SEP 6DA3: Data Analytics and Big Data

Final Project

Submitted by
Kenil Sachapara (400551600)
Hem Gondaliya (400551597)
Om Kakadiya (400551508)

Instructor:
Siqi Zhao

McMaster University, Hamilton

Dec 16, 2024

Abstract

The Walmart Store Sales Forecasting project aims to predict weekly sales for various departments across 45 Walmart stores using historical sales data, promotional markdowns, and external factors. This report discusses the challenges, data preparation, exploratory data analysis (EDA), feature engineering, and the models considered for forecasting.

Contents

1 Introduction

The dataset presented here contains historical sales data for 45 Walmart stores across different regions, each containing several departments. The goal of this analysis is to predict the department-level sales for each store. The dataset spans several years and includes key features such as promotional markdowns, store-specific details, and other socio-economic factors. The primary challenge in this dataset arises from modeling the impact of holiday-specific markdowns, which are crucial for accurate sales forecasting.

Walmart runs several promotional markdown events each year, which directly affect sales. These markdowns often precede major holidays such as the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks that include these holidays are weighted five times higher in the evaluation metric than non-holiday weeks, emphasizing the importance of capturing the impact of these promotions in the prediction model. This presents a challenge, as some holidays are not consistently present in the dataset, and there is incomplete historical data regarding promotional markdowns.

The dataset is provided in the following files:

- **stores.csv**: Contains anonymized information about the 45 stores, including store type and size.
- **train.csv**: The historical training data covering the period from 2010-02-05 to 2012-11-01, containing weekly sales, store and department numbers, and holiday status.
- **test.csv**: Identical to the training data, but with withheld sales values, which need to be predicted.
- **features.csv**: Contains additional store and region-specific data, including temperature, fuel prices, markdowns, CPI, and unemployment rates.

The file **features.csv** also includes information on the four major holidays and their respective weeks in the dataset, which can be crucial for identifying periods of sales spikes due to markdowns. These holidays include:

- **Super Bowl**: Occurs around mid-February each year.
- **Labor Day**: Occurs around the first Monday in September.
- **Thanksgiving**: Occurs around the fourth Thursday in November.
- **Christmas**: Occurs on December 25th.

This dataset provides a comprehensive view of Walmart's sales patterns, influenced by both external and internal factors, making it an ideal challenge for applying machine learning techniques to predict future sales.

2 Data Preparation

2.1 Dataset Overview

The dataset includes historical sales data for 45 Walmart stores and consists of the following files:

- `stores.csv:` Contains anonymized information about the 45 stores, indicating the type and size of each store.
- `train.csv:` Covers historical weekly sales data for the period 2010-02-05 to 2012-11-01. It contains the following fields:
  - Store: The store number.
  - Dept: The department number.
  - Date: The week.
  - Weekly_Sales: Sales for the given department in the given store.
  - IsHoliday: Whether the week is a special holiday week.
- `test.csv:` Similar to train.csv but with the Weekly_Sales field withheld for prediction.
- `features.csv:` Provides additional data related to store, department, and regional activity for the given dates. It includes:
  - Store: The store number.
  - Date: The week.
  - Temperature: Average temperature in the region.
  - Fuel_Price: Cost of fuel in the region.
  - MarkDown1-5: Anonymized data related to promotional markdowns (available after Nov 2011 and not for all stores/times).
  - CPI: The Consumer Price Index.
  - Unemployment: The unemployment rate.
  - IsHoliday: Whether the week is a special holiday week.

2.2 Data Cleaning Steps

- **Handling Missing Values:**
  - MarkDown columns were filled with 0, as missing values represent no promotions.
  - Missing values in CPI and Unemployment were imputed using forward-fill and backward-fill methods to ensure consistency over time.
- **Outlier Detection:** Sales outliers were identified using the Interquartile Range (IQR) method and capped to mitigate their influence on model training.

- **Date Formatting:** Converted the Date column to a standard datetime format for efficient feature extraction and analysis.
- **Merging Datasets:** Combined `train.csv` and `features.csv` on Store and Date to create a unified dataset. Joined with `stores.csv` to incorporate store type and size.
- **Holiday Mapping:** Added a Holiday_Type feature to distinguish between major holidays (e.g., Super Bowl, Labor Day, Thanksgiving, and Christmas) and regular weeks.

3 Exploratory Data Analysis (EDA)

3.1 Overview of Sales Data

- **Distribution of Weekly Sales:** Analyzed the distribution of Weekly_Sales across departments and stores. Sales showed significant variability, with some departments exhibiting consistent trends while others displayed high volatility.

- **Holiday Effects:** Weeks marked as holidays had higher average weekly sales. Sales during Christmas and Thanksgiving outperformed other holidays like Labor Day and the Super Bowl.

3.2 Trends in Time Series Data

- **Seasonality:** Clear seasonal trends were observed, with sales peaking during holiday periods. Non-holiday weeks showed stable yet lower average sales.

- **Year-over-Year Comparison:** Sales trends increased year-over-year, indicating growth in store performance and customer base.

3.3 Feature Relationships

- **Correlation Analysis:** Weekly_Sales positively correlated with MarkDown features and IsHoliday.

- **Fuel Price and Sales:** Sales were moderately affected by changes in Fuel_Price, with higher prices slightly dampening sales trends.

3.4 Store Segmentation

- **Store Size and Sales:** Larger stores consistently achieved higher weekly sales than smaller ones.

- **Store Types:** Store types (Type A, B, C) demonstrated distinct sales patterns, with Type A stores leading in revenue.

3.5 Missing Data Patterns

- **Markdown Data:** Significant missing values were observed in MarkDown columns, particularly before November 2011. Visualizations revealed that smaller stores were more likely to lack markdown data.

- **Imputation Strategies:** Missing MarkDown values were treated as zeros, reflecting the absence of promotions. Other missing values in CPI and Unemployment were imputed using time-based forward and backward fills.

4 Feature Engineering

- **Date Features:** Extracted year, month, week, and weekday from the date to capture any seasonal patterns and temporal trends.
- **Promotions Impact:** Created a feature to represent the number of promotional markdowns during each week, allowing the model to understand the impact of promotions.
- **Holiday Flags:** Incorporated binary features to mark holidays such as Christmas, Thanksgiving, Super Bowl, and Labor Day to capture the effect of special events.
- **Lag Features:** Created lag features to capture the past week's sales and temperature, which can influence current sales.
- **Rolling Statistics:** Calculated rolling averages and rolling standard deviations for weekly sales and temperature to help capture trends and volatility in the data.

