ANLY-501, Fall 2017

Group 7: Jiajia Liu Wupeng Han Jiaxu He

**Project Assignment 1**

**Data Science Problem:**

1. Explain the data science problem you plan to investigate.

Basically, the data science question for our group is to ask “Does the news’ major headline have influences on market index?”. As we notice that, the fluctuations of market’s stock price could be effected by various reasons, such as policies, economy, imports and exports, and so on. Hence, we are going to search for that the internal relationships between our daily news’s major headline associated with the stock market, and also determine whether there is a pattern or evidence existed or not to predict the trend of the stock price based on some historical data and current data as well.

1. Provide sufficient context and background information about why this problem is meaningful or adds insight.

As we mentioned a little bit in the first question, in general, we consider there are many different forces to move the stock price up and down. More specifically, those forces are more likely any logical reasons we think they will impact the stock price certainly, including interest rates, inflation and deflation, changes in economic policy, substitute, economic strength of market and peers, political shocks, change of demographics, etc. In a word, we could clearly observe that there is a naturally coherent flow with those reasons with the stock price itself, which those logical flows could be persuasively explained well to the public. Furthermore, what is the terms of “news” itself? We defined the term “news” as some new information or a report about something that has happened recently whether it is important and true or not. “News” also could be characterize as some latest comments and point of views towards to a certain topic. In general, news chosen reported is categorized as in several typical board areas, which is economy, politics, laws, science, real estate, health, sports, and arts. However, we would guess that there could be some sort of interrelations stand out behind the screens, since the movement of stock price is studied by a natural market behavior and also usually decided by human or groups’ a series of reactions. We consider this data science problem as meaningful and explainable, because that news typically brings emotions, feelings, and thoughts to the public. Consequently, all those emotions, feelings and thoughts would affect people and the public when they monitor the market and then control the stock price and expectations from the stock market. As a matter of fact, news is just a social media which gives a board direction to people what to think about instead of telling people how to think. Hence, all the decisions of the stock market are made are still based on people’s natural reactions after they hear about the daily news.

In addition, if we assume there is some kind of interrelationships between the news’ major headline and the stock market itself, we also would assume that those investors or large corporations would anticipate the patterns and issues therefore make changes on their stocks before the whole stock market collapsed. Some breaking news might cause the price of stock market suddenly dropped to an off-peak point in less than one day, while other some general news might cause the price of certain commodity to change in three to four days. For example, the breaking news of massive shooting in Las Vegas leads to the stock price of Storm Ruger Corporation increased 3.97% to $53.75 per share now on the same day, while the historical stock price of Storm Ruger Corporation is always stayed around $50 per share. Accordingly, if we could find the estimated average gap time of the reactions that would result in heavy volume of price drop, we could surely seek out a number of alternative ways and adopt important measures in order to minimize the corporation’s losses and still maintain least possible of shareholders at the same time. This is a microeconomic view to deeply discover this data science question, but we could regard the whole nation as a macroeconomic role to take actions whenever there is a breaking news happened.

On the contrary, another question would arouse our ponderation is that does our directions of investments and strategies will be highly different when there is no news as resources? As an ordinary people, news typically is our first choice of hearing about what is going on with the state, country, and world. Even not everybody would pay attentions to our daily news, major events that just happened constantly attract people’s attentions. Overall, we believe this data science problem would be deeply important and meaningful to discover about through our semester long project due to the thoughts and insights above.

3.Have a citation or two that supports your “problem ideas”. In other words, what research has already been done with respect to this problem.

Work cited:

Lien, Kathy. “How To Trade Forex On News Releases.” *Investopedia*, 21 Nov. 2005, www.investopedia.com/articles/forex/05/tradingonnews.asp.

Fedyk, Anastassia. “Research: How Investors' Reading Habits Influence Stock Prices.”*Harvard Business Review*, 2 Sept. 2016, hbr.org/2016/09/research-how-investors-reading-habits-influence-stock-prices.

