1. Introduction

Understanding the relationship between air quality measures and climate dynamics in urban settings is essential. Urban areas, characterized by high population density, traffic, and industrial activities, are significant sources of air pollution. This pollution impacts public health and contributes to climate change. Analyzing variations in air pollutant levels and their correlation with climate change indicators provides critical insights into the effectiveness of environmental policies and urban planning strategies.

The primary research question addressed in this project is: "How do variations in air pollutant levels across different urban environments correlate with climate change indicators, and what predictive models can be developed to forecast these effects on urban air quality?" This project compares air pollutant levels, such as PM2.5, PM10, NO2, and O3, from Beijing and various European locations with climate indicators like temperature and humidity. The goal is to identify patterns and develop models that accurately predict future air quality based on these parameters, ultimately guiding better environmental policies and urban planning.

2. Data Sources

2.1. Beijing Multi-Site Air Quality Data

This dataset, sourced from the UCI Machine Learning Repository, provides detailed hourly measurements of pollutants like PM2.5, PM10, SO2, NO2, and O3 from multiple monitoring stations in Beijing from 2013 to 2017. It also includes meteorological data such as temperature and humidity.

Data Structure and Quality: Structured in CSV format, the data contains columns for various pollutants and meteorological parameters with timestamps. Missing values are marked as "NA." The dataset is comprehensive and well-documented.

License: Available under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Compliance is ensured by acknowledging the source in all project documentation and publications.

2.2. Inorganic Gases-2017

Sourced from the European Commission's Joint Research Centre (JRC), this dataset includes measurements of PM2.5, PM10, NO2, and O3 from various European locations, collected during 2017.

Data Structure and Quality: Also in CSV format, with columns for pollutants and timestamps. The data is consistently formatted and has minimal missing values.

License: Available under the European Union Open Data Portal license. Compliance is ensured by citing the source in all project outputs and publications.

Both datasets were selected for their relevance and high quality, enabling analysis of urban air quality and climate dynamics across different regions.

3. Data Pipeline

3.1. Overview
The data pipeline was implemented using Python to automate the downloading, extracting, cleaning, normalizing, and storing of data from multiple sources into an SQLite database. This process ensures a structured and efficient workflow, making it easier to handle large datasets and perform subsequent analysis. The pipeline utilized several libraries, including ‘requests’ for data download, ‘pandas’ for data manipulation, ‘sqlalchemy’ for database management, and ‘zipfile’ and ‘os’ for file handling. The processed data is stored in an SQLite database.

3.2. Download and Extraction
The first step involves downloading and extracting the data. The Beijing dataset is provided as a ZIP file, which is downloaded and extracted to obtain the CSV files. The European dataset is downloaded directly as a CSV file. Custom user-agent headers are used to ensure successful requests. Network issues are managed using try-except blocks, ensuring that the pipeline can continue running or retry downloads if necessary.

3.3. Data Processing
For accurate temporal analysis, separate date and time columns are combined into a single datetime column. This step implemented using pandas, ensures that the data is consistently formatted and can be easily integrated for time-series analysis. The Beijing dataset contains missing values denoted as "NA," which are replaced with pandas' NA values. Incomplete rows are then dropped. For the European dataset, rows with missing values are similarly dropped to ensure data integrity. Handling missing data is crucial for maintaining the quality and consistency of the datasets. To standardize the numerical columns, the data is normalized using pandas. This involves scaling the values based on their mean and standard deviation, which is essential for accurate analysis and model building. Normalization helps in comparing data from different sources on a common scale.

3.4. Saving Data to SQLite Database
Once the data is cleaned and normalized, it is saved into SQLite databases using sqlalchemy. Each dataset is stored in a table named after the file, facilitating structured storage and easy querying. This step ensures that the data is organized and readily accessible for further analysis.

3.5. Error Handling and Adaptability
The pipeline is designed to handle errors gracefully. Multiple try-except blocks are implemented to catch exceptions during data download, extraction, and processing. This ensures that the pipeline can adapt to changes in input data structure or content without crashing. The use of dynamic file processing allows the pipeline to handle varying datasets efficiently.

By implementing this automated pipeline, data from multiple sources is efficiently processed and stored, ready for analysis to understand the impact of air quality measures on climate dynamics in urban settings.

4. Results and Limitations

4.1. Output Data
The output of the data pipeline consists of cleaned and normalized datasets stored in SQLite databases. Each table within the database corresponds to a specific dataset and contains

columns for datetime, pollutant levels (such as PM2.5, PM10, NO2, O3), and meteorological data (such as temperature and humidity). This structured storage format ensures that the data is organized and easily accessible for analysis.

4.2. Data Structure and Quality

The output data is structured in a tabular format, with each row representing a unique datetime entry and columns representing various pollutant levels and meteorological parameters. This structure allows for efficient querying and analysis of time-series data. According to the data quality dimensions, the output data exhibits the following characteristics:

Accuracy: The data reflects accurate measurements of air pollutants and meteorological conditions, sourced from reliable monitoring stations.
Completeness: Missing values have been handled by replacing "NA" with pandas' NA and dropping incomplete rows, ensuring the datasets are as complete as possible.
Consistency: The data is consistently formatted, with standardized scales for numerical columns achieved through normalization.
Timeliness: The data covers relevant periods, with the Beijing dataset spanning 2013 to 2017 and the European dataset covering 2017.
Relevancy: The datasets are highly relevant to the research question, capturing critical pollutants and meteorological data necessary for the analysis.

4.3. Data Format

SQLite was chosen as the output format for its lightweight and efficient handling of structured data. This format facilitates easy integration with data analysis tools and scripts, allowing for efficient querying and manipulation of the data. SQLite databases are also portable and can be easily shared or integrated into various applications.

4.4. Critical Reflection on Data and Potential Issues

While the pipeline ensures high-quality data, several potential issues and limitations need to be considered:

Data Completeness: Despite handling missing values, some gaps remain due to dropped incomplete rows. This could affect the comprehensiveness of the analysis.
Geographic and Temporal Coverage: Differences in regulatory policies and geographic conditions between Beijing and European locations may introduce variability that requires careful interpretation. The datasets cover different periods, which might complicate direct comparisons.
Model Generalization: Predictive models developed from these datasets might not generalize well to other urban environments without similar datasets for validation. The specific environmental and regulatory contexts of Beijing and European cities might limit the applicability of the findings to other regions.
Initial Focus Shift: Initially, the project aimed to analyze the impact of agricultural practices on methane emissions and climate goals. However, restrictive licenses on relevant datasets necessitated a shift to urban air quality measures. This change, while ensuring compliance, also redirected the focus and may have introduced new complexities and considerations.

By addressing these limitations and continuously refining the data pipeline, the project aims to provide accurate and reliable insights into the impact of air quality measures on climate dynamics in urban settings.