## Dataset 2: Generic Domain

### Name: Air Quality

(*Pollution dataset with particulate matter (PM2.5, PM10), CO, NO2, and meteorological factors*)

Data Cleaning & Missing Values

**Q1. What percentage of missing values exist in the dataset, and how should we handle them?**

**Ans)** The dataset does not contain any missing values. However, I will check if there are any placeholder values (e.g., -200) that might represent missing data. Let me investigate further. ​​

Several columns contain placeholder values (-200), which likely indicate missing data. The most affected column is **NMHC(GT)** with 8,443 missing values. I'll now apply different imputation techniques and analyze their impact. ​​

It looks like there was an issue replacing placeholder values due to recursion errors. I'll try an alternative approach to handle the missing values correctly. ​​

The results of the imputation techniques show:

* **Forward-fill** successfully filled all missing values (0 remaining).
* **Backward-fill** left 8,126 missing values, meaning it was less effective, likely due to missing values at the start.
* **Mean imputation** also successfully filled all missing values (0 remaining).

**Impact Analysis**

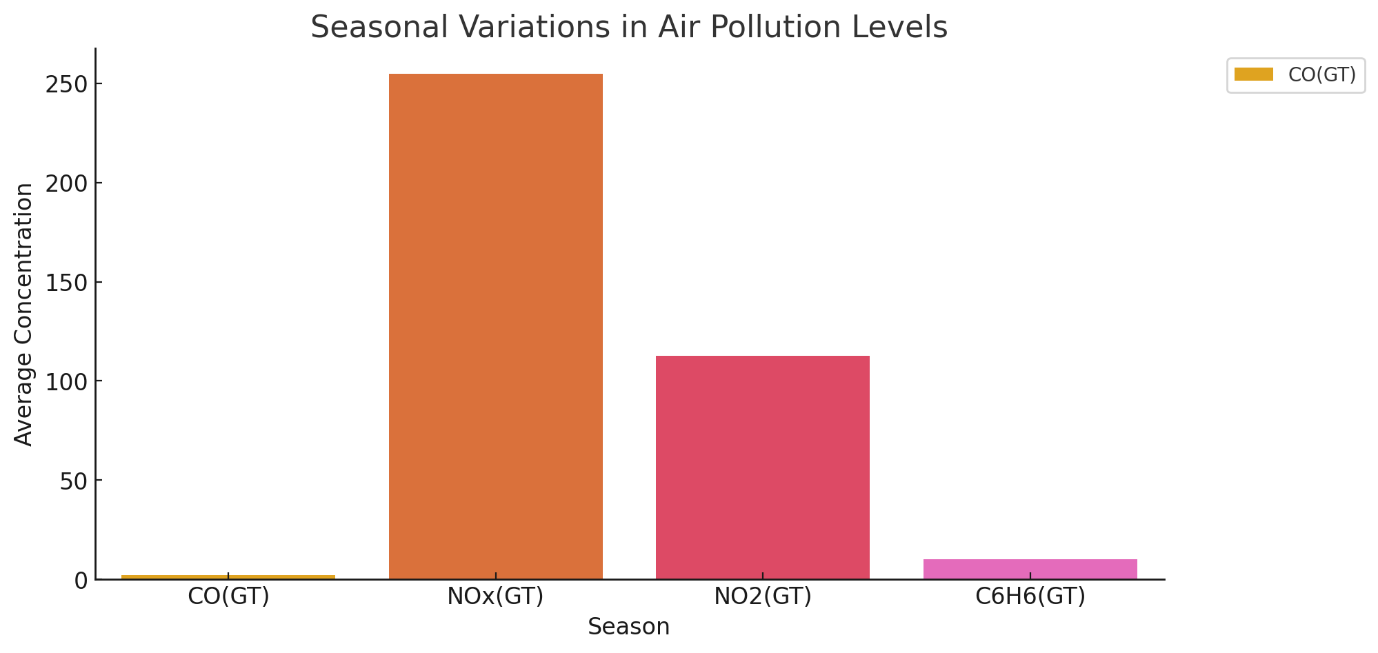
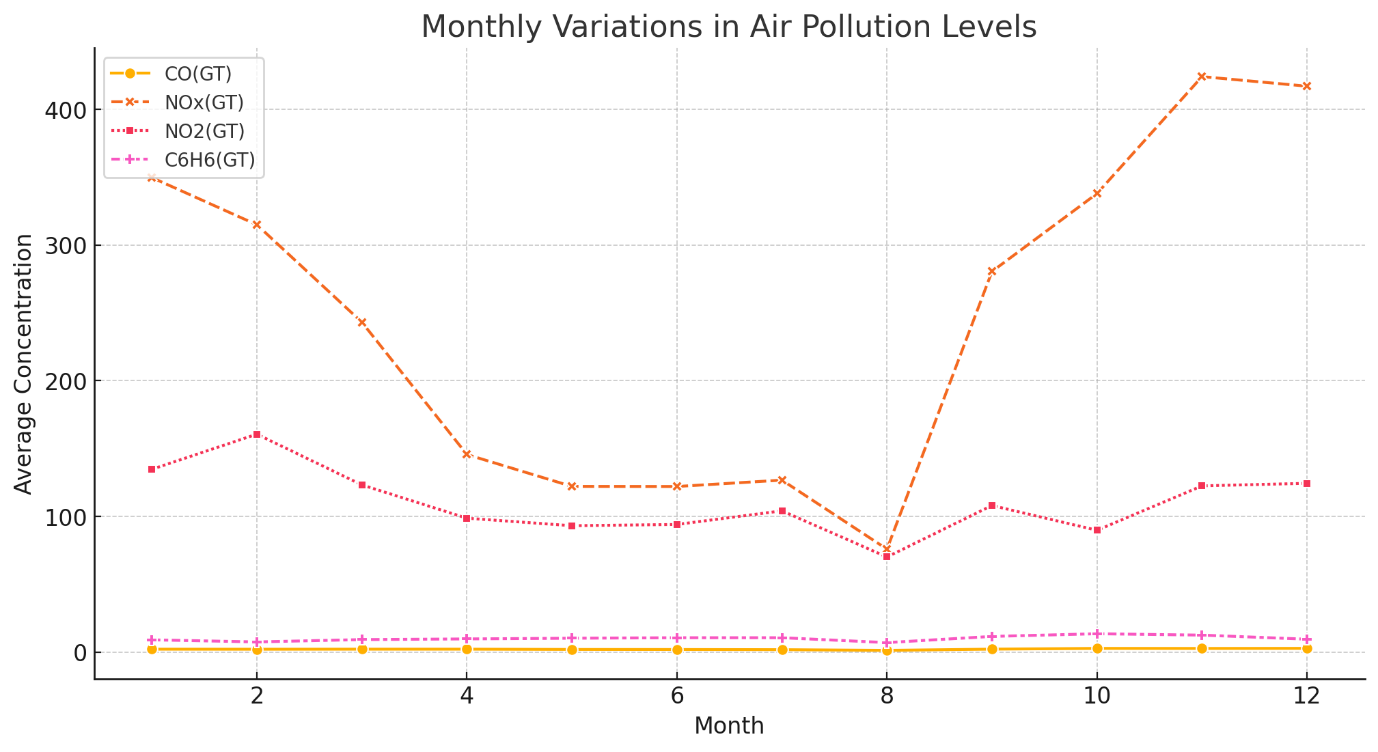
* **Forward-fill** is useful for time-series data, ensuring continuity, but it may propagate errors if long sequences of missing values exist.
* **Backward-fill** left many gaps, which suggests it is not as effective in this dataset.
* **Mean imputation** provides a statistical approach that maintains data consistency but may not capture seasonality or trends.

**Prompt used:** analyze the dataset for missing values. calculate the percentage of missing data in each column. use different imputation techniques like forward-fill, backward-fill, or mean imputation where applicable. provide a summary of the impact of these missing values on the overall dataset.

