Walmart Sales Forecasting
Analysis and Prediction of Weekly Sales

Miguel Ángel Bravo Martínez del Valle
Fernanda Carrillo Escárcega
Riley Rutan

December 10, 2024

Contents

1 Introduction **2**

2 Executive Summary **2**

3 Methodology **3**

4 Exploratory Data Analysis **3**
4.1 Seasonal Differencing 4
4.2 Additional Differencing 6

5 Model Development **8**
5.1 Models Proposed After Seasonal Differencing 8
5.1.1 SARIMA(0,0,7)$(0,1,0)_{52}$ 8
5.1.2 SARIMA(1,0,0)$(0,1,0)_{52}$ 10
5.2 Models Proposed After Additional Differencing 11
5.2.1 SARIMA(1,1,1)$(0,1,0)_{52}$ 11
5.2.2 SARIMA(0,1,1)$(0,1,0)_{52}$ 12
5.2.3 SARIMA(0,1,1)$(0,1,0)_{52}$ 13

6 Forecasting and Discussion **15**

7 Conclusions and Future Work **16**

1 Introduction

This project aims to analyze and forecast weekly sales data for a single Walmart store using time series analysis techniques. The primary objective is to develop a robust forecasting model that accurately captures seasonal patterns and trends, enabling data-driven decision-making for optimizing inventory management, workforce allocation, and marketing strategies.

To achieve this, historical sales data from Kaggle's "Walmart Recruiting - Store Sales Forecasting" dataset has been utilized. This dataset contains detailed weekly sales data for multiple Walmart stores across various departments over a two-year period. The analysis focuses on transforming raw sales data into actionable insights to improve operational efficiency and predict future sales with precision.

2 Executive Summary

Walmart, the largest global retailer, operates over 10,500 stores worldwide and serves as a leader in data-driven decision-making. This project analyzes weekly sales data from a single Walmart store, spanning February 2010 to November 2012. The dataset, sourced from Kaggle's "Walmart Recruiting - Store Sales Forecasting" competition, comprises detailed sales records across 45 stores and approximately 100 departments, totaling over 421,000 rows of historical data.

The goal of this project is to develop an advanced time series forecasting model to predict future sales trends with high accuracy. By identifying seasonal patterns and demand fluctuations, the analysis aims to provide actionable insights for optimizing three key areas of store operations:

- **Inventory Management**: Preventing stockouts while minimizing overstocking costs to ensure optimal inventory levels.
- **Staffing Efficiency**: Adjusting workforce schedules to align with predicted sales peaks and troughs, reducing operational inefficiencies.
- **Targeted Marketing Campaigns**: Enhancing promotional effectiveness by focusing efforts on high-demand periods, such as holiday seasons.

The analysis reveals clear seasonal trends, including sharp sales increases during holiday periods, and identifies recurring weekly patterns that drive revenue. Advanced time series methods, including ARIMA and SARIMA models, were employed to transform raw sales data into accurate forecasts. These forecasts not only support operational decision-making for the analyzed store but also offer scalable insights applicable across Walmart's broader network of stores.

Through this project, Walmart can leverage predictive analytics to reduce costs, improve customer satisfaction, and maintain its competitive edge in the retail industry.

3 Methodology

The following steps were undertaken:

1. Data preprocessing to handle missing values and aggregate weekly sales.
2. Exploratory data analysis to uncover trends and seasonality.
3. Model selection and validation using ARIMA techniques.
4. Forecast evaluation with metrics like RMSE and residual diagnostics.

4 Exploratory Data Analysis

The dataset includes historical weekly sales data from 45 Walmart stores, each comprising approximately 100 departments. For this analysis, Store 13 was selected as it contained no missing values, ensuring the reliability of the data. The dataset for Store 13 consisted of 143 weekly observations.

To simplify the analysis, the weekly sales from all departments in Store 13 were aggregated, creating a new dataset with a single variable representing the total weekly sales for the store. This approach allows for a more streamlined exploration of trends and seasonality at the store level.

The time series plot of weekly sales for Store 13 is shown in Figure 1. It highlights sharp seasonal peaks, particularly during holiday periods, as well as variability in weekly sales.

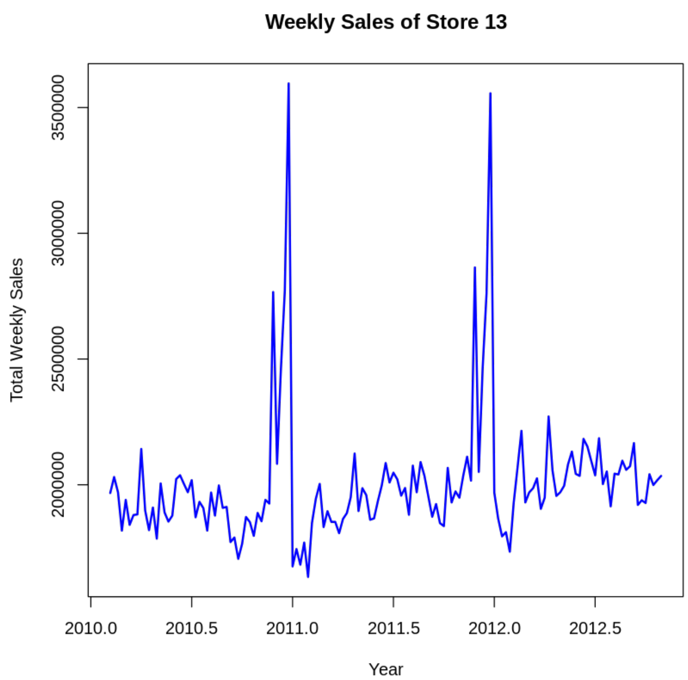


Figure 1: Time Series of Total Weekly Sales for Store 13

Augmented Dickey-Fuller Test

To assess the stationarity of the weekly sales time series for Store 13, we applied the Augmented Dickey-Fuller (ADF) Test. This test evaluates the presence of a unit root, which indicates whether a time series is non-stationary.

The ADF test produced a test statistic of -5.1234 with a p-value less than 0.01, providing strong evidence to reject the null hypothesis of non-stationarity. This result suggests that the series is stationary at the 1% significance level.

Despite the stationarity detected by the test, the observed seasonality in the data may still require additional transformations, such as seasonal differencing, to fully capture the underlying patterns. This observation informs the subsequent steps in model development.

