Predictive Analytics Mini Project

Spring 2022

Project Report – Walmart Sales Forecasting

Group 3:

Shreyashi Mukhopadhyay

Shalmali Joshi

Namit Srivastava

Nhi Nguyen

Introduction:

Walmart (https://www.walmart.com/) is one of the largest retailers in the world and it is very important for them to have accurate forecasts for their sales in various departments. Since there can be many factors that can affect the sales for every department, it becomes imperative that we identify the key factors that play a part in driving the sales and use them to develop a model that can help in forecasting the sales with some accuracy. This project aims to describe the exploratory study of historical time series sales data for 45 Walmart stores spanning across 99 departments located in different regions of the country and tuning the best algorithms to obtain the best possible results for weekly sales forecasting. The overall data showed a 52-week seasonal frequency.

Holidays can create a huge impact on sales. So, if there is a good prediction on Sales then Walmart can calculate how much product to order during Holiday time. It will help in predicting which products need to be purchased during the holiday time as customers are planning to buy the products and they need to be available immediately. Through this prediction they can figure out which product to be made available at what time. This problem can also solve the issue of Marketing Campaigns as forecasting is often used to adjust ads and marketing campaigns can influence the number of sales. Walmart runs several markdown events throughout the year and these markdown events precede to the prominent holidays and it becomes even more important to schedule Markdown events preceding or at the time of popular Holiday events more efficiently.

Project description:

Predicting future sales for a company is one of the most important aspects of strategic planning.

This project aims to analyze how internal and external factors would affect the future Weekly Sales of Walmart, one of the biggest companies in the US.

This project also encompasses a complete analysis of data through data visualization, data exploration, pattern discovery and time series analysis in R as the programming language.

Data Availability:

The project data and relevant information are available at https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting.

The data collected ranges from 2010 to 2012, where 45 Walmart stores across the country were included in this analysis. Each store contains several departments, and this data aims to are tasked with predicting the department-wide sales for each store. It is important to note that we also have external data available like CPI, Unemployment Rate and Fuel Prices in the region of each store which, hopefully, helps us to make a more detailed analysis.

The following four holidays fall within the following weeks in the dataset (not all

holidays are in the data):

- Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
- Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
- Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
- Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Data description:

We have utilized three data files for our analysis from Kaggle.com, namely Train.csv, Stores.csv and Features.csv which contain the following information:

Train.csv

- Store: The store number. Range from 1–45.
- Type: Three types of stores ‘A’, ‘B’ or ‘C’.
- Size: Sets the size of a Store would be calculated by the no. of products available in the particular store ranging from 34,000 to 210,000.

Stores.csv

-Date: The date of the week where this observation was taken.

-Weekly_Sales: The sales recorded during that Week.
-Store: The store which observation in recorded 1–45.
-Dept: One of 1–99 that shows the department.
-IsHoliday: Boolean value representing a holiday week or not.

Features.csv

-Temperature: Temperature of the region during that week.
-Fuel_Price: Fuel Price in that region during that week.
-MarkDown 1:5: Represents the Type of markdown and what quantity was available during that week.
-CPI: Consumer Price Index during that week.
-Unemployment: The unemployment rate during that week in the region of the store.

Data Manipulation:

→ We first read the data files Train.csv, Stores.csv and Features.csv in R and convert them into individual dataframes in R.

→ We then merge the 3 files into a single file **`dfTrainMerged`** and remove NA values.

→ We convert the Boolean field IsHoliday to a numeric field 0 or 1.

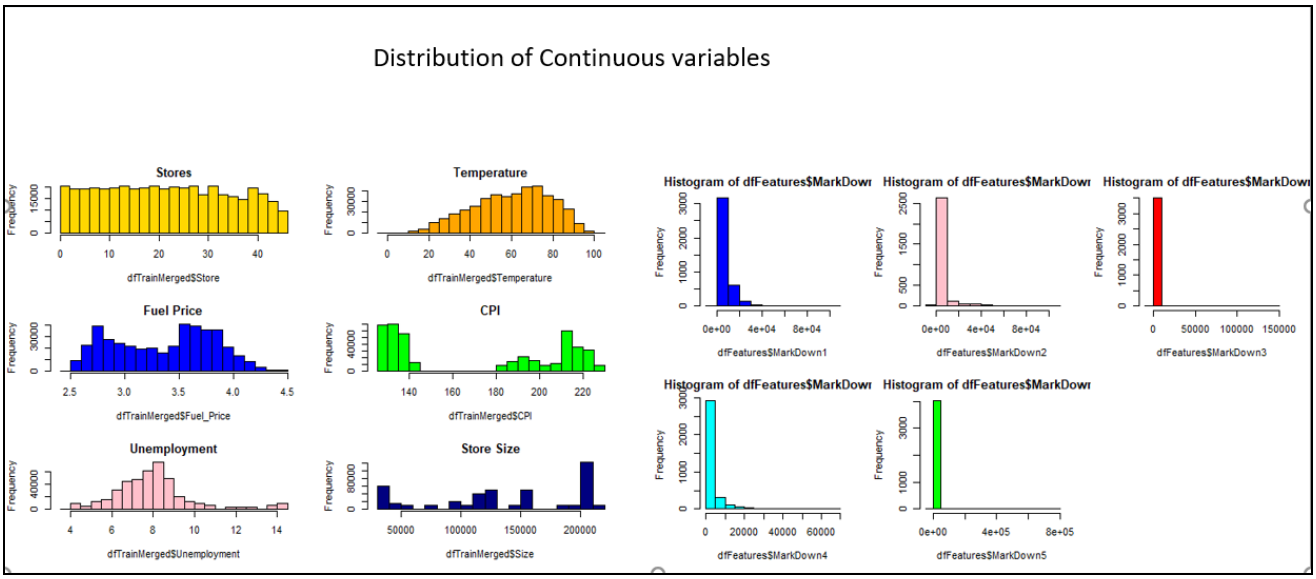
→ We then split the date field into four data columns of Day, Month, Year and Week.

