TRƯỜNG KỸ THUẬT VÀ CÔNG NGHỆ

KHOA CÔNG NGHỆ THÔNG TIN

HỌC PHẦN: THỰC TẬP ĐỒ ÁN CHUYÊN NGÀNH

HỌC KỲ I, NĂM HỌC 2025-2026

TÌM HIỂU VỀ PHÂN TÍCH DỮ LIỆU (DATA ANALYTICS)

Giáo viên hướng dẫn:

TS. Nguyễn Bảo Ân

Sinh viên thực hiện:

Họ tên: Cô Nhân Quý
MSSV: 110122150
Lớp DA22TTB

Vĩnh Long, tháng 12 năm 2025

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Vĩnh Long, ngày .... tháng …… năm ……

Giảng viên hướng dẫn

(Ký tên và ghi rõ họ tên)

NHẬN XÉT CỦA THÀNH VIÊN HỘI ĐỒNG

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Vĩnh Long, ngày .... tháng …… năm ……

Thành viên hội đồng

(Ký tên và ghi rõ họ tên)

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến TS. Nguyễn Bảo Ân, giáo viên hướng dẫn, người đã tận tình hướng dẫn, chỉ bảo và tạo điều kiện thuận lợi để em trong suốt quá trình thực hiện Thực tập Đồ án chuyên ngành với đề tài " Tìm hiểu về Phân tích Dữ liệu ". Với những kiến thức quý báu của thầy đã giúp em vượt qua những khó khăn trong quá trình nghiên cứu và hoàn thiện báo cáo.

Sự nhiệt tình và kiên nhẫn của thầy đã giúp em hoàn thành đồ án một cách suôn sẻ và hiệu quả hơn. Em thật sự rất biết ơn và mong rằng sẽ tiếp tục nhận được sự giúp đỡ từ thầy trong những lần sau.

Xin chân thành cảm ơn thầy!

Sinh viên ký và ghi rõ họ và tên

Sinh viên thực hiện

Cô Nhân Quý

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

TÓM TẮT ĐỒ ÁN

Trong kỷ nguyên số, dữ liệu đóng vai trò then chốt trong việc định hình chiến lược kinh doanh, đặc biệt là trong lĩnh vực thương mại điện tử. Đồ án "Tìm hiểu về Phân tích dữ liệu (Data Analytics)" tập trung nghiên cứu và ứng dụng quy trình phân tích dữ liệu chuẩn mực để khai thác giá trị từ bộ dữ liệu giao dịch thực tế.

Nội dung chính của đồ án bao gồm:

1. **Nghiên cứu lý thuyết:** Tìm hiểu quy trình Data Analytics từ thu thập, làm sạch, xử lý đến trực quan hóa dữ liệu, cùng với các kỹ thuật thống kê và mô hình phân tích hành vi khách hàng (RFM).
2. **Thực nghiệm:** Sử dụng ngôn ngữ lập trình Python và các thư viện chuyên dụng (Pandas, Matplotlib, Seaborn) để xử lý bộ dữ liệu Online Retail (từ UCI Machine Learning Repository).
3. **Kết quả:** Xây dựng quy trình làm sạch dữ liệu hiệu quả (xử lý giá trị thiếu, ngoại lai), thực hiện phân tích thống kê mô tả về doanh thu, xu hướng tiêu dùng, và áp dụng mô hình RFM để phân khúc khách hàng.

Kết quả của đồ án cung cấp cái nhìn sâu sắc về hiệu quả hoạt động kinh doanh, giúp hỗ trợ ra quyết định chiến lược về Marketing và quản lý khách hàng, đồng thời là tài liệu tham khảo hữu ích cho việc ứng dụng Khoa học dữ liệu trong thực tiễn.

MỞ ĐẦU

Lý do chọn đề tài

Các chuyên gia và nhà quản lý đã nhấn mạnh, dữ liệu là nguồn tài nguyên chiến lược của nền kinh tế số và chuyển đổi số quốc gia, cần được bảo đảm an toàn thông tin, an ninh mạng và bảo vệ dữ liệu cá nhân. Việc nâng cao năng lực phân tích, xử lý và quản trị dữ liệu là yêu cầu cấp thiết với mọi tổ chức, doanh nghiệp và cá nhân [7].

Đó là một trong những nội dung của Tọa đàm "Xử lý và phân tích dữ liệu: Động lực cho chuyển đổi số quốc gia" diễn ra vào ngày 28/05/2025 đã cho thấy tầm quan trọng của việc phân tích dữ liệu.

Phân tích dữ liệu vốn là một môn học cũng như một ngành khoa học cung cấp phương pháp xử lý các dữ liệu thô thông qua các công cụ nhằm tìm hiểu các quy luật của dữ liệu được cho là bất thường hoặc không rõ ràng trong các tập dữ liệu.

Và mục tiêu của đề tài này là Tìm hiểu về phân tích dữ liệu (Data Analytics) - bao gồm các bước: thu thập, làm sạch, xử lý, và trực quan hóa dữ liệu. Thông qua việc sử dụng các thư viện Python phổ biến như Pandas, Matplotlib, Seaborn, sinh viên sẽ xây dựng một ứng dụng nhỏ phân tích dữ liệu CSV (ví dụ: điểm sinh viên hoặc doanh thu bán hàng), kèm dashboard trực quan đơn giản để thể hiện kết quả phân tích.

Bên cạnh đó, đề tài phù hợp với định hướng đào tạo của ngành Công nghệ thông tin, giúp sinh viên vận dụng kiến thức lập trình Python và xử lý dữ liệu vào bài toán phân tích dữ liệu thực tế. Tập dữ liệu Online Retail là dữ liệu công khai, có cấu trúc rõ ràng, đảm bảo tính khả thi trong quá trình thực hiện và phù hợp với phạm vi của một đồ án chuyên ngành.

Về mặt kinh tế và thực tiễn:

- Về mặt kinh tế, việc phân tích dữ liệu bán lẻ giúp doanh nghiệp tối ưu hóa hoạt động kinh doanh, giảm chi phí vận hành và nâng cao hiệu quả sử dụng nguồn lực thông qua việc ra quyết định dựa trên dữ liệu.

- Về mặt thực tiễn, đề tài mô phỏng một bài toán phân tích dữ liệu trong lĩnh vực thương mại điện tử, phản ánh dữ liệu giao dịch thực tế và có thể áp dụng cho các hệ thống bán lẻ trực tuyến hiện nay.

Về mặt kỹ thuật và công nghệ:

Về mặt kỹ thuật, đề tài sử dụng các thư viện Python phổ biến như Pandas, Matplotlib và Seaborn, giúp sinh viên rèn luyện kỹ năng xử lý, phân tích và trực quan hóa dữ liệu theo quy trình phân tích dữ liệu hiện đại.

Mục tiêu đề tài

Mục tiêu tổng quát:

Tìm hiểu và áp dụng quy trình phân tích dữ liệu (Data Analytics) trên tập dữ liệu bán lẻ trực tuyến dạng CSV (Online Retail), nhằm khai thác thông tin có giá trị phục vụ cho việc đánh giá hoạt động kinh doanh và hỗ trợ ra quyết định.

Mục tiêu cụ thể:

- Nghiên cứu quy trình phân tích dữ liệu, bao gồm các bước: thu thập dữ liệu, làm sạch dữ liệu, xử lý dữ liệu và trực quan hóa dữ liệu.

- Thực hiện làm sạch và tiền xử lý dữ liệu CSV, bao gồm xử lý giá trị thiếu, loại bỏ dữ liệu bất thường và chuẩn hóa dữ liệu giao dịch.

- Áp dụng các kỹ thuật phân tích thống kê và trực quan hóa dữ liệu để đánh giá doanh thu, sản phẩm bán chạy và hành vi mua sắm của khách hàng theo thời gian và theo khu vực.

- Sử dụng các thư viện Python phổ biến như Pandas, Matplotlib và Seaborn để xây dựng ứng dụng phân tích dữ liệu và dashboard trực quan đơn giản.

Đối tượng và dữ liệu nghiên cứu

Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là quy trình phân tích dữ liệu (Data Analytics) và các kỹ thuật xử lý, phân tích dữ liệu bán lẻ trực tuyến. Cụ thể, đề tài tập trung nghiên cứu cách thức thu thập, làm sạch, xử lý, phân tích và trực quan hóa dữ liệu giao dịch nhằm khai thác các thông tin có giá trị phục vụ cho việc đánh giá hoạt động kinh doanh và hành vi mua sắm của khách hàng.

Dữ liệu nghiên cứu:

Nguồn dữ liệu: UCI Machine Learning Repository – Online Retail Dataset [1]

Link: https://archive.ics.uci.edu/static/public/352/data.csv

Số lượng mẫu: 541909

Số lượng biến: 8

STT	Tên biến	Vai trò	Kiểu dữ liệu	Mô tả	Đơn vị
1	InvoiceNo	ID	Categorical	Mã hóa đơn, định danh duy nhất cho mỗi giao dịch. Nếu bắt đầu bằng "C" thì là hóa đơn hủy	
2	StockCode	ID	Categorical	Mã định danh cho từng sản phẩm	
3	Description	Feature	Categorical	Tên sản phẩm	
4	Quantity	Feature	Integer	Số lượng sản phẩm trong mỗi giao dịch	Sản phẩm
5	InvoiceDate	Feature	Date	Thời điểm phát sinh giao dịch	
6	UnitPrice	Feature	Continuous	Giá bán của một đơn vị sản phẩm	Bảng Anh (Sterling)
7	CustomerID	Feature	Categorical	Mã định danh khách hàng	
8	Country	Feature	Categorical	Quốc gia nơi khách hàng sinh sống	

Bảng 1 Bảng cấu trúc dữ liệu dataset

Ưu điểm của dữ liệu:

- Dữ liệu có nguồn gốc thực tế, phản ánh hoạt động bán lẻ trực tuyến.
- Dữ liệu có cấu trúc rõ ràng, dễ xử lý bằng Pandas.
- Chứa đầy đủ thông tin về giao dịch, sản phẩm, khách hàng, thời gian và quốc gia.
- Phù hợp để phân tích doanh thu, sản phẩm bán chạy và hành vi khách hàng.

Hạn chế của dữ liệu:

- Cột CustomerID có nhiều giá trị thiếu, gây hạn chế cho phân tích hành vi khách hàng.
- Tồn tại các giao dịch bất thường như:
  + Quantity âm (đơn hàng bị hủy)
  + InvoiceNo bắt đầu bằng "C"
- Một số cột chưa đúng kiểu dữ liệu (InvoiceDate đang là object).

Phương pháp xử lý dữ liệu:

- Chuyển đổi kiểu dữ liệu:

Chuyển InvoiceDate sang kiểu datetime.

- Xử lý dữ liệu thiếu:
  + Loại bỏ các bản ghi thiếu CustomerID khi thực hiện phân tích hành vi khách hàng (RFM).
  + Giữ lại các bản ghi này khi phân tích doanh thu tổng thể.
- Loại bỏ dữ liệu bất thường:
  + Loại bỏ các hóa đơn bị hủy (InvoiceNo bắt đầu bằng "C").
  + Loại bỏ các giao dịch có Quantity ≤ 0 hoặc UnitPrice ≤ 0.
- Tạo thuộc tính mới:

  Tính doanh thu cho mỗi giao dịch:

$$\text{Revenue} = \text{Quantity} \times \text{UnitPrice}$$

- Chuẩn hóa dữ liệu thời gian:

Trích xuất tháng, năm từ InvoiceDate để phân tích theo thời gian.

- Phân tích và trực quan hóa:

Sử dụng biểu đồ cột, biểu đồ đường, biểu đồ phân bố để rút ra nhận định.

Phương pháp thực hiện

Phương pháp lý thuyết:

Đề tài áp dụng phương pháp phân tích dữ liệu (Data Analytics) nhằm khai thác thông tin có giá trị từ dữ liệu bán lẻ trực tuyến. Quy trình phân tích dữ liệu được thực hiện theo các bước cơ bản gồm: thu thập dữ liệu, làm sạch dữ liệu, xử lý dữ liệu, phân tích thống kê và trực quan hóa dữ liệu.

Ưu điểm:

- Giúp khai thác hiệu quả thông tin có giá trị từ dữ liệu bán lẻ trực tuyến thông qua quy trình phân tích dữ liệu rõ ràng và có hệ thống.
- Dễ triển khai với dữ liệu dạng bảng (CSV), phù hợp với các tập dữ liệu thực tế trong thương mại điện tử.
- Cho phép trực quan hóa kết quả phân tích, hỗ trợ người dùng dễ dàng quan sát xu hướng, hành vi khách hàng và hiệu quả kinh doanh.
- Sử dụng các công cụ phổ biến như Pandas [2][3], Matplotlib [4], Seaborn [5], giúp kết quả phân tích dễ tái sử dụng và mở rộng.
- Phù hợp với phạm vi đồ án chuyên ngành, giúp sinh viên rèn luyện kỹ năng xử lý và phân tích dữ liệu thực tế.

Hạn chế:

- Kết quả phân tích phụ thuộc nhiều vào chất lượng dữ liệu đầu vào; dữ liệu thiếu hoặc bất thường có thể ảnh hưởng đến độ chính xác.
- Chủ yếu dừng lại ở mức phân tích mô tả và khám phá dữ liệu, chưa tập trung vào dự đoán hay học máy nâng cao.
- Việc xử lý dữ liệu lớn có thể tốn thời gian và tài nguyên nếu không tối ưu tốt.
- Khả năng khái quát hóa kết quả còn hạn chế do dữ liệu chỉ phản ánh một tập giao dịch trong một giai đoạn nhất định.

Các hạn chế trên được khắc phục thông qua quá trình làm sạch dữ liệu và lựa chọn phương pháp phân tích phù hợp với mục tiêu nghiên cứu

Phương pháp thu thập dữ liệu:

- Nguồn dữ liệu: Tập dữ liệu Online Retail Dataset
- Sử dụng dữ liệu có sẵn từ dataset: Online Retail Dataset từ UCI Machine Learning Repository, áp dụng trong giai đoạn chuẩn bị dữ liệu;
- URL: https://archive.ics.uci.edu/static/public/352/data.csv ;
- Phương thức thu thập: Download trực tiếp từ repository công khai.

Phương pháp thực nghiệm:

Môi Trường Thực Nghiệm: Google Colab:

- Ngôn ngữ: Python 3.x.
- Thư viện chính:
  + pandas: Xử lý và phân tích dữ liệu
  + numpy: Tính toán số học
  + matplotlib, seaborn: Trực quan hóa dữ liệu
  + scikit-learn: Machine learning algorithms

Quy trình thực nghiệm:

1. **Khảo sát và kiểm tra dữ liệu**

   Đánh giá kích thước dữ liệu, kiểu dữ liệu và tình trạng giá trị thiếu.

2. **Làm sạch và tiền xử lý dữ liệu**
   - Xử lý giá trị thiếu
   - Loại bỏ dữ liệu bất thường (hóa đơn hủy, số lượng âm, đơn giá bằng 0)
   - Chuẩn hóa dữ liệu.

3. **Phân tích thống kê dữ liệu**

   Tính toán doanh thu, sản phẩm bán chạy và các chỉ số thống kê cơ bản.

4. **Trực quan hóa dữ liệu**

   Thể hiện kết quả phân tích thông qua các biểu đồ như biểu đồ cột, biểu đồ đường và biểu đồ phân bố.

5. **Phân tích hành vi khách hàng**

   Áp dụng phương pháp RFM để phân nhóm và đánh giá khách hàng.

6. **Tổng hợp và đánh giá kết quả**

   Rút ra nhận định về xu hướng kinh doanh và hành vi tiêu dùng.

Trực quan hóa kết quả:

Kết quả phân tích được trình bày thông qua các biểu đồ trực quan nhằm hỗ trợ người dùng dễ dàng quan sát và đánh giá, bao gồm:

- Biểu đồ doanh thu theo thời gian;
- Biểu đồ sản phẩm bán chạy;
- Biểu đồ phân bố khách hàng theo quốc gia;

CHƯƠNG 1: TỔNG QUAN

1.1. Tổng quan về Phân tích dữ liệu (Data Analytics)

Khái niệm

Phân tích dữ liệu (Data Analytics) là quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu nhằm mục đích khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định. Đây là một lĩnh vực đa ngành kết hợp giữa khoa học máy tính, thống kê và kiến thức nghiệp vụ để chuyển đổi dữ liệu thô thành tri thức hành động (actionable insights).

Trong bối cảnh chuyển đổi số, phân tích dữ liệu không chỉ dừng lại ở việc báo cáo những gì đã xảy ra mà còn hướng tới việc giải thích tại sao nó xảy ra và dự báo xu hướng trong tương lai.

Vai trò và Ý nghĩa thực tiễn:

Dữ liệu hiện nay được xem là nguồn tài nguyên chiến lược của mọi tổ chức. Việc ứng dụng phân tích dữ liệu mang lại những lợi ích thiết thực:

- **Hỗ trợ ra quyết định:** Chuyển từ việc ra quyết định dựa trên cảm tính sang ra quyết định dựa trên dữ liệu (Data-driven decision making), giúp giảm thiểu rủi ro và tăng độ chính xác.
- **Tối ưu hóa vận hành:** Giúp doanh nghiệp nhận diện các điểm nghẽn trong quy trình, quản lý tồn kho hiệu quả và cắt giảm chi phí.
- **Thấu hiểu khách hàng:** Phân tích hành vi mua sắm giúp cá nhân hóa trải nghiệm khách hàng, từ đó gia tăng lòng trung thành và doanh thu (như việc áp dụng mô hình RFM trong đồ án này).

1.2. Phân loại các dạng phân tích

Trong đồ án này, chúng ta tập trung vào hai loại hình phân tích nền tảng:

- **Phân tích mô tả (Descriptive Analytics):** Sử dụng các chỉ số thống kê (trung bình, trung vị, độ lệch chuẩn) và biểu đồ để tóm tắt dữ liệu lịch sử (ví dụ: Doanh thu tháng trước là bao nhiêu? Sản phẩm nào bán chạy nhất?).
- **Phân tích chẩn đoán (Diagnostic Analytics):** Đi sâu vào tìm kiếm nguyên nhân gốc rễ của các xu hướng thông qua việc phân tích tương quan và khám phá dữ

liệu đa chiều (ví dụ: Tại sao doanh thu giảm vào…? Có phải do yếu tố mùa vụ hay thiếu hàng?).

1.3. Quy trình phân tích dữ liệu chuẩn (Data Analysis Process)

Đồ án tuân theo quy trình chuẩn gồm các bước sau:

1. **Xác định vấn đề:**

Tìm hiểu rõ mục tiêu (ví dụ đối với Kinh doanh: Tăng doanh thu, phân nhóm khách hàng).

2. **Thu thập dữ liệu (Data Collection):**

Tập hợp dữ liệu từ các nguồn uy tín (dữ liệu đồ án là bộ dữ liệu Online Retail từ UCI Repository).

3. **Tiền xử lý dữ liệu (Data Preprocessing):**

Đây là bước quan trọng nhất, bao gồm:

- Làm sạch dữ liệu (Data Cleaning): Xử lý giá trị thiếu (Missing values), loại bỏ dữ liệu trùng lặp và các giá trị nhiễu (Outliers).
- Chuyển đổi dữ liệu (Data Transformation): Chuẩn hóa định dạng thời gian, tạo các biến mới (Feature Engineering) như tính tổng doanh thu từ đơn giá và số lượng.

4. **Phân tích và Mô hình hóa (Analysis & Modeling):**

Áp dụng các kỹ thuật thống kê và thuật toán (như RFM) để tìm ra các mẫu (patterns) trong dữ liệu.

5. **Trực quan hóa (Data Visualization):**

Biểu diễn kết quả thông qua các biểu đồ trực quan (Bar chart, Line chart, Scatter plot) giúp thông tin trở nên dễ hiểu.

CHƯƠNG 2: NGHIÊN CỨU LÝ THUYẾT

2.1. Cơ sở lý thuyết về Phân tích thống kê

Phân tích thống kê là cơ sở lý thuyết thứ nhất được áp dụng trong đề tài "Tìm hiểu về Phân tích dữ liệu" dựa vào dữ liệu bán lẻ trực tuyến (Online Retail CSV) nhằm khám phá và hiểu rõ đặc điểm của dữ liệu giao dịch.

Phương pháp này tập trung vào việc tóm tắt và mô tả các thông tin quan trọng của khách hàng, sản phẩm và giao dịch thông qua các chỉ số trung tâm, độ biến thiên và phân phối dữ liệu.

