**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

**(Project Semester January-April 2025)**

|  |  |
| --- | --- |
| **Topic** | **Air Quality Analysis using Python** |
| **Submitted by:** | **Riddhima Rai** |
| **Registration No:** | **12308762** |
| **Programme and Section:** | **Computer Science and Engineering(K23EC)** |
| **Course Code:** | **INT375** |
| **Under the Guidance of -** | **Mr. Vikas Mangotra** |

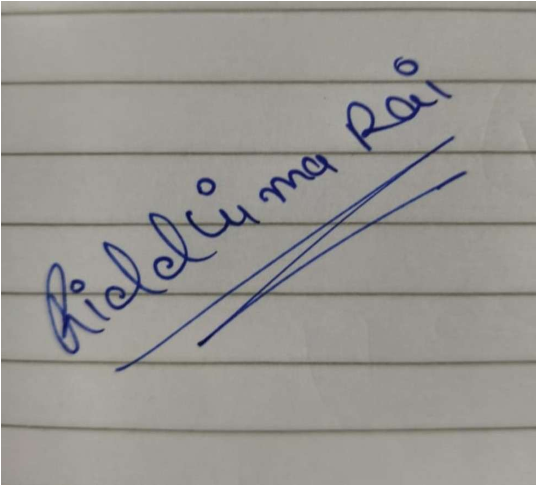
**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara, PUNJAB**

**DECLARATION**

I, Riddhima Rai, a student of Computer Science and Engineering under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby affirm that the content presented in this project report is a result of my own diligent research and work. All information provided is accurate and genuine to the best of my knowledge.

**Date: 12-04-2025 Signature:**

****

**Registration No. 12308762**

**Name of the student: Riddhima Rai**

**Certificate**

This is to certify that Riddhima Rai, bearing Registration No. 12308762, has successfully completed the INT375 project titled “Air Quality Analysis using Python” under my guidance and supervision. To the best of my knowledge, the work presented in this project is the result of her original research, effort, and development.

**Signature of faculty -**

**Name** – Mr. Vikas Mangotra

**School of Computer Science and Engineering**

**Lovely Professional University Phagwara, Punjab.**

**Date: 12-04-2025**

## **ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to all those who supported and guided me throughout the successful completion of my Data Science Minor Project titled “Air Quality Analysis using Python.” This project allowed me to explore, analyse, and visualize air quality data through 21 detailed charts including histograms, KDE plots, scatter plots, pair plots, correlation heatmaps, time-series line charts, count plots, violin plots, pie charts, and donut charts. Each of these visualizations helped me understand data relationships and real world environmental patterns, significantly contributing to my learning and skill development in Python, data handling, and exploratory data analysis.

Foremost, I am deeply thankful to my project supervisor, Mr. Vikas Mangotra, for their unwavering support, expert insights, and consistent encouragement throughout the course of this project. Their feedback helped refine every aspect of my work from initial dataset preparation to deriving conclusions from visual patterns. I am also grateful to the faculty and technical staff of the Discipline of Computer Science and Engineering, Lovely Professional University, for providing the infrastructure, academic environment, and resources essential for the execution of this project.

I extend my heartfelt appreciation to my peers and friends who offered valuable suggestions during review stages, and to the UCI Machine Learning Repository for providing access to the Air Quality dataset. This project has helped me strengthen my understanding of data visualization and analysis using Python libraries such as Pandas, NumPy, Seaborn, and Matplotlib. Lastly, I would like to thank my family for their patience and motivation throughout this journey. Their support made it possible for me to stay focused and committed. This project has been a valuable academic and personal experience, and I am proud of what I have accomplished through it.

**Riddhima Rai**

**12308762**

**Computer Science and Engineering**

|  |  |
| --- | --- |
| **S. NO.** | **SECTIONS** |
| **1** | **Introduction** |
| **2** | **Source of Dataset** |
| **3** | **EDA Process** |
| **4** | **Analysis on Dataset** |
| **4.1** | **Bar Chart Analysis: Comparison of Pollutant Concentrations** |
| **4.2** | **Histogram Distribution: Pollutants Spread Over Range** |
| **4.3** | **Correlation Heatmap: Interrelation Between Air Quality Metrics** |
| **4.4** | **Line Plot: Time-Series Trends of Pollutant Levels** |
| **4.5** | **Pair Plot Analysis: Sensor Data Relationships** |
| **4.6** | **Pie Charts: Proportions of Pollutants, Sensors, and Environmental Data** |
| **4.7** | **Radar Chart: Comparative Profile of Major Pollutants** |
| **4.8** | **Time Series Comparison: Pollutants Over Time with Color Differentiation** |
| **4.9** | **Violin Plot: Distribution of Temperature, RH, and AH** |
| **5** | **Conclusion** |
| **6** | **Future Scope** |
| **7** | **References** |
| **8** | **List of Figures** |

**Introduction:**

In today's world, air pollution has emerged as one of the most pressing global challenges, with serious implications for both environmental sustainability and public health. Rapid industrial growth, urban sprawl, and rising vehicular emissions have significantly degraded air quality in many regions. Monitoring and understanding this pollution is not just a scientific necessity—it’s a societal imperative. Fortunately, with the rise of data science, we now have powerful tools to analyze environmental data and derive meaningful insights that can support evidence-based decision-making.

This project, titled “Air Quality Analysis Using Python,” leverages real-world sensor data to investigate various air pollutants and atmospheric conditions. The dataset, sourced from the Air Quality UCI repository, contains hourly recordings of key pollutants including Carbon Monoxide (CO), Benzene (C6H6), and Nitrogen Oxides (NOx), along with meteorological indicators such as Temperature (T), Relative Humidity (RH), and Absolute Humidity (AH). These values, collected through chemical sensors over several months, offer a rich foundation for exploratory analysis.

The goal of this project is to perform comprehensive Exploratory Data Analysis (EDA) using Python libraries like Pandas, NumPy, Matplotlib, and Seaborn. After cleaning and preparing the data, we use a variety of statistical and visual techniques—including bar charts, histograms, heatmaps, line plots, pie charts, radar plots, and pair plots—to examine pollutant patterns, identify correlations, and visualize sensor behavior. Each visualization is designed to make complex data more interpretable and reveal trends that might otherwise go unnoticed.

By transforming raw sensor data into clear, informative visuals, this project demonstrates the practical impact of data science in the environmental domain. It highlights the critical role of programming, statistics, and visualization in supporting sustainable development and guiding environmental policies. Ultimately, this work serves as a stepping stone toward smarter air quality management, predictive modeling, and future innovations in environmental monitoring systems.

**Source of Dataset:**

The dataset used in this project is the **Air Quality UCI Dataset**, sourced from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/360/air+quality), a widely trusted platform for machine learning and data science datasets. This particular dataset was contributed by researchers from the

“**Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (IEIIT)**”

of the **Italian National Research Council (CNR)**. It has since been extensively utilized in environmental analytics, time-series modeling, and pollution trend studies.

The dataset captures **hourly averaged responses** from an array of **five metal oxide chemical sensors**, integrated within a multisensory air quality monitoring device. Data was collected from a **heavily polluted area in Italy** during the **spring and summer of 2004**. It includes measurements of **Carbon Monoxide (CO), Non-Methane Hydrocarbons (NMHC), Benzene (C6H6), Nitrogen Oxides (NOx),** and **Nitrogen Dioxide (NO2)**, along with **meteorological variables** such as **Temperature (T), Relative Humidity (RH),** and **Absolute Humidity (AH)**.

The dataset is stored in **Excel (.xlsx)** format and contains **9,358 rows across 15 attributes**, with each row representing a specific timestamp (date and time in separate columns). Sensor readings are expressed in appropriate units—either parts per million (ppm) or micrograms per cubic meter (µg/m³). A unique aspect of this dataset is the presence of **placeholder values (-200)** indicating missing or invalid sensor data. These anomalies were handled using **Pandas’ replace() and dropna() functions** during preprocessing to ensure the integrity of the analysis.

