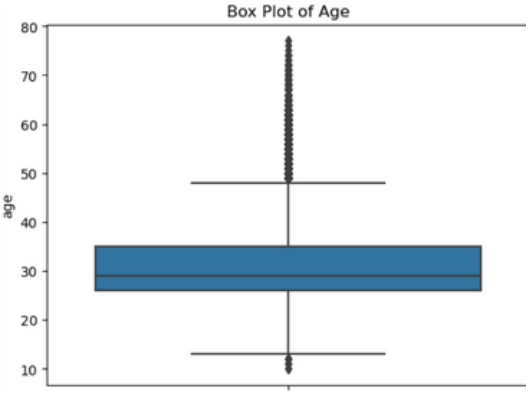
	Column	Null Count	Dtype
0	user_id	0	int64
1	age	95367	float64
2	sex	5518	object
3	phone	21208	object
4	job	21208	object
5	carrier	21208	object
6	marital_status	357	object

Hình 1: Dữ liệu bị null của user_info



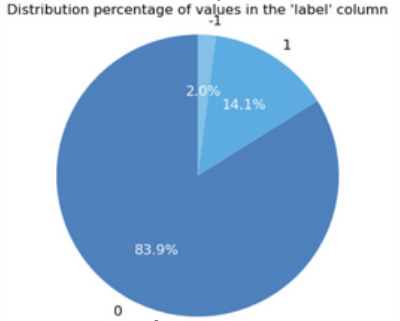
Hình 2: Box Plot của Age

```
carrier
vinaphone      65737
other          65640
mobiphone      65602
viettel        65258
vietnamobile   65158
Name: count, dtype: int64
```

Hình 3: Phân bố carrier sau khi xử lí

	Column	Non-Null Count	Data Type
0	user_id	327395	int64
1	age	327395	int32
2	sex	327395	object
3	job	327395	object
4	carrier	327395	object
5	marital_status	327395	object
6	age_segment	327395	category

Hình 4: User_info sau khi được làm sạch



Hình 5: Phân bố label trong tập train

[TIỀN XỬ LÝ DỮ LIỆU]

Bộ dữ liệu gồm 4 bảng nhóm đã xử lý lần lượt như sau:

1. Bảng user_info: có 424170 dòng, 7 cột

-Data Cleaning:

- Xóa cột **Phone** vì không có ý nghĩa phân tích
- Cột **Age** có 2 vấn đề:

+ M**ising values** (95367 dòng): Điền bằng mode của **Age** theo từng **Marital_status** để giữ phân phối tự nhiên (Nhóm không dùng mean hay drop vì gây mất dữ liệu và lệch phân phối).

+ **Outlier**: Dùng IQR để phát hiện ngoại lệ. Các giá trị ngoại lai sai lệch như: -1, 0, 150, 999 → loại bỏ hoàn toàn khỏi dữ liệu.

- Tạo thêm cột "age_segment": chia tuổi ra thành 4 nhóm (10-26, 26-29, 29-35, 35-77)

-Data Transformation:

- **Cột Sex:** Gom dữ liệu lại thành 3 nhóm: Male, Female, Unknown
- **Cột Job:** Gom dữ liệu lại thành 14 nhóm:
- **Cột Marital_status:** Gom dữ liệu lại thành 4 nhóm: Married, Single, Divorce, Unknown
- **Cột carrier:** sử dụng phương pháp Random để điền vào các giá trị bị thiếu (nan) của mỗi nhà mạng vì các giá trị về nhà mạng có phân bố đều nhau

2. Bảng user_log:

- **Cột Brand_ id:** Tỷ lệ missing values chỉ chiếm 0.17% tổng số dòng chủ yếu gồm: add-to-cart và purchase. Tuy nhiên ảnh hưởng lại không nhỏ nếu đến hành vi add-to-cart, do chiếm 99.78% dữ liệu add-to-cart nằm trong nhóm này.

