Trying to understand the Impact of News on the Stock closing prices

# Introduction

First, provide the context of the problem and then state the problem (your main research question). Second, write briefly that what are you proposing to solve this problem (don’t write details of the solution here). (You can use part of your abstract here)

The stock prices are usually affected with many factors like momentum, PE ratio, but one of the most important factors that affect stock prices is the news feeds featuring this stock.

For example the word “Nissan Recalls” would lead to a drop in the price of the Nissan’s stock, while it might represent a raise in the price of Toyota & Honda stocks.

Although the news feeds have many sources but we tried to focus here on the news snippets, summary and abstracts retrieved from New York Times.

In this capstone we will try to build a model to predict the stock prices OR predict the changes in stock prices for two stocks (Apple and Microsoft) based on the news feeds.

To predict the stock prices we would need to go through many regression prediction models like ()

While for more accurate results we might try to just predict the changes in the stock prices (UP - Down) by applying classification models like ()

# Literature Review

Write summary of the related papers that you reviewed here. Write the summary in your own words—don’t use the technical jargon from the paper that you don’t understand. Keep this section short—a short paragraph or few sentences about each paper you reviewed should be sufficient.

# Dataset

Give the description of the dataset that you are using along with the individual attributes you will or will not use in your analysis. Also mention the source of the dataset (where did you get it from). In case the data is curated and created by you please explain the details. Descriptive statistics of the attributes and datasets can also be provided here.

The Dataset here consists of Stock Values of Apple(AAPL) and Microsoft(MSFT) from 2006 to 2016 and News summary, abstract and snippets on News featuring these two tech giants during the same period.

The **final data set** was created by joining the Market data and the news data sets after applying sentiment analysis on those News string and converting them into a score, so that the final dataset can further be used for a Regression Model.

While the Sentiment analysis was already prepared and provided with the dataset on kaggle but here is what the sentiment values represents here:

* The pos, neu, and neg scores are ratios for proportions of text that fall in each category (so these should all add up to be 1... or close to it with float operation).
* The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). In other words it is a 'normalized, weighted composite score.

1. **positive sentiment**: compound score >= 0.05
2. **neutral sentiment**: (compound score > -0.05) and (compound score < 0.05)
3. **negative sentiment**: compound score <= -0.05

Although this data set is downloaded from Kaggle.com, The News snippets, summary and abstracted were originally retrieved from The New York Times API, while the stock values were obtained from Yahoo Finance.

**Dataset Link**:

https://www.kaggle.com/BidecInnovations/stock-price-and-news-realted-to-it

Our Dataset consist of four files:

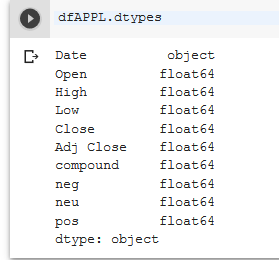
1. AppleFinalData.csv:

This file contains the Stock Price of Apple from 2006 to 2016 and the sentiment score of the news about Apple.

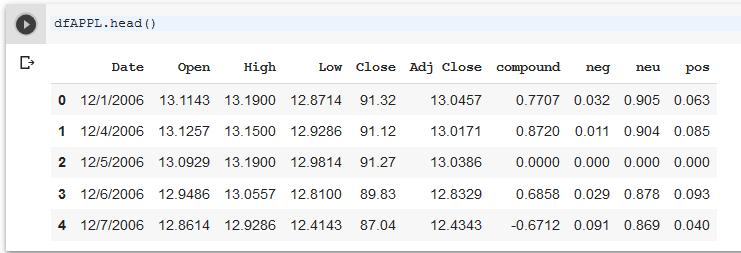
The dataset contains the following attributes (columns):

1. Date: In YYYY-MM-DD format, note that the dates here represent the trade dates where the Market is open (week days), that means that the days in a trading year are 252 days. (object)
2. Open: Opening weighted average stock value in USD (float)
3. High: All day high in USD (float)
4. Law: All day low in USD (float)
5. Close: Closing weighted average stock value in USD (float)
6. Adj Close: Adjusted closing prices - adjusted for both dividends and splits - in USD (float)
7. Neg: The sentiment analysis measurement of the conditional probability of this news being in the negative news class. (float)
8. Neu: The sentiment analysis measurement of the conditional probability of this news being in the neutral news class. (float)
9. Pos: The sentiment analysis measurement of the conditional probability of this news being in the positive news class. (float)
10. Compound: This is the compound score of the news (float)

