Contents

**[Phần I: Mở đầu](#_heading=h.n3r8akh97hbe)** [6](#_heading=h.n3r8akh97hbe)

**[1.](#_heading=h.wuhma4a8cr31)****Lý do chọn đề tài:** 6

**[2.](#_heading=h.kyz3gastv7ue)****Mục tiêu Phân tích Dự án** 6

**[3.](#_heading=h.rv12vco4ucuv)****Công cụ phân tích** 7

**[4.](#_heading=h.6ukaohogw9yt)****Dữ liệu sử dụng trong phân tích** 7

**[Giới thiệu chung:](#_heading=h.5pt8ryy45v1m)** [7](#_heading=h.5pt8ryy45v1m)

**[Các đặc điểm nổi bật:](#_heading=h.ex2x7zwll9ei)** [7](#_heading=h.ex2x7zwll9ei)

**[Một số cột tiêu biểu:](#_heading=h.5nbklnnd3921)** [7](#_heading=h.5nbklnnd3921)

**[Phần II: Phân tích](#_heading=h.ircnw7wmicme)** [8](#_heading=h.ircnw7wmicme)

**[2.](#_heading=h.k3ww9yncnkts)****Làm sạch dữ liệu (Data Cleaning)** 13

**[3.](#_heading=h.oycs00tr9nl5)****Phân tích dữ liệu khám phá (EDA - Exploratory Data Analysis)** 16

**[3.1. Độ tuổi (Age](#_heading=h.6yuoxh4968rz)**[) 16](#_heading=h.6yuoxh4968rz)

**[3.2. Giới tính (Gender)](#_heading=h.853zszd2boed)** [17](#_heading=h.853zszd2boed)

**[Sở thích Mua sắm theo Giới tính](#_heading=h.l3btha9b2cl6)** [18](#_heading=h.l3btha9b2cl6)

**[Phân bố trình độ học vấn](#_heading=h.hbaijodo4cb8)** [19](#_heading=h.hbaijodo4cb8)

**[3.3. Nghề nghiệp (Occupation)](#_heading=h.hbxkgo8qinvp)** [20](#_heading=h.hbxkgo8qinvp)

**[Phân bố nghề nghiệp khách hang](#_heading=h.831h4veh7uxv)** [20](#_heading=h.831h4veh7uxv)

**[3.4. Tình trạng hôn nhân (Marital\_Status)](#_heading=h.4pgiofirk15u)** [20](#_heading=h.4pgiofirk15u)

**[Tình trạng hôn nhân của khách hang](#_heading=h.6c0qzvijl4e4)** [20](#_heading=h.6c0qzvijl4e4)

**[3.5. Mức thu nhập (Income\_Level)](#_heading=h.c1qbk5sir2oe)** [21](#_heading=h.c1qbk5sir2oe)

**[Mức thu nhập của khách hàng](#_heading=h.p2ks33jlrh4)** [21](#_heading=h.p2ks33jlrh4)

**[Tần suất Mua hàng theo Nhóm Thu nhập](#_heading=h.k95m6wd67pd3)** [22](#_heading=h.k95m6wd67pd3)

**[Mức độ hài lòng theo Nhóm Thu nhập](#_heading=h.5dnr25q87w89)** [23](#_heading=h.5dnr25q87w89)

**[Kết luận:](#_heading=h.x79nlescad6f)** [23](#_heading=h.x79nlescad6f)

**[4.Hành vi mua sắm (Purchase Behavior)](#_heading=h.hfpurce9zqbz)** [24](#_heading=h.hfpurce9zqbz)

**[4.1. Tổng số đơn hàng (Frequency\_of\_Purchase)](#_heading=h.b3030rd05mhu)** [24](#_heading=h.b3030rd05mhu)

**[4.2. Tổng chi tiêu (Purchase\_Amount)](#_heading=h.q73jignqyw4g)** [25](#_heading=h.q73jignqyw4g)

**[4.3. Thời điểm mua hàng (Time\_of\_Purchase)](#_heading=h.3wpygidaf5vi)** [26](#_heading=h.3wpygidaf5vi)

**[4.4. Hạng mục sản phẩm (Purchase\_Category)](#_heading=h.5rbrwktm62tt)** [27](#_heading=h.5rbrwktm62tt)

**[4.5. Thiết bị sử dụng (Device\_Used\_for\_Shopping)](#_heading=h.cv7f9xnetpk1)** [28](#_heading=h.cv7f9xnetpk1)

**[4.6. Phương thức thanh toán (Payment\_Method)](#_heading=h.z9wyz4s8jojb)** [29](#_heading=h.z9wyz4s8jojb)

**[4.7. Sở thích vận chuyển (Shipping\_Preference)](#_heading=h.y5e3kapasm0m)** [31](#_heading=h.y5e3kapasm0m)

**[4.8. Kết luận:](#_heading=h.w4jdnlvh3s2n)** [32](#_heading=h.w4jdnlvh3s2n)

**[5.Phần phân tích mức độ tương tác và hành vi của khách hàng đối với sản phẩm và quảng cáo](#_heading=h.x9orblme1jd3)** [32](#_heading=h.x9orblme1jd3)

**[5.1. Phân tích biểu đồ mật độ phân bố thời gian tìm hiểu sản phẩm (Male vs Female)](#_heading=h.rzpmz6unndvn)** [32](#_heading=h.rzpmz6unndvn)

**[5.2. Phân tích biểu đồ mức độ tương tác với quảng cáo theo giới tính](#_heading=h.f7qi4lq1tms8)** [35](#_heading=h.f7qi4lq1tms8)

**[5.3. Phân tích mức độ hài lòng của khách hàng theo mức độ tương tác với quảng cáo](#_heading=h.ic6d3flkc692)** [36](#_heading=h.ic6d3flkc692)

**[5.4. Phân tích mối liên hệ giữa ảnh hưởng mạng xã hội và mục đích mua hàng](#_heading=h.68w6d5iaotqz)** [38](#_heading=h.68w6d5iaotqz)

**[5.5. Kết Luận](#_heading=h.3ix3sloxqdwz)** [39](#_heading=h.3ix3sloxqdwz)

**[6.Phân cụm khách hàng](#_heading=h.3b9pp0wanvui)** [39](#_heading=h.3b9pp0wanvui)

**[6.1. Chuẩn bị](#_heading=h.sofk3p1be34s)** [39](#_heading=h.sofk3p1be34s)

**[6.2. Tính/Chuẩn bị các đặc trưng](#_heading=h.3m2f6d7gfoar)** [40](#_heading=h.3m2f6d7gfoar)

**[6.3. One-hot encode cột Gender](#_heading=h.o465cydbh750)** [41](#_heading=h.o465cydbh750)

**[6.4. Kết hợp features](#_heading=h.twpk92kknhnn)** [41](#_heading=h.twpk92kknhnn)

**[6.5. Chuẩn hóa dữ liệu](#_heading=h.pkg2umm73vr6)** [41](#_heading=h.pkg2umm73vr6)

**[6.6. Áp dụng KMeans](#_heading=h.2o09urlud6kg)** [42](#_heading=h.2o09urlud6kg)

**[6.6.1 Tính profile trung bình của mỗi cụm](#_heading=h.xk7d57jbuayb)** [42](#_heading=h.xk7d57jbuayb)

**[6.6.2. Giảm chiều với PCA 2D](#_heading=h.azavyuhhpapa)** [43](#_heading=h.azavyuhhpapa)

**[6.6.3. Vẽ scatter với centroids và tên segment](#_heading=h.pub5laef002n)** [43](#_heading=h.pub5laef002n)

**[6.7. Áp dụng DBSCAN](#_heading=h.9zr91i5urt24)** [45](#_heading=h.9zr91i5urt24)

[Giải thích & nhận xét: 47](#_heading=h.35yqlar1q8lk)

[Đề xuất hành động: 48](#_heading=h.eb6rb1h41rhf)

**[6.8. Thống kê trung bình đặc trưng theo cụm KMeans](#_heading=h.m757ikm6rze2)** [49](#_heading=h.m757ikm6rze2)

**[Giải thích:](#_heading=h.6dbtfl8z50jw)** [49](#_heading=h.6dbtfl8z50jw)

**[Nhận xét:](#_heading=h.qa00suigqbm8)** [50](#_heading=h.qa00suigqbm8)

**[Đề xuất hành động:](#_heading=h.whoru7h39dfd)** [50](#_heading=h.whoru7h39dfd)

**[6.9. Thống kê trung bình đặc trưng theo cụm DBSCAN](#_heading=h.qcah8tpfffer)** [50](#_heading=h.qcah8tpfffer)

**[Giải thích bảng](#_heading=h.pkdpy8d0tk0t)** [51](#_heading=h.pkdpy8d0tk0t)

**[Nhận xét](#_heading=h.4q3ecp584i8z)** [51](#_heading=h.4q3ecp584i8z)

**[Đề xuất hành động](#_heading=h.xw6ng6ntvck0)** [51](#_heading=h.xw6ng6ntvck0)

**[7.Dự đoán hành vi](#_heading=h.4l02xk6jn0eo)** [52](#_heading=h.4l02xk6jn0eo)

**[7.1. Mô hình Phân loại:](#_heading=h.qzo2bie9gmt2)** [52](#_heading=h.qzo2bie9gmt2)

**[7.2. Mô hình Hồi quy:](#_heading=h.12gudbtgwqhi)** [52](#_heading=h.12gudbtgwqhi)

**[7.3. Mục đích sử dụng mô hình](#_heading=h.9a29nkpu41e8)** [53](#_heading=h.9a29nkpu41e8)

**[7.4.Kết quả sau khi huấn luyện mô hình](#_heading=h.kouur4m4pt3v)** [53](#_heading=h.kouur4m4pt3v)

**[7.4.1. Đối với bài toán Phân loại nhóm chi tiêu:](#_heading=h.cdybf36fubaj)** [53](#_heading=h.cdybf36fubaj)

**[7.4.2. Đối với bài toán Dự báo chi tiêu](#_heading=h.uzdvcufiuu6z)** [54](#_heading=h.uzdvcufiuu6z)

**[7.5. Đánh giá hiệu quả mô hình](#_heading=h.oqai60e1ku67)** [54](#_heading=h.oqai60e1ku67)

**[7.6.Đề xuất:](#_heading=h.lqto4jt400sk)** [55](#_heading=h.lqto4jt400sk)

**[7.6.1. Phân khúc khách hàng để cá nhân hóa chiến dịch](#_heading=h.srxuf8vutope)** [55](#_heading=h.srxuf8vutope)

**[7.6.2. Tối ưu trải nghiệm mua sắm và kênh tiếp cận](#_heading=h.wa7u34ki0twx)** [55](#_heading=h.wa7u34ki0twx)

**[8.Kết luận và Đề xuất](#_heading=h.gce1psp341ui)** [55](#_heading=h.gce1psp341ui)

**[8.1. Kết luận](#_heading=h.pr4wj5npeadv)** [55](#_heading=h.pr4wj5npeadv)

**[8.2. Đề xuất](#_heading=h.sdesrxlptxik)** [56](#_heading=h.sdesrxlptxik)

**[8.2.1. Tăng cường cá nhân hóa trên cơ sở phân khúc khách hàng](#_heading=h.a6ze6wvy4jna)** [56](#_heading=h.a6ze6wvy4jna)

**[8.2.2.](#_heading=h.tx72qjlhnbvi)****[Tối ưu hóa trải nghiệm trên thiết bị di động](#_heading=h.tx72qjlhnbvi)** [56](#_heading=h.tx72qjlhnbvi)

**[8.2.3. Kết nối quảng cáo – trải nghiệm – sản phẩm](#_heading=h.mgymr4g7h3qa)** [56](#_heading=h.mgymr4g7h3qa)

**[8.2.4. Đề xuất dài hạn](#_heading=h.qsmkaavksuqd)** [56](#_heading=h.qsmkaavksuqd)

# **Phần I: Mở đầu**

1. **Lý do chọn đề tài:**

Trong bối cảnh chuyển đổi số diễn ra mạnh mẽ, thương mại điện tử không chỉ đơn thuần là một kênh bán hàng mà đã trở thành một phần tất yếu trong hành vi tiêu dùng hiện đại. Việc người tiêu dùng dễ dàng tiếp cận, so sánh và đưa ra quyết định mua sắm chỉ trong vài giây đã đặt ra thách thức lớn cho các doanh nghiệp: liệu họ đã thật sự "hiểu" khách hàng của mình? Sự thay đổi nhanh chóng trong xu hướng tiêu dùng đòi hỏi doanh nghiệp phải liên tục cập nhật, phân tích và điều chỉnh chiến lược kinh doanh. Đề tài "Phân tích hành vi tiêu dùng trong thương mại điện tử” từ dữ liệu Ecommerce\_Consumer\_Behavior\_Analysis\_Data.csv được chọn nhằm tiếp cận bài toán đó từ góc nhìn dữ liệu, giúp hiểu sâu sắc hơn về đặc điểm nhân khẩu học, hành vi mua sắm và mức độ tương tác của khách hàng – từ đó hỗ trợ doanh nghiệp xây dựng chiến lược phù hợp, nâng cao trải nghiệm người dùng và gia tăng mức độ gắn bó thương hiệu.

1. **Mục tiêu Phân tích Dự án**

Dự án nhằm phân tích hành vi tiêu dùng trong thương mại điện tử dựa trên dữ liệu khách hàng đã được làm sạch, với các mục tiêu chính:

* Hiểu rõ đặc điểm nhân khẩu học của khách hàng (tuổi, giới tính, học vấn, thu nhập, v.v.).
* Phân tích hành vi mua sắm như tần suất, chi tiêu, thời gian mua, thiết bị và phương thức thanh toán.
* Đánh giá mức độ hài lòng, trung thành và sự tương tác với quảng cáo/truyền thông.
* Phân cụm khách hàng để phát hiện các nhóm hành vi tiêu biểu và xây dựng chiến lược cá nhân hóa.
* Dự đoán hành vi tiêu dùng, như mức chi tiêu trong tương lai.

Đưa ra đề xuất chiến lược giúp doanh nghiệp nâng cao trải nghiệm, tối ưu marketing và tăng trưởng doanh thu.

1. **Công cụ phân tích**:

Jupyter Notebook

1. **Dữ liệu sử dụng trong phân tích**

**Giới thiệu chung:**

Dataset này cung cấp dữ liệu toàn diện về hành vi người tiêu dùng, hỗ trợ nghiên cứu thị trường, phân tích thống kê, phân khúc khách hàng và mô hình dự đoán. Nó chứa thông tin về hành vi mua sắm, nhân khẩu học, mức độ hài lòng, sở thích sản phẩm và các yếu tố ảnh hưởng đến quyết định mua hàng.

**Các đặc điểm nổi bật:**

* Nhân khẩu học: Tuổi, giới tính, thu nhập, học vấn, tình trạng hôn nhân, nghề nghiệp, vị trí.
* Hành vi mua sắm: Số lần mua, số tiền chi tiêu, loại sản phẩm, kênh mua sắm.
* Trung thành khách hàng: Mức độ trung thành thương hiệu, tham gia chương trình khách hàng thân thiết, tương tác với quảng cáo.
* Phản hồi sản phẩm: Đánh giá sản phẩm, mức độ hài lòng.
* Quyết định mua hàng: Thời gian nghiên cứu sản phẩm, ý định mua hàng, thời gian đưa ra quyết định.
* Ảnh hưởng bên ngoài: Ảnh hưởng của mạng xã hội, nhạy cảm với giảm giá, tỉ lệ trả hàng.

**Một số cột tiêu biểu:**

Customer\_ID, Age, Gender, Income\_Level, Education\_Level, Purchase\_Amount, Purchase\_Channel, Brand\_Loyalty, Product\_Rating, Social\_Media\_Influence, Purchase\_Intent, Time\_of\_Purchase, Customer\_Satisfaction, v.v.

Ecommerce\_Consumer\_Behavior\_Analysis\_Data.csv, là dữ liệu thương mại điện tử đã được làm sạch và xử lý sơ bộ để đảm bảo tính chính xác và nhất quán. Bộ dữ liệu bao gồm các thông tin chi tiết về đặc điểm nhân khẩu học của khách hàng như tuổi, giới tính, trình độ học vấn, nghề nghiệp, tình trạng hôn nhân và mức thu nhập. Bên cạnh đó, dữ liệu còn ghi nhận các hành vi mua sắm như tần suất mua hàng, tổng chi tiêu, thời điểm mua hàng, loại sản phẩm, phương thức thanh toán và thiết bị sử dụng. Ngoài ra, các chỉ số liên quan đến mức độ hài lòng, trung thành, sự tương tác với quảng cáo, và các yếu tố ảnh hưởng đến quyết định mua hàng cũng được lưu trữ trong bộ dữ liệu này. Nhờ sự đa dạng và phong phú của các thuộc tính, dữ liệu cung cấp nền tảng vững chắc để phân tích toàn diện hành vi tiêu dùng và đưa ra những nhận định, đề xuất phù hợp cho doanh nghiệp.

* Dữ liệu trước khi làm sạch: Ecommerce\_Consumer\_Behavior\_Analysis\_Data.csv
* Dữ liệu sau khi làm sạch: Cleaned\_Ecommerce\_Data.csv
* Nguồn: [Ecommerce Consumer Behavior Analysis Data](https://www.kaggle.com/datasets/salahuddinahmedshuvo/ecommerce-consumer-behavior-analysis-data)

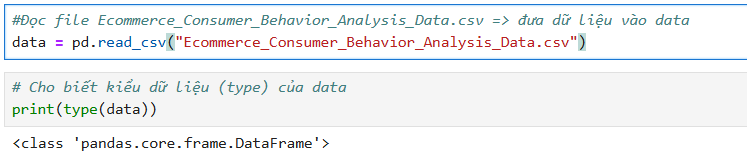
# **Phần II: Phân tích**

1. **Nhập dữ liệu và khám phá sơ bộ (Data Import và Exploration)**

-Thêm thư viện cần thiết:

****

-Đọc file và cho biết kiểu dữ liệu:

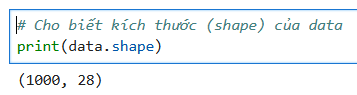
****

Kết quả cho biết data là một đối tượng thuộc lớp DataFrame của thư viện pandas.

