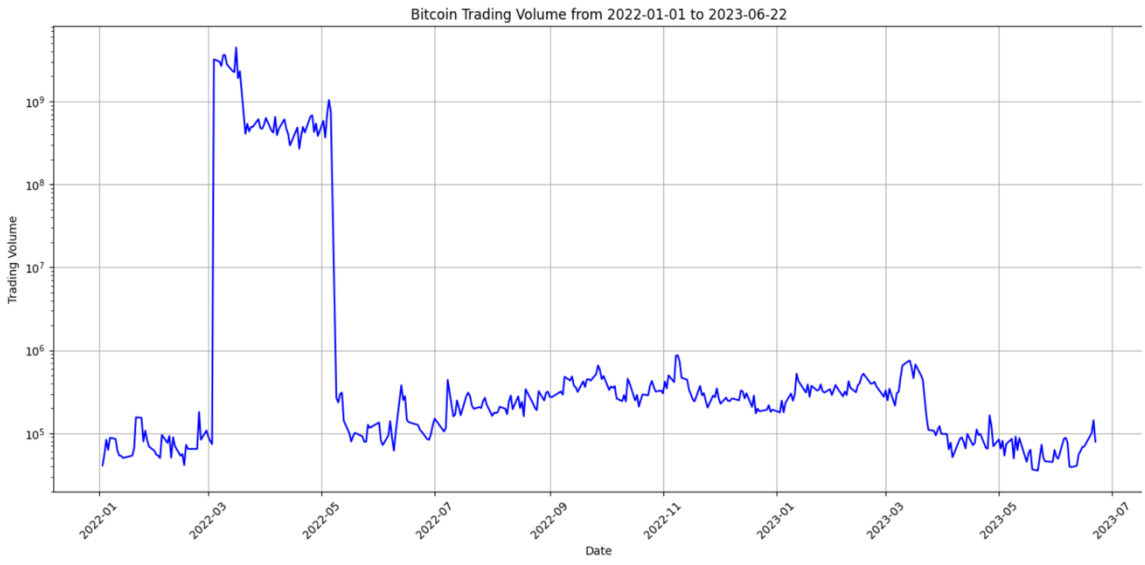
CS 210 Project Step:3

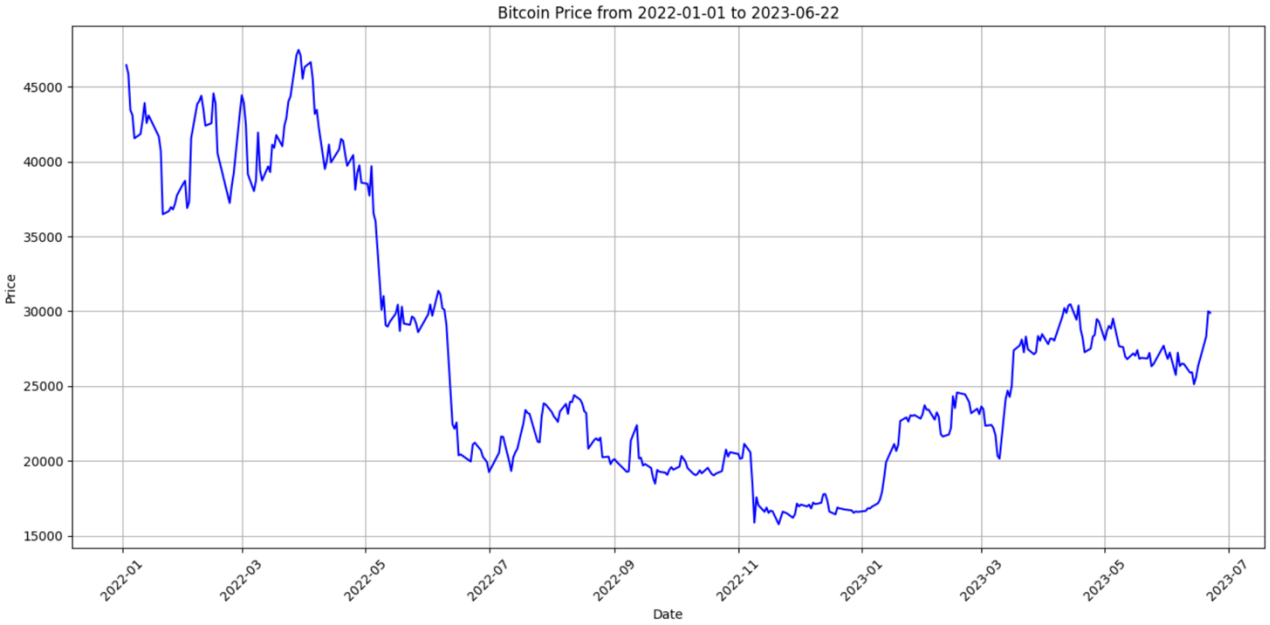
Yağız Uçar 31204

Stock Market Dataset:

Dataset includes the prices and trading volumes of various financial assets such as natural gas, crude, oil, copper, Bitcoin, Berkshire, Netflix, Amazon, Meta, and gold. It contains dates along with relevant prices and volume data. Financial analysis, academic research, behavioral finance, developing financial products and machine learning and data science applications could be given as an instance of possible areas of usage of this dataset.



