



TRƯỜNG KỸ THUẬT CÔNG NGHỆ  
KHOA CÔNG NGHỆ THÔNG TIN

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:

Cô Nhân Quý - 110122150 - DA22TTB

# MỤC LỤC

- 
- 1 Giới thiệu đề tài
  - 2 Nghiên cứu lý thuyết
  - 3 Hiện thực hóa nghiên cứu
  - 4 Kết quả đạt được
  - 5 Kết luận và Hướng phát triển
-

# Giới thiệu đề tài

---

## Lý do chọn đề tài:

- Xu hướng tất yếu: Dữ liệu là tài nguyên chiến lược trong nền kinh tế số và chuyển đổi số.
- Giá trị thực tiễn: Giúp doanh nghiệp chuyển dịch sang ra quyết định dựa trên dữ liệu (Data-driven) để tối ưu vận hành.
- Phù hợp chuyên môn: Vận dụng kỹ năng lập trình Python vào bài toán thực tế của ngành thương mại điện tử.

# Giới thiệu đề tài

---

## Mục tiêu:

- Nắm vững quy trình: Nghiên cứu và áp dụng trọn vẹn quy trình Data Analytics (Thu thập -> Làm sạch -> Xử lý -> Trực quan hóa).
- Xử lý dữ liệu thực tế: Làm sạch và chuẩn hóa bộ dữ liệu Online Retail (xử lý dữ liệu thiếu, ngoại lai).
- Phân tích chuyên sâu: Đánh giá xu hướng doanh thu và phân khúc khách hàng bằng mô hình RFM.
- Xây dựng ứng dụng: Tạo Dashboard trực quan hóa kết quả phân tích bằng Python (Streamlit).

# Nghiên cứu lý thuyết

---

- Dữ liệu nghiên cứu: Sử dụng bộ dữ liệu chuẩn Online Retail (từ UCI Machine Learning Repository) gồm hơn 540.000 dòng giao dịch thực tế để thực nghiệm.
- Phân tích thống kê (Statistical Analysis): Sử dụng các chỉ số mô tả (Mean, Median, Std) và trực quan hóa để khám phá đặc điểm, xu hướng phân phối của dữ liệu giao dịch.
- Mô hình RFM: Áp dụng kỹ thuật phân tích 3 chỉ số: Recency (Gần nhất) - Frequency (Tần suất) - Monetary (Giá trị) để phân khúc và đánh giá hành vi khách hàng.

# Hiện thực hóa nghiên cứu

## Công cụ và môi trường phát triển

- Ngôn ngữ lập trình:** Python 3.10+ (Ngôn ngữ chuẩn cho Khoa học dữ liệu).
- Môi trường thực nghiệm:**
  - Google Colab: Dùng để chạy mã phân tích và huấn luyện mô hình nhanh chóng.
- Thư viện chính (Libraries):**
  - Xử lý dữ liệu: Pandas, NumPy.
  - Trực quan hóa: Matplotlib, Seaborn (Vẽ biểu đồ).
  - Ứng dụng Web: Streamlit (Tạo Dashboard tương tác).

# Hiện thực hóa nghiên cứu

## Quy trình Phân tích dữ liệu (Workflow)

---

**1. Thu thập dữ liệu:** Tải bộ dữ liệu Online Retail từ UCI Machine Learning Repository (541.909 dòng giao dịch).

### 2. Tiền xử lý (Preprocessing):

- Làm sạch giá trị thiếu (Gán nhãn "GUEST" cho khách vãng lai).
- Xử lý dữ liệu ngoại lai (Outliers) và chuẩn hóa định dạng thời gian.

# Hiện thực hóa nghiên cứu

## Quy trình Phân tích dữ liệu (Workflow)

---

- Phân tích thống kê:** Tính toán các chỉ số doanh thu, xác định xu hướng theo thời gian và sản phẩm bán chạy.
- Phân tích RFM:** Phân khúc khách hàng dựa trên 3 chỉ số: Thời gian mua gần nhất (R) - Tần suất mua (F) - Giá trị đơn hàng (M).
- Đánh giá & Trực quan hóa:** Xây dựng Dashboard hiển thị kết quả và rút ra nhận định kinh doanh.

# Kết quả thực nghiệm

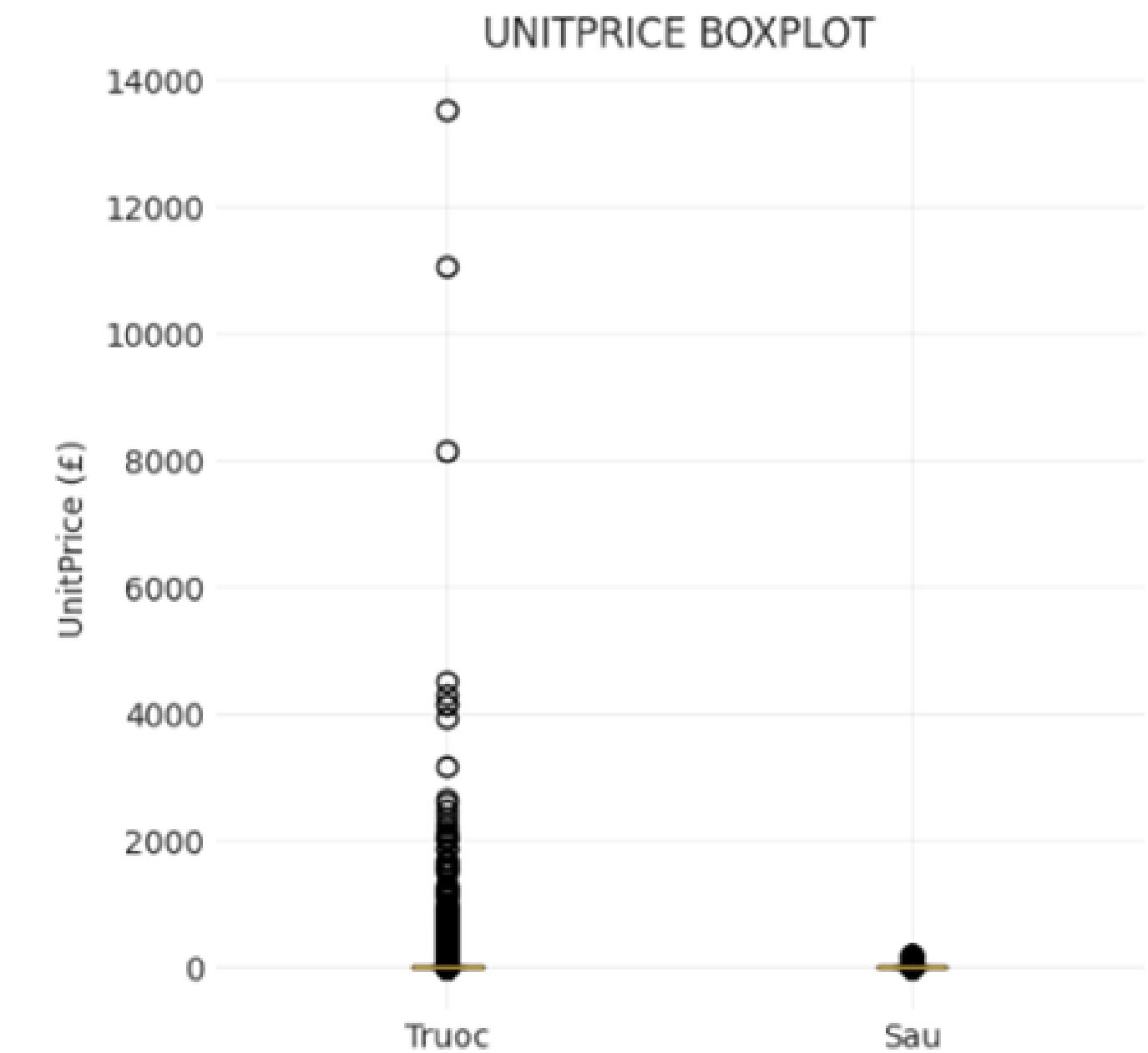
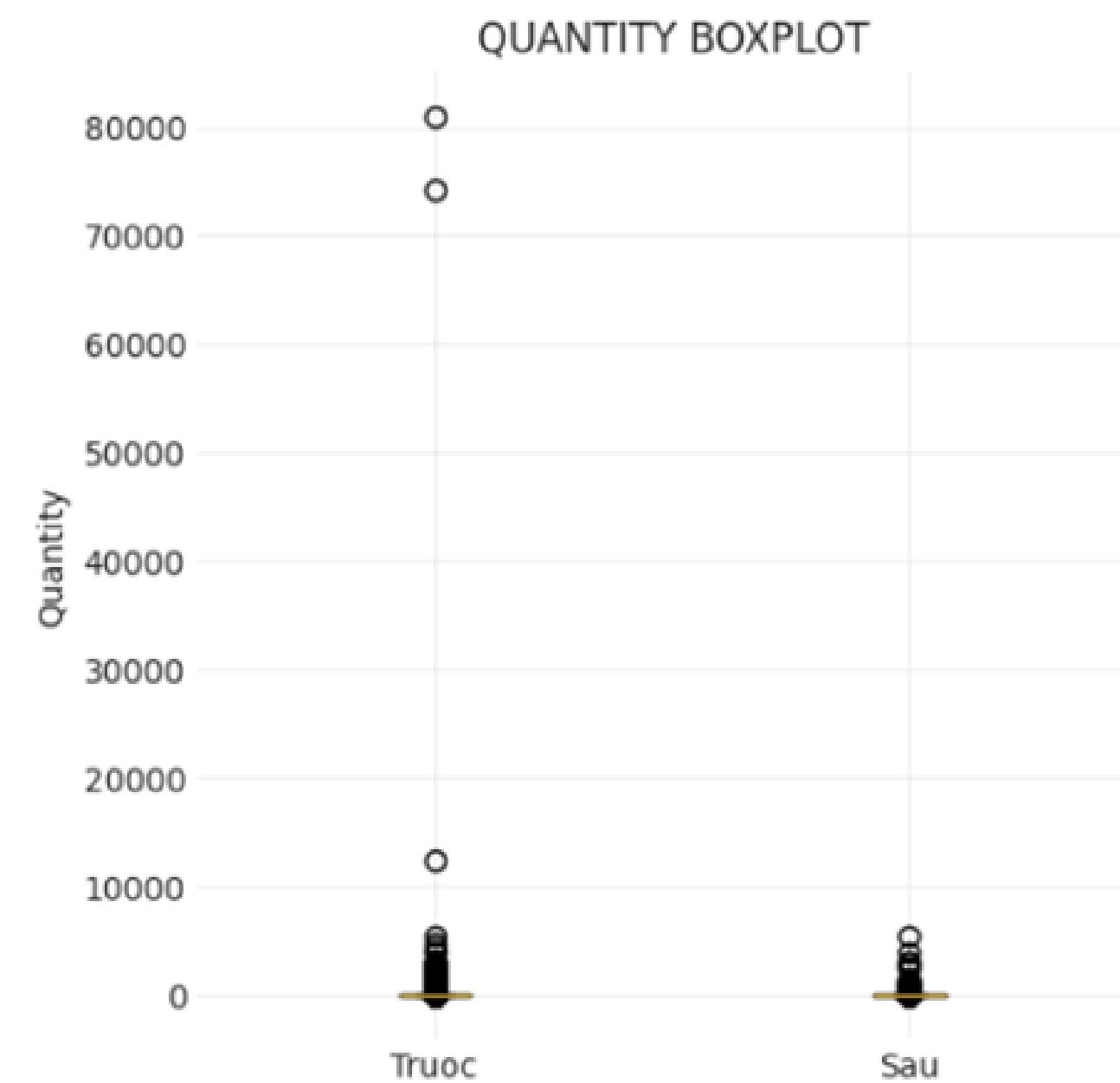
## Tiền xử lý & Phân tích xu hướng

