**2.1 Background and Project Goal**

**2.1.1 Project Background**

Air quality is an important index to indicate the effect of Climate change and Global warming as the overall air quality is gradually changing with the change of the rates of greenhouse gases and other factors which decide the air quality. It has been a critical issue due to drastic shift to urbanization and rapid industrialization worldwide. Due to the decrease of the air quality some dangerous emissions of pollutants are arising, which have negative impacts on health and environment. It will be noted that researchers and scientists are monitoring the important contributors such as Particulate Matter (PM2.5, PM10); Nitrogen dioxide (NO2); Sulfur dioxide (SO2); Carbon monoxide (CO); Ozone (O3) because these pollutants directly impact respiratory and cardiovascular diseases. Even these pollutants are the key ones which cause global warming which leads to climate change. People all over the world are being conscious about these matters. As a result, there are some open doors to discuss the solution regarding Air Quality Index (AQI), which also needs studies about these key pollutants.

At the moment we are in the process of devising a dataset which is known as the ‘Global Air Quality’; this dataset entails the following important measurements of air quality from some of the most distinguished cities all over the world. This dataset also has significant predictors related to environmental standards of air fitness for example Particulate Matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3) besides the climatic factors such as temperature, relative humidity and wind speed. This dataset has 10000 records highly important for research, data analysts and policy makers to track arial quality, determine the effects of pollution on the health of citizens, and looking for chance to improve the air quality.

**2.1.2 Explanation of Project Goal**

The dataset which is going to be used includes several vital pollutions like PM2.5, PM10, NO2, SO2, CO, O3 and climate, temperature, humidity, wind speed and other factors involved in the project are going to develop a model to monitor air quality. The model aims to provide relevant information concerning the factors behind the emissions of poor air and possible impact on human health, especially the cardio-respiratory diseases. Thus, utilizing this dataset, the project wants to contribute to understanding the link between pollution and health by providing researchers, data scientists, and policymakers with information and reference data for monitoring and increasing air quality in industrialized and urban environments around the world.

**2.2 Data Set Description**

**2.2.1 Data Characteristics and Source**

Our dataset provides comprehensive air quality and environmental data across various global cities, featuring a blend of pollutant measurements and weather indicators. Each row represents data from a specific location and date, with the following 4 key characteristics:

**Location Information**

* **City** and **Country**: Indicate where the measurements were taken, allowing for geographic comparisons and insights into pollution levels across different urban areas worldwide.

**Date**

* **Date**: The specific day on which the air quality and environmental conditions were recorded, in a dd-mm-yyyy format. This allows for time-based analysis, such as identifying trends or seasonal variations in air quality.

**Air Pollutant Levels**

* **PM2.5** and **PM10**: Particulate matter measurements in micrograms per cubic meter (µg/m³). PM2.5 (fine particles) and PM10 (coarse particles) are crucial indicators of air pollution, with significant impacts on health. Fine particles, in particular, can penetrate deep into the lungs, making these readings vital for assessing air quality.
* **NO2** (Nitrogen Dioxide), **SO2** (Sulfur Dioxide), **CO** (Carbon Monoxide), and **O3** (Ozone): Concentrations of these gases, measured in µg/m³. Each of these pollutants affects air quality and has distinct sources and impacts. For instance, NO2 is typically associated with vehicle emissions, while high levels of O3 are linked to sunlight reacting with other pollutants, posing respiratory health risks.

**Environmental Conditions**

* **Temperature** (°C): The ambient temperature, which can influence pollutant dispersion and chemical reactions in the air.
* **Humidity** (%): The relative humidity level, as moisture in the air can impact how pollutants are suspended or dispersed.
* **Wind Speed** (m/s): The speed of wind at the time of measurement, which affects pollutant distribution. Higher wind speeds can help disperse pollutants, potentially improving air quality.

**2.2.2 High-Level Statistics**

Our analysis provides a robust statistical summary of the numeric data in the dataset, ensuring relevance and clarity.

**1. Central Tendencies (Mean & Median):**

* **Mean**: The average values across numeric columns indicate the general level of the data. This highlights the overall magnitude of each variable, which is essential for understanding trends.
* **Median**: This measures the middle value, offering a sense of the typical data point and helping to mitigate the effects of extreme values (outliers).

#### **2. Variability (Standard Deviation & Variance):**

* **Standard Deviation**: This quantifies the spread of data around the mean. Columns with high standard deviation have significant variability, signaling inconsistencies in data.
* **Variance**: Representing the squared dispersion, variance complements the standard deviation and helps in identifying highly fluctuating columns.

#### **3. Range (Minimum & Maximum):**

* These metrics outline the bounds of data, indicating the smallest and largest values. This range provides insights into data scale and identifies potential anomalies.

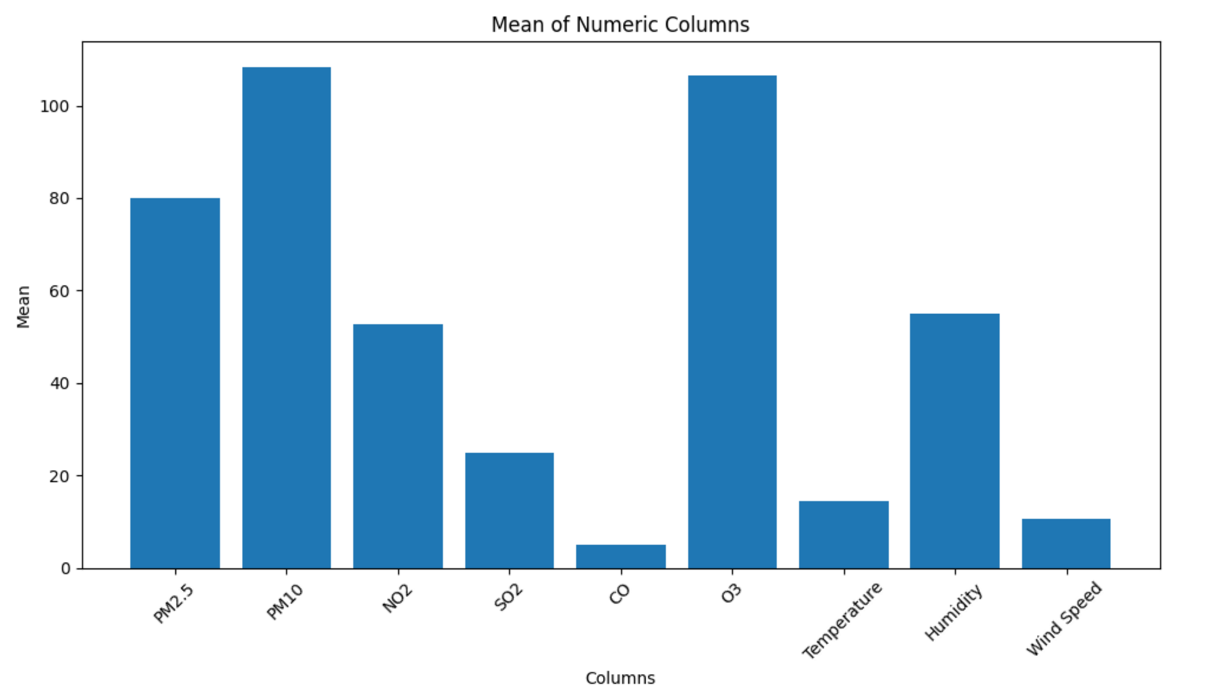
#### **4. Distribution Shape (Skewness & Kurtosis):**

* **Skewness**: Captures the asymmetry in data distribution. Positive skewness indicates a tail to the right, while negative skewness signifies a tail to the left. This helps understand if the data is balanced or biased toward certain values.
* **Kurtosis**: Highlights the "tailedness" or the extremity of data points. A higher kurtosis signifies more extreme values (heavy tails), while lower kurtosis indicates a flatter distribution.

