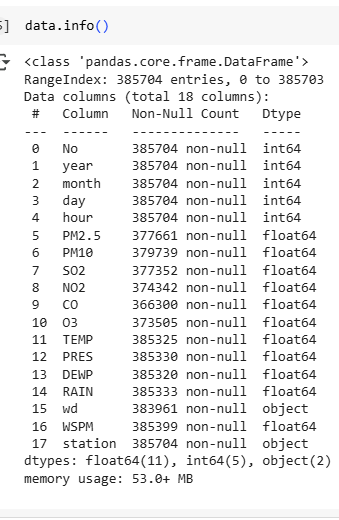
# Exploratory Data Analysis (EDA)

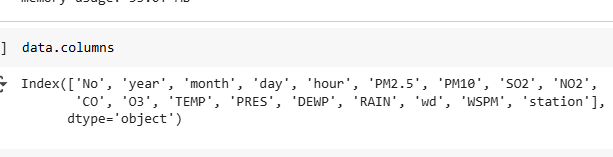
## Fundamental data understanding



**Figure 2: Information about the data**

(Source: Google colab)

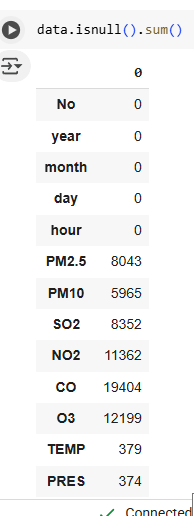
This dataset includes information on pollutants, weather parameters, and station characteristics with 385,704 entries and 18 columns. It has missing values for some columns, such as PM2.5, CO, and NO2. Data types are of three types, namely integer, floating point, and categorical (or object) type.



**Figure 3: Columns of the data**

(Source: Google colab)

The dataset includes 18 columns capturing various features such as date and time, air pollutants (e.g., PM2.5, SO2, NO2), meteorological data (e.g., TEMP, PRES, WSPM), wind direction (wd), and station details for each recorded observation.



**Figure 4: Null values of the dataset**

(Source: Google colab)

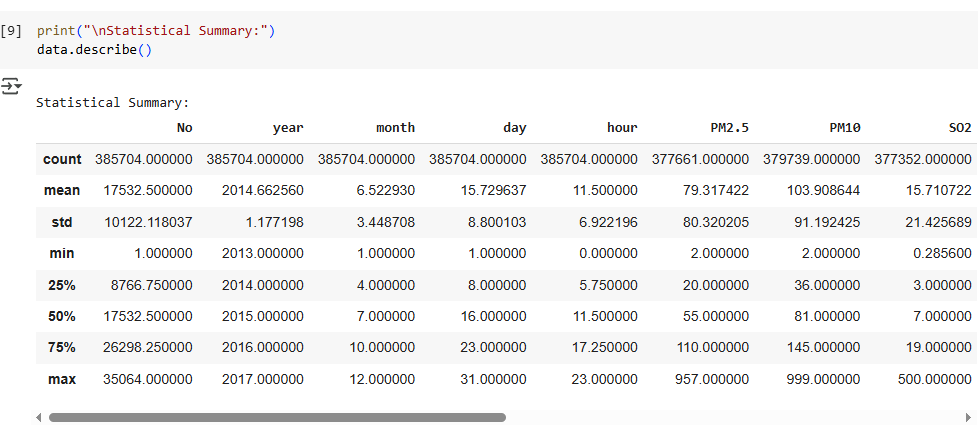
The dataset contains several missing values, notably in key pollutant columns: PM2.5 (8,043), CO (19,404), and O3 (12,199). Weather-related fields like TEMP, DEWP, and wind direction (wd) also have minor gaps that require preprocessing.



**Figure 5: Data types**

(Source: Google colab)

The dataset includes 18 columns with various data types: 5 integer fields for temporal values, 11 float fields representing pollutant levels and weather conditions, and 2 object types for wind direction (wd) and station names, requiring appropriate encoding.

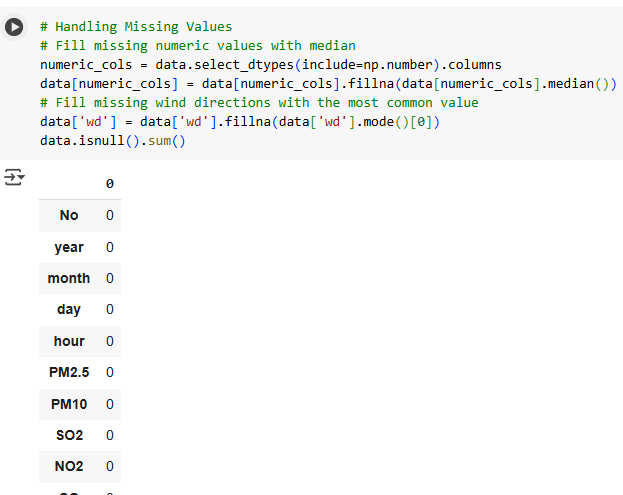


**Figure 6: Statistical summary**

(Source: Google colab)

The statistical summary reveals the central tendencies and spread of each numeric variable. PM2.5 values range from 2 to 957 µg/m³, while temperature spans from -19.9°C to 41.6°C. Wind speed varies up to 12.9 m/s, indicating significant meteorological variation.

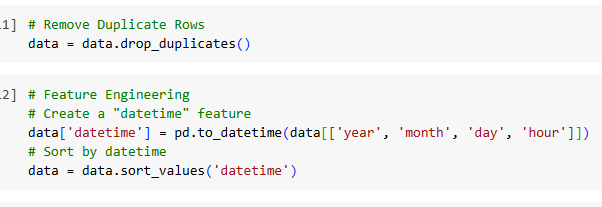
## Data preprocessing



**Figure 7: Handling missing values**

(Source: Google colab)

Missing values in the dataset were addressed using appropriate strategies. Numeric columns were filled with their respective median values to minimize distortion. For the categorical 'wd' (wind direction) column, the most frequent category was used. This ensured no missing values remained in the dataset.



**Figure 8: Removing duplicates and feature engineering**

(Source: Google colab)

Duplicate records were removed to maintain data quality. A new “datetime” feature was created by combining the year, month, day, and hour columns, enabling time-based analysis. The dataset was then sorted by this datetime column to ensure chronological order for further time series processing.