BÁO CÁO ĐỒ ÁN CUỐI KỲ

MÔN: TRỰC QUAN HÓA DỮ LIỆU

Nhóm 10 - Chủ đề 1: Tài chính cá nhân

Trong bối cảnh ngày nay, khi chi phí sinh hoạt liên tục biến động và lạm phát ảnh hưởng đến mọi khía cạnh của cuộc sống, việc quản lý tài chính cá nhân không còn là lựa chọn – mà đã trở thành một kỹ năng thiết yếu. Đặc biệt đối với giới trẻ hiện đại, hiểu biết về tài chính và mong muốn đạt được tự do tài chính sớm đang ngày càng trở nên phổ biến. Họ không chỉ muốn kiểm soát thu nhập – chi tiêu hàng tháng mà còn hướng đến những mục tiêu dài hạn như mua nhà, đầu tư hay nghỉ hưu sớm. Để làm được điều đó, việc phân tích dữ liệu tài chính cá nhân đóng vai trò quan trọng. Thông qua việc theo dõi xu hướng chi tiêu, nhận diện các giai đoạn chi tiêu cao bất thường, và phân bổ ngân sách hợp lý, người dùng có thể chủ động đưa ra các quyết định tài chính phù hợp với từng giai đoạn cuộc sống. Để có cái nhìn trực quan và toàn cảnh, nhóm quyết định phân tích và vẽ dashboard, phần nào có thể khái quát và đưa ra cái nhìn tốt nhất cho các cá nhân.

I. Tổng quan về dữ liệu và mục tiêu phân tích

1. Financial Literacy Dataset

Là một tập dữ liệu tài chính cá nhân mô phỏng, phản ánh các yếu tố liên quan đến thu nhập, chi tiêu, thói quen tiết kiệm và tiềm năng tiết kiệm của các cá nhân.

- Dữ liệu gồm: 20000 dòng và 27 cột.
- Không có các giá trị thiếu hay trùng lặp.

Trong tập dữ liệu có nhiều cột với ý nghĩa khác nhau có thể phân thành 4 nhóm chính:

1. Thông tin cá nhân và nhân khẩu học: Income, Age, Dependents, Occupation, City_Tier.
2. Chi tiêu hàng năm: Rent, Loan_Repayment, Insurance, Groceries, Transport, Eating_out, Entertainment, Utilities, Education, Miscellaneous.
3. Thông tin về tiết kiệm: Desired_Savings_Percentage, Desired_Savings, Disposable_Income.
4. Tiềm năng tiết kiệm từ từng hạng mục: các cột còn lại.

Mục tiêu phân tích:

- Phân tích và tổng quan được xu hướng tài chính của 20,000 cá nhân khác nhau.

- Hiểu rõ sự khác biệt tài chính giữa các nhóm nhân khẩu học.
- Giúp cá nhân so sánh bản thân với cộng đồng tương đồng.
- Hỗ trợ tổ chức tài chính phân khúc và định hướng sản phẩm phù hợp.

2. <u>Personal Finance Dataset</u>

Là một tập dữ liệu mô phỏng lịch sử giao dịch của một cá nhân từ năm 2020-2024.

- Dữ liệu gồm: 1500 dòng và 5 cột
- Không có các giá trị thiếu hay trùng lặp

Trong đó các cột có ý nghĩa sau:

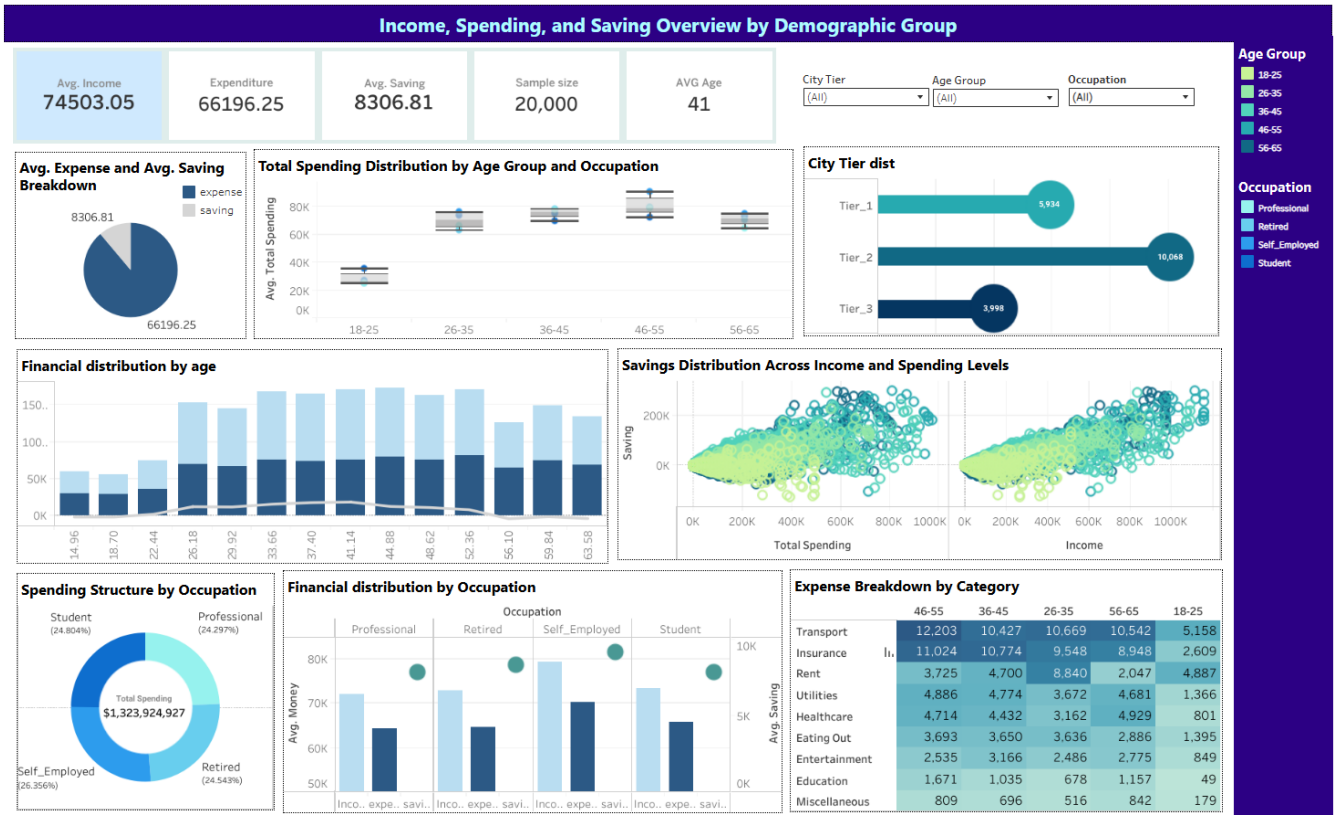
- Date: Ngày giao dịch
- Transaction Description: Mô tả nội dung giao dịch
- Category: Nhóm chi tiêu hoặc thu nhập của giao dịch
- Amount: Tổng chi hoặc thu
- Type: Phân loại nhóm chi tiêu hoặc thu nhập

