Walmart Weekly Sales

Time Series Final Project 2025

Authors: Lily Campbell, Pat O'Hea, Sajan Mehta, Ana Sy-Quia

Abstract

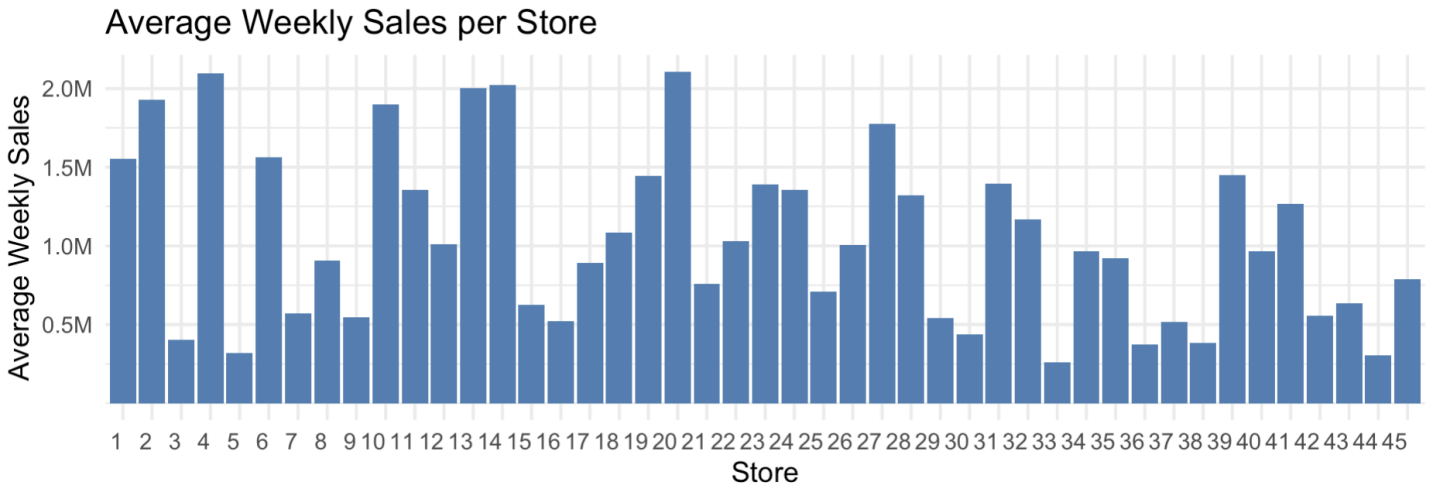
This analysis evaluates four approaches to forecasting total weekly sales for Walmart retail stores. It uses data on sales volumes for 45 stores reported weekly for two and a half years. Evaluated methods include simple ARIMA hierarchical time series forecasting, Fourier transformations, Prophet, and state space models with hierarchical time series forecasting. The best approach based on RMSE was a state space model with a hierarchical bottom-up approach. The analysis also suggests that Walmart needs to incorporate store level sales metrics and exogenous variables to best forecast their total future sales.

Problem Overview

This report seeks to identify the best method of forecasting total sales volume for all Walmart retail stores based on weekly store data. This task is vital for Walmart to be able to properly manage its stocking of products and staffing, as well as plan for future investment in more stores. Our dataset includes weekly sales volumes for 45 Walmart stores for February 2010- October 2012, as well as potentially explanatory variables including the Consumer Price Index, average temperature, unemployment, fuel price and whether the week included a holiday. We hypothesized that the exogenous variables will help to explain total sales volume, and total sales will have significant seasonal trends that need to be incorporated into the modeling approach.

