Tesla Stock Price Prediction

Data and Source

The project uses three datasets: 'Elon_Tweets_2011_2021.csv', 'Tesla_Stock_2014_2023.csv', and 'Dogecoin 2017-2024.csv' from Kaggle.com. The first dataset contains tweets from Elon Musk, the second one has data on Tesla's stock prices, and the third one includes data on Dogecoin.

Insight into the Datasets

- The Elon_Tweets_2011_2021.csv dataset contains 6 fields namely date, time, tweet, replies_count, retweets_count, and likes_count.
- The 'Tesla_Stock_2014_2023.csv' dataset includes opening, highest, lowest, and closing prices, along with trading volume. Momentum indicators like RSI and CCI, moving averages (SMA and EMA), MACD, Bollinger Bands, and ATR provide insights. The dataset aims to predict future stock movements.
- The 'Dogecoin 2017-2024.csv' dataset contains 7 fields namely Date, Open, High, Low, Close, Adj Close, Volume.

Data Exploration, Cleaning and Preprocessing Steps

1. **Data Inspection**: Inspected the datasets for completeness of information like missing values or nulls, duplicates, accuracy, and consistency of information across all the tweets, focusing on key metrics such as tweet. The
2. **Data Cleaning**: The data seems to be clean without any null or duplicates. But it needed some transformations.
3. **Data Preprocessing**: In our analysis we used all the three datasets, so we did some data preprocessing to all of those. To prepare the text data, we followed several preprocessing steps as below:

    - Removed punctuation, special characters, URLs, and numbers.
    - Eliminated whitespace and stop words.
    - Performed tokenization and removed any empty tokens.
    - Converted the text to lowercase.
    - Applied lemmatization.

    To prepare the numeric data, we followed the steps below:

- Change the date fields to datetime.
- Created new columns like year, month, day, daily returns for further calculations.
- Created subset datasets exclusive to Dogecoin tweets for analysis.

Research Questions:

1. What was the effect of the Covid pandemic on Tesla's stock value? & Do Tesla's stock prices and Dogecoin's market value exhibit similar responses to Elon Musk's tweets? Alternatively, does the market's response to Dogecoin affect the share price of Tesla?
2. How will Tesla stock prices trend in the future? Construct a predictive model.
3. How do Elon Musk's tweets affect stock market Movements? Can this sentiment be measured though sentiment analysis?

Program Description:

The Python program designed for analyzing the three files is structured to perform comprehensive data analysis tasks. The program is composed of several modules, each responsible for a specific part of the analysis process.

Modules:

1. Data Loading: Utilizes pandas to load the CSV file into a Data Frame for easy manipulation.
2. Data Exploration: Inspecting metadata using various methods like info(), describe() to understand the structure of dataset, its data types, and shape.
3. Data Cleaning: Implements functions to clean and preprocess the data, including handling missing values, standardizing text, and removing unwanted columns.
4. Data Visualization: Created visualizations to explore Elon Musk's tweet activity, top tweets, and total engagement over time. It also provides insights into Tesla stock trends, including closing prices, volume, and daily returns. Additionally, it offers visualizations for Dogecoin's price and trading volume. These visuals help our team gain insights for analysis and decision-making.
5. Data Preprocessing: Conversion of date columns to datetime format, addition of columns for year, month, day, and daily returns in the Tesla stock data, performed text preprocessing for Elon tweets by removing punctuation, special characters, URLs, and numbers. We also eliminated whitespace, stop words, and empty tokens, converted the text to lowercase, and applied lemmatization.

6. Data Modeling: Performed Binary Classifier model, Prophet time series, sentiment analysis, trend analysis to determine answers for various research questions.
7. Reporting: Generates an output file for the above models to be stored and referred to for later use.

.

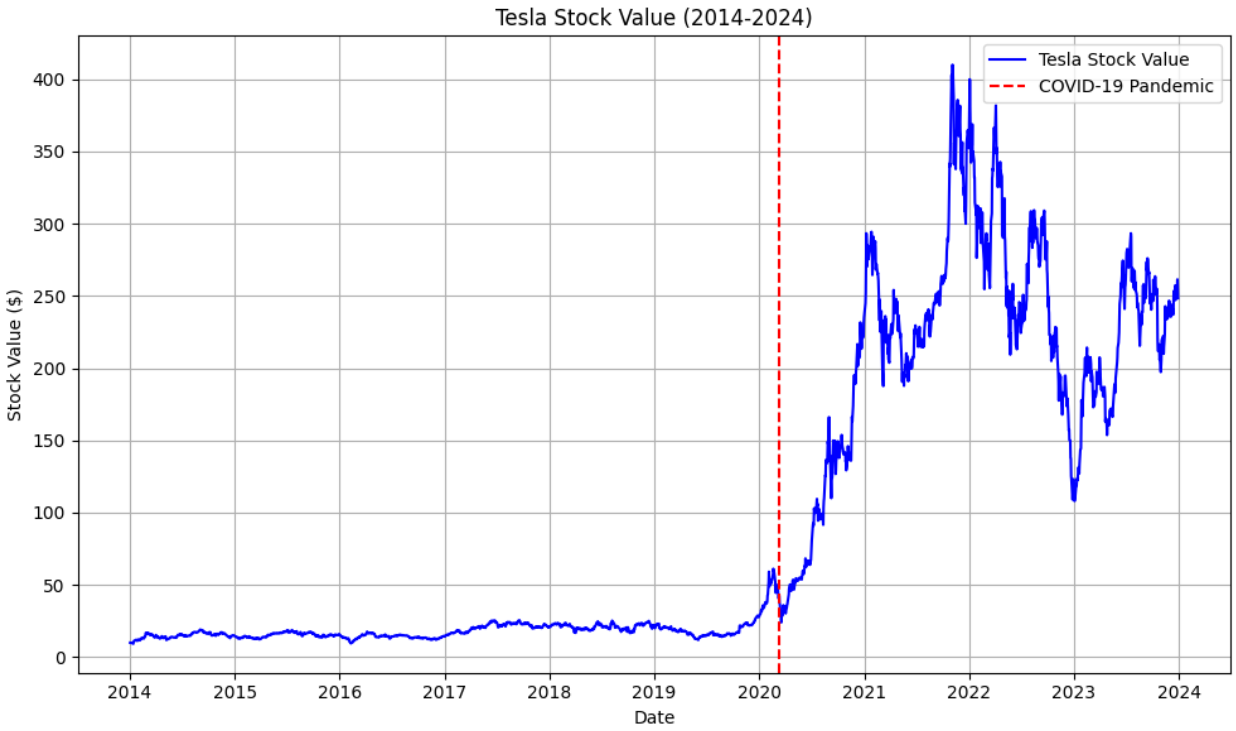
Output and Analysis: Delving deep into these datasets, here are the results for our research questions.

