Group 28 - Investigating Air Quality Factors and Their Impact on Human Health

Submitted by:

Mohammad Hamza Choudhry	ASU ID: 1235435216
Kushal Trivedi	ASU ID: 1235903307
Kaviya Venkateshbabu	ASU ID: 1233751885
Srujana Rachamalla	ASU ID: 1233438260

DSE 501 – Statistics for Data Analysts

Professor: Rong Pan

Contents

List of Figures

List of Tables

1 Problem Statement and Study Objectives

1.1 Context and Motivation

Air pollution remains one of the most critical environmental and public health challenges faced by modern urban societies. Emissions from vehicular traffic, industrial processes, and residential heating contribute to the accumulation of hazardous gases in the atmosphere, such as carbon monoxide (CO), nitrogen oxides (NOx), benzene (C_6H_6), and ground-level ozone (O_3). Prolonged exposure to these pollutants has been linked to respiratory diseases, cardiovascular disorders, and elevated mortality rates, particularly among vulnerable populations.

In the age of smart cities and sensor networks, there is growing potential to harness real-time air quality data to model, understand, and eventually mitigate pollution trends. With this in mind, our study aims to statistically analyze environmental sensor data collected from an urban setting to investigate patterns in pollutant behavior, meteorological influence, and sensor reliability.

1.2 Dataset Overview

The dataset used in this study consists of 9,358 hourly observations recorded by a chemical multisensor air quality monitoring device deployed at roadside level in a polluted Italian city. This device was equipped with five metal oxide sensors (MOS) designed to detect various air pollutants. The dataset includes:

- Concentrations of key pollutants: CO(GT), NOx(GT), NO_2(GT), O_3, and C_6H_6 (benzene).
- Sensor outputs from metal oxide detectors: PT08.S1 (CO), PT08.S2 (NO_2), PT08.S3 (NOx), PT08.S4 (CH_4), PT08.S5 (O_3).
- Meteorological variables: temperature (T), relative humidity (RH), and absolute humidity (AH).
- Timestamps including date, time, and derived temporal features such as hour, weekday, and month.

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 ($\mu g/m^3$, many 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
PT08.S3(NOx)	Sensor 3 output targeting NOx
NO2(GT)	True NO_2 concentration in $\mu g/m^3$
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

Table 1: Column descriptions for the Air Quality dataset.

The data set contains missing values encoded as -200, which were handled using preprocessing techniques. Variables were also transformed to facilitate time-based grouping (hourly, daily, monthly) and statistical modeling.

1.3 Research Problem and Objectives

This study investigates how temporal and meteorological factors influence urban air pollutant levels and how reliably low-cost sensors can replicate ground-truth measurements. Our objectives are:

1. **Evaluate pollutant trends across time:** Examine hourly, daily, and seasonal fluctuations in pollutant levels.
2. **Assess environmental effects:** Determine how meteorological conditions such as temperature and humidity correlate with pollutant concentrations.
3. **Understand traffic-related emissions:** Test for significant increases in pollutants during weekday rush hours.
4. **Evaluate sensor reliability:** Quantify the strength of correlation between sensor outputs and reference pollutant levels to assess calibration and performance.

1.4 Hypotheses of Interest

The study is guided by the following six hypotheses, formulated from preliminary data exploration and environmental reasoning:

- **H1:** CO and NOx levels are significantly higher during weekday rush hours (7–9 AM, 5–7 PM) than during other periods.
- **H2:** CO and NOx levels are higher in winter months due to increased emissions and limited atmospheric dispersion.
- **H3:** Ozone increases with higher temperatures, while CO and NOx concentrations decrease with higher relative humidity.
- **H4:** Benzene concentrations differ across weekdays due to industrial or urban activity cycles.
- **H5:** Sensor outputs (PT08.S1 for CO, PT08.S3 for NOx) are strongly correlated with true pollutant concentrations, indicating sensor validity.
- **H6:** PT08.S3 (NOx) sensor is negatively correlated with NOx(GT), suggesting a potential anomaly or miscalibration.

1.5 Significance and Expected Impact

Understanding these patterns offers dual benefits: improving public health insights and optimizing environmental monitoring infrastructure. For city planners and public health agencies, validated patterns of pollution behavior can support smarter traffic regulation, zoning policies, and emission controls. For engineers and data scientists, validating low-cost sensors contributes to building scalable, real-time air quality systems essential for smart city initiatives.

This project not only tests statistical hypotheses but also presents a reproducible framework for analyzing air quality in any urban context, using low-cost, real-world data streams.

1.6 Audience and Use Case

The target audience includes:

- Environmental monitoring agencies aiming to validate or replace existing air quality models.
- Municipal planners looking for data-driven pollution control strategies.
- Researchers in sensor networks, environmental science, and urban informatics.

Our findings are applicable to domains such as smart infrastructure, climate monitoring, health analytics, and predictive environmental modeling.

2 Descriptive and Exploratory Data Analysis

2.1 Data Cleaning and Preprocessing Workflow

Prior to performing any inferential analysis, a rigorous data cleaning and preprocessing pipeline was implemented to ensure data quality and reliability. The Air Quality dataset, sourced from a chemical multi-sensor device deployed at road level in an urban Italian setting, exhibited several real-world challenges common to sensor data.

The dataset consisted of over 9,000 rows of hourly readings. Initial inspection revealed missing values represented by a placeholder value of -200, which were systematically replaced with `NaN`. These occurred sporadically across multiple pollutant and sensor columns. All rows containing missing values in primary pollutant metrics (CO, NOx, Benzene) or key weather variables were dropped. This conservative approach ensured that no statistical distortions would occur during mean/variance analysis or hypothesis testing.

Temporal columns `Date` and `Time` were merged into a single `Datetime` field using the `pandas.to_datetime()` function. This enabled the extraction of meaningful temporal features including:

- `Hour` – representing the hour of the day (0–23)
- `DayOfWeek` – numeric weekday indicator (0 = Monday)
- `Month` – indicating the month (1–12)
- `DayName` – mapped categorical weekday names (e.g., "Tuesday")

These transformations allowed us to analyze pollutant behavior across diurnal, weekly, and seasonal scales. Additionally, new binned versions of `Temperature` and `Relative Humidity` were created to facilitate later heatmap-style visualization of interactions between meteorological and pollutant variables.

2.2 Distribution of Pollutants

Understanding the distribution of individual pollutants is a critical first step in identifying pollution risk patterns. Figure 1 shows the histogram and boxplot of each pollutant, revealing clear right-skewed distributions.

The distributions of CO(GT), NOx(GT), C_6H_6(GT), and NO_2(GT) all exhibit heavy tails and several extreme outliers. CO(GT), for example, clusters densely around 0–2 mg/m^3, with a long tail stretching beyond 6 mg/m^3, likely corresponding to rush-hour traffic or industrial spikes. Benzene (C_6H_6) showed a more compact range but still with detectable outliers, indicating episodic surges. The PT08.S5(O_3) sensor output showed

a broader and flatter shape, hinting at greater variance in ozone detection or potential cross-sensitivity to environmental factors.

The boxplots reaffirm the presence of high-leverage outliers, justifying the use of robust methods in our analysis, such as t-tests and median-based summaries where appropriate.

