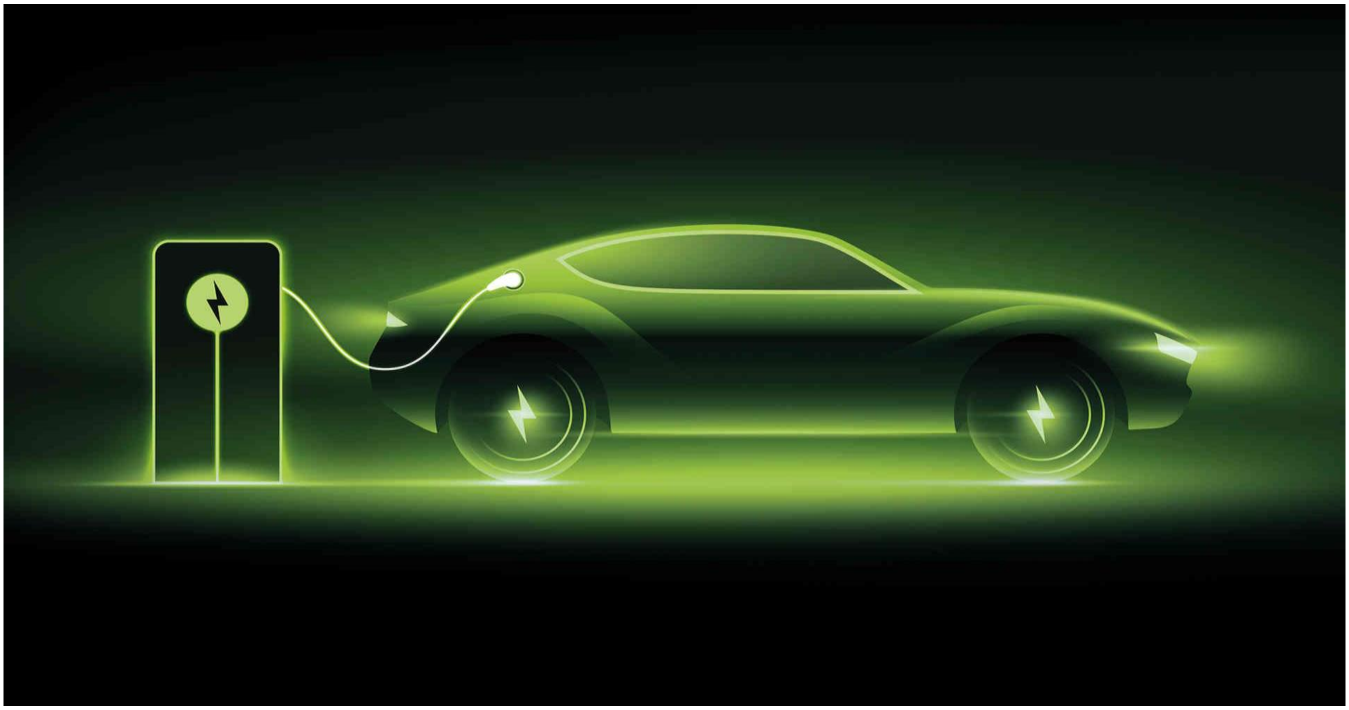


Assignment 1: Project data mosaic

Group Details

- **Group Number**: 26
- **Student IDs**:
  - Student 1: [24280070] - Contributions: Data gathering using api, scripting, data storage, report
  - Student 2: [24280018] - Contributions: Pipeline diagram, reporting , theoretical questions

---

1. Overview of Our Topic

We chose the topic of **Electric Vehicles (EVs)** because of their growing impact on the automotive industry and sustainability. By analyzing EV-related discussions on Reddit and stock performance of major EV manufacturers using Yahoo Finance, we aim to gain insights into market trends and public sentiment.

We expect to see:

- Public discussions on EV performance, charging infrastructure, and policies.
- Stock price trends of leading EV manufacturers.
- Correlations between consumer sentiment and stock performance.

2. Data Collection Process

Reddit Data (Using PRAW)