Mục tiêu phân tích:

- Khám phá cơ cấu chi tiêu, tỷ trọng giữa các nhóm thiết yếu và không thiết yếu.
- Tìm kiếm những cơ hội điều chỉnh để tối ưu hóa ngân sách, tỷ lệ tiết kiệm và gợi ý điều chỉnh thói quen cá nhân.

I. Dashboard

1. Dashboard Financial Literacy Dataset



Dashboard: Income, Spending, and Saving Overview by Demographic Group

Link dashboard:
https://public.tableau.com/app/profile/nguy.n.phan.ho.ng.ph.c/viz/Book1_17516920190310/General?publish=yes

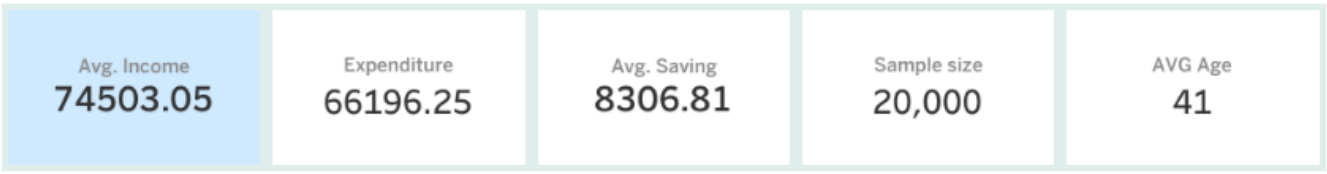
Các thành phần chính trong dashboard

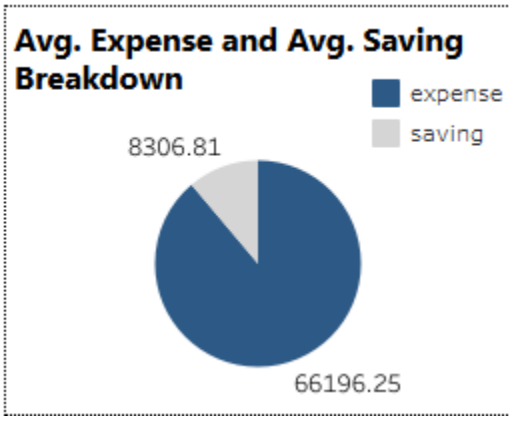
- Tổng quan tài chính (KPI cards): trung bình thu nhập, chi tiêu, tiết kiệm, tuổi trung bình và kích thước mẫu
- Phân phối chi tiêu theo độ tuổi và nghề nghiệp (Boxplot): so sánh mức chi tiêu giữa các nhóm nghề và độ tuổi
- Phân bố cư dân theo City Tier: ảnh hưởng của môi trường sống đến hành vi tài chính
- Thu nhập, chi tiêu và tiết kiệm theo độ tuổi (Stacked): xu hướng tích lũy theo chu kỳ tuổi.

- Mối quan hệ giữa thu nhập, chi tiêu và tiết kiệm (Scatter): nhận diện mối quan hệ giữa các yếu tố.
- Cơ cấu chi tiêu theo nghề nghiệp: xác định các nghề có tỷ trọng chi tiêu như thế nào.
- Chi tiêu theo danh mục và độ tuổi (Heatmap): nhận diện mối liên hệ giữa độ tuổi và các danh mục.

Những điểm nổi bật của dashboard:

- Thu nhập trung bình của mẫu dữ liệu là: 74503.05
- Chi tiêu là: 66196 chiếm gần 89% thu nhập
- Khoản tiết kiệm trung bình là: 8307 chiếm khoảng 11% thu nhập
- Mẫu dữ liệu: 20000 cá nhân
- Tuổi trung bình là 41 tuổi cho thấy lực lượng lao động trung niên chiếm tỷ trọng cao.
- Trung bình chi phí gấp khoảng 8 lần trung bình tiết kiệm → xu hướng tiết kiệm còn rất thấp.



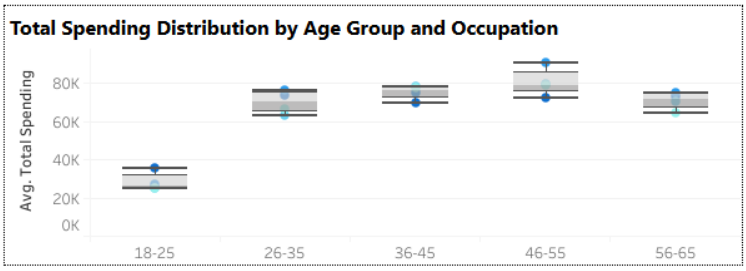


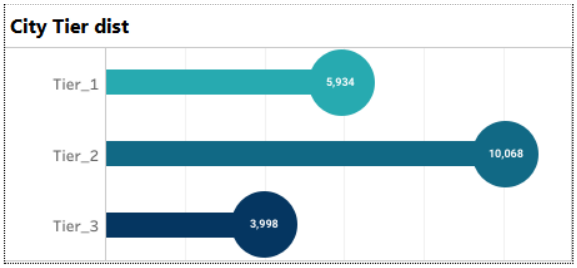
Cơ cấu dân số & vùng sinh sống:

City Tier:

- Tier-2 chiếm khoảng 50% mẫu (khoảng 10 nghìn người) → cho thấy kết quả tổng hợp dễ bị chi phối bởi nhóm này.
- Tier-1 khoảng 5934 nghìn người, Tier-3 khoảng 4000 nghìn người.

