Retail Demand forecasting

Surbhi Rathore

December 18, 2023

Abstract

Consumer Packaged Goods (CPG) are the products that are purchased by consumers on a regular basis and requires frequent replacement, for eg. packed/ unpacked food and beverages, household goods, personal care products, etc. In today's world, demand and supply chain loop in terms of CPGs is the most competitive space. The demands for CPGs is on the rise as it's directly proportional to the increasing population rate. The more the consumption the more the demand following more supply. The project is focused on implementing linear regression model along with fitting seasonal arima model for forecasting future weekly sales for walmart (up to 23 weeks in advance) to understand the science behind increasing or decreasing dynamics of demand. The project also considers other exogenous factors that may contribute in rise or fall of the demand. Demand forecasting plays a very significant role in pricing the product correctly, inventory management strategy and maximizing the profitability.

Keywords: Demand Forecasting, Linear Regression, ARIMA, Seasonal ARIMA, Exogenous variables.

1 Introduction

In the dynamic world of retail, accurate demand forecasting is very critical for optimizing inventory management, ensuring customer satisfaction, and ultimately, enhancing business profitability. Walmart is a retail giant, with an extensive and diverse product offering capability, their systematic functioning rely on forecasting models to anticipate consumer demand trends. This project aims to focus on demand forecasting for Walmart, utilizing both traditional time series analysis, specifically the Seasonal ARIMA model (which looks at past sales patterns), and a regression-based approach with additional predictors. The primary objectives of this project are as follows:

1. Examine a linear regression with relevant predictors. These predictors include factors such as unemployment rate, fuel price, temperature, and other external variables that could influence consumer purchasing behavior.
2. Implement a robust Seasonal ARIMA model to capture and forecast the temporal patterns inherent in Walmart's sales data.

1.1 Scientific question

To guide through analysis and model development, I would be interested in uncovering the following scientific questions:

1. What insights can be gained by integrating predictors into a linear regression model for demand forecasting? This question seeks to explore the value of incorporating external factors in improving the accuracy of sales predictions in Seasonal ARIMA model.
2. How well does the Seasonal ARIMA model captures the inherent seasonality in Walmart's sales data? This question addresses the effectiveness of the time series model in accounting for recurring patterns and fluctuations in sales over time.

1.2 Literature review

In the world of retail, accurately predicting customer demand is helpful for businesses. Previous researches have underscored the pivotal role of demand forecasting in enhancing operational efficiency and customer satisfaction. The ability to foresee trends, especially in large-scale companies like Walmart, not only streamlines inventory management but also ensures that customers find what they need when they need it.

Linear regression, a straightforward yet powerful tool, allows us to go beyond historical sales data. By considering exogenous factors, I aim to explore a more holistic approach to demand forecasting. Time series analysis with predictors can provide a nuanced understanding of the various influences on customer purchasing behavior.

The application of time series analysis in retail demand forecasting has been widely explored. Previous studies emphasize the significance of capturing temporal patterns, seasonality, and trends in sales data. The Seasonal ARIMA model, known for its capability to handle time-dependent variations, has been successfully employed in various retail contexts. With this project I seek to contribute to the existing body of knowledge by examining how well Seasonal ARIMA aligns with the unique sales dynamics of Walmart.

While much has been explored in the realm of demand forecasting, the application of these methods to the specifics of Walmart's sales forecasting is a unique challenge. Walmart's vast product range, diverse customer base, and supply chain dynamics require tailored approaches. By using insights from existing literature, I aim to adapt forecasting model to contribute to the development of effective and practical forecasting strategy.

The rest of the paper is structured as follows, first section 2, talks about the data being used for analysis and it's description. Section 3 shows the preliminary steps in analysis for understanding the nature of the data and implementation strategy in terms of predictive modeling. Section 4 elaborates on the methodologies implemented along with their equations. Section 5 highlights the final results obtained. Section 6 emphasizes on the conclusions.

2 Data

Data used for demand forecasting is Walmart Dataset is obtained from Kaggle.com. Walmart Data is a historical data that covers weekly sales from 2010-02-05 to 2012-11-01. The dataset includes information about 45 stores. Sales vary widely, with the median at 960,746 and the mean at 1,046,965. About 7 percent of the weeks are flagged as holiday weeks. It is a structured format data with 6,435 rows and 8 columns. All the variables listed below:

1. Store - it represents the store number
2. Date - the week of sales
3. Weekly Sales - sales for the given store recorded weekly
4. Holiday Flag - whether the week is a special holiday week (1 : Holiday week and 0 : Non-holiday week)
5. Temperature - Temperature on the day of the sale
6. Fuel_Price - Cost of fuel in the region during the week
7. CPI – consumer price index, CPI can be defined as a measure that examines the average change in prices paid by consumers for goods and services over time. It is an indicator of inflation.
8. Unemployment - Prevailing unemployment rate

Data description comes with holiday events that historically fell in respective weeks. 1. Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13 2. Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13 3. Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13 4. Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

3 Explanatory analysis

Time series analysis is a concept of analyzing data that has been recorded chronologically (strongly associated with time). For example, sensor data, weather data, sales data, etc. Strong association with time in terms of a weekly sales data, can reveal variations with respect to level of usage in different seasons or increasing trend due to the quality/ extreme demand of the product. Initial analysis steps involved understanding the underlying distribution of the data. In the first step of explanatory analysis I chose to examine if there is any visual relation between the response variable and predictor variables. Figure 1 shows that there isn't very strong relation between the response variable and respective predictor variable. There is no strong relationship between fuel price and sales, there is higher sale when unemployment rate is lower. Temperature shows slight variation in sale for lower and higher temperature. Different stores showing different sales amount and different ranges of CPI have similar sales distributions.

