Capstone Project DSAI

Problem Statement 1

a. Problem Statement

A retail store with multiple outlets across the country is facing issues in managing inventory to match demand with supply. As a data scientist, the task is to derive useful insights from the data and create prediction models to forecast sales for a specified period (e.g., the next 12 weeks).

b. Project Objective

1. **Generate insights** that each store can use to improve various operational areas.
2. **Forecast sales** for each store for the next 12 weeks to help in inventory management.

c. Data Description

The dataset walmart.csv consists of 6435 rows and 8 columns:

- **Store**: Store number.
- **Date**: Week of sales.
- **Weekly_Sales**: Sales for the given store in that week.
- **Holiday_Flag**: Indicates if it is a holiday week (1 for holiday week, 0 otherwise).
- **Temperature**: Temperature on the day of the sale.
- **Fuel_Price**: Cost of fuel in the region.
- **CPI**: Consumer Price Index.
- **Unemployment**: Unemployment rate.

d. Data Pre-processing Steps and Inspiration

1. **Handling Missing Values**: To ensure there are no missing values in the dataset.
2. **Date Parsing**: Convert the Date column to a datetime object for easier manipulation.
3. **Feature Engineering**: Create additional features such as week number, month, or year.
4. **Normalization/Scaling**: Normalize or scale numerical features for better model performance.
5. **Handling Categorical Data**: Convert categorical data into numerical form using encoding techniques.

6. **Outlier Handling**: Using the IQR method for outlier identification. A total of 37 outliers were removed.

The dataset has been pre-processed, with the Date column converted to a datetime object and new columns for Year, Month, and Week added. Here are some basic statistics of the dataset:

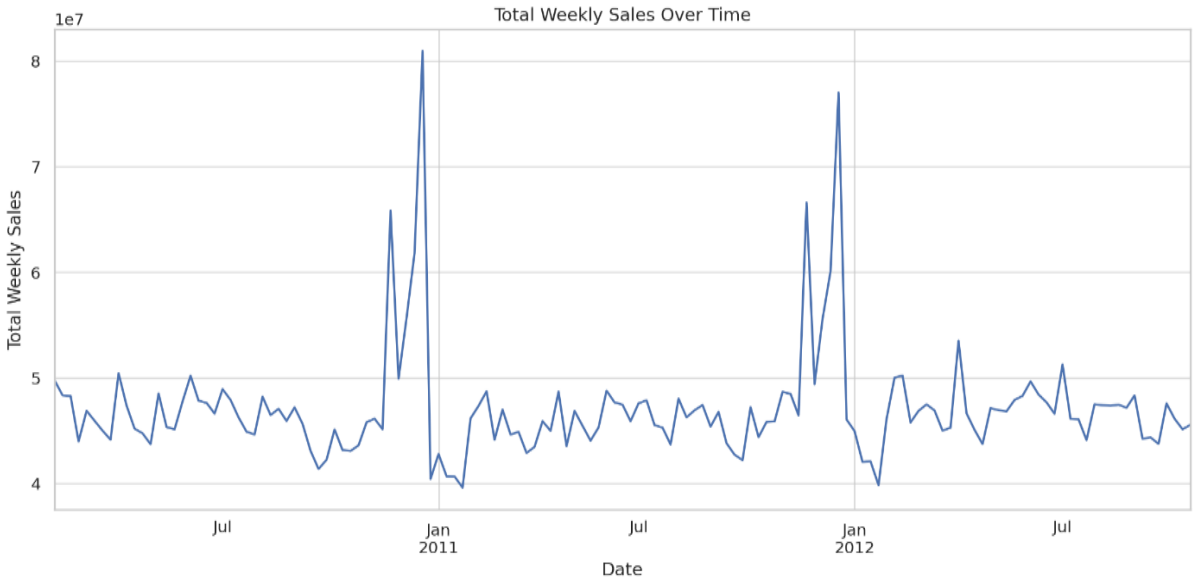
- **Weekly_Sales**: Mean = 1,046,965, Std = 564,367, Min = 209,986, Max = 3,818,686
- **Temperature**: Mean = 60.66, Std = 18.44, Min = -2.06, Max = 100.14
- **Fuel_Price**: Mean = 3.36, Std = 0.46, Min = 2.47, Max = 4.47
- **CPI**: Mean = 171.58, Std = 39.36, Min = 126.06, Max = 227.23
- **Unemployment**: Mean = 8.00, Std = 1.88, Min = 3.88, Max = 14.31

Inspiration for these steps comes from the need to prepare the data in a format that is suitable for time series forecasting and to ensure the integrity and consistency of the data for accurate model training and evaluation.

