ANLY-501, Fall 2017

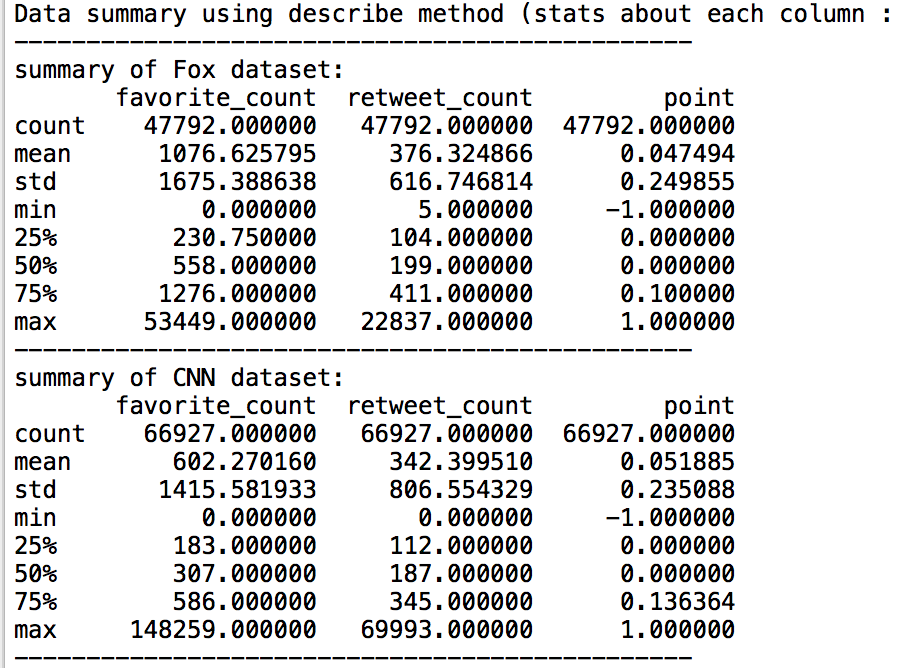
Group 7: Jiajia Liu Wupeng Han Jiaxu He

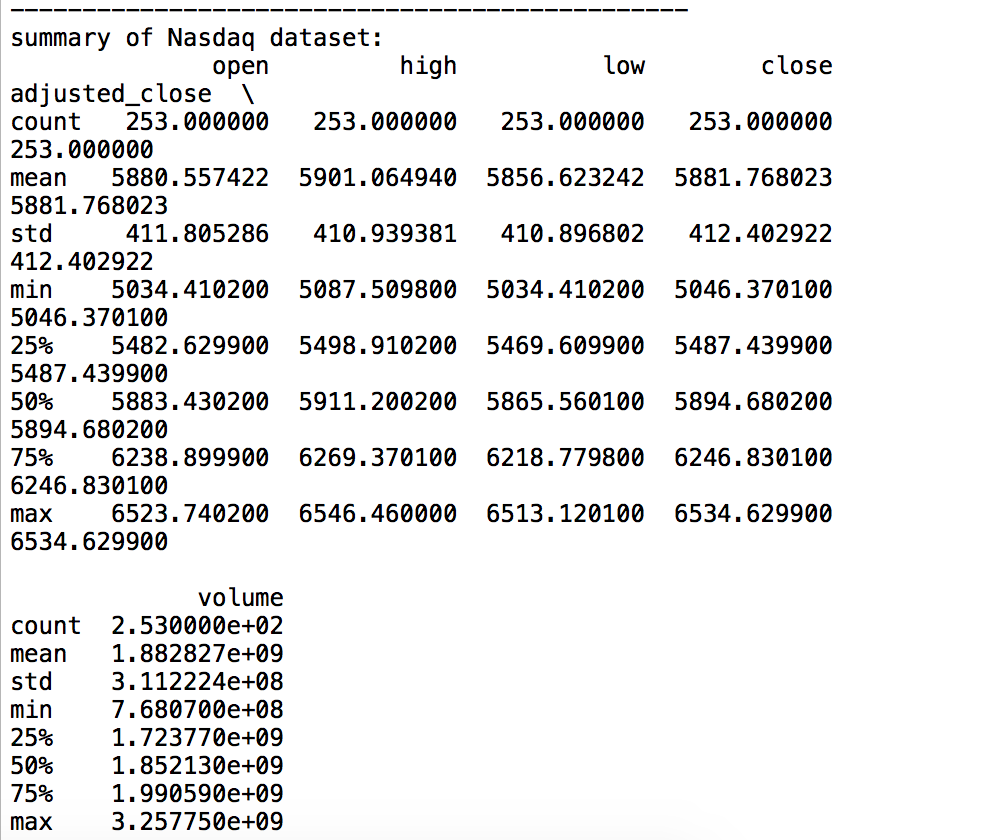
**Project Assignment 2**

**Determine the mean (mode if categorical), median, and standard deviation of at least 10 attributes in your data sets. Use Python to generate these results and use the project report to show and explain each.**

By the outlook of the datasets from CNN and Foxnews (*displayed in graph\_1*), there are 47992 samples of Foxnews datasets and 66927 samples of CNN dataset. For the Foxnews datasets, the means of the attributes “favorite\_count”, “retweet\_count”, and “point” are: 1076.63, 376.325, while the standard deviations are 1675.39, 616.75, 0.2498. For the CNN datasets, the means of the attributes “favorite\_count”, “retweet\_count”, and “point” are: 602.27, 342. 4, 0.0519, while the standard deviations are 1415.58, 806.55, 0.24. In these results, the value of the means tends to be greater than the value of the medians, so it is obvious that the distribution of the data appears to be skewed to the right. Then, the summary statistics are calculated separately by python. It is easy to see the differences in the center and spread of the data for each variable. For example, Fox datasets have a higher mean of “favorite \_count” and “retweet\_count” than CNN datasets.

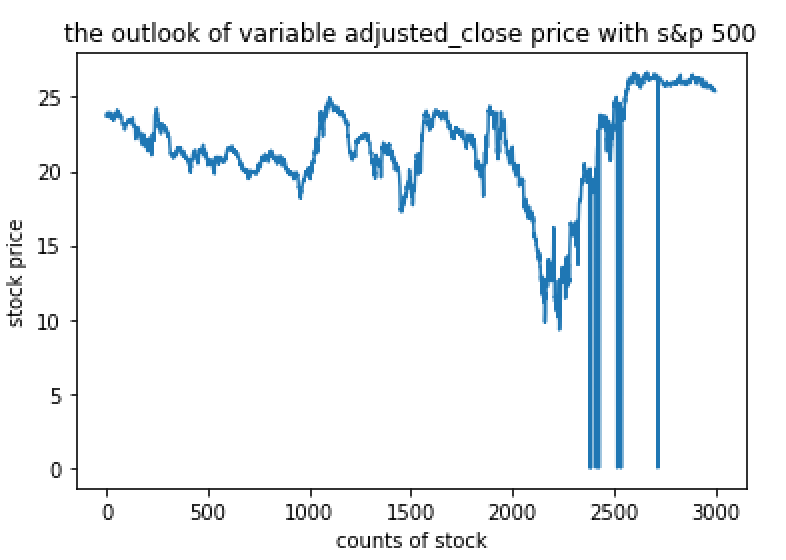
By the outlook of the datasets from three market datasets, which are gold, Nasdaq, S&P 500, there are 297 samples of gold dataset, 253 samples of Nasdaq dataset, and 252 samples of S&P 500 dataset. The results of mean, median, and standard deviation are presented in python output (*displayed in graph\_2*). In these results, the value of the means tends to be very close to the value of the medians for dataset gold and dataset S&P 500, so it is obvious that the distribution of the data appears to be normal. However, it is easy to see that the value of the means tends to be greater than the value of the medians for dataset Nasdaq, so the conclusion is that the distribution of the data appears to be skewed to the right.