5 Model Selection and Implementation

5.1 Models Considered

1. Baseline Model: Linear Regression.

2. Tree-Based Models:

    - Decision Tree Regressor
    - Gradient Boosting Regressor (XGBoost)

5.2 Hyperparameter Tuning

GridSearchCV optimized parameters such as:

- Number of trees.

- Depth of the tree.

- Learning rate for boosting models.

5.3 Pipeline

Imputation, scaling, and encoding were done by a preprocessing pipeline prior to data being fed into models.

5.4 Cross-Validation

5-fold cross-validation made sure that the model evaluation was resilient.

6 Results and Evaluation

6.1 Linear Regression

Training Set:

- WMAE Loss: 14,776.36
- MAE: 14,576.65
- MSE: 475,155,405.51
- RMSE: 21,798.06
- R-squared: 0.0857

Validation Set:

- WMAE Loss: 14,884.37

6.2 Decision Tree Regressor

Training Set:

- WMAE Loss: 0.0
- MAE: 0.0
- MSE: 0.0
- R-squared: 1.0

Validation Set:

- WMAE Loss: 1,942.89
- MAE: 1,722.37
- MSE: 24,803,215.97
- R-squared: 0.95

6.3 Gradient Boosting Machine

Training Set:

- WMAE Loss: 3,059.97

Validation Set:

- WMAE Loss: 3,157.3

7 Insights and Business Implications

7.1 Insights from Model Importances

- **Random Forest:** The most important features are department, store size, and store number.
- **Gradient Boosting Machine:** The three main contributing factors are department, store size, and store type. Store Type is more significant than Store Number, in contrast to Decision Trees and Random Forests.

7.2 Business Implications

1. **Promotion Strategies:** Sales can be considerably increased by using strategic markdowns during holiday weeks.
2. **Holiday Planning:** Improved staffing and inventory planning during significant holidays.
3. **Store Optimization:** Targeted methods are necessary for larger retailers (Type A) to maintain higher sales volumes.
4. **Department Sales Strategy:** Revenue can be maximized by customized promotional initiatives based on department-specific sales trends.

8 Conclusion and Future Work

Weekly sales were accurately forecasted by the Walmart Store Sales Forecasting initiative. Through the utilization of sophisticated feature engineering and the integration of external aspects, the models yielded practical insights for strategic decision-making.

8.1 Model Comparison

- Linear Regression: Validation WMAE - 14,884.37
- Decision Tree Regressor: Validation WMAE - 1,938.54
- Gradient Boosting Machine: Validation WMAE - 1,339.29

With its hyperparameters adjusted, the Gradient Boosting Machine performed better than the other models and was the most accurate forecaster of future sales.

8.2 Future Enhancements

1. Include more outside data, including consumer demographics or economic trends.
2. Try using deep learning models to make better predictions.
3. Test theories in the real world to confirm hypotheses and improve tactics.

9 Key Findings and Recommendations

9.1 Key Findings

1. 'A' stores perform better than 'B' and 'C' stores in terms of size and average weekly sales.
2. Holiday weeks, especially Thanksgiving and Christmas, see a sharp increase in weekly sales.
3. Department and store size have a big impact on sales patterns.
4. The best performance was obtained by the Gradient Boosting Machine, which is why it is the recommended model for predictions.

9.2 Recommendations

1. Concentrate marketing efforts on Type 'A' stores and high-performing departments.
2. To enhance inventory management during busy times, apply the predictive model.
3. To keep models accurate, add new data on a regular basis.

10 References

- **Walmart Store Sales Forecasting Dataset:** Available on Kaggle [Kaggle Walmart Store Sales Forecasting].
- **XGBoost Documentation:** [XGBoost Documentation].

