Assignment\_4

Pranay Kumar Kodeboyina

2023-03-19

#install.packages("flexclust")  
#install.packages("cluster")  
#install.packages("tidyverse")  
#install.packages("factoextra")  
#install.packages("FactoMineR")  
#install.packages("ggcorrplot")  
  
library(tinytex)  
library(flexclust)

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library (cluster)  
library (tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library (factoextra)

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

library (FactoMineR)  
library(ggcorrplot)

library(readr)  
  
#Extracting the current working directory  
getwd()

## [1] "/Users/kodeboyina/Documents/Kent State/Sem1/Fundamentals of ML/Assignment4"

Pharma<-read\_csv("/Users/kodeboyina/Documents/Kent State/Sem1/Fundamentals of ML/Assignment4/Pharmaceuticals.csv")

## Rows: 21 Columns: 14  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Symbol, Name, Median\_Recommendation, Location, Exchange  
## dbl (9): Market\_Cap, Beta, PE\_Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(Pharma)

#Transpose of a matrix  
t(t(names(Pharma)))

## [,1]   
## [1,] "Symbol"   
## [2,] "Name"   
## [3,] "Market\_Cap"   
## [4,] "Beta"   
## [5,] "PE\_Ratio"   
## [6,] "ROE"   
## [7,] "ROA"   
## [8,] "Asset\_Turnover"   
## [9,] "Leverage"   
## [10,] "Rev\_Growth"   
## [11,] "Net\_Profit\_Margin"   
## [12,] "Median\_Recommendation"  
## [13,] "Location"   
## [14,] "Exchange"

#summary Of the Data  
summary(Pharma)

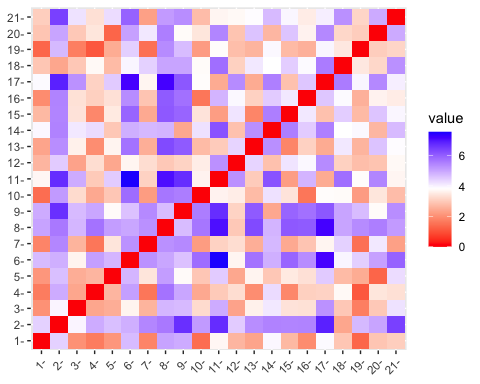
## Symbol Name Market\_Cap Beta   
## Length:21 Length:21 Min. : 0.41 Min. :0.1800   
## Class :character Class :character 1st Qu.: 6.30 1st Qu.:0.3500   
## Mode :character Mode :character Median : 48.19 Median :0.4600   
## Mean : 57.65 Mean :0.5257   
## 3rd Qu.: 73.84 3rd Qu.:0.6500   
## Max. :199.47 Max. :1.1100   
## PE\_Ratio ROE ROA Asset\_Turnover Leverage   
## Min. : 3.60 Min. : 3.9 Min. : 1.40 Min. :0.3 Min. :0.0000   
## 1st Qu.:18.90 1st Qu.:14.9 1st Qu.: 5.70 1st Qu.:0.6 1st Qu.:0.1600   
## Median :21.50 Median :22.6 Median :11.20 Median :0.6 Median :0.3400   
## Mean :25.46 Mean :25.8 Mean :10.51 Mean :0.7 Mean :0.5857   
## 3rd Qu.:27.90 3rd Qu.:31.0 3rd Qu.:15.00 3rd Qu.:0.9 3rd Qu.:0.6000   
## Max. :82.50 Max. :62.9 Max. :20.30 Max. :1.1 Max. :3.5100   
## Rev\_Growth Net\_Profit\_Margin Median\_Recommendation Location   
## Min. :-3.17 Min. : 2.6 Length:21 Length:21   
## 1st Qu.: 6.38 1st Qu.:11.2 Class :character Class :character   
## Median : 9.37 Median :16.1 Mode :character Mode :character   
## Mean :13.37 Mean :15.7   
## 3rd Qu.:21.87 3rd Qu.:21.1   
## Max. :34.21 Max. :25.5   
## Exchange   
## Length:21   
## Class :character   
## Mode :character   
##   
##   
##

#Question A 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.  
set.seed(123)  
Pharma1<-Pharma[,3:11]  
t(t(names(Pharma1)))

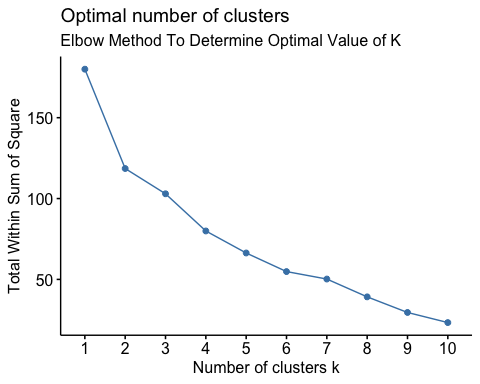
## [,1]   
## [1,] "Market\_Cap"   
## [2,] "Beta"   
## [3,] "PE\_Ratio"   
## [4,] "ROE"   
## [5,] "ROA"   
## [6,] "Asset\_Turnover"   
## [7,] "Leverage"   
## [8,] "Rev\_Growth"   
## [9,] "Net\_Profit\_Margin"

#Using Euclidean distance formula which is given by $d = $

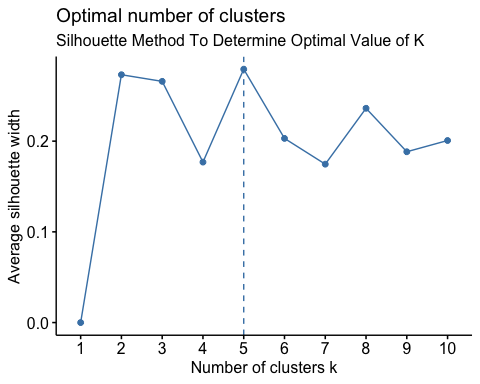
## Normalizing the data frame  
Pharma2<-scale(Pharma1)  
Dist\_Pharma<-get\_dist(Pharma2, method = "euclidean", stand = FALSE)  
  
#Visualizing the distance metrics to identify the clusters and distance from each metric  
fviz\_dist(Dist\_Pharma, order = FALSE, show\_labels = TRUE, lab\_size = NULL, gradient = list(low = "red", mid = "white", high = "blue"))



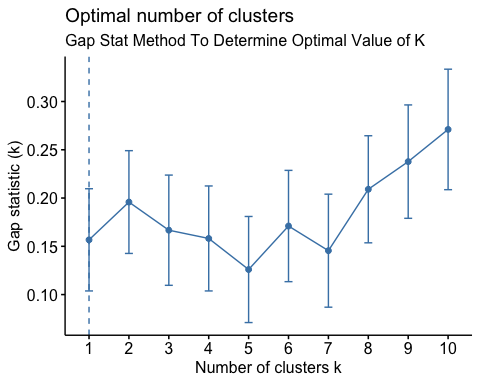
#In order determine the optimal value of k, employing different methods to determine k  
pharma\_elbow <- scale(Pharma1)  
fviz\_nbclust(pharma\_elbow, kmeans, method = "wss") +labs(subtitle = "Elbow Method To Determine Optimal Value of K")



pharma\_Silhouette <- scale(Pharma1)  
fviz\_nbclust(pharma\_Silhouette, kmeans, method = "silhouette") + labs(subtitle = "Silhouette Method To Determine Optimal Value of K")

