# **Project Manager: Motamedi, ali**

# **Air Quality Analysis**

**Abstract**

Air quality has a significant impact on human health, the environment, and the economy. The fluctuating levels of pollutants such as Carbon Monoxide (CO), Nitrogen Oxides (NOx), and Benzene (C6H6) in urban areas pose a challenge for environmental policies and public health. This project addresses the opportunity to utilize a comprehensive dataset on air quality to analyse the patterns, correlations, and impacts of various pollutants. The primary research question focuses on understanding how these pollutants vary over time and their relationship with environmental conditions such as temperature, humidity, and atmospheric pressure.

**Literature review**

Air pollution is a significant global issue, with 91% of the world’s population living in areas exceeding WHO air quality limits (WHO, 2016). Ambient air pollution accounts for an estimated 4.2 million deaths per year due to stroke, heart disease, lung cancer, and chronic respiratory diseases (WHO, 2018). With increasing urbanization and industrialization, air quality has become a growing concern for public health and environmental policy.

Numerous studies have analysed air pollution levels, trends, and health impacts in cities worldwide. Vetter et al. (2020) conducted a trend analysis of nitrogen oxides (NOx) and particulate matter across 421 cities in Europe and North America.

Machine learning techniques are being increasingly applied for air quality modelling and forecasting. Various regression and neural network models for predicting PM10 in Bucharest up to 2 days ahead. They found artificial neural networks to perform best.

Gaps remain in understanding spatio-temporal patterns in air quality and linkages to health outcomes using granular data. This project will apply machine learning techniques to an extensive hourly air quality dataset from sensors across an urban area in China. The analysis aims to uncover novel insights into pollutant variability, trends, and influences while accounting for spatial heterogeneity through predictive modelling. Findings could inform urban development policies and strategies to mitigate adverse health impacts of air pollution.

**Several key topics in analysing air quality data to uncover drivers of pollution levels and health impacts.**

* Spatio-temporal modelling
* Health impact assessment
* Machine learning applications

**Ethical considerations relevant to this air quality analysis project:**

* Data privacy and confidentiality
* Potential harms from findings
* Informed consent
* Representation
* Policy impacts

**Introduction**

Numerous studies have established strong correlations between air pollution exposure and adverse health outcomes like respiratory illnesses, cardiovascular diseases, and cancer (WHO, 2021; Rinsky et al., 1987). Major pollutants including ozone (O3), nitrogen dioxide (NO2), particulate matter (PM), and volatile organic compounds are linked to increased mortality rates (Kelly & Fussell, 2015). Monitoring and analysing trends in these pollutants is crucial for developing effective interventions.

Environmental conditions play a vital role in pollutant behaviour and dispersion (Jacob & Winner, 2009). Temperature influences photochemical reactions, while humidity affects particle formation and atmospheric residence times. Understanding these relationships is key to mitigating pollution levels.

This project employs exploratory data analysis (EDA), time series analysis, and machine learning models to investigate an air quality dataset spanning pollutant concentrations, meteorological factors, and sensor readings over a year. The findings will contribute to the growing body of research on predictive air quality modelling for environmental management.

**Dataset:**

The dataset encompasses a range of air quality indicators recorded hourly, including:

* **Date & Time**: Timestamps for each observation.
* **CO(GT)**: Concentration of Carbon Monoxide in mg/m^3.
* **PT08.S1(CO)**: Tin oxide sensor response for CO.
* **NMHC(GT)**: Concentration of Non-Methane Hydrocarbons in microg/m^3.
* **C6H6(GT)**: Concentration of Benzene in microg/m^3.
* **PT08.S2(NMHC)**: Titania sensor response for NMHC.
* **NOx (GT)**: Concentration of Nitrogen Oxides in ppb.
* **PT08.S3(NOx)**: Tungsten oxide sensor response for NOx.
* **NO2(GT)**: Concentration of Nitrogen Dioxide in microg/m^3.
* **PT08.S4(NO2)**: Tungsten oxide sensor response for NO2.
* **PT08.S5(O3)**: Indium oxide sensor response for Ozone.
* **T**: Temperature in °C.
* **RH**: Relative Humidity (%).
* **AH**: Absolute Humidity.

