Cryptocurrency Analysis Report: Predicting Bitcoin Price Movements

Group Number: 9
Roll No.: 24280050, 24280005

1. Introduction

This project focuses on cryptocurrencies, particularly Bitcoin, due to its significant influence on the financial market and its role as a leading digital asset. By analyzing data from multiple sources, including Reddit communities, public crypto datasets, and Google Trends, we aim to understand market sentiment, price fluctuations, and the impact of news or regulations on the crypto market.

2. Why Choose This Topic?

Bitcoin's price is highly volatile and influenced by a variety of factors, including market sentiment, trading volumes, and regulatory news. Understanding these factors could provide valuable insights for investors and traders.

Expected Data:

- Bitcoin price movements and trading volumes.
- Sentiment analysis from Reddit discussions.
- On-chain data, such as transaction volumes and active addresses.
- Google Trends data for search terms like "Bitcoin price" and "crypto regulation," reflecting public interest and market sentiment.

3. Data Collection Process

1. Reddit Data

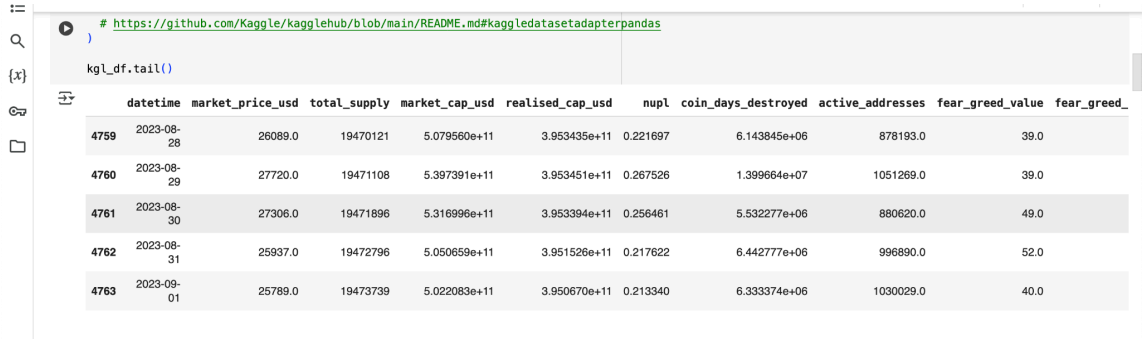
- **Source:** Reddit communities (r/CryptoCurrency and r/Bitcoin).
- **Steps Taken:**
  - Imported PRAW (Python Reddit API Wrapper) to access Reddit data.
  - No actual data extraction was implemented.
- **Challenges:**
  - API rate limits and Terms of Service constraints on storing or redistributing user-generated content.

	Subreddit	Title	Author	Upvotes	Datetime	Text
0	crypto	[Meta] Regarding the future of the subreddit	Natanael_L	108	1.686522e+09	A bit late notice compared to a lot of the oth...
1	crypto	Best Resources To Learn Mathematics and Notati...	silene0259	10	1.739364e+09	What are the best resources to learn mathemati...
2	crypto	International Cryptographic Module Conference:...	fosres	7	1.739317e+09	For those of you that have attended the Intern...
3	crypto	Could I Use Homomorphic Encryption On Decentra...	silene0259	2	1.739318e+09	Is this possible?
4	crypto	Possibility of TDA showing up in cryptography	Dangerous-Relative-7	1	1.739308e+09	Last semester, I had to write a paper about th...
...	...	...	...	...	...	...
766	CryptoTechnology	How does exchange work with existing chains? A...	droaak	2	1.732234e+09	I am trying to understand the technical system...
767	CryptoTechnology	Cere Network vs. Filecoin: Is Cere the Next St...	SpacKingz	5	1.732134e+09	I've been diving deep into decentralized stora...
768	CryptoTechnology	Daggers Over Chains	0basse1	7	1.732079e+09	I've put together a whitepaper introducing a n...
769	CryptoTechnology	What Are the Most Promising Advanced Crypto Sy...	Faith1_2	20	1.732042e+09	With blockchain technology evolving rapidly, w...
770	cryptocurrencymemes	Mod applications are open!	CryptoMaximalist	2	1.737687e+09	With the crypto market heating up again, crypt...