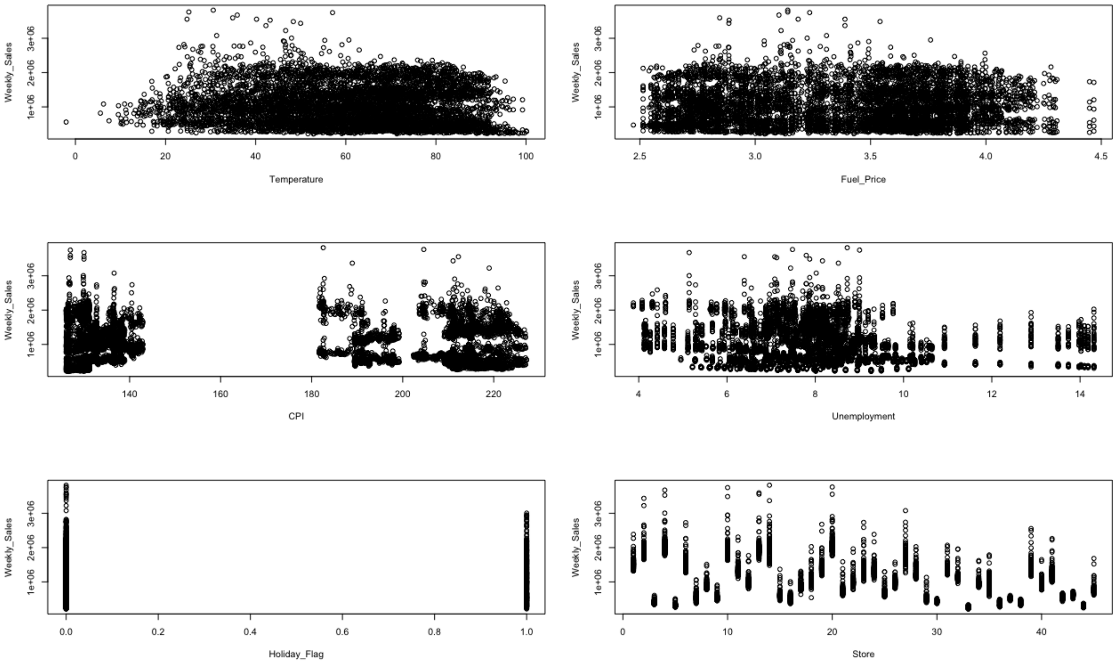


Figure 1: Weekly sales versus exogenous variables

To strengthen the analysis, I plotted a correlation plot, figure 2 clearly shows a weak

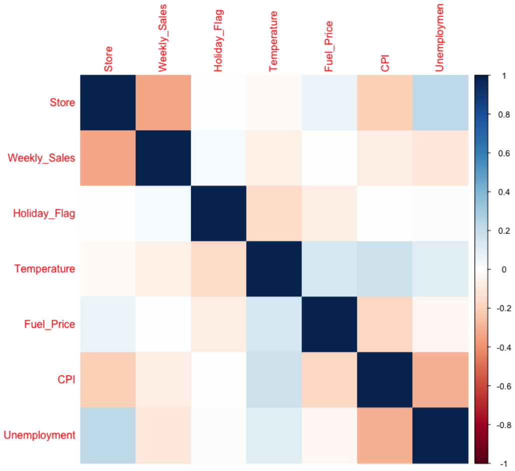


Figure 2: Correlation between response and predictor variables

positive and negative correlation with the response variable which does not explain much about it's respective effect on sales. The goal behind initial analysis was to fit a regression model and examine the predictors that significantly influence the rise or fall in sales rate. Linear regression follows an assuption that data should be normally distributed for the model to fit and produce effective results hence, Figure 3 shows the distributions of each variable and it exhibits that weekly sales has a right tail (right skewed) and it can be log transformed to achieve a normalized distribution. Unemployment is right skewed too. Further in the analysis I log transformed the variables. Figure 4 shows the distribution of all the log transformed features. I chose to only keep log(weeklysales) and log(unemployment) as transformed features and use others as it is. To understand the nature of response variable weekly sales, I plotted 5 log transformed weekly sales over time and it revealed a hint of yearly and monthly seasonality. 6 explains the seasonal aspect in the data and shows that festivals will lead to high weekly sales, especially a week before the Christmas and the Thanksgiving. Figure 7 shows the auto correlation (acf) and partial auto correlation (pacf) plots for response variable weekly sales. It explains how observations are correlated with the it's lagged version. ACF plot slowly decays and PACF shows a few partial peaks confirming the relation between observation and the errors introduced.

4 Statistical Methods

Figure 5 shows that sales feature in walmart dataset carry's a hint of yearly seasonality due to presence of yearly holidays. The statistical Method that fits the specifications learnt through explanatory analysis over walmart data suggests to implement regression to analyze the significant predictors followed by implementation of seasonal ARIMA/ seasonal ARMA. Below is the detailed description of the methodologies used.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an extension