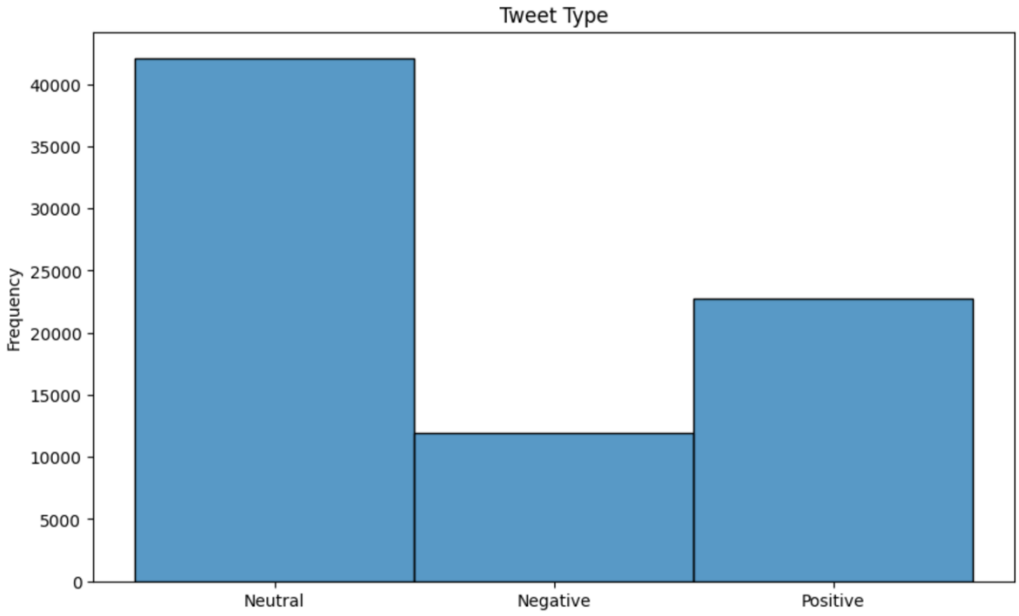
- The graph shows the changes in the Bitcoin trading volume from the dates between 2022/01/01 and 2023/06/22.



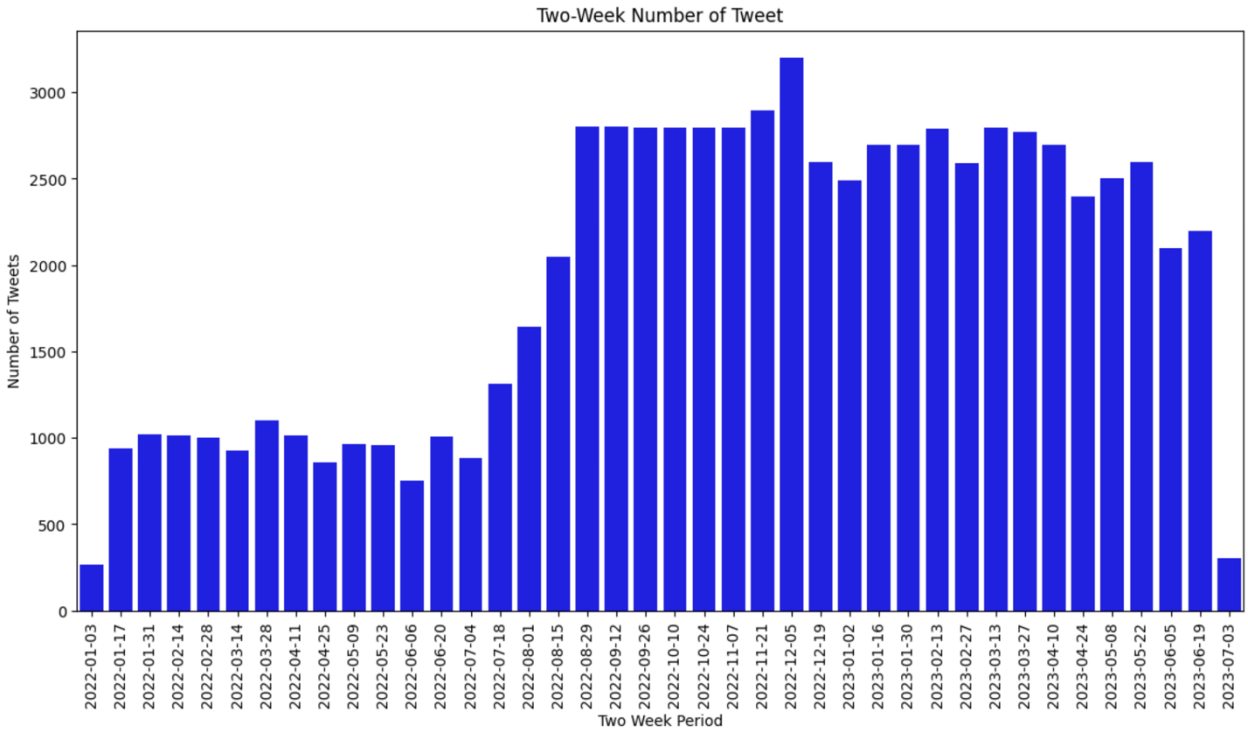
- The graph shows the changes in the Bitcoin prices from the dates between 2022/01/01 and 2023/06/22.

