

Cuộc thi phân tích dữ liệu 2024

VÒNG 2.2 PRESENTATION

3G

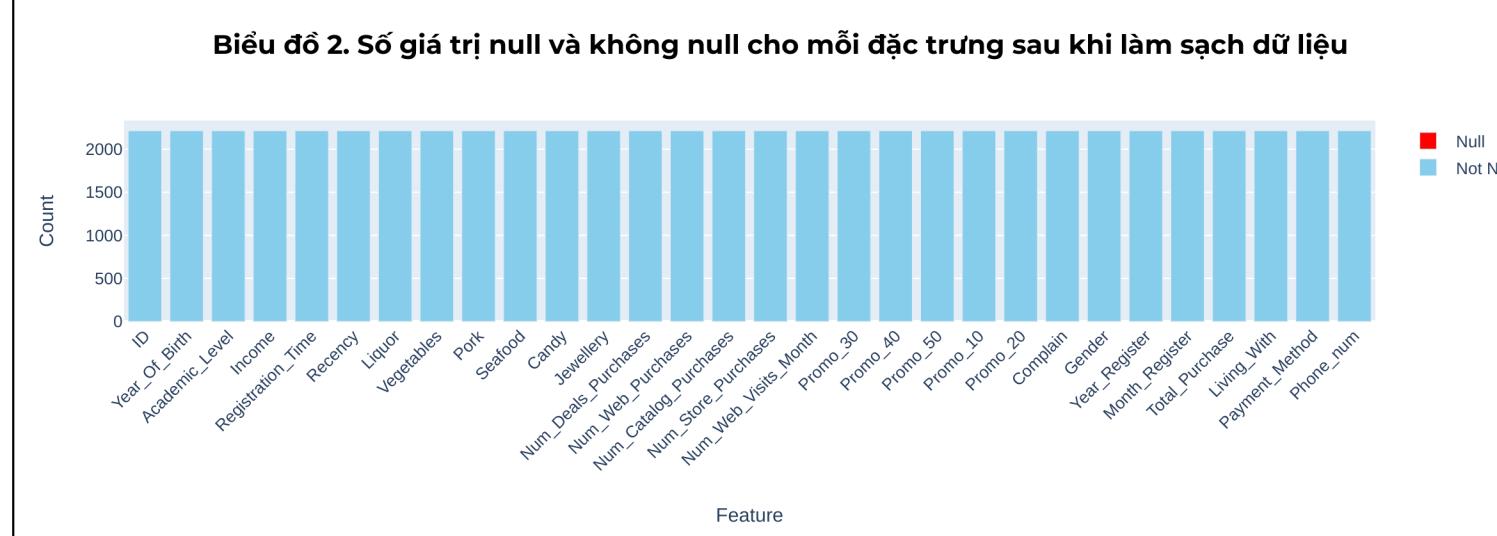
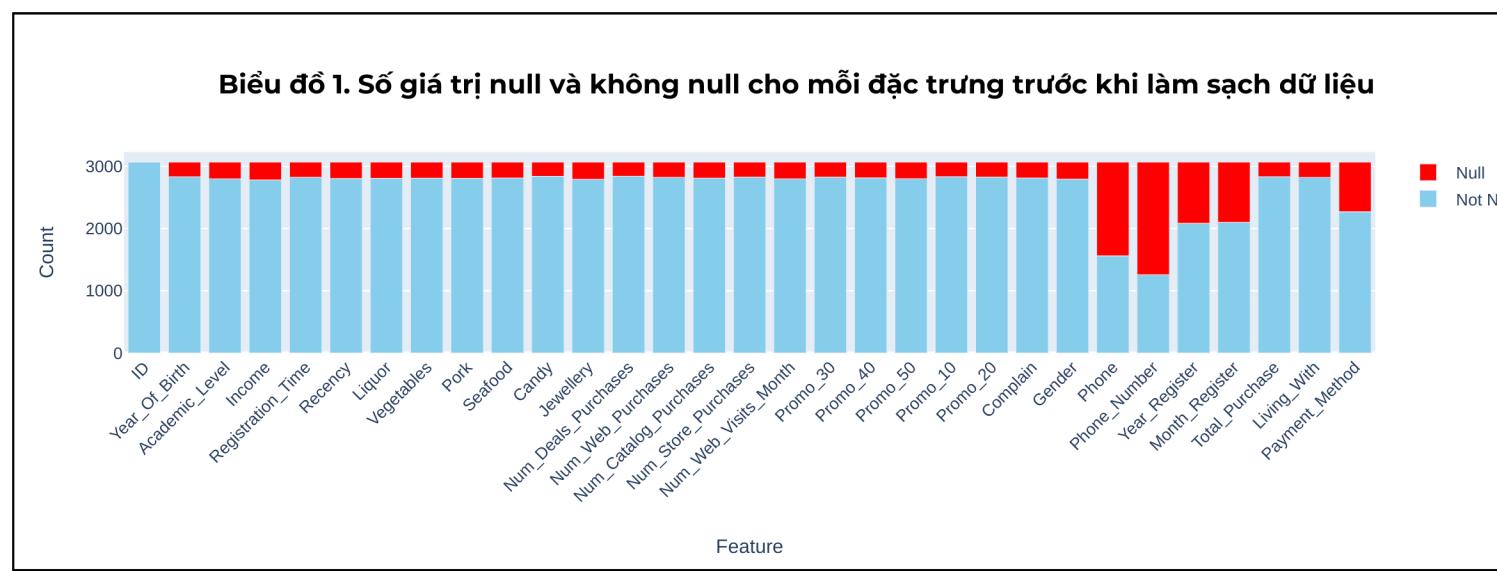
Gia Quyên - Lệ Ngọc - Hoa Viên



Nội dung



Bộ dữ liệu chứa thông tin tổng hợp của khách hàng được thu thập trong 2 năm, gồm thông tin cá nhân, chi tiêu và hành vi mua hàng.



	Trước	Sau
Số quan sát	3069	2211
Số đặc trưng	31	35

Mô tả dữ liệu sau khi được xử lý

#	Column	Non-Null Count	Dtype
0	ID	2211 non-null	int64
1	Year_of_Birth	2211 non-null	float64
2	Academic_Level	2211 non-null	category
3	Income	2211 non-null	float64
4	Registration_Time	2211 non-null	datetime64[ns]
5	Recency	2211 non-null	float64
6	Liquor	2211 non-null	float64
7	Vegetables	2211 non-null	float64
8	Pork	2211 non-null	float64
9	Seafood	2211 non-null	float64
10	Candy	2211 non-null	float64
11	Jewellery	2211 non-null	float64
12	Num_Deals_Purchases	2211 non-null	float64
13	Num_Web_Purchases	2211 non-null	float64
14	Num_Catalog_Purchases	2211 non-null	float64
15	Num_Store_Purchases	2211 non-null	float64
16	Num_Web_Visits_Month	2211 non-null	float64
17	Promo_30	2211 non-null	category
18	Promo_40	2211 non-null	category
19	Promo_50	2211 non-null	category
20	Promo_10	2211 non-null	category
21	Promo_20	2211 non-null	category
22	Complain	2211 non-null	float64
23	Gender	2211 non-null	category
24	Year_Register	2211 non-null	int64
25	Month_Register	2211 non-null	int64
26	Total_Purchase	2211 non-null	float64
27	Living_With	2211 non-null	category
28	Payment_Method	2211 non-null	category
29	Phone_num	2211 non-null	string
30	Age	2211 non-null	int64
31	Monetary	2211 non-null	float64
32	Marriage_Status	2211 non-null	category
33	Number_child	2211 non-null	int64
34	Age_quartile	2211 non-null	category