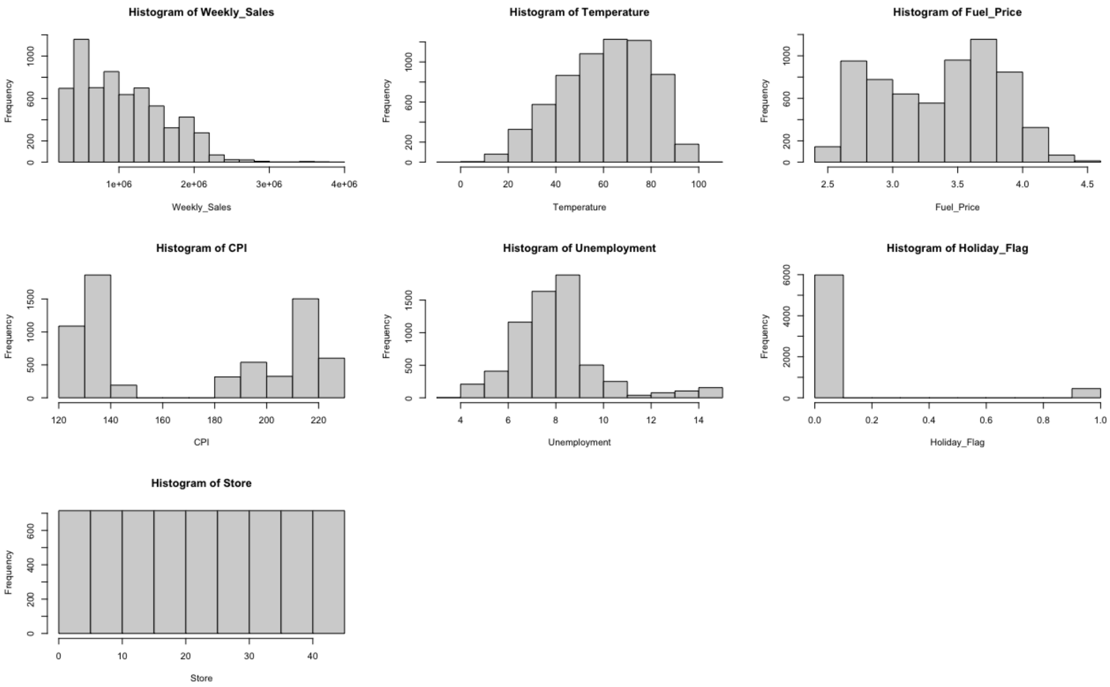


Figure 3: Distributions of each variable

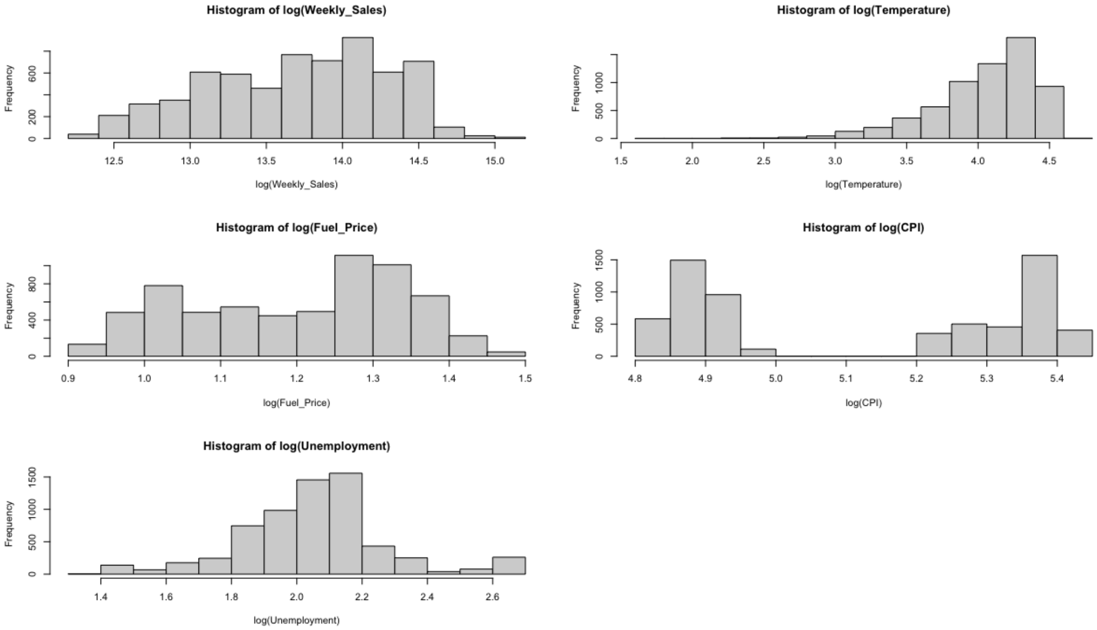


Figure 4: Distributions of log transformed variables

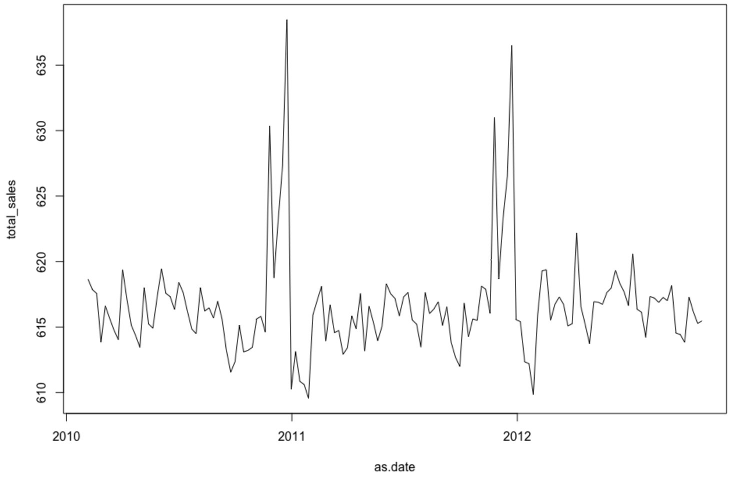


Figure 5: Time series plot of response weekly sales

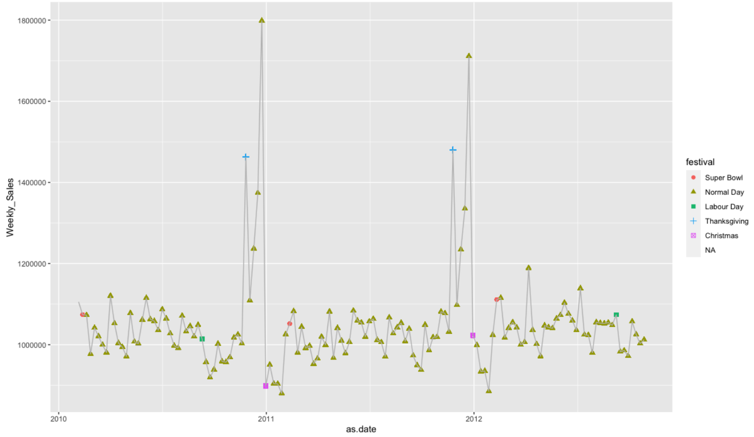


Figure 6: Time series plot of response weekly sales emphasizing the yearly peaks informing of seasonality in the data

