**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**



**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**BÁO CÁO CUỐI KÌ**

**ĐỀ TÀI: PHÂN TÍCH TẬP DỮ LIỆU VỀ**

**HÀNH VI MUA SẮM CỦA KHÁCH HÀNG**

Danh sách sinh viên thực hiện:

Phan Cao Bằng21133006

Phan Khải Huyền 21133041

Đào Lê Huy Giáp 21133028

Nguyễn Phương Khoa 21133048

**GVHD: Thầy Trần Trọng Bình**

**Thành Phố Hồ Chí Minh, Tháng 05 năm 2024**

|  |  |
| --- | --- |
|  |  |

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**Họ và tên sinh viên thực hiện**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **HỌ VÀ TÊN SINH VIÊN** | **Mã số sinh viên** | **Tỉ lệ tham gia %** | **Kí tên** |
| 1 | Phan Khải Huyền | 21133041 | 100% |  |
| 2 | Phan Cao Bằng | 21133006 | 100% |  |
| 3 | Nguyễn Phương Khoa | 21133048 | 100% |  |
| 4 | Đào Lê Huy Giáp | 21133028 | 100% |  |

**Chuyên ngành:** Kỹ thuật dữ liệu

**Môn học:** Khai phá dữ liệu

**Nhận xét**

Thành phố Hồ Chí Minh, / / 2023

Giảng viên hướng dẫn

(Tên và chữ ký)

**BẢNG PHÂN CÔNG CÔNG VIỆC**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | MSSV | Họ và tên | Phân công nhiệm vụ |
| 1 | 21133041 | Phan Khải Huyền | Làm file word.  Tiền xử lý dữ liệu, trực quan hóa dữ liệu,  tìm kiếm mô hình phù hợp. |
| 2 | 21133006 | Phan Cao Bằng | Tìm tài liệu, chỉnh sửa nội dung  Thực hiện các mô hình phân lớp, so sánh và đánh giá mô hình. |
| 3 | 21133048 | Nguyễn Phương Khoa | Tìm nội dung.  Thực hiện các mô hình luật kết hợp, so sánh và đánh giá mô hình. |
| 4 | 21133028 | Đào Lê Huy Giáp | Tìm nội dung.  Thực hiện các mô hình phân cụm, so sánh và đánh giá mô hình. |

**LỜI CẢM ƠN**

*Kính thưa thầy Trần Trọng Bình, nhóm em xin gửi lời cảm ơn chân thành đến thầy vì đã hướng dẫn và giúp đỡ chúng em trong quá trình hoàn thiện đồ án môn Khai phá dữ liệu. Nhờ vào sự chỉ dẫn và giúp đỡ của thầy, nhóm em đã có thể hoàn thành đồ án một cách hiệu quả và đạt được kết quả như mong đợi. Chúng em rất biết ơn thầy đã dành thời gian và tâm huyết để truyền đạt kiến thức và kinh nghiệm cho nhóm.*

*Thầy đã giúp chúng em có được những kiến thức và kỹ năng cần thiết để làm việc một cách hiệu quả, và nhóm em sẽ sử dụng những kiến thức này để phát triển bản thân và đóng góp vào công việc sau này.*

*Một lần nữa, nhóm em xin chân thành cảm ơn thầy vì sự giúp đỡ và hướng dẫn của thầy trong quá trình học tập của nhóm.*

*Trân trọng,*

**Nhóm 16**

Nguyễn Phương Khoa

Phan Cao Băng

Đào Lê Huy Giáp

Phan Khải Huyền

**Mục Lục**

[CHƯƠNG I: TỔNG QUAN VỀ ĐỀ TÀI 1](#_Toc167974752)

[1.1. Lý do chọn đề tài 1](#_Toc167974753)

[1.2. Tổng quan về tập dữ liệu 1](#_Toc167974754)

[1.2.1. Nguồn dữ liệu 1](#_Toc167974755)

[1.2.2. Mô tả chi tiết tập dữ liệu 1](#_Toc167974756)

[CHƯƠNG II: KIỂM TRA VÀ TIỀN XỬ LÝ DỮ LIỆU 4](#_Toc167974757)

[2.1 Kiểm tra dữ liệu 4](#_Toc167974758)

[2.2. Tiền xử lý dữ liệu 7](#_Toc167974759)

[2.3. Trực quan hóa dữ liệu: 11](#_Toc167974760)

[CHƯƠNG III: MÔ HÌNH PHÂN TÍCH VÀ KẾT QUẢ 17](#_Toc167974761)

[3.1. K-Means Clustering: 17](#_Toc167974762)

[3.1.1. Chuẩn bị dữ liệu 17](#_Toc167974763)

[3.1.2. Xác định số cụm tối ưu bằng phương pháp Elbow 17](#_Toc167974764)

[3.1.3. Xác định số cụm tối ưu bằng phương pháp Silhouette 18](#_Toc167974765)

[3.1.4. Huấn luyện và dự đoán bằng thuật toán K-Means 19](#_Toc167974766)

[3.1.5. Phân cụm 20](#_Toc167974767)

[3.1.6. Vẽ biểu đồ phân phối các cụm 20](#_Toc167974768)

[3.1.7. Đánh giá mô hình 28](#_Toc167974769)

[3.2. ASSOCIATION RULE MINING: APRIORI ALGORITHM: 30](#_Toc167974770)

[3.2.1. Chuẩn bị dữ liệu cho thuật toán Apriori: 30](#_Toc167974771)

[Thực hiện phân đoạn (segmentation) các dữ liệu dựa trên độ tuổi, thu nhập và số ngày tương tác. 30](#_Toc167974772)

[3.2.2. Applying Apriori Algorithm 33](#_Toc167974773)

[3.2.3. Setting the Association Rules 34](#_Toc167974774)

[3.2.4. Tìm kiếm "Highly Active Customers" theo các sản phẩm: 35](#_Toc167974775)

[3.2.5. Đánh giá mô hình: 44](#_Toc167974776)

[3.3. CLASSIFICATION MODEL 44](#_Toc167974777)

[3.3.1. SVC 45](#_Toc167974778)

[3.3.2. KNeighborsClassifier 46](#_Toc167974779)

[3.3.3. LogisticRegression 48](#_Toc167974780)

[3.3.4. So sánh các mô hình 50](#_Toc167974781)

[CHƯƠNG IV: KẾT LUẬN 52](#_Toc167974782)

[CHƯƠNG VI: TÀI LIỆU THAM KHẢO 53](#_Toc167974783)

# CHƯƠNG I: TỔNG QUAN VỀ ĐỀ TÀI

## 1.1. Lý do chọn đề tài

Ngày nay, việc mua sắm không chỉ đơn thuần là một hoạt động cần thiết trong cuộc sống hàng ngày, mà còn trở thành một nguồn dữ liệu quan trọng để khai thác. phân tích hành vi mua sắm của khách hàng là một công cụ quan trọng để tăng cường hiểu biết về khách hàng và tương tác với họ một cách hiệu quả. Qua việc tối ưu hóa chiến dịch tiếp thị, tạo sản phẩm tùy chỉnh và nắm bắt xu hướng thị trường, doanh nghiệp có thể đạt được sự thành công và phát triển bền vững trong môi trường kinh doanh ngày càng cạnh tranh.

Việc chọn đề tài phân tích tập dữ liệu hành vi mua sắm của khách hàng giúp chúng em liệu nắm vững kiến thức, nâng cao kỹ năng phân tích dữ liệu và giải quyết các vấn đề thực tế trong cuộc sống.

## 1.2. Tổng quan về tập dữ liệu

### 1.2.1. Nguồn dữ liệu

* Tập dữ liệu “Customer Personality Analysis” được lấy từ trang web Kaggle, một nền tảng cung cấp các tập dữ liệu tin cậy và miễn phí cho mục đích học tập và nghiên cứu.
* Đường dẫn tập dữ liệu: <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

### 1.2.2. Mô tả chi tiết tập dữ liệu

Tập dữ liệu bao gồm:

- 29 cột.

- 10240 dòng.

Các thuộc tính của tập dữ liệu:

Thông tin về người:

|  |  |
| --- | --- |
| ID | Mã khách hàng |
| Year\_Birth | Năm sinh của 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 hộ gia đình khách hàng |
| Kidhome | Số lượng trẻ em trong hộ gia đình khách hàng |
| Teenhome | Số lượng thanh thiếu niên trong hộ gia đình của khách hàng |
| Dt\_Customer | Ngày khách hàng bắt đầu giao dịch với công ty |
| Recency | Số ngày kể từ lần mua hàng cuối cùng của khách hàng |
| Complain | 1 nếu khách hàng đã khiếu nại trong 2 năm qua, 0 nếu không |

Thông tin về sản phẩm:

|  |  |
| --- | --- |
| MntWines | Số tiền chi cho rượu vang trong 2 năm qua |
| MntFruits | Số tiền chi cho trái cây trong 2 năm qua |
| MntMeatProducts | Số tiền chi cho thịt trong 2 năm qua |
| MntFishProducts | Số tiền chi cho cá trong 2 năm qua |
| MntSweetProducts | Số tiền chi cho đồ ngọt trong 2 năm qua |
| MntGoldProds | Số tiền chi cho vàng trong 2 năm qua |

Thông tin về khuyến mãi:

|  |  |
| --- | --- |
| NumDealsPurchases | Số lần mua hàng được giảm giá |
| AcceptedCmp1 | 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch đầu tiên, 0 nếu không |
| AcceptedCmp2 | 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ hai, 0 nếu không |
| AcceptedCmp3 | 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ ba, 0 nếu không |
| AcceptedCmp4 | 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ tư, 0 nếu không |
| AcceptedCmp5 | 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ năm, 0 nếu không |
| Response | 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch cuối cùng, 0 nếu không |

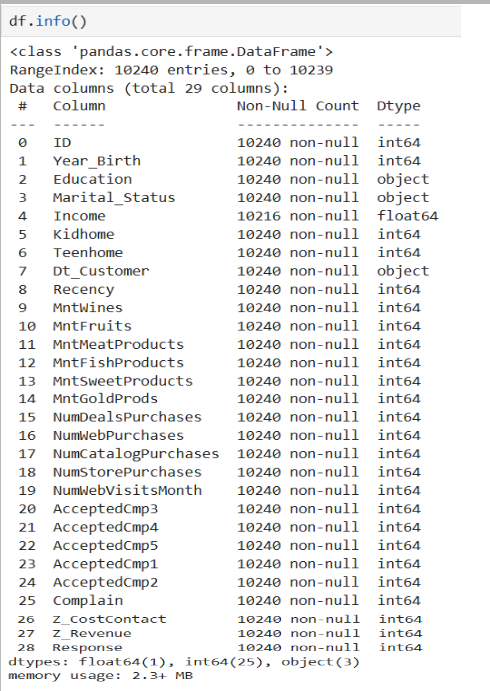
Thông tin về cách thức mua hàng:

|  |  |
| --- | --- |
| NumWebPurchases | Số lần mua hàng thông qua website của công ty |
| NumCatalogPurchases | Số lần mua hàng bằng cách sử dụng danh mục |
| NumStorePurchases | Số lần mua hàng trực tiếp tại cửa hàng |
| NumWebVisitsMonth | Số lần ghé thăm website của công ty trong tháng qua |

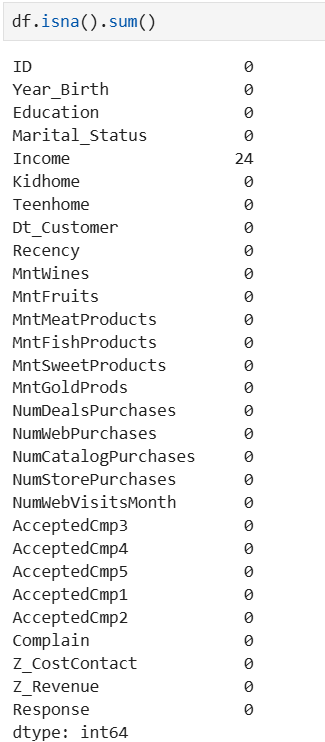
# CHƯƠNG II: KIỂM TRA VÀ TIỀN XỬ LÝ DỮ LIỆU

## 2.1 Kiểm tra dữ liệu

- Kiểm tra kiểu giá trị của dữ liệu:



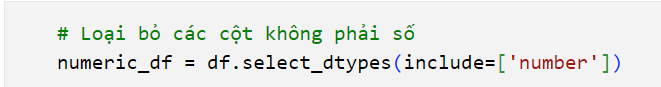
- Kiểm tra các giá trị null:



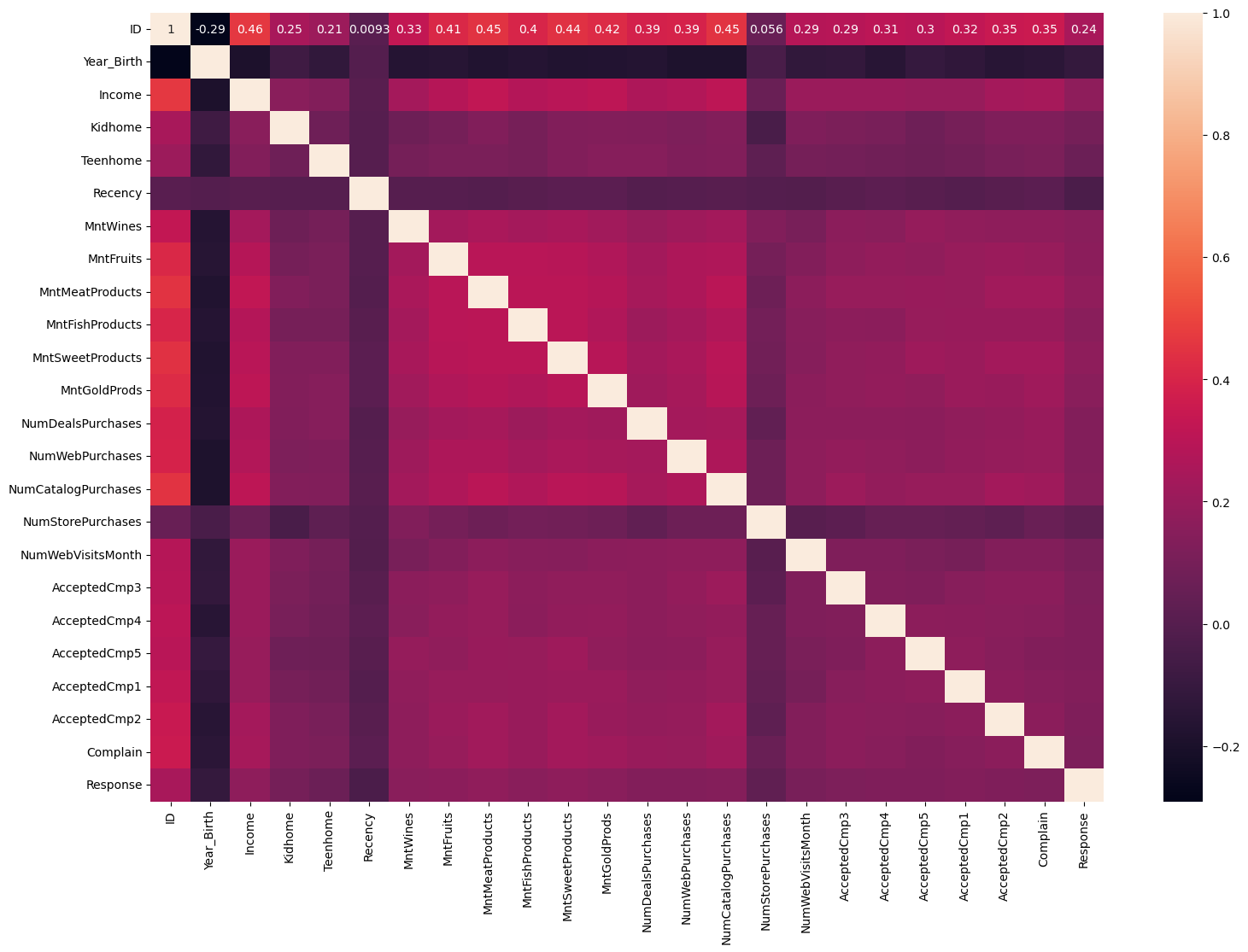
* Thuộc tính ['Income'] có 24 giá trị null.

\* Kiểm tra mức độ tương quan giữa các thuộc tính

* Loại bỏ các cột có kiểu dữ liệu không phải số

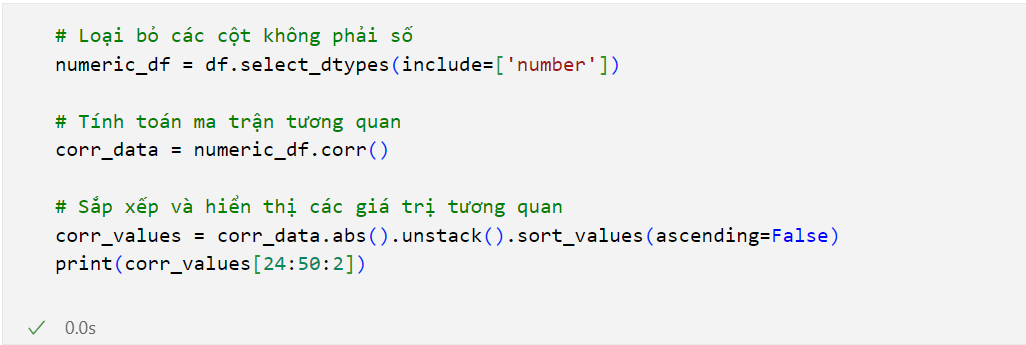


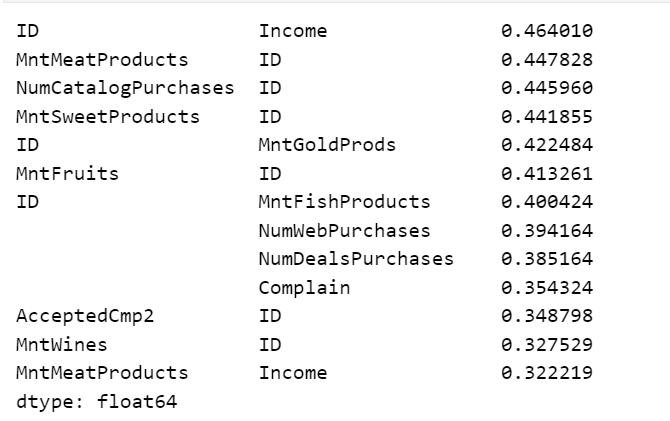
* Biểu đồ heatmap để trực quan hóa mức độ tương quan giữa các cột số



🡪 Mỗi cột mang thông tin riêng và không quá trùng lặp với các cột khác. Do đó, không cần loại bỏ bất kỳ cột nào dựa trên tương quan giữa chúng.

