**Project Question:**

Are states with a higher adoption of electric vehicles in the USA experiencing a reduction in CO2 emissions from the transportation sector? Additionally, how do these emissions compare to vehicle-related CO2 emissions in Canada? To answer this question, I selected three datasets that focus on electric vehicles in the USA and CO2 emissions in both the USA and Canada. The first dataset, named electric_cars, is sourced from Data.gov, which provides open governmental data for the USA. The other two datasets, co2_can and co2_usa, were obtained from Kaggle. I chose these datasets because I am interested in exploring the impact of electric vehicles on CO2 emissions. By analyzing these data sources, I aim to uncover trends and draw comparisons between the adoption of electric vehicles and their effect on transportation-related emissions in the USA and Canada.

**Data Structure and Data Quality:**

All three datasets are structured data organized in a fixed schema, with each in CSV format.
Data Quality: **Completeness:** The electric_cars dataset has a small number of missing values—only 19 missing values for Electric Range and Base MSRP, and just 5 missing values for Electric Utility. The co2_usa and co2_can datasets, however, do not have any missing data.

**Timeliness:** The data in all three datasets is regularly updated. The co2_usa and co2_can datasets include data spanning over 40 years. This long-term data is essential for my analysis as it allows me to track how CO2 emissions have evolved over time, which is crucial for understanding trends and changes in vehicle-related emissions.

**Consistency:** I have also checked the Consistency of the datasets and found that all of them maintain a consistent format in each row. Occasionally, I identified some outliers, which I have addressed and fixed to ensure the accuracy and reliability of the data.

**Licenses and Permissions for Data Usage**

This project utilizes three datasets under different licenses, each specifying the terms and conditions for use, modification, and distribution:

**Dataset: electric_cars**
**License:** Open Database License (ODbL) v1.0
**Usage Rights:**

- Free to use, modify, and distribute the dataset, provided that the original dataset is attributed, and any derivative works are shared under the same license.
- You may use the dataset for any purpose, including commercial purposes, if the terms of the ODbL license are followed.
- Any modifications or adaptations of the dataset must be made available under the same ODbL license.

**Plan to fulfill license's obligations:**

- I will include the ODbL license text and attribute the source.
- Any modifications to the dataset will be shared under the same license.
- I will include a disclaimer of warranty ("as-is").

**Link of the License:** https://opendatacommons.org/licenses/odbl/1-0/
**Link of the Dataset:** https://catalog.data.gov/dataset/electric-vehicle-population-data/resource/fa51be35-691f-45d2-9f3e-535877965e69

**Dataset: co2_usa**
**License:** Apache License, Version 2.0
**Usage Rights:**

- Free to use, modify, reproduce, and distribute, including creating derivative works, provided the conditions of the license are met.
- The software or dataset must include a copy of the license and provide an attribution to the original source.

**Plan to fulfill license's obligations:**

- I will include the Apache License text in the project.
- Provide proper attribution to the dataset source.
- If modified, I will document changes and include a disclaimer of warranty ("as-is")

**Link of the License:** https://www.apache.org/licenses/LICENSE-2.0
**Link of the Dataset:** https://www.kaggle.com/datasets/abdelrahman16/co2-emissions-usa

**Dataset: co2_can**
**License:** MIT License
**Usage Rights:**

- Free to use, modify, merge, publish, distribute, sublicense, and/or sell copies of the dataset, if the original copyright notice and permission notice are included in all copies or substantial portions of the dataset.

**Plan to fulfill license's obligations:**

- I will include the MIT License text and provide attribution to the source.
- Include a disclaimer that the dataset is provided "as-is.

**Link of the License:** https://www.mit.edu/~amini/LICENSE.md
**Link of the Dataset:** https://www.kaggle.com/datasets/isaacfemiogunniyi/co2-emission-of-vehicles-in-canada

**Which technology did I use to implement it?**

I used Python as the programming language to extract, transform, and load the data, creating an automated pipeline. Initially, I used Jupyter Notebook to perform checks, such as identifying duplicates and missing data. I also created visualizations to detect outliers. Afterward, I moved to Microsoft Visual Studio to build the entire pipeline, enabling it to be executed in a single step.

**Cleaning:**

First, I dropped two columns from the *electric_cars* dataset, *DOL Vehicle ID*, as they were not relevant to my analysis. I also found some missing data in the *electric_cars* dataset. For columns where the missing data could not be reasonably inferred from the mode or mean of other data (e.g., *County*, *City*, *Postal Code*, and *Vehicle Location*), I filled these values with "Unknown." For missing data in the *Electric Utility*, *Electric Range*, and *Base MSRP* columns, I grouped the data by *Make* and *Model Year* and used the mode to fill in the missing values. This approach is based on the assumption that cars manufactured by the same company and with the same model year typically share similar motor performance characteristics. During the cleaning process, I also identified and handled outliers in the *Base MSRP* column. I defined outliers as any rows where the *Base MSRP* was greater than 200,000. For each of these outliers, I replaced the *Base MSRP* with the mode (most frequent value) of the *Base MSRP* for the same *Make* and *Model Year*.
In the *co2_can* dataset, I dropped three columns—*Fuel Consumption Hwy (L/100 km)*, *Fuel*

*Consumption City (L/100 km)*, and *Fuel Consumption Comb (mpg)*—since they were already represented in other columns using different measurement units. I also found no duplicates in any of the datasets.

**Meta-quality measures:**
In my data pipeline, I used the mode to handle missing values for the Electric Utility, Electric Range, and Base MSRP columns. For missing data, I grouped the dataset by Make (for Electric Utility) and by both Make and Model Year (for Electric Range and Base MSRP). I filled missing values with the most frequent value within each group, ensuring consistency for cars of the same make and model year. If the mode was empty, I set the value to None to avoid incorrect imputations.
The pipeline is designed to handle changes in the data, automatically adjusting to new manufacturers or model years. It ensures that missing values are filled appropriately, maintaining data integrity even if the input data changes or new values are added.

**Output Data of the Data Pipeline:** The pipeline produces three cleaned and processed datasets: co2_usa, cars, and co2_canada. These datasets are saved in an SQLite database, making them structured, organized, and ready for further analysis.

**Data Structure and Quality**

The output data is structured and cleaned, with duplicates removed and missing values addressed. For instance, missing values were filled based on grouped attributes such as Make and Model Year. Outliers, like unusually high Base MSRP values, were corrected using similar grouping logic. Overall, the data meets important quality criteria such as consistency, accuracy, and completeness.

**Chosen Data Format and Rationale**

I chose SQLite as the output format because it is simple and makes the data easy to store and query. This format works well for structured datasets and is convenient for integration with other tools.

**Critical Reflection on Data and Limitations:**

If the input data structure changes, some adjustments in the pipeline will be necessary.