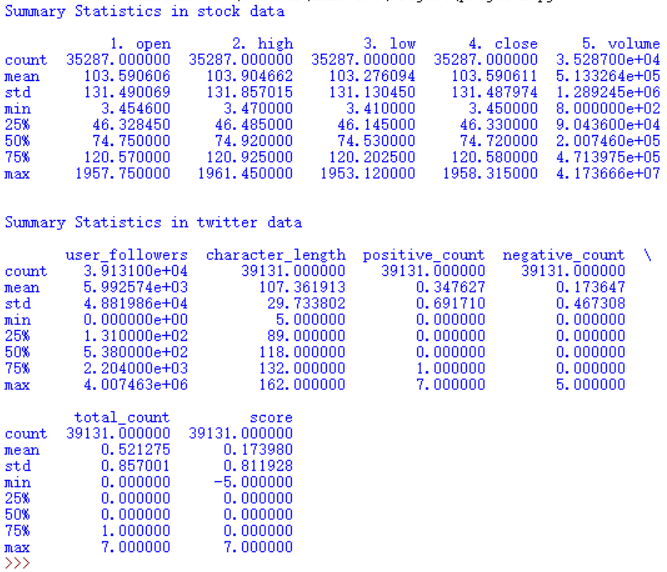
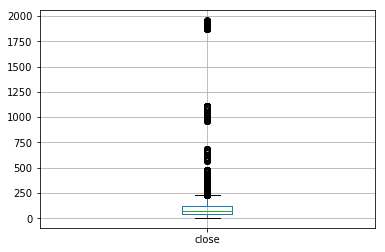
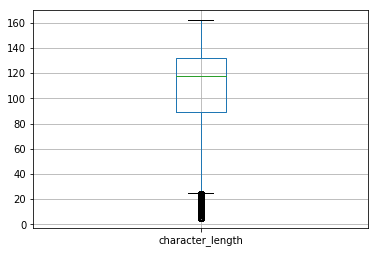
**The Impact of Social Media on Stock Prices**

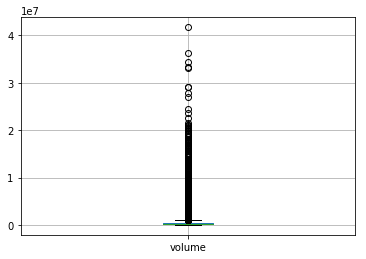
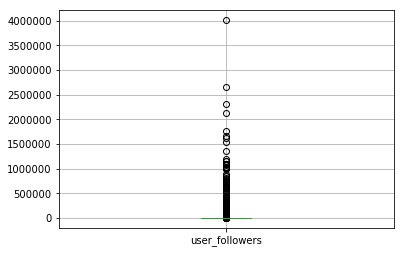
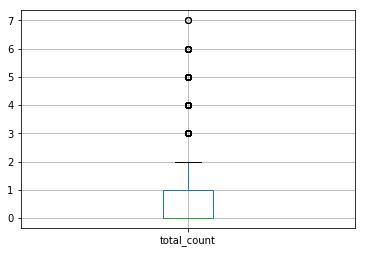
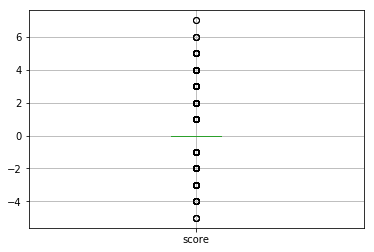
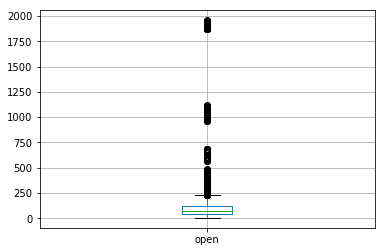
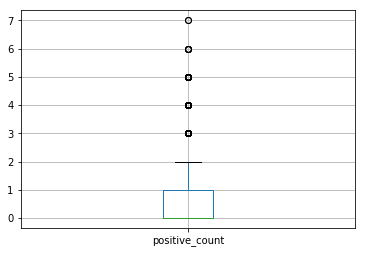
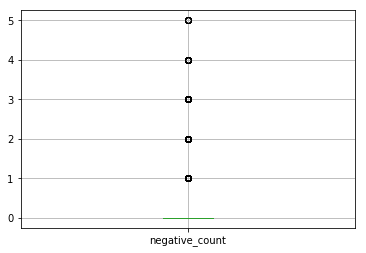
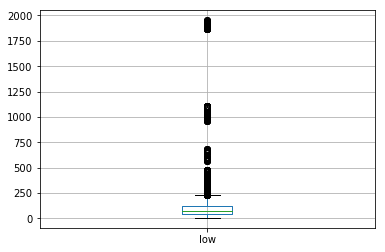
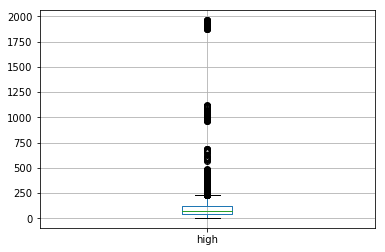
**Exploratory analysis:**

**Basic Statistical Analysis and data cleaning insight:**

To have a basic understanding of the collected data and the distribution of each quantitative attribute, it is important to start with a summary statistic for the quantitative attributes. The following screenshots are the statistical summaries of 11 quantitative attributes from stock and Twitter data.

****

****

****

After having the basic understanding of quantitative attributes, this project moves to the next step: data cleaning. Data cleaning is a crucial task which cannot be ignored, especially for the notoriously noisy Twitter data. During project 1, part of the data cleaning had been finished already. In this part of the project, data cleaning and aggregation will be continued.

The attributes from the Twitter data that may contain outliers are user\_id and user\_followers. From the stock data, possible outliers may be in open, high, low, close, and volume.

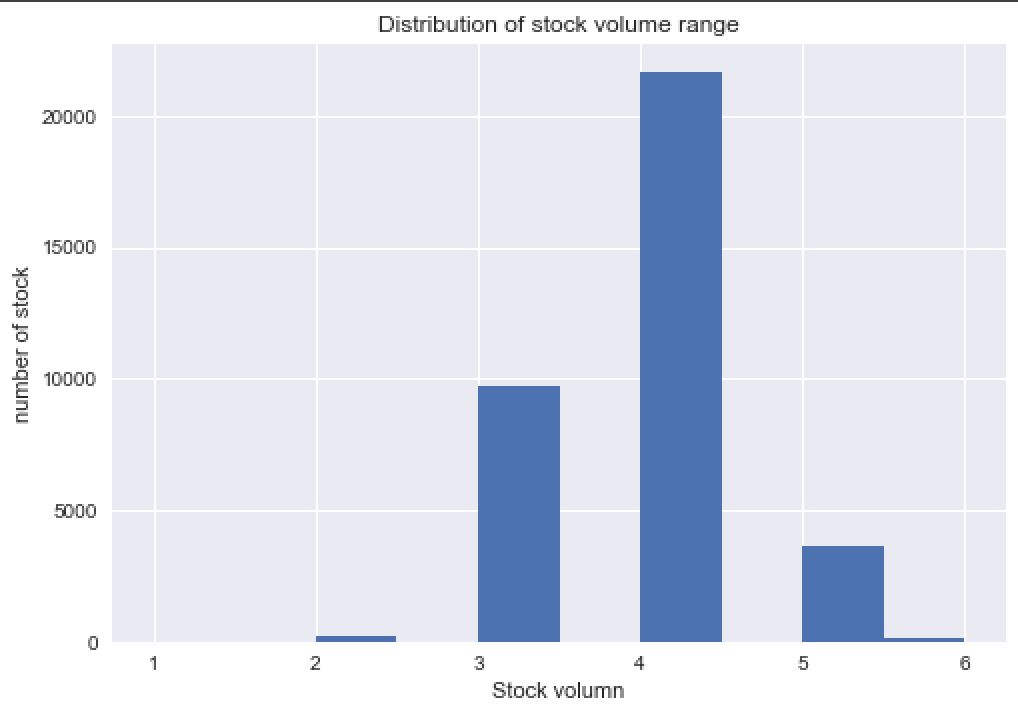
Since the projects focus on the relationship between Twitter content and stock price, “text” is the attribute that directly influences the conclusion. And so only the attribute “text” will be cleaned in this project. The method that is used to detect outliers is visualizing outliers with boxplots, but since “text” is not a quantitative attribute, it is impossible to develop a boxplot. Therefore another attribute “character\_count” is created to indicate the length of the Twitter text. The “text” with small values (less than 10) of “character\_count” is considered invalid since it has a high probability of not delivering mood with such little content. Before counting the words in “text,” the code removed all punctuations and emojis to perform the task more accurately.

After removing the invalid data, the twitter data is still not “pure clean” because missing values and noisy values still exist in other attributes such as “user\_name,” “user\_location” and “user\_description.” Since these attributes are not related to the analysis, the missing values will be left with “NAN.”

To simplify the hypothesis test in the following steps, it is a good idea to bin some of the quantitative attributes from each dataset. For the Twitter dataset, the quantitative attribute “user\_follower” will be binned into 5 bins: 0 to 1,000, 1,001 to 10,000, 10,001 to 100,000, 100,001 to 1,000,000, 1,000,001 to 10,000,000. Binning strategies are used in two of the attributes in stock dataset: “close” and “volume.” “close” is binned into 4 bins: 0 to 10, 11 to 100, 101 to 1,000, 1,001 to 10,000. “volume” is binned into 6 bins: 0 to 1000; 1,001 to 10,000; 10,001 to 100,000; 100,001 to 1,000,000; 1,000,001 to 10,000,000; and 10,000,001 to 100,000,000.