\* Kiểm tra tương quan bằng cách unstack dữ liệu

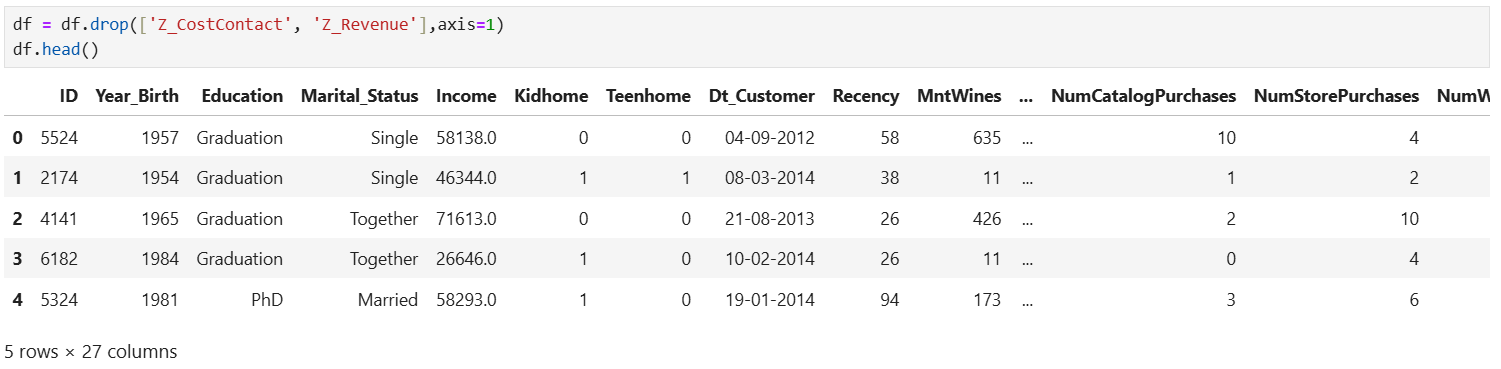




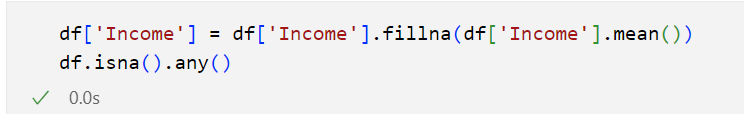
Khi giá trị tương quan gần với 1 hoặc -1, điều đó chỉ ra mức độ tương quan cao. Giá trị gần 1 cho thấy mối tương quan dương mạnh, trong khi giá trị gần -1 cho thấy mối tương quan âm mạnh. Nếu giá trị tương quan gần 0, tức là không có mối tương quan hoặc mối tương quan yếu giữa hai biến.

## 2.2. Tiền xử lý dữ liệu

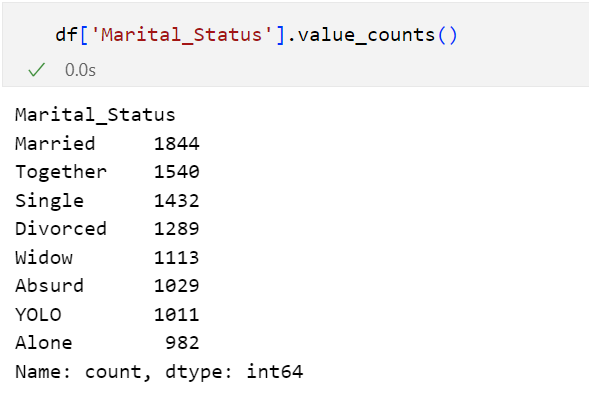
- Xóa 2 cột 'Z\_CostContact' và 'Z\_Revenue' vì không sử dụng cho mô hình dự đoán:



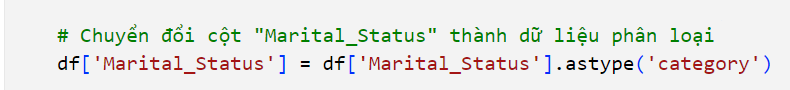
- Điền giá trị thiếu vào cột “Income” bằng giá trị trung bình



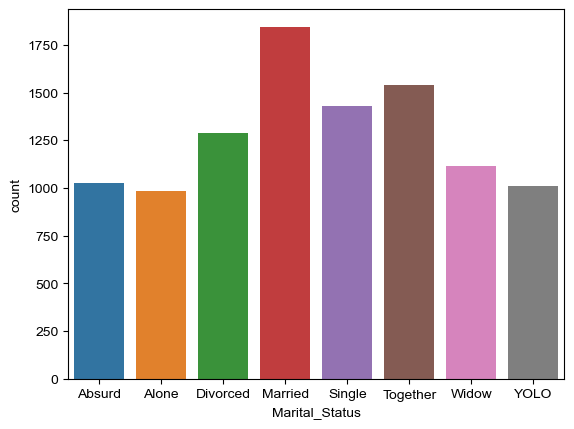
- Kiểm tra số lượng danh mục duy nhất có trong cột "Marital\_Status"



Chuyển đổi cột "Marital\_Status" thành dữ liệu phân loại và hiển thị số lượng mẫu trong mỗi danh mục

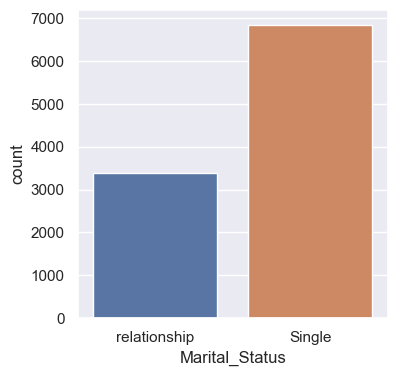


Trước

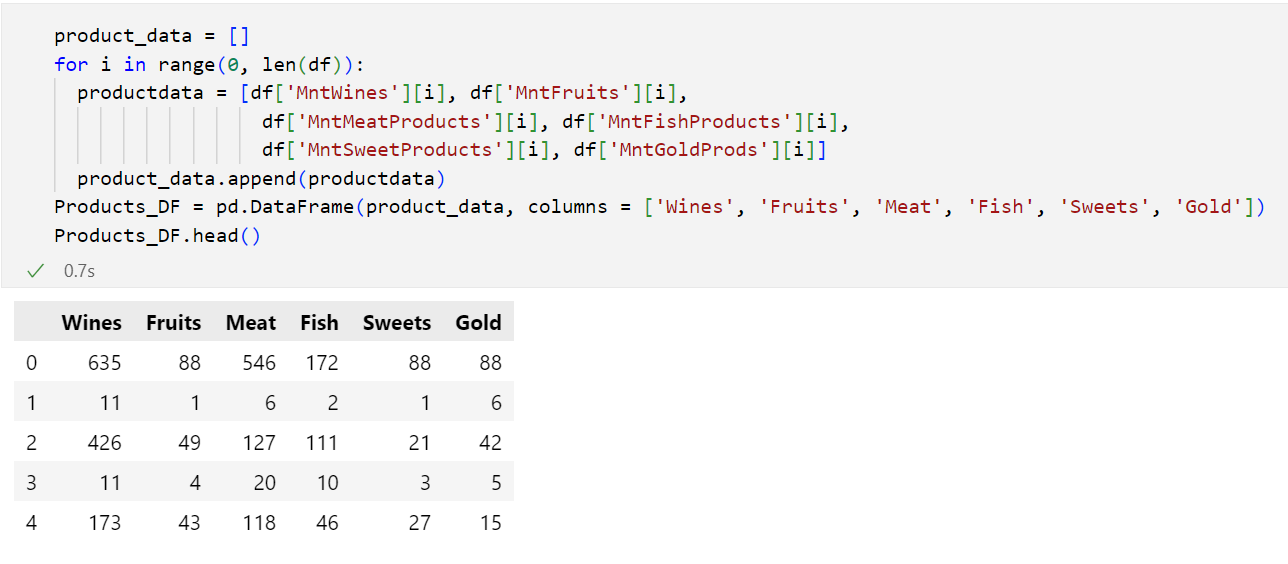


Sau khi nhóm 'Married', 'Together' lại thành "relationship"

'Divorced', 'Widow', 'Alone', 'YOLO', 'Absurd' được nhóm lại thành "Single"

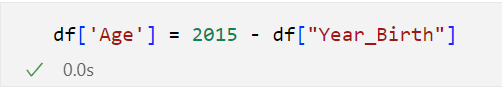


- Tách các sản phẩm thành các DataFrame riêng biệt để sử dụng cho việc khai thác luật kết hợp

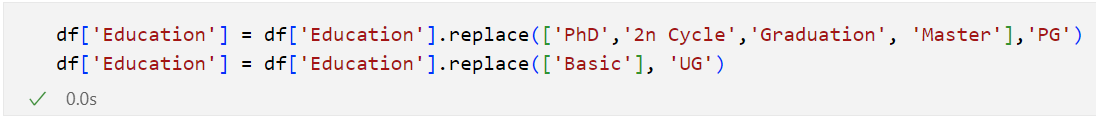


- Kết hợp các DataFrame khác nhau thành một cột duy nhất để giảm số chiều.

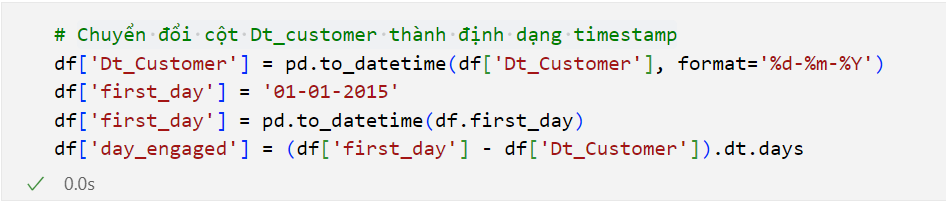
- Thêm cột 'Age'



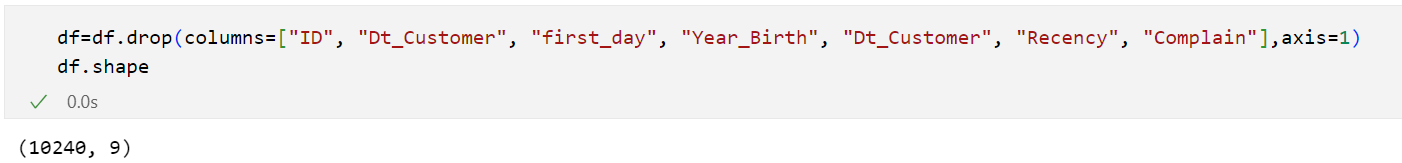
- Chuyển đổi các giá trị trong cột Education thành UG và PG



- Chuyển đổi cột Dt\_customer thành định dạng timestamp

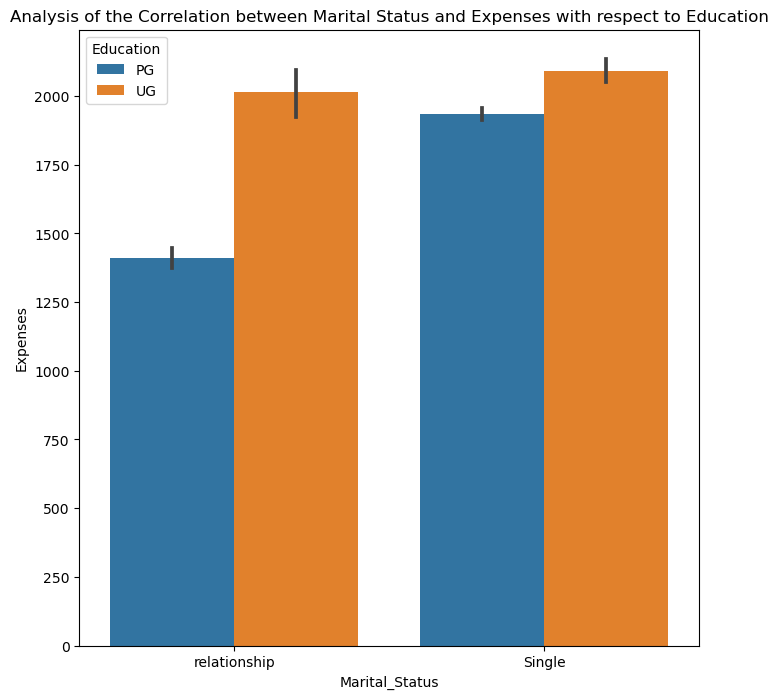


- Loại bỏ các cột không cần thiết



## 2.3. Trực quan hóa dữ liệu:

- Mối tương quan giữa tình trạng hôn nhân và các khoản chi tiêu dựa trên trình độ học vấn



🡪Biểu đồ được chia thành hai phần:

Phần bên trái hiển thị chi tiêu của các nhóm đã có gia đình theo mức độ học vấn.

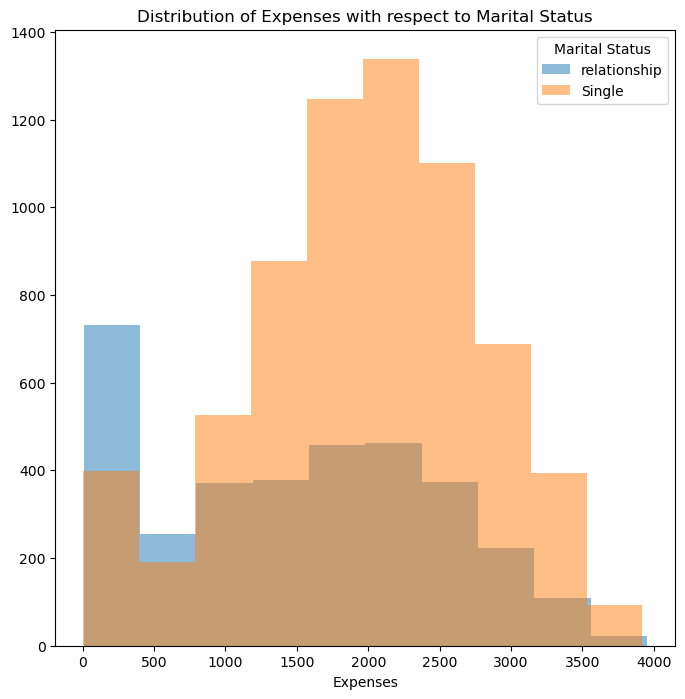
Phần bên phải hiển thị chi tiêu của các nhóm độc thân theo mức độ học vấn.

Nhìn chung, biểu đồ cho thấy:

Chi tiêu tăng theo mức độ học vấn đối với cả hai nhóm độc thân và đã có gia đình.

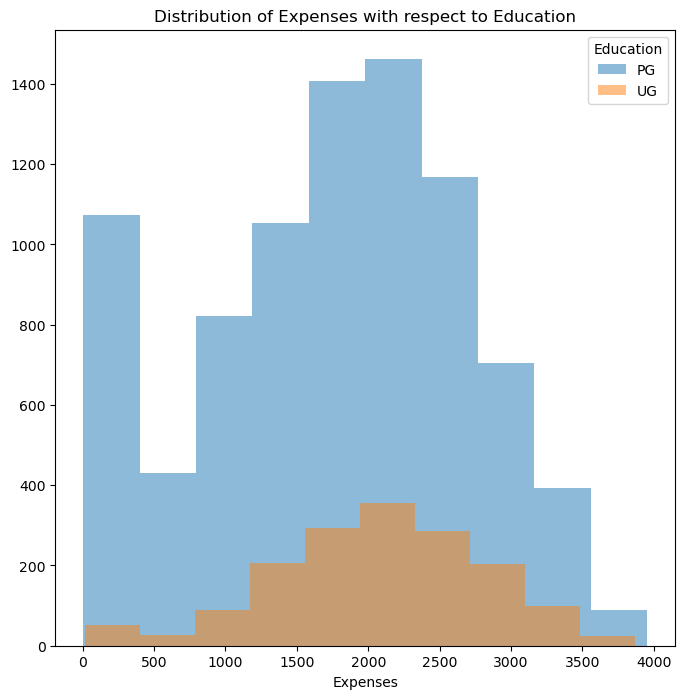
Nhóm đã có gia đình có mức chi tiêu thấp hơn nhóm độc thân ở tất cả các mức độ học vấn.