Swedroe, Larry. “The Impact of News Events on Market Prices.” *CBS News*, CBS Interactive, 17 Oct. 2013, [www.cbsnews.com/news/the-impact-of-news-events-on-market-prices/](http://www.cbsnews.com/news/the-impact-of-news-events-on-market-prices/).

After searching articles and papers online, we found that there are some analysis and researches have been accomplished by different study areas of groups. Definitely, there is an unsure claim says that reported news released result in a rapid increase in volatility of the market price, but the problem is no one could really assert that how efficiently it reflected in stock prices. One of the interesting parts is that sometimes the market price immediately moves up and down after the new information released, but the new stock price somehow would adjust itself to a price which is very close to the old price at the end of the day. So maybe we would think that current market price is in analogy with a mirror that reflects the total knowledge and expectations from all investors after they processed the released new information. In a nutshell, since news’ major headline brings a lot of variations and explanations that causes the movements of market price, it would be interesting to know and investigate this data science problem.

**Potential Analyzes that Can Be Conducted Using Collected Data:**

1. You should first briefly describe the data you plan to collect. How many variables/attributes? What are their data types and levels of measurement? What does the data describe?

Since we are interested in whether there is an internal relationship between news’ headline and the stock market or not, so we plan to collect two datasets for variable news and three datasets for variable stock market, which are contains Nasdaq, S&P 500, gold, CNN news, and Fox news. We believe at least two datasets for each variable would be more representative in order to minimize the bias. For dataset Nasdaq(IXIC.csv), there are 4469 objects with 9 different variables in total. Nasdaq’s dataset includes two different data types which are 8 numerical variables and 1 categorical variable, and it displays the movements of a stock market in a given day, such as open price, close price, high price, etc. For the levels of measurements, 8 numerical variables are ratio variables. Similarly, dataset S&P 500 (GSPC.csv) has 3001 rows with 9 different variables in total, and it has the same attributes, same data type and same levels of measurement as dataset Nasdaq. Also, dataset Gold (gold.csv) has 523 objects with one categorical variable and one numerical variable. Gold dataset mainly tell us what is the stock price of gold in that day.

Furthermore, we are planning to divided the news’ headline into two big categories later on, which is “relevant news” and “irrelevant news”, because we assume that only some news’ headline would have influences on the stock market. Hence, the levels of measurement of news’ headline are binary variables, 1 represents relative news while 0 represents irrelevant news. Dataset CNN (cnn.csv) has 47989 counts of variables with 8 variables in total, which has 3 numerical variables and 5 categorical variables. Some of the attributes are favorites\_count, source, text, in\_reply\_to\_screen\_name, is\_retweet, etc. For variable “is\_retweet”, it is a very typical binary variable since the levels of measurement only contain 2 levels, which is true or false. Dataset CNN primarily tells us the actions associated with the contents of the tweet, such has how many people liked this tweet, whether they retweet the tweet or not, how many people retweet the tweet after they saw it, which source does the tweet come from, and so on***.*** Consequently, dataset Fox News has 47990 objects with 8 different variables in total, and it has the same attributes, same data type and same levels of measurement as dataset CNN. The existence of both dataset CNN and dataset Fox News serves the same purpose despite its contents of tweet.

1. How will your data support your data science problem?

The time frame of all datasets is one year based, so all the information we have has the same length of standard time. Hence, since our majority of stock market datasets and news datasets display the price of stock and new reported information per each day, then we would proposal an associated relationship between them, and build a method later on in second portion of project to observe the fluctuations of the market’s price whenever there is a new released of reported news.

