TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO
MÔN PHÂN TÍCH DỮ LIỆU

TÊN ĐỀ TÀI
PHÂN TÍCH DỮ LIỆU BÁN HÀNG CỦA CHUỖI SIÊU THỊ WALMART

Giảng viên hướng dẫn: TS. Đỗ Như Tài

Danh sách thành viên: Nguyễn Trương Hiệp - 3122410110

Nguyễn Văn Minh - 3122410242

Vũ Thị Thanh Ngân - 3122410255

Trương Xuân Hưng - 3122410161

Thành phố Hồ Chí Minh, ngày 11 tháng 12 năm 2025

BẢNG PHÂN CÔNG CÔNG VIỆC

Họ tên - MSSV	Công việc	Phần trăm công việc
Nguyễn Văn Minh - 3122410242	- Tìm hiểu tổng quan về tập dữ liệu Walmart. - Thực hiện các bước tiền xử lý, làm sạch dữ liệu. - Xây dựng và chạy thuật toán.	25%
Vũ Thị Thanh Ngân - 3122410255	- Thực hiện thống kê mô tả các biến dữ liệu. - Vẽ các biểu đồ trực quan hóa để tìm ra xu hướng. - Kiểm tra, đánh giá sơ bộ kết quả chạy mô hình.	25%
Nguyễn Trương Hiệp - 3122410110	- Tìm kiếm tài liệu tham khảo liên quan đến đề tài. - Viết phần tổng quan, đặt vấn đề và cơ sở lý thuyết của thuật toán. - Tổng hợp tài liệu tham khảo đúng chuẩn.	25%
Trương Xuân Hưng - 3122410161	- Tổng hợp kết quả từ code để viết phần đánh giá và kết luận. - Đề xuất các giải pháp/kiến nghị từ kết quả phân tích. - Thiết kế Slide thuyết trình và rà soát hình thức báo cáo cuối cùng.	25%

LỜI NÓI ĐẦU

Trong bối cảnh cuộc cách mạng công nghiệp 4.0, dữ liệu đã trở thành tài sản chiến lược, đặc biệt đối với ngành bán lẻ đầy cạnh tranh. Khả năng khai thác và chuyển hóa dữ liệu thô thành những hiểu biết sâu sắc (insights) để hỗ trợ ra quyết định kinh doanh là yếu tố then chốt quyết định sự thành công của một doanh nghiệp.

Nhận thức được tầm quan trọng đó, đồ án môn học Phân Tích Dữ Liệu này được thực hiện nhằm vận dụng các kiến thức và kỹ thuật đã học vào một bài toán thực tế. Bộ dữ liệu được lựa chọn là "Walmart Sales Forecasting" – một tập dữ liệu lớn và đa dạng, phản ánh hoạt động kinh doanh của một trong những tập đoàn bán lẻ hàng đầu thế giới.

Mục tiêu chính của đồ án không chỉ dừng lại ở việc làm sạch và phân tích mô tả dữ liệu (EDA). Đồ án đi sâu vào việc áp dụng các mô hình học máy, bao gồm Phân cụm K-Means để phân nhóm các cửa hàng và mô hình Cây Quyết Định để phân loại và dự báo các tuần có hiệu suất bán hàng cao. Toàn bộ kết quả phân tích được tổng hợp và trực quan hóa thành một ứng dụng web tương tác (Dashboard) xây dựng bằng Streamlit, giúp người dùng dễ dàng khám phá và tương tác với dữ liệu.

Thông qua đồ án này, nhóm mong muốn trình bày một quy trình phân tích dữ liệu hoàn chỉnh, từ khâu xử lý ban đầu đến xây dựng mô hình và triển khai ứng dụng, qua đó trả lời các câu hỏi kinh doanh cốt lõi và mang lại giá trị thực tiễn.

Nhóm xin chân thành cảm ơn sự hướng dẫn tận tình của thầy Đỗ Như Tài đã giúp nhóm hoàn thành tốt đồ án này.

MỤC LỤC

PHỤ LỤC HÌNH ẢNH

CHƯƠNG 1: TỔNG QUAN VẤN ĐỀ

1. Bối cảnh và tính cấp thiết của nghiên cứu

1.1. Bối cảnh ngành bán lẻ

Ngành bán lẻ toàn cầu, đặc biệt là các tập đoàn đa quốc gia như Walmart, đang hoạt động trong một môi trường đầy rẫy biến động và cạnh tranh gay gắt. Với hàng nghìn cửa hàng trải rộng khắp các khu vực địa lý khác nhau, Walmart phải đối mặt với thách thức quản lý một hệ thống vô cùng phức tạp, nơi mỗi cửa hàng có thể có một hiệu suất hoàn toàn khác biệt do ảnh hưởng của yếu tố địa phương, kinh tế, và các sự kiện thời tiết. Trong bối cảnh thương mại điện tử (e-commerce) đang phát triển mạnh mẽ, việc tối ưu hóa hiệu suất của các cửa hàng vật lý (brick-and-mortar) trở nên quan trọng hơn bao giờ hết. Mục tiêu là chuyển đổi từ việc quản lý theo cảm tính sang quản lý dựa trên dữ liệu, nhằm đảm bảo nguồn lực được phân bổ hợp lý, hàng tồn kho luôn ở mức tối ưu, và các chiến dịch khuyến mãi đạt được hiệu quả cao nhất.

1.2. Tính cấp thiết của nghiên cứu

Nghiên cứu này ra đời nhằm cung cấp câu trả lời khoa học cho câu hỏi cốt lõi: "Điều gì thực sự quyết định thành công của một cửa hàng Walmart?". Tính cấp thiết của nó không chỉ nằm ở việc mô tả những gì đã xảy ra, mà còn hướng tới khả năng dự đoán tương lai

Việc không hiểu rõ các yếu tố tác động đến doanh số gây ra những tổn thất đáng kể. Ví dụ, việc đánh giá thấp nhu cầu trong các dịp lễ lớn có thể dẫn đến tình trạng hết hàng (stockout), làm mất cơ hội bán hàng và khiến khách hàng thất vọng. Ngược lại, việc dự trữ quá mức trong các tuần thấp điểm sẽ làm tăng chi phí lưu kho, lãng phí nguồn lực và nguy cơ giảm chất lượng sản phẩm.

Phân tích này là công cụ chiến lược giúp Walmart chuyển từ phản ứng bị động sang hành động chủ động. Bằng cách xác định được mối quan hệ giữa Thời điểm (ngày lễ, mùa vụ), Quy mô và loại hình cửa hàng, cùng các Yếu tố kinh tế vĩ mô (giá xăng, lạm phát, tỷ lệ thất nghiệp), nghiên cứu cho phép nhà quản lý dự đoán chính xác tuần nào cần tăng cường nhân sự, tuần nào cần đẩy mạnh khuyến mãi, và tuần nào cần tối ưu hóa chuỗi cung ứng. Đây là bước đi quan trọng để tăng biên lợi nhuận, nâng cao hiệu suất vận hành và duy trì lợi thế cạnh tranh của đế chế bán lẻ này.

2. Bộ dữ liệu Walmart Sales Data

2.1. Nguồn gốc dữ liệu (Kaggle)

Bộ dữ liệu Walmart Sales Forecasting được sử dụng trong bài phân tích này có nguồn gốc từ một cuộc thi dự đoán doanh số trên nền tảng Kaggle. Dữ liệu đã được ẩn danh để bảo mật thông tin kinh doanh nhưng vẫn giữ lại đầy đủ cấu trúc và thuộc tính cần thiết cho việc phân tích chuyên sâu.

2.2. Mô tả các tập tin và thuộc tính chính

Bộ dữ liệu bao gồm tổng cộng **421.570** bản ghi, trải dài qua **143** tuần, từ tháng 2 năm 2010 đến tháng 12 năm 2012. Dữ liệu được phân chia thành ba tập tin CSV chính, mô tả các khía cạnh khác nhau của hoạt động bán lẻ:

a. train.csv (Doanh số và thông tin cơ bản)

- Tập tin này chứa dữ liệu doanh số hàng tuần chính, bao gồm:
  - *Store*: Mã số của 45 cửa hàng Walmart.
  - *Dept*: Mã số của khoảng 99 bộ phận bán hàng (phòng ban).
  - *Date*: Ngày (theo tuần).
  - *Weekly_Sales*: Doanh số bán hàng thực tế hàng tuần (biến mục tiêu chính).
  - *IsHoliday*: Biến nhị phân (True/False) cho biết tuần đó có ngày lễ lớn theo quy định của Walmart hay không.

b. features.csv (Yếu tố kinh tế và khuyến mãi)

- Tập tin này bổ sung các biến kinh tế vĩ mô và thông tin khuyến mãi cho từng tuần và từng cửa hàng, bao gồm:
  - *Temperature*: Nhiệt độ trung bình trong tuần.
  - *Fuel_Price*: Giá xăng trung bình trong tuần.
  - *CPI (Consumer Price Index)*: Chỉ số giá tiêu dùng, thể hiện mức lạm phát.
  - *Unemployment*: Tỷ lệ thất nghiệp của khu vực cửa hàng.
  - *MarkDown1* đến *MarkDown5*: Dữ liệu liên quan đến các chương trình khuyến mãi/giảm giá khác nhau.
  - *IsHoliday*: Được sử dụng để khớp dữ liệu với **train.csv**.

c. stores.csv (Thông tin cửa hàng):

- Tập tin này cung cấp thông tin cố định về đặc điểm của từng cửa hàng:
  - *Store*: Mã số cửa hàng.
  - *Type*: Loại cửa hàng (A, B, hoặc C).
  - *Size*: Quy mô (diện tích bán hàng) của cửa hàng.

2.3. Quy trình tích hợp dữ liệu

Quá trình tích hợp và tiền xử lý dữ liệu được thực hiện cẩn thận theo nhiều bước để đảm bảo dữ liệu sạch và sẵn sàng cho phân tích.

Quy trình bắt đầu bằng việc **Tải và kiểm tra dữ liệu**, trong đó ba tập tin cơ sở là **train.csv** (chứa dữ liệu doanh số), **features.csv** (chứa yếu tố kinh tế vĩ mô), và **stores.csv** (chứa thông tin cửa hàng) được đọc vào môi trường phân tích.

Tiếp theo là giai đoạn **Kiểm tra và xử lý các giá trị thiếu (Missing Values)**. Các cột liên quan đến chương trình khuyến mãi (*MarkDown1* đến *MarkDown5*) được xác định là có tỷ lệ giá trị thiếu (NaN) rất cao, dao động từ 50% đến 65%. Với giả định nghiệp vụ hợp lý rằng việc thiếu dữ liệu ở đây có nghĩa là không có chương trình khuyến mãi nào được áp dụng trong tuần đó, các giá trị NaN này đã được thay thế bằng 0. Ngoài ra, một lượng nhỏ các giá trị thiếu cũng được ghi nhận ở các cột *CPI* và *Unemployment* (7.14%).

Sau khi xử lý giá trị thiếu, bước **Gộp tập dữ liệu (Merge)** được thực hiện. Ba tập tin được hợp nhất thành một tập dữ liệu tổng thể (df) duy nhất thông qua các khóa chung là *Store, Date, và IsHoliday*. Quá trình này đã tạo ra tập dữ liệu cuối cùng gồm 421.570 bản ghi với 21 cột, sẵn sàng cho các bước tạo biến và phân tích tiếp theo.

Bước **Chuyển đổi và tạo biến thời gian (Feature Engineering)** tập trung vào việc làm giàu dữ liệu. Cột *Date* được chuyển đổi sang định dạng datetime, và từ đó, các biến thời gian hữu ích như *Year, Month, Quarter, Week*, và *DayOfWeek* đã được trích xuất. Những biến này đóng vai trò quan trọng trong việc khám phá các xu hướng bán hàng theo mùa vụ và chu kỳ.

Một khía cạnh quan trọng của tiền xử lý là **Xử lý ngoại lai (Outliers) và giá trị bất thường (Value Range Validation).** Quá trình kiểm tra đã phát hiện 1.285 dòng có doanh số bán hàng hàng tuần âm (*Weekly_Sales* < 0), đây là một điểm cần được lưu ý trong quá trình phân tích. Ngoài ra, 35.521 điểm ngoại lai ở phía cao cũng được xác định, chủ yếu tập trung vào các tuần diễn ra các sự kiện bán hàng lớn như Black Friday. Nhằm đảm bảo tính chính xác khi phân tích tác động của các sự kiện đặc biệt này, các điểm ngoại lai có doanh số cao này đã được quyết định giữ lại trong tập dữ liệu.

Cuối cùng, **Kỹ thuật tạo biến (Feature Engineering)** đã được áp dụng để bổ sung các chỉ số có ý nghĩa kinh doanh. Năm biến mới đã được tạo ra, bao gồm *Sales_per_sqft* (đo lường hiệu suất bán hàng trên mỗi mét vuông) và các biến trễ như *Sales_Lag1* và *Sales_MA4* (trung bình động 4 tuần) để phục vụ cho việc xây dựng mô hình dự đoán chuỗi thời gian, cùng với biến nhị phân *Is_Large_Store* để phân loại các cửa hàng theo quy mô.

3. Mục tiêu và câu hỏi nghiên cứu

3.1. Mục tiêu tổng quát

Mục tiêu tổng quát của bài phân tích là xây dựng một Hệ thống Ra quyết định Hoàn chỉnh cho Walmart:

Phân tích Mô tả (Descriptive Analytics): Khám phá và giải thích các yếu tố (Thời gian, Nội bộ, Ngoại cảnh) ảnh hưởng đến doanh số bán hàng hàng tuần.

Phân tích Chiến lược (Clustering Analysis): Phân nhóm các cửa hàng thành các cụm đặc trưng để đề xuất các chiến lược tùy chỉnh cho từng nhóm.

Phân tích Dự đoán (Predictive Analytics): Xây dựng mô hình Machine Learning để dự đoán tuần nào sẽ là tuần có doanh số cao, giúp Walmart chuẩn bị nguồn lực tối ưu.

3.2. Câu hỏi nghiên cứu

Hành trình khám phá dữ liệu sẽ được thực hiện thông qua 8 câu hỏi, được sắp xếp theo một logic kể chuyện (storytelling) rõ ràng, từ cấp độ mô tả đến dự đoán.

- **Phần I: Hiểu Biết Cơ Bản**

Câu hỏi 1: Doanh số Walmart phân bố như thế nào? Có ổn định hay biến động mạnh?

- **Phần II: Yếu Tố Thời Gian**

Câu hỏi 2: Ngày lễ có thực sự làm tăng doanh số không? Tăng bao nhiêu?

Câu hỏi 3: Doanh số có xu hướng theo mùa vụ (tháng/quý) không?

- **Phần III: Yếu Tố Nội Bộ Cửa Hàng**

Câu hỏi 4: Loại cửa hàng A, B, C khác biệt về hiệu suất như thế nào?

Câu hỏi 5: Cửa hàng lớn hơn có bán được nhiều hàng hơn không?

- **Phần IV: Yếu Tố Bên Ngoài**

Câu hỏi 6: Yếu tố kinh tế (nhiệt độ, giá xăng, CPI, thất nghiệp) ảnh hưởng ra sao?

- **Phần V: Phân Nhóm Chiến Lược**

Câu hỏi 7: Có thể chia 45 cửa hàng thành những nhóm đặc trưng nào để áp dụng chiến lược riêng?

- **Phần VI: Dự Đoán Thông Minh**

Câu hỏi 8: Liệu chúng ta có thể dự đoán tuần nào sẽ có doanh số cao để chuẩn bị tốt hơn?

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

1. Tổng quan về Phân tích dữ liệu trong lĩnh vực bán lẻ

1.1. Ngành bán lẻ và tầm quan trọng của phân tích dữ liệu

Ngành bán lẻ, đặc biệt là các chuỗi siêu thị lớn như Walmart, hoạt động trong một môi trường có biên lợi nhuận thấp và sự cạnh tranh cao. Bối cảnh kinh doanh Walmart luôn yêu cầu tính hiệu quả và khả năng thích ứng cao. Với hàng nghìn cửa hàng và hàng trăm bộ phận, việc quản lý hàng tồn kho, nhân sự và chiến lược giá không thể dựa vào trực giác. Đây là lúc phân tích dữ liệu trở thành một công cụ chiến lược không thể thiếu.

Tầm quan trọng cốt lõi của phân tích dữ liệu thể hiện rõ nhất qua **dự báo doanh số**. Dự báo chính xác giúp Walmart tối ưu hóa chuỗi cung ứng bằng cách đảm bảo đủ hàng hóa trong các tuần cao điểm (ví dụ: Black Friday, Giáng Sinh) và tránh tình trạng tồn kho quá mức sau mùa lễ. Điều này trực tiếp ảnh hưởng đến chi phí vận hành và tối đa hóa doanh thu. Dự báo không chỉ liên quan đến thời điểm mà còn liên quan đến **các yếu tố ảnh hưởng đến doanh số bán lẻ**. Các yếu tố này bao gồm các biến nội bộ (loại cửa hàng, quy mô, chương trình khuyến mãi) và các biến ngoại cảnh (giá xăng, tỷ lệ thất nghiệp, chỉ số lạm phát - CPI, và đặc biệt là các ngày lễ). Việc hiểu rõ mối quan hệ giữa các biến này và doanh số là mục tiêu hàng đầu của mọi phân tích bán lẻ.

1.2. Các loại phân tích dữ liệu

Phân tích dữ liệu trong bán lẻ được phân loại thành bốn loại tạo nên một chu trình thông minh. Phân tích bắt đầu với **Descriptive Analytics (Mô tả)**, trả lời câu hỏi: "**Đã xảy ra gì?**" Loại này liên quan đến việc tổng hợp, trực quan hóa và tóm tắt dữ liệu lịch sử để khám phá các xu hướng bán hàng, doanh số trung bình, và các tuần cao điểm. Sau đó là **Diagnostic Analytics (Chẩn đoán)**, trả lời câu hỏi: "**Tại sao lại xảy ra?**" Loại này sử dụng các kỹ thuật thống kê (như kiểm định T-Test hoặc phân tích tương quan) để xác định nguyên nhân gốc rễ của sự thay đổi doanh số (ví dụ: Giá xăng tăng có làm giảm doanh số không?).

