1.1. DATASET OVERVIEW

Bộ dữ liệu là thông tin tổng hợp của các khách hàng trong khoảng 2 năm (2021-2023), bao gồm thông tin khách hàng, chi tiêu đối với các mặt hàng, hành vi mua hàng và giao dịch.

Dataset characteristics

Loại dữ liệu:
- 8 categorical fields
- 23 numeric fields

Nội dung dữ liệu:
- Thông tin cá nhân của Khách hàng
- Tương tác giữa Khách hàng và Doanh nghiệp (hình thức thanh toán, thời gian đăng ký, complain...)
- Lượng chi tiêu đối với 5 mặt hàng thức ăn và 1 mặt hàng trang sức
- Hành vi phản hồi với ưu đãi
- Nguồn traffic

Data quality

- Missing values: 13.06%

Trong đó, tất cả các hàng đều có ít nhất 1 missing value.

Dataset limitation

- Bộ dữ liệu có **nhiều missing values** và **outliers**.
- **Thời gian thu thập dữ liệu khá ngắn, ngắt quãng giữa năm**. Từ đó chưa thể đưa ra được xu hướng theo mùa.

1.2. DATA CLEANING

Xử lý Missing Values

Vấn đề	Giải pháp
Các ID bị lặp lại nhiều hàng nhưng bị khuyết giá trị ở các cột một cách bù trừ cho nhau (dạng duplicate đục lỗ), gây ra nhiều missing values.	Dùng phương pháp ffill và bfill cho từng nhóm ID bị lặp lại

Xử lý Duplicates

	Giải pháp
Phone và Phone_Number Hai cột này đều là số điện thoại, bị khuyết giá trị một cách bù cho nhau.	Chập hai cột lại, và bỏ đi 2 cột Phone và Phone_Number cũ
Registration_Time, Year_Register, Month_Register Cột Year_Register và Month_Register có nhiều giá trị khuyết, nhưng 2 cột này có giá trị được trích xuất từ cột Registration_Time	Bỏ đi 2 cột Year_Register và Month_Register vì đã có đầy đủ trong Registration_Time
Income Cột này có 24 giá trị khuyết	Fill giá trị khuyết bằng giá trị trung bình (mean) của Income
Payment_Method Có nhiều giá trị khuyết ở cột này (446), đây có thể là do thông tin bị thu thập thiếu	Đổi giá trị khuyết thành 'unknown' để dễ phân tích

1.2. DATA CLEANING

Xử lý Data Type

Cột	Giải pháp
Registration_Time	datetime
Academic_Level, Gender, Marital_Status, Phone	string
Year_Of_Birth, Recency, Num_Deals_Purchases, Num_Web_Purchases, Num_Catalog_Purchases, Num_Store_Purchases, Num_Web_Visits_Month, Promo_30, Promo_40, Promo_50, Promo_10, Promo_20, Complain, Total_Purchase, Children, Age	int

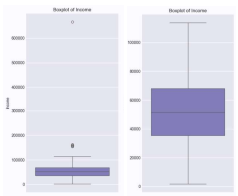
Xử lý Error Values

- So sánh xu hướng chấp nhận promo với nhau
- Thay giá trị '-1' ở Promo_40 thành '0'

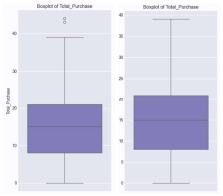
Xử lý Outliers

Bước 1: Kiểm tra outlier bằng IQR của 3 biến: Income, Total_Purchase, Total_Expense
Bước 2: Bỏ đi các row có outliers

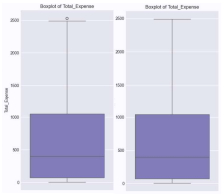
Income



Total_Purchase



Total_Expense



1.3. DATA TRANSFORMING

Biến đổi Columns

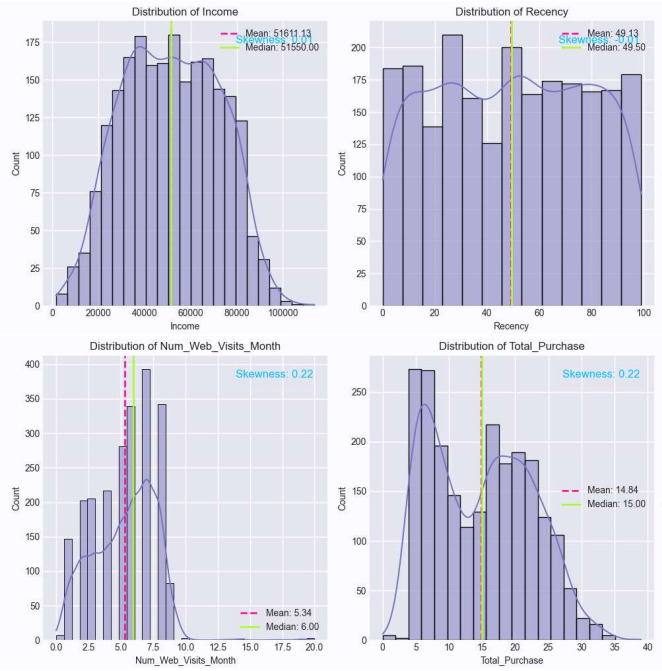
- Tạo thêm cột Age từ cột Year_Of_Birth
- Tách Living_With ra làm 2 cột Maritual_Status và Children
- Tách các biến trong cột Marital_Status ra làm 2 category: single và relationship
- Tạo cột Farmily_Size bằng tổng của Children và số cha/mẹ ứng với Maritual_Status
- Tạo cột Total_Expense bằng tổng các cột: 'Liquor', 'Vegetables', 'Pork', 'Seafood', 'Candy', 'Jewellery'
- Tạo cột Is_Parent với giá trị 0 nếu Children bằng 0, 1 nếu Children > 0
- Sửa các biến trong cột Academic_Level ra thành Under_Graduate, Graduate, Post_Graduate ứng với giá trị cũ (Post_Graduate cho PhD, Master, và 2n Cycle, Graduate cho Graduation, còn lại là Under_Graduate).
- Tạo cột 'Promote_Respone' là tổng số lần tham gia các promotion
- Tạo cột Customer_For bằng hiệu của Registration_Time mới nhất với từng Registration_Time

Tạo bảng rfm_df

- Lấy các cột Recency, Total_Purchase, và Total_Expense.
- Tính Recency_Score, Frequency_Score, Monetary_Score bằng cách tính phần trăm điểm của giá trị Recency, Total_Purchase, và Total_Expense ở từng hàng.
- Tính RFM bằng tổng của Recency_Score, Frequency_Score, Monetary_Score
- Tính RFM_Score bằng cách tính phần trăm điểm của giá trị RFM ở từng hàng
- Tính correlation giữa RFM_Score và Income.