 (graph\_1)

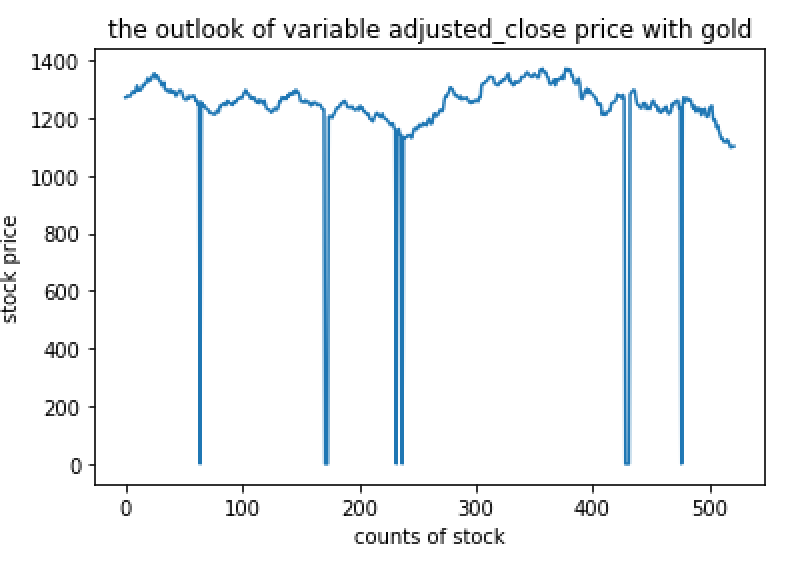
(graph\_2)

**Explain how you detected the outliers, and how you made the decision to keep or remove them.**

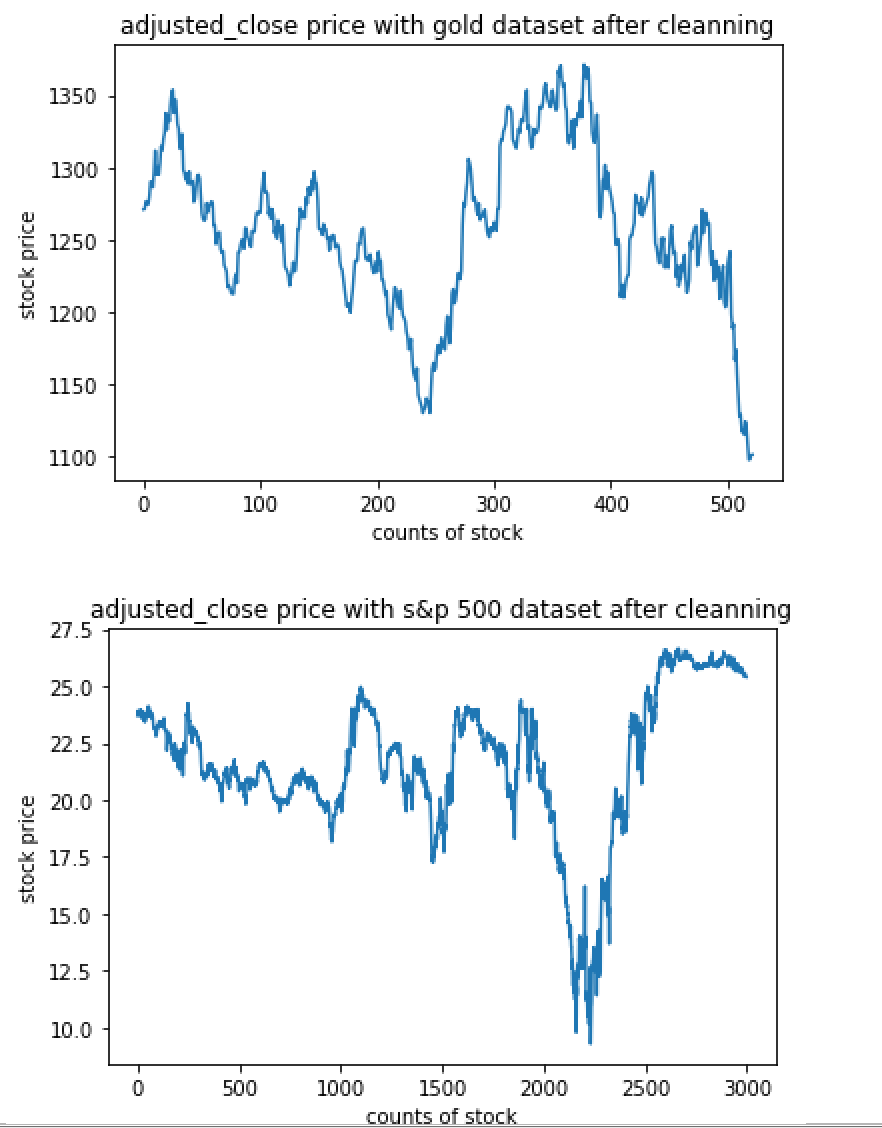
Boxplot and other regular plots are used to check the existence of outliers in the datasets. Outliers are points that diverge greatly from the overall pattern. It turns out that only the S&P 500 datasets and the gold datasets have 4 and 11 extreme outliers respectively (*displayed in graph\_3 and graph\_4*), since the stock price cannot be zero dollars in general. In fact, it might be a typo error. Therefore, in this analysis, it was decided to remove any points that have a price of zero. (*displayed in graph\_5*) However, none of the outliers were deleted for Fox dataset and CNN dataset, because the outliers for the variable “retweet” may signify some important news taking place.



(graph\_3)



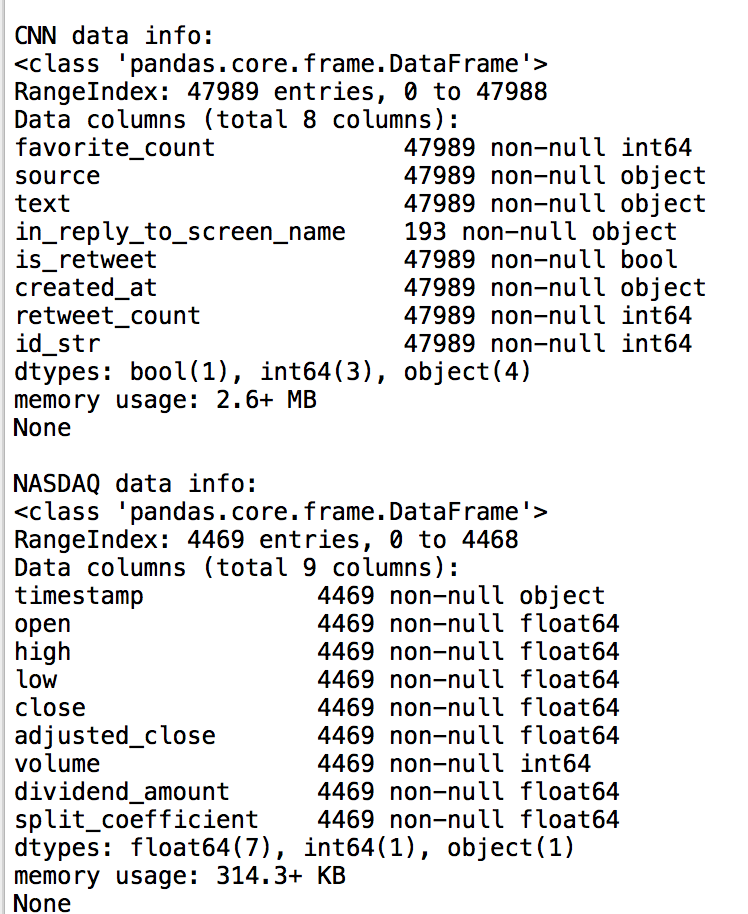
(graph\_4)



(graph\_5)

**From the cleaning phase of Project 1, also discuss which attributes had missing values and explain your strategy for handling them.**

Missing value means that there is no number or no content in a given column. Basically, the *print info ()* function would check the presence of missing values in the first step, and it turns out all the columns are filled with either objects, integers, float, or bool (*displayed in graph\_6*). This result demonstrates that there are zero missing values for those five datasets.