Dataset and Feature Engineering

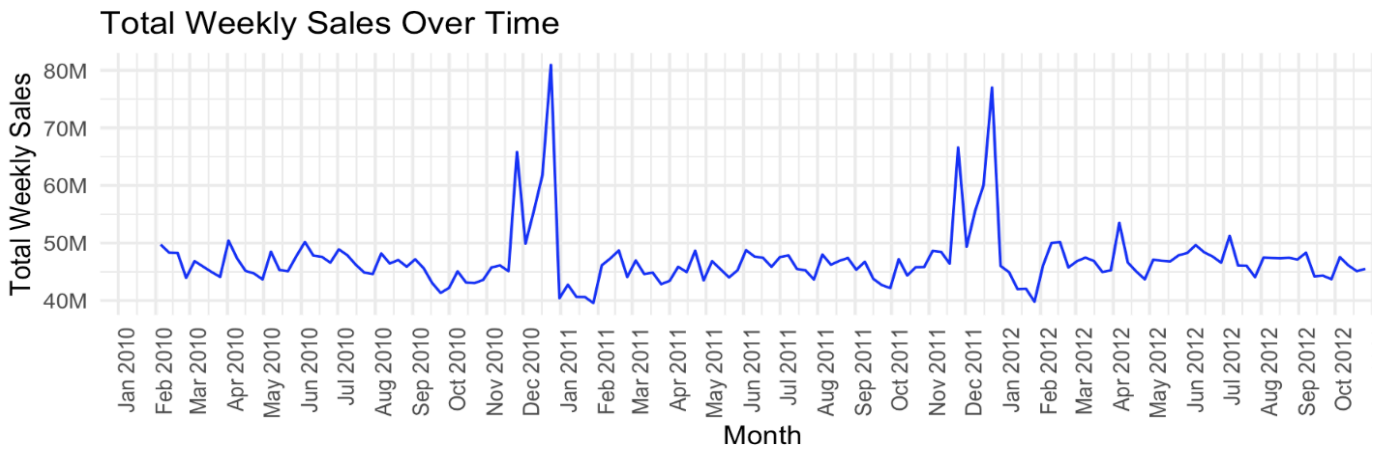
Our dataset consists of weekly sales data for 45 Walmart retail stores for 143 weeks. There was no missing data, and no extreme outliers were detected. The store level average weekly sales range from $260K to $2.1M, with an average of $1M weekly sales for a Walmart store.



The total volume of sales over all stores in the dataset ranges from 39.5M to 80.9M over the 143 weeks of data, with an average of 47.1M. Total volume of sales shows significant seasonality, as can be seen in the plot below. It appears that Walmart sales spike annually around Thanksgiving and December holidays.

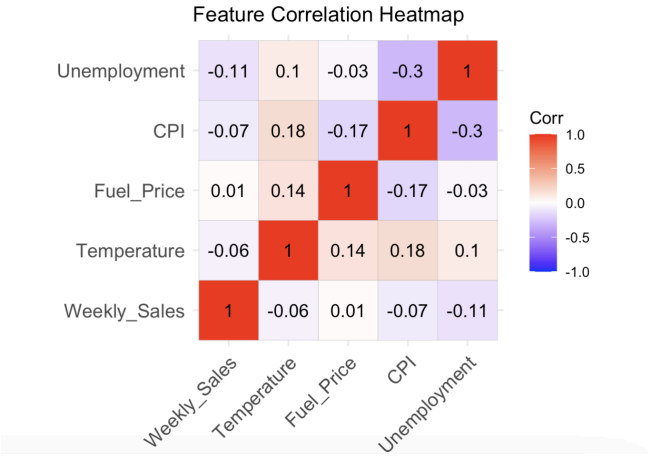
To assess stationarity, we conducted tests such as the Augmented Dickey-Fuller (ADF) test and examined differencing requirements. Total weekly sales is stationary in its original form, meaning that its mean and variance remain stable over time. The absence of a trend or strong seasonality suggests that no transformation is required for modeling. Temperature is also stationary, which aligns with the expectation that temperature follows seasonal fluctuations but does not exhibit long-term trends or unit root behavior.

Unemployment required one difference to achieve stationarity, indicating that it exhibits a persistent trend over time and that a first-order differencing transformation was necessary to remove it. Fuel price and CPI both required two differences to become stationary, suggesting that they have a strong long-term trend, likely due to economic inflation and external market forces. Second-order differencing effectively removed these trends, making them suitable for time series modeling.



The dataset also included the week's unemployment rate, average fuel price, average temperature, and the consumer price index (CPI) for each store. We hypothesized that high unemployment in an area will have a negative effect on sales if consumers have less disposable income. Unemployment had a - 0.11 correlation with weekly sales, but did not have significant granger causality of weekly sales. We hypothesized that fuel price could also have a negative impact on weekly sales, as customers may be less willing to travel to a store if the cost is greater. Fuel price had little correlation with weekly sales, but did have significant granger causality. We hypothesized that CPI, which is a common measurement of inflation, may signify a drop in weekly sales. CPI had a -0.07 correlation with weekly sales, and had significant granger causality.