11 Feature Engineering

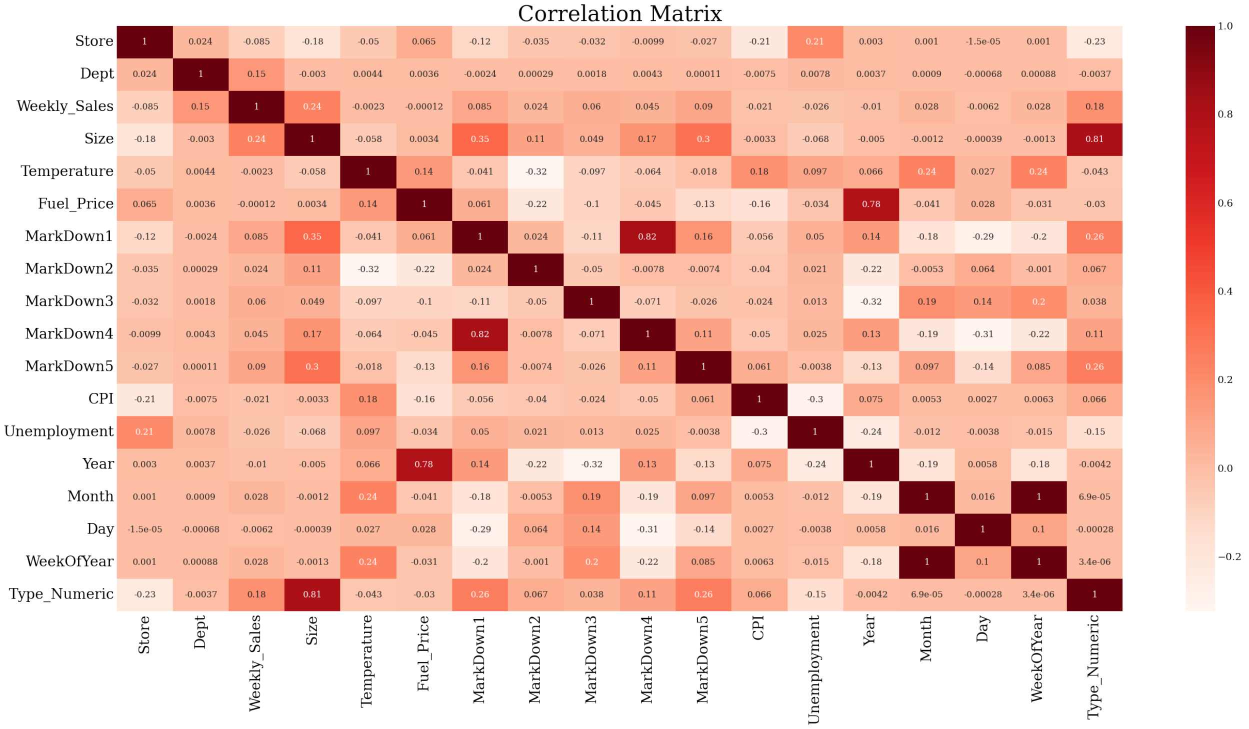


Figure 1: Coorelation Matrix

12 EDA Images

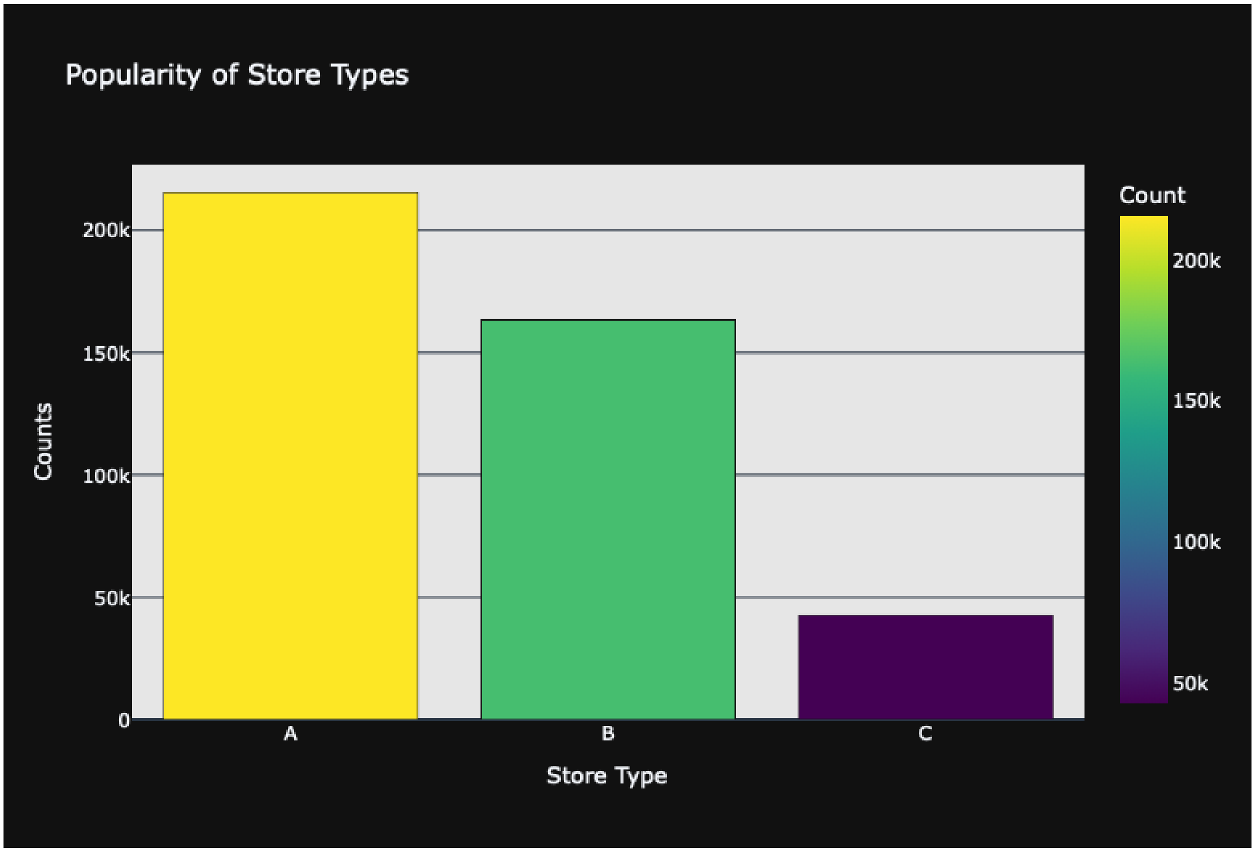


Figure 2: Popularity of Store Types

