Sand and Sun Public Library

Memo

|  |  |
| --- | --- |
| To: | Sand and Sun Public Library Staff |
| From: | Jordan Spector |
| Date: | 2/19/21 |
| Re: | Project Proposal |

**Topic: Analyze Bitcoin Price Movements**

**Data Description:**

The main data set contains the daily exchange rate of US dollar vs Bitcoin prices from 2013 to 2021 in CSV format

|  |  |  |
| --- | --- | --- |
| Fields | Description | Example |
| Date | Date of trading day for Bitcoin | 2/11/2021 |
| Open | Opening BTC price at start of day | 44807.58 |
| Close | Closing BTC price at end of day | 47969.51 |
| Low | Lowest BTC price recorded during day | 43994.02 |
| High | Highest BTC price recorded during day | 48678.9 |
| Volume | Amount of Bitcoin traded | 81,388,911,810 |

I will obtain supplemental datasets containing tweets as well news headlines that are relevant to Bitcoin—Kaggle seems to have some aggregations and twitter’s API could directly provide data. Once the data is prepared, a combination of NLP techniques like sentiment analysis and modeling will attempt to predict future price movements or identify interesting patterns.

**Research Questions:**

* Can sentiment analysis or key phrases from relevant news and tweets predict future movements in Bitcoin’s price?
* How do Bitcoin halvings (a reduction in the amount of Bitcoin rewarded for mining) affect price trajectory over time? (Group price history by halving dates)
* Are there any noticeable long-term trends in price? Is it rising, declining or stabilizing?
* How has trade volume varied during downward and upward price movements?
* What does price movement and sentiment look like on average at a rollup level—either monthly or yearly—and does it indicate when Bitcoin was more bullish or bearish?

**Data Preparation Plan**

1. Load the datasets into Python
2. Filter and clean the datasets, accounting for missing values, anomalies, etc.
3. Outer join the tweet and news datasets with the main Bitcoin historical dataset by date and form one large dataframe
4. Ensure the dates match up correctly with the joined columns of data and calculate any relevant features to add (like log difference in price per day or sentiment score of tweets/news headlines)