➔ Thay thế giá trị null trong brand_id = "unknown"

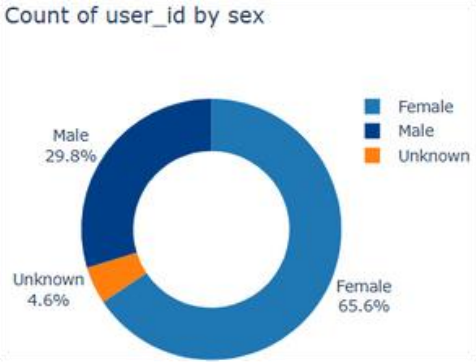
- G**iá trị bị trùng lặp:** Tỷ lệ bị trùng lặp chiếm 37.5%,

trong đó các giá trị của action khi các dòng bị trùng lặp là click (41.42%), purchase (14.75%), add-to-cart (4.76%). Những giá trị bị trùng lặp hướng xử lý sẽ không xóa vì người dùng có thể nhấp chuột, mua hàng hoặc thêm vào giỏ hàng nhiều lần trong ngày.

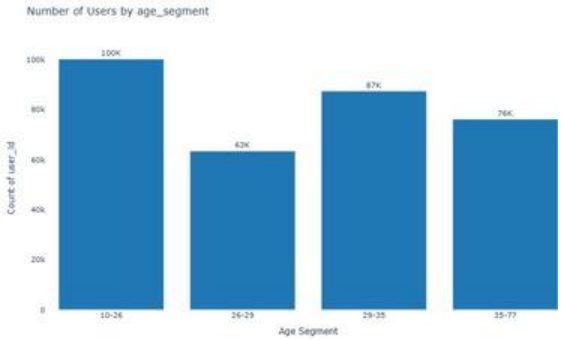
3. Bảng train và test:

- B**ảng train:** không có giá trị trống và trùng lặp, cột label có 3 giá trị 0,1 và -1
- B**ảng test:** không có giá trị trống và trùng lặp, cột label chỉ có 2 giá trị 0 và 1

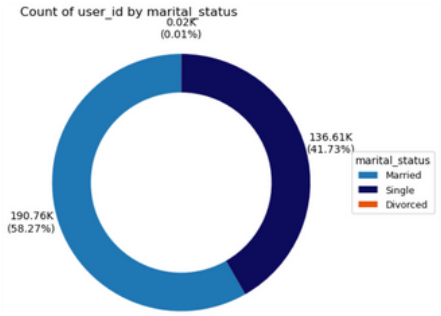
=> Không xóa giá trị -1 ở bảng train do nó có thể là giá trị chưa xác định được



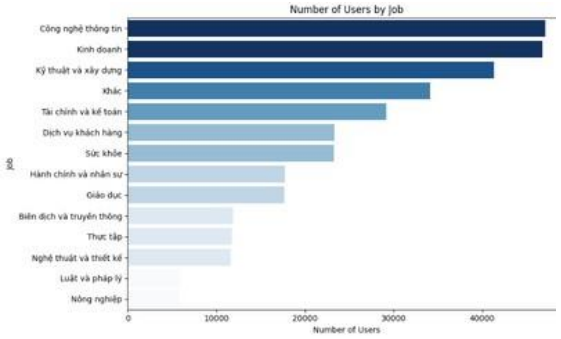
Hình 6: Phân bố giới tính



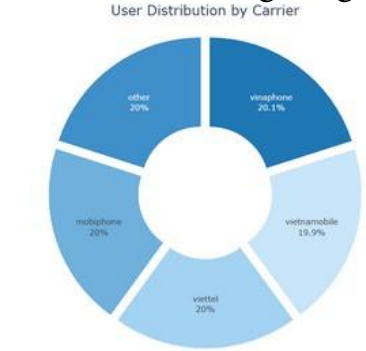
Hình 7: Phân bố nhóm tuổi



Hình 8: Phân bố tình trạng hôn nhân



Hình 9: Phân bố nghề nghiệp



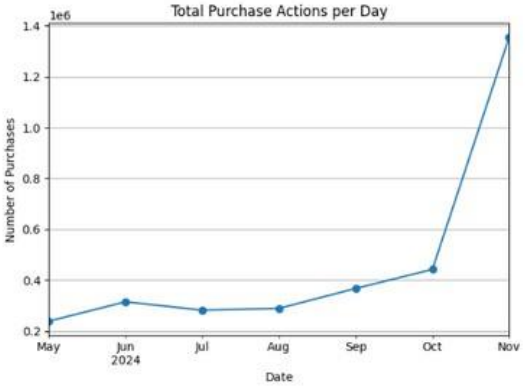
[PHÂN TÍCH KHÁM PHÁ DỮ LIỆU]

