Fictitious Names

Introduction:

This time you will create a data again

Special thanks to Chris Albon for sharing the dataset and materials. All the credits to this exercise belongs to him.

In order to understand about it go here.

Step 1. Import the necessary libraries

```
import pandas as pd
```

Step 2. Create the 3 DataFrames based on the following raw data

+ Mã + Văn bản

```
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]}
```

Step 3. Assign each to a variable called data1, data2, data3

```
data1 = pd.DataFrame(raw_data_1)
data2 = pd.DataFrame(raw_data_2)
data3 = pd.DataFrame(raw_data_3)
print("data1:\n", data1)
print("data2:\n", data2)
print("data3:\n", data3)
```

```
data1:
   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
data2:
   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
data3:
   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
```

Step 4. Join the two dataframes along rows and assign all_data

```
all_data = pd.concat([data1, data2], axis=0, ignore_index=True)
print("all_data:\n", all_data)
```

```
all_data:
   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
5           4      Billy    Bonder
6           5      Brian     Black
7           6       Bran   Balwner
8           7      Bryce     Brice
9           8      Betty    Btisan
```

Step 5. Join the two dataframes along columns and assing to all_data_col

```
all_data_col = pd.concat([data1, data2], axis=1)
print("all_data_col:\n", all_data_col)
```

```
all_data_col:
   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
```

Step 6. Print data3

```
print("data3:\n", data3)
```

```
data3:
   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
```

Step 7. Merge all_data and data3 along the subject_id value

```
merged_data = pd.merge(all_data, data3, on='subject_id', how='inner')
print("merged_data (step 7):\n", merged_data)
```

```
merged_data (step 7):
   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           5     Ayoung   Atiches       16
5           4      Billy    Bonder       61
6           5      Brian     Black       16
7           7      Bryce     Brice       14
8           8      Betty    Btisan       15
```

Step 8. Merge only the data that has the same 'subject_id' on both data1 and data2

```
inner_merge = pd.merge(data1, data2, on='subject_id', how='inner')
print("inner_merge (step 8):\n", inner_merge)
```

```
inner_merge (step 8):
   subject_id first_name_x last_name_x first_name_y last_name_y
0           4        Alice        Aoni        Billy      Bonder
1           5       Ayoung     Atiches        Brian       Black
```

Step 9. Merge all values in data1 and data2, with matching records from both sides where available.

```
outer_merge = pd.merge(data1, data2, on='subject_id', how='outer')
print("outer_merge (step 9):\n", outer_merge)
```