Khai báo các thư viện:

- pandas: Làm việc với cấu trúc dữ liệu, hỗ trợ đọc ghi file dữ liệu với nhiều định dạng khác nhau như csv, xls, sql, html...
- matplotlib: Thư viện hỗ trợ vẽ các đồ thị trong Python
- Numpy cung cấp các đối tượng và phương thức để làm việc với mảng nhiều chiều và các phép toán đại số tuyến tính.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

- Đọc file dữ liệu "bank-data-1.1-6thuoctinh.csv" gồm 300 mẫu, 6 thuộc tính : id, age, sex, region, income, married. Dữ liệu bị thiếu tại thuộc tính sex, income. Tuổi có giá trị ngoại lai 200.
- Đọc file dữ liệu "bank-data-1.2-7thuoctinh.csv" gồm 300 mẫu, 7 thuộc tính : id, children, car, save_act, current_act, mortgage,pep.
- Đọc file dữ liệu "bank-data-2-12thuoctinh.csv" gồm 300 mẫu, 12 thuộc tính : id, age, sex, region, income, married, children, car, save_act, current_act, mortgage, pep. Dữ liệu cột sex là NAM và NU

```
df1_1 = pd.read_csv("bank-data-1.1-6thuoctinh.csv")
df1_2 = pd.read_csv("bank-data-1.2-7thuoctinh.csv")
df2 = pd.read_csv("bank-data-2-12thuoctinh.csv")
```

```
print("Kích thước tập dữ liệu 1.1 ",df1_1.shape)
print("Kích thước tập dữ liệu 1.2 ",df1_2.shape)
print("Kích thước tập dữ liệu 2 ",df2.shape)
```

```
Kích thước tập dữ liệu 1.1  (300, 6)
Kích thước tập dữ liệu 1.2  (300, 7)
Kích thước tập dữ liệu 2  (300, 12)
```

Đọc 10 dòng dữ liệu đầu tiên của df1_1

- Dữ liệu khuyết thiếu có giá trị NaN

```
df1_1.head(10)
```

	ID	age	sex	region	income	married
0	1	48	FEMALE	INNER_CITY	17546.0	NO
1	2	40	MALE	TOWN	30085.1	YES
2	3	51	FEMALE	INNER_CITY	16575.4	YES
3	4	23	FEMALE	TOWN	20375.4	YES
4	5	57	FEMALE	RURAL	50576.3	YES
5	6	57	FEMALE	TOWN	37869.6	YES
6	7	22	NaN	RURAL	NaN	NO
7	8	58	MALE	TOWN	24946.6	YES
8	9	37	FEMALE	SUBURBAN	25304.3	YES
9	10	200	MALE	TOWN	24212.1	YES

Kiểm tra giá trị thiếu

Để phát hiện các giá trị khuyết thiếu Pandas cung cấp các hàm isnull(), notnull(), value_counts, count

```
df1_1Na= df1_1.count()
print(df1_1Na)

df1_1Na= df1_1['sex'].value_counts(dropna=False)
print(df1_1Na)

df1_1Na= df1_1['sex'].count()
print(df1_1Na)

df1_1Na= df1_1['sex'].isnull()
```

```
print(df1_1Na)

df1_1Na=df1_1.isnull()
print(df1_1Na)
```

```
ID         300
age        300
sex        297
region     300
income     298
married    300
dtype: int64
MALE      154
FEMALE    143
NaN         3
Name: sex, dtype: int64
297
0      False
1      False
2      False
3      False
4      False
       ...
295    False
296    False
297    False
298    False
299    False
Name: sex, Length: 300, dtype: bool
        ID    age    sex  region  income  married
0    False  False  False   False   False    False
1    False  False  False   False   False    False
2    False  False  False   False   False    False
3    False  False  False   False   False    False
4    False  False  False   False   False    False
..     ...    ...    ...     ...     ...      ...
295  False  False  False   False   False    False
296  False  False  False   False   False    False
297  False  False  False   False   False    False
298  False  False  False   False   False    False
299  False  False  False   False   False    False
```

```
[300 rows x 6 columns]
```

