BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC ĐẠI NAM

BÀI TẬP LỚN

TÊN HỌC PHẦN: XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU

ĐỀ TÀI:PHÂN TÍCH DỮ LIỆU GIAO DỊCH CỦA MỘT SIÊU THỊ ĐỂ ĐÁNH GIÁ HIỆU QUẢ CHƯƠNG TRÌNH KHUYẾN MÃI

Giáo viên hướng dẫn: ThS. Lê Diệu Anh

Sinh viên thực hiện:

STT	Mã sv	Họ và tên	Lớp
1	1771020707	Trần Anh Tú	CNTT 17-15
2	1771020729	Nguyễn Thanh Tùng	CNTT 17-15
3	1771020663	Phạm Đức Duy Tiến	CNTT 17-15

Hà Nội, năm 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC ĐẠI NAM

BÀI TẬP LỚN

TÊN HỌC PHẦN: XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU

ĐỀ TÀI:PHÂN TÍCH DỮ LIỆU GIAO DỊCH CỦA MỘT SIÊU THỊ ĐỂ ĐÁNH GIÁ HIỆU QUẢ CHƯƠNG TRÌNH KHUYẾN MÃI

STT	Mã Sinh Viên	Họ và Tên	Ngày Sinh	Điểm	
				Bằng Số	Bằng Chữ
1	1771020707	Trần Anh Tú	05/06/2005		
2	1771020729	Nguyễn Thanh Tùng	01/09/2004		
3	1771020663	Phạm Đức Duy Tiến	09/10/2005		

CÁN BỘ CHẤM THI

Hà Nội, năm 2025

LỜI NÓI ĐẦU

Trong thời đại số hóa hiện nay, dữ liệu đóng vai trò vô cùng quan trọng trong mọi lĩnh vực, đặc biệt là trong ngành bán lẻ. Việc tận dụng dữ liệu một cách hiệu quả không chỉ giúp doanh nghiệp hiểu rõ hành vi mua sắm của khách hàng mà còn hỗ trợ ra quyết định chiến lược một cách chính xác và khoa học hơn. Một trong những yếu tố quan trọng ảnh hưởng đến doanh số của siêu thị chính là các chương trình khuyến mãi. Các chiến dịch khuyến mãi không chỉ giúp tăng cường doanh thu ngắn hạn mà còn tác động trực tiếp đến thói quen mua sắm, khả năng quay lại của khách hàng và mức độ trung thành đối với thương hiệu.

Tuy nhiên, không phải chương trình khuyến mãi nào cũng mang lại hiệu quả như mong đợi. Một số chương trình có thể giúp tăng doanh thu nhưng lại làm giảm lợi nhuận, trong khi một số khác có thể thu hút nhiều khách hàng mới nhưng không duy trì được sự gắn kết lâu dài. Do đó, việc phân tích dữ liệu giao dịch để đánh giá hiệu quả của các chương trình khuyến mãi là điều cần thiết.

Với đề tài "Phân tích dữ liệu giao dịch của một siêu thị để đánh giá hiệu quả chương trình khuyến mãi", nhóm nghiên cứu sử dụng bộ dữ liệu giao dịch thực tế của một siêu thị, áp dụng các phương pháp thống kê mô tả, mô hình hồi quy tuyến tính, trực quan hóa dữ liệu và dự đoán xu hướng doanh thu để phân tích tác động của chương trình khuyến mãi đến hành vi mua sắm của khách hàng. Qua đó, đề tài không chỉ giúp đánh giá hiệu quả của từng chương trình khuyến mãi mà còn đưa ra các khuyến nghị giúp tối ưu hóa chiến lược bán hàng trong tương lai.

MỤC LỤC

MỤC LỤC HÌNH ẢNH

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1. Giới thiệu chung về đề tài lựa chọn

1.1.1. Tổng quan về phân tích dữ liệu trong kinh doanh

Trong thời đại số hóa, dữ liệu đóng vai trò quan trọng trong việc hỗ trợ ra quyết định kinh doanh. Đặc biệt, trong ngành bán lẻ và hệ thống siêu thị, dữ liệu giao dịch được thu thập thông qua hệ thống POS giúp doanh nghiệp theo dõi doanh số, xu hướng mua sắm và hiệu quả của các chương trình khuyến mãi.

Phân tích dữ liệu giúp doanh nghiệp:

+ Hiểu rõ xu hướng mua sắm của khách hàng: Khách hàng thường mua gì, mua khi nào, tần suất mua hàng ra sao?

+ Đánh giá hiệu quả của các chương trình khuyến mãi: Liệu chương trình có thực sự làm tăng doanh số hay chỉ làm giảm biên lợi nhuận?

+ Tối ưu hóa doanh thu và lợi nhuận: Tìm ra loại sản phẩm có sức hút lớn nhất trong các chương trình khuyến mãi.

+ Dự báo doanh số và lập kế hoạch kinh doanh: Dự đoán doanh số trong tương lai dựa trên các dữ liệu lịch sử.

Hình 1.1 Ảnh minh hoạ

1.1.2. Chương trình khuyến mãi trong kinh doanh siêu thị

Chương trình khuyến mãi là một trong những chiến lược quan trọng của siêu thị để thu hút khách hàng và tăng doanh số bán hàng. Khuyến mãi không chỉ giúp kích cầu mà còn tạo dựng lòng trung thành của khách hàng.

* Các hình thức khuyến mãi phổ biến trong siêu thị

+ Giảm giá trực tiếp: Giảm một phần trăm (%) trên tổng giá trị hóa đơn hoặc trên từng sản phẩm.

+ Mua 1 tặng 1: Khách hàng mua một sản phẩm sẽ nhận được một sản phẩm miễn phí.

+ Tích điểm đổi quà: Khách hàng tích lũy điểm từ các giao dịch để đổi quà hoặc giảm giá cho lần mua tiếp theo.

+ Voucher giảm giá: Cung cấp mã giảm giá hoặc phiếu mua hàng cho khách hàng.

Ảnh hưởng của chương trình khuyến mãi đến doanh thu

+ Tăng doanh số ngắn hạn: Khi có khuyến mãi, khách hàng có xu hướng mua nhiều hơn.

+ Tăng số lượng khách hàng mới: Khuyến mãi có thể thu hút thêm khách hàng mới, đặc biệt là những khách hàng nhạy cảm với giá.

+ Tác động đến hành vi mua hàng: Khách hàng có thể thay đổi lựa chọn sản phẩm dựa trên chương trình khuyến mãi.

+ Rủi ro giảm lợi nhuận: Một số chương trình khuyến mãi có thể làm tăng doanh số nhưng giảm biên lợi nhuận do giá bán thấp

1.2. Lý do chọn đề tài

Trong bối cảnh cạnh tranh khốc liệt giữa các chuỗi siêu thị và sự thay đổi trong thói quen mua sắm của khách hàng, các doanh nghiệp cần liên tục đánh giá và điều chỉnh chiến lược khuyến mãi để tối ưu hóa hiệu quả.

Lý do chọn đề tài này gồm:

+ Nhu cầu tối ưu hóa chiến lược khuyến mãi

+ Các chương trình khuyến mãi khác nhau có thể ảnh hưởng đến doanh thu theo cách khác nhau.

+ Cần xác định loại hình khuyến mãi nào thực sự hiệu quả, tránh lãng phí nguồn lực.

+ Dữ liệu giúp đánh giá mức độ ảnh hưởng của khuyến mãi đến từng danh mục sản phẩm.

1.1.1. Ứng dụng phân tích dữ liệu trong kinh doanh

Dữ liệu giao dịch giúp xác định thói quen mua hàng của khách hàng. Phân tích dữ liệu giúp doanh nghiệp đưa ra quyết định dựa trên dữ liệu thực tế thay vì cảm tính.

1.1.2. Hỗ trợ ra quyết định dựa trên dữ liệu thực tế

Cung cấp thông tin để tối ưu hóa nguồn lực và quản lý hàng tồn kho: Phân tích dữ liệu lịch sử về việc sử dụng nguồn lực để điều chỉnh chiến lược quản lý, đảm bảo tối ưu hóa hiệu suất hoạt động. Đồng thời, theo dõi tình trạng tồn kho theo thời gian thực, giúp tránh tình trạng thiếu hoặc dư thừa hàng hóa.

Hỗ trợ xác định thời gian thích hợp để triển khai các chương trình khuyến mãi hiệu quả nhất: Dựa trên các xu hướng thị trường, dữ liệu về hành vi khách hàng và hiệu suất khuyến mãi trong quá khứ, đề xuất thời điểm và chiến lược phù hợp để tối đa hóa doanh thu. Điều này giúp doanh nghiệp nâng cao khả năng cạnh tranh trên thị trường và đáp ứng nhu cầu khách hàng một cách hiệu quả.

