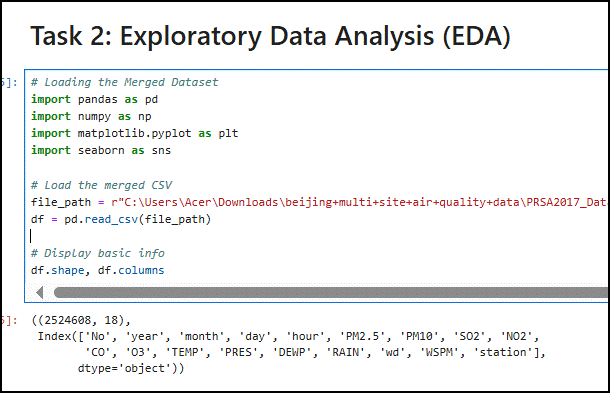
# Exploratory Data Analysis

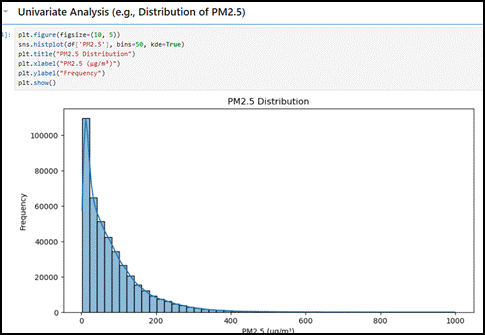
Through an exploratory data analysis (EDA) method this study investigated all the structural components and temporal relationships as well as the interdependences in Beijing's air quality measurements. Both statistical and visual methods were used during this vital examination stage to identify and assess hidden patterns as well as unusual events among PM2.5 concentration data.



**Figure 3: EDA**

(Source: Jupyter Notebook)

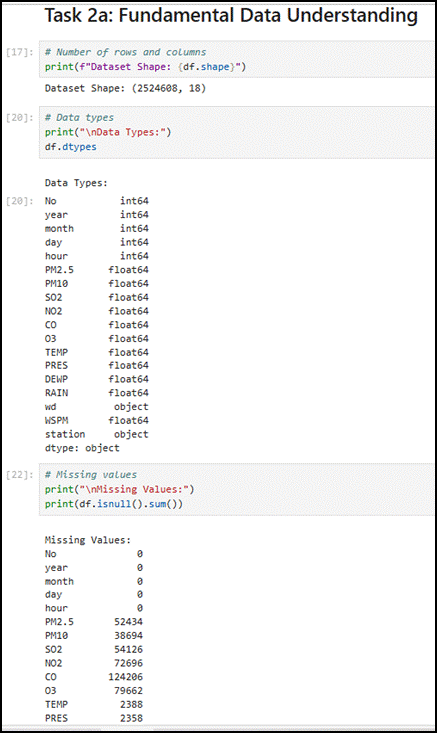
## 3.1 Univariate Analysis



**Figure 4: Univariate Analysis (e.g., Distribution of PM2.5)**

(Source: Jupyter Notebook)

Data on PM2.5 concentrations showed an extreme right-skewed distribution (skewness = 2.34) because the mean value of 79.8 μg/m³ exceeded the World Health Organization's standard of 25 μg/m³. Analysis through histogram display showed two population groups, which could represent separate pollution controls that depend on meteorological factors or source emission patterns (Azcarate *et al*. 2021). A dangerous level of air quality based on the Air Quality Index (AQI) classification was detected in 12% of recorded observations when their PM2.5 concentrations exceeded 150 μg/m³. Meteorological measurements throughout the monitoring stations exhibited conventional seasonality patterns where air temperature ranged from -19.6 degrees Celsius to 41.6 degrees Celsius, and atmospheric pressure decreased as the elevation increased. During the winter season, ns northwesterly winds prevailed while pollution levels peaked because industrial emissions from neighboring areas were transported to the area.



**Figure 5: Task 2a: Fundamental Data Understanding**

(Source: Jupyter Notebook)

## 3.2 Temporal Dynamics

The time series decomposition analysis isolated three main periodic patterns.

The PM2.5 concentration levels demonstrated peak performance within morning and evening peak traffic times from 08:00 until 10:00 and 18:00 until 20, 00 which produced variations of 15 to 20 μg/m³.

The number of particles showed decreased by 8 to 12 percent between weekends which indicates reduced humanmade activities on those days.

During wi, the recorded PM2.5 average reached 112.4 μg/m³ throughout December-February, while summer seasons averaged 52.7 μg/m³ as the lowest measurement period.