ACF and PACF Analysis

To further investigate the structure of the time series, we analyzed the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF). These tools provide insight into the temporal dependencies within the data and guide the selection of appropriate model parameters for SARIMA.

The ACF and PACF plots are presented in Figure 2. The ACF reveals significant seasonal autocorrelations, confirming the presence of recurring patterns at lags corresponding to multiples of 52 weeks. The PACF plot highlights specific lags with strong partial correlations, indicating the potential presence of autoregressive components.

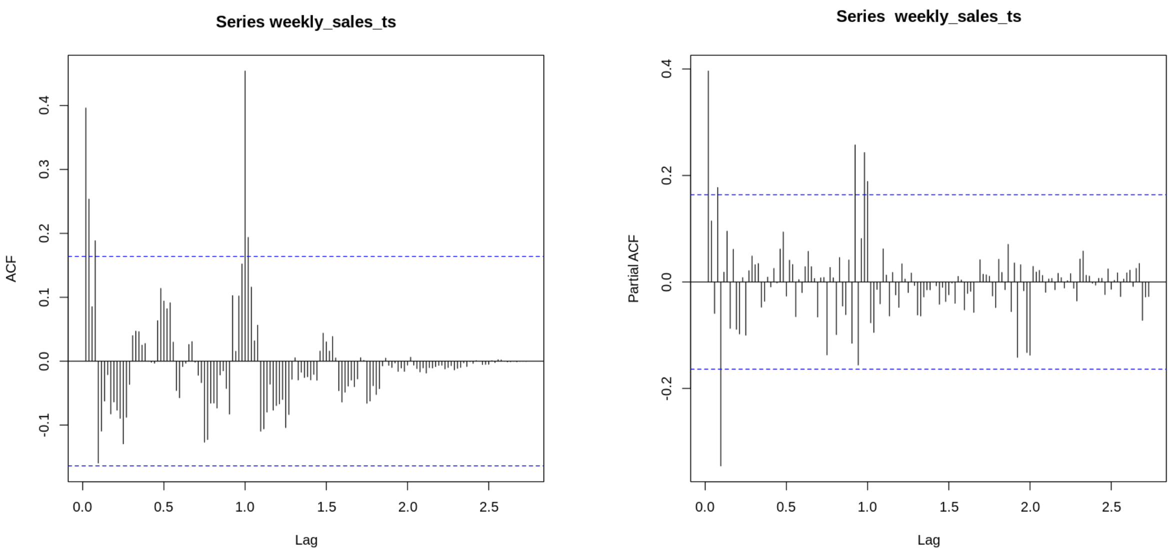


Figure 2: ACF (left) and PACF (right) for Weekly Sales of Store 13

4.1 Seasonal Differencing

Despite the stationarity suggested by the initial Augmented Dickey-Fuller test, the strong seasonal patterns observed in the time series and corroborated by the ACF and PACF plots necessitate a seasonal differencing step. Seasonal differencing removes recurring patterns by subtracting the value from the same season in the previous cycle. After applying seasonal differencing, the resulting time series is shown in Figure 3.

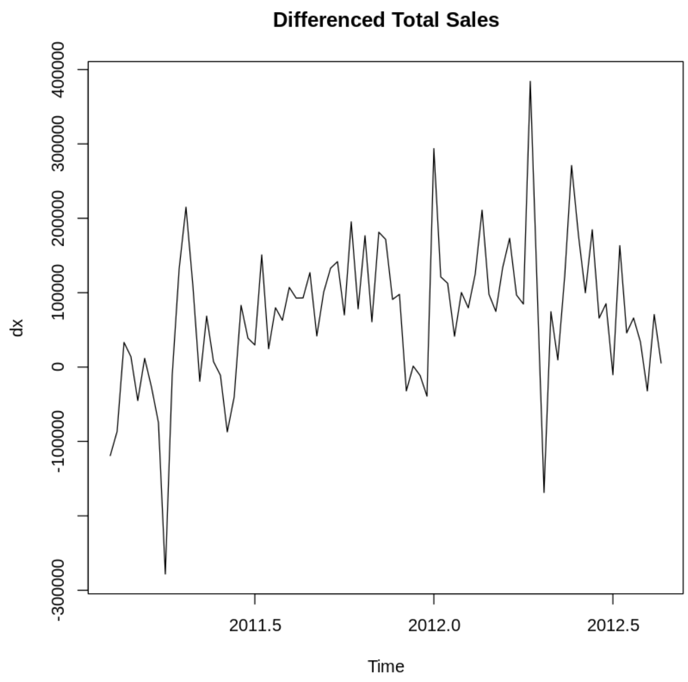


Figure 3: Seasonally Differenced Time Series for Weekly Sales of Store 13

Following seasonal differencing, we applied the Augmented Dickey-Fuller (ADF) test again to confirm stationarity. The test results indicated that the differenced series is stationary, with a p-value of 0.034. While this is below the 5% threshold, it is relatively close, suggesting a borderline case of stationarity.

To further analyze the properties of the seasonally differenced series, we computed the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), shown in Figure 4. The ACF and PACF plots suggest the following:

- The seasonal components (P and Q) of the SARIMA model are both zero ($P = Q = 0$).
- No regular differencing ($d = 0$) is required as only seasonal differencing was applied.