1.3.Mục tiêu

1.3.1. Đánh giá tác động của chương trình khuyến mãi lên doanh số bán hàng

So sánh doanh thu trung bình trong các giai đoạn trước, trong và sau khi thực hiện khuyến mãi để xác định mức độ ảnh hưởng đến tổng doanh thu.

Phân tích số lượng hóa đơn phát sinh trong từng giai đoạn để làm rõ mối quan hệ giữa khuyến mãi và hành vi mua sắm của khách hàng.

1.3.2. Phân tích hành vi mua sắm của khách hàng

Xác định sự thay đổi về giá trị trung bình của đơn hàng: So sánh giá trị trung bình của mỗi đơn hàng trước và sau thời gian khuyến mãi để đánh giá tác động của các chương trình ưu đãi đến hành vi tiêu dùng của khách hàng.

Đánh giá mức độ tham gia của khách hàng mới và khách hàng thân thiết: Phân tích số liệu để xác định tỷ lệ khách hàng mới tham gia trong thời gian khuyến mãi và so sánh với sự đóng góp của khách hàng thân thiết. Điều này giúp xác định hiệu quả của các chương trình khuyến mãi trong việc thu hút và giữ chân khách hàng.

Phân tích phương thức thanh toán nào phổ biến trong thời gian khuyến mãi: Theo dõi và so sánh tần suất sử dụng các phương thức thanh toán như tiền mặt, thẻ tín dụng, chuyển khoản hoặc ví điện tử. Việc này cung cấp dữ liệu để tối ưu hóa trải nghiệm thanh toán cho khách hàng trong tương lai.

1.3.3. So sánh hiệu quả giữa các loại hình khuyến mãi

Việc phân tích hiệu quả của các loại hình khuyến mãi nhằm xác định những chiến lược mang lại doanh thu tốt nhất. Nhóm nghiên cứu tập trung vào việc đánh giá doanh thu tổng quan từ từng loại hình như giảm giá trực tiếp, mua một tặng một, miễn phí vận chuyển, hoặc tặng quà kèm theo. Thông qua các số liệu thu thập được, nhóm có thể nhận định loại khuyến mãi nào thu hút khách hàng tốt nhất và góp phần lớn nhất vào việc tăng doanh thu.

Bên cạnh đó, nhóm còn tiến hành so sánh mức độ tác động của từng loại hình khuyến mãi lên các danh mục sản phẩm khác nhau. Điều này giúp xác định các chương trình phù hợp nhất với từng nhóm sản phẩm, từ đó tối ưu hóa hiệu quả kinh doanh. Ví dụ, một số loại hình khuyến mãi có thể phù hợp hơn cho các sản phẩm thiết yếu, trong khi những chương trình khác lại mang lại kết quả tốt hơn cho các mặt hàng xa xỉ hoặc thời vụ.

=> Việc so sánh không chỉ dựa trên doanh thu, mà còn trên các yếu tố khác như số lượng giao dịch, mức độ tăng trưởng khách hàng mới, và hiệu quả giữ chân khách hàng thân thiết. Kết quả phân tích này cung cấp cơ sở dữ liệu quan trọng để xây dựng các chiến lược khuyến mãi dài hạn và cải thiện trải nghiệm khách hàng.

1.3.4. Dự báo tác động của chương trình khuyến mãi trong tương lai

Ứng dụng mô hình phân tích chuỗi thời gian để dự báo doanh số. Đề xuất chiến lược khuyến mãi tối ưu dựa trên kết quả phân tích.

1.4.Phạm vi nghiên cứu của đề tài

1.4.1. Phạm vi không gian

Phạm vi nghiên cứu tập trung vào dữ liệu được thu thập từ một hệ thống siêu thị cụ thể. Việc này đảm bảo nguồn dữ liệu có tính nhất quán, phù hợp với mục tiêu phân tích. Để cung cấp cái nhìn toàn diện hơn, nghiên cứu bao gồm các chi nhánh ở nhiều khu vực khác nhau, giúp so sánh và đánh giá hiệu quả của các chương trình khuyến mãi theo từng địa điểm.

Bằng cách phân tích sự khác biệt về doanh thu, số lượng giao dịch, và hành vi mua sắm tại các chi nhánh, nghiên cứu mang lại dữ liệu thực tế hữu ích để đưa ra những chiến lược tối ưu hóa khuyến mãi phù hợp với đặc thù từng khu vực. Điều này cũng giúp nhận diện rõ ràng hơn các yếu tố như sức mua của khách hàng, đặc điểm dân cư, hay mức độ cạnh tranh tại các khu vực khác nhau, từ đó đề xuất các giải pháp điều chỉnh linh hoạt và hiệu quả hơn.

1.4.2. Phạm vi thời gian

Dữ liệu được thu thập trong khoảng 6 tháng đến 1 năm, đảm bảo tính đại diện và đủ lớn để phân tích. Nghiên cứu chia thành các giai đoạn: trước, trong và sau khuyến mãi, nhằm đánh giá chính xác tác động theo từng thời điểm.

1.4.3. Phạm vi nội dung

- Dữ liệu đầu vào:

1) Thông tin giao dịch: Bao gồm ID hóa đơn, ngày thực hiện giao dịch, chi nhánh và thành phố nơi giao dịch diễn ra.
2) Thông tin khách hàng: Phân loại khách hàng và phương thức thanh toán mà họ sử dụng.
3) Thông tin sản phẩm: Liệt kê danh mục sản phẩm, số lượng hàng hóa được mua, doanh thu thu về và lợi nhuận đạt được.

- Dữ liệu đầu ra:

    1) Đánh giá hiệu quả chương trình khuyến mãi: Thông qua chỉ số doanh thu và số lượng giao dịch đạt được.

    2) Phân tích hành vi và xu hướng mua sắm: Nhằm hiểu rõ hơn về thói quen tiêu dùng của khách hàng.

    3) So sánh hiệu quả các loại hình khuyến mãi: Xác định loại hình nào đạt được kết quả tốt nhất.

    4) Dự báo doanh số tương lai: Đưa ra dự đoán cho các chương trình khuyến mãi tiếp theo, dựa trên dữ liệu phân tích hiện có.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Trình bày về các khái niệm thống kê

2.1.1. Tổng thể và mẫu nghiên cứu

Trong thống kê, hai khái niệm quan trọng đóng vai trò nền tảng trong việc thu thập và phân tích dữ liệu là tổng thể và mẫu nghiên cứu . Việc hiểu rõ và áp dụng đúng hai khái niệm này giúp đảm bảo quá trình nghiên cứu đạt được kết quả chính xác và đáng tin cậy.

Tổng thể: Là tập hợp tất cả các phần tử có liên quan đến một nghiên cứu.Tổng thể thường có quy mô lớn và chứa rất nhiều điểm dữ liệu, do đó khó có thể phân tích toàn bộ.

Ví dụ:

- Trong bài toán phân tích hiệu quả chương trình khuyến mãi tại siêu thị, tổng thể sẽ bao gồm toàn bộ giao dịch mua hàng tại siêu thị trong khoảng thời gian nghiên cứu. Nếu nghiên cứu tập trung vào năm 2023, tổng thể dữ liệu sẽ bao gồm tất cả các giao dịch mua bán diễn ra trong 12 tháng của năm 2023.
- Nếu nghiên cứu về mức độ hài lòng của khách hàng với một dịch vụ, tổng thể sẽ là toàn bộ khách hàng đã từng sử dụng dịch vụ trong khoảng thời gian xác định
- Tổng thể có thể được chia thành các loại sau:
- Tổng thể hữu hạn : Gồm một số lượng phần tử có thể đếm được. Ví dụ: Tổng số khách hàng đã mua sắm tại một siêu thị trong năm.
- Tổng thể vô hạn: Không thể đếm được hết các phần tử. Ví dụ: Tập hợp tất cả các giao dịch có thể xảy ra trong tương lai.
- Mẫu nghiên cứu là một tập hợp con được chọn từ tổng thể để tiến hành phân tích. Mục tiêu của việc chọn mẫu là thu thập thông tin một cách hiệu quả mà không cần phải phân tích toàn bộ tổng thể, trong khi vẫn đảm bảo tính chính xác khi suy luận.
- Giảm chi phí và thời gian: Việc thu thập và xử lý toàn bộ dữ liệu của tổng thể có thể tốn rất nhiều tài nguyên. Lựa chọn một mẫu nhỏ hơn giúp tối ưu hóa chi phí và tiết kiệm thời gian.

- Dễ dàng quản lý và phân tích: Dữ liệu của tổng thể có thể rất lớn, gây khó khăn cho quá trình lưu trữ, tính toán và phân tích. Một mẫu có kích thước hợp lý giúp công việc phân tích trở nên đơn giản hơn.
- Đảm bảo độ chính xác cao: Nếu mẫu được chọn đúng cách, nó có thể phản ánh chính xác đặc điểm của tổng thể, giúp suy luận thống kê có giá trị.

