Investigating Air Quality Factors and Their Impact on Human Health

Group 28 – DSE 501 (Spring 2025)

Mohammad Hamza Choudhry • Kushal Trivedi

Kaviya Venkateshbabu • Srujana Rachamalla

Instructor: Prof. Rong Pan | Arizona State University

Agenda

- Introduction and Motivation
- Dataset Overview
- Data Cleaning and Preprocessing
- Exploratory Data Analysis
- Hypothesis Testing (H1–H6)
- Key Findings
- Public Health Implications and Policy Recommendations
- Future Work

Introduction

Air pollution remains a major global health challenge.
Urban populations are especially at risk due to:

- Heavy traffic and industrial emissions
- Seasonal factors like winter heating
- Weather conditions that trap pollutants near the ground

Our project focuses on:

- Analyzing when and why pollution peaks occur
- Evaluating the reliability of sensor-based air monitoring
- Connecting patterns in pollution to human health concerns

Dataset Overview

We used the UCI Air Quality dataset, which contains:

- 9,358 hourly records from an urban location in Italy
- Ground-truth pollutant values: CO, NOx, NO_2, Benzene
- Sensor readings: 5 metal oxide sensors (PT08.S1–S5)
- Weather data: Temperature, Relative Humidity, Absolute Humidity
- Time features: Date and Time (used to extract hour, weekday, month)

This dataset allowed us to explore pollutant behavior, sensor accuracy, and environmental effects over time.

Air Quality Dataset: Attribute Definitions and Descriptions

The dataset includes hourly measurements from an air quality monitoring station. Each column represents a pollutant, sensor reading, or environmental variable.

Column Name	Description
Date	Date of measurement (DD/MM/YYYY)
Time	Time of measurement (HH.MM.SS)
CO(GT)	True CO concentration in mg/m^3
PT08.S1(CO)	Sensor 1 output in response to CO
NMHC(GT)	Non-Methane Hydrocarbons in $\mu g/m^3$ (many values missing)
C6H6(GT)	Benzene concentration in $\mu g/m^3$
PT08.S2(NMHC)	Sensor 2 output targeting NMHC
NOx(GT)	True NOx concentration in ppm

Air Quality Dataset: Attribute Definitions and Descriptions

Column Name	Description
PT08.S3(NOx)	Sensor 3 output targeting NOx
NO2(GT)	True NO_2 concentration in μg/m^3
PT08.S4(NO2)	Sensor 4 output targeting NO_2
PT08.S5(O3)	Sensor 5 output targeting O_3
T	Ambient temperature in °C
RH	Relative humidity (%)
AH	Absolute humidity in g/m^3
Hour / Month	Extracted from timestamp for temporal trend analysis

This structured dataset enabled detailed analysis of pollutant behavior, sensor reliability, and temporal trends.

Data Cleaning & Preprocessing

To ensure data quality and consistency, we performed the following steps:

- Replaced all invalid values (e.g., **-200**) with missing value markers
- Dropped incomplete rows to retain only valid hourly records
- Combined Date and Time columns to form a single **Datetime** column
- Extracted new features:
  - **Hour of day**
  - **Day of week**
  - **Month**
- Binned temperature and humidity values to analyze joint effects on pollutants

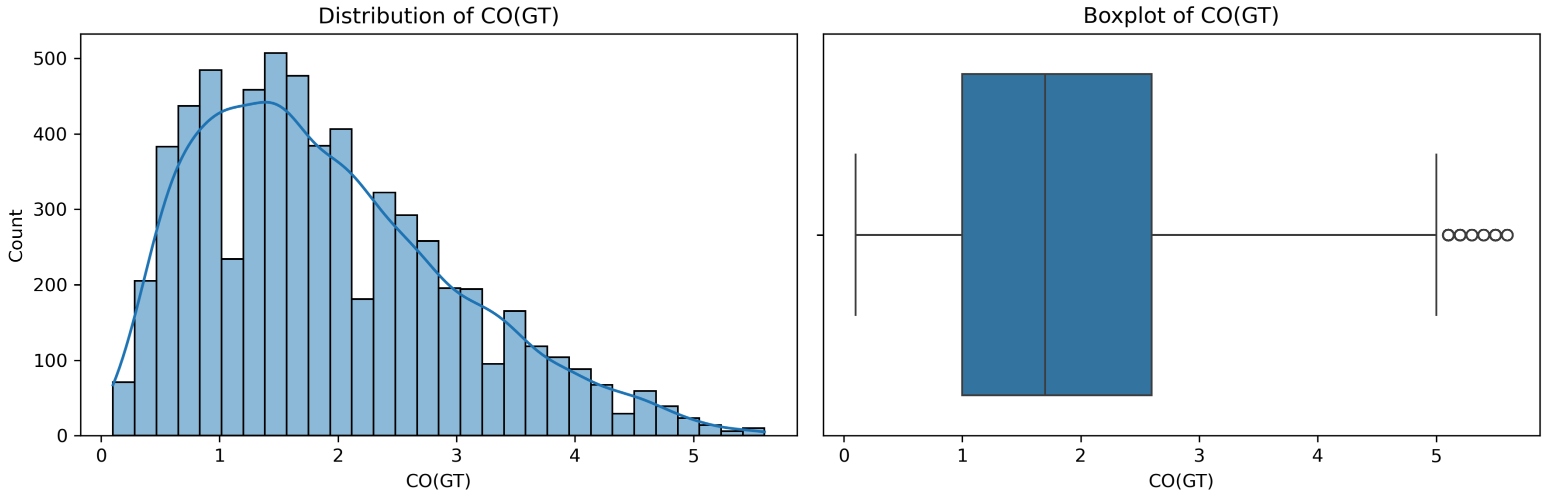
These steps prepared the dataset for time-based analysis and hypothesis testing.

Pollutant Distributions

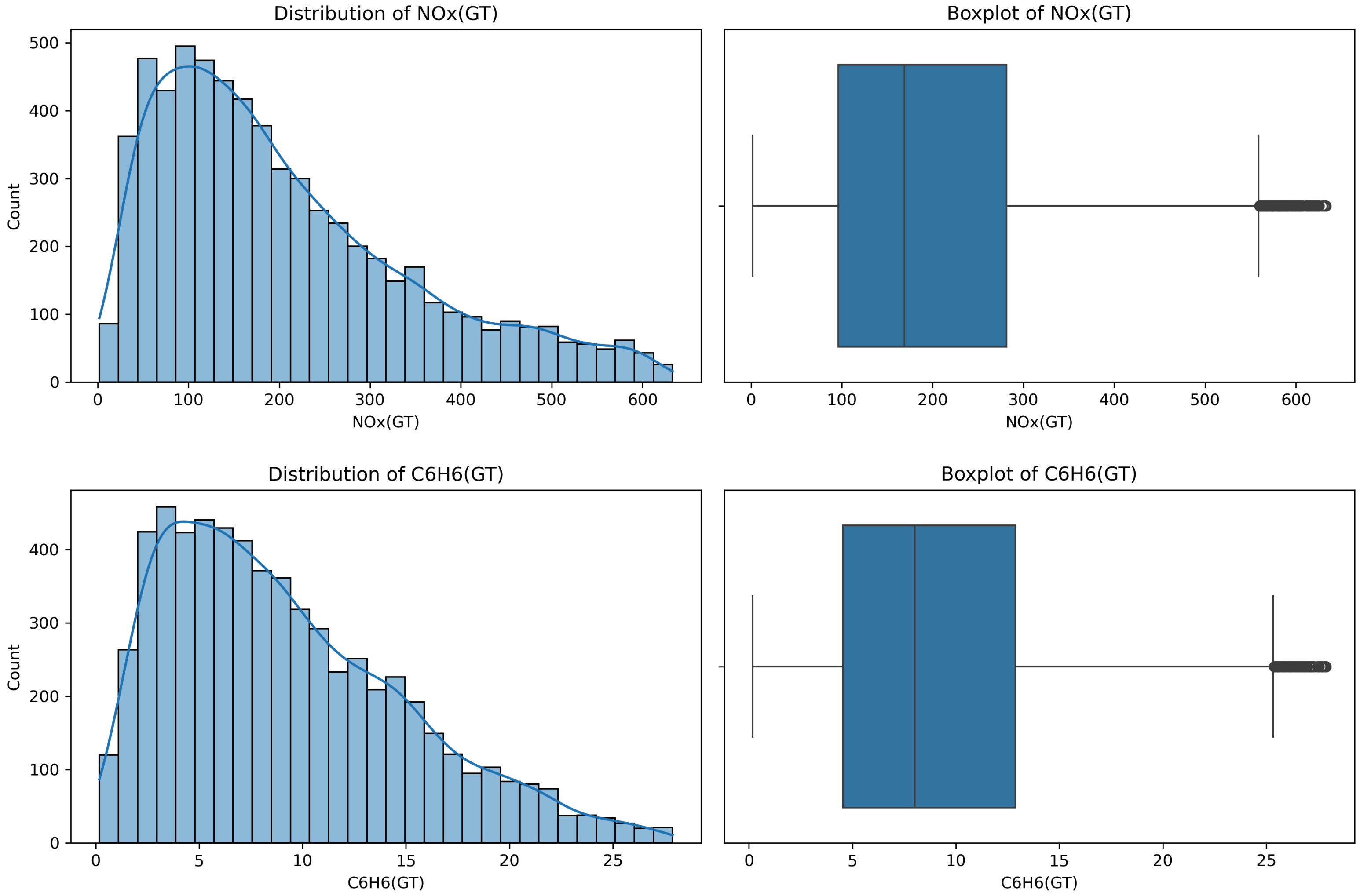
We examined the distribution of each key pollutant:

- Most pollutants showed **right-skewed patterns**
- A majority of values were low to moderate
- Some **high outliers** reflect traffic or industrial spikes

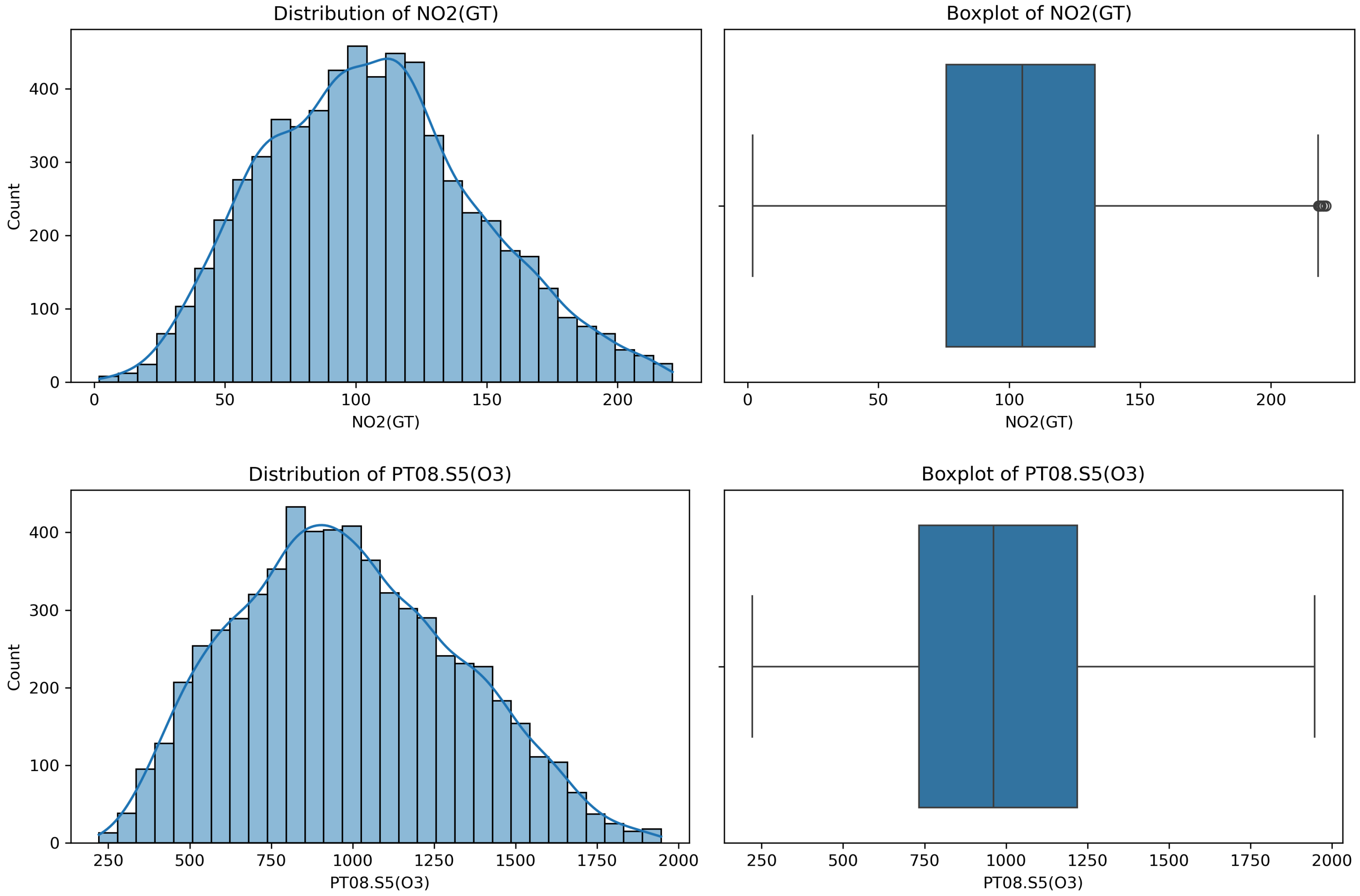
Visuals:



Pollutant Distributions



Pollutant Distributions



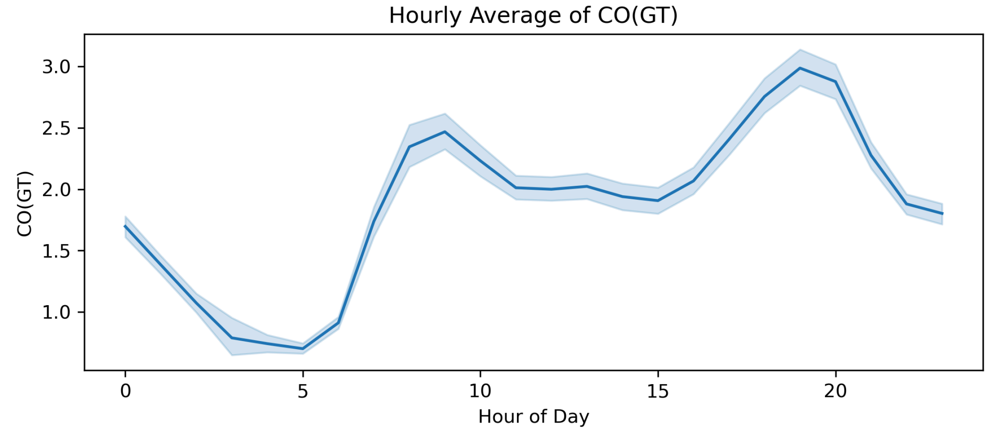
Daily and Weekly Trends

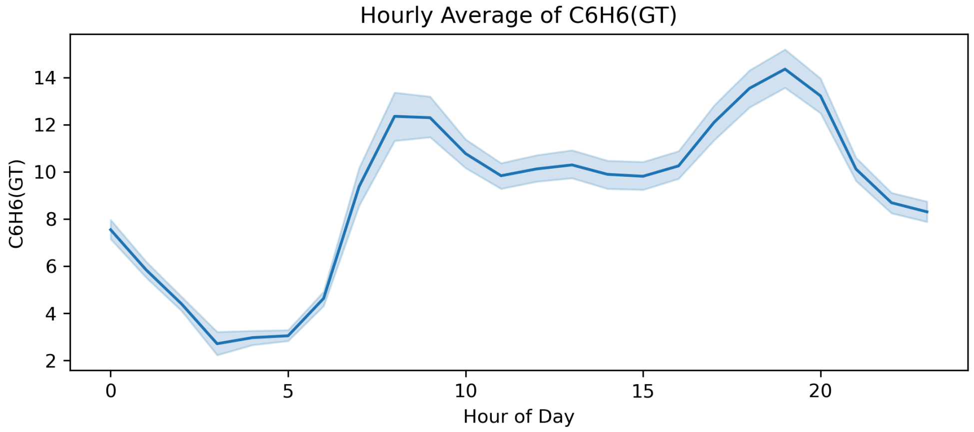
We observed clear temporal patterns in pollutant levels:

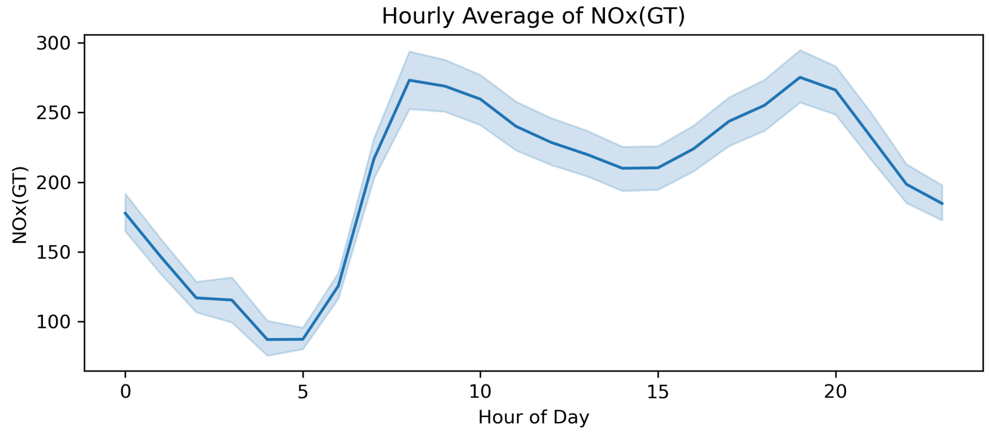
- **Rush hour peaks** around 8 AM and 6 PM for CO, NOx and C6H6
- **Weekday levels** are consistently higher than weekends
- **Ozone (O_3)** tends to rise during mid-day sunlight hours