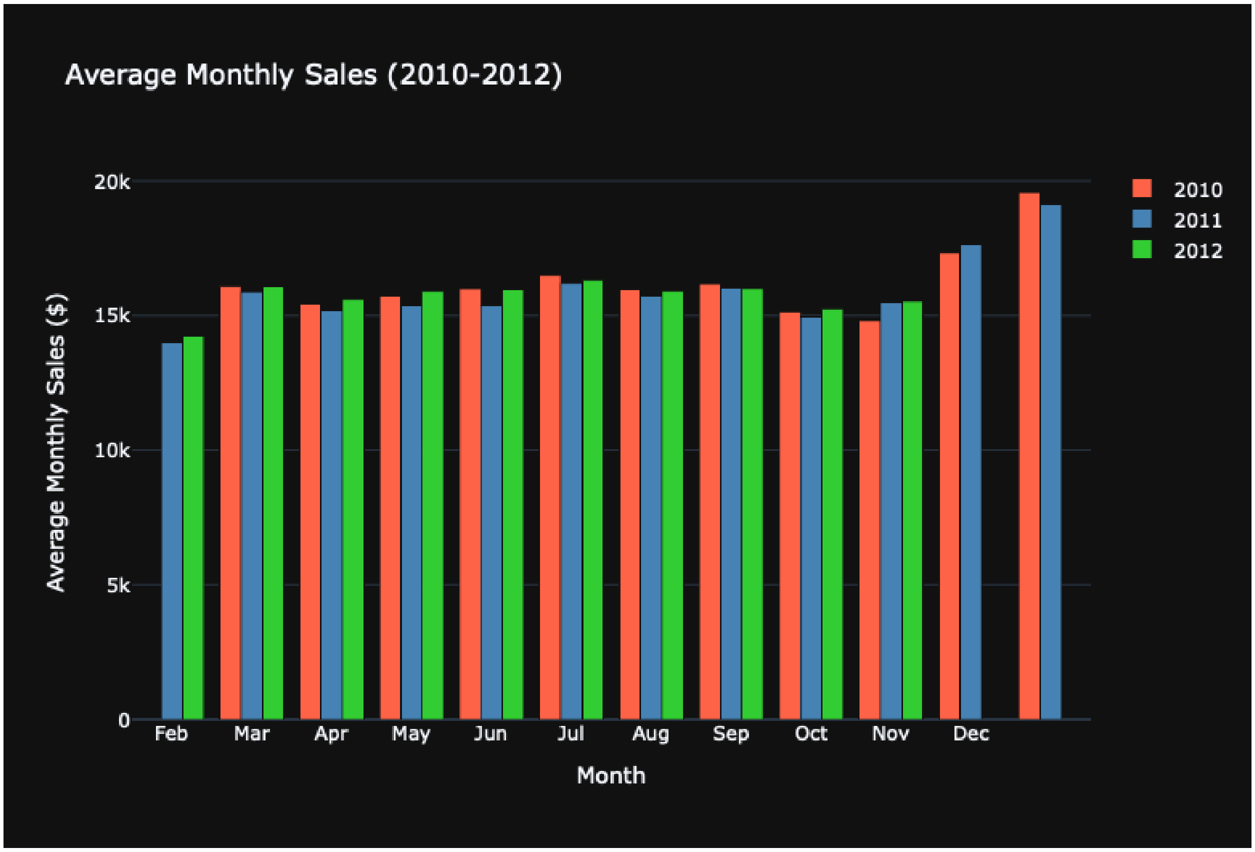


Figure 3: Average Monthly Sales

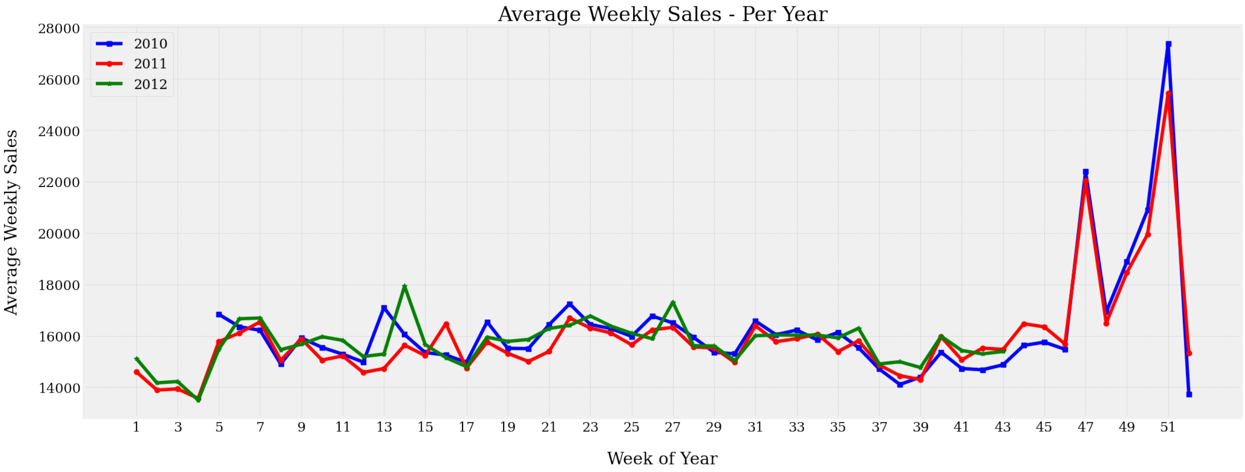
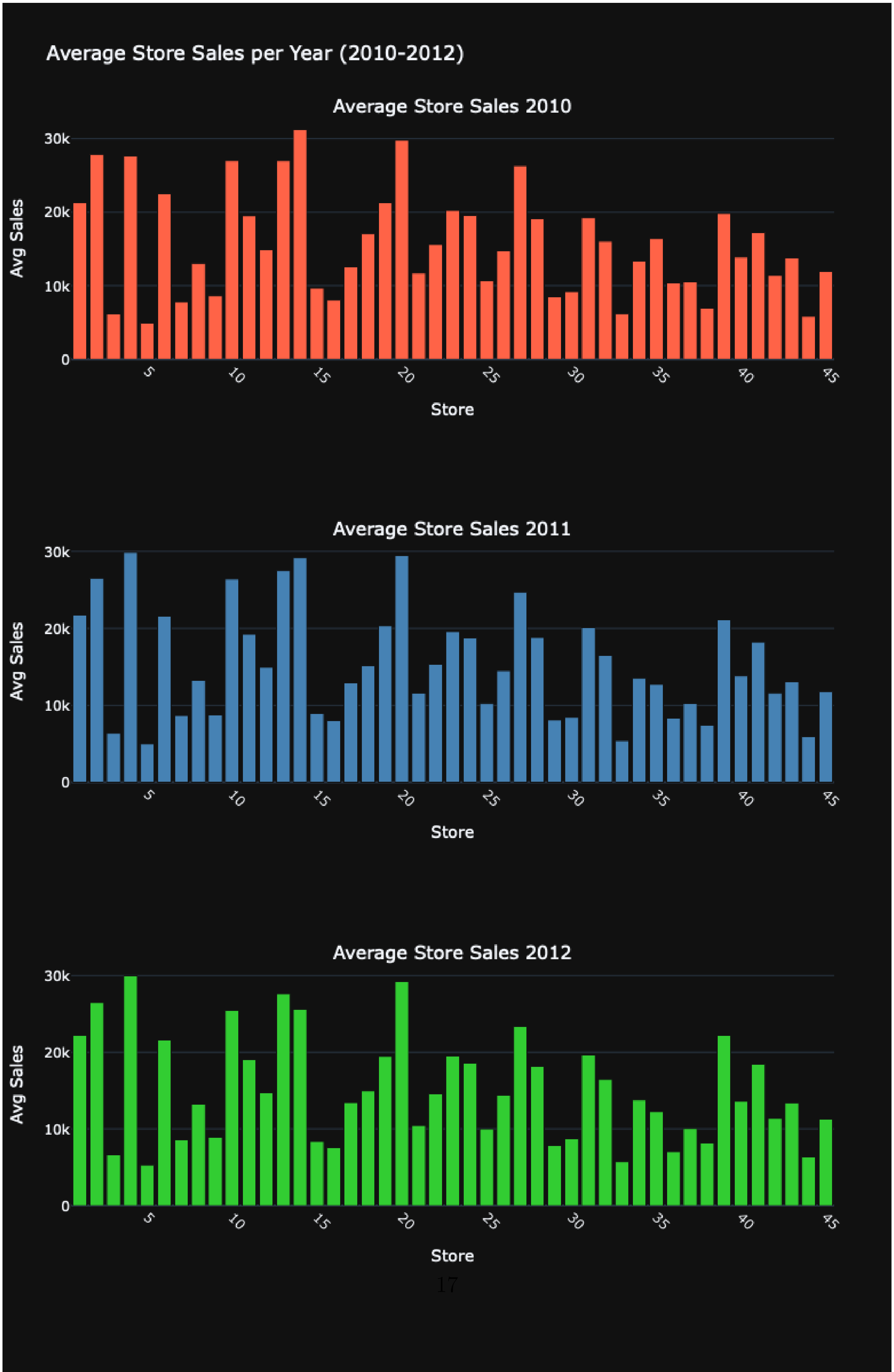