Ứng dụng cụ thể:

- Phân tích phân phối số lượng giao dịch theo thời gian (ngày, tháng, năm);
- Phân tích số lượng khách hàng, tần suất mua hàng và giá trị đơn hàng;
- So sánh doanh thu trung bình giữa các quốc gia và nhóm khách hàng;
- Xác định các sản phẩm bán chạy và sản phẩm có doanh thu cao.

Ý nghĩa:

Phân tích thống kê mô tả giúp cung cấp cái nhìn tổng quan về hoạt động kinh doanh bán lẻ trực tuyến, hỗ trợ doanh nghiệp nhận diện xu hướng tiêu dùng, hành vi mua sắm của khách hàng và các yếu tố ảnh hưởng đến doanh thu, làm cơ sở cho các bước phân tích chuyên sâu tiếp theo.

Phân tích tương quan và kiểm định giả thuyết:

Phân tích tương quan được sử dụng để:

- Đánh giá mối quan hệ giữa các biến số: số lượng mua, giá sản phẩm và doanh thu;
- Phát hiện hiện tượng đa cộng tuyến giữa các biến định lượng;
- Xác định các biến có ảnh hưởng mạnh đến tổng doanh thu hoặc giá trị đơn hàng.

2.2. Cơ sở lý thuyết Phân tích dữ liệu và Mô hình RFM

Phân tích dữ liệu (Data Analytics) là nền tảng lý thuyết chính được sử dụng trong đề tài để khai thác thông tin giá trị từ dữ liệu bán lẻ trực tuyến. Phân tích dữ liệu cho phép xử lý, khám phá và trực quan hóa dữ liệu nhằm hỗ trợ ra quyết định kinh doanh.

Các kỹ thuật phân tích chính được áp dụng:

- Phân tích mô tả (Descriptive Analytics): Tổng hợp và mô tả dữ liệu lịch sử;
- Phân tích xu hướng (Trend Analysis): Đánh giá sự thay đổi doanh thu và hành vi mua sắm theo thời gian;
- Phân tích phân nhóm khách hàng (Customer Segmentation) dựa trên tần suất mua và giá trị đơn hàng.

Ý nghĩa:

Phân tích dữ liệu bán lẻ trực tuyến giúp doanh nghiệp hiểu rõ nhu cầu khách hàng, tối ưu chiến lược marketing, quản lý tồn kho hiệu quả và nâng cao năng lực cạnh tranh trên thị trường thương mại điện tử.

2.3. Giới thiệu về bộ dữ liệu

2.3.1. Mô tả bộ dữ liệu

Vấn đề nghiên cứu:

Bộ dữ liệu Online Retail được sử dụng để phân tích hành vi mua sắm của khách hàng trong môi trường thương mại điện tử. Mục tiêu của đề tài là khai thác dữ liệu giao dịch nhằm nhận diện xu hướng tiêu dùng, đánh giá hiệu quả kinh doanh và hỗ trợ doanh nghiệp đưa ra quyết định chiến lược.

Nguồn dữ liệu và tác giả gốc:

Dữ liệu được thu thập từ UCI Machine Learning Repository [1], đây là bộ dữ liệu giao dịch gồm tất cả các giao dịch diễn ra từ ngày 01/12/2010 đến 09/12/2011 của một doanh nghiệp bán lẻ trực tuyến không có cửa hàng, có trụ sở và đăng ký tại Vương quốc Anh. Công ty chủ yếu bán các món quà độc đáo cho mọi dịp. Nhiều khách hàng của công ty là các nhà buôn bán.

Bộ dữ liệu được tạo ra bởi: Daqing Chen (Trường Kỹ thuật, Đại học South Bank London)

Tình trạng dữ liệu hiện có:

Bộ dữ liệu Online Retail gồm 541.909 bản ghi và phản ánh dữ liệu giao dịch bán lẻ thực tế. Qua kiểm tra, đa số các cột có dữ liệu đầy đủ, tuy nhiên cột CustomerID có tình trạng thiếu hụt dữ liệu khá cao (khoảng 24,93%), cho thấy sự tồn tại của các giao dịch từ khách hàng vãng lai.

Cột Description chỉ thiếu một tỷ lệ nhỏ dữ liệu (0,27%). Dữ liệu xuất hiện các giá trị bất thường như số lượng và đơn giá âm, chủ yếu liên quan đến các giao dịch hủy hoặc hoàn trả.

Cột InvoiceDate ban đầu ở dạng chuỗi đã được chuyển đổi sang kiểu thời gian để phục vụ phân tích theo thời gian.

Ngoài ra, dữ liệu có sự mất cân đối về mặt địa lý khi phần lớn giao dịch đến từ Vương quốc Anh..

2.3.2. Danh sách các biến

Bộ dữ liệu chứa 8 thuộc tính chính:

STT	Tên biến	Vai trò	Kiểu dữ liệu	Mô tả	Đơn vị
1	InvoiceNo	ID	Categorical	Mã hóa đơn, định danh duy nhất cho mỗi giao dịch. Nếu bắt đầu bằng “C” thì là hóa đơn hủy	
2	StockCode	ID	Categorical	Mã định danh cho từng sản phẩm	
3	Description	Feature	Categorical	Tên sản phẩm	
4	Quantity	Feature	Interger	Số lượng sản phẩm trong mỗi giao dịch	Sản phẩm
5	InvoiceDate	Feature	Date	Thời điểm phát sinh giao dịch	
6	UnitPrice	Feature	Continuous	Giá bán của một đơn vị sản phẩm	Bảng Anh (Sterling)

STT	Tên biến	Vai trò	Kiểu dữ liệu	Mô tả	Đơn vị
7	CustomerID	Feature	Categorical	Mã định dnah khách hàng	
8	Country	Feature	Categorical	Quốc gia nơi khách hàng sinh sống	

Bảng 2 Mô tả các biến trong Bộ dữ liệu

- Xét về cấu trúc bộ dữ liệu này gồm nhiều loại biến khác nhau:
  + Biến định danh: InvoiceNo, StockCode, CustomerID.
  + Biến định lượng: Quantity, UnitPrice.
  + Biến định tính: Description, Country.
- Xét về chất lượng dữ liệu thì bộ dữ liệu tồn tại nhiều thiếu sót:
  + Cột CustomerID có tỷ lệ giá trị thiếu tương đối lớn (khoảng 25%), phản ánh thực tế rằng một phần giao dịch đến từ khách hàng vãng lai không đăng nhập hoặc không cung cấp thông tin định danh.
  + Cột Description cũng tồn tại một tỷ lệ nhỏ giá trị thiếu, đã được xử lý bằng cách thay thế bằng giá trị mặc định, đảm bảo tính nhất quán của dữ liệu. Cột InvoiceDate được chuẩn hóa sang định dạng thời gian (datetime), tạo điều kiện thuận lợi cho việc phân tích xu hướng theo ngày, tháng và mùa vụ.
  + Các biến định lượng như Quantity và UnitPrice, dữ liệu cho thấy sự xuất hiện của các giá trị bất thường (ví dụ số lượng hoặc đơn giá âm), chủ yếu do các giao dịch hủy hoặc trả hàng.

2.3.3. Quá trình tiền xử lý dữ liệu

Kiểm tra và xử lý giá trị thiếu:

Dữ liệu được kiểm tra giá trị thiếu bằng thư viện *pandas*. Kết quả cho thấy cột CustomerID có tỷ lệ thiếu dữ liệu khá cao (khoảng 24,93%), trong khi cột *Description* chỉ thiếu dữ liệu một tỷ lệ nhỏ (0,27%).

Cách xử lý được áp dụng như sau:

- Đối với CustomerID: không loại bỏ dữ liệu mà tạo biến đánh dấu IsCustomerMissing để phân biệt khách hàng đã đăng ký và khách vãng lai; đồng thời gán giá trị "GUEST" vào cột CustomerID_Filled để phục vụ phân tích tổng hợp.

- Đối với Description: các giá trị thiếu được điền bằng chuỗi "Unknown Description".

Lý do: Việc giữ lại các giao dịch thiếu CustomerID giúp bảo toàn quy mô dữ liệu và phản ánh đúng thực tế hoạt động bán lẻ trực tuyến.

Chuẩn hóa và chuyển đổi dữ liệu:

- Cột InvoiceDate được chuyển đổi từ kiểu chuỗi sang kiểu datetime để phục vụ phân tích theo thời gian (ngày, tháng, giờ).

- Các cột số như Quantity và UnitPrice được kiểm tra định dạng để đảm bảo sẵn sàng cho các phép tính doanh thu và thống kê mô tả.

Lý do: Dữ liệu thời gian và dữ liệu số cần ở đúng định dạng để tránh sai lệch trong quá trình phân tích.

Xử lý dữ liệu bất thường:

Trong bộ dữ liệu Online Retail, một số giao dịch có giá trị *Quantity* hoặc *UnitPrice* âm được xác định là các đơn hàng bị hủy hoặc hoàn trả. Thay vì loại bỏ hoàn toàn, các bản ghi này được giữ lại và phân tích riêng biệt nhằm phản ánh đúng nghiệp vụ bán lẻ thực tế. Việc này giúp tránh làm sai lệch kết quả phân tích doanh thu và đồng thời cung cấp thông tin quan trọng về hành vi hoàn trả của khách hàng.

Loại bỏ giá trị ngoại lai:

Do dữ liệu ban đầu tồn tại một số giao dịch có số lượng, đơn giá hoặc doanh thu chênh lệch rất lớn so với phần lớn các giao dịch còn lại (có thể do lỗi nhập liệu, đặc thù thu mua số lượng lớn,... không thể phản ánh đúng hoạt động kinh doanh làm ảnh hưởng đến quá trình phân tích dữ liệu).

Để phát hiện các giá trị ngoại lai:

- So sánh giá trị lớn nhất (Max) với phân vị 99% (99th percentile);
- Kết hợp kiến thức nghiệp vụ bán lẻ, đặt các ngưỡng hợp lý cho Quantity, UnitPrice và DoanhThu;
- Tham khảo thêm phương pháp IQR (Interquartile Range) để đánh giá mức độ phân tán dữ liệu.

Thay vì loại bỏ toàn bộ các giá trị cực đoan thì lựa chọn:

- Loại bỏ các giá trị vượt ngưỡng phi thực tế (ví dụ số lượng hoặc giá bán quá lớn);
- Giữ lại phần lớn dữ liệu hợp lệ bằng cách giới hạn theo phân vị (percentile);
- Ưu tiên bảo toàn dữ liệu để đảm bảo tính đại diện của tập dữ liệu.

Lý do:

- Trong phân tích dữ liệu bán hàng, các giá trị được xem là "ngoại lai" không phải lúc nào cũng là lỗi dữ liệu.
- Một số trường hợp phản ánh các giao dịch có giá trị cao hoặc hành vi mua sắm đặc biệt.

Do đó, việc xử lý cần được thực hiện cẩn thận để tránh làm mất các thông tin quan trọng.

Cân bằng dữ liệu (Class Balancing):

Dữ liệu bán hàng thường có:

- Phân phối không đồng đều giữa các mức doanh thu;
- Một số ít giao dịch chiếm tỷ trọng lớn trong tổng doanh thu.
- Sự mất cân bằng này có thể:
- Làm sai lệch các chỉ số thống kê tổng hợp;
- Khiến các phân tích xu hướng và hành vi khách hàng bị thiên lệch.

Hướng tiếp cận xử lý

- Không áp dụng các kỹ thuật cân bằng lớp như SMOTE, do không phù hợp với bài toán phân tích dữ liệu bán hàng;
- Tập trung giảm ảnh hưởng của các giao dịch cực đoan thông qua bước xử lý ngoại lai;
- Ưu tiên các phương pháp thống kê bền vững (median, percentile) trong quá trình phân tích.

Lý do:

Việc áp dụng các kỹ thuật cân bằng dữ liệu không phù hợp có thể làm sai bản chất dữ liệu ban đầu. Do đó, đồ án lựa chọn cách tiếp cận phù hợp hơn với đặc thù dữ liệu bán lẻ, nhằm đảm bảo kết quả phân tích phản ánh đúng thực tế kinh doanh.

2.3.4. Bộ dữ liệu hoàn chỉnh

Tổng quan bộ dữ liệu

Sau quá trình tiền xử lý, bộ dữ liệu có:

- **Số quan sát:** 526.003 giao dịch
- **Số biến:** 14 biến (13 biến độc lập + 1 biến phụ thuộc). Từ 8 biến ban đầu loại bỏ biến CustomerID và bổ sung thêm 7 biến: IsCustomerMissing, CustomerID_Filled, Nam, Thang, Ngay, Tuan và Doanh Thu.
    + 13 biến độc lập (InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, Country, IsCustomerMissing, CustomerID_Filled, Nam, Thang, Ngay, Tuan)
    + 1 biến phụ thuộc (DoanhThu)
- **Tình trạng dữ liệu:**
    + Không còn giá trị bị thiếu (Missing = 0)
    + Kiểu dữ liệu đã được chuẩn hóa phù hợp cho phân tích và mô hình hóa

Bộ dữ liệu sau tiền xử lý phản ánh tương đối đầy đủ và ổn định các giao dịch bán lẻ hợp lệ, sẵn sàng cho các bước phân tích sâu hơn.

Thống kê mô tả

- **Mean và Median:**

  + Các biến Quantity, UnitPrice và DoanhThu có giá trị trung bình lớn hơn trung vị, cho thấy dữ liệu có xu hướng phân phối lệch phải (right-skewed).

  + Ví dụ, Quantity có trung vị là 3 nhưng trung bình là 8.63, cho thấy phần lớn giao dịch mua số lượng nhỏ, trong khi một số ít giao dịch có số lượng lớn kéo giá trị trung bình tăng lên.

- **Độ lệch chuẩn (Std):**

  + Quantity (22.85) và DoanhThu (21.30) có độ lệch chuẩn tương đối lớn, phản ánh sự đa dạng trong hành vi mua sắm giữa các giao dịch.

  + UnitPrice có độ lệch chuẩn 5.14, cho thấy mức giá sản phẩm dao động đáng kể giữa các mặt hàng.

- **Min và Max:**

  + Quantity dao động từ 1 đến 5.568 đơn vị;

  + UnitPrice dao động từ 0 đến 183.55;

  + DoanhThu dao động từ 0 đến 183.60.

Sự chênh lệch này cho thấy phạm vi rộng của giá trị giao dịch, từ các đơn hàng nhỏ lẻ đến các giao dịch có giá trị tương đối cao.

- **Các biến thời gian:** Các biến Nam, Thang, Ngay và Tuan được trích xuất từ InvoiceDate, giúp hỗ trợ phân tích xu hướng theo thời gian (theo năm, tháng, tuần và ngày).

- **Biến nhị phân:** Biến IsCustomerMissing chỉ nhận hai giá trị True/False, cho phép phân biệt giữa giao dịch có thông tin khách hàng và giao dịch khách vãng lai, thuận lợi cho việc phân tích hành vi khách hàng.

Phân phối của bộ dữ liệu

Quantity (Số lượng sản phẩm):

Phân phối lệch phải rõ rệt, phần lớn giao dịch có số lượng nhỏ (1–5 sản phẩm), chỉ một tỷ lệ rất nhỏ có số lượng mua lớn.

UnitPrice (Đơn giá):

Phân phối lệch phải, với đa số sản phẩm có mức giá thấp đến trung bình, trong khi một số ít sản phẩm có giá cao hơn đáng kể.

Doanh Thu:

Phân phối lệch phải mạnh, phản ánh đặc trưng phổ biến trong dữ liệu bán lẻ, nơi phần lớn giao dịch có giá trị thấp nhưng một số giao dịch giá trị cao đóng góp đáng kể vào tổng doanh thu.

Thời gian giao dịch:

Phân phối theo tháng và tuần cho thấy tính chu kỳ trong hoạt động mua sắm, phù hợp với đặc điểm kinh doanh bán lẻ theo mùa và thời điểm.

CHƯƠNG 3: HIỆN THỰC HÓA NGHIÊN CỨU

3.1. Công cụ và môi trường phát triển

- N**gôn ngữ lập trình:** Python (Phiên bản 3.10+). Đây là ngôn ngữ phổ biến nhất hiện nay cho Khoa học dữ liệu nhờ cú pháp đơn giản và cộng đồng hỗ trợ lớn.
- **Nền tảng thực nghiệm:**
  - Google Colab: Sử dụng để chạy các đoạn mã phân tích, huấn luyện mô hình và kiểm thử nhanh nhờ ưu điểm không cần cài đặt môi trường cục bộ và hỗ trợ GPU miễn phí.

Hình 1 Công cụ Google Colab

  - Visual Studio Code (VS Code): Môi trường phát triển tích hợp (IDE) dùng để tổ chức cấu trúc dự án và chỉnh sửa mã nguồn ứng dụng Streamlit trên máy cục bộ.

Hình 2 Phần mềm Visual Studio Code

- **Quản lý mã nguồn:** Git và GitHub để lưu trữ và quản lý phiên bản mã nguồn.

- **Các thư viện Python sử dụng (Libraries)**

  - **Thu thập và Xử lý dữ liệu:**

    + Pandas: Thư viện nòng cốt dùng để đọc file CSV (pd.read_csv), làm sạch dữ liệu (xử lý Null, Outlier), và thao tác trên DataFrame (bảng dữ liệu).

    + NumPy: Hỗ trợ các phép toán số học trên mảng và ma trận, phục vụ cho các tính toán thống kê nền tảng.

    + ucimlrepo: Thư viện chuyên dụng để tải trực tiếp bộ dữ liệu *Online Retail* từ kho lưu trữ UCI Machine Learning Repository.

  - **Trực quan hóa dữ liệu (Data Visualization):**

    + Matplotlib: Thư viện vẽ biểu đồ nền tảng, dùng để tạo các biểu đồ cơ bản.

    + Seaborn: Được xây dựng trên nền Matplotlib, giúp vẽ các biểu đồ thống kê đẹp mắt và trực quan hơn (như Heatmap, Boxplot, Distribution plot).

  - **Học máy và Thống kê (Machine Learning & Statistics):**

    + Scikit-learn (sklearn): Cung cấp các công cụ mạnh mẽ cho:

    + Tiền xử lý: StandardScaler (chuẩn hóa dữ liệu), SimpleImputer (xử lý giá trị thiếu).

    + Mô hình hóa: Hỗ trợ các thuật toán như K-Means (cho phân cụm) hoặc Logistic Regression (nếu có sử dụng cho dự đoán).

    + Đánh giá: Cung cấp các chỉ số như accuracy_score, confusion_matrix.

    + Graphviz: Dùng để trực quan hóa các cấu trúc đồ thị hoặc cây quyết định (nếu có).

  - **Xây dựng Ứng dụng Web (Web Application):**

    + Streamlit: Framework giúp chuyển đổi các đoạn mã phân tích dữ liệu thành ứng dụng web tương tác (Dashboard) một cách nhanh chóng mà không cần kiến thức chuyên sâu về Front-end.

    + Pyngrok: Công cụ hỗ trợ tạo đường hầm (tunnel) để public ứng dụng Streamlit đang chạy trên Google Colab ra internet để người dùng có thể truy cập.

3.2. Phân tích dữ liệu

Thu thập dữ liệu →Tiền xử lí dữ liệu →Phân tích thống kê mô tả →Phân tích hành vi khách hàng (RFM) →Đánh giá và biện luận.

Chi tiết các bước trong Workflow

Bước 1: Thu thập dữ liệu

- Mô tả: Dữ liệu được sử dụng trong nghiên cứu là bộ dữ liệu giao dịch bán lẻ (Online Retail Dataset), chứa thông tin chi tiết về các giao dịch mua hàng của khách hàng trong một khoảng thời gian nhất định.
- Chi tiết thực hiện:
  - Tải xuống dữ liệu từ Kho lưu trữ UCI: https://archive.ics.uci.edu/static/public/352/data.csv
  - Dữ liệu gốc chứa 8 thuộc tính:
    + InvoiceNo: Mã hóa đơn
    + StockCode: Mã định danh sảnh phẩm
    + Description: Tên sản phẩm
    + InvoiceDate: Thời gian giao dịch
    + CustomerID: Mã khách hàng
    + Quantity: Số lượng sản phẩm
    + UnitPrice: Giá bán đơn vị
    + Country: Quốc gia
  - Doanh thu (DoanhThu) được tính từ: DoanhThu = Quantity × UnitPrice
- Mục tiêu:

  Đảm bảo dữ liệu đầu vào:
  - Có quy mô đủ lớn để phân tích thống kê;
  - Đại diện cho hành vi mua sắm đa dạng của khách hàng;
  - Phù hợp để thực hiện các phân tích doanh thu và phân khúc khách hàng.