(graph\_6)

**If you find that you data needs to be further cleaned or differently cleaned based on analyses, include explanations here. Be specific about what you did and why.**

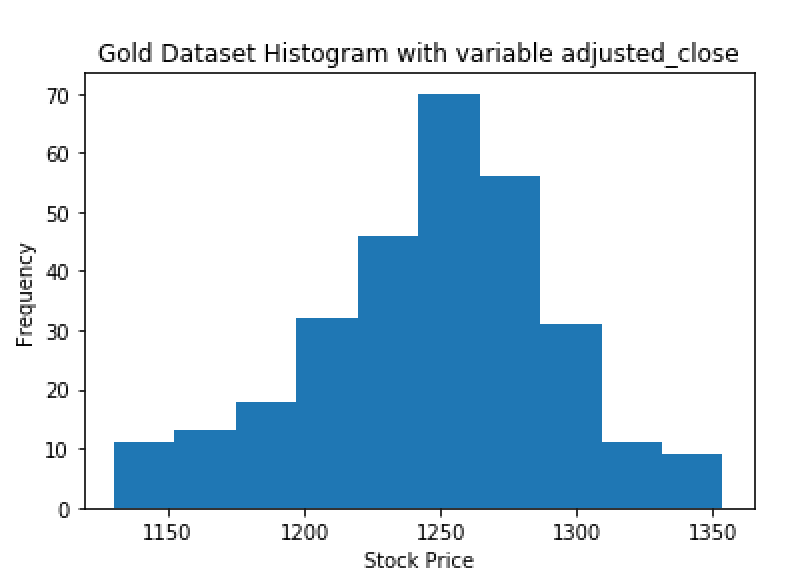
Most of the cleaning processes were finished in the project 1, except the size of the CNN dataset is a little bit different than other four datasets. Since the timelines are from October 2016 to October 2017 for other four datasets, so the data earlier than October 2017 were removed from CNN dataset. After cleaning, all of them are consistent with the same length of date.

**For at least one of the numeric variables in one of the datasets, write code to bin the data. This will create a new column. Use the binning strategy that is most intuitive for your data. Explain your decision. Include why you chose to bin the specific attribute selected, the binning method used, and why that method makes sense for your data.**

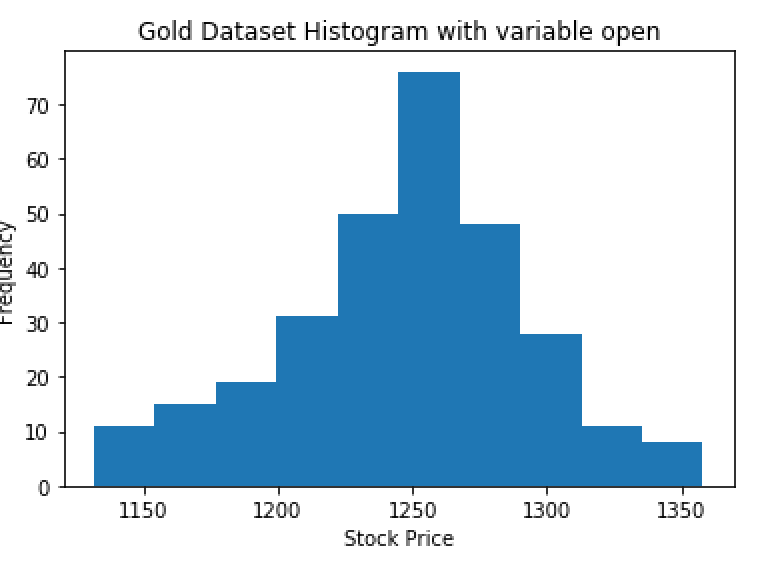
Two different binning strategies were applied in this project. First of all, both CNN dataset and Foxnew dataset’s variable “point” were binned into two different categorical levels: positive and negative. If the value of variable “point” exceeded 0, it was defined as “cpos”(point of CNN news is positive) and “fpos”(point of Fox news is positive), the other condition was defined as “cneg” (point of CNN news is negative) and “fneg”(point of Fox news is negative). Since the variable “point” is a continuous variable, it is not convenient to implement the continuous variables when doing the association rule analysis. Moreover, a new variable was created for other three stock datasets. This new variable named “increasing rate” was calculated by (today’s adjusted\_close – yesterday’s adjusted\_close) / yesterday’s adjusted\_close. When the value of this new variable exceeded 0, it was categorized as “nasin”(Nasdaq increasing) and “sp5in”(S&P500 increasing), the other condition was defined “nasde”(Nasdaq decreasing) and “sp5de”(S&P500 decreasing). For both binning strategies, they were very useful and handy while dealing with associations rules and other decisions.

**Use a histogram to plot at least three (3) of the variables (attributes) in either dataset. Discuss the insight generated by the histograms. What do they show or suggest?**

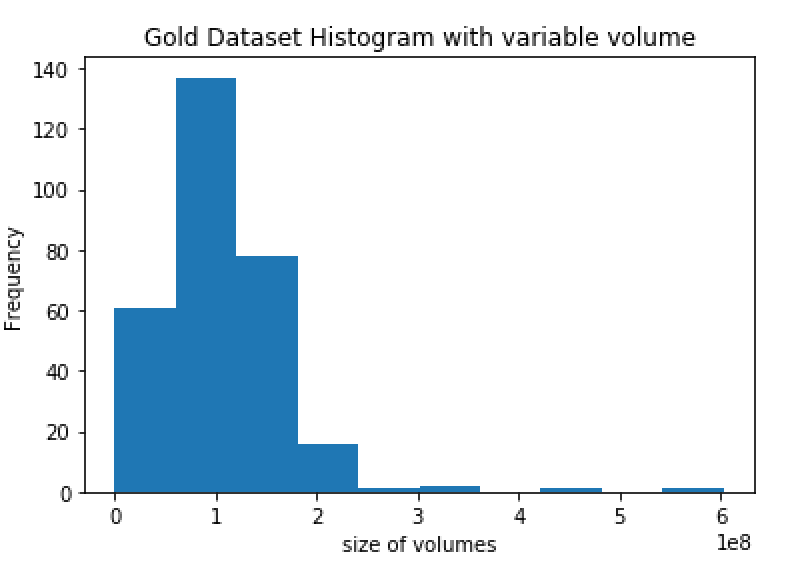
Variable “adjusted\_close” , variable “open”, and variable “volume” are chosen, because it were frequently used through the whole analysis. Also, those three categories were continuous and numeric variable, so it is essential to know the distribution of them in order to perform the linear regression model and ANOVA test in further steps. The histogram of variable “adjusted\_close” and “open” appears to be normal due to a strong “bell shaped”, while the histogram of variable “volume” shows a skewed right trend (*displayed in graph7, graph 8, and graph 9).* Then, for Nasdaq dataset, the histogram of variable “adjust close” appears to be normal with a strong spike on the right side, and the histogram of variable “volume” also presents a normal distribution with a strong “bell-shaped.” (*displayed in graph 10 and graph 11*)



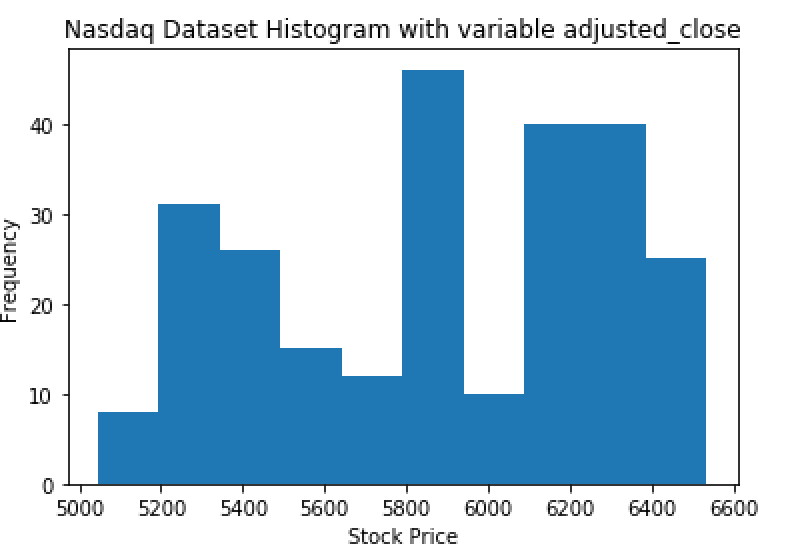
(graph 7)