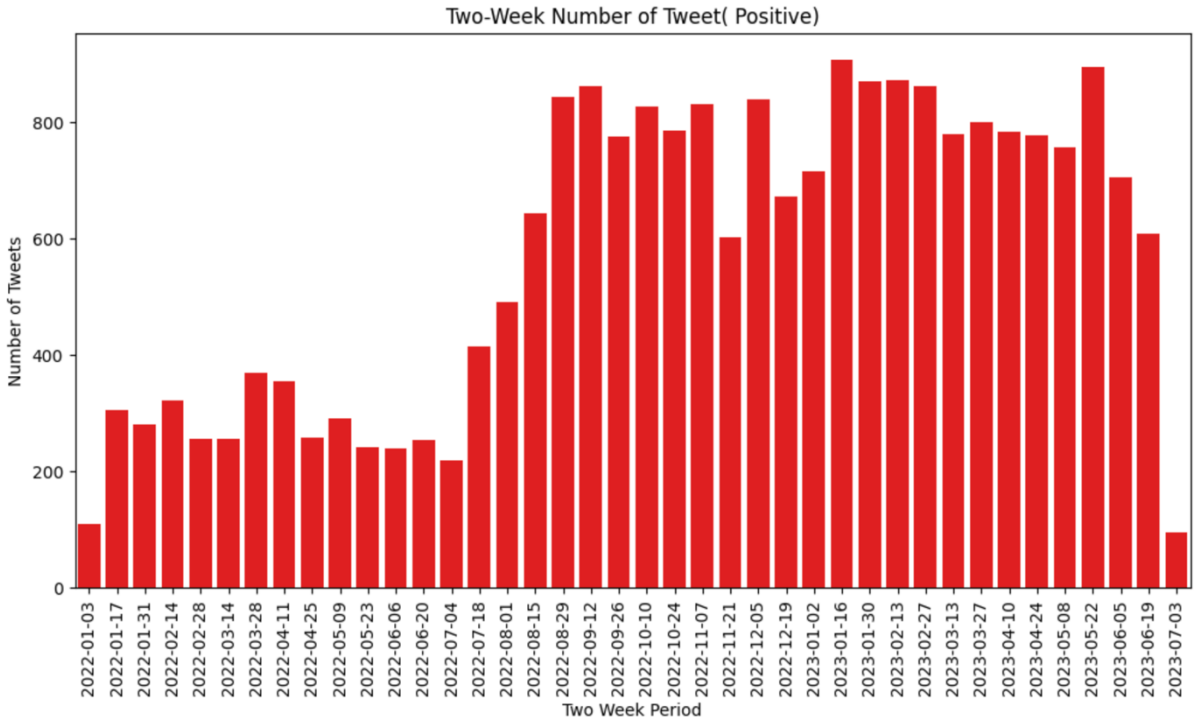
Social Media Dataset:

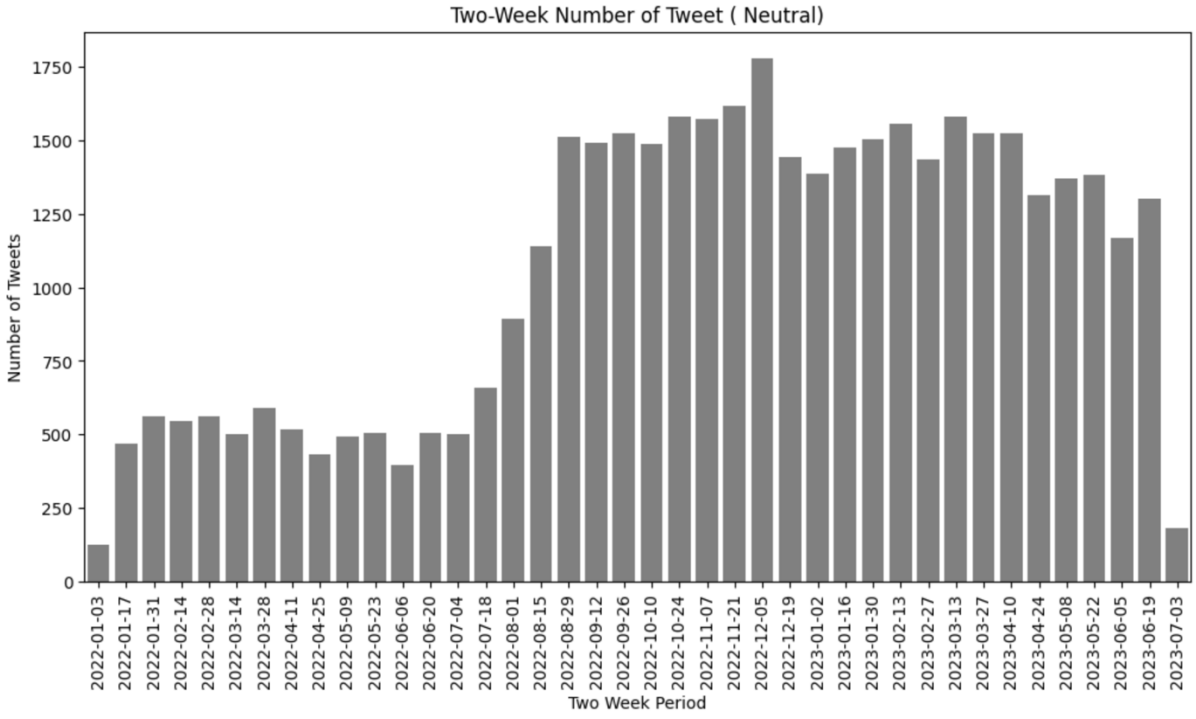
Dataset includes the present an analytic complication of tweets collected from various users. Moreover, it includes not only the tweets but also details such as the number of likes, retweets, quote counts, and sentiments analysis labels. In dataset sentiment analysis is categorized as a “Positive”, “Negative” and “Neutral” for every tweet. In the data set content of tweets provides rich source of data reflecting the sentiment and conversational trends on social media. This dataset can be utilized to understand social media dynamics and user interaction.

