1. **Group Information**

**Group Number: 39**

Student IDs and Names:

**Muhammad Naveed Ashfaq: 24280042**

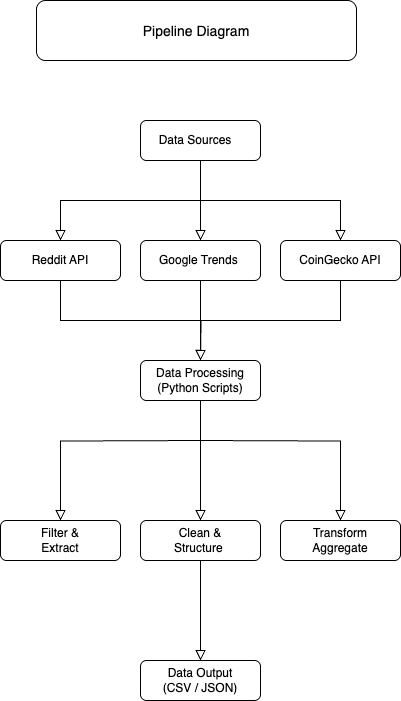
Contribution: Worked on data collection, API integration and report writing

**Syed Feroz Raza: 24030012**

Contribution: Worked on data cleaning, visualization

[Github Link](https://github.com/Naveed333/crypto_data_pipeline) : <https://github.com/Naveed333/crypto_data_pipeline>

**Pipeline Diagram**

****

1. **Overview of the Topic**

We chose Cryptocurrency Trends & Market Data because of its relevance in financial markets and public interest. This dataset includes real-time market prices, Google search trends, and Reddit discussions. Also, we have an interest in buying and selling crypto

We expect to analyze:

**Google Trends:** Public interest over time

**Reddit:** User discussions and sentiment

**CoinGecko API:** Real-time market prices and trading volume

1. **Data Collection Process:**

**Reddit Data Collection:**

* Used praw API to fetch posts related to cryptocurrency.
* Filtered by upvotes to identify the most popular discussions.

Challenges:

* API rate limits
* Some subreddits had restricted data access

**Google Trends Data Collection:**

* Used pytrends to extract search trends for keywords like "Bitcoin," "Ethereum."

Challenges:

* Limited historical data
* API request failures due to Google blocking frequent queries

**Public Data (CoinGecko API):**

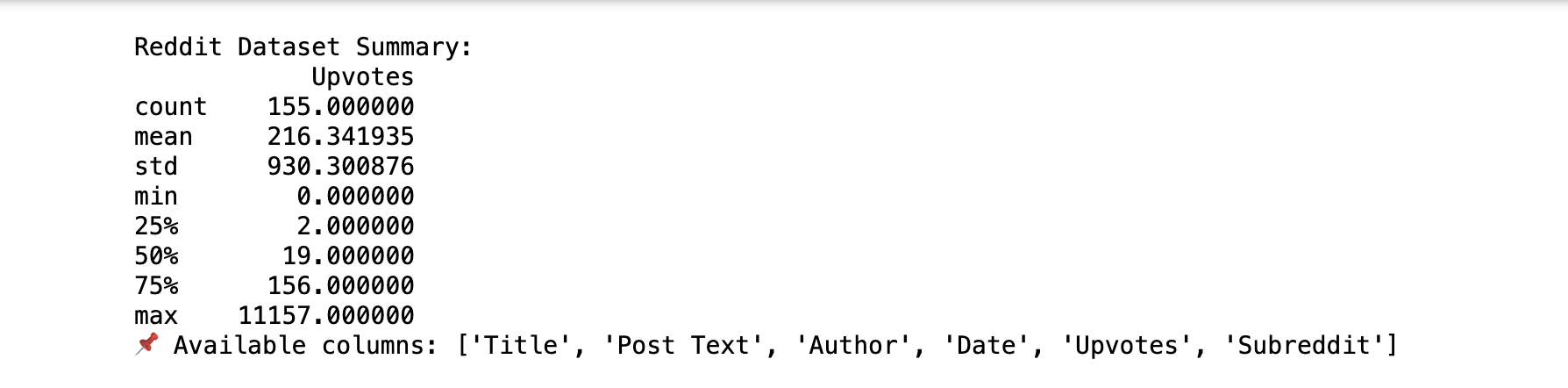
* Used requests to fetch real-time market data.