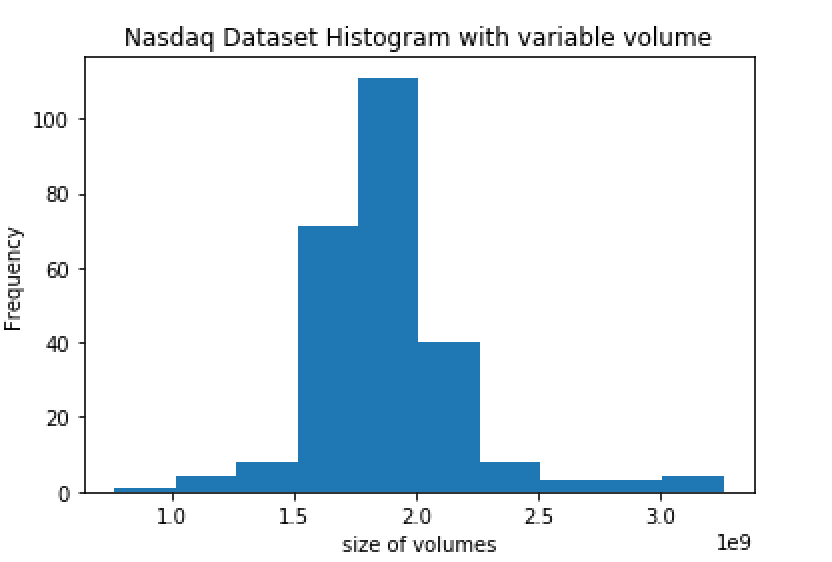
(graph 8)



(graph 9)

****

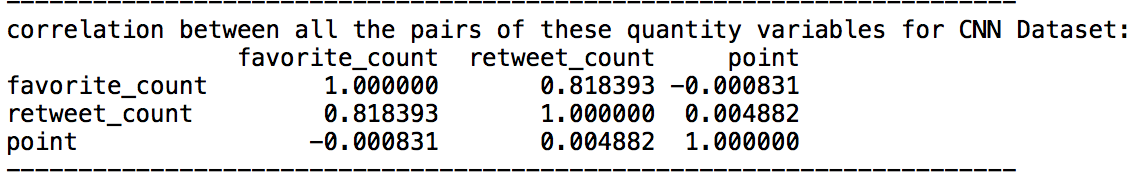
(graph 10)

****

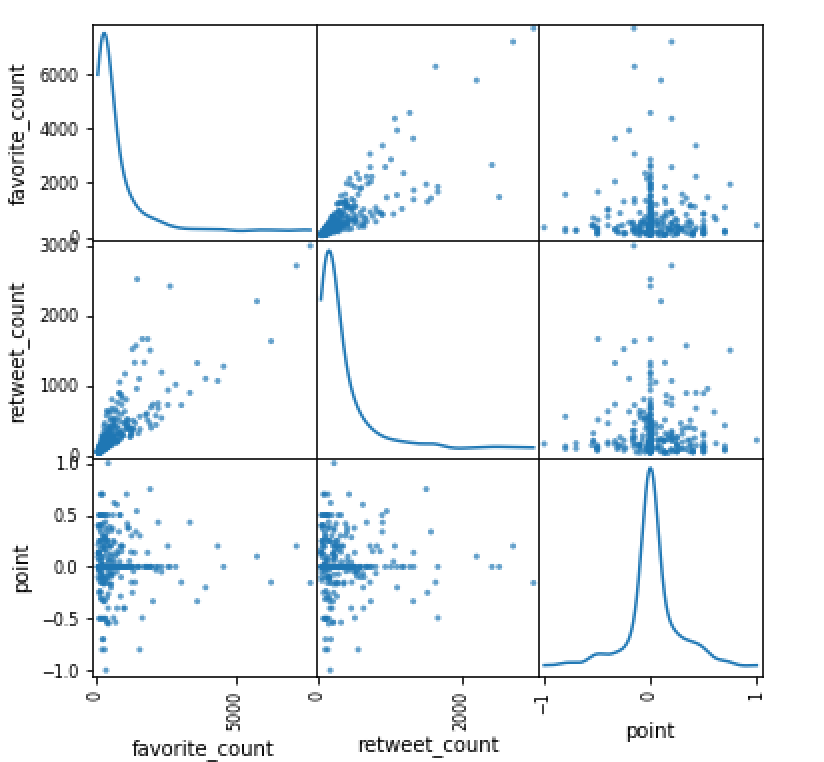
(graph 11)

**Identify three (3) quantitative variables from either data set. Find the correlation between all the pairs of these quantity variables. Include a table of the output in your report, and explain your findings – what does this indicate about your data? Use scatterplots to display the results. Ideally, create a set of scatterplot subplots.**

From the *table 12,* variable “favorite\_count” and variable “retweet\_count” are highly correlated with almost a strong positive relationship, which means that one variable increases while the other increases at the same time. On the other hand, since the correlation coefficient is -0.00083 for variable “point” and variable “favorite\_count”, therefore those two variables do not have a correlational association of any kind. Similar to other datasets, linear relationship and non-linear relationship can be observed by the presences of scatterplot’s subplots. If the shape of the scatterplot looks like a straight line, it means linear relationship exists while others are not. (*displayed in graph\_13*) In this scatterplot subplots, only variable “favorite\_count” and “retweet\_count” are correlated, and the diagonal plots show the distribution of each variable.



(table 12)



(graph 13)

**Cluster Analysis:**

In the Cluster Analysis, three cluster analyses were used to cluster the stock datasets. Three functions “KMEANS”“HIERAR”and “DBS”were defined in this project’s code to do the k-means partition clustering, hierarchical clustering and dbscan clustering respectively. Those functions will return the cluster labels and silhouette scores as a list, also printing the silhouette score and giving out a scatter plot if the parameter verbose were set True.

For the stock dataset, the price is highly depending on the time. Though the price will be fluctuant in a short term, it usually stays an increasing trend for a long term. If the stock is clustered based on price, the result will turn out with all first half stock data labeled 0 and the other labeled 1, which is meaningless. Therefore, volume, differences of open and close price, and increasing rate were used as the standards to cluster stock datasets, since those variables are less depending on time.

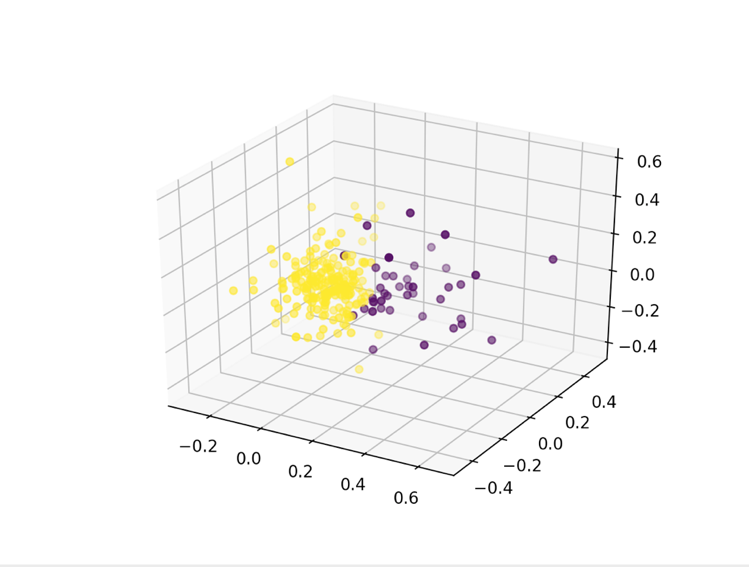
Then, the function “findbestpara” was defined to find the parameters, leading to the highest silhouette score for our 3 cluster methods. After that, the parameters were founded to use to do the three clustering for the 2 stock datasets. The program generated two csv files with 3 types of clustering labels.

