Data Report

Air Quality Analysis in Major North American Cities with the Rise of Electric Vehicles.

Yeasin Arafat Shampod (23080363)

Question:

The main question for this project is:
"How has the rise of electric vehicles (EVs) affected air quality in major North American cities?"
The objective of this project is to analyze air quality trends in major North American cities and compare them with the rise in electric vehicle usage. The project involves extracting air quality data and electric vehicle usage statistics, transforming the data to account for inconsistencies, and analyzing whether the increase in electric vehicles has led to improvements in air quality, particularly in pollutants such as CO2, NO2, and PM2.5.

Data Sources:

Source 1: Air Quality Data from the U.S. Environmental Protection Agency (EPA)

- **Description**: This dataset contains air quality measurements from monitoring stations across North America, focusing on key pollutants such as CO2, NO2, particulate matter (PM2.5), and O3.
- **Data Content**: The dataset includes daily readings from air quality monitoring stations in various cities across the U.S. and Canada, as well as metadata about the monitoring stations.
- **Link**: EPA Air Quality Data

Source 2: Electric Vehicle Adoption Data from the U.S. Department of Energy

- **Description**: This dataset provides information about the rise of electric vehicle adoption across major North American cities, including the number of registered electric vehicles by city, state, and year.
- **Data Content**: The dataset includes information about the number of electric vehicles in various U.S. cities and the rate at which they are being adopted over time.
- **Link**: DOE Electric Vehicle Data

Data Structure and Quality

- **Structure**: Both datasets are tabular. The air quality dataset contains columns for the pollutant type, location (city or station), date, and pollutant concentration. The electric vehicle dataset includes columns for city, year, and the number of electric vehicles registered.
- **Quality**: The air quality dataset is generally reliable but contains missing data for certain cities or time periods. The electric vehicle dataset has relatively
- consistent data, though some cities may have incomplete records, especially in earlier years.

Licenses

- **EPA Air Quality Data**: This dataset is publicly available under an open data license for non-commercial use.
- **DOE Electric Vehicle Data**: This dataset is licensed under an open data agreement for public use, particularly for research and educational purposes.
-

License Links:

- EPA License Information
- DOE License Information

Data Pipeline:

High-Level Overview

The data pipeline for this project is designed to automate the extraction, transformation, and loading of air quality and electric vehicle adoption data. The technologies used include:

- **Pandas**: For data manipulation and transformation.
- **SQLite**: To store the cleaned data for analysis.
- **CSV**: For backup storage of the cleaned data.

Transformation and Cleaning Steps

1. **Extracting Data**: Data is fetched from the EPA and DOE datasets using pd.read_csv() for CSV files.
2. **Cleaning Data**:
   - **Standardizing Column Names**: Columns are standardized to lowercase with underscores replacing spaces.
   - **Handling Missing Data**: Missing values in both datasets are filled using appropriate methods, such as forward-filling for missing air quality data and zero-imputation for missing EV data.
   - **Date Formatting**: Dates in the air quality dataset are standardized to YYYY-MM-DD format for consistency.
   - **Removing Duplicates**: Any duplicate rows, particularly in the air quality data, are removed.
3. **Loading Data**: The cleaned datasets are saved into both CSV files and SQLite databases for easy querying and integration with future analytical tools.

Problems Encountered and Solutions

- **Missing Data**: Both datasets had missing entries, particularly in earlier years for the electric vehicle data. For air quality, we used forward-filling for missing pollutant values where appropriate.
- **Date Format Inconsistencies**: The date format in the air quality data was inconsistent. This was resolved by converting all dates to YYYY-MM-DD format during the transformation step.

- **Data Normalization**: Electric vehicle data was provided by year and city, but the air quality data had city-specific readings at varying intervals. The solution was to normalize the air quality data by month for better comparison with EV adoption trends.

Meta-Quality Measures

- **Error Handling**: The pipeline includes error handling to log any missing files or data discrepancies during the extraction and loading phases.

- **Data Validation**: After loading the data, a validation step checks for outliers or unrealistic values (e.g., negative pollutant levels or incorrect EV counts) and flags them for review.

Result and Limitations:

Output Data

- **Air Quality Data**: The output dataset contains cleaned air quality data for pollutants such as CO2, NO2, and PM2.5 for cities in the U.S. and Canada, spanning multiple years.
- **Electric Vehicle Data**: This dataset contains the number of electric vehicles registered by city and year, showing the rise in adoption over time.

Data Quality

- **Before Cleaning**: The raw air quality data had missing entries for some pollutants in certain cities and inconsistent date formats. The EV adoption data had gaps in early years for some cities.
- **After Cleaning**: The cleaned air quality data is free of duplicates, missing values are handled, and date formats are consistent. The electric vehicle dataset is now complete, with all cities having consistent yearly data.

Data Format

- **CSV Files**: Both datasets are saved in CSV format for portability and future analysis.
- **SQLite Databases**: The data is also saved in SQLite format to enable easy querying and efficient storage.

Critical Reflection

- **Limitations**: The analysis is dependent on the accuracy of the data provided by external sources (EPA and DOE). While the datasets are fairly robust, discrepancies in early data years could introduce bias in the analysis.
- **Future Improvements**: Future versions of the pipeline could integrate real-time data collection and more sophisticated techniques for dealing with missing values. Additionally, integrating more cities or longer time frames could provide a more comprehensive analysis of the impact of EV adoption.