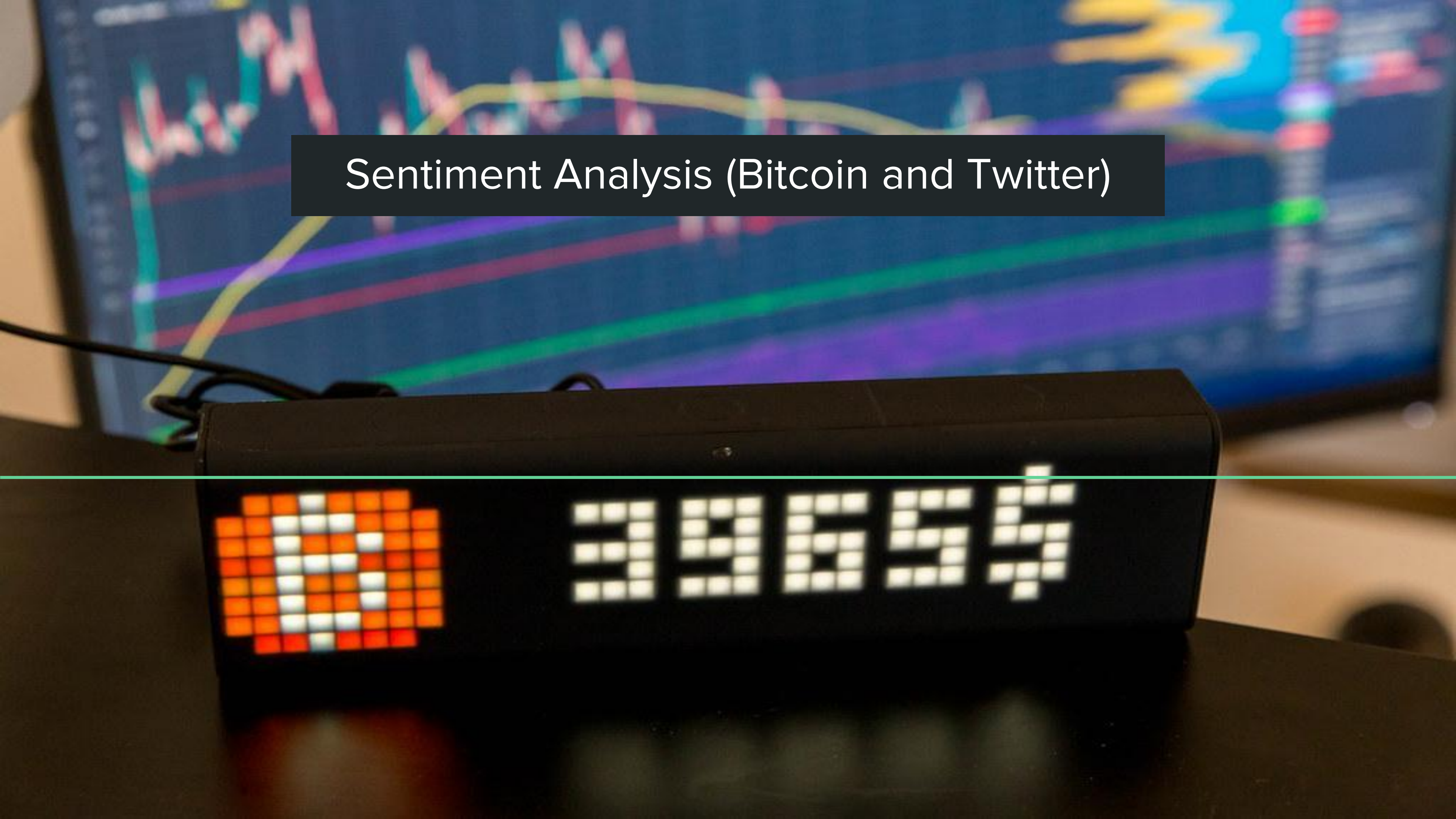
Sentiment Analysis (Bitcoin and Twitter)



Business Problem

Cryptocurrency usage has dramatically increased over the past few years and with more crypto buyers emerging, companies are offering digital currency options. Social media and other media outlets can influence the price of cryptocurrency to fluctuate as it has the power to control the general public’s opinion. We are interested in analyzing social media sentiment and its effects on the price of cryptocurrencies.