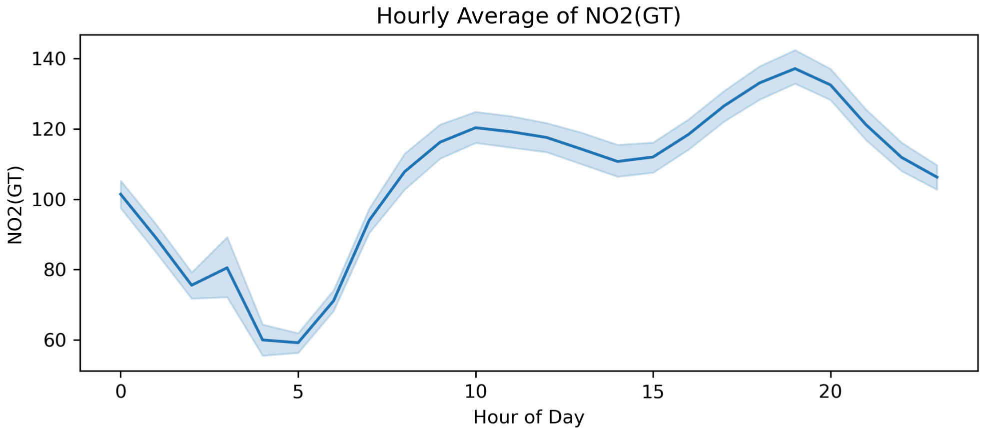
Hourly Trends: Pollution Peaks

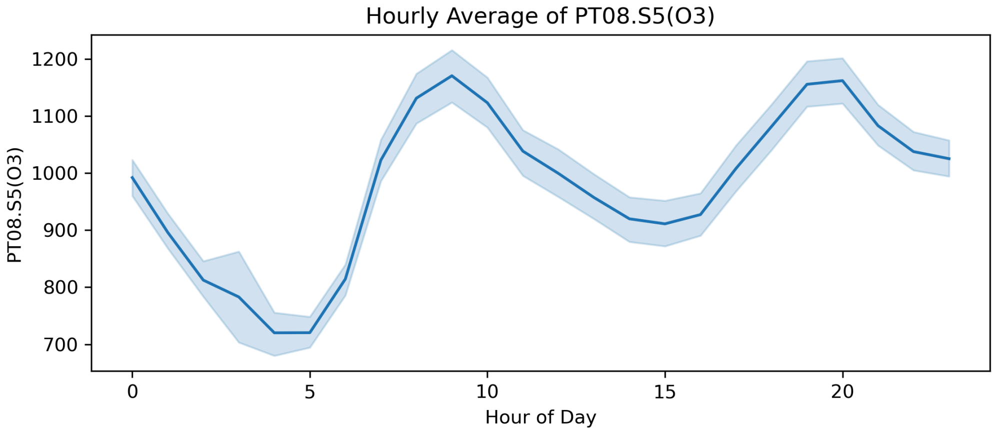
- **CO, NOx, NO_2, and Benzene (C6H6)** show clear peaks around **8 AM and 6 PM**, aligning with rush hours
- **O_3 levels** rise steadily in the morning and **peak mid-day**, consistent with **sunlight-driven reactions**


Hourly Average of CO(GT)
CO(GT)
1.0
1.5
2.0
2.5
3.0
0
5
10
15
20
Hour of Day


Hourly Average of C6H6(GT)
C6H6(GT)
2
4
6
8
10
12
14
0
5
10
15
20
Hour of Day

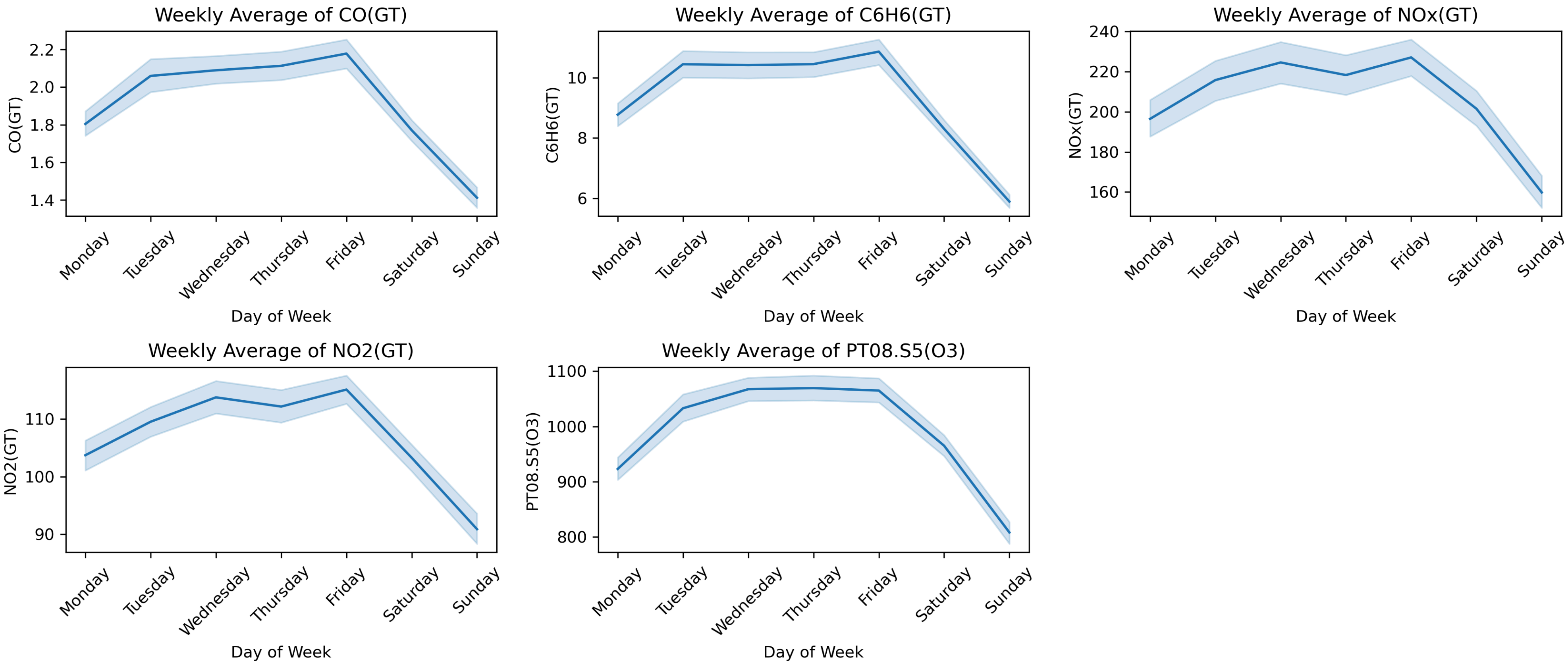

Hourly Average of NOx(GT)
NOx(GT)
100
150
200
250
300
0
5
10
15
20
Hour of Day


Hourly Average of NO2(GT)
NO2(GT)
60
80
100
120
140
0
5
10
15
20
Hour of Day


Hourly Average of PT08.S5(O3)
PT08.S5(O3)
700
800
900
1000
1100
1200
0
5
10
15
20
Hour of Day

