Paris Vu

# Final Report:
# Twitter Sentiment Analysis

## Problem Statement

There is evidence to suggest that there can be a correlation between tweets and stock prices, particularly for publicly traded companies. The idea behind this correlation is that tweets can influence public perception, which in turn can impact stock prices.

For example, if a company releases positive news, such as a new product launch or strong financial results, this may lead to positive sentiment on social media and an increase in stock prices. On the other hand, if a company is involved in a scandal or there is negative news about it, this can lead to negative sentiment on social media and a decrease in stock prices.
Some studies have shown that sentiment analysis of tweets can predict stock prices with a moderate level of accuracy. However, it is important to note that correlation does not necessarily imply causation, and there are many factors that can influence stock prices.

By scraping live tweets for the month of December 2022. I created a model to help Apple Inc. quantify the sentiment of the tweets written regarding them. Measuring sentiments in tweets can provide valuable insights into the attitudes and opinions of people on their products. For example, Having a specific scores on Apple related tweets can help the company with:

1. Understanding customer feedback: Many businesses use social media platforms, including Twitter, to connect with their customers and receive feedback. Sentiment analysis can help these businesses to understand the opinions and attitudes of their customers towards their products or services.
2. Monitoring brand reputation: By monitoring sentiments on Twitter, businesses can also track the sentiment towards their brand and products. Negative sentiment could indicate a potential PR crisis or highlight areas that require improvement.
3. Tracking public opinion: Sentiment analysis can provide insight into the opinions of people on various topics, including politics, social issues, and current events. This can help organizations, including governments, to make informed decisions that are in line with public opinion.

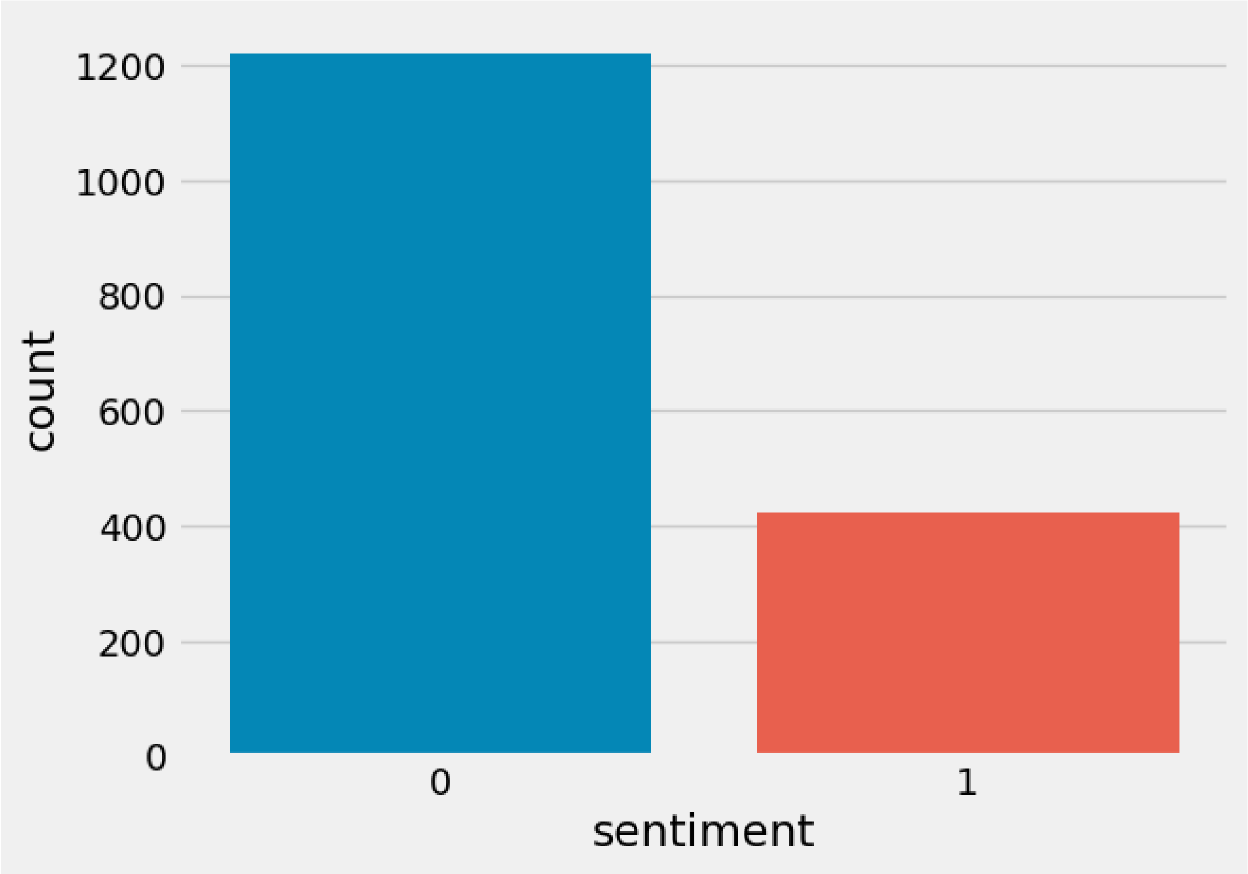
4. Improving social media engagement: By analyzing the sentiment of tweets, businesses can identify the topics and issues that resonate with their audience. This information can then be used to develop targeted content that increases engagement.
5. Research and analysis: Researchers and analysts can use sentiment analysis to gain insights into public opinion and attitudes towards various topics. This can be useful for social science research, marketing research, and political analysis.