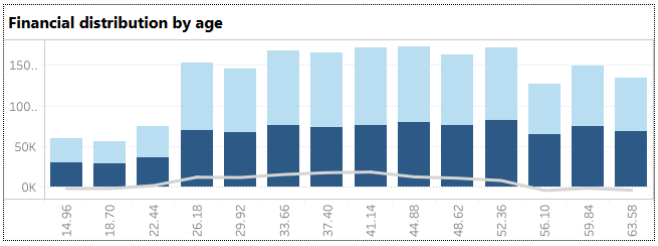
Age group: năm nhóm tuổi được phân bố khá đều, trung vị khoảng 36-45.

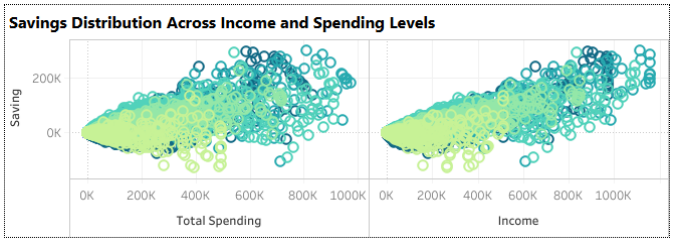




Thu nhập – chi tiêu – tiết kiệm theo độ tuổi

- Thu nhập và chi tiêu tăng dần từ 18-25 tới đỉnh 46-55, sau đó giảm nhẹ ở độ tuổi khoảng 56-65.
- Khoảng cách giữa thu và chi rộng nhất ở nhóm tuổi 46-55.
- Nhóm tuổi 18-25 có tỷ lệ tiết kiệm thấp nhất.



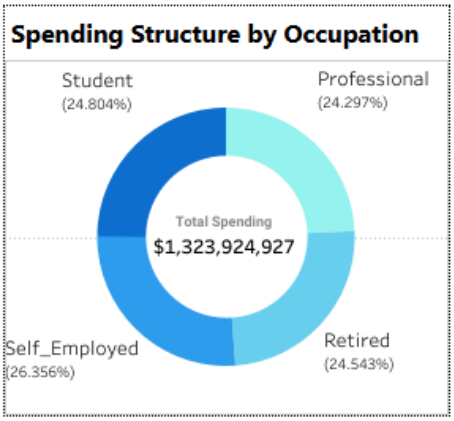


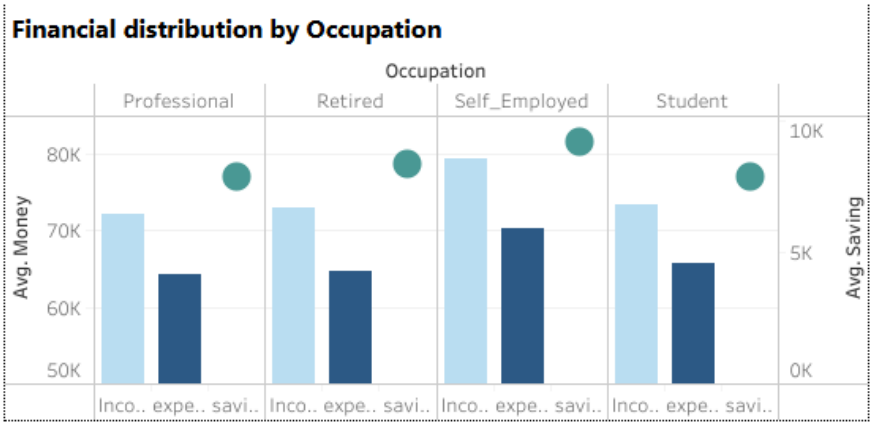
Sự khác biệt theo nghề nghiệp

Cấu trúc chi tiêu theo nghề gần như được chia đều khoảng 25% mỗi nhóm, cho thấy kích cỡ mẫu nghề nghiệp khá cân xứng.

- Self-Employed là nghề có thu nhập trung bình, chi tiêu trung bình và tiết kiệm trung bình cao nhất trong tất cả các khối nghề. Cho thấy nhóm nghề này có thu nhập cao, chi tiêu lớn nhưng vẫn duy trì được mức tiết kiệm tốt.

- Nhóm nghành Professional có thu nhập trung bình, mức chi tiêu trung bình cao và mức tiết kiệm trung bình ở mức khá. Có thể thấy nhóm ngành này đang cân bằng và giữ được tỷ lệ tiết kiệm ổn định.
- Ritred là nhóm ngành có mức thu nhập trung bình, chi tiêu ở mức cao vừa và tiết kiệm ở mức thấp. Dựa vào những điểm trên có thể thấy nhóm này có thu nhập khá hạn chế nên mức tiết kiệm ở mức co hẹp so với những nhóm ngành khác.
- Student có mức thu nhập, chi tiêu thấp và mức tiết kiệm thấp nhất. Cho thấy rằng đây là nhóm còn phụ thuộc nhiều vào yếu tố khác, chưa có tích lũy cho bản thân.





Nhóm chi tiêu phổ biến đối với từng độ tuổi:

- Transport và Insurance là hai nhóm chi tiêu phổ biến nhất đối với tất cả các độ tuổi. Trong đó độ tuổi khoảng 46-55 chi tiêu nhiều nhất vào hai nhóm này.
- Rent là nhóm chi tiêu phổ biến đối với độ tuổi 26-35 và 18-25. Đây là hai độ tuổi đang trong giai đoạn lập nghiệp, thuê nhà.
- Độ tuổi 56-65 chi tiêu nhiều vào nhóm Utilities và Healthcare. Đây là chi tiêu hợp lý phù hợp với xu hướng tuổi tác.