771 rows × 6 columns

2. Public Crypto Datasets

- **Source:** Kaggle (Bitcoin Network On-Chain Blockchain Data).
- **Steps Taken:**
    - Loaded dataset using KaggleHub, focusing on historical Bitcoin price, volume, and on-chain metrics.
    - Preprocessed data by converting the 'datetime' column and one-hot encoding the 'fear_greed_category' column after filtering out '0' values.
- **Challenges:**
    - Data inconsistencies in the 'fear_greed_category' field.
    - Missing values in some on-chain metrics.



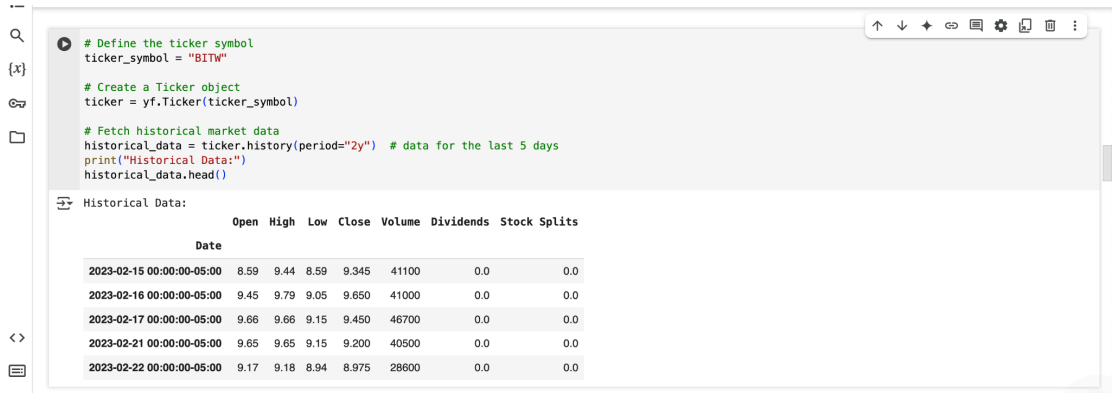
```
  # https://github.com/Kaggle/kagglehub/blob/main/README.md#kaggledatasetadapterpandas
)

kgl_df.tail()
```

	datetime	market_price_usd	total_supply	market_cap_usd	realised_cap_usd	nupl	coin_days_destroyed	active_addresses	fear_greed_value	fear_greed_
4759	2023-08-28	26089.0	19470121	5.079560e+11	3.953435e+11	0.221697	6.143845e+06	878193.0	39.0	
4760	2023-08-29	27720.0	19471108	5.397391e+11	3.953451e+11	0.267526	1.399664e+07	1051269.0	39.0	
4761	2023-08-30	27306.0	19471896	5.316996e+11	3.953394e+11	0.256461	5.532277e+06	880620.0	49.0	
4762	2023-08-31	25937.0	19472796	5.050659e+11	3.951526e+11	0.217622	6.442777e+06	996890.0	52.0	
4763	2023-09-01	25789.0	19473739	5.022083e+11	3.950670e+11	0.213340	6.333374e+06	1030029.0	40.0	

3. yfinance (Yahoo Finance)

- **Source:** Yahoo Finance (Bitcoin historical price data).
- **Steps Taken:**
    - Used yfinance library to extract historical Bitcoin price data.
    - Analyzed correlation between historical price movements and on-chain metrics.
- **Challenges:**
    - Incomplete data for certain date ranges.
    - Discrepancies between on-chain metrics and market prices.



```
# Define the ticker symbol
ticker_symbol = "BITW"

# Create a Ticker object
ticker = yf.Ticker(ticker_symbol)

# Fetch historical market data
historical_data = ticker.history(period="2y")  # data for the last 5 days
print("Historical Data:")
historical_data.head()
```

Historical Data:

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
2023-02-15 00:00:00-05:00	8.59	9.44	8.59	9.345	41100	0.0	0.0
2023-02-16 00:00:00-05:00	9.45	9.79	9.05	9.650	41000	0.0	0.0
2023-02-17 00:00:00-05:00	9.66	9.66	9.15	9.450	46700	0.0	0.0
2023-02-21 00:00:00-05:00	9.65	9.65	9.15	9.200	40500	0.0	0.0
2023-02-22 00:00:00-05:00	9.17	9.18	8.94	8.975	28600	0.0	0.0

