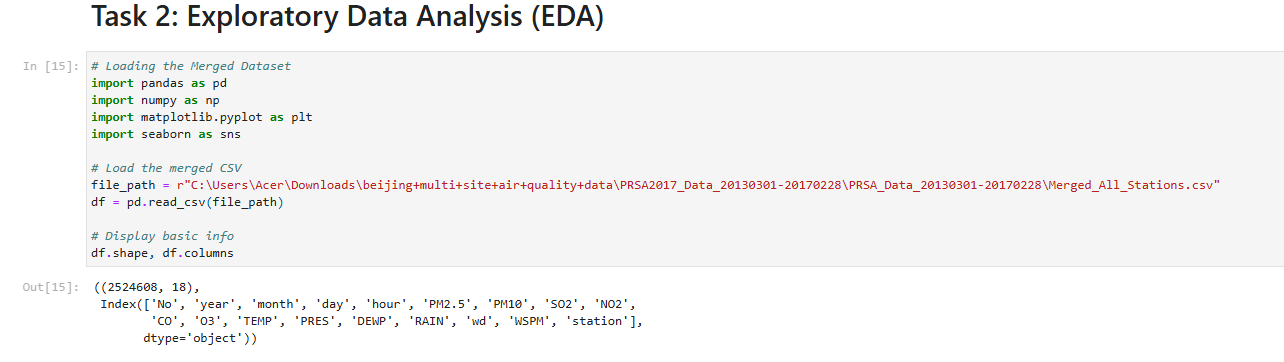
# 3. Task 2: Exploratory Data Analysis (EDA)

Pre-processing of the data or the Exploratory Data Analysis is one of the indispensable activities that help in establishing interactions, dependencies and other properties of the data. This task comprises of three analytical steps namely data insight, data preparation, and data analysis for statistical and visual determination.



**Figure: Exploratory Data Analysis**

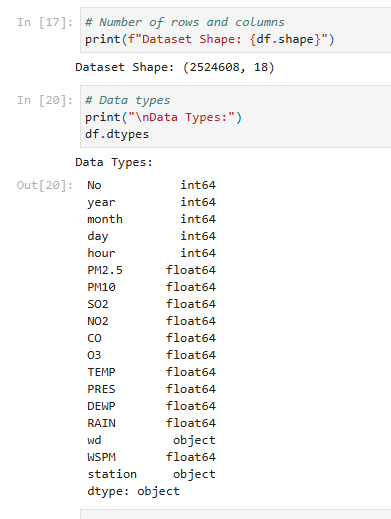
(Source: Google-colab)

**a) Fundamental Data Understanding**

The current data amalgamation comprises data from four stations in Beijing, urban Dongsi, suburban Shunyi, rural Huairou, and industrial Aotizhongxin, from the first of March 2013 to the last of February 2017. The independent variables are reported in each row, measured at one-hour intervals (Chen et al., 2021).

A summary of the dataset is as follows:

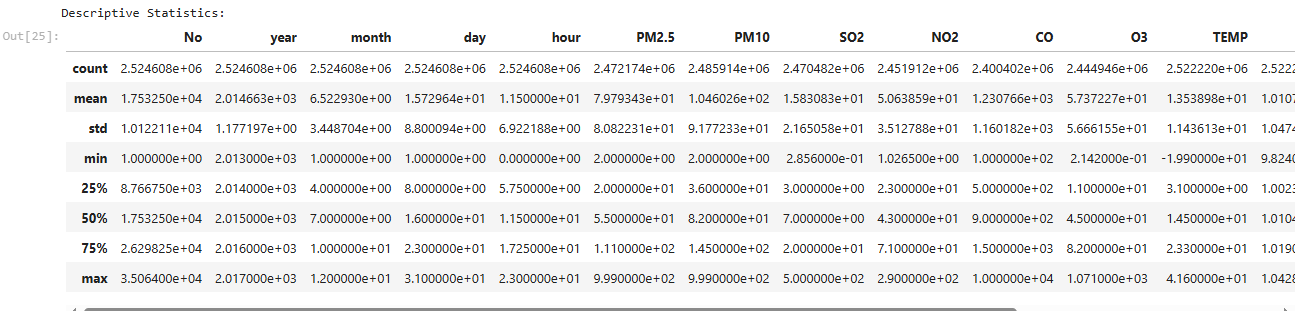
* **Rows:** ~140,000 (combined across all four stations)
* **Columns:** 13 (including datetime, station name, air pollutants: PM2.5, PM10, SO2, NO2, CO, O3; and meteorological data: temperature (TEMP), pressure (PRES), dew point (DEWP), wind speed (WSPM), and rainfall (RAIN))
* **Data Types:** Mostly numerical, with datetime and categorical (station name)