Figure 4: Average Weekly Sales - per Year


Average Store Sales per Year (2010-2012)
Average Store Sales 2010
Avg Sales
30k
20k
10k
0
5
10
15
20
25
30
35
40
45
Store
Average Store Sales 2011
Avg Sales
30k
20k
10k
0
5
10
15
20
25
30
35
40
45
Store
Average Store Sales 2012
Avg Sales
30k
20k
10k
0
5
10
15
20
25
30
35
40
45
Store

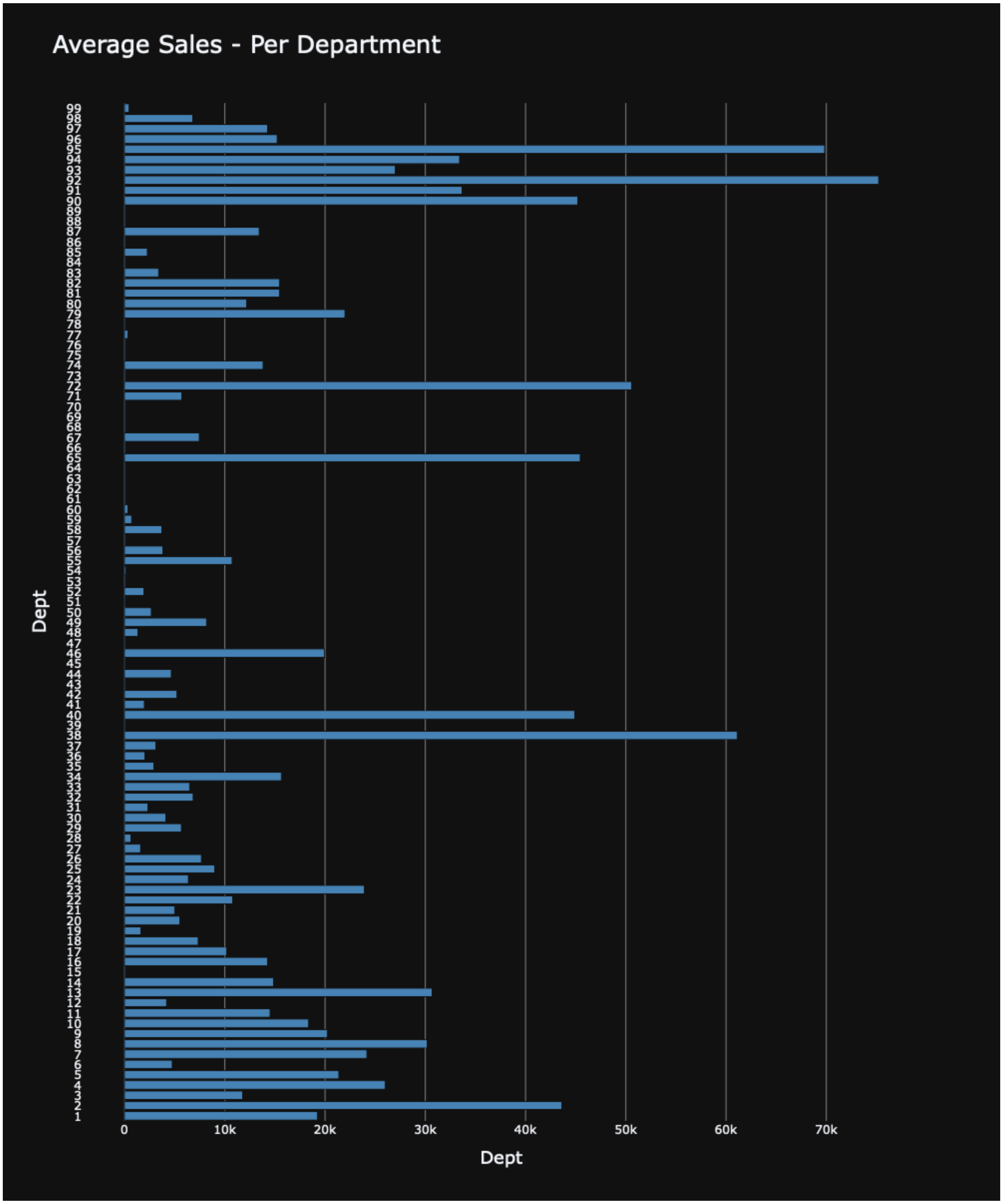


Figure 6: Average Department Sales

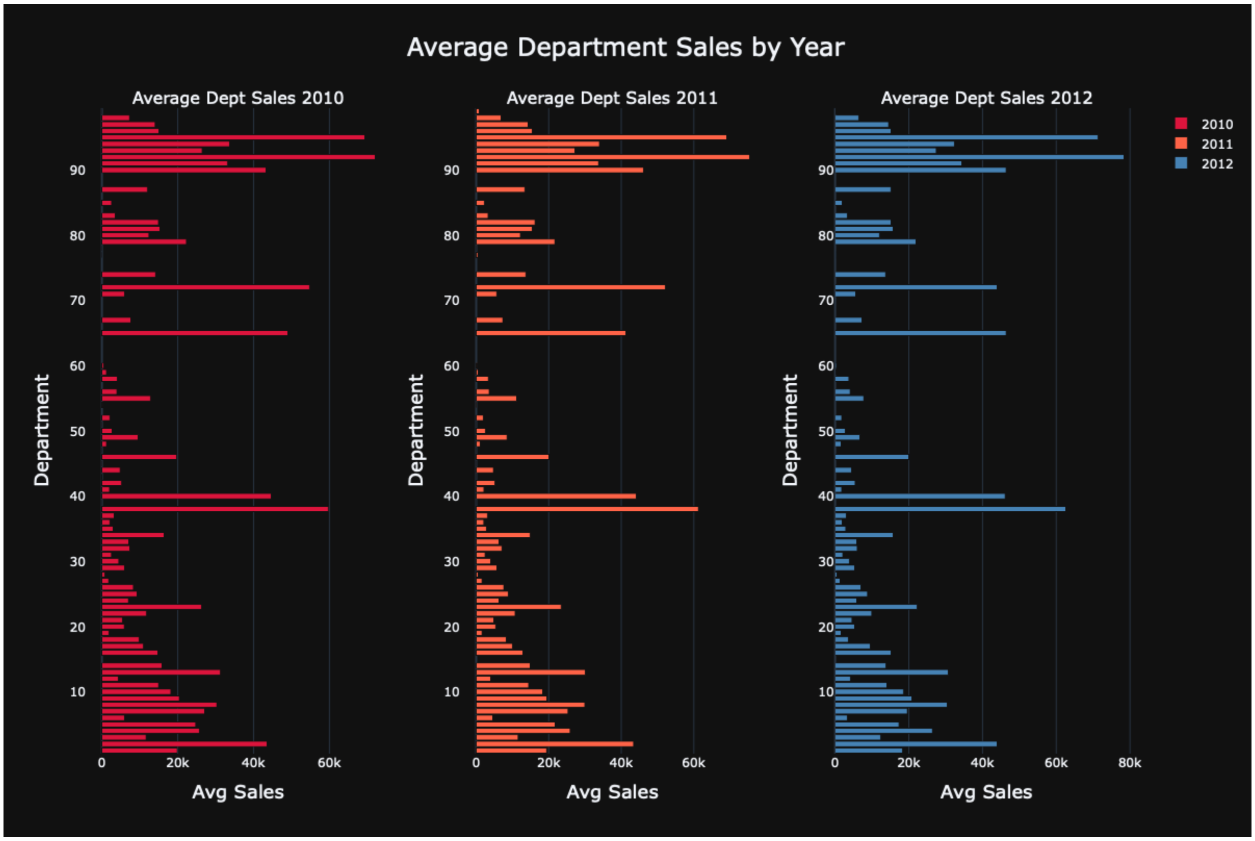


Figure 7: Average Department Sales - Per Year