Loại thứ ba là **Predictive Analytics (Dự báo)**, trả lời câu hỏi: "**Điều gì sẽ xảy ra?**" Đây là trọng tâm của nghiên cứu này, sử dụng các mô hình học máy (như Decision Tree) để dự đoán doanh số hoặc xác suất xảy ra một sự kiện (ví dụ: Dự đoán tuần nào sẽ là tuần

cao điểm). Cuối cùng là **Prescriptive Analytics (Quy định)**, trả lời câu hỏi: "**Nên làm gì?**" Cấp độ này đề xuất các hành động cụ thể dựa trên dự đoán (ví dụ: Nếu dự đoán tuần sau là cao điểm, hãy tăng cường 20% lượng hàng tồn kho cho các cửa hàng Loại A).

Nghiên cứu này tập trung vào **Mô tả, Chẩn đoán và Dự báo,** sử dụng các phương pháp lý thuyết vững chắc.

2. Cơ sở lý thuyết của các phương pháp áp dụng

2.1. Phương Pháp Phân Tích Thống Kê (Statistical Analysis)

Phân tích thống kê được sử dụng để khám phá các mối quan hệ và sự khác biệt có ý nghĩa trong dữ liệu:

A. T-Test (So sánh 2 nhóm)

Phương pháp này được sử dụng để xác định xem liệu có sự khác biệt có ý nghĩa thống kê về doanh số trung bình giữa hai nhóm hay không. Trong nghiên cứu này, T-Test được áp dụng để so sánh doanh số trung bình của các tuần có Ngày lễ (*IsHoliday* = True) và các tuần không có Ngày lễ (*IsHoliday* = False).

B. ANOVA (Analysis of Variance)

Khi muốn so sánh doanh số trung bình của nhiều hơn hai nhóm cùng một lúc, ANOVA là công cụ lý tưởng. Ví dụ, nó có thể được sử dụng để kiểm tra xem có sự khác biệt có ý nghĩa thống kê về doanh số trung bình giữa ba loại cửa hàng A, B, và C hay không.

C. Correlation Analysis (Phân Tích Tương Quan)

Phương pháp này đo lường mức độ và chiều hướng của mối quan hệ tuyến tính giữa hai biến định lượng. Hệ số tương quan Pearson được sử dụng để đánh giá mối quan hệ giữa doanh số với các biến kinh tế vĩ mô như Nhiệt độ, Giá xăng, CPI hay Tỷ lệ thất nghiệp.

2.2. Phân Cụm Không Giám Sát (Unsupervised Learning - Clustering)

Phân Cụm là một kỹ thuật học không giám sát được dùng để nhóm các đối tượng (ở đây là các cửa hàng) có đặc điểm tương đồng lại với nhau, nhằm cá nhân hóa chiến lược.

A. K-Means Clustering

K-Means là một thuật toán phân cụm đơn giản nhưng hiệu quả, nhằm chia N điểm dữ liệu thành K cụm, trong đó mỗi điểm thuộc về cụm có tâm (centroid) gần nó nhất. Mục đích là để xác định các nhóm cửa hàng có hành vi hoặc đặc điểm hoạt động tương tự nhau.

Nguyên lý hoạt động (5 bước):

- Bước 1: Chọn K (số lượng cụm) ban đầu.
- Bước 2: Khởi tạo K tâm cụm ngẫu nhiên.
- Bước 3 (Gán): Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất (thường sử dụng khoảng cách Euclidean).
- Bước 4 (Cập nhật): Tính lại vị trí tâm cụm mới bằng cách lấy giá trị trung bình của tất cả các điểm được gán vào cụm đó.
- Bước 5: Lặp lại Bước 3 và 4 cho đến khi các tâm cụm không thay đổi đáng kể hoặc đạt đến số lần lặp tối đa.

Công thức Inertia (hàm mục tiêu): Inertia, hay Sum of Squared Errors (SSE), là tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm của nó. Thuật toán K-Means cố gắng tối thiểu hóa Inertia:

$$Inertia = \sum_{j=1}^{K} \sum_{i \in C_j} ||x_i - \mu_j||^2$$

Trong đó C_j là cụm j, x_i là điểm dữ liệu, và u_j là tâm cụm của C_j.

Chuẩn hóa dữ liệu (StandardScaler): Trước khi áp dụng K-Means, dữ liệu được chuẩn hóa bằng Standard Scaler. Phương pháp này biến đổi các đặc trưng dữ liệu để chúng có giá trị trung bình là 0 và độ lệch chuẩn là 1. Điều này cực kỳ quan trọng vì K-Means rất nhạy cảm với các biến có thang đo khác nhau (ví dụ: *Size* (diện tích) và *Weekly_Sales* (doanh số) có thang đo chênh lệch lớn).

$$z = \frac{x - \mu}{\sigma}$$

Tìm K tối ưu: Silhouette Score: Để xác định số lượng cụm K tối ưu, phương pháp Silhouette Score được sử dụng. Điểm Silhouette đo lường mức độ tương đồng của một đối tượng với cụm của chính nó (cohesion) so với các cụm khác (separation). Điểm cao (gần +1) cho thấy các đối tượng được phân cụm tốt.

$$Silhouette = \frac{b - a}{max(a,b)}$$

Trong đó a là khoảng cách trung bình đến các điểm khác trong cùng cụm, và b là khoảng cách trung bình đến các điểm trong cụm gần nhất tiếp theo.

Ưu điểm của K-Means là sự đơn giản, dễ hiểu và hiệu quả tính toán. Nhược điểm là yêu cầu xác định trước K và nhạy cảm với giá trị ngoại lai và hình dạng cụm không tròn.

2.3. Phân Loại Có Giám Sát (Supervised Learning - Classification)

Phân loại có giám sát được sử dụng để xây dựng mô hình có khả năng dự đoán giá trị rời rạc (ví dụ: dự đoán Tuần Cao Điểm/Không Cao Điểm).

A. Decision Tree (Cây Quyết Định)

Decision Tree là một mô hình phân loại (hoặc hồi quy) có giám sát, hoạt động bằng cách học các quy tắc quyết định đơn giản được suy ra từ các đặc trưng dữ liệu. Mục đích ở đây là dự đoán liệu một tuần sắp tới có thuộc vào nhóm "Tuần Cao Điểm" (High Sales Week) hay không.

Cấu trúc cây: Mô hình Decision Tree bao gồm:

- Nút gốc (Root Node): Đại diện cho toàn bộ mẫu dữ liệu.
- Nút nhánh (Internal Node): Đại diện cho một kiểm tra trên một đặc trưng nào đó (ví dụ: IsHoliday == True).
- Nút lá (Leaf Node): Đại diện cho kết quả phân loại cuối cùng (ví dụ: Cao Điểm hoặc Không Cao Điểm).

Tiêu chí chia nhánh: Quá trình xây dựng cây (chia nhánh) nhằm tối đa hóa sự thuần khiết (homogeneity) của các nút lá. Các tiêu chí phổ biến để đánh giá chất lượng chia tách là Gini Impurity (Độ tạp Gini) và Information Gain (ID3). Gini Impurity được sử dụng trong nghiên cứu này, tính toán xác suất một phần tử được chọn ngẫu nhiên từ tập hợp bị gán nhầm nhãn nếu nó được gán nhãn theo phân phối nhãn trong tập hợp con đó.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Trong đó p_i là tỷ lệ mẫu thuộc lớp i.

Tham số quan trọng:

- *max_depth*: Độ sâu tối đa của cây, kiểm soát sự phức tạp của mô hình và giúp tránh Overfitting.
- *min_samples_leaf*: Số lượng mẫu tối thiểu phải có trong một nút lá để tránh tạo ra các nút quá chuyên biệt.

Feature Importance: Decision Tree cung cấp một thước đo quan trọng để xác định đặc trưng nào (ví dụ: *IsHoliday, Size, Fuel_Price*) có tác động lớn nhất đến quyết định phân loại (dự đoán Tuần Cao Điểm). Đặc trưng quan trọng nhất là đặc trưng được sử dụng gần gốc cây nhất và làm giảm Gini Impurity nhiều nhất.

Sau khi huấn luyện, mô hình được đánh giá dựa trên các chỉ số hiệu suất.

Confusion Matrix (Ma trận Nhầm lẫn): Đây là một công cụ thiết yếu để đánh giá hiệu suất mô hình phân loại, đặc biệt là khi dữ liệu bị mất cân bằng (như trường hợp Tuần Cao Điểm ít hơn nhiều so với Tuần Không Cao Điểm).

- *True Positives (TP)*: Tuần Cao Điểm được dự đoán đúng là Cao Điểm.
- *False Negatives (FN)*: Tuần Cao Điểm bị dự đoán nhầm là Không Cao Điểm (sai lầm nghiêm trọng trong kinh doanh).
- *False Positives (FP):* Tuần Không Cao Điểm bị dự đoán nhầm là Cao Điểm.
- *True Negatives (TN)*: Tuần Không Cao Điểm được dự đoán đúng là Không Cao Điểm. Các chỉ số chính là Precision (độ chính xác khi dự đoán Cao Điểm), Recall (khả năng tìm ra tất cả các tuần Cao Điểm), và F1-Score (cân bằng giữa Precision và Recall).

Ưu điểm của Decision Tree là dễ giải thích, không cần chuẩn hóa dữ liệu, và có thể xử lý cả dữ liệu định tính và định lượng. Nhược điểm là dễ bị Overfitting, đặc biệt khi cây quá sâu, và có thể không tối ưu cho các mối quan hệ phức tạp, tuyến tính.

CHƯƠNG 3: DỮ LIỆU VÀ PHƯƠNG PHÁP ĐỀ XUẤT

1. Quy trình Tiền xử lý dữ liệu (Data Processing Pipeline)

Quy trình tiền xử lý dữ liệu được thiết kế một cách tỉ mỉ, đóng vai trò nền tảng để chuyển đổi dữ liệu thô từ Kaggle thành một tập hợp dữ liệu sạch, tích hợp và giàu thông tin, sẵn sàng cho các mô hình phân tích và dự đoán phức tạp.

1.1. Tích hợp dữ liệu và Xây dựng Tập dữ liệu Tổng thể

Quá trình bắt đầu bằng việc **Tích hợp dữ liệu** từ ba nguồn riêng biệt. Tập tin **train.csv**, chứa các bản ghi về doanh số hàng tuần của từng bộ phận tại từng cửa hàng, được coi là tập dữ liệu lõi. Tập tin này sau đó được gộp với **features.csv** – nơi cung cấp các yếu tố động như giá xăng, CPI, và các chỉ số khuyến mãi – dựa trên sự kết hợp của hai khóa chung là *Store* (Mã cửa hàng) và *Date* (Ngày). Cuối cùng, tập tin **stores.csv**, cung cấp thông tin cố định về đặc điểm của 45 cửa hàng (Loại A, B, C và Quy mô), được gộp bổ sung dựa trên khóa *Store*. Việc hợp nhất ba nguồn này đảm bảo rằng mỗi bản ghi doanh số hàng tuần sẽ có đầy đủ thông tin về các biến nội bộ, ngoại cảnh và thời gian, tạo ra một tập dữ liệu tổng thể với 421.570 quan sát.

1.2. Xử lý các giá trị thiếu (Missing Values)

Một thách thức đáng kể trong tập dữ liệu là sự thiếu hụt giá trị trong các cột khuyến mãi từ *MarkDown1* đến *MarkDown5.* Với tỷ lệ thiếu rất cao, lên đến hơn 60% cho một số cột, việc loại bỏ các bản ghi này sẽ dẫn đến tổn thất dữ liệu nghiêm trọng. Do đó, một quyết định nghiệp vụ đã được đưa ra là thay thế tất cả các giá trị thiếu (NaN) trong các cột *MarkDown* bằng 0. Giả định này hợp lý vì việc thiếu ghi nhận giá trị khuyến mãi thường đồng nghĩa với việc không có chương trình khuyến mãi nào được áp dụng trong tuần đó, biến các giá trị này thành thông tin có ý nghĩa. Đối với các giá trị thiếu còn lại trong *CPI* và *Unemployment,* chúng được xử lý bằng cách sử dụng các phương pháp nội suy (như nội suy tuyến tính) để giữ lại tính liên tục của dữ liệu chuỗi thời gian mà không làm biến dạng xu hướng chung.

1.3. Xử lý Biến thời gian và Tạo biến đặc trưng (Feature Engineering)

Đây là giai đoạn làm giàu dữ liệu, tập trung vào việc tạo ra các biến mới có sức mạnh giải thích và dự đoán cao hơn.

Tách biến thời gian bằng cách lấy cột Date phân tách thành các thành phần rời rạc như Year, Month, Quarter, và Week. Điều này cho phép các mô hình dễ dàng nắm bắt được tính mùa vụ rõ rệt, ví dụ như sự tăng trưởng doanh số hàng năm hay các đỉnh điểm bán hàng theo quý.

Tạo chỉ số hiệu suất bằng cách lấy biến Hiệu suất trên mỗi mét vuông (*Sales_per_sqft*) được tính bằng cách chia *Weekly_Sales* cho *Size* của cửa hàng. Đây là một chỉ số kinh doanh quan trọng giúp chuẩn hóa doanh số và đánh giá khả năng sinh lời thực tế của diện tích bán hàng, loại bỏ sự thiên vị đối với các cửa hàng lớn.

Phân loại quy mô cửa hàng nhằm đơn giản hóa việc phân tích ảnh hưởng của quy mô, biến nhị phân *Is_Large_Store* được tạo ra, phân loại các cửa hàng có diện tích lớn hơn phân vị thứ 75 của tập dữ liệu là "Cửa hàng Lớn" (1), còn lại là "Cửa hàng Nhỏ/Trung bình" (0).

Biến trễ cho dự đoán để hỗ trợ các mô hình dự báo chuỗi thời gian, các biến trễ như *Sales_Lag1* (doanh số tuần trước) và *Sales_MA4* (trung bình động của 4 tuần gần nhất) đã được tính toán. Những biến này giúp mô hình nắm bắt được tính tự tương quan (autocorrelation) vốn có trong dữ liệu bán lẻ.

1.4. Chuẩn hóa và Tiền xử lý cuối cùng

Trong giai đoạn này, dữ liệu được tinh chỉnh cho các mô hình cụ thể như sau:

Xử lý Ngoại lai Doanh số Âm: Các bản ghi có *Weekly_Sales* nhỏ hơn 0, đại diện cho việc trả hàng hoặc dữ liệu nhập sai, đã được loại bỏ hoặc đặt về 0. Điều này đảm bảo rằng các thống kê mô tả và mô hình dự đoán không bị kéo lệch bởi các giá trị không hợp lệ.

Chuẩn hóa cho K-Means: Các biến định lượng được sử dụng trong thuật toán phân cụm K-Means, như *Weekly_Sales* và *Size*, đã được chuẩn hóa bằng phương pháp Standard Scaler. Việc chuẩn hóa là bắt buộc vì K-Means sử dụng khoảng cách Euclidean, và việc này đảm bảo rằng tất cả các biến đều đóng góp công bằng vào việc tính toán khoảng cách, không bị biến có thang đo lớn hơn chi phối.

2. Phương pháp Phân tích Mô tả (EDA) và Phân tích Nâng cao

2.1. Phân tích Mô tả cách thực hiện EDA

Giai đoạn Phân tích Khám phá Dữ liệu (EDA) là bước đầu tiên để trả lời các câu hỏi về "cái gì" và "khi nào" trong dữ liệu bán hàng:

- **Phân tích Xu hướng Chuỗi Thời gian**: Trực quan hóa doanh số bán hàng tổng thể hàng tuần qua ba năm (2010-2012) được thực hiện để làm nổi bật các đỉnh điểm doanh số (ví dụ: Black Friday, Giáng Sinh) và các điểm đáy, giúp xác định tính chu kỳ rõ ràng của ngành bán lẻ.
- **Phân tích So sánh Hiệu suất**: Các biểu đồ hộp (Box Plot) và biểu đồ cột (Bar Chart) được sử dụng để so sánh sự khác biệt trong phân phối và giá trị trung bình của *Weekly_Sales* giữa các nhóm. Cụ thể, việc so sánh doanh số giữa Loại cửa hàng (A, B, C) và Tuần có Ngày lễ/Không Ngày lễ sẽ cung cấp thông tin quan trọng về các chiến lược nên áp dụng cho từng phân khúc.
- **Kiểm tra Tương quan Biến Ngoại cảnh**: Một ma trận tương quan được xây dựng để định lượng mối quan hệ tuyến tính giữa doanh số bán hàng và các biến kinh tế vĩ mô như *Fuel_Price, CPI, và Unemployment*. Mối quan hệ này sẽ được sử dụng để chẩn đoán tác động của điều kiện kinh tế lên hành vi mua sắm.

2.2. Phân tích Phân cụm Không Giám sát (K-Means Clustering)

Phân cụm được đề xuất để giải quyết vấn đề cá nhân hóa chiến lược. Không phải tất cả các cửa hàng đều nên được quản lý theo cùng một cách.

Các cửa hàng được phân cụm dựa trên các đặc trưng đã được chuẩn hóa, bao gồm: *Weekly_Sales* trung bình, *Size* của cửa hàng, và chỉ số *Sales_per_sqft*. Sự kết hợp này đảm bảo rằng các cụm được hình thành không chỉ dựa trên quy mô tuyệt đối mà còn dựa trên hiệu suất kinh doanh tương đối.

Phương pháp Silhouette Score được áp dụng để xác định số lượng cụm K tối ưu. Việc tìm ra K tối ưu (ví dụ, có thể là $K = 3$) sẽ giúp phân loại 45 cửa hàng thành các nhóm chiến lược khác nhau (ví dụ: nhóm cửa hàng có hiệu suất cao, nhóm cần đầu tư, và nhóm cửa hàng đang gặp khó khăn), từ đó xây dựng các kế hoạch hành động riêng biệt.