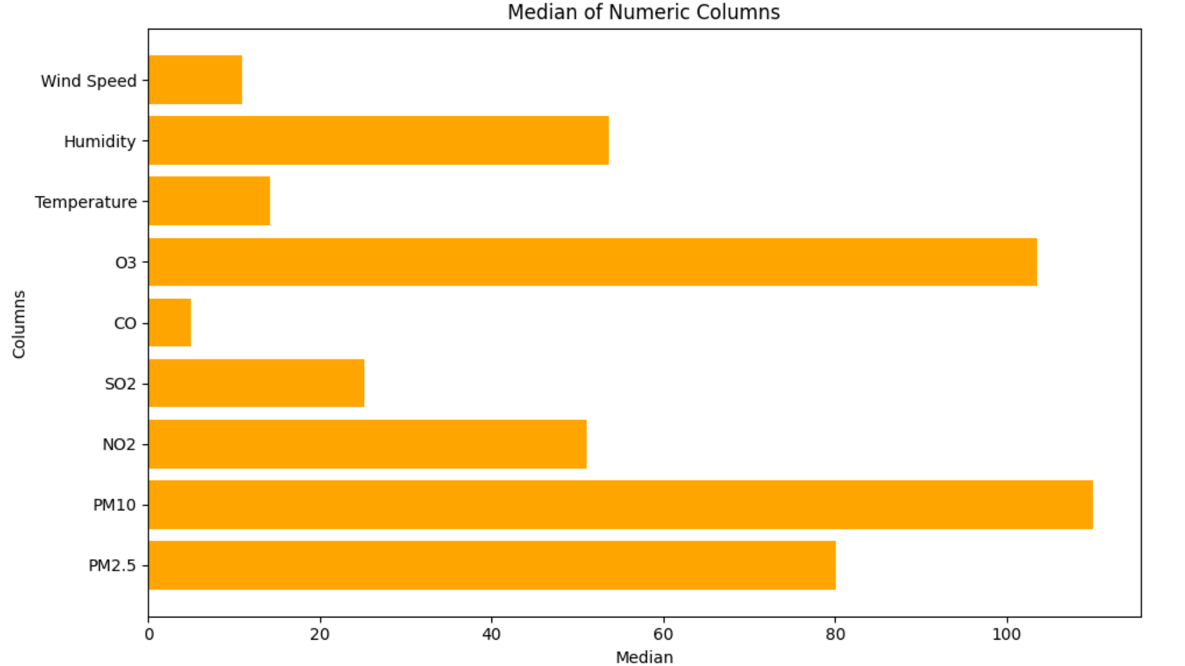
#### **5. Visual Representations:**

We employed diverse visualizations for each statistical aspect:

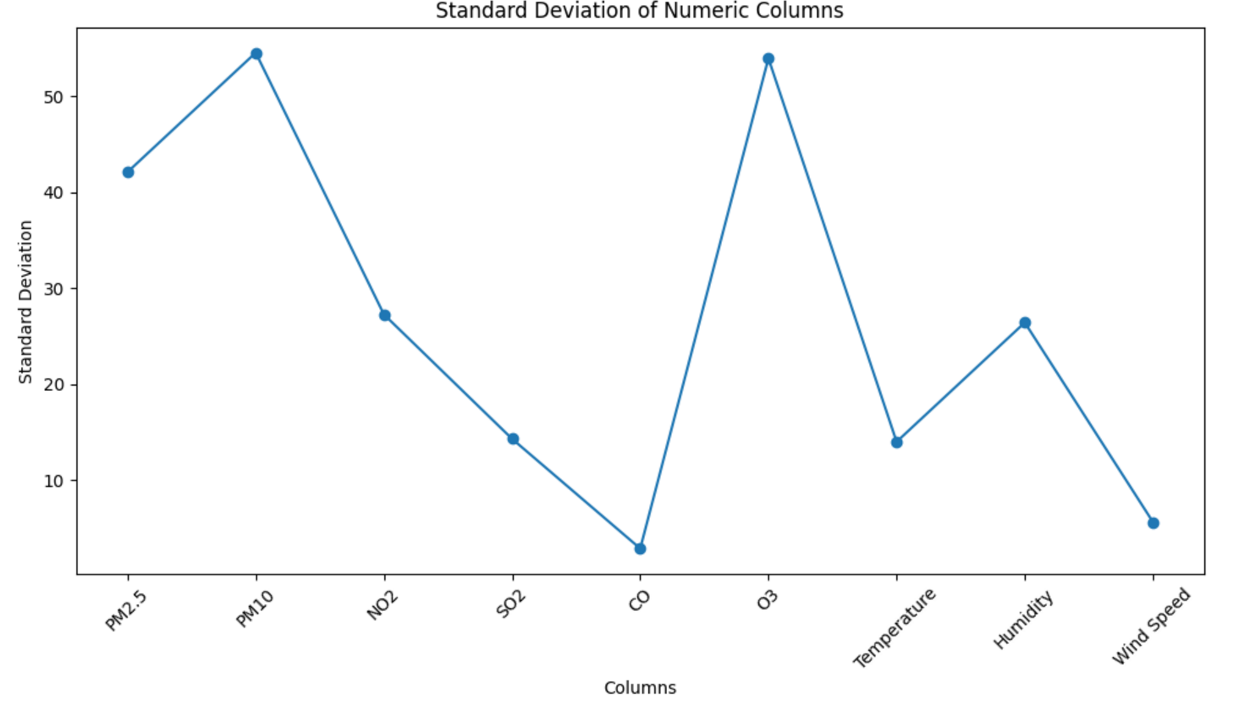
* **Bar Charts**: To compare mean and range values across columns.



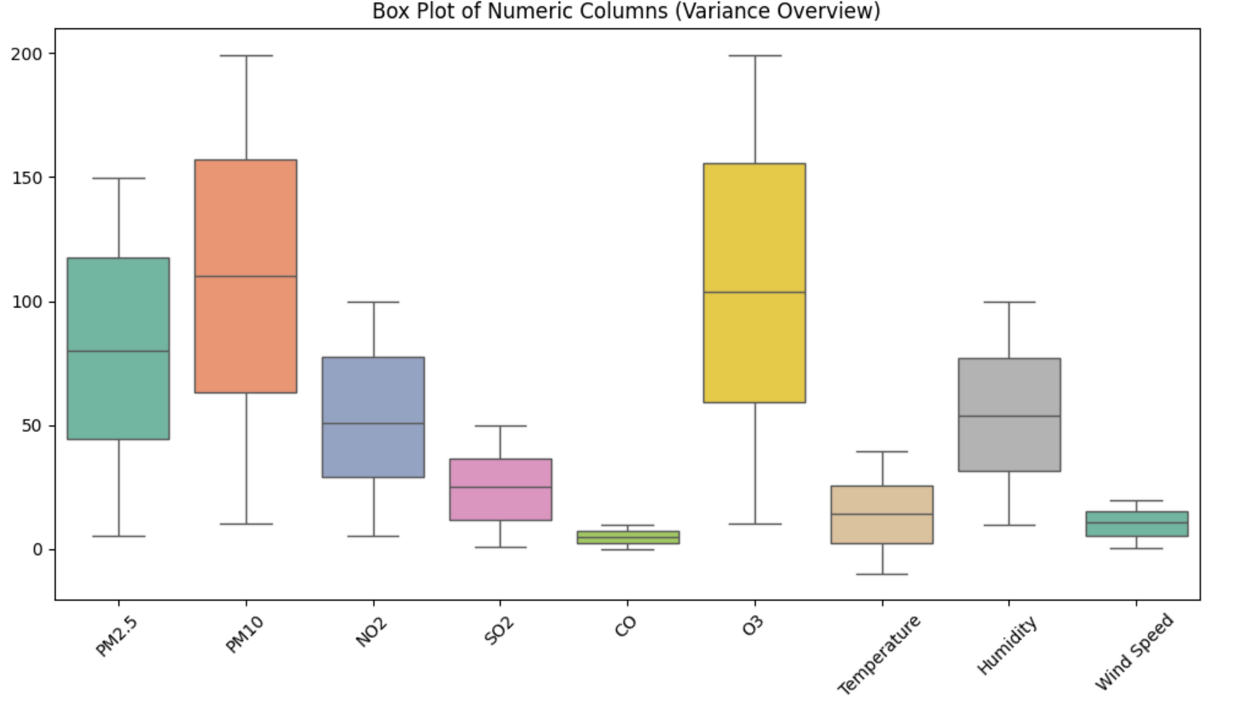
* **Horizontal Bar Charts**: Used for median values to provide an alternative orientation.



