1. **Group Information**

Group Number: 39

Student IDs and Names:

Muhammad Naveed Ashfaq: 24280042

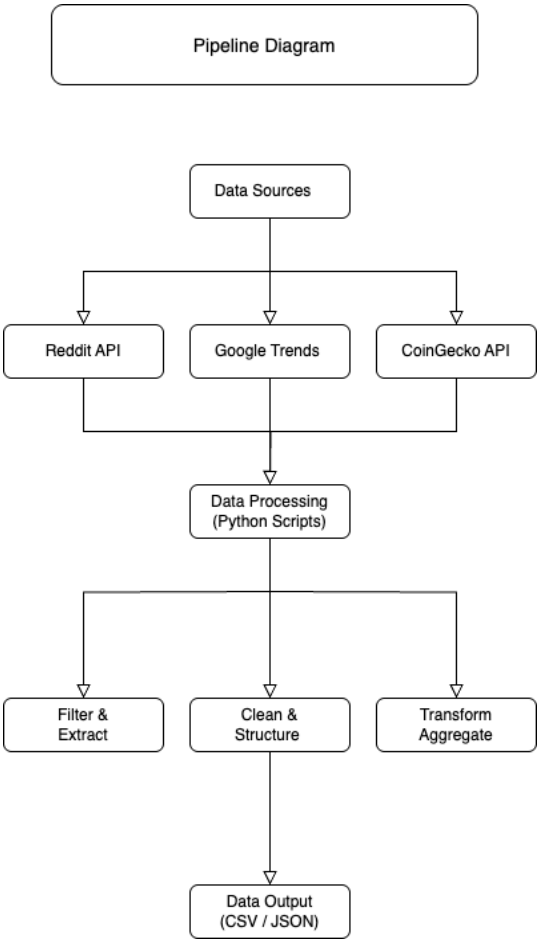
Contribution: Worked on data collection, API integration and report writing

Syed Feroz Raza: 24030012

Contribution: Worked on data cleaning, visualization

Github Link : https://github.com/Naveed333/crypto_data_pipeline

Pipeline Diagram



2. 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

3. 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

4. Initial Observations:

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

```
Reddit Dataset Summary:
            Upvotes
count    155.000000
mean     216.341935
std      930.300876
min        0.000000
25%        2.000000
50%       19.000000
75%      156.000000
max    11157.000000
📌 Available columns: ['Title', 'Post Text', 'Author', 'Date', 'Upvotes', 'Subreddit']


Google Trends Dataset Summary:
       Interest Score
count      106.000000
mean        20.943396
std         24.566268
min          0.000000
25%          0.000000
50%         12.000000
75%         34.000000
max        100.000000
📌 Available columns: ['Date', 'Keyword', 'Interest Score']


Crypto Public Data Summary:
       current_price    market_cap  market_cap_rank  fully_diluted_valuation  \
count     100.000000  1.000000e+02       100.000000             1.000000e+02
mean     5983.174582  3.150270e+10        50.500000             3.449839e+10
std     22649.151070  1.918638e+11        29.011492             1.925580e+11
min         0.000009  8.728800e+08         1.000000             8.983034e+08
25%         0.717148  1.387729e+09        25.750000             1.685589e+09
50%         3.285000  2.563398e+09        50.500000             3.894153e+09
75%        59.750000  7.529968e+09        75.250000             9.785829e+09
max     95468.000000  1.892473e+12       100.000000             1.892473e+12

       total_volume      high_24h       low_24h  price_change_24h  \
count  1.000000e+02    100.000000    100.000000        100.000000
mean   1.369474e+09   6192.097337   5956.594424       -145.582937
std    5.674277e+09  23419.998199  22546.193253        545.810207
min    6.597100e+04      0.000010      0.000009      -2810.129137
25%    4.112378e+07      0.777660      0.712088         -0.544212
50%    1.228133e+08      3.560000      3.260000         -0.046505
```

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:

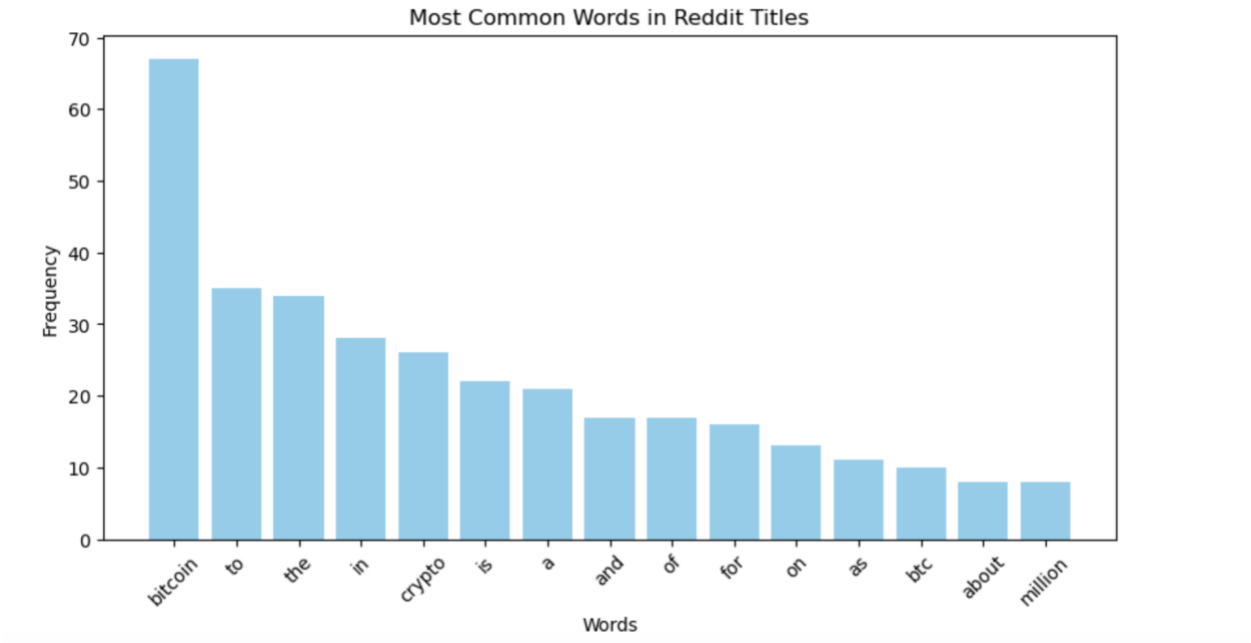
- **Raw Data:**
  - /datasets/raw/reddit_posts.csv
  - /datasets/raw/google_trends.csv
  - /datasets/raw/crypto_market.csv
- **Processed Data:**
  - /datasets/processed/combined_data.csv (merging all sources)