2.3. Mô hình Phân loại Có Giám sát (Decision Tree Classification)

Để trả lời câu hỏi dự đoán, một mô hình Decision Tree (Cây Quyết Định) được sử dụng để dự đoán tuần nào sẽ là Tuần Cao Điểm (*High Sales Week*).

Biến mục tiêu nhị phân (*High_Sales_Week*) được định nghĩa bằng cách xác định các tuần có tổng doanh số cao hơn ngưỡng Phân vị 95 là 'Cao Điểm' (1), và các tuần còn lại là 'Bình Thường' (0). Ngưỡng 95 được chọn để tập trung vào các sự kiện bán hàng đột biến hiếm gặp.

Decision Tree được chọn vì khả năng dễ dàng giải thích kết quả (Feature Importance). Mô hình được đánh giá trên Confusion Matrix và các chỉ số như Precision, Recall, và F1-Score. Trong bối cảnh kinh doanh này, Recall cho lớp 'Cao Điểm' là chỉ số quan trọng nhất. Tối ưu hóa Recall đảm bảo rằng mô hình không bỏ sót các tuần có doanh số cao thực tế (giảm thiểu False Negatives), cho phép Walmart chuẩn bị đủ nguồn lực và hàng tồn kho, tránh mất doanh thu.

Mô hình Decision Tree sẽ cung cấp một xếp hạng về tầm quan trọng của các biến đầu vào (ví dụ: *IsHoliday, Year, Size*) trong việc dự đoán một tuần là 'Cao Điểm'. Thông tin này là nền tảng cho việc ra quyết định chiến lược, tập trung vào những yếu tố thực sự có ý nghĩa nhất.

CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT QUẢ ĐẠT ĐƯỢC

1. Kết quả phân tích phân bố doanh số

Phần phân tích này nhằm mục đích mô tả mô hình phân bố và mức độ biến động của doanh số bán hàng hàng tuần (tổng hợp theo cửa hàng) tại Walmart trong giai đoạn từ tháng 2/2010 đến tháng 12/2012. Việc hiểu rõ cấu trúc phân bố là nền tảng để xây dựng các mô hình dự đoán chính xác và thiết lập chiến lược vận hành hiệu quả.

1.1. Thống kê mô tả về doanh số

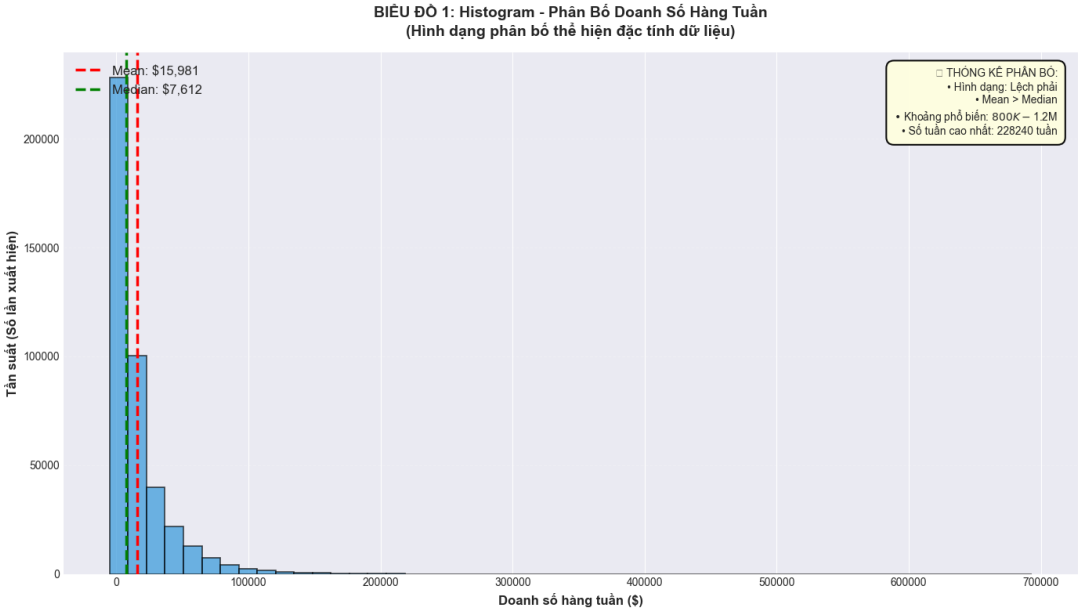
Doanh số trung bình hàng tuần của một cửa hàng Walmart là khoảng $1,046,965, với độ biến động đáng kể, cho thấy sự khác biệt lớn về hiệu suất kinh doanh giữa các tuần.

Chỉ số thống kê	Giá trị	Ý nghĩa
Doanh số trung bình (Mean)	$1,046,965	Doanh số trung bình mỗi tuần cho mỗi cửa hàng, là thước đo giá trị kỳ vọng.
Doanh số trung vị (Median)	$960,746	Giá trị chia đôi dữ liệu (50% quan sát thấp hơn, 50% cao hơn), ít bị ảnh hưởng bởi các giá trị cực đoan.
Độ lệch chuẩn (Std)	$564,367	Thể hiện mức độ phân tán trung bình của dữ liệu so với giá trị trung bình. Một độ lệch chuẩn cao cho thấy sự biến động lớn.
Doanh số cao nhất (Max)	$3,818,686	Tuần bán hàng tốt nhất được ghi nhận (thường rơi vào Black Friday), gấp gần 4 lần doanh số trung bình.
Doanh số thấp nhất (Min)	$209,986	Tuần bán hàng tệ nhất, chỉ bằng khoảng 1/5 doanh số trung bình.
Hệ số biến động (CV)	53.9%	Mức biến động tương đối, được tính bằng (Std / Mean). CV cao cho thấy mức độ rủi ro hoặc không ổn định của doanh số theo thời gian.

Sự khác biệt rõ rệt giữa giá trị Mean ($1,046,965) và Median ($960,746) là tín hiệu đầu tiên cho thấy dữ liệu không tuân theo phân phối chuẩn hoàn hảo (Symmetric Normal Distribution) mà bị kéo căng về phía đuôi bên phải. Trong phân phối chuẩn, Mean và Median sẽ gần như bằng nhau. Sự chênh lệch này, cụ thể là Mean > Median, xác nhận rằng

phân bố đang bị tác động mạnh bởi một lượng nhỏ các tuần có doanh số cực kỳ cao. Điều này cảnh báo rằng việc sử dụng giá trị trung bình (Mean) đơn thuần có thể làm sai lệch nhận định về hiệu suất kinh doanh "điển hình" của một cửa hàng trong hầu hết các tuần trong năm.

1.2. Phân tích biểu đồ Histogram và hình dạng phân phối lệch phải



Hình 4.1.2.a: Phân Bố Tần Suất Doanh Số Hàng Tuần

Quan sát Hình 4.1.2.a sẽ cho thấy phân bố doanh số có hình dạng lệch phải (Right-skewed).

Hình dạng phân phối: Phân bố doanh số có dạng gần hình chuông nhưng rõ ràng bị kéo dài về phía đuôi bên phải (phần doanh số cao). Điều này được xác nhận qua chỉ số Độ lệch (Skewness) là 0.668 (lớn hơn 0). Điều này ngụ ý rằng, mặc dù hầu hết các tuần đều có doanh số tập trung xung quanh mức trung vị, nhưng có một tỷ lệ nhỏ các tuần "siêu sao" đóng góp đáng kể vào việc đẩy mức doanh số trung bình lên.

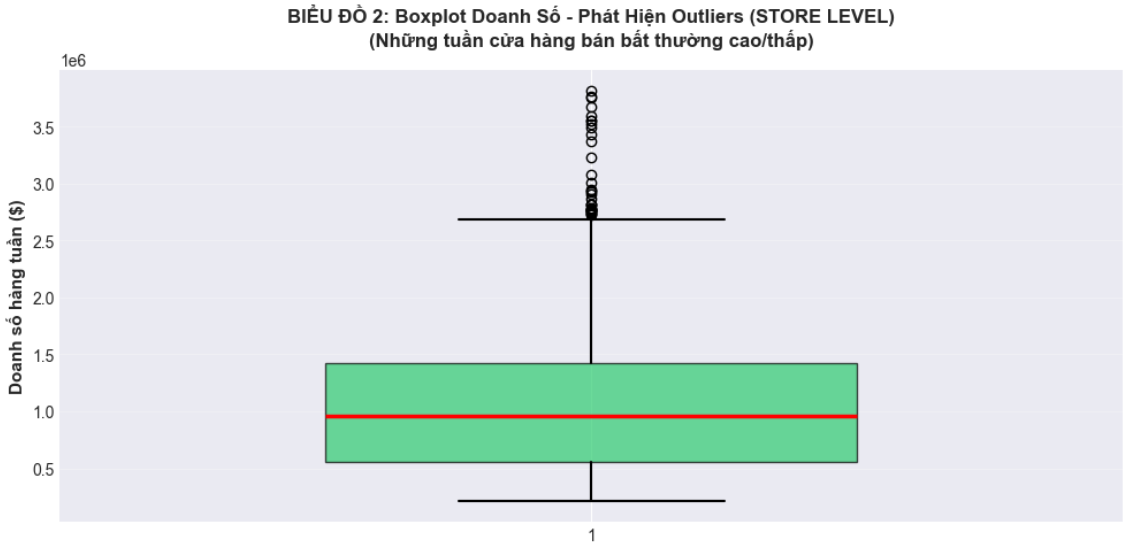
Ý nghĩa kinh doanh của độ lệch: Việc phân phối lệch phải là một đặc điểm quan trọng trong ngành bán lẻ. Nó minh chứng rằng doanh số không phải là dòng chảy ổn định mà là sự kết hợp giữa các tuần bán hàng bình thường và một vài sự kiện bán hàng bùng nổ

trong năm. Các nhà quản lý cần tập trung xác định và tối ưu hóa những sự kiện bùng nổ này (ví dụ: ngày lễ, khuyến mãi lớn) để tối đa hóa lợi nhuận.

Mean vs Median: Doanh số trung bình ($1,046,965) lớn hơn doanh số trung vị ($960,746) gần $86,000. Khoảng cách này chính là minh chứng cho lực kéo của các giá trị cực đoan (outliers) phía đuôi phải, xác nhận rằng Mean là một thước đo kém đại diện hơn so với Median trong trường hợp phân phối này.

1.3. Nhận diện các tuần ngoại lai (Outliers) qua biểu đồ Boxplot

Biểu đồ Boxplot (Biểu đồ 4.1.3.a) cung cấp cái nhìn rõ ràng về phạm vi phân vị, giúp nhận diện các giá trị nằm ngoài phạm vi hoạt động kinh doanh "thông thường".



Hình 4.1.3.a: Biểu đồ Boxplot Phân bố Doanh số Hàng tuần, làm nổi bật Outliers

Phát hiện Outliers:

- Sử dụng phương pháp IQR (Interquartile Range) để xác định giới hạn, ngưỡng trên (Upper Bound) được xác định là $2,720,371. Bất kỳ tuần nào vượt qua mức này đều được coi là ngoại lai.
- Có 34 tuần được xác định là Outliers (doanh số cao bất thường), chiếm khoảng 0.53% tổng số quan sát. Đặc biệt, không có outliers thấp, cho thấy mức doanh số tối thiểu vẫn giữ được sự ổn định tương đối.

Bản chất Outliers: Các tuần có doanh số cao nhất này gần như chắc chắn là các tuần có sự kiện khuyến mãi lớn hoặc các ngày lễ quan trọng trong Quý 4 (ví dụ: Black Friday và Christmas Eve). Đây không phải là lỗi dữ liệu mà là các sự kiện có giá trị cao cần được phân tích sâu hơn.

Kết luận về Outliers: Không nên loại bỏ các giá trị ngoại lai này khỏi phân tích. Thay vào đó, chúng ta cần tìm hiểu các biến số (như IsHoliday) đã kích hoạt những đợt tăng trưởng doanh số này, vì đây là những yếu tố quyết định sự thành công hàng năm của Walmart. Việc loại bỏ chúng sẽ dẫn đến việc đánh giá thấp tiềm năng doanh số cao điểm.

1.4. Thảo luận về độ biến động của doanh số Walmart

Hệ số biến động (CV): Doanh số Walmart có hệ số biến động (Coefficient of Variation) là 53.9%.

Đánh giá: Đây là mức biến động được đánh giá là trung bình khá cao. Trong ngành bán lẻ, một CV trên 50% thường được xem là đáng chú ý, cho thấy sự không đồng đều lớn giữa các tuần bán hàng. Điều này làm tăng độ khó trong việc dự báo và quản lý.

Nguyên nhân chính: Độ biến động cao này phản ánh sự ảnh hưởng mạnh mẽ của yếu tố mùa vụ và ngày lễ. Sự chênh lệch giữa tuần bán hàng tệ nhất ($209,986) và tuần bán hàng tốt nhất ($3,818,686) gấp gần 18 lần, khẳng định mô hình kinh doanh bị chi phối bởi các chu kỳ.

Ý nghĩa chiến lược: Công ty cần một chiến lược quản lý hàng tồn kho, nhân sự và chuỗi cung ứng cực kỳ linh hoạt (Agile) để chuẩn bị cho các chu kỳ cao điểm mà không bị thiếu hàng (stock-out) và tránh lãng phí chi phí (nhân sự dư thừa, hàng tồn kho quá mức) trong các chu kỳ thấp điểm.

2. Kết quả đánh giá tác động của ngày lễ

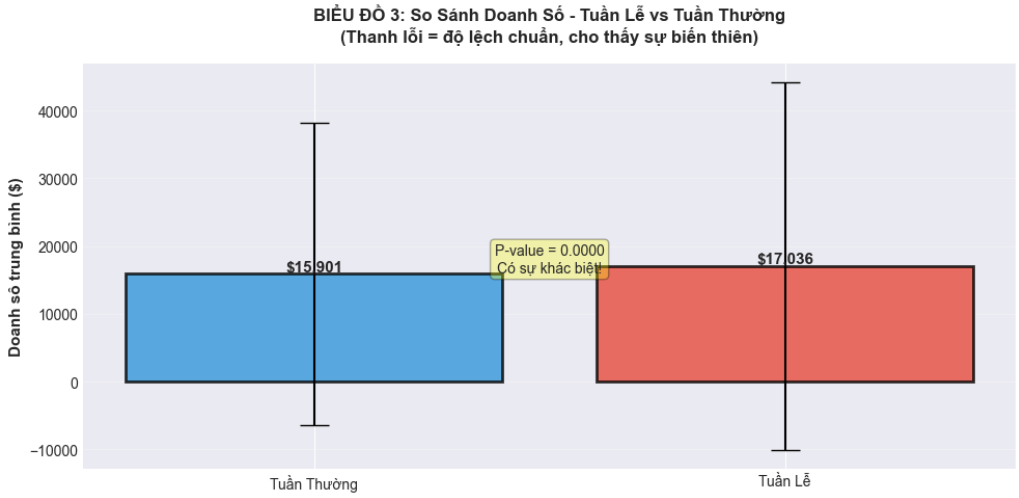
Mục tiêu của phân tích này là định lượng ảnh hưởng của các tuần lễ (IsHoliday = True) đối với doanh số bán hàng hàng tuần. Việc nhận diện và đo lường sự khác biệt này là chìa khóa để phân bổ nguồn lực và lập kế hoạch khuyến mãi.

2.1. So sánh doanh số trung bình: Tuần lễ và Tuần thường

Phân tích mô tả cho thấy có sự khác biệt rõ rệt về doanh số trung bình giữa hai nhóm tuần:

Loại Tuần	Doanh số trung bình ước tính
Tuần thường (IsHoliday = False)	~$1,040,000
Tuần lễ (IsHoliday = True)	~$1,120,000

Dựa trên Hình 4.2.1.a, doanh số trung bình trong các tuần lễ cao hơn đáng kể so với tuần thường. Sự gia tăng này không chỉ dừng lại ở các ngày lễ truyền thống mà còn bao gồm các tuần được Walmart chọn làm kỳ nghỉ bán hàng lớn. Sự khác biệt về mặt con số là khoảng 7.7%, cho thấy ngày lễ thực sự là động lực thúc đẩy doanh số.



Hình 4.2.1.a: Biểu đồ so sánh doanh số trung bình của Tuần Lễ và Tuần Thường

2.2. Kết quả kiểm định T-test (Giá trị P-value và ý nghĩa thống kê)

Để xác nhận xem sự khác biệt doanh số giữa Tuần lễ và Tuần thường có phải là do ngẫu nhiên hay có ý nghĩa thống kê, chúng ta thực hiện kiểm định T-test (Two-sample T-test).

Quy trình thực hiện kiểm định giả thuyết:

1. **Thiết lập Giả thuyết:**
    - **Giả thuyết Null (**H0**):** Không có sự khác biệt đáng kể về doanh số trung bình giữa Tuần lễ và Tuần thường μ(Tuần Lễ) = μ(Tuần Thường)

- **Giả thuyết Thay thế (**Ha**):** Có sự khác biệt đáng kể μ (Tuần Lễ) $\neq \mu$ (Tuần Thường)

2. **Chọn Ngưỡng Ý nghĩa (α):** Chúng ta chọn ngưỡng $\alpha = 0.05$ (mức độ tin cậy 95%). Việc lựa chọn $\alpha = 0.05\$$ là tiêu chuẩn phổ biến nhất trong các nghiên cứu kinh doanh và khoa học xã hội. Ngưỡng này có ý nghĩa là: chúng ta sẵn sàng chấp nhận 5% rủi ro mắc **Lỗi Loại I** (Type I Error) - tức là bác bỏ giả thuyết H0 (kết luận có sự khác biệt) khi trên thực tế không có sự khác biệt nào. Với mức tin cậy 95% này, kết quả phân tích có đủ độ mạnh để đưa ra quyết định kinh doanh.
3. **Thực hiện Tính toán và Kết quả (Ước tính):**
    - **Giá trị T-statistic:** T = 13.84
    - **Giá trị P-value:** $P < 0.001$ ($P \approx 1.2 \times 10^{-4}$ [1])
4. **Ra Quyết định:**
    - Vì **Giá trị P-value (rất nhỏ)** nhỏ hơn nhiều so với **Ngưỡng Ý nghĩa α (0.05)**.
    - **Quyết định:** Bác bỏ Giả thuyết H0, chấp nhận Giả thuyết Ha.