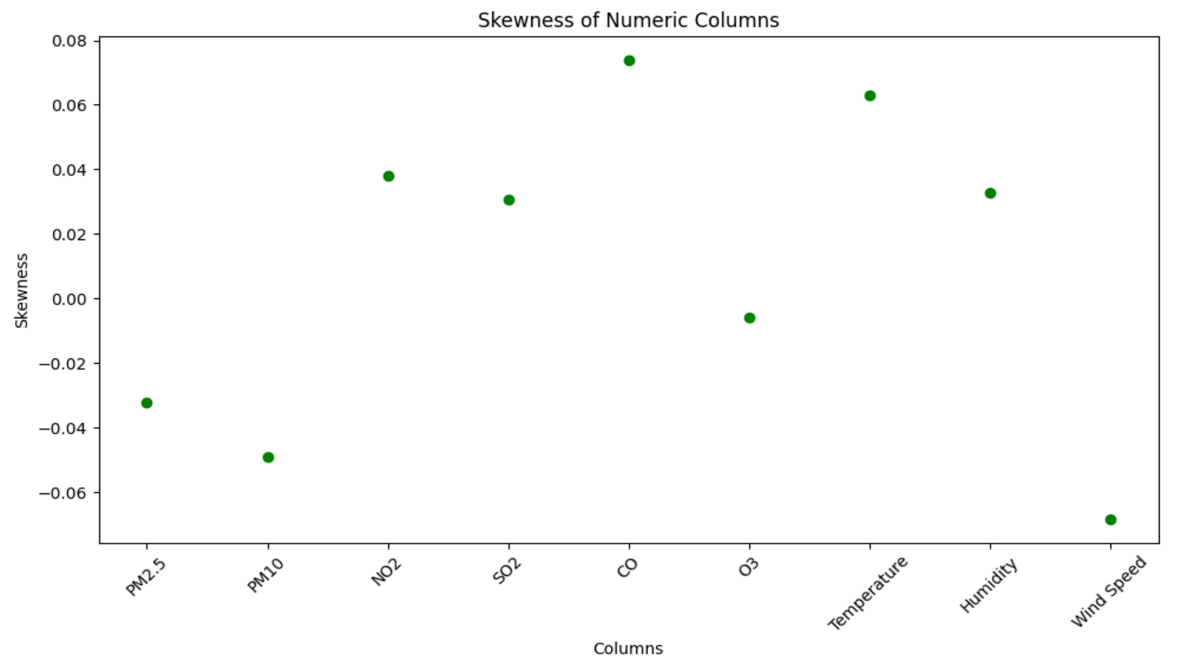
* **Line Charts**: To showcase the dynamic trends in standard deviation.



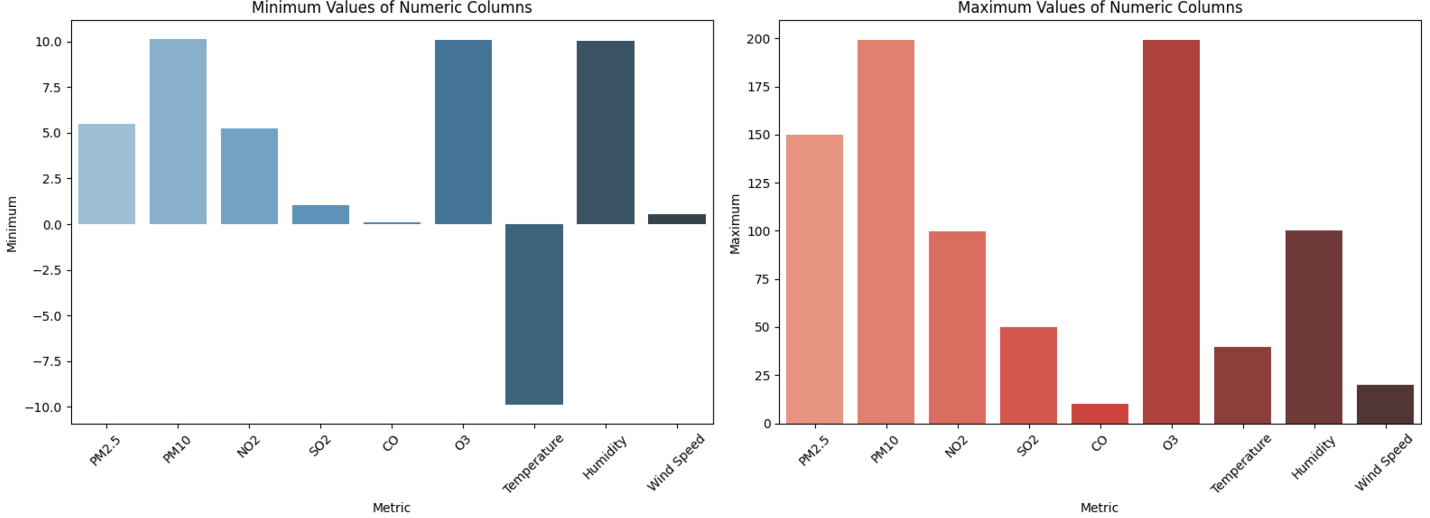
* **Box Plots**: To visually represent variance and identify potential outliers.

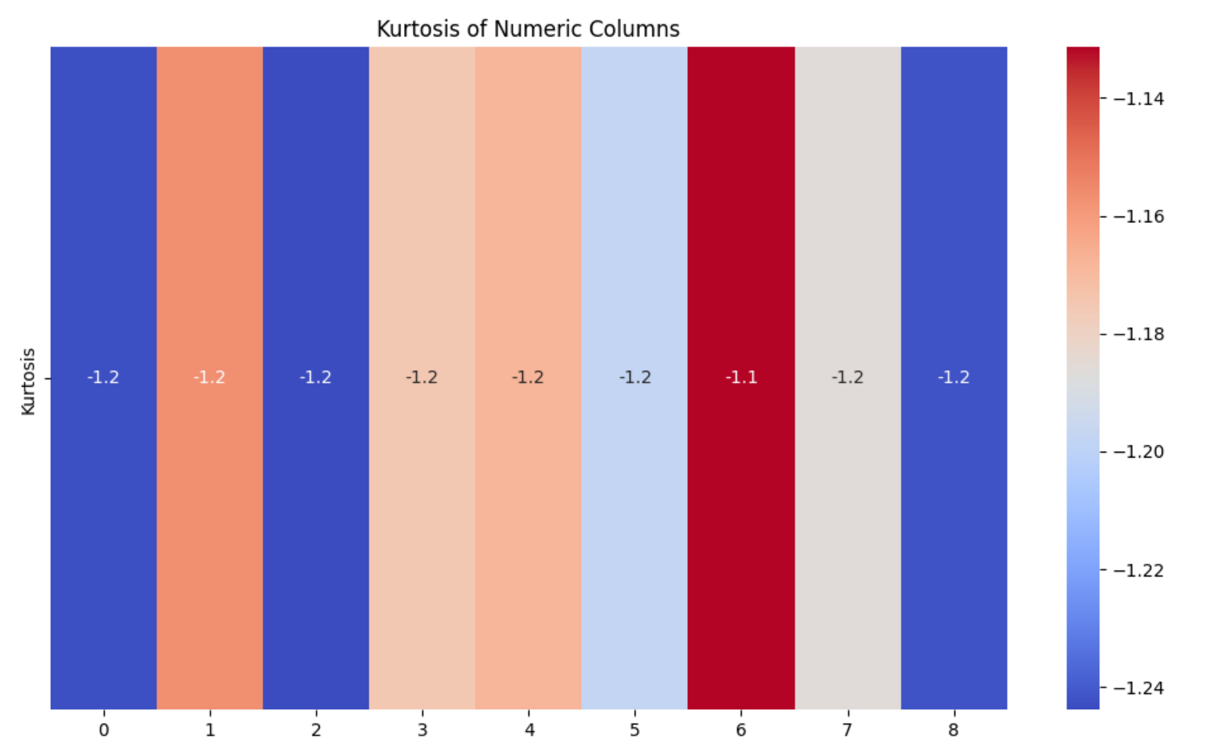


* **Scatter Plots**: To analyze skewness and highlight data symmetry.

