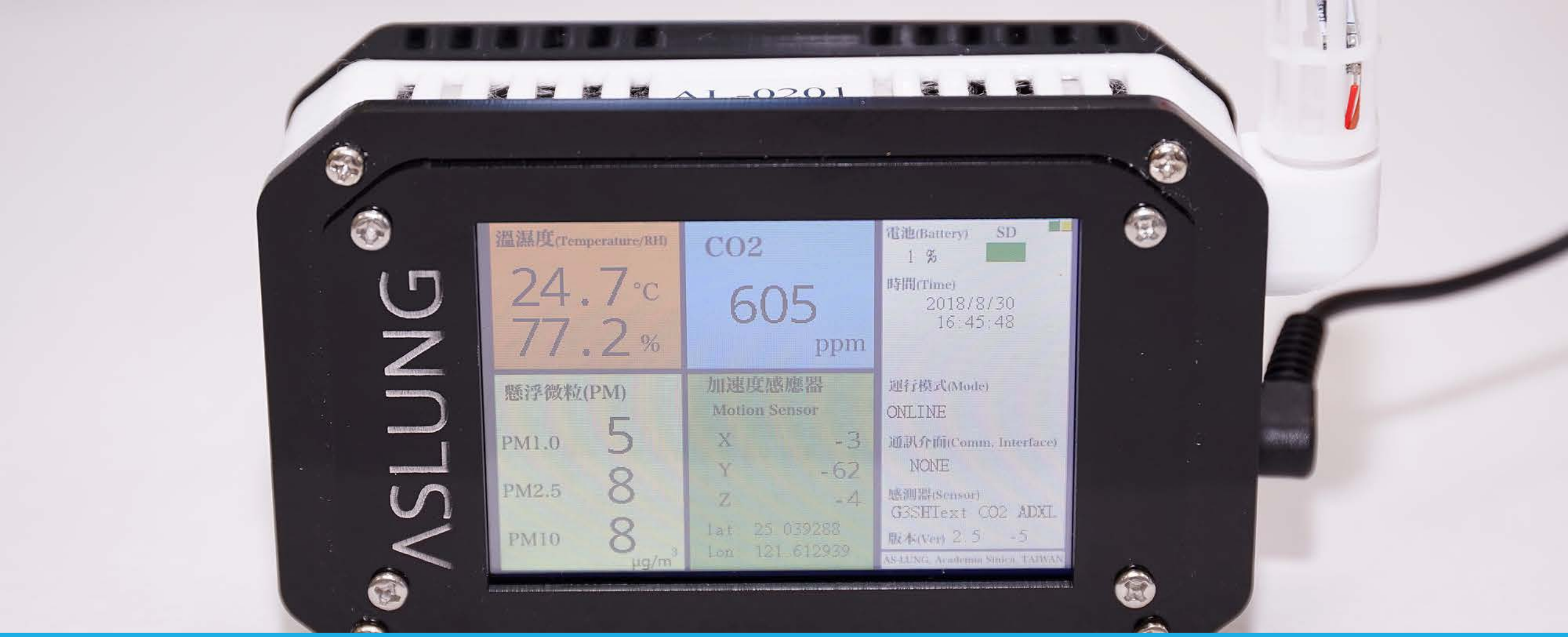


Advanced Institute on Health Investigation and Air Sensing for Asian Pollution (AI on Hi-ASAP)

Shih-Chun Candice LUNG & Chun-Hu Liu 2020/10/05

Outline

- Get calibration factor from reference instrument
  - Rational of calefaction factor
  - How to run python code to set calibration factor
- Data cleaning
  - What data cleaning tasks need to be done
  - How to run python code to do data cleaning

Get calibration factor from reference instrument

Rationale of calibration factor

AS-Lung compare with reference instrument

Reference instrument AS-Lung

Rename the filename to fit the python code

Rename the filename of AS-Lung

Ex: 2020-09-27.csv ➔AL-0001_2020-09-27.csv

The filename should be contain AS-Lung ID (AL-0001 or AL0001)

Rename the filename of reference standard

Ex: Grimm026.xls ➔standard_Grimm026.xls

The filename should be contain “standard” and with excel data file

Calibration factor

Factors of PM_1 and $PM_{2.5}$ will be reported since PM10 of G3 sensor is not reliable.

