THỰC HÀNH: LÀM SẠCH DỮ LIỆU CƠ BẢN

Import các thư viện

In [26]:
```
import pandas as pd
```

Đặt tên các cột dữ liệu cần thiết

In [27]:
```
column_name = ['Id', 'Name', 'Age', 'Weight', 'm0006', 'm0612', 'm1218', 'f0006', 'f0612', 'f1218']
```

3. Tiến hành tải dữ liệu vào chương trình ứng dụng Python và giải quyết vấn đề “Missing header in the csv file”

Đọc dữ liệu lên DataFrame

In [28]:
```
df = pd.read_csv('../data/patient_heart_rate.csv', names = column_name)
```

Hiển thị 10 dòng đầu tiên

In [29]:
```
df.head(10)
```

Out[29]:

	Id	Name	Age	Weight	m0006	m0612	m1218	f0006	f0612	f1218
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scöopy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

4. Xử lý vấn đề một cột lưu hỗn hợp nhiều dữ liệu, ở đây là cột “Name” chứa bao gồm “Firstname” và “Lastname”, giải pháp là ta sẽ tách ra làm 2 cột

In [30]:
```
df[['FirstName', 'LastName']] = df['Name'].str.split(expand=True)
df = df.drop('Name', axis=1)
```

Hiển thị 5 dòng đầu tiên

In [31]:
```
df.head(5)
```

Out[31]:

	Id	Age	Weight	m0006	m0612	m1218	f0006	f0612	f1218	FirstName	LastName
0	1.0	56.0	70kgs	72	69	71	-	-	-	Mickéy	Mousé
1	2.0	34.0	154.89lbs	-	-	-	85	84	76	Donald	Duck
2	3.0	16.0	NaN	-	-	-	65	69	72	Mini	Mouse
3	4.0	NaN	78kgs	78	79	72	-	-	-	Scrooge	McDuck
4	5.0	54.0	198.658lbs	-	-	-	69	NaN	75	Pink	Panther

5. Cột Weight có vấn đề về không thống nhất các đơn vị đo lường trong dữ liệu. Ta sẽ chuyển các đơn vị về thành đơn vị chuẩn “kg”

In [41]:
```
weight = df['Weight']

for i in range(0, len(weight)):
  x = str(weight[i])
  if 'lbs' in x[-3:]:
    x = x[:-3:]
    float_x = float(x)
    y = int(float_x/2.2)
    y = str(y)
    weight[i] = y
  if 'kgs' in x[:-3:]:
    x = x[:-3:]
```

In [42]:
```
df[['Weight']]
```

Out[42]:

	Weight
0	70kgs
1	70
2	NaN
3	78kgs
4	90
5	85
6	56kgs
7	78kgs
8	NaN
9	NaN
10	85
11	45kgs
12	60kgs
13	NaN
14	NaN
15	NaN
16	81

Chuyển kiểu dữ liệu của cột "Weight"

In [44]:
```
# Xoá các ký tự không phải số và không phải dấu thập phân từ cột 'Weight'
df['Weight'] = df['Weight'].str.replace('[^0-9.]', '', regex=True)

# Chuyển đổi kiểu dữ liệu của cột 'Weight' thành float
df['Weight'] = df['Weight'].astype(float)
```

In [46]:
```
df[['Weight']]
```

Out[46]:

	Weight
0	70.0
1	70.0
2	NaN
3	78.0
4	90.0
5	85.0
6	56.0
7	78.0
8	NaN
9	NaN
10	85.0
11	45.0
12	60.0
13	NaN
14	NaN
15	NaN
16	81.0

In [45]:
```
df["Weight"].dtype
```

Out[45]:
```
dtype('float64')
```

Đổi tên cột Weight thành Weight_kgs

In [47]:
```
df.rename(columns={'Weight': 'Weight_kgs'}, inplace=True)
```

Hiển thị 5 dòng đầu tiên

In [48]:
```
df.head(5)
```

Out[48]:

	Id	Age	Weight_kgs	m0006	m0612	m1218	f0006	f0612	f1218	FirstName	LastName
0	1.0	56.0	70.0	72	69	71	-	-	-	Mickéy	Mousé
1	2.0	34.0	70.0	-	-	-	85	84	76	Donald	Duck
2	3.0	16.0	NaN	-	-	-	65	69	72	Mini	Mouse
3	4.0	NaN	78.0	78	79	72	-	-	-	Scrooge	McDuck
4	5.0	54.0	90.0	-	-	-	69	NaN	75	Pink	Panther