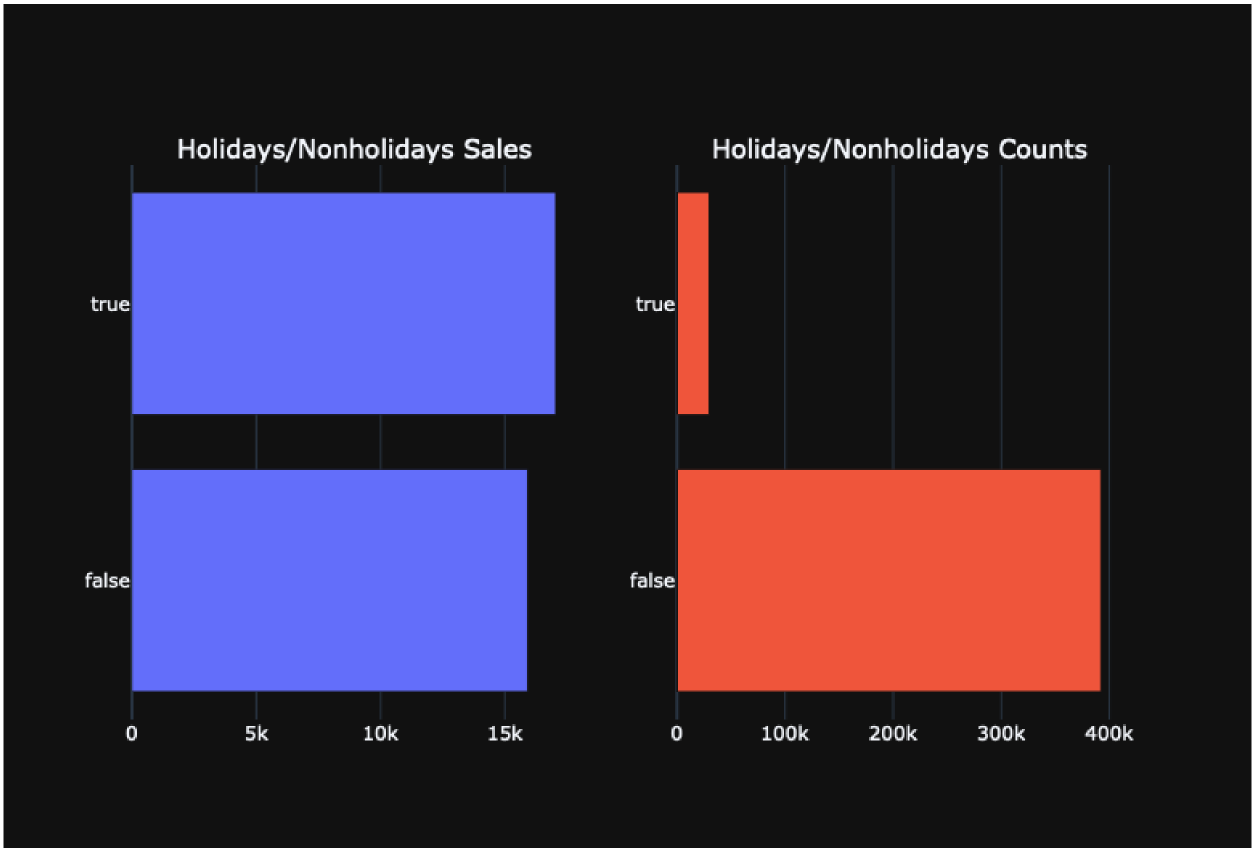


Figure 8: Holidays Vs Nonholidays Sales

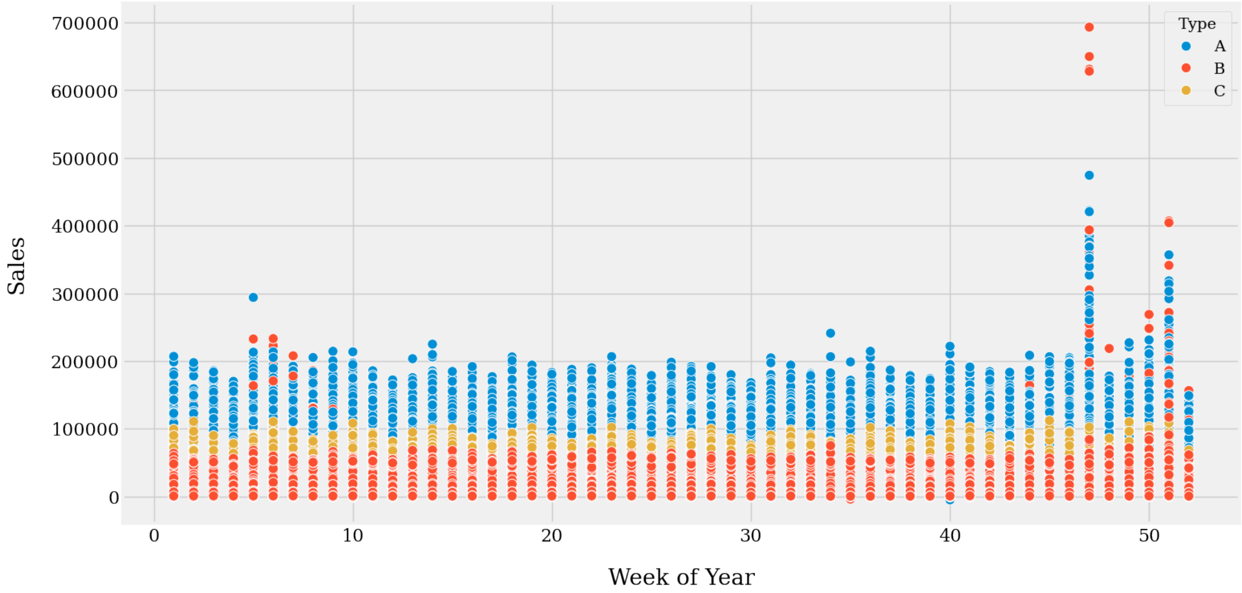


Figure 9: Week of Year vs Sales

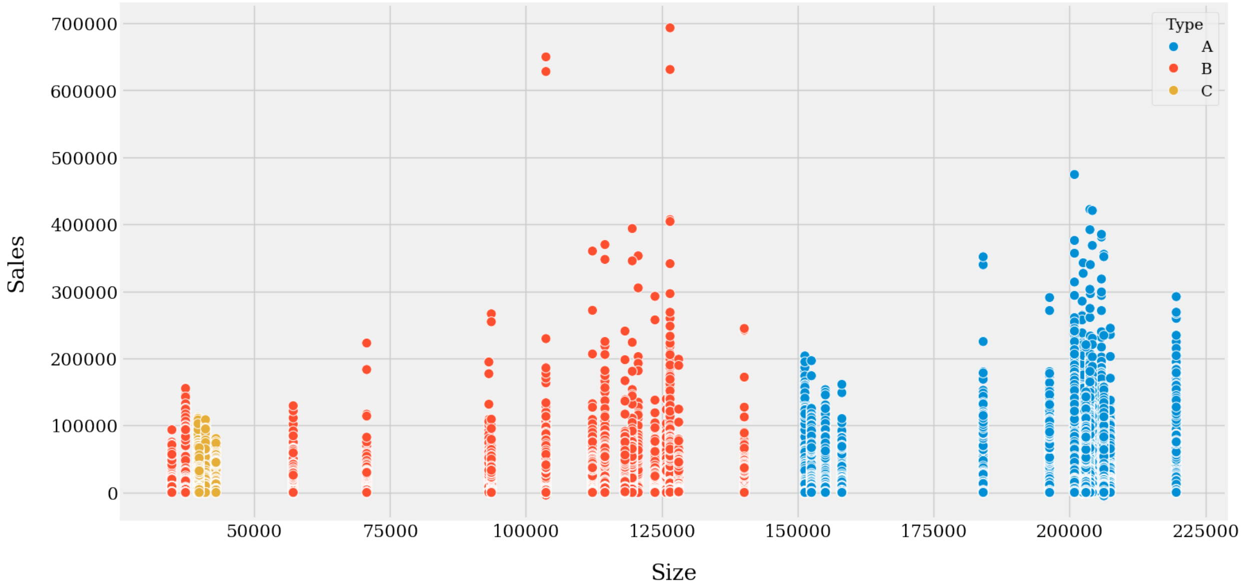


Figure 10: Size of Store vs Sales

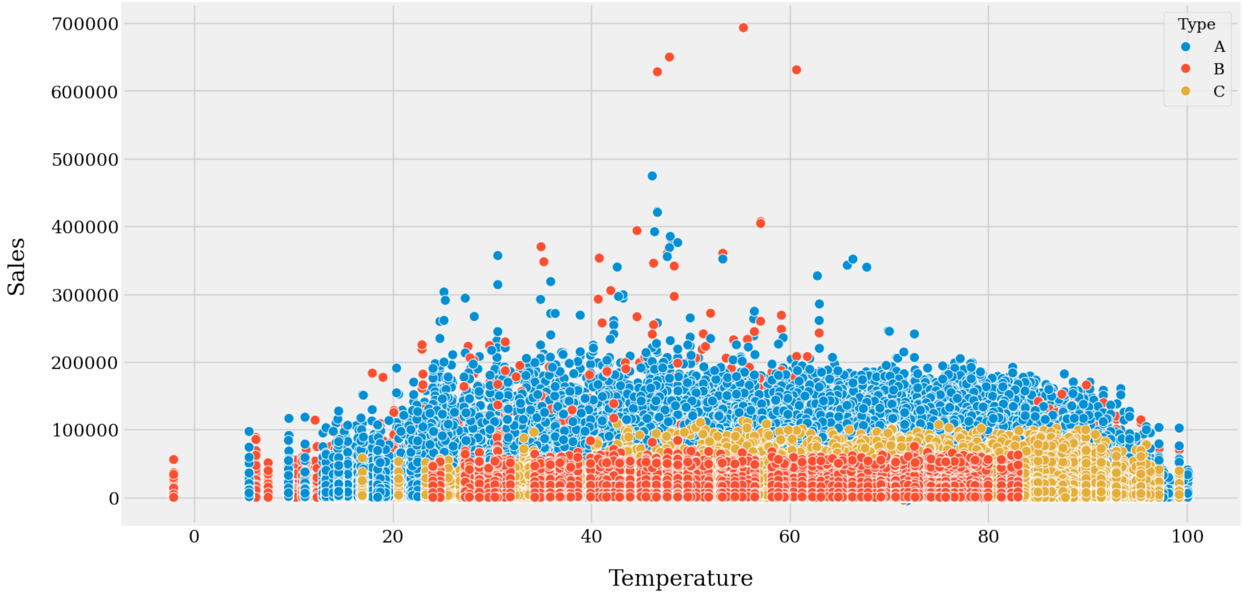