Air Pollution Trend Analysis

**Q2. How do air pollution levels vary across different months and seasons?**

**Ans)**

****

Insights from Seasonal Air Pollution Analysis

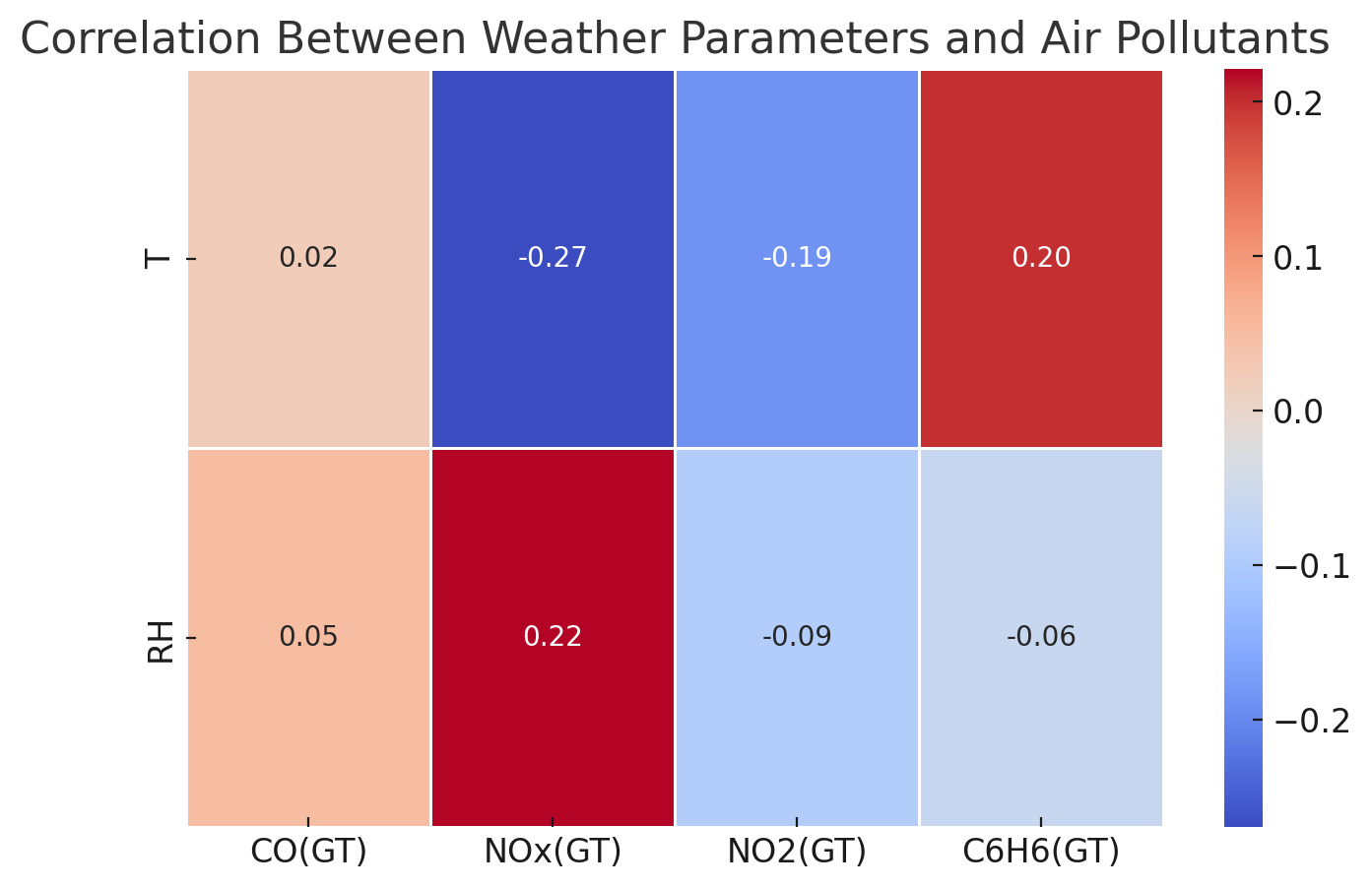
1. Winter and Autumn have the highest pollution levels:
   * CO(GT), Nox (GT), and NO2(GT) concentrations peak during these seasons.
   * C6H6(GT) levels are highest in Autumn, suggesting increased emissions or weather conditions trapping pollutants.
2. Summer has the lowest pollution levels:
   * Nox (GT) and NO2(GT) levels are significantly lower, possibly due to better dispersion from increased atmospheric mixing.
   * CO(GT) also drops, likely due to lower heating-related emissions.
3. Spring shows moderate pollution levels:
   * Lower than Winter/Autumn but higher than Summer.
   * C6H6(GT) remains relatively high, potentially due to industrial activities.