Select regression model

Simple linear regression or Two segments regression

Rename the column name to fit the python code

Data format of AS-Lung

Do not modify the data format of AL-Lung

Data format of reference instrument

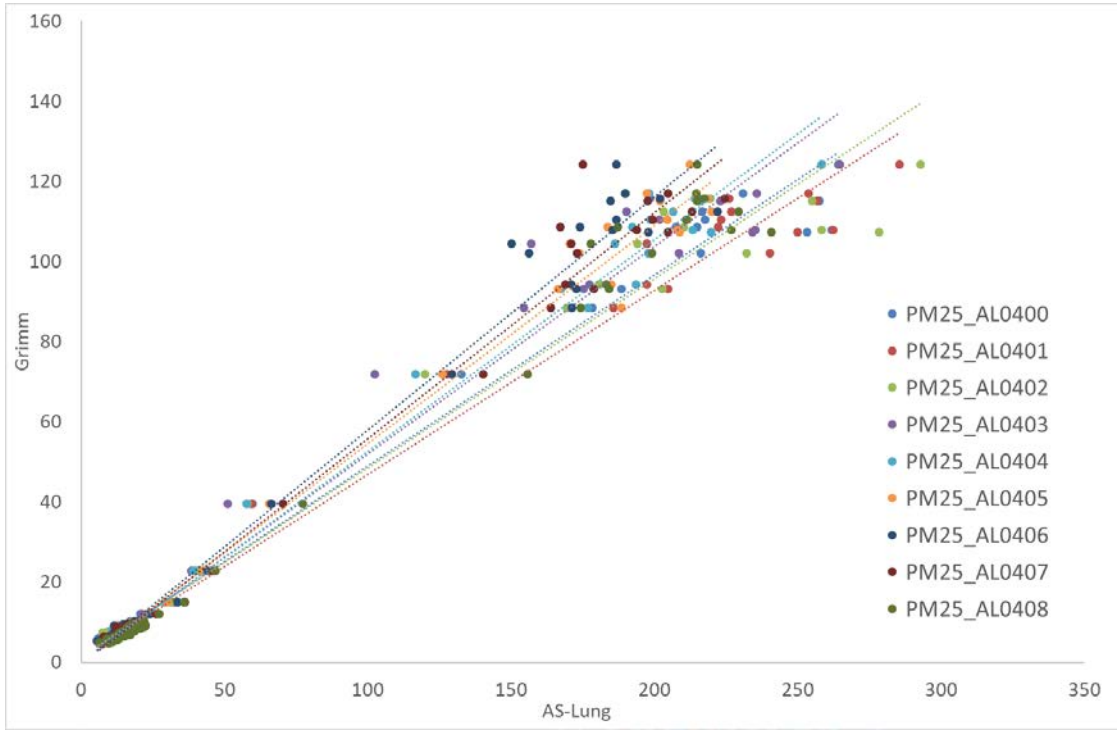
Rename the column name of time and PM

Modify the data time format of time column

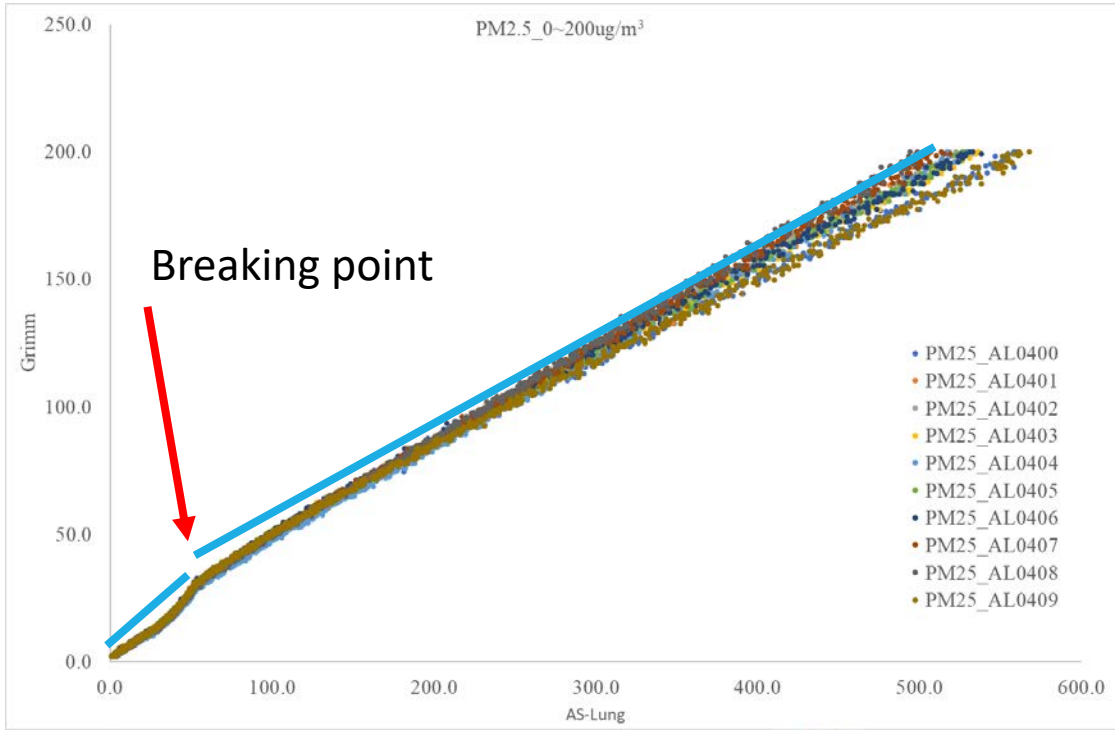
A	B	C	D
datatime	std_PM10	std_PM2.5	std_PM1
2020/4/8 13:46:01	5.5	5.4	5.1
2020/4/8 13:46:01	5.3	5.2	5

Regression model

Simple linear regression



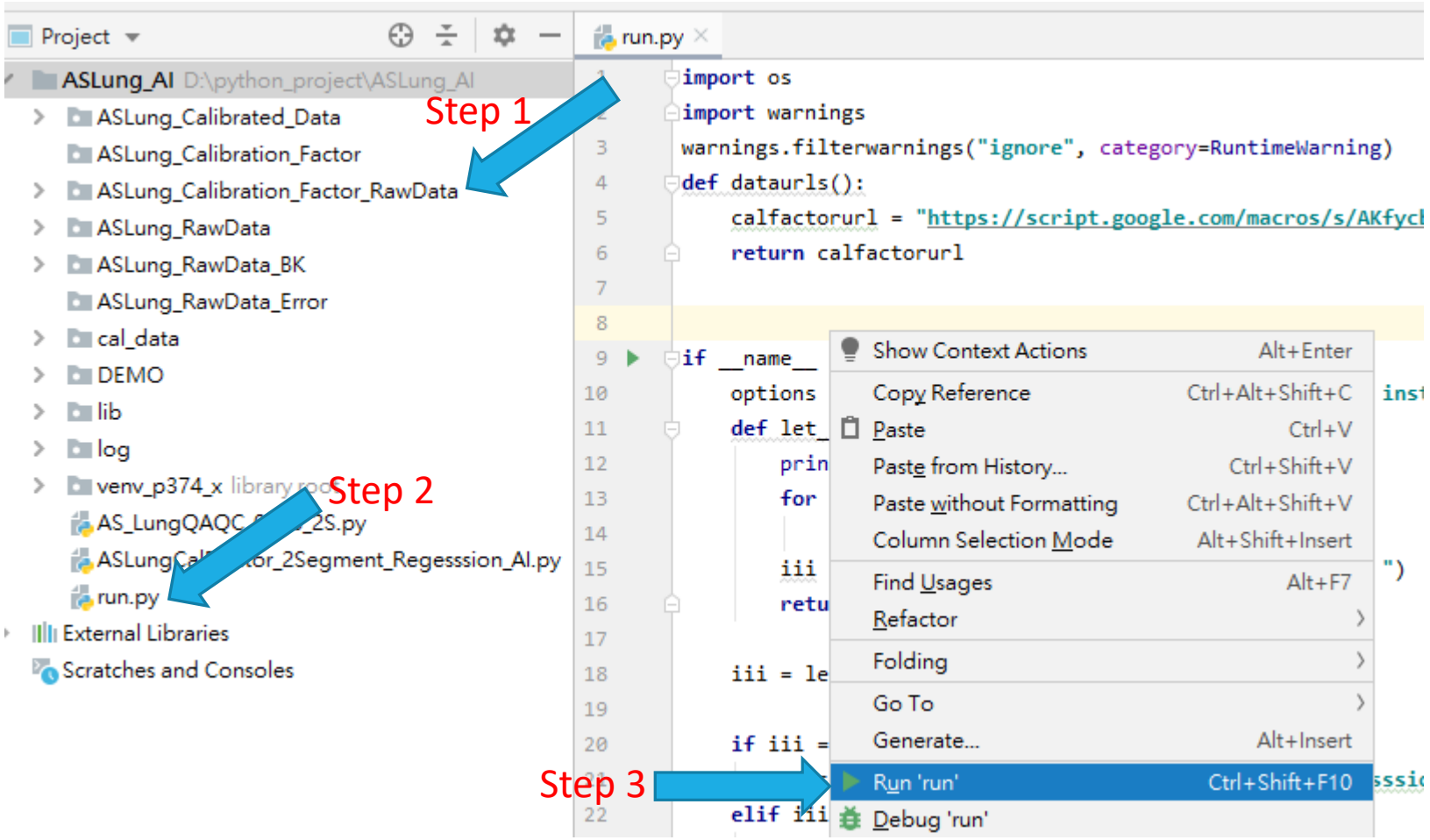
Two segments regression



When run the python code, we can select regression model.

