Unveiling the Climate Crisis: A Data Science Investigation of Climate Related Issues (CO_2 emissions and climate related disasters)

Introduction:

This project investigates the relationship between global warming trends (specifically CO_2 emissions) and the increase in climate-related disasters, analyzed on a yearly and country-wise basis. Utilizing data science techniques, including statistical and machine learning methods, we examine the impact of various factors (e.g., greenhouse gas emissions, land use changes, population growth) on the observed increase in disasters.

Main Question

How did global warming accelerate, what are the associated climate-related disasters, and how can it be mitigated?

Data Sources

CO_2 Emission by Countries Year Wise (1750-2022): Selected for its long-term, country-level detail on CO_2 emissions, crucial for understanding historical trends.
Climate Change Data (IMF): Chosen for its focus on the relationship between climate change and natural disasters, aligning with our project's core question.

Origin & Content:
CO_2 Emissions: The selected Kaggle dataset provides a comprehensive record of annual CO_2 emissions by country from 1750 to 2022, including details on emissions, population, area, and density, chosen due to the lack of comparable long-term data on the platform.

Climate Disaster Data: IMF dataset, sourced from a variety of climate change literature, includes metrics on natural disasters, potentially linked to climate indicators like temperature or sea level. The reason for selecting this dataset is that it takes various verified sources as a reference and used to update yearly. The links between climate change and natural disasters are well documented in a wide variety of climate change literature. The dataset covers disasters that killed ten (10) or more people, affected hundred (100) or more people, led to declaration of a state of emergency and led to call for international assistance.

Structure & Quality:

Data Source	Structure	Quality	Notes
CO_2 Emissions	CSV	High: Consistent, well-structured	Potential for gaps in early or less-developed countries
Climate Change Data (IMF)	CSV	Medium: Variable, dependent on source literature	Need to carefully assess methodology of included studies

Licenses & Obligations:

Both datasets appear to be available for open use (Kaggle dataset requires attribution). We will:

Attribution: Clearly cite both sources in our project.
Kaggle:(CC0: Public Domain) Per the author, the dataset is intended to help researchers and environemnt experts to predict about the global warming.
IMF: Per IMF, the data is made available by EM-DAT, CRED / UCLouvain, Brussels, Belgium. Regarding the liscense permission, we are allowed to download, create link to the IMF website for noncommercial usage only, without any right to resell or redistribute or to compile or create derivative works.

Data Pipeline

In order to build a robust pipeline, we are using Python programming language. The major steps our pipeline follows are as follows:

1. **Ingestion:**
    - Python scripts load CSV data from Kaggle (CO_2 emissions) and the IMF platform (disaster data).
2. **Cleaning & Transformation:**
    - **Both Datasets:**
        - Unnecessary columns are dropped and column names are made more meaningful. Date formats and country codes are standardized. Missing values are handled:
            - Some rows with entirely missing data are removed.
            - Missing values in specific columns (e.g., disaster numbers) are filled with '0.0' as it doesn't affect the analysis.
            - Relevant substrings are extracted for focused analysis.
    - **CO_2 Emissions:**
        - Data is aggregated to regional or global levels for broader insights.
    - **Filtering by Date:**
        - We have CO_2 data from 1750, but disaster data only from 1980 onwards. The data is filtered to include only dates after 1980 for consistency.

3. **Joining:**
    - The cleaned datasets are merged based on country/region code, creating a combined table with emissions and disaster data per year.
4. **Feature Engineering:**
    - Additional features are calculated, such as rolling averages of CO_2 emissions to smooth out yearly fluctuations, trends in disaster occurrences to identify patterns over time, climate change indicators derived from scientific literature etc.
5. **Output:**
    - The processed and enriched data is stored as an SQLite database file for efficient analysis and potential modeling in later stages.

Problems Encountered:

- Kaggle Dataset authentication: For ease of use and centralized configuration, Kaggle credentials were stored as an environment variable, and a single config file managed parameters for the data pipeline's loading, cleaning, and subsequent stages.
- Missing country codes in kaggle data set: A dictionary was used to fill in missing country codes in the dataset, ensuring standardization and consistency with the two-digit format.
- Data Availability: In climate related dataset, we have the data only after 1980. Therefore, we can only consider the data after 1980 for both datasets.
- Data Type Conversion: The presence of special characters and symbols in certain columns posed a challenge for accurate data type conversion, necessitating careful selection and extraction of relevant data subsets to ensure meaningful analysis.

Result and Limitations:

- Data Output: The data pipeline generates two output files: a CSV for sharing and initial review, and an SQLite database for efficient analysis and correlation exploration, both containing only relevant, filtered data.
- Output preferance file (sqlite): SQLite is ideal for data analysis tasks due to its fast read performance, standard SQL language support, and convenient Python interface through the sqlite3 module, making it easy to query and manipulate data directly within your analysis code.
- Potential Issues: Potential challenges remain despite a successful data pipeline implementation. Removing rows with null values could introduce bias and data loss. Inherent biases in data sources, collection methods, and underlying studies may affect results. While correlation between CO_2 emissions and climate disasters might be observed, establishing causality is complex due to potential confounding factors, requiring further exploration and consideration of broader climate science contexts for accurate interpretation.