---

- Làm sạch dữ liệu: Xử lý triệt để dữ liệu nhiễu và ngoại lai, giữ lại được 97% dữ liệu hợp lệ (526,003 dòng). Độ lệch chuẩn giảm hơn 85%, giúp dữ liệu ổn định và tin cậy.
- Tính mùa vụ: Phát hiện doanh thu tăng trưởng mạnh vào cuối năm, đạt đỉnh điểm vào Tháng 11 (>£1.15 triệu) để chuẩn bị cho kỳ nghỉ lễ.

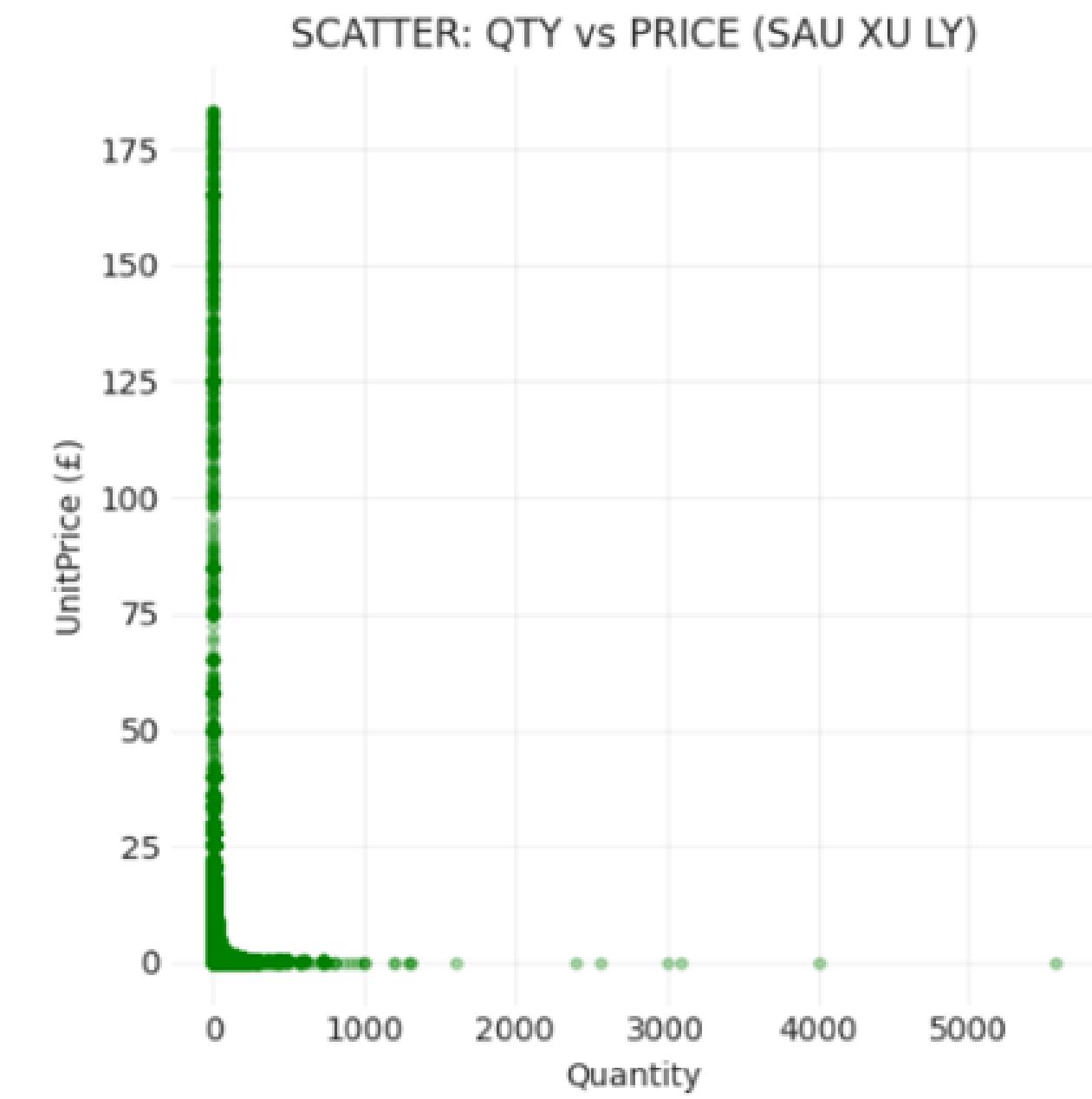
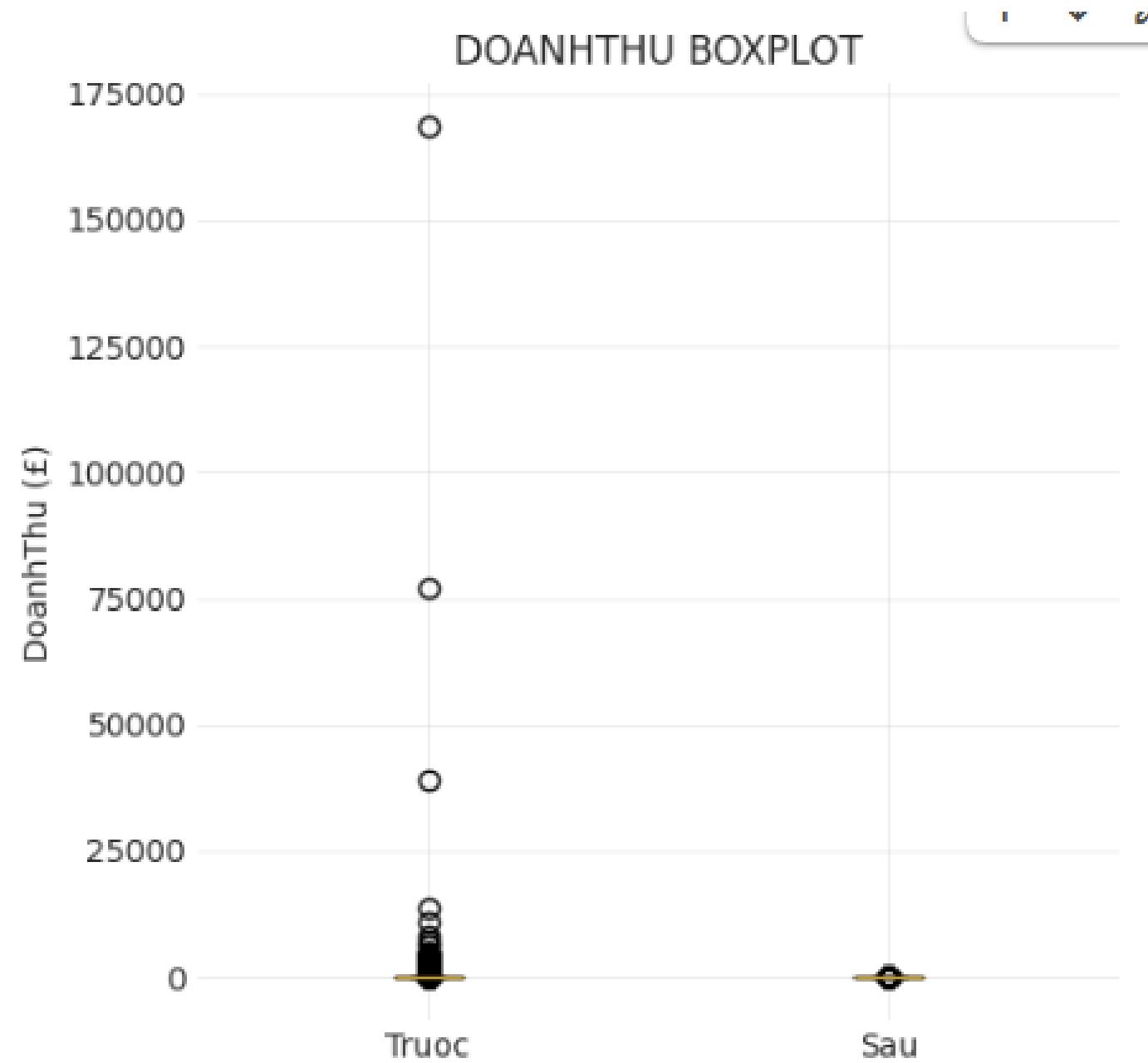
# Kết quả thực nghiệm

