CLUSTERING

SANIA FATIMA

2023-11-11

SUMMARY:

OBJECTIVE:Understand the structure of pharmaceutical industry using some financial measures. APPROACH:Organising the financial measures of pharmaceutical industry basing on similar groups.

##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.

It's important to load the libraries for performing the functions and features provided by each package to perform specific tasks such as data manipulation, Clustering etc.

Loading the csv files and reading the dataset.After reading the dataset,we found there are 21 observation with 14 variables.

The reason for excluding the specific columns(2,12,13,14) is that they contain categorical variables that are not suitable for forming the clusters.It's a preprocessing step before applying the clustering techniques.

#SCALING THE DATA As there might be variables which is unit less,the scaling is done to bring all variables to a common scale,which is very important for machine learning algorithms.It transforms the data so that each variable has a mean of 0 and standard deviation of 1.By standardizing the data,the variables will become comparable and the scale of variable will have no longer influences on the results of certain analyzes.It is therefore customary to normalize continuous measurements before computing the Euclidean distance. This converts all measurements to the same scale.

#EUCLIDEAN DISTANCE The distance from a point to itself will be 0, and dij = dji. Given that, we only need to specify distances as an upper or lower matrix. An important point to note is that the measure computed above is highly influenced by the scale of each variable, so that variables with larger scales have a much greater influence over the total distance. The choice of the distance measure plays a major role in cluster analysis. The main guideline is domain dependent.

Although Euclidean distance is the most widely used distance, it has three main features that need to be kept in mind. 1. It is scale dependent. 2. It completely ignores any relationship between measurements, so if the measurements are correlated, maybe other measures of distances might better. 3. It is sensitive to outliers.

#ELBOW METHOD K=5 An elbow method is a line chart that shows how much the data is spread out within each cluster as we add more clusters. It is used to find the best number of clusters for our Pharmaceuticals industry by looking for the point where the line starts to flatten out, which is called the "elbow". This is the point where adding more clusters doesn't significantly improve the way the data is clustered.

Here,also comes the concept of WSS(within sums of squares) which means,the measure of variations within the clusters.The lower the variations within the clusters, the similar the datapoints are to each other.

The chart shows that the elbow point 5 provides the best value for k. While WSS will continue to drop for larger values of k, we have to make the tradeoff between overfitting, i.e., a model fitting both noise and signal, to a model having bias. Here,the elbow point provides that compromise where WSS, while still decreasing beyond k = 5, decreases at a much smaller rate. In other words, k=5 provides the best value between bias and overfitting.

#SILHOUETTE METHOD To guarantee the validity and consistency of record assignments to clusters, we can employ the "Silhouette Method". This approach is straightforward, assessing the similarity of an object to its designated cluster relative to other clusters. The Silhouette Method yields values within the -1 to +1 range.High values indicate a good matching of data to clusters.From the Graph,it is clear that 5 is the ideal number of clusters moreover,we look for large values for the Silhouette Width (Y Axis).

#KMEANS K-Means:The k-means algorithm is a partitional clustering algorithm to group 'n' objects based on attributes into K partitions, where k<n.

The choice of the number of clusters can either be driven by external considerationsor the methods we have opted such as elbow,silhouette or we can try a few values stochastically. After choosing k, the n records are partitioned into these initial clusters. If there is external reasoning for the initial assignment, we can apply that, but otherwise, this is done randomly(nstart=25). In these cases, the algorithm can be return with different randomly generated starting partitions to reduce chances of the heuristic producing a poor solution.

By viewing the figure,we can say that we formed 5 clusters as by elbow and silhouette method ,we found our best K=5.

#EPSILON VALUE According to the plot the optimal epsilon value is 3.1

The method proposed here uses the concept of k-nearest neighbor distances. We calculate the average distance of every point to its k nearest neighbors. The value of k is chosen by the modeler and corresponds to MinPts. We will then plot these distances against the sampled points, and then identify the value at which there is a sharp change in values. This is similar to identifying the elbow when determining k in k-means.This distance can be seen as a measure of the local density of points.

#DBSCAN DBSCAN requires two parameters: epsilon and the minimum number of points required to form a dense region[a] (minPts). It starts with an arbitrary starting point that has not been visited. This point's epsilon-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized epsilon-environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its epsilon-neighborhood is also part of that cluster. Hence, all points that are found within the epsilon-neighborhood are added, as is their own epsilon-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

###In this case of Pharmaceutical industry, the dbscan is inappropriate as it is not forming the clusters and taking the important data points as noise points. DBSCAN cannot cluster data sets well with large differences in densities(3.1),since the minPts and epsilon combination cannot then be chosen appropriately for all clusters. If the data and scale are not well understood, choosing a meaningful distance threshold epsilon can be difficult. The quality of DBSCAN depends on the distance measure . The most common distance metric used is Euclidean distance. Especially for high-dimensional data, this metric can be rendered almost useless due to the so- called "Curse of dimensionality", making it difficult to find an appropriate value for epsilon.

#HIERARCHICAL CLUSTERING Hierarchical methods can be either agglomerative or divisive. Agglomerative methods begin with n clusters and sequentially merge similar clusters until a single cluster is obtained. Divisive methods work in the opposite direction, starting with one cluster that includes all records.

##AGNES

A dendrogram is a treelike diagram that summarizes the process of clustering.Similar records are joined by lines whose vertical length reflects the distance between the records. The figure shows the dendrogram of AGNES for our example and here also we formed the 5 clusters. CLUSTER NO:1:It includes the companies such as ABT,LLY,NVS,AZN,SGP,BMY,WYE and it has a height length greater than 2 CLUSTER NO:2:The companies included in this cluster are GSK,PFE,JNJ,MRK which represents that their heights are similar in nature within the cluster however have larger difference between the cluster. CLUSTER NO:3:The pharmaceutical industry included in this cluster are AGN,PHA,BAY.The cluster is formed again on the basis of similar heights. CLUSTER NO:4:The Pharmaceutical industry included in this cluster are

AHM,IVX,AVE,WPI,ELN,MRX. CLUSTER NO:5: It has only one company that is CHTT which shows 0 variations in itself as it considered it as an outlier.Therefore, this dataset has no outliers as every company is important in Pharmaceutical industry. Heirarchical clustering is very sensitive to outliers. Note that the figure is affected by the choice of distance, e.g., Euclidean versus City, and the type of linkage, centroid, single,etc

With respect to the choice of distance between clusters, single and complete linkage are robust to changes in the distance metric (e.g., Euclidean, statistical distance) as long as the relative ordering is kept. In contrast, average linkage is more influenced by the choice of distance metric, and might lead to completely different clusters when the metric is changed.

when we tested the methods employ such as SINGLE,COMPLETE,WARD,AVERAGE.It can be inferred that ward is the best approach here as we look for the high agglomerative coefficient value where it has 0.79431 as compared to other methods.So,the higher the agglomerative coefficient value,higher the cohesion of clusters are.

##DIANA Diana is also the part of hierarchical clustering and it is basically used with large amount of data.AS the dataset is very small of 21 firms Diana will not be suitable to cluster the groups .Though it will form the clusters it starts with taking all the records into consideration from the first step.

#AGNES VERSUS DIANA Generally speaking,the output of both AGNES and DIANA are comparable. Both have similar output complexity. Agnes is most commonly used technique. It is empirically suggested that AGNES is used for identifying small clusters where as DIANA is used for identifying the large clusters.

As a result,we can say that Hierarchical clustering is inappropriate to select because of outliers and AGNES which is suitable for forming small clusters are considering the vital records are outliers as there are no outliers present in this dataset.On the other hand, because of small dataset and formation of small clusters,Diana is not at all suitable for the Pharmaceuticals industry.Moreover,Hierarchical methods are especially useful when the goal is to arrange the clusters into a natural hierarchy.

#KMEANS VS DBSCAN VS HIERARCHICAL

KMEANS are easy to implement and by elbow and silhouette method we found k value is 5 and its groups the clusters in an effective way without involving any data as outliers..It reiterates till the centroid value doesnot get change. Whereas,DBSCAN is not forming the clusters properly and it is considering the important records as noisy data. HIERARCHICAL is suitable to arrange the clusters into a natural hierarchy.Though, its forming the clusters appropriately by both AGNES and DIANA but its considering the crucial datapoints are outliers.

#Which Clustering technique is Appropriate to choose for pharmaceutical industry?

Kmeans is the best clustering technique to choose for this Pharmaceutical industry .The reason behind this is the dataset doesn't have outliers and it is correctly forming the clusters based on the objective that the similarities/variations within the clusters are less and between the clusters are high.The distance is a fundamental concept for knowing the diference between the clusters. K-means provides easily interpretable results, with each data point assigned to a specific cluster.This algorithm aims to minimize the within sum of squared(WSS) distances between data points and their respective cluster centroids.

