Assignment-4_FML

Spandana Sodadasi

2023-11-10

```
knitr::opts_chunk$set(echo = TRUE, comment = NULL)
```

Summary:-

This assignment can be summarized in four steps:

Step - 1: OBJECTIVE

To understand the structure of Pharmaceutical industry from 21 different firms using the given financial measures.

Step - 2: APPROACH

To organize these firms into similar groups basing on the given financial measures.

Step - 3: METHOD

Applying different clustering algorithms and choosing a specific algorithm that is appropriate for the given dataset.

Step - 4: INTERPRETATION

1.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.

Step - 1: The first step that we need to do before getting into the actual clustering is to decide the number of clusters to be formed that means we basically need to decide the 'k' value. We here are using the two most commonly used methods i.e., the elbow method and the silhouette method to decide the optimum value of k.

In the elbow method, we pick the K-value where the elbow actually gets created and then we call it an elbow point. Beyond the elbow point, the increasing value of 'K' does not lead to a significant reduction in WCSS. In the graph, although we cannot find a clear elbow point we can still observe that the values of k after 5 on x-axis are reducing in linear fashion. So, we can consider the value of k as '5'.

To confirm the value of k we are also going to use another method called the silhouette method. It is a measure of similarity of the objects within its own cluster which is the 'Cohesion' when compared to other clusters. So, we here tend to consider the value of k that shows the highest average silhouette width as higher the value the better is the separation between groups and better is the cohesion within the group, which in this case is also 5. Therefore, after using both the methods we can conclude that k=5.

Step - 2: Using the value of 'k' and the numerical variables we will create clusters for 21 firms. This clustering can be done by using various methods such as k means, DBSCAN and Hierarchical clustering. We will run all the three methods and then decide which one would be appropriate enough to apply for the given data.

K-Means Clustering:

K-Means is a clustering technique which is stochastic in nature whose objective is to group similar data points together by identifying k number of centroids, and then allocate every data point to the nearest cluster, while keeping the centroids as small as possible.

We here are applying the k-means technique on the normalized data where k=5.

K means clustering groups the 21 firms into 5 clusters accordingly basing on the similar characteristics of the 9 different variables. Also, k=5 is the optimal value of k as it leads to a meaningful division of clusters.Here, the first cluster represents 8 companies with moderate market capitalization, efficient asset utilization, and solid profitability, where the distance within the cluster is approximately 21.9. Similarly, the second cluster represents 4 companies with very low market capitalization, low returns, and challenges in profitability where the distance within the cluster is approximately 12.8. The third cluster includes 3 companies that have very low market capitalization, high volatility, and the one's struggle for generating returns, where the distance within the cluster is approximately 15.6. The fourth cluster represents 4 companies having high market capitalization, strong returns, and robust profitability where the distance within the cluster is approximately 9.3. The fifth cluster represents 2 companies with low market capitalization but high valuation, yet they face challenges in generating returns but as there are only two firms involved in it the Within - Cluster sum of square distance is very low with a value of 2.8.

DBSCAN Clustering:

DBSCAN is a density-based clustering algorithm that groups together points that are closely packed together, while filtering out noise points that lie in low-density regions. It determines the density of an area based on two parameters i.e., Epsilon which is a measure of radius around a data points and the Minimum Points specify The minimum number of data points required to be within the radius of a given point.

The DBSCAN clustering method needs a certain number of points to be close together to form clusters. The given dataset consists of only 21 records with 9 variables, achieving the required density for meaningful clusters becomes challenging, as the similarity between data points might get diluted. Hence, there are only fewer instances where a sufficient number of points are close together within the specified radius (epsilon) to form dense clusters. As a result, the algorithm may struggle to form distinct clusters. Also, in the given dataset there are no specific outliers or boundaries but DBSCAN by default considers certain boundary points and a noise point. Therefore, we can conclude that DBSCAN is not an appropriate clustering algorithm for the given data.

Hierarchical Clustering:

Hierarchical Clustering establishes a hierarchy of clusters in a dataset, beginning with each data point as an individual cluster. It progressively merges the nearest clusters until a specified criterion is met. The result of hierarchical clustering is a tree-like structure, called a Dendrogram.

Hierarchical Clustering is divided into two types:

(a) Agglomerative Clustering: It is a bottom-up approach

(b) Divisive Clustering: It is a top-down approach

We will further plot the clusters using both Agglomerative Clustering (Agnes) and Divisive Clustering (Diana) and then decide which one is better to apply for the given data.

In the process of Hierarchical Clustering, we initially calculate agglomerative coefficients using various methods and opt for the one that returns the highest value which is 'Ward linkage' in this instance. This coefficient serves as an indicator of the distance between clusters i.e., dissimilarity between the clusters. Upon analyzing

the dendrogram, we can analyze the five clusters, with the first encompassing 7 firms, the second containing 4 firms, and so forth. Notably, the last cluster comprises only one firm, suggesting the potential identification of outliers. However, it's crucial to note that outliers may not be applicable to the provided data, where each firm is considered a data point rather than an outlier. Additionally, choosing k as 5 leads to a meaningful cluster division, as clusters are separated based on similar heights, ensuring minimal distances among them and consequently enhancing overall similarity.

Similarly, Divisive Clustering would also give us outliers, hence applying Divisive Clustering would also not be suitable for the given dataset.

Conclusion:

We have applied three clustering algorithms, and the most suitable one among them for this dataset is the "K-Means" algorithm. It effectively groups firms with similar characteristics into tighter clusters and is notably sensitive to outliers and noisy data, providing meaningful results for a dataset without outliers.

On the other hand, the DBCAN Clustering method, by default, considers boundary points and noise points. Additionally, the small size of the dataset poses a challenge in forming meaningful clusters. Similarly, Hierarchical Clustering also considers outliers, which may not be applicable to the provided data. This clustering algorithm is better suited for dataset where the underlying structure involves natural groupings at different levels of granularity. Therefore, both DBSCAN and Hierarchical Clustering might not be as suitable for the given Pharmaceutical dataset.