Phân tích tổng quan về khách hàng (phân tích đơn biến):

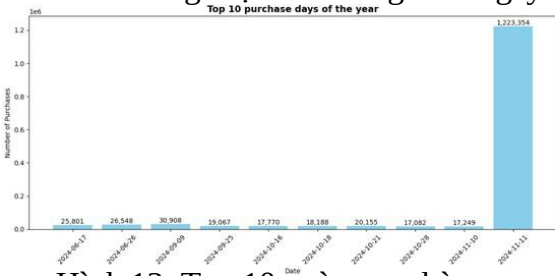
- **Sex**: Phái **nữ** là nhóm khách hàng chính (chiếm **65**%)

➔ Chiến lược sản phẩm/dịch vụ nên nhấn mạnh yếu tố cảm xúc, thiết kế, trải nghiệm người dùng (UX)

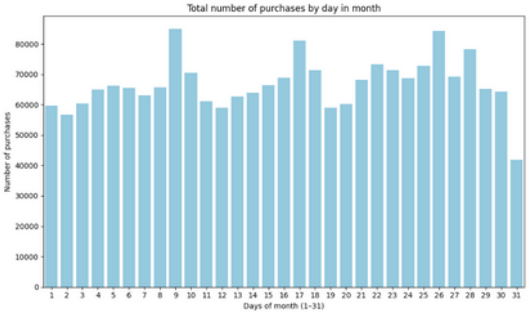
- **Age_segment:** Tệp khách hàng trẻ (**gen Z và gen Y**) là chủ lực, có hành vi tiêu dùng linh hoạt, quen thuộc với công nghệ và thích mua sắm trực tuyến, trải nghiệm mới, các app hiện đại.
- **Job**: Phần lớn người dùng đến từ nhóm ngành có thu nhập ổn định và am hiểu công nghệ, đặc biệt là CNTT và Kinh doanh, Kỹ thuật xây dựng→ dễ tiếp cận qua kênh digital, email, ví điện tử, nền tảng fintech...
- **Marital_status**: Đa phần khách hàng là người độc thân (58%), có khả năng chi tiêu cho bản thân cao hơn, thường ưu tiên trải nghiệm, công nghệ, học tập, phát triển cá nhân → thích hợp cho các dịch vụ như học online, du lịch, công nghệ mới.
- **Carrier**: Viettel, Vinaphone, Mobifone,Vietnamobile, phân bố đồng đều (~mỗi nhà mạng chiếm 20%)



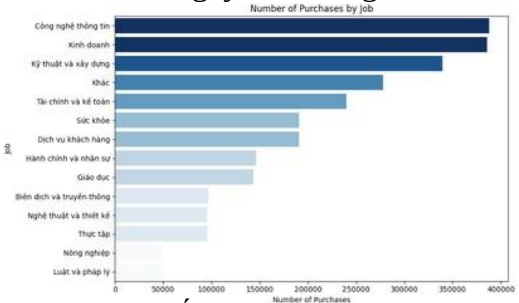
Hình 11: Tổng lượt mua hàng theo ngày



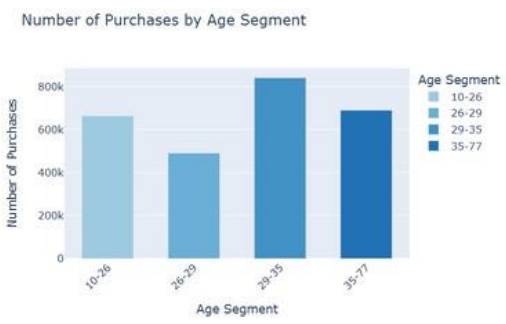
Hình 12: Top 10 ngày mua hàng nhiều nhất



