

## LAB: LÀM SẠCH DỮ LIỆU CƠ BẢN

**Mục tiêu:** Sinh viên biết cách sử dụng gói Pandas để xử lý dữ liệu và thống kê cơ bản dữ liệu

**Nội dung:** Sử dụng bộ dữ liệu `patient_heart_rate.csv`

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. Vấn đề 1: Tiến hành tải dữ liệu vào chương trình ứng dụng Python và giải quyết vấn đề “Missing header in the csv file”

```
#Problem 1
# Thêm header vào dataframe để diễn giải dữ liệu
column_names= ["Id", "Name", "Age", "Weight", 'm0006', 'm0612', 'm1218', 'f0006', 'f0612', 'f1218']
# Đọc file dữ liệu
df = pd.read_csv("patient_heart_rate.csv", names = column_names)
#Hiển thị một vài dòng dữ liệu đầu tiên ra màn hình
print(df.head())
```

4. Vấn đề 2: 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

```
#Problem 2
df[['Firstname', 'Lastname']] = df['Name'].str.split(expand=True)
df = df.drop('Name', axis=1)
print (df)
```

5. Vấn đề 3: 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”

```
# Problem 3
#Get the Weight column
weight = df['Weight']

for i in range(0, len(weight)):
    x = str(weight[i])
    #Incase lbs is part of observation remove it
    if "lbs" in x[-3:]:
        #Remove the lbs from the value
        x = x[:-3]
        #Convert string to float
        float_x = float(x)
        #Covert to kgs and store as int
        y = int(float_x/2.2)
        #Convert back to string
        y = str(y)+"kgs"
        weight[i] = y

print(df)
```

6. Vấn đề 4: Vấn đề về xuất hiện dòng dữ liệu rỗng (không có giá trị: NaN). Giải pháp có thể đưa ra là xóa bỏ

```
# Problem 4:
df.dropna(how="all", inplace=True)
print(df)
```

7. Vấn đề 5: Có nhiều dòng dữ liệu bị trùng lặp thông tin hoàn toàn[fullname, lastname, age, weight,...], 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'])
print(df)
```

8. Vấn đề 6: 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 ký 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. Vấn đề “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:

- Deletion:** Remove records with missing values
- Dummy substitution:** Replace missing values with a dummy but valid value: e.g.: 0 for numerical values.
- 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.

.....

Yêu cầu thay thế như sau: Nếu dòng nào có Age hoặc Weight có dữ liệu thì phần Age hoặc Weight được tính như bên dưới, nếu thiếu cả 2 thông tin thì xóa dòng

Age: Giá trị thay thế là mean của các giá trị trong cột Age

Weight: Giá trị thay thế là mean của các giá trị trong cột Weight

10. Vấn đề “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, 1218 ~ 12-18. 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 (tháng-ngày) 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

Mã nguồn:

```
##Melt the Sex + time range columns in single column
df = pd.melt(df, id_vars=['Id', 'Age', 'Weight', 'Firstname', 'Lastname'], value_name="PulseRate", var_name="sex_and_time").sort_values(['Id', 'Age', 'Weight', 'Firstname', 'Lastname'])

# Extract Sex, Hour Lower bound and Hour upper bound group
tmp_df = df["sex_and_time"].str.extract("(\\D)(\\d+)(\\d{2})", expand=True)

# Name columns
tmp_df.columns = ["Sex", "hours_lower", "hours_upper"]

# Create Time column based on "hours_lower" and "hours_upper" columns
tmp_df["Time"] = tmp_df["hours_lower"] + "-" + tmp_df["hours_upper"]

# Merge
df = pd.concat([df, tmp_df], axis=1)




# Drop unnecessary columns and rows
df = df.drop(['sex_and_time', 'hours_lower', 'hours_upper'], axis=1)
df = df.dropna()
df.to_csv('outputcleanup.csv', index=False)
print (df)
```

11. Hãy rút gọn dữ liệu phù hợp và reindex lại dữ liệu.

Lưu ý: Ngoài ra còn rất nhiều vấn đề về mặt xử lý dữ liệu dựa trên nhiều khía cạnh khác nhau tùy vào sự am hiểu về dữ liệu của các chuyên gia như:

1. Handling dates
2. Correcting character encodings (a problem you hit when you scrape data off the web)

12. Phân tích dữ liệu giá trị huyết áp theo bảng phân loại và vẽ các biểu đồ liên quan

	Tuổi	Huyết áp thấp	Huyết áp bình thường	Huyết áp cao
	1 - 12 Tháng	75 / 50	90 / 60	100 / 75
	1 - 5 Tuổi	80 / 55	95 / 65	110 / 79
	6 - 13 Tuổi	90 / 60	105 / 70	115 / 80
	14 - 19 Tuổi	105 / 73	117 / 77	120 / 81
	20 - 24 Tuổi	108 / 75	120 / 79	132 / 83
	25 - 29 Tuổi	109 / 76	121 / 80	133 / 84
	30 - 34 Tuổi	110 / 77	122 / 81	134 / 85
	35 - 39 Tuổi	111 / 78	123 / 82	135 / 86
	40 - 44 Tuổi	112 / 79	125 / 83	137 / 87
	45 - 49 Tuổi	115 / 80	127 / 84	139 / 88
	50 - 54 Tuổi	116 / 81	129 / 85	142 / 89
	55 - 59 Tuổi	118 / 82	131 / 86	144 / 90
	60 - 64 Tuổi	121 / 83	134 / 87	147 / 91

13. Vẽ biểu đồ thể hiện mối quan hệ của huyết áp với cân nặng, tuổi và giới tính

14. Mô tả phân phối của giá trị huyết áp.