Ví dụ:

- Trong nghiên cứu về hiệu quả của chương trình khuyến mãi tại siêu thị, thay vì phân tích toàn bộ dữ liệu giao dịch của 12 tháng, ta có thể chọn một mẫu nghiên cứu bao gồm dữ liệu giao dịch trong 6 tháng để phân tích.
- Trong khảo sát mức độ hài lòng của khách hàng, thay vì hỏi ý kiến tất cả khách hàng đã từng mua hàng, ta có thể chọn ngẫu nhiên 1.000 khách hàng để khảo sát.

2.1.2. Các đặc trưng của mẫu và tổng thể

Để mô tả và phân tích dữ liệu, thống kê sử dụng nhiều chỉ số quan trọng như trung bình, trung vị, phương sai, độ lệch chuẩn và hệ số tương quan.

Trung bình: Đây là chỉ số đo lường giá trị trung tâm của dữ liệu, được tính bằng tổng tất cả các giá trị chia cho số lượng quan sát. Ví dụ, nếu doanh thu trung bình mỗi ngày của siêu thị là 4 triệu đồng, điều này có nghĩa là doanh số hàng ngày dao động quanh mức này. Tuy nhiên, trung bình có thể bị ảnh hưởng bởi các giá trị ngoại lệ, ví dụ một ngày doanh thu đột biến cao có thể làm tăng trung bình nhưng không phản ánh đúng thực tế chung.Công thức trung bình mẫu:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Trung vị : Là giá trị chính giữa khi sắp xếp dữ liệu theo thứ tự tăng dần. So với trung bình, trung vị có ưu điểm là không bị ảnh hưởng bởi giá trị quá cao hoặc quá thấp. Nếu doanh thu trong 5 ngày là [1 triệu, 2 triệu, 3 triệu, 4 triệu, 10 triệu], trung bình là 4 triệu, nhưng trung vị là 3 triệu, phản ánh thực tế hơn khi có ngoại lệ.

Phương sai và độ lệch chuẩn: Phương sai đo lường mức độ phân tán của dữ liệu so với trung bình, trong khi độ lệch chuẩn là căn bậc hai của phương sai. Nếu độ lệch chuẩn cao, dữ liệu có sự dao động mạnh. Ví dụ, nếu doanh thu dao động từ 1 triệu đến 10 triệu, độ lệch chuẩn sẽ cao, cho thấy sự biến động lớn.

Công thức tính phương sai tổng thể :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$$

Công thức tính phương sai mẫu ngẫu nhiên:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Độ lệch chuẩn tổng thể:

$$\sigma = \sqrt{\sigma^2}$$

Độ lệch chuẩn mẫu ngẫu nhiên:

$$S = \sqrt{S^2}$$

Hệ số tương quan: Chỉ số này giúp đo lường mối quan hệ giữa hai biến số. Nếu hệ số tương quan gần bằng 1, hai biến có quan hệ tích cực (khuyến mãi càng cao, doanh số càng tăng). Nếu gần -1, hai biến có quan hệ nghịch biến (giảm giá càng nhiều, lợi nhuận có thể giảm theo).

2.2. Trình bày các khái niệm về phân tích dữ liệu

2.2.1. Khái niệm về dữ liệu và phân tích dữ liệu

Dữ liệu có thể được chia thành hai loại chính:

+ *Dữ liệu định lượng* : Đây là các số liệu có thể đo lường, chẳng hạn như doanh thu hàng ngày, số lượng sản phẩm bán ra, hoặc số lần khách hàng quay lại mua sắm. Các giá trị này có thể được xử lý thông qua các phép toán để tính toán xu hướng hoặc dự báo.

Phân loại dữ liệu định lượng:

- Dữ liệu rời rạc): Là dữ liệu chỉ nhận các giá trị nguyên, không thể chia nhỏ hơn nữa.
- Dữ liệu liên tục: Là dữ liệu có thể nhận bất kỳ giá trị nào trong một khoảng liên tục, thường xuất hiện trong các phép đo lường vật lý.

+ *Dữ liệu định tính* : Đây là các thông tin không thể đo lường trực tiếp, chẳng hạn như danh mục sản phẩm, phương thức thanh toán (tiền mặt, thẻ tín dụng, ví điện tử), hoặc loại khách hàng (thành viên hay khách vãng lai). Những dữ liệu này thường được sử dụng để phân loại và tìm ra đặc điểm của nhóm khách hàng.

Phân loại dữ liệu định tính:

+ Dữ liệu danh mục (Nominal Data): Là dữ liệu dùng để phân loại mà không có thứ tự hoặc mức độ hơn kém giữa các giá trị.

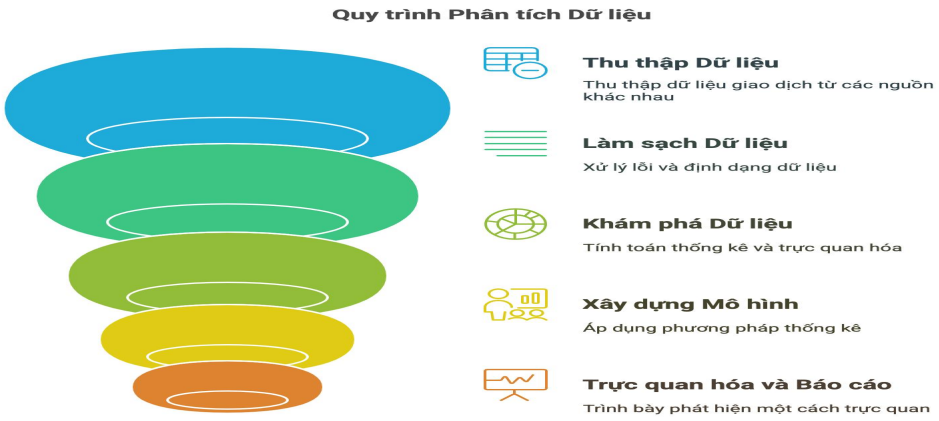
+ Dữ liệu thứ tự (Ordinal Data): Là dữ liệu có thể sắp xếp theo thứ tự, nhưng khoảng cách giữa các giá trị không đồng đều.

Trong bài toán phân tích chương trình khuyến mãi, dữ liệu giao dịch của siêu thị bao gồm cả hai loại trên. Doanh thu và số lượng đơn hàng là dữ liệu định lượng, trong khi danh mục sản phẩm và phương thức thanh toán là dữ liệu định tính. Phân tích dữ liệu sẽ giúp xác định xem chương trình khuyến mãi có làm thay đổi xu hướng mua sắm hay không, từ đó tối ưu hóa chiến lược kinh doanh.

2.3. Quy trình phân tích dữ liệu

Phân tích dữ liệu thường được thực hiện theo một quy trình gồm năm bước quan trọng:

1. Thu thập dữ liệu: Dữ liệu giao dịch được thu thập từ hệ thống POS của siêu thị, bao gồm doanh số bán hàng, danh mục sản phẩm, loại khách hàng, phương thức thanh toán và thời gian giao dịch.
2. Làm sạch dữ liệu: Trước khi phân tích, dữ liệu cần được kiểm tra và xử lý lỗi. Các bước bao gồm loại bỏ dữ liệu trùng lặp, xử lý giá trị bị thiếu và định dạng lại dữ liệu ngày tháng.
3. Khám phá dữ liệu : Giai đoạn này bao gồm việc tính toán các chỉ số thống kê mô tả (trung bình, độ lệch chuẩn) và vẽ biểu đồ để trực quan hóa dữ liệu.
4. Xây dựng mô hình phân tích: Trong bước này, các phương pháp thống kê như hồi quy tuyến tính, kiểm định giả thuyết, hoặc phân tích chuỗi thời gian được áp dụng để tìm ra mối quan hệ giữa chương trình khuyến mãi và doanh số bán hàng. Ví dụ, nếu mô hình hồi quy cho thấy rằng cứ mỗi lần giảm giá 10%, doanh thu tăng 5%, điều đó có thể giúp siêu thị điều chỉnh mức giảm giá phù hợp.
5. Trực quan hóa và báo cáo kết quả: Kết quả phân tích được trình bày dưới dạng biểu đồ hoặc báo cáo để dễ dàng diễn giải. Các doanh nghiệp có thể dựa trên những phát hiện này để điều chỉnh chiến lược khuyến mãi nhằm tối ưu hóa lợi nhuận.



Hình 2.1. Quy trình phân tích dữ liệu

CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU GIAO DỊCH CỦA MỘT SIÊU THỊ ĐỂ ĐÁNH GIÁ HIỆU QUẢ CHƯƠNG TRÌNH KHUYẾN MÃI

3.1. Thu thập dữ liệu

Để đánh giá hiệu quả của chương trình khuyến mãi, chúng tôi sử dụng tập dữ liệu mô phỏng các giao dịch bán hàng tại một siêu thị . Dữ liệu được lưu trữ trong tệp CSV với tổng cộng 500 dòng, bao gồm cả dữ liệu hợp lệ và dữ liệu "bẩn" để phục vụ quá trình làm sạch và phân tích.

Các biến chính trong tập dữ liệu bao gồm:

+ Ngày giao dịch: Thời điểm phát sinh giao dịch (một số dòng bị thiếu).

+ Khuyến mãi: Trạng thái có áp dụng chương trình khuyến mãi hay không, gồm nhiều biến thể như "Có", "Không", "CÓ", "ko", "NaN".

+ Thời gian trong ngày: Khoảng thời gian giao dịch diễn ra, có thể là "Sáng", "Trưa", "Chiều", "Tối", có dòng bị thiếu hoặc lỗi chính tả.

+ Khu vực quầy: Vị trí quầy hàng thực hiện giao dịch, ví dụ: "Quầy 1", "quầy 3", "QUẦY 4" (không đồng nhất cách viết).

+ Ngành hàng: Loại sản phẩm được mua như "Thực phẩm", "Đồ uống", "Hóa mỹ phẩm", "Gia dụng", có cả "gia dụng" viết thường và giá trị thiếu.

+ Phương thức thanh toán: Hình thức thanh toán như "Tiền mặt", "Thẻ tín dụng", "Ví điện tử", "thẻ", có dòng thiếu dữ liệu.

+ Loại khách hàng: Phân loại như "Thành viên", "Thường", "VIP", với một số lỗi viết hoa/thường và giá trị bị thiếu.

+ Số lượng sản phẩm: Tổng số sản phẩm trong mỗi giao dịch, có một số dòng bị thiếu.

+ Doanh thu (nghìn đồng): Giá trị doanh thu ghi nhận cho mỗi giao dịch, bao gồm cả giá trị không hợp lệ như "Không rõ" hoặc thiếu số liệu.

3.2. Làm sạch dữ liệu

Dữ liệu thu thập ban đầu thường có nhiều lỗi như thiếu giá trị, kiểu dữ liệu không đúng hoặc chứa thông tin trùng lặp. Việc làm sạch dữ liệu giúp loại bỏ các lỗi này, đảm bảo dữ liệu có thể sử dụng để phân tích mà không bị sai lệch.

	A	B	C	D	E	F	G	H	I
1	Ngày giao dịch	Khuyến mãi	Thời gian trong ng	Khu vực quầy	Ngành hàng	Phương thức th	Loại khách hàng	Số lượng sản phẩm	Doanh thu (nghìn
2	21/02/2023	CÓ	Sáng	QUẦY 4	gia dụng	nan	VIP	5	315.57
3	15/01/2023	ko	Tối	quầy 3	Đồ uống	thẻ	thường	1	41.11
4	13/03/2023	nan	Sáng	quầy 3	gia dụng	Tiền mặt	VIP	5	178.36
5	02/03/2023	Không	Tối	QUẦY 4	Thực phẩm	thẻ	Thường	4	191.01
6	21/01/2023	CÓ	nan	quầy 3	Hóa mỹ phẩ	thẻ	VIP	5	368.04
7	24/03/2023	có	Sáng	Quầy 1	Đồ uống	thẻ	nan	11	Không rõ
8	28/03/2023	Không	Trưa	quầy 3	Gia dụng	tiền mặt	Thường	2	73.77
9	16/03/2023	có	tối	nan	nan	Chuyển khoản	VIP	12	1040.83
10	16/03/2023	Có	tối	quầy 3	Đồ uống	Ví điện tử	thường	6	401.45
11	29/03/2023	Không	sáng	Quầy 2	Gia dụng	Ví điện tử	thường	3	127.26
12	24/01/2023	CÓ	tối	Quầy 1	nan	Tiền mặt	thường	6	391.15
13	03/01/2023	Không	Chiều	Quầy 2	nan	thẻ	vip	4	140.22
14	22/01/2023	Không	nan	Quầy 2	Đồ uống	Tiền mặt	Thành viên	1	42.92
15	22/02/2023	Không	Tối	nan	gia dụng	tiền mặt	thường	1	42.15
16	02/01/2023	nan	sáng	Quầy 1	gia dụng	tiền mặt	Thường	4	198.9
17	29/03/2023	Không	Chiều	Quầy 1	nan	Thẻ tín dụng	thường	2	70.34
18	30/01/2023	có	Chiều	Quầy 1	gia dụng	tiền mặt	vip	11	573.95
19	07/02/2023	Có	Sáng	Quầy 1	Thực phẩm	Thẻ tín dụng	nan	14	1110.07
20	02/01/2023	ko	tối	Quầy 1	nan	Chuyển khoản	Thường	1	50.58
21									
22	01/03/2023	CÓ	Trưa	quầy 3	Gia dụng	nan	nan	12	750.56
23	21/01/2023	Không	sáng	Quầy 2	gia dụng	Chuyển khoản	Thường	3	120.34
24	02/02/2023	CÓ	nan	Quầy 1	nan	tiền mặt	VIP	14	893.6
25	17/03/2023	có	Trưa	quầy 3	Thực phẩm	Chuyển khoản	Thành viên	10	644.34
26	27/02/2023	CÓ	Tối	quầy 3	Gia dụng	Chuyển khoản	Thành viên	8	589.41
27	22/01/2023	ko	Trưa	Quầy 2	Đồ uống	Chuyển khoản	vip	1	44.85

Hình 3.1.Dữ liệu trước khi làm sạch

Code làm sạch dữ liệu :

```
# Xóa dòng trùng
df.drop_duplicates(inplace=True)

# Xử lý cột 'Doanh thu (nghìn đồng)' có giá trị như 'Không rõ'
df = df[df['Doanh thu (nghìn đồng)'] != 'Không rõ']
df['Doanh thu (nghìn đồng)'] = pd.to_numeric(df['Doanh thu (nghìn đồng)'], errors='coerce')

# Xử lý giá trị thiếu
```

```
num_cols = df.select_dtypes(include=[np.number]).columns
cat_cols = df.select_dtypes(include=['object']).columns

for col in num_cols:
df[col] = df[col].fillna(df[col].median())

for col in cat_cols:
df[col] = df[col].fillna(df[col].mode()[0])

df['Khuyến mãi'] = df['Khuyến mãi'].str.lower().str.strip()
df['Khuyến mãi mã hóa'] = df['Khuyến mãi'].apply(lambda x: 1 if x in ['có', 'có.', 'có khuyến mãi'] else 0)
df['Ngành hàng'] = df['Ngành hàng'].str.lower().str.strip()
df['Khu vực quầy'] = df['Khu vực quầy'].str.lower().str.strip()
```

Ngày giao dịch	Khuyến mãi	Thời gian trong r	Khu vực quầy	Ngành hàng	Phương thức t	Loại khách hà	Số lượng sản	Doanh thu (nghìn	Khuyến mã
21/02/2023	có	Sáng	quầy 4	gia dụng	tiền mặt	VIP	5	315.57	1
15/01/2023	ko	Tối	quầy 3	đồ uống	thẻ	thường	1	41.11	0
13/03/2023	có	Sáng	quầy 3	gia dụng	Tiền mặt	VIP	5	178.36	1
02/03/2023	không	Tối	quầy 4	thực phẩm	thẻ	Thường	4	191.01	0
21/01/2023	có	sáng	quầy 3	hóa mỹ phẩm	thẻ	VIP	5	368.04	1
28/03/2023	không	Trưa	quầy 3	gia dụng	tiền mặt	Thường	2	73.77	0
16/03/2023	có	tối	quầy 2	đồ uống	Chuyển khoản	VIP	12	1040.83	1
16/03/2023	có	tối	quầy 3	đồ uống	Ví điện tử	thường	6	401.45	1
29/03/2023	không	sáng	quầy 2	gia dụng	Ví điện tử	thường	3	127.26	0
24/01/2023	có	tối	quầy 1	đồ uống	Tiền mặt	thường	6	391.15	1
03/01/2023	không	Chiều	quầy 2	đồ uống	thẻ	vip	4	140.22	0
22/01/2023	không	sáng	quầy 2	đồ uống	Tiền mặt	Thành viên	1	42.92	0
22/02/2023	không	Tối	quầy 2	gia dụng	tiền mặt	thường	1	42.15	0
02/01/2023	có	sáng	quầy 1	gia dụng	tiền mặt	Thường	4	198.9	1
29/03/2023	không	Chiều	quầy 1	đồ uống	Thẻ tín dụng	thường	2	70.34	0
30/01/2023	có	Chiều	quầy 1	gia dụng	tiền mặt	vip	11	573.95	1
07/02/2023	có	Sáng	quầy 1	thực phẩm	Thẻ tín dụng	vip	14	1110.07	1
02/01/2023	ko	tối	quầy 1	đồ uống	Chuyển khoản	Thường	1	50.58	0
02/02/2023	có	sáng	quầy 2	đồ uống	tiền mặt	vip	5	333.69500000000	1
01/03/2023	có	Trưa	quầy 3	gia dụng	tiền mặt	vip	12	750.56	1
21/01/2023	không	sáng	quầy 2	gia dụng	Chuyển khoản	Thường	3	120.34	0
02/02/2023	có	sáng	quầy 1	đồ uống	tiền mặt	VIP	14	893.6	1
17/03/2023	có	Trưa	quầy 3	thực phẩm	Chuyển khoản	Thành viên	10	644.34	1
27/02/2023	có	Tối	quầy 3	gia dụng	Chuyển khoản	Thành viên	8	589.41	1
22/01/2023	ko	Trưa	quầy 2	đồ uống	Chuyển khoản	vip	1	44.85	0
30/03/2023	có	sáng	quầy 2	đồ uống	tiền mặt	vip	14	733.46	1