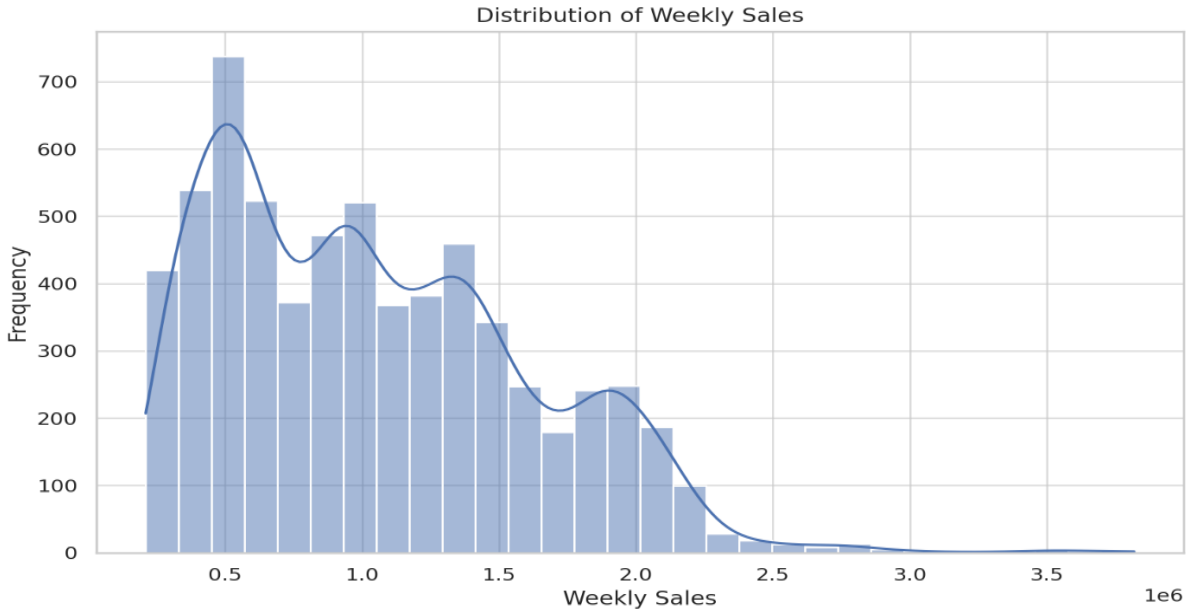
Exploratory Data Analysis (EDA)

Here are the visualizations and insights from the data:

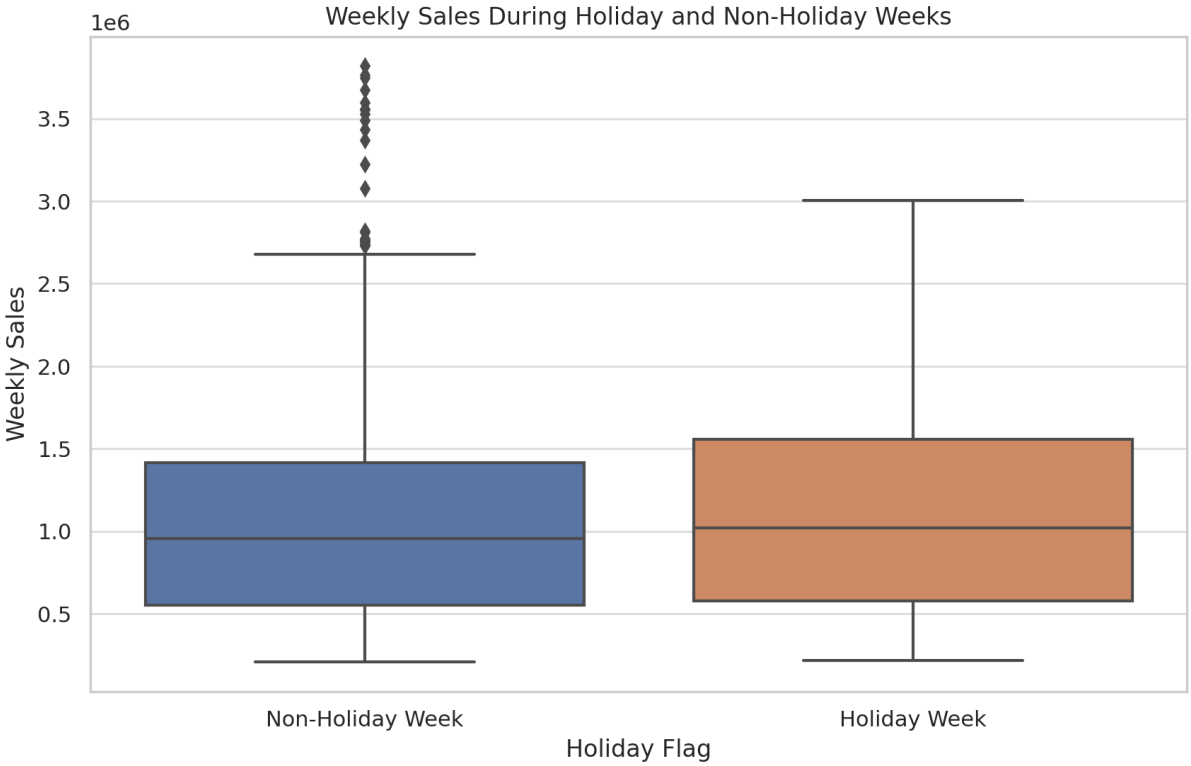
1. **Sales Trends Over Time**:
    - The total weekly sales show a clear trend over time with noticeable spikes and dips.
    - There is a seasonal pattern in the sales data, indicating periods of higher and lower sales throughout the year.



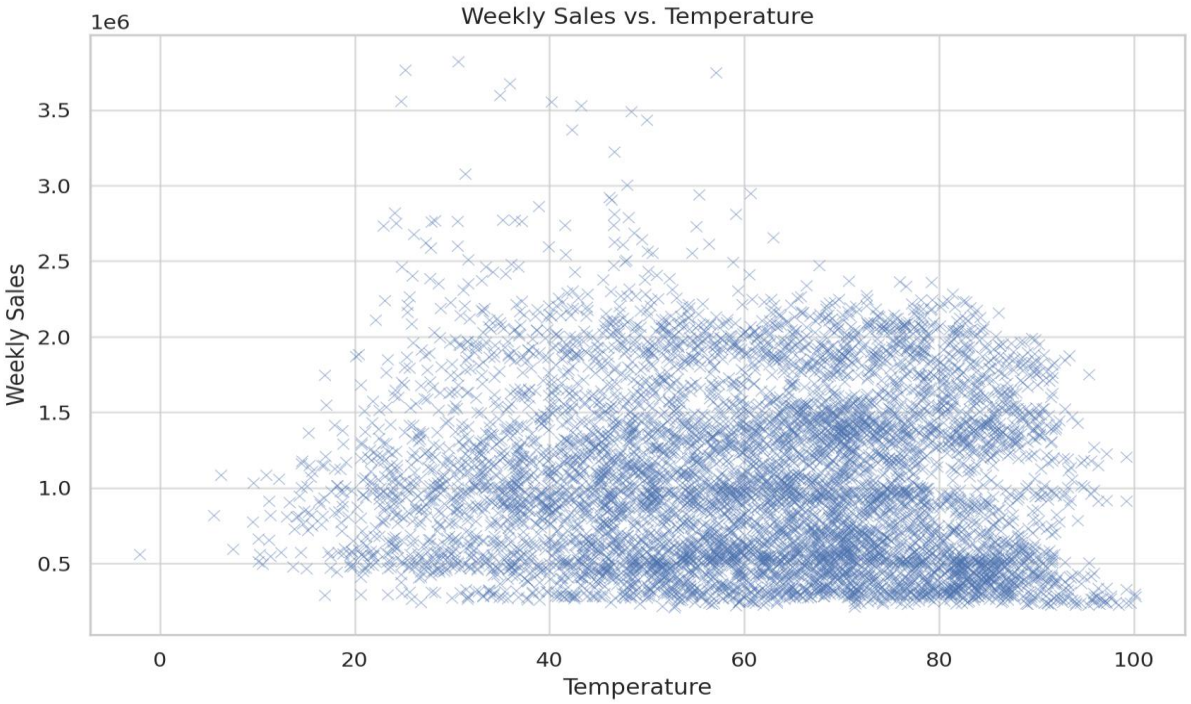
2. **Sales Distribution**:
    - The distribution of weekly sales is right-skewed, with most sales concentrated around the lower end and a few weeks with very high sales.