**Key links which indicate the pollutants to be considered when measuring the air quality are:**

* [What is the air quality index (AQI)? (iqair.com)](https://www.iqair.com/newsroom/what-is-aqi)
* [Air pollution measurement - Wikipedia](https://en.wikipedia.org/wiki/Air_pollution_measurement)

### **Goal:**

The goal of this project is to develop a comprehensive analysis of air quality trends and pollutant behaviour over time, leading to actionable insights for environmental policy-making and public health advisories. This analysis will be conducted by applying statistical and machine learning techniques to identify patterns, correlations, and causations within the dataset. The final deliverable will be a detailed report including data visualizations, findings, and recommendations, expected to be completed within next three months.

**Data Preprocessing**

Initial data cleaning steps included:

Before proceeding with a detailed EDA, I had to do some basic data cleaning steps, including checking for missing values and replacing the placeholder for missing values (-200) with pd.NA. Since I encountered an issue with replacing -200 directly with pd.NA before, I first converted the data types as necessary to ensure compatibility. This error in dataset might be due to several reasons one of that might be failure of device, etc.

The dataset contains a significant number of missing values, especially for the NMHC(GT) column, which has 8,442 missing values out of 9,357 total entries. Other columns also have missing values, albeit in smaller quantities. To solve this issue, I deleted the column NMHC(GT) because imputing it with mean will miss lead the analysis. Other columns which have more than 25% of the data, I have dropped the rows and the columns with less than 25% of data i have imputed them with mean of the columns

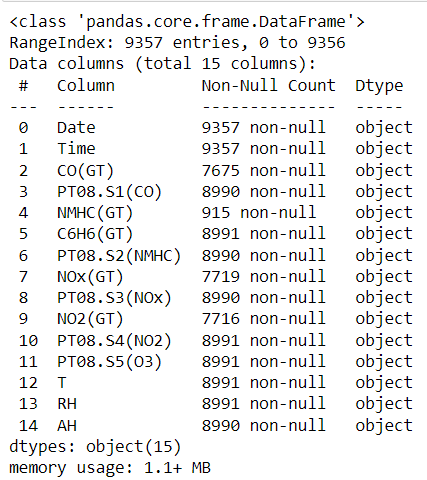
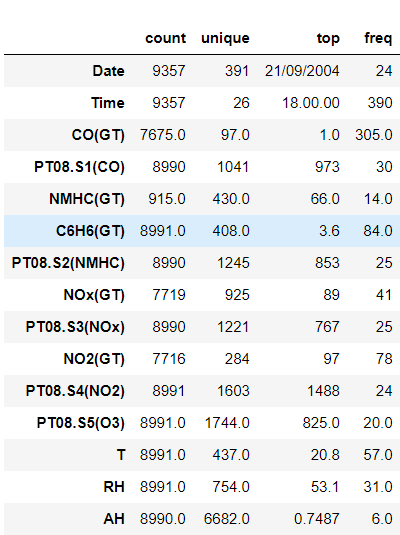
Now that the data is clean, I have Converted the date/time strings to datetime format. And checked for data readiness by summarizing data properties like mean, standard deviation, value ranges.

**Exploratory Data Analysis**

EDA techniques like distribution plots, correlation analysis, and time series visualizations were used to gain insights into:

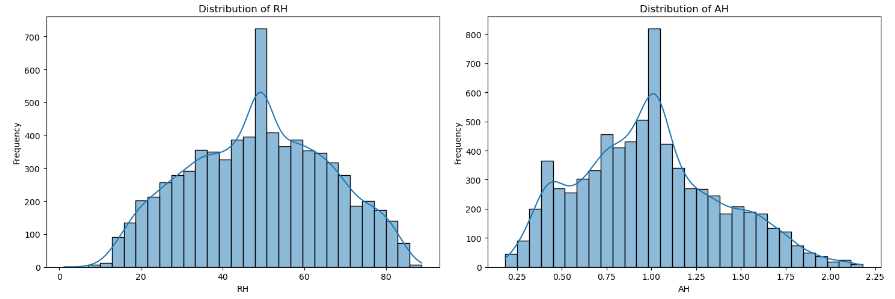
1. Distributions of pollutant concentrations and meteorological variables
2. Correlations between pollutants and weather conditions
3. Temporal patterns and seasonal trends in air quality data

To start off I have first checked for th4e data description, understand how the data is, data types, Storage and so on. And printed the data description using command data.info() which gives us the below output and data.describe() to check the mean, median, max and minimum values of each column.