#Reasons for choosing cluster Number=4 in KMEAN

In this case,the cluster number=4 is recommended because,as we compared the other clusters,the Net profit margin which is the essential characteristic for every firm to infer; whether the firm is landing towards the gains or losses or it could survive or not in the market .So, cluster number=4 of Pharmaceutical industry gains the net profit of 0.59 which is comparatively high when compared to other clusters.

Though the Revenue growth of the firm is little less when compared to cluster 5,we cannot eliminate the cluster based on determination of one factor of industry for choosing the right grouping.Moreover,The debts of purchasing the assets of this industry is literally low (-0.468) as compared to other clusters.

The Assets_Turnover(sales income to the company) of cluster number=4 is 1.153, which records high and its mandatory for the company to have large sales income to gain high amount of profit.

ROA(Return on Assets) is a type of return on investment metric that measures the profitability of a business in relation to its total asset.The contribution of ROA is 1.350 which is a positive sign for pharmaceutical industry to earn profits.

The Return on equity(ROE) falls for cluster number 4 is 1.2349 ,from which we can infer that the pharmaceutical industry is efficient at generating profits.The fundamentality of PE(profit earning ratio) is that the lower it is, the better.Keeping in mind the other factors to determine which cluster is right to choose,we might have drawbacks of the company as there is no company formed without a single drawback.So,The PE Ratio of the cluster number 4 is -0.198.

The ideal beta(risk) for the companies is 1 where our pharamaceutical industry is having -0.178 that can be consider as the good feature for selecting the cluster.

The Higher Market_capital (1.695) typically represents stability and its a type of corporate assets for the industry which is peak when compared to other clusters.

These are the reasons for choosing the cluster number =4 as the best.The companies in cluster number 4 are GlaxoSmithKline plc(GSK),Pfizer Inc(PFE),Merck & Co.,Inc.(MRK),Johnson & Johnson (JNJ) which represents the variations and distance within the cluster is less and the variations or distances, between the clusters are large in comparison with other clusters. In Pharmaceutical industry, there are no outliers as every datapoints(companies) are important and it forms the clusters approriately.

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

#KMEANS CLUSTERS WITH RESPECT TO NUMERICAL VARIABLES

CLUSTER NUMBER:1 -> the size of cluster number 1 is 8 and the companies included in this cluster are Bristol-Mysers Squibb Company(BMY),AstraZeneca PLC(AZN),Eli Lily and Company(LLY),Schering-plough corporation(SGP),Novartis AG(NVS),Abbott Laboratories(ABT),Amersham plc(AHM),Wyeth(WYE),where the variations within this cluster (WWS) is 21.879 that means there are a little huge amount of dissimilarities present within the clusters with different centers/centroid value of every aspect of Pharmaceutical Industry such as Beta[-0.436],ROA[0.408],Net profit margin[0.556] and so on.As a result, I can say that this cluster is maintaining its current operations.

CLUSTER NUMBER:2 -> the size of cluster number 2 is 4 and the companies included in this cluster are Aventis(AVE),Watson Pharamaceuticals, Inc.(WPI),Medicis Pharmaceuticals Corporation(MRX),Elan Corporation,plc(ELN) ,where the variations within this cluster (WWS) is 12.791 (where the distance or variation within the cluster is low in comparison with cluster 1) with different centers/centroid values of every aspect of Pharmaceutical Industry such as Market capital[-0.760],Beta[0.279],Leverage[0.60],Revenue Growth[1.518] and so on. Overall,I can say that this cluster focuses on improving profitability and efficiency.

CLUSTER NUMBER:3 -> the size of cluster number 3 is 3 and the companies included in this cluster are IVAX Corporation(IVX),Chattern,Inc(CHTT),Bayer AG(BAY) ,where the variations within this cluster (WWS) is 15.595(that means it has quite high dissimilarity within the cluster when compared to cluster 2) with different centers value of every aspect of Pharmaceutical Industry such as Market capital[-0.870],leverage[1.366],Beta[1.340] and so on.As a result,I can say that this particular cluster is addressing financial challenges and reduce leverage.

CLUSTER NUMBER:4 -> the size of cluster number 4 is 4 and the companies included in this cluster are GlaxoSmithKline plc(GSK),Pfizer Inc(PFE),Merck & Co.,Inc.(MRK),Johnson & Johnson (JNJ),where the variations within this cluster (WWS) is 9.284 (it has minimal similarities within the clusters) with different centers value of every aspect of Pharmaceutical Industry such as Market capital[1.695],Beta[-0.178],Net profit margin[0.591] and so on.Overall, I can interpret that the pharmaceutical industry present in this cluster are continuing current strategies and monitoring for sustained growth.

CLUSTER NUMBER:5 -> the size of cluster number 5 is 2 and the companies included in this cluster are Allergan Inc(AGN),Pharmacia Corporation (PHA),where the variations within this cluster (WWS) is

2.803(which has very little dissimilarities within its grouping) with different centers/centroid value of every aspect of Pharmaceutical Industry such as Market capital[-0.439],Beta[-0.470],PE ratio[2.700] and so on.To intrepret on this cluster number 5 , I can say that this cluster evaluate and possibly adjust pricing or cost structure for better profitability.

##2b)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 where the Exchanges are from NYSE,NASDAQ,AMEX and the location of the companies include US,CANADA,UK,FRANCE,GERMANY,IRELAND,SWITZ and the median recommendation includes MODERATE BUY,MODERATE SELL,HOLD,STRONG BUY.The table shows the three categorical variables at the last such as Median_Recommendation,Location,Exchange to see whether there is an impact of these variables or not. we selected the variables which are categorical in nature to identify the pattern in the clusters used in forming the clusters:

###MEDIAN RECOMMENDATION: The MEDIAN RECOMMENDATION graph shows that the analysts of the Pharmaceutical industry suggest that the majority of the products is on HOLD except cluster number 2.This implies a neutral stance. Analysts believe investors may want to neither buy more nor sell their existing holdings. It could suggest a period of stability or uncertainty. There is a MODERATE BUY in almost all clusters(1 to 5) which represents the positive outlook, but with some reservations. It could mean that analysts see potential for growth, but there might be certain risks or uncertainties associated with the investment. The cluster number 1 and 2 have MODERATE SELL which represents the negative outlook for the stock, but with some moderation. Analysts may see reasons for concern, but they might not view the situation as extremely dire. It could indicate a more cautious recommendation to reduce exposure to the stock. CLuster Number 1 is only the one where the analyst suggests the STRONG BUY.Analysts believe the stock has strong potential for appreciation, and investors are encouraged to consider buying.

###LOCATION: After Analyzing the graph of LOCATION ACROSS CLUSTER,the US is the dominant country where it has huge amount of Pharmaceutical industry.In other words,every cluster has the industries from US.Cluster 1 and 4 are the the companies from UK .IRELAND and FRANCE Pharmaceutical companies are present only in cluster 2.The less covered location part across cluster of pharmaceutical industry is CANADA which is present in Cluster 5.Moreover, Cluster 1 has the companies where the location is from SWITZERLAND.

###EXCHANGE The EXCHANGE graph shows that all clusters 1,2,4,5 have the exchange from NYSE.Though cluster 3 have the exchange from NYSE however it is also having the echanges from two other exchange listing securities, they are NASDAQ and AMEX. The advantages of having single exchange listing is simplicity and focus,clear market identity,exclusivity etc

To interpret all type of categorical variables , I can say that these three variables (Median Recommendation,location,exchange) have the patterns while forming the clusters with respect to numerical variables.

#3)Provide an appropriate name for each cluster using any or all of the variables in the dataset. ###NAMING THE CLUSTERS Cluster 1: Stable Growth Companies Characteristics: Moderate Market Cap, Low Beta, Moderate PE Ratio, Positive ROE and ROA, Moderate Asset Turnover, Low Leverage, Moderate Revenue Growth, High Net Profit Margin.

Cluster 2: High Risk, High Leverage Companies Characteristics: Low Market Cap, Moderate Beta, Low PE Ratio, Negative ROE and ROA, Low Asset Turnover, Moderate to High Leverage, High Revenue Growth, Variable Net Profit Margin.

Cluster 3: Financial Conservative Companies Characteristics: Very Low Market Cap, High Beta, High PE Ratio, Negative ROE and ROA, Low Asset Turnover, High Leverage, Very High Revenue Growth, Negative Net Profit Margin.

Cluster 4: Large,Profitable Companies Characteristics: High Market Cap, Low to Moderate Beta, Moderate PE Ratio, Very Positive ROE and ROA, High Asset Turnover, Low Leverage, Moderate Revenue Growth, High Net Profit Margin.

Cluster 5: High Volatility Characteristics: Moderate Market Cap, Low Beta, High PE Ratio, Negative ROE and ROA, Moderate Asset Turnover, Low Leverage, High Revenue Growth, Negative Net Profit Margin.

