1. Main Question

Which countries have the lowest amount of increasing CO2 emissions compared to their population growth?

2. Data Sources

To answer the question, two data sources have been selected for this project: both of them are from The World Bank. First one provides CO2 emissions(in kilotons) data for all countries in the world and the second one shows the total population numbers for all countries in the world.

Data source 1: CO2 Emissions Dataset

Metadata URL: https://data.worldbank.org/indicator/EN.ATM.CO2E.KT

Data URL: https://api.worldbank.org/v2/en/indicator/EN.ATM.CO2E.KT?downloadformat=csv

Description: The World Bank dataset, accessed through the indicator code EN.ATM.CO2E.KT, contains annual CO2 emissions data for all countries, providing a comprehensive view of global emissions. This dataset includes historical data (1990 - 2020), allowing for the analysis and comparison of CO2 emission trends over time. Each entry in the dataset includes the country name, country code, year and the corresponding CO2 emission value. The World Bank is a reputable and authoritative source, ensuring that the data is reliable and has been collected using standardized methodologies.

Source: Climate Watch Historical GHG Emissions (1990-2020). 2023. Washington, DC: World Resources Institute.

Data Structure & Quality: The dataset is structured as a CSV directory which is a text file format that uses commas to separate values and newlines to separate records. This dataset has tabular data (numbers and text) in strings, where each line of the file represents one data record. The dataset contains real world data and no duplicate or invalid information. However, some rows and columns have missing values. It has consistency in its format, all variables are in string format. Dataset has no data for the last 3 years, the most recent year is 2020. The data aligns with the needs of the project.

License and Obligations: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). Link to the license: https://creativecommons.org/licenses/by-nc/4.0/

I will use/share/adapt the data by filling following obligations: giving appropriate credit , providing a link to the license and indicating if changes were made. Also, not using the material for commercial purposes.

Data source 2: The Total Population Dataset

Metadata URL: https://data.worldbank.org/indicator/SP.POP.TOTL

Data URL: https://api.worldbank.org/v2/en/indicator/SP.POP.TOTL?downloadformat=csv

Description: The World Bank dataset, accessed through the indicator code SP.POP.TOTL, contains

annual total population data for all countries, providing a comprehensive view of global population. This dataset includes historical data (1960 - 2022), allowing for the analysis and comparison of population growth over time. Each entry in the dataset includes the country name, country code, year and the corresponding population value. The World Bank is a reputable and authoritative source, ensuring that the data is reliable and has been collected using standardized methodologies.

Source: (1) United Nations Population Division. World Population Prospects: 2022 Revision. (2) Census reports and other statistical publications from national statistical offices, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Population and Vital Statistics Reprot (various years), (5) U.S. Census Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Programme.

Data Structure & Quality: The dataset is structured as a CSV directory. This dataset has tabular data (numbers and text) in strings. The dataset contains real world data and no duplicate or invalid information. However, some rows and columns have missing values. It has consistency in its format, all variables are in string format. Dataset has no data for the last year(2023), the most recent year is 2022. The presentation of data fits the project's objective.

License and Obligations: Creative Commons Attribution 4.0 (CC-BY 4.0). Link to the license: https://datacatalog.worldbank.org/public-licenses#cc-by

I will copy, modify and distribute data in any format for any purpose, including commercial use by filling following obligations: giving appropriate credit , providing a link to the license and indicating if changes were made including translations.

3. Data Pipeline

The data pipeline described here involves several key steps: data acquisition, extraction, transformation, cleaning and storage. The pipeline is implemented in Python using various libraries and tools:

- **Requests**: For downloading data from the web.
- **Pandas**: For data manipulation and cleaning.
- **Zipfile**: For extracting zipped data files.
- **SQLite3**: For storing cleaned data in a database.
- **OS**: For file and directory operations.

3.1. Transformation and Cleaning Steps

- ➢ **Remove Columns with All NaN Values**: Eliminates irrelevant data.
- ➢ **Remove Rows with Few Non-NaN Values**: Ensures data quality.
- ➢ **Keep Specific Columns**: Focuses on relevant indicators and years (1990-2020).
- ➢ **Remove All Remaining NaN Values**: Ensures completeness.

3.2. Issues Encountered and Solutions

- ➢ **Unnamed Column Issue**: Both datasets contained an 'Unnamed: 68' column, which was unnecessary. I resolved this by deleting the column from both datasets.
- ➢ **Missing Data for 2023**: The datasets did not include data for the year 2023. Therefore, I removed this column to ensure consistency.

- ➢ **Zero Values in CO2 Emissions**: Initially, I deleted zero values in the CO2 emissions dataset, assuming they were inaccurate. However, I later learned that some countries had zero CO2 emissions in certain years. Consequently, I decided to retain these zero values.
- ➢ **Mismatched Timeframes**: The CO2 emissions dataset covered the years from 1990 to 2020, while the population dataset spanned from 1960 to 2022. To align the datasets, I filtered the population data to match the timeframe of the CO2 emissions dataset, focusing only on the years from 1990 to 2020.

3.3. Error Handling

The pipeline checks the HTTP status code during data download to ensure successful retrieval. If the download fails, an error message is printed, and the function returns False. During data extraction, the zipfile module assumes a correct ZIP file structure, and any extraction issues will raise exceptions. The load_data function processes only valid CSV files, excluding metadata, and returns None if no suitable files are found.

3.4. Handling Changing Input Data:

- **Flexible Column Handling**: The clean_data function dynamically constructs a list of year columns (1990-2020). This can be easily adjusted for different time ranges by changing the range.
- **Dynamic Column Selection**: The pipeline processes any columns that match the expected patterns (e.g., year columns), making it adaptable to changes in the dataset's structure.
- **Modular Functions**: Each step (download, extract, load, clean, store) is modular, making it easy to adjust or replace parts of the pipeline to accommodate new requirements or data formats.

4. Result and Limitations

The resulting data structure is a pandas DataFrame containing information on CO2 emissions and population from World Bank indicators. The quality of the output is ensured for further analysis through various data transformation and cleaning steps. Data reflects the real world and is correct. Also, it contains all necessary information and aligns with the needs of the project. All the data needed for analysis is consistent in format. Age of data is appropriate. Overall, the result is a cleaned and structured DataFrame containing reliable data on CO2 emissions and population. The output format chosen is SQLite databases due to their structured storage, ease of integration, query capabilities, portability, and suitability for small to medium-sized datasets.

Although the data pipeline effectively processed the data without encountering any missing values for the specified countries and years, it's important to note that potential outliers might emerge during the analysis phase. Furthermore, given that the datasets encompass data from all countries worldwide and span a broad timeframe, it's reasonable to assume that the data is unbiased, allowing for the potential generalizability of conclusions.