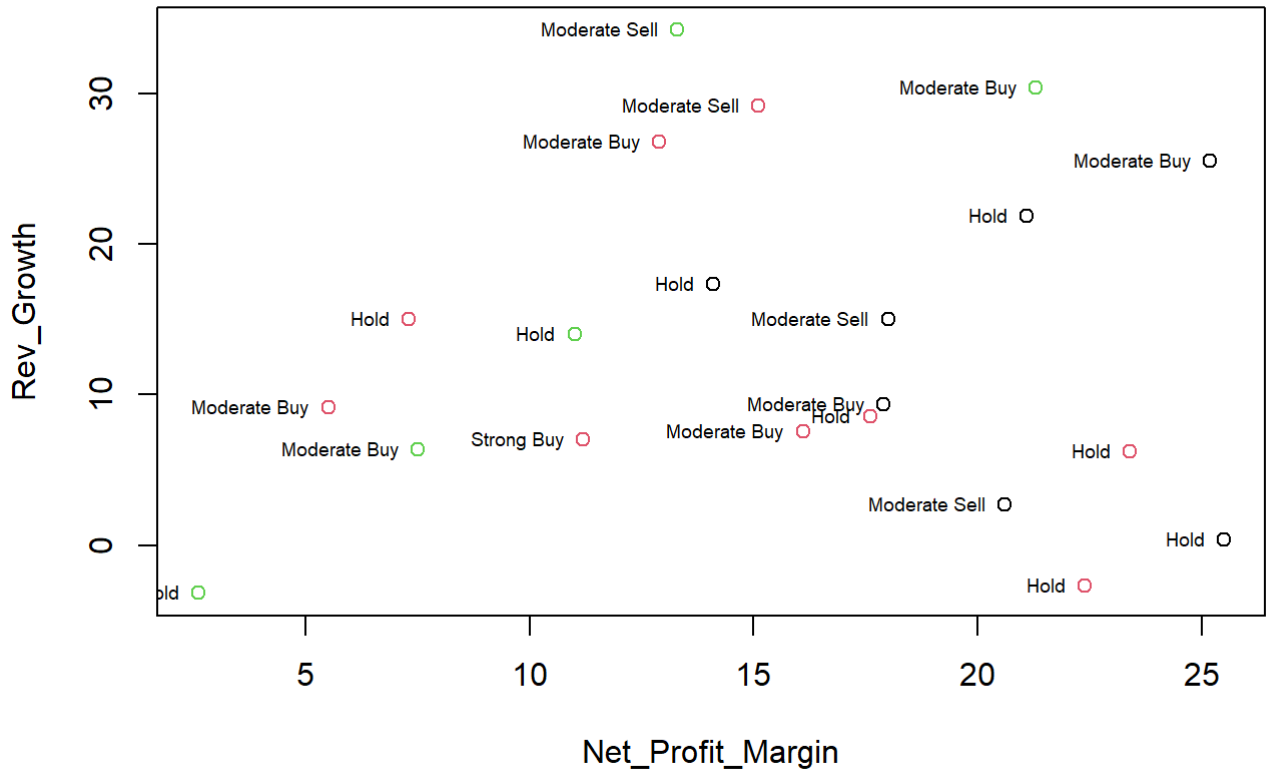
```
plot(ROE~Net_Profit_Margin, df, col= kc$cluster)
with(df,text(df$Rev_Growth ~ df$Net_Profit_Margin, labels=df$Name,pos=2, cex=0.7))
```