Hình 13: Tổng lượt mua hàng mỗi ngày của tháng



Hình 14: Số lượt mua hàng theo nghành nghề



Hình 15: Số lượt mua hàng theo nhóm tuổi

[PHÂN TÍCH KHÁM PHÁ DỮ LIỆU]

-Phân tích hành vi khách hàng (phân tích đa biến):

1/ Theo thời gian:

Insights 1: Ngày **11/11/2024** (ngày đôi) ghi nhận số lượng mua hàng tăng đột biến (**~1.2 triệu lượt**) – cao vượt trội so với các ngày còn lại trong năm.

->Giải thích: Đây là ngày lễ độc thân (Single Day) ngày sale đôi của năm có khả năng mang lại doanh số tăng đột biến.

Insights 2: Tổng số lượt mua cao nhất của 7 tháng là ngày 9, 17, 26.

-> Ý tưởng: Không chỉ tập trung vào 1 ngày sale lớn mà có thể chia nhỏ thời gian vào các ngày khác trong tháng như: ngày 9,17,26

> Giải pháp chung: Phân loại người tiêu dùng dựa trên hành vi mua sắm trong các ngày sale lớn bằng Machine Learning:

- Xây dựng mô hình phân loại để xác định nhóm "Deal Hunters" vs nhóm khách hàng trung thành.
- Dựa trên phân loại, thiết kế chiến lược khuyến mãi cá nhân hóa bằng cách: Giảm ngân sách khuyến mãi cho nhóm Deal Hunters

2/ Theo hành vi mua hàng (purchase)

Job x purchase

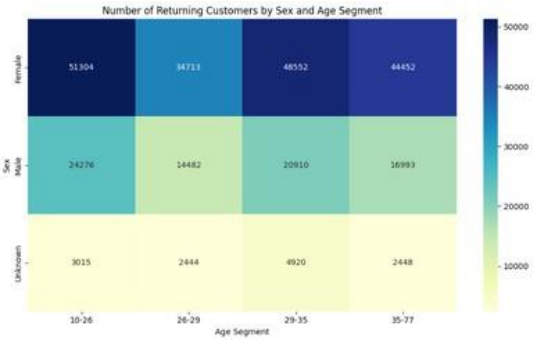
- Công nghệ thông tin và Kinh doanh là hai ngành có sức mua mạnh nhất, cho thấy tiềm năng cao trong các chiến dịch tiếp thị. Nhóm ngành như Kỹ thuật, Tài chính, và Sức khỏe có mức tiêu dùng ổn định, là tệp khách hàng tiềm năng cần duy trì.

Age-segment x purchase

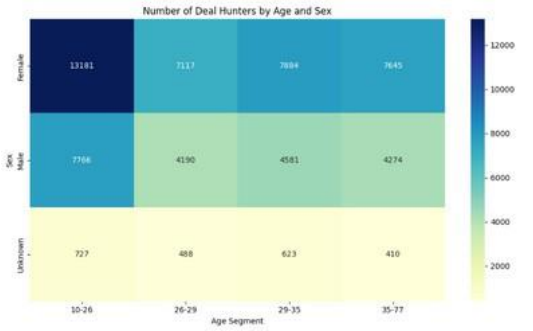
+ Nhóm tuổi 29–35 mua nhiều nhất, vượt mốc 900K lượt mua –> cho thấy đây là phân khúc có sức mua cao nhất và khả năng chi tiêu mạnh.

+ Nhóm 35–77 cũng có sức mua ổn định, gần ngang với nhóm 10–26, chứng minh nhóm tuổi lớn vẫn tham gia tích cực mua sắm online.

