[[1]](#footnote-2)

Urban Air Quality Pattern & Sensor Assessment: Insights from Exploratory Data Analysis

Sushant Regmi, B.Sc. Computer Science, Kathmandu University, sr51041719@student.ku.edu.np

*Abstract*—In this comprehensive study, we conducted an in-depth Exploratory Data Analysis (EDA) [1]of air quality parameters in an urban setting of Italy, delving into the dynamics of various pollutants and their intricate relationships with meteorological and temporal factors. Further, we showed the need for upgrading various sensors for measurement of toxic gases since the ones overestimated values of almost all the toxic gases. Our dataset, comprising hourly measurements, offered insights into the complex interplay among pollutants, such as carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene (C6H6), total nitrogen oxides (NOx), nitrogen dioxide (NO2), and ozone (O3), along with corresponding sensor responses, including PT08.S1 (tin oxide) for CO, PT08.S2 (titania) for NMHC, PT08.S3 (tungsten oxide) for Nitrogen Oxides, PT08.S4 (tungsten oxide) for Nitrogen Dioxide, and PT08.S5 (indium oxide) for Ozone. Our analysis extended to the realm of meteorological variables, revealing intricate parameter relationships. We observed that temperature and time of the day played pivotal roles in influencing pollutant concentrations, with clear dependencies emerging across the dataset. Employing a rich array of analytical techniques, including scatterplots and heatmaps, we uncovered distinct temporal patterns, pinpointing weekday peak-hour vehicular emissions as the primary source of toxic gases. Notably, our findings underscore the urgency of promoting electric vehicles in urban areas to mitigate air pollution, fostering a healthier, more sustainable environment. This study advocates for cleaner air, resulting in better lung health, and overall well-being, emphasizing the crucial relationship between human activity, air quality, meteorological factors, and public health.

*Impact Statement* — The impact of this exploratory data analysis (EDA) is multifaceted and holds significant implications for urban environmental quality and public health. By unraveling the intricate relationships between air quality parameters and meteorological and temporal factors, our study underscores the urgent need for targeted interventions to combat air pollution in urban areas. Specifically, the identification of vehicular emissions as a primary contributor to toxic gas levels, especially during peak weekday hours, highlights the critical role of transitioning to electric vehicles. Implementing cleaner transportation options not only promises to enhance urban air quality but also holds the potential to improve respiratory health and overall well-being for city residents. Further, evaluation of various sensor values with true values sheds light towards the correctness and reliability of the sensors. This EDA serves as a vital foundation for evidence-based policy decisions aimed at fostering cleaner, healthier urban environments, ultimately contributing to a sustainable and vibrant future for urban communities.

*Index Terms*— Exploratory data analysis, Air quality parameters, pollutants, electric vehicles

# INTRODUCTION

T

HIS study embarks on an exploratory data analysis (EDA) journey to delve into the complex interplay of air quality parameters and meteorological factors in an urban setting. The dataset under scrutiny encompasses a rich tapestry of hourly measurements, including carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene (C6H6), nitrogen oxides (NOx), nitrogen dioxide (NO2), and ozone (O3). The inclusion of sensor responses, such as PT08.S1 (tin oxide) for CO, PT08.S2 (titania) for NMHC, PT08.S3 (tungsten oxide) for Nitrogen Oxides, PT08.S4 (tungsten oxide) for Nitrogen Dioxide, and PT08.S5 (indium oxide) for Ozone, adds layers of intricacy to our analysis. As we navigate the depths of this data, we also consider the meteorological variables at play, including temperature, relative humidity, and absolute humidity. These meteorological factors are pivotal, as they are known to influence pollutant concentrations and their dispersion in urban environments.

The objective of this study is to uncover hidden patterns, correlations, and dependencies within this vast dataset. By doing so, we aim to shed light on the sources and temporal variations of air pollutants and their relationship with meteorological conditions. Such insights hold the promise of informing evidence-based policy decisions and interventions to mitigate air pollution's adverse effects on public health and the urban environment. Furthermore, we will also evaluate the credibility of various multisensory devices which are specialized to measure the ppb value of various toxic gases in the air and