Weekly Trends: Weekdays vs. Weekends

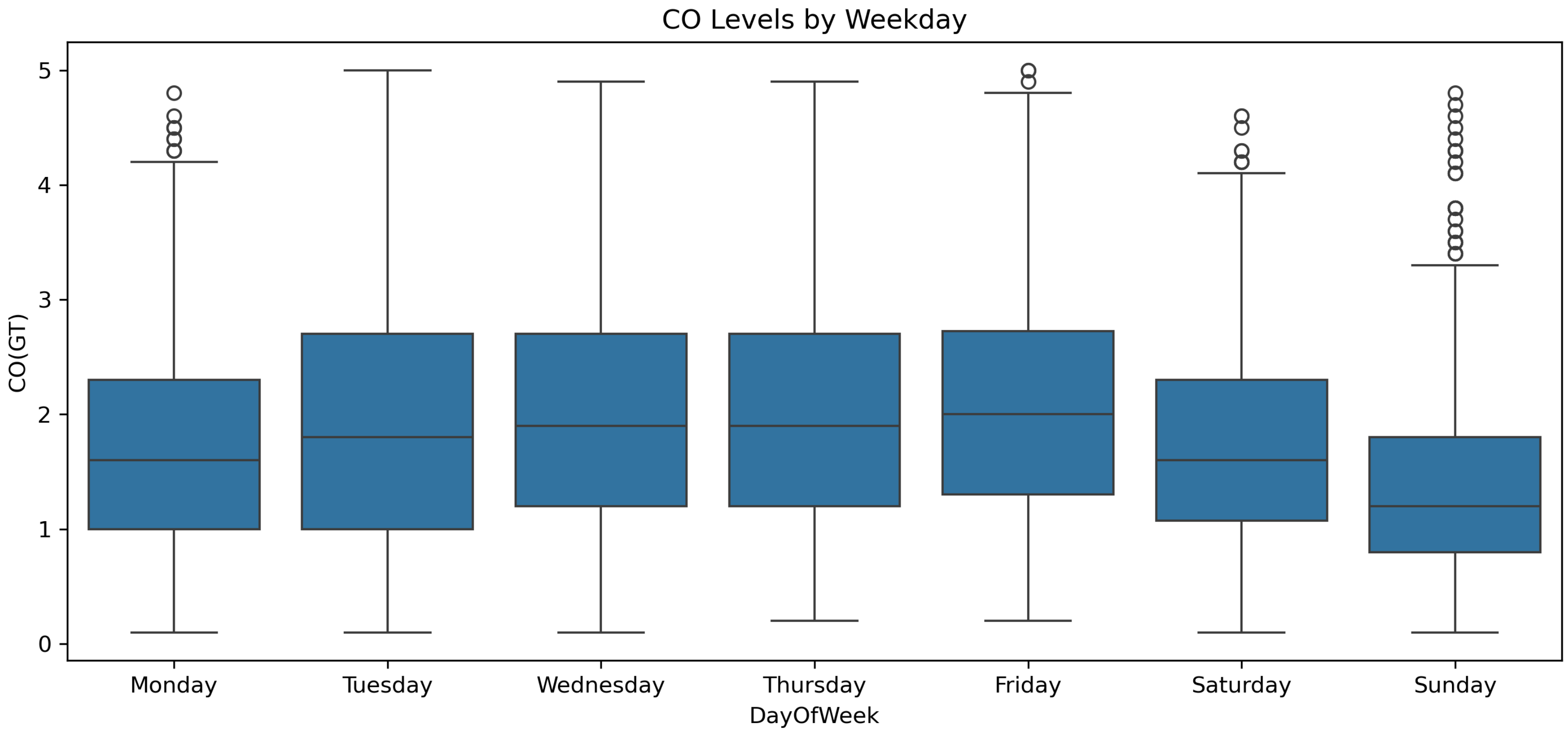
- Pollutant levels are consistently **higher on weekdays**, especially Tuesday to Friday
- Sharp drop on **Saturdays and Sundays**, likely due to reduced traffic and industrial output



Weekly Average of CO(GT)
CO(GT)
2.2
2.0
1.8
1.6
1.4
Weekly Average of C6H6(GT)
C6H6(GT)
10
8
6
Weekly Average of NOx(GT)
NOx(GT)
240
220
200
180
160
Weekly Average of NO2(GT)
NO2(GT)
110
100
90
Weekly Average of PT08.S5(O3)
PT08.S5(O3)
1100
1000
900
800
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

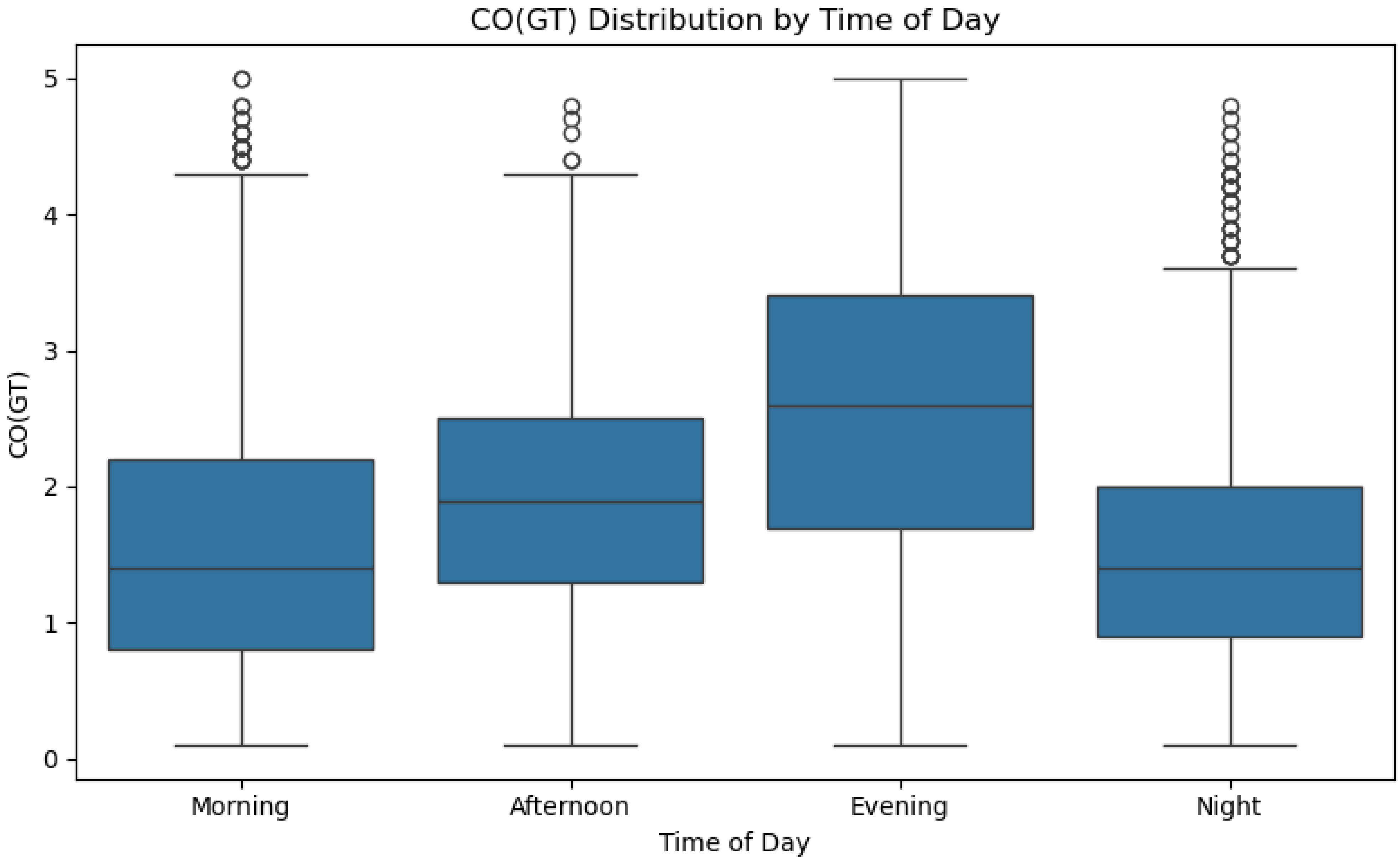
Weekly CO Variation: Boxplot View

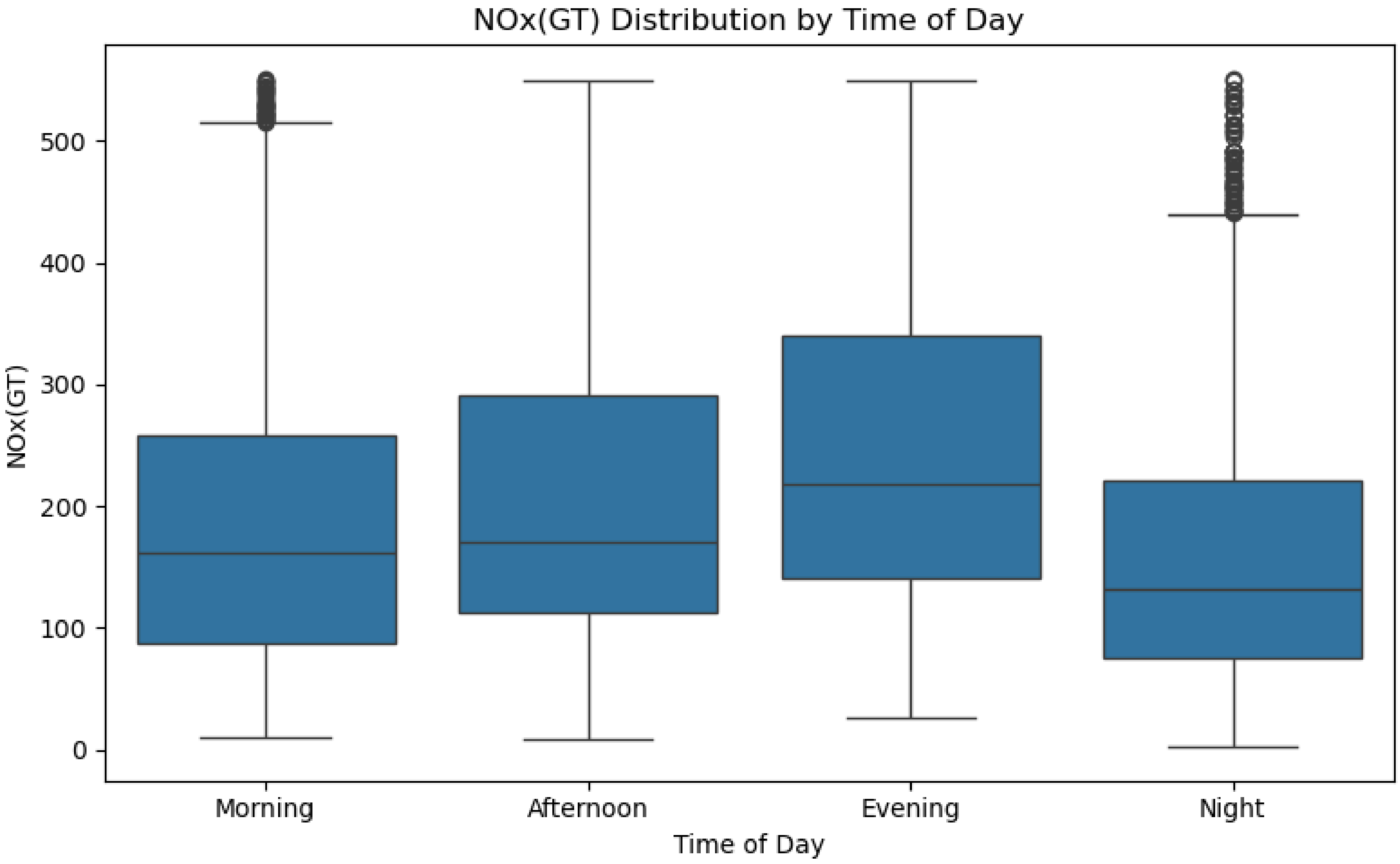
- **CO concentrations** are higher midweek and lower on weekends
- **Sunday** shows the **cleanest air** overall
- The boxplot reveals variability and confirms traffic-related pollution trends



