



DAZONE 2024
BÁO CÁO
VÒNG 2.2

SPARKLE



LÊ TUẤN ANH



NGUYỄN TUYẾT ANH



DƯƠNG THANH DŨNG

NỘI DUNG

01

TÓM TẮT
BÁO CÁO

02

XỬ LÝ CƠ BẢN

03

PHÂN CỤM
KHÁCH HÀNG

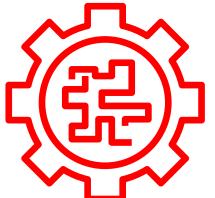
04

KHẮC HỌA
CHÂN DUNG



Situation

Công ty kinh doanh bao gồm các mặt hàng: rượu, thịt lợn, cá, rau củ, đồ ngọt và trang sức. Công ty cung cấp các ưu đãi cho khách hàng.



Complication

Doanh thu công ty giảm do chi tiêu mỗi khách hàng giảm mạnh.



Question

Làm thế nào để tối ưu hóa chiến lược tiếp thị và dịch vụ, cá nhân hóa và nâng cao trải nghiệm của khách hàng nhằm tăng chi tiêu mỗi khách hàng?



Answer

Phân cụm khách hàng để có cái nhìn tổng quan về nhóm khách hàng có đặc điểm tương tự nhau.

Phân 5 nhóm khách hàng

- Recent Big Customers
- At Risk customers
- Loyal Customers
- Promise Customers
- Lost Customers

Biến phân loại

ID	Mã số duy nhất của khách hàng
Year_Of_Birth	Năm sinh của khách hàng
Gender	Giới tính của khách hàng
Phone_Number/ Phone	Số điện thoại của khách hàng
Payment_Method	Hình thức thanh toán nhiều nhất
Academic_Level	Trình độ học vấn của khách hàng
Living_With	Tình trạng hôn nhân và số con trong nhà
Income	Thu nhập hàng năm của khách hàng
Recency	Số ngày kể từ lần mua hàng cuối cùng của khách hàng
Complain	1 nếu khách hàng phàn nàn trong vòng 2 năm qua; 0 nếu không.

Biến sản phẩm

Liquor, Vegetables, Pork, Seafood, Candy, Jewellery thể hiện số tiền đã chi tiêu trong 2 năm qua cho từng sản phẩm.

Khuyến mãi

Num_Deals_Purchases	Số lượng giao dịch mua hàng với giảm giá
Promo_10/20/30/40/50	1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ nhất/ thứ hai/thứ ba/thứ tư/ thứ năm. 0 nếu không

Kênh mua hàng

Num_Web_Purchases	Số giao dịch mua hàng qua Website.
Num_Catalog_Purchases	Số giao dịch mua sử dụng danh mục trên Web.
Num_Store_Purchases	Số giao dịch mua hàng trực tiếp tại cửa hàng.
Num_Web_Visits_Month	Số lần truy cập vào trang Web trong tháng qua.

01 Xử lý dữ liệu trùng lặp

ID	Num_Deals_Purchase	Num_Web_Purchases	Num_Catalog_Purchase	Num_Store_Purchase	Num_...
1128	4	3	1	5	
1128	4			5	
1128	4		1	5	

- Xác định dữ liệu trùng lặp.
- Loại bỏ và giữ lại dữ liệu có đầy đủ các giá trị hơn.

02 Xử lý dữ liệu không tương thích

ID	Num_Deals_Purchase	Num_Web_Purchases	Num_Catalog_Purchase	Num_Store_Purchase	Num_...
1128	4	3	1	5	
1128	4			5	
1128	4		1	5	

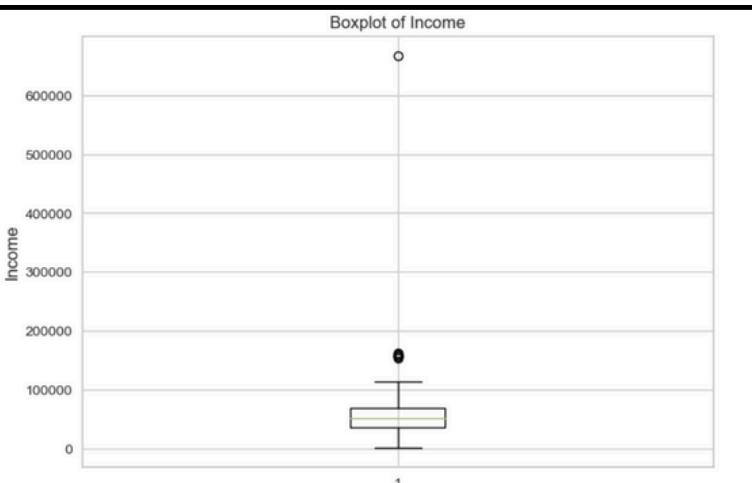
- Year_Register, Month_Register phụ thuộc vào Registration_Time.
- Year_Register/Month_Register có khác biệt giá trị so với Registration_Time.

04 Thêm bớt và chỉnh sửa các cột

- Xóa cột Phone Number do có cùng giá trị với cột Phone.
- Chia cột Living_With thành Living Status và Number_Of_Child.

05 Kiểm tra Outliers

ID	Promo_30	Promo_40
9037	0	-1
3161	0	-1
9153	0	-1
4545	1	-1
5865	0	-1
5870	0	-1
8922	0	-1
2804	0	-1
1431	0	-1



Hình 1. Phân tích Income Outlier.

- Thay thế giá trị Promo_40<0 bằng 0 (vì promo chỉ có giá trị phù hợp 0 và 1).
- Loại bỏ giá trị Income = 666666, cao bất thường.

03 Xử lý dữ liệu trống

1355	Together_Mobile
1362	Divorced_0
1365	Married_3 Online
1373	Married_1
1376	Married_1 Cash
1382	Single_1 Online
1386	Together_2
1401	Single_2 Card
1406	Together_1
1411	Together_Online
1413	Divorced_2

Trước

Sau

1355	Together_Mobile
1362	Divorced_0
1365	Married_3 Online
1373	Married_1
1376	Married_1 Cash
1382	Single_1 Online
1386	Together_2
1401	Single_2 Card
1406	Together_Online
1411	Together_Online
1413	Divorced_2

Hình 2. Trước và sau khi xử lý dữ liệu trống.

- Xử lý các dữ liệu trống của trường Income và trường Payment_Method bằng **phương pháp MICE**:

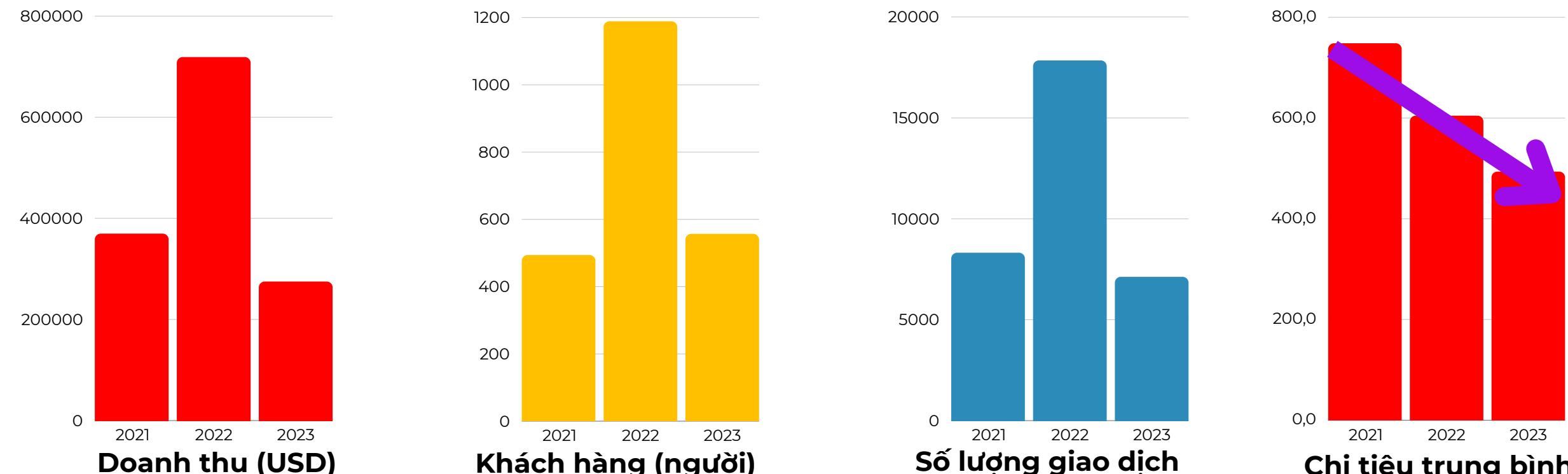
***Phương pháp MICE:** Dựa trên các thông số tương đồng của các khách hàng khác nhau, lựa chọn giá trị Income và Payment_Method phổ biến nhất để điền vào giá trị bị trống bên cạnh việc sử dụng Mode và giá trị trung bình.

Dữ liệu đã sẵn sàng

Thống kê tổng quan trong vòng 3 năm

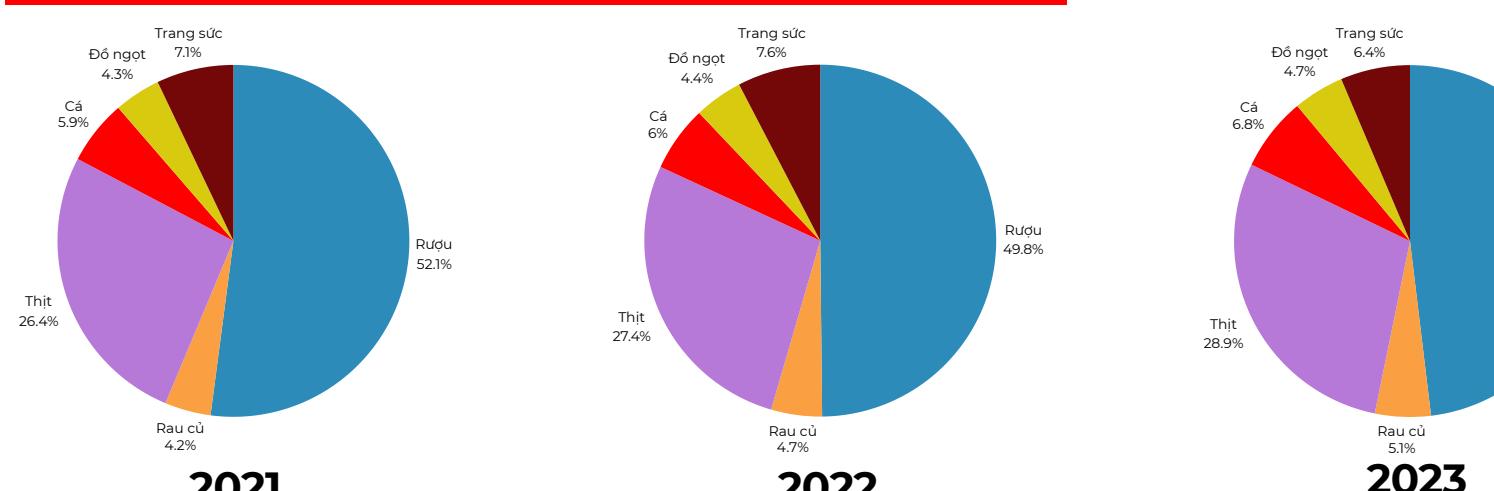
-  **1,36M** USD doanh thu
-  **2240** khách hàng
-  **608.8** USD / khách hàng
-  **06** mặt hàng
-  **33,291** giao dịch
-  **44** số lượng đơn hàng lớn nhất /khách hàng
-  **> 50%** doanh thu

Hiệu suất doanh nghiệp



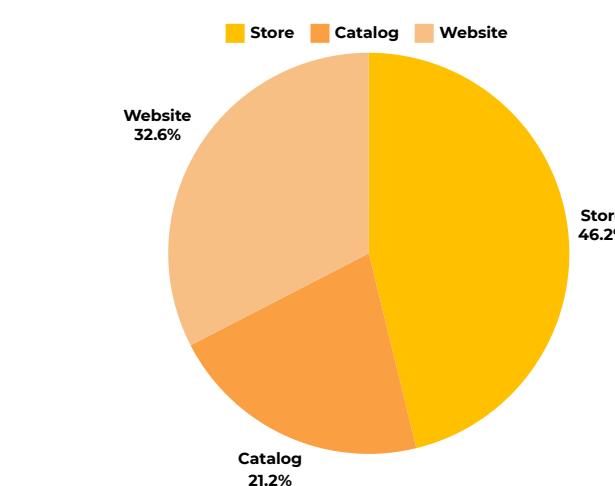
Doanh thu và số lượng giao dịch biến động mạnh và giảm, trong khi lượng khách hàng biến động mạnh và tăng nhẹ. Tuy nhiên, chi tiêu trên đầu người lại giảm mạnh qua từng năm.

Mức độ tiêu thụ sản phẩm qua 3 năm



Khách hàng có xu hướng tiêu thụ thịt nhiều hơn, trong khi mức độ tiêu thụ rượu giảm.

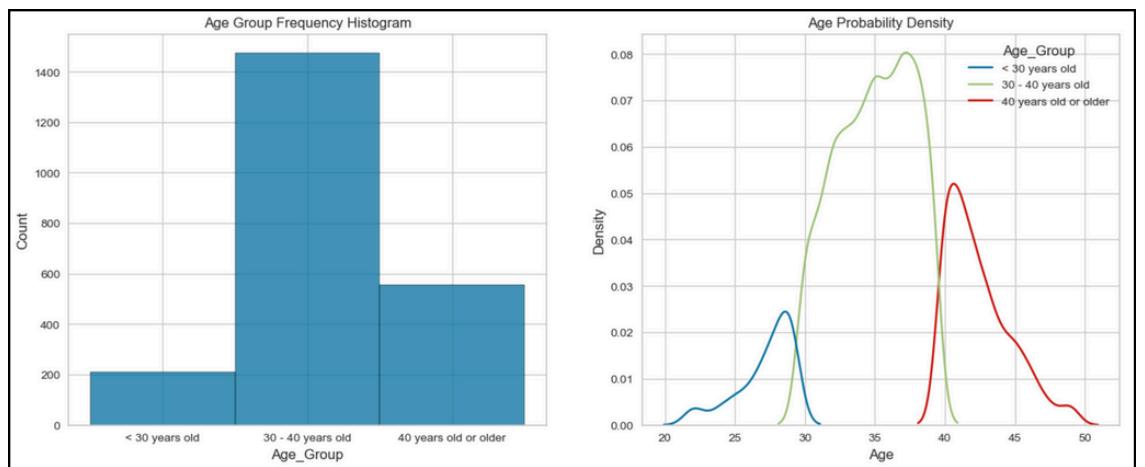
Tỷ lệ mua hàng qua nền tảng



Cửa hàng và Website là hai kênh bán hàng chính của doanh nghiệp

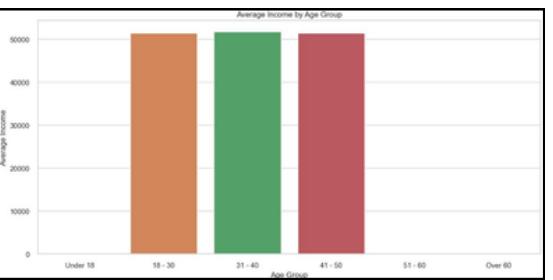
Khách hàng chính chủ yếu ở độ tuổi 30 - 40. Ở nhóm tuổi này, thu nhập càng cao, chi tiêu càng nhiều. Trong khi đó, học vấn và chi tiêu có quan hệ tỉ lệ nghịch.

Phân tích theo TUỔI

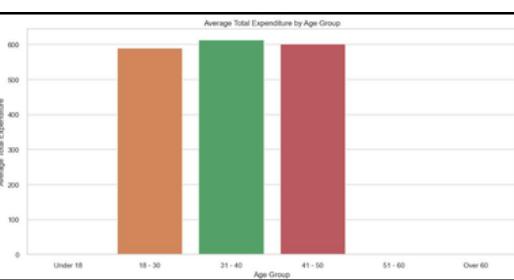


Hình 3. Biểu đồ phân phối tuổi khách hàng

- Nhóm khách hàng công ty có độ tuổi từ khoảng **20 - 50 tuổi**.
- Trong đó, nhóm khách hàng chủ yếu tập trung ở tuổi 30 - 40.

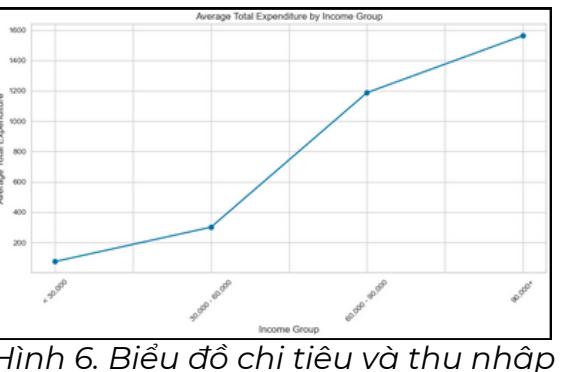


Hình 4. Biểu đồ thu nhập theo tuổi



Hình 5. Biểu đồ chi tiêu theo tuổi

- Các nhóm tuổi có **thu nhập và chi tiêu trung bình** tương đối **bằng nhau**.
- Trong đó **nhóm tuổi 30 - 40** có **thu nhập và chi tiêu cao nhất**.

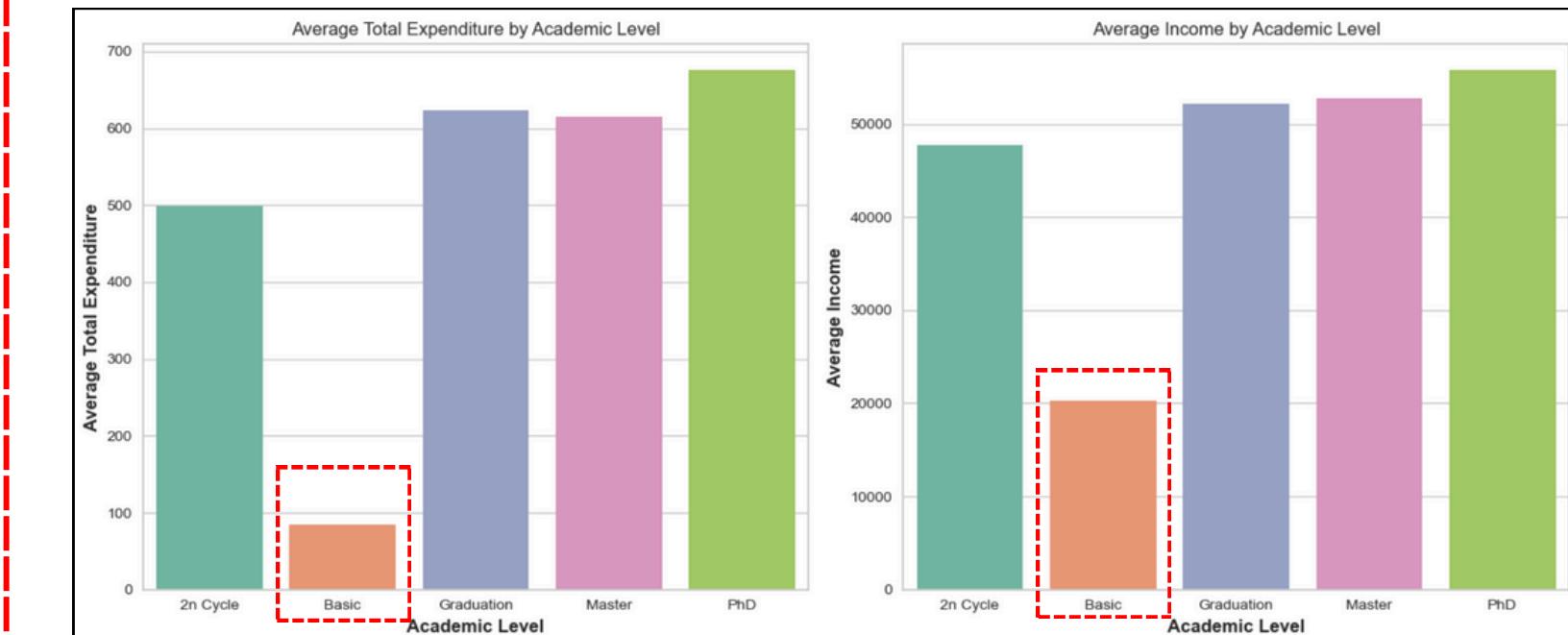


Hình 6. Biểu đồ chi tiêu và thu nhập

Khách hàng của công ty tập trung ở độ tuổi **30 - 40 tuổi** có mức độ thu nhập và chi tiêu cao

**Thu nhập càng cao,
Chi tiêu càng nhiều.**

Phân tích theo HỌC VẤN



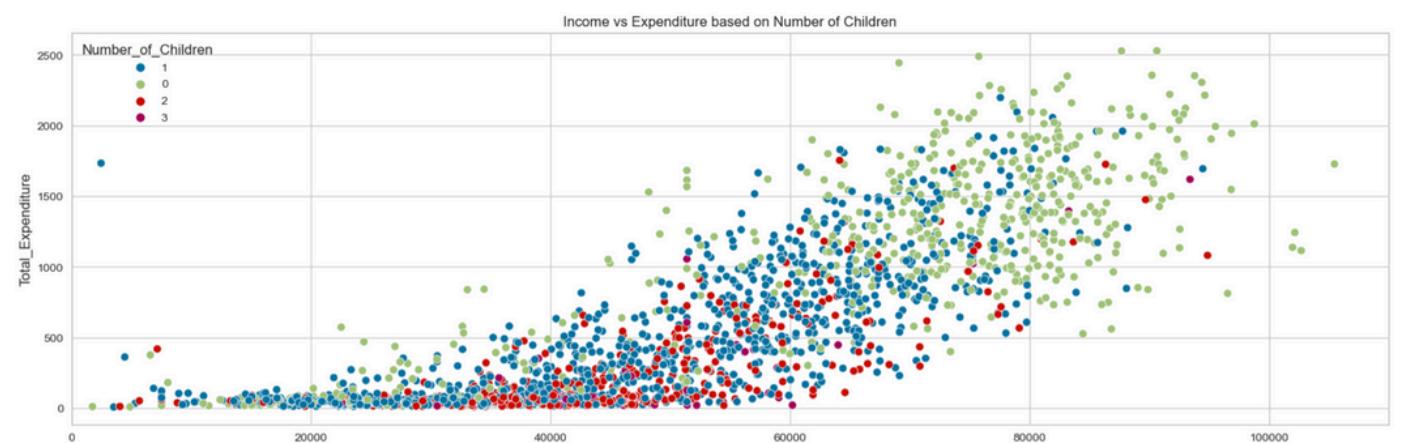
Hình 7. Biểu đồ chi tiêu theo học vấn

Hình 8. Biểu đồ thu nhập theo học vấn

- Mức **thu nhập và chi tiêu** trung bình của các nhóm người có trình độ **học vấn thấp (nhóm basic)** **thấp** hơn so với các nhóm người có trình độ **học vấn cao hơn**.
- Tỷ lệ chi tiêu so với thu nhập giảm** theo trình độ **học vấn**: Biểu đồ cho thấy nhóm người có trình độ **học vấn thấp** có **tỷ lệ chi tiêu so với thu nhập cao** hơn so với các nhóm người có trình độ **học vấn cao hơn**.

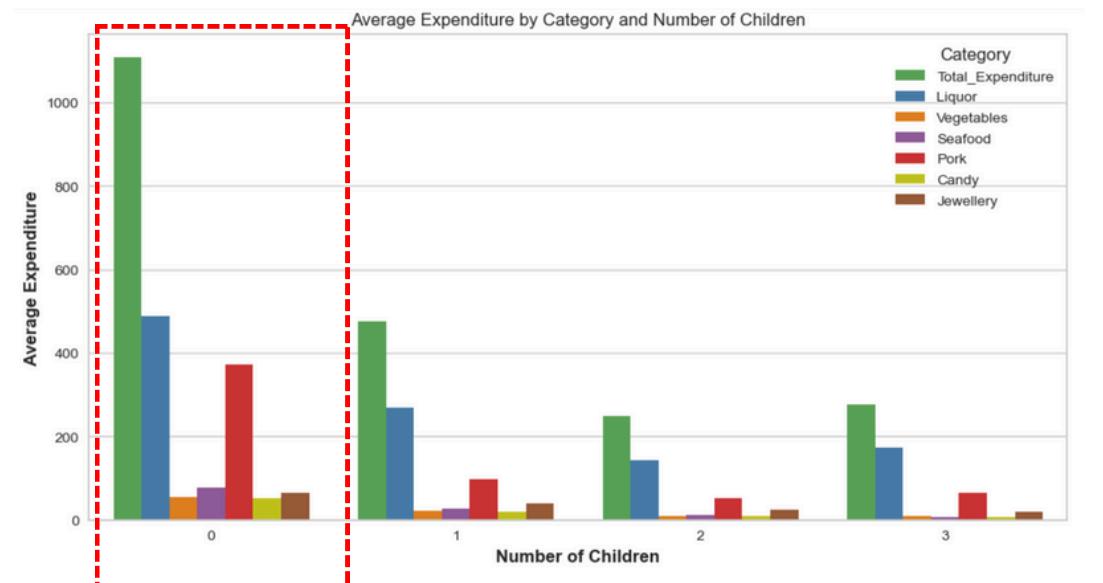
Có thể thấy những người có **học vấn càng cao** càng **chi tiêu nhiều** và “**mạnh tay**” hơn những người **học vấn cơ bản**.

Phân tích theo SỐ LƯỢNG CON



Hình 9. Biểu đồ chi tiêu và thu nhập theo số lượng con

Mức **thu nhập** và **chi tiêu trung bình** của gia đình **không con** cao hơn gia đình **đông con**

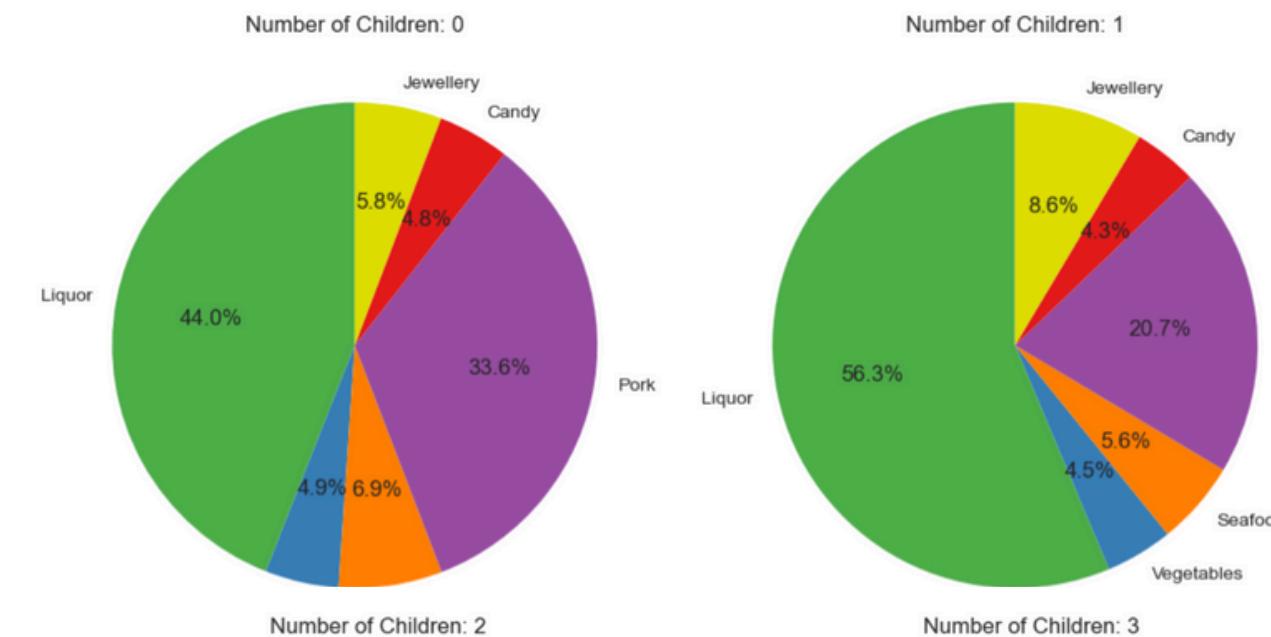


Hình 10. Biểu đồ chi tiêu các mặt hàng theo số lượng con

Hộ gia đình, cá nhân **không con** tiêu thụ **nhiều hơn** so với hộ gia đình **có con**, đặc biệt ở mặt hàng **rượu**.

Các gia đình **đông con** chi nhiều tiền hơn cho **thực phẩm**: thịt, cá, rau củ thay vì rượu

Phân tích theo SỐ LƯỢNG CON



Tương đồng

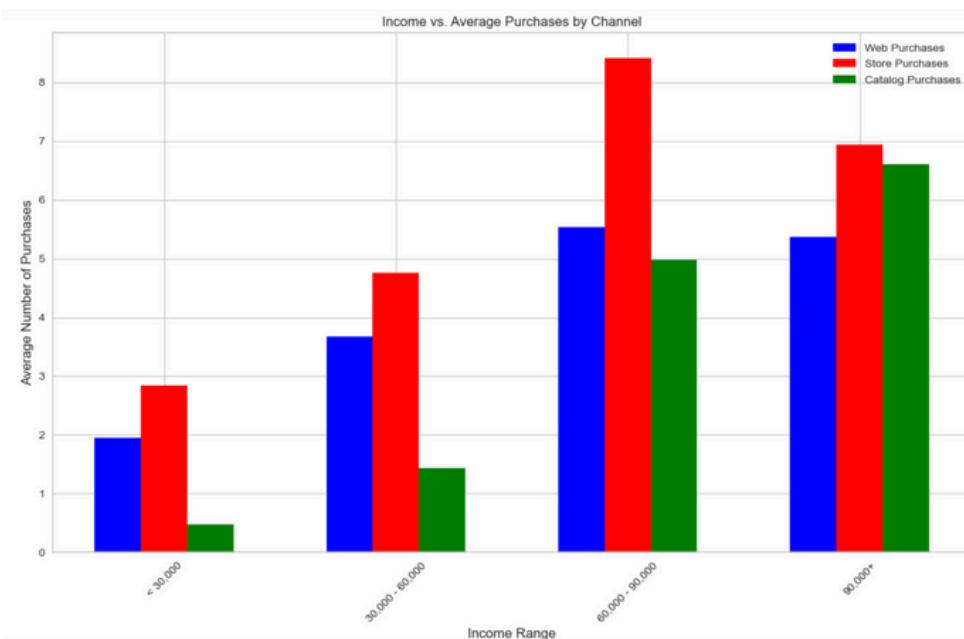
- Các hộ gia đình đều chi tiêu **phần lớn rượu** trong tổng chi tiêu.
- Chi tiêu cho **đồ ngọt** của 4 nhóm đều chiếm **tỷ lệ ít nhất**.

Khác nhau

- Các nhóm **có con** có xu hướng **chi tiêu trang sức** trong tổng mức chi tiêu nhiều hơn nhóm **không con**.

Phân tích theo KÊNH MUA HÀNG

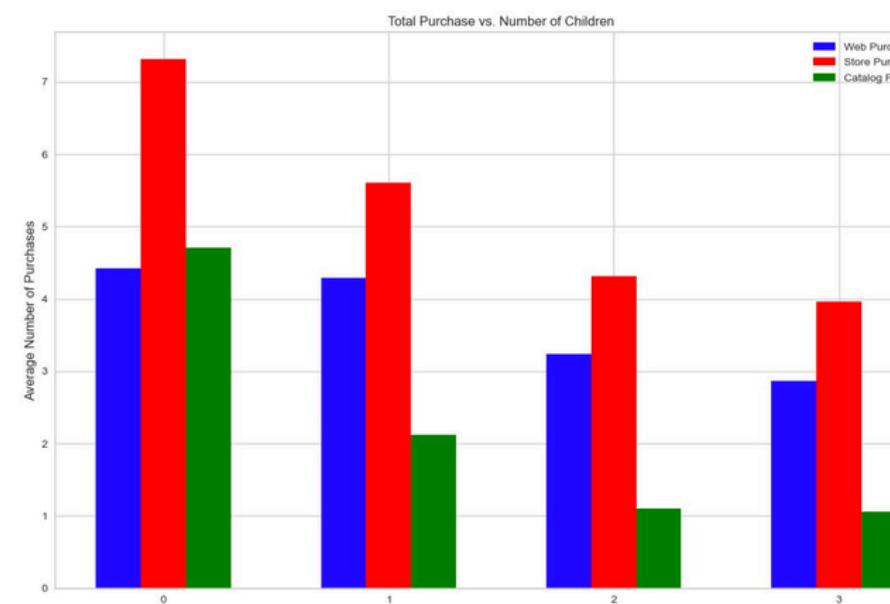
Thu nhập



Hình 12. Biểu đồ kênh mua hàng theo thu nhập

- Dù thu nhập **cao hay thấp**, khách hàng vẫn mua hàng thông qua **cửa hàng** nhiều hơn so với mua hàng qua **web và catalog**.
- Thu nhập càng tăng**, khách hàng càng có xu hướng mua qua **catalog** nhiều hơn.

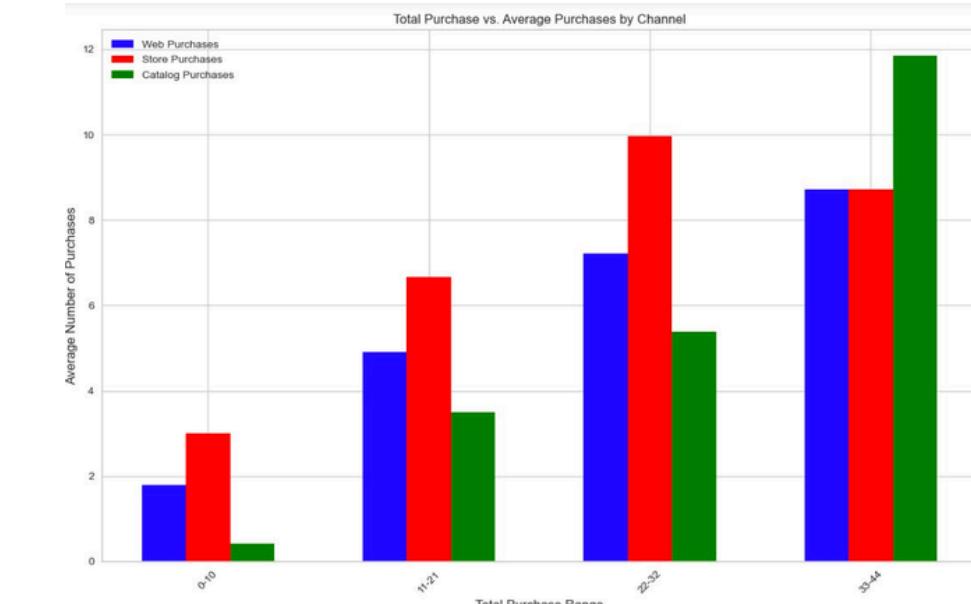
Số lượng con



Hình 13. Biểu đồ kênh mua hàng theo số lượng con

- Các khách hàng **không con** và **đông con** mua hàng qua **cửa hàng** nhiều hơn.
- Tuy nhiên những gia đình **không con** mua qua **catalog** nhiều hơn gia đình đông con.

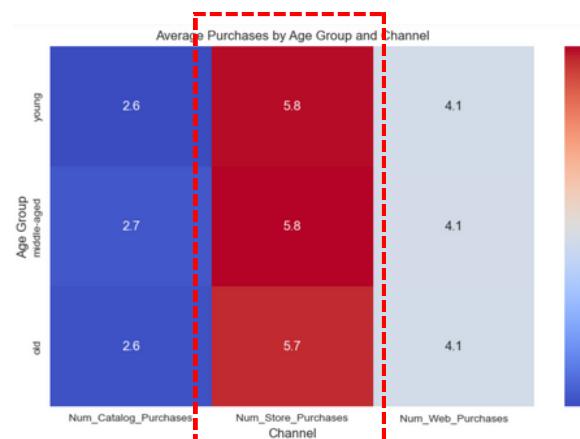
Giao dịch



Hình 14. Biểu đồ kênh mua hàng theo số lượng giao dịch

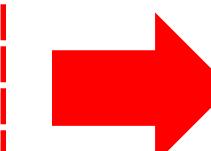
- Phần lớn khách hàng mua qua **cửa hàng** dù **giao dịch nhiều hay ít**
- Khách hàng có tổng lượng giao dịch **nhiều nhất (40 - 44 giao dịch)** mua qua **catalog** nhiều nhất

Phân tích theo KÊNH MUA HÀNG



Độ tuổi

Dù ở **độ tuổi** nào, khách hàng vẫn mua hàng thông qua **cửa hàng** nhiều nhất



Bất kể **thu nhập, độ tuổi, số lượng con, số lượng giao dịch**, khách hàng mua hàng thông qua **cửa hàng** nhiều nhất. Tuy nhiên, những khách hàng có **thu nhập cao, không con và giao dịch nhiều** có xu hướng mua hàng qua **catalog** nhiều hơn các nhóm còn lại.

Hình 15. Biểu đồ kênh mua hàng theo độ tuổi

KIỂM NGHIỆM A/B

Hai ưu đãi với mức **giảm giá 10%** và mức **giảm giá 30%** có tỷ lệ chấp nhận ưu đãi từ khách hàng **cao nhất**



Tiến hành thử nghiệm A/B so sánh **tính hiệu quả** của **Promo_10** và **Promo_30** đối với **số lần giao dịch giảm giá** của khách hàng

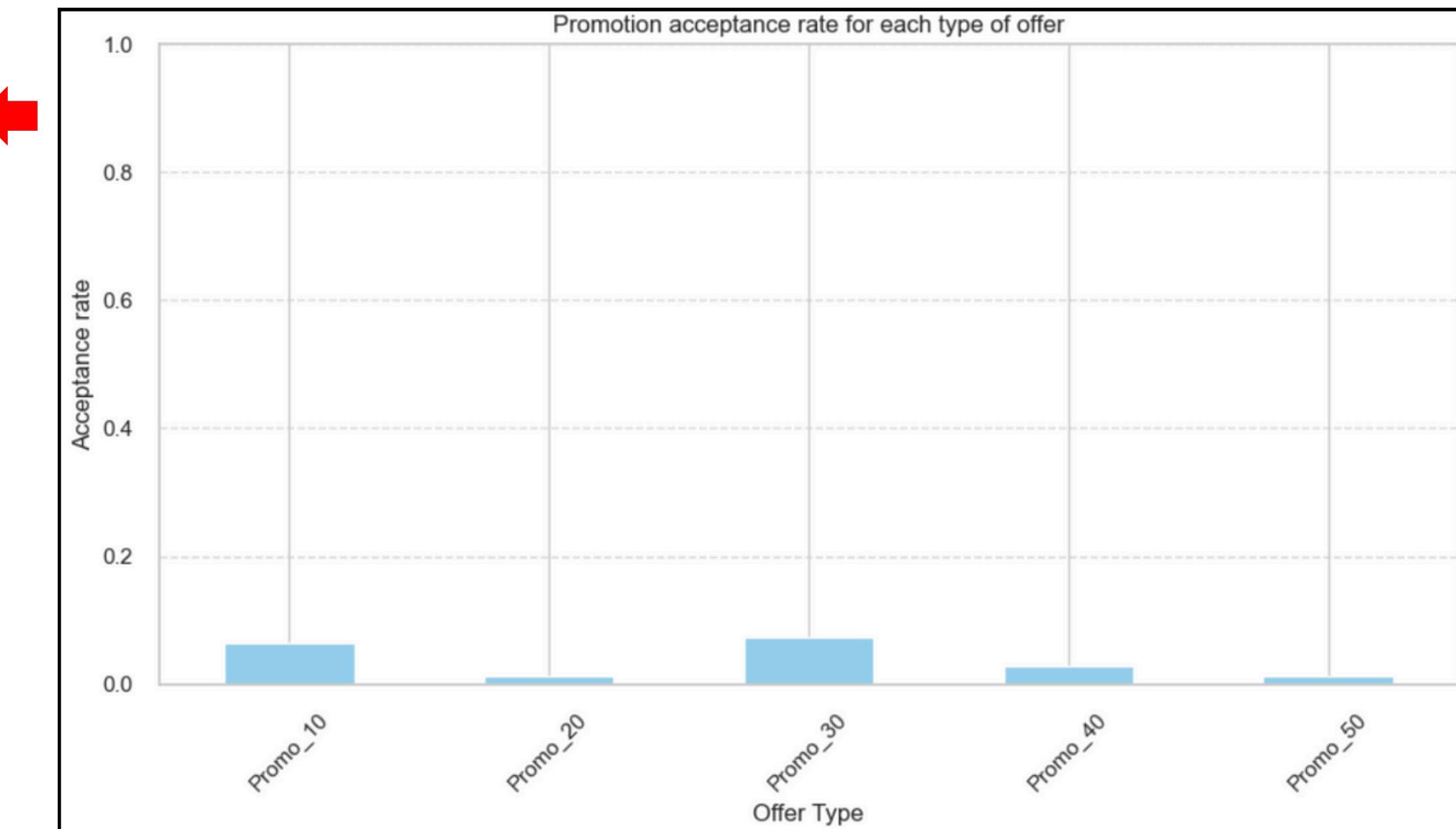


H0: Số lượng giao dịch mua trung bình với Promo_30 nhỏ hơn hoặc bằng với Promo_10.
H1: Số lượng giao dịch mua trung bình với Promo_30 nhiều hơn so với Promo_10.



P_value < 5%

Power = 97,24% > 80%



Hình 16. Tỷ lệ khách hàng chấp nhận cho từng ưu đãi

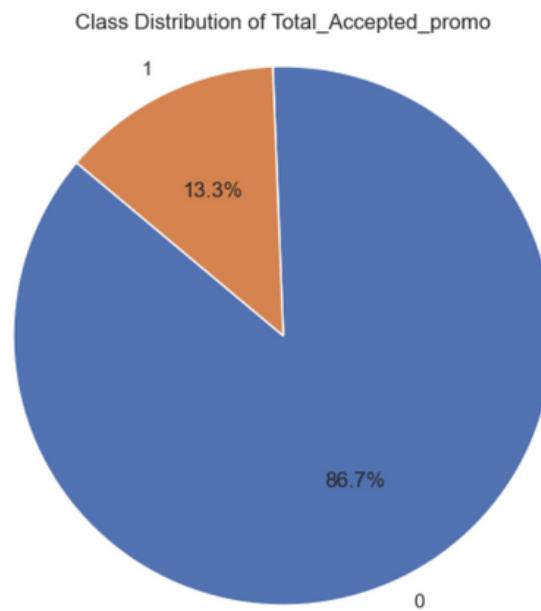
Số lượng giao dịch mua trung bình với Promo_30 nhiều hơn so với Promo_10.

▶ **Promo_30 hiệu quả hơn Promo_10.**

Tiến hành phân tích các giả thuyết về số lượng giao dịch theo giảm giá giữa các gói Promo khác nhau nhưng chưa thể kết luận mối quan hệ hiệu quả giữa các gói Promo còn lại (Phụ lục 3)

Tuy nhiên, phần lớn khách hàng không chấp nhận ưu đãi vì một số yếu tố như Num_Catalog_Purchase, Age_Group, Num_Store_Purchase.

Số lượng khách hàng không chấp nhận ưu đãi nhiều hơn



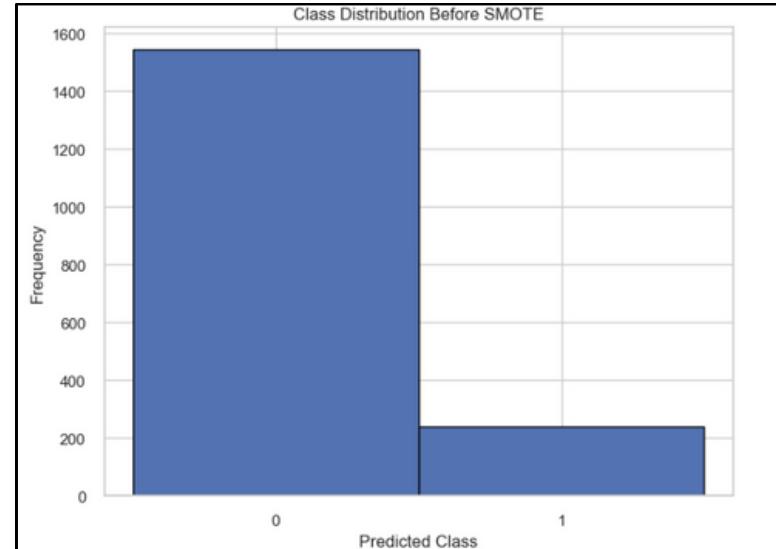
Dự đoán X

Dự đoán: Mức độ không chấp nhận ưu đãi của khách hàng cao hơn chấp nhận ưu đãi

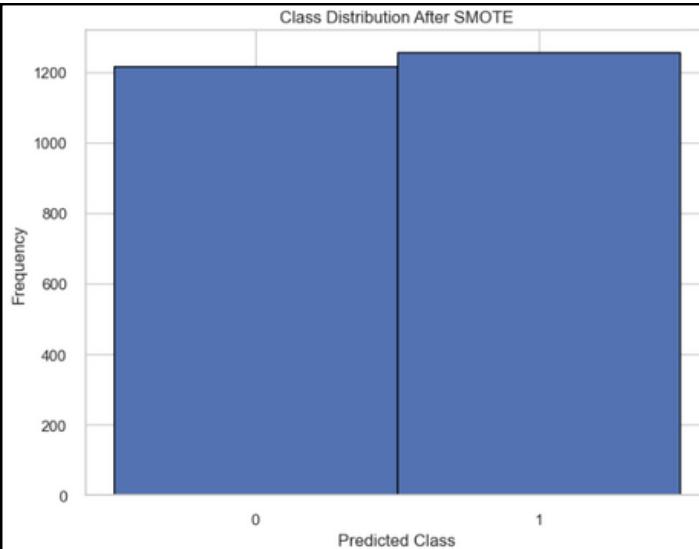
Hình 17. Tỉ lệ chấp nhận và không chấp nhận khuyến mãi của khách hàng.

Xử lý dữ liệu lớp cân bằng bằng phương pháp SMOTE để chuẩn bị cho các mô hình Feature Selection

Trước



Sau



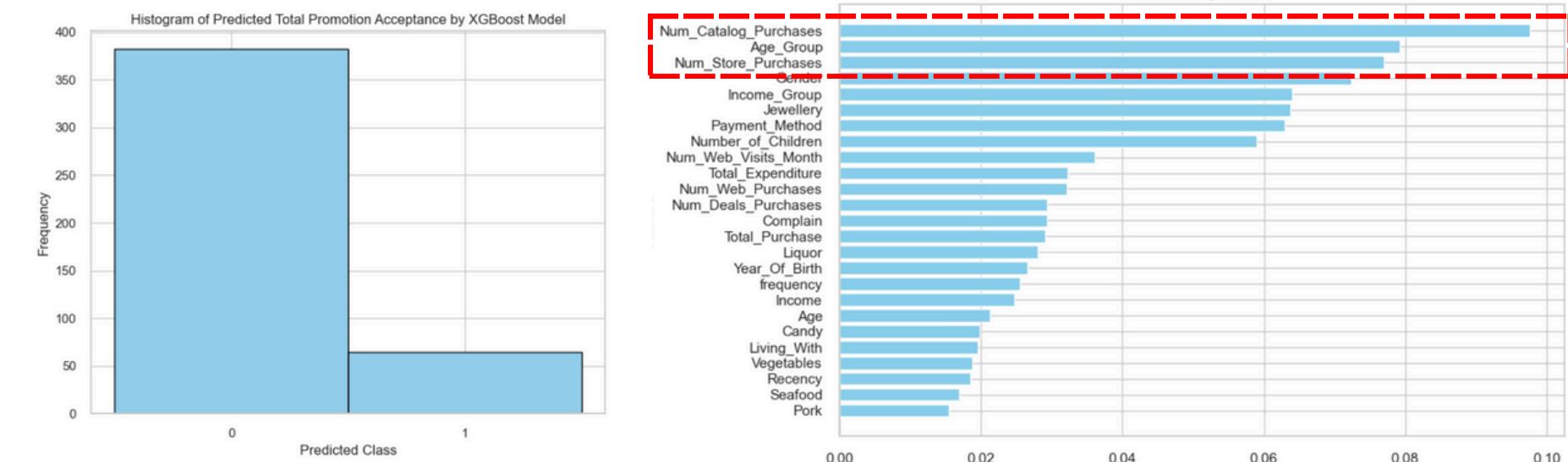
Hình 18. Bộ dữ liệu huấn luyện cho các mô hình Feature Selection trước và sau

Kết quả phân tích chỉ số đánh giá độ hiệu quả của các mô hình kiểm tra

Model	Accuracy	Balance Accuracy	Precision	Recall	F1 Score	ROC -AUC
XGBoost	0.86	0.71	0.44	0.50	0.47	0.83
Random Forest	0.86	0.71	0.44	0.52	0.48	0.83
Logistic Regression	0.77	0.72	0.31	0.64	0.42	0.79
Gradient Boosting	0.84	0.72	0.41	0.55	0.47	0.82
Neural Network	0.85	0.60	0.38	0.27	0.32	0.76

So sánh các chỉ số thu được về độ hiệu quả, **XGBoost là mô hình hiệu quả nhất cho Feature Selection.**

FEATURE SELECTION = XGBOOST



Hình 19. Dự đoán tỉ lệ chấp nhận khuyến mãi (XGBoost)

Dự đoán X chính xác với độ chính xác 86% của XGBoost.

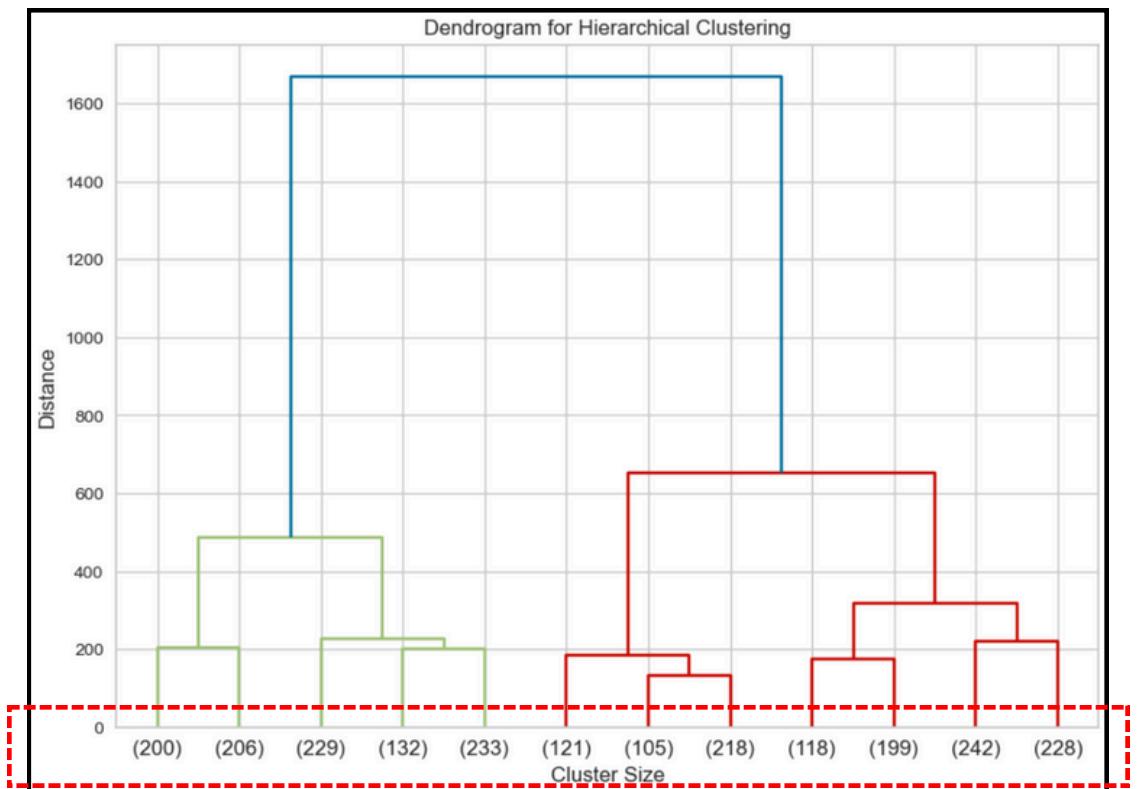
Hình 20. Dự đoán yếu tố tác động đến ý định không chấp nhận khuyến mãi (XGBoost)

Num_Catalog_Purchase, Age_Group, Num_Store_Purchase là các yếu tố chính thúc đẩy khách hàng không chấp nhận khuyến mãi giảm giá.

Hierarchical clustering

Tiếp cận theo hướng phân cụm phân chia (divisive clustering):

- Bắt đầu = 1 cụm chứa tất cả các điểm dữ liệu.
- Chia cụm này thành 2 cụm con có tổng khoảng cách trong các cụm là nhỏ nhất.
- Lặp lại bước 2 đến khi mỗi điểm dữ liệu là một cụm riêng lẻ hoặc đạt đến tiêu chí dừng nào đó.

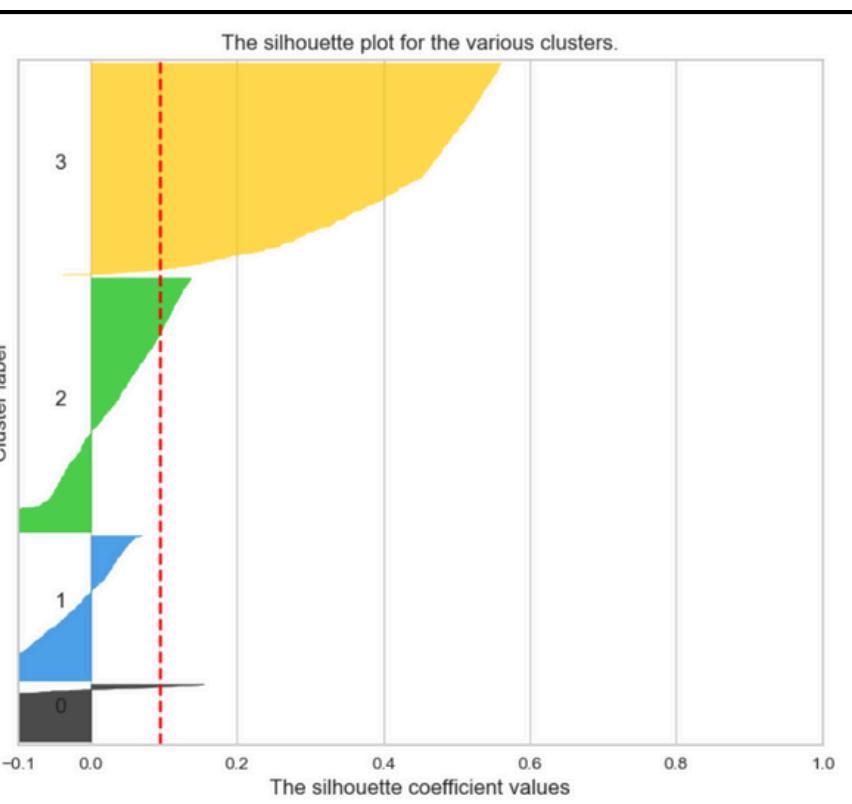


Hình 21. Sử dụng Dendrogram để xác định số cụm

k = 12 và số lượng khách hàng mỗi cụm không đều nhau

Gaussian Mixture Model (GMM)

- Giả định rằng dữ liệu của bạn là sự kết hợp của nhiều phân phối Gaussian (nhiều chuông).
- Xác định các tham số của các "chuông": GMM tìm ra vị trí trung tâm (mean), độ rộng (covariance), và chiều cao (weight) của mỗi chuông để mô tả dữ liệu một cách tốt nhất.

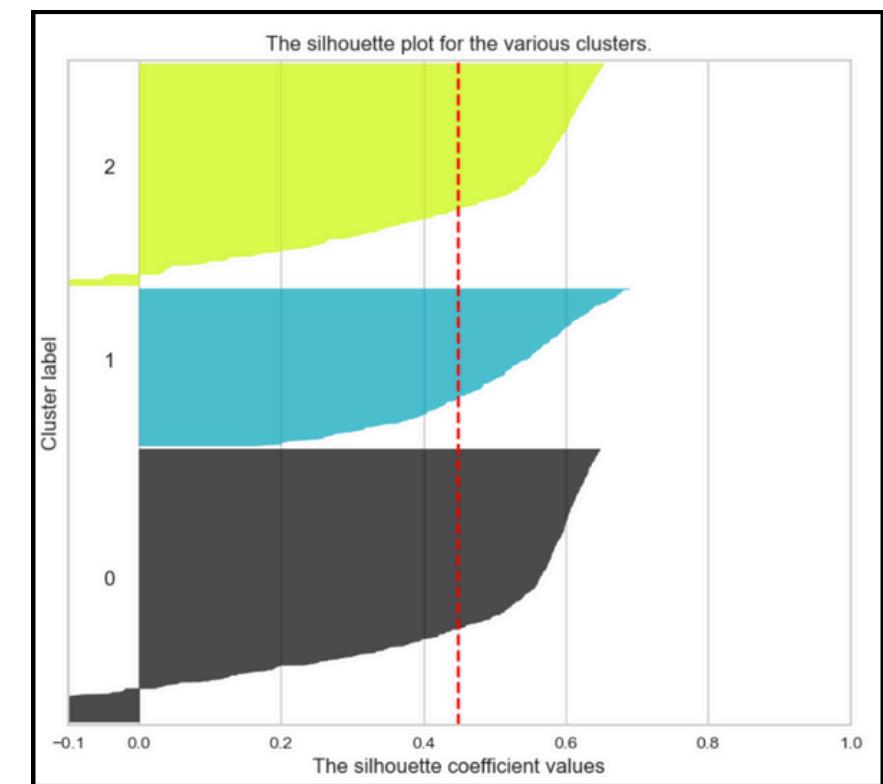


Hình 22. Phân tích Sihoutte cho GMM Clustering

k = 4 và độ rộng các cụm chênh lệch lớn, đặc biệt giữa cụm 3 và cụm 0.

Mean Shift Clustering

- Là một thuật toán phân cụm không tham số, không yêu cầu xác định trước số lượng cụm.
- Ý tưởng:** Tìm kiếm các đỉnh của mật độ dữ liệu để xác định các cụm.



Hình 23. Phân tích Sihoutte cho Mean Shift Clustering

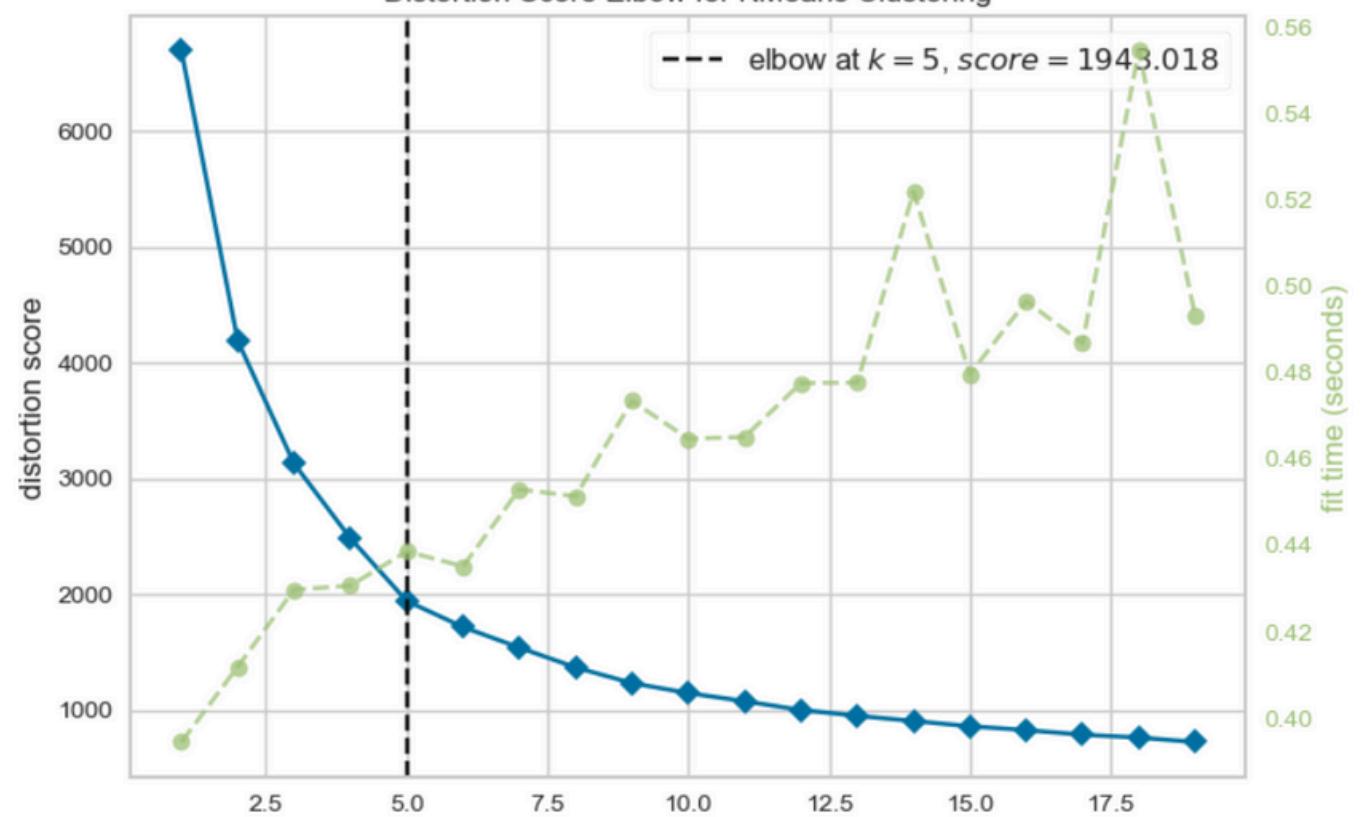
k = 3; phân bố không đồng đều giữa các cụm

Kết quả thu được đa dạng về số lượng cụm khách hàng nhưng phân phối giữa các cụm không đồng đều và chưa hiệu quả.

K-Means clustering

- Xác định các yếu tố quan trọng để tiến hành chạy phân cụm
- Sử dụng phương pháp Elbow đưa ra chỉ số phân cụm k
- Sử dụng thêm phương pháp Sihoutte để củng cố kết quả cho phương pháp Elbow
- Từ chỉ số phân cụm k tiến hành xây dựng mô hình k-mean cluster

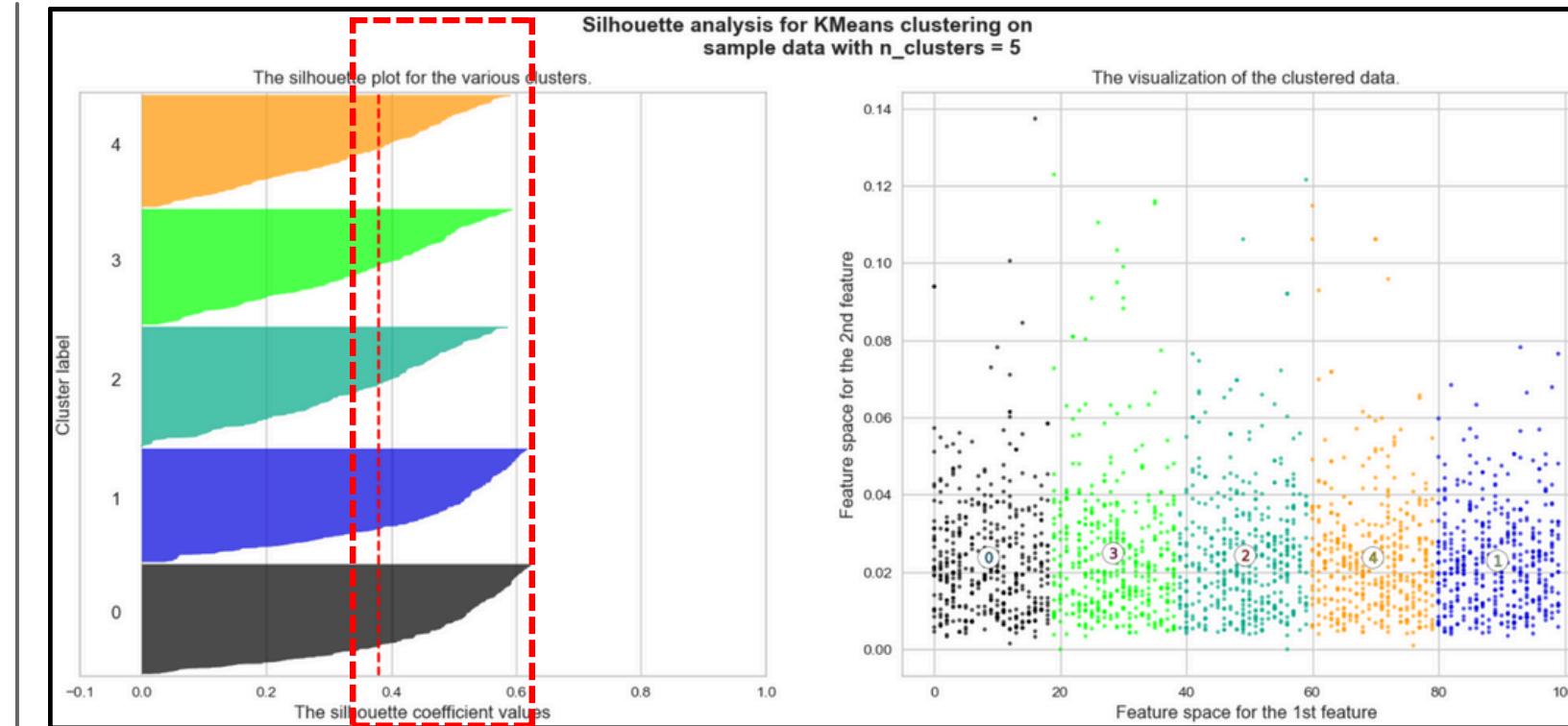
Distortion Score Elbow for KMeans Clustering



Hình 24. Khảo sát số cụm bằng Elbow Rule

* Phân tích Elbow:

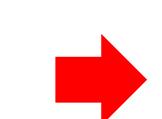
Chạy quá trình phân cụm và các điểm centroids lặp lại 300 lần để tăng độ chính xác, thu được **k = 5**.



Hình 25. Phân tích Sihoutte với n = 5

Kết quả phân tích chỉ số đánh giá hiệu quả của thuật toán phân cụm

Model	Sihoutte score	Mean Sihoutte score	Std Sihoutte score
K-Means	0.45	0.42	0.01
Hierarchical	0.34	0.34	0.02
GMM	0.09	0.16	0.07
Mean Shift	0.45	0.53	0.05



K-Means

Lý do kết hợp thuật toán K-Means và mô hình RFM

- RFM tập trung trực tiếp vào **hành vi mua sắm** khách hàng.
 - RFM **tiếp cận rõ ràng** để đánh giá **giá trị và tiềm năng** khách hàng dựa trên tần suất và giá trị của giao dịch.
 - RFM là **mô hình** được sử dụng **rộng rãi** trong **việc tối ưu hóa CRM**
- => **Tiền đề cho bước sau:** Tạo ra các chiến dịch nhắm mục tiêu cụ thể và cải thiện quan hệ với khách hàng.

Tiến hành phân cụm khách hàng bằng phương pháp K-Means theo 3 tiêu chí

Monetary

Monetary (M): tổng số tiền khách hàng đã chi trả cho công ty trong suốt vòng đời mua hàng

Frequency

Frequency (F): tần suất giao dịch

Recency

Recency (R): khoảng thời gian tính từ lần cuối giao dịch tới hiện tại

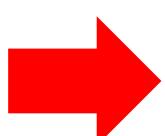
Tiến hành phân cụm khách hàng theo các chỉ số mô hình RFM

M_score: phân cụm monetary thành 5 cụm theo ngũ phân vị: từ 1 đến 5 với chỉ số càng cao Monetary càng lớn

F_score: phân cụm frequency thành 5 cụm theo ngũ phân vị: từ 1 đến 5 với chỉ số càng cao Frequency càng lớn

R_score: phân cụm Recency thành 5 cụm theo ngũ phân vị: từ 1 đến 5 với chỉ số càng cao Recency càng nhỏ

Sau khi kết hợp phương pháp K-means và mô hình RFM.
=> Xác định được cluster và chỉ số RFM của từng khách hàng.



ID	recency	frequency	monetary	income	Education	Catalog	Store	Web	r	f	m	rfm_sum	Cluster	
831	1001	37.00	0.03	1,105.00	61074	Graduation	5.00	8.00	11.00	4	4	4	12	1
1132	1002	92.00	0.03	738.00	60093	Graduation	2.00	10.00	7.00	1	4	4	9	0
301	1005	65.00	0.05	1,318.00	79689	2n Cycle	9.00	13.00	4.00	2	5	5	12	0
1101	1006	12.00	0.01	67.00	41021	Master	0.00	3.00	2.00	5	1	2	8	3
2888	1007	55.00	0.03	1,665.00	57333	PhD	5.00	9.00	8.00	3	4	5	12	0
...

Hình 26. Cluster và chỉ số RFM của từng khách hàng

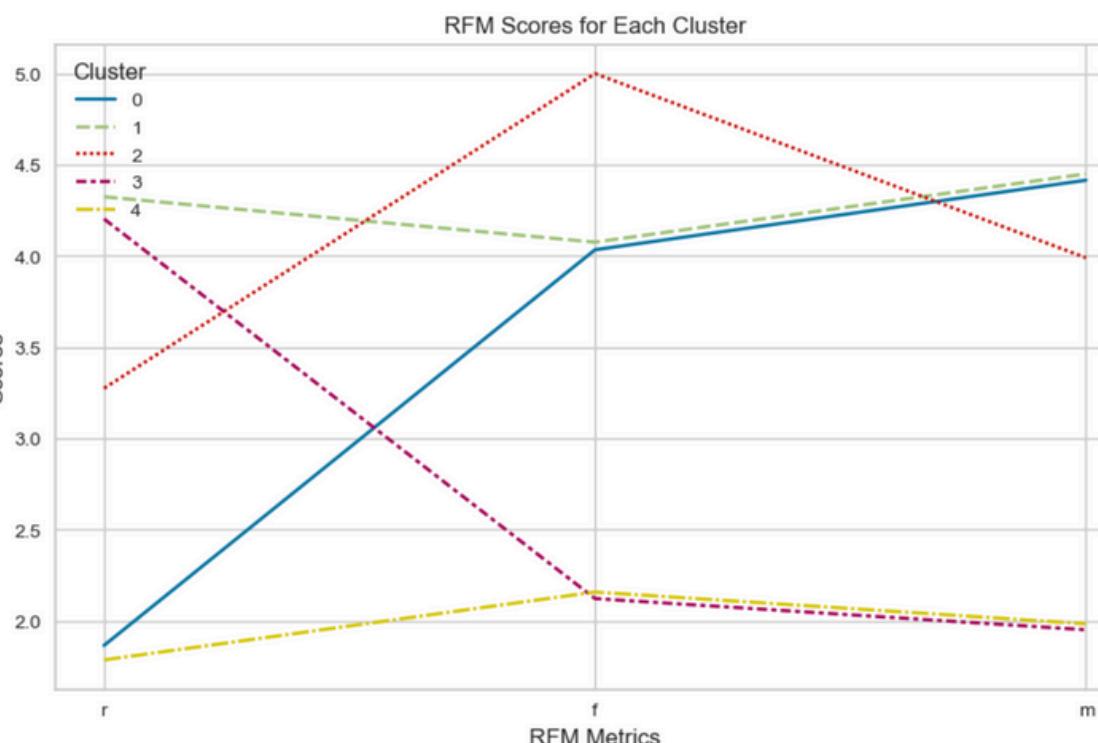
Tiến hành xác định giá trị trung bình (mean) của các chỉ số R, F, M cho mỗi cluster

cluster	r	f	m
0	1.87	4.03	4.42
1	4.32	4.08	4.45
2	3.27	5.00	3.99
3	4.20	2.12	1.95
4	1.79	2.16	1.98

Hình 27. Giá trị mean R, F, M theo từng cluster

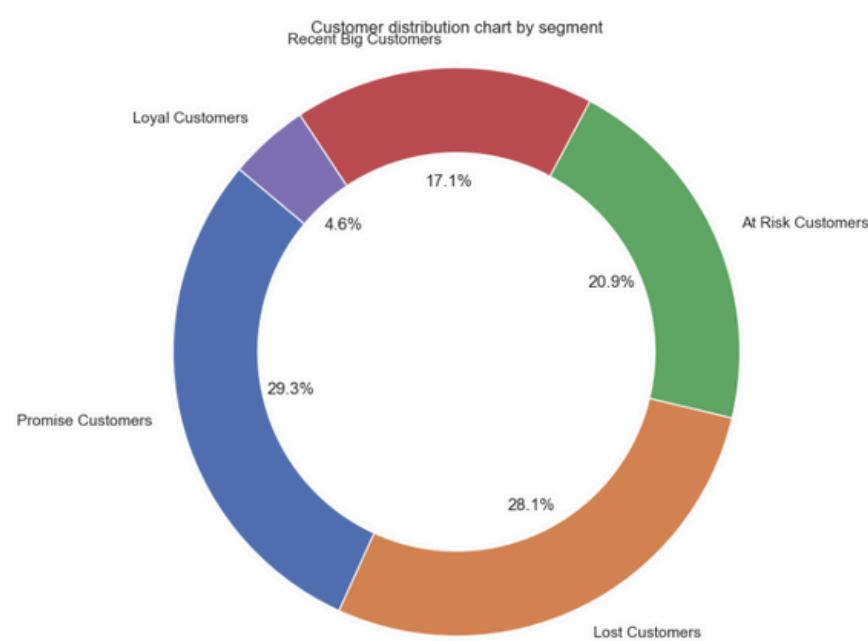
=> **Mục tiêu: Tìm được tính chất của mỗi cluster**

Vẽ biểu đồ từ hình 27 để xác định được nhóm khách hàng của công ty



Hình 28. Biểu đồ thể hiện chỉ số R, F, M từng nhóm

- Cluster 1: Recent Big Customers** là những khách hàng **chi tiêu nhiều** cho công ty, **mua gần đây nhất** nhưng mua hàng **chưa thường xuyên**
- Cluster 0: At Risk customers** là những khách hàng **chi tiêu nhiều** cho công ty, mua **chưa thường xuyên** và **lần cuối mua khá lâu**
- Cluster 2: Loyal Customers** là khách hàng mua **cực kì thường xuyên**, mức độ **chi tiêu trung bình** và thời gian từ **lần mua gần nhất trung bình**
- Cluster 3: Promise Customers** là khách hàng mua **không thường xuyên**, **chi tiêu ít** nhưng mua **gần đây nhất**
- Cluster 4: Lost Customers** là khách hàng mua **không thường xuyên**, **chi tiêu ít** nhưng **lần cuối mua hàng đến nay xa**



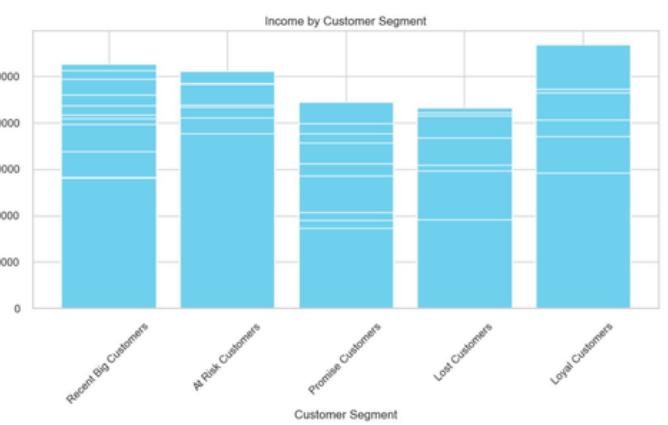
Hình 29. Tỷ lệ khách hàng theo từng nhóm

- Khách hàng đem phần lớn doanh thu cho công ty là Recent Big Customers, At Risk Customers và Loyal Customers : 42.6%
- Khách hàng mua hàng thường xuyên chỉ chiếm 4.6% (Loyal customers)
- Khách hàng có xu hướng rời bỏ chiếm 47% (At Risk Customers và Lost Customers)

Sau khi phân nhóm khách hàng theo chỉ số RFM, doanh nghiệp cần thực hiện những chiến lược thích hợp để cải thiện các chỉ số này.

=> Vì vậy, bước sau sẽ **phân tích sâu hơn về đặc điểm** của **từng nhóm** khách hàng => Chiến dịch **hiệu quả và đáp ứng** mục tiêu cụ thể.

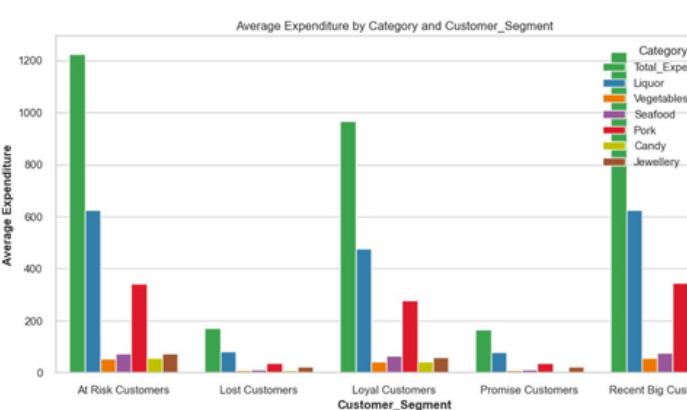
Phân tích theo THU NHẬP



Hình 30. Biểu đồ thể hiện thu nhập theo nhóm khách hàng

- Loyal Customers:** thu nhập cao
- Recent Big Customers và At Risk Customers:** thu nhập tương đối cao
- Promise Customers và Lost Customers:** thu nhập trung bình

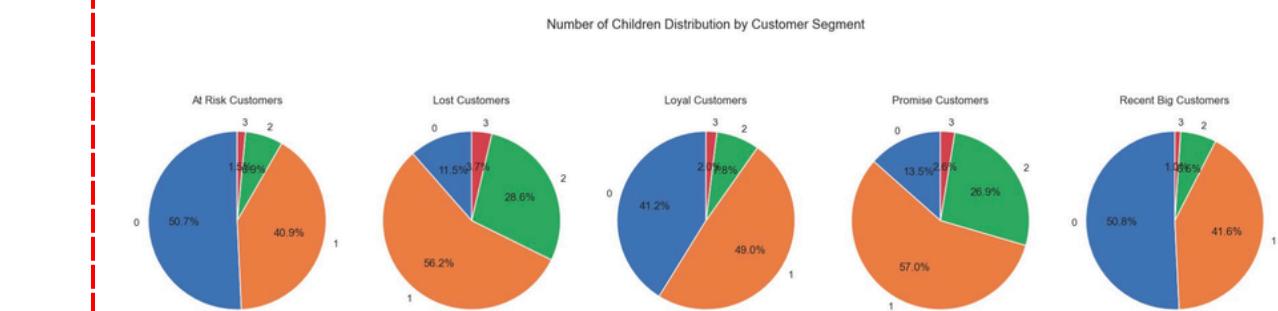
Phân tích theo CHI TIÊU



Hình 31. Biểu đồ thể hiện chi tiêu theo nhóm khách hàng

- Recent Big customers, Loyal customers và At risk customers :** chi tiêu nhiều, “mạnh tay”, đặc biệt là rượu
- Promise customers và Lost customers:** chi tiêu ít và chi tiêu rượu ít

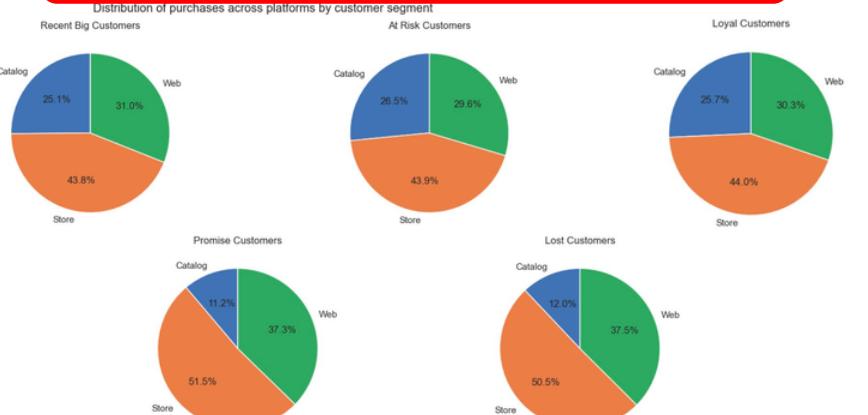
Phân tích theo SỐ LƯỢNG CON



Hình 32. Biểu đồ thể hiện số lượng con theo nhóm khách hàng

- Recent Big customers, Loyal customers và At risk customers:** chủ yếu là gia đình không con
- Promise customers và Lost customers:** gia đình đông con

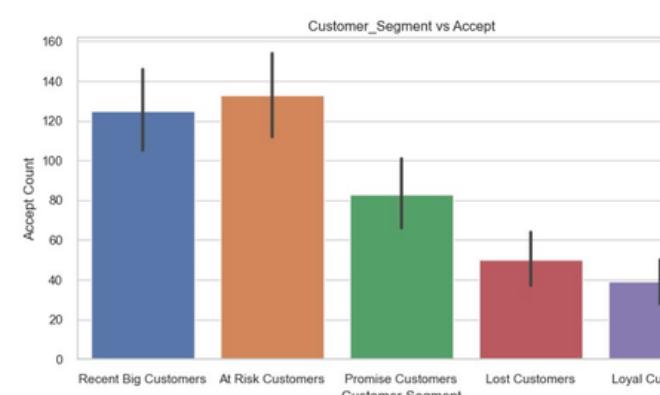
Phân tích theo KÊNH MUA HÀNG



Hình 33. Biểu đồ thể hiện tỷ lệ mua qua kênh mua hàng theo nhóm khách hàng

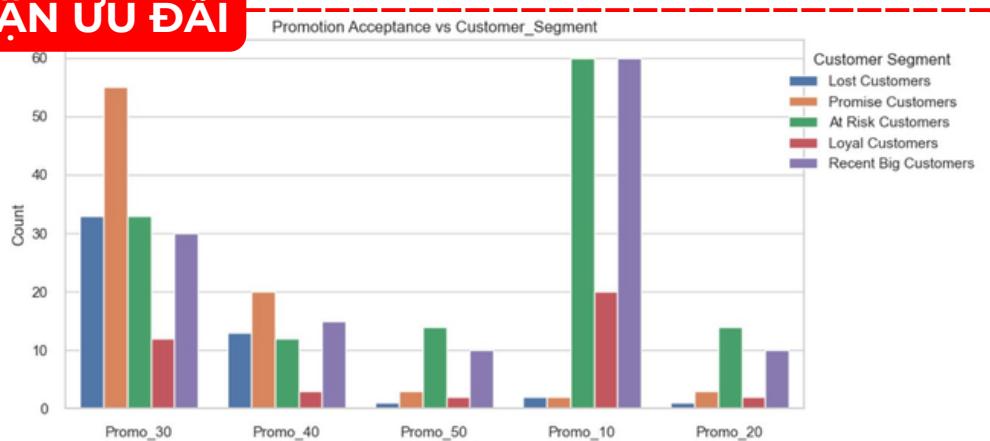
- Recent Big customers, Loyal customers và At risk customers:** mua hàng chủ yếu qua cửa hàng. Tỷ lệ mua hàng qua catalog cao hơn nhóm còn lại
- Promise customers và Lost customers:** chủ yếu mua qua cửa hàng

Phân tích theo CHẤP NHẬN ƯU ĐÃI



Hình 34. Biểu đồ thể hiện tỷ lệ chấp nhận ưu đãi theo nhóm khách hàng

- Recent Big customers và At risk customers:** tỷ lệ chấp nhận ưu đãi cao
- Promise customers:** chấp nhận ưu đãi trung bình
- Lost Customers và Loyal Customers:** tỷ lệ chấp nhận ưu đãi thấp



Hình 35. Biểu đồ thể hiện tỷ lệ chấp nhận từng loại ưu đãi theo nhóm khách hàng

- Recent Big Customers, At risk customers và Loyal Customers:** tỷ lệ chấp nhận ưu đãi 10% nhiều nhất
- Promise customers và Lost Customers:** đa phần chấp nhận ưu đãi 30% và 40%

Segmentation	Recency	Frequency	Monetary	Đặc điểm	Đề xuất
Recent Big Customers	21.46	0.03	1232.42	<ul style="list-style-type: none"> Thu nhập tương đối cao và chi tiêu nhiều, “mạnh tay” Gia đình không con, chi tiêu rượu nhiều Mua qua cửa hàng chủ yếu và có xu hướng mua qua catalog cao Tỷ lệ chấp nhận ưu đãi cao Chấp nhận ưu đãi thấp: 10% 	CHIẾN DỊCH PHÁT TRIỂN <ul style="list-style-type: none"> Tăng mức độ thường xuyên mua hàng Phát triển mối quan hệ để duy trì recent big customers CHIẾN DỊCH GIỮ CHÂN <ul style="list-style-type: none"> Tăng mức độ thường xuyên mua hàng Giữ chân khách hàng vì họ có xu hướng rời bỏ doanh nghiệp
At Risk Customers	73.03	0.03	1223.64	At Risk customer là những khách hàng chi tiêu nhiều cho công ty, mua chưa thường xuyên và lần cuối mua khá xa	
Loyal Customers	43.51	0.07	968.26	Loyal Customers là khách hàng mua cực kì thường xuyên , mức độ chi tiêu trung bình và thời gian từ lần mua gần nhất trung bình	CHIẾN DỊCH PHÁT TRIỂN <ul style="list-style-type: none"> Nhắc nhở khách hàng mua bằng các ưu đãi, chiến dịch Tăng mức độ chi tiêu cho công ty
Promise Customers	24.62	0.02	166.50	Promise Customers là khách hàng mua không thường xuyên , chi tiêu ít nhưng mua gần đây nhất	CHIẾN DỊCH GIỮ CHÂN & PHÁT TRIỂN <ul style="list-style-type: none"> Giữ chân khách hàng mới đáp ứng nhu cầu, cung cấp ưu đãi, Phát triển mối quan hệ khi đã giữ chân khách hàng
Lost Customers	74.64	0.02	170.64	Lost Customers là khách hàng mua không thường xuyên , chi tiêu ít nhưng lần cuối mua hàng đến nay xa	Cân nhắc giữa việc tìm hiểu lý do họ rời bỏ, khơi dậy sự quan tâm bằng chiến dịch tiếp cận hoặc bỏ qua. Vì chi phí giữ chân có thể cao hơn chi phí thu hút khách hàng mới

Chỉ số R, F, M trung bình của 5 nhóm khách hàng (Phụ lục 5)

CẢM ƠN VÌ ĐÃ
LĂNG NGHE

SPARKLE - DAZONE 2024

PHỤ LỤC

SPARKLE - DAZONE 2024

Minh họa: Khách hàng có ID 1128

ID	Year_of_Birth	Academic_Level	Income	Registration_Time	Recency	Liquor	Vegetables	Pork	Seafood	Candy	Jeweller	Num_Deals_Purchased
1128	1987	Graduation	54414.0	23/01/2022	49	110	20	16	24	26	18	4
1128	1987	Graduation	54414.0	23/01/2022		110		16	24	26	18	4
1128	1987	Graduation	54414.0	23/01/2022	49	110	20	16		26	18	4

ID	Num_Deals_Purchased	Num_Web_Purchases	Num_Catalog_Purchase	Num_Store_Purchase	Num_Web_Visits_Month	Promo_30	Promo_40	Promo_50	Promo_10	Promo_20	Complain
1128	4	3		1	5	4	0	-1	0	0	0
1128	4				5				0	0	
1128	4			1	5		0	-1	0	0	0

ID	Gender	Phone	Phone_Number	Year_Registered	Month_Registered	Total_Purchases	Living_Status	Payment_Method
1128	Other	8,46E+10				1	13	Together_2
1128	Other	8,46E+10				1	13	Together_2
1128		8,46E+10					Together_2	

Row Labels	Sum of Total_Expenditure						
2021	369751						
2022	718992						
2023	274965						
Grand Total	1363708						
Row Labels	Sum of Liquor	Sum of Veg	Sum of Pork	Sum of Seafood	Sum of Candy	Sum of Jewellery	
2021	192563	15669	97577	21970	15784	26188	
2022	358266	33638	196828	43414	31990	54856	
2023	132227	14090	79563	18673	12847	17565	
Grand Total	683056	63397	373968	84057	60621	98609	
Row Labels	Average of Total_Expenditure	Count of ID	Sum of Total_Purchase	Sum of Num_Store_Purchases	Sum of Num_Catalog_Purchases	Sum of Num_Web_Purchases	Sum of Num_Web_Visits_Month
2021	748,483,8057	494	8319	3049	1486	2334	3115
2022	604,703,1119	1189	17848	6992	3206	4937	6272
2023	493,653,5009	557	7124	2929	1271	1879	2522
Grand Total	608,798,2143	2240	33291	12970	5963	9150	11909

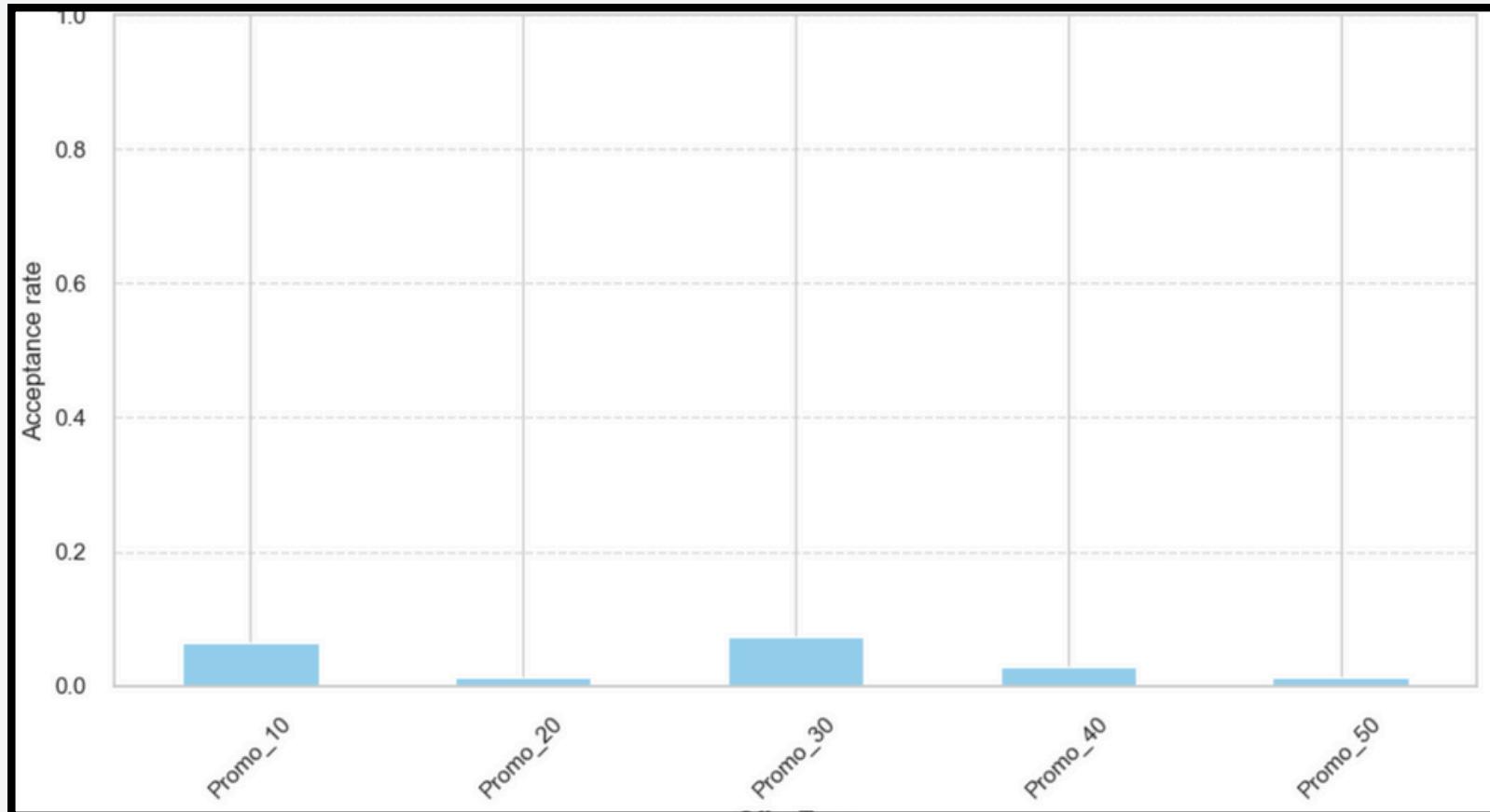
- **Phương pháp MICE (Multiple Imputation by Chained Equations)** là một kỹ thuật phổ biến được sử dụng để xử lý dữ liệu thiếu trong phân tích thống kê.
- Khi dữ liệu có giá trị bị thiếu, sử dụng MICE bao gồm việc điền vào những giá trị thiếu này bằng cách tạo ra một loạt các bản sao của dữ liệu, mỗi bản được điền vào một cách độc lập, và sau đó kết hợp chúng lại. Quá trình này được thực hiện qua nhiều lần lặp cho đến khi đạt được sự hội tụ.
- **Ý tưởng chính** đằng sau MICE là sử dụng các mô hình hồi quy hoặc các mô hình dự đoán khác để điền vào các giá trị thiếu. Trong mỗi bước, một biến được chọn để điền giá trị thiếu dựa trên các biến khác có sẵn trong dữ liệu. Sau đó, một mô hình được tạo ra để dự đoán giá trị của biến đã chọn dựa trên các biến khác. Quá trình này lặp đi lặp lại cho đến khi đạt đến số lần lặp định trước hoặc cho đến khi hội tụ.
- MICE có thể được sử dụng cho nhiều loại dữ liệu khác nhau và với các loại mô hình dự đoán khác nhau, tùy thuộc vào yêu cầu cụ thể của vấn đề và đặc điểm của dữ liệu.

Các bước triển khai phương pháp MICE:

- Khởi tạo dữ liệu với các giá trị chỗ trống:
- Ban đầu, các giá trị bị thiếu được thay thế bằng các giá trị tạm thời (chỗ trống). Thông thường, chỗ trống có thể là giá trị trung bình hoặc giá trị phổ biến nhất của cột tương ứng.

Sử dụng mô hình hồi quy để điền các giá trị bị thiếu:

- Đối với mỗi biến bị thiếu (ví dụ như thu nhập và phương thức thanh toán) lần lượt, một mô hình hồi quy được xây dựng bằng cách sử dụng các biến còn lại để dự đoán các giá trị bị thiếu.
- Lặp lại quá trình điền giá trị:
- Quá trình này được lặp lại nhiều lần (ví dụ: 10 lần lặp) để cập nhật các giá trị dự đoán cho đến khi hội tụ (các giá trị không thay đổi nhiều nữa).



Hình 1: Kiểm nghiệm xác định Promo nào sẽ hiệu quả hơn.

Độ hiệu quả giữa Promo_30 vs Promo_10 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_30 nhỏ hơn hoặc bằng với Promo_10.	Reject
H1		

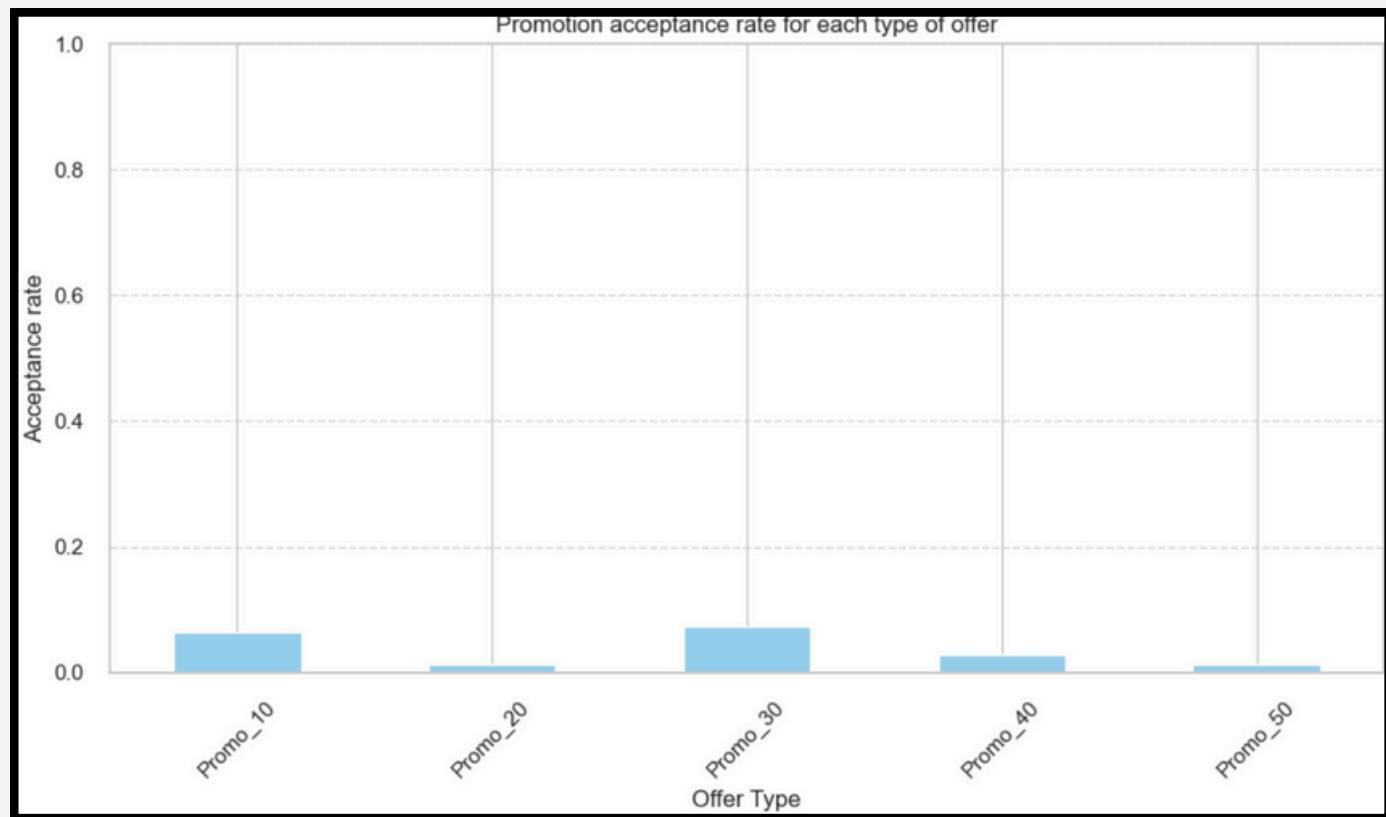
Độ hiệu quả giữa Promo_20 vs Promo_40 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_20 nhỏ hơn hoặc bằng với Promo_40.	None
H1		

Độ hiệu quả giữa Promo_40 vs Promo_30(P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_40 nhỏ hơn hoặc bằng với Promo_30.	None
H1		

Độ hiệu quả giữa Promo_10 vs Promo_20 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_10 nhỏ hơn hoặc bằng với Promo_20.	None
H1		

Độ hiệu quả giữa Promo_10 vs Promo_40 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_10 nhỏ hơn hoặc bằng với Promo_40.	None
H1		

Độ hiệu quả giữa Promo_10 vs Promo_50(P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_10 nhỏ hơn hoặc bằng với Promo_50.	Reject
H1		



Hình 1: Kiểm nghiệm xác định Promo nào sẽ hiệu quả hơn.

Độ hiệu quả giữa Promo_50 vs Promo_20 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_50 nhỏ hơn hoặc bằng với Promo_20.	None
H1		

Độ hiệu quả giữa Promo_50 vs Promo_30(P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_50 nhỏ hơn hoặc bằng với Promo_30.	None
H1		

Độ hiệu quả giữa Promo_50 vs Promo_40 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_50 nhỏ hơn hoặc bằng với Promo_40.	None
H1		

Độ hiệu quả giữa Promo_30 vs Promo_20 (P<0.05)		
H0	Số lượng giao dịch mua trung bình với Promo_30 nhỏ hơn hoặc bằng với Promo_20.	None
H1		

Gaussian Mixture Model (GMM) là một phương pháp thống kê dùng để tìm ra các nhóm (hay cụm) trong dữ liệu bằng cách giả định rằng dữ liệu được tạo ra từ sự kết hợp của nhiều "chuông" (phân phối Gaussian) khác nhau. Dưới đây là cách diễn giải dễ hiểu nhất về GMM:

Khái niệm cơ bản

Cụm (Cluster): Nhóm các điểm dữ liệu giống nhau.

Phân phối Gaussian: Đường cong hình chuông biểu diễn cách các giá trị dữ liệu phân bố xung quanh một trung bình (mean) nào đó.

Ý tưởng chính của GMM

Mô hình hóa dữ liệu bằng các "chuông": GMM giả định rằng dữ liệu của bạn là sự kết hợp của nhiều phân phối Gaussian (nhiều chuông). Mỗi phân phối Gaussian này đại diện cho một cụm trong dữ liệu.

Xác định các tham số của các "chuông": GMM tìm ra vị trí trung tâm (mean), độ rộng (covariance), và chiều cao (weight) của mỗi chuông để mô tả dữ liệu một cách tốt nhất.

Các bước thực hiện GMM

Khởi tạo: Bắt đầu với các giá trị ngẫu nhiên cho các tham số của mỗi chuông (mean, covariance, weight).

Expectation (E-step): Tính toán xác suất rằng mỗi điểm dữ liệu thuộc về mỗi chuông.

Maximization (M-step): Cập nhật các tham số của mỗi chuông dựa trên các xác suất vừa tính.

Lặp lại: Tiếp tục thực hiện E-step và M-step cho đến khi các tham số không thay đổi nhiều nữa (hội tụ).

Ví dụ đơn giản

Hãy tưởng tượng bạn có một bát kẹo với nhiều loại kẹo khác nhau. Bạn muốn chia kẹo thành các loại khác nhau dựa trên kích thước và màu sắc của chúng. GMM sẽ giúp bạn làm điều này bằng cách:

Giả sử rằng mỗi loại kẹo được phân bố xung quanh một trung bình kích thước và màu sắc nhất định.

Xác định các cụm (loại kẹo) và các đặc điểm của mỗi cụm.

Minh họa

Giả sử bạn có dữ liệu 2D (ví dụ: chiều dài và chiều rộng của các viên kẹo). GMM sẽ:

Tìm ra các cụm trong dữ liệu.

Xác định vị trí trung tâm và độ rộng của mỗi cụm.

Gán mỗi viên kẹo vào một cụm dựa trên xác suất nó thuộc về cụm đó.

Tóm lại

GMM là một công cụ mạnh mẽ giúp bạn tìm ra các nhóm trong dữ liệu bằng cách giả định rằng dữ liệu được tạo ra từ sự kết hợp của nhiều phân phối Gaussian. Nó không chỉ xác định các cụm mà còn cung cấp thông tin chi tiết về các cụm này.

Mean Shift Clustering là một phương pháp phân cụm dữ liệu dựa trên việc di chuyển các "điểm trung tâm" đến vị trí trung bình của các điểm dữ liệu xung quanh, cho đến khi không có thay đổi đáng kể nào nữa. Dưới đây là cách diễn giải đơn giản nhất về Mean Shift Clustering:

Ý tưởng chính

Khởi điểm: *Bắt đầu với mỗi điểm dữ liệu là một "điểm trung tâm" ứng với một cụm.*

Di chuyển trung tâm: *Dịch chuyển mỗi điểm trung tâm đến vị trí trung bình của các điểm dữ liệu trong một khu vực xung quanh, thường được xác định bởi một hàm kernel (ví dụ: hàm Gaussian).*

Lặp lại: Lặp lại quá trình này cho đến khi không có sự thay đổi đáng kể nào trong vị trí của các điểm trung tâm nữa.

Phân cụm: Gán mỗi điểm dữ liệu vào cụm của điểm trung tâm gần nhất sau khi hội tụ.

Ví dụ dễ hiểu

Giả sử bạn có một số lượng lớn các điểm trên một bản đồ và bạn muốn chia chúng thành các khu vực dựa trên mật độ dân số.

Bạn đặt một hòn đảo nhỏ (điểm trung tâm) ở mỗi điểm dữ liệu.

Sau đó, bạn di chuyển mỗi hòn đảo đến vị trí trung bình của một vùng xung quanh, dựa trên mật độ dân số.

Bạn lặp lại quá trình này cho đến khi không còn sự di chuyển đáng kể nữa.

Cuối cùng, mỗi điểm dữ liệu được gán vào khu vực của hòn đảo gần nhất.

Ứng dụng

Mean Shift Clustering thường được sử dụng trong việc phân cụm hình ảnh, phân tích dữ liệu không gian, nhận dạng đối tượng trong thị giác máy tính và nhiều lĩnh vực khác. Điều quan trọng là nó không yêu cầu bạn biết trước số lượng cụm cần tìm và có thể tìm ra các cụm có hình dạng phức tạp.

PHỤ LỤC 5: Chỉ số R, F, M trung bình của 5 nhóm khách hàng