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

Interpretation of the clusters for Normalized data:

(i) Cluster -1: This cluster is characterized by 8 firms, where the distance within the cluster is approximately 21.9 with relatively low values in Market Capital, Beta, PE Ratio, ROE, and ROA. The Asset Turnover is moderately positive, indicating a reasonable efficiency in asset utilization. The Leverage and Rev Growth is significantly negative, and Net Profit Margin is positive, implying stable profitability.

(ii) Cluster -2: The 4 firms in this cluster are having distance within the cluster of approximately 12.8. They exhibit low values in Market Capital, slightly positive Beta, and negative values in PE Ratio, ROE, and ROA. The Asset Turnover is notably negative, indicating inefficiency in asset utilization. The Rev Growth is very high, and Net Profit Margin is close to zero, indicating challenges in profitability.

(iii) Cluster -3: This cluster represents 3 companies where the distance within the cluster is approximately 15.6 with extremely low Market Capital, high positive Beta, and moderately negative values in PE Ratio, ROE, and ROA. The Asset Turnover is negative, indicating potential inefficiency in asset utilization. The Leverage is notably positive, suggesting higher financial leverage. The Rev Growth is moderately negative and Net Profit Margin is strongly negative, indicating challenges in profitability.

(iv) Cluster -4: This cluster represents 4 Companies with a distance within the cluster of approximately 9.3 having high values in Market Capital, slightly negative Beta, and moderately negative values in PE Ratio, ROE, and ROA and the Rev Growth is positive, and Net Profit Margin is moderately positive, indicating stable profitability. Also the Asset Turnover is positive which indicates efficient asset utilization.

(v) Cluster -5: This cluster includes 2 companies which comparatively involves the Within - Cluster sums of square distance of 2.8 with negative values in Market Capital, Beta, and ROA. The PE Ratio and ROE are very high, suggesting high valuation and strong returns. The Rev Growth is moderately positive, and Net Profit Margin is strongly negative, indicating challenges in profitability.

Conclusion:

Therefore, after looking at all the clusters we can say that the ideal cluster would be cluster no-4 as it shows a decent Within - Cluster sums of square distance of 9.3 for 4 different firms and it also shows high market capitalization, strong profitability and relatively lower risk.

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

Interpretation of the clusters with respect to categorical variables:

(i) Cluster -1 is primarily dominated by companies based in the United States and then Uk and Switzerland that are listed on the New York Stock Exchange (NYSE). Analysts recommend to hold their stocks as it indicates stability and relatively low-risk investment prospects.

(ii) Cluster -2 includes companies listed on the NYSE from various locations such as US, Ireland and France and they have a recommendation of moderate buy or sell , indicating potential growth opportunities for these firms.

(iii) Cluster -3 comprises a combination of American and Germany companies listed on NYSE, AMEX AND NASDAQ stock exchange market. Analysts recommend a hold or moderate buy, indicating a balanced outlook for these companies.

(iv) Cluster -4 consists of companies from the UK and USA, with a mixed recommendation of partially hold and buy for their stocks listed on the NYSE. This suggests a potential for growth accompanied by some level of risk.

(v) Cluster -5 comprises a blend of American and Canadian companies listed on the NYSE. They carry a moderate buy or hold recommendation, indicate a chance of both growth and also some level of risk.

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

(i) Cluster -1 - “Stable Firms”: Companies with balanced financial metrics operating efficiently within its industry.

(ii) Cluster -2 - “Growth Oriented Firms”: Companies with Low asset turnover and high revenue growth suggest growth potential but sub-optimal efficiency.

(iii) Cluster -3 - “High Risk Firms”: Companies with High leverage, low net profit margin, and ROA indicate a company relying on debt with inadequate profitability and returns. This raises investor concerns about meeting debt obligations and potential financial distress.

(iv) Cluster -4 - “Profitable Firms”: These are typically the large and well-established companies that have a significant market presence and a strong financial position. High market capitalization means that the company has a large number of outstanding shares and a high stock price, resulting in a high valuation and as the net profit margin is moderately positive it indicates stable profitability.

(v) Cluster -5 - “Overvalued - Risky Firms”: High PE ratio and low net profit margin indicate the market values and the company’s stock at a premium compared to its earnings, despite lower profitability. Investors paying a premium for each dollar of earnings may pose a risk, as the company might not meet market expectations, potentially leading to a future decline in stock price.

Problem Statement:-

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. For each firm, the following variables are recorded:

1. Market capitalization (in billions of dollars)
2. Beta
3. Price/earnings ratio
4. Return on equity
5. Return on assets
6. Asset turnover
7. Leverage
8. Estimated revenue growth
9. Net profit margin
10. Median recommendation (across major brokerages)
11. Location of firm's headquarters
12. Stock exchange on which the firm is listed

Use cluster analysis to explore and analyze the given dataset as follows:

1.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.

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

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

Data Importing and Cleaning:

1.Loading the required libraries.

```
library(tidyverse, warn.conflicts = FALSE)
```

```
Warning: package 'tidyverse' was built under R version 4.3.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(factoextra, warn.conflicts = FALSE)
```

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

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

```
library(caret, warn.conflicts = FALSE)
```

```
Loading required package: lattice
```

```
library(e1071, warn.conflicts = FALSE)
library(cluster, warn.conflicts = FALSE)
```

```
Warning: package 'cluster' was built under R version 4.3.2
```

```
library(dplyr, warn.conflicts = FALSE)
library(tinytex, warn.conflicts = FALSE)
library(dbscan, warn.conflicts = FALSE)
```

```
Warning: package 'dbscan' was built under R version 4.3.2
```

```
library(fpc, warn.conflicts = FALSE)
```

```
Warning: package 'fpc' was built under R version 4.3.2
```

2.Importing and reading the dataset.

```
library(readr)
Pharmaceuticals <- read.csv("C:/Users/spand/Downloads/Pharmaceuticals.csv")
dim(Pharmaceuticals)
```

```
[1] 21 14
```

3.Dropping the categorical variables.

```
set.seed(1)
Pharma <- Pharmaceuticals[,-c(2,12,13,14)]
row.names(Pharma) <- Pharma[,1]
Pharma <- Pharma[,-1]
head(Pharma)
```

```
    Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
ABT      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
AGN       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
AHM       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
AZN      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
AVE      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
BAY      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
```

```
    Net_Profit_Margin
ABT              16.1
AGN               5.5
AHM              11.2
AZN              18.0
AVE              12.9
BAY               2.6
```

All the categorical variables have been dropped.