2.1. TỔNG QUAN PHÂN PHỐI

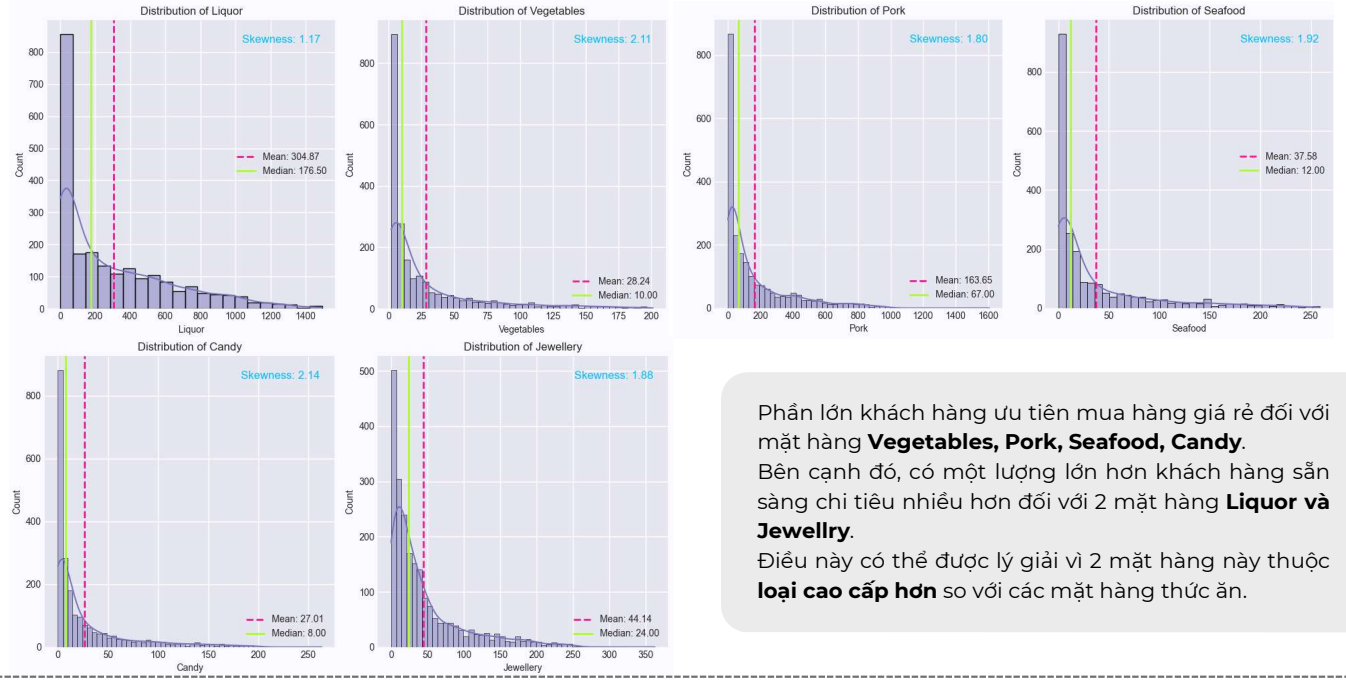
2.1.1. Phân phối gần chuẩn



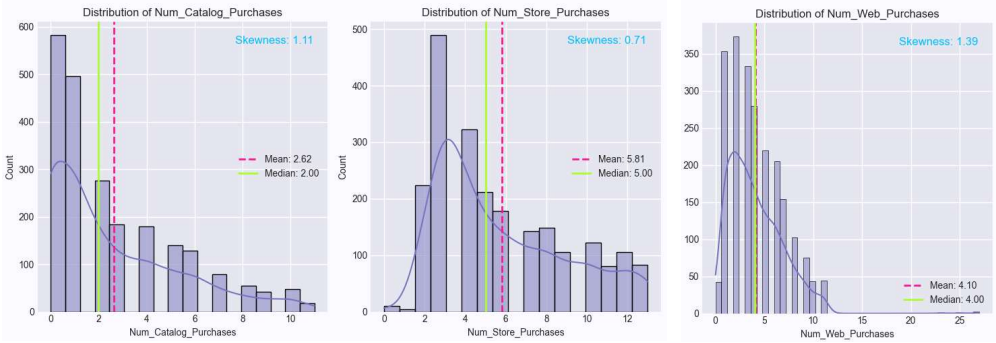
Hình 1. Các phân phối gần chuẩn

- **Trung bình thu nhập** của các khách hàng là **khoảng 51.6K USD**
- Mỗi khách **trung bình mua hàng 15 lần trong 2 năm**, **truy cập vào Website trung bình 5-6 lần/tháng** và **chưa quay lại mua hàng trong 49 ngày**

2.1.2. Phân phối lệch



Phần lớn khách hàng ưu tiên mua hàng giá rẻ đối với mặt hàng **Vegetables, Pork, Seafood, Candy**.
Bên cạnh đó, có một lượng lớn hơn khách hàng sẵn sàng chi tiêu nhiều hơn đối với 2 mặt hàng **Liquor và Jewellry**.
Điều này có thể được lý giải vì 2 mặt hàng này thuộc **loại cao cấp hơn** so với các mặt hàng thức ăn.



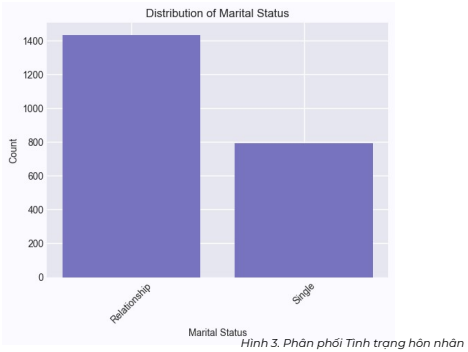
Hình 2. Các phân phối lệch

Lượt mua hàng từ cửa hàng (Store) cao hơn so với từ Catalog hay Website

2.2. PHÂN TÍCH ĐƠN BIẾN

2.2.1. Biến nhân khẩu học

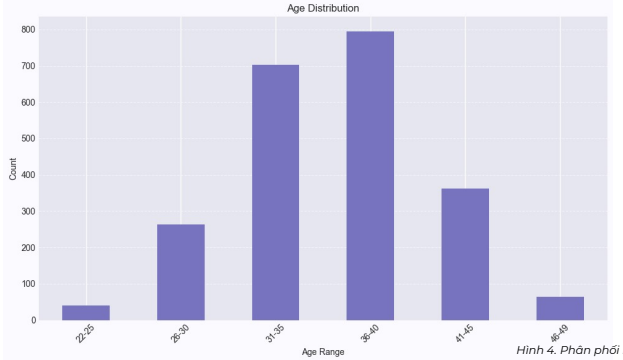
Tình trạng hôn nhân (Marital Status)



Hình 3. Phân phối Tình trạng hôn nhân

Nhóm khách hàng trong trạng thái có mối quan hệ (Relationship) chiếm tỉ lệ áp đảo.