6. Vấn đề về xuất hiện dòng dữ liệu rỗng (không có giá trị: NaN). Giải pháp có thể đưa ra là xóa bỏ

In [49]:
```
df.dropna(how='all', inplace=True)
```

7. Có nhiều dòng dữ liệu bị trùng lắp thông tin hoàn toàn [fullname, lastname, age, weight_kgs,....], giải pháp đưa ra là chỉ giữ lại một dòng dữ liệu, tuy nhiên giải pháp phải dựa trên nghiệp vụ của tập dữ liệu và quan sát của người xử lý.

In [50]:
```
df = df.drop_duplicates(subset=['FirstName', 'LastName', 'Age', 'Weight_kgs'])
```

In [51]:
```
df.head(5)
```

Out[51]:

	Id	Age	Weight_kgs	m0006	m0612	m1218	f0006	f0612	f1218	FirstName	LastName
0	1.0	56.0	70.0	72	69	71	-	-	-	Mickéy	Mousé
1	2.0	34.0	70.0	-	-	-	85	84	76	Donald	Duck
2	3.0	16.0	NaN	-	-	-	65	69	72	Mini	Mouse
3	4.0	NaN	78.0	78	79	72	-	-	-	Scrooge	McDuck
4	5.0	54.0	90.0	-	-	-	69	NaN	75	Pink	Panther

8. Xuất hiện dữ liệu bị ảnh hưởng bởi lỗi non-ASCII, không định dạng ASCII. Giải pháp: Tùy vào nghiệp vụ ta có thể: xóa dữ liệu tại đó, thay thế bằng dữ liệu khác hoặc thay bằng việc đánh dấu bằng một kí tự khác (ví dụ: ‘warning’)

In [52]:
```
# Problem 6:
df.FirstName.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.LastName.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
print(df)
```

```
     Id   Age  Weight_kgs m0006 m0612 m1218 f0006 f0612 f1218 FirstName  \
0   1.0  56.0        70.0    72    69    71     -     -     -     Micky
1   2.0  34.0        70.0     -     -     -    85    84    76    Donald
2   3.0  16.0         NaN     -     -     -    65    69    72      Mini
3   4.0   NaN        78.0    78    79    72     -     -     -   Scrooge
4   5.0  54.0        90.0     -     -     -    69   NaN    75      Pink
5   6.0  52.0        85.0     -     -     -    68    75    72      Huey
6   7.0  19.0        56.0     -     -     -    71    78    75     Dewey
7   8.0  32.0        78.0    78    76    75     -     -     -      Scpy
11 10.0  12.0        45.0     -     -     -    92    95    87     Louie
12 11.0   NaN        60.0    78    75    72     -     -     -     Henry
13 12.0  34.0         NaN    65    67    55     -     -     -    Michel
14 13.0   NaN         NaN     -     -     -    68    72    70      Tana
15 14.0   NaN         NaN   NaN   NaN   NaN   NaN   NaN   NaN  UniLever
16 15.0  52.0        81.0     -     -     -    68    75    72       NaN

   LastName
0      Mous
1      Duck
2     Mouse
3    McDuck
4   Panther
5    McDuck
6    McDuck
7       Doo
11   McDuck
12      Nam
13     Long
14    Ricky
15     None
16      NaN
```

9. Thay giá trị thiếu của tuổi bằng giá trị yếu vị.

In [53]:
```
df['Age'].fillna(df['Age'].mode()[0], inplace=True)
```

10. Thay giá trị thiếu của cân nặng bằng giá trị trung vị

In [54]:
```
df['Weight_kgs'].fillna(df['Weight_kgs'].median(), inplace=True)
```

11. “Một cột chứa quá nhiều thông tin cần được phân rã”, như trong bài toán này ta thấy header “m0006” chứa các nội dung bao gồm: m → male, 0006 ~ 00-06 (lần đo chỉ số huyết áp từ 00h- đến 06h). Còn giá trị thì là kết quả huyết áp.

