Walmart Sales Prediction

Project Overview

With its vast network of stores and diverse product offerings, Walmart faces the challenge of accurately predicting sales, which is crucial for efficient supply chain management and meeting customer demands.

Project Objective:

- Develop a reliable sales predicting model for Walmart.

Challenges:

- Sales predictions at the product level are intricate due to factors like seasonality, marketing campaigns, holidays, and contextual influences.
- Limited historical data adds complexity to modeling retail sales accurately.

Dataset

The dataset used for this project is sourced from Kaggle

Data Coverage:

- Historical weekly sales data.
- Includes data for 81 departments.
- Covers 45 Walmart stores located in diverse regions.
- Spans from February 5, 2010, to November 1, 2012.

Features:

- The feature file contains supplementary details for each date, including Temperature, Fuel Price, MarkDown, CPI, Unemployment, and Holiday information.

Store Information:

- The store file provides insights into the size and type of each Walmart store.

Merged Dataset:

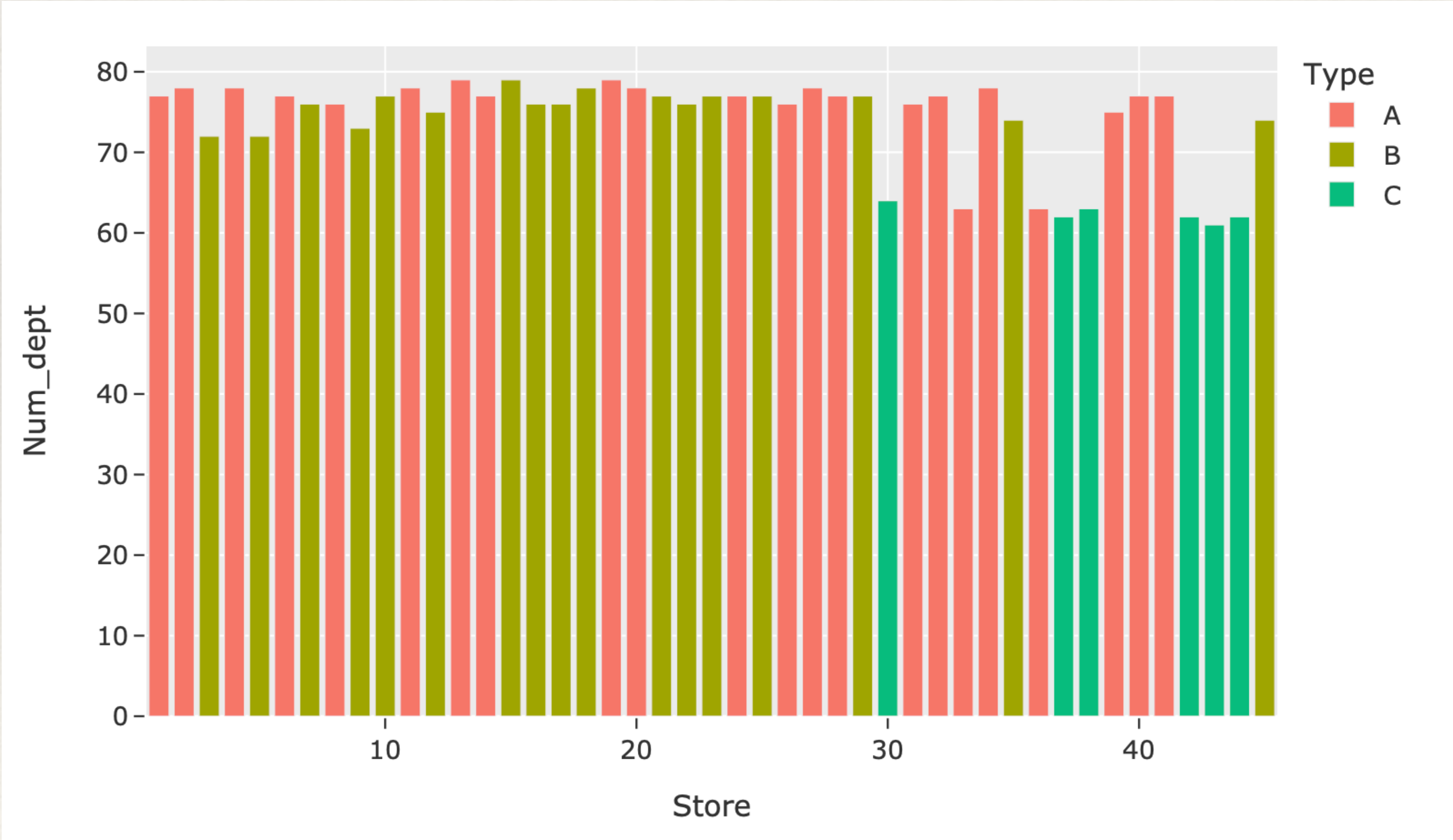
- Before preprocessing, the merged dataset contains 421,570 records and 16 columns.

The list of columns

- Store: The ID of the store.
- Dept: The ID of the department.
- Date: The date of the record.
- Weekly_Sales: The amount of sales in the given week
- IsHoliday: A boolean that indicates if the week is a holiday week.
- Temperature: The average temperature in the region.
- Fuel_Price: The cost of fuel in the region.
- MarkDown1 to MarkDown5: Anonymized data related to promotional markdowns.
- CPI: The consumer price index.
- Unemployment: The unemployment rate.
- Type: The type of the store.
- Size: The size of the store.

Exploratory Data Analysis - Store and Department

- In this dataset, we observe a total of 81 available departments across Walmart stores.
- However, the number of departments varies from store to store, creating interesting insights:

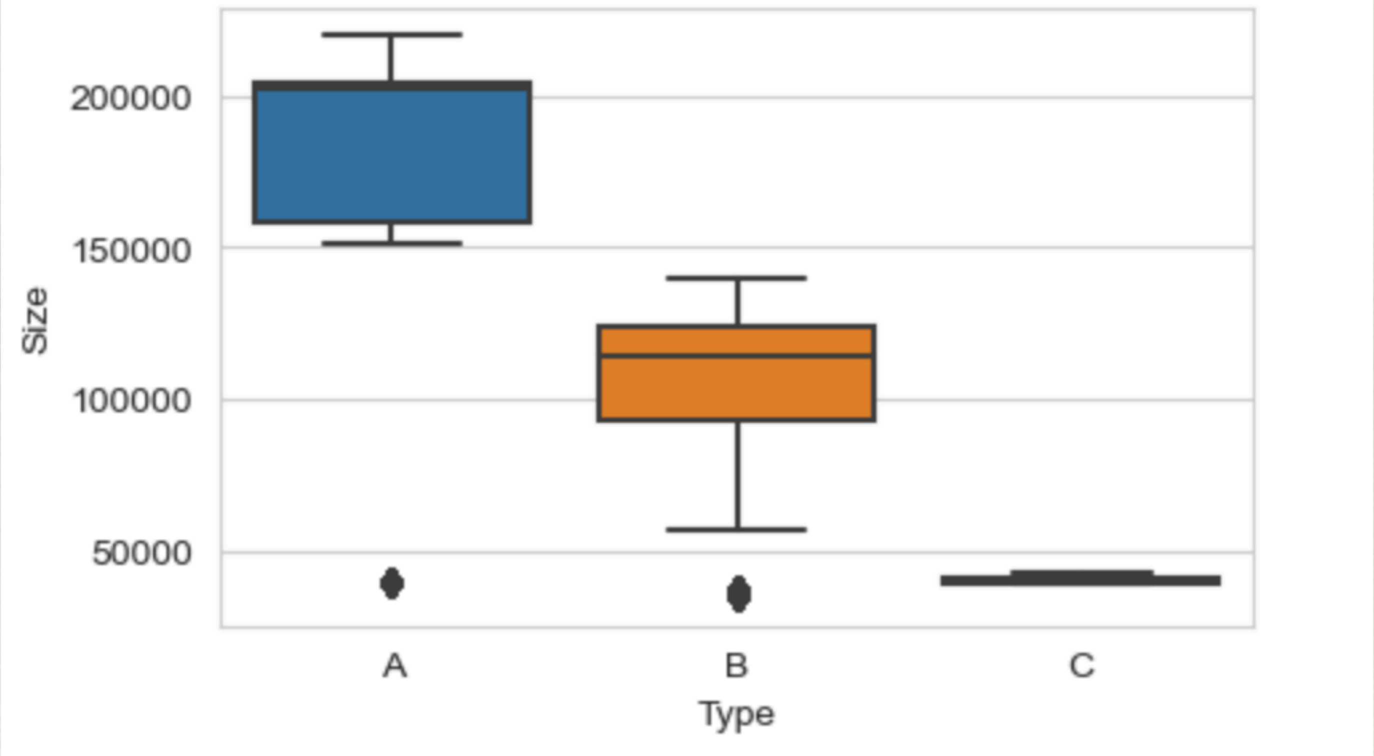


- Some stores feature as many as 79 departments, while others have as few as 61.
- Specific departments are not consistently available in all stores.
- The number of departments and store types
  - Type A and B stores generally have around 75 to 79 departments.
  - Type C stores typically have 61 to 64 departments, with exceptions in stores 33 and 36.

Exploratory Data Analysis - Store Type and Size

Store Type as an Indicator of Size:

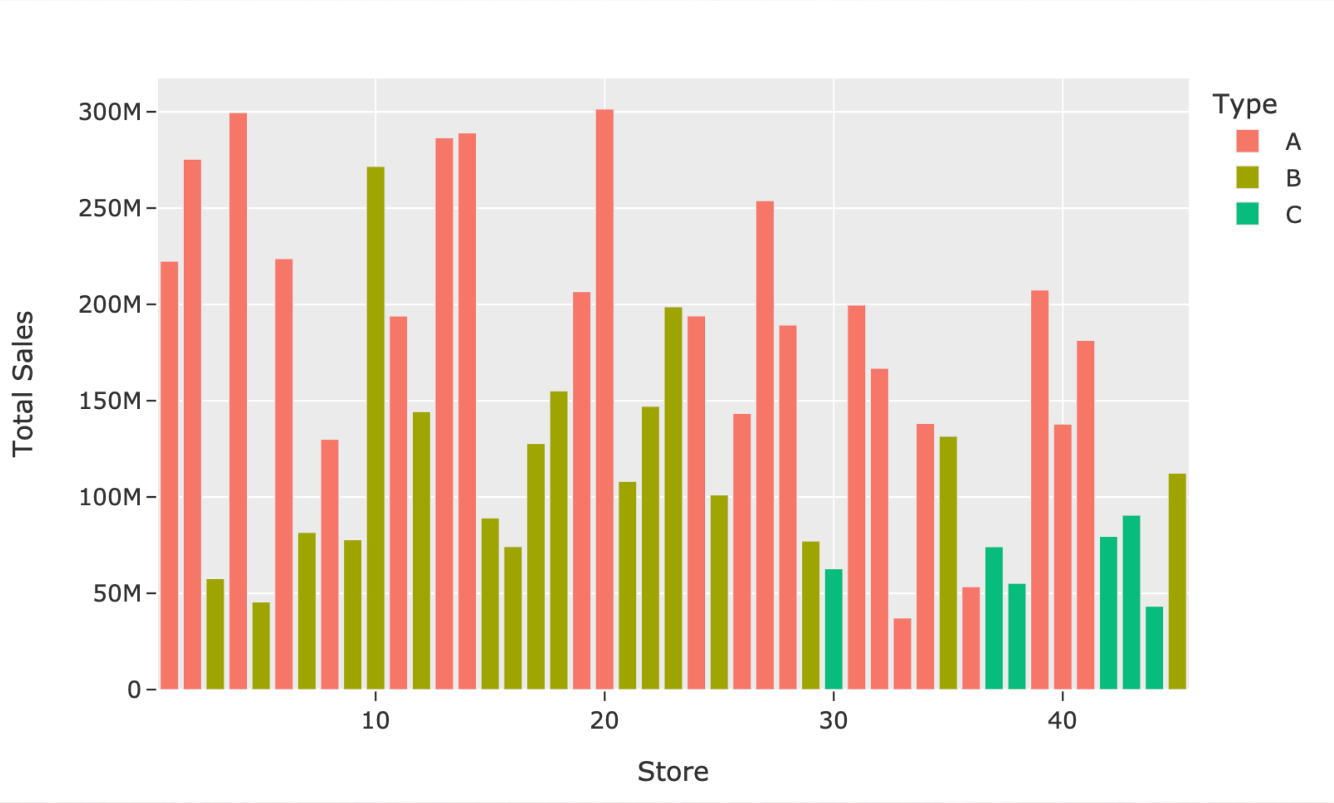
- The store types correspond to their sizes, with Type A being the largest and Type C being the smallest

