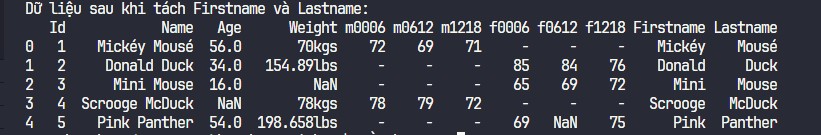
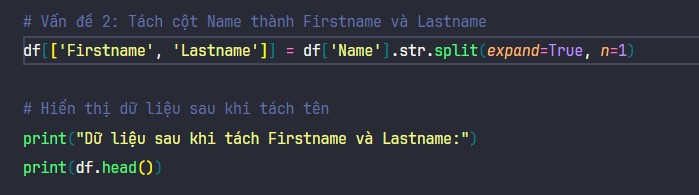
**THỰC HÀNH LAB 03: LÀM SẠCH DỮ LIỆU CƠ BẢN**

***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”



***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



**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”

A computer screen shot of text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**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

A computer screen with text

AI-generated content may be incorrect.

A screenshot of a computer generated image

AI-generated content may be incorrect.

**Vấn đề 5:** Có nhiều dòng dữ liệu bị trùng lắp 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.

A black screen with colorful text

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

**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 kí tự khác (ví dụ: ‘warning’)

A computer screen with text

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

**Vấn đề 7:** “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.

Yêu cầu:

- Thống kê thông tin dữ liệu thiếu trên từng biến Age và Weight

- Yêu cầu xử lý dữ liệu thiếu 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

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

A computer screen shot of text

AI-generated content may be incorrect.

A purple and yellow rectangle with white text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Vấn đề 8:** “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 (mm-dd).



**Vấn đề 11:** Hãy khảo sát tỉ lệ dữ liệu thiếu trên biến huyết áp. Dữ liệu bị thiếu thì hãy xử lý bằng phương pháp sau

• Thay thế bằng giá trị trung bình liền trước và liền sau của người đó. Nếu không được thì dùng 2)

• Thay thế bằng giá trị trung bình 2 giá liền trước của người đó. Nếu không được thì dùng 3)

• Thay thế bằng giá trị trung bình 2 giá liền sau của người đó. Nếu không được thì dùng 4)

• Trung bình của các giá trị huyết áp của người đó. Nếu không được thì dùng 5).

• Trung bình của các giá trị huyết áp của nhóm giới tính. Nếu không được thì dùng 6)

• Trung bình của các giá trị dữ liệu. Nếu không được thì thay bằng mức ổn định trong y học.

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a code

AI-generated content may be incorrect.

**Vấn đề 12:** Hãy rút gọn dữ liệu phù hợp và reindex lại dữ liệu. Sau đó, lưu trữ dữ liệu đã xử lý thành công với tên file patient\_heart\_rate\_clean.csv

