PROJECT REPORT

Weekly Sales Prediction for Walmart Stores

Department of Mathematics, IIT Kharagpur

Group Members:

Member 1 : Aryam Shankar 22EC3FP37
Member 2 : Vipul Vaibhaw 22CH3FP30
Member 3 : Hrishikesh Tiwari 22CH3FP39
Member 4 : Arunava Bhattacharya 22CH3FP48
Member 5 : Sreelekshmi Kishore 22BT3FP36
Member 6 : Pranjal Paliwal 22BT3FP51

Acknowledgments

We would like to express our sincere appreciation to our instructor, Dr. Buddhananda Banerjee for his invaluable guidance, feedback, and support throughout this project. His expertise and encouragement greatly contributed to the successful completion of this analysis.

1. Introduction

The objective of this project is to develop and evaluate time series regression models for forecasting the overall weekly sales across 45 Walmart stores. Accurate forecasting of weekly sales is vital for inventory planning, staffing, and promotional decision-making. This report outlines the data preparation, exploratory analysis, feature engineering, stationarity testing, decomposition, autocorrelation diagnostics, and the implementation of ARIMA and SARIMA models.

2. Dataset Description

Data Sources:

- **train.csv:** Historical weekly sales per department per store (2010-02-05 to 2012-11-01).
- **features.csv:** Store-level time series features such as Temperature, Fuel Price, CPI, Unemployment, and promotional markdowns (MarkDown1–5), along with IsHoliday flags.
- **stores.csv:** Store metadata including store size and type.

3. Data Preprocessing and Merging

Merging Strategy:

1. Read each CSV into pandas DataFrames.
2. Merge train with features on Store and Date (left join).
3. Merge the result with stores on Store (left join).
4. Convert Date to datetime and sort chronologically.

This consolidation yields a unified dataset containing sales, environmental, economic, and store-specific attributes for further analysis.

- **Missing Markdown Data:** Columns MarkDown1 through MarkDown5 contain too many missing values (NA) to have any valuable contribution to the model and hence are dropped to avoid distorting models.
- **Time Delta Check:** By computing differences between consecutive Date entries, we confirm weekly frequency and detect any irregularities.

4. Feature Engineering

To capture temporal patterns, we created the following features:

- **Calendar Components:** year, month, day, day_of_year, week_of_year, quarter, and season.
- **Cyclical Encoding:** month and week_of_year are transformed into sine and cosine pairs (month_sin, month_cos, week_sin, week_cos) to respect their circular nature.

This enriches the dataset with interpretable information on seasonal and periodic effects.

5. Exploratory Data Analysis (EDA)

To understand the underlying patterns and relationships in the data before modeling, we conducted several analyses and visualizations below.