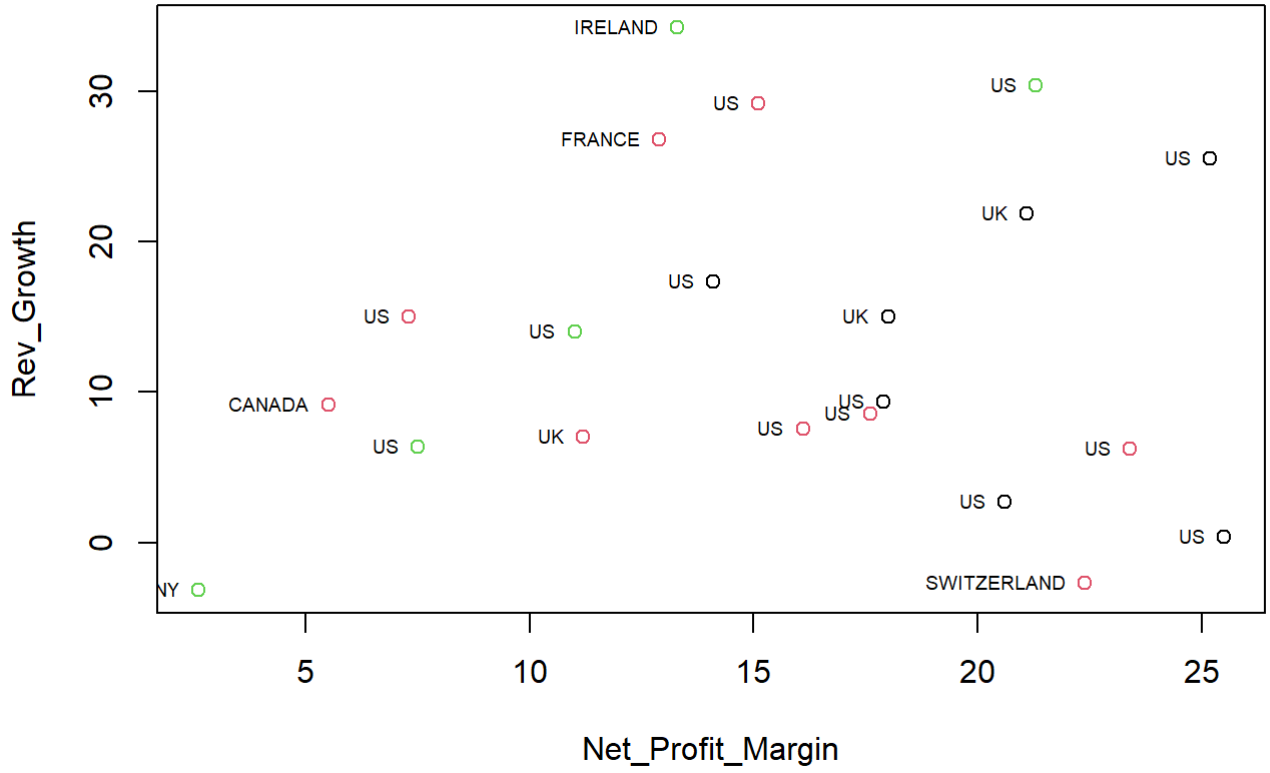

```
plot(Rev_Growth~Net_Profit_Margin, df, col= kc$cluster)
with(df,text(df$Rev_Growth ~ df$Net_Profit_Margin, labels=df$Symbol,pos=2, cex=0.6))
```



