In [1]:
```
import pandas as pd
import numpy as np
```

1&2處理airbox資料

In [2]:
```
# PM2.5資料

# tw1001 = pd.read_csv('iis_airbox_20201001.csv')
# tw1002 = pd.read_csv('iis_airbox_20201002.csv')
# tw1003 = pd.read_csv('iis_airbox_20201003.csv')
# tw1004 = pd.read_csv('iis_airbox_20201004.csv')
# tw1005 = pd.read_csv('iis_airbox_20201005.csv')

# 如果檔案很多，且檔名雷同，可參考以下方式快速讀檔
for i in range(1,6):
    locals()['tw100{}'.format(i)] = pd.read_csv('iis_airbox_2020100{}.csv'.format(i))
# 另外也可以import os，使用os.listdir，因下次作業有可能需要使用，先請同學自行研究。
```

In [3]:
```
# 取得高雄市PM2.5資料
ks1001 = tw1001[tw1001.SiteName.str.contains('高雄市')]
ks1002 = tw1002[tw1002.SiteName.str.contains('高雄市')]
ks1003 = tw1003[tw1003.SiteName.str.contains('高雄市')]
ks1004 = tw1004[tw1004.SiteName.str.contains('高雄市')]
ks1005 = tw1005[tw1005.SiteName.str.contains('高雄市')]
```

In [4]:
```
ks1001
```

Out[4]:

	device_id	SiteName	PM25	timestamp
3	74DA38F20DD0	高雄市市立樂群國小(2018)	4	2020-10-01 00:00:00
6	74DA38F20B0C	高雄市市立民生國小(2018)	44	2020-10-01 00:00:02
10	74DA38F20F06	高雄市市立佛公國小(2018)	4	2020-10-01 00:00:02
15	74DA38F20F26	高雄市市立華山國小(2018)	6	2020-10-01 00:00:03
24	74DA38F20D6C	高雄市市立明華國中(2018)	29	2020-10-01 00:00:05
...	...	...	...	...
367942	74DA38F20F1C	高雄市市立福安國小(2018)	28	2020-10-01 23:59:45
367973	74DA38F20FEA	高雄市市立大社國小(2018)	36	2020-10-01 23:59:53
367975	74DA38F20A12	高雄市市立油廠國小(2018)	26	2020-10-01 23:59:53
367981	74DA38F20B82	高雄市市立溪洲國小(2018)	26	2020-10-01 23:59:54
367987	74DA38F2080C	高雄市市立新庄國小(2018)	20	2020-10-01 23:59:55

41749 rows × 4 columns

使用pd.concat去串聯5個資料

In [5]:
```
# pd.concat
ks1001_5 = pd.concat([ks1001, ks1002, ks1003, ks1004, ks1005], join='outer')
ks1001_5   #高雄市要有166598筆資料
```

Out[5]:

	device_id	SiteName	PM25	timestamp
3	74DA38F20DD0	高雄市市立樂群國小(2018)	4	2020-10-01 00:00:00
6	74DA38F20B0C	高雄市市立民生國小(2018)	44	2020-10-01 00:00:02
10	74DA38F20F06	高雄市市立佛公國小(2018)	4	2020-10-01 00:00:02
15	74DA38F20F26	高雄市市立華山國小(2018)	6	2020-10-01 00:00:03
24	74DA38F20D6C	高雄市市立明華國中(2018)	29	2020-10-01 00:00:05
...	...	...	...	...
358378	74DA38F20B76	高雄市市立橫山國小(2018)	37	2020-10-05 23:59:55
358381	74DA38F20B7A	高雄市市立梓官國小(2018)	25	2020-10-05 23:59:56
358382	74DA38F20FC0	高雄市市立大樹國小(2018)	38	2020-10-05 23:59:56
358391	74DA38F20DFE	高雄市市立梓官國中(2018)	23	2020-10-05 23:59:57
358392	74DA38F2114C	高雄市市立大義國中(2018)	21	2020-10-05 23:59:57