DataFrame là một cấu trúc dữ liệu dạng bảng hai chiều, tương tự như một bảng trong Excel hoặc một bảng trong cơ sở dữ liệu.

Mỗi hàng tương ứng với một bản ghi khách hàng (customer record), còn mỗi cột là một thuộc tính hành vi tiêu dùng (như tuổi, giới tính, thu nhập, mức độ tương tác với quảng cáo, v.v.).

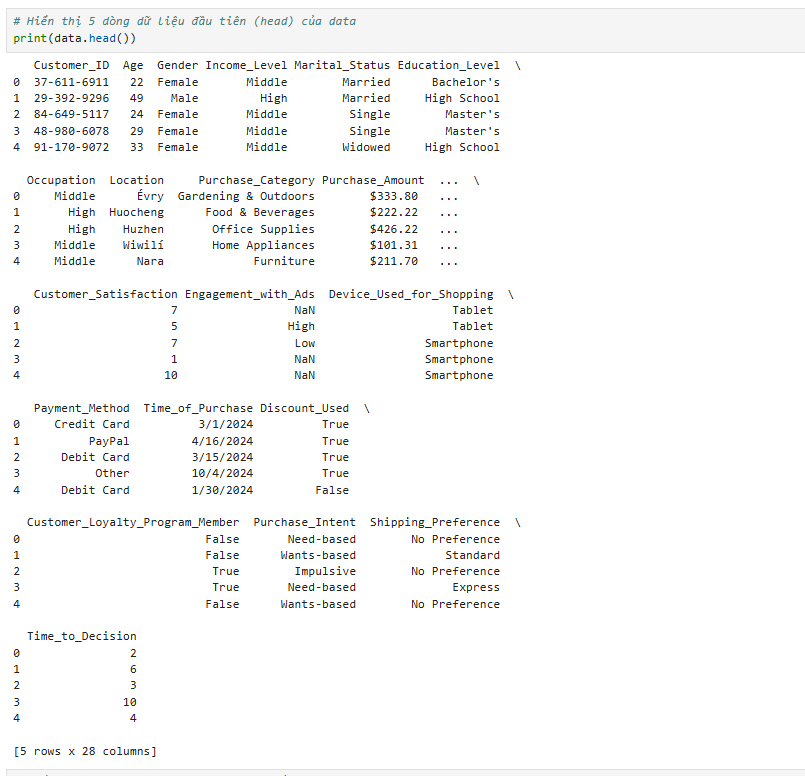
-Trả về kích thước của bảng dữ liệu dưới dạng một bộ tuple (số\_dòng, số\_cột).



1000 dòng: Bảng dữ liệu chứa thông tin của 1000 khách hàng khác nhau.

28 cột: Mỗi khách hàng được mô tả bởi 28 thuộc tính (feature).

-Hiển thị 5 dòng đầu tiên của bảng dữ liệu

****

Kết quả trả về gồm 5 bản ghi đầu tiên, mỗi bản ghi tương ứng với một khách hàng, với đầy đủ 28 cột thuộc tính. Dưới đây là một số nhóm thông tin nổi bật:

1. Thông tin cá nhân:

* *Customer\_ID*: Mã định danh của khách hàng.
* *Age, Gender, Income\_Level, Marital\_Status, Education\_Level*: Đặc điểm nhân khẩu học.
* *Occupation, Location*: Nghề nghiệp và địa điểm sinh sống.

2. Hành vi mua sắm:

* *Purchase\_Category, Purchase\_Amoun*t: Danh mục và số tiền đã mua.
* *Customer\_Satisfaction*: Mức độ hài lòng (thang điểm).
* *Payment\_Method*: Phương thức thanh toán.
* *Time\_of\_Purchase*: Ngày mua hàng.
* *Discount\_Used*: Có dùng mã giảm giá hay không.
* *Shipping\_Preference*: Hình thức giao hàng ưu tiên.
* *Device\_Used\_for\_Shopping*: Thiết bị sử dụng khi mua hàng.

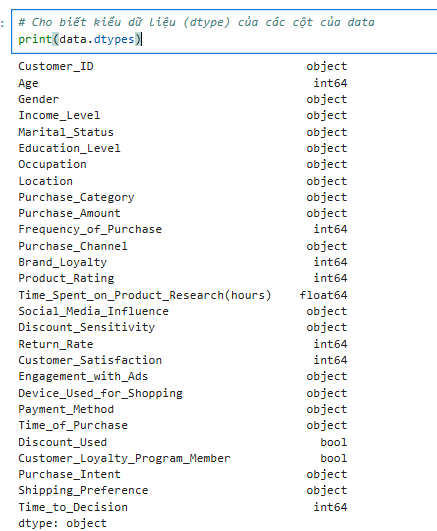
3. Tương tác và ảnh hưởng bên ngoài:

* *Engagement\_with\_Ads*: Mức độ tương tác với quảng cáo.
* *Social\_Media\_Influence*: Mức độ ảnh hưởng từ mạng xã hội.
* *Customer\_Loyalty\_Program\_Member*: Là thành viên của chương trình khách hàng thân thiết hay không.
* *Purchase\_Intent*: Động cơ mua hàng (Need-based, Wants-based, Impulsive, ...).
* *Time\_to\_Decision*: Thời gian cân nhắc trước khi quyết định mua hàng.

-Tương tự hiển thị 5 dòng cuối của bảng dữ liệu

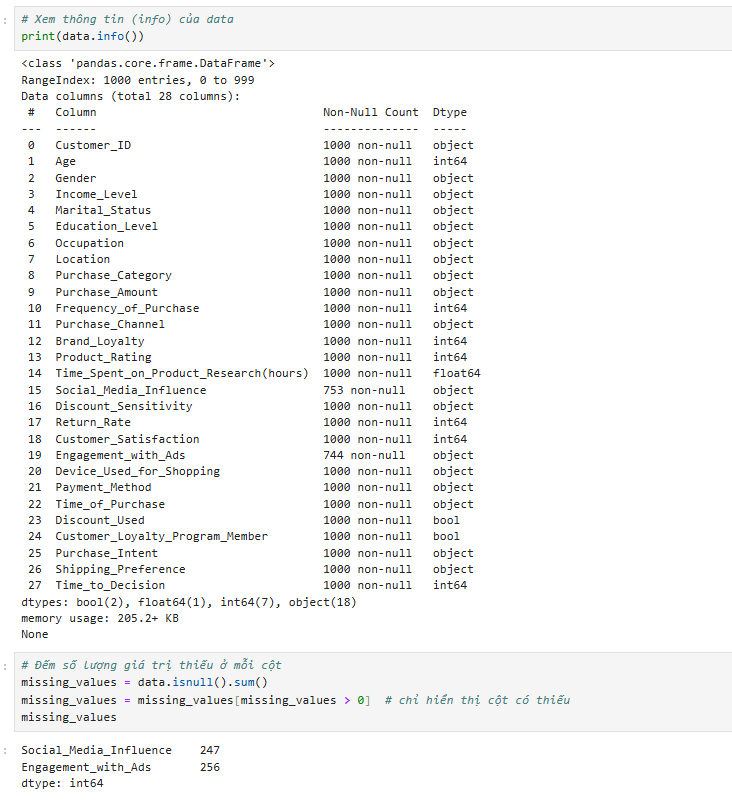


-Kiểu dữ liệu (data type) của từng cột (thuộc tính) trong bảng data



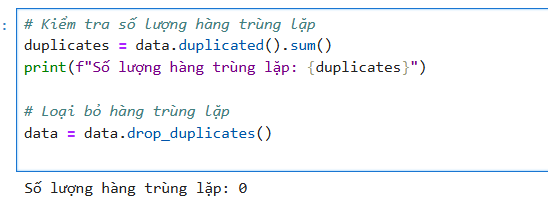
Kiểu dữ liệu được phân loại theo 4 nhóm: kiểu chuỗi (object), kiểu số nguyên (int64), kiểu số thực (float64), kiểu logic (bool)

- Thông tin tổng quan về dữ liệu – data.info() và kiểm tra giá trị thiếu – data.isnull().sum()

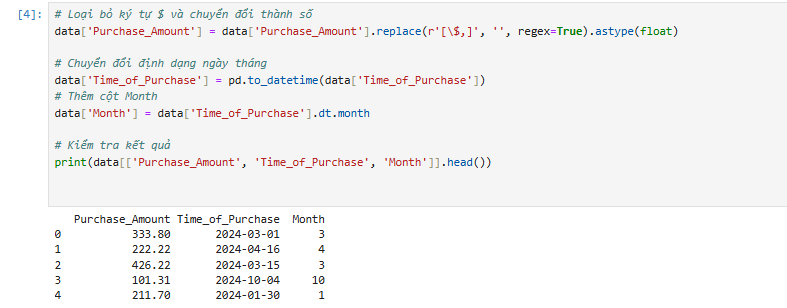


1. **Làm sạch dữ liệu (Data Cleaning)**

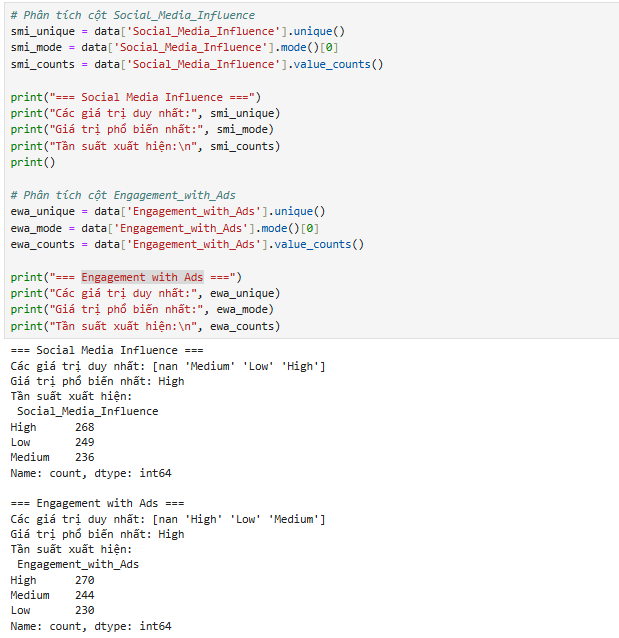
-Kiểm tra số lượng hàng trùng lặp và loại bỏ hàng trùng lặp



- Loại bỏ ký tự $ và chuyển đổi thành số, định dạng ngày tháng và them cột Month



- Phân tích cột Social\_Media\_Influence và Engagement with Ads



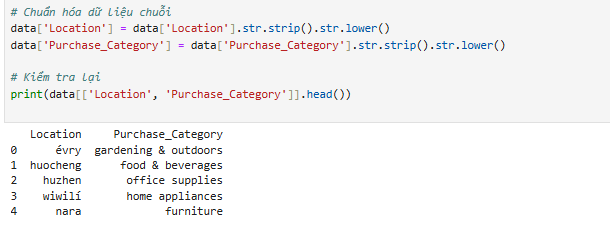
-Điền giá trị phổ biến nhất cho các cột bị thiếu



-Hàm loại bỏ giá trị ngoại lai bằng IQR và áp dụng loại bỏ ngoại lai cho các cột số



- Chuẩn hóa dữ liệu chuỗi



- Lưu lại thành file mới có tên Cleaned\_Ecommerce\_Data.csv



1. **Phân tích dữ liệu khám phá (EDA - Exploratory Data Analysis)**

Phân tích các yếu tố nhân khẩu học như độ tuổi, giới tính, học vấn, nghề nghiệp, thu nhập, tình trạng hôn nhân và khu vực địa lý.

Mục tiêu: Tạo chân dung khách hàng rõ ràng để phục vụ việc phân khúc và cá nhân hóa trải nghiệm.

Các yếu tố được phân tích:

## **3.1. Độ tuổi (Age**)

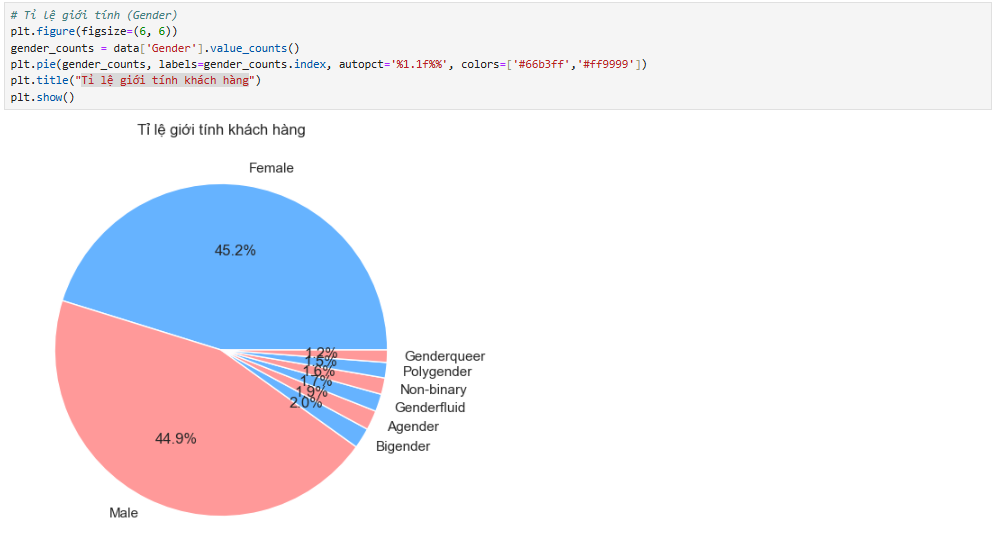
**Phân phối độ tuổi (Age)**



Khách hàng phân bổ khá đều từ 18 đến 50 tuổi, trong đó nổi bật là hai nhóm tuổi 34–37 và 46–50 với lượng khách cao nhất. Tuy nhiên, nhóm 24–30 tuổi cũng chiếm tỷ trọng lớn, cho thấy đây là nhóm khách hàng trẻ tiềm năng, kết hợp giữa độ tuổi sẵn sàng chi tiêu và sự quen thuộc với công nghệ.Phân bố độ tuổi khách hàng trải rộng, cho thấy doanh nghiệp đang phục vụ đa dạng thế hệ. Nhóm tuổi 24–30 và 34–37 có số lượng cao, là các độ tuổi đang trong giai đoạn ổn định thu nhập và chi tiêu mạnh. Nhóm 46–50 cũng bất ngờ chiếm tỷ trọng lớn, cho thấy sức mua từ nhóm khách trung niên không thể bỏ qua. Chiến lược nên đồng thời nhắm đến cả người trẻ thành thị và người tiêu dùng lớn tuổi có khả năng tài chính ổn định.

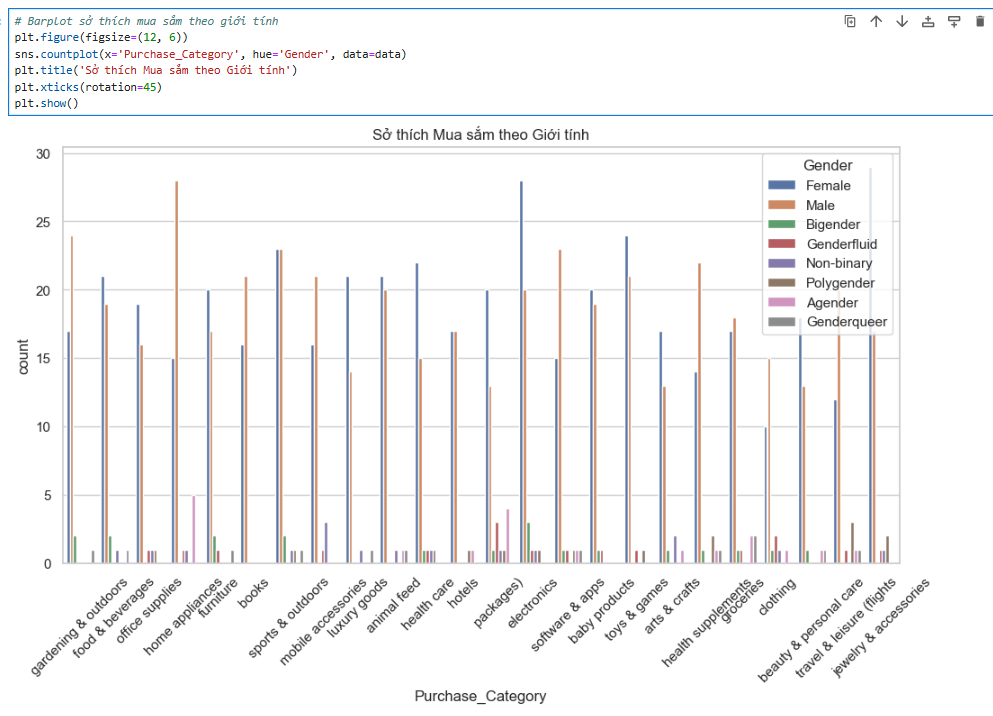
## **3.2. Giới tính (Gender)**

**Tỉ lệ giới tính khách hang**

****

Nam và nữ chiếm tỷ lệ gần tương đương, lần lượt là 45.2% và 44.9%, cho thấy không có sự chênh lệch đáng kể giữa hai giới tính phổ biến này. Nhóm khách hàng phi nhị nguyên giới (non-binary và các biến thể khác) chiếm tổng cộng 9.9%, phản ánh sự đa dạng giới tính trong tập khách hàng – một yếu tố cần được lưu tâm trong truyền thông và thiết kế trải nghiệm số. Trái với giả định ban đầu rằng “nữ chiếm 60%”, dữ liệu thực tế cho thấy doanh nghiệp nên định hướng chiến lược trung hòa giới tính, đồng thời đảm bảo sự hòa nhập cho nhóm LGBTQ+. Tập khách hàng có giới tính đa dạng với tỷ lệ nam và nữ gần như ngang nhau (~45%), cùng một tỷ lệ đáng kể thuộc các nhóm phi nhị nguyên. Doanh nghiệp nên xây dựng chiến lược trung lập giới và cá nhân hóa trải nghiệm để phù hợp với sự đa dạng này.