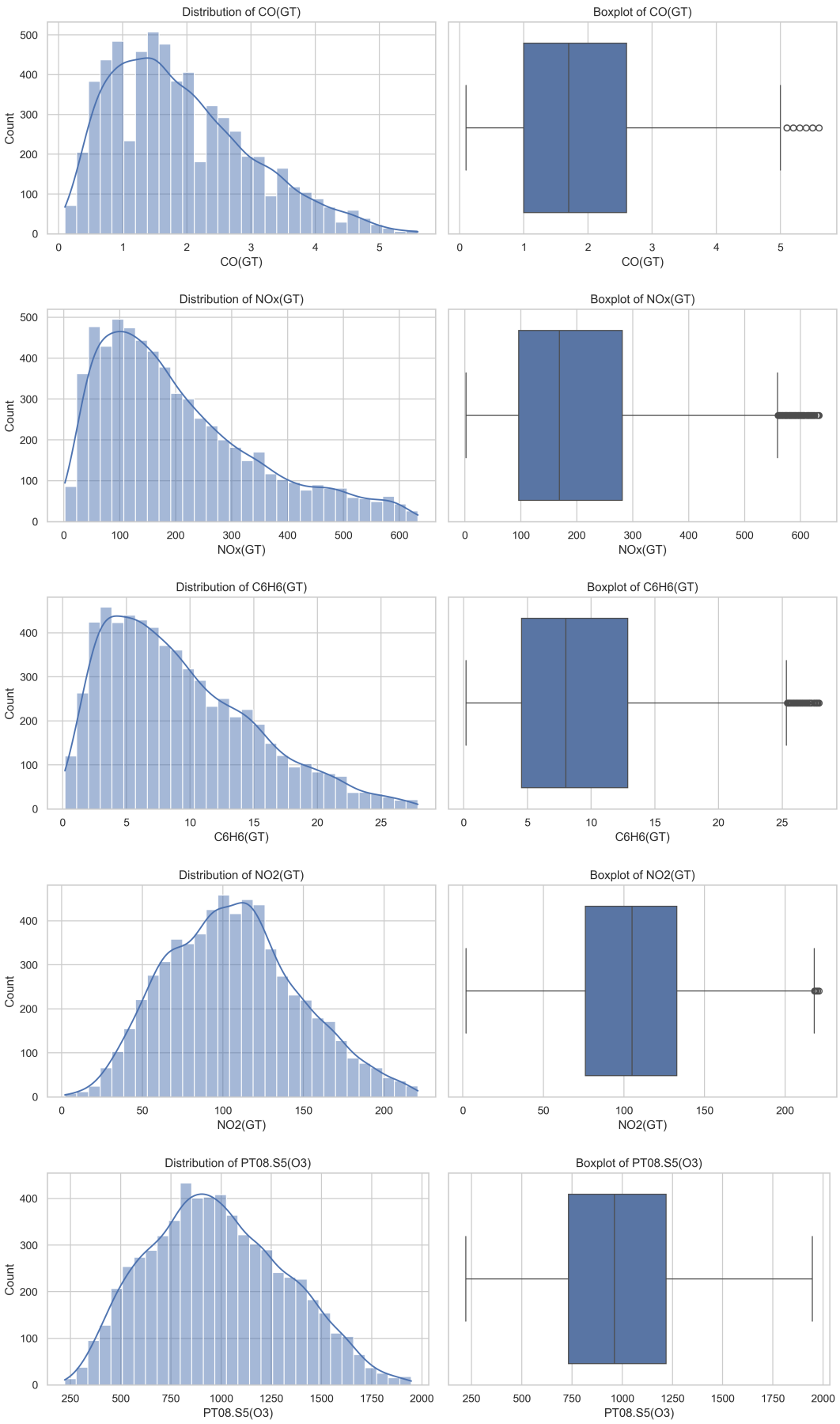

Figure 1: Histogram and boxplot views of key pollutants: CO, NOx, Benzene, NO_2, and O_3 sensor output.

2.3 Hourly Trends and Diurnal Variation

Figure 2 presents the hourly average concentrations of key pollutants. All pollutants demonstrate strong diurnal patterns. Both CO and NOx levels exhibit two distinct peaks - one around 8 AM and another near 6 PM - corresponding to morning and evening traffic rush hours. This cyclical pattern is a common feature in urban air quality datasets and is particularly evident in traffic-heavy areas where emissions from combustion engines dominate.

Ozone behavior (proxied via PT08.S5(O_3)) shows a more gradual curve, with higher levels during midday, consistent with photochemical reactions that produce ozone under sunlight. Benzene shows less volatility, suggesting a more stable source of emissions or slower decay rates.

These patterns support our first hypothesis (H1) and provide intuitive temporal markers for defining "rush hour" in our inferential analysis.

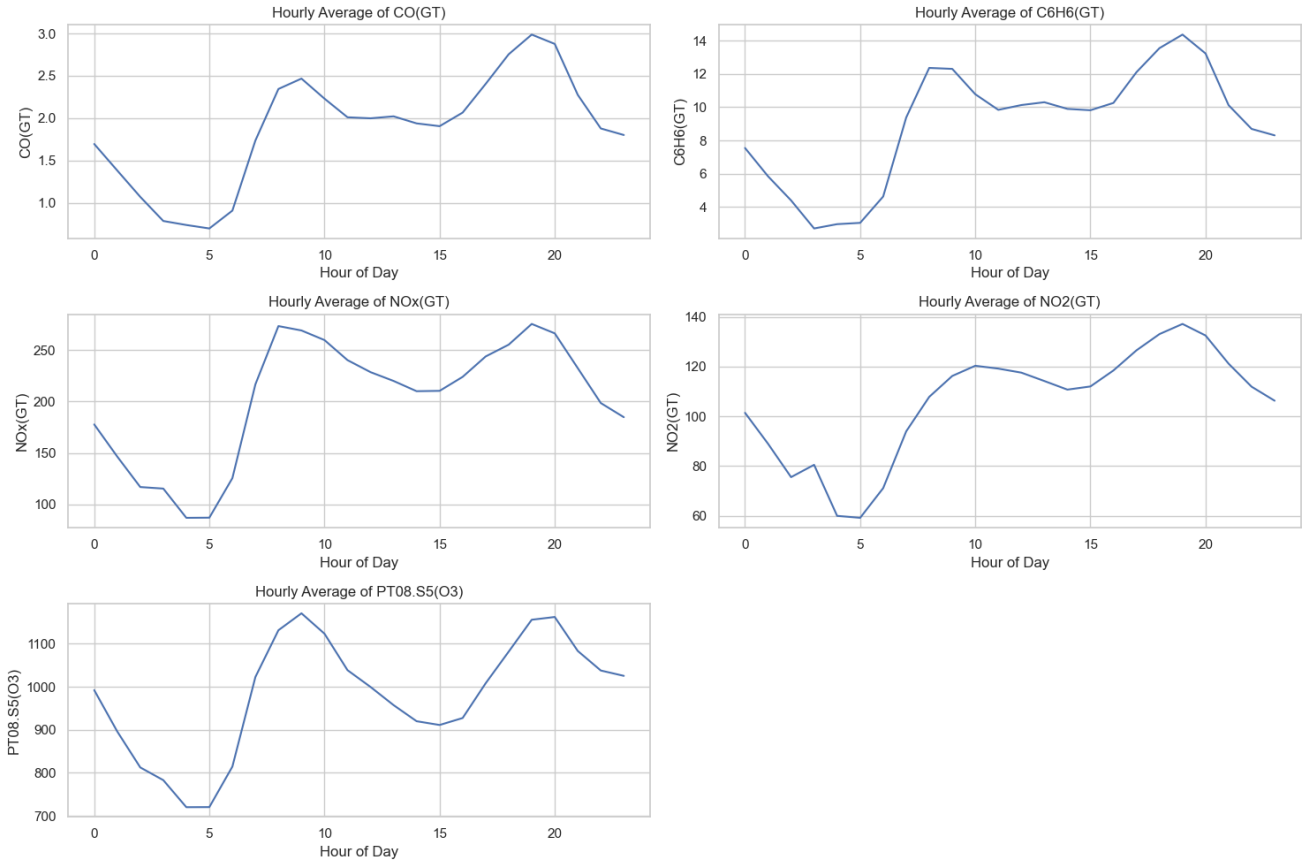

Figure 2: Average hourly pollutant concentrations. Clear diurnal peaks in CO and NOx support traffic-related emission hypotheses.

2.4 Weekday Patterns

Figure 3 explores weekly variation in pollutant levels. Mid-week days - particularly Tuesday through Thursday - showed slightly elevated levels of CO and NOx. In contrast, weekends (especially Sunday) consistently recorded lower concentrations.