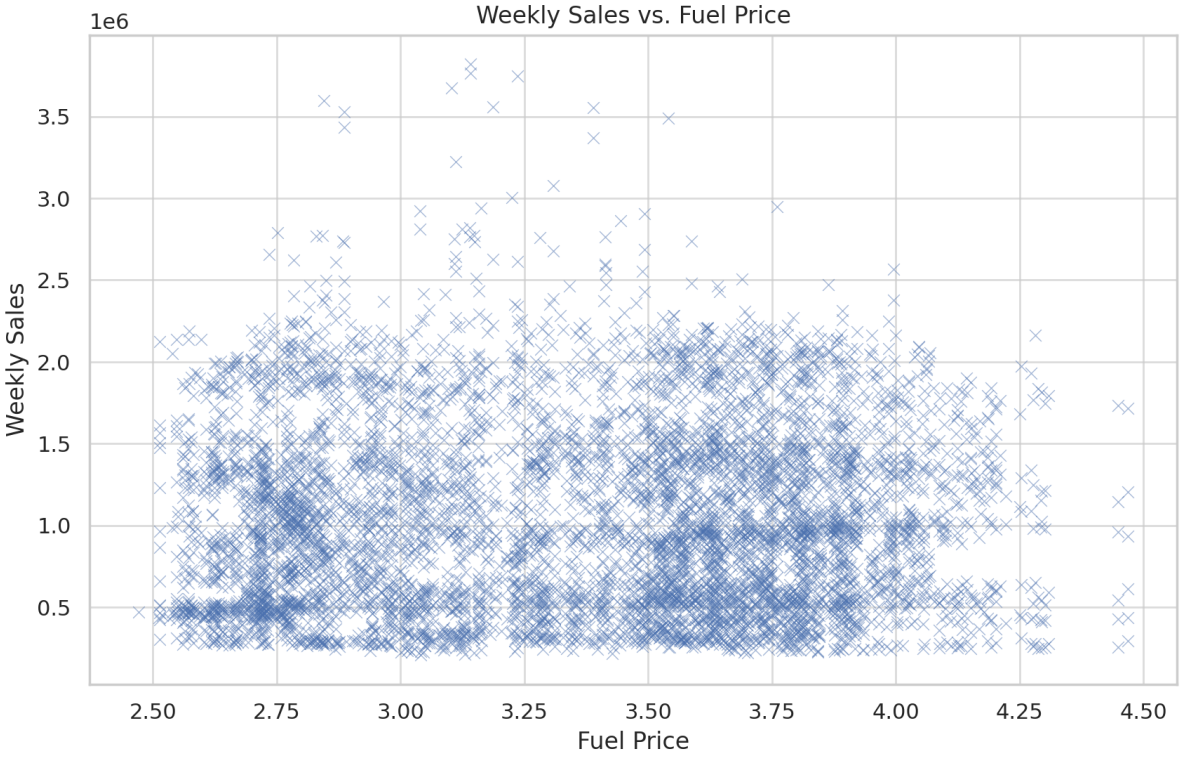
3. **Impact of Holidays on Sales**:
    - Sales during holiday weeks tend to be higher than non-holiday weeks. This indicates a positive impact of holidays on sales.