Cryptocurrency is volatile and for that reason there are many pitfalls an investor can encounter in the market. Analyzing social media sentiment surrounding crypto allows investors to make strategic investment decisions and prevent them from investing blindly.

Datasets (Bitcoin price)

```
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Open Time           51311 non-null  object
 1   Open                51311 non-null  float64
 2   High                51311 non-null  float64
 3   Low                 51311 non-null  float64
 4   Close               51311 non-null  float64
 5   Volume              51311 non-null  float64
 6   Close Time          51311 non-null  object
 7   Quote Asset Volume  51311 non-null  float64
 8   Number of Trades    51311 non-null  int64
 9   TB Base Volume      51311 non-null  float64
 10  TB Quote Volume     51311 non-null  float64
 11  Ignore              51311 non-null  int64
dtypes: float64(8), int64(2), object(2)
```

- The first dataset is the bitcoin historical pricing dataset. You can use the Binance API Websocket to pull historical pricing data.
- We are mainly interested in getting a daily Opening Price and Number of trades
- Use a Python script to get this data and upload it to S3 bucket.
- Use an AWS EC2 instance to additionally clean then resample the dataset to daily price data.

Datasets (Bitcoin tweets)

```
Data columns (total 13 columns):
 #   Column            Dtype
---  ------            -----
 0   user_name         object
 1   user_location     object
 2   user_description  object
 3   user_created      object
 4   user_followers    object
 5   user_friends      object
 6   user_favourites   object
 7   user_verified     object
 8   date              object
 9   text              object
 10  hashtags          object
 11  source            object
 12  is_retweet        object
dtypes: object(13)
```

- For the second dataset, use the Twitter API in conjunction with the Tweepy library to scrape tweets related to #bitcoin or #btc. The tweets were scraped in an AWS EC2 instance with a daily cronjob that saves to a csv and batch uploaded to s3 bucket.

- In the EC2 instance part of the script cleans the dataset of nulls and drops any tweets that came from a source titled “Bot”

Data Wrangling

```
#Use the tweets list to clean up all tweets, removing any extra characters
def processTweet(tweet): #start process_tweet
    # process the tweets
    #Convert to lower case
    tweet = tweet.lower()
    #Convert www.* or https?://* to URL
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',tweet)
    #Convert @username to AT_USER
    tweet = re.sub('@[^\s]+','AT_USER',tweet)
    #Remove additional white spaces
    tweet = re.sub('[\s]+', ' ', tweet)
    #Replace #word with word
    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)#trim
    tweet = tweet.strip('\'"')
    return tweet
#end
tweets_cleaned = []
#Read the tweets one by one and process it
for i in range (0,len(tweets)):
    processedTweet = processTweet(tweets[i])
    tweets_cleaned.append(processedTweet)
print(len(tweets_cleaned))
tweets_cleaned[0:11]
```

```
#Import and use vaderSentiment to calculate polarity scores of tweets
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

```
vader = SentimentIntensityAnalyzer()
def getCompoundScore(Tweets):
    score = vader.polarity_scores(Tweets)['compound']
    return score
scores = []
#Read the tweets one by one and get the sentiment score
for i in range (0,len(tweets_cleaned)):
    tweets_score = getCompoundScore(tweets_cleaned[i])
    scores.append(tweets_score)
```