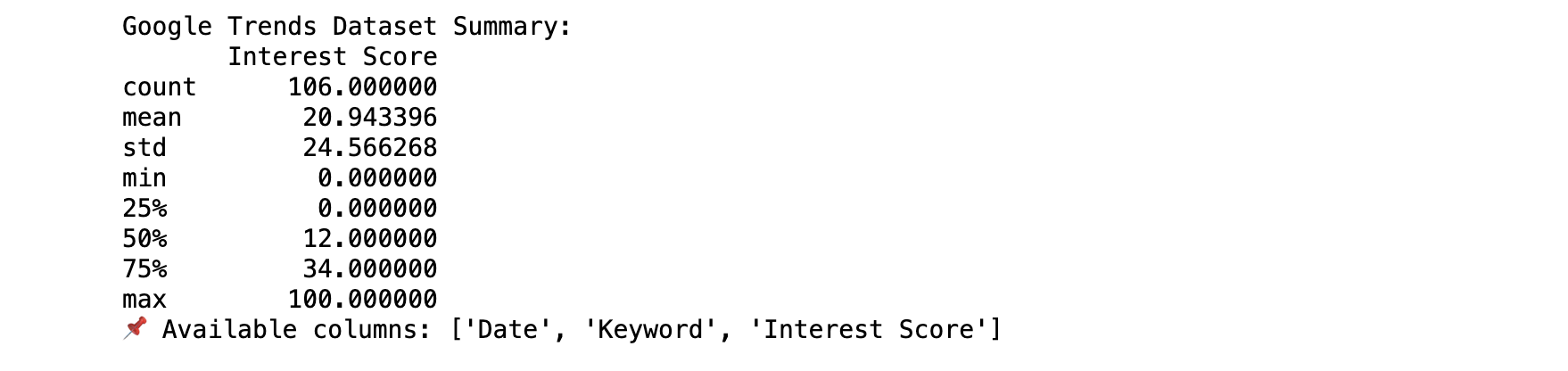
Challenges:

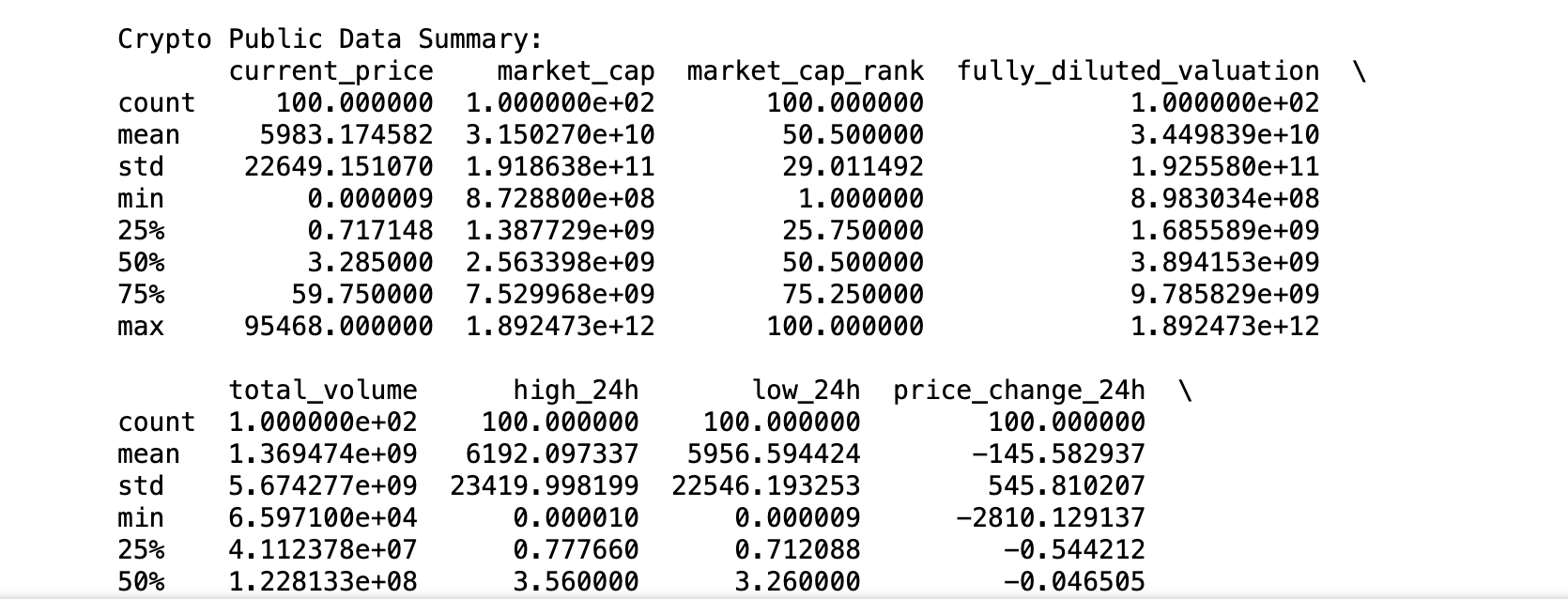
* API rate limits restricting frequent updates