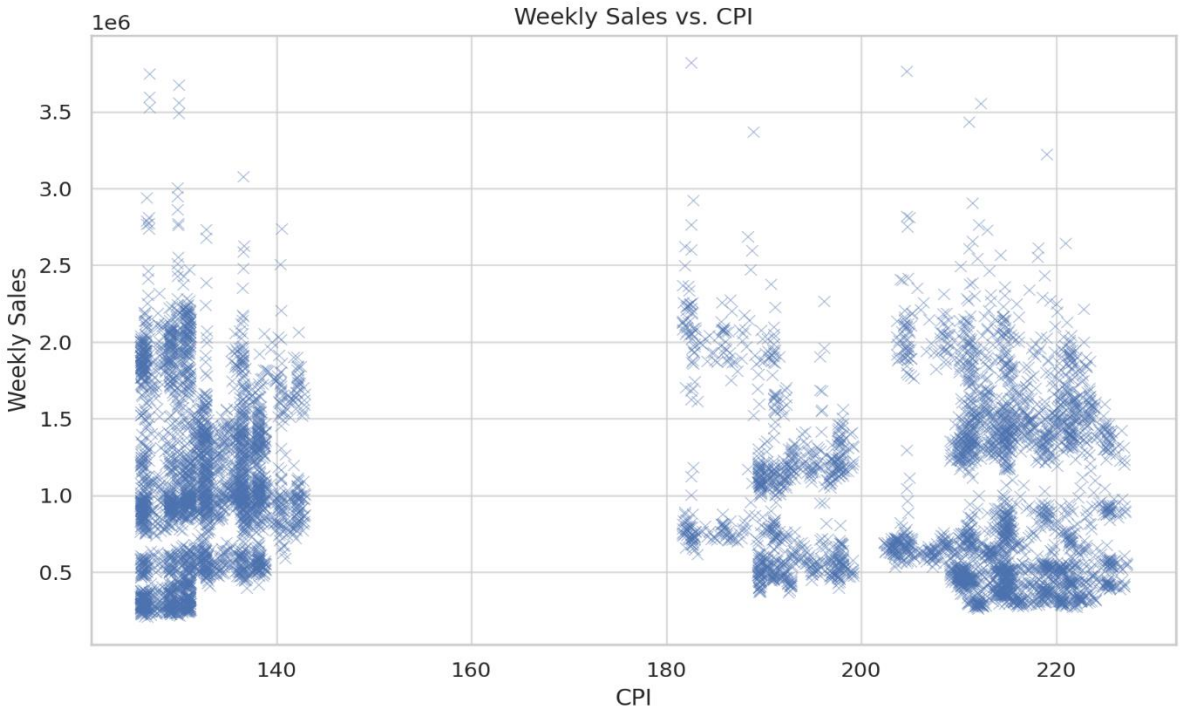
4. **Impact of Temperature on Sales**:
    - There is a weak relationship between temperature and weekly sales. Sales do not show a strong correlation with temperature.



5. **Impact of Fuel Price on Sales**:
    - Sales do not show a strong correlation with fuel prices, indicating that changes in fuel prices have minimal impact on weekly sales.



6. **Impact of CPI on Sales**:
    - There is no significant correlation between CPI and weekly sales. The consumer price index does not appear to influence weekly sales significantly.



7. **Impact of Unemployment on Sales**:
    - There is a slight negative correlation between unemployment and weekly sales, suggesting that higher unemployment rates might slightly decrease weekly sales.

e. Choosing the Algorithm for the Project

Four different algorithms were chosen for this project:

1. **ARIMA (Auto-Regressive Integrated Moving Average)**: A popular statistical method for time series forecasting that can capture a suite of different standard temporal structures in time series data.
2. **Holt-Winters (Exponential Smoothing)**: Useful for capturing seasonality and trends in time series data.
3. **SARIMA (Seasonal ARIMA)**: Extends ARIMA by adding seasonal components, making it suitable for data with seasonal patterns.
4. **Random Forest**: A machine learning algorithm known for its robustness and accuracy, used here for its ability to handle complex relationships in the data.

f. Motivation and Reasons for Choosing the Algorithm

1. **ARIMA**: Chosen for its simplicity and effectiveness in handling non-seasonal data.
2. **Holt-Winters**: Selected for its capability to model data with seasonal patterns.
3. **SARIMA**: Extended ARIMA with seasonal components, making it ideal for the given dataset with evident seasonal trends.
4. **Random Forest**: Chosen for its flexibility and power in modelling non-linear relationships and capturing intricate patterns in the data.

