Air Quality Data Analysis and Visualization Report

Dataset: UCI Machine Learning Repository - Air Quality Data Set
Tools Used: MATLAB
Project Scope: Data cleaning, exploratory data analysis (EDA), visualization, and predictive modeling of air pollution metrics in relation to environmental conditions.

---

1. Objective

The project aims to explore, visualize, and model air quality data to uncover trends, outliers, seasonal patterns, and relationships between pollutants and environmental variables like temperature and humidity.

2. Dataset Description

The dataset contains hourly measurements from an air quality monitoring station in an Italian city, with features such as:

- **Pollutants:** CO(GT), NOx(GT), NO2(GT), NMHC(GT), C6H6(GT)
- **Sensor Responses:** PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3)
- **Environmental Variables:** Temperature (T), Relative Humidity (RH), Absolute Humidity (AH)
- **Time Features:** Date, Time (converted to unified Datetime)

Missing values are encoded as `-200` and were converted to `NaN`.

3. Data Cleaning

- Converted `Date` and `Time` to a unified `Datetime` column.
- Replaced `-200` values with `NaN` for all relevant features.
- Extracted daily and monthly time components for aggregation.

4. Exploratory Data Analysis (EDA)

- **Time Series Visualization:** Plotted time series for each pollutant and sensor.
- **Monthly Trends:** Displayed mean and standard deviation per month.
- **Histograms:** Showed the distribution of pollutant concentrations.
- **Autocorrelation Analysis:** ACF plots were used to detect repeating patterns:
  Key Observations:
  - **Strong 24-hour cycles:** Found in CO(GT), C6H6(GT), NOx(GT), NO2(GT), and all sensors except PT08.S5(O3).
  - **NMHC(GT):** No clear periodicity; showed irregular spikes.
  - **Sensor responses:** Generally followed expected pollutant patterns, with minor deviations (e.g., PT08.S3(NOx) showed strong periodicity despite low correlation).

  - **Weak periodicity:** PT08.S5(O3) exhibited noisy cycles, possibly near 24–26 hours.

5. Outlier Detection

- **Periodic Variables:** Outliers identified using IQR method on a per-day basis.
- **Non-Periodic Variables:** Outliers identified globally using IQR.
- **Visualization:** Outliers overlaid on time series plots to validate detection.

6. Correlation and Relationship Analysis

- **Correlation Heatmap:** Pearson correlation revealed that CO(GT) is strongly correlated with NMHC(GT), C6H6(GT), and PT08.S1(CO), suggesting related sources or reactions. NOx(GT) and NO2(GT) also showed high mutual correlation. Sensor variables generally aligned with their target pollutants, though PT08.S3(NOx) showed weak or negative correlation, potentially due to sensor drift. Environmental variables (T, RH, AH) correlated well with each other but weakly with pollutants.
- **Scatter Plot Matrix:** Confirmed linear relationships like CO(GT) vs C6H6(GT) and NMHC(GT) vs C6H6(GT). Nonlinear patterns appeared between sensor and ground-truth values. Histograms showed skewed distributions and potential outliers, while scatter plots revealed seasonal clusters.

7. Environmental Influence (T & RH)

- Created 2D heatmaps showing average pollutant values under various temperature and humidity bins.
- Showed strong environmental influence on pollutant concentrations.

8. Predictive Modeling

- **Model:** Multiple Linear Regression using predictors: PT08.S1(CO), Temperature, and RH.
- **Target:** CO(GT)
- **Performance:**
  - R-squared: 0.782
  - Root Mean Squared Error: 0.649
- **Insights:**
  - PT08.S1(CO) is the most significant predictor.
  - Temperature and RH negatively influence CO concentration.
  - Model is reasonably accurate for simple linear prediction.

9. Conclusion

This MATLAB-based project effectively demonstrated data preprocessing, EDA, visualization, and modeling for air quality data. Visualizations revealed pollutant behavior and periodicity. Correlation and regression analysis indicated environmental factors are predictive of CO levels. The analysis pipeline can be expanded for other pollutants or used in real-time monitoring systems.

Deliverables:

- MATLAB Live Scripts & Figures
- Summary report (this document)

Author: Yuanyuan Feng
Date: 08/08/2025