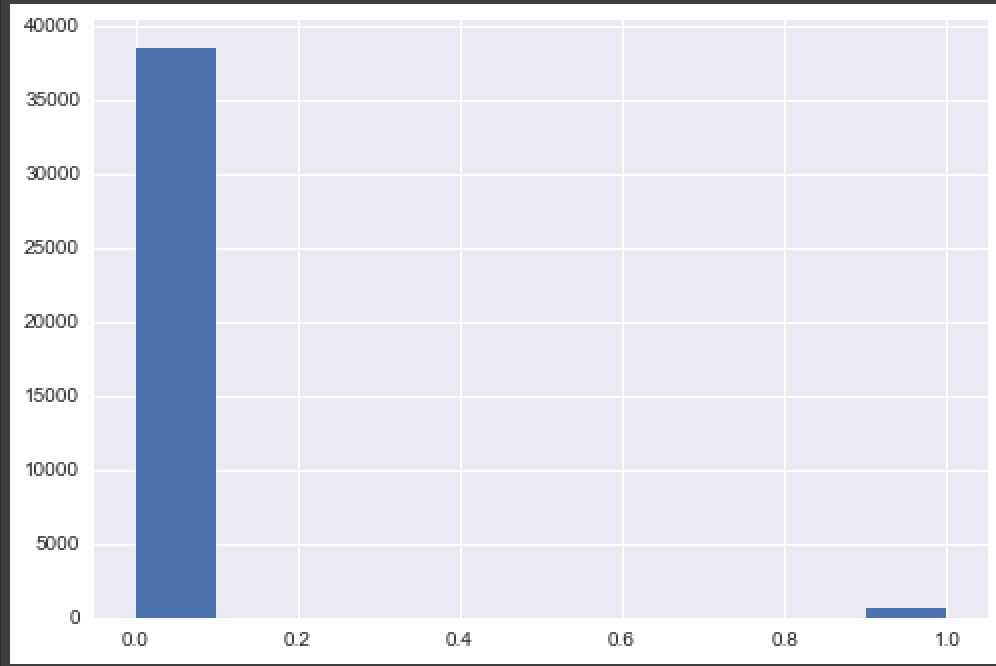
**Histograms and Correlations:**

In order to plot the histogram, three variables were picked: followers, verified users and binning stock prices. Figure A shows the distribution of stock volume and its amount. To be specific, there is no stock under the volume 1000. The volume from 100,001 to 1,000,000 reaches the largest amount, which is over 20,000



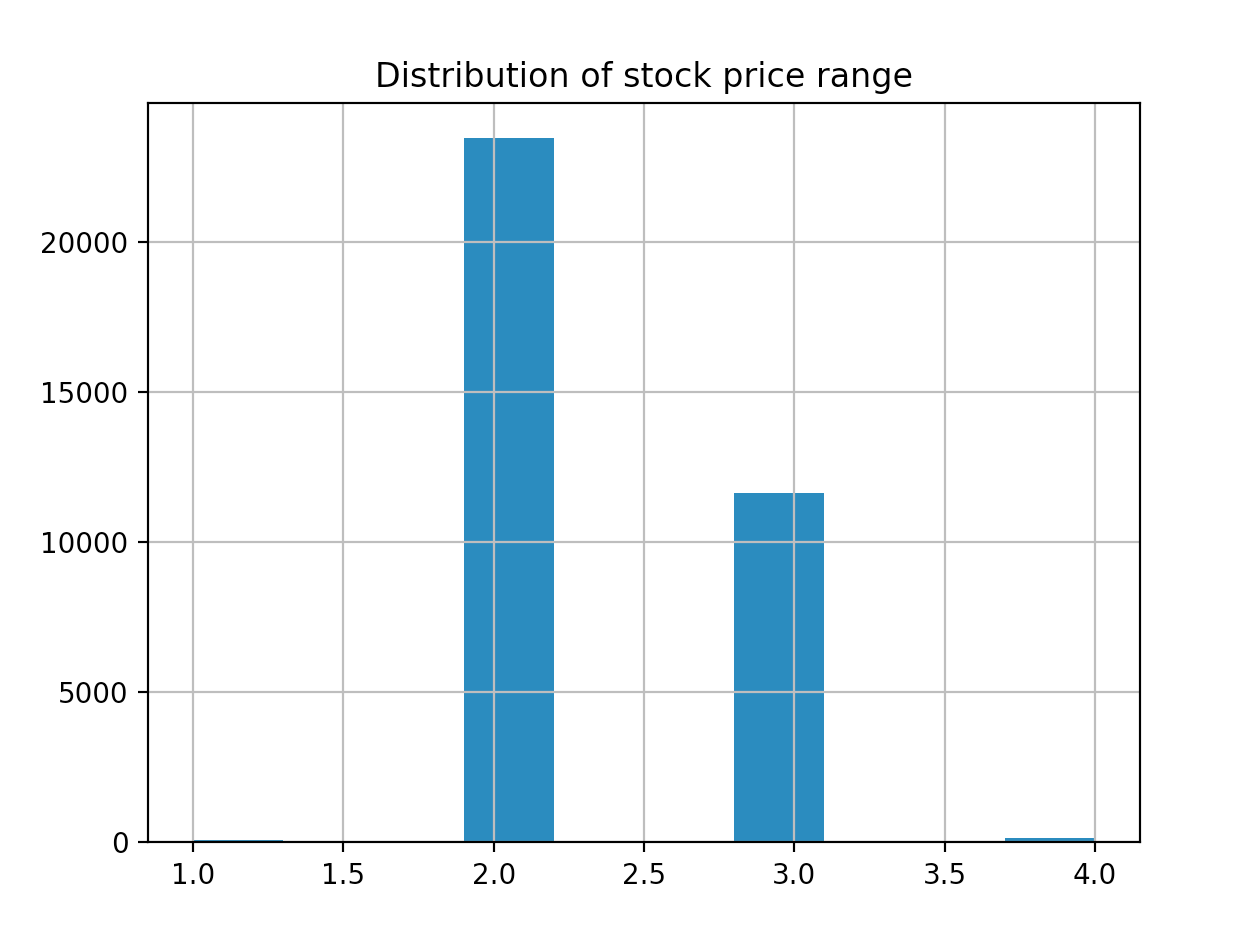
**Figure A**

Figure B shows the distribution of non-verified users and verified users. The x-axis represents the status of this Twitter user (non-verified or verified), while the y-axis is the number of Twitter users. It is easy to conclude that the number of non-verified users is far greater than the number of verified users. The difference between non-verified users and verified users is about 30,000.



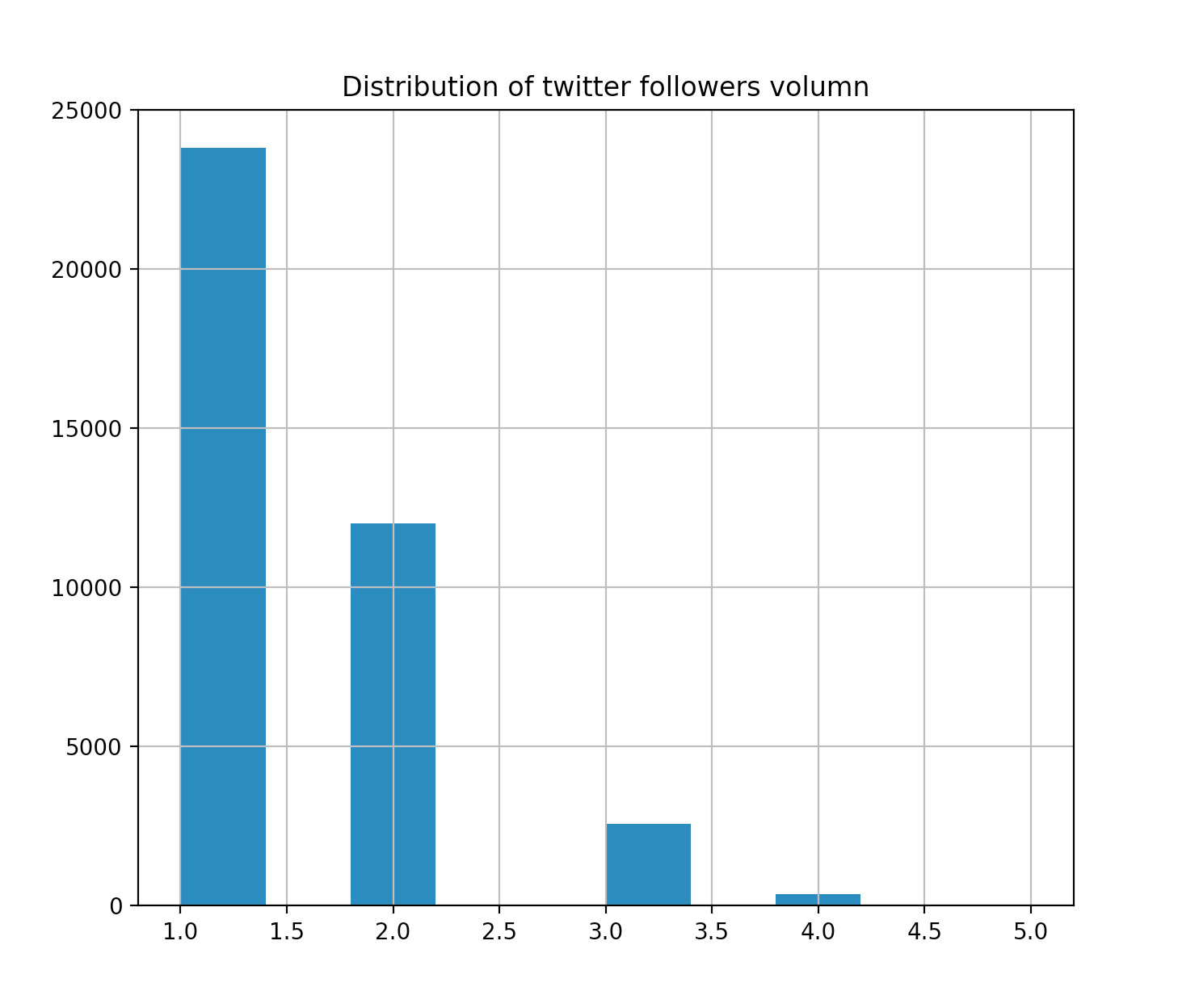
**Figure B**: Distribution of non-verified users and verified users

Figure C shows the distribution of the stock price range. The price range under 1,000, and from 100,001 to 1,000,000 have very low amounts. The majority of stock prices is concentrated on the scale from 1,001 to 10,000 dollars, which over 20,000 volumes.