How to run the Python Code (1)

1. Open PyCharm -> project -> run.py



Step 1:
Copy raw data to the folder of "ASLung_Calibration_Factor_RawData"

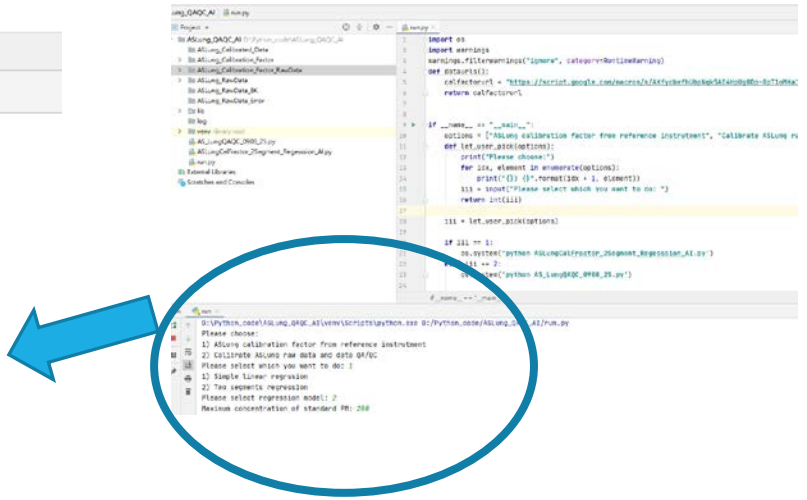
Step 2:
Open run.py (double click)

Step 3:
Run the python code

How to run the Python Code (2)

2. Select 1) ASLung calibration factor from reference instrument to run the python code

```
24
if __name__ == "__main__"  ›  let_user_pick()
run
D:\Python_code\ASLung_QAQC_AI\venv\Scripts\python.exe D:/Python_code/ASLung_QAQC_AI/run.py
Please choose:
1) ASLung calibration factor from reference instrutment
2) Calibrate ASLung raw data and data QA/QC
Please select which you want to do: 1
1) Simple linear regrssion
2) Two segments regression
Please select regression model: 2
Maximum concentration of standard PM: 200
```

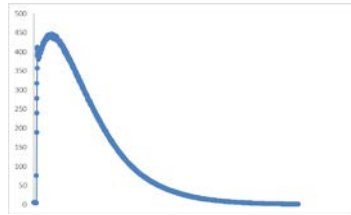


3. Select Regression model and input PM value after “Maximum concentration of standard PM”

   3.1 Maximum concentration of standard PM is the highest concentration of the regression.

   3.2 We use 200 (chamber) or 150 (hood) as maximum value.

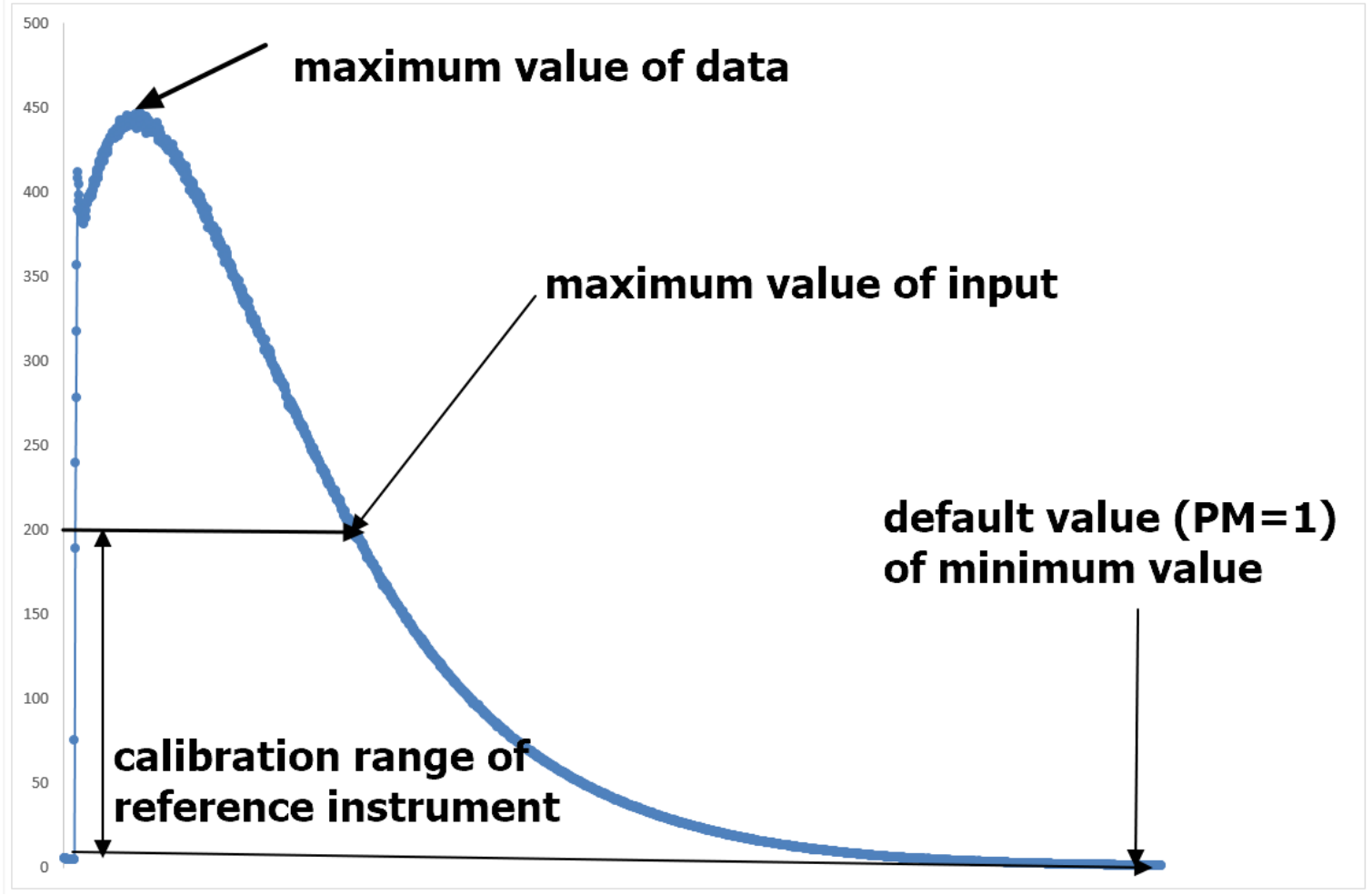
   3.2 If max value is 200, the calibration range is from 1 to 200.



4. The python code will automatically average data to 1 minutes.

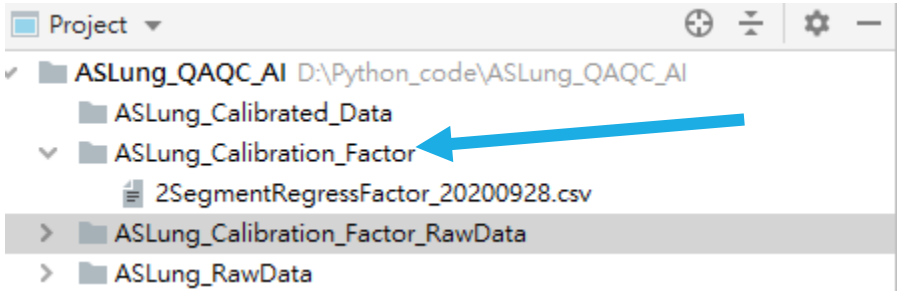
5. The python code also select the calibration range from 1 to max value of input after the maximum value of the data set

How to run the Python Code (2)

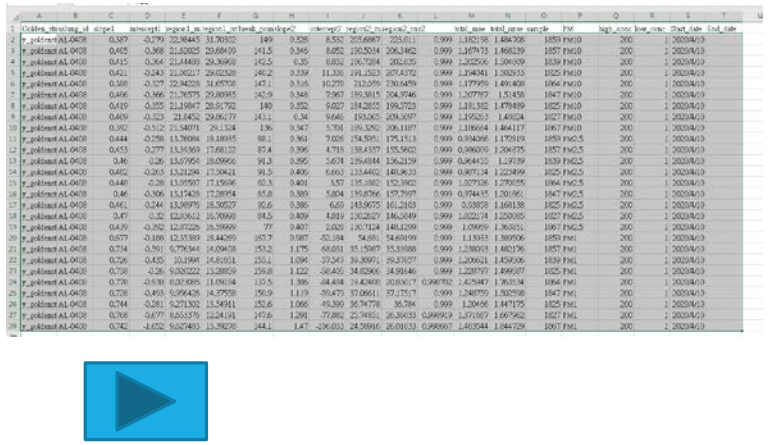


How to run the Python Code (3)

