Impact of Greenhouse Gas Emissions on Global Temperatures

Mahtab Ranji

June 6, 2024

1 Introduction

Climate change is a major global concern, driven by greenhouse gas emissions from human activities. This study examines how these emissions have influenced global temperatures over the past 25 years, from 1995 to 2020. We use two key datasets for our analysis:
Global Temperature Data: This dataset provides annual average temperatures for all countries, allowing us to track temperature changes over time and across regions.
CO2 and Greenhouse Gas Emissions Data: This dataset includes annual CO2 and other greenhouse gas emissions by region, giving us insight into emission trends and their potential impact on global temperatures.
Our goal is to identify patterns and connections between greenhouse gas emissions and temperature changes. By understanding these relationships, we can gain insights into the effects of human activities on climate change, which can help guide future policies and actions.
The main question I aim to answer is: How do greenhouse gas emissions correlate with global temperature changes from 1995 to 2020?

2 Data Sources

I used two datasets in this projects and both datasets appear to have high completeness and accuracy, with consistent formatting and minimal missing values.:

2.1 Dataset Description

CO2 and Greenhouse Gas Emissions

- **Source:** `https://www.kaggle.com/datasets/imtkaggleteam/co-and-greenhouse-gas-emissions`
- **Contains:** Annual CO2 emissions by region
- **License:** Other
  The license for this dataset is categorized as "Other," with no specific license specified. According to Kaggle's guidelines for "Other" licenses, users must specify the details in the dataset summary. Based on the author's statements in the dataset descriptions, it appears that the datasets are open for use in analysis.
  PS: Initially, I considered using the "Greenhouse Gas Giants" dataset, which specifies CO2 production by different commodities(Oil, Natural Gas, ...). However, I needed CO2 emissions data produced by countries, so I switched to a more suitable dataset that provides national-level emissions.

Temperature of All Countries (1995-2020)

- **Source:** `https://www.kaggle.com/datasets/subhamjain/temperature-of-all-countries-19952020`
- **Contains:** Contains data of various major cities of different countries in the world.
- **License:** [Database Contents License (DbCL) v1.0]
  Open Data Commons is not a law firm and does not provide legal services of any kind. Open Data Commons has no formal relationship with you. Your receipt of this document does not create any kind of agent-client relationship. Please seek the advice of a suitably qualified legal professional licensed to practice in your jurisdiction before using this document.

3 Data Pipeline

3.1 Overview

Python was the primary technology used for data manipulation libraries and ease of use..

- `os` module manages the file system interactions.
- `pandas` handles data loading, cleaning, transformation, and merging.
- `KaggleApi` facilitates data acquisition from Kaggle.
- `sqlite3` manages data storage in a lightweight, disk-based database, providing an efficient way to query and analyze the merged dataset.

3.2 Cleaning and Transformation Steps

1. **Loading Data**: The data files were downloaded from Kaggle using the Kaggle API.
2. **Data Cleaning**: The temperature data was cleaned by removing rows with missing or erroneous temperature values (-99). Additionally, country names were standardized (e.g., 'US' to 'United States').
3. **Data Transformation**: The cleaned temperature data was aggregated to compute the annual average temperature for each country. The CO2 emissions data was renamed and prepared for merging.
4. **Merging Datasets**: The transformed temperature and CO2 emissions datasets were merged based on the common attributes of year and country.
5. **Database Storage**: Saved the merged dataset into an SQLite database for structured querying and analysis.

3.3 Error Handling

The pipeline includes checks for missing values and handles mismatches that are:
Implemented checks for missing values and standardized country names.
Ensured that datasets were correctly aligned on the common keys (year and country).

3.4 Challenges and Solutions

- **Challenge:** Inconsistent country naming conventions across datasets.
- **Solution:** Standardized country names to ensure proper merging.
- **Challenge:** Handling invalid temperature values.
- **Solution:** Removed or corrected invalid entries and converted temperature units.

4 Result and Limitations

4.1 Output Data

The output of the data pipeline is a merged dataset containing information on annual average temperatures and CO2 emissions for various countries from 1995 to 2020. The data structure consists of columns for year, country, average temperature, and CO2 emissions.

- **Data Structure:** The data is structured with columns for Year, Country, AvgTemperature, and Annual CO2 emissions. The quality is high due to rigorous cleaning steps.
- **Format:** SQLite database and pandas DataFrame.
- **Reason:** SQLite provides a structured format for efficient querying, while DataFrame allows for easy manipulation and visualization.

4.2 Data Quality

The dataset is clean, standardized, and suitable for correlation analysis. However, it may still have limitations due to missing historical data for certain countries or years.

4.3 Critical Reflection

While the data pipeline successfully integrates and analyzes temperature and CO2 emissions data, potential limitations may include in the original datasets.

- **Data Gaps:** Some countries or years might have missing data, which can affect the analysis.
- **Accuracy:** The accuracy of the results depends on the quality of the source datasets. Any inaccuracies in the original data will propagate through the analysis.
- **Bias:** The datasets might not cover all countries equally, leading to potential bias in the analysis.

5 Figures and Tables

In the last step, I performed some checks, to ensure the integrity of the analysis and avoid potential errors or biases caused by mismatched data coverage. It helps in validating the datasets and provides confidence in the subsequent analyses and conclusions drawn from them.

5.1 Checking Years Coverage

- **Purpose:** This ensures that both datasets cover the same time period, which is essential for meaningful analysis.
- **Importance:** If the datasets have different years of coverage, it may lead to incorrect interpretations or incomplete analyses. By identifying common years, you ensure that you're analyzing data for the same timeframe.
- **Action Taken:** The code extracts unique years from both datasets and determines the common years. It then prints out the years covered in each dataset and the common years.

5.2 Checking Countries Coverage

- **Purpose:** Similar to checking years coverage, this step ensures that both datasets cover the same set of countries.
- **Importance:** If there are discrepancies in the countries covered, it could lead to biased results or incomplete insights.
- **Action Taken:** The code extracts unique countries from both datasets and identifies the common countries. It then prints out the countries covered in each dataset and the common countries.