WRITTEN BY MR. HUYNH NAM

Website: www.yelo.vn

THỰC HÀNH: LÀM SẠCH DỮ LIỆU CƠ BẢN

Nội dung: Xử lý dữ liệu y khoa về huyết áp của bệnh nhân

Mục tiêu: Sinh viên biết cách sử dụng gói Pandas để xử lý dữ liệu

1. Tiến hành hiểu dữ liệu từ chuyên gia
   "*The data set has been kept small enough for you to be able to grok it all at once. The data is in csv format. Each row in the dataset has data about different individuals and their heart rate details for different time intervals. The columns contain information such as individual's Age, Weight, Sex and Heart Rates taken at different time intervals.*"
2. Thông thường ta thường xử lý các vấn đề sau về dữ liệu
   1. Missing headers in the csv file
   2. Multiple variables are stored in one column
   3. Column data contains inconsistent unit values
   4. An empty row in the data
   5. Duplicate records in the data
   6. Non-ASCII characters
   7. Missing values
   8. Column headers are values and not variable names
3. Tiến hành tải dữ liệu vào chương trình ứng dụng Python và giải quyết vấn đề "Missing header in the csv file"

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

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

```
[3] df = pd.read_csv('/content/patient_heart_rate.csv',names = column_name)
```

```
[4] df.head(10)
```

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

```
[11] df[['Firstname','Lastname']] = df['Name'].str.split(expand=True)
     df = df.drop('Name',axis=1)
```

```
[12] df.head(5)
```

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"

```
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:]
    float_x = float(x)
    weight[i]=x
```

Đổi tên cột Weight thành Weight_kgs

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

```
df.head(5)
```

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

```
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.

```
df = df.drop_duplicates(subset=['Firstname','Lastname','Age','Weight_kgs'])
```

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’)

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

9. "Missing values", vấn đề này xảy ra tại các cột "Age", "Weight" và "Heart Rate". Thiếu dữ liệu (dữ liệu không đầy đủ) là vấn đề xảy ra nhiều trong các nguồn dữ liệu do nhiều nguyên nhân chủ quan lẫn khách quan. Có một vài giải pháp để xử lý vấn đề này, chủ yếu dựa trên kinh nghiệm và nghiệp vụ về tập dữ liệu đó. Một số giải pháp đưa đề xuất từ chuyên gia như sau:
    a. **Deletion**: Remove records with missing values
    b. **Dummy substitution**: Replace missing values with a dummy but valid value: e.g.: 0 for numerical values.
    c. **Mean substitution**: Replace the missing values with the mean.
    d. **Frequent substitution**: Replace the missing values with the most frequent item.
    e. **Improve the data collector**: Your business folk will talk to the clients and inform them about why it is worth fixing the problem with the data collector.

    ……

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

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

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

```
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.

3	4.0	NaN	78kgs	78	79	72	-	-	-	Scrooge	McDuck
4	5.0	54.0	90kgs	-	-	-	69	NaN	75	Pink	Panther
5	6.0	52.0	85kgs	-	-	-	68	75	72	Huey	McDuck
6	7.0	19.0	56kgs	-	-	-	71	78	75	Dewey	McDuck
7	8.0	32.0	78kgs	78	76	75	-	-	-	Scööpy	Doo

Chúng ta sẽ tách nội dung của cột này ra làm 3 cột sau: PulseRate : giá trị huyết áp, Sex: giới tính (m: male, f: female) và time: thời gian (từ giờ-đến giờ) như sau:

	Id	Age	Weight	Firstname	Lastname	PulseRate	Sex	Time
0	1.0	56.0	70kgs	Micky	Mous	72	m	00-06
9	1.0	56.0	70kgs	Micky	Mous	69	m	06-12
18	1.0	56.0	70kgs	Micky	Mous	71	m	12-18
27	1.0	56.0	70kgs	Micky	Mous	-	f	00-06
36	1.0	56.0	70kgs	Micky	Mous	-	f	06-12
45	1.0	56.0	70kgs	Micky	Mous	-	f	12-18

Gợi ý:

Bước 1: Tạo melt dữ liệu để có cột gender_time

```
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','Lastname'])
```

	Id	Age	Weight_kgs	Firstname	Lastname	gender_time	PulseRate
0	1.0	56.0	70	Micky	Mous	m0006	72
14	1.0	56.0	70	Micky	Mous	m0612	69
28	1.0	56.0	70	Micky	Mous	m1218	71
42	1.0	56.0	70	Micky	Mous	f0006	-
56	1.0	56.0	70	Micky	Mous	f0612	-
...	...	...	...	...	...	...	...

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

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

	0	1	2
0	m	00	06
14	m	06	12
28	m	12	18
42	f	00	06
56	f	06	12
...	...	...	...

Bước 3: Đặt tên cột cho data frame tạm

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

	Gender	Lower_hour	Upper_hour
0	m	00	06
14	m	06	12
28	m	12	18
42	f	00	06
56	f	06	12
...	...	...	...

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

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

	Id	Age	Weight_kgs	Firstname	Lastname	gender_time	PulseRate	Gender	Lower_hour	Upper_hour
0	1.0	56.0	70	Micky	Mous	m0006	72	m	00	06
14	1.0	56.0	70	Micky	Mous	m0612	69	m	06	12
28	1.0	56.0	70	Micky	Mous	m1218	71	m	12	18
42	1.0	56.0	70	Micky	Mous	f0006	-	f	00	06
56	1.0	56.0	70	Micky	Mous	f0612	-	f	06	12

Bước 5: Bỏ cột gender_time

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

	Id	Age	Weight_kgs	Firstname	Lastname	PulseRate	Gender	Lower_hour	Upper_hour
0	1.0	56.0	70	Micky	Mous	72	m	00	06
14	1.0	56.0	70	Micky	Mous	69	m	06	12
28	1.0	56.0	70	Micky	Mous	71	m	12	18
42	1.0	56.0	70	Micky	Mous	-	f	00	06
56	1.0	56.0	70	Micky	Mous	-	f	06	12
...	...	...	...	...	...	...	...	...	...

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

```
import numpy as np
```

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

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

```
df['Firstname'].isnull().sum()
```

```
df['Lastname'].isnull().sum()
```

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

```
df['Firstname'].fillna('Unknown', inplace=True)
```

```
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

```
df=df.reset_index()
```

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

```
df.to_csv('patient_heart_rate_clean.csv')
```