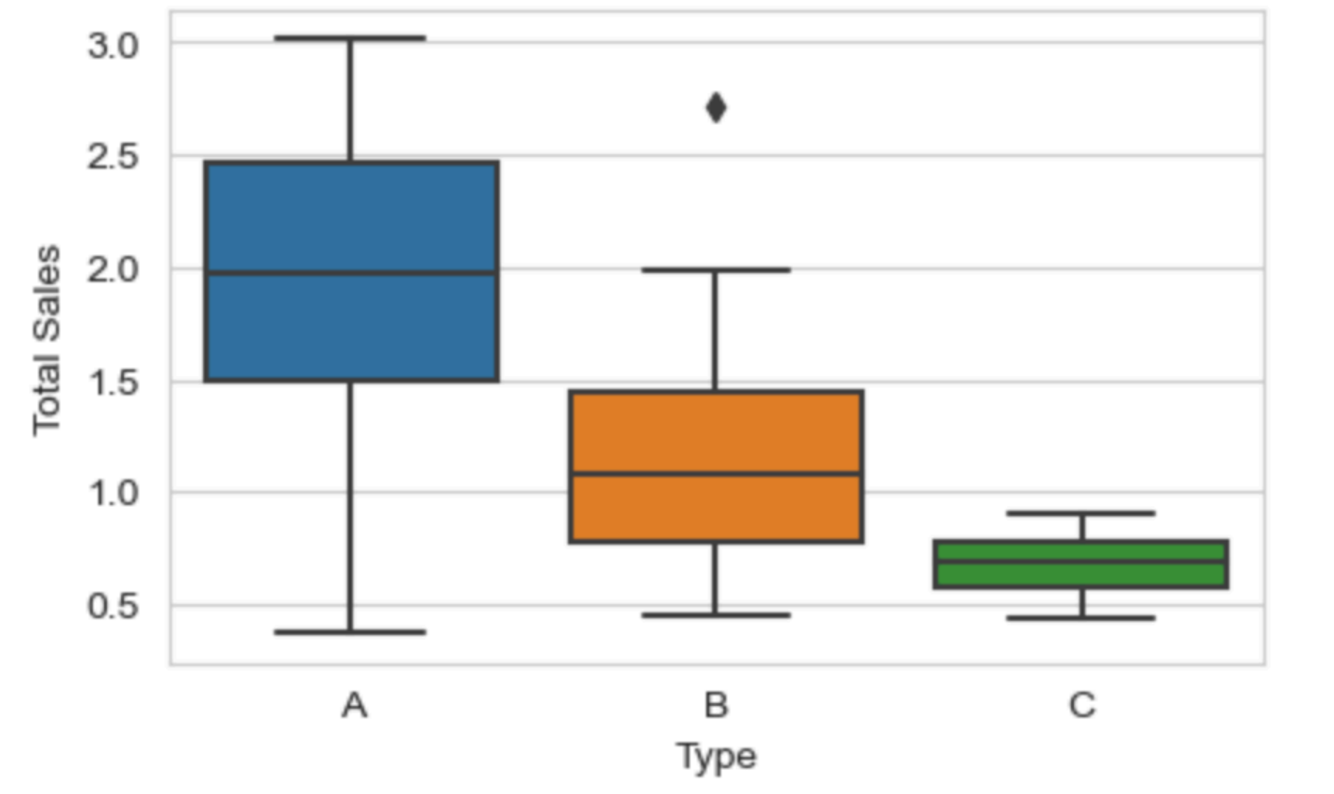


Findings:

- Most stores align with expected size patterns.
- Stores 33 and 35 significantly deviate from the expected patterns, indicating potential mislabeling as Type A.
- Stores 3 and 5 do not distinctly stand out from other Type B stores.

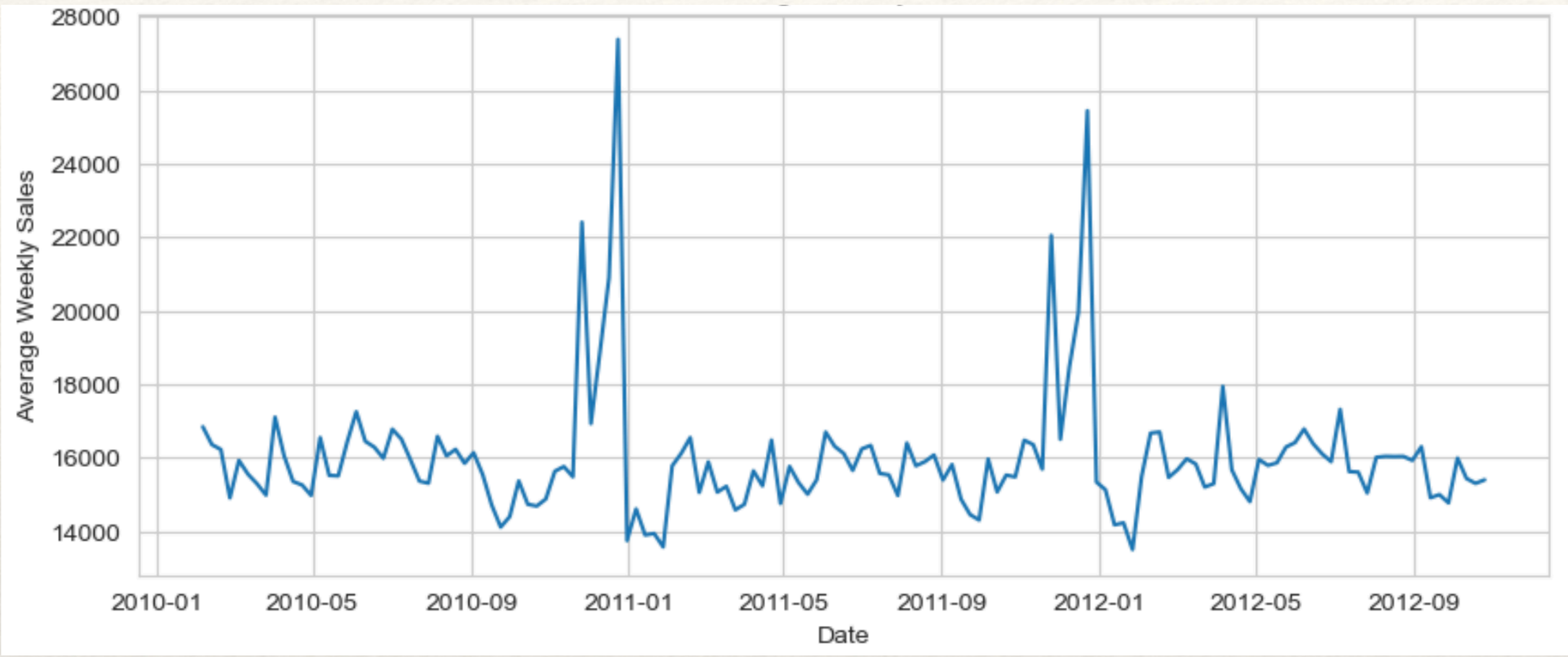
Exploratory Data Analysis - Total sales by store and type