Unemployment, Fuel Price and CPI may have mixed effects on weekly sales because Walmart can be seen as a discount store. In more affluent areas where Walmart's competitors may be more expensive, consumers may be more likely to shop at Walmart when they have relatively less spending power. However in less affluent and rural areas where Walmart is the primary shopping choice, lower purchasing power may make consumers less likely to shop at Walmart. This may explain why correlation between these variables and total week sales is low, but could still be useful in forecasting sales at a local store level.



Feature Engineering: We tried leveraging lags of the independent variables in many of our models, but found that these lags provided little benefit. We hypothesize that this is the result of two factors: First, the weekly nature of the data allows for near-real-time adjustments to variables such as temperature and fuel price. Second, CPI and unemployment are not very volatile variables and do not typically change radically on a week-to-week basis.

Model Approach - ARIMA Hierarchical Time Series Forecasting

The dataset has a hierarchical structure, with each store's weekly sales adding up to the total volume of sales for all walmart stores. We experimented with using a bottom-up approach, using ARIMA to forecast each store's sales for the test period, then summing them to forecast total sales. We hypothesized that this approach would help to capture local trends in each store's sales which may be overlooked while only forecasting the total sales based on past total sales. The auto.arima function was used on each store's time series, then the sum of all stores forecasts was used to determine the total weekly sales forecast.

This approach resulted in an RMSE of $4.3M. Because ARIMA was used for each individual store, the monthly and annual seasonality may not have been adequately captured. Further discussion of this model's results are in the appendix. We decided that this approach may benefit by trying different forecasting methods like X-SARIMA or state-space transformations on each store before aggregating to total sales.

Model Approach - Prophet Model

We applied Facebook's Prophet model to forecast weekly total sales for all Walmart stores. Prophet is well-suited for handling time series data with strong seasonal trends, making it a good candidate for this problem. Using a bottom-up approach, we trained separate Prophet models for each store incorporating both weekly and yearly seasonality and then summed the individual forecasts to obtain the total sales prediction. We also included holiday indicators as a regressor, and testing showed that additional regressors such as temperature, fuel prices, CPI, and unemployment slightly degraded RMSE. Setting Prophet's seasonality mode to multiplicative better captured proportional seasonal effects and long-term growth patterns.

This approach resulted in an RMSE of $2.28 million for 2012, with a relative RMSE of 4.9% of mean sales, showing that Prophet effectively captured the overall trend and seasonal patterns in sales. However, some deviations between actual and predicted sales indicate that Prophet may not fully capture localized variations or extreme fluctuations in the data. Additionally, early 2012 predictions still showed some underestimation.

Model Approach - Hierarchical Dynamic Regression with Fourier Seasonality and ARMA Errors

Since there were 45 different stores, we decided to go for another bottom up hierarchical approach, but instead using the following regressors, Holiday Flag, CPI, and Fuel Price with ARMA errors and Fourier seasonality. This allowed for the model to pick up on changes that occurred to individual branches, as some were not as affected by the regressors as others, and some stores had different seasonality patterns. Rather than using the hts package, it made sense to develop the seasonality and regressors in a function which was then applied to each store's individual data.

