使用Python進行資料整理 – Pandas初探

In [1]:

```
# 載入函式庫
import pandas as pd
pd.__version__ #查看版本
```

Out[1]:

```
'0.23.0'
```

1.用read_csv()來讀取csv檔

In [2]:

```
# 讀取csv檔案
df = pd.read_csv("C:/ArkEase Pro/py_pandas/data/SRDA20181107.csv") #請自行依檔案位置調整
df
```

Out[2]:

	id	sex	age	a1	a2	a3	a4	ka4
0	1001	1.0	43	2	1	1	5	打零工
1	1002	2.0	46	3	1	2	3	NaN
2	1003	1.0	28	1	2	3	5	NaN
3	1004	NaN	25	4	2	4	2	NaN
4	1005	2.0	77	2	9	5	1	幫家裡做家庭代工
5	1006	2.0	55	3	3	6	3	NaN
6	1006	3.0	60	2	3	7	5	育嬰假
7	1008	NaN	18	1	4	8	4	NaN
8	1009	1.0	15	2	4	9	4	NaN
9	1010	2.0	65	2	9	10	3	NaN

2.用duplicated()來檢查重複值

In [3]:

```
# 樣本編號檢查
cd1 = df.duplicated('id')
df.id[cd1] #僅輸出id變項
```

Out[3]:

```
6    1006
Name: id, dtype: int64
```

In [4]:

```
# 亦可整合成一列程式碼，差別在於有沒有把條件存成物件。
df.id[df.duplicated('id')]
```

Out[4]:

```
6    1006
Name: id, dtype: int64
```

3.用isna()來檢查遺漏值

In [5]:

```
# 檢查遺漏值/缺失值(missing value/ data)
cd2 = df['sex'].isna()
df.sex[cd2] #僅輸出sex變項
```

Out[5]:

```
3   NaN
7   NaN
Name: sex, dtype: float64
```

4.用isin()來檢查指定值

(1)檢查特定樣本編號

In [6]:

```
# 檢查特定樣本編號
cd3 = df['id'].isin([1006])
df[cd3] #輸出整筆資料
```

Out[6]:

	id	sex	age	a1	a2	a3	a4	ka4
5	1006	2.0	55	3	3	6	3	NaN
6	1006	3.0	60	2	3	7	5	育嬰假

In [7]:

```
# 依檢查需求，也可僅列出id變項。
df.id[cd3] #僅輸出id變項
```

Out[7]:

```
5    1006
6    1006
Name: id, dtype: int64
```

In [8]:

```
# 亦可整合成一列程式碼
df[df['id'].isin([1006])]
```

Out[8]:

	id	sex	age	a1	a2	a3	a4	ka4
5	1006	2.0	55	3	3	6	3	NaN
6	1006	3.0	60	2	3	7	5	育嬰假

In [9]:

```
# 或是結合其他函式，如duplicated()
df[df['id'].isin([df.id[df.duplicated('id')]])]
```

Out[9]:

	id	sex	age	a1	a2	a3	a4	ka4
5	1006	2.0	55	3	3	6	3	NaN
6	1006	3.0	60	2	3	7	5	育嬰假

(2)檢查類別變項

In [10]:

```
# 檢查類別變項
cd4 = ~df['sex'].isin(['1', '2'])
df[cd4] #輸出整筆資料
```

Out[10]:

	id	sex	age	a1	a2	a3	a4	ka4
3	1004	NaN	25	4	2	4	2	NaN
6	1006	3.0	60	2	3	7	5	育嬰假
7	1008	NaN	18	1	4	8	4	NaN

In [11]:

```
# 亦可一列程式碼表示
df[~df['sex'].isin(['1', '2'])]
```

Out[11]:

	id	sex	age	a1	a2	a3	a4	ka4
3	1004	NaN	25	4	2	4	2	NaN
6	1006	3.0	60	2	3	7	5	育嬰假
7	1008	NaN	18	1	4	8	4	NaN

(3)題目間關聯性

In [12]:

```
# 檢查跳續答
# 依照問卷設計a1變項回答(4)無法選擇，a2變項要跳答
# 條件5：當a1變項回答4，a2變項卻不是9（跳答碼）；而當a1變項不是回答4，a2變項卻回答9。
cd5 = ((df['a1'].isin(['4']) & ~df['a2'].isin(['9'])) |
       (~df['a1'].isin(['4']) & df['a2'].isin(['9'])))
df[cd5]
```

