2021 Advanced Institute on Health Investigation and Air Sensing for Asian Pollution (AI on Hi-ASAP)

Q & A for Exam I

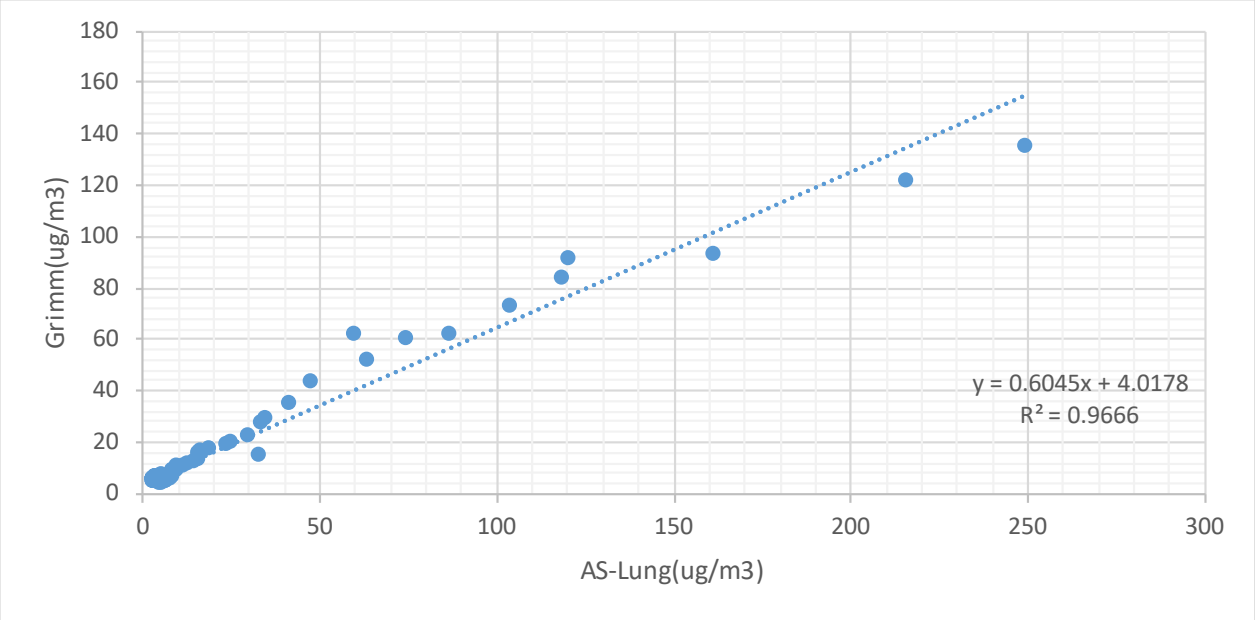
Speaker: Chun-Hu Liu

Data cleaning

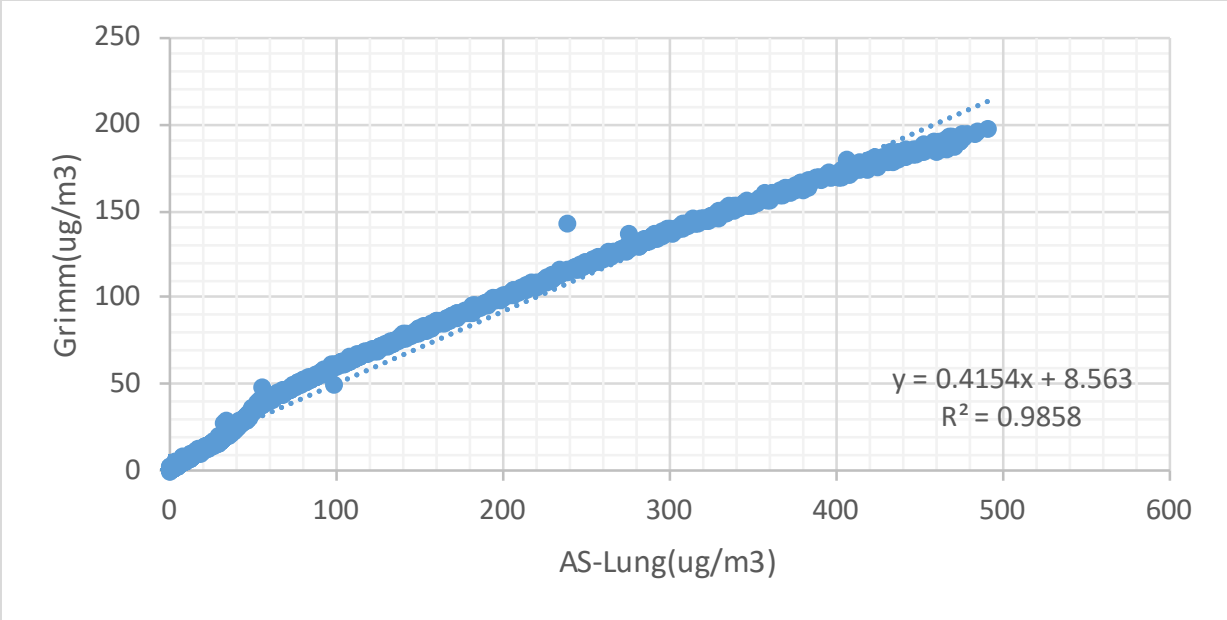
Q1.
There are two calibration data set. One is for simple linear regression and another is for two segments regression. Please get the calibration factor from the two data set and past them to the google sheet (Maximum concentration is 200). You can get the google sheet link from “GoogleSheetAndAPI.xlsx”.