Figure 11: Temperature vs Sales

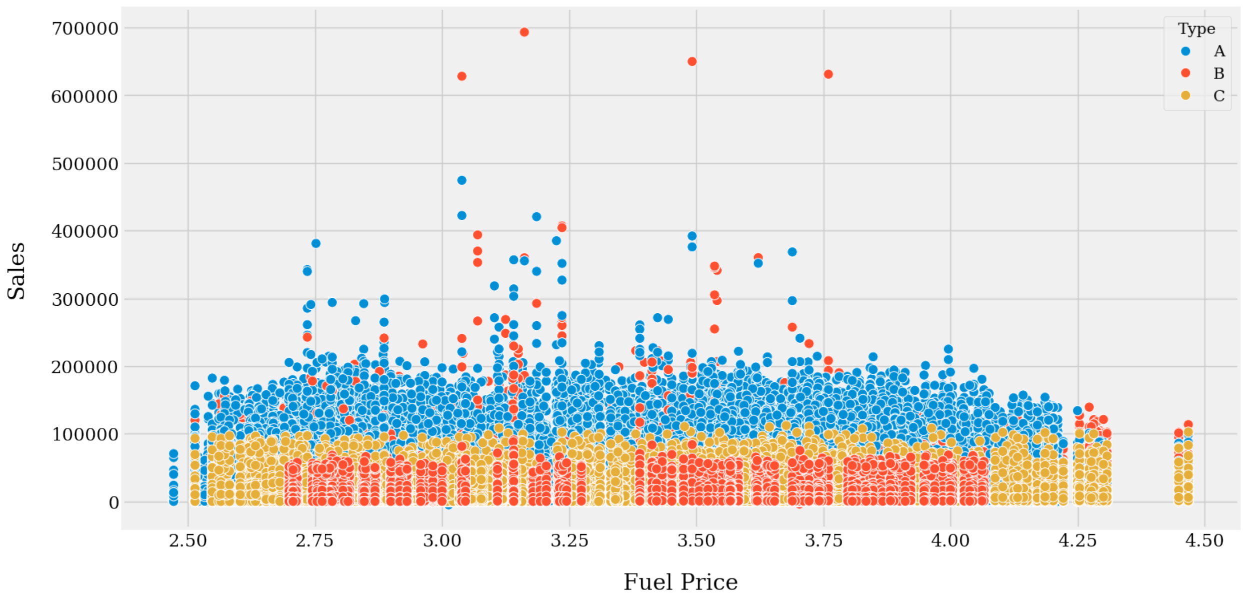


Figure 12: Fuel Price vs Sales

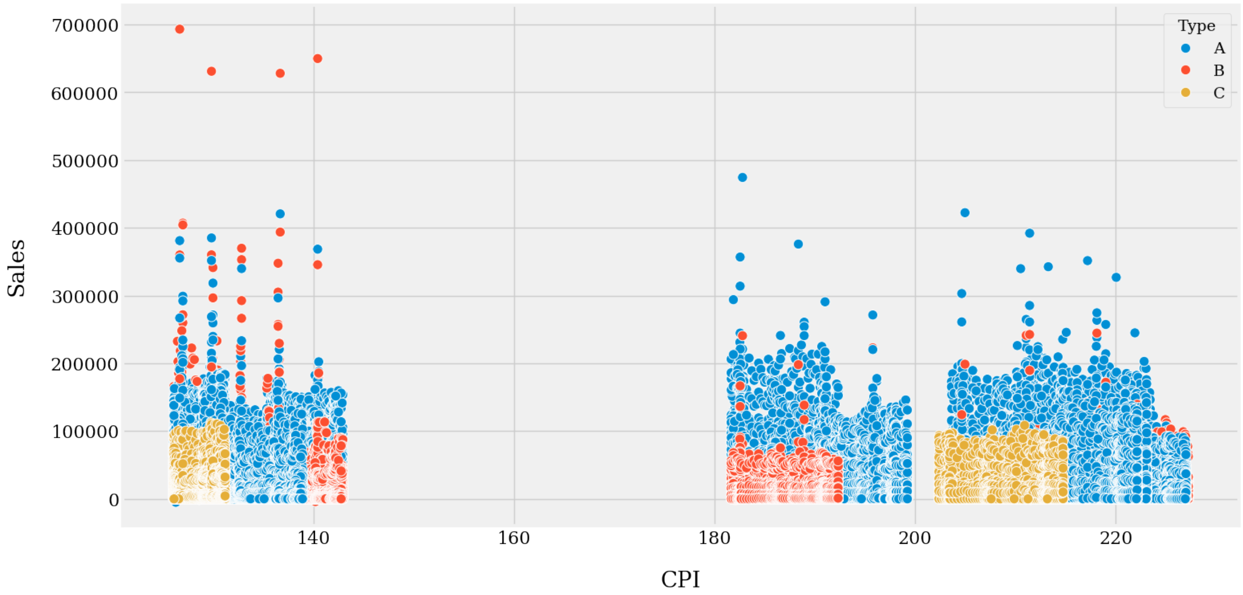


Figure 13: CPI vs Sales

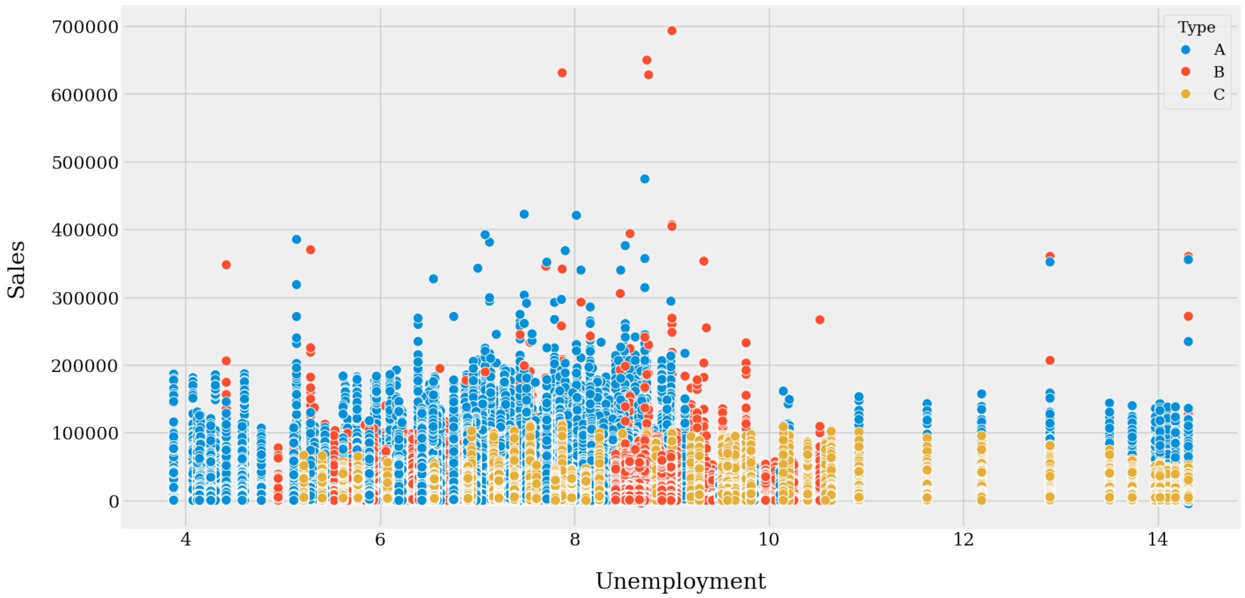


Figure 14: Unemployment vs Sales