→ The resultant final data frame **`dfTrainMerged1`** has the following data structure.

```r
str(dfTrainMerged1)
```

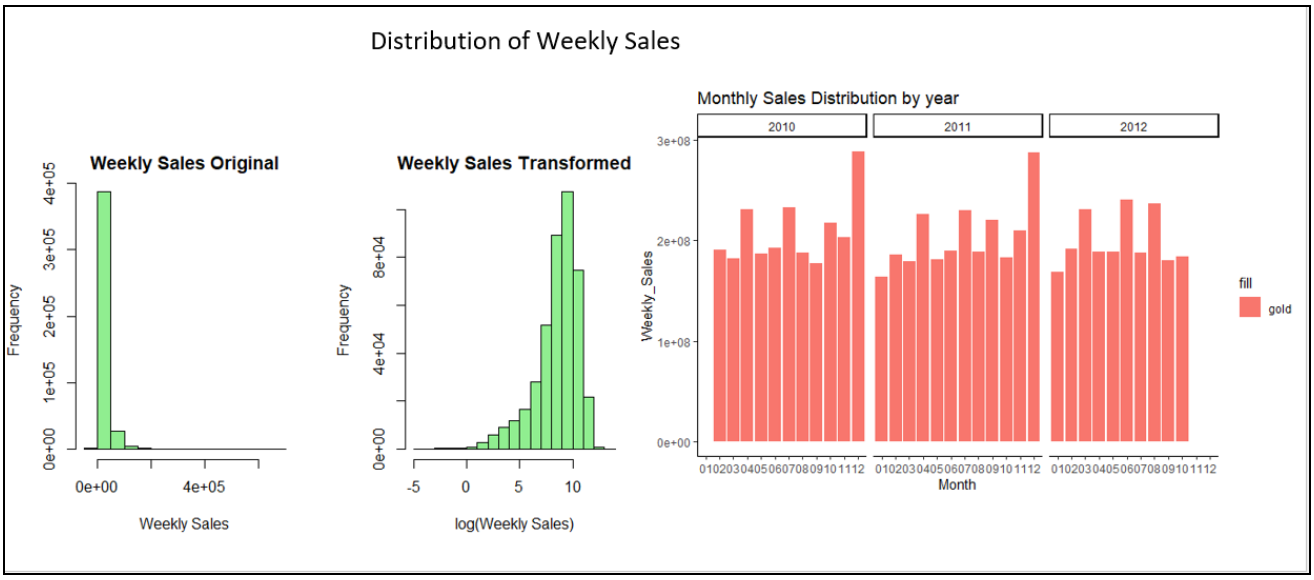
```
'data.frame':   421570 obs. of  19 variables:
 $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Year        : chr  "2012" "2012" "2012" "2012" ...
 $ Month       : chr  "01" "01" "01" "01" ...
 $ Day         : chr  "13" "13" "13" "13" ...
 $ IsHoliday   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Dept        : int  1 25 55 92 21 17 19 42 4 44 ...
 $ Weekly_Sales: num  16894 6384 10418 158155 7091 ...
 $ Type        : chr  "A" "A" "A" "A" ...
 $ Size        : int  151315 151315 151315 151315 151315 151315 151315 151315 151315 151315 151315 ...
 $ Temperature : num  48.5 48.5 48.5 48.5 48.5 ...
 $ Fuel_Price  : num  3.26 3.26 3.26 3.26 3.26 ...
 $ MarkDown1   : num  5183 5183 5183 5183 5183 ...
 $ MarkDown2   : num  8026 8026 8026 8026 8026 ...
 $ MarkDown3   : num  42.2 42.2 42.2 42.2 42.2 ...
 $ MarkDown4   : num  453 453 453 453 453 ...
 $ MarkDown5   : num  3719 3719 3719 3719 3719 ...
 $ CPI         : num  220 220 220 220 220 ...
 $ Unemployment: num  7.35 7.35 7.35 7.35 7.35 ...
 $ Week        : chr  "02" "02" "02" "02" ...
```

Data Visualization:



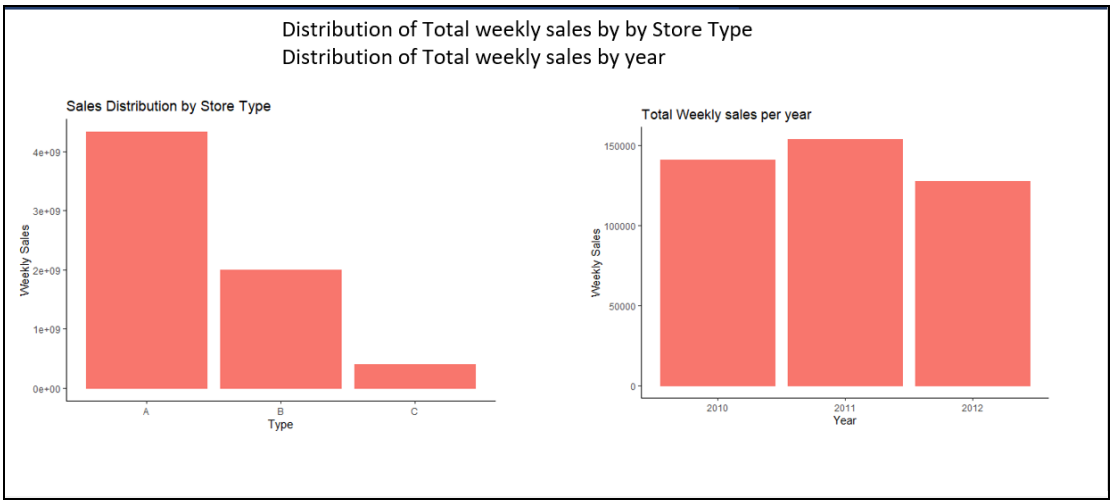
→ The plots of the continuous variables in the data show us that except Temperature, all the other variables are either positively skewed or have uneven distribution or have extreme values.