1) **What was the effect of the Covid pandemic on Tesla's stock value? & Do Tesla's stock prices and Dogecoin's market value exhibit similar responses to Elon Musk's tweets? Alternatively, does the market's response to Dogecoin affect the share price of Tesla?**

   Covid-19 pandemic had a significant impact on financial markets creating a roller-coaster ride of uncertainty and resilience. Amid global uncertainty, Tesla's stock surged significantly, defying expectations. Let's delve into Tesla's stock surge and examine the influence of Elon Musk's tweets, as well as the role of Dogecoin in this scenario.

   To analyze the pandemic's impact on Tesla, the program utilizes data from the first 10 years of Tesla's dataset. It then plots a trend analysis, incorporating a Covid marker line to highlight the impact.

   - **Tesla Stock Performance During the Pandemic: A Trend Analysis**

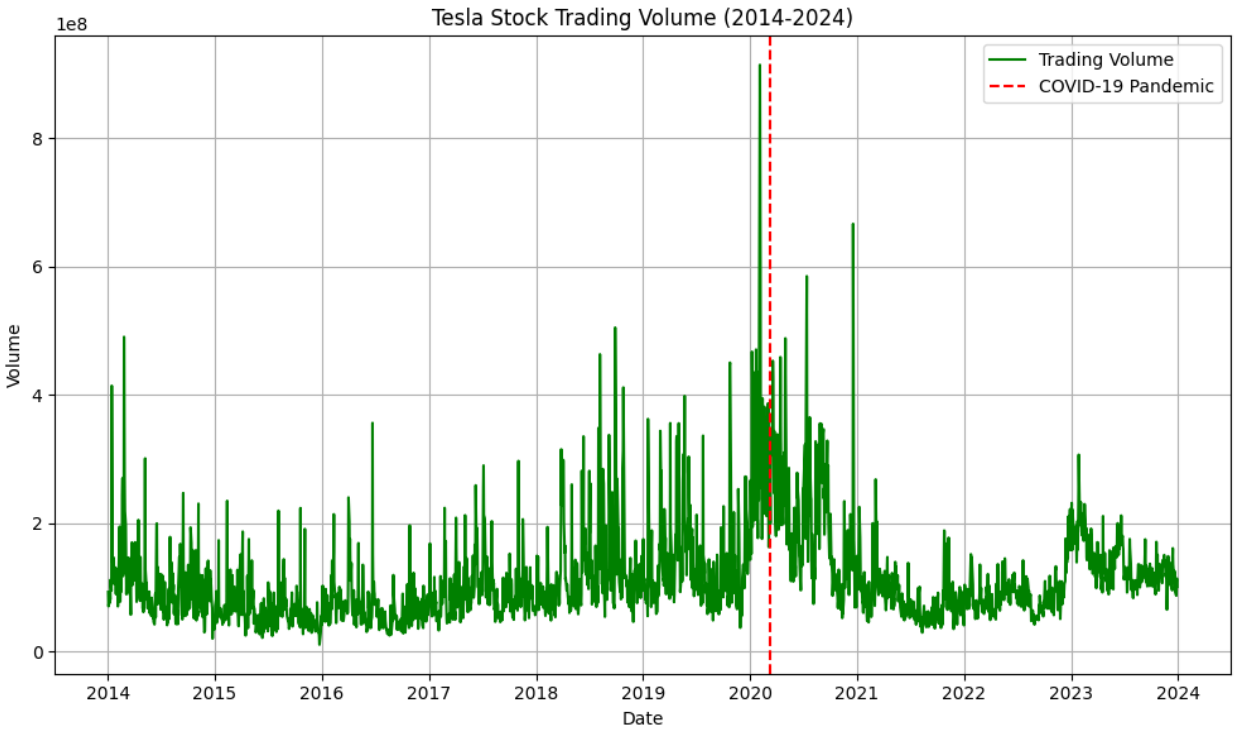


Overall Trend: The stock value starts at a lower point in 2014 and remains relatively stable until around late 2019. Starting around early 2020, there is a sharp increase in stock value. The value peaks around 2021 before displaying volatility with several peaks and troughs.

Covid-19 Impact: The red dashed vertical line represents the start of the Covid-19 pandemic (around 2020). Despite initial fluctuations, Tesla's stock value surged during the pandemic.

Post-Pandemic Recovery: After reaching a peak in late 2021, the stock value showed fluctuations. It remained above pre-pandemic levels, indicating sustained recovery.

- **Analyzing Tesla Stock Volume During the Pandemic**



Overall Trend: The green line shows the trading volume over time. From 2014 to 2016, the volume remained relatively low. Around 2017, there was a significant increase in trading volume. The period from 2017 to 2020 saw higher volatility in volume.

Covid-19 Impact: The red dashed vertical line around 2020 represents the start of the Covid-19 pandemic. Around this time, there was a sharp spike in trading volume, possibly due to market uncertainty and investor reactions.

Post-Pandemic Recovery: After the initial surge, trading volume gradually stabilized. From 2021 to 2024, the volume remained relatively consistent.

- **Analyzing Tesla Stock Opening Prices During the Pandemic**

Overall Trend: The opening price shows significant volatility over the years. From 2014 to 2016, the opening price remained relatively stable. Around 2017, there was an upward trend, leading to higher opening prices.

Covid-19 Impact: The red dashed vertical line around 2020 represents the start of the Covid-19 pandemic. Despite initial fluctuations, Tesla's opening price surged during the pandemic.

