Electricity and its Environmental Impact in Canada

DATA 604 Final Report

Yifeng Liang, UCID: 30072968
Gulshan Laskar, UCID: 30256534
Sai Namratha Govindu, UCID: 30262718
Hemanth Kaleswara Chowday Dhanekula, UCID: 30262699
Lalith Nandakumar, UCID: 30262992

Table of Contents

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

The continually evolving demand for electricity in Canada poses significant challenges and opportunities in the journey toward sustainable energy systems. This project explores how electricity consumption and generation impact the environment, focusing on key factors such as population growth, the adoption rate of electric vehicles (EVs), and the interplay between renewable and non-renewable energy sources.

As Canada strives to achieve net-zero emissions from electricity generation by 2035, understanding the dynamics of electricity demand and supply is critical. Population growth and the increasing popularity of EVs are reshaping electricity consumption patterns, placing pressure on existing generation capacities and influencing policy decisions. Concurrently, the balance between renewable and non-renewable energy sources directly affects greenhouse gas (GHG) emissions, a pivotal metric in assessing progress toward climate goals.

This analysis aims to determine whether trends in electricity consumption align with population growth and EV adoption rates. It also evaluates the current energy mix in Canada's electricity generation and its corresponding GHG emissions. By examining these factors, the project provides insights into whether Canada is on track to meet its environmental commitments while adapting to changing societal and technological landscapes.

1.1.2. Problem

As the world grapples with climate change, Canada's ambitious commitment to achieving net-zero emissions from electricity generation by 2035 highlights the critical need for informed decision-making in energy policy. Meeting this target requires a thorough understanding of the complex relationships between population growth, technological adoption (e.g., electric vehicles), electricity consumption trends, the balance of renewable versus non-renewable energy sources, and greenhouse gas (GHG) emissions.

1.1.3. Challenges

This study encounters several challenges stemming from the nature of the data, the complexity of the relationships between variables, and regional variations within Canada. These challenges are outlined below:

Factors influencing electricity generation, consumption, and greenhouse gas (GHG) emissions are often interdependent. For example, population growth drives increased energy demand, while the adoption of electric vehicles (EVs) impacts both electricity consumption and emissions reductions. These interrelationships complicate the isolation of individual factors. To address this, statistical

techniques such as multicollinearity tests (e.g., Variance Inflation Factor) and interaction models will be employed to better understand and adjust for these dependencies.

The integration of five diverse datasets poses significant challenges due to variations in structure, temporal coverage, and geographic granularity. For instance, population data spans 2016–2024, while GHG emissions data covers 2009–2021. Aligning these datasets requires meticulous preprocessing, including time-series alignment and unit standardization, to ensure consistency and facilitate accurate analysis.

Although the datasets provide key insights, they do not encompass all variables that impact electricity generation and emissions. For example, behavioral factors such as energy conservation practices or policy compliance levels are not included, limiting the comprehensiveness of the analysis. This constraint highlights the need for cautious interpretation of results.

Challenges related to data quality, such as missing values and inconsistent formats, are prevalent across the datasets. Additionally, varying levels of data granularity—such as electricity consumption data segmented by sector and province—necessitate advanced imputation techniques and thorough validation to mitigate the impact of these gaps.

Canada's energy landscape varies significantly by region due to differing resource availability and provincial energy policies. For instance, Quebec relies heavily on hydroelectric power, whereas Alberta predominantly uses fossil fuels. This regional variability requires the development of tailored models to capture provincial differences accurately, ensuring that analyses reflect the diversity of Canada's energy systems.

1.2 OBJECTIVES

1.2.1 Overview

The primary objective of this study is to analyze the interconnected trends of electricity demand and supply in Canada, with a focus on understanding their environmental implications. The research seeks to determine whether electricity consumption patterns align with population growth and the adoption rate of electric vehicles, both of which significantly influence energy demand. Furthermore, it evaluates the current energy mix of renewable and non-renewable sources in electricity generation, examining their respective contributions to greenhouse gas (GHG) emissions. This study also investigates regional and sectoral variations in electricity consumption and emissions, shedding light on disparities across provinces and industries. Ultimately, the analysis aims to assess whether Canada is progressing toward its 2035 goal of achieving net-zero emissions from electricity generation, providing insights into the effectiveness of current energy policies and identifying areas for improvement.

1.2.2 Goals & Research Questions

The primary goal of this project is to analyze the electricity landscape in Canada by examining key factors influencing electricity demand, generation, and consumption patterns. The study aims to investigate the interplay between population growth, electric vehicle adoption, greenhouse gas emissions, and electricity generation to provide insights into achieving sustainable energy practices. Through a comprehensive exploration of trends and relationships, this analysis seeks to address the challenges and opportunities for Canada's transition toward a greener future, including the feasibility of achieving net-zero electricity emissions by 2035.

To guide this analysis, several research questions were formulated to provide a focused approach to the study. The first area of inquiry examines the relationship between population growth and electricity usage across provinces in Canada. Specifically, it explores whether provinces with higher population growth rates are experiencing proportionally higher increases in electricity usage during the study period. This question aims to highlight the demographic factors driving electricity demand.

Another critical focus is the impact of electric vehicle adoption on electricity demand. This includes investigating how the adoption rate of electric cars influences electricity consumption at the provincial level and whether the increase in electricity usage outpaces the combined growth rates of population and electric car adoption. These questions are crucial for understanding the evolving energy needs driven by technological and lifestyle changes.

The study also delves into electricity generation patterns, identifying which province leads in electricity generation and analyzing the balance between renewable and non-renewable resources in the energy mix. An additional inquiry investigates whether the province with the largest electricity generation also produces the most greenhouse gas (GHG) emissions, highlighting potential trade-offs between energy production and environmental impact.

A key overarching question addresses the possibility of Canada achieving net-zero electricity emissions by 2035. This includes assessing current trends in population growth, electric car adoption, electricity demand, and generation. By projecting these trends, the study seeks to evaluate the feasibility of meeting ambitious sustainability goals within the stipulated timeline.

Finally, the project explores broader patterns in electricity consumption and vehicle adoption. This includes examining sectoral and provincial trends in electricity consumption from 2005 to 2023, as well as identifying shifts in the adoption of electric and alternative fuel vehicles compared to traditional fuel types such as gasoline and diesel. The study also considers seasonal or yearly changes in vehicle counts across different categories, providing a nuanced understanding of emerging trends in the transportation sector.

These research questions collectively form the foundation of the project, guiding the analysis and ensuring a comprehensive examination of Canada’s electricity ecosystem. By addressing these inquiries, the study aims to contribute meaningful insights to support evidence-based policymaking and sustainable energy initiatives.

2. METHODOLOGY

This section outlines the datasets utilized in the study and the methodological approach adopted for data cleaning, preprocessing, and exploratory analysis then merging the datasets.

2.1 Datasets

The project involved five datasets that provided comprehensive information about electricity generation, consumption, and related factors across Canada. These datasets, which include the Electricity Generated Dataset, Electricity Car Dataset, Population Dataset, Electricity Consumption Dataset, and Greenhouse Gas Emissions Dataset, were carefully examined and processed to ensure accuracy and relevance for the analysis.

Table 1: Original Datasets Overview and Attributes

Datasets Name	**Attributes**	**Attribute Type**	**Dataset Cleaned**
Population Dataset	'REF_DATE', 'GEO', 'DGUID', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS'	Object Int64 Float64	Yifeng
Electricity Generated Dataset	'REF_DATE', 'GEO', 'DGUID', 'Class of electricity producer', 'Type of electricity generation', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS'	Object Int64 Float64	Gulshan
Electricity Car Dataset	'REF_DATE', 'GEO', 'DGUID', 'Fuel type', 'Vehicle type', 'Statistics', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS', 'Other metadata'	Object Int64 Float64	Hemanth
Electricity Consumption	'Date', 'Province', 'Total_electricity', 'Residential_electricity', 'Commercial_electricity', 'Industrial_electriciry', 'Transportation_electricity'	Object Int64 Float64	Lalith
Greenhouse Gas Emissions	'REF_DATE', 'GEO', 'DGUID', 'Sector', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS'.	Object Int64 Float64	Namratha

From Table 1: Original Datasets Overview and Attributes, we proceed to explain the datasets used in this study in detail.

2.1.1 Dataset 1: Electricity Generated Dataset

This dataset from Statistics Canada includes detailed data on electricity generation by type across Canadian provinces, covering sources such as hydro, wind, nuclear, and more. Licensed under the Open Government License – Canada, it supports unrestricted analysis of regional electricity generation trends. Key features include columns for each energy type, generation capacity, and provincial designation. It is a CSV file with 65326 rows and 16 columns, the dataset is manageable for efficient processing and analysis. The 'REF_DATE' column represents the temporal aspect of the dataset, indicating the month and year for each record, stored as a string. Geographical information is captured in the 'GEO' column, which identifies the location of electricity generation, ranging from the national level ("Canada") to individual provinces. Each record is uniquely associated with a 'DGUID' (Geographic Unique Identifier), also stored as a string.

The dataset categorizes electricity producers through the 'Class of Electricity Producer' column, which specifies whether the data corresponds to total electricity producers or a specific subset. The method or source of electricity generation is detailed in the 'Type of Electricity Generation' column, providing classifications such as hydraulic turbine, nuclear steam turbine, and conventional steam turbine. Both these columns are stored as strings.

The measurement of electricity generated is represented in the 'UOM (Unit of Measure)' column, which standardizes data in megawatt-hours (MWh). The associated numerical code for the unit is recorded in the 'UOM_ID' column as an integer. A scaling factor applied to the data values is indicated in the 'SCALAR_FACTOR' column (stored as a string), with its corresponding code provided in the 'SCALAR_ID 'column (stored as an integer).

Additionally, the dataset includes a 'VECTOR' column, which acts as a unique identifier for each data series. The electricity generation figures themselves are captured in the 'VALUE' column as floating-point numbers, with some entries containing missing values. Metadata about the status and annotations for specific entries are found in the 'STATUS' and 'SYMBOL' columns, although these columns have a significant proportion of missing values. A 'TERMINATED' column, which is uniformly null in this dataset, may signify records marked as discontinued or inactive. Lastly, the 'DECIMALS' column indicates the level of precision for the recorded data, stored as an integer.

2.1.2 Dataset 2: Electricity Car Dataset

The project investigates patterns and relationships in vehicle data, focusing on fuel types, vehicle types, and usage statistics over time in Canada. The dataset provides insights into trends in fuel consumption, vehicle adoption, and shifts towards sustainable energy. This dataset aims about the number of vehicles used by the people in the Canada region, this dataset not only contains the information about electric vehicles but also provides information about the gasoline vehicle as well as hybrid vehicles. The CSV file provides data on vehicle counts and related statistics, organized by fuel type, vehicle type, geographical region, and time period. It includes detailed information about the number of vehicles in various categories, helping to understand the composition and trends of the automotive landscape over time. This dataset contains 11,550 rows and 17 columns. The dataset includes, 'REF_DATE', it provides the time period of the data (e.g., monthly intervals

starting in 2017), ‘GEO’ this column provides the geographic region (e.g., Canada), ‘DGUID’ this is a unique identifier for geographical and administrative boundaries.

The most important attributes which we used for this project are, ‘Fuel type’ it contains Types of fuel, such as gasoline, diesel, hybrid electric, or battery electric, ‘Vehicle type’ indicates the Categories of vehicles (e.g., passenger cars, pickup trucks, vans). ‘Statistics’ contains the Type of metric, such as the number of vehicles. ‘VALUE’ is the measured value corresponding to the combination of fuel and vehicle type. ‘Other metadata’ has Units of measure, scalar factors, and vector codes for data processing. These different attributes are of Object, Int64 and Float64 data types.

Additionally, ‘UOM’ represents the units, ‘UOM_ID’ indicates the id for each unit (it has the same value), ‘SCALAR_FACTOR’ also indicates the units whereas ‘SCALAR_ID’ represents the id for each unit (it has the same value), ‘VECTOR’ represents the int values which are not useful in this dataset and ‘COORDINATE’ has the float values which represent the coordinate values which are considered as outliers. ‘VALUE’ contains the measure of each column but it’s also an unused attribute in the dataset. ‘STATUS’, ‘SYMBOL’, ‘TERMINATED’ all the attributes are null values so we removed those attributes. ‘DECIMALS’ contains all the values as zero.

2.1.3 Dataset 3: Population Dataset

Population dataset is from Statistics Canada, and it covers annual population counts by province across Canada from 2016 to 2024. The license is Open Government License – Canada, which permits free use. This dataset includes date, geographic information like provinces, and number counts of population which we are particularly interested in. Population dataset is available as a CSV file in 14 columns, and 463 rows well formatted table, and it is stored in one dataset file.

The dataset under consideration provides detailed population statistics across Canada and its provinces from 2016 to 2024. It consists of 462 observations, with each entry corresponding to a specific geographical region and a particular time period. The temporal aspect of the dataset is captured by the column "REF_DATE," which records the reference year and quarter (e.g., "2016-07"). The geographical information is represented by the column "GEO," which identifies locations such as "Canada" or individual provinces. Additionally, a unique geographic identifier is provided in the "DGUID" column, ensuring unambiguous identification of each region.

The core population data is contained in the "VALUE" column, which records the population counts for each region and time. These values are measured in units of "Persons," as specified in the "UOM" column, with a uniform measurement unit identifier provided in the "UOM_ID" column (value 249 across all entries). To facilitate consistency in data interpretation, the "SCALAR_FACTOR" column specifies that no scaling has been applied to the population counts, denoted as "units," and its corresponding identifier, "SCALAR_ID," consistently holds the value 0.

The dataset includes several metadata attributes, such as "VECTOR" and "COORDINATE," which act as statistical markers, with the former always set to "v1." Furthermore, the column

"DECIMALS" indicates the precision of the population data, and its value, set uniformly to 0, confirms that the population counts are reported without decimal places. However, the dataset also includes columns such as "STATUS," "SYMBOL," and "TERMINATED," which contain no data for any of the observations. These empty columns appear to be placeholders or relics from the dataset's original structure and do not contribute to the analysis.

Overall, the dataset is structured to provide essential information for understanding population dynamics across Canada. However, several columns, including "STATUS," "SYMBOL," "TERMINATED," "SCALAR_ID," and "UOM_ID," are redundant or non-informative in the context of this analysis. After appropriate cleaning and preprocessing, the dataset can be effectively utilized to analyze temporal and geographical trends in Canada's population over the designated period.

2.1.4 Dataset 4: Electricity Consumption Dataset

Electricity_Use dataset comprises seven subsets from Statistics Canada, one for each province, along with a combined dataset detailing annual electricity consumption across Canada from 2005 onward. It is licensed under the Open Government License – Canada, permitting free use. Each dataset includes features for the year, province, sector, and quantity of electricity consumed per hour in kilowatts. The datasets are available as individual .xlsx files for each province. Each individual dataset contains 35 features and 45 records even though the data is transposed, the features and records are labeled accordingly.

2.1.5 Dataset 5: Greenhouse Gas Emissions Dataset

This dataset from Statistics Canada includes detailed data on greenhouse gas emission by sector across Canadian provinces, licensed under the Open Government License. It is a CSV file with 23423 rows and 13 columns, from 2009 – 2021 among all sectors round Canada. The ‘REF_DATE’ column indicating the year for each record, stored as a string. Geographical information is captured in the ‘GEO’ column, which identifies the individual provinces in Canada. Each record is uniquely associated with a ‘DGUID’ (Geographic Unique Identifier), also stored as a string.

The dataset includes ‘UOM (Unit of Measure)’ column, which standardizes data in kilotons. The associated numerical code for the unit is recorded in the ‘UOM_ID’ column as an integer. A scaling factor applied to the data values is indicated in the ‘SCALAR_FACTOR’ column (stored as a string), with its corresponding code provided in the ‘SCALAR_ID ‘column (stored as an integer). the ‘VECTOR’ column, acts as a unique identifier for each data series. The emission figures named as ‘VALUE’ column as floating-point numbers. The ‘STATUS’ and ‘SYMBOL’ columns, with significant missing values. A ‘TERMINATED’ column is an inactive cell. Lastly, the ‘DECIMALS’ column as helping hand to ‘VALUE’ column. For better analysis and combining with other datasets we decided to maintain the data from the year 2017.

2.2 Approach

2.2.1 Data Cleaning and Preprocessing

Dataset 1: Electricity Generated Dataset

The initial dataset, consisting of 65,326 rows and 16 columns, presented several challenges that required careful cleaning and restructuring to ensure it was suitable for analysis. Redundant columns were a significant issue, as several contained constant or uninformative values. For instance, columns such as 'UOM_ID', 'SCALAR_ID', 'SCALAR_FACTOR', and 'DECIMALS' exhibited fixed values that provided no additional insight. Similarly, the 'TERMINATED' column contained only null values, and the 'STATUS' and 'SYMBOL' columns had an overwhelmingly high proportion of missing entries, with 77% and 99% of their values absent, respectively.

Metadata fields like 'VECTOR' and 'COORDINATE' were also deemed irrelevant for the scope of this analysis and subsequently removed.

Another critical issue was the presence of negative values in the 'VALUE' column, which represents electricity generation. These entries were logically invalid and were replaced with null values before being removed. The dataset also required formatting adjustments like the 'REF_DATE' column, initially stored as a string, was converted to a datetime format to facilitate temporal analysis. Additionally, several columns, such as 'GEO' and 'Class of electricity producer', were converted to categorical data types to optimize memory usage and streamline analysis.

Addressing missing values was another key step in the cleaning process. Approximately 23% of the entries in the 'VALUE' column were missing, representing a significant gap in the dataset. Handling these missing values required careful consideration, as imputing them could introduce substantial bias. The 'VALUE' column reflects specific electricity generation measurements tied to unique temporal, spatial, and production type contexts. Variability in electricity generation across provinces, years, and production types means that imputing values, such as using the mean or median, risked distorting the data by masking genuine patterns or trends. To maintain analytical integrity and ensure the accuracy of insights, rows with missing values in the 'VALUE' column were removed.

After removing the missing values, the dataset was fully cleaned, with no missing entries remaining. The final cleaned dataset comprised 49,405 rows and 6 columns. Each column was verified to contain only non-null values, ensuring the dataset's integrity for further analysis. The 'REF_DATE' column was converted to a datetime format, while the 'GEO,' 'Class_of_electricity_producer,' and 'Type_of_electricity_generation' columns were stored as categorical data types to optimize memory usage. The 'DGUID' column retained its object data type, and the 'VALUE' column was maintained as a float for accurate numerical computations. The cleaned dataset occupied approximately 1.7 MB of memory, reflecting the efficiency achieved through data type conversions.

The dataset also underwent category simplification to improve readability and usability. In the 'Class of electricity producer' column, verbose categories were renamed; for instance, "Total all classes of electricity producer" was simplified to "Total all classes," while "Electricity producers, electric utilities" and "Electricity producers, industries" were shortened to "Utilities" and "Industries," respectively.

After cleaning, the dataset was filtered to focus on data from 2017 to 2024, aligning with the project's objective of analyzing recent electricity generation patterns. An additional 'Year' column was derived from 'REF_DATE' to facilitate aggregation and analysis at the annual level. The 'VALUE' column, which originally contained monthly electricity generation data, was aggregated to calculate annual electricity generation values. For each unique combination of 'Class_of_electricity_producer' and 'Type_of_electricity_generation', the monthly values were summed, providing a consolidated annual figure. This transformation ensured that the dataset reflected yearly electricity generation metrics, simplifying trend analysis and facilitating comparisons across years.