1. Then write a brief explanation of possible directions, goals, and hypotheses that you may be able to investigate with the data you have collected***.***

In order to investigate the data science question more deeply, we would create some data mining processes to analyze the content of the headlines. Therefore, we are going to code a sentimental analysis to test the positivity and negativity of the headlines. Typically, good news brings positive attitude to the public, leading to a probable rise in stock price caused by investors, and vice versa. If we set the scale of this sentimental analysis as [-1,1], we suppose that upon release, the best news has influence that outputs value 1, and the worst news outputs value -1. If the news tends to be irrelevant or has very little influence to the public, the output from the sentimental analysis will be close to a value that equals to zero. In addition, if we assume two datasets of news as x variable or a explanatory variable, then the three stock market datasets would be treated as y variable or a response variable. In this case, due to the large volume of news released, our explanatory variables include the counts of favorite likes, the counts of retweet, and the weighted score from the sentimental analysis per day; our response variables are the movements of stock price and the volumes of trade per day. Consequently, we are planning to use statistics techniques to build a multiple linear regression model to test how the mean value of the response relates to specific variable of the predictors. Our major hypothesis is whether the value of coefficient β is equal to zero or not. If the coefficient β is equal to zero, we believe that there is not a significant linear relationship between our predictors and responses, and we reject the null hypothesis, and otherwise, we would confirm the hypothesis. we are going to set up 6 models to discover the relationship between headlines’ influence and stock market. In order to determine the existence of significance, we will use ANOVA test to find the answer since ANOVA uses an F test statistics to test if two variances are equal. In addition, we will use 2 sample T- test as well. If both ANOVA test and T test give us the same results, that means our model and estimated analysis are accurate.

Before we perform linear regression model, we notice that there might be a potential multicollinearity issue and the uncertainty distribution of all the variables. First of all, we might have variables that provide redundant information. Multicollinearity issues cause issues like generation of different significance in F- test and T-test, low value of partial etc. For instance, variable favorite counts might be highly correlated with variable retweet. Thus, we will need to diagnose the VIF value and try to eliminate the highly-correlated variables. Moreover, if the data is not normally distributed, we have to transform it to enhance its convenience, reduce its skewness, and strengthen its linear relationship for the model.

1. Ideas here may not end up being your final question set - these are part of the exploration and thinking phase. At this stage, you are generating possible directions.
   1. News has been posted more and more on twitter in recently years. So if our theory was true, would it be possible that news’ channels have realized the influence of news’ contents, and they are already trying to influence or even manipulate stock market or prices?
   2. Would the News only affect people’s decisions on items that are flexible and light, like stock holdings and/or travel plans? For things that are less flexible and more essential investments such as gold, according to our latest techniques, the News may not have too much influence on them.
   3. Would the influences of the News not appear immediately, but show up after a period of time? In addition, what if some other news comes out during the period and has quicker influences on the market which may cause the result to overlay the impact from the previous news?
   4. The number of viewers on twitter changes every day. Someday there could be a lot of people watching the news and someday they are fewer. In this sense, the influences of the news change accordingly.