- Phân phối của các khoản chi tiêu theo tình trạng hôn nhân



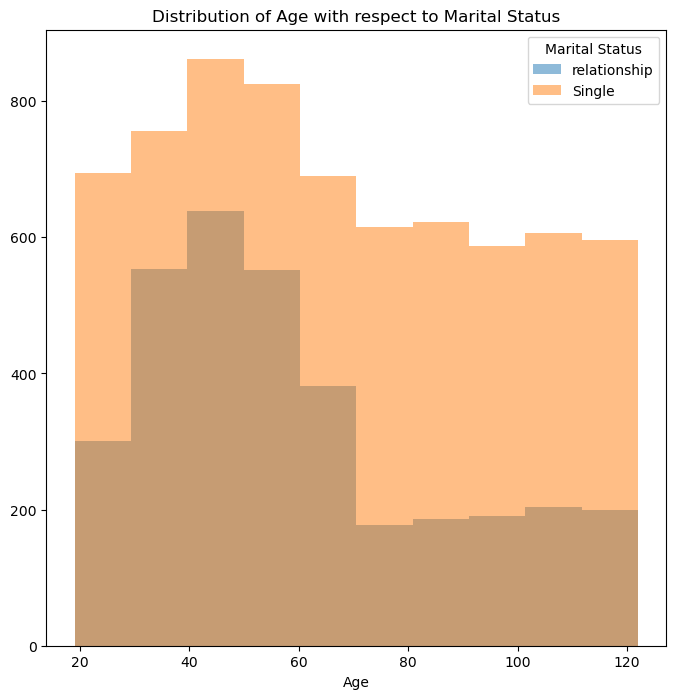
Nhóm đã có gia đình có xu hướng chi tiêu ít hơn nhóm độc thân

- Phân phối của các khoản chi tiêu theo trình độ học vấn



Nhóm có trình độ học vấn cao hơn có xu hướng chi tiêu nhiều hơn nhóm có trình độ học vấn thấp hơn

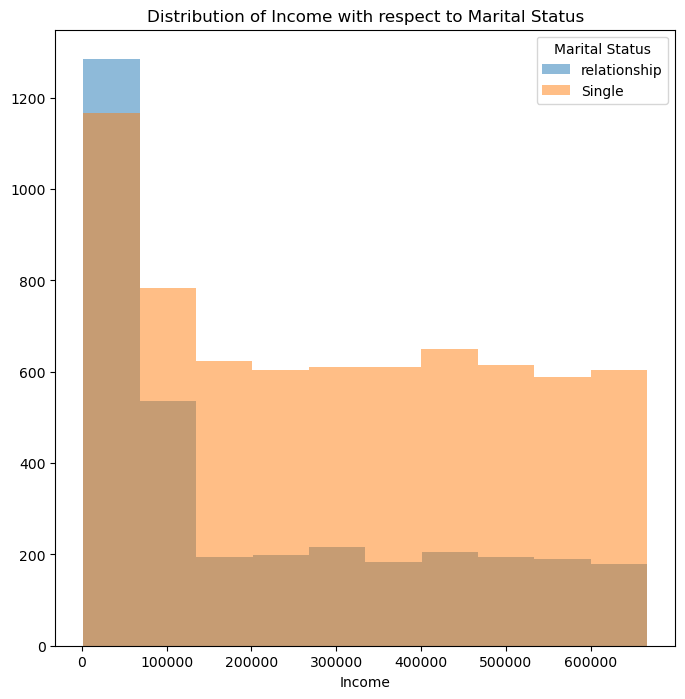
- Phân phối tuổi theo tình trạng hôn nhân



Tỷ lệ khách hàng độc thân cao nhất ở độ tuổi 40-50

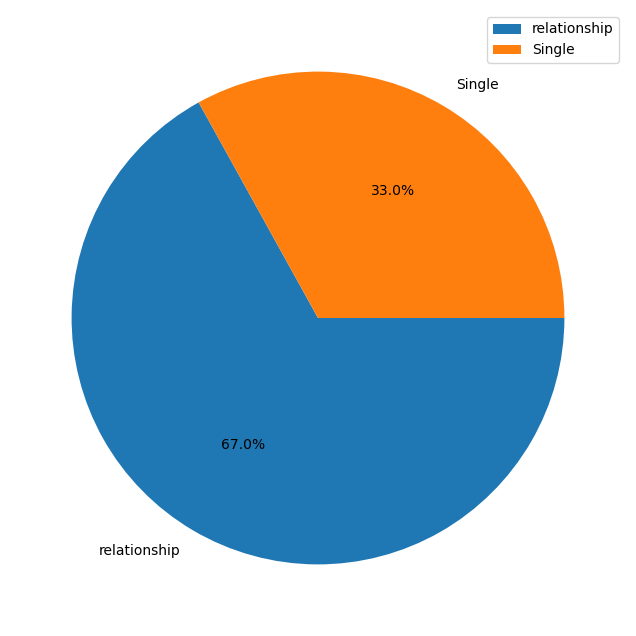
Tỷ lệ khách hàng đã có gia đình cao nhất ở độ tuổi 40-50

- Phân phối thu nhập theo tình trạng hôn nhân



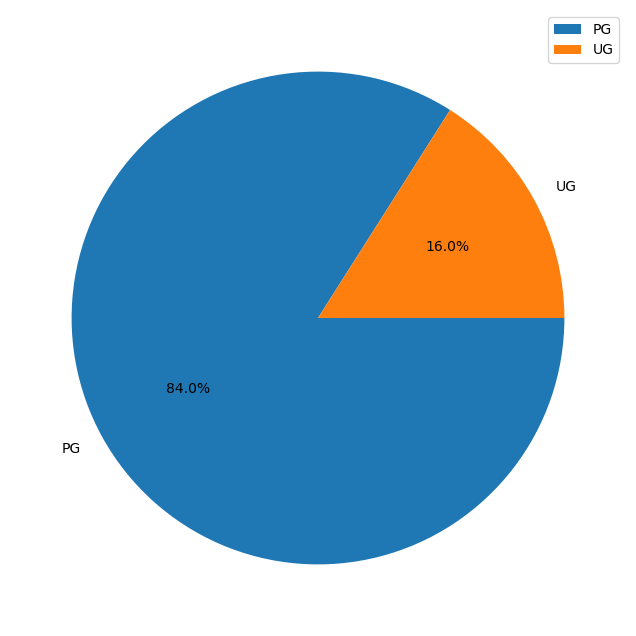
Thu nhập của nhóm khách hàng đã có gia đình cao hơn so với nhóm độc thân

- Tỉ lệ khách hàng theo tình trạng hôn nhân



Tỷ lệ nhóm đã có gia đình (67%) cao hơn nhiều so với tỷ lệ nhóm độc thân (33%).

- Tỉ lệ khách hàng theo trình độ học vấn

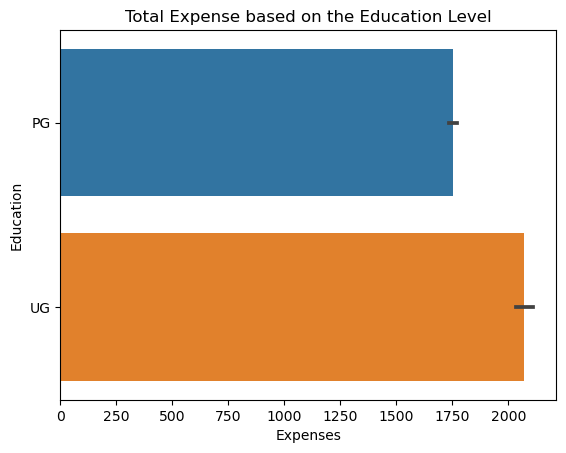


Tỷ lệ phần trăm của PG cao hơn so với UG.

Tỷ lệ phần trăm của PG là 84%.

Tỷ lệ phần trăm của UG là 16%.

- Tổng chi tiêu dựa trên trình độ học vấn



- Tổng thu nhập dựa trên trình độ học vấn

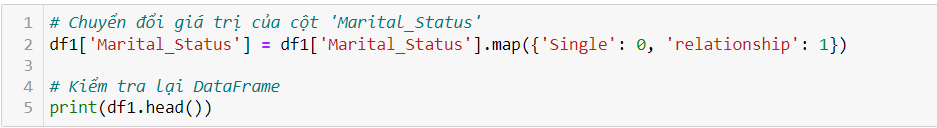


# CHƯƠNG III: MÔ HÌNH PHÂN TÍCH VÀ KẾT QUẢ

## 3.1. K-Means ****Clustering****:

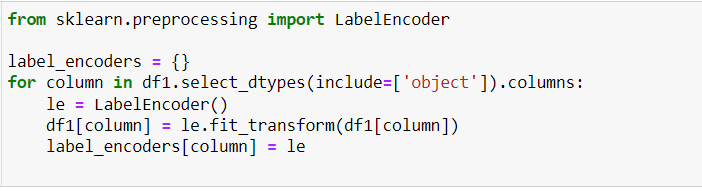
### 3.1.1. Chuẩn bị dữ liệu

**Chuyển đổi giá trị của cột 'Marital\_Status'**



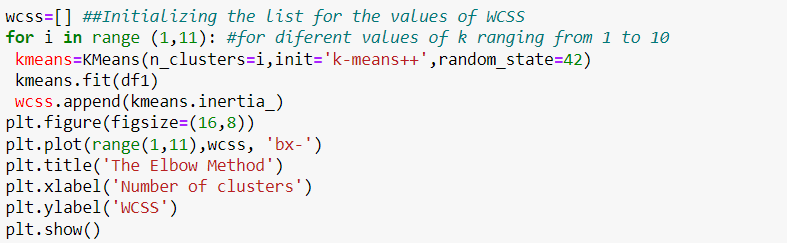
Chuyển đổi các giá trị của cột 'Marital\_Status' thành các giá trị số để dễ dàng sử dụng trong các thuật toán học máy. Trong trường hợp này, 'Single' được chuyển thành 0 và 'relationship' được chuyển thành 1.

**Mã hóa các cột dạng chuỗi**



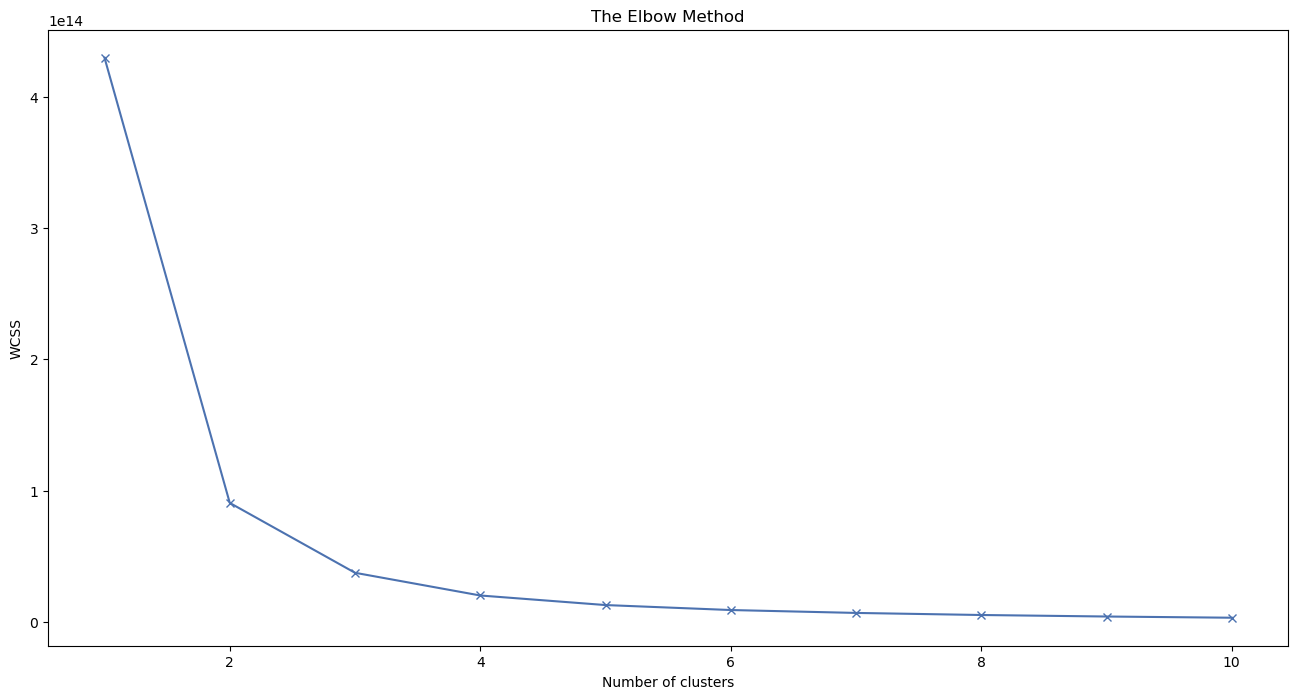
Mã hóa tất cả các cột dạng chuỗi (object) trong DataFrame thành các giá trị số bằng cách sử dụng 'LabelEncoder' từ thư viện 'sklearn'.

### 3.1.2. Xác định số cụm tối ưu bằng phương pháp Elbow



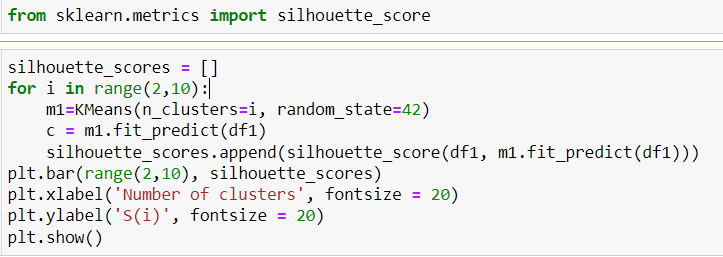
* Sử dụng phương pháp Elbow để xác định số cụm (k) tối ưu cho thuật toán K-Means.
* Tính toán Within-Cluster Sum of Squares (WCSS) cho các giá trị k từ 1 đến 10.
* Vẽ đồ thị WCSS để tìm điểm khủy tay (Elbow), nơi mà tăng thêm số cụm không làm giảm nhiều WCSS, từ đó xác định số cụm tối ưu.

**Đồ thị WCSS:**

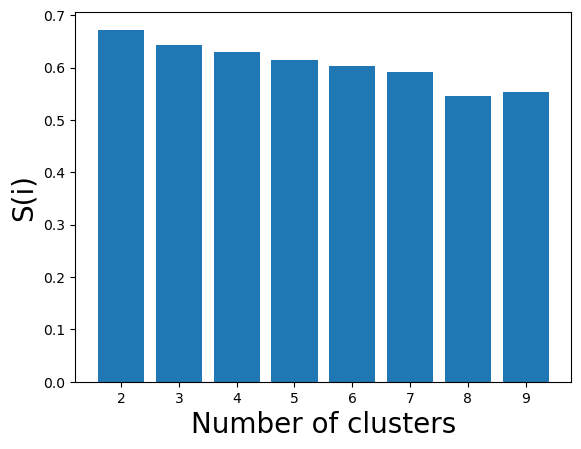


Trong hình trên thì ta thấy vị trí của điểm khuỷ tay chính là k=2 vì khi số lượng cụm lớn hơn 2 thì tốc độ suy giảm của hàm biến dạng dường như không đáng kể so với trước đó.

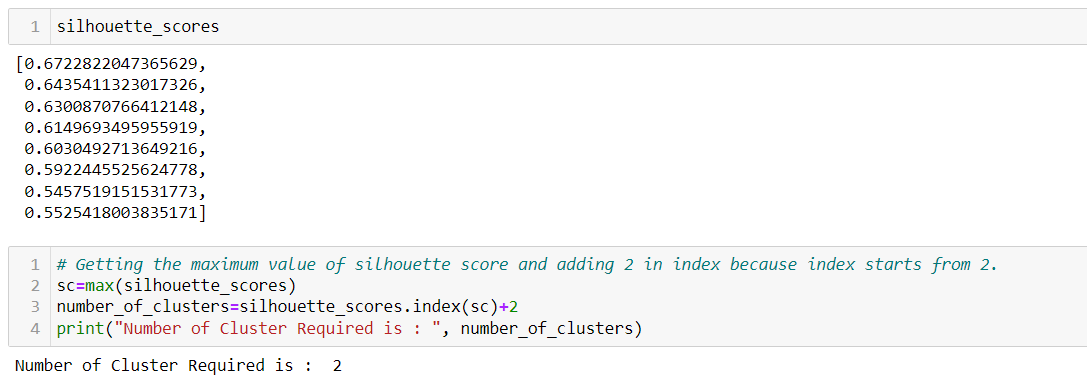
### 3.1.3. Xác định số cụm tối ưu bằng phương pháp Silhouette



- Biểu đồ cột chỉ số Silhouette theo số cụm:

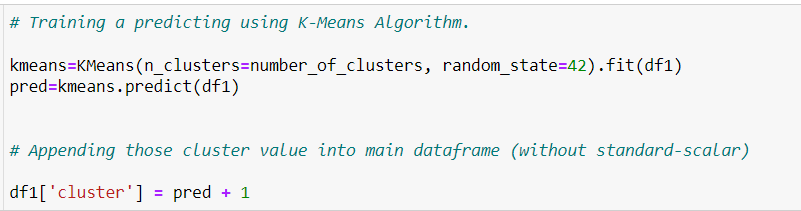


- Các chỉ số Silhouette và số cụm tối ưu cho thuật toán K-Means:



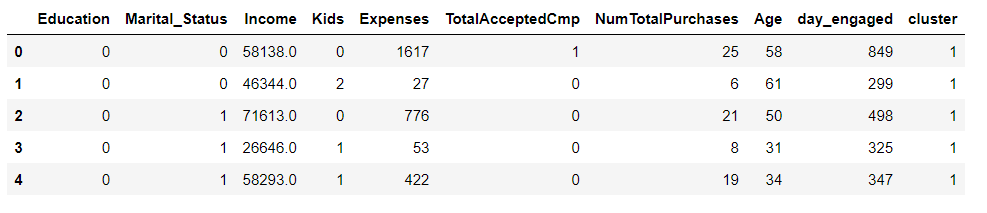
‘+2’: Vì vòng lặp bắt đầu từ 2 (range(2,10)), chỉ số của giá trị Silhouette trong danh sách silhouette\_scores tương ứng với số lượng cụm trừ đi 2. Do đó, cần cộng thêm 2 vào chỉ số để tìm đúng số lượng cụm tối ưu.