## Tiền xử lý & Phân tích xu hướng



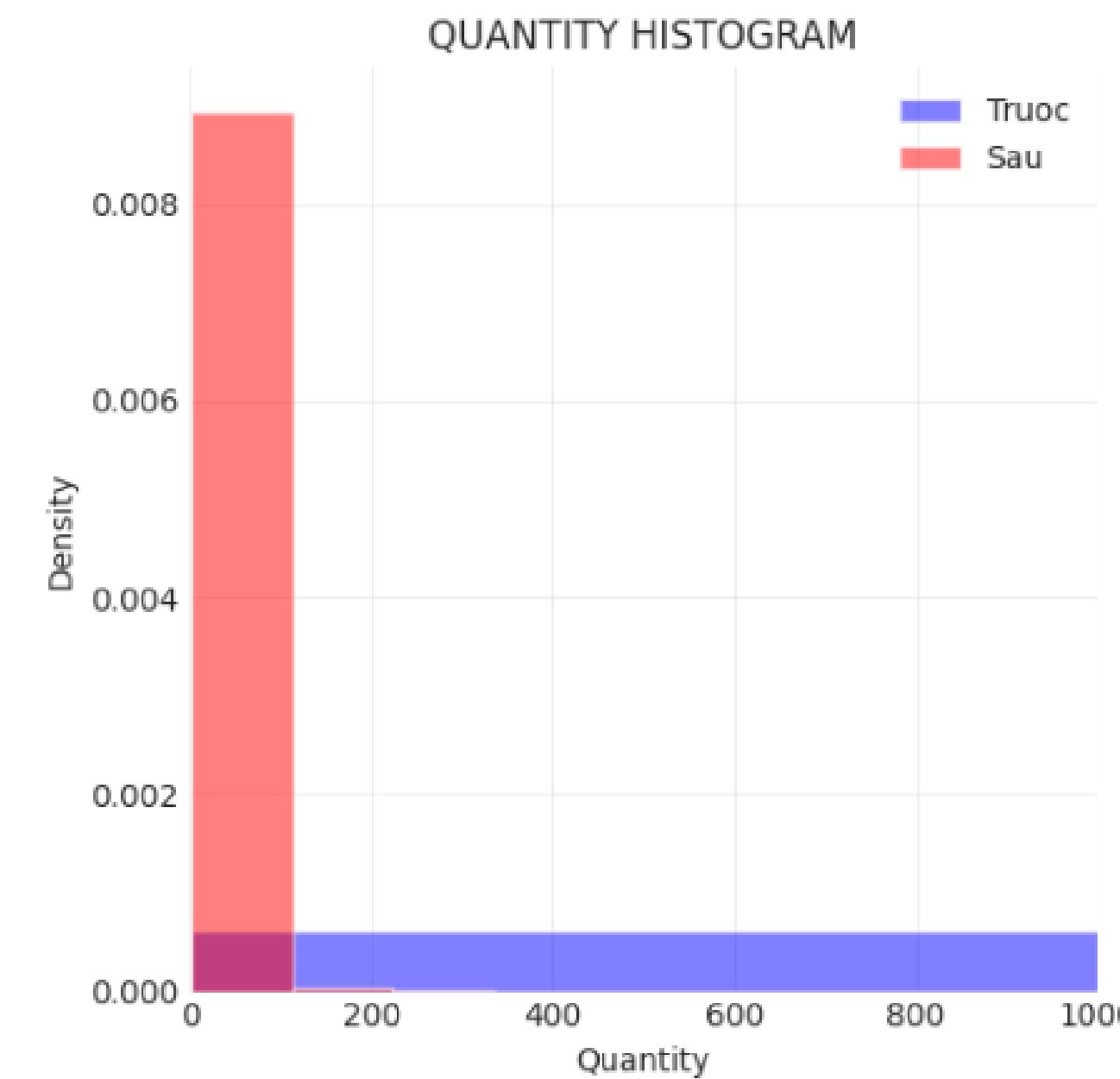
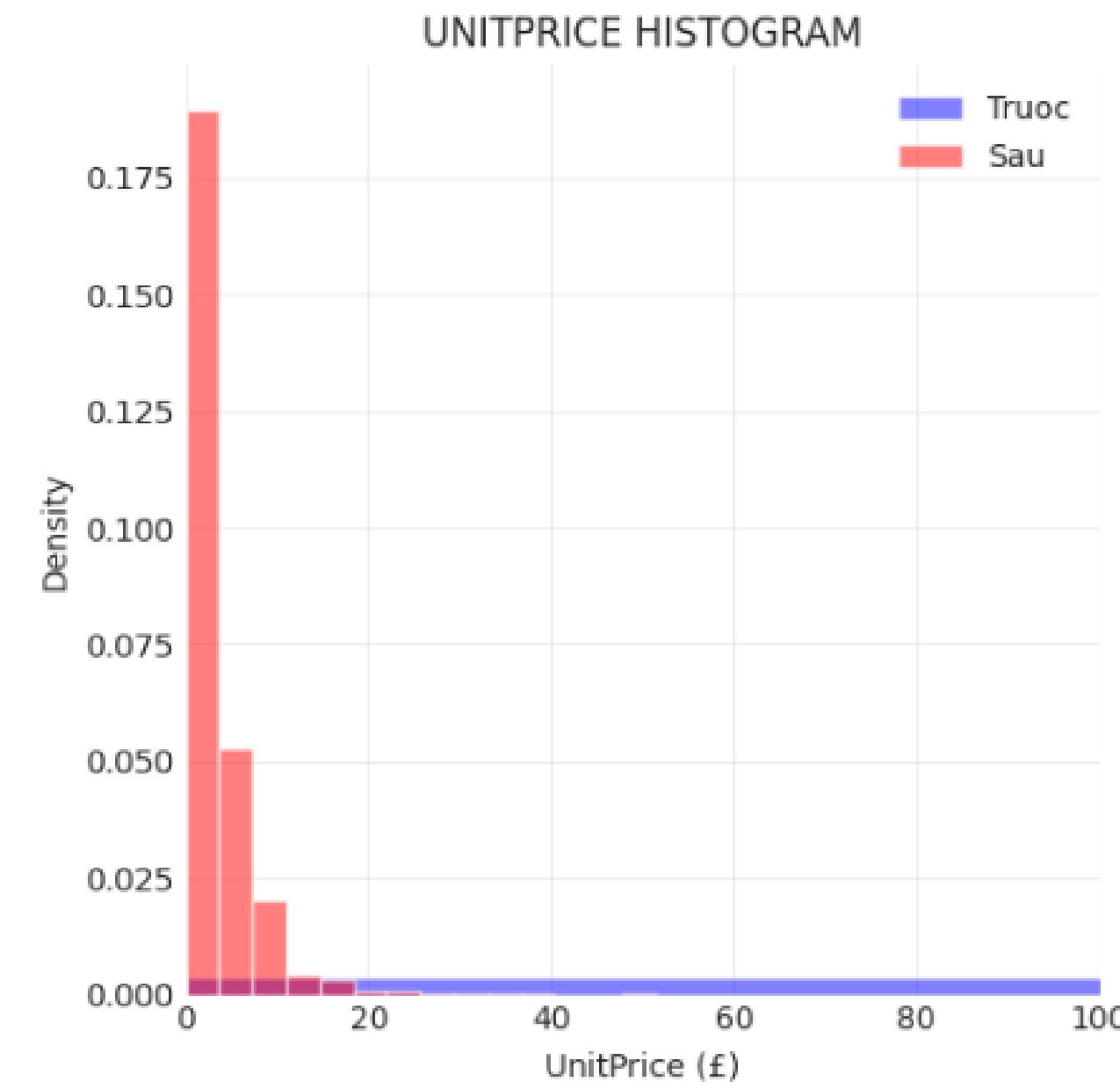
# Kết quả thực nghiệm

# Tiền xử lý & Phân tích xu hướng



# Kết quả thực nghiệm

## Tiền xử lý & Phân tích xu hướng



DOANH THU THEO THANG (x1000 £)