→ We thus would need to apply log transformation to the above plots to view a more even spread of the data.



→ After applying log transformation, we now see a more evenly distributed spread of the data.

→ Also, upon visualizing the total weekly sales per year, we can observe clear seasonality in the data wherein the data peaks starting October, November, and December for both years 2010 and 2011 owing to the festive shopping during the Thanksgiving and Christmas Holidays. (Note that we have not been provided data for 11/12 and 12/12 since we need to forecast the sales for that time of the year).



→ The distribution of the number of stores by store type shows that the store belonging to format A is highest in number followed by stores of format B and format C.

→ The distribution of the store size by store type shows that the store of format A has the highest store size with the highest number of items sold in that store followed by store type B and C.

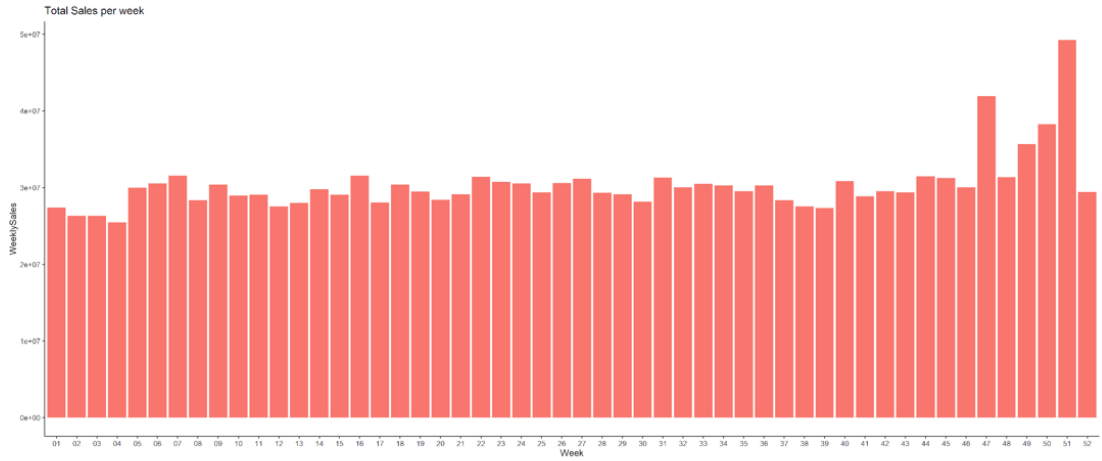
→ The plot of distribution of total weekly sales per week for 52 weeks shows that the weekly sales starts peaking from weeks 47 through 52 showing a clear seasonality in the sales data.

→ The plot of distribution of department wise mean weekly sales shows that department 92 and 95 have high mean weekly sales.

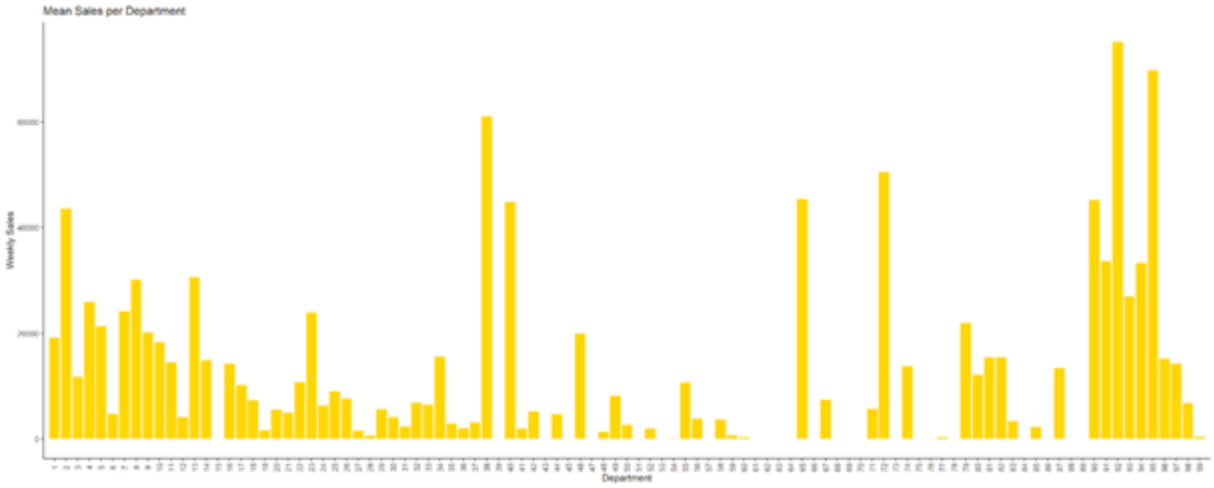
→ The plot of the distribution of weekly sales by store shows that store 20 has the highest weekly sales.

→ The plot of the distribution of weekly sales by day shows that the sales are higher on weekdays compared to weekends and dip significantly on the last day of the month.