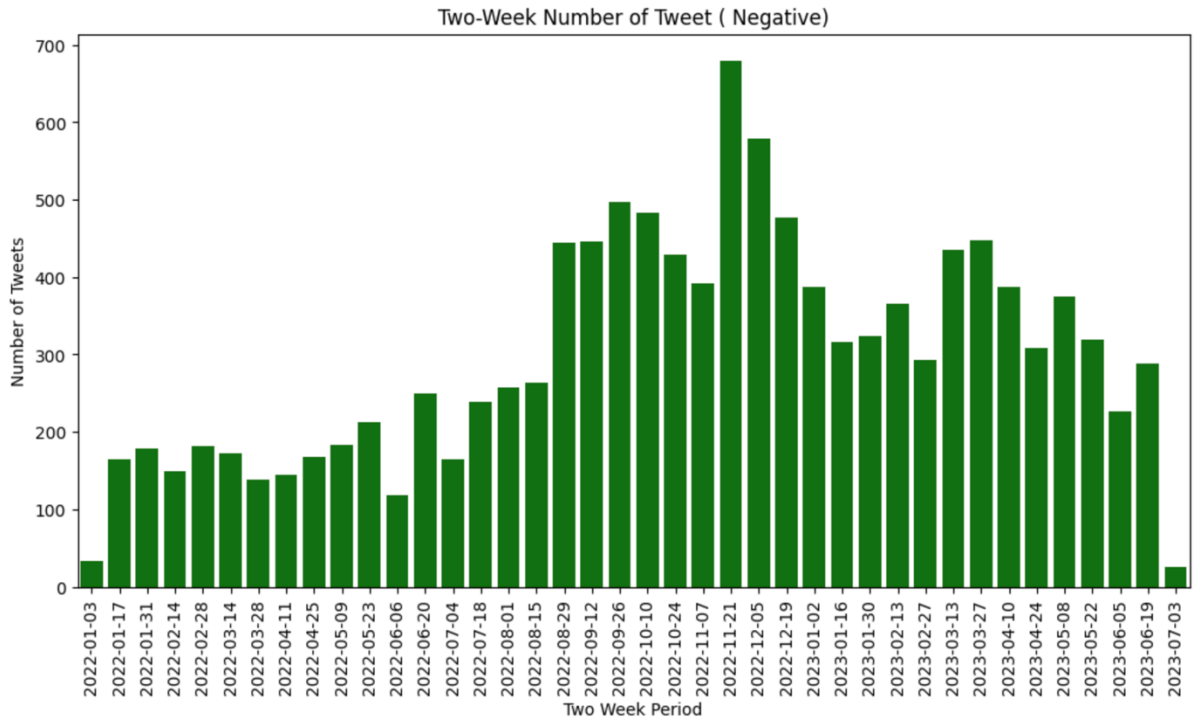


- The graph shows the frequency of neutral, positive, and negative tweets which are related with the Bitcoin.




Two-Week Number of Tweet(Positive)
Number of Tweets
0
200
400
600
800
2022-01-03
2022-01-17
2022-01-31
2022-02-14
2022-02-28
2022-03-14
2022-03-28
2022-04-11
2022-04-25
2022-05-09
2022-05-23
2022-06-06
2022-06-20
2022-07-04
2022-07-18
2022-08-01
2022-08-15
2022-08-29
2022-09-12
2022-09-26
2022-10-10
2022-10-24
2022-11-07
2022-11-21
2022-12-05
2022-12-19
2023-01-02
2023-01-16
2023-01-30
2023-02-13
2023-02-27
2023-03-13
2023-03-27
2023-04-10
2023-04-24
2023-05-08
2023-05-22
2023-06-05
2023-06-19
2023-07-03
Two Week Period


Two-Week Number of Tweet (Neutral)
Number of Tweets
0
250
500
750
1000
1250
1500
1750
2022-01-03
2022-01-17
2022-01-31
2022-02-14
2022-02-28
2022-03-14
2022-03-28
2022-04-11
2022-04-25
2022-05-09
2022-05-23
2022-06-06
2022-06-20
2022-07-04
2022-07-18
2022-08-01
2022-08-15
2022-08-29
2022-09-12
2022-09-26
2022-10-10
2022-10-24
2022-11-07
2022-11-21
2022-12-05
2022-12-19
2023-01-02
2023-01-16
2023-01-30
2023-02-13
2023-02-27
2023-03-13
2023-03-27
2023-04-10
2023-04-24
2023-05-08
2023-05-22
2023-06-05
2023-06-19
2023-07-03
Two Week Period



- These 4 graphs above divide the dates between 2022/01/01 and 2023/06/22 into 2-week periods and shows the number of negative, positive, neutral, and total number of tweets about bitcoin during these 2-week periods.

