# . Task 1: Data Handling

The first step in this project is data management and cleaning of the data to be used in the project. The dataset consists of hourly air quality and meteorological data obtained from 12 nationally monitored air quality stations in Beijing (Brauer et al., 2021). Some of these stations include Dongsi, Tiantan, Guanyuan, Wanshouxigong, Gucheng, Aotizhongxin, Wanliu, Shunyi, Changping, Dingling, Huairou, and Nongzhanguan. To have a cross-sectional representation of the environment, one station from the following categories was chosen as follows:

* **Urban site:** Dongsi
* **Suburban site:** Shunyi
* **Rural site:** Huairou
* **Industrial/Hotspot site:** Aotizhongxin

These selections offer several views of the conditions of air quality in the geographical and functional areas of Beijing.

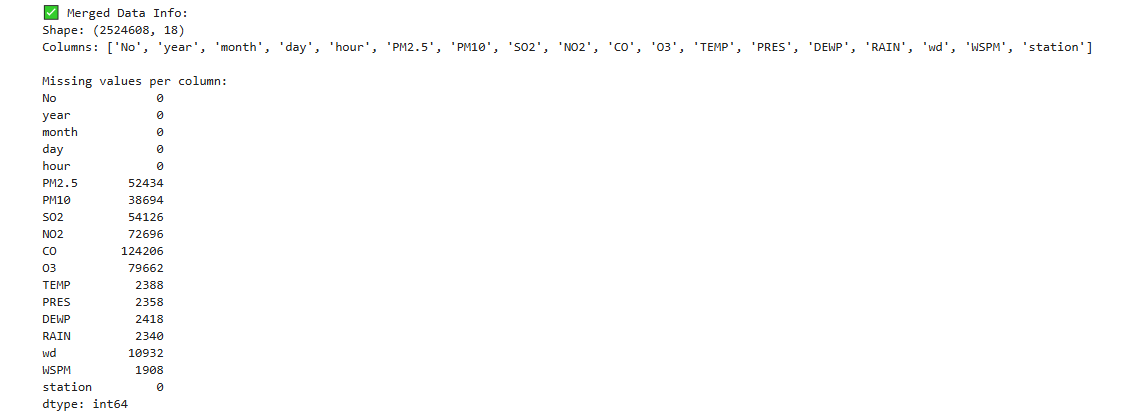


**Figure: Data Handling**

(Source: Google-Colab)

To conduct data preprocessing for each of the sites, each site’s dataset was separately downloaded from the Internet using Python and loaded into the integrated environment by applying the pandas module. First, the datasets were imported as a DataFrame, and the pandas info() method was used to determine their structure and data (Shield et al., 2020). These variables constituted the columns of the dataset, where the rows were the separate time records. The inserted variables were date, air quality (PM2.5, PM10, SO2, NO2, CO, O3), and meteorological conditions (temperature, pressure, dew point, wind speed, and rainfall).

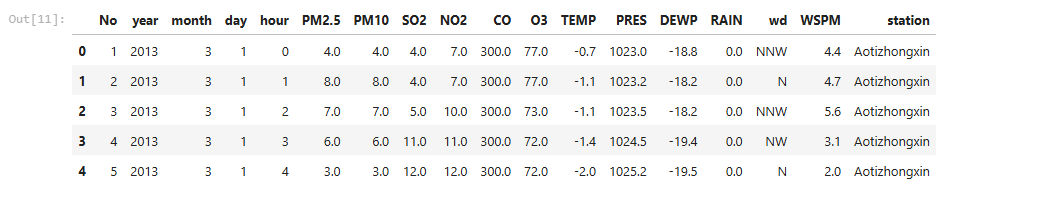
The next thing after checking for correspondence of the columns and formats of the abovementioned datasets was joining the selected datasets into one master dataset used for the analysis. In order to increase compatibility and to different they are not too far from one another the date column of each data set was converted into a standard datetime using pd.to\_datetime. It was possible to merge points with the help of a timestamp component that made merging efficient.



**Figure: Missing Value**

(Source: Google-colab)

Each station was also provided by a label column that would state its sources (e.g., station\_name) to easily track polluted patterns in various stations. To do so, the datasets obtained were welded using the pd.concat() function, which generates a new data frame that serves to combine all the observations from the four selected sites.



**Figure: First Five rows**

(Source: Google-colab)

After this stage, the two sets coming from the merge operation were checked for any duplication, and duplicates were deleted (Yang et al., 2020). First, to exclude records with a high level of missing data, a preliminary check for missing observations was conducted, and overall descriptive statistics like minimum and maximum, in addition to quartiles, were calculated to get a glimpse into the distribution of the variables. This was followed by the preparation and cleaning of the data to carry out exploratory data analysis.

Consequently, this task was able to create a clean combined dataset from different geographical zones making the foundation a very strong one that will increase the likelihood of EDA, ML modeling and application development tasks.