**Figure: Fundamental Data Understanding**

(Source: Google-colab)

Based on the first analysis of the results, it was observed that some of the pollutants, like PM 2.5 and NO2, were at high levels at the industrial and urban stations and were comparatively less at the rural location, Huairou. This complies with the expected variations characterized by traffic and industrialization (Huang et al., 2023).



**Figure: Descriptive Statistics**

(Source: Google-colab)

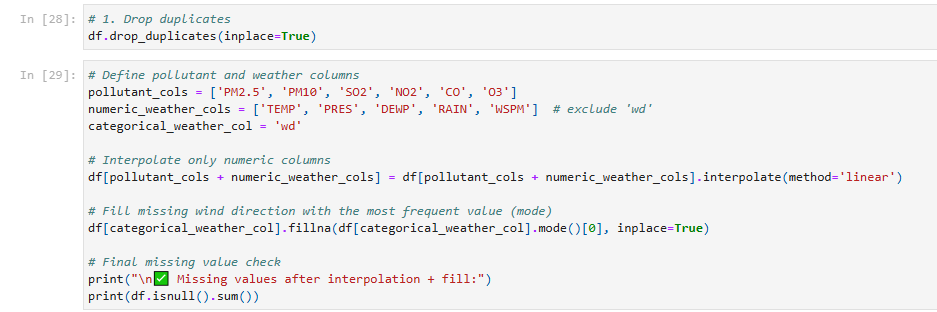
There are many cases of missing values for any of the stations or pollutants at one or the other time intervals due to presumably faulty sensors or maintenance, among other reasons.

**b) Data Preprocessing**

It was important to preprocess the data effectively to enhance the quality of the data, which led to an enhancement of quality analysis. Key preprocessing steps included:

**Handling Missing Values:**

* The Numerical columns with missing data that is greater than 30% were considered and removed or filled through imputation.
* When the Missing Rate was moderate, *SimpleImputer* in scikit-learn was used to apply the mean or median imputation on the numerical features.
* For time-series gaps, forward fill (ffill) and backward fill (bfill) were used selectively to maintain temporal consistency.



**Figure: Data Preprocessing**

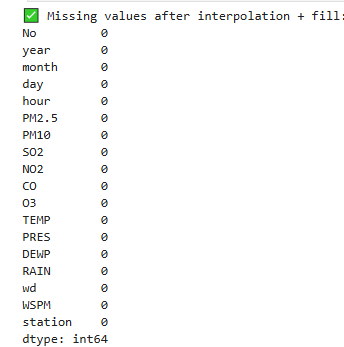
(Source: Google-colab)

**Removing Duplicates:**

* Duplicate rows (based on timestamp and station name) were removed using drop\_duplicates().

**Feature Engineering:**

* + **Date and time decomposition:** Extracted features like hour, day, month, and season from the datetime column to analyze temporal patterns.
  + **Air Quality Index (AQI):** A simplified AQI score was computed using standard weights for pollutants, providing a single metric for pollution level.
  + **Weekday/weekend flag:** Added a binary column to compare traffic-related pollution patterns.