Formulating Hypothesis and Hypothesis Testing:

The graphs which were visualized based on the datasets lead to understanding the information contained how datasets will used during the hypothesis testing process.

Formulating Hypothesis:

- Null Hypothesis (H0):

  Social media activities (tweets) have no effect on both Bitcoin's price and trading volume.

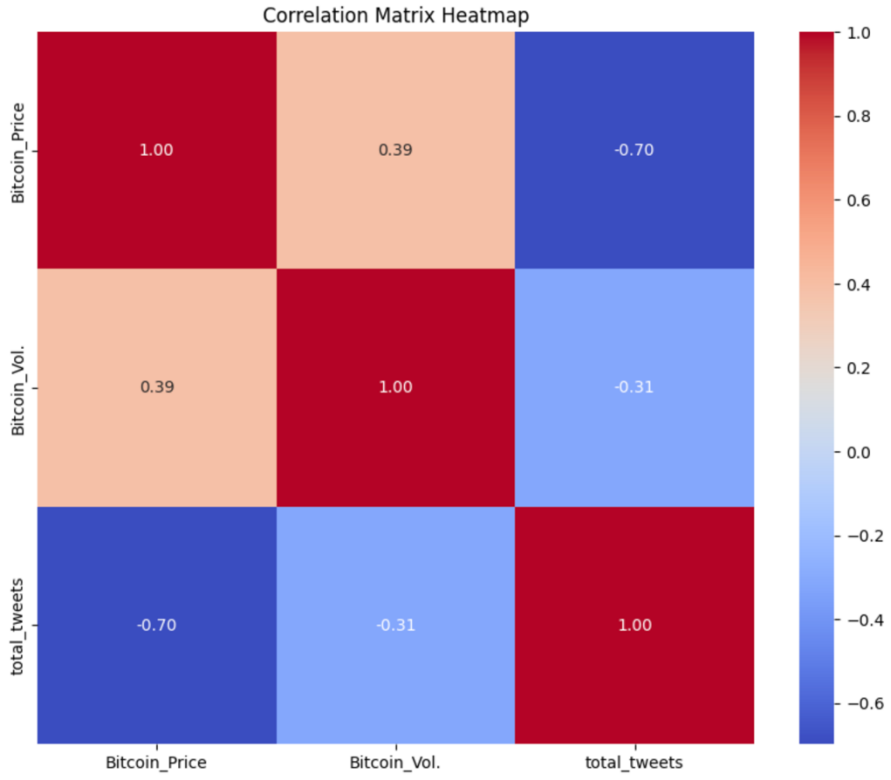
- Alternative Hypothesis (HA):

  Social media activities (tweets) have effect on both Bitcoin's price and trading volume.

Hypothesis Testing:

- Finding Correlation Matrix:

In this step the correlation between Bitcoin Price, Bitcoin trading volume, and the total number of tweets from 2022/01/01 to 2023/06/22, was analyzed. After, merging the datasets, a correlation matrix was created, and heatmap was generated.



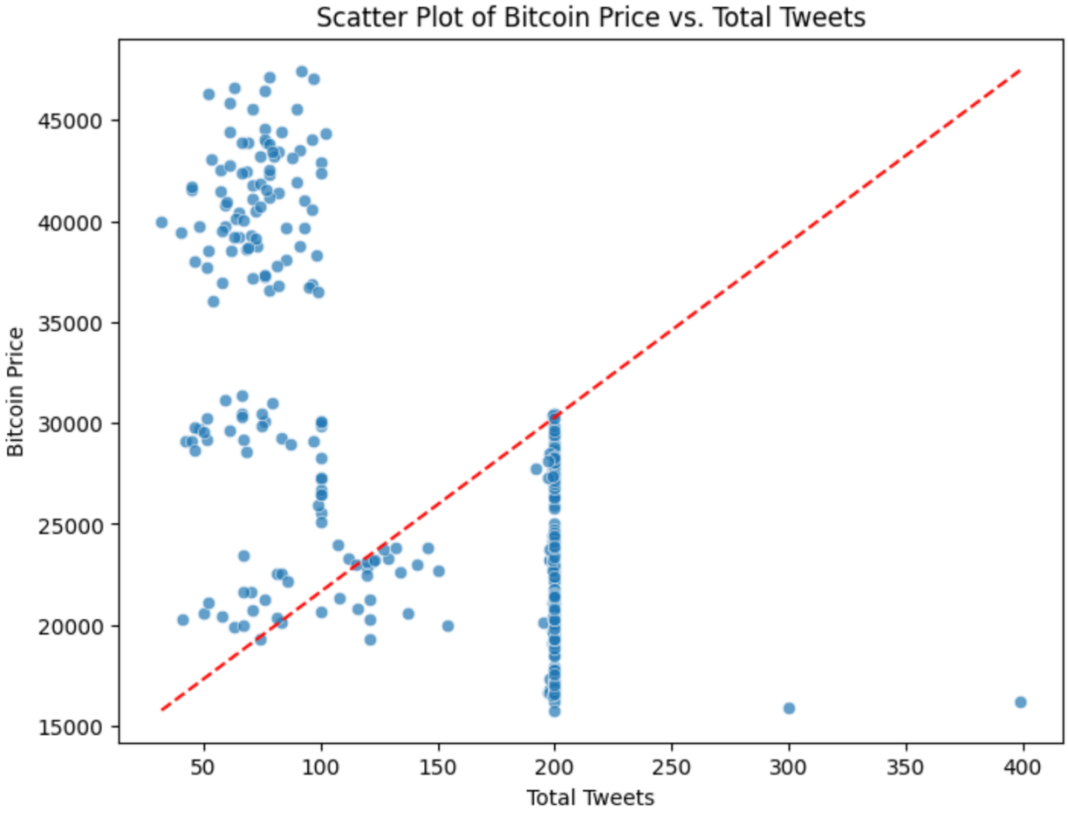
The result showed that:

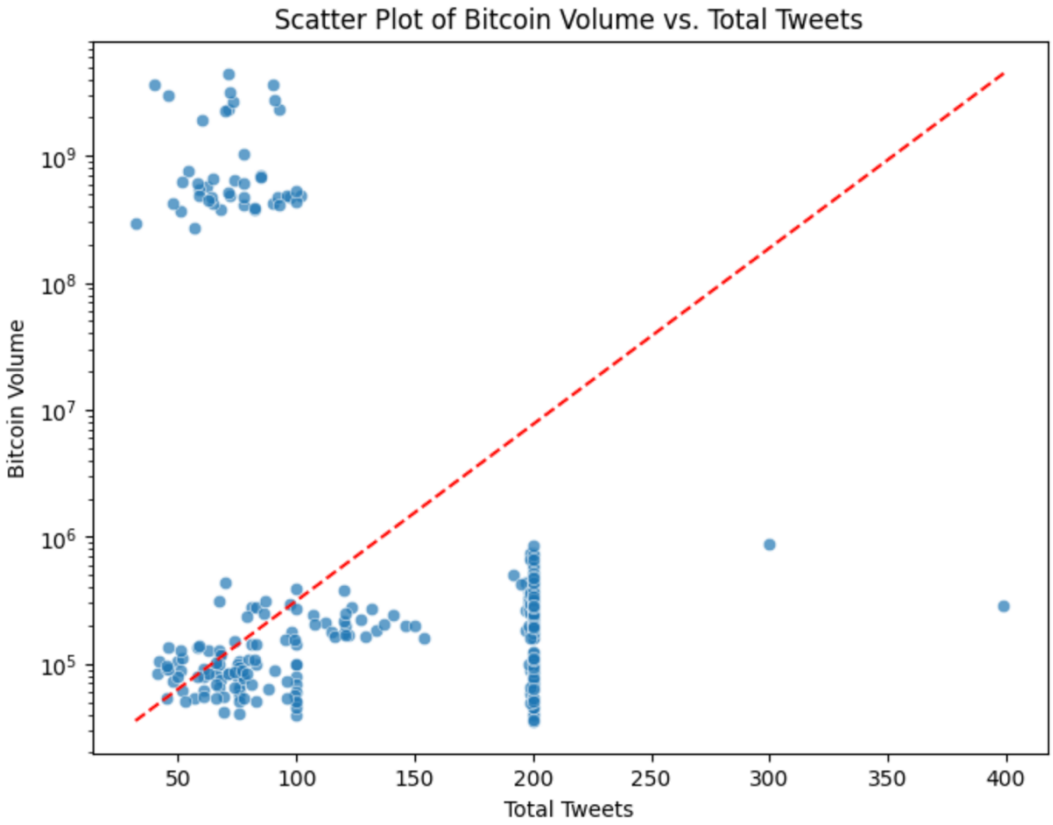
- A strong negative correlation was found between Bitcoin Price and Total number of tweets. (-0.70)
- A negative correlation was found between Bitcoin Trading Volume and total number of tweets. (-0.31)

- Correlation Test:

Correlation test was conducted to examine the relationship between “Bitcoin price and Total number of tweets” and “Bitcoin trading volume and Total number of tweets” from 2022/01/01 to 2023/06/22.

A scatter plot was created to visualize both bitcoin price and number of tweets relationship and bitcoin trading volume and number of tweets. Correlation coefficient shown and p-values were calculated.