Following this aggregation, the 'VALUE' column was renamed to 'Electricity_VALUE_Annual' to clearly represent the annualized electricity generation figures. The 'REF_DATE' column was replaced by the 'Year' column to further standardize the dataset for yearly analysis. The cleaned and restructured dataset was then saved in a new CSV file, titled 'Cleaned_Electricity_Generated_Dataset.csv', imported into an SQL database ensuring it was readily accessible for subsequent analysis and advanced integration with other datasets. This final dataset contained 4,368 rows and 5 columns, each with complete and non-null data. The columns included 'Year' (stored as an integer), 'GEO,' 'Class_of_electricity_producer,' and 'Type_of_electricity_generation' (all stored as categorical data types), and 'Electricity_VALUE_Annual' (stored as a float to represent annual electricity generation values). Through efficient restructuring and the use of categorical data types, the memory usage of the dataset was optimized to approximately 65.5 KB, ensuring it was lightweight and ready for advanced analysis.

Dataset 2: Electricity Car Dataset

Initially original dataset contains **11**,550 entries and 17 attributes, capturing detailed information on vehicle statistics across Canada. Other attributes include metadata like 'DGUID' (geographical identifiers), 'Statistic**s'** (constant as "Number of vehicles"), 'STATUS', and 'SCALAR_FACTOR'. The dataset contains redundant or sparse columns, such as 'SYMBOL' and 'TERMINATED', which are entirely null. The original dataset had 17 columns, many of which were irrelevant or redundant such as 'SYMBOL', 'TERMINATED', 'STATUS'. These were removed to retain only the essential attributes: 'REF_DATE', 'GEO', 'Fuel type', and 'VALUE'.

Columns like 'SYMBOL' and 'TERMINATED' were entirely null, and other fields, such as 'DGUID', had significant missing values. By removing such columns, the dataset became more concise and easier to analyze. The 'REF_DATE' column was standardized into numeric years for better temporal analysis. Redundant identifiers like 'SCALAR_FACTOR' and 'SCALAR_ID' were dropped, as they added no analytical value. The original dataset contained 11,550 row**s**, many of which were aggregated or not directly relevant to the analysis of electric car trends. The cleaned dataset, with 2,310 rows, focuses on meaningful and specific entries, improving processing time and analysis quality.

By removing these columns, the dataset became cleaner and more focused on the usable data. The 'REF_DATE' column was originally in a "YYYY-MM" format. Converting it to a year-only numeric format improved consistency and reduced redundancy for trend analysis. The 'Fuel type' attribute retained its original categories but ensured alignment for comparing vehicle counts across fuel types. The cleaned dataset specifically supports the project's goal which is analyzing the trends in electric car adoption and comparing it with other fuel types.

Non-essential rows and columns that didn't directly contribute to this objective were eliminated, streamlining the dataset for efficient and targeted analysis. The project's aim is to compare the trend of electric car usage over time against other fuel types. Cleaning ensured that only the data relevant to this

goal was retained, removing distractions and inconsistencies. By cleaning the dataset, the information became easier to work with, more reliable, and better aligned with the objectives of the project. The size of the dataset is 89.53KB. This compact size reflects the streamlined structure, focusing only on the essential data for analysis.

Dataset 3: Population Dataset

The data cleaning process involved several critical steps to prepare the dataset for analysis. Initially, the dataset was loaded into a pandas DataFrame for processing, ensuring compatibility with Python-based data manipulation techniques. To streamline the dataset and focus on relevant information, only three columns—"REF_DATE," "GEO," and "VALUE"—were selected, as these represent the temporal, geographical, and population data required for the analysis.

The "REF_DATE" column, which originally contained date information in a string format, was converted to a datetime data type to enable efficient temporal filtering and manipulation. This transformation allowed for filtering data entries where the month was July, focusing on mid-year population statistics. To further standardize the temporal information, the "REF_DATE" column was adjusted to retain only the year component, enhancing clarity and simplifying analysis.

Subsequently, the "VALUE" column, which represents population counts, was renamed to "Population" for improved clarity and interpretability. This step ensures consistency and aligns the dataset with the intended focus of the analysis. A check for missing values was performed to identify potential data gaps, which is a crucial step in ensuring the reliability and completeness of the dataset.

Finally, the cleaned and filtered dataset was exported to a new CSV file, "Population_Canada.csv," preserving the modifications for subsequent analysis. This export ensures that the refined dataset can be easily shared and reused in future research, with all unnecessary information removed and the data appropriately structured. Overall, the cleaning process enhanced the dataset's usability, clarity, and analytical focus, enabling precise exploration of Canada's population trends.

Dataset 4: Electricity Consumption Dataset

Initially the dataset comprised of seven subsets from Statistics Canada, one for each province, along with a combined dataset detailing annual electricity consumption across Canada from 2005 onward. It is licensed under the Open Government License – Canada, permitting free use. Each dataset includes features for the year, province, sector, and quantity of electricity consumed per hour in kilowatts. This dataset will be used to uncover insights into the proportion of electricity utilized by year, sector, and province, with these key features serving as the primary focus in data exploration.

After cleaning the dataset has year, province and 4 sectors (Residential, Commercial, Industrial, Transportation) and electricity consumed per hour in kilowatts from the year 2005 to 2050 (where the data from 2024-2050 are forecasted data synthesized by Canada Energy Regulator)

This dataset will be used to uncover insights into the proportion of electricity utilized by year, sector, and province, with these key features serving as the primary focus in data exploration.

This dataset was cleaned by dropping various columns like 'Natural Gas', 'RPP', 'Biofuels', 'Hydrogen', 'Other' from each sector as we only concentrate on electricity part of the dataset and then transposing it pivoting, renaming the columns and merging it in such a way that it has the electricity consumed in all 4 sectors Residential, Commercial, Industrial, Transportation and total electricity consumed for each province

and year from 2005 to 2050. (the data from 2024 to 2050 are synthetic data forecasted by the Canada Energy Regulator)

Dataset 5: Greenhouse Gas Emissions Dataset

As listed in the description of dataset, the greenhouse gas emission data set contains inactive (null) columns, some missing values and most importantly the repeated columns. So, the cleaning process began by inspecting the dataset for missing values, inconsistencies, and outliers. Missing values were addressed by either imputation or removal, depending on the proportion and significance of the missing data. Then I jumped into identifying outliers, and found out that the outliners don't have significant affect during the analysis. The 'VALUE' column has some missing values, I removed them to make things easier. Then I moved to columns like 'DECIMALS', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', having repeated values so i directly removed these columns, whereas columns like 'STATUS', 'SYMBOL', 'TERMINATED' all have null values so I dropped them. Then came the 'DGUID' column having the same meaning as Geo location, so I just terminated the column. These steps ensured a clean, reliable dataset for accurate modeling and analysis.

2.2.2 Exploratory Data Analysis (EDA)

EDA was performed to uncover patterns and trends across all datasets. Techniques such as aggregation, visualization, and statistical summaries were applied to understand various dynamics related to electricity generation, consumption, and their environmental impact. For the Electricity Generated Dataset, the focus was on comparing renewable versus non-renewable production across different provinces and years. The Electricity Car Dataset was analyzed to assess the impact of electric vehicle adoption on electricity demand patterns. The Population Dataset was explored to examine correlations between population growth and electricity consumption trends. The Electricity Consumption Dataset provided insights into regional disparities in electricity use, while the Greenhouse Gas Emissions Dataset helped evaluate the relationship between electricity generation and environmental impact. Specific attention was also given to temporal trends, highlighting changes in electricity generation and consumption over time.

As part of the exploratory data analysis, we performed five distinct queries, one for each of the datasets, to explore various aspects of electricity generation, consumption, and related factors. Each query was designed to uncover patterns and provide insights into how different variables interact across time, location, and energy sources. In this section, we will discuss each query and its corresponding analysis in detail, covering all five datasets: Electricity Generated, Electricity Car, Population, Electricity Consumption, and Greenhouse Gas Emissions.

Dataset 1: Electricity Generated Dataset

As part of the exploratory data analysis for the Electricity Generated dataset, we performed five distinct queries to uncover various trends in electricity generation across Canada. These queries aimed to analyze electricity generation patterns, identify shifts in energy production types (renewable vs. non-renewable), and explore regional disparities. The queries were designed to answer critical questions regarding electricity generation across provinces over time, focusing on

factors such as energy source types, trends in generation volume, and the temporal distribution of electricity production.

Below, we elaborate on the queries:

Query 1: Annual Trends in Electricity Generation by Type Across Provinces

To explore these trends, we used a SQL query to group the data by year, province, and type of electricity generation. This allowed us to calculate the total electricity generation for each combination of these variables, providing a comprehensive overview of how electricity generation evolved over time in different provinces. The data was aggregated by summing the 'Electricity_VALUE_Annual' values, which represent the electricity generation for each specific year. The query results were ordered by year, province, and type of electricity generation to facilitate a chronological and regional analysis.

The purpose of this query was to examine yearly patterns in electricity generation across provinces, with a specific focus on identifying the types of electricity generation (renewable vs. non-renewable) that dominate in each region. By aggregating the data in this way, we could observe how the share of renewable and non-renewable energy sources changed over time, providing insights into regional disparities and shifts towards cleaner energy. The query was executed using the following SQL statement is below:

The query in figure 1 groups the data by three key dimensions 'Year', 'Province', and 'Type of electricity generation' and calculates the total electricity generation for each combination. By doing this, we are able to capture the annual changes in the total electricity generated in each province, as well as the relative contribution of different types of electricity generation over time.

From the results of this query, we gained several key insights. First, we were able to identify yearly trends in electricity generation across Canadian provinces. For example, we observed that provinces like 'Quebec' and 'Ontario' consistently produced high volumes of electricity, with Quebec having a significant reliance on renewable energy sources such as hydroelectric power. In contrast, provinces like 'Alberta' and 'Saskatchewan' had a greater reliance on non-renewable energy sources such as coal and natural gas.

Figure 1: SQL Query to Aggregate Electricity Generation Data by Year and Province

```
query_1 = """
SELECT
    Year,
    GEO AS Province,
    Type_of_electricity_generation,
    SUM(Electricity_VALUE_Annual) AS Total_Electricity_Generation
FROM
    Electricity_Generated_Data_by_provience
GROUP BY
    Year, GEO, Type_of_electricity_generation
ORDER BY
    Year ASC, Province ASC, Type_of_electricity_generation ASC;

"""
```

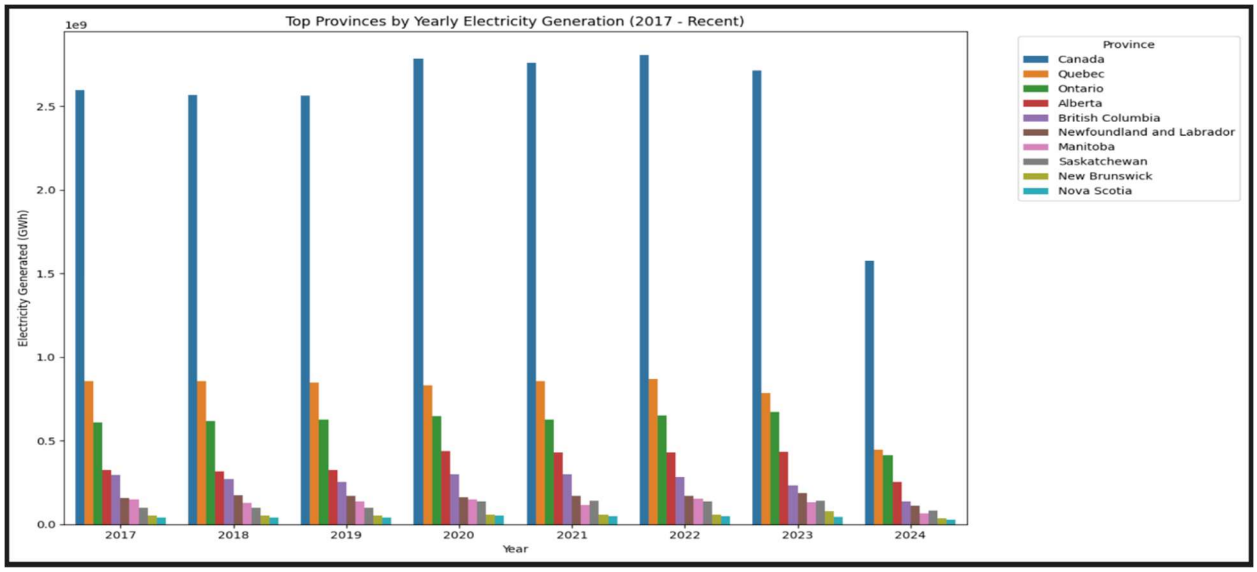
This insight helped us understand the energy mix in different regions and the extent to which each province is transitioning towards renewable energy.

Moreover, the query allowed us to observe how certain provinces had increased or decreased their electricity generation over time, particularly in response to economic, policy, and technological changes. For example, British Columbia showed a steady increase in renewable electricity generation, while other provinces displayed fluctuations in their energy mix depending on the availability of natural resources and regional policies.

To better visualize these trends, we plotted the results using a bar plot that showed the total electricity generation by year and province. The visualization clearly highlights the top provinces by electricity generation, allowing for an easier comparison of regional disparities and trends over time. In the plot, different provinces are represented by different colors, and the total electricity generation is plotted on the y-axis, with years on the x-axis. The plot provides a clear view of the changes in electricity generation over time, with an emphasis on identifying provinces that contributed significantly to the national electricity generation in figure 2.

This visualization in figure 2 complements the SQL query by providing a clear, graphical representation of the data, making it easier to identify key trends, changes, and regional variations. Notably, Quebec and Ontario consistently dominate electricity production, contributing significantly more than other provinces throughout the observed period from 2017 to 2024. This observation underscores the substantial infrastructure and natural resources these provinces have dedicated to electricity generation, making them vital players in meeting Canada’s national energy demands. While Canada's total electricity generation appears relatively stable over the years, there is noticeable variation among the provinces. Smaller provinces, such as Nova Scotia, New Brunswick, and Saskatchewan, contribute significantly less to the overall electricity generation.

Figure 2: *Annual Trends in Electricity Generation by Type Across Provinces (2017 - Recent)*



These disparities may be influenced by differences in population size, energy resource availability, and provincial energy policies. Additionally, provinces like British Columbia and Manitoba, which rely heavily on renewable resources such as hydroelectric power, show steady generation trends compared to fossil-fuel-reliant provinces like Alberta. This reinforces the idea that resource availability and energy mix play critical roles in determining provincial energy outputs.

A subtle trend of decreased electricity generation in certain provinces over the years aligns with the challenges previously identified in the analysis. This decline highlights the growing pressure to expand generation capacity, particularly in response to rising demand from population growth and electric vehicle (EV) adoption. It also underscores the urgency of diversifying electricity sources to mitigate dependency on fossil fuels, reduce greenhouse gas emissions, and align with Canada's net-zero ambitions.

Query 2: Total Electricity Generation for Each Province and Ranks

As part of the exploratory data analysis, our second query sought to identify the total electricity generation across various provinces in Canada. This analysis aimed to rank the provinces based on their total electricity generation and uncover the distribution of generation capabilities across the country. By summing the annual electricity generation values for each province, this query aimed to provide a comprehensive overview of which provinces are the largest contributors to Canada's overall electricity supply.

The SQL query executed for this analysis aggregated the Electricity_VALUE_Annual data by province, summing the total electricity generation for each province over the available years. The provinces were ranked in descending order based on their total electricity generation, and the top 11 provinces were selected for further analysis. The query used is outlined as below:

Figure 3: SQL Query for Total Electricity Generated for Each province and their ranks.

```
query_2 = '''
SELECT
    GEO AS Province,
    SUM(Electricity_VALUE_Annual) AS Total_Electricity_Generated
FROM
    Electricity_Generated_Data_by_provience
GROUP BY
    GEO
ORDER BY
    Total_Electricity_Generated DESC
LIMIT 11;
'''
```

The primary objective of this query was to rank the provinces based on their electricity generation, highlighting the regions that contribute the most to the national energy supply. By aggregating the data in this way, we were able to capture the total generation figures and observe how different provinces compare in their electricity production capacity. The results were then ordered in descending order of total generation, providing a clear ranking of the largest electricity producers in Canada.

The results of this query revealed significant disparities in electricity generation across Canada's provinces. Quebec emerged as the undisputed leader, generating a total of 6.35 billion units. This can be attributed to Quebec's vast hydroelectric resources, which allow the province to generate a substantial amount of renewable energy. Quebec's dominance in electricity generation underscores its important role in Canada's renewable energy sector.

Following Quebec, Ontario ranked second with 4.86 billion units of electricity generated. Ontario's energy mix is diverse, consisting of a combination of nuclear, hydroelectric, and natural gas generation. This diversification in energy sources has allowed Ontario to maintain a high level of electricity generation despite the ongoing transition to cleaner, more sustainable energy.

Alberta ranked third, with a total generation of 2.95 billion units. Alberta primarily relies on fossil fuels such as natural gas and coal for its electricity generation, although the province has been making strides in increasing its renewable energy capacity in recent years. The dominance of fossil fuels in Alberta's energy mix is reflected in its ranking as the third-largest producer of electricity in Canada.

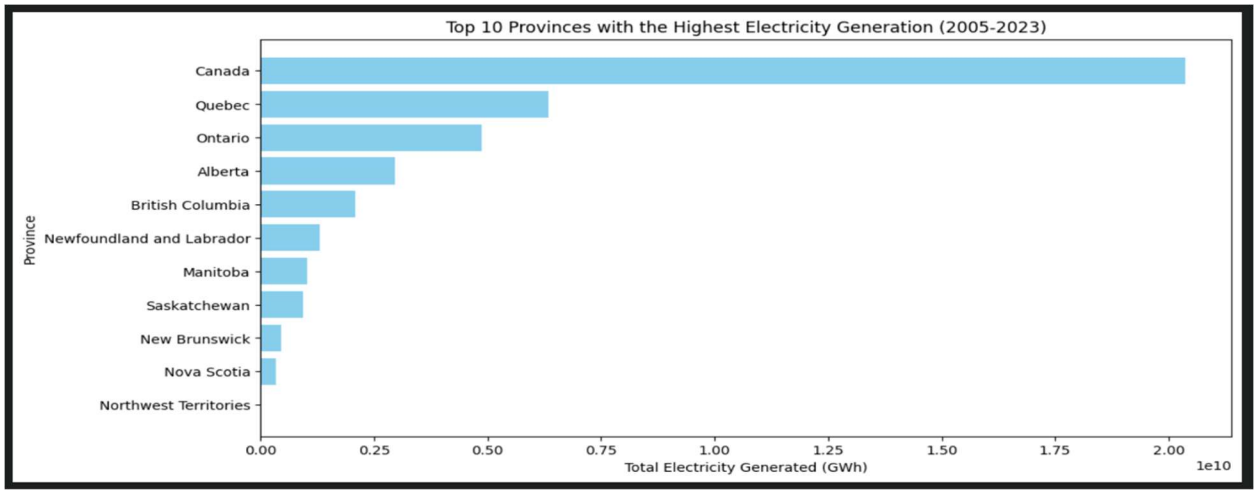
British Columbia ranked fourth, generating 2.08 billion units of electricity. The province's geography, particularly its mountainous terrain, is highly conducive to hydroelectric power generation, which forms the backbone of British Columbia's electricity generation strategy. Other provinces such as Newfoundland and Labrador, Manitoba, and Saskatchewan also play important roles in Canada's electricity generation, although their outputs are significantly smaller than those of Quebec and Ontario. Newfoundland and Labrador generated 1.30 billion units, primarily from hydroelectric resources, while Manitoba generated 1.03 billion units, largely from the same source. Saskatchewan, with 929.45 million units, continues to rely heavily on fossil fuels, although its renewable sector is expanding.

