# Data Report

## 1. Question

**How has the adoption of electric vehicles influenced energy grids and infrastructure development in North American cities?**

## 2. Data Sources

### 2.1. Estimated U.S. Consumption of Electricity by Light-Duty Electric Vehicle Types

- **Why Chosen:**
  - Provides historical and current data on electricity consumption specific to electric vehicles (EVs), essential for analyzing the impact on energy grids.
- **Source Information:**
  - **Provider:** U.S. Energy Information Administration (EIA)
  - **Access Link:** EIA Table D1 Excel File
- **Data Content:**
  - Monthly and annual estimates of electricity consumption by different types of light-duty EVs across the United States.
  - Includes data on energy consumption patterns, which can be correlated with EV adoption rates.
- **Structure and Quality:**
  - Data is organized in an Excel file with a single sheet containing annual and monthly data.
  - Very concise data, only limited to the year 2022 and 2024. Sourced from a reputable government agency.
- **License and Usage Rights:**
  - **License:** Public Domain (as per EIA's data usage policy)
  - **License Details:** EIA Terms of Use https://www.eia.gov/about/copyrights_reuse.php
  - **Obligations and Compliance:**
    - Free to use with proper attribution.
    - Plan to cite EIA as the data source in all project materials.

### 2.2. Alternative Fuel Stations API

- **Why Chosen:**
  - Offers real-time data on the location and characteristics of electric charging stations, reflecting infrastructure development in response to EV adoption.
- **Source Information:**
  - **Provider:** National Renewable Energy Laboratory (NREL)
  - **Access Link:** NREL Alternative Fuel Stations API https://developer.nrel.gov/docs/transportation/vehicles-v1/
- **Data Content:**
  - Information on electric vehicle charging stations, including location, operational status, and station details across North America.

    - Facilitates analysis of infrastructure growth and distribution patterns.
- **Structure and Quality:**
    - Data is accessible via API in JSON format, enabling automated and up-to-date data retrieval.
    - Maintained by a reputable institution, ensuring reliability and accuracy.
- **License and Usage Rights:**
    - **License:** NREL Data License https://www.nrel.gov/disclaimer.html
    - **Obligations and Compliance:**
        - Requires attribution and adherence to usage guidelines.
        - Plan to include NREL attribution and comply with rate limits and other terms.

# 3. Data Pipeline

## 3.1. High-Level Description

- **Technologies Used:**
    - Programming Language: Python
    - Libraries: Pandas, Requests, JSON
    - Tools: Jupyter Notebook for development and testing

## 3.2. Data Extraction

- **EIA Excel Data:**
    - Automated download of the Excel file from the EIA website.
    - Used `pandas.read_excel()` to load data into DataFrames.
- **NREL API Data:**
    - Utilized the Requests library to make API calls.
    - Parsed JSON responses to extract relevant data fields.

## 3.3. Data Transformation and Cleaning

- **Transformations Applied:**
    - **EIA Data:**
        - Standardized date formats to datetime objects.
        - Renamed columns for consistency.
        - Filtered data to include only relevant vehicle types and time frames.
    - **NREL API Data:**
        - Extracted necessary fields such as station location, status, and capacity.
        - Converted location data into geospatial formats if needed.
- **Reasons for Transformations:**
    - To ensure compatibility between datasets.
    - Facilitate accurate merging and analysis.
    - Improve data quality by handling inconsistencies and missing values.

## 3.4. Challenges and Solutions

- **Issue:** Inconsistent date formats in the Excel data.
    - **Solution:** Implemented a function to parse and standardize date formats.
- **Issue:** API rate limits causing incomplete data retrieval.
    - **Solution:** Added error handling and retry logic, implemented caching where appropriate.
- **Issue:** Missing or null values in key data fields.
    - **Solution:** Applied data imputation techniques and filtered out unreliable records.

### 3.5. Meta-Quality Measures

- **Error Handling:**
    - Used try-except blocks to catch and log errors during data fetching and processing.
- **Data Validation:**
    - Implemented checks for data completeness and correctness after each transformation step.
- **Adaptability:**
    - Designed the pipeline to handle changes in input data structure by dynamically reading headers and schema.

## 4. Results and Limitations

### 4.1. Output Data Description

- **Final Dataset:**
    - A merged dataset combining EV electricity consumption data with charging infrastructure details.
    - Contains time-series and geospatial components for comprehensive analysis.
- **Data Structure:**
    - **Columns:** Date, Electricity Consumption, Number of Charging Stations, Location Coordinates, Station Capacity, etc.
    - **Format:** Stored as a SQLite database for scalability and ease of access.

### 4.2. Data Format Choice

- **SQLite Database:**
    - Allows for efficient querying and data manipulation.
    - Handles larger datasets without the overhead of a full-fledged database server.

### 4.4. Critical Reflection and Potential Issues

- **Data Limitations:**
    - **Geographical Scope:** Some datasets may lack granularity at the city level.
    - **Temporal Alignment:** Discrepancies in data update frequencies between sources.
- **Anticipated Challenges:**
    - **Data Integration:** Merging datasets with differing structures and levels of detail.
    - **Biases:** Possible overrepresentation or underrepresentation of certain regions due to data availability.