Air Quality Data Analysis in Urban Monitoring: A Case Study

Author: Atakan Tatar

Middle East Technical University

Statistics

August 1, 2025

**Abstract**

This report analyzes hourly air pollution data collected from an urban monitoring station. The aim is to identify potential data loss, evaluate pollutant trends over time, and detect relationships between specific chemical compounds. Data cleaning and exploratory data analysis (EDA) were conducted using Python libraries such as pandas, matplotlib, and seaborn. The report concludes with suggestions for data collection improvement and further modeling.

**Introduction**

Urban air pollution poses a serious threat to public health and environmental quality. Accurate and continuous monitoring of pollutants such as nitrogen oxides (NOx), carbon monoxide (CO), and non-methane hydrocarbons (NMHC) is essential for evidence-based policy making. In this analysis, we examine an air quality dataset collected from a European city to explore hourly trends and assess the presence of missing or erroneous data.

**Dataset Description**

The dataset contains 9,358 observations and 13 variables, including gas concentrations (in µg/m³), temperature (°C), and relative humidity (%). Key variables are:

- CO(GT): True hourly average of CO concentration.

- NOx(GT): True hourly average of NOx concentration.

- NMHC(GT): True hourly average of non-methane hydrocarbons.

- T: Ambient temperature.

- RH: Relative humidity.

Negative values such as -200 indicate sensor errors or missing data points. These values require special treatment during cleaning.

**Data Cleaning**

To ensure accurate analysis, rows containing -200 in the major pollutant columns (NOx(GT), NMHC(GT), CO(GT)) were dropped. This eliminated approximately 18% of the total rows. Additionally, timestamps were converted into datetime format to allow temporal analysis.

**Exploratory Data Analysis**

- **Hourly trends** of pollutant levels were analyzed using line plots to detect patterns throughout the day. In particular, NO(GT) and NO₂(GT) values were examined to identify any recurring measurement anomalies. çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, metin içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

**Figure 1**  
Hourly average of NO(GT) values.çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, metin içeren bir resim

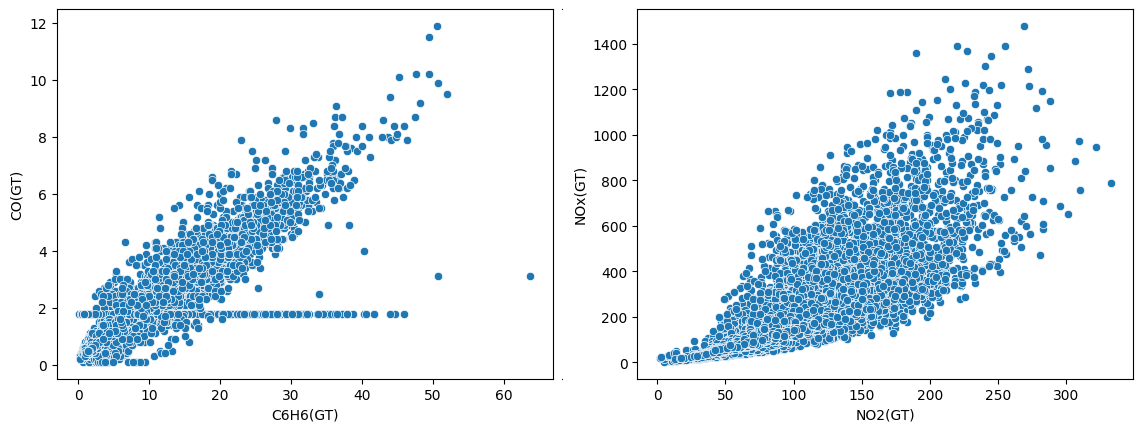
Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

**Figure 2**  
Hourly average of NO₂(GT) values.  
- A **time series plot** of NO₂(GT) was generated to explore long-term changes in nitrogen dioxide levels across the dataset timeline.metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü, diyagram içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

**Figure 3**  
Time series of NO₂(GT) concentrations across the observed period.

- To examine relationships between pollutants, **scatterplots** were created. The correlation between C6H6(GT) (benzene) and CO(GT) (carbon monoxide), as well as between NO₂(GT) and NOx(GT), was visually assessed.



**Figure 4**

Scatterplots of Pollutant Relationships:  
(a) C6H6(GT) versus CO(GT); (b) NO₂(GT) versus NOx(GT).

**Findings**

- A recurring data loss was observed specifically around **03:00 AM**, where both **NO (Nitric Oxide)** and **NO₂ (Nitrogen Dioxide)** readings were either missing or recorded as **-200**, indicating sensor failure or measurement gaps. This pattern points to a **systematic issue** during early morning hours, possibly related to device calibration or operational downtime.

- Throughout the dataset, **NO₂(GT)** values showed a **gradual upward trend**, suggesting increasing nitrogen dioxide levels during the initial monitoring period. However, a **noticeable decline occurred after February 2005**, which may indicate a shift in emission sources, environmental regulations, or monitoring protocols.

- The analysis revealed a **strong positive correlation** between **C6H6(GT)** (benzene) and **CO(GT)** (carbon monoxide), implying that these pollutants are likely to share common sources, such as **vehicle exhaust** or **industrial combustion**.

- A similar pattern was observed between **NO₂(GT)** and **NOx(GT)**, with a **clear linear relationship**. This is consistent with the fact that **NO₂ is a major component of total nitrogen oxides (NOx)**, and their levels typically rise and fall together under similar emission conditions.

**Conclusion**

This data analysis provided insights into patterns and potential issues within the air quality monitoring dataset. The exploratory analysis highlighted systematic data loss occurring during early morning hours, specifically around 03:00 AM, which may indicate a sensor or calibration problem requiring further investigation. Additionally, NO₂(GT) concentrations showed a gradual upward trend during the monitored period, followed by a decline after February 2005, possibly reflecting changes in local emission sources or regulatory practices.

The strong positive correlations identified between C6H6(GT) and CO(GT), as well as between NO₂(GT) and NOx(GT), suggest common emission sources, most likely related to vehicular traffic and industrial combustion. These relationships may inform targeted air quality control strategies in urban areas.

Future research could expand on these findings by applying predictive modeling techniques, investigating meteorological influences, and incorporating traffic flow data to better understand the dynamics of pollutant concentrations over time. Overall, ensuring consistent sensor calibration and data validation will be essential for maintaining the reliability of future monitoring efforts.

**References**

Air Quality UCI Dataset. (2008). UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Air+Quality

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics (4th ed.). SAGE Publications.

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.).