The pharmaceutical industry can opt the method of kmeans clustering as it is suitable and helping it to form the groups appropriately and it can choose cluster number=4 companies as they have little variations and more similarities within the cluster.

Loading the required libraries.

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

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

```
## Warning: package 'lubridate' 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   3.4.4     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)
```

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

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

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

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

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

```
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
```

It's important to load the libraries for performing the functions and features provided by each package to perform specific tasks such as data manipulation, Clustering etc.

Importing and reading the dataset.

```
library(readr)
equity.Pharma<- read.csv("C:/Users/Sania fatima/Desktop/Clustering/Pharmaceuticals.csv")
dim(equity.Pharma)
```

```
## [1] 21 14
```

Loading the csv files and reading the dataset.After reading the dataset,we found there are 21 observation with 14 variables.

Dropping the categorical variables.

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

```
##     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
```

The reason for excluding the specific columns(2,12,13,14) is that they contain categorical variables that are not suitable for forming the clusters.It's a preprocessing step before applying the clustering techniques.

Normalizing the data by using the scale function.

```
Pharma.scaling.norm <- scale(Pharma.1)
head(Pharma.scaling.norm)
```

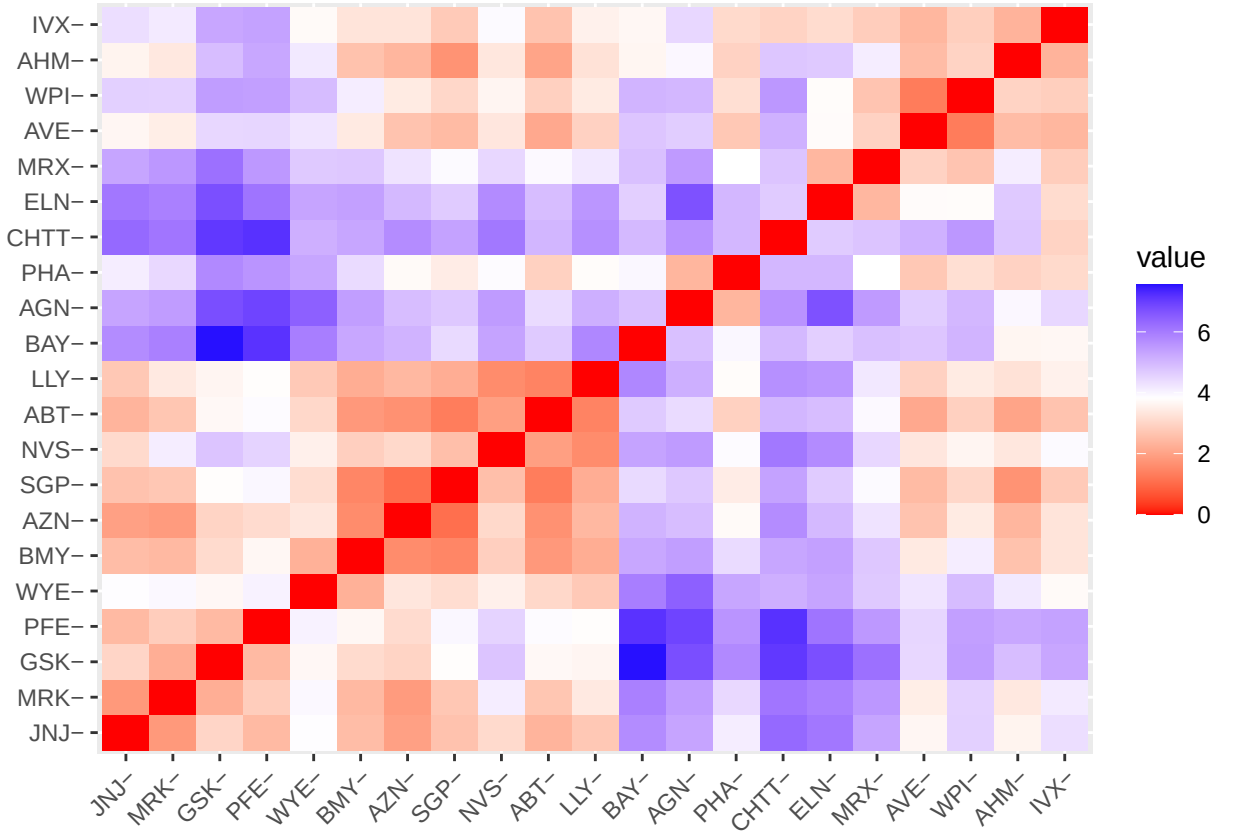
```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      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
```

As there might be variables which is unit less,the scaling is done to bring all variables to a common scale,which is very important for machine learning algorithms.It transforms the data so that each variable has a mean of 0 and standard deviation of 1.By standardizing the data,the variables will become comparable and the scale of variable will have no longer influences on the results of certain analyzes.It is therefore customary to normalize continuous measurements before computing the Euclidean distance. This converts all measurements to the same scale.

Clustering the data by using euclidean distance and plotting the graph.

```
Euclidean.distance <- get_dist(Pharma.scaling.norm)

Euclidean.distance %>%
  fviz_dist(order = TRUE, show_labels = TRUE)
```

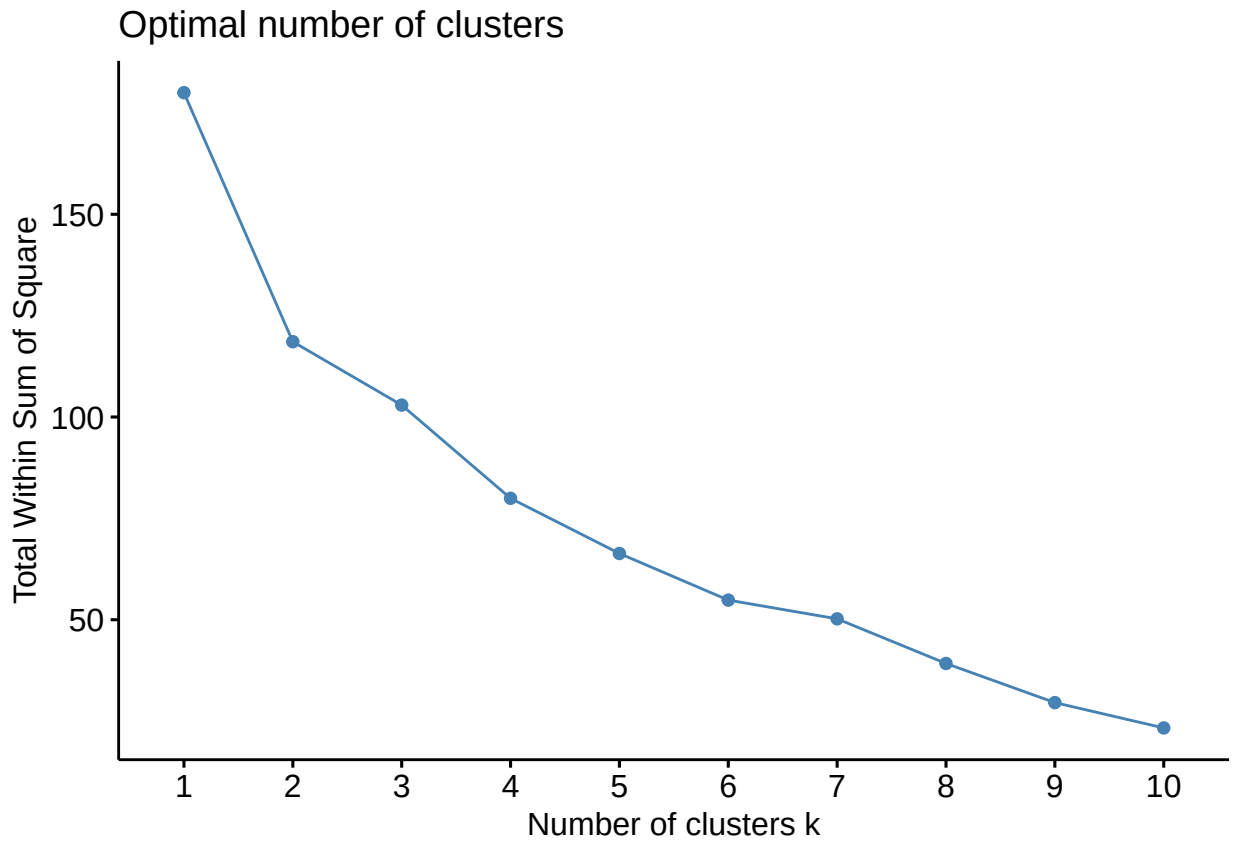


The distance from a point to itself will be 0, and dij = dji. Given that, we only need to specify distances as an upper or lower matrix. An important point to note is that the measure computed above is highly influenced by the scale of each variable, so that variables with larger scales have a much greater influence over the total distance. The choice of the distance measure plays a major role in cluster analysis. The main guideline is domain dependent.