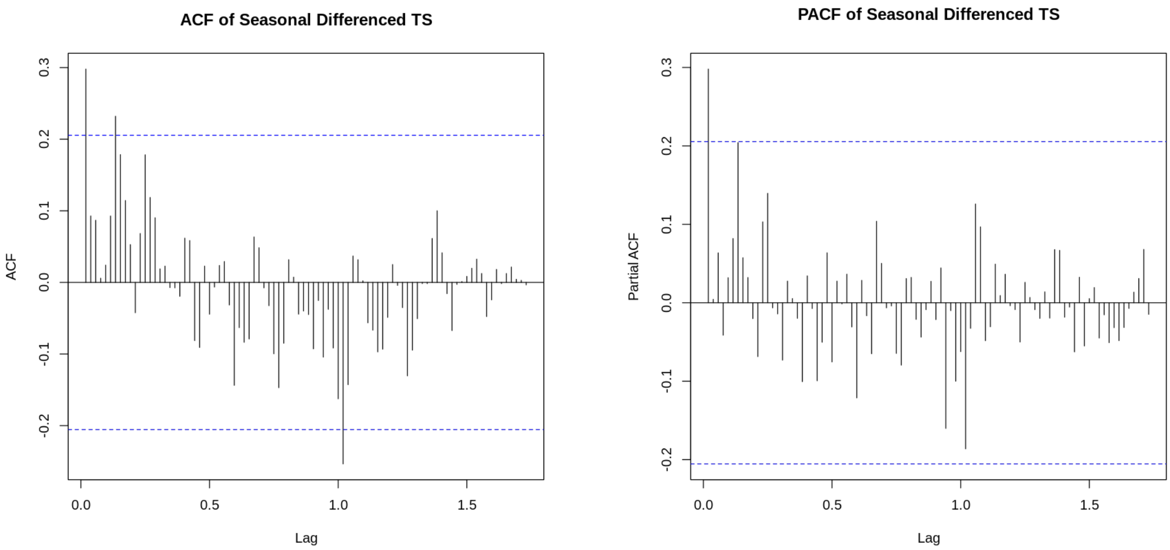


Figure 4: ACF (left) and PACF (right) for Seasonally Differenced Series

Based on these findings, the following SARIMA configurations are proposed for further evaluation:

- **SARIMA(1,0,0)(0,1,0)**$_{52}$: Includes a seasonal differencing component ($D = 1$) and an autoregressive ($AR(1)$) term for the non-seasonal part.
- **SARIMA(0,0,7)(0,1,0)**$_{52}$: Incorporates a moving average ($MA(7)$) term for the non-seasonal component.

4.2 Additional Differencing

Because the seasonally differenced data was only borderline stationary (p-value = 0.034 in the Augmented Dickey-Fuller test), an additional differencing step was applied to ensure stationarity. This step involved taking a regular (non-seasonal) difference of the time series, which removes any remaining linear trends. The resulting differenced series is shown in Figure 5.

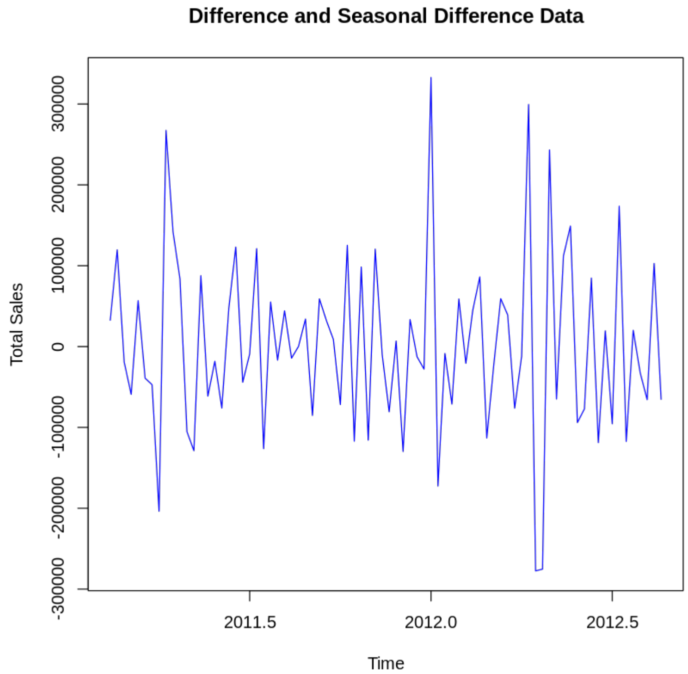


Figure 5: Differenced Time Series after Additional Differencing

Following this transformation, the Augmented Dickey-Fuller (ADF) test was conducted again. The test results confirm that the series is now stationary, with a test statistic of -6.4552 and a p-value well below 0.01, providing strong evidence to reject the null hypothesis of non-stationarity.

To further analyze the properties of this doubly differenced series, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were computed, as shown in Figure 6. The ACF and PACF plots provide insights into the remaining autocorrelation structure and inform the selection of appropriate ARIMA model parameters.

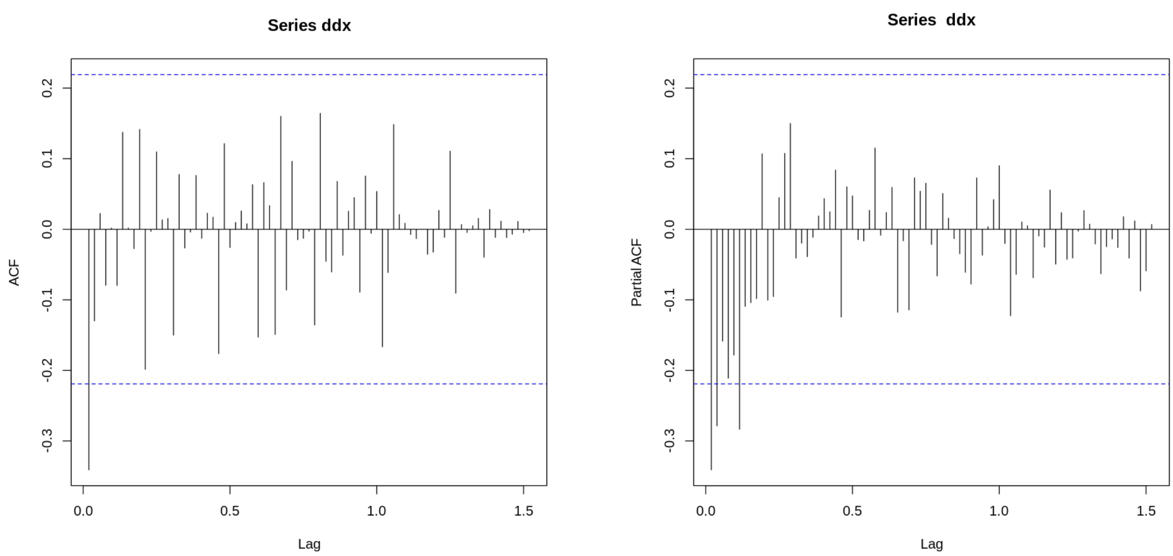


Figure 6: ACF (left) and PACF (right) for Doubly Differenced Series

In addition to the ACF and PACF, the Extended Autocorrelation Function (EACF) was computed to refine the selection of model parameters. The EACF matrix (Table 1) suggests two possible ARIMA configurations for further evaluation.

