

****

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KĨ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KỲ**

**MÔN HỌC: TƯƠNG TÁC DỮ LIỆU TRỰC QUAN**

**ĐỀ TÀI: TRỰC QUAN VÀ PHÂN TÍCH HÀNH VI KHÁCH HÀNG**

**Giảng viên hướng dẫn: Lê Quang Thái**

**Mã môn học: 222IDVI333677**

**Người thực hiện: Nguyễn Thanh Hùng MSSV: 20133045**

**Lê Hoàng Khang MSSV: 20133050**

**Vũ Trung Kiên MSSV: 20133060**

**Tp. Hồ Chí Minh, tháng 5 năm 2023**

**MỤC LỤC**

[**1.** **Tóm tắt (Abstract):** 3](#_Toc134563011)

[**2.** **Giới thiệu (Introduction):** 3](#_Toc134563012)

[**3.** **Dữ liệu (Data):** 4](#_Toc134563013)

[**4.** **Trực quan hóa dữ liệu (Data Visulization):** 8](#_Toc134563014)

[**5.** **Xây dựng mô hình** 13](#_Toc134563015)

[**5.1.** **Sử dụng PCA để giảm chiều của dữ liệu:** 13](#_Toc134563016)

[**5.2.** **Plot PCA với 3 thành phần:** 14](#_Toc134563017)

[**5.3.** **Thuật toán phân cụm**: 14](#_Toc134563019)

[**6.** **Nhận xét và kết luận (Comment, Conclusions):** 17](#_Toc134563021)

[**6.1.** **Sử dụng thuật toán Kmeans** 17](#_Toc134563022)

[**6.2.** **Phân cụm dựa trên thu nhập và chi tiêu:** 18](#_Toc134563025)

[**6.3.** **Phân cụm dựa trên chi tiêu của sản phẩm:** 18](#_Toc134563028)

[**6.4.** **Phân cụm dựa trên tổng quảng cáo** 19](#_Toc134563033)

[**6.5.** **Tổng kết (Summary):** 20](#_Toc134563038)

[**7.** **Đóng góp (Contributions):** 22](#_Toc134563059)

**Report Guides**

1. **Tóm tắt (Abstract):**

Đây là bài report của nhóm chúng tôi về phân tích tập dữ liệu "Marketing Data" trên Kaggle. Tập dữ liệu này cung cấp thông tin về chiến lược marketing, hành vi khách hàng và hiệu quả các chiến dịch quảng cáo của một công ty thương mại điện tử.

Trong bài report này, chúng tôi sẽ trình bày về các kết quả phân tích của chúng tôi về tập dữ liệu này. Chúng tôi đã sử dụng các kỹ thuật trực quan hóa dữ liệu để giải thích và minh họa các kết quả phân tích của mình. Bên cạnh đó, chúng tôi cũng đã áp dụng các phương pháp thống kê và máy học để trả lời các câu hỏi liên quan đến tập dữ liệu này.

Những nội dung chính của bài report bao gồm:

Đánh giá các yếu tố ảnh hưởng đến hành vi khách hàng.

Phân tích hành vi và thuộc tính khách hàng để tìm kiếm những nhóm khách hàng tiềm năng.

Đánh giá hiệu quả của các chiến dịch quảng cáo và đưa ra các đề xuất cải thiện.

Hy vọng bài report của chúng tôi sẽ cung cấp cho bạn những thông tin hữu ích và giúp bạn hiểu rõ hơn về tập dữ liệu Marketing Data này.

1. **Giới thiệu (Introduction):**

Các câu hỏi nghiên cứu của nhóm:

* Trình độ học vấn có ảnh hưởng đến thu nhập cá nhân hay không?
* Các độ tuổi khác nhau có ảnh hưởng đến việc mua của từng loại sản phẩm hay không?
* Sức mua hàng của cá nhân có bị ảnh hưởng bởi quốc gia hay không?
* Sự tương quan giữa thu nhập và chi tiêu dựa trên số trẻ con trong gia đình có khác biệt nhau như thế nào?
* Dựa vào những câu hỏi trên, nhóm chúng em đặt ra vấn đề về phân loại khách hàng dựa trên các yếu tố liên quan.

1. **Dữ liệu (Data):**

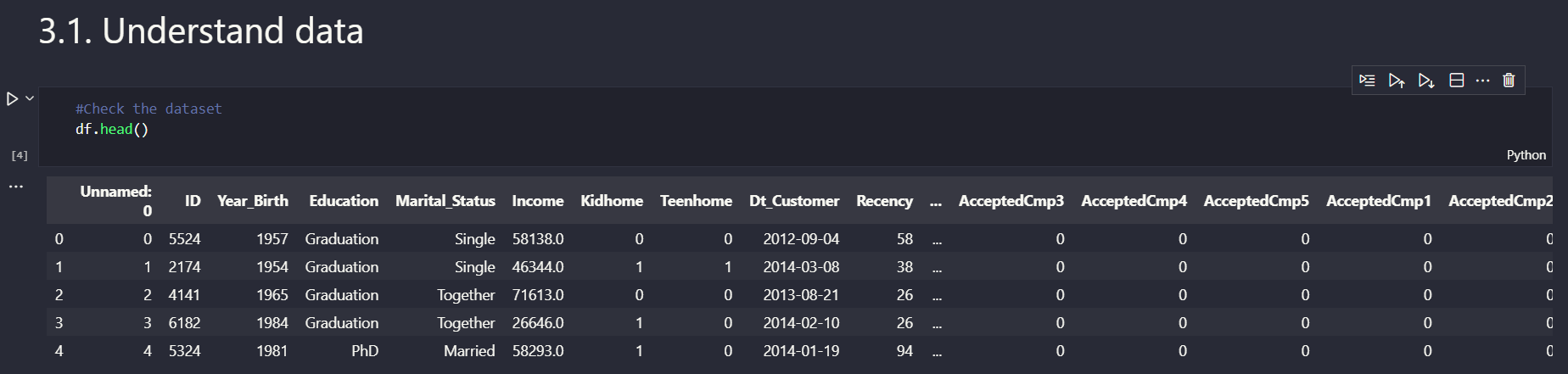
Nguồn dữ liệu thu thập: https://www.kaggle.com/datasets/jackdaoud/marketing-data

Tập dữ liệu mà nhóm chúng em chọn có tên là "Marketing Data" và được lưu trữ trên trang Kaggle. Tập dữ liệu này bao gồm thông tin về các chiến dịch tiếp thị của một công ty trong vòng 3 năm (2012-2014). Tập dữ liệu này bao gồm các biến sau:

* ID: Mã số khách hàng.
* Year\_Birth: Năm sinh khách hàng.
* Education: Trình độ học vấn của khách hàng.
* Marital\_Status: Tình trạng hôn nhân của khách hàng.
* Income: Thu nhập hàng năm của khách hàng.
* Kidhome: Số lượng trẻ em trong gia đình của khách hàng dưới 18 tuổi.
* Teenhome: Số lượng trẻ em trong gia đình của khách hàng từ 18 đến 25 tuổi.
* Dt\_Customer: Ngày đăng ký thành viên của khách hàng.
* Recency: Số ngày kể từ khi khách hàng mua sản phẩm của công ty lần cuối cùng.
* MntWines: Số tiền khách hàng đã chi tiêu cho rượu vang trong 2 năm qua.
* MntFruits: Số tiền khách hàng đã chi tiêu cho các loại trái cây trong 2 năm qua.
* MntMeatProducts: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ thịt trong 2 năm qua.
* MntFishProducts: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ hải sản trong 2 năm qua.
* MntSweetProducts: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ kẹo và đồ ngọt trong 2 năm qua.
* MntGoldProds: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ vàng, bạc và kim cương trong 2 năm qua.
* NumDealsPurchases: Số lượng giao dịch mà khách hàng đã tham gia với giá khuyến mãi trong 2 năm qua.
* NumWebPurchases: Số lượng sản phẩm mà khách hàng đã mua trên trang web của công ty trong 2 năm qua.
* NumCatalogPurchases: Số lượng sản phẩm mà khách hàng đã mua thông qua các catalog trong 2 năm qua.
* NumStorePurchases: Số lượng sản phẩm mà khách hàng đã mua trực tiếp tại cửa hàng của công ty trong 2 năm qua.
* NumWebVisitsMonth: Số lượng truy cập trung bình của khách hàng trên trang web của công ty trong một tháng.