6. When finish, the file will be save in the folder of “ASLung_Calibration_Factor “



7. Open the file, copy the calibration factor and past them to the google sheet.



This google sheet only for demo and course, all the research group has individual google sheet
https://docs.google.com/spreadsheets/d/1yuvjPvsr1sEzm_pXpZWpnMxEC3WbsKn65V-apohp86E/edit#gid=0

The format between csv file and google sheet are the same, DONOT change the data format

Google sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Golden_	aslung_	slope1	intercep	region1_	region1_	break_p	slope2	intercep	region2_	region2_	r2	total_ma	total_rm	sample	PM	high_co	low_con	Start_date	End_date
2	y_golde	AL-0045	0.498	2.982			10000					0.931				PM1	150	1	2018/1/19	2020/2/15
3	y_golde	AL-0045	0.53	7.472			10000					0.987				PM1	150	1	2020/2/15	
4	y_golde	AL-0077	0.776	1.93			10000					0.943				PM1	150	1	2018/1/19	2020/5/29
5	y_golde																		2020/5/29	
6	y_golde																		2018/3/22	2019/1/30
7	y_golde																		2019/1/30	
8	y_golde	AL-0107	0.377	6.113			10000					0.777				PM1	150	1	2018/3/28	
9	y_golde	AL-0120	0.663	4.441			10000					0.994				PM1	150	1	2018/4/9	
10	y_golde	AL-0124	0.681	2.18			10000					0.994				PM1	150	1	2018/4/12	
11	y_golde	AL-0125	0.697	3.723			10000					0.993				PM1	150	1	2018/4/17	2019/1/31
12	y_golde	AL-0125	0.383	3.42			10000					0.979				PM1	150	1	2019/1/31	
13	y_golde	AL-0128	0.709	2.662			10000					0.994				PM1	150	1	2018/4/17	
14	y_golde	AL-0131	0.536	0.353			10000					0.994				PM1	150	1	2018/9/10	2019/1/31
15	y_golde	AL-0131	0.501	2.731			10000					0.986				PM1	150	1	2019/1/31	
16	y_golde	AL-0136	0.61	0.422			10000					0.992				PM1	150	1	2018/9/10	
17	y_golde	AL-0138	0.694	2.29			10000					0.992				PM1	150	1	2018/4/26	2020/7/7
18	y_golde	AL-0138	0.647	2.663			10000					0.997				PM1	150	1	2020/7/7	
19	y_golde	AL-0139	0.591	2.215			10000					0.988				PM1	150	1	2018/4/26	2019/1/31
20	y_golde	AL-0139	0.662	2.389			10000					0.978				PM1	150	1	2019/1/31	2020/7/7
21	y_golde	AL-0139	0.549	5.613			10000					0.994				PM1	150	1	2020/7/7	

1. If you renew the calibration factor, please key in the end date in the “End_date” column.
2. Add a new row to add new calibration factor, DO NOT replace the old factor

Data cleaning

What data cleaning tasks need to be done

AS-Lung monitoring data

There are two ways toe get raw data
1.data in SD card or 2.Data from database

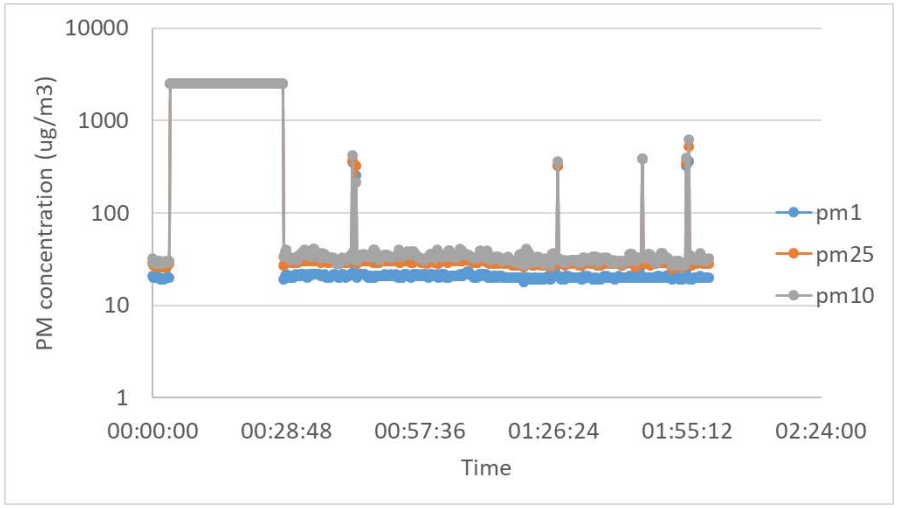
Data cleaning criteria

Step of data cleaning

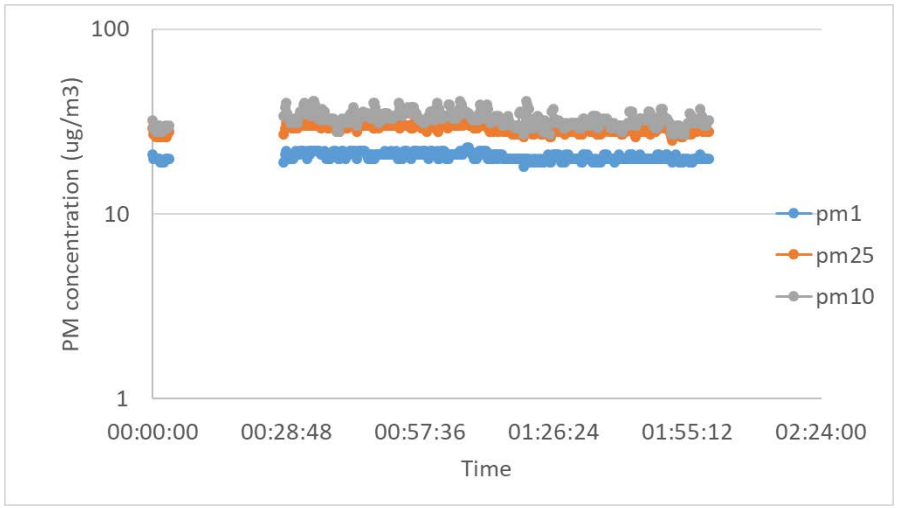
1. PM < 1
2. PM > 50 and PM1=PM2.5=PM10
3. Remove ghost peak

