# Dataset Information

The information used in the context of this work is the hourly measurements of air quality, with a particular focus on the PM2.5 rate. This involve such factors as temperature (TEMP), pressure (PRES), relative humidity (DEWP), and wind speed (WSPM), and other pollutants such as PM 10, SO2, NO2, CO and O3. The other variable is for wind Direction (Wd) and datetime including year, month, day and hour. Handling of missing data were done using the ‘mode’ on the nominal ‘Wind Direction’ and the ‘median’ for the numerical values. It was then sorted by datetime to have the data presented in a chronological manner in this format. This dataset can be used effectively for both exploratory analysis as well as the prediction analysis for PM2.5 concentration along with other variables of the surroundings.

# Data Handling

The data comprises 11 files, the observation of air quality at different stations in China, for the period of March 1 2013, to February 28, 2017. All of these files contain hourly readings for PM2.5, PM10, SO2, NO2, CO, and O3 and environmental parameters of temperature, pressure, dew point, wind direction and wind speed, respectively. These datasets offer an extensive insight into the air quality in its present trends in pollution levels as well as its relation to various weather parameters.



**Figure 1: Merging the datasets**

(Source: Google colab)

The code starts with the fact that there must be a folder with the CSV files and uses glob to get all the paths to the CSV files. It then utilises coded statements that read each CSV file into a DataFrame and save this in an array. Subsequently, all of the DataFrames are joined into one DataFrame using pd.concat. The shape of the resulting dataset is illustrated, and some of the first rows of the data are shown by using data.head(). This helped to combine the data from the many single files into a large and more manageable file to analyse.