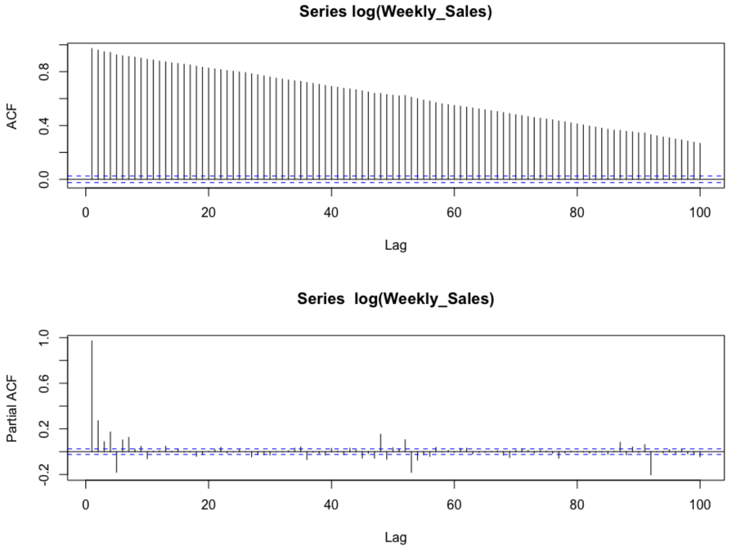


Figure 7: ACF and PACF of the weekly sales

of the ARIMA model that specifically considers seasonality in time series data. SARIMA is particularly useful for capturing and forecasting time-dependent patterns that occur at regular intervals, such as daily, monthly, or yearly seasonality. The general form of a SARIMA model is denoted as SARIMA(p, d, q)(P, D, Q)[s], where:

ARIMA(p, d, q) × SARIMA(P, D, Q)[s] ARIMA(p, d, q)×SARIMA(P, D, Q)[s]

Components of SARIMA are:

1. Autoregressive (AR) terms: These terms predict the current value (yt) based on its past values $(yt - 1, yt - 2, ..., yt - p)$. The order (p) specifies the number of lag terms used.

2. Moving Average (MA) terms: It accounts for random errors (ϵ_t) in the past ($\epsilon_{t-1}, \epsilon_{t-2}, \ldots \epsilon_{t-q}$). The order (q) specifies the number of lag terms used.

3. Seasonal AR (SAR) terms: These terms predict the current value based on past values at seasonal lags $(yt - S, yt - 2S, ..., yt - pS)$. The seasonal order (P) specifies the number of lag terms used.

4. Seasonal MA (SMA) terms: These terms account for random errors at seasonal lags (ϵ_{t-S}, ϵ_{t-2S}, $\ldots \epsilon_{t-QS}$). The seasonal order (Q) specifies the number of lag terms used.

5. Integrated component (I): This component represents the differencing, necessary to make the time series stationary. It involves taking the difference between consecutive observations to remove trends.

The general equation for a SARIMA model is given by:

$$yt = c+\phi_1 y_{t-1}+\phi_2 y_{t-2}+\ldots+\phi_p y_{t-p}+\theta_1 \epsilon_{t-1}+\theta_2 \epsilon_{t-2}+\ldots+\theta_q \epsilon_{t-q}+\Phi_1 y_{t-S}+\Phi_2 y_{t-2S}+\ldots+\Phi_P y_{t-PS}+\Theta_1 \epsilon_{t-S}+\Theta_2 \tag{4.1}$$

where: yt is the value at time t c is the constant term ϕ_i are the AR coefficients θ_i are the MA coefficients Φ_i are the SAR coefficients Θ_i are the SMA coefficients ϵ_t is the white noise error at time t S is the seasonal period (e.g., 12 for monthly data) p, q, P, Q are the orders of the AR, MA, SAR, and SMA terms, respectively

Choosing the appropriate orders (p, d, q, P, Q) is crucial for a good fit. I relied on methods like AIC and BIC to compare models based on their statistical information and additionally, analyzed autocorrelation (ACF) and partial autocorrelation (PACF) plots to identify the significant lags for AR and MA components. ARIMA model handles non-stationary data by differencing, producing a stationary data for further forecasting. Before implementing the model we must ensure the stationarity of the data.

Another method is, linear regression with predictors. It is a statistical model used for predicting the value of a dependent variable based on one or more independent variables (predictors). The model assumes a linear relationship between the dependent variable and the predictors. Linear Regression Model Equation for single predictor is given by:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4.2}$$

Where, Y is the dependent variable (the variable we want to predict) X is the predictor variable (the variable used for prediction) β_0 is the intercept (Value of Y when, X = 0) β_1 is the slope ϵ is the error term

Linera Regression with mutiple predictor is given by :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_k X_k + \epsilon \tag{4.3}$$

The underlying assumptions of Linear Regression are:

1. Linearity: The relationship between the dependent variable and the predictors is assumed to be linear.

2. Residuals (the differences between observed and predicted values) should be independent of each other.

3. The variance of the residuals should be constant across all levels of the predictors.

4. The residuals are assumed to be normally distributed.

5. The predictors should not be perfectly correlated with each other.

5 Results

The first model that I implemented was linear regression with multiple predictors, figure 8 shows the residual plot along with it's distribution which is not perfectly normal but close to a normal distribution. The acf decays showing auto correlations withs lagged version and pacf plot shows significant 4 peaks informing of the correlation with errors. Figure 9 shows the predictors that have significant impacts on the sales. Unemployment is a significant predictor with a p-value less than 0.05 and for every one-unit increase in the unemployment rate, the sales decreases by 0.25 units. Another significant predictor is temperature and CPI, although impact on these variables is not on a very large scale. While the model is statistically significant, the low R-squared values and some insignificant coefficients suggest that it might not be a strong overall fit and robust in explaining the variation in y.