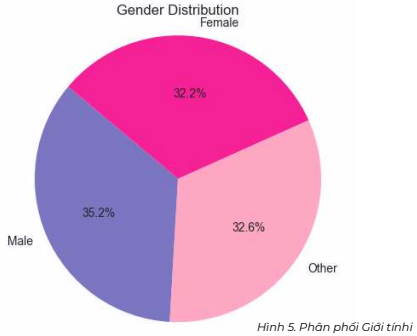
Tuổi tác (Age)



Hình 4. Phân phối Khoảng tuổi

Nhóm tuổi khách hàng mục tiêu chiếm ưu thế là từ 31-40 tuổi. Đây là nhóm tuổi quan tâm nhiều đến chi tiêu cho các mặt hàng ăn uống.

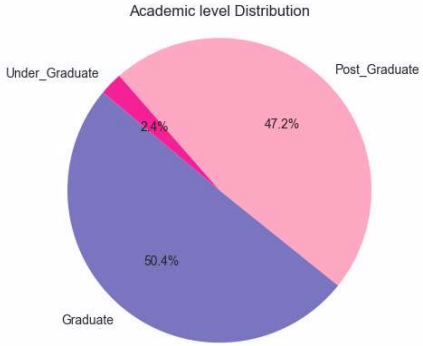
Giới tính (Gender)



Hình 5. Phân phối Giới tính

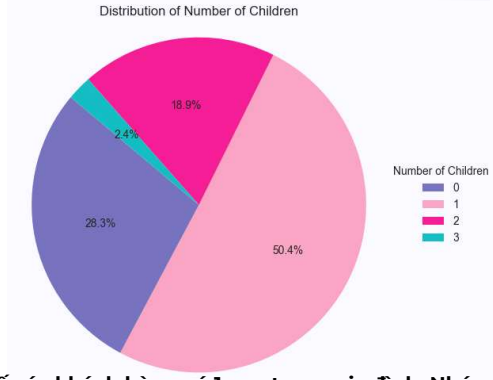
Tệp dữ liệu phân bổ khá đều giữa 3 loại giới tính Male, Female và Others.

Trình độ học vấn (Academic Level)



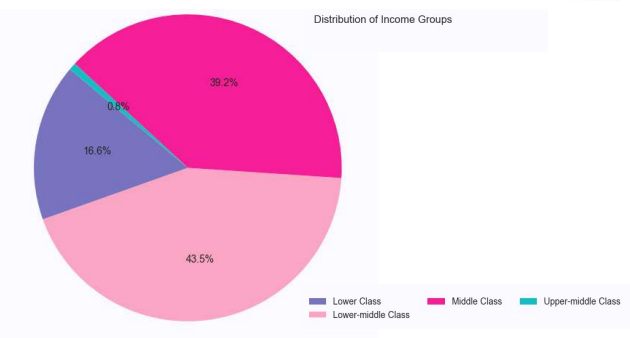
Phần lớn khách hàng đều có trình độ học vấn từ bậc Đại Học trở lên. Nhóm khách hàng với trình độ dưới bậc Đại Học chỉ chiếm tỉ lệ rất nhỏ.

Số lượng con (Children)



Đa số các khách hàng có 1 con trong gia đình. Nhóm không có con chiếm vị trí thứ 2.

Thu nhập (Income)

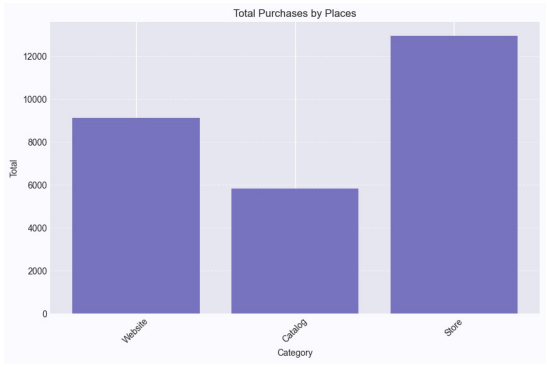


Nhóm thu nhập cận TB và TB chiếm tỉ lệ cao nhất. Điều này lý giải cho hành vi mua sắm hàng giá rẻ của khách hàng.

2.2. PHÂN TÍCH ĐƠN BIẾN

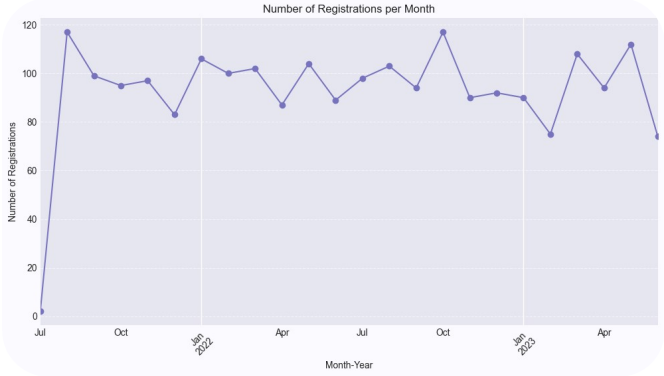
2.2.2. Biến hành vi mua sắm

Kênh mua sắm (Purchasing channel)



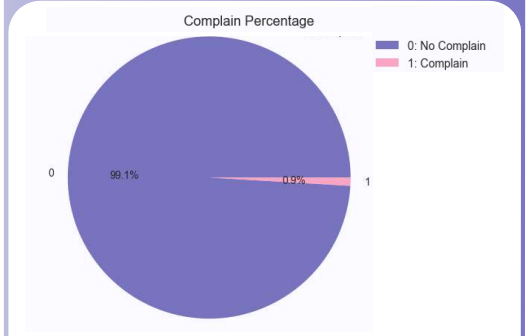
Khách hàng ưu tiên kênh mua sắm trực tiếp từ cửa hàng (Store), sau đó đến Website và Catalog.

Thời gian đăng ký (Registration Time)



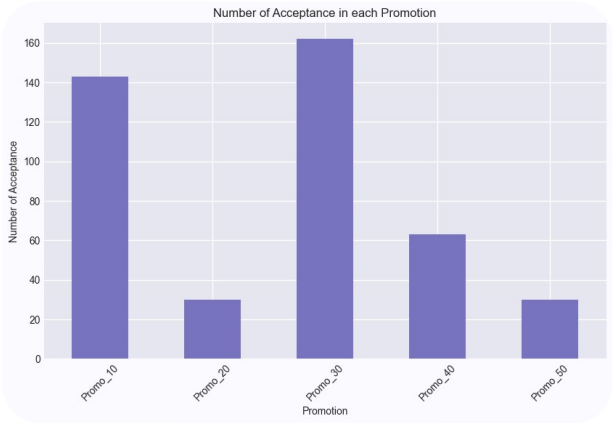
Lượt đăng ký theo tháng thường giảm vào giai đoạn cuối năm, ngoài ra chưa thể hiện xu hướng cao điểm cụ thể.