4.Normalizing the data by using the scale function.

```
Pharma_Norm <- scale(Pharma)
head(Pharma_Norm)
```

```
    Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
      Leverage Rev_Growth Net_Profit_Margin
ABT -0.2120979 -0.5277675        0.06168225
AGN  0.0182843 -0.3811391       -1.55366706
AHM -0.4040831 -0.5721181       -0.68503583
AZN -0.7496565  0.1474473        0.35122600
AVE -0.3144900  1.2163867       -0.42597037
BAY -0.7496565 -1.4971443       -1.99560225
```

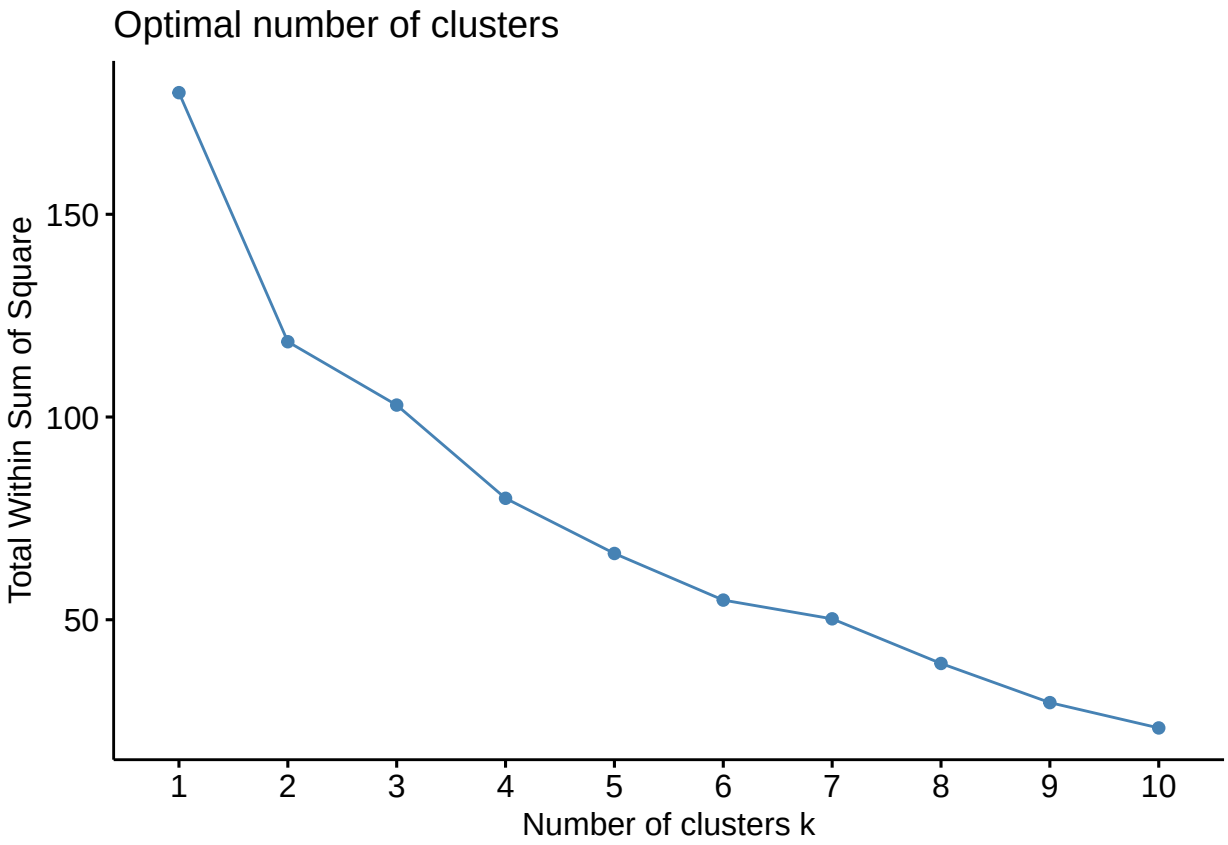
Questions:-

1.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.