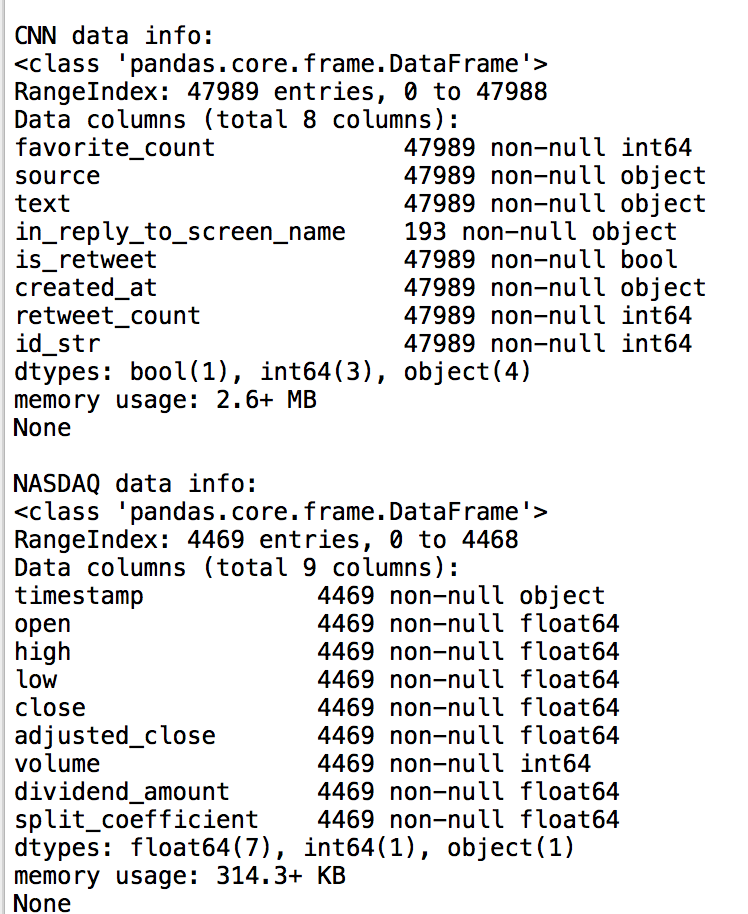
**Data Issues:**

1. For this part, please explain the issues that you see with the data, e.g. noise, missing values, level of cleanliness, etc. List all issues that you see - so that you will be able to clean the data accordingly.

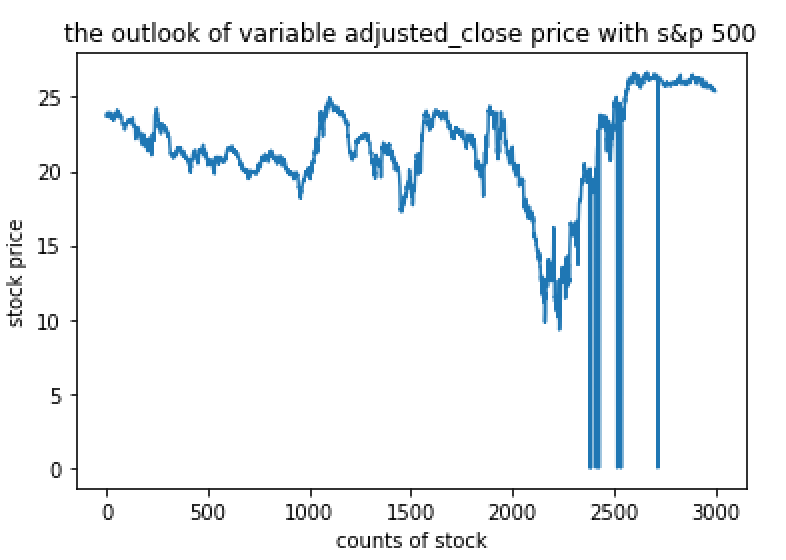
Even if the majority of the datasets looks quite clean for us, there are several issues that we detect after using pandas to read the datasets as data frame in python: messy contents, untidy formats of the timeline, potential outliers, and incorrect values. Missing value means that there is no number or no content in a given column. Basically, we used the *print info ()* function to check the presence of missing values in the first step, and it turns out all the columns are filled with either objects, integers, float, bool (*displayed in graph\_1*). This result tells us that there are zero missing values for those five datasets. Next, a few tweets’ contents look very messy and incomprehensible, so we generated a code to delete all the obscure contents which has zero influence on the market index. The timelines of the stock market datasets and news datasets have totally different formats, so we created a code to define all the dates in the format of *year/month/day* for both news’ datasets. By formatting the dates, it is easier for us to read and more convenient for mergence purposes in the future. Additionally, we used boxplot and other regular plots to check whether there were outliers in the datasets. Outliers are points that diverge greatly from the overall pattern. It turned out that only the S&P 500 dataset and the gold dataset had 4 and 11 extreme outliers respectively (*displayed in graph\_2 and graph\_3*) since the stock price cannot be zero dollars in general. In fact, we believe it might be a typo error. We decided to not delete any outliers for Fox dataset and CNN dataset, because the potential point of variable retweet that far away from the distribution might be regarded as some breaking news taken place.

