# Homework 1

- Use Python/Pandas/Matplotlib to analyze the data **"airBoxData.csv"**. The file contains air quality records from airBox devices in Taiwan and the time span is from July 18, 2020 to July 24, 2020. Each row is the air quality record of one hour from one airBox device. The following is the meaning of each column. Note that this dataset has not gone through the data cleaning process and you may find some unreasonable data values. You just assume all record is correct to answer questions in this homework.

| **Name of columns** | **Meaning** |
|---|---|
| siteID | airBox device ID |
| year | year of the record |
| month | month of the record (1: January, 2: February, …11: November, 12: December) |
| day | day of the record |
| hour | hour of the record (24 hour clock) |
| weekday | Weekday of the record (1: Monday, 2 Tuesday, …….6: Saturday, 7, Sunday) |
| gps_lat | Latitude of the device location |
| gps_lon | Longitude of the device location |
| PM10 | Measure of average PM10 level of one hour |
| Tmp | Measure of average temperature level of one hour |
| PM2.5 | Measure of average PM2.5 level of one hour |
| PM1.0 | Measure of average PM1.0 level of one hour |
| RH | Measure of average relative humility level of one hour |
| area | city |

- What we teach in class may not be sufficient to answer all the questions. You should look up documents online to learn the instructions you may need.

- Submit an ipython notebook file that answers the following questions with python commands. In the ipython notebook file that you submit, also include the original question and its answer (if applicable) as a comment before the code for each answer.

## Question 1

- Read the data into a data frame. (3pts)
- Display all records sorted by the PM2.5 values. (3pts)

- Calculate and show how many devices in each city. (7pts)
- Display the average PM2.5 values (over the whole time span) of each device and sort them by the PM2.5 values. You can identity which device location with the best/worst PM2.5 quality and observe some unreasonable measurements. (7pts)
- Display the average PM2.5 values (over the whole time span) of each city and sort them by the PM2.5 values. You can identity which city has the best/worst air quality. (10pts)
- Calculate the average PM2.5 of each day within each city. List the results by sorting average PM2.5 values in an ascending order within each city, i.e. you should put average PM2.5 values of the same city together. (We do not mind the order among cities). You can identity which day has the best/worst air quality in each city. (10pts)
- Calculate average PM 2.5 values (all devices in Taiwan) of each hour on Friday and do the same thing for Saturday. List the results by a table. The table has three columns whose names should be "hour", "Friday_PM2.5" and "Saturday_PM2.5". The result should be sorted by "hour" in an ascending order You can observe the difference of air quality change over time between Friday and Saturday. (10pts)

## Question 2

- Is the PM2.5 related to PM1.0? You should choose and create a visualization to support your answer. (15pts)
- Among Taipei, Pingtung, Nantou and Taichuang, which city has the most significant difference PM2.5 change pattern over the whole day on July 24 from the other three cities? You should choose and create a visualization to support your answer. (15pts)
- (Assume we only consider PM2.5 to evaluate the air quality) Please implement the following steps to answer which city (Taipei or Tainan) has a better air quality. (20pts)
  - Remove all records whose PM2.5 is 0 (simple data cleaning process to remove impossible data values)
  - Collect all records from Taipei and draw a histogram to shows how many hour PM2.5 records in each histogram bin interval. Set the histogram bin count to 100, the min value to 0 and max value to 100 to create the histogram.
  - Repeat the above step to draw a histogram for Tainan.
  - Observe these two histograms to answer which city has a better air quality in the time span of the dataset and explain your answer.