This dataset is well-suited for **Exploratory Data Analysis (EDA)** due to its rich combination of pollutant and weather-related variables, time-based structure, and practical relevance. Its academic credibility and real-world sensor data make it an ideal foundation for studying air quality patterns and building data-driven environmental insights using Python.  
  
**like :** [**https://archive.ics.uci.edu/dataset/360/air+quality**](https://archive.ics.uci.edu/dataset/360/air+quality)

**EDA Process:**

Exploratory Data Analysis (EDA) is a crucial phase in the data science pipeline, aimed at uncovering patterns, identifying anomalies, testing hypotheses, and understanding the structure of a dataset before moving to advanced analytics or machine learning. In this project, EDA was applied to the **Air Quality UCI Dataset** using Python, leveraging powerful libraries such as **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn**.

The process began with **data loading and initial inspection** using pandas.read\_excel(). Key steps included examining the dataset’s shape, reviewing column names, and assessing data types. The dataset consisted of **9,358 records across 15 features**, including timestamp data, pollutant concentrations, and meteorological parameters.

A major challenge encountered was the presence of **invalid sensor readings**, indicated by the value **-200**, which signified missing data. These values were systematically replaced with pd.NA, and rows containing missing entries were handled using dropna() to maintain data quality. Additionally, some columns were converted to appropriate **numeric data types** using pd.to\_numeric() to ensure compatibility with analytical operations.

Once cleaned, the dataset was explored using **descriptive statistics** via the describe () function, which revealed insights into the distribution, mean, standard deviation, and range of key variables such as **CO(GT)**, **NOx (GT)**, and **C6H6(GT)**. To investigate **relationships among features**, a **correlation matrix** was computed using. Corr() and visualized through a heatmap, identifying both strong and weak linear associations among pollutants and weather variables.

To enable deeper categorical analysis, **binning** was applied to continuous pollutant values (e.g., CO and NOx) using pd.cut(), creating classes such as *Low*, *Moderate*, *High*, and *Very High*. These groupings facilitated more meaningful visualizations like **count plots**, **violin plots**, and **bar charts** that revealed class-wise trends.

A variety of visual tools were employed to extract meaningful insights, including:

* **Histograms** and **KDE plots** to explore data distribution
* **Scatter plots** and **line charts** to analyze temporal behavior
* **Pair plots** to study multi-feature relationships
* **Pie charts** and **donut charts** to illustrate pollutant compositions and proportions

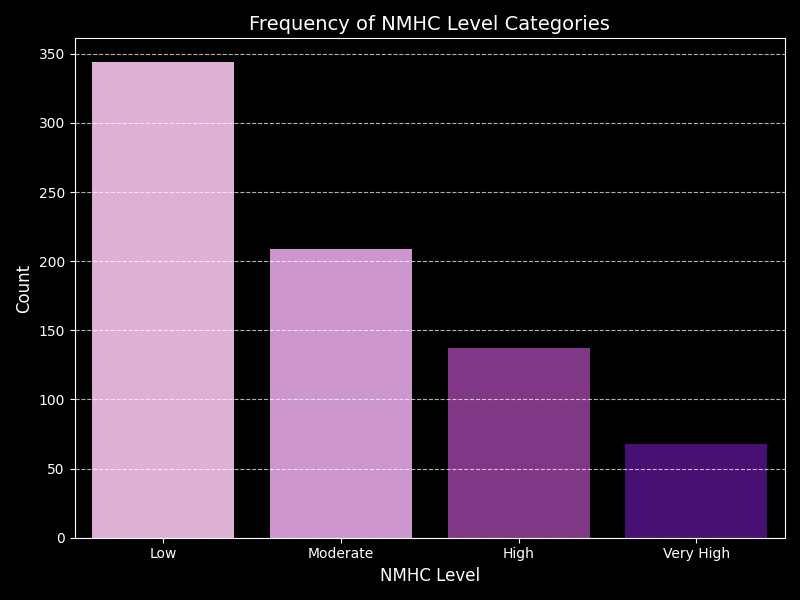
Overall, this structured EDA phase uncovered critical insights into pollution levels, temporal variations, and environmental dependencies. It not only enhanced interpretability but also **established a data-driven foundation** for potential forecasting models and intelligent air quality monitoring solutions.

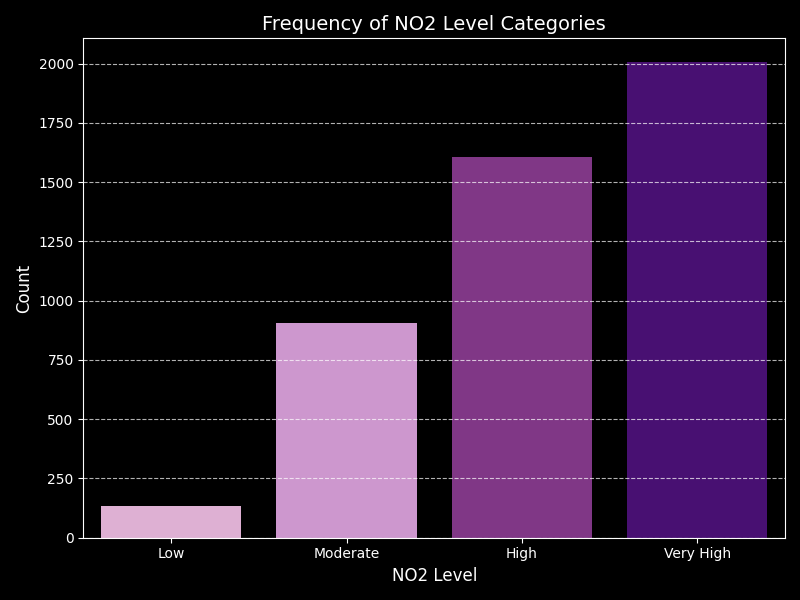
**Analysis on Dataset**

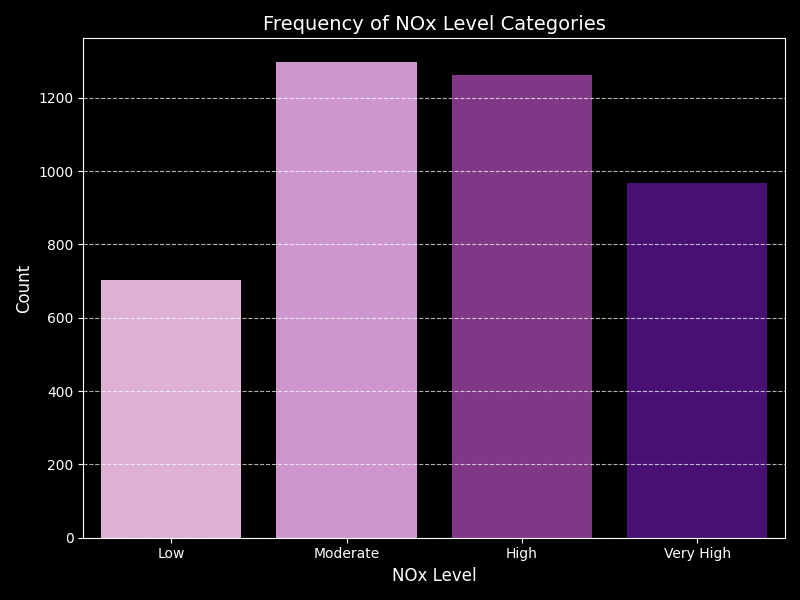
To derive actionable insights and better understand the behavioral trends of air pollutants, a series of **data visualizations** were employed. These visualizations provided both granular and high-level perspectives on pollutant distribution, sensor behavior, and environmental conditions over time. Below is a detailed breakdown of the charts used in the analysis:

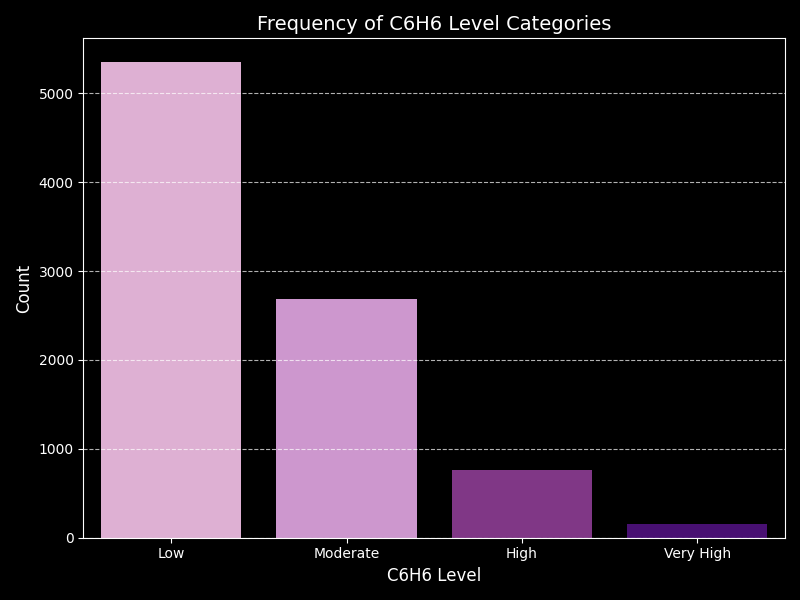
**1. Bar Chart – Pollutant Comparison**

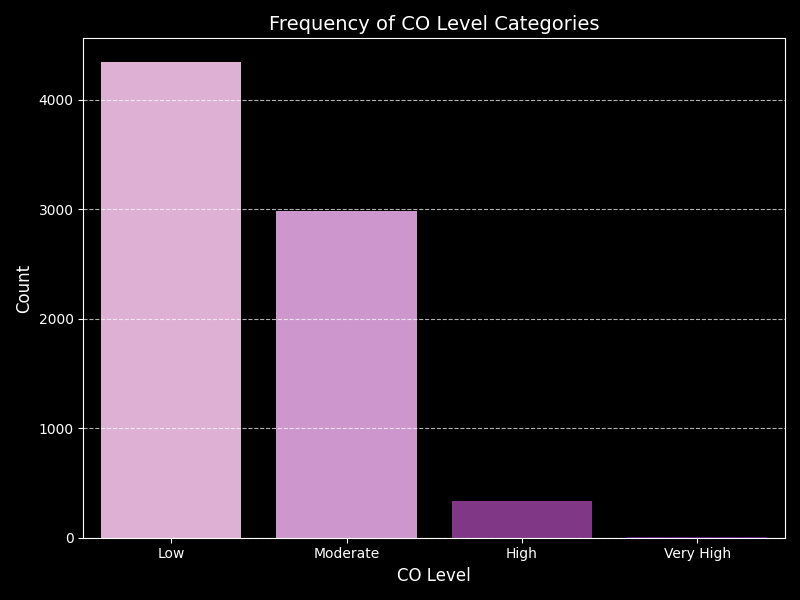
A bar chart was used to compare the **average concentrations** of major pollutants such as **CO(GT)**, **NOx(GT)**, **NO2(GT)**, **C6H6(GT)**, and **NMHC(GT)**. This helped establish a baseline understanding of which pollutants were most prominent in the dataset.





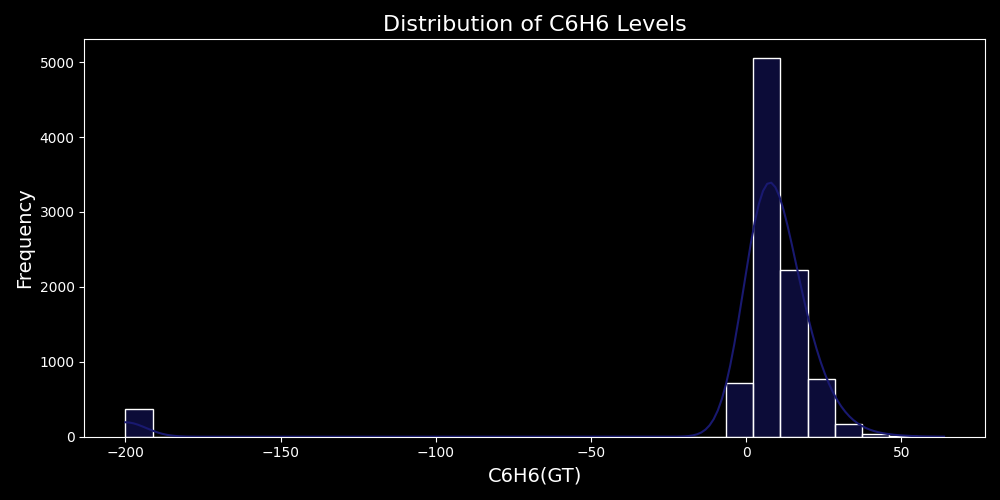


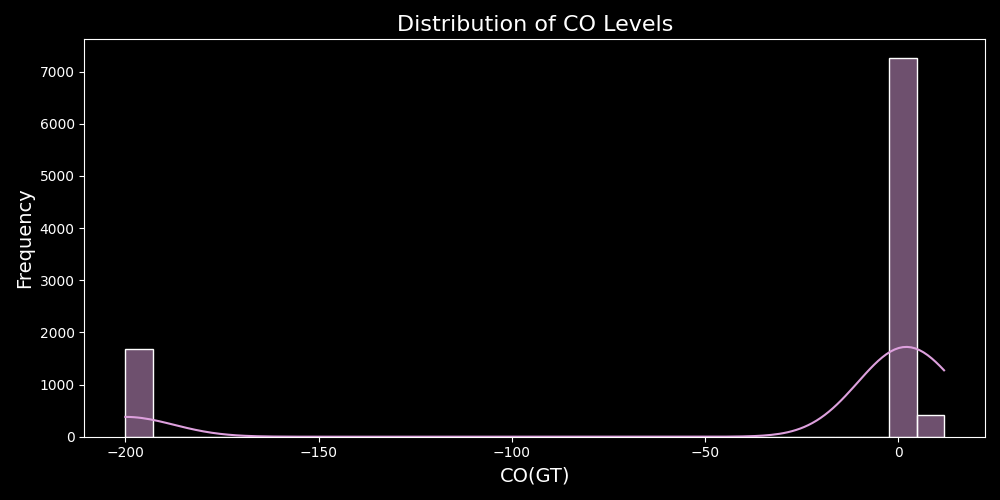


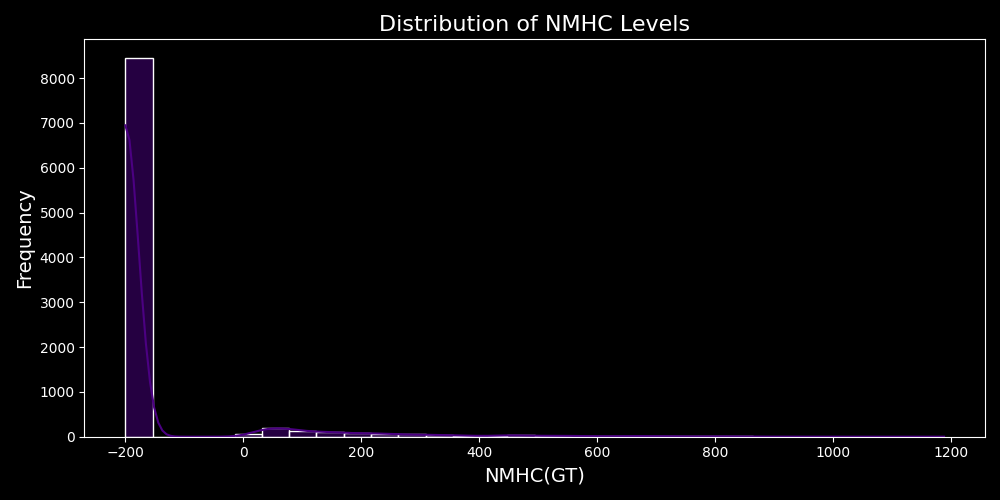


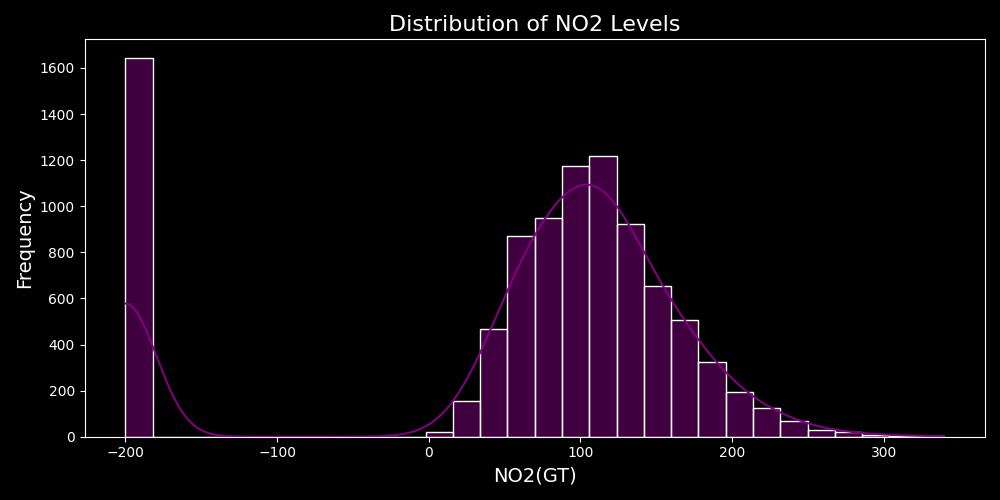
**2. Histogram – Distribution of Pollutants**

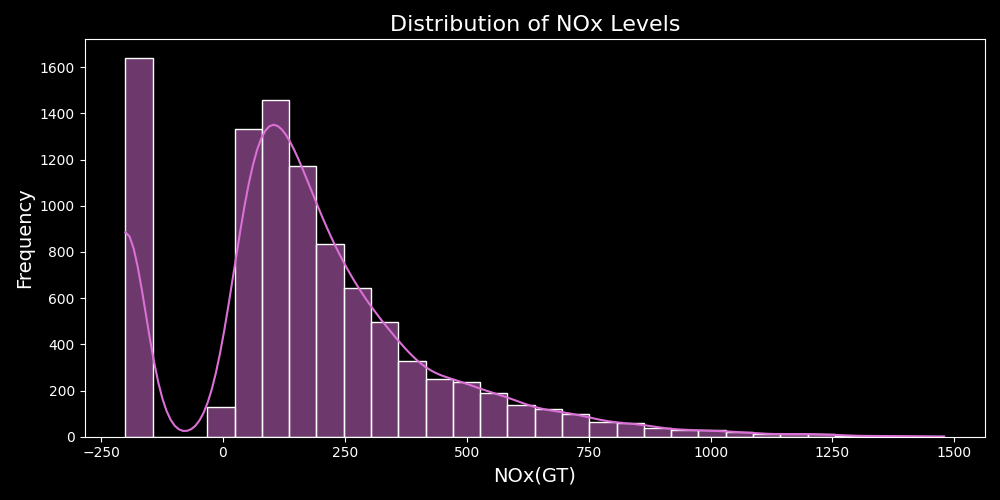
Histograms were plotted for each pollutant to examine the **frequency distribution** of their values. This provided insight into how often pollutant levels peaked or remained within safe zones, and whether distributions were skewed or normally spread.