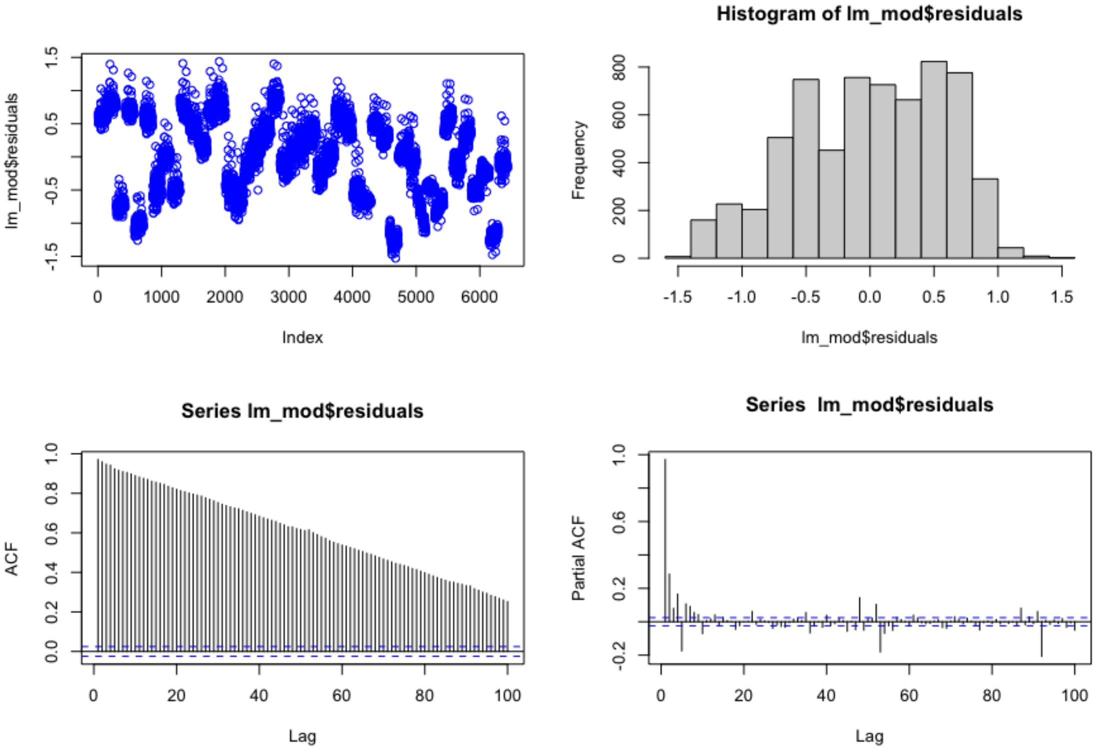
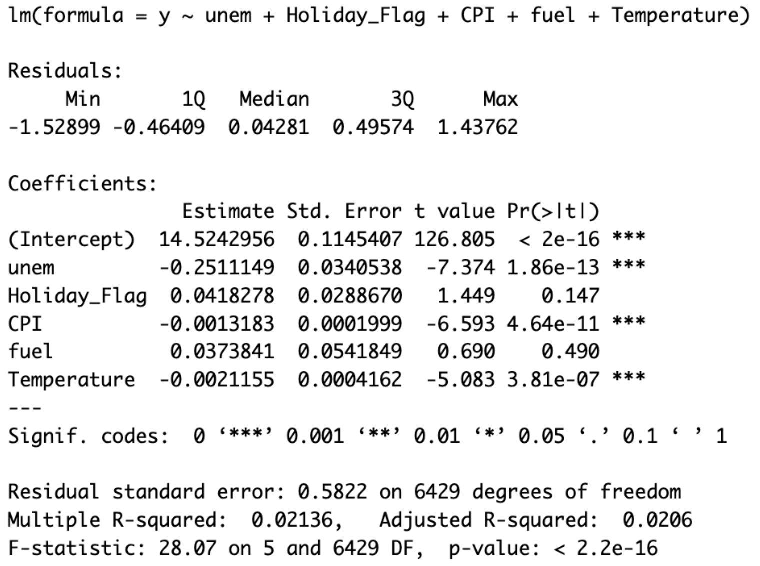


Figure 8: Linear rehression residual plots

Following the results, I started conversion of variables to a time series object, figure **??** shows the nature of plots over time. We can clearly observe the seasonal pattern in weekly sales plot, temperature shows the variation in temperature during different seasons, consumer price index (a measure of inflation) shows an increasing trend over time, unemployment does not follow an accurate trend but does show how unemployment has increased and decreased over



```
lm(formula = y ~ unem + Holiday_Flag + CPI + fuel + Temperature)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.52899 -0.46409  0.04281  0.49574  1.43762 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  14.5242956  0.1145407 126.805  < 2e-16 ***
unem         -0.2511149  0.0340538  -7.374 1.86e-13 ***
Holiday_Flag  0.0418278  0.0288670   1.449    0.147    
CPI          -0.0013183  0.0001999  -6.593 4.64e-11 ***
fuel          0.0373841  0.0541849   0.690    0.490    
Temperature  -0.0021155  0.0004162  -5.083 3.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5822 on 6429 degrees of freedom
Multiple R-squared:  0.02136,	Adjusted R-squared:  0.0206 
F-statistic: 28.07 on 5 and 6429 DF,  p-value: < 2.2e-16
```

Figure 9: Linear regression statistics

time. Fuel price also shows increasing variation over time.

In order for check if the data is stationary, I performed Augmented Dickey-Fuller Test and KPSS test which revealed that data is stationary with constant mean and a significant p value and hence, it does not require difference (current (t) - previous (t-1)). Differencing is a concept used to stabilize the mean and variance over time in the data for fitting the model better. Figure 12 shows that weekly sales has significant non-zero peaks in acf and a few peaks in pacf too. Q-Q plot suggest of a few outliers in the data. While fitting the model I chose the significant predictors - temperature, fuel price, unemployment and CPI. The data was split into 80-20 for train and test respectively. Considering the acf and pacf plots I chose the ordered (1,0,2) and (2,0,2) for respective AR, MA and seasonal components to fit seasonal arima model. Figure 13 shows the residual acf and pacf plots from the results, the peaks at respective lags are confined to the significant non-zero interval suggesting good fit of the model. Residual plot shows constant mean and the histogram shows a normalized distribution. It does give a hit of scattered points, but over all it suggesting a good fit. I also chose to test the function (introduced in class) to find the best seasonal arima model by looping over various values for different orders and return the one that best fits the model. I found order for (p,d,q) to be (1,0,1) and seasonal (P,D,Q) order to be (2,0,2). Figure 14 shows the residual plot suggesting of of good fit. ACF has one peak touching the confidence interval boundary yet the results show there isn't any correlation. Distribution plots show a normal distribution with a few dispersed residuals. Figure 15 shows the difference in residual plots obtained from each model.