Bước 2: Tiền xử lý dữ liệu

- Mô tả: Làm sạch và chuẩn hóa dữ liệu từ tệp gốc, xử lý các ngoại lệ và chuẩn bị dữ liệu cho phân tích.
- Chi tiết thực hiện:

  - Xử lý giá trị thiếu: Các giao dịch không có CustomerID được gán nhãn GUEST để phân biệt khách vãng lai;
  - Xử lý dữ liệu bất thường: Kiểm tra doanh thu âm và doanh thu bằng 0;
  - Phát hiện và xử lý outlier
  - Chuẩn hóa dữ liệu thời gian: Trích xuất các đặc trưng như ngày, giờ giao dịch để phục vụ phân tích hành vi theo thời gian.

- Mục tiêu:

Đảm bảo dữ liệu:

- Không bị nhiễu bởi các giá trị cực đoan;
- Phản ánh đúng hành vi mua sắm thực tế;
- Sẵn sàng cho phân tích thống kê và phân khúc khách hàng.

Bước 3: Phân tích thống kê mô tả

- Mô tả: Tóm tắt các đặc tính cơ bản của mô tả thống kê thông tin dữ liệu, trực quan hóa và phân tích tương quan.
- Chi tiết thực hiện:

  - **Thống kê mô tả doanh thu**:
    + Doanh thu trung bình;
    + Trung vị (Median);
    + Phân bố lệch phải mạnh;
  - **Phân tích phân phối doanh thu**:
    + Phần lớn giao dịch nằm khoảng nào?
    + Phần trăm giá trị giao dịch;

Trực quan hóa dữ liệu:

+ Histogram, boxplot và log-scale histogram để quan sát phân bố;

+ ECDF (Empirical Cumulative Distribution Function) để phân tích phân vị;

- **Phân tích theo nhóm doanh thu**:

+ Chia giao dịch thành các nhóm (£0–10, £10–50, £50–100, …);

+ So sánh giữa các nhóm doanh thu;

- Mục tiêu:

- Hiểu rõ đặc điểm phân bố doanh thu;
- Xác định hành vi mua sắm phổ biến;
- Làm cơ sở cho phân tích phân khúc khách hàng.

Bước 4: Phân tích phân khúc khách hàng bằng RFM

- Mô tả: Phân tích RFM (Recency – Frequency – Monetary) được sử dụng để đánh giá giá trị và hành vi của từng khách hàng dựa trên lịch sử giao dịch..

- Chi tiết thực hiện:

- **Chuẩn bị dữ liệu**:

+ Loại bỏ khách hàng GUEST;

+ Tổng hợp dữ liệu theo từng CustomerID;

- **Tính toán chỉ số RFM**:

+ **Recency (R)**: Số ngày kể từ lần mua gần nhất;

+ **Frequency (F)**: Số hóa đơn mua hàng;

+ **Monetary (M)**: Tổng doanh thu khách hàng tạo ra;

- **Chấm điểm RFM**:

+ Chia mỗi chỉ số thành 5 mức (quintiles);

+ Gán điểm từ 1 đến 5;

- **Phân khúc khách hàng**:
  + Champions (khách VIP);
  + Loyal Customers (khách trung thành);
  + Potential Loyalist;
  + New Customers;
  + At Risk;
  + Hibernating.

- Mục tiêu:
  - Nhận diện khách hàng giá trị cao;
  - Phân loại khách hàng theo hành vi mua sắm;
  - Hỗ trợ chiến lược marketing và chăm sóc khách hàng.

Bước 5: Đánh giá và biện luận

- Mô tả: Đánh giá hiệu suất của các bước trên và đưa ra kết luận, đề xuất cải tiến.

- Chi tiết thực hiện:
  - **Tiền xử lý dữ liệu**:
    + Xử lý outlier hiệu quả bằng percentile và log-scale;
    + Dữ liệu không có doanh thu âm;
  - **Phân tích thống kê**:
    + Phân bố doanh thu lệch phải rõ rệt;
    + Phần lớn giao dịch có giá trị thấp;
  - **Phân khúc RFM**:
    + Nhận diện rõ nhóm khách hàng VIP và nhóm có nguy cơ rời bỏ;
    + RFM phù hợp với dữ liệu bán lẻ;
  - **Đề xuất cải tiến**:
    + Áp dụng clustering (K-Means) để so sánh với RFM;

+ Kết hợp thêm biến thời gian và quốc gia;

+ Phân tích hành vi mua lặp lại theo mùa.

- Mục tiêu : Đánh giá toàn diện và đề xuất hướng phát triển cho hệ thống hỗ trợ ra quyết định kinh doanh.

CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU

4.1 Tổng quan

Quá trình thực hiện đã chứng minh tính hiệu quả của quy trình phân tích dữ liệu, từ khâu tiền xử lý đến khai thác thông tin chuyên sâu:

- **Về chất lượng dữ liệu:** Công tác làm sạch dữ liệu đã loại bỏ triệt để các yếu tố nhiễu (đơn hàng ảo, phí vận hành), giữ lại 526,003 giao dịch hợp lệ. Việc xử lý ngoại lai giúp giảm độ lệch chuẩn của dữ liệu xuống hơn 85%, đảm bảo các chỉ số thống kê phản ánh trung thực hành vi mua sắm thực tế của khách hàng thay vì bị chi phối bởi các điểm dữ liệu bất thường.

- **Về giá trị thông tin (Insights):** Các phương pháp phân tích đã chuyển hóa thành công dữ liệu thô thành tri thức kinh doanh cụ thể:

  + **Hiệu quả kinh doanh:** Xác định chính xác tính mùa vụ (đỉnh điểm tháng 11) và "khung giờ vàng" (10h-15h), cung cấp cơ sở khoa học để tối ưu hóa thời điểm chạy chiến dịch Marketing.

  + **Phân khúc khách hàng:** Mô hình RFM đã chứng minh khả năng phân loại vượt trội khi tách biệt rõ ràng các nhóm khách hàng. Đặc biệt, việc phát hiện nhóm At Risk (có nguy cơ rời bỏ nhưng mức chi tiêu trung bình rất cao ~£1,400) là phát hiện quan trọng nhất, chỉ ra "điểm rò rỉ doanh thu" mà doanh nghiệp cần khắc phục ngay lập tức.

4.2 Kết quả thực hiện

4.2.1. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được thực hiện nhằm làm sạch và chuẩn hóa bộ dữ liệu bán lẻ ban đầu, tạo tiền đề cho các bước phân tích thống kê và phân khúc khách hàng.

Bộ dữ liệu sau khi tổng hợp gồm 526.003 dòng giao dịch, với các thuộc tính liên quan đến thời gian, khách hàng và doanh thu.

Bước 1. Đọc và kiểm tra dữ liệu ban đầu:

- Sử dụng các hàm:
  - df.info() để kiểm tra:

+ Số lượng dòng và cột;

+ Kiểu dữ liệu của từng biến;

+ Tình trạng thiếu dữ liệu;

- df.describe() để thống kê mô tả các biến số định lượng.

- Kết quả kiểm tra cho thấy:

- Dữ liệu có quy mô lớn, phù hợp cho phân tích;

- Các biến số như Quantity, UnitPrice, DoanhThu có độ phân tán cao;

- Phân phối doanh thu có xu hướng lệch phải (right-skewed), đặc trưng của dữ liệu bán lẻ.

- **Mục đích**:

Nắm bắt cấu trúc tổng thể của dữ liệu và phát hiện sớm các vấn đề tiềm ẩn trước khi xử lý chi tiết.

Bước 2. Kiểm tra và xử lý giá trị thiếu:

- Sử dụng isnull().sum() để xác định số lượng giá trị thiếu trong từng cột;

- Kết quả cho thấy:

+ Một số giao dịch không có CustomerID;

+ Thay vì loại bỏ các dòng này, các giá trị thiếu ở CustomerID được: Gán nhãn GUEST, đại diện cho khách hàng vãng lai.

Lý do xử lý:

- Việc loại bỏ các dòng thiếu CustomerID có thể làm mất thông tin doanh thu;

- Gán nhãn GUEST giúp:

+ Bảo toàn dữ liệu giao dịch;

+ Phân biệt rõ khách hàng đăng ký và khách vãng lai trong các phân tích sau.

Bước 3. Chuyển đổi biến mục tiêu:

- Biến DoanhThu được tính toán từ:

  DoanhThu = Quantity × UnitPrice

- Các biến thời gian (InvoiceDate) được chuẩn hóa về định dạng datetime;
- Trích xuất thêm các đặc trưng phục vụ phân tích:
  - Ngày giao dịch;
  - Giờ giao dịch;
  - Tháng giao dịch.

Mục đích:

- Đảm bảo các biến được đưa về dạng phù hợp cho phân tích thống kê;
- Hỗ trợ phân tích hành vi mua sắm theo thời gian.

Bước 4. Xử lý giá trị ngoại lai (Outliers):

- Phân tích phân vị (percentiles) của biến DoanhThu cho thấy:
  - 95% giao dịch có doanh thu ≤ £49.92;
  - 99% giao dịch có doanh thu ≤ £122.04;
  - Giá trị lớn nhất là £183.60;
- Dữ liệu có phân phối lệch phải mạnh, xuất hiện một số giao dịch có doanh thu cao hơn mặt bằng chung.

Phương pháp xử lý:

- Không loại bỏ hoàn toàn các giá trị ngoại lai do: Đây có thể là các giao dịch hợp lệ có giá trị cao;
- Áp dụng:
  - Capping theo percentile (95% – 99%) khi trực quan hóa;
  - Log transformation (log-scale) để giảm ảnh hưởng của outliers trong phân tích phân bố.

Lý do:

Giữ lại thông tin quan trọng của các giao dịch lớn;

Tránh làm sai lệch kết quả phân tích thống kê.

Bước 5. Kiểm tra độ lệch:

- Tính toán các chỉ số:
  - Skewness ≈ 4.08;
  - Kurtosis ≈ 21.41;
- Kết quả cho thấy:
  - Phân phối doanh thu lệch phải rất mạnh;
  - Phần lớn giao dịch có giá trị thấp, trong khi một số ít giao dịch có giá trị cao.

Ý nghĩa:

- Đây là đặc trưng phổ biến của dữ liệu bán lẻ;
- Cần sử dụng median, percentiles và log-scale thay vì chỉ dùng mean trong phân tích.

Bước 6. Thống kê mô tả:

- Mô tả: Tóm tắt các đặc tính cơ bản của mô tả thống kê thông tin dữ liệu, trực quan hóa và phân tích tương quan.
- Chi tiết thực hiện:
  - Thống kê mô tả doanh thu:
    + Doanh thu trung bình: ~£15.14;
    + Trung vị (Median): ~£9.90;
    + Phân bố lệch phải mạnh (Skewness ≈ 4.08);

- Phân tích phân phối doanh thu:

  + Phần lớn giao dịch nằm trong khoảng £0 – £50;

  + Giao dịch giá trị cao chiếm tỷ lệ rất nhỏ;

- Trực quan hóa dữ liệu:

  + Histogram, boxplot và log-scale histogram để quan sát phân bố;

  + ECDF (Empirical Cumulative Distribution Function) để phân tích phân vị;

- Phân tích theo nhóm doanh thu:

  + Chia giao dịch thành các nhóm (£0–10, £10–50, £50–100, …);

  + Hơn 90% giao dịch thuộc nhóm doanh thu thấp và trung bình.

- Mục tiêu:

- Hiểu rõ đặc điểm phân bố doanh thu;
- Xác định hành vi mua sắm phổ biến;
- Làm cơ sở cho phân tích phân khúc khách hàng.

Bước 7. Kiểm tra cộng tuyến (Multicollinearity):

- Kiểm tra mối quan hệ giữa các biến số định lượng như:
  - Quantity;
  - UnitPrice;
  - DoanhThu;
- Doanh thu có quan hệ tuyến tính trực tiếp với Quantity và UnitPrice;
- Trong các bước phân tích tiếp theo (RFM), các biến được tổng hợp theo khách hàng, giúp:
  - Giảm ảnh hưởng của đa cộng tuyến;
  - Tập trung vào hành vi mua sắm tổng thể thay vì từng giao dịch đơn lẻ.

- **Mục đích**:
  - Đảm bảo các biến đầu vào không gây nhiễu cho phân tích;
  - Tăng độ tin cậy của kết quả phân khúc khách hàng.

Kết quả

Xem thông tin dataset

```
=== THÔNG TIN DATASET ===
Kích thước dataset: (541909, 8)
Số dòng: 541909
Số cột: 8


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Hình 3 Thông tin dataset

Xem tỉ lệ giá trị thiếu

```
Tỉ lệ missing value (%):
Description     0.27
CustomerID     24.93
dtype: float64
```

Hình 4 Tỉ lệ giá trị thiếu

Điền giá trị thiếu, xóa cột CustomerID

```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   InvoiceNo          541909 non-null  object
 1   StockCode          541909 non-null  object
 2   Description        541909 non-null  object
 3   Quantity           541909 non-null  int64
 4   InvoiceDate        541909 non-null  datetime64[ns]
 5   UnitPrice          541909 non-null  float64
 6   Country            541909 non-null  object
 7   IsCustomerMissing  541909 non-null  bool
 8   CustomerID_Filled  541909 non-null  object
dtypes: bool(1), datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 33.6+ MB
```

Hình 5 Điền giá trị thiếu và xóa cột CustomerID

Chuyển đổi "InvoiceDate" sang định dạng Datetime

	Quantity	InvoiceDate	UnitPrice
count	541909.000000	541909	541909.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114
min	-80995.000000	2010-12-01 08:26:00	-11062.060000
25%	1.000000	2011-03-28 11:34:00	1.250000
50%	3.000000	2011-07-19 17:17:00	2.080000
75%	10.000000	2011-10-19 11:27:00	4.130000
max	80995.000000	2011-12-09 12:50:00	38970.000000
std	218.081158	NaN	96.759853

Hình 6 Chuyển đổi InvoiceDate sang Datetime

Xử lý giá trị âm

```
=== XỬ LÝ GIÁ TRỊ ÂM ===
PHÂN TÍCH GIÁ TRỊ ÂM:
- Số giao dịch Quantity âm: 10624
- Số giao dịch UnitPrice âm: 2
- Số giao dịch DoanhThu âm: 9290

CHI TIẾT GIAO DỊCH UNITPRICE ÂM:
       StockCode      Description  Quantity  UnitPrice  DoanhThu
299983         B  Adjust bad debt         1  -11062.06 -11062.06
299984         B  Adjust bad debt         1  -11062.06 -11062.06

SAU KHI XỬ LÝ:
- Dataset gốc: 541909 dòng
- Sau khi loại UnitPrice âm: 541907 dòng
- Dataset phân tích (không có trả hàng): 531283 dòng
```

Hình 7 Xử lý giá trị âm

Phân tích và xử lý outlier

```
KIEM TRA OUTLIERS
==========================================

1. THONG KE CO BAN:
----------------------------------------
Quantity: Max=80,995, 99th=100
   So giao dich > 100: 4,950
   So giao dich > 1000: 116

UnitPrice: Max=$13,541.33, 99th=$16.98
   So giao dich > $16.98: 5,075
   So giao dich > $1000: 54

DoanhThu: Max=$168,469.60, 99th=$183.60
   So giao dich > $183.60: 5,279
   So giao dich > $5000: 9

2. IQR OUTLIERS DETECTION:
----------------------------------------
Quantity: 56,635 outliers (10.7%)
UnitPrice: 37,999 outliers (7.2%)
DoanhThu: 42,651 outliers (8.0%)
```

```
3. TOP 3 OUTLIERS CAO NHAT:
----------------------------------------
Quantity outliers:
  1. Invoice 581483: 80,995 cai
  2. Invoice 541431: 74,215 cai
  3. Invoice 578841: 12,540 cai

UnitPrice outliers:
  1. AMAZONFEE - AMAZON FEE...: $13,541.33
  2. B - Adjust bad debt...: $11,062.06
  3. POST - POSTAGE...: $8,142.75

DoanhThu outliers:
  1. Invoice 581483: $168,469.60 (80,995 x $2.08)
  2. Invoice 541431: $77,183.60 (74,215 x $1.04)
  3. Invoice 556444: $38,970.00 (60 x $649.50)
```

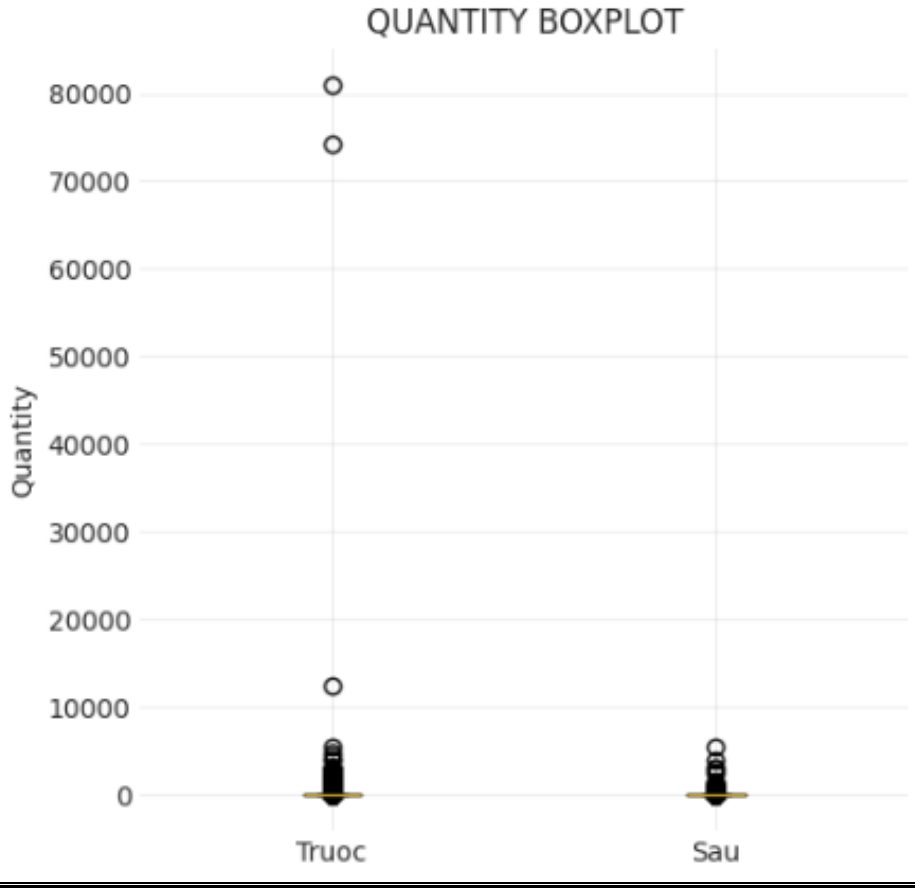
Hình 8 Kiểm tra và phân tích outlier

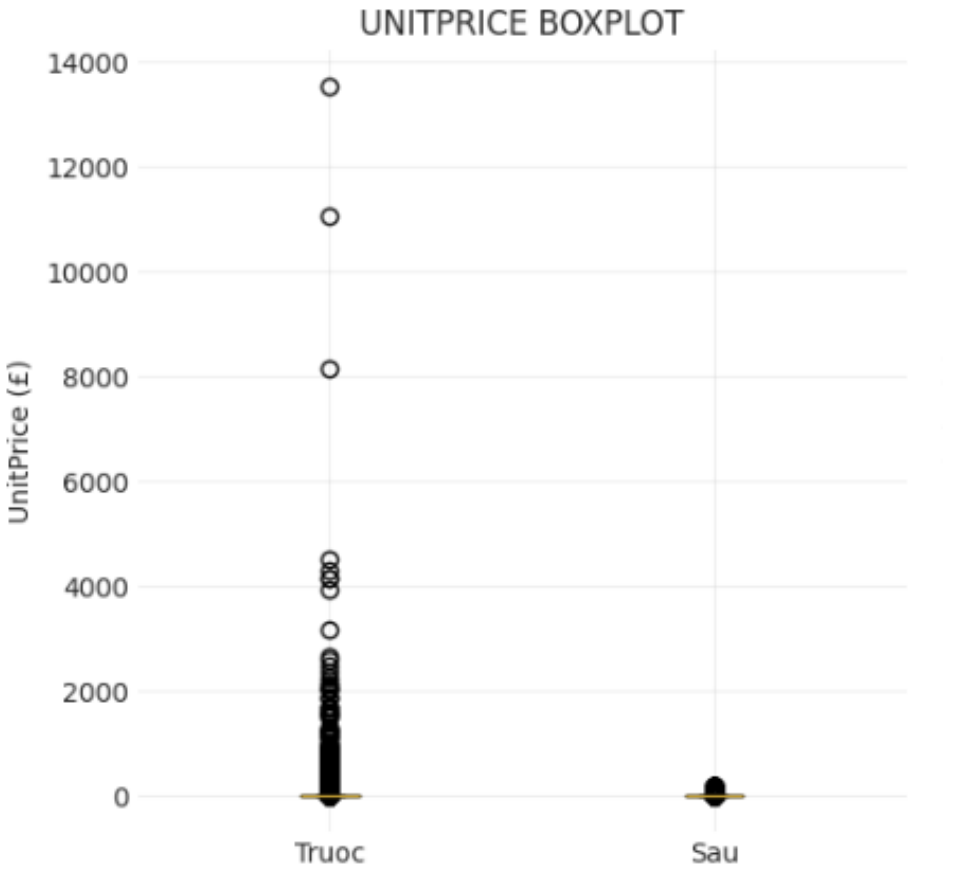
Xử lý Outlier và So sánh trước sau

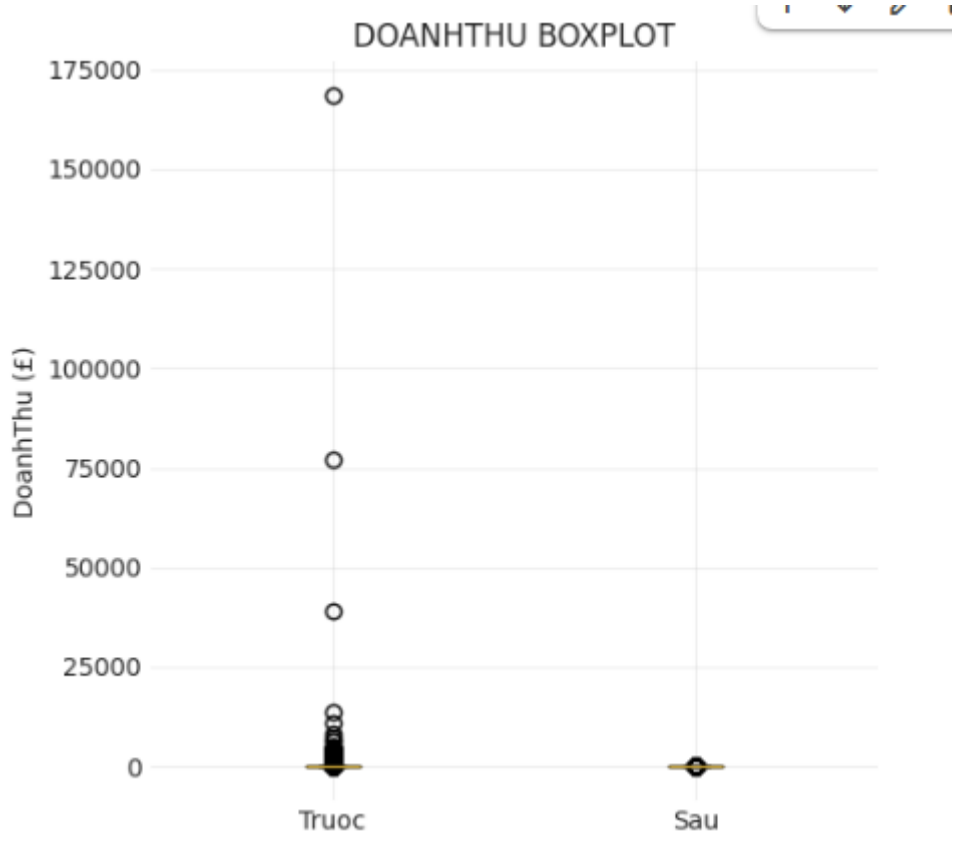
```
--------------------------------------------------
XU LY OUTLIERS
==================================================
Du lieu ban dau: 531,283 giao dich