Xóa dòng dữ liệu khuyết thiếu dropna()

```
drop_df1_1 = df1_1.dropna()
print(drop_df1_1)
```

```
      ID  age     sex      region    income married
0      1   48  FEMALE  INNER_CITY  17546.00      NO
1      2   40    MALE        TOWN  30085.10     YES
2      3   51  FEMALE  INNER_CITY  16575.40     YES
3      4   23  FEMALE        TOWN  20375.40     YES
4      5   57  FEMALE       RURAL  50576.30     YES
..   ...  ...     ...         ...       ...     ...
295  296   34    MALE        TOWN  32548.90     YES
296  297   54  FEMALE       RURAL  24583.40      NO
297  298   18    MALE       RURAL   8639.24     YES
298  299   47  FEMALE  INNER_CITY  17139.50      NO
299  300   24  FEMALE  INNER_CITY  13667.70     YES

[296 rows x 6 columns]
```

Điền thay thế giá trị khuyết thiếu fillna()

- Thay thế bởi một giá trị xác định, ví dụ thay thế giới tính thiếu bằng giá trị "LGBT"

```
df1_1['sex'] = df1_1['sex'].fillna('LGBT')
df1_1.head(10)
```

	ID	age	sex	region	income	married
0	1	48	FEMALE	INNER_CITY	17546.0	NO
1	2	40	MALE	TOWN	30085.1	YES
2	3	51	FEMALE	INNER_CITY	16575.4	YES
3	4	23	FEMALE	TOWN	20375.4	YES
4	5	57	FEMALE	RURAL	50576.3	YES
5	6	57	FEMALE	TOWN	37869.6	YES
6	7	22	LGBT	RURAL	NaN	NO

Bài tập

Thay thế các giá trị khuyết thiếu trong thuộc tính income bởi giá trị trung bình của chúng

Ghép các tệp dữ liệu (mở rộng thuộc tính)

```
print(df1_1)
print(df1_2)
df1=pd.merge(df1_1,df1_2, on='ID')
print(df1)
```

```
        ID  age     sex      region    income married
0        1   48  FEMALE  INNER_CITY  17546.00      NO
1        2   40    MALE        TOWN  30085.10     YES
2        3   51  FEMALE  INNER_CITY  16575.40     YES
3        4   23  FEMALE        TOWN  20375.40     YES
4        5   57  FEMALE       RURAL  50576.30     YES
..     ...  ...     ...         ...       ...     ...
295    296   34    MALE        TOWN  32548.90     YES
296    297   54  FEMALE       RURAL  24583.40      NO
297    298   18    MALE       RURAL   8639.24     YES
298    299   47  FEMALE  INNER_CITY  17139.50      NO
```

```
299  300   24  FEMALE  INNER_CITY  13667.70     YES

[300 rows x 6 columns]
      ID  children  car save_act current_act mortgage  pep
0      1         1   NO       NO          NO       NO  YES
1      2         3  YES       NO         YES      YES   NO
2      3         0  YES      YES         YES       NO   NO
3      4         3   NO       NO         YES       NO   NO
4      5         0   NO      YES          NO       NO   NO
..   ...       ...  ...      ...         ...      ...  ...
295  296         0  YES      YES         YES      YES   NO
296  297         2  YES      YES         YES      YES   NO
297  298         2   NO       NO          NO       NO   NO
298  299         2  YES       NO         YES       NO   NO
299  300         0   NO      YES         YES       NO   NO

[300 rows x 7 columns]
      ID  age     sex      region    income married  children  car save_act  \
0      1   48  FEMALE  INNER_CITY  17546.00      NO         1   NO       NO
1      2   40    MALE        TOWN  30085.10     YES         3  YES       NO
2      3   51  FEMALE  INNER_CITY  16575.40     YES         0  YES      YES
3      4   23  FEMALE        TOWN  20375.40     YES         3   NO       NO
4      5   57  FEMALE       RURAL  50576.30     YES         0   NO      YES
..   ...  ...     ...         ...       ...     ...       ...  ...      ...
295  296   34    MALE        TOWN  32548.90     YES         0  YES      YES
296  297   54  FEMALE       RURAL  24583.40      NO         2  YES      YES
297  298   18    MALE       RURAL   8639.24     YES         2   NO       NO
298  299   47  FEMALE  INNER_CITY  17139.50      NO         2  YES       NO
299  300   24  FEMALE  INNER_CITY  13667.70     YES         0   NO      YES

    current_act mortgage  pep
0            NO       NO  YES
1           YES      YES   NO
2           YES       NO   NO
3           YES       NO   NO
4            NO       NO   NO
..          ...      ...  ...
295         YES      YES   NO
296         YES      YES   NO
297          NO       NO   NO
298         YES       NO   NO
299         YES       NO   NO
```

```
[300 rows x 12 columns]
```