Các biến này được sử dụng để phân tích hành vi tiêu dùng của khách hàng và thiết kế các chiến dịch tiếp thị hiệu quả.

Thực hiện in vài dòng trong tập dữ liệu:

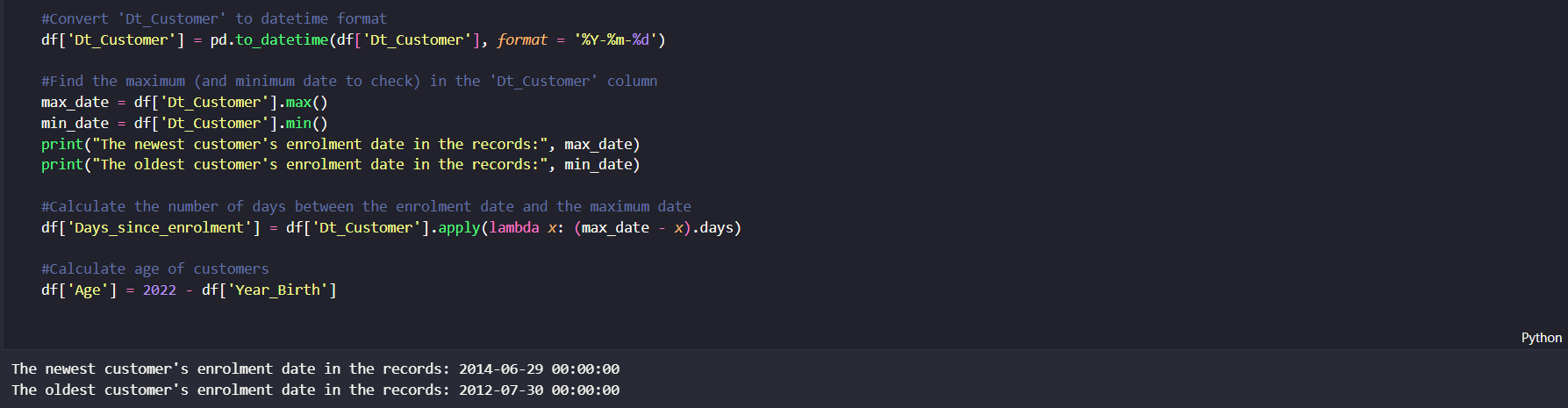


Tiền xử lý dữ liệu:

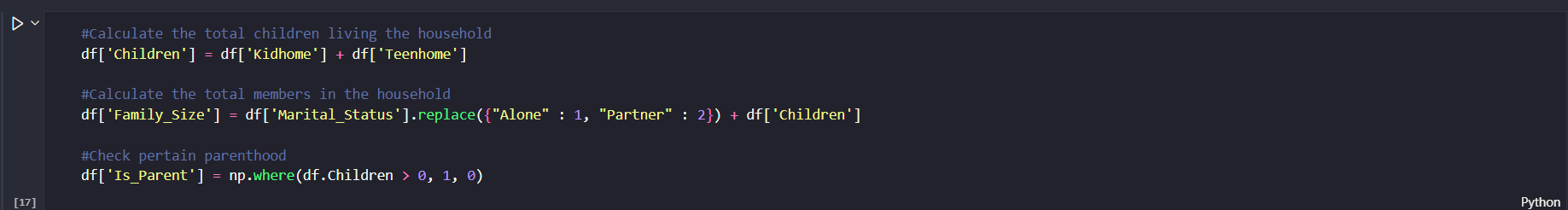
* Đổi tên các biến cho rõ ràng:

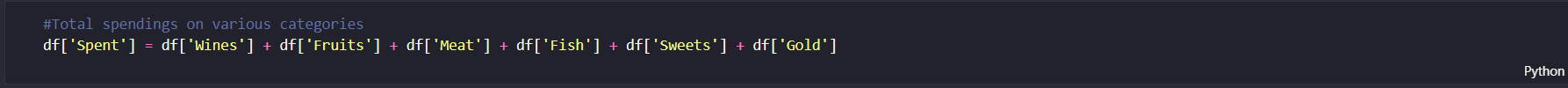


* Làm sạch dữ liệu theo các bước sau:
  + Tạo một thuộc tính mới gọi là "Days\_since\_enrolment" để cho biết số ngày kể từ khi khách hàng bắt đầu mua sắm tại siêu thị.
  + Trích xuất "Age" của khách hàng bằng cách sử dụng thuộc tính "Year\_Birth" của họ.
  + Tạo một thuộc tính mới có tên là "Spent" để cho biết tổng số tiền mà một khách hàng đã chi tiêu cho nhiều danh mục khác nhau trong khoảng thời gian 2 năm.
  + nhóm thuộc tính "Mariage\_Status" thành hai loại: "alone" và "partner".
  + Tạo một thuộc tính mới gọi là "children" để cho biết tổng số trẻ em trong một hộ gia đình, bao gồm cả trẻ em và thanh thiếu niên.
  + Tạo một thuộc tính mới gọi là "Family\_Size" để làm rõ hơn quy mô của một hộ gia đình.
  + Tạo một thuộc tính mới gọi là "Is\_Parent" để cho biết khách hàng có phải là cha mẹ hay không.
  + Đơn giản hóa thuộc tính "Education" bằng cách nhóm các giá trị của nó thành ba loại.
  + Loại bỏ một số biến dư thừa không cần thiết cho phân tích của nhóm.