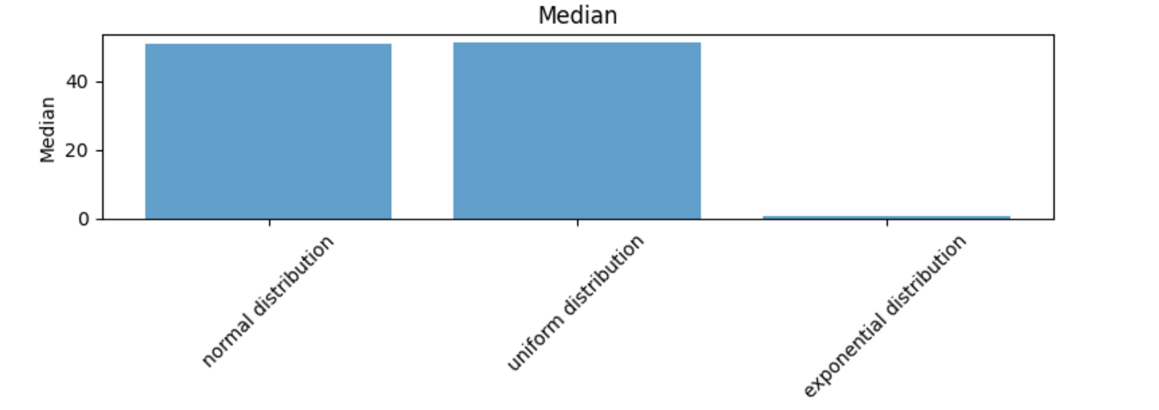
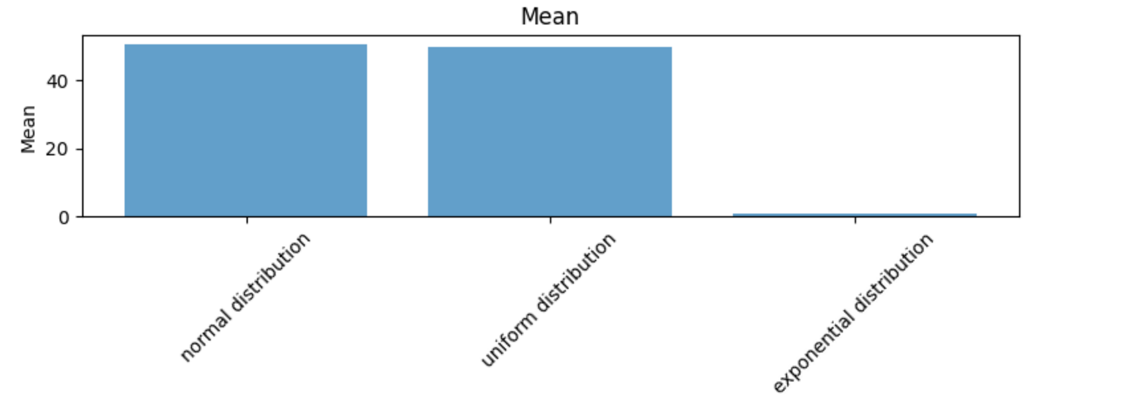