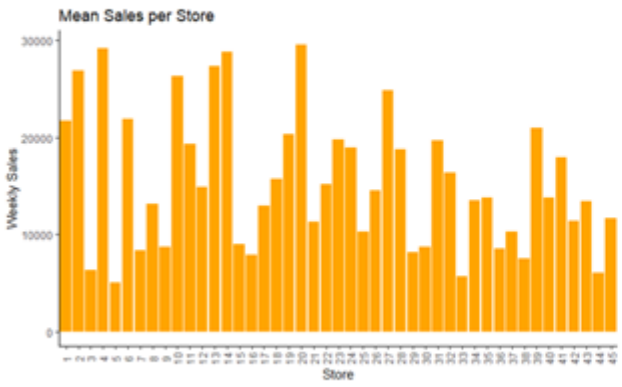
Distribution of Total Sales per week for 52 weeks

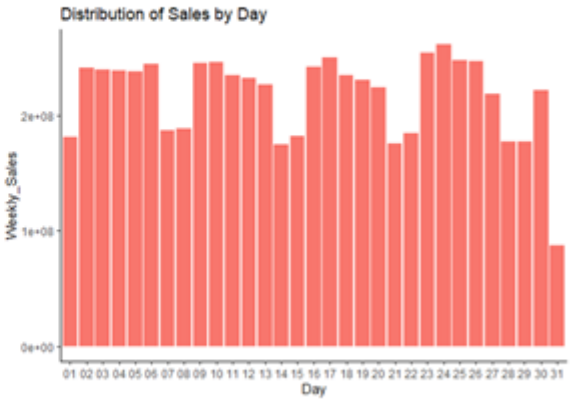


Distribution of Department wise Mean Sales

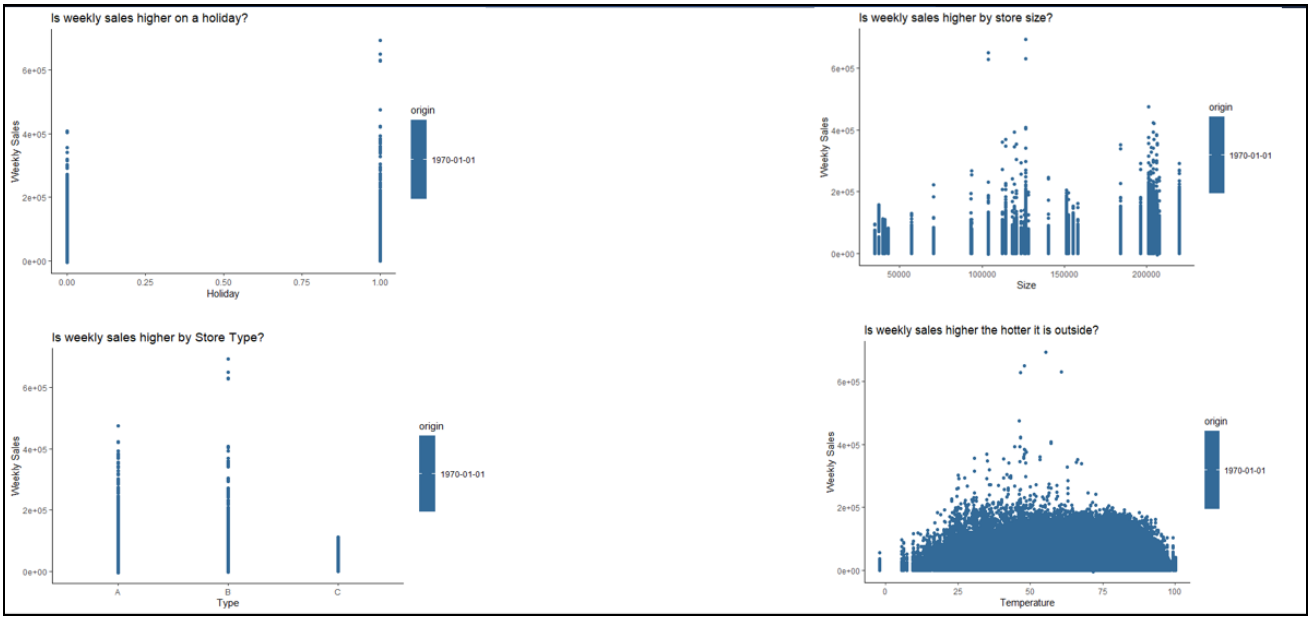


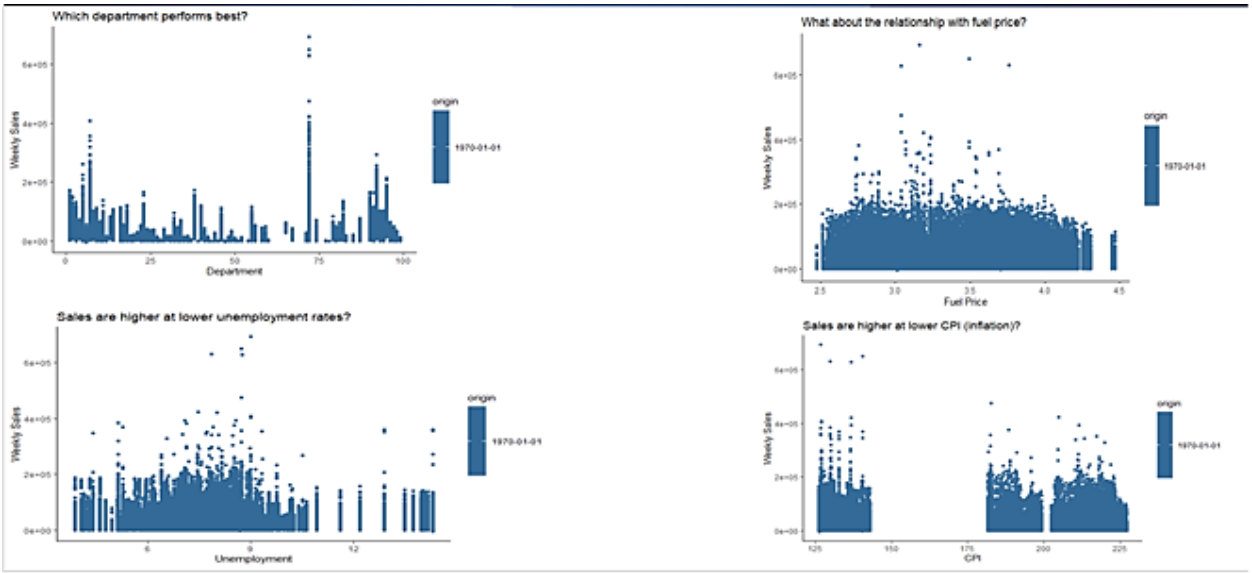
Distribution of Mean weekly sales per store and
Distribution of sales by days of the month