## Historical Data (2014)

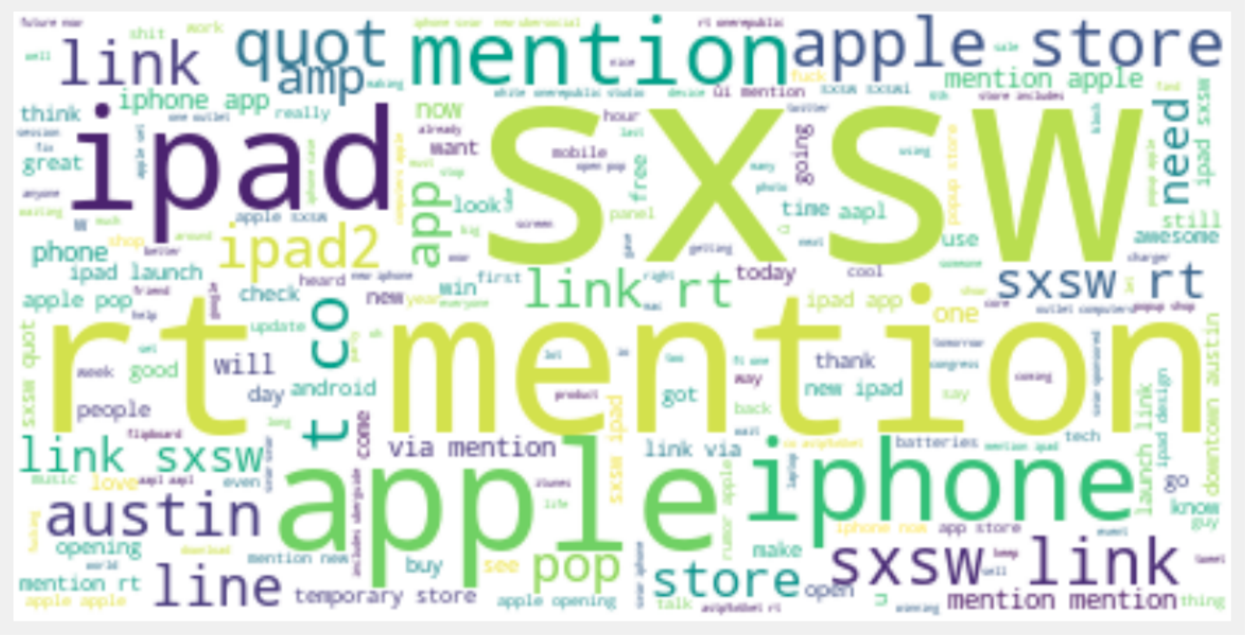
Before diving into tweets scraping, I first explored a pre-scraped dataset (Dec. 5 to Dec. 10, 2014) that contains 3886 rows and 12 columns. Although processed, the dataset is still not perfect and has a lot of null values. Because we only care about the contents of the tweets, we can slowly eliminate columns that we do not need such as "_unit_id, "_golden", "_unit_state", etc.
The author also provides a column with the keyword they used to pull their tweets: #AAPL OR @Apple. Knowing this information, I was able to apply it to my own queries in the Data Wrangling of this project.