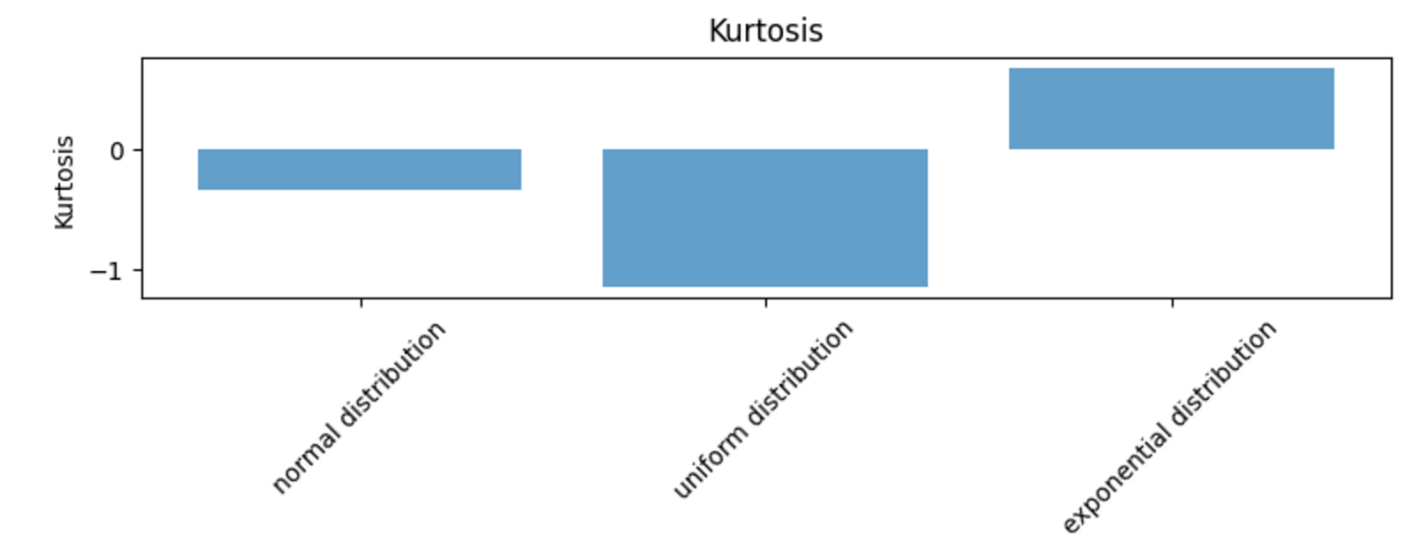
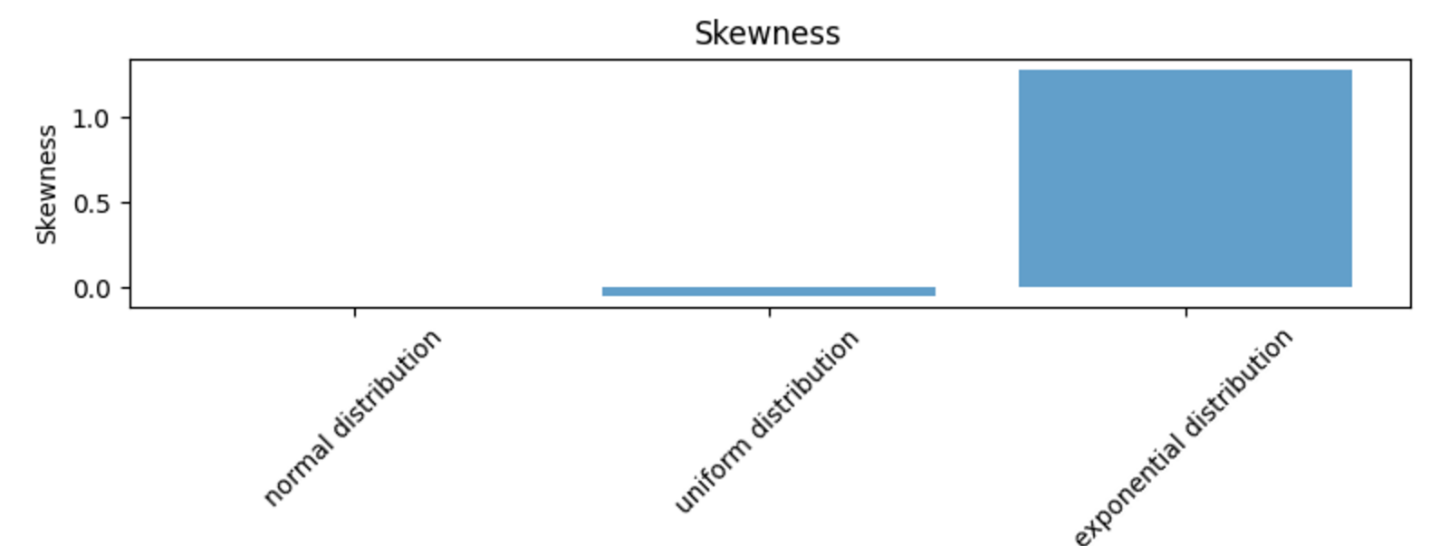
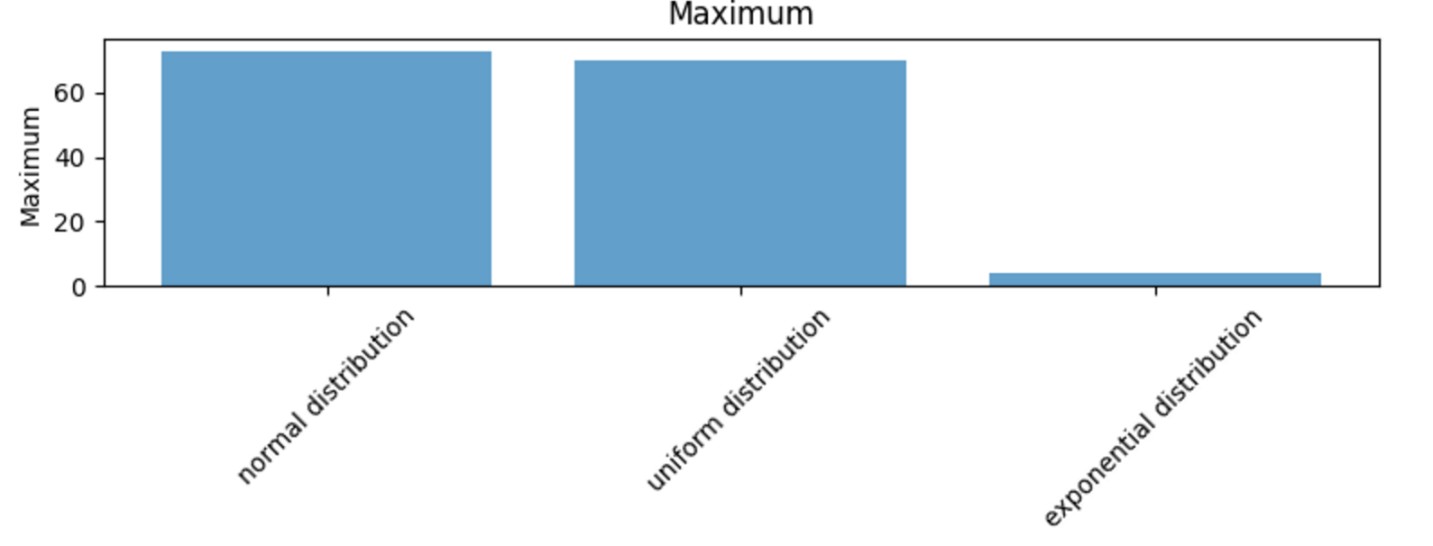
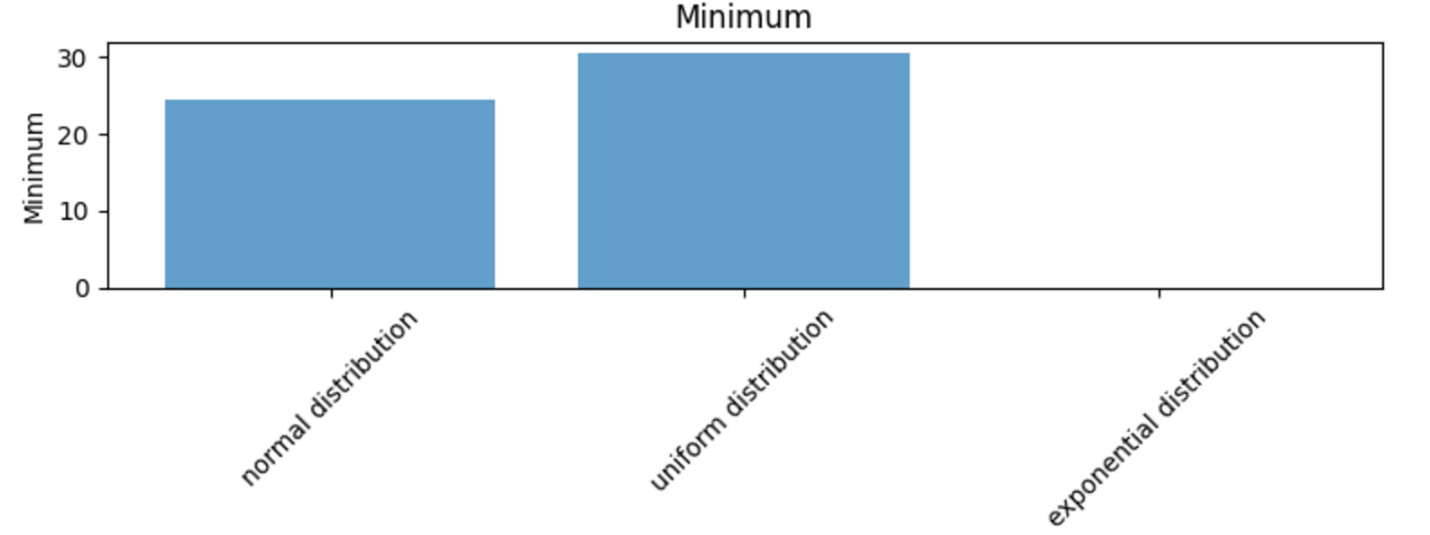
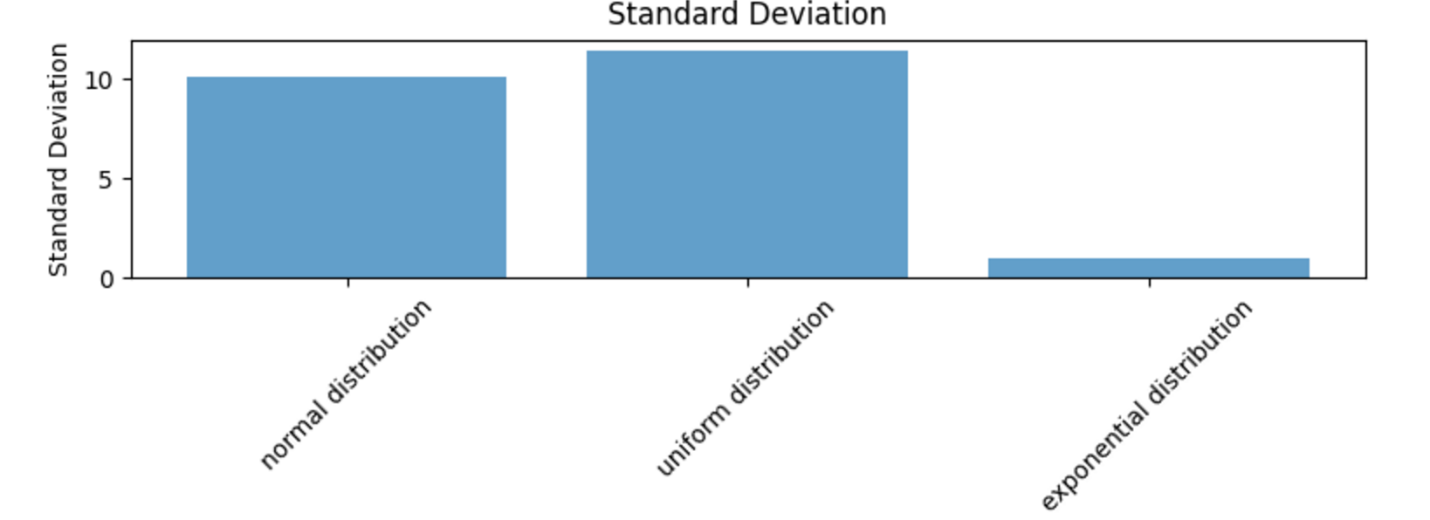
* **Side-by-Side Bar Charts:**To compare minimum and maximum values within the same chart, highlighting the range and variability for each numeric column.