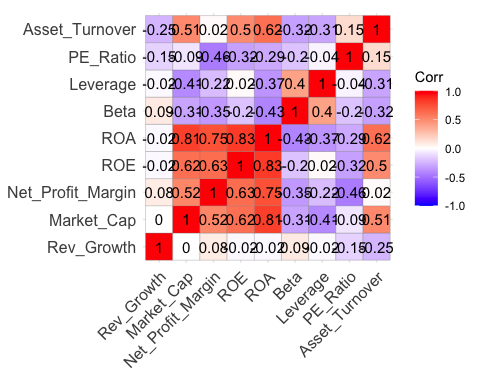


pharma\_gap\_stat <- scale(Pharma1)  
fviz\_nbclust(pharma\_gap\_stat, kmeans, method = "gap\_stat") + labs(subtitle = "Gap Stat Method To Determine Optimal Value of K")



##In this case, we have three different recommendations for the optimal value of K: K=5 from the elbow and silhouette methods, and K=1 from the gap statistic method. The final value of K = 5 from the above two methods it also depends on the additional factors such as interpretability, computational efficiency, and the stability of the clustering solution.

correlation <- cor(Pharma2)  
  
#correlation matrix plot using the correlation matrix.  
ggcorrplot(correlation, outline.color = "grey50", lab = TRUE, hc.order = TRUE, type = "full")



set.seed(123)  
kmean <- kmeans(Pharma2, centers = 5, nstart = 25)  
  
  
## Below finding cluster center for all rows and colomns  
kmean$centers

## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 2 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 3 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 4 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 5 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.27449312 -0.7041516 0.556954446  
## 2 1.36644699 -0.6912914 -1.320000179  
## 3 -0.14170336 -0.1168459 -1.416514761  
## 4 -0.46807818 0.4671788 0.591242521  
## 5 0.06308085 1.5180158 -0.006893899

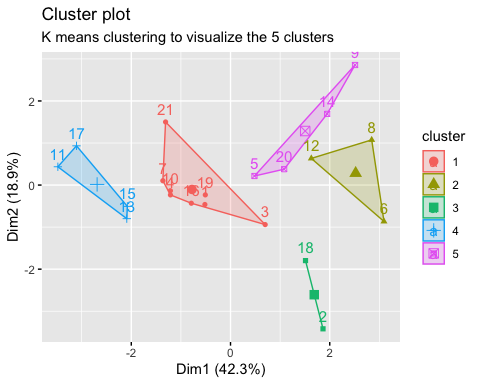
## Number of observation in each cluster  
  
kmean$size

## [1] 8 3 2 4 4

## Observation and their respective cluster label.  
kmean$cluster[c(1:21)]

## [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1

##For the above observations applyng the K means clustering to visualize the 5 clusters  
fviz\_cluster(kmean, data=Pharma2) + labs(subtitle = "K means clustering to visualize the 5 clusters")

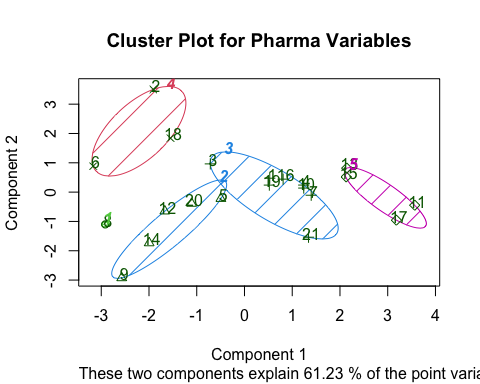


kmean

## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4  
##   
## Cluster means:  
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915 0.1729746  
## 2 -0.87051511 1.3409869 -0.05284434 -0.6184015 -1.1928478 -0.4612656  
## 3 -0.43925134 -0.4701800 2.70002464 -0.8349525 -0.9234951 0.2306328  
## 4 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431 1.1531640  
## 5 -0.76022489 0.2796041 -0.47742380 -0.7438022 -0.8107428 -1.2684804  
## Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -0.27449312 -0.7041516 0.556954446  
## 2 1.36644699 -0.6912914 -1.320000179  
## 3 -0.14170336 -0.1168459 -1.416514761  
## 4 -0.46807818 0.4671788 0.591242521  
## 5 0.06308085 1.5180158 -0.006893899  
##   
## Clustering vector:  
## [1] 1 3 1 1 5 2 1 2 5 1 4 2 4 5 4 1 4 3 1 5 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 21.879320 15.595925 2.803505 9.284424 12.791257  
## (between\_SS / total\_SS = 65.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