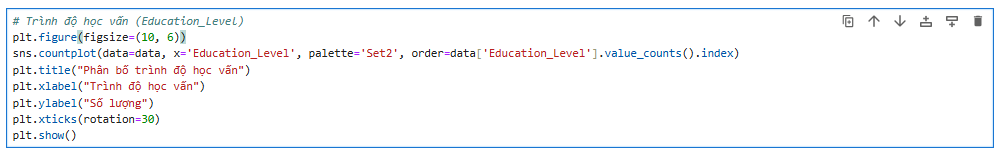
### **Sở thích Mua sắm theo Giới tính**

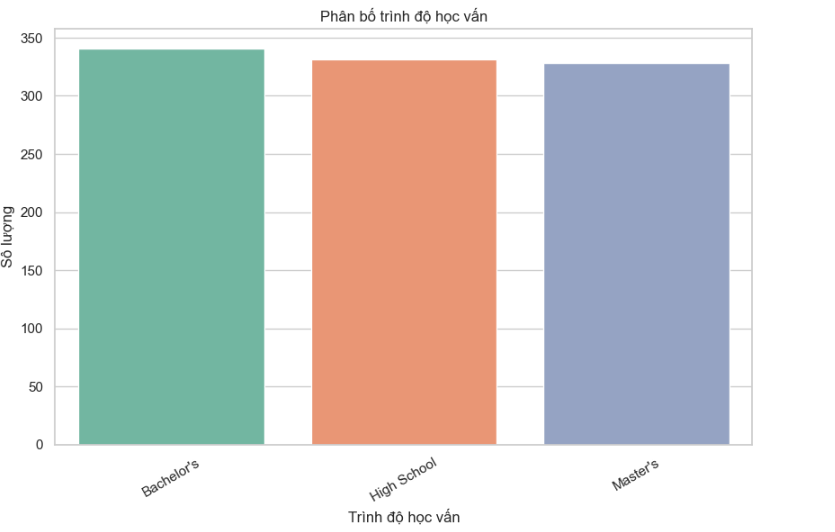


Biểu đồ thể hiện sự khác biệt rõ rệt về sở thích mua sắm giữa các nhóm giới tính. Nam giới thường ưa chuộng các sản phẩm công nghệ, du lịch và đồ ăn thức uống, trong khi nữ giới quan tâm nhiều hơn đến thời trang, làm đẹp và nội thất. Các nhóm giới tính khác như Non-binary hay Genderqueer tuy chiếm tỷ lệ nhỏ nhưng thể hiện sự đa dạng trong hành vi tiêu dùng. Kết quả cho thấy giới tính là yếu tố quan trọng ảnh hưởng đến hành vi mua sắm, từ đó giúp doanh nghiệp xây dựng chiến lược tiếp thị phù hợp cho từng nhóm khách hàng.

- Trình độ học vấn (Education\_Level)

### **Phân bố trình độ học vấn**

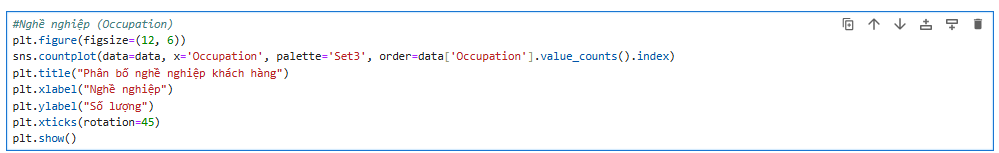


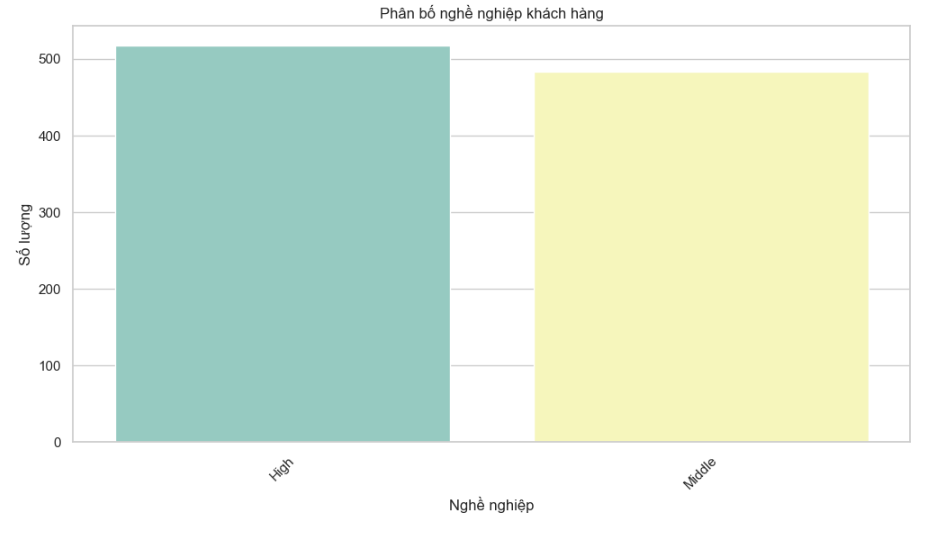


Khách hàng trên nền tảng thương mại điện tử có trình độ học vấn cao, với số lượng gần như ngang nhau giữa ba nhóm: Cử nhân (khoảng 341 người), Trung học (khoảng 331 người) và Thạc sĩ (khoảng 328 người). Điều này cho thấy doanh nghiệp đang tiếp cận một tệp khách hàng có hiểu biết, quen thuộc với công nghệ và có khả năng chi tiêu ổn định. Với đặc điểm này, hành vi mua sắm của họ nhiều khả năng chịu ảnh hưởng bởi các yếu tố lý tính như thông tin sản phẩm chi tiết, đánh giá người dùng và độ tin cậy của thương hiệu. Do đó, doanh nghiệp nên tập trung phát triển nội dung truyền thông chuyên sâu, chú trọng đến trải nghiệm mua sắm thông minh và minh bạch. Ngoài ra, đây cũng là nhóm tiềm năng cho các sản phẩm cao cấp hoặc gói dịch vụ giá trị gia tăng.

## **3.3. Nghề nghiệp (Occupation)**

### **Phân bố nghề nghiệp khách hang**

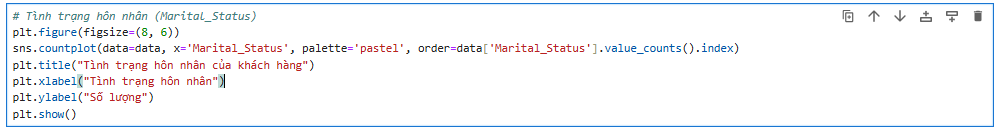
****

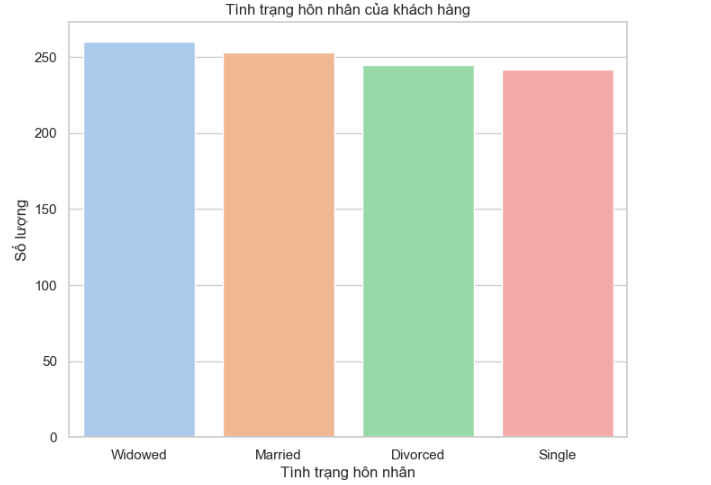
****

Biểu đồ cho thấy phân bố khách hàng theo nghề nghiệp tập trung chủ yếu ở hai nhóm: "High" và "Middle". Nhóm "High" chiếm tỷ lệ lớn hơn một chút với hơn 520 người, trong khi nhóm "Middle" có gần 500 người. Sự chênh lệch không đáng kể cho thấy cơ cấu nghề nghiệp của khách hàng khá cân bằng. Dữ liệu nghề nghiệp có thể đã được phân loại theo mức thu nhập hoặc cấp bậc chuyên môn. Đây là cơ sở tốt để kết hợp với các yếu tố khác như thu nhập, độ tuổi hoặc hành vi chi tiêu nhằm phân tích sâu hơn.

## **3.4. Tình trạng hôn nhân (Marital\_Status)**

### **Tình trạng hôn nhân của khách hang**

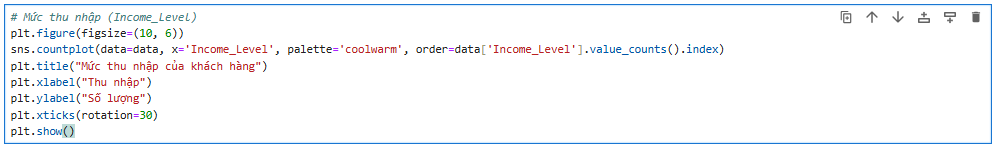


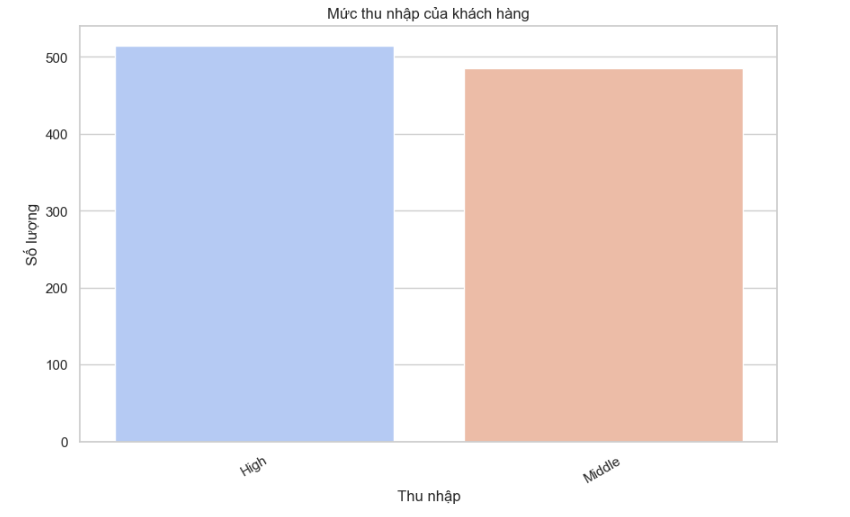


Dữ liệu về tình trạng hôn nhân cho thấy sự phân bổ khá đồng đều giữa bốn nhóm: Đã kết hôn (253), Độc thân (242), Ly hôn (245) và Góa (260). Điều này phản ánh một tập khách hàng đa dạng về mặt lối sống và nhu cầu mua sắm. Nhóm đã kết hôn có xu hướng chi tiêu cho gia đình, trong khi người độc thân có thể ưu tiên trải nghiệm cá nhân và các sản phẩm tiện ích. Đáng chú ý, tỷ lệ người đã ly hôn và góa chiếm tỷ trọng lớn, mở ra cơ hội tiếp cận một nhóm khách hàng có xu hướng độc lập tài chính và nhiều tiềm năng chi tiêu. Việc nắm bắt đúng nhu cầu theo từng tình trạng hôn nhân sẽ giúp doanh nghiệp tối ưu hóa thông điệp truyền thông và thiết kế gói sản phẩm phù hợp với từng phân khúc.

## **3.5. Mức thu nhập (Income\_Level)**

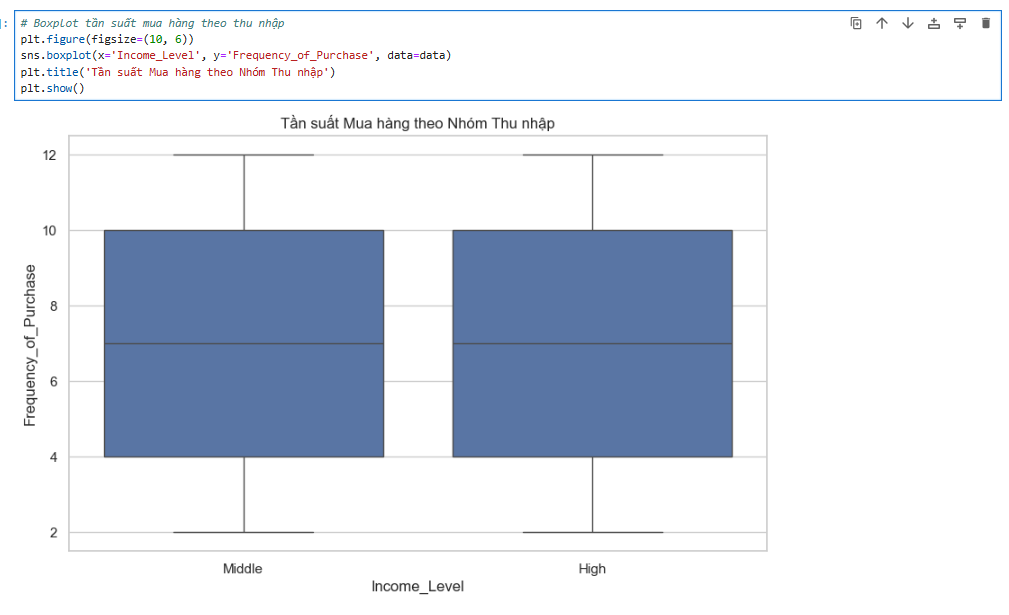
### **Mức thu nhập của khách hàng**

****

****

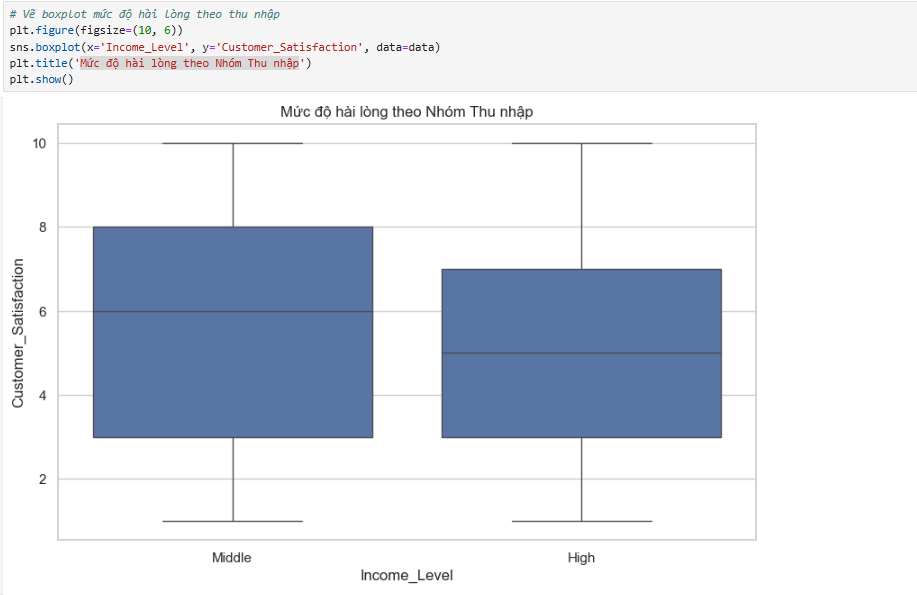
Biểu đồ thể hiện mức thu nhập của khách hàng chủ yếu rơi vào hai nhóm: “High” và “Middle”. Nhóm thu nhập cao (“High”) chiếm tỷ lệ nhỉnh hơn với hơn 510 người, trong khi nhóm thu nhập trung bình (“Middle”) có khoảng 490 người. Sự phân bố này tương đối đồng đều, cho thấy tệp khách hàng có sự đa dạng nhất định về khả năng chi tiêu. Điều này mở ra cơ hội để doanh nghiệp phân loại và xây dựng các chiến lược tiếp thị cá nhân hóa phù hợp với từng phân khúc thu nhập.

### **Tần suất Mua hàng theo Nhóm Thu nhập**



Biểu đồ boxplot về Tần suất mua hàng theo Nhóm Thu nhập cho thấy mức độ tương đồng đáng kể giữa hai nhóm Middle và High. Trung vị tần suất mua hàng của cả hai nhóm đều nằm ở mức trung bình khá (khoảng 7 lần), với phạm vi tần suất trải dài từ 2 đến 12 lần. Điều này chỉ ra rằng thu nhập không phải là yếu tố quyết định rõ ràng đến tần suất mua hàng – cả khách hàng thu nhập trung bình và cao đều có hành vi mua sắm thường xuyên tương đương. Từ đó, chiến lược tiếp cận nên tập trung nhiều hơn vào yếu tố hành vi và nhu cầu thực tế, thay vì chỉ dựa trên mức thu nhập.

### **Mức độ hài lòng theo Nhóm Thu nhập**

****

Biểu đồ boxplot thể hiện mức độ hài lòng của khách hàng theo nhóm thu nhập cho thấy sự khác biệt đáng chú ý giữa hai nhóm Middle và High. Trung vị mức độ hài lòng của nhóm thu nhập trung bình cao hơn so với nhóm thu nhập cao. Đồng thời, nhóm thu nhập trung bình cũng có mức độ phân tán rộng hơn, nghĩa là có nhiều khách hàng cực kỳ hài lòng, trong khi nhóm thu nhập cao lại tập trung quanh mức trung bình thấp hơn. Điều này có thể phản ánh kỳ vọng cao hơn từ nhóm thu nhập cao, dẫn đến mức độ hài lòng không cao như mong đợi. Doanh nghiệp nên chú trọng cải thiện trải nghiệm và giá trị dịch vụ cho nhóm khách hàng thu nhập cao để đáp ứng tốt hơn kỳ vọng của họ.