Khiếu nại (Complain)



Hầu hết (99.1%) khách hàng không khiếu nại đối với trải nghiệm mua sắm.

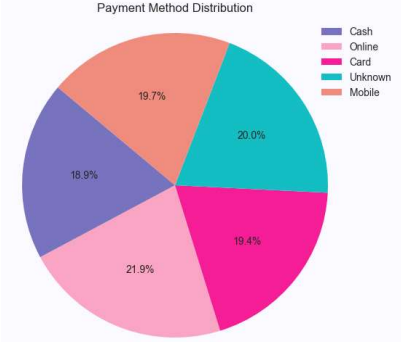
Phản hồi với ưu đãi (Promotion Acceptance)



Phân phối cho thấy **khách hàng tiếp nhận ưu đãi nhiều ở lần thứ nhất và thứ ba**, với giá trị ưu đãi 10% và 30%.

- Lần đầu: thu hút lượt tiếp nhận cao do hiệu ứng FOMO.
- Lần thứ 2: chính vì lượt tiếp nhận quá cao ở lần đầu tiên, khách hàng đã chi tiêu đủ và không còn nhu cầu. Bên cạnh đó, một vài khách hàng sẽ có xu hướng chờ đợi lần ưu đãi tiếp theo.
- Lần thứ 3: lượt tiếp nhận tăng đỉnh điểm do tâm lý "hời"
- Lần thứ 4 và 5: khách hàng đã chi tiêu đủ, mất đi hiệu ứng FOMO, tâm lý sợ hàng giá quá rẻ sẽ đi kèm với chất lượng kém. Chỉ còn lại lượng khách hàng ưu tiên giá siêu rẻ hoặc đã bỏ lỡ các lần ưu đãi trước đó.

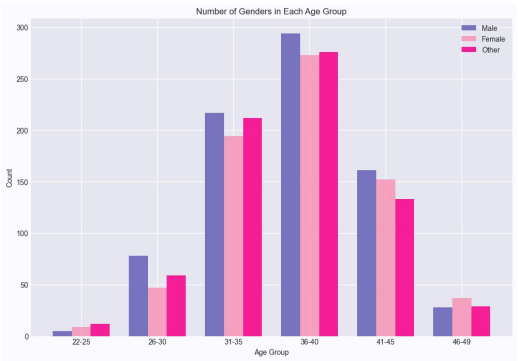
Phương thức thanh toán (Payment Method)



Các phương thức thanh toán chiếm tỉ lệ tương đương nhau. Phương thức thanh toán Online có tỉ lệ nhỉnh hơn nhẹ.

2.3. PHÂN TÍCH ĐA BIẾN

Age X Gender



Qua phân phối, ta thấy **đối tượng khách hàng mục tiêu chính** là **31-40 tuổi,** với Nam chiếm tỉ trọng cao nhất, sau đó đến Others và Nữ. **Nhóm khách hàng mục tiêu phụ** là từ **41-45 tuổi. Ở cả 2 nhóm này, đối tượng khách hàng chính đều là Nam.**

Đây là nhóm **tuổi trưởng thành**, quan tâm đến việc mua sắm đồ ăn thức uống.

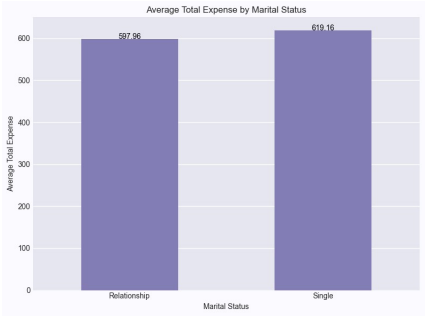
Purchase Channel X Total Expense X Purchase Quantity

Giữa các kênh mua sắm, **kênh cửa hàng (Store) vẫn mang lại kết quả khả quan nhất** cả về số lượng đơn hàng lẫn giá trị mua hàng. Điều này có thể được lý giải bởi đa số các mặt hàng là đồ ăn hoặc các mặt hàng có giá trị cao, vì vậy tâm lý khách hàng sẽ muốn **mua trực tiếp** để kiểm chứng chất lượng, độ tươi, độ giá trị v.v. của mặt hàng

Income X Total_Expense (Correlation)

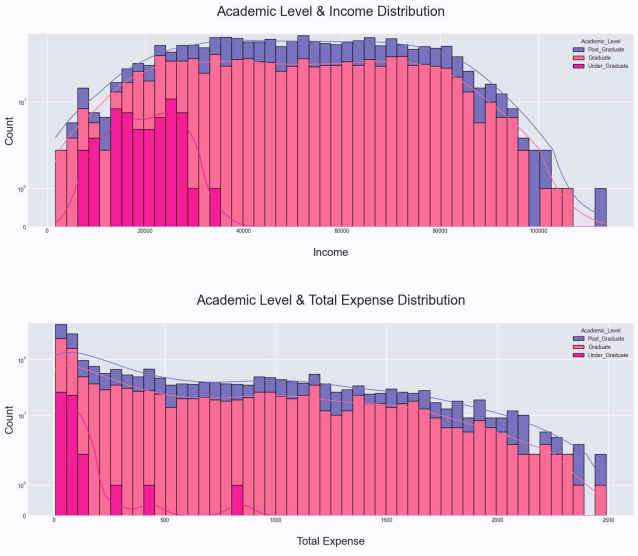
Giữa 2 biến Income và Total_Payment có hệ số tương quan khá cao (0.8). Điều này có nghĩa **khách hàng có thu nhập càng cao thì sẽ chi tiêu càng nhiều**.