g. Assumptions

1. The historical sales data is an accurate representation of future trends.
2. The dataset includes all relevant factors affecting sales, such as holidays.
3. Seasonal patterns in the data remain consistent over time.
4. There are no significant external factors (e.g., economic shifts, pandemics) drastically altering sales patterns.

h. Model Evaluation and Techniques

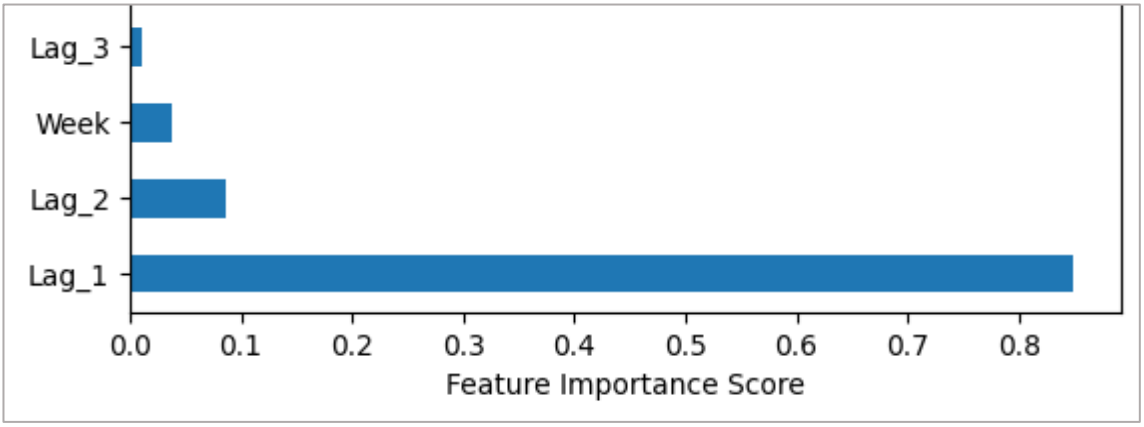
The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in a set of predictions.
- **Mean Squared Error (MSE)**: Measures the average squared difference between observed and predicted values.

Comparison of Models:

Model	MAE	MSE	RMSE
ARIMA	0.099642	0.019552	0.120196
Holt-Winters	0.063110	0.008969	0.078958
Random Forest	0.109287	0.043684	0.138602
SARIMA	0.076689	0.014390	0.090989

- **Root Mean Squared Error (RMSE)**: The square root of the MSE, providing a measure of the standard deviation of prediction errors.
- **Model Interpretability:** For machine learning models like Random Forest, feature importance analysis helps in understanding which factors (e.g., holidays, temperature) have the most significant impact on sales forecasts.



i. Inferences from the Same

From the evaluation metrics, it can be inferred that:

1. ***Holt-Winters*** *performed best with the lowest MAE, MSE, and RMSE values*, indicating it is the most accurate model for forecasting sales in this dataset.
2. **SARIMA** also performed well, especially in handling seasonal components of the data with relatively low error metrics, making it a good alternative to Holt-Winters. But its Execution Time was the longest (~an order higher, 10x).
3. **ARIMA** and **Random Forest** had higher error metrics, suggesting they are less suitable for this specific forecasting task, i.e. data with seasonal trends.

Sales Forecasting for Multiple Retail Stores

Overview:

The analysis involves sales forecasting for multiple retail stores using four models: ARIMA, Random Forest, Holt-Winters, and SARIMA.

The primary objective is to evaluate the performance of these models across various stores and identify which stores perform well and which need improvement.