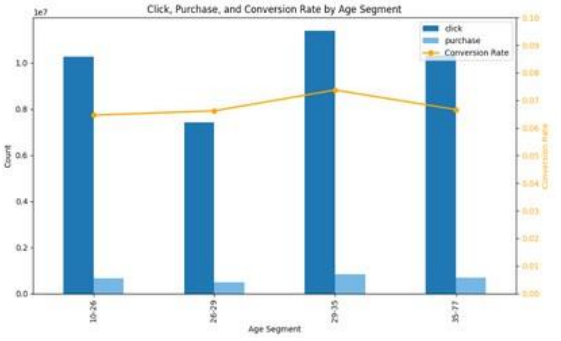
+ Nhóm 26–29 có lượt mua thấp nhất, gợi ý đây là nhóm cần được tăng cường kích thích hành vi mua hàng (ví dụ: ưu đãi cá nhân hóa hoặc remarketing).



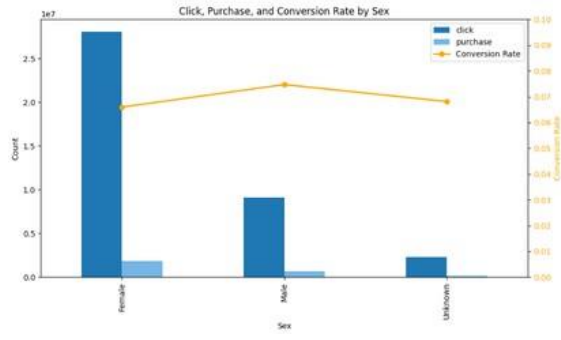
Hình 16: Số lượng khách hàng quay lại theo nhóm tuổi và giới tính



Hình 17: Deal Hunters theo nhóm tuổi và giới tính



Hình 18: Tỷ lệ chuyển đổi theo nhóm tuổi



Hình 19: Tỷ lệ chuyển đổi theo giới tính

[DASHBOARD STORYTELLING]

-Phân tích hành vi khách hàng (phân tích đa biến):

- **Returning Customer by Age-Sex**

Nữ giới là nhóm có số lượng khách hàng quay lại cao nhất ở tất cả các độ tuổi, nổi bật nhất là nhóm **10–26** và **29–35**. Điều này cho thấy nữ giới là tệp khách hàng trung thành và tiềm năng cần được ưu tiên trong các chiến dịch giữ chân. Nam giới có tỷ lệ quay lại thấp hơn rõ rệt, cao nhất ở nhóm **10–26** , giảm dần ở các độ tuổi lớn hơn. Do đó, chiến lược chăm sóc khách hàng nên được cá nhân hóa theo giới tính và độ tuổi để tối ưu hiệu quả.

3/ **Phát hiện Deal Hunter theo độ tuổi và giới tính**

Insights:

- **Nữ giới** (đặc biệt **10–26** tuổi) là nhóm Deal Hunters mạnh nhất – cho thấy họ nhạy với ưu đãi và tích cực mua sắm vào dịp sale.
- Nam giới có xu hướng săn deal ít hơn, thể hiện hành vi mua nhanh – ít tương tác nhưng hiệu quả.
- Số lượng Deal Hunters giảm dần theo độ tuổi, phản ánh nhóm tuổi lớn ít bị thu hút bởi các chương trình khuyến mãi.

4/ Tỷ lệ chuyển đổi Conversion Rate theo Sex và Age-Segment:

Insight : Nữ giới click và mua nhiều nhất, nhưng nam giới có tỷ lệ chuyển đổi cao hơn

-> Giải thích:

- Nữ giới có số lượng click và đơn hàng cao nhất → tương tác mạnh, sản phẩm phù hợp nhu cầu.
- Nam giới có tỷ lệ chuyển đổi cao hơn → quyết định mua nhanh khi đã quan tâm.