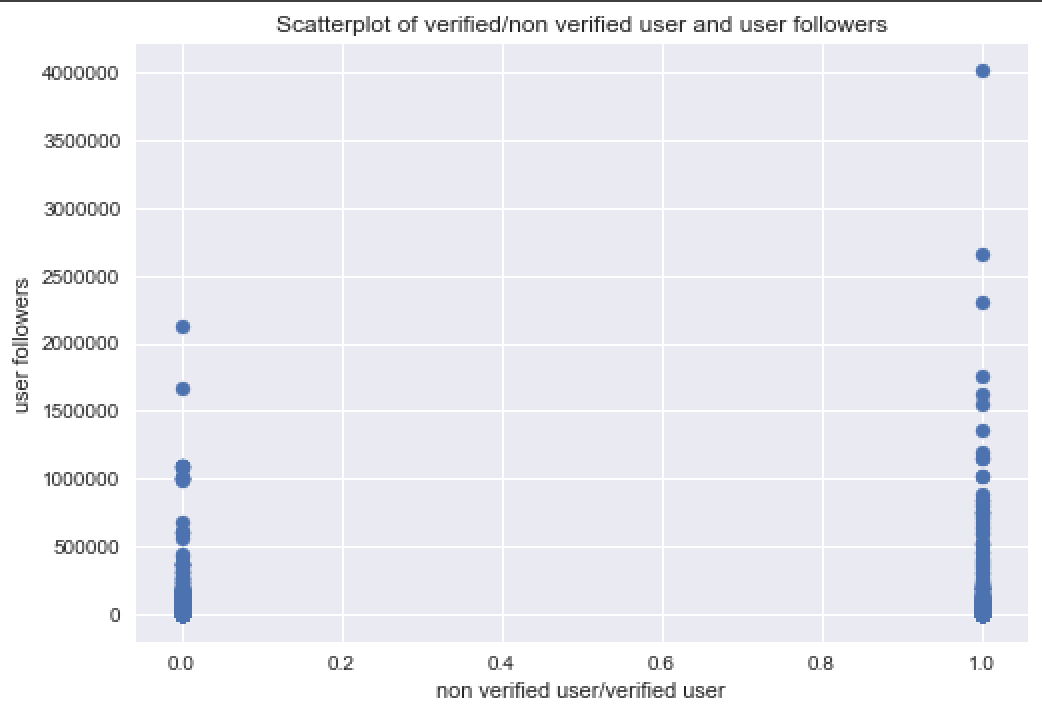
**Figure C**

Figure D shows the distribution of Twitter followers volume. Twitter users with followers from 0 to 1000 has over 20,000 users, number of users with followers from 1001 to 10,000 is about 12,000 , users’ followers from 10,001 to 100,000 is about 3000, and the number of Twitter users’ followers from 100,001 to 1,000,000 is in the hundreds.

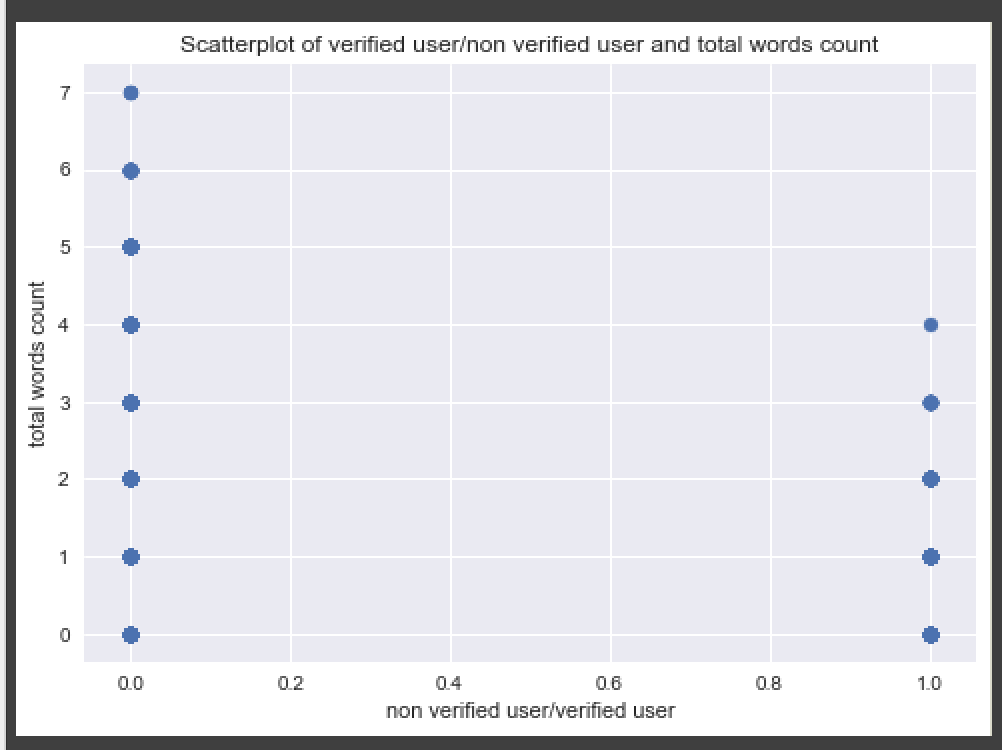


**Figure D**

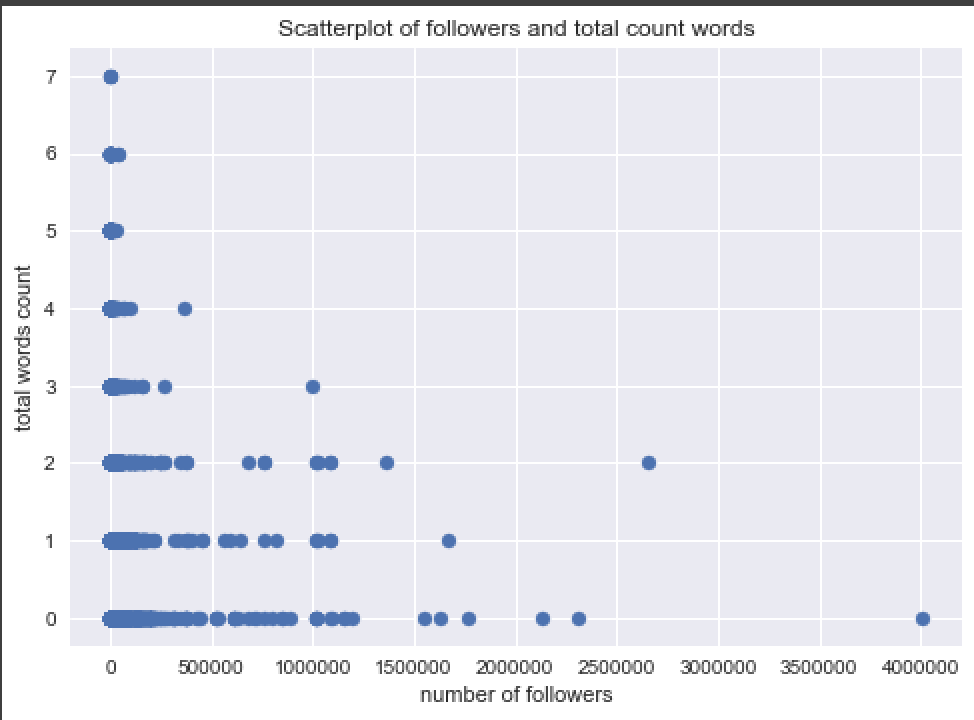
In the Twitter dataset, three variables are selected: user followers, verified users, and total words count. Then it obtains scatter plots as below:

****

**Figure 1**

****

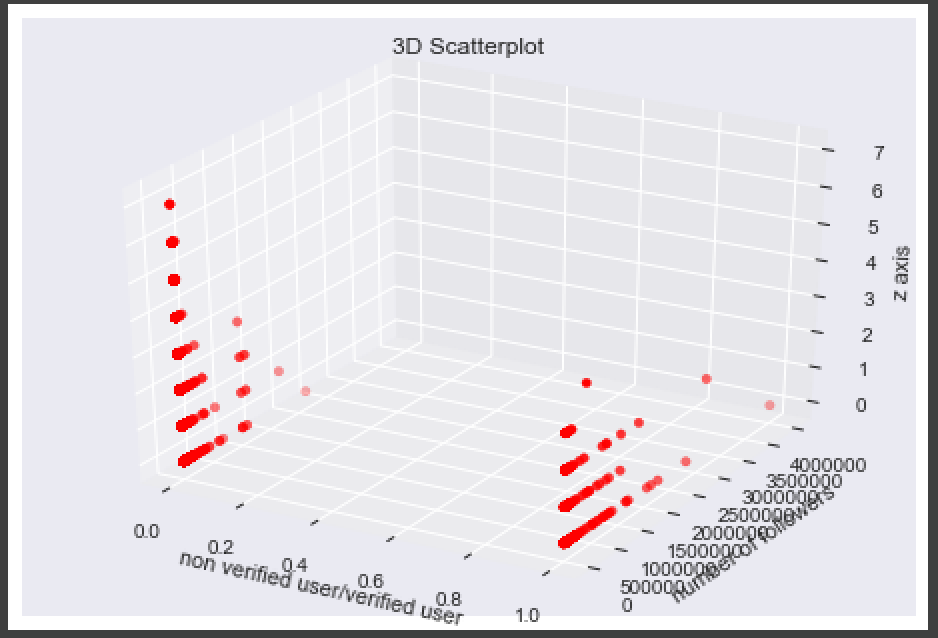
**Figure 2**



**Figure 3**

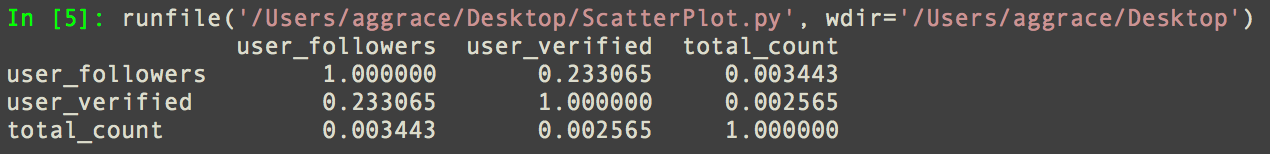
Figure 1 shows the scatter plot between verified or non-verified users and the number of followers. It is easy to see that the number of followers of verified Twitter user is greater than the number of non-verified followers. Figure 2 is the scatter plot of verified users and total word count. This plot shows that non-verified Twitter users are more likely to express their sentiment on Twitter than are verified users. Here, the total word count is the sum of positive and negative attitude words in a Tweet. Figure 3 shows the correlation between the number of followers and total word count. It shows the trend that as the number of followers increased, the total words count decreased.

Figure 4 shows a scatterplot among three variables in 3-dimension: it is easy to see that the verified Twitter users have more followers than the non-verified Twitter users, however, their willingness to express public sentiment is less than that of the non-verified Twitter users. This phenomenon is easy to understand since users with lots of followers and verified accounts would want to be more concise with what they publish.



**Figure 4**

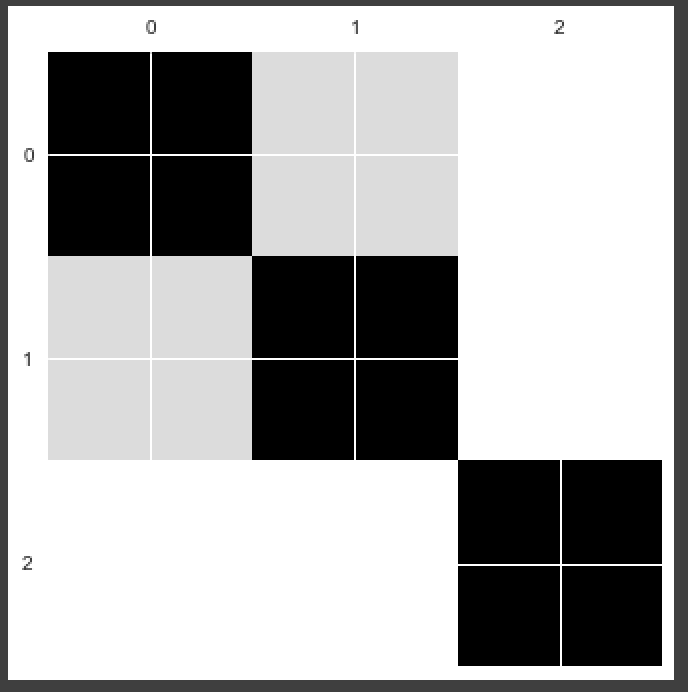
To be more accurate about the correlation among these three variables, the correlation matrix is shown below.



**Figure 5**

It is clear that the correlation between followers and verified users (0.233065), total word count and user followers (0.003443), total word count and verified users (0.002565) are all positively correlated. The most correlated variables are followers and verified users, which is 0.233065, and the least correlated variables are total count and verified users, at 0.002565.

There is also a correlation visualization as below:



**Figure 6**

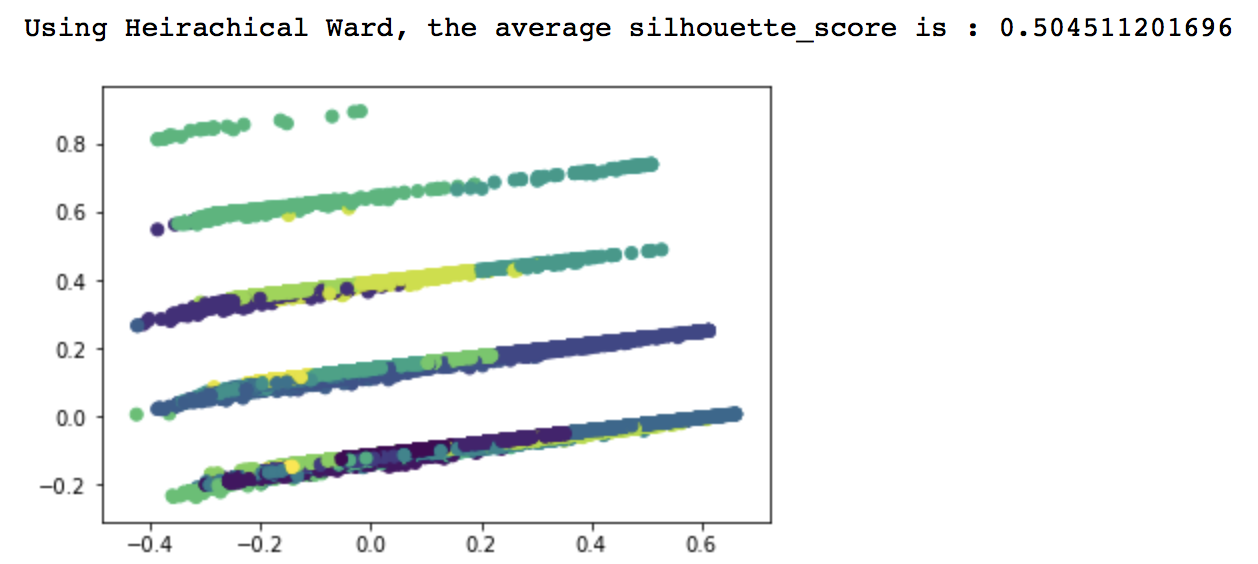
The different shades show the different correlation among variables. The darker the color, the closer the correlation is to 1. In white areas, the correlation approximates to 0.

**Cluster Analysis:**

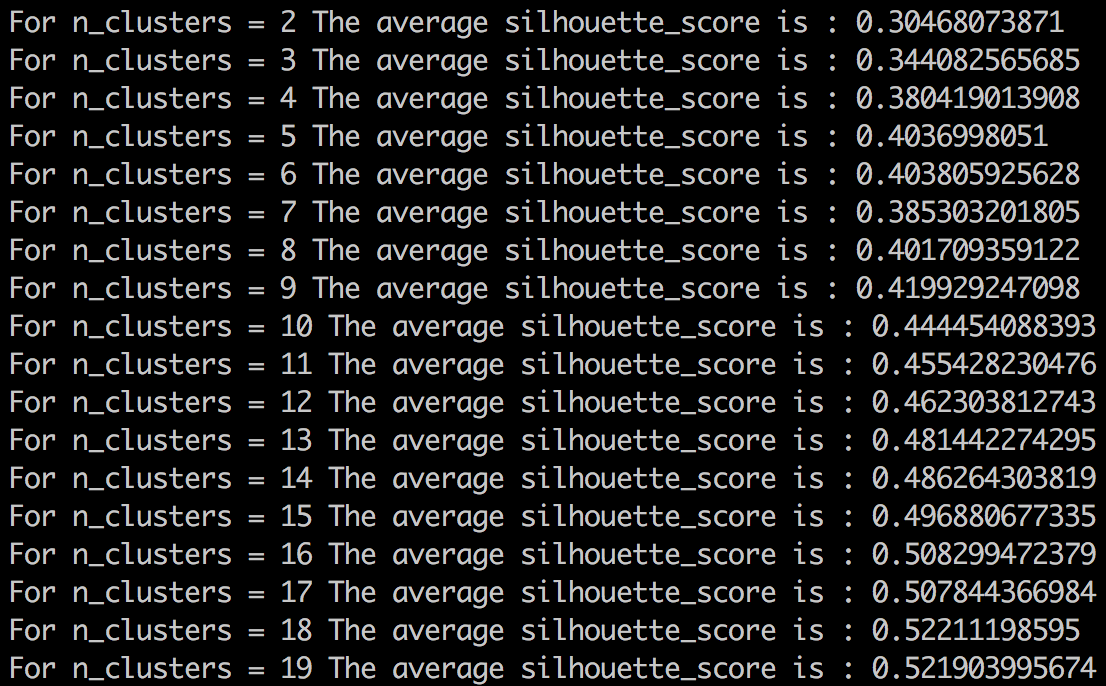
Clustering is an important process in data analysis because it reveals the hidden subsets. This process could help people discover patterns and develop a better understanding of the collected data. During cluster analysis, three different clustering methods will be applied to the data: hierarchical clustering (using Ward, which means Euclidean distance), k-means, and dbscan clustering.