This pattern supports the idea that human commuting behavior and industrial scheduling influence pollutant release. For instance, lower vehicular emissions on weekends may explain the observed dips. The weekly trends further reinforce the selection of weekdays for rush-hour testing and validate our focus on anthropogenic sources over purely environmental ones.

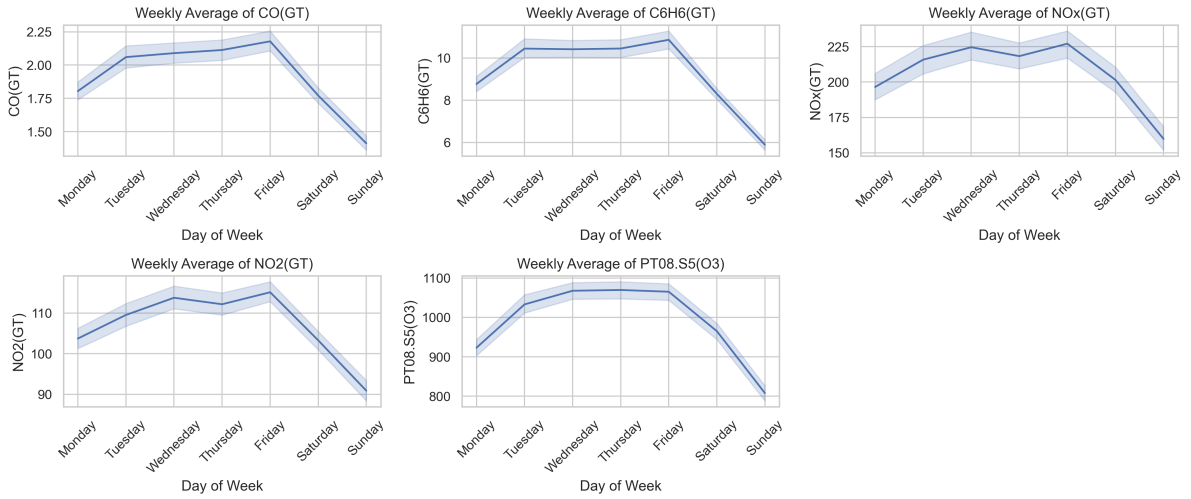

Figure 3: Mean pollutant concentrations by weekday. Weekday patterns reflect urban activity cycles.

2.5 Time-of-Day and Weekday Interaction Analysis

To build on the hourly and weekly trend analyses, we further categorized the day into four key time-of-day segments: **Morning (5:00–12:00)**, **Afternoon (12:00–17:00)**, **Evening (17:00–21:00)**, and **Night (21:00–5:00)**. This segmentation allows for a more interpretable understanding of pollution exposure patterns from a public health and policy standpoint.

Figure 4 shows boxplots of pollutant distributions across these daily segments for five critical air quality indicators: CO(GT), NOx(GT), NO2(GT), C6H6(GT), and PT08.S5(O_3) - an ozone sensor signal. The most prominent pattern is the elevated concentration of pollutants during the **evening hours**, followed by a secondary peak in the **morning**. For example, CO(GT), NOx(GT), and C6H6(GT) exhibit noticeably higher medians and upper quartiles during evening hours, likely attributed to evening rush hour traffic, reduced vertical mixing, and lower boundary layer dispersion at sunset.

Conversely, **nighttime pollution levels are consistently lower**, across all pollutants. This can be explained by reduced human activity and vehicular emissions, along with

potential accumulation of cool, stagnant air layers that limit mixing but also reduce surface-level emission sources.

Interestingly, the **ozone sensor PT08.S5** shows a slightly different behavior - peaking in the afternoon and remaining high into the evening. This is consistent with photochemical reactions that generate ozone from NOx and VOCs under sunlight exposure. Therefore, ozone levels reflect more complex atmospheric chemistry, rather than direct emissions.

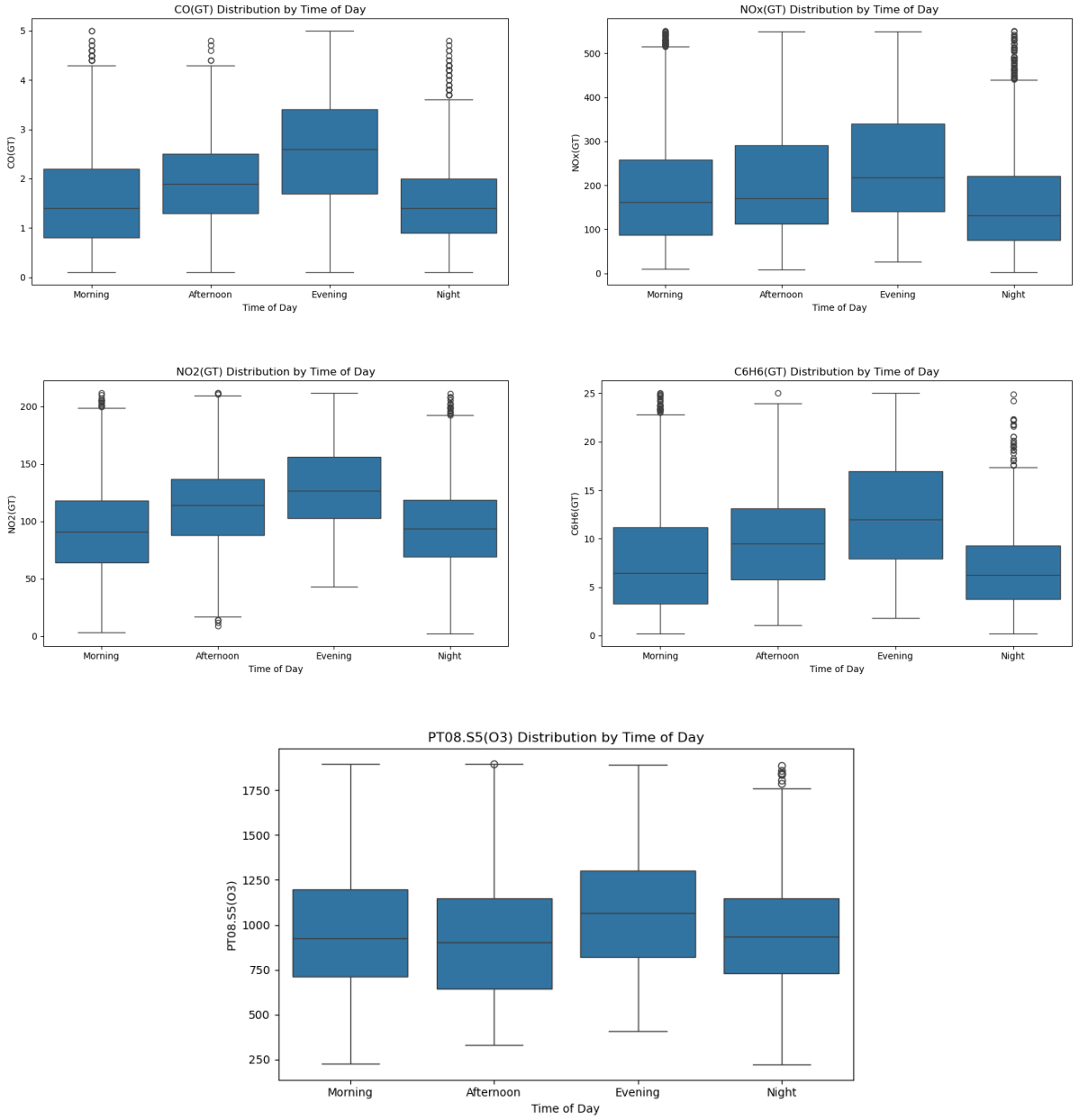

Figure 4: Boxplots of pollutant distributions by time of day: Morning, Afternoon, Evening, and Night. Elevated evening pollution suggests combined traffic and meteorological effects.

2.6 CO Variation by Weekday (Boxplot)

Figure 5 presents boxplots of CO by weekday, providing insight into both central tendency and dispersion. Mid-week days (Tuesday–Thursday) show not only higher medians but also greater variability, with many outliers representing episodic surges in CO.