Tổng lượt mua
32813.0

Tổng chi tiêu
1340789.0

Tổng số khách hàng
2211

% khách hàng phàn nàn
0,95%

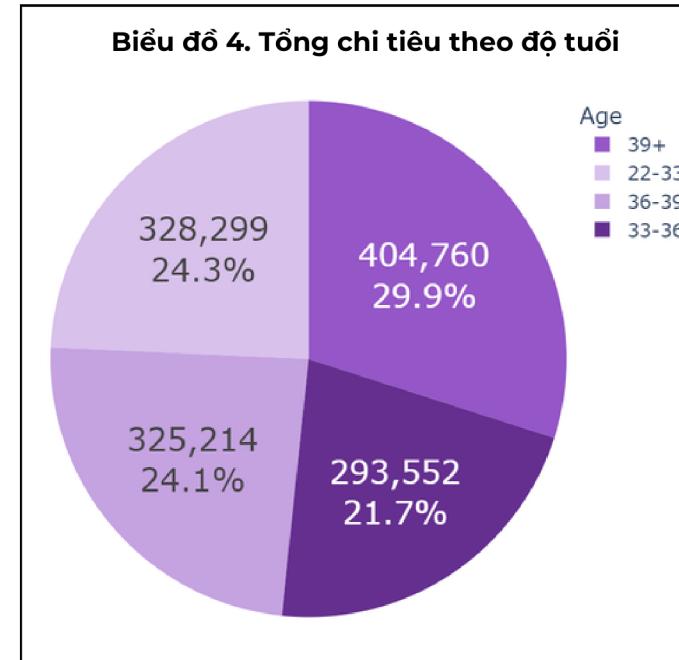
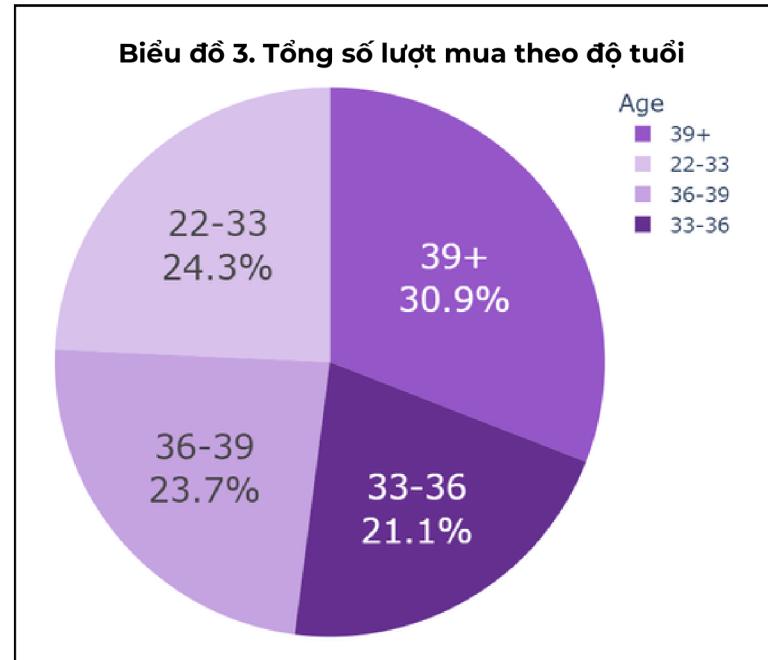
% khách hàng sử dụng giảm giá
98,1%

% khách hàng truy cập vào web trong tháng qua
99,6%

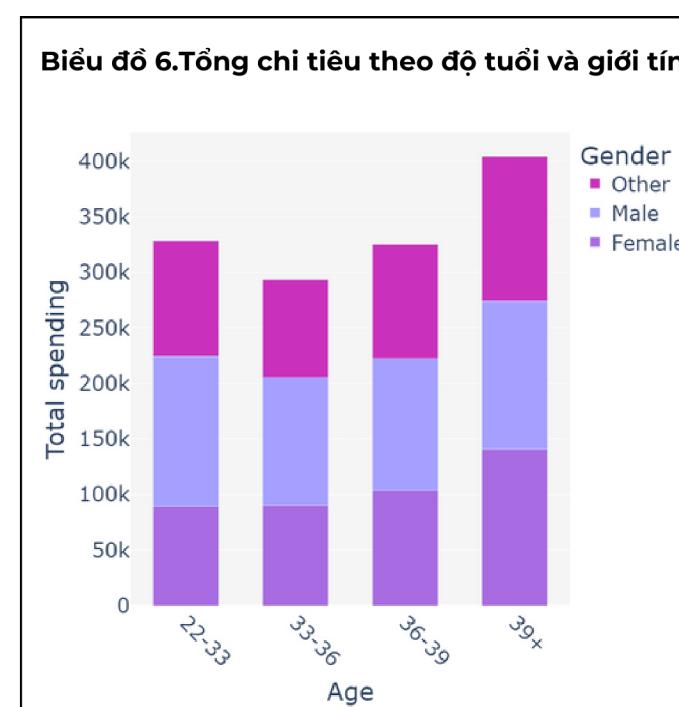
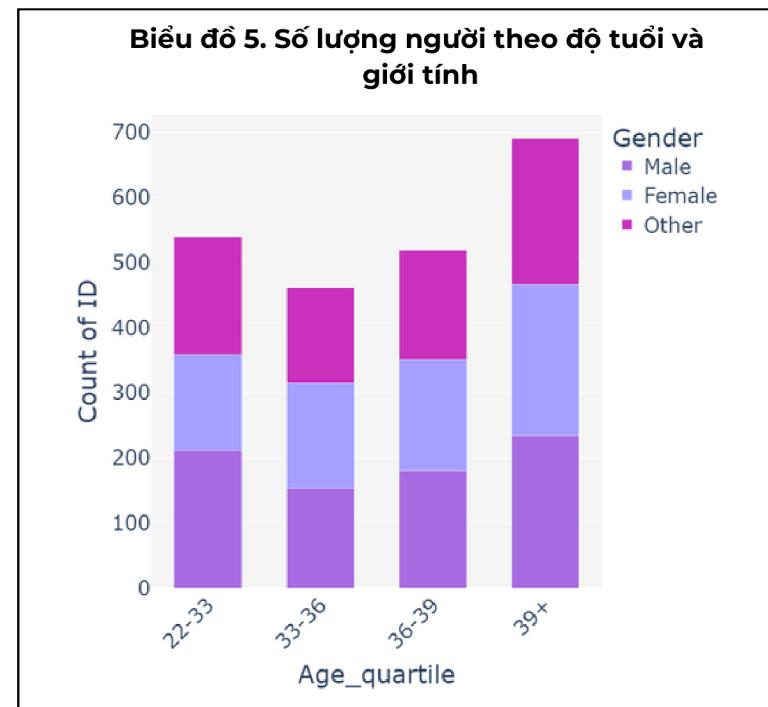
Theo tuổi và giới tính

Phân tích nhân khẩu học

Khách hàng chủ yếu ở độ tuổi trung niên trẻ (30-40 tuổi), không có sự phân hóa rõ ràng về giới tính.



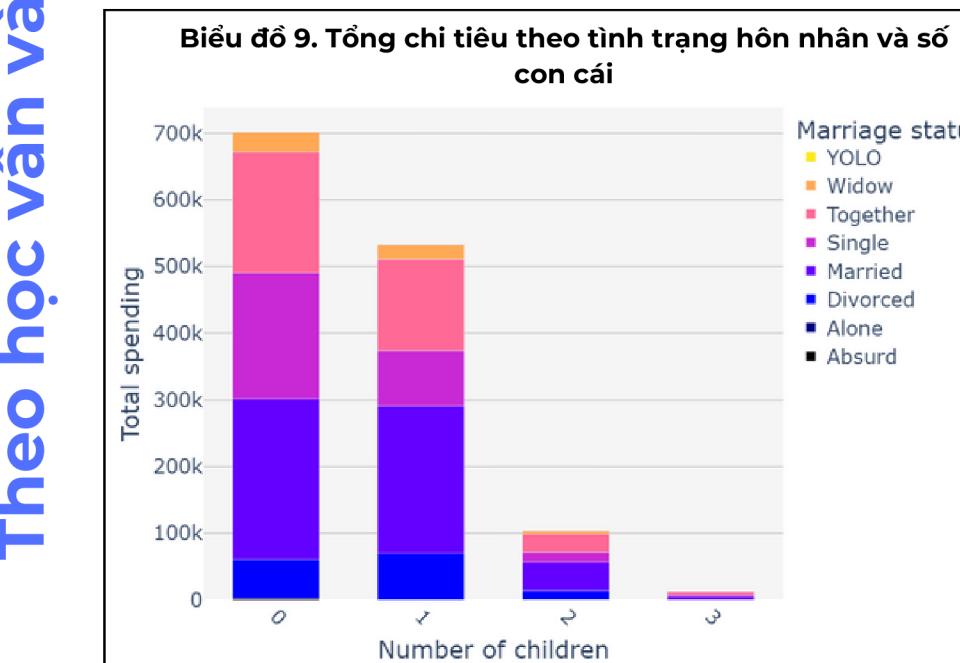
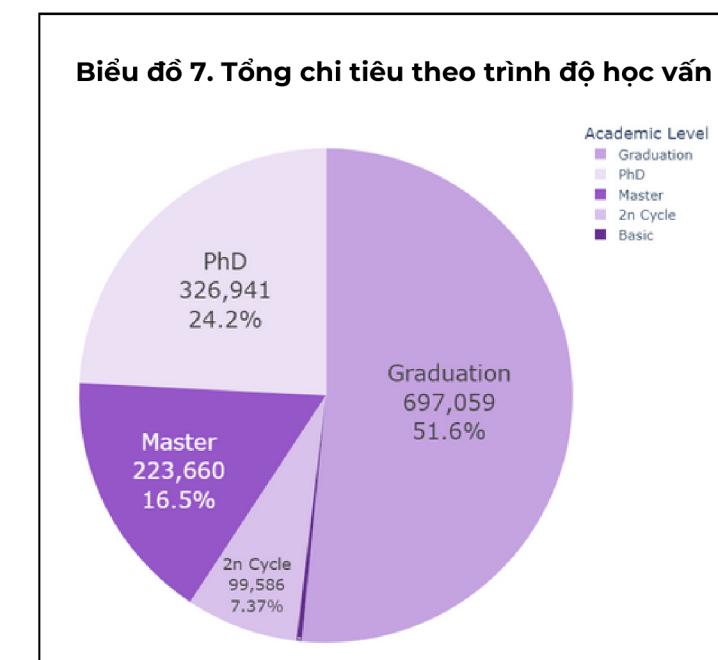
Không có sự khác biệt giữa số lượt mua và tổng chi tiêu theo độ tuổi, trong đó nhóm 39+ có sự đóng góp nhiều hơn.