A. Which data set is simple linear regression?
B. What are the regression factors of AL-0216 in data set 2? Slope, intercept and R^2

Q1.a Which data set is simple linear regression?



Dataset 1



Dataset 2

Step 1. plot simple linear regression via excel

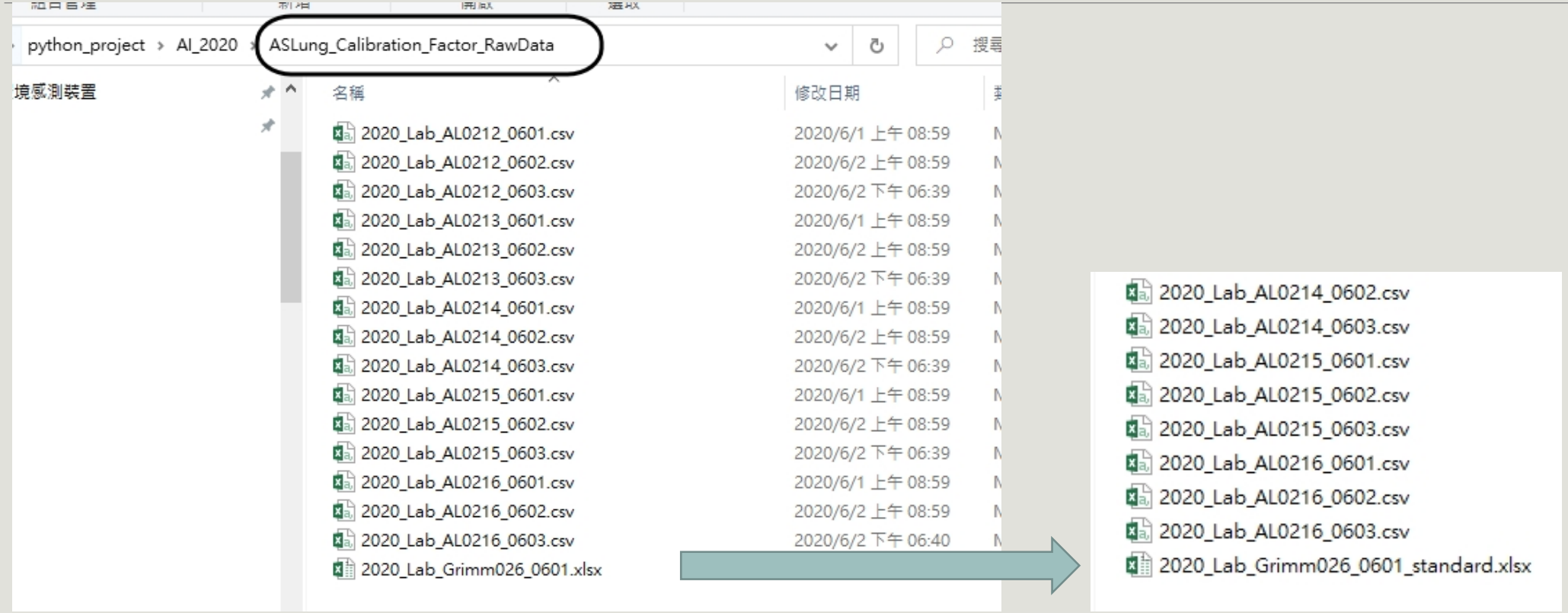
Step 2. there are not enough data point at high concentration

Step 3. there is a recurve point at dataset 2 and enough data point data at high concentration

Get calibration factor

WHAT ARE THE REGRESSION FACTORS OF AL-0216 IN DATA SET 2? SLOPE, INTERCEPT AND R^2

Q1.b What are the regression factors of AL-0216 in data set 2? Slope, intercept and R^2



Step 1. Copy Dataset to ASLung_Calibration_Factor_RawData and rename the file name of reference PM with keyword “standard”