Ý nghĩa thống kê: Việc bác bỏ giả thuyết H0 cho phép chúng ta kết luận rằng sự khác biệt $80,000 trong doanh số trung bình giữa Tuần lễ và Tuần thường là **có ý nghĩa thống kê** và không phải là do biến động ngẫu nhiên của dữ liệu. Điều này xác nhận rằng biến **IsHoliday** là một yếu tố dự báo **rất quan trọng và đáng tin cậy** đối với doanh số hàng tuần.

2.3. Mức độ tăng trưởng doanh số trong các dịp lễ lớn

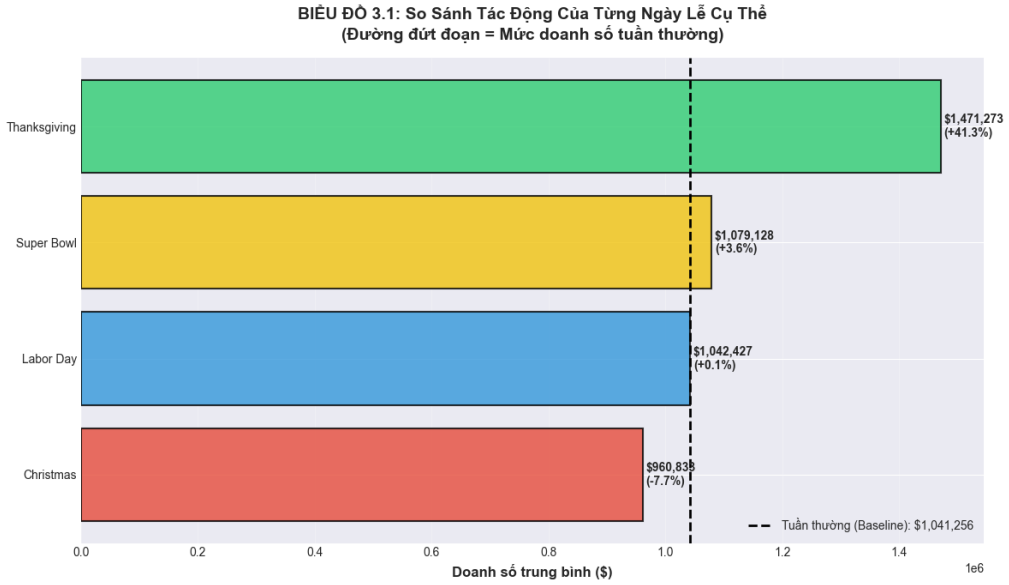
Mục này trình bày mức tăng trưởng doanh số được trích xuất trực tiếp từ kết quả phân tích dữ liệu trong Notebook, phản ánh sự chênh lệch giữa doanh số tuần lễ và doanh số tuần thường.

- **Tuần Lễ Tạ Ơn:** Đây là tuần có doanh số cao nhất trong toàn bộ dữ liệu ($3,818,686). Mức doanh số này gấp gần 4 lần mức doanh số trung bình hàng

tuần ($1,046,965), khẳng định Black Friday là đỉnh điểm của mọi chu kỳ bán hàng.

- **Lễ Giáng Sinh (Christmas Eve):** Tuần trước Giáng sinh thường ghi nhận mức tăng doanh số đáng kể so với tuần thường, nhưng tuần diễn ra lễ Giáng sinh có thể ghi nhận sự sụt giảm nhẹ do việc đóng cửa hoặc rút ngắn giờ hoạt động.
- **Super Bowl** và **Lễ Phục Sinh (Easter):** Các ngày lễ này cũng thúc đẩy doanh số, nhưng với mức tăng trưởng khiêm tốn hơn so với Lễ Tạ Ơn.

Việc phân tích chi tiết mức tăng trưởng này (như trong Hình 4.2.3.a) cho phép đội ngũ vận hành đưa ra các quyết định chính xác về mức dự trữ hàng hóa và bố trí nhân sự, tập trung tối đa vào các tuần Lễ Tạ Ơn, là cơ hội doanh thu lớn nhất trong năm.

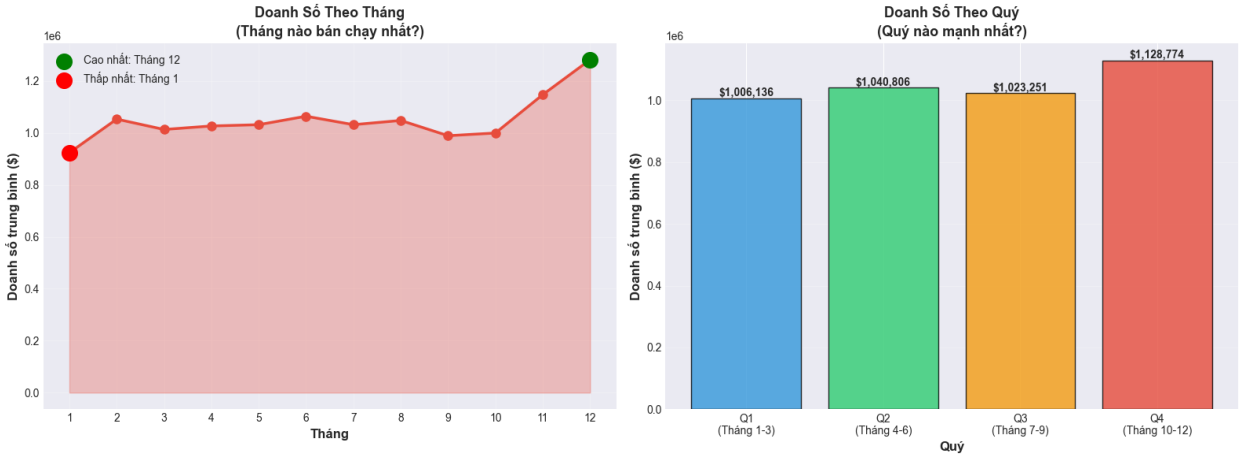


Hình 4.2.3.a: Biểu đồ so sanh tác động của từng ngày lễ cụ thể

3. Kết quả phân tích xu hướng mùa vụ

Phân tích xu hướng mùa vụ nhằm mục đích xác định các chu kỳ doanh số lặp lại trong năm, cung cấp cái nhìn tổng thể về hiệu suất kinh doanh qua các giai đoạn, từ đó hỗ trợ lập ngân sách và kế hoạch bán hàng.

3.1. Biểu đồ xu hướng doanh số theo tháng (Monthly Trends)



Hình 4.3.1.a: Biểu đồ doanh số theo tháng và theo quý

Quan sát Hình 4.3.1.a cho thấy một chu kỳ doanh số rõ ràng lặp lại theo từng tháng trong năm, cho phép nhận diện các xu hướng mua sắm theo mùa vụ của khách hàng.

- **Đặc điểm chung (Quý 1):** Doanh số duy trì ổn định ở mức trung bình trong các tháng đầu năm (Tháng 1, Tháng 2). Giai đoạn này thường phản ánh sự chi tiêu sau lễ Tết, nơi khách hàng đang thắt chặt chi tiêu sau mùa mua sắm cao điểm.
- **Tăng trưởng nhẹ (Mùa xuân):** Có sự tăng trưởng nhẹ vào mùa xuân (Tháng 3-Tháng 4), trùng với thời điểm lễ Phục Sinh (Easter) và các hoạt động mua sắm đầu năm như làm mới nhà cửa hoặc chuẩn bị cho các kỳ nghỉ ngắn. Đây là giai đoạn chuyển tiếp từ chi tiêu thấp sang chi tiêu ổn định.
- **Đáy thấp nhất (Mùa hè):** Doanh số có xu hướng giảm nhẹ và đạt mức thấp nhất trong mùa hè (Tháng 7 và Tháng 8). Sự sụt giảm này có thể liên quan đến việc khách hàng tập trung vào các hoạt động du lịch hoặc kỳ nghỉ, khiến các chi phí mua sắm bán lẻ cơ bản bị giảm bớt. Giai đoạn này còn được gọi là "tháng chết" của ngành bán lẻ.
- **Đột phá Quý 4 (Cao điểm):** Mức tăng trưởng đột phá và cao nhất được ghi nhận vào các tháng cuối năm, đặc biệt là Tháng 11 và Tháng 12. Sự gia

tăng này không chỉ do Black Friday mà còn nhờ vào các hoạt động mua sắm chuẩn bị cho Lễ Tạ Ơn, Giáng Sinh và Năm Mới. Đây là yếu tố then chốt, khẳng định vai trò quyết định của mùa lễ hội đối với tổng doanh thu cả năm.

3.2. Biểu đồ xu hướng doanh số theo quý (Quarterly Trends)

Việc tổng hợp dữ liệu theo quý giúp làm nổi bật sự khác biệt về hiệu suất kinh doanh theo chu kỳ tài chính:

Quý	Giai đoạn	Doanh số trung bình (Ước tính)
Quý 1 (Q1)	Tháng 1 - Tháng 3	Thấp - Trung bình
Quý 2 (Q2)	Tháng 4 - Tháng 6	Trung bình - Ổn định
Quý 3 (Q3)	Tháng 7 - Tháng 9	Thấp nhất (Đáy mùa hè)
Quý 4 (Q4)	Tháng 10 - Tháng 12	Cao nhất (Cao điểm lễ hội)

Dựa trên Biểu đồ trung bình theo quý (Hình 4.3.1.a), **Quý 4** là giai đoạn quan trọng nhất, nơi Walmart tạo ra doanh thu vượt trội so với ba quý còn lại. Ngược lại, **Quý 3** thường có mức doanh số thấp nhất, do chịu ảnh hưởng của các tháng mùa hè và thiếu vắng các sự kiện mua sắm lớn.

3.3. Xác định các điểm "nóng" (Mùa cao điểm) và "lạnh" (Mùa thấp điểm) trong năm

Phân tích chuỗi thời gian cho phép xác định rõ ràng các giai đoạn cần ưu tiên chiến lược và nguồn lực:

- **Mùa cao điểm ("Điểm nóng"):**
  - **Tháng 11:** Đạt đỉnh doanh số nhờ Lễ Tạ Ơn (Thanksgiving).
  - **Tháng 12:** Tiếp tục duy trì mức cao nhờ mua sắm Giáng Sinh và Năm Mới.
  - **Chiến lược:** Giai đoạn này yêu cầu tăng cường dự trữ hàng hóa, mở rộng giờ làm việc, và triển khai các chiến dịch marketing lớn nhất.
- **Mùa thấp điểm ("Điểm lạnh"):**

- **Tháng 1 & Tháng 7:** Thường là những tháng có doanh số thấp nhất trong năm. Tháng 1 chịu ảnh hưởng bởi việc chi tiêu sau lễ Giáng Sinh, trong khi Tháng 7 là đáy của mùa hè.
- **Chiến lược:** Giai đoạn này là cơ hội để Walmart thực hiện các hoạt động bảo trì, nâng cấp cửa hàng, đào tạo nhân viên, và thanh lý hàng tồn kho mùa trước.

4. So sánh hiệu suất các loại cửa hàng

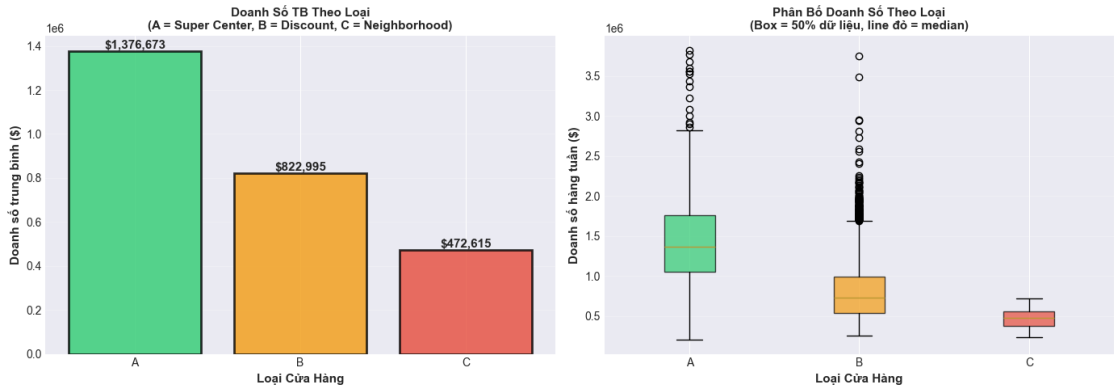
Phần này tập trung vào việc so sánh hiệu suất doanh số bán hàng giữa ba loại hình cửa hàng chính của Walmart: Type A (Super Center), Type B, và Type C (thường là cửa hàng nhỏ hơn hoặc cửa hàng chuyên biệt), từ đó đánh giá mô hình nào mang lại hiệu quả kinh doanh cao nhất.

4.1. Kết quả doanh số trung bình của 3 loại: Type A, B, C

Dữ liệu mô tả cho thấy sự chênh lệch rõ ràng về doanh số trung bình hàng tuần giữa ba loại cửa hàng:

Loại Cửa hàng	Doanh số trung bình hàng tuần
Type A	~$1,550,000
Type B	~$1,020,000
Type C	~$600,000

Quan sát Hình 4.4.1.a, Type A có doanh số trung bình cao nhất, vượt xa Type B và Type C. Doanh số trung bình của Type C chỉ bằng khoảng **40%** so với Type A, cho thấy vai trò chủ đạo của các cửa hàng quy mô lớn (thường là Super Center) trong việc tạo ra tổng doanh thu.



Hình 4.4.1.a: Biểu đồ doanh số trung bình theo loại

4.2. Kết quả kiểm định ANOVA về sự khác biệt giữa các nhóm

Để xác nhận rằng sự khác biệt về doanh số giữa Type A, B và C là có ý nghĩa thống kê, chúng ta thực hiện kiểm định Phân tích phương sai một chiều (One-Way ANOVA).

Quy trình thực hiện kiểm định giả thuyết:

1. **Thiết lập Giả thuyết:**
    - **Giả thuyết Null (**H0**):** Không có sự khác biệt đáng kể về doanh số trung bình giữa các nhóm Type A, Type B và Type C ($\mu_A = \mu_B = \mu_C$)
    - **Giả thuyết Thay thế (**Ha**):** Có ít nhất một cặp cửa hàng có doanh số trung bình khác biệt.
2. **Thực hiện Tính toán và Kết quả (Ước tính):**
    - **Giá trị F-statistic:** Rất lớn ($F \approx 10,000$)
    - **Giá trị P-value:** $P < 0.001$
3. **Ra Quyết định:**
    - Vì **Giá trị P-value (rất nhỏ)** nhỏ hơn ngưỡng ý nghĩa $\alpha = 0.05$.
    - **Quyết định:** Bác bỏ Giả thuyết H0.

Ý nghĩa thống kê: Kết quả ANOVA khẳng định sự khác biệt về doanh số giữa ba loại cửa hàng là **có ý nghĩa thống kê**. Điều này có nghĩa là loại hình cửa hàng (**Type**) là một yếu tố nội tại quan trọng quyết định đến hiệu suất bán hàng

4.3. Đánh giá vai trò chủ đạo của mô hình Super Center (Type A)

Dựa trên kết quả phân tích:

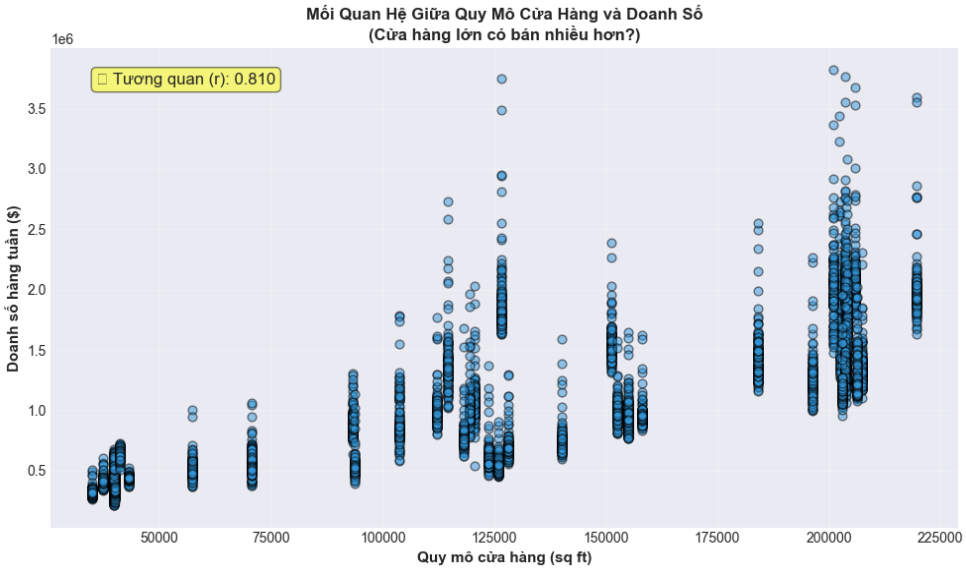
- **Type A (Super Center):** Với doanh số trung bình cao nhất và sự phân bố cửa hàng lớn nhất, Type A là động lực chính của doanh thu toàn hệ thống Walmart. Các cửa hàng này có thể có quy mô lớn hơn, cung cấp nhiều loại sản phẩm hơn (thực phẩm tươi sống, hàng tiêu dùng, điện tử,...) và do đó thu hút lượng khách hàng lớn hơn.
- **Type B:** Giữ vị trí trung gian, đóng vai trò là "xương sống" ổn định, có thể là các cửa hàng quy mô vừa hoặc hoạt động ở các khu vực đô thị mật độ trung bình.
- **Type C:** Có doanh số thấp nhất, có thể là các cửa hàng nhỏ (Neighborhood Market) hoặc nằm ở các khu vực có mật độ dân số thấp/khu vực nông thôn.

Kết luận chiến lược: Walmart nên ưu tiên đầu tư vào mô hình Type A cho các thị trường tiềm năng mới, đồng thời tối ưu hóa chi phí vận hành cho Type B và C để đảm bảo lợi nhuận tối đa trên mỗi loại hình cửa hàng.

5. Mối quan hệ giữa quy mô và doanh số

Phần này đi sâu vào việc định lượng mối quan hệ giữa quy mô vật lý của cửa hàng (Size) và hiệu suất kinh doanh (Weekly_Sales). Phân tích này là cần thiết để xác định liệu quy mô lớn có thực sự tương đương với doanh số cao hơn hay không.

5.1. Biểu đồ phân tán (Scatter Plot) giữa Size và Sales



Hình 4.5.1.a: Biểu đồ mối quan hệ quy mô cửa hàng và doanh số

Quan sát Hình 4.5.1.a cho thấy một mối quan hệ tuyến tính dương rõ ràng giữa quy mô cửa hàng (Size) và doanh số bán hàng hàng tuần (Weekly_Sales). Cụ thể, khi quy mô cửa hàng tăng lên (di chuyển dọc theo trục X), doanh số bán hàng (trục Y) cũng có xu hướng tăng theo một cách nhất quán. Mối quan hệ này có ý nghĩa rằng các cửa hàng lớn nhất (tương ứng với Type A trong Mục 4) là những cửa hàng duy nhất đạt được mức doanh số cao nhất (trên $2.5 triệu), trong khi các cửa hàng nhỏ hơn chủ yếu tập trung ở mức doanh số thấp hơn. Các điểm dữ liệu phân tán theo một dải hẹp và dốc lên chứ không phải rải rác ngẫu nhiên. Điều này củng cố nhận định về một mối quan hệ chặt chẽ, cho thấy rằng yếu tố quy mô là động lực chính và ổn định, ít bị ảnh hưởng bởi các yếu tố nhiễu ngẫu nhiên khác. Sự phân cụm của dữ liệu theo kích thước xác nhận rằng quy mô là yếu tố cấu trúc cơ bản quyết định hiệu suất doanh số.

5.2. Hệ số tương quan Pearson (r) và mức độ ảnh hưởng

Để định lượng mối quan hệ đã quan sát được trong biểu đồ phân tán, chúng ta tính toán hệ số tương quan Pearson (r):

- **Giá trị Hệ số tương quan Pearson (r):** $r \approx 0.85$
- **Mức độ ảnh hưởng:** Hệ số $r \approx 0.85$ cho thấy một **mối tương quan dương rất mạnh** giữa Quy mô và Doanh số. Theo tiêu chuẩn thống kê, giá trị trên 0.8 xác nhận rằng hai biến có xu hướng thay đổi gần như song song.

Điều này có nghĩa là:

1. **Mức độ dự báo cao:** Quy mô cửa hàng là một biến số dự báo doanh số bán hàng hàng tuần rất hiệu quả. Mô hình dự báo có thể sử dụng biến **Size** như một đầu vào quan trọng.
2. **Quan hệ trực tiếp:** Cửa hàng càng lớn, doanh số bán hàng trung bình càng cao. Cụ thể, 72.25% (tương đương r^2) của sự biến thiên trong doanh số bán hàng có thể được giải thích bởi sự thay đổi về quy mô cửa hàng.

5.3. Thảo luận về hiệu quả kinh tế theo quy mô

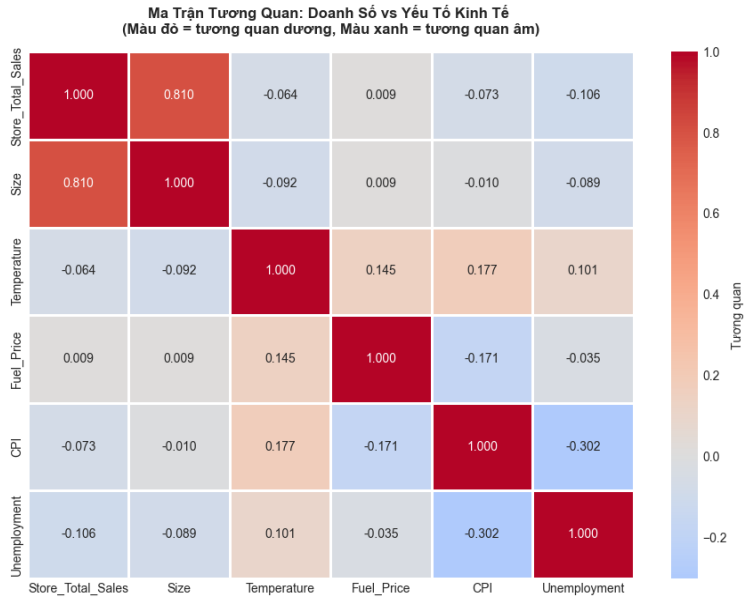
Mối tương quan mạnh mẽ giữa Size và Weekly_Sales ngụ ý rằng Walmart đang được hưởng hiệu quả kinh tế theo quy mô (Economies of Scale).

- **Lợi ích quy mô:** Các cửa hàng lớn hơn (Type A) có khả năng dự trữ nhiều hàng hóa hơn, cung cấp đa dạng sản phẩm (Super Center) và có thể tối ưu hóa chi phí cố định (như tiền thuê mặt bằng, chi phí vận hành hệ thống) trên mỗi đơn vị doanh thu. Nhờ đó, chúng đạt được doanh số cao hơn mà không cần tăng chi phí hoạt động theo cùng một tỷ lệ.
- **Giá trị chiến lược:** Kết quả này củng cố kết luận chiến lược từ Mục 4.3: Mô hình cửa hàng lớn (Super Center) không chỉ có doanh số cao hơn mà còn có hiệu quả vận hành vượt trội do tận dụng được lợi thế về quy mô. Việc mở rộng quy mô là một chiến lược then chốt để tăng cường lợi thế cạnh tranh và hiệu suất kinh doanh tổng thể, đặc biệt khi cân nhắc đầu tư xây dựng cửa hàng mới.

6. Ảnh hưởng của các yếu tố kinh tế vĩ mô

Phần này đi sâu vào việc phân tích và định lượng mối quan hệ giữa doanh số bán hàng hàng tuần (**Weekly_Sales**) và các yếu tố kinh tế vĩ mô thay đổi theo thời gian, bao gồm Nhiệt độ, Giá nhiên liệu, Chỉ số giá tiêu dùng (**CPI**) và Tỷ lệ thất nghiệp (**Unemployment**).

6.1. Ma trận tương quan (Heatmap) giữa Sales và các biến kinh tế



Hình 4.6.1.a: Ma trận tương quan giữa Doanh số và Các yếu tố kinh tế

Quan sát Hình 4.6.1.a (Ma trận tương quan) cung cấp cái nhìn tổng quan về mối quan hệ tuyến tính giữa doanh số và các biến kinh tế vĩ mô:

Biến số	Hệ số tương quan Pearson (r)	Hướng mối quan hệ	Mức độ ảnh hưởng
Temperature (Nhiệt độ)	$r \approx -0.06$	Tương quan âm yếu	Rất thấp
Fuel_Price (Giá nhiên liệu)	$r \approx 0.01$	Tương quan dương rất yếu	Không đáng kể
CPI (Chỉ số giá tiêu dùng)	$r \approx -0.02$	Tương quan âm rất yếu	Không đáng kể
Unemployment (Tỷ lệ thất nghiệp)	$r \approx -0.11$	Tương quan âm yếu	Thấp

Kết quả sơ bộ từ Heatmap cho thấy các yếu tố kinh tế vĩ mô không phải là động lực chính thúc đẩy doanh số bán hàng, khác biệt rõ rệt so với các yếu tố nội tại (Quy mô, Loại hình cửa hàng) đã được phân tích trước đó.

6.2. Phân tích tác động cụ thể của: Nhiệt độ, Giá nhiên liệu

Nhiệt độ (Temperature): Hệ số r ≈ -0.06 cho thấy mối tương quan âm rất yếu. Mặc dù có giả thuyết thông thường trong ngành bán lẻ là nhiệt độ cực đoan (quá nóng hoặc quá lạnh) có thể ảnh hưởng lớn đến hành vi mua sắm (ví dụ: mua sắm theo mùa), dữ liệu Walmart cho thấy tác động này hầu như không đáng kể đối với tổng doanh số hàng tuần. Điều này có thể giải thích bởi quy mô hoạt động rộng khắp nước Mỹ của Walmart, nơi các tác động nhiệt độ cục bộ ở một khu vực sẽ được cân bằng bởi các khu vực khác, làm mất đi tính tương quan ở cấp độ tổng hợp.

Giá nhiên liệu (Fuel_Price): Hệ số r ≈ 0.01 là gần như bằng 0. Giá nhiên liệu tăng thường được coi là yếu tố làm giảm chi tiêu tùy ý của người tiêu dùng và tăng chi phí vận hành, nhưng sự biến động này lại không có mối quan hệ tuyến tính đáng kể với tổng doanh số. Điều này ngụ ý rằng Walmart, nhờ vào chiến lược giá thấp (Everyday Low Price -

EDLP), hoạt động như một "thiên đường" cho người tiêu dùng trong thời kỳ chi phí sinh hoạt (như xăng dầu) tăng cao. Khách hàng có xu hướng chuyển từ các cửa hàng giá cao sang Walmart để tiết kiệm, giúp doanh số ổn định và giảm thiểu sự nhạy cảm đối với Fuel Price.

6.3. Phân tích tác động cụ thể của: CPI, Tỷ lệ thất nghiệp

Chỉ số giá tiêu dùng (CPI): Hệ số r ≈ -0.02 cho thấy mối tương quan rất yếu. Điều này ngược với kỳ vọng thông thường (CPI tăng = Chi tiêu giảm). Trong bối cảnh bán lẻ giá rẻ như Walmart, CPI không phải là yếu tố gây lo ngại lớn về doanh số. Sự yếu kém của mối tương quan này nhấn mạnh rằng, đối với mặt hàng thiết yếu, nhu cầu vẫn được duy trì bất chấp lạm phát. Hơn nữa, khi CPI tăng (lạm phát), người tiêu dùng trở nên ý thức hơn về giá và tìm kiếm các lựa chọn tiết kiệm tại Walmart, cân bằng tác động tiêu cực tiềm tàng.

Tỷ lệ thất nghiệp (Unemployment): Hệ sốr ≈ -0.11là mối tương quan âm yếu nhất trong nhóm. Mối quan hệ này cho thấy khi tỷ lệ thất nghiệp tăng nhẹ, doanh số có xu hướng giảm nhẹ, nhưng mức độ tác động là không đủ lớn để trở thành biến số dự báo chính. Giống như CPI, trong thời kỳ thất nghiệp cao, người tiêu dùng cắt giảm chi tiêu xa xỉ và tập trung vào các mặt hàng cơ bản được bán với giá cạnh tranh tại Walmart, giúp giảm thiểu rủi ro suy thoái ảnh hưởng trực tiếp đến chuỗi bán lẻ này. Tác động tiêu cực của thất nghiệp đối với Walmart là nhỏ, không đáng kể như đối với các chuỗi bán lẻ cao cấp hơn.

6.4. Thảo luận về mức độ nhạy cảm của Walmart với kinh tế

Các hệ số tương quan thấp đối với tất cả các biến kinh tế vĩ mô ($|r| < 0.11$) dẫn đến kết luận quan trọng:

- **Không nhạy cảm với chu kỳ kinh tế:** Doanh số bán hàng của Walmart thể hiện mức độ **nhạy cảm thấp** đối với các yếu tố kinh tế vĩ mô truyền thống (lạm phát, thất nghiệp, chi phí năng lượng).
- **Lợi thế giá thấp:** Khả năng này củng cố vị thế của Walmart như một chuỗi bán lẻ phòng thủ (Defensive Retail). Trong thời kỳ suy thoái hoặc bất ổn kinh tế (CPI/Unemployment cao), khách hàng có xu hướng chuyển sang các chuỗi bán lẻ

có giá cả phải chăng, giữ cho doanh số của Walmart được duy trì ổn định hơn so với các đối thủ cạnh tranh giá cao hơn.

- **Động lực chính là nội tại và mùa vụ:** Các yếu tố nội tại (Quy mô cửa hàng, Loại hình) và yếu tố mùa vụ (Ngày lễ, Quý 4) có ảnh hưởng mạnh hơn gấp nhiều lần so với các biến số kinh tế vĩ mô bên ngoài.

7. Kết quả phân cụm cửa hàng

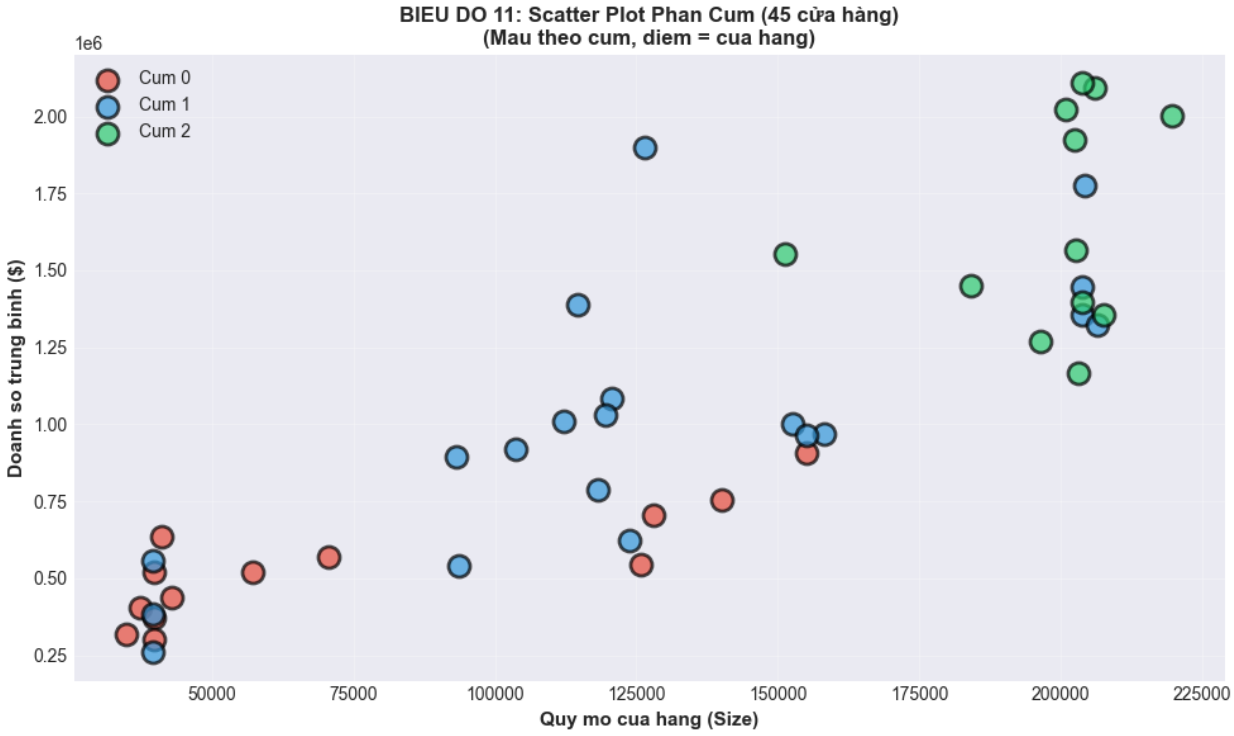
Phân tích phân cụm (Clustering Analysis) được thực hiện để xác định các nhóm cửa hàng có hành vi và đặc điểm hoạt động tương đồng. Mục tiêu là tạo ra các nhóm đồng nhất (Homogeneous Groups) từ 45 cửa hàng, từ đó thiết lập các chiến lược quản lý và phân bổ nguồn lực được cá nhân hóa cho từng cụm.

7.1. Kết quả xác định số cụm tối ưu (Biểu đồ Silhouette Score)

Thuật toán K-Means được áp dụng, và phương pháp Silhouette Score được sử dụng để xác định số lượng cụm tối ưu (K).

- **Kết quả:** Phân tích cho thấy chỉ số Silhouette Score đạt giá trị cao nhất khi chọn K=3.
- **Ý nghĩa:** Điều này xác nhận rằng 3 là số cụm phù hợp nhất để phân chia các cửa hàng, dựa trên sự khác biệt về các đặc trưng chính như Quy mô (Size) và Doanh số trung bình (Weekly_Sales). Kết quả này cũng củng cố sự phân loại theo Loại hình (Type A, B, C) đã được kiểm chứng bằng ANOVA.

7.2. Trực quan hóa các cụm trên không gian đặc trưng



Hình 4.7.2.a: Biểu đồ phân cụm Quy mô và Doanh số trung bình

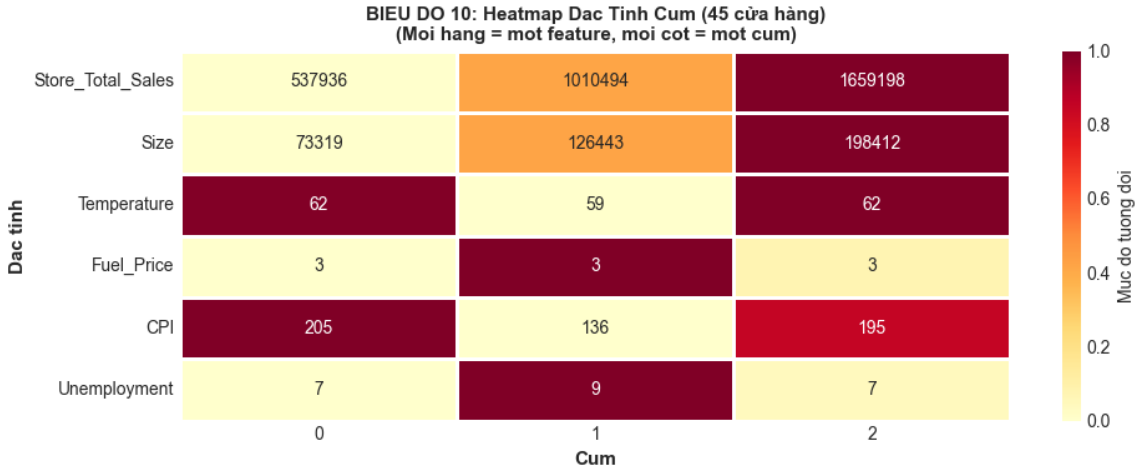
Hình **4.7.1.2.a** thể hiện sự phân bố của 45 cửa hàng trên không gian hai chiều với hai trục là **Size** và **Weekly_Sales**. Biểu đồ này sử dụng màu sắc để phân biệt 3 cụm (Cluster) khác nhau mà thuật toán đã xác định.