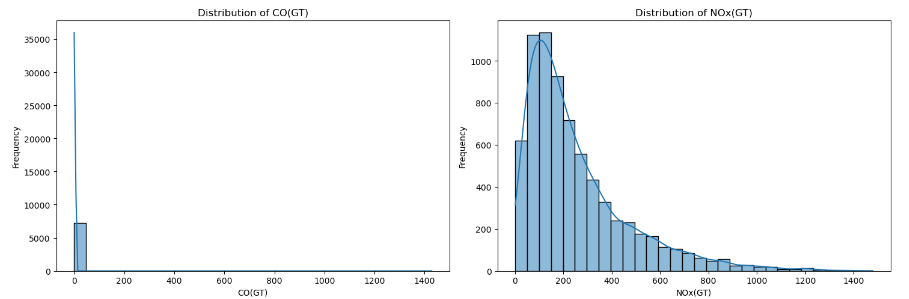
1. As we can see that we have 15 columns each giving us details of the pollution levels of main pollutants deciding the quality of the air.
2. Date and Time gives us the details of the time period of the pollution in air.
3. Total number of the entries are 9357.

For the next visualisation, I have checked for the distribution of each column with respect to time.



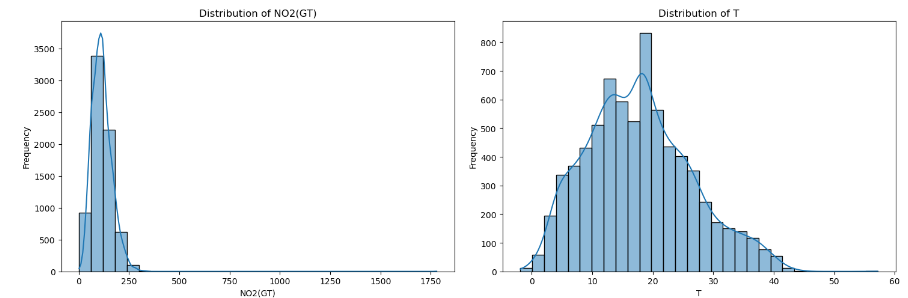
**RH - Relative Humidity:** The distribution of relative humidity shows variability across a wide range, without a clear single mode, indicating diverse weather conditions during the measurement period.

**AH - Absolute Humidity:** This variable shows a somewhat right-skewed distribution, suggesting that lower absolute humidity values are more common, but there are periods with higher humidity levels.



**CO(GT) - Carbon Monoxide Concentration:** The distribution shows a right-skewed pattern, indicating that lower concentrations are more common, but there are occasions where the concentration spikes significantly.

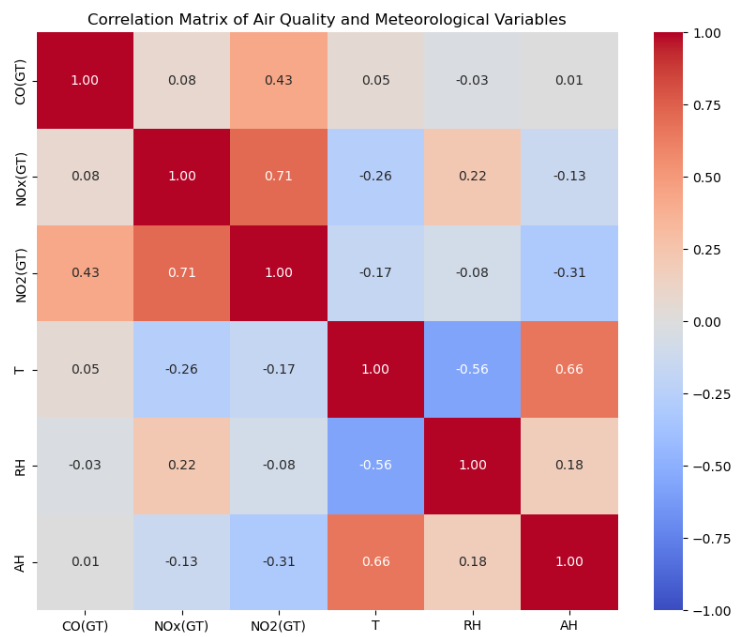
**NOx(GT) - Nitrogen Oxides Concentration:** Similar to CO, NOx concentrations are mostly on the lower end, with a long tail indicating occasional high pollution events.