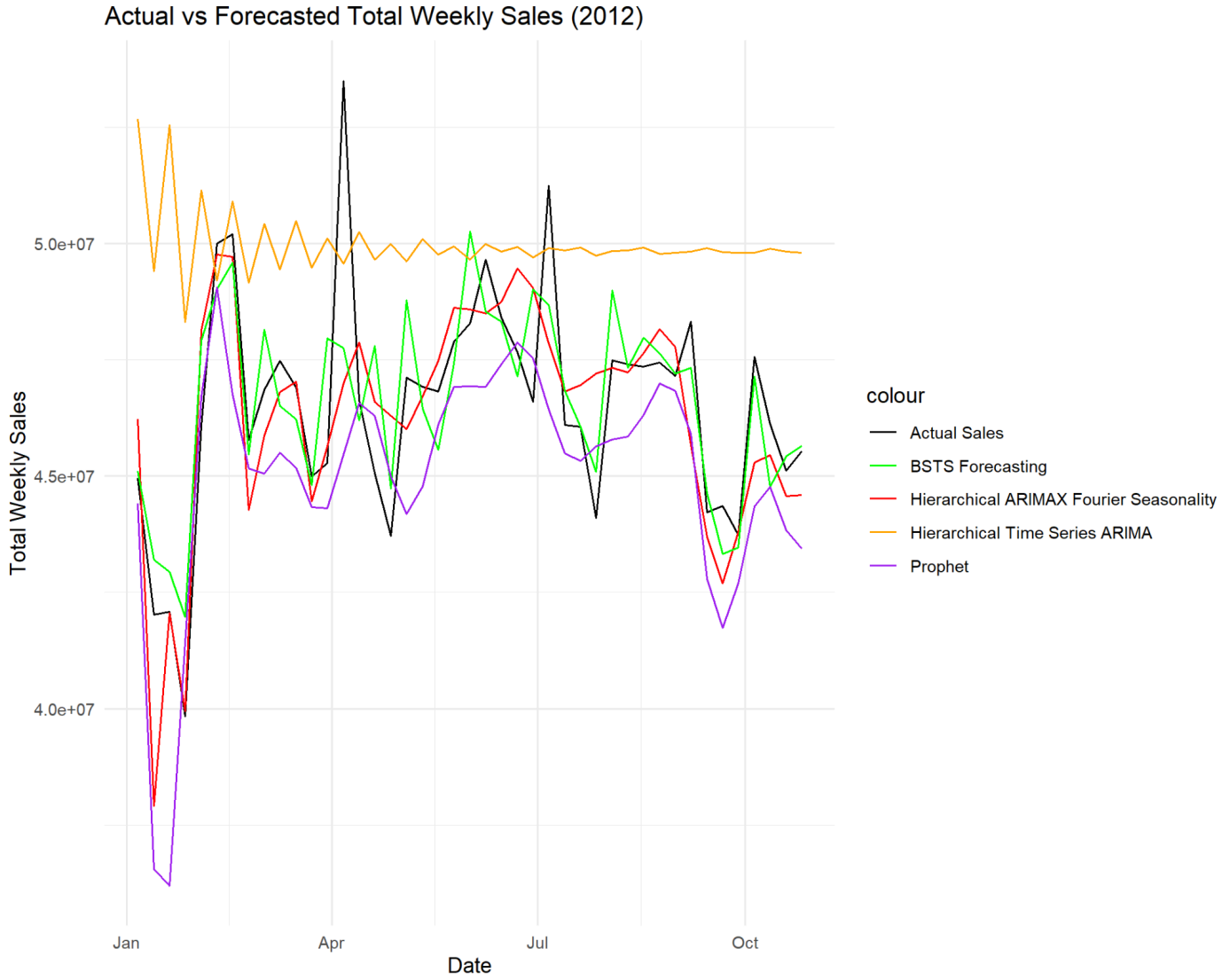
The model with the best RMSE used 8 Fourier harmonics on a 52 week frequency. The ARMA order was generally around (3,0,0) or (0,0,3), not having both autoregressive and moving average components. The RMSE of this modeling approach was

$1.76 million for 2012, providing a relative RMSE of 3.8% of mean sales. The Dynamic Regression model with Fourier Seasonality and ARMA errors performed well at capturing the trends, seasonality, and shocks in the data.

Model Approach - Bayesian Structural Time Series Modeling

After visualizing the trend decomposition of the time series we decided to implement a Bayesian Structural Time Series (BSTS) state space model with hierarchical time series incorporation. We tried feature and model engineering strategies including incorporation of lagged regressors and log transformation of our target variable, and altering the state space to include different trend and seasonal components.

Ultimately, we found that a yearly seasonal component with a local level trend and two autoregressive terms resulted in the best model performance (lowest RMSE). We found that some of the regressors (Temperature, Holiday Flag) showed low inclusion probabilities, so we chose to exclude them from our final model in favor of a simpler more generalizable model.



Selected Model

Our final selected model was a BSTS state space model using hierarchical time-series to aggregate individual stores in order to forecast total weekly sales across all stores (bottom-up approach).

