Exploring the Association Between Climate Change Indicators and Air Pollution Levels

1 Introduction

Climate change and air pollution are critical environmental issues, both driven by human activity and significantly impacting public health and ecosystems. This project investigates the correlation between these two factors specifically in European countries, where industrial activities and dense urban areas contribute to notable pollution levels and climate variations.

1.1 Main Questions

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

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

This dataset provides historical records of monthly temperature changes, structured as a CSV file with columns for country names, months, years, and temperature changes. The data, sourced from the reputable Food and Agriculture Organization (FAO), is of high quality and covers an extensive temporal range. However, the dataset may include missing values and duplicated columns for years, which require preprocessing.

Licensing: The dataset is licensed under "CC BY-NC-SA 3.0 IGO," allowing for use, sharing, and adaptation for non-commercial purposes with appropriate credit. To comply with this license, the project will include proper citations and links to the original data source, and the data will be used solely for educational purposes.

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

This dataset compiles data on ground measurements of annual mean concentrations of nitrogen dioxide (NO2) and particulate matter(PM10, PM2.5), which aim at representing an average for the cities. Both groups of pollutants originate mainly from human activities related to fossil fuel combustion.
It is structured as an Excel file with multiple sheets including data and metadat, with columns for country codes, country names, city names, years, pollutant concentrations, and types of measurement stations. The data provided by the reputable World Health Organization (WHO), is of high quality. However, it contains missing values for some of the countires in some years, which require preprocessing.

Licensing: The dataset is licensed under is licensecd under "CC BY-NC-SA 3.0 IGO" which allows users to freely copy for non-commercial purposes, provided WHO is acknowledged as the source. To comply, the project include 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 functions:

- **Data Extraction**: Downloads datasets from FAO and WHO using the requests library. FAOSTAT Climate Change dataset is downloaded as a ZIP file, which is extracted using the zipfile library, where the first CSV file inside the ZIP archive is loaded into a pandas DataFrame. WHO Air Quality dataset is obtained as an Excel file. The relevant sheet and columns are directly read into a pandas DataFrame using the pandas library with the openpyxl engine.

- **Data Transformation**: Cleans and processes the data to align formats and performs necessary calculations.

- **Data Storage**: 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 dataset includes years as separate columns, so it is reshaped to store years as one column.

-WHO Air Quality dataset includes pollutant concentrations for different cities within each country. The data is aggregated by calculating the mean pollutant concentrations for each country 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. - Filter to include data for only selected European countries. - 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
WHO Air Quality	- Aggregate air quality data on country level.

3.2 Error Handling and Dynamic Input

The pipeline includes basic error handling mechanisms using try-except blocks to manage potential issues, such as network problems during data download or errors while connecting to the database which ensures it can provide informative error messages.
However, the pipeline currently does not fully support automatic adaptation to changes in the data source structure. Adjustments to the pipeline code would be required if there are significant changes in the structure of the input data, such as new columns or different formats.

4 Result and Limitations

4.1 Output

The output data of the pipeline is stored as two tables in a SQLite database. Storing the data in a relational database facilitates efficient data retrieval and manipulation for subsequent analysis.

4.2 Limitations

Temporal Inconsistency: While the temperature dataset offers a long-term perspective (1960-2023), the air quality data is limited to recent years (2010-2022). The differing ranges of years between the two datasets may pose challenges in interpreting trends and limit the comprehensiveness of the analysis.

Data Accuracy: Despite the reputability of the data sources, the accuracy of measurements and reporting can vary between countries and over time. These potential inaccuracies and variations must be acknowledged when comparing data between different countries and periods.