Provinces like New Brunswick and Nova Scotia contribute less to the national generation total, with 450.27 million units and 344.77 million units, respectively. The smaller provinces have less

generation capacity, often relying on a combination of fossil fuels and renewable energy sources. At the bottom of the list, Northwest Territories generated only 25.78 million units, reflecting the province's smaller population and energy demands.

To visualize these results, we created a horizontal bar chart that ranks the top 10 provinces based on their total electricity generation (figure 4). This visualization clearly illustrates the disparities between the leading provinces and those with smaller contributions. The chart is organized with the largest values at the top, making it easy to compare the total electricity generation for each province.

Figure 4: Top 10 Provinces with the Highest Electricity Generation (2005-2023)



The bar chart above (figure 4) provides a visual representation of the total electricity generation for each of the top 10 provinces, highlighting the magnitude of contribution from Quebec, Ontario, and Alberta. The horizontal bars make it easy to compare the provinces side by side, while the inversion of the y-axis places the largest electricity generators at the top for easier identification.

Query 3: Proportional Balance Between Renewable and Non-Renewable Energy by Province

The third query in our analysis aimed to explore the proportional balance between renewable and non-renewable electricity generation across different provinces in Canada. This analysis was crucial in understanding the energy mix in each province and how each contributes to the ongoing transition to renewable energy sources. By categorizing electricity generation into renewable and non-renewable sources, we could evaluate how each province is balancing these sources to meet its energy demands and how this mix contributes to Canada's overall sustainability goals.

To achieve this, the SQL query used a conditional summation approach. Renewable energy was defined as electricity generated from hydraulic turbines, wind turbines, and solar power, while non-renewable energy encompassed sources such as coal, natural gas, and nuclear. The query calculated the total renewable and non-renewable generation for each province and the total electricity generated by summing the values across all energy types. The SQL code is as follows:

Figure 5: SQL Query for Proportional balance between renewable and non-renewable energy by province

```
query_3= '''
SELECT
    GEO AS Province,
    SUM(CASE WHEN Type_of_electricity_generation IN ('Hydraulic turbine',
    'Wind turbine', 'Solar') THEN Electricity_VALUE_Annual ELSE 0 END) AS Renewable_Generation,
    SUM(CASE WHEN Type_of_electricity_generation NOT IN ('Hydraulic turbine',
    'Wind turbine', 'Solar') THEN Electricity_VALUE_Annual ELSE 0 END) AS Non_Renewable_Generation,
    SUM(Electricity_VALUE_Annual) AS Total_Generation
FROM
    Electricity_Generated_Data_by_provience
GROUP BY
    GEO
ORDER BY
    Total_Generation DESC;
'''
```

The results of this query were used to generate two key visualizations: a stacked bar chart and a heatmap. These visualizations provide insight into the distribution of renewable and non-renewable electricity generation across provinces, helping to identify regions that rely heavily on non-renewable energy sources and those that have successfully integrated renewable energy into their energy mix.

The stacked bar chart in figure 6 below visualizes the proportional balance between renewable and non-renewable generation for each province. In this chart, the height of each bar represents the total electricity generation in a province, and the segments of each bar represent the proportions of renewable and non-renewable generation. Quebec, for example, stands out with a large proportion of its electricity generated from renewable sources, primarily from hydraulic turbines. This is in line with Quebec's rich hydropower resources, which contribute significantly to the province's overall electricity generation. Ontario, on the other hand, presents a more balanced energy mix. While the province generates substantial renewable energy, primarily from wind and solar power, a considerable share still comes from non-renewable sources, particularly nuclear and natural gas. This balance reflects Ontario's diverse energy infrastructure and its ongoing efforts to reduce its reliance on fossil fuels.

In contrast, British Columbia shows a considerable share of renewable energy in its generation mix, largely due to its abundant hydropower resources. The province's mountainous terrain makes it ideally suited for hydropower generation, which is a key contributor to its electricity generation. Smaller provinces like Prince Edward Island and Yukon, while contributing relatively little to total national generation, still maintain a mix of renewable and non-renewable sources. However, the renewable generation share in these provinces is relatively lower due to their smaller energy infrastructure and lower total generation capacity.

Figure 6: Bar chart to show Proportional Balance Between Renewable and Non-Renewable Electricity Generation by Province

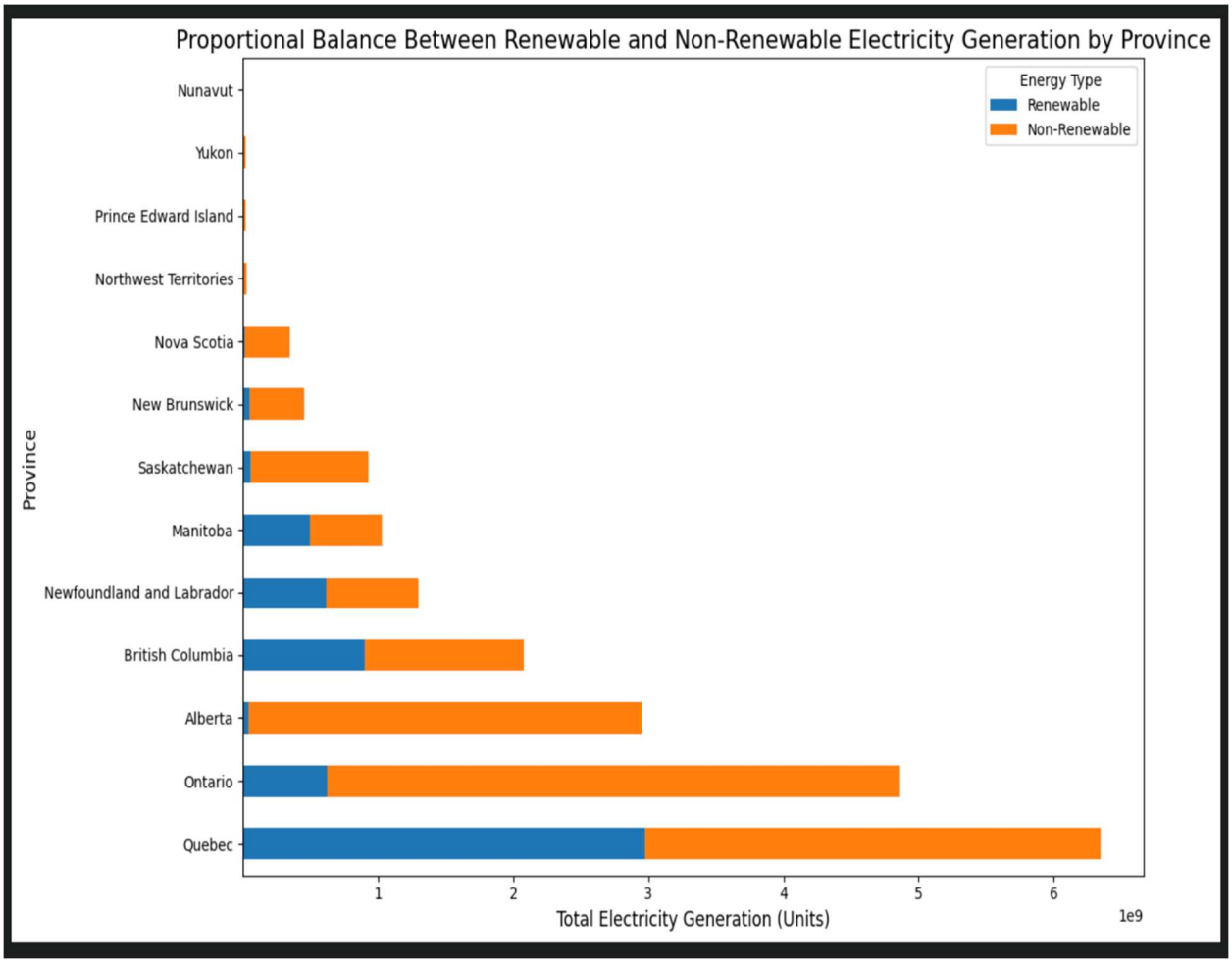
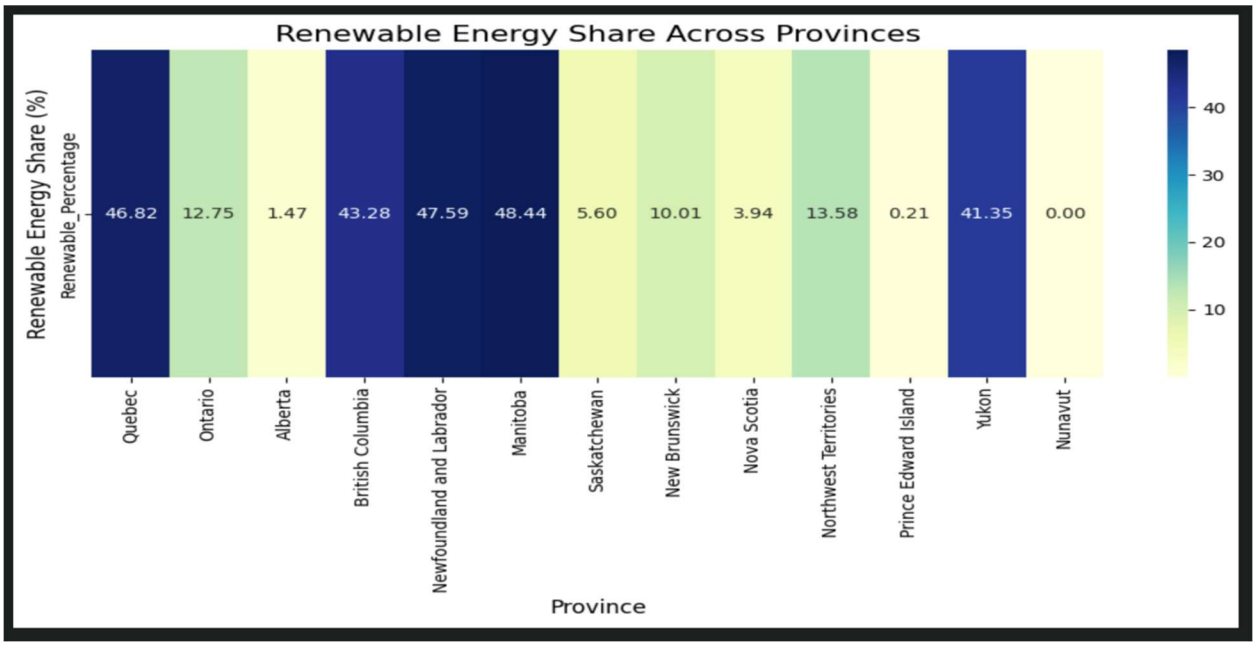


Figure 7: Heatmap showing percentage of renewable generation for each province



In addition to the stacked bar chart in figure 6, the percentage of total electricity generated from renewable sources was calculated for each province. This percentage was visualized using a heatmap in figure 7, where darker shades of blue represent provinces with a higher share of renewable energy in their electricity generation. Quebec, once again, stands out with the highest renewable energy share, while provinces like Alberta and Saskatchewan are represented by lighter shades, indicating their lower renewable generation share. The heatmap effectively illustrates the disparities in renewable energy adoption across Canada's provinces.

The results of this query underscore the ongoing transition in Canada's energy landscape. While provinces like Quebec and British Columbia demonstrate a clear commitment to renewable energy, others, such as Alberta and Saskatchewan, are still heavily dependent on non-renewable sources. These findings are crucial for understanding the regional variations in Canada's energy mix and the challenges faced by certain provinces in achieving a more sustainable energy future.

Query 4: Peak electricity Generation Provinces by Type

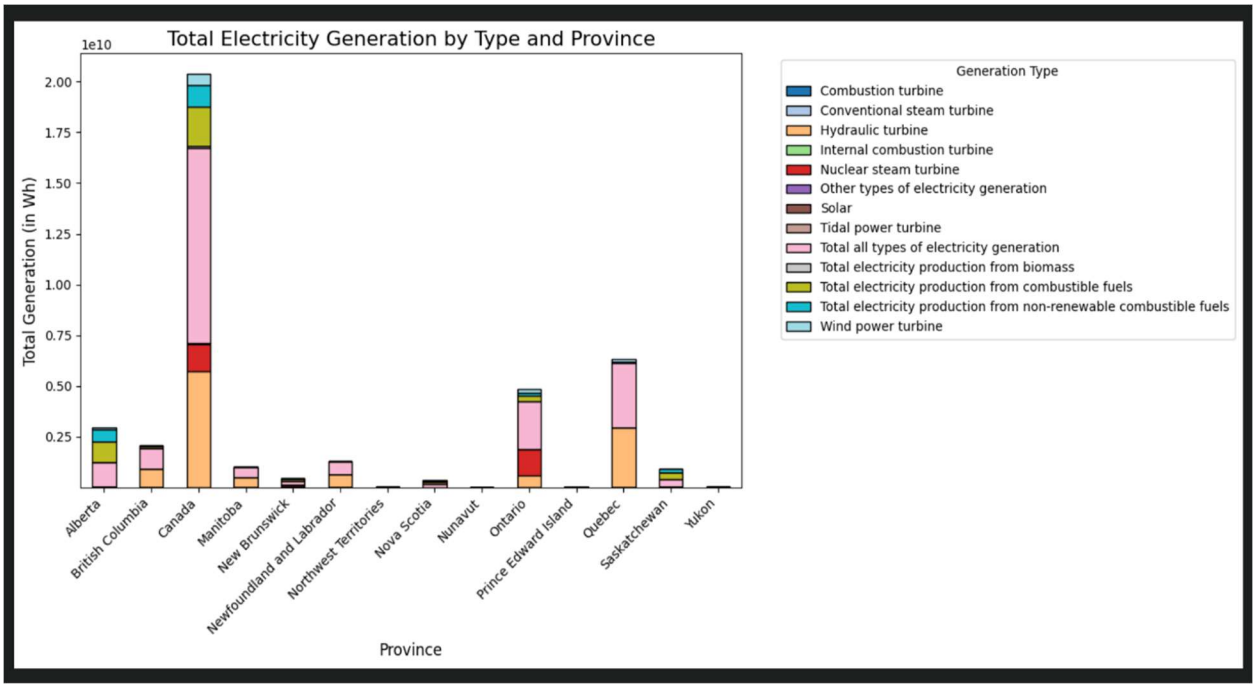
The SQL query in figure 8 below aims to analyze the peak electricity generation by province, grouped by the type of electricity generation. Specifically, the query selects the type of electricity generation (Type_of_electricity_generation), the province (GEO), and the sum of the annual electricity values (SUM(Electricity_VALUE_Annual)) for each type and province. This aggregation is achieved using the GROUP BY clause, which organizes the data by generation type and province. The query also sorts the results in descending order of total generation (Total_Generation) within each generation type to identify the provinces with the highest contributions to each electricity type. This sorting provides an intuitive understanding of which provinces dominate in specific types of electricity generation.

Figure 8: SQL Query on Analysis of Provincial Electricity Generation by Type in Canada

```
query_4 = '''
SELECT
    Type_of_electricity_generation AS Generation_Type,
    GEO AS Province,
    SUM(Electricity_VALUE_Annual) AS Total_Generation
FROM
    Electricity_Generated_Data_by_provience
GROUP BY
    Type_of_electricity_generation, GEO
ORDER BY
    Generation_Type, Total_Generation DESC;
'''
```

The visualization code transforms the query results into a stacked bar chart to provide a comparative and detailed view of electricity generation across provinces and types. The process begins by converting the query results into a pandas DataFrame (df4). This DataFrame is then pivoted using the pivot_table function, where the provinces are set as rows (index), the generation types as columns (columns), and the total generation values as the cell values. The aggfunc='sum' ensures that if there are multiple entries for a particular province and generation type, they are aggregated correctly.

Figure 9: Stacked Bar Chart of Total Electricity Generation by Type and Province in Canada



The visualization in figure 9 clearly represents the total electricity generation by type across Canadian provinces. Notable patterns emerge, such as Quebec's dominance in total electricity production, driven by renewable sources like hydropower. Similarly, provinces like Ontario showcase a substantial contribution from nuclear energy, alongside other generation types. Alberta and Saskatchewan, on the other hand, display significant reliance on non-renewable combustible fuels, highlighting their dependency on fossil fuels. British Columbia's generation profile is also dominated by renewable sources, including hydropower and biomass.

The stacked nature of the chart enables a quick comparison of total generation within each province and the relative contributions of different electricity generation types. This provides valuable insights into the energy mix of each province, aiding in understanding regional strengths and dependencies in the electricity generation landscape. The national total generation bar further contextualizes provincial data, offering a macro view of Canada's energy profile.

Query 5: Identifying Outliers in Provincial Electricity Generation by Year

The objective of this query in figure 10 is to identify outlier provinces in terms of their electricity generation over multiple years. Outliers are determined by comparing each province's annual electricity generation to the average generation across all provinces in the same year. Provinces are classified as high outliers if their generation exceeds 1.5 times the average and as low outliers if their generation falls below 0.5 times the average. To achieve this, the query first calculates the total annual electricity generation for each province by grouping the data based on GEO (province) and Year.

Figure 10: SQL query on Identifying Outliers in Provincial Electricity Generation by Year

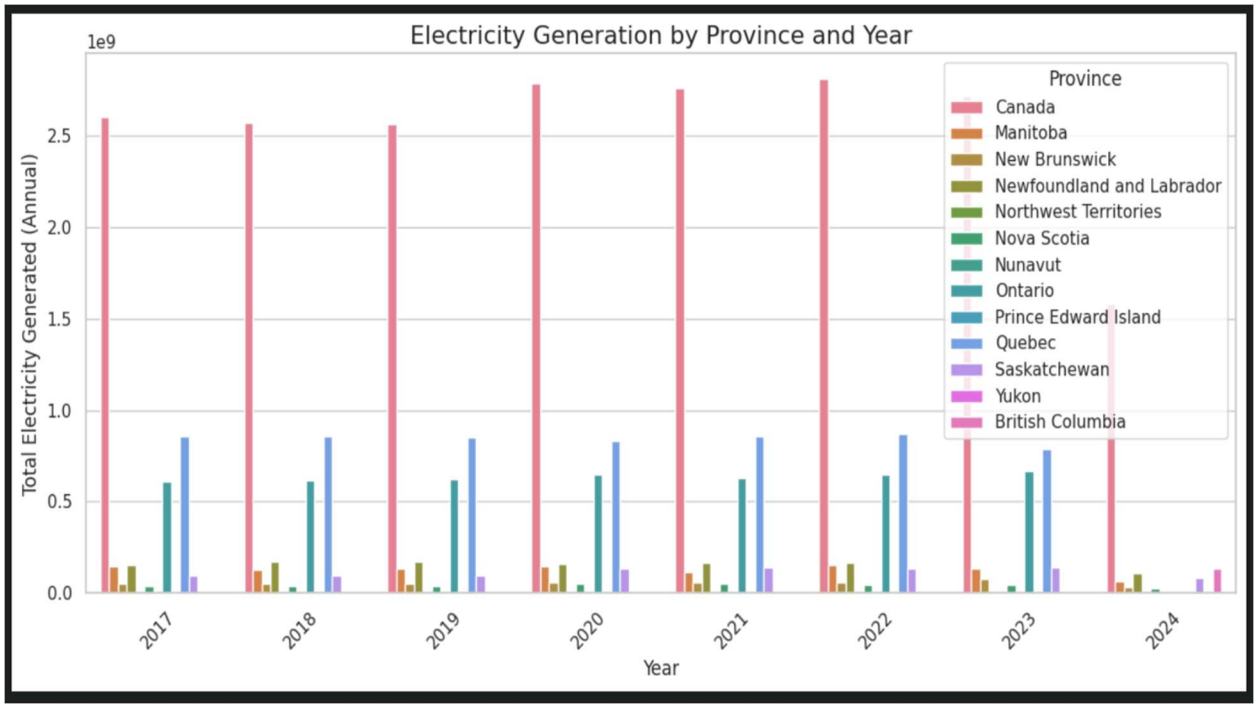
```
query_5 = '''
SELECT
    GEO AS Province,
    Year,
    SUM(Electricity_VALUE_Annual) AS Total_Electricity_Generated
FROM
    Electricity_Generated_Data_by_provience
GROUP BY
    GEO, Year
HAVING
    Total_Electricity_Generated > (
        SELECT AVG(Total_Electricity_Generated) * 1.5
        FROM (
            SELECT SUM(Electricity_VALUE_Annual) AS Total_Electricity_Generated
            FROM Electricity_Generated_Data_by_provience
            GROUP BY GEO, Year
        ) AS Subquery
    )
    OR
    Total_Electricity_Generated < (
        SELECT AVG(Total_Electricity_Generated) * 0.5
        FROM (
            SELECT SUM(Electricity_VALUE_Annual) AS Total_Electricity_Generated
            FROM Electricity_Generated_Data_by_provience
            GROUP BY GEO, Year
        ) AS Subquery
    )
ORDER BY
    Province, Year;

'''
```

The HAVING clause is then used to filter provinces based on the thresholds for high and low outliers. These thresholds are computed using subqueries that derive the average total generation for all provinces across each year. Finally, the results are ordered by province and year, enabling a clear chronological and geographic comparison.