* **Heatmaps**: To provide a concise view of kurtosis values, facilitating comparisons across variables.

# **6. Numeric Data Distribution in Each Statistics**

* **Normal Distribution** is balanced and ideal for modeling symmetric real-world phenomena.
* **Uniform Distribution** shows equal probability, making it suitable for random sampling scenarios.
* **Exponential Distribution** captures phenomena with rapid decay and occasional extreme values, often used in reliability and survival a



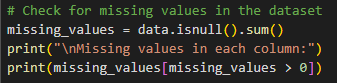
**2.3 Data Preprocessing and Issues**

**2.3.1 Identification of Data Issues**

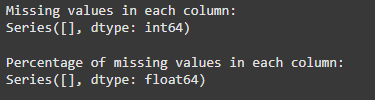
**Handling Missing Values**

Missing values inside any dataset is a very crucial problem for data analysis which results in reducing the strength of statistical analysis and it can create bias in estimations of the parameters. As a result, it reduces representatives of the samples also. So, identifying the missing values and imputing the suitable value is a crucial step for data preprocessing.

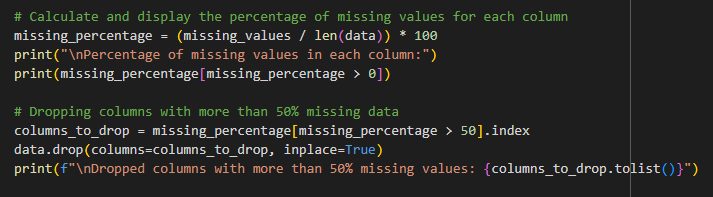
i) By “*data.isnull().sum()*”, we checked for missing values across all the columns. Then we gave the command to print all the missing values.



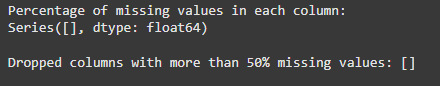
Implementing the code, no missing values were found in this dataset:



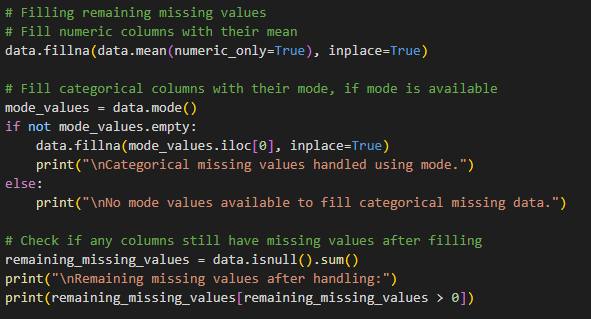
ii) Next step was calculating the percentage of missing values by the percentage formula. And then the dropping column code is there because we wanted to drop the columns with the missing values.

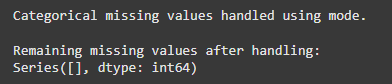


But, as there is no missing values in our dataset, no columns were dropped.



iii) Even though, there was no missing values, there was the mechanism to handle potential issues like numeric columns would have been filled by their mean value and categorical missing values would be filled with mode. And at last, we implemented code to check if any columns still have missing values after imputation.



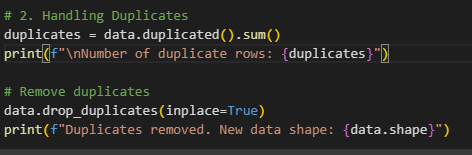


**Handling Duplicates**

Addressing duplicate data is crucial for data quality and analysis. Handling duplicate data involves identifying and addressing repeated or similar records in a dataset. Utilize techniques like data profiling to detect duplicates and employ tools or code to generate summary statistics.

Implementation:

1. Initial checks for duplicates using “data.duplicated().sum()
2. Duplicate removal by “data.drop\_duplicates()”

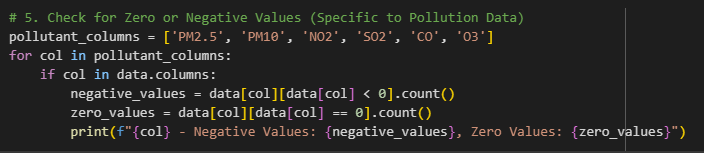


Here no duplicates were found, but still the removal function ensures the robustness of dataset and maintains a clean dataset if there were duplicates. As a result, the data quality would be ensured.



**Check for Zero or Negative values**

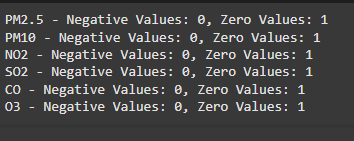
The usage of zero or negative value constructs as key in a database is often frowned at based on practicability and more importantly theoretical realms. These values may contradict and default semantic expectations of by-positive-increment identifiers and can possibly affect auto-increment or indexing performance as well as result in granularity of the store. Further, zero is sometimes used as a flag and negative values are erroneous and ambiguous. Although other database systems permit such keys for more flexibility and in certain instances, sticking to positive integer avoids compatibility issue, data consistency and adheres to common database practice.



**Check Column Presence**: Before proceeding, the code checks if the column exists in the dataset using the condition if col in data.columns.

**Count Negative Values**: For each column, it calculates the number of negative values using data[col][data[col] < 0].count(). This filters rows where the value is less than zero and counts them.

* **Count Zero Values**: Similarly, it calculates the number of zero values using data[col][data[col] == 0].count().
* **Print Results**: The results for each column are printed in the format "{col} - Negative Values: {negative\_values}, Zero Values: {zero\_values}"



The output indicates the following about the pollutant data:

Negative Values: Each of the pollutant columns including PM2.5, PM10, NO2, SO2, CO, O3 does not contain any negative values which is desirable since negative values for any pollutant does not make any physical sense.

Zero Values: In each column there is a zero valued entry. While zero values are not inherently invalid, they might signify:

Missing Data: Lack of documented measurements of the extant risk.

Legitimate Zero: For instance, a location with insignificant levels of the presence of a particular pollutant within the uptake range of the measuring instrument .

Regarding negative values it says that there is no problem with data integrity, however in regard to zero values, it recommends to look at them more closely whether they are of any realistic value in the given data set and according to the goals and objectives of the analysis.

**2.3.2 Preprocessing Techniques**

Provides a comprehensive explanation of preprocessing techniques.

* Encoding Categorical Data
* Feature Scaling
* Feature Engineering

There are some general data preprocessing steps by which one dataset can turn into clean and consolidated dataset. Data preprocessing finds out the issues of the dataset and make the dataset convenient for the analytical work.

Data Preprocessing Steps:

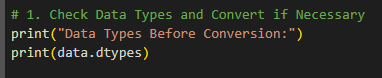
1. Acquiring the Dataset
2. Import all the Libraries
3. Import the Dataset
4. Handling the missing values and issues
5. Encoding the Categorical Values to Numerical Values
6. Feature Scaling
7. Splitting the Dataset

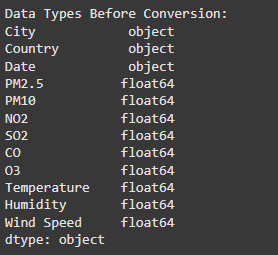
For our dataset, we have tried to implement preprocessing techniques. Among them we tried to do handle the missing values. Also, we needed to check encoding the categorical data as, machine cannot understand categorical data other than the numerical data. Also, handling duplicates, scaling, feature engineering, checking for zero or negative values and making some visualizations to identify distribution to check skewness, outliers or patterns.

**Encoding Categorical Data**

To implement a machine learning algorithm, we have to convert the categorical datas into numerical data because machine doesn’t understand categorical data.

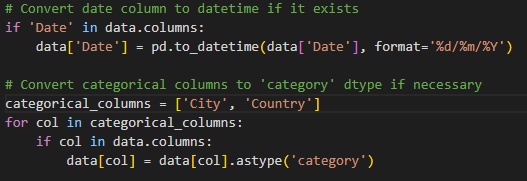
i) First of all, we identified the initial data types of all columns using “data.dtypes” which showed the output of all the initial data typed of all columns.



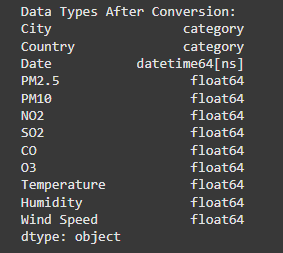


ii) Conversion

* Date Columns: Converted date column from ‘Object’ to ‘datetime64[ns]’ using pd.to\_datetime().
* Categorical Columns: Transformed City and Country columns from ‘object’ to the ‘category’ datatype.



Output:



**Scaling**

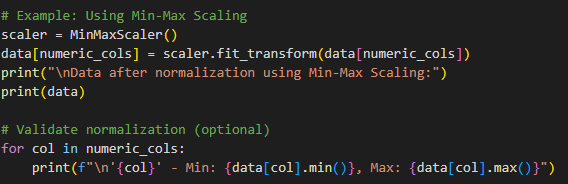
The process of changing a dataset's feature values until they fall inside a predetermined range, such as 0 to 1 or -1 to 1, is known as data scaling. This can help to increase the algorithm's performance by ensuring that no single characteristic dominates the distance calculations.

Implementation:

i) Identify the numeric columns through ‘data.select\_dtypes(include=np.number).column.tolist()’ 

ii) Application of Min-max scaler:

One well-liked feature scaling method in machine learning is the MinMax Scaler. It scales a dataset's features to a predetermined range, usually 0–1. The primary benefit of the MinMax Scaler is that it brings the values into a desired range while maintaining the original distribution's structure.



Implementing this, all numerical column was successfully normalized to the range [0,1]. The scaled data was consistent and non-biased.

**Feature Engineering**

Rescaling each feature to have a mean of 0 and a standard deviation of 1 is known as feature scaling. Nevertheless, feature scaling is not necessary for tree-based techniques such as decision trees, random forests, and gradient boosting because they are invariant to the scale of the features.

Identified Feature Interaction:

Developed a new feature called Temp\_Humidity\_Interaction in order to measure the amount of interaction between temperature and humidity.

Formula applied:



Purpose of the New Feature:

This interaction term is very useful in environmental data sets because cold served as a tracer for temperature and humidity because both variables affect the level of pollution and are interrelated. Many a time the models require an understanding of how these two variables correlate concerning the target variable from which the residuals arise.

Feature Integration:

The new feature was inserted as a new column to the dataset. It was applied on the temperature and humidity dataset without changing its format so it’s usable in isolation if the need arises.

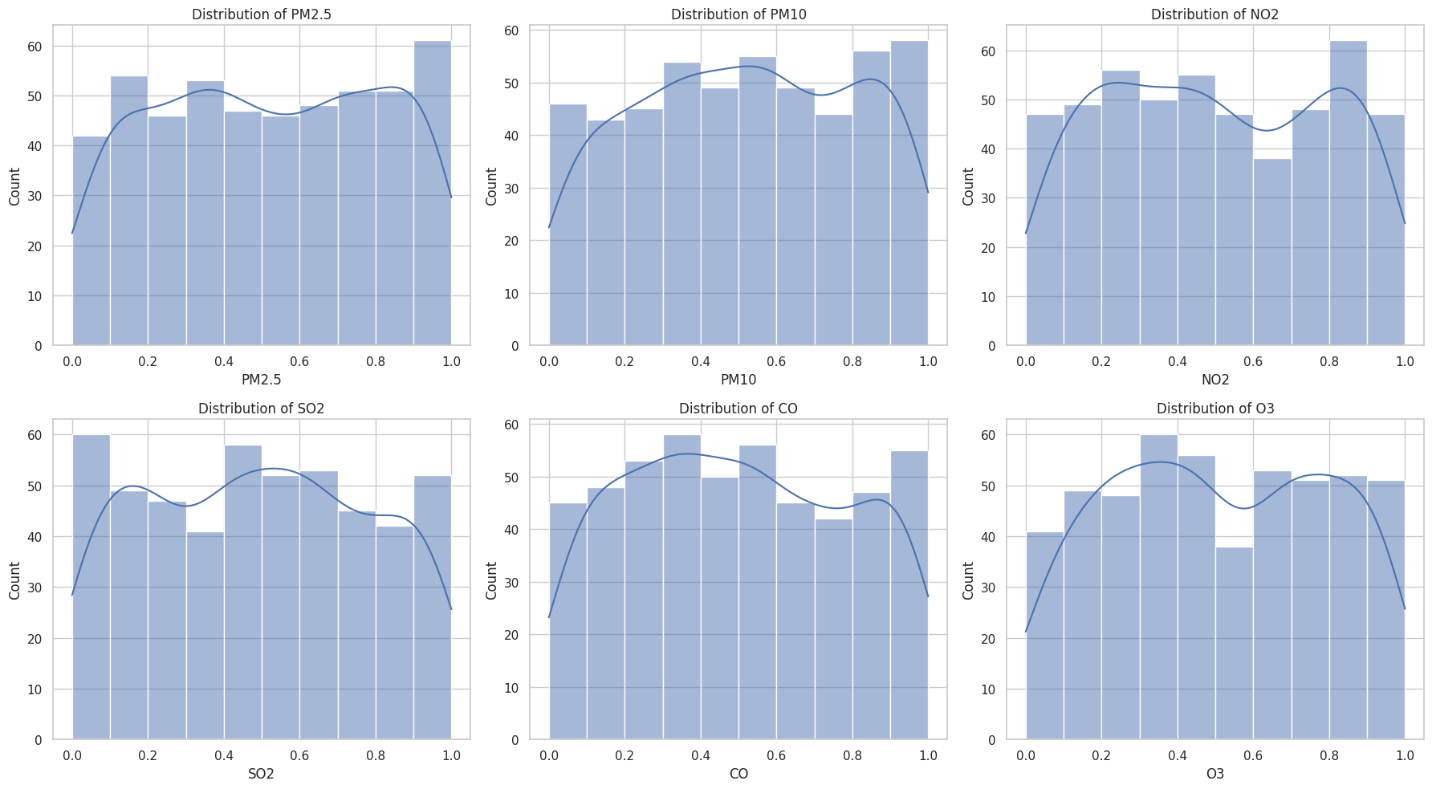
**2.4 Data Science Techniques (35%)**

**2.4.1 Used Techniques**

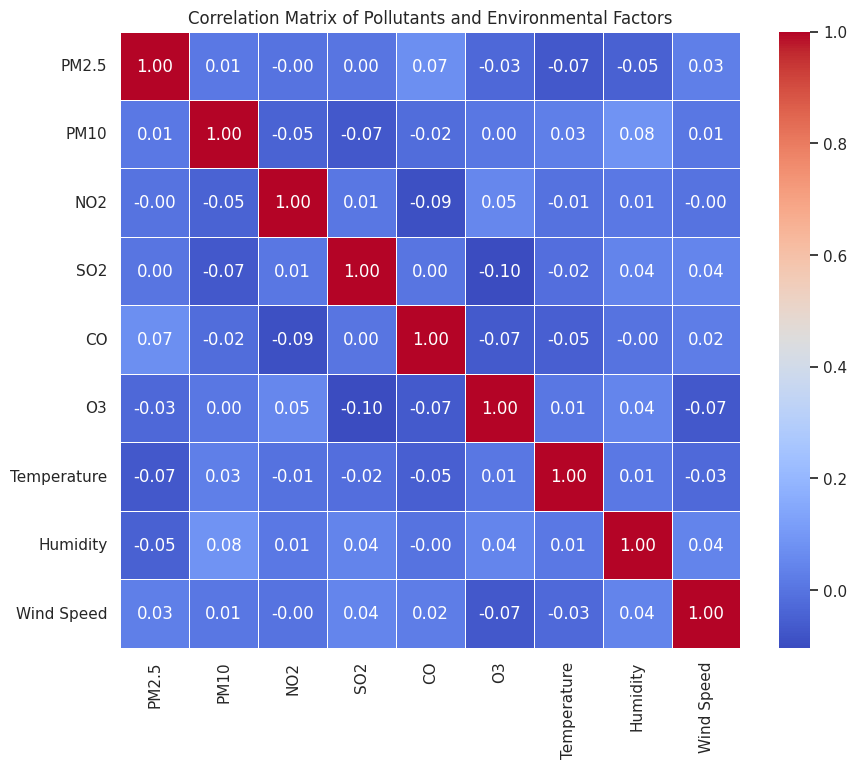
i) **Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is a crucial phase in the data analysis process, which aims to understand the structure, patterns, and key characteristics of the dataset. To find connections, spot inconsistencies, and spot patterns, it involves gathering and displaying the data. EDA ensures that the dataset is prepared for additional analysis and modeling tasks by assisting in the decision-making process around data pretreatment and model selection. **Histogram and Correlation matrix were utilized to visualize the distribution of pollutant levels and identify patterns in the data.**