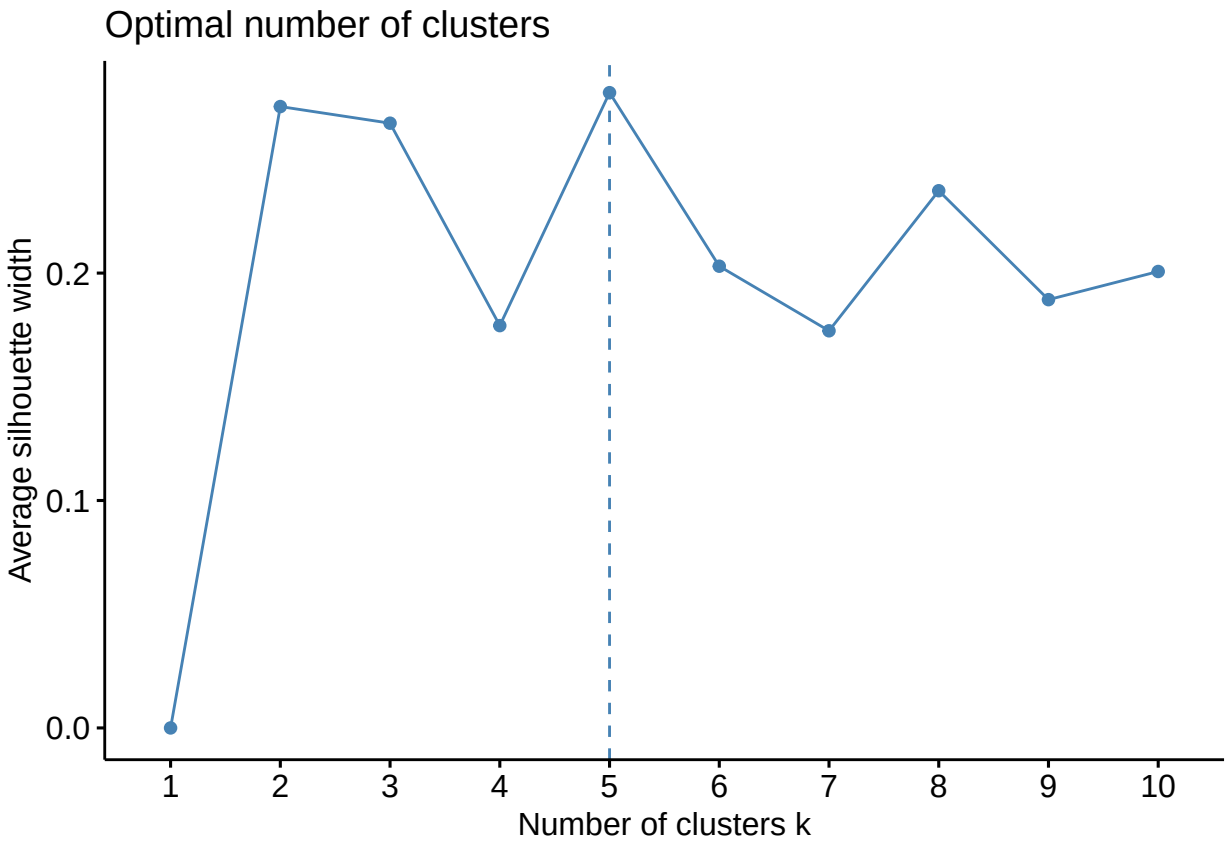
Step - 1: The first step that we need to do before getting into the actual clustering is to decide the number of clusters to be formed that means we basically need to decide the 'k' value. We here are using the two most commonly used methods i.e., the elbow method and the silhouette method to decide the optimum value of k.

Finding the value of "K" by using the elbow and silhouette method.

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




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



In the elbow method, we pick the K-value where the elbow actually gets created and then we call it an elbow point. Beyond the elbow point, the increasing value of 'K' does not lead to a significant reduction in WCSS. In the graph, although we cannot find a clear elbow point we can still observe that the values of k after 5 on x-axis are reducing in linear fashion. So, we can consider the value of k as '5'.

To confirm the value of k we are also going to use another method called the silhouette method. It is a measure of similarity of the objects within its own cluster which is the 'Cohesion' when compared to other clusters. So, we here tend to consider the value of k that shows the highest average silhouette width as higher the value the better is the separation between groups and better is the cohesion within the group, which in this case is also 5. Therefore, after using both the methods we can conclude that k=5.

(Note: This value can be changed if does not provide a better insight.)

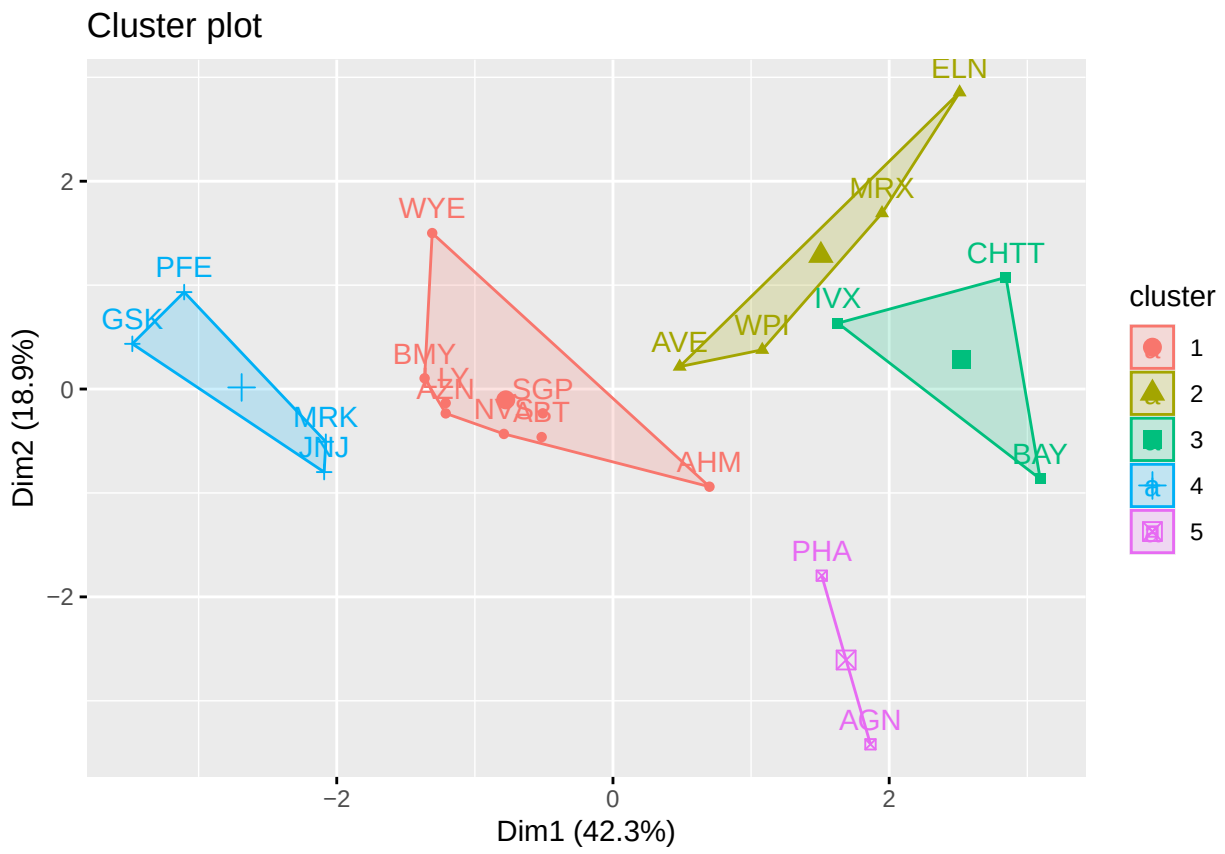
Step - 2: Using the value of 'k' and the numerical variables we will create clusters for 21 firms. This clustering can be done by using various methods such as k means, DBSCAN and Hierarchical clustering. We will run all the three methods and then decide which one would be appropriate enough to apply for the given data.

