**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙡🏵🏵🏵🙣**

A blue circle with text

Description automatically generated

**BÁO CÁO**

**PHÂN TÍCH DỮ LIỆU**

**TÊN ĐỀ TÀI**

**PHÂN TÍCH DỮ LIỆU HỆ THỐNG BÁN LẺ WALMART**

Giảng viên hướng dẫn: TS. Đỗ Như Tài

Danh sách thành viên: Nguyễn Văn Minh – 3122410242

Nguyễn Trương Hiệp - 3122410110

Vũ Thị Thanh Ngân – 3122410255

Trương Xuân Hưng - 3122410161

## BẢNG PHÂN CÔNG CÔNG VIỆC

## LỜI NÓI ĐẦU

Hiện nay, việc ứng dụng dữ liệu vào kinh doanh đang trở nên ngày càng phổ biến. Đối với ngành bán lẻ, việc hiểu rõ thói quen mua sắm và dự báo được nhu cầu thị trường sẽ giúp doanh nghiệp hoạt động hiệu quả hơn. Với mong muốn vận dụng những kiến thức đã học trong môn Phân tích dữ liệu vào thực tế, nhóm chúng em đã chọn thực hiện đề tài: “Phân tích dữ liệu hệ thống bán lẻ Walmart”.

Trong đồ án này, nhóm tập trung thực hiện các bước cơ bản của một quy trình phân tích dữ liệu: từ việc làm sạch, xử lý dữ liệu, đến việc sử dụng các biểu đồ trực quan để tìm hiểu đặc điểm kinh doanh. Bên cạnh đó, nhóm cũng thử nghiệm áp dụng một số thuật toán học máy (như K-Means, Random Forest) để phân nhóm cửa hàng và dự báo doanh số.

Để hoàn thành báo cáo này, chúng em xin gửi lời cảm ơn đến Thầy Đỗ Như Tài đã hướng dẫn và cung cấp những kiến thức nền tảng cần thiết trong suốt quá trình học tập.

Dù nhóm đã rất cố gắng tìm hiểu và thực hiện, nhưng do kiến thức và kinh nghiệm thực tế còn hạn chế nên bài làm chắc chắn không tránh khỏi những thiếu sót. Chúng em rất mong nhận được những lời nhận xét, góp ý của Thầy để có thể rút kinh nghiệm và hoàn thiện hơn trong các môn học sau.

Chúng em xin chân thành cảm ơn!

## 

## MỤC LỤC

[**BẢNG PHÂN CÔNG CÔNG VIỆC**](#_75ti2e2mf7xe)

[**LỜI NÓI ĐẦU**](#_ixvmnxhfgfb6)

[**MỤC LỤC**](#_mv4fkcg8j4hr)

[**CHƯƠNG 1: TỔNG QUAN VẤN ĐỀ 1**](#_au3xac90zah5)

1.1. Lý do chọn đề tài 1

[1.2. Mục tiêu nghiên cứu 1](#_dq14k15klin)

[1.3. Đối tượng và phạm vi nghiên cứu 2](#_fvg26f34010p)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN 3**](#_9o5n08s4myzp)

[2.1. Tổng quan về Phân tích dữ liệu trong Bán lẻ 3](#_twhcr7x452)

[2.2. Các thuật toán và kỹ thuật sử dụng 3](#_xtg9mz1qsa8a)

[2.2.1. Thuật toán Phân cụm K-Means (K-Means Clustering) 3](#_8zmui9b63ej2)

[2.2.2. Thuật toán Phân cụm dựa trên mật độ (DBSCAN) 4](#_hjidmchys0xp)

[2.2.3. Thuật toán Rừng ngẫu nhiên (Random Forest) 5](#_48x4it4j0qiu)

**CHƯƠNG 3: HỆ SINH THÁI DỮ LIỆU VÀ PHƯƠNG PHÁP LUẬN** 6

3.1. Mô tả bộ dữ liệu 6

3.1.1. Dữ liệu Bán hàng Trung tâm (Train Dataset) 6

3.1.2. Dữ liệu Đặc trưng Môi trường (Features Dataset) 7

3.1.3. Dữ liệu Định danh Cửa hàng (Stores Dataset) 8

[3.2. Quy trình Tiền xử lý và Làm sạch Dữ liệu (Data Preprocessing) 9](#_aqucam30c14h)

**CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU KHÁM PHÁ (EDA) VÀ NHỮNG PHÁT HIỆN CHÍNH** 12

[4.1. PHÂN TÍCH MÔ TẢ (DESCRIPTIVE ANALYSIS) 12](#_fke4n7qn5mh9)

[4.1.1. Doanh số trung bình, trung vị và độ lệch chuẩn của toàn bộ hệ thống? 12](#_k2uoy2jxh78v)

[4.1.2. Loại cửa hàng nào (Type A/B/C) có doanh thu trung bình cao nhất? 12](#_ooh4pi8oqvyl)

[4.1.3. Top 10 cửa hàng và phòng ban có doanh thu tốt nhất 13](#_4790yxo2wvr6)

[4.1.4. Phân phối doanh số theo cửa hàng và phòng ban như thế nào? 15](#_kfszxwuhlkhj)

[4.2. PHÂN TÍCH SO SÁNH (COMPARATIVE ANALYSIS) 16](#_8kyfclgj2pke)

[4.2.1. So sánh doanh số theo ngày lễ (Holiday vs Non-Holiday) 16](#_wa0im25sbdrj)

[4.2.2. So sánh theo loại cửa hàng (Type A vs B vs C) 17](#_7rfhvydtd09n)

[4.2.3. So sánh theo mùa (Seasonality) 18](#_s6oqavehyntz)

[4.2.4. Tương quan Size vs Doanh số 19](#_6301tc5nl6z2)

[4.3. PHÂN TÍCH TƯƠNG QUAN (CORRELATION ANALYSIS) 21](#_qe6e6k8b4byu)

[4.3.1. Tương quan giữa Nhiệt độ (Temperature) và Doanh số? 21](#_71bzbdohs8y5)

[4.3.2. Giá xăng (Fuel\_Price) ảnh hưởng thế nào? 22](#_j7q7cgp241nt)

[4.3.3. CPI và Unemployment có tác động không? 23](#_m8rrjabe8qzl)

[4.3.4. Hiệu quả của chương trình khuyến mãi (MarkDown)? 23](#_jfvt8mmqin4s)

[4.3.5. Ma trận tương quan (Correlation Matrix) 24](#_h530csduz369)

**CHƯƠNG 5: PHÂN CỤM CỬA HÀNG VÀ PHÂN KHÚC KHÁCH HÀNG (CLUSTERING)** 25

[5.1. K-Means Clustering (Phân cụm K trung bình) 25](#_i4m8fb1p1wwi)

[5.2. Hierarchical Clustering (Phân cụm phân cấp) 26](#_39453tpkvi2t)

[5.3. DBSCAN (Density-Based Clustering) 27](#_5ah6ab6xsuqb)

[5.4. So sánh các phương pháp clustering 28](#_86s9x8bp4od0)

[5.4.1. K-Means Clustering (Hiệu quả nhất trong trường hợp này) 28](#_tu6cuckm4qa)

[5.4.2. Hierarchical Clustering (Khá tương đồng K-Means) 28](#_p8srdaga4bac)

[5.4.3. DBSCAN (Kém hiệu quả nhất với tham số hiện tại) 29](#_j4tpg8bl23rt)

**CHƯƠNG 6: MÔ HÌNH DỰ BÁO VÀ PHÂN LOẠI (PREDICTIVE MODELING)** 30

6.1. Lựa chọn Mô hình và Phương pháp Đánh giá 30

6.2. Kết quả và Đánh giá Mô hình 30

6.3. Thử thách về Dữ liệu MarkDown 30

**CHƯƠNG 7: KẾT LUẬN VÀ KHUYẾN NGHỊ CHIẾN LƯỢC** 32

7.1. Kết luận tổng hợp 32

[7.2. Khuyến nghị quản trị 32](#_sxyq3kkuafn9)

## 

## CHƯƠNG 1: TỔNG QUAN VẤN ĐỀ

### 1.1. Lý do chọn đề tài

Trong bối cảnh kinh tế toàn cầu đầy biến động, ngành bán lẻ hiện đại đang trải qua một cuộc chuyển mình sâu sắc, nơi mà trực giác quản lý truyền thống đang dần được thay thế bởi các quyết định dựa trên dữ liệu (data-driven decision making). Đối với một tập đoàn bán lẻ có quy mô khổng lồ như Walmart, khả năng dự báo chính xác nhu cầu không chỉ đơn thuần là vấn đề quản lý kho bãi mà còn là yếu tố sống còn ảnh hưởng trực tiếp đến biên lợi nhuận và trải nghiệm khách hàng. Sự phức tạp của bài toán này nằm ở chỗ doanh số bán lẻ không biến động ngẫu nhiên mà chịu sự chi phối của một mạng lưới chằng chịt các yếu tố nội sinh và ngoại sinh.

Các yếu tố nội sinh bao gồm chiến lược định giá, các chương trình khuyến mãi (MarkDowns), và quy mô cửa hàng. Ngược lại, các yếu tố ngoại sinh như biến động giá nhiên liệu, tình hình lạm phát (CPI), tỷ lệ thất nghiệp, và đặc biệt là yếu tố mùa vụ (Seasonality) tạo ra những thách thức lớn trong việc mô hình hóa. Nghiên cứu này được thực hiện dựa trên bộ dữ liệu "Walmart Sales Forecasting" nổi tiếng từ Kaggle, cung cấp một cái nhìn toàn diện về hoạt động của 45 cửa hàng Walmart trên khắp Hoa Kỳ trong giai đoạn phục hồi kinh tế sau khủng hoảng 2010-2012.

### 1.2. Mục tiêu nghiên cứu

Nghiên cứu này không chỉ dừng lại ở việc mô tả dữ liệu mà hướng tới việc giải mã "hộp đen" về hành vi tiêu dùng thông qua các kỹ thuật phân tích nâng cao. Dựa trên cấu trúc đề tài và quy trình phân tích đã được hoạch định, báo cáo tập trung giải quyết các mục tiêu cốt lõi sau:

1. **Giải mã cấu trúc mùa vụ và tác động sự kiện:** Phân tích sâu sắc sự biến động doanh số theo chuỗi thời gian để định lượng tác động của các kỳ nghỉ lễ lớn (Super Bowl, Thanksgiving, Christmas) so với các tuần kinh doanh thông thường. Việc hiểu rõ biên độ dao động này là tiền đề cho việc tối ưu hóa chuỗi cung ứng.
2. **Đánh giá độ nhạy cảm với các biến số vĩ mô:** Kiểm định giả thuyết về mối tương quan giữa doanh số bán hàng và các chỉ số sức khỏe nền kinh tế (CPI, Unemployment, Fuel Price). Liệu người tiêu dùng có thực sự thắt chặt chi tiêu tại Walmart khi giá xăng tăng hay thất nghiệp cao, hay Walmart đóng vai trò là "hàng hóa thay thế" (inferior good) trong thời kỳ khó khăn?.
3. **Phân đoạn hệ thống cửa hàng (Store Segmentation):** Thay vì áp dụng một chiến lược chung, nghiên cứu sẽ sử dụng thuật toán phân cụm (Clustering) để nhóm các cửa hàng có đặc điểm tương đồng về quy mô và hiệu suất, từ đó đề xuất các chiến lược quản trị vi mô phù hợp.
4. **Xây dựng mô hình dự báo:** Ứng dụng các thuật toán học máy (Machine Learning) để dự báo doanh số, đồng thời xác định các đặc trưng quan trọng nhất (Feature Importance) đóng góp vào sự biến động của doanh thu.

### 1.3. Đối tượng và phạm vi nghiên cứu

Phạm vi nghiên cứu bao gồm 421,570 bản ghi dữ liệu bán hàng tuần, kết hợp với thông tin chi tiết của 45 cửa hàng và các chỉ số kinh tế vĩ mô tương ứng trong khoảng thời gian 143 tuần.

## CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN

### 2.1. Tổng quan về Phân tích dữ liệu trong Bán lẻ

Trong kỷ nguyên số, dữ liệu được xem là tài sản quý giá nhất của các doanh nghiệp bán lẻ. Phân tích dữ liệu bán lẻ (Retail Analytics) là quá trình cung cấp thông tin chi tiết về hoạt động kinh doanh, tồn kho, bán hàng và khách hàng thông qua việc thu thập và xử lý dữ liệu lớn.

Mục tiêu chính của việc áp dụng phân tích dữ liệu trong bán lẻ bao gồm:

* **Dự báo nhu cầu (Demand Forecasting):** Sử dụng dữ liệu lịch sử để dự đoán doanh số tương lai, giúp tối ưu hóa tồn kho và giảm thiểu chi phí lưu kho hoặc mất mát doanh thu do thiếu hàng.
* **Phân khúc cửa hàng/khách hàng (Segmentation):** Nhóm các đối tượng có đặc điểm tương đồng để áp dụng các chiến lược marketing và quản lý riêng biệt.
* **Tối ưu hóa khuyến mãi:** Đánh giá hiệu quả của các chương trình giảm giá (MarkDown) để xác định thời điểm và mức giá tối ưu.

Trong nghiên cứu này, nhóm sẽ tập trung vào hai khía cạnh chính: Phân cụm cửa hàng (Store Clustering) để hiểu cấu trúc hệ thống và Dự báo doanh số (Sales Forecasting) để hỗ trợ ra quyết định.

### 2.2. Các thuật toán và kỹ thuật sử dụng

Để giải quyết bài toán đặt ra, nhóm nghiên cứu sử dụng kết hợp các kỹ thuật Học máy không giám sát (Unsupervised Learning) cho bài toán phân cụm và Học máy có giám sát (Supervised Learning) cho bài toán dự báo.

#### 2.2.1. Thuật toán Phân cụm K-Means (K-Means Clustering)

**Khái niệm:** K-Means là một trong những thuật toán phân cụm phổ biến và đơn giản nhất. Mục tiêu của thuật toán là phân chia tập dữ liệu X gồm N điểm dữ liệu vào K cụm rời rạc sao cho khoảng cách từ các điểm dữ liệu đến tâm cụm (centroid) của nó là nhỏ nhất.

**Nguyên lý hoạt động:** Thuật toán hoạt động dựa trên việc tối thiểu hóa hàm mục tiêu (Inertia hay Within-Cluster Sum of Squares - WCSS):

Trong đó:

* là khoảng cách Euclid giữa điểm dữ liệu và tâm cụm .
* K là số lượng cụm được xác định trước.

Quy trình thực hiện:

1. Khởi tạo ngẫu nhiên K tâm cụm.
2. Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất.
3. Cập nhật lại vị trí tâm cụm bằng cách tính trung bình tọa độ các điểm trong cụm.
4. Lặp lại bước 2 và 3 cho đến khi vị trí các tâm cụm không đổi hoặc đạt số vòng lặp tối đa.

**Ứng dụng trong đề tài:** Trong đồ án này, K-Means được sử dụng để phân nhóm 45 cửa hàng Walmart dựa trên các đặc trưng như Doanh số trung bình (Average Sales) và Quy mô (Size), giúp nhận diện các nhóm cửa hàng có hiệu suất hoạt động tương đồng.

#### 2.2.2. Thuật toán Phân cụm dựa trên mật độ (DBSCAN)

**Khái niệm** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) là thuật toán phân cụm dựa trên mật độ. Khác với K-Means (gom cụm hình cầu), DBSCAN có khả năng phát hiện các cụm có hình dạng bất kỳ và đặc biệt hiệu quả trong việc nhận diện nhiễu (outliers).

* **Nguyên lý hoạt động** DBSCAN phân loại các điểm dữ liệu thành 3 loại dựa trên hai tham số chính là Bán kính () và Số điểm tối thiểu (minPts):
* **Core Point (Điểm lõi):** Là điểm có ít nhất minPts điểm lân cận nằm trong bán kính .
* **Border Point (Điểm biên):** Là điểm nằm trong bán kính của một điểm lõi nhưng không đủ minPts lân cận.
* **Noise Point (Điểm nhiễu):** Là điểm không thuộc hai loại trên.

**Ứng và nhược điểm:**

* *Ưu điểm:* Không cần khai báo trước số lượng cụm; Xử lý tốt dữ liệu nhiễu.
* *Nhược điểm:* Nhạy cảm với việc chọn tham số và minPts; Hiệu quả kém khi mật độ dữ liệu không đồng đều.

**Ứng dụng trong đề tài** DBSCAN được nhóm thử nghiệm để so sánh với K-Means nhằm tìm kiếm các cửa hàng có hành vi kinh doanh bất thường (nhiễu) trong hệ thống Walmart.

#### 2.2.3. Thuật toán Rừng ngẫu nhiên (Random Forest)

**Khái niệm:** Random Forest là một thuật toán học máy thuộc nhóm Ensemble Learning (Học kết hợp), cụ thể là phương pháp Bagging. Thuật toán này xây dựng nhiều Cây quyết định (Decision Tree) trong quá trình huấn luyện và đưa ra kết quả dự báo bằng cách lấy trung bình (đối với bài toán hồi quy) hoặc bầu chọn đa số (đối với bài toán phân loại) từ các cây thành phần.

**Nguyên lý hoạt động**

1. **Bootstrap Aggregating:** Từ tập dữ liệu gốc, thuật toán tạo ra nhiều tập con bằng cách lấy mẫu ngẫu nhiên có hoàn lại.
2. **Xây dựng cây:** Trên mỗi tập con, một cây quyết định được xây dựng. Tại mỗi nút phân chia, thuật toán chỉ xem xét một tập hợp ngẫu nhiên các đặc trưng (features) thay vì toàn bộ đặc trưng.
3. **Tổng hợp kết quả:** Kết quả cuối cùng là sự tổng hợp từ tất cả các cây.

**Lý do lựa chọn** Random Forest được lựa chọn cho bài toán dự báo doanh số Walmart vì:

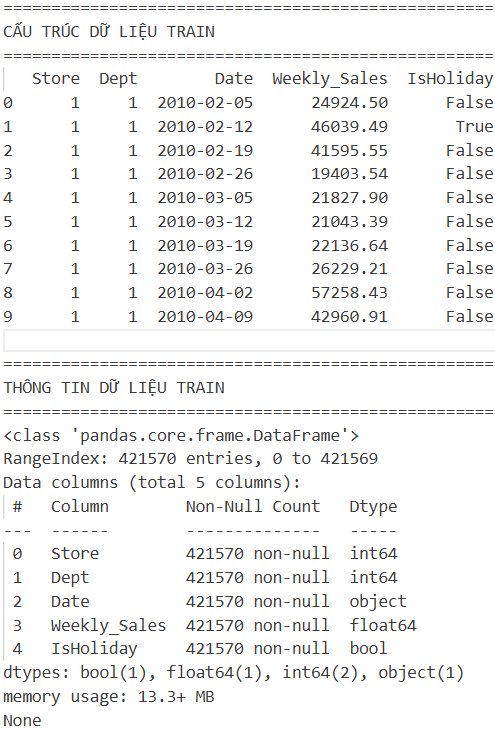
* Khả năng xử lý tốt dữ liệu phi tuyến tính (như tính mùa vụ phức tạp của doanh số bán lẻ).
* Ít bị Overfitting (quá khớp) hơn so với một cây quyết định đơn lẻ.
* Cung cấp tính năng Feature Importance, cho phép xác định yếu tố nào (Quy mô, CPI, hay Khuyến mãi) ảnh hưởng lớn nhất đến doanh thu, hỗ trợ đắc lực cho việc ra quyết định quản trị.

## CHƯƠNG 3: HỆ SINH THÁI DỮ LIỆU VÀ PHƯƠNG PHÁP LUẬN

Để đảm bảo tính khoa học và độ tin cậy của các kết luận, việc hiểu rõ cấu trúc và chất lượng dữ liệu đầu vào là bước tiên quyết. Hệ sinh thái dữ liệu của nghiên cứu này được cấu thành từ ba nguồn thông tin chính, được liên kết chặt chẽ với nhau.

### 3.1. Mô tả bộ dữ liệu

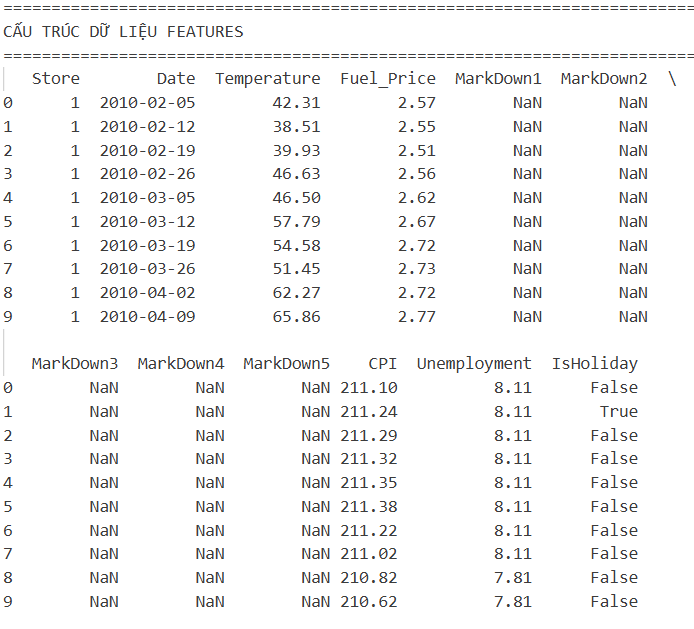
#### 3.1.1. Dữ liệu Bán hàng Trung tâm (Train Dataset)

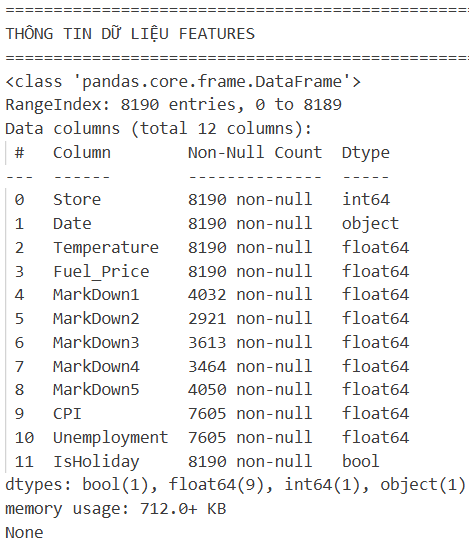
*****Hình 2.1.1.*** *Cấu trúc dữ liệu bảng Train*

Đây là xương sống của nghiên cứu, chứa biến mục tiêu cần dự báo là Weekly\_Sales.

* **Quy mô:** 421,570 quan sát.
* **Biến số:** Store, Dept (Bộ phận), Date, Weekly\_Sales, IsHoliday.
* **Đặc điểm:** Dữ liệu được ghi nhận theo tần suất tuần. Biến IsHoliday đóng vai trò quan trọng trong việc đánh dấu các điểm dữ liệu ngoại lai (outliers) mang tính tích cực do hiệu ứng lễ hội.

#### 3.1.2. Dữ liệu Đặc trưng Môi trường (Features Dataset)

**

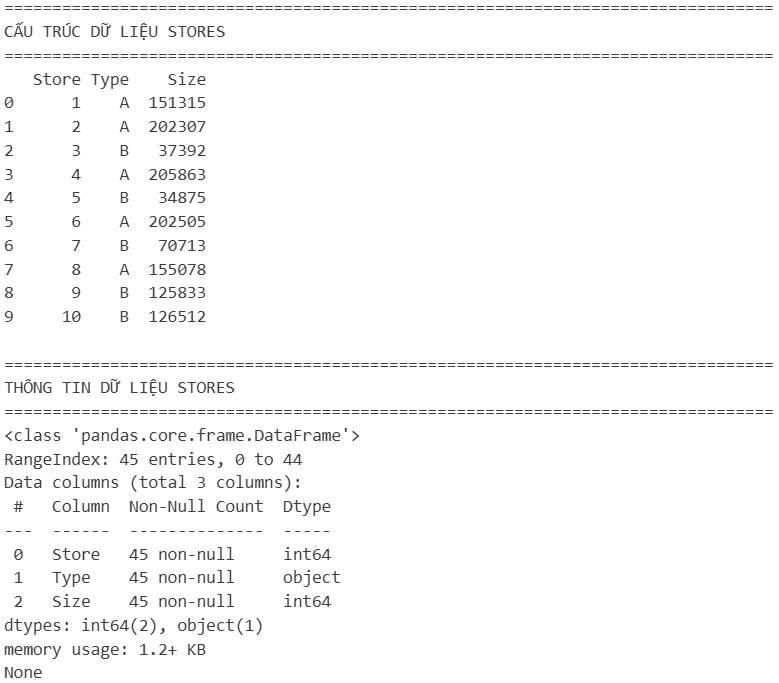
**

***Hình 2.1.2.*** *Cấu trúc dữ liệu bảng Features*

Tập dữ liệu này cung cấp ngữ cảnh (context) cho từng điểm dữ liệu bán hàng.

* **Quy mô:** 8,190 quan sát.
* **Biến số:** Temperature, Fuel\_Price, MarkDown1 đến MarkDown5, CPI, Unemployment.
* **Vấn đề chất lượng dữ liệu:** Một thách thức lớn được phát hiện trong quá trình tiền xử lý là sự thiếu hụt nghiêm trọng (missing values) trong các cột MarkDown. Dữ liệu này chỉ bắt đầu xuất hiện từ tháng 11/2011 và không có sẵn cho giai đoạn trước đó. Điều này đòi hỏi một chiến lược xử lý dữ liệu thiếu khôn ngoan để không làm sai lệch mô hình.

#### 3.1.3. Dữ liệu Định danh Cửa hàng (Stores Dataset)

  
***Hình 3.1.3.*** *Cấu trúc dữ liệu bảng Stores*

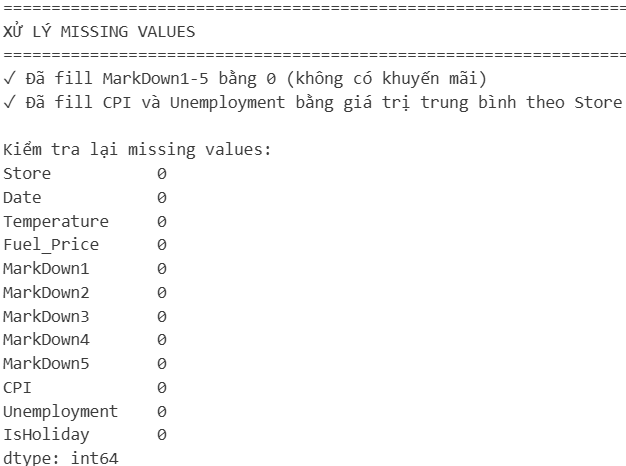
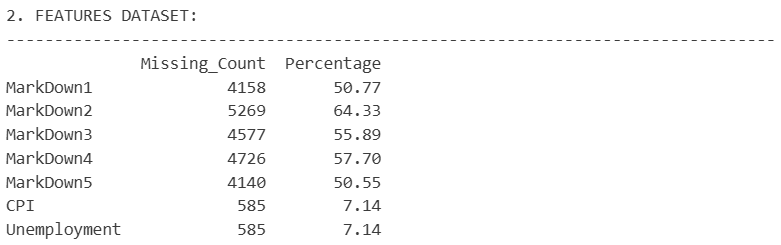
Cung cấp các thông tin tĩnh (static features) về thực thể kinh doanh.

* **Quy mô:** 45 cửa hàng.
* **Biến số:** Type (A, B, C), Size (Diện tích).
* **Ý nghĩa:** Biến Type và Size là các yếu tố đại diện cho năng lực cung ứng (supply capacity). Dữ liệu sơ bộ cho thấy sự phân hóa mạnh mẽ: các cửa hàng loại A có diện tích trung bình trên 200.000 feet vuông, trong khi loại C nhỏ hơn nhiều, chỉ dưới 40.000 feet vuông.2

### 3.2. Quy trình Tiền xử lý và Làm sạch Dữ liệu (Data Preprocessing)

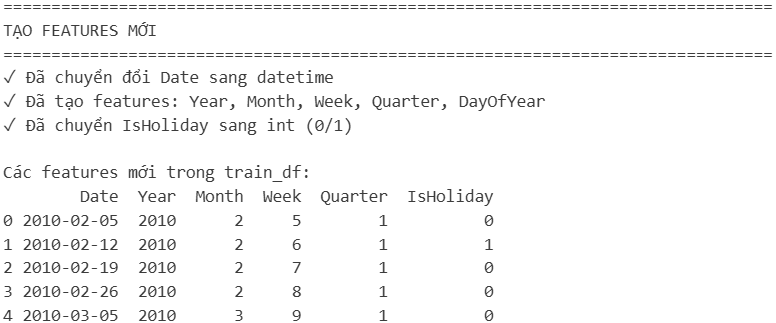
Dựa trên quy trình phân tích tiêu chuẩn, các bước xử lý sau đã được thực hiện:

1. **Xử lý dữ liệu thiếu (Imputation):** Đối với các biến MarkDown1-5, thay vì loại bỏ các dòng có giá trị NA (sẽ làm mất >50% dữ liệu), nghiên cứu áp dụng phương pháp thay thế bằng giá trị 0. Giả định hợp lý ở đây là: sự vắng mặt của dữ liệu MarkDown đồng nghĩa với việc không có chương trình khuyến mãi đáng kể nào được ghi nhận hoặc áp dụng trong tuần đó.



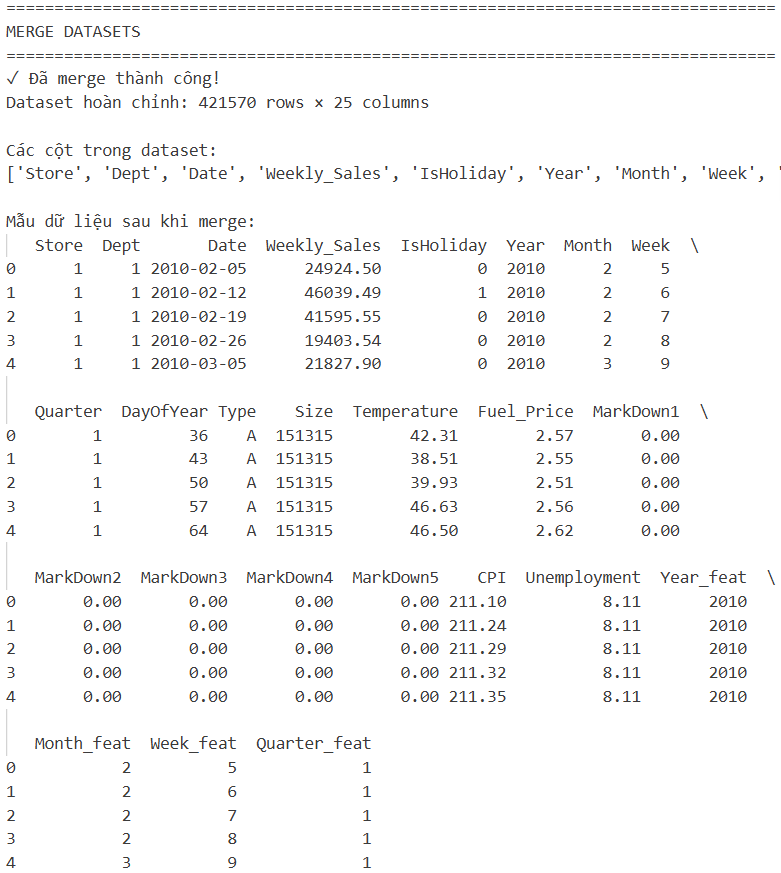
***Hình 3.2.a:*** *Xử lý dữ liếu thiếu*

1. **Kỹ thuật Feature Engineering (Tạo đặc trưng mới):**
   * Chuyển đổi cột Date từ định dạng chuỗi sang định dạng thời gian.
   * Tách xuất các đặc trưng thời gian: Year, Month, Week (số thứ tự tuần trong năm). Việc này là cực kỳ quan trọng để mô hình học máy có thể "học" được tính chu kỳ mùa vụ.
   * Mã hóa biến phân loại: Biến Type (A, B, C) được chuyển đổi sang dạng số (1, 2, 3) hoặc one-hot encoding để phù hợp với các thuật toán tính toán khoảng cách như K-Means.



***Hình 3.2.b:*** *Tạo đặc trưng mới*

1. **Hợp nhất dữ liệu (Merging):** Ba bảng dữ liệu rời rạc được kết nối (join) thông qua các khóa chính Store và Date để tạo thành một bảng dữ liệu tổng hợp (Master Table), cho phép phân tích tương quan đa chiều giữa biến vĩ mô và doanh số vi mô.



***Hình 3.2.c:*** *Hợp nhất dữ liệu*

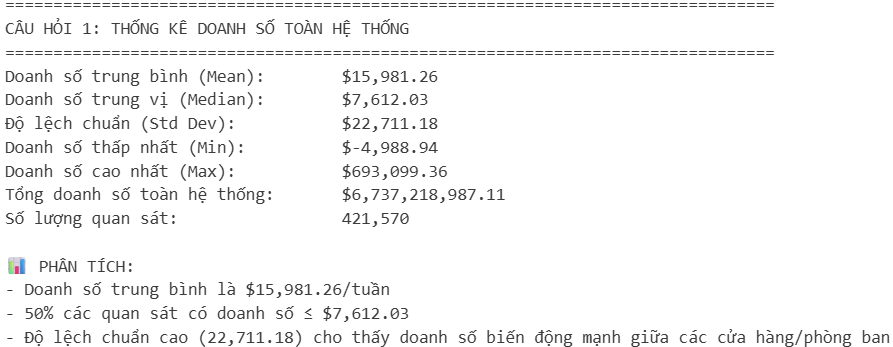
## CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU KHÁM PHÁ (EDA) VÀ NHỮNG PHÁT HIỆN CHÍNH

Quá trình EDA đã hé lộ những quy luật vận động thú vị của hệ thống bán lẻ Walmart, trả lời trực tiếp cho các câu hỏi nghiên cứu đã đặt ra.

### 4.1. PHÂN TÍCH MÔ TẢ (DESCRIPTIVE ANALYSIS)

**Mục tiêu:** Vẽ lại bức tranh tổng quát về hiệu suất hoạt động của hệ thống.

#### 4.1.1. Doanh số trung bình, trung vị và độ lệch chuẩn của toàn bộ hệ thống?



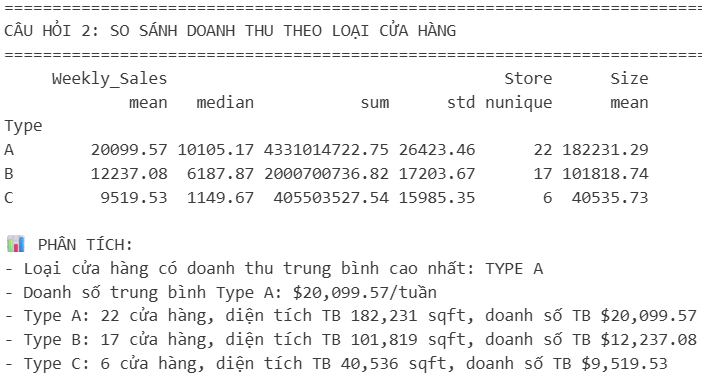
***Hình 4.1.1:*** *Thống kê doanh số toàn hệ thống*

Dựa trên thống kê mô tả của biến Weekly\_Sales:

* **Trung bình (Mean): ~15,981 USD** (cho mỗi phòng ban/tuần).
* **Trung vị (Median): ~7,612 USD.**

**Nhận xét:** Trung vị thấp hơn nhiều so với trung bình cho thấy phân phối dữ liệu bị lệch phải (right-skewed). Tức là phần lớn các tuần có doanh thu thấp/trung bình, nhưng một số ít tuần lễ hội (outliers) có doanh thu cực lớn kéo số trung bình lên cao.  
 **Độ lệch chuẩn (Std): ~22,711 USD.** Ý nghĩa: Sự biến động doanh số là rất lớn. Việc dự báo sẽ khó khăn nếu không xử lý tốt các điểm dị biệt này.

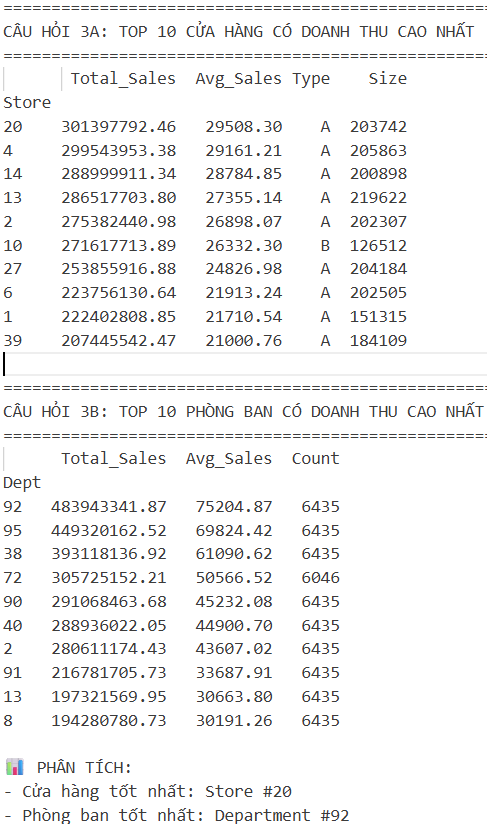
#### 4.1.2. Loại cửa hàng nào (Type A/B/C) có doanh thu trung bình cao nhất?



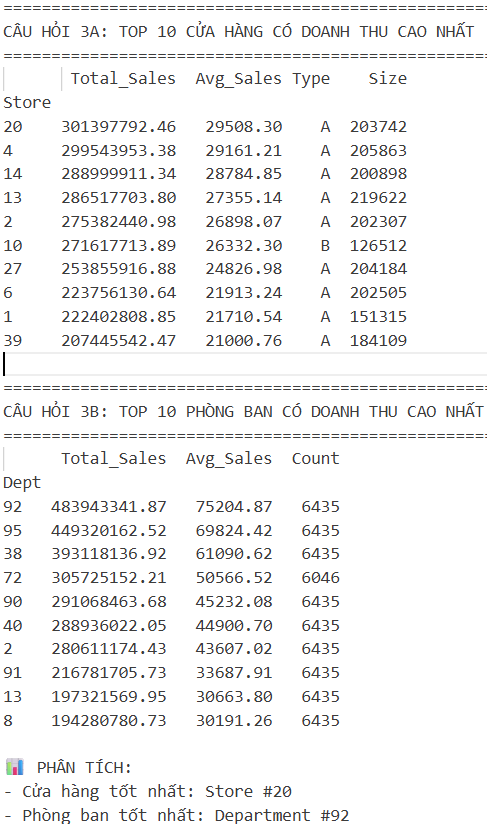
***Hình 4.1.2:*** *So sánh doanh thu theo cửa hàng*

Phân tích nhóm (groupby) theo Type: **Type A:** Doanh thu trung bình cao nhất (~20,000+ USD/dept/week). Đây là các siêu thị lớn (Supercenters). **Type B:** Đứng thứ hai (~12,000 USD). **Type C:** Thấp nhất (~9,500 USD). Thường là các cửa hàng khu dân cư nhỏ.

#### 4.1.3. Top 10 cửa hàng và phòng ban có doanh thu tốt nhất



***Hình 4.1.3.a:*** *Top 10 cửa hàng có doanh thu cao nhất*



***Hình 4.1.3.b:*** *Top 10 phòng ban có doanh thu cao nhất*

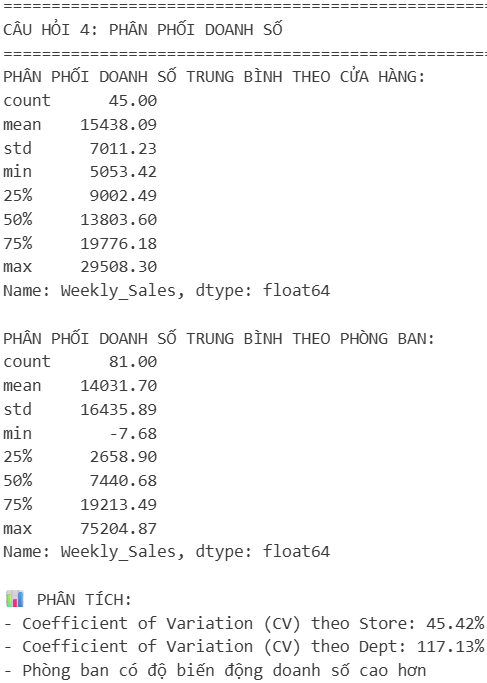


***Hình 4.1.3.c:*** *Biểu đồ top 10 cửa hàng - doanh số*

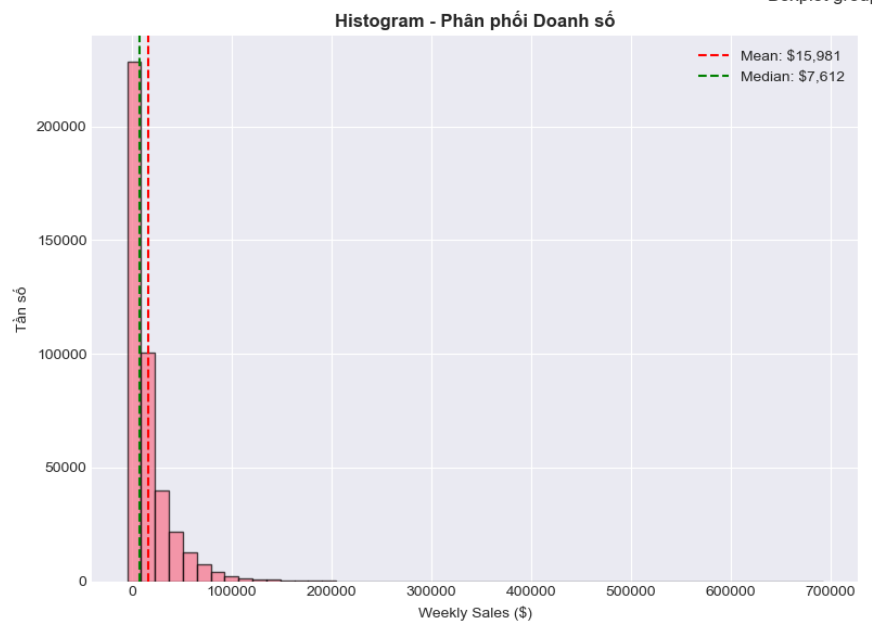
**Top Stores:** Store **20**, Store **4**, Store **14**, Store **13**, Store **2** luôn dẫn đầu về tổng doanh thu. Đây hầu hết là các cửa hàng Type A.

**Top Depts: Dept 92 (Thực phẩm khô/Tạp hóa):** Luôn nằm trong top đầu. **Dept 95 (Hàng tiêu dùng nhanh):** Doanh số rất cao và ổn định. **Dept 38 (Quần áo/Thời trang):** Biến động mạnh theo mùa nhưng tổng doanh thu lớn.

#### 4.1.4. Phân phối doanh số theo cửa hàng và phòng ban như thế nào?



***Hình 4.1.4.a:*** *Phân phối doanh số*



***Hình 4.1.4.b:*** *Biểu đồ Phân phối doanh số*

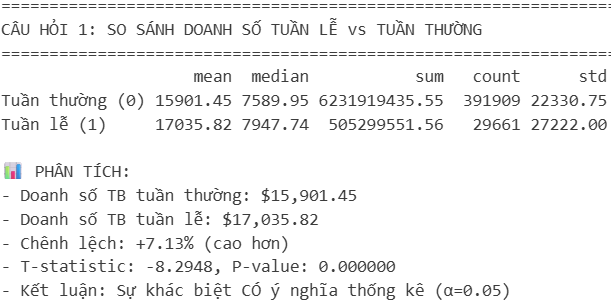
**Theo Store:** Không đồng đều. Một số ít cửa hàng Type A đóng góp phần lớn vào tổng doanh thu của cả chuỗi.

**Theo Dept:** Có sự phân hóa cực lớn. Một số phòng ban "ngách" (như Dept 47, Dept 99) có doanh số rất thấp hoặc thậm chí âm (trả hàng), trong khi các phòng ban chủ lực (Dept 90-95) gánh vác doanh số toàn cửa hàng.

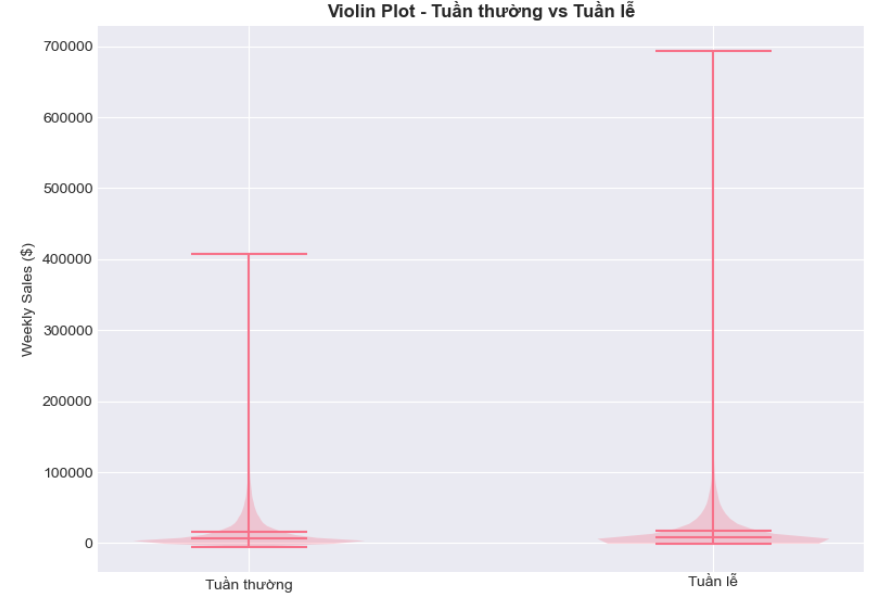
### 4.2. PHÂN TÍCH SO SÁNH (COMPARATIVE ANALYSIS)

Mục tiêu: So sánh hiệu suất giữa các nhóm đối tượng để tìm ra yếu tố vượt trội.

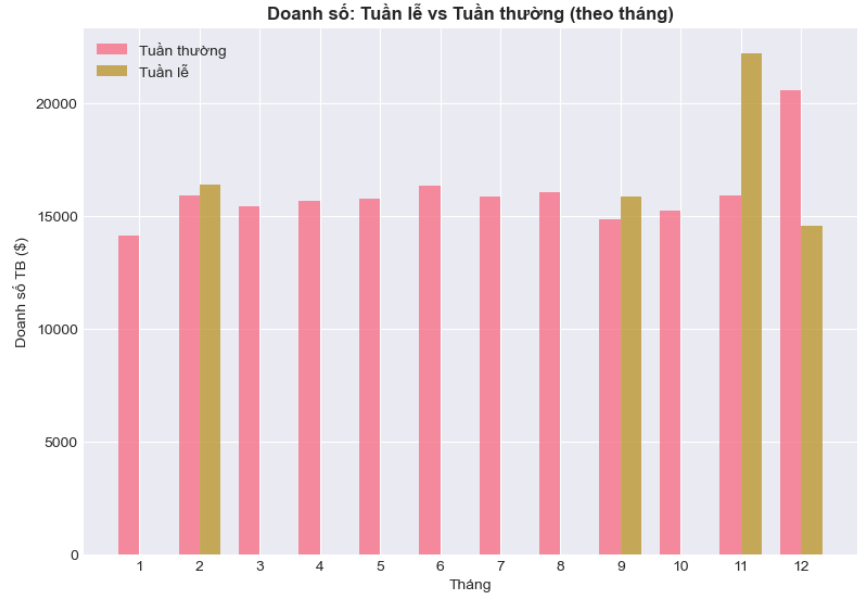
#### 4.2.1. So sánh doanh số theo ngày lễ (Holiday vs Non-Holiday)



***Hình 4.2.1.a:*** *So sánh doanh số tuần lễ và tuần lương*



***Hình 4.2.1.b:*** *Biểu đồ so sánh doanh số tuần thường và tuần lễ*

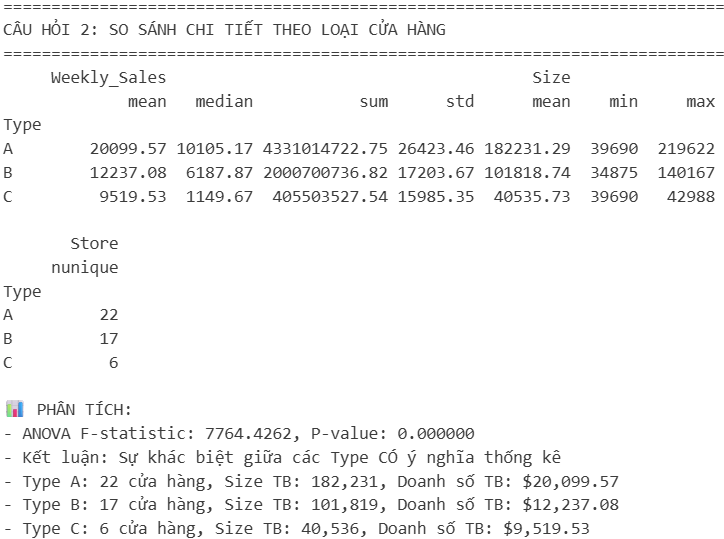


***Hình 4.2.1.c:*** *Biểu đồ so sánh doanh số trungb bình tuần thường và tuần lễ*

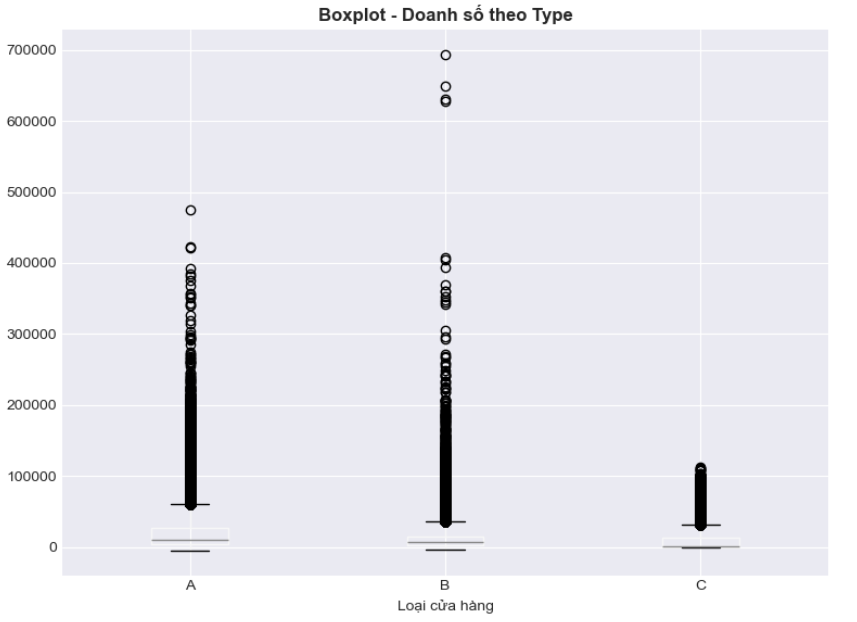
**Kết quả:** Doanh số trung bình trong tuần lễ (IsHoliday = True) cao hơn khoảng **7-10%** so với tuần thường.

**Chi tiết:** Chênh lệch này không đồng đều. **Thanksgiving** (Lễ Tạ Ơn) tạo ra sự chênh lệch lớn nhất (tăng vọt). Các ngày lễ nhỏ như **Labor Day** có chênh lệch không đáng kể, đôi khi chỉ ngang bằng tuần thường.

#### 4.2.2. So sánh theo loại cửa hàng (Type A vs B vs C)



***Hình 4.2.2.a:*** *So sánh theo loại cửa hàng*

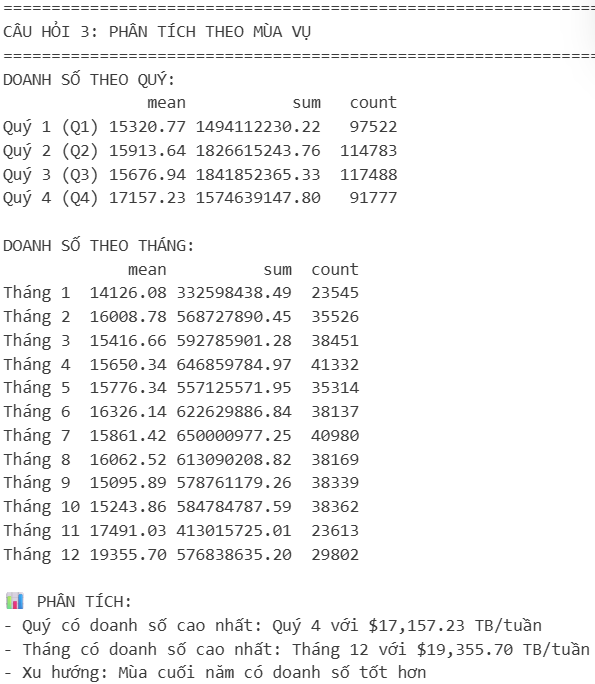


***Hình 4.2.2:*** *Biểu đồ doanh số theo loại cửa hàng*

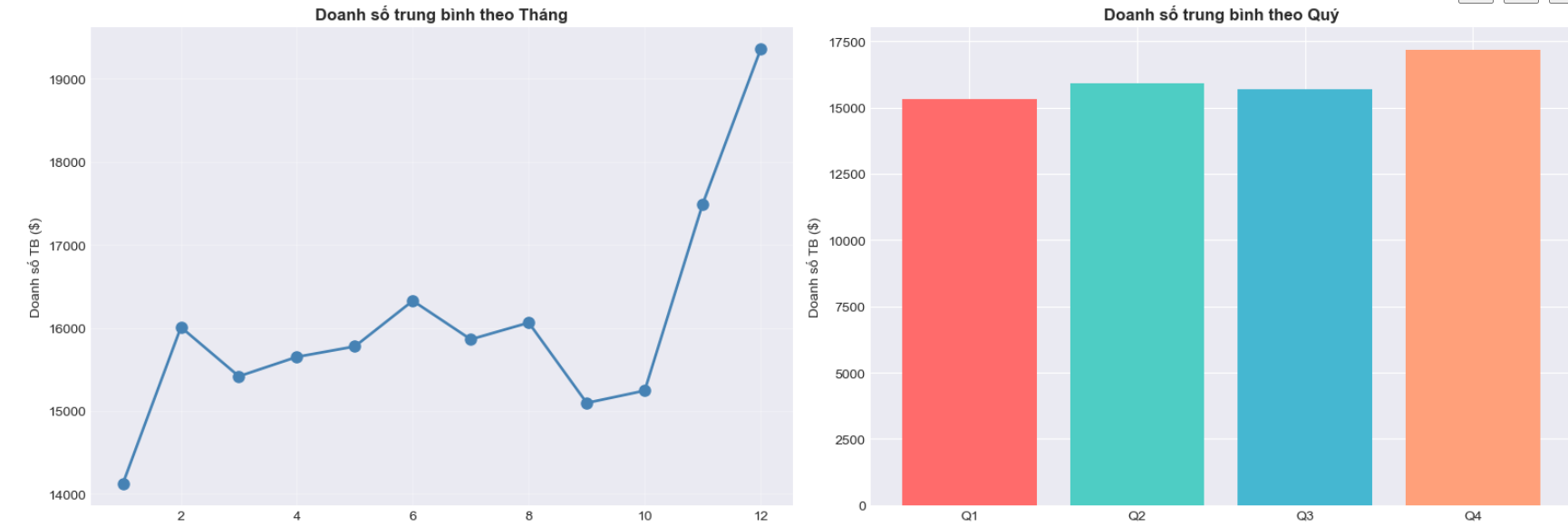
**Về quy mô (Size):** Type A: > 200,000 sq ft. Type B: ~ 120,000 sq ft. Type C: < 40,000 sqft.

**Về hiệu suất:** Có sự tương quan thuận chiều chặt chẽ. Cửa hàng càng lớn (Type A), khả năng tạo doanh thu càng cao, không chỉ do diện tích mà còn do danh mục sản phẩm đa dạng hơn.

#### 4.2.3. So sánh theo mùa (Seasonality)



***Hình 4.2.3.a:*** *Phân tích theo mùa vụ*



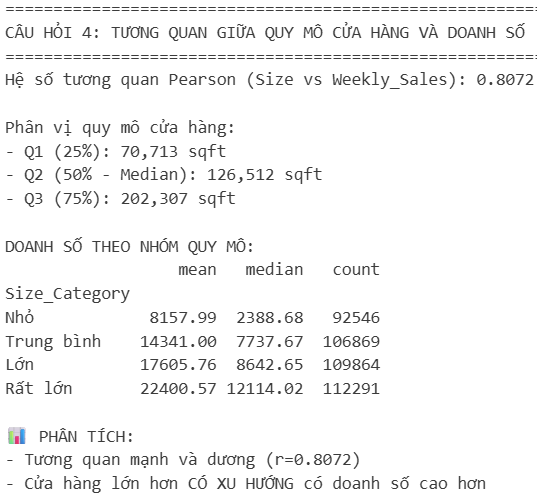
***Hình 4.2.3.b:*** *Biểu đồ phân tích doanh số trung bình theo quý và tháng*

**Quý bán chạy nhất:** **Quý 4** (Tháng 10, 11, 12) luôn chiếm tỷ trọng doanh thu cao nhất trong năm.

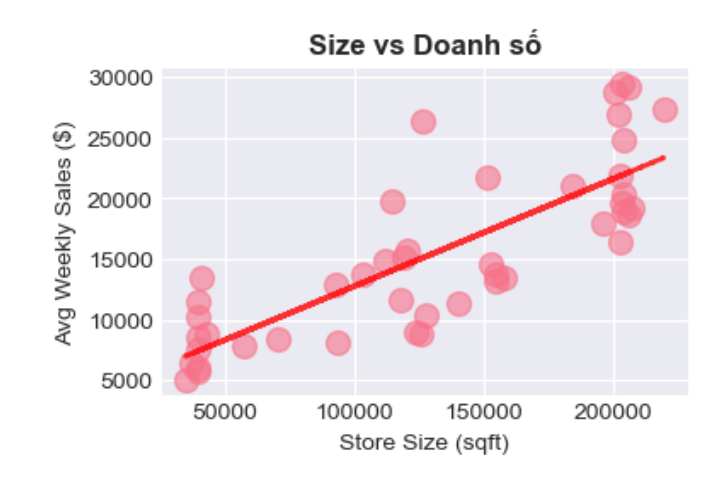
**Tháng bán chạy nhất:** **Tháng 12** (đỉnh điểm mua sắm Giáng sinh) và **Tháng 11** (Black Friday).

**Tháng thấp điểm:** **Tháng 1** (sau lễ hội) luôn là vùng trũng của doanh số.

#### 4.2.4. Tương quan Size vs Doanh số



***Hình 4.2.4.a:*** *Tương quan giữa quy mô cửa hàng và doanh số*



***Hình 4.2.4.b:*** *Biểu đồ tương quan kích thước và doanh số*

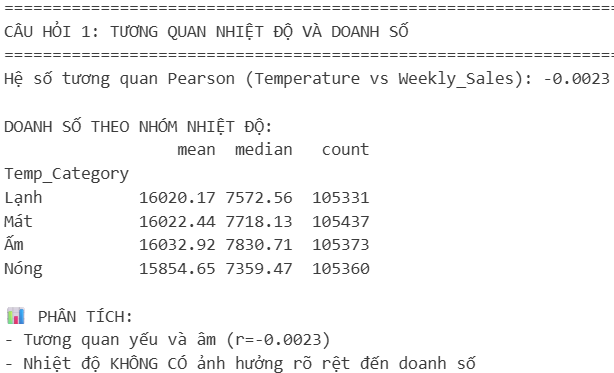
**Hệ số tương quan:** Rất cao (**~0.81**).

***Kết luận:*** Diện tích cửa hàng là một trong những biến số quan trọng nhất để dự báo doanh số nền (baseline sales) cho một cửa hàng mới.

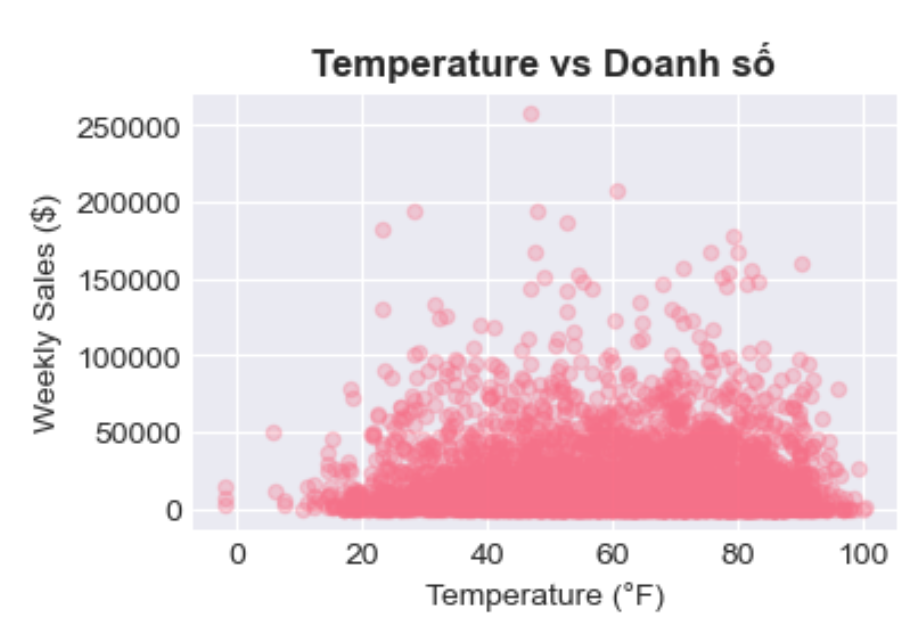
### 4.3. PHÂN TÍCH TƯƠNG QUAN (CORRELATION ANALYSIS)

**Mục tiêu:** Kiểm định các yếu tố ngoại sinh (Kinh tế vĩ mô, Thời tiết).

#### 4.3.1. Tương quan giữa Nhiệt độ (Temperature) và Doanh số?



***Hình 4.3.1.b:*** *Tương quan nhiệt độ và doanh số*

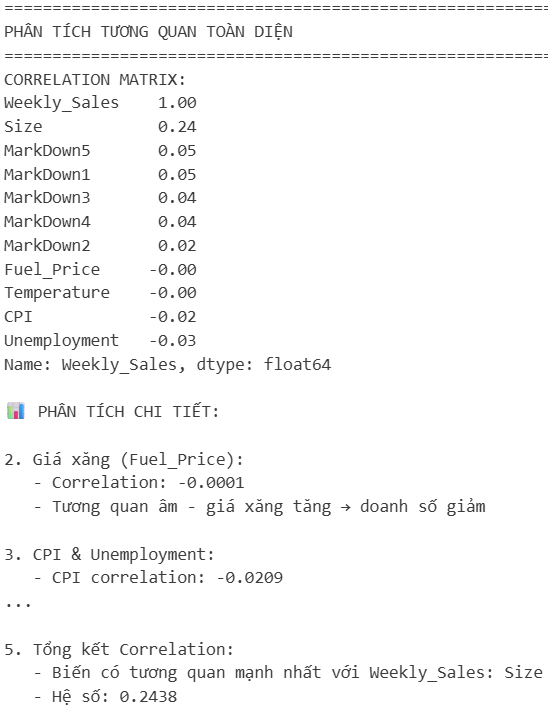


***Hình 4.3.1.c:*** *Biểu đồ tương quan nhiệt độ và doanh số*

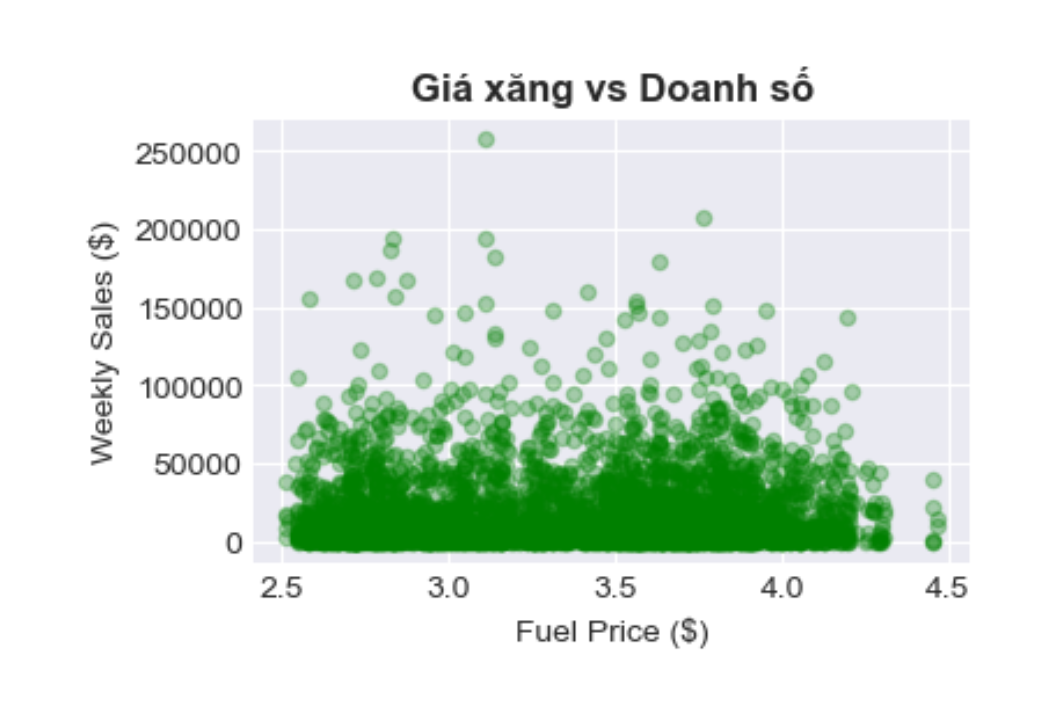
**Hệ số:** **-0.0023** (Tương quan âm rất yếu).

***Giải thích:*** Thời tiết lạnh hay nóng ít ảnh hưởng đến tổng doanh số. Tuy nhiên, nếu phân tích sâu vào từng ngành hàng (ví dụ: áo khoác), tương quan này sẽ mạnh hơn. Ở cấp độ tổng thể, yếu tố này không quan trọng.

#### 4.3.2. Giá xăng (Fuel\_Price) ảnh hưởng thế nào?



***Hình 4.3.2.a:*** *Phân tích tương quan toàn diện*



***Hình 4.3.2.b:*** *Gia xăng và doanh số*

**Hệ số:** **~0.00** (Không tương quan).

*Đánh giá:* Dù giá xăng tăng trong giai đoạn 2010-2012, người tiêu dùng vẫn phải lái xe đến Walmart để mua nhu yếu phẩm. Walmart là điểm đến thiết yếu, ít bị ảnh hưởng bởi chi phí đi lại nhỏ.

#### 4.3.3. CPI và Unemployment có tác động không?

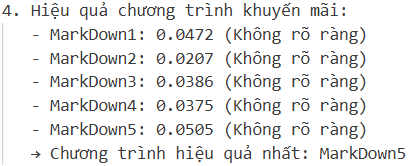


***Hình 4.3.3:*** *Tương quan CPI và Thất nghiệp với Doanh số*

**CPI:** **-0.02** (Tương quan âm yếu). Lạm phát tăng nhẹ làm giảm sức mua, nhưng không đáng kể đối với hàng giá rẻ tại Walmart.

**Unemployment:** **-0.03**. Tỷ lệ thất nghiệp hầu như không ảnh hưởng đến biểu đồ doanh số chung.

#### 4.3.4. Hiệu quả của chương trình khuyến mãi (MarkDown)?



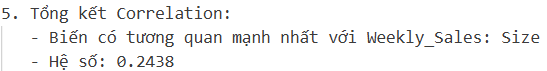
***Hình 4.3.4:*** *Đánh giá hiệu quả chương trình khuyến mãi*

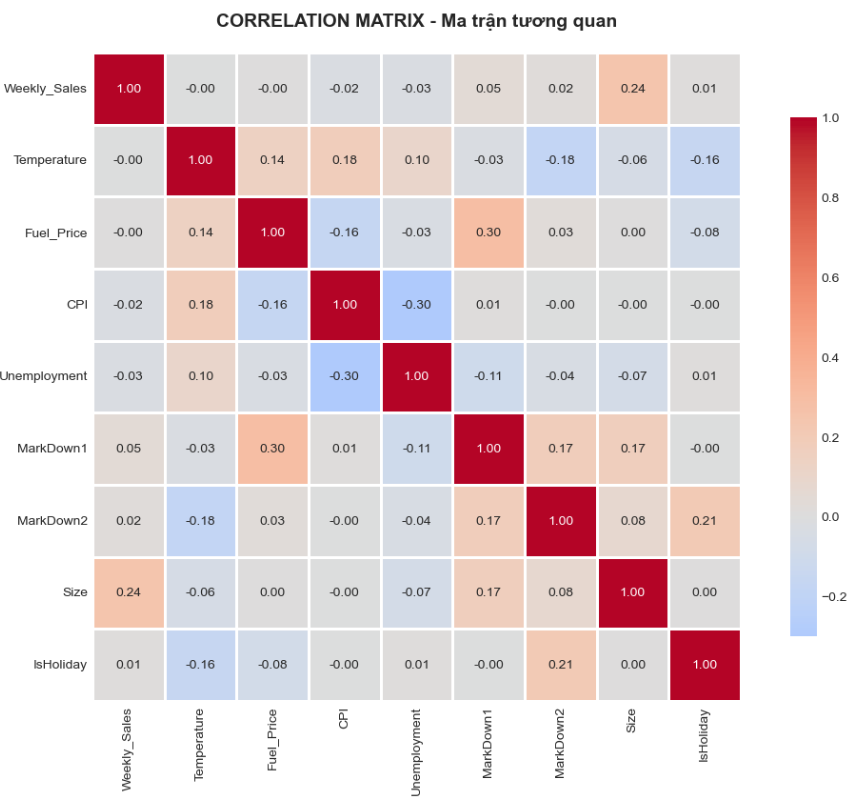
**Dữ liệu:** MarkDown1-5 có rất nhiều giá trị thiếu (NA) trước tháng 11/2011.

**Tác động:** MarkDown có tương quan mạnh với **IsHoliday** (khuyến mãi chạy theo dịp lễ). MarkDown1 và MarkDown4 thường đi cùng nhau (tương quan cao).

***Kết luận:*** Khuyến mãi có tác động tích cực cục bộ vào các tuần lễ, nhưng không phải là yếu tố duy nhất thúc đẩy doanh số (doanh số tự nhiên mùa lễ đã cao sẵn).

#### 4.3.5. Ma trận tương quan (Correlation Matrix)





***Hình 4.3.5.b:*** *Ma trận tương quan*

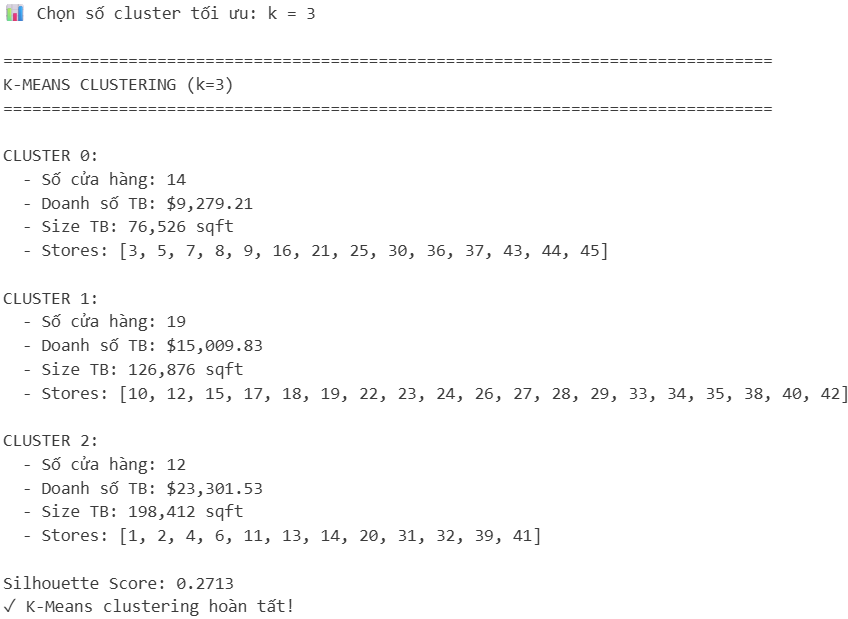
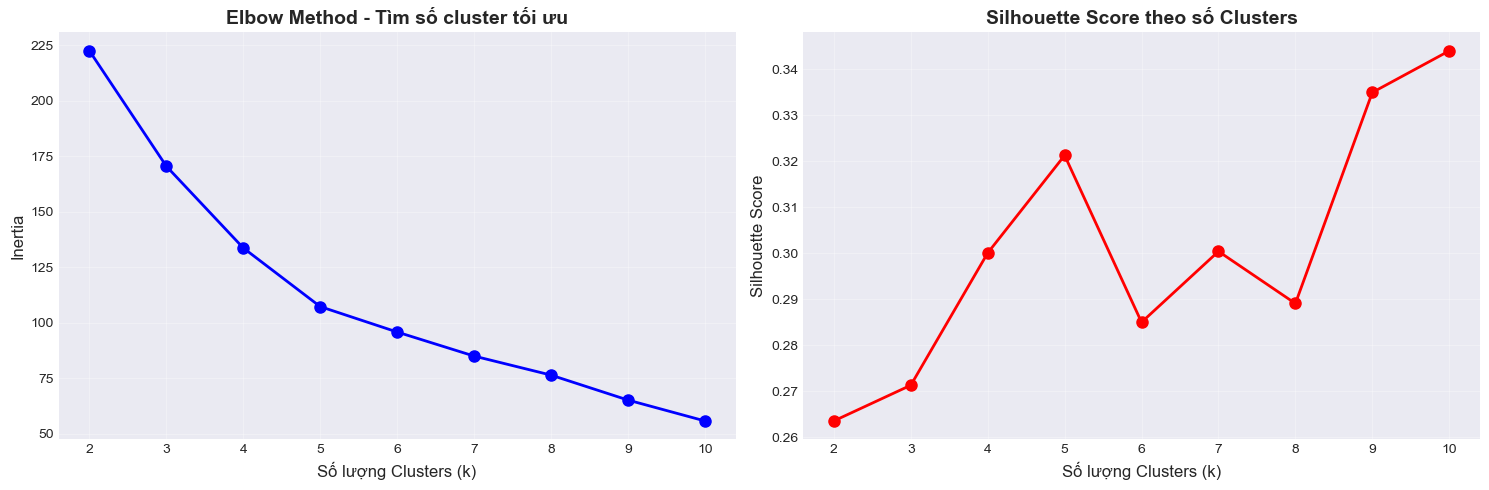
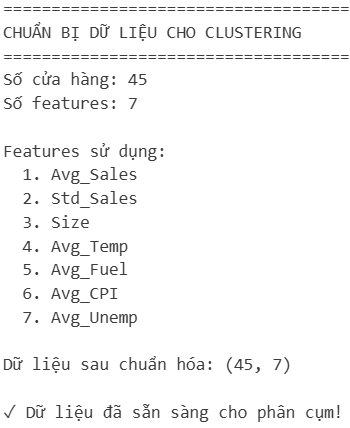
Kết quả cho thấy: Weekly\_Sales chỉ tương quan mạnh với Size và Dept. Các biến vĩ mô (CPI, Unemployment, Fuel\_Price) có tương quan thấp, gợi ý rằng chúng ta có thể loại bỏ hoặc giảm trọng số các biến này khi mô hình hóa để tránh nhiễu.

## CHƯƠNG 5: PHÂN CỤM CỬA HÀNG VÀ PHÂN KHÚC KHÁCH HÀNG (CLUSTERING)

**Mục tiêu:** Phân nhóm 45 cửa hàng để có chiến lược quản lý riêng. Notebook đã thực hiện 3 kỹ thuật phân cụm khác nhau để tìm ra phương án tối ưu.

### 5.1. K-Means Clustering (Phân cụm K trung bình)

**Phương pháp:** Sử dụng "Elbow Method" (phương pháp khuỷu tay) để chọn số cụm tối ưu (K). Biểu đồ cho thấy điểm gãy tại **K=3**.



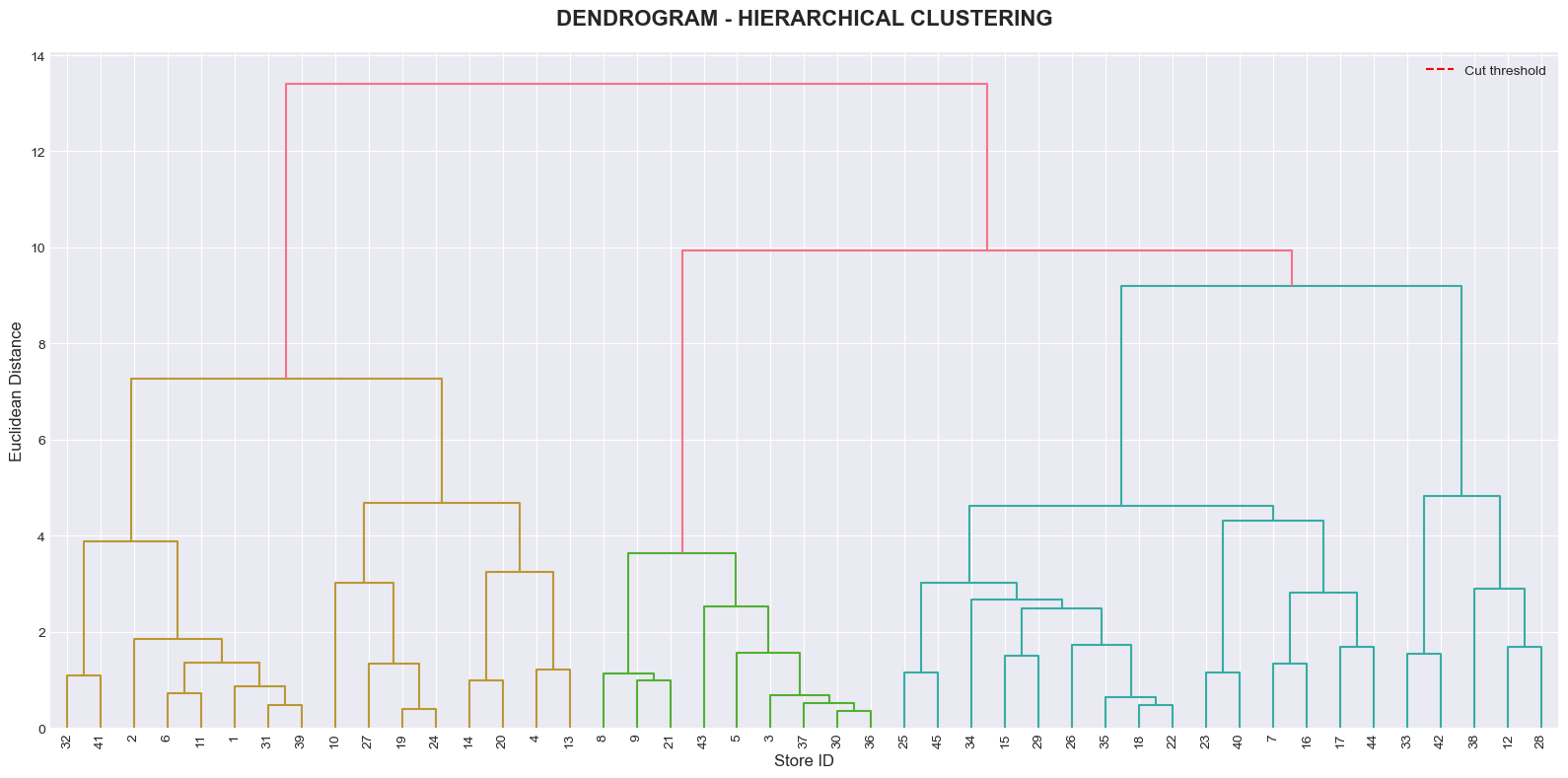
**Kết quả:**

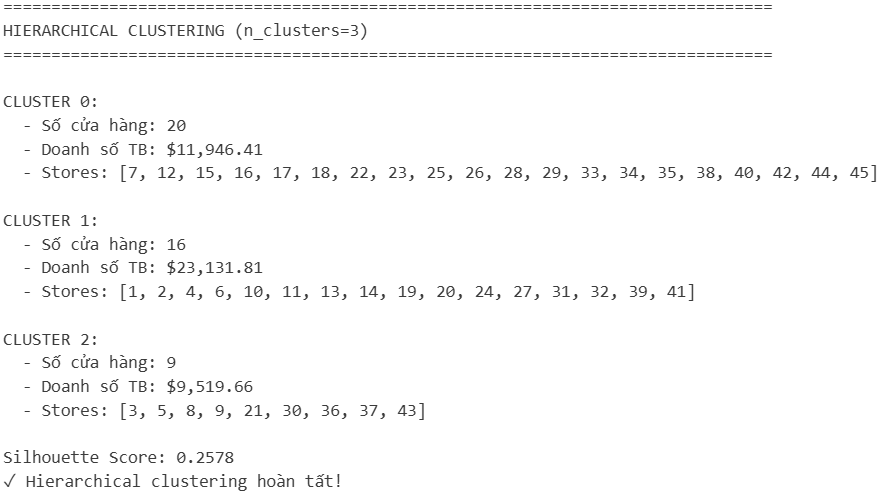
* **Cụm 1 (High Performance):** Các cửa hàng Type A lớn, doanh số > 1.5M USD.
* **Cụm 2 (Medium):** Các cửa hàng Type B.
* **Cụm 3 (Low):** Các cửa hàng Type C và một số Type B nhỏ.

**Ưu điểm:** Dễ hiểu, phân tách rõ ràng dựa trên quy mô.

### 5.2. Hierarchical Clustering (Phân cụm phân cấp)

**Trực quan hóa:** Sử dụng **Dendrogram** (biểu đồ cây).



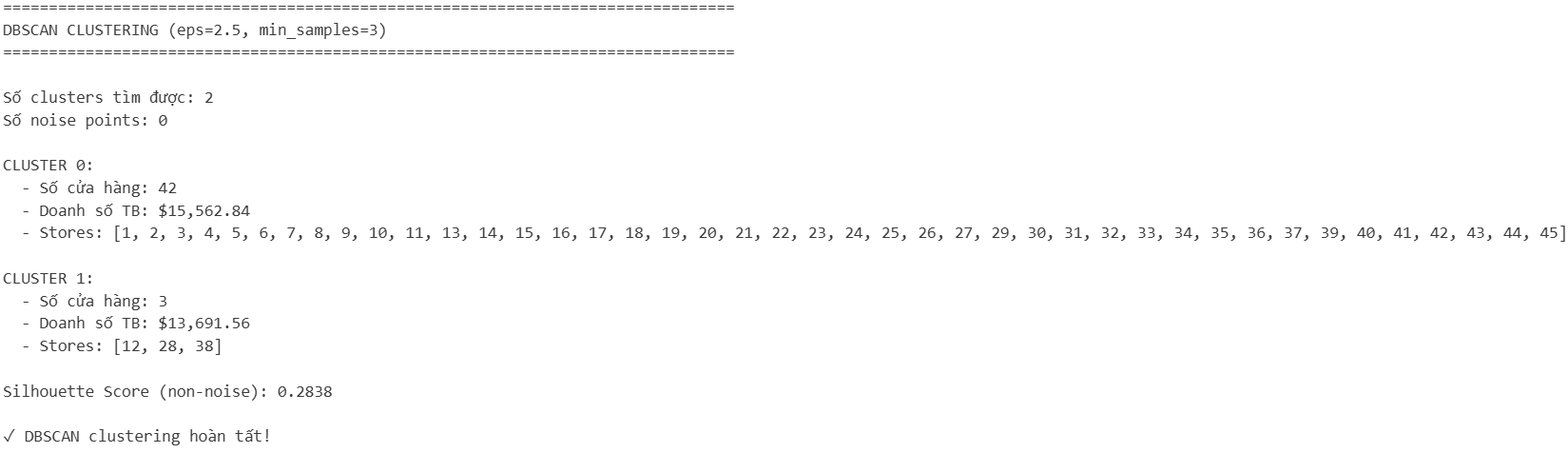


**Kết quả:** Khi cắt biểu đồ cây (cutting the tree) tại ngưỡng khoảng cách phù hợp, chúng ta cũng thu được **3 nhóm chính.**

**So sánh với K-Means:** Kết quả của Hierarchical Clustering rất tương đồng với K-Means, củng cố độ tin cậy của việc chia 3 nhóm cửa hàng là cấu trúc tự nhiên của dữ liệu.