1. **Initial Observations:**

We used pandas to generate a summary of our dataset. Below is a sample output:







**5. AI Product Idea**

Using this dataset, we plan to develop a Crypto Sentiment & Prediction Model that:

* Analyzes Reddit discussions for sentiment analysis
* Correlates search interest with market trends
* Predicts short-term price movements based on combined insights

**6. Terms of Service & Privacy Constraints**

* **Reddit:** API data can be stored but cannot be **redistributed** as per their TOS.
* **Google Trends:** Only aggregated, non-personal data is available, making it compliant.
* **CoinGecko:** Allows public use of their market data with attribution but enforces rate limits.

**Main Concerns:**

* **User Privacy:** Reddit posts might include identifiable information.
* **Data Redistribution:** Certain datasets cannot be openly shared.

**7. Impact of Multi-Source Data Collection**

**Advantages:**

* Provides a more comprehensive analysis
* Reduces bias by balancing different sources

**Challenges:**

* Data format inconsistencies (e.g., different time zones, missing values)
* Merging structured (CoinGecko) vs. unstructured (Reddit) data

**8. Storing & Combining Data**

We structured the datasets as follows:

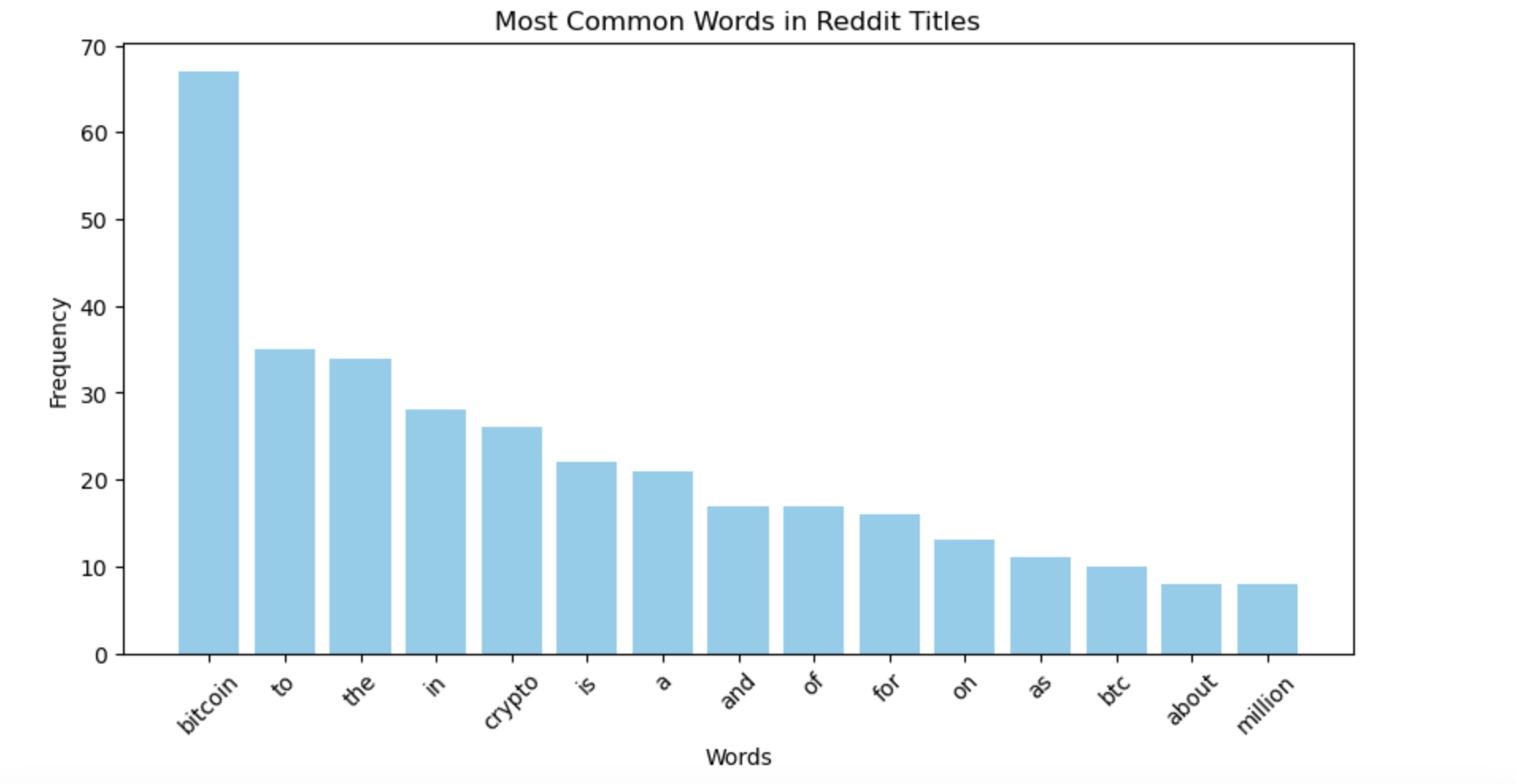
* **Raw Data:**
  + /datasets/raw/reddit\_posts.csv
  + /datasets/raw/google\_trends.csv
  + /datasets/raw/crypto\_market.csv
* **Processed Data:**
  + /datasets/processed/combined\_data.csv (merging all sources)