### 3.1.4. Huấn luyện và dự đoán bằng thuật toán K-Means

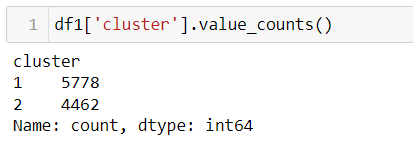


* Huấn luyện mô hình K-Means với số cụm tối ưu là 2.

**Dữ liệu trong df1:**



### 3.1.5. Phân cụm



- Số điểm dữ liệu trong cluster 1: 5778 , cluster 2: 4462

- Biểu đồ hiển thị sự phân bố các điểm dữ liệu trong 2 cụm

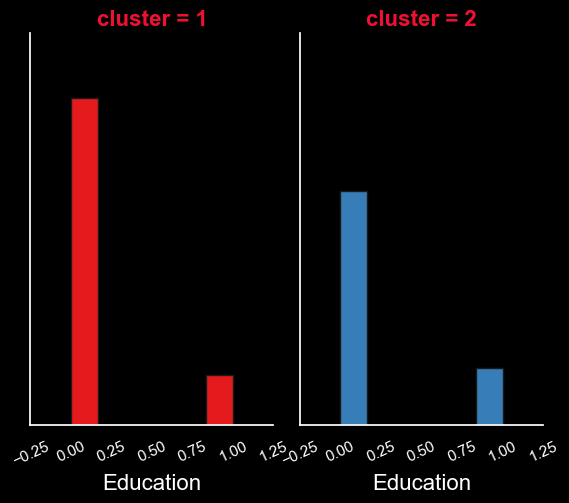


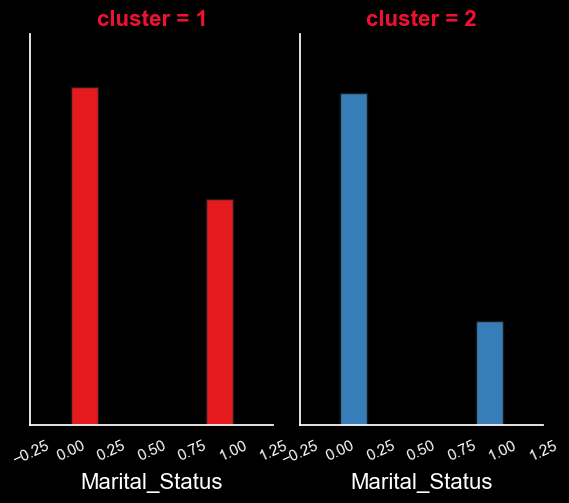
### 3.1.6. Vẽ biểu đồ phân phối các cụm



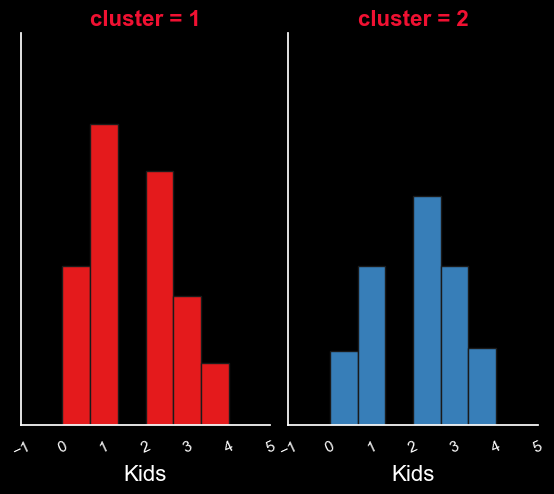
* Vẽ biểu đồ phân phối (histogram) cho từng cột trong DataFrame df1, phân theo

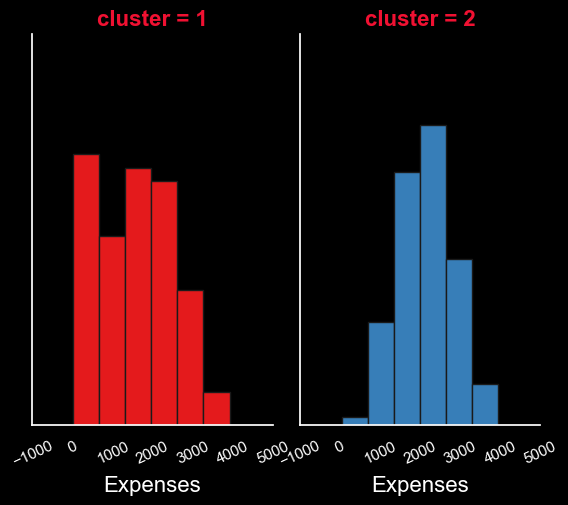
**Kết quả:**

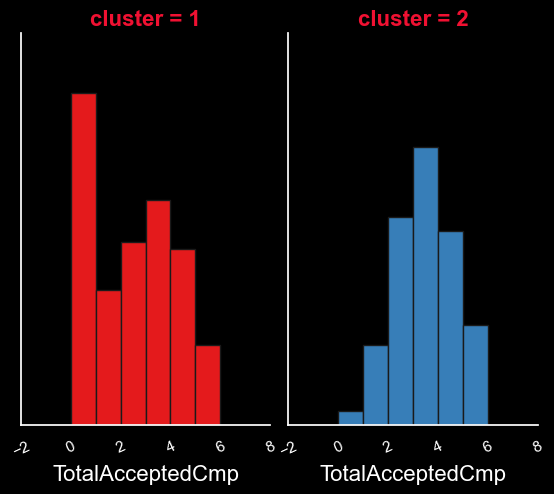




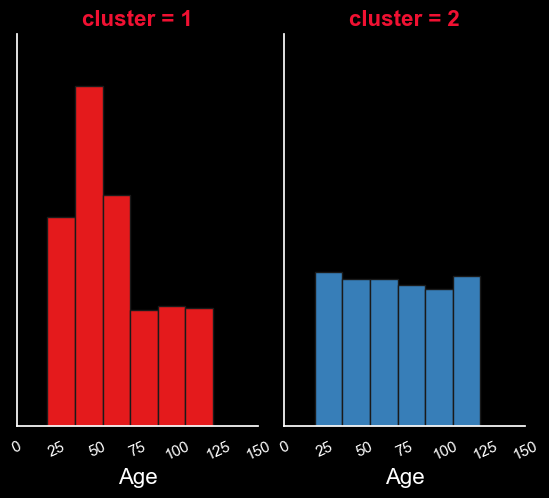


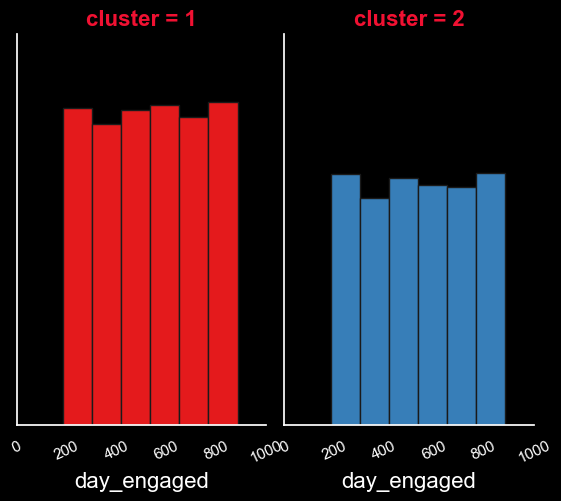


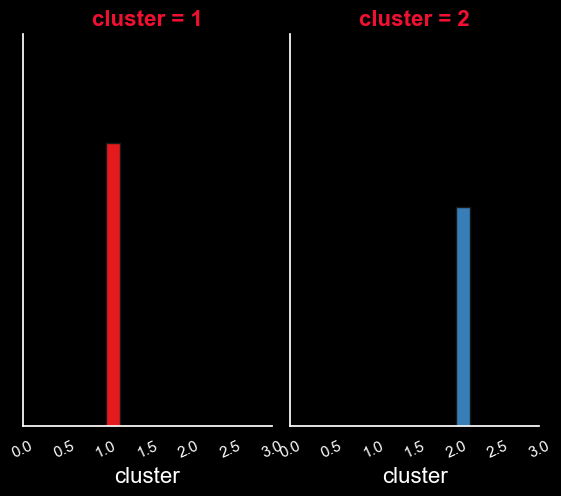












**Dựa trên thông tin trên, chúng ta có thể chia khách hàng thành hai nhóm chính:**

- Khách hàng hoạt động cao thuộc cụm 1.

- Khách hàng ít hoạt động thuộc cụm 2.

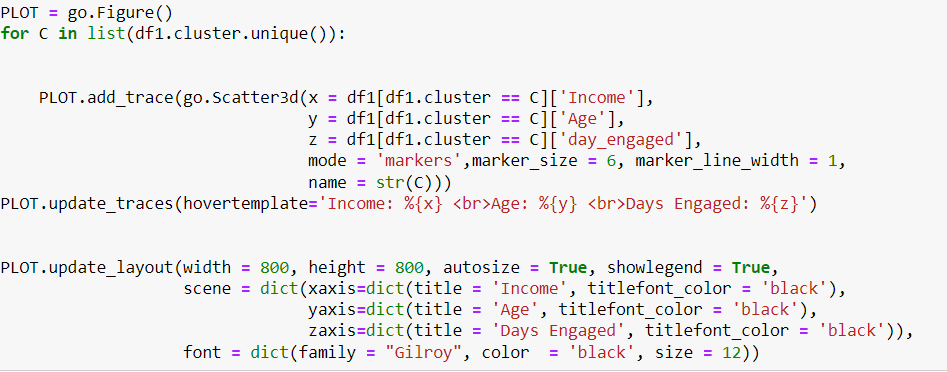
**1. Đặc điểm của khách hàng hoạt động cao**

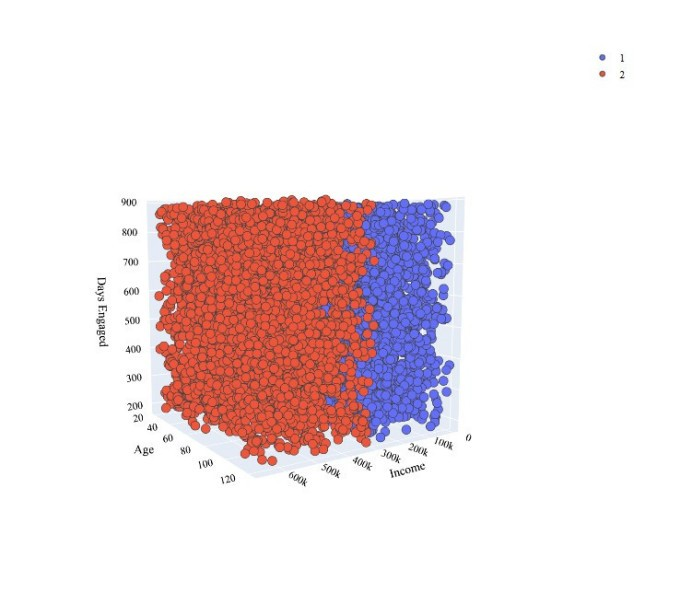
* Về giáo dục: Khách hàng hoạt động cao thường có trình độ sau đại học.
* Về tình trạng hôn nhân: Số người trong mối quan hệ gần gấp đôi số người độc thân.
* Về thu nhập: Thu nhập của khách hàng hoạt động cao thấp hơn một chút so với khách hàng ít hoạt động.
* Về con cái: Khách hàng hoạt động cao có nhiều con hơn so với các khách hàng khác (trung bình khoảng 1 con).
* Về chi tiêu: Chi tiêu của khách hàng hoạt động cao ít hơn so với khách hàng ít hoạt động. Những khách hàng này chi tiêu trung bình khoảng 100-200 đơn vị tiền.
* Về tuổi: Tuổi của những khách hàng này dao động từ 25 đến 125. Độ tuổi phổ biến nhất là từ 40 đến 50.
* Về mức độ gắn bó: Khách hàng hoạt động cao trung thành hơn và gắn bó với công ty trong thời gian dài hơn.

**2. Đặc điểm của khách hàng ít hoạt động**

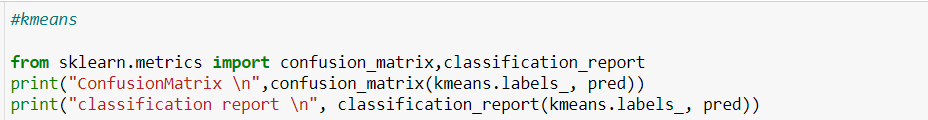
* Về giáo dục: Khách hàng ít hoạt động thường có trình độ đại học.
* Về tình trạng hôn nhân: Số người trong mối quan hệ và số người độc thân xấp xỉ nhau.
* Về thu nhập: Thu nhập của khách hàng ít hoạt động rất thấp hoặc gần như không có.
* Về con cái: Chỉ có một số ít khách hàng này có con.
* Về chi tiêu: Chi tiêu của khách hàng ít hoạt động rất thấp hoặc gần như không có.
* Về tuổi: Tuổi của những khách hàng này dao động từ 25 đến 125.
* Về mức độ gắn bó: Khách hàng ít hoạt động không tham gia nhiều và không gắn bó với công ty trong thời gian dài.

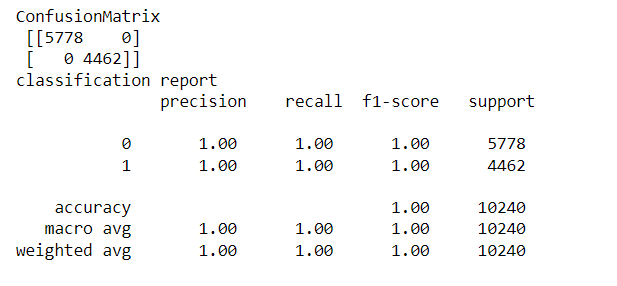
- Biểu Đồ 3D Phân Tán hai cụm theo các cột ‘Income’, ‘Age’, ‘day\_engaged’:





### 3.1.7. Đánh giá mô hình





- Precision (Độ chính xác): Tỷ lệ của các dự đoán chính xác trên tổng số các dự đoán cho mỗi lớp. Công thức tính: Precision = TP / (TP + FP)

Precision cho cụm 0 và 1 đều là 1.00, tức là tất cả các mẫu dự đoán cho cả hai cụm đều chính xác.

- Recall (Độ nhạy): Tỷ lệ của các dự đoán chính xác trên tổng số các mẫu thực tế cho mỗi lớp. Công thức tính: Recall = TP / (TP + FN)

Recall cho cụm 0 và 1 đều là 1.00, tức là mô hình đã nhận diện chính xác tất cả các mẫu thực tế cho cả hai cụm.

- F1-Score: Trung bình điều hòa của Precision và Recall. Công thức tính: F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)

F1-Score cho cụm 0 và 1 đều là 1.00, cho thấy mô hình cân bằng tốt giữa Precision và Recall.

- Support: Số lượng mẫu thực tế trong mỗi lớp.

Support cho cụm 0 là 5778.

Support cho cụm 1 là 4462.

- Accuracy (Độ chính xác chung): Tỷ lệ của các dự đoán chính xác trên tổng số mẫu. Công thức tính: Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy tổng thể là 1.00, cho thấy tất cả các mẫu đã được phân loại chính xác.

- Macro Avg: Trung bình số học của Precision, Recall và F1-Score trên các lớp.

Tất cả các giá trị đều là 1.00, cho thấy mô hình hoạt động tốt đồng đều trên cả hai lớp.

- Weighted Avg: Trung bình có trọng số của Precision, Recall và F1-Score trên các lớp, có tính đến support.

Tất cả các giá trị đều là 1.00, cho thấy mô hình hoạt động tốt trên toàn bộ dữ liệu.

## 3.2. ASSOCIATION RULE MINING: APRIORI ALGORITHM:

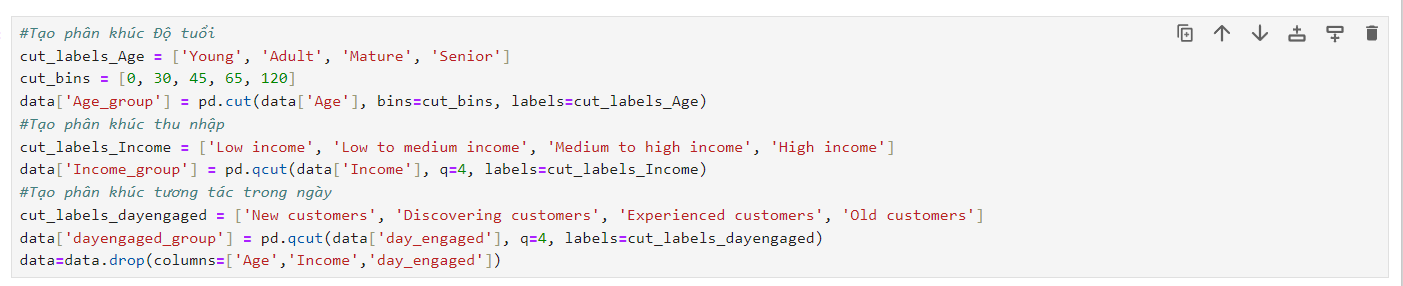
Khai phá quy tắc kết hợp được sử dụng khi chúng ta muốn tìm mối liên kết giữa các đối tượng khác nhau trong một tập hợp, tìm các mẫu thường xuyên trong cơ sở dữ liệu giao dịch.

Thuật toán Apriori là kỹ thuật đơn giản nhất để xác định mối quan hệ cơ bản giữa các loại phần tử khác nhau.

Ở đây chúng ta sử dụng thuật toán này để tìm ra khách hàng nào phù hợp nhất cho một mặt hàng nhất định. Qua đó giúp doanh nghiệp quảng bá đúng đối tượng khách hàng mục tiêu để tăng hiệu quả và tiết kiệm chi phí.