K-Means Clustering:

K-Means is a clustering technique which is stochastic in nature whose objective is to group similar data points together by identifying k number of centroids, and then allocate every data point to the nearest cluster, while keeping the centroids as small as possible.

Clustering the data using the K-Means Algorithm.

```
set.seed(2)
Pharma_Kmeans <- kmeans(Pharma_Norm, centers = 5, nstart = 25)
fviz_cluster(Pharma_Kmeans, data = Pharma_Norm)
```




```
Pharma_Kmeans$centers
```

```
   Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
3 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
5 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
     Leverage Rev_Growth Net_Profit_Margin
1 -0.27449312 -0.7041516       0.556954446
2  0.06308085  1.5180158      -0.006893899
3  1.36644699 -0.6912914      -1.320000179
4 -0.46807818  0.4671788       0.591242521
5 -0.14170336 -0.1168459      -1.416514761
```

```
Pharma_Kmeans$size
```

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

```
Pharma_Kmeans$withinss
```

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

We here are applying the k-means technique on the normalized data where k=5.

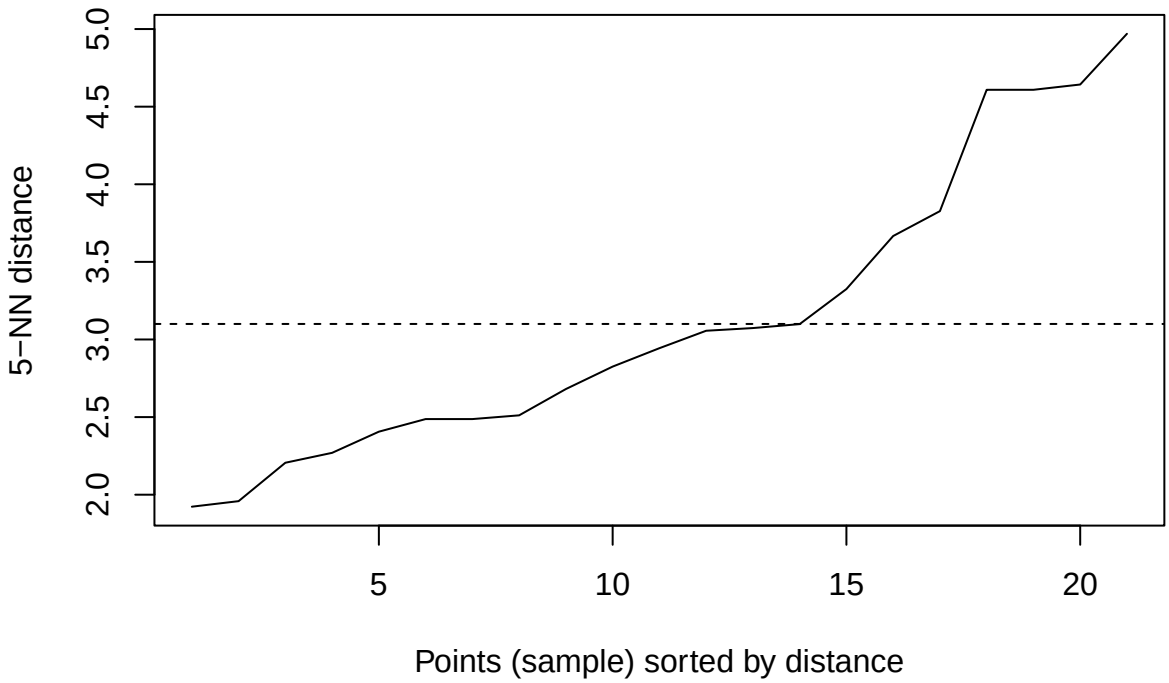
K means clustering groups the 21 firms into 5 clusters accordingly basing on the similar characteristics of the 9 different variables. Also, k=5 is the optimal value of k as it leads to a meaningful division of clusters.Here, the first cluster represents 8 companies with moderate market capitalization, efficient asset utilization, and solid profitability, where the distance within the cluster is approximately 21.9. Similarly, the second cluster represents 4 companies with very low market capitalization, low returns, and challenges in profitability where the distance within the cluster is approximately 12.8. The third cluster includes 3 companies that have very low market capitalization, high volatility, and the one's struggle for generating returns, where the distance within the cluster is approximately 15.6. The fourth cluster represents 4 companies having high market capitalization, strong returns, and robust profitability where the distance within the cluster is approximately 9.3. The fifth cluster represents 2 companies with low market capitalization but high valuation, yet they face challenges in generating returns but as there are only two firms involved in it the Within - Cluster sum of square distance is very low with a value of 2.8.

DBSCAN Clustering:

DBSCAN is a density-based clustering algorithm that groups together points that are closely packed together, while filtering out noise points that lie in low-density regions. It determines the density of an area based on two parameters i.e., Epsilon which is a measure of radius around a data points and the Minimum Points specify The minimum number of data points required to be within the radius of a given point.

Finding the epsilon value.