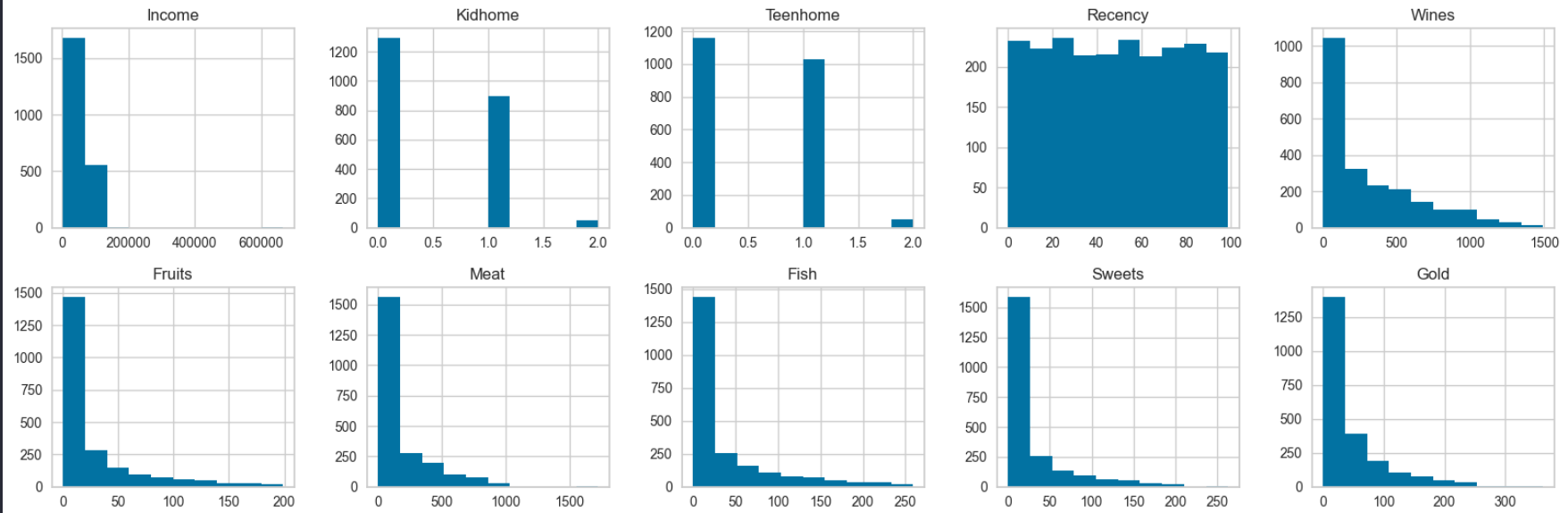


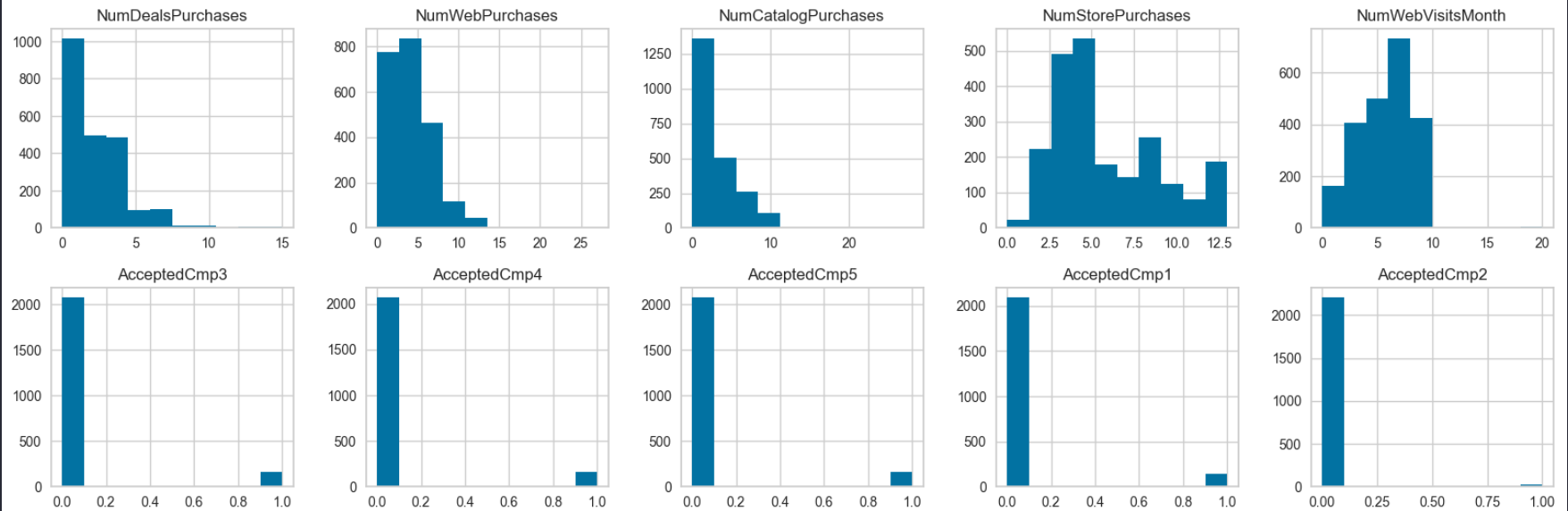


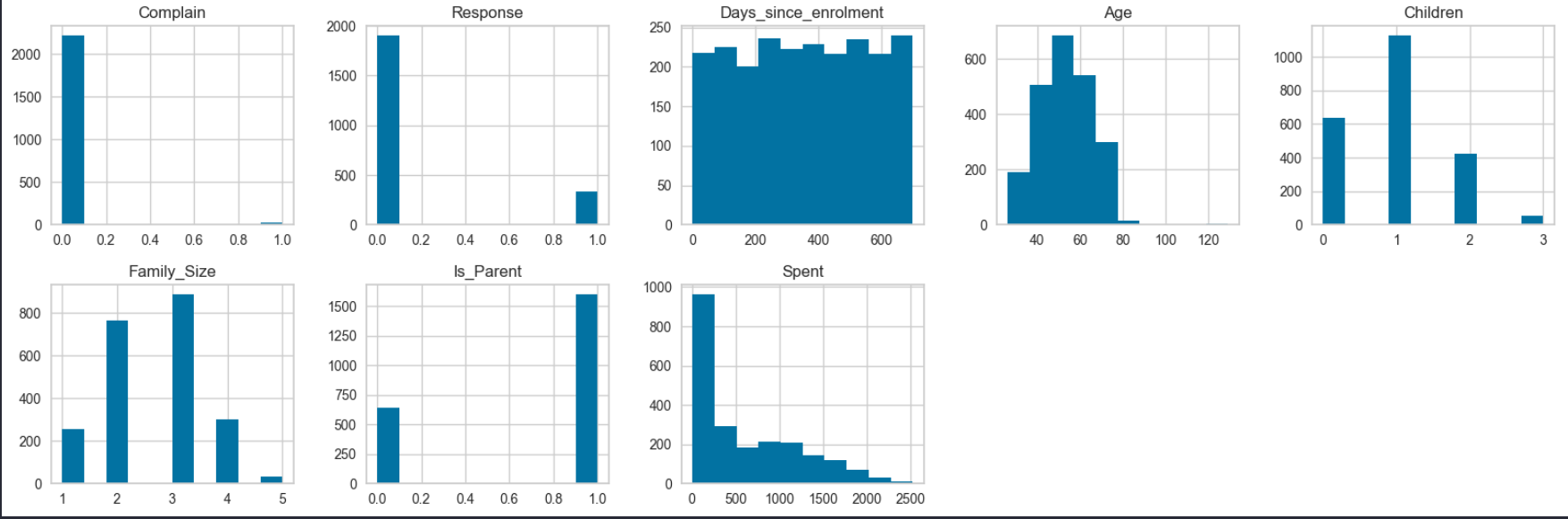


1. **Trực quan hóa dữ liệu (Data Visulization):**

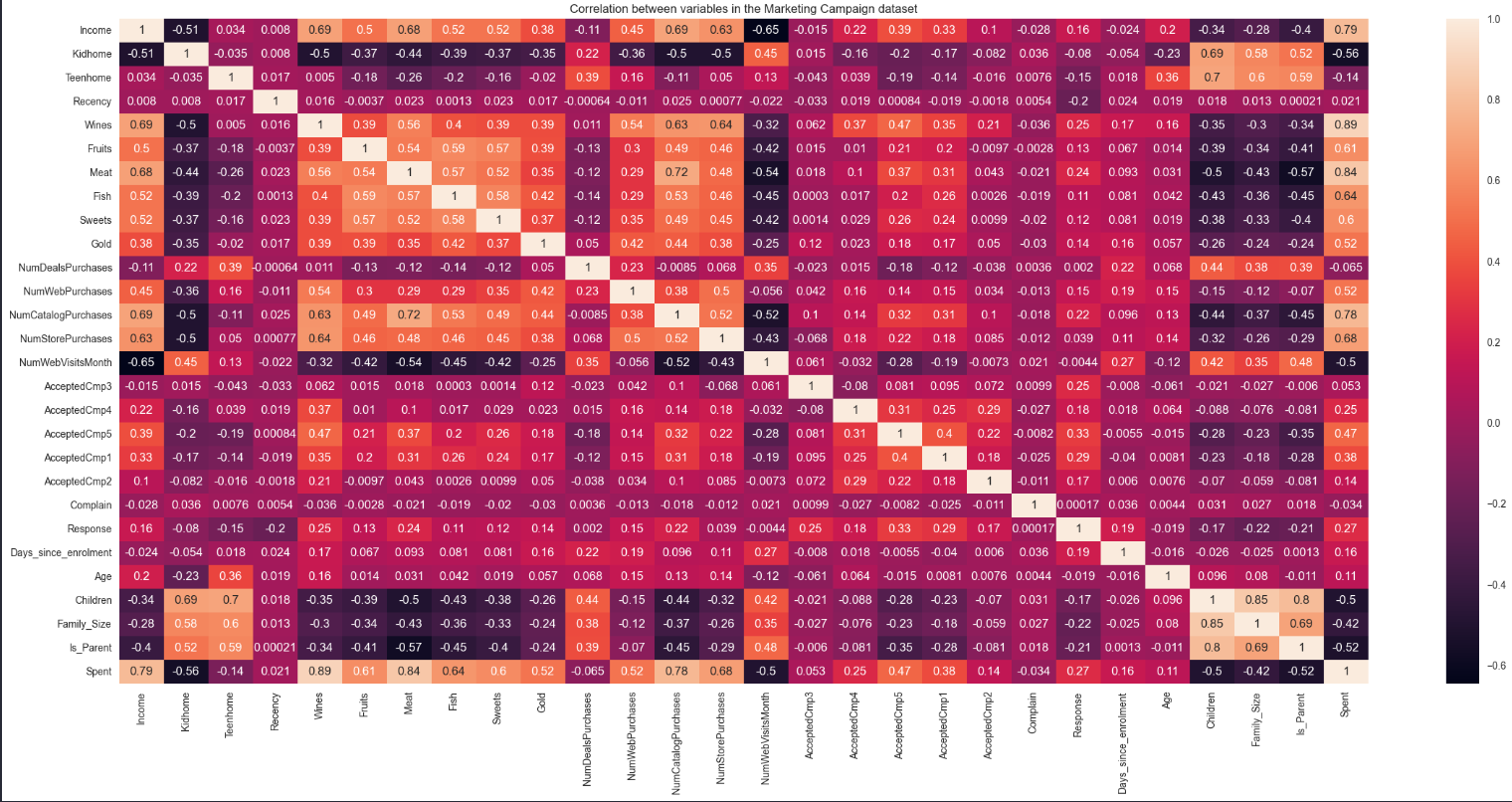
* Biểu đồ Histogram cho các biến số:

****

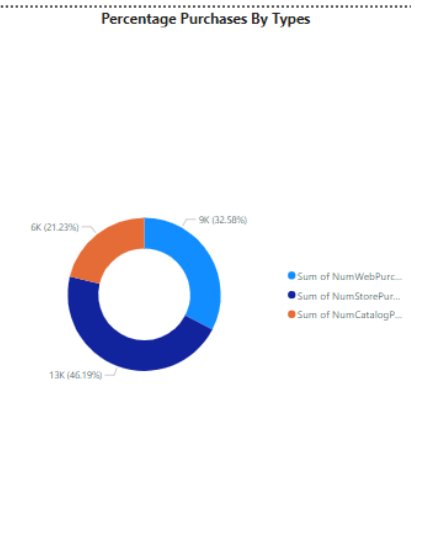
****

****

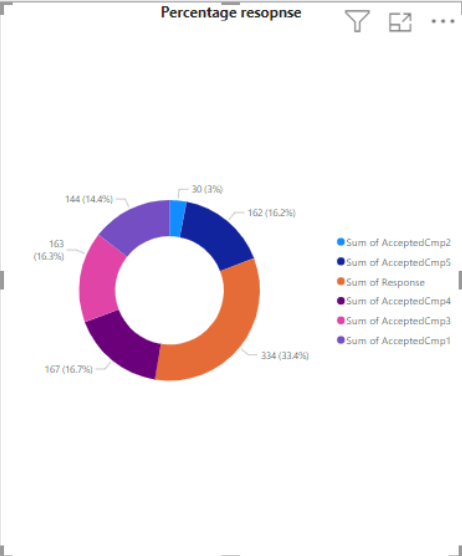
