



CAPSTONE PROJECT



PROJECT TITLE:

WALMART SALES ANALYSIS AND FORECASTING

Sekhar Chandra Padhi

Table of Contents:

Introduction

- 1.1 Problem Statement
- 1.2 Project Objective
- 1.3 Dataset

Data Cleaning and Preprocessing

- 2.1 Data Loading
- 2.2 Data Type Conversion
- 2.3 Handling Missing Values
- 2.4 Handling Duplicates
- 2.5 Outlier Detection
- 2.6 Outlier Treatment

Exploratory Data Analysis (EDA) and Insights

- 3.1 Unemployment Rate Impact
- 3.2 Seasonal Trends
- 3.3 Temperature Impact
- 3.4 Consumer Price Index (CPI) Effect
- 3.5 Top Performing Stores
- 3.6 Worst Performing Stores
- 3.7 Performance Difference Significance

Sales Forecasting

- 4.1 Data Preparation
- 4.2 Stationarity Check
- 4.3 Model Selection
- 4.4 Model Training and Evaluation

Conclusion & FAQ

1. Introduction

- Walmart, one of the largest retail giants globally, faces the ongoing challenge of efficiently managing inventory across its vast network of stores. The goal is to meet customer demand while keeping costs in check. However, predicting sales accurately can be quite complex due to various external factors such as economic indicators, weather patterns, and holiday seasons.
- This project aims to forecast Walmart's weekly sales for the upcoming 12 weeks using historical sales data from February 2010 to November 2012. The dataset includes information from 45 Walmart stores, with attributes like weekly sales, temperature, fuel prices, consumer price index (CPI), unemployment rates, and a holiday flag indicating special sales periods.
- The project begins with data cleaning, exploratory data analysis (EDA), and visualization to uncover relationships between various factors and weekly sales. Key insights are derived from analyzing the influence of temperature, CPI, holidays, and unemployment rates on sales performance. Following this analysis, the SARIMAX model is trained and evaluated for its predictive accuracy in forecasting future sales.
- The chosen forecasting method is time series analysis, specifically using the SARIMAX model, which is adept at handling datasets with seasonal patterns and external regressors.

1. Introduction

1.1 Problem Statement

The Retail giant- Walmart which has multiple outlets across the country is facing issues in managing the inventory in order to match the demand with respect to supply

1.1 Project Objective

This project aims to analyze historical sales data from Walmart stores to understand the factors influencing weekly sales and to develop a predictive model for forecasting sales for the next 12 weeks. The insights gained from this analysis will help Walmart optimize inventory management, pricing strategies, and promotional activities to improve business performance.



1. Introduction

1.2 Dataset

- The sales data is recorded starting from Feb-2010 till Nov-2012.
- There are total 6435 rows (143 rows each for 45 stores) and 8 columns (Features as mentioned above in the table)
- Temperature recorded is in degree Fahrenheit.
- Date column consists of one date of a particular week. Unemployment rate is indicated on every week record per store.

Feature Name & Description

- Store: Store number.
- Date: Week of sales.
- Weekly_Sales: Sales for the given week.
- Holiday_Flag: Whether the week is a special holiday week (1 – Holiday week, 0 – Non-holiday week).
- Temperature: Average temperature in the region.
- Fuel_Price: Cost of fuel in the region.
- CPI: Consumer Price Index.
- Unemployment: Unemployment rate.





2. Data Cleaning and Preprocessing



2.1 Data Loading


- The dataset was loaded into a pandas DataFrame using the `pd.read_csv()` function.

2.2 Data Type Conversion

- The 'Date' column was converted to datetime format using `pd.to_datetime()` to facilitate time series analysis.

2.3 Handling Missing Values




- 
- The dataset was checked for missing values using `df.isnull().sum()`. There were no missing values found.

2.4 Handling Duplicates

- The dataset was checked for duplicate rows using `df.duplicated().sum()`. No duplicates were found.