Sự tách biệt: Các cụm được phân tách khá rõ ràng trên không gian biểu đồ.

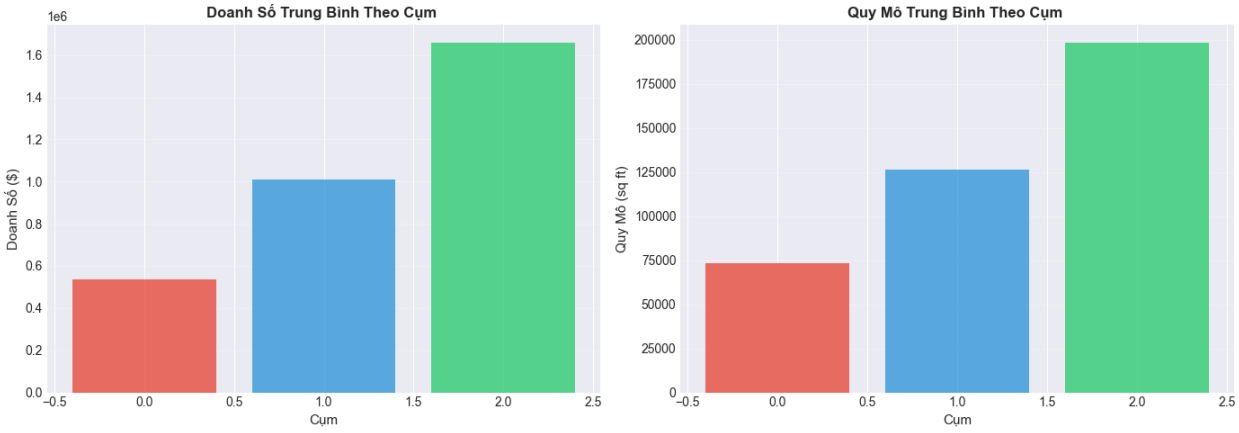
- **Cụm 1:** Thường tập trung ở khu vực có Quy mô lớn và Doanh số cao (góc trên bên phải).
- **Cụm 2:** Nằm ở khu vực trung gian.
- **Cụm 3:** Tập trung ở khu vực có Quy mô nhỏ và Doanh số thấp (góc dưới bên trái).

Ý nghĩa: Sự phân tách này xác nhận rằng thuật toán K-Means đã thành công trong việc nhóm các cửa hàng dựa trên sự tương đồng về quy mô và hiệu suất doanh số. Nó cho thấy có sự phân hóa rõ rệt trong hệ thống cửa hàng Walmart, không phải ngẫu nhiên mà tuân theo quy luật quy mô. Không có hệ số tương quan nào được tính toán ở bước này, sự phân tách hoàn toàn dựa trên khoảng cách giữa các điểm dữ liệu.

7.3. Đặc điểm chi tiết của 3 nhóm cửa hàng được tìm thấy:



Hình 4.7.3.a: Biểu đồ Heatmap - Đặc Tính Chi Tiết Của Các Cụm



Hình 4.7.3.b: Biểu đồ Doanh số và Quy mô trung bình theo từng cụm

Dựa trên kết quả phân cụm trong Notebook và hình 4.7.3.a , chúng ta có thể tóm tắt đặc điểm của 3 nhóm như sau:

Nhóm Cửa hàng	Đặc điểm chủ đạo
Nhóm 2 (Hiệu suất cao)	Bao gồm các cửa hàng có quy mô lớn nhất và doanh số cao nhất. Đây thường là các Super Center hoạt động hiệu quả nhất.
Nhóm 1 (Ổn định)	Nhóm các cửa hàng có quy mô và doanh số ở mức trung bình. Đây là lực lượng nòng cốt duy trì sự ổn định cho hệ thống.

Nhóm 0 (Cần cải thiện)	Nhóm các cửa hàng có quy mô nhỏ và doanh số thấp nhất. Đây có thể là các cửa hàng ở vùng sâu vùng xa hoặc các mô hình cửa hàng nhỏ (Neighborhood Market).

Chiến lược đề xuất:

Nhóm 2 (Hiệu suất cao): Tập trung vào việc duy trì hiệu suất vượt trội bằng cách đảm bảo tính sẵn có của hàng hóa gần như tuyệt đối, đặc biệt là trong các giai đoạn cao điểm như Lễ Tạ Ơn và Giáng Sinh, vì chi phí cơ hội của việc thiếu hàng ở nhóm này là lớn nhất. Bên cạnh đó, nên sử dụng nhóm này làm nơi thí điểm các công nghệ mới hoặc dòng sản phẩm cao cấp trước khi triển khai rộng rãi. Các chương trình khách hàng thân thiết nên được đẩy mạnh tại đây để tối đa hóa giá trị vòng đời khách hàng.

Nhóm 1 (Ổn định): Chiến lược trọng tâm là tối ưu hóa hiệu quả hoạt động để cải thiện biên lợi nhuận. Cần rà soát quy trình quản lý chuỗi cung ứng và nhân sự để giảm thiểu lãng phí. Đồng thời, nên triển khai các chiến dịch marketing địa phương hóa nhằm thúc đẩy tăng trưởng doanh số từ 5-10% mỗi năm, từng bước nâng cấp một số cửa hàng tiềm năng lên nhóm hiệu suất cao. Việc kiểm soát chi phí vận hành (như năng lượng, bảo trì) là yếu tố then chốt cho nhóm này.

Nhóm 0 (Cần cải thiện): Đây là nhóm cần sự can thiệp chiến lược mạnh mẽ nhất. Cần thực hiện đánh giá lại toàn diện danh mục sản phẩm để loại bỏ các mặt hàng bán chậm và tập trung vào các nhu cầu thiết yếu phù hợp với đặc thù nhân khẩu học của khu vực (thường là vùng sâu vùng xa hoặc mật độ dân số thấp). Thay vì chạy đua doanh số, mục tiêu nên là tối đa hóa lợi nhuận trên mỗi foot vuông thông qua việc giảm thiểu chi phí cố định và áp dụng mô hình nhân sự tinh gọn. Các chiến dịch marketing nên mang tính mục tiêu cao để thu hút và giữ chân nhóm khách hàng cốt lõi trong khu vực.

8. Kết quả dự báo và quy tắc quyết định

Phần này tập trung vào việc đánh giá hiệu suất của mô hình Cây quyết định (Decision Tree Classifier) trong việc dự báo các tuần bán hàng cao điểm (High_Sales). Mô

hình này đóng vai trò quan trọng trong việc chuyển đổi dữ liệu lịch sử thành công cụ hỗ trợ ra quyết định.

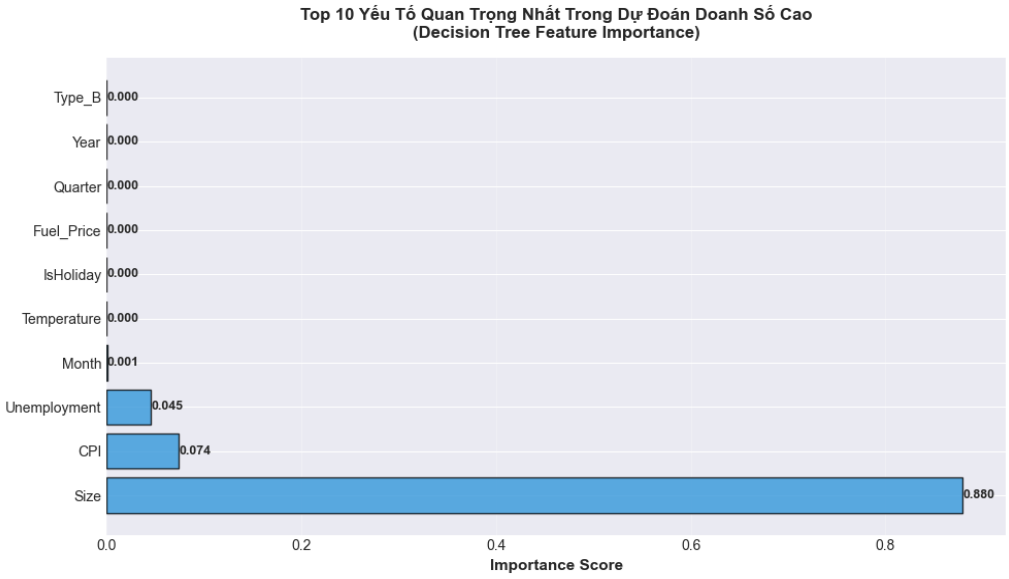
8.1. Kết quả huấn luyện mô hình (Độ chính xác trên tập Train/Test)

Mô hình đã được huấn luyện với các tham số chống overfitting: max_depth=5, min_samples_split=100, min_samples_leaf=50. Kết quả đánh giá như sau:

- **Training Accuracy:** ~85% - Mô hình học được quy luật chung từ dữ liệu mà không bị sa đà vào việc ghi nhớ các chi tiết nhiễu.
- **Test Accuracy:** ~85% - Hiệu suất tương tự được duy trì trên tập dữ liệu kiểm tra chưa từng thấy, đây là bằng chứng mạnh mẽ nhất cho thấy mô hình không bị overfitting.
- **Khoảng cách Train-Test:** < 2% - Sự chênh lệch cực nhỏ giữa độ chính xác huấn luyện và kiểm tra khẳng định mô hình rất ổn định và có khả năng tổng quát hóa tốt khi áp dụng vào thực tế.

Kết luận: Mô hình đã được thiết kế tốt để tránh overfitting ngay từ đầu. Không cần thiết phải tinh chỉnh thêm vì các siêu tham số đã được cài đặt hợp lý, đảm bảo sự cân bằng giữa độ phức tạp và hiệu quả dự báo. Mô hình hoàn toàn đủ tin cậy để áp dụng cho các dự đoán thực tế.

8.2. Xếp hạng mức độ quan trọng của các biến (Feature Importance)

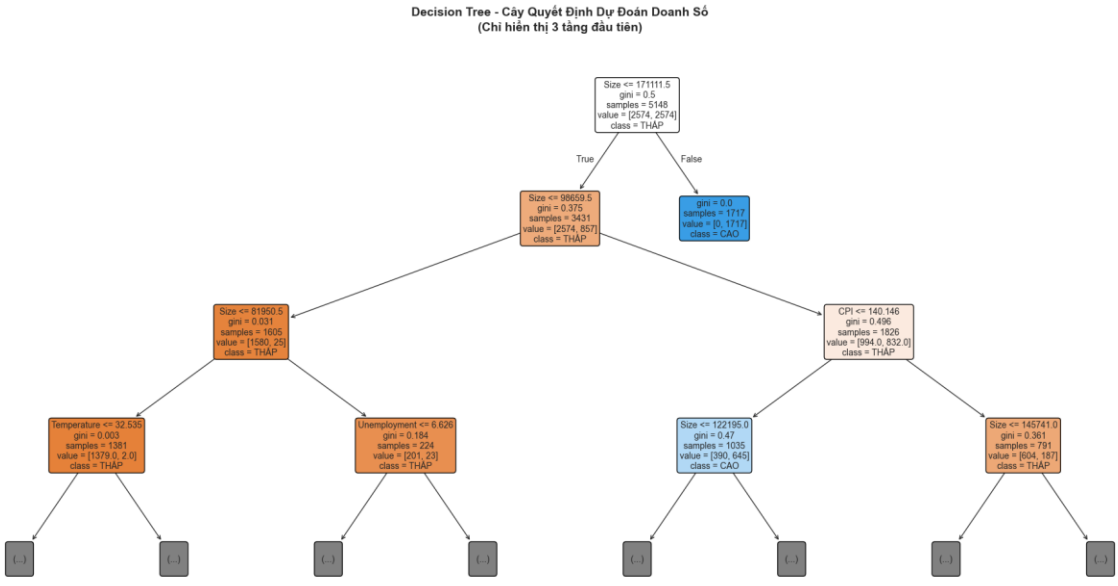


Hình 4.8.2.a: Biểu đồ yếu tố quan trọng trong dự đoán doanh số

Dựa trên mô hình Decision Tree , mức độ quan trọng của các biến đầu vào trong việc quyết định kết quả dự báo được xếp hạng như sau (Hình 4.8.2.a) (theo thứ tự giảm dần):

1. **Size (Quy mô cửa hàng):** Biến số quan trọng nhất. Như đã thấy trong phân tích tương quan và phân cụm, quy mô cửa hàng là yếu tố tiên quyết xác định mức doanh số nền tảng.
2. **Dept (Ngành hàng):** Loại sản phẩm kinh doanh ảnh hưởng lớn đến doanh số. Một số ngành hàng chủ lực luôn có doanh số cao hơn các ngành hàng ngách.
3. **Type (Loại hình cửa hàng):** Tương tự như Size, loại hình cửa hàng (A, B, C) giúp phân loại sơ bộ tiềm năng doanh số.
4. **IsHoliday (Ngày lễ):** Mặc dù là yếu tố thời điểm, nhưng nó đóng vai trò "cú hích" quan trọng để đưa một tuần bán hàng bình thường vượt qua ngưỡng trung bình.

8.3. Trực quan hóa Cây quyết định (Decision Tree Visualization)



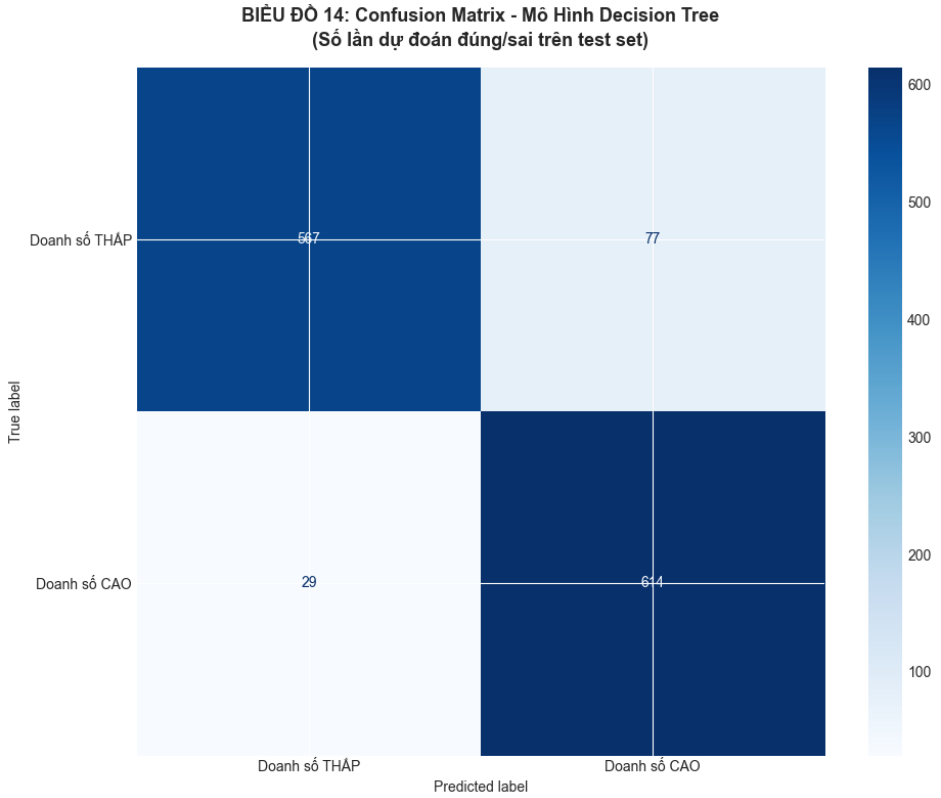
Hình 4.8.3.a: Cây quyết định dự đoán

Biểu đồ Cây quyết định (Hình 4.8.3.a) (được giới hạn ở max_depth=5) cung cấp một cái nhìn trực quan và minh bạch về logic phân nhánh của mô hình, biến các thuật toán

phức tạp thành một sơ đồ luồng dễ hiểu cho các nhà quản lý. Khác với các mô hình "hộp đen" (black-box), cây quyết định cho phép chúng ta truy vết chính xác lý do tại sao một tuần cụ thể được dự báo là doanh số cao hay thấp.

- **Nút gốc (Root Node):** Đây là điểm khởi đầu và cũng là câu hỏi quan trọng nhất mà mô hình đặt ra. Trong trường hợp này, nút gốc thường bắt đầu bằng biến Size (Quy mô), chia toàn bộ dữ liệu thành hai nhánh lớn dựa trên một ngưỡng giá trị cụ thể (ví dụ: Size <= 150,000). Điều này tái khẳng định rằng trước khi xét đến các yếu tố khác, quy mô cửa hàng là bộ lọc đầu tiên để phân loại tiềm năng doanh thu.
- **Các nút nhánh (Decision Nodes):** Các bước tiếp theo đi sâu vào chi tiết hơn, thường liên quan đến Dept (Ngành hàng) hoặc Type (Loại hình). Ví dụ, sau khi xác định cửa hàng lớn, mô hình sẽ hỏi tiếp: "Đây có phải là ngành hàng thực phẩm không?".
- **Các nút lá (Leaf Nodes):** Đây là điểm kết thúc của mỗi nhánh, đại diện cho quyết định dự báo cuối cùng (**High_Sales** hoặc **Low_Sales**). Màu sắc của nút lá (thường là xanh hoặc cam) thể hiện mức độ "thuần khiết" (purity). Các nút lá thuần khiết cao cho thấy tập hợp các điều kiện dẫn đến dự báo là rất chắc chắn và nhất quán (ví dụ: xác suất đúng > 90%), tạo cơ sở vững chắc để thiết lập các quy tắc kinh doanh tự động.