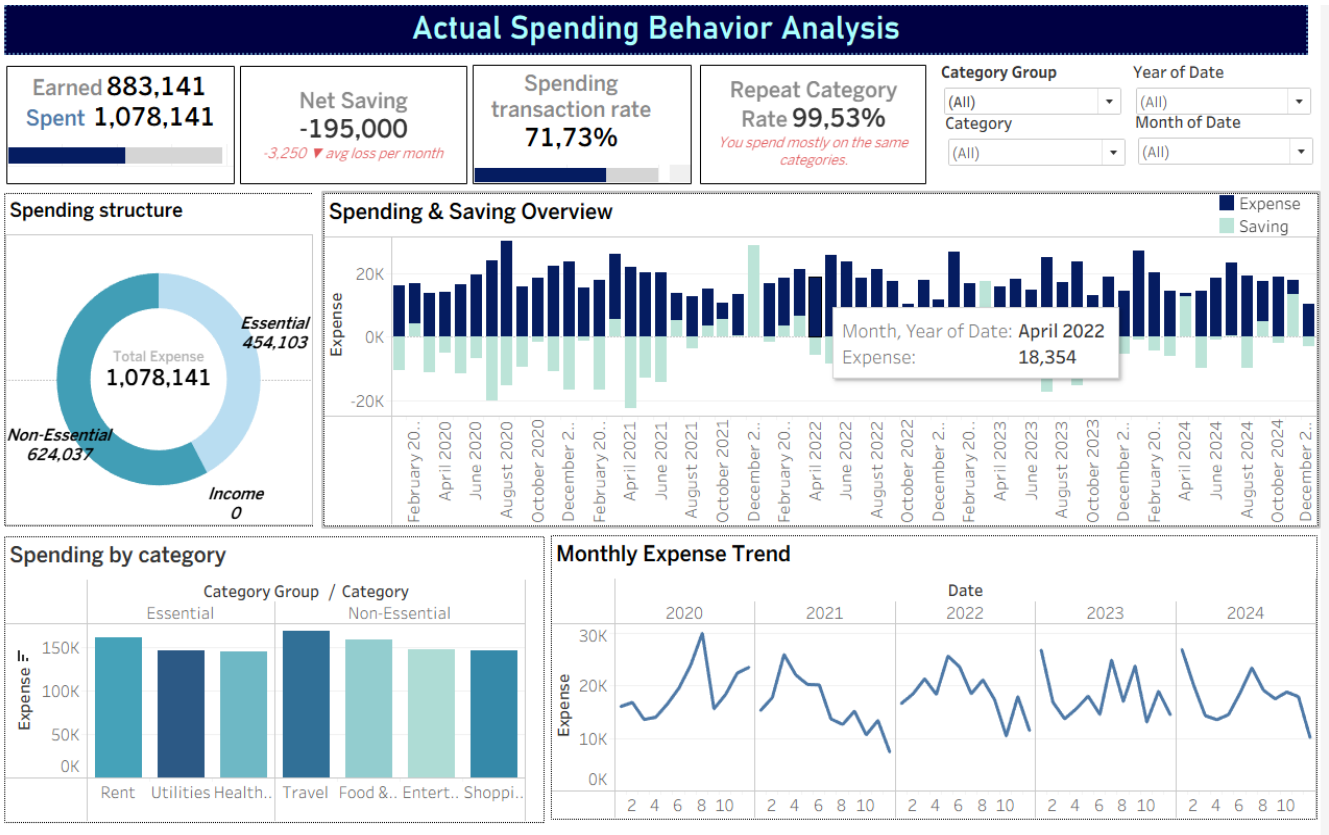
Expense Breakdown by Category

	46-55	36-45	26-35	56-65	18-25
Transport	12,203	10,427	10,669	10,542	5,158
Insurance	11,024	10,774	9,548	8,948	2,609
Rent	3,725	4,700	8,840	2,047	4,887
Utilities	4,886	4,774	3,672	4,681	1,366
Healthcare	4,714	4,432	3,162	4,929	801
Eating Out	3,693	3,650	3,636	2,886	1,395
Entertainment	2,535	3,166	2,486	2,775	849
Education	1,671	1,035	678	1,157	49
Miscellaneous	809	696	516	842	179

Từ những điểm nổi bật trên của dashboard, có thể đưa ra được các hướng insight hành động như:

- Định hướng tiết kiệm:
    - Nhóm tuổi 18-25 và 26-35 nên cần được truyền đạt và giáo dục tài chính sớm để nâng cao tỷ lệ tiết kiệm.
    - Sử dụng các công cụ quản lý tài chính để quản lý tài chính cá nhân. Ví dụ "tự động trích lập" khoảng 10-15% thu nhập mỗi kỳ vào khoảng tiết kiệm cá nhân.
- Tối ưu các danh mục chi tiêu:
    - Transport và Insurance là hai khoản chi tiêu "đòn bẩy" lớn nhất, cần tối ưu hóa. Ví dụ: thay đổi phương tiện đi lại, mua các gói bảo hiểm phù hợp giúp giảm chi phí đáng kể.
    - Các nhóm tuổi 26-35 có thể giảm một phần tiền thuê nhà sang quỹ mua nhà hoặc các quỹ đầu tư dài hạn.

2. Dashboard Personal Finance Dataset



Dashboard: Actual Spending Behavior Analysis

Link dashboard:
https://public.tableau.com/app/profile/nguyen.nhi8170/viz/CuoiKy_17519870918010/Dashboard1?publish=yes

Mục tiêu của Dashboard.

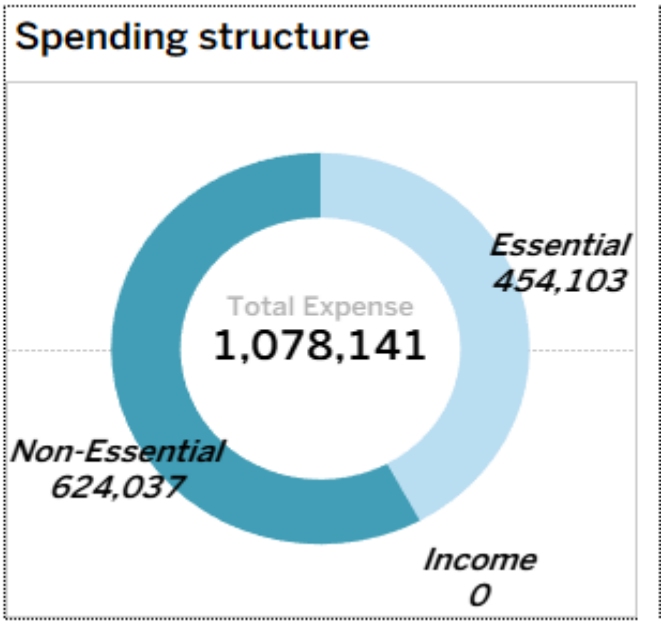
Được thiết kế nhằm phân tích hành vi chi tiêu thực tế của người dùng theo thời gian. Giúp người dùng hiểu rõ hơn về cách họ kiếm tiền và tiêu tiền, cũng như nhận biết các xu hướng tiết kiệm, loại chi tiêu, tỷ lệ chi tiêu theo danh mục

Cấu trúc và thành phần chính.