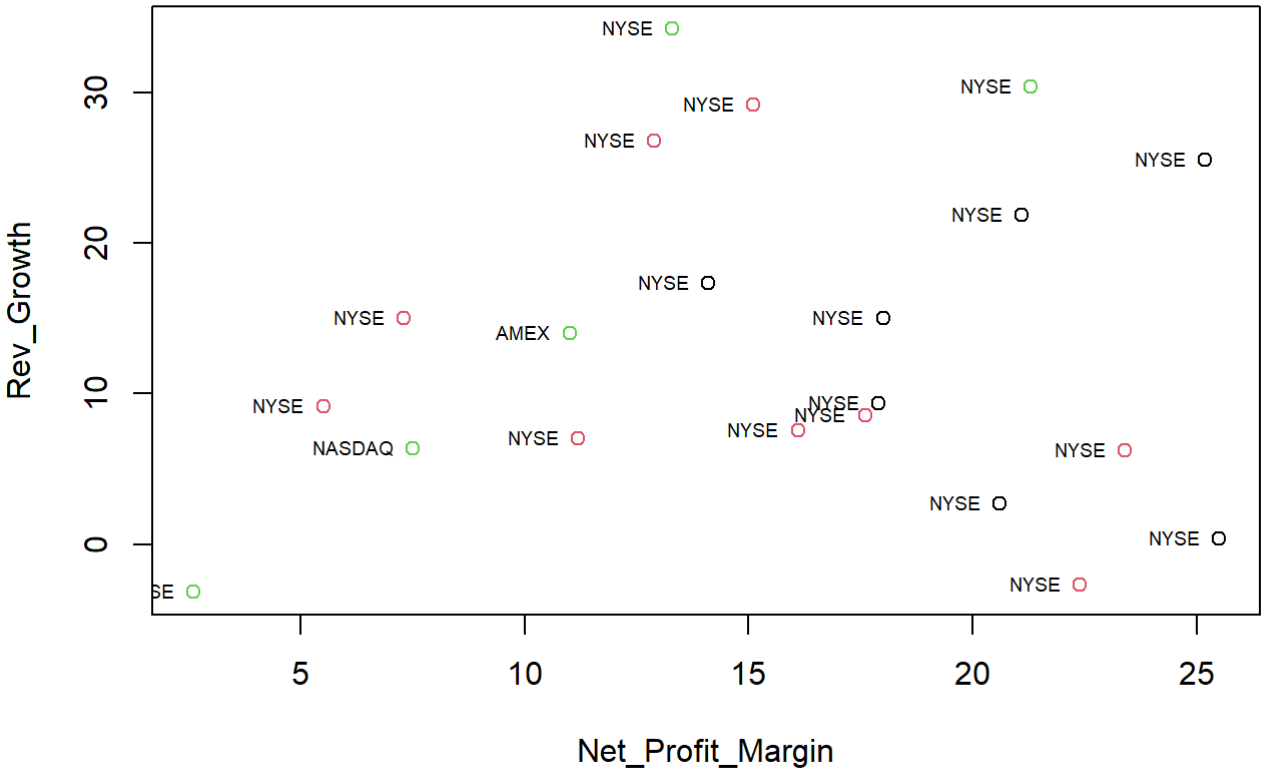

```
plot(Rev_Growth~Net_Profit_Margin, df, col= kc$cluster)
with(df,text(df$Rev_Growth ~ df$Net_Profit_Margin, labels=df$Median_Recommendation,pos=2, cex
=0.6))
```




```
plot(Rev_Growth~Net_Profit_Margin, df, col= kc$cluster)
with(df,text(df$Rev_Growth ~ df$Net_Profit_Margin, labels=df$Location,pos=2, cex=0.6))
```



```
plot(Rev_Growth~Net_Profit_Margin, df, col= kc$cluster)
with(df,text(df$Rev_Growth ~ df$Net_Profit_Margin, labels=df$Exchange,pos=2, cex=0.6))
```




```
# Load necessary libraries
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.2
```