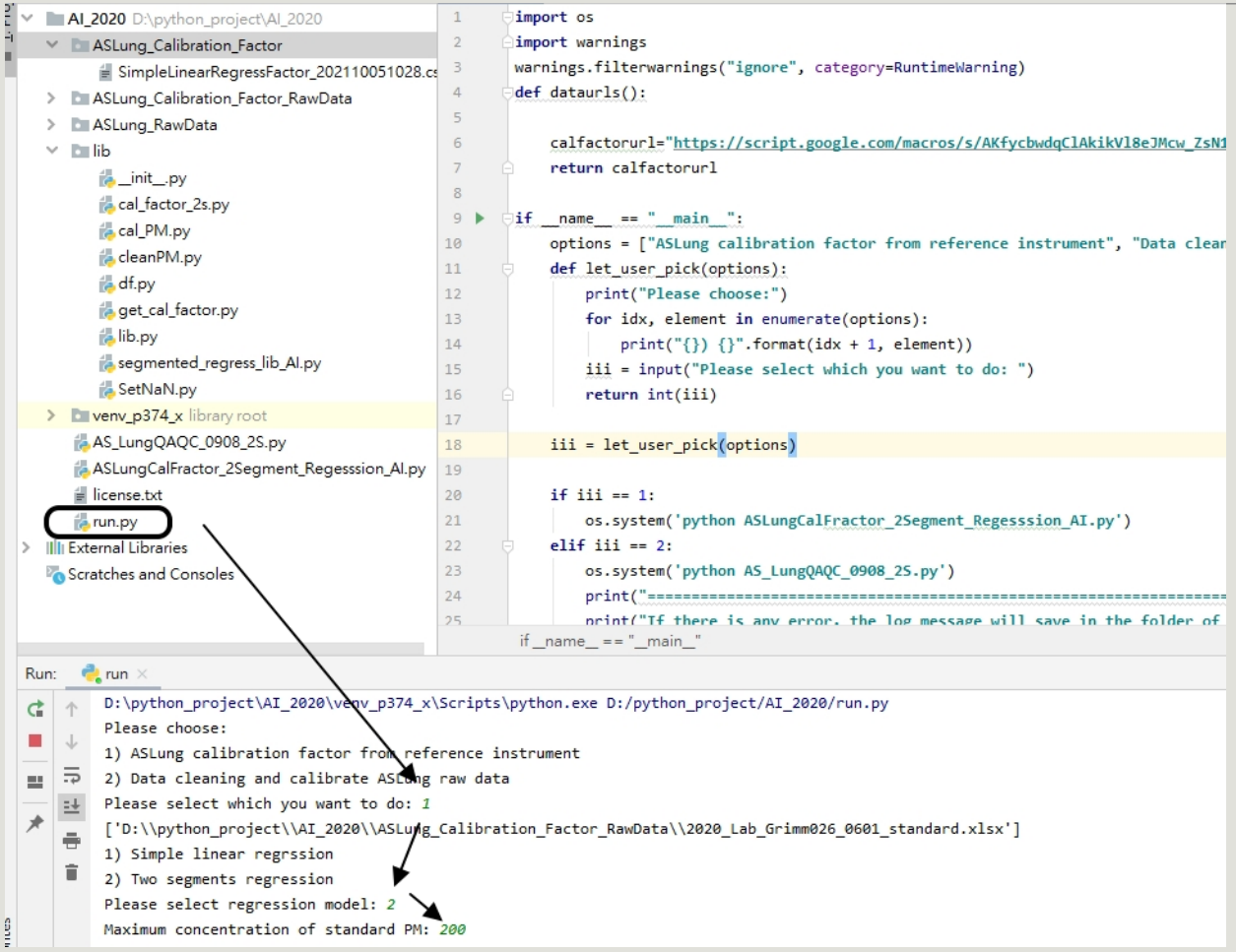
Q1.b What are the regression factors of AL-0216 in data set 2? Slope, intercept and R^2

	A	B	C	D	E	F	G	H
1	<Header> User name: Location: Model: 11-A Serial No.: 11A16026 Firmware revision: 12.30E Software revision: 4-1 Rev XIX (24-05-2016) Unit: ug/m3							
2								
3								
4								
5	date & time	PM10 [ug/r	PM2.5 [ug/	PM1 [ug/m	Inhalable [	Thoracic [u	Alveolic [ug/m3]	
6	2018/04/16 17:37:01	32.6	11.8	5.5	59.4	37.9	15.6	
7	2018/04/16 17:37:07	45.4	16.2	6.2	45.7	41.7	21.2	
8	2018/04/16 17:37:13	37	19.5	10.8	37.3	34.5	22.8	
9	2018/04/16 17:37:19	34.9	11.5	5.8	39.9	33.1	16.3	
10	2018/04/16 17:37:25	36.7	9.4	6.9	63.5	40.7	13	
11	2018/04/16 17:37:31	27.5	8.5	5.9	32.5	26.1	11.3	
12	2018/04/16 17:37:37	18.8	10.7	6.2	18.8	18.7	13.6	
13	2018/04/16 17:37:43	30.8	10.1	7.2	31.1	28.1	13.7	
14	2018/04/16 17:37:49	30.7	9.8	7	52.4	36	14.7	
15	2018/04/16 17:37:55	30.8	12	7.5	57.5	35.7	14.9	
16	2018/04/16 17:38:01	28.8	11.7	9.4	29.2	26.7	14.5	

	A	B	C	D	E	F	G	H
1	datatime	std_PM10	std_PM2.5	std_PM1	Inhalable [	Thoracic [u	Alveolic [ug/m3]	
2	2018/04/16 17:37:01	32.6	11.8	5.5	59.4	37.9	15.6	
3	2018/04/16 17:37:07	45.4	16.2	6.2	45.7	41.7	21.2	
4	2018/04/16 17:37:13	37	19.5	10.8	37.3	34.5	22.8	
5	2018/04/16 17:37:19	34.9	11.5	5.8	39.9	33.1	16.3	
6	2018/04/16 17:37:25	36.7	9.4	6.9	63.5	40.7	13	
7	2018/04/16 17:37:31	27.5	8.5	5.9	32.5	26.1	11.3	
8	2018/04/16 17:37:37	18.8	10.7	6.2	18.8	18.7	13.6	

Step 2. rename column of reference PM and delete row 1 to row 4

Q1.b What are the regression factors of AL-0216 in data set 2? Slope, intercept and R^2