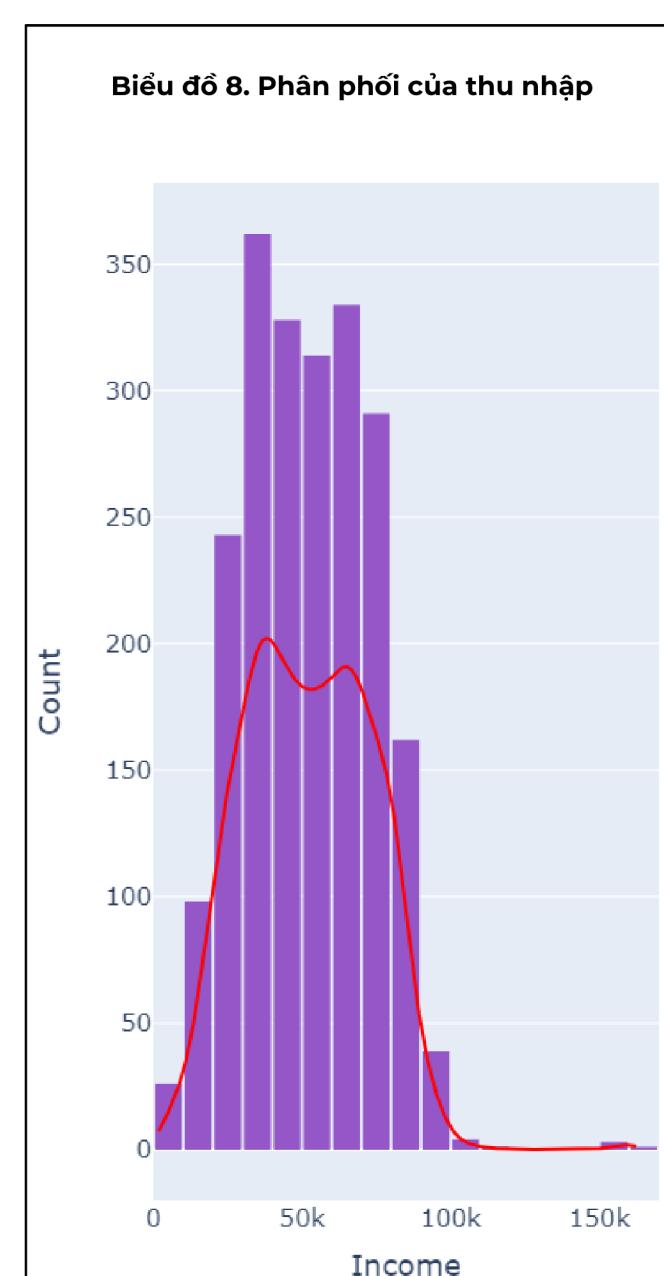
Giữa số khách hàng và tổng chi tiêu không có sự phân hóa rõ ràng theo giới tính.

Theo học vấn và hôn nhân

Khách hàng là người đã có gia đình, có trình độ học vấn cao và mức thu nhập từ trung bình khá trở lên.

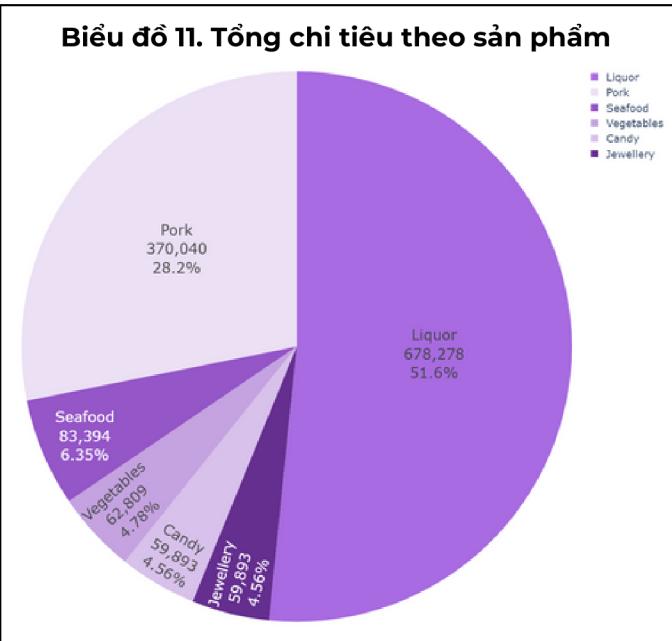
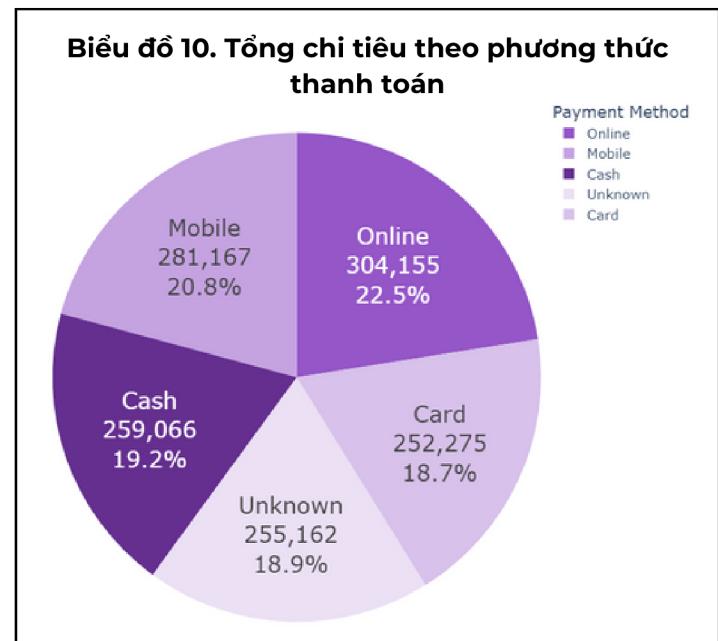


Hơn 90% khách hàng có trình độ đã tốt nghiệp, trong đó có sự tương quan cao giữa trình độ học vấn và thu nhập. Mức thu nhập dao động ở khoảng 30k-70k USD. Thường là người đã có gia đình và có 1 con, tuy nhiên, những cặp đôi chưa có con có xu hướng chi tiêu nhiều hơn.

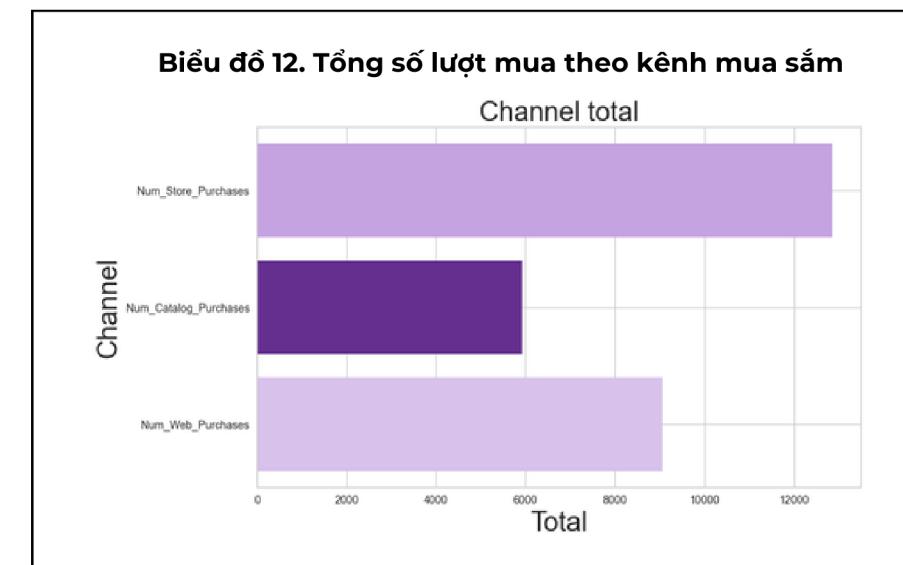


Phân tích hành vi mua hàng

Khách hàng thích thanh toán qua Online và Mobile hơn là trực tiếp, trong đó, Rượu là sản phẩm mang lại nhiều doanh thu nhất qua kênh mua sắm trực tiếp.

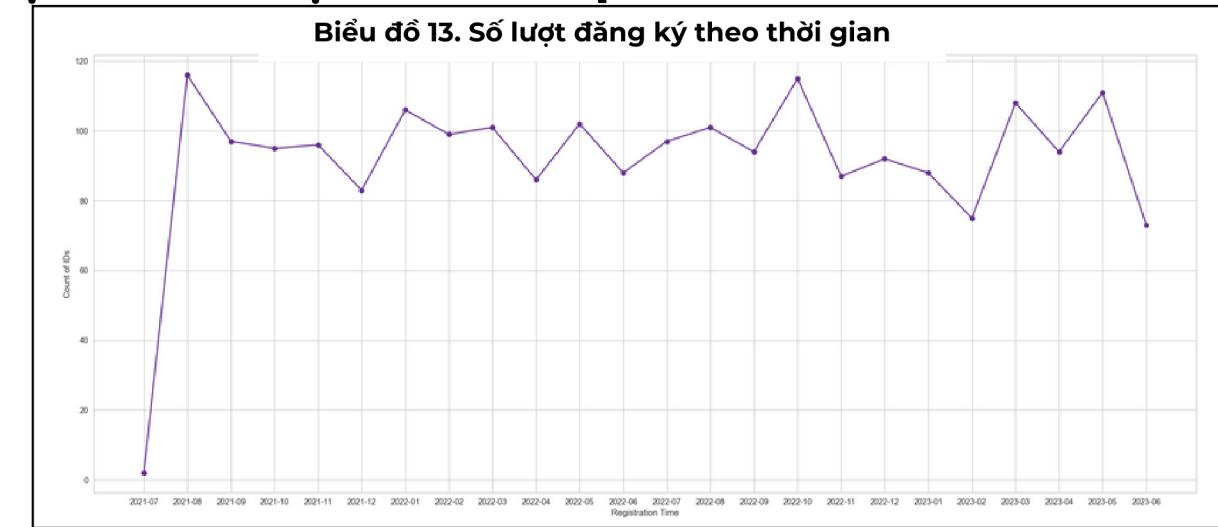


Không có sự chênh lệch lớn giữa các phương thức thanh toán, Online và Mobile chiếm gần 50%. Rượu là sản phẩm được chi tiêu nhiều nhất.

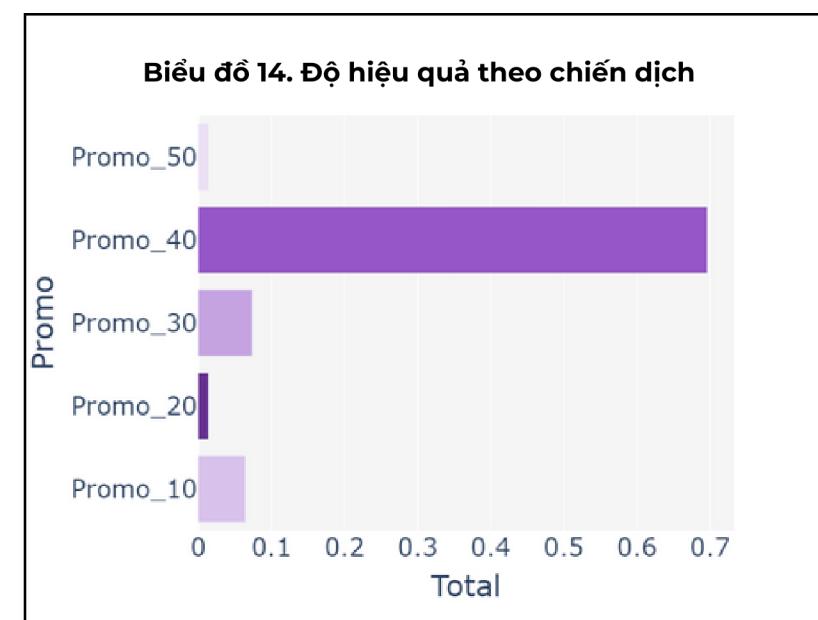


Kênh mua sắm trực tiếp được khách hàng ưu chuộng.

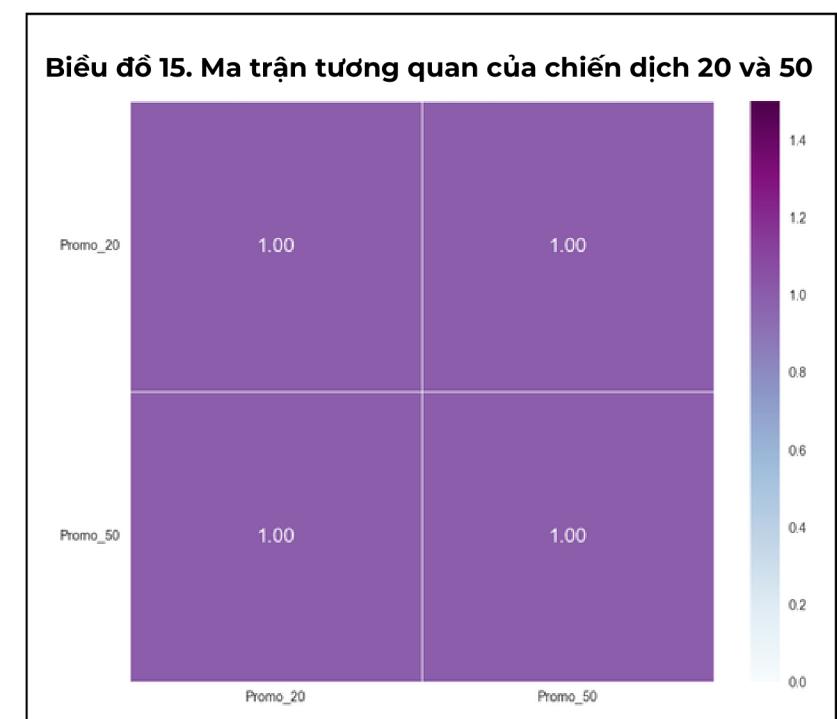
Các chiến dịch dường như chưa đem lại sự hiệu quả, hầu như tỷ lệ chấp nhận chiến dịch rất thấp.



Số lượt đăng ký trong 2 năm qua không có sự biến động mạnh, dao động từ 80 - 110 người qua các tháng.



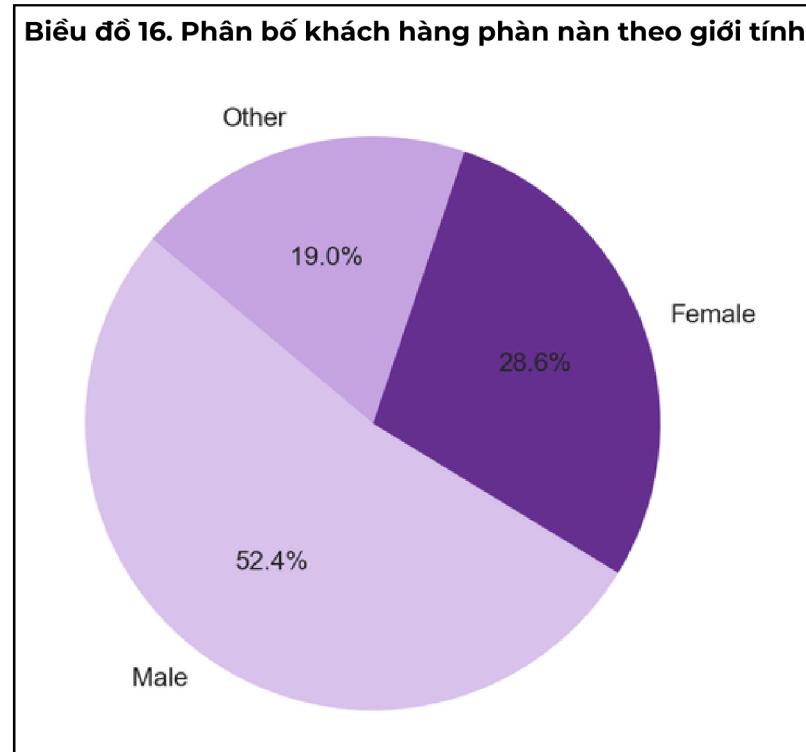
Chỉ có chiến dịch 40 đem lại sự hiệu quả, chiến dịch 20 và 50 có sự tương quan cao.
=> Nhìn chung, các chiến dịch đang không tiếp cận được đến khách hàng, điều này có thể đến từ nhiều nguyên nhân như khách hàng chưa tiếp cận được chiến dịch - phương tiện thông tin không hiệu quả (không có thông tin), hoặc đã tiếp cận nhưng không bị thu hút bởi thông điệp của chiến dịch.



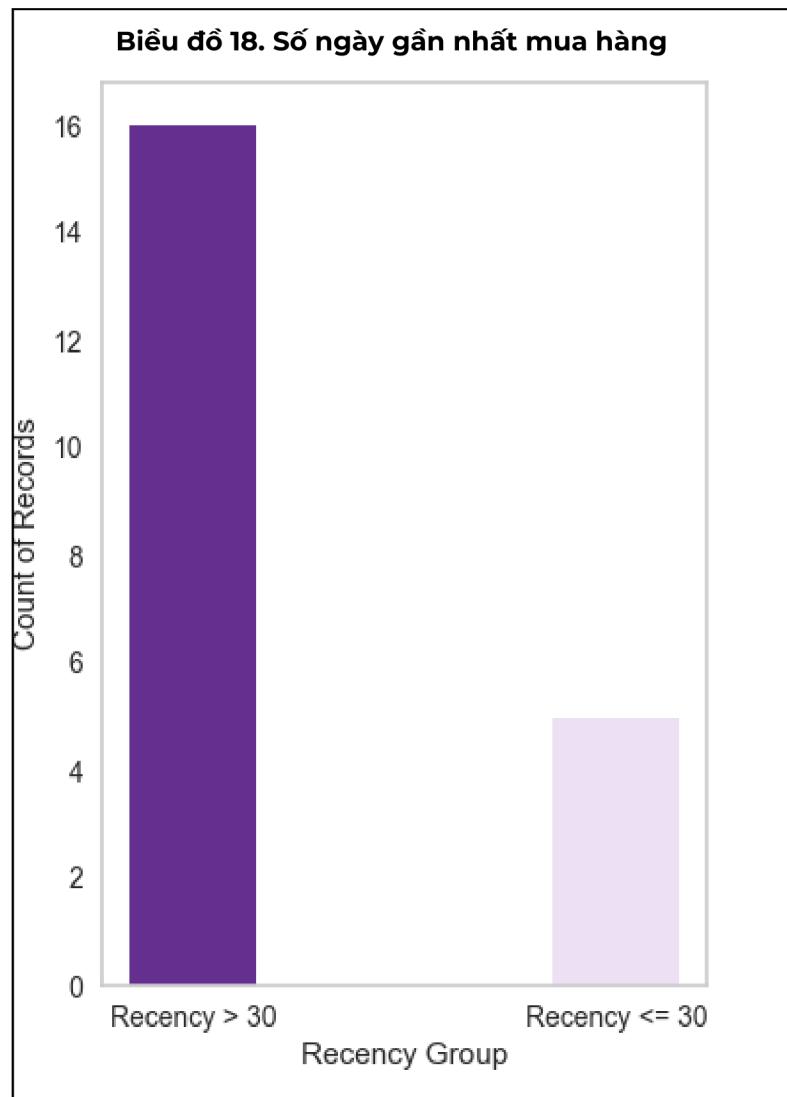
Phân tích tệp khách hàng chưa hài lòng với doanh nghiệp

Nhìn chung, hầu hết khách hàng đều hài lòng với doanh nghiệp, tuy nhiên có 0,95% khách hàng đã liên hệ để phàn nàn.

Chủ yếu là nam, đã có gia đình và 1 con.



Hầu hết đây là những khách hàng đã gắn bó với doanh nghiệp trên 1 năm, tuy nhiên đa số họ đã không quay trở lại mua sắm trong 30 ngày qua.



	Số tháng gắn bó	Số deal đã mua
count	21	21
mean	23.1	2.33
std	6.28	1.43
min	12	1
25%	19	1
50%	26	2
75%	28	3
max	32	7

Nhìn chung, tỷ lệ khách hàng chưa hài lòng hiện tại là không đáng kể, do đó doanh nghiệp cần tiếp tục duy trì để số lượng này không tăng lên.