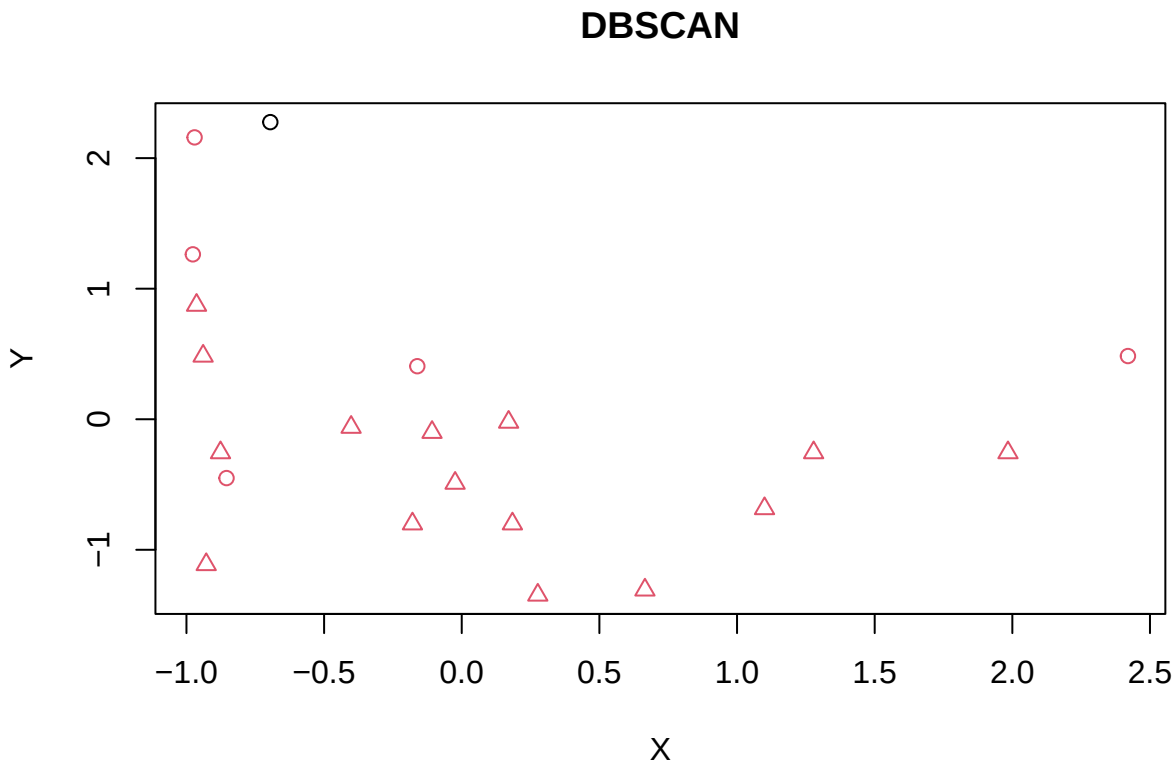
```
dbscan::kNNdistplot(Pharma_Norm, k=5)
abline(h=3.1, lty=2)
```



According to the plot the optimal epsilon value would be '3.1'.

Clustering the data using the DBSCAN Algorithm.

```
Pharma_DBscan <- fpc::dbscan(Pharma_Norm, eps = 3.1, MinPts = 5)
plot(Pharma_DBscan, Pharma_Norm, main="DBSCAN", frame= TRUE, xlab = "X", ylab = "Y")
```



The DBSCAN clustering method needs a certain number of points to be close together to form clusters. The given dataset consists of only 21 records with 9 variables, achieving the required density for meaningful clusters becomes challenging, as the similarity between data points might get diluted. Hence, there are only fewer instances where a sufficient number of points are close together within the specified radius (epsilon) to form dense clusters. As a result, the algorithm may struggle to form distinct clusters. Also, in the given dataset there are no specific outliers or boundaries but DBSCAN by default considers certain boundary points and a noise point. Therefore, we can conclude that DBSCAN is not an appropriate clustering algorithm for the given data.

Hierarchical Clustering

Hierarchical Clustering establishes a hierarchy of clusters in a dataset, beginning with each data point as an individual cluster. It progressively merges the nearest clusters until a specified criterion is met. The result of hierarchical clustering is a tree-like structure, called a Dendrogram.

Hierarchical clustering is divided into two types:

(a) Agglomerative Clustering: It is a bottom-up approach

(b) Divisive Clustering: It is a top-down approach

We will further plot the clusters using both Agglomerative clustering (Agnes) and Divisive Clustering (Diana) and then decide which one is better to apply for the given data.

Plotting a Dendrogram using AGNES()

```
hc_single <- agnes(Pharma_Norm, method = "single")
hc_complete <- agnes(Pharma_Norm, method = "complete")
hc_ward <- agnes(Pharma_Norm, method = "ward")
hc_average <- agnes(Pharma_Norm, method = "average")

print(hc_single$ac)
```

```
[1] 0.4600348
```

```
print(hc_complete$ac)
```

```
[1] 0.6990833
```