- **Chỉ số tổng quan KPI**

- Earned (883,141): Tổng thu nhập trong toàn bộ thời gian theo dõi
- Spent (1,078,141): Tổng chi tiêu -> Vượt mức thu nhập khoảng 22%
- Net Saving (-195,000): Hiệu số giữa thu và chi cho thấy người dùng đang chi tiêu vượt quá số tiền kiếm được
- Spendin Transaction Rate (71,73%) : Tỷ lệ giao dịch chi tiêu trên tổng số giao dịch
- Repeat Category Rate (99,53%): Tần suất lặp lại danh mục chi tiêu cho thấy người dùng có xu hướng chi tiêu lặp lại ở các danh mục giống nhau

- **Biểu đồ 1: Spending Structure**

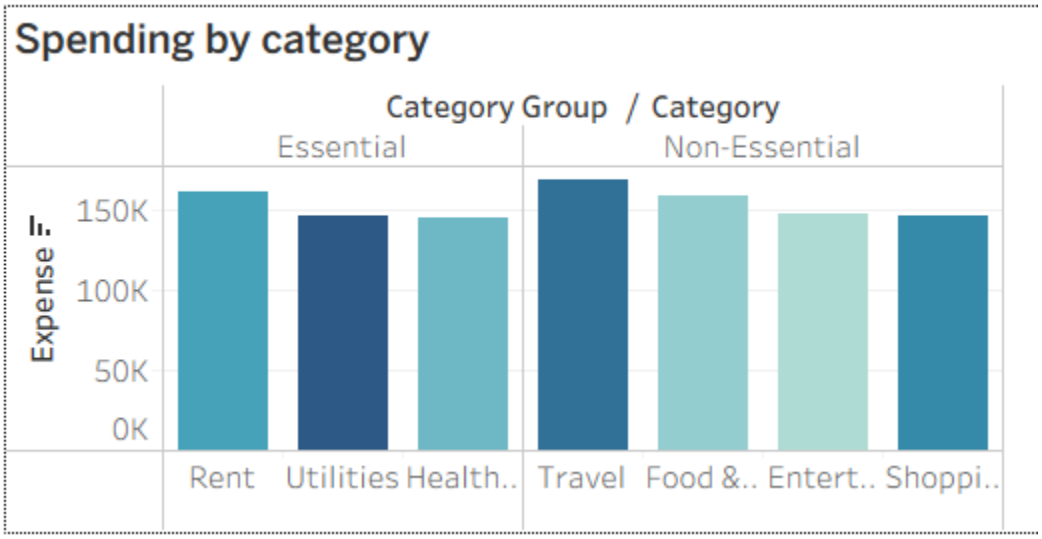


Chi tiêu không thiết yếu (Non – Essential) chiếm tỷ lệ cao hơn ~58% so với chi tiêu thiết yêu (Essential) ~ 42%

Phần lớn chi phí rơi vào nhóm không thiết yêu cho thấy mực độ ưu tiên chi tiêu chưa hợp lý, có khả năng ảnh hưởng đến mục tiêu tiết kiệm dài hạn

➔ Người dùng có thể xem xét vàe cắt giảm các khoản chi tiêu không thiết yếu để cải thiện khả năng tiết kiệm

- **Biểu đồ 2: Spending by Category**



Nhóm Essential (Chi tiêu thiết yếu).

- Gồm 3 danh mục chính: Rent (Thuê nhà), Utilities (Tiện tích), Health (Y tế)
- Chi tiêu cho Reat là cao nhất trong nhóm và cho thấy đây là khoản bắt buộc và chiếm trọng lớn trong ngân sách hàng tháng
- Untilities và Health có mức chi tiêu khác tương đồng, xấp xỉ nhau và chỉ thấp hơn Rent một chút

➔ Tổng thể, các khoản chi tiêu thiết yếu này có chi phí ổn định và khó cắt giảm được cho liên quan đến nhu cầu sinh hoạt cơ bản

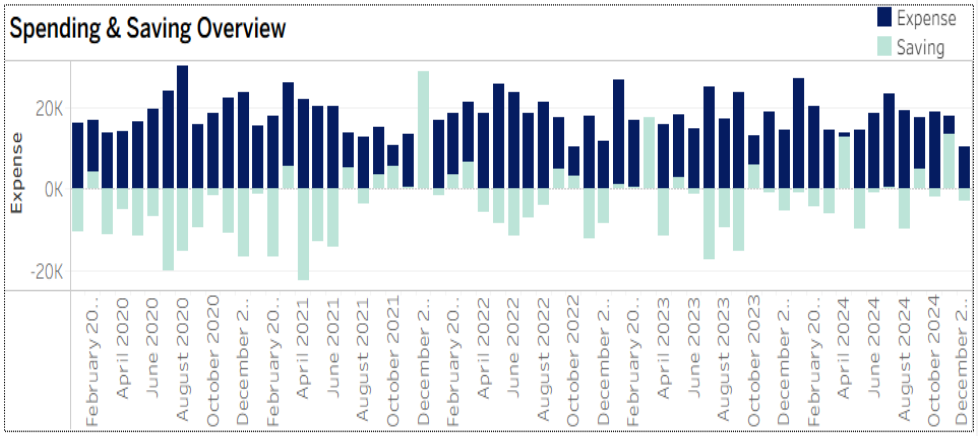
Nhóm Non – Essential (Chi tiêu không thiết yếu)

- Gồm các danh mục như Travel, Food & Drik, Entertainment, Shopping.
- Travel chiếm tỷ trọng cao nhất trong cả nhóm và vượt qua cả tiền thuê nhà -> Điều này cho thấy người dùng có xu hướng ưu tiên trải nghiệm du lịch, mặc dù đây là khoản chi tiêu không bắt buộc.

- Các danh mục Food & Drink, Entertainment và Shopping cũng có mức chi cao tương đương nhau -> Phản ánh được hành vi chi tiêu thiên về hưởng thụ.

➔ Mức chi cho các danh mục không thiết yếu đang ở mức ngang bằng, thậm chí vượt cả chi tiêu thiết yếu -> Cần xem lại mức độ ưu tiên tài chính

- **Biểu đồ 3: Spending và Saving Overview**



Expense diễn ra đều đặn qua toàn bộ giai đoạn.