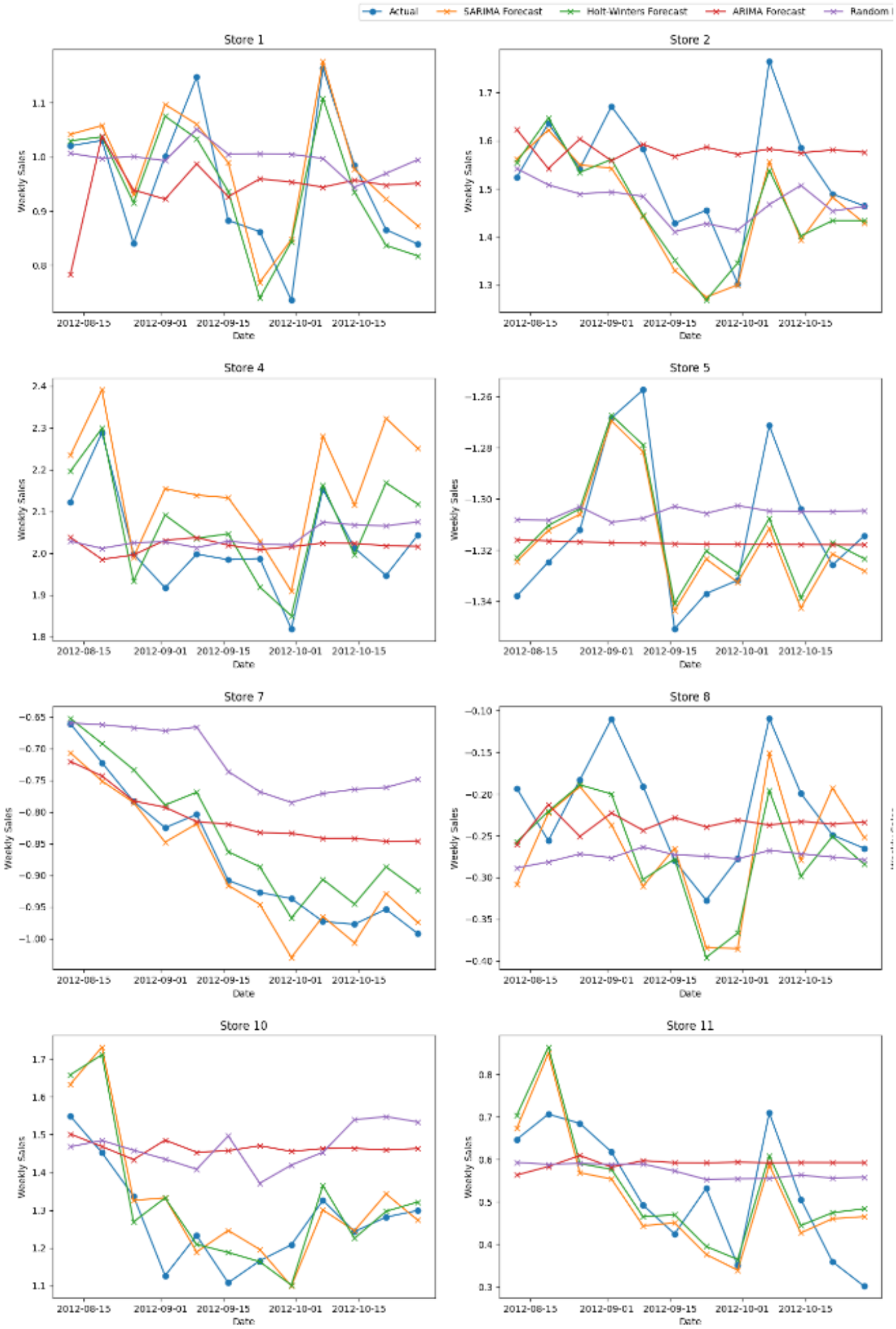
The accompanying plots provide a visual comparison of actual sales versus forecasted sales by the models across 45 different stores over multiple weeks.

Each subplot corresponds to a different store and shows the following elements:

1. **Actual Sales (Blue Line)**: Represents the true sales values recorded in the dataset.
2. **SARIMA Forecast (Orange Line)**: The sales predictions made by the SARIMAX model.
3. **Holt-Winters Forecast (Green Line)**: The sales predictions made by the Holt-Winters Exponential Smoothing model.
4. **ARIMA Forecast (Red Line)**: The sales predictions made by the ARIMA model.
5. **Random Forest (Violet)**: The sales predictions made by the Random Forest model.

