Chapter 5 - exercise 2: Merge dữ liệu

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

Cung cấp các dictionary sau:

In [2]:
```
raw_data_1 = {
        'subject_id': ['1', '2', '3', '4', '5'],
        'first_name': ['Alex', 'Amy', 'Allen', 'Alice', 'Ayoung'],
        'last_name': ['Anderson', 'Ackerman', 'Ali', 'Aoni', 'Atiches']}

raw_data_2 = {
        'subject_id': ['4', '5', '6', '7', '8'],
        'first_name': ['Billy', 'Brian', 'Bran', 'Bryce', 'Betty'],
        'last_name': ['Bonder', 'Black', 'Balwner', 'Brice', 'Btisan']}

raw_data_3 = {
        'subject_id': ['1', '2', '3', '4', '5', '7', '8', '9', '10', '11'],
        'test_id': [51, 15, 15, 61, 16, 14, 15, 1, 61, 16]}
```

Yêu cầu:

1. Tạo data1, data2, data3 từ 3 dictionary trên. In data1, data2, data3
2. Tạo all_data bằng cách gộp data1 và data2 theo dòng. In all_data
3. Tạo all_data_col bằng cách gộp data1 và data2 theo cột. In all_data_col.
4. Tạo all_data_3 bằng cách gộp all_data và data3, với cột chung là 'subject_id'. In all_data_3
5. In thông tin của all_data_3
6. In thống kê chung của all_data_3

In [3]:
```
# Câu 1: Tạo data1, data2, data3 từ 3 dictionary trên. In data1, data2, data3
data1 = pd.DataFrame(raw_data_1, columns = ['subject_id', 'first_name', 'last_name
data2 = pd.DataFrame(raw_data_2, columns = ['subject_id', 'first_name', 'last_name
data3 = pd.DataFrame(raw_data_3, columns = ['subject_id','test_id'])
```

In [4]:
```
data1
```

Out[4]:

	subject_id	first_name	last_name
0	1	Alex	Anderson
1	2	Amy	Ackerman
2	3	Allen	Ali
3	4	Alice	Aoni
4	5	Ayoung	Atiches

In [5]:
```
data2
```

Out[5]:

	subject_id	first_name	last_name
0	4	Billy	Bonder
1	5	Brian	Black
2	6	Bran	Balwner
3	7	Bryce	Brice
4	8	Betty	Btisan

In [6]:
```
data3
```

Out[6]:

	subject_id	test_id
0	1	51
1	2	15
2	3	15
3	4	61
4	5	16
5	7	14
6	8	15
7	9	1
8	10	61
9	11	16

In [7]:
```
# Câu 2: Tạo all_data bằng cách gộp data1 và data2 theo dòng. In all_data
all_data = pd.concat([data1, data2])
all_data
```

Out[7]:

	subject_id	first_name	last_name
0	1	Alex	Anderson
1	2	Amy	Ackerman
2	3	Allen	Ali
3	4	Alice	Aoni
4	5	Ayoung	Atiches
0	4	Billy	Bonder
1	5	Brian	Black
2	6	Bran	Balwner
3	7	Bryce	Brice
4	8	Betty	Btisan

In [8]:
```
# Câu 3: Tạo all_data_col bằng cách gộp data1 và data2 theo cột. In all_data_col.
all_data_col = pd.concat([data1, data2], axis = 1)
all_data_col
```

Out[8]:

	subject_id	first_name	last_name	subject_id	first_name	last_name
0	1	Alex	Anderson	4	Billy	Bonder
1	2	Amy	Ackerman	5	Brian	Black
2	3	Allen	Ali	6	Bran	Balwner
3	4	Alice	Aoni	7	Bryce	Brice
4	5	Ayoung	Atiches	8	Betty	Btisan

In [9]:
```
# Câu 4: Tạo all_data_3 bằng cách gộp all_data và data3, với cột chung là 'subject
all_data_3 = pd.merge(all_data, data3, on='subject_id')
all_data_3
```

Out[9]:

	subject_id	first_name	last_name	test_id
0	1	Alex	Anderson	51
1	2	Amy	Ackerman	15
2	3	Allen	Ali	15
3	4	Alice	Aoni	61
4	4	Billy	Bonder	61
5	5	Ayoung	Atiches	16
6	5	Brian	Black	16
7	7	Bryce	Brice	14
8	8	Betty	Btisan	15

In [10]:
```
# câu 5: In thông tin của all_data_3
all_data_3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9 entries, 0 to 8
Data columns (total 4 columns):
subject_id    9 non-null object
first_name    9 non-null object
last_name     9 non-null object
test_id       9 non-null int64
dtypes: int64(1), object(3)
memory usage: 360.0+ bytes
```

In [11]:
```
# Câu 6: In thống kê chung của all_data_3
all_data_3.describe()
```

Out[11]:

	test_id
count	9.000000
mean	29.333333
std	21.453438
min	14.000000
25%	15.000000
50%	16.000000
75%	51.000000
max	61.000000

In [ ]:
```

```