BÀI TẬP DATA CLEANING

BÀI 1: Cho file dữ liệu **data_obsity.csv**

1. Đọc file dữ liệu và hiển thị 5 hàng đầu, hiển thị các column và kích thước dữ liệu
2. 3 hàng đầu dường như lặp lại các thông tin, hãy remove thông tin bị lặp lại
3. Hai hàng đầu chứa các giá trị giống nhau trong tất cả các cột ngoại trừ cột đầu tiên và giá trị cột này là NaN. Cột thứ 3 chỉ ra giới tính nhưng chúng ta có thể lấy các thông tin như nhau từ các tên cột, ví dụ 2016.1 là male, 2016.2 là female và 2016 là trung bình và đây là giá trị hợp lệ cho tất cả các năm. 3 hàng đầu là dư thừa vì vậy hãy xóa chúng.
4. Hãy xóa các cột dư thừa
5. Hãy xóa cột "variable" column

BÀI 2: Cho dữ liệu kiểu từ điển sau :

```
{'Name': ['Nik', 'Kate', 'Evan', 'Kyra', 'Nik', 'Kate'], 'Age': [33, 32, 40, 57, 33, 32], 'Location': ['Toronto', 'London', 'New York', 'Atlanta', 'Toronto', 'Paris'], 'Date Modified': ['2022-01-01', '2022-02-24', '2022-08-12', '2022-09-12', '2022-01-01', '2022-12-09']}
```

1. Xác định giá trị nào là missing hay không từ dữ liệu trên
2. Thực hiện đếm giá trị missing có trong dữ liệu
3. Thực hiện xóa dữ liệu missing data từ dữ liệu trên
4. Thực hiện fill dữ liệu missing data thành giá trị 0
5. Thực hiện fill dữ liệu missing data theo phương pháp thống kê sử dụng SimpleImputer
6. Thực hiện fill dữ liệu missing data theo phương pháp thống kê sử dụng K láng giềng gần nhất (KNNImputer)
7. Thực hiện fill dữ liệu missing data theo phương pháp thống kê sử dụng phương pháp lặp Iterative (IterativeImputer)

8. Xóa dữ liệu duplicate theo hàng và cột nếu có

BÀI 3: Load dữ liệu dataframe sau và thực hiện xử lý :

```
{ 'Name': ['Tranter; Melvyn', 'Lana; Courtney', 'Abel; Shakti', 'Vasu; Imogene', 'Aravind; Shelly', 'Tranter; Melvyn'], 'Location': ['TORONTO', 'LONDON', 'New york', np.NaN, 'toronto', 'Madrid'], 'Sales': [123, 243, 654, np.NaN, 345, np.NaN] }
```

1. Tạo cột First Name và Last Name. Lưu ý rằng có một dấu chấm phẩy giữa các tên.
2. Loại bỏ bản ghi trùng lặp dựa trên cột Name, giữ lại bản ghi cuối cùng.
3. Tính tỷ lệ % bản ghi bị thiếu trong mỗi cột.

BÀI 4: Cho file data **unclean_data.csv**

1. Đọc file dữ liệu với định dạng encoding='utf8'
2. Hiển thị tên các cột, chuyển tên tất cả các cột thành chữ HOA
3. Rename tên cột 'DURATION' thành 'TIME'
4. Thêm giá trị mặc định cho dữ liệu bị thiếu (missing data) hoặc sử dụng giá trị trung bình (mean) để fill dữ liệu missing đó
5. Xóa hàng/cột thiếu dữ liệu
6. Xóa các hàng cột bị duplicate