166598 rows × 4 columns

In [6]:
```
# 高雄各站pm2.5資料(2018以前，篩掉2019的資料)
ks1001_5_2018 = ks1001_5[ks1001_5.SiteName.str.contains('2019')==False]
ks1001_5_2018
# 高雄市2018要有166598筆資料，與篩選前一樣數量(因為原資料只包含2018的資料)
# (評分方式只看是否有篩選的動作)
```

Out[6]:

	device_id	SiteName	PM25	timestamp
3	74DA38F20DD0	高雄市市立樂群國小(2018)	4	2020-10-01 00:00:00
6	74DA38F20B0C	高雄市市立民生國小(2018)	44	2020-10-01 00:00:02
10	74DA38F20F06	高雄市市立佛公國小(2018)	4	2020-10-01 00:00:02
15	74DA38F20F26	高雄市市立華山國小(2018)	6	2020-10-01 00:00:03
24	74DA38F20D6C	高雄市市立明華國中(2018)	29	2020-10-01 00:00:05
...	...	...	...	...
358378	74DA38F20B76	高雄市市立橫山國小(2018)	37	2020-10-05 23:59:55
358381	74DA38F20B7A	高雄市市立梓官國小(2018)	25	2020-10-05 23:59:56
358382	74DA38F20FC0	高雄市市立大樹國小(2018)	38	2020-10-05 23:59:56
358391	74DA38F20DFE	高雄市市立梓官國中(2018)	23	2020-10-05 23:59:57

1&2處理測站資料

In [7]:
```
#測站資料
station = pd.read_csv('iis_airbox_station.csv')
```

In [8]:
```
#高雄市測站資料
station_ks = station[station.station_name.str.contains('高雄市')]
station_ks
#高雄市要有336個測站
```

Out[8]:

	station_id	station_name	station_address	lat	lon	source	create_year
2	74DA38F20A0C	高雄市市立七賢國小	高雄市七賢國小 高雄市新興區七賢一路393號	22.634	120.306	Kaohsiung18	2018
4	74DA38F20EC8	高雄市市立紅毛港國小	高雄市紅毛港國小 高雄市前鎮區明鳳7街1號	22.583	120.334	Kaohsiung18	2018
9	74DA38F21146	高雄市仁武特殊教育學校	仁武特殊教育學校	22.683	120.346	Kaohsiung18	2018
11	74DA38F20B12	高雄市市立多納國小	高雄市多納國小 高雄市茂林區多納里1-2號	22.910	120.716	Kaohsiung18	2018
13	74DA38F20F1A	高雄市市立建國國小	高雄市建國國小 高雄市前金區自立二路111號	22.635	120.298	Kaohsiung18	2018
...	...	...	...	...	...	...	...
3095	74DA38F20F6E	高雄市市立甲仙國小	高雄市甲仙國小 高雄市甲仙區西安里文化路45號	23.081	120.589	Kaohsiung18	2018
3096	74DA38F20F78	高雄市市立木柵國小	高雄市木柵國小 高雄市內門區木柵里木柵3號	22.975	120.467	Kaohsiung18	2018
3098	74DA38F20FC6	高雄市市立忠孝國小	高雄市忠孝國小 高雄市鳳山區新富路630號	22.617	120.349	Kaohsiung18	2018
3099	74DA38F20FC8	高雄市市立鼎金國小	高雄市鼎金國小 高雄市三民區鼎山街375號	22.656	120.321	Kaohsiung18	2018
3103	74DA38F20FE4	高雄市市立崇德國小	高雄市崇德國小 高雄市田寮區崇德里崇德路101號	22.877	120.379	Kaohsiung18	2018

336 rows × 7 columns

In [9]:
```
#高雄市測站位置資料(2018以前)
station_ks_2018 = station_ks[station_ks.create_year<=2018]
station_ks_2018
#同樣的，測站資料只有2018以前的資料，評分看是否篩選。
```

