TÓM TẮT QUY TRÌNH MÔ PHỎNG DỮ LIỆU TÀI CHÍNH CÁ NHÂN

I. Synthetic data generation based on real data

Mục tiêu: Mục tiêu **tạo ra một bộ dữ liệu tài chính cá nhân tổng hợp** (giả lập) có cấu trúc giống dữ liệu thực, dựa trên các thông tin công khai từ **Bureau of Labor Statistics (BLS.gov)** – cụ thể là **bảng chi tiêu theo độ tuổi trong năm 2023**.

Ý tưởng tổng quát

- Sử dụng hai nguồn dữ liệu:
    - **CSV ([1] financial-literacy-data.csv)**: là bảng dữ liệu gốc, mô phỏng chi tiêu cá nhân đã được làm sạch, làm cơ sở để tạo lại các biến chi tiêu.
    - **XLSX (reference-person-age-ranges-2023.xlsx)**: là bảng **lookup** từ BLS.gov, chứa các giá trị **mean, SE, N** theo độ tuổi – dùng để **lấy gần đúng phân phối thực tế**.

II. Phương pháp mô phỏng

1. Phân phối ban đầu

- Tất cả các mục chi tiêu là **dữ liệu dương**, nên **phân phối log-normal** được lựa chọn là hợp lý trong bước đầu tiên.

2. Vấn đề phát sinh

- Các chi tiêu được sinh theo từng mục riêng biệt → **mất tương quan giữa các mục chi tiêu**.
- Trong thực tế, các khoản chi tiêu thường **có quan hệ với nhau** (corr), đặc biệt theo nhóm tuổi → sinh dữ liệu như trên là **thiếu hợp lý**.

3. Giải pháp

- Chuyển sang dùng phân phối **đa biến (multivariate normal)** sau khi log-transform dữ liệu → giữ được **ma trận tương quan (correlation matrix)**.
- Ma trận này được tính từ bảng lookup của BLS, áp dụng ngược để sinh dữ liệu mô phỏng thực tế hơn.

Xử lý file CSV financial-literacy-data.csv

- File có **~20.000 dòng**, được **làm sạch sẵn** bởi thành viên nhóm:
    - Phân chia độ tuổi
    - Loại bỏ outlier

- **Không có chỉnh sửa thêm** trong notebook này.

Xử lý file XLSX (lookup table từ BLS)

B1: Lấy dữ liệu từ BLS.gov, chọn các chỉ tiêu phù hợp với dữ liệu CSV gốc
B2: Làm sạch lần 1 (thủ công): loại bỏ các mục không cần thiết
B3: Làm sạch lần 2 bằng Python:

- Chuẩn hóa định dạng
- Chuyển bảng từ dạng **wide** sang **long** (pivot các cột độ tuổi thành 1 cột)

B4: Tính **ma trận tương quan (corr_matrix)** giữa các mục chi tiêu trong bảng lookup
B5: Viết hàm sinh dữ liệu:

- Chọn nhóm tuổi (hoặc tuổi cụ thể)
- Sinh vector chi tiêu từ phân phối **multivariate normal**
- Gán lại các cột chi tiêu vào bảng mô phỏng

B6: Kiểm tra lại dữ liệu sinh ra, đảm bảo tính hợp lý và kết thúc bước xử lý.

Ưu điểm

- Mô phỏng dữ liệu **sát thực tế**, dựa trên dữ liệu chính thống từ BLS.gov
- Giữ được **mối quan hệ giữa các khoản chi tiêu** (corr giữa các mục)
- Dữ liệu có thể dùng để huấn luyện mô hình, xây dashboard, hoặc phân tích hành vi chi tiêu

Hạn chế

- Phân phối log-normal hoặc multivariate-normal chỉ **xấp xỉ** được thực tế, không phản ánh được các mối quan hệ phi tuyến
- Việc tính toán ma trận tương quan từ bảng mean có thể **chưa phản ánh đúng phương sai thực sự**, do thiếu dữ liệu gốc
- Dữ liệu mô phỏng **không có yếu tố thời gian** hoặc khu vực địa lý cụ thể (nếu cần mở rộng)