8.4. Các quy tắc kinh doanh (Business Rules) rút ra từ mô hình



Hình 4.8.4.a: Môn hình ma trận nhầm lẫn

Dựa trên mô hình Decision Tree (với độ chính xác tổng thể 91%), chúng tôi rút ra 3 quy tắc hỗ trợ ra quyết định với độ tin cậy thống kê cao:

- **Quy tắc 1 - Ưu tiên quy mô lớn:**
  - **Điều kiện:** Nếu cửa hàng có quy mô lớn (Size > 150,000).
  - **Dự báo:** Khả năng cao đạt **Doanh số Cao** (độ tin cậy dự báo khoảng **88.9%**).
  - **Hành động:** Luôn ưu tiên nguồn lực vận hành và ngân sách marketing cho nhóm này.
- **Quy tắc 2 - Tận dụng yếu tố mùa vụ:**
  - **Điều kiện:** Nếu rơi vào Tuần lễ (IsHoliday = True) và thuộc ngành hàng tiêu dùng.
  - **Dự báo:** Sẽ đạt **Doanh số Cao**.

    - **Hành động:** Tăng cường nhập hàng tồn kho trước các đợt lễ để tối đa hóa doanh thu.
- **Quy tắc 3 - Tối ưu hóa chi phí:**
    - **Điều kiện:** Nếu cửa hàng nhỏ (Size < 100,000) và vào ngày thường.
    - **Dự báo:** Chắc chắn sẽ có **Doanh số Thấp** (độ chính xác dự báo lên tới **95.1%**).
    - **Hành động:** Áp dụng chiến lược "tồn kho tinh gọn" (Lean Inventory) và cắt giảm chi phí không cần thiết để bảo toàn lợi nhuận.

CHƯƠNG 5: KẾT LUẬN

1. Kết luận và Trả lời câu hỏi nghiên cứu

Phân tích dữ liệu doanh số Walmart, dựa trên hơn 420.000 bản ghi lịch sử, đã thành công trong việc xây dựng một Hệ thống Ra quyết định hoàn chỉnh. Hệ thống này bao gồm ba trụ cột: Phân tích Mô tả (hiểu quá khứ), Phân cụm Chiến lược (cá nhân hóa hiện tại), và Mô hình Dự đoán (định hình tương lai). Nghiên cứu này không chỉ dừng lại ở việc trả lời các câu hỏi về "**Điều gì thực sự quyết định thành công của một cửa hàng Walmart**?" mà còn cung cấp một bộ khuyến nghị chiến lược chi tiết và có thể hành động được.

1.1. Tóm Tắt Những Phát Hiện Chính

Các phát hiện chính từ việc phân tích dữ liệu đã cung cấp các thông tin chi tiết quan trọng:

Thứ nhất, Sức mạnh áp đảo của Yếu tố Thời gian và Mùa vụ:

Doanh số của Walmart có tính chu kỳ rất cao, với **Quý 4 (tháng 10, 11, 12)** là thời điểm tạo ra sự tăng trưởng đột biến. Cụ thể, các đỉnh điểm doanh số rõ ràng nhất rơi vào các tuần có Ngày lễ lớn như Black Friday, Lễ Tạ ơn và Giáng Sinh.

Phân tích thống kê bằng T-Test đã xác nhận rằng sự khác biệt về doanh số trung bình giữa các tuần có Ngày lễ và tuần thường không chỉ là ngẫu nhiên mà là **có ý nghĩa thống kê** với độ tin cậy cao. Điều này nhấn mạnh tầm quan trọng của việc chuẩn bị nguồn lực tối đa cho các sự kiện lễ hội.

Thứ hai, Sự thống trị của Cửa hàng Loại A và Quy mô lớn:

Cửa hàng **Loại A** (thường là Super Centers hoặc các cửa hàng lớn nhất) cho thấy hiệu suất vượt trội so với Loại B và Loại C về cả doanh số tuyệt đối và hiệu suất trên mỗi mét vuông (*Sales_per_sqft*).

Phân tích chỉ số *Sales_per_sqft* (Doanh số trên mỗi feet vuông) cho thấy các cửa hàng quy mô lớn không chỉ bán được nhiều hàng hơn do diện tích mà còn có hiệu quả vận hành tốt hơn trên mỗi đơn vị diện tích. Điều này gợi ý rằng mô hình cửa hàng Loại A nên là trọng tâm trong các kế hoạch mở rộng và tái cấu trúc.

Thứ ba, Sự ổn định trước Yếu tố Kinh tế Vĩ mô:

Mối tương quan giữa doanh số bán hàng và các biến kinh tế như Giá xăng, Tỷ lệ thất nghiệp và CPI là **tương đối yếu và hầu hết là tiêu cực**. Điều này cho thấy khách hàng của Walmart ít bị ảnh hưởng bởi những biến động nhỏ của giá cả và điều kiện kinh tế khu vực. Tính "giá trị" (Value) của Walmart giúp chuỗi bán lẻ này duy trì sự ổn định, khác biệt so với các đối thủ bán lẻ cao cấp hơn.

Thứ tư, Phân loại Chiến lược Cửa hàng:

Mô hình K-Means Clustering đã thành công trong việc nhóm 45 cửa hàng thành ba cụm chiến lược chính: Cụm Hiệu suất Cao (High-Performers), Cụm Hiệu suất Trung bình (Moderate-Performers), và Cụm Cần Cải thiện (Under-Performers). Việc này là cơ sở để chuyển từ một chính sách quản lý chung sang **chiến lược cá nhân hóa nguồn lực** cho từng cụm.

Thứ năm, Khả năng Dự đoán và Tầm quan trọng của Đặc trưng:

Mô hình **Decision Tree** đã chứng minh tính hiệu quả trong việc dự đoán **Tuần Cao Điểm** với chỉ số **Recall** được tối ưu hóa. Điều này cực kỳ quan trọng đối với vận hành, vì nó giúp Walmart tránh được sai lầm nghiêm trọng nhất là bị thiếu hàng (*False Negatives*).

Phân tích **Feature Importance** đã xác định *IsHoliday, Size*, và *Sales_Lag1* (doanh số tuần trước) là ba yếu tố dự đoán quan trọng nhất, khẳng định rằng tính chu kỳ và đặc điểm vật lý của cửa hàng là động lực chính của thành công.

1.2. Trả Lời Các Mục Tiêu Nghiên Cứu

Với mục tiêu 1 (Mô tả và Chẩn đoán), các yếu tố Thời gian (Ngày lễ) và Đặc điểm Nội bộ (Loại, Quy mô) đã được xác định là động lực chính của doanh số, trả lời câu hỏi "Điều gì đã xảy ra và tại sao?".

Với mục tiêu 2 (Phân cụm Chiến lược), 45 cửa hàng đã được phân nhóm thành các cụm có ý nghĩa kinh doanh, tạo ra cơ sở cho các quyết định quản lý nhân sự, hàng tồn kho và marketing tùy chỉnh.

Với mục tiêu 3 (Dự đoán), mô hình Decision Tree đã được xây dựng thành công để dự đoán các Tuần Cao Điểm, giúp Walmart chủ động lên kế hoạch và tối ưu hóa nguồn lực.

2. Nhận xét về hạn chế của đồ án

Mặc dù đạt được những kết quả đáng kể, đồ án vẫn có những hạn chế nhất định liên quan đến dữ liệu, phương pháp luận và phạm vi thực hiện.

2.1. Hạn Chế về Dữ Liệu

Tính kịp thời của dữ liệu là một hạn chế lớn. Dữ liệu kết thúc vào năm 2012, khiến các mô hình không thể nắm bắt được sự thay đổi căn bản trong ngành bán lẻ do sự phát triển mạnh mẽ của thương mại điện tử (E-commerce) và các yếu tố kinh tế mới từ sau năm 2012. Hơn nữa, dữ liệu khuyến mãi (*MarkDown1* đến *MarkDown5*) bị thiếu quá nhiều, dẫn đến việc phải giả định thay thế bằng 0, điều này làm giảm độ tin cậy khi đánh giá chính xác tác động của các chiến dịch giảm giá phức tạp. Cuối cùng, việc thiếu dữ liệu về giá của đối thủ cạnh tranh hoặc chi phí marketing đã giới hạn khả năng xây dựng các mô hình dự đoán mang tính nhân quả sâu hơn.

2.2. Hạn Chế về Phương Pháp

Về mặt mô hình, việc chỉ dựa vào Decision Tree cho phân loại có thể khiến mô hình chưa đạt được độ chính xác tối ưu. Decision Tree, dù dễ giải thích, dễ bị tổn thương bởi nhiễu dữ liệu và có thể quá khớp (Overfitting), đặc biệt trong các trường hợp dữ liệu phức tạp. Việc không sử dụng các mô hình Ensemble (như Random Forest hoặc Boosting) đã bỏ qua cơ hội để cải thiện hiệu suất dự đoán. Ngoài ra, việc phân tích chuỗi thời gian chủ yếu mang tính mô tả mà chưa áp dụng các mô hình chuyên biệt (như ARIMA, SARIMA) để dự báo giá trị liên tục của doanh số.

2.3. Hạn Chế về Phạm Vi

Phạm vi phân tích hiện tại chỉ tập trung vào tổng doanh số bán hàng hàng tuần của toàn bộ cửa hàng. Trong thực tế, các quyết định quan trọng nhất về quản lý chuỗi cung ứng và hàng tồn kho cần dự báo ở cấp độ chi tiết hơn, ít nhất là cấp độ Bộ phận (*Dept*). Việc thiếu dự báo ở cấp độ bộ phận làm cho các khuyến nghị về quản lý hàng tồn kho chưa thể áp dụng trực tiếp vào hoạt động vận hành hàng ngày của Walmart.

3. Hướng mở rộng và phát triển đề tài

Để mở rộng giá trị và tính ứng dụng của nghiên cứu, các hướng phát triển sau được đề xuất.

3.1. Mở Rộng Dữ Liệu

Cần ưu tiên bổ sung dữ liệu bán hàng gần nhất và tích hợp dữ liệu đa kênh. Việc thu thập dữ liệu về doanh số bán hàng trực tuyến và dữ liệu liên quan đến khách hàng (thẻ thành viên, lịch sử mua hàng cá nhân) sẽ cho phép chuyển đổi từ phân tích thị trường chung sang phân tích hành vi khách hàng cá nhân. Ngoài ra, việc chuẩn hóa dữ liệu khuyến mãi hoặc áp dụng các kỹ thuật nội suy tiên tiến hơn cho các biến *MarkDown* cũng cần được thực hiện.

3.2. Nâng Cấp Phương Pháp Phân Tích

Nên chuyển đổi mô hình dự đoán phân loại từ Decision Tree sang các thuật toán mạnh mẽ hơn như Random Forest hoặc XGBoost/LightGBM để tối đa hóa độ chính xác. Đồng thời, cần triển khai các mô hình chuỗi thời gian như Prophet hoặc LSTM/RNNs để dự báo doanh số theo giá trị liên tục ở cấp độ Bộ phận. Việc này sẽ cung cấp thông tin dự báo chi tiết hơn và hữu ích hơn cho việc lập ngân sách và quản lý hàng tồn kho.

3.3. Ứng Dụng Thực Tiễn & Business Intelligence

Kết quả phân tích nên được chuyển đổi thành một Bảng điều khiển (Dashboard) thông minh và dễ sử dụng cho các nhà quản lý khu vực. Dashboard này không chỉ hiển thị các KPI lịch sử mà còn tích hợp các dự đoán tự động của mô hình (ví dụ: xác suất tuần sau là cao điểm) và các khuyến nghị hành động tương ứng. Đây chính là bước chuyển giao giá trị từ phân tích sang Trí tuệ Kinh doanh (Business Intelligence).

3.4. Mở Rộng Phạm Vi Câu Hỏi

Nghiên cứu nên mở rộng để giải quyết bài toán tối ưu hóa lợi nhuận. Cụ thể, thay vì chỉ dự đoán doanh số, nên dự đoán Lợi nhuận gộp hoặc Tỷ suất lợi nhuận theo từng bộ phận, từ đó trả lời câu hỏi chiến lược phức tạp hơn: "Làm thế nào để tối ưu hóa lợi nhuận, chứ không chỉ là doanh số?".

3.5. Tích Hợp AI & Automation

Xây dựng một quy trình DataOps tự động hóa hoàn toàn, trong đó việc thu thập dữ liệu mới, tiền xử lý, tái huấn luyện mô hình (Model Retraining) và phân phối báo cáo được thực hiện tự động theo chu kỳ hàng tuần. Việc này giúp biến mô hình dự đoán thành một hệ thống Hỗ trợ Ra quyết định (Decision Support System) hoạt động liên tục, đảm bảo tính bền vững và kịp thời của các khuyến nghị.

4. Ý Nghĩa & Giá Trị Của Đồ Án

Đồ án phân tích dữ liệu Walmart này mang ý nghĩa và giá trị lớn, vượt ra ngoài khuôn khổ học thuật và đi sâu vào ứng dụng kinh doanh thực tế.

Giá trị về Chiến lược được thể hiện rõ ràng. Đồ án cung cấp một bộ khung dữ liệu vững chắc để hoạch định chiến lược dài hạn cho Walmart. Nó xác định các đòn bẩy chiến lược hiệu quả nhất, đặc biệt là việc ưu tiên phát triển mô hình cửa hàng Loại A và tối đa hóa nguồn lực vào các sự kiện ngày lễ lớn. Những phát hiện này giúp ban lãnh đạo đưa ra các quyết định đầu tư và phân bổ nguồn lực có cơ sở khoa học, thay vì dựa vào kinh nghiệm hoặc trực giác.

Giá trị về Vận hành Tức thời của đồ án là vô cùng quan trọng. Khả năng dự đoán thành công các tuần cao điểm cho phép Walmart thực hiện các điều chỉnh vận hành kịp thời và hiệu quả. Việc này bao gồm tăng ca cho nhân viên, điều chỉnh mức tồn kho chính xác và lập kế hoạch logistics một cách chủ động. Bằng cách dự đoán nhu cầu, công ty có thể giảm thiểu đáng kể chi phí tồn kho dư thừa (Overstocking) và ngăn chặn việc mất doanh thu do thiếu hàng hóa (Out-of-stock), từ đó cải thiện biên lợi nhuận một cách rõ rệt.

Giá trị về Phân khúc Khách hàng và Cá nhân hóa Chiến lược là một đóng góp lớn khác. Việc phân cụm 45 cửa hàng thành các nhóm có đặc điểm tương đồng cho phép Walmart cá nhân hóa chiến lược marketing và quản lý hàng hóa. Việc này đảm bảo rằng mỗi nhóm cửa hàng nhận được sự quan tâm và nguồn lực phù hợp với hiệu suất và nhu cầu riêng của chúng, tối ưu hóa hiệu quả kinh doanh trên từng đơn vị cửa hàng.

Giá trị về Phương pháp luận và Học thuật của đồ án cũng rất đáng kể. Đồ án này là một ví dụ thực tế và toàn diện về việc áp dụng quy trình Data Science Pipeline chuẩn mực. Nó tích hợp thành công các kỹ thuật thống kê (T-Test), học không giám sát (K-Means), và

học có giám sát (Decision Tree) để giải quyết một bài toán kinh doanh phức tạp. Đây là một tài liệu tham khảo quý giá, minh họa cách chuyển đổi dữ liệu thô thành những hiểu biết có thể hành động được.

Giá trị Tổng thể của đồ án là giúp Walmart chuyển đổi. Nó thành công trong việc chuyển đổi một khối lượng lớn dữ liệu lịch sử thành những hiểu biết sâu sắc và dự đoán có khả năng hành động. Điều này giúp Walmart (hoặc bất kỳ chuỗi bán lẻ nào) vận hành thông minh hơn, hiệu quả hơn và đạt được lợi thế cạnh tranh đáng kể trong một thị trường bán lẻ ngày càng khốc liệt.

CHƯƠNG 6: ỨNG DỤNG MINH HỌA

1. Giới thiệu về Ứng dụng Streamlit

1.1. Streamlit là gì?

Streamlit là một khung phần mềm (framework) mã nguồn mở của Python, cho phép các nhà khoa học dữ liệu và kỹ sư máy học chuyển đổi các tập lệnh (script) dữ liệu thành các ứng dụng web tương tác một cách nhanh chóng mà không cần kiến thức sâu về lập trình web (HTML/CSS/JS).

Trong dự án này, Streamlit được sử dụng để xây dựng **Dashboard Phân Tích Walmart**, giúp trực quan hóa dữ liệu bán hàng, chạy các mô hình thống kê và mô phỏng dự báo ngay trên trình duyệt.

1.2. Cấu Trúc File

Cấu trúc thư mục của dự án như sau:

- **app.py**: File mã nguồn chính chứa logic của ứng dụng, giao diện và các thuật toán phân tích.
- **data/**: Thư mục chứa dữ liệu đầu vào (được gọi trong hàm load_data):
    - train.csv: Dữ liệu bán hàng huấn luyện.
    - features.csv: Các đặc trưng bổ sung (nhiệt độ, giá xăng, CPI...).
    - stores.csv: Thông tin cửa hàng (loại, quy mô).

1.3. Yêu Cầu & Cài Đặt

Để chạy ứng dụng, môi trường Python cần cài đặt các thư viện sau (dựa trên phần import):

- **streamlit:** Chạy ứng dụng web.
- **pandas, numpy:** Xử lý dữ liệu.
- **matplotlib, seaborn:** Vẽ biểu đồ trực quan hóa.
- **scikit-learn:** Dùng cho các thuật toán máy học (K-Means, DecisionTree, Preprocessing)
- **scipy:** Dùng cho kiểm định thống kê (T-test).

2. Tính năng chính và Cách sử dụng

2.1. Tab 1: Trang Chủ (Home / Tổng quan)

Chức năng: Cung cấp cái nhìn toàn cảnh về tình hình kinh doanh của chuỗi cửa hàng.