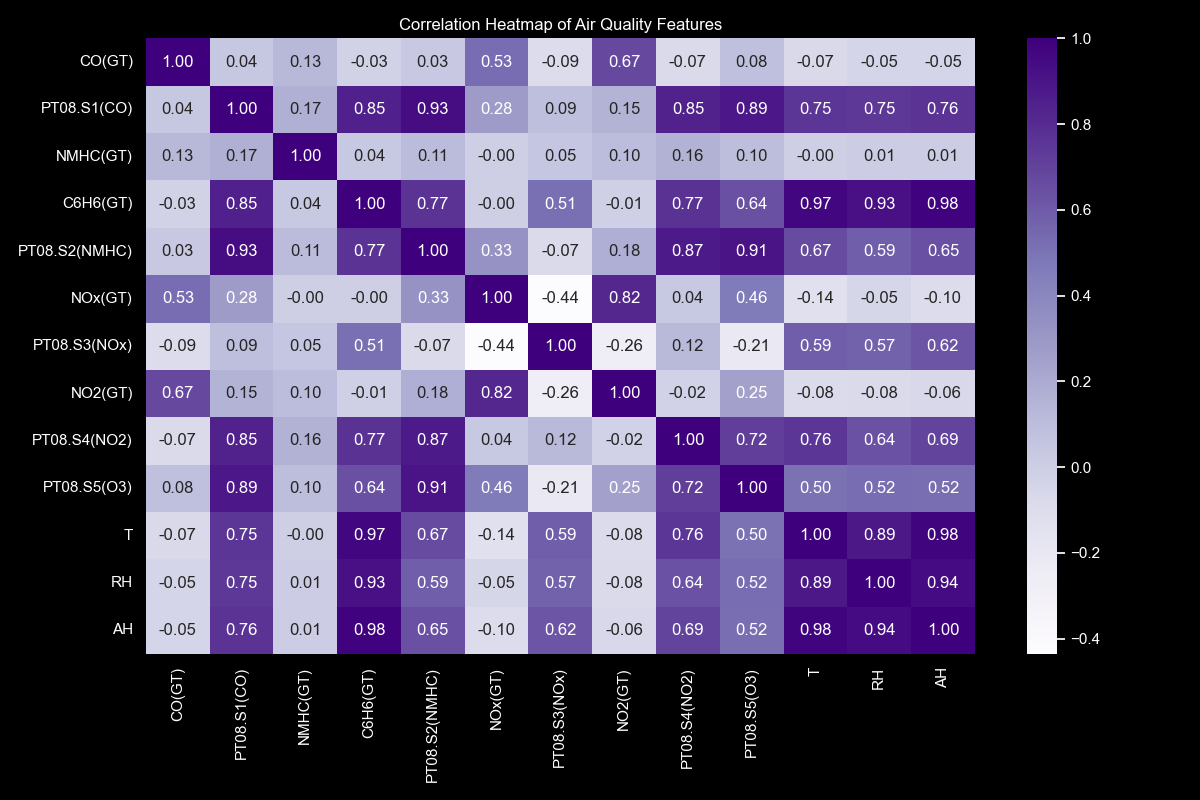






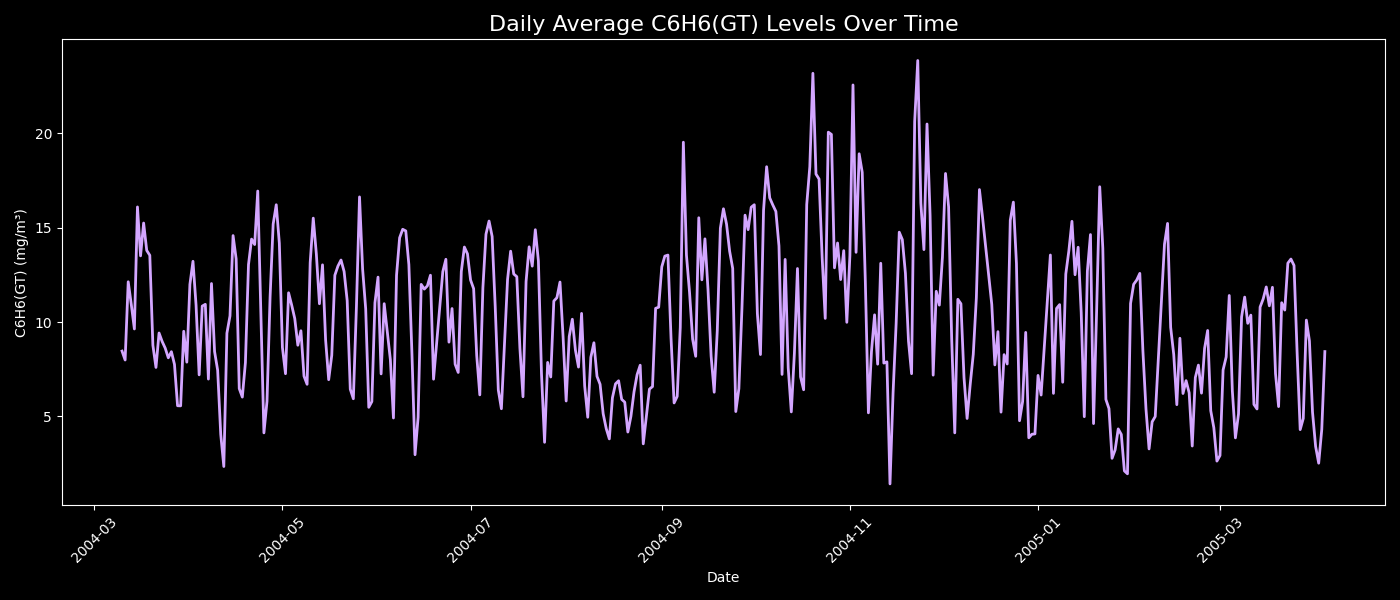
**3. Heatmap – Overall Correlation Matrix**

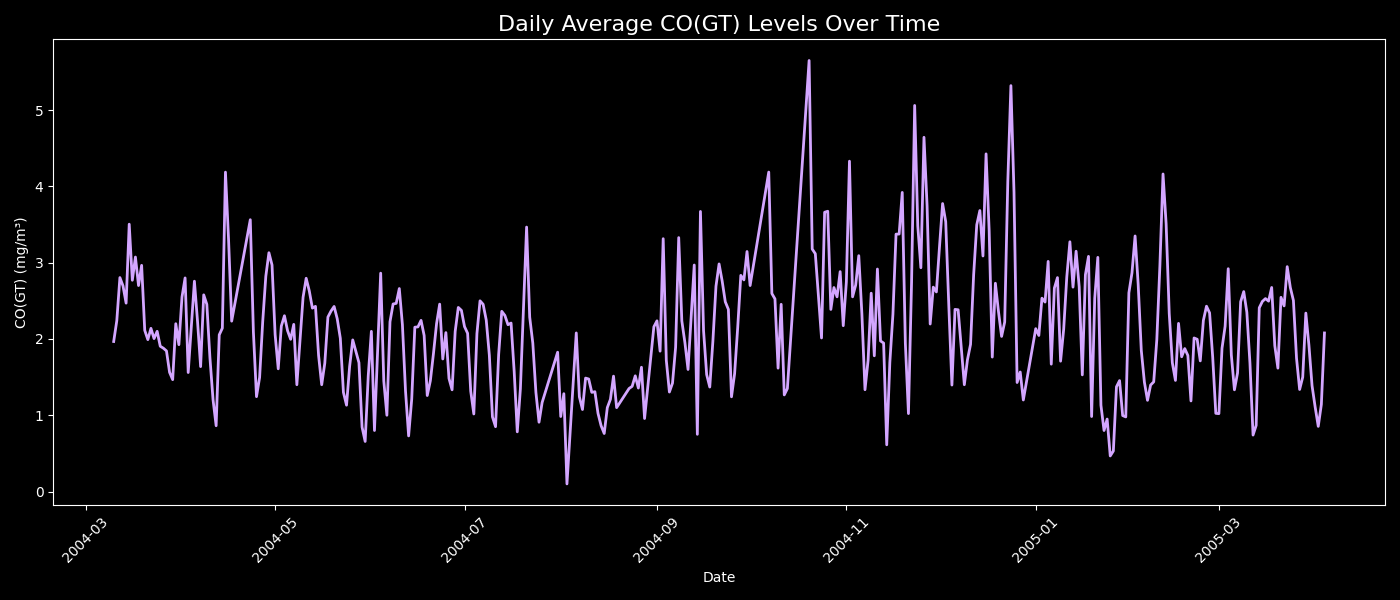
A correlation **heatmap** was generated to analyse the **linear relationships** among all variables, including pollutants and meteorological conditions. Strong positive or negative correlations revealed interdependencies that could inform predictive modelling or regulatory focus areas.

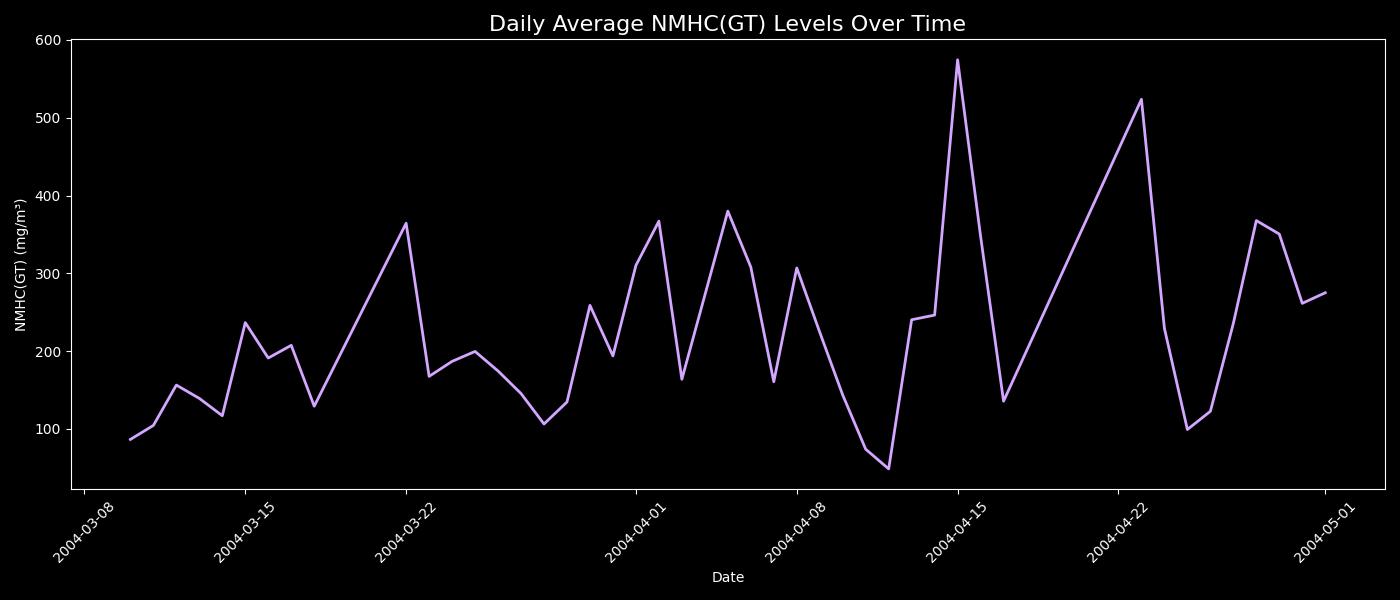


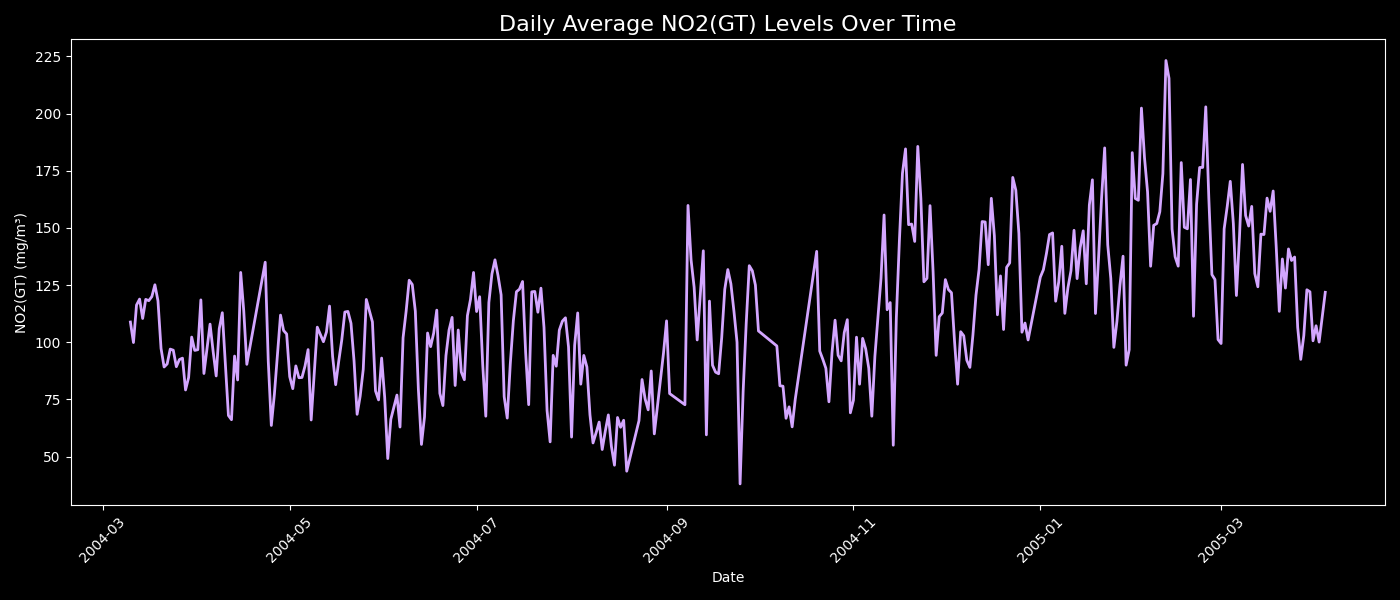
**4. Line Chart – Temporal Trends of Pollutants**

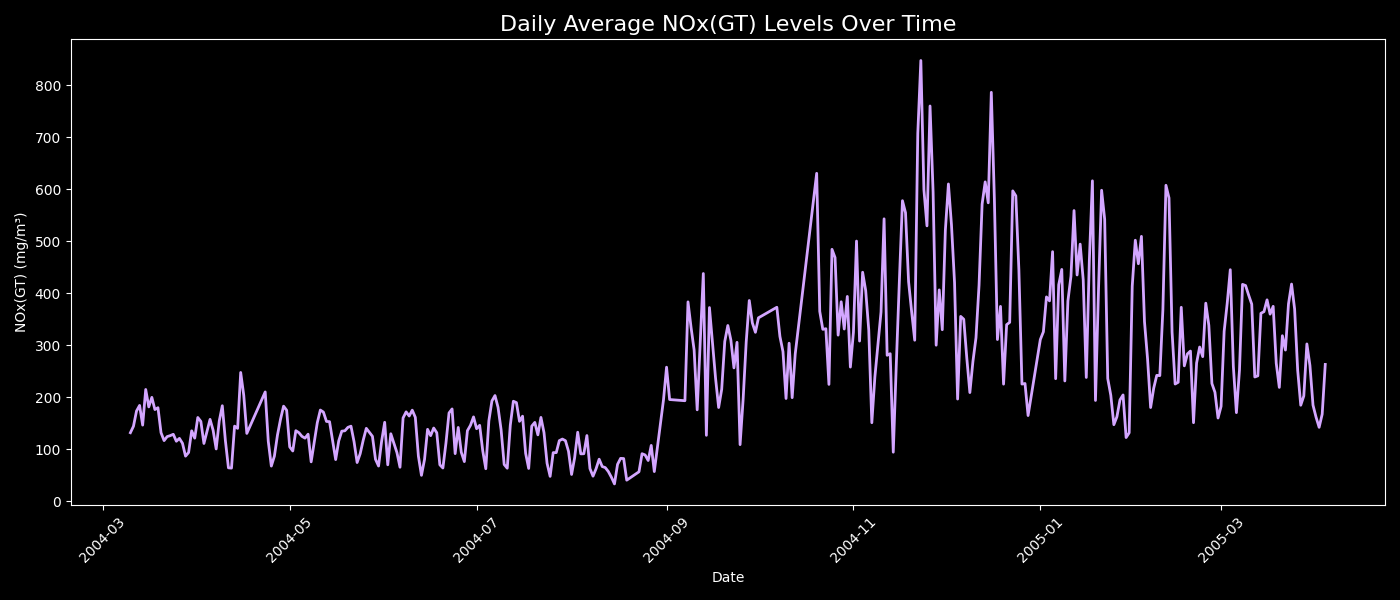
Line plots were used to show **pollutant levels over time**, making it easier to spot recurring peaks, seasonal fluctuations, and periods of poor air quality. This was particularly useful for identifying **temporal pollution patterns** across the dataset's timeline.