Hierarchical clustering(Ward):

Using Hierarchical clustering, it is possible to have as many clusters as the programmer want because it is in the form of a dendrogram. The PCA result gives a comparatively clear visualization of the whole dataset, and the silhouette score is very high. The PCA projection is shown below:

****

In the Twitter users information data, it works the nicely since the data size is big, which is an advantage of Heirachical clustering. The plot is also showing five stripes of scatterplots, which is indicating that there is approximately five subgroups. However, the Hierarchical clustering method does not give 5 clusters.

K-mean clustering: when k = 18, k-means clustering reaches the highest Silhouette score. 

**Fig 7.** The silhouette score for different number of clusters from 2 to 19.

However, PCA cluster projection gives an outcome beyond expectation. The scatterplot shows that there are roughly 5 subgroups instead of 18.

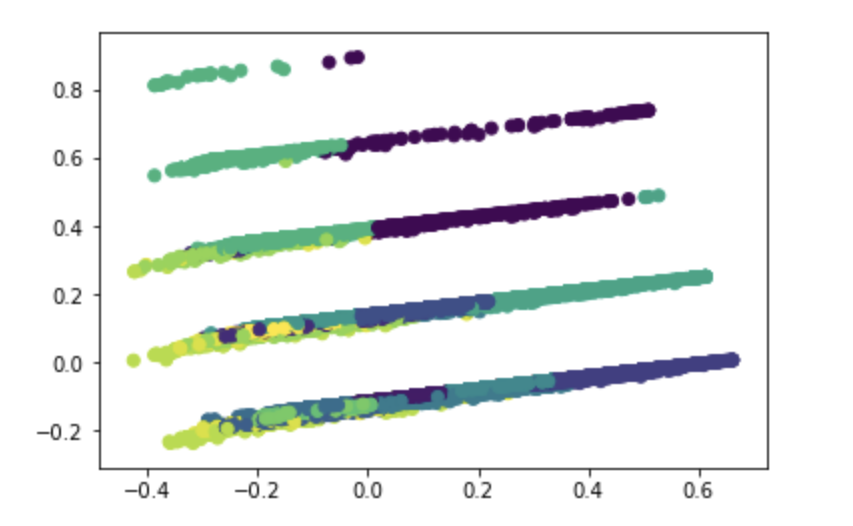


Fig 8.PCA cluster projection when cluster number k=18.

Since the visualization indicates the number of clusters could be 5, it is worth trying to plug in k=5 and project the data again. However, when k=5 is applied to the PCA cluster projection, the colored clustering still does not provide a satisfying result.

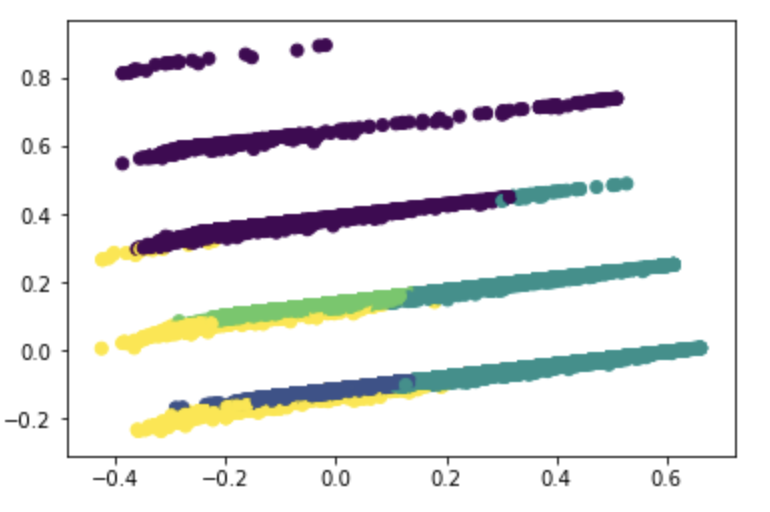
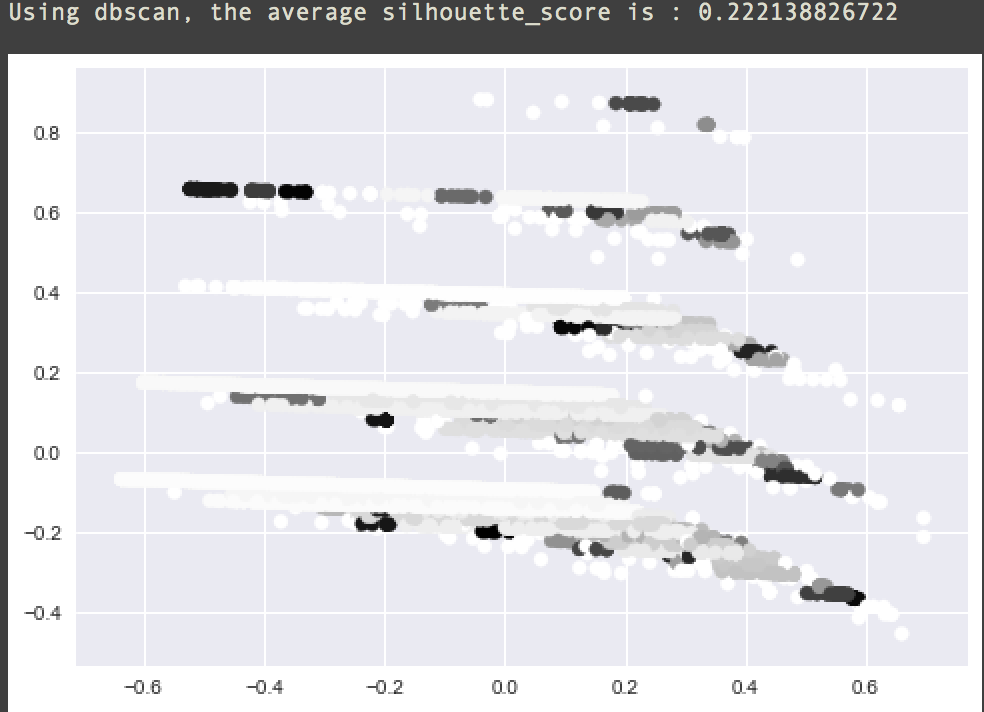


Fig 9. PCA cluster projection when cluster number k=5

This outcome can be explained by the drawbacks of k-means: it does not deal with non-globular shapes. The outcome could also be explained by the algorithm of k-means clustering, which is selecting the centroid randomly, then joining other vertices by the shortest distance. Therefore, the result of the final clustering depends heavily on the initial centroids.

DBSCAN clustering:

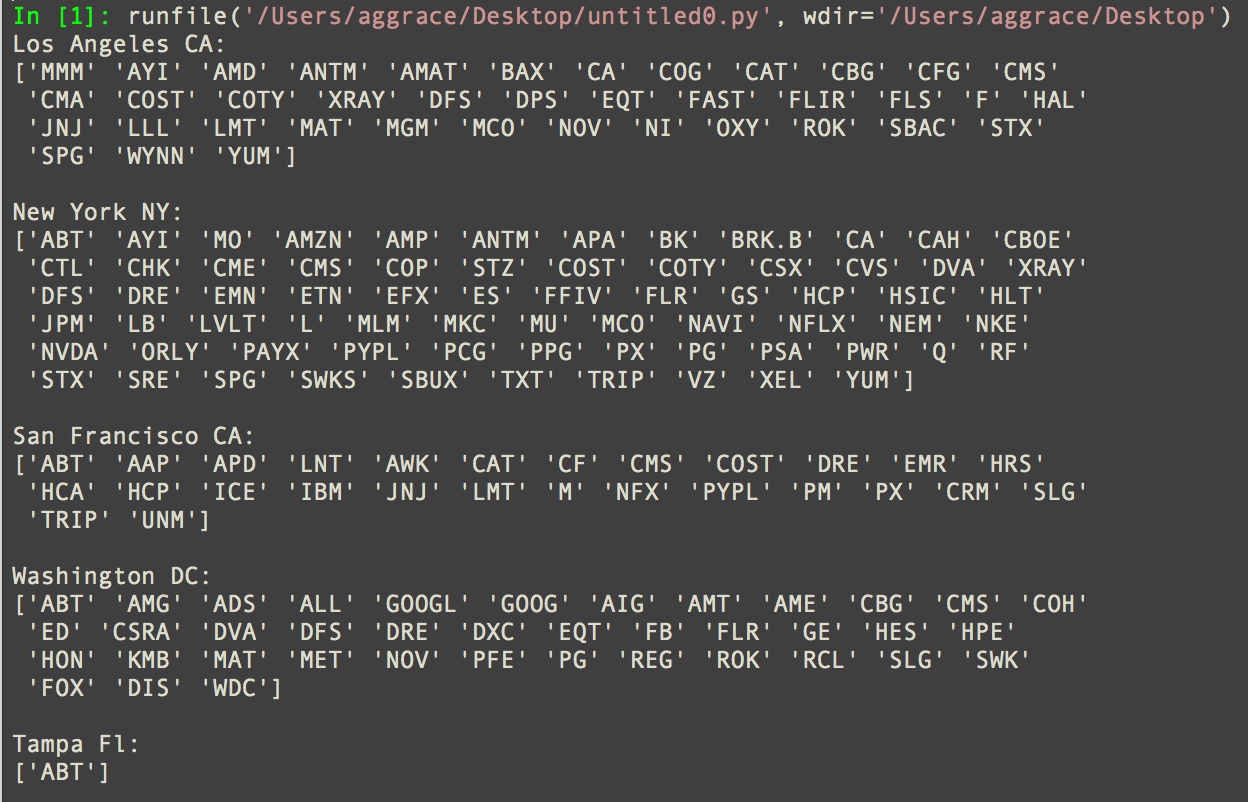
DBSCAN clustering is based on density clustering. It can be used for arbitrary shape. In order to cluster the dataset by the DBSCAN method, the first step is to select variables. Then, this process requires 'userFollowersNumber,' 'character\_length,' 'positive\_count,' 'negative\_count,' 'total\_count,' and 'score' as clustering variables. However, since the dimension is too high to cluster, we need to reduce the dimension by using PCA. Therefore we get a data frame with 2-dimension and cluster labels. We set up the Eps equal to 0.03 and have result as below:

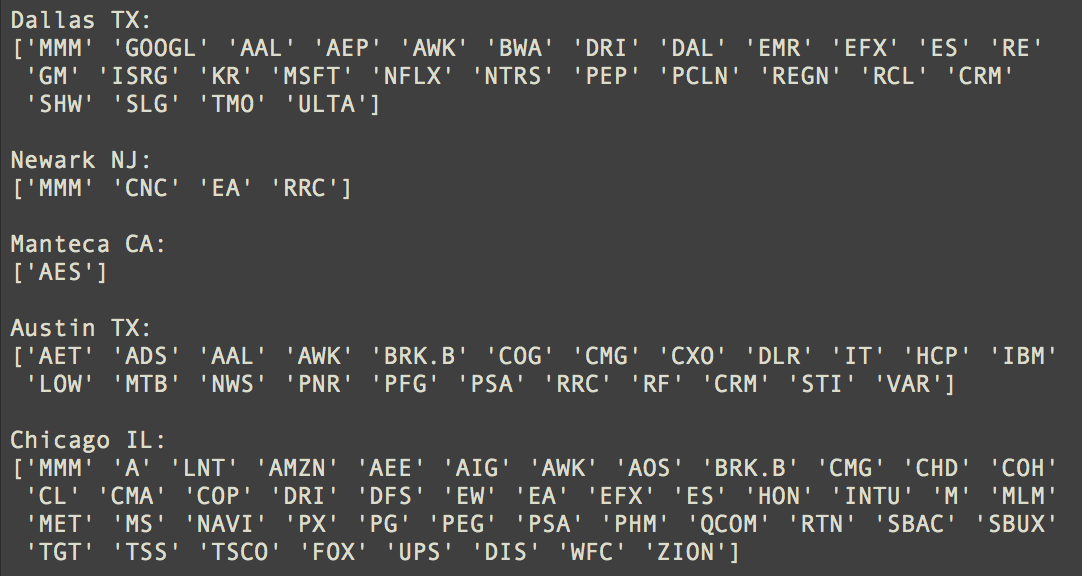


The silhouette coefficient of DBSCAN cluster is 0.222, which is not as high as the other two clustering methods. Also, the number of cluster by using DBSCAN equal to 5

**Association Rules / Frequent Itemset Mining Analysis**

The goal here is to figure out a group of stock indexes with high frequency on Twitter dataset. Firstly, we select 10 cities of United States as transactions from the column ‘User\_location’: Los Angeles CA, New York NY, San Francisco CA, Washington DC, Tampa Fl, Dallas TX, Newark NJ, Manteca CA, Austin TX, and Chicago IL respectively. After having these cities’ names, we gather stock index information from the same row with ‘ticker’ column. However, there exists a data cleaning problem: duplicate values. Therefore, the project used a unique method to remove those duplicate values and obtain the following results:





**Figure 10**

Here, 10 cities represent the transactions when applying the association rule. Then, the stock indexes of 10 cities become the itemset to investigate. Before using the apriori algorithm,

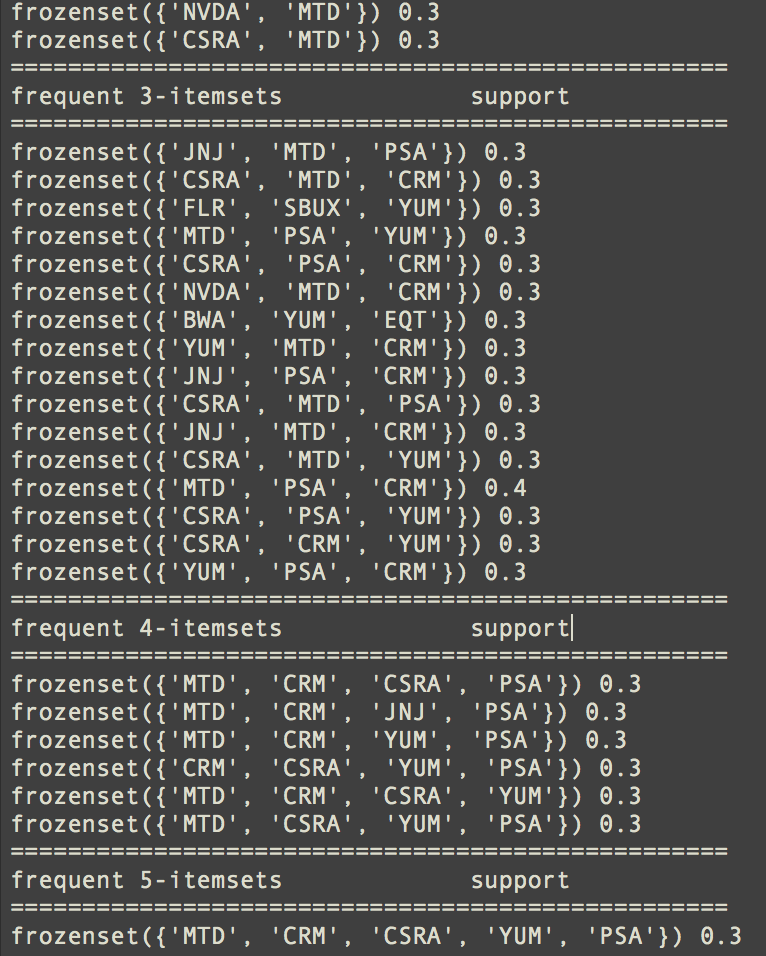
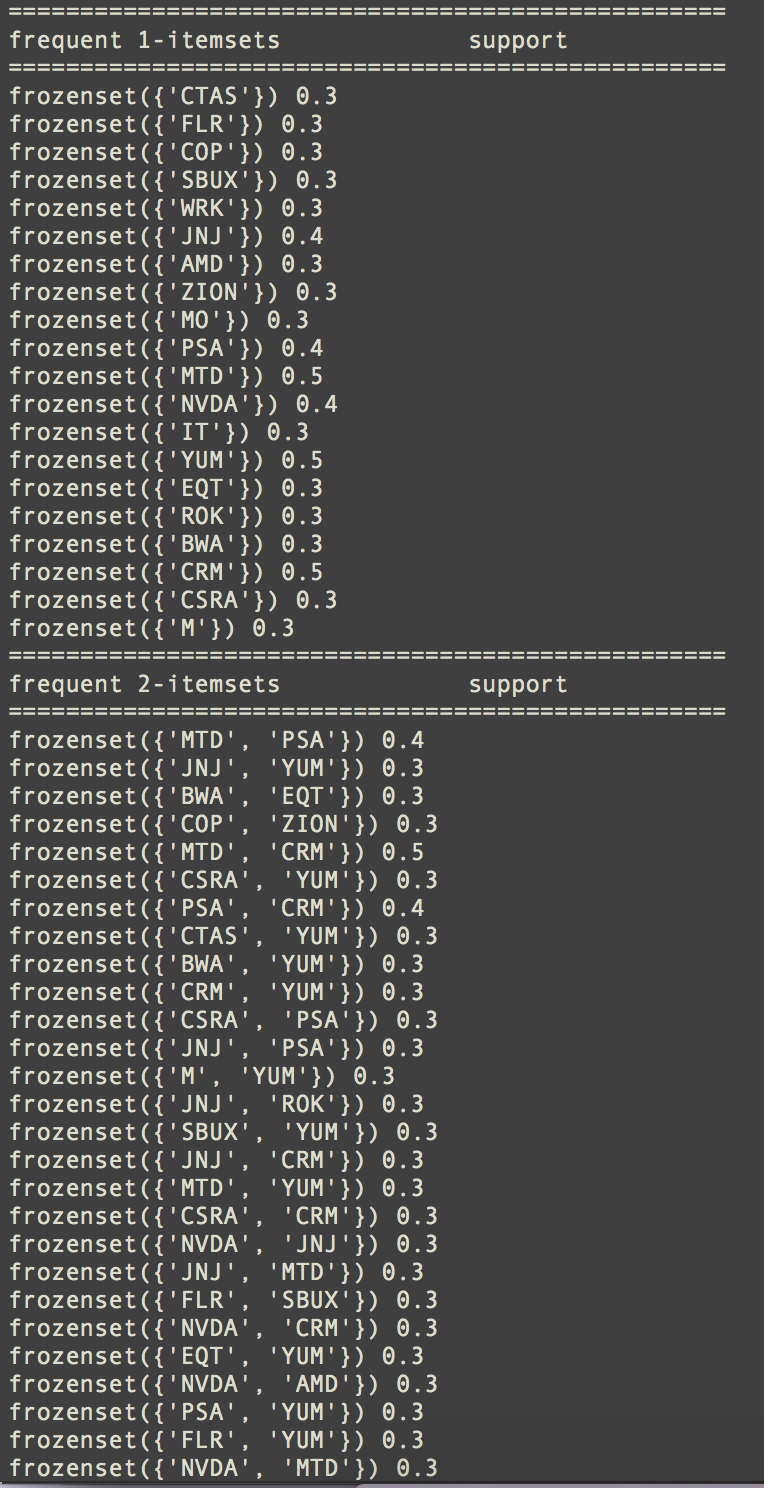
the first support level is 0.5, and the frequent itemset is 2. By applying the apriori algorithm, the first step is to find the frequent 1-itemset that satisfied the minimum support level. Then, the algorithm prunes the itemset that is not satisfied with support level 0.5, and begins to search for the frequent 2-itemset which satisfies the support level. After reaching frequent 2-itemsets, it is able to find out the frequent 3-itemsets. However, because of this relatively high support level, it cannot find the frequent 3-itemsets with support level 0.5 for those transactions. Hence, the program yields the following result:



**Figure 11, k=2, support level=0.5**

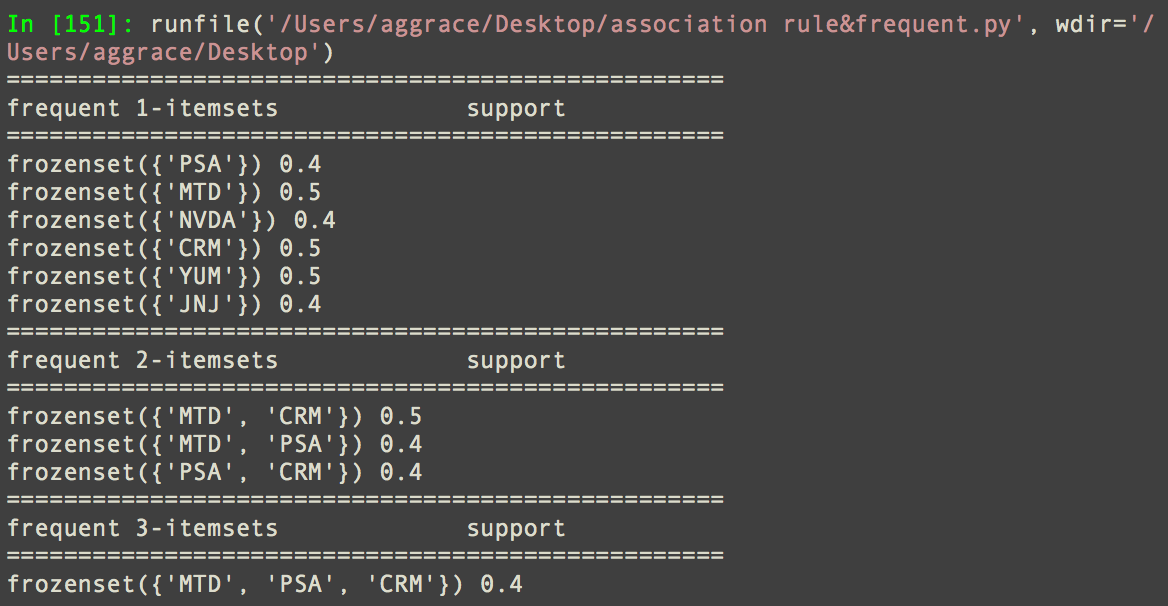
**(k stands for the frequent k itemsets)**

This result shows that the possibility that MTD and CRM always get mentioned together is at least 50%.



**Figure 12(k=5, support level=0.3)**

From Figure 12, it is clear that if assigning support level=0.3 and k=5 as the conditions, then the program can find the most frequent 5-itemset, which contains stock indexes {‘MTD’, ‘CRM’, ‘CSRA’, ‘YUM’, ‘PSA’}. This result shows that the probability of stock indexes with “MTD,” “CRM,” “CSRA,” “YUM,” and “PSA” listed together is at least 30%.

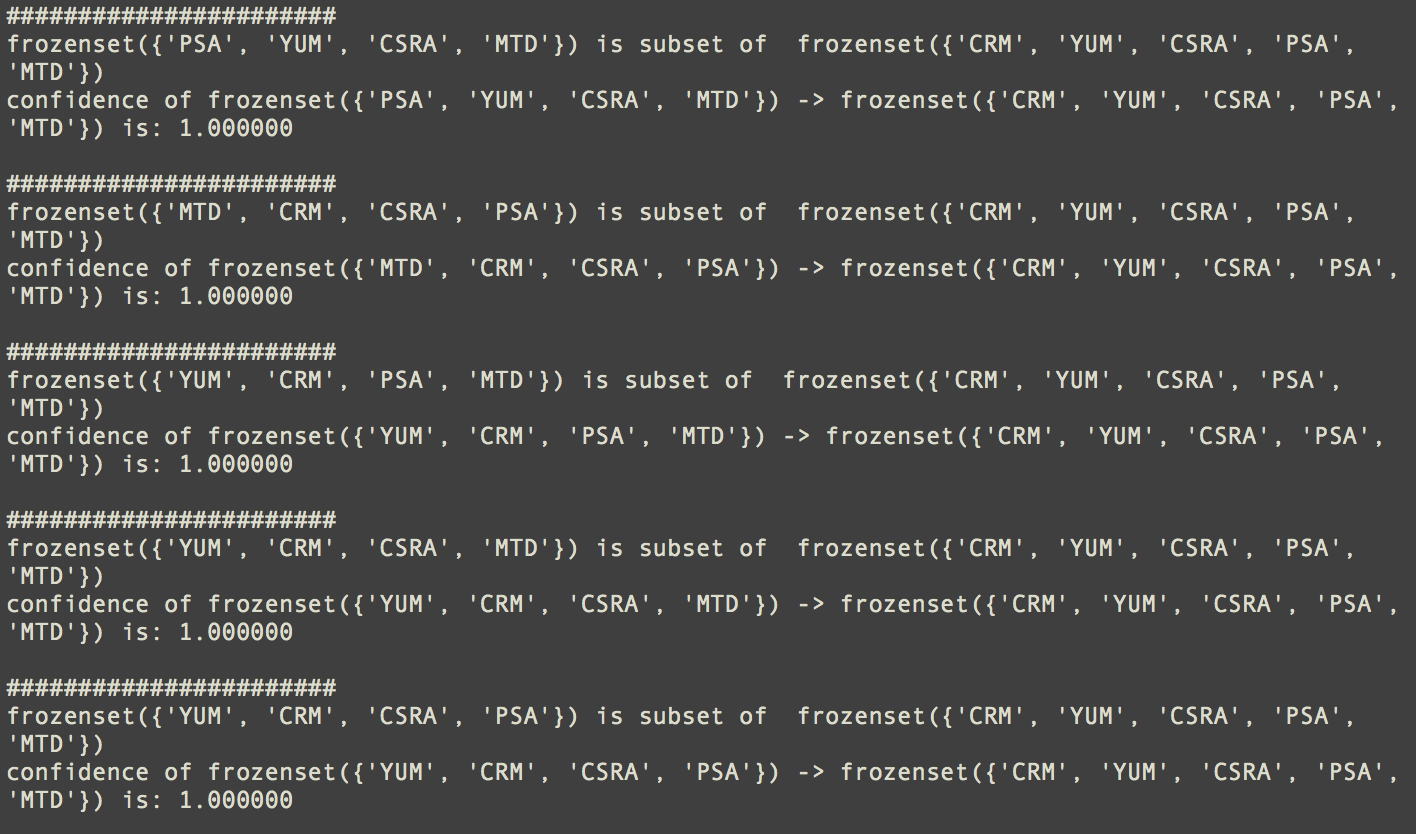


**Figure 13 (k=3, support level=0.4)**

Figure 13 shows the frequent itemsets with support level 0.4 are {‘MTD’, ‘PSA’, ‘CRM’}, which means the least possibility of showing “MTD,” “PSA,” and “CRM” together is 40%

The most frequent pattern with these three different support levels is k=5 and support level=0.3. Since the support level is relatively low when comparing with support levels at 0.4 and 0.5, it can find more frequent itemsets than at other levels. Also, by using apriori algorithm, it improves the efficiency of pruning infrequent itemsets and it makes the process of finding frequent itemsets easier. The most interesting part of applying the apriori association rule is that this support level cannot exceed 0.5, since in the first frequent 1-itemset, the largest support level is 0.5. Therefore, it is easy to know that the frequent itemsets like k=2, 3, 4 and 5 would not exceed 0.5. However, if the support level is equal to 0.2, the frequent k-itemsets would be very large and the python output cannot be shown completely.

Moreover, the confidence level of frequent itemsets would show later in output.txt. Figure 11 only shows part of the confidence level results with k=5 and support level=0.3.



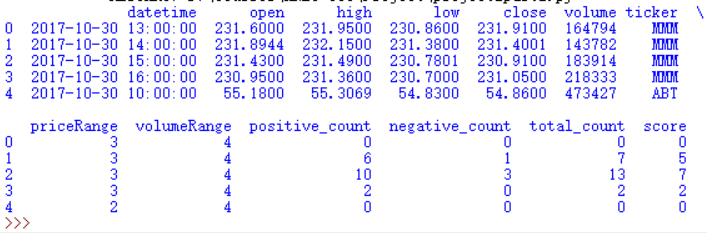
**Figure 14**

**Predictive Analysis – Part 1**

**Hypothesis Testing**

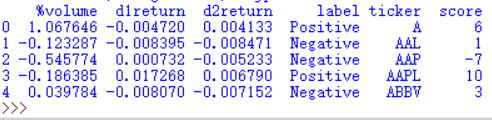
To answer the primary data science question, there are three aspects of interests that are highly correlated to the question. First, Twitter users share their feelings about companies by posting or retweeting tweets. Second, individual investors buy or sell some shares of a stock. These two behaviors are both decisive, fast, or even real-time reactions. Second, a small fraction of Twitter users are marked as Verified Users. Tweets from Verified Users, who are very likely to have a large number of followers, can receive wide concentration from fans. Sometimes, they need to be very careful while posting a tweet. Third and most important is whether an outsider, who can hardly touch information from company insiders, can really predict the stock price change in the future if they can only find tweets publicly posted on the Internet.