- Trung bình mỗi tháng chi từ 15K – 25K
- Một số tháng đỉnh điểm như tháng 8/2020, 4/2021, 5/2022, 1/2023, 1/2024. Cho thấy chi tiêu tăng mạnh vào các dịp lễ, Tết

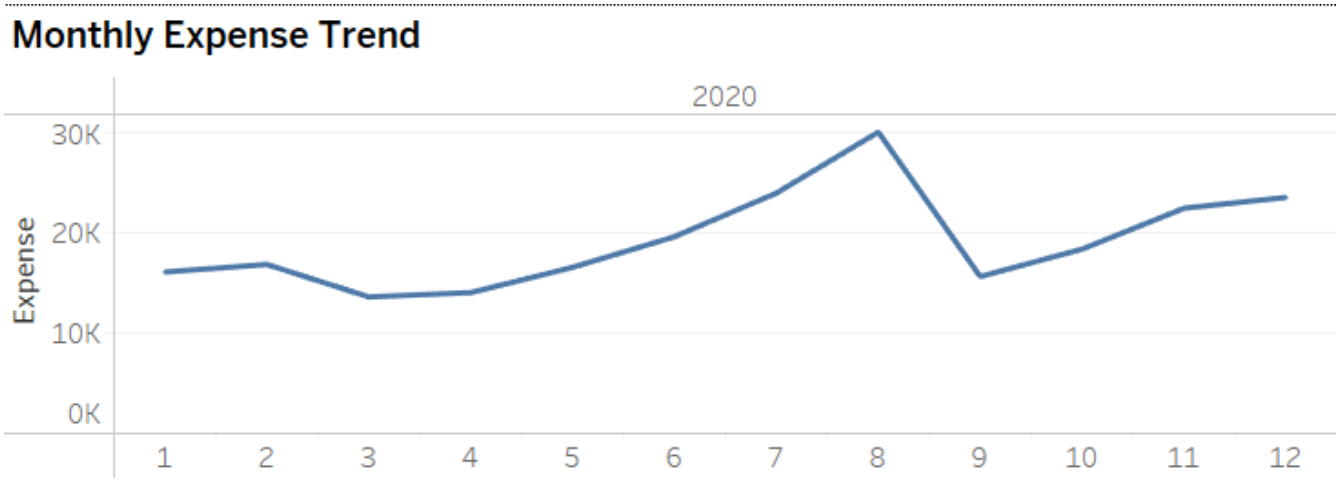
Khoản tiết kiệm (Saving) có xu hướng âm liên tục.

- Hầu hết các tháng đều có Saving âm, chứng tỏ chi tiêu vượt thu nhập thường xuyên
- Ít tháng có Saving dương -> Người dùng không duy trì được thói quen tiết kiệm ổn định. Việc chi tiêu đang phụ thuộc nhiều vào thời điểm, không có ngân sách kiểm soát hiệu quả

➔ Người dùng cần thiết lập hạn mức chi tiêu hàng tháng, đồng thời tăng tỷ lệ tiết kiệm đều đặn thay vì chi tiêu vượt quá mức thu nhập.

- **Biểu đồ 4: Monthly Expense Trend**

  Năm 2020:



Chi tiêu có xu hướng tăng dần từ tháng 3 đến tháng 8:

- Bắt đầu từ mức thấp nhất vào tháng 3 (~13K), sau đó tăng mạnh và đạt đỉnh vào tháng 8 (~30K).
- Điều này cho thấy có thể có các hoạt động đặc biệt, dự án hoặc chi phí bất thường phát sinh vào giữa năm.

Giảm đột ngột vào tháng 9:

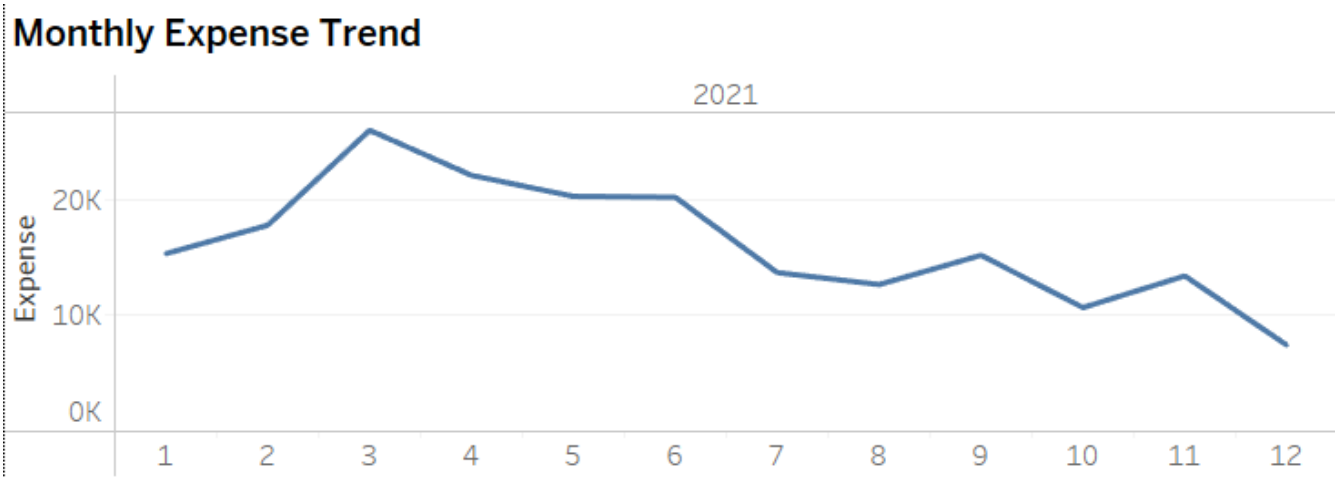
- Chi tiêu rơi mạnh xuống còn khoảng 16K, gần bằng mức của các tháng đầu năm. Có thể do cắt giảm ngân sách, kết thúc dự án, hoặc kiểm soát chi phí.

Tăng nhẹ trở lại từ tháng 10 đến tháng 12:

- Từ khoảng 17K (tháng 10) lên gần 24K (tháng 12), cho thấy một mức độ hồi phục nhưng không bằng mức đỉnh tháng 8.