### 3.2.1. Chuẩn bị dữ liệu cho thuật toán Apriori:

Thực hiện phân đoạn (segmentation) các dữ liệu dựa trên độ tuổi, thu nhập và số ngày tương tác.



**Phân đoạn độ tuổi**:

 cut\_labels\_Age: danh sách các nhãn cho các nhóm tuổi (Young, Adult, Mature, Senior).

 cut\_bins: danh sách các điểm cắt cho các nhóm tuổi, tức là 0-30, 31-45, 46-65, 66-120.

 pd.cut(): hàm này phân đoạn dữ liệu data['Age'] dựa trên các điểm cắt cut\_bins và gán nhãn từ cut\_labels\_Age cho các nhóm tương ứng.

 Kết quả được lưu vào cột Age\_group.

**Phân đoạn thu nhập**:

 cut\_labels\_Income: danh sách các nhãn cho các nhóm thu nhập (Low income, Low to medium income, Medium to high income, High income).

 pd.qcut(): hàm này phân đoạn dữ liệu data['Income'] thành 4 phần bằng nhau (quartiles) và gán nhãn từ cut\_labels\_Income.

 Kết quả được lưu vào cột Income\_group.

**Phân đoạn số ngày tương tác**:

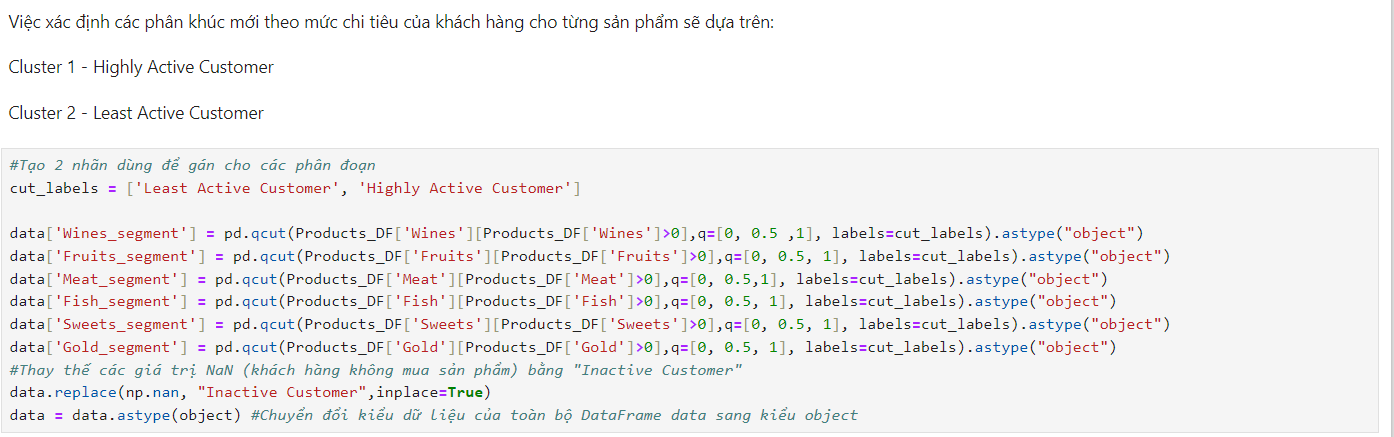
 cut\_labels\_dayengaged: danh sách các nhãn cho các nhóm số ngày tương tác (New customers, Discovering customers, Experienced customers, Old customers).

 pd.qcut(): hàm này phân đoạn dữ liệu data['day\_engaged'] thành 4 phần bằng nhau và gán nhãn từ cut\_labels\_dayengaged.

 Kết quả được lưu vào cột dayengaged\_group.  
Việc xác định các phân khúc mới theo mức chi tiêu của khách hàng cho từng sản phẩm sẽ dựa trên:

Cluster 1 - Highly Active Customer

Cluster 2 - Least Active Customer



**Tạo các phân đoạn**:

cut\_labels: danh sách các nhãn cho các nhóm hoạt động (Least Active Customer, Highly Active Customer). Đây là các nhãn dùng để gán cho các phân đoạn.

**Phân đoạn mức độ hoạt động của từng danh mục sản phẩm**:

pd.qcut(): hàm này phân đoạn dữ liệu thành 2 phần bằng nhau (0-50% và 50-100%) và gán nhãn từ cut\_labels.

Products\_DF['<Product>'][Products\_DF['<Product>']>0]: lọc ra những giá trị lớn hơn 0 cho từng sản phẩm để chỉ phân đoạn những khách hàng đã mua sản phẩm đó.

.astype("object"): chuyển đổi kiểu dữ liệu của cột mới sang kiểu object.

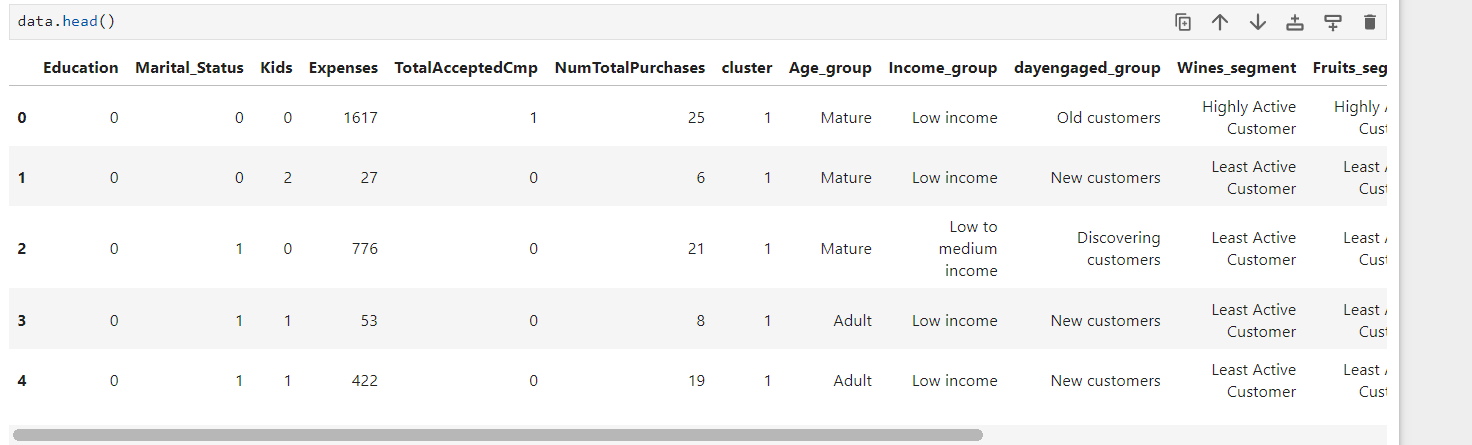
**Thay thế các giá trị NaN**:

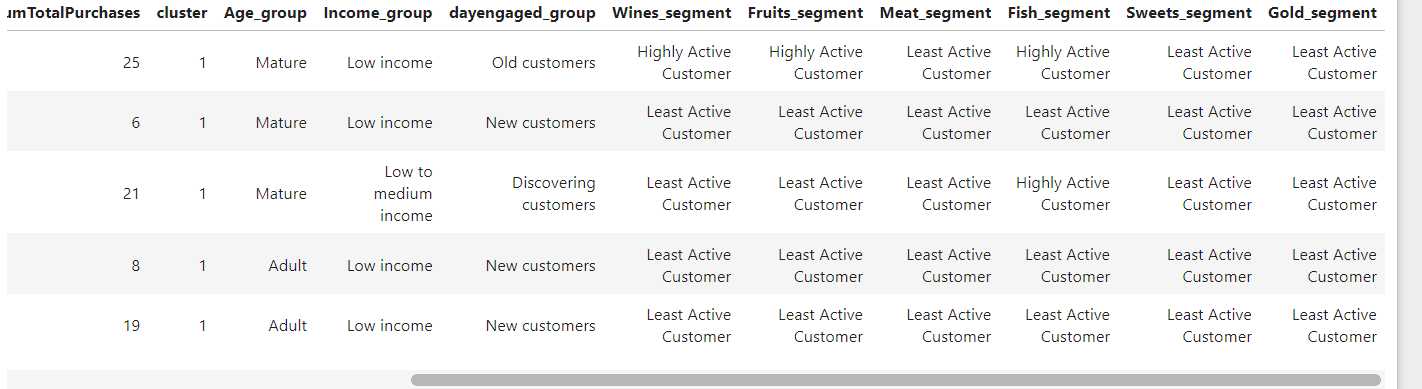
Thay thế các giá trị NaN (khách hàng không mua sản phẩm) bằng "Inactive Customer".

**Chuyển đổi kiểu dữ liệu của DataFrame**:

Chuyển đổi kiểu dữ liệu của toàn bộ DataFrame data sang kiểu object.

**Kết quả của các phân đoạn:**





**Dòng 0**:

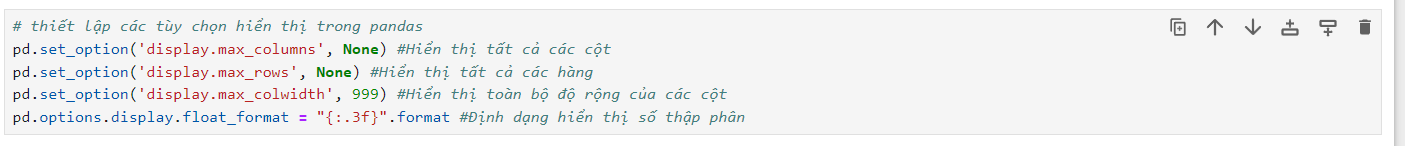
* Khách hàng không có dữ liệu về trình độ học vấn và tình trạng hôn nhân, không có con.
* Tổng chi tiêu là 1617, đã chấp nhận 1 chiến dịch tiếp thị, tổng cộng có 25 lần mua hàng.
* Thuộc nhóm phân đoạn 1.
* Nhóm độ tuổi: Mature, thu nhập: Low income, tương tác: Old customers.
* Mức độ hoạt động: Rượu vang, trái cây, và cá: Highly Active Customer, thịt, bánh kẹo và vàng: Least Active Customer.

**Dòng 1**:

* Khách hàng không có dữ liệu về trình độ học vấn và tình trạng hôn nhân, có 2 con.
* Tổng chi tiêu là 27, không chấp nhận chiến dịch tiếp thị nào, có 6 lần mua hàng.
* Thuộc nhóm phân đoạn 1.
* Nhóm độ tuổi: Mature, thu nhập: Low income, tương tác: New customers.
* Mức độ hoạt động: tất cả sản phẩm: Least Active Customer.

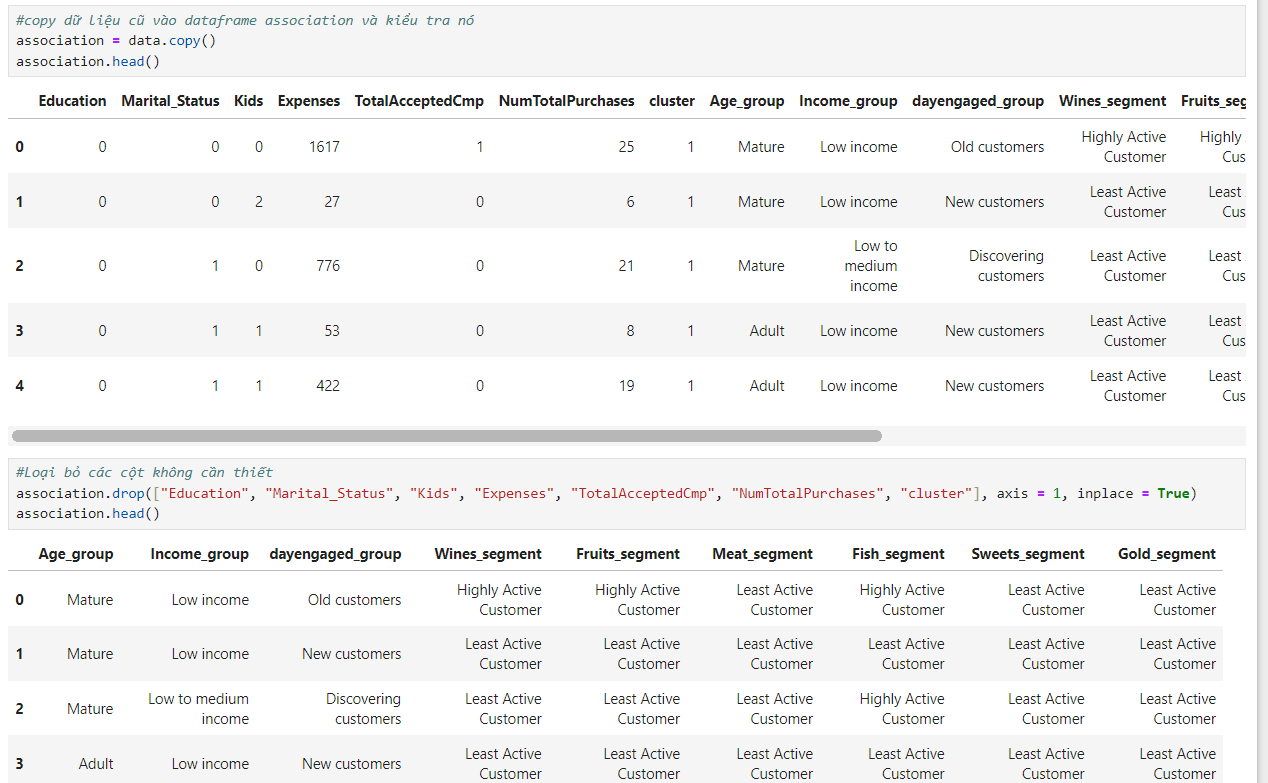
### 3.2.2. Applying Apriori Algorithm

Thiết lập các tùy chọn hiển thị trong pandas:

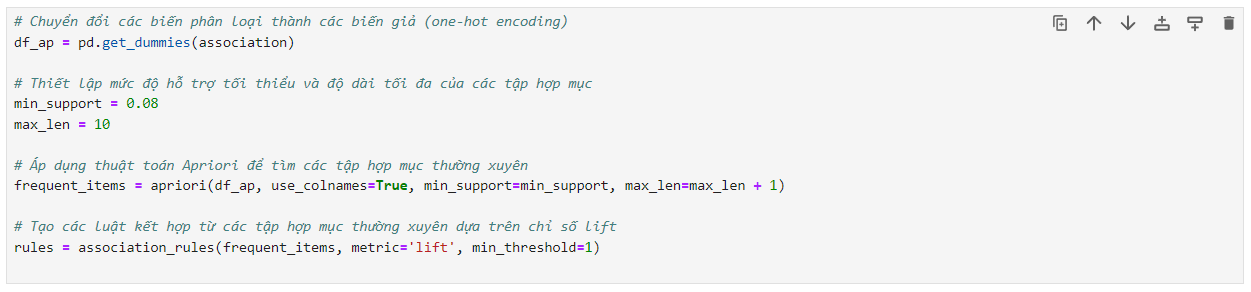


* Thiết lập độ rộng tối đa của các cột hiển thị. Đặt giá trị lớn (999) để đảm bảo rằng toàn bộ nội dung của mỗi cột được hiển thị mà không bị cắt ngắn.
* Định dạng hiển thị số thập phân, đảm bảo rằng các giá trị số sẽ được hiển thị với 3 chữ số thập phân. Điều này giúp chuẩn hóa và làm rõ hơn dữ liệu số trong DataFrame.

Coppy dữ liệu vào dataframe mới và loại bỏ các cột không cần thiết:



### 3.2.3. Setting the Association Rules



**Chuyển đổi các biến phân loại thành các biến giả**:

 pd.get\_dummies(association): Chuyển đổi các cột phân loại trong DataFrame association thành các cột biến giả (one-hot encoding). Mỗi giá trị trong các cột phân loại sẽ được biểu diễn dưới dạng cột riêng biệt với giá trị 0 hoặc 1.

 Kết quả là DataFrame df\_ap, chứa các cột mới là các biến giả cho mỗi giá trị phân loại ban đầu.

**Thiết lập mức độ hỗ trợ tối thiểu và độ dài tối đa của các tập hợp mục**:

 min\_support = 0.08: Đặt mức độ hỗ trợ tối thiểu là 0.08 (8%). Điều này có nghĩa là chỉ những tập hợp mục xuất hiện trong ít nhất 8% của tất cả các giao dịch mới được xem xét.

 max\_len = 10: Đặt độ dài tối đa của các tập hợp mục là 10.

**Áp dụng thuật toán Apriori để tìm các tập hợp mục thường xuyên**:

apriori(df\_ap, use\_colnames=True, min\_support=min\_support, max\_len=max\_len + 1): Hàm này thực hiện thuật toán Apriori trên DataFrame df\_ap.

* use\_colnames=True: Giữ lại tên cột ban đầu trong kết quả.
* min\_support=min\_support: Sử dụng mức độ hỗ trợ tối thiểu đã thiết lập trước đó.
* max\_len=max\_len + 1: Đặt độ dài tối đa của các tập hợp mục là max\_len + 1, ở đây là 11.