4. Initial Observations

- **Basic Statistics and Information:**

    - Utilized Pandas to generate summary statistics, including mean, median, and standard deviation.
    - Observed historical Bitcoin price, volume, and on-chain metrics.
- **Key Findings:**
    - Presence of missing values in some on-chain metrics.
    - Inconsistencies in the 'fear_greed_category' field.
    - Data types appropriately assigned after preprocessing.
- **Correlation Analysis:**
    - Indicated potential relationships between on-chain metrics and Bitcoin price movements.

5. AI Product Idea

Predictive Model for Bitcoin Price Movements

- **Objective:**
  To forecast Bitcoin price movements using historical trends, on-chain metrics, and sentiment analysis from Reddit discussions.
- **Potential Features:**
    - Historical price data and trading volumes from Yahoo Finance.
    - On-chain metrics such as transaction volume and fear/greed index.
    - Sentiment analysis scores from Reddit comments and posts.
- **Application:**
  The model could be integrated into a financial advisory platform, helping traders and investors make informed decisions based on predictive analytics.

6. Terms of Service Constraints and Privacy Issues

Reddit

- User-generated content is subject to copyright, prohibiting redistribution without permission.
- API rate limits impact data collection efficiency.
- Privacy concerns require anonymization to protect user identities.

yfinance (Yahoo Finance)

- Data usage is typically allowed for personal and educational purposes but may have restrictions for commercial use.
- Redistribution or republishing of financial data without consent is prohibited.

General Considerations

- Compliance with GDPR and other data protection regulations is necessary when storing user data.
- Transparent data usage policies should be established to maintain user trust.

7. Multi-Source Data Collection: Benefits and Challenges

Benefits

- Provides a more comprehensive view of market trends and public sentiment.
- Cross-referencing enhances data validation and accuracy.
- Diversified dataset improves model robustness and generalization capabilities.

Challenges

- Conflicts or discrepancies may arise due to differences in data collection methodologies or time zones.
- Data consistency issues, such as varying update frequencies, can hinder analysis.
- Normalization of metrics from multiple sources is required to maintain uniformity.

Potential Discrepancies

- Reddit sentiment may not always correlate with price movements on yfinance due to market manipulation or external influences.
- On-chain metrics might show a different narrative compared to historical price trends, leading to conflicting signals for predictive modeling.

8. Ways of Data Storage and Integration

Centralized Database

- **SQL Databases:** MySQL or PostgreSQL for structured financial and on-chain data.
- **NoSQL Databases:** MongoDB for unstructured Reddit data, supporting flexible schemas and rapid retrieval.

Data Lake Architecture

- Using cloud-based data lakes (e.g., AWS S3, Google Cloud Storage) for scalability and easy access.
- Enables analytics directly on raw data without extensive ETL processes.

Data Integration and ETL

- Using ETL pipelines (e.g., Apache Airflow) to extract, transform, and load data from multiple sources.
- Data normalization and cleaning are integrated into the pipeline to maintain consistency.

Combining Data

- Merging datasets based on common fields like date or timestamp.
- Aligning time zones and data frequencies to ensure temporal consistency.
- Aggregating sentiment scores and on-chain metrics to create a unified feature set for modeling.

9. Conclusion

This project aims to predict Bitcoin price movements using historical trends, on-chain metrics, and sentiment analysis from Reddit discussions. Collecting data from multiple sources provides a comprehensive perspective on market trends and sentiment. However, challenges like data inconsistencies, discrepancies, and Terms of Service constraints require careful handling and processing.
The proposed predictive model could be a valuable tool for investors and traders, enabling data-driven decision-making in the highly volatile cryptocurrency market.

10. Future Work and Recommendations

- Implementing sentiment analysis using NLP models on Reddit data.
- Enhancing data preprocessing techniques to handle inconsistencies and missing values.
- Fine-tuning the predictive model using advanced machine learning algorithms.
- Continuously updating the model with real-time data for improved accuracy.