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

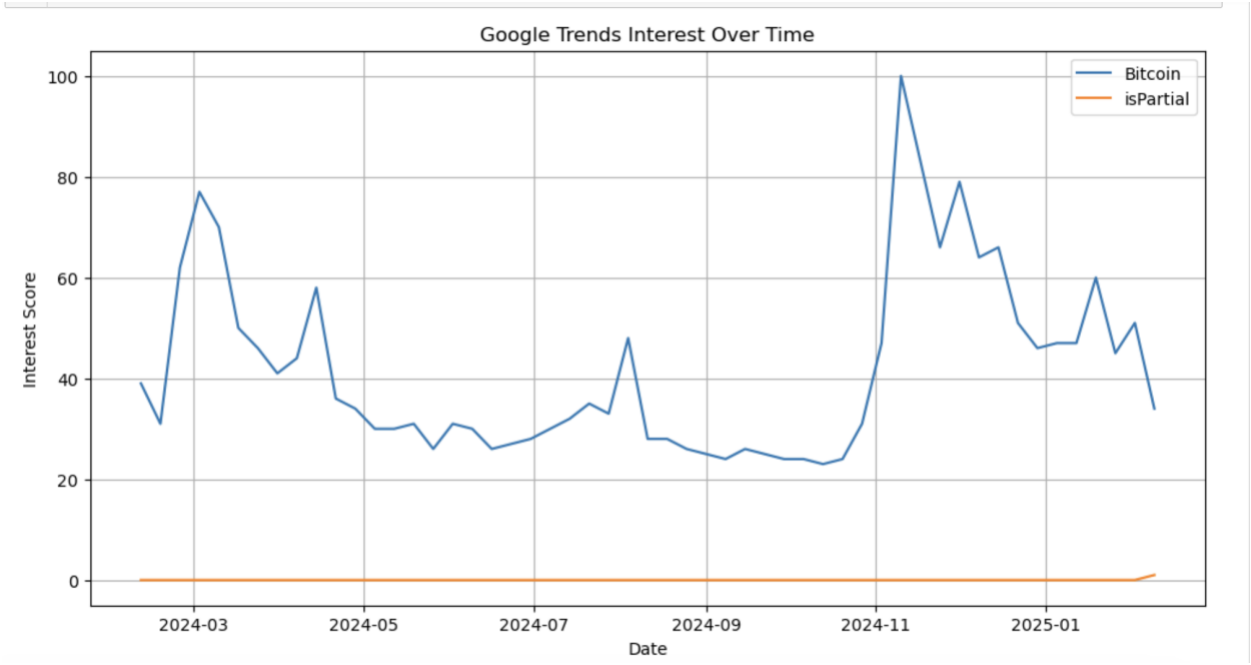
- **Reddit: A word frequency chart or average upvotes over time.**



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

	current_price	market_cap	market_cap_rank	fully_diluted_valuation	total_volume	high_24h	low_24h	price_change_24h	price_change_percer
count	100.000000	1.000000e+02	100.000000	1.000000e+02	1.000000e+02	100.000000	100.000000	100.000000	1
mean	5983.174582	3.150270e+10	50.500000	3.449839e+10	1.369474e+09	6192.097337	5956.594424	-145.582937	
std	22649.151070	1.918638e+11	29.011492	1.925580e+11	5.674277e+09	23419.998199	22546.193253	545.810207	
min	0.000009	8.728800e+08	1.000000	8.983034e+08	6.597100e+04	0.000010	0.000009	-2810.129137	-
25%	0.717148	1.387729e+09	25.750000	1.685589e+09	4.112378e+07	0.777660	0.712088	-0.544212	
50%	3.285000	2.563398e+09	50.500000	3.894153e+09	1.228133e+08	3.560000	3.260000	-0.046505	
75%	59.750000	7.529968e+09	75.250000	9.785829e+09	2.843449e+08	61.845000	58.677500	-0.003235	
max	95468.000000	1.892473e+12	100.000000	1.892473e+12	4.174124e+10	98632.000000	94962.000000	15.570000	

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


Google Trends Interest Over Time
Bitcoin
isPartial
Interest Score
100
80
60
40
20
0
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Date