Although Euclidean distance is the most widely used distance, it has three main features that need to be kept in mind. 1. It is scale dependent. 2. It completely ignores any relationship between measurements, so if the measurements are correlated, maybe other measures of distances might better. 3. It is sensitive to outliers.

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

```
# Using the "wss" method for determining the number of clusters
Pharma.scaling.norm %>%
  fviz_nbclust(FUNcluster = kmeans, method = "wss")
```

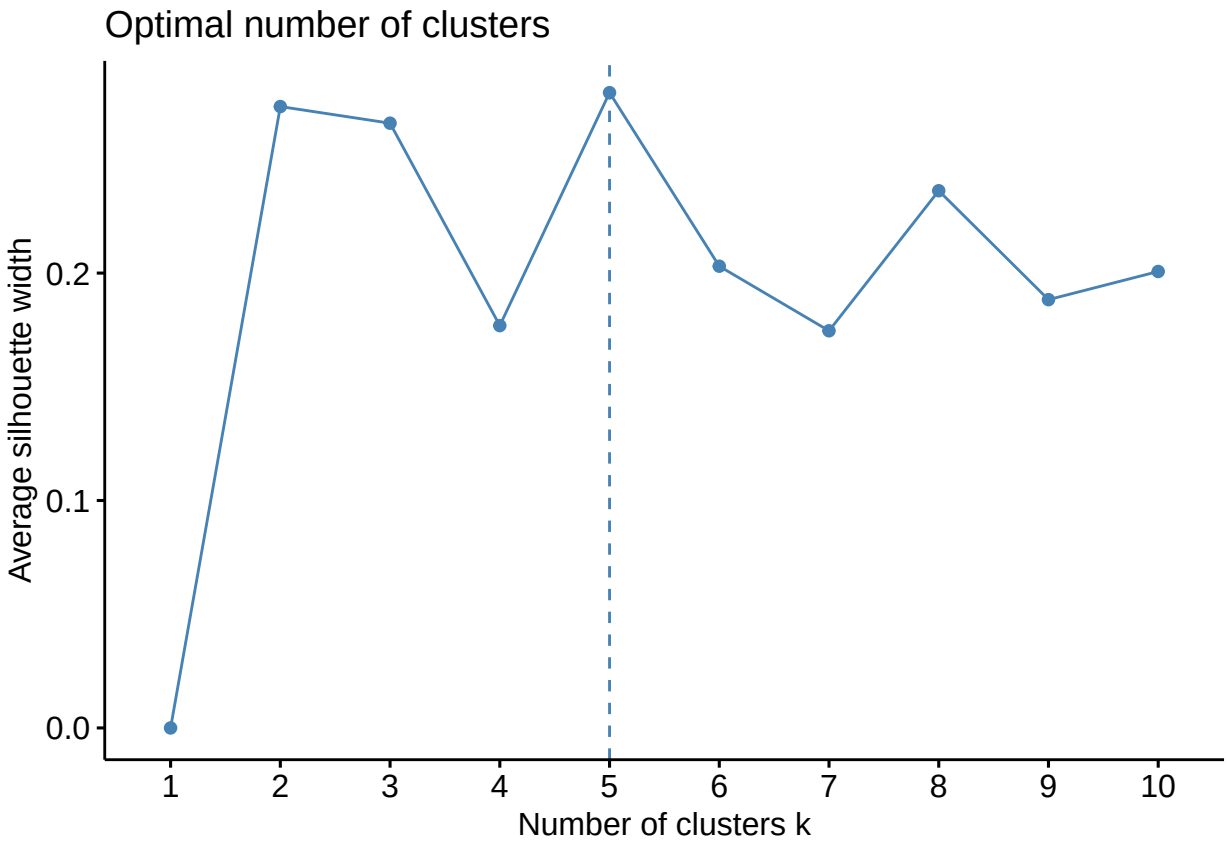


K=5 An elbow method is a line chart that shows how much the data is spread out within each cluster as we add more clusters. It is used to find the best number of clusters for our Pharmaceuticals industry by looking for the point where the line starts to flatten out, which is called the "elbow". This is the point where adding more clusters doesn't significantly improve the way the data is clustered.

Here,also comes the concept of WSS(within sums of squares) which means,the measure of variations within the clusters.The lower the variations within the clusters, the similar the datapoints are to each other.

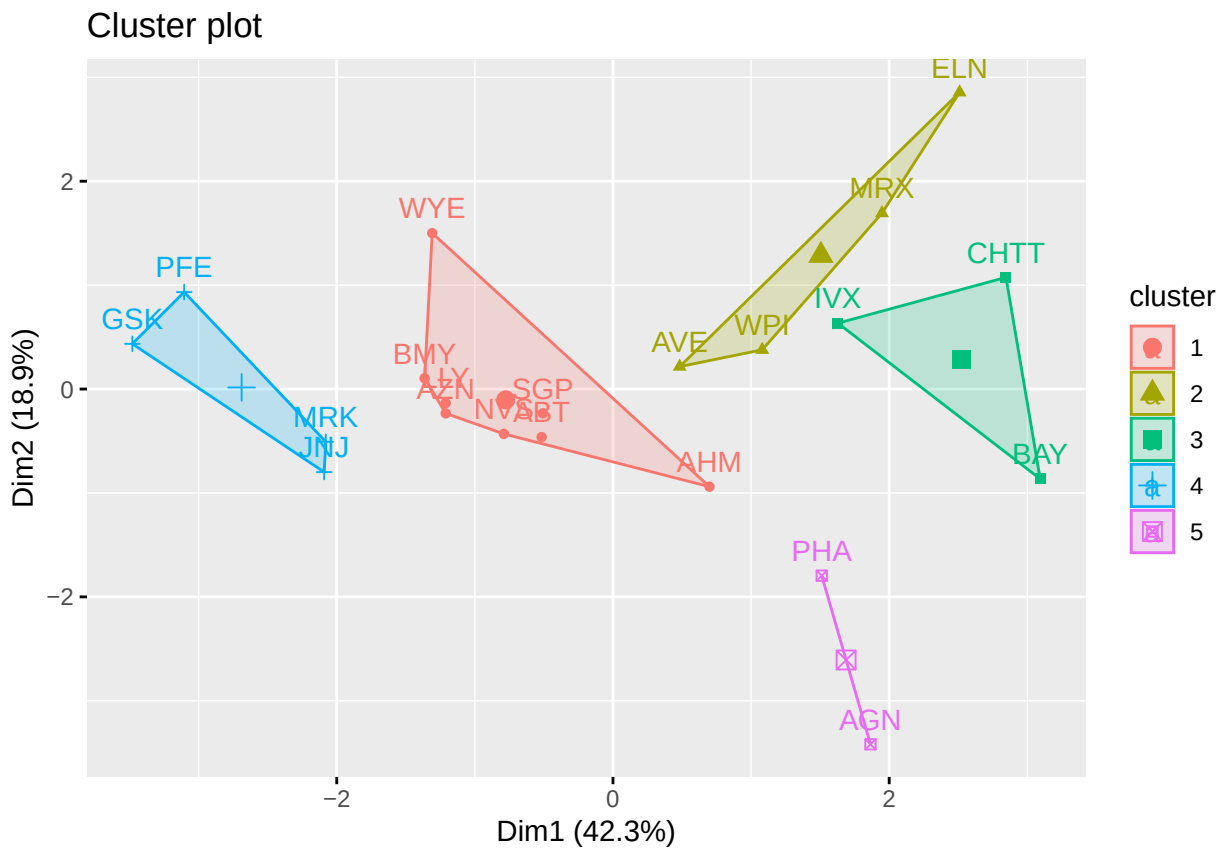
The chart shows that the elbow point 5 provides the best value for k. While WSS will continue to drop for larger values of k, we have to make the tradeoff between overfitting, i.e., a model fitting both noise and signal, to a model having bias. Here,the elbow point provides that compromise where WSS, while still decreasing beyond k = 5, decreases at a much smaller rate. In other words, k=5 provides the best value between bias and overfitting.

```
# Using the "silhouette" method for determining the number of clusters
Pharma.scaling.norm %>%
  fviz_nbclust(FUNcluster = kmeans, method = "silhouette")
```



To guarantee the validity and consistency of record assignments to clusters, we can employ the "Silhouette Method". This approach is straightforward, assessing the similarity of an object to its designated cluster relative to other clusters. The Silhouette Method yields values within the -1 to +1 range.High values indicate a good matching of data to clusters.From the Graph,it is clear that 5 is the ideal number of clusters moreover,we look for large values for the Silhouette Width (Y Axis).

```
set.seed(2)
kmeans.Pharma <- kmeans(Pharma.scaling.norm, centers = 5, nstart = 25)
fviz_cluster(kmeans.Pharma, data = Pharma.scaling.norm)
```




```
kmeans.Pharma$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
```

```
kmeans.Pharma$size
```

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

```
kmeans.Pharma$withinss
```

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

K-Means:The k-means algorithm is a partitional clustering algorithm to group 'n' objects based on attributes into K partitions, where k<n.