From the two csv files, many fun facts can be founded: firstly, Dbscan did not do good on clustering this kind of datasets, but it could handle the noise better than other two methods. Because the results of dbscan has only one cluster, rest of points were considered as noisy points. From the generated csv file, it is obvious that the points with -1 label having at least one abnormally large or small parameters. Secondly, k-means clustering, hierarchical clustering cannot handle noise, because they cannot label noise. In most conditions, k-means and hierarchical clustering will give same point as the same labels. This may because those two clustering method are both clustering points that based on distance. In S&P 500 dataset, most points were labeled 0 and few points were labeled 1, it may because the S&P500 dataset is less fluctuant. Next, for the Nasdaq dataset, there is interesting finding that most labeled points “1” are the turning point for increasing rate, which means the increasing rate of Nasdaq started turning from increasing to decreasing or from decreasing to increasing after the 1 labeled point.

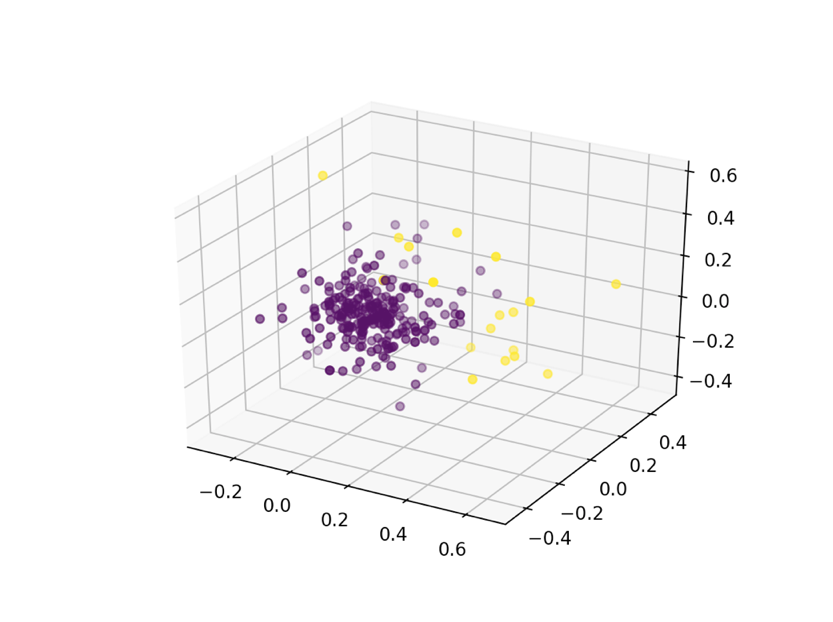
When silhouette score was used to measure the quality of the clusters, followed the results were obtained. Nasdaq: Kmeans: 0.39, Hierarchical:0.51, DBscan:0.60. S&P500: Kmeans:0.54, Hierarchical: 0.60, DDscan:0.7. The score of k-means were the lowest, because there was noise in this dataset and k-means cannot handle noise so well.

Then the scatter graphs were plotted. From the graph, it is obvious that k-means cluster and hierarchical cluster have almost the same scatter plots. DBscan plot labeled the outlier points but it only generated 1 cluster. *(displayed in graph\_14)*