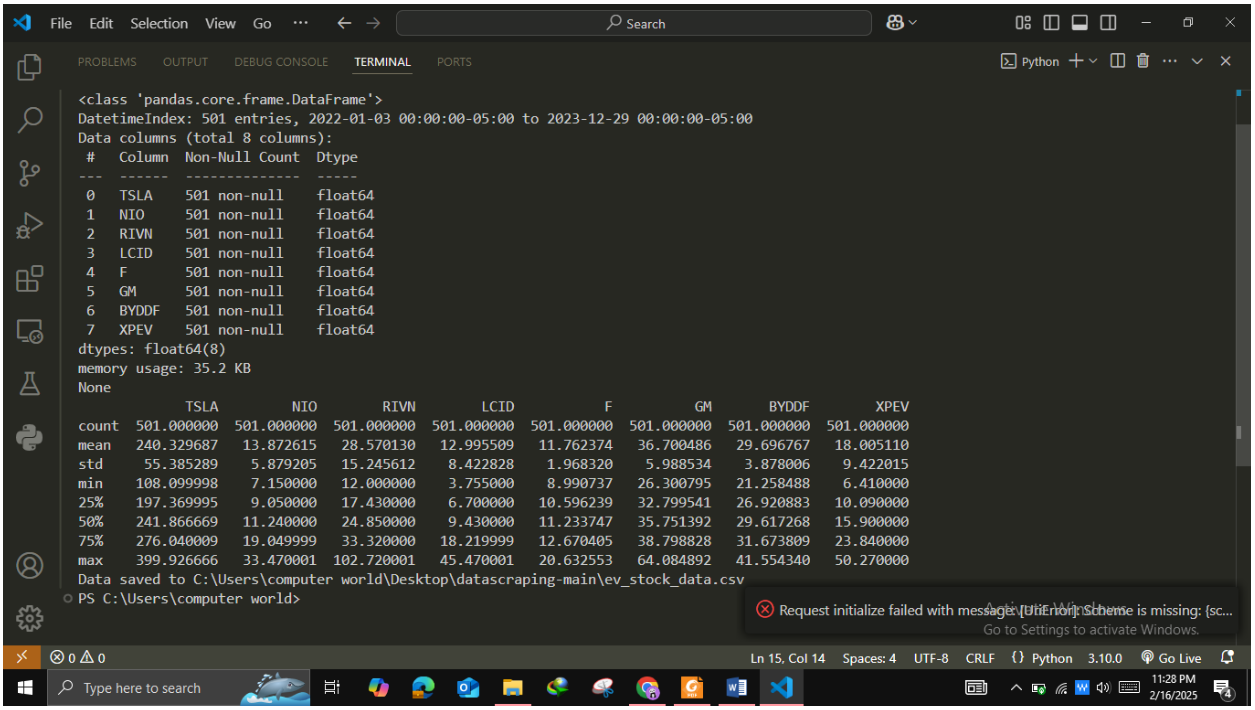
- Extracted posts from the r/electricvehicles subreddit using the **PRAW library**.
- Collected attributes such as title, text, author, upvotes, date, and comments.
- Challenges faced:
    - **API Rate Limits**: Had to introduce time delays to avoid exceeding API request limits.
    - **Incomplete Data**: Some posts lacked detailed content or were removed by moderators.

Yahoo Finance Data (Using yfinance)

- Fetched stock data for major EV companies (Tesla, NIO, Rivian, etc.) over the past 2 years.
- Kept only the ‘Close’ price for analysis.
- Challenges faced:
    - **Data Gaps**: Some stocks had missing days due to market holidays.
    - **Module Issues**: Had to update `yfinance` to ensure smooth functionality.

3. Initial Observations

We processed the data using pandas and generated summary statistics. Below is an example output of our stock dataset:



```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 501 entries, 2022-01-03 00:00:00-05:00 to 2023-12-29 00:00:00-05:00
Data columns (total 8 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   TSLA    501 non-null    float64
 1   NIO     501 non-null    float64
 2   RIVN    501 non-null    float64
 3   LCID    501 non-null    float64
 4   F       501 non-null    float64
 5   GM      501 non-null    float64
 6   BYDDF   501 non-null    float64
 7   XPEV    501 non-null    float64
dtypes: float64(8)
memory usage: 35.2 KB
None
             TSLA         NIO        RIVN        LCID           F          GM       BYDDF        XPEV
count  501.000000  501.000000  501.000000  501.000000  501.000000  501.000000  501.000000  501.000000
mean   240.329687   13.872615   28.570130   12.995509   11.762374   36.700486   29.696767   18.005110
std     55.385289    5.879205   15.245612    8.422828    1.968320    5.988534    3.878006    9.422015
min    108.099998    7.150000   12.000000    3.755000    8.990737   26.300795   21.258488    6.410000
25%    197.369995    9.050000   17.430000    6.700000   10.596239   32.799541   26.920883   10.090000
50%    241.866669   11.240000   24.850000    9.430000   11.233747   35.751392   29.617268   15.900000
75%    276.040009   19.049999   33.320000   18.219999   12.670405   38.798828   31.673809   23.840000
max    399.926666   33.470001  102.720001   45.470001   20.632553   64.084892   41.554340   50.270000
Data saved to C:\Users\computer world\Desktop\datascraping-main\ev_stock_data.csv
PS C:\Users\computer world>
```

Request initialize failed with message: [UriError]: Scheme is missing: {sc...

Activate Windows
Go to Settings to activate Windows.

4. Potential AI Product

Using this dataset, we could develop an **AI-powered EV Market Predictor**, which:

- Uses sentiment analysis on Reddit posts to gauge consumer perception.
- Predicts stock price fluctuations of EV companies based on trends.
- Provides investors and enthusiasts with real-time insights into the EV market.