1. Xu ly Quantity outliers:
   - 3 giao dich Quantity > 10,000

2. Xu ly UnitPrice outliers:
   - 3 giao dich UnitPrice > $5,000

3. Xu ly DoanhThu outliers (method: percentile):
   - 5,279 giao dich DoanhThu > $183.60 (99th)
```

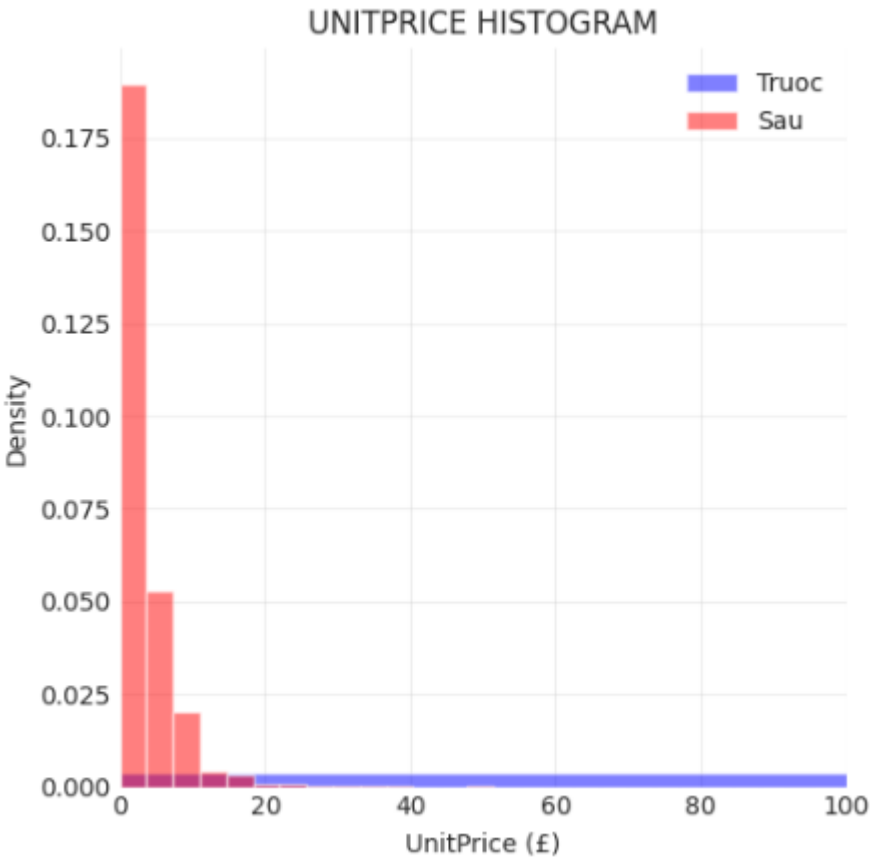
```
4. Ket qua xu ly:
----------------------------------------
Tong outliers: 5,280
Ty le outliers: 1.0%
Dataset co outliers: 531,283 giao dich
Dataset khong outliers: 526,003 giao dich

5. So sanh truoc/sau:
----------------------------------------
Truoc xu ly:
  Quantity: max=80,995, mean=10.7
  UnitPrice: max=$13,541.33, mean=$3.90

Sau xu ly (khong outliers):
  Quantity: max=5,568, mean=8.6
  UnitPrice: max=$183.55, mean=$3.36
```

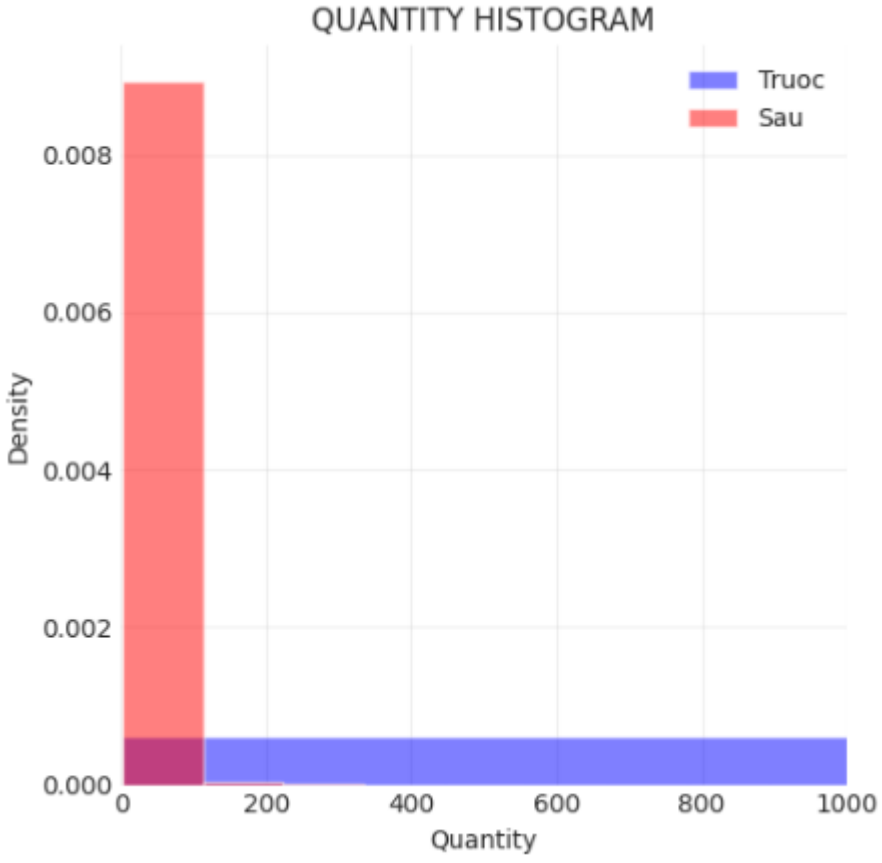
Hình 9 Kết quả và so sánh trước, sau khi xử lý outlier

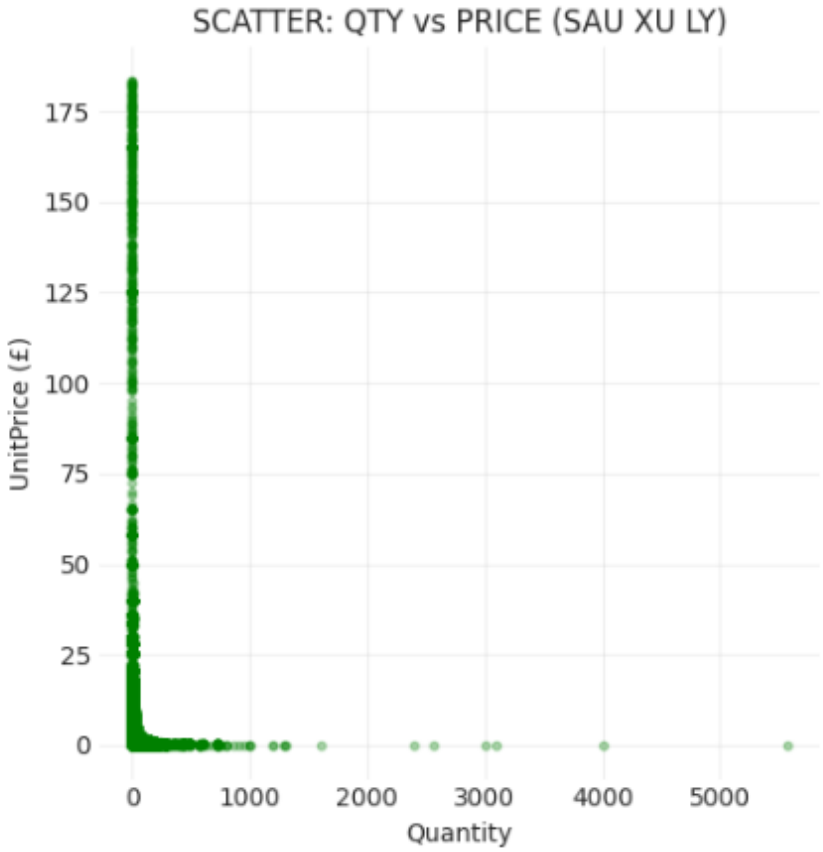



UNITPRICE BOXPLOT
UnitPrice (£)
14000
12000
10000
8000
6000
4000
2000
0
Truoc
Sau


DOANHTHU BOXPLOT
DoanhThu (£)
175000
150000
125000
100000
75000
50000
25000
0
Truoc
Sau


UNITPRICE HISTOGRAM
Truoc
Sau
Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
20
40
60
80
100
UnitPrice (£)


QUANTITY HISTOGRAM
Truoc
Sau
Density
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
1000
Quantity



Hình 10 Mô hình so sánh trước, sau khi xử lý outlier