Time of Day Patterns: CO and NOx

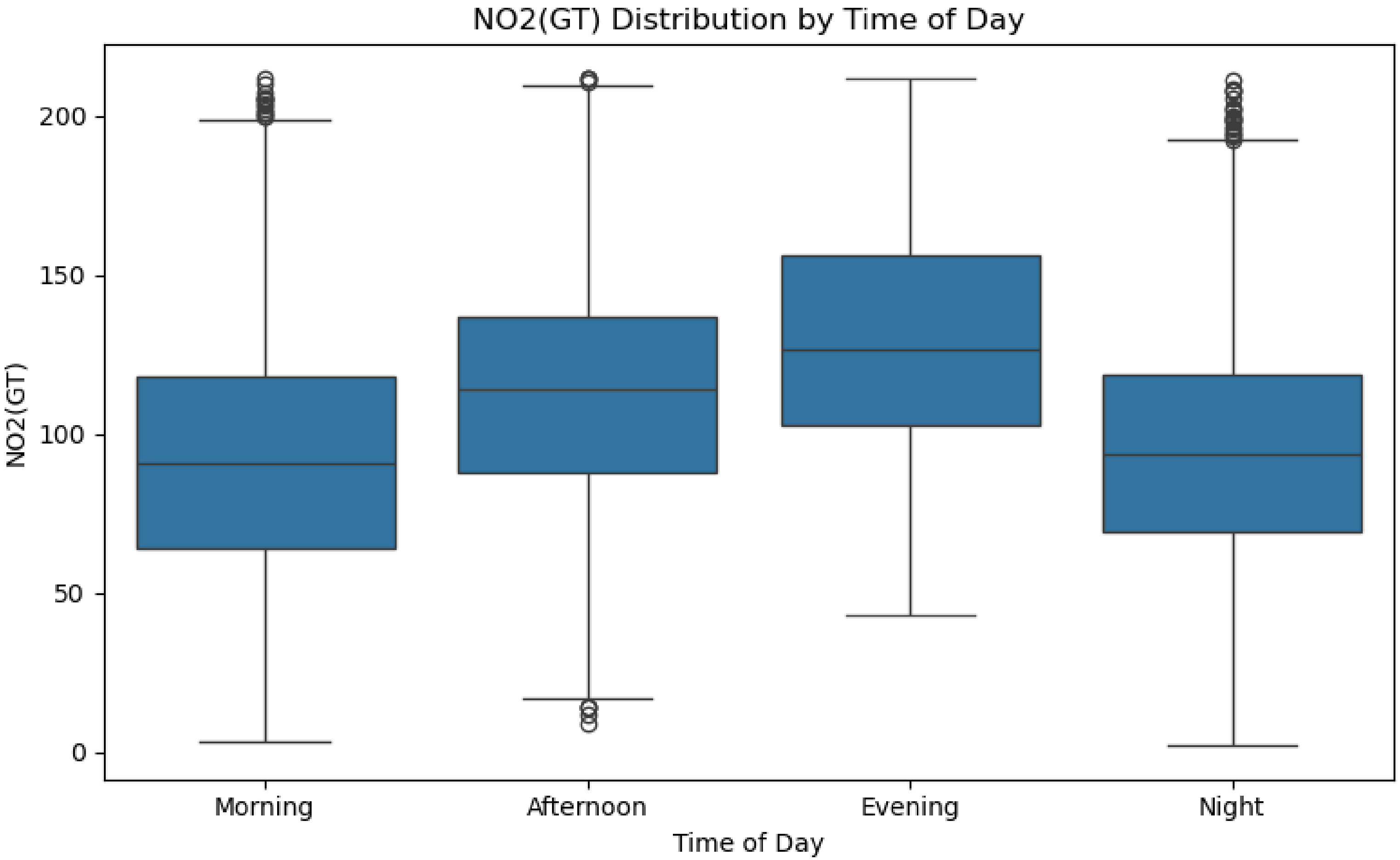
- Both **CO** and **NOx** concentrations peak during the **evening** period.
- This reflects typical urban patterns tied to **evening rush hour** and reduced dispersion after sunset.
- CO levels are slightly higher in the afternoon than in the morning, but **evening is the dominant peak**.


CO(GT) Distribution by Time of Day
CO(GT)
Time of Day
Morning
Afternoon
Evening
Night
0
1
2
3
4
5

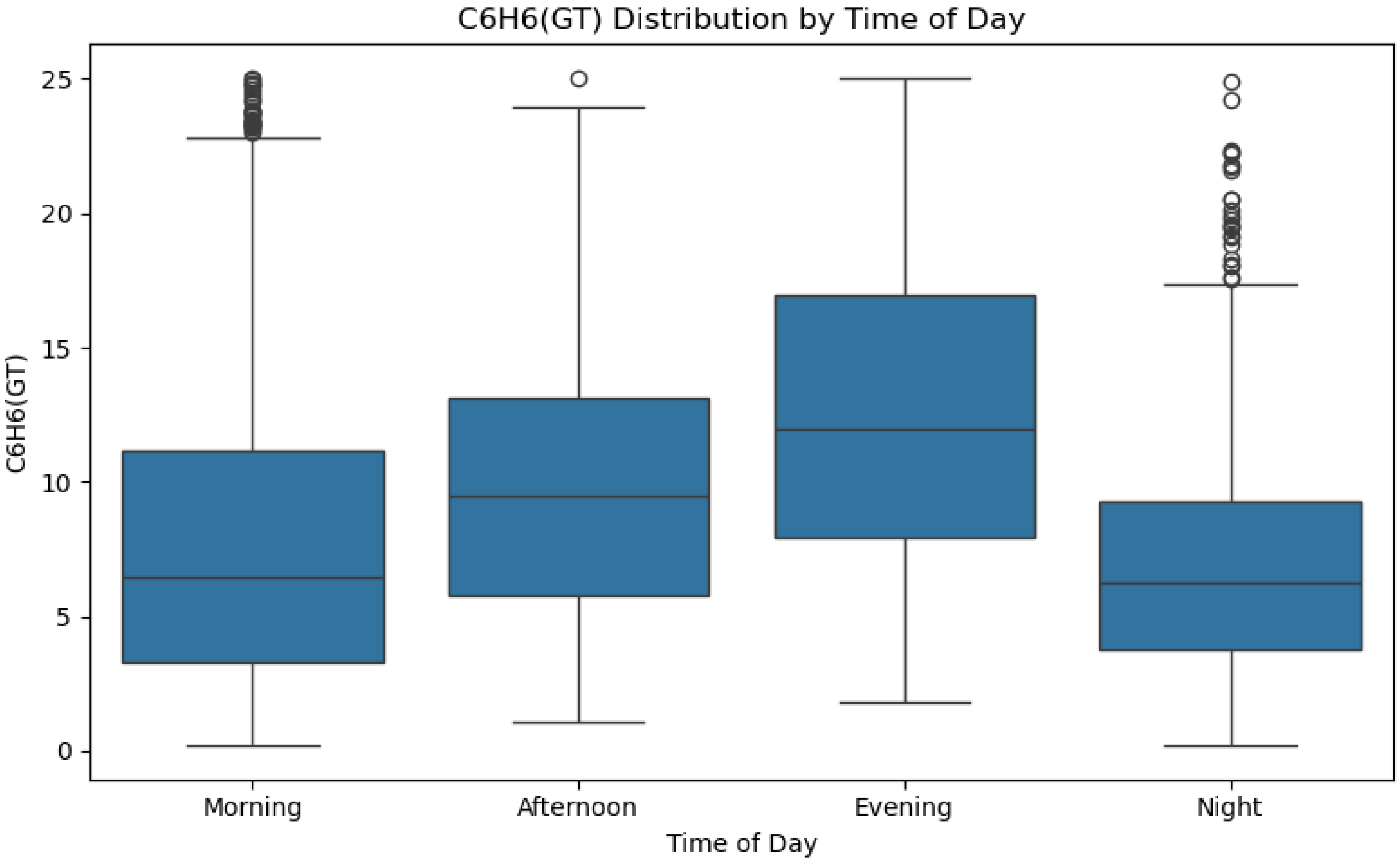

NOx(GT) Distribution by Time of Day
NOx(GT)
0
100
200
300
400
500
Morning
Afternoon
Evening
Night
Time of Day

Time of Day Patterns: NO_2

- **NO_2 concentrations** increase throughout the day, reaching a **clear peak in the evening**.
- Follows similar dynamics to NOx, suggesting shared traffic-based sources.

