1 Question

How does the change in forest areas affect the global climate?

2 Data Sources

2.1 Description of Data Sources

- **Dataset 1: World Forest Area (km and %)**

  This dataset contains information about the world's forest area changes since 1990. Studying global forest area changes is crucial for assessing environmental health, informing conservation strategies, and understanding the impact of human activities on biodiversity and climate regulation.

- **Dataset 2: Climate Insights Dataset**

  This dataset provides valuable insights into the ongoing changes in our climate. It encompasses a comprehensive collection of temperature records, CO2 emissions data, and sea level rise measurements.

2.2 Data Structure and Quality

The dataset "World Forest Area (km and %)" contains the following columns:

- **Country Name:** The name of the country.
- **Year:** The specific year of the recorded data.
- **Forest Area (km):** The total forest area in hectares.
- **Forest Area (% of land area):** The percentage of the total land area that is covered by forests.

  The dataset covers a wide range of countries and years (1990-2021). Some countries or years might have missing data due to lack of reporting.

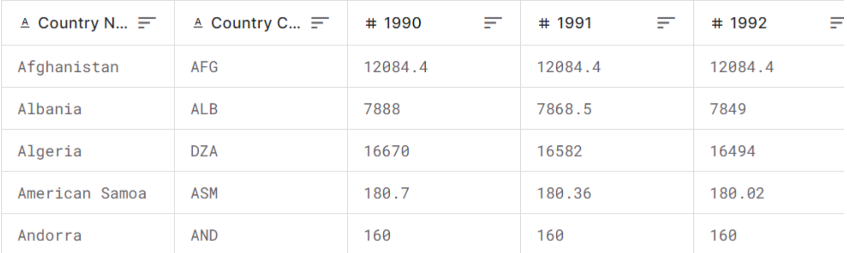
  Standardized units (e.g., kilometer square for forest area) ensure uniformity. Consistent formatting across years and countries facilitates comparative analysis. Country-level data provides geographic specificity, aiding in comparative studies.

The "Climate Insights Dataset" on Kaggle contains multiple files, with the primary file being climate$_hangedata.csv.Thestructureofthisfileincludes$:

- **Year:** The specific year of the recorded data.
- **Country:**The name of the country.
- **CO2 Emissions:** CO2 emissions in metric tons.
- **Temperature Anomalies:** Deviations from a baseline temperature, indicating climate change.
- **GDP:** Gross Domestic Product of the country.
- **Population:** The population of the country for the given year.

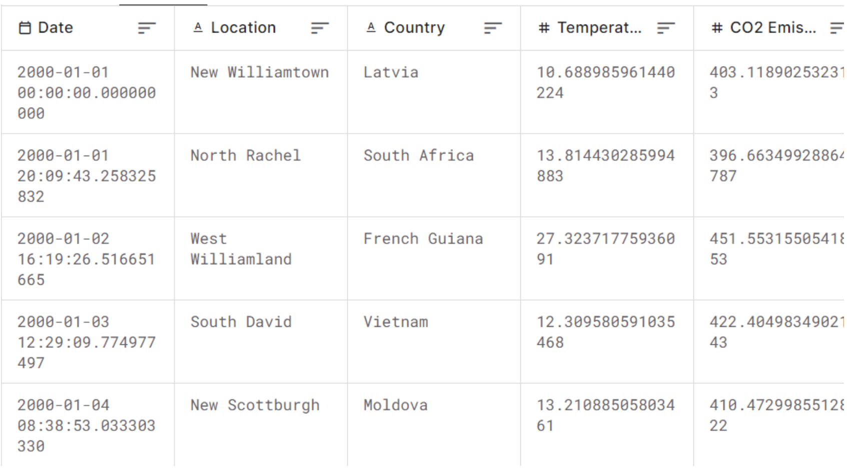
- **Energy Consumption:** Energy consumption data.

  Data is likely sourced from reputable global environmental agencies and institutions. Ensuring cross-references with authoritative sources can enhance reliability.Standardized units (e.g., metric tons for emissions, degrees Celsius for temperature anomalies) ensure uniformity. Consistent formatting across years and countries facilitates comparative analysis



Country N...	Country C...	# 1990	# 1991	# 1992
Afghanistan	AFG	12084.4	12084.4	12084.4
Albania	ALB	7888	7868.5	7849
Algeria	DZA	16670	16582	16494
American Samoa	ASM	180.7	180.36	180.02
Andorra	AND	160	160	160

Figure 1: First 5 rows of the forest area by km dataset



Date	Location	Country	# Temperat...	# CO2 Emis...
2000-01-01 00:00:00.000000000	New Williamtown	Latvia	10.688985961440224	403.118902532313
2000-01-01 20:09:43.258325832	North Rachel	South Africa	13.814430285994883	396.66349928864787
2000-01-02 16:19:26.516651665	West Williamland	French Guiana	27.32371775936091	451.5531550541853
2000-01-03 12:29:09.774977497	South David	Vietnam	12.309580591035468	422.4049834902143
2000-01-04 08:38:53.033303330	New Scottburgh	Moldova	13.21088505803461	410.4729985512822

Figure 2: First 5 rows of climate insights dataset.

2.3 Licenses and Permissions

The data sources are publicly available on Kaggle under open-data licenses CC0: Public Domain and CC by 4.0: Creative Commons Attribution 4.0. Detailed license information can be found at: CC0 and CC by 4.0

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.

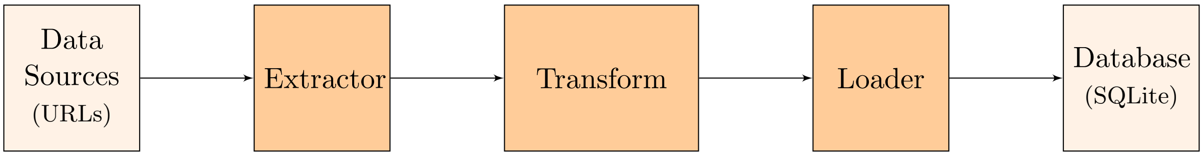


Figure 3: ETL Pipeline Diagram

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:

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

Climate insights and forest area indicator can be compared and checked for correlation and similarly the other two datasets can be compared. The only limitation is that the incomplete or missing data in either dataset can hamper analysis. Since, Both datasets may have missing values or incomplete records for certain years or regions, it can affect the analysis.