The choice of the number of clusters can either be driven by external considerationsor the methods we have opted such as elbow,silhouette or we can try a few values stochastically. After choosing k, the n records are partitioned into these initial clusters. If there is external reasoning for the initial assignment, we can apply

that, but otherwise, this is done randomly(nstart=25). In these cases, the algorithm can be return with different randomly generated starting partitions to reduce chances of the heuristic producing a poor solution.

By viewing the above figure,we can say that we formed 5 clusters as by elbow and silhouette method ,we found our best K=5.

By seeing the clusters formed,we can interpret as follows: CLUSTER NUMBER:1 -> the size of cluster number 1 is 8 and the companies included in this cluster are Bristol-Mysers Squibb Company(BMY),AstraZeneca PLC(AZN),Eli Lily and Company(LLY),Schering-plough corporation(SGP),Novartis AG(NVS),Abbott Laboratories(ABT),Amersham plc(AHM),Wyeth(WYE),where the variations within this cluster (WWS) is 21.879 that means there are a little huge amount of dissimilarities present within the clusters with different centers/centroid value of every aspect of Pharmaceutical Industry such as Beta[-0.436],ROA[0.408],Net profit margin[0.556] and so on.As a result, I can say that this cluster is maintaining its current operations.

CLUSTER NUMBER:2 -> the size of cluster number 2 is 4 and the companies included in this cluster are Aventis(AVE),Watson Pharamaceuticals, Inc.(WPI),Medicis Pharmaceuticals Corporation(MRX),Elan Corporation,plc(ELN) ,where the variations within this cluster (WWS) is 12.791 (where the distance or variation within the cluster is low in comparison with cluster 1) with different centers/centroid values of every aspect of Pharmaceutical Industry such as Market capital[-0.760],Beta[0.279],Leverage[0.60],Revenue Growth[1.518] and so on. Overall,I can say that this cluster focuses on improving profitability and efficiency.

CLUSTER NUMBER:3 -> the size of cluster number 3 is 3 and the companies included in this cluster are IVAX Corporation(IVX),Chattern,Inc(CHTT),Bayer AG(BAY) ,where the variations within this cluster (WWS) is 15.595(that means it has quite high dissimilarity within the cluster when compared to cluster 2) with different centers value of every aspect of Pharmaceutical Industry such as Market capital[-0.870],leverage[1.366],Beta[1.340] and so on.As a result,I can say that this particular cluster is addressing financial challenges and reduce leverage.

CLUSTER NUMBER:4 -> the size of cluster number 4 is 4 and the companies included in this cluster are GlaxoSmithKline plc(GSK),Pfizer Inc(PFE),Merck & Co.,Inc.(MRK),Johnson & Johnson (JNJ),where the variations within this cluster (WWS) is 9.284 (it has minimal similarities within the clusters) with different centers value of every aspect of Pharmaceutical Industry such as Market capital[1.695],Beta[-0.178],Net profit margin[0.591] and so on.Overall, I can interpret that the pharmaceutical industry present in this cluster are continuing current strategies and monitoring for sustained growth.

CLUSTER NUMBER:5 -> the size of cluster number 5 is 2 and the companies included in this cluster are Allergan Inc(AGN),Pharmacia Corporation (PHA),where the variations within this cluster (WWS) is 2.803(which has very little dissimilarities within its grouping) with different centers/centroid value of every aspect of Pharmaceutical Industry such as Market capital[-0.439],Beta[-0.470],PE ratio[2.700] and so on.To intrepret on this cluster number 5 , I can say that this cluster evaluate and possibly adjust pricing or cost structure for better profitability.

#Reasons for choosing cluster Number=4

In this case,the cluster number=4 is recommended because,as we compared the other clusters,the Net profit margin which is the essential characteristic for every firm to infer; whether the firm is landing towards the gains or losses or it could survive or not in the market .So, cluster number=4 of Pharmaceutical industry gains the net profit of 0.59 which is comparatively high when compared to other clusters.

Though the Revenue growth of the firm is little less when compared to cluster 5,we cannot eliminate the cluster based on determination of one factor of industry for choosing the right grouping.Moreover,The debts of purchasing the assets of this industry is literally low (-0.468) as compared to other clusters.

The Assets_Turnover(sales income to the company) of cluster number=4 is 1.153, which records high and its mandatory for the company to have large sales income to gain high amount of profit.

ROA(Return on Assets) is a type of return on investment metric that measures the profitability of a business in relation to its total asset.The contribution of ROA is 1.350 which is a positive sign for pharmaceutical industry to earn profits.

The Return on equity(ROE) falls for cluster number 4 is 1.2349 ,from which we can infer that the pharmaceutical industry is efficient at generating profits.The fundamentality of PE(profit earning ratio) is that the lower it is, the better.Keeping in mind the other factors to determine which cluster is right to choose,we might have drawbacks of the company as there is no company formed without a single drawback.So,The PE Ratio of the cluster number 4 is -0.198.

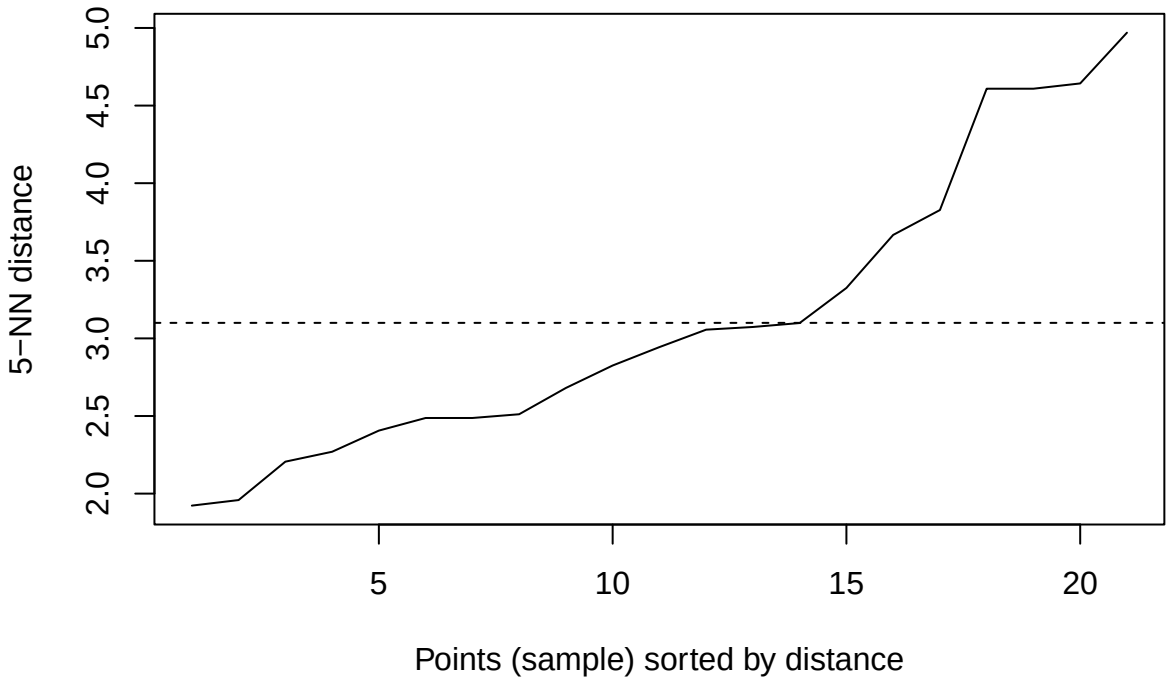
The ideal beta(risk) for the companies is 1 where our pharamaceutical industry is having -0.178 that can be consider as the good feature for selecting the cluster.

The Higher Market_capital (1.695) typically represents stability and its a type of corporate assets for the industry which is peak when compared to other clusters.

These are the reasons for choosing the cluster number =4 as the best.The companies in cluster number 4 are GlaxoSmithKline plc(GSK),Pfizer Inc(PFE),Merck & Co.,Inc.(MRK),Johnson & Johnson (JNJ) which represents the variations and distance within the cluster is less and the variations or distances, between the clusters are large in comparison with other clusters. In Pharmaceutical industry, there are no outliers as every datapoints(companies) are important and it forms the clusters approriately.