Data Exploration





→ Exploring the data at a more granular level shows us that:

1. Weekly Sales are higher on a holiday than a non-holiday.

2. Weekly sales is higher by store size.

3. Weekly sales is highest for store type A compared to B and C except a few outliers.

4. Weekly sales are highest when the temperature is moderate around 50 -60 degrees.

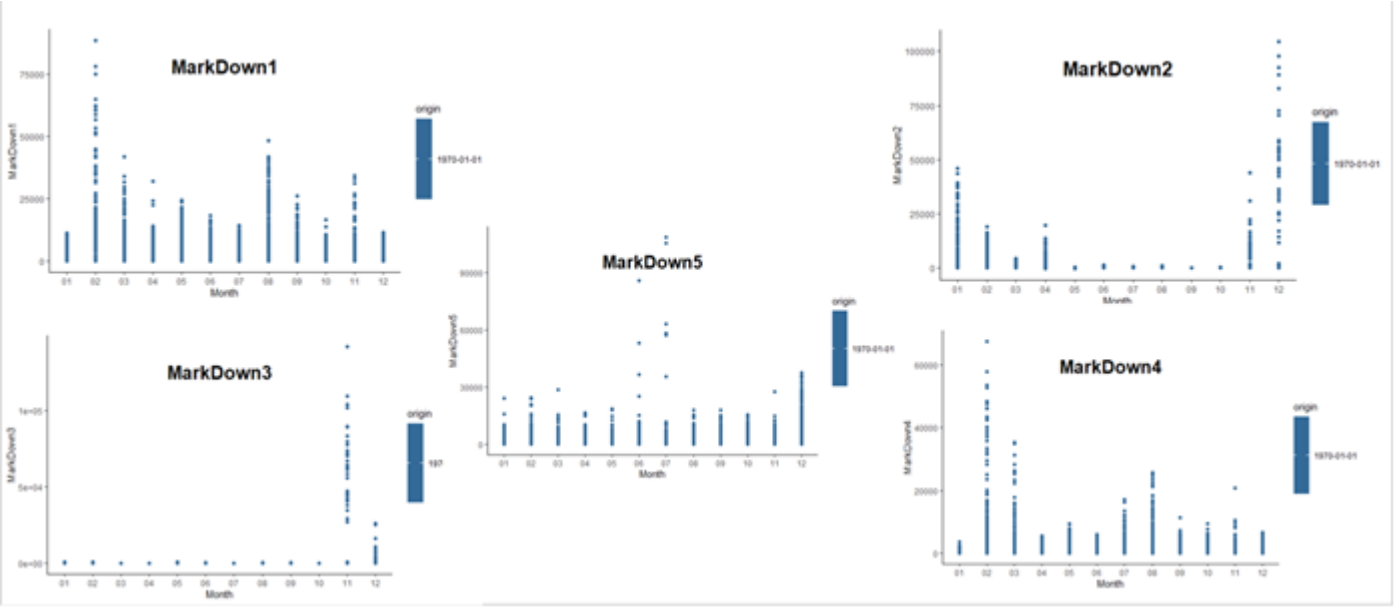
5. Department 75 performs best in terms of weekly sales.

6. Weekly sales are higher when the fuel prices are lower.

7. Weekly sales also reported higher levels when the unemployment was lower.

8. Weekly Sales were higher for lower CPI values.

HOLIDAY AND MARKDOWN ANALYSIS



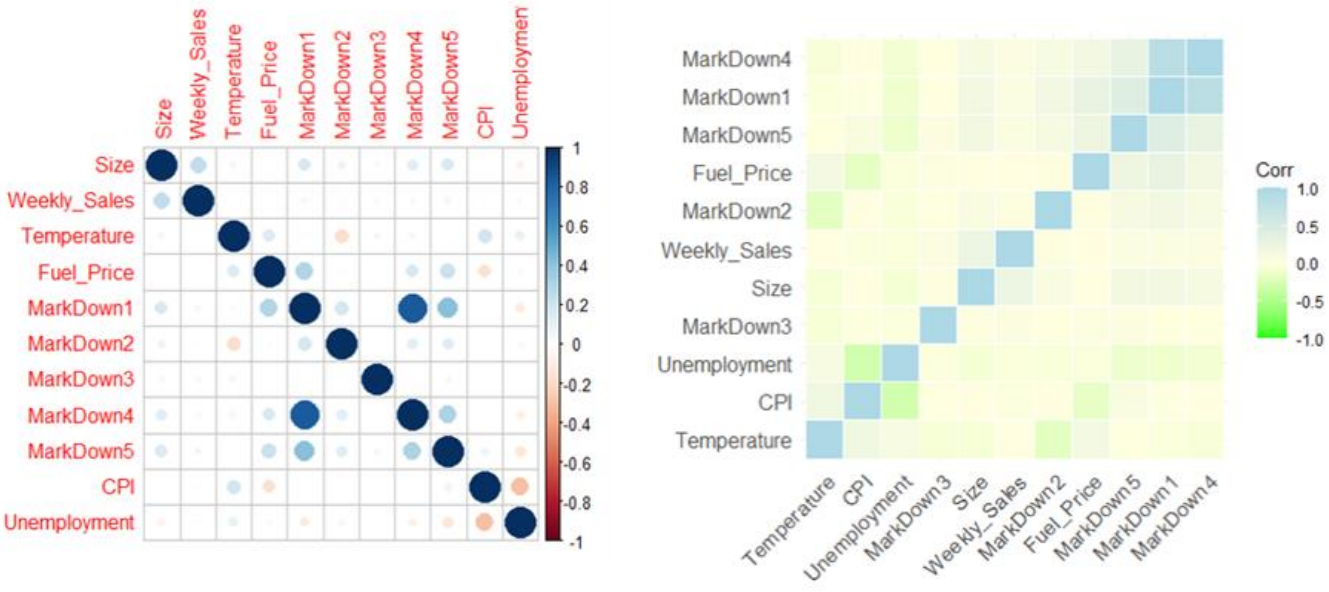
9. Markdown1 and Markdown4 are offered in February during the Super Bowl Holiday, and it shows the sale of around 75,000 – 80,000 units.