After I've dropped all of the columns that I believe were irrelevant and null values, the only sentiment values left in the dataframe are 5 (positive), and 1 (negative). To simplify the numbers, we created the function new_sentiment(x) to convert 5 to 1, signifying positive, and 1 to 0, signifying negative. By plotting the values count of these numbers, we can see that there were three times the amount of negative tweets to the amount of positive tweets. This result is not surprising since historical records show that there were a mix of positive and negative indicators in the global economy. For instances, some economic highlights at the time were:

1. The US economy continued to show signs of improvement, with the unemployment rate falling to 5.6%, its lowest level since 2008. This was seen as a positive sign for the US economy.
2. The European Central Bank (ECB) announced a new stimulus program to help boost the Eurozone economy, which had been struggling with low inflation and weak growth. This program included buying government bonds, known as quantitative easing.
3. Falling oil prices continued to put pressure on oil-producing countries, such as Russia and Venezuela, which experienced economic difficulties as a result.
4. The Chinese economy showed signs of slowing down, which led to concerns about the impact on global growth.



To help me get a better understanding of what words users were tweeting in these tweets, I used another dataset with a completely different set of tweets, but this time it also contains their mentioned apple products and the emotions associated with it. Looking at the data, we can see that the owner has done some natural language processing to get the positive and negative emotions associated with the tweets. We will use this data, together with the first csv, to generate a word cloud.

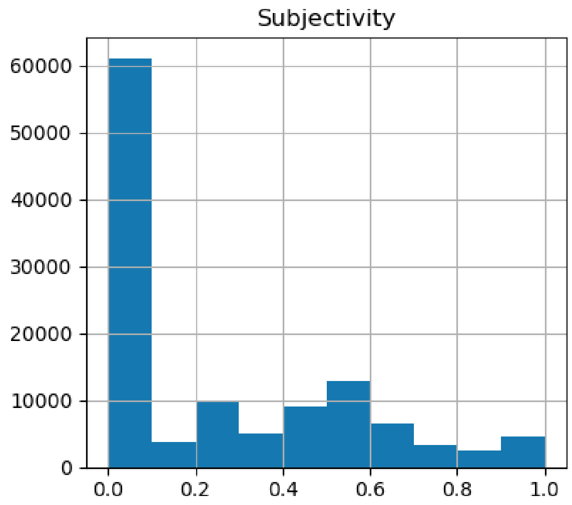


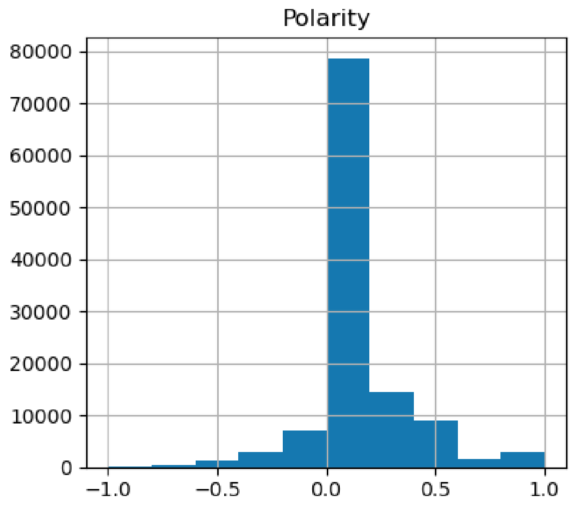
One common 'big' word that appear in both of these dataset is **SXSW**. After doing a simple google search, SXSW is a music festival that is usually held in mid-March. It is surprising to see that both the first dataset (with dates in December 2014) and second dataset (unknown dates) mention this festival. This means that tweets are 3 months ahead of real life events!