5.1 Distribution of Weekly Sales

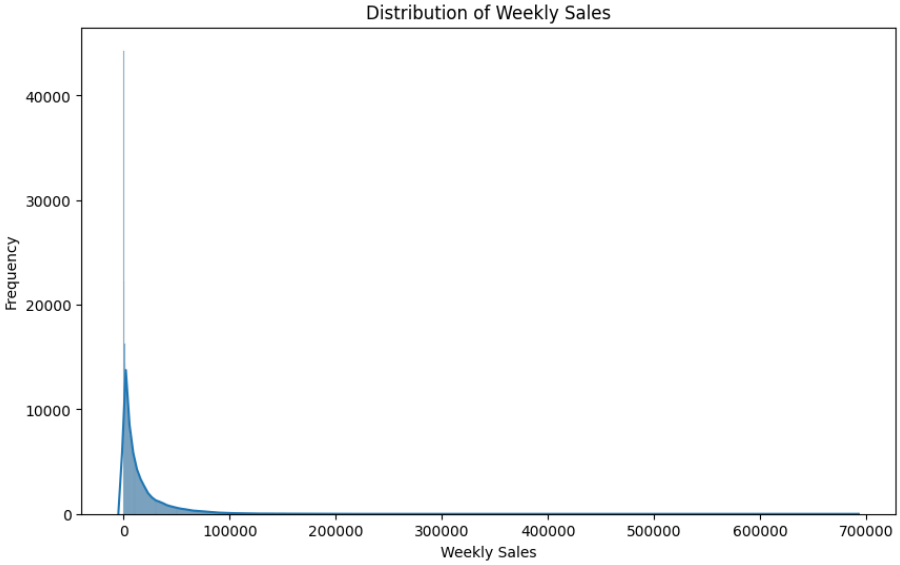

Figure 5.1 : Distribution of Weekly sales

- A histogram with a kernel density estimate shows that weekly sales are right-skewed, with most weeks having sales between 0 and 200,000 but a long tail extending beyond 500,000.
- The median sales value lies around 100,000, suggesting typical weekly demand.
- A few extreme outlier weeks exceed 700,000 in sales—likely corresponding to major holiday periods (e.g., Christmas week).

Inference: The skew and heavy tail indicate that extreme sales events occur infrequently but must be accounted for, possibly via log-transformations or robust loss functions in regression models.

5.2 Weekly Sales by Store Type

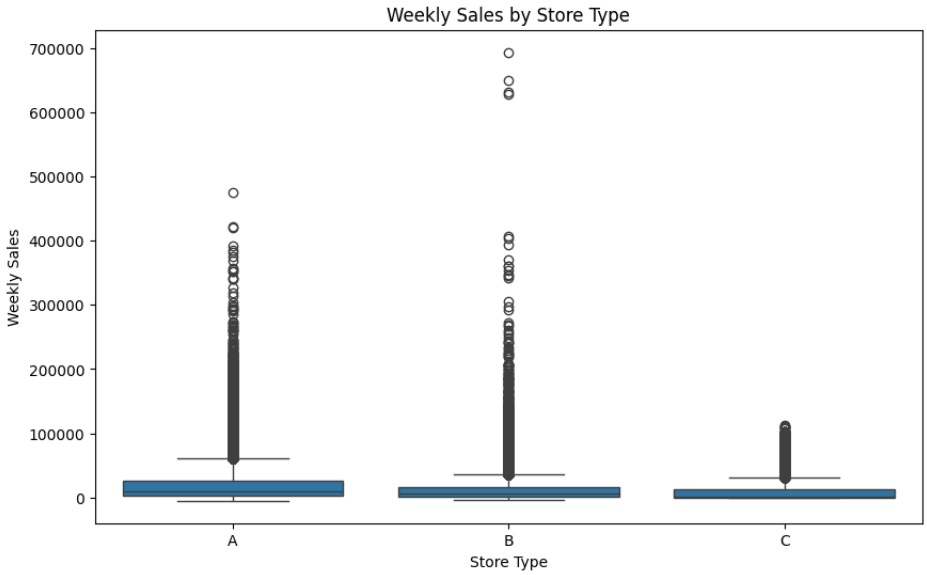

Figure 5.2 : Weekly Sales by Store type

- Box plots of Weekly_Sales grouped by store Type (A, B, C) reveal that Type A stores have the highest median weekly sales (~120,000) and the widest interquartile range, indicating greater variability. Type B stores sit in the middle, while Type C stores show the lowest medians (~80,000) and tighter IQRs.
- All types exhibit outliers at the high end, but Type A outliers reach above 800,000, reinforcing their role as high-volume locations.

Inference: Store type significantly influences sales volume and variability. Including Type as a categorical feature in forecasting models is likely to improve accuracy by capturing store-specific demand profiles.