10. Markdown 2 is offered in December during the Christmas Holiday, and it shows the sale of around 100,000 units.

11. Markdown3 is offered during the Thanksgiving Holiday, and it shows the maximum sale of more than 1 million units.

12. Markdown 5 is offered during the 4th of July Holiday, and it shows a sale of around 30,000 to 90,000 units.

CORRELATION PLOTS AND HEATMAP:



13. The correlation plot and the correlation heatmaps show that Markdown1 and Markdown 4 are highly positively correlated since both the Markdowns are offered during the Super Bowl Holiday in February.

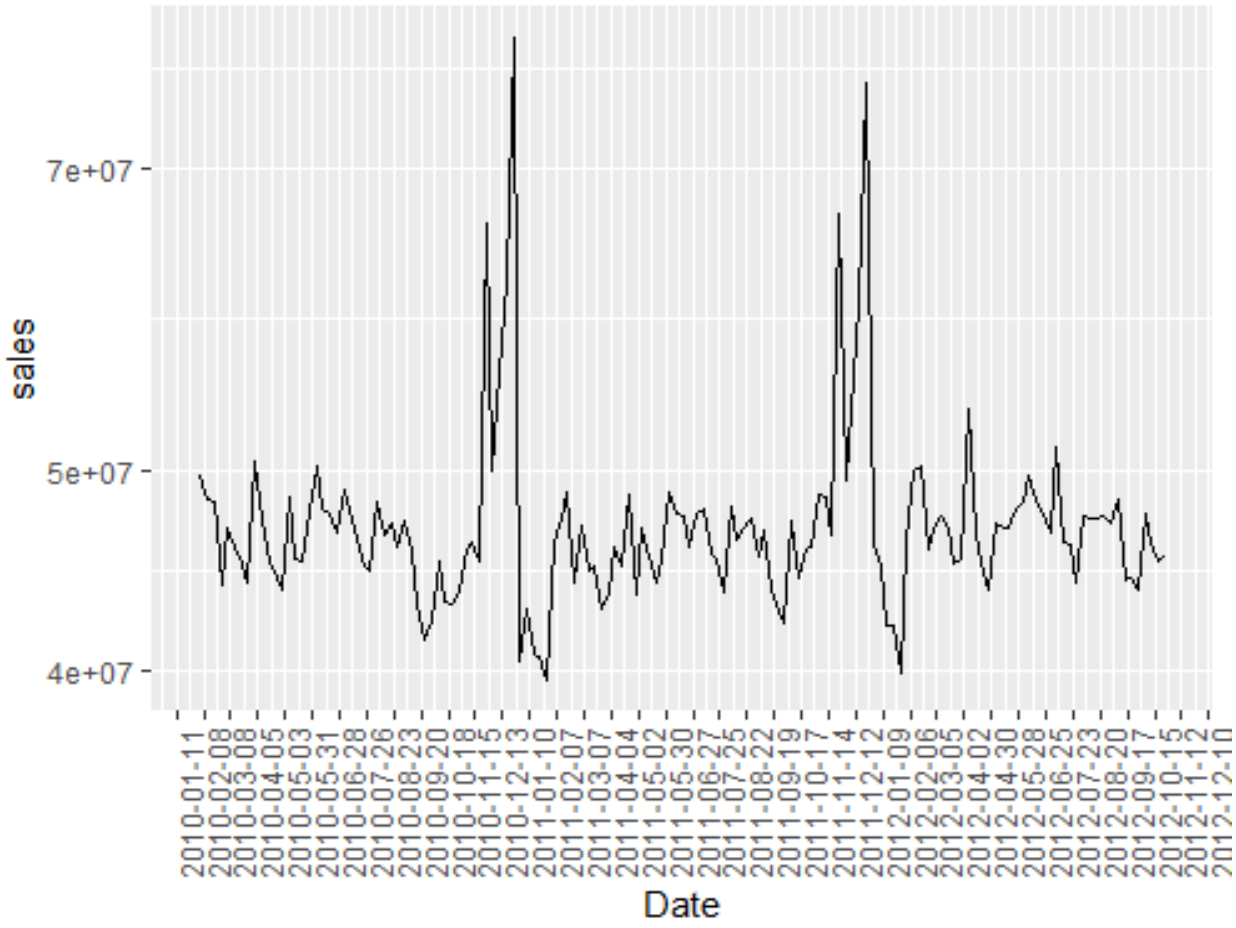
14. Markdown5 and Markdown1 are positively correlated. Markdown5 is offered during the 4th of July which follows the Markdown 1 offered during February and March for the Super Bowl Holiday, which shows that Markdown items unsold during the Markdown1 are further Marked down during the 4th of July holiday to enhance sales.

15. CPI and unemployment are highly negatively correlated since the CPI normally remains low when the unemployment is lower.

EDA, Relationship, Pattern discovery, Sarima Modelling and Sales Forecasting:

Aggregating Weekly Sales Data:

Cleaned, pre-processed, and validated the train dataset to ensure proper format and no missing values.
Next, aggregated and summarized the Walmart Weekly Sales in a plot so that trend, cyclic patterns, and seasonality could be observed.