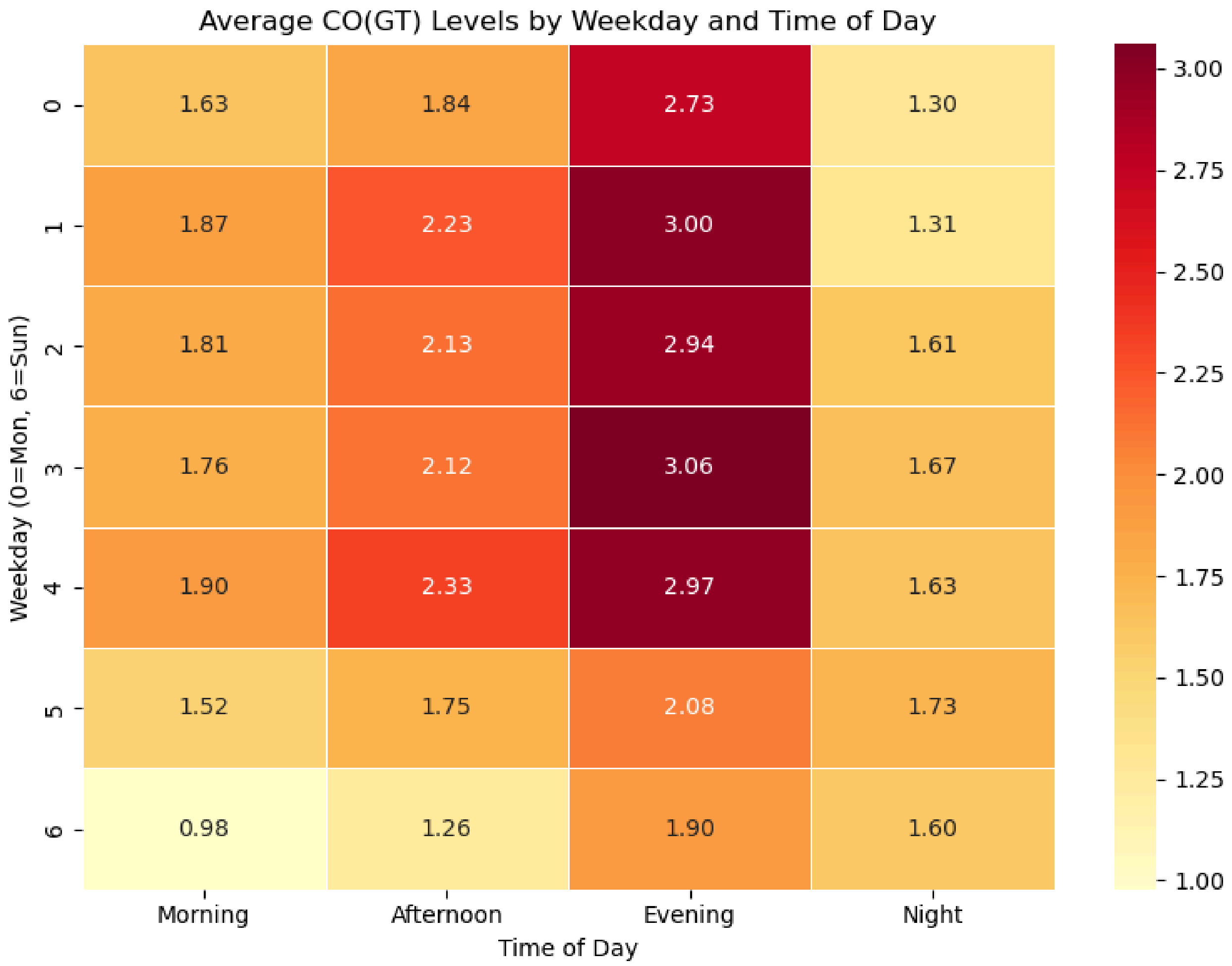


- **Benzene (C6H6)** also peaks in the **evening**, likely due to workday industrial and traffic emissions.



Cross-Time Heatmap: Weekday × Hour

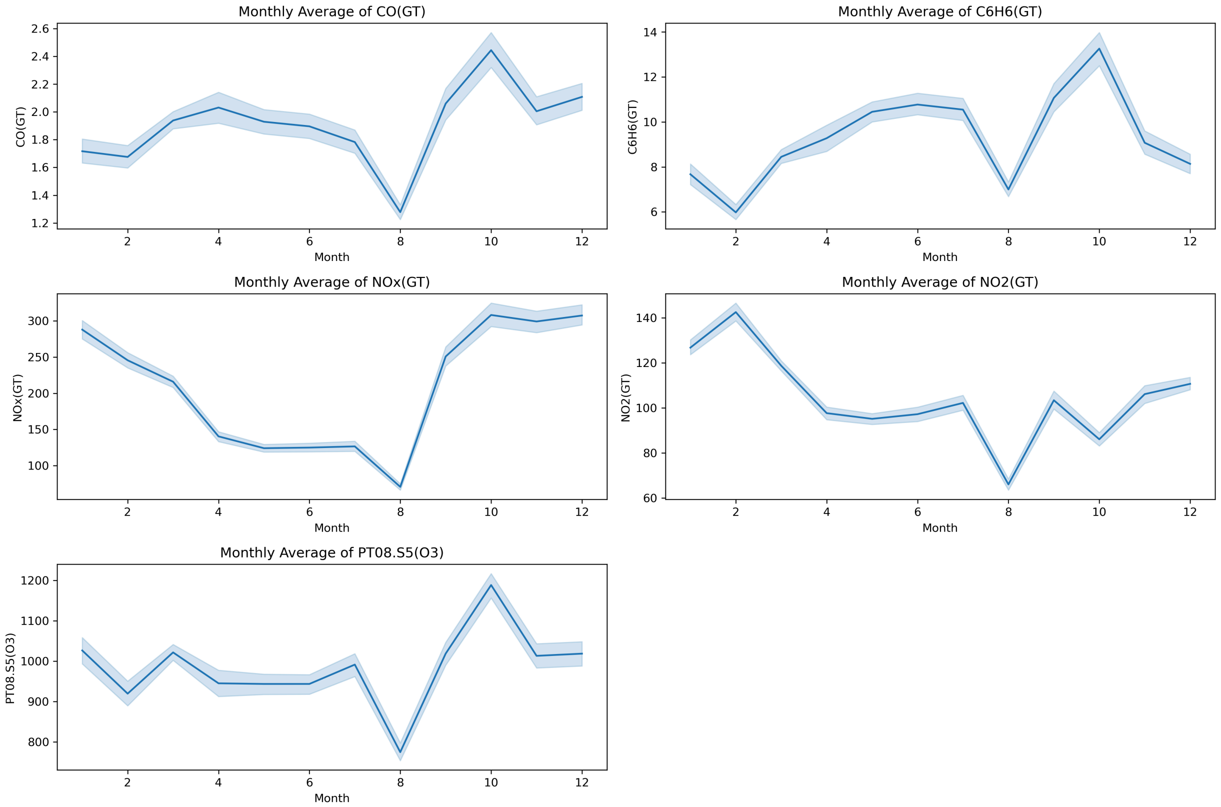
Highest CO concentrations are observed on **weekday evenings** — especially Tuesday to Thursday.



Seasonal Pollution Trends

Pollutants like CO and NOx are significantly higher in **winter months**.