I ran the model through outlier detection functions - Additive outlier (AO) and Innovative outliers (IO) to check if there is any outlier detected. The results showed that in model1 IO identified two points with innovation outliers at indices 43 and 47, with estimated magnitudes

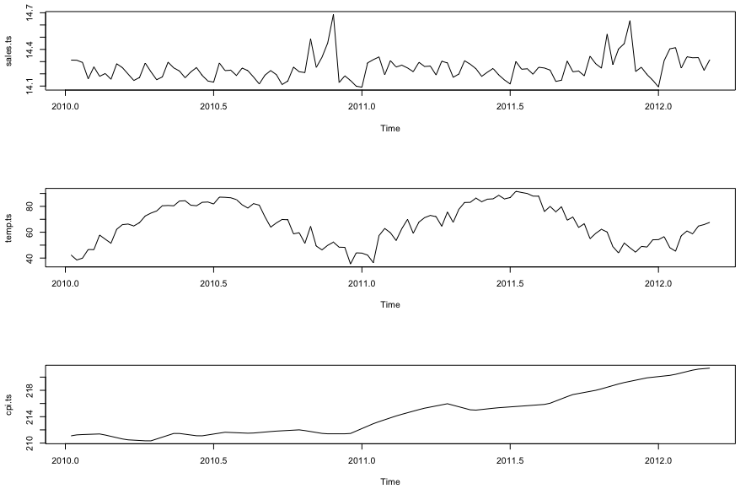


Figure 10: weekly sales, temperature and cpi plots as time series object over time