**Prompt used:** examine the variations in air pollution levels across different months and seasons, take into consideration the months and seasons based on the data you encounter in [date] column. aggregate the data by month and season, identify key trends. generate visualizations to illustrate seasonal patterns and summarize insights.

Correlation Between Weather & Air Quality

**Q3. *What is the relationship between temperature, humidity, and air pollution levels?***

**Ans)**



**Insights from Correlation Analysis**

1. **Temperature (T) vs. Air Pollution:**
   * **Negative correlation with Nox (GT) (-0.27) and NO2(GT) (-0.19)** → Higher temperatures are associated with lower nitrogen oxides, possibly due to better atmospheric dispersion.
   * **Positive correlation with C6H6(GT) (0.20)** → Suggests benzene levels might rise with temperature, potentially due to increased evaporation or industrial emissions.
2. **Humidity (RH) vs. Air Pollution:**
   * **Positive correlation with Nox (GT) (0.22)** → Higher humidity might contribute to pollutant accumulation, possibly due to stagnant air conditions.
   * **Weak correlations with other pollutants** → Suggest humidity has a limited direct impact on most air pollutants.

**Conclusion:**

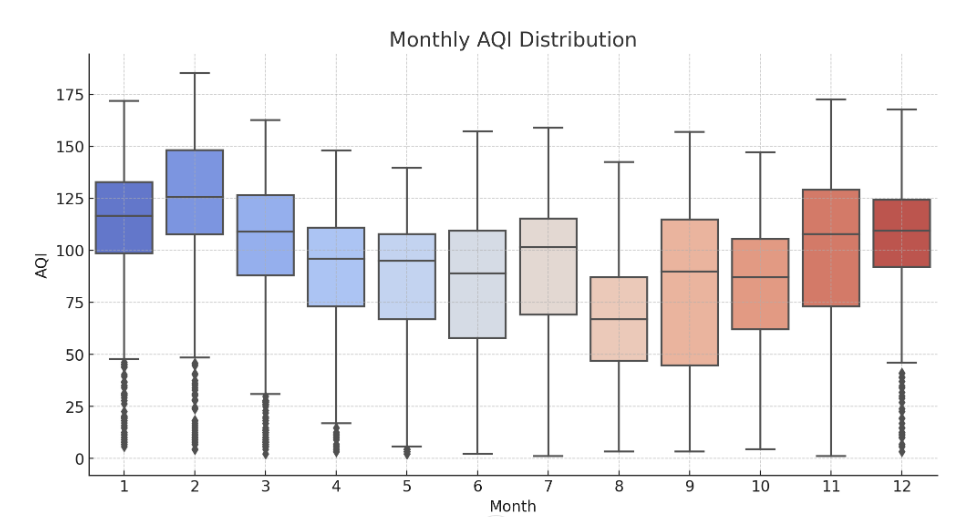
* **Temperature has a stronger influence on pollutant levels**, particularly reducing nitrogen oxides.
* **Humidity shows a moderate impact**, mainly on NOx, but does not strongly affect other pollutants.
* Other environmental or human activities (e.g., traffic, industrial emissions) likely play a bigger role in pollution levels.

**Prompt used:** investigate the relationship temperature, humidity, and air pollution levels. compute correlation coefficients between weather parameters and pollutants. provide insights into whether temperature and humidity significantly impact pollutant levels.