### **Kết luận:**

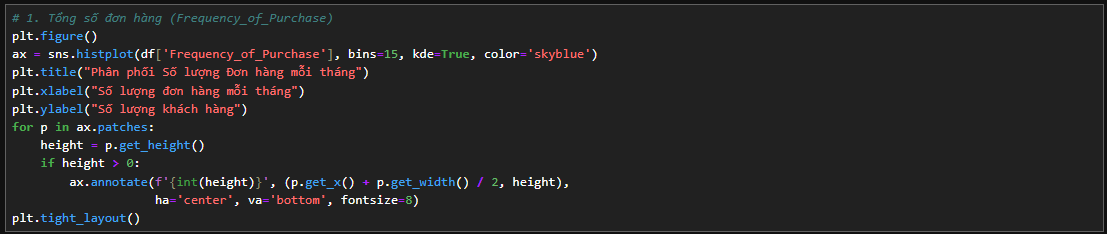
Qua quá trình phân tích dữ liệu khám phá, chúng ta đã thu được nhiều hiểu biết giá trị về đặc điểm nhân khẩu học và hành vi tiêu dùng của khách hàng. Độ tuổi khách hàng trải dài từ 18 đến 50, nổi bật là các nhóm 24–30, 34–37 và 46–50 – đại diện cho ba phân khúc tiềm năng: người trẻ thành thị, nhóm ổn định thu nhập và người tiêu dùng trung niên có tiềm lực tài chính. Phân tích theo giới tính cho thấy sở thích mua sắm khác biệt rõ rệt giữa nam và nữ ở nhiều danh mục, đồng thời làm nổi bật sự hiện diện ngày càng đa dạng của các nhóm giới tính phi nhị nguyên.

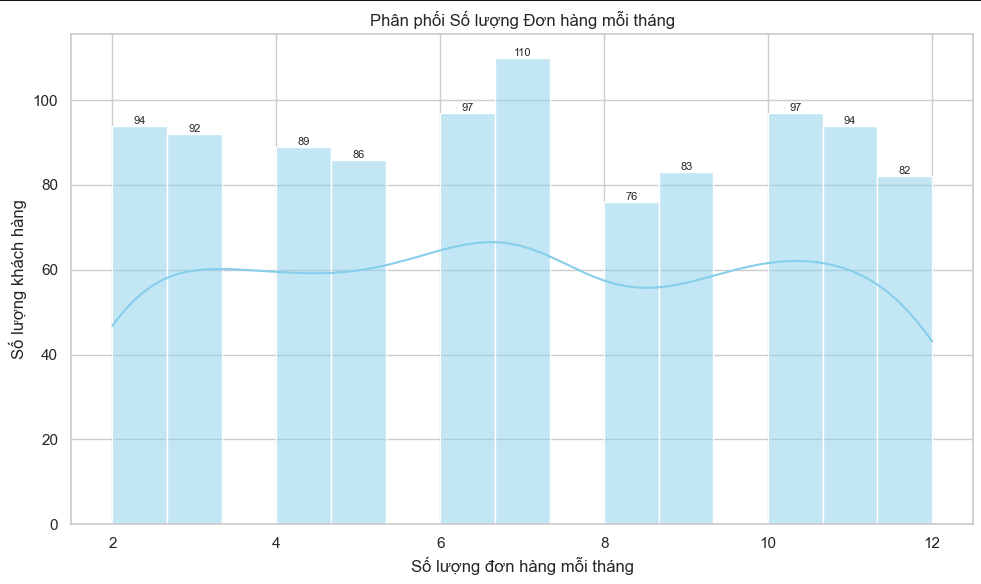
Bên cạnh đó, việc so sánh mức độ hài lòng và tần suất mua sắm theo thu nhập cho thấy khách hàng thu nhập trung bình và cao có hành vi khá tương đồng. Điều này khẳng định rằng các yếu tố như trải nghiệm cá nhân hóa, chất lượng dịch vụ và sản phẩm phù hợp sẽ đóng vai trò quan trọng hơn trong việc duy trì lòng trung thành và thúc đẩy mua sắm, thay vì chỉ tập trung vào thu nhập.

Những phát hiện từ EDA là nền tảng vững chắc để định hướng các bước phân tích tiếp theo, đặc biệt trong việc xây dựng mô hình phân khúc khách hàng và cá nhân hóa chiến lược tiếp thị hiệu quả hơn.

# **4.Hành vi mua sắm (Purchase Behavior)**

## **4.1. Tổng số đơn hàng (Frequency\_of\_Purchase)**

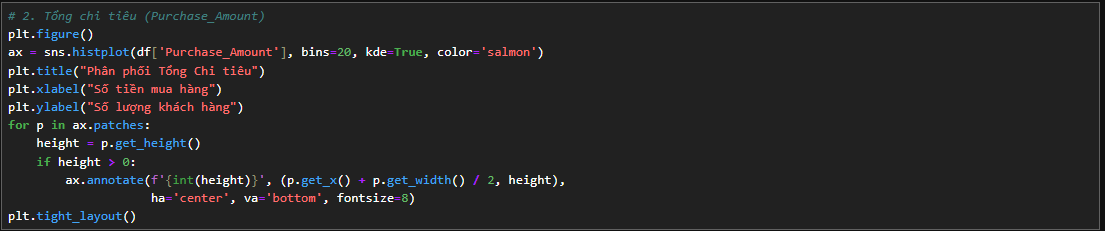


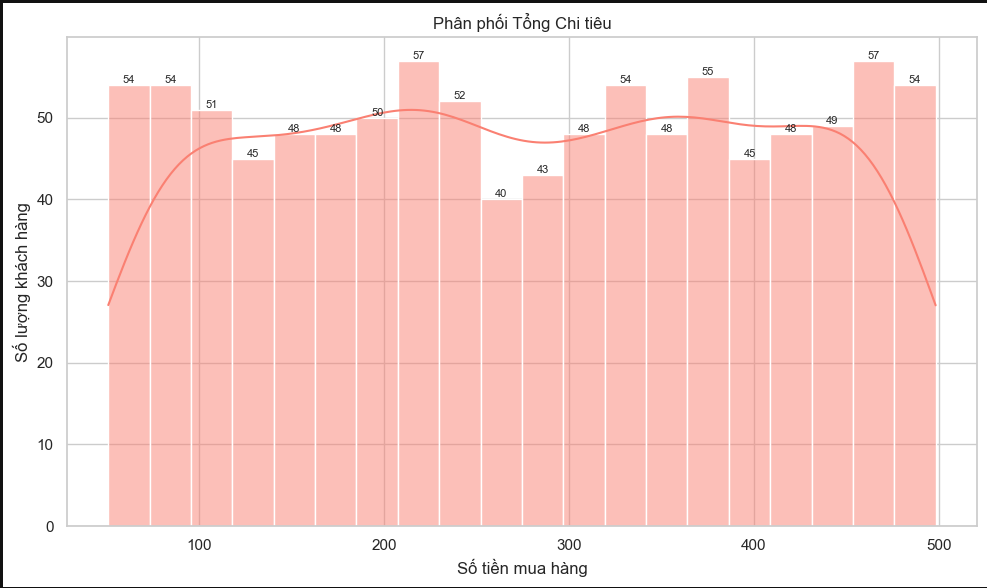


Biểu đồ cho thấy số lượng đơn hàng mà khách hàng thực hiện mỗi tháng khá đồng đều không có sự chênh lệch lớn với nhau, số lượng đơn hàng mỗi tháng có hơn 90 khách hàng thực hiện mua là những số lượng như 2-3, 6-7 và 10-11 đơn hàng, với đỉnh điểm là 7 đơn hàng có đến 110 người mua và 8 đơn là thấp nhất chỉ có 76 khách hàng. Các mức thấp hơn như 4,5,9 và 12 đơn hàng cũng có lượng khách khá lớn, từ 80–90 người. Tuy nhiên, từ mức 11 trở đi, số người giảm nhanh, và rất ít khách mua tới 15–20 đơn mỗi tháng. Phân phối nghiêng về phía số lượng mua trung bình.

Điều này phản ánh rằng phần lớn khách hàng có thói quen mua sắm tương đối định kỳ, nhưng không quá thường xuyên, phù hợp với các sản phẩm tiêu dùng vừa túi tiền hoặc có vòng đời sử dụng ngắn. Mức độ mua tập trung ở khoảng 6–7 đơn hàng cho thấy hành vi mua có kế hoạch, thay vì ngẫu nhiên. Nguyên nhân có thể đến từ việc người dùng chủ yếu mua hàng tiêu dùng hoặc sản phẩm phổ thông, không phải hàng cao cấp. Ngoài ra, những chương trình khuyến mãi thường xuyên và giao diện mua hàng dễ dùng cũng khuyến khích mua đều đặn nhưng không quá dồn dập.

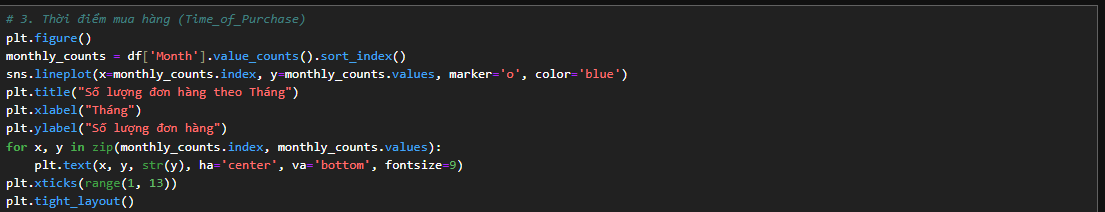
## **4.2. Tổng chi tiêu (Purchase\_Amount)**

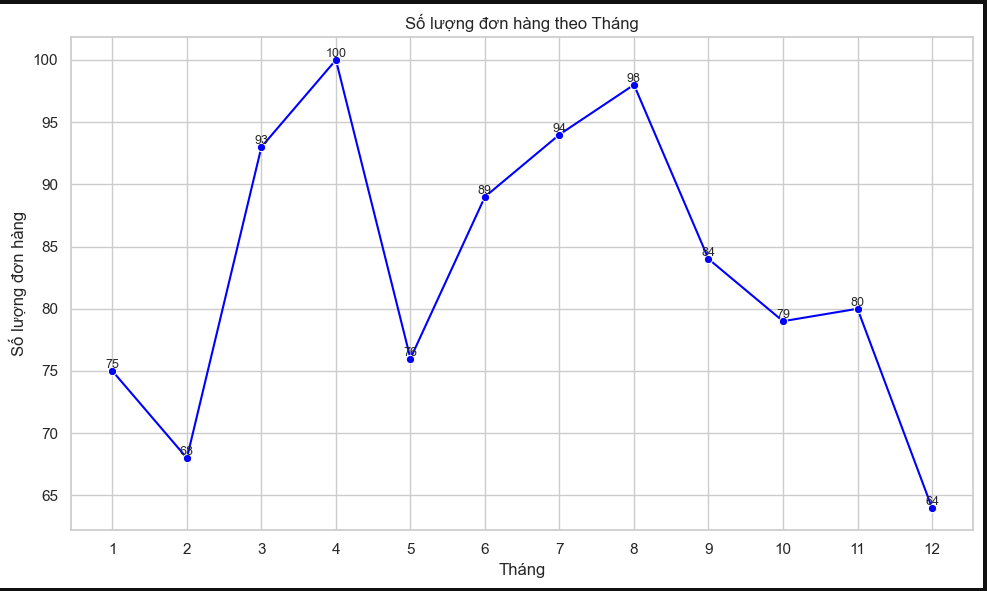


  
Biểu đồ cho thấy phần lớn khách hàng có tổng chi tiêu nằm trong khoảng 50 đến 500 đô la, với mật độ phân phối ở các khoảng 50–100, 200-250, 450-500 đô la. Các mức chi tiêu 100-200, 250-325, 400-450 đô la thì có sự thụt giảm đi so với các mức chi tiêu khác, và rất ít khách chi vượt quá 500 đô la, tạo ra một biểu đồ có sự nhấp nhô rõ ràng.

Hành vi này cho thấy phần lớn người mua có xu hướng chi tiêu trung bình và tiết kiệm, ít có những đơn hàng có giá trị quá cao. Nguyên nhân có thể đến từ việc khách hàng chủ yếu chọn những sản phẩm có giá phổ thông, phù hợp với ngân sách cá nhân hoặc hộ gia đình. Đồng thời, sự đa dạng của các mặt hàng trong khoảng giá này (điện tử, thời trang, thể thao…) đáp ứng đúng nhu cầu tiêu dùng. Việc chỉ có một nhóm nhỏ khách hàng chi tiêu cao cũng cho thấy nhóm “big spender” chưa phải là đối tượng chính hoặc các sản phẩm cao cấp chưa được khai thác mạnh.

## **4.3. Thời điểm mua hàng (Time\_of\_Purchase)**

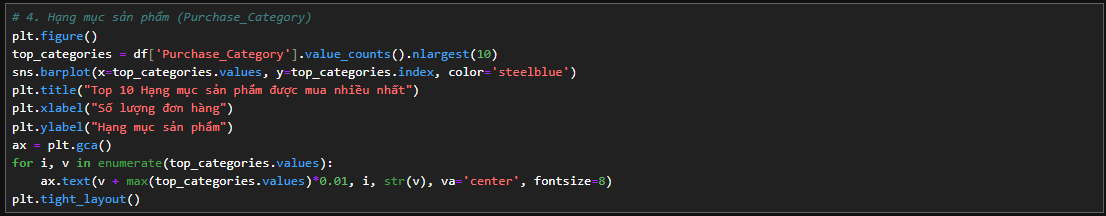


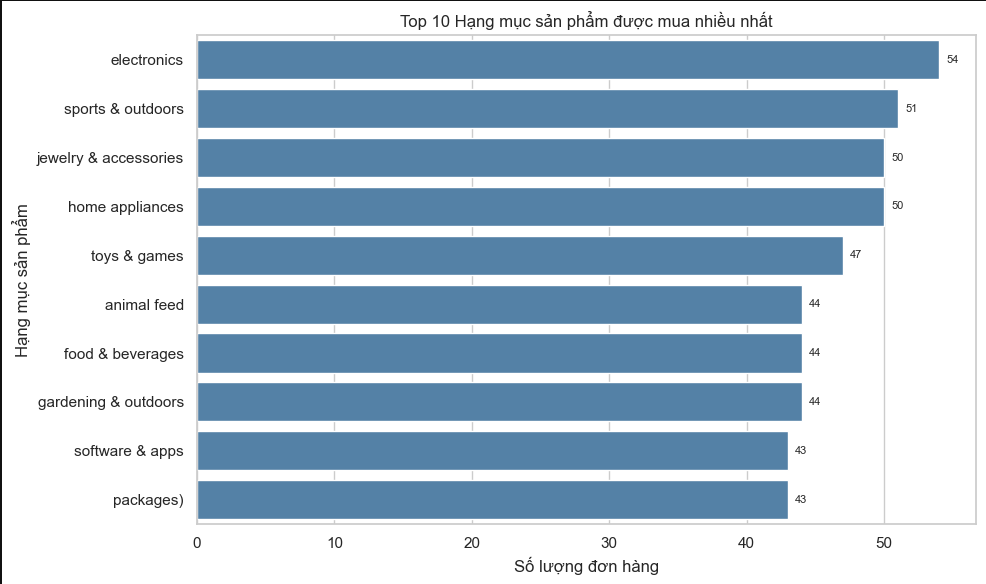


Biểu đồ thể hiện xu hướng số lượng đơn hàng theo tháng trong năm. Số lượng đơn hàng tăng mạnh từ tháng 2 đến tháng 4, trong đó tháng 4 đạt đỉnh với 100 đơn hàng. Sau đó, có sự dao động nhẹ và tiếp tục tăng trở lại vào tháng 7 và 8 (94 và 98 đơn hàng), rồi giảm dần từ tháng 9 đến tháng 12, với tháng 12 có số đơn hàng thấp nhất – chỉ 64 đơn. Nhìn chung, hoạt động mua sắm không quá đột biến nhưng thể hiện rõ những chu kỳ cao điểm vào mùa xuân và cuối mùa hè.