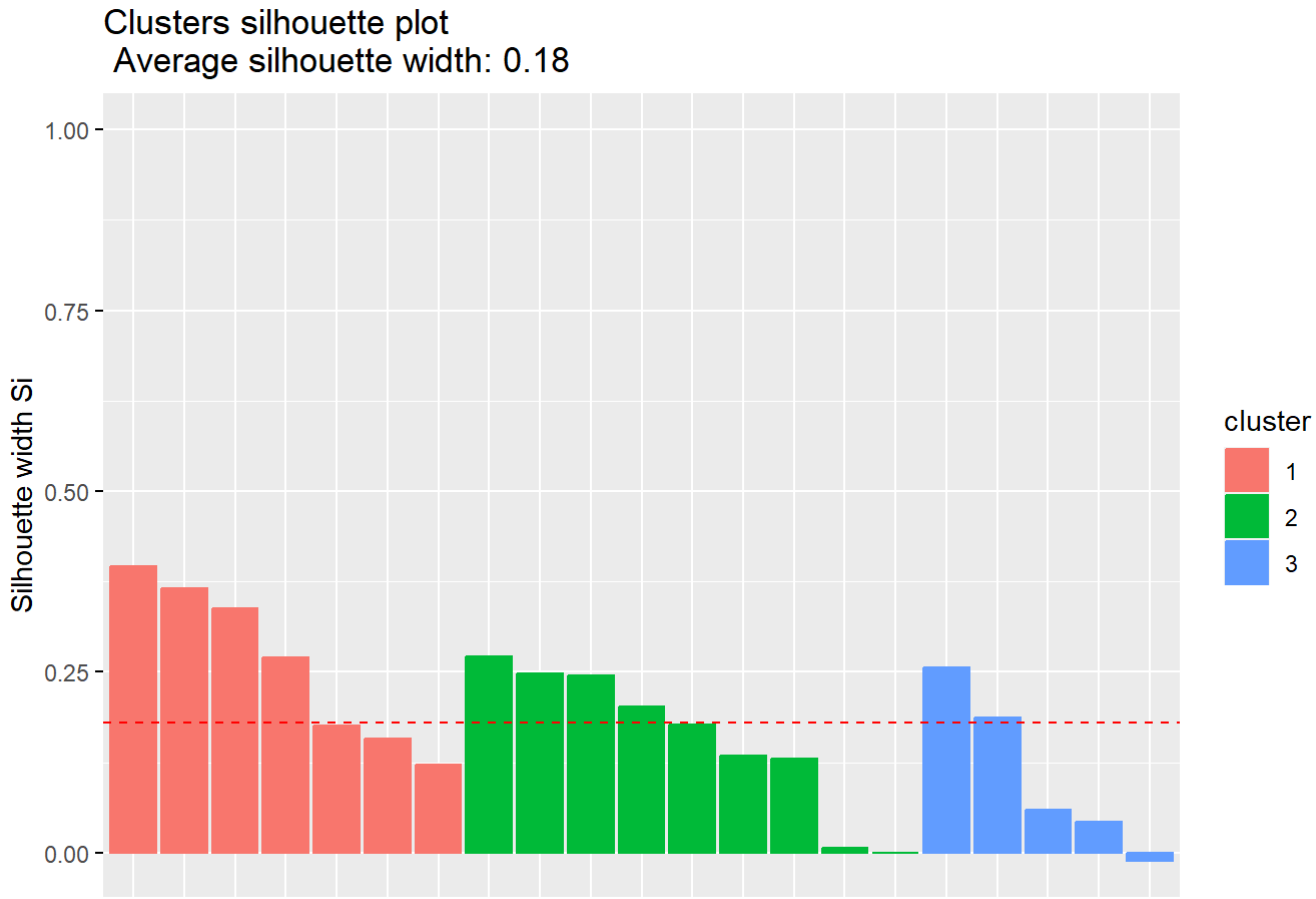
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

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

```
library(cluster)
# Create a silhouette object
silhouette_kmeans <- silhouette(kc$cluster, dist(df_scale))  # Calculate distances using 'dis
t' function

# Plot the silhouette
fviz_silhouette(silhouette_kmeans)
```

