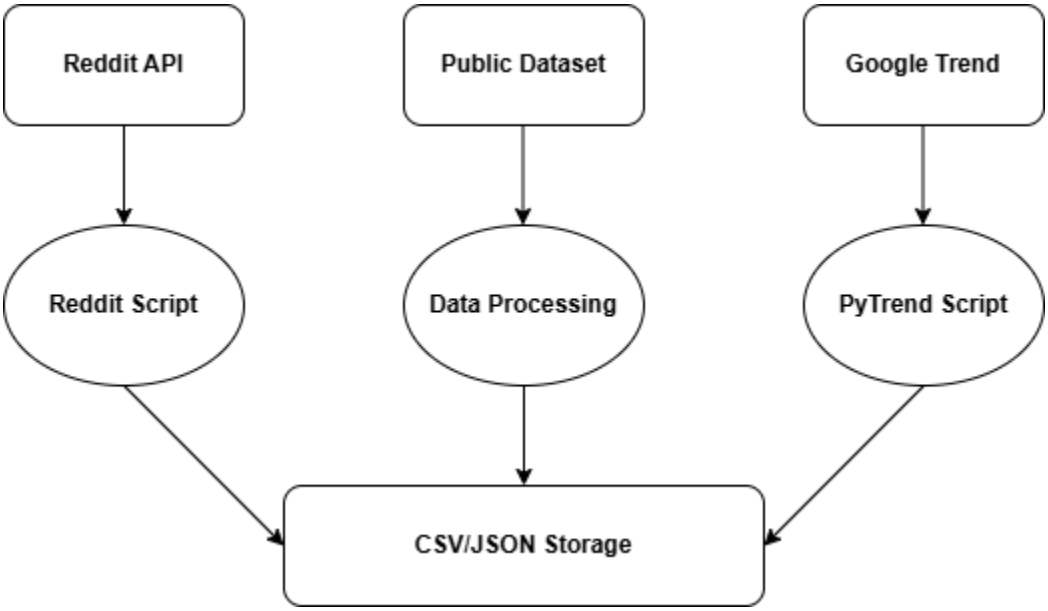
Assignment 1: Project Data Mosaic

Group 13

24280022 *24280009*

Part 3:

3. Pipeline Diagram



4. Git Submission Link

https://github.com/eeman122/assignment1.git

Part 4:

1. Write your group number, student ids, and summarize contributions of both students in the report.

Student IDs: 24280009 (Kashaf Gohar), 24280022 (Eeman Adnan)
Contributions of each member:
24280009: collected data from reddit and public dataset and performed cleaning and q1 to q4.
24280022: collected data from google trend and performed code optimization and q5 to q8.

2. Overview of Your Topic: Why did you choose it? What data do you expect to see?

Reason for choosing this topic:EVs is a rapidly evolving industry and with its global adoption everyone is talking about them. We wanted to take a topic which has plenty of updated data, so we found EVs to be an ideal subject for real-world discussions and market trends. Also, this technology impacts transportation, sustainability and consumer trends, making it well worth our time.
Expectations regarding data: From Reddit, we expect to find user discussions, concerns, reviews on EVs, and sentiment analysis on customer experiences
From Kaggle, we expect to find the structured datasets with market trends, sales figures, registration of EVs compared to internal combustion engine vehicles.

3. Data Collection Process: Summarize the steps you took for each source and any challenges (API rate limits, incomplete data, TOS constraints).

Explanation of how we collected data from each source:Reddit data was collected using PRAW library to fetch posts from electricvehicles(subreddit name). The extracted fields include title, content, subreddit, data, and upvotes(scores). Then, save data as reddit_posts.csv.
Public data was collected by downloading a structured dataset (csv) related to EVs from Kaggle API by downloading Kaggle API in kaggle.json and then loaded it into pandas for analysis
Challenges faced:Some Reddit posts had missing content, so fillna() was used to replace missing content with “no content”. Also, reddit data was unstructured so in order to make it structured data was saved into csv after converting data from unix timestamp to datetime.

4. Initial Observations: Generate a summary of the datasets using pandas. Add the screenshot of the console output of pandas DataFrame in the document.

df.describe() and df.info() have been used to check dataset structure.
Reddit data: The mean upvotes per post is 130.99, but some viral posts reach 1742 upvotes showing high variance. Usually, Posts have on average 27 comments but the viral posts get high engagement reaching 411 comments on average. The dataset spans Feb 9 - Feb 13 2025 showing the fresh discussions.
Kaggle Data:- Vehicles model year ranges from 1997 to 2025 with most from 2022 or later. Average Base MSRP is $992 and Average electric range is 52.16 miles.

Dropped “Legislative District” due to excessive missing values. Filled categorical missing values with “unknown”. Replaced missing numeric values in “ electric range” and “base mrsp” with the median. Dropped rows where “Vehicle Location” was missing.
(Attach screenshot here)
Include a screenshot of the summary output.

```
reddit_df = pd.read_csv("reddit_posts.csv")
print(reddit_df.describe())
# print(reddit_df)
```

[59] ✓ 0.0s

```
                Date       Upvotes
count   1.000000e+02    100.000000
mean    1.694182e+09   5086.420000
std     4.707386e+07   7612.531735
min     1.494662e+09      0.000000
25%     1.665072e+09    995.000000
50%     1.714671e+09   2176.000000
75%     1.730871e+09   4721.000000
max     1.739549e+09  38545.000000
```

```
trends_df = pd.read_csv("datasets/raw/google_trends_data.csv")
print(trends_df.describe())
```

[57] ✓ 0.0s

```
       electric vehicle       tesla  ev charging
count         53.000000   53.000000         53.0
mean           1.566038   78.037736          2.0
std            0.500363    9.980678          0.0
min            1.000000   64.000000          2.0
25%            1.000000   70.000000          2.0
50%            2.000000   75.000000          2.0
75%            2.000000   86.000000          2.0
max            2.000000  100.000000          2.0
```

5. What AI product will you make using this data?

The AI product from this data can be a prediction model that predicts the trends for the adoption of EVs. It can help show trends and discussion and based on that we can predict what will be the further increase or decrease in sales. It will help analyse sentiments regarding a product.

6. Which terms of service constraints or privacy issues might arise when collecting data from Reddit and Google? Consider limitations on storing or redistributing user-generated content.

Reddit doesn't allow storage of personal user data and it requires the anonymization of posts. For google trends the data is in a collection however it has limitations in its usage. Google's TOS doesn't allow automated scraping.

7. How does collecting from multiple sources help or hinder data quality? What conflicts or discrepancies might you face?

The benefits of taking data from multiple sources are that there is great data quality and variation in the data. It allows data to be used for cross-verification and can make up for missing data.
The challenges associated with this are that there are only opinions on reddit and hardly any factual data which can change predictions. Public data may not be updated and can have historical data which also may not keep up with real-time trends. Google trends may not show or reflect the true intent of the buyer.

8. Can you think of ways to store and combine all of this data?

We can combine this data based on region as that will cause varying trends. The data can be stored in CSV or JSON format if it's kept for a specific time period. In the case we want to keep historical data we can move it to SQL Databases.