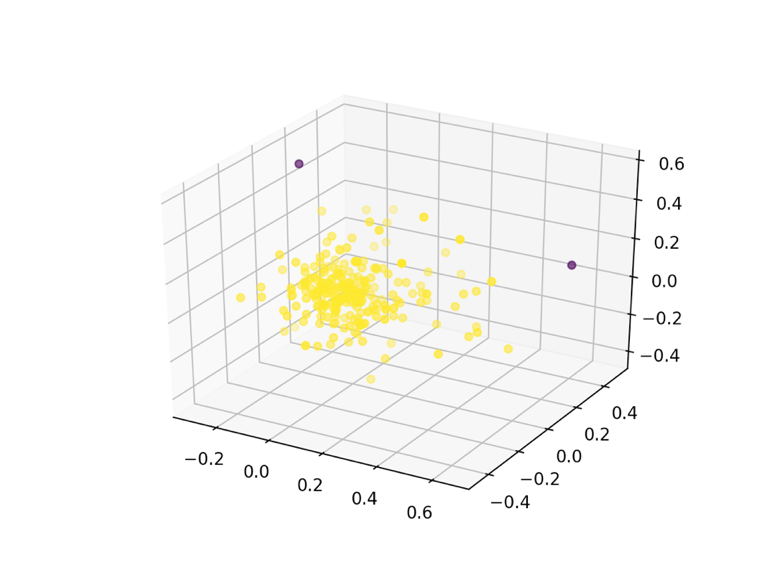
Nasdaq Kmeans



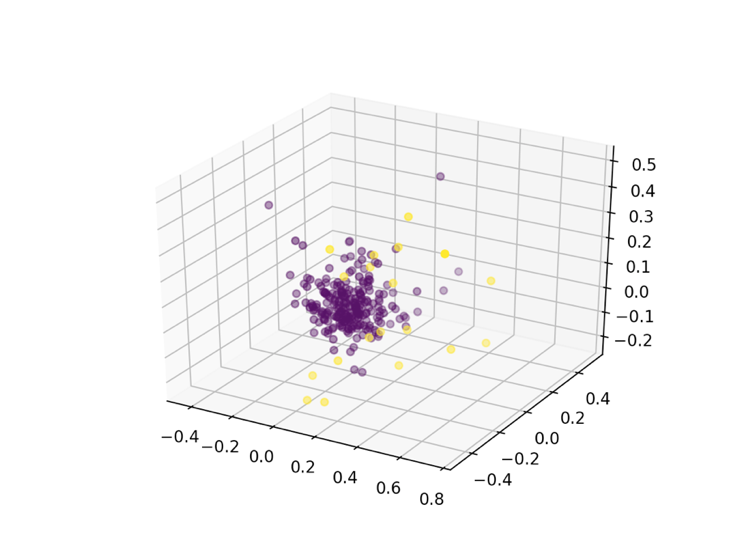
Nasdaq Hierarchical



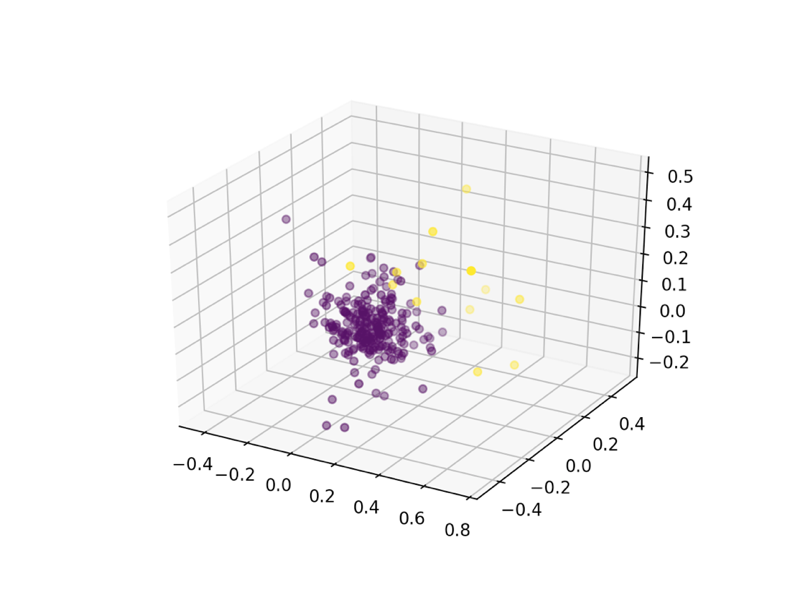
Nasdaq DBscan



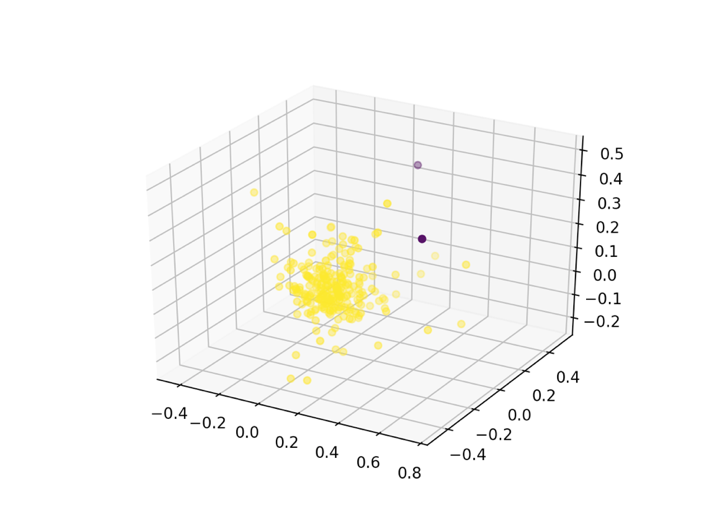
SP500 Kmeans



SP500 Hierarchical



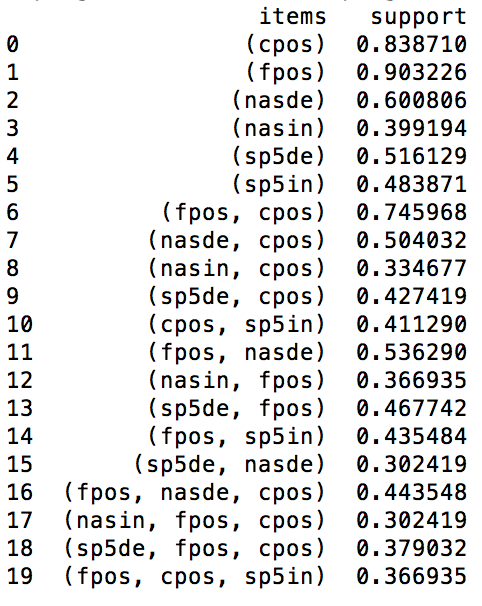
SP500 DBscan



(graph 14)

**Association Rules / Frequent Itemset Mining Analysis:**

In the Association Rules part. To find out the possible relationships between one day’s CNN, Fox news’ sentiment score and the next day’s Nasdaq, S&P500’s increasing rate, all of those four columns were combined into one data frame. Because association rule cannot handle continuous variables, so the first step of the analysis was to bin the four columns according to positive and negative values. In the data binning process, “cneg” was used to label CNN news with a negative sentiment score，“cpos” was used to label CNN news with a positive sentiment score. Similarly, “fneg” and “fpos” indicate certain day’s Fox news has a negative or positive sentiment score. “nasde” was used to label Nasdaq with a negative increasing rate, “nasin” was used to label Nasdaq with a positive increasing rate. Similarly, “sp5de”and “sp5in” indicate certain day’s S&P500 has a negative and positive increasing rate. After the associate rules analysis (minsupport=0.3), results were displayed in the graph below was generated (*displayed in graph\_15*). From these results, many interesting facts were obtained: firstly, the main trend of CNN news and Fox news were positive. Because the result showed the support of “(cpos)”is 0.84 and “(fpos)”is 0.9. Next, the support of (cpos,fpos) is 0.75, which means there was a high possibility that CNN and Fox news both reported overall positive news in the same day. Then, there were almost no effects of certain day’s overall news’ sentiment scores to next day’s Nasdaq and S&P500 increasing rate. Because row 18 and 19, 16 and 17 almost have the same support in these results, which mean the positive or the negative sentiment scores of CNN and Fox news don’t affect the increasing or decreasing of Nasdaq and S&P500.



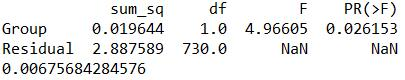
(graph\_15)

**Hypothesis Testing**:

**The first hypothesis:**

Since the goal of this project is to detect the relationship between the price of the stock market and News’ headlines, then figuring out any possible variables or changes that could influence each other are very important. The News’ headlines that have been collected come from two different resources, which are CNN and FoxNews. So, the first hypothesis of this project is CNN and FoxNews do not have significant difference between TextBlob points.

This project uses a One-Way ANOVA to test this hypothesis since One-Way ANOVA can directly compare the differences of sample means in different groups. The null hypothesis will be rejected if the difference between the sample means is too large or too small. The results can be found by setting the significance level. In this case, the significance level has been set as 0.05, and the p-value has been calculated to be 0.026153 through the code*. (displayed in table 16)*



(table\_16)

And since the p-value is less than the significance level, the null hypothesis has been rejected. In other words, CNN and FoxNews have significant difference between TextBlob points, which shows that they have significant different opinions and attitudes to the news and events that happened last year.

**The second hypothesis:**

Since the news dataset came from Tweet, it contained some other interesting variables such as favorites and retweets. Although the main measurement in this project for the tweets is the TextBlob points, it can be interesting to find out any other factors that may have linear relationship with the TextBlob points. So, the second hypothesis in this project is that TextBlob point does not have linear relationship with variable “favorites” and variable “retweets”.

TextBlob points directly show the positive attitudes or opinions if the points are positive, and show the negative attitudes or opinions if the points are negative. In either way, higher absolute values of the points shows the stronger attitudes and opinions. So, the data has been cleaned to show the absolute values of the points, which is proper for this hypothesis test.

For CNN news, set:

Null Hypothesis: The TextBlob points do not have linear relationship with favorites and retweets. (Slope = 0 )

Alternative Hypothesis: The TextBlob points do have linear relationship with favorites and retweets. (Slope is not equal to 0)

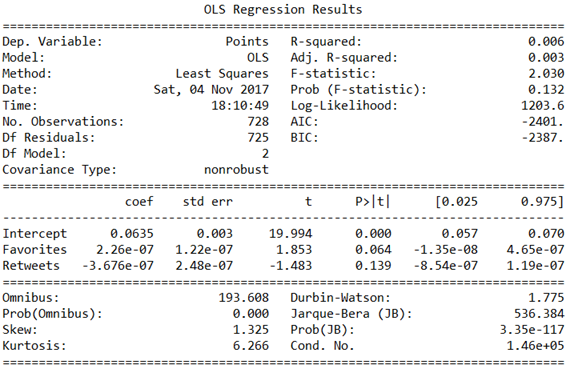
For Fox News, set:

Null Hypothesis: The TextBlob points do not have linear relationship with favorites and retweets. (Slope = 0 )

Alternative Hypothesis: The TextBlob points do have linear relationship with favorites and retweets. (Slope is not equal to 0)

After running through the model and code, the results have been showed below: *. (displayed in table 17 and 18)*

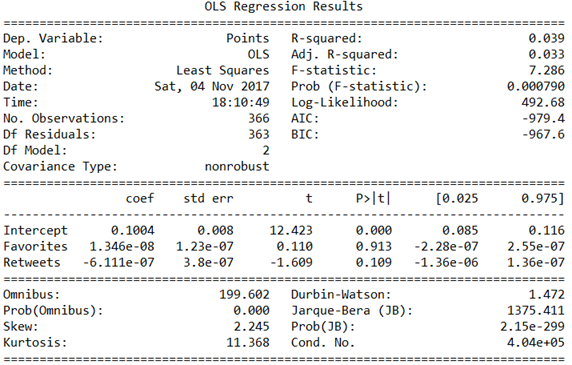
For the CNN news:



(table\_17)

Since the p-value is 0.132, which is larger than 0.05, it fails to reject the null hypothesis. So, the TextBlob points for CNN do not have linear relationship with favorites and retweets.

For the Fox News:



(table\_18)

Since the p-value is 0.000790, which is less than 0.05, it rejects the null hypothesis. So, the TextBlob points for Fox News have linear relationship with favorites and retweets.