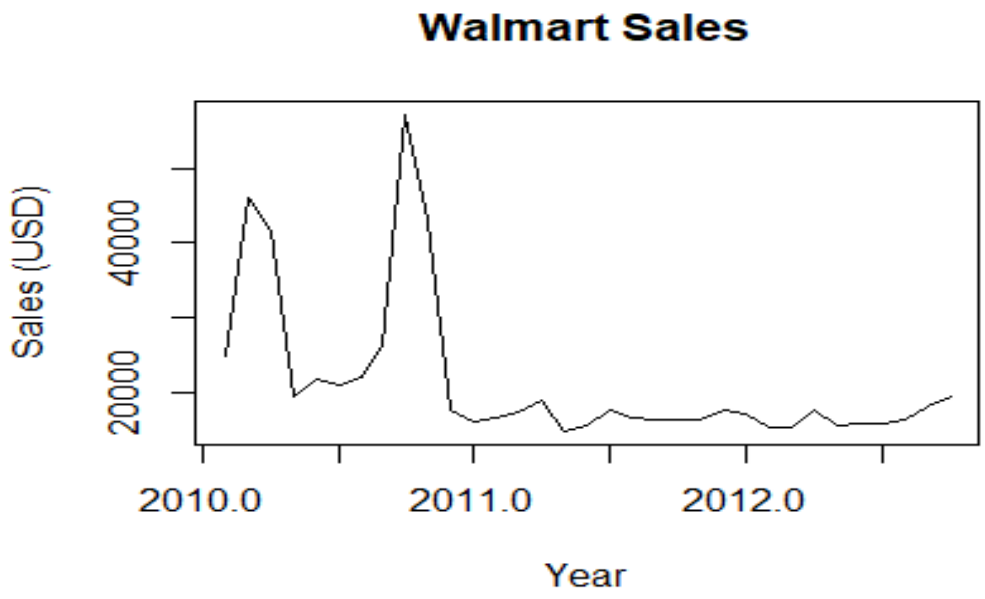
Some significant observations were as follows:
There is a definite seasonality observed with aggregation of the weekly Walmart sales.

- Sales peak maximum in the winter months (November and December) every year. This is the holiday season: festive weeks of Thanksgiving, Christmas, and New Year.
- There are lesser spikes in sales during April-May around Easter or Labor Day. It indicates that these are not popular festivals for retail shopping.

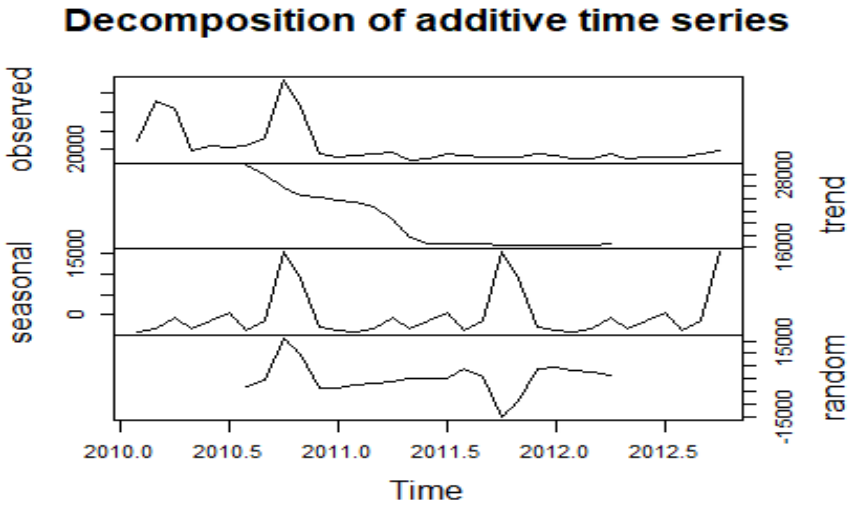
With above observations, it was necessary to plot the time -series and do the transformations to decide further course of action.

Plotting, decomposing the time series and applying Transformations:

Plotted the Walmart Weekly Sales in a time series to have the following observations:

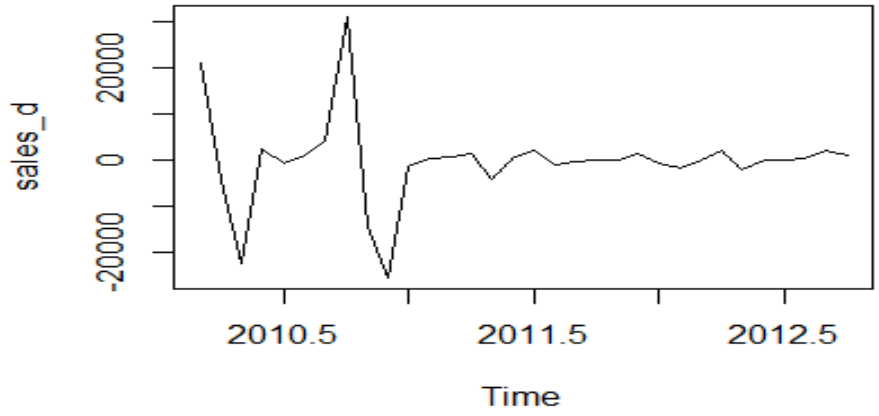


- Sales are the highest towards Dec 2010 and Jan 2011.
- There is a steep drop in sales after that.