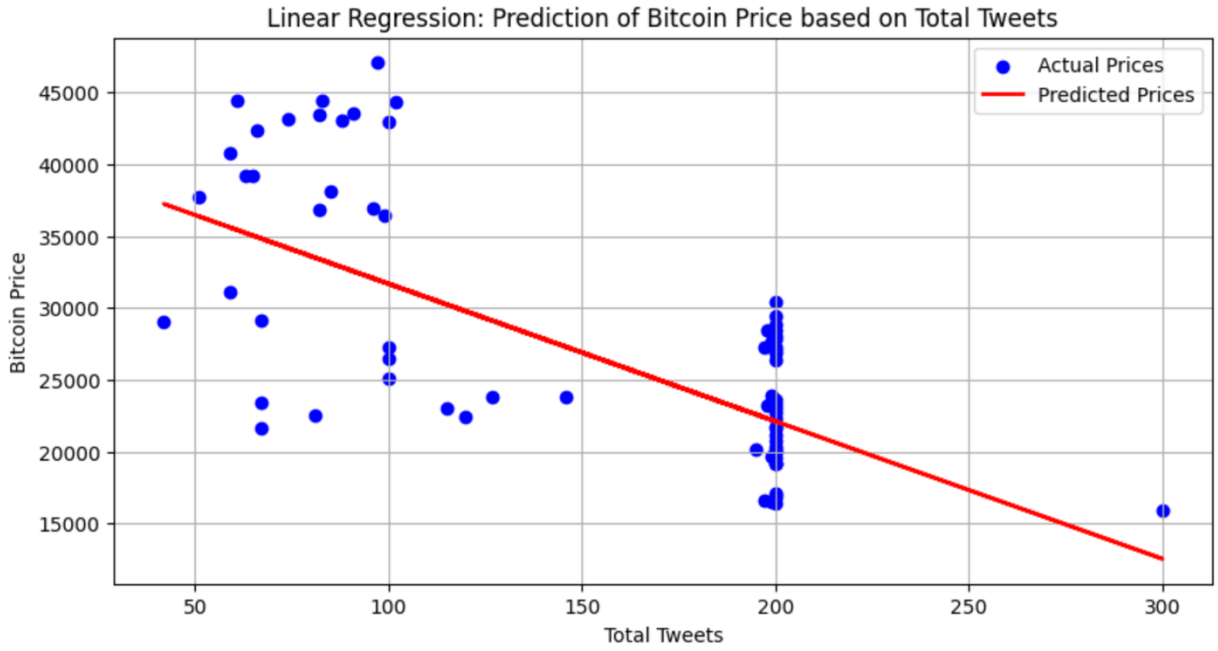


The result showed that:

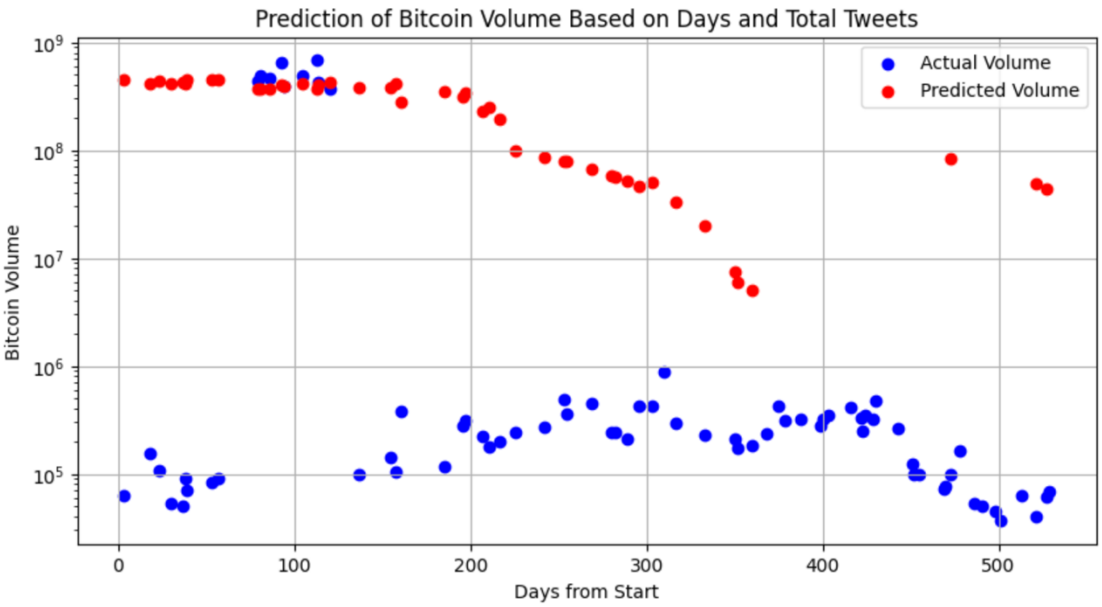
- Both p-values were calculated less than the 0.05 threshold.
- The null hypothesis is rejected.
- Strong negative correlation between Bitcoin Price and number of tweets.
- Negative correlation between Bitcoin Trading Volume and number of tweets.

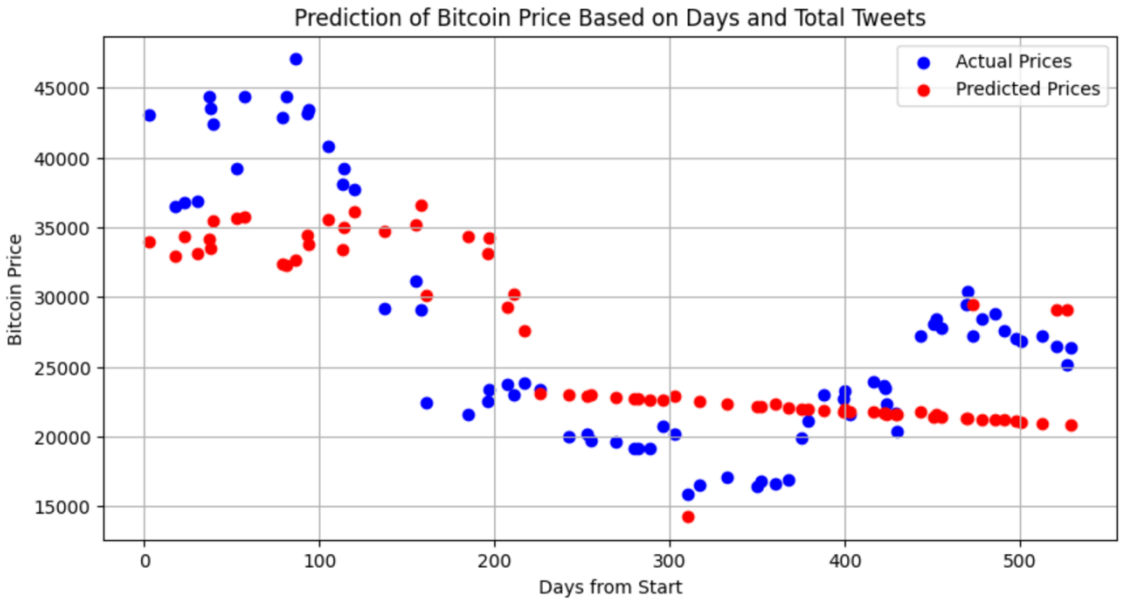
Linear Regression Model for Prediction:

Combining Bitcoin price and number of tweets data, a regression model was created. A “Linear Regression Model” was used to make a prediction for bitcoin price based on number of tweets. Mean Squared Error (MSE) was calculated, and prediction result visualized.



Combining both “Bitcoin price and number of tweets” and “Bitcoin Volume and number of tweets” data, a regression model was created based on dates. Mean Squared Error (MSE) was calculated, and prediction result visualized.


Prediction of Bitcoin Volume Based on Days and Total Tweets
Actual Volume
Predicted Volume
Bitcoin Volume
Days from Start
10^9
10^8
10^7
10^6
10^5
0
100
200
300
400
500


Prediction of Bitcoin Price Based on Days and Total Tweets
Actual Prices
Predicted Prices
Bitcoin Price
Days from Start
45000
40000
35000
30000
25000
20000
15000
0
100
200
300
400
500

Machine Learning Techniques

KNN (K-Nearest Neighbors) Model:

- Predicting Tweet Count Using Bitcoin Price

  Best K Value: 2
  Mean Squared Error (MSE): 1561.8219178082193

- Predicting Bitcoin Price Using Tweet Count

  Best K Value: 1
  Mean Squared Error (MSE): 834276890.7600001

Decision Tree Model:

- Predicting Tweet Count Using Bitcoin Price

  Mean Squared Error (MSE): 1788.266008546683

- Predicting Bitcoin Price Using Tweet Count

  Mean Squared Error (MSE): 42626464.18395548

Model Comparison and Evaluation

- Tweets and Market data were merged.
- For predicting both "Tweet Number" and "Bitcoin Prices":
    - Feature (X): Bitcoin Price
    - Target (y): Tweet Number
    - Feature (X): Tweet Number
    - Target (y): Bitcoin Price
- Data were split into training and testing sets using an 80/20 split.

KNN (K-Nearest Neighbors) Model:

- Loop used to iterate over range 1 to 30 to find best k for both predicting Tweets Number and Bitcoin Price.
- MSE calculated for every k values. Selecting the k which has lowest MSE.
- The final MSE the best KNN model was calculated on the test set.

Decision Tree Model:

Two Methods were used for Decision Tree Model:

Method 1:

- GridSearchCV were used to hyperparameter tuning for decision tree.
- Finding max depth which minimizes the Mean Squared Value.

Method 2:

- Standard Decision Tree model used without hyperparameter tuning.

This shows that using GridSearchCV that hyperparameter the decision tree significantly improves model performance.

Final Evaluation

KNN (K-Nearest Neighbors) Model:

“Predicting Tweet Count Using Bitcoin Price” with KNN gives better results than “Predicting Bitcoin Price Using Tweet Count”. The main reason is bitcoin prices have more straightforward relationship. Furthermore, while predicting “Bitcoin Price” with using “Tweet Count” significance complex relationship occurs which KNN cannot capture.

Decision Tree Model:

“Predicting Tweet Count Using Bitcoin Price” with KNN gives better results than “Predicting Bitcoin Price Using Tweet Count”. The main reason is there is a complex and nonlinear relationship while predicting “Bitcoin Price” with using “Tweet Count”.

Best Performing Model:

KNN for predicting "Number of Tweets" using "Bitcoin Price".

The Reasons for Poor Model Performance:

- **Data Quality and Noise:** High noise levels and Outliers has significant negative affects on ML models.
- **Feature Selection:** Limited features mostly do not capture complex relationship.
- **Temporal Dependencies:** The machine learning models do not account time and event-based trends and dependencies.

There are several ways to improve ML models. Such as:

- Removing the noise and outliers.
- Expanding feature set.
- Using ML models which has higher complexity.
- Cross Validation