Phân bố số lượng đơn hàng theo tháng phản ánh rõ rệt các chu kỳ mua sắm theo mùa và lễ hội quốc tế, tương ứng với thói quen tiêu dùng tại các quốc gia như Hoa Kỳ, Anh, Ấn Độ, Đức và Canada – là những nơi chiếm phần lớn giao dịch trong dữ liệu. Cụ thể, tháng 3 và 4 ghi nhận số lượng đơn hàng tăng mạnh, có thể xuất phát từ các dịp lễ như Holi (Ấn Độ) hoặc Easter (Phục Sinh) ở các nước phương Tây – những thời điểm mà người tiêu dùng có xu hướng chi tiêu nhiều cho quà tặng, thời trang và sản phẩm gia dụng. Tương tự, tháng 7 và 8 cũng là giai đoạn mua sắm cao điểm khi trùng với mùa tựu trường (Back-to-School) tại Mỹ, châu Âu và Canada, cùng với sự kiện Amazon Prime Day vào tháng 7, thường thúc đẩy doanh số bán hàng toàn cầu. Ngược lại, tháng 12 lại có số đơn hàng thấp nhất, điều này trái với giả định thông thường về mùa lễ hội cuối năm. Tuy nhiên, nguyên nhân có thể đến từ việc người tiêu dùng đã hoàn tất việc mua sắm từ cuối tháng 11 với các sự kiện lớn như Black Friday và Cyber Monday, hoặc do lo ngại về thời gian vận chuyển trong kỳ nghỉ lễ nên không phát sinh nhiều đơn hàng mới

## **4.4. Hạng mục sản phẩm (Purchase\_Category)**

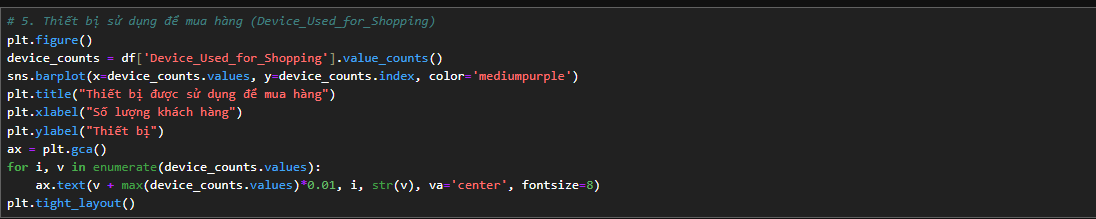


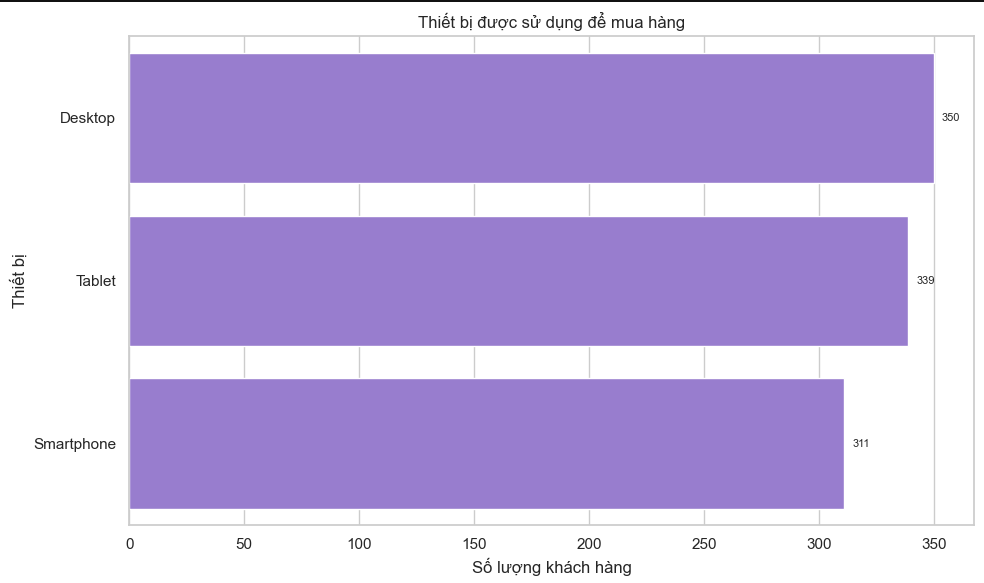


Biểu đồ cho thấy các sản phẩm thuộc nhóm công nghệ như **Electronics** chiếm tỷ lệ lớn nhất trong số các đơn hàng, chứng tỏ khách hàng rất ưa chuộng các thiết bị điện tử. Ngoài ra, các hạng mục liên quan đến thể thao, thời trang như **Sports & Outdoors** và **Jewelry & Accessories** cũng nằm trong top đầu, phản ánh nhu cầu đa dạng về sức khỏe và phong cách cá nhân. Các sản phẩm phục vụ gia đình và giải trí như **Home Appliances** và **Toys & Games** cũng có lượng đơn hàng cao, cho thấy người mua quan tâm đến việc nâng cao chất lượng cuộc sống và nhu cầu giải trí cho trẻ em. Những nhóm hàng thiết yếu như **Animal Feed** và **Food & Beverages** cũng giữ tỷ lệ ổn định trong tổng số đơn hàng. Đặc biệt, sự xuất hiện của **Software & Apps** cho thấy xu hướng chuyển dịch sang sản phẩm kỹ thuật số và dịch vụ trực tuyến đang ngày càng phát triển.

Sự phân bố đa dạng này phản ánh xu hướng tiêu dùng hiện nay, khi khách hàng không chỉ tập trung vào một nhóm sản phẩm mà mở rộng sang nhiều lĩnh vực khác nhau. Việc Electronics chiếm ưu thế có thể do sự phát triển mạnh mẽ của công nghệ và thói quen làm việc, học tập trực tuyến ngày càng phổ biến. Các nhóm sản phẩm thể thao và thời trang được ưa chuộng do người dùng ngày càng quan tâm đến sức khỏe và phong cách sống. Mặt hàng gia dụng và đồ chơi thể hiện nhu cầu cải thiện chất lượng cuộc sống gia đình và đáp ứng nhu cầu giải trí cho trẻ nhỏ. Các sản phẩm thiết yếu như thức ăn cho động vật và thực phẩm luôn giữ sức mua ổn định vì chúng là nhu cầu cơ bản hàng ngày. Sự gia tăng trong việc mua **Software & Apps** phản ánh xu hướng chuyển đổi số và ứng dụng công nghệ số ngày càng thâm nhập sâu vào đời sống. Ngoài ra, sự tiện lợi trong mua sắm online và khả năng tiếp cận đa dạng sản phẩm đã góp phần làm tăng sự phân bổ đa dạng này.

## **4.5. Thiết bị sử dụng (Device\_Used\_for\_Shopping)**

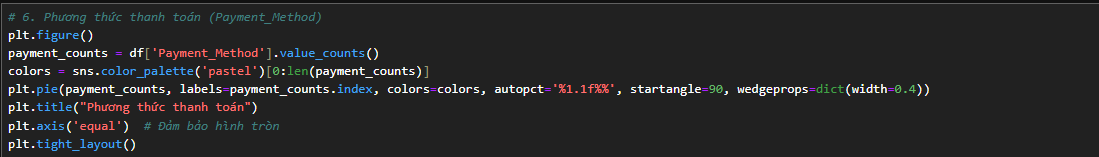


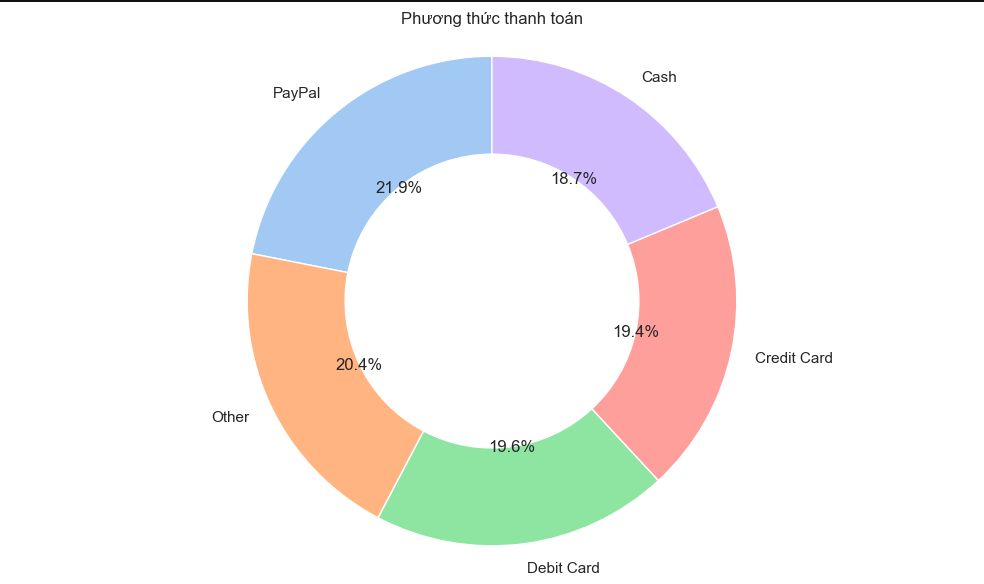


Biểu đồ cho thấy ba loại thiết bị chính được khách hàng sử dụng để thực hiện giao dịch mua sắm là **Desktop**, **Tablet** và **Smartphone** với số lượng khách hàng khá cân bằng. Desktop dẫn đầu với 350 khách hàng, tiếp theo là Tablet với 339 khách hàng, và Smartphone chiếm 311 khách hàng. Sự chênh lệch về số lượng giữa ba thiết bị không quá lớn, cho thấy người dùng có xu hướng mua sắm trên nhiều nền tảng thiết bị khác nhau thay vì tập trung vào một loại duy nhất. Điều này phản ánh sự đa dạng trong thói quen truy cập và tương tác của khách hàng với nền tảng thương mại điện tử.

Sự phân bố đồng đều về các thiết bị này xuất phát từ nhiều nguyên nhân. Desktop thường được sử dụng trong môi trường làm việc hoặc tại nhà, nơi khách hàng có thể dễ dàng thao tác mua sắm với màn hình lớn và bàn phím tiện lợi, phù hợp với các giao dịch phức tạp hoặc cần tìm hiểu kỹ sản phẩm. Tablet với tính di động cao, màn hình lớn hơn smartphone, được nhiều khách hàng lựa chọn vì sự tiện lợi khi vừa di chuyển vừa có trải nghiệm xem sản phẩm tốt hơn. Smartphone tuy màn hình nhỏ hơn nhưng nổi bật về tính di động và truy cập nhanh chóng, phục vụ nhu cầu mua sắm tức thì hoặc khi khách hàng không có nhiều thời gian. Sự đa dạng trong lựa chọn thiết bị cũng phản ánh sự phát triển mạnh mẽ của công nghệ và thói quen sử dụng đa thiết bị của người dùng hiện nay, khi họ linh hoạt chuyển đổi giữa các thiết bị tùy theo hoàn cảnh và nhu cầu cá nhân. Ngoài ra, sự phát triển của các ứng dụng và giao diện tối ưu trên từng loại thiết bị cũng góp phần nâng cao trải nghiệm người dùng và thúc đẩy việc mua sắm trên mọi nền tảng.

## **4.6. Phương thức thanh toán (Payment\_Method)**

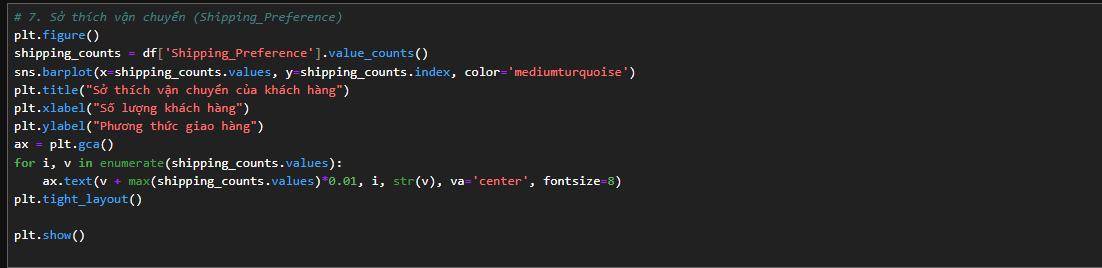


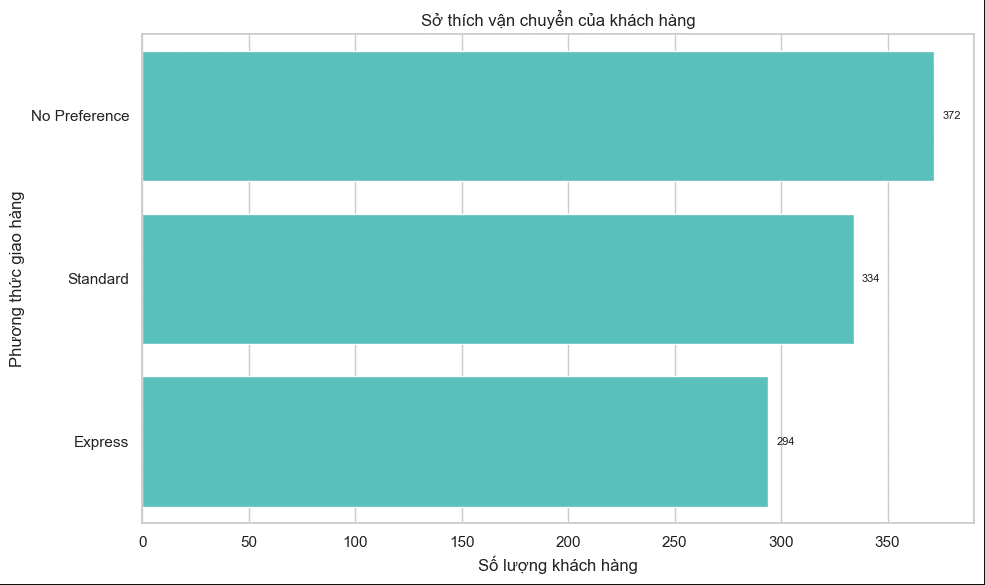


Biểu đồ phân bổ tỷ lệ các phương thức thanh toán cho thấy sự đa dạng và cân bằng khá đều giữa các hình thức phổ biến hiện nay. PayPal dẫn đầu với 21.9%, tiếp theo là các hình thức thanh toán thẻ Debit Card (19.6%) và Credit Card (19.4%), cùng với phương thức thanh toán bằng tiền mặt (Cash) chiếm 18.7%. Một phần đáng kể (20.4%) thuộc nhóm “Other” – có thể bao gồm các hình thức thanh toán khác như ví điện tử, chuyển khoản ngân hàng, hay các phương thức thanh toán mới nổi. Tỷ lệ tương đối đồng đều giữa các phương thức cho thấy không có một hình thức nào thực sự chiếm ưu thế áp đảo mà khách hàng có xu hướng đa dạng hóa cách thanh toán.

Sự phân bố này phản ánh thói quen tiêu dùng linh hoạt và đa dạng của khách hàng trên nền tảng thương mại điện tử. PayPal dẫn đầu bởi tính tiện lợi, bảo mật và phổ biến rộng rãi trên toàn cầu, phù hợp với các giao dịch online nhanh chóng. Thanh toán thẻ Debit và Credit được sử dụng nhiều do sự tiện lợi và các chính sách ưu đãi của ngân hàng, đồng thời đây là những phương thức thanh toán quen thuộc và an toàn đối với nhiều người dùng. Tiền mặt vẫn chiếm tỷ trọng đáng kể do một phần khách hàng có thể ưu tiên thanh toán khi nhận hàng (COD) hoặc hạn chế sử dụng các hình thức thanh toán điện tử do thói quen hoặc hạn chế về công nghệ. Nhóm “Other” thể hiện sự xuất hiện của các phương thức mới hoặc các ví điện tử ngày càng phổ biến, đặc biệt trong bối cảnh thanh toán không tiền mặt đang được khuyến khích phát triển. Ngoài ra, sự đa dạng về phương thức thanh toán giúp tạo điều kiện thuận lợi cho nhiều đối tượng khách hàng khác nhau, từ người thích công nghệ đến người vẫn quen với hình thức truyền thống, qua đó góp phần tăng khả năng tiếp cận và cải thiện trải nghiệm mua sắm.

## **4.7. Sở thích vận chuyển (Shipping\_Preference)**





Biểu đồ thể hiện rõ ràng rằng đa số khách hàng không có sở thích vận chuyển cụ thể, với 372 người chọn “No Preference”. Tiếp theo là nhóm khách hàng chọn phương thức giao hàng tiêu chuẩn (Standard) với 334 người, và số lượng ít hơn một chút chọn giao hàng nhanh (Express) với 294 người. Sự chênh lệch giữa các nhóm không quá lớn, cho thấy khách hàng có sự phân bổ tương đối đều về các lựa chọn phương thức giao hàng, nhưng phần lớn có xu hướng thoải mái, không đặt nặng yêu cầu về thời gian giao hàng.

Sự phân bổ này phản ánh tâm lý và thói quen mua sắm của khách hàng hiện nay. Nhóm “No Preference” chiếm đa số có thể do nhiều khách hàng không quá quan trọng về thời gian nhận hàng, họ ưu tiên sản phẩm và giá cả hơn là tốc độ vận chuyển. Phương thức giao hàng tiêu chuẩn được nhiều người lựa chọn bởi giá cước thường thấp hơn so với giao hàng nhanh, phù hợp với nhu cầu tiết kiệm chi phí trong khi vẫn đảm bảo thời gian nhận hàng hợp lý. Giao hàng nhanh tuy hấp dẫn nhờ tốc độ giao nhận, nhưng chi phí cao hơn có thể là rào cản với một bộ phận khách hàng, dẫn đến số lượng người chọn phương thức này thấp hơn. Ngoài ra, sự phát triển của hệ thống logistics và mạng lưới giao hàng hiện đại giúp nâng cao chất lượng dịch vụ vận chuyển, khiến khách hàng cảm thấy thoải mái hơn khi không phải chọn phương thức cụ thể. Thói quen này cũng có thể chịu ảnh hưởng từ việc nhiều nhà bán hàng áp dụng chính sách giao hàng miễn phí hoặc giá ưu đãi cho các đơn hàng, làm giảm bớt sự khác biệt về chi phí giữa các lựa chọn vận chuyển.

## **4.8. Kết luận:**

Dựa trên các phân tích, có thể rút ra rằng thói quen tiêu dùng hiện nay mang tính cẩn trọng, linh hoạt và phụ thuộc công nghệ. Khách hàng mua sắm với tần suất vừa phải, chi tiêu ở mức trung bình, chủ yếu lựa chọn các sản phẩm thiết thực như thời trang, điện tử, tiêu dùng cá nhân. Họ thường mua vào những dịp lễ hội hay những tháng tựu trường, phản ánh việc ưu tiên mua khi cần thiết, hạn chế mua vào những dịp cuối năm

Về phương tiện, người dùng ưu tiên desktop, cho thấy hành vi tiêu dùng di động chiếm ưu thế. Hình thức thanh toán rất đa dạng, nhưng phần lớn dùng PayPal và thẻ, phản ánh sự tin tưởng vào thanh toán số. Phần lớn khách hàng không yêu cầu đặc biệt về vận chuyển, cho thấy họ sẵn sàng đợi lâu hơn nếu tiết kiệm chi phí.