```
THAY DOI CHINH:
----------------------------------------
Do lech chuan Quantity giam: 85.4%
Do lech chuan UnitPrice giam: 85.7%
```

Hình 11 Kiểm tra tính chuẩn

Thông tin dữ liệu sau khi làm sạch:

Số dòng: 526,003 dòng. (Từ 541,909 dòng ban đầu, đã loại bỏ các giao dịch hủy, nợ xấu và ngoại lai cực đoan).

Số cột:

Bao gồm 8 biến gốc và 6 biến tạo mới phục vụ phân tích (IsCustomerMissing, CustomerID_Filled, DoanhThu, Nam, Thang, Ngay, Tuan).

Kiểu dữ liệu:

- Các cột số (Quantity, UnitPrice, DoanhThu) có kiểu int64 hoặc float64.
- Cột thời gian InvoiceDate đã được chuyển đổi chuẩn sang định dạng datetime.

- Các cột phân loại (Country, Description) đã được chuẩn hóa.

Đánh giá hiệu suất (của quá trình tiền xử lý):

Xử lý giá trị thiếu:

- Vấn đề lớn nhất là cột CustomerID (thiếu ~25%) đã được xử lý bằng cách gán nhãn "GUEST".
- Cột Description được điền giá trị mặc định "Unknown".

Hiệu quả:

Không làm mất mát dữ liệu doanh thu tổng thể, đồng thời vẫn phân biệt được khách hàng định danh và khách vãng lai.

Xử lý ngoại lai (outliers):

Các giá trị cực đoan (như phí Amazon £13,541 hay đơn hàng sỉ 80,000 cái) đã được xử lý bằng phương pháp lọc ngưỡng (Thresholding) và Percentile Capping.

Hiệu quả:

Độ lệch chuẩn (Std) của Quantity giảm 85.4% và UnitPrice giảm 85.7%, giúp biểu đồ phân phối trở nên rõ ràng và tin cậy hơn.

Kiểm tra phân phối và độ lệch:

Xác định rõ đặc điểm phân phối lệch phải (Right-skewed) nặng của dữ liệu bán lẻ. Đã áp dụng thang đo Logarit (Log-scale) khi trực quan hóa để khắc phục vấn đề này.

Biện luận:

- Kết quả tiền xử lý đã đạt được mục tiêu làm sạch và chuẩn bị bộ dữ liệu Online Retail một cách hiệu quả.
- Bộ dữ liệu sạch với 526,003 dòng đã loại bỏ hoàn toàn các yếu tố gây nhiễu (đơn hàng ảo, phí vận hành), đảm bảo tính chính xác cho các bước phân tích thống kê doanh thu và phân khúc khách hàng (RFM) tiếp theo.

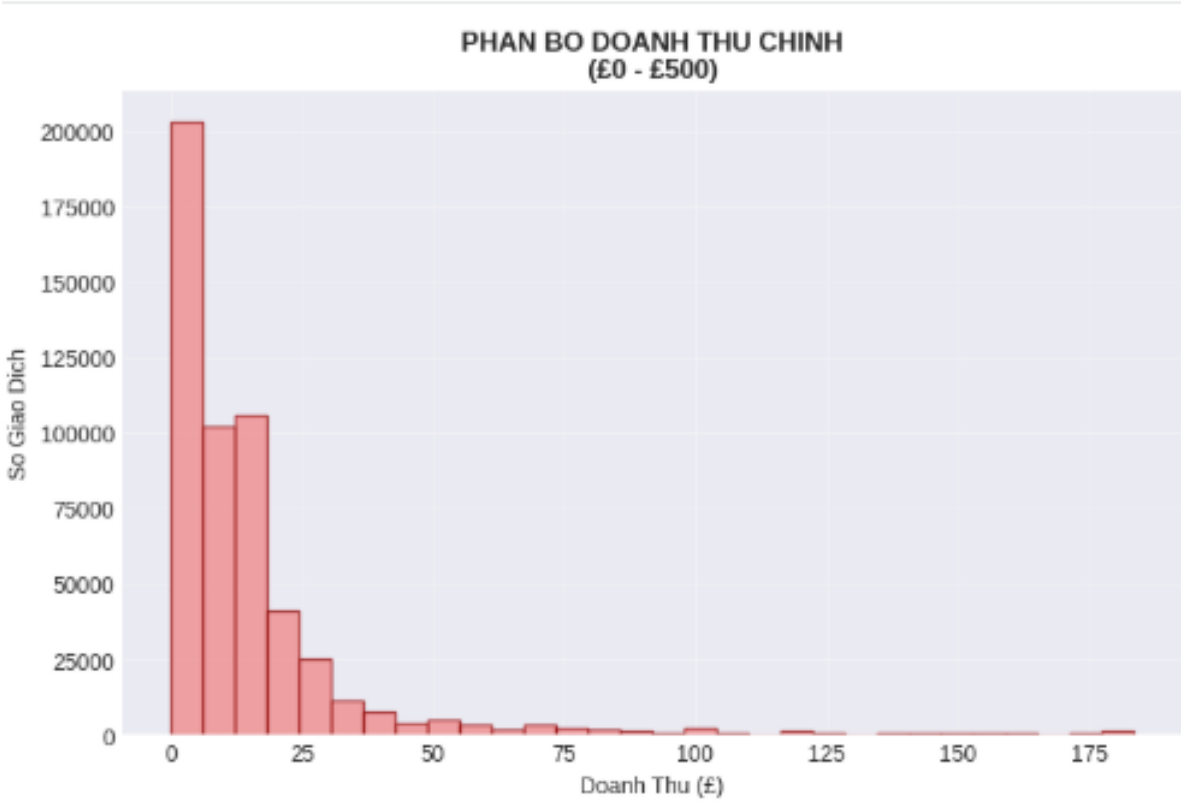
4.2.2. Thống kê mô tả và phân tích xu hướng

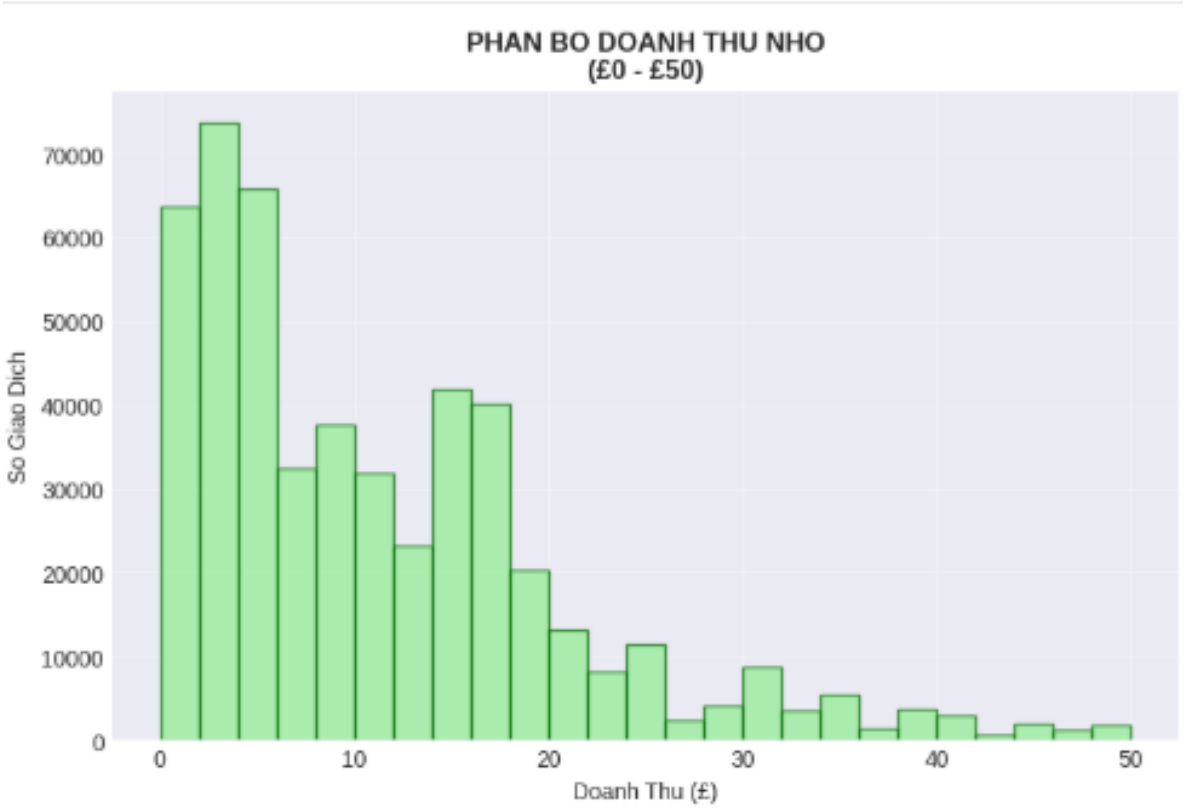
4.2.2.1. Thống kê mô tả về giá trị đơn hàng:

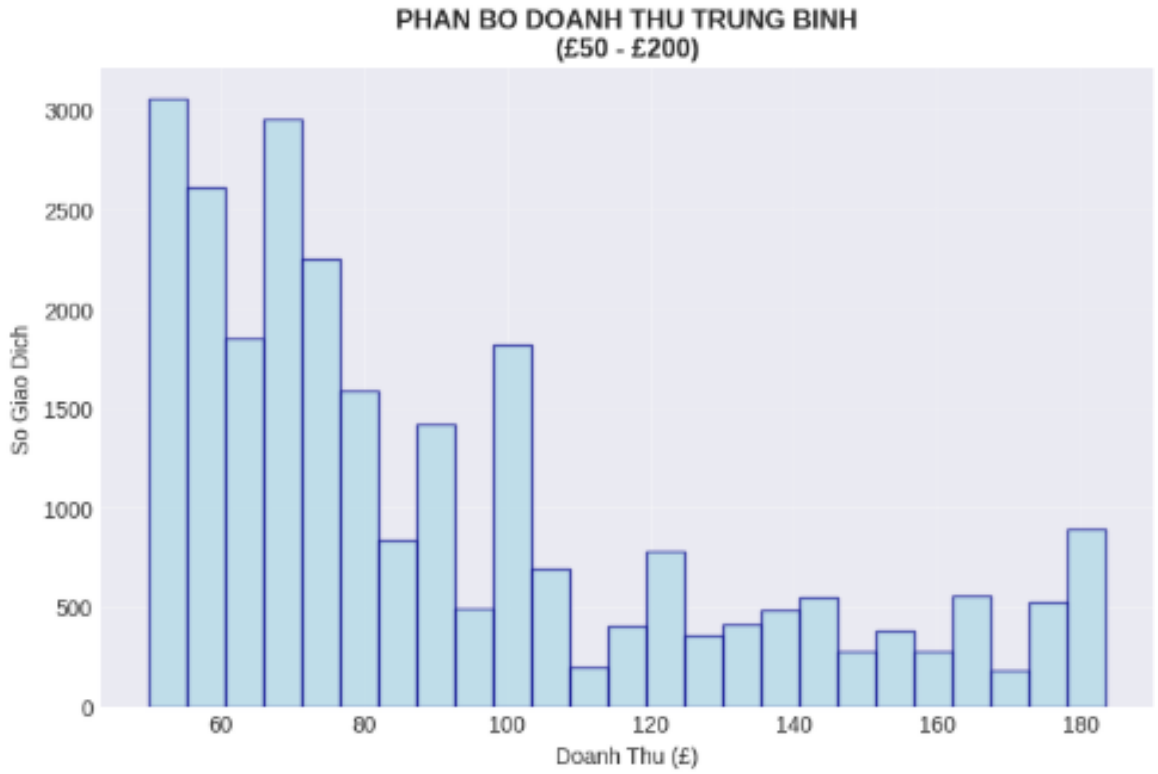
Sau khi tiền xử lý, dữ liệu của bộ Online Retail cho thấy sự phân phối lệch phải (Right-skewed) rất mạnh. Dựa vào biểu đồ phân tích chi tiết phân bố doanh thu cung cấp cái nhìn tổng quan về dữ liệu:

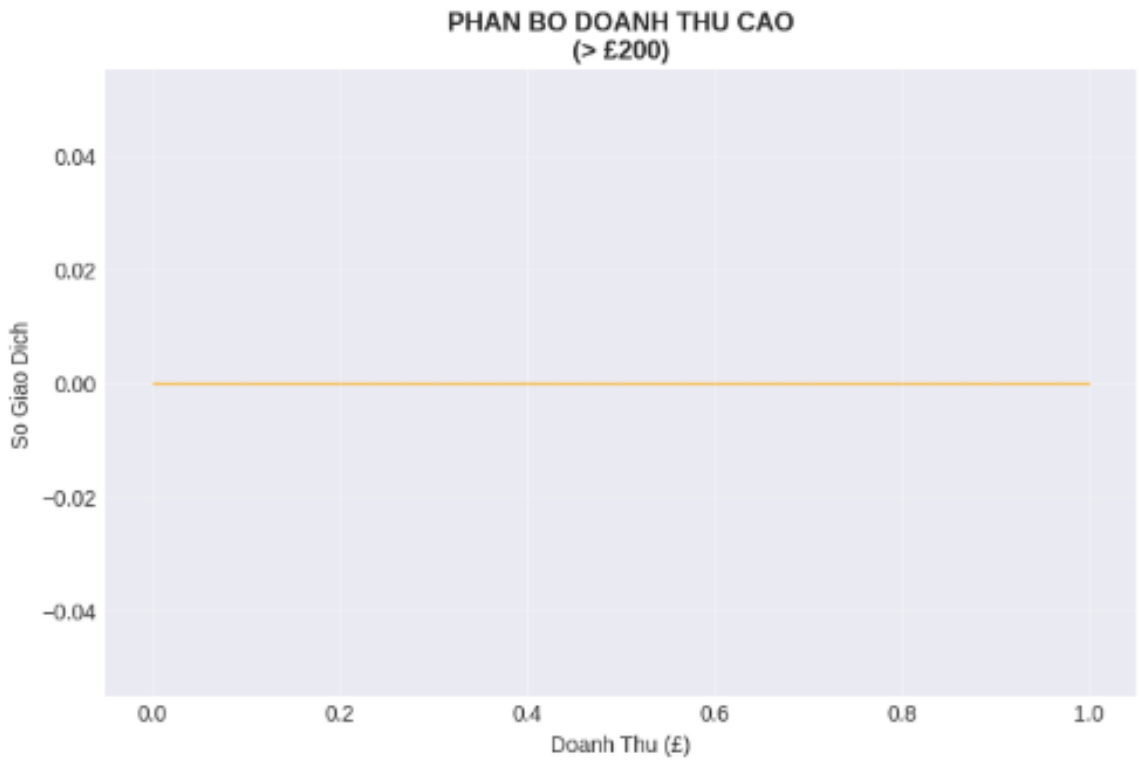
- Phân khúc phổ thông (£0 - £50): Đây là phân khúc chiếm tỷ trọng áp đảo với hơn 95% tổng số giao dịch. Đỉnh của phân phối nằm ở khoảng giá trị £10 - £15. Điều này phản ánh hành vi mua sắm đặc trưng của khách hàng là mua các món quà lưu niệm nhỏ, giá rẻ hoặc mua lẻ tẻ.
- Phân khúc trung bình (£50 - £200): Số lượng giao dịch giảm dần. Đây thường là các đơn hàng gom chung (group buying) hoặc khách hàng mua sỉ quy mô nhỏ.
- Phân khúc cao (> £200): Xuất hiện rất hiếm hoi (dưới 0.1%). Tuy nhiên, nhờ bước xử lý ngoại lai (Outlier Handling) ở mục 3.2.1, các giá trị này là doanh thu thực tế (hợp lệ), không phải do lỗi dữ liệu.

Kết quả




PHAN BO DOANH THU NHO
(£0 - £50)
So Giao Dich
70000
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
Doanh Thu (£)


PHAN BO DOANH THU TRUNG BINH
(£50 - £200)
So Giao Dich
3000
2500
2000
1500
1000
500
0
60
80
100
120
140
160
180
Doanh Thu (£)

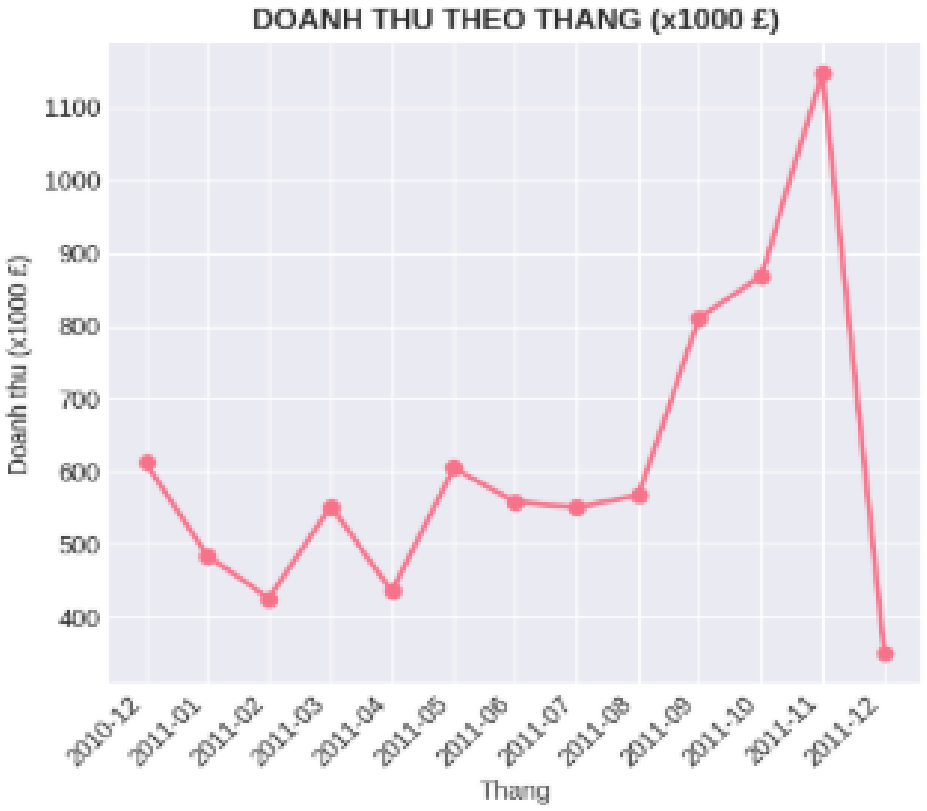


Hình 12 Phân bố doanh thu

4.2.2.2. Phân tích xu hướng kinh doanh:

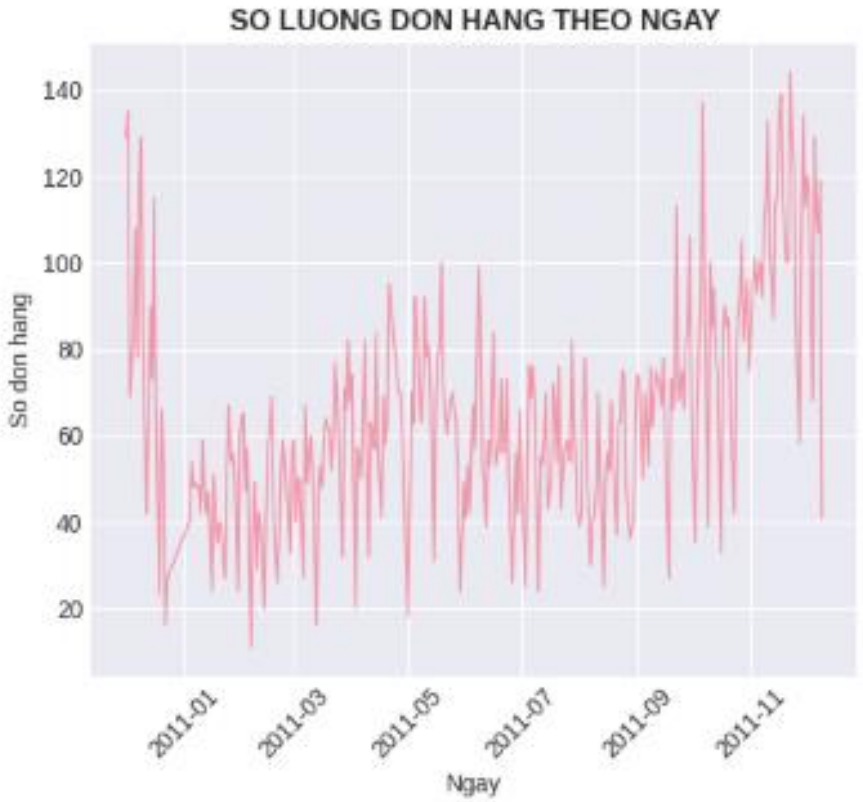
Việc phân tích dữ liệu chuỗi thời gian (Time-series) giúp nhận diện rõ tính mùa vụ và thói quen mua sắm:

- **Xu hướng theo tháng:**
  - Doanh thu duy trì ổn định ở mức trung bình trong 8 tháng đầu năm.
  - Bắt đầu tăng trưởng mạnh từ tháng 9 và đạt đỉnh điểm vào Tháng 11/2011 (doanh thu vượt mốc £1.15 triệu).
  - Nhận định*:*
    + Đây là hiệu ứng mua sắm chuẩn bị cho kỳ nghỉ lễ lớn nhất trong năm (Giáng sinh và Năm mới).
    + Sau tháng 11, doanh thu tháng 12 có xu hướng giảm nhẹ do dữ liệu chưa cập nhật hết tháng hoặc do sức mua đã bão hòa.



Hình 13 Doanh thu theo tháng

- **Xu hướng theo ngày:**

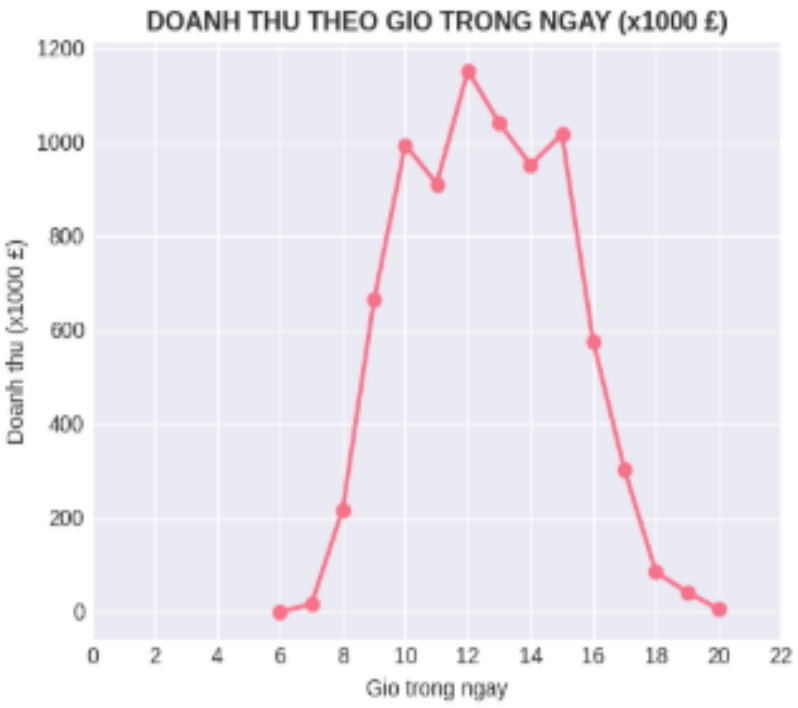


Hình 14 Số lượng đơn theo ngày

- Quan sát biểu đồ "Số lượng đơn hàng theo ngày", ta thấy xu hướng tăng trưởng rõ rệt về tần suất mua sắm vào giai đoạn cuối năm (từ tháng 9 đến đầu tháng 12).
- Số lượng đơn hàng trong một ngày có sự biến động lớn, tuy nhiên mật độ các ngày có lượng đơn hàng cao (>100 đơn/ngày) xuất hiện dày đặc vào Tháng 11.
- Nhận định: Điều này khẳng định tính mùa vụ của ngành bán lẻ: Khách hàng bắt đầu mua sắm tích trữ cho lễ hội từ đầu Quý 4.

- **Xu hướng theo giờ:**
  - Khung giờ mua sắm sôi động nhất là từ 10:00 sáng đến 15:00 chiều.
  - Sau 15:00, lượng đơn hàng giảm mạnh.
  - Khuyến nghị:

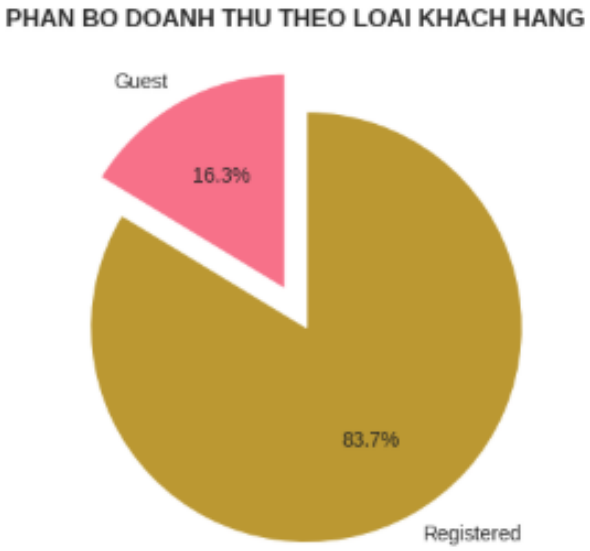