➔ Năm 2020 cho thấy hành vi chi tiêu bất ổn với sự chênh lệch lớn giữa các tháng cao điểm và thấp điểm

Năm 2021:

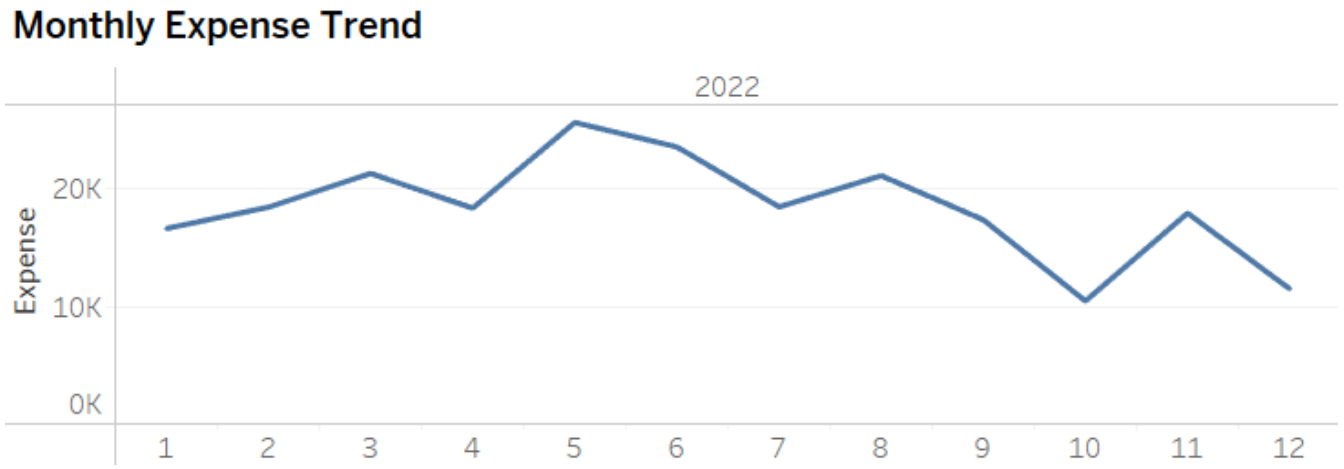


- Chi tiêu tăng mạnh từ tháng 1 đến tháng 3, đạt đỉnh vào tháng 3 (~25K).
- Sau tháng 3, chi tiêu giảm dần đều, đặc biệt rõ rệt từ tháng 6 trở đi.
- Tháng 12 là tháng có chi tiêu thấp nhất trong năm (~9K).

➔ Năm 2021 có xu hướng **thắt chặt chi tiêu** sau quý I

➔ Người dùng đã quản lý chi tiêu tốt theo giai đoạn, tập trung chi cho mùa cao điểm và tiết kiệm về sau

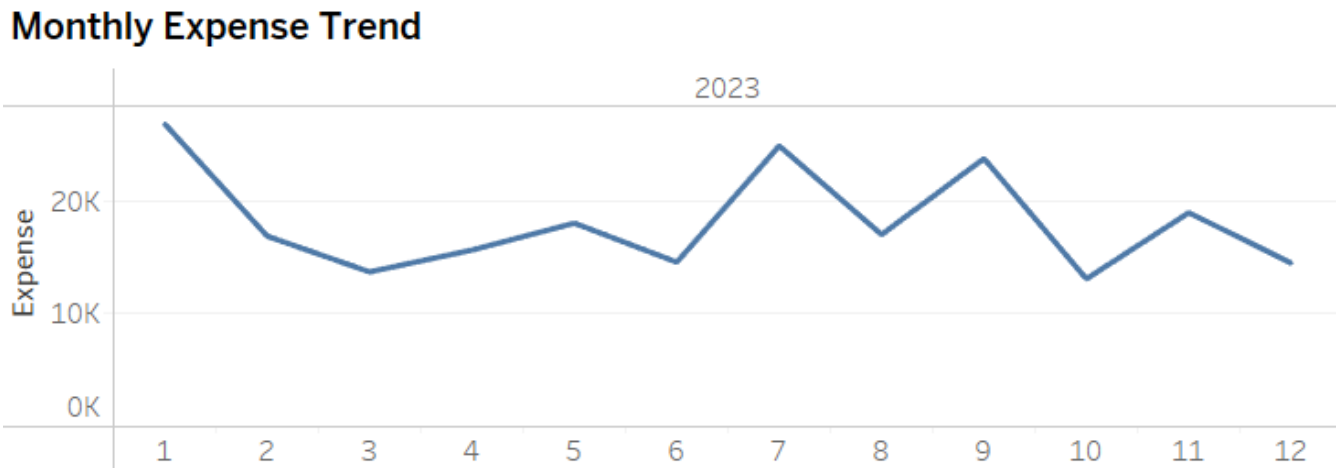
Năm 2022:



- Chi tiêu tăng đều từ tháng 1 đến tháng 3 và đạt đỉnh cao nhất vào tháng 5 (~27K).
- Sau tháng 5, chi tiêu dao động giảm, với một số điểm tăng nhẹ vào tháng 8 và tháng 11.
- Tháng 10 là thời điểm có chi tiêu thấp nhất (~11K).
- Tổng thể năm 2022 có xu hướng dao động nhiều, không ổn định như năm 2021.

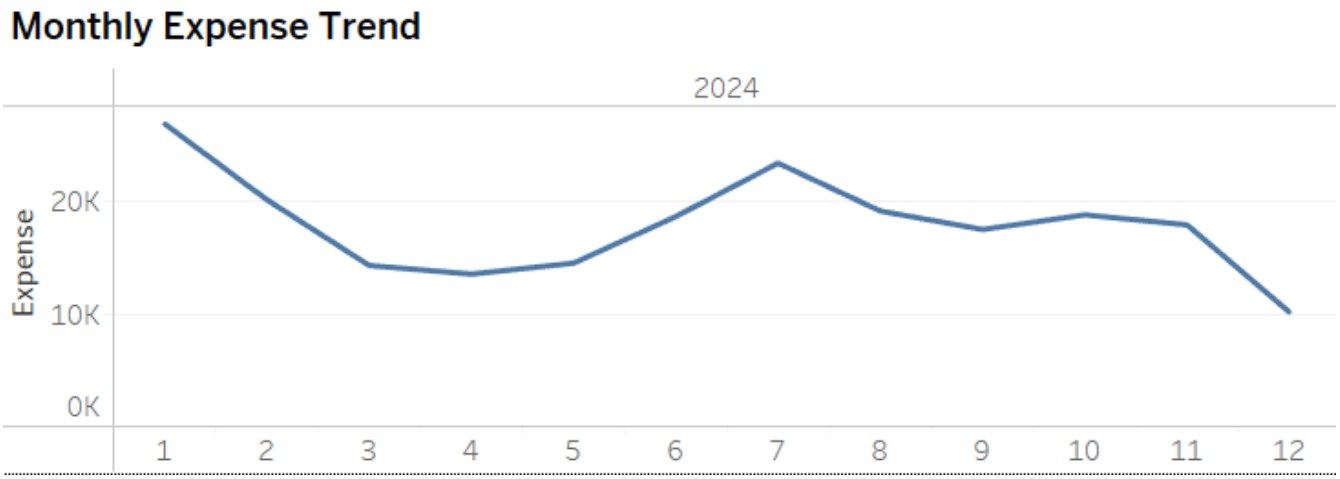
➔ Chi tiêu năm 2022 **không ổn định**, với nhiều đợt tăng giảm xen kẽ. Cần phân tích kỹ các tháng biến động mạnh (tháng 5, 10 và 11) để hiểu rõ nguyên nhân và điều chỉnh kế hoạch ngân sách phù hợp hơn trong tương lai.