5.3 Time Series of Total Sales

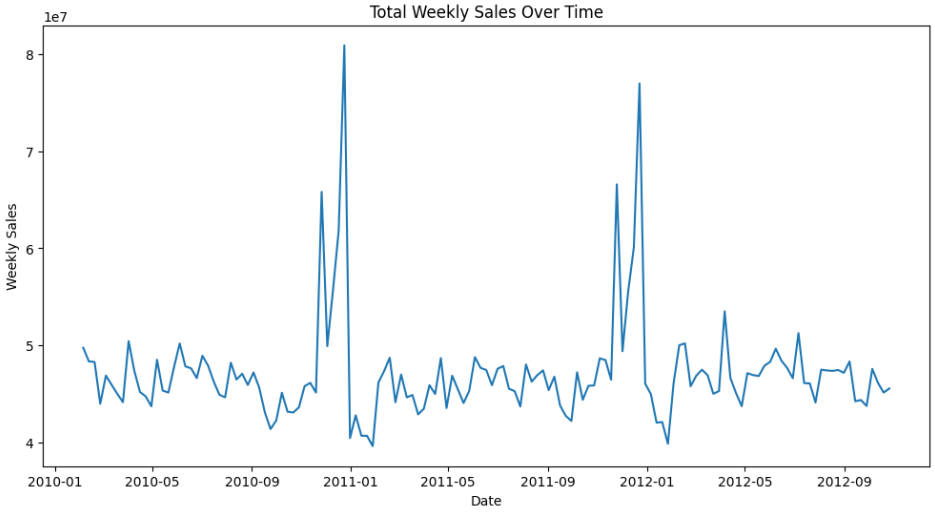

Figure 5.3 : Total Weekly Sales Over Time

- Aggregating sales across all stores and plotting over time uncovers a clear upward trend, with total weekly sales growing at an approximate rate of 5–10% year-over-year.
- Pronounced seasonal peaks occur in late November and December, aligning with Black Friday and Christmas shopping—and smaller peaks around February (Super Bowl) and early September (Labor Day).
- There is also a mild mid-summer plateau, possibly reflecting back-to-school promotions.

Inference: The series exhibits strong trend and multiple seasonal cycles tied to holidays. Forecasting models must incorporate both elements, for example via seasonal differencing or explicit holiday regressors.

5.4 Correlation Analysis

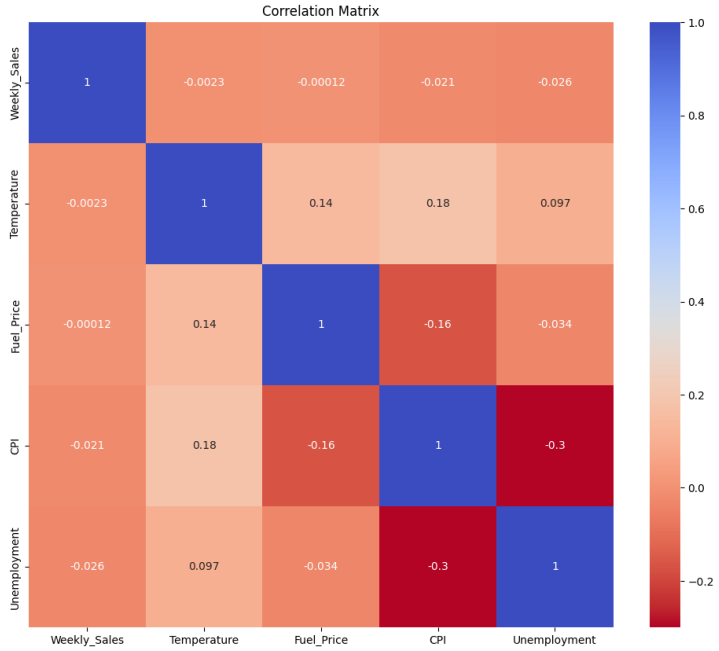

Figure 5.4 : Correlation Matrix

A heatmap of pairwise correlations among Weekly_Sales, Temperature, Fuel_Price, CPI, and Unemployment reveals:

- Weak negative associations between weekly sales and both CPI (r = -0.02) and Unemployment (r = -0.03), suggesting modest sensitivity of sales to macroeconomic factors.
- Negligible correlation with Temperature (r = -0.002) and Fuel_Price (r = -0.0001), indicating these variables alone may not drive overall sales fluctuations.
- A moderate negative correlation between CPI and Unemployment (r = -0.30), consistent with economic theory (higher CPI often co-occurs with lower unemployment).