* Biểu đồ thể hiện độ tương quan giữa các biến trong dữ liệu

****

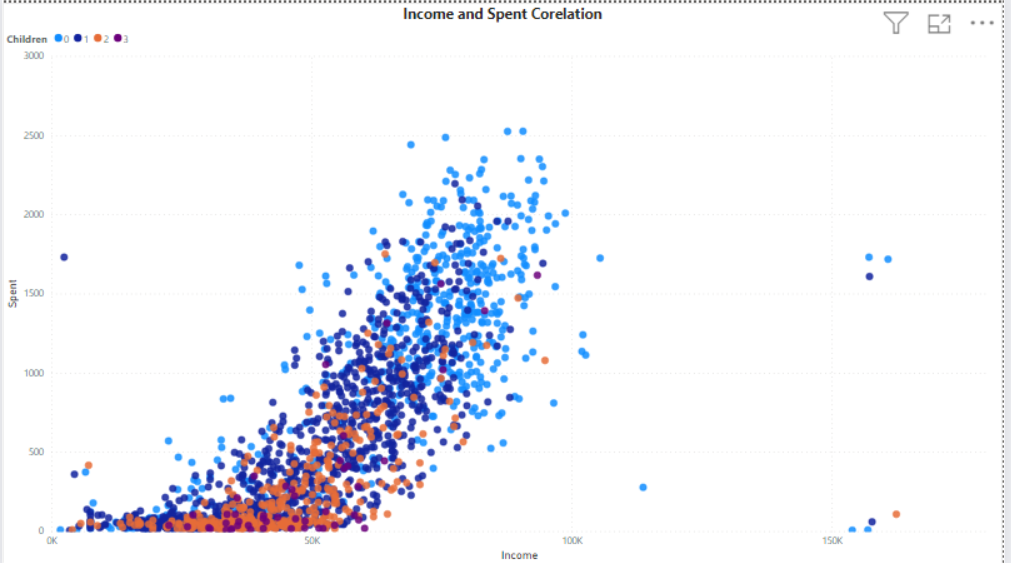
* Biểu đồ biểu thị phần trăm phân bố dữ liệu của các kiểu đặt hàng

****

* Biểu đồ biểu thị phần trăm phân của sự chấp nhận các chiến dịch marketing



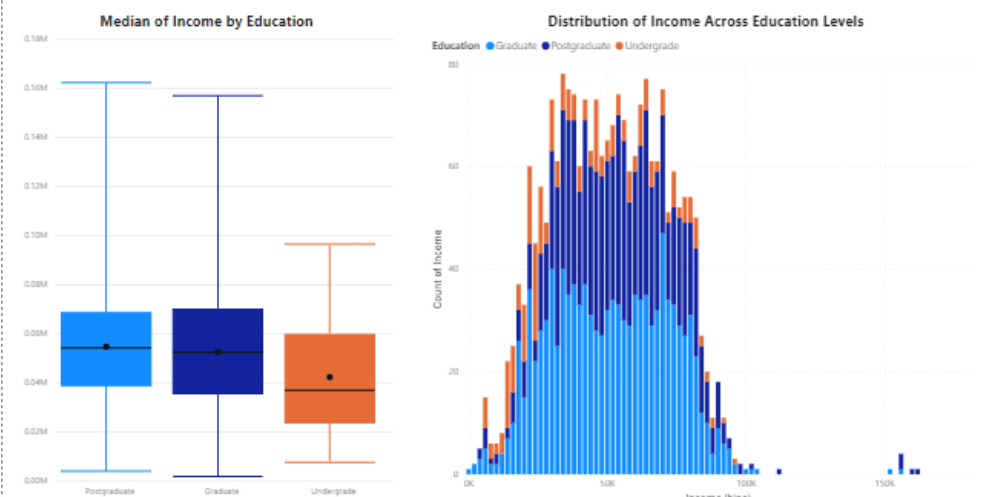
* Biểu đồ thể hiện sự tương quan giữa thu nhập và chi tiêu