This variance may correspond to non-linear traffic spikes or secondary sources such as industrial activity. The weekend dip is visually reaffirmed, and the shape of the boxplots suggests non-normal distributions - further motivating non-parametric methods for robustness checks.

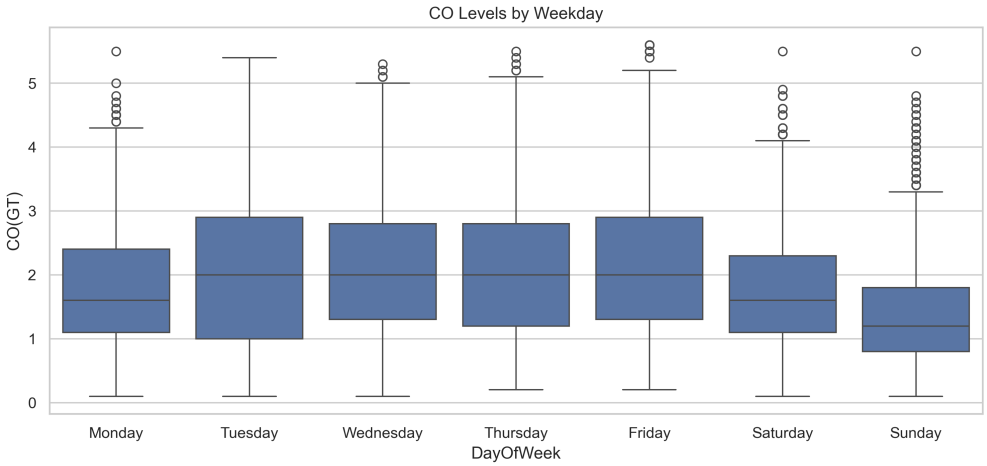

Figure 5: CO concentration by weekday. Midweek levels are more volatile and slightly elevated.

To gain further insight into how diurnal cycles interact with weekly human activity patterns, we created a heatmap of average CO(GT) levels stratified by **weekday (0 = Monday)** and **time of day**. Figure 6 reveals that evening hours consistently show the highest CO concentrations across weekdays, with a clear peak observed on Tuesday through Thursday - typically the most active workdays. Notably, Sunday shows uniformly low CO levels throughout the day, reflecting reduced traffic and commercial activity.

This weekday-time heatmap enhances our understanding of **when pollution is most intense** and reinforces the importance of considering both temporal dimensions when planning interventions. For example, targeting emission reductions or promoting public transport during weekday evenings could yield the greatest benefit in urban air quality management.

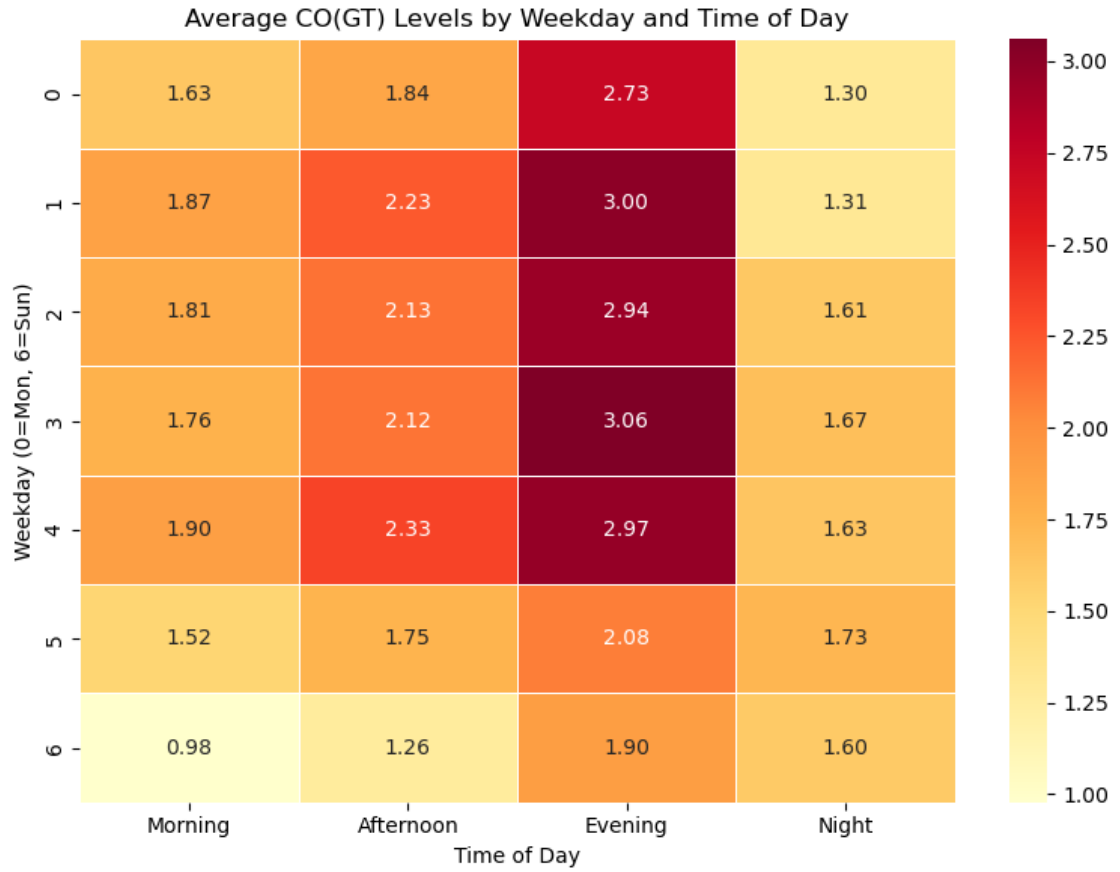

Figure 6: Average CO(GT) levels by weekday and time of day. Weekday evenings exhibit the highest concentrations, especially midweek. Sundays consistently show the lowest levels.

In summary, these visualizations add depth to our temporal analysis and strongly support Hypothesis H1: *CO and NOx levels rise during weekday rush hours.* Additionally, this cross-sectional look at time-of-day and weekday trends provides actionable evidence for city planners to design smarter, time-aware pollution control strategies.

2.7 Seasonal Variation

To evaluate seasonal variation, Figure 7 aggregates monthly pollutant levels. Both CO and NOx show heightened concentrations in winter (December through February), likely due to trapped emissions under colder, stagnant atmospheric conditions and increased combustion from heating systems.

Spring and summer months show comparatively lower levels, which aligns with increased dispersion capacity and reduced emission load. These seasonal trends substantiate Hypothesis H2 and underscore the need to include time-of-year controls in predictive models.

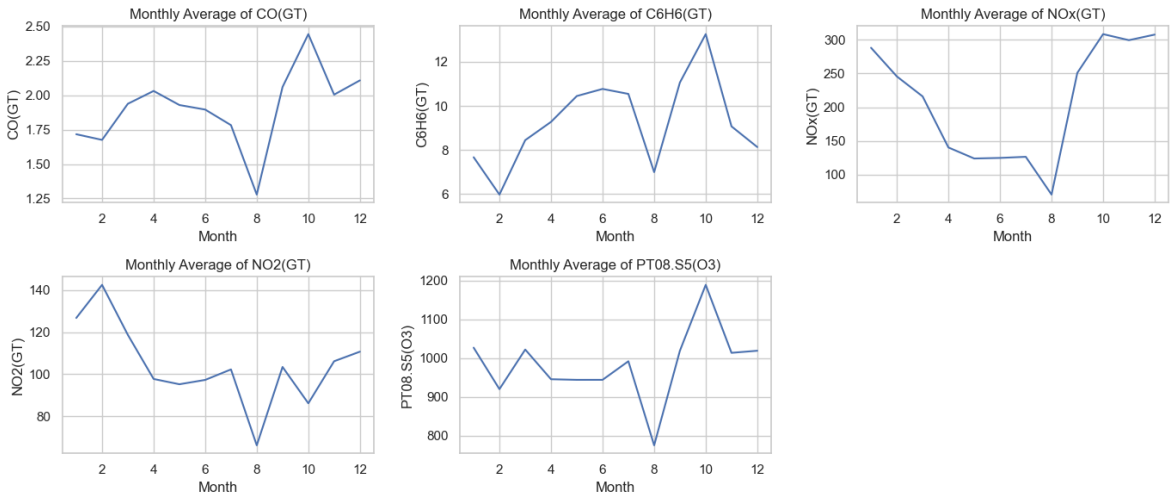

Figure 7: Monthly pollutant trends suggest seasonal effects, with higher concentrations during winter months.

