
📓 Module 1: Data Preprocessing & Exploratory Data Analysis (EDA)

Project Title: *Air Aware – Smart Air Quality Prediction System*

Introduction

This module focuses on the crucial initial steps of any data science project: **data acquisition, cleaning, exploration, and preparation**. Air quality data is often raw, noisy, and inconsistent. Therefore, before applying forecasting techniques, it must be transformed into a structured and meaningful format.

The three major components of this module are:

1. Download and preprocess datasets (CPCB, OpenAQ)
2. Conduct Exploratory Data Analysis (EDA)
3. Resample and prepare features for forecasting

---

1. Download and Preprocess Air Quality Datasets

Objective:

To obtain raw air quality data from reliable sources and make it usable for analysis.

---

a) CPCB (Central Pollution Control Board, India)

- **Data Source:**
  CPCB publishes air quality data for Indian cities via official portals. Data formats may include CSV, Excel, or APIs.
- **Preprocessing Steps:**
  - **File Handling:** Scripts are required to read multiple files (if split by time/location).
  - **Missing Values:**
    - Drop rows/columns with excessive gaps.
    - Apply imputation methods (mean, median, linear interpolation, or KNN imputation).
  - **Data Type Conversion:**

        - Ensure timestamps are in datetime format.
        - Convert pollutant concentrations into numerical values.
    - **Standardization of Column Names:**
        - Harmonize variations like “pm25”, “PM2.5”, and “Particulate Matter 2.5” into a consistent schema.

---

b) OpenAQ (Global Platform)

- **Data Source:**
  OpenAQ aggregates **real-time and historical air quality data** from global monitoring stations. The data is accessible via API.
- **Preprocessing Steps:**
    - **API Calls:** Use Python libraries (e.g., requests) to fetch location- and pollutant-specific data.
    - **JSON Parsing:** Extract timestamps, pollutant type, location, and values.
    - **Data Structuring:** Store parsed data into a pandas DataFrame for analysis.
    - **Unit Standardization:** Convert inconsistent measurement units into standard form ($\mu g/m^3$).

---

2. Conduct Exploratory Data Analysis (EDA)

Objective:

To visualize and summarize the dataset, identify patterns, and detect anomalies or correlations.

---

a) Time Series Analysis

- **Temporal Trends:**
  Plot PM2.5, PM10, SO_2, NO_2 concentrations across time to observe daily, weekly, and seasonal fluctuations.
- **Diurnal Patterns:**
  Identify pollutant variations across a 24-hour cycle (e.g., traffic-related peaks).

- **Seasonal Variation:**
  Use boxplots to compare seasonal/monthly levels, showing winter peaks or festival-related surges.

---

b) Statistical Summaries

- **Descriptive Statistics:**
  Compute mean, median, standard deviation, min, and max for each pollutant.
- **Distribution Plots:**
  Histograms or KDE plots highlight skewness and outliers.

---

c) Correlation Analysis

- **Correlation Matrix & Heatmaps:**
  Measure relationships between pollutants (e.g., PM2.5 and PM10 usually correlate strongly).
- **Scatter Plots:**
  Visualize interactions (e.g., PM2.5 vs. NO_2).
- **Pollutant–Weather Correlation (if available):**
  Study associations with temperature, humidity, and wind speed.

---

3. Resample Data and Prepare Features for Forecasting

Objective:

To transform raw, irregular air quality data into a structured format suitable for predictive modeling.

---

a) Resampling

- **Why Resample?**
  Raw measurements may be recorded at 5–15 min intervals. For forecasting, fixed intervals (hourly/daily) are needed.
- **Methods:**
  - **Downsampling:** Aggregate frequent readings into larger intervals.
    - mean() → hourly/daily averages (most common for pollutants).

        - max() → capture peak values.
    - **Upsampling (less common):** Fill higher-frequency intervals with imputed data.

---

b) Feature Engineering

- **Time-Based Features:**
    - Hour of day → captures diurnal cycles.
    - Day of week → distinguishes weekday vs. weekend traffic.
    - Month/season → highlights seasonal variation.
- **Lag Features:**
    - Example: PM2.5_lag1 = PM2.5 concentration one hour/day ago.
- **Rolling Window Features:**
    - Rolling mean (e.g., last 6 hours average).
    - Rolling standard deviation (pollutant variability).
- **External Features (if integrated):**
    - Weather data (temperature, humidity, wind).
    - Holiday/event flags (e.g., Diwali, New Year).

Expected Outcomes of Module 1

1. **Clean, standardized dataset** ready for analysis.
2. **Resampled time-series dataset** at hourly/daily frequency.
3. **Visual insights** into pollutant behavior through plots and heatmaps.
4. **Feature-engineered dataset** prepared for forecasting models in Module 2.

✅ **In summary:**
Module 1 builds the foundation of the *Air Aware Project* by ensuring the dataset is clean, reliable, and structured. Through EDA, it uncovers patterns, correlations, and temporal behaviors of pollutants. Finally, it prepares meaningful features that improve the accuracy of future forecasting models