# TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO TIỂU LUẬN PHÂN TÍCH DỮ LIỆU

**PHÂN TÍCH KHÁM PHÁ VỀ DỮ LIỆU MUA SẮM**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thông tin thành viên:**

Lê Nguyễn Quốc Hưng 3121410240

Trần Bá Tiên 3122410410

Dương Huỳnh Gia 3121410167

Bùi Thành Công 3122410042

***TP. Hồ Chí Minh, tháng 10/2025***

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..……………………………………….……………………..…………………………………........................................................................................................................

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại Học Sài Gòn, đặc biệt là các thầy cô ở khoa Công Nghệ Thông Tin đã tạo điều kiện cho em tiếp cận và tìm hiểu để hoàn thành đồ án môn học lần này.

Và chúng em cũng xin chân thành đặc biệt cảm ơn thầy Đỗ Như Tài, thầy là giảng viên giảng dạy đã nhiệt tình hướng dẫn giúp chúng em hoàn thành được đồ án lần này.

Trong quá trình nghiên cứu và làm bài báo cáo đồ án, kiến thức cũng như kinh nghiệm thực tế còn nhiều hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được ý kiến đóng góp của thầy, cô để em học hỏi được nhiều kỹ năng, kinh nghiệm và sẽ hoàn thành tốt hơn cho những bài báo cáo sắp tới.

Chúng em xin chân thành cảm ơn!

### I. ĐỊNH NGHĨA VẤN ĐỀ (PROBLEM DEFINITION)

Trong bối cảnh thương mại điện tử ngày càng cạnh tranh, việc thấu hiểu hành vi khách hàng và tối ưu hóa hoạt động kinh doanh là yếu tố sống còn. Bạn đã bao giờ tự hỏi: **Ai là những khách hàng giá trị nhất của mình? Sản phẩm nào đang mang lại doanh thu cao nhất và sản phẩm nào chỉ bán chạy về số lượng? Khách hàng thường mua sắm vào những thời điểm nào trong ngày hoặc trong tuần?**

Bộ dữ liệu "Online Retail" cung cấp dữ liệu giao dịch chi tiết từ một cửa hàng bán lẻ trực tuyến tại Anh, mở ra cơ hội để chúng ta trả lời những câu hỏi đó.

Mục tiêu của dự án này là thực hiện Phân tích Dữ liệu Khám phá (EDA) trên bộ dữ liệu giao dịch này. Chúng ta sẽ làm sạch, phân tích và trực quan hóa dữ liệu để khám phá các xu hướng mua sắm, xác định các sản phẩm và thị trường chủ lực, và cuối cùng là phân loại các nhóm khách hàng khác nhau dựa trên hành vi của họ. Những insight thu được sẽ là nền tảng để đề xuất các chiến lược kinh doanh, marketing và quản lý kho hàng hiệu quả hơn.

Mô tả bài toán: Đây là một bài toán phân tích khám phá và phân cụm khách hàng. Từ dữ liệu giao dịch thô, chúng ta cần rút ra những hiểu biết kinh doanh có giá trị và xác định các phân khúc khách hàng riêng biệt.

**Dữ liệu vào (Features):** Bộ dữ liệu bao gồm 8 thuộc tính giao dịch chính:

* **InvoiceNo:** Mã số hóa đơn.
* **StockCode:** Mã sản phẩm trong kho.
* **Description:** Tên mô tả sản phẩm.
* **Quantity:** Số lượng sản phẩm được mua.
* **InvoiceDate:** Ngày và giờ giao dịch.
* **UnitPrice:** Đơn giá của sản phẩm.
* **CustomerID:** Mã định danh duy nhất cho mỗi khách hàng.
* **Country:** Quốc gia của khách hàng.

**Kết quả mong muốn (Desired Outcomes):**

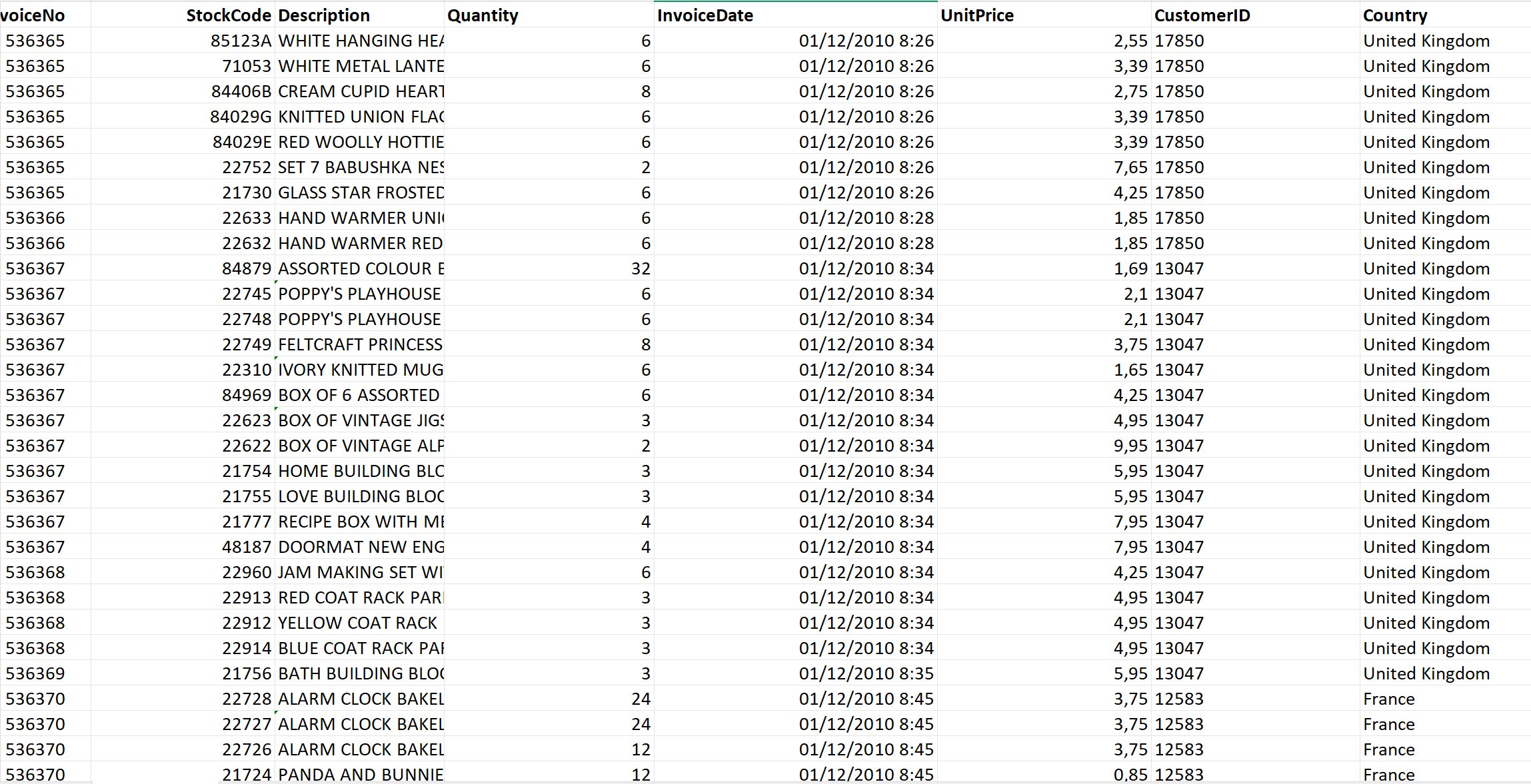
Xác định các sản phẩm, quốc gia và khách hàng mang lại doanh thu cao nhất.

Phân tích các xu hướng mua sắm theo thời gian (giờ, ngày, tháng).

Phân cụm khách hàng (Customer Segments) thành các nhóm có ý nghĩa (ví dụ: Khách hàng Vô địch, Khách hàng Trung thành, Khách hàng cần chú ý) để có chiến lược chăm sóc phù hợp.

