Data Report: CO2 Emissions and Temperature Rise: A Global and Regional Analysis

1. Question

- What is the correlation between CO2 emissions and global temperature rise, and how do regional differences affect this relationship?

2. Data Sources

Our World in Data - CO2 Emission Dataset

- Why Chosen: This dataset provides comprehensive data on CO2 and greenhouse gas emissions, allowing for an analysis of CO2 emissions trends by country and region.
- Origin: The dataset is provided by Our World in Data and is accessible through their GitHub repository.
- Data Content: The dataset contains data on global CO2 and greenhouse gas emissions by country and region over time.
- Data Structure and Quality: The data is in CSV format, with columns for country, year, CO2 emissions, and other metrics like population, GDP, oil_co2_per_capita, land_use_change_co2_per_capita. Because the dataset is regularly updated, it is well-maintained and has a clear structure.
- Licenses and Compliance: The data is available under the Creative Commons BY license, which means that the data can be shared (copied aand redistributed) and adapted (mixed and transformed) in any way as long as appropriate credit and a link to the license is given.

HadCRUT 5 - Temperature Anomaly Dataset

- Why Chosen: This dataset provides comprehensive data on global temperature through gridded temperature anomalies across the world.
- Origin: This dataset is provided by the Climatic Research Unit (CRU) at the University of East Anglia and the Met Office Hadley Centre.
- Data Content: This dataset contains monthly global temperature anomalies from 1850 to the present on a 5 degree grid.
- Data Structure and Quality: The data is in netCDF format, containing variables such as time, latitude, longitude, and temperature anomalies (tas_mean). The quality of the dataset is high, with well-documented variables and reliable historical data.
- Licenses and Compliance: The data is available under the Open Government License, which means that the data can be used, shared, adapted, and exploited commercially and non-commercially as long as appropriate credit and a link to the license is given

By combining the CO2 emission dataset from *Our World in Data* and the global temperature anomaly dataset from *Met Office Hadley Centre and CRU*, I can analyze the relationship between CO2 emissions and temperature changes. The CO2 emission dataset provides detailed information about emissions on a global and regional level, while the temperature anomaly dataset provides global and regional temperature data. This combination allows to investigate if and how CO2 emissions correlate with the global temperature rise. By further taking the regional information into account, I can also analyze how this relationship varies across selected regions. This allows to identify potential regional variations in the impact of CO2 emissions on temperature change, providing a deeper understanding of the global climate change.

3. Data Pipeline

The data pipeline is written in Python and makes use of following libaries to extract, transform, clean, and store the updated data: numpy, pandas, sqlite3, cftime, and netCDF4. **Pipeline Steps:**

-
    a. **Download Data:** Automatically download the HadCRUT 5 and Our World in Data CO2 datasets from the source URLs.
-
    a. **Load Data:** Read the datasets using the netCDF4 and the pandas library.
    - The **HadCRUT 5 dataset** was provided in the netCDF format, so it first had to be read in as a netCDF file using *netCDF4* and was then converted to a pandas DataFrame for efficient handling.
    - The **CO2 emission dataset** was provided in CSV format, so it was directly read into a pandas DataFrame.
-
    a. **Transform Data:** Convert time units for the HadCRUT 5 dataset and extract relevant variables and columns
    - To be able to efficiently use the time information, it was converted from "days since 1850-01-01" to a standard datetime format using the *cftime* library, resulting in following format: "%Y-%m-%d %H:%M:%S".
    - To ensure only relevant columns are stored and further used in the upcoming analysis, only the important columns were selected from the datasets. From HadCRUT 5, the columns '**time**', '**latitude**', '**longitude**', '**temperature anomaly**' are keept. From the CO2 emission dataset, the identified key columns are '**country**', '**year**', '**co2**', '**co2_per_capita**', '**co2_per_gdp**','**temperature_change_from_co2**'.
    - To be able to efficiently compare both datasets, the multi-dimensional HadCRUT 5 data had to be converted into a 2D DataFrame. This involved repeating the '**time**' dimension to match each combination of latitude and longitude, tiling the '**latitude**' and '**longitude**' dimension to align with the repeated '**time**' dimension and flattening the '**temperature_anomaly**' dimension such that each temperature anomaly entry corresponds to the appropriate '**time**', '**latitude**', and '**longitude**' value.
-
    a. **Save Data:** Store the transformed data in an SQLite database for efficient querying and analysis.

4. Problems Encountered and Solutions

- The HadCRUT 5 dataset is provided in the netCDF format with a '**.nc**' file extension. Direct download was problematic, requiring the addition of '**mode=bytes**' to the URL to be able to download it successfully.

- Furthermore, the HadCRUT 5 dataset uses a non-standard time format ("days since 1850-01-01"), which differs from the CO2 emission dataset. This discrepancy in time formats didn't allow direct comparison and had to be converted to a matching time format to ensure time consistency among both datasets.
- Also, the HadCRUT 5 dataset was structured in multi-dimensional arrays, which wasn't suitable for direct analysis and comparison with the CO2 emission dataset, which is in tabular CSV format. Therefore, the structure had to be converted into a 2D DataFrame.

5. Results and Limitations

Results: The output of the data pipeline consists of the following two tables:

- co2_emission_dataset.sqlite
- temperature_dataset.sqlite

Limitations: The main limitation of the HadCRUT 5 dataset is the spatial resolution of the temperature data, as this data is represented through a 5x5 degree grid, which may not be sufficient for a country-level analysis. While the resolution will probably be sufficient for global trends, it may smooth over the country-specific variations in temperature and therefore might limit the accuracy for the regional analysis.

Completness: Both datasets selected for this project are comphrehensive and conver a large period of time. While the HadCRUT 5 dataset spans from 1850 to the present, the CO2 emission dataset includes detailed annual data on country-level basis. Combining both datasets will result in the necessary information needed to analyze the correlation between CO2 emission and tempererature rise, and are therefore regarded as complete for this project.

Representativeness: Both datasets are a collection of global measurements, making them highly representative for studying global and regional trends. The CO2 emissions dataset includes data from a wide range of countries and regions, resulting in a comprehensive view of global emissions. Similarly, the HadCRUT 5 dataset provides temperature data on a global basis, allowing for a broad analysis of temperature changes.

Correctness: The datasets are obtained from known institutes such as the Met Office Hadley Centre, the Climatic Research Unit and verified national records which should ensure a high accuracy and reliability.

Availability: Both datasets are publicly available and easily accessible. The HadCRUT 5 dataset can be downloaded from the Climatic Research Unit's website, while the CO2 emission dataset is available on GitHub. So as long as the dataset publishers don't remove or restrict access to the datasets, either by removing the dataset or by changing the license, the data can be used for my further analysis.