A possible approach for **data fusion**:

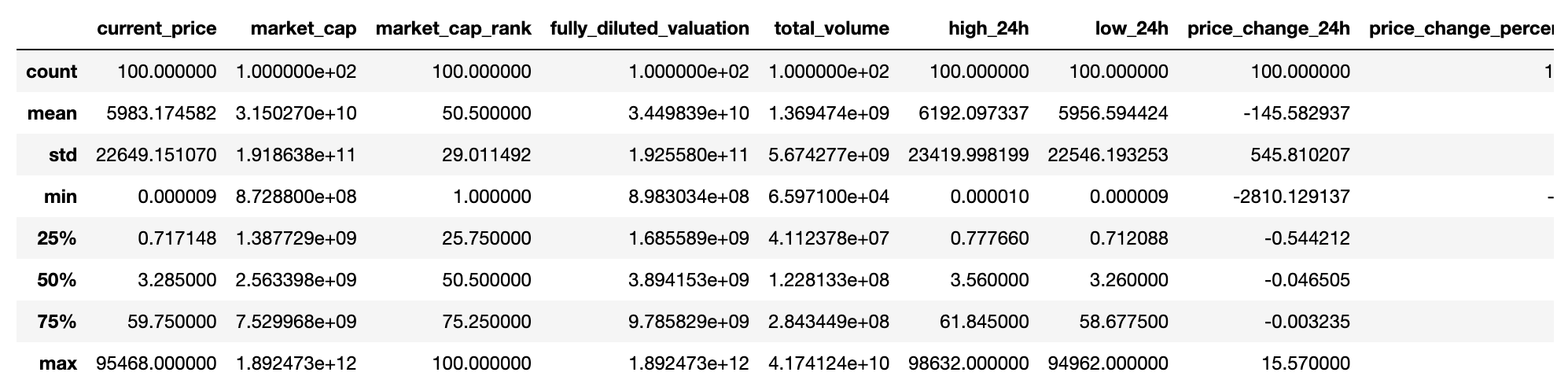
* Convert all timestamps to a **common format (UTC)**
* Normalize data by **scaling market prices & sentiment scores**
* Store data in a **single database (PostgreSQL or MongoDB)** for easy querying

**9. Data Visualization**

* **Reddit: A word frequency chart or average upvotes over time.**



* **Public Data: Basic descriptive stats (count, mean, min, max of relevant fields).**



* **Google Trends: A line chart of interest over time for your keywords.**