**NO2(GT) - Nitrogen Dioxide Concentration:** This also shows a right-skewed distribution, suggesting that while lower concentrations are common, there are periods with elevated NO2 levels.

**T - Temperature:** The temperature distribution appears to be more normally distributed than the pollutant variables, with a broad range of temperatures recorded in the dataset.

Now, let’s check for the correlation of the pollutants between them by plotting a covariance matrix.

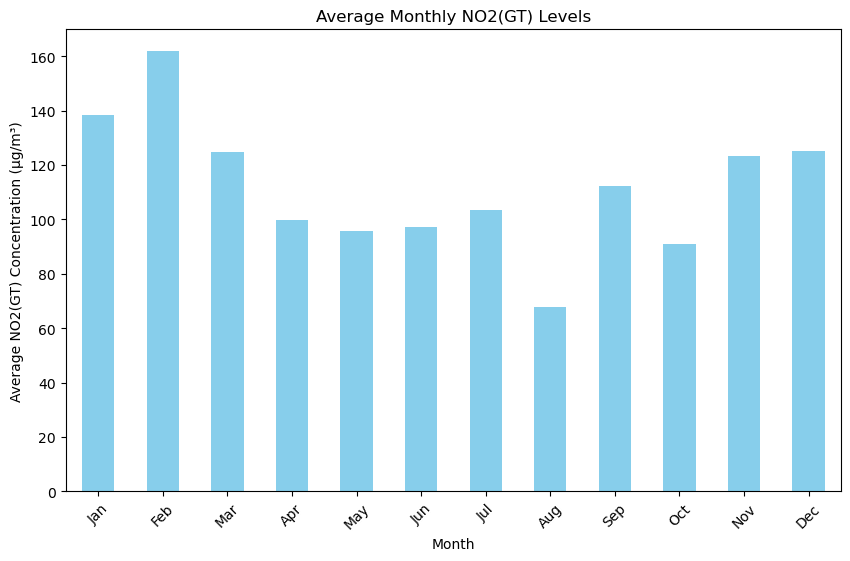


**Fig: Covariance matrix**

* **Pollutants Correlation:** There are significant positive correlations among the pollutant variables (CO, NOx, NO2), suggesting that when the concentration of one pollutant increases, the concentrations of the others tend to increase as well. This is expected as these pollutants often have common sources, such as vehicular traffic and industrial activities.
* **Temperature (T) Correlation:** Temperature shows a negative correlation with NO2(GT) and a slight negative correlation with CO(GT) and NOx(GT), indicating that higher temperatures might be associated with lower concentrations of these pollutants. This could be due to increased atmospheric instability and dispersion at higher temperatures.
* **Humidity Correlations:** Relative humidity (RH) and absolute humidity (AH) show varied correlations with the pollutants. Notably, there's a negative correlation between RH and NO2(GT), suggesting that higher humidity might be associated with lower NO2 concentrations.