Table 1: Extended Autocorrelation Function (EACF) Matrix for Doubly Differenced Series

AR/MA	0	1	2	3	4	5	6	7	8
0	x	o	o	o	o	o	o	o	o
1	x	o	o	o	o	o	o	o	o
2	x	x	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o
4	x	x	o	x	o	o	o	o	o
5	x	x	o	x	o	o	o	o	o
6	x	x	o	x	o	o	o	o	o
7	x	o	o	o	o	o	o	o	o

Based on the EACF analysis, two potential SARIMA configurations are suggested for further evaluation:

- **SARIMA(0,1,1)(0,1,0)$_{52}$**: Includes a single moving average ($MA(1)$) term for the non-seasonal component.
- **SARIMA(1,1,1)(0,1,0)$_{52}$**: Incorporates both an autoregressive ($AR(1)$) and moving average ($MA(1)$) term for the non-seasonal component.

Both models share the same seasonal configuration, with no seasonal autoregressive ($P = 0$) or moving average ($Q = 0$) terms, a seasonal differencing order of $D = 1$, and a seasonal period of $s = 52$.

5 Model Development

This section presents the development, evaluation, and selection of SARIMA models for forecasting the weekly sales time series. Based on the differencing steps and subsequent analyses, two sets of candidate models were identified. The final subsection details the selection process for identifying the single best model.

5.1 Models Proposed After Seasonal Differencing

5.1.1 SARIMA(0,0,7)(0,1,0)$_{52}$

The initial model included seven non-seasonal moving average (MA) terms and an intercept. However, after evaluating the statistical significance of the coefficients, only $MA(1)$ and the intercept were retained, as the remaining terms were found to be insignificant. This simplification improved the model's interpretability and reduced the Akaike Information Criterion (AIC) score from 2096.78 to 2087.91, indicating a better balance between model fit and complexity.

Residual diagnostics confirmed that the simplified model captures the structure of the time series effectively:

- **Residual Plot**: The residuals (Figure 7) show no visible patterns, suggesting that the model adequately explains the data.
- **Histogram**: The residuals follow an approximately normal distribution (Figure 8).
- **ACF and PACF**: The autocorrelation and partial autocorrelation plots (Figure 9) show no significant lags, supporting the assumption of white noise.
- **Box-Ljung Test**: The test confirmed no significant autocorrelation in the residuals ($p > 0.05$).

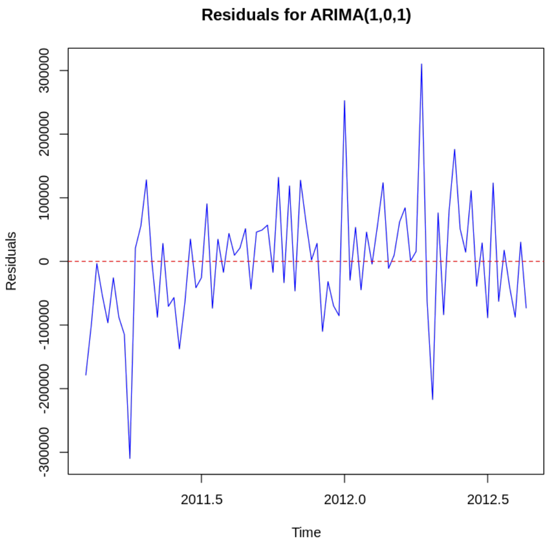


Figure 7: Residual Plot

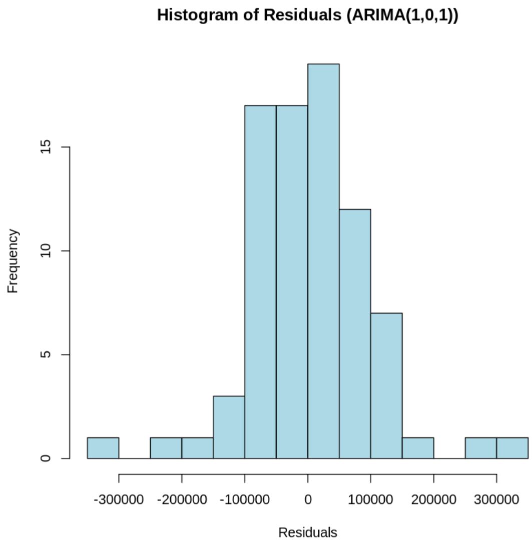
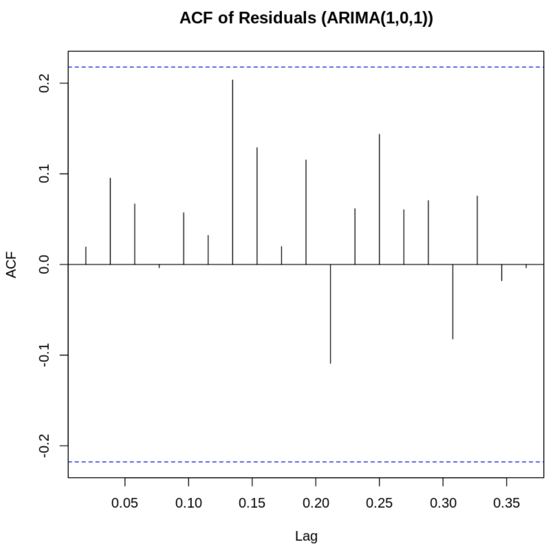


Figure 8: Histogram of Residuals



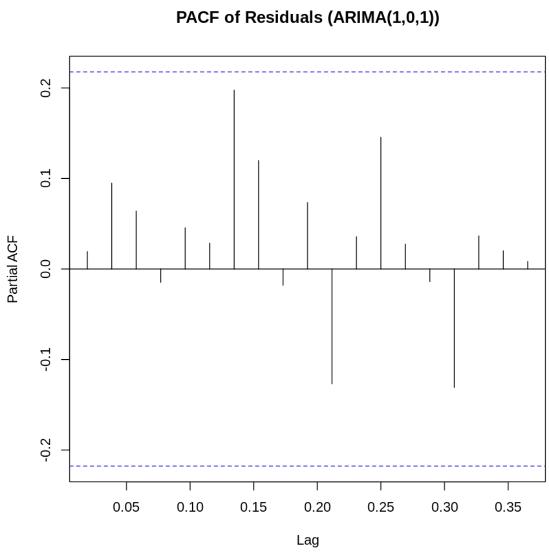


Figure 9: ACF (left) and PACF (right) of Residuals for SARIMA(0,0,7)$(0,1,0)_{52}$

5.1.2 SARIMA(1,0,0)(0,1,0)$_{52}$

The initial model for this configuration included one non-seasonal autoregressive ($AR(1)$) term and an intercept. Both parameters were found to be statistically significant, confirming their importance in modeling the time series. The model achieved an Akaike Information Criterion (AIC) score of 2086.95, demonstrating its efficiency in balancing model fit and complexity.