2.5 Outlier Detection



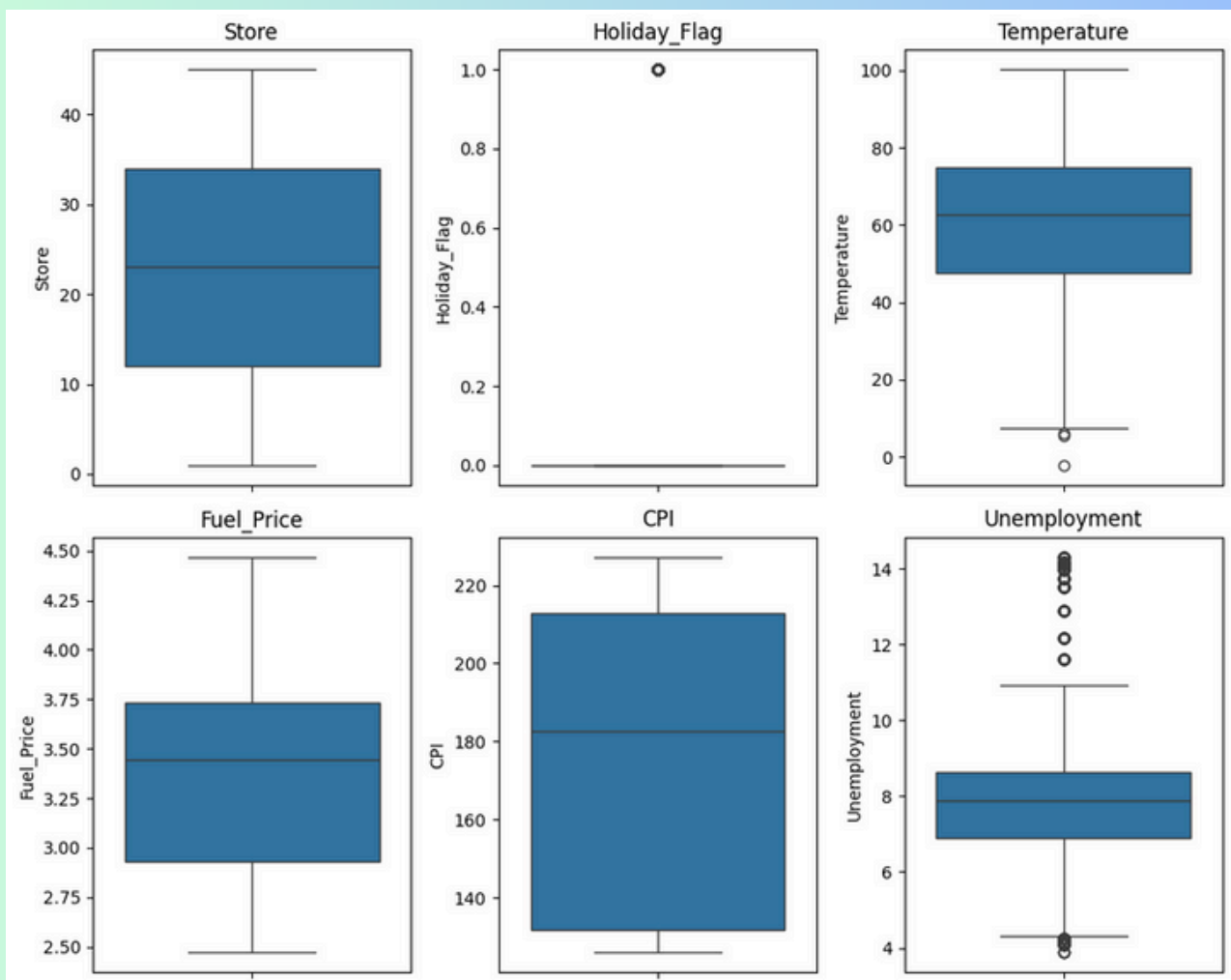
- Box plots were used to visually identify potential outliers in numerical columns like 'Weekly_Sales', 'Temperature', 'Fuel_Price', 'CPI', and 'Unemployment'.
- 

2. Data Cleaning and Preprocessing

2.6 Outlier Treatment

- The decision to not remove outliers using IQR or other techniques in this analysis is based on the understanding that outliers often represent genuine events and insights in retail sales data. By retaining outliers, the analysis aims to capture a more complete picture of sales dynamics and ensure a more accurate and robust understanding of the factors driving sales.