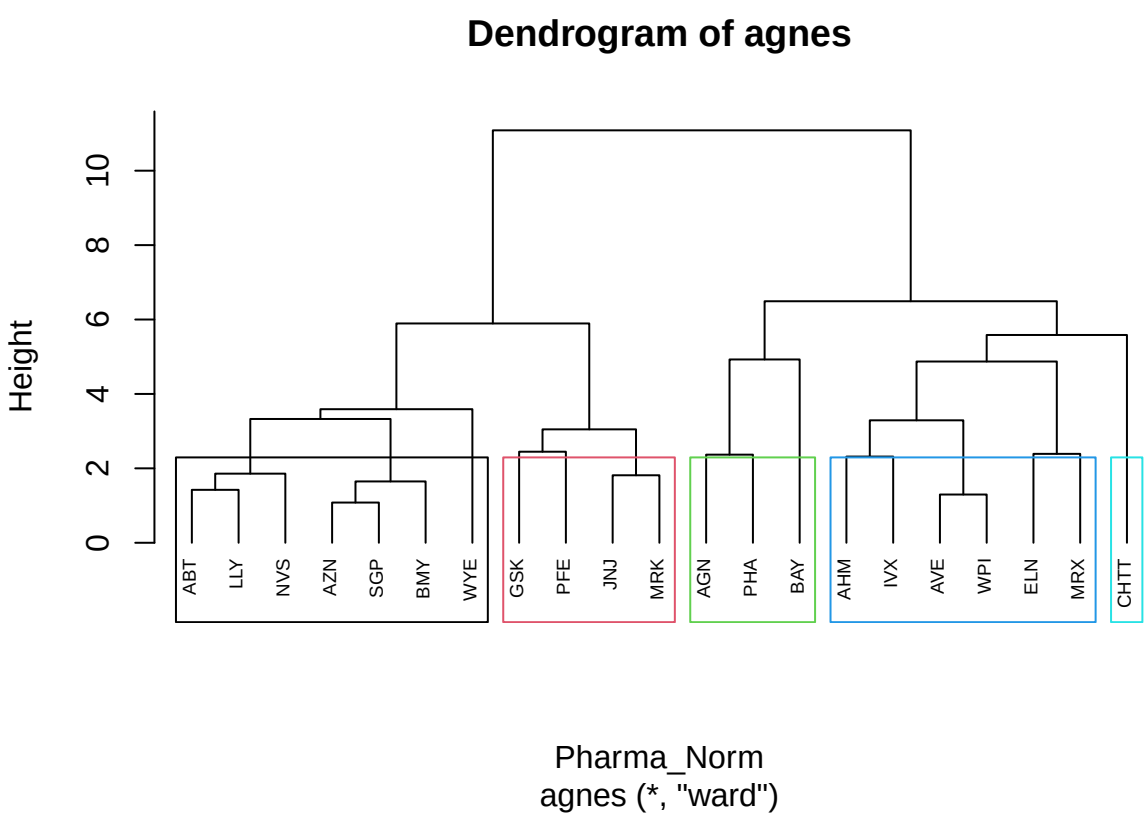
```
print(hc_ward$ac)
```

```
[1] 0.7943164
```

```
print(hc_average$ac)
```

```
[1] 0.5600652
```

```
pltree(hc_ward, cex = 0.6, hang = -1, main = "Dendrogram of agnes")
rect.hclust(hc_ward, k=5, border = 1:5)
```



In the process of Hierarchical Clustering, we initially calculate agglomerative coefficients using various methods and opt for the one that returns the highest value which is 'Ward linkage' in this instance. This coefficient serves as an indicator of the distance between clusters i.e., disimilarity between the clusters. Upon analyzing the dendrogram, we can analyze the five clusters, with the first encompassing 7 firms, the second containing 4 firms, and so forth. Notably, the last cluster comprises only one firm, suggesting the potential identification of outliers. However, it's crucial to note that outliers may not be applicable to the provided data, where each firm is considered a data point rather than an outlier. Additionally, choosing k as 5 leads to a meaningful cluster division, as clusters are separated based on similar heights, ensuring minimal distances among them and consequently enhancing overall similarity.

Similarly, Divisive Clustering would also give us outliers, hence applying Divisive Clustering would also not be suitable for the given dataset.

Conclusion:

We have applied three clustering algorithms, and the most suitable one among them for this dataset is the "K-Means" algorithm. It effectively groups firms with similar characteristics into tighter clusters and is notably sensitive to outliers and noisy data, providing meaningful results for a dataset without outliers.

On the other hand, the DBSCAN Clustering method, by default, considers boundary points and noise points. Additionally, the small size of the dataset poses a challenge in forming meaningful clusters. Similarly, Hierarchical Clustering also considers outliers, which may not be applicable to the provided data. This clustering algorithm is better suited for dataset where the underlying structure involves natural groupings at different levels of granularity. Therefore, both DBSCAN and Hierarchical Clustering might not be as suitable for the given Pharmaceutical dataset.

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

Calculating the centroid and size values for normalized data.

```
set.seed(2)
Pharma_Kmeans <- kmeans(Pharma_Norm, centers = 5, nstart = 25)
Pharma_Kmeans$centers
```

```
   Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
3 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
5 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
     Leverage Rev_Growth Net_Profit_Margin
1 -0.27449312 -0.7041516       0.556954446
2  0.06308085  1.5180158      -0.006893899
3  1.36644699 -0.6912914      -1.320000179
4 -0.46807818  0.4671788       0.591242521
5 -0.14170336 -0.1168459      -1.416514761
```

```
Pharma_Kmeans$size
```

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

```
Pharma_Kmeans$withinss
```

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

Interpretation of the clusters for Normalized data:

(i) Cluster -1: This cluster is characterized by 8 firms, where the distance within the cluster is approximately 21.9 with relatively low values in Market Capital, Beta, PE Ratio, ROE, and ROA. The Asset Turnover is moderately positive, indicating a reasonable efficiency in asset utilization. The Leverage and Rev Growth is significantly negative, and Net Profit Margin is positive, implying stable profitability.

(ii) Cluster -2: The 4 firms in this cluster are having distance within the cluster of approximately 12.8. They exhibit low values in Market Capital, slightly positive Beta, and negative values in PE Ratio, ROE, and ROA. The Asset Turnover is notably negative, indicating inefficiency in asset utilization. The Rev Growth is very high, and Net Profit Margin is close to zero, indicating challenges in profitability.

(iii) Cluster -3: This cluster represents 3 companies where the distance within the cluster is approximately 15.6 with extremely low Market Capital, high positive Beta, and moderately negative values in PE Ratio, ROE, and ROA. The Asset Turnover is negative, indicating potential inefficiency in asset utilization. The Leverage is notably positive, suggesting higher financial leverage. The Rev Growth is moderately negative and Net Profit Margin is strongly negative, indicating challenges in profitability.