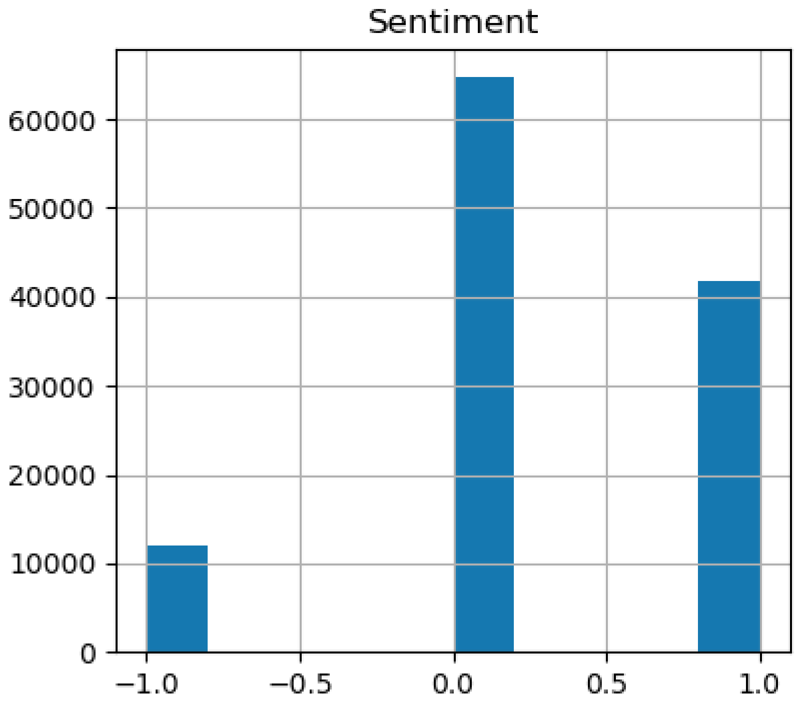
## Data Wrangling

After exploring historical datasets, I scraped live tweets for the month of December 2022. I chose this month because it was the month that I was in while carrying out this project and I want to see whether the statement that tweets are 3 months ahead of real life events stays true in March 2023.

I scapred Tweets with specific hashtags (#appl, #apple, #ipad, & #iphone) using the SNScrape method, and generated Apple stock data for the month of December 2022. The raw dataset first had the following columns 'user', 'likes', 'source', 'text', 'Time'. After having the contents of these tweets, I extracted their subjectivity and polarity, then quantified it into numbers representing the following: -1 = Negative; 0 = Neutral; 1 = Positive. Subjectivity detection and polarity detection are subtasks under sentiment analysis. Subjectivity detection aims to remove 'factual' or 'neutral' content, i.e., objective text that does not contain any opinion. Polarity detection aims to differentiate the opinion into 'positive' and 'negative'.





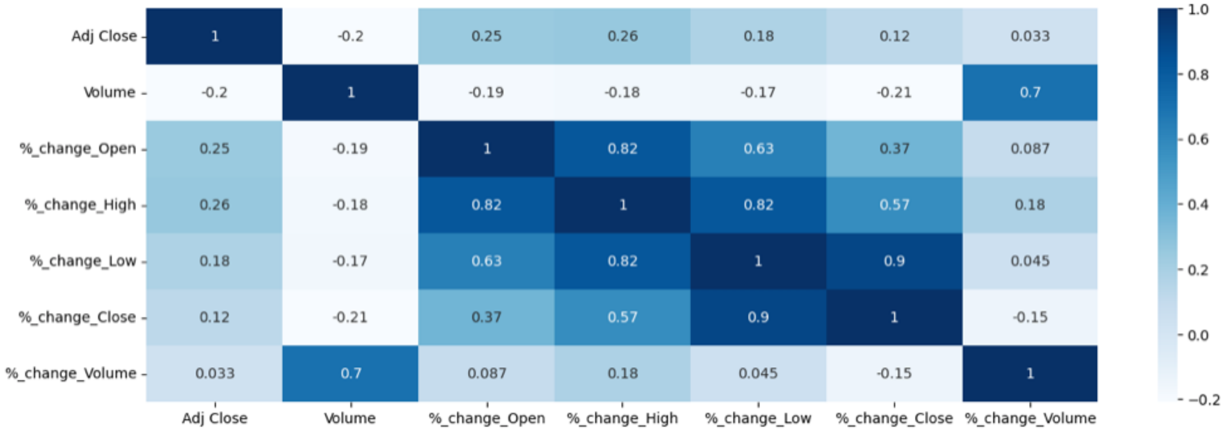


At the end of this notebook, I was able to produce two datasets: live tweets with quantified sentiments from Dec. 2022 (01_tweets_data) and historical stock price from Dec. 2022 (01_stock_data).

## Exploratory Data Analysis

In this notebook, we extracted some features from stock data and visualized their correlation. We also looked at the tweets from the previous notebook (01_Data_Wrangling) to extract the most common words, likes, time of tweet, and frequent user.