Hình 3.2.Dữ liệu sau khi làm sạch

3.3. Phân tích dữ liệu

- Xác định biến trong mô hình hồi quy

Biến phụ thuộc :

Doanh thu (nghìn đồng)

Đây là biến phản ánh tổng giá trị giao dịch tại siêu thị. Biến này được chọn làm biến phụ thuộc vì nó là yếu tố mà siêu thị muốn phân tích, dự báo và tối ưu hóa.
Việc thay đổi các chính sách như khuyến mãi kỳ vọng sẽ tạo ra sự thay đổi trong doanh thu, do đó doanh thu được mô hình hóa là kết quả đầu ra (output) phụ thuộc vào các yếu tố đầu vào.

Biến độc lập :

Khuyến mãi mã hóa

Là biến nhị phân (giá trị 0 hoặc 1) biểu thị trạng thái có hay không có khuyến mãi trong mỗi giao dịch. Biến này được tạo ra từ cột "Khuyến mãi" sau khi chuẩn hóa văn bản.
Đây là biến độc lập quan trọng nhất trong mô hình, nhằm kiểm định giả thuyết rằng việc áp dụng khuyến mãi có ảnh hưởng tích cực đến doanh thu bán hàng.

Từ đó, mô hình hồi quy tuyến tính đơn giản được xây dựng như sau:

$$\text{Doanh thu} = \beta_0 + \beta_1 * \text{Khuyến mãi mã hoá} + \epsilon$$

Trong đó:

β_0 :Doanh thu trung bình khi không có khuyến mãi.
β_1: Mức tăng doanh thu trung bình khi có khuyến mãi
ϵ: Sai số ngẫu nhiên.

- **Phân phối doanh thu**

```
sns.histplot(df['Doanh thu (nghìn đồng)'], kde=True, bins=30, color='skyblue')
```

\+ Mục tiêu là xem doanh thu có bị lệch phân phối hay không.

+ Kết quả cho thấy doanh thu có phân phối lệch phải, nghĩa là đa số giao dịch rơi vào khoảng 100k–600k đồng, một số ít cao bất thường.

- So sánh doanh thu theo khuyến mãi

```
sns.boxplot(x='Khuyến mãi mã hóa', y='Doanh thu (nghìn đồng)', data=df, palette='Set2')
```

+ Dễ dàng nhận thấy nhóm có khuyến mãi có mức doanh thu cao hơn so với nhóm không khuyến mãi.

+ Biểu đồ hộp cũng cho thấy mức độ dao động trong doanh thu tăng khi có khuyến mãi.

- Ma trận tương quan

```
corr = df[['Doanh thu (nghìn đồng)', 'Số lượng sản phẩm', 'Khuyến mãi mã hóa']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
```

+ Tính toán tương quan giữa các biến định lượng.

+ Doanh thu tương quan mạnh nhất với Số lượng sản phẩm (~0.6).

+ Khuyến mãi mã hóa có tương quan dương (~0.4), khẳng định tác động tích cực đến doanh thu.

- Giao dịch theo ngành hàng

```
sns.countplot(y='Ngành hàng', data=df, order=df['Ngành hàng'].value_counts().index, palette='pastel')
```

+ Giúp xác định ngành hàng được mua nhiều nhất.

+ Các ngành thực phẩm, đồ uống, và gia dụng dẫn đầu.

+ Gợi ý chiến lược: tập trung khuyến mãi vào các ngành có sức mua lớn.

- Hồi quy tuyến tính trực quan

+ Vẽ biểu đồ phân tán kèm đường hồi quy.

+ Xu hướng rõ ràng: nhóm được khuyến mãi có doanh thu cao hơn, đường hồi quy dốc lên → xác nhận giả thuyết ban đầu.

3.3. Hồi Quy Tuyến Tính

Để đánh giá mức độ ảnh hưởng của chương trình khuyến mãi đến hiệu quả kinh doanh, nhóm tiến hành xây dựng mô hình hồi quy tuyến tính đơn biến. Mục tiêu của mô hình là xác định mối quan hệ giữa việc có áp dụng khuyến mãi hay không với giá trị doanh thu trong từng giao dịch.

Trong mô hình này, biến phụ thuộc được lựa chọn là "Doanh thu (nghìn đồng)", đại diện cho kết quả tài chính mà siêu thị thu được từ mỗi lượt mua hàng. Đây là chỉ số quan trọng phản ánh hiệu quả vận hành và là cơ sở để đo lường thành công của các chính sách kích cầu.

Biến độc lập là "Khuyến mãi mã hóa", được mã hóa dưới dạng nhị phân với hai giá trị: 0 nếu giao dịch không áp dụng khuyến mãi và 1 nếu có khuyến mãi. Biến này là đại diện cho yếu tố mà doanh nghiệp có thể chủ động điều chỉnh, từ đó phù hợp để đưa vào phân tích hồi quy.

Mô hình được xây dựng với giả thuyết rằng: chương trình khuyến mãi có thể tạo ra sự khác biệt trong hành vi tiêu dùng, qua đó ảnh hưởng đến doanh thu bán hàng. Hồi quy tuyến tính giúp định lượng mối liên hệ này và kiểm định xem sự thay đổi doanh thu khi có khuyến mãi có mang ý nghĩa thống kê hay không.

Sau khi huấn luyện mô hình với tập dữ liệu đã được xử lý sạch, nhóm sử dụng hai phương pháp phân tích hồi quy: một là thông qua thư viện scikit-learn để nhanh chóng ước lượng hệ số, hai là sử dụng statsmodels để kiểm định giả thuyết thống kê và phân tích chi tiết các tham số như độ tin cậy, mức ý nghĩa và phương sai sai số. Việc sử dụng song song hai công cụ giúp tăng cường độ tin cậy và tính minh bạch trong quá trình xây dựng mô hình.

Kết quả hồi quy tuyến tính từ mô hình là nền tảng quan trọng để đánh giá hiệu quả thực sự của chương trình khuyến mãi, đồng thời làm căn cứ để đưa ra các khuyến nghị chiến lược phù hợp cho siêu thị trong việc tối ưu hóa doanh thu và thu hút khách hàng.