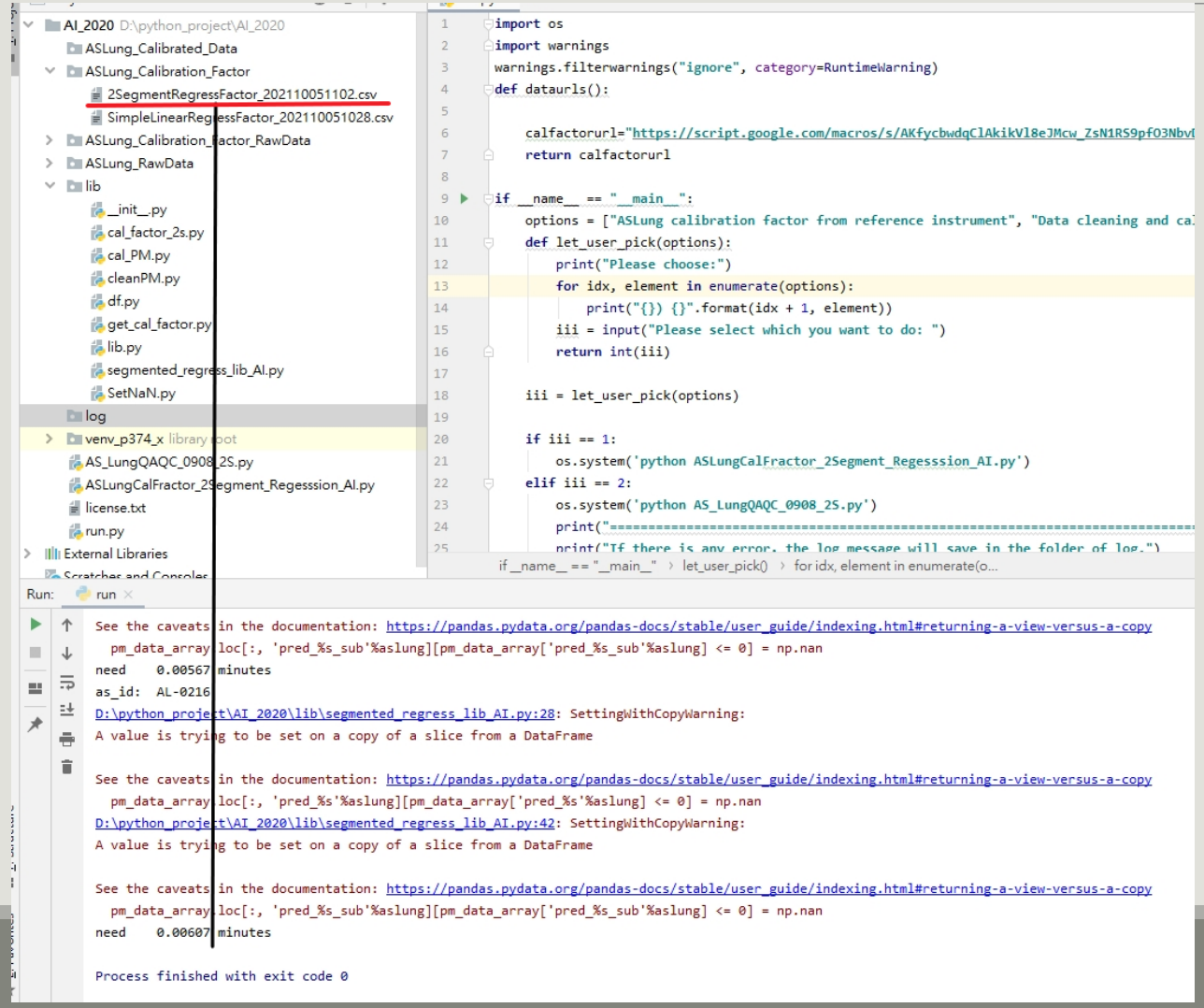
Step 3. run python code

Step 4. Select 1: ASLung Calibration factor from reference instrutment

Step 5. Select 2: Two segments regression

Step 6. Input 200: Maximum concentration of standard PM

Q1.b What are the regression factors of AL-0216 in data set 2? Slope, intercept and R^2



Step 7. When python finish, it will show "Process finished exit code 0"

Step 8. The calibration factions will save in "2segmentRegressionFactionxxx.csv" file