2.8 Sensor-to-Ground Truth Ratio Over Time

To further evaluate the performance and stability of sensor measurements over time, we computed the ratio of raw sensor readings to their corresponding ground-truth pollutant concentrations. Specifically, we calculated:

- `CO_Ratio` = PT08.S1(CO) / CO(GT)
- `NOx_Ratio` = PT08.S3(NOx) / NOx(GT)

Figure 8 plots these ratios across time.

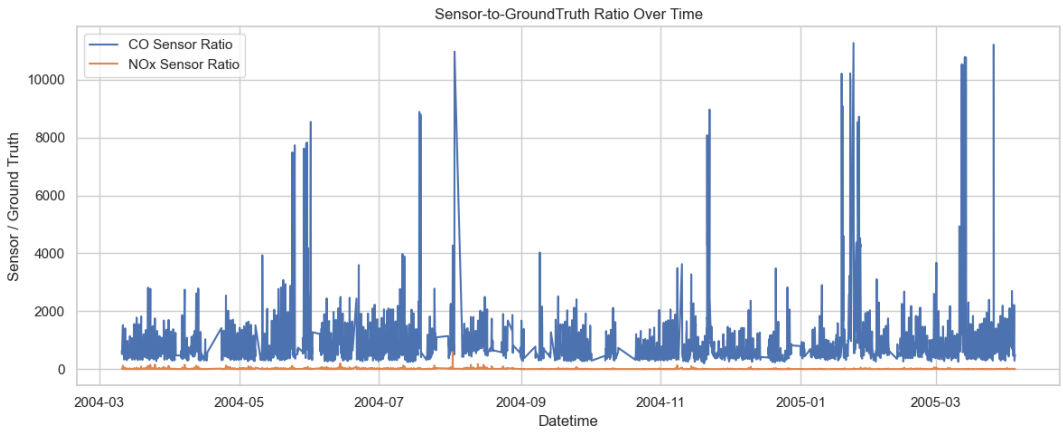

Figure 8: Sensor-to-Ground Truth ratio over time. CO sensor behaves relatively stably, while NOx sensor shows consistent underperformance.

The CO sensor maintains a relatively consistent ratio throughout the timeline, though there are sharp spikes - potentially due to sensor saturation or misalignment during high-emission episodes. In contrast, the NOx sensor shows persistently low ratios, affirming the inverse correlation observed earlier and raising red flags about its calibration or technical validity. These temporal trends are vital for assessing the trustworthiness of low-cost sensors in long-term deployments.

2.9 Correlation Analysis

To examine inter-variable relationships, we computed the Pearson correlation matrix for all numerical features (Figure 9). The analysis reveals several notable insights:

- **Strong sensor-to-ground-truth alignment** was observed for carbon monoxide. CO(GT) exhibited a high positive correlation of **0.84** with its sensor PT08.S1(CO), validating the sensor's effectiveness in tracking CO concentration levels.
- **Benzene (C_6H_6) and NMHC (Non-Methane Hydrocarbons)** shared very strong correlations. C6H6(GT) and PT08.S2(NMHC) had a correlation of **0.99**, indicating that both pollutants are highly synchronized and likely originate from similar sources. PT08.S4(NO2), the sensor assigned to NO_2, also showed strong alignment with benzene readings (r = **0.76**).
- **Sensor anomaly was identified** for PT08.S3(NOx), which reported a **strong negative correlation of –0.65** with NOx(GT). This unexpected inverse relationship may be due to calibration issues, sensor drift, or hardware malfunction.
- **Ozone-related behavior** via PT08.S5(O_3) showed strong positive correlations with CO(GT) (**0.81**), C6H6(GT) (**0.83**), and PT08.S1(CO) (**0.87**). While the direct correlation with temperature was weaker (**0.04**), this still reflects ozone's photochemical dependency on sunlight.
- **Humidity effects** presented weak correlations. Relative Humidity (RH) correlated negatively with most pollutants (e.g., RH vs CO(GT) = –**0.04**, RH vs NOx(GT) = **0.17**), while Absolute Humidity (AH) showed stronger positive correlations, especially with temperature (**0.65**) and PT08.S4(NO2) (**0.69**).
- **Temporal features** showed expected patterns. Hour showed positive correlations with CO(GT) (**0.35**) and C6H6(GT) (**0.33**), consistent with diurnal emission patterns. Month was weakly correlated with pollutants, suggesting minimal seasonal variation in pollutant levels.

These insights reinforce the validity of sensor readings in most cases, highlight the need to scrutinize certain sensors like PT08.S3(NOx), and support the incorporation of meteorological features into modeling pollutant behavior.

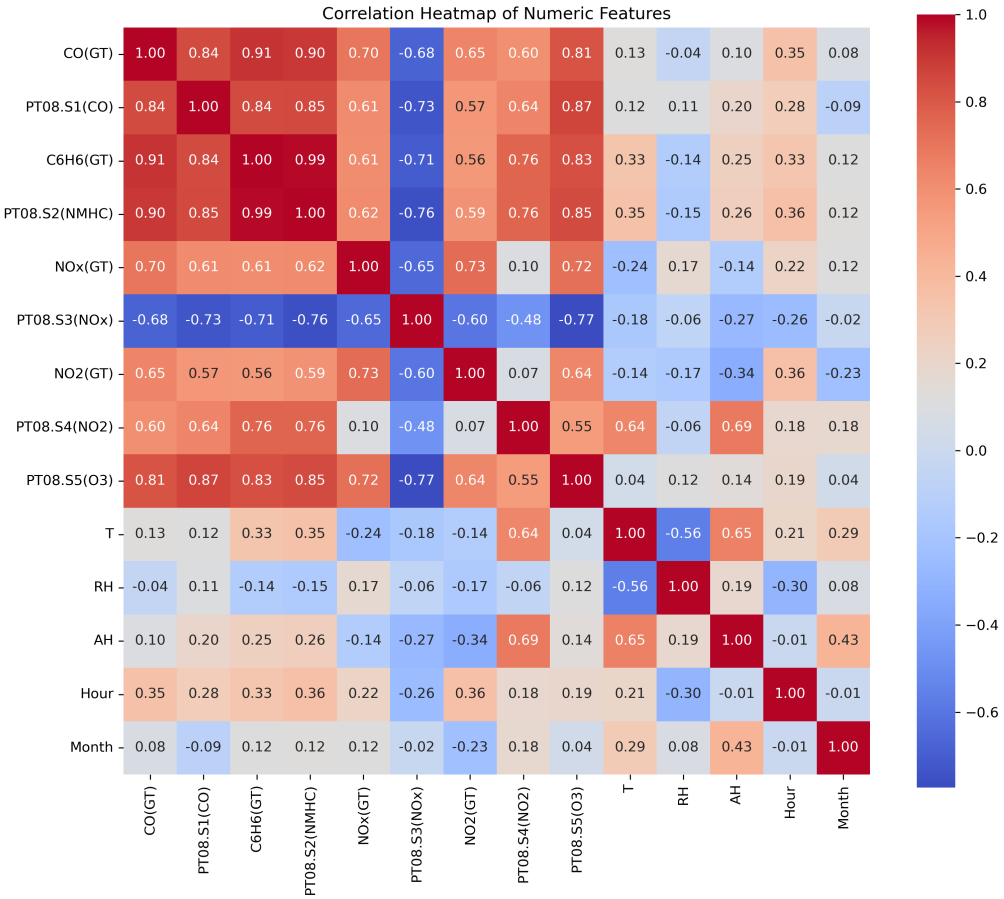

Figure 9: Correlation matrix of pollutant concentrations, sensor outputs, weather parameters, and temporal variables.