The visualization of the query results was designed to highlight trends and anomalies in electricity generation over time. The query output was converted into a pandas DataFrame, which was subsequently used to create a time-series plot. The plot displays each province's total annual electricity generation, with distinct colors representing different provinces for clarity. The figure was scaled appropriately to ensure that even smaller provinces with lower generation levels are visible. A legend was included to help identify the provinces, and the x-axis, which represents the years, was labeled to align with the analyzed timeline. This approach ensures that trends, variations, and outliers in electricity generation are visually emphasized.

The visualization in figure 11 reveals notable insights into electricity generation trends and outliers across provinces from 2017 to 2024. Quebec consistently stands out as a high outlier, with its electricity generation significantly surpassing that of other provinces every year, reflecting its reliance on hydropower and large-scale production capacity. On the other hand, provinces like Prince Edward Island, Yukon, and Nunavut are consistently identified as low outliers due to their smaller populations and limited generation capacity. The trends also reveal stability in electricity generation for larger provinces like Ontario and Alberta, while smaller provinces exhibit low but steady generation levels. This analysis highlights disparities in energy production across Canada, emphasizing regional differences influenced by factors such as resource availability, population size, and energy policies.

Figure 11: Time-Series Plot of Provincial Electricity Generation Trends and Outliers (2017–2024)



Dataset 2: Electricity Car Dataset

From the Exploratory Data Analysis (EDA), I worked on some queries,

Figure 12: Annual Trends in Electric Vehicle Adoption

```
# Query 1: Annual Trends in Electric Vehicle Adoption
query1 = """
SELECT REF_DATE, SUM(VALUE) AS Total_Battery_Electric
FROM data
WHERE `Fuel type` = 'Battery electric'
GROUP BY REF_DATE
ORDER BY REF_DATE;
"""
```

This SQL code uses Pandas and pandasql libraries to analyze trends in electric vehicle adoption. pandas is imported to handle data as a DataFrame. pandasql is imported as sqldf, enabling the use of SQL queries directly on Pandas Data Frames. The dataset 'electricity_car_UPDATED.csv' is loaded into a Pandas DataFrame named data. It contains data on vehicle adoption by 'fuel type', 'geography', and 'time'. 'pysqldf' is a lambda function that executes SQL queries on Pandas DataFrames using 'pandasql.sqldf'. The 'globals()' ensures that the SQL query can access all variables in the global scope, including the DataFrame data.

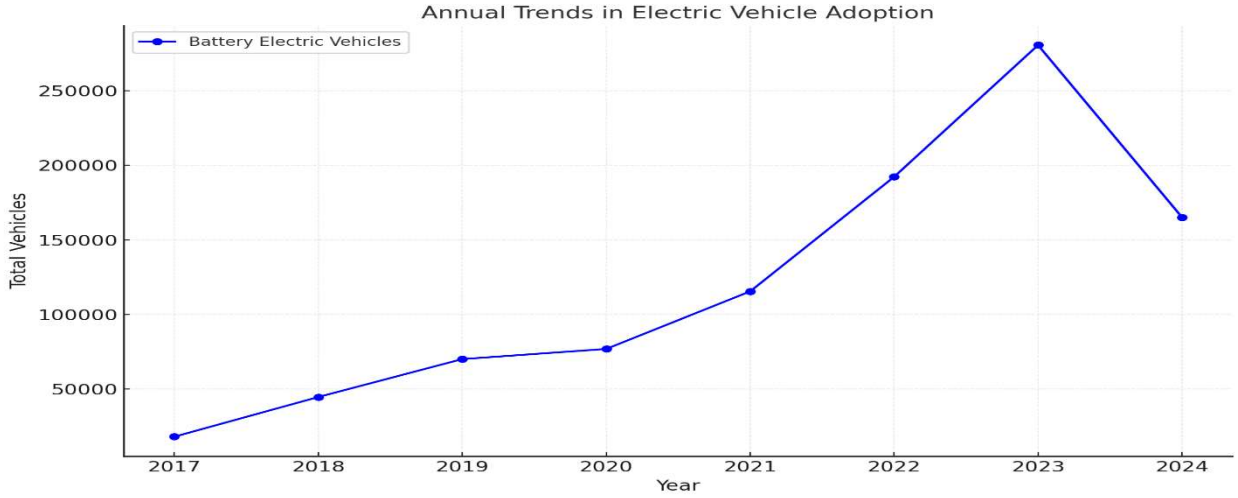
The query calculates annual trends in battery electric vehicle adoption are, SELECT: Selects 'REF_DATE' (year) and calculates the total number of battery electric vehicles (VALUE) as

'Total_Battery_Electric'. FROM data: Specifies the dataset to query. WHERE: Filters rows where the 'Fuel type' is Battery electric. GROUP BY: Groups the data by 'REF_DATE '(year) to aggregate values. ORDER BY REF_DATE: Orders the results chronologically by year.

pysqldf(query1) executes the SQL query on the DataFrame data.The result, stored in result1, is a new DataFrame showing the yearly total of battery electric vehicles. print(result1) outputs the result to the console.

Figure 13 below depicts the annual trends in electric vehicle adoption, the blue line indicates the battery electric vehicle. From the year 2017 to 2024 the total battery electric vehicles usage has increased. As we can see in the year 2017 the usage of battery electric vehicles is less than 50000, there's a slight increase in the year 2018 which is 50000. Likewise, the adoption of battery electric vehicles has been increased in the years 2019 and 2020 to 100000. But there is a rapid increase in the usage of electric vehicles by the people from the year 2021 to 2023 from 100000 to more than 250000. And decreased to 170000 in the year 2024.

Figure 13: Annual Trends in Electric Vehicle Adoption



The reason why there is less percentage in the adoption for the battery electric vehicles in the years of 2017 and 2018 is, people might not have knowledge of electric vehicles at that time. Because electric vehicles have just started to emerge. As we see the trend increased in the year 2019 and 2020 because people got some knowledge of electric vehicles also it was the pandemic (Covid-19) period, people got some knowledge about those vehicles. And the years passed, people got more concerned about the environment and reduced the usage of gasoline vehicles, as an alternative people adopted electric vehicles. So, this might be the reason why there is an increase in trends from 2021 to 2023.

Figure 14: Comparison of Fuel Types in the Latest Year

```
query = """
SELECT `Fuel type`, VALUE
FROM dataset
WHERE REF_DATE = (SELECT MAX(REF_DATE) FROM dataset);
"""
```

This code demonstrates the use of SQLite to query a dataset directly in Python. It utilizes an in-memory database to perform operations without creating a persistent database file, making it efficient for temporary computations. The dataset 'electricity_car_UPDATED.csv' is first loaded into a Pandas DataFrame. Then, a connection to an SQLite database is established using 'sqlite3.connect(': memory), which sets up a database in the system's memory. The dataset is stored as a table named dataset in the SQLite database using to_sql, allowing SQL queries to interact with it.

The SQL query retrieves the Fuel type and its corresponding 'VALUE' for the most recent year (REF_DATE) in the dataset. This is achieved by Sub selecting the maximum value of 'REF_DATE' using 'SELECT MAX(REF_DATE)'. Filtering the data to only include rows matching this most recent date. The query's result is fetched as a Pandas DataFrame using 'pd.read_sql_query' and printed to display the values of different fuel types for the most recent year available in the dataset. This approach efficiently combines SQLite's robust querying capabilities with Pandas' data-handling flexibility.

Figure 15: Comparison of Fuel Types in 2024

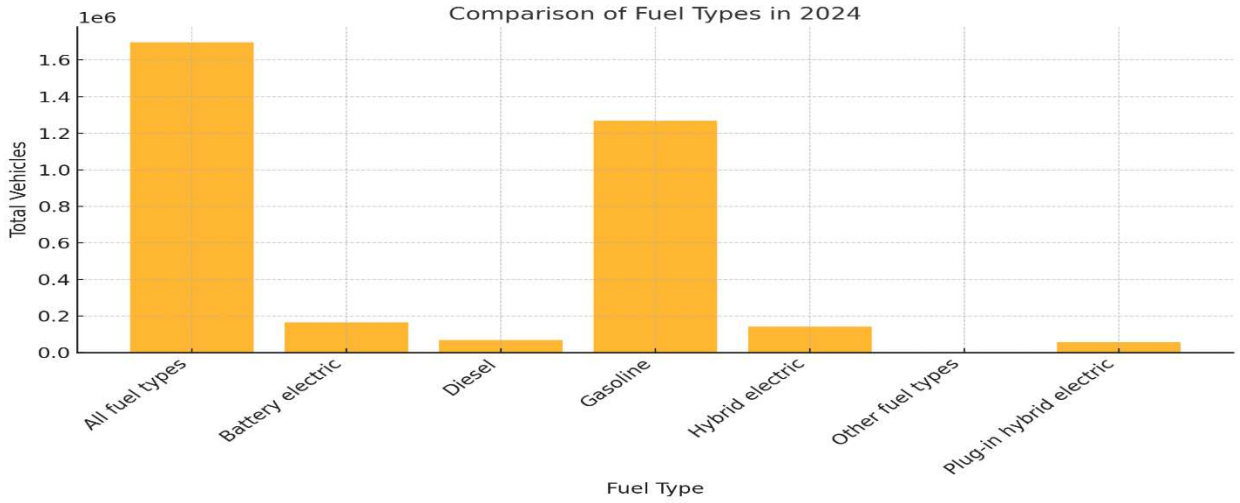
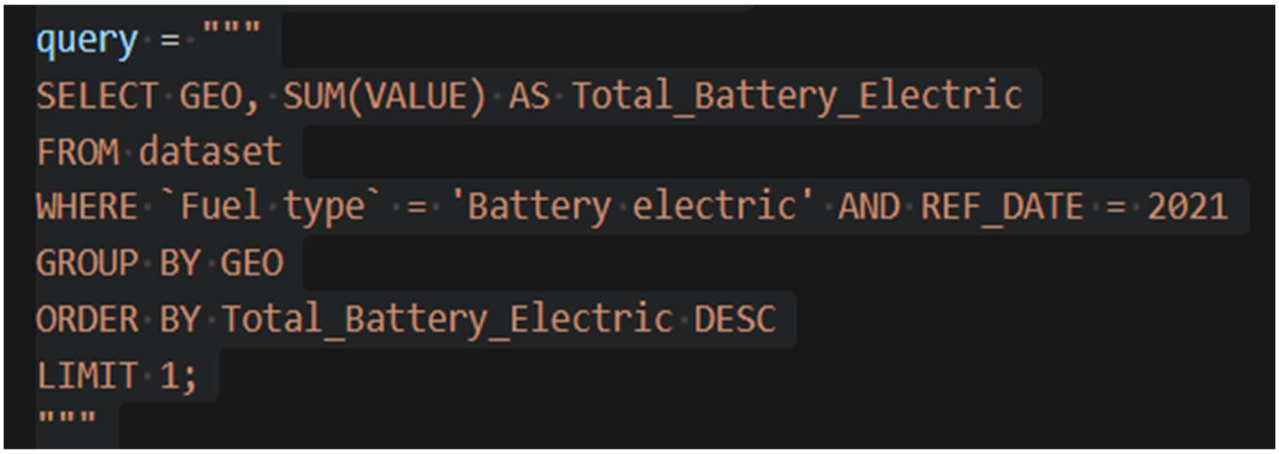


Figure 15 indicates the comparison of fuel types in 2024, as we can see the x-axis indicates the 'Fuel Type', y-axis indicates the 'Total Vehicles'. Under the 'Fuel Type' section it contains, All

fuel types, Battery electric, Diesel, Gasoline, Hybrid electric, other fuel types, Plug-in hybrid electric. ‘Total Vehicles’ contains the values from 0.0 to 1.6. When compared to other fuel types, ‘All fuel types’ has the highest value of more than 1.6. It contains all types of vehicles such as electric, gasoline, diesel, hybrid etc. So, it has the highest value of more than 1.6.

When we investigate the ‘Battery electric’, it has the lower value of 0.1 as the usage of electric vehicles has decreased in the year of 2024 compared to previous years. Compared to the other fuel types ‘Diesel’ has the lowest value of 0.0. Due to the increase in the price and less availability of diesel, might be the reason for the smaller number of ‘Diesel’ fuel type vehicles in the year of 2024. Out of all the fuel types, ‘Gasoline’ has the highest value that means more population are using ‘Gasoline’ vehicles rather than ‘Battery electric’ and ‘Diesel’ vehicles. Whereas the ‘Hybrid electric’ vehicles have the value of 0.1, these hybrid electric vehicles have the ability to work with either electricity as well as gasoline. ‘Other fuel types’ of vehicles are being used by the people in the year 2024. And lastly, ‘Plug-in hybrid electric’ vehicles have the value of 0.0, it means that people in 2024 are not using this type of vehicle.

Figure 16: Regional Insights



```
query = """
SELECT GEO, SUM(VALUE) AS Total_Battery_Electric
FROM dataset
WHERE `Fuel type` = 'Battery electric' AND REF_DATE = 2021
GROUP BY GEO
ORDER BY Total_Battery_Electric DESC
LIMIT 1;
"""
```

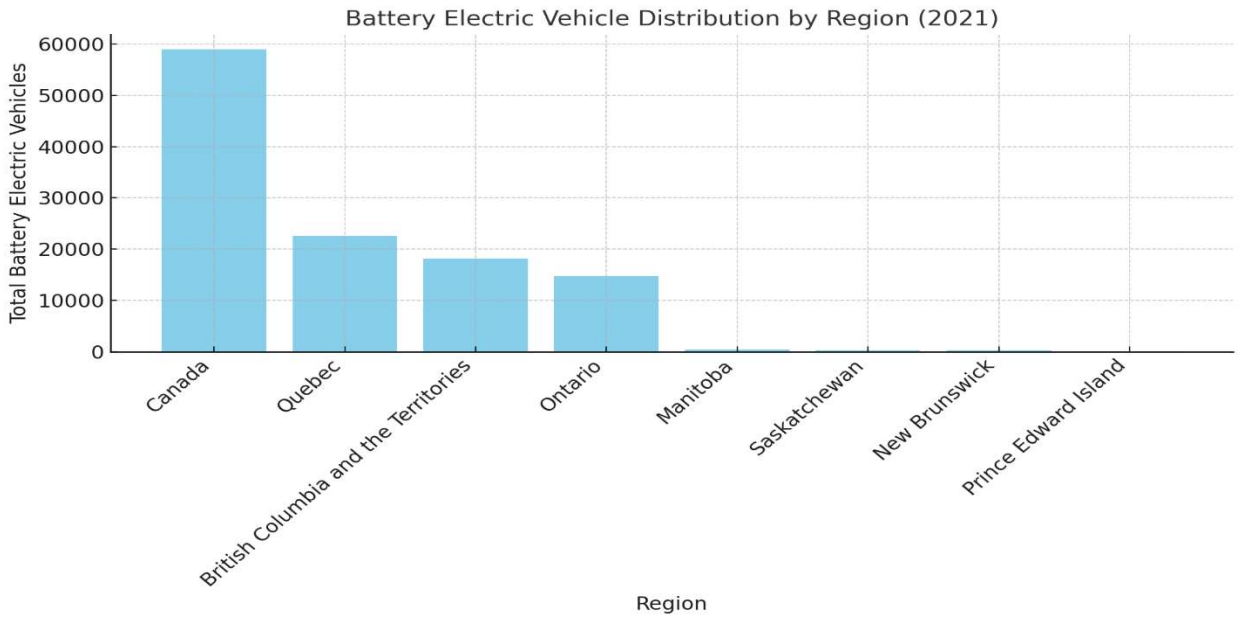
This Python script analyzes a dataset containing information about electric vehicles using pandas and SQLite. It begins by loading a CSV file ‘electricity_car_UPDATED.csv’ into a panda DataFrame. The data is then transferred to an in-memory SQLite database for efficient querying. A SQL query is executed to identify the geographical region ‘GEO’ with the highest total number of battery electric vehicles ‘Fuel type’ = 'Battery electric' in the year 2021 ‘REF_DATE = 2021’. The query calculates the total number of such vehicles per region, sorts the results in descending order, and retrieves the top region. The dataset, stored in a CSV file named ‘electricity_car_UPDATED.csv’, is read into a panda DataFrame. This dataset is expected to include details such as geographical regions ‘GEO’, fuel types ‘Fuel type’, reference years ‘REF_DATE’, and values ‘VALUE’, which likely represent the number of vehicles.

An SQLite in-memory database (: memory:) was created to allow SQL queries on the dataset. The entire DataFrame is loaded into the database as a table named dataset. This setup is particularly useful for performing complex queries on large data sets without permanent storage. A SQL query is defined and executed to: Filter rows where the Fuel type is 'Battery electric'. Focus on records from the year 2021 ‘REF_DATE = 2021’. Group the data by ‘GEO’ (geographical region) and

compute the total number of battery electric vehicles '(SUM(VALUE))' for each region. Sort the results in descending order by the total and retrieve only the top entry '(LIMIT 1)'.

The result of the query, a panda DataFrame, contains two columns: the geographical region 'GEO' with the highest adoption of battery electric vehicles and its corresponding total 'Total_Battery_Electric'. This output provides valuable insights into regional trends in electric vehicle adoption for policymakers, researchers, or businesses. The output, showing the region and its corresponding total, is displayed as a panda DataFrame. This approach effectively combines data manipulation with SQL-based analysis to provide insights into electric vehicle adoption by geography.

Figure 17: Battery Electric Vehicle Distribution by Region (2021)



The bar chart displays the distribution of Battery Electric Vehicles (BEVs) by region in Canada for 2021. It highlights that Canada as a whole has the highest total BEVs, followed by Quebec and British Columbia and the Territories. Ontario also shows significant BEV numbers, while regions such as Manitoba, Saskatchewan, New Brunswick, and Prince Edward Island have comparatively lower BEV distributions. The data indicates a concentration of BEVs in larger provinces, likely reflecting population density and infrastructure availability.