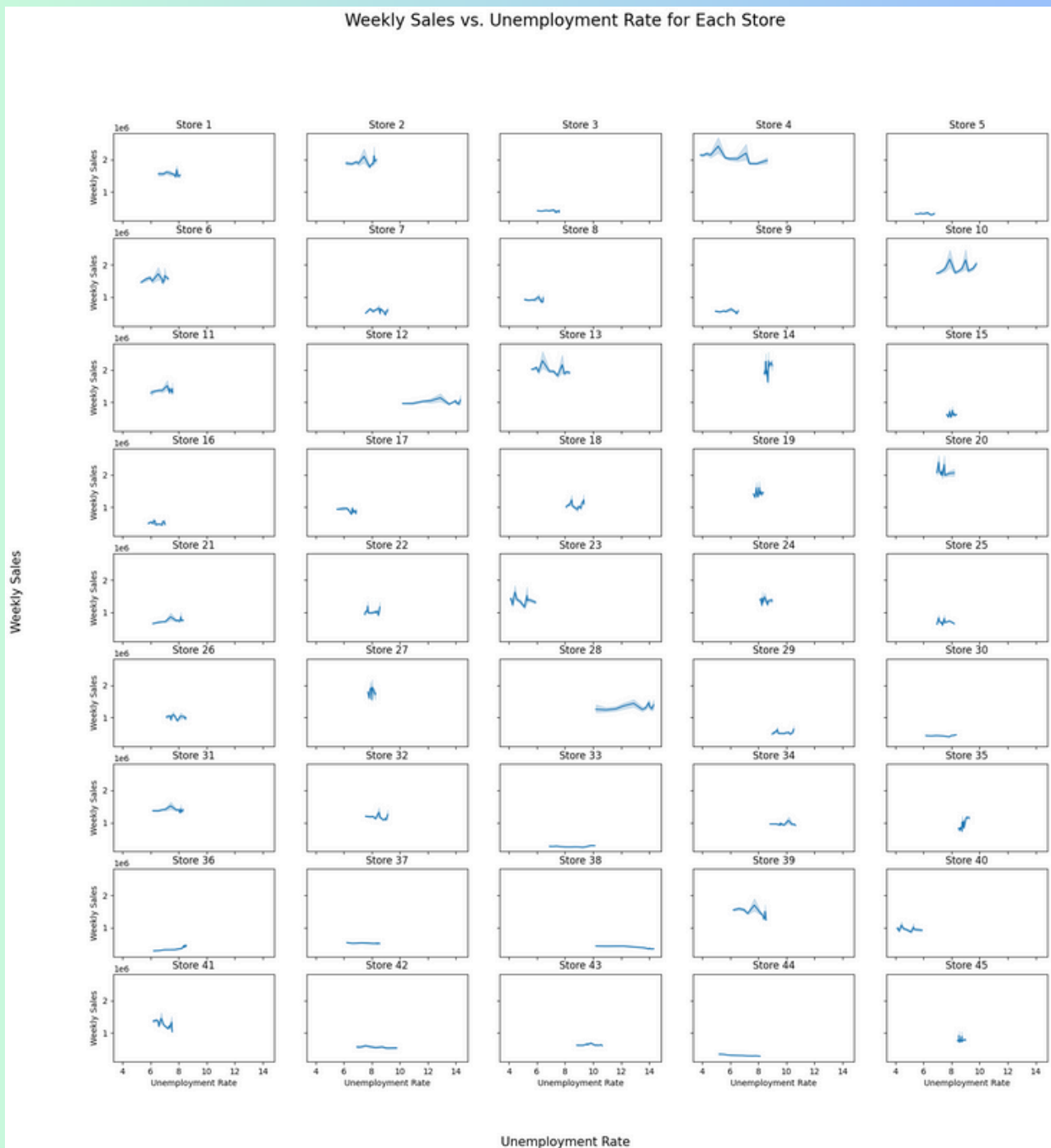
Boxplot for outliers visualization



3. Exploratory Data Analysis (EDA) and Insights

3.1 Unemployment Rate Impact

- Based on the analysis conducted, it appears that stores **3, 29, 30, 33, 38, 42, 43** and **45** are likely to be among those most significantly affected by changes in the unemployment rate.