Quy trình thực hiện

**DATASET ĐÃ
ĐƯỢC XỬ LÝ**



CHUẨN BỊ DỮ LIỆU

Monetary

Frequency

Recency

DATA SCALING

R, F, M values

Standard Scaler

CHUẨN HÓA DỮ LIỆU

DISTRIBUTION TRANSFORMATION

Monetary

Frequency

Box-Cox transformation

Cubic transformation

DROP OUTLIERS

Monetary

Frequency

MÔ HÌNH PHÂN CỤM

K-means

BIRCH

Hierarchical

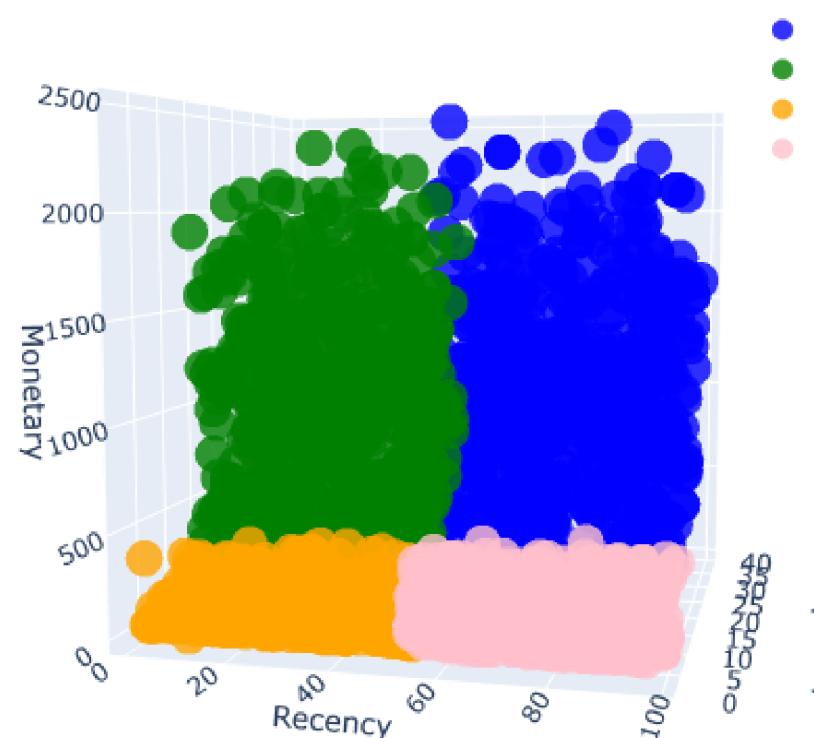
Spectral

Gaussian mixture

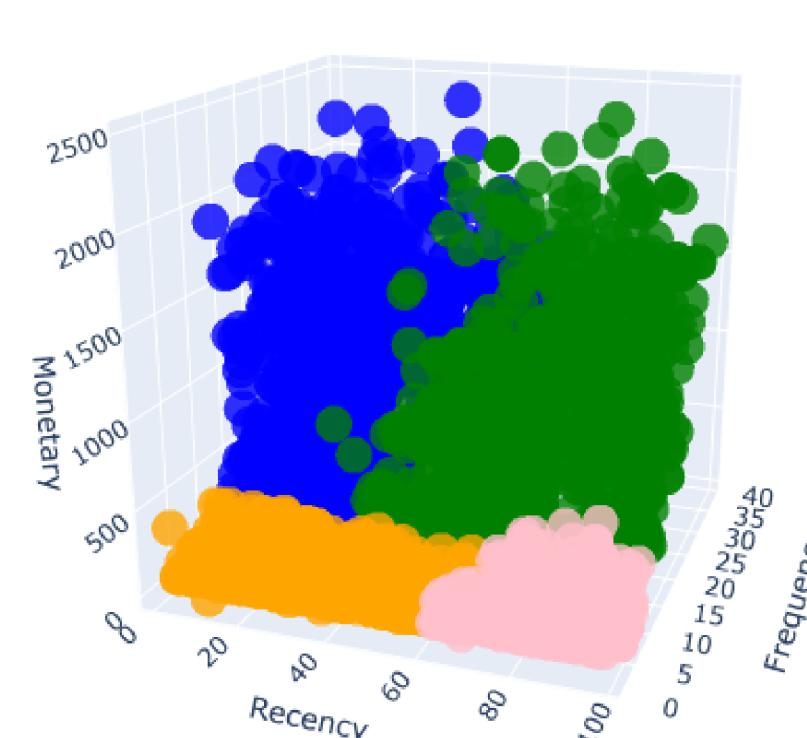
Xác định số cụm k
tối ưu

Một số mô hình phân cụm

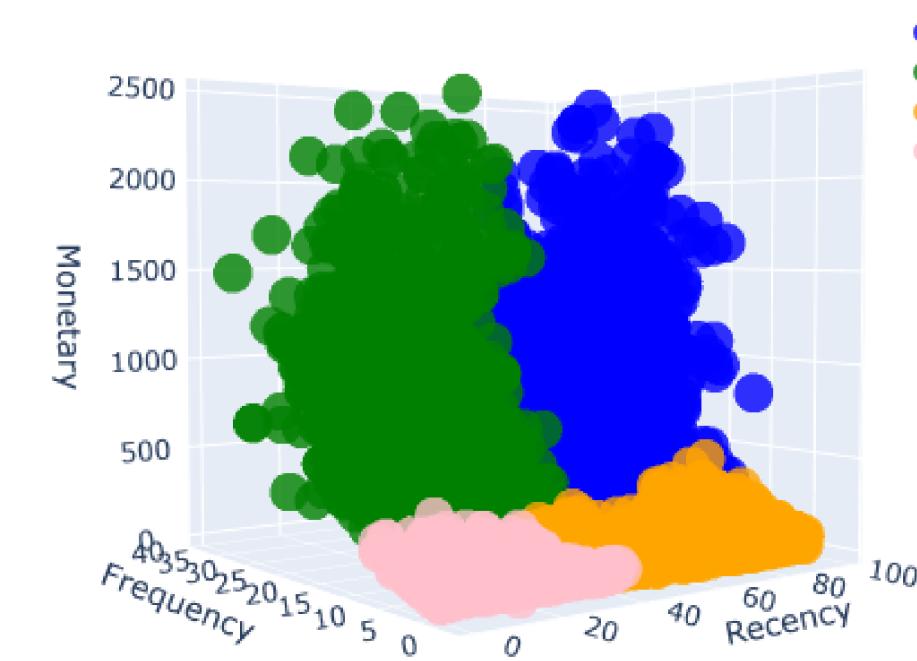
K-means



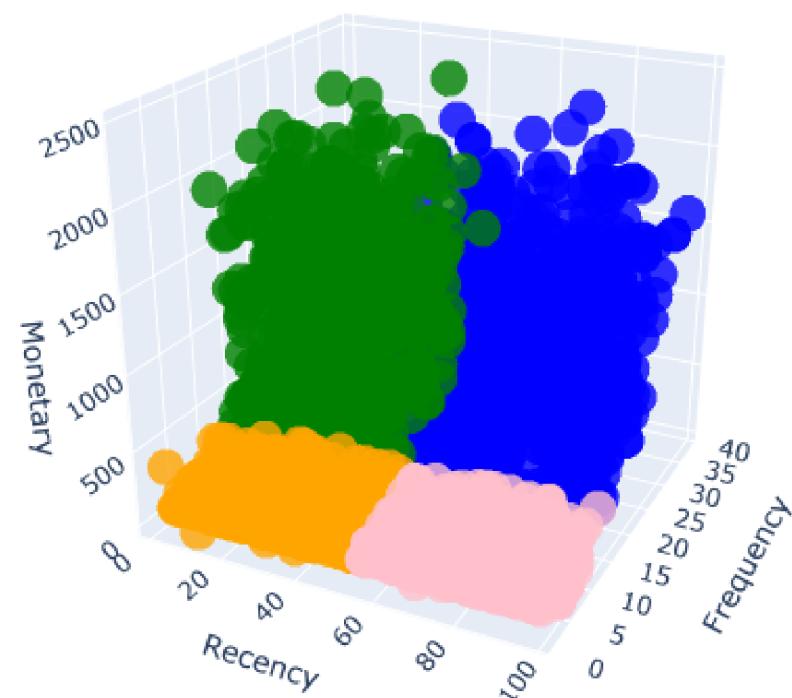
BIRCH



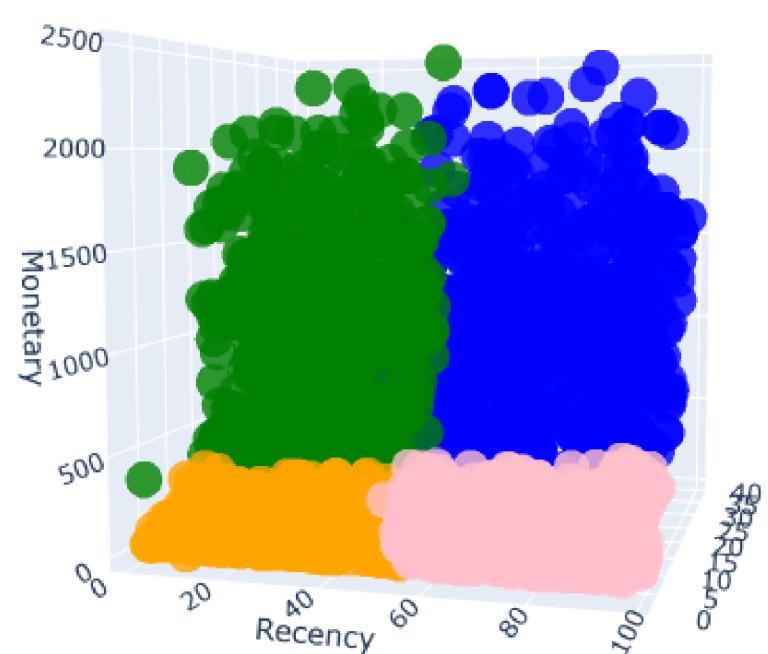
Hierarchical



Spectral



Gaussian Mixture



So sánh dùng độ đo silhouette

Thuật toán	Điểm silhoutte
K-means	0.400763
BIRCH	0.364869
Hierarchical	0.380105
Spectral	0.400161
Gaussian Mixture	0.393695

So sánh kết quả phân cụm

GAUSSIAN MIXTURE MODEL

	R	F	M	% by M
Cluster 0	26.26	7.35	80.60	3.54%
Cluster 1	21.10	20.55	993.99	43.62
Cluster 2	70.94	21.10	1086.37	47.67
Cluster 3	76.04	8.47	117.84	5.17

HIERACHICAL MODEL

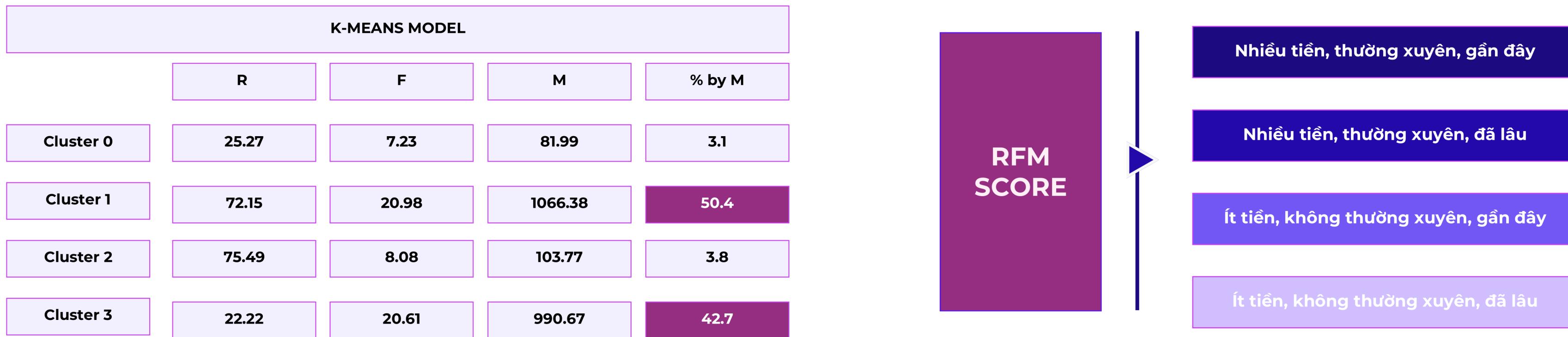
	R	F	M	% by M
Cluster 0	23.32	20.96	992.37	45.74
Cluster 1	70.14	7.67	90.10	4.15
Cluster 2	72.01	20.15	1011.78	46.64
Cluster 3	19.74	6.93	75.28	3.47

BIRCH MODEL

	R	F	M	% by M
Cluster 0	71.89	19.46	921.91	42.86
Cluster 1	29.55	7.14	75.79	3.52
Cluster 2	26.08	21.39	1062.01	49.37
Cluster 3	81.06	7.49	91.26	4.24

SPECTRAL MODEL

	R	F	M	% by M
Cluster 0	22.13	20.70	1003.69	45.43
Cluster 1	76.19	7.65	92.32	4.18
Cluster 2	72.24	20.67	1028.50	46.55
Cluster 3	25.86	7.37	84.97	3.85