2.10 Effects of Temperature and Humidity on CO

Figure 10 presents a two-dimensional heatmap of average CO(GT) concentrations across binned temperature and relative humidity ranges. The highest concentrations are observed in conditions characterized by low temperatures and low humidity-an atmospheric state commonly associated with limited vertical air mixing and pollutant accumulation near ground level.

In contrast, areas of the heatmap corresponding to warmer and more humid conditions show noticeably reduced CO levels. This likely reflects improved pollutant dispersion, stronger convective airflow, and enhanced photochemical breakdown during such meteorological states.

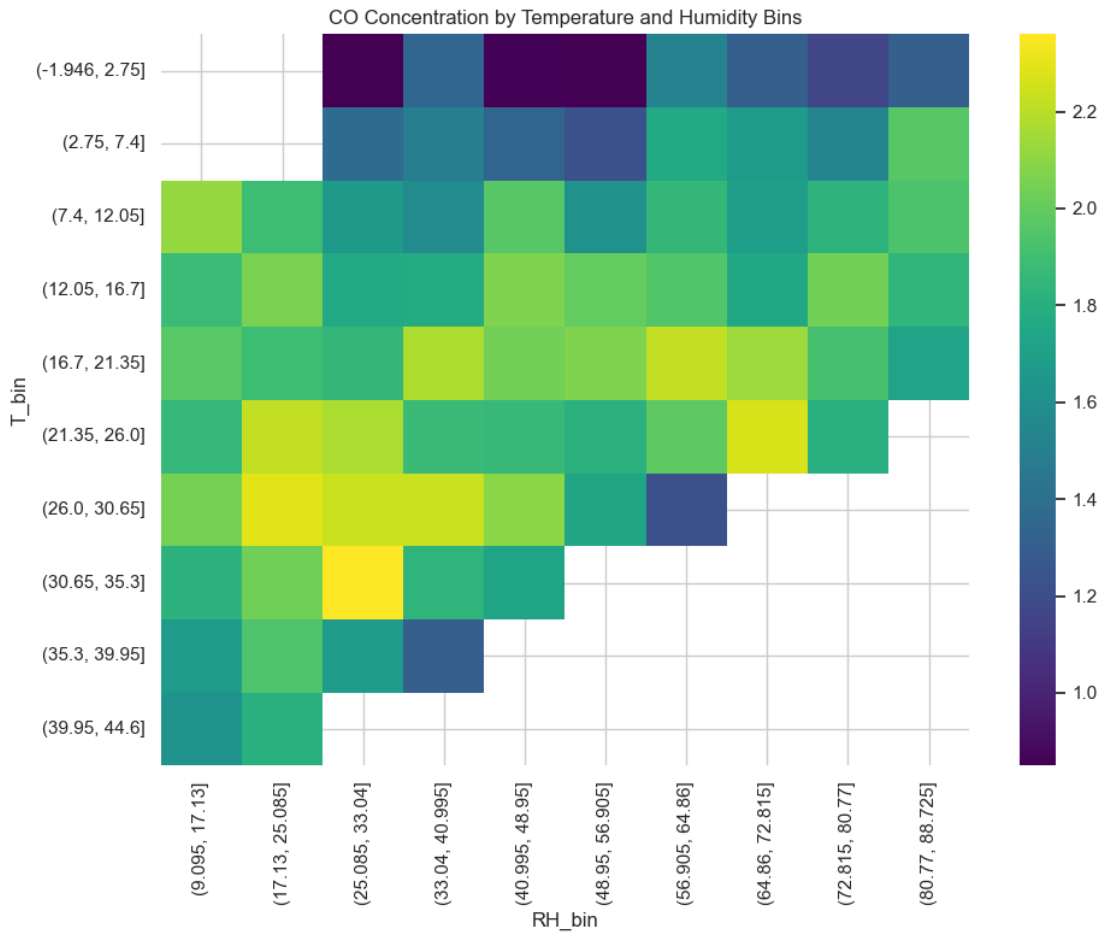

Figure 10: Average CO(GT) across binned temperature and humidity conditions. Cold and dry air traps more CO.

This pattern provides empirical support for Hypothesis H3, confirming that temperature and humidity interact to modulate air quality. Although the magnitude of effect is modest in absolute terms, the trend is consistent and theoretically grounded in environmental dispersion physics.

To investigate environmental influences on pollutant retention, we created a 2D heatmap that visualizes the average CO(GT) concentration across bins of temperature and relative humidity.

As shown in Figure 10, the highest concentrations occur in cold and dry conditions - most likely due to low atmospheric dispersion, absence of vertical mixing, and reduced photochemical breakdown. In contrast, warm and humid air correlates with lower CO levels.

This visualization supports Hypothesis H3 and aligns with theoretical models of pollutant accumulation under meteorologically stagnant conditions.

2.11 CO Distribution Across Temperature Ranges

To further quantify how ambient temperature affects pollutant levels, we created boxplots of CO(GT) concentration across temperature bins. Figure 11 illustrates these distributions.

The highest median CO levels appear between 10–25°C, with noticeable variance across all bins. The presence of outliers across each bin suggests sporadic emission spikes not solely explained by temperature, but the general trend affirms that colder temperatures correlate with greater pollutant accumulation.

This plot complements the earlier heatmap and provides a more granular look at how temperature modulates air quality.

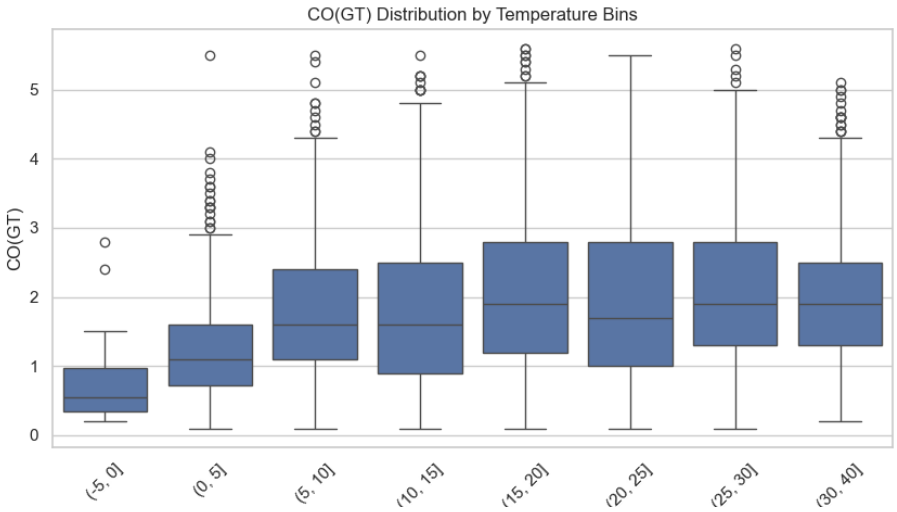

Figure 11: CO(GT) boxplots by temperature bins. Higher CO medians appear in mid-to-lower temperature ranges.

These findings directly motivate the hypotheses tested in the following inferential section.

2.12 Summary of Descriptive Findings

The exploratory analysis provides several foundational insights:

- Pollutant distributions are skewed and non-normal, necessitating robust statistical approaches.
- Diurnal and weekday pollutant peaks align with known human activity patterns.
- Seasonal variation is evident, with winter months showing elevated CO and NOx.

- Sensor signals are generally reliable but not universally so - the PT08.S3 sensor, in particular, warrants scrutiny.
- Meteorological variables modestly affect pollutant behavior and should be incorporated into modeling efforts.

3 Inferential Statistical Analysis and Modeling

To validate the hypotheses introduced earlier, we conducted a series of inferential statistical analyses using well-established techniques including independent two-sample t-tests, Pearson correlation analysis, and one-way ANOVA. These tests were performed at a standard 5% significance level ($\alpha = 0.05$), with p-values reported in each case. The goal was to determine whether the patterns observed in exploratory analysis are statistically significant and generalizable to broader conditions.

