Assessing the relationship between renewable energy consumption and CO2 emissions across nations

I. Motivation

The combustion of fossil fuels such as coal, petroleum, and natural gas emits a significant amount of carbon dioxide (CO_2), accounting for the largest share of greenhouse gases, which are associated with global warming. To reverse or at least mitigate climate change, many countries are shifting their main energy source away from fossil fuels and towards renewable, sustainable alternatives such as solar or wind energy. This project aims to analyze the global consumption of renewable energy and assess its influence on the total CO_2 emissions, exploring how renewable energy supports the fight against climate change. The main question of this project is:

What is the correlation between renewable energy consumption and CO_2 emissions in various countries?

II. Data Sources

To analyze the relationship posed in the question, it is crucial to have a dataset containing CO_2 emission data and another containing renewable energy consumption data from various countries around the world.

The CO_2 emission dataset is provided by The World Bank, with the original data sourced from Climate Watch – Historical GHG Emissions (1990–2020) (2023, Washington, DC: World Resources Institute). Similarly, the renewable energy consumption dataset is also supplied by The World Bank. This data is collected from multiple sources, including the IEA, IRENA, UNSD, World Bank, and WHO (2023 Tracking SDG 7: The Energy Progress Report, Washington DC).

Both datasets comprise of data from 266 countries and associations from 1990 to 2020, making them suitable for analyzing the relationship in question.

The datasets are generally of high quality, adhering to several key data quality criteria:

1. **Accuracy**: The collected data, based on official reports by national governments under the UNFCCC and gathered by reputable institutions and research organizations, ensures that the figures reflect actual expert measurements and estimations.
2. **Completeness**: The datasets, covering the period from 1990 to 2020/2021, include 266 countries and associations. However, there are still a few missing values for certain countries and specific years.
3. **Consistency**: The dataset is presented in a tabular format, where each row corresponds to a country and each column represents a specific year from 1990 to 2020/2021. Each cell contains the value of CO2 emissions or renewable energy consumption for a country in a year.
4. **Timeliness**: The data was last updated in 2023 with a 2-to-3-year lag. Although the most recent years are not included, the dataset's coverage from 1990 to 2020/2021 is sufficient for understanding long-term underlying patterns.
5. **Relevancy:** The data is highly relevant for users involved in environmental studies, policy making, climate change and renewable energy consumption analysis. It is essential for developing strategies to tackle global environmental challenges.

The CO2 Emission dataset is licensed under CC BY-NC 4.0, which can be found under the Metadata URL here: World Bank CO2 Emission Metadata.

The renewable energy consumption dataset is licensed under CC BY-4.0, which can be found under the Metadata URL here: World Bank Renewable Energy Consumption Metadata.

To adhere to the license requirements, I will:

- Ensure that proper credit is given to the original data providers

- Include a clear statement of the CC BY-NC 4.0 / CC BY-4.0 license in all work;
- Indicate adaptations, changes and ensure that the original data source is acknowledged;
- Ensure that the CO_2 emission data is only used for this project which is study-related, non-commercial purposes

The overview of data sources can be summarized in the table below

	CO_2 Emission Dataset	**Renewable Energy Consumption Dataset**
Geographical Coverage	266 countries and associations	266 countries and associations
Temporal Coverage	1990 - 2020	1990 - 2021
License	CC BY-NC 4.0	CC BY-4.0
Source	Climate Watch – Historical GHG Emissions (1990–2020) (2023, Washington, DC: World Resources Institute)	IEA, IRENA, UNSD, World Bank, and WHO (2023 Tracking SDG 7: The Energy Progress Report, Washington DC)

III. Data Pipeline

The Python-based data pipeline was constructed with support from the Pandas library. This pipeline for the project consists of several key stages:

1. **Data Extraction**
   - The raw data on CO_2 emissions and renewable energy consumption were downloaded from the World Bank website in ZIP format.
   - The archive is then extracted, and the CSV data files are collected and loaded into Pandas data frame in Python environment.
   - An obstacle encountered during CSV extraction is the variability in the data file's filename. The data file's name isn't fixed but ends with a timestamp reflecting the download time of the archives. Thus, to extract the desired file, it is required to add logic to identify files beginning with specific prefixes before loading.
   - Additionally, the CSV files come with metadata rows, which need to be eliminated when using Pandas to read the data.
   - The dataset configurations are stored in a Python dictionary that includes the URL, prefix file name, number of rows to skip, and table name.

2. **Data Cleaning**
   - While the data's temporal coverage is from 1990 – 2020/2021, the CSV files contain data from 1960 – 2023. Nevertheless, all data falling outside of the specified coverage range is empty and can be safely removed from the data frame. The year periods are then set to 1990 – 2020 for both datasets to ensure consistency between them.
   - Some of the countries do not have any values. They got excluded from the data frame.
     - CO2 emission dataset: We started with 266 countries and associations and ended up with 239 countries and associations.
     - Renewable energy consumption dataset: We started with 266 countries and associations and ended up with 260 countries and associations.
     - There's a difference between the countries in the two datasets, so only the common countries are considered. This leaves us with 239 countries.
   - Additional missing values:
     - In the CO_2 emission dataset, there was one missing data point for Namibia (1990). This was filled with the emission value from 1991, which is the closest available value.
     - In the renewable energy dataset, there were several missing data points for different countries and time periods e.g. Cambodia (1990–1994), Eritrea (1990–1991), Namibia (1990), Montenegro (1990–2004), South Sudan (1990–2011), and Suriname (1990–1999). For those countries, where

there was a prolonged period without data, it's possible they did not participate in the survey during those times. In such cases, missing values were also filled with the first available value.