# METHODOLOGY

## Data Cleaning and Preprocessing:

1. *Data Collection*: The dataset used in this study comprises hourly measurements of air quality parameters, including carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene (C6H6), nitrogen oxides (NOx), nitrogen dioxide (NO2), ozone (O3), and various sensor responses, along with meteorological variables like temperature, relative humidity, and absolute humidity. Ground truth hourly averaged concentrations for these parameters were provided by a reference certified analyzer.
2. *Data Inspection*: The analysis began with an initial inspection of the dataset to understand its structure, data types, and missing values. This step provided insights into the dataset's quality and helped in planning the data cleaning process.
3. *Outlier Detection:* To identify and handle outliers, box plots were generated for each of the key parameters, such as CO, NMHC, C6H6, NOx, and NO2. Outliers were identified based on the Interquartile Range (IQR) method and subsequently trimmed from the dataset to ensure the quality of the data used for analysis*.*

## Data Mining:

1. *Correlation Analysis*: The next step involved evaluating the correlations between different columns in the dataset. Correlation matrices and heatmaps were generated to visualize the relationships between variables, providing insights into potential dependencies.
2. *Comparison Charts*: Scatterplots were employed to visualize and quantify the relationships between key variables, such as CO, NO2, MMHC and NOx. This graphical representation allowed for a deeper understanding of the correlations and patterns within the data.
3. *Temporal Analysis*: Time series plots were created to visualize the variations in the true levels of toxic gases (CO, NO2, NMHC and NOx) with respect to time. This analysis revealed temporal patterns and allowed for the identification of correlations between time of day and toxic gas levels.

## Sensor Credibility Assessment:

1. *Comparison with Sensor Values:* The reliability of the chemical sensors embedded in the multisensory device was evaluated by comparing the sensor responses with the ground truth values. This analysis aimed to assess the accuracy and credibility of the sensors in measuring air quality parameters.
2. *Reliability Assessment*: Based on the comparisons, conclusions were drawn regarding the reliability and credibility of the sensors. The study addressed whether the sensor values provided accurate representations of the true levels of air pollutants.

# Exploratory Data analysis

The dataset [2] consists of 9,358 hourly averaged responses from 5 metal oxide chemical sensors within an urban Air Quality Chemical Multisensory Device. Deployed at road level in an Italian city, data were collected continuously from March 2004 to February 2005. Ground truth concentrations for CO, NMHC, benzene, NOx, and NO2 were concurrently measured using reference certified analyzers co-located with the sensor device.

The following steps were followed for Exploratory Data Analysis.

## Data Cleaning and Preprocessing

The given dataset had multiple elements with value as ‘-200’ which were changed to ‘NaN’ for consistency purposes. With that, the following was the situation of null values in various columns in the dataset.

A screenshot of a computer screen

Description automatically generated

*Fig: Count of null values of each column in the dataset*

There is a decrease in the number of values, but the remaining values will be of more validity and importance.

Now, we remove the outliers that are present in the true value of the gases.