3. Exploratory Data Analysis (EDA) and Insights

3.2 Seasonal Trends

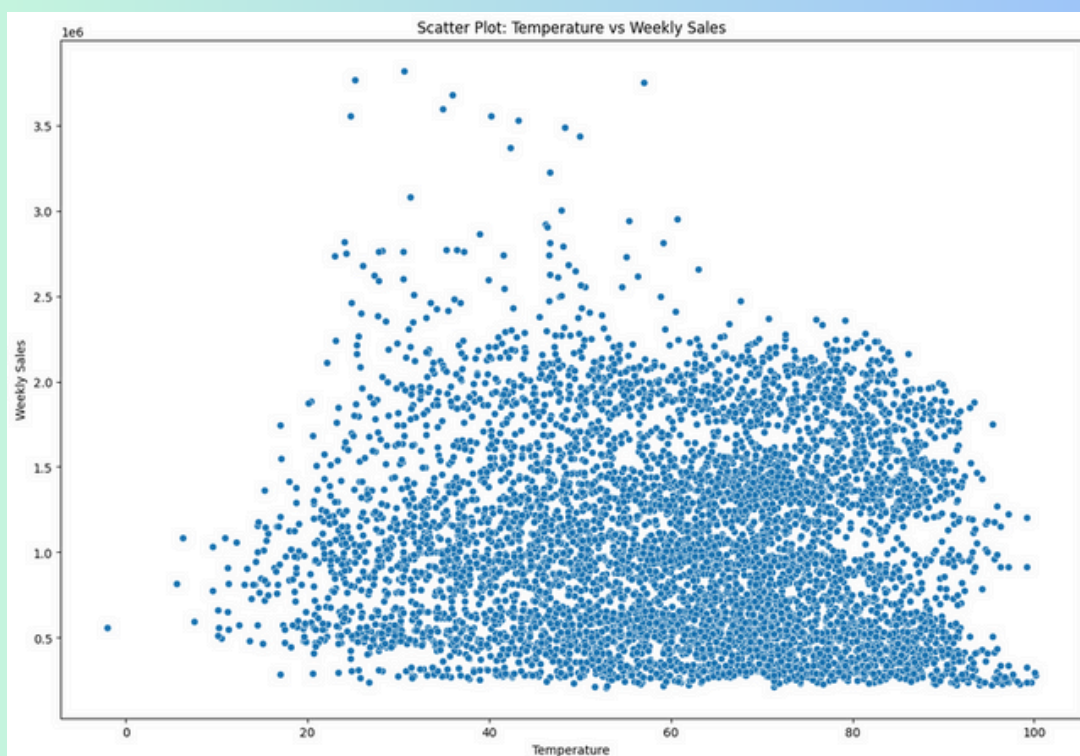
The weekly sales data exhibits a clear seasonal trend characterized by peaks and troughs occurring at regular intervals.