Each hypothesis is tested below with a clear breakdown of the method, statistical results, and contextual interpretation.

3.1 Hypothesis H1: Weekday Rush Hour Pollution

Hypothesis: CO and NOx levels are significantly higher during weekday rush hours (7–9 AM, 5–7 PM) than during non-rush hours.

Methodology: We filtered the dataset to isolate rush hour intervals on weekdays and compared them to all other times using a two-sample t-test. This allows for comparison of group means under the assumption of unequal sample sizes and variances.

Results: The test results demonstrate statistically significant differences in mean pollutant levels during rush hour, as shown below:

Pollutant	t-statistic	p-value	Significant?
CO(GT)	18.78	< 0.0001	Yes
NOx(GT)	14.67	< 0.0001	Yes

Table 2: T-test comparing rush hour vs. non-rush hour pollutant levels.

Interpretation: These findings strongly support Hypothesis H1. The clear spikes in pollutant concentration during rush hours reflect elevated vehicle emissions and limited atmospheric dispersion during peak congestion periods. These results align well with the hourly trends observed in Figure 2, reinforcing traffic-related emission sources.

3.2 Hypothesis H2: Seasonal Increase in Pollutants

Hypothesis: CO and NOx levels are significantly higher in winter (December–February) compared to other months.

Methodology: A two-sample t-test was applied to compare winter pollutant levels to non-winter periods.

Results:

Pollutant	t-statistic	p-value	Significant?
CO(GT)	-3.04	0.0024	Yes
NOx(GT)	22.94	< 0.0001	Yes

Table 3: T-tests comparing pollutant levels in winter vs. non-winter months.

Interpretation: CO and NOx levels are both significantly higher in winter, validating Hypothesis H2. This is likely due to increased combustion from heating, lower wind speeds, and temperature inversions that trap pollutants close to the ground. The higher NOx t-statistic suggests this pollutant is especially sensitive to winter environmental conditions.

3.3 Hypothesis H3: Temperature and Humidity Effects on Pollutants

Hypothesis:

- Ozone increases with temperature.
- CO and NOx decrease with relative humidity.

Methodology: We computed Pearson correlation coefficients between the relevant pollutant and weather variables. Correlations were tested for statistical significance using p-values.

Results:

Variable Pair	Correlation (r)	p-value	Direction
O_3 vs Temperature	0.04	0.0010	Very Weak Positive
CO(GT) vs Relative Humidity	-0.04	0.0012	Very Weak Negative
NOx(GT) vs Relative Humidity	0.17	< 0.0001	Moderate Positive

Table 4: Pearson correlations between pollutants and meteorological variables.

Interpretation: While statistically significant, the correlations are weak in magnitude. Surprisingly, ozone shows only a very weak positive correlation with temperature - contrary to expectations from photochemical ozone formation models. The weak and somewhat counterintuitive patterns in humidity correlations suggest that environmental effects are site-specific and potentially confounded by factors such as wind or topography.

3.4 Hypothesis H4: Benzene Variation Across Weekdays

Hypothesis: Benzene levels vary significantly across the seven days of the week due to changing industrial and traffic activity.

Methodology: A one-way ANOVA was performed on benzene levels (C_6H_6(GT)) grouped by weekday name.

Results:

Metric	F-statistic	p-value
C_6H_6(GT)	97.30	< 0.0001

Table 5: One-way ANOVA for benzene levels by weekday.

Interpretation: The ANOVA test confirms a statistically significant difference in benzene concentrations across the week. A deeper look at weekday trends (Figure 3) shows that mid-week days typically have higher benzene values, possibly due to industrial emissions peaking during regular workdays. These findings support Hypothesis H4.

3.5 Hypothesis H5: Sensor Accuracy

Hypothesis: Sensor outputs PT08.S1(CO) and PT08.S3(NOx) correlate strongly with their respective ground-truth gas concentrations.

Methodology: We computed Pearson correlation between sensor signals and pollutant values. Strong, positive correlations were expected for accurate sensors.

Results:

Sensor vs. Pollutant	Correlation (r)	p-value	Interpretation
PT08.S1(CO) vs CO(GT)	0.84	< 0.0001	Strong Positive
PT08.S3(NOx) vs NOx(GT)	-0.65	< 0.0001	Strong Negative

Table 6: Correlation between sensor output and actual pollutant concentration.

Interpretation: PT08.S1 appears to be a reliable sensor for CO detection. However, the inverse correlation of PT08.S3(NOx) with NOx levels raises concerns. While statistically strong, the negative sign contradicts expected behavior and points to calibration or sensor degradation issues. Hypothesis H5 is therefore partially supported.

3.6 Hypothesis H6: NOx Sensor Anomaly

Hypothesis: PT08.S3(NOx) is negatively correlated with NOx(GT), indicating potential malfunction or calibration error.

Results: As previously shown, the correlation coefficient between PT08.S3(NOx) and NOx(GT) is $r = -0.65$, with a p-value < 0.0001.

Interpretation: This statistically significant inverse relationship supports Hypothesis H6. A well-functioning sensor should have a positive correlation with the pollutant it measures. The strong negative value suggests either cross-sensitivity, inversion of sensor signal behavior, or drift over time. This highlights the need for recalibration and sensor diagnostics in smart city deployments.

3.7 Summary of Hypothesis Testing

Hypothesis	**Supported?**	**Test Used**
H1: Rush Hour Effect	Yes	T-test
H2: Seasonal Variation	Yes	T-test
H3: Temp/Humidity Effect	Partially	Pearson Correlation
H4: Benzene Weekday Change	Yes	One-way ANOVA
H5: Sensor Accuracy	Partially	Pearson Correlation
H6: NOx Sensor Anomaly	Yes	Pearson Correlation

Table 7: Summary of hypothesis testing results.

4 Conclusions

This study investigated the relationships between various atmospheric pollutants and external factors including human activity, environmental conditions, and sensor reliability. By leveraging a comprehensive dataset of over 9,000 hourly observations from a chemical multisensor device deployed in a heavily trafficked Italian urban setting, we performed both descriptive and inferential statistical analyses to address six carefully formulated hypotheses. The conclusions drawn from this analysis carry important implications for urban air quality monitoring, policy intervention, and sensor network design.

Summary of Major Findings

Our analysis demonstrated consistent and statistically significant temporal patterns in pollutant concentrations, especially carbon monoxide (CO) and nitrogen oxides (NOx). The following key findings emerged:

- **Rush Hour Effect:** Pollutant levels were significantly elevated during weekday rush hours (7–9 AM and 5–7 PM), with t-tests confirming large differences in mean values between rush and non-rush hours. This supports the strong link between vehicular traffic and urban air pollution.
- **Seasonal Effects:** CO and NOx levels were both significantly higher in winter months (December to February), likely due to increased heating emissions, low wind speeds, and atmospheric conditions such as temperature inversions that trap pollutants.
- **Temperature and Humidity:** Weak-to-moderate correlations were observed between pollutant concentrations and meteorological variables. While statistically significant, these relationships were less impactful than initially hypothesized, suggesting that meteorological factors alone do not dictate air quality levels and must be modeled alongside human activity indicators.
- **Benzene Weekday Variation:** One-way ANOVA revealed significant variation in benzene levels across weekdays. Elevated concentrations during midweek likely reflect industrial work schedules and peak weekday traffic flows.
- **Sensor Accuracy and Anomalies:** While the CO sensor (PT08.S1) correlated strongly with CO(GT), the NOx sensor (PT08.S3) showed an unexpected inverse correlation with NOx(GT). This indicates possible calibration issues or nonlinear sensor response and highlights the necessity of frequent recalibration and diagnostic checks in deployed sensing networks.

These results not only validate multiple hypotheses but also paint a coherent picture of air quality fluctuations in dense urban environments. The combination of anthropogenic

emissions, seasonal variability, and meteorological influences interact in complex ways, making robust monitoring and modeling essential.