Nhìn chung, các yếu tố ảnh hưởng đến quyết định mua hàng gồm: giá cả, thời điểm khuyến mãi, tính tiện lợi của thiết bị di động, phương thức thanh toán đa dạng và dịch vụ giao hàng hợp lý. Người tiêu dùng hiện đại hành động có cân nhắc, đề cao sự thuận tiện và an toàn trong mua sắm.

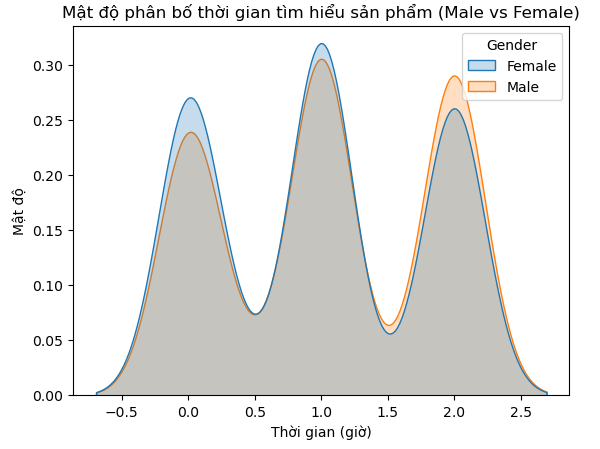
# **5. Phần phân tích mức độ tương tác và hành vi của khách hàng đối với sản phẩm và quảng cáo**

Trong thời đại số, việc phân tích tương tác và hành vi của khách hàng khi tiếp cận với quảng cáo trở nên cực kỳ quan trọng. Những hành động như nhấp vào quảng cáo, thời gian dừng lại trên nội dung, hay mức độ tương tác với thương hiệu đều phản ánh mức độ quan tâm thực sự của họ. Thông qua đó, chúng ta có thể xác định ai là người tiếp cận quảng cáo nhiều nhất và mức độ gắn kết của họ với thông điệp truyền tải. Và giờ, ta sẽ làm rõ điều này qua các biểu đồ phân tích sau đây để hiểu rõ hơn về hành vi tương tác của khách hàng.

Đầu tiên, có thể rất nhiều người đã từng đặt câu hỏi: “Liệu nam và nữ có sự khác biệt trong cách họ dành thời gian để tìm hiểu sản phẩm trước khi mua hay không?”

Để trả lời cho câu hỏi này, ta cùng phân tích biểu đồ mật độ phân bố thời gian tìm hiểu sản phẩm theo giới tính (Male vs Female).

## **5.1. Phân tích biểu đồ mật độ phân bố thời gian tìm hiểu sản phẩm (Male vs Female)**



Từ biểu đồ, ta có thể rút ra một số nhận định đáng chú ý về hành vi tìm hiểu sản phẩm của khách hàng:

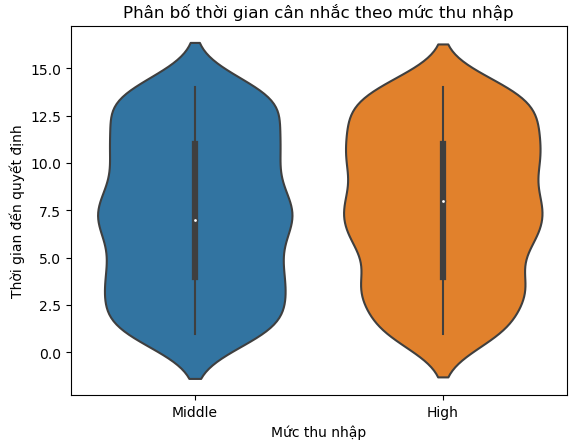
1. Đa số khách hàng không dành quá nhiều thời gian để tìm hiểu sản phẩm Có một mật độ rất cao tập trung quanh mức 0 giờ — điều này cho thấy phần lớn người tiêu dùng ra quyết định khá nhanh, hoặc gần như không tìm hiểu trước khi mua.
2. Tuy nhiên, khi đã tìm hiểu, họ thường dành hẳn hàng giờ để nghiên cứu kỹ Biểu đồ có thêm hai đỉnh rõ rệt tại khoảng 1 giờ và 2 giờ, cho thấy có những nhóm khách hàng cụ thể dành một lượng lớn thời gian để tìm hiểu, có thể là những người kỹ tính, mua hàng giá trị cao, hoặc thiên về lý trí.
3. So sánh nam và nữ: không quá chênh lệch, nhưng có xu hướng khác biệt nhẹ

- Trong khoảng dưới 1.5 giờ, đường mật độ của nữ cao hơn → nữ giới thường là những người dành thời gian ngắn nhưng vẫn tìm hiểu kỹ lưỡng.

- Ngược lại, từ sau mốc 1.5 giờ trở đi, mật độ của nam bắt đầu nhỉnh hơn, cho thấy một nhóm nam giới dành thời gian nghiên cứu rất kỹ trước khi quyết định.

- Bên cạnh đó thì việc cân nhắc có nên mua một sản phẩm hay không thì ta cũng rất hay đắn đo suy nghĩ, dù là đã tìm hiểu kỹ nhưng hầu như mọi người sẽ có xu hướng cân nhắc lại một lần nữa, vì thế ta nên tìm hiểu về vấn đề đó bằng biểu đồ dưới đây

Phân tích biểu đồ violin: **Thời gian cân nhắc mua hàng theo mức thu nhập**



Nhận định chi tiết:

1. Phân bố thời gian từ 0–5 giờ: Phần đáy của violin (0–5 giờ) của nhóm thu nhập Middle phình to hơn, cho thấy nhiều khách hàng ở nhóm này quyết định nhanh hơn - họ dành ít thời gian để cân nhắc.
2. Phân bố từ 5–7.5 giờ: Cả hai nhóm đều có sự phình ra, nhưng từ mốc 5 giờ trở đi, nhóm High lại chiếm ưu thế rõ rệt với phần thân violin rộng hơn, phản ánh nhiều người thu nhập cao có xu hướng dành nhiều thời gian hơn để suy nghĩ trước khi mua.
3. Boxplot bên trong violin: Trung vị (median) của nhóm High cao hơn so với nhóm Middle, cho thấy trung bình người có thu nhập cao cân nhắc lâu hơn.

Người tiêu dùng có thu nhập cao thường dành nhiều thời gian hơn để cân nhắc trước khi mua hàng. Điều này có thể phản ánh sự cẩn trọng, kỳ vọng cao hơn về sản phẩm, hoặc liên quan đến việc họ mua các mặt hàng giá trị lớn hơn.

Hầu như hiện nay các doanh nghiệp đều tập trung vào quảng cáo để thu hút thêm người dùng, vậy ta sẽ đi tìm hiểu thêm vài vấn đề giữa người dùng và quảng cáo để biết chi tiết hơn, liệu quảng cáo có thật sự hiệu quả? Và người dùng cảm thấy như thế nào giữa quảng cáo và sản phẩm thực tế?

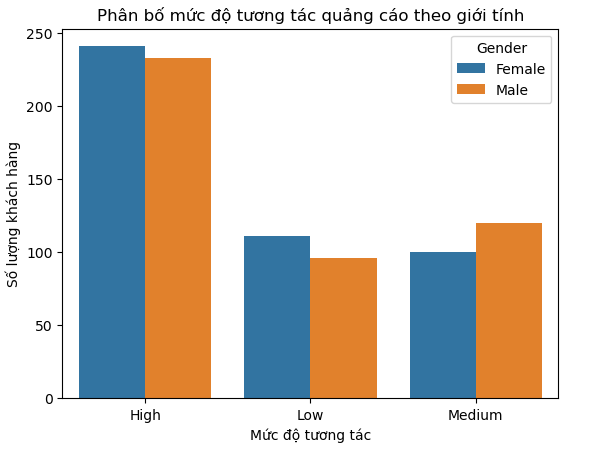
## **5.2. Phân tích biểu đồ mức độ tương tác với quảng cáo theo giới tính**

Biểu đồ grouped bar chart thể hiện sự khác biệt về mức độ tương tác với quảng cáo (Engagement\_with\_Ads) giữa hai giới tính: Male và Female.

- Trục X: Phân loại mức độ tương tác thành 3 nhóm: Low, Medium, High.

- Trục Y: Số lượng khách hàng trong từng nhóm.

- Mỗi nhóm được chia thành 2 cột theo giới tính: nữ (màu xanh) và nam (màu cam).



Nhận xét sơ qua ta thấy được các điều sau:

1. Tương tác cao (High): Số lượng nữ có mức tương tác cao nhỉnh hơn nam, cho thấy phụ nữ có xu hướng dễ bị thu hút hơn bởi quảng cáo khi cảm thấy hứng thú.
2. Tương tác trung bình (Medium): Ở mức này, nam lại chiếm ưu thế nhẹ, cho thấy họ có thể thường xuyên tiếp xúc với quảng cáo nhưng chưa chắc đã bị thuyết phục mạnh.

3. Tương tác thấp (Low): Bất ngờ là số lượng nữ trong nhóm tương tác thấp lại cao hơn nam, có thể phản ánh rằng không phải tất cả phụ nữ đều quan tâm đến quảng cáo - sự quan tâm có tính phân cực hơn.

Tổng thể:

- Sự chênh lệch giữa nam và nữ ở từng mức tương tác là không lớn, cho thấy quảng cáo ảnh hưởng đến cả hai giới tương đối đồng đều, tuy có một số xu hướng nhỏ theo từng mức độ.

- Phụ nữ có xu hướng nằm ở hai thái cực — hoặc rất quan tâm, hoặc gần như không để ý đến quảng cáo. Trong khi đó, nam giới phân bổ đều hơn, đặc biệt ở nhóm tương tác trung bình.

Vì thế việc phân loại nội dung quảng cáo phù hợp với từng nhóm là vô cùng quan trọng

Ví dụ:

- Nữ giới: Tăng cường quảng cáo cảm xúc và hình ảnh bắt mắt để thu hút nhóm tương tác cao, đồng thời cần cải thiện nội dung với nhóm ít quan tâm.

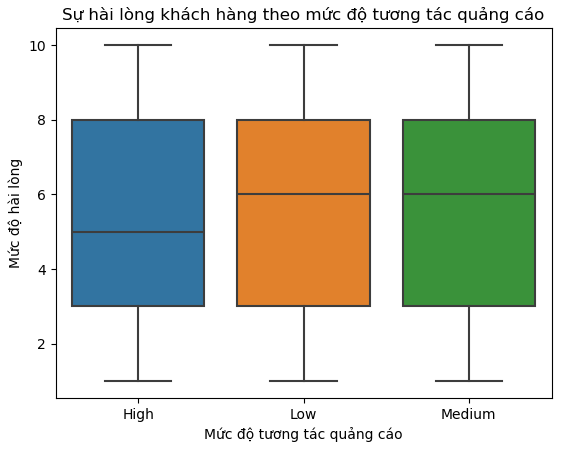
- Nam giới: Nội dung nên rõ ràng, trực tiếp, dễ tiếp cận để giữ sự chú ý và thúc đẩy chuyển đổi từ nhóm tương tác trung bình.

## **5.3. Phân tích mức độ hài lòng của khách hàng theo mức độ tương tác với quảng cáo**

Biểu đồ sử dụng boxplot để thể hiện sự phân bố điểm hài lòng của khách hàng (Customer Satisfaction, thang điểm 1–10) theo từng mức độ tương tác với quảng cáo (Engagement\_with\_Ads) gồm ba nhóm:

- Trục X: Nhóm tương tác Low, Medium, High.

- Trục Y: Mức độ hài lòng với sản phẩm.



Nhận xét chi tiết:

- Các hộp boxplot giữa các nhóm có chiều cao và phạm vi phân bố gần như tương tự nhau, cho thấy sự phân tán điểm hài lòng là tương đối giống nhau giữa các nhóm.

- Tuy nhiên, điểm khác biệt đáng chú ý nằm ở đường trung vị (median):

- Nhóm Low và Medium đều có median ở khoảng mức 6, cho thấy khách hàng trong hai nhóm này nhìn chung có mức độ hài lòng trung bình-khá.

- Trong khi đó, nhóm High lại có median thấp hơn, khoảng mức 5, cho thấy khách hàng tương tác quảng cáo nhiều hơn lại không hài lòng bằng nhóm còn lại.

Quảng cáo có thể tạo kỳ vọng cao, nhưng nếu sản phẩm không đáp ứng được thì dẫn đến sự thất vọng. Hoặc khách hàng tương tác nhiều với quảng cáo là những người còn đang cân nhắc, hoài nghi hơn với sản phẩm, và sau cùng có trải nghiệm không tốt như kỳ vọng.

Qua đó ta thấy các chiến lược quảng cáo cần đảm bảo sự nhất quán giữa thông điệp quảng cáo và chất lượng sản phẩm, tránh việc “thổi phồng” làm kỳ vọng vượt thực tế. Với nhóm tương tác cao, doanh nghiệp có thể đầu tư vào trải nghiệm sau mua, như hỗ trợ khách hàng, hậu mãi… để cải thiện mức độ hài lòng.

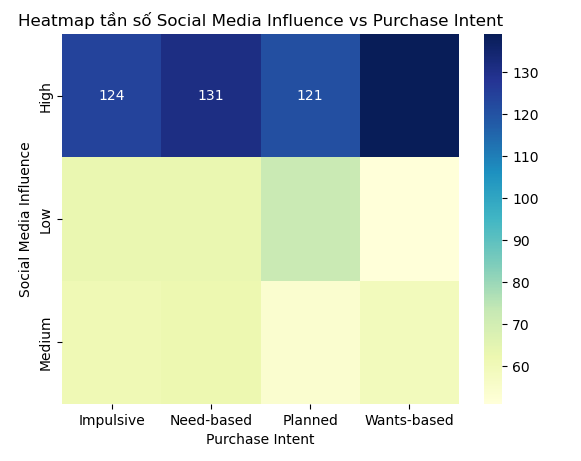
## **5.4. Phân tích mối liên hệ giữa ảnh hưởng mạng xã hội và mục đích mua hàng**

Biểu đồ thể hiện mức độ ảnh hưởng của mạng xã hội đến mục đích mua hàng

Biểu đồ là một heatmap (bản đồ nhiệt) thể hiện tần số khách hàng theo 2 chiều:

- Trục Y: Mức độ ảnh hưởng bởi mạng xã hội (Social Media Influence) - gồm các mức: Low, Medium, High.

- Trục X: Mục đích mua hàng (Purchase Intent) - gồm các nhóm: Impulsive: Mua bốc đồng, Need-based: Mua vì nhu cầu thiết yếu, Planned: Mua có kế hoạch trước, Wants-based: Mua vì mong muốn/cảm xúc cá nhân



Màu sắc càng đậm thể hiện số lượng khách hàng trong nhóm đó càng nhiều.

- Ở mức ảnh hưởng mạng xã hội thấp (Low): Khách hàng chủ yếu mua theo kiểu Planned (có kế hoạch). Điều này cho thấy khi ít bị ảnh hưởng bởi mạng xã hội, khách hàng thường mua sắm có chủ đích, suy nghĩ kỹ hơn.

- Ở mức Medium: Phân bố khá đều giữa các kiểu mua, không có nhóm nào vượt trội. Điều này cho thấy mức ảnh hưởng trung bình thì khó dự đoán xu hướng rõ ràng.

- Ở mức ảnh hưởng cao (High): Số lượng khách hàng mua theo kiểu Wants-based (mong muốn/cảm xúc) là cao nhất. Ngược lại, nhóm Planned lại rất thấp, gần như ít có khách hàng lập kế hoạch rõ ràng

Điều này cho thấy khi bị mạng xã hội ảnh hưởng mạnh, khách hàng dễ bị cảm xúc chi phối, dễ mua theo cảm hứng hoặc mong muốn hơn là theo nhu cầu thực tế.

Mạng xã hội càng ảnh hưởng mạnh, hành vi mua càng thiên về cảm xúc và ít tính toán. Điều này mở ra cơ hội cho các chiến dịch quảng bá cảm xúc, truyền cảm hứng, đặc biệt trên nền tảng mạng xã hội.

## **5.5. Kết Luận**

Dựa trên các biểu đồ đã phân tích, có thể kết luận rằng hành vi tương tác với quảng cáo và hành vi mua hàng của khách hàng chịu ảnh hưởng bởi nhiều yếu tố như giới tính, thu nhập, thời gian cân nhắc, và mức độ tác động từ mạng xã hội.

Cụ thể, nữ giới có xu hướng tương tác với quảng cáo nhiều hơn nam ở cả mức thấp và cao, trong khi nam giới lại chiếm ưu thế ở mức trung bình. Tuy nhiên, tương tác cao với quảng cáo không đồng nghĩa với sự hài lòng cao. Trái lại, nhóm khách hàng có mức độ tương tác cao lại ghi nhận điểm hài lòng thấp hơn nhóm còn lại, cho thấy rằng quảng cáo nếu không đi đôi với trải nghiệm thực tế tốt thì dễ gây thất vọng.   
 Về hành vi mua hàng, nam giới thường dành thời gian tìm hiểu sản phẩm lâu hơn, cho thấy sự cân nhắc kỹ lưỡng hơn so với nữ giới. Người có thu nhập cao cũng có xu hướng ra quyết định chậm hơn, chứng tỏ họ thận trọng hơn trước khi mua. Đáng chú ý, khách hàng bị ảnh hưởng mạnh bởi mạng xã hội lại có xu hướng mua sắm theo cảm xúc hoặc mong muốn (wants-based), trong khi người ít bị ảnh hưởng lại có hành vi mua hàng có kế hoạch rõ ràng hơn (planned).