    Doanh nghiệp nên tập trung nhân sự trực chat/chốt đơn và tung ra các chương trình Flash Sale vào khung giờ vàng (11h-13h) để tối ưu hóa tỷ lệ chuyển đổi.



Hình 15 Phân bố Doanh thu theo tháng, ngày, giờ

- **Xu hướng theo loại khách hàng:**

Biểu đồ tròn (Pie chart) cho thấy sự áp đảo của nhóm khách hàng đã đăng ký (Registered - 83.7% doanh thu) so với khách vãng lai (Guest - 16.3%). Điều này chứng minh chiến lược tập trung chăm sóc khách hàng thành viên là hoàn toàn đúng đắn.



Hình 16 Phân bố Doanh thu theo Loại khách hàng

4.2.3. Phân tích phân nhóm khách hàng

Để hiểu sâu hơn về giá trị khách hàng đồ án đã áp dụng mô hình RFM (Recency - Frequency - Monetary) [6], để phân loại 4,290 khách hàng thành các nhóm chiến lược.

Kết quả phân nhóm

- Nhóm Champions:

  - Đặc điểm: Là nhóm khách hàng "vô địch" - mua gần đây nhất, thường xuyên nhất và chi tiêu nhiều nhất.

  - Số liệu: Mặc dù số lượng chỉ khoảng 420 khách, nhưng mức chi tiêu trung bình đạt tới >£6,000/người.

  - Vai trò: Đây là nhóm gánh vác doanh thu chính cho doanh nghiệp (theo nguyên lý Pareto 80/20).

- Nhóm At Risk (Nguy cơ rời bỏ):

  - Đặc điểm: Từng mua hàng rất nhiều trong quá khứ nhưng đã lâu không quay lại.

  - Phát hiện quan trọng: Mức chi tiêu trung bình của nhóm này lên tới ~£1,400/người (cao thứ 3 toàn hệ thống, chỉ sau Champions và Loyal).

  - Cảnh báo: Doanh nghiệp đang có nguy cơ mất đi một lượng doanh thu khổng lồ từ nhóm khách hàng này nếu không có biện pháp can thiệp.

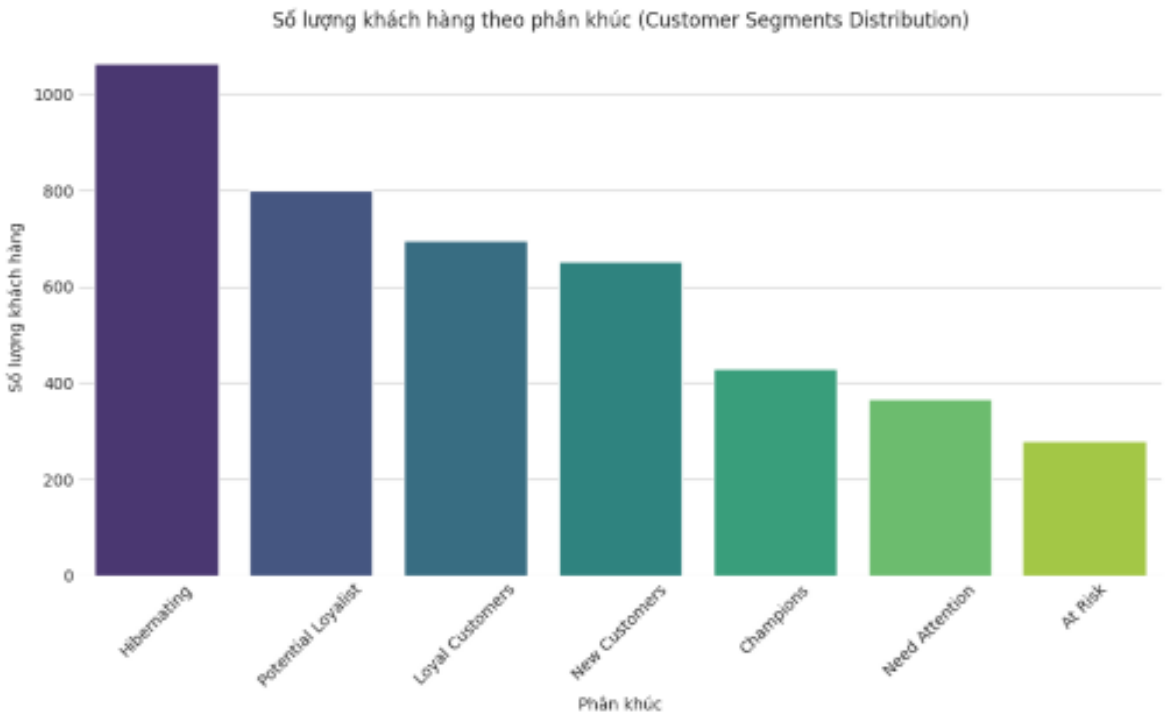
- Nhóm Hibernating:

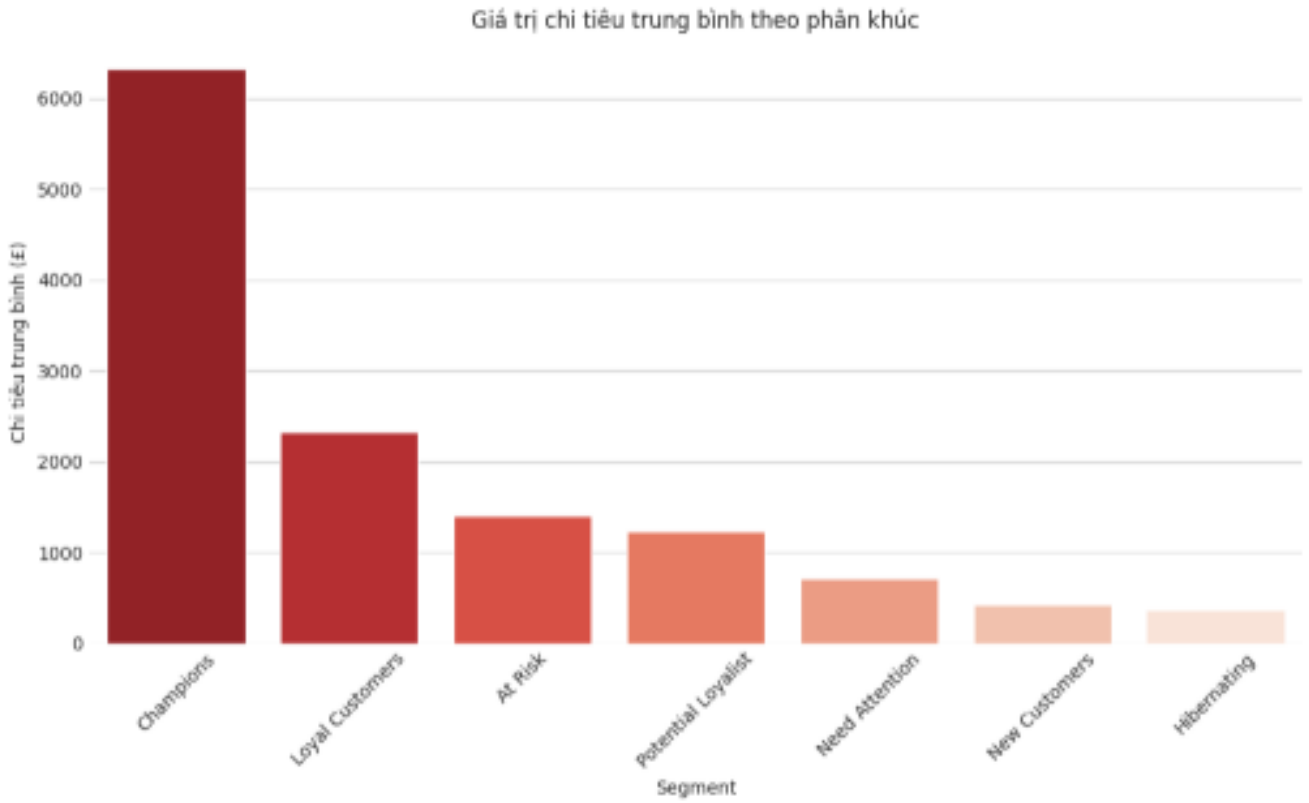
  - Đặc điểm: Chiếm số lượng đông nhất (>1,000 khách) nhưng giá trị chi tiêu trung bình rất thấp (<£400).

  - Đánh giá: Đây thường là khách vãng lai hoặc săn khuyến mãi một lần.

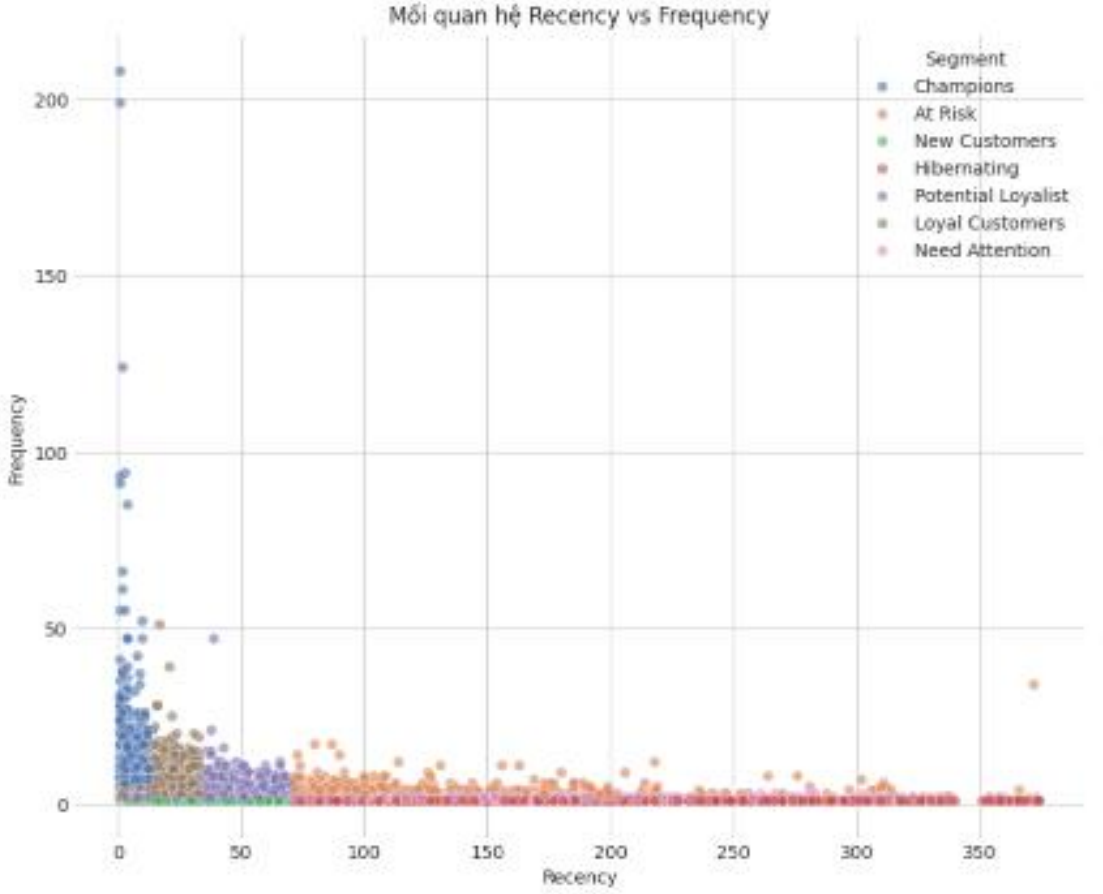
- Nhóm Potential Loyalist (Tiềm năng):

Đặc điểm: Mua hàng gần đây với tần suất khá tốt. Đây là nguồn dự bị để chuyển đổi thành khách hàng trung thành.





Hình 17 Thống kê số lượng và giá trị chi tiêu trung bình theo phân khúc khách hàng



Hình 18 Phân bố hành vi khách hàng dựa trên tần suất và thời gian mua gần nhất

Hình 19 Chạy thử ứng dụng trên Streamlit

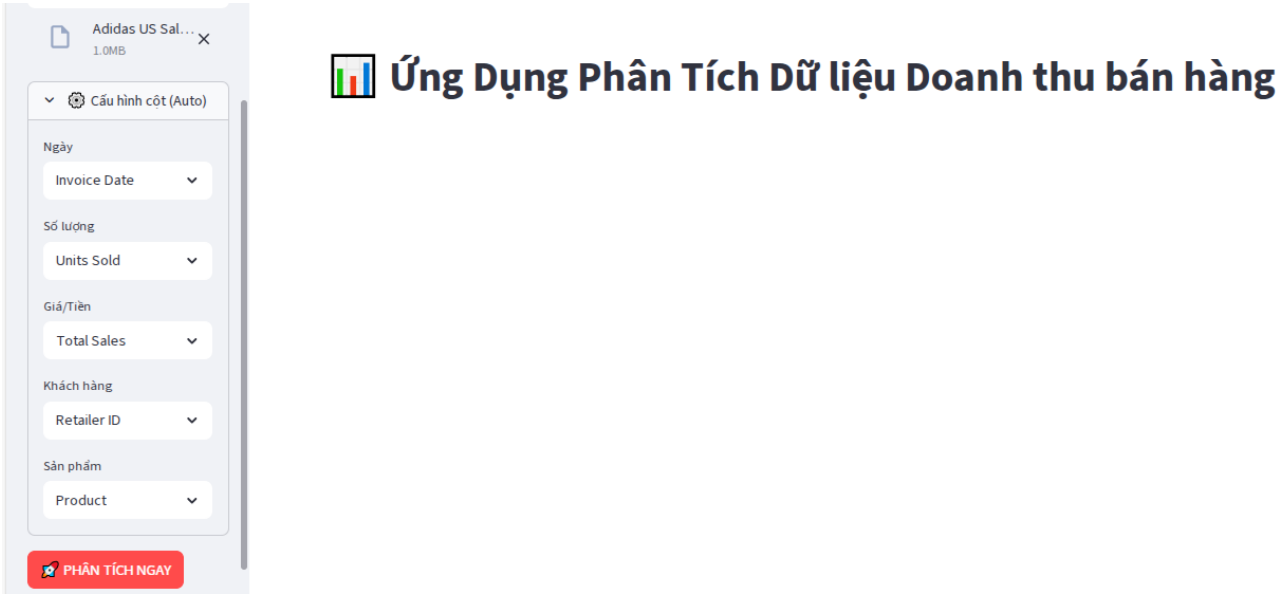
4.2.4. Chạy thử nghiệm lần 1

CSV: Adidas US Sales Datasets

Số cột: 13 (Retailer, Retailer ID, Invoice Date, Region, State, City, Product, Price per Unit, Untis Sold, Total Sales, Operating Profit, Operating Margin, Sales Method)

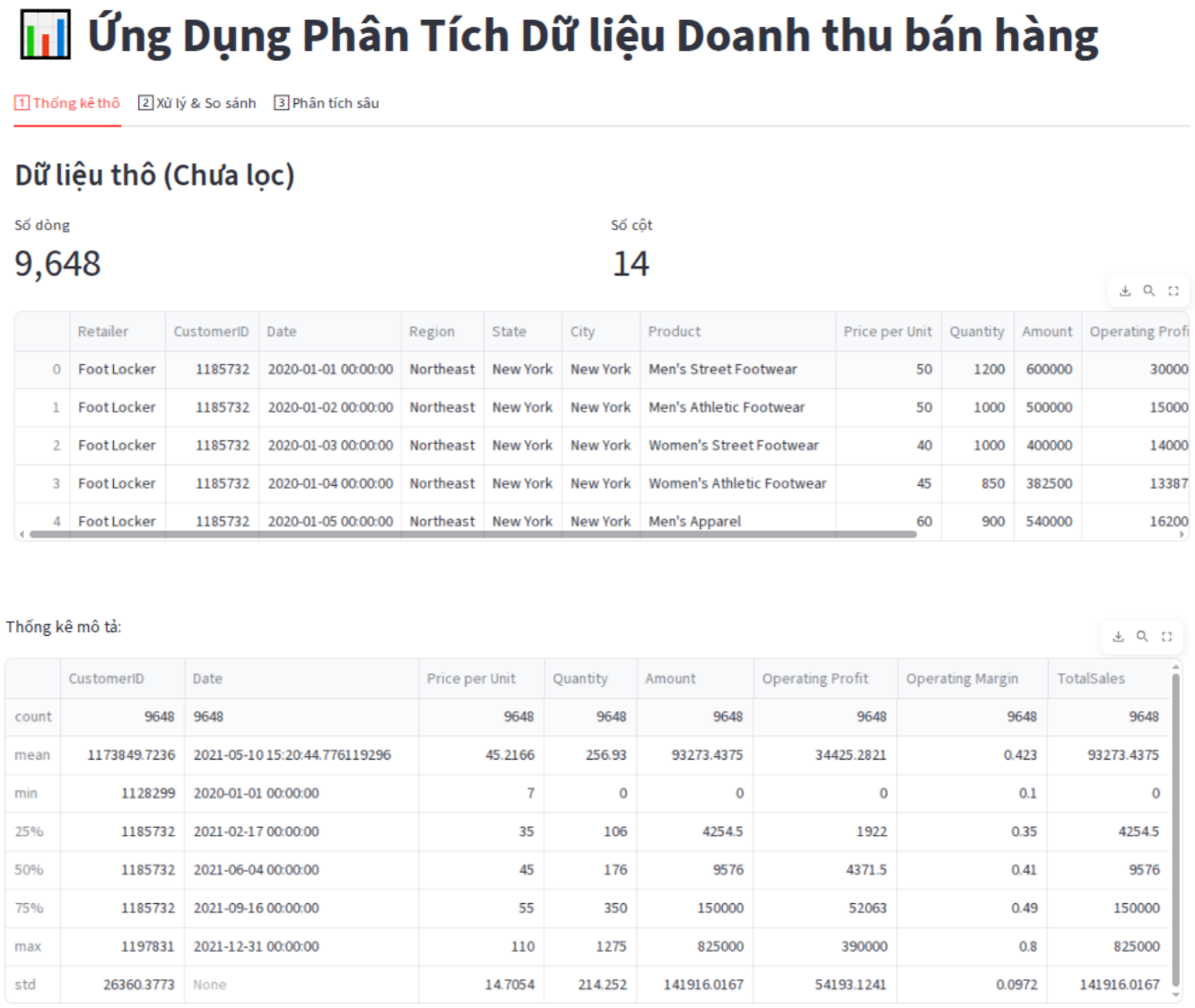
Các cột để kiểm tra:

- Ngày: Invoice Date
- Số lượng: Units Sold
- Giá tiền: Total Sales
- Khách hàng: Retailer ID
- Sản phẩm: Product



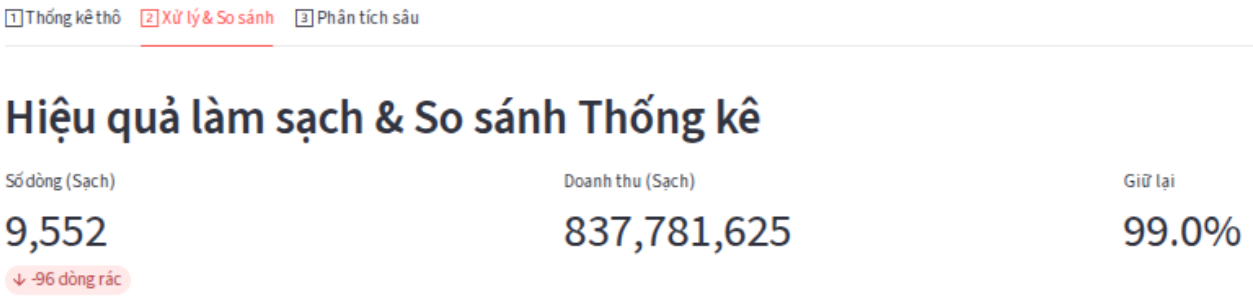
Hình 20 Thử nghiệm lần 1 và cấu hình cột

Kết quả thử nghiệm lần 1:



Hình 21 Thống kê dữ liệu thô của csv Adidas trên ứng dụng

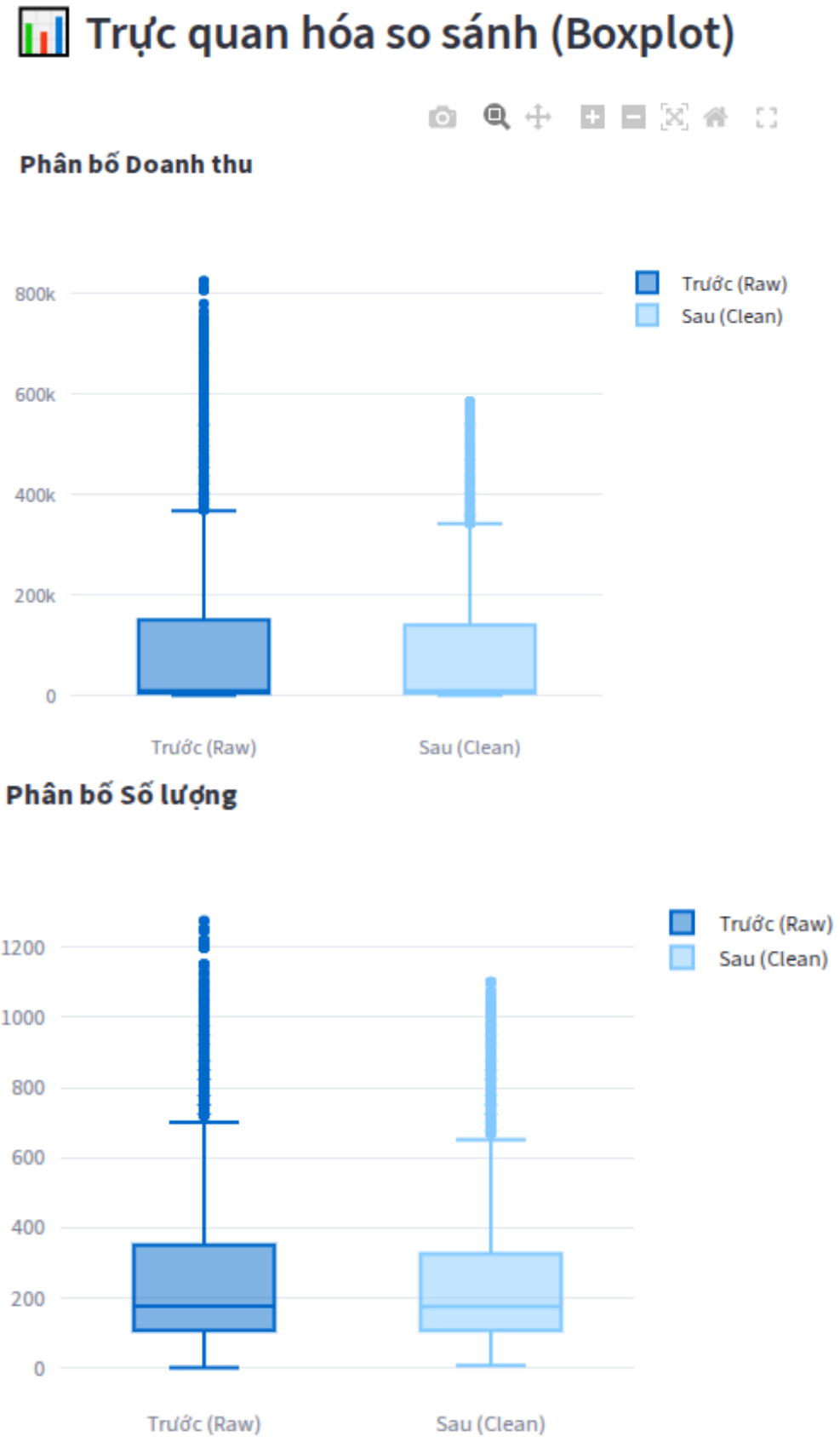
Sau khi tiền xử lý và so sánh:



Bảng so sánh chỉ số thống kê (Trước vs Sau)

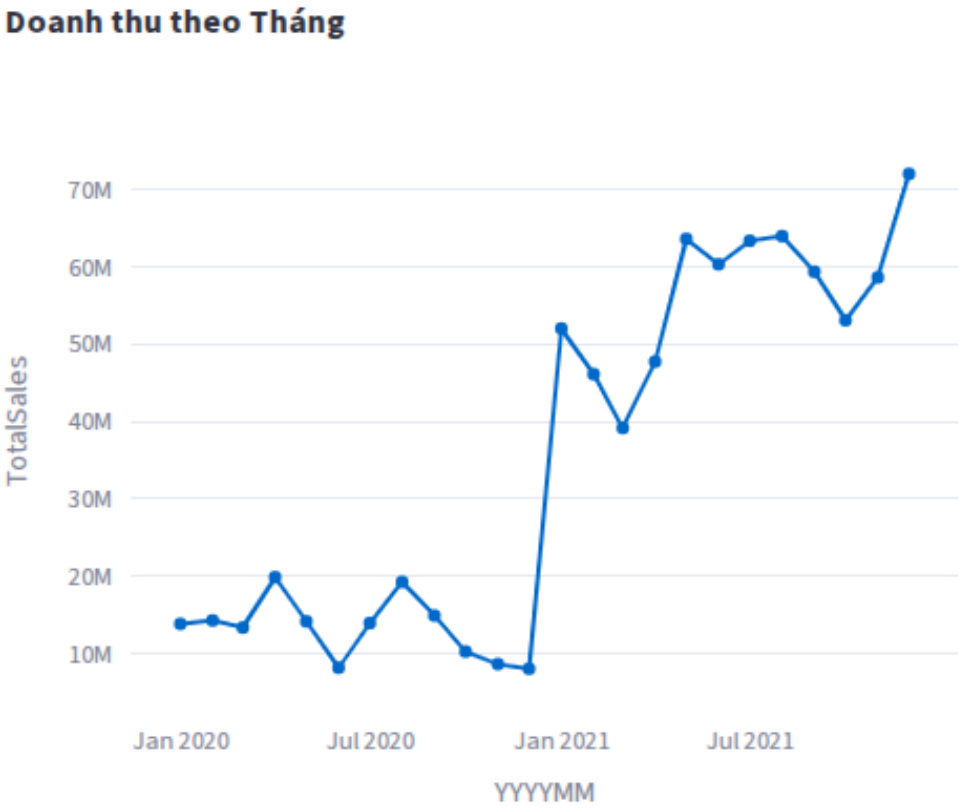
	SL (Trước)	SL (Sau)
count	9,648.00	9,552.00
mean	256.93	250.85
std	214.25	204.96
min	0.00	6.00
25%	106.00	106.00
50%	176.00	175.00
75%	350.00	325.00
max	1,275.00	1,100.00

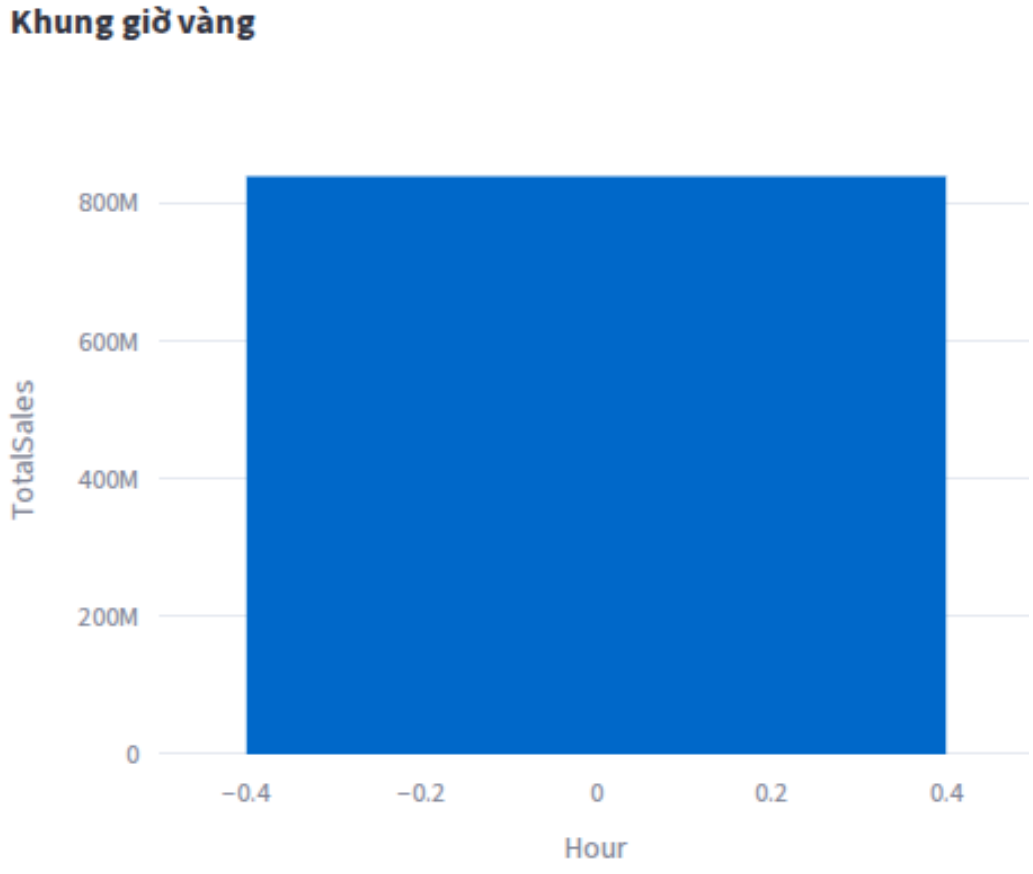
Tiền (Trước)	Tiền (Sau)
9,648.00	9,552.00
93,273.44	87,707.46
141,916.02	130,418.25
0.00	160.00
4,254.50	4,217.50
9,576.00	9,433.00
150,000.00	140,000.00
825,000.00	585,000.00

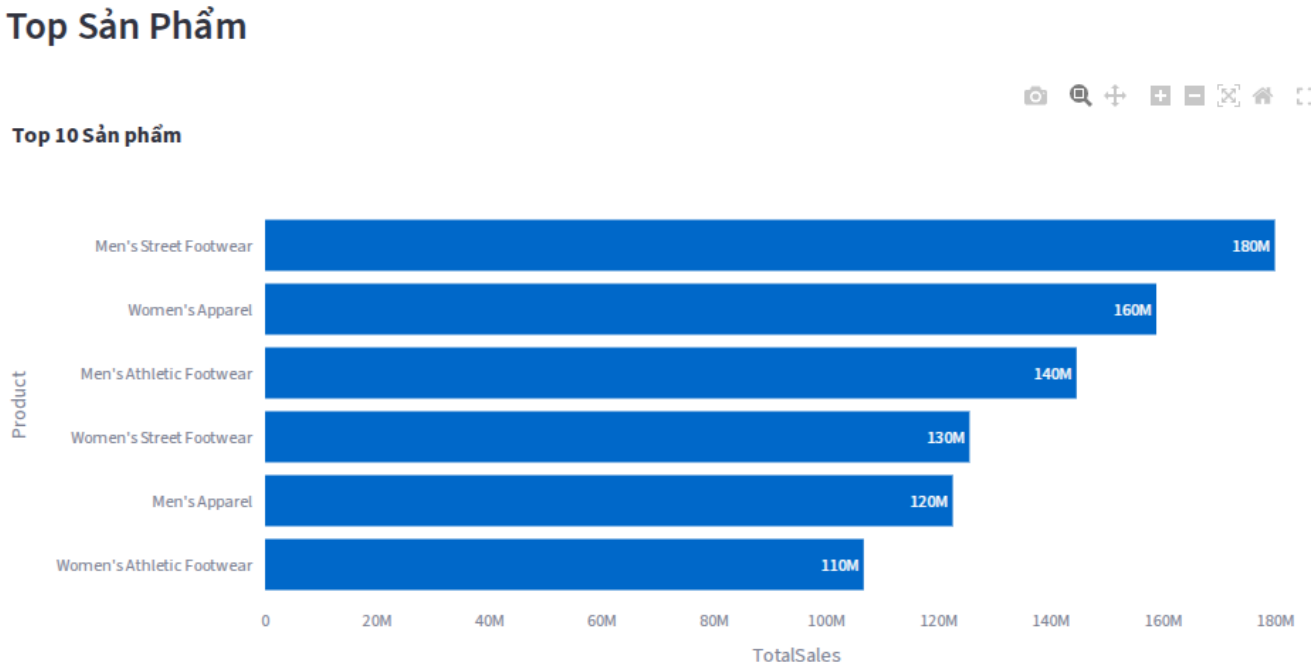


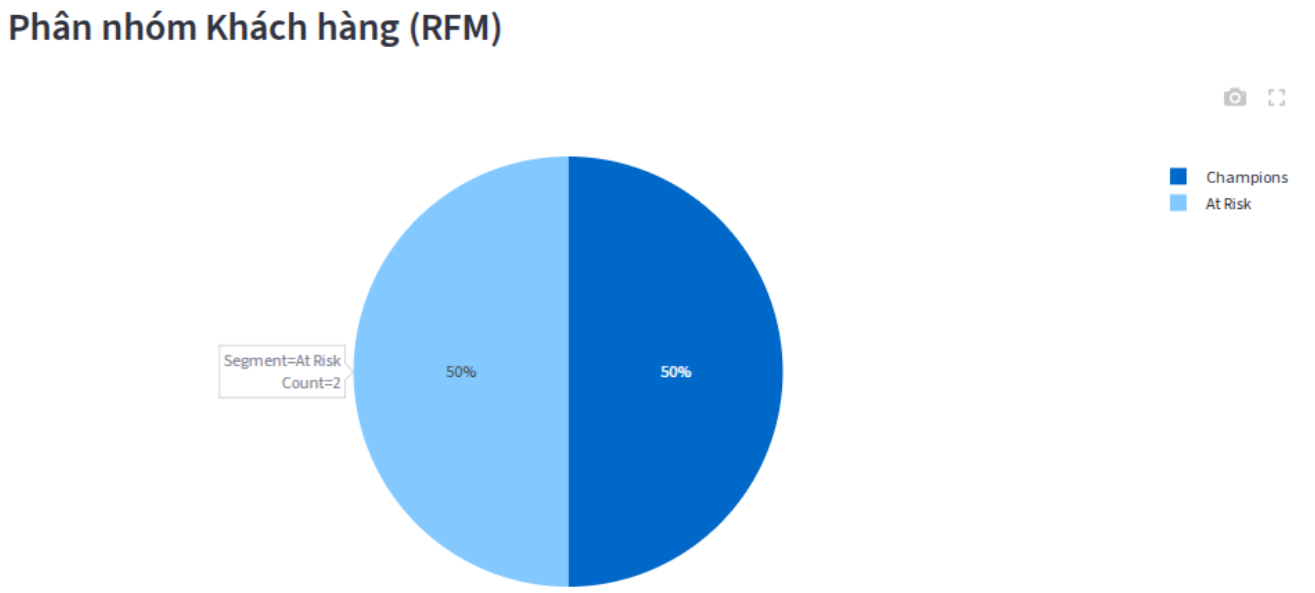
Hình 22: Bảng và mô hình so sánh trước và sau khi xử lý dữ liệu

Phân tích chuyên sâu:









Hình 23 Phân tích chuyên sâu (xu hướng và phân nhóm khách hàng)

Chạy thử nghiệm lần 2

CSV Online Retail

Số cột: 8 (Invoice No, Stock Code, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country)

Các cột để kiểm tra:

- Ngày: Invoice Date
- Số lượng: Quantity
- Giá tiền: UnitPrice
- Khách hàng: CustomerID
- Sản phẩm: Description

Kết quả thử nghiệm lần 2:

Thống kê dữ liệu thô

📊 Ứng Dụng Phân Tích Dữ Liệu Doanh thu bán hàng

1 Thống kê thô | 2 Xử lý & So sánh | 3 Phân tích sâu

Dữ liệu thô (Chưa lọc)

Số dòng: 541,909

Số cột: 9

	InvoiceNo	StockCode	Product	Quantity	Date	Amount	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

Thống kê mô tả (Raw):

	Quantity	Date	Amount	CustomerID	TotalSales
count	541909	541909	541909	406829	541909
mean	9.5522	2011-07-04 13:34:57.156386048	4.6111	15287.6906	17.9878
min	-80995	2010-12-01 08:26:00	-11062.06	12346	-168469.6
25%	1	2011-03-28 11:34:00	1.25	13953	3.4
50%	3	2011-07-19 17:17:00	2.08	15152	9.75
75%	10	2011-10-19 11:27:00	4.13	16791	17.4
max	80995	2011-12-09 12:50:00	38970	18287	168469.6
std	218.0812	None	96.7599	1713.6003	378.8108

Hình 24 Thống kê mô tả CSV Online Retail trên ứng dụng

