Data Report: Air-Extreme: Analyzing the Impact of Air Quality on Extreme Weather Events

Contents:

Question

How does air quality impact the frequency and severity of extreme weather events globally?

Data Sources

Air Quality Data

Source: WHO Air Quality Database

Description: The WHO Air Quality Database provides global air quality data collected from monitoring stations worldwide. The dataset includes measurements of PM2.5 and PM10 particulate matter concentrations, which are key indicators of air pollution.

Structure and Quality: The dataset contains columns for country, ISO3 codes, city or locality, measurement year, and PM2.5 and PM10 concentrations, along with temporal coverage. The data is structured in tabular form and has some missing values which were handled during preprocessing.

License: The data is available for public use with proper attribution to WHO. We comply by clearly attributing WHO and sharing our derived dataset with proper acknowledgment.

Disaster Data

Source: EM-DAT (The International Disaster Database)

Description: EM-DAT provides data on the occurrence and effects of over 22,000 mass disasters in the world from 1900 to the present day. The dataset includes information on the type and frequency of disasters by country.

Structure and Quality: The dataset is structured in tabular form and contains columns for disaster type, ISO country codes, disaster occurrence dates, and impacts (e.g., total deaths, total affected). The quality is generally high, with comprehensive coverage of natural disasters.

Filtering Criteria: Only natural disasters with subgroups Climatological and Meteorological are included to maintain relevance to air quality impacts.

License: The data is available for academic and non-commercial use with attribution to EM-DAT. We adhere to this requirement by properly attributing EM-DAT in our project documentation and reports.

Data Pipeline

High-Level Description

The data pipeline was implemented using Python, leveraging libraries such as pandas for data manipulation and SQLite for storing the processed data. The pipeline involves data extraction, transformation, and loading (ETL) processes to integrate and prepare the datasets for analysis.

Transformation and Cleaning Steps

In [ ]:
```
import os
import pandas as pd
import sqlite3
```

Loading and Cleaning Air Quality Data

In [ ]:
```
climate_csv = '/Users/asfand/Downloads/who_aap_2021_v9_11august2022.xlsx'
climate_df = pd.read_excel(climate_csv, sheet_name=1)
climate_df.rename(columns={'WHO Country Name': 'Country'}, inplace=True)

columns_to_keep = [
    'ISO3',
    'Country',
    'PM2.5 (µg/m3)',
    'PM10 (µg/m3)'
]
climate_df = climate_df[columns_to_keep]

# Handle missing values by filling them with the mean of the respective c
climate_df['PM2.5 (µg/m3)'].fillna(climate_df['PM2.5 (µg/m3)'].mean(), in
climate_df['PM10 (µg/m3)'].fillna(climate_df['PM10 (µg/m3)'].mean(), inpl

climate_df = climate_df.groupby(['Country', 'ISO3']).agg({
    'PM2.5 (µg/m3)': 'mean',
```

```
    'PM10 (µg/m3)': 'mean'
}).reset_index()
```

Loading and Cleaning Disaster Data

In [ ]:
```
disaster_csv = '/Users/asfand/Downloads/public_emdat_custom_request_2024-
disaster_df = pd.read_excel(disaster_csv)

# Filter to keep only natural disasters and relevant subgroups
relevant_subgroups = ['Climatological', 'Meteorological']
filtered_disasters = disaster_df[
    (disaster_df['Disaster Group'] == 'Natural') &
    (disaster_df['Disaster Subgroup'].isin(relevant_subgroups))
]

# Calculate the count of disasters per country
disaster_counts = filtered_disasters['ISO'].value_counts().reset_index()
disaster_counts.columns = ['ISO', 'Disasters']
```

Merging Datasets

In [ ]:
```
# Merge the two DataFrames on ISO3 and ISO
merged_df = pd.merge(climate_df, disaster_counts, left_on='ISO3', right_o

# Fill NaN values in the Disasters column with 0 (countries with no disas
merged_df['Disasters'].fillna(0, inplace=True)

# Drop the redundant ISO column from the merged DataFrame
merged_df.drop(columns=['ISO'], inplace=True)
```

Problems Encountered and Solutions

- Missing Values: Many rows had missing PM2.5 and PM10 values. This was addressed by using mean imputation.
- Data Matching: Discrepancies in country codes between datasets were handled by ensuring consistent ISO3 codes.
- Licensing: Ensuring compliance with open data licenses by properly attributing sources and sharing derived data under the same license.

Error Handling

- The pipeline includes checks for missing or inconsistent data and logs errors during data loading and transformation. This ensures robustness against changing input data or errors in the datasets.

Result and Limitations

Output Data

The final output is a merged dataset stored in an SQLite database. It contains the average air quality measurements (PM2.5 and PM10) and the count of natural disasters for each country.

Data Structure and Quality

The output data is structured in a relational database format, facilitating efficient queries and analysis. The quality of the data is high due to thorough cleaning and preprocessing steps.

Critical Reflection

- Potential Issues: Some countries may have missing or incomplete air quality data, which could affect the analysis. Additionally, the disaster data may not capture all incidents accurately, especially in less documented regions.
- Future Work: Future work could involve integrating additional data sources to improve coverage and accuracy, and exploring more sophisticated imputation methods for missing data.