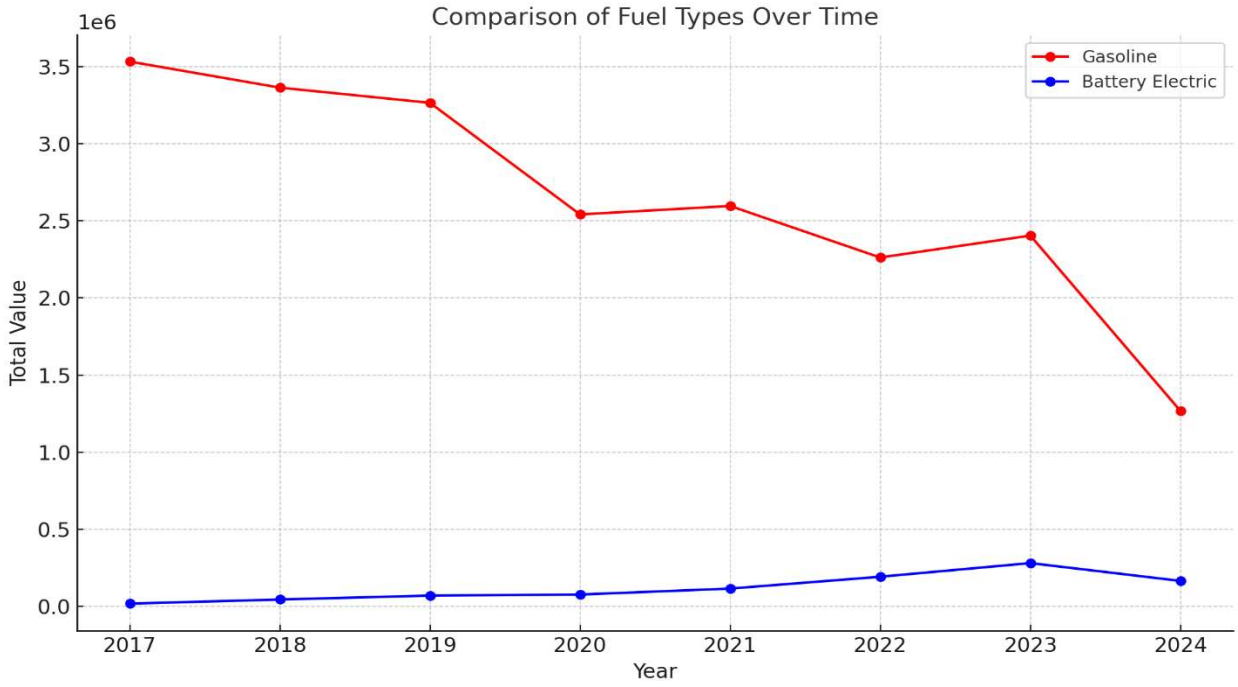
Figure 18: Shift from Gasoline to Electric

```
query = """
SELECT REF_DATE, `Fuel type`, SUM(VALUE) AS Total_Value
FROM dataset
WHERE `Fuel type` IN ('Gasoline', 'Battery electric')
GROUP BY REF_DATE, `Fuel type`
ORDER BY REF_DATE, `Fuel type`;
"""
```

This Python script is designed to analyze trends in vehicle fuel types over time using a dataset of vehicle data stored in a CSV file 'electricity_car_UPDATED.csv'. It begins by loading the dataset into a pandas DataFrame, which is then transferred into an in-memory SQLite database for efficient querying. The script focuses on comparing the usage of two specific fuel types: 'Gasoline' and 'Battery electric'. By using SQL, it calculates the total number of vehicles for each fuel type '(SUM(VALUE))' grouped by year '(REF_DATE)' and 'fuel type', providing insights into trends across different time periods. The results are ordered by year and fuel type for easy interpretation.

The final output is a pandas DataFrame containing the year, fuel type, and total number of vehicles for the selected categories. This allows for a clear comparison of the adoption patterns of gasoline and battery electric vehicles. The use of SQL enhances the script's ability to handle structured data, making it a versatile tool for examining the evolution of fuel type preferences. This analysis could be particularly useful for tracking the transition from traditional gasoline-powered vehicles to electric ones, offering valuable insights to researchers, policymakers, and industry stakeholders.

Figure 19: Comparison of Fuel Types Over Time



The line graph compares the total values of gasoline and battery electric vehicles (BEVs) from 2017 to 2024. Gasoline vehicles show a consistent decline over the years, dropping significantly between 2023 and 2024. In contrast, BEVs display a gradual increase, indicating a steady rise in adoption. While gasoline vehicles still dominate in total value, the growing trend for BEVs highlights a shift towards more sustainable fuel types over time.

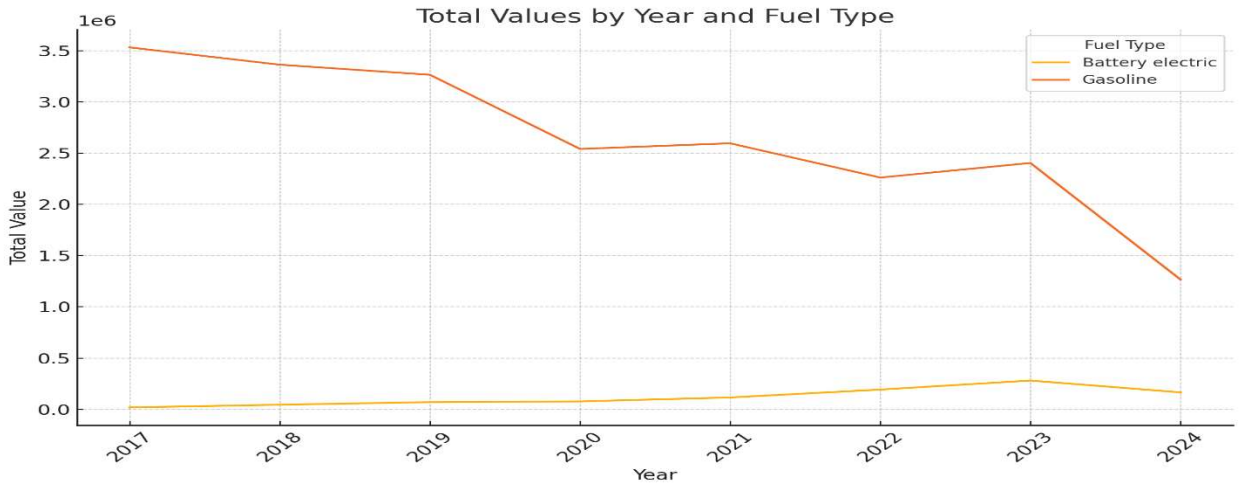
Figure 20: Market Share of Electric Vehicles

```
query = """
SELECT REF_DATE, `Fuel type`, SUM(VALUE) AS Total_Value
FROM data
WHERE `Fuel type` IN ('Gasoline', 'Battery electric')
GROUP BY REF_DATE, `Fuel type`
ORDER BY REF_DATE, `Fuel type`;
"""
```

This Python script demonstrates the use of the pandasql library to analyze trends in vehicle fuel types from a dataset stored in a CSV file 'electricity_car_UPDATED.csv'. The dataset is loaded into a pandas DataFrame, and the pandasql library allows SQL queries to be executed directly on the DataFrame, enabling users to leverage the power of SQL for structured data analysis. The script focuses on comparing two fuel types: "Gasoline" and "Battery electric," and calculates the total number of vehicles (SUM(VALUE)) for each fuel type grouped by year 'REF_DATE'. The results are ordered by year and fuel type to present a clear, time-ordered view of the data.

The output is a DataFrame containing three columns: year, fuel type, and the total number of vehicles for each fuel type in each year. By combining Python's pandas library with SQL queries, the script simplifies the process of analyzing structured datasets, making it accessible to users familiar with SQL. This approach is particularly useful for visualizing trends in fuel usage over time and identifying the growth or decline of different vehicle technologies. Policymakers, automotive analysts, and researchers can use this data to track the transition to electric vehicles and evaluate the impact of policies promoting cleaner transportation.

Figure 21: Total Values by Year and Fuel Type

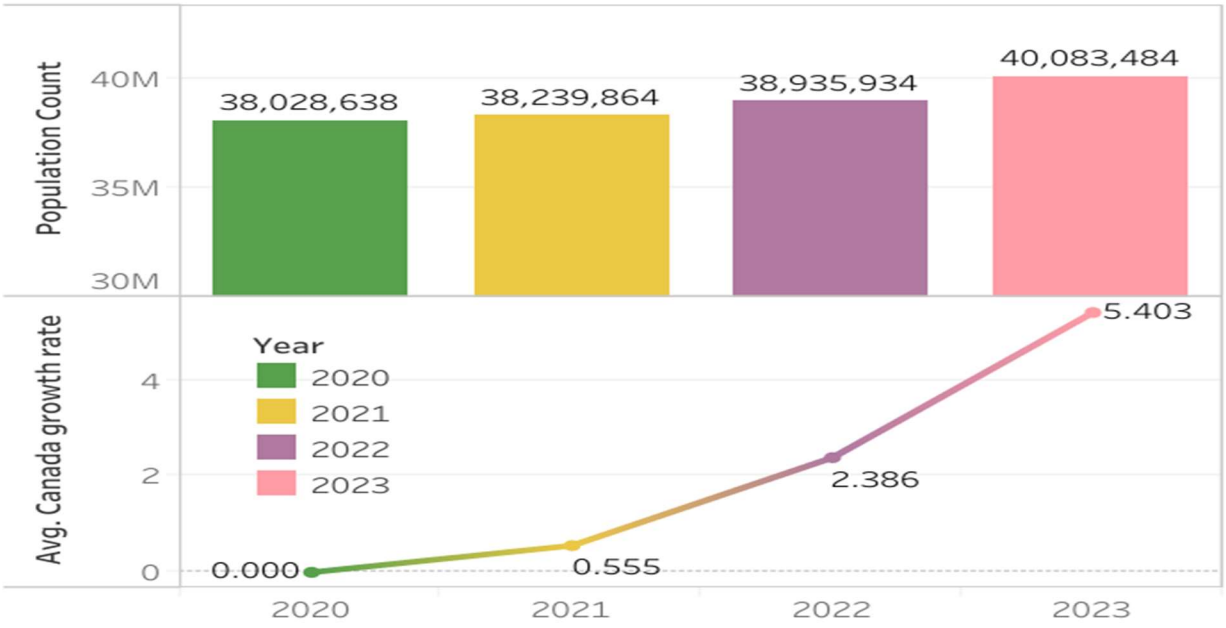


The chart shows the total values over time (from 2017 to 2024) categorized by fuel type: "Battery Electric" and "Gasoline." The orange line, representing gasoline, indicates a significant decline in total value from over 3.5 million in 2017 to below 1.5 million in 2024. In contrast, the yellow line for battery electric vehicles shows a relatively low but steady increase in total value over the years, although its scale remains much smaller compared to gasoline. This trend highlights a possible gradual shift toward battery electric vehicles, while gasoline vehicles face a substantial decrease in total value.

Dataset 3: Population Dataset

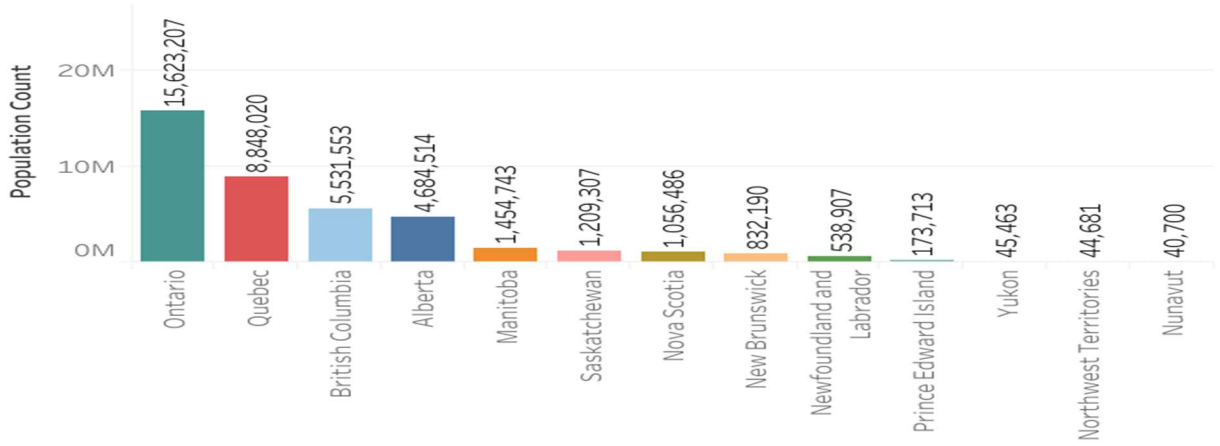
The exploratory data analysis focuses on understanding Canada's population trends from 2020 to 2023 and the distribution of population among its provinces in 2023. Two visualizations were created to capture these insights effectively.

Figure 22: Population and population growth of Canada from 2020 to 2023.



The first visualization from figure examines Canada's total population and annual growth rates from 2020 to 2023. The bar chart indicates a steady rise in the population, starting from 38,028,638 in 2020 and reaching 40,003,404 by 2023. This growth is further contextualized by the line graph, which highlights the average annual population growth rates. The analysis shows no growth in 2020 (0.000%), likely influenced by the pandemic's effects, followed by a moderate increase in 2021 (0.555%) and 2022 (2.386%). A significant surge is observed in 2023, with a growth rate of 5.403%. This reflects a sharp recovery and possibly the impact of increased immigration or natural growth in the post-pandemic period. The combined use of population figures and growth rates provides a clear understanding of national trends.

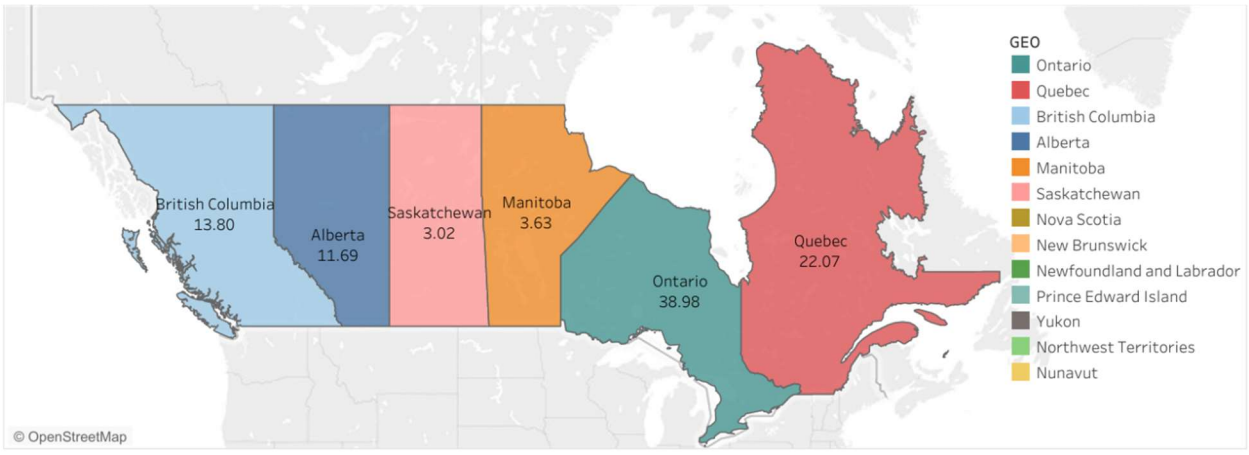
Figure 23: Population of provinces in Canada in 2023.



The second visualization in figure focuses on the population distribution among Canada's provinces in 2023. Ontario has the largest population, with 15,623,207 residents, followed by Quebec with 8,948,020 and British Columbia with 5,531,553. Alberta also has a substantial population of 4,684,514. In contrast, provinces and territories like Yukon (45,463), Northwest Territories (44,681), and Nunavut (40,700) have the smallest populations, reflecting the sparse settlement in these northern regions. This significant disparity in population distribution highlights regional differences, which could stem from variations in economic opportunities, accessibility, and living conditions.

This analysis provides a comprehensive view of Canada's demographic changes and the unequal population distribution across regions. These insights can inform decision-making in areas such as urban development, resource planning, and regional policy initiatives.

Figure 24: Population percentage in the top 6 provinces in Canada in 2023.



Next, the visualization in figure we created focuses on the population distribution across the top six provinces in Canada for 2023. The map illustrates the percentage share of each province's population within the overall Canadian population. The provinces are color-coded for clarity, with specific percentages displayed directly on the map to facilitate easy interpretation.

Ontario emerges as the most populous province, accounting for 38.98% of the population. It is followed by Quebec with 22.07%, highlighting the concentration of population in central Canada.

British Columbia and Alberta, with 13.80% and 11.69% respectively, represent significant population hubs in western Canada. Manitoba and Saskatchewan, though smaller contributors, still feature prominently with 3.63% and 3.02% respectively. Together, these provinces comprise the majority of Canada's population, underscoring their importance in demographic studies and policy planning.

This visualization provides an intuitive spatial representation of population concentration and regional disparities in Canada. By integrating demographic percentages with geographic data, this map highlights population distribution patterns that are critical for urban planning, resource allocation, and economic strategies at both provincial and federal levels. Moreover, the color scheme enhances comprehension, making it easier for stakeholders to identify population density variations across the provinces.

To gain meaningful insights from the population dataset, I performed several SQL queries, each targeting specific aspects of the data. These queries are designed to analyze population distribution, growth rates over different time periods, and the impact of external factors such as the COVID-19 pandemic. In this section, I will explain each query individually, outlining its purpose, methodology, and insights gained. The detailed breakdown of each query will help demonstrate the steps taken to uncover trends and patterns in the dataset.

Query 1: Identifying Regions with the Largest Populations as of July 1, 2024

This query identifies the regions in Canada with the largest populations as of July 1, 2024. By grouping the data by region (GEO) and selecting the maximum population value (MAX(VALUE)), it provides a snapshot of how the population is distributed across the country.

Figure 25: SQL query on Identifying Regions with the Largest Populations as of July 1, 2024

```
Query = '''
SELECT GEO, MAX(VALUE) AS Max_population
FROM Population
WHERE REF_DATE='2024-07-01'
GROUP BY GEO
ORDER BY Max_population DESC;
'''
```

Ordering the results in descending order ensures that the regions with the highest populations are listed first, making it easy to analyze the demographic structure. This information is critical for understanding population hotspots and serves as a baseline for further demographic or resource allocation studies.

Figure 26: SQL query for Analyzing Post-Pandemic Recovery Growth Rates (2020-2023)

```
Query = '''
SELECT GEO, (MAX(VALUE) - MIN(VALUE)) / MIN(VALUE) * 100 AS Growth_rate
FROM Population
WHERE GEO IN ('Alberta', 'British Columbia', 'Manitoba', 'Saskatchewan', 'Quebec', 'Ontario','Canada')
AND REF_DATE BETWEEN '2020-07-01' AND '2023-07-01'
GROUP BY GEO ORDER BY Growth_rate DESC;
'''
```

This query calculates the population growth rate for selected regions between July 2020 and July 2023. By considering both the maximum and minimum population values over this period and calculating the percentage growth rate, the query highlights regions that experienced the fastest or slowest recovery during the post-pandemic period. Ordering by growth rate in descending order ensures that the most rapidly growing regions are listed first. The analysis is valuable for understanding how demographic trends shifted after COVID-19, which impacted factors like migration, urbanization, and birth rates.

Figure 27: SQL query for Capturing Growth Trends in the Post-Pandemic Stabilization Period (2022-2023)

```
Query = '''
SELECT GEO, (MAX(VALUE) - MIN(VALUE)) / MIN(VALUE) * 100 AS Growth_rate
FROM Population
WHERE GEO IN ('Alberta', 'British Columbia', 'Manitoba', 'Saskatchewan', 'Quebec', 'Ontario')
AND REF_DATE >'2022-05-31'
GROUP BY GEO ORDER BY Growth_rate DESC;
'''
```

This query focuses on population growth from mid-2022 onwards, capturing demographic changes during the post-pandemic recovery period. It calculates the growth rate for selected provinces by comparing the maximum and minimum population values from June 2022 onward. By limiting the date range to this period, the query provides insights into the recovery momentum of each region after the initial impacts of the pandemic had subsided. Understanding these growth patterns helps highlight the demographic resilience and economic recovery of different provinces.