(iv) Cluster -4: This cluster represents 4 Companies with a distance within the cluster of approximately 9.3 having high values in Market Capital, slightly negative Beta, and moderately negative values in PE Ratio, ROE, and ROA and the Rev Growth is positive, and Net Profit Margin is moderately positive, indicating stable profitability. Also the Asset Turnover is positive which indicates efficient asset utilization.

(v) Cluster -5: This cluster includes 2 companies which comparatively involves the Within - Cluster sums of square distance of 2.8 with negative values in Market Capital, Beta, and ROA. The PE Ratio and ROE are very high, suggesting high valuation and strong returns. The Rev Growth is moderately positive, and Net Profit Margin is strongly negative, indicating challenges in profitability.

Conclusion:

Therefore, after looking at all the clusters we can say that the ideal cluster would be cluster no-4 as it shows a decent Within - Cluster sums of square distance of 9.3 for 4 different firms and it also shows high market capitalization, strong profitability and relatively lower risk.

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

Dropping the first two categorical variables.

```
set.seed(3)
Pharma_Pattern <- Pharmaceuticals[,-c(1,2)]
head(Pharma_Pattern)
```

```
  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 Median_Recommendation Location Exchange
1              16.1          Moderate Buy       US     NYSE
2               5.5          Moderate Buy   CANADA     NYSE
3              11.2            Strong Buy       UK     NYSE
4              18.0         Moderate Sell       UK     NYSE
5              12.9          Moderate Buy   FRANCE     NYSE
6               2.6                  Hold  GERMANY     NYSE
```

Finding the pattern in the clusters with respect to the numerical variables for those not used in forming the clusters.

```
Cluster_Pattern <-  Pharma_Pattern %>% select(c(10,11,12)) %>% mutate(Cluster = Pharma_Kmeans$cluster)
print(Cluster_Pattern)
```

```
   Median_Recommendation    Location Exchange Cluster
1           Moderate Buy          US     NYSE       1
2           Moderate Buy      CANADA     NYSE       5
3             Strong Buy          UK     NYSE       1
4          Moderate Sell          UK     NYSE       1
5           Moderate Buy      FRANCE     NYSE       2
6                   Hold     GERMANY     NYSE       3
7          Moderate Sell          US     NYSE       1
8           Moderate Buy          US   NASDAQ       3
9          Moderate Sell     IRELAND     NYSE       2
10                  Hold          US     NYSE       1
11                  Hold          UK     NYSE       4
12                  Hold          US     AMEX       3
13          Moderate Buy          US     NYSE       4
14          Moderate Buy          US     NYSE       2
15                  Hold          US     NYSE       4
16                  Hold SWITZERLAND     NYSE       1
17          Moderate Buy          US     NYSE       4
```

```
18                  Hold          US     NYSE       5
19                  Hold          US     NYSE       1
20         Moderate Sell          US     NYSE       2
21                  Hold          US     NYSE       1
```

Visualizing the distribution of firms grouped by clusters by using the bar charts.

```
# Bar chart for Median Recommendation
Median_Recommendation <- ggplot(Cluster_Pattern, aes(x = factor(Cluster),
                         fill = Median_Recommendation)) + geom_bar() +
                         labs(x = 'Clusters', y = 'Frequency',
                         title = 'Median Recommendation Across the Clusters') +
                         theme_minimal()
Median_Recommendation
```


```
# Bar Chart for Location
Location <- ggplot(Cluster_Pattern, aes(x = factor(Cluster), fill = Location)) +
            geom_bar() + labs(x = 'Clusters', y = 'Frequency',
            title = 'Location Across the Clusters') +
            theme_minimal()
Location
```


```
# Bar Chart for Exchange
Exchange <- ggplot(Cluster_Pattern, aes(x = factor(Cluster), fill = Exchange)) +
            geom_bar() + labs(x = 'Clusters', y = 'Frequency',
            title = 'Exchange Across the Clusters') +
            theme_minimal()
Exchange
```

Interpretation of the clusters with respect to categorical variables:

(i) Cluster -1 is primarily dominated by companies based in the United States and then Uk and Switzerland that are listed on the New York Stock Exchange (NYSE). Analysts recommend to hold their stocks as it indicates stability and relatively low-risk investment prospects.

(ii) Cluster -2 includes companies listed on the NYSE from various locations such as US, Ireland and France and they have a recommendation of moderate buy or sell , indicating potential growth opportunities for these firms.

(iii) Cluster -3 comprises a combination of American and Germany companies listed on NYSE, AMEX AND NASDAQ stock exchange market. Analysts recommend a hold or moderate buy, indicating a balanced outlook for these companies.

(iv) Cluster -4 consists of companies from the UK and USA, with a mixed recommendation of partially hold and buy for their stocks listed on the NYSE. This suggests a potential for growth accompanied by some level of risk.

(v) Cluster -5 comprises a blend of American and Canadian companies listed on the NYSE. They carry a moderate buy or hold recommendation, indicate a chance of both growth and also some level of risk.

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

(i) Cluster -1 - “Stable Firms”: Companies with balanced financial metrics operating efficiently within its industry.

(ii) Cluster -2 - “Growth Oriented Firms”: Companies with Low asset turnover and high revenue growth suggest growth potential but sub-optimal efficiency.

(iii) Cluster -3 - "High Risk Firms": Companies with High leverage, low net profit margin, and ROA indicate a company relying on debt with inadequate profitability and returns. This raises investor concerns about meeting debt obligations and potential financial distress.

(iv) Cluster -4 - "Profitable Firms": These are typically the large and well-established companies that have a significant market presence and a strong financial position. High market capitalization means that the company has a large number of outstanding shares and a high stock price, resulting in a high valuation and as the net profit margin is moderately positive it indicates stable profitability.

(v) Cluster -5 - "Overvalued - Risky Firms": High PE ratio and low net profit margin indicate the market values and the company's stock at a premium compared to its earnings, despite lower profitability. Investors paying a premium for each dollar of earnings may pose a risk, as the company might not meet market expectations, potentially leading to a future decline in stock price.