Q1.b What are the regression factors of AL-0216 in data set 2? Slope, intercept and R^2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Golden_standard	aslung_id	slope1	intercept1	region1_mae	region1_rmse	break_point1	slope2	intercept2	region2_mae	region2_rmse	r2	total_mae	total_rmse	sample	PM	high_conc	low_conc	Start_date	End_date
2	y_goldenstand	AL-0212	0.636	1.309	10.69110668	15.59651992	101.6	0.35	30.344	162.9780282	179.2623209	0.998929847	1.452288042	1.971093021	2009	PM2.5	200	1	2020/6/3	
3	y_goldenstand	AL-0213	0.597	0.948	10.85181733	15.71952827	98.1	0.343	25.929	171.0987424	189.8476768	0.999	1.253312292	1.681788926	2079	PM2.5	200	1	2020/6/3	
4	y_goldenstand	AL-0214	0.594	1.046	11.39977227	16.5237682	100.2	0.334	27.077	177.6203625	196.768013	0.999	1.268840056	1.703799943	2058	PM2.5	200	1	2020/6/3	
5	y_goldenstand	AL-0215	0.618	1.458	11.86555889	17.38533841	104.3	0.306	34.005	194.5645703	214.7304115	0.998846211	1.486286028	2.046609234	2010	PM2.5	200	1	2020/6/3	
6	y_goldenstand	AL-0216	0.574	1.053	11.84882477	17.17266265	98.9	0.318	26.358	191.6033689	213.0114306	0.999	1.11899021	1.491030187	2070	PM2.5	200	1	2020/6/3	
7	y_goldenstand	AL-0212	0.944	0.846	0.494928947	0.59237292	50.2	0.634	16.426	42.48252661	50.46678468	0.999	1.234079622	1.718069495	2009	PM1	200	1	2020/6/3	
8	y_goldenstand	AL-0213	0.851	0.618	1.589543601	2.30257094	47.7	0.601	12.555	53.85883231	63.05176569	0.999	1.302254739	1.772557772	2079	PM1	200	1	2020/6/3	
9	y_goldenstand	AL-0214	0.87	0.704	1.297618055	1.887068814	47.7	0.609	13.138	51.13278237	60.13584487	0.999	1.258186334	1.698384261	2058	PM1	200	1	2020/6/3	
10	y_goldenstand	AL-0215	0.944	1.014	0.68596275	0.809721274	55.4	0.583	21.015	51.90423515	61.21700718	0.999	1.276783183	1.765149894	2010	PM1	200	1	2020/6/3	
11	y_goldenstand	AL-0216	0.818	0.583	2.394982176	3.418482095	53.1	0.525	16.14	73.77153099	85.79145656	0.999	1.056191168	1.407530833	2070	PM1	200	1	2020/6/3	
12																				

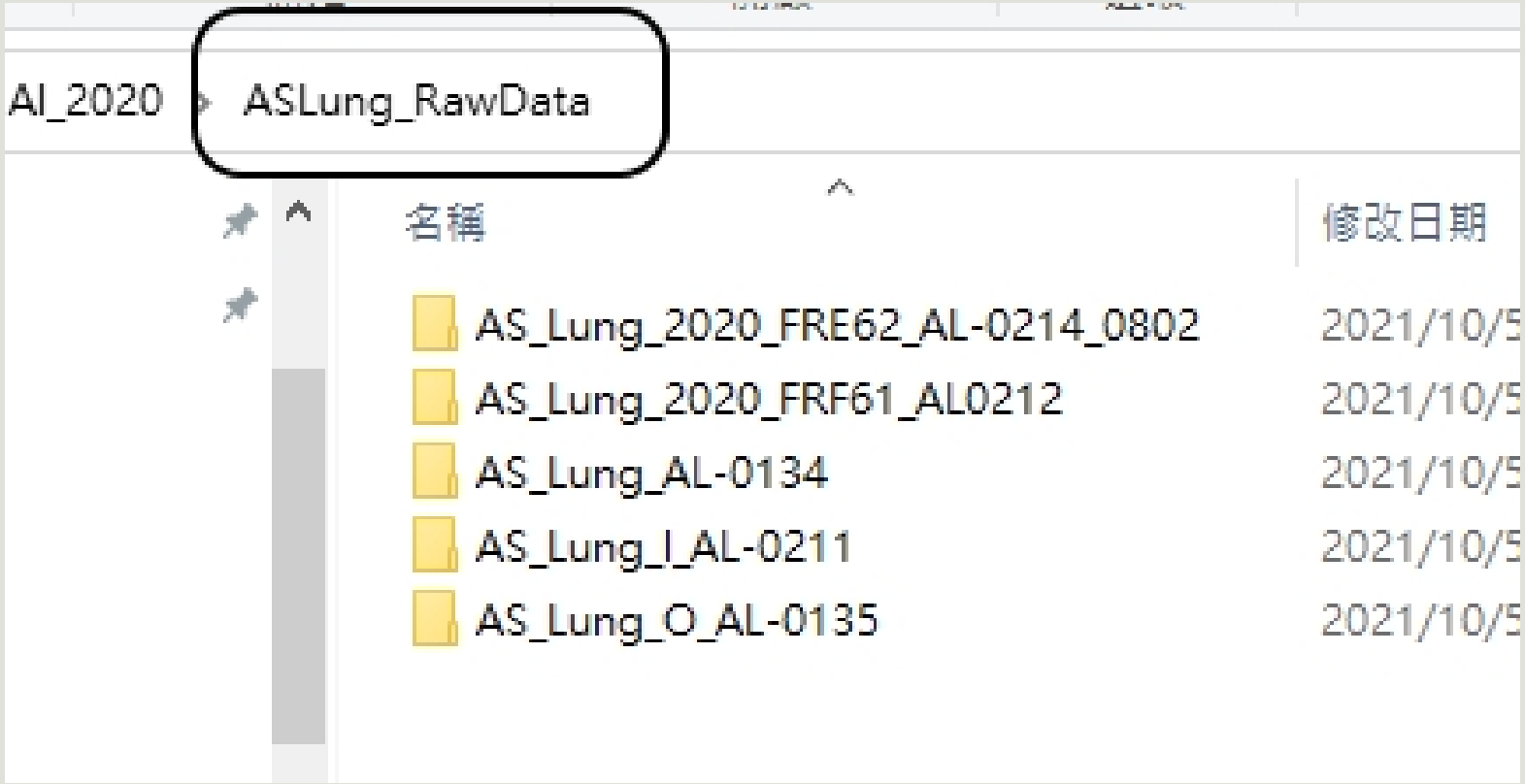
	A	B	C	D	E	F	G	H	I
	Golden_standard	aslung_id	slope1	intercept1	break_point1	slope2	intercept2	r2	PM
	y_goldenstand	AL-0216	0.574	1.053	98.9	0.318	26.358	0.999	PM2.5
	y_goldenstand	AL-0216	0.818	0.583	53.1	0.525	16.14	0.999	PM1