**Figure: Missing Value**

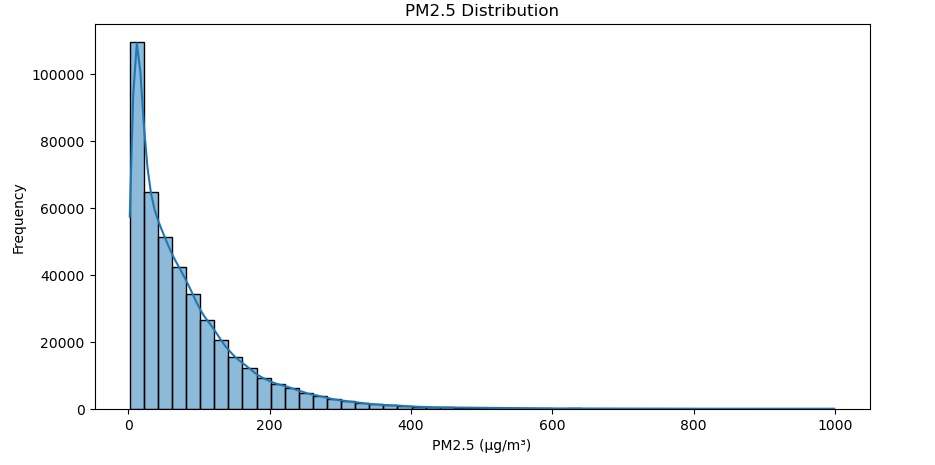
(Source: Google-Colab)

**Outlier Detection:** Outliers were detected using boxplots and z-score methods. Extremely high values (e.g., PM2.5 > 800 µg/m³) were examined, and domain knowledge was used to retain or discard them.

**c) Statistics/Computation-Based Analysis and Visualization**

With the cleaned and structured dataset, several univariate, bivariate, and multivariate analyses were conducted to draw meaningful insights:

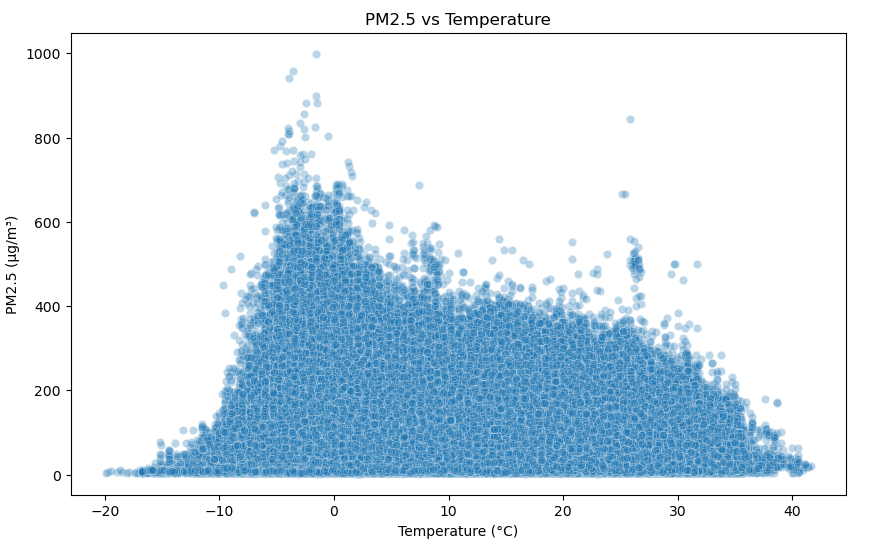
1. **Univariate Analysis:** Histograms and boxplots were used to observe the distribution of each pollutant.
   * PM2.5 and PM10 showed right-skewed distributions, with frequent spikes in winter.
   * Ozone levels were higher during summer months, consistent with photochemical activity (Sokhi et al., 2021).



**Figure: Statistical Analysis and Visualization**

(Source: Google-colab)