Out[12]:

	id	sex	age	a1	a2	a3	a4	ka4
3	1004	NaN	25	4	2	4	2	NaN
4	1005	2.0	77	2	9	5	1	幫家裡做家庭代工
9	1010	2.0	65	2	9	10	3	NaN

In [13]:

```
# 檢查開放題
# 依照問卷設計a4變項回答(5)其他，ka4變項要鍵入文字。
# 條件6:當a4變項回答5，ka4變項卻無資料(NaN)；而當a4變項不是回答5，ka4變項有鍵入文字。
cd6 = (((df['a4'] == 5) & (df['ka4'].isna())) |
       ((df['a4'].isin([1, 2, 3, 4])) & (df['ka4'].notna())))
df[['id', 'a4', 'ka4']][cd6] #輸出id、a4和ka4變項
```

Out[13]:

	id	a4	ka4
2	1003	5	NaN
4	1005	1	幫家裡做家庭代工

5.用between()來檢查區間值

In [14]:

```
# 檢查連續變項
cd7 = ~df['age'].between(18, 65)
df.age[cd7] #輸出age變項
```

Out[14]:

```
4    77
8    15
Name: age, dtype: int64
```

6.用loc[]修改資料

(1)利用索引值（index）

In [15]:

```
# 利用索引值修改資料
df.loc[6, 'id'] = 1007
df.id #輸出id變項
```

Out[15]:

```
0    1001
1    1002
2    1003
3    1004
4    1005
5    1006
6    1007
7    1008
8    1009
9    1010
Name: id, dtype: int64
```

(2)利用樣本編號

In [16]:

```
# 利用樣本編號修改資料
df.loc[df['id']==1007, 'sex'] = 2
df[['id', 'sex']] #輸出id和sex變項
```

Out[16]:

	id	sex
0	1001	1.0
1	1002	2.0
2	1003	1.0
3	1004	NaN
4	1005	2.0
5	1006	2.0
6	1007	2.0
7	1008	NaN
8	1009	1.0
9	1010	2.0

7.用**rename()**來更改變項名稱

In [17]:

```
# 更改變項名稱
df = df.rename(columns={'ka4':'ka4_o'})
df.columns #輸出columns
```

Out[17]:

```
Index(['id', 'sex', 'age', 'a1', 'a2', 'a3', 'a4', 'ka4_o'], dtype='obje
ct')
```

8.用**replace()**重新編碼

(1)同一變項

In [18]:

```
# 重新編碼：同一變項
df['a2'].replace({1:1, 2:1, 3:2, 4:2, 9:9}, inplace=True) #inplace=True才會寫入
df
```

Out[18]:

	id	sex	age	a1	a2	a3	a4	ka4_o
0	1001	1.0	43	2	1	1	5	打零工
1	1002	2.0	46	3	1	2	3	NaN
2	1003	1.0	28	1	1	3	5	NaN
3	1004	NaN	25	4	1	4	2	NaN
4	1005	2.0	77	2	9	5	1	幫家裡做家庭代工
5	1006	2.0	55	3	2	6	3	NaN
6	1007	2.0	60	2	2	7	5	育嬰假
7	1008	NaN	18	1	2	8	4	NaN
8	1009	1.0	15	2	2	9	4	NaN
9	1010	2.0	65	2	9	10	3	NaN

(2)新增變項

In [19]:

```
# 重新編碼：新增變項
df['new_a3'] = df['a3'].replace({1:1, 2:1, 3:1, 4:1, 5:1, 6:2, 7:2, 8:2, 9:2, 10:2})
df
```

Out[19]:

	id	sex	age	a1	a2	a3	a4	ka4_o	new_a3
0	1001	1.0	43	2	1	1	5	打零工	1
1	1002	2.0	46	3	1	2	3	NaN	1
2	1003	1.0	28	1	1	3	5	NaN	1
3	1004	NaN	25	4	1	4	2	NaN	1
4	1005	2.0	77	2	9	5	1	幫家裡做家庭代工	1
5	1006	2.0	55	3	2	6	3	NaN	2
6	1007	2.0	60	2	2	7	5	育嬰假	2
7	1008	NaN	18	1	2	8	4	NaN	2
8	1009	1.0	15	2	2	9	4	NaN	2
9	1010	2.0	65	2	9	10	3	NaN	2