Possible reasons for this seasonal trend

- Holiday Season
- Promotions & Sales
- Summer Vacations
- Economic Factors

3.3 Temperature Impact

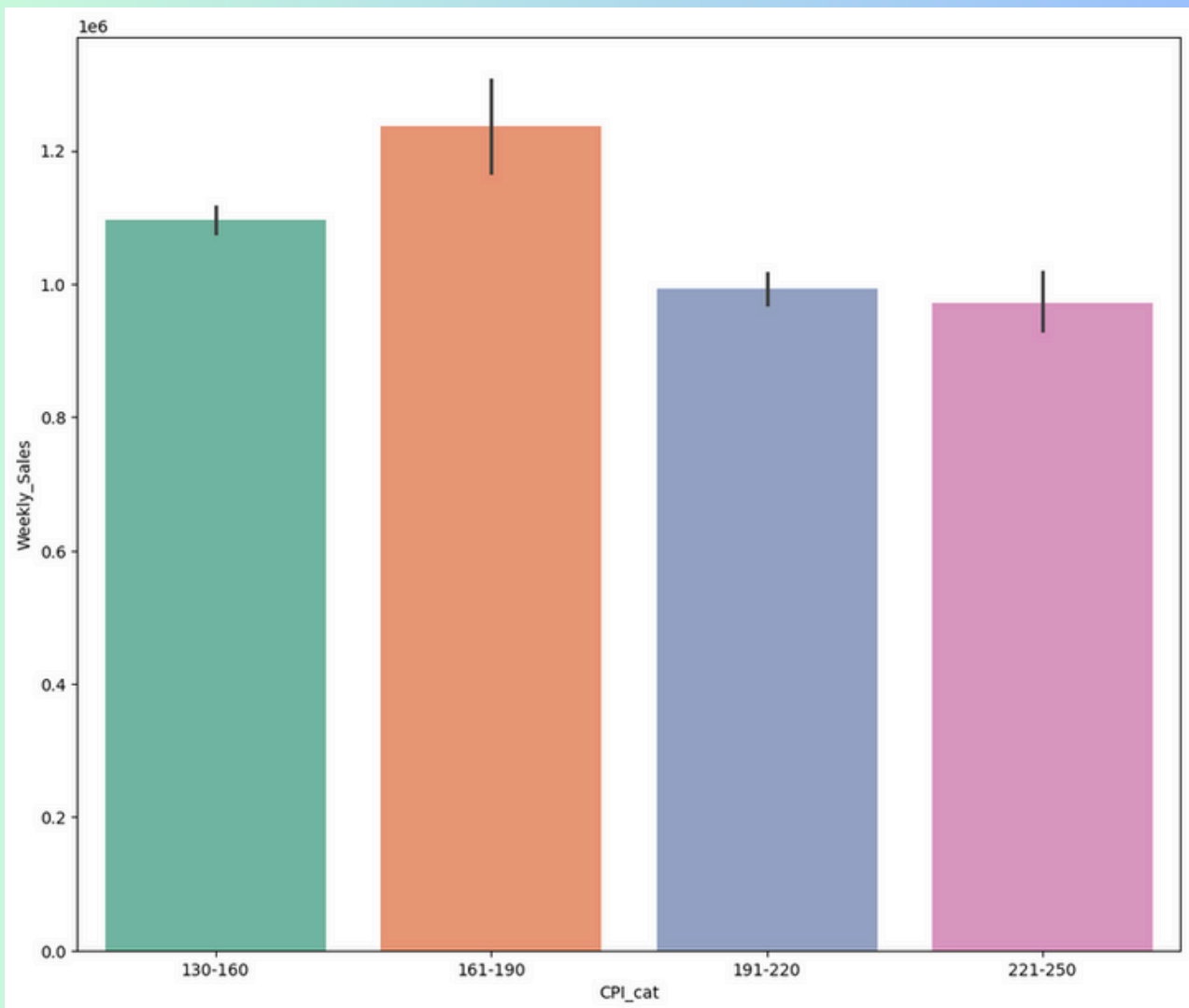
- A weak negative correlation (-0.042) was observed between temperature and weekly sales. While temperature does have an effect, it is likely minimal compared to other factors.



3. Exploratory Data Analysis (EDA) and Insights

3.4 Consumer Price Index (CPI) Effect

- Analysis revealed that as the CPI increases, weekly sales tend to decrease, highlighting the impact of inflation on consumer spending.



3. Exploratory Data Analysis (EDA) and Insights

3.5 Top Performing Stores

- Stores 2,4,6,10,13,14,20,27 were identified as the top-performing stores based on their total weekly sales.