Step 9. Open the CSV file and **you can see the answer of Q1.b**

Do the same steps for dataset 1

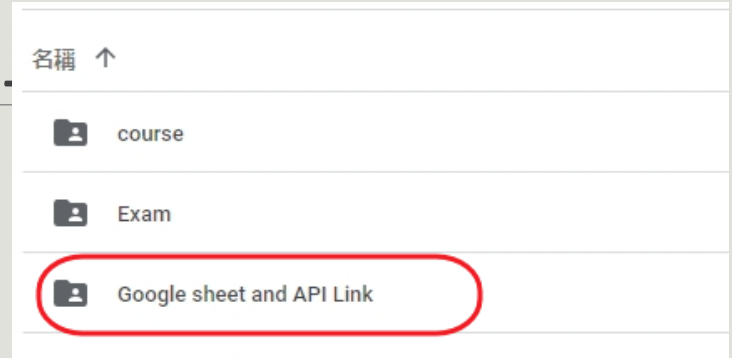
Data cleaning

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is ___



Step 1. copy data to ASLung_RawData

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is _



https://script.google.com/macros/s/AKfycby1uS0h6zIIKvo6E49hhYSz7dbrCdhClhQga3cmzg/exec

B	C	D	E	F	G
Title and Affiliation	Country	First and Middle Name (Given Name)	Last Name (Surname/Family	Google sheet	Calibration factor API
Professor, Department of Chemistry, University of Dhaka	Bangladesh	Abdus	Salam	https://docs.google.com/spreadsheets/d/1d9xsQXHfdPBZ1oUKJoBs5ywPZtr08WjN	https://script.google.com/macros/s/AKfycbw4xvHB_Hilwtz2j8qxi2tc7-okoetr
MS Student, University of Dhaka	Bangladesh	Md Riad Sarkar	Pavel		
Master's student	Bangladesh	Shahid Uz	Zaman		
Assistant Professor, Department of Public Health, Faculty of Medicine, Universitas Padjadjaran	Indonesia	Dwi	Agustian	https://docs.google.com/spreadsheets/d/1Xituc0scPwTBjNv-uBae4V7MEosMjQC9	https://script.google.com/macros/s/AKfycby1uS0h6zIIKvo6E49hhYSz7dbrCdl
Lecturer/Padjadjaran University	Indonesia	Trianing Tyas Kusuma	Anggaeni		
Lecturer at Institute of Technology Bandung	Indonesia	Haryo Satriyo	Tomo		
Dr. Murnira, Institut for Environment and Development, Universiti Kebangsaan Malaysia,	Malaysia	Murnira	Othman	https://docs.google.com/spreadsheets/d/1Wms7CFwjbHbQc5uHpjwVW6zuBvAdor	https://script.google.com/macros/s/AKfycbxo908Pq7fubLcctHnhkM8RpFMe
National University of Mongolia	Mongolia	Ariundelger	Ariunsaikhan	https://docs.google.com/spreadsheets/d/13suC9zyVxWkVShWYsmMpTJYdpEfbFyt3	https://script.google.com/macros/s/AKfycbwjTe66YL2Xa67GPQm4szNQDEv
Atmosphere sience	Mongolia	Batdelger	Byambaa		
National University of Mongolia	Mongolia	Sonomdagva	Chonokhuu		

Step 2. Find the "GoogleSheetAndApi.xlsx" file, you can see the google sheet and API link

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is ___

A	B	C	D	E	F	G	H	I	J	K	
Golden_stand	aslung_id	slope1	intercept1	region1_mae	region1_rmse	break_point1	slope2	intercept2	region2_mae	region2_rmse	r2

Step 3 . Open the google sheet, you will see a new sheet without any calibration factor data.

