1 Question

How does the Impact of Diverse National Emissions on Global Temperature Trends Over Time?

2 Data Sources

2.1 Description of Data Sources

- **Dataset 1: Emissions by Country Dataset**

  This dataset provides an in-depth look into the global CO_2 emissions at the country-level, allowing for a better understanding of how much each country contributes to the global cumulative human impact on climate.

- **Dataset 2: All Countries Temperature Statistics 1970-2021.**

  This dataset provides information on changes in global surface temperature across all countries from 1970 to 2021. It includes data on temperature variations over a 51-year period and is based on information from various sources, including weather stations, satellites, and ocean buoys.

2.2 Data Structure and Quality

The dataset "Global Fossil CO2 Emissions by Country (2002-2022)" contains the following columns:

- **Country:** The name of the country.
- **Year:** The specific year of the recorded data.
- **Emissions:** CO2 emissions in metric tons.
- **Per Capita Emissions:** CO2 emissions per capita.
- **GDP:** Gross Domestic Product of the country.
- **Population:** The population of the country for the given year.
- **Emission sources:** Breakdown of emissions by source, such as coal, oil, gas, etc.

  The dataset covers a wide range of countries and years (2002-2022). Some countries or years might have missing data due to lack of reporting.

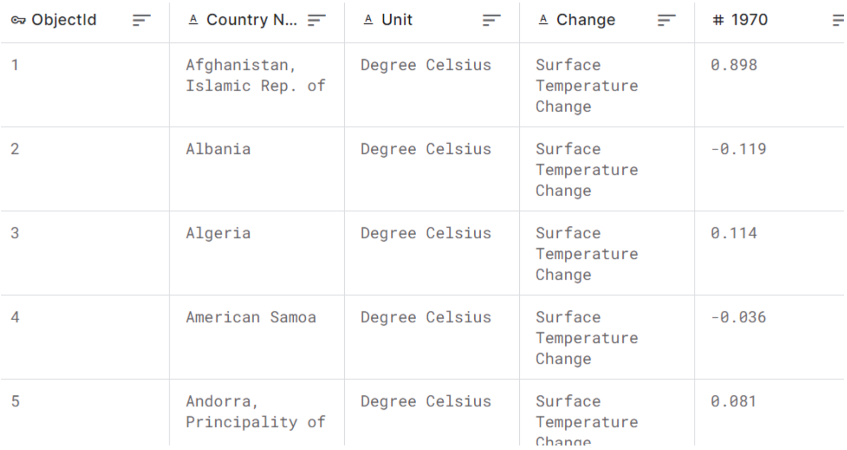
  Standardized units (e.g., metric tons for emissions, USD for GDP) ensure uniformity. Consistent formatting across years and countries facilitates comparative analysis. Country-level data provides geographic specificity, aiding in comparative studies.

The dataset "All Countries Temperature Statistics (1970-2021)" on Kaggle features the following data structure and quality aspects:

- **Country:** The name of the country.
- **Year:** The specific year of the recorded data.
- **Average Temperature:** The average temperature for the year.
- **Minimum Temperature:** The minimum recorded temperature for the year.
- **Maximum Temperature:** The maximum recorded temperature for the year.

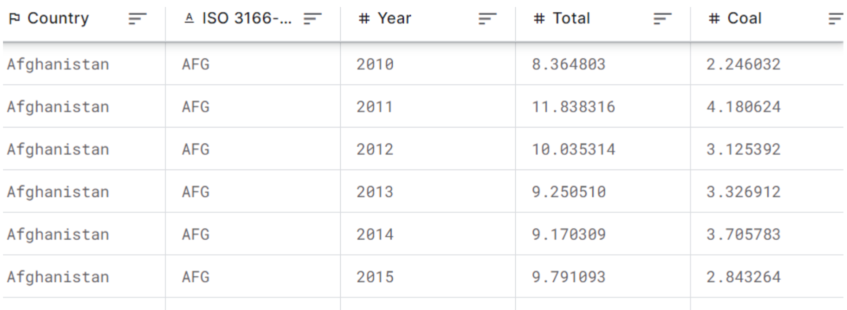
- **Temperature Anomaly:** Deviations from a baseline temperature, indicating climate change.

  Standardized temperature units (Celsius) and consistent formatting facilitate comparative analysis. Consistent measurement methodologies across years and countries.



ObjectId	Country N...	Unit	Change	1970
1	Afghanistan, Islamic Rep. of	Degree Celsius	Surface Temperature Change	0.898
2	Albania	Degree Celsius	Surface Temperature Change	-0.119
3	Algeria	Degree Celsius	Surface Temperature Change	0.114
4	American Samoa	Degree Celsius	Surface Temperature Change	-0.036
5	Andorra, Principality of	Degree Celsius	Surface Temperature Change	0.081

Figure 1: First 5 rows of annual surface temperature change dataset.



Country	ISO 3166-...	Year	Total	Coal
Afghanistan	AFG	2010	8.364803	2.246032
Afghanistan	AFG	2011	11.838316	4.180624
Afghanistan	AFG	2012	10.035314	3.125392
Afghanistan	AFG	2013	9.250510	3.326912
Afghanistan	AFG	2014	9.170309	3.705783
Afghanistan	AFG	2015	9.791093	2.843264

Figure 2: First 5 rows of emissions by the countries over the years.

2.3 Licenses and Permissions

The data sources are publicly available on Kaggle under open-data licenses CC0:Public Domain. Detailed license information can be found at: CC0

3 Data Pipeline

The data pipeline has three main modules: extractor, transform, and loader. Each of the modules has their respective functions. First `extract_csv` from extractor module is used to extract the data source from URL, then `delete_columns` from transform module deletes the list of useless columns specified for every dataset, once all the transformations have been applied, dataset is then loaded to sqlite database using `load_df_to_sqlite` from loader module.

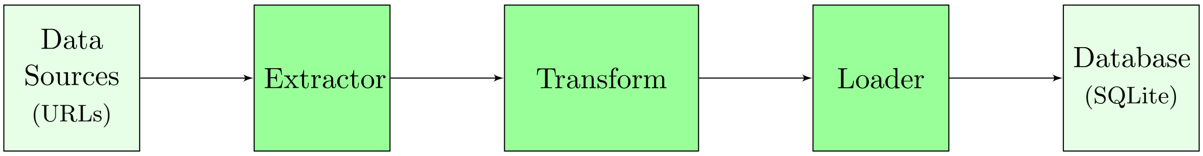


Figure 3: ETL Pipeline Diagram

4 Result and Limitations

Output datasets of the pipeline for all data sources are stored in sqlite database as tables as it was faster and easier to handle as a collective database, The pipeline is coded in a way that data quality dimensions were of the upmost priority and that the output datasets of the pipeline

- contain all necessary information which is required to answer selected questions
- presentation of the datasets aligns with the requirements of the questions need to be answered
- reflect the real word and are correct indicators
- are consistent in their formats

Annual surface temperature change and emissions country wise indicator can be compared and checked for correlation. The limitation is temporal mismatch i.e in the emissions dataset where the year column has too ancient values that are useless to be mapped on the temperature change dataset. Also, Emissions data is annual and country-specific, while temperature data might also be annual but could lack fine-grained temporal detail.