## On the cluster graph above, we can see that there are 5 clusters, each with its own color and shape. The center of the cluster is the centriod or the center point. We have reached the final center points after 25 restarts as there is no change until and unless the new data is added.

library(cluster)  
kmean2 <- kmeans(Pharma2,5)  
clusplot(Pharma2, kmean2$cluster, color=TRUE, shade=TRUE, labels=2, lines=0, main="Cluster Plot for Pharma Variables")



#Question B. Interpret the clusters with respect to the numerical variables used in forming the clusters.  
  
#Calculating the means of the clusters based o its components to intrepret the data set  
aggregate(Pharma2,by=list(kmean2$cluster),FUN=mean)

## Group.1 Market\_Cap Beta PE\_Ratio ROE ROA  
## 1 1 -0.97676686 1.2630872 0.03299122 -0.1123792 -1.1677918  
## 2 2 -0.79605926 0.3205014 -0.45014035 -0.6533148 -0.7881923  
## 3 3 -0.03142211 -0.4360989 -0.31724852 0.1950459 0.4083915  
## 4 4 -0.52462814 0.4451409 1.84984387 -1.0404550 -1.1865838  
## 5 5 1.69558112 -0.1780563 -0.19845823 1.2349879 1.3503431  
## Asset\_Turnover Leverage Rev\_Growth Net\_Profit\_Margin  
## 1 -4.612656e-01 3.7427970 -0.6327607 -1.2488842  
## 2 -1.107037e+00 0.2717048 1.2256188 -0.1486179  
## 3 1.729746e-01 -0.2744931 -0.7041516 0.5569544  
## 4 -3.330669e-16 -0.3443544 -0.5769454 -1.6095439  
## 5 1.153164e+00 -0.4680782 0.4671788 0.5912425