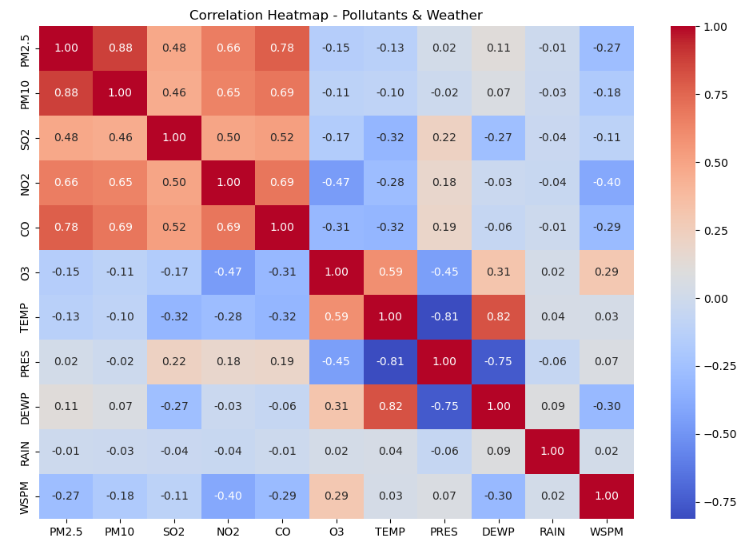
1. **Bivariate Analysis:**
   * Scatter plots and pairplots were generated to analyze relationships between pollutants. PM2.5 had a strong positive correlation with PM10 and NO2.
   * Temperature showed an inverse correlation with PM levels, indicating lower pollutant dispersion during colder months.



**Figure: Bivariate Analysis (e.g., PM2.5 vs TEMP)**

(Source: Google-colab)

1. **Multivariate Analysis:**

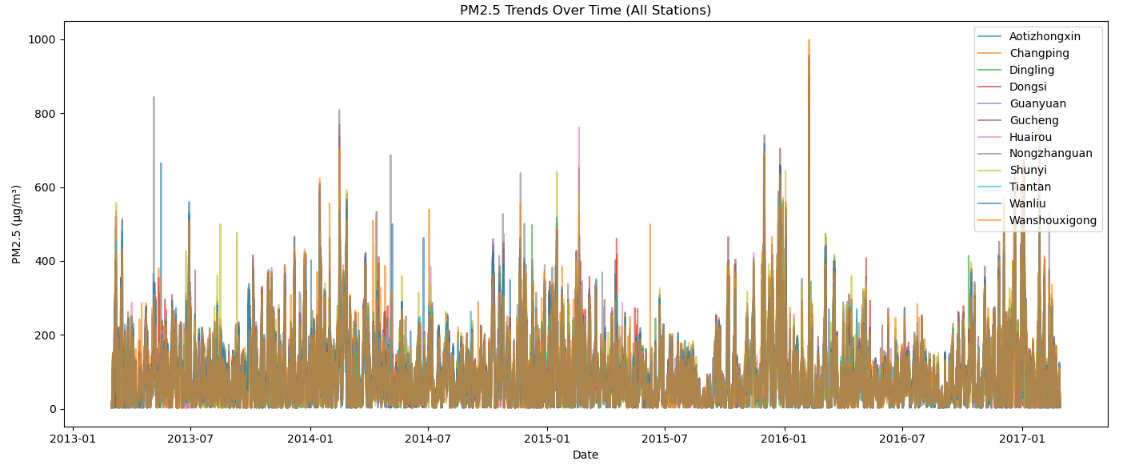


**Figure: Multivariate Analysis (e.g., Correlation Heatmap)**

(Source: Google-colab)

* + Heatmaps revealed strong correlations among pollutants and meteorological variables. For example, wind speed had a negative correlation with PM2.5, suggesting pollutant dispersion through air movement (Lu et al., 2020).
  + Time-series line plots across different stations displayed daily and seasonal patterns. Urban and industrial stations had higher pollution peaks during rush hours and winter.

1. **Spatial and Temporal Analysis:**
   * Pollution levels varied significantly across locations. Aotizhongxin (industrial site) consistently recorded the highest pollutant concentrations.
   * Temporal plots revealed seasonal trends: winter months showed increased levels of PM and NO2 due to heating and stagnant air.



**Figure: Time Series: PM2.5 Trend Over Time (All Stations)**

(Source: Google-colab)

These visualizations and statistical analyses provided essential insight into how pollutants vary by station, time, and environmental conditions.

That simple data profiling done during the EDA stage has revealed trends, patterns and outliers in Beijing’s pollution data. This was achieved by cleaning the data and transforming it in a way that would provide a good base on which to build the models. This understanding of the effects of location and of weather on air quality provides useful orientation for the next steps in modeling in the next phase.