Figure 28: SQL query for Assessing COVID-19 Impact on Population Growth (2020 to Mid-2022)

```
Query = '''
SELECT GEO, (MAX(VALUE) - MIN(VALUE)) / MIN(VALUE) * 100 AS Growth_rate
FROM Population
WHERE GEO IN ('Alberta', 'British Columbia', 'Manitoba', 'Saskatchewan', 'Quebec', 'Ontario')
AND REF_DATE BETWEEN '2020-01-01' AND '2022-05-31'
GROUP BY GEO ORDER BY Growth_rate DESC;
'''
```

This query examines the impact of the COVID-19 pandemic on population growth rates between January 2020 and May 2022. It calculates the percentage growth rate for selected regions during this period, revealing how the pandemic disrupted population trends through reduced immigration, urban-to-rural migration, and other factors. Analyzing this timeframe helps contextualize demographic shifts caused by the crisis, showing which regions experienced stagnation or slowdowns during the pandemic and offering a basis for comparison with the post-pandemic recovery phase.

Figure 29: SQL query for Cleaning and Streamlining the Dataset

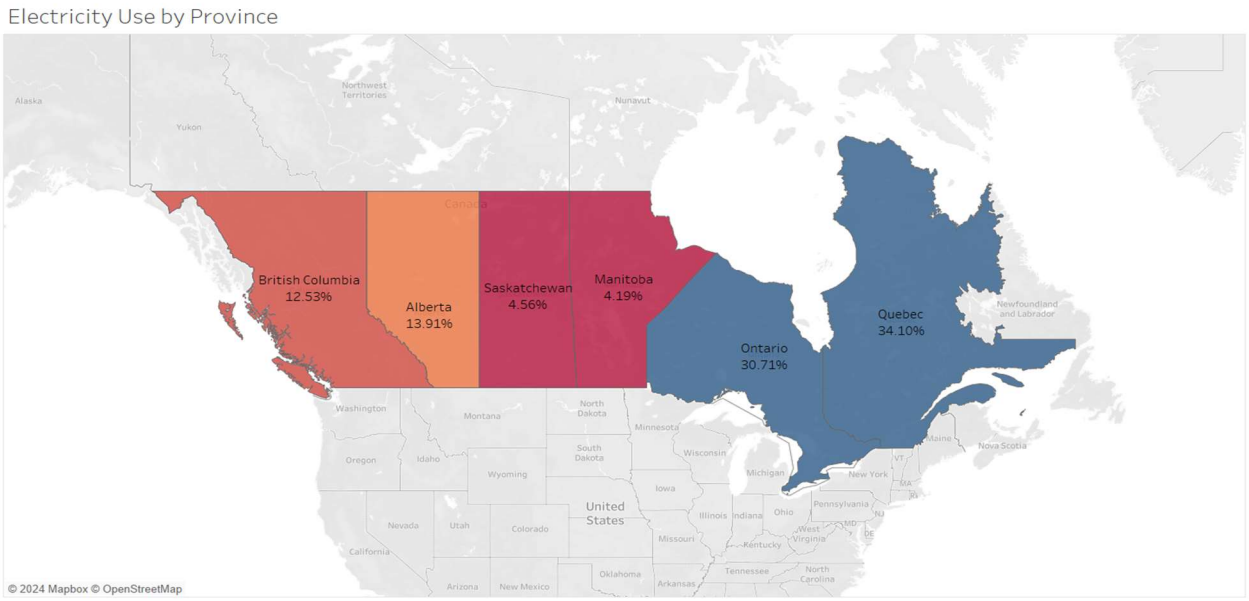
```
Query = '''
SELECT REF_DATE, GEO, VALUE
FROM Population
WHERE VALUE IS NOT NULL;
'''
```

This query extracts only the key columns (REF_DATE, GEO, and VALUE) while filtering out any rows where population values are null. By ensuring that the data contains no missing values, this step creates a clean and reliable dataset for subsequent analysis. This is essential for reducing errors, improving accuracy, and simplifying further exploration of population trends. Clean data forms the foundation of any robust analysis and ensures that insights drawn from the dataset are valid and actionable.

These queries together provide a comprehensive view of population trends, the impact of COVID-19, and post-pandemic recovery dynamics, enabling meaningful analysis of Canada's demographic evolution.

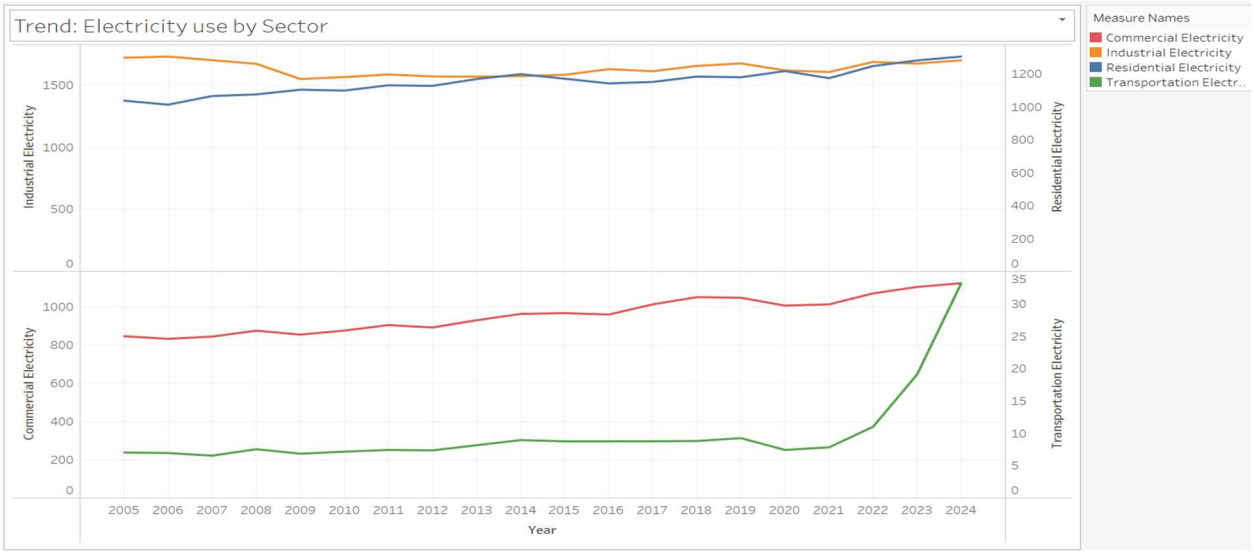
Dataset 4: Electricity Consumption Dataset

Figure 30: Electricity consumption percentage in the top 6 provinces in Canada



The map shows electricity use by Canadian provinces as a percentage of the total. Quebec leads with 34.10%, followed by Ontario at 30.71%, while Alberta and British Columbia account for 13.91% and 12.53%, respectively. Smaller shares come from Saskatchewan (4.56%) and Manitoba (4.19%). The distribution reflects population density and industrial activity, with Quebec and Ontario consuming the most.

Figure 31: Trends in electricity consumption for each sector

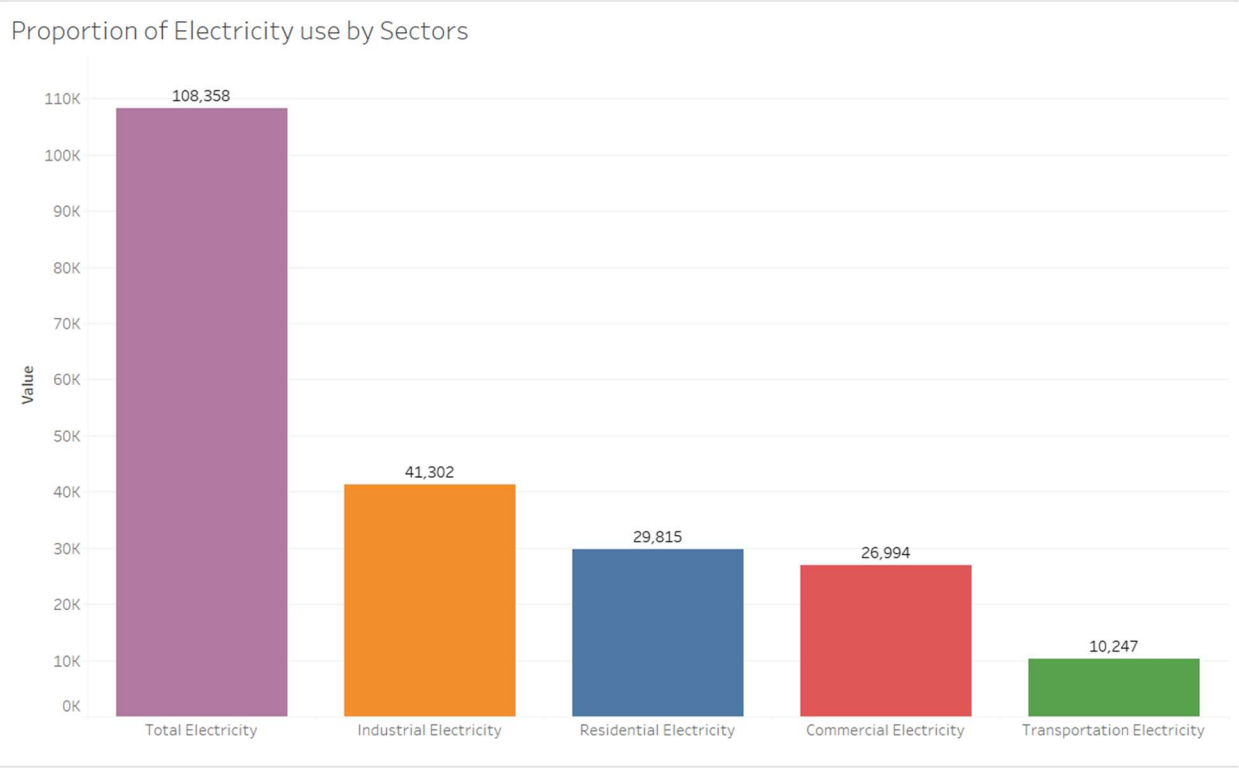


The chart illustrates trends in electricity consumption across four sectors: industrial, commercial, residential, and transportation, from 2005 to 2024. The industrial and commercial sectors exhibit the highest levels of electricity consumption, with industrial usage consistently leading. While both have shown gradual increases, their growth has been steady and linear over the years. Residential electricity consumption follows a similar, albeit slightly less pronounced, trend of steady growth.

In contrast, transportation electricity consumption remained relatively low and stable until a significant surge beginning around 2020. This sharp rise likely reflects the increasing adoption of

electric vehicles and advancements in transportation electrification. Overall, the chart highlights the growing electricity demands of all sectors, with notable acceleration in the transportation sector in recent years.

Figure 32: Sector and their electricity consumption



This bar chart compares the electricity consumption across different sectors, highlighting their contributions to the total electricity usage. The total electricity usage is the highest at 108,358 units, serving as an aggregate of all sectors. Among the individual sectors, industrial electricity consumption leads at 41,302 units, followed by residential at 29,815 units and commercial at 26,994 units.

Transportation electricity consumption is the smallest contributor at 10,247 units, which aligns with its relatively recent growth trajectory. This distribution underscores the dominance of industrial usage in electricity demand, while transportation, despite its rising trend, still accounts for a smaller share overall.

Figure 33: SQL query for finding the sector that has highest electricity consumption and output

Identify the Top Contributing Sector to Total Electricity Usage by Province and Year

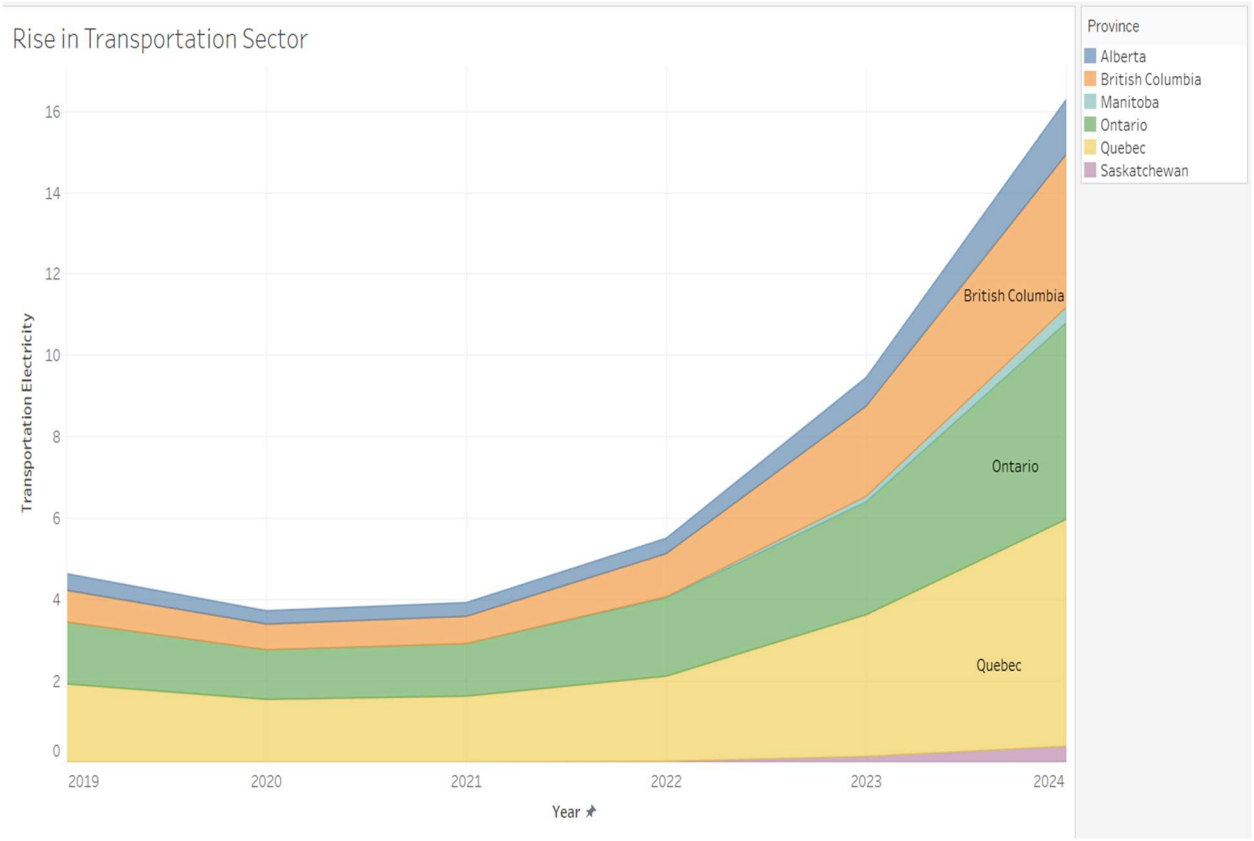
```
[ ] Top_Contributors = """
    SELECT
        Province,
        Year,
        CASE
            WHEN Residential_Electricity = MAX(Residential_Electricity, Commercial_Electricity, Industrial_Electricity, Transportation_Electricity)
            THEN 'Residential'
            WHEN Commercial_Electricity = MAX(Residential_Electricity, Commercial_Electricity, Industrial_Electricity, Transportation_Electricity)
            THEN 'Commercial'
            WHEN Industrial_Electricity = MAX(Residential_Electricity, Commercial_Electricity, Industrial_Electricity, Transportation_Electricity)
            THEN 'Industrial'
            ELSE 'Transportation'
        END AS Top_Sector,
        MAX(Residential_Electricity, Commercial_Electricity, Industrial_Electricity, Transportation_Electricity) AS Top_Contribution
    FROM
        ElectricityData;
    """
    q4 = pd.read_sql_query(Top_Contributors, connection)
    print(q4)
```

```
          Province  Year  Top_Sector  Top_Contribution
0          Alberta  2005  Industrial            137.60
1          Alberta  2006  Industrial            144.27
2          Alberta  2007  Industrial            142.37
3          Alberta  2008  Industrial            142.78
4          Alberta  2009  Industrial            142.24
..             ...   ...         ...               ...
317  Saskatchewan  2046  Industrial             58.30
318  Saskatchewan  2047  Industrial             60.13
319  Saskatchewan  2048  Industrial             62.15
320  Saskatchewan  2049  Industrial             64.38
321  Saskatchewan  2050  Industrial             66.81

[322 rows x 4 columns]
```

This query finds the sector—residential, commercial, industrial, or transportation—that contributes most to electricity consumption in each province for every year. It compares the consumption values for each sector in every row and, based on the highest value, assigns that sector as the "Top Sector" and its consumption as the "Top Contribution." all of this is moslty accomplised using CASE when then and MAX keywords This does the analysis of which sector is contributing most to the overall electricity usage within a province for each year.

The query results show that the **industrial sector** consistently accounts for the highest electricity consumption across Canadian provinces over time, measured in kilowatt-hours (kWh). For example, Alberta's industrial electricity usage was the top contributor from 2005 (137.60 kWh) to 2009, while in Saskatchewan, the industrial sector remains dominant, with consumption projected to rise from 58.30 kWh in 2046 to 66.81 kWh in 2050. These values reflect the energy-intensive nature of industrial activities and their significant role in driving electricity demand over the years and into the future.

Figure 34: Rise in Transportation Sector between the year 2019 to 2024



The chart illustrates the rise in electricity consumption within the transportation sector across various provinces from 2019 to 2024. Notably, British Columbia leads the increase, showing a sharp growth trajectory beginning in 2022, followed closely by Ontario and Quebec, which also experience significant increases. The combined contributions of Alberta, Manitoba, and Saskatchewan are relatively smaller but steadily rising, reflecting regional adoption trends.

This dramatic increase in transportation electricity usage, particularly post-2022, likely correlates with the accelerated adoption of electric vehicles and expanded infrastructure investments in provinces like British Columbia and Ontario. The data underscores the growing role of clean transportation initiatives in provincial energy consumption profiles.