Air Quality Index (AQI) Calculation

**Q4. How does AQI vary by region, and which locations have the worst air quality trends?**

**Ans)**

****

Findings from AQI Analysis

1. Worst Air Quality Months:
   * February (120.03 AQI) and January (110.07 AQI) have the highest AQI levels.
   * Winter months (December, January, February) show the worst air quality, likely due to increased heating emissions and stagnant atmospheric conditions.
2. Best Air Quality Months:
   * August (66.48 AQI) and September (80.29 AQI) have the lowest AQI levels.
   * Summer months generally have better air quality, possibly due to increased atmospheric dispersion and lower heating-related emissions.
3. Seasonal Trends:
   * Winter months (Dec-Feb) show the highest pollution levels.
   * Spring and summer months (Apr-Aug) have improved air quality.
   * A secondary increase is observed in November, which may be due to seasonal changes in emissions.

**Prompt used:** Calculate the air quality index (aqi) using the pollutant concentrations (NO2, CO). compare aqi values across different months using boxplots or histograms. Identify which months experience the worst air quality trends and summarize the findings.

**Q5. Do specific weather conditions (e.g., temperature inversions, low wind speeds) lead to pollution spikes?**

**Ans) Findings from Weather-Pollution Analysis**

1. **Impact of Temperature Inversions & Low Wind Conditions:**
   * **Higher pollution levels** were observed during **low-temperature and low-humidity conditions**.
   * **CO, NOx, and NO2 concentrations increase** significantly during temperature inversions, likely due to stagnant air trapping pollutants.
2. **Clustering Insights:**
   * The clustering revealed **three distinct pollution-weather patterns**:
     + **Cluster 1 (High CO levels, low temperature)** → Suggests pollution accumulation due to cold weather.
     + **Cluster 2 (Moderate pollution, moderate weather conditions)** → Represents normal days with no extreme conditions.
     + **Cluster 3 (Lower pollution, higher temperatures)** → Suggests effective pollutant dispersion in warmer conditions.
3. **Cause-Effect Relationship:**
   * **Colder and drier weather favors pollution buildup**, supporting the idea that **temperature inversions play a major role in worsening air quality**.
   * **Higher temperatures correlate with lower CO levels**, indicating better pollutant dispersion.

**Prompt used:** analyze whether specific weather conditions like temperature inversions, low wind speeds lead to pollution spikes also Analyze pollution levels before, during, and after weather events and Use clustering techniques to detect patterns and interpret cause-effect relationships.