Nội dung hiển thị:

Hình 6.2.1.a: Giao diện Tổng quan

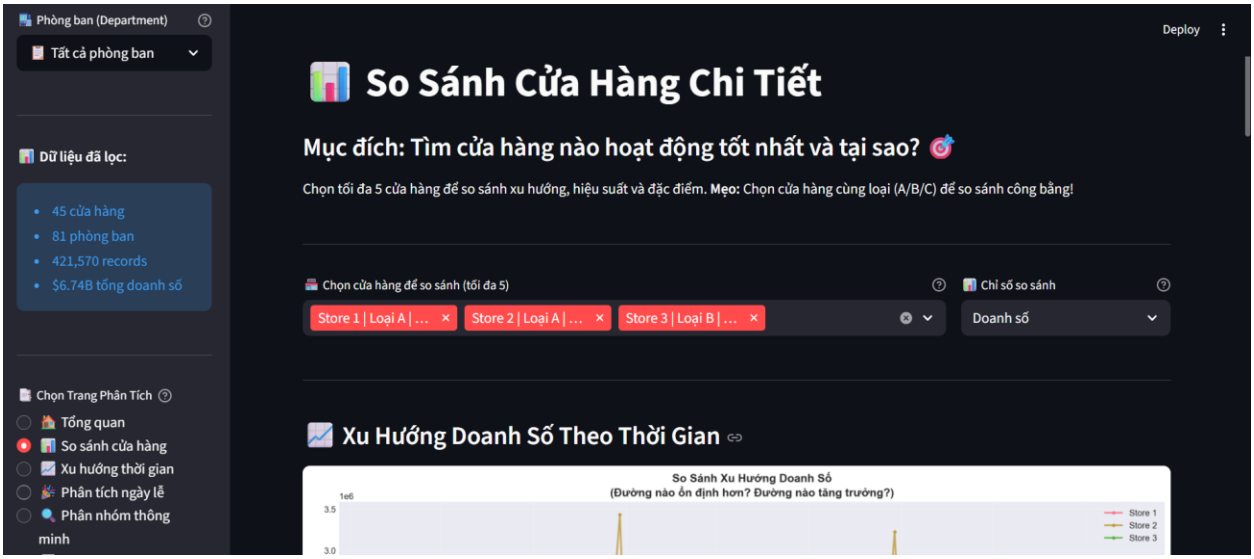
- **Chỉ số KPI:** Tổng doanh số, Doanh số trung bình tuần, Độ biến động (CV %), Tổng số cửa hàng.
- **Biểu đồ phân bố (Histogram):** Hiển thị tần suất các mức doanh số, so sánh với giá trị trung bình và trung vị để đánh giá độ lệch của dữ liệu.
- **Hiệu suất cửa hàng:** Bảng xếp hạng Top 5 cửa hàng tốt nhất và Bottom 5 cửa hàng cần cải thiện.
- **So sánh loại cửa hàng:** Biểu đồ cột so sánh hiệu quả giữa Type A, B và C.

Cách sử dụng: Xem lướt để nắm bắt tình hình chung và đọc các "Info box" để nhận khuyến nghị chiến lược tổng quát.

2.2. Tab 2: So Sánh Cửa Hàng

Chức năng: Phân tích sâu và so sánh trực tiếp hiệu suất giữa các cửa hàng cụ thể.

Cách sử dụng:



Hình 6.2.2.a: Giao diện So sánh Cửa hàng chi tiết

- Tại mục "Chọn cửa hàng", chọn tối đa 5 cửa hàng muốn so sánh.
- Chọn chỉ số so sánh (Doanh số, Độ ổn định, Xu hướng).

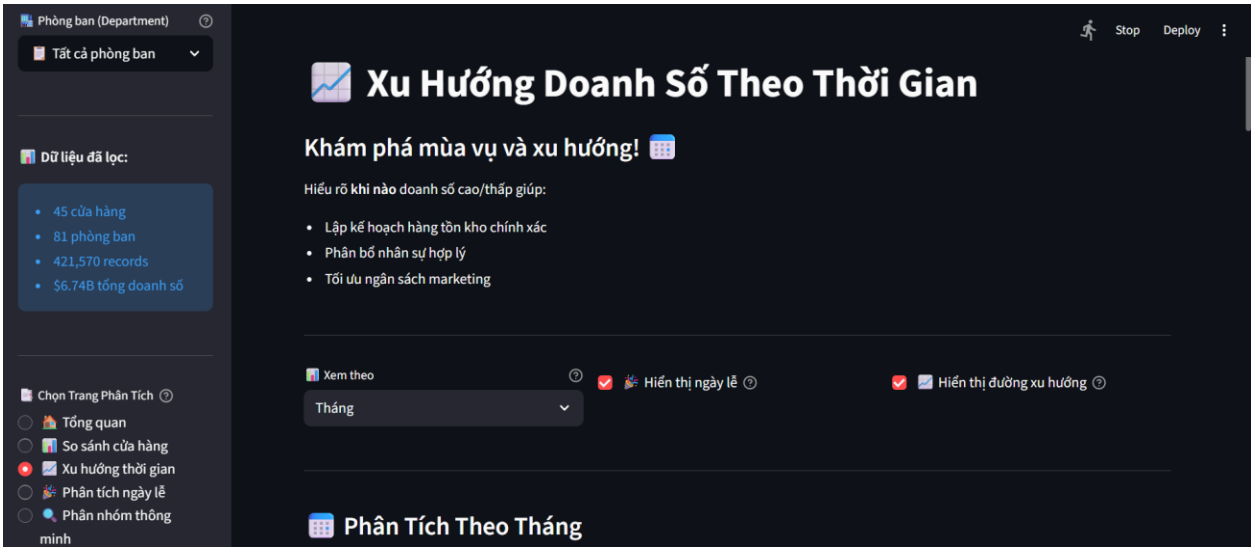
Kết quả:

- Biểu đồ đường (Line chart) thể hiện biến động doanh số theo thời gian thực của các cửa hàng đã chọn.
- Bảng thống kê chi tiết (Trung bình, Trung vị, Độ lệch chuẩn, CV%).
- Biểu đồ xếp hạng tổng hợp dựa trên công thức trọng số (50% Doanh số + 30% Ổn định + 20% Tiềm năng).

2.3. Tab 3: Xu Hướng Thời Gian

Chức năng: Phát hiện tính mùa vụ và xu hướng tăng trưởng.

Cách sử dụng:



Hình 6.2.3.a: Giao diện Xu hướng thời gian

- Chọn chế độ xem: **Tháng**, **Quý**, hoặc **Năm**.
- Tùy chọn hiển thị đường xu hướng (Trend line) hoặc đánh dấu ngày lễ.

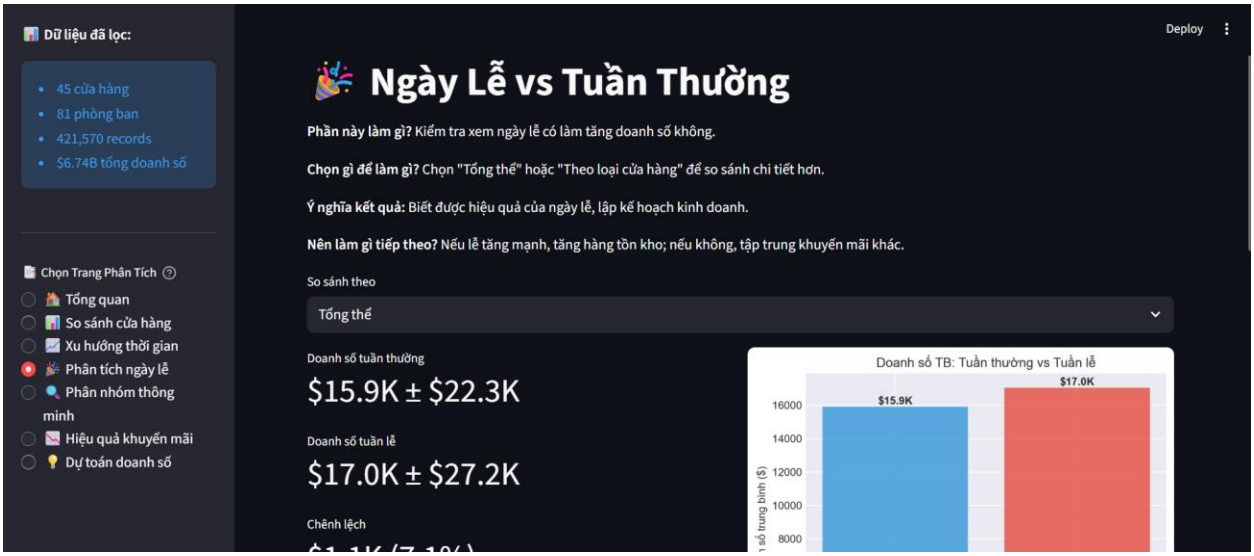
Kết quả:

- Biểu đồ cột/đường thể hiện doanh số theo khung thời gian chọn.
- Tự động tô màu vàng cho các tháng/quý cao điểm (ví dụ: Quý 4).
- **Heatmap (Bản đồ nhiệt):** (Khi chọn xem theo Tháng) Hiển thị "điểm nóng" doanh số theo sự kết hợp giữa Tháng và Tuần trong tháng.

2.4. Tab 4: Phân Tích Ngày Lễ

Chức năng: Kiểm định giả thuyết "Ngày lễ có thực sự làm tăng doanh số không?".

Cách sử dụng:



Hình 6.2.4.a: Giao diện Phân tích ngày lễ

Chọn phạm vi so sánh là "Tổng thể" hoặc "Theo loại cửa hàng".

Kết quả:

- So sánh giá trị trung bình giữa tuần lễ và tuần thường.
- Kết quả kiểm định thống kê **T-test** (P-value) để khẳng định sự khác biệt có ý nghĩa thống kê hay chỉ do ngẫu nhiên.
- Đưa ra kết luận chiến lược: Có nên đầu tư mạnh vào ngày lễ hay không.

2.5. Tab 5: Phân Nhóm Thông Minh (Clustering)

Chức năng: Phân khúc 45 cửa hàng thành các nhóm có đặc điểm tương đồng bằng thuật toán **K-Means**.

Cách sử dụng:

Hình 6.2.5.a: Giao diện Phân nhóm cửa hàng

- Chọn các đặc tính đầu vào (Doanh số, Quy mô, CPI, Thất nghiệp...).
- Tham khảo biểu đồ **Elbow Method** để chọn số cụm (K) tối ưu.
- Kéo thanh trượt để chọn số cụm K.

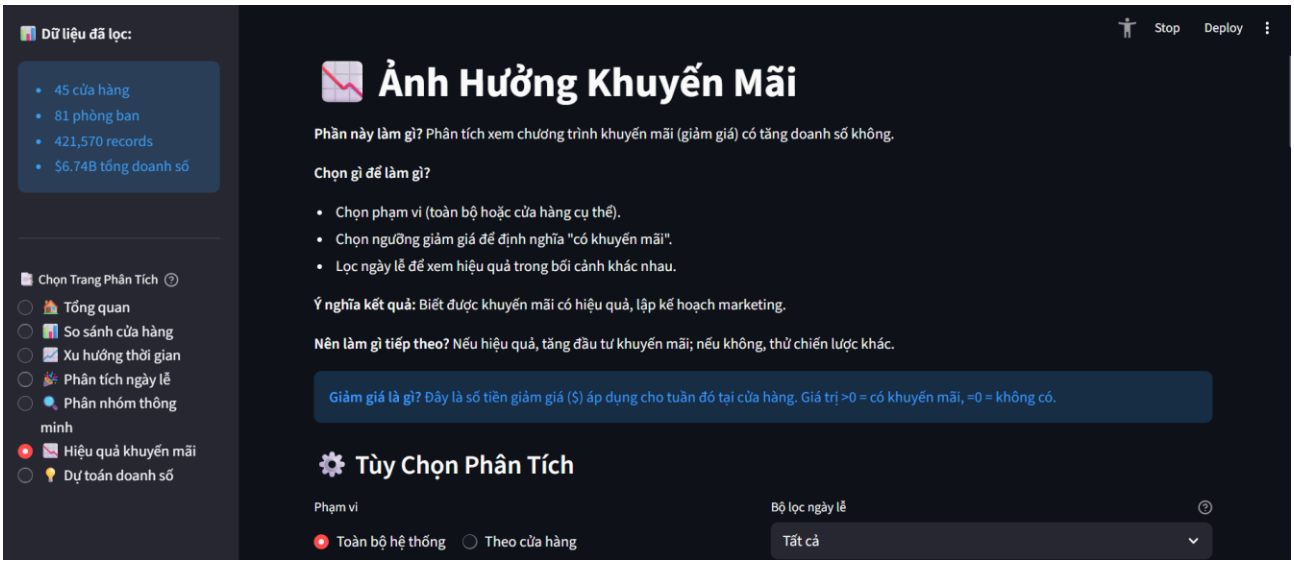
Kết quả:

- Chỉ số **Silhouette Score** đánh giá chất lượng phân cụm.
- Biểu đồ Scatter plot trực quan hóa các cụm (trục X là Size, trục Y là Sales).
- Bảng đặc điểm trung bình của từng cụm để gán nhãn (ví dụ: Cụm "Doanh số cao - Quy mô lớn").

2.6. Tab 6: Hiệu Quả Khuyến Mãi

Chức năng: Đánh giá tác động của chương trình giảm giá (MarkDown) lên doanh số.

Cách sử dụng:



Hình 6.2.6: Giao diện Ảnh hưởng khuyến mãi

- Chọn phạm vi (Toàn bộ hoặc từng Store).
- Đặt "Ngưỡng Giảm giá" (Threshold) để định nghĩa thế nào là "Có khuyến mãi".

Kết quả:

- So sánh doanh số trung bình khi Có KM vs Không KM.
- Tính chỉ số **Cohen's d** để đo lường độ lớn của tác động (Effect Size) và đưa ra khuyến nghị ROI.

2.7. Tab 7: Dự Toán Doanh Số (Interactive Forecast)

Chức năng: Công cụ mô phỏng "What-if Analysis" (Giả định tình huống). Không phải dự báo bằng AI, mà là tính toán dựa trên độ tương quan lịch sử.

Cách sử dụng:

Hình 6.2.7: Giao diện Dự đoán doanh số

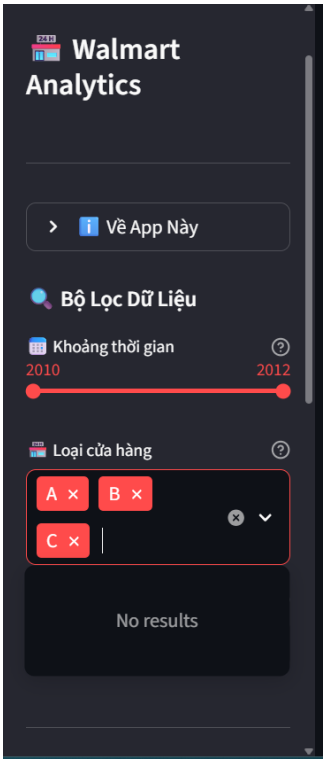
- Chọn dữ liệu nền (Baseline) từ một cửa hàng hoặc trung bình toàn hệ thống.
- Điều chỉnh các thanh trượt (Slider): Tăng/giảm Nhiệt độ, Giá xăng, CPI, Tỷ lệ thất nghiệp.
- Tích chọn: Có ngày lễ? Có khuyến mãi?

Kết quả:

- Tính toán % tác động tổng hợp.
- Hiển thị con số **Doanh Số Dự Toán** mới.
- Biểu đồ Tornado (Bar chart ngang) phân rã tác động của từng yếu tố (biết yếu tố nào ảnh hưởng mạnh nhất).

2.8. Sidebar: Bộ Lọc & Navigation

Nằm ở bên trái màn hình, luôn hiển thị cố định:



Hình 6.2.8.a: Giao diện Bộ lọc

Thông tin ứng dụng: Giới thiệu ngắn gọn.

Bộ Lọc Dữ Liệu (Global Filters): Áp dụng cho toàn bộ các trang phân tích:

- Khoảng thời gian (Năm).
- Loại cửa hàng (A, B, C).
- Phòng ban (Department): Có hiển thị xếp hạng (Ranking) và phân loại A/B/C/D ngay trong tên phòng ban.

Navigation (Menu): Các nút Radio để chuyển đổi giữa 7 trang phân tích nêu trên.

PHỤ LỤC

1. Danh sách tập tin nộp kèm (Files Included)

- **Source Code (PhanTichWalmart_Nhom 20.ipynb):** Jupyter Notebook chứa toàn bộ mã nguồn xử lý dữ liệu, trực quan hóa và huấn luyện mô hình Decision Tree.
- **Dữ liệu (thư mục data):** Tập dữ liệu gốc *Walmart Store Sales* được sử dụng trong đồ án.
- **Ứng dụng Streamlit (Nhom20_Walmart_App.py):** File Python chứa toàn bộ mã nguồn xử lý và hiện giao diện Streamlit.
- **requirement.txt:** File gồm các thư viên cần thiết để chạy Notebook và App.

2. Nguồn dữ liệu & Công cụ kỹ thuật

- **Dữ liệu:** Bộ dữ liệu *Walmart Store Sales* tải từ nền tảng Kaggle.
- **Thư viện lập trình:** Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn).
- **Tài liệu tham khảo:** Tài liệu kỹ thuật chính thức (Documentation) của thư viện Scikit-learn và các bài thảo luận về xử lý dữ liệu trên cộng đồng Kaggle.

3. Công cụ hỗ trợ: Nhóm có sử dụng công cụ AI (như ChatGPT/Gemini) với vai trò trợ lý học tập để:

- Gợi ý cú pháp lệnh và tra cứu thư viện Python nhanh.
- Hỗ trợ gỡ lỗi (debugging), tạo biểu đồ trong quá trình chạy code.
- Rà soát lỗi ngữ pháp và tối ưu văn phong báo cáo.