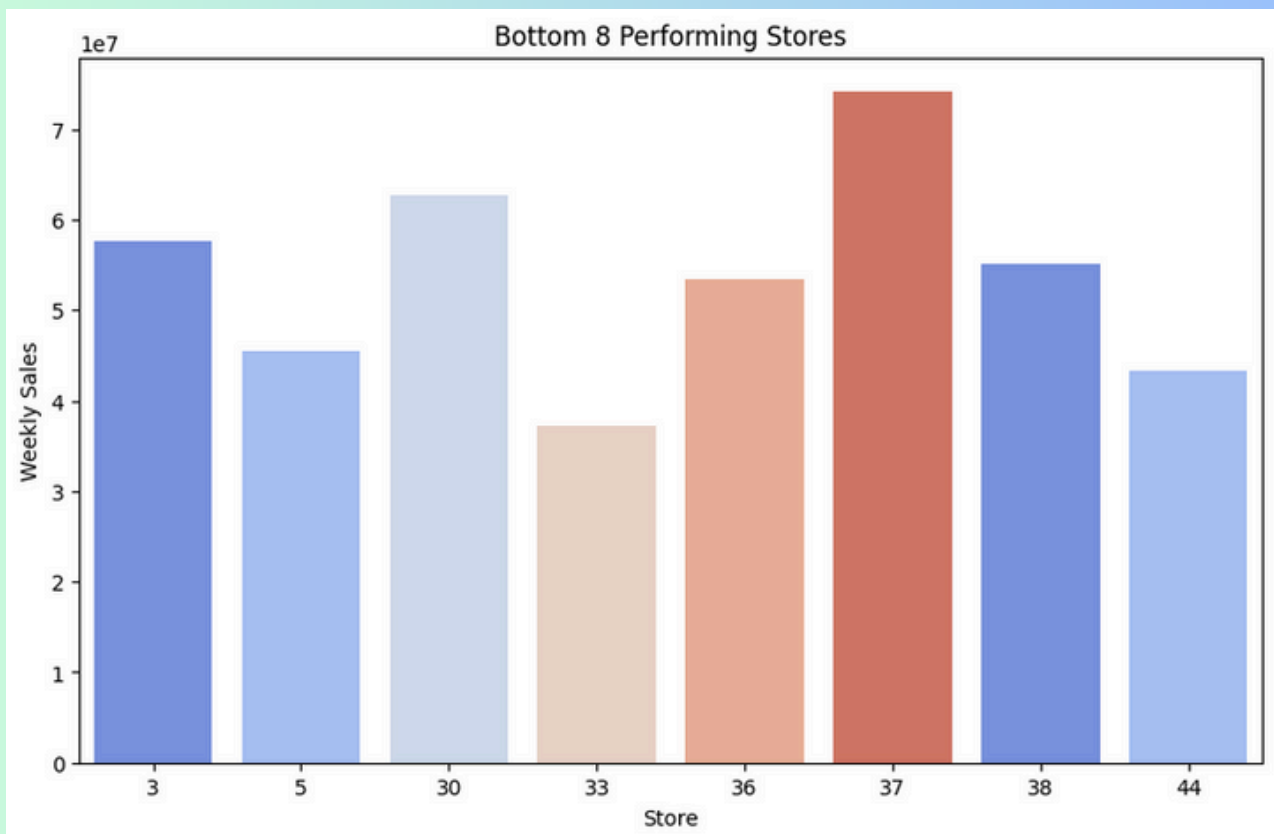
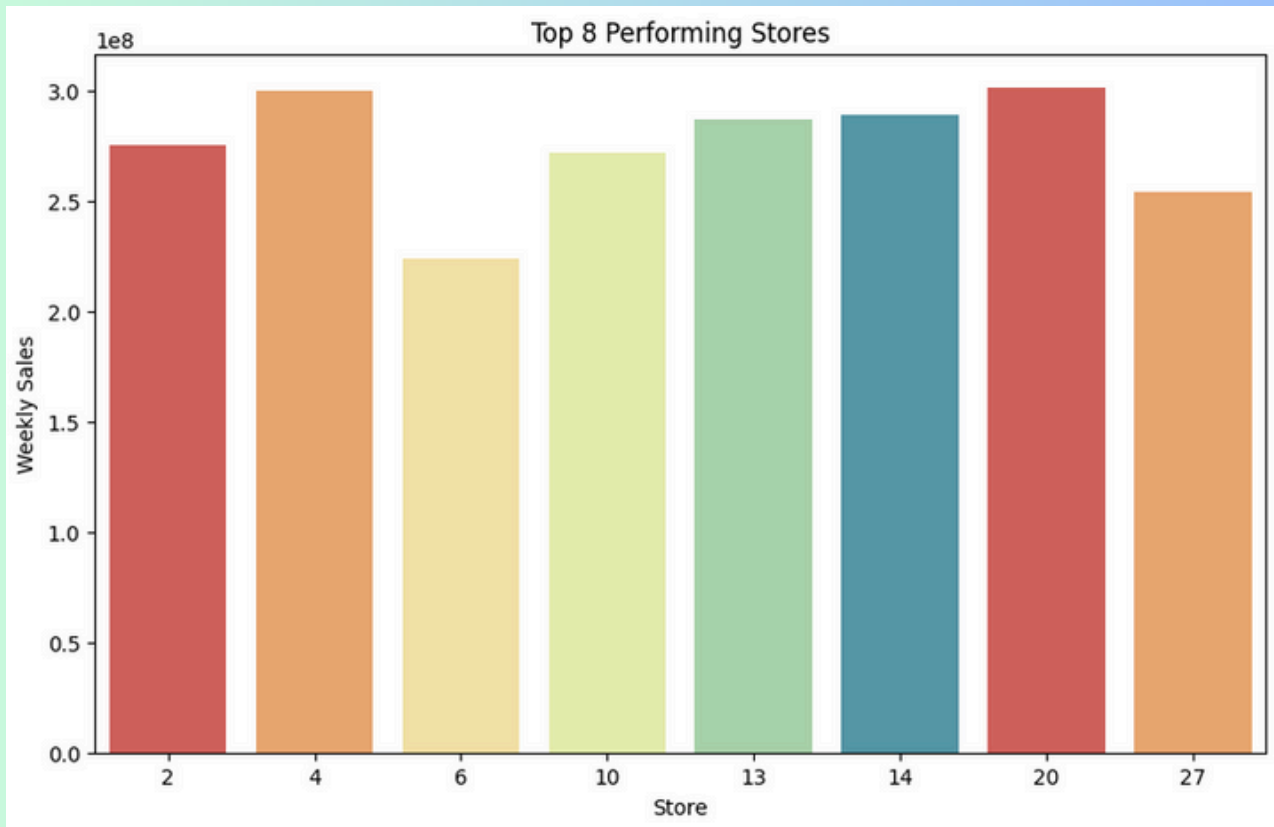
3.6 Worst Performing Stores