Set PM value as NaN
NaN is equal null

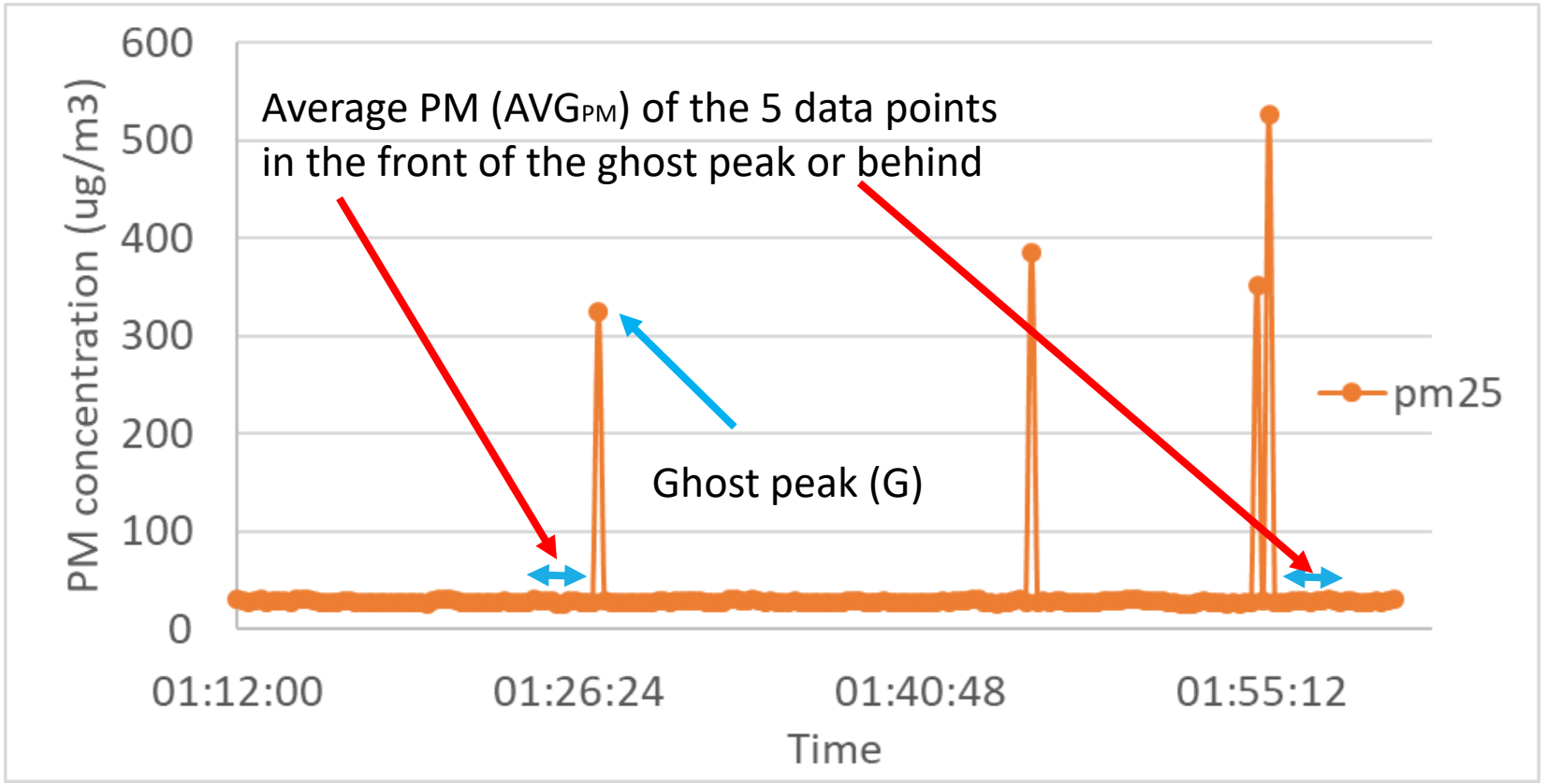
Before



After



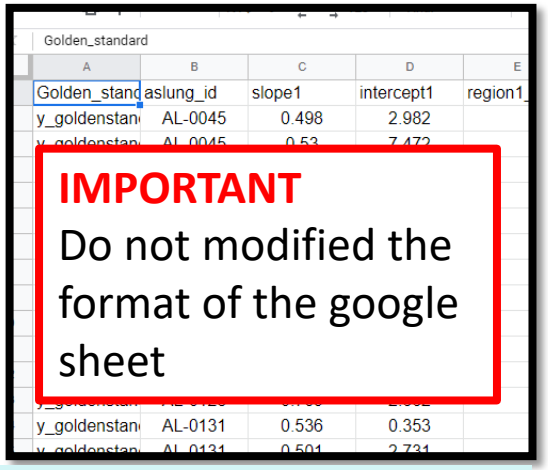
Remove ghost peak



$$\frac{G}{AVG_{PM}} > 10$$

What data cleaning tasks need to be done

4. Remove data of temperature, humidity and CO_2 when values are less than 1
5. Get calibration factor from google drive and calibrate AS-Lung data


Golden_standard

Golden_stand	aslung_id	slope1	intercept1	region1_
y_goldenstan	AL-0045	0.498	2.982	
y_goldenstan	AL-0131	0.536	0.353	

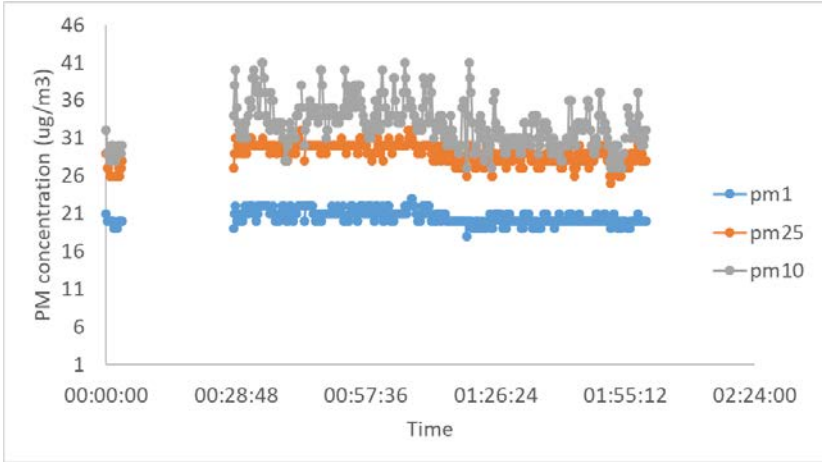
IMPORTANT
Do not modified the format of the google sheet

Calibration factor in google sheet

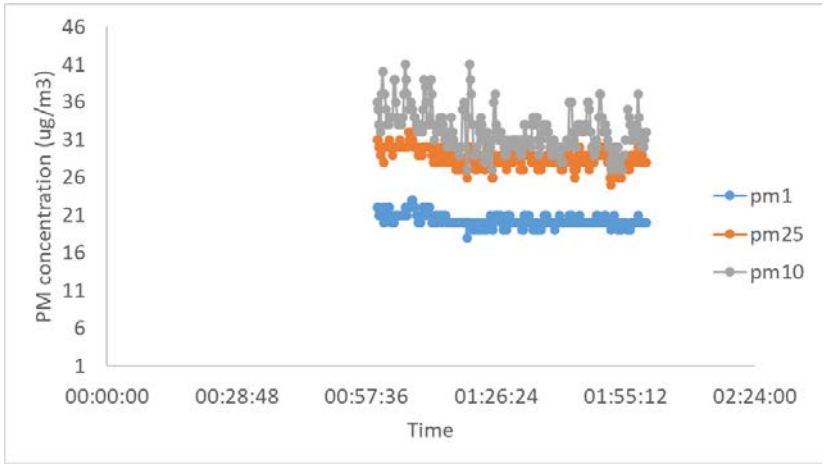
{"ASLUNG":
[{"Golden_standard":"Golden_standard","aslung_id":"aslung_i
","region2_rmse":"region2_rmse","r2":"r2","total_mae":"tota
{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0045","slope1":0.498,"intercept1":2.982,"region1_mae":"","re
e":"2018/1/19","End_date":"2020/2/15"},{"Golden_standard":"
0045","slope1":0.53,"intercept1":7.472,"region1_mae":"","re
":"2020/2/15","End_date":""},{"Golden_standard":"y_goldenst
0077","slope1":0.776,"intercept1":1.93,"region1_mae":"","re
":"2018/1/19","End_date":"2020/5/29"},{"Golden_standard":"y
0077","slope1":0.719,"intercept1":3.705,"region1_mae":"","re
e":"2020/5/29","End_date":""},{"Golden_standard":"y_goldens
0102","slope1":0.696,"intercept1":2.259,"region1_mae":"","re
e":"2018/3/22","End_date":"2019/1/30"},{"Golden_standard":"
0102","slope1":0.438,"intercept1":2.515,"region1_mae":"","re
te":"2019/1/30","End_date":""},{"Golden_standard":"y_golden
0107","slope1":0.377,"intercept1":6.113,"region1_mae":"","re
te":"2018/3/28","End_date":""},{"Golden_standard":"y_golden
0120","slope1":0.663,"intercept1":4.441,"region1_mae":"","re
te":"2018/4/9","End_date":""},{"Golden_standard":"y_goldens
0124","slope1":0.681,"intercept1":2.18,"region1_mae":"","re
e":"2018/4/12","End_date":""},{"Golden_standard":"y_goldens
0125","slope1":0.697,"intercept1":3.723,"region1_mae":"","re
te":"2018/4/17","End_date":"2019/1/31"},{"Golden_standard":
0125","slope1":0.383,"intercept1":3.42,"region1_mae":"","re
e":"2019/1/31","End_date":""},{"Golden_standard":"y_goldens
0128","slope1":0.709,"intercept1":2.662,"region1_mae":"","re
te":"2018/4/17","End_date":""},{"Golden_standard":"y_golden