Finding the epsilon Value

```
dbscan::kNNdistplot(Pharma.scaling.norm, k=5)
abline(h=3.1, lty=2)
```

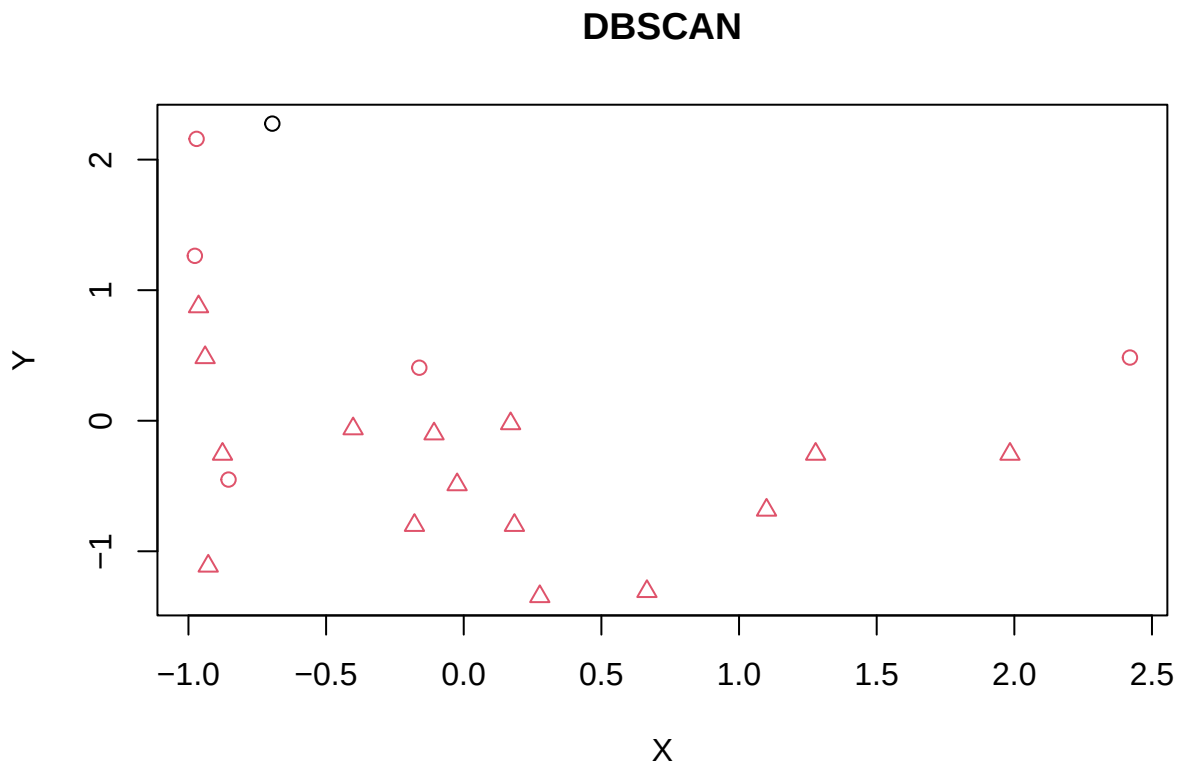


According to the plot the optimal epsilon value is 3.1

The method proposed here uses the concept of k-nearest neighbor distances. We calculate the average distance of every point to its k nearest neighbors. The value of k is chosen by the modeler and corresponds to MinPts. We will then plot these distances against the sampled points, and then identify the value at which there is a sharp change in values. This is similar to identifying the elbow when determining k in k-means.This distance can be seen as a measure of the local density of points.

Clustering using DBScan.

```
DBscan.Pharma <- fpc::dbscan(Pharma.scaling.norm, eps = 3.1, MinPts = 5)
plot(DBscan.Pharma, Pharma.scaling.norm, main="DBSCAN", frame= TRUE, xlab = "X", ylab = "Y")
```



DBSCAN requires two parameters: epsilon (eps) and the minimum number of points required to form a dense region (minPts). It starts with an arbitrary starting point that has not been visited. This point's epsilon-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized epsilon-environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its epsilon-neighborhood is also part of that cluster. Hence, all points that are found within the epsilon-neighborhood are added, as is their own epsilon-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

In this case of Pharmaceutical industry, the dbscan is inappropriate as it is not forming the clusters and taking the important data points as noise points. DBSCAN cannot cluster data sets well with large differences in densities(3.1),since the minPts and epsilon combination cannot then be chosen appropriately for all clusters. If the data and scale are not well understood, choosing a meaningful distance threshold epsilon can be difficult. The quality of DBSCAN depends on the distance measure . The most common distance metric used is Euclidean distance. Especially for high-dimensional data, this metric can be rendered almost useless due to the so- called "Curse of dimensionality", making it difficult to find an appropriate value for epsilon.

Plotting a dendogram using agnes()

```
Hierarchical.single <- agnes(Pharma.scaling.norm, method = "single")
Hierarchical.complete <- agnes(Pharma.scaling.norm, method = "complete")
Hierarchical.ward <- agnes(Pharma.scaling.norm, method = "ward")
```

```
Hierarchical.average <- agnes(Pharma.scaling.norm, method = "average")

print(Hierarchical.single$ac)
```

```
## [1] 0.4600348
```

```
print(Hierarchical.complete$ac)
```

```
## [1] 0.6990833
```

```
print(Hierarchical.ward$ac)
```

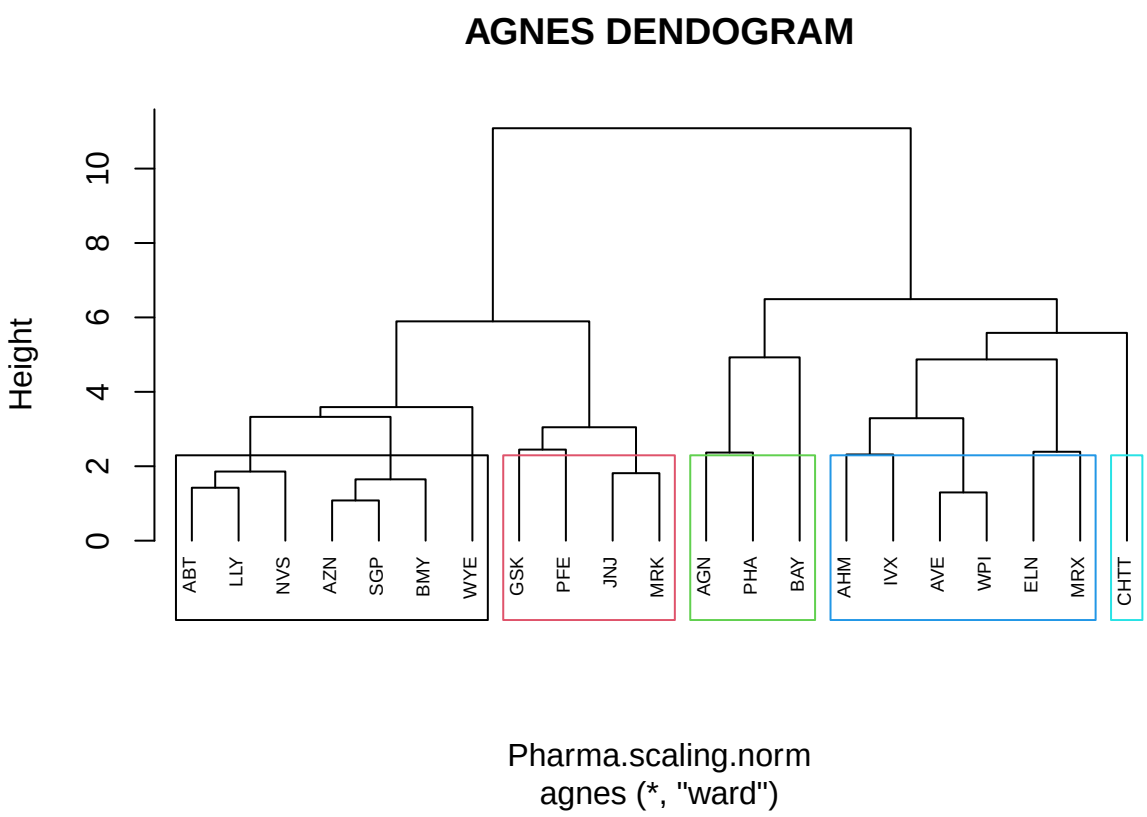
```
## [1] 0.7943164
```

```
print(Hierarchical.average$ac)
```

```
## [1] 0.5600652
```

```
pltree(Hierarchical.ward, cex = 0.6, hang = -1, main = "AGNES DENDOGRAM")

rect.hclust(Hierarchical.ward, k=5, border = 1:5)
```



Hierarchical methods can be either agglomerative or divisive. Agglomerative methods begin with n clusters and sequentially merge similar clusters until a single cluster is obtained. Divisive methods work in the opposite direction, starting with one cluster that includes all records.

