**Assignment 1: Project Data Mosaic**

**Group Details**

* **Group Number:** [Your Group Number]
* **Student IDs:** [Student 1 ID], [Student 2 ID]
* **Contributions:**
  + Student 1: [Describe contributions]
  + Student 2: [Describe contributions]

**1. Overview of the Topic**

**Chosen Topic:** Electric Vehicles (EVs)  
**Reason for Choice:** The increasing adoption of EVs and their impact on sustainability.  
**Expected Data Insights:** Trends in public discussions, search interest, and real-world EV adoption statistics.

**2. Data Collection Process**

**2.1 Reddit Data Collection**

* **Source:** r/ElectricVehicles subreddit
* **Method:** Used praw to fetch relevant posts based on keywords.
* **Challenges:** API rate limits, filtering relevant content.
* **Output:** datasets/raw/reddit\_posts.csv

**2.2 Google Trends Data Collection**

* **Source:** Google Trends
* **Method:** Used pytrends to fetch interest data over 12 months.
* **Challenges:** Trends data fluctuates based on regional selection.
* **Output:** datasets/raw/google\_trends.csv

**2.3 Public Dataset Collection**

* **Source:** [Specify dataset source]
* **Method:** Downloaded and processed CSV dataset.
* **Challenges:** Missing values, data cleaning.
* **Output:** datasets/raw/public\_ev\_data.csv

**3. Initial Observations**

* **Reddit Data:** Summary statistics (word count, average upvotes, frequent topics).
* **Google Trends Data:** Keyword trends over time (Tesla, EV charging, etc.).
* **Public Dataset:** Statistical insights (average registrations, growth rates).

**Console Output (Screenshots):** *(Include Pandas summary statistics screenshots)*

**4. Potential AI Product**

* **Idea:** EV market prediction tool
* **How It Works:** Combines search trends, discussion analysis, and real-world data.
* **Use Case:** Helps policymakers, businesses, and researchers understand EV adoption patterns.

**5. Ethical and Privacy Considerations**

* **Reddit & Google Data Constraints:** Respecting TOS, avoiding direct user data exposure.
* **Public Data Limitations:** Ensuring accuracy and relevance.

**6. Data Integration Challenges**

* **Conflicts in Data:** Reddit discussions might be opinion-driven, while public datasets are factual.
* **Combining Different Formats:** Handling structured/unstructured data.
* **Storage and Access:** Organizing in a structured manner for future retrieval.

**7. Data Visualization**

*(Include at least one chart per dataset)*

* **Reddit:** Word frequency or sentiment analysis.
* **Google Trends:** Line chart showing interest over time.
* **Public Data:** Histogram or bar chart for EV adoption rates.

**8. GitHub Repository**

* **Link:** [Provide GitHub repo link]
* **Structure:**
  + assignment1/report.pdf
  + assignment1/scripts/reddit.py
  + assignment1/scripts/google\_trends.py
  + assignment1/scripts/public\_data.py
  + assignment1/datasets/raw/\*.csv