Từ đó, có thể thấy rằng việc xây dựng chiến lược marketing hiệu quả không chỉ dựa vào việc tạo quảng cáo bắt mắt mà cần hiểu rõ tâm lý, thói quen và đặc điểm hành vi của từng nhóm khách hàng. Quảng cáo cần được thiết kế phù hợp với từng nhóm mục tiêu, đồng thời đảm bảo chất lượng sản phẩm/dịch vụ để giữ vững sự hài lòng và lòng trung thành của khách hàng.

# **6.Phân cụm khách hàng**

## **6.1. Chuẩn bị**

|  |
| --- |
| ## 1. Import thư viện và load dữ liệu  import pandas as pd  from sklearn.preprocessing import StandardScaler, OneHotEncoder  from sklearn.cluster import KMeans, DBSCAN  from sklearn.decomposition import PCA  from sklearn.manifold import TSNE  import matplotlib.pyplot as plt |

|  |
| --- |
| # 1. Load dữ liệu  df = pd.read\_csv(r'C:\Users\nguye\Documents\Zalo Received Files\Nhóm 15 Final Project\Cleaned\_Ecommerce\_Data.csv')  df.columns = df.columns.str.strip()  # loại bỏ khoảng trắng thừa có thể có  print("Shape:", df.shape)  print("Columns:", df.columns.tolist())  df.head() |

**Kết quả:**

**A screenshot of a computer

AI-generated content may be incorrect.**

## **6.2. Tính/Chuẩn bị các đặc trưng**

|  |
| --- |
| # Aggregate theo khách hàng  agg = (      df.groupby('Customer\_ID')        .agg(            Total\_Orders = ('Customer\_ID', 'count'),            Total\_Spending = ('Purchase\_Amount', 'sum'),            Frequency\_of\_Purchase = ('Frequency\_of\_Purchase', 'mean'),            Age = ('Age', 'first'),            Gender = ('Gender', 'first')        )        .reset\_index(drop=True)  )  agg.head() |

**Kết quả:**

**A screenshot of a computer

AI-generated content may be incorrect.**

- Tổng chi tiêu: sum của `Purchase\_Amount`

- Tần suất mua: lấy từ `Frequency\_of\_Purchase`

- Age (tuổi), Gender (giới tính)

## **6.3. One-hot encode cột Gender**

|  |
| --- |
| # 3. One-hot encode cột Gender  ohe = OneHotEncoder(sparse\_output=False, drop='if\_binary')  gender\_ohe = ohe.fit\_transform(agg[['Gender']])  gender\_cols = ohe.get\_feature\_names\_out(['Gender'])  df\_gender = pd.DataFrame(gender\_ohe, columns=gender\_cols, index=agg.index) |

- Ta cần chuyển Gender thành biến nhị phân (0/1) để mô hình Thuật toán phân cụm (như KMeans, DBSCAN) hiểu được.

- Male sẽ là 1, Female là 0

## **6.4. Kết hợp features**

Tạo một bảng dữ liệu X gồm các đặc trưng số đã chọn, chuẩn bị cho bước chuẩn hóa và phân cụm (KMeans/DBSCAN).

|  |
| --- |
| X = pd.concat([      agg[['Total\_Orders','Total\_Spending','Frequency\_of\_Purchase','Age']],      df\_gender  ], axis=1)  display(X.head()) |

Kết quả:

A screenshot of a computer screen

AI-generated content may be incorrect.

## **6.5. Chuẩn hóa dữ liệu**

Chuẩn hóa dữ liệu để:

Đưa các đặc trưng (feature) về cùng một thang đo (scale).

Tránh hiện tượng đặc trưng có đơn vị lớn (ví dụ: "Total\_Spending") chi phối kết quả phân cụm so với đặc trưng nhỏ (ví dụ: "Age", "Gender\_Male").

Là bắt buộc nếu bạn dùng các thuật toán phân cụm dựa trên khoảng cách, như KMeans hoặc DBSCAN.

|  |
| --- |
| # 5.  scaler = StandardScaler()  X\_scaled = scaler.fit\_transform(X) |

## **6.6. Áp dụng KMeans**

|  |
| --- |
| k = 4  kmeans = KMeans(n\_clusters=k, random\_state=42)  labels\_km = kmeans.fit\_predict(X\_scaled)  agg['Cluster\_KM'] = labels\_km  print("KMeans cluster sizes:")  print(pd.Series(labels\_km).value\_counts().sort\_index()) |

Kết quả:  
A screenshot of a computer program

AI-generated content may be incorrect.

### **6.6.1 Tính profile trung bình của mỗi cụm**

|  |
| --- |
| profile = agg.groupby('Cluster\_KM').agg({      'Total\_Orders':'mean',      'Total\_Spending':'mean',      'Frequency\_of\_Purchase':'mean',      'Age':'mean'  }).round(2)  print("Cluster profiling:\n", profile)  # 6c. Đặt tên cho từng cụm dựa trên profile  name\_map = {      0: 'High-Spend, Low-Freq',      1: 'Low-Spend, Low-Freq',      2: 'Medium-Spend, High-Freq',      3: 'High-Spend, High-Freq'  }  agg['Segment'] = agg['Cluster\_KM'].map(name\_map) |

Kết quả:  
A screenshot of a computer screen

AI-generated content may be incorrect.

### **6.6.2. Giảm chiều với PCA 2D**

|  |
| --- |
| pca = PCA(n\_components=2, random\_state=42)  pc = pca.fit\_transform(X\_scaled)  agg['PC1'], agg['PC2'] = pc[:,0], pc[:,1]  # 6e. Tọa độ centroids trên không gian PCA  centroids\_2d = pca.transform(kmeans.cluster\_centers\_) |

- Giảm chiều dữ liệu chuẩn hóa (X\_scaled) xuống 2 chiều (2D) bằng PCA.

- Vẽ biểu đồ phân cụm theo kết quả của thuật toán KMeans.

### **6.6.3. Vẽ scatter với centroids và tên segment**

|  |
| --- |
| plt.figure(figsize=(10,6))  colors = ['tab:blue','tab:orange','tab:green','tab:red']  for c, color in zip(sorted(agg['Cluster\_KM'].unique()), colors):      pts = agg[agg['Cluster\_KM']==c]      plt.scatter(pts['PC1'], pts['PC2'], label=name\_map[c], alpha=0.6, color=color)  # Centroids  #plt.scatter(centroids\_2d[:,0], centroids\_2d[:,1],              #marker='X', s=200, c='black', label='Centroids')  # Đẩy legend ra ngoài  plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)  plt.title('Customer Segments (KMeans + PCA)')  plt.xlabel('PC1')  plt.ylabel('PC2')  plt.grid(True)  plt.tight\_layout()  plt.show() |

A screen shot of a graph

AI-generated content may be incorrect.

Biểu đồ trên mô tả kết quả phân cụm KMeans (k=4) sau khi giảm chiều về 2D bằng PCA, với mỗi cụm đã được gán tên (“segment”) dựa trên profile trung bình:

- High-Spend, Low-Freq (xanh dương):

Vị trí: phân tán bên trái, PC1 thấp, PC2 quanh 0

Đặc trưng: chi tiêu cao nhất, tần suất mua thấp nhất

Đề xuất: đây là nhóm “big spender” nhưng mua không thường xuyên.

Tạo ưu đãi định kỳ (ví dụ voucher tháng),

Mời tham gia membership cao cấp,

Gửi reminder hoặc sản phẩm “best-seller” để kích hoạt mua thêm.

- Low-Spend, Low-Freq (cam):

Vị trí: bên phải, PC1 cao, PC2 quanh 0

Đặc trưng: chi tiêu thấp, tần suất cũng thấp

Đề xuất: nhóm khách lẻ, chưa có gắn kết.

Khuyến mãi giảm giá sâu cho mặt hàng phù hợp,

Chiến dịch gửi ưu đãi chào mừng,

Khảo sát nhanh để hiểu nguyên nhân họ ít quay lại.

- Medium-Spend, High-Freq (xanh lá):

Vị trí: cụm nhỏ, PC1 gần 0, PC2 cao (~3)

Đặc trưng: chi tiêu trung bình, mua rất đều đặn

Đề xuất: nhóm “regular shopper”.

Cross-sell / Upsell sản phẩm bổ sung,

Khuyến khích dùng subscription (dịch vụ định kỳ),

Tặng điểm thưởng cho mỗi lần mua.

- High-Spend, High-Freq (đỏ):

Vị trí: xung quanh cụm xanh lá nhưng tách ra chút

Đặc trưng: vừa chi tiêu lớn, vừa mua nhiều

Đề xuất: VIP customers.

Ưu tiên dịch vụ cá nhân hóa,

Mời tham gia event exclusive,

Quà tặng sinh nhật, access test sản phẩm mới.

## **6.7. Áp dụng DBSCAN**

|  |
| --- |
| dbscan = DBSCAN(eps=1.5, min\_samples=5)  labels\_db = dbscan.fit\_predict(X\_scaled)  agg['Cluster\_DB'] = labels\_db  print("\nDBSCAN cluster sizes (–1 = noise):")  print(pd.Series(labels\_db).value\_counts().sort\_index()) |

A screenshot of a computer program

AI-generated content may be incorrect.

|  |
| --- |
| # 1. Tính profile trung bình mỗi cụm DBSCAN  profile\_db = agg.groupby('Cluster\_DB').agg({      'Total\_Orders': 'mean',      'Total\_Spending': 'mean',      'Frequency\_of\_Purchase': 'mean',      'Age': 'mean'  }).round(2)  print("DBSCAN cluster profile:")  display(profile\_db)  # 2. Đặt tên cho từng cụm (tự chỉnh theo kết quả profile\_db của bạn)  name\_map\_db = {      -1: 'Noise/Unclassified',       0: 'High-Spend, Low-Freq',       1: 'Low-Spend, Low-Freq',       2: 'Medium-Spend, Medium-Freq',       3: 'High-Spend, High-Freq',       4: 'Young & High-Spend',       5: 'Older & Low-Spend',       6: 'Medium-Spend, Medium-Freq',       7: 'Mature & Medium-Freq',       8: 'Young, Big-Spender',       9: 'Young & Frequent',      10: 'Frequent, Low-Spend'  }  # Áp map vào DataFrame  agg['Segment\_DB'] = agg['Cluster\_DB'].map(name\_map\_db)  # 3. t-SNE 2D  tsne = TSNE(n\_components=2, random\_state=42, init='pca')  tsne\_comp = tsne.fit\_transform(X\_scaled)  agg['TSNE1'], agg['TSNE2'] = tsne\_comp[:,0], tsne\_comp[:,1]  # 4. Vẽ scatter với segment names  plt.figure(figsize=(12,6))  for c in sorted(agg['Cluster\_DB'].unique()):      pts = agg[agg['Cluster\_DB']==c]      label = name\_map\_db.get(c, f'Cluster {c}')      plt.scatter(pts['TSNE1'], pts['TSNE2'], label=label, alpha=0.6)  plt.title('Customer Segments by DBSCAN (t-SNE 2D)')  plt.xlabel('TSNE1')  plt.ylabel('TSNE2')  plt.grid(True)  # Legend ngoài khung  plt.legend(bbox\_to\_anchor=(1.02, 1), loc='upper left', fontsize='small', frameon=True)  plt.tight\_layout()  plt.show() |

Kết quả:  
A screenshot of a computer

AI-generated content may be incorrect.

A graph with different colored dots

AI-generated content may be incorrect.

Biểu đồ trên cho ta cái nhìn tổng quan về 12 phân khúc khách hàng được DBSCAN tìm ra (cộng thêm nhóm noise), trực quan hóa trên mặt phẳng t-SNE:

### Giải thích & nhận xét:

  - Nhóm lớn bên trái (DB 0: “High-Spend, Low-Freq”)

    + Màu cam, tập trung quanh TSNE1 ≈ –25.

    + Chi tiêu rất cao (Total\_Spending lớn) nhưng mua không thường xuyên (Frequency thấp).

  - Nhóm lớn bên phải (DB 1: “Low-Spend, Low-Freq”)

    + Màu xanh lá, tập trung quanh TSNE1 ≈ +25.

    + Chi tiêu và tần suất đều thấp → khách hàng lẻ, chưa gắn kết.

  - Hàng loạt “micro-segments” xung quanh:

    + Tái hiện các nhóm nhỏ như:

      - Medium-Spend, Medium-Freq

      - High-Spend, High-Freq – khách VIP thường xuyên

      - Young & High-Spend, Young & Frequent, Young, Big-Spender

      - Older & Low-Spend, Mature & Medium-Freq

      - Frequent, Low-Spend

    + Mỗi nhóm này có hành vi rất đặc thù về độ tuổi, chi tiêu, tần suất.

- Nhiễu:

+ Một số khách không vào được cụm nào: có thể do dữ liệu thiếu, hành vi thất thường hoặc nhóm quá nhỏ.

### Đề xuất hành động:

  - High-Spend, Low-Freq

    + Gửi ưu đãi đặc biệt khuyến khích mua lần 2, 3…

    + Mời tham gia membership cao cấp (VIP, free shipping)

  - Low-Spend, Low-Freq

    + Chiến dịch nhắc nhở: email/SMS với coupon hấp dẫn

    + Khảo sát nhanh về lý do ít quay lại

  - Medium-Spend, Medium-Freq

    + Cross-sell sản phẩm liên quan

    + Tặng điểm tích lũy mỗi lần mua

  - High-Spend, High-Freq (VIP)

    - Tri ân VIP bằng quà tặng, event exclusive

    - Mời làm ambassador chia sẻ trải nghiệm

  - Young & High-Spend / Frequent

    + Chiến dịch social media nhắm đúng độ tuổi

    + Flash sale mobile-app push

  - Older & Low-Spend

    + Ưu đãi theo dịp (sinh nhật, lễ)

    + Hỗ trợ tư vấn để tăng trải nghiệm mua sắm

  - Frequent, Low-Spend

    + Khuyến khích upsell: bundle, gói combo giá tốt

    + Giới thiệu membership để tăng giá trị trung bình đơn

  - Nhóm Nhiễu:

    + Tái phân loại sau khi thu thập thêm dữ liệu.

    + Triển khai chiến dịch khám phá nhu cầu (survey, mini-game).

## **6.8. Thống kê trung bình đặc trưng theo cụm KMeans**

|  |
| --- |
| print("\nMean features by KMeans cluster:")  display(agg.groupby('Cluster\_KM')[['Total\_Orders','Total\_Spending','Frequency\_of\_Purchase','Age']].mean()) |

Kết quả:  
A screenshot of a computer

AI-generated content may be incorrect.

### **Giải thích:**

- Cluster\_KM: Mỗi số đại diện cho một nhóm (cụm) khách hàng được phân theo KMeans.

- Total\_Orders: Trung bình số đơn hàng của khách hàng trong cụm. Tất cả đều bằng 1.0, có thể do dữ liệu ban đầu hoặc cách nhóm — nghĩa là mỗi khách hàng trong cụm chỉ có 1 đơn hàng trung bình.

- Total\_Spending: Trung bình tổng chi tiêu của khách hàng trong từng cụm. Nhóm 0 chi tiêu cao nhất (~283), nhóm 2 chi tiêu thấp nhất (~243), các nhóm còn lại nằm giữa.

- Frequency\_of\_Purchase: Tần suất mua sắm trung bình theo cụm. Nhóm 2 và 3 có tần suất mua sắm cao hơn (7.4-7.5 lần), nhóm 0 và 1 thấp hơn (~6.9-6.94).

- Age: Tuổi trung bình của khách hàng trong từng cụm khá đồng đều, dao động quanh 33-35 tuổi.

### **Nhận xét:**

- Các cụm có số đơn hàng trung bình bằng nhau, nhưng khác nhau về tổng chi tiêu và tần suất mua sắm.

- Cụm có tổng chi tiêu cao nhất không nhất thiết có tần suất mua sắm cao nhất.

- Tuổi khách hàng tương đối đồng đều, nên có thể tuổi không phải là yếu tố phân biệt lớn giữa các cụm trong trường hợp này.

### **Đề xuất hành động:**

- Tập trung marketing theo đặc điểm chi tiêu và tần suất: Ví dụ, nhóm có chi tiêu cao nhưng tần suất thấp có thể được khuyến khích mua thường xuyên hơn.

- Thiết kế ưu đãi hoặc chương trình khách hàng thân thiết dựa trên tần suất mua sắm để tăng sự gắn kết.

- Đi sâu phân tích các yếu tố khác như giới tính, thu nhập, hay loại sản phẩm mua, để có chiến lược cá nhân hóa hơn.

## **6.9. Thống kê trung bình đặc trưng theo cụm DBSCAN**

|  |
| --- |
| print("\nMean features by DBSCAN cluster:")  display(agg.groupby('Cluster\_DB')[['Total\_Orders', 'Total\_Spending', 'Frequency\_of\_Purchase', 'Age']].mean()) |

Kết quả:

A screenshot of a computer screen

AI-generated content may be incorrect.

### **Giải thích bảng**

Bảng thể hiện giá trị trung bình của 4 đặc trưng chính (Total\_Orders, Total\_Spending, Frequency\_of\_Purchase, Age) cho từng cụm DBSCAN.

Cluster\_DB = -1 là nhóm noise (các điểm bị DBSCAN đánh dấu là nhiễu, không thuộc cụm nào).

Các cụm từ 0 đến 10 là các nhóm khách hàng được DBSCAN phân loại dựa trên đặc trưng đã chuẩn hóa.

### **Nhận xét**

Total\_Orders luôn bằng 1.0 cho tất cả các cụm, nghĩa là mỗi khách hàng chỉ có một đơn hàng trong dữ liệu gốc hoặc cột này chưa đa dạng.

Total\_Spending biến thiên khá lớn giữa các cụm, từ khoảng 149 đến hơn 400, cho thấy sự khác biệt lớn về mức chi tiêu.