Out[9]:

	station_id	station_name	station_address	lat	lon	source	create_year
2	74DA38F20A0C	高雄市市立七賢國小	高雄市七賢國小 高雄市新興區七賢一路393號	22.634	120.306	Kaohsiung18	2018
4	74DA38F20EC8	高雄市市立紅毛港國小	高雄市紅毛港國小 高雄市前鎮區明鳳7街1號	22.583	120.334	Kaohsiung18	2018
9	74DA38F21146	高雄市仁武特殊教育學校	仁武特殊教育學校	22.683	120.346	Kaohsiung18	2018
11	74DA38F20B12	高雄市市立多納國小	高雄市多納國小 高雄市茂林區多納里1-2號	22.910	120.716	Kaohsiung18	2018
13	74DA38F20F1A	高雄市市立建國國小	高雄市建國國小 高雄市前金區自立二路111號	22.635	120.298	Kaohsiung18	2018
...	...	...	...	...	...	...	...
3095	74DA38F20F6E	高雄市市立甲仙國小	高雄市甲仙國小 高雄市甲仙區西安里文化路45號	23.081	120.589	Kaohsiung18	2018
3096	74DA38F20F78	高雄市市立木柵國小	高雄市木柵國小 高雄市內門區木柵里木柵3號	22.975	120.467	Kaohsiung18	2018
3098	74DA38F20FC6	高雄市市立忠孝國小	高雄市忠孝國小 高雄市鳳山區新富路630號	22.617	120.349	Kaohsiung18	2018
3099	74DA38F20FC8	高雄市市立鼎金國小	高雄市鼎金國小 高雄市三民區鼎山街375號	22.656	120.321	Kaohsiung18	2018
3103	74DA38F20FE4	高雄市市立崇德國小	高雄市崇德國小 高雄市田寮區崇德里崇德路101號	22.877	120.379	Kaohsiung18	2018

336 rows × 7 columns

3&4合併資料(加入經緯度)

In [10]:
```
# 對測站資料的 'station_id' 改名稱為 'device_id'
# 改名稱方便使用pd.merge
station_ks_2018 = station_ks_2018.rename(columns={'station_id':'device_id'})
#station_ks_2018 = station_ks_2018.rename(columns={'station_id':'ks1001_5_2018.columns[0]'})
station_ks_2018.columns
```

Out[10]:
```
Index(['device_id', 'station_name', 'station_address', 'lat', 'lon', 'source',
       'create_year'],
      dtype='object')
```

In [11]:
```
# 只挑 ['device_id','lat','lon'] 為想要 merge的 Dataframe
station_ks_2018 = station_ks_2018.loc[:,['device_id','lat','lon']]
station_ks_2018
```

Out[11]:

	device_id	lat	lon
2	74DA38F20A0C	22.634	120.306
4	74DA38F20EC8	22.583	120.334
9	74DA38F21146	22.683	120.346
11	74DA38F20B12	22.910	120.716
13	74DA38F20F1A	22.635	120.298
...	...	...	...
3095	74DA38F20F6E	23.081	120.589
3096	74DA38F20F78	22.975	120.467
3098	74DA38F20FC6	22.617	120.349
3099	74DA38F20FC8	22.656	120.321
3103	74DA38F20FE4	22.877	120.379

336 rows × 3 columns

In [13]:
```
df = pd.merge(ks1001_5_2018,station_ks_2018, on='device_id', how = 'left')
```

輸出json檔與feather檔並比較檔案大小

In [14]:
```
%timeit df.to_json(path_or_buf = 'df.json')
%timeit df.to_feather(path = 'df.feather')
```

```
344 ms ± 9.26 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
92.7 ms ± 1.52 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

比較檔案大小(法1)

In [15]:
```
import os
print('\'json檔\'的文件大小為',os.path.getsize('df.json'),'位元組')
print('\'feather檔\'的文件大小為',os.path.getsize('df.feather'),'位元組')
```

```
'json檔'的文件大小為 28006396 位元組
'feather檔'的文件大小為 5339994 位元組
```

比較檔案大小(法2)

In [16]:
```
import os
print('\'json檔\'的文件大小為',os.stat('df.json').st_size,'位元組')
print('\'json檔\'的文件大小為',os.stat('df.feather').st_size,'位元組')
```

```
'json檔'的文件大小為 28006396 位元組
'json檔'的文件大小為 5339994 位元組
```

In [ ]: