**Name: Samiksha Sandeep Zokande**

**2448354**

**4MDS – B**

**Exploratory Data Analysis**

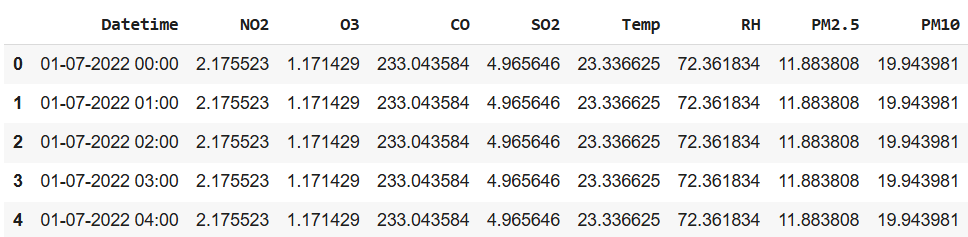
*Introduction:-*

Datasets are raw information which when processed provide insightful information. The trends, patterns, relations etc. in a data are not visible to common eye to overcome this Exploratory Data Analysis plays a huge part. EDA is used mainly by data scientists to build visuals of a dataset which summarize characteristics of a data.

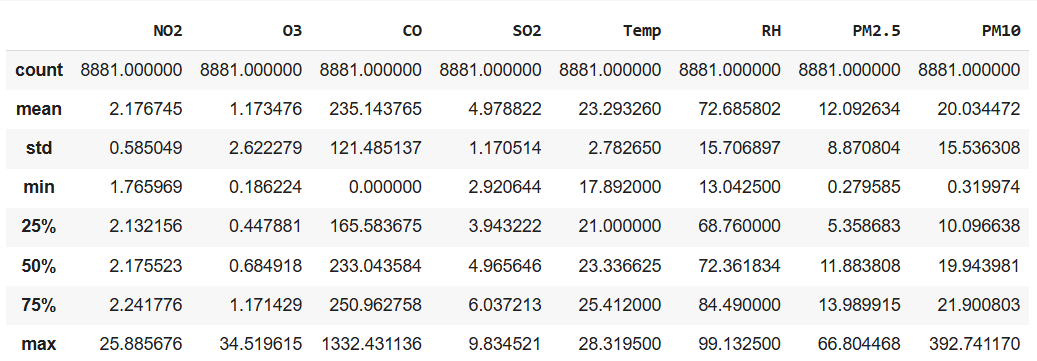
*Dataset:-*

The Central Pollution Control Board (CPCB) of India provides this data through its air quality monitoring stations across the country. The dataset contains measurements of nitrogen dioxide (NO2) and ozone (O3) and carbon monoxide (CO) and sulfur dioxide (SO2) and particulate matter with a diameter of 2.5 micrometers or less (PM2.5) and particulate matter with a diameter of 10 micrometers or less (PM10). The study receives its data from measurements taken every 15 minutes which generates detailed time series information. The dataset contains 8881 entries and **7 columns**, including six key air pollutants (PM2.5, PM10, NO₂, SO₂, O₃, CO) and timestamps across one year while the Excel file occupies 896 KB of storage space.

The dataset analyzes data from July 1 2022 through July 31 2023. The twelve-month period contains all four seasons which are important to analyze patterns in air quality and develop accurate predictions. The below tables give the overview and statistical information about the data. Below image shows a few entries of the dataset.



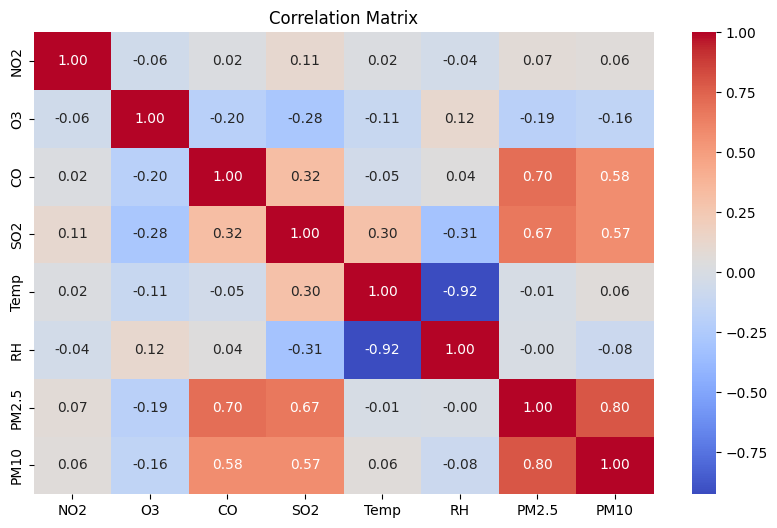
The statistical properties of the dataset are as follows:



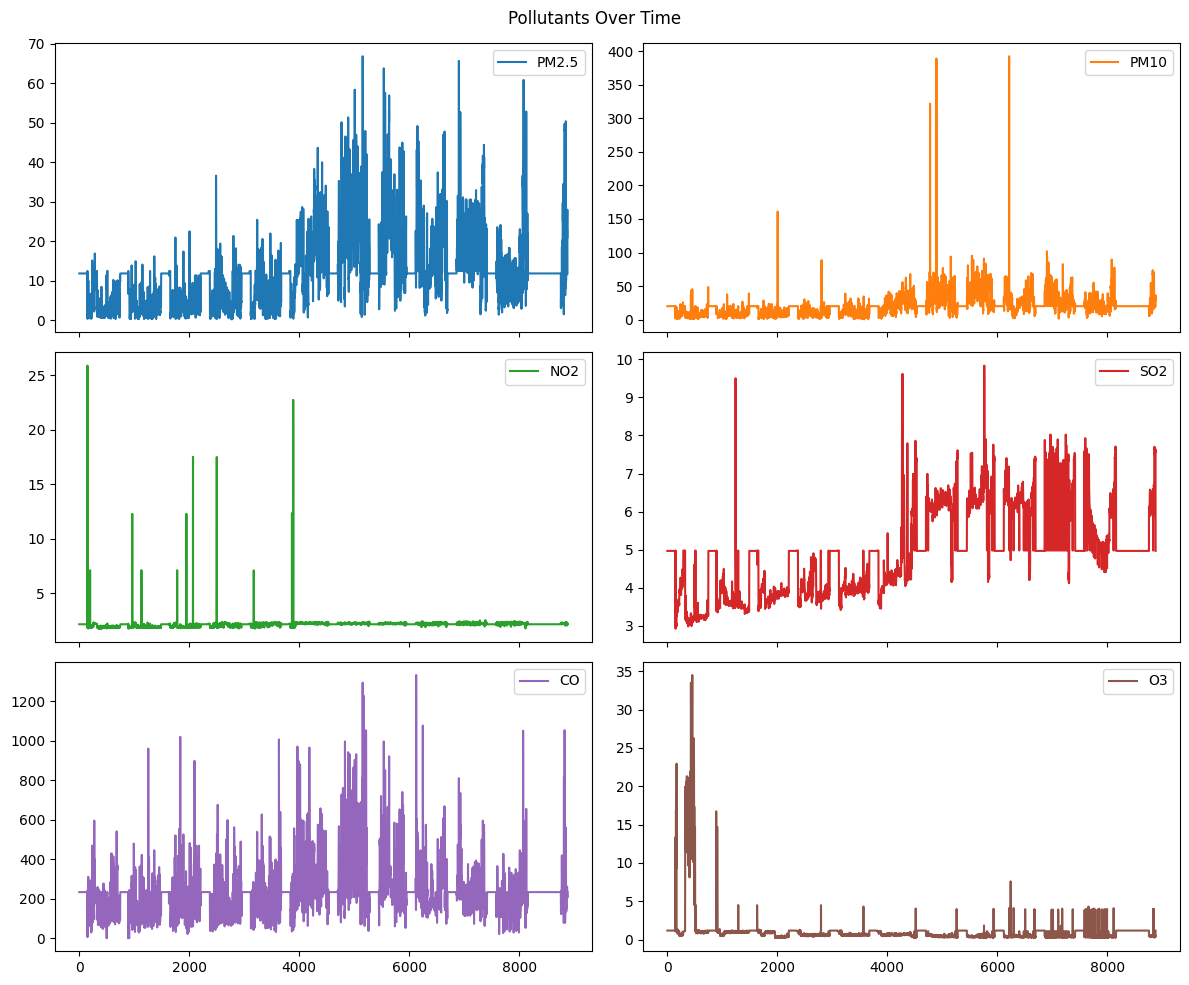
*Data pre-processing:-*

The first step involved importing the necessary Python libraries (e.g., Pandas, NumPy, Matplotlib) and loading the datasets. Dataset at first contained missing values due to sensor malfunctions or communication issues. The preprocessing stage used suitable imputation methods such as linear interpolation and forward/backward fill to fix missing values before analysis. The time series denoising process used Rolling Window Smoothing to maintain essential patterns while reducing noise. The Rolling Window method utilizes a moving average operation to minimize short-term oscillations. Outlier detection stands as a vital step to achieve strong time series modelling results. The Interquartile Range (IQR) method was employed to identify and handle outliers. The method of Winsorization was selected instead of removing or marking outliers as missing data because it prevented distortions to the time series by restricting extreme values to the IQR bounds. The method protects the time-based sequence while reducing the effects of extreme values.