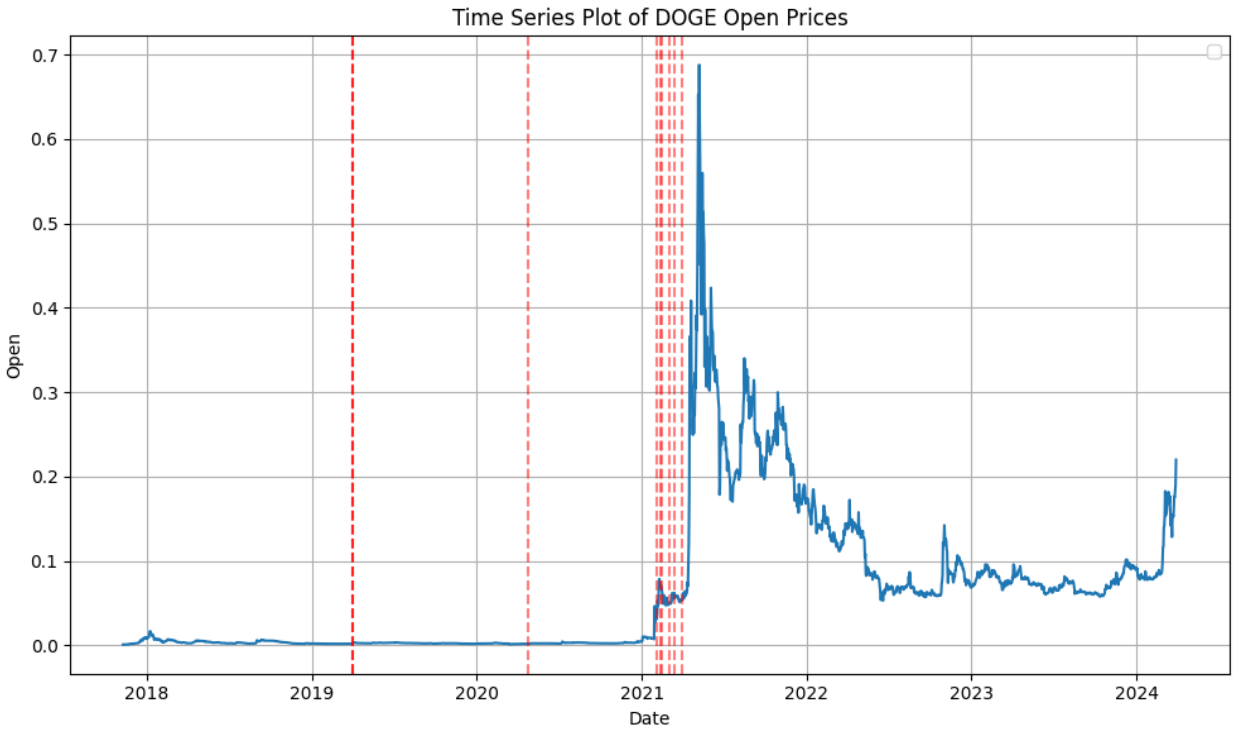
Post-Pandemic Recovery: After reaching a peak in late 2021, the opening price showed fluctuations. It remained above pre-pandemic levels, indicating sustained recovery.

In the subsequent part of the analysis, we incorporate the Dogecoin dataset and Elon Musk's tweets alongside our original data. We begin by filtering tweets that mention Dogecoin, creating a separate dataframe for our analysis. Next, we plot Tesla's stock price in conjunction with Elon's tweets to observe trends. Additionally, we analyze the Dogecoin time series, specifically focusing on the impact of Elon Musk's tweets.

- **Tesla Stock Price Trends in Relation to Elon Musk's Tweets**

Elon Musk's Tweets and Tesla Price: Whenever Elon Musk tweets about Dogecoin, it tends to have a downside impact on the tesla prices. The dip in the Tesla price coincides with Musk's tweets, effectively making his tweets impact the stock price.

- **Dogecoin Price Trends in Relation to Elon Musk's Tweets**



Elon Musk's Tweets and DOGE Price: Whenever Elon Musk tweets about Dogecoin, it tends to have a significant impact on the cryptocurrency's price. The spikes in the DOGE price coincide with Musk's tweets, effectively making his tweets act as trading signals for DOGE holders.

Price Movement: The DOGE price surges sharply after Musk's tweet, sending the cryptocurrency's supporters into a flutter. We see a spike in early 2021 after a series of tweets.

Elon Musk's influence on Dogecoin's price is evident, and his tweets have become a recurring trend affecting the cryptocurrency market.

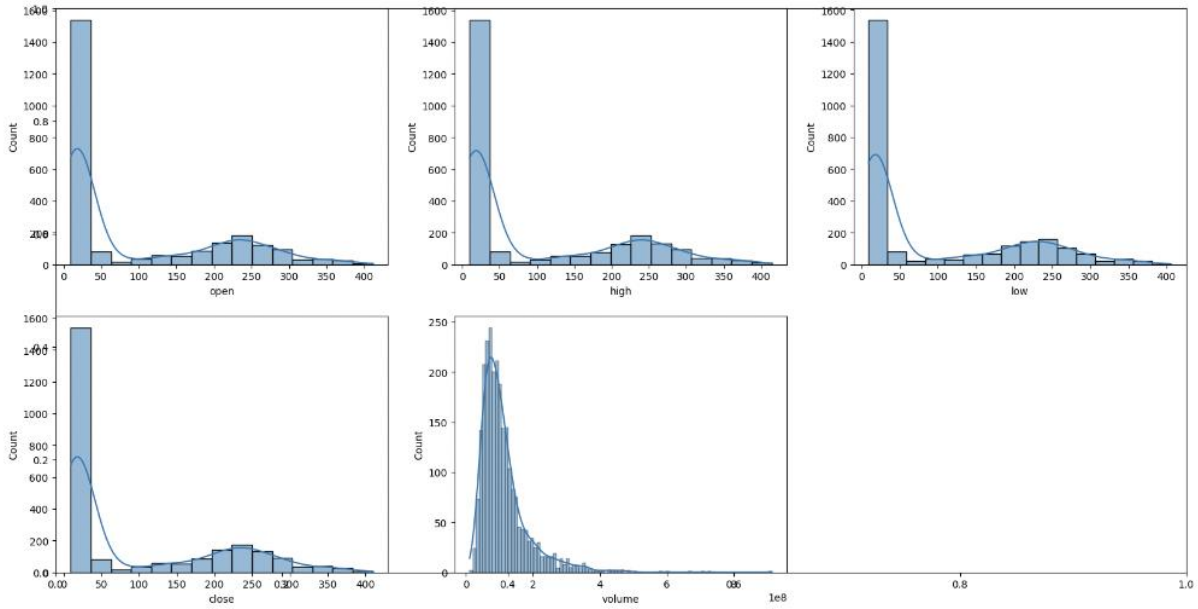
.

2) **How will Tesla stock prices trend in the future? Construct a predictive model.**