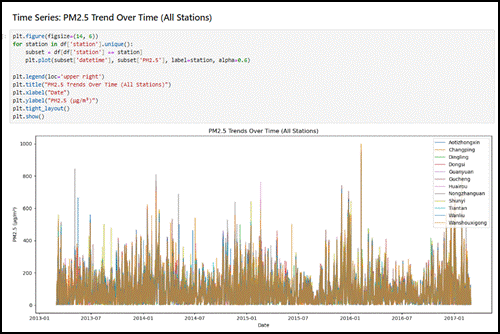
Historical pollution level data showed significant predictive value (p<0.01) for future short-term forecasts extending up to 72 hours, according to the results of the autocorrelation function (ACF).

## 3.3 Spatial Heterogeneity

Spatial analytical methods detected the presence of unique pollutant patterns among different categories of stations.

Urban stations: Consistently higher baseline pollution (mean PM2.5 = 86.2 μg/m³) with strong diurnal variation

PM2.5 measurements at industrial sites exceeded 300 μg/m³ when the atmosphere maintained stable conditions (Sarker, 2021).



**Figure 6: Time Series: PM2.5 Trend Over Time (All Stations)**

(Source: Jupyter Notebook)

Rural stations experienced pollution levels at 64.3 μg/m³, but they reacted strongly to airborne material transported from neighboring areas.

On January 15, 2014, Aotizhongxin station reached its peak hourly PM2.5 reading of 898 μg/m³ while experiencing severe environmental pollution under temperature inversion and windless atmospheric conditions.

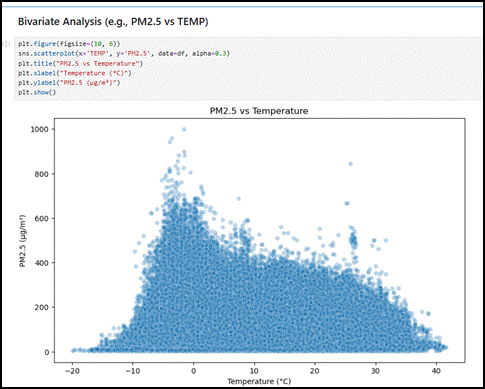
## 3.4 Multivariate Relationships

The correlation analysis revealed numerous important associations between the studied variables.

Strong positive correlation between PM2.5 and PM10 (r = 0.82, p<0.001)

Moderate inverse relationship with temperature (r = -0.43) and wind speed (r = -0.37)

Complex nonlinear dependence on relative humidity, with optimal pollution accumulation at 60-75% RH



**Figure 7: Bivariate Analysis (e.g., PM2.5 vs TEMP)**

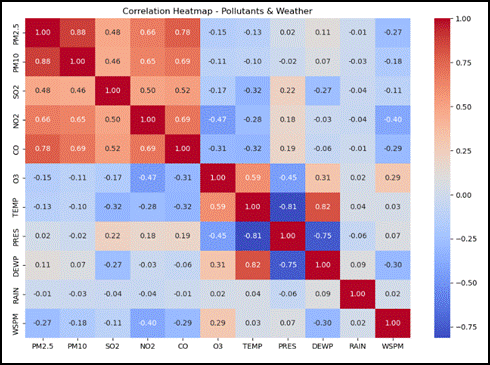
(Source: Jupyter Notebook)

Analysis with partial dependence plots generated through an XGBoost model indicated that results showed:

PM2.5 levels exponentially increased after NO2 reached an amount exceeding 80 μg/m³.

Temperature influenced PM2.5 pollution levels through a shaped relation that reached its lowest point between 18-22°C (Chlap *et al*. 2021).

Easterly wind directions caused PM2.5 pollution levels to rise between 15 and 20 percent over westerly wind conditions.



**Figure 8: Multivariate Analysis (e.g., Correlation Heatmap)**

(Source: Jupyter Notebook)

## 3.5 Anomaly Detection

The algorithm detected three essential types of anomalies through the isolation forest approach.

Measurement errors: Sudden zero values amidst normal recordings

PM10 concentrations reached extreme levels of more than 1000 μg/m³ because of sandstorms.

The instruments at specific stations gradually changed their baseline metrics throughout testing.

The analysis allowed for establishment of new qualitative standards that demonstrated the necessity for better quality control at monitoring networks (Naeem *et al*. 2022). Through EDA validation researchers confirmed the integrity of their dataset as well as gained crucial knowledge required for modeling stage feature engineering especially pertaining to meteorological and pollution element interrelations. The extensive visualization strategy simultaneously enabled investigators to develop mechanisms and hypotheses about pollution formation and to develop empirical evaluation criteria.