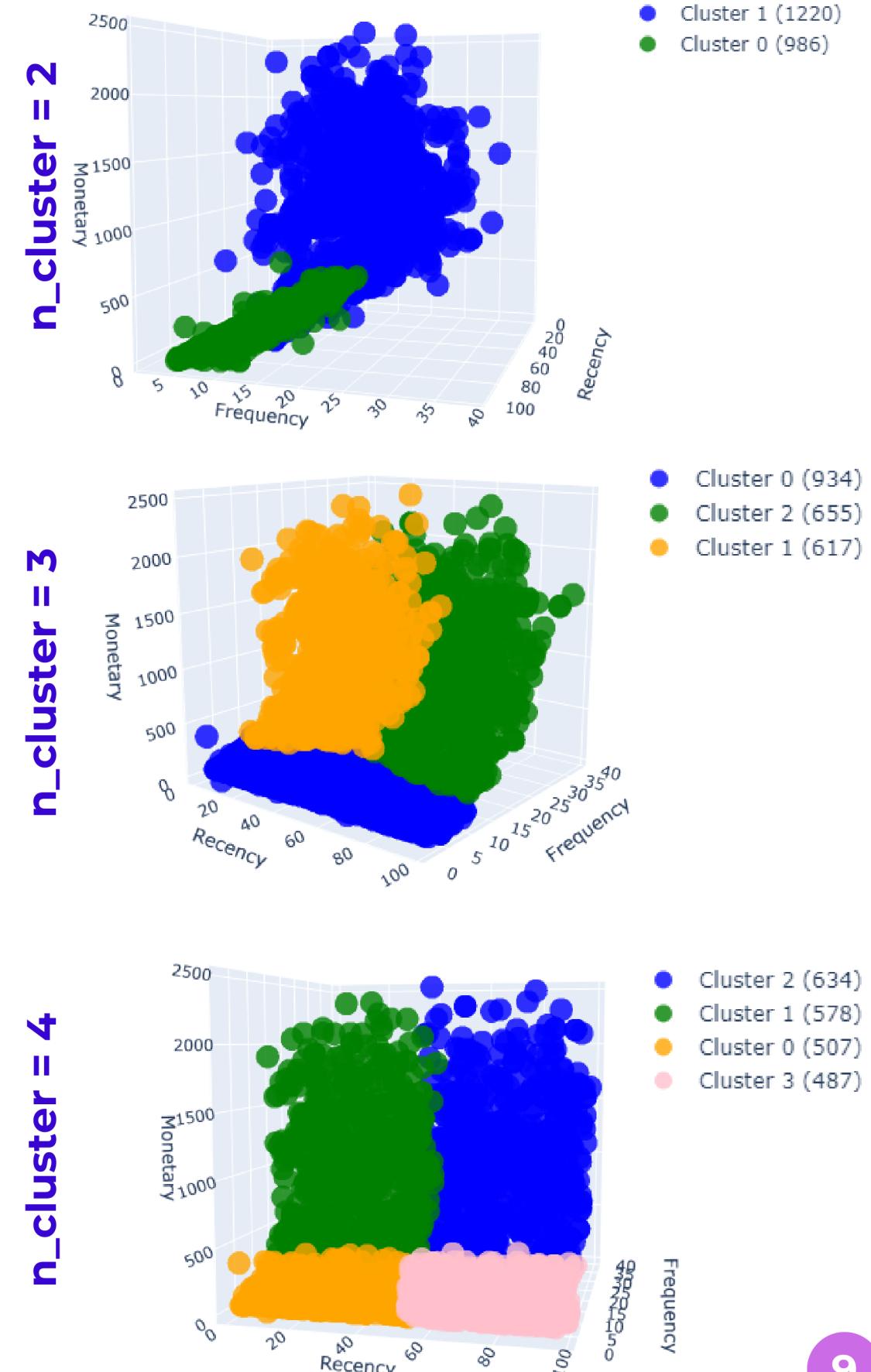
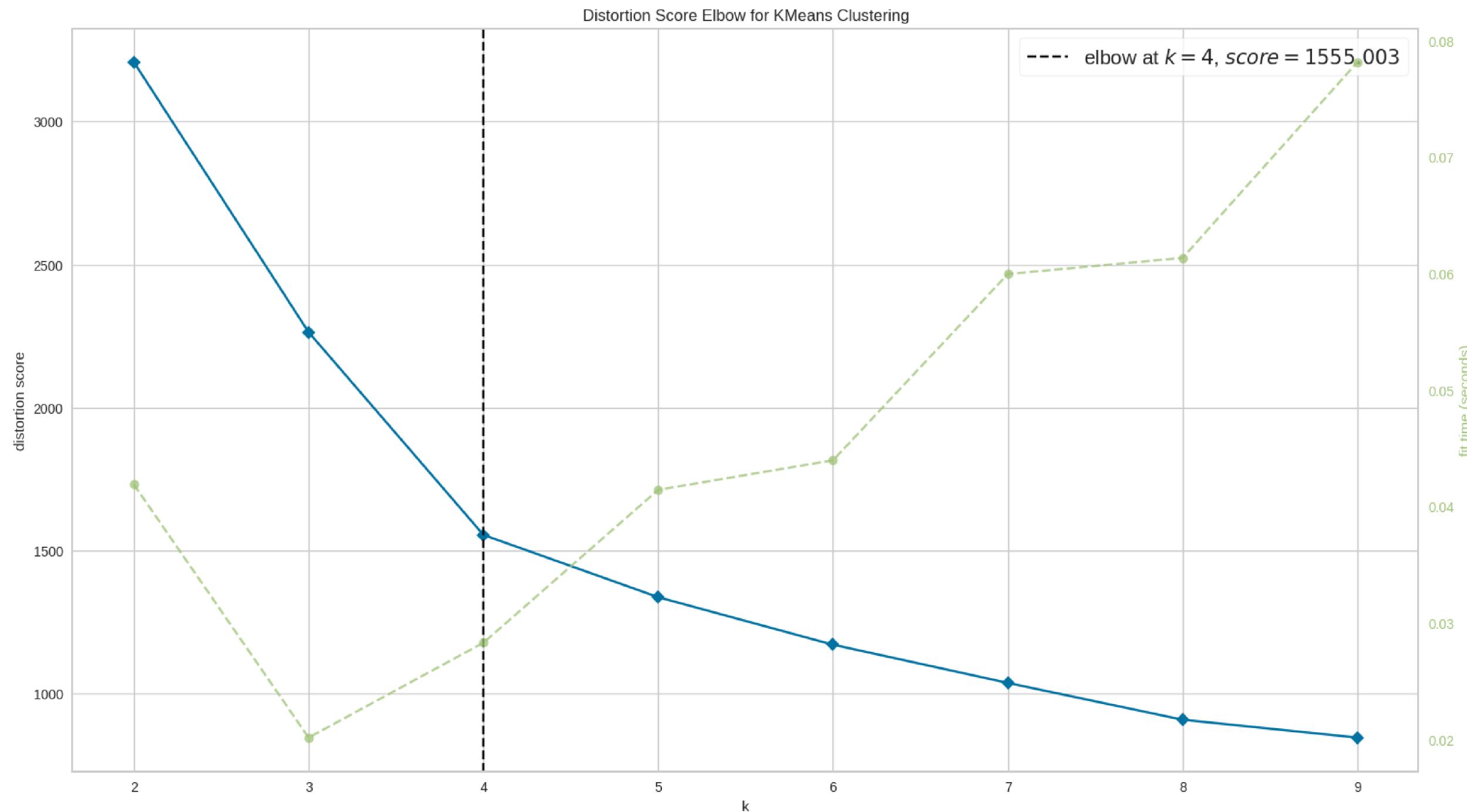