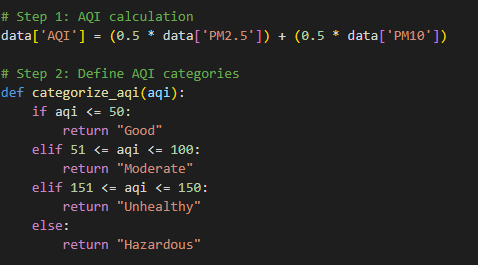
Histograms help identify skewness, variability, and the existence of outliers by clearly displaying the frequency of values within designated ranges. The frequency distribution of the main pollutants (PM2.5, PM10, NO2, SO2, CO, and O3) was revealed. The histograms, for most contaminants, showed skewed distributions, indicating that levels of the pollutants were dispersed unevenly, with sporadic severe values that might be pollution spikes. These representations gave an early comprehension of the variation in pollution levels, which also indicated the existence of possible outliers.



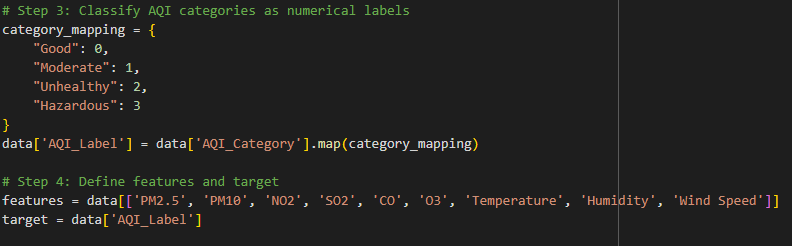
The correlation matrix helped to investigate the connections between environmental variables (temperature, humidity, wind speed) and pollutants (PM2.5, PM10, NO2, SO2, CO, and O3). An intuitive grasp of how these variables affect one another was provided by the heatmap visualization, which showed the direction and strength of linear correlations. For example, there was a strong positive connection between pollutants like PM2.5 and PM10, suggesting that they probably come from comparable sources or share contributing characteristics. Conversely, weaker associations with weather factors like humidity, temperature, and wind speed indicated more intricate or indirect effects.



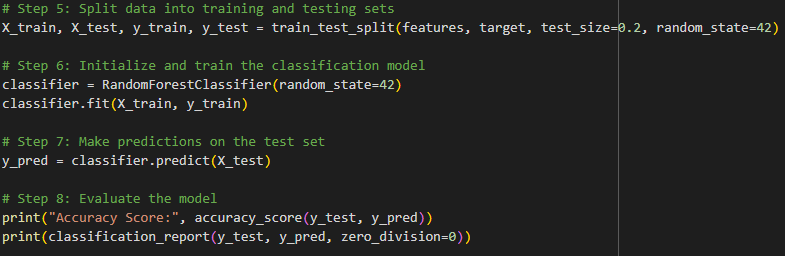
ii) **Classification analysis:**   
  
The classification analysis predicts AQI categories such as "Good," "Moderate," "Unhealthy," and "Hazardous" based on pollutant and environmental data. The AQI combines important pollutant levels into a single metric by calculating the weighted average of PM2.5 and PM10. Stakeholder interpretation is made simpler by a custom method that allocates AQI categories according to specified thresholds. To make sure these categories are compatible with machine learning techniques, they are subsequently transferred to numerical labels.

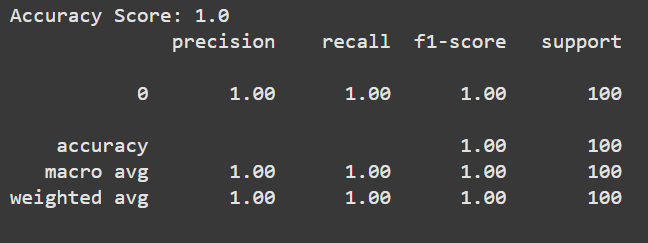


To capture both direct and indirect influences on air quality, features like pollution levels and environmental factors are included. A Random Forest Classifier is selected because of its resilience, capacity to manage a variety of data kinds, and efficiency in identifying intricate patterns through ensemble learning while lowering the possibility of overfitting. With an accuracy score of 1.0, the classification model accurately predicted the AQI categories on the test set. Its perfect ability to detect each category without false positives or negatives is shown in its precision, recall, and F1 score. Although this high accuracy may require confirmation to make sure it isn't the consequence of overfitting or an unbalanced dataset, the result indicates that the model successfully captured the correlations between characteristics and AQI labels.



Random Forest Classifier:





**iii) Regression Model**

**2.4.2 Rationale for Technique Selection (15%)** - Rozin

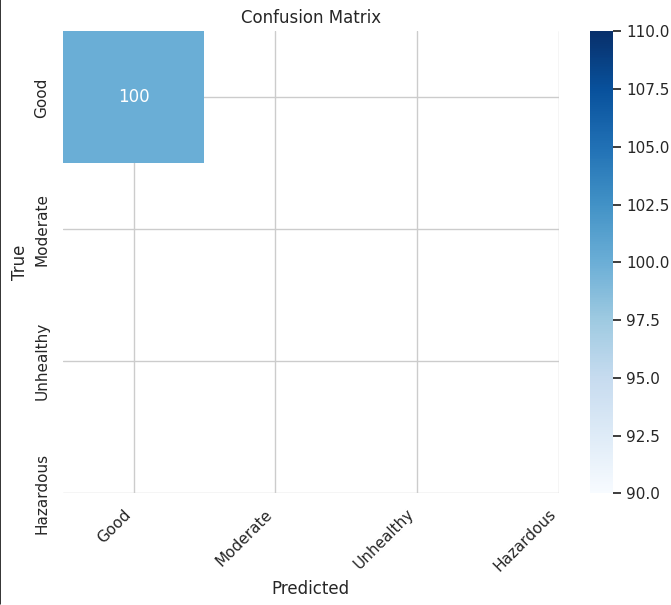
Provides a comprehensive rationale for technique selection

**2.5 Model Validation**

**Confusion Matrix:**

The confusion matrix is a performance evaluation tool used in classification tasks that provides a thorough analysis of the model's predictions across all potential categories. The link between actual AQI categories and those predicted by the model is shown graphically in this context. Every column shows the anticipated category, while every row displays the actual category. This enables us to evaluate the model's capacity to accurately categorize every group and spot incorrect classifications.

The code provided generates the confusion matrix after applying the categorize\_aqi function to transform the actual (y\_test) and predicted (y\_pred) AQI values into categorized labels. The mapping of continuous AQI values into meaningful categories, such as "Good," "Moderate," "Unhealthy," and "Hazardous," makes the judgment more understandable.

  
  
With no misclassifications or predictions for other categories, the output displays a confusion matrix with all 100 cases falling into the "Good" category. The model forecasts all test samples as "Good" with complete accuracy, which is consistent with the previous classification output (accuracy score of 1.0). This finding, however, probably points to a problem with the model's capacity for generalization, as it might have been trained on unbalanced data that was primarily classified as "Good." The absence of other categories ("Moderate", "Unhealthy", "Hazardous") in both true and forecasted values makes this clear.

To conclude, the confusion matrix, when combined with the model's output, functions as a diagnostic tool. It emphasizes how crucial it is to examine both the distribution of forecasts across all categories and overall accuracy in order to guarantee accurate and useful classification of air quality levels.

**2.6 Conclusion – Mridul**

Provides insightful observations and comprehensive suggestions