Inference: While environmental and macroeconomic features show limited direct linear relationships with aggregate sales, they may still contribute nonlinearly or as exogenous regressors in models like SARIMAX. Holiday indicators or interaction terms could further reveal their conditional effects.

6. Stationarity Analysis

6.1 Rolling Statistics

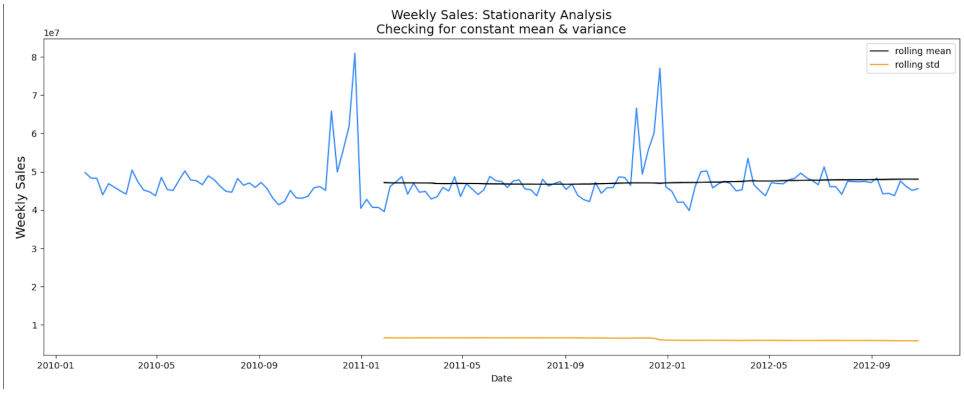

Figure 6.1(a) : Stationarity Analysis of Weekly Sales

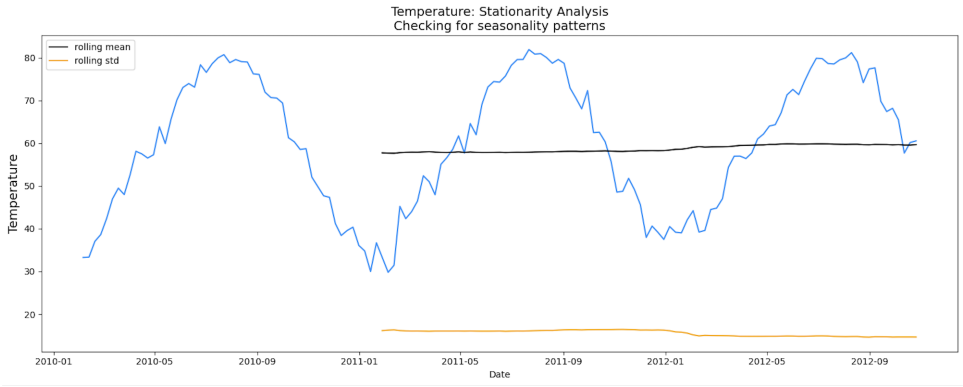

Figure 6.1(b) : Stationarity Analysis of Temperature

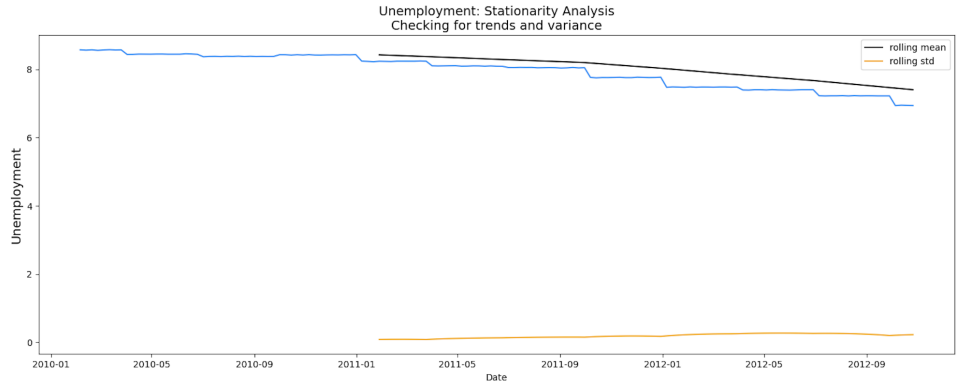

Figure 6.1(c) : Stationarity Analysis of Unemployment

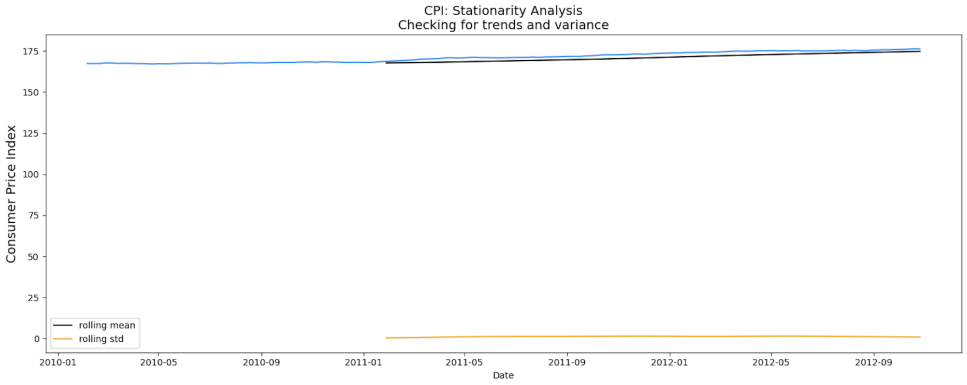

Figure 6.1(d) : Stationarity Analysis of CPI

Plots of the rolling mean and standard deviation (with a 52-week window) for weekly sales and temperature show only minor fluctuations over time, indicating weak stationarity. Since unemployment, CPI, and temperature are not parameters of primary interest, our focus remains solely on weekly sales.

