1 Question

How does the change in forest areas affect the global temperatures on a yearly basis?

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: 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 "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 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.

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

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 2: First 5 rows of annual surface temperature change 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.

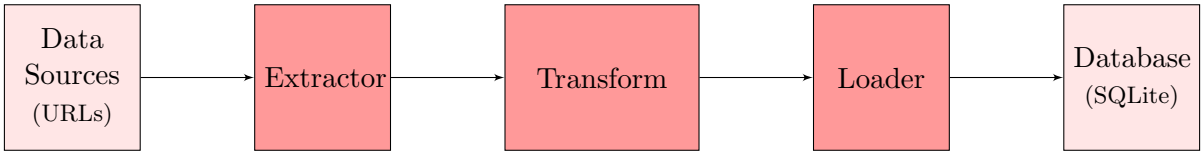

Figure 3: ETL Pipeline Diagram

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

Annual surface temperature change 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 have significant gaps for certain years or regions, it can lead to biased or inaccurate conclusions.