Group No.14

Student id # 24280053, 24280074

Overview

I chose the topic of cryptocurrency and stock market trends to analyze the discussions on Reddit about crypto market and financial data from Yahoo Finance. This data is valuable for understanding market sentiment and price fluctuations. I expect to see discussions about Bitcoin (BTC) and Ethereum (ETH) on Reddit, as well as historical stock prices for various assets, including traditional stocks (AAPL, TSLA, GOOGL, MSFT) and cryptocurrencies (BTC-USD, ETH-USD).

Data Collection

1. Reddit Data Collection:

- Used the praw API to extract posts from r/CryptoCurrency and r/Bitcoin/Ethereum.
- Extracted fields: title, text, author, date, upvotes, and subreddit.
- Challenges: API rate limits and potential missing or deleted posts.

2. Yahoo Finance Data Collection:

- Used the yfinance library to fetch 2 years of historical closing prices.
- Included stocks and cryptocurrencies.

3. Public Dataset Collection:

- o fetched open-source json dataset from public GitHub repository.
- Challenges: Data format inconsistencies and missing values.

Public Dataset Link

Initial Observations:

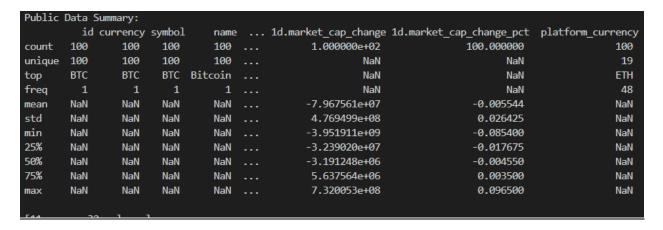
 The Reddit dataset contains multiple discussions with varying engagement levels (upvotes, comments).

```
Data collection and processing complete! Files saved in 'datasets' folder.
Reddit Data Summary:
                                                                                      subreddit
                                                   title text ...
                                                                        upvotes
                                                                    400.000000
                                                                                          400
count
                                                    400 400 ...
                                                    400 162
unique
                                                                           NaN
       How many of you guys actually made money with ...
top
                                                                           NaN CryptoCurrency
                                                        239
freq
                                                                           NaN
                                                    NaN NaN ... 4526.012500
                                                                                          NaN
mean
min
                                                    NaN NaN ... 17.000000
25%
                                                    NaN NaN ... 1191.500000
                                                                                          NaN
50%
                                                    NaN
                                                        NaN
                                                                  2708.500000
                                                                                          NaN
75%
                                                    NaN
                                                                   5898.250000
                                                                                          NaN
max
                                                    NaN
                                                        NaN
                                                                  53426.0000000
                                                                                          NaN
                                                    NaN NaN
                                                                   5628.402679
                                                                                          NaN
std
```

 Financial data shows expected market fluctuations, with crypto being more volatile

```
Finance Data Summary:
                 Close
         3470.000000
count
        10636.151309
mean
std
        21477.374661
min
           85.888641
25%
          184.645088
50%
          372.747620
75%
         3317.024414
       106140.601562
max
```

Public data needs additional cleaning to align with our structured datasets.



Al Product Idea:

I can develop an AI-powered sentiment analysis tool that predicts potential market movements based on Reddit discussions.

Terms of Service & Privacy Issues:

- Reddit: User-generated content must comply with Reddit's API TOS, and direct reposting may violate rules.
- Yahoo Finance: Data cannot be redistributed in bulk; only processed insights should be shared.

Data Quality Considerations:

- Collecting from multiple sources enhances robustness but introduces potential discrepancies.
- Reddit sentiment may not directly be able to predict price changes due to speculative behavior of the users
- Public datasets often need additional pre-processing to match structured financial data.

Storage & Integration Strategy:

- Combine data using timestamps to align Reddit discussions with financial data.
- Implement ETL pipelines to clean and normalize data before analysis.

GitHub Link

Github link