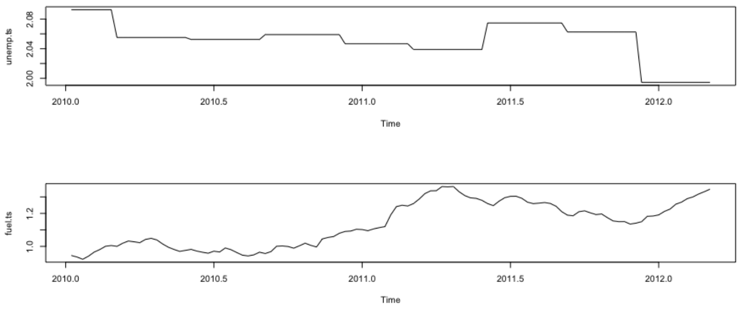


Figure 11: Unemployment and fuel price plotted as time series object over time

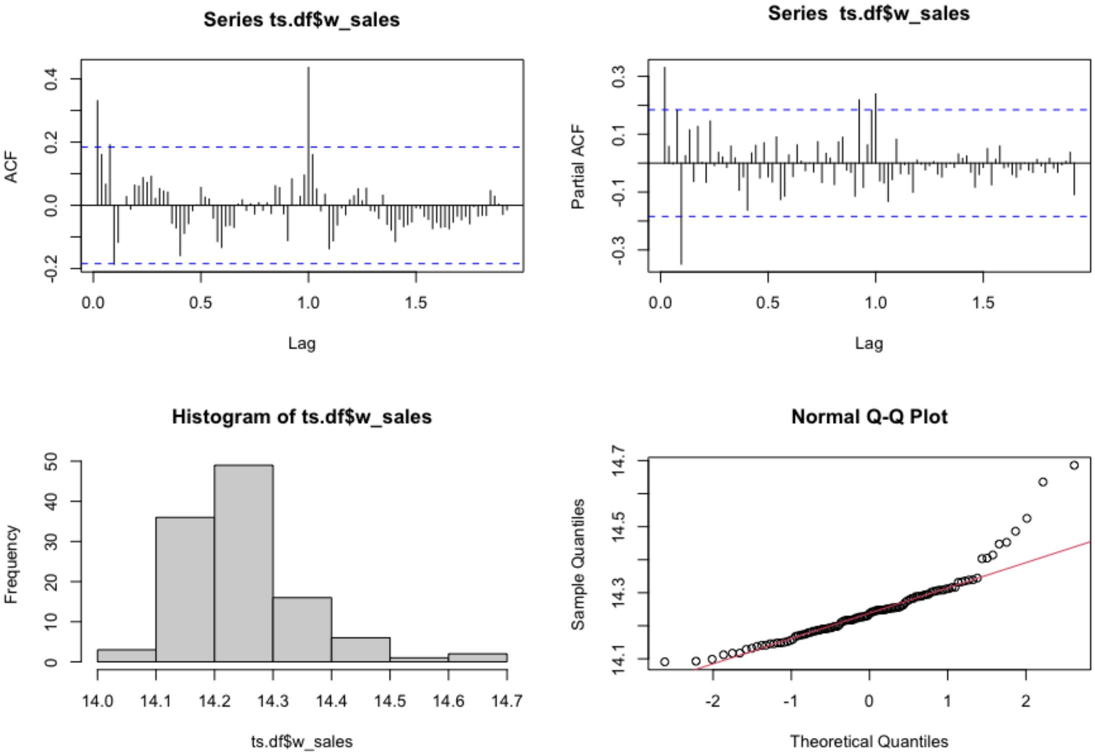


Figure 12: ACF, PACF and normality statistics for weekly ts

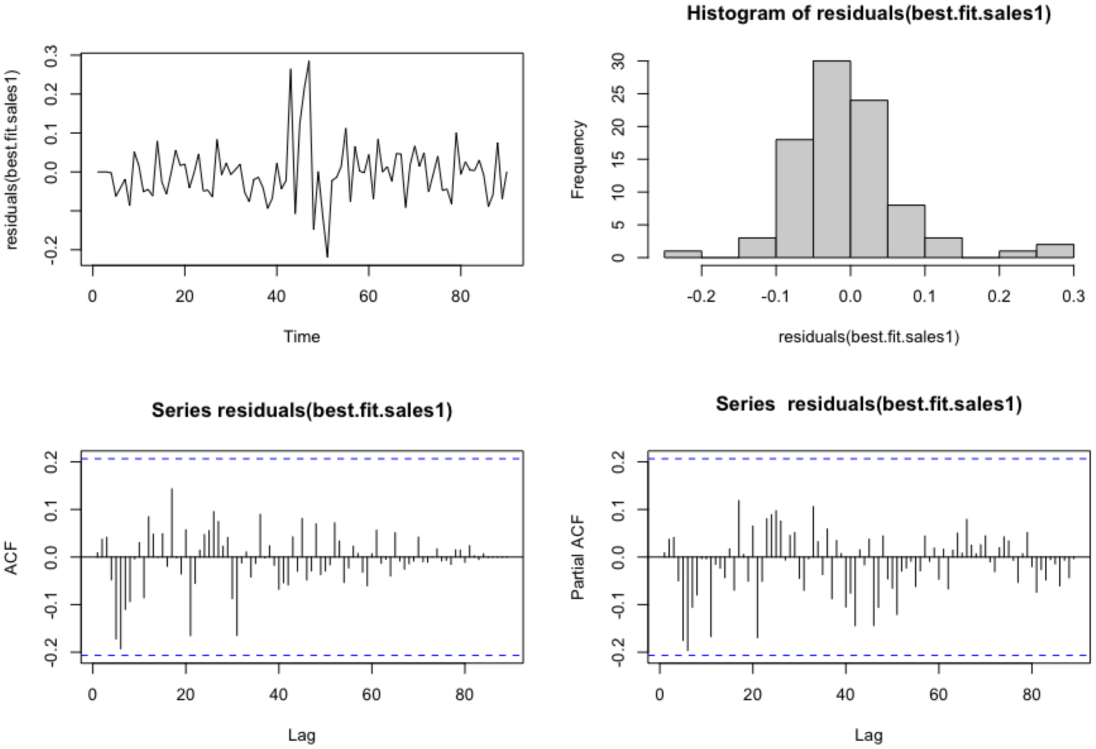


Figure 13: Distribution, ACF and PACF plots for Residual obtained from the model

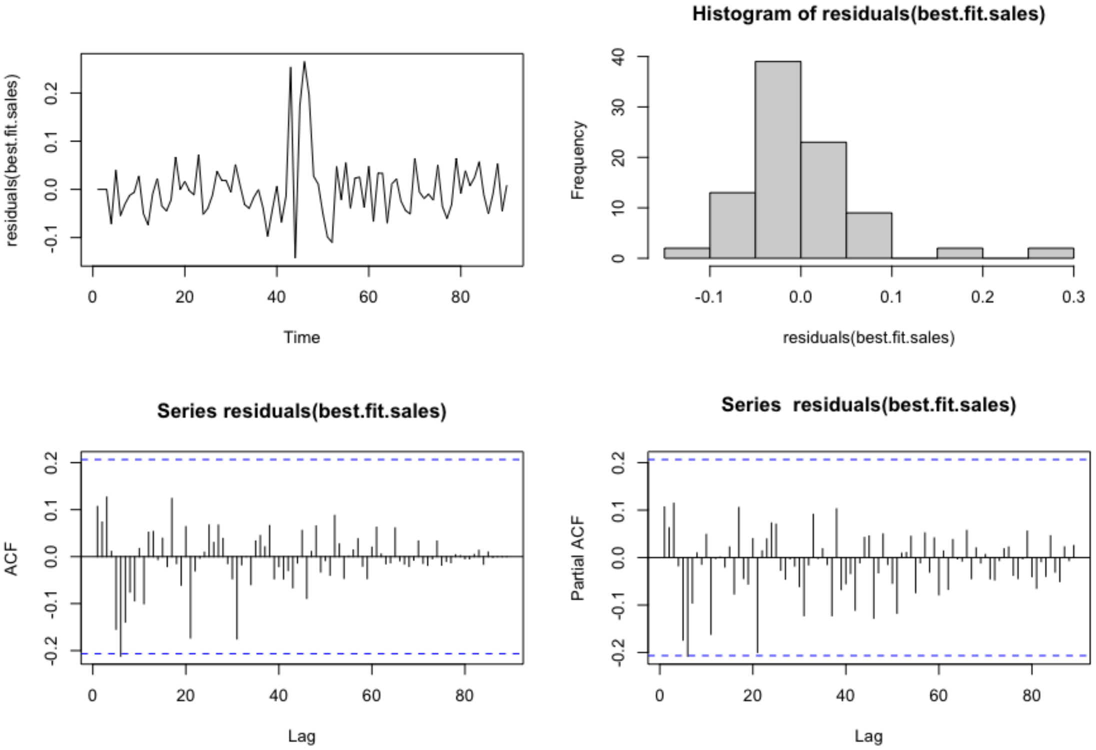


Figure 14: Distribution, ACF and PACF plots for Residual obtained from the model using pre defined function for finidng best fit model

