1 Question

How do changes in CO2 emissions affect global temperature anomalies?

2 Data Sources

2.1 Description of Data Sources.

Data source 1: NASA GISS Surface Temperature Analysis (GISTEMP)

- Data URL: https://data.giss.nasa.gov/gistemp/tabledata_v4/GLB.Ts+dSST.csv
- Data Type: CSV
- This data set addresses the problem of global surface temperature measurement and analysis. It provides a comprehensive record of temperature anomalies over time, which is crucial for understanding trends in global warming and climate change.

Datasource2: Our World in Data CO2 dataset

- Data URL: https://github.com/owid/co2-data/raw/master/owid-co2-data.csv
- Data Type: CSV
- This data set addresses the problem of tracking global carbon dioxide emissions. It provides detailed records of CO2 emissions from different countries and regions, which is critical for understanding the sources of greenhouse gasses contributing to climate change.

2.2 Data Structure and Quality:

- **NASA GISS Surface Temperature Analysis (GISTEMP):** The data is structured in a tabular CSV format. It is a global time series data with monthly frequency, where there are year and months columns. This data shows temperature change of the world in degree- celsius from 1880 to 2024. Column J-D shows the temperature change from January to December for each year. The data contains missing values till 1950, accurate, it is up-to-date and it reflects the real world

```
Downloaded data from https://data.giss.nasa.gov/gistemp/tabledata_v4/GLB.Ts+dSST.csv:
   Year   Jan   Feb   Mar   Apr   May  ...   Aug   Sep   Oct   Nov   Dec   J-D
0  1950 -0.26 -0.27 -0.07 -0.21  -.11  ...  -.16  -.11  -.20  -.34  -.21  -.17
1  1951 -0.34 -0.41 -0.20 -0.14   .00  ...   .06   .05   .08  -.01   .16  -.07
2  1952  0.11  0.11 -0.08  0.03  -.03  ...   .05   .07   .00  -.13  -.02   .01
3  1953  0.07  0.15  0.11  0.19   .11  ...   .05   .05   .08  -.03   .05   .08
4  1954 -0.24 -0.10 -0.15 -0.14  -.20  ...  -.18  -.10  -.02   .08  -.18  -.13

[5 rows x 14 columns]
```

Figure 1: First five rows of global surface temperature changes.

- **Our World in Data CO2 dataset:** The data is structured in tabular csv format. It is the global time series data. It shows the CO2 emission of the countries by each year from 1850 to 2024. There are different types of CO2 emissions. The indicator value is expressed in billion tons. The data contains missing values till 1950, accurate, up-to-date and it reflects the real world.

```
Downloaded data from https://github.com/owid/co2-data/raw/master/owid-co2-data.csv:
       country  year  cement_co2  ...  gas_co2  oil_co2  share_global_co2
0  Afghanistan  1950         0.0  ...      0.0    0.063             0.001
1  Afghanistan  1951         0.0  ...      0.0    0.066             0.001
2  Afghanistan  1952         0.0  ...      0.0    0.060             0.001
3  Afghanistan  1953         0.0  ...      0.0    0.068             0.002
4  Afghanistan  1954         0.0  ...      0.0    0.064             0.002

[5 rows x 9 columns]
```

Figure 2: First five rows of carbon dioxide emissions by countries

2.3 Licenses:

Our World in Data CO2 Dataset: All data, visualizations and articles produced by Our World in Data are open access under the Creative Commons by license. Detail information about license can be found here.

NASA GISS Surface Temperature Analysis (GISTEMP): This database is made available under the public domain dedication and license v1.0 . Detailed information about license can be found here.

I am planning to use them for educational project only.

3. Data Pipeline

The data pipeline is designed to download climate-related datasets from provided URLs, clean and transform the data, and then store it into SQLite databases. This pipeline is implemented using Python, using libraries such as pandas, os, urllib, io and sqlite3 for database interactions. **Transformation and Cleaning Steps:** Downloading Data: Data is downloaded from the given URLs.Data Transformation: **First Dataset**: Used the second row as the header.
Filtered out rows before the year 1950. Dropped unnecessary columns ('D-N', 'DJF', 'MAM', 'JJA', 'SON'). **Second Dataset:** Filtered out rows before the year 1950.
Kept only the relevant columns ('country', 'year', 'cement_co2', 'co2', 'co2_growth_prct', 'coal_co2', 'gas_co2', 'oil_co2', 'share_global_co2').

4. Results and Limitations

The output data of the climate data pipeline consists of 2 SQLite database files containing cleaned and transformed data.

4.1 Data Structure: **First Dataset**: have columns related to global temperature anomalies over time, starting from the year 1950. Irrelevant columns have been dropped. **Second Dataset**: have columns related to CO2 emissions by country and year, starting from 1950. Only relevant columns are kept.

4.2 Data Quality: **Accuracy**: The data is obtained from reputable organizations (NASA and OWID), ensuring a high accuracy.**Completeness**: The data is filtered to include years from 1950 onwards, ensuring completeness.**Consistency**: The data has been processed to ensure consistent formats, such as numeric values for years.**Timeliness**: The datasets are up-to-date as of the latest available data from the sources.**Relevancy**: The selected columns and years ensure that the data is relevant to our project.

4.3 Output Data Format: SQLite Database: The format for the output data is SQLite. This format is ideal because it is portable and compatible.

4.4 Potential Issues: Data Completeness: the datasets are comprehensive within their scope, there may still be gaps or missing values that could impact analysis. Ensuring completeness in the initial download and cleaning steps helps solve it, but some limitations may persist. Data Correctness: The data sources are reputable, but it's always important to consider potential biases or errors introduced during data collection or processing stages.