# II. CHUẨN BỊ VẤN ĐỀ (PREPARE PROBLEM)

* Bộ dữ liệu được lấy từ Kaggle .

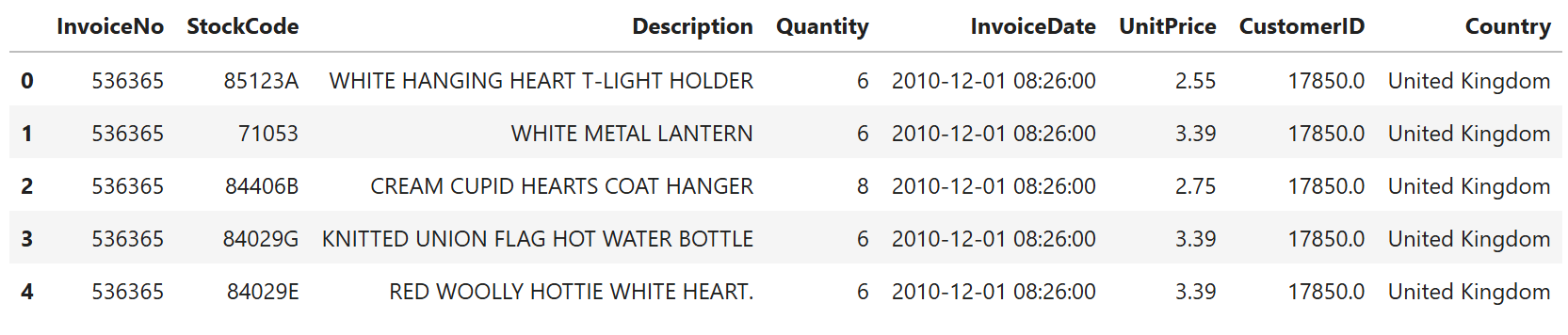


*Hình 1: Bộ dữ liệu*

**III. PHÂN TÍCH**

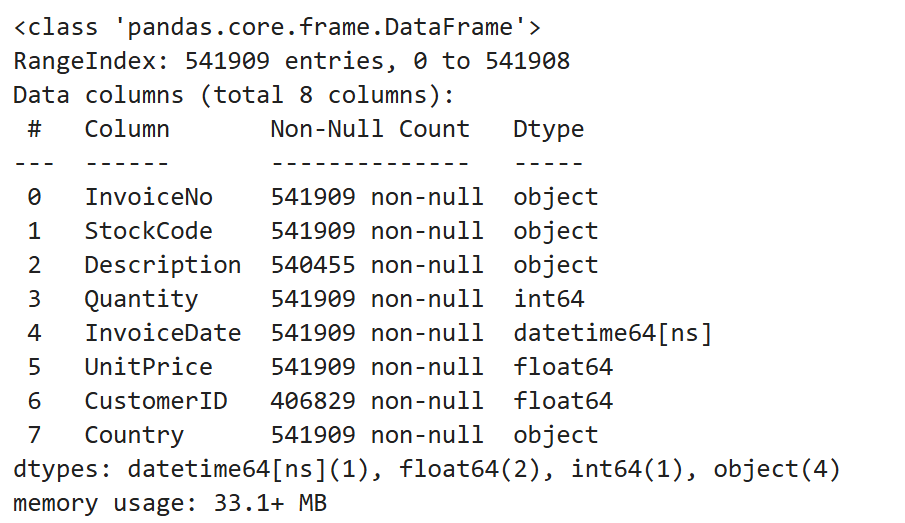
## 1. EDA không dùng biểu đồ (Non-graphie EDA)

### 1.1 In 5 dòng data đầu



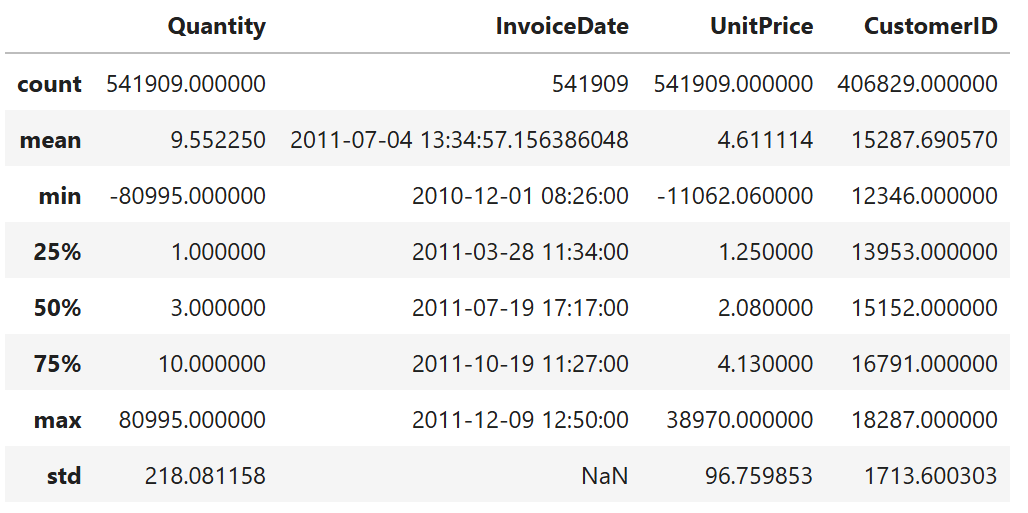
*Hình 2: In 5 dòng dữ liệu đầu tiên*

### 1.2 Kiểm tra thông tin khái quát



*Hình 3: Hình thông tin khái quát*

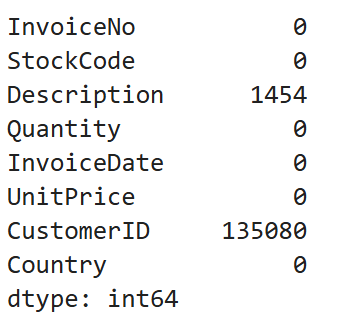
### 1.3 Thống kê mô tả: Xem các thống kê cơ bản cho các cột số.



*Hình 4: Thống kê cơ bản của các cột*

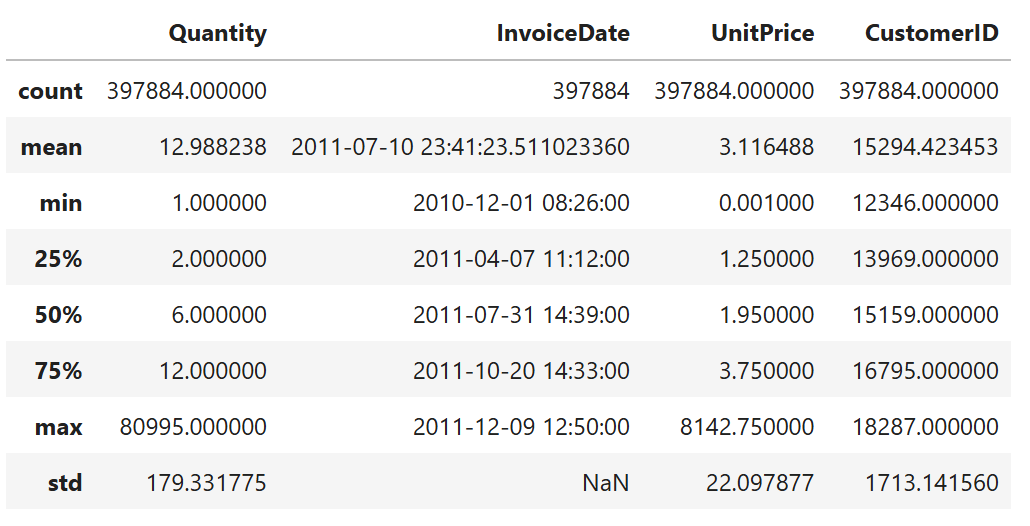
* Count: số lượng giá trị
* Mean: giá trị trung bình
* Std: độ lệch chuẩn (mức độ phân tán)
* Min, max: giá trị nhỏ nhất, lớn nhất
* 25%, 50%, 75%: các phân vị (Q1, median, Q3)