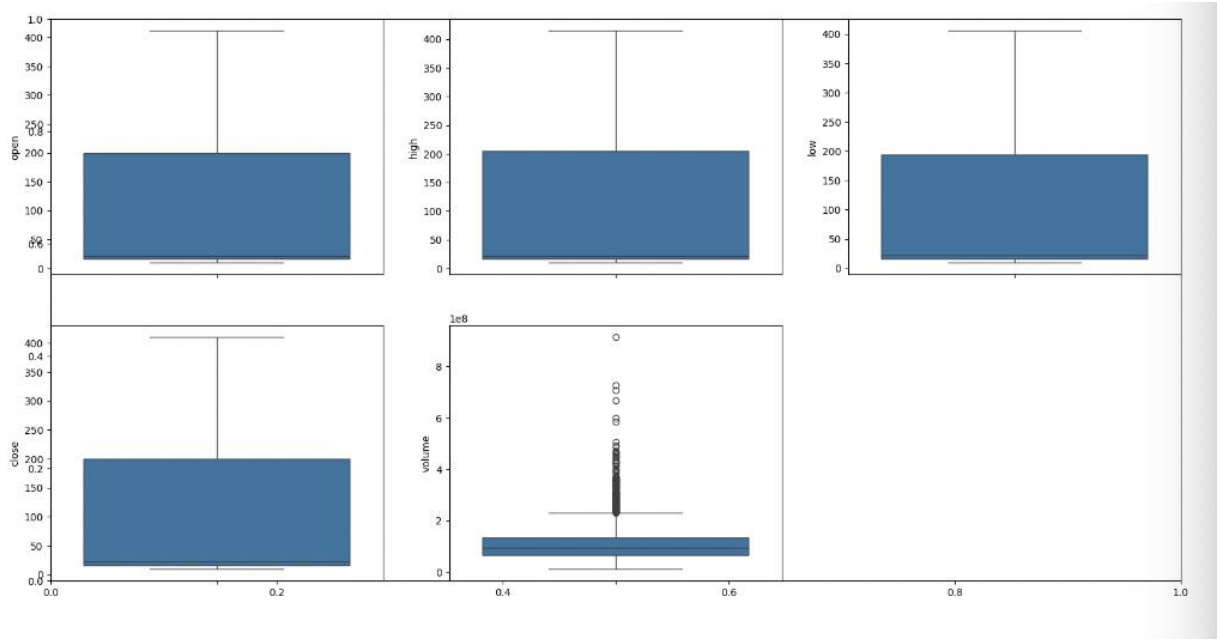
Predicting stock value can be a tricky task, especially for volatile stock like Tesla whose share price has fluctuated more than 25% for five of the last eight quarters. Forbes even named Tesla the "biggest S&P 500 loser" of 2024 due to year over year declining profits, emerging competitors in the electric vehicle market such as Rivian, and Elon Musk's polarizing public image following his controversial Twitter takeover in late 2022.1.

Nevertheless, machine learning models can provide insight into binary “to buy or not buy” decision making using previous years’ data, while more sophisticated models can even build future timeseries trendlines.

- **Binary Classifier Models**: Before building prediction models, Exploratory Data Analysis was conducted on the Tesla Stock dataset to understand variable distributions and outliers that may disrupt the prediction models.

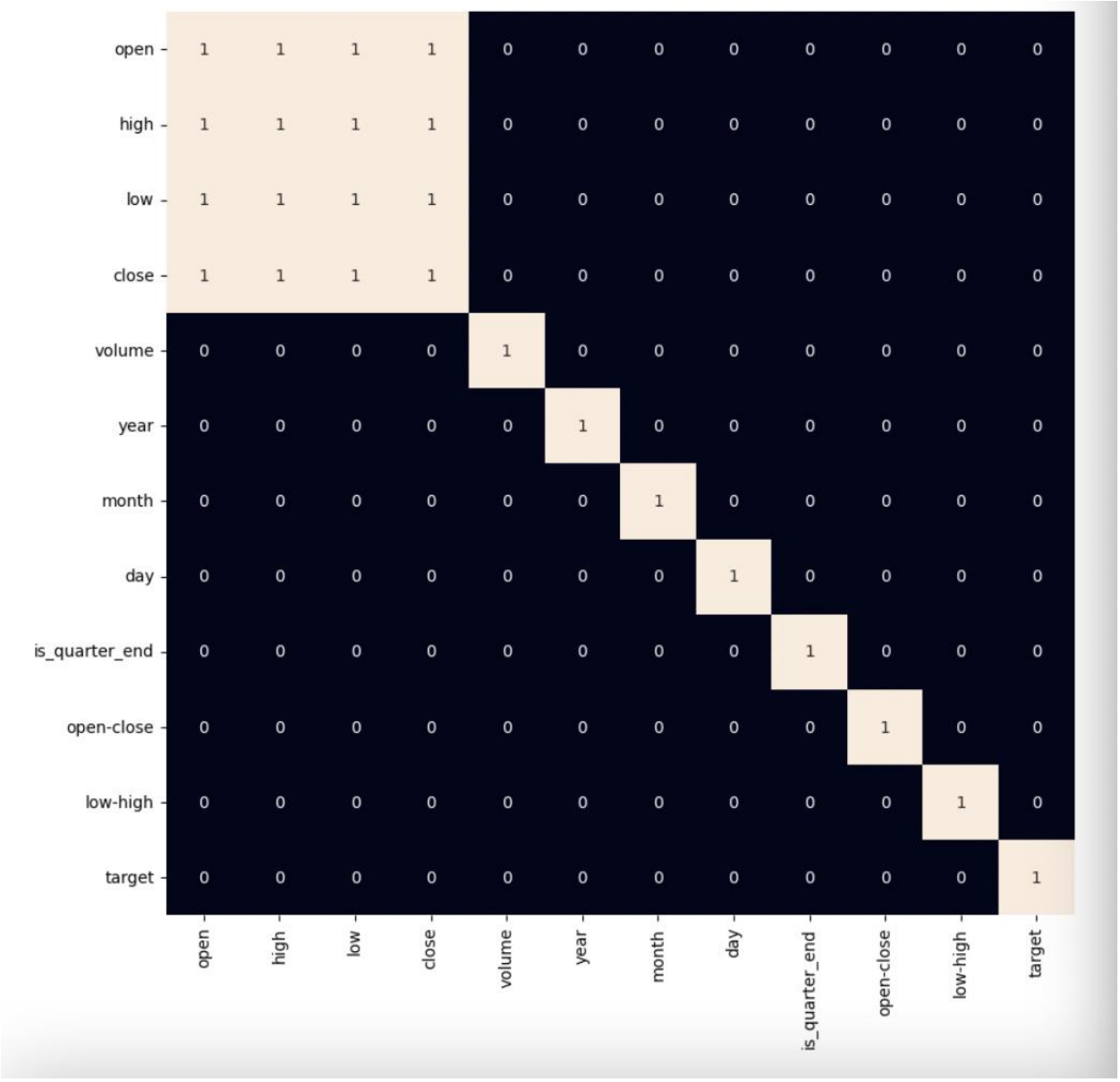


All variables (open, high, low, close, and volume) showed a right-side skew in the data distribution. In other words, the mean of each variable was higher than the majority of data observations for each attribute.



A simple box plot chart for each variable showed a similar result. Most notably, the “volume” variable had many outliers above the third quartile.

Next, variables were plotted on a correlation heatmap to ensure no collinearity between variables would disrupt the model.



As expected, the strongest collinearity was between open, high, low and close for each data observation.

Finally, additional variables were added to the dataset for analysis, including a binary variable “is_quarter_end”, and numeric variables “open-close” and “low-high” to be used for the classifier model training.