A dendrogram is a treelike diagram that summarizes the process of clustering.Similar records are joined by lines whose vertical length reflects the distance between the records. The above figure shows the dendrogram for our example. CLUSTER NO:1:It includes the comapnies such as ABT,LLY,NVS,AZN,SGP,BMY,WYE

and it has a height length greater than 2 CLUSTER NO:2:The companies included in this cluster are GSK,PFE,JNJ,MRK which represents that their heights are similar in nature within the cluster however have larger difference between the cluster. CLUSTER NO:3:The pharmaceutical industry included in this cluster are AGN,PHA,BAY.The cluster is formed again on the basis of similar heights. CLUSTER NO:4:The Pharmaceutical industry included in this cluster are AHM,IVX,AVE,WPI,ELN,MRX. CLUSTER NO:5: It has only one company that is CHTT which shows 0 variations in itself as it considered it as an outlier.Therefore, this dataset has no outliers as every company is important in Pharmaceutical industry. Heirarchical clustering is very sensitive to outliers. Note that the figure is affected by the choice of distance, e.g., Euclidean versus City, and the type of linkage, centroid, single,etc

With respect to the choice of distance between clusters, single and complete linkage are robust to changes in the distance metric (e.g., Euclidean, statistical distance) as long as the relative ordering is kept. In contrast, average linkage is more influenced by the choice of distance metric, and might lead to completely different clusters when the metric is changed.

when we tested the methods employ such as SINGLE,COMPLETE,WARD,AVERAGE.It can be inferred that ward is the best approach here as we look for the high agglomerative coefficient value where it has 0.79431 as compared to other methods.So,the higher the agglomerative coefficient value,higher the cohesion of clusters are.

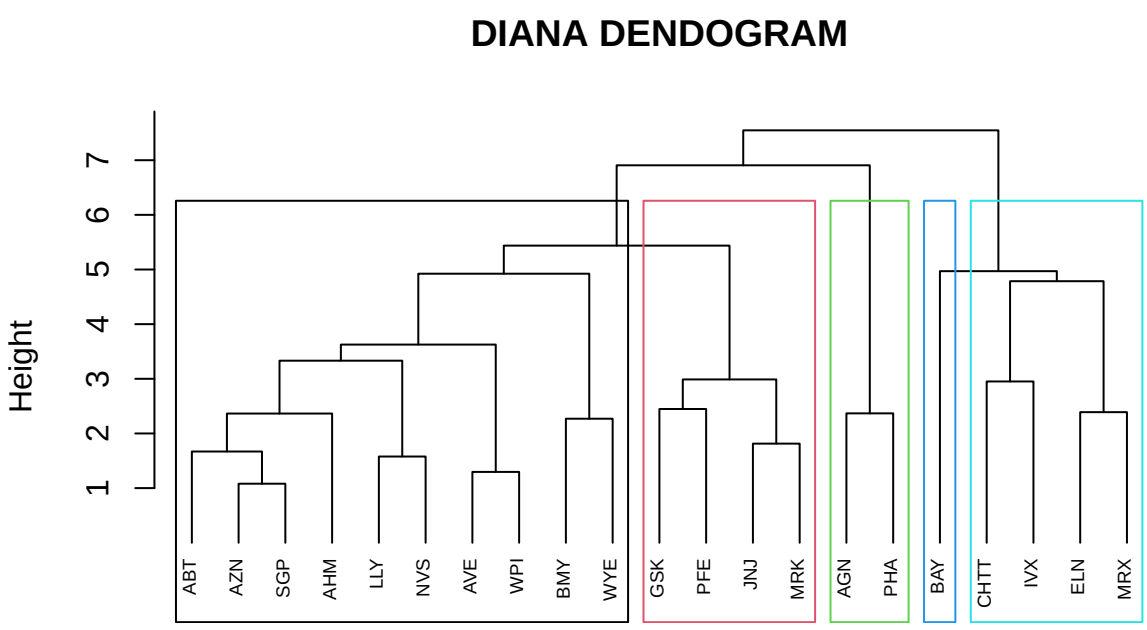
Plotting a dendogram using diana()

```
Hierarchical.diana <- diana(Pharma.scaling.norm)

print(Hierarchical.diana$dc)
```

```
## [1] 0.7135475
```

```
pltree(Hierarchical.diana, cex = 0.6, hang = -1, main = "DIANA DENDOGRAM")
rect.hclust(Hierarchical.diana, k=5, border = 1:5)
```



Diana is also the part of hierarchical clustering and it is basically used with large amount of data.AS the dataset is very small of 21 firms Diana will not be suitable to cluster the groups .Though it will form the clusters it starts with taking all the records into consideration from the first step.

AGNES VERSUS DIANA Generally speaking,the output of both AGNES and DIANA are comparable. Both have similar output complexity. Agnes is most commonly used technique. It is empirically suggested that AGNES is used for identifying small clusters where as DIANA is used for identifying the large clusters.

As a result,we can say that Hierarchical clustering is inappropriate to select because of outliers and AGNES which is suitable for forming small clusters are considering the vital records are outliers as there are no outliers present in this dataset.On the other hand, because of small dataset and formation of small clusters ,Diana is not at all suitable for the Pharmaceuticals industry.Moreover,Hierarchical methods are especially useful when the goal is to arrange the clusters into a natural hierarchy.

#KMEANS VS DBSCAN VS HIERARCHICAL

KMEANS are easy to implement and by elbow and silhouette method we found k value is 5 and its groups the clusters in an effective way without involving any data as outliers..It reiterates till the centroid value doesnot get change. Whereas,DBSCAN is not forming the clusters properly and it is considering the important records as noisy data. HIERARCHICAL is suitable to arrange the clusters into a natural hierarchy.Though, its forming the clusters appropriately by both AGNES and DIANA but its considering the crucial datapoints are outliers.

#Which Clustering technique is Appropriate to choose for pharmaceutical industry?

Kmeans is the best clustering technique to choose for this Pharmaceutical industry .The reason behind this is the dataset doesn't have outliers and it is correctly forming the clusters based on the objective that the similarities/variations within the clusters are less and between the clusters are high.The distance is a fundamental concept for knowing the diference between the clusters. K-means provides easily interpretable results, with each data point assigned to a specific cluster.This algorithm aims to minimize the within sum of

squared(WSS) distances between data points and their respective cluster centroids. (b) Dropping the first two variables While dropping the numerical variables, we can have the correct column as a categorical variables

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

```
##   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
```

The table shows the three categorical variables at the last such as Median_Recommendation,Location,Exchange to see whether there is an impact of these variables or not.

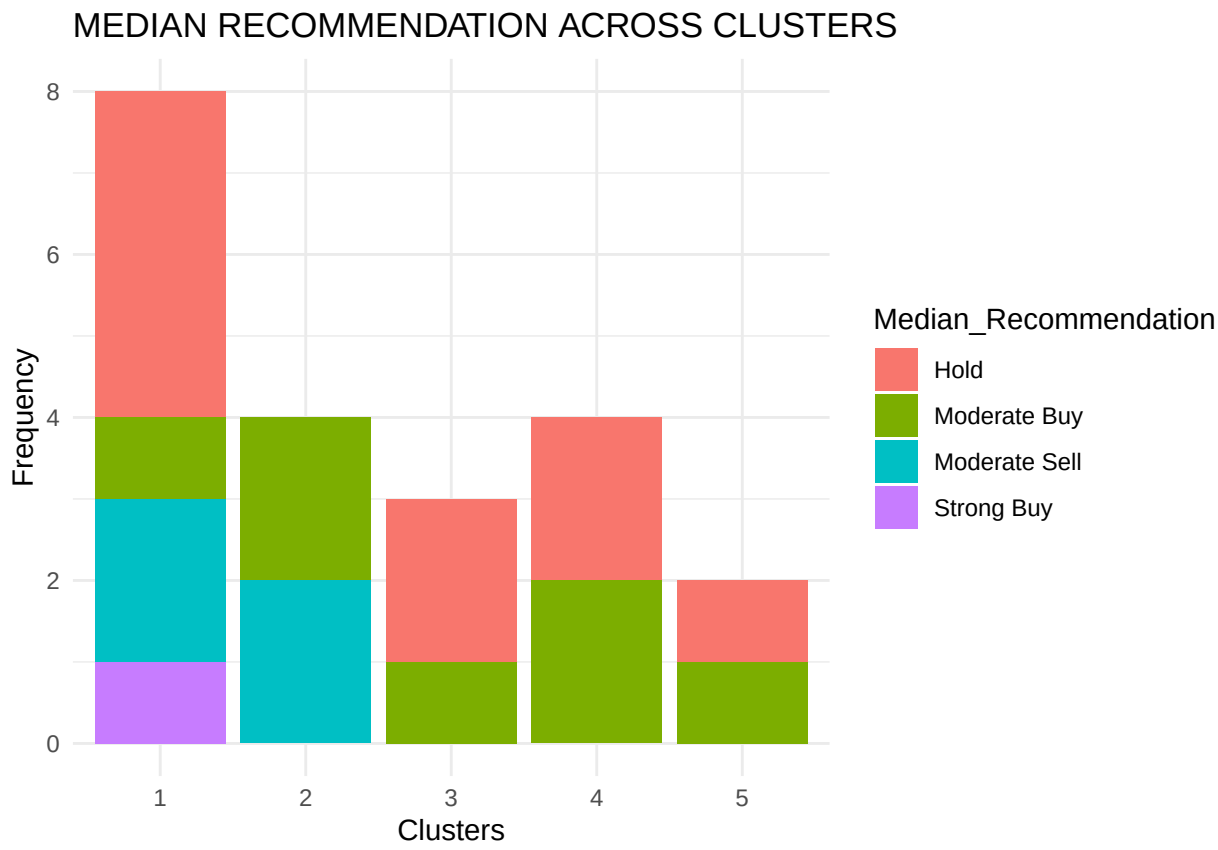
Categorical variables: In this chunck,we selected the variables which are categorical in nature to identify the pattern in the clusters used in forming the clusters:

```
Pharmaceuticals.Pattern <-  Pharmaceuticals %>% select(c(10,11,12)) %>% mutate(Cluster = kmeans.Pharma$
print(Pharmaceuticals.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
```