A graph of a graph of a graph

Description automatically generated with medium confidence

*Fig: Box plots of true value of various gases in the dataset with outliers*

A graph of a diagram

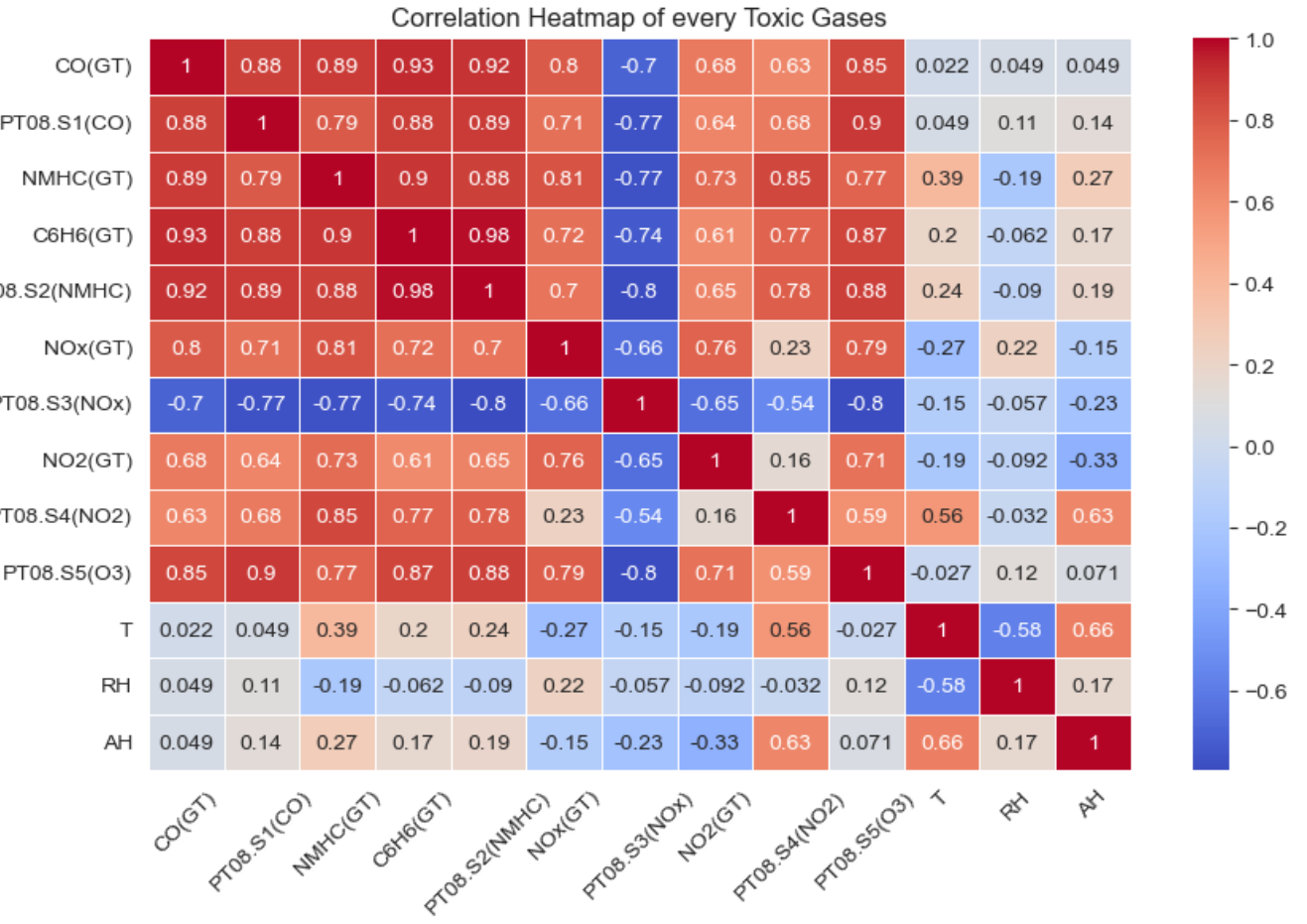
Description automatically generated with medium confidence

*Fig: Box plots of true value of various gases in the dataset without outliers*

After removing the outliers, the dataset is ready for evaluation.

## Data Mining

Firstly, we generated a correlation heatmap between every column in the dataset.



*Fig: Correlation heatmap of every Toxic Gases*

From the correlation diagram, we can infer that a lot of toxic gases have high levels of linear correlation.

Here is a view of different levels of concentration of toxic gases in the air throughout the measurement lifecycle.

A graph showing the number of toxic gases

Description automatically generated

*Fig: Frequency of toxic gases throughout the measurement lifecycle*

Total Nitrogen Oxide (NOx) and Nitrogen Dioxide (NO2) is the highlight of the graph. Research [3] on production of such nitrogen oxides seems to point at emissions from cars, trucks and buses, power plants, and off-road equipment as the culprit for higher rate of these toxic gases. Similarly, emissions of carbon monoxide (CO) and Non-Methane Hydrocarbon (NMHC) are also one of the side effects of burning fuels to produce energy. Each of these toxic gases are responsible for an increase in respiratory health problems for people living in such urban areas.

This research is also backed by following graph that shows that during the daytime when the roads are busy with fuel gobbling automobiles, the rate of NOx, NO2, CO and NMHC skyrocket and during the mid night (12:00 AM) onwards to early morning time (6:00 AM), the rate of these toxic chemicals seem to decrease.

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

*Fig: Change in levels of toxic gases throughout the day*

Now, let’s look at the findings from the sensors that were integrated into an Air Quality Chemical Multisensory Device which nominally took the measurements of various toxic gases.

## Sensor Credibility Assessment

1. *PT08.S1(CO):* This tin oxide sensor was used to nominally sense the hourly measurement of CO in the air. Here is the Bland Altman Plot [4] which shows the agreement between both the measurements.

A graph with a line

Description automatically generated

*Fig: Bland Altman Plot for CO vs PT08.S1(CO)*

This shows that the tin oxide sensor overestimates the true value of the Carbon Monoxide in the air, and it may not be a good way to measure the CO content in the air.

1. *PT08.S2(NMHC):* This titania sensor was used to nominally sense the hourly measurement of NMHC in the air. Here is the Bland Altman Plot which shows the agreement between both the measurements.

A graph of blue dots

Description automatically generated

*Fig: Bland Altman Plot for NMHC vs PT08.S2(NMHC)*

This shows that the titania sensor overestimates the true value of the NMHC in the air and it may not be a good way to measure the NMHC content in the air.