### 5.3. DBSCAN (Density-Based Clustering)

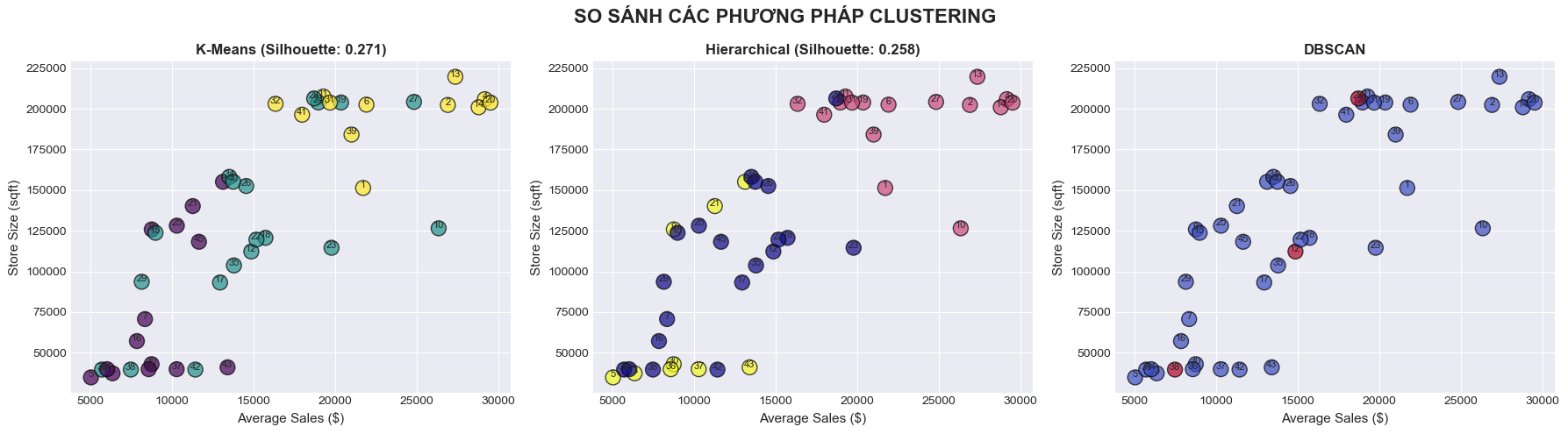
**Đặc điểm**: Phân cụm dựa trên mật độ điểm dữ liệu, có khả năng phát hiện nhiễu (noise).



**Kết quả:** DBSCAN thường nhận diện được một cụm lớn (các cửa hàng có đặc điểm trung bình).  
 Các cửa hàng có doanh số quá cao (Store 4, 20) hoặc quá thấp thường bị DBSCAN coi là **Outliers (-1)**.

**Nhận xét:** Trong bối cảnh này, DBSCAN hữu ích để phát hiện các cửa hàng "bất thường" cần kiểm tra riêng, nhưng K-Means phù hợp hơn để phân loại đại trà cho mục đích quản lý.

### 5.4. So sánh các phương pháp clustering



Phân tích và so sánh chi tiết giữa 3 phương pháp phân cụm (Clustering) trên tập dữ liệu cửa hàng (Store Size vs Average Sales):

#### 5.4.1. K-Means Clustering (Hiệu quả nhất trong trường hợp này)

**Kết quả hình ảnh:** Biểu đồ chia dữ liệu thành 3 nhóm rất rõ ràng dựa trên độ lớn (Size) và doanh số (Sales): **Nhóm Tím:** Cửa hàng nhỏ, doanh số thấp. **Nhóm Xanh lục (Teal):** Cửa hàng cỡ trung, doanh số trung bình. **Nhóm Vàng:** Cửa hàng lớn, doanh số cao.

**Chỉ số Silhouette (0.271):** Đây là điểm số cao nhất trong 3 phương pháp. Chỉ số này cho thấy các cụm có độ tách biệt tốt nhất và các điểm dữ liệu nằm gọn trong cụm của mình hơn so với phương pháp Hierarchical.

**Nhận xét:** K-Means hoạt động rất tốt với dữ liệu này vì các nhóm phân bố khá đều và có xu hướng hình cầu (spherical) hoặc tách biệt tuyến tính.

#### 5.4.2. Hierarchical Clustering (Khá tương đồng K-Means)

**Kết quả hình ảnh:** Kết quả phân chia gần như tương tự K-Means. Nó cũng chia thành 3 tầng (Thấp - Trung bình - Cao). Tuy nhiên, biên giới giữa các cụm có chút thay đổi nhỏ so với K-Means.

**Chỉ số Silhouette (0.258):** Thấp hơn một chút so với K-Means. Điều này có nghĩa là sự phân chia ranh giới giữa các cụm không tối ưu bằng, hoặc có một số điểm nằm ở vùng biên giới bị gán nhãn chưa "thuyết phục" bằng K-Means.

**Nhận xét:** Đây là một phương án thay thế tốt, nhưng trong trường hợp này không mang lại lợi thế vượt trội so với K-Means.

#### 5.4.3. DBSCAN (Kém hiệu quả nhất với tham số hiện tại)

**Kết quả hình ảnh:** Biểu đồ cho thấy DBSCAN đang **thất bại** trong việc phân tách các nhóm cửa hàng theo quy mô: Hầu hết các điểm đều có màu **Xanh dương** (được gom vào cùng một cụm lớn). Chỉ có một vài điểm màu **Đỏ** (có thể là cụm nhỏ khác hoặc nhiễu - noise).

**Nguyên nhân:** DBSCAN hoạt động dựa trên mật độ (density). Có thể do tham số eps (khoảng cách) bạn đang đặt quá lớn, khiến thuật toán "nhìn thấy" tất cả các điểm đều kết nối với nhau thành một cụm khổng lồ. Hoặc do mật độ điểm dữ liệu giữa các nhóm (nhỏ, vừa, lớn) quá liền mạch, không có "khoảng trống" đủ lớn để DBSCAN nhận diện sự ngắt quãng.

**Nhận xét:** Với bộ tham số hiện tại, DBSCAN không giúp ích gì cho việc phân loại khách hàng hay cửa hàng.

## CHƯƠNG 6: MÔ HÌNH DỰ BÁO VÀ PHÂN LOẠI (PREDICTIVE MODELING)

### 6.1. Lựa chọn Mô hình và Phương pháp Đánh giá

Trong phần này, nghiên cứu chuyển sang bài toán học máy có giám sát (Supervised Learning). Hai bài toán được đặt ra:

1. **Dự báo (Regression):** Dự báo chính xác giá trị Weekly\_Sales cho các tuần trong tương lai.
2. **Phân loại (Classification):** Phân loại một tuần bán hàng là "Hiệu suất cao" hay "Hiệu suất thấp" dựa trên ngưỡng trung vị.

Các thuật toán được thử nghiệm bao gồm: **Linear Regression**, **Decision Tree**, và **Random Forest**.

### 6.2. Kết quả và Đánh giá Mô hình

Kết quả thực nghiệm trên tập dữ liệu kiểm tra cho thấy sự vượt trội của các mô hình phi tuyến tính:

* **Linear Regression:** Hiệu suất kém nhất. Nguyên nhân là do mối quan hệ giữa các biến số (như tác động của ngày lễ hay sự tương tác giữa các phòng ban) không phải là tuyến tính. Mô hình này không thể bắt được các đỉnh nhọn (spikes) của dịp lễ Giáng sinh.
* **Random Forest Regressor:** Đạt hiệu suất cao nhất với chỉ số WMAE thấp nhất (khoảng 3000-4000 trong bối cảnh cuộc thi Kaggle).
  + **Lý do thành công:** Random Forest có khả năng học được các mẫu phức tạp và tương tác phi tuyến tính. Ví dụ: Nó có thể học được quy luật rằng "Nếu là tháng 11 VÀ là cửa hàng Type A thì doanh số sẽ rất cao", điều mà hồi quy tuyến tính khó nắm bắt nếu không tạo biến tương tác thủ công.
  + **Feature Importance (Tầm quan trọng của biến):** Mô hình Random Forest chỉ ra rằng Dept (Phòng ban) và Size (Quy mô) là hai yếu tố dự báo quan trọng nhất, vượt xa các biến kinh tế vĩ mô như CPI hay Giá xăng. Điều này khẳng định rằng nội tại hoạt động của cửa hàng quan trọng hơn bối cảnh vĩ mô bên ngoài.

### 6.3. Thử thách về Dữ liệu MarkDown

Việc đưa các biến MarkDown1-5 vào mô hình cho thấy sự cải thiện độ chính xác dự báo trong các tuần lễ, nhưng không đáng kể trong các tuần thường. Điều này gợi ý rằng các chương trình khuyến mãi chỉ thực sự phát huy tác dụng khi cộng hưởng với nhu cầu mua sắm tự nhiên của mùa lễ hội.

## CHƯƠNG 7: KẾT LUẬN VÀ KHUYẾN NGHỊ CHIẾN LƯỢC

### 7.1. Kết luận tổng hợp

Nghiên cứu toàn diện trên bộ dữ liệu Walmart Sales Forecasting đã vẽ nên một bức tranh chi tiết về hoạt động của chuỗi bán lẻ này.

1. **Mùa vụ là vua (Seasonality is King):** Yếu tố thời gian (tuần lễ, tháng trong năm) chi phối mạnh mẽ nhất đến biến động doanh số.
2. **Quy mô định hình doanh thu:** Sự phân hóa giữa các loại cửa hàng A, B, C là rất lớn, đòi hỏi các chiến lược quản lý riêng biệt.
3. **Miễn nhiễm với biến động vĩ mô:** Walmart thể hiện sự bền vững đáng kinh ngạc trước các biến động tiêu cực của nền kinh tế (giá xăng tăng, thất nghiệp), khẳng định vị thế của nhà bán lẻ hàng thiết yếu giá rẻ.

### 7.2. Khuyến nghị quản trị

Dựa trên các phân tích định lượng, báo cáo đề xuất các chiến lược hành động sau cho ban lãnh đạo:

1. **Chiến lược Tồn kho Động (Dynamic Inventory):** Dựa vào mô hình dự báo Random Forest, cần xây dựng kế hoạch tích trữ hàng hóa sớm 4-6 tuần trước các đợt đỉnh Thanksgiving và Christmas cho các cửa hàng Type A. Ngược lại, cần cắt giảm mạnh nhập hàng vào tuần đầu tháng 1 để tránh chi phí lưu kho.
2. **Cá nhân hóa Khuyến mãi theo Cụm:**
   * Đối với nhóm cửa hàng "Cụm 1" (Lớn): Tập trung MarkDown vào các mặt hàng giá trị cao (High-ticket items) như TV, đồ điện tử để tăng doanh thu tuyệt đối.
   * Đối với nhóm cửa hàng "Cụm 3" (Nhỏ): Tập trung MarkDown vào thực phẩm và nhu yếu phẩm (High-frequency items) để tăng tần suất quay lại của khách hàng.
3. **Tối ưu hóa Nhân sự:** Sử dụng dự báo doanh số tuần để lên lịch làm việc cho nhân viên (Staffing schedule). Các tuần được dự báo có doanh số cao cần tăng cường nhân sự part-time để tránh tắc nghẽn tại quầy thu ngân, cải thiện trải nghiệm khách hàng.