The biggest problem for having 5 datasets was that we had a few incorrect values inherent in the stock price and volume of trades. Incorrect values mean that the value of that given variable is logically wrong. Here, we made two considerations and some additional viewpoints. First of all, the volume of trade being equal to zero reflects that there are no changes or fluctuations in the whole market. Possible reasons include: holiday breaks and the existence of an incompletely competitive market. Therefore, we built a code to check and delete all the columns when the variable volume was equal to zero. The second condition was that there was no changes of market index when the highest price was equivalent to the lowest price in a given day, so we chose to delete any columns where two prices are the same. Lastly, we also checked the existence of duplicated observations. Duplicated observations mean that the information of that given column is the same as other columns, which is meaningless. After we designed a code to test it, it turned out all of the columns are different than each other.

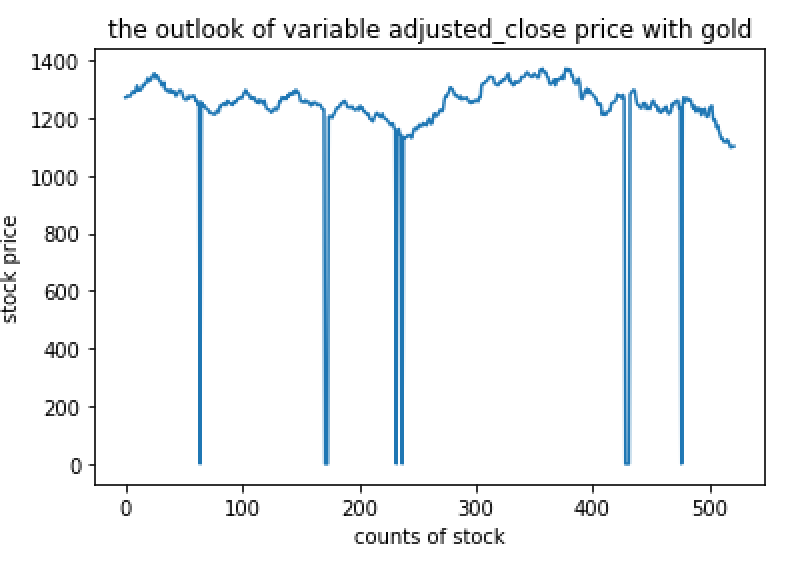
References:



(graph\_1)



(graph\_2)



(graph\_3)

**Data Cleaning:**

1. As part of the cleaning requirement, you will create a way to measure and score how clean your data is at the beginning and after cleaning. Explain these measures (quality scores) in your project in words, and also include the methods in your code. \*\*Always show and discuss ALL WORK in your project write-up involved in creating these quality measures.

When we first got our raw data of tweets from News, the contents of tweets were really messy: they contained a lot of obscure characters, symbols and even URLs. Our first goal was to integrate all the tweets into the format that contained only the information and words we need. The first way we created to measurement was the ratio of clean rows/total rows. At the beginning, ratio of clean rows/total rows was close to 0, since almost every tweet had problems. after the cleaning, we deleted all URLs, and generated a measurement ratio greater than 0, since some rows were ready for use. Next, we deleted all symbols and nonsenses characters, and left the contents only with letters and numbers, which helped our next step analysis for news’ datasets. At this point, the updated ratio was approximately close to 1, which meant that all data had been cleaned according to our requirements and instructions. When we were dealing with the stock market and gold data, we created another measurement method to compare the variances before the cleaning and after the cleaning, since half of the useful data was in the form of numbers. and we performed the analysis on numbers directly. We calculated the mean first, and then the variance of the whole samples. Next, we executed the cleaning process. We deleted the rows with low price being equal to the high price and ones where the volume equaled to 0. We also fixed some rows to assist our comprehension process. After the cleaning, we calculated new means and new variances. We found new variance is much lower than the previous one by a big percentage, which meant that we got cleaner data after we performed cleaning process.