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is ___

ASLung_2SCF_Taiwan_Pang

檔案 編輯 查看 插入 格式 資料 工具 外掛程式 說明 上次編輯是在數秒前

A2:T19 | y_goldenstand

Golden_stanc	aslung_id	slope1	intercept1	region1_mae	region1_rmse	break_point1	slope2	intercept2	region2_mae	region2_rmse	r2	total_mae	total_rmse	sample	PM	high_conc	low_co
y_goldenstan	AL-0212	0.636	1.309	10.69111	15.59652	101.6	0.35	30.344	162.978	179.2623	0.99893	1.452288	1.971093	2009	PM2.5	200	
y_goldenstan	AL-0213	0.597	0.948	10.85182	15.71953	98.1	0.343	25.929	171.0987	189.8477	0.999	1.253312	1.681789	2079	PM2.5	200	
y_goldenstan	AL-0214	0.594	1.046	11.39977	16.52377	100.2	0.334	27.077	177.6204	196.768	0.999	1.26884	1.7038	2058	PM2.5	200	
y_goldenstan	AL-0215	0.618	1.458	11.86556	17.38534	104.3	0.306	34.005	194.5646	214.7304	0.998846	1.486286	2.046609	2010	PM2.5	200	
y_goldenstan	AL-0216	0.574	1.053	11.84882	17.17266	98.9	0.318	26.358	191.6034	213.0114	0.999	1.11899	1.49103	2070	PM2.5	200	
y_goldenstan	AL-0212	0.944	0.846	0.494929	0.592373	50.2	0.634	16.426	42.48253	50.46678	0.999	1.23408	1.718069	2009	PM1	200	
y_goldenstan	AL-0213	0.851	0.618	1.589544	2.302571	47.7	0.601	12.555	53.85883	63.05177	0.999	1.302255	1.772558	2079	PM1	200	
y_goldenstan	AL-0214	0.87	0.704	1.297618	1.887069	47.7	0.609	13.138	51.13278	60.13584	0.999	1.258186	1.698384	2058	PM1	200	
y_goldenstan	AL-0215	0.944	1.014	0.685963	0.809721	55.4	0.583	21.015	51.90424	61.21701	0.999	1.276783	1.76515	2010	PM1	200	
y_goldenstan	AL-0216	0.818	0.583	2.394982	3.418482	53.1	0.525	16.14	73.77153	85.79146	0.999	1.056191	1.407531	2070	PM1	200	
y_goldenstan	AL-0125	0.60057	3.742313			10000					0.972045			635	PM2.5	150	
y_goldenstan	AL-0128	0.50361	3.591891			10000					0.975369			615	PM2.5	150	
y_goldenstan	AL-0134	0.419875	4.343777			10000					0.974212			824	PM2.5	150	
y_goldenstan	AL-0135	0.446472	4.341272			10000					0.97964			816	PM2.5	150	
y_goldenstan	AL-0125	0.702705	3.69809			10000					0.978766			635	PM1	150	
y_goldenstan	AL-0128	0.616252	3.258302			10000					0.982444			615	PM1	150	
y_goldenstan	AL-0134	0.568007	3.398774			10000					0.983401			824	PM1	150	
y_goldenstan	AL-0135	0.568888	3.493351			10000					0.987019			816	PM1	150	

AI_2020 D:\python_project\AI_2020
- ASLung_Calibrated_Data
- ASLung_Calibration_Factor
  - 2SegmentRegressFactor_202110051102.csv
  - SimpleLinearRegressFactor_202110051028.csv

Step 4. copy the calibration factors to the sheet, which are generated from dataset 1 and dataset 2

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is ___

run.py

```python
import os
import warnings
warnings.filterwarnings("ignore", category=RuntimeWarning)
def dataurls():

    calfactorurl="https://script.google.com/macros/s/AKfycbwdqClAkikVl8eJMcw_ZsN1RS9pfO3NbvDJYgRuRsZrOKpIVCMG/exec"
    return calfactorurl
```

Step 5. open “run.py” and update calibration factor API

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is ___

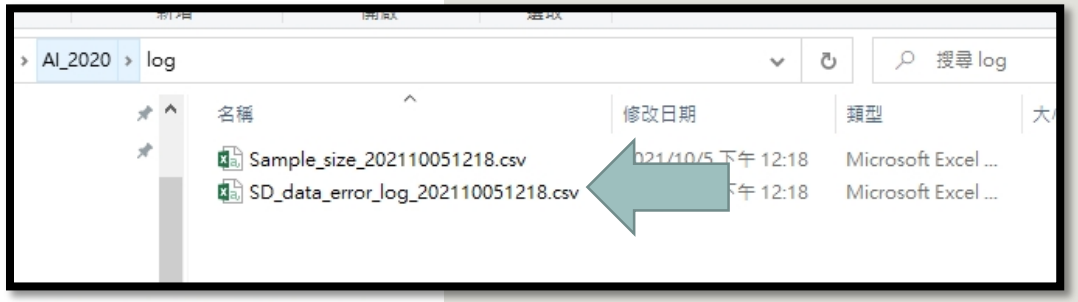
```
un:  run
D:\python_project\AI_2020\venv_p374_x\Scripts\python.exe D:/python_project/AI_2020/run.py
Please choose:
1) ASLung calibration factor from reference instrument
2) Data cleaning and calibrate ASLung raw data
Please select which you want to do: 2
1) Calculate AS-Lung data from SD card
2) Calculate 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 automatically remove all the data in the hour
==========================================================================================
Calculate AS-Lung data from SD card
```

Step 6. run the python code

Step 7. select 2(Data cleaning and calibrate ASLung raw data) then select 1(Calculate AS-Lung data from SD card)

Q1. There is a AS-Lung device do not have calibration factor. The AS-Lung id is _AL-0211_