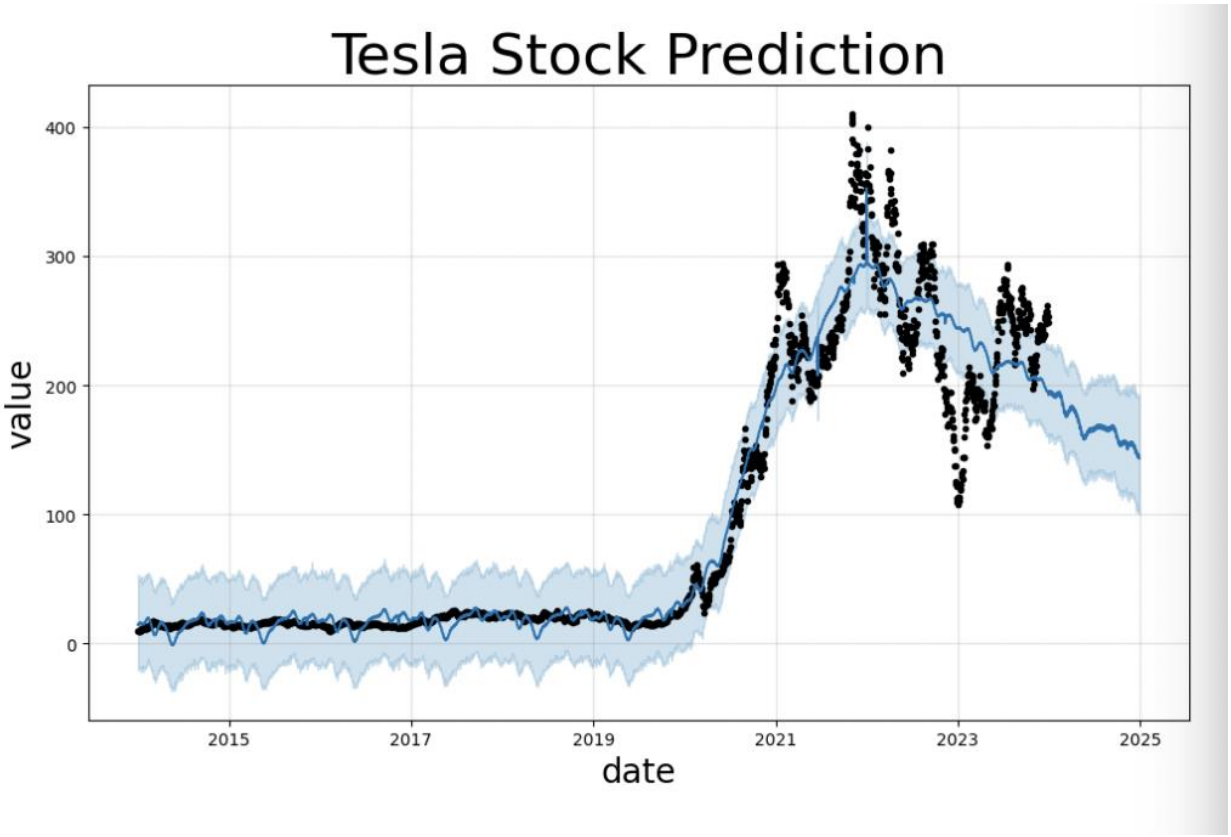
After the initial data analysis was complete, three binary classifier models (Logistic Regression, SVC, and XGBoost Classifier) were trained to determine whether to buy based on previous year’s data from the Tesla 2014-2023 stock value dataset. All three

models used daily open/close spread, daily low/high spread, and end-of-quarter data as input variables, trying to predict a binary output target variable of whether to buy.

Model	Logistic Regression	SVC (Support Vector Classification)	XGBoost Classifier
Training Accuracy	51.50%	49.67%	93.33%
Test Accuracy	53.38%	51.97%	54.34%

Ultimately, the models performed at par or only slightly better than a 50-50 guess. The XGBoost Classifier, while performing the best, was severely overfitted to the training data, resulting in a large accuracy discrepancy between training and test results. To improve the accuracy of the models, more data should be included in the model and measures should be taken to prevent model overfitting on training data.

- **Prophet Timeseries Forecasting**: In addition to the binary classifier models, a timeseries forecast plot was built using Meta tool “Prophet”. Prophet is a forecasting tool for non-linear time series data with strong seasonality. It functions well with missing data and handles outliers well, which is useful for the “volume” attribute of the Tesla dataset.



The Prophet results showed an increasing downward trend for Tesla Stock value, which is aligned with current reports of Tesla's decreasing market value through 2024 and beyond. However, given the volatile nature of Tesla's stock, the predictions will only be proven or disproven with time.

3) How does Elon Musk's tweets affect Tesla Stock and Dogecoin value? Can this be measured though sentiment analysis.

Elon Musk's tweets have been known to influence financial markets, particularly the stock price of Tesla and the value of Dogecoin. This analysis examines the extent of this impact using sentiment analysis of all of Musk's tweets and subsequent statistical analysis to determine correlations between tweet sentiments and market movements. The steps include exploratory data analysis (EDA) on Elon Musk's Twitter activity, data cleaning, and merging multiple data sources into a comprehensive dataset for analysis.

Data Collection and Cleaning

1. **Elon Musk's Tweets**:
    - A comprehensive dataset of Elon Musk's tweets was obtained from Kaggle via https://www.kaggle.com/datasets/andradaolteanu/all-elon-musks-tweets

```
Total Engagement: 306590084
Total Number of Tweets: 12562
Total Likes: 269706831
Total Replies: 8534246
Total Retweets: 28349007
Number of Rows: 12562
Number of Columns: 10
Date Range: 2010-06-04 00:00:00 to 2021-04-17 00:00:00
```

    - The tweet data was cleaned by removing retweets, replies, non-English tweets, URLs, hashtags, and mentions. This preprocessing ensured that the dataset contained only relevant and clean text for sentiment analysis.

```python
# Function to clean text
def clean_text(text):
    text = re.sub(r'\n', '', text)
    text = re.sub(r'\xa0', '', text)
    text = re.sub(r'\t', '', text)
    text = re.sub(r'http\S+', '', text)
    text = re.sub(r'https\S+', '', text)
    text = re.sub(r'pic.twitter\S+', '', text)
    text = re.sub(r'#\S+', '', text)
    text = re.sub(r'@\S+', '', text)
    text = remove_emojis(text)  # Remove emojis
    text = re.sub(r'[^\w\s]', '', text)  # Remove all punctuation
    text = text.strip()
    text = ' '.join(text.split())
    text = text.lower()
    return text

# Function to remove stopwords
def remove_stopwords(text):
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(text)
    filtered_text = ' '.join([word for word in word_tokens if word not in stop_words])
    return filtered_text
```

    - Sentiment analysis tool TextBlob was applied to each clean tweet to generate sentiment scores, classifying tweets as positive, negative, or neutral.
    - For instance, a tweet with the text "love this beautiful shot" received a high sentiment score of 0.675, indicating a positive sentiment.

```
        date      time                                tweet  replies_count  \
0 2021-04-11  18:50:33  for now. costs are decreasing rapidly.            640
1 2021-04-11  18:48:58               love this beautiful shot           2464
2 2021-04-11  17:49:38                        trust the shrub            115
3 2021-04-11  15:23:49      the art in cyberpunk is incredible           8437
4 2021-04-11   9:18:47                                  🤣🤣            446

   retweets_count  likes_count  total_engagement  year  \
0             444        15281             16365  2021
1            1517        71161             75142  2021
2              48         1380              1543  2021
3           10329       228144            246910  2021
4             542         7489              8477  2021

              cleaned_tweet  sentiment
0  costs decreasing rapidly      0.000
1       love beautiful shot      0.675
2               trust shrub      0.000
3  art cyberpunk incredible      0.900
4                                0.000
```