6.2 Augmented Dickey–Fuller (ADF) Test

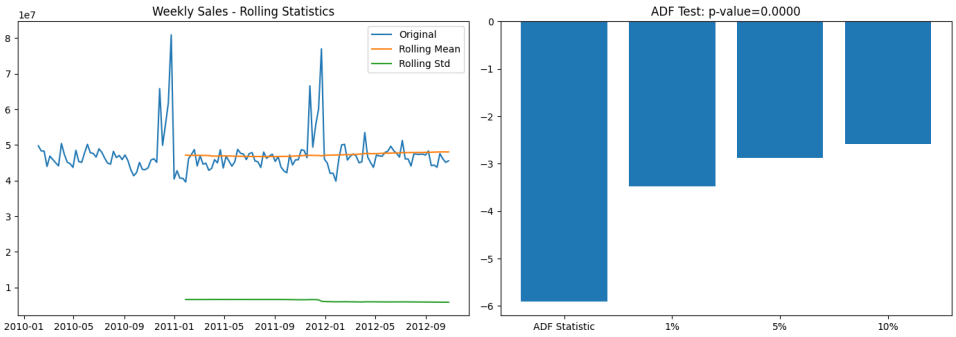

Figure 6.2(a): Weekly Sales Rolling Statistics and corresponding critical values

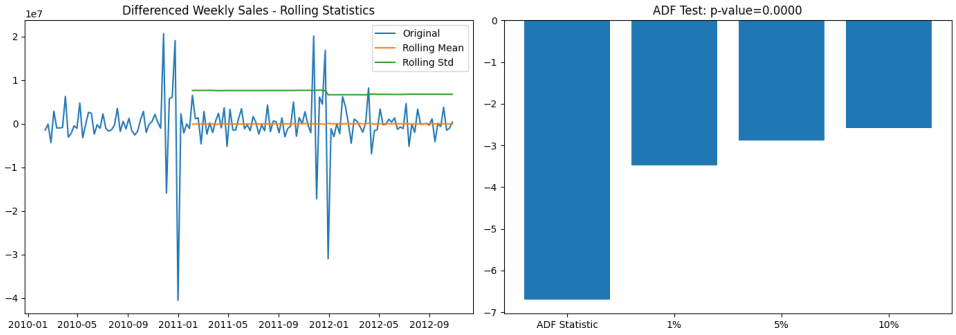

Figure 6.2(b): Differenced Weekly Sales Rolling Statistics and corresponding critical values

- **Original Series:** Low p-value (<0.05) fails to reject the unit-root null hypothesis, confirming stationarity.

7. Time Series Decomposition

An additive seasonal decomposition with period = 52 weeks splits the series into:

- **Trend:** Underlying growth.
- **Seasonality:** Recurring annual patterns.
- **Residual:** Noise and unexplained fluctuations

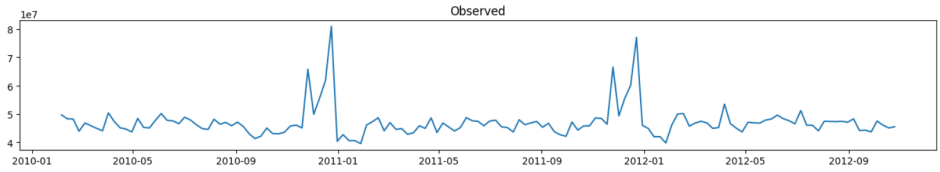

Figure 7(a): Observed Values

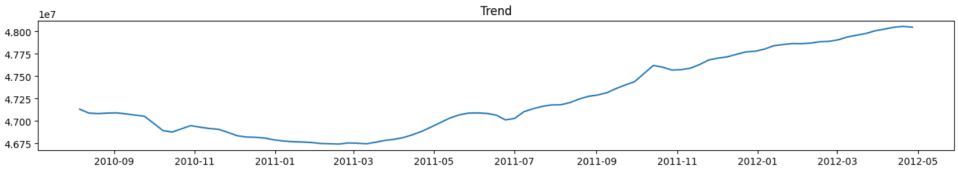

Figure 7(b): Trend Series Split

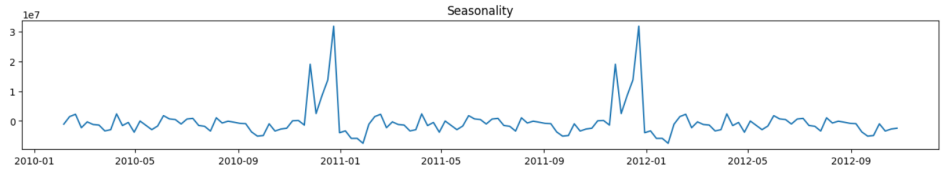

Figure 7(c): Seasonality Series Split

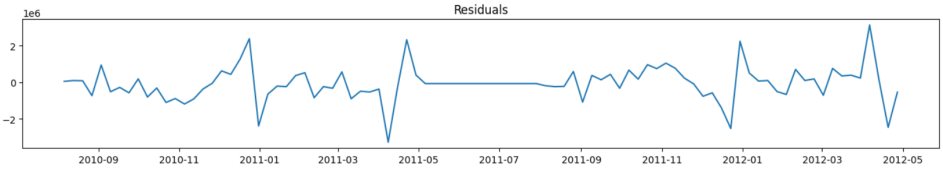

Figure 7(d): Residuals Series Split

Visualization of each component confirms **strong yearly cycles** and a **gradually rising trend**.