Marital status X Average Expense



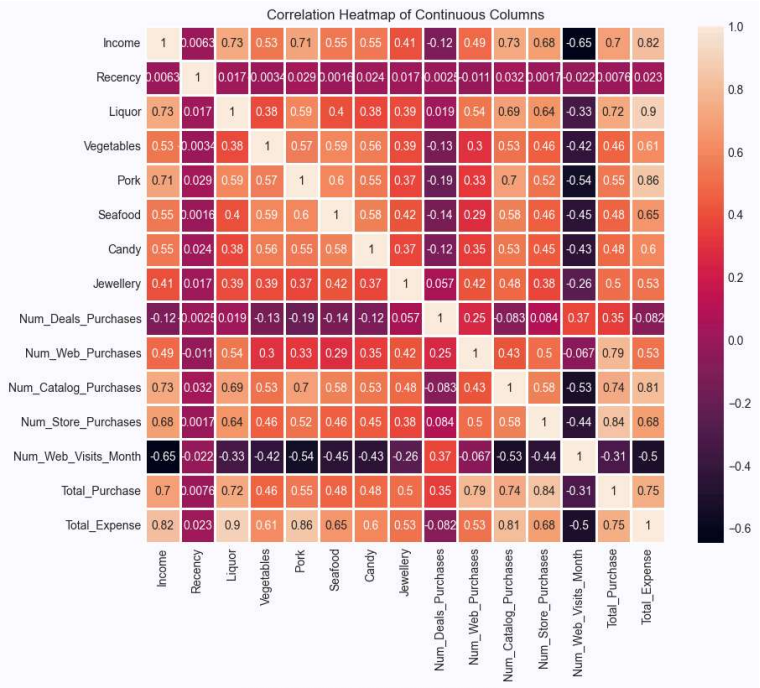
Dù số lượng khách hàng nhóm Relationship chiếm số lượng lớn hơn nhóm Single, **trung bình chi tiêu của đối tượng Single** lại **nhiều hơn** là **đối tượng Relationship**.

Academic Level X Income X Expense



Nhóm **Under_Graduate** rơi vào nhóm đối tượng **thu nhập thấp**, đối tượng **Graduate và Post_Graduate** thuộc nhóm **thu nhập cận TB**.

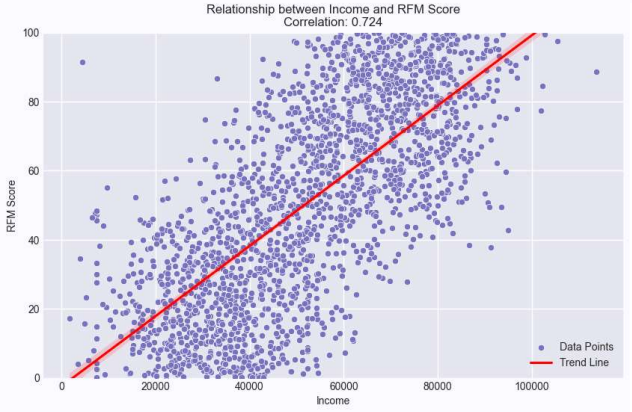
Tuy nhiên, các nhóm thuộc trình độ học vấn khác nhau đều có **khoảng chi tiêu không mấy khác biệt**, trừ nhóm **Under_graduate có lượng chi tiêu khá thấp**.



2.4. TƯƠNG QUAN GIỮA CÁC BIẾN

Các cặp biến có tương quan cao:

- **Total Expense và Income**: đã được giải thích ở mục trên.
- **Total Expense và Liquor/Vegetables/Pork/Seafood/Candy/Jewellry**: do Total Expense là tổng giá trị chi tiêu cho các mặt hàng Liquor, Vegetables, Pork, Seafood, Candy, Jewellry. Tương tự đối với tương quan giữa **Total Purchase và Liquor/Vegetables/Pork/Seafood/Candy/Jewellry**.
- **Total Purchase và Num_Web_Purchase/Num_Catalog_Purchase/Num_Store_Purchase**: do Total Purchase là tổng lượt mua hàng từ các các kênh Website, Catalog và Store.
- **Income và Liquor, Pork**: Liquor thuộc mặt hàng xa xỉ hơn, do đó độ co giãn (elasticity) khá lớn, dễ tăng tiêu thụ khi thu nhập (Income) tăng. Pork tuy là mặt hàng thiếu yếu nhưng lượng tiêu thụ lớn hơn so với Seafood hay Vegetables.



2.5. PHÂN TÍCH RFM

Từ phần phân tích đơn biến và đa biến, ta nhận thấy sự tương quan nổi bật giữa thu nhập (Income) đến hành vi mua hàng. Do đó, nhóm đề nghị **kiểm chứng chỉ số RFM** để đưa ra kết luận rõ ràng hơn.

Sau khi tính chỉ số RFM và kiểm tra tương quan giữa Income và chỉ số RFM, ta nhận thấy rằng, khách hàng có thu nhập càng cao thì chỉ số RFM càng cao, với chỉ số tương quan là 0.724. Điều này có nghĩa là **thu nhập càng cao** thì **họ mua hàng càng thường xuyên, chi tiêu càng nhiều, và thường xuyên mua hàng lặp lại**.

3. Feature Selection & Feature Engineering

Biến đổi Cột

Cột	Ý nghĩa
Age	Năm hiện tại - Year_Of_Birth
Children	Suy ra từ cột Living_With, gồm giá trị: 0,1,2,3
Marital_Status	Suy ra từ cột Living_With, gồm 2 giá trị: single, relationship
Farmily_Size	Tổng của Children và số cha/mẹ ứng với Maritual_Status
Total_Expense	Tổng các cột: 'Liquor', 'Vegetables', 'Pork', 'Seafood', 'Candy', 'Jewellery'
Is_Parent	Với giá trị 0 nếu Children bằng 0, 1 nếu Children > 0
Academic_Level	Sửa các biến thành Under_Graduate, Graduate, Post_Graduate ứng với giá trị cũ (Post_Graduate cho PhD, Master, và 2n Cycle, Graduate cho Graduation, còn lại là Under_Graduate)
Promote_Respone	Tổng số lần tham gia các promotion
Customer_For	Hiệu của Registration_Time mới nhất với từng Registration_Time

Kỹ thuật

- **Chuẩn hóa dữ liệu (StandardScaler)**

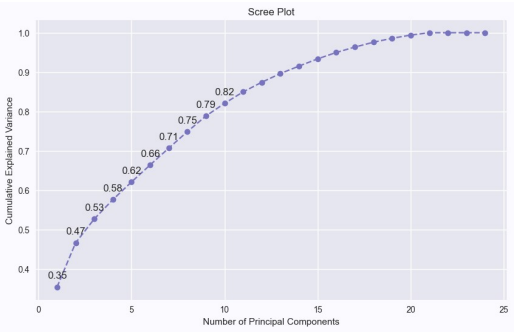
Là một công cụ chuẩn hóa các đặc tính của dữ liệu bằng cách loại bỏ giá trị trung bình và chia cho độ lệch chuẩn để đảm bảo rằng mỗi đặc tính có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1.

- **Phân tích thành phần chính (Probably approximately correct learning)**

Là một kỹ thuật giảm chiều dữ liệu được sử dụng trong lĩnh vực học máy và khai phá dữ liệu

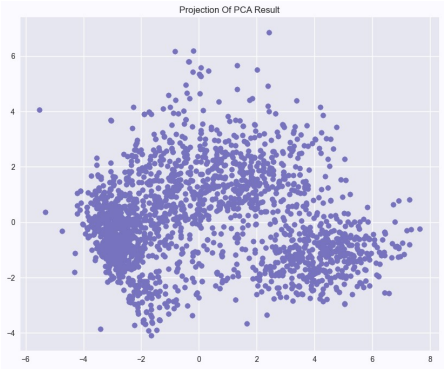
  - *Giảm số chiều dữ liệu đầu vào* trong khi vẫn *giữ lại phần lớn thông tin quan trọng.*
  - *Giúp loại bỏ những biến không quan trọng hoặc có độ tương quan cao* với các biến khác.

Xác định chiều dữ liệu



Phân tích phương sai được giải thích, ta thấy điểm đứt gãy là 2.

Kết quả PCA



Ta thấy rằng dữ liệu phân phối thành các **đám mây dày đặc nằm san sát nhau**. Ngoài ra, ta còn thấy **một số điểm ngoại lai nằm rải rác quanh đám mây**.

4.1. MODEL PREPARATION

Đề xuất mô hình

Các mô hình được đề xuất gồm Connectivity models (Agglomerative, Hierarchical,...), Centroid models (K-Means,...), Density models (DBSCAN,...), và Distribution models (GMM,...).

Mô hình GMM có thể không phù hợp vì các giả định về đặc điểm cụm, hình dạng cụm và mật độ cụm không phù hợp với tập dữ liệu đã cho. Dựa vào đồ thị 2 chiều của kết quả PCA, ta thấy rằng các điểm phân bố san sát nhau với mật độ cao từ đó khiến cho mô hình DBSCAN không hoạt động hiệu quả, tuy nhiên nhóm tác giả sẽ lợi dụng đặc điểm này để lọc bớt các outliers không phù hợp với "đám mây" chính.

Model	Cluster Characteristics	Cluster Shape	Cluster Density	Number of Clusters
K-Means	Spherical, similar size	Spherical	Uniform	Predefined
DBSCAN	Any, varying density	Any	Non-uniform	Doesn't need to be known
GMM	Gaussian distribution, elliptical	Elliptical	Uniform	Predefined
Agglomerative	Hierarchical, undefined	Any	Any	Doesn't need to be known

Chỉ số đánh giá

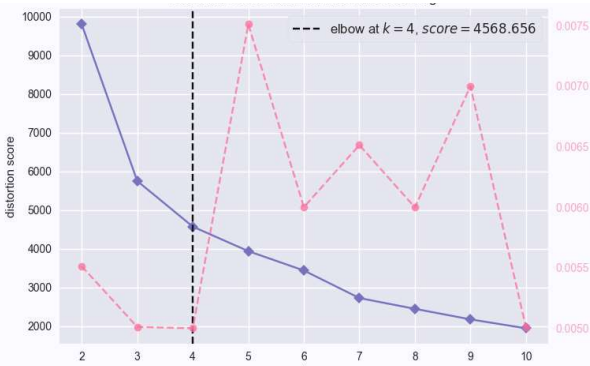
- **Inertia** (Within-cluster Sum of Squares - WCSS): Đo lường tổng khoảng cách bình phương từ mỗi điểm đến trung tâm cụm của nó.
- **Silhouette Score**: Đo lường sự tương đồng của các điểm trong cùng một cụm so với các điểm trong các cụm khác.
- **Davies-Bouldin Index**: Đánh giá sự tương đồng giữa các cụm. Giá trị nhỏ hơn cho thấy phân cụm tốt, với các cụm có khoảng cách lớn giữa nhau và các điểm trong cùng một cụm gần nhau.
- **Calinski-Harabasz Index (Variance Ratio Criterion)**: Đo lường tỷ lệ phân tán giữa các cụm và phân tán trong cụm.

4.2. OUTLIERS HANDLING

Tính toán khoảng cách k tối ưu với từng giá trị min_samples để làm tham số eps cho mô hình DBSCAN, phục vụ lọc các điểm không phù hợp.

4.3. OPTIMAL CLUSTERS

Sử dụng phương pháp Elbow để xác định số cluster tối ưu



4.4. MODEL COMPARISION

Model	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.455638	0.855567	3191.82
Agglomerative Clustering	0.526953	0.77084	3499.92

- **Silhouette Score** của mô hình AC gần 1 hơn, chứng tỏ các điểm được phân đúng cụm hơn, và khoảng cách giữa các cụm lớn.
- **Davies - Bouldin Index** của mô hình AC nhỏ hơn cho thấy phân cụm tốt, các cụm có khoảng cách lớn giữa nhau và các điểm trong cùng một cụm gần nhau hơn.
- **Calinski-Harabasz Index** của mô hình AC lớn hơn cho thấy phân cụm tốt, có sự phân tán lớn giữa các cụm và sự phân tán nhỏ trong các cụm.

Nhóm quyết định sử dụng mô hình Agglomerative để làm thuật toán phân cụm khách hàng chính.

4.5. HYPERPARAMETER TUNNING

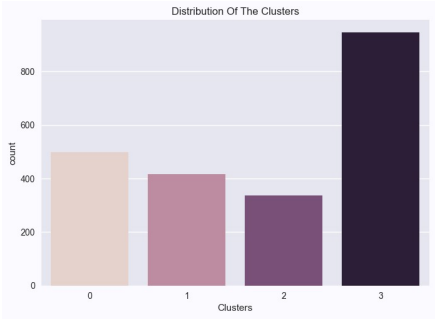
Thực hiện grid search cho mô hình với từng giá trị của các parameter metric, linkage, distance_threshold và compute_distances như sau:

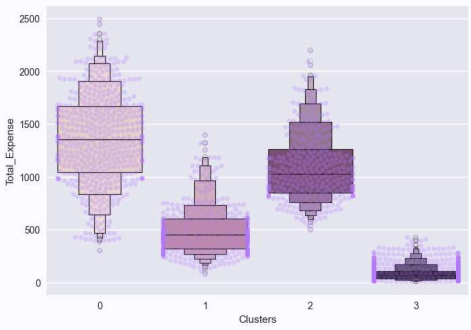
- **metric**: 'euclidean', 'manhattan', 'cosine'
- **linkage**: 'ward', 'complete', 'average', 'single' ('ward' chỉ hoạt động với metric 'euclidean')
- **distance_threshold**: None, 175, 180, 185 (n_cluster là None khi distance_threshold khác None)
- **compute_distances**: True, False

```
Best Silhouette Score: 0.6009046862893253
Best Davies-Bouldin Index: 0.6415143559374005
Best Calinski-Harabasz Index: 4434.399986168784
Best Parameters: {'metric': 'euclidean', 'linkage': 'ward', 'distance_threshold': None, 'compute_distances': True}
```

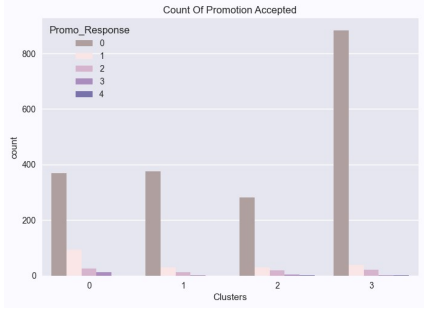
Sau khi tìm được các tham số tối ưu cho mô hình. thực hiện vẽ kết quả lên mặt phẳng hai chiều để quan sát các cluster được đánh dấu

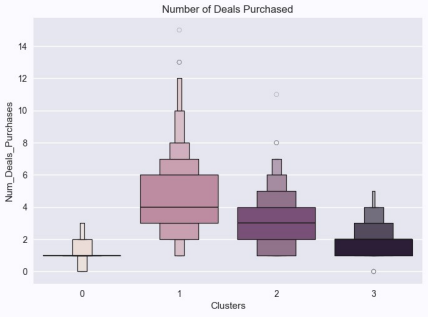
5. CUSTOMER PROFILING



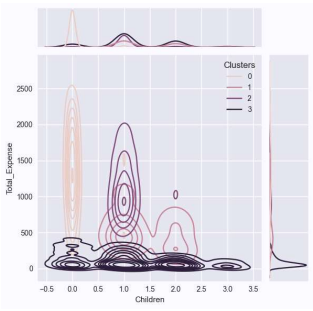


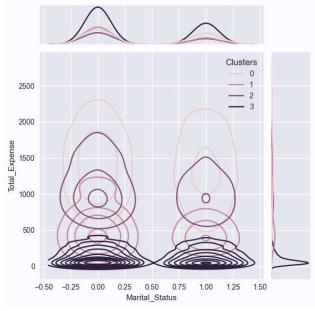
Tiêu chí	Nhóm 0	Nhóm 1	Nhóm 2	Nhóm 3
Tỉ trọng	Thứ 2	Thứ 3	Ít nhất	Lớn nhất
Thu nhập	Trung bình trở lên (đa số <$50K)	Cận trung bình - trung bình ($30K-80K)	Cận trung bình - trung bình ($40K - $90K)	Thấp - cận trung bình (<$62K)
Chi tiêu	Trung bình - cao (đa số $700-$2K) linh hoạt	Ít - trung bình ($200-$1.1K)	Trung bình - cao (đa số $700-$1.5K)	Ít (đa số <$250)
Phản hồi với ưu đãi	Không mua nhiều ưu đãi	Quan tâm đến ưu đãi nhất trong 4 nhóm	Có quan tâm đến ưu đãi	Quan tâm đến ưu đãi nhưng không đáng kể

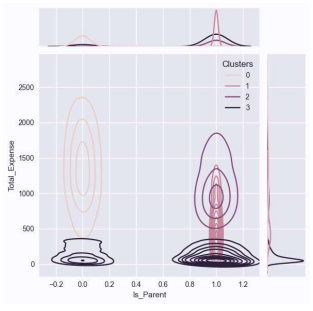


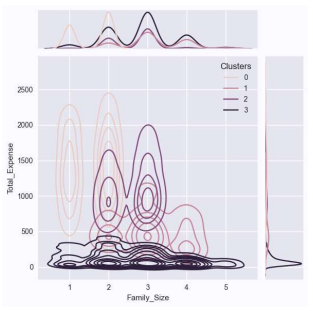


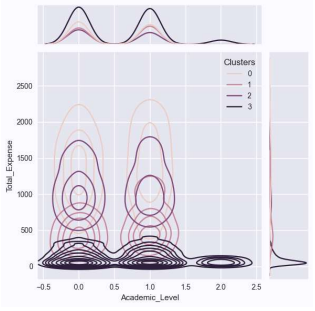
5. CUSTOMER PROFILING

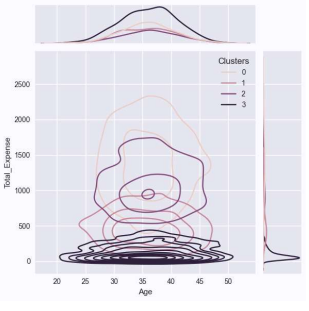












Tiêu chí	Nhóm 0	Nhóm 1	Nhóm 2	Nhóm 3
Tuổi	27-47	24-47	26-46	20-51
Là cha mẹ	Chắc chắn không	Chắn chắn	Chắc chắn	Phần lớn
Trình trạng hôn nhân	Hơn một nửa là độc thân	Phần lớn là độc thân	Phần lớn là độc thân	Hơn một nửa là độc thân
Số con cái	0	Đa số là 1 con, một số có 2 con	Đa số là 1 con, số ít có 2 con	Đa số có 1 - 2 con, rất ít có 3 con.
Family size	Lớn nhất là 2, nhỏ nhất là 1	Lớn nhất là 4, nhỏ nhất là 2	Lớn nhất là 3, nhỏ nhất là 2	Lớn nhất là 5, nhỏ nhất là 1
Trình độ học vấn	Phần lớn là Under-graduate và Graduate	Phần lớn là Under-graduate và Graduate	Phần lớn là Under-graduate và Graduate	Đa số là Under-graduate và Graduate, số ít là Post-graudate

5. CUSTOMER PROFILING

NHÓM 0

KHÁCH HÀNG TRUNG THÀNH

"Tôi có thể độc thân hoặc đang trong một mối quan hệ. Tôi không có con.

Tôi có thu nhập tốt nên không quá quan tâm đến các ưu đãi. Tôi chi tiêu khá linh hoạt, thường xuyên chi tiêu cao."

NHÓM 1

KHÁCH HÀNG SĂN DEAL

Tôi độc thân và là cha/mẹ. Gia đình tôi có khá nhiều thành viên.

Tuy vậy, tôi có thu nhập và chi tiêu không quá nổi bật. Do đó, tôi chỉ quan tâm nhiều đến các chương trình ưu đãi.

NHÓM 2

KHÁCH HÀNG TIỀM NĂNG

Tôi độc thân và là cha/mẹ. Gia đình tôi có khá nhiều thành viên.

Tôi có thu nhập tốt và quan tâm nhiều đến các chương trình ưu đãi. Tôi có thể chi tiêu cao nhưng thường chỉ chi tiêu trung bình.

NHÓM 3

KHÁCH HÀNG BÌNH DÂN

Tôi có thể độc thân hoặc đang trong một mối quan hệ. Tôi có từ 1-2 con.

Tôi có thu nhập thấp, chi tiêu thấp. Tôi chỉ mua sắm đủ nhu cầu và không quan tâm nhiều đến ưu đãi.

6. EXTENSION

6.1. Suggesting variables

Để tăng tính đa dạng và chính xác của dữ liệu, có thể bổ sung các trường dữ liệu sau:

Thông tin địa lý

Thêm thông tin về vị trí địa lý của khách hàng như địa chỉ, thành phố, quốc gia, vĩ độ, kinh độ, v.v. Điều này có thể giúp phân tích mối quan hệ vùng miền và ảnh hưởng của vị trí đến hành vi mua sắm.

Dữ liệu giao dịch

Bổ sung thông tin về giao dịch cụ thể của khách hàng như thời gian giao dịch. Thông tin này sẽ cung cấp cái nhìn chi tiết hơn về hành vi mua sắm của khách hàng theo thời gian trong ngày, trong tuần.

Các bài toán có thể mở rộng bao gồm:

Phân tích hành vi mua sắm

Sử dụng dữ liệu để phát hiện xu hướng mua hàng, giỏ hàng và dự đoán hành vi mua sắm tương lai của khách hàng.

Dự đoán doanh số bán hàng

Dự đoán doanh số bán hàng tương lai dựa trên các yếu tố như mùa, xu hướng mua sắm và chiến lược tiếp thị.

Phân loại khách hàng

Sử dụng các thuật toán phân cụm để phân loại khách hàng thành các nhóm dựa trên các đặc điểm chung hoặc hành vi mua sắm.

Tối ưu hóa chiến lược tiếp thị

Đánh giá hiệu quả của các chiến lược tiếp thị hiện tại và tối ưu hóa việc phân phối nguồn lực tiếp thị dựa trên dữ liệu về Promo, giỏ hàng của các loại khách hàng.

6. EXTENSION

6.2. Market Basket Analysis

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
Candy	Jewellery	0.812388	0.972621	0.797127	0.981215	1.008836
Candy	Vegetables	0.812388	1.000000	0.812388	1.000000	1.000000
Candy	Seafood	0.812388	0.827648	0.732047	0.901105	1.088754
Candy	Liquor	0.812388	1.000000	0.812388	1.000000	1.000000
Jewellery	Vegetables	0.972621	1.000000	0.972621	1.000000	1.000000
Jewellery	Seafood	0.972621	0.827648	0.810144	0.832949	1.006404
Jewellery	Candy	0.972621	0.812388	0.797127	0.819566	1.008836
Jewellery	Liquor	0.972621	1.000000	0.972621	1.000000	1.000000
Liquor	Jewellery	1.000000	0.972621	0.972621	0.972621	1.000000
Liquor	Seafood	1.000000	0.827648	0.827648	0.827648	1.000000
Liquor	Pork	1.000000	0.999551	0.999551	0.999551	1.000000
Liquor	Vegetables	1.000000	1.000000	1.000000	1.000000	1.000000
Liquor	Candy	1.000000	0.812388	0.812388	0.812388	1.000000
Pork	Vegetables	0.999551	1.000000	0.999551	1.000000	1.000000
Pork	Liquor	0.999551	1.000000	0.999551	1.000000	1.000000
Seafood	Liquor	0.827648	1.000000	0.827648	1.000000	1.000000
Seafood	Vegetables	0.827648	1.000000	0.827648	1.000000	1.000000
Seafood	Jewellery	0.827648	0.972621	0.810144	0.978850	1.006404
Seafood	Candy	0.827648	0.812388	0.732047	0.884490	1.088754
Vegetables	Liquor	1.000000	1.000000	1.000000	1.000000	1.000000
Vegetables	Candy	1.000000	0.812388	0.812388	0.812388	1.000000
Vegetables	Seafood	1.000000	0.827648	0.827648	0.827648	1.000000
Vegetables	Jewellery	1.000000	0.972621	0.972621	0.972621	1.000000
Vegetables	Pork	1.000000	0.999551	0.999551	0.999551	1.000000

Antecedents: Các mặt hàng xuất hiện trước.
Consequents: Các mặt hàng xuất hiện sau.
Antecedent Support: Tỷ lệ giao dịch chứa mặt hàng xuất hiện trước.
Consequent Support: Tỷ lệ giao dịch chứa mặt hàng xuất hiện sau.
Support: Tỷ lệ giao dịch chứa cả mặt hàng.
Confidence: Xác suất tìm thấy mặt hàng xuất hiện sau trong một giao dịch khi mặt hàng xuất hiện trước đã xuất hiện.
Lift: Tỷ lệ giữa hỗ trợ quan sát được và hỗ trợ mong đợi nếu mặt hàng xuất hiện trước và sau độc lập với nhau.

Phân tích giỏ hàng là một kỹ thuật khai thác dữ liệu phân tích các mẫu đồng xuất hiện và xác định độ mạnh của mối liên kết giữa các sản phẩm được mua cùng nhau.

Khi mọi người mua hàng A, rõ ràng rằng họ cũng có thể mua hàng B cùng với nó. Mối quan hệ này được mô tả dưới dạng thuật toán điều kiện, như dưới đây.

NẾU {A} THÌ {B}

Thuật toán Apriori là một thuật toán được sử dụng phổ biến trong khai phá dữ liệu (data mining) để xác định các tập mục phổ biến và tạo ra các luật kết hợp. Ta sử dụng thuật toán này để thực hiện Market Basket Analysis.

Với tỷ lệ giao dịch có chứa có cả 2 mặt hàng (support) lớn hơn 0.5, ta có các cặp mặt hàng như sau:

antecedents	consequents
Candy	[Liquor, Vegetables, Seafood, Jewellery]
Jewellery	[Liquor, Vegetables, Seafood, Candy]
Liquor	[Vegetables, Pork, Seafood, Candy, Jewellery]
Pork	[Liquor, Vegetables]
Seafood	[Liquor, Vegetables, Candy, Jewellery]
Vegetables	[Liquor, Pork, Seafood, Candy, Jewellery]

- Rau củ hoặc đồ uống có cồn luôn mặt trong mọi giỏi hàng.
- Khi mua rau củ hoặc đồ uống có cồn trước thì xác suất mua thêm các mặt hàng khác là rất cao.
- Sau khi mua thịt heo thì chỉ có xác suất cao mua 2 mặt hàng là đồ uống có cồn và rau.

Lợi ích:

- Phân tích hành vi khách hàng
- Áp dụng để đưa ra khuyến mãi và chiến dịch
- Tối ưu hóa hàn tồn kho
- Gợi ý cá nhân hóa

Thank You

Huynh Cong Danh
FTU2

Nguyen Dang Khoa
FTU2

Phan Vo Diem Trang
FTU2