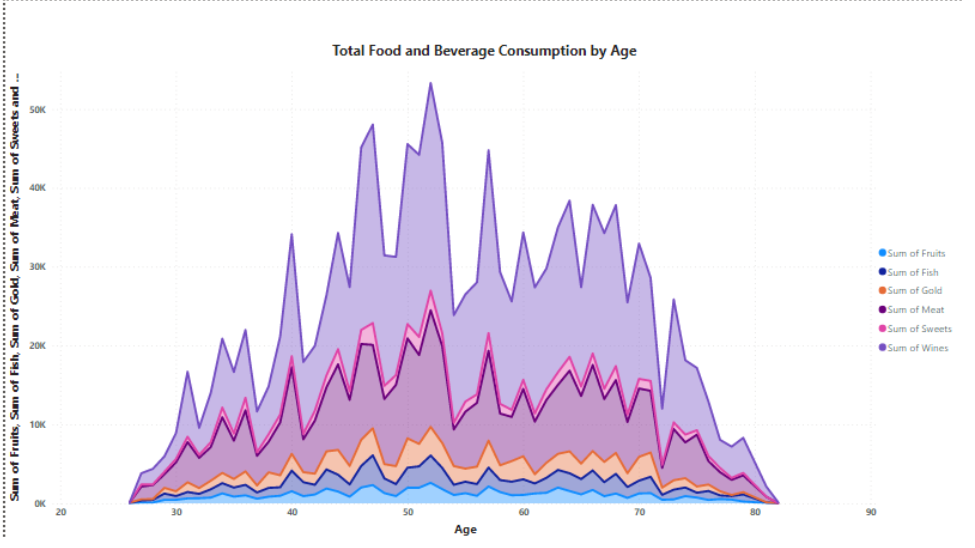
* Các biểu đồ cột thể hiện sự phân bố của các biến Age, Education, Mariage\_Status, Children



* Biểu đồ boxplot và cột cho thu nhập dựa trên học vấn



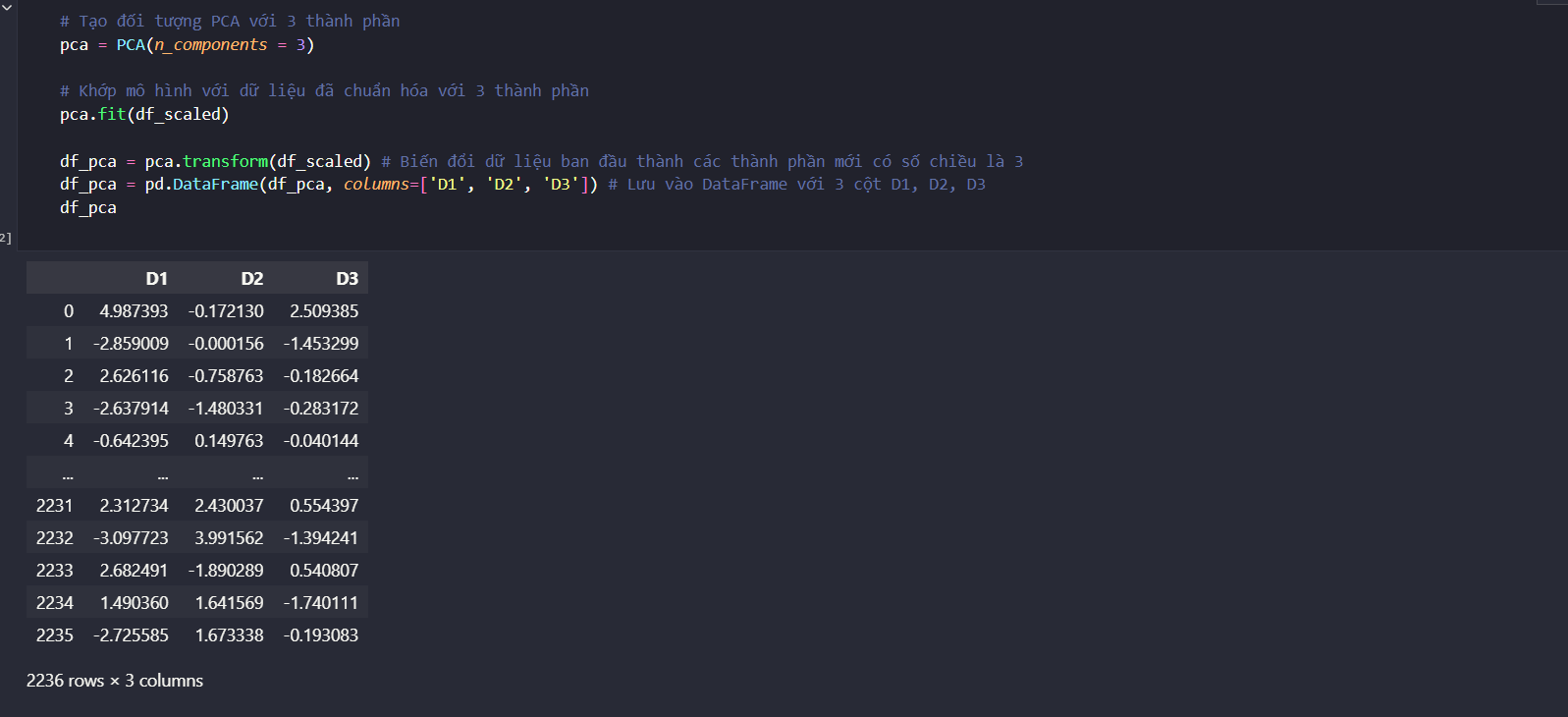
* Biểu đồ đường cho chi tiêu của các loại sản phẩm dựa trên độ tuổi