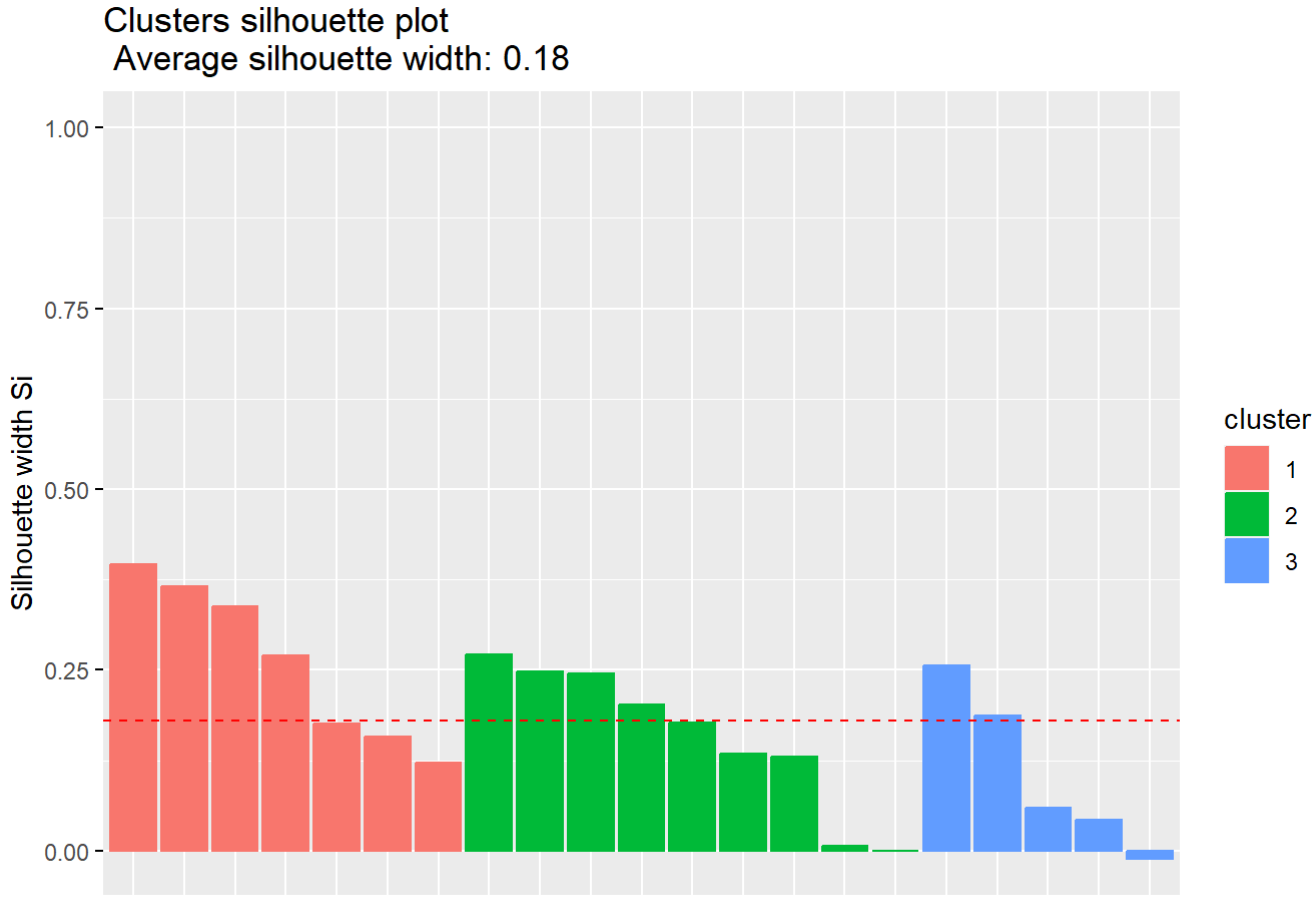
```
##   cluster size ave.sil.width
## 1       1    7          0.26
## 2       2    9          0.16
## 3       3    5          0.11
```



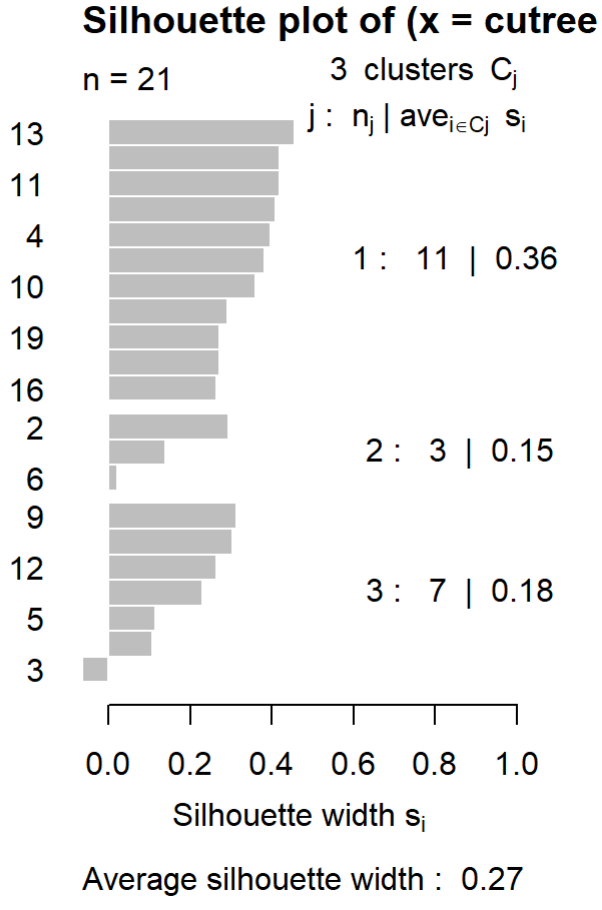
I have compared the silhouette score for both methods and its higher in the hierarchical agglomerative clustering with complete linkage when compare to k means. IN the hierarchical agglomerative clustering with complete linkage its is 0.27 and in k means its 0.18 so the objects in the hierarchical clustering result are better matched to their own clusters than those in the K-means clustering . So clusters that were formed by hierarchical agglomerative clustering using complete linkage is more distinct or well-separated compared to the clusters formed by K-means . So I decide to choose hierarchical agglomerative clustering with complete linkage for my clustering analysis. To better understand the above point look at the silhouette plot.

```
# Plot both silhouettes side by side
par(mfrow=c(1, 2))  # Set the layout to display plots side by side

# Plot the silhouette for K-means
fviz_silhouette(silhouette_kmeans)
```

```
##   cluster size ave.sil.width
## 1       1    7          0.26
## 2       2    9          0.16
## 3       3    5          0.11
```




```
# Silhouette Plot for Hierarchical Clustering
plot(silhouette(cutree(hc.c,3), distance))
```