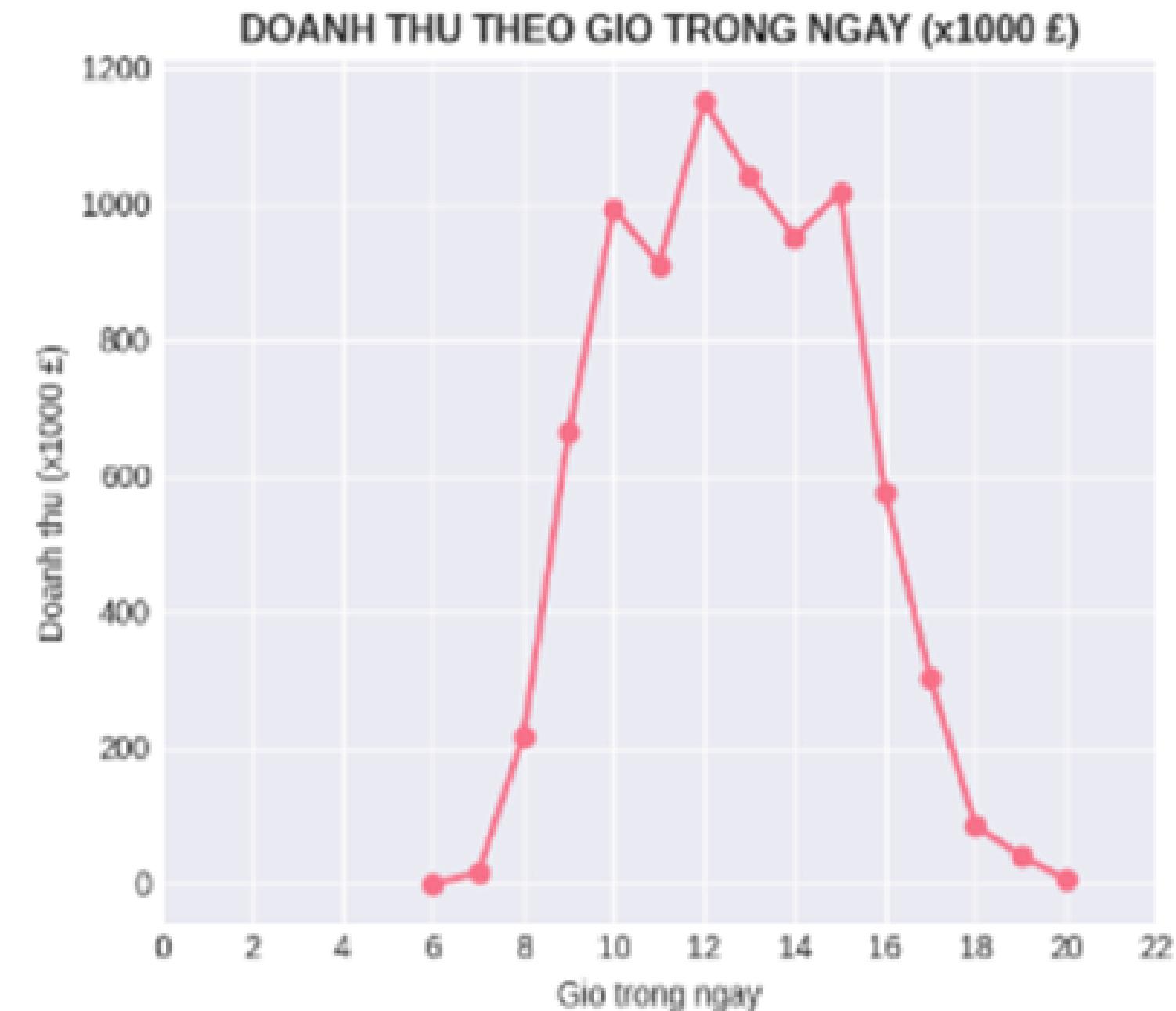
SO LUONG DON HANG THEO NGAY



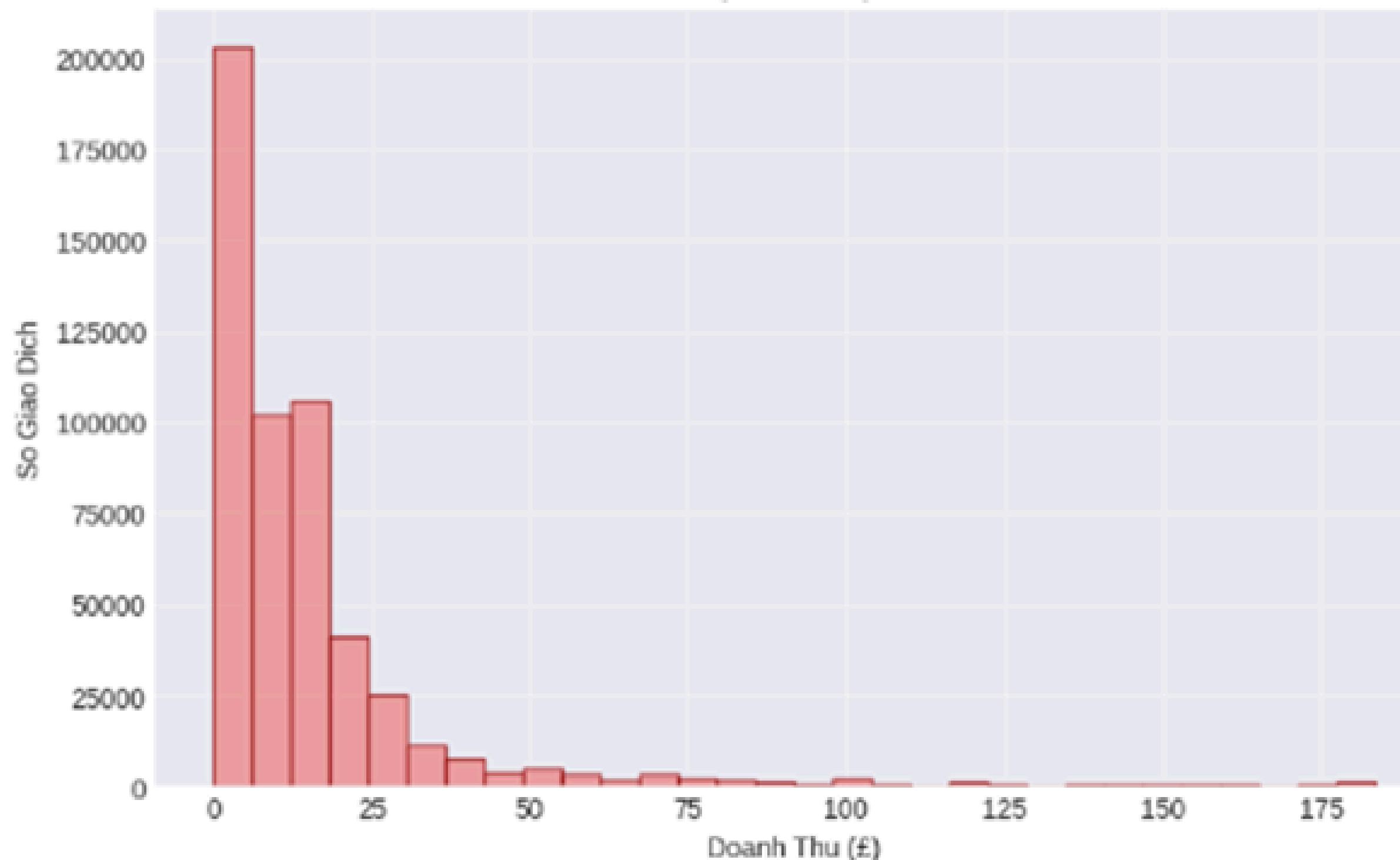
# Kết quả thực nghiệm

## Tiền xử lý & Phân tích xu hướng

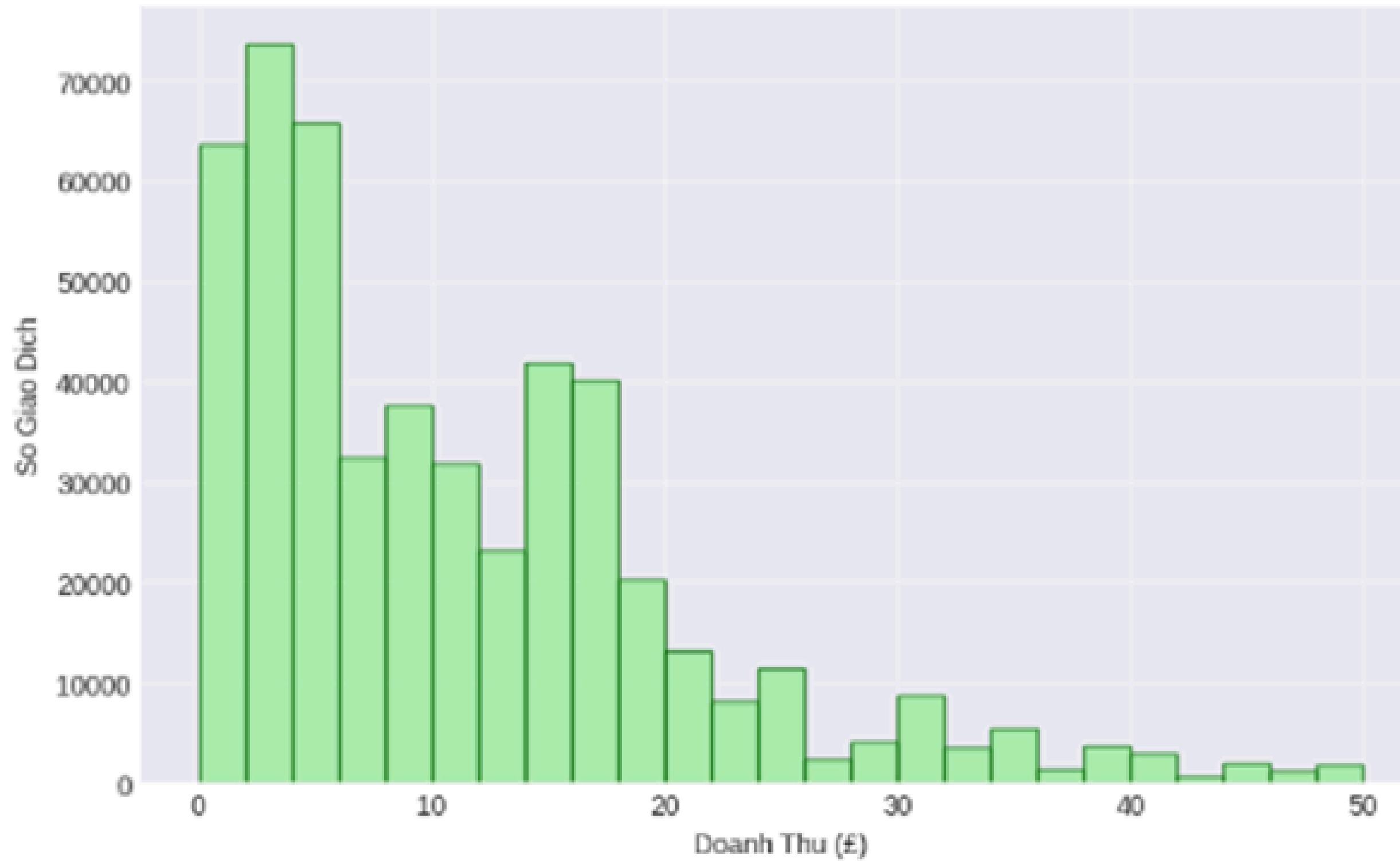
- Khung giờ vàng: Xác định thời điểm mua sắm sôi động nhất là từ 10:00 - 15:00, hỗ trợ tối ưu hóa nhân sự và chiến dịch Flash Sale.
- Mô hình kinh doanh: 95% đơn hàng có giá trị nhỏ (<£50), phản ánh đặc thù bán lẻ các món quà lưu niệm/vật dụng nhỏ.



