Assessing Renewable Energy's Impact on Environmental Health and Pollutant Concentrations in the U.S. within North American region

1 Introduction

1.1 Motivation for selecting the North American region

The USA is an ideal region for this project due to its comprehensive and high-quality datasets, diverse environmental conditions, and significant air quality challenges. With robust data from organizations like the EPA, the USA offers a unique opportunity to explore the impact of pollutants across various climates and urban settings. Its global influence and focus on sustainability make it a critical case study for understanding the relationship between air quality, pollutant emissions, and renewable energy adoption. Insights gained here can drive data-driven policies and solutions applicable worldwide.

1.2 The question of interest

To assess the impact of renewable energy on the Air Quality Index (AQI) and pollutant emissions, this project concentrates on one main question.

“Does increased renewable energy consumption lead to measurable improvements in environmental health and pollutant concentrations?”

2 Dataset

The project utilizes two comprehensive datasets that provide insights into renewable energy consumption and air quality in the United States.

2.1 Datasource 1: U.S Renewable Energy Consumption

- Metadata URL: https://www.kaggle.com/datasets/alistairking/renewable-energy-consumption-in-the-u-s
- Data URL: https://www.kaggle.com/datasets/alistairking/renewable-energy-consumption-in-the-u-s?select=dataset.csv
- Data Type: CSV
- License: U.S. Government Work

This dataset provides monthly data on renewable energy consumption in the United States from January 1973 to January 2024, broken down by energy source and consumption sector. The data is originally sourced from the U.S. Energy Information Administration (EIA). This dataset contains the consumption of renewable energy (sources: Hydroelectric Power, Solar Energy, Wind Energy, Wood Energy, Waste

Energy, Biomass Energy, etc) in the given sector (Commercial, Electric Power, Industrial, Residential, or Transportation) and month, units measured in trillion BTUs.

2.2 Datasource 2: U.S. Pollution Data 2000 - 2023

- Metadata URL: https://www.kaggle.com/datasets/guslovesmath/us-pollution-data-200-to-2022
- Data URL: https://www.kaggle.com/datasets/guslovesmath/us-pollution-data-200-to-2022?select=pollution_2000_2023.csv
- Data Type: CSV
- License: U.S. Government Work

This dataset spans from 2000 to 2023, comprising around 665,414 observations across 21 columns. It provides an analysis of air quality in the United States, with an emphasis on daily data with features like mean value, first max value, first max hour, and AQI for all four components Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2), Carbon Monoxide (CO), and Ozone (O3) for cities of United States. This dataset was originally provided by the U.S. Environmental Protection Agency (EPA).

2.3 Datasource3: U.S. Emissions Data 1990 - 2023

- Metadata URL: https://www.epa.gov/air-emissions-inventories/air-pollutant-emissions-trends-data#
- Data URL: https://www.epa.gov/system/files/other-files/2024-02/state_tier1_08feb2024_ktons.xlsx
- Data Type: XLSX
- License: EPA Data License

This dataset spans from the year 1990 to 2023. It provides an analysis of pollutant emissions in the United States, with an emphasis on pollutants like Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2), Carbon Monoxide (CO), PM10, PM2.5, and others generated from various sources (fuel combustion, wildfires, vehicles, industry, etc.).

3 Data Pipeline

The data pipeline was implemented using Python, leveraging several key technologies and libraries:

3.1 Pipeline Workflow: Extract, Transform, Load (ETL)

3.1.1 Extract

- Kaggle datasets retrieved via API (func: setup_kaggle_credentials, download_dataset)
- EPA emissions data fetched directly from EPA's website (func: download_emissions_data)

3.2.2 Data Transformation & Cleaning Steps

Renewable Energy Dataset	Pollution Dataset	Emissions Dataset
Duplicate removal	Date-time standardization	Data reshaping from wide to long format
Missing value handling using forward/backward fill	State name standardization (whitespace removal)	State abbreviation conversion to full names
Data validation to ensure non-empty datasets	Temporal decomposition (Year, Month, Day columns)	Zero-filling for missing emissions data
Func: preprocess_renewable_energy	Func: preprocess_pollution	Func: process_emissions_data

3.2.3 Data Load

Data is stored temporary in `data/temp` directory during processing and final processed data stored in SQLite database. The temporary files are automatically cleaned up after processing.

3.2.4 Meta-quality Implementation

Error Handling	Data Validation	Download Resilience	Processing Resilience	Storage Resilience
Try-except blocks with specific error types, Automatic retry mechanism for transient failures	Input data verification, Schema validation, Empty dataset checks, Required column verification	Retry mechanism for failed downloads, Multiple data source handling, Connection error handling	Flexible data cleaning that adapts to varying input formats, Robust missing value handling, Type conversion error handling	Database connection error handling, Automatic directory creation, Clean up of temporary files

4 Conclusion and Limitation

Data are stored in SQLite database with all the fields required for data analysis. However, there are some limitations on missing data regarding the Pollution dataset. Much rigorous preprocessing will be required.