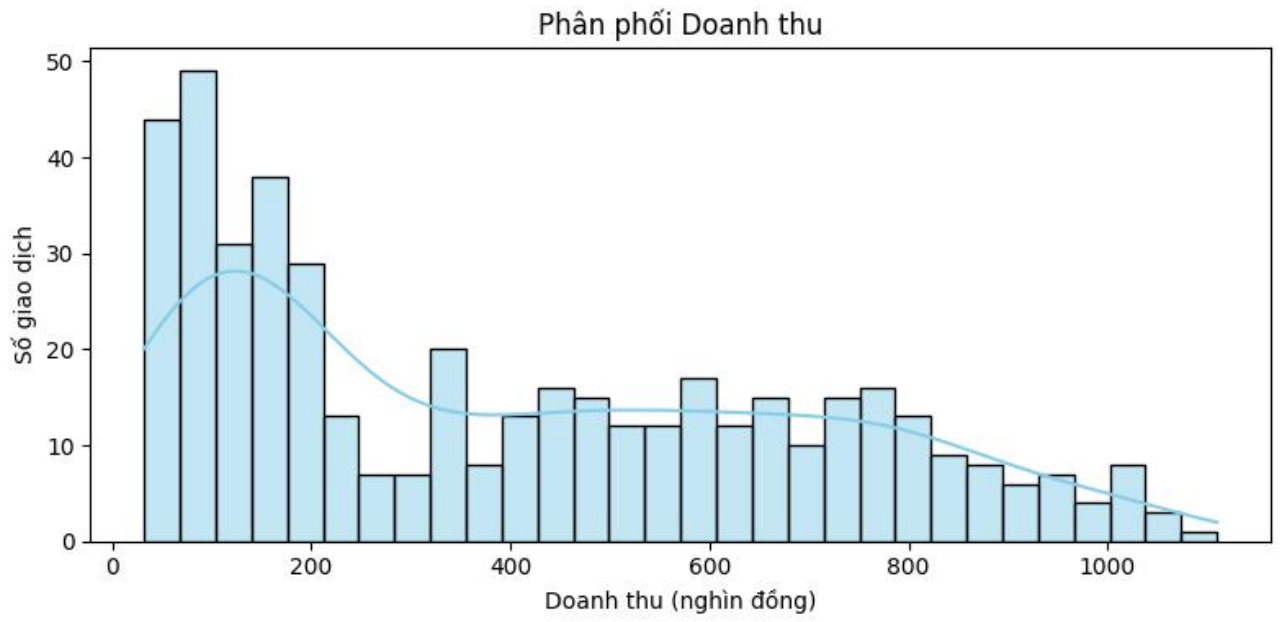
```
X = df[['Khuyến mãi mã hóa']]
y = df['Doanh thu (nghìn đồng)']

# sklearn
model = LinearRegression()
model.fit(X, y)
print("\n📈 Kết quả hồi quy (sklearn):")
print(f"Hệ số hồi quy (beta): {model.coef_[0]:.2f}")
print(f"Intercept: {model.intercept_:.2f}")
print(f"R²: {model.score(X, y):.4f}")

# statsmodels
X_const = sm.add_constant(X)
model_sm = sm.OLS(y, X_const).fit()
print("\n📊 Bảng phân tích hồi quy chi tiết (statsmodels):")
print(model_sm.summary())
```

Hình 3.3. Mã nguồn hồi quy tuyến tính

3.4. Kết quả mô hình



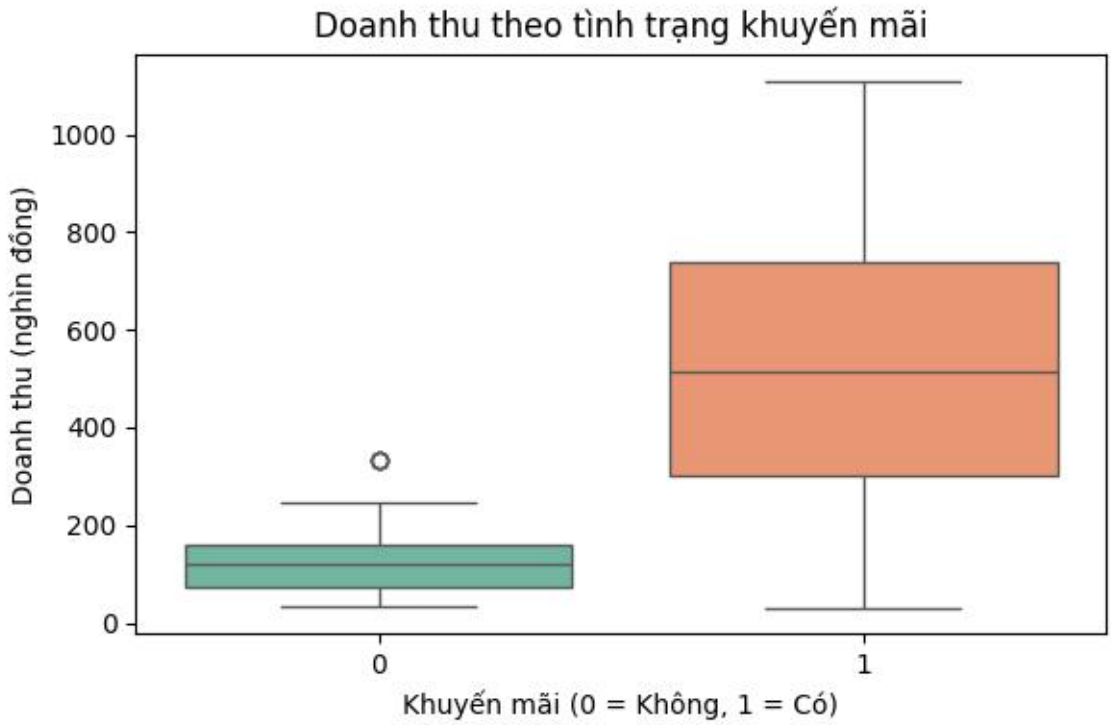
+ *Phân phối doanh thu có dạng lệch phải* , tức là phần lớn các giao dịch tập trung ở mức doanh thu thấp, trong khi số lượng giao dịch có doanh thu cao giảm dần. Điều này là phổ biến trong ngành bán lẻ, khi đa số khách hàng mua số lượng sản phẩm nhỏ, trong khi chỉ có một số ít đơn hàng có giá trị cao.

+ *Mức doanh thu phổ biến nằm trong khoảng từ 50.000 đến 250.000 đồng,* với đỉnh histogram rơi vào khoảng 100.000 – 150.000 đồng. Đây là mức chi tiêu trung bình của đa phần người tiêu dùng, cho thấy xu hướng mua sắm vừa phải, phù hợp với nhu cầu cơ bản.

+ *Tần suất giao dịch giảm rõ rệt từ mức doanh thu 300.000 đồng trở lên,* tuy nhiên vẫn tồn tại một lượng không nhỏ các giao dịch có doanh thu từ 500.000 đến hơn 1 triệu đồng, cho thấy vẫn có một nhóm khách hàng sẵn sàng chi tiêu cao – có thể là khách hàng trung thành, khách mua sỉ hoặc giao dịch tích lũy nhiều mặt hàng.

+ *Không xuất hiện các giá trị cực đoan rõ rệt,* cho thấy dữ liệu doanh thu khá nhất quán và không bị ảnh hưởng bởi các giao dịch bất thường.

+ *Dạng phân phối này cũng hỗ trợ tốt cho phân tích hồi quy tuyến tính*, vì mô hình có thể ước lượng tương quan tuyến tính một cách tương đối chính xác, đặc biệt khi kết hợp với biến số lượng sản phẩm và khuyến mãi.



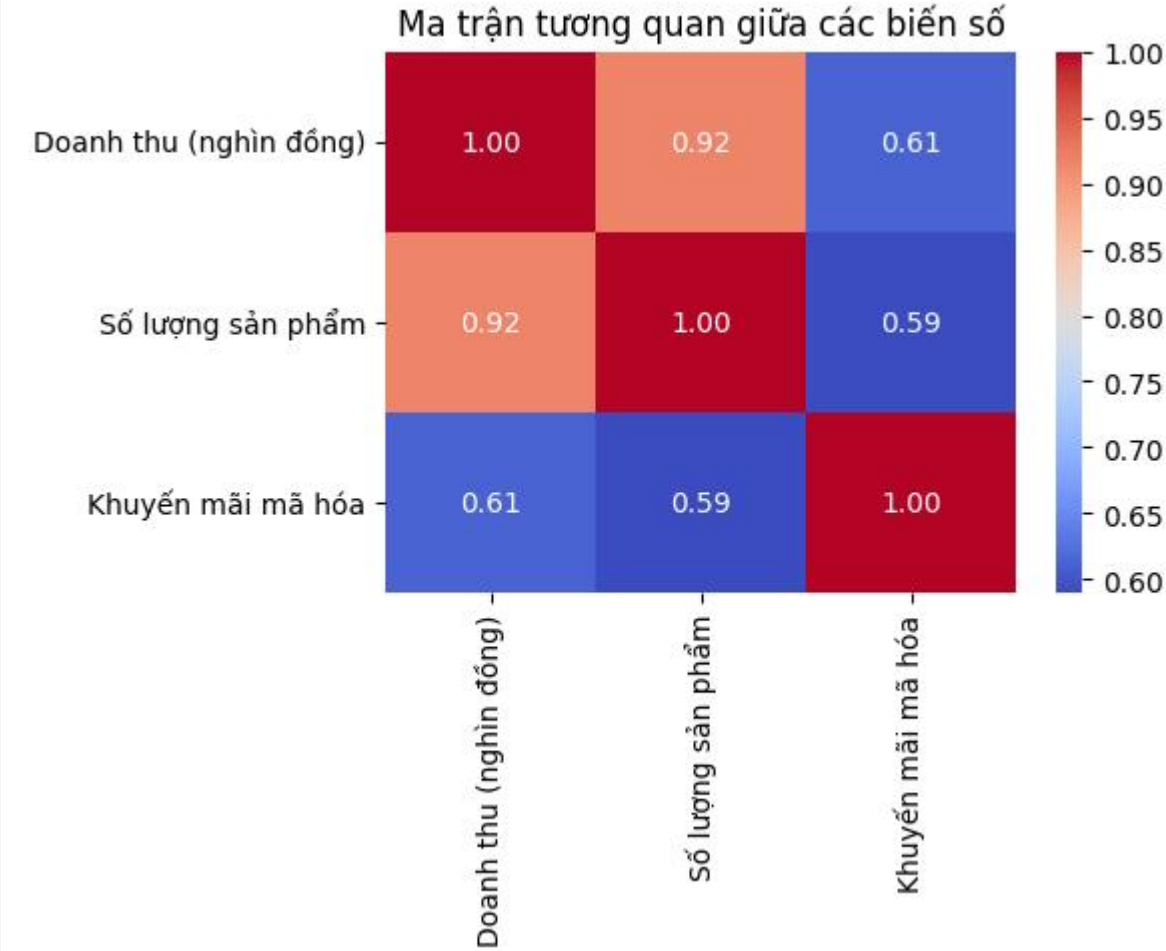
+ *Nhóm có khuyến mãi đạt doanh thu trung bình và trung vị cao hơn đáng kể so với nhóm không có khuyến mãi.* Điều này cho thấy chương trình khuyến mãi có khả năng kích thích hành vi chi tiêu của khách hàng, dẫn đến giá trị đơn hàng lớn hơn.