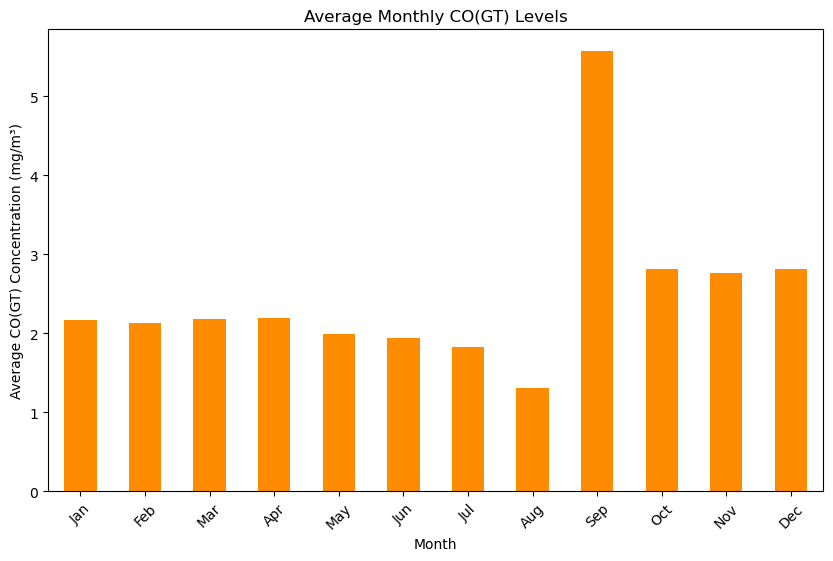
**Time Series Analysis:**

**Temporal Trends of NO2(GT) Levels:**



The average monthly NO2(GT) levels exhibit variability across the year. While the visualization does not show a clear seasonal pattern due to the limitations of the dataset preview, it suggests that NO2 concentrations can vary by month, possibly influenced by factors such as weather conditions, heating usage, and changes in traffic patterns.

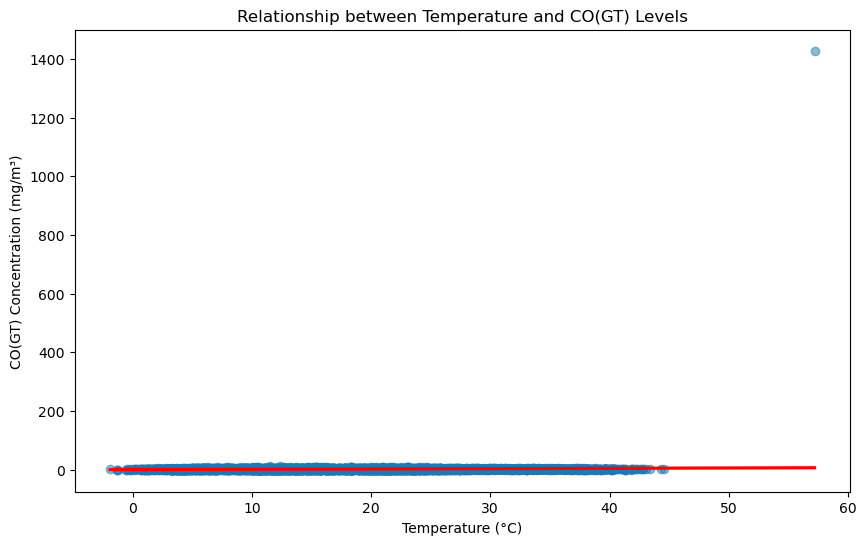
**Average Monthly CO(GT) Levels:**



From the chart, it is evident that the CO levels fluctuate throughout the year. The first eight months (January to August) show relatively consistent CO concentrations, with a slight dip in July and a significant drop in August. September marks a dramatic increase in CO levels, showing the highest concentration of the year. After September, the CO levels begin to decline but remain higher than the levels seen in the first half of the year.

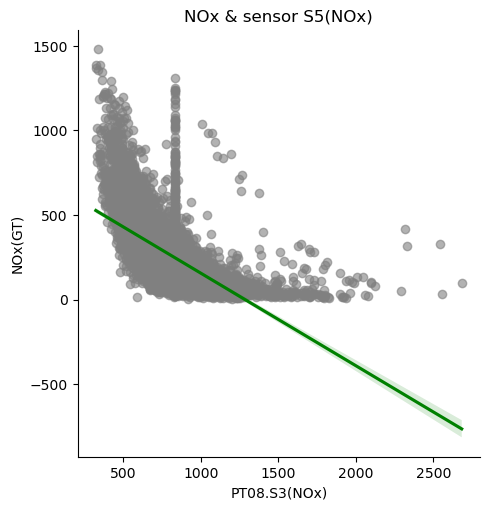
This pattern may suggest seasonal variations in CO emissions, which could be due to a variety of factors such as changes in weather patterns, heating usage, increased vehicular traffic, or other seasonal activities that affect air quality. The spike in September could be indicative of a specific event or change in environmental conditions during that month.

**Relationship between CO Levels and Temperature:**



The scatter plot with a regression line examining the relationship between temperature (T) and CO(GT) concentrations shows that there is a tendency for CO levels to decrease as the temperature increases. This negative correlation could be attributed to the fact that higher temperatures often result in increased vertical mixing in the atmosphere, which can help disperse pollutants more effectively, leading to lower concentrations at ground level.