Calibration factor API from google sheet

Before



After

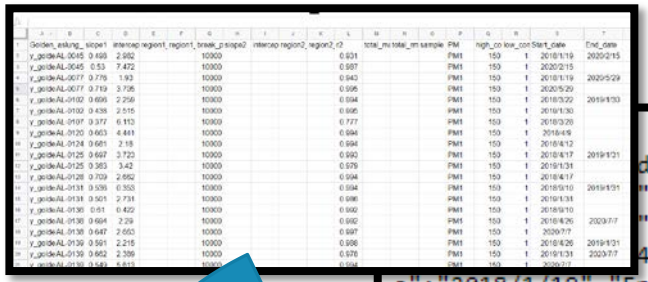


6. If calibrated $PM_1 > PM_{2.5}$, $PM_1 = PM_{2.5}$
7. If the missing data is more than 1/3 in an hour, the python code will auto remove all the data in the hour

Google sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Golden_	aslung_	slope1	intercep	region1_	region1_	break_p	slope2	intercep	region2_	region2_	r2	total_ma	total_rm	sample	PM	high_co	low_con	Start_date	End_date
2	y_golde	AL-0045	0.498	2.982			10000					0.931				PM1	150	1	2018/1/19	2020/2/15
3	y_golde	AL-0045	0.53	7.472			10000					0.987				PM1	150	1	2020/2/15	
4	y_golde	AL-0077	0.776	1.93			10000					0.943				PM1	150	1	2018/1/19	2020/5/29
5	y_golde	AL-0077	0.719	3.705			10000					0.995				PM1	150	1	2020/5/29	
6	y_golde	AL-0102	0.696	2.259			10000					0.994				PM1	150	1	2018/3/22	2019/1/30
7	y_golde	AL-0102	0.438	2.515			10000					0.995				PM1	150	1	2019/1/30	
8	y_golde	AL-0107	0.377	6.113			10000					0.777				PM1	150	1	2018/3/28	
9	y_golde	AL-0120	0.663	4.441			10000					0.994				PM1	150	1	2018/4/9	
10	y_golde	AL-0124	0.681	2.18			10000					0.994				PM1	150	1	2018/4/12	
11	y_golde	AL-0125	0.697	3.723			10000					0.993				PM1	150	1	2018/4/17	2019/1/31
12	y_golde	AL-0125	0.383	3.42			10000					0.979				PM1	150	1	2019/1/31	
13	y_golde	AL-0128	0.709	2.662			10000					0.994				PM1	150	1	2018/4/17	
14	y_golde	AL-0131	0.536	0.353			10000					0.994				PM1	150	1	2018/9/10	2019/1/31
15	y_golde	AL-0131	0.501	2.731			10000					0.986				PM1	150	1	2019/1/31	
16	y_golde	AL-0136	0.61	0.422			10000					0.992				PM1	150	1	2018/9/10	
17	y_golde	AL-0138	0.694	2.29			10000					0.992				PM1	150	1	2018/4/26	2020/7/7
18	y_golde	AL-0138	0.647	2.663			10000					0.997				PM1	150	1	2020/7/7	
19	y_golde	AL-0139	0.591	2.215			10000					0.988				PM1	150	1	2018/4/26	2019/1/31
20	y_golde	AL-0139	0.662	2.389			10000					0.978				PM1	150	1	2019/1/31	2020/7/7
21	y_golde	AL-0139	0.549	5.613			10000					0.994				PM1	150	1	2020/7/7	

Calibration factor API



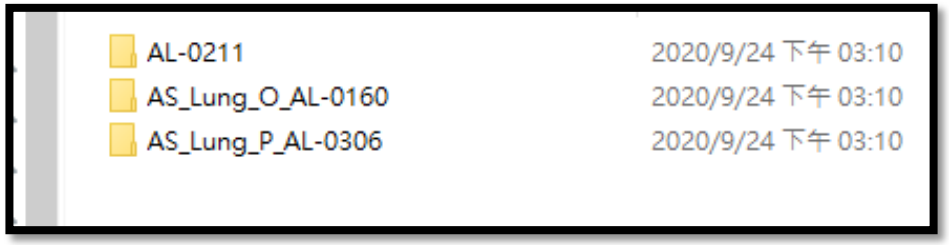
https://script.google.com/macros/s/AKfycbwfhUbpNqk5AE4HpUg0Dp-0pT1oMKa1mxLzWWAXb3dlnhTYRN8/exec


```
d":"Golden_standard","aslung_id":"aslung_id","slope1":"slope1","intercept1":"intercept1","region1_mae":"region1_mae","region1_rms
"region2_rmse","r2":"r2","total_mae":"total_mae","total_mse":"total_rmse","sample":"sample","PM":"PM","high_conc":"high_conc","lo
":"y_goldenstand","aslung_id":"AL-
498,"intercept1":2.982,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
e":"2018/1/19","End_date":"2020/2/15"},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0045","slope1":0.53,"intercept1":7.472,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regio
":"2020/2/15","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0077","slope1":0.776,"intercept1":1.93,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regio
":"2018/1/19","End_date":"2020/5/29"},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0077","slope1":0.719,"intercept1":3.705,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
e":"2020/5/29","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0102","slope1":0.696,"intercept1":2.259,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
e":"2018/3/22","End_date":"2019/1/30"},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0102","slope1":0.438,"intercept1":2.515,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
e":"2019/1/30","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0107","slope1":0.377,"intercept1":6.113,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
te":"2018/3/28","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0120","slope1":0.663,"intercept1":4.441,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
te":"2018/4/9","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0124","slope1":0.681,"intercept1":2.18,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regio
e":"2018/4/12","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0125","slope1":0.697,"intercept1":3.723,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regi
te":"2018/4/17","End_date":"2019/1/31"},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
0125","slope1":0.383,"intercept1":3.42,"region1_mae":"","region1_rmse":"","break_point1":10000,"slope2":"","intercept2":"","region2_mae":"","regio
e":"2019/1/31","End_date":""},{"Golden_standard":"y_goldenstand","aslung_id":"AL-
```


We create calibration factor API from google sheet. Keep the format of the google sheet is very important.