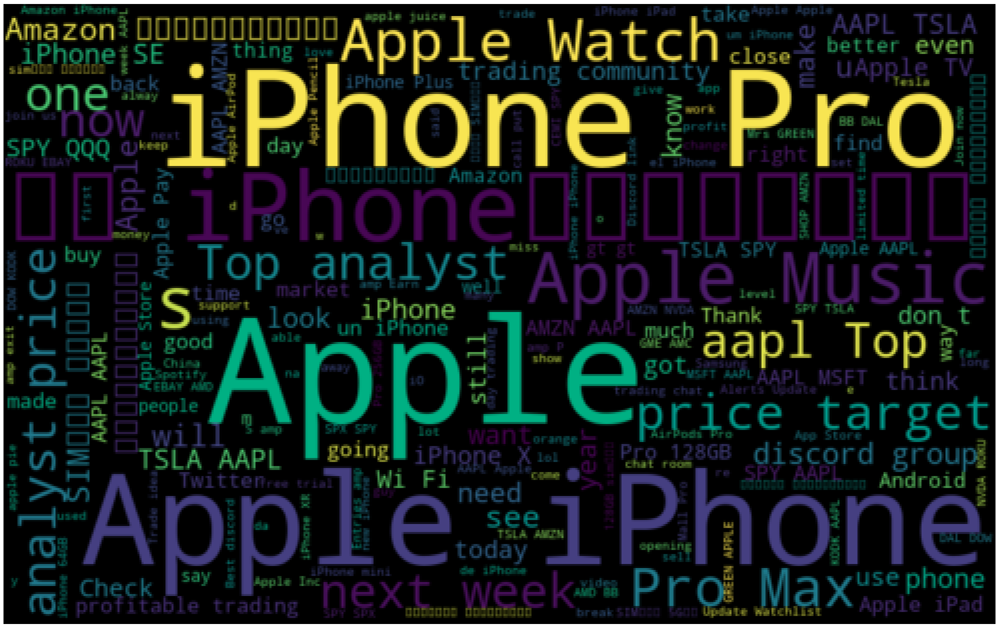
From the stock data that we generated in Data Wrangling, we examined the change in price between a given date and its next day's pricing (one period into the future) using the function .shift(). For example, "shifted_Open" implies the next day's opening price for the current date. From there, we were able to find the percentage change of AAPL opening, high, low, closing prices, and daily trading volume.

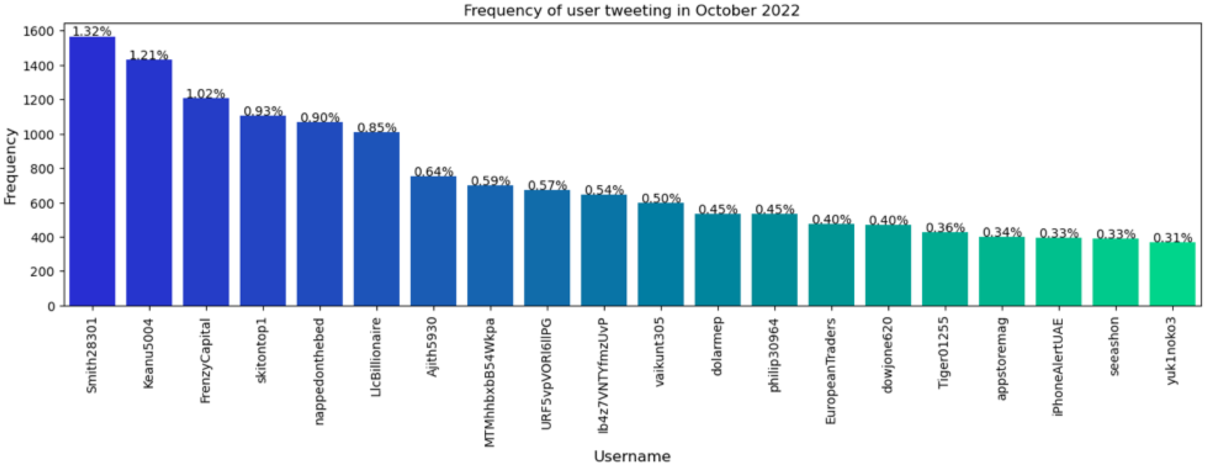


The main variable we are going to focus on in this dataset is the ‘Volume’. Variables having a large correlation value with volume represent that those numbers might have intrigued a large number of buyers and sellers. Correlation between those features and the volume feature will tell us how a change in that feature impacts the number of stocks traded that day.