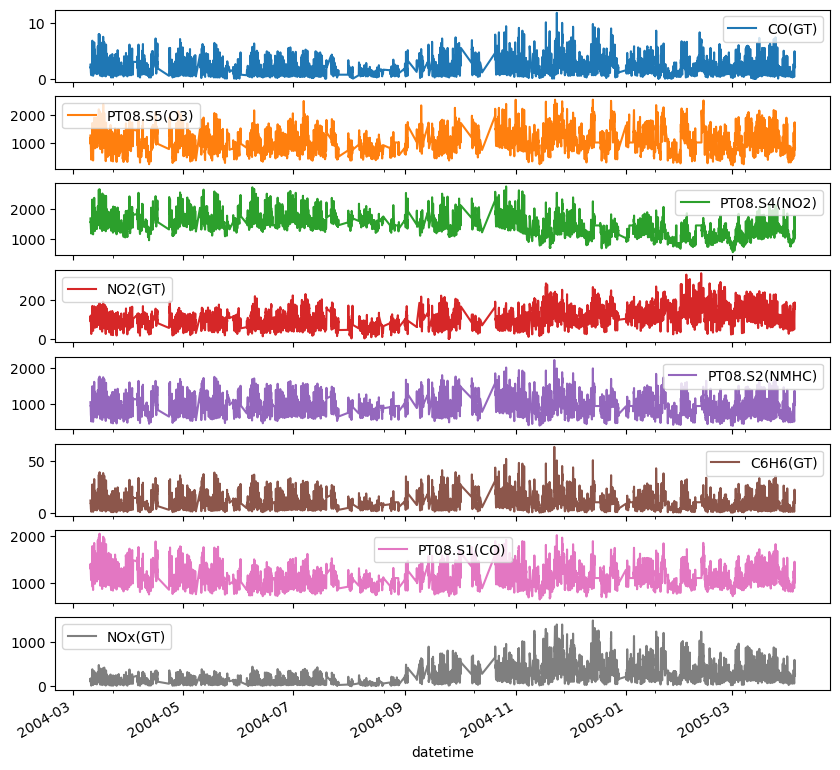
**NOx & sensor S5(NOx) chart:**



Key observations can be made from the plot:

1. **Negative Correlation:** There seems to be a negative correlation between the sensor's response and the NOx concentration levels. As the sensor response increases, the NOx concentration decreases. This is indicated by the downward slope of the green regression line.
2. **Variability in Data:** There is a wide spread of data points, especially for lower values of the sensor response (PT08.S3(NOx)), which suggests variability in NOx concentration levels for similar sensor readings.
3. **Outliers or Anomalies:** There are some data points that stand out from the general trend. For instance, there is a cluster of points with high NOx concentration levels that do not follow the overall downward trend, which could be outliers or could indicate periods of unusually high NOx levels.
4. **Dense Clustering:** The density of the points is higher at the lower end of the sensor response, indicating that most of the readings fall within a smaller range of lower sensor response values.