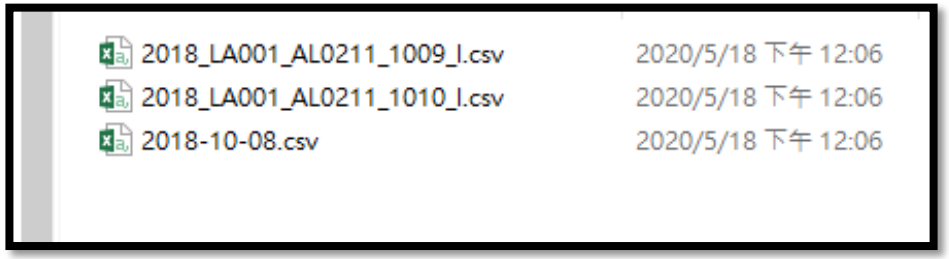
Data prepare

1.Folder name of the data set should be contain AS-Lung ID, EX:AL-0001



AS-Lung ID only
Or
Lab ID before AS-Lung ID ,
separate by “_”

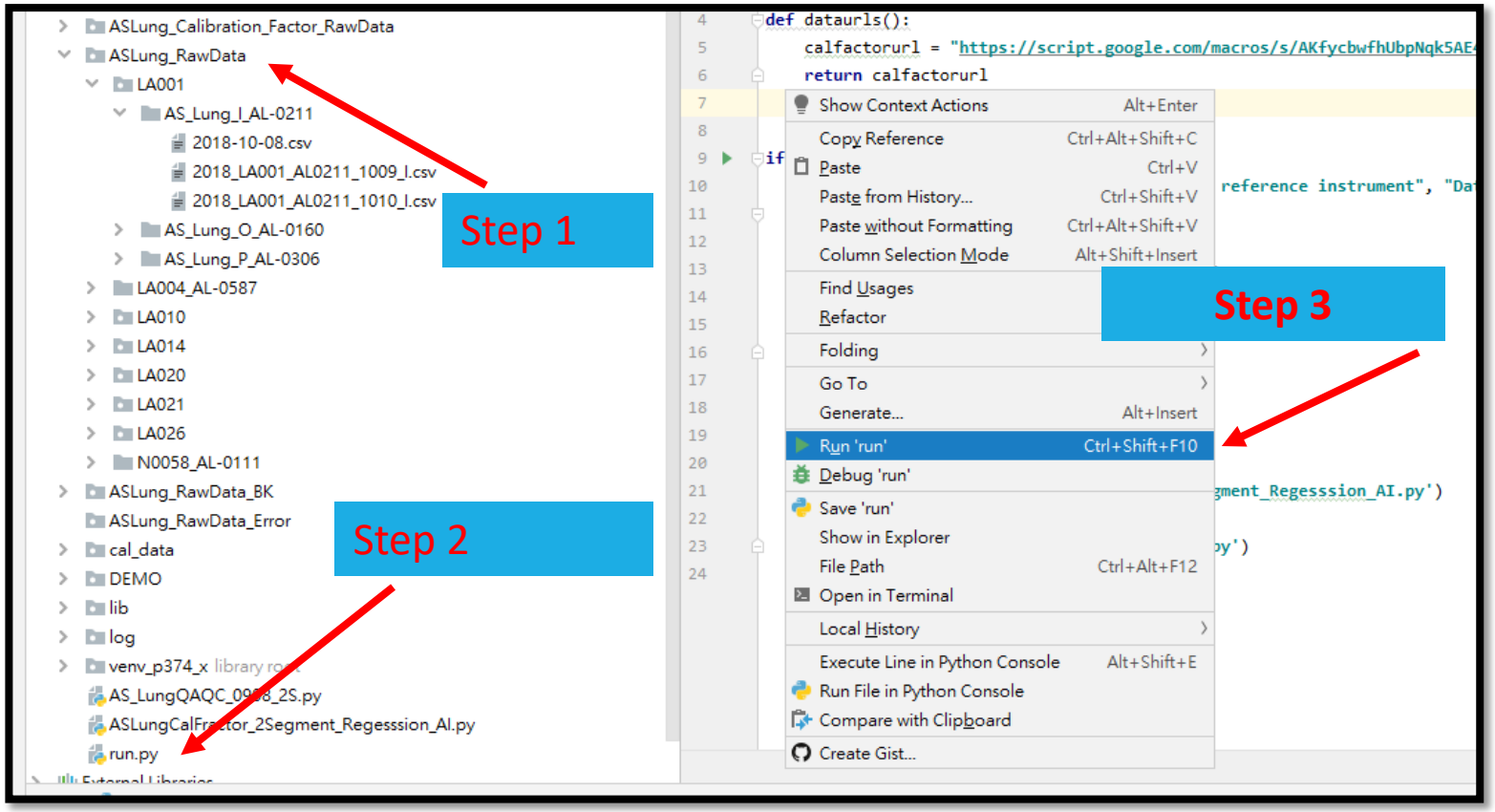
2.Rename the filename is optional



Keep the original filename
Or
Rename the filename
it is optional

How to run data cleaning

3. Open PyCharm -> project -> run.py



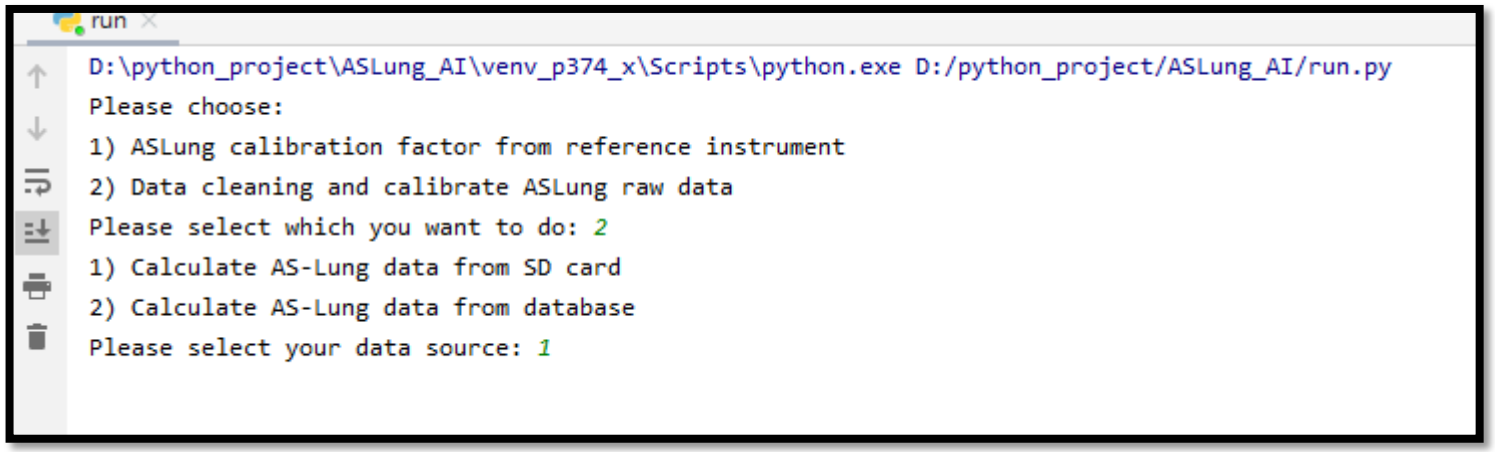
Step 1:
Copy raw data to the folder of "ASLung_RawData"