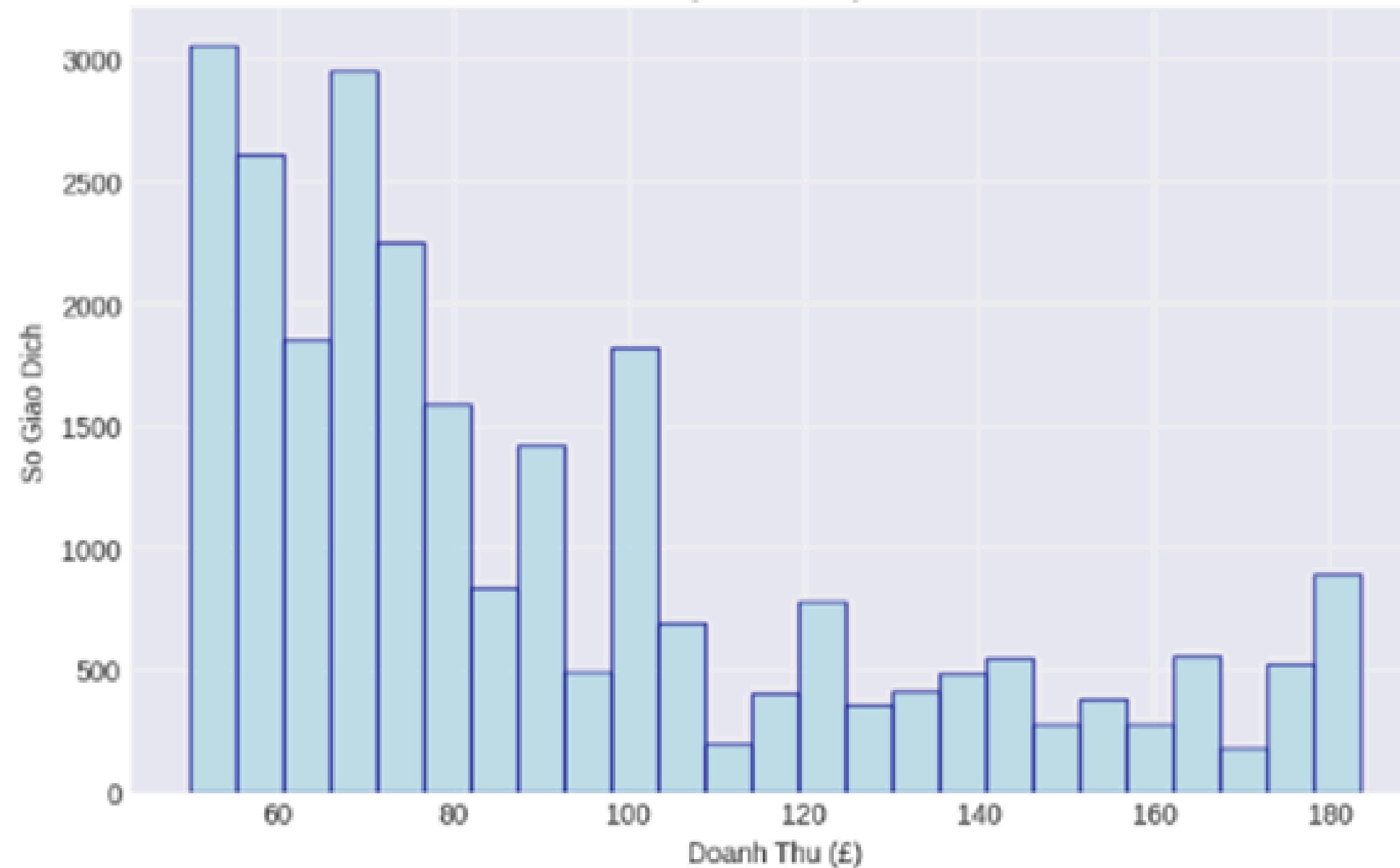
## PHAN BO DOANH THU CHINH (£0 - £500)



## PHAN BO DOANH THU NHO (£0 - £50)



## PHAN BO DOANH THU TRUNG BINH (£50 - £200)



# Kết quả thực nghiệm

## Phân khúc khách hàng

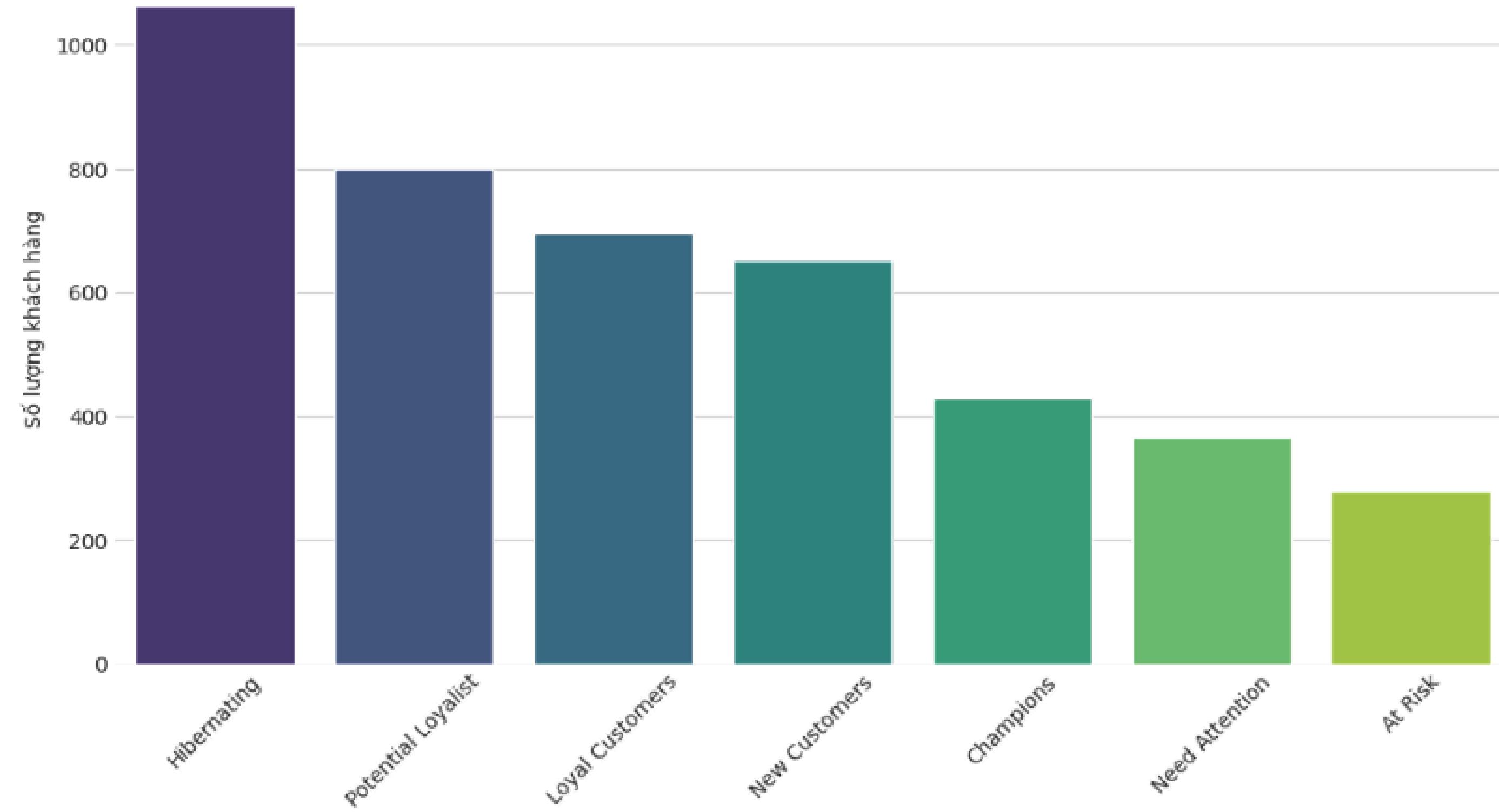
---

Phân loại chiến lược: Phân nhóm thành công 4,290 khách hàng thành các nhóm riêng biệt dựa trên hành vi mua sắm.