2. **Market Data**:
    - Historical stock prices for Tesla and Dogecoin prices were also obtained.
    - Both datasets were standardized to a consistent format and time zone for accurate alignment and analysis.
3. **Merging Data**:
    - Tweets were aligned with market data based on date and time.
    - Daily and weekly aggregation of tweets was performed to match the frequency of stock and cryptocurrency data, facilitating a coherent analysis.

```python
# Calculate weekly percentage change for Tesla and Dogecoin
tesla_weekly_percentage_change = tesla_weekly_avg.pct_change() * 100
dogecoin_weekly_percentage_change = dogecoin_weekly_avg.pct_change() * 100

# Ensure all arrays have the same length
length = min(len(tesla_weekly_avg), len(dogecoin_weekly_avg), len(tesla_weekly_change), len(dogecoin_weekly_change),
             len(tesla_weekly_percentage_change), len(dogecoin_weekly_percentage_change))

# Create a new DataFrame with weekly averages, changes, and percentage changes
weekly_summary_df = pd.DataFrame({
    'Date': tesla_weekly_avg.index[:length],
    'Tesla_Weekly_Avg_Close': tesla_weekly_avg.values[:length],
    'Dogecoin_Weekly_Avg_Close': dogecoin_weekly_avg.values[:length],
    'Tesla_Weekly_Price_Change': tesla_weekly_change.values[:length],
    'Dogecoin_Weekly_Price_Change': dogecoin_weekly_change.values[:length],
    'Tesla_Weekly_Percentage_Change': tesla_weekly_percentage_change.values[:length],
    'Dogecoin_Weekly_Percentage_Change': dogecoin_weekly_percentage_change.values[:length]
})

# Print the new DataFrame
print(weekly_summary_df.head())
```

Exploratory Data Analysis (EDA)

2. **Tweet Activity and Engagement**:
    - An intensive EDA was conducted on Elon Musk's Twitter activity, analyzing tweet frequency, engagement metrics (likes, retweets, replies), and their distribution over time.

Engagement Metrics:

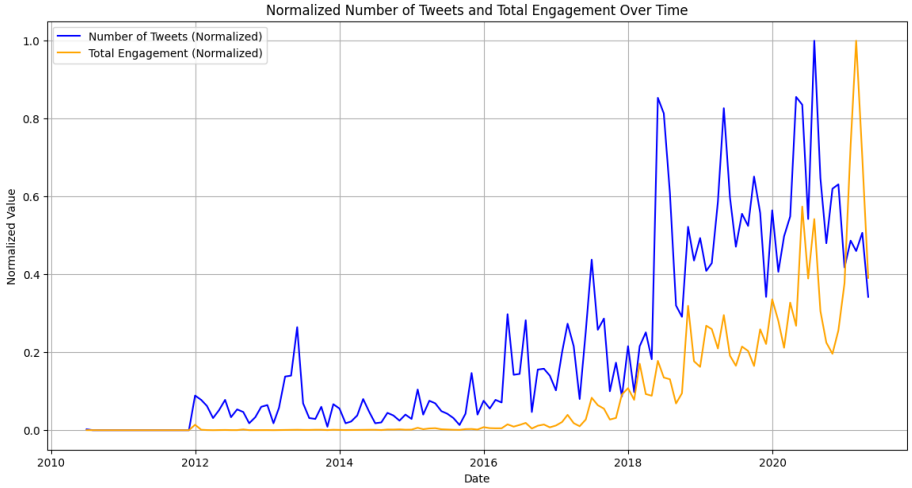
- Tweets with positive sentiment had an average length of 85.67 characters, while those with negative sentiment averaged 101.24 characters.
- The average total engagement for tweets varied significantly, with tweets like "the art in cyberpunk is incredible" reaching high engagement levels (246,910 total engagements).

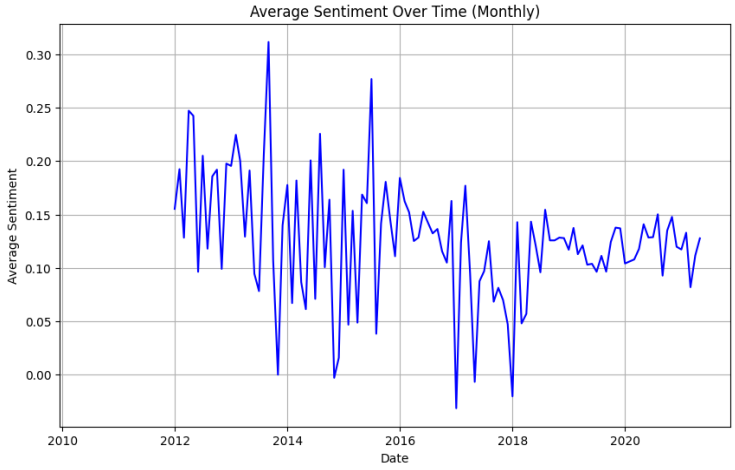
```
Average Tweets per Day: 3.16
Average Tweets per Month: 96.63
Average Tweets per Year: 1046.83
```

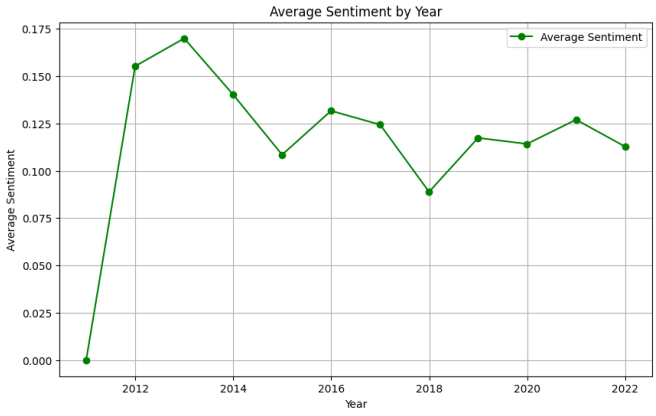
```
Top 10 Months of Positive Sentiment:
date
2013-08-31    0.311496
2015-06-30    0.276817
2012-03-31    0.247171
2012-04-30    0.242375
2014-07-31    0.225532
2013-01-31    0.224583
2013-07-31    0.205647
2012-06-30    0.205000
2014-05-31    0.200665
2013-02-28    0.200108
```