Step 2:
Open run.py (double click)

Step 3:
Run the python code

How to run data cleaning

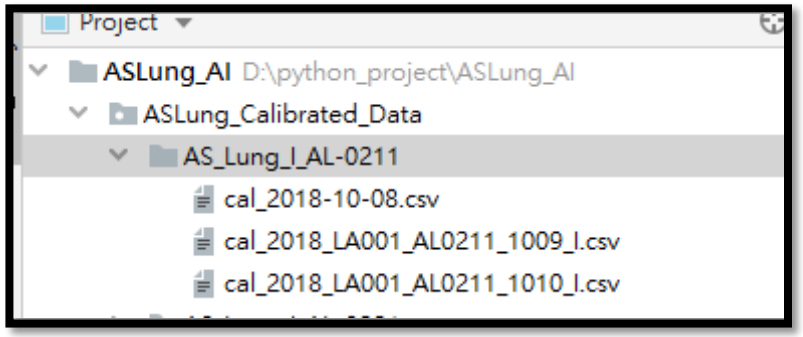
4. Select 2) Data cleaning and calibrate ASLung raw data



5. Select your data source, SD card or database

6 . Calibrated data will be save in the folder of

“ASLung_Calibrated_Data”, the filename will add prefix of “cal_” in the front of original filename

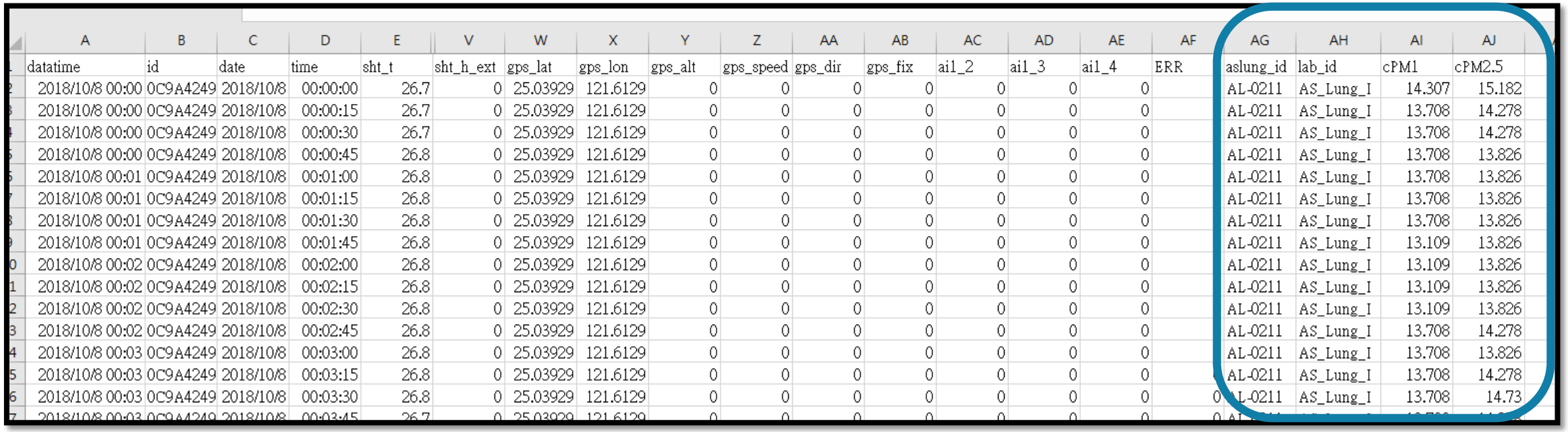


Step of data cleaning will remind again when run the python code

```
run
D:\python_project\ASLung_AI\venv_p374_x\Scripts\python.exe D:/python_project/ASLung_AI/run.py
Please choose:
1) ASLung calibration factor from reference instrument
2) Data cleaning and calibrate ASLung raw data
Please select which you want to do: 2
1) Calculate AS-Lung data from SD card
2) Calculate AS-Lung data from database
Please select your data source: 1
================================================================================================
Data cleaning and calibrate AS-Lung data, Please wait!
Setp of data cleaning and calibration
1. Set raw data of PM as NaN when PM >50 and PM1=PM2.5=PM10 or PM <1
2. Set ghost Peak as NaN
3. Set raw data of temperature, humidity and CO2 as NaN when values are less than 1
4. Get calibration factor from google drive and calibrate AS-Lung data
5. If calibrated PM1 > PM2.5, PM1 value will be set as PM2.5
6. If the missing data is more than 1/3 in an hour, the python code will auto remove all the data in the hour
================================================================================================
```

Open CSV file after calibrated

Open the calibrated csv file, AS-Lung ID, lab id , calibrated PM_1 and $PM_{2.5}$ will add in the end column of the data set



	A	B	C	D	E	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	datatime	id	date	time	sht_t	sht_h_ext	gps_lat	gps_lon	gps_alt	gps_speed	gps_dir	gps_fix	ai1_2	ai1_3	ai1_4	ERR	aslung_id	lab_id	cPM1	cPM2.5
2	2018/10/8 00:00	0C9A4249	2018/10/8	00:00:00	26.7	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	14.307	15.182
3	2018/10/8 00:00	0C9A4249	2018/10/8	00:00:15	26.7	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	14.278
4	2018/10/8 00:00	0C9A4249	2018/10/8	00:00:30	26.7	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	14.278
5	2018/10/8 00:00	0C9A4249	2018/10/8	00:00:45	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	13.826
6	2018/10/8 00:01	0C9A4249	2018/10/8	00:01:00	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	13.826
7	2018/10/8 00:01	0C9A4249	2018/10/8	00:01:15	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	13.826
8	2018/10/8 00:01	0C9A4249	2018/10/8	00:01:30	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	13.826
9	2018/10/8 00:01	0C9A4249	2018/10/8	00:01:45	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.109	13.826
10	2018/10/8 00:02	0C9A4249	2018/10/8	00:02:00	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.109	13.826
11	2018/10/8 00:02	0C9A4249	2018/10/8	00:02:15	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.109	13.826
12	2018/10/8 00:02	0C9A4249	2018/10/8	00:02:30	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.109	13.826
13	2018/10/8 00:02	0C9A4249	2018/10/8	00:02:45	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	14.278
14	2018/10/8 00:03	0C9A4249	2018/10/8	00:03:00	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	13.826
15	2018/10/8 00:03	0C9A4249	2018/10/8	00:03:15	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0		AL-0211	AS_Lung_I	13.708	14.278
16	2018/10/8 00:03	0C9A4249	2018/10/8	00:03:30	26.8	0	25.03929	121.6129	0	0	0	0	0	0	0	0	AL-0211	AS_Lung_I	13.708	14.73
17	2018/10/8 00:03	0C9A4249	2018/10/8	00:03:45	26.7	0	25.03929	121.6129	0	0	0	0	0	0	0	0				

How to update calibration factor API link

1.Open run.py

2.Go to line 5, find **calfactorurl = “xxxxxxxxxxx”**

run.py

```
1  import os
2  import warnings
3  warnings.filterwarnings("ignore", category=RuntimeWarning)
4  def dataurls():
5      calfactorurl = "https://script.google.com/macros/s/AKfycbwfhUbpNqk5AE4HpUg0Dp-0pT1oMKa1mxLzWWAXb3dlnhTYRN8/exec"
6      return calfactorurl
7
8
```

3. Past the new link after **calfactorurl=“**

You can get your google sheet and calibration factor API link from “GoogleSheetAndAPI.xlsx

Thanks!