+ *Độ phân tán của nhóm có khuyến mãi rộng hơn*, phản ánh sự đa dạng trong hành vi mua sắm. Có những khách hàng mua với giá trị rất cao khi có ưu đãi, trong khi số còn lại dao động ở mức trung bình – điều này phù hợp với thực tế các chương trình giảm giá thường thu hút cả khách hàng mua ít và người mua sắm số lượng lớn.

+ *Nhóm không có khuyến mãi có doanh thu thấp và khá đồng đều*, phần lớn các giao dịch tập trung quanh mức thấp hơn 200.000 đồng. Ngoài ra, một số điểm ngoại lai vẫn xuất hiện nhưng không đáng kể.

+ *Khoảng dao động của nhóm khuyến mãi mở rộng lên tới hơn 1 triệu đồng*, trong khi

nhóm không khuyến mãi chỉ dao động đến khoảng 300.000 đồng. Điều này càng khẳng định tác động tích cực của khuyến mãi đến giá trị chi tiêu.



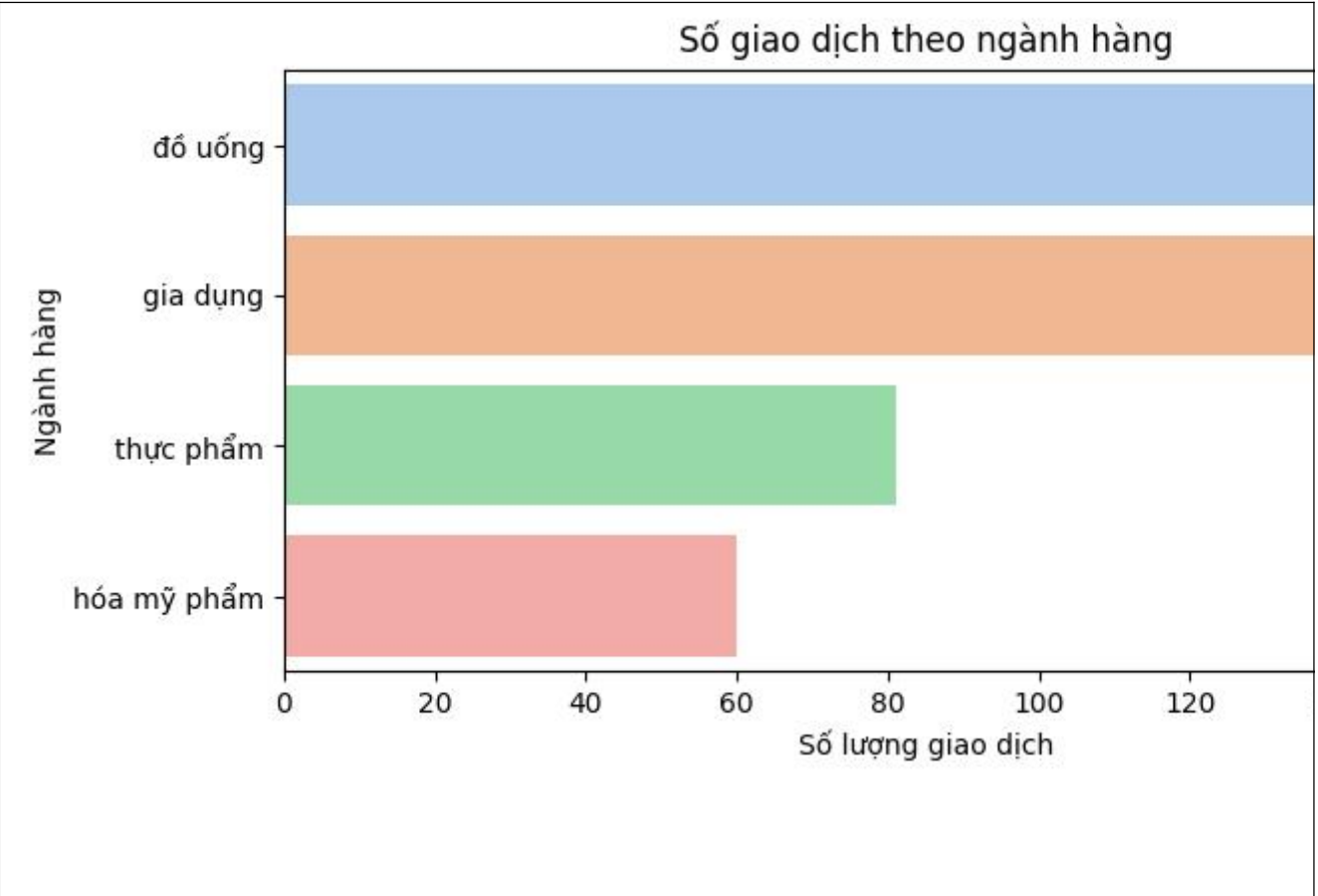
+ *Mối tương quan mạnh giữa "Doanh thu" và "Số lượng sản phẩm" (hệ số tương quan ~ 0.92) cho thấy rằng khi số lượng sản phẩm trong mỗi giao dịch tăng lên thì doanh thu cũng tăng theo tỷ lệ thuận*. Đây là một mối quan hệ dễ hiểu và hợp lý trong bối cảnh bán lẻ, bởi vì tổng doanh thu thường phụ thuộc trực tiếp vào số sản phẩm được mua.

+ *Biến "Khuyến mãi mã hóa" có tương quan trung bình với "Doanh thu" (0.61) và với "Số lượng sản phẩm" (0.59)*. Điều này cho thấy việc áp dụng khuyến mãi có thể không chỉ làm tăng giá trị đơn hàng, mà còn có xu hướng thúc đẩy khách hàng mua nhiều sản phẩm

hơn trong mỗi giao dịch.

+ *Các hệ số tương quan đều mang dấu dương, cho thấy các mối quan hệ cùng chiều*: khi có khuyến mãi → số lượng sản phẩm mua tăng → doanh thu tăng. Điều này hỗ trợ cho giả thuyết rằng chương trình khuyến mãi là một yếu tố tác động tích cực đến hiệu quả kinh doanh.

+ *Tuy nhiên, vì các hệ số tương quan đều < 1*, ta hiểu rằng vẫn còn những yếu tố khác ảnh hưởng đến doanh thu, như thời gian trong ngày, ngành hàng, loại khách hàng, v.v. Điều này mở ra hướng nghiên cứu sâu hơn với các mô hình hồi quy tuyến tính đa biến trong tương lai.