Therefore, this project brings up three one-to-one assumptions, constructs three hypothesis tests, and hopefully gains a clear result to either prove or disprove these assumptions. However, before the start of testing, appropriate datasets are required for either statistical or machine learning analysis. In order to study sentiment expressions on Twitter and actions of investors in stock market, it is necessary to merge two datasets from Twitter and the S&P 500 by adding all numbers of sentiment words together through tweets about a given stock for each hour and then transfer those data to stock datasets, because the time interval between two consecutive data points in the stock dataset is also 60 minutes. **Figure 15** is a sample of this new dataset:



**Figure 15. Merged Dataset for Hypothesis Testing 1**

The 4th row indicates MMM’s stock data on October 30, 2017 from 3:00:00 pm to 4:00:00 pm as well as its corresponding performance on Twitter during the last hour of the trading day. Though the second hypothesis does not require extra dataset to study because all the variables it needs are already in the original Twitter dataset, the last hypothesis still needs an additional dataset that contains daily returns, volume percent change, as well as a sentiment score, which is calculated by subtracting negative word counts from positive ones. The sentiment score is a positive integer if a user posts more positive words than negative words, and vice versa. If the score is 0, either that user is not expressing any feelings or the amount of positive words and negatives are equal. **Figure 16** is portion of such dataset after merging two original datasets in a different logic:



**Figure 16: Merged Dataset for Hypothesis Testing 3**

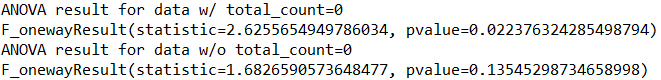
For example, line 4 for Apple, Inc. Its stock return is approximately +1.7268% on October 30th, 2017 and +0.679% on the next day. Its volume on Day 2 (October 31st, 2017) is 18.6385% less than the previous day. During Day 1 (October 30th, 2017), while looking through all tweets about Apple, the number of positive words is 10 more than that of negative words. “label” represents stocks’ behavior on Day 2, which is a label of “d2return”.

These new datasets are crucial to continue hypothesis testing. In the first hypothesis, it is interesting to know to what extent, if at all, the different volume bins differ in total sentiment expressions. That is, we want to test the following hypotheses (μ is the mean number of total sentiment word counts):

H0: μvolumebin1 = μvolumebin2 = μvolumebin3 = μvolumebin4 = μvolumebin5 = μvolumebin6

H1: μi = μj for any i, j (Or The hypothesis H0 is not true)

To test this hypothesis, one-way ANOVA test is the best statistical analysis method. The Python code result is on **Figure 17**:



**Figure 17: One-way ANOVA Test**

At the beginning, ANOVA result shows p-value of 0.022, so the null hypothesis would be rejected at most standard significance levels. However, after further considerations, it seems that a large portion of emotionless Tweets may pull down averages. In addition, a new ANOVA test is made for a dataset with sentiment tweets. The result is quite different, with a new p-value of 0.135, which is larger than 0.05. We therefore fail to reject the null hypothesis and more data is needed.

The second assumption is whether sentiment expression is related to Tweets’ character lengths, as well as whether or not the user is verified. As a surrogate for these factors, a linear function can be included as a predictor. In the second hypothesis testing, a multiple linear regression model is applied. The linear model has the mathematical form (note that, “total\_count” is the total number of positive and negative words in a single Tweet. “user\_verified” is a binary data. If yes, it equals 1. If no, 0. “character\_length” is the number of characters in a single Tweet):

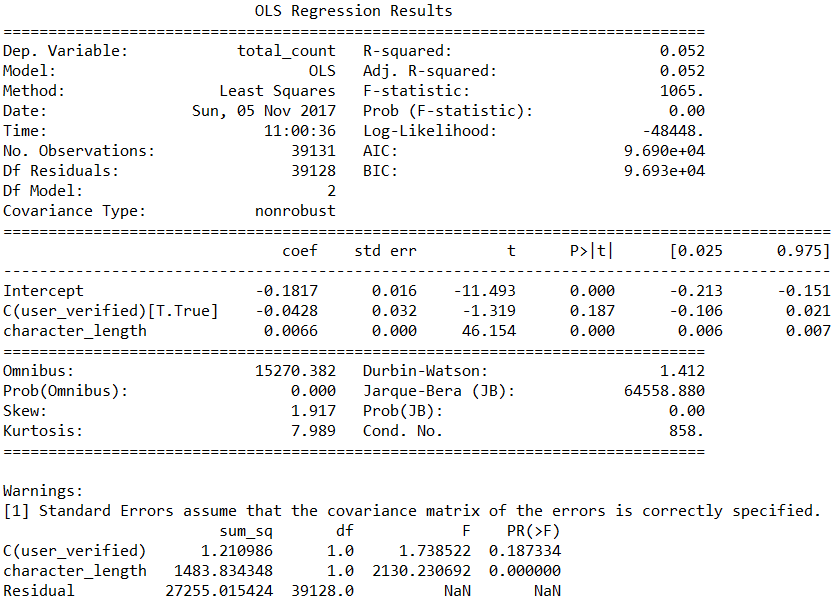
total\_count = β0 + β1 user\_verified + β2 character\_length

Therefore, hypotheses are as follows:

H0: β1 ≠ β2 ≠ 0

H1: some βi = 0

Inputting original Twitter data to Python code gives the following summary statistics and ANOVA results:



**Figure 18: Multiple Linear Regression**

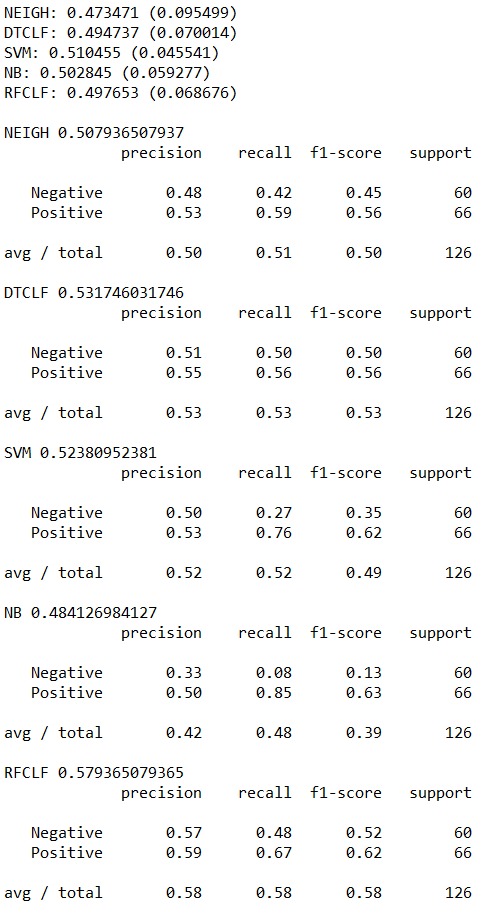
The linear regression results show that the mathematical linear models are

total\_count = – 0.1817 – 0.0428 + 0.0066 character\_length (if this user is verified)

total\_count = – 0.1817 + 0.0066 character\_length (if this user is not verified)

This result is quite interesting because the coefficient of “user\_verified” is negative, which means that holding all other factors the same, a verified user is likely to share fewer personal feelings with respect to a stock on public platform. This verifies the previous assumption on verified users’ caution while writing Tweets. It is interesting to note that p-value for variable “character\_length” is 0.000, which is very statistically significant. This is obviously understandable because when a Tweet is longer, it is more possible to show sentiment words.

Finally, the most important prediction in the entire project is to find whether people can predict stock change in the future based on history Tweets. Null hypothesis is that supervised learning can be used to predict future stock price, and it won’t work as an alternative hypothesis. To study this, machine learning and many classification methods are introduced. Methods used are Nearest Neighbors (NEIGH), Decision Tree Classifiers (DTCLF), Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest Classifier (RFCLF). One of the merged dataset has three variables (“%volume”: volume percent change; “d1return”: stock return in Day1; “score”: sentiment score) and one class (“label”: stock return in Day2 is “Positive” or “Negative”). By default, 75% of the dataset is training data. Applying machine learning analysis methods give the following output:



**Figure 19. Supervised Learning Analysis Results**

In general, the result is not satisfying because all accuracy scores are around 0.5, which means that using the method learning from 75% of the data can only predict approximately half of the test data. These machine learning methods are not even better than flipping a coin. Among all 5 classification methods, Random Forest Classifier best predicts future stock return with around 60% accuracy.

In conclusion of all hypothesis tests, twitter users react variously to different sizes of companies. Sentiment expressions has a close relation with users type as verified or not, as well as tweet length they post. In addition, currently, it is impossible for us to predict future stock price given by sentiments from social media.

**Limitation:**

There are a few drawbacks in the project. First, the tickers are abbreviations, which could lead to a misunderstanding and result in inaccuracy. For example, #AAP could stand for American Academy of Pediatrics instead of Advance Auto Parts inc. Second, programs do not have sense of humor. The word count and score may not deliver the true meaning of the twitter user. Especially nowadays people use sarcastic tones in tweeter.

**Conclusion:**

During this project, it is disappointing that the result of all supervised learning methods that has been tried do not have a satisfying result. The result could also be due to the reason that the collected data is only within a week, which could not accurately represent the big picture.