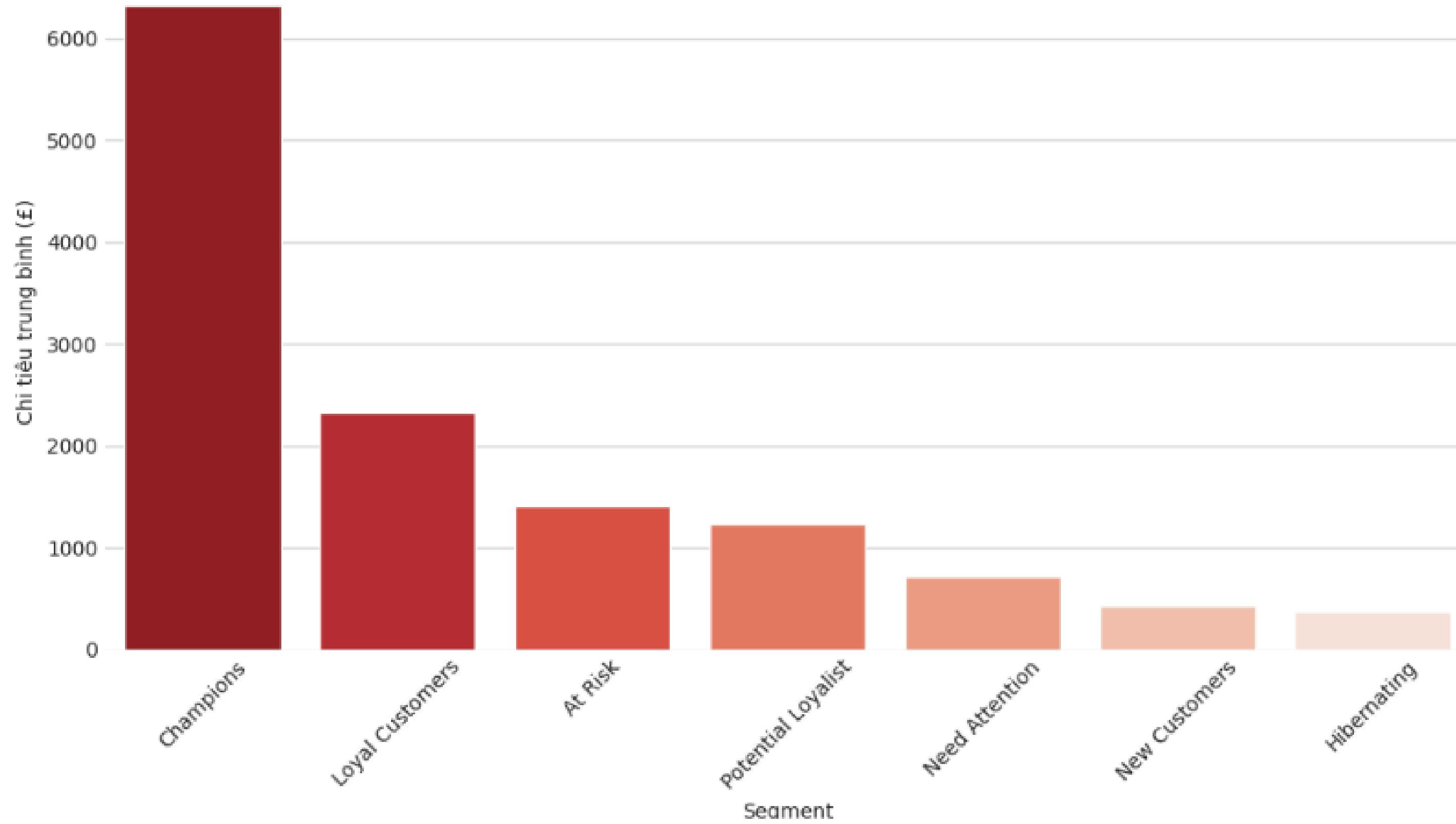
- Nhóm Champions (Vô địch): Số lượng ít (~420 khách) nhưng chi tiêu "khủng" (>£6,000/người), là nhóm gánh doanh thu chính.
- Nhóm At Risk (Nguy cơ): Chi tiêu cao thứ 3 hệ thống (~£1,400) nhưng đang rời bỏ doanh nghiệp.

Ý nghĩa: Chuyển từ tiếp thị đại trà sang chăm sóc khách hàng cá nhân hóa theo từng nhóm.

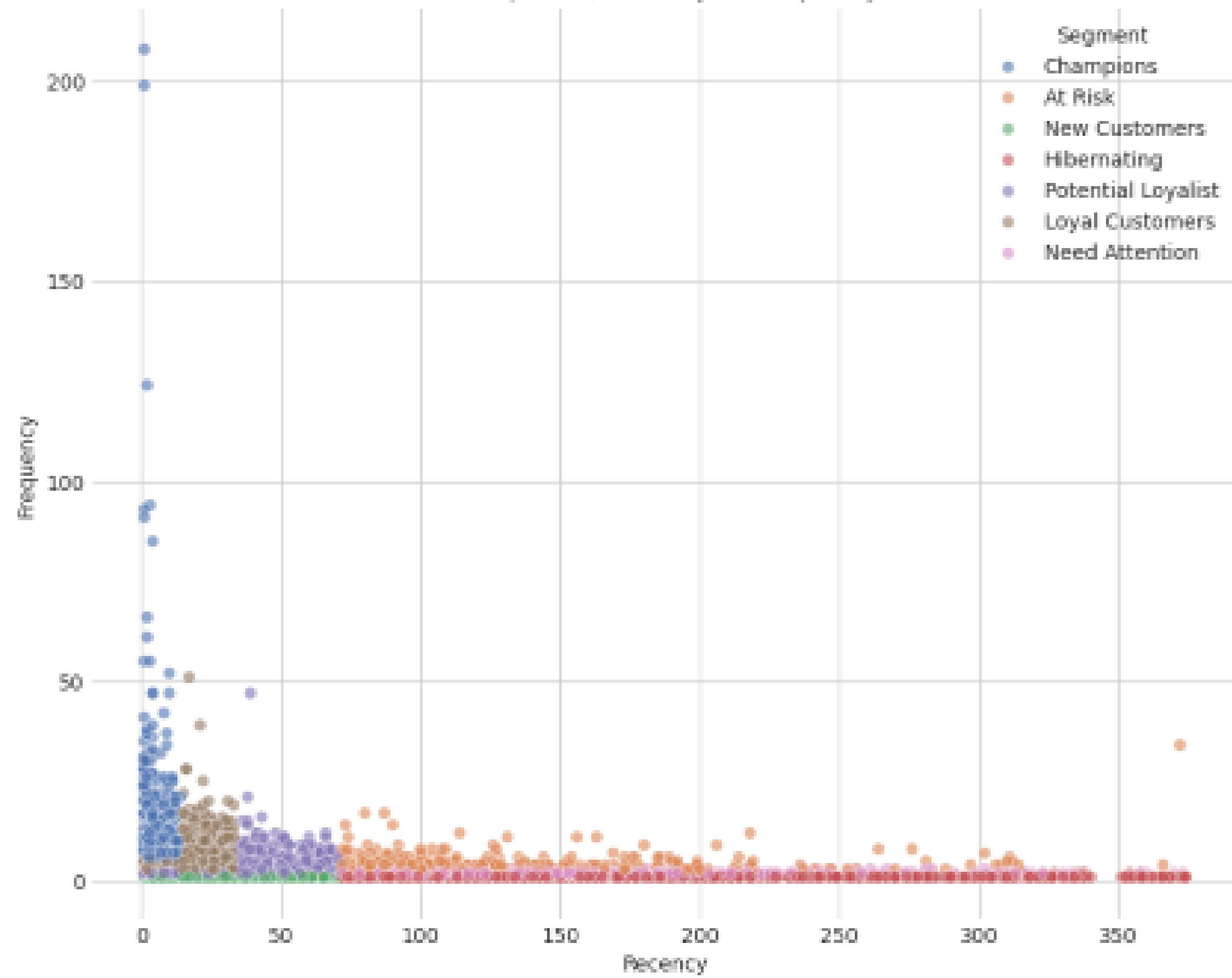
## Số lượng khách hàng theo phân khúc (Customer Segments Distribution)



Giá trị chi tiêu trung bình theo phân khúc



Mối quan hệ Recency vs Frequency



# Kết quả thực nghiệm

## Xây dựng ứng dụng và Demo

---

1. Tính đa năng: Ứng dụng Python/Streamlit đã chạy thử nghiệm thành công trên 2 bộ dữ liệu khác nhau: Adidas US Sales (Báo cáo tổng hợp) và Online Retail (Giao dịch chi tiết).
2. Tính năng nổi bật:
  - Tự động hóa quy trình làm sạch và so sánh thống kê trước/sau xử lý.
  - Dashboard tương tác trực quan (Biểu đồ xu hướng, Phân phối RFM) giúp nhà quản lý dễ dàng nắm bắt thông tin.

Online Retail.csv X  
43.5MB

▼ Cấu hình cột (Auto)

Ngày  
InvoiceDate

Số lượng  
Quantity

Giá/Tiền  
UnitPrice

Khách hàng  
CustomerID

Sản phẩm  
Description

 PHÂN TÍCH NGAY

# 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
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United

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

Online Retail.csv X

43.5MB

Cấu hình cột (Auto)