Model structure:
Dependent Variable: Log of individual store weekly sales
Independent Variables: Fuel Price, Consumer Price Index & Unemployment Rate.
State Space Components:

- Trend: Local Level
- Seasonal: Yearly Seasonal (52 weeks)
- Autoregressive Terms: Two

We chose this model because of its relative simplicity and accurate forecasting performance. RMSE from this model was about $1,485,000, compared to the total average of $47,000,000 weekly sales. Of note, this model's residuals showed significant autocorrelation and abnormality, which is observable in the forecast plot where the predictions are close to the actual values - however they are consistently above or below the actual values. This is an important note because while this model is useful for point-estimates on forecasts, it shouldn't be relied upon for confidence intervals or statistical inference regarding these forecasts. This limitation is partially a symptom of the model, but also a limitation of the data. This model could certainly be improved, but there are likely exogenous variables not included in this dataset that influence sales patterns.

Evaluation

Our analysis aimed to determine the best forecasting approach for Walmart's total weekly sales using a variety of time series models. We explored four different methodologies: ARIMA hierarchical forecasting, a bottom-up Prophet model, Fourier-transformed ARIMA, and a state-space hierarchical model. Each approach had strengths and weaknesses in capturing the trends, seasonality, and external influences on Walmart's sales.

The ARIMA hierarchical approach struggled with capturing long-term seasonality, resulting in an RMSE of $4.3 million. The bottom-up Prophet model, which incorporated weekly and yearly seasonality along with holiday indicators, improved performance, achieving an RMSE of $2.28 million with a relative RMSE of 4.9%. However, Prophet's forecasts still showed some underestimation in early 2012 and did not fully capture localized store-level variations. The Fourier-transformed ARIMA model further improved accuracy, achieving an RMSE of $1.76 million, benefiting from the ability to model store-specific seasonal effects. Finally, the Bayesian Structural Time Series (BSTS) state-space model with a hierarchical bottom-up approach delivered the best performance, with an RMSE of $1.49 million. This model successfully incorporated store-level dynamics while maintaining generalizability, making it the most effective method for forecasting Walmart's total sales.

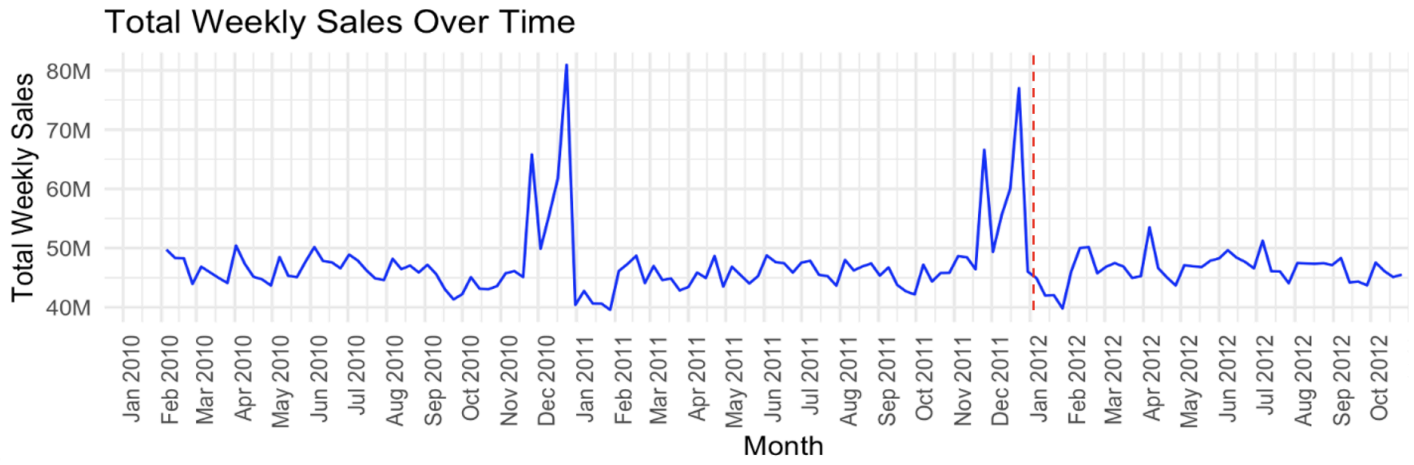
One of the key takeaways from this analysis is that bottom-up approaches were more effective than top-down forecasting, as they allowed models to capture store-specific variations that would otherwise be lost. Additionally, our results showed that external regressors do not always improve forecasting performance. While we initially hypothesized that fuel price, CPI, and unemployment would provide meaningful insights, their inclusion sometimes degraded RMSE, suggesting that their impact on sales is more complex than simple correlation measures might indicate. Finally, this analysis reinforced the importance of evaluating multiple forecasting approaches rather than relying on a single method, as different models varied in their ability to capture long-term seasonality and localized store trends.

