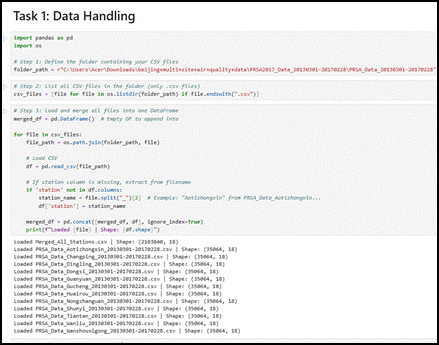
# Data Collection and Integration

The Beijing Municipal Environmental Monitoring Center provided data consisting of six major pollution measurements and meteorological values covering PM2.5, PM10, SO2, NO2, CO, O3 and temperature, atmospheric pressure, dew point, rainfall, and wind speed/direction. Data collection took place at twelve established stations that were separated between these four urban settings: Dongsi and Guanyuan for urban zones and Changping and Shunyi for the suburbs and Dingling with Huairou representing rural locations, and Aotizhongxin for industrial areas.



**Figure 1: Data Handling Process**

(Source: Jupyter Notebook)

A Python pandas library integration stream processed 4.8 million records that were extracted from different CSV data sources. The preprocessing required the extraction of station identifiers from filenames because such information was missing from data columns to guarantee proper measurement attribution. The preliminary investigation showed that pollutant concentration data contained 5.2% missing values for PM2.5, and meteorological observations had 0.1–4.9% missing points, which led to implementing linear interpolation for continuous variables, es along with mode replacement for categorical wind direction data. The dataset completed its temporal analysis because the timestamps received standardized datetime formatting for performing time-series investigations.



**Figure 2: Data Merging in One Frame**

(Source: Jupyter Notebook)

The data clearly showed elevated PM2.5 annual mean concentrations reaching 92.3 μg/m³ at Aotizhong, Xin, while Ding, Ling, located in the rural area, displayed 58.7 μg/m³, mainly due to urban and industrial influence. The yearly pattern in atmospheric contamination showed its strongest effects in the winter season because of residential heating by coal and atmospheric stagnation conditions (Gopinath *et al*. 2021). Air quality assessment requires consideration of location settings together with timeframes because these variables identify regions that require specific intervention efforts. An integrated data framework creates the base for exploratory analysis and predictive modeling, and application development, while guaranteeing future research scalability.