Figure 35: SQL query for the fastest-growing electricity consumption sector for each province and output.

```
fastest_growing_sector = """
WITH Sector_Contribution AS (
    SELECT
        Province,
        Year,
        Residential_Electricity * 100.0 / Total_Electricity AS Residential_Percentage,
        Commercial_Electricity * 100.0 / Total_Electricity AS Commercial_Percentage,
        Industrial_Electricity * 100.0 / Total_Electricity AS Industrial_Percentage,
        Transportation_Electricity * 100.0 / Total_Electricity AS Transportation_Percentage
    FROM
        ElectricityData
),
Lagged_Contribution AS ( -- Calculate LAG values before aggregation
    SELECT
        Province,
        Year,
        Residential_Percentage,
        LAG(Residential_Percentage, 1, 0) OVER (PARTITION BY Province ORDER BY Year) AS Previous_Residential_Percentage,
        Commercial_Percentage,
        LAG(Commercial_Percentage, 1, 0) OVER (PARTITION BY Province ORDER BY Year) AS Previous_Commercial_Percentage,
        Industrial_Percentage,
        LAG(Industrial_Percentage, 1, 0) OVER (PARTITION BY Province ORDER BY Year) AS Previous_Industrial_Percentage,
        Transportation_Percentage,
        LAG(Transportation_Percentage, 1, 0) OVER (PARTITION BY Province ORDER BY Year) AS Previous_Transportation_Percentage
    FROM
        Sector_Contribution
),
```

```
Yearly_Growth AS (
    SELECT
        Province,
        AVG(Residential_Percentage - Previous_Residential_Percentage) AS Residential_Growth,
        AVG(Commercial_Percentage - Previous_Commercial_Percentage) AS Commercial_Growth,
        AVG(Industrial_Percentage - Previous_Industrial_Percentage) AS Industrial_Growth,
        AVG(Transportation_Percentage - Previous_Transportation_Percentage) AS Transportation_Growth
    FROM
        Lagged_Contribution
    WHERE
        Year > 2005 -- To avoid null values from LAG
    GROUP BY
        Province
)
SELECT
    Province,
    CASE
        WHEN Residential_Growth = MAX(Residential_Growth, Commercial_Growth, Industrial_Growth, Transportation_Growth) THEN 'Residential'
        WHEN Commercial_Growth = MAX(Residential_Growth, Commercial_Growth, Industrial_Growth, Transportation_Growth) THEN 'Commercial'
        WHEN Industrial_Growth = MAX(Residential_Growth, Commercial_Growth, Industrial_Growth, Transportation_Growth) THEN 'Industrial'
        ELSE 'Transportation'
    END AS Fastest_Growing_Sector,
    MAX(Residential_Growth, Commercial_Growth, Industrial_Growth, Transportation_Growth) AS Growth_Rate
FROM
    Yearly_Growth
GROUP BY
    Province;
"""
q5 = pd.read_sql_query( fastest_growing_sector , connection)
```

	Province	Fastest_Growing_Sector	Growth_Rate
0	Alberta	Transportation	0.479508
1	British Columbia	Transportation	0.502989
2	Canada	Transportation	0.464891
3	Manitoba	Transportation	0.615312
4	Ontario	Transportation	0.511811
5	Quebec	Transportation	0.328583
6	Saskatchewan	Transportation	0.652415

The following query calculates the fastest-growing electricity consumption sector for each province by analyzing year-over-year percentage changes in residential, commercial, industrial, and transportation electricity usage. It first computes the contribution of each sector to the total electricity usage (Sector_Contribution), then it calculates the values of the previous year using the lag function (Lagged_Contribution) and computes average yearly growth rates for each sector (Yearly_Growth). Lastly, the query identifies, for each province, the sector with the highest growth

rate and outputs the province, the fastest-growing sector, and its growth rate, thus helping in getting the consumption trends across various sectors

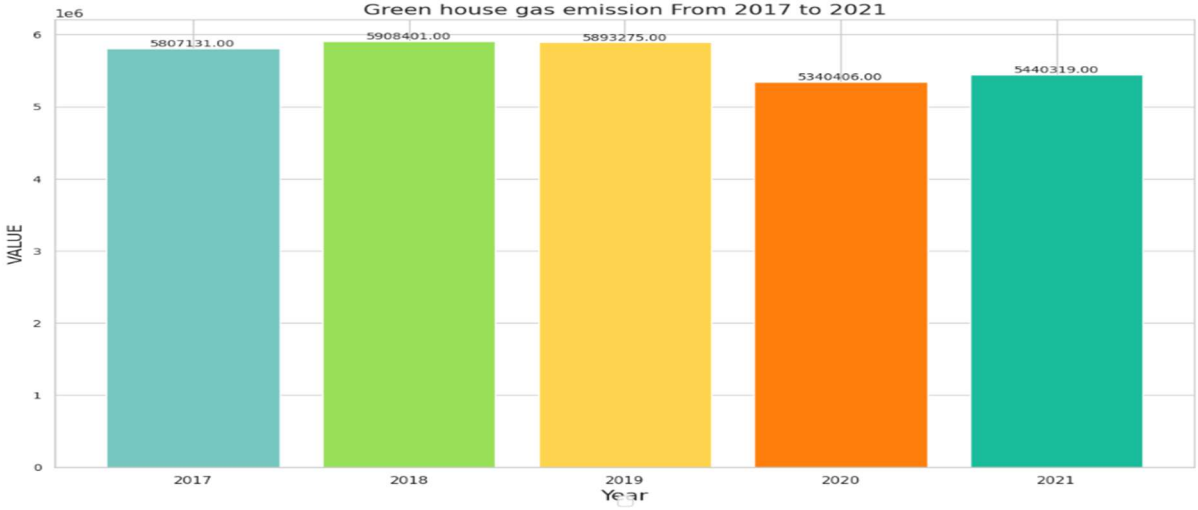
The results indicate that the transportation sector is the fastest-growing electricity-consuming sector across all Canadian provinces. Growth rates are particularly high in **Saskatchewan (0.652%)** and **Manitoba (0.615%)**, reflecting substantial advancements in transportation electrification in these regions. British Columbia and Ontario also show strong growth rates, at **0.503%** and **0.512%**, respectively, likely driven by widespread electric vehicle (EV) adoption and supportive infrastructure investments. Meanwhile, Quebec, despite having the lowest growth rate in transportation electricity use (**0.329%**), still leads in this sector, possibly due to its already significant base of EV usage.

This trend demonstrates a consistent nationwide shift toward electrifying transportation, aligning with sustainability goals and climate policies. Provinces are likely investing in clean energy initiatives, such as EV incentives, public transportation electrification, and charging infrastructure, to support this transition. These efforts not only reduce reliance on fossil fuels but also position Canada as a leader in sustainable energy practices.

Dataset 5: Greenhouse Gas Emissions Dataset

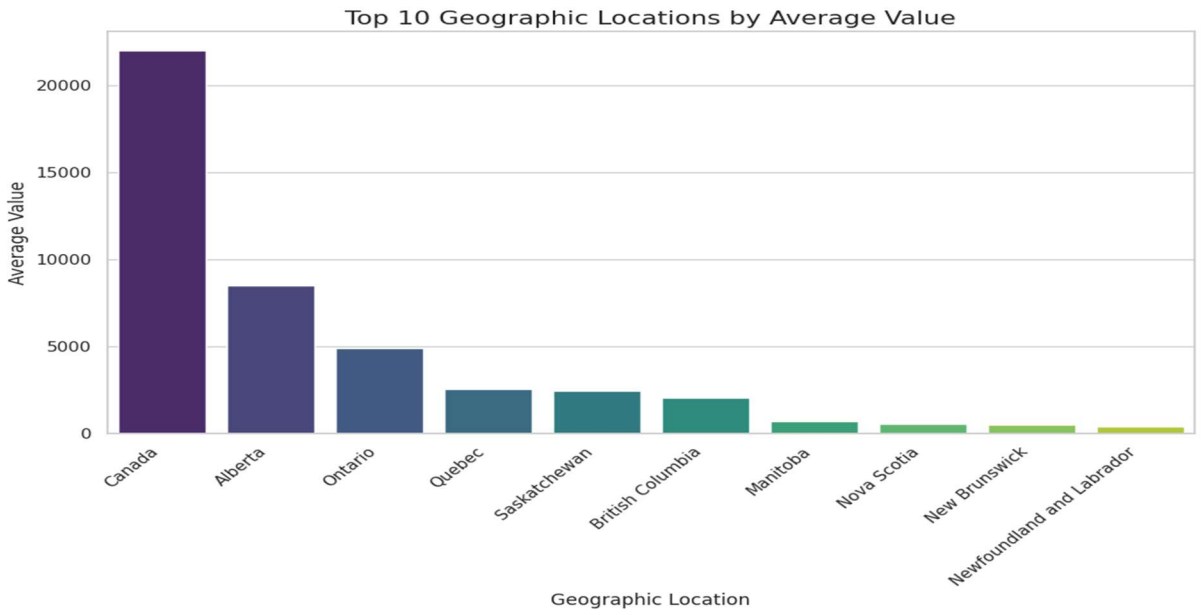
Analysis is the most required for every data set as it shows the flow of data and the information inside it in a clear and explicit way. Thus, the visualizations that had my eye from the greenhouse gas emission dataset are:

Figure 36: Graph representing the emission values with years from 2017 to 2021



This graph shows a significant value number over the years indicating the problem of today's world. But we can see a slight decrease in values from 2019- 2020 indicating covid19 and shutting down major sectors of the world had made some changes, stressing on the value, we can say that the topic of our project electricity of household is still intact.

Figure 37: Analysis of average emission value within the Canadian provinces



Here, from the plot we can clearly say that alberta has the highest emissions compared to the most populated provinces like Ontario and British Columbia due to more number of industries round Alberta.

Now, can this dataset put up more insights which can help in much more clear analysis and make the project more meaningful. Here are the SQL queries that I made during analysis part

Query1: Retrieve all data for a specific year (2019)

Figure 38: SQL query for Selecting the Highest emission year 2019 and output

```
query = """
SELECT *
FROM dataset
WHERE REF_YEAR = 2019
"""
pd.read_sql(query, con=engine)
```

	REF_YEAR	GEO	SECTOR	COORDINATE	VALUE
0	2019	Canada	Total, industries and households	1.100	775612.0
1	2019	Canada	Total, industries	1.200	639121.0
2	2019	Canada	Crop and animal production (except cannabis) [...	1.128	70290.0
3	2019	Canada	Cannabis production (licensed) [BS111CL]	1.129	1185.0
4	2019	Canada	Cannabis production (unlicensed) [BS111CU]	1.130	112.0
...	...	...	...	...	...
1478	2019	Nunavut	Balancing item: Motor fuels	14.122	-438.0
1479	2019	Nunavut	Balancing item: Aviation	14.123	49.0
1480	2019	Nunavut	Balancing item: Synthetic fluorinated gases	14.124	24.0
1481	2019	Nunavut	Balancing item: Non-energy Products from Fuels...	14.125	-1.0
1482	2019	Nunavut	Balancing item: Other differences	14.126	155.0

This shows all the rows involving 2019, the different provinces, the sectors and their emissions making sure that the values how did the emission value is higher in this year 2019.

Query 2: Calculate the total value for each sector

Figure 39: SQL query To know the emissions from each sector and output

```
query2 = """
SELECT SECTOR, SUM(VALUE) as Total_Value
FROM dataset
GROUP BY SECTOR
ORDER BY Total_Value DESC
"""
pd.read_sql(query, con=engine)
```

	SECTOR	Total_Value
0	Total, industries and households	7451177.0
1	Total, United Nations Framework Convention on ...	6979485.0
2	Total, industries	6177906.0
3	Oil and gas extraction [BS21100]	1686656.0
4	Total, households	1273266.0
...	...	...
124	Balancing item: Other differences	-35956.0
125	Balancing item: Motor fuels	-114646.0
126	Balancing item: Aviation	-119978.0
127	Total, Reconciliation with Canada's submission...	-471690.0
128	Balancing item: Biomass	-567838.0

From the above analysis we can say that both industries and households are making major contributions like about 7451177 to greenhouse gas emissions all these years. While dividing the sectors individually, the industries sector out of 7451177, contributes nearly 6177906 kilotons of emission. Then follows oils and chemical extraction and then the household with 1273266 kilotons.

Query 3: Filter records for a specific location ("Alberta")

Figure 40: SQL query for Analyzing the current living province Alberta, contributing for emission and output.

```
query = """
SELECT *
FROM dataset
WHERE GEO = 'Alberta'
"""
pd.read_sql(query, con=engine)
```

	REF_YEAR	GEO	SECTOR	COORDINATE	VALUE
0	2017	Alberta	Total, industries and households	10.100	278551.0
1	2017	Alberta	Total, industries	10.200	258619.0
2	2017	Alberta	Crop and animal production (except cannabis) [...	10.128	19966.0
3	2017	Alberta	Cannabis production (unlicensed) [BS111CU]	10.130	30.0
4	2017	Alberta	Forestry and logging [BS11300]	10.400	709.0
...	...	...	...	...	...
625	2021	Alberta	Balancing item: Motor fuels	10.122	-1107.0
626	2021	Alberta	Balancing item: Aviation	10.123	-826.0
627	2021	Alberta	Balancing item: Synthetic fluorinated gases	10.124	1613.0
628	2021	Alberta	Balancing item: Non-energy Products from Fuels...	10.125	2065.0
629	2021	Alberta	Balancing item: Other differences	10.126	-6986.0

630 rows × 5 columns

Alberta has the highest average emission value among all the provinces leaving the most populated provinces like British Columbia and Ontario. Indicating that the household or population are not the major greenhouse gas emissions, it is the industries sector that has the major contribution like about 258619 kilotons average emission value.

Query 4: Identify the sector with the highest value in a given year (2020)

Figure 41: Provides SQL query about finding the highest emission sector and output.

```
query = """
SELECT SECTOR, MAX(VALUE) as Max_Value
FROM dataset
WHERE REF_YEAR = 2020
GROUP BY SECTOR
ORDER BY Max_Value DESC
LIMIT 4
"""
pd.read_sql(query, con=engine)
```

	SECTOR	Max_Value
0	Total, industries and households	693849.0
1	Total, United Nations Framework Convention on ...	658788.0
2	Total, industries	579689.0
3	Oil and gas extraction [BS21100]	161248.0

This query is very useful to get the result clear and easy to look into the highest emission results for a conclusion and report work'

Query 5: Compute the average value for each year

Figure 42: SQL query to Analysing yearly threads along the years 2017 – 2021 and output.

```
query = """
SELECT REF_YEAR, AVG(VALUE) as Average_Value
FROM dataset
GROUP BY REF_YEAR
ORDER BY REF_YEAR
"""
pd.read_sql(query, con=engine)
```

	REF_YEAR	Average_Value
0	2017	3810.453412
1	2018	3925.847841
2	2019	3973.887390
3	2020	3613.265223
4	2021	3698.381373

Over the years, the average value of greenhouse gas emission is shown in the above query, where 2019 has the highest emission and 2020 has a decrease due to covid19 and shutting of most

major sectors round the globe. But still the emission figures around all the years are high and make a big question of Canada's challange of making the Net 0 possible!

2.2.3 Merging Datasets

To ensure a seamless and accurate analysis of our datasets, we employed a structured methodology for integrating multiple sources of data into a unified format. After performing extensive cleaning and conducting individual exploratory analyses, we merged the datasets to facilitate deeper insights and comprehensive data exploration. This section provides a detailed explanation of the steps and techniques used for this integration process, highlighting the use of Python and SQL to manage and manipulate the data.

The integration process began with the importation of essential libraries, including pandas for data manipulation and SQL alchemy for managing database operations. A SQL engine was created using the "sqlalchemy.create_engine" function, which provided a bridge between the Python environment and the SQL database. This setup allowed us to efficiently store, query, and combine the datasets using SQL's relational database capabilities.

We then loaded five key datasets into Pandas DataFrames: "Population", "Electricity Demand", "Electricity Car Usage", "Greenhouse Gas Emissions", and "Electricity Generation". During this step, special attention was given to parsing date columns accurately. For example, fields such as "DATE", "Year", and "REF_DATE" were parsed as datetime objects, ensuring the temporal integrity of the data. This was a crucial step, as it allowed us to maintain consistency in time-based analyses and facilitated precise filtering of records based on specific time ranges.

Once loaded, each dataset was imported into the SQL database using the ".to_sql()" method provided by Pandas. This process assigned each dataset to its corresponding table in the database, ensuring a clear structure and enabling easy retrieval of data for subsequent operations. The "if_exists='replace'" parameter was used during this step to overwrite any existing data in the tables, ensuring that the most recent and cleaned versions of the datasets were used throughout the analysis.

To merge the datasets, we designed and executed two SQL queries tailored to integrate relevant datasets based on common fields. The first query in figure 43, we combined the "Population" and "Electricity Demand" datasets. This integration was based on matching the "Date" and "Province" columns from both tables, ensuring alignment between population data and electricity demand across provinces. The query selected fields such as total electricity demand and its components (e.g., residential, commercial, and industrial electricity) and filtered records from 2017 onward using the "WHERE YEAR (Population.DATE) > 2016" condition. The resulting dataset, named "population_demand", provided a comprehensive view of electricity consumption trends at the provincial level in recent years.

Figure 43: Combine Population and Electricity Demand datasets

```
#Query 1: Combine Population and Electricity Demand datasets
Query = '''SELECT
    YEAR(Population.DATE) AS Year,
    GEO AS Province,
    POPULATION,
    Total_Electricity,
    Residential_Electricity,
    Commercial_Electricity,
    Industrial_Electricity,
    Transportation_Electricity
FROM
    Population
JOIN
    Demand
ON
    Population.DATE = Demand.Date
    AND Population.GEO = Demand.Province
WHERE
    YEAR(Population.DATE) > 2016
ORDER BY
    Year;
'''
```

The second query in figure 44 we integrated the "Emission" and "Generation" datasets by joining them on the "Year" and "GEO_LOCATION" columns.

Figure 44: Combine Emission and Electricity Generation datasets

```
#Query 1: Combine Emission and Electricity Generation datasets
Query = '''
SELECT
    YEAR(Generation.Year) AS Year,
    GEO,
    Electricity_VALUE_Annual AS TotalGenerated,
    VALUE AS Emission
FROM
    Generation
JOIN
    Emission
ON
    Emission.REF_YEAR = Generation.Year
    AND Emission.GEO_LOCATION = Generation.GEO
WHERE
    YEAR(Generation.Year) > 2016
    AND SECTOR = 'Electric power generation, transmission and distribution [BS22110]'
    AND Type_of_electricity_generation = 'Total types of electricity generation'
    AND Class_of_electricity_producer = 'Total all classes'
ORDER BY
    Year;
'''
```

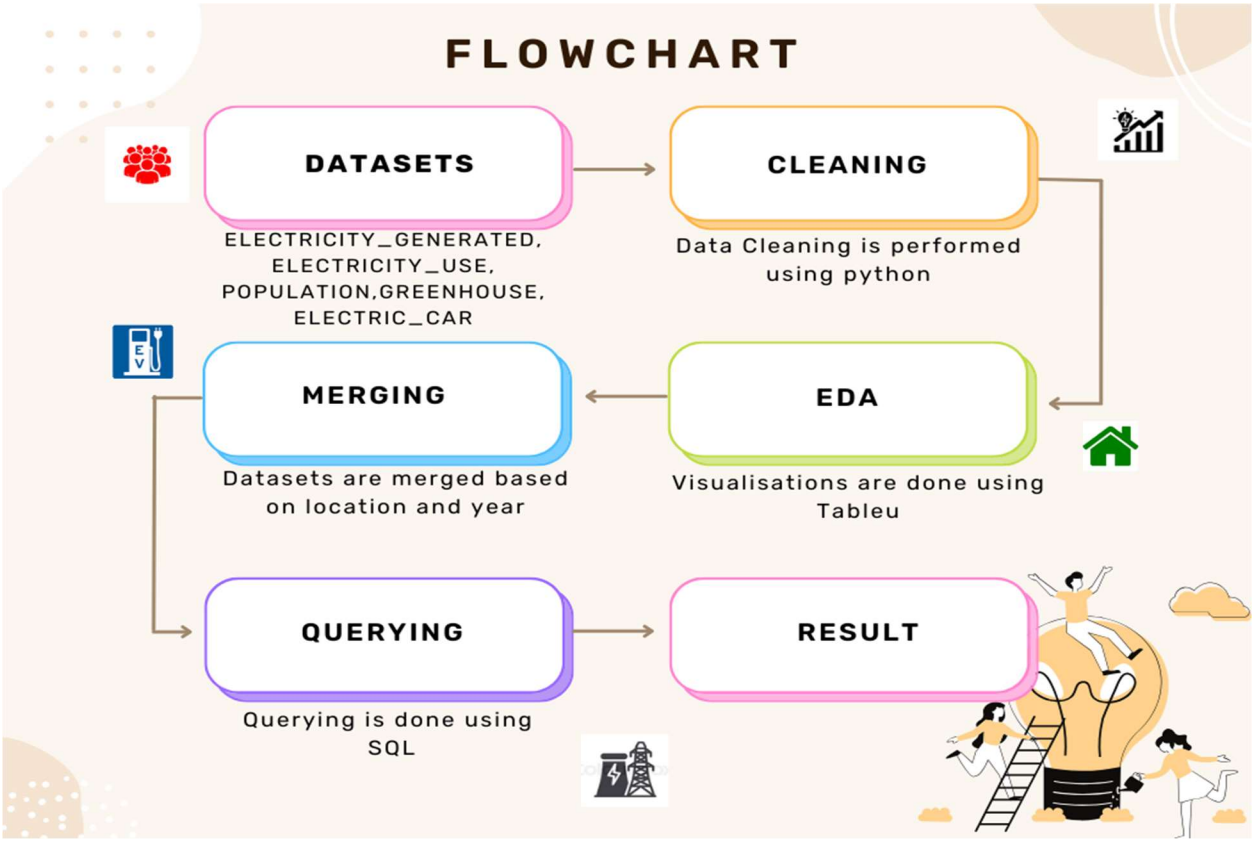
This query focused on electricity-related emissions by filtering records within the "Electric power generation, transmission, and distribution (BS22110)" sector and selecting only the relevant types of electricity generation classified under "total generation". The output of this query, stored as

"emission_generation", included variables such as annual electricity generation ("TotalGenerated") and associated greenhouse gas emissions ("Emission").