### 1.5 Kiểm tra giá trị thiếu ở mỗi cột



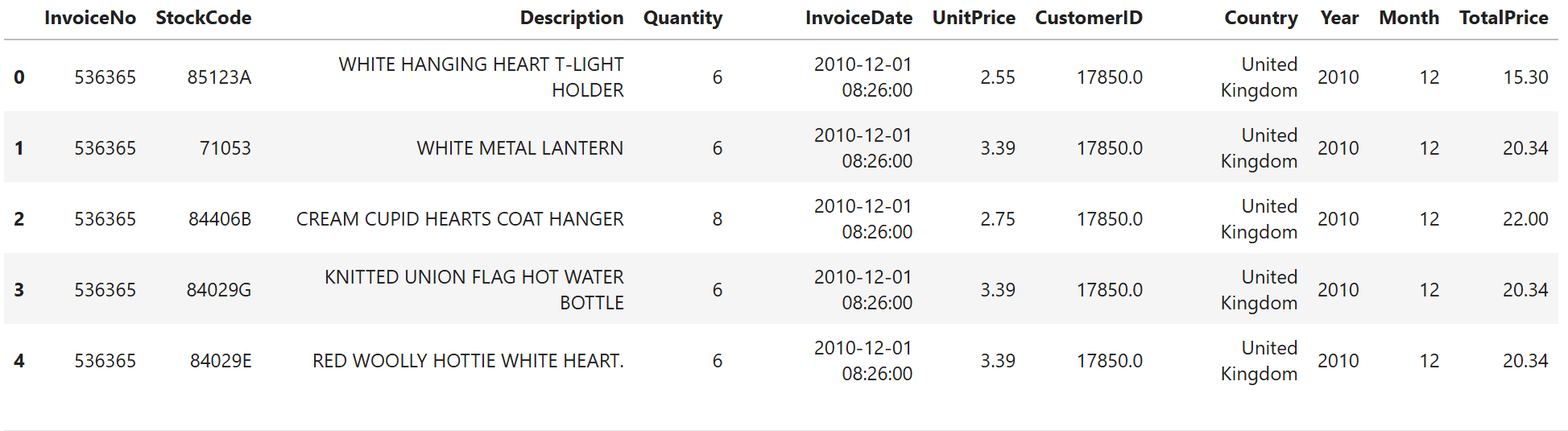
*Hình 5: Kiểm tra giá trị thiếu ở mỗi cột*

- Cột Customer ID thiếu rất nhiều giá trị. Vì phân tích hành vi khách hàng là rất quan trọng, chúng ta sẽ xóa các dòng không có ID, tương tự với những đơn hàng bị hủy và các trường hợp giá nhỏ hơn 0.



Hình 6: Thống kê cơ bản của các cột sau khi xóa

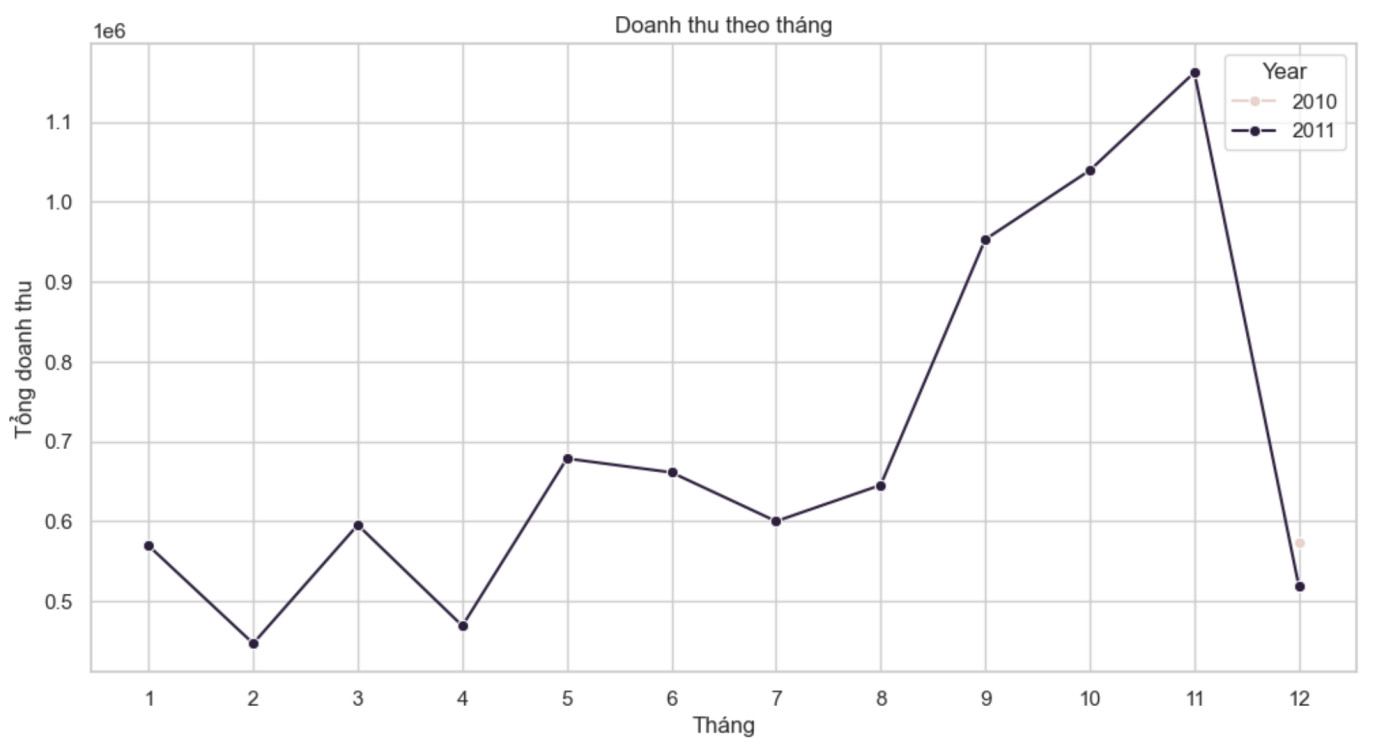
### 1.6 Thêm các cột cần thiết



Hình 7: Thêm các cột Year, Month và TotalPrice

## **2. EDA trực quan bằng sơ đồ (Graphical EDA)**

### 2.1 Tổng doanh thu theo tháng



**Hình 8: Doanh thu theo tháng**

**Nhận xét: Sự tăng trưởng mạnh vào cuối năm:**

Điểm đáng chú ý nhất là doanh thu đạt đỉnh vào tháng 11 năm 2011, vượt mức 1,1 triệu. Đây là mức cao nhất trong cả năm.

Diễn giải: Sự tăng trưởng đột biến này rất có thể là do nhu cầu mua sắm cho các dịp lễ lớn cuối năm như Black Friday, Cyber Monday và chuẩn bị cho Giáng sinh. Đây là mùa mua sắm quan trọng nhất trong năm.

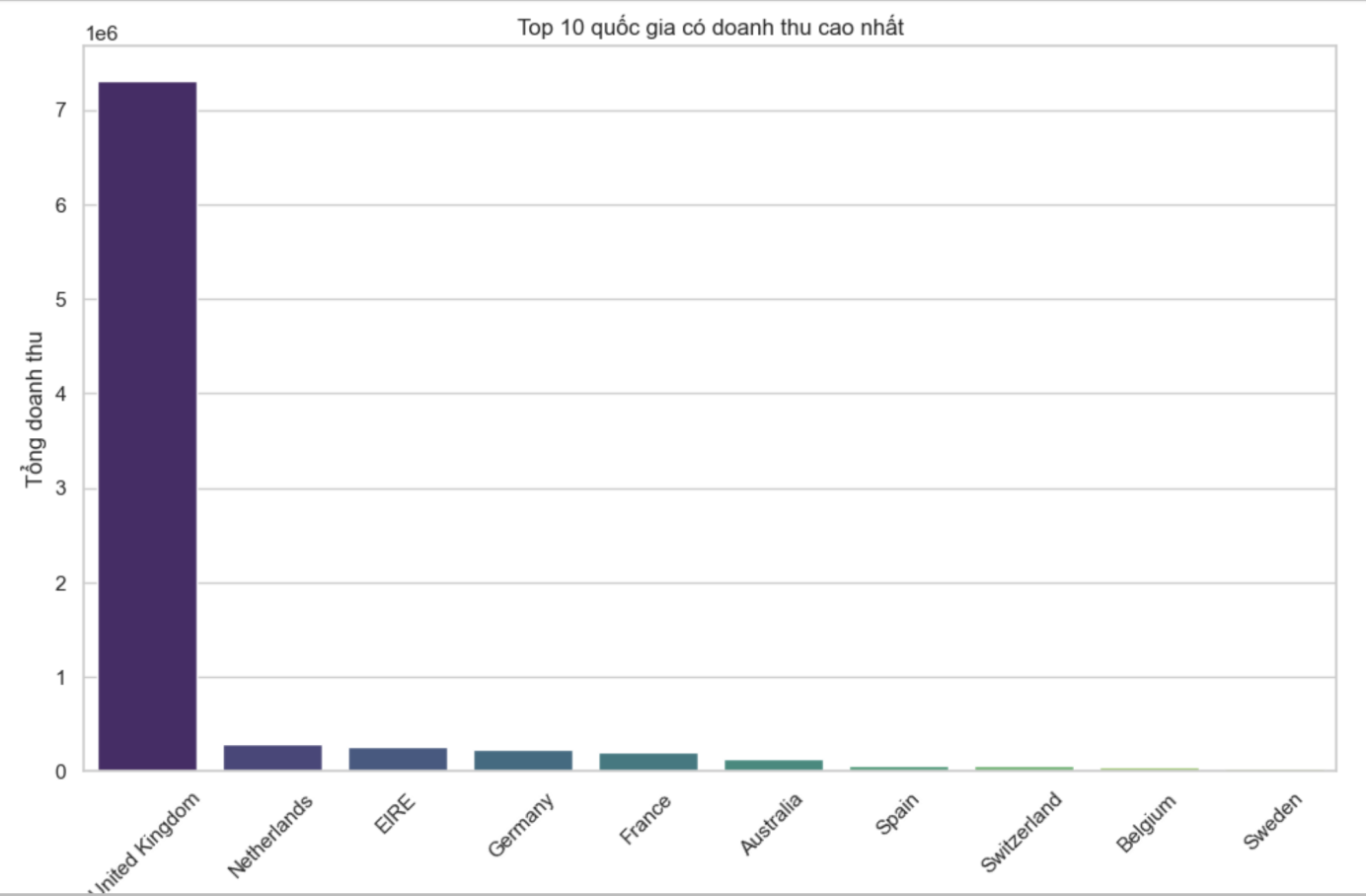
**- Sự sụt giảm vào tháng 12:**

Ngay sau khi đạt đỉnh vào tháng 11, doanh thu sụt giảm mạnh xuống mức rất thấp vào tháng 12 năm 2011 (chỉ còn khoảng hơn 500,000).

+ Dữ liệu không đầy đủ: Rất có thể bộ dữ liệu cho tháng 12/2011 chỉ ghi nhận giao dịch trong một khoảng thời gian ngắn (ví dụ: chỉ tuần đầu tiên của tháng) và không phản ánh toàn bộ hoạt động kinh doanh của tháng đó.

+ Khách hàng đã mua sắm phần lớn vào tháng 11 nên nhu cầu tháng 12 giảm. Tuy nhiên, mức giảm sâu như vậy vẫn là một dấu hiệu cần điều tra thêm.

**2.2 Top 10 quốc gia có doanh thu cao nhất**



**Hình 9: Top các quốc gia có doanh thu cao nhất**

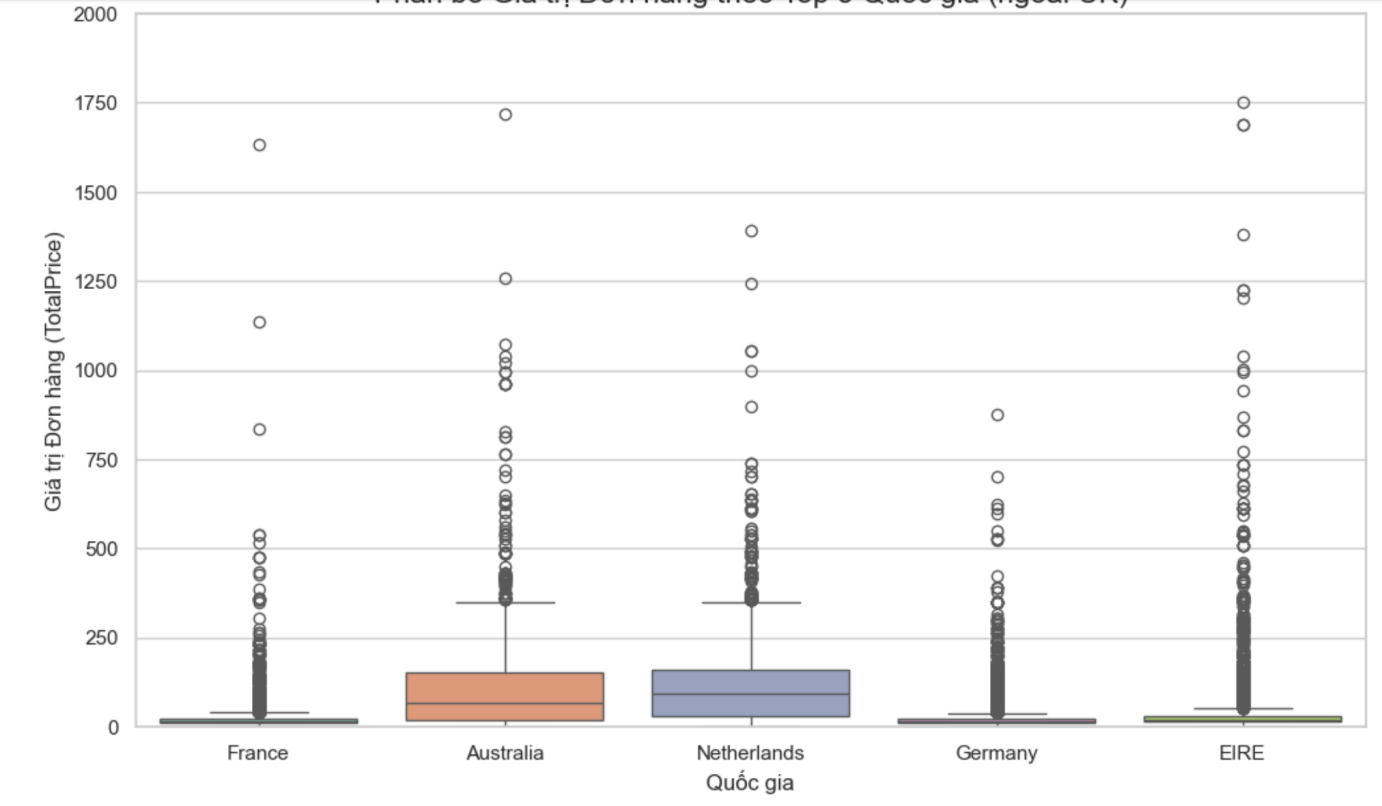
- Rủi ro về sự phụ thuộc: Việc phụ thuộc quá lớn vào thị trường UK là một rủi ro chiến lược. Bất kỳ sự suy thoái kinh tế hay thay đổi chính sách nào tại UK cũng sẽ ảnh hưởng nặng nề. Báo cáo cần nhấn mạnh đến việc đa dạng hóa thị trường là một mục tiêu quan trọng.

- Tiềm năng phát triển tại các thị trường lân cận: Sự thành công bước đầu tại Hà Lan, Ireland và Đức cho thấy tiềm năng tăng trưởng rất lớn. Đề xuất:

+ Phân tích sâu hơn: Tìm hiểu xem sản phẩm nào bán chạy nhất tại các thị trường này để xây dựng chiến lược marketing tập trung.

+ Đầu tư có chọn lọc: Dồn nguồn lực marketing và logistics để đẩy mạnh tăng trưởng ở 3-4 thị trường quốc tế hàng đầu thay vì dàn trải.

**2.3 Biểu đồ hộp (Boxplot) để so sánh phân bố doanh thu quốc gia (ngoài UK)**



**Hình 10: Phân bố Doanh thu theo quốc gia ngoài UK**

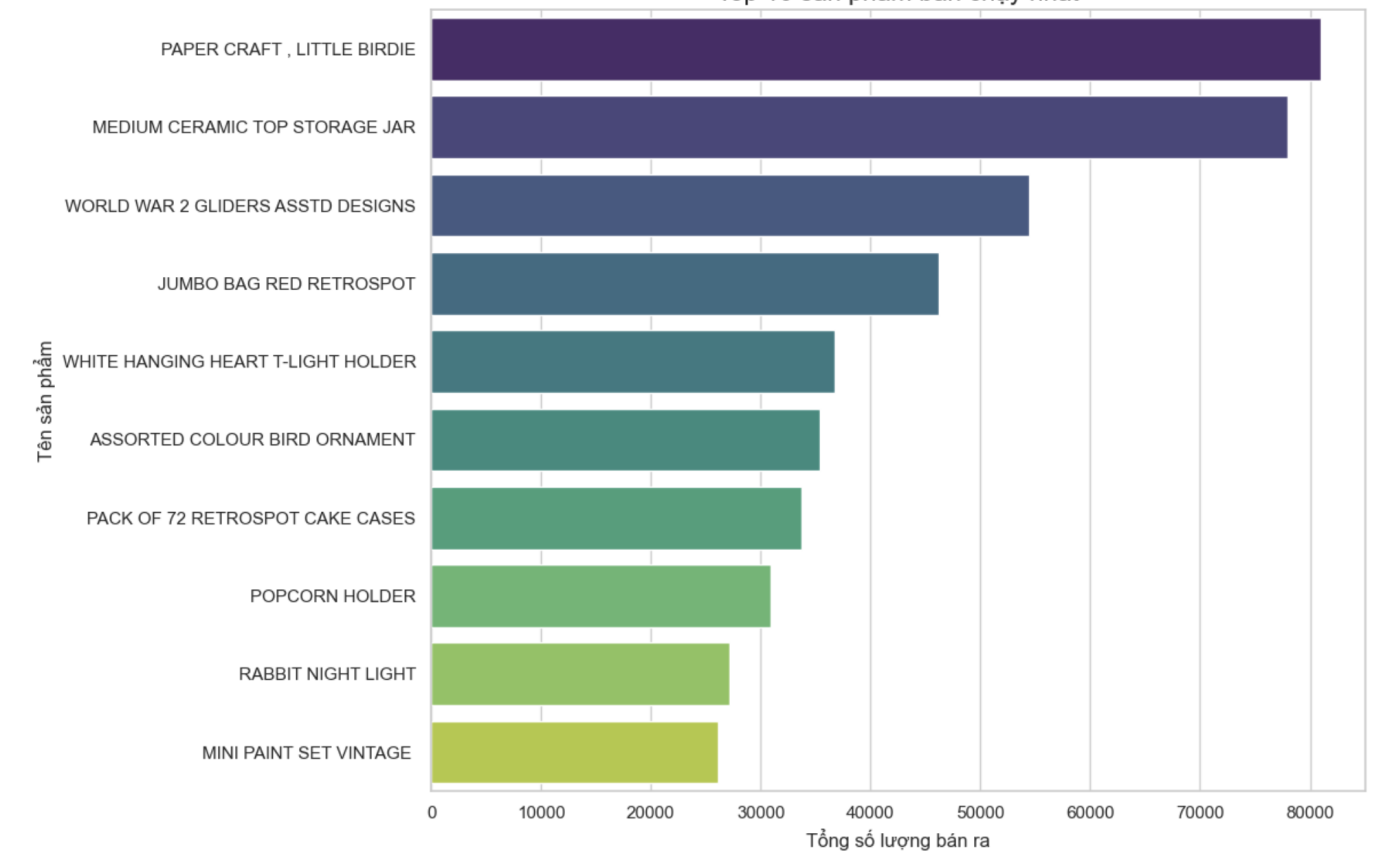
* **Sự khác biệt về mức chi tiêu phổ biến:**

- Úc (Australia) và Hà Lan (Netherlands) có "hộp" (đại diện cho 50% số đơn hàng ở giữa) lớn hơn và nằm ở vị trí cao hơn so với các nước còn lại. Điều này cho thấy khách hàng ở Úc và Hà Lan có xu hướng chi tiêu cho mỗi đơn hàng cao hơn và đa dạng hơn. Mức giá của các đơn hàng "điển hình" ở hai quốc gia này cao hơn so với Pháp và Đức.

- Thị trường có mức chi tiêu thấp và ổn định: Pháp (France) và Đức (Germany) có hộp rất hẹp và nằm gần mức 0. Phần lớn các đơn hàng ở hai quốc gia này có giá trị thấp và ít biến động. Khách hàng ở đây có thể chủ yếu mua các mặt hàng nhỏ, giá rẻ.

- Tất cả năm quốc gia đều có rất nhiều điểm tròn nằm bên ngoài "râu" của hộp, kéo dài lên các mức giá trị rất cao (trên 500, 1000, thậm chí gần 1750). Đây là những đơn hàng "khủng", có giá trị cao bất thường. Sự tồn tại của chúng cho thấy ở mỗi thị trường đều có một nhóm nhỏ khách hàng sẵn sàng chi tiêu rất mạnh tay. Đây là những khách hàng cực kỳ giá trị.

**2.4 Những sản phẩm bán chạy nhất**

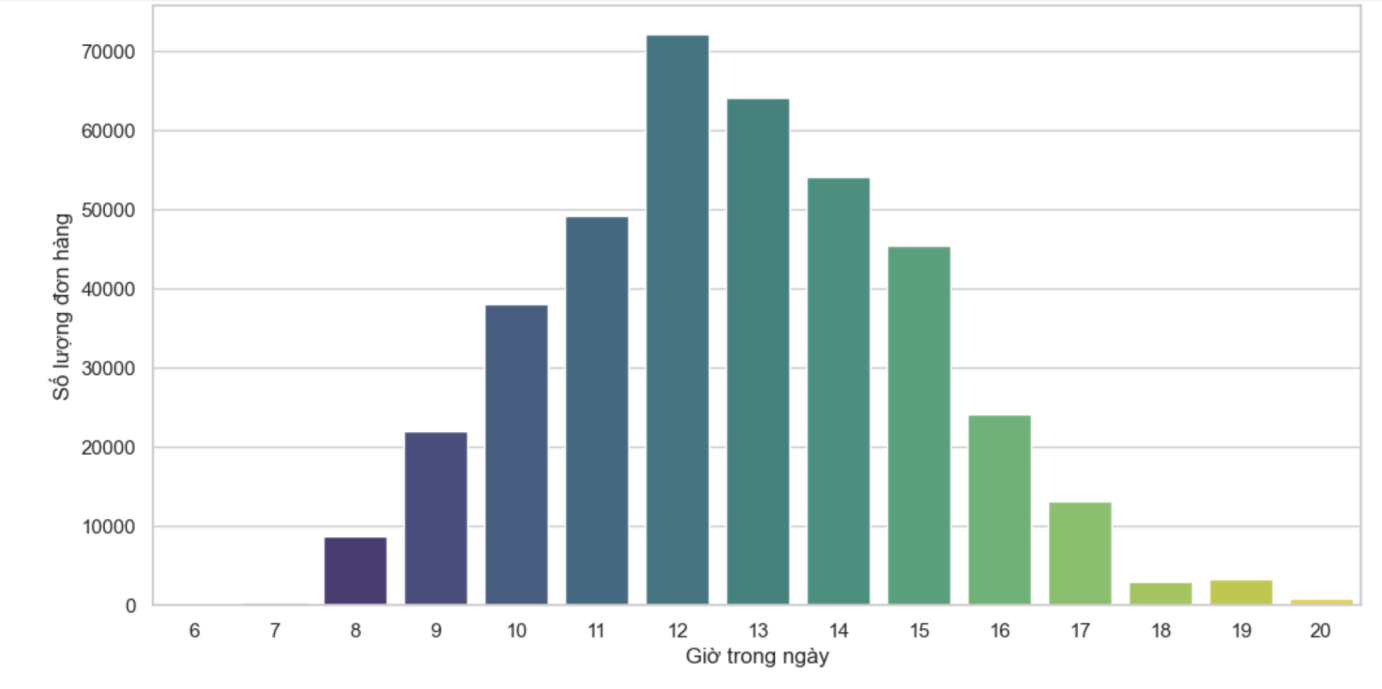


**Hình 11: Top 10 sản phẩm bán chạy nhất**

- Hầu hết các sản phẩm trong top 10 đều có đặc điểm chung là đồ trang trí, đồ gia dụng, văn phòng phẩm.Đây là những sản phẩm đặc trưng, giá rẻ và được yêu thích rộng rãi.

- Sự vắng mặt của các mặt hàng giá trị cao: Không có sản phẩm nào trong danh sách này có vẻ là một mặt hàng đắt tiền. Điều này cũng cho thấy rằng số lượng bán ra cao không đồng nghĩa với doanh thu cao.

**2.5 Số lượng đơn hàng trong ngày theo giờ:**



Hình 12: Số lượng đơn hàng trong ngày theo giờ

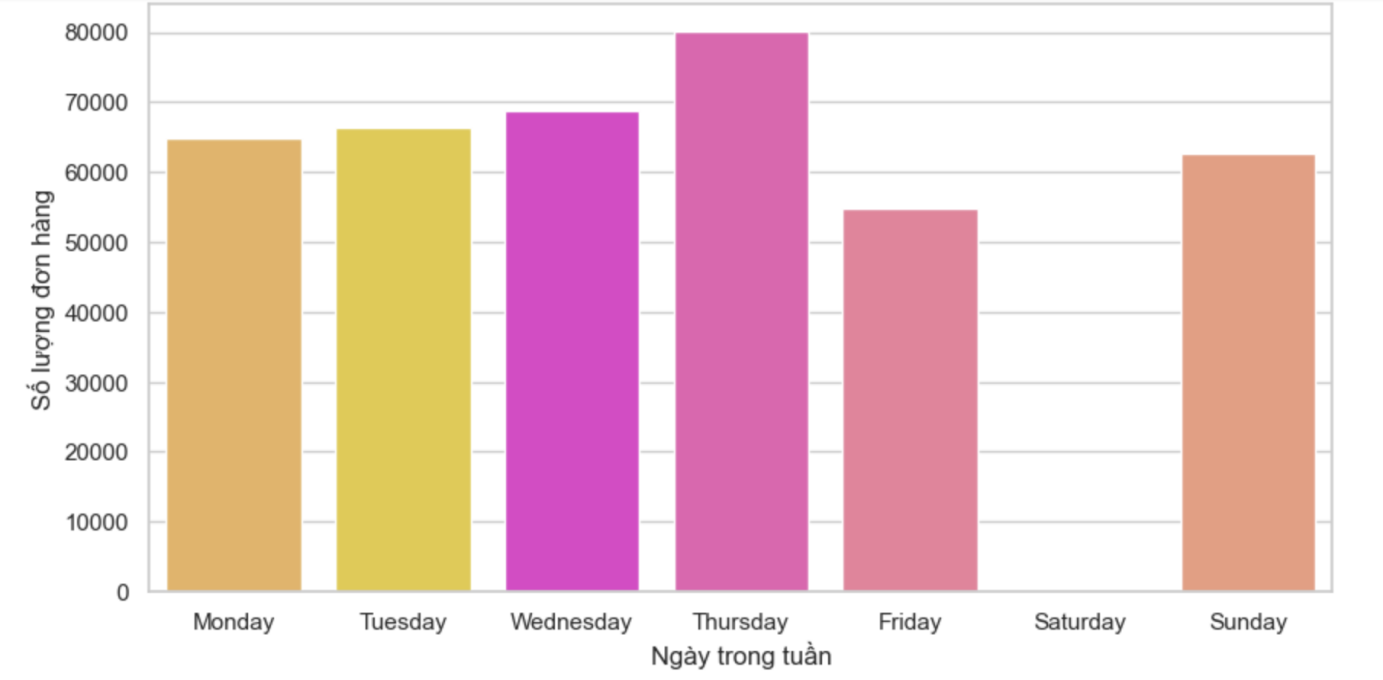
**- Giờ cao điểm:** Lượng đơn hàng bắt đầu tăng mạnh từ 10 giờ sáng và đạt đỉnh tuyệt đối vào lúc 12 giờ trưa.

+ Khung giờ vàng kéo dài từ 11 giờ sáng đến 15:00 giờ, chiếm phần lớn tổng số đơn hàng trong ngày.

+ Điều này hoàn toàn phù hợp với lịch trình sinh hoạt phổ biến. 12 giờ trưa là giờ nghỉ trưa của đa số nhân viên văn phòng và người lao động, họ tận dụng thời gian này để mua sắm. Khung giờ đầu giờ chiều cũng là lúc mọi người có thời gian rảnh rỗi hơn sau khi đã giải quyết các công việc buổi sáng.

**- Giờ thấp điểm:** Lượng đơn hàng rất thấp vào buổi sáng trước 9 giờ và giảm mạnh sau 16 giờ chiều. Vì buổi sáng là thời gian di chuyển và bắt đầu công việc, trong khi chiều tối là thời gian tan làm và dành cho gia đình, do đó nhu cầu mua sắm trực tuyến giảm đi.

### 2.6 Số lượng đơn hàng theo ngày trong tuần

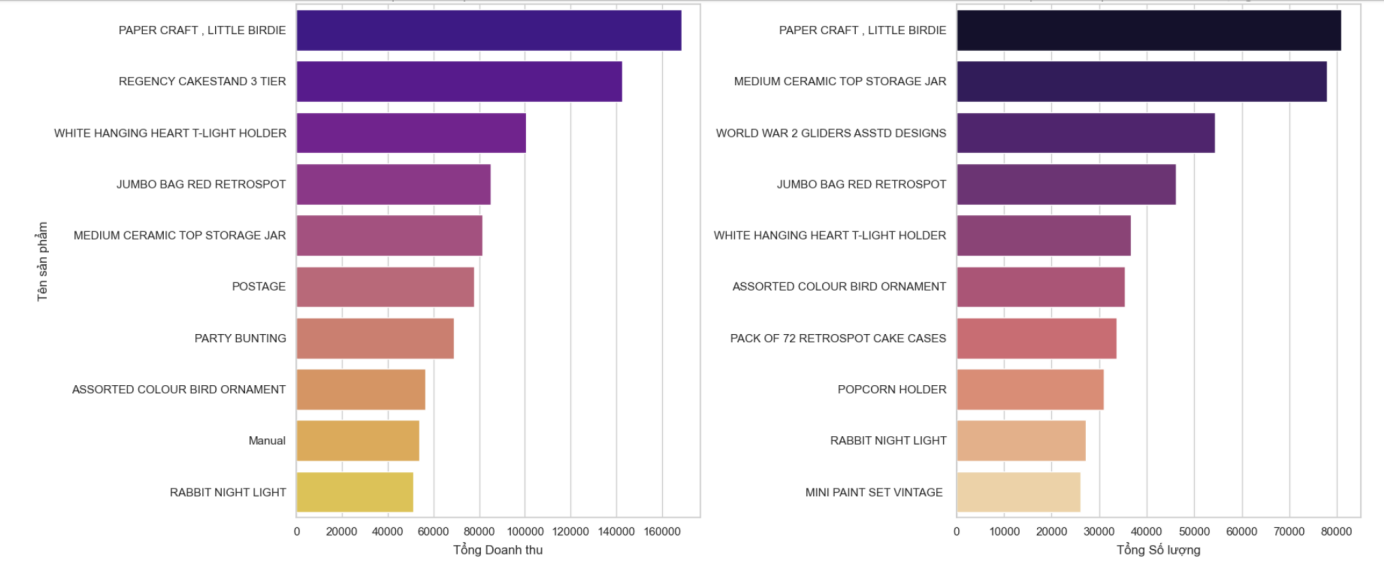


**Hình 13: Số lượng đơn hàng theo ngày trong tuần**

- Lượng đơn hàng giảm mạnh vào thứ sáu và đặc biệt là thứ bảy không có bất cứ đơn hàng nào. Điều này cho thấy có nhiều khả năng như lỗi dữ liệu (sau khi xóa bớt), công ty không xử lý/vận chuyển vào ngày thứ 7 hoặc là do khách hàng là các doanh nghiệp không hoạt động vào thứ 7.

- Qua đó cần phải xác định lại đối tượng khách hàng, chính sách làm việc vào thứ 7 vì đây là ngày mua sắm cao điểm của khách hàng cá nhân, marketing theo ngày, nhất là thứ 5 nhắm vào ngày có lượng khách hàng mua sắm cao nhất.

### 2.7 Biểu đồ phân tán sử dụng seaborn



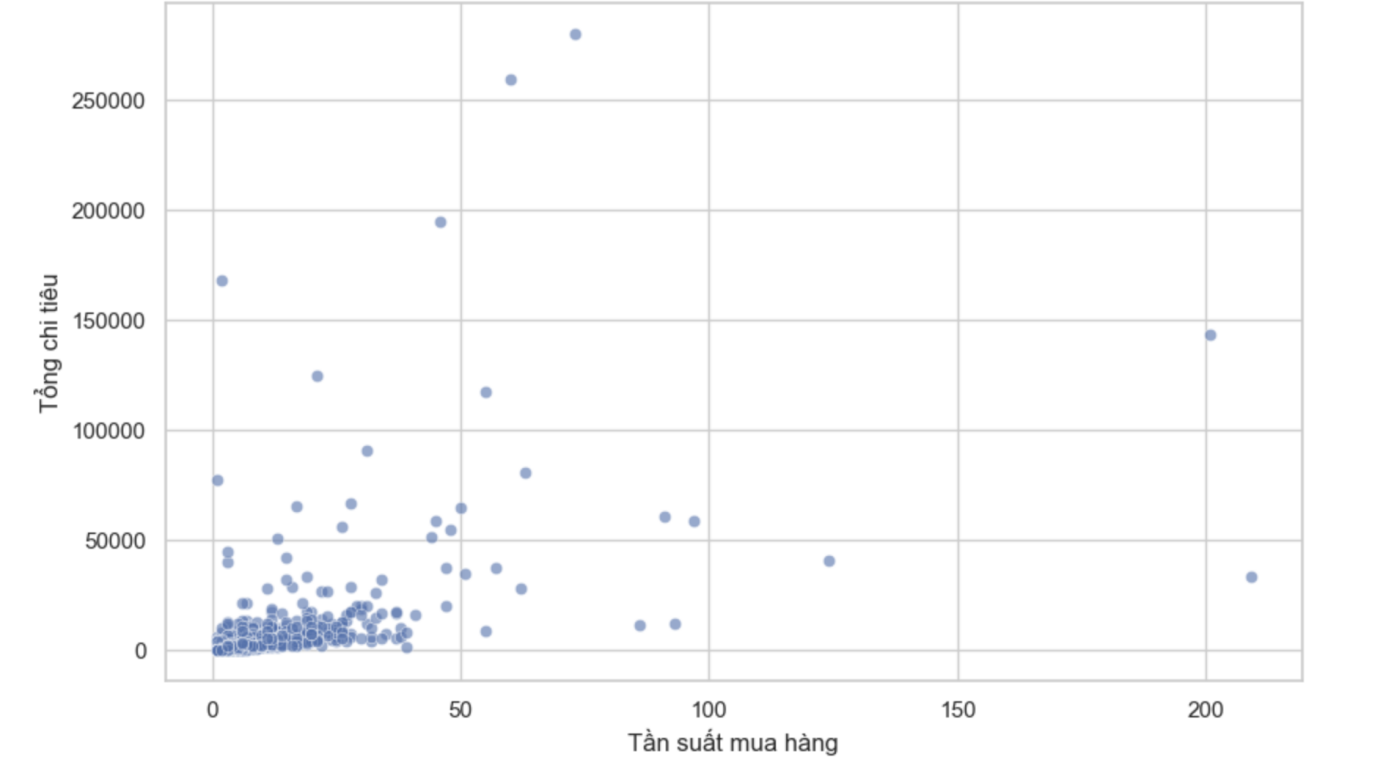
**Hình 14: Bảng so sánh giữa tổng doanh thu và tổng sản phẩm bán ra**

- Có sự khác biệt rõ rệt giữa hai danh sách. Mặc dù một số sản phẩm xuất hiện ở cả hai bên, vị trí xếp hạng của chúng thay đổi đáng kể. Điều này cho thấy cơ cấu sản phẩm của công ty bao gồm hai nhóm chính: nhóm sản phẩm phổ biến, giá rẻ (bán được số lượng lớn) và nhóm sản phẩm cao cấp, giá trị cao (mang lại doanh thu lớn).

- "REGENCY CAKESTAND 3 TIER" và "POSTAGE" xuất hiện ở vị trí rất cao trong top doanh thu nhưng không có mặt trong top số lượng. Điều này cho thấy đây là những sản phẩm có đơn giá cao. Dù không bán được nhiều về số lượng, mỗi sản phẩm bán ra đều đóng góp một phần doanh thu đáng kể.

- Ngược lại: "WORLD WAR 2 GLIDERS ASSTD DESIGNS" và "PACK OF 72 RETROSPOT CAKE CASES" nằm trong top cao của số lượng nhưng lại không có mặt trong top doanh thu. Đây là những sản phẩm giá rẻ, hấp dẫn, có vai trò thu hút khách hàng. Chúng giúp tăng số lượng giao dịch và tạo cơ hội để bán kèm các sản phẩm khác.