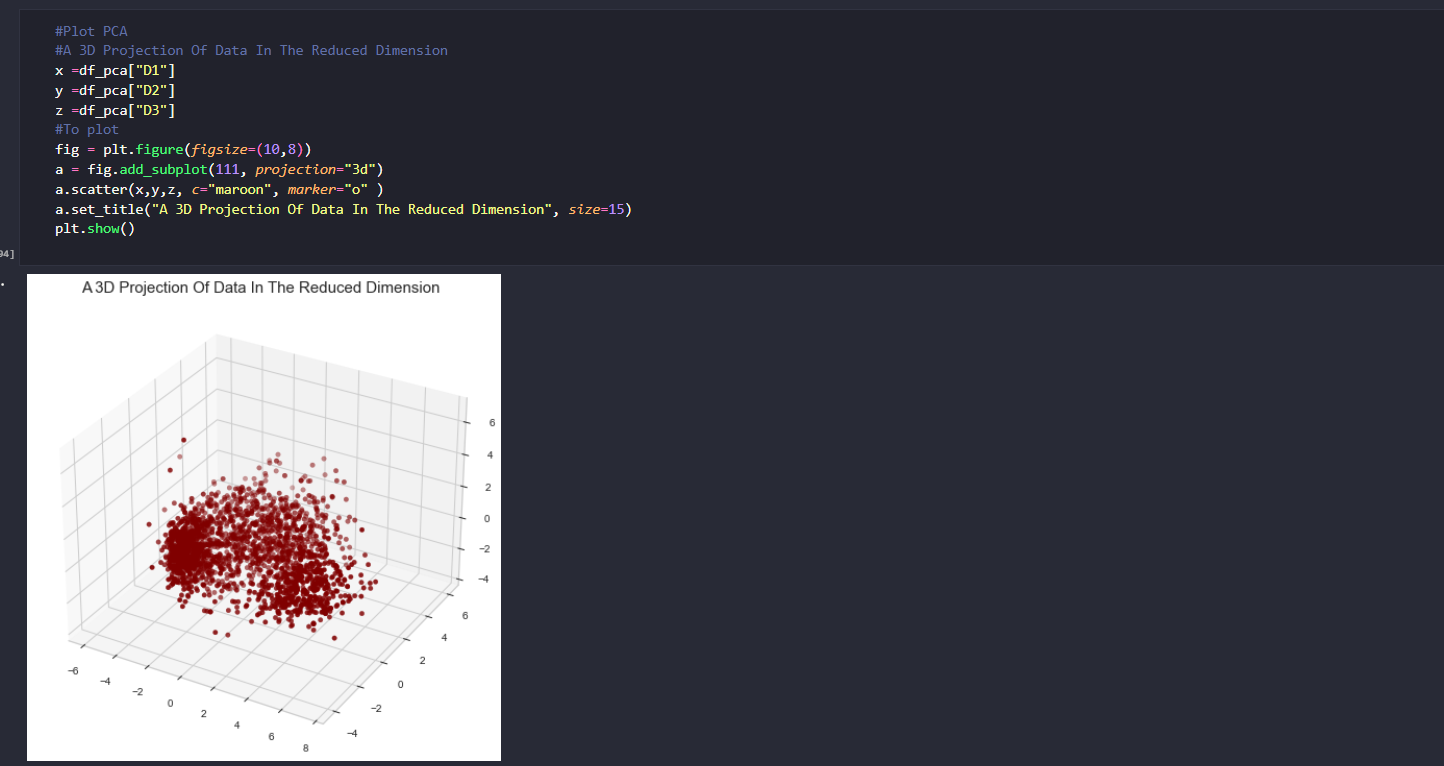
1. **Xây dựng mô hình**
   1. **Sử dụng PCA để giảm chiều của dữ liệu:**

PCA (Principal Component Analysis) là một phương pháp phân tích dữ liệu thống kê để giảm số chiều của dữ liệu mà không mất đi sự đa dạng của các thuộc tính. Nó thực hiện việc biến đổi không gian của dữ liệu ban đầu bằng cách tìm ra các trục chính (principal components) để giảm sự phụ thuộc giữa các thuộc tính. Các trục chính này tương ứng với các thành phần quan trọng nhất trong dữ liệu và có thể được sử dụng để xây dựng mô hình, tìm kiếm các mẫu và tạo ra các biểu đồ dữ liệu hiệu quả hơn.

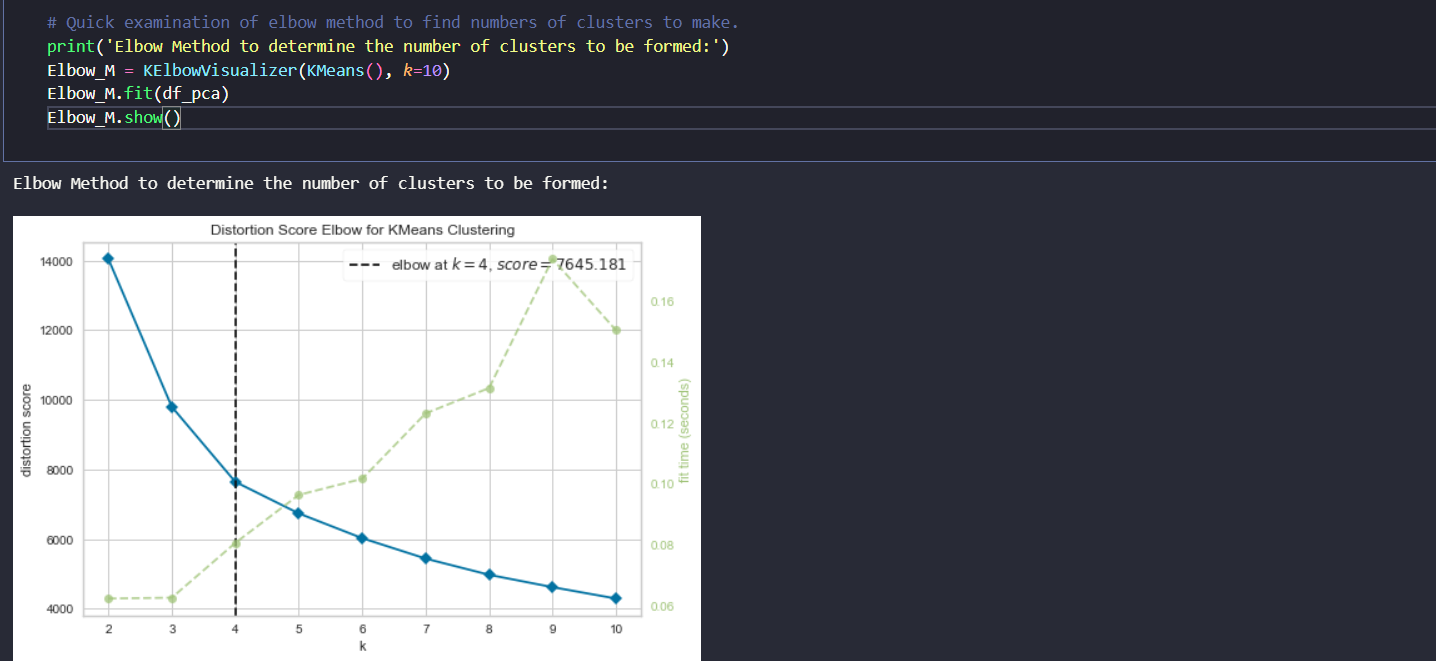
Trong dữ liệu của nhóm chúng em sau khi tiền xử lý có rất nhiều thuộc tính. Vì vậy nhóm chúng em đã quyết định sử dụng PCA vào tệp dữ liệu của nhóm



* 1. **Plot PCA với 3 thành phần:**



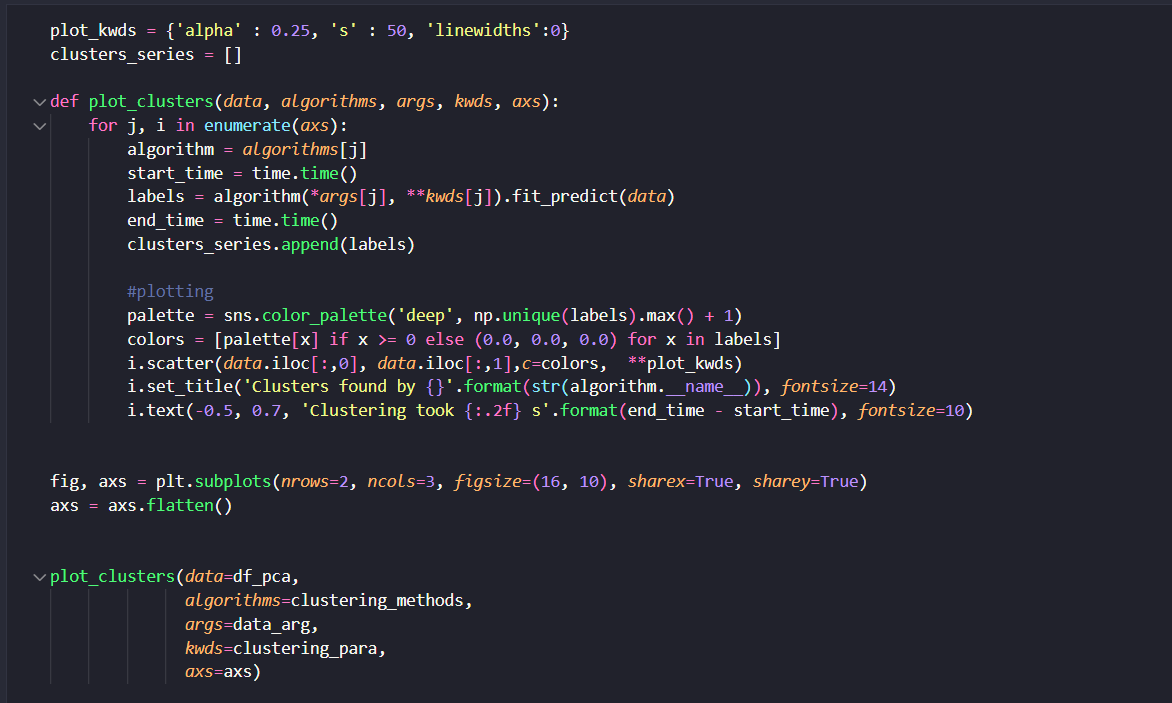
* 1. **Thuật toán phân cụm**Tìm số cụm bằng phương pháp elbow:



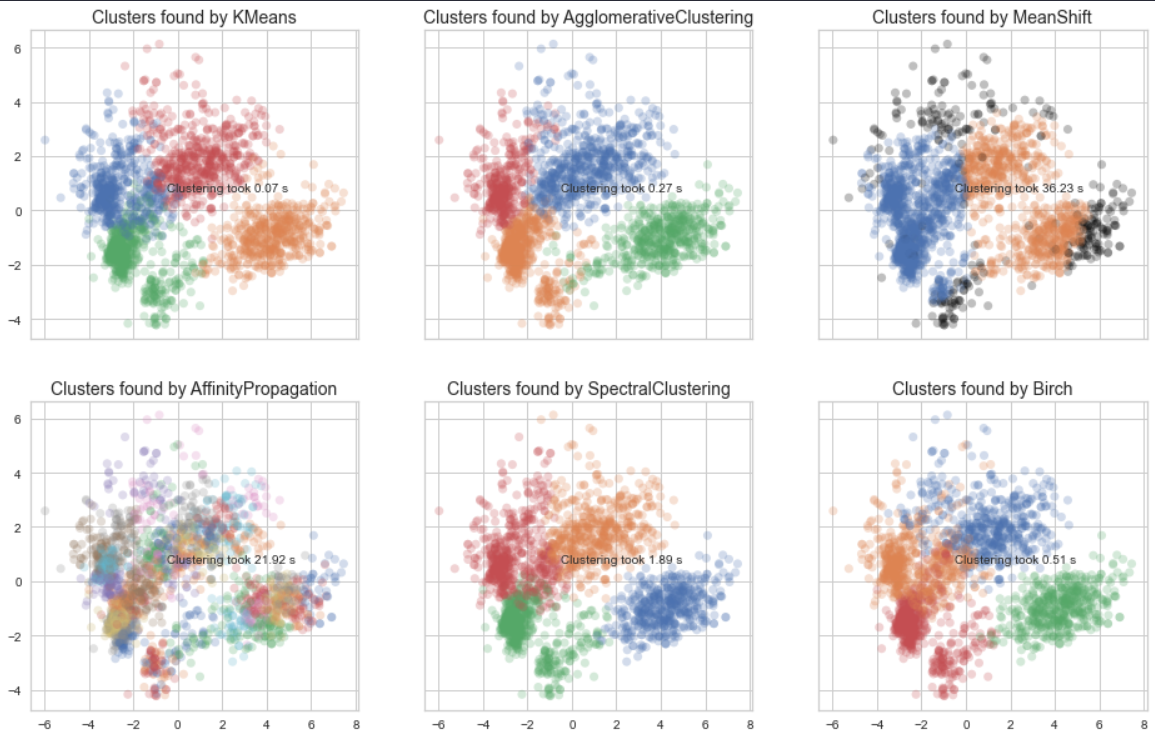
Nhìn vào biểu đồ, ta có thể dễ dàng chọn số cụm là 4.

Tìm kiếm thuật toán phân cụm tốt nhất:

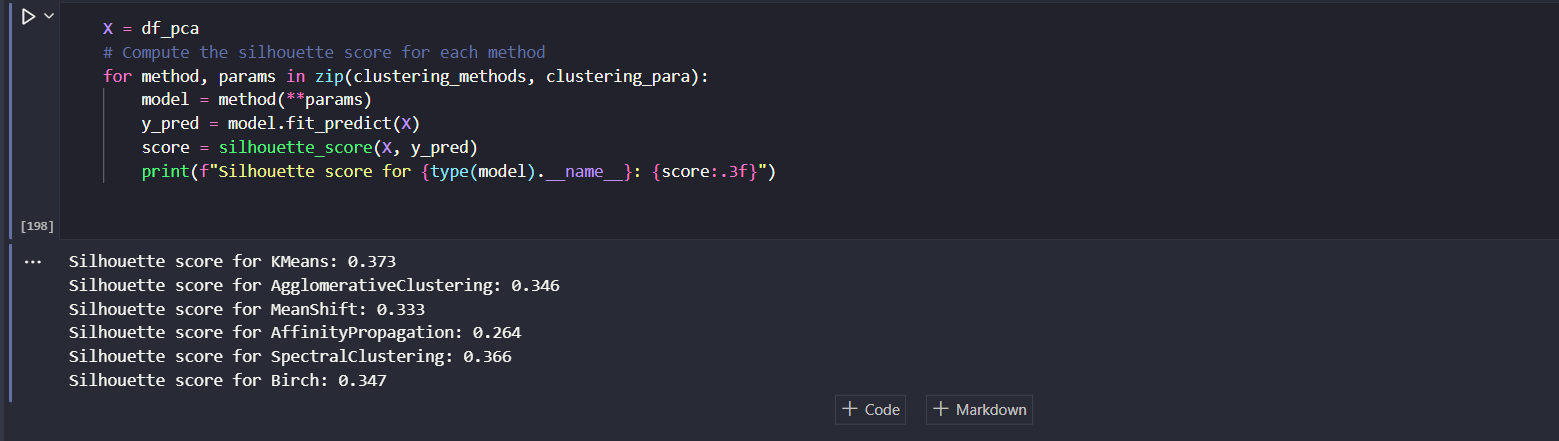




Biểu đồ trực quan các mô hình phân cụm

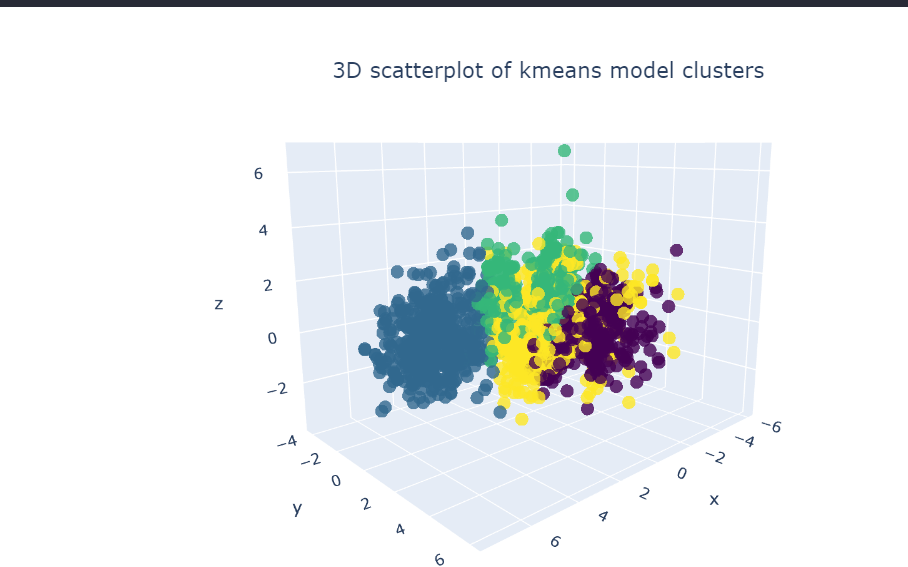


Độ chính xác của từng thuật toán:



1. **Nhận xét và kết luận (Comment, Conclusions):**
   1. **Sử dụng thuật toán Kmeans**

Với kết quả của độ chính xác từng thuật toán thu được ở trên. Kmeans có độ chính xác cao nhất với 0.373. Do đó, nhóm quyết định chọn thuật toán Kmeans để tiến hành phân cụm các nhóm khách hàng.