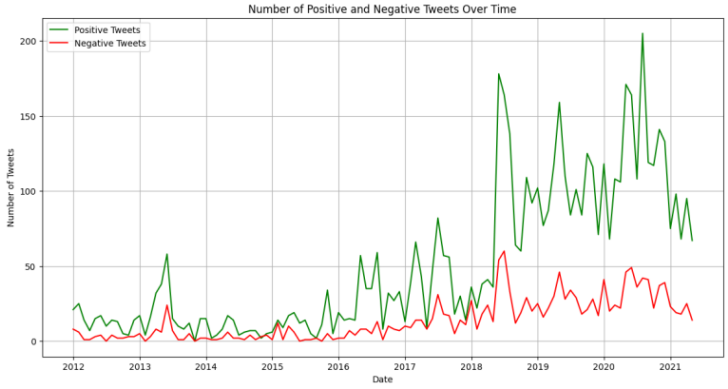
```
Top 10 Months of Negative Sentiment:
date
2016-12-31   -0.031371
2017-12-31   -0.020239
2017-04-30   -0.006690
2014-10-31   -0.002879
Name: sentiment, dtype: float64
```

```
Average length of positive tweets: 85.571
Average length of negative tweets: 103.53
```


Normalized Number of Tweets and Total Engagement Over Time
Number of Tweets (Normalized)
Total Engagement (Normalized)
Normalized Value
Date
0.0
0.2
0.4
0.6
0.8
1.0
2010
2012
2014
2016
2018
2020

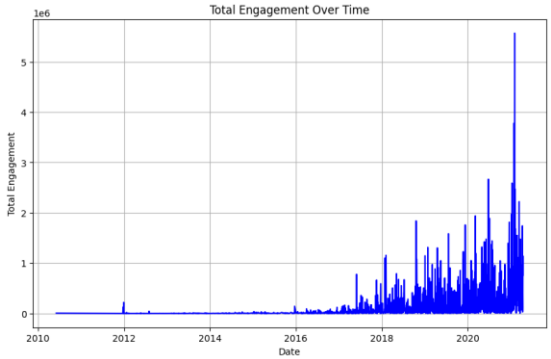

Average Sentiment Over Time (Monthly)
Average Sentiment
Date
0.00
0.05
0.10
0.15
0.20
0.25
0.30
2010
2012
2014
2016
2018
2020


Average Sentiment by Year
Average Sentiment
Average Sentiment
Year
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
2012
2014
2016
2018
2020
2022



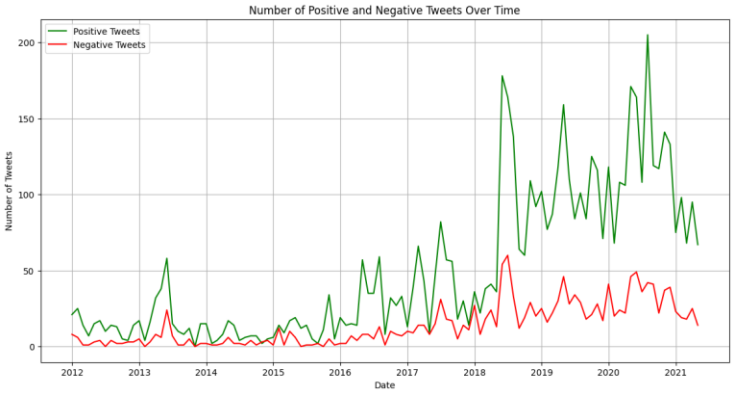
- New columns for engagement metrics were created to quantify the influence and reach of each tweet. For example, on April 11, 2021, a tweet stating "the art in cyberpunk is incredible" garnered 8,437 replies, 10,329 retweets, 228,144 likes, and had a sentiment score of 0.900.

```
# Calculate total engagement by summing replies, likes, and retweets
elon_tweets_df['total_engagement'] = elon_tweets_df['replies_count'] + elon_tweets_df['retweets_count'] + elon_tweets_df['likes_count']
 (variable) total_engagement_over_time: Series[Any]
total_engagement_over_time = elon_tweets_df.groupby('date')['total_engagement'].sum()

# Print the total engagement over time
print(total_engagement_over_time)

import matplotlib.pyplot as plt

# Plot total engagement over time
plt.figure(figsize=(10, 6))
plt.plot(total_engagement_over_time.index, total_engagement_over_time.values, color='b')
plt.title('Total Engagement Over Time')
plt.xlabel('Date')
plt.ylabel('Total Engagement')
plt.grid(True)
plt.show()
```



3. **Sentiment Analysis**:

- Sentiment analysis tool TextBlob was applied to each clean tweet to generate sentiment scores, classifying tweets as positive, negative, or neutral.
- For instance, a tweet with the text "love this beautiful shot" received a high sentiment score of 0.675, indicating a positive sentiment.



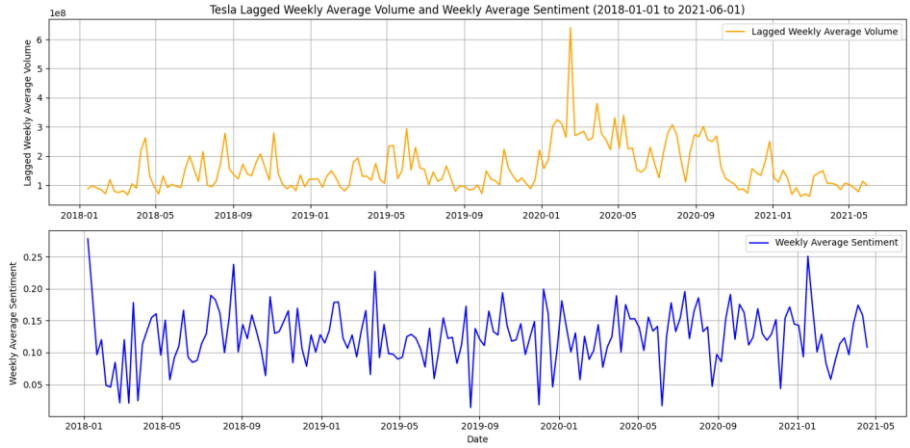
4. **Additional columns added**:
   - Additional columns were added for weekly average sentiment scores, engagement metrics, and percentage changes in stock and cryptocurrency prices. These features helped in capturing trends and their potential correlation with market movements.