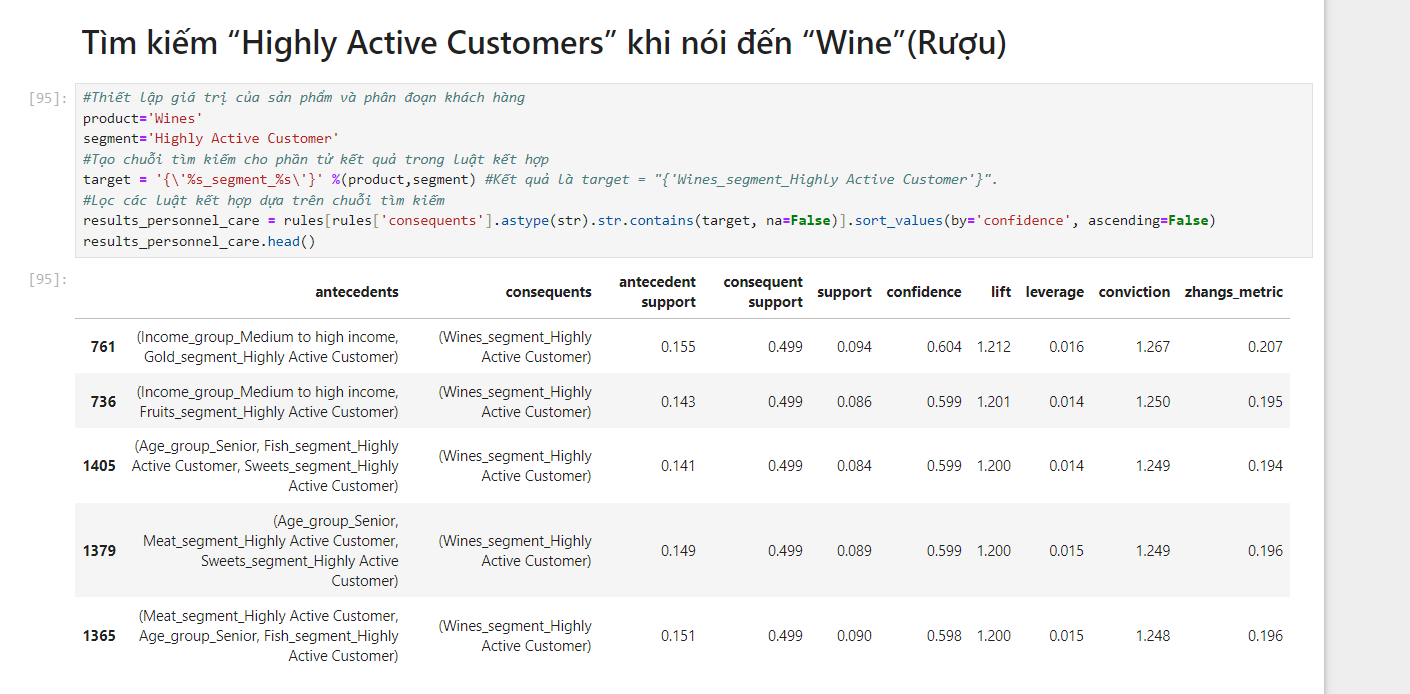
**Tạo các luật kết hợp từ các tập hợp mục thường xuyên**:

association\_rules(frequent\_items, metric='lift', min\_threshold=1): Hàm này tạo các luật kết hợp từ các tập hợp mục thường xuyên.

* metric='lift': Sử dụng chỉ số lift để đánh giá các luật kết hợp.
* min\_threshold=1: Chỉ những luật kết hợp có lift lớn hơn hoặc bằng 1 mới được xem xét.

### 3.2.4. Tìm kiếm "Highly Active Customers" theo các sản phẩm:

#### 3.2.4.1. Tìm kiếm "Highly Active Customers" khi nói đến "Wine"



**Thiết lập giá trị của sản phẩm và phân đoạn khách hàng**:

* product='Wines': Đặt giá trị sản phẩm là "Wines".
* segment='Highly Active Customer': Đặt giá trị phân đoạn khách hàng là "Highly Active Customer".

**Tạo chuỗi tìm kiếm cho phần tử kết quả trong luật kết hợp**:

* target = '{\'%s\_segment\_%s\'}' %(product,segment): Tạo chuỗi tìm kiếm cho phần tử kết quả trong luật kết hợp.
  + Sử dụng định dạng chuỗi (%s\_segment\_%s) để chèn các giá trị của product và segment vào trong chuỗi.
  + Kết quả là target = "{'Wines\_segment\_Highly Active Customer'}"

**Lọc các luật kết hợp dựa trên chuỗi tìm kiếm**:

* rules['consequents'].astype(str).str.contains(target, na=False): Chuyển đổi cột 'consequents' trong DataFrame rules thành chuỗi và kiểm tra xem có chứa target hay không.
  + rules['consequents'].astype(str): Chuyển đổi giá trị của cột 'consequents' thành chuỗi.
  + .str.contains(target, na=False): Kiểm tra xem chuỗi có chứa target không. na=False để bỏ qua giá trị NaN.
* rules[rules['consequents'].astype(str).str.contains(target, na=False)]: Lọc các luật kết hợp mà phần tử kết quả (consequents) chứa target.
* .sort\_values(by='confidence', ascending=False): Sắp xếp các luật đã lọc theo giá trị 'confidence' giảm dần.

Dưới đây là giải thích chi tiết về các cột trong bảng kết quả và ý nghĩa của các giá trị được hiển thị:

1. **antecedents**: Các mục hoặc tập hợp mục xảy ra trước trong luật kết hợp (điều kiện của luật).
2. **consequents**: Các mục hoặc tập hợp mục xảy ra sau trong luật kết hợp (kết quả của luật).
3. **antecedent support**: Tỷ lệ các giao dịch chứa các mục trong antecedents.
4. **consequent support**: Tỷ lệ các giao dịch chứa các mục trong consequents.
5. **support**: Tỷ lệ các giao dịch chứa cả antecedents và consequents. Đây là xác suất xảy ra đồng thời của cả hai tập hợp mục trong toàn bộ dữ liệu.
6. **confidence**: Xác suất xảy ra của consequents khi antecedents đã xảy ra. Đây là tỉ lệ giữa support và antecedent support.
7. **lift**: Tỷ lệ giữa confidence và consequent support. Lift đo lường mức độ cải thiện dự đoán sự xuất hiện của consequents khi biết antecedents so với khi không biết gì.
8. **leverage**: Sự khác biệt giữa support của cả hai tập hợp mục (antecedents và consequents) và support dự kiến nếu chúng độc lập. Leverage càng cao thì mối quan hệ giữa hai tập hợp mục càng mạnh.
9. **conviction**: Đo lường xác suất mà luật sẽ đúng dựa trên xác suất sai của nó. Conviction cao hơn 1 cho thấy mối quan hệ mạnh hơn giữa antecedents và consequents.
10. **zhangs\_metric**: Một thước đo đánh giá sức mạnh của luật kết hợp, dựa trên các giá trị hỗ trợ, độ tin cậy, và lift.

Dòng 761:

* **antecedents**: Income\_group\_Medium to high income, Gold\_segment\_Highly Active Customer
* **consequents**: Wines\_segment\_Highly Active Customer
* **antecedent support**: 0.155 (15.5% các giao dịch chứa các mục trong antecedents)
* **consequent support**: 0.499 (49.9% các giao dịch chứa các mục trong consequents)
* **support**: 0.094 (9.4% các giao dịch chứa cả antecedents và consequents)
* **confidence**: 0.604 (60.4% khả năng consequents xảy ra khi antecedents đã xảy ra)
* **lift**: 1.212 (Khả năng xảy ra của consequents cao hơn 1.212 lần khi biết antecedents)
* **leverage**: 0.016 (Sự khác biệt giữa support của cả hai tập hợp mục và support dự kiến nếu chúng độc lập)
* **conviction**: 1.267 (Xác suất mà luật sẽ đúng dựa trên xác suất sai của nó)
* **zhangs\_metric**: 0.207 (Thước đo đánh giá sức mạnh của luật)

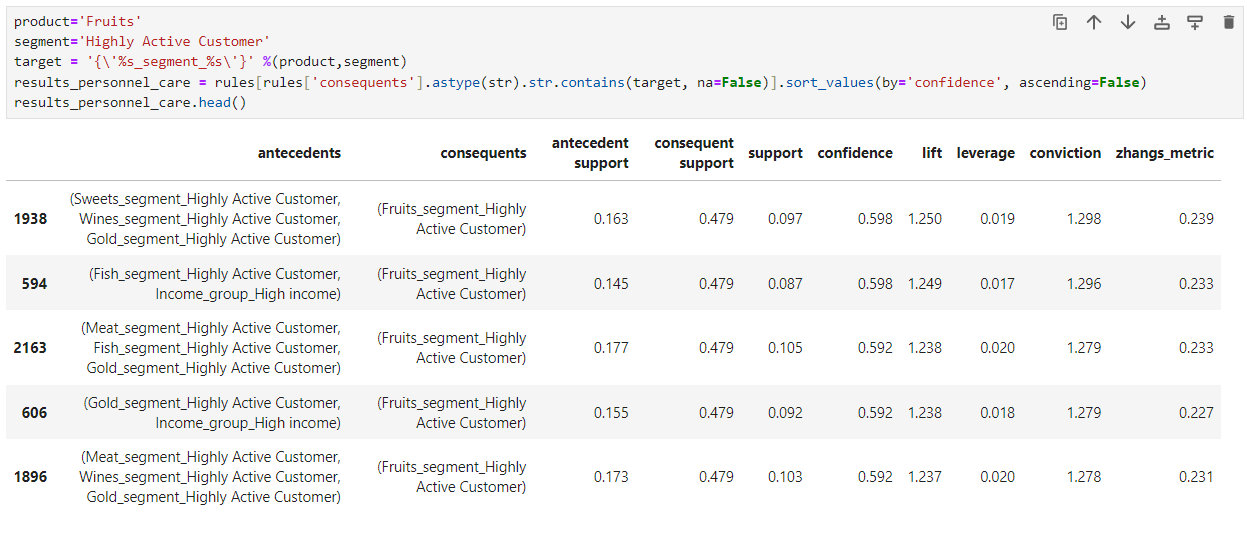
**Mối quan hệ và ý nghĩa:**

Kết quả của luật kết hợp cung cấp những hiểu biết quan trọng về các mẫu hành vi của khách hàng, đặc biệt là trong việc xác định các khách hàng "Highly Active" đối với sản phẩm rượu (Wine). Dưới đây là những lợi ích cụ thể:

1. **Xác định nhóm khách hàng có khả năng mua rượu cao**:
   * Luật kết hợp chỉ ra rằng nếu khách hàng thuộc các nhóm thu nhập trung bình đến cao (Income\_group\_Medium to high income) và cũng là khách hàng "Highly Active" trong các danh mục khác như vàng (Gold\_segment), trái cây (Fruits\_segment), hoặc các sản phẩm khác, thì khả năng cao họ cũng sẽ là khách hàng "Highly Active" đối với rượu.
2. **Phân tích tập khách hàng mục tiêu**:
   * Bằng cách biết các yếu tố liên quan đến nhóm khách hàng "Highly Active" đối với rượu, bạn có thể tập trung vào các chiến lược tiếp thị nhắm đến các phân khúc này. Ví dụ: Những khách hàng cao tuổi (Age\_group\_Senior) và có hành vi mua sắm tích cực đối với cá (Fish\_segment) hoặc thịt (Meat\_segment) cũng có khả năng cao sẽ mua rượu.
3. **Tối ưu hóa chiến lược marketing**:

* Những kết quả này cho phép bạn điều chỉnh các chiến dịch tiếp thị để nhắm vào các phân khúc khách hàng cụ thể. Chẳng hạn kết hợp sản phẩm thịt, cá và rượu cho khách hàng cao tuổi.
* Thiết kế ưu đãi đặc biệt hoặc chương trình khách hàng thân thiết cho các khách hàng "Highly Active" trong các danh mục như trái cây hoặc vàng để khuyến khích họ mua thêm rượu.

#### 3.2.4.2. Tìm kiếm "Highly Active Customers" khi nói đến "Fruits"(Trái cây)

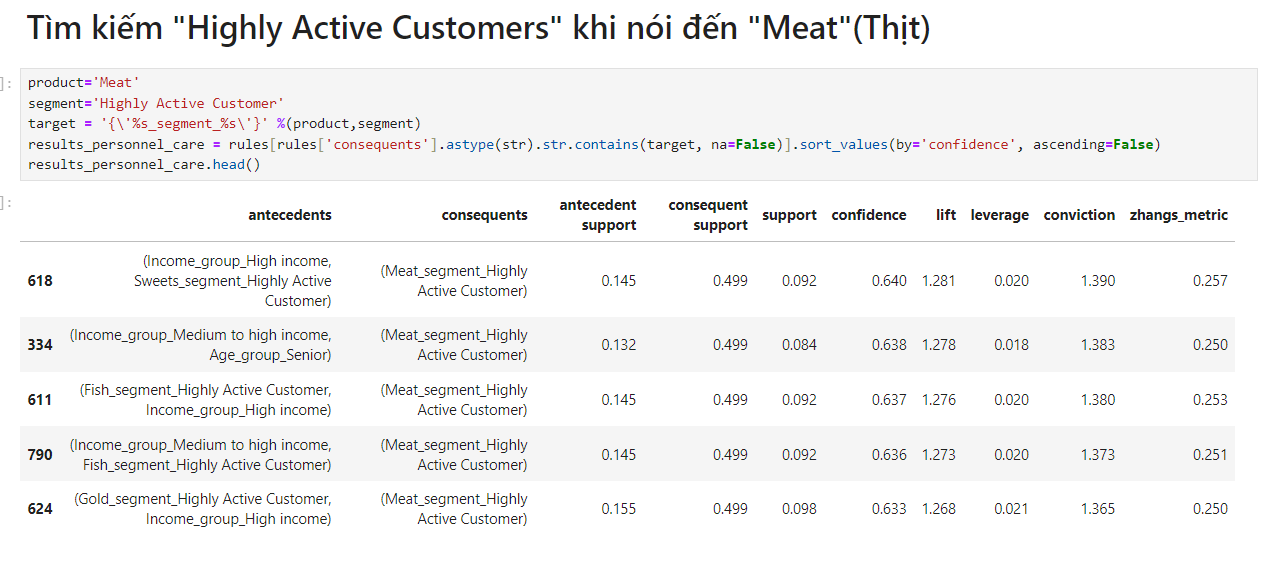


**Mối quan hệ và ý nghĩa:**

Kết quả của các luật kết hợp cung cấp cái nhìn sâu sắc về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm trái cây (Fruits). Dưới đây là những lợi ích cụ thể:

1. **Xác định nhóm khách hàng có khả năng mua trái cây cao**:
   * Các luật kết hợp chỉ ra rằng nhóm khách hàng "Highly Active" đối với các danh mục sản phẩm như kẹo (Sweets\_segment), rượu (Wines\_segment), và vàng (Gold\_segment) có khả năng cao sẽ cũng là nhóm khách hàng "Highly Active" đối với trái cây.
   * Nhóm khách hàng "Highly Active" đối với các sản phẩm khác như cá (Fish\_segment) hoặc thịt (Meat\_segment) cũng có khả năng mua trái cây cao.
2. **Phân tích tập khách hàng mục tiêu**:
   * Bằng cách hiểu các nhóm khách hàng có khả năng cao mua trái cây, bạn có thể tập trung vào việc phát triển các chiến lược tiếp thị và chương trình khuyến mãi đặc biệt để thu hút và duy trì nhóm khách hàng này.
   * Ví dụ, việc tạo ra các gói ưu đãi hoặc chương trình thưởng cho nhóm khách hàng "Highly Active" trong các danh mục khác như rượu và vàng có thể tăng cường việc mua sắm trái cây từ họ.
3. **Tối ưu hóa chiến lược marketing**:
   * Hiểu biết về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm trái cây giúp bạn tối ưu hóa chiến lược tiếp thị. Ví dụ, bạn có thể tạo ra các chiến dịch tiếp thị đa kênh đặc biệt để quảng bá sản phẩm trái cây đến nhóm khách hàng "Highly Active" trong các danh mục sản phẩm khác.
4. **Hiểu biết sâu sắc về hành vi mua sắm của khách hàng**:
   * Các luật kết hợp giúp bạn hiểu rõ hơn về các đặc điểm và hành vi mua sắm của nhóm khách hàng "Highly Active". Điều này giúp bạn phân khúc khách hàng một cách chi tiết và chính xác hơn, từ đó tạo ra các chiến lược kinh doanh phù hợp.

#### 3.2.4.3. Tìm kiếm "Highly Active Customers" khi nói đến "Meat"(Thịt)



**Mối quan hệ và ý nghĩa:**

Kết quả của các luật kết hợp cung cấp cái nhìn sâu sắc về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm thịt (Meat). Dưới đây là những lợi ích cụ thể:

1. **Xác định nhóm khách hàng có khả năng mua thịt cao**:
   * Các luật kết hợp chỉ ra rằng nhóm khách hàng "Highly Active" đối với các sản phẩm như kẹo (Sweets\_segment) và có thu nhập cao (Income\_group\_High income) có khả năng cao sẽ cũng là nhóm khách hàng "Highly Active" đối với thịt.
   * Nhóm khách hàng "Highly Active" đối với các danh mục khác như cá (Fish\_segment) hoặc thu nhập trung bình đến cao (Income\_group\_Medium to high income) cũng có khả năng mua thịt cao.
2. **Phân tích tập khách hàng mục tiêu**:
   * Bằng cách hiểu các nhóm khách hàng có khả năng cao mua thịt, bạn có thể tập trung vào việc phát triển các chiến lược tiếp thị và chương trình khuyến mãi đặc biệt để thu hút và duy trì nhóm khách hàng này.
   * Ví dụ, việc tạo ra các gói ưu đãi hoặc chương trình thưởng cho nhóm khách hàng "Highly Active" trong các danh mục khác như vàng có thể tăng cường việc mua sắm thịt từ họ.
3. **Tối ưu hóa chiến lược marketing**:
   * Hiểu biết về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm thịt giúp bạn tối ưu hóa chiến lược tiếp thị. Ví dụ, bạn có thể tạo ra các chiến dịch tiếp thị đa kênh đặc biệt để quảng bá sản phẩm thịt đến nhóm khách hàng "Highly Active" trong các danh mục sản phẩm khác.