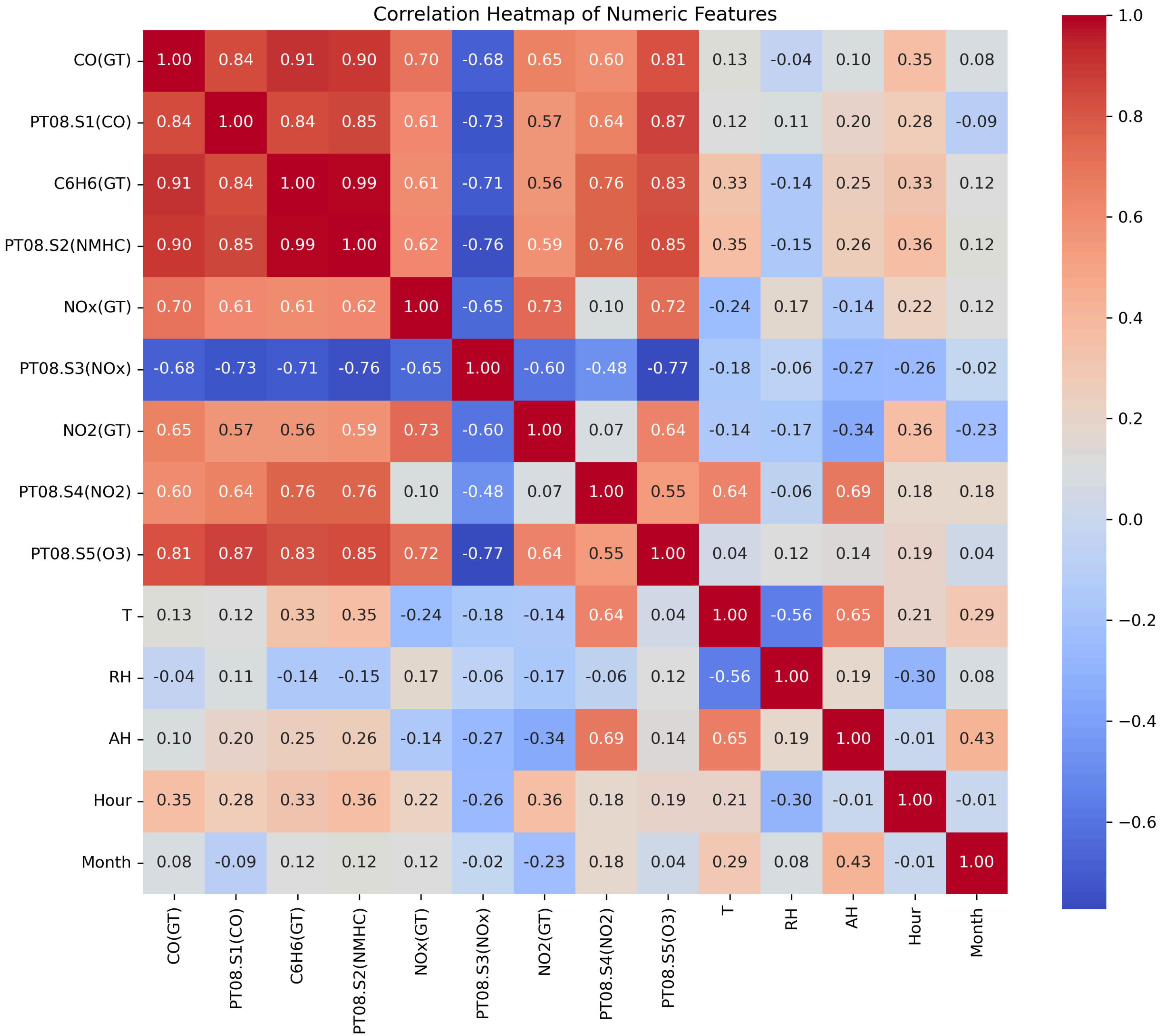
- August shows a marked decline in all pollutants, possibly from reduced industrial activity and better atmospheric dispersion. These trends highlight seasonal variation influenced by weather and human activity.



Correlation Analysis: Sensor Validity and Environmental Influence

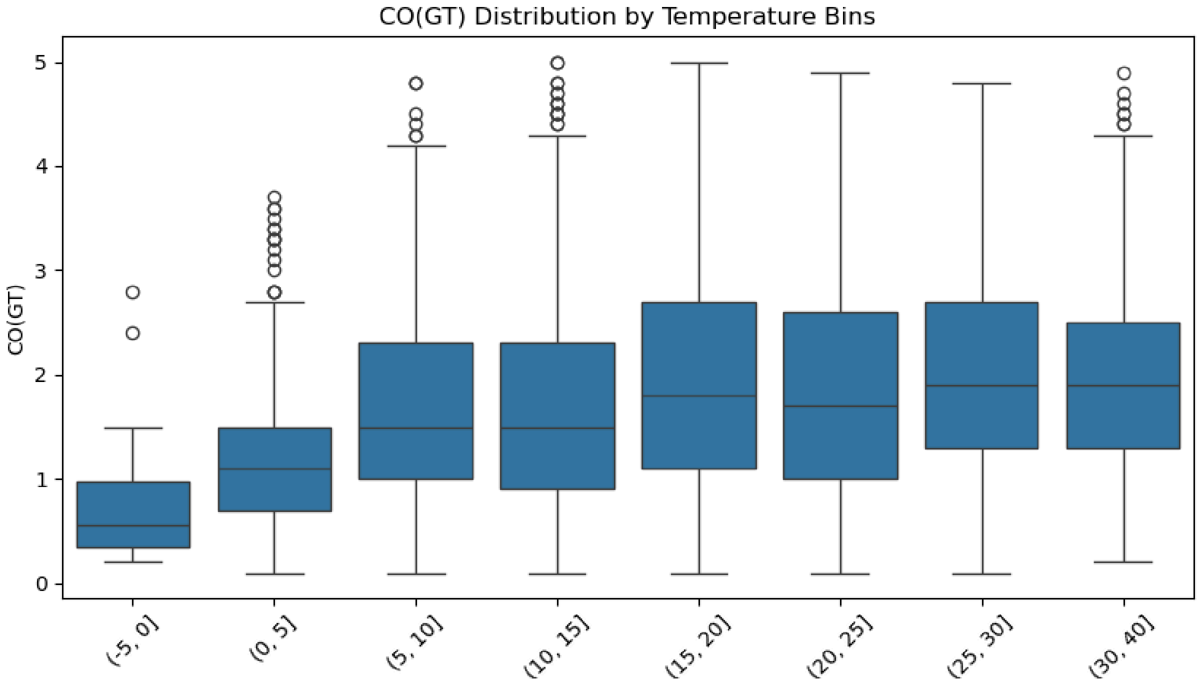
We analyzed how well each sensor correlates with its corresponding ground-truth pollutant. Key insights:

- **CO Sensor (PT08.S1):** Shows a **strong positive correlation** with CO(GT) (r = 0.84).
  → Indicates reliable performance and consistent tracking of CO levels.
- **NOx Sensor (PT08.S3):** Displays a **strong negative correlation** with NOx(GT) (r = –0.65).
  → This inverse relationship suggests a **serious calibration issue**, making the sensor unreliable for real-world measurements.
- **O_3 Sensor (PT08.S5):** Shows weak correlation with temperature (r = 0.04) and humidity (r = 0.12). Suggests that O_3 behavior in this dataset is influenced by additional environmental or chemical factors not captured here.