K <- data.frame(Pharma2, kmean2$cluster)  
view(K)  
K

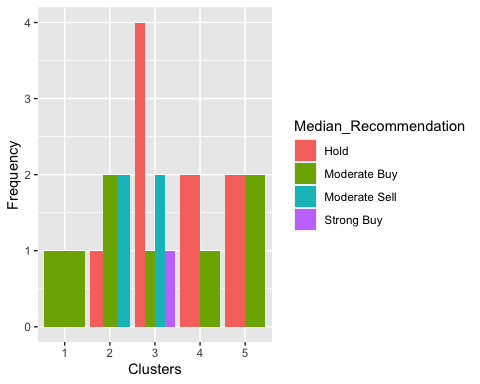
## Market\_Cap Beta PE\_Ratio ROE ROA Asset\_Turnover  
## 1 0.1840960 -0.80125356 -0.04671323 0.04009035 0.2416121 -5.121077e-16  
## 2 -0.8544181 -0.45070513 3.49706911 -0.85483986 -0.9422871 9.225312e-01  
## 3 -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700 9.225312e-01  
## 4 0.1702742 -0.02225704 -0.24290879 0.10638147 0.9181259 9.225312e-01  
## 5 -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461 -4.612656e-01  
## 6 -0.6953818 2.27578267 0.14948233 -1.45146000 -1.7127612 -4.612656e-01  
## 7 -0.1078688 -0.10015669 -0.70887325 0.59693581 0.8617498 9.225312e-01  
## 8 -0.9767669 1.26308721 0.03299122 -0.11237924 -1.1677918 -4.612656e-01  
## 9 -0.9704532 2.15893320 -1.34037772 -0.70899938 -1.0174553 -1.845062e+00  
## 10 0.2762415 -1.34655112 0.14948233 0.34502953 0.5610770 -4.612656e-01  
## 11 1.0999201 -0.68440408 -0.45749769 2.45971647 1.8389364 1.383797e+00  
## 12 -0.9393967 0.48409069 -0.34100657 -0.29136529 -0.6979905 -4.612656e-01  
## 13 1.9841758 -0.25595600 0.18013789 0.18593083 1.0872544 9.225312e-01  
## 14 -0.9632863 0.87358895 0.19240011 -0.96753478 -0.9610792 -1.845062e+00  
## 15 1.2782387 -0.25595600 -0.40231769 0.98142435 0.8429577 1.845062e+00  
## 16 0.6654710 -1.30760129 -0.23677768 -0.52338423 0.1288598 -9.225312e-01  
## 17 2.4199899 0.48409069 -0.11415545 1.31287998 1.6322239 4.612656e-01  
## 18 -0.0240846 -0.48965495 1.90298017 -0.81506519 -0.9047030 -4.612656e-01  
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593 0.5234929 4.612656e-01  
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905 -9.225312e-01  
## 21 -0.1614497 0.40619104 -0.75792214 1.92938746 0.5422849 -4.612656e-01  
## Leverage Rev\_Growth Net\_Profit\_Margin kmean2.cluster  
## 1 -0.21209793 -0.52776752 0.06168225 3  
## 2 0.01828430 -0.38113909 -1.55366706 4  
## 3 -0.40408312 -0.57211809 -0.68503583 3  
## 4 -0.74965647 0.14744734 0.35122600 3  
## 5 -0.31449003 1.21638667 -0.42597037 2  
## 6 -0.74965647 -1.49714434 -1.99560225 4  
## 7 -0.02011273 -0.96584257 0.74744375 3  
## 8 3.74279705 -0.63276071 -1.24888417 1  
## 9 0.61983791 1.88617085 -0.36501379 2  
## 10 -0.07130879 -0.64814764 1.17413980 3  
## 11 -0.31449003 0.76926048 0.82363947 5  
## 12 1.10620040 0.05603085 -0.71551412 2  
## 13 -0.62166634 -0.36213170 0.33598685 5  
## 14 0.44065173 1.53860717 0.85411776 2  
## 15 -0.39128411 0.36014907 -0.24310064 5  
## 16 -0.67286239 -1.45369888 1.02174835 3  
## 17 -0.54487226 1.10143723 1.44844440 5  
## 18 -0.30169102 0.14744734 -1.27936246 4  
## 19 -0.74965647 -0.43544591 0.29026942 3  
## 20 -0.49367621 1.43089863 -0.09070919 2  
## 21 0.68383297 -1.17763919 1.49416183 3

#From the above data interpreting the data based on the values  
  
#Clusters 1:- JNJ, MRK, GSK, PFE  
#(Highest Market\_Cap and Highest Net\_Proft\_Margin, Highest ROE/ROA, Highest Asset\_Turnover, Higheest Net\_Profit\_Margin)  
  
##Cluster 2:- AVE, ELN, IVX, MRX, WPI  
#(Highest Revenue Growth and lowest PE/Asset Turnover Ratio)  
  
#Cluster 3:- AGN,BAY, PHA  
#(Highest PE ratio and lowest ROE/ROA and Net\_Profit\_Margin)  
  
#Cluster 4:- CHTT  
#(Lowest Market\_Cap, Highest Beta, Highest Levarage)  
  
#Cluster 5:- ABT, AHM, AZN, BMY, LLY, NVS, SGP, WYE  
# (Lowest Net\_Proft\_Margin and lowest Rev\_Growth)

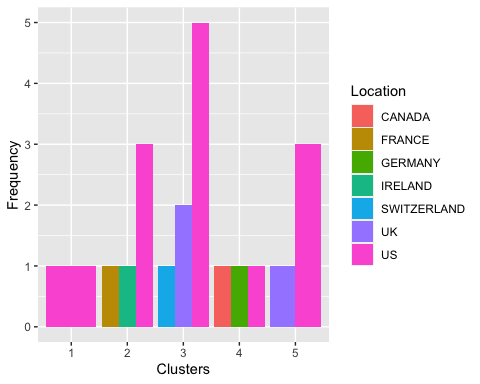
#Question C. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)  
  