Residual diagnostics confirmed the adequacy of this model:

- **Residual Plot**: The residuals (Figure 10) show no visible patterns, suggesting the model effectively captures the structure of the data.
- **Histogram**: The residuals follow an approximately normal distribution (Figure 11).
- **ACF and PACF**: The autocorrelation and partial autocorrelation plots (Figure 12) indicate no significant lags, supporting the assumption of white noise.
- **Box-Ljung Test**: The test confirmed no significant autocorrelation in the residuals ($p > 0.05$).

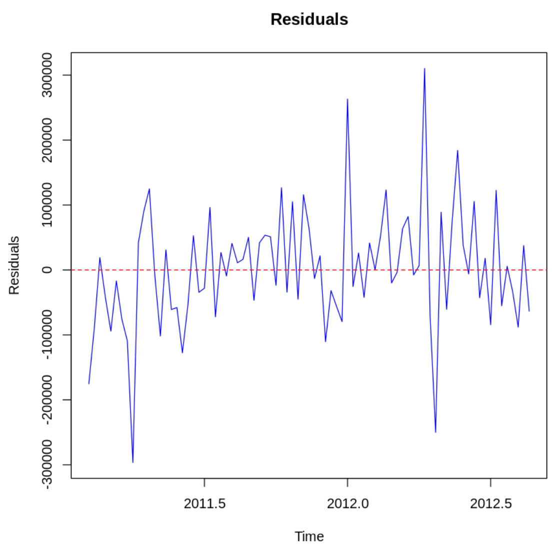


Figure 10: Residual Plot

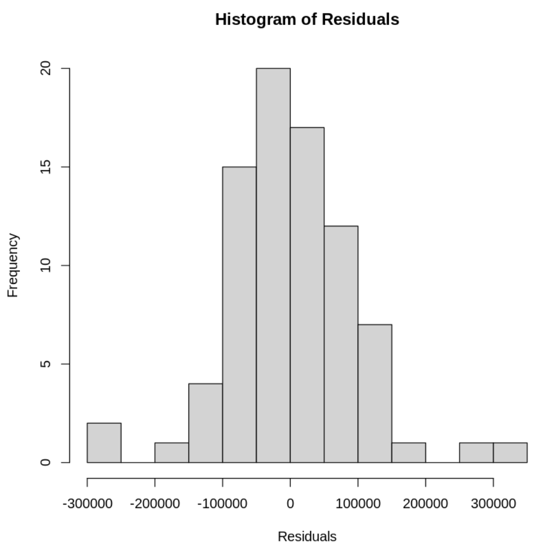


Figure 11: Histogram of Residuals

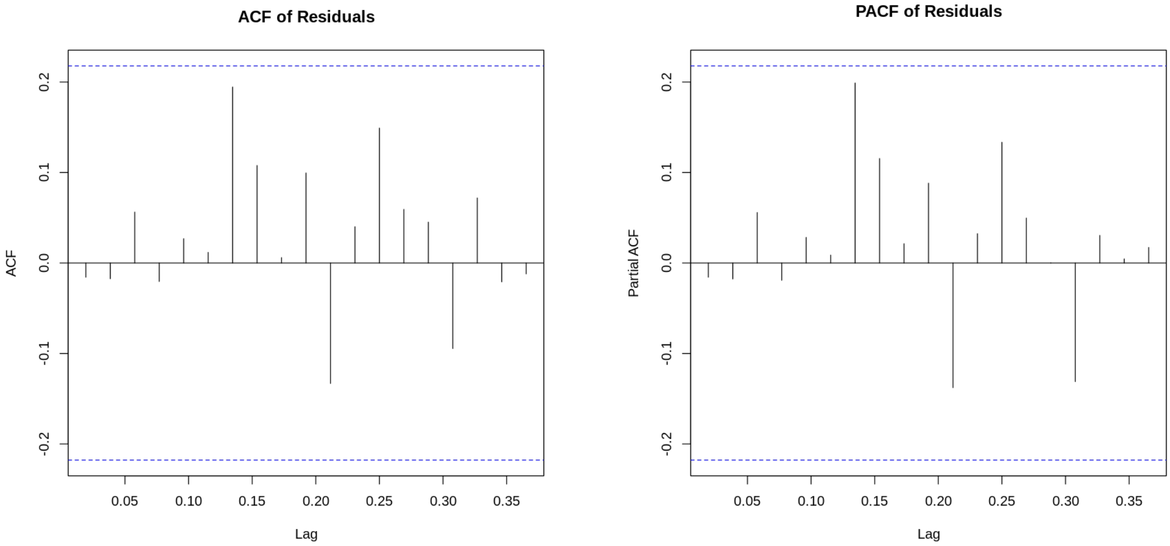


Figure 12: ACF (left) and PACF (right) of Residuals for SARIMA(1,0,0)(0,1,0)$_{52}$

5.2 Models Proposed After Additional Differencing

5.2.1 SARIMA(1,1,1)(0,1,0)$_{52}$

The SARIMA(1,1,1)(0,1,0)$_{52}$ model includes one non-seasonal autoregressive ($AR(1)$) term, one non-seasonal moving average ($MA(1)$) term, and a seasonal differencing component ($D = 1$) with a seasonal period of 52 weeks. After fitting the model, both $AR(1)$ and $MA(1)$ coefficients were found to be statistically significant. The model achieved an Akaike Information Criterion (AIC) score of 2061.71, the lowest among all tested configurations.

Residual diagnostics were conducted to evaluate the model's adequacy:

- **Residual Plot**: The residuals (Figure 13) show no visible patterns, indicating that the model captures the structure of the data effectively.
- **Histogram**: The residuals follow an approximately normal distribution (Figure 14).
- **ACF and PACF**: The autocorrelation and partial autocorrelation plots (Figures 15) reveal no significant lags, supporting the assumption of white noise.
- **Box-Ljung Test**: The test confirmed no significant autocorrelation in the residuals ($p > 0.05$).