Tiền xử lý dữ liệu và so sánh:

Hiệu quả làm sạch & So sánh Thống kê

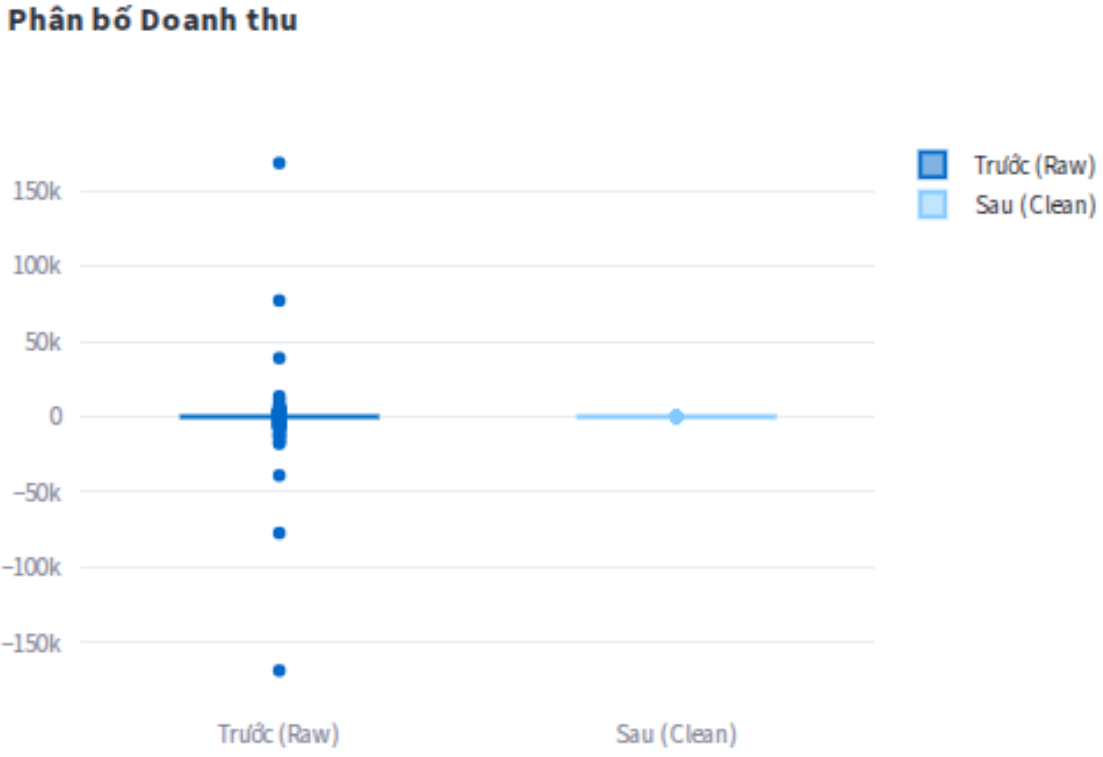
Số dòng (Sạch)	Doanh thu (Sạch)	Giữ lại
524,825	7,963,272	96.8%
↓ -17084 dòng rác		

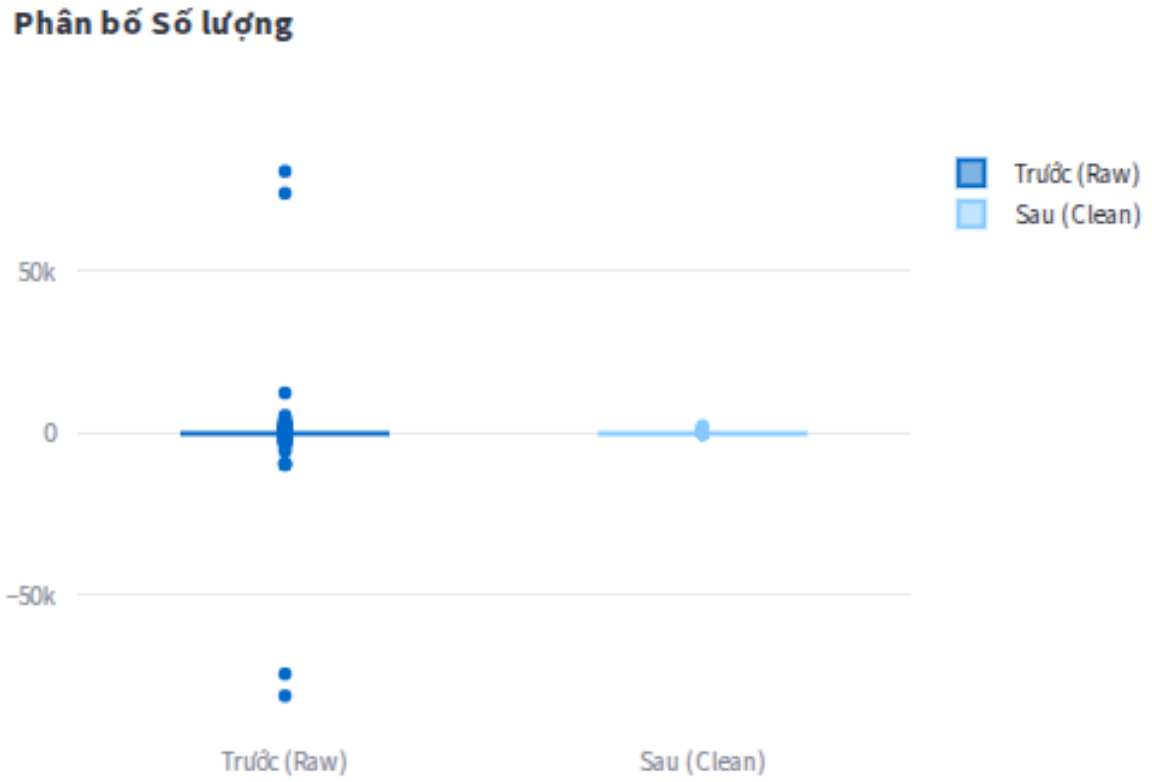
Bảng so sánh chỉ số thống kê (Trước vs Sau

	SL (Trước)	SL (Sau)
count	541,909.00	524,825.00
mean	9.55	8.54
std	218.08	18.83
min	-80,995.00	1.00
25%	1.00	1.00
50%	3.00	3.00
75%	10.00	10.00
max	80,995.00	2,400.00

Tiền (Trước)	Tiền (Sau)
541,909.00	524,825.00
17.99	15.17
378.81	21.31
-168,469.60	0.00
3.40	3.75
9.75	9.90
17.40	17.40
168,469.60	183.60

Trực quan hóa so sánh (Boxplot)

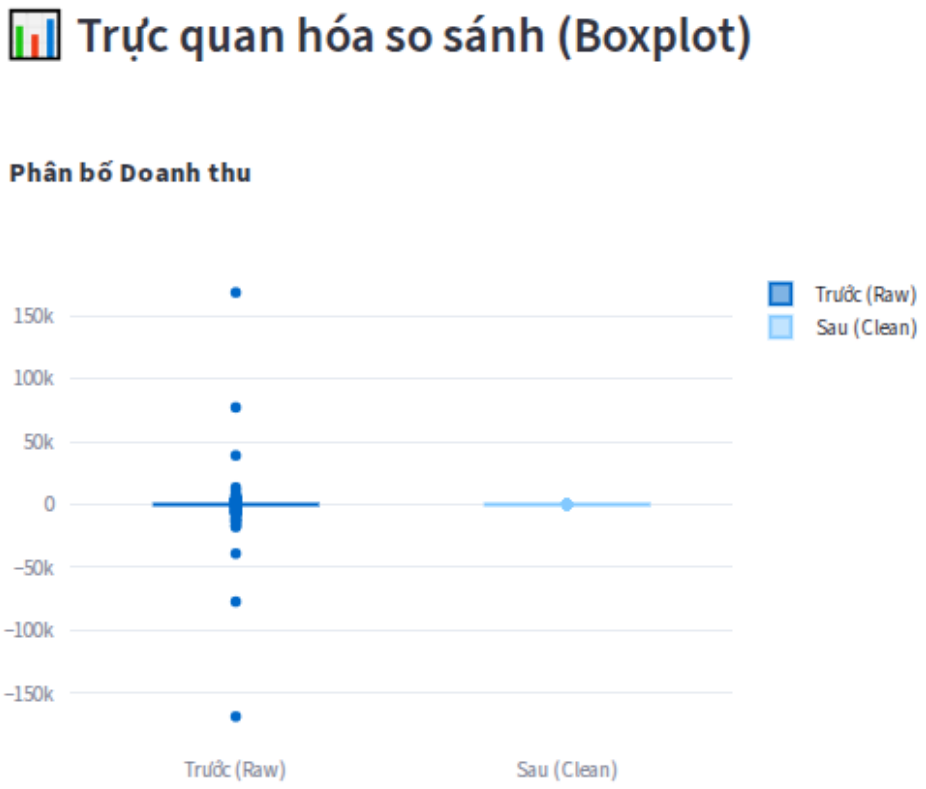


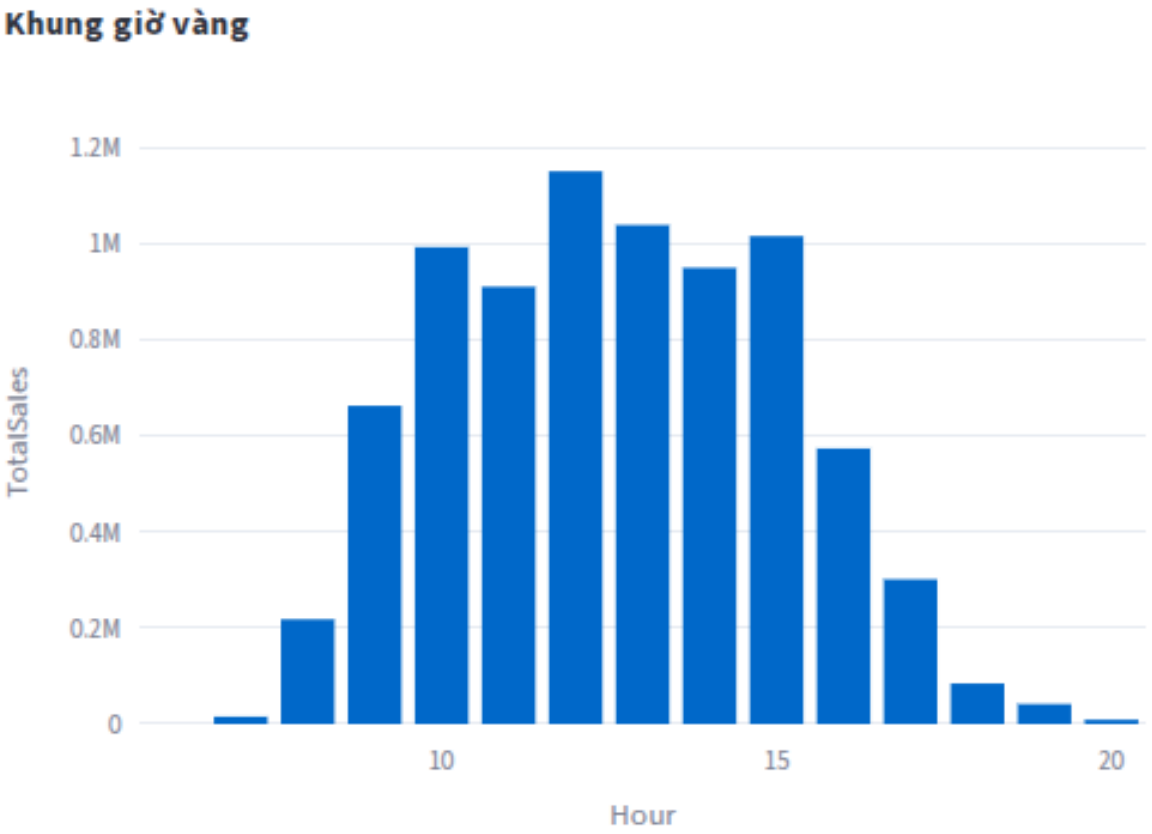


Hình 25 Kết quả và so sánh dữ liệu trước và sau tiền xử lý

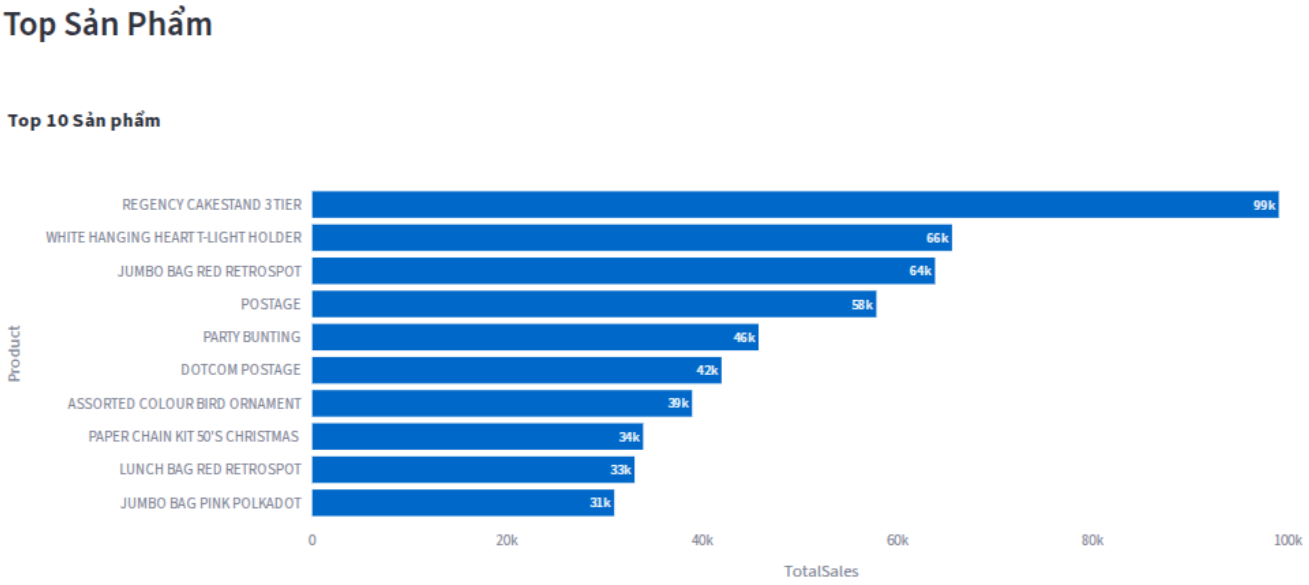
Phân tích chuyên sâu:

- Doanh thu theo thời gian:

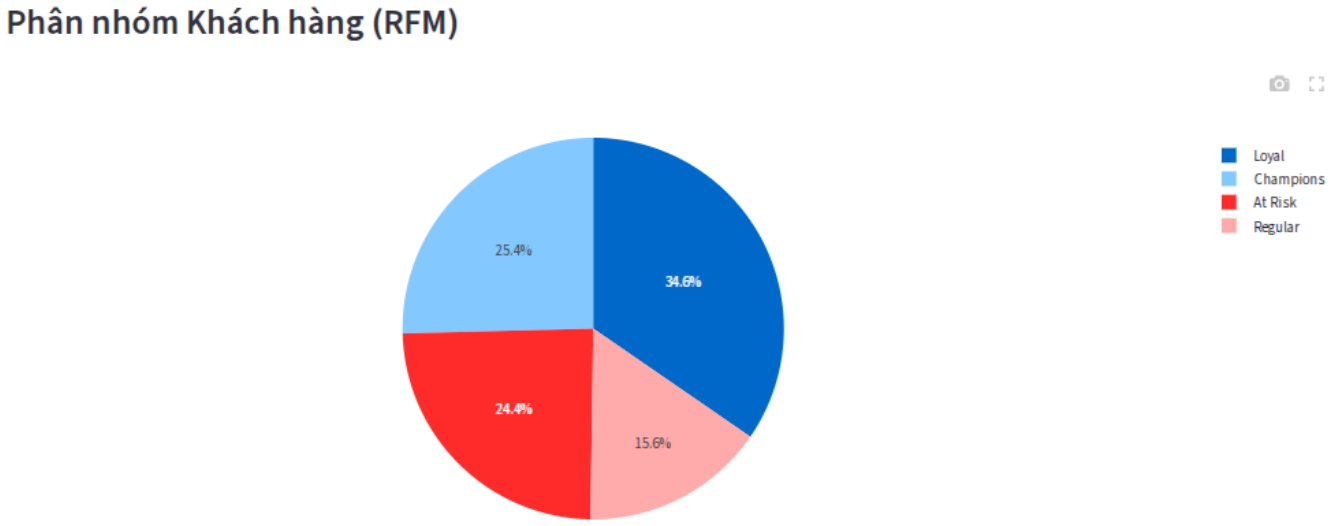




- Top 10 sản phẩm



- Phân nhóm khách hàng



Hình 26 Kết quả Phân tích chuyên sâu thử nghiệm lần 2 trên ứng dụng

Tổng kết:

Qua quá trình thực nghiệm trên hai bộ dữ liệu thực tế khác nhau là Adidas US Sales (Lần 1) và Online Retail (Lần 2) đã rút ra những đánh giá tổng quan sau:

- Khả năng thích ứng và xử lý đa dạng
- So sánh chiều sâu phân tích

Khả năng thích ứng và xử lý đa dạng:

Quy trình phân tích dữ liệu (Data Analytics) được xây dựng đã chứng minh tính linh hoạt và hiệu quả khi áp dụng trên các cấu trúc dữ liệu khác nhau:

- Lần 1 (Adidas US Sales):

  Xử lý tốt dữ liệu báo cáo kinh doanh tổng hợp với nhiều biến định tính (Khu vực, Phương thức bán hàng) và định lượng (Doanh số, Lợi nhuận). Hệ thống đã trích xuất thành công các chỉ số KPI quan trọng về hiệu quả hoạt động.

- Lần 2 (Online Retail):

  Xử lý thành công dữ liệu giao dịch chi tiết (Transaction-level) với khối lượng lớn (>500,000 dòng) và độ nhiễu cao. Đây là thử thách lớn hơn về mặt kỹ thuật làm sạch và xử lý ngoại lai.

So sánh chiều sâu phân tích:

- Ở thử nghiệm 1 (Adidas):
  + Kết quả dừng lại ở mức Phân tích mô tả (Descriptive Analytics).
  + Kết quả trực quan hóa giúp nhìn rõ bức tranh doanh thu và lợi nhuận theo khu vực địa lý và dòng sản phẩm.
- Ở thử nghiệm 2 (Online Retail):
  + Kết quả đã tiến sâu hơn sang Phân tích chẩn đoán (Diagnostic Analytics).
  + Không chỉ thống kê doanh thu, đã đi sâu phân tích hành vi khách hàng thông qua mô hình RFM.
  + Việc xử lý triệt để các giá trị ngoại lai (giảm độ lệch chuẩn >85%) và phát hiện nhóm khách hàng "At Risk" (Nguy cơ rời bỏ) là những điểm nhấn quan trọng, mang lại giá trị thực tiễn cao cho doanh nghiệp.

Kết luận chung về thực nghiệm

Cả hai lần chạy thử nghiệm đều đạt được mục tiêu đề ra: chuyển đổi dữ liệu thô thành thông tin có giá trị (Insights). Việc nâng cấp từ phân tích thống kê cơ bản (Adidas) sang phân tích hành vi chuyên sâu (Online Retail) cho thấy sự hoàn thiện dần của quy trình và kỹ năng phân tích. Đây là cơ sở vững chắc để khẳng định tính khả thi và hiệu quả của ứng dụng hỗ trợ ra quyết định kinh doanh mà đồ án hướng tới.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả đạt được

Tiền xử lý dữ liệu thành công:

- Làm sạch hiệu quả: Bộ dữ liệu gốc (541,909 dòng) chứa nhiều nhiễu đã được xử lý triệt để, giữ lại 526,003 giao dịch hợp lệ (chiếm ~97%).
- Xử lý giá trị thiếu: Vấn đề thiếu thông tin định danh khách hàng (khoảng 25% CustomerID) đã được giải quyết hợp lý bằng cách gán nhãn "GUEST" để không làm mất mát dữ liệu doanh thu tổng thể.
- Xử lý ngoại lai (Outliers): Các giá trị bất thường như đơn hàng hủy (số lượng âm), nợ xấu (giá âm) hoặc các đơn hàng sỉ cực lớn (>80,000 sản phẩm) đã được xử lý bằng phương pháp lọc ngưỡng và IQR. Kết quả làm giảm độ lệch chuẩn của biến số lượng và đơn giá xuống hơn 85%, giúp dữ liệu ổn định và tin cậy hơn.
- Chuẩn hóa dữ liệu: Chuyển đổi thành công dữ liệu thời gian (InvoiceDate) sang định dạng chuẩn để phục vụ phân tích chuỗi thời gian (Time-series).

Phân tích thống kê mô tả và xu hướng:

- Về hành vi mua sắm: Xác định được mô hình kinh doanh dựa trên số lượng lớn các đơn hàng giá trị nhỏ (95% đơn hàng có giá trị dưới £50).
- Tính mùa vụ: Chứng minh được xu hướng tăng trưởng doanh thu mạnh vào cuối năm, đạt đỉnh điểm vào Tháng 11/2011 (doanh thu > £1.15 triệu) để chuẩn bị cho kỳ nghỉ lễ.
- Khung giờ vàng: Xác định chính xác khung giờ mua sắm sôi động nhất là từ 10:00 sáng đến 15:00 chiều, cung cấp cơ sở khoa học cho các chiến dịch Flash Sale.

Phân tích và đánh giá phân khúc khách hàng:

- Mô hình RFM:

Đã áp dụng thành công mô hình RFM (Recency - Frequency - Monetary) để phân loại 4,290 khách hàng định danh thành các nhóm chiến lược.

- Nhận diện nhóm khách hàng trọng yếu:

  + Phát hiện nhóm Champions (Vô địch): Số lượng ít (~420 khách) nhưng chi tiêu trung bình cực cao (>£6,000).

  + Phát hiện nhóm At Risk (Nguy cơ): Có mức chi tiêu trung bình cao thứ 3 (~£1,400) nhưng đang rời bỏ doanh nghiệp. Đây là phát hiện quan trọng nhất để kích hoạt chiến lược giữ chân khách hàng.

- Trực quan hóa: Hệ thống Dashboard trực quan (Biểu đồ phân phối, Scatter plot RFM) giúp nhìn rõ bức tranh toàn cảnh về "sức khỏe" của tệp khách hàng.

5.2. Ưu nhược điểm

Ưu điểm

- Dữ liệu thực tế và Quy mô lớn: Đồ án làm việc trên bộ dữ liệu thực tế với hơn 500,000 dòng, phản ánh đúng những thách thức của dữ liệu lớn (Big Data) trong thương mại điện tử.

- Quy trình chuẩn chỉnh: Thực hiện đầy đủ và bài bản quy trình ETL (Trích xuất - Chuyển đổi - Tải), từ xử lý thô đến phân tích chuyên sâu.

- Giá trị thực tiễn cao (Actionable Insights): Các kết quả không chỉ là con số thống kê mà được chuyển hóa thành các khuyến nghị kinh doanh cụ thể (như thời điểm chạy khuyến mãi, danh sách khách hàng cần chăm sóc).

- Trực quan hóa ấn tượng: Sử dụng các kỹ thuật trực quan hóa nâng cao (Zoom-in histogram, Log-scale, Dashboard tổng hợp) giúp thông tin trở nên dễ hiểu với nhà quản lý.

Nhược điểm

- Mất cân bằng dữ liệu: Dữ liệu bị lệch quá nhiều về thị trường Vương quốc Anh (>85%), làm hạn chế khả năng phân tích hành vi tiêu dùng toàn cầu.

- Giới hạn của phương pháp RFM: Việc phân nhóm hiện tại dựa trên các quy tắc thống kê (Rule-based) và phân vị (Quintiles), có thể chưa tối ưu bằng các thuật toán học máy không giám sát (như K-Means Clustering) để tìm ra các nhóm ẩn.

- Chưa có mô hình dự báo: Đồ án mới dừng lại ở phân tích mô tả (Descriptive Analytics) và chẩn đoán (Diagnostic), chưa áp dụng các mô hình học máy (Machine Learning) để dự báo doanh thu tương lai.

5.3. Hướng phát triển

Để cải thiện các nhược điểm và phát triển có những đề xuất sau:

- Ứng dụng Học máy (Machine Learning):

  + Sử dụng thuật toán K-Means Clustering để phân nhóm khách hàng tự động, so sánh hiệu quả với mô hình RFM truyền thống nhằm tìm ra các phân khúc ngách.

  + Xây dựng Hệ thống gợi ý (Recommendation System) sử dụng thuật toán Apriori hoặc Collaborative Filtering để gợi ý sản phẩm bán chéo (Cross-sell) cho nhóm khách hàng Tiềm năng.

- Dự báo nâng cao (Forecasting):

Áp dụng các mô hình chuỗi thời gian như ARIMA hoặc Prophet để dự báo doanh thu tháng 12 và Quý 1 năm sau, hỗ trợ doanh nghiệp tối ưu hóa kế hoạch nhập kho và nhân sự.

- Tối ưu hóa chiến lược Marketing tự động:

Dựa trên kết quả phân khúc RFM, xây dựng các kịch bản Marketing tự động (Marketing Automation): Gửi mã giảm giá cho nhóm "At Risk", gửi thông tin sản phẩm mới cho nhóm "Champions".

- Mở rộng phạm vi dữ liệu:

Kết hợp thêm các nguồn dữ liệu bên ngoài (như dữ liệu kinh tế vĩ mô, ngày lễ, thời tiết) để phân tích các yếu tố ngoại cảnh ảnh hưởng đến sức mua.

5.4. Kết luận

Quá trình triển khai đồ án "Tìm hiểu về Phân tích dữ liệu (Data Analytics)" trên bộ dữ liệu Online Retail đã đạt được những kết quả quan trọng. Đã hoàn thành việc xây dựng một quy trình xử lý và phân tích dữ liệu hoàn chỉnh, từ việc làm sạch dữ liệu thô đến việc trích xuất các tri thức kinh doanh giá trị.

Kết quả phân tích đã chứng minh rằng dữ liệu không chỉ là những con số vô tri mà là tài sản chiến lược. Việc xác định được tính mùa vụ, khung giờ vàng và phân loại chính xác các nhóm khách hàng (đặc biệt là nhóm Champions và At Risk) cung cấp cơ sở vững chắc cho doanh nghiệp chuyển đổi từ việc "bán hàng theo cảm tính" sang "ra quyết định dựa trên dữ liệu" (Data-driven decision making).

Tuy vẫn còn một số hạn chế về phương pháp dự báo, nhưng đồ án đã đặt nền móng vững chắc cho việc ứng dụng Khoa học dữ liệu vào thực tế doanh nghiệp. Những hướng phát triển tiếp theo như ứng dụng AI/Machine Learning hứa hẹn sẽ mang lại những bước đột phá lớn hơn nữa trong việc tối ưu hóa hiệu quả kinh doanh và trải nghiệm khách hàng.

TÀI LIỆU THAM KHẢO

[1] D. Chen, "Online Retail Data Set," *UCI Machine Learning Repository*, 2015. [Online]. Available: https://archive.ics.uci.edu/dataset/352/online+retail.

[2] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 51-56.

[3] The Pandas Development Team, "pandas documentation," 2024. [Online]. Available: https://pandas.pydata.org/docs/.

[4] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, 2007.

[5] M. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[6] A. M. Hughes, *Strategic Database Marketing*. Chicago, IL, USA: Probus Publishing Company, 1994. (Tài liệu gốc về mô hình RFM).

[7] Xử lý và phân tích dữ liệu: Động lực cho chuyển đổi số quốc gia: https://baochinhphu.vn/xu-ly-va-phan-tich-du-lieu-dong-luc-cho-chuyen-doi-so-quoc-gia-102250528164243578.htm