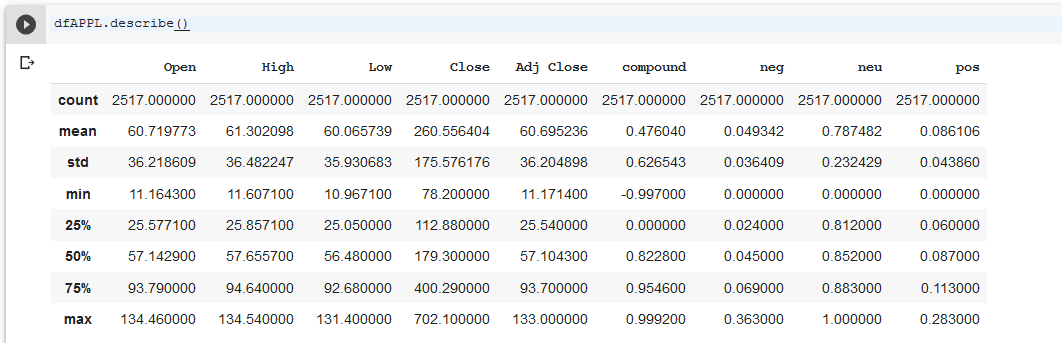
Below is the data types of the attributes:



Below is a sample of the data head



Below is the statistics description for each column



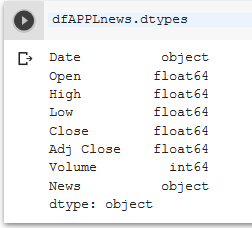
1. AppleNewsStock.csv

This file contains the Stock Price of Apple from 2006 to 2016 and the string of News summary and snippets published by the New York Times on that day about Apple.

The dataset contains the following attributes (columns):

* 1. Date: In YYYY-MM-DD format, note that the dates here represent the trade dates where the Market is open (week days), that means that the days in a trading year are 252 days. (date)
  2. Open: Opening weighted average stock value in USD (float)
  3. High: All day high in USD (float)
  4. Law: All day low in USD (float)
  5. Close: Closing weighted average stock value in USD (float)
  6. Adj Close: Adjusted closing prices - adjusted for both dividends and splits - in USD (float)
  7. Volume: Number of trades in that day. (int)
  8. News: It’s the combination of all the news headlines and there abstract into a single string. (object)

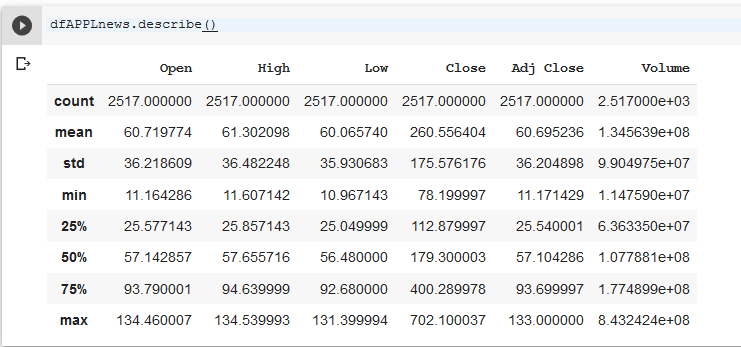
Data types of the Apple news dataset:



Below is a sample of the data head



Below is the statistics description for each column



# Approach

Create a block diagram for the steps of your approach to clearly provide an overview. For example, if you first scrapped twitter, second applied NLP techniques to extract keywords, third labelled the tweets as positive and negative using a set of keywords, and fourth build a classifier, then you should create a box for each of the steps with arrows connecting one step to the next one. A sample block diagram is shown below.

Once this is done, explain each of the steps in detail. What are you planning to do in each step or have already done. For example, in the above case you would create subheadings for each of the steps.

## Step 1: <Name of the step>

Write details of the step 1. If there is any source code that you’d like to share then provide the link of the Github.

## Step 2: <Name of the step>

Write details of the step 2. If there is any source code that you’d like to share then provide the link of the Github.

………………………………………………………………………………………………………………………..

………………………………………………………………………………………………………………………..

## Step N: <Name of the step>

Write details of the step N. If there is any source code that you’d like to share then provide the link of the Github.