To find the pattern in the clusters with respect to columns Median_Recommendation, Location, and Exchange we can calculate the median values for each cluster for the feature;

```
# Cluster memberships for each method
hc_c_clusters <- cutree(hc.c, 3)
hc_a_clusters <- cutree(hc_a, 3)
kc_clusters <- kc$cluster
```

```
# Columns to analyze
selected_columns <- c('Median_Recommendation', 'Location', 'Exchange')

# Subset categorical columns along with cluster memberships
cluster_data <- cbind(
  df[selected_columns],
  hc_c = hc_c_clusters,
  hc_a = hc_a_clusters,
  kc = kc_clusters
)
```

```
# Analyzing 'Median_Recommendation', 'Location', and 'Exchange' within clusters
for (method in c('hc_c', 'hc_a', 'kc')) {
  cat("\nCluster analysis for method:", method, "\n")
  for (col in selected_columns) {
    cat("Variable:", col, "\n")
    table_result <- table(cluster_data[[method]], cluster_data[[col]])
    print(table_result)
    # Optionally, perform chi-square tests here
    # chi_result <- chisq.test(table_result)
    # print(chi_result)
  }
}
```

```
##
## Cluster analysis for method: hc_c
## Variable: Median_Recommendation
##
##     Hold Moderate Buy Moderate Sell Strong Buy
##   1    6            3             2          0
##   2    2            1             0          0
##   3    1            3             2          1
## Variable: Location
##
##     CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US
##   1      0      0       0       0           1  2  8
##   2      1      0       1       0           0  0  1
##   3      0      1       0       1           0  1  4
## Variable: Exchange
##
##     AMEX NASDAQ NYSE
##   1    0      0   11
##   2    0      0    3
##   3    1      1    5
##
## Cluster analysis for method: hc_a
## Variable: Median_Recommendation
##
##     Hold Moderate Buy Moderate Sell Strong Buy
##   1    7            4             3          1
##   2    2            1             0          0
##   3    0            2             1          0
## Variable: Location
##
##     CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US
##   1      0      1       0       0           1  3 10
##   2      1      0       1       0           0  0  1
##   3      0      0       0       1           0  0  2
## Variable: Exchange
##
##     AMEX NASDAQ NYSE
##   1    1      0   14
##   2    0      0    3
##   3    0      1    2
##
## Cluster analysis for method: kc
## Variable: Median_Recommendation
##
##     Hold Moderate Buy Moderate Sell Strong Buy
##   1    3            2             2          0
##   2    4            3             1          1
##   3    2            2             1          0
## Variable: Location
##
##     CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US
##   1      0      0       0       0           0  2  5
##   2      1      1       0       0           1  1  5
##   3      0      0       1       1           0  0  3
## Variable: Exchange
```

```
##
##     AMEX NASDAQ NYSE
##   1    0      0    7
##   2    0      0    9
##   3    1      1    3
```

There is some clear patterns and similarities across clusters within each method:

Across Hierarchical Clustering (hc_c):

Cluster 1: Dominated by ‘Hold’ recommendations and ‘NYSE’ exchanges, mostly from the US. Cluster 2: Quite distinct, with less variability in recommendations and locations but still heavily tied to ‘NYSE’. Cluster 3: A mix of various recommendations, more varied locations, but still heavily associated with ‘NYSE’.