#To examine the patterns by visualizing clusters against the variables grouped by clusters and to identify any trends in the data  
# Pharma is the original dataset and df1 is the subset of df  
# col 12,13 and 14 are the col of 10,11 and 12 of the subset of df  
pattern <- Pharma %>% select(c(12,13,14)) %>% mutate(Cluster = kmean2$cluster)  
print(pattern)

## # A tibble: 21 × 4  
## Median\_Recommendation Location Exchange Cluster  
## <chr> <chr> <chr> <int>  
## 1 Moderate Buy US NYSE 3  
## 2 Moderate Buy CANADA NYSE 4  
## 3 Strong Buy UK NYSE 3  
## 4 Moderate Sell UK NYSE 3  
## 5 Moderate Buy FRANCE NYSE 2  
## 6 Hold GERMANY NYSE 4  
## 7 Moderate Sell US NYSE 3  
## 8 Moderate Buy US NASDAQ 1  
## 9 Moderate Sell IRELAND NYSE 2  
## 10 Hold US NYSE 3  
## # … with 11 more rows

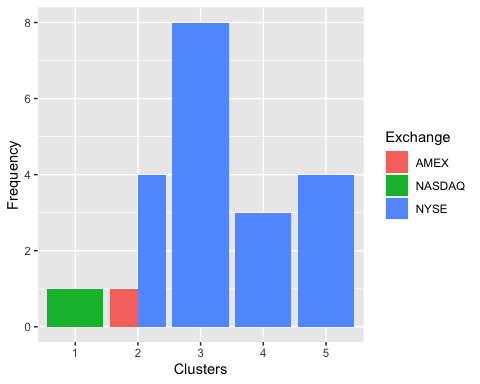
view(pattern)  
  
ggplot(pattern, mapping = aes(factor(Cluster), fill=Median\_Recommendation)) + geom\_bar(position = 'dodge') + labs(x='Clusters', y='Frequency')



ggplot(pattern, mapping = aes(factor(Cluster), fill=Location)) + geom\_bar(position = 'dodge') + labs(x='Clusters', y='Frequency')



ggplot(pattern, mapping = aes(factor(Cluster), fill=Exchange)) + geom\_bar(position = 'dodge') + labs(x='Clusters', y='Frequency')



#Based on the recommendation there is a pattern in the data  
  
#Cluster 1 Offers Moderate buy medians, as well as a different count from the US, but the businesses are evenly dispersed on the NASDAQ.  
  
#Cluster 2 Offers Hold, Moderate Buy, Moderate Sell medians, as well as a different count from the France, Ireland & US, but the businesses are evenly dispersed on the AMEX & NASDAQ.  
  
#Cluster 3 Offers High Hold, Moderate Buy/Sell, Strong Buy medians, as well as a different count from the Switzerland, UK & US, but the businesses are evenly dispersed on the NYSE.  
  
#Cluster 4 Offers Hold, Moderate Buy medians, as well as a different count from the Germany, Canada & US, but the businesses are evenly dispersed on the NYSE.  
  
#Cluster 5 Offers Hold, Moderate Buy/Sell, Strong Buy medians, as well as a different count from the UK & US, but the businesses are evenly dispersed on the NYSE.

## Question D. Provide an appropriate name for each cluster using any or all of the variables in the dataset.  
  
#Cluster 1: Moderate Buy and Highest Market Cap  
  
#Cluster 2: Moderate Buy/Sell and Highest Revenue Growth   
  
#Cluster 3: High Hold and Highest PE ratio  
  
#Cluster 4: Moderate Hold and Lowest Market\_Cap  
  
#Cluster 5: Hold/Moderate Buy and Lowest Net\_Proft\_Margin