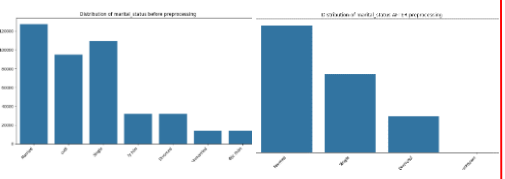
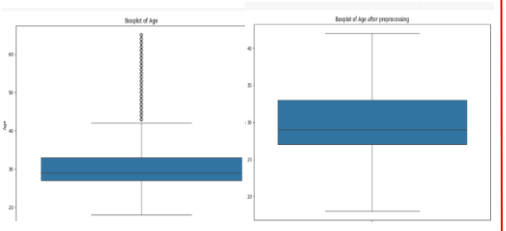
	user_id	age	sex	phone	job
count	424170.000000	328803.000000	418652	402962	402962
unique	NaN	NaN	40	30000	53
top	NaN	NaN	F	******4312	Chuyên viên nhân sự
freq	NaN	NaN	47868	39	7798
mean	212085.500000	32.064674	NaN	NaN	NaN
std	122447.476178	32.842554	NaN	NaN	NaN
min	1.000000	-1.000000	NaN	NaN	NaN
25%	106043.250000	26.000000	NaN	NaN	NaN
50%	212085.500000	29.000000	NaN	NaN	NaN
75%	318127.750000	35.000000	NaN	NaN	NaN
max	424170.000000	999.000000	NaN	NaN	NaN

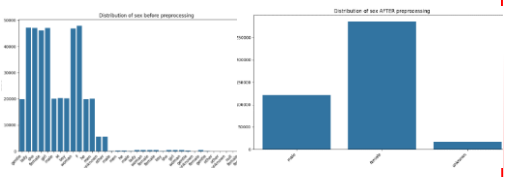
Hình 1. Thống kê mô tả bảng user_info trước tiền xử lí



Hình 2. Thống kê mô tả 'age' trước và sau tiền xử lí



Hình 3. Thống kê mô tả 'sex' trước và sau tiền xử lí



Hình 4. Thống kê mô tả 'marital status' trước và sau tiền xử lí

```
Data columns (total 5 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   user_id         424170 non-null  int64
 1   age             424170 non-null  float64
 2   sex             424170 non-null  object
 3   job             424170 non-null  object
 4   marital_status  424170 non-null  object
```

	user_id	age	sex	job	marital_status
count	424170.000000	424170.000000	424170	424170	424170
unique	NaN	NaN	3	54	4
top	NaN	NaN	female	unknown	Married
freq	NaN	NaN	285632	21208	222625
mean	212085.500000	30.032270	NaN	NaN	NaN
std	122447.476178	5.893977	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN
25%	106043.250000	27.000000	NaN	NaN	NaN
50%	212085.500000	29.000000	NaN	NaN	NaN
75%	318127.750000	33.000000	NaN	NaN	NaN
max	424170.000000	42.000000	NaN	NaN	NaN

Hình 5. Bảng user_info sau tiền xử lí, không có giá trị null và đầy đủ dữ liệu

TIỀN XỬ LÍ DỮ LIỆU – DATA PREPROCESSING

Bộ dữ liệu có 4 bảng: user_info (424170 hàng × 7 cột), user_log (54925330 hàng × 7 cột), train (233782 hàng × 3 cột) và test (52695 hàng × 3 cột).

1.1. Hiểu dữ liệu và Xử lí bảng user_info:

- **Cột Age:** 95,367 giá trị bị thiếu (~22%) và nhiều giá trị ngoại lai.

Phương pháp:

+ Xử lí outlier: xác định IQR (Interquartile Range), cắt các giá trị cực trị về ngưỡng trên, dưới là 12, 48.
+ Điền giá trị bị thiếu bằng trung vị thay vì trung bình, vì tuổi phân bố đồng đều từ 25–45, ít giá trị ngoại lai. Quan trọng hơn, trung vị ít bị ảnh hưởng bởi giá trị ngoại lai.

-Cột Sex: có rất nhiều giá trị không chuẩn như F, lady, girl, she, woman, female, M, boy, men, male, he,... và giá trị bị thiếu

Phương pháp:

+ Điền giá trị bị thiếu bằng 'unknown'.
+ Nhóm về 3 nhóm chính: male, female, unknown.

-Cột marital_status: có nhiều giá trị không đồng nhất (cưới, ly hôn, độc thân, Divorced, Unmarried,…), và giá trị bị thiếu.

Phương pháp:

- Nhóm về 3 nhóm chính:
  + **Married:** 'Married', 'cưới'.
  + **Divorced:** 'Divorced', 'ly hôn'.
  + **Single:** 'Single', 'Unmarried', 'độc thân'.
- Điền giá trị bị thiếu bằng 'unknown'.

-Cột Phone: số lượng giá trị bị thiếu là 21,208 (~5%), cột này *không đóng góp trực tiếp* trong quá trình phân tích

Phương pháp: **xem xét loại bỏ cột này**.

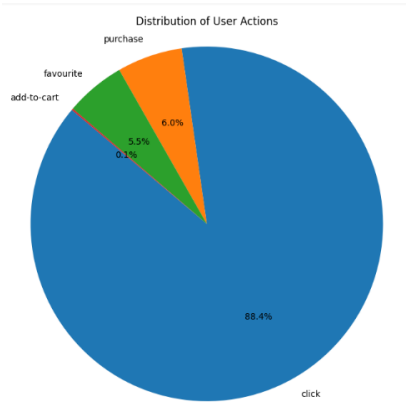
-Cột Job: số lượng giá trị bị thiếu là 21,208 (~5%) chiếm tỉ trọng không nhiều.

Phương pháp: Điền giá trị bị thiếu bằng 'unknown'.

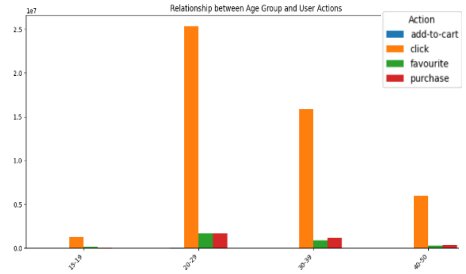
Cột Carrier: số lượng giá trị bị thiếu là 21,208 (~5%), cột này *không đóng góp trực tiếp* trong quá trình phân tích

Phương pháp: **xem xét loại bỏ cột này**.

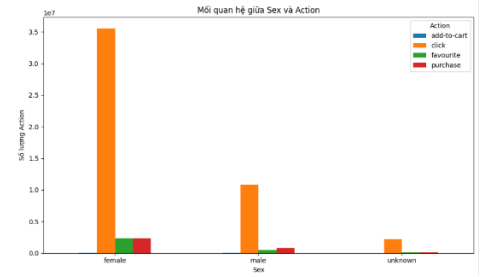
1.2. Xử lí bảng user_log: Tất cả các cột trong bảng user_log không có giá trị thiếu, ngoại trừ **brand_id.** Không xét giá trị ngoại lai vì hầu hết thuộc tính cột trong bảng là id, datetime.**- Cột brand_id:** dữ liệu bị thiếu là 91015 (~0.17%**)** chiếm tỉ trọng nhỏ.



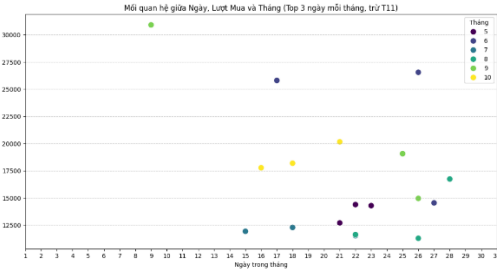
Hình 6. Tỉ trọng cột 'Action'



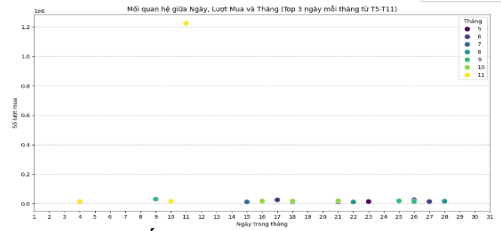
Hình 7. Số 'Action' với nhóm tuổi



Hình 8: Số 'Action' với giới tính



Hình 9. Số lượng mua hàng các tháng



Hình 10. Số lượng mua hàng các tháng

DATA VISUALIZATION

2.1. Phân tích sơ lược hành vi, nhóm khách hàng tiềm năng:

Hình 6: Mô tả tỉ lệ các loại hành động người dùng trên nền tảng TMĐT, cho thấy phần lớn người dùng hành động "click" chiếm **88.4%**, trong khi hành động "purchase" chỉ chiếm **6%** và "favourite" chiếm **5.5%**. Đặc biệt, hành động "add-to-cart" chiếm tỉ lệ rất nhỏ **0.1%**.

Insight: nền tảng hiện tại đang gặp vấn đề lớn bước chuyển đổi người dùng từ "click" sang "purchase". Để giải quyết, xây dựng các chiến dịch remarketing nhắm vào nhóm người dùng "favourite" và "add-to-cart", những người đã thể hiện sự quan tâm nhưng chưa hoàn tất giao dịch.

Hình 7: Dễ dàng nhận thấy nhóm tuổi **20-29** là nhóm hoạt động nhiều nhất, 'click' xuất hiện gấp **~10 lần** so với Action khác. Tương tự Nhóm tuổi còn lại cũng có đặc điểm lượt 'click' cao ngất ngưỡng.

Hình 8: Người dùng nữ (female) có số lượng hành động vượt trội hơn hẳn so với các giới tính khác, đặc biệt ở hành động "click".

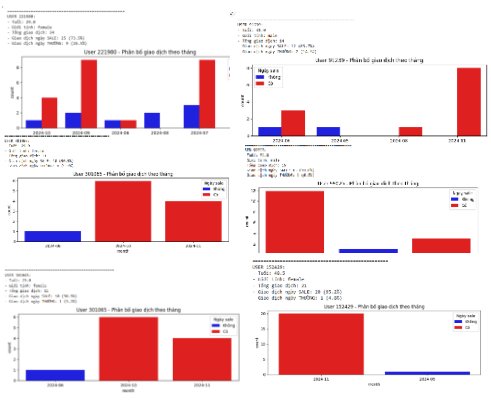
Insight: Khách hàng ở *bất kì độ tuổi nào, giới tính nào* cũng dành ra 10 lần click để quy đổi được 1 lần mua hàng hoặc 1 thêm vào yêu thích, điều này vì sở thích "dạo chơi" các gian hàng để có lựa chọn ưng ý nhất trước khi quyết định chốt đơn.
→ Nên phát triển hệ thống recommendation với sự lựa chọn đa dạng hơn để kéo dài thời gian dạo chơi, *vì khách hàng càng "lướt", càng "chốt" đơn!*

➔Xác định nhóm tuổi **20-29** và **giới tính Nữ** là **khách hàng tiềm năng nhất, có khả năng quay lại mua hàng cao nhất,** vì lượt click + purchase + favourite cao ➔ cần tập trung khai thác, vì họ vừa có khả năng chi trả, và nhu cầu mua sắm cao chiếm 40% tổng doanh số bán lẻ trực tuyến tại châu Á (McKinsey & Company, 2023)

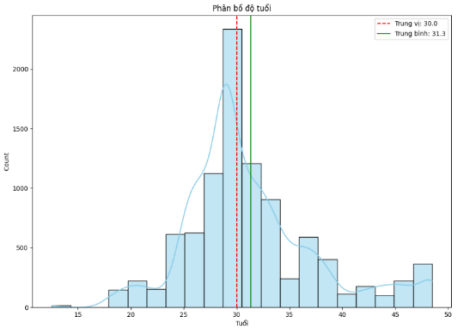
2.2. Phân tích sơ lược những hành vi mua hàng trong năm:

Hình 9: Ngày 9/9 có lượt mua đột biến, cao nhất toàn biểu đồ (sale ngày đôi 9/9). Cuối tháng (ngày 20-28) là thời điểm mua sắm sôi động nhất ở nhiều tháng, có thể do chu kỳ lương. Tháng 6 cũng rất mạnh, với các ngày mua cao điểm rơi vào giữa và cuối tháng. Mức độ "đỉnh" của top 3 ngày mua sắm rất khác nhau giữa các tháng, cho thấy ảnh hưởng của mùa vụ hoặc hiệu quả chiến dịch từng tháng.

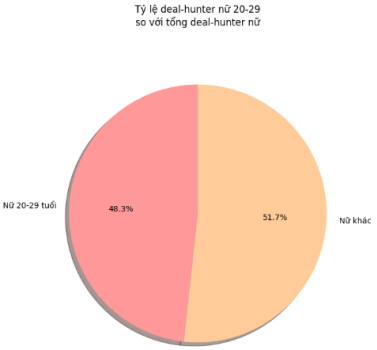
Hình 10: Ngày 11/11 là Đỉnh Điểm Mua Sắm Tuyệt Đối: Lượt mua ngày này vượt trội hoàn toàn, cho thấy sức ảnh hưởng khổng lồ của



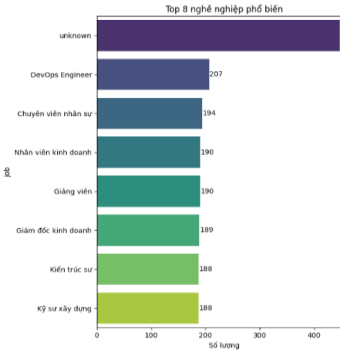
Hình 11. So sánh lượt mua hàng của Deal-hunter giả định trong ngày sale và không sale



Hình 12:Phân phối độ tuổi của deal-hunter



Hình 13: Tỉ lệ deal-hunter Nữ 20-29



Hình 14. Top 8 nghề nghiệp phổ biến của deal-hunter

các sự kiện siêu khuyến mãi (như Lễ Độc Thân). Sự khác biệt về số lượt mua giữa ngày cao điểm nhất của tháng 11 và các tháng khác là rất lớn (1,250,000 so với mức cao nhất khoảng 30,000 - 50,000 ở các tháng khác như tháng 9 hoặc 6)

➔ **Giả thiết:** Nếu những khách hàng tiềm năng quay lại cao, với điều kiện có phát sinh >= 10 giao dịch (tất cả tháng) và tỉ lệ giao dịch ngày sale phải ít nhất gấp đôi ngày thường mới, thì người đó là **Deal-hunter.**

Để kiểm chứng giả thiết **"Deal-hunters là những người khách hàng tiềm năng quay lại cao chỉ mua hàng vào dịp sale, ít mua vào ngày thường"**, nhóm phân tích đã lựa chọn ***ngẫu nhiên 6 user trong tập hơn 9000 khách hàng được gán nhãn deal-hunter*** (dựa trên hành vi mua hàng).

Hình 11: Cho thấy những người "được cho" là deal-hunter có số lượng giao dịch vào các ngày sale (cột đỏ) cao vượt trội hơn những tháng không có ngày sale (cột xanh).

→ Điều này có thể kết luận rằng giả thuyết một phần tin tưởng được.

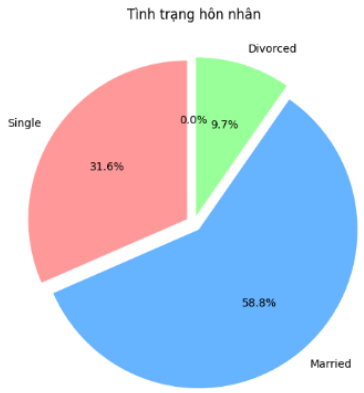
Hình 12: Số lượng deal-hunter đông nhất nằm ở nhóm tuổi ngay trước 30 (khoảng 27-29 tuổi), với một đỉnh rất rõ rệt. Deal-hunter chủ yếu là những người trưởng thành trẻ tuổi, tập trung mạnh nhất ở độ tuổi từ cuối 20 đến đầu 30

Hình 13: Phụ nữ trong độ tuổi 20-29 chiếm tới 35.1% tổng số người săn deal. Đây là một tỷ lệ rất đáng kể, cho thấy nhóm này đóng góp hơn một phần ba vào toàn bộ nhóm "deal-hunter".

⇨ **Kết luận 12 và 13 một lần nữa ủng hộ cho giả thiết Deal-hunter là sự kết hợp của Khách hàng tiềm năng (20-29 tuổi ở hình 12 và giới tính Nữ ở hình 13), cùng với tần suất mua hàng thấp vào ngày có sự kiện sale.**

Hình 14: Nghề nghiệp nổi bật nhất là DevOps Engineer, với số lượng vượt trội so với các ngành khác.Tiếp theo là các ngành có tính chất công việc văn phòng, thu nhập trung bình–cao và gắn với nhu cầu mua sắm online cao, như: Chuyên viên nhân sự,, Nhân viên kinh doanh,,…

Các ngành nghề này thường liên quan đến: Thói quen làm việc văn phòng, thường xuyên sử dụng internet, mua sắm online tiện lợi. Có thu nhập ổn định, nhu cầu chi tiêu cao, đặc biệt là vào các dịp sale lớn hoặc khuyến mãi nhỏ lẻ.Họ cũng là nhóm quen thuộc với việc săn deal, có kỹ năng tìm kiếm thông tin và lựa chọn sản phẩm phù hợp.



Hình 15: Phân phối Tình trạng hôn nhân của deal-hunter

Hình 15: Đa số người "săn deal" là những người đã lập gia đình. Điều này phản ánh rằng người đã kết hôn thường có nhiều khoản chi tiêu cố định và phát sinh (cho gia đình, con cái, nhu cầu sinh hoạt...). Họ có xu hướng tìm kiếm các ưu đãi, giảm giá để tiết kiệm chi phí.

- ➔ Hành vi "săn deal" có thể được xem là chiến lược chi tiêu hợp lý của nhóm khách hàng này.
- ➔ Các chiến dịch sale/khuyến mãi được thiết kế nhắm đến đối tượng người đã kết hôn, với các gói sản phẩm/dịch vụ phù hợp (đồ gia dụng, thực phẩm, đồ dùng gia đình, sản phẩm cho trẻ em...).

Nhóm Single (31.6%) cũng chiếm tỷ trọng đáng kể → đây có thể là nhóm trẻ tuổi, độc lập tài chính, thích săn sale để tận hưởng tiêu dùng nhưng không phải là nhóm chính.

- ➔ Cần cá nhân hóa thông điệp marketing cho nhóm này, nhấn mạnh vào lợi ích tiết kiệm và giá trị lâu dài khi mua sắm trong các dịp sale lớn.
- ➔ Thiết kế các gói ưu đãi phù hợp (combo thời trang, công nghệ, giải trí...) để thu hút họ trở thành loyal customer thay vì chỉ săn deal.

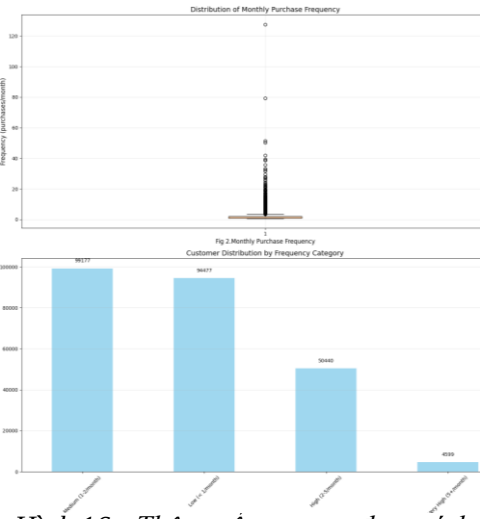
Nhóm Divorced có tỷ lệ thấp nhất (9.7%), cho thấy nhóm này có thể có nhu cầu chi tiêu thấp hơn hoặc ít tham gia săn deal so với các nhóm khác.

2.3. Phân tích cụ thể phương pháp tìm khách hàng trung thành:

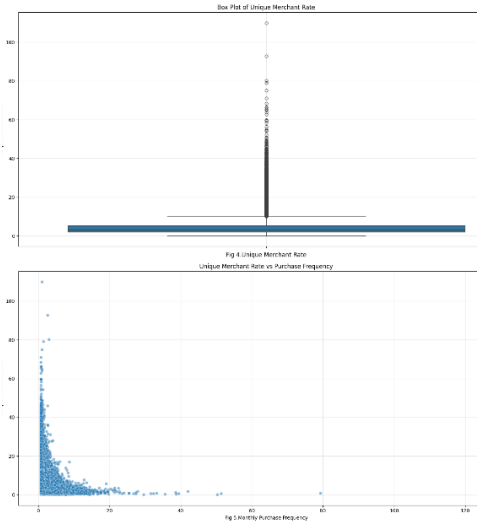
Sau khi đã xác định được nhóm deal-hunter, những người chỉ mua vào các dịp sale lớn, có hành vi săn giá rẻ– việc xác định khách hàng trung thành trở nên dễ dàng hơn, vì **loyal customer** thường có đặc điểm **ngược lại** với deal-hunter.

Engineering Features hữu ích trong việc xác định loyal customer có:

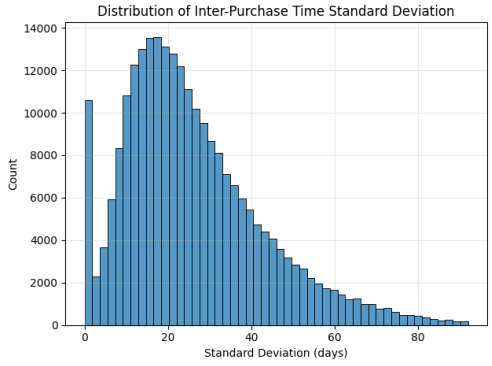
- Recency thấp , mới mua gần đây
- Frequency: Những khách hàng có tần suất mua hàng cao (>1 lần một tháng?), càng cao thì càng tốt
- Total purchase: càng cao càng tốt (mua nhiều), nhưng cũng phải trải đều (để tránh tránh deal-hunter)
- Unique merchant rate: là khi tỉ lệ này thấp nghĩa là họ sẽ không phải là các deal hunter (vì deal-hunter sẽ kiếm nhiều shop để kiếm giá thấp nhất), kết hợp với các yếu tố trên có thể giúp cải thiện hơn việc tìm ra loyal customer *(Corporate Finance Institute)*



Hình 16: Thông số recency: được tính từ ngày mua hàng cuối cùng cho đến cuối tháng của data (ở đây là ngày 30/11/2024)



Hình 17: Unique Merchant Rate và Unique Merchant



Hình 18: Thông số standard deviation between purchases

Hình 16: Biểu đồ trên cho thấy 100% những người dùng sau khi qua các tiêu chí lọc thì đều có Recency bằng nhau, điều này cho thấy 100% người dùng đều có ít nhất một hành động mua vào ngày 11/11/2024 (lí do là vì data chỉ thu thập đến ngày 11/11/2024). Qua biểu đồ trên, có thể nhận xét rằng sàn có một số lượng lớn người dùng hiện vẫn hoạt động (xấp xỉ 250 ngàn người).

Tuy vậy dữ liệu Recency này sẽ không đóng góp nhiều giá trị trong model đánh giá RMF do tất cả người dùng đều có điểm giống nhau.

Hình 17: Unique Merchant Rate và Unique Merchant: Qua 2 biểu đồ trên dễ có thể nhận thấy rằng phần lớn người mua hàng ở sàn hiện tại có xu hướng tìm nhiều mua hàng từ nhiều shop khác nhau, tuy nhiên những khách hàng có tiềm năng trung thành và khách hàng vip lại có xu hướng trung thành với một vài shop (tỉ lệ unique merchant rate thấp), điều này khớp với dự đoán về tệp khách hàng trung thành và khách hàng vip phía trên.

Hình 18: Thông số standard deviation between purchases: thông số này sẽ tính ra độ lệch chuẩn của khoảng thời gian giữa các lần mua (còn có thể hiểu là thói quen mua hàng của người dùng). Thông số này thấp sẽ cho thấy khách hàng có thói quen mua hàng đều đặn (chỉ thị của một khách hàng trung thành) và ngược lại khi nó cao cho thấy khách hàng mua hàng không thường xuyên, có thể chỉ mua khi có các đợt giảm giá lớn (chỉ thị của một deal-hunter).

Cho thấy một sự lệch về bên trái, cho thấy đa số khách hàng của sàn có thói quen mua hàng trong mỗi 15 - 30 ngày. Đây là một thông số không quá tệ, nó vẫn cho thấy rằng phần lớn khách hàng vẫn là một khách hàng thường xuyên của sàn, tuy nhiên biểu đồ cũng cho thấy khoảng ⅓ người dùng hiện tại không có thói quen mua sắm thường xuyên. Tuy nhiên để có thể chuyển đổi tệp khách hàng thường xuyên thành khách hàng trung thành, sàn cần bổ sung thêm các chiến lược để giảm số lượng ngày giãn cách giữa các lần mua của khách hàng xuống và tăng số lượng mua hàng lên. Có thể giảm các khuyến mãi lớn vào siêu sale, thay vì đó đưa ra các khuyến mãi nhỏ và vừa đều hơn để kích thích cảm giác mua hàng của khách hàng.

References:

(1) Corporate Finance Institute. (n.d.). Share of Wallet (SOW). Retrieved from https://corporatefinanceinstitute.com/resources/wealth-management/share-of-wallet-sow/
(2) Platzer, M., & Reutterer, T. (2016). Ticking Away the Moments: Timing Regularity Helps to Better Predict Customer Activity. Marketing Science, 35(5), 779–799.
(3) McKinsey & Company (2023) *The State of Fashion 2023*. Available at: https://www.mckinsey.com/industries/retail/our-insights/state-of-fashion