Yes,there is a pattern in the clusters where the Exchanges are from NYSE,NASDAQ,AMEX and the location of the companies include US,CANADA,UK,FRANCE,GERMANY,IRELAND,SWITZERLAND and the median recommendation includes MODERATE BUY,MODERATE SELL,HOLD,STRONG BUY.

```
Median.Recommendation <- ggplot(Pharmaceuticals.Pattern, aes(x = factor(Cluster), fill = Median_Recommer
  geom_bar() +
  labs(x = 'Clusters', y = 'Frequency', title = 'MEDIAN RECOMMENDATION ACROSS CLUSTERS') +
  theme_minimal()
Median.Recommendation
```



The MEDIAN RECOMMENDATION graph shows that the analysts of the Pharmaceutical industry suggest that the majority of the products is on HOLD except cluster number 2.This implies a neutral stance. Analysts believe investors may want to neither buy more nor sell their existing holdings. It could suggest a period of stability or uncertainty.

There is a MODERATE BUY in almost all clusters(1 to 5) which represents the positive outlook, but with some reservations. It could mean that analysts see potential for growth, but there might be certain risks or uncertainties associated with the investment.

The cluster number 1 and 2 have MODERATE SELL which represents the negative outlook for the stock, but with some moderation. Analysts may see reasons for concern, but they might not view the situation as extremely dire. It could indicate a more cautious recommendation to reduce exposure to the stock.

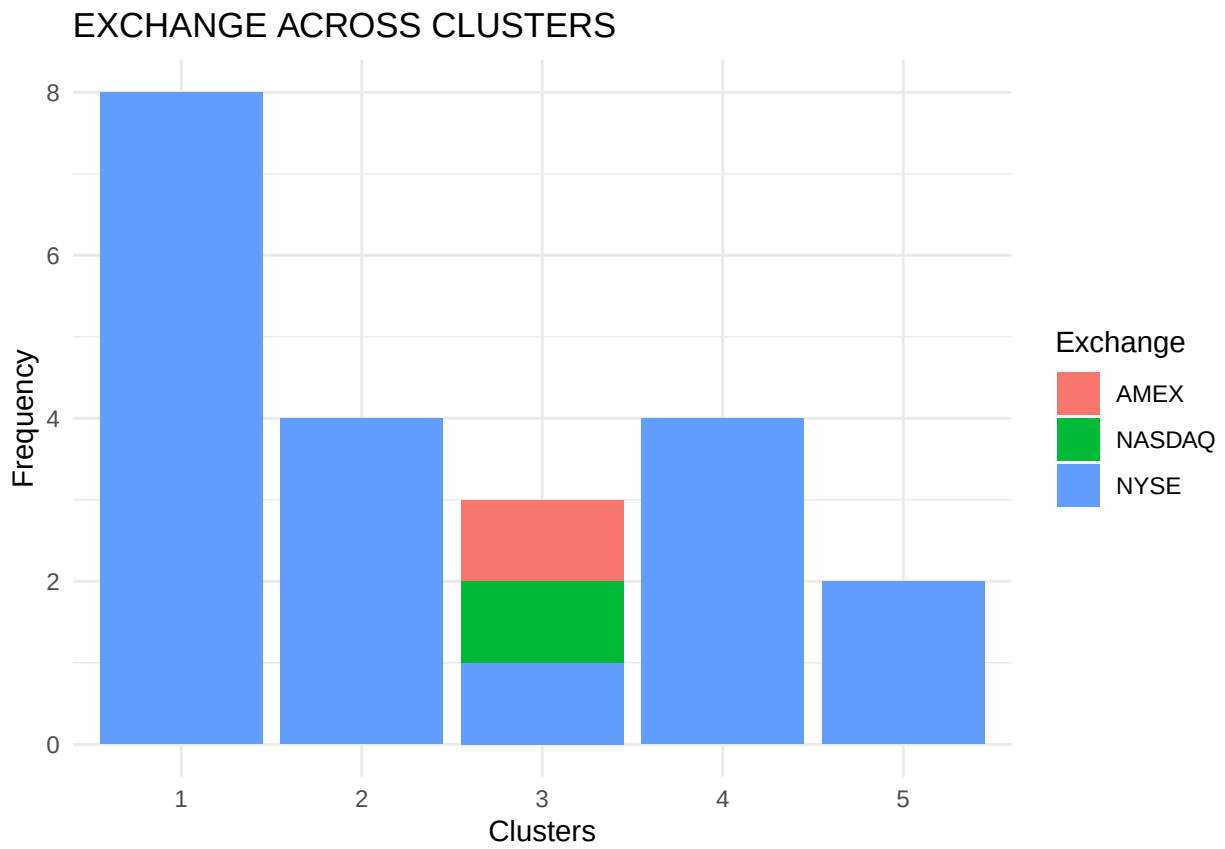
CLuster Number 1 is only the one where the analyst suggests the STRONG BUY.Analysts believe the stock has strong potential for appreciation, and investors are encouraged to consider buying.

```
# Bar plot for Location across Clusters
location.plot <- ggplot(Pharmaceuticals.Pattern, aes(x = factor(Cluster), fill = Location)) +
  geom_bar() +
  labs(x = 'Clusters', y = 'Frequency', title = 'LOCATION ACROSS CLUSTERS') +
  theme_minimal()
location.plot
```

After Analyzing the graph of LOCATION ACROSS CLUSTER,the US is the dominant country where it has huge amount of Pharmaceutical industry.In other words,every cluster has the industries from US.Cluster 1 and 4 are the the companies from UK .IRELAND and FRANCE Pharmaceutical companies are present only in cluster 2.The less covered location part across cluster of pharmaceutical industry is CANADA which is present in Cluster 5.Moreover, Cluster 1 has the companies where the location is from SWITZERLAND.

```
# Bar plot for Exchange across Clusters
Exchange.plot <- ggplot(Pharmaceuticals.Pattern, aes(x = factor(Cluster), fill = Exchange)) +
  geom_bar() +
  labs(x = 'Clusters', y = 'Frequency', title = 'EXCHANGE ACROSS CLUSTERS') +
  theme_minimal()

Exchange.plot
```



The EXCHANGE graph shows that all clusters 1,2,4,5 have the exchange from NYSE.Though cluster 3 have the exchange from NYSE however it is also having the echanges from two other exchange listing securities, they are NASDAQ and AMEX. The advantages of having single exchange listing is simplicity and focus,clear market identity,exclusivity etc

To interpret all type of categorical variables , I can say that these three variables (Median Recommendation,location,exchange) have the patterns while forming the clusters with respect to numerical variables.

#NAMING THE CLUSTERS Cluster 1: Stable Growth Companies Characteristics: Moderate Market Cap, Low Beta, Moderate PE Ratio, Positive ROE and ROA, Moderate Asset Turnover, Low Leverage, Moderate Revenue Growth, High Net Profit Margin.

Cluster 2: High Risk, High Leverage Companies Characteristics: Low Market Cap, Moderate Beta, Low PE Ratio, Negative ROE and ROA, Low Asset Turnover, Moderate to High Leverage, High Revenue Growth, Variable Net Profit Margin.

Cluster 3: Financial Conservative Companies Characteristics: Very Low Market Cap, High Beta, High PE Ratio, Negative ROE and ROA, Low Asset Turnover, High Leverage, Very High Revenue Growth, Negative Net Profit Margin.

Cluster 4: Large,Profitable Companies Characteristics: High Market Cap, Low to Moderate Beta, Moderate PE Ratio, Very Positive ROE and ROA, High Asset Turnover, Low Leverage, Moderate Revenue Growth, High Net Profit Margin.

Cluster 5: High Volatility Characteristics: Moderate Market Cap, Low Beta, High PE Ratio, Negative ROE and ROA, Moderate Asset Turnover, Low Leverage, High Revenue Growth, Negative Net Profit Margin.