Ngày  
InvoiceDate

Số lượng  
Quantity

Giá/Tiền  
UnitPrice

Khách hàng  
CustomerID

Sản phẩm  
Description

**Ứng Dụng Phân Tích Dữ Liệu Doanh thu bán hàng**

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

## 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

**PHÂN TÍCH NGAY**

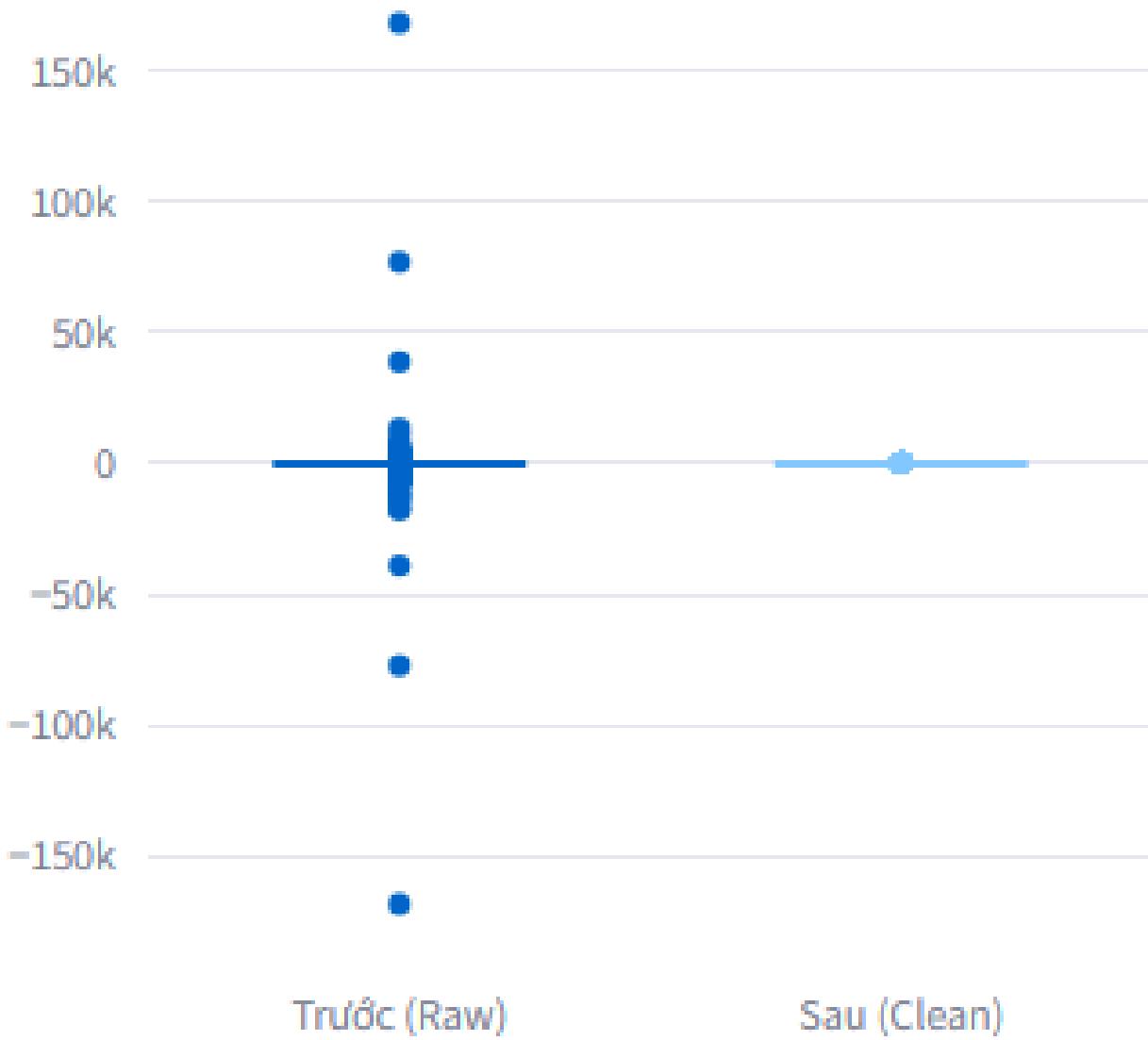
**Bảng so sánh chỉ số thống kê (Trước vs Sau)**

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

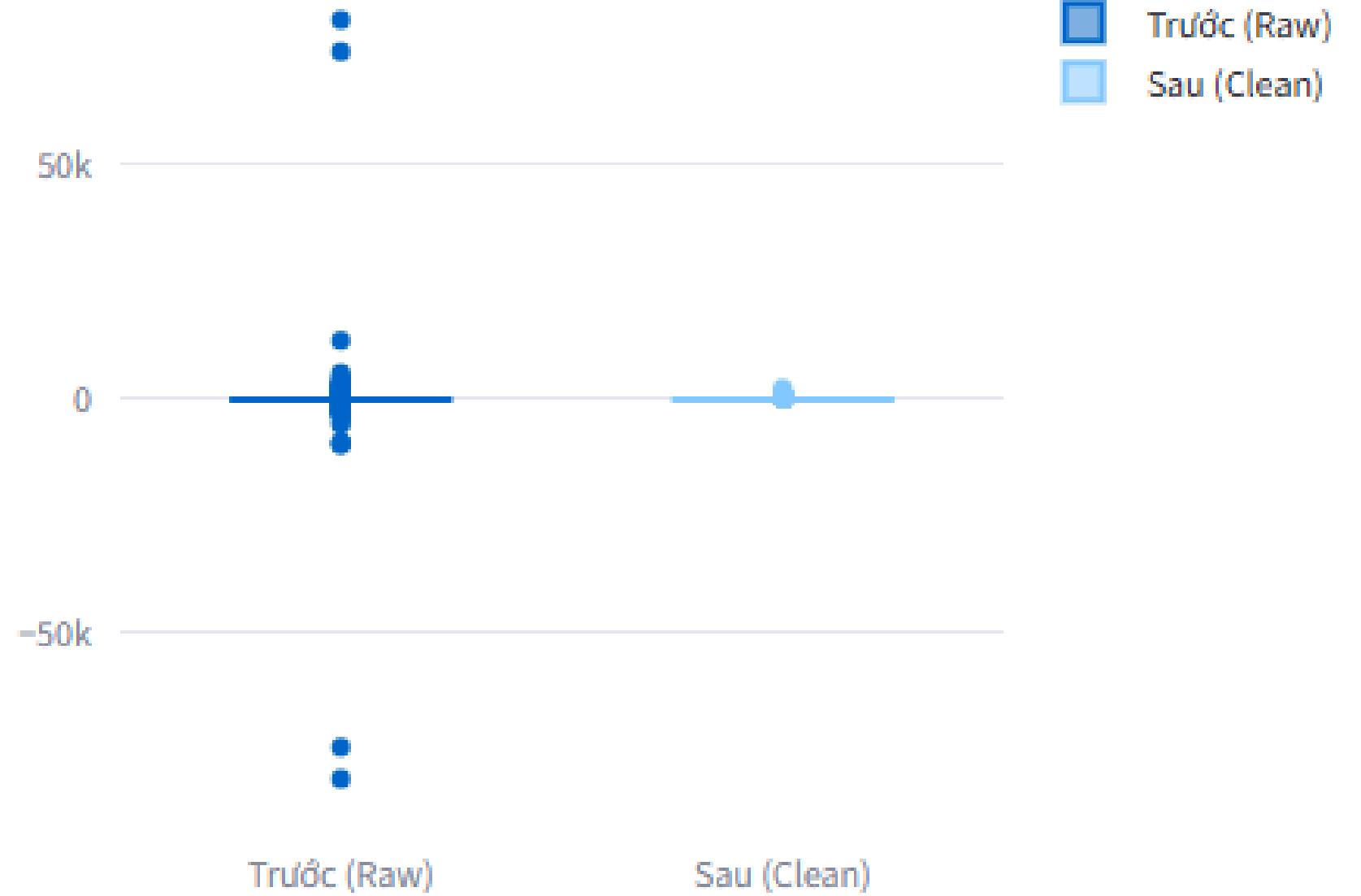
	SL (Trước)	SL (Sau)	Tiền (Trước)	Tiền (Sau)
count	541,909.00	524,825.00	541,909.00	524,825.00
mean	9.55	8.54	17.99	15.17
std	218.08	18.83	378.81	21.31
min	-80,995.00	1.00	-168,469.60	0.00
25%	1.00	1.00	3.40	3.75
50%	3.00	3.00	9.75	9.90
75%	10.00	10.00	17.40	17.40
max	80,995.00	2,400.00	168,469.60	183.60

Ghi chú: Bảng trên giúp so sánh các chỉ số như Trung bình (mean), Độ lệch (std) và Cực đại (max) thay đổi thế nào sau khi loại bỏ nhiễu.