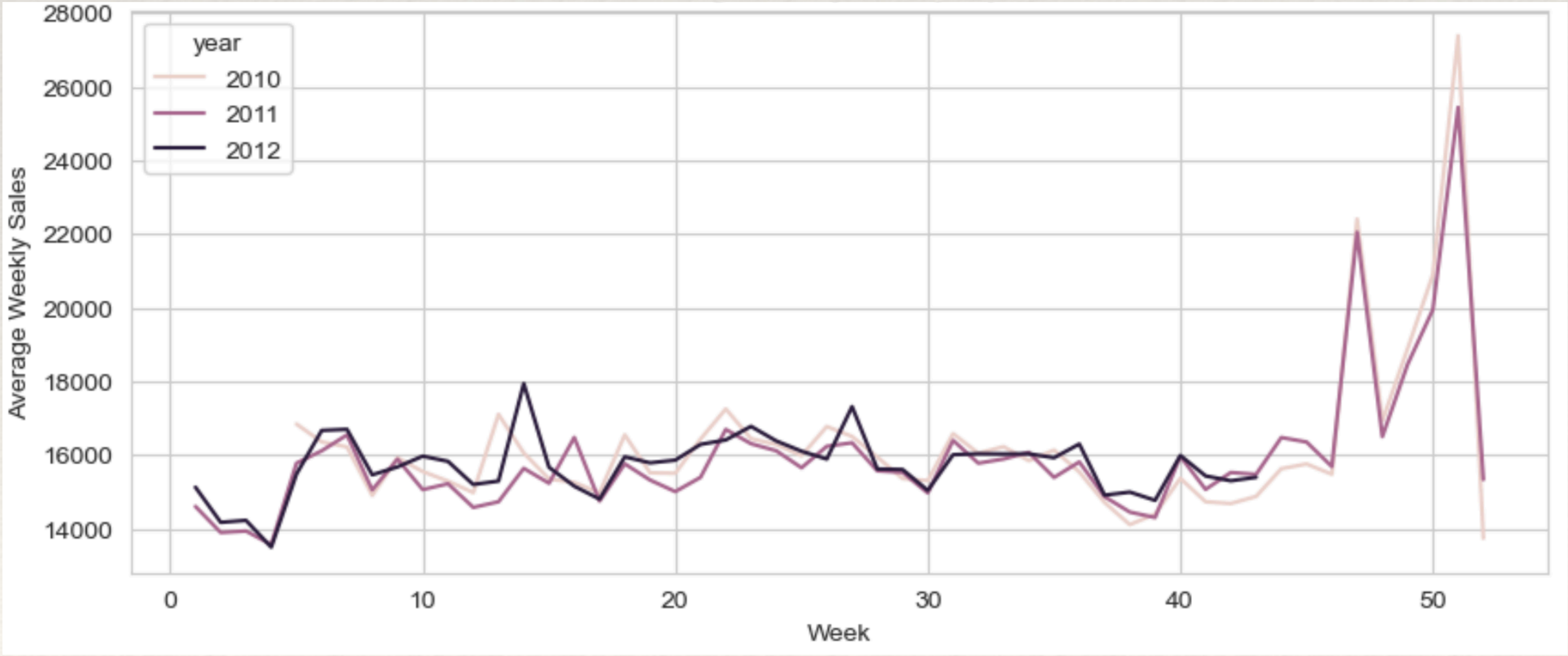




- Type A stores have the highest median sales.
- Type B stores follow with slightly lower sales.
- Type C stores tend to cluster around the median.
- Stores 33 and 35 have sales levels similar to Type C stores, supporting the earlier conclusion of potential mislabeling.
- However, there isn't a straightforward correlation between store type and total sales, as other factors influence sales.

Exploratory Data Analysis - Seasonal Trends in Sales





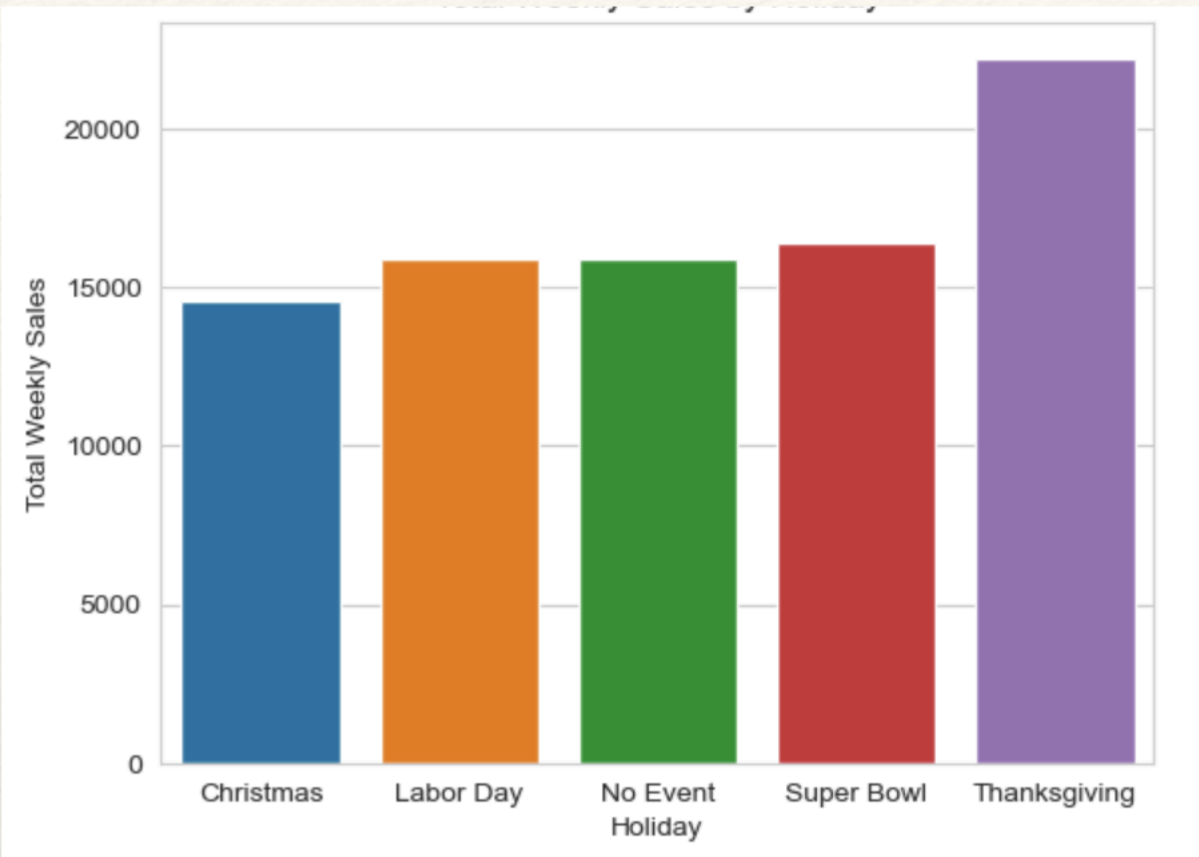
Recurring Seasonal Trends:

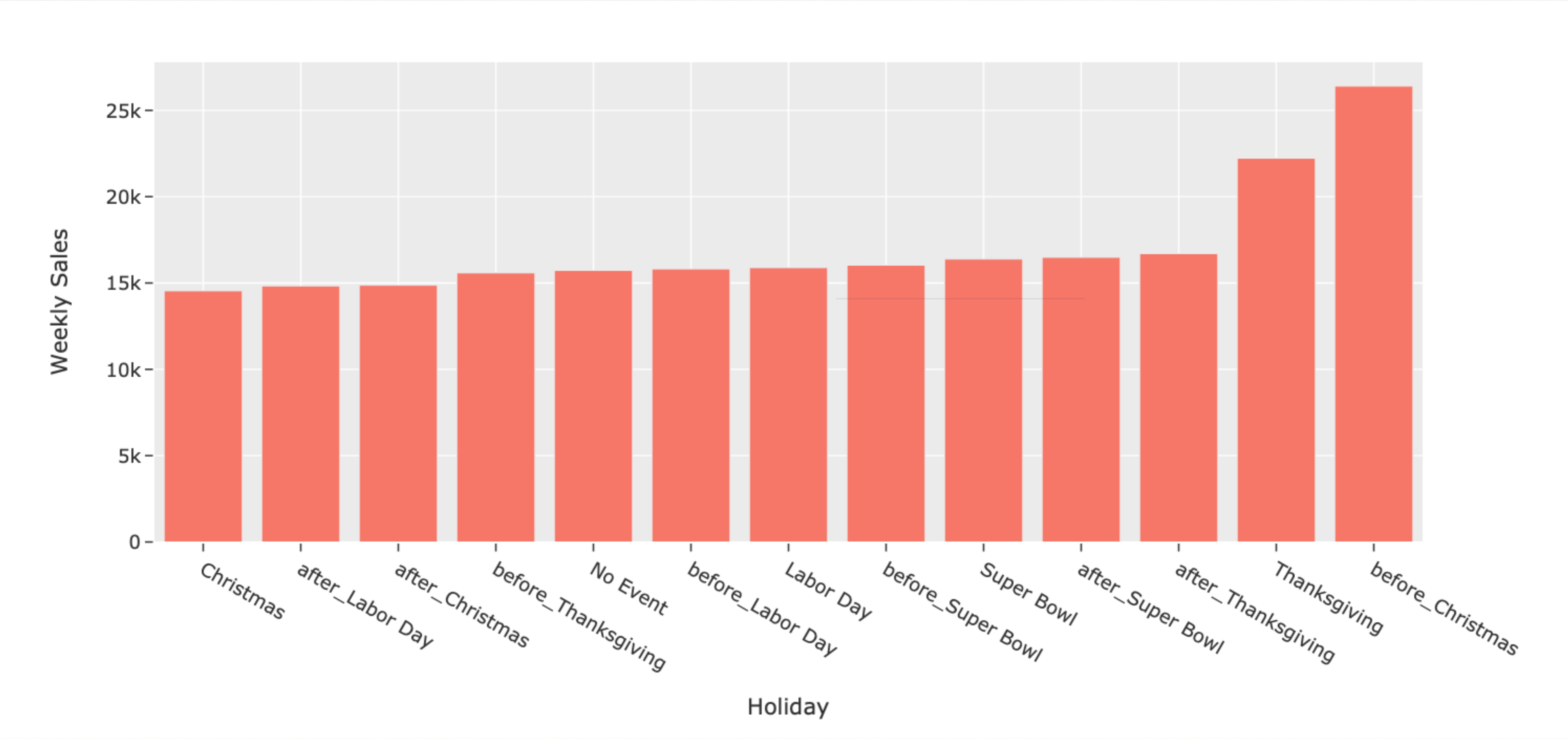
- This peak is associated with increased consumer spending during the holiday seasons.
- The average weekly sales per year plot showcases a stable and consistent pattern over three years, with minimal variation.

Implications:

- The presence of recurring seasonal trends suggests that specific factors drive increased consumer spending during the holiday seasons.
- The underlying factors driving sales remain relatively constant over time.

Exploratory Data Analysis - Holiday





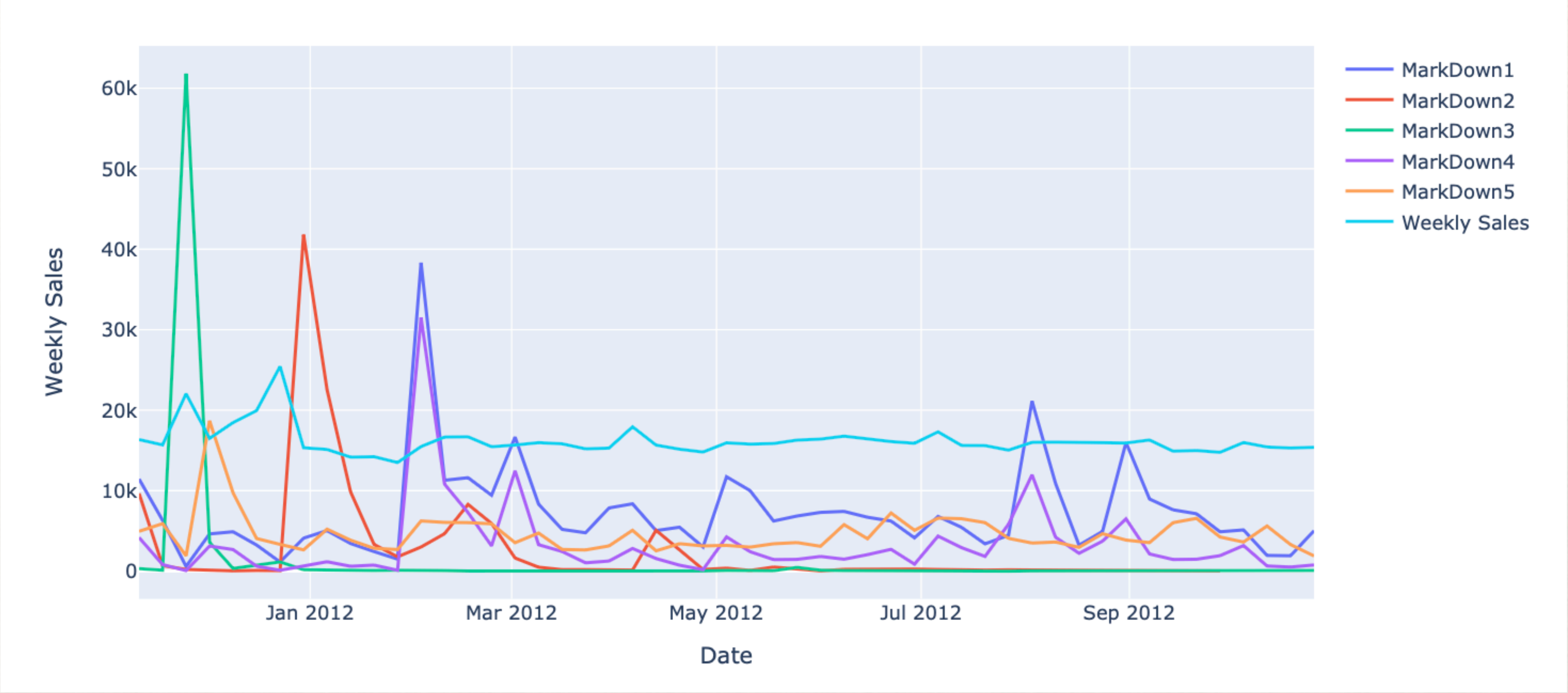
Holiday Sales Impact:

- Among these holidays, Thanksgiving stands out with the highest average weekly sales impact.
- Interestingly, Christmas has a surprisingly lower impact on sales, but December records the highest peak in sales throughout the year.

Sales Timing for Christmas:

- The sales effect of Christmas primarily occurs before Christmas day, as people tend to shop for gifts in advance.
- To investigate further, we added an extra column to analyze the sales impact of the week before and after each holiday. This analysis confirmed that the week leading up to Christmas saw the highest average weekly sales of the year.

Exploratory Data Analysis - Markdowns



- The Markdown columns contain anonymized data that relates to the promotional markdowns that Walmart is running.
- These columns are only available from November 2011 onward and were not present before that time.

Exploratory Data Analysis - Markdowns

Markdown Patterns:

- Markdowns generally exhibit a consistent pattern throughout the year, except for a significant surge leading up to the holidays.

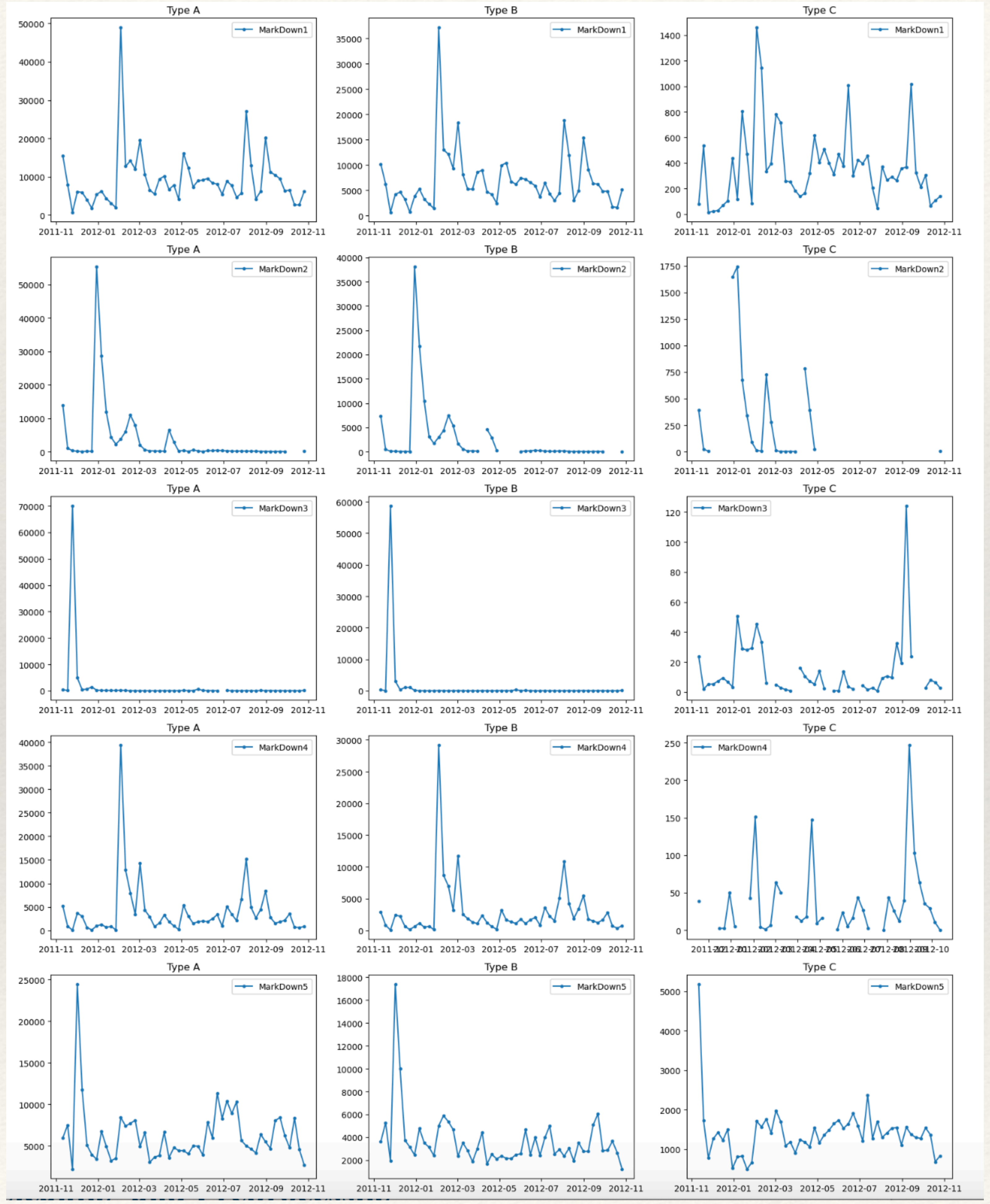
Holiday-Related Sales Impact:

- This holiday-related peak is particularly extreme, with substantial differences between the maximum and mean values and high standard deviations.

Markdown Impact:

- Markdown 3 experiences a dramatic increase of 61,817.05 during Thanksgiving but remains stable during the rest of the year.
- Markdowns 1 and 4 have peaked during the Super Bowl, while Markdown 4 specifically targets Christmas.
- Markdown 5 falls between Thanksgiving and Christmas, aligning with the holiday season.

Exploratory Data Analysis - Markdowns



Type A and B Stores:

- These stores exhibit similar markdown patterns over time.
- They generally have low percentages of missing values in most markdown columns.

Type C Stores:

- Despite being smaller in size, Type C stores have the highest percentage of missing values.
- The plot indicates a different pattern compared to the other store types.

Differences in the distribution of markdown values are noticeable among store types:

- Type A and B stores have wider ranges, higher mean, and median values compared to Type C stores.

Data Cleaning and Preprocessing

Feature engineering techniques to account for temporal aspects in the data

- date and time features, lag features, and a window feature

Categorical Features and Holiday-Related Features

- Type Feature: Through exploratory data analysis, it was observed that store type is related to store size. Hence, Ordinal Encoding was applied to the Type feature.
- IsHoliday Feature: This feature denotes whether a holiday is observed or not, with each holiday having a distinct impact on sales. A new Holiday column was created to indicate the name of the holiday on each date. One-Hot Encoding was employed to generate binary columns for each holiday.
- Weeks Before Christmas: During the Christmas season, significant sales occur in the weeks leading up to the holiday. To account for this, the WeeksBeforeChristmas column was introduced, assigning corresponding values up to four weeks before Christmas.

Imputing Missing Value

In markdown columns, missing values are not random, so it may not be appropriate to use imputation techniques such as mean, median, or mode.

Imputing Values while Preserving Temporal Patterns

- An effective strategy for imputing missing Markdown values is to utilize the Markdown during the corresponding week for each store.
- This approach preserves the temporal patterns in the Markdown data and aligns with the average Markdown pattern over time, which is related to specific holidays.

Consistency in Markdown Patterns:

- Assuming that Markdown patterns remain consistent, the strategies observed after November 2011 can also be applied to the period before November 2011.

Modeling

Data Splitting:

- When working with time series data, maintaining chronological order during splitting is crucial to avoid train-test contamination.
- Rather than using a ratio, we opted for a specific time point to separate the data.
- The training set comprises 80% of the total weeks in the data, ensuring a chronological split.

Random Forest Baseline:

- A Random Forest model was used as a baseline.
- The RMSE for the Random Forest model was approximately 3,554.79.

Model Improvement

Overfitting Challenge:

- A significant disparity between the RMSE of the training set and the test set suggests overfitting, where the model is fitting too closely to the training data.

XGBoost Experimentation:

- To address overfitting and improve model performance, I experimented with the XGBoost model, which showed the second-highest initial performance.

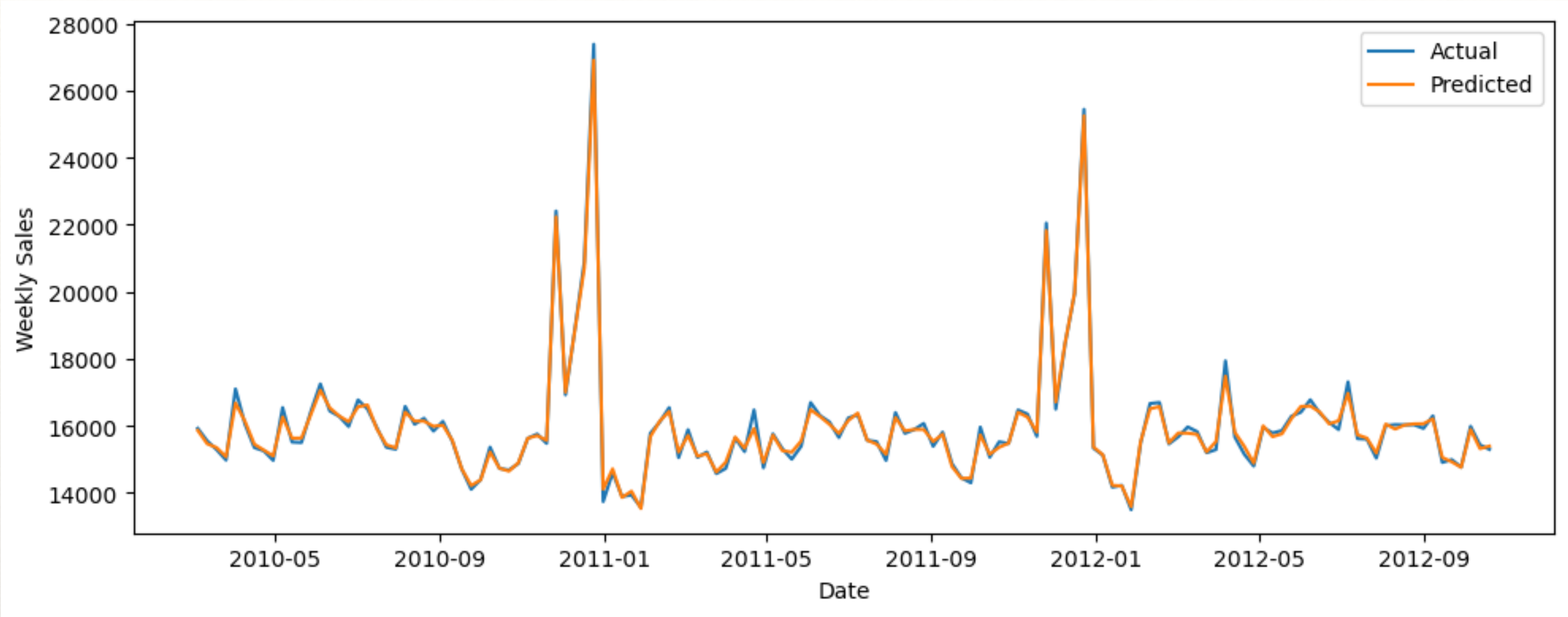
Hyperparameter Tuning:

- Hyperparameters of the XGBoost model were systematically tuned to optimize its performance.

Improved RMSE:

- After identifying the optimal hyperparameters, the RMSE of the XGBoost model significantly improved, achieving an RMSE of 2,747.75.

Model Performance and Sales Spikes



The weekly sales predicted by the XGBoost model are compared to actual sales. The data is divided into five folds, preserving the temporal order.

Predictive Insights and Limitations

Underestimation During Peak Periods: Notably, the XGBoost model tends to underestimate weekly sales during peak periods.

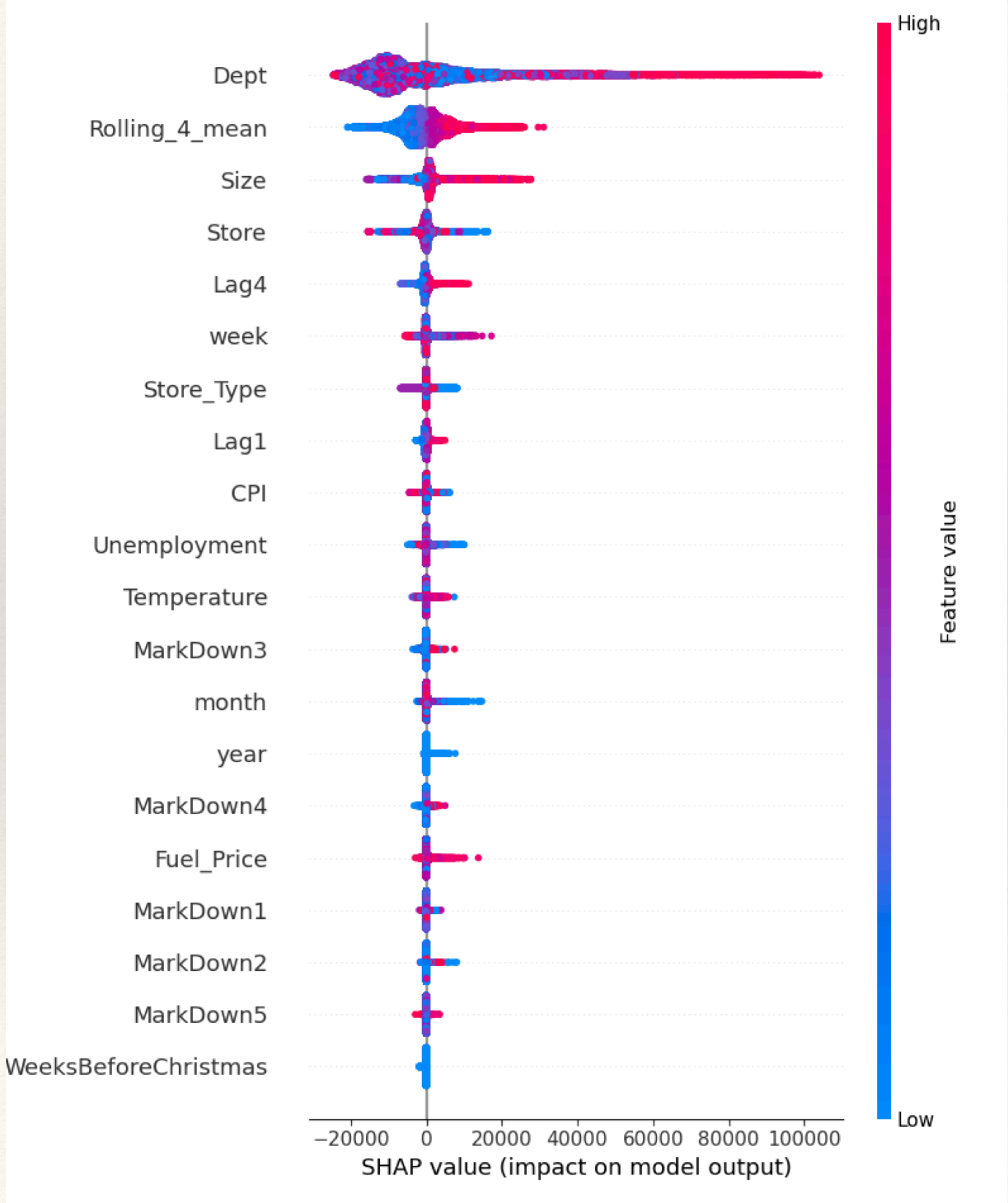
High Residuals in Specific Departments:

- Certain departments exhibit high residuals, with some exceeding $50,000.
- Approximately half of these high residuals are concentrated in Departments 72 and 7.
- During the holiday season, these two exhibit exceptional sales performance.

Model Limitations:

Unfortunately, the XGBoost model cannot accurately capture sudden jumps in sales like those observed in these departments.

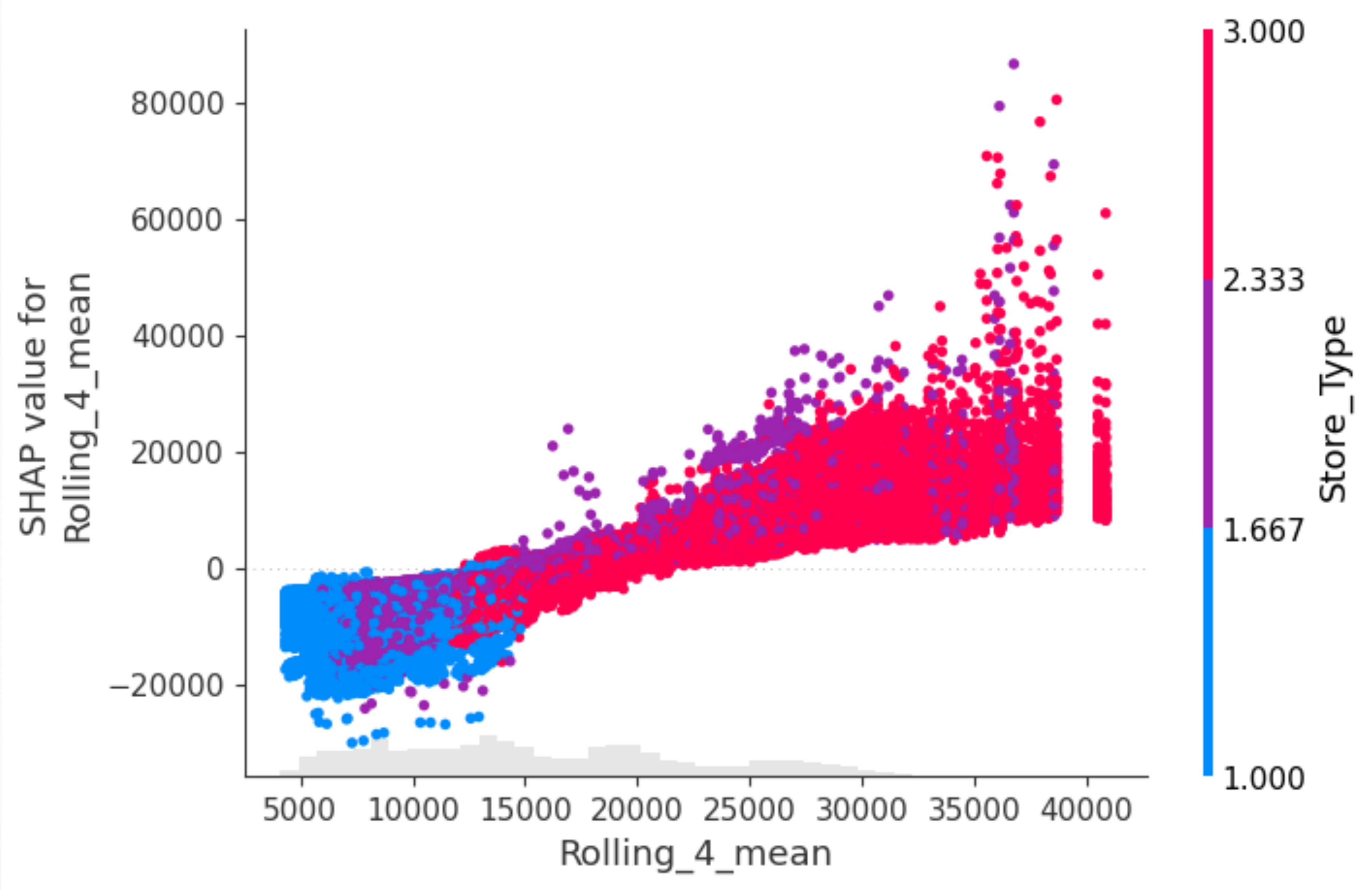
SHAP Analysis and Feature Importance



Key Insights:

- **Department**: Identified as the most important feature on average. Departments with high values have large positive SHAP values, primarily driven by Department 72.
- **Rolling_4_mean**: Rolling features are valuable in time series forecasting as they enable models to analyze historical patterns and dependencies in the data. Positive SHAP values associated with higher rolling averages indicate that when there are high recent sales trends, as captured by the rolling average, the model tends to predict higher weekly sales.

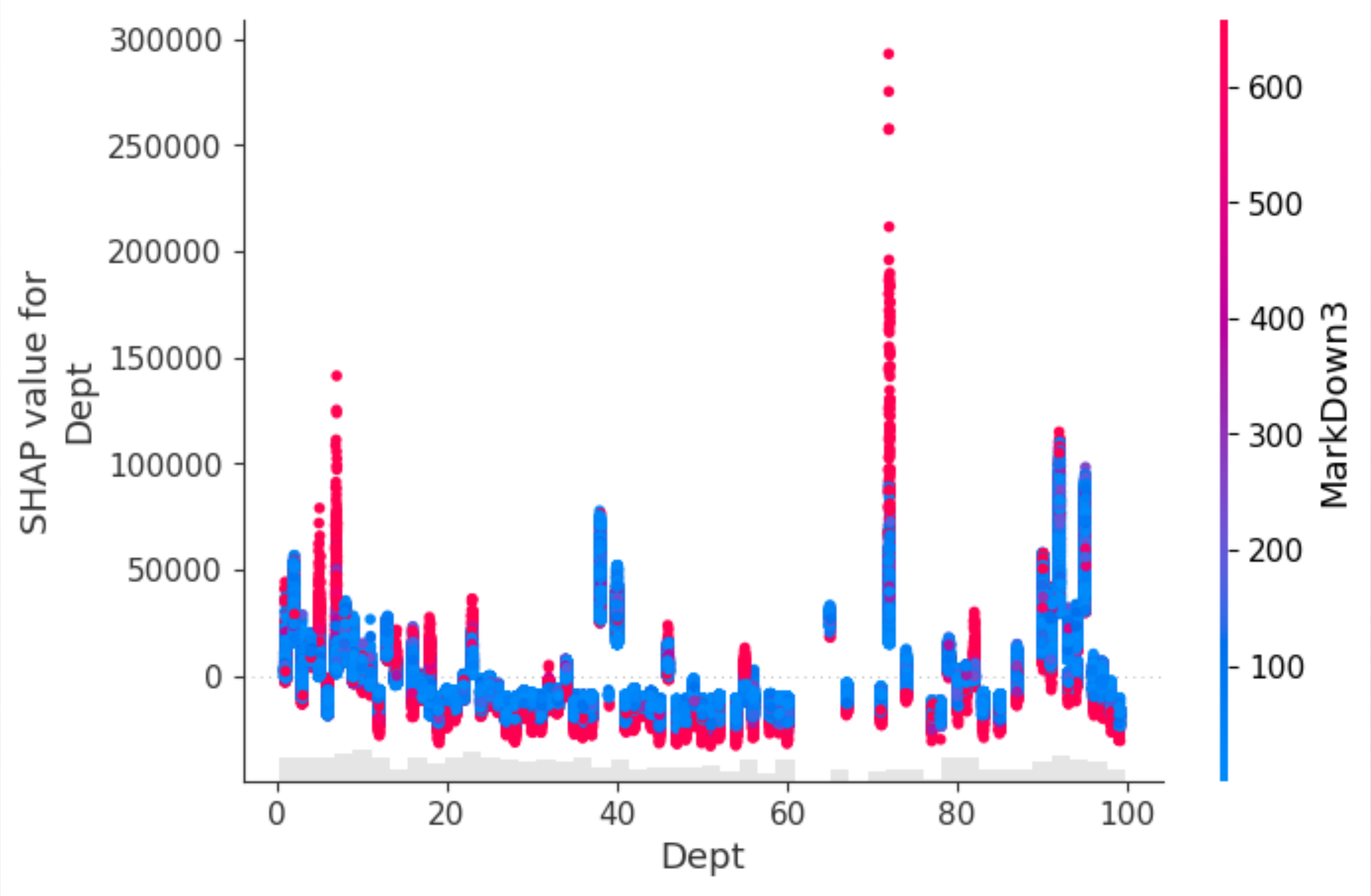
SHAP Dependency Plot



Store Type Influence:

- Red color (Store Type A) is predominant in the plot, indicating that larger stores (Type A) tend to have higher rolling averages.
- This finding aligns with earlier EDA observations that larger stores generally have higher weekly sales.
- The SHAP analysis supports the insight that store type plays a significant role in influencing the positive correlation between rolling averages and weekly sales.

SHAP Dependency Plot



The figure illustrates how Markdown 3 impacts different departments and contributes to SHAP values.

Impact on Large Positive SHAP Values:

- It's evident that certain departments with high values are affected by Markdown 3, leading to large positive SHAP values.
- This interaction plays a crucial role in explaining why the model underestimates certain prediction points.

Recommendations for Further Improvements

Additional Features:

- Consider incorporating additional features such as holidays and more detailed department categories.

Hyperparameter Tuning:

- Conduct extensive hyperparameter tuning for the XGBoost model. Optimizing model parameters can further improve its predictive performance.

Ensemble Methods:

- Experiment with ensemble methods, such as stacking multiple models. Combining different models can enhance prediction accuracy, especially for challenging scenarios like certain departments during holiday seasons.

Time Series Forecasting Techniques:

- Explore time series forecasting techniques to capture seasonality and trend patterns more precisely.These techniques can provide more accurate insights into sales patterns over time.