Ghép các tệp dữ liệu (mở rộng mẫu)

```
df=pd.concat([df1,df2])
print(df)
```

```
      ID  age     sex      region    income married  children  car save_act  \
0      1   48  FEMALE  INNER_CITY  17546.00      NO         1   NO       NO
1      2   40    MALE        TOWN  30085.10     YES         3  YES       NO
2      3   51  FEMALE  INNER_CITY  16575.40     YES         0  YES      YES
3      4   23  FEMALE        TOWN  20375.40     YES         3   NO       NO
4      5   57  FEMALE       RURAL  50576.30     YES         0   NO      YES
..   ...  ...     ...         ...       ...     ...       ...  ...      ...
295  296   61      NU  INNER_CITY  47025.00      NO         2  YES      YES
296  297   30      NU  INNER_CITY   9672.25     YES         0  YES      YES
297  298   31      NU        TOWN  15976.30     YES         0  YES      YES
298  299   29     NAM  INNER_CITY  14711.80     YES         0   NO      YES
299  300   38     NAM        TOWN  26671.60      NO         0  YES       NO

    current_act mortgage  pep
0            NO       NO  YES
1           YES      YES   NO
2           YES       NO   NO
3           YES       NO   NO
4            NO       NO   NO
..          ...      ...  ...
295         YES      YES   NO
296         YES       NO   NO
297          NO       NO  YES
298          NO      YES   NO
299         YES      YES  YES

[600 rows x 12 columns]
```

```
df=pd.concat([df1,df2], ignore_index=True)
```

```
print(df)
```

```
        ID  age     sex      region    income married  children  car save_act  \
0        1   48  FEMALE  INNER_CITY  17546.00      NO         1   NO       NO
1        2   40    MALE        TOWN  30085.10     YES         3  YES       NO
2        3   51  FEMALE  INNER_CITY  16575.40     YES         0  YES      YES
3        4   23  FEMALE        TOWN  20375.40     YES         3   NO       NO
4        5   57  FEMALE       RURAL  50576.30     YES         0   NO      YES
..     ...  ...     ...         ...       ...     ...       ...  ...      ...
595    296   61      NU  INNER_CITY  47025.00      NO         2  YES      YES
596    297   30      NU  INNER_CITY   9672.25     YES         0  YES      YES
597    298   31      NU        TOWN  15976.30     YES         0  YES      YES
598    299   29     NAM  INNER_CITY  14711.80     YES         0   NO      YES
599    300   38     NAM        TOWN  26671.60      NO         0  YES       NO

    current_act mortgage  pep
0            NO       NO  YES
1           YES      YES   NO
2           YES       NO   NO
3           YES       NO   NO
4            NO       NO   NO
..          ...      ...  ...
595         YES      YES   NO
596         YES       NO   NO
597          NO       NO  YES
598          NO      YES   NO
599         YES      YES  YES

[600 rows x 12 columns]
```

BTVN: Chuẩn hóa dữ liệu trên cột sex