```
#store the tweet with polarity score in a dictionary
df ={"tweets": tweets_cleaned, "score" : scores}
data = pd.DataFrame(df)#transfer into a dataframe
```

- We had to clean all tweets of extra characters, emojis etc
- We used vaderSentiment for our Natural Language Processing to analyze and and retrieve the polarity score of each tweet. Based on this score we could then categorize each tweet into Positive, Negative or Neutral tweet
- The next step was to merge the daily tweets with their score and sentiment together with their corresponding daily price data and rate of change

Data Wrangling (continued)

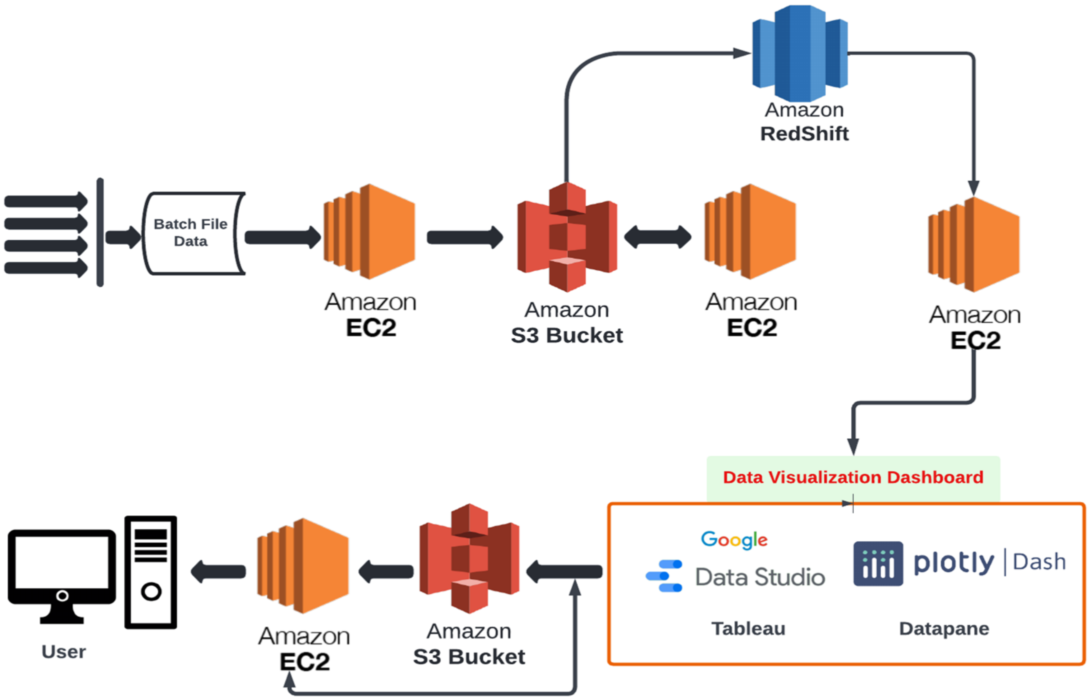
```python
#Creating a merged dataframe containing tweets and their daily price and rate of change
finaldf = pd.merge(merged_tweets,newdf)
```