#### 3.2.4.4. Tìm kiếm "Highly Active Customers" khi nói đến "Fish"(Cá)



**Mối quan hệ và ý nghĩa:**

Kết quả của các luật kết hợp cung cấp cái nhìn sâu sắc về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm cá (Fish). Dưới đây là những lợi ích cụ thể:

1. **Xác định nhóm khách hàng có khả năng mua cá cao**:
   * Các luật kết hợp chỉ ra rằng nhóm khách hàng "Highly Active" đối với các sản phẩm như kẹo (Sweets\_segment) và có thu nhập cao (Income\_group\_High income) có khả năng cao sẽ cũng là nhóm khách hàng "Highly Active" đối với cá.
   * Các nhóm khách hàng "Highly Active" đối với các danh mục khác như thịt (Meat\_segment), rượu (Wines\_segment), hoặc trái cây (Fruits\_segment) cũng có khả năng mua cá cao.
2. **Phân tích tập khách hàng mục tiêu**:
   * Bằng cách hiểu các nhóm khách hàng có khả năng cao mua cá, bạn có thể tập trung vào việc phát triển các chiến lược tiếp thị và chương trình khuyến mãi đặc biệt để thu hút và duy trì nhóm khách hàng này.
   * Ví dụ, việc tạo ra các gói ưu đãi hoặc chương trình thưởng cho nhóm khách hàng "Highly Active" trong các danh mục khác như rượu và thịt có thể tăng cường việc mua sắm cá từ họ.
3. **Tối ưu hóa chiến lược marketing**:
   * Hiểu biết về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm cá giúp bạn tối ưu hóa chiến lược tiếp thị. Ví dụ, bạn có thể tạo ra các chiến dịch tiếp thị đa kênh đặc biệt để quảng bá sản phẩm cá đến nhóm khách hàng "Highly Active" trong các danh mục sản phẩm khác.

#### 3.2.4.5. Tìm kiếm "Highly Active Customers" khi nói đến "Sweets"(Đồ ngọt)



**Mối quan hệ và ý nghĩa:**

Kết quả của các luật kết hợp cung cấp cái nhìn sâu sắc về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm đồ ngọt (Sweets). Dưới đây là những lợi ích cụ thể:

1. **Xác định nhóm khách hàng có khả năng mua đồ ngọt cao**:
   * Các luật kết hợp chỉ ra rằng nhóm khách hàng "Highly Active" đối với các sản phẩm như vàng (Gold\_segment) và có thu nhập từ trung đến cao (Income\_group\_Medium to high income) có khả năng cao sẽ cũng là nhóm khách hàng "Highly Active" đối với đồ ngọt.
   * Nhóm khách hàng "Highly Active" đối với các danh mục khác như cá (Fish\_segment), rượu (Wines\_segment), hoặc trái cây (Fruits\_segment) cũng có khả năng mua đồ ngọt cao.
2. **Phân tích tập khách hàng mục tiêu**:
   * Bằng cách hiểu các nhóm khách hàng có khả năng cao mua đồ ngọt, bạn có thể tập trung vào việc phát triển các chiến lược tiếp thị và chương trình khuyến mãi đặc biệt để thu hút và duy trì nhóm khách hàng này.
   * Ví dụ, việc tạo ra các gói ưu đãi hoặc chương trình thưởng cho nhóm khách hàng "Highly Active" trong các danh mục khác như vàng có thể tăng cường việc mua sắm đồ ngọt từ họ.
3. **Tối ưu hóa chiến lược marketing**:
   * Hiểu biết về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua sắm đồ ngọt giúp bạn tối ưu hóa chiến lược tiếp thị. Ví dụ, bạn có thể tạo ra các chiến dịch tiếp thị đa kênh đặc biệt để quảng bá sản phẩm đồ ngọt đến nhóm khách hàng "Highly Active" trong các danh mục sản phẩm khác.

#### 3.2.4.6. Tìm kiếm "Highly Active Customers" khi nói đến "Gold"



**Mối quan hệ và ý nghĩa**

Kết quả của các luật kết hợp cung cấp cái nhìn sâu sắc về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua vàng. Dưới đây là những lợi ích cụ thể:

1. **Xác định nhóm khách hàng có khả năng mua vàng cao**:
   * Các luật kết hợp chỉ ra rằng nhóm khách hàng "Highly Active" đối với các sản phẩm như rượu (Wines\_segment) và có thu nhập từ trung đến cao (Income\_group\_Medium to high income) có khả năng cao sẽ cũng là nhóm khách hàng "Highly Active" đối với vàng.
   * Nhóm khách hàng "Highly Active" đối với các danh mục khác như trái cây (Fruits\_segment) hoặc đồ ngọt (Sweets\_segment) cũng có khả năng mua vàng cao.
2. **Phân tích tập khách hàng mục tiêu**:
   * Bằng cách hiểu các nhóm khách hàng có khả năng cao mua vàng, bạn có thể tập trung vào việc phát triển các chiến lược tiếp thị và chương trình khuyến mãi đặc biệt để thu hút và duy trì nhóm khách hàng này.
   * Ví dụ, việc tạo ra các gói ưu đãi hoặc chương trình thưởng cho nhóm khách hàng "Highly Active" trong các danh mục khác như rượu và trái cây có thể tăng cường việc mua vàng từ họ.
3. **Tối ưu hóa chiến lược marketing**:
   * Hiểu biết về mối quan hệ giữa các nhóm khách hàng "Highly Active" và việc mua vàng giúp bạn tối ưu hóa chiến lược tiếp thị. Ví dụ, bạn có thể tạo ra các chiến dịch tiếp thị đa kênh đặc biệt để quảng bá sản phẩm vàng đến nhóm khách hàng "Highly Active" trong các danh mục sản phẩm khác.

### 3.2.5. Đánh giá mô hình:

Bằng cách sử dụng luật kết hợp, nếu chúng ta có thể tìm thấy Người tiêu dùng lớn nhất cho một sản phẩm cụ thể, thì nhiều cách tiếp thị những sản phẩm này có thể được thu hẹp cho phù hợp với những khách hàng này và nhu cầu của họ.

Do đó, thuật toán Apriori giúp doanh nghiệp tìm ra nhóm khách hàng lớn nhất cho một sản phẩm cụ thể, từ đó thu hẹp chiến lược tiếp thị và định giá sản phẩm phù hợp. Điều này giúp cải thiện định vị sản phẩm và nâng cao hiệu quả tiếp thị dựa trên nhu cầu của những khách hàng tiềm năng.

Tuy nhiên, Association Rule Mining có thể bị ảnh hưởng nghiêm trọng bởi dữ liệu nhiễu hoặc dữ liệu không đồng nhất và gây khó khăn trong việc diễn giải kết quả vì một số luật kết hợp có thể rất phức tạp và khó hiểu, đặc biệt khi có nhiều biến và điều kiện.

## 3.3. CLASSIFICATION MODEL

*Dự đoán Hoạt động Mua hàng của Khách hàng với 2 Nhóm: Cluster 1 - Khách hàng tích cực mua hàng và Cluster 2 - Khách hàng ít hoạt động mua hàng."*

**Đầu vào (input)** là các biến đầu vào được sử dụng để dự đoán hành vi mua hàng của khách hàng. Các biến đầu vào bao gồm:

* Education (Giáo dục)
* Marital\_Status (Tình trạng hôn nhân)
* Income (Thu nhập)
* Kids (Con cái)
* Expenses (Chi phí)
* TotalAcceptedCmp (Tổng số chiến dịch quảng cáo được chấp nhận)
* NumTotalPurchases (Tổng số lần mua hàng)
* Age (Tuổi)
* day\_engaged (Số ngày đã tương tác)

**Đầu ra (output)** là nhãn phân loại cho từng khách hàng, xác định liệu họ thuộc vào nhóm mua hàng tích cực (Highly Active Customer) hay nhóm ít hoạt động mua hàng (Least Active Customer), được gọi là cluster.

Trong một bài toán phân loại như vậy, dữ liệu đầu vào sẽ được sử dụng để dự đoán nhãn phân loại cho các trường hợp mới hoặc chưa biết trước. Cụ thể, các thuộc tính của mỗi khách hàng (ví dụ: giáo dục, thu nhập, tuổi, vv.) sẽ được sử dụng để dự đoán xác suất rằng khách hàng đó thuộc vào nhóm mua hàng nào. Đầu ra sẽ là nhãn của nhóm mua hàng mà mô hình dự đoán cho mỗi khách hàng.

**Chuẩn bị dữ liệu**

df2=df1.copy()

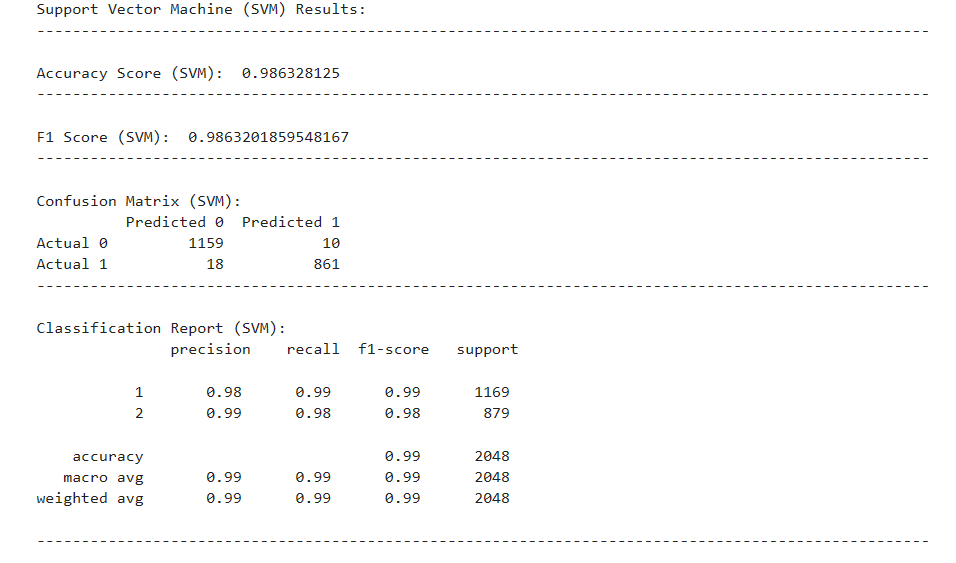
x = df2.drop('cluster', axis=1)

y = df2['cluster']

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=2)

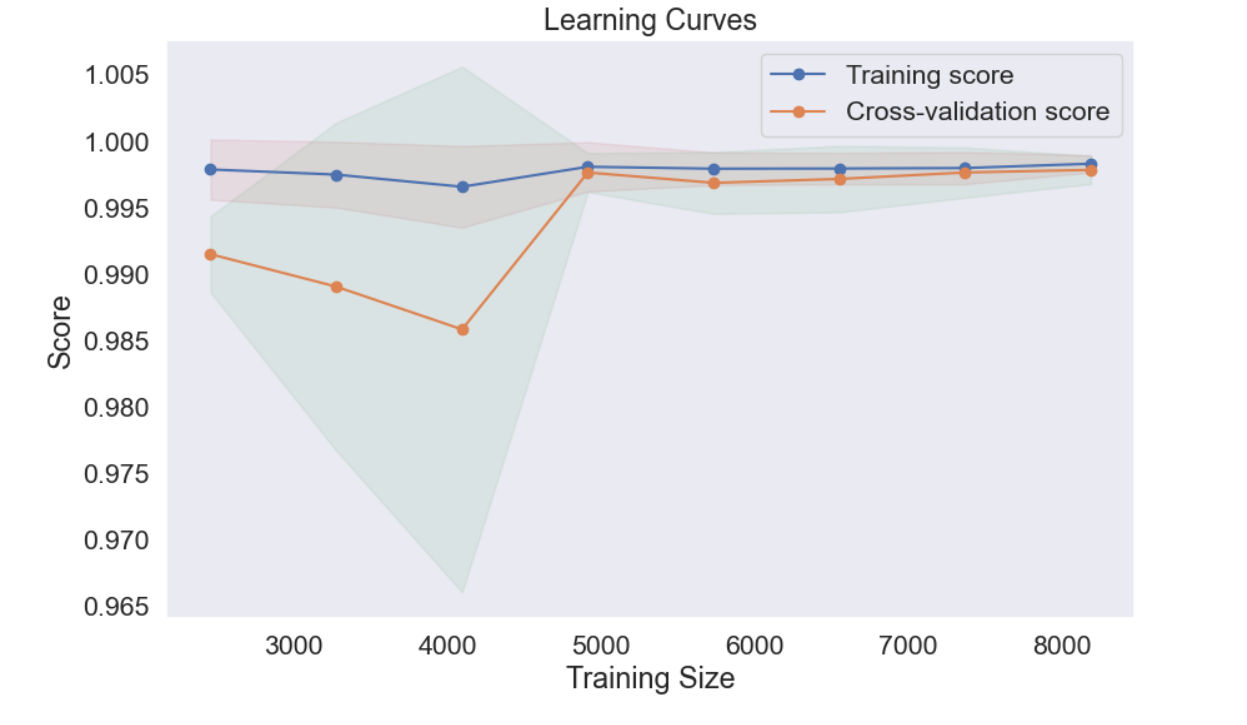
### 3.3.1. SVC

Lý do chọn mô hình: SVM thường hoạt động tốt trên các tập dữ liệu có cấu trúc phức tạp và khi có một ranh giới phân chia rõ ràng giữa các lớp dữ liệu. Điều này phù hợp với nhu cầu của bài toán.



Mô hình SVM hoạt động rất tốt trong việc phân loại khách hàng thành hai nhóm: Highly Active Customer và Least Active Customer. Các chỉ số đo lường đều cao, cho thấy mô hình không chỉ chính xác mà còn cân bằng tốt giữa việc dự đoán đúng và phản ứng với tất cả các nhóm khách hàng. Điều này chứng tỏ mô hình SVM là một công cụ mạnh mẽ và hiệu quả trong việc phân loại hành vi mua hàng của khách hàng.

**Kiểm tra overfiting**



**Nhận xét:**

*Overfitting:*

Không có dấu hiệu overfitting rõ ràng. Điểm số huấn luyện và kiểm tra chéo đều rất cao (~1.0 và 0.99) và gần nhau, cho thấy mô hình tổng quát hóa tốt trên dữ liệu mới.

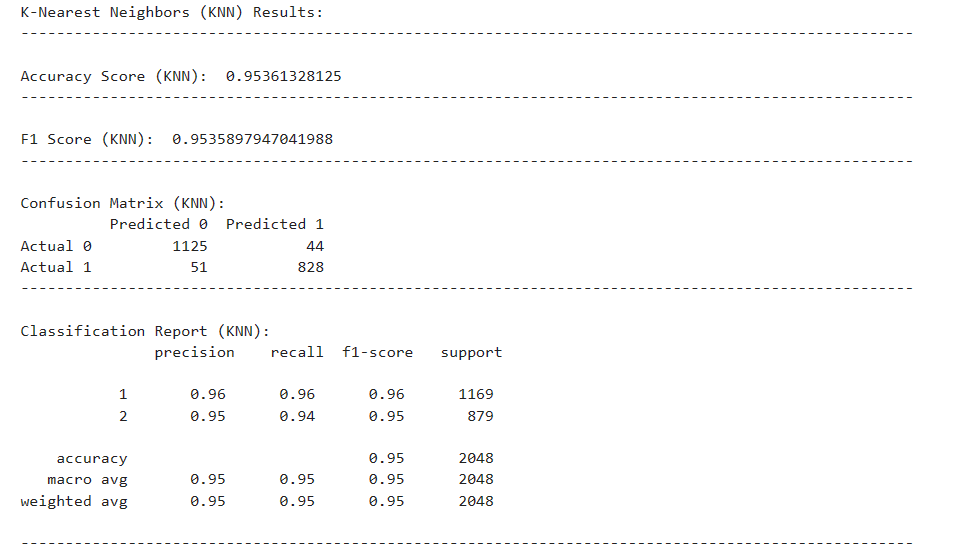
*Underfitting:*

Không có dấu hiệu underfitting trong đồ thị này, vì cả điểm số huấn luyện và kiểm tra chéo đều rất cao từ đầu.

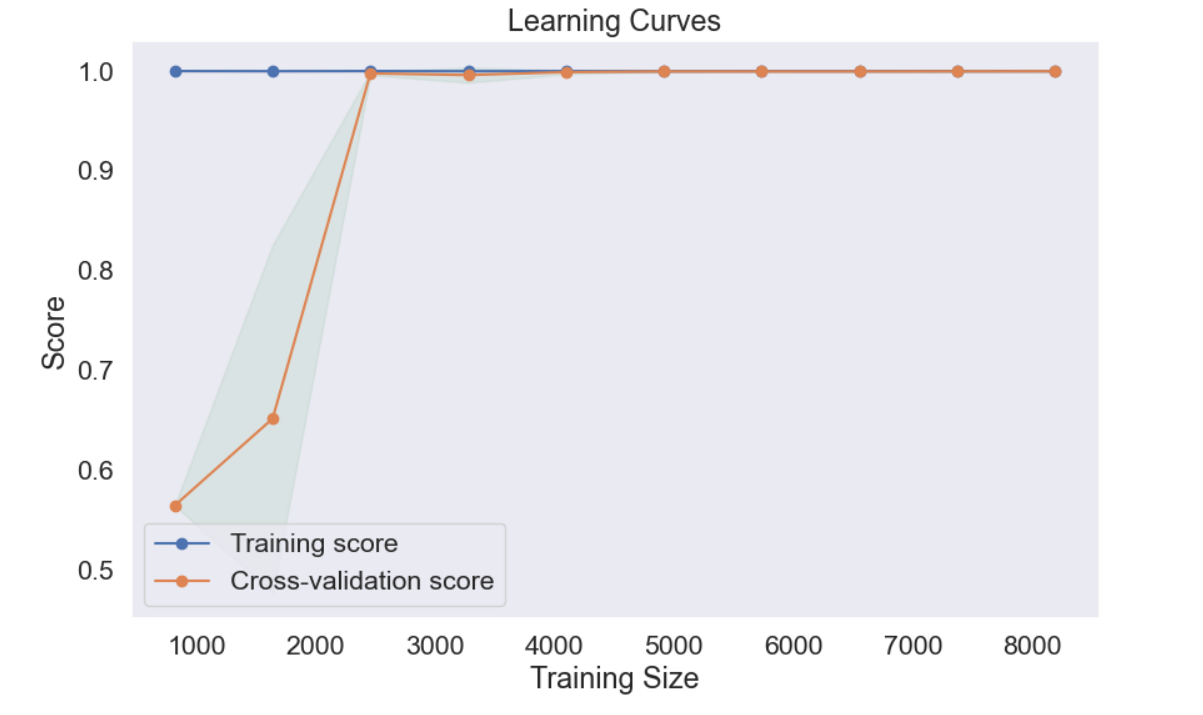
Không có dấu hiệu của overfitting hay underfitting, mô hình tổng quát hóa tốt ngay từ đầu.

### 3.3.2. KNeighborsClassifier

Lý do chọn mô hình: KNN đơn giản và không yêu cầu giả định về phân phối của dữ liệu, có thể phù hợp với việc xác định nhóm khách hàng dựa trên hành vi mua hàng tương tự



Mô hình KNN hoạt động hiệu quả trong việc phân loại khách hàng thành hai nhóm: Highly Active Customer và Least Active Customer. Các chỉ số đo lường đều cao, cho thấy mô hình không chỉ chính xác mà còn cân bằng tốt giữa việc dự đoán đúng và phản ứng với tất cả các nhóm khách hàng. Điều này chứng tỏ mô hình KNN là một công cụ mạnh mẽ và hiệu quả trong việc phân loại hành vi mua hàng của khách hàng.

Kiểm tra overfiting

**Nhận xét:**

*Overfitting:*

Ban đầu, mô hình có thể gặp phải underfitting (với điểm số kiểm tra chéo thấp, khoảng 0.55) nhưng không phải overfitting.

Sau khi tăng kích thước tập huấn luyện lên đến khoảng 2000 mẫu, điểm số kiểm tra chéo tăng lên và gần bằng điểm số huấn luyện (cả hai đều gần 1.0), cho thấy không có hiện tượng overfitting.

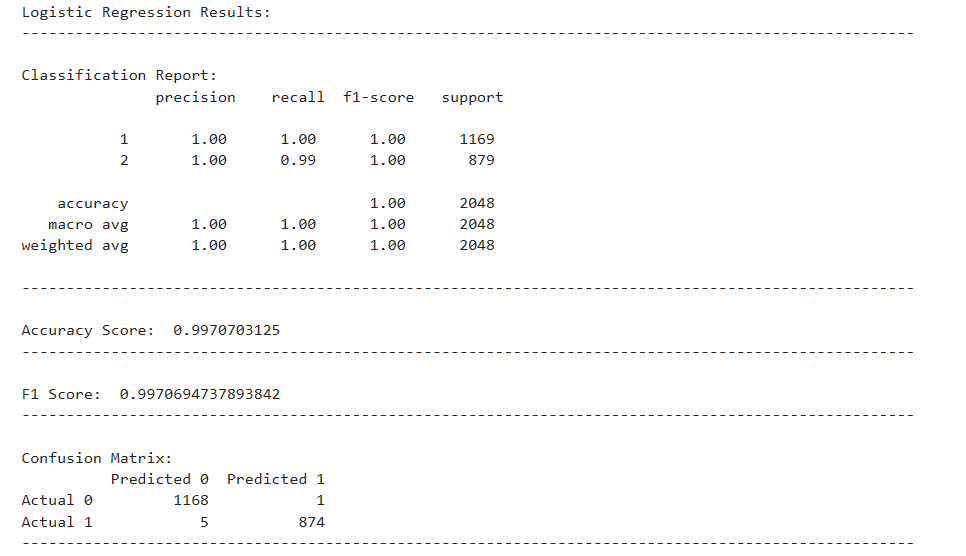
*Underfitting:*

Ban đầu có dấu hiệu underfitting khi điểm số kiểm tra chéo rất thấp và cách xa điểm số huấn luyện. Tuy nhiên, điều này nhanh chóng được cải thiện khi kích thước tập huấn luyện tăng lên.

Ban đầu có dấu hiệu underfitting nhưng cải thiện nhanh chóng và không có overfitting khi kích thước tập huấn luyện tăng.

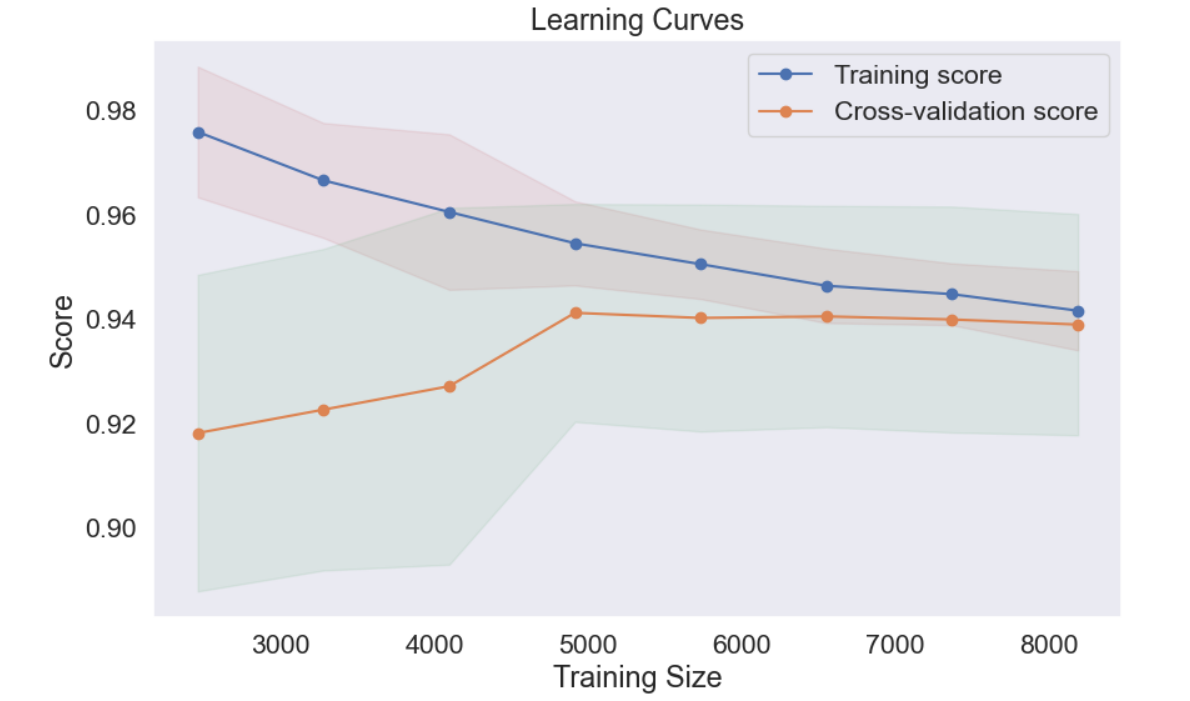
### 3.3.3. LogisticRegression

Lý do chọn mô hình: Logistic Regression thường được sử dụng trong các vấn đề phân loại nhị phân và cung cấp xác suất dự đoán, phù hợp với bài toán này.



Mô hình Logistic Regression hoạt động cực kỳ hiệu quả trong việc phân loại khách hàng thành hai nhóm: Highly Active Customer và Least Active Customer. Các chỉ số đo lường đều gần như hoàn hảo, cho thấy mô hình không chỉ chính xác mà còn có sự cân bằng tốt giữa việc dự đoán đúng và phản ứng với tất cả các nhóm khách hàng. Điều này chứng tỏ mô hình Logistic Regression là một công cụ mạnh mẽ và hiệu quả trong việc phân loại hành vi mua hàng của khách hàng.

**Kiểm tra overfiting¶**



**Nhận xét:**

*Overfitting:*

Ban đầu, mô hình có dấu hiệu overfitting. Điểm số huấn luyện rất cao (~0.98) nhưng điểm số kiểm tra chéo thấp hơn (~0.92), cho thấy mô hình học rất tốt trên tập huấn luyện nhưng không tổng quát hóa tốt trên dữ liệu mới.

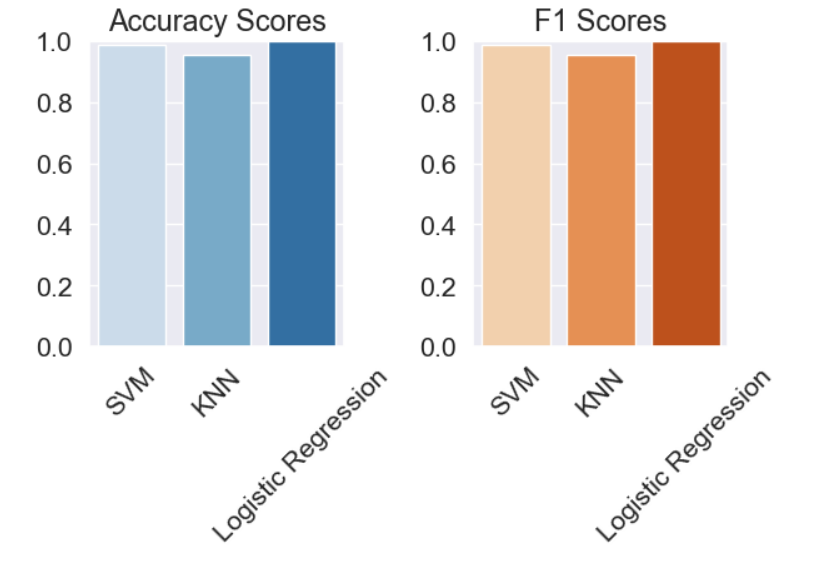
Khi kích thước tập huấn luyện tăng, khoảng cách giữa điểm số huấn luyện và kiểm tra chéo giảm, cho thấy hiện tượng overfitting giảm dần.

*Underfitting:*

Không có dấu hiệu rõ ràng của underfitting trong đồ thị này, vì cả điểm số huấn luyện và kiểm tra chéo đều khá cao.

Ban đầu có dấu hiệu overfitting nhưng giảm dần khi kích thước tập huấn luyện tăng.

### 3.3.4. So sánh các mô hình



Có thể thấy các mô hình đều có độ chính xác cao và điểm f1 cũng cao, thể hiện các mô hình hoạt động rất tốt trên tập dữ liệu, cũng như các biểu đồ trực quan hóa cho thấy việc tăng kích thước tập huấn luyện giúp giảm thiểu overfitting trong đồ thị 1 và underfitting trong đồ thị 2, trong khi đồ thị 3 cho thấy một mô hình mạnh mẽ với khả năng tổng quát hóa tốt ngay từ đầu..

**Ưu và nhược điểm:**

1. Support Vector Machine (SVM):

Nguyên tắc: Tìm đường phân cách tối ưu (hyperplane) giữa các nhóm dữ liệu

Ưu điểm: Hiệu quả trong xử lý dữ liệu có chiều cao, khả năng tổng quát hóa tốt, xử lý tốt dữ liệu phi tuyến tính (khi sử dụng kernel

Nhược điểm: Có thể gặp khó khăn trong xử lý các tập dữ liệu rất lớn, việc lựa chọn kernel phù hợp có thể phức tàng của khách hàng.

2. K-Nearest Neighbors (KNN):

Nguyên tắc: Dự đoán nhãn của một điểm dựa trên nhãn của các điểm gần nhất (k điểm lân cận).

Ưu điểm: Dễ dàng triển khai, hiệu quả trong các trường hợp dữ liệu có cấu trúc phức tạp.

Nhược điểm: Có thể bị ảnh hưởng bởi sự hiện diện của dữ liệu nhiễu, cần lựa chọn k phù hợp, tính toán có thể chậm với tập dữ liệu lớn.

3. Logistic Regression:

Nguyên tắc: Sử dụng hàm sigmoid để dự đoán xác suất thuộc về một lớp cụ thể.

Ưu điểm: Dễ hiểu, dễ giải thích, hiệu quả trong việc xử lý dữ liệu tuyến tính.

Nhược điểm: Có thể gặp khó khăn trong xử lý dữ liệu phi tuyến tính.

**Ý nghĩa mô hình:**

1. Hiểu rõ hành vi khách hàng:

Phân loại khách hàng: Xác định khách hàng tích cực và khách hàng ít hoạt động dựa trên hành vi mua hàng.

Nhận diện nhu cầu: Phân tích hành vi của từng nhóm khách hàng để hiểu rõ nhu cầu, sở thích và động lực mua hàng.

Xây dựng chiến lược marketing phù hợp: Tập trung vào những nhóm khách hàng tiềm năng, tạo ra các chương trình khuyến mãi, ưu đãi phù hợp với từng nhóm.

2. Tăng hiệu quả kinh doanh:

Tối ưu hóa chi phí marketing: Tập trung vào các nhóm khách hàng có khả năng mua hàng cao, hạn chế lãng phí vào những khách hàng ít tiềm năng.

Tăng doanh thu: Dự đoán hoạt động mua hàng giúp doanh nghiệp tập trung vào các chiến lược thúc đẩy mua hàng cho khách hàng tiềm năng.

Cải thiện tỷ lệ chuyển đổi: Xây dựng các chiến lược marketing và bán hàng phù hợp với từng nhóm khách hàng, giúp tăng tỷ lệ chuyển đổi từ khách hàng tiềm năng thành khách hàng thực tế.

3. Nâng cao trải nghiệm khách hàng:

Cá nhân hóa dịch vụ: Cung cấp các dịch vụ và thông tin phù hợp với nhu cầu của từng nhóm khách hàng.

Tăng sự hài lòng: Tạo ra những trải nghiệm mua hàng tích cực, giúp khách hàng cảm thấy được tôn trọng và quan tâm.

Thúc đẩy lòng trung thành: Xây dựng mối quan hệ lâu dài với khách hàng bằng cách cung cấp những giá trị phù hợp với nhu cầu của họ.

Ví dụ ứng dụng:

Cửa hàng thời trang trực tuyến: Phân loại khách hàng theo phong cách thời trang, sở thích mua hàng để đưa ra các khuyến mãi và sản phẩm phù hợp.

Nền tảng thương mại điện tử: Dự đoán nhu cầu mua hàng của khách hàng để giới thiệu sản phẩm phù hợp, cá nhân hóa trải nghiệm mua sắm.

Công ty dịch vụ du lịch: Phân loại khách hàng theo nhu cầu du lịch, giới thiệu các tour du lịch phù hợp với sở thích và ngân sách của họ.

# CHƯƠNG IV: KẾT LUẬN

Trong đồ án này, nhóm chúng em đã tiến hành phân tích tập dữ liệu về hành vi mua sắm của khách hàng và áp dụng nhiều phương pháp machine learning để hiểu sâu hơn về ngữ cảnh mua sắm và dự đoán hành vi mua sắm trong tương lai. Thông qua việc sử dụng các mô hình như Kmeans, KNN, SVC, Logistic Regression và khai phá luật kết hợp, chúng em đã có được cái nhìn tổng quan và chi tiết về hành vi mua sắm của khách hàng.

Kết quả thu được từ đồ án này đã cung cấp thông tin quý giá về các nhóm khách hàng, các yếu tố quyết định trong quá trình mua sắm, cũng như dự đoán khả năng mua sắm của khách hàng trong tương lai. Các mô hình machine learning đã cho thấy khả năng dự đoán đáng kể và có thể được áp dụng trong các chiến lược kinh doanh thực tế để tối ưu hóa trải nghiệm khách hàng và tăng hiệu quả kinh doanh.

Về ưu điểm của nhóm, nhóm có tinh thần làm việc tập thể cao: Mọi thành viên đều trách nhiệm và đóng góp ý kiến tích cực, Ứng dụng đa dạng mô hình: Sử dụng Kmeans, KNN, SVC, Logistic Regression, và khai phá luật kết hợp cho ra kết quả đáng khích lệ: Các mô hình cho thông tin quý giá và áp dụng thực tế cao.

Về nhược điểm: thời gian còn hạn chế: không đủ thời gian để tối ưu hóa tất cả mô hình, kỹ năng chuyên môn chưa đồng đều: một số thành viên còn thiếu kinh nghiệm,

dữ liệu chưa hoàn hảo: hạn chế về chất lượng và số lượng dữ liệu.

Nếu có đủ thời gian, nhóm chúng em sẽ quản lý thời gian hiệu quả hơn, nâng cao kỹ năng chuyên môn và chia sẻ kiến thức nội bộ, cải thiện chất lượng dữ liệu: Thu thập dữ liệu phong phú và áp dụng kỹ thuật tiền xử lý tốt hơn và tăng cường kiểm thử và tối ưu hóa mô hình.

# CHƯƠNG VI: TÀI LIỆU THAM KHẢO

* 1. Giáo trình môn Khai phá dữ liệu
  2. VISHAL MADLE, Data Mining-Behaviour Analysis, [https://www.kaggle.com/code/vishalmadle/data-mining-behaviour-analysis/notebook#Objective-of-Association-Rule-Mining,](https://www.kaggle.com/code/vishalmadle/data-mining-behaviour-analysis/notebook#Objective-of-Association-Rule-Mining, 25/5/2024)  25/5/2024
  3. AKASH PATEL, Customer Personality Analysis, <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>, 25/5/2024