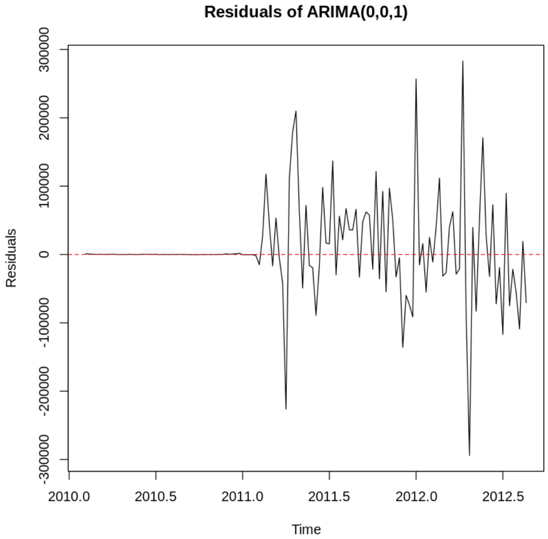


Figure 13: Residual Plot

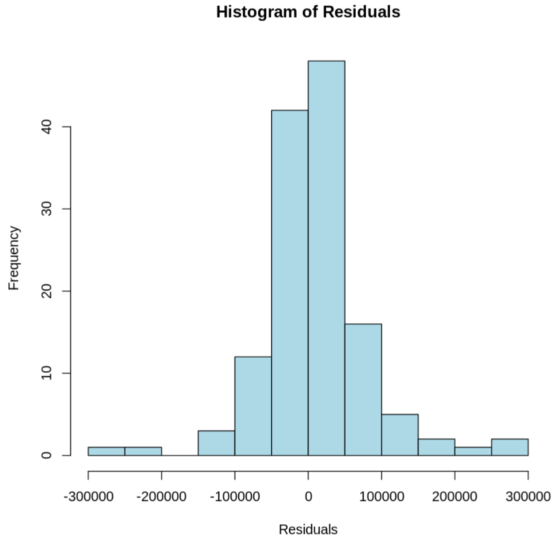
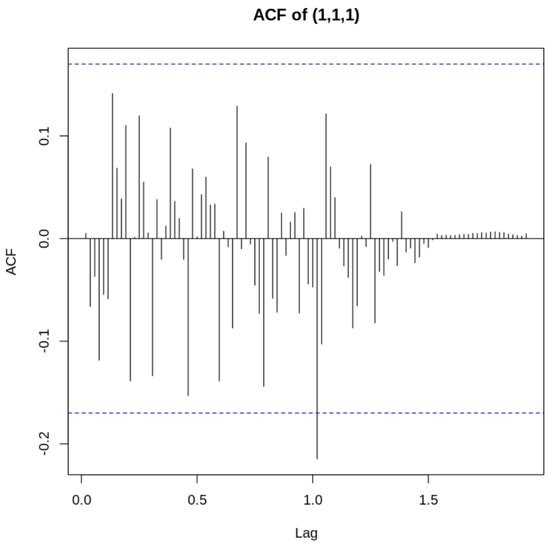


Figure 14: Histogram of Residuals



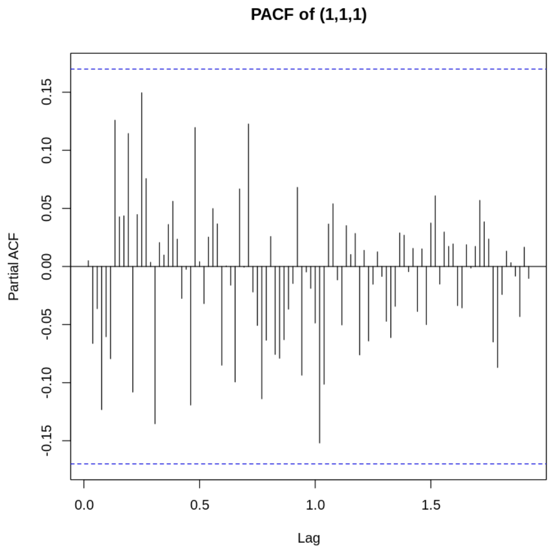


Figure 15: ACF (left) and PACF (right) of Residuals for SARIMA(1,1,1)(0,1,0)$_{52}$

5.2.2 SARIMA(0,1,1)(0,1,0)$_{52}$

The SARIMA(0,1,1)(0,1,0)$_{52}$ model includes one non-seasonal moving average ($MA(1)$) term and a seasonal differencing component ($D = 1$) with a seasonal period of 52 weeks. After fitting the model, $MA(1)$ was found to be statistically significant ($p < 0.001$), while the intercept was not significant ($p > 0.05$). The model achieved an Akaike Information Criterion (AIC) score of 2064.25 and a Bayesian Information Criterion (BIC) score of 2073.40.

Residual diagnostics confirmed that the model performs adequately:

- **Residual Plot**: The residuals (Figure 19) show no visible patterns, suggesting the model effectively captures the structure of the data.

- **Histogram**: The residuals follow an approximately normal distribution (Figure 20).
- **ACF and PACF**: The autocorrelation and partial autocorrelation plots (Figures 21) indicate no significant lags, supporting the assumption of white noise.
- **Box-Ljung Test**: The test confirmed no significant autocorrelation in the residuals ($p > 0.05$).