Chúng ta sẽ tách nội dung của cột này ra làm 3 cột sau: PulseRate : giá trị huyết áp, Sex: giới

Bước 1: Tạo melt dữ liệu để có cột gender_time

In [55]:
```
df = pd.melt(df, id_vars=['Id', 'Age', 'Weight_kgs', 'FirstName', 'LastName'],
             value_name='PulseRate', var_name='gender_time').sort_values(['Id', 'Age', 'Weight_kgs', 'FirstName',
```

In [56]:
```
df.head(5)
```

Out[56]:

	Id	Age	Weight_kgs	FirstName	LastName	gender_time	PulseRate
0	1.0	56.0	70.0	Micky	Mous	m0006	72
14	1.0	56.0	70.0	Micky	Mous	m0612	69
28	1.0	56.0	70.0	Micky	Mous	m1218	71
42	1.0	56.0	70.0	Micky	Mous	f0006	-
56	1.0	56.0	70.0	Micky	Mous	f0612	-

Bước 2: Tạo data frame tạm là kết quả của việc tách cột gender_time

In [57]:
```
df_temp = df['gender_time'].str.extract("(\D)(\d+)(\d{2})", expand=True)
```

Bước 3: Đặt tên cột cho data frame tạm

In [58]:
```
df_temp.columns = ['Gender', 'Lower_hour', 'Upper_hour']
```

Bước 4: Nối data frame tạm vào data frame ban đầu

In [59]:
```
df = pd.concat([df, df_temp], axis=1)
```

In [60]:
```
df.head(5)
```

Out[60]:

	Id	Age	Weight_kgs	FirstName	LastName	gender_time	PulseRate	Gender	Lower_hour	U
0	1.0	56.0	70.0	Micky	Mous	m0006	72	m	00	
14	1.0	56.0	70.0	Micky	Mous	m0612	69	m	06	
28	1.0	56.0	70.0	Micky	Mous	m1218	71	m	12	
42	1.0	56.0	70.0	Micky	Mous	f0006	-	f	00	
56	1.0	56.0	70.0	Micky	Mous	f0612	-	f	06	

Bước 5: Bỏ cột gender_time

In [61]:
```
df = df.drop(['gender_time'], axis=1)
```

12. Loại bỏ hết các dòng dữ liệu thừa là những dòng có phần PulseRate có dấu -

In [62]:
```
import numpy as np
```

In [63]:
```
df = df.replace('-', np.nan).dropna(subset=['PulseRate'])
df.head(10)
```

Out[63]:

	Id	Age	Weight_kgs	FirstName	LastName	PulseRate	Gender	Lower_hour	Upper_hour
0	1.0	56.0	70.0	Micky	Mous	72	m	00	06
14	1.0	56.0	70.0	Micky	Mous	69	m	06	12
28	1.0	56.0	70.0	Micky	Mous	71	m	12	18
43	2.0	34.0	70.0	Donald	Duck	85	f	00	06
57	2.0	34.0	70.0	Donald	Duck	84	f	06	12
71	2.0	34.0	70.0	Donald	Duck	76	f	12	18
44	3.0	16.0	74.0	Mini	Mouse	65	f	00	06
58	3.0	16.0	74.0	Mini	Mouse	69	f	06	12
72	3.0	16.0	74.0	Mini	Mouse	72	f	12	18
3	4.0	34.0	78.0	Scrooge	McDuck	78	m	00	06

13. Nhận thấy có những bệnh nhân chưa ghi nhận họ tên (ví lý do nào đó)

In [64]:
```
df['FirstName'].isnull().sum()
```

Out[64]: 3

In [65]:
```
df['LastName'].isnull().sum()
```

Out[65]: 3

Nhưng giá trị huyết áp và thời gian đo huyết áp thì đầy đủ nên dữ liệu quan tâm là trị số huyết áp vẫn dùng được, nên ta thay họ, tên bị thiếu thành Unknown

In [66]:
```
df['FirstName'].fillna('Unknown', inplace=True)
```

In [67]:
```
df['LastName'].fillna('Unknown', inplace=True)
```

14. Sau khi xử lý thì index của dòng dữ liệu đã thay đổi lung tung, ta cần reset index lại cho theo khuôn mẫu

In [68]:
```
df = df.reset_index()
```

15. Sau đó, lưu trữ dữ liệu đã xử lý thành công với tên file patient_heart_rate_clean.csv

In [69]:
```
df.to_csv('../data/patient_heart_rate_clean.csv')
```