Frequency\_of\_Purchase cũng khá biến động, có những cụm có tần suất mua hàng cao (như cụm 10: 10.67 lần), có cụm thấp hơn (cụm 4, 8).

Age phân bố từ khoảng 20 tuổi đến gần 46 tuổi, một số cụm có khách hàng trẻ hơn (cụm 9: 20.8 tuổi), số khác trung niên hơn (cụm 5: 45.6 tuổi).

Cụm noise (-1) có mức chi tiêu trung bình, tần suất mua hàng và độ tuổi khá tương đồng với các cụm khác nhưng là những điểm không thể gom nhóm tốt bởi DBSCAN.

### **Đề xuất hành động**

Phân tích sâu từng cụm khách hàng để cá nhân hóa chiến dịch marketing:

Cụm có mức chi tiêu cao + tần suất mua thấp (vd: cụm 8, 4) có thể là khách hàng tiềm năng cần kích thích mua hàng thường xuyên hơn.

Cụm có tần suất mua cao nhưng chi tiêu thấp (vd: cụm 10, 5) có thể hướng đến chiến lược upsell, cross-sell.

Cụm khách hàng trẻ tuổi (vd: cụm 9) có thể cần các chiến dịch tiếp cận bằng kênh truyền thông số, mạng xã hội.

Cụm khách hàng lớn tuổi hơn (vd: cụm 5) có thể quan tâm đến các ưu đãi đặc biệt hoặc sản phẩm phù hợp theo độ tuổi.

Xử lý cụm noise (-1):

Đây là nhóm khách hàng không dễ phân loại, có thể cần xem xét thêm các đặc trưng khác hoặc dữ liệu bổ sung để phân loại tốt hơn.

Hoặc có thể làm riêng chiến dịch ưu đãi để thử tiếp cận và quan sát phản hồi.

Kiểm tra lại cột Total\_Orders vì dữ liệu có vẻ không đa dạng (luôn bằng 1.0), nếu có thể thu thập thêm dữ liệu về số đơn hàng thực tế sẽ giúp phân nhóm chính xác hơn.

Tiếp tục áp dụng thêm các thuật toán phân nhóm khác để so sánh kết quả, hoặc điều chỉnh tham số DBSCAN (eps, min\_samples) để có nhóm cụm tốt hơn.

# **7.Dự đoán hành vi**

Để dự đoán và phân loại hành vi chi tiêu của khách hàng dựa trên dữ liệu có sẵn, ta sử dụng các mô hình học máy cho hai bài toán chính: Phân loại nhóm chi tiêu và Dự báo số tiền chi tiêu.

## **7.1. Mô hình Phân loại:**

Nhóm mô hình này được thiết kế để gán một mẫu dữ liệu vào một trong nhiều lớp định sẵn.

Logistic Regression: Đây là một trong những mô hình phân loại cơ bản và được sử dụng rộng rãi nhất. Nó hoạt động bằng cách ước tính xác suất một mẫu thuộc về một lớp cụ thể thông qua một hàm sigmoid, sau đó sử dụng ngưỡng xác suất để đưa ra quyết định phân loại. Mặc dù có bản chất tuyến tính, Logistic Regression là điểm khởi đầu tốt để đánh giá khả năng phân tách tuyến tính của dữ liệu.

Random Forest Classifier: Mô hình này xây dựng một "rừng" gồm nhiều cây quyết định độc lập trong quá trình huấn luyện. Kết quả phân loại cuối cùng được quyết định dựa trên sự đồng thuận (bỏ phiếu) của đa số các cây. Đặc điểm nổi bật của Random Forest là khả năng xử lý tốt các mối quan hệ phi tuyến tính và ít bị ảnh hưởng bởi nhiễu hay ngoại lai.

## **7.2. Mô hình Hồi quy:**

Nhóm mô hình này được sử dụng để dự đoán một giá trị số liên tục (ví dụ: giá tiền, số lượng...).

Linear Regression: Mô hình hồi quy tuyến tính là phương pháp đơn giản nhất để dự đoán một biến mục tiêu dạng số. Nó giả định rằng có một mối quan hệ tuyến tính giữa các biến đầu vào (đặc trưng) và biến mục tiêu. Mô hình tìm cách vẽ một đường (trong trường hợp một đặc trưng) hoặc một mặt phẳng/siêu mặt phẳng (trong trường hợp nhiều đặc trưng) phù hợp nhất với dữ liệu.

Random Forest Regressor: Tương tự như phiên bản phân loại, Random Forest Regressor sử dụng tập hợp các cây quyết định. Tuy nhiên, thay vì bỏ phiếu cho một lớp, mỗi cây dự đoán một giá trị số, và kết quả cuối cùng thường là giá trị trung bình của các dự đoán này. Mô hình này kế thừa khả năng xử lý phi tuyến và tương tác đặc trưng từ Random Forest.

XGBoost Regressor: XGBoost (Extreme Gradient Boosting) là một thuật toán Boosting hiệu năng cao. Nó xây dựng các cây quyết định một cách tuần tự, trong đó mỗi cây mới được thêm vào nhằm mục đích sửa chữa lỗi mà các cây trước đó mắc phải. XGBoost nổi tiếng với tốc độ xử lý nhanh, hiệu suất dự đoán mạnh mẽ và khả năng xử lý nhiều loại dữ liệu khác nhau.

Việc áp dụng đa dạng các mô hình cho phép có cái nhìn toàn diện hơn về mức độ phức tạp của mối quan hệ giữa các đặc trưng khách hàng và hành vi chi tiêu của họ.

## **7.3. Mục đích sử dụng mô hình**

Các mô hình được lựa chọn và áp dụng trong phân tích này nhằm giải quyết hai mục tiêu cụ thể, trực tiếp liên quan đến việc hiểu và tận dụng thông tin về hành vi chi tiêu khách hàng cho các mục đích kinh doanh:

Sử dụng các mô hình phân loại như Logistic Regression và Random Forest Classifier để phân loại khách hàng vào các nhóm chi tiêu cụ thể. Việc chia khách hàng thành các phân khúc như "Low", "Medium", "High" dựa trên dữ liệu hành vi (không bao gồm trực tiếp số tiền chi tiêu hoặc thời gian mua hàng) cho phép doanh nghiệp áp dụng các chiến lược marketing và chăm sóc khách hàng mục tiêu hóa. Ví dụ, nhóm khách hàng "High" có thể nhận được các ưu đãi độc quyền hoặc dịch vụ hỗ trợ cao cấp hơn, trong khi nhóm "Low" có thể được khuyến khích mua sắm thêm thông qua các chương trình khuyến mãi phù hợp.

Sử dụng các mô hình hồi quy như Linear Regression, Random Forest Regressor, và XGBoost Regressor để dự báo trực tiếp số tiền chi tiêu mà một khách hàng có thể thực hiện. Khả năng dự đoán một giá trị số cụ thể này có ý nghĩa lớn trong việc định lượng giá trị tiềm năng của từng khách hàng, tối ưu hóa chiến lược định giá, hoặc cá nhân hóa các đề xuất sản phẩm với mức giá phù hợp.

## **7.4.Kết quả sau khi huấn luyện mô hình**

Sau khi thực hiện các bước tiền xử lý dữ liệu cần thiết (bao gồm tạo biến mục tiêu "Spending\_Group" bằng cách chia định lượng Purchase\_Amount, mã hóa các biến phân loại, và chia dữ liệu thành tập huấn luyện/kiểm thử), các mô hình đã được huấn luyện và đánh giá trên tập dữ liệu kiểm thử.

### **7.4.1. Đối với bài toán Phân loại nhóm chi tiêu:**

**Logistic Regression:**

Mô hình này cho độ chính xác (Accuracy) trên tập kiểm thử là khoảng 0.35. Các chỉ số khác như Precision, Recall, và F1-score (đo lường hiệu quả trên từng lớp và trung bình) đều ở mức rất thấp, xấp xỉ 0.35.

Ma trận nhầm lẫn (Confusion Matrix) cho thấy số lượng dự đoán đúng trên đường chéo (ví dụ: True=Low, Predicted=Low) rất nhỏ so với số lượng các trường hợp bị phân loại sai sang các nhóm khác.

**Random Forest Classifier:**

Tương tự, mô hình Random Forest cho kết quả rất khiêm tốn với độ chính xác khoảng 0.34. Các chỉ số Precision, Recall, F1-score trung bình cũng chỉ đạt khoảng 0.34.

Ma trận nhầm lẫn của mô hình này cũng cho thấy sự chồng lấn và nhầm lẫn đáng kể giữa các nhóm Low, Medium, High, tương tự như Logistic Regression.

### **7.4.2. Đối với bài toán Dự báo chi tiêu**

Linear Regression: Mô hình hồi quy tuyến tính cho chỉ số R² là 0.0173 và sai số tuyệt đối trung bình (MAE) là 113.38.

Random Forest Regressor: Mô hình Random Forest cho chỉ số R² là -0.0443 và MAE là 116.44.

XGBoost Regressor: Mô hình XGBoost cho chỉ số R² là -0.2946 và MAE là 127.39.

## **7.5. Đánh giá hiệu quả mô hình**

Dựa trên các kết quả huấn luyện, có thể khẳng định rằng tất cả các mô hình được thử nghiệm đều cho thấy hiệu suất thấp trên cả hai bài toán phân loại nhóm chi tiêu và dự báo số tiền chi tiêu trên tập dữ liệu này.

-Đối với bài toán phân loại, độ chính xác chỉ khoảng 34-35%, trong khi với 3 lớp có kích thước tương đương nhau, một mô hình đoán ngẫu nhiên cũng có thể đạt độ chính xác xấp xỉ 33.3%. Điều này chỉ ra rằng các mô hình này hầu như không học được cách phân biệt các nhóm chi tiêu dựa trên các đặc trưng được cung cấp. Các chỉ số Precision và Recall thấp cho từng lớp (ví dụ: Recall 0.27 cho lớp High của Logistic Regression) đồng nghĩa với việc mô hình bỏ sót rất nhiều mẫu thuộc lớp đó hoặc phân loại nhầm rất nhiều mẫu từ lớp khác vào lớp đó. Ma trận nhầm lẫn là bằng chứng hình ảnh rõ nét nhất cho thấy sự chồng lấn nghiêm trọng giữa các nhóm khách hàng trong không gian đặc trưng.

-Đối với bài toán hồi quy, các chỉ số R² là rất đáng báo động. Giá trị R² đo lường tỷ lệ phương sai của biến mục tiêu được giải thích bởi mô hình. R² = 1 là dự đoán hoàn hảo, R² = 0 nghĩa là mô hình không tốt hơn việc chỉ dự đoán giá trị trung bình. Việc cả ba mô hình đều có R² rất gần 0 (Linear Regression) hoặc thậm chí âm (Random Forest, XGBoost) cho thấy chúng hoạt động kém hơn hoặc chỉ tương đương với một mô hình đơn giản nhất là luôn dự đoán biến mục tiêu bằng giá trị trung bình của tập huấn luyện. MAE cao (trên 100) chỉ ra rằng trung bình, dự đoán của mô hình lệch khỏi giá trị thực tế một khoảng đáng kể.

Bên cạnh hiệu suất kém, mô hình Logistic Regression còn xuất hiện cảnh báo về việc không hội tụ hoàn toàn trong quá trình huấn luyện. Điều này thường là do các đặc trưng số không được chuẩn hóa/điều chỉnh tỷ lệ, hoặc số vòng lặp huấn luyện không đủ. Dù đã tăng max\_iter, cảnh báo vẫn tồn tại, gợi ý vấn đề về tỷ lệ đặc trưng hoặc sự khó khăn cố hữu trong việc tìm lời giải tối ưu.

Việc cả các mô hình tuyến tính (Logistic Regression, Linear Regression) và các mô hình phi tuyến mạnh mẽ hơn (Random Forest, XGBoost) đều thất bại đồng loạt củng cố giả thuyết rằng vấn đề cốt lõi nằm ở bản thân dữ liệu và các đặc trưng đầu vào. Các đặc trưng hiện có có thể chưa đủ thông tin, không phản ánh đúng bản chất phức tạp, phi tuyến của hành vi chi tiêu, hoặc các nhóm chi tiêu được định nghĩa ban đầu thực sự chồng lấn rất lớn dựa trên các yếu tố đo lường được.

## **7.6.Đề xuất:**

### **7.6.1. Phân khúc khách hàng để cá nhân hóa chiến dịch**

Thông qua phương pháp phân cụm khách hàng, xác định được các nhóm người dùng với hành vi rõ rệt như: nhóm trung thành với thương hiệu, nhóm nhạy cảm với khuyến mãi và nhóm ít tương tác. Điều này cho phép doanh nghiệp định hướng nội dung và hình thức quảng bá phù hợp với từng đối tượng cụ thể.

Chiến lược đề xuất:

- Thiết kế chiến dịch cá nhân hóa cho từng nhóm: ưu đãi độc quyền cho khách trung thành, giảm giá hoặc combo hấp dẫn cho nhóm nhạy giá.

- Áp dụng remarketing thông minh để nhắm lại đúng nhóm có tiềm năng cao, giảm lãng phí ngân sách quảng cáo.

### **7.6.2. Tối ưu trải nghiệm mua sắm và kênh tiếp cận**

Phân tích hành vi người tiêu dùng cho thấy phần lớn khách hàng sử dụng smartphone và tablet để mua sắm, đồng thời các yếu tố như thời gian nghiên cứu sản phẩm, sự hài lòng, lòng trung thành và tương tác với quảng cáo có ảnh hưởng trực tiếp đến mức chi tiêu.

Chiến lược đề xuất:

-Tối ưu giao diện và trải nghiệm mua sắm trên thiết bị di động.

-Triển khai chương trình tích điểm, gợi ý sản phẩm dựa trên thời gian nghiên cứu, gửi email hoặc thông báo đẩy khi khách bỏ giỏ hàng.

-Ưu tiên quảng bá sản phẩm có đánh giá tốt và phù hợp với thói quen tìm hiểu của từng nhóm.

# **8.Kết luận và Đề xuất**

## **8.1. Kết luận**

Qua quá trình phân tích toàn diện từ dữ liệu thô đến khai thác chuyên sâu, dự án đã làm rõ được nhiều khía cạnh quan trọng về hành vi tiêu dùng trong môi trường thương mại điện tử hiện đại. Những kết quả thu được cho thấy:

Người tiêu dùng hiện đại hành động có cân nhắc, ưu tiên sự thuận tiện, giá trị thực tế và trải nghiệm mua sắm liền mạch – đặc biệt trên thiết bị di động.

Đặc điểm nhân khẩu học như độ tuổi, giới tính và thu nhập có tác động rõ rệt đến hành vi mua hàng và tương tác với quảng cáo, từ đó khẳng định tầm quan trọng của việc phân khúc khách hàng.

Phân tích mức độ tương tác và hành vi quảng cáo chỉ ra rằng việc đầu tư vào quảng cáo cần đi đôi với chất lượng sản phẩm và trải nghiệm thực tế, nếu không sẽ gây tác dụng ngược.

Các cụm khách hàng được xác định có hành vi và tiềm năng khác biệt, là nền tảng cho chiến lược tiếp cận cá nhân hóa.

Mô hình dự đoán hành vi tiêu dùng hiện tại chưa đạt hiệu quả mong muốn, phần lớn do giới hạn trong dữ liệu và đặc trưng đầu vào chưa đủ thông tin hoặc chưa phản ánh đúng bản chất hành vi người dùng.

Tổng thể, dự án đã đạt được mục tiêu cốt lõi: cung cấp bức tranh dữ liệu rõ nét về hành vi người tiêu dùng, đồng thời chỉ ra hướng cải tiến cho doanh nghiệp trong việc cá nhân hóa trải nghiệm, tối ưu chiến dịch tiếp thị và nâng cao mức độ gắn bó của khách hàng.

## **8.2. Đề xuất**

Dựa trên các phân tích đã thực hiện, nhóm đề xuất một số hướng chiến lược cụ thể như sau:

### **8.2.1. Tăng cường cá nhân hóa trên cơ sở phân khúc khách hàng**

Xây dựng nội dung và chiến dịch riêng biệt cho từng cụm khách hàng:

Cụm chi tiêu cao nhưng ít mua → chiến lược thúc đẩy tần suất mua bằng ưu đãi định kỳ.

Cụm mua nhiều nhưng chi tiêu thấp → tập trung vào chương trình upsell/cross-sell.

Cụm trẻ tuổi → tận dụng mạng xã hội, KOLs và quảng cáo dạng ngắn.

Cụm lớn tuổi → chú trọng sản phẩm an toàn, khuyến mãi chu đáo, giao diện dễ dùng.

### **8.2.2.** **Tối ưu hóa trải nghiệm trên thiết bị di động**

Đầu tư vào giao diện, tốc độ và chức năng thanh toán trên điện thoại.

Phát triển ứng dụng mobile riêng hoặc tích hợp sâu với các nền tảng thanh toán số.

### **8.2.3. Kết nối quảng cáo – trải nghiệm – sản phẩm**

Không chạy quảng cáo dàn trải, thay vào đó sử dụng remarketing thông minh dựa trên hành vi và thời điểm người dùng quan tâm.

Kết hợp quảng cáo với đánh giá thực tế, minh bạch hóa thông tin sản phẩm.

Đảm bảo trải nghiệm hậu quảng cáo (chất lượng sản phẩm, giao hàng, CSKH) đúng với kỳ vọng mà quảng cáo tạo ra.

### **8.2.4. Đề xuất dài hạn**

Thiết lập hệ thống phân tích hành vi người dùng theo thời gian thực để phản ứng nhanh với thay đổi hành vi tiêu dùng.

Kết hợp dữ liệu thương mại điện tử với dữ liệu mạng xã hội, feedback, review,... để có cái nhìn toàn diện hơn.

Đào tạo đội ngũ phân tích dữ liệu nội bộ, giúp doanh nghiệp quyết định dựa trên dữ liệu thay vì cảm tính.