5. Terms of Service & Privacy Issues

- **Reddit**: User-generated content should not be redistributed without permission. We ensure compliance by only analyzing aggregate data without exposing individual posts or usernames.
- **Yahoo Finance**: Stock data is public, but excessive automated scraping may violate API terms. We followed API request limits to avoid restrictions.

6. Multi-Source Data Quality Considerations

Benefits:

- Cross-verifying trends from different sources improves reliability.
- Combining financial and social data gives a holistic view of market dynamics.

Challenges:

- **Data Discrepancies**: Reddit discussions are opinion-based, while stock data is numerical.
- **Different Update Frequencies**: Social discussions are real-time, while financial markets follow trading hours.

---

7. Data Storage & Integration

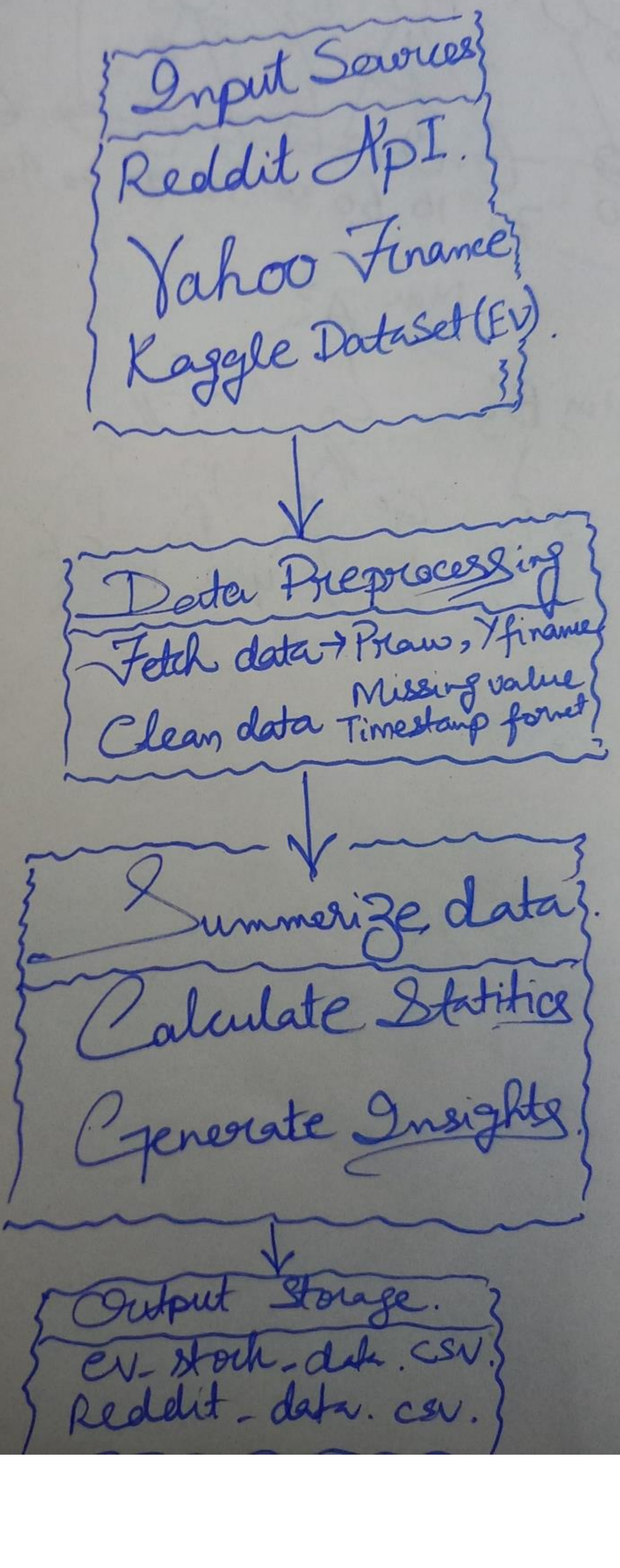
Collecting data from multiple sources improves quality by providing more complete insights, cross-verifying information, and capturing diverse perspectives. For example, combining Reddit sentiment with Yahoo Finance stock data helps understand both market trends and public opinion. However, challenges arise due to inconsistent data formats, time misalignment, and conflicting insights—Reddit discussions may not always reflect actual stock performance. Additionally, API restrictions and rate limits can hinder data collection. To ensure accuracy, preprocessing techniques like data cleaning, timestamp alignment, and normalization are essential, along with cross-verifying sources before analysis.

8. Data Storage & Integration

To effectively store and combine this data, we could:

- Use a **relational database (SQL)** to maintain structured historical records.
- Store unstructured Reddit text data in a **NoSQL database (MongoDB)**.
- Use **ETL pipelines** to clean, transform, and integrate both datasets into a unified analytical framework.

Pipeline DiaGram.



Conclusion

Our project successfully integrates social sentiment from Reddit with financial stock trends from Yahoo Finance. Future work could involve training machine learning models to predict stock trends based on online discussions and news sentiment analysis.

GitHub Repository: https://github.com/AliRaza514/assignment_1_EV_Group-26_ROLL-18-70