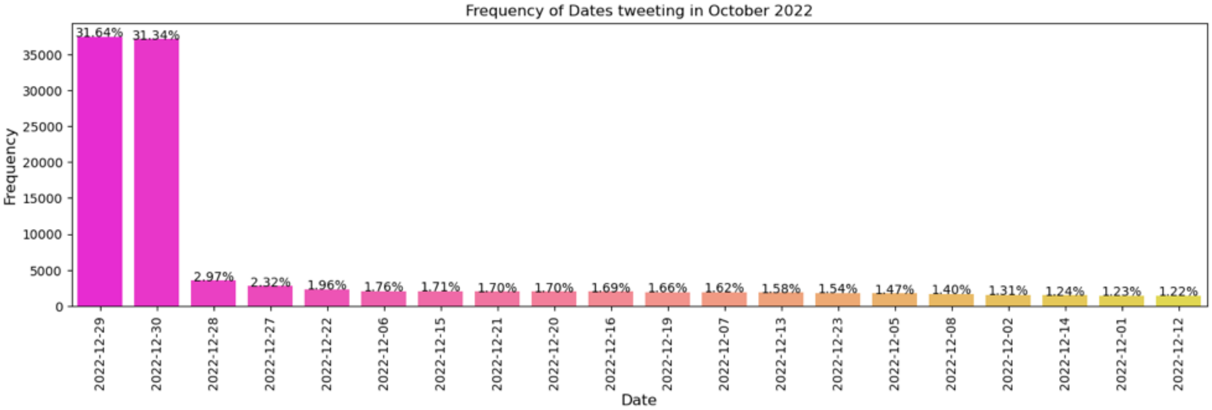
The %_change_Volume shows the most positive correlation with the volume feature. This means that the greater the difference between the volume of the stock today and the volume of the stock price of yesterday, the greater will be the stocks traded that day. As there is no information about the number of buyers and sellers, we can only guess that a high difference in volume might attract more buyers. Whereas a fewer difference may attract more sellers. With this in mind, we know to focus on days with high percentage change in trading volume.

Now to actually understand the content of the tweets which people tweet, we created a Word Cloud of the top most words used in the tweets, and some plots representing frequency of these tweets.