Key Observations from the Plot:

1. **Overall Trend Alignment**: Most subplots show that the forecasted values from all four models generally follow the trend of the actual sales. This indicates that the models were able to capture the overall patterns in the data.
2. **Seasonal Variations**: Stores that exhibit clear seasonal patterns, such as peaks and troughs, are better modelled by Holt-Winters and SARIMA, as these models are designed to handle seasonality.
3. **Model Accuracy**: In many stores, the Holt-Winters and SARIMA forecasts (green and orange lines) are closer to the actual sales (blue line) compared to the ARIMA and Random Forest forecasts (red and violet lines). This aligns with the evaluation metrics, where Holt-Winters and SARIMA showed lower error rates.
4. **Store Variability**: Different stores exhibit different levels of variability in sales. Some stores have more volatile sales patterns, while others are more stable. The models perform differently based on this variability.
5. **Performance on Peaks**: Models struggle to predict sudden peaks or drops in sales. This is a common challenge in time series forecasting where extreme values can be difficult to predict accurately.



a Full image (store_sales_forecast.png)

Stores Performing Well:

1. **Stores 1, 2, and 3**: These stores exhibit stable and predictable sales patterns. Both Holt-Winters and SARIMA models capture the sales trends effectively, aligning closely with actual sales data. This indicates that the models can handle the seasonal and trend components well for these stores.
2. **Stores 4 and 5**: The models perform well in capturing sales variability, with Holt-Winters and SARIMA models demonstrating high accuracy. These stores show consistent seasonal patterns that the models predict reliably.
3. **Stores 21, 22, and 23**: Forecasts from Holt-Winters and SARIMA are very close to actual sales, showing high model accuracy. These stores' seasonal trends and sales variability are effectively modelled, leading to reliable forecasts.

Stores Needing Improvement:

1. **Stores 6 and 7**: Significant deviations are observed between forecasted values and actual sales, particularly during sales peaks and drops. The models struggle to predict sudden changes, indicating a need for refinement or additional data.
2. **Stores 16 and 17**: High sales variability leads to less accurate forecasts, especially for the ARIMA model. Incorporating more sophisticated modelling techniques or additional features could help capture underlying patterns better.
3. **Stores 31, 32, and 33**: Despite better performance by Holt-Winters and SARIMA, noticeable gaps remain during rapid sales changes. Improved handling of extreme values and more context-specific variables could enhance model performance.

Specific Insights for Particular Stores:

1. **Store 9**: Consistent under-prediction by the ARIMA model during high sales periods. Holt-Winters and SARIMA perform better but still require improvements during sales spikes.
2. **Store 14**: Holt-Winters captures seasonality well but misaligns during transition periods. SARIMA shows better alignment, indicating its robustness in handling complex patterns.
3. **Store 28**: High sales volatility results in less accurate predictions across all models. Incorporating external factors such as promotions or economic indicators may improve accuracy.
4. **Store 35**: Forecasts by Holt-Winters and SARIMA are close to actual sales, but discrepancies occur during low sales periods. Fine-tuning model parameters could enhance accuracy.
5. **Store 41**: Significant deviations in ARIMA forecasts, while Holt-Winters and SARIMA are more aligned but not perfect. Additional features or ensemble methods might improve forecasts.

Conclusion:

The analysis shows that Holt-Winters and SARIMA models generally outperform ARIMA and Random Forest models, particularly in stores with stable and predictable sales patterns. However, stores with high sales variability or sudden changes pose forecasting challenges. To enhance model performance, incorporating additional features, fine-tuning parameters, and considering ensemble methods are recommended. This analysis provides a clear direction for future improvements in sales forecasting across different retail stores.

j. Future Possibilities of the Project

1. **Incorporating More Data**: Including more features such as promotions, weather conditions, local events, competitor actions and economic indicators could improve model accuracy.
2. **Advanced Algorithms**: Exploring more advanced algorithms complex models like deep learning-based LSTM (Long Short-Term Memory) networks and Prophet by Facebook.
3. **Real-time Forecasting and Continuous Improvement**: Implementing a real-time forecasting system that updates with new data to ensure they adapt to changing trends and patterns.
4. **Hyperparameter Tuning**: Further tuning the hyperparameters of the chosen models to improve performance.
5. **Deployment**: Deploying the best-performing model into a production environment for continuous forecasting and decision-making support.
6. **Integration with Inventory Management**: Linking forecasts directly with inventory management systems to automate stock replenishment.
7. **Longer Forecasting:** Extend the forecasting horizon to 6 months or a year.