**The third hypothesis:**

The third hypothesis of this project is that the headlines of CNN and FoxNews have influences on the changes of the stock market and gold prices. The project is assuming that CNN and FoxNews could predict the increasing or the decreasing situations through the TextBlob points of news headlines. In this way, the project used five supervised predictive models, which are Decision Tree, KNN, Naïve Bayes, SVM and Random Forest, to test the accuracy of the prediction. These predictive models were chosen for this hypothesis because these models can use the statistics and data provided to predict the outcomes. In other words, these models can directly test the necessary relationship between the variables and final results. If there is necessary relationship, predictive models can get high accuracy score and the results can be used to predict the future.

KNN means the K-nearest Neighbors, which uses the algorithm to calculate the measurement like distances in order to find the closest training samples to do the classification.

Decision Tree is the model that contains nodes, which denote several tests to lead to different branches. Each node represents a class label. And after building the tree through the training, it can predict the results of classification by going through the nodes(tests).

Gaussian Naive Bayes is the model that uses Bayesian theorem. The NB model is very useful if the dimensionality is high, and it will assume that the variables are independent. NB uses a family of algorithms to consider each of features that contribute to the final classification result.

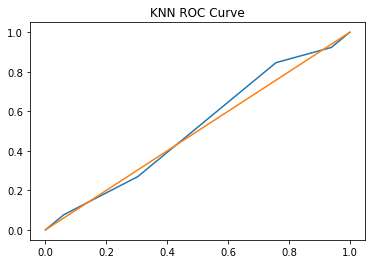
Support vector machines are a set of supervised learning models that uses algorithms to predict the result of classification, regression and even outlier detection.

RandomForest is the model that build several trees after learning the training data to generate different classifications. It is like a forest that contains several decision trees, and for each votes on the classification and class with the most votes wins and generate the result.

For the data set that was being used to do the test, this project combined the TextBlob column in the datasets of CNN and FoxNews. Meanwhile, the project cleaned the data of Nasdaq, Gold and S&P 500 to add a new column that shows the increased rate of adjusted close value. Since the prediction models need the classification of the data for the outcome, the increase rate has been simply changed into “increase” if the number is positive or “decrease” if the number is negative. The increase rate columns which come from Nasdaq, Gold and S&P 500 were combined with the TextBlob points from CNN and FoxNews to prepare the three datasets for the testing. Below are the results after running the program for each dataset:

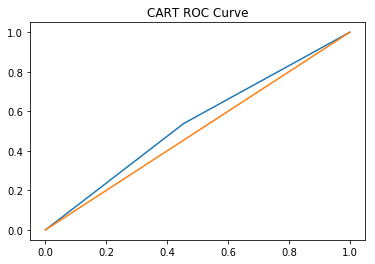
For Nasdaq:

KNN has the prediction accuracy of 0.50847, and confusion matrix is [[15 11] [18 15]], which means KKN has the accuracy of 15/(15+11) = 0.58 to predict a “decrease” to be “decrease”, and has the accuracy of 11/(15+11) = 0.42 to predict a “increase” to be “decrease”. Also, KKN has has the accuracy of 18/(18+15) = 0.55 to predict a “decrease” to be “increase”, and has the accuracy of 15/(18+15) = 0.45 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



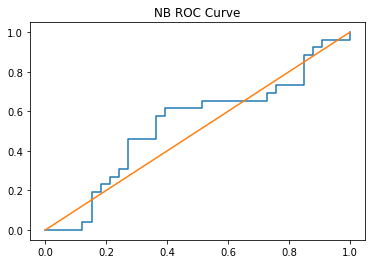
Decision Tree has the prediction accuracy of 0.47458, and confusion matrix is [[14 12] [19 14]]

, which means CART has the accuracy of 14/(14+12) = 0.54 to predict a “decrease” to be “decrease”, and has the accuracy of 12/(14+12) = 0.46 to predict a “increase” to be “decrease”. Also, CART has has the accuracy of 19/(19+14) = 0.58 to predict a “decrease” to be “increase”, and has the accuracy of 14/(19+14) = 0.42 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



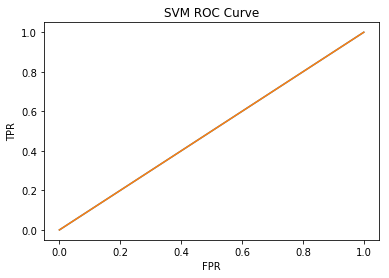
Naïve Bayes has the prediction accuracy of 0.45763, and confusion matrix is [[25 1] [31 2]]

, which means NB has the accuracy of 25/(25+1) = 0.96 to predict a “decrease” to be “decrease”, and has the accuracy of 1/(25+1) = 0.04 to predict a “increase” to be “decrease”. Also, NB has has the accuracy of 31/(31+2) = 0.94 to predict a “decrease” to be “increase”, and has the accuracy of 2/(31+2) = 0.06 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



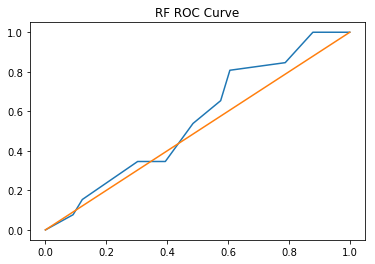
SVM has the prediction accuracy of 0.44068, and confusion matrix is [[26 0][33 0]]

, which means SVM has the accuracy of 26/(26+0) = 1 to predict a “decrease” to be “decrease”, and has the accuracy of (0/26+0) = 0 to predict a “increase” to be “decrease”. Also, SVM has has the accuracy of 33/(33+0) = 1 to predict a “decrease” to be “increase”, and has the accuracy of 0/(33+0) = 0 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



Random Forest has the prediction accuracy of 0.50848, and confusion matrix is [[15 11][18 15]]

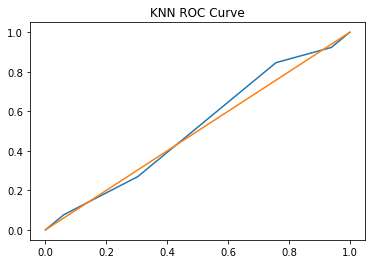
, which means RF has the accuracy of 15/(15+11) = 0.58 to predict a “decrease” to be “decrease”, and has the accuracy of (11/15+11) = 0.42 to predict a “increase” to be “decrease”. Also, RF has has the accuracy of 18/(18+15) = 0.55 to predict a “decrease” to be “increase”, and has the accuracy of 15/(18+15) = 0.45 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



In general, the average prediction accuracy is 0.477968, which is really close to 0.5. And for each method, the accuracy rate is close to 0.5. They do not have significant accuracy rate to classify the increase or decrease situation base on the CNN and FoxNews. Therefore, the CNN and FoxNews cannot predict the changes of Nasdaq.

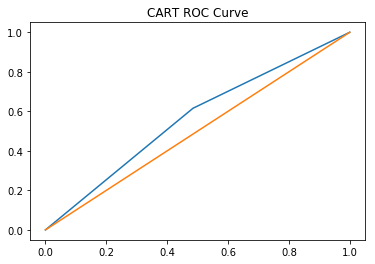
For Gold:

KNN has the prediction accuracy of 0.50847, and confusion matrix is [[15 11] [18 15]], which means KKN has the accuracy of 15/(15+11) = 0.58 to predict a “decrease” to be “decrease”, and has the accuracy of 11/(15+11) = 0.42 to predict a “increase” to be “decrease”. Also, KKN has has the accuracy of 18/(18+15) = 0.55 to predict a “decrease” to be “increase”, and has the accuracy of 15/(18+15) = 0.45 to predict a “increase” to be “increase”.Below is the ROC curve that generated from python:



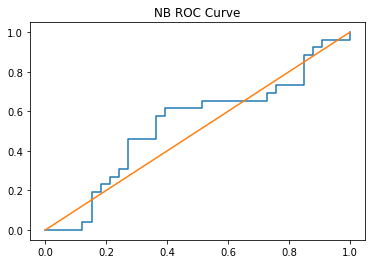
Decision Tree has the prediction accuracy of 0.47458, and confusion matrix is [[14 12] [20 13]]