Finally, these merged datasets were exported as CSV files using the ".to_csv()" method. The "population_demand.csv" file captured key electricity demand metrics by province, while the "emission_generation.csv" file highlighted the relationship between electricity generation and emissions. These structured outputs provided a solid foundation for subsequent exploratory and statistical analyses, ensuring the alignment of all data points with the study's objectives. This systematic integration process not only streamlined data management but also allowed for a scalable and reproducible framework for analyzing complex datasets.

2.3 Workflow

Figure 45: Flow chart



The flowchart in figure 45 illustrates the sequential process followed in the electricity data analysis project, providing a clear overview of the key stages, tools, and methodologies employed. The workflow begins with the **Datasets** stage, where multiple datasets related to electricity generation, consumption, population, greenhouse gas emissions, and electric vehicle usage are gathered. These datasets form the foundation of the analysis, offering a comprehensive view of various factors influencing energy production and consumption.

The next stage is Cleaning, which involves preprocessing the collected data to ensure accuracy and consistency. This step is crucial as raw data often contains missing values, duplicates, or inconsistencies. Python was employed for data cleaning, leveraging libraries such as Pandas and

NumPy for handling large datasets efficiently. This step prepares the data for subsequent analysis by eliminating errors and standardizing formats.

Following data cleaning, the project transitions to **Merging**, where datasets are combined based on shared attributes such as location and year. This integration facilitates a holistic analysis by linking information across different datasets, enabling deeper insights. For example, merging electricity generation data with population data allows for examining per capita electricity generation trends. This step ensures that the data is structured in a way that supports exploratory and advanced analyses.

The merged datasets are then subjected to Exploratory Data Analysis (EDA), a critical phase where insights and patterns are derived using visualization tools such as Tableau. EDA serves as a preliminary analysis to uncover relationships, trends, and anomalies within the data. Visual representations such as bar charts, line graphs, and heatmaps are created to make complex data understandable and to guide further analysis.

The Querying stage follows, where SQL is employed to perform detailed data analysis and extraction. SQL enables querying specific subsets of the data, such as identifying peak electricity generation provinces or analyzing outliers in electricity production. This stage allows for precise, targeted analysis that aligns with the project objectives.

Finally, the workflow concludes with the “Result” stage, where the insights derived from the analysis are consolidated and interpreted. The results are communicated using visualizations and narratives that capture the key findings of the project. This stage not only provides answers to the research questions but also highlights actionable recommendations for improving electricity generation and management in Canada.

In summary, the flowchart reflects a structured, iterative approach to data analysis, emphasizing the importance of each stage in achieving meaningful outcomes. From data cleaning to querying and visualization, each step plays a vital role in transforming raw data into actionable insights.

2.4 Contributions

This project was a collaborative effort by a team of five members: Yifeng (Team Leader), Gulshan, Sai Namratha, Hemanth, and Lalith. Each member played a vital role in the successful execution of the project, contributing their unique skills and insights to various phases of the analysis. The work was structured in a way that allowed individual ownership of specific datasets, followed by collaborative efforts to integrate and analyze the data effectively.

The project began with each team member working independently on their assigned datasets. Yifeng, the team leader, took charge of the "Population" dataset, ensuring its accuracy and conducting preliminary analysis. Gulshan managed the "Electricity Generation" dataset, focusing on cleaning and exploring data trends related to different electricity sources. Sai Namratha worked with the "Greenhouse Gas Emission" dataset, identifying key emission patterns and preparing the

data for integration. Hemanth handled the "Electric Car" dataset, which involved examining trends in electric vehicle adoption and its correlation with electricity demand. Lalith managed the "Electricity Consumption" dataset, providing insights into provincial and sectoral consumption patterns. Each team member conducted thorough data cleaning and exploratory analysis to ensure their datasets were accurate and ready for integration.

After completing individual analyses, the team collaborated to determine compatibility between datasets for merging. This required detailed discussions to identify common fields and ensure the integrity of the combined data. Yifeng led the dataset integration process, building the logic and executing the SQL queries to merge datasets effectively. For example, the "Population" and "Electric Car" datasets were merged, and their visualization was handled primarily by Yifeng and Lalith. Their efforts highlighted significant trends in population growth and its impact on electric vehicle adoption.

Meanwhile, the "Greenhouse Gas Emission" and "Electricity Generation" datasets were merged through the combined efforts of Gulshan and Sai Namratha. This required close collaboration to address compatibility issues, with Hemanth providing valuable assistance in troubleshooting and resolving technical challenges. The resulting merged dataset offered insights into the relationship between electricity generation and its environmental impact, forming a critical part of the analysis.

Visualization and result interpretation were conducted as a team effort. Each member contributed to the design and refinement of visualizations, ensuring they effectively conveyed the trends and patterns identified in the data. Discussions were held to finalize the findings and draw conclusions, leveraging the diverse perspectives and expertise of all team members. These collaborative discussions played a pivotal role in achieving a comprehensive understanding of the data and deriving meaningful insights.

In summary, this project was a true team effort, with every member contributing significantly to the success of the analysis. The structured division of tasks, combined with a collaborative approach to integration and visualization, ensured that the project was executed efficiently and effectively. The shared responsibility for discussions and decision-making fostered a collective sense of achievement, culminating in a well-rounded and insightful analysis of Canada's electricity landscape.

3. MAIN RESULTS OF THE ANALYSIS

3.1 Results

Following the integration of datasets, as outlined in the methodology, the analysis yielded significant insights into electricity consumption, generation, and the broader environmental and societal impacts of these trends in Canada. By merging datasets on population growth, electricity demand, greenhouse gas emissions, and electric vehicle (EV) adoption, a comprehensive picture of Canada’s energy landscape was revealed.

The integrated data revealed a notable disparity between electricity consumption and generation. Electricity consumption has shown a steady increase over the years, driven primarily by population growth and the rising adoption of EVs. However, electricity generation has not kept pace with this demand. This growing gap highlights a critical concern about the sustainability of Canada's energy supply in meeting its rising needs, especially with ambitious environmental goals like achieving net-zero emissions by 2035.

One of the key drivers of electricity demand identified in the analysis was population growth. Since 2020, Canada's population has increased by 5.4%. This population growth directly correlates with an increased need for electricity, not only for residential and industrial purposes but also for supporting the rapidly growing EV market. EV adoption has seen significant growth, rising from 2.5% of vehicles in 2020 to 8.4% in 2023, marking a 5.9% increase in just three years. This growth, while promising from a sustainability perspective, places additional strain on electricity generation, as EVs require substantial charging infrastructure and electricity supply to function effectively.

The analysis also shed light on the environmental challenges associated with electricity generation. While there has been a notable decline in greenhouse gas (GHG) emissions and an increased focus on renewable energy sources, the rate of decline is insufficient to meet Canada's net-zero targets. The data indicates that renewable energy resources, though growing in prominence, are not replacing non-renewable resources at the pace necessary to offset emissions from electricity generation. Moreover, the environmental benefits of EV adoption are somewhat offset by the slow transition to clean energy, as electricity generation continues to rely on significant non-renewable energy sources in some provinces.

In light of this, it is essential to reference the perspectives shared in *The Globe and Mail* article, *What does Canada's journey to net zero look like?*, which discusses Canada's ambitious net-zero target. Experts from the Net Zero Advisory Body (NZAB), including scientists and economists, expressed a mix of urgency and optimism regarding Canada's ability to meet its climate goals. While the advisory body acknowledges that Canada is fully equipped with the tools needed to achieve net zero, there is a shared consensus on the need for immediate and aggressive action. As one NZAB member, Abreu, notes, governments and industries must take the lead, but individual Canadians also have a role to play in driving this transformation.

These expert views align closely with our findings. While the data illustrates a concerning gap between electricity demand and generation, the optimism expressed by the NZAB reflects the broader potential for renewable, zero-carbon energy to reshape Canada's electricity landscape. Donner, another NZAB member, highlighted how renewable energy sources have become cheaper than fossil fuel generation—a shift unimaginable two decades ago. This underscores the importance of accelerating the transition to renewables, a sentiment echoed by our analysis, which identifies the insufficient growth of renewable resources as a key barrier to achieving net zero.

Finally, a predictive analysis based on the trends in electricity demand, generation, and environmental impacts reveals a sobering conclusion. While the Canadian government has set an ambitious target of achieving net-zero electricity generation by 2035, the current trajectory suggests that this goal is unlikely to be achieved within the set timeline. The interplay of increasing

electricity demand, insufficient generation growth, and the slow adoption of renewable resources presents a substantial challenge to realizing this vision.

These insights emphasize the need for immediate and accelerated policy interventions to address the widening gap between electricity consumption and generation, promote renewable energy infrastructure, and reduce GHG emissions. While optimism about achieving net zero exists, as the NZAB experts highlight, it is clear from both our analysis and the perspectives shared in the article that success depends on a collaborative effort involving governments, industries, and citizens. Without significant advancements in clean energy technology and more aggressive policy implementation, Canada's net-zero targets will remain aspirational rather than achievable.

4. DISCUSSION AND CONCLUSION

4.1 Discussion

The analysis conducted in this study has revealed critical insights into the dynamics of Canada's electricity consumption, generation, and associated environmental impacts. By integrating datasets on population growth, electricity demand, greenhouse gas emissions, and electric vehicle (EV) adoption, we developed a comprehensive understanding of the challenges and opportunities within Canada's energy landscape. This section synthesizes these findings and highlights their implications, with a focus on actionable recommendations.

One of the most significant findings of our analysis is the growing disparity between electricity demand and generation in Canada. As population growth and EV adoption have surged in recent years, electricity consumption has followed a steady upward trajectory. This trend is largely driven by a 5.4% increase in population since 2020 and a dramatic rise in EV usage, which grew from 2.5% of vehicles in 2020 to 8.4% in 2023. However, electricity generation has not kept pace with this increased demand, raising concerns about the sustainability of Canada's energy supply. Without substantial investment in generation capacity, this gap is expected to widen, potentially compromising economic growth and energy security.

The environmental implications of these trends further compound the challenges. While our analysis indicates a decline in greenhouse gas emissions and an increasing reliance on renewable energy sources, the rate of progress remains insufficient to meet Canada's ambitious net-zero emissions target by 2035. Non-renewable energy sources continue to account for a significant share of electricity generation, particularly in certain provinces. This reliance undermines the environmental benefits of EV adoption, as clean transportation initiatives are only as effective as the energy sources that power them. Accelerating the transition to renewable energy is therefore critical to achieving Canada's climate goals. Despite these challenges, our findings also highlight opportunities for progress. Renewable energy sources have become increasingly cost-effective, presenting a viable path toward sustainable energy generation. The decreasing cost of renewables, coupled with technological advancements, underscores the feasibility of a rapid transition away

from fossil fuels. Moreover, coordinated efforts by governments, industries, and citizens can create a foundation for achieving net-zero emissions within the next decade.

To address these challenges, several recommendations emerge from this study. First, policymakers must accelerate investments in renewable energy infrastructure to meet the growing demand while reducing greenhouse gas emissions. This includes expanding solar, wind, and hydroelectric capacity and modernizing the grid to accommodate renewable integration. Second, more aggressive and enforceable policy interventions are necessary to incentivize renewable energy adoption, improve energy efficiency, and enhance EV infrastructure. Third, fostering collaborative efforts among stakeholders as governments, industries, and the public is essential to driving the cultural and structural changes needed to achieve net-zero emissions. Finally, continuous monitoring and analysis of electricity demand, generation, and emissions will be crucial for guiding and refining policy decisions.

4.2 Conclusion

In conclusion, the findings of this analysis emphasize the urgency of addressing the growing gap between electricity consumption and generation, the slow transition to renewable energy, and the environmental challenges posed by continued reliance on non-renewable resources. While the challenges are significant, the data suggests that with immediate and decisive action, Canada's net-zero emissions target for 2035 remains within reach. Achieving this vision requires a shared commitment to renewable energy, robust policy frameworks, and collective action across all levels of society. This study underscores the importance of leveraging data-driven insights to inform strategies and pave the way for a sustainable energy future.

4.3 Approach

To achieve the objectives of this study, a systematic and data-driven approach was employed to collect, clean, integrate, and analyze multiple datasets related to electricity consumption, generation, and environmental impacts in Canada. The process began with the acquisition of five key datasets that are "*P*opulation", "Electricity Demand", "Electric Vehicle Adoption", "Greenhouse Gas Emissions", and "Electricity Generation". Each dataset underwent an extensive cleaning process to ensure accuracy and consistency, with a particular focus on addressing missing values, standardizing column names, and ensuring the integrity of temporal fields such as dates. Python and SQL were the primary tools used for data manipulation and integration. Libraries such as Pandas were employed for initial data cleaning and exploratory analysis, while SQL Alchemy facilitated the creation of a relational database environment to manage and combine datasets effectively.

After preparing the data, a structured integration process was implemented. Each dataset was imported into SQL using Pandas' .to_sql() method, assigning them to corresponding tables in the database. This relational structure allowed for efficient querying and joining of datasets based on common fields such as dates and geographic locations. Two key SQL queries were designed to integrate the data: the first combined *Population* and *Electricity Demand* data to analyze electricity trends by province and year, while the second merged *Greenhouse Gas Emissions* and *Electricity*

Generation data to explore the relationship between emissions and energy production. Special attention was given to filtering and aggregating records to align with the study's objectives, such as focusing on electricity generation in the "Electric Power Generation" sector and limiting the analysis to records from 2017 onward.

Finally, the integrated datasets were exported as structured CSV files, facilitating further exploratory and statistical analyses. The outputs included a merged dataset on provincial electricity demand and population growth and another on emissions and electricity generation trends. This systematic approach ensured that the analysis was both scalable and reproducible, providing a robust foundation for deriving insights and supporting evidence-based recommendations. If we wanted to do anything different, we would have liked to find datasets that has various other factors included like economy, weather and climate and government policies as these also tend to influence the electricity generated and consumed which in turn affects the environment.

4.4 Future Work

This study has provided valuable insights into electricity consumption, generation, and their environmental impacts in Canada, but there are opportunities for further exploration to enhance the understanding and application of the findings. Future work could focus on investigating additional factors that influence electricity demand, such as seasonal variations, socioeconomic trends, industrial activities, and advancements in energy-efficient technologies. Incorporating these variables into the analysis could provide a more comprehensive understanding of the dynamics driving electricity demand and help policymakers design more targeted interventions.

Another avenue for future research involves developing improved predictive models to estimate greenhouse gas emissions. While this study focused on trends and associations, the incorporation of machine learning techniques or advanced statistical models could yield more accurate forecasts. These models could account for nonlinear relationships between electricity generation methods, energy policies, and emissions, offering deeper insights into the pathways toward achieving net-zero targets.

Moreover, future studies could explore the regional disparities in electricity generation and renewable energy adoption. A granular analysis of provincial energy policies, infrastructure investments, and resource availability would provide a clearer picture of the challenges and opportunities faced by different regions. Additionally, a sector-based analysis of electricity consumption and emissions leading to examining residential, commercial, and industrial sectors could help identify specific areas where interventions would be most effective.

Lastly, further work could integrate international benchmarks and comparisons to assess Canada's progress in transitioning to sustainable energy practices. Analyzing data from countries with advanced renewable energy systems and net-zero policies could offer valuable lessons and strategies for accelerating Canada's energy transition. By addressing these areas, future research can build on the findings of this study, offering more robust insights and actionable recommendations to support sustainable energy development.

5. REFERENCES

5.1 Reference

1. Residential Electricity and Natural Gas Plans. EnergyRates.ca. (2020, September 1). https://energyrates.ca/residential-electricity-natural-gas/ .

2. Government of Canada, C. E. R. (2024, September 10). Canada energy regulator / Régie de l'énergie du Canada. CER. https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-canada.html .

3. Canada, S. (2024, September 3). Government of Canada. Canada.ca. https://www.canada.ca/en/services/environment/weather/climatechange/climate-plan/net-zero-emissions-2050.html .

4. Canada, E. and C. C. (2022, March 16). Government of Canada. Canada.ca. https://www.canada.ca/en/environment-climate-change/services/canadian-environmental-protection-act-registry/achieving-net-zero-emissions-electricity-generation-discussion-paper.html .

5. Government of Canada, Statistics Canada. (2024, September 25). Population Estimates, quarterly. Population estimates, quarterly. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901&cubeTimeFrame.startMonth=07&cubeTimeFrame.startYear=2016&cubeTimeFrame.endMonth=07&cubeTimeFrame.endYear=2024&referencePeriods=20160701%2C20240701%29 .

6. Statistics Canada. (2024, November). Electricity generation in Canada and provinces from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510001501&pickMembers%5B0%5D=1.1&pickMembers%5B1%5D=2.1&cubeTimeFrame.startMonth=03&cubeTimeFrame.startYear=2018&cubeTimeFrame.endMonth=03&cubeTimeFrame.endYear=2024&referencePeriods=20180301%2C20240301 .

7. Statistics Canada Open Data Portal (2024, October). https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2310030801.

8. *Canada's Energy Future 2023: Energy Supply and Demand Projections to 2050 - end-use-demand-2023 - Open Government Portal*. (n.d.). https://open.canada.ca/data/en/dataset/7643c948-d661-4d90-ab91-e9ac732fc737/resource/9003d40e-087b-4af0-b2c8-2217fd697a28 .

9. Government of Canada Open Data (2024, November) https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64?_gl=1%2A1vojehp%2A_ga%2AMTAxOTA2NzU0Ni4xNzI3MDQ0MjIw%2A_ga_S9JG8CZVYZ%2AMTczMDQ5NzMxMi4xLjEuMTczMDQ5NzM1Mi4yMC4wLjA.

10. *Canada's Official Greenhouse Gas Inventory*. Open Government Portal. (n.d.).

https://open.canada.ca/data/en/dataset/779c7bcf-4982-47eb-af1b-a33618a05e5b

11. Government of Canada, C. E. R. (2024, May 17). *Canada energy regulator / Régie de l'énergie du Canada*. CER. https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/market-snapshots/2024/market-snapshot-ghg-emissions-from-on-site-electricity-generation-and-cogeneration-at-canadas-energy-intensive-facilities-are-increasing.html

12. Pope, S. (2023, March 28). *Canada's journey to net zero: What does it look like?*. Canadian Geographic. Retrieved December 8, 2024, from https://canadiangeographic.ca/articles/canadas-journey-net-zero/

End of Project Report