```python
finaldf
```

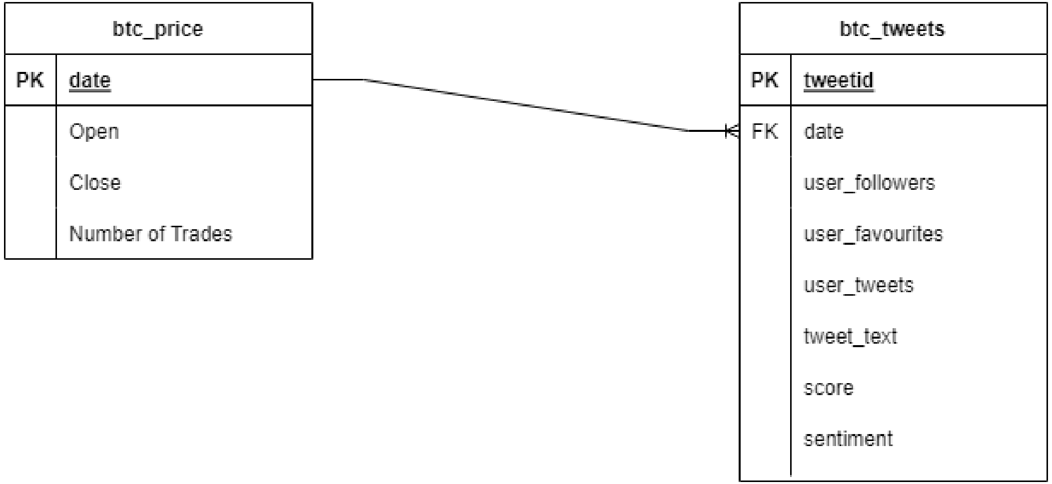
	date	user_followers	user_friends	user_favourites	tweets	score	sentiment	timestamp	Open	High	Low	Close	Volume	Quote Asset Volume	Number of Trades	TB Base Volume	TB Quote Volume	Ignore	roc
0	2021-07-04	32.0	19.0	82.0	bitcoin current price: $ 34682.16 € 29214.49 c...	0.0000	neutral	2021-07-04 00:00:00	35269.082708	35370.196667	35180.610625	35281.847292	910.489079	3.217621e+07	24270.208333	441.256981	1.558724e+07	0.0	NaN
1	2021-07-04	163.0	1.0	0.0	bitcoin: $34682.84 💚 +99.16 last 1 hour (+0.29...	0.9274	positive	2021-07-04 00:00:01	35269.082708	35370.196667	35180.610625	35281.847292	910.489079	3.217621e+07	24270.208333	441.256981	1.558724e+07	0.0	NaN
2	2021-07-04	610.0	1103.0	6270.0	exposed: congressman trying to 'shut down' cry...	-0.0772	negative	2021-07-04 00:00:02	35269.082708	35370.196667	35180.610625	35281.847292	910.489079	3.217621e+07	24270.208333	441.256981	1.558724e+07	0.0	NaN
3	2021-07-04	55401.0	60571.0	77.0	6:00 pm >> $btc price: $34673.75000000 &...	0.0000	neutral	2021-07-04 00:00:02	35269.082708	35370.196667	35180.610625	35281.847292	910.489079	3.217621e+07	24270.208333	441.256981	1.558724e+07	0.0	NaN
4	2021-07-04	310.0	114.0	943.0	dogecoin (doge) failed to cross the 100 dma! U...	-0.5562	negative	2021-07-04 00:00:03	35269.082708	35370.196667	35180.610625	35281.847292	910.489079	3.217621e+07	24270.208333	441.256981	1.558724e+07	0.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3413407	2022-07-03	1239.0	430.0	3563.0	victims of luna are creating a cluna nft colle...	-0.0258	negative	2022-07-03 23:59:48	19179.100208	19233.236458	19124.494375	19180.412500	1064.322215	2.042200e+07	18349.708333	528.374228	1.014068e+07	0.0	-0.003195