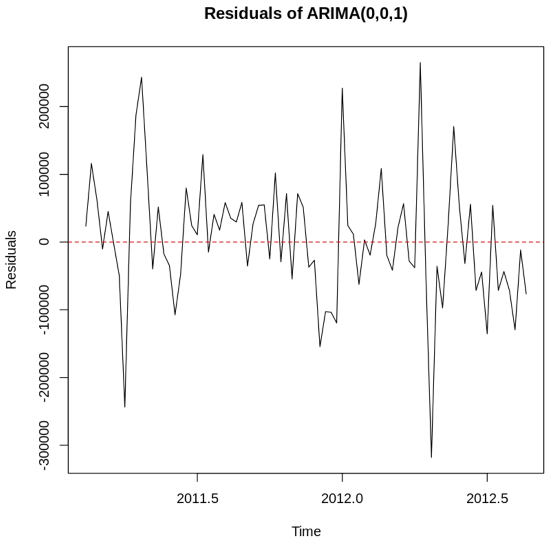


Figure 16: Residual Plot

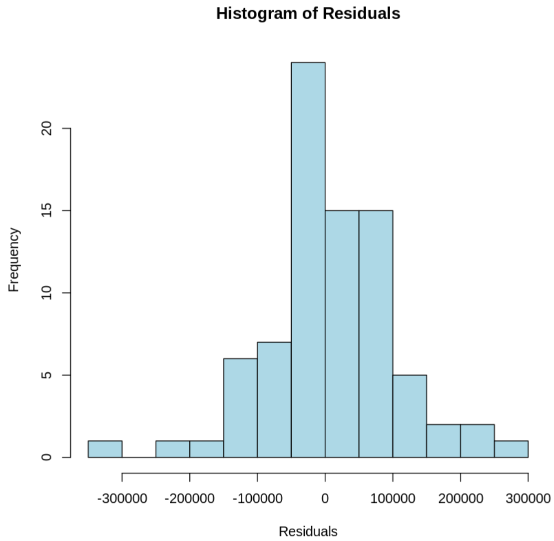


Figure 17: Histogram of Residuals

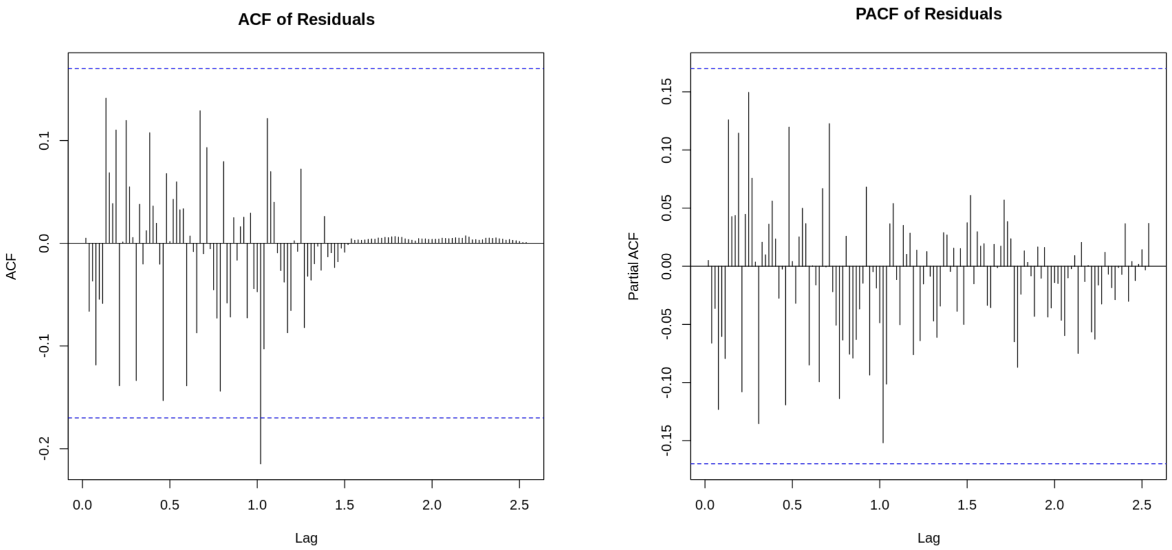


Figure 18: ACF (left) and PACF (right) of Residuals for SARIMA(0,1,1)(0,1,0)$_{52}$

5.2.3 SARIMA(0,1,1)(0,1,0)$_{52}$

The SARIMA(0,1,1)(0,1,0)$_{52}$ model includes one non-seasonal moving average ($MA(1)$) term and a seasonal differencing component ($D = 1$) with a seasonal period of 52 weeks.

After fitting the model, $MA(1)$ was found to be statistically significant ($p < 0.001$), while the intercept was not significant ($p > 0.05$). The model achieved an Akaike Information Criterion (AIC) score of 2064.25 and a Bayesian Information Criterion (BIC) score of 2073.40.

Residual diagnostics confirmed that the model performs adequately:

- **Residual Plot**: The residuals (Figure 19) show no visible patterns, suggesting the model effectively captures the structure of the data.
- **Histogram**: The residuals follow an approximately normal distribution (Figure 20).
- **ACF and PACF**: The autocorrelation and partial autocorrelation plots (Figures 21) indicate no significant lags, supporting the assumption of white noise.
- **Box-Ljung Test**: The test confirmed no significant autocorrelation in the residuals ($p > 0.05$).

