Project C13: Bitcoin and Reddit

Analyzing correlation between Bitcoin trading and Reddit comments that contain the word “bitcoin”

Project repository: https://github.com/BraitLehepuu/RedditBitcoinAnalysis

TEAM:
Anne-Mari Kasemetsa
Ralf Brait Lehepuu

Task 2: Business understanding

Background:

Social media has the power to influence the value of bitcoin in a matter of seconds. This is the case with everything nowadays, but the case is even stronger with cryptocurrency because social media is the main platform for discussion among young people interested in investing. One popular platform for crypto related discussion is Reddit. This raised a question. We know how effectively an opinion leader's tweet about bitcoin can increase/decrease the value of bitcoin, but how much does the overall discussion of regular internet users affect the value of bitcoin and vice-versa - how does the value of bitcoin affect the discussion about bitcoin on social media?

Business goals:

There are three business goals for this project.

1. How does the price of bitcoin affect the score of the comments on Reddit
2. How are the volume of comments and volume of bitcoin trading correlated
3. How are bitcoin trades and Reddit comment scores related

Business success criteria:

1. Gain insight if the price of bitcoin (high or low) has an effect on the score of comments. Possible outcomes:
   higher value of bitcoin → higher scores on Reddit comments, lower value of bitcoin → lower scores on Reddit comments, lower value of bitcoin → higher scores on Reddit comments, higher value of bitcoin → lower scores on Reddit comments, the value of bitcoin doesn’t correlate with change in Reddit comment scores (if the latter, then to find if it is always the case or dependent on some reason)
2. Gain insight if the price of bitcoin affects how many comments are made about bitcoin on Reddit. Possible outcomes:
   higher value of bitcoin → more comments, higher value → less comments, lower value → more comments, lower value → less comments, value of bitcoin doesn’t correlate with change on the volume of comments made on the topic (if the latter, then to find if it is always the case or dependent on some reason)
3. Gain insight if the amount of bitcoin trades made are related to the score of the comments. Possible outcomes:
   higher volume of trades → more popular comments, higher volume of trades → less popular comments, lower volume of trades → less popular comments, lower volume of trades → more popular comments, the volume of trades doesn’t affect the score of the comments greatly (if the latter, then to find if it is always the case or dependent on some reason)

Inventory of resources:

For analysis we are using two datasets:

Dataset 1 (937 MB): Public dataset from Kaggle about reddit comments containing the word “Bitcoin”.
Link to dataset 1

Dataset 2 (105 MB): Public dataset from Kaggle about bitcoin historical OHLC and volume values every minute.
Link to dataset 2

Requirements, assumptions, and constraints:

Schedule for the project:

- 28.11 first report on the project + cleaning data - data feature engineering
- 05.12 business goals 1, 2
- 12.12 business goal 3 + poster for the project
- 15.12 presentation of the project

Both of the datasets were collected from Google BigQuery.

Risks and contingencies:

The only cause that could delay the completion of the project is personal problems of the team members.

Terminology:

- Bitcoin - a decentralized digital currency
- OHLC - short-form of open (highest price of the observed period), high (highest price of the day), low (lowest price of the day), close (final price where the asset trades at).
- BTC - The ticker symbol representing the digital asset or cryptocurrency bitcoin. Also used as shorthand for Bitcoin Core.
- upvote - Reddit term for ‘liking’ a post/comment - increasing a cumulative tally of popularity
- downvote - Reddit term for ‘disliking’ a post/comment - decreasing a cumulative tally of popularity
- subreddit - a forum dedicated to a specific topic on the website Reddit

Costs and benefits:

This project has no costs and is not for profit.
The project will benefit any researcher/enthusiast interested in the correlation between social media influence on cryptocurrency

Data-mining goals:

To find patterns between the peaks of bitcoin values and/or volumes and Reddit comment values and/or volumes

Data-mining success criteria:

To see clear trends on graphs

Task 3: Data understanding

Gathering data:
outline data requirements:

- date of reddit comment
- score of reddit comment
- average score of reddit comments within a time window
- volume of reddit comments within a time window
- volume of reddit comments within a time window

verify data availability:
all required data is available in the two datasets used for this project
define selection criteria:
Reddit dataset: datetime, date, subreddit, score, controversiality
Bitcoin dataset: timestamp, open, high, low, close, volume(BTC), volume(currency), price

Describing data:

Daily and weekly data of Reddit comments: start time, volume of comments, maximum score of all comments, minimum score of all comments, average score of comments, median score of comments

Daily and weekly data of bitcoin trading: start time, volume, open price, high price, low price, low price, close price

Exploring data:

Bitcoin dataset:
CSV files for select bitcoin exchanges for the time period of Jan 2012 to December March 2021, with minute to minute updates of OHLC (Open, High, Low, Close), Volume in BTC and indicated currency, and weighted bitcoin price. Timestamps are in Unix time. Timestamps without any trades or activity have their data fields filled with NaNs.

Reddit dataset:
4M+ Comments from Reddit that contain the word "bitcoin" from 2009 to 2019.

Task 4: Planning of our project

Tasks:

- ☐ cleaning the data
- ☐ create a bitcoin price fluctuation graph over a selected time window
- ☐ create a bitcoin trading volume graph over a selected time window
- ☐ create a reddit comment volume graph over a selected time window
- ☐ create a reddit comment score fluctuation graph over a selected time window

- [ ] join the graphs
- [ ] analysis of the outcome