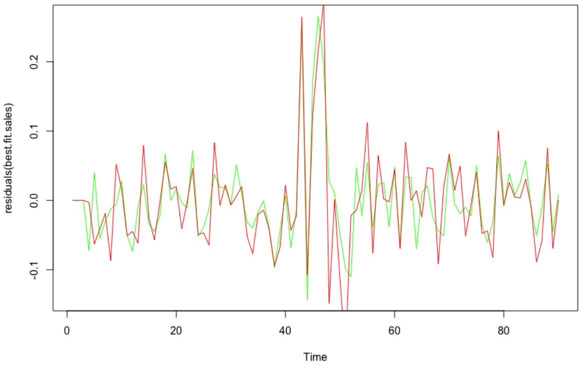


Figure 15: Residual plot for both model1 and model2

of 4.088329 and 4.400629, respectively and AO identified two points with additive outliers at indices 47 and 51, with estimated magnitudes of 5.2825 and -3.691476, respectively. In model2 IO identified that there is a sudden and temporary change in the behavior of the data at three intervention points (43,46 and 47) and no AO is detected. This informs that these points shows significant deviations from the expected behavior. I chose to forecast the sales for 23 weeks ahead using both model1 and model2. Figure **??** shows the forecasted data points which lies under the forecast interval for both models, both models produce very similar predictions for test data. The original data in blue is almost under the forecast interval except of two major peaks suggesting a close good forecast against the test data points. The Root Mean Squared Forecast Error (RMSFE) is a metric used to evaluate the accuracy of forecasting models. It is calculated as the square root of the mean of the squared forecast errors. In the context of forecasting with a seasonal ARIMA model, RMSFE for model1 and model2 are 1.024729 and 1.037951, respectively. The fact that both models have RMSFE values close to 1 suggests a reasonable level of forecasting accuracy.

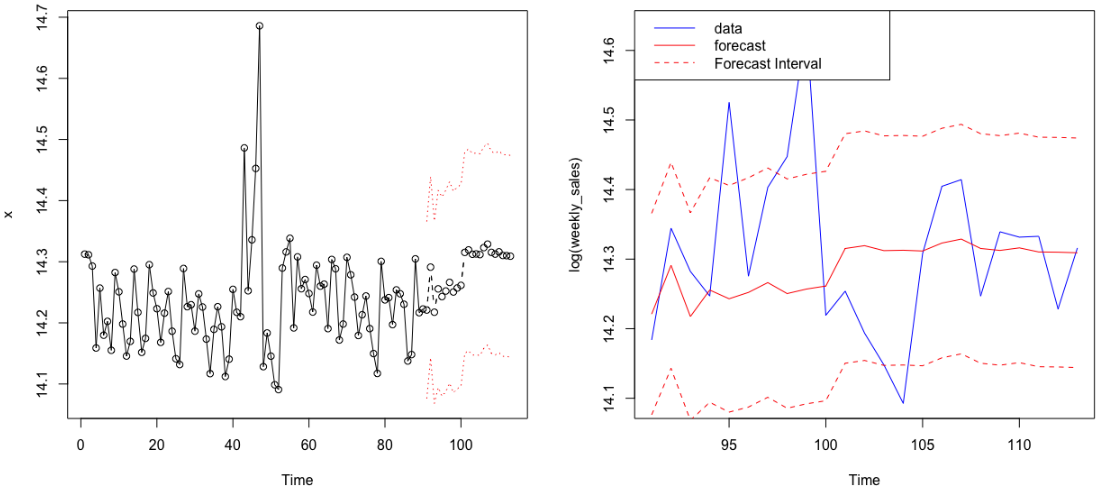


Figure 16: Residual plot for both model1 and model2

6 Discussion and Conclusion

The analysis began with a thorough examination of the data, utilizing Augmented Dickey-Fuller and KPSS tests results indicated that the data is stationary, suggesting no need for differencing. Key predictors used for model fitting are, temperature, fuel price, unemployment, and CPI. I fitted two seasonal ARIMA models with the chosen predictors and corresponding orders based on ACF and PACF plots. The residual analysis for both models shows great results, with ACF and PACF plots of the residuals exhibiting peaks confined to the significant non-zero interval. In forecasting, the models produced similar predictions for the test data, demonstrating consistency

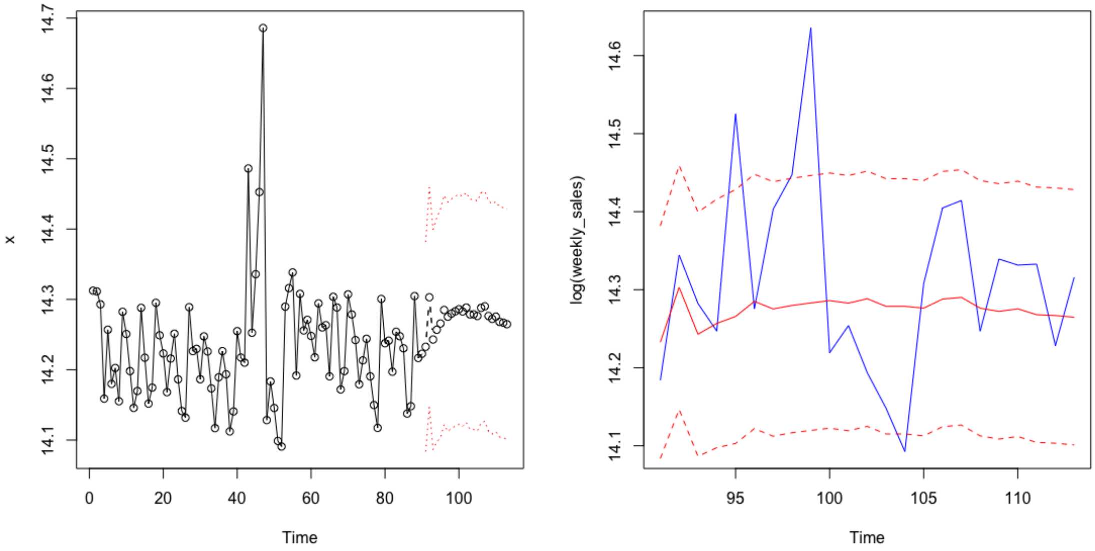


Figure 17: Residual plot for both model1 and model2

in their forecasting capabilities.

While the forecasting results are encouraging, it's important to acknowledge the inherent uncertainty associated with predicting future periods. Differences between the forecasted and actual data are expected, particularly for unseen periods during model training. I believe exploring advanced forecasting techniques can be a direction to consider, considering additional predictors, or implementing ensemble models for enhancing accuracy could be persued as an extension to this project.