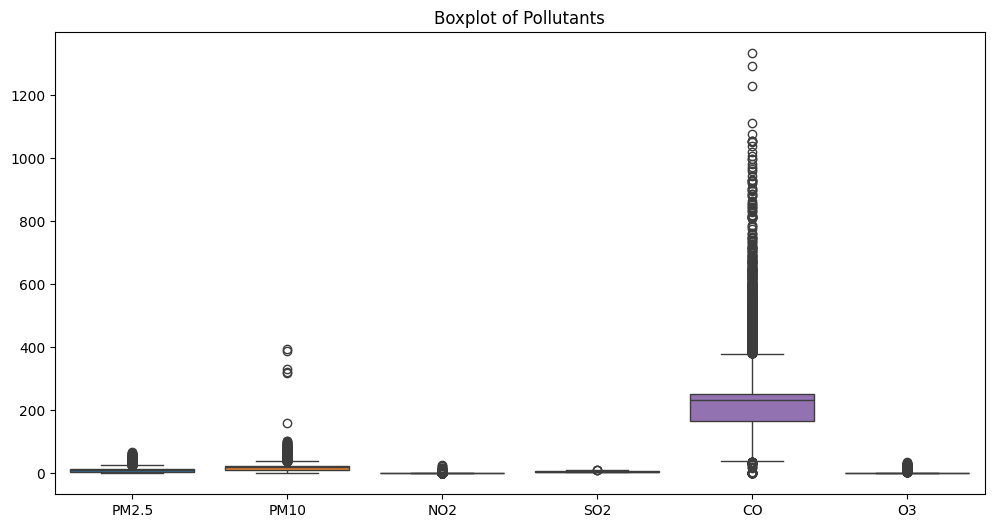
*Graphs and Interpretations:-*



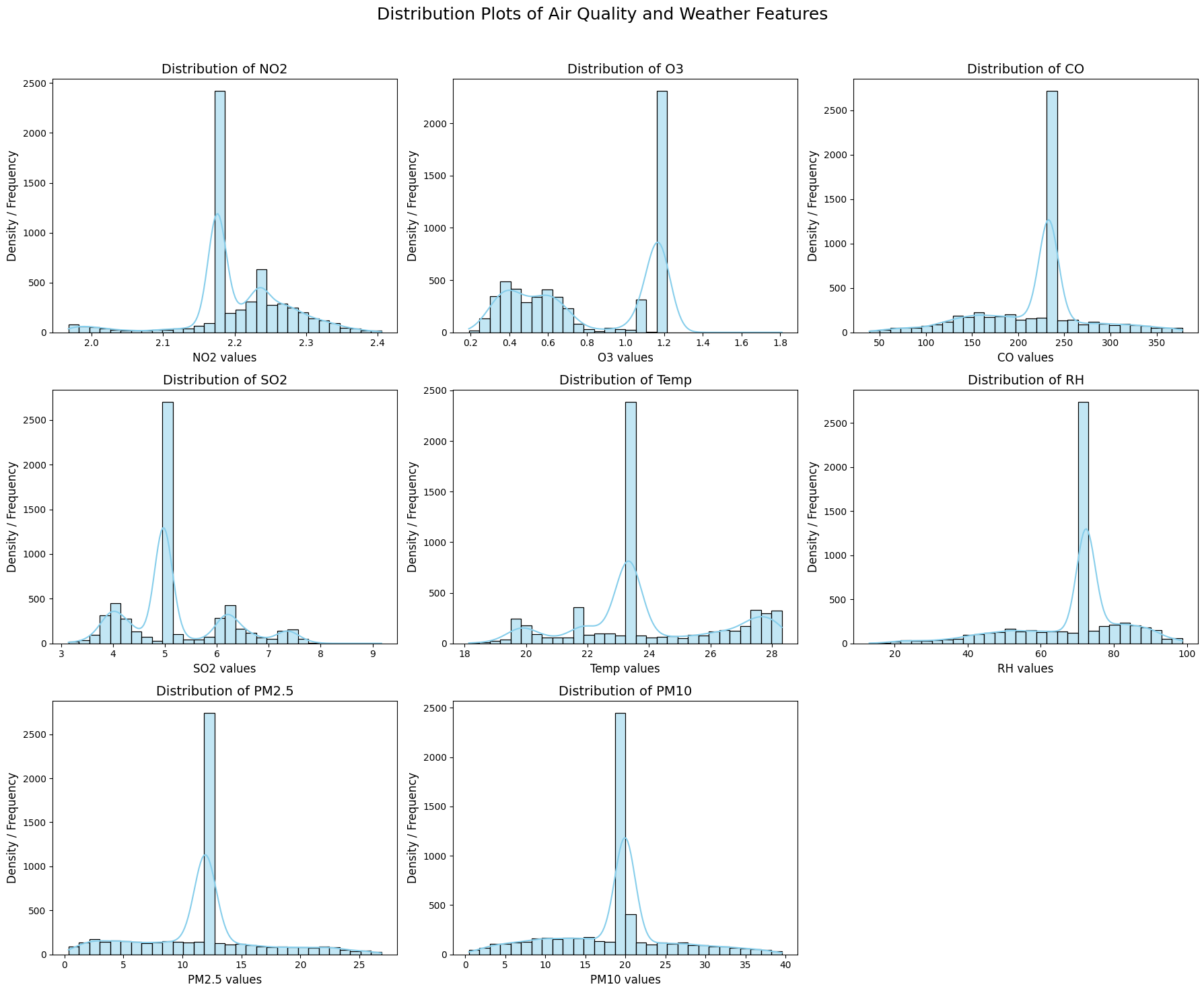
Correlation matrix is used to show any possible relationship between two features. Here it can be clearly seen that O3 and NO2 show very low correlation with all of the other pollutants. In further work these pollutants will not be considered as they may contribute very less to multivariate analysis. CO, SO2, PM2.5, PM10 are the pollutants which are mutually highly correlated with each other. These are the kind of insights given by a correlation matrix.



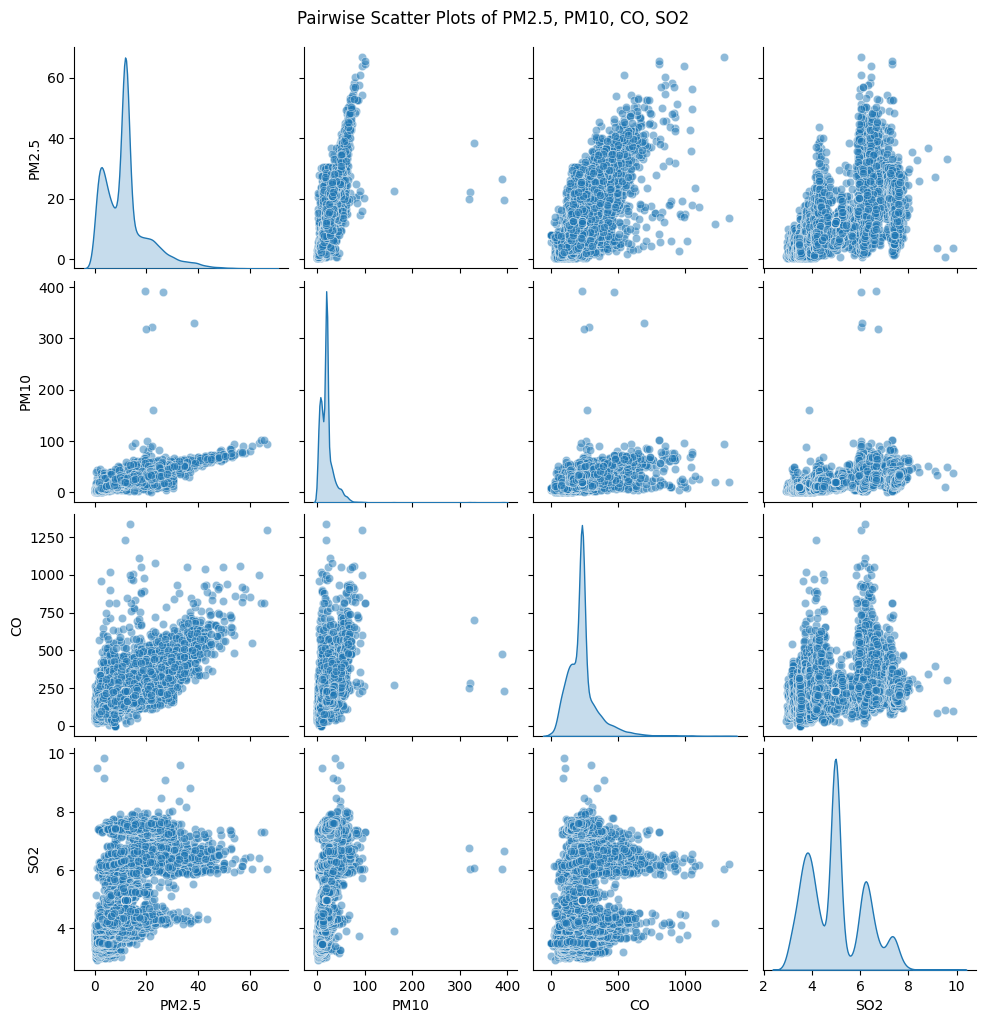
The graph above shows the pollutant measure over a period of time. Multiple sharp spikes can be seen in PM2.5 with an upward trend indicating a rise in pollution. This gradually at the end drops down. For PM10 trend is not observed but a sudden spike is seen exceeding 350 µg/m³. CO shows significant volatility the middle portion has dense peaks likely due to traffic. This above graph helps in understanding the pattern of pollutants. This will ultimately help in finding a solution.



The graph shown here is a box plot. Box plots are used to represent outliers. The middle portion is called a box with 3 quartiles first quartile(Q1), second quartile or median (Q2) and third quartile (Q3). The lines extending are called the whiskers . All pollutants show certain degree of outliers especially CO and PM10. As explained in the pre-processing the outliers were handled using Winsorization.



Distribution graph as shown above is used to understand the distribution of the features. These values help understand how values are spread, whether normally distributed, skewed etc. Each subplot visualizes the distribution (histogram + KDE). CO indicates heavy central peak around 250. And extremely narrow spike indicating many identical values. For PM2.5 the graph is slight right skew indicates occasional high values (pollution spikes) and spike around 12 µg/m³. Overall, values are low/moderate which indicates decent air quality for most hours. Similar to PM2.5, PM10 has major peak around 15µg/m³.



Taking into consideration the correlation graph scatter plot has been plotted with correlated features. This helps understand how one feature rises in accordance with the other. PM2.5 shows an increase in levels in relation with PM10 and CO. This shows a clear positive trend. While PM10 has shown a rise in accordance with PM2.5. Similarly, CO also has shown a upward trend with respect to PM2.5