- Stores 3,5,30,33,36,37,38,44

3.7 Performance Difference Significance

- A t-test was conducted to compare the sales of the top-performing and worst-performing stores. The results indicated a statistically significant difference ($p\text{-value} < 0.05$) between the two groups, confirming that the performance gap is substantial.
- There is a significant difference between the highest and lowest performing stores.
- $p_value: 2.956591936392104e-12$

visualization





4. Sales Forecasting

4.1 Data Preparation

- As we are going to use this for time series analysis, let's convert the date column to an index.

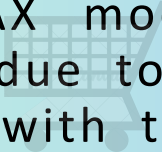
4.2 Stationarity Check

- The Augmented Dickey-Fuller (ADF) test was performed to check the stationarity of the time series data. The results indicated that the data is stationary, which means there is no need for differencing.



4.3 Model Selection

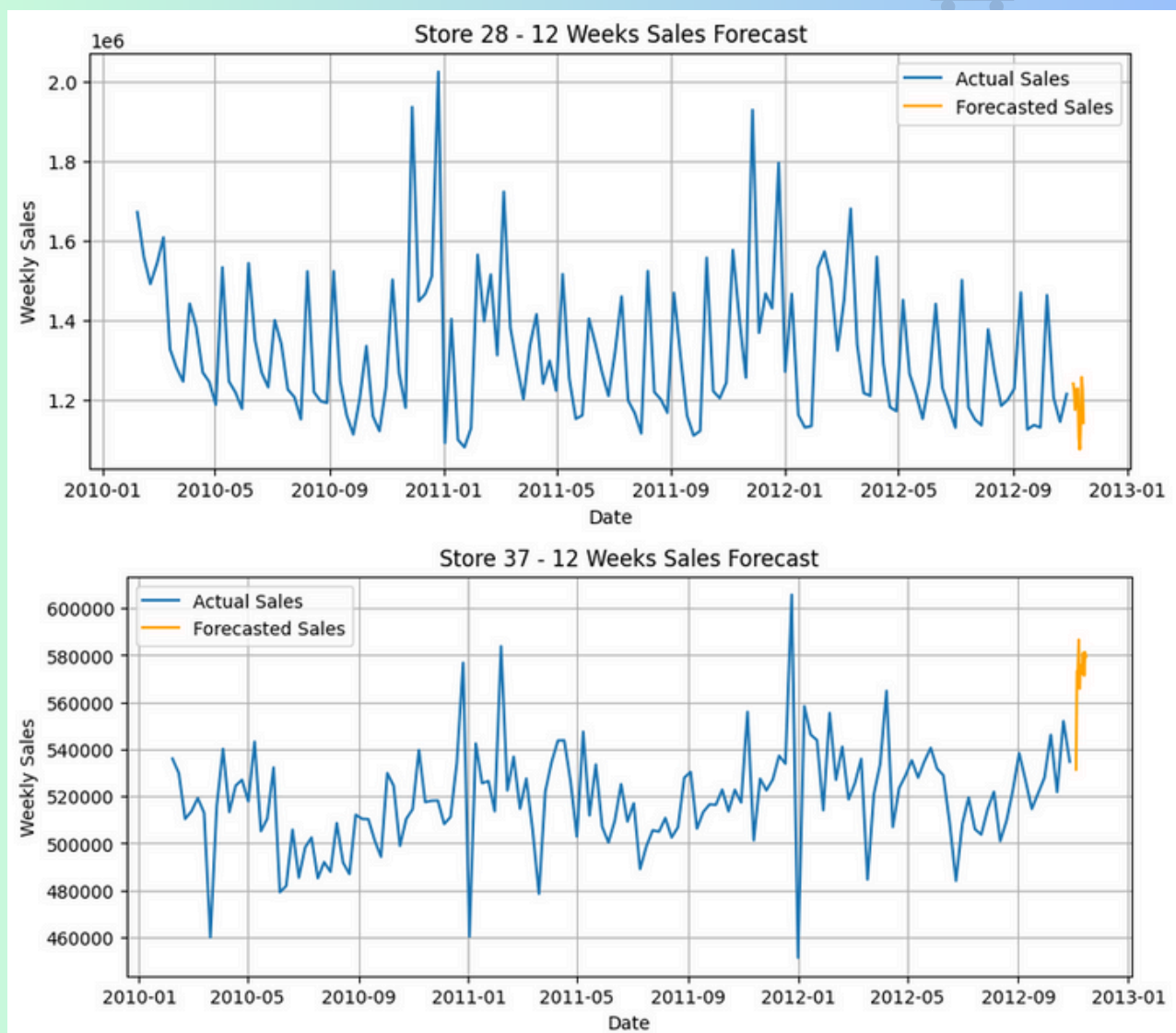
- To forecast sales for the next 12 weeks, opting for time series analysis was the most effective strategy. While one might initially consider Linear Regression due to the numerical nature of the target column, the time-dependent characteristic of the target variable requires more specialized algorithms. Time series methods like ARIMA and SARIMAX are tailored for such tasks.
- The SARIMAX model was chosen for sales forecasting due to its ability to handle time series data with trends and seasonality. The `auto_arima` function was used to automatically determine the optimal parameters (p, d, q) for the SARIMA model.



4. Sales Forecasting

4.4 Model Evaluation

- The code iterates through each store, fits a SARIMAX model to its historical sales data, and generates a 12-week forecast. The chosen model parameters (order=(4, 1, 5), seasonal_order=(4, 1, 5, 12)) might require optimization based on specific data characteristics. The forecasts, along with confidence intervals, are plotted for each store, providing a visual representation of predicted sales trends.







5. Conclusion