**2.8 Biểu đồ Phân tán (Scatter Plot): Tần suất vs tổng chi tiêu**

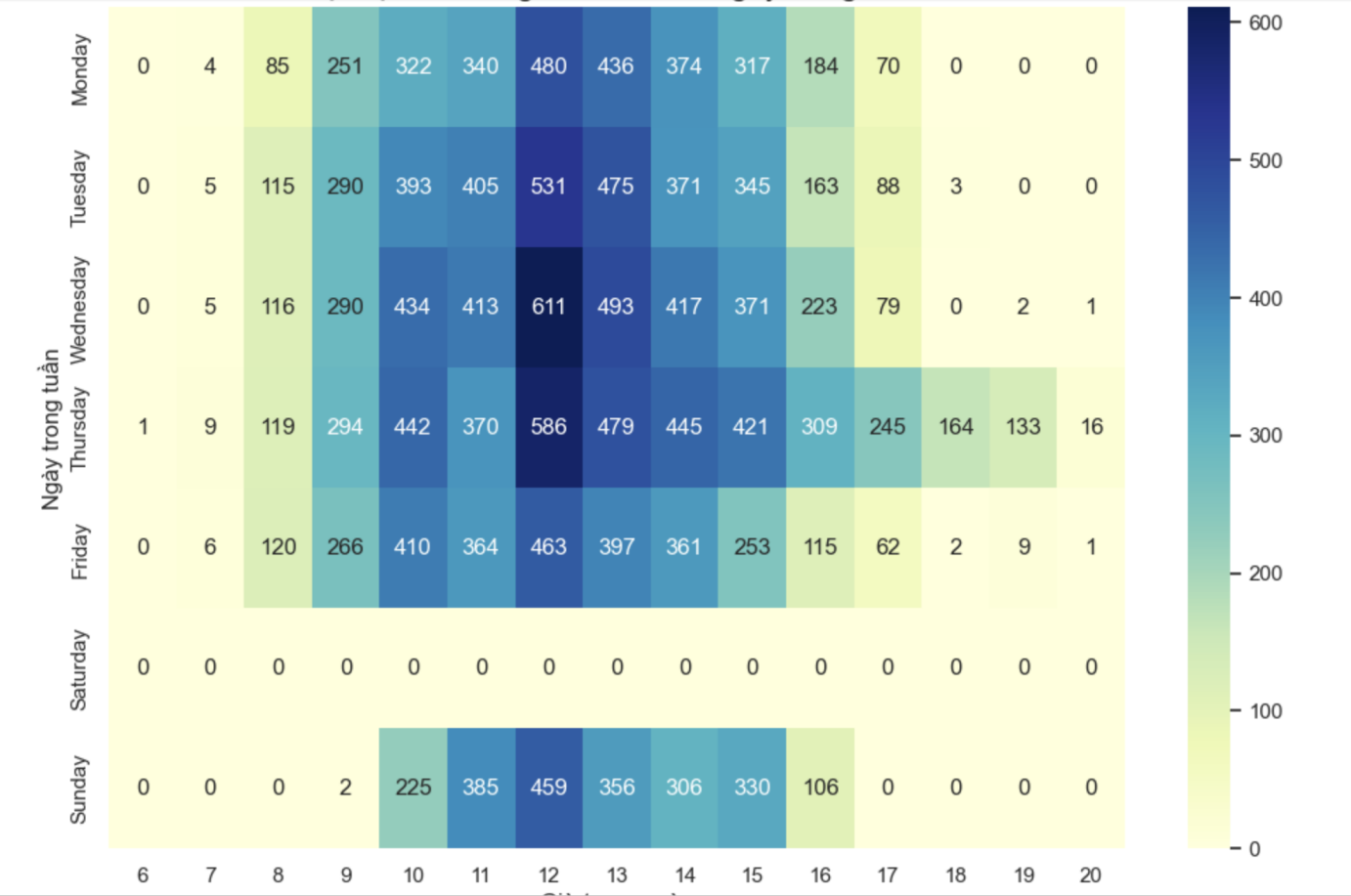


**Hình 15: Mối quan hệ giữa tuần suất và chi tiêu**

- Phần lớn khách hàng tập trung ở khu vực có Tần suất < 25 và Tổng chi tiêu < 25,000. Đây là nhóm khách hàng phổ thông, bao gồm cả khách hàng mới và khách hàng mua sắm không thường xuyên. Tệp khách hàng này cần được nuôi dưỡng để tăng tần suất và giá trị mua hàng.

- Khách hàng giá trị cao: Có một vài điểm dữ liệu nằm tách biệt ở phía trên và phía bên phải của biểu đồ. Đặc biệt là các điểm có Tổng chi tiêu > 100,000 hoặc tần suất > 50. Dù chỉ chiếm một số lượng rất nhỏ, họ lại đóng góp một phần doanh thu cực kỳ lớn. Đây là một trong những khách hàng giá trị nhất của công ty.

**2.9 Heatmap về mật độ đơn hàng theo giờ và ngày trong tuần**



**Hình 16: Mật độ đơn hàng theo giờ và ngày trong tuần**

- Tối ưu hóa hoạt động dựa trên "Vùng nóng": Mọi hoạt động quan trọng nhất từ marketing, chạy quảng cáo, hỗ trợ khách hàng, đến xử lý đơn hàng cần được dồn toàn lực vào khung giờ 10:00 - 16:00 từ Thứ Hai đến Thứ Sáu. Đây là thời điểm vàng để tối đa hóa doanh thu.

- Ưu tiên hàng đầu: Điều tra dữ liệu ngày Thứ Bảy: Cần phải xác minh ngay lập tức tại sao không có dữ liệu cho ngày Thứ Bảy. Nếu đây là một lỗi kỹ thuật hoặc một lỗ hổng trong quy trình vận hành, công ty đang bỏ lỡ một lượng doanh thu tiềm năng rất lớn, đặc biệt là từ tệp khách hàng cá nhân.

- Khai thác tiềm năng ngày Chủ Nhật: Dữ liệu cho thấy Chủ Nhật là một cơ hội. Công ty có thể thử nghiệm các chiến dịch email marketing hoặc quảng cáo nhỏ vào sáng Chủ Nhật, nhắm riêng vào các sản phẩm tiêu dùng cá nhân (thay vì các sản phẩm B2B) để xem có thể thúc đẩy doanh số vào ngày này hay không.

# III. KẾT LUẬN

Qua quá trình Phân tích Dữ liệu Khám phá (EDA) và xây dựng mô hình, dự án đã rút ra được những insight kinh doanh quan trọng từ bộ dữ liệu bán lẻ trực tuyến và chứng minh tiềm năng của việc ứng dụng khoa học dữ liệu để tối ưu hóa hoạt động kinh doanh.

**- Kết quả chính:**

* **Phân tích khám phá (EDA)** đã xác định thành công các mẫu hành vi mua sắm cốt lõi: khách hàng tập trung mua sắm vào **giờ trưa các ngày trong tuần**, doanh thu có **tính mùa vụ và đạt đỉnh vào tháng 11**. Phân tích cũng cho thấy sự **phụ thuộc lớn vào thị trường nội địa (UK)**.
* **Phân khúc khách hàng theo mô hình RFM** đã phân loại hiệu quả tệp khách hàng thành các nhóm có ý nghĩa như "Champions" (khách hàng vô địch), "Loyal Customers" (khách hàng trung thành), và "Hibernating" (khách hàng đang ngủ đông), tạo tiền đề cho các chiến dịch marketing cá nhân hóa.
* **Mô hình dự đoán Giá trị Vòng đời Khách hàng (CLV)** đã chứng tỏ khả năng dự báo doanh thu tương lai từ một khách hàng dựa trên hành vi quá khứ, giúp doanh nghiệp xác định các khách hàng giá trị cao.

**- Hướng phát triển:**

* **Nâng cấp mô hình:** Ứng dụng các thuật toán mạnh mẽ hơn như **XGBoost, LightGBM** để cải thiện độ chính xác của mô hình dự đoán CLV.
* **Xây dựng hệ thống gợi ý sản phẩm (Recommendation System):** Dựa trên lịch sử mua hàng để gợi ý các sản phẩm phù hợp, tăng giá trị trung bình đơn hàng.
* **Phân tích sâu hơn:** Kết hợp dữ liệu này với các nguồn dữ liệu khác (ví dụ: dữ liệu marketing, chi phí sản phẩm) để tính toán lợi nhuận trên từng phân khúc khách hàng.

**- Ý nghĩa thực tế:**

* **Hỗ trợ ra quyết định kinh doanh:** Cung cấp cơ sở dữ liệu vững chắc để Ban lãnh đạo đưa ra quyết định về phân bổ ngân sách marketing, quản lý hàng tồn kho, và chiến lược mở rộng thị trường.
* **Tối ưu hóa Marketing & Bán hàng:** Giúp đội ngũ marketing thiết kế các chiến dịch nhắm đúng mục tiêu cho từng phân khúc khách hàng, từ đó tăng tỷ lệ chuyển đổi và giữ chân khách hàng hiệu quả hơn.
* **Nâng cao trải nghiệm khách hàng:** Hiểu rõ hành vi khách hàng để cá nhân hóa trải nghiệm mua sắm, góp phần xây dựng lòng trung thành và nâng cao lợi thế cạnh tranh của doanh nghiệp.

# TÀI LIỆU THAM KHẢO

[1] https://www.kaggle.com/code/rajatkumar30/eda-online-retail.