Exploring the Association Between Climate Change Indicators and Air Pollution Levels

1 Introduction

Climate change and air pollution, both products of human activity, represent urgent global challenges. Climate change brings rising temperatures, altered weather patterns, and increased extreme events, threatening ecosystems and human well-being. Meanwhile, air pollution, stemming from industrial processes and transportation, poses significant health risks and ecological harm. In light of these intertwined challenges, this project seeks to explore the correlation between climate change and air pollution. By unraveling the complex interplay between these critical environmental factors, we aim to inform evidence-based strategies for mitigating their adverse impacts on human health, ecosystems, and the planet's resilience.

1.1 Main Questions

1. How have temperature and air pollution levels changed over the years?
2. Is there a correlation between key air pollutant levels (PM10, PM2.5, NO2) and temperature changes?

2 Data Sources

2.1 FAOSTAT Climate Change Data

- Metadata URL: https://www.fao.org/faostat/en/#data/ET/metadata
- Data URL: https://fenixservices.fao.org/faostat/static/bulkdownloads/Environment_Temperature_change_E_All_Data.zip
- Data Type: CSV
- Licensing: CC BY-NC-SA 3.0 IGO

This dataset provides historical records of temperature changes and is structured as a CSV file with columns for country names, months, years, and temperature changes. The data is high quality, provided by a reputable source (FAO), and covers a broad temporal range. However, the dataset may contain missing values and duplicated columns for certain years, which require preprocessing.
It comes under the license "CC BY-NC-SA 3.0 IGO," which allows for the data to be used, shared, and adapted for non-commercial purposes, provided appropriate credit is given. To comply, my project includes proper citations and links to the original data source, and it will be ensured that the data is used for educational purposes only.

2.2 WHO Ambient Air Quality Data

- Metadata URL: https://cdn.who.int/media/docs/default-source/air-pollution-documents/air-quality-and-health/who_ambient_air_quality_database_version_2024_(v6.1).xlsx?sfvrsn=c504c0cd_3&download=true
- Data URL: https://cdn.who.int/media/docs/default-source/air-pollution-documents/air-quality-and-health/who_ambient_air_quality_database_version_2024_(v6.1).xlsx?sfvrsn=c504c0cd_3&download=true
- Data Type: Excel
- Licensing: Open Data

This dataset contains information on air pollution indicators, including particulate matter (PM10, PM2.5) and nitrogen dioxide (NO2), and is structured as an Excel file with multiple sheets, which includes metadata as well. The data is high quality, provided by a reputable source (WHO), and covers a wide range of countries and cities.
WHO data can be used for research and educational purposes, provided proper attribution is given, and the data is not used for commercial purposes. To comply, my project includes proper citations and links to the original data source, ensuring that the data is used solely for educational purposes, in line with WHO's licensing terms.

3 Data Pipeline

The data pipeline, implemented in Python, automates the process of data extraction, transformation, and loading (ETL). It consists of several stages

- Data Extraction: Downloads temperature data which is a csv file in Zip format from FAO and air quality data which is an excel file with multiple sheets from WHO extracting only the sheet related to data.
- Data Transformation: Cleans and processes the data to align formats and performs necessary calculations.
- Data Loading: Saves the processed data into an SQLite database for further analysis.

3.1 Transformation Steps

Since both datasets have missing values that could skew the analysis, the missing values are dropped to ensure data integrity.
FAOSTAT Climate Change includes years as different columns; therefore, the dataset is reshaped to store them as one column called "year". Since the dataset includes temperature changes for each month of the years, data is aggregated (mean) on a yearly basis to ensure consistency with other datasets.
WHO Air Quality dataset includes pollutant concentrations in different cities in a country, so the data is aggregated (mean) on a country level to ensure consistency.

Dataset	Transformations
FAOSTAT Climate Change/WHO Air Quality	- Rename columns to ensure consistency across datasets. - Drop unnecessary columns from the datasets. - Round numeric columns (pollutant concentrations and temperature changes) to two decimals. - Drop missing values.
FAOSTAT Climate Change	- Filter temperature dataset to include only data related to temperature change. - Reshape temperature dataset to store years as a single column called "Year". - Aggregate temperature change data on a yearly basis by calculating the mean temperature change.
WHO Air Quality	- Filter air quality dataset to include data for only selected European countries. - Aggregate air quality data on a country level by calculating the mean pollutant concentrations. - Drop missing values from the air quality dataset to ensure data integrity.

3.2 Error Handling and Dynamic Input

The pipeline includes try-except blocks to handle potential errors such as network issues during data downloading or database connection errors. This ensures that the pipeline can fail gracefully and provide informative error messages. It is designed to be flexible with changing input data by using dynamic filtering and transformation steps. If the structure of the source data changes, the pipeline can be adjusted with minimal modifications to accommodate the new structure.

4 Result and Limitations

The output data of the pipeline is stored in a SQLite database. We chose SQLite as the output format due to its simplicity, portability, and compatibility with Python. Storing the data in a relational database facilitates efficient data retrieval and manipulation for subsequent analysis.

The temperature dataset covers the years from 1960 to 2022, while the WHO air quality dataset only spans from 2013 to 2022. This inconsistency in the range of years may pose challenges in interpreting trends and limit the comprehensiveness of the analysis. While the data sources are reputable, the accuracy of measurements and reporting may vary between countries and over time. It's essential to acknowledge potential inaccuracies and limitations when comparing between different countries.