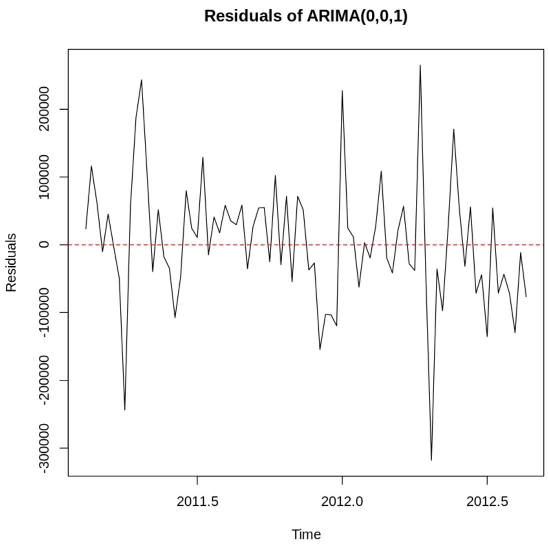


Figure 19: Residual Plot

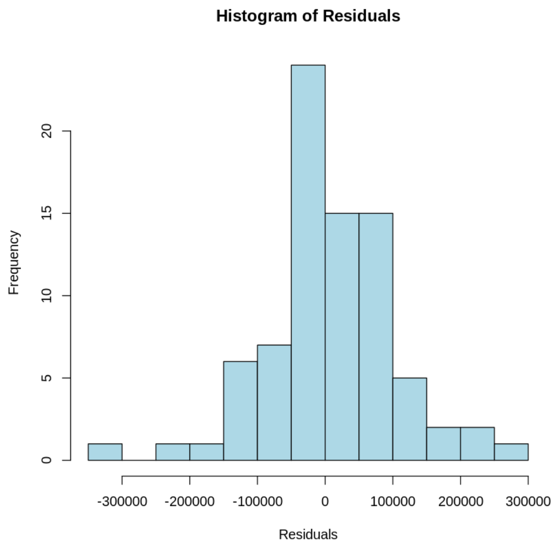


Figure 20: Histogram of Residuals

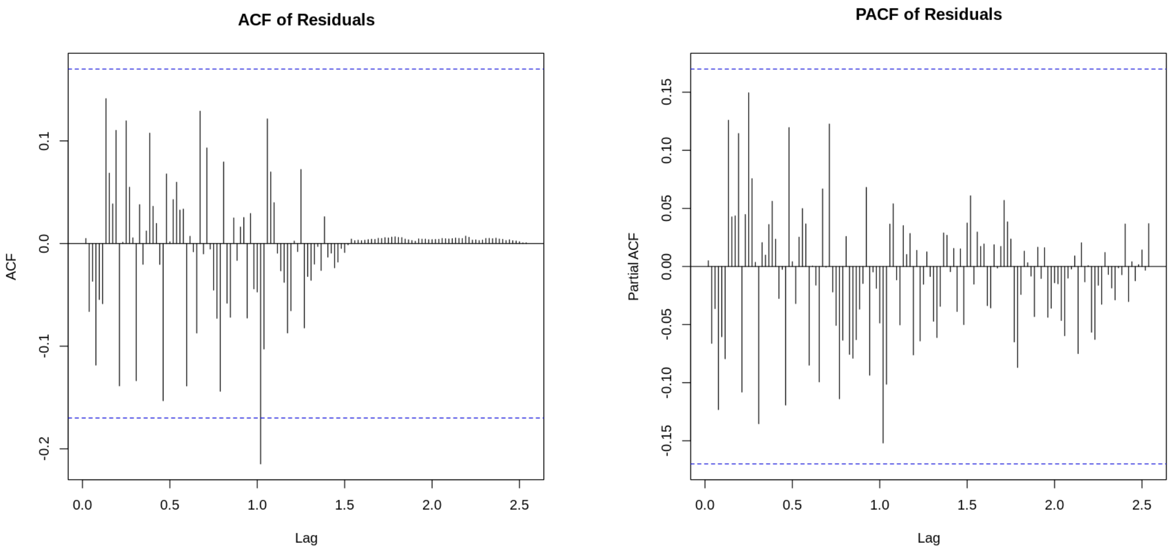


Figure 21: ACF (left) and PACF (right) of Residuals for SARIMA(0,1,1)(0,1,0)$_{52}$

Selection of the Best Model

The SARIMA(1,1,1)(0,1,0)$_{52}$ model was selected as the best-performing model for forecasting weekly sales due to the following reasons:

- **Lowest AIC Score**: Achieved the lowest AIC score (2061.71), indicating the best balance between model fit and complexity.
- **Residual Analysis**: Residuals showed no patterns, followed a normal distribution, and had no significant autocorrelation ($p > 0.05$).
- **Robustness**: Captures key time series dynamics with both $AR(1)$ and $MA(1)$ components.

This model will be used for the final forecasting of weekly sales.

6 Forecasting and Discussion

The selected SARIMA(1,1,1)(0,1,0)$_{52}$ model was used to generate forecasts for weekly sales. The results indicate that the model effectively captures the overall trend and seasonal patterns in the data. Key observations from the analysis are as follows:

- **Model Performance**: The fitted values align closely with the actual data, as shown in Figure 22. This highlights the model's ability to accurately replicate historical trends and seasonality.
- **Forecast Accuracy**: The forecasted values (Figure 23) are reliable, with only one actual observation falling outside the 95% prediction interval, indicating strong predictive performance.
- **Root Mean Square Error (RMSE)**: The RMSE for the test data confirms the model's robustness in minimizing forecast errors.

- **Sensitivity to Recent Trends**: The model is sensitive to abrupt changes in recent trends, which could slightly affect short-term predictions.
- **Potential for Improvement**: While the model performs well, incorporating additional periods of data could further enhance its accuracy and robustness.

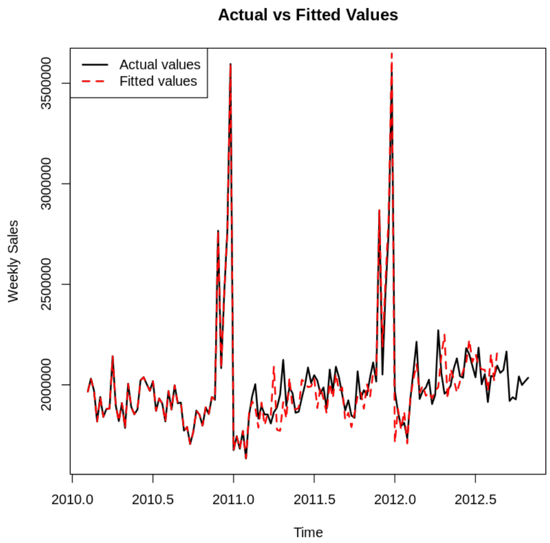


Figure 22: Actual vs Fitted Values for Weekly Sales

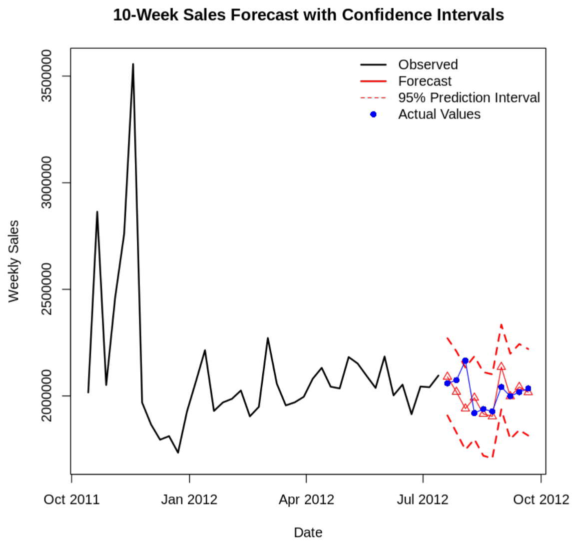


Figure 23: Forecasted Weekly Sales with 95% Prediction Interval

7 Conclusions and Future Work

The ARIMA(1,1,1)x(0,1,0)$_{52}$ model demonstrated effective sales forecasting capabilities. Future improvements could include:

- Incorporating more data to enhance model robustness.
- Extending the analysis to other Walmart stores.