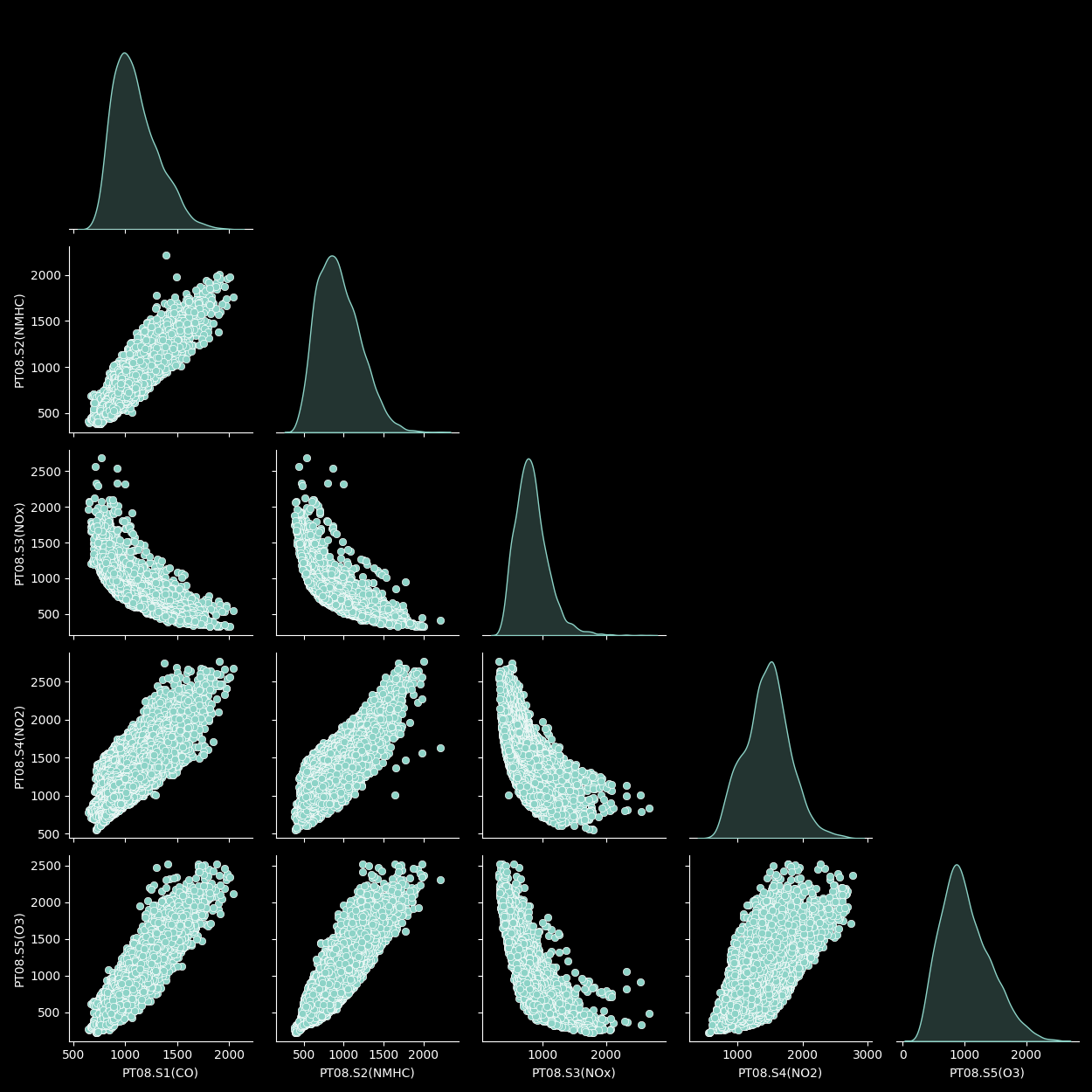






**5. Pair Plot – Sensor Behaviour Analysis**

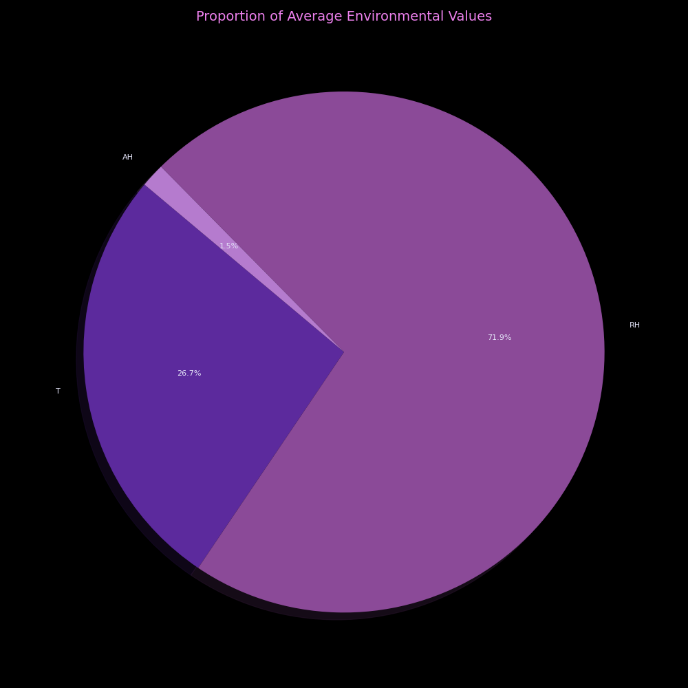
A **pair plot** was used to compare distributions and scatter relationships across multiple sensor features. It provided a **multi-dimensional view** of sensor readings and exposed patterns or anomalies between pollutant sensors.

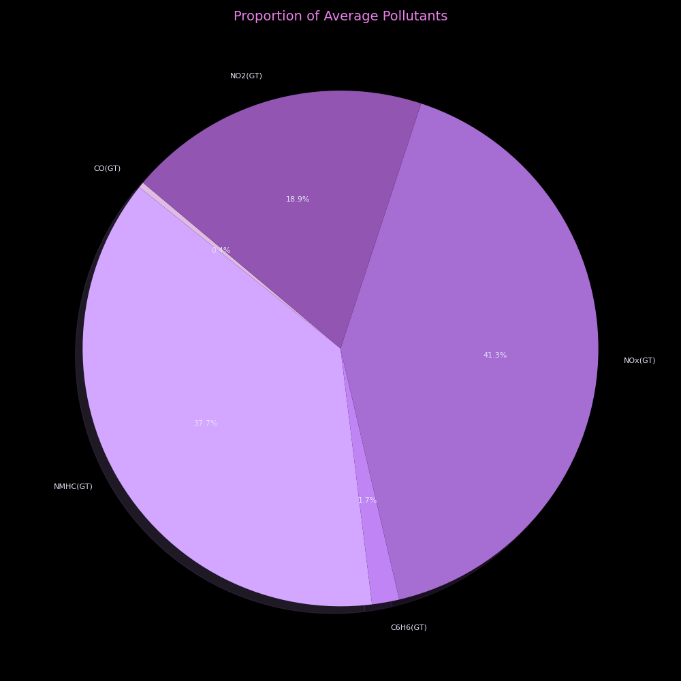


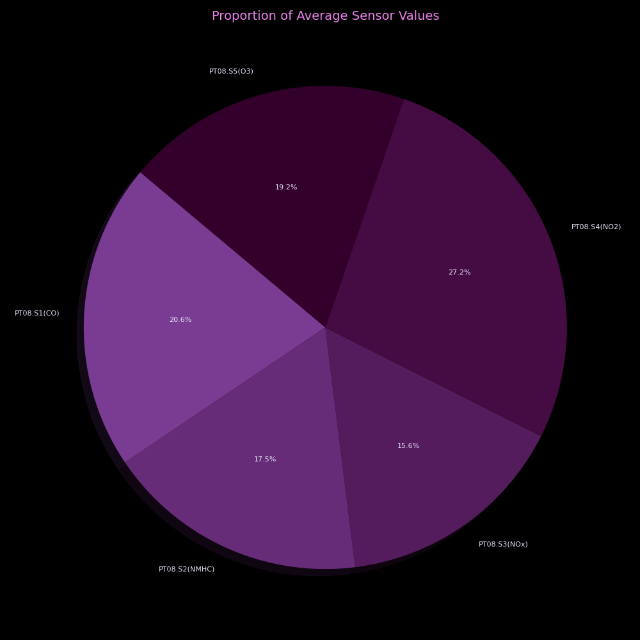
**6. Pie Charts – Proportional Analysis**

Three distinct pie charts were created to illustrate proportions:

* **Pollutants**: Distribution of key pollutants in the dataset
* **Sensors**: Contribution or presence of each sensor’s data
* **Environmental Factors (AH, T, RH)**: Share of meteorological