- This project successfully analyzed historical sales data from Walmart stores, identified key factors influencing weekly sales, and developed predictive models to forecast sales for the next 12 weeks.

Key Findings:

- Unemployment rate and CPI have a significant impact on weekly sales.
- There is an inverse relationship between temperature and weekly sales, indicating seasonal trends.
- There is a significant difference in performance between the top and worst-performing stores.

Recommendations:

- Walmart should closely monitor unemployment rates and CPI to adjust pricing and inventory strategies accordingly. Promotional activities should be aligned with seasonal trends to capitalize on sales opportunities.
 - Strategies for improving the performance of the worst-performing stores should be explored.
- 
- 



5. Conclusion

Limitations:

- The models were trained on a limited dataset, and their accuracy may vary with new data. External factors not included in the dataset (e.g., competitor actions, economic conditions) could influence future sales.
- This analysis provides valuable insights for Walmart to optimize its operations and improve business performance. Further research and model refinement can enhance the accuracy of sales forecasts and enable more effective decision-making.



FAQ

Given that our data does not exhibit any obvious trend or pattern, why we have used the SARIMAX model?

Hidden Seasonality:

While initial analysis might not reveal obvious seasonality, there could be subtle seasonal patterns that SARIMA can capture. The weekly sales data shows some recurring fluctuations, particularly during holiday seasons or specific times of the year (refer to question (b) inferences), and temperature which is an external factor which usually causes seasonality in this case showing a weak correlation but overall with external factor data and date we had some seasonal pattern and trends. These patterns might not be prominent but can still influence the model's accuracy. SARIMA is designed to identify and model such hidden or complex seasonal patterns.



*Thank
you!*