## Phân bố Doanh thu



## Phân bố Số lượng

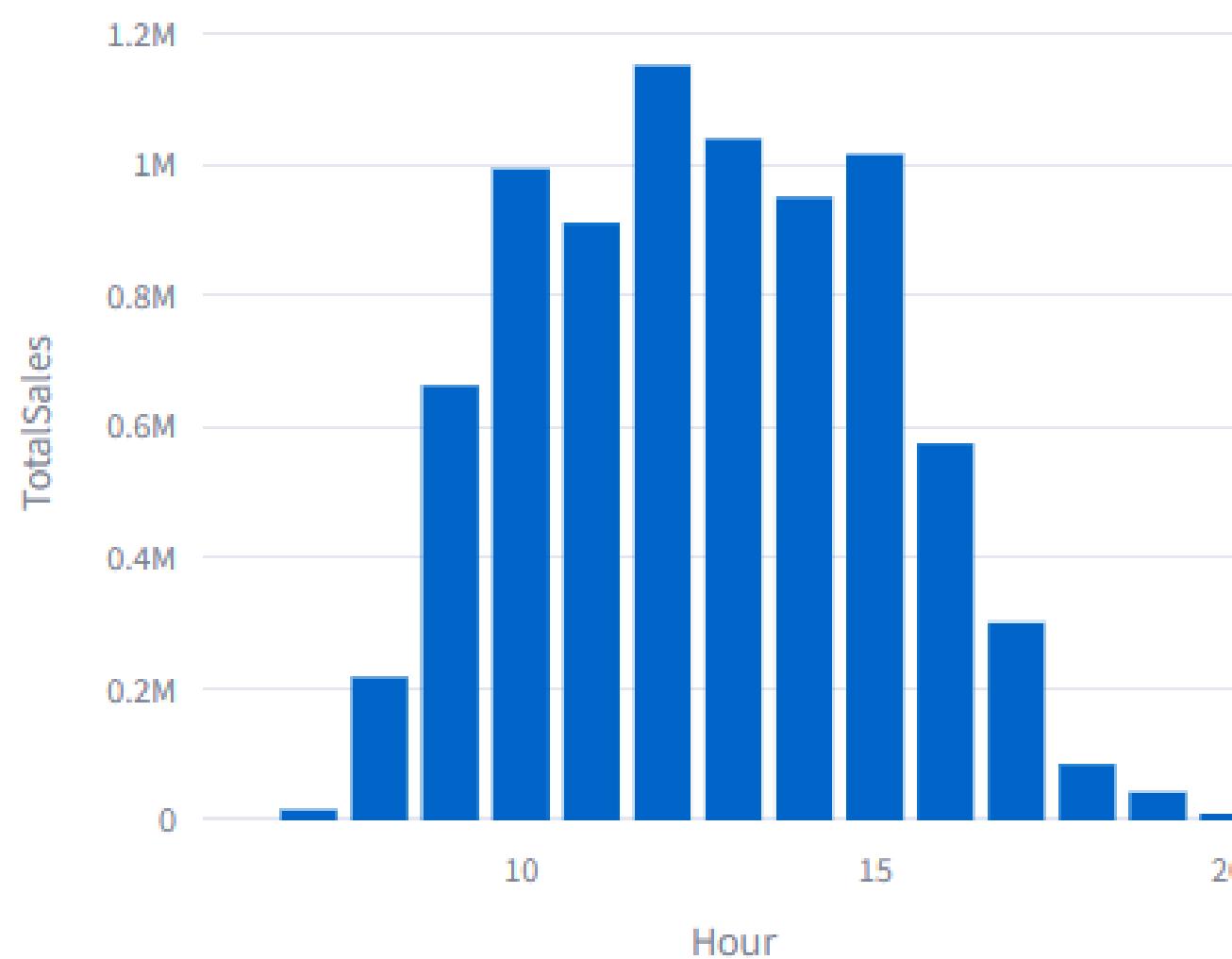


# Phân tích Chuyên sâu

Doanh thu theo Tháng



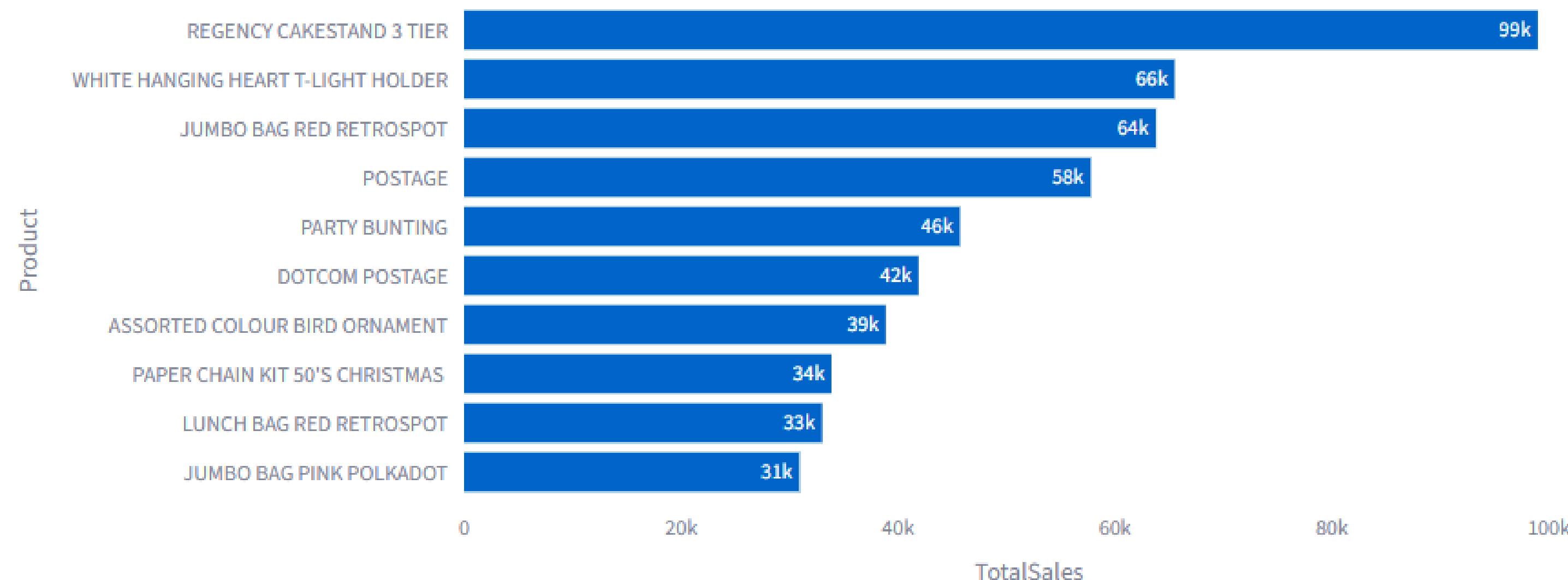
Khung giờ vàng



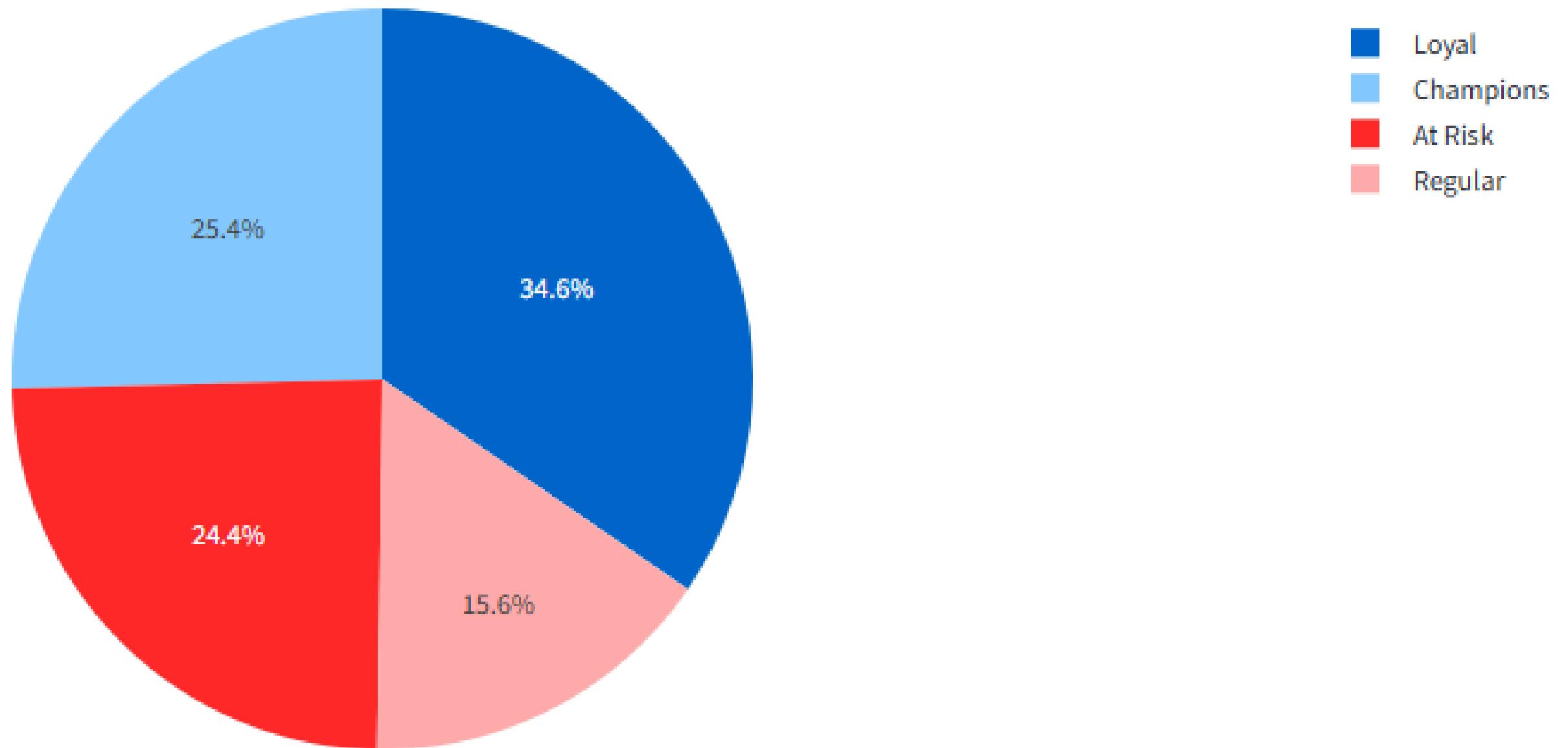
# Top Sản Phẩm



## Top 10 Sản phẩm



## Phân nhóm Khách hàng (RFM) ↪



# Kết luận và hướng phát triển

## Kết luận

---

Quy trình hoàn chỉnh: Xây dựng thành công quy trình xử lý dữ liệu khép kín, làm sạch triệt để nhiễu và ngoại lai, giữ lại 97% dữ liệu hợp lệ (526,003 dòng).

- Xác định chính xác tính mùa vụ (đỉnh điểm tháng 11) và "khung giờ vàng" (10:00 - 15:00) để tối ưu chiến dịch bán hàng.
- Phân loại thành công 4,290 khách hàng qua mô hình RFM, đặc biệt là nhận diện nhóm "At Risk" (chi tiêu cao nhưng đang rời bỏ).

# Kết luận và hướng phát triển

## Hướng phát triển

---

- **Ứng dụng Học máy (Machine Learning):** Sử dụng thuật toán K-Means Clustering để phân nhóm khách hàng tự động thay vì các quy tắc cố định, giúp tìm ra các phân khúc ngách.
- **Dự báo nâng cao (Forecasting):** Áp dụng các mô hình chuỗi thời gian như ARIMA hoặc Prophet để dự báo doanh thu tương lai, hỗ trợ lập kế hoạch nhập kho.

# Kết luận và hướng phát triển

## Hướng phát triển

---

- **Hệ thống gợi ý (Recommendation System):** Xây dựng tính năng gợi ý sản phẩm bán chéo (Cross-sell) cho từng nhóm khách hàng tiềm năng.
- **Tự động hóa Marketing:** Kết hợp kết quả RFM để kích hoạt các kịch bản chăm sóc khách hàng tự động (gửi mã giảm giá, email marketing).

**CẢM ƠN THẦY CÔ VÀ  
CÁC BẠN ĐÃ XEM**