3413408	2022-07-03	40.0	105.0	20.0	AT_USER bitcoin fixes this	0.0000	neutral	2022-07-03 23:59:50	19179.100208	19233.236458	19124.494375	19180.412500	1064.322215	2.042200e+07	18349.708333	528.374228	1.014068e+07	0.0	-0.003195
3413409	2022-07-03	2419.0	4.0	18.0	bitcoin last price $19318 btc 🚀 daily indicato...	0.6369	positive	2022-07-03 23:59:52	19179.100208	19233.236458	19124.494375	19180.412500	1064.322215	2.042200e+07	18349.708333	528.374228	1.014068e+07	0.0	-0.003195
3413410	2022-07-03	550.0	91.0	9.0	orion money (orion) went up 11.7 percent in th...	0.0000	neutral	2022-07-03 23:59:55	19179.100208	19233.236458	19124.494375	19180.412500	1064.322215	2.042200e+07	18349.708333	528.374228	1.014068e+07	0.0	-0.003195
3413411	2022-07-03	83.0	122.0	1054.0	AT_USER AT_USER AT_USER yeah but bitcoin sweat...	0.1531	positive	2022-07-03 23:59:56	19179.100208	19233.236458	19124.494375	19180.412500	1064.322215	2.042200e+07	18349.708333	528.374228	1.014068e+07	0.0	-0.003195

3413412 rows × 19 columns

Data Pipeline & Warehousing

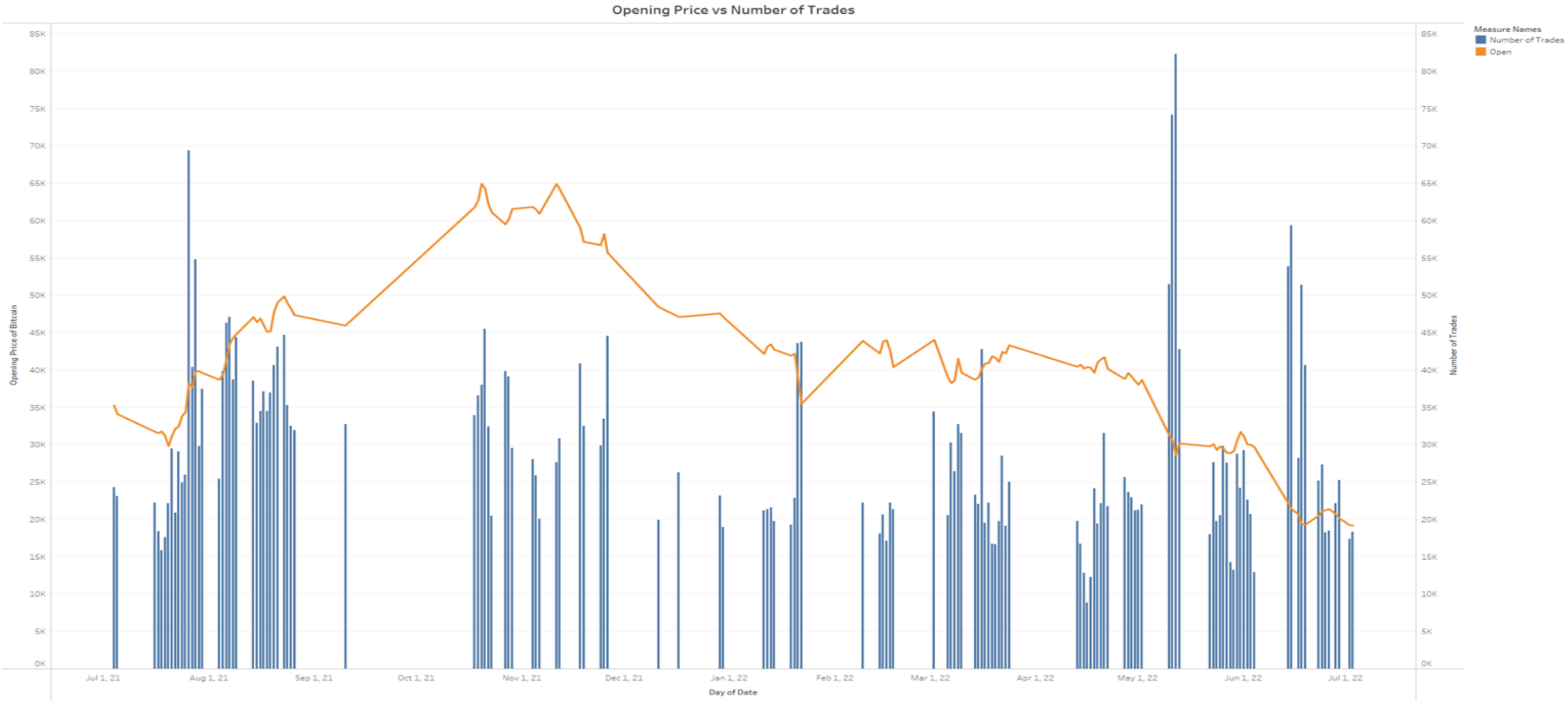


Data Schema



btc_price
PK
date
Open
Close
Number of Trades
btc_tweets
PK
tweetid
FK
date
user_followers
user_favourites
user_tweets
tweet_text
score
sentiment

Exploratory Data Analysis (Example)



The trends of Open and Number of Trades for Date Day. Color shows details about Open and Number of Trades. The data is filtered on Sentiment (Sentimentcountperdate.Csv), which keeps negative and positive.

Exploratory Data Analysis (Example)

Rate of Change (Open Price) vs Sentiment Score

The trends of Rateofchange as an attribute and sum of Scoresbythousandth for Date Day. For pane Sum of Scoresbythousandth: Color shows details about Sentiment. The view is filtered on Sentiment, which keeps negative and positive.