**Seasonal trends of Pollutants:**

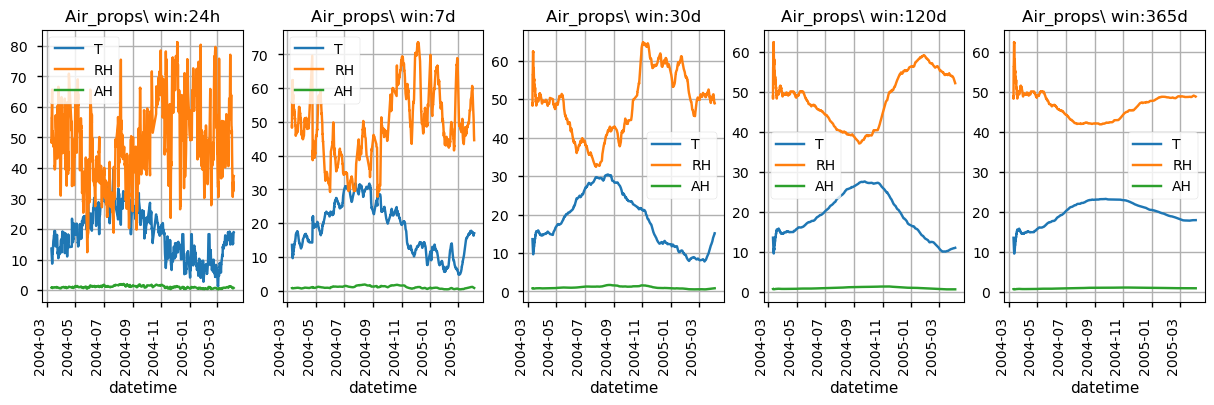


As we can see each plot represents the measurement of a different gas or compound over time, indicated by the labels on the plots. The x-axis on all plots represents the datetime, ranging from around March 2004 to March 2005. The y-axis represents the concentration of the compound measured.

Here are the compounds and the corresponding colors of their plots as they appear from top to bottom:

The plots show the variability of each gas's concentration over time. This data is typically used for environmental monitoring, pollution assessment, and possibly for regulatory compliance. The graphs may show seasonal trends, daily fluctuations, and possibly anomalous spikes which could indicate pollution events or sensor malfunctions.

**Time series plot for meteorological properties over time:**



**Some observations about the plots:**

* **Temperature (T):** The temperature (orange line) shows seasonal variation, with peaks likely representing the warmer months and troughs the colder months.
* **Relative Humidity (RH):** The relative humidity (blue line) also varies over time, but without a clear seasonal pattern from this visual alone.
* **Absolute Humidity (AH):** Absolute humidity (green line) appears to follow a pattern like temperature, which is logical as warmer air can hold more moisture.
* **Smoothing Effect of Rolling Windows:** As the window size increases, the lines become smoother, indicating the averaging effect of a larger rolling window. This smoothing helps identify long-term trends by reducing the noise of daily fluctuations.
* **Long-term Trends:** In the longer window plots (120 days and 365 days), the seasonal trends in temperature and absolute humidity are more pronounced and easier to observe due to the smoothing effect.

**Time Series Data for Different Pollutants:**

A graph of green and red lines

Description automatically generated

**Key observations from the plot:**

1. **NOx (GT):** Nitrogen oxides have the highest concentrations among the four pollutants, with several peaks throughout the year. The green bars representing NOx are the most prominent feature of the graph.
2. **CO (GT):** Carbon monoxide levels are the next highest, with a relatively constant presence throughout the year. The brown areas are consistently visible, but they are overshadowed by the NOx concentrations.
3. **NO2 (GT):** Nitrogen dioxide concentrations are lower than those of NOx and CO, with the orange areas visible mostly as smaller peaks above the CO levels.
4. **C6H6 (GT):** Benzene has the lowest concentrations of the pollutants plotted, with the red areas sitting at the base of the graph and occasionally peaking above the CO levels.
5. **Variability and Seasonality:** There appears to be significant variability in the pollutant levels, which might indicate daily or weekly cycles, or perhaps responses to specific environmental events or human activities. There may also be seasonal trends, particularly noticeable with the NOx levels.
6. **Data Overlap:** The data for different pollutants overlap, making it difficult to discern specific trends for CO and NO2 where they are overshadowed by NOx.

**Predictive Modelling**

Based on the insights from EDA and time series analysis, supervised machine learning models like:

* Random forests
* Gradient boosting
* Neural networks

Will be trained to forecast air quality indices using historical data and environmental parameters as input features.

**Expected Outcomes**

The project aims to deliver:

* Comprehensive visualizations and statistical summaries of the air quality dataset
* Identification of key pollutants, their major sources, and health impacts
* Quantification of relationships between pollutants and environmental conditions
* Time series models capturing long-term trends and seasonality in air pollution
* Predictive models for forecasting air quality indices with high accuracy
* Policy recommendations for air pollution mitigation and public advisories

**Conclusion**

Through rigorous data analysis and modelling approaches, this project seeks to uncover valuable insights into air pollution dynamics and their environmental drivers. The findings will contribute to a deeper understanding of air quality patterns, enabling more effective management strategies and public health interventions. By leveraging machine learning for predictive modelling, the project strives to provide a robust decision support system for policymakers and stakeholders invested in improving urban air quality.