3. **Data Loading**
   - The two data frames are then saved into 2 tables and stored in an SQLite Database.

The high-level overview of the data pipeline can be described as in Figure 1.

During the execution of the pipeline, any encountered errors will be caught by the try-except block and subsequently logged.

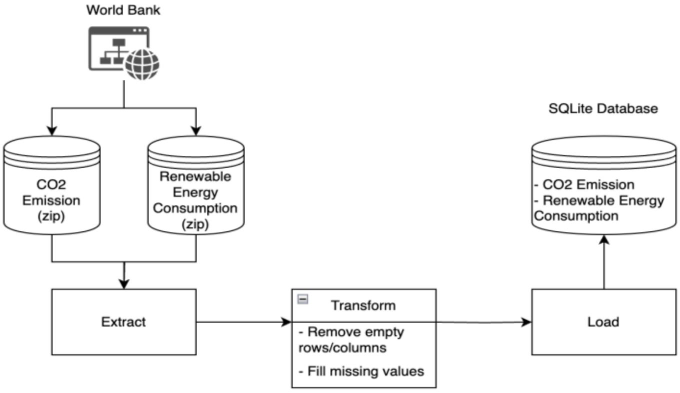


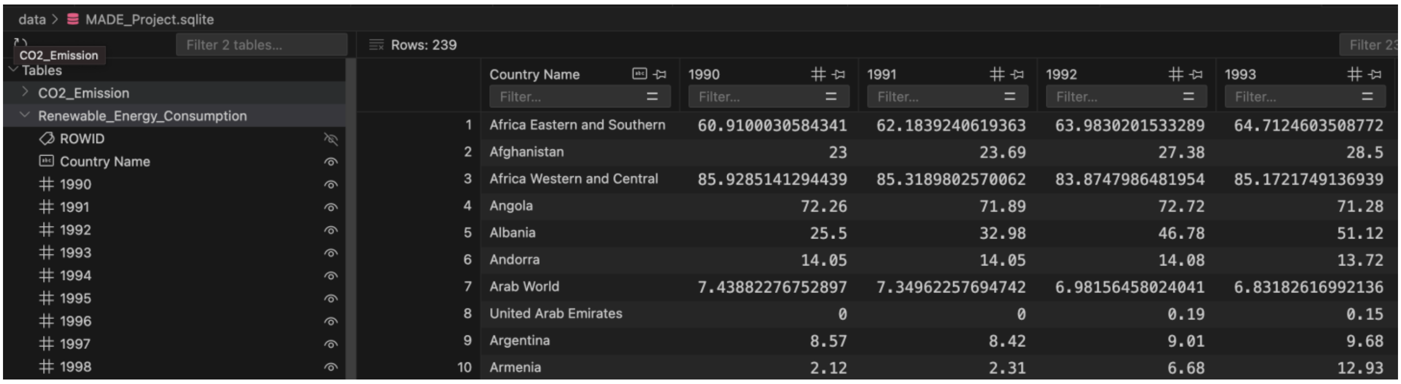
Figure 1. High-level overview of the Data Pipeline

IV. Results and Limitations

1. Results

Given the already high quality of the original datasets, the cleaning step serves to further enhance and guarantee their integrity and consistency. As a result, the data pipeline generates refined and processed output datasets.

The output data is saved within 2 tables in a relational database file, specifically in SQLite format. Rows correspond to individual countries or associations, while columns represent different years. Each cell within the table contains emissions or consumption value. This general-purpose format allows easy interpretation, sharing as well as compatibility with many analysis tools. A quick glimpse of the database can be seen in Figure 2.



Rows: 239

	Country Name	1990	1991	1992	1993
1	Africa Eastern and Southern	60.9100030584341	62.1839240619363	63.9830201533289	64.7124603508772
2	Afghanistan	23	23.69	27.38	28.5
3	Africa Western and Central	85.9285141294439	85.3189802570062	83.8747986481954	85.1721749136939
4	Angola	72.26	71.89	72.72	71.28
5	Albania	25.5	32.98	46.78	51.12
6	Andorra	14.05	14.05	14.08	13.72
7	Arab World	7.43882276752897	7.34962257694742	6.98156458024041	6.83182616992136
8	United Arab Emirates	0	0	0.19	0.15
9	Argentina	8.57	8.42	9.01	9.68
10	Armenia	2.12	2.31	6.68	12.93

Figure 2. Sample database result

2. Limitations

One potential issue is erroneous data in the original sources, such as biased reporting due to political or economic reasons and inaccuracies in measured values. Additionally, imputed values may also be incorrect. Another potential limitation comes from variations in data sources. Inaccuracy and inconsistency in data collection methods may affect the validity, reliability and comparability of the analysis.

Also, renewable energy consumption is a relative value which may not accurately reflect absolute levels of energy usage, potentially obscuring the actual scale of renewable energy adoption in different countries. During the initial phases of the project, a dataset on Renewable Energy Output was introduced to analyze the adoption level of renewable energy adoption. Unfortunately, due to licensing restrictions, the dataset could not be utilized.

Concerning the final outcome, there may be no correlation observed between CO2 emissions and renewable energy consumption.