Analyzing the Impact of Air Quality on Crop Production in Germany

Xinia Apchora

June 6, 2024

1 Main Question

This project examines the correlation between various air pollutants (e.g., CO, NO2, O3, PM10, PM2.5, SO2) and crop production metrics (e.g., area harvested, yield, production) in Germany for the years 2018-2022. Main questions includes How do various air pollutants impact crop production metrics in Germany from 2018 to 2022, and How do these recent data trends compare with historical data from 1961 to 2017?

2 Data Sources

2.1 World Air Quality Data 2024

Source: Kaggle - World Air Quality Data 2024
Description: This dataset provides comprehensive air quality data, including various pollutants (e.g CO, NO2, O3, PM10, PM2.5, SO2)
Data Structure and Quality: The dataset is in CSV format, with columns representing different air pollutants, their values, and timestamps. The quality is generally good, but some missing values and anomalies are expected.
License: Open-data license. Ensure compliance by citing the source and adhering to any data usage restrictions.

2.2 Production Crops and Livestock Products

Source: FAO - Production Crops and Livestock Products
Description: This dataset offers detailed crop production data for Germany, essential for analyzing long-term trends and correlations with air quality.
Data Structure and Quality: The dataset is in CSV format, with columns for country, crop type, production metrics, and years. The data is high-quality but includes missing values that need handling.
License: Open-data license. Ensure compliance by citing the source and adhering to any data usage restrictions.

3 Data Pipeline

3.1 Technology Used

The data pipeline for this project is implemented using Python, leveraging several libraries for data handling, analysis, and storage. Key libraries used include: pandas For data manipulation. requests for downloading data. sqlite3 for database operations. os for directory management. zipfile, urllib3 for handling HTTP requests and extracting zip files.

3.2 Pipeline Overview

The data pipeline consists of several stages:

Download datasets from Kaggle and FAO using HTTP requests. Extract data from zip files and read CSV files. To clean data, Handle missing values, filter relevant data (e.g., Germany-specific data),

and ensure consistent formatting. Store cleaned data in an SQLite database for structured access and analysis.

3.3 Challenges and Solutions

Several challenges were encountered during the pipeline implementation:

Handled by ensuring proper API authentication and error handling in the download functions. Implemented error handling to catch and resolve parsing issues in CSV files. Ensured consistent data types and formats across different datasets.

3.4 Error Handling and Adaptability

The pipeline includes mechanisms to handle errors related to non-numeric values and ensures adaptability to changes in input data by dynamically checking and processing files. Using fillna(0) to set NaN values as 0 and handle date format for last updated column in air_quality dataset.

4 Results and Limitations

4.1 Output Data

The output of the data pipeline consists of cleaned datasets for air quality and crop production stored in an SQLite database. The cleaned datasets are stored in tables named `air_quality` and `crop_production`.

4.2 Data Quality and Structure

The cleaned datasets have high quality with no missing values and consistent formats. The data is stored in an SQLite database, which allows for efficient querying and manipulation.

4.3 Data Format

The chosen format for storing the output data is an SQLite database. This format was selected for its ability to handle large datasets efficiently and ease of integration with analysis tools.

4.4 Critical Reflection

Potential issues include data gaps for certain years, challenges in ensuring data consistency across multiple datasets. Replacing missing values with 0 might not always be appropriate. This can lead to inaccuracies, especially if missing values represent different meanings in different contexts The findings from this project are specific to Germany and may not be generalizable to other regions or countries. Different environmental conditions, agricultural practices, and pollution levels in other areas may lead to different results. Regular updates to the data pipeline will be necessary to accommodate new data and changes in data structure.

5 Tables

	CO	NO2	O3	PM10	PM2.5	Area Harvested	Yield	Production
CO	1.00	0.25	0.30	0.15	0.20	0.10	0.25	0.40
NO2	0.25	1.00	0.15	0.40	0.35	0.15	-0.10	0.20
O3	0.30	0.15	1.00	0.25	0.30	-0.20	0.30	-0.10
PM10	0.15	0.40	0.25	1.00	0.35	0.05	0.22	0.35
PM2.5	0.20	0.35	0.30	0.35	1.00	0.12	0.15	0.25
Area Harvested	0.10	0.15	-0.20	0.05	0.12	1.00	0.30	0.40
Yield	0.25	-0.10	0.30	0.22	0.15	0.30	1.00	0.35
Production	0.40	0.20	-0.10	0.35	0.25	0.40	0.35	1.00

Table 1: Correlation Matrix between Air Quality Indicators and Crop Production Metrics

6 Future Work

Future work will include generating and incorporating visualizations to better understand the distribution of air pollutants and crop production metrics. Additionally, more advanced statistical analyses and machine learning models will be explored to predict crop yields based on air quality indicators. Extending the analysis to include more recent data as it becomes available. Investigating the impact of additional environmental factors, such as temperature and precipitation, on crop production. Exploring regional differences within Germany to understand how local variations in air quality affect crop production. Finally, Collaborating with environmental scientists and agronomists to validate findings and develop actionable recommendations for farmers and policymakers.