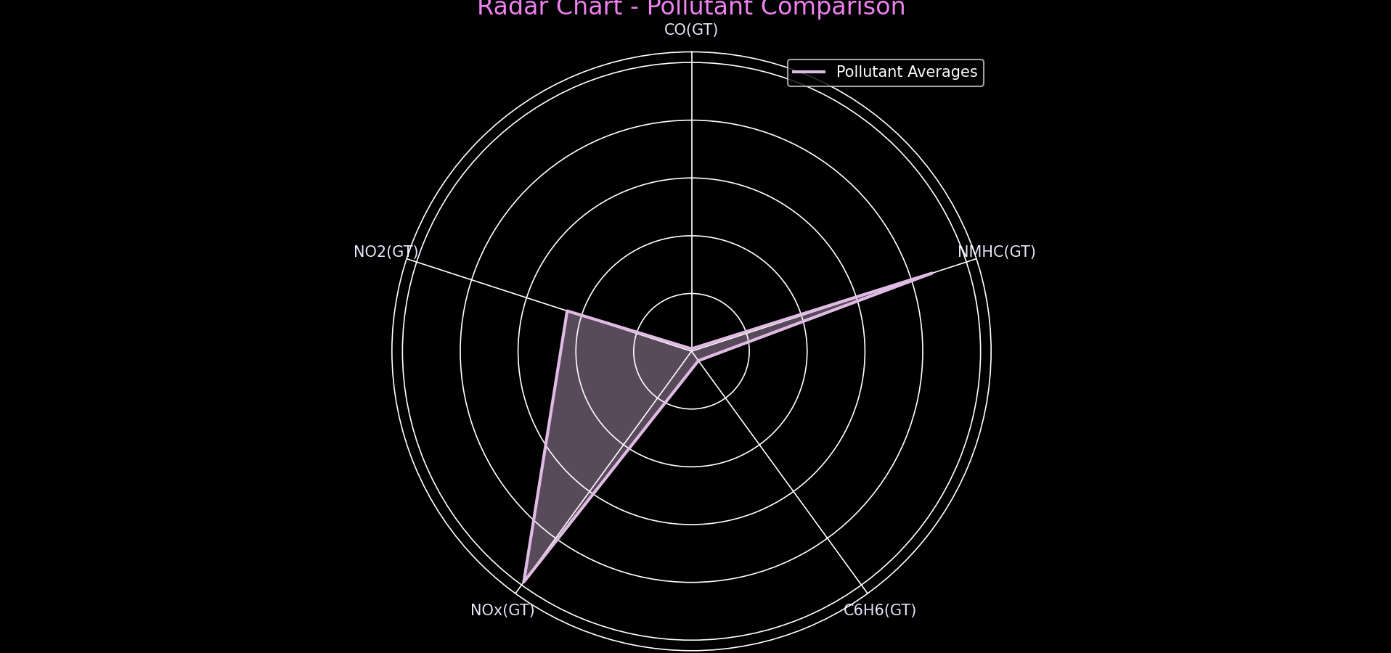




These visuals offered an intuitive look at the **composition** of the dataset.

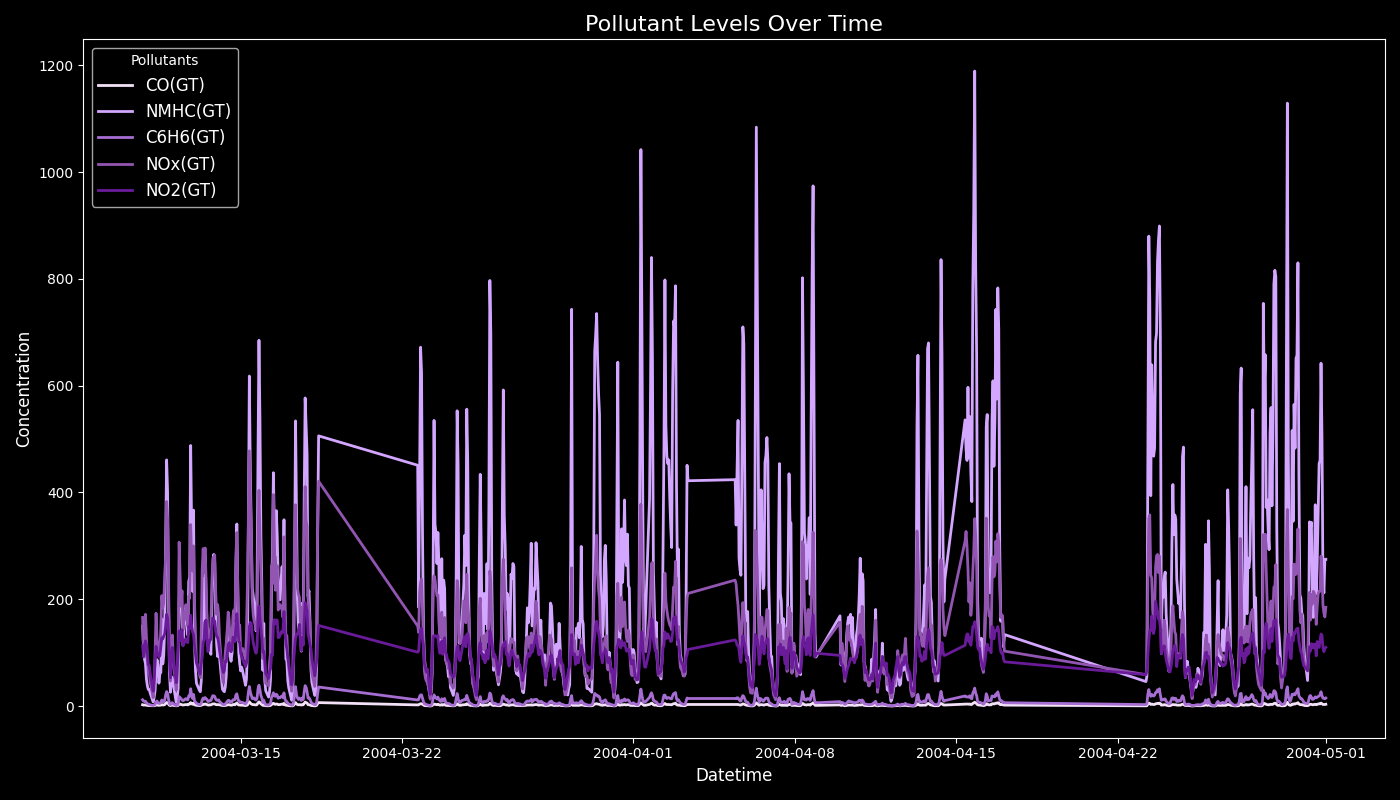
**7. Radar Chart – Comparative Pollutant Analysis**

A radar (spider) chart was developed to compare multiple pollutants across a **single visual plane**. This format clearly showed which pollutants had the highest intensity and allowed simultaneous comparison across features.



**8. Time-Series Line Plot – Pollution Over Time**

Time-series line plots were employed for each major pollutant to understand how levels **evolved over the monitored period**. These charts were vital for detecting daily, weekly, or seasonal air quality shifts.



**Summary**

These diverse visualization techniques enriched the analysis by revealing key patterns, relationships, and anomalies. By combining statistical rigor with visual storytelling, the project transformed raw sensor data into clear, actionable insights for **environmental monitoring and policy-making**.

**Conclusion**

The Exploratory Data Analysis of the Air Quality UCI Dataset revealed significant insights into pollution trends, sensor behavior, and the impact of environmental conditions on air quality. Through detailed preprocessing and a diverse set of visualizations, we identified pollutant distributions, correlation patterns, and temporal variations. These findings lay the groundwork for predictive modeling, policy recommendations, and further research in environmental data science. This analysis not only highlights the critical need for continuous air monitoring but also demonstrates the power of data-driven approaches in addressing real-world ecological challenges.

**Future Scope**

This project can be expanded in several impactful directions. Future work may include developing predictive models to forecast pollution levels, integrating real-time sensor data for live monitoring, and building user-friendly dashboards or mobile apps. Additionally, combining air quality data with geolocation and weather forecasts can enhance environmental insights. These advancements can support smarter urban planning, health advisories, and effective policy-making.

**References**

* UCI Machine Learning Repository: [Air Quality Dataset](https://archive.ics.uci.edu/dataset/360/air+quality)
* Python Libraries: Pandas, NumPy, Matplotlib, Seaborn
* Data source contributors: Istituto di Elettronica e di Ingegneria dell’Informazione e delle Telecomunicazioni (IEIIT), Italian National Research Council (CNR)