Năm 2023:



- Chi tiêu cao nhất vào tháng 1 (~26K), sau đó giảm mạnh vào tháng 2 và chạm đáy ở tháng 3 (~13K).
- Từ tháng 4 đến tháng 12, chi tiêu dao động thất thường với nhiều đợt tăng giảm xen kẽ:
  - Tăng mạnh vào tháng 7 và 9 (~25K),
  - Giảm sâu vào tháng 10 (~13K).

Năm 2024:



- Tháng 1 có chi tiêu cao nhất (~26K), sau đó giảm mạnh liên tục đến tháng 3 (~14K).
- Chi tiêu bắt đầu tăng trở lại từ tháng 4, đạt đỉnh thứ hai vào tháng 7 (~24K).
- Từ tháng 8 trở đi, chi tiêu giảm dần và chạm đáy vào tháng 12 (~10K).

➔ Khác với năm 2023 có chi tiêu rất cao đầu năm, năm 2024 có vẻ như người tiêu dùng đã có sự điều chỉnh, bắt đầu năm với chi tiêu được kiểm soát hơn.

➔ Tháng 6-7 là mùa cao điểm của các hoạt động du lịch, nghỉ hè. Đây có thể là động lực chính khiến chi tiêu tăng vọt cho các chuyến đi, ăn uống, vui chơi giải trí. Mặc dù Tháng 12 là mùa lễ hội, nhưng chi tiêu lại thấp. Điều này có thể cho thấy người tiêu dùng đang **tiết kiệm một cách có chủ đích** cho các khoản chi lớn hơn vào cuối năm (như mua sắm Tết Nguyên Đán cho năm kế tiếp, hoặc các khoản đầu tư cuối năm).

Tổng kết tình hình chi tiêu qua các năm:

- Người dùng có thu nhập ổn định nhưng chi tiêu thường vượt mức và thiếu tiết kiệm.
- Cơ cấu chi tiêu chưa hợp lý, với gần 60% tổng chi dành cho các danh mục không thiết yếu.
- Chi tiêu cá nhân trong giai đoạn 2020-2024 cho thấy một mô hình **chu kỳ biến động theo mùa/lễ hội**, với các đỉnh chi tiêu tập trung vào đầu năm, giữa năm (mùa hè) và cuối năm. Tuy nhiên, hành vi chi tiêu tổng thể **thiếu kỷ luật và ổn định**, thường xuyên xảy ra tình trạng "quá đà" dẫn đến phải thắt chặt đột ngột. Năm 2023 và 2024 cho thấy có dấu hiệu của sự **điều chỉnh và kiểm soát chi tiêu** sau các giai đoạn chi tiêu cao điểm, nhưng nhìn chung vẫn còn nhiều dư địa để cải thiện sự ổn định tài chính.

3. Mối liên hệ giữa hai dashboard

Dashboard: Income, Spending, and Saving Overview by Demographic Group:

- Cung cấp cái nhìn toàn cảnh về thu nhập – chi tiêu – tiết kiệm của các nhóm nhân khẩu học (tuổi, nghề, vùng)
- Mặc dù thu nhập trung bình khá cao (74,503 USD), nhưng mức tiết kiệm vẫn thấp (~11%) và nhiều nhóm có xu hướng chi tiêu gần bằng thu nhập.

Dashboard: Actual Spending Behavior Analysis:

- Qua phân tích cho thấy một cá nhân hoặc một hộ gia đình cụ thể đang gặp tình trạng chi tiêu vượt mức thu nhập với mức lỗ 195,000 USD trong 5 năm
- 58% chi tiêu là không thiết yếu, và 99,53% chi tiêu lặp lại theo thói quen và chưa tối ưu theo mục tiêu tài chính.
- Dựa vào những thông tin trên có thể thấy dù có thu nhập cao, nhưng hành vi chi tiêu lặp lại và tập trung vào các khoản không thiết yếu là nguyên nhân khiến nhiều cá nhân không tích lũy được tài sản – thậm chí chi tiêu vượt mức thu nhập.

Để hạn chế tình trạng này xảy ra cần có giải pháp như:

Mỗi cá nhân cần tự động hóa kiểm soát chi tiêu và sửa đổi thói quen tài chính của mình: Lập một mô hình phân bổ rõ ràng cho từng mức chi tiêu, tiết kiệm. Giới hạn rõ từng mức chi cho các nhu cầu không thiết yếu như du lịch, shopping và giải trí.

Ngoài ra, đối với các doanh nghiệp phát triển công cụ tài chính có thể tạo ra các công cụ tài chính tự động giúp người dùng có thể phân bổ rõ ràng từng mức chi thu của mình. Và tạo ra các xu hướng marketing quảng cáo sản phẩm tiết kiệm dựa trên các phân khúc độ tuổi đặc biệt là độ tuổi từ 18-36.

TÀI LIỆU THAM KHẢO

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)

BẢNG PHÂN CÔNG NHIỆM VỤ

Họ tên	Nhiệm vụ	Mức độ hoàn thành
Nguyễn Thị Nhật Liên	Xử lý dữ liệu, báo cáo, thuyết trình	100%

Đoàn Thị Diệu Linh	Báo cáo, Slide	100%
Nguyễn Phương Nhi	Dashboard	100%
Nguyễn Phan Hoàng Phúc	Dashboard	100%