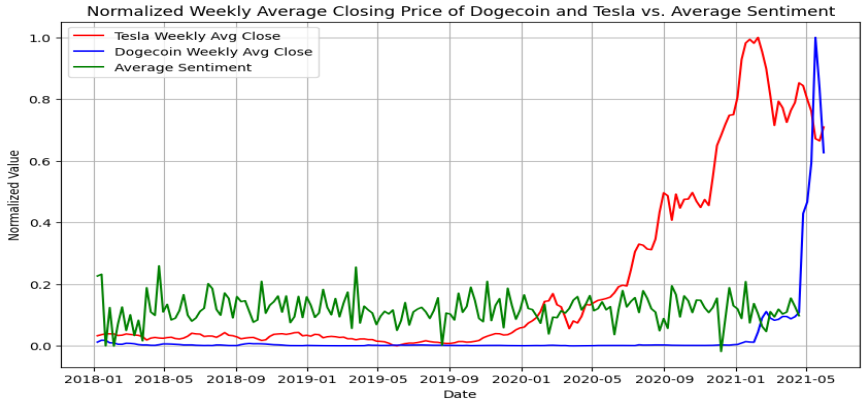
Analysis and Visualization

1. **Correlation Analysis**:

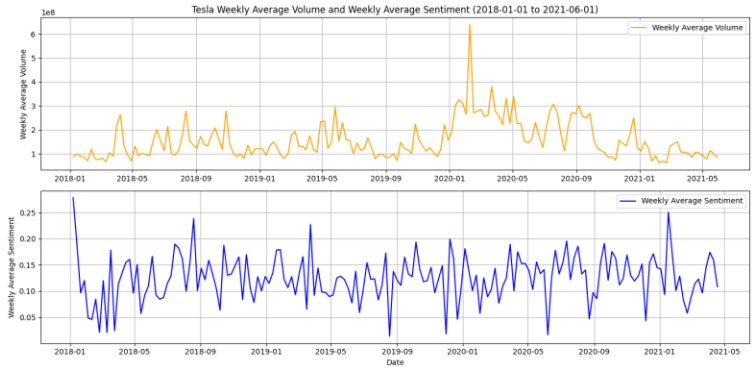
- Correlations between tweet sentiments, engagement metrics, and market price movements were computed. Both same-week and lagged correlations were calculated to explore immediate and delayed impacts.
- The correlation between previous week's sentiment and next week's Tesla price change was 0.128, suggesting a modest positive relationship. Conversely, the correlation between sentiment and same week's Dogecoin price change was -0.061, indicating a slight negative relationship.

```
The correlation between lagged average weekly volume and average weekly sentiment is: 0.12515
```

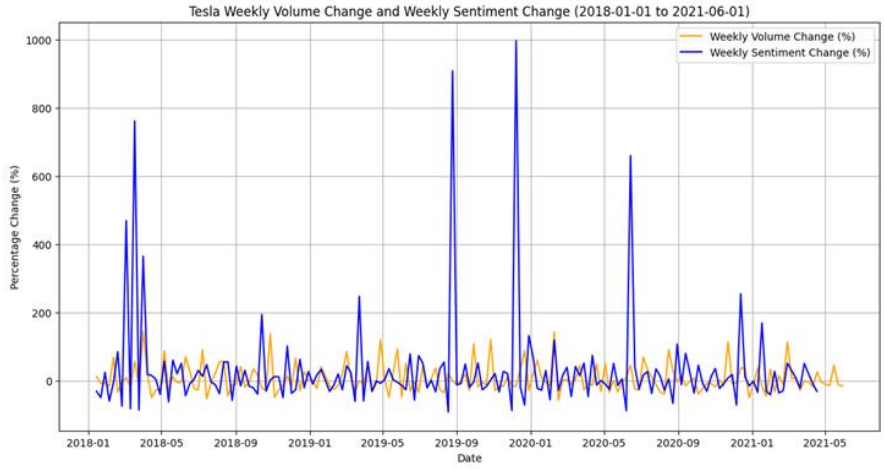


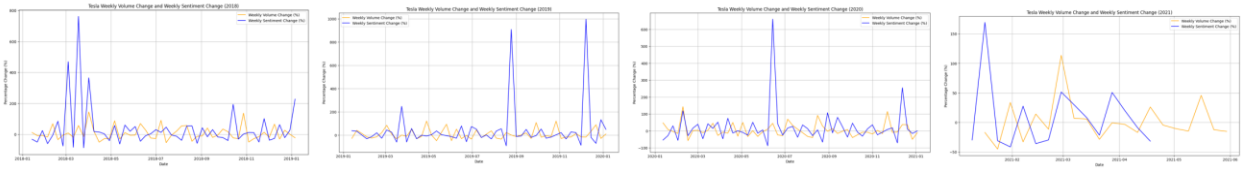


2. **Volume and Sentiment Correlations**:
   - The correlation between average weekly sentiment and average weekly volume for Tesla was 0.028, indicating a very weak positive relationship.
   - The correlation between lagged average weekly sentiment and average weekly volume was -0.086, suggesting a slight negative relationship.
   - The correlation between lagged average weekly volume and average weekly sentiment was 0.125, showing a modest positive relationship.



- The analysis shows that weekly volume changes closely follow weekly sentiment changes. Peaks in sentiment often coincide with spikes in trading volume, suggesting that sentiment in Elon Musk's tweets significantly influences investor behavior and trading activity. Analysis was also broken down annually and shows a very close relationship.





Tasks and roles of each member of the group:

- **Padmaja**: Worked on Research question 1, performed all the required programming steps like preprocessing, cleaning, and created the final project report.
- **Sarah**: Worked on Research question 2, performed all the required programming steps like preprocessing, cleaning, created the final project presentation.
- **Sachi**: Worked on Research question 3, performed all the required programming steps like preprocessing, cleaning, merged all our code and created the final project jupyter notebook.

Final Conclusions about the data based on Results:

This analysis demonstrates that while there is some correlation between Elon Musk's tweet sentiments and the market movements of Tesla stock and Dogecoin, the relationships are generally weak. Sentiment analysis of tweets shows that Musk's tweets can have an immediate but often negligible impact on market prices and trading volumes. The findings suggest that Musk's tweets are slightly influential. Results also showed that during the early 2020 to 2021 increased tweets, twitter engagement occurred during the same time as Tesla Stock and Dogecoin prices increased significantly. This analysis can be improved if an hourly analysis is conducted, to see the immediate impact of Elon's tweets.

Reference:

All Elon Musk's Tweets. www.kaggle.com.
https://www.kaggle.com/datasets/andradaolteanu/all-elon-musks-tweets

Dogecoin Historical Data. www.kaggle.com. Accessed June 15, 2024.
https://www.kaggle.com/datasets/dhruvildave/dogecoin-historical-data

Tesla (TSLA) Stock 2015 - 2024. www.kaggle.com. Accessed June 15, 2024.
https://www.kaggle.com/datasets/saadatkhalid/tesla-tsla-stock-2015-2024

Merge, join, concatenate and compare — pandas 2.2.2 documentation. (n.d.).
https://pandas.pydata.org/docs/user_guide/merging.html

Mulani, S. (2022, August 3). *Using StandardScaler() Function to Standardize Python Data*. DigitalOcean. https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python

NLTK :: Natural Language Toolkit. (n.d.). https://www.nltk.org/

Tutorial: Quickstart¶. Tutorial: Quickstart - TextBlob 0.18.0.post0 documentation. (n.d.). https://textblob.readthedocs.io/en/dev/quickstart.html

W3Schools.com. (n.d.). https://www.w3schools.com/python/pandas/default.asp