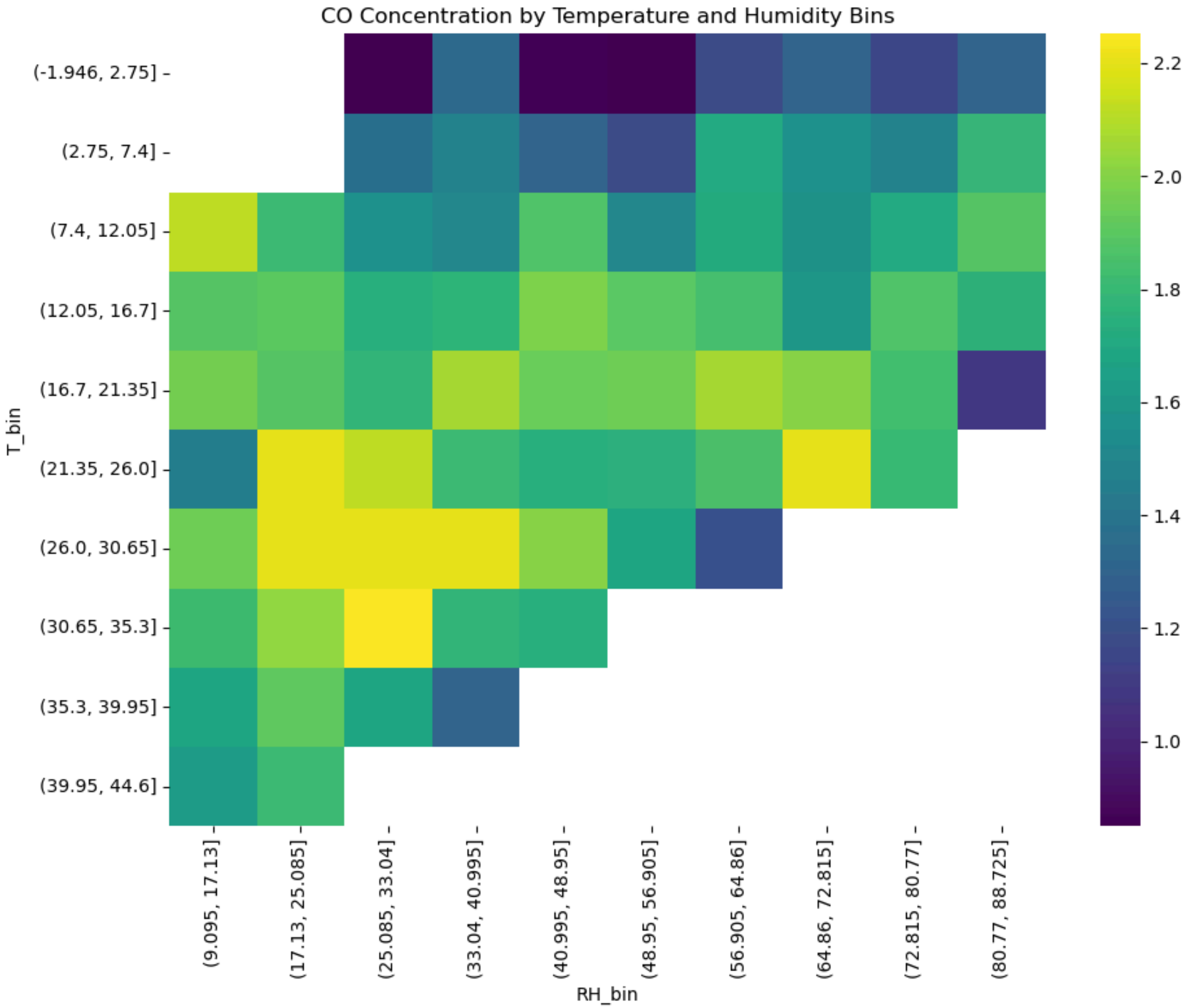

Correlation Heatmap of Numeric Features
CO(GT) 1.00 0.84 0.91 0.90 0.70 -0.68 0.65 0.60 0.81 0.13 -0.04 0.10 0.35 0.08
PT08.S1(CO) 0.84 1.00 0.84 0.85 0.61 -0.73 0.57 0.64 0.87 0.12 0.11 0.20 0.28 -0.09
C6H6(GT) 0.91 0.84 1.00 0.99 0.61 -0.71 0.56 0.76 0.83 0.33 -0.14 0.25 0.33 0.12
PT08.S2(NMHC) 0.90 0.85 0.99 1.00 0.62 -0.76 0.59 0.76 0.85 0.35 -0.15 0.26 0.36 0.12
NOx(GT) 0.70 0.61 0.61 0.62 1.00 -0.65 0.73 0.10 0.72 -0.24 0.17 -0.14 0.22 0.12
PT08.S3(NOx) -0.68 -0.73 -0.71 -0.76 -0.65 1.00 -0.60 -0.48 -0.77 -0.18 -0.06 -0.27 -0.26 -0.02
NO2(GT) 0.65 0.57 0.56 0.59 0.73 -0.60 1.00 0.07 0.64 -0.14 -0.17 -0.34 0.36 -0.23
PT08.S4(NO2) 0.60 0.64 0.76 0.76 0.10 -0.48 0.07 1.00 0.55 0.64 -0.06 0.69 0.18 0.18
PT08.S5(O3) 0.81 0.87 0.83 0.85 0.72 -0.77 0.64 0.55 1.00 0.04 0.12 0.14 0.19 0.04
T 0.13 0.12 0.33 0.35 -0.24 -0.18 -0.14 0.64 0.04 1.00 -0.56 0.65 0.21 0.29
RH -0.04 0.11 -0.14 -0.15 0.17 -0.06 -0.17 -0.06 0.12 -0.56 1.00 0.19 -0.30 0.08
AH 0.10 0.20 0.25 0.26 -0.14 -0.27 -0.34 0.69 0.14 0.65 0.19 1.00 -0.01 0.43
Hour 0.35 0.28 0.33 0.36 0.22 -0.26 0.36 0.18 0.19 0.21 -0.30 -0.01 1.00 -0.01
Month 0.08 -0.09 0.12 0.12 0.12 -0.02 -0.23 0.18 0.04 0.29 0.08 0.43 -0.01 1.00
CO(GT)
PT08.S1(CO)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
Hour
Month
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Temperature and Humidity Effects on CO

- CO levels increase with temperature until about **25°C**, then plateau.
- This supports the idea that **cool to mild temperatures trap pollutants** closer to the ground.
- Outliers in these bins may represent **morning traffic** or **evening inversions**.



- CO concentration is **not linearly dependent** on temperature or humidity.
- Highest values cluster around **20–30°C** and **40–70% RH**
- This suggests pollutants **accumulate under moderate weather conditions** — possibly due to human activity and limited dispersion.



Hypothesis Testing

Our statistical investigation was guided by six research hypotheses.
Each hypothesis aimed to explore a distinct environmental or technical factor based on air quality literature and sensor evaluation objectives.

- **H1: Weekday Rush Hour Impact**
  CO and NOx levels are significantly higher during weekday rush hours compared to non-peak hours, reflecting traffic congestion patterns.
- **H2: Seasonal Variation in Pollutants**
  Pollution levels for CO and NOx are elevated in winter months due to low atmospheric dispersion and increased combustion-related heating.

- **H3: Influence of Temperature and Humidity**
  Meteorological variables modulate pollutant behavior:
  - O_3 increases with temperature
  - CO and NOx concentrations decrease with relative humidity
- **H4: Weekly Trends in Benzene (C6H6)**
  Benzene concentrations vary significantly by day of the week, possibly due to variations in weekday industrial or traffic-related activity.
- **H5: Sensor–Ground Truth Agreement**
  Sensor readings should exhibit strong linear correlations with their corresponding pollutant ground-truth values if functioning properly.
- **H6: Detection of Sensor Faults**
  A significant negative or inconsistent correlation between a sensor and its reference value indicates possible sensor drift or miscalibration.

Hypothesis Test Results (1/2)

Hypothesis	Analytical Focus	Statistical Test	Result
H1	Difference in pollutant means between peak and non-peak hours	Independent t-test	Supported
H2	Comparison of pollutant means between winter and non-winter periods	Independent t-test	Supported
H3	Correlation of CO, NOx with RH; O_3 with Temperature	Pearson correlation	Not Supported (Weak)

Hypothesis Test Results (2/2)

Hypothesis	Analytical Focus	Statistical Test	Result
H4	Variation in benzene (C6H6) across weekdays	One-way ANOVA	Supported
H5	Correlation between sensor and ground-truth values	Pearson correlation	Partially Supported
H6	Detection of sensor anomaly (NOx sensor)	Pearson correlation	Supported

Key Findings

This study provided a data-driven investigation into urban air pollution using real-world sensor data.
Key conclusions include:

- **Pollutant levels peak during rush hours and winter months**, increasing health risks for urban populations.
- **CO and Benzene sensors performed reliably**, while the **NOx sensor showed significant calibration issues**.
- Weather conditions such as temperature and humidity influence pollutant levels, though the correlations are modest.
- **Statistical testing confirmed key patterns**, supporting targeted interventions in urban traffic and sensor monitoring.

Overall, this project demonstrates how environmental data can be used to guide public health policies and smarter city planning.

Public Health Implications

Our analysis highlights specific time periods and environmental conditions where air quality poses elevated health risks:

- **Weekday rush hours** (8 AM and 6 PM) show sharp increases in CO and NOx levels
- **Winter months** exhibit sustained high pollution due to heating and stagnant air
- These patterns raise concerns for:
    - **Children**, who may commute during peak hours
    - **Elderly individuals** and **asthma patients**, who are more vulnerable to pollutants
- **Unreliable sensors**, especially for NOx, pose a risk for delayed response and misinformed policy decisions

Policy Recommendations

Based on our findings, we propose the following interventions to improve urban air quality and protect public health:

- **Traffic Management**
  - Implement congestion pricing or restrict traffic during peak hours
  - Promote public transportation and carpooling incentives
- **Urban Planning**
  - Avoid placing schools and hospitals near major roads
  - Enforce zoning laws that consider air quality patterns
- **Sensor Maintenance**
  - Establish calibration protocols for all air quality sensors, particularly for NOx
  - Deploy multiple sensors per location to detect faults early
- **Public Awareness**
  - Promote awareness campaigns around peak pollution times and seasons

Future Work

To extend the impact of this study, future research can explore:

- **Spatio-temporal analysis** using data from multiple locations or mobile sensors
- **Integration with healthcare data** to study links between pollution exposure and hospital admissions
- **Machine learning models** to predict pollutant surges and detect sensor anomalies
- **City-scale pollution mapping** with GIS tools and real-time feeds
- **Simulation of policy scenarios** to estimate the effect of interventions like car bans or zoning reforms

Thank You!

Group 28 – Spring 2025