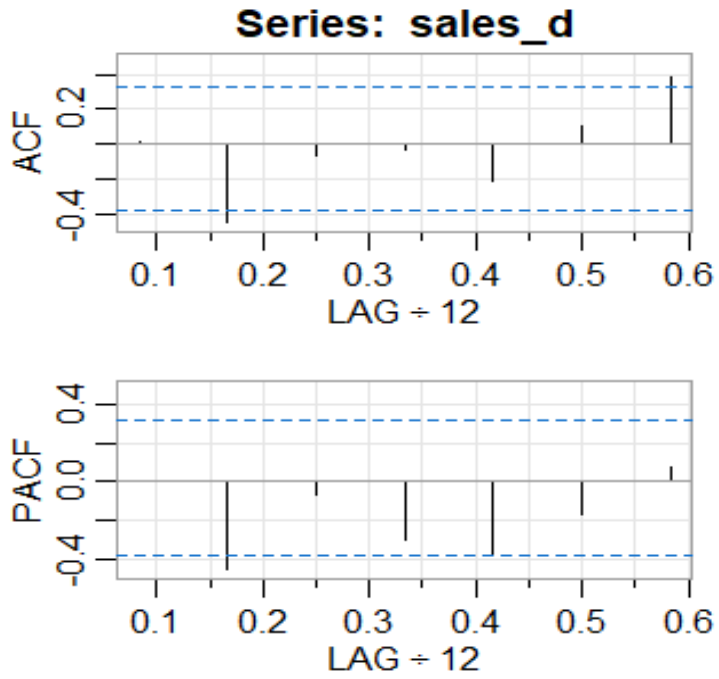
Observations: There is a definite seasonality observed in the data. The trend in Walmart sales decreased rapidly from mid-year 2010 and seems to be constant from mid-year 2011 onwards.

Make the time series stationary, we decided to just apply the difference. Log transformation is not necessary as we must take care of the seasonality.

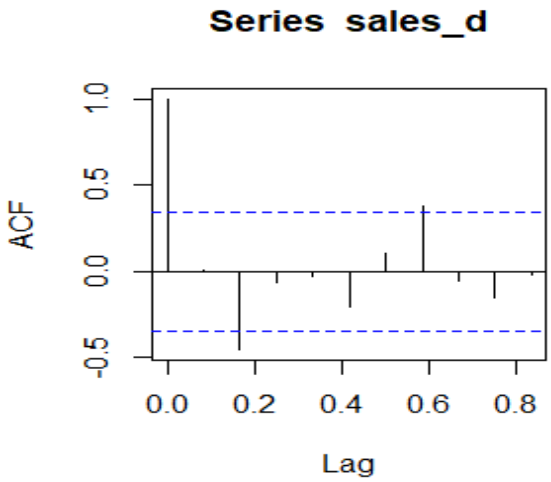


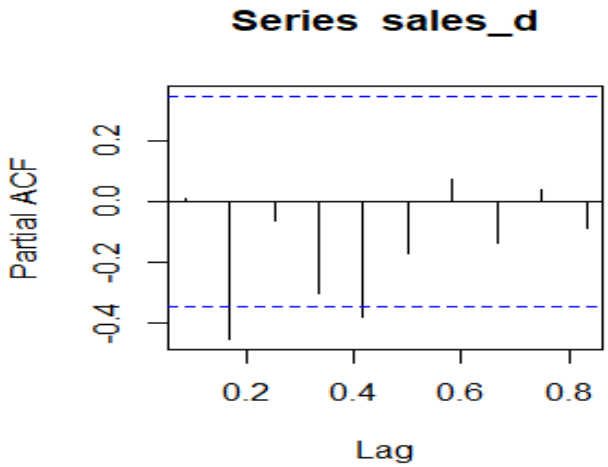
Above plot shows that the time series is stationery since seasonality has been removed. With Dickey-Fuller Test, P value is 0.02 indicating we can reject the null hypothesis 'series is non-stationary'. Thus, the time series is **stationary after the transformation.**

Plotting the ACF and PACF using the differenced series:



(This seems to be a seasonal arima model. The ACF spikes at 2 places, while, the PACF has a spike at lag 1.)



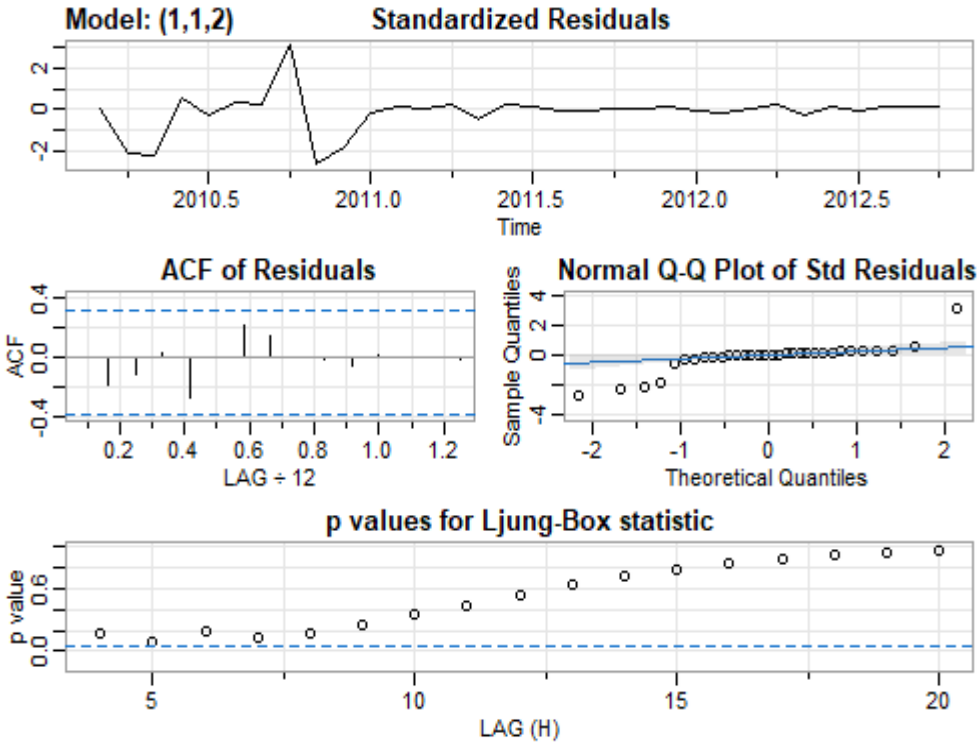


(ACF shows an AR1 model. PACF spikes at 2 positions, indicating an MA (2) model)