8. Autocorrelation Analysis

ACF and PACF Plots (Differenced Series)

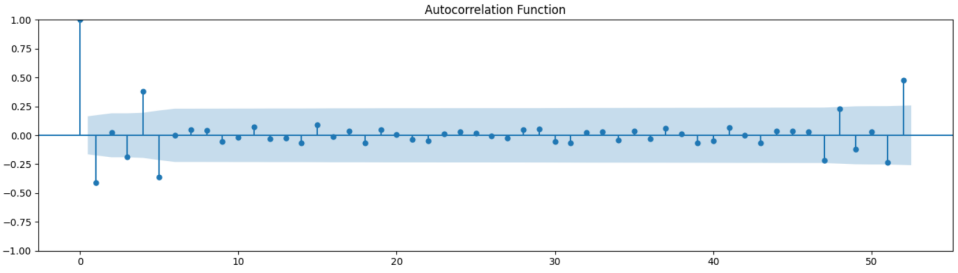

Figure 8(a) : ACF Plot

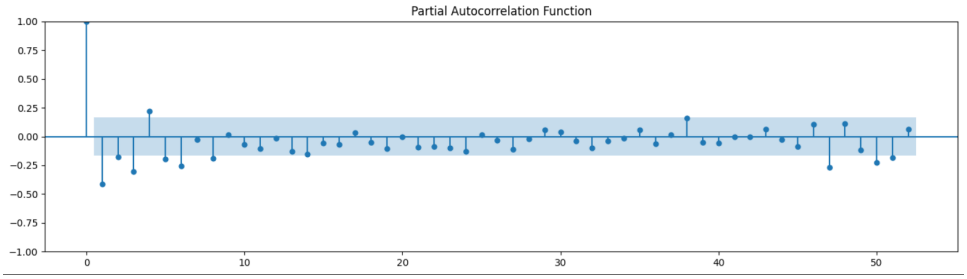

Figure 8(b) : PACF Plot

- **ACF:** Significant spikes at lag 1 and lag 52 suggest persistence and yearly seasonality.
- **PACF:** Cutoff after lag 2 indicates an AR(2) component.

These diagnostics support ARIMA(p=2, d=0, q=1) with seasonal order (P=2, D=0, Q=1, s=52). d=0 and D=0 are supported by the fact that a visual inspection also does not reveal any clear trend in the data.

9. Model Development

9.1 Data Split

Dates are split chronologically into training (first 85%) and testing (last 15%) to simulate forecasting on future data.

9.2 ARIMA Model

- Fitted ARIMA(2,0,1) on training aggregate sales.
- Forecasted on test period and plotted against actuals to assess performance.

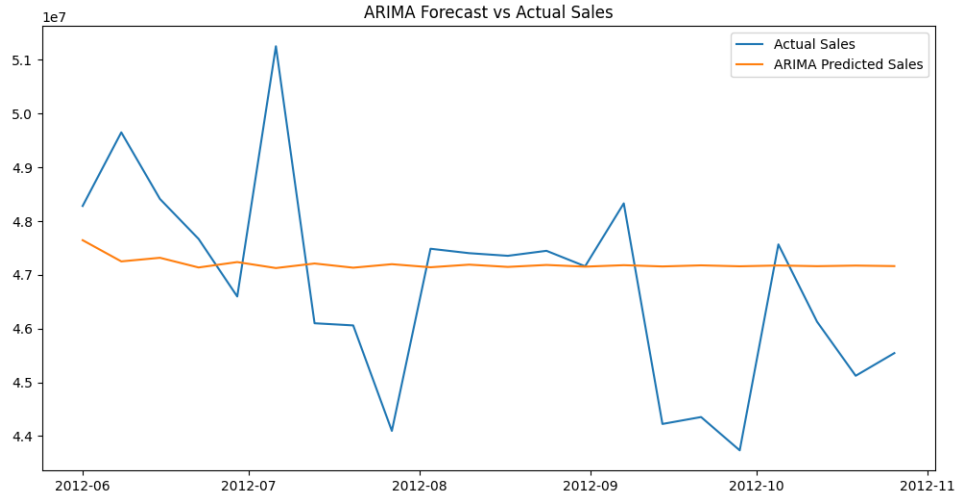

Figure 9.2: ARIMA Forecast vs Actual Sales

9.3 SARIMA Model

- Fitted SARIMA (2,0,1) (2,0,1,52) incorporating annual seasonality.
- Forecast of test period aligns more closely with actual sales patterns, capturing seasonal peaks and troughs.