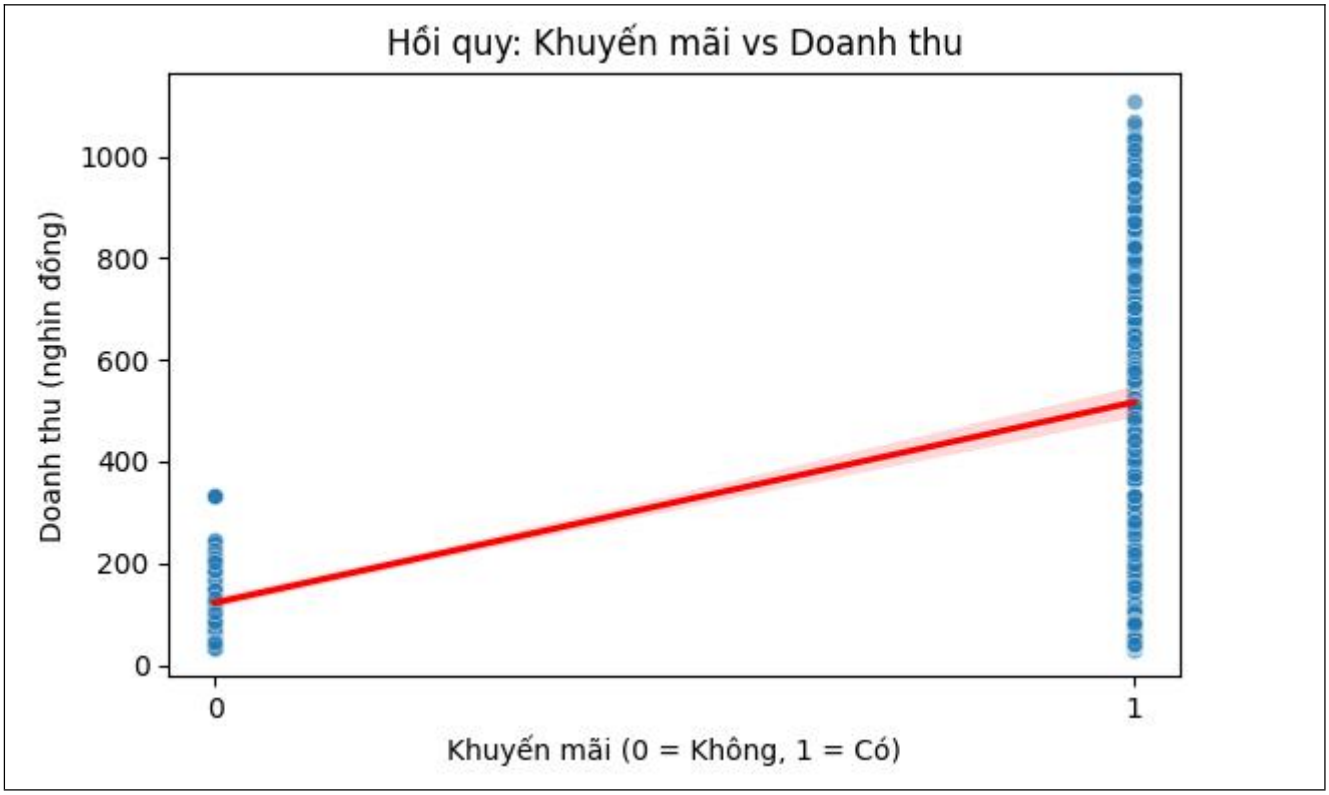
+ *Ngành "Đồ uống" ghi nhận số lượng giao dịch cao nhất,* vượt mốc 160 giao dịch. Điều này phản ánh đặc thù của ngành hàng tiêu dùng nhanh, có nhu cầu ổn định, tần suất

mua sắm cao và dễ tiếp cận với nhiều nhóm khách hàng. Đây cũng là nhóm sản phẩm thường được mua lẻ với giá trị thấp, nhưng tần suất mua cao.

+ *Ngành "Gia dụng" đứng thứ hai về số lượng giao dịch,* với con số chỉ nhỉnh hơn một chút so với "Đồ uống". Sự phổ biến của ngành hàng này có thể đến từ nhu cầu thiết yếu, đặc biệt trong bối cảnh khuyến mãi, người tiêu dùng có xu hướng tích trữ hoặc thay mới các mặt hàng gia dụng.

+ *"Thực phẩm" và "Hóa mỹ phẩm" có số lượng giao dịch thấp hơn rõ rệt,* lần lượt ở khoảng 80 và 60 giao dịch. Mặc dù đây cũng là những nhóm hàng thiết yếu, nhưng có thể do mức độ ưu đãi, chương trình tiếp thị, hoặc đặc tính sử dụng (chu kỳ mua lặp lại dài hơn) khiến chúng ít được lựa chọn hơn trong giai đoạn khảo sát.

+ *Từ góc nhìn chiến lược*, những ngành hàng có số lượng giao dịch thấp nhưng doanh thu trung bình cao (nếu có) vẫn có thể là các phân khúc tiềm năng. Do đó, việc kết hợp phân tích cả số lượng và giá trị giao dịch là rất quan trọng để có cái nhìn toàn diện.



+ *Xu hướng tuyến tính dương rõ rệt:* Đường hồi quy màu đỏ cho thấy xu hướng tăng của doanh thu khi chuyển từ nhóm không khuyến mãi (0) sang nhóm có khuyến mãi (1). Điều này phản ánh rằng việc áp dụng khuyến mãi có tác động tích cực đến giá trị giao dịch.

+ *Doanh thu của nhóm có khuyến mãi cao hơn đáng kể:* Tập dữ liệu cho thấy phần lớn các điểm dữ liệu ở nhóm "1" (có khuyến mãi) có giá trị doanh thu cao hơn so với nhóm "0". Khoảng giá trị của nhóm có khuyến mãi trải dài và bao phủ nhiều mức doanh thu khác nhau, thậm chí có những điểm vượt mức 1.000 nghìn đồng.

+ *Nhóm không khuyến mãi có doanh thu ổn định nhưng thấp:* Các điểm dữ liệu của nhóm "0" (không khuyến mãi) tập trung chặt ở mức doanh thu thấp, cho thấy ít biến động và giá trị trung bình thấp hơn rõ rệt so với nhóm có khuyến mãi.

+ *Ý nghĩa thống kê:* Đường hồi quy đi lên chứng minh mô hình đã tìm được mối quan hệ cùng chiều giữa khuyến mãi và doanh thu. Dù có sự phân tán trong dữ liệu (thể hiện bằng khoảng tin cậy), xu hướng tổng thể vẫn rất rõ ràng.

3.5. Kết luận chung

Phân tích dữ liệu giao dịch siêu thị cho thấy rằng chương trình khuyến mãi có tác động tích cực và rõ rệt đến doanh thu. Các giao dịch có khuyến mãi đạt mức doanh thu trung bình cao hơn đáng kể so với những giao dịch không có ưu đãi. Mô hình hồi quy tuyến tính đã xác nhận mối quan hệ cùng chiều giữa khuyến mãi và doanh thu, với mức ý nghĩa thống kê cao.

Ngoài ra, các ngành hàng như "Đồ uống" và "Gia dụng" ghi nhận số lượng giao dịch lớn, cho thấy tiềm năng tiếp tục mở rộng. Ma trận tương quan cũng chỉ ra rằng không chỉ khuyến mãi, mà số lượng sản phẩm cũng là yếu tố ảnh hưởng mạnh đến doanh thu.

Từ đó, siêu thị nên tiếp tục duy trì và tối ưu hóa các chương trình khuyến mãi, đồng thời phân tích thêm các yếu tố như ngành hàng, phân khúc khách hàng và thời điểm mua sắm để nâng cao hiệu quả kinh doanh trong tương lai.

KẾT LUẬN

Qua quá trình thu thập và phân tích dữ liệu, nhóm đã xác định được các yếu tố quan trọng như mức độ ảnh hưởng của chương trình khuyến mãi đến doanh thu và hành vi mua sắm của khách hàng. Các phương pháp phân tích hiện đại như hồi quy tuyến tính đã được áp dụng để đưa ra những đề xuất chiến lược có giá trị, giúp tối ưu hóa hoạt động kinh doanh trong tương lai.

Bên cạnh đó, nghiên cứu cũng cho thấy một số ưu điểm nổi bật, bao gồm tính khách quan, độ chính xác cao và khả năng cung cấp các giải pháp thiết thực dựa trên dữ liệu thực tế. Tuy nhiên, vẫn còn một số hạn chế như phạm vi nghiên cứu còn hẹp, kỹ năng phân tích dữ liệu chưa sâu, và nguồn lực hạn chế, khiến nhóm chưa thể ứng dụng các mô hình tiên tiến hơn như Machine Learning vào bài báo cáo. Nhìn chung, kết quả nghiên cứu là nền tảng vững chắc để xây dựng và cải thiện chiến lược khuyến mãi trong tương lai, đặc biệt nếu các hạn chế hiện tại được khắc phục.

PHÂN CÔNG NHIỆM VỤ

STT	MSV	Họ và tên	Nhiệm vụ
1	1771020707	Trần Anh Tú	Viết word,thu thập datatset
2	1771020729	Nguyễn Thanh Tùng	Viết code
3	1771020663	Phạm Đức Duy Tiến	Thu thập dữ liệu

DANH MỤC TÀI LIỆU THAM KHẢO

[1]. https://nguyentonhu2000.violet.vn/entry/khai-niem-tong-the-mau-va-chon-mau-11701119.html

[2]. https://fptshop.com.vn/tin-tuc/thu-thuat/cach-tinh-phuong-sai-159197

[3]. https://cellphones.com.vn/sforum/phuong-sai

[4]. https://cellphones.com.vn/sforum/do-lech-chuan-la-gi

[5]. https://mindx.edu.vn/blog/phan-tich-du-lieu-la-gi-quy-trinh-va-vi-du-thuc-te-ve-phan-tich-du-lieu