, which means CART has the accuracy of 14/(14+12) = 0.54 to predict a “decrease” to be “decrease”, and has the accuracy of 12/(14+12) = 0.46 to predict a “increase” to be “decrease”. Also, CART has has the accuracy of 20/(20+13) = 0.61 to predict a “decrease” to be “increase”, and has the accuracy of 13/(20+13) = 0.39 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



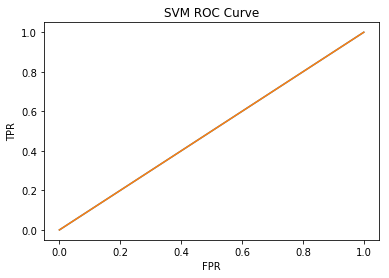
Naïve Bayes has the prediction accuracy of 0.45763, and confusion matrix is [[25 1] [31 2]]

, which means NB has the accuracy of 25/(25+1) = 0.96 to predict a “decrease” to be “decrease”, and has the accuracy of 1/(25+1) = 0.04 to predict a “increase” to be “decrease”. Also, NB has has the accuracy of 31/(31+2) = 0.94 to predict a “decrease” to be “increase”, and has the accuracy of 2/(31+2) = 0.06 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



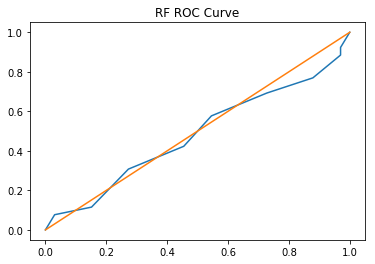
SVM has the prediction accuracy of 0.44068, and confusion matrix is [[26 0][33 0]]

, which means SVM has the accuracy of 26/(26+0) = 1 to predict a “decrease” to be “decrease”, and has the accuracy of (0/26+0) = 0 to predict a “increase” to be “decrease”. Also, SVM has has the accuracy of 33/(33+0) = 1 to predict a “decrease” to be “increase”, and has the accuracy of 0/(33+0) = 0 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



Random Forest has the prediction accuracy of 0.52542, and confusion matrix is [[19 7][21 12]]

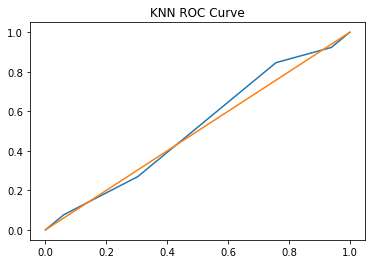
, which means RF has the accuracy of 19/(19+7) = 0.73 to predict a “decrease” to be “decrease”, and has the accuracy of (7/19+7) = 0.27 to predict a “increase” to be “decrease”. Also, RF has has the accuracy of 21/(21+12) = 0.64 to predict a “decrease” to be “increase”, and has the accuracy of 12/(21+12) = 0.36 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



In general, the average prediction accuracy is 0.481356, which is really close to 0.5. And for each method, the accuracy rate is close to 0.5. They do not have significant accuracy rate to classify the increase or decrease situation base on the CNN and FoxNews. Therefore, the CNN and FoxNews cannot predict the changes of Gold.

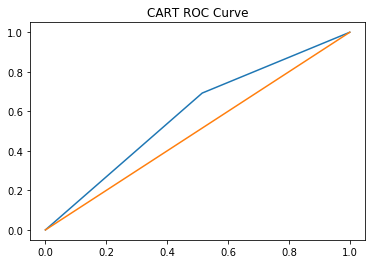
For S&P 500,

KNN has the prediction accuracy of 0.50847, and confusion matrix is [[15 11] [18 15]], which means KKN has the accuracy of 15/(15+11) = 0.58 to predict a “decrease” to be “decrease”, and has the accuracy of 11/(15+11) = 0.42 to predict a “increase” to be “decrease”. Also, KKN has has the accuracy of 18/(18+15) = 0.55 to predict a “decrease” to be “increase”, and has the accuracy of 15/(18+15) = 0.45 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



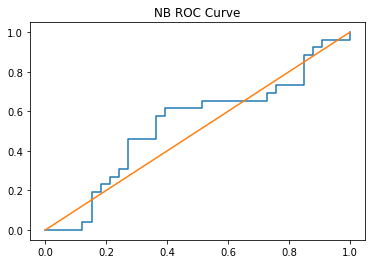
Decision Tree has the prediction accuracy of 0.44068, and confusion matrix is [[13 13] [20 13]]

, which means CART has the accuracy of 13/(13+13) = 0.5 to predict a “decrease” to be “decrease”, and has the accuracy of 13/(13+13) = 0.5 to predict a “increase” to be “decrease”. Also, CART has has the accuracy of 20/(20+13) = 0.61 to predict a “decrease” to be “increase”, and has the accuracy of 13/(20+13) = 0.39 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



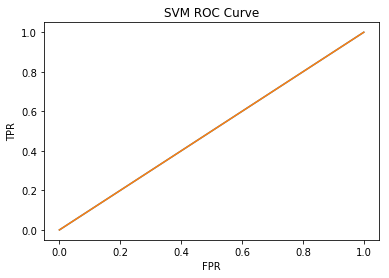
Naïve Bayes has the prediction accuracy of 0.45763, and confusion matrix is [[25 1] [31 2]]

, which means NB has the accuracy of 25/(25+1) = 0.96 to predict a “decrease” to be “decrease”, and has the accuracy of 1/(25+1) = 0.04 to predict a “increase” to be “decrease”. Also, NB has has the accuracy of 31/(31+2) = 0.94 to predict a “decrease” to be “increase”, and has the accuracy of 2/(31+2) = 0.06 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



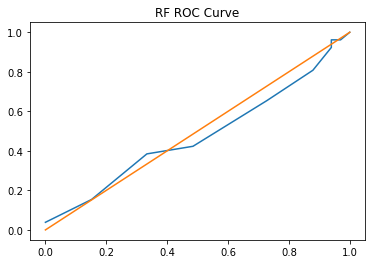
SVM has the prediction accuracy of 0.44068, and confusion matrix is [[26 0][33 0]]

, which means SVM has the accuracy of 26/(26+0) = 1 to predict a “decrease” to be “decrease”, and has the accuracy of (0/26+0) = 0 to predict a “increase” to be “decrease”. Also, SVM has has the accuracy of 33/(33+0) = 1 to predict a “decrease” to be “increase”, and has the accuracy of 0/(33+0) = 0 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



Random Forest has the prediction accuracy of 0.47458, and confusion matrix is [[17 9][22 11]]

, which means RF has the accuracy of 17/(17+9) = 0.65 to predict a “decrease” to be “decrease”, and has the accuracy of (9/17+9) = 0.35 to predict a “increase” to be “decrease”. Also, RF has has the accuracy of 22/(22+11) = 0.66 to predict a “decrease” to be “increase”, and has the accuracy of 11/(22+11) = 0.33 to predict a “increase” to be “increase”. Below is the ROC curve that generated from python:



In general, the average prediction accuracy is 0.464408. And for each method, the accuracy rate is close to 0.5. They do not have significant accuracy rate to classify the increase or decrease situation base on the CNN and FoxNews. Therefore, the CNN and FoxNews cannot predict the changes of S&P 500.

**Conclusions:**

The purpose of this project is to find out the relation between the attitude of news and the increasing rate of stock index. CNN and Fox news were chosen as the news resource, TextBlob sentiment score was used to measure the attitude of news. Nasdaq and S&P500 were chosen as the stock index in this project. Before performing the predictive analysis, most of the variables of the 5 datasets were tested to show its distribution, it turns out that most of them are normally distributed. Then, the next logical step was to test the correlation relationships for the frequent variables that used quite a lot in this analysis, it appears that they tend to have very little dependencies to each other. During this predictive analysis, this project did find out that there was a significant difference of attitudes and opinions between CNN and Fox news . Also, the number of favorites and retweets for the tweets they posted each day have different influences. There is linear relationship between the attitudes and opinions of Fox news and favorites and retweets, and CNN does not show obvious linear relationship. After using predictive models to test the relationship between the attitude of news and increasing rate of gold, Nasdaq and S&P500 directly, neither model could give good predict results in gold, Nasdaq and S&P500.