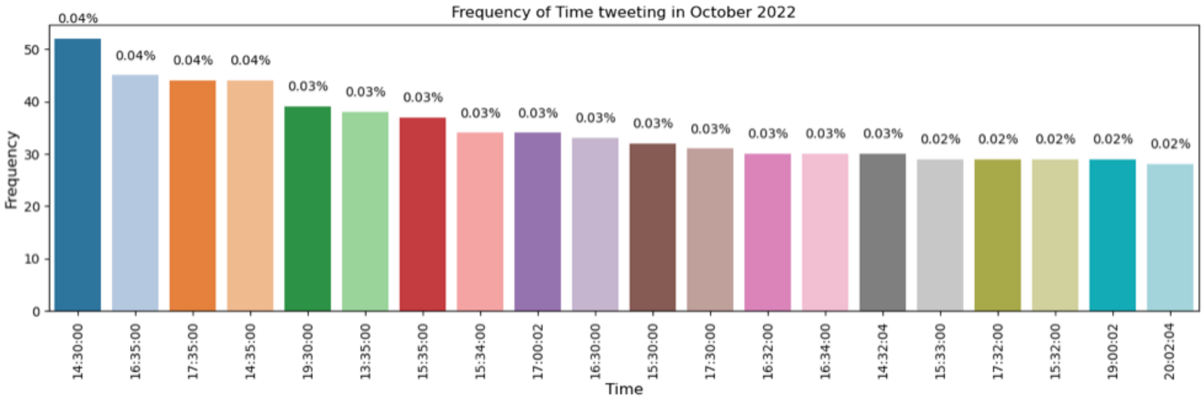


1.32% of the tweets were posted by user Smith28301

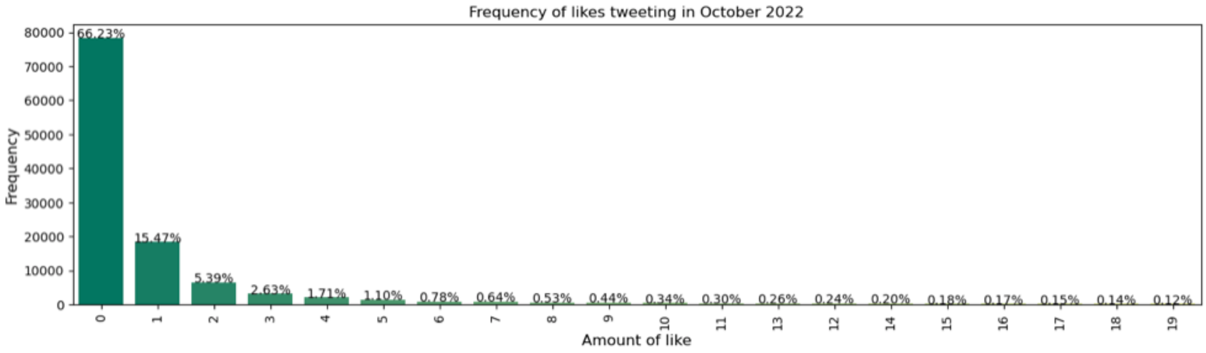


Over 60% of the tweets were posted on December 29th, and December 30th, 2022 as traders were anticipating a Santa Rally. A "Santa rally" is a term used to describe a phenomenon in which the stock market experiences a rise in stock prices in the final weeks of the calendar year, typically in the last week of December. The name "Santa rally" comes from the fact that it often coincides with the holiday season, and some traders believe that the surge in buying during this time is due to increased consumer spending and a generally positive mood.

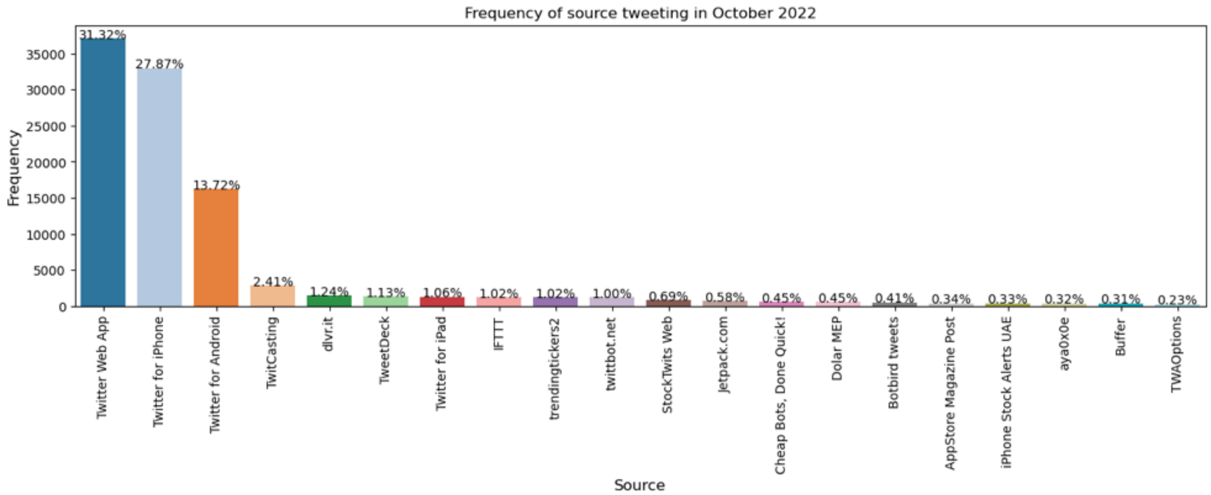
The Santa rally is not a guaranteed occurrence and is not supported by any particular economic theory. However, some analysts believe that the rally may be driven by a number of factors, including year-end tax planning, window dressing by fund managers, and optimism about the coming year.



Most tweets were published around 2:30pm to 2:35pm UTC (6:00am to 6:30am PST)



Over 65% of tweets received 0 likes and about 15% received 1 like



Over 30% of tweets were published using the Twitter app

## Pre-processing & Training Data

The objective of this notebook is to prepare data for fitting models. In notebook 02_Exploratory_Data_Analysis, we have already featured new columns for the stock data, measuring the changes each day. In this notebook, we will be using word embedding to convert words to vectors, then use that as the main input to our training set. Word Embedding is a language modeling technique used for mapping words to vectors of real numbers. It represents words or phrases in vector space with several dimensions.

## Modeling

## Takeaways

## Future Research