For future work, alternative ensemble methods that combine the strengths of different models, such as integrating Prophet's trend-capturing ability with state-space modeling, could be explored. Additionally, testing alternative time series models, such as XGBoost or LSTM-based approaches, could provide insight into whether machine learning models offer advantages over traditional statistical forecasting methods within the constraints of the existing dataset.

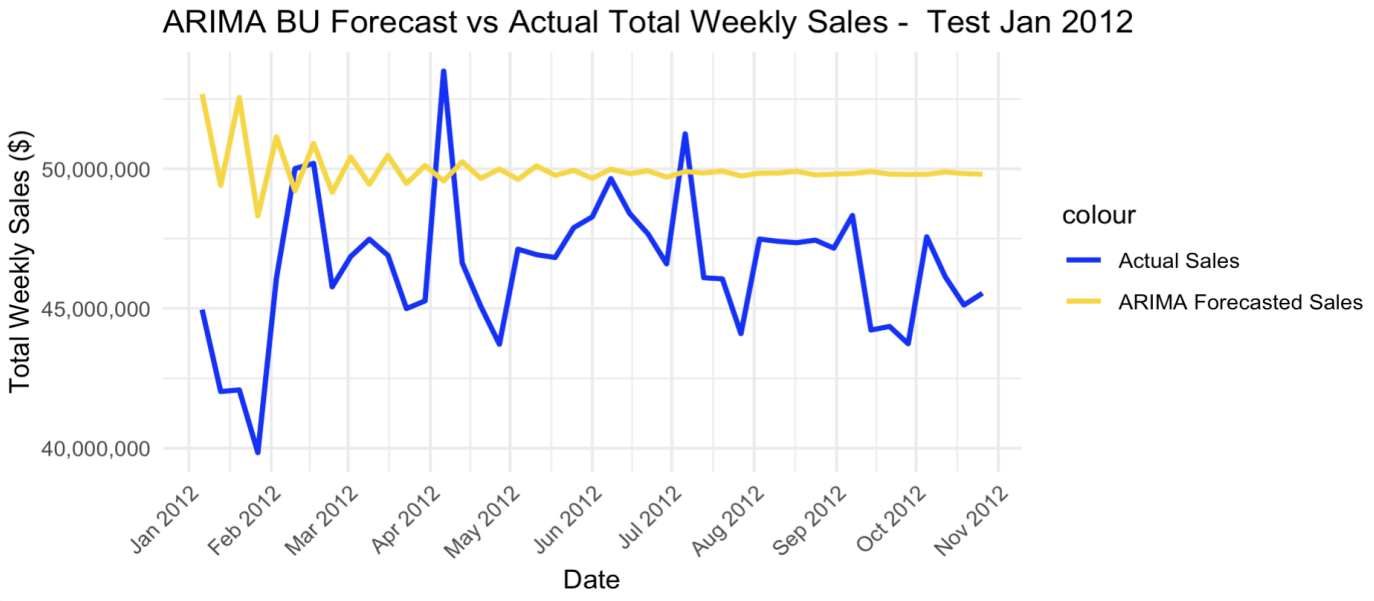
Appendix

Notes on ARIMA Model:

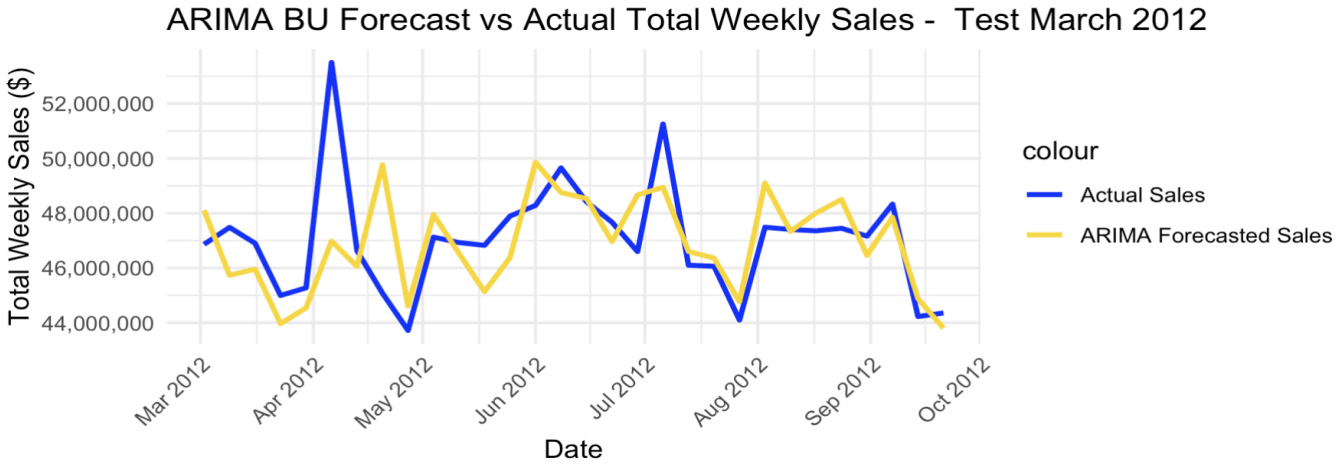
We specified our test period to begin January of 2012, which unfortunately coincided with a sharp drop in the total sales after the December 2011 holiday season. The training data only had one prior spike and subsequent drop in sales in December 2010. We hypothesize that the auto.arima model used in the first ARIMA model proposed was unable to capture this seasonal trend successfully, and was overly influenced by the preceding peak in data. When the train/test threshold was adjusted, then the ARIMA model shows a significantly better fit to the test data.

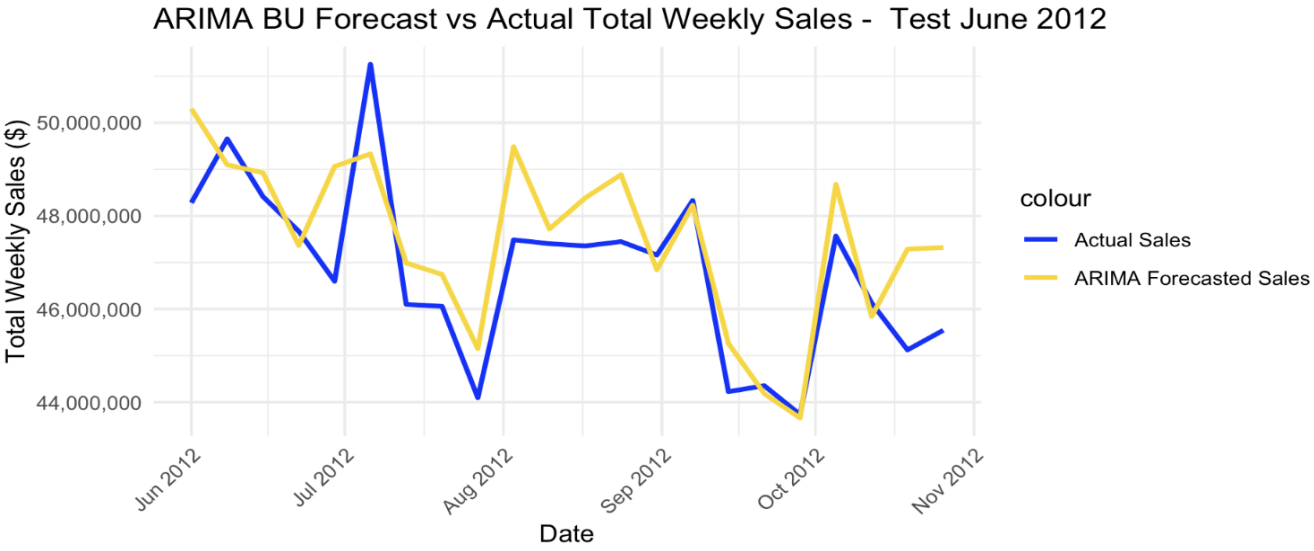


Current Test/Train Split Results:



Adjusted Test/Train thresholds:





Group Contributions:

Lily: EDA, Project Overview, Hierarchical ARIMA

Ana: EDA, Feature Engineering, Prophet

Pat: EDA, Selected Model, BSTS Hierarchical

Sajan: EDA, Plots, Hierarchical Fourier Dynamic Regression