Public Health and Environmental Implications

Elevated levels of CO, NOx, and benzene are well-documented risk factors for respiratory and cardiovascular diseases. The confirmation of rush hour spikes and winter pollution plateaus suggests that at-risk populations - particularly those with pre-existing conditions - may benefit from targeted interventions such as:

- Implementation of **low-emission zones** or car-free periods during high-risk hours
- Promotion of **public transit or non-motorized commuting** (e.g., cycling lanes, pedestrian zones)
- **Seasonal emission control** strategies, such as subsidized cleaner heating systems during winter

Understanding these pollutant patterns is also essential for urban planning. Air quality insights can guide the placement of schools, hospitals, or parks away from high-exposure zones, and influence the scheduling of municipal operations such as street cleaning and traffic rerouting.

Sensor Deployment and Data Quality Considerations

Our findings also reveal critical insights into sensor deployment and monitoring strategies. While metal oxide sensors are widely favored for their affordability and real-time capability, their reliability must be continuously validated. The observed inverse correlation between NOx(GT) and PT08.S3(NOx) sensor readings demonstrates the potential for measurement drift, environmental interference, or miscalibration.

Sensor networks must therefore incorporate:

- Periodic recalibration cycles with reference-grade equipment
- Real-time anomaly detection using statistical baselines
- Redundancy in critical locations to cross-verify pollutant levels

Accurate data collection is foundational to any predictive air quality model. Without addressing sensor performance inconsistencies, downstream inferences and policy decisions may be skewed or misdirected.

Study Limitations and Future Work

While this study yielded valuable findings, several limitations remain:

- The dataset spans a single year and location, which may not capture broader regional or longitudinal variations.
- The dataset does not contain explicit data on traffic volume or industrial schedules, limiting our ability to directly link emissions to source activities.
- Sensor signals are based on electrical resistance or output voltage, which may exhibit nonlinear behavior under extreme conditions.

Future research should integrate additional data layers such as GPS-tagged traffic patterns, wind direction and speed, and topographical features. Moreover, machine learning approaches (e.g., clustering or time-series forecasting) could further improve pollutant prediction and event detection.

Final Remarks

In conclusion, this project successfully combines data cleaning, visualization, hypothesis testing, and domain knowledge to deliver a robust analysis of urban air quality. The insights produced offer practical pathways for pollution mitigation and underscore the importance of reliable sensor networks. As cities worldwide grapple with environmental and public health challenges, data-driven studies like this one serve as essential tools in designing informed, targeted, and sustainable interventions.

5 Policy Recommendations and Future Work

This study provides a data-driven foundation for informed policymaking and system design in urban air quality management. Drawing from both statistical evidence and domain understanding, we propose a series of recommendations for government bodies, environmental agencies, and smart city planners. We also outline critical research and infrastructure gaps that should be addressed to enhance the accuracy, coverage, and interpretability of air pollution analytics in future studies.

5.1 Urban Policy Recommendations

The findings of our study clearly demonstrate the temporal concentration of air pollutants around rush hours, mid-week workdays, and winter months. Based on these observations, we suggest the following urban governance and environmental policy strategies:

- **Rush Hour Vehicle Restrictions:** Implementing vehicle use limitations during high-emission periods (7–9 AM, 5–7 PM) can effectively reduce short-term exposure in high-traffic zones. This may include rotating license plate bans, congestion pricing, or the expansion of high-occupancy vehicle (HOV) lanes.
- **Promotion of Public Transit and Active Mobility:** Our findings support the urgent need to reduce private vehicle reliance. Policies incentivizing electric buses, dedicated bike lanes, and expanded metro access-especially during peak pollution hours-could reduce emissions while improving accessibility.
- **Seasonal Emission Regulations:** Since winter emissions were significantly higher, localized seasonal restrictions (e.g., on coal-based heating or industrial output) may yield effective reductions. Governments could also subsidize energy-efficient heating systems during colder months.
- **Zoning Reforms and Urban Planning:** Air quality data can guide decisions on zoning regulations. Residential areas, schools, and hospitals should be located away from pollution hotspots (e.g., highways, intersections, industrial corridors). Air pollution maps derived from this analysis can directly inform city master plans.
- **Industrial Activity Oversight:** Variations in benzene levels across weekdays indicate significant contributions from industrial processes. Requiring real-time emissions disclosure, periodic third-party audits, and enforcing daily caps could help reduce peak pollution episodes.

These recommendations represent low-cost, high-impact strategies that can be tailored to the socio-economic and logistical realities of local municipalities. Data science models such as ours can serve as real-time decision support tools for both long-term planning and short-term emergency responses.

5.2 Sensor Calibration and Infrastructure Investment

One critical insight from our analysis was the unexpected inverse correlation between the PT08.S3(NOx) sensor and true NOx concentrations. This anomaly raises concerns about the reliability of low-cost sensors used in modern urban IoT deployments.

To improve data quality and trustworthiness, we recommend:

- **Sensor Redundancy:** Critical intersections or zones with health-sensitive populations should be equipped with multiple sensor nodes to triangulate true concentrations and detect malfunction patterns.
- **Calibration Protocols:** All sensors should undergo routine recalibration against high-precision laboratory-grade reference devices. Calibration drift must be monitored and corrected in firmware and analytics pipelines.
- **Edge Anomaly Detection:** Deploying real-time anomaly detection models at the edge can prevent polluted or corrupted sensor values from influencing public dashboards or triggering false alerts.
- **Open Data Standardization:** Agencies should publish sensor meta-information (e.g., model type, calibration dates, maintenance logs) alongside pollutant data to ensure interpretability and transparency in public-facing datasets.

Investments in calibration infrastructure and failover mechanisms are crucial if low-cost sensor arrays are to serve as the backbone of national air quality monitoring networks.

5.3 Directions for Future Research

While this project yields significant insights, several areas remain underexplored due to the limitations of our dataset and scope. Future extensions of this study may include:

- **Multi-Year Longitudinal Analysis:** Analyzing pollutant trends over multiple years would reveal how air quality evolves with changing traffic patterns, infrastructure, or climate events. This would allow seasonal and policy impacts to be measured longitudinally.
- **Source Attribution Modeling:** By integrating traffic flow data, satellite imagery, and industrial emission registries, researchers can build models to attribute observed pollutants to specific source classes (vehicles, factories, domestic combustion).
- **Spatio-Temporal Interpolation:** Current data is single-location and time-series based. Expanding the sensor network and applying spatio-temporal kriging or deep learning interpolation would allow full 2D pollution heatmaps across urban areas.

- **Health Risk Forecasting Models:** Future work can link pollutant concentrations to real-world hospitalization and respiratory complaint data using time-series regression or causal inference models, leading to targeted public health alerts.
- **Policy Simulation Environments:** Creating simulation environments that allow city officials to test "what-if" scenarios-such as introducing traffic-free zones or modifying bus routes-would enable evidence-backed policymaking.

These extensions would not only enhance the scope and scientific rigor of air quality studies but also help bridge the gap between academic research and civic impact.

Final Outlook

This project underscores the power of integrating statistical modeling, sensor technology, and environmental science to gain actionable insight into urban air quality. The techniques employed here-from time-series visualization to hypothesis testing-can be generalized to other cities and pollutants. More importantly, the findings and recommendations serve as a vital guide for city planners, public health officials, and environmental engineers striving to build healthier, more resilient communities.

The future of air quality monitoring lies not just in collecting more data, but in making smarter, more interpretable use of it. Through collaborative interdisciplinary efforts, such as this one, we move closer to that goal.

References

1. De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2), 750–757.

2. UCI Machine Learning Repository. (2008). Air Quality Data Set. Retrieved from `https://archive.ics.uci.edu/ml/datasets/Air+Quality`

3. Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (3rd ed.). John Wiley & Sons.

4. Cressie, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience.

5. U.S. Environmental Protection Agency (EPA). (2023). Health Effects of Air Pollution. Retrieved from `https://www.epa.gov/clean-air-act-overview/health-and-environme`

6. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

7. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.

8. Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

9. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

10. World Health Organization (WHO). (2021). Ambient (outdoor) air pollution. Retrieved from `https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`