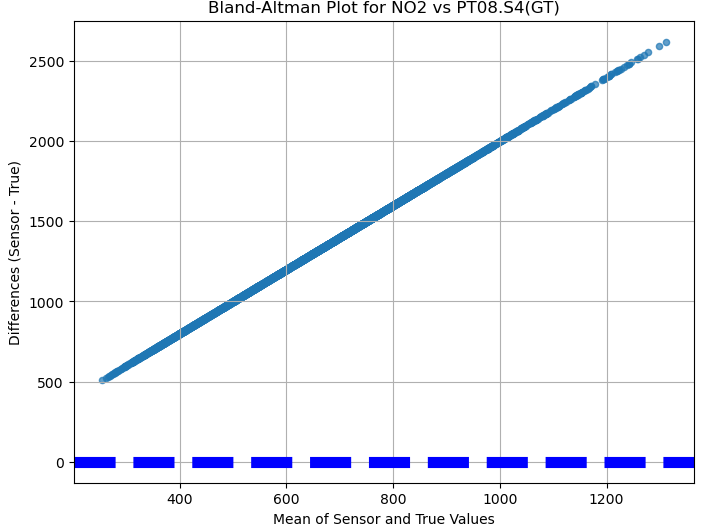
1. *PT08.S3(NOx):* This tungsten oxide sensor was used to nominally sense the hourly measurement of NOx in the air. Here is the Bland Altman Plot which shows the agreement between both the measurements. A graph with blue lines and a white grid

   Description automatically generated

*Fig: Bland Altman Plot for NOx vs PT08.S3(NOx)*

This shows that the tungsten oxide sensor is somewhat better than other sensors in the dataset however, it is still not good enough to be reliable for measuring the Nitrogen Oxides in the air.

1. *PT08.S4(NO2):* This tungsten oxide sensor was used to nominally sense the hourly measurement of NO2 in the air. Here is the Bland Altman Plot which shows the agreement between both the measurements.



*Fig: Bland Altman Plot for NO2 vs PT08.S4(NO2)*

This shows that the tungsten oxide sensor overestimates the true value of the Carbon Monoxide in the air, and it may not be a good way to measure the CO content in the air.

In the webpage of the dataset, the author talks about the data exhibiting cross-sensitivities, where the response of one sensor is influenced by the presence of other gases. Furthermore, dataset might reflect concept drift, indicating changes in the underlying data distribution, and sensor drift, suggesting variations in sensor performance over time resulting in these factors potentially affecting the accuracy of the sensors' concentration estimations.

# Conclusion

## Our dataset, comprising over 9,000+ hourly measurements from reference analyzers and a network of metal oxide chemical sensors, provided a rich source of information about the dynamics of air pollutants and their interactions with meteorological factors like temperature. Our EDA journey began with meticulous data cleaning and preprocessing, ensuring the integrity of our analyses. We wielded the power of visualization, employing box plots to uncover outliers and trimming them from key parameters like CO(GT), NMHC(GT), C6H6(GT), NOx(GT), and NO2(GT). With our data cleansed and ready, we ventured into the heart of EDA.

## Correlation analyses illuminated intricate relationships between variables, shedding light on dependencies that underpin air quality dynamics. Scatterplots visualized these correlations, revealing intricate patterns among pollutants and meteorological variables. Temporal analysis further enriched our understanding, unveiling the profound impact of time on toxic gas levels. Daytime peaks, especially on weekdays, underscored the influence of vehicular emissions on urban air quality. By comparing sensor values to true values, we gauged the reliability of these chemical sensors.

References

[1] Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. (2016). Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham. <https://doi.org/10.1007/978-3-319-43742-2_>

[2] Vito,Saverio. (2016). Air Quality. UCI Machine Learning Repository. <https://doi.org/10.24432/C59K5F>

[3] Wan, N., Xiong, X., Kluitenberg, G. J., Hutchinson, J. M. S., Aiken, R., Zhao, H., and Lin, X.: Estimation of biomass burning emission of NO2 and CO from 2019–2020 Australia fires based on satellite observations, Atmos. Chem. Phys., 23, 711–724, <https://doi.org/10.5194/acp-23-711-2023>

[4] M.A. Mansournia et al. Bland-Altman methods for comparing methods of measurement and response to criticisms, <https://doi.org/10.1016/j.gloepi.2020.100045>

1. This paper was submitted on September 8, 2023, for assessment. [↑](#footnote-ref-2)