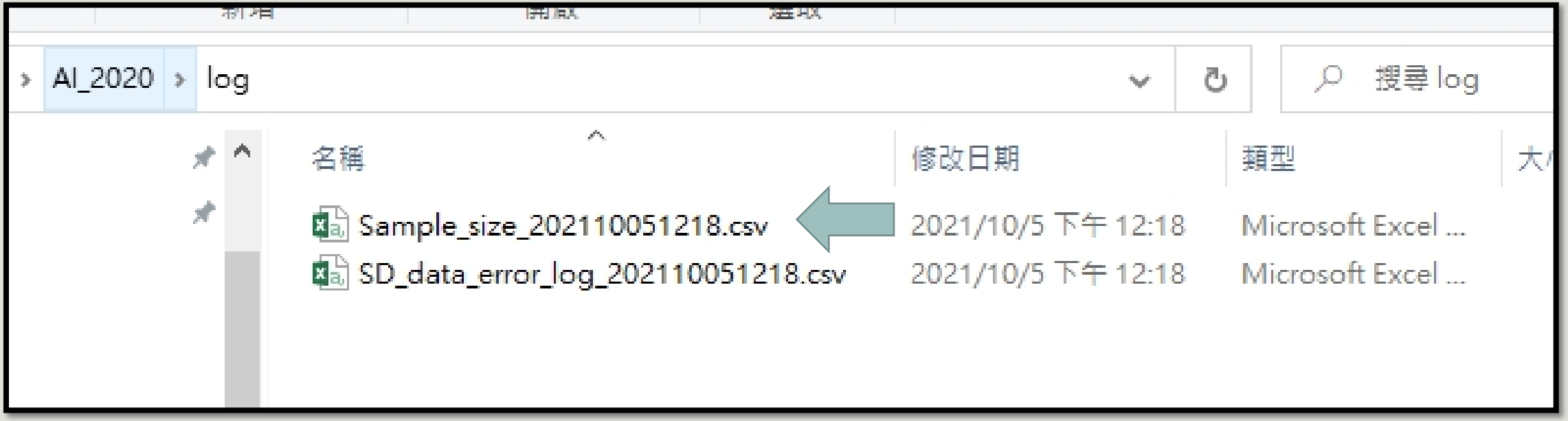
```
Calculate data file of :  2018-08-30.csv
Data file:  D:\python_project\AI_2020\ASLung_RawData\AS_Lung_O_AL-0135\2018-08-31.csv
AS-Lung ID:  AL-0135
Log interval:  60  secs
Lab ID:  AS_Lung_O
Save Folder:  AS_Lung_O_AL-0135
Calculate data file of :  2018-08-31.csv
  aslung_id    DataDate                  Error Message             Folder
0   AL-0211  2019-03-11  There is no calibration factor  AS_Lung_I_AL-0211
1   AL-0211  2019-03-12  There is no calibration factor  AS_Lung_I_AL-0211
2   AL-0211  2019-03-13  There is no calibration factor  AS_Lung_I_AL-0211
End Time:  2021-10-05 12:18:57
==================================================================================================
If there is any error, the log message will save in the folder of log.
Then check the dataset format.
==================================================================================================

Process finished with exit code 0
```



Step 8. when python finish, it will shoe the error message on the window. The answer also save in the log file of SD_data_Error_log file

So, the answer of Q2 is AL-0211

Q2. How many sample size of AS-Lung device at the sampling date of 2018-08-30?



Step 1. open the log file of Sample_size file

Q2. How many sample size of AS-Lung device at the sampling date of 2018-08-30?

	A	B	C	D
1	aslung_id	Data file	Samples	Data recovery
2	AL-0214	2020-08-03.csv	5712	99.2
3	AL-0214	2020-08-04.csv	2175	37.8
4	AL-0214	2020-08-05.csv	5760	100
5	AL-0212	2020-08-14.csv	5760	100
6	AL-0212	2020-08-15.csv	5760	100
7	AL-0212	2020-08-16.csv	5760	100
8	AL-0212	2020-08-17.csv	5760	100
9	AL-0134	2018-09-05.csv	1408	97.8
10	AL-0134	2018-09-06.csv	1413	98.1
11	AL-0134	2018-09-07.csv	1408	97.8
12	AL-0135	2018-08-29.csv	1386	96.2
13	AL-0135	2018-08-30.csv	1374	95.4
14	AL-0135	2018-08-31.csv	1388	96.4
15				

Step 2. The answer of Q2 is 1374

Q3. Which of the sampling date and AS-Lung ID is less than 80% data recovery?

	A	B	C	D
1	aslung_id	Data file	Samples	Data recovery
2	AL-0214	2020-08-03.csv	5712	99.2
3	AL-0214	2020-08-04.csv	2175	37.8
4	AL-0214	2020-08-05.csv	5760	100
5	AL-0212	2020-08-14.csv	5760	100
6	AL-0212	2020-08-15.csv	5760	100
7	AL-0212	2020-08-16.csv	5760	100
8	AL-0212	2020-08-17.csv	5760	100
9	AL-0134	2018-09-05.csv	1408	97.8
10	AL-0134	2018-09-06.csv	1413	98.1
11	AL-0134	2018-09-07.csv	1408	97.8
12	AL-0135	2018-08-29.csv	1386	96.2
13	AL-0135	2018-08-30.csv	1374	95.4
14	AL-0135	2018-08-31.csv	1388	96.4
15				

Step 1. open the log file of Sample_size file

Step 2. The answer is 2020-08-04 and AL-0214

Thank you for your attention!

Any question and comment are welcome

Chun-Hu Liu

SC Candice Lung