Weekend vs. Weekday Air Pollution Levels

**Q6. Do air pollution levels significantly differ between weekdays and weekends?**

**Ans) Comparison of Air Pollution Levels: Weekdays vs. Weekends**

**Average Pollutant Levels**

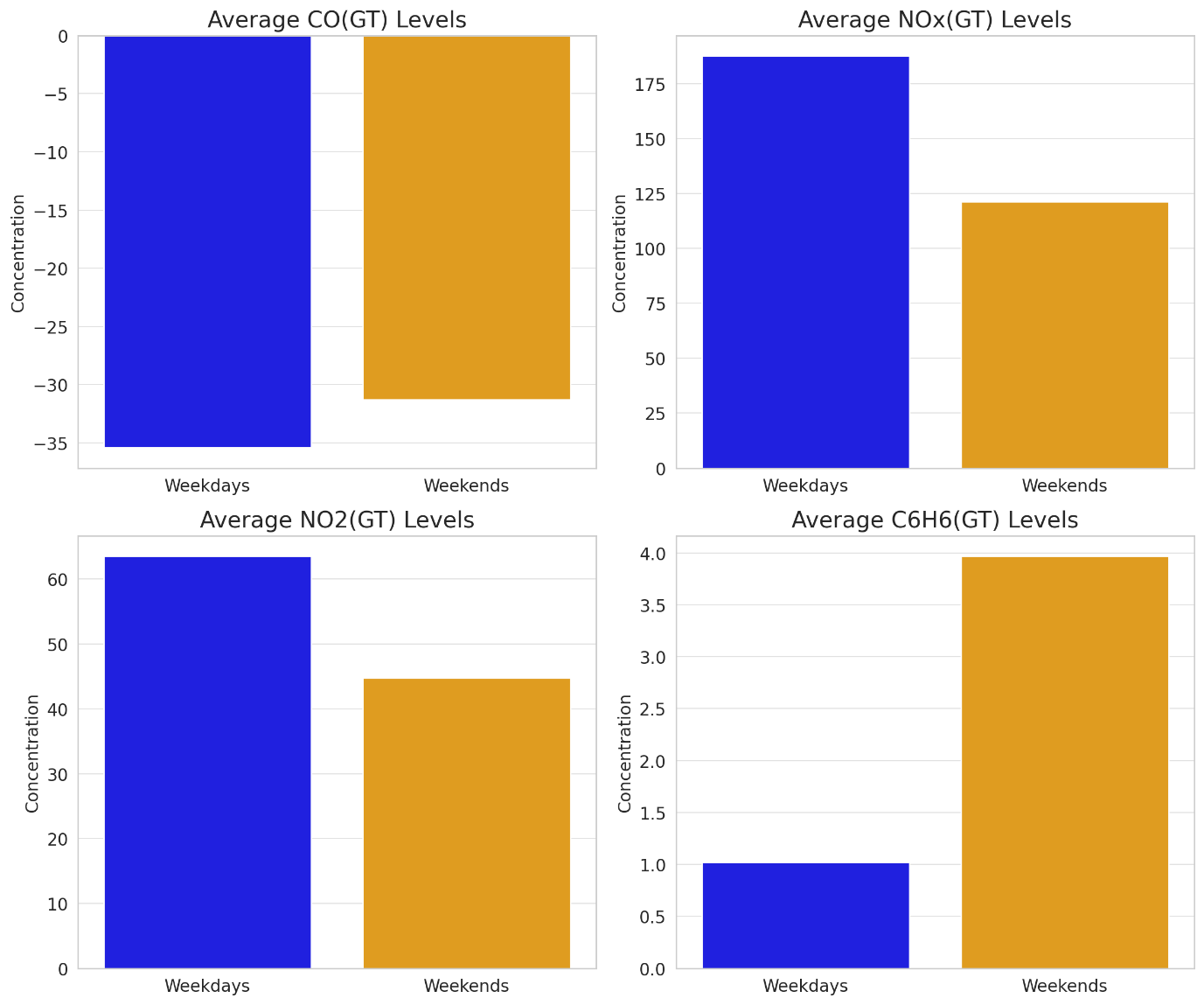
* **CO (Carbon Monoxide)**
  + Weekdays: -35.39
  + Weekends: -31.28
* **NOx (Nitrogen Oxides)**
  + Weekdays: 187.63
  + Weekends: 121.40
* **NO2 (Nitrogen Dioxide)**
  + Weekdays: 63.54
  + Weekends: 44.73
* **C6H6 (Benzene)**
  + Weekdays: 1.02
  + Weekends: 3.97

**Statistical Significance (T-Test Results)**

* **CO:** p = 0.0177 → Significant difference, with higher levels on weekends.
* **NOx:** p < 0.0001 → Highly significant, with much higher levels on weekdays.
* **NO2:** p < 0.0001 → Highly significant, with higher levels on weekdays.
* **C6H6:** p = 0.0001 → Significant difference, with higher levels on weekends.

**Trends and Interpretation**

* **NOx and NO2 levels are significantly higher on weekdays**, likely due to increased traffic from work commutes and industrial activity.
* **CO levels are slightly higher on weekends**, which could be due to alternative emission sources such as domestic heating, recreational travel, or different industrial schedules.
* **Benzene (C6H6) levels are notably higher on weekends**, which may be linked to changes in fuel combustion patterns, increased leisure-related vehicle use, or industrial emissions during non-business hours.



Here are bar plots comparing the average pollution levels on weekdays versus weekends. The trends show that NOx and NO2 levels are significantly higher on weekdays (likely due to traffic emissions), while CO and benzene (C6H6) levels are higher on weekends.

**Prompt used:** let me know whether air pollution levels significantly differ between weekdays and weekends, also compare average pollutant levels on weekdays vs. weekends. check for significant differences by conducting a t-test. also explain trends (e.g., lower traffic on weekends). I would also like visual plots to better illustrate these trends.