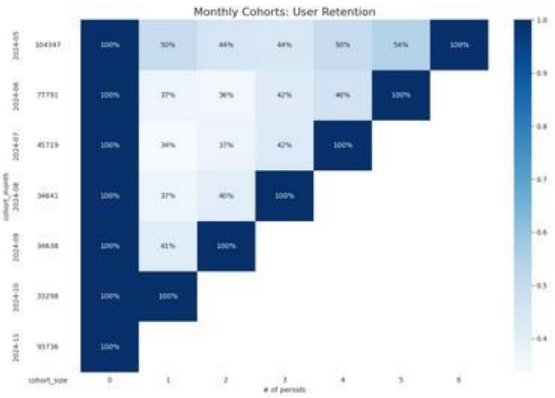
Insight : Người dùng cần trung bình ***13.61 lượt click*** trước khi thực hiện 1 lần mua hàng, cho thấy hành vi mua có xu hướng tìm hiểu kỹ trước khi ra quyết định.

-> Giải thích:

- Tỷ lệ này phản ánh hành trình mua hàng dài, có thể do sản phẩm phức tạp, thông tin chưa đủ rõ ràng hoặc chưa tạo được niềm tin nhanh chóng.
- Người dùng đang tương tác nhiều nhưng chưa đủ thuyết phục để chuyển đổi sớm.

-> Khuyến nghị:

Tối ưu nội dung hiển thị ngay từ lần click đầu Cải thiện trải nghiệm trang đích



Hình 20: Tỷ lệ giữ chân khách hàng theo nhóm tháng (Cohort Analysis)

[DASHBOARD STORYTELLING]

Monthly Cohort: User Retension

Tỷ lệ Khách hàng quay lại theo tháng tfi tháng 5 đến tháng 11

- Tỷ lệ khách hàng quay lại của sàn TMDT giao động từ 34% đến 50%, cao hơn so với trung bình ngành trong khi tỷ lệ khách hàng quay lại của trung bình ngành e-commerce năm 2024 là 38%. Điều này phản ánh chiến lược giữ chân khách hàng hiện tại hiệu quả hơn so với trung bình ngành.
- Số lượng khách hàng mới (cohort_size) giảm dần theo thời gian. Tháng 2024-05 có cohort_size cao nhất (104,347), giảm dần tới 33,298 vào 2024-10. Có thể do chiến dịch marketing ban đầu mạnh mẽ, sau đó giảm dần về sau.
- Retention rate vào tháng 11 lên đến 100% → vì tháng 11 là ngày sale lớn + chiến lược marketing hiệu quả nên hầu như khách hàng cũ quay lại mua hàng vào ngày này

TỔNG KẾT

- Nữ giới là nhóm khách hàng chính, chiếm tỷ trọng lớn (65%), có mức độ tương tác cao và xu hướng quay lại mạnh mẽ, đặc biệt trong các dịp khuyến mãi. Điều này cho thấy chiến lược tiếp thị hiện tại đang phát huy hiệu quả trong việc thu hút và giữ chân nhóm khách hàng nữ.
- Nhóm tuổi trẻ (nhóm 10-35 tuổi) là lực lượng chủ đạo trong hành vi mua sắm trực tuyến, với sức mua lớn và tỷ lệ quay lại cao. Đây là tệp khách hàng quan trọng cần tiếp tục khai thác thông qua các kênh digital và chiến lược trải nghiệm mới mẻ.
- Ngành nghề chủ đạo là CNTT, Kinh doanh và Kỹ thuật – cho thấy người tiêu dùng chủ yếu là nhóm có thu nhập ổn định, am hiểu công nghệ và dễ tiếp cận các kênh digital như email, ví điện tử, fintech.
- Ngày sale lớn như 11/11 đóng vai trò quan trọng trong việc thúc đẩy doanh số và tỷ lệ quay lại. Tuy nhiên, cần kết hợp chiến lược duy trì khách hàng sau ngày sale để tránh tình trạng khách hàng chỉ mua hàng một lần (Deal Hunters).
- Tỷ lệ giữ chân khách hàng hiện tại dao động từ 34%–50% – cao hơn so với trung bình ngành, phản ánh chiến lược giữ chân hiệu quả hơn so với thị trường. Đây là lợi thế cạnh tranh cần phát huy.
- Cần chú trọng cải thiện tỷ lệ chuyển đổi, đặc biệt từ nhóm nữ giới, bằng cách tối ưu nội dung, giao diện, và tăng trải nghiệm tích cực ngay từ lần tương tác đầu tiên.