* 1. **Phân cụm dựa trên thu nhập và chi tiêu:**



Như đã trình bày ở trên, Thu nhập và Chi tiêu có mối quan hệ chặt chẽ. Các cụm hiển thị thông tin chi tiết hơn về các mẫu của chúng:

Nhóm 0: chi tiêu thấp và thu nhập thấp

Nhóm 3: chi tiêu cao và thu nhập trung bình

Nhóm 2: chi tiêu thấp và thu nhập thấp

Nhóm 1: chi tiêu cao và thu nhập cao

Để dễ hiểu hơn, chúng ta sẽ phân loại Nhóm 0 là Khách hàng thông thường, Nhóm 1 là Khách hàng ưu tú, Nhóm 2 là Khách hàng tốt và Nhóm 3 là Khách hàng tiềm năng.

* 1. **Phân cụm dựa trên chi tiêu của sản phẩm:**

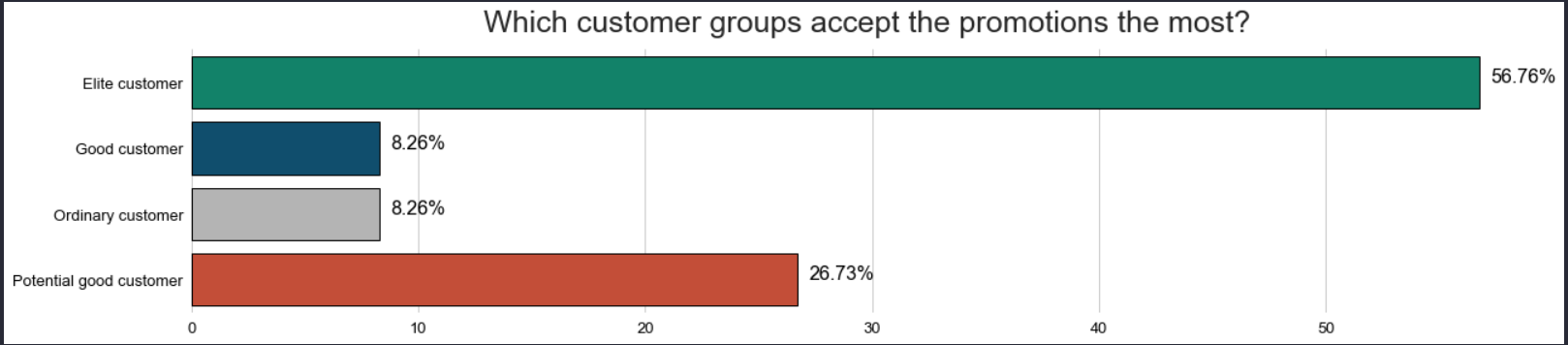


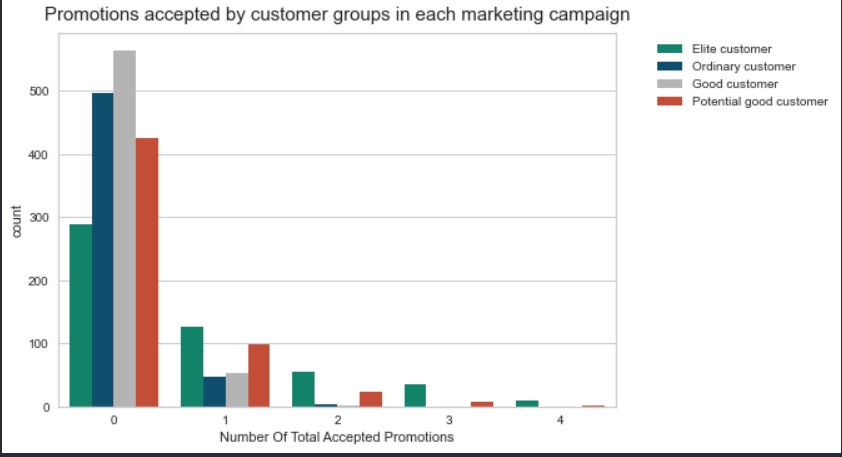
Tất cả bốn nhóm đều chi tiêu nhiều nhất cho Rượu, tiếp theo là Thịt, trong số sáu sản phẩm khác nhau.

Về mua Vàng, nhóm khách hàng phổ thông và khách hàng tiềm năng tốt có xu hướng mua nhiều hơn, với tỷ lệ lần lượt là 17,7% và 10,5%, trong khi hai nhóm còn lại chỉ có tỷ lệ là 8% và 5,4%.

Mặt khác, những khách hàng ưu tú có xu hướng mua Thịt nhiều hơn, chiếm khoảng 240.000 bảng Anh hay 33,9% tổng chi tiêu của họ, cao hơn mức chi tiêu dưới 26% của các nhóm khác.

* 1. **Phân cụm dựa trên tổng quảng cáo**





Đối với chiến dịch tiếp thị ban đầu, mỗi nhóm khách hàng bình thường và tiềm năng tốt có khoảng 500 lượt chấp nhận, trong khi nhóm khách hàng tốt và uu tú có lần lượt là 400 và 300 lượt chấp nhận.

Trong các chiến dịch tiếp theo, số lượng khuyến mãi được chấp nhận giảm đáng kể. Những khách hàng bình thường và có tiềm năng tốt có khuynh hướng không chấp nhận đề nghị, trong khi chỉ một phần nhỏ khách hàng tốt và ưu tú chấp nhận nó.

* 1. **Tổng kết (Summary):**

1. Nhóm ưu tú

Chi tiêu cao và thu nhập cao

Họ chắc chắn không phải là cha mẹ

Tối đa các thành viên trong gia đình là hai người, không có con

Họ là những cặp đôi nhiều hơn một chút so với những người độc thân

2. Nhóm tốt

Chi tiêu cao và thu nhập trung bình

Nhiều khả năng là cha mẹ

Quy mô gia đình của họ có thể từ hai đến bốn thành viên

Chủ yếu có một con (hầu hết là thanh thiếu niên)

3. Nhóm tiềm năng tốt

Chi tiêu thấp và thu nhập trung bình

Tất cả đều là cha mẹ

Quy mô gia đình của họ dao động từ hai đến năm thành viên

Độ tuổi của họ là từ 35 đến khoảng 80 tuổi

4. Nhóm thường

Chi tiêu thấp và thu nhập thấp

Phần lớn trong số họ là cha mẹ

Quy mô gia đình tối đa là ba

Họ thường có một con (chủ yếu là một đứa trẻ)

1. **Đóng góp (Contributions):**

|  |  |  |
| --- | --- | --- |
| **Tên thành viên** | **Nhiệm vụ** | **Mức độ hoàn thành** |
| Nguyễn Thanh Hùng | + Chọn công cụ cho đề tài + Xây dựng thuật toán cho đề tài | 100% |
| Lê Hoàng Khang | + Chọn dữ liệu cho đề tài  + Tiền xử lý dữ liệu | 100% |
| Vũ Trung Kiên | + Sử dụng PowerBI để trực quan và tạo dashboard cho đề tài  + Tổng hợp bài làm, viết báo cáo | 100% |