Across Hierarchical Clustering (hc_a): Cluster 1: Dominated by ‘Hold’, ‘Order Buy’, and ‘NYSE’, majorly from the US. Cluster 2: Similar to Cluster 1 but less varied in recommendations and locations. Cluster 3: Again, significant ‘Hold’ recommendations and ‘NYSE’, also concentrated in the US.

Across K-means Clustering (kc): Cluster 1: ‘Hold’ is dominant, locations are primarily in the US, and ‘NYSE’ is the preferred exchange. Cluster 2: Similar to Cluster 1 in terms of recommendation and exchange preference, still mostly from the US. Cluster 3: A mix of ‘Hold’ and ‘Moderate Buy’, primarily from the US, and again favoring ‘NYSE’.

In general, the ‘Hold’ recommendation and the presence of ‘NYSE’ are recurrent across clusters for all three clustering methods. Additionally, the dominance of the US in location across clusters is notable.

#How the above information helps equities analyst studying the pharmaceutical industry? This information can be particularly insightful for an equities analyst studying the pharmaceutical industry in several ways:

Investment Strategies: Understanding Market Trends: Recurrent trends like ‘Hold’ recommendations and prevalence on ‘NYSE’ might indicate stability or common investor sentiment across clusters. This insight can guide decisions on long-term investments or cautious strategies.

Regional Focus: Dominance of the US in location might emphasize the concentration of pharmaceutical companies or key players in specific regions. It signals where significant market activity or potential growth might occur.

Risk Assessment: Risk Diversification: While ‘Hold’ recommendations suggest a certain level of stability, an equities analyst can delve deeper into other clusters with varied recommendations to diversify risk. Understanding the risk-reward ratio across different clusters aids in decision-making.

Industry Insights: Comparative Analysis: By observing how different clusters and preferences overlap or diverge across methodologies, analysts can gain a comparative view of how different investment strategies or clusters align within the pharmaceutical industry.

Investment Allocation: Portfolio Distribution: Insights into recurring patterns can help balance investment portfolios across clusters, optimizing for stability, growth, and potential.

Strategy Adaptation: Adapting to Market Sentiment: Recognizing recurrent preferences enables analysts to adapt their strategies based on evolving market sentiments, allowing for proactive investment decisions.

Risk Mitigation: Identifying Outliers: Recognizing outliers within clusters or patterns diverging from the norm helps identify potential risks or unique opportunities. In essence, this information offers crucial guidance for investment strategies, risk assessment, industry insights, and adapting to market sentiments within the pharmaceutical industry. It allows for informed decision-making .

Considering the patterns and dominant characteristics observed across the clusters from the analysis of the pharmaceutical industry data:

1. **Cluster Stability and Reliability**

    - *Key Characteristics*: Dominance of ‘Hold’ recommendations, high presence on ‘NYSE’, and significant US presence in location.
    - *Rationale*: Reflects stability, consistent performance, and a strong market presence.
2. **Global Consistency Cluster**
    - *Key Characteristics*: Consistent ‘Hold’ recommendations, dominance of ‘NYSE’, and a strong US presence alongside international diversity.
    - *Rationale*: Represents stability with a global footprint, balancing US market influence with international opportunities.
3. **Dynamic Investment Diversity**
    - *Key Characteristics*: Varied recommendations (‘Hold’, ‘Moderate Buy’), significant ‘NYSE’ presence, and a strong US market presence across clusters.
    - *Rationale*: Indicates a mix of cautious investment (Hold), moderate risk-taking (Moderate Buy), and a consistent market interest in the US.

These cluster names aim to capture the overarching characteristics observed across the different clustering methodologies and offer insight into the underlying investment trends, regional influences, and risk profiles within the pharmaceutical industry.