MADE- Project Report-SS24
Automated Data Pipeline for Enhancing Environmental Data Analysis: AQI index of Country

Author: Asit Mishra

DataSource_1:

DataURL:"https://www.kaggle.com/datasets , Data taken from kaggle, The data has been made publicly available by the Central Pollution Control Board (CPCB), which is the official portal of the Government of India: CPCB Website.

DataSource_2:

- Data URL: "https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india"
- Data URL: "https://www.kaggle.com/datasets/fedesoriano/air-quality-data-in-india"
- Data URL: "https://www.kaggle.com/datasets/neomatrix369/air-quality-data-in-india-extended"
- Data URL: "https://public.opendatasoft.com/explore/dataset/openaq/export/?disjunctive.city&disjunctive.location&disjunctive.measurements_parameter&sort=measurements_lastupdated&refine.country=IN"
- Data CSV

Question: Objective

How has the air quality in India changed over the years due to factors such as changes in environmental policy, economic growth, and industrialization? Are local trends in air quality detectable? Can we establish correlations between fluctuations in air quality and shifts in environmental policy within India? Additionally, can we predict the Air Quality Index (AQI) based on levels of CO2, NH3, Benzene, SO2, PM2.5, and PM10?

License Compliance:

- Public Data
- All data source used in this project are available under the standard open data license, **"Creative Common CC"** which allow us to use the data for both commercial and non-commercial purposes.

1. Data Pipeline

Technologies Used:

- **Data Loading**: Pandas for data loading.
- **Storage**: Intermediate storage using Pandas Data Frames.
- **Processing**: Pandas and NumPy for data transformation and cleaning.
- **Automation**: Jupiter Notebook for orchestrating the pipeline.

Transformation and Cleaning Steps:

- Data **Loading**: Load CSV files into Pandas Data Frames.
- Data **Cleaning**:
    - **Deduplication Operation:** Remove duplicate entries based on country, city, and date fields.
    - **Handling Missing Values**: Fill or drop missing values based on analysis needs.
    - **Normalization**: Standardize date formats and numerical values.
- **Data Enrichment**: Calculate additional metrics such as AQI over different parameter i.e.,

Error Handling and Adaptability:

- **Error Logging**: Use Python's logging module to capture and log errors during pipeline execution.
- **Data Validation**: Implement validation checks at each stage to ensure data integrity.
- **Scalability**: Design pipeline to handle increasing data volumes and potential new data sources by modularizing components.

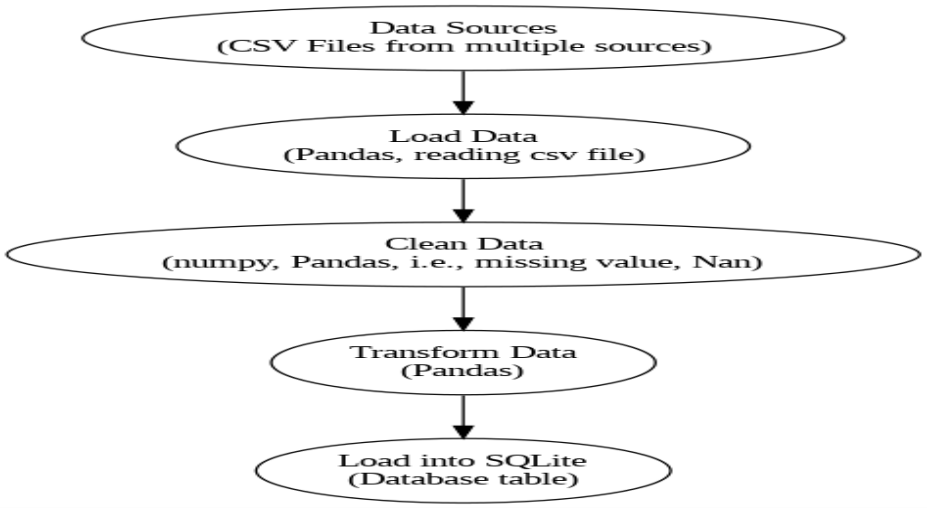
Result and Limitations

Outputs:

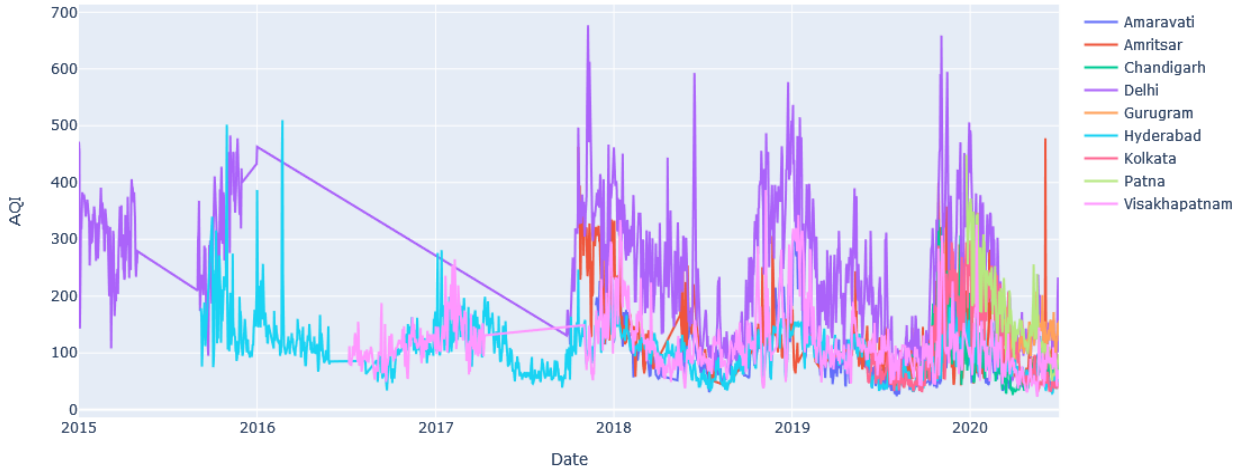
- **Structure**: Combined and cleaned datasets with fields for country, City, Date, PM2.5, PM10, NO, NO2, Nox, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, and additional calculated metrics.
- **Quality**: High-quality data with reduced noise, standardized formats, and enriched information for better analysis.

Critical Issues:

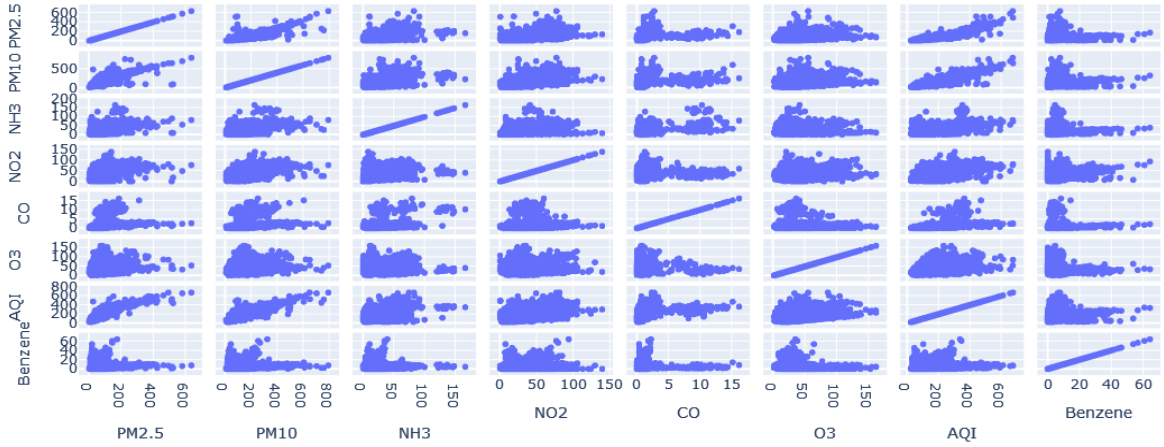
- **Data Issues**: Potential biases in cities and their scale of industrialization may affect the affecter like PM2.5, PM10, NO, NO2, Nox, NH3, CO, SO2, O3, Benzene, Toluene, Xylen inconsistencies in emission reporting across cities in country which may affect the AQI index.
- Future **Improvements**:
    - Enhancement of anomaly detection in cities PM2.5, PM10, NO, NO2, NH3, CO, SO2, O3, Benzene.
    - Integration of additional environmental datasets for a more comprehensive analysis.
    - And will Continuously monitor and adapt the pipeline for evolving data.



AQI Trend Over Time



Scatter Plot Matrix



AQI Distribution by City