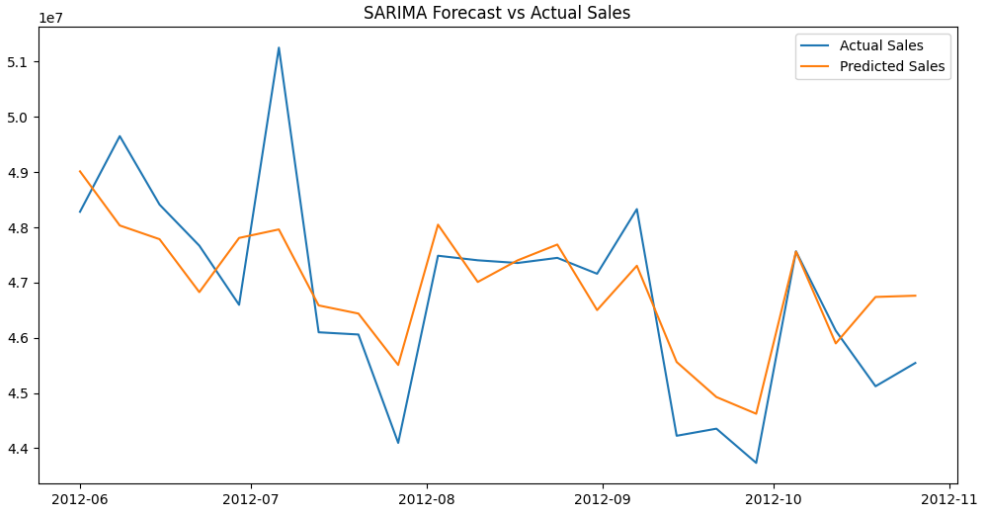

Figure 9.3 : SARIMA Forecast vs Actual Sales

10. Results and Evaluation

In this section, we will compare the performance of ARIMA and SARIMA models applied to the dataset in question, with a focus on their ability to capture underlying patterns and provide accurate forecasts.

1. Model Characteristics

Feature	ARIMA	SARIMA
Seasonality Handling	Does not model seasonality	Explicitly models seasonal patterns
Parameter Complexity	Fewer parameters, simpler model	More parameters due to seasonal terms
Forecast Behavior	Captures overall trend	Captures both trend and seasonality
Use Case Suitability	Appropriate for non-seasonal data	More appropriate for seasonal time series

2. Visual Inspection

Upon analyzing the plots of the forecast outputs from both models:

- The **ARIMA model** captures the general trend of the data but fails to accurately reflect the recurring seasonal fluctuations observed in the time series.
- The **SARIMA model**, on the other hand, more effectively tracks both the trend and the periodic seasonal variations. The model's forecasts align more closely with the actual data, particularly in areas where seasonal peaks and troughs occur.

3. Residual Analysis

If we consider the residuals (model errors):

- The residuals of the **ARIMA model** exhibit visible patterns, suggesting that some structure remains unaccounted for — likely due to the unmodeled seasonality.
- The **SARIMA model** residuals appear more randomly distributed and centered around zero, indicating a better fit and capturing of the underlying structure of the time series.

11. Conclusion

This project demonstrates a comprehensive pipeline for time series forecasting of weekly sales:

1. Data integration and cleaning ensure reliable inputs.
2. Feature engineering captures temporal cycles.
3. Stationarity checks guide appropriate differencing.
4. Decomposition and autocorrelation plots inform model specification.
5. Seasonal ARIMA outperforms the basic ARIMA in capturing annual patterns.