Kết quả phân cụm RFM

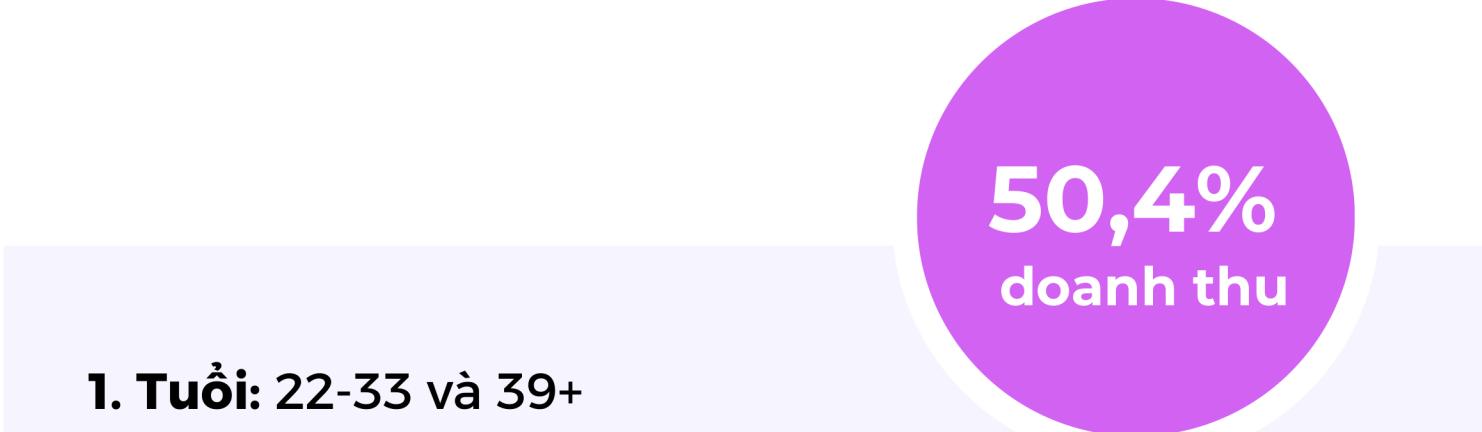
Kết quả phân cụm RFM giữa các mô hình học máy cho thấy không có sự khác biệt quá lớn giữa kết quả và khoảng cách các cụm, các mô hình cho điểm số silhouette khá đồng đều giữa các cụm. Đối với bộ dữ liệu được cho, **K-Means** được xem là mang lại hiệu quả tốt nhất giữa các phương pháp.

So sánh chỉ số các cụm dựa trên RFM score, các phương pháp đều phân thành 4 cụm với các đặc điểm RFM giống nhau ở mỗi cụm. Theo nguyên tắc Pareto, 20% khách hàng của cửa hàng mang lại 80% doanh thu cho cửa hàng. Cụ thể, ở đây, tệp khách hàng tiềm năng cho cửa hàng là cụm "**Nhiều tiền, thường xuyên, gần đây**" - **Khách hàng tiềm năng hiện tại** và cụm "**Nhiều tiền, thường xuyên, đã lâu**" - **Khách hàng tiềm năng trong quá khứ**.

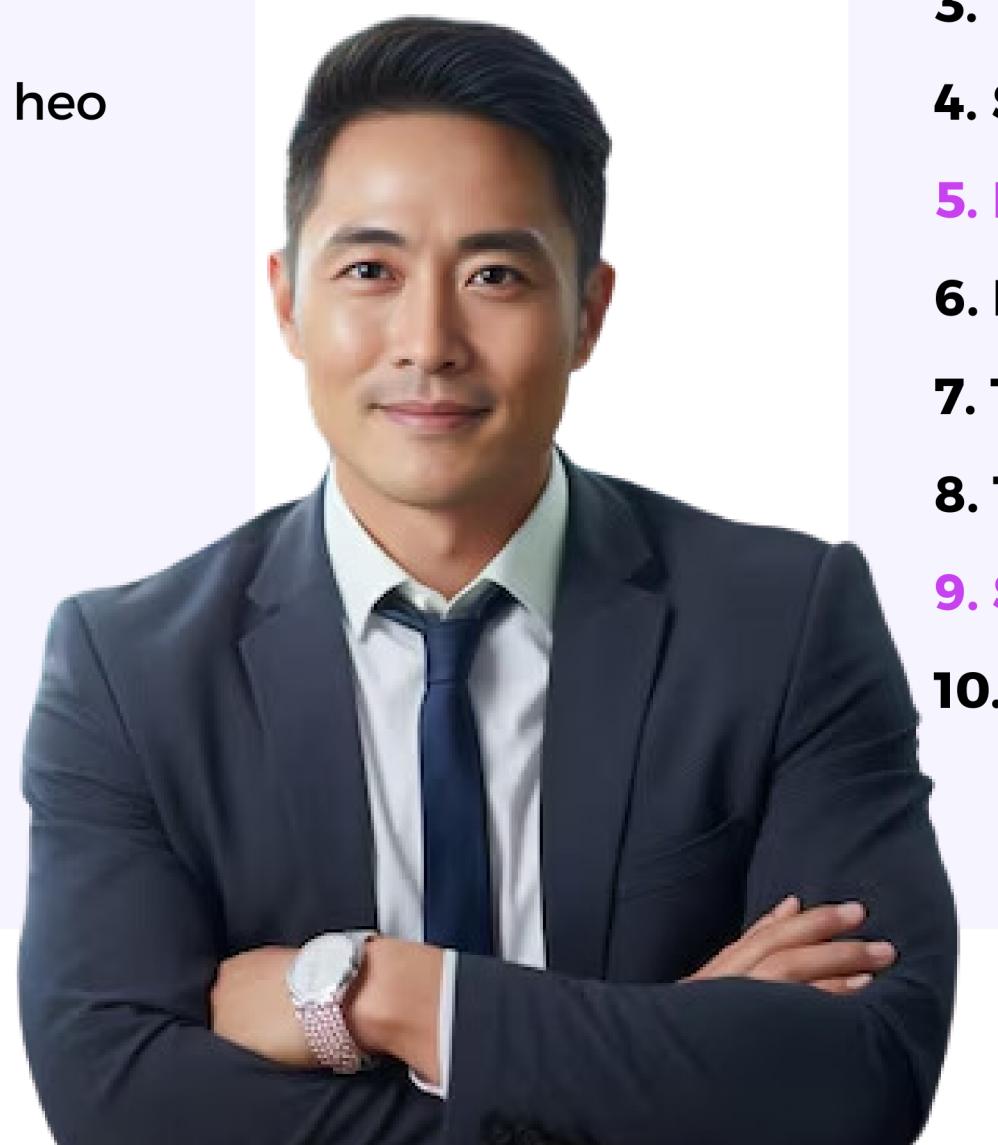
Thử nghiệm các hyperparameter trên thuật toán K-Means



Khách hàng tiềm năng trong quá khứ



- 1. Tuổi:** 22-33 và 39+
- 2. Giới tính:** nam (36,8%)
- 3. Thu nhập:** 70000-85000 USD
- 4. Sản phẩm chi tiêu nhiều nhất:** Rượu và thịt heo
- 5. Phương thức thanh toán:** Online và card
- 6. Kênh mua sắm:** cửa hàng trực tiếp
- 7. Trình độ học vấn:** đã tốt nghiệp
- 8. Tình trạng hôn nhân:** đã kết hôn
- 9. Số con cái:** 0 và 1 con
- 10. Giá trị trung bình đơn hàng:** 51 USD



Khách hàng tiềm năng ở hiện tại



- 1. Tuổi:** 36+
- 2. Giới tính:** nam (37,8%)
- 3. Thu nhập:** 70000-85000 USD,
- 4. Sản phẩm chi tiêu nhiều nhất:** Rượu và thịt heo
- 5. Phương thức thanh toán:** Online và mobile
- 6. Kênh mua sắm:** cửa hàng trực tiếp
- 7. Trình độ học vấn:** đã tốt nghiệp
- 8. Tình trạng hôn nhân:** đã kết hôn
- 9. Số con cái:** 1 con
- 10. Giá trị trung bình đơn hàng:** 48 USD

Painpoint

Mặc dù, tệp quá khứ đóng góp tới 50,4% trên tổng doanh thu 2 năm vừa qua, nhưng họ là những khách hàng đã quá lâu chưa mua sắm trở lại.

Giá trị trung bình đơn hàng giảm, tổng doanh thu của tệp quá khứ lớn hơn tệp hiện tại.

Chạy rất nhiều chiến lược nhưng hầu như chưa mang lại sự hiệu quả.

Vấn đề

Các chiến lược marketing và chăm sóc khách hàng để giữ chân khách hàng chưa hiệu quả.

Hoạt động bán hàng của doanh nghiệp không đạt được hiệu quả, dẫn đến sự giảm giá trị chi tiêu của khách hàng.

Chưa tiếp cận đến được khách hàng tiềm năng

Câu hỏi

Làm thế nào để giữ chân, xác định được giá trị khách hàng và tiếp cận đến đúng đối tượng để thúc đẩy mua hàng?

Đề xuất

Sử dụng các **mô hình học máy phân loại nhị phân** để dự đoán xác suất khách hàng sẽ rời bỏ, từ đó có những chiến lược phù hợp để giữ chân khách hàng.

Sử dụng **mô hình thống kê Gamma-Gamma** để ước tính giá trị của khách hàng, từ đó xác định đúng tệp khách hàng cần chú trọng để đưa ra các chiến lược gia tăng tiêu dùng.

Sử dụng **mô hình A/B testing** để giảm rủi ro và tối ưu hóa tỷ lệ chuyển đổi, từ đó giúp doanh nghiệp ra quyết định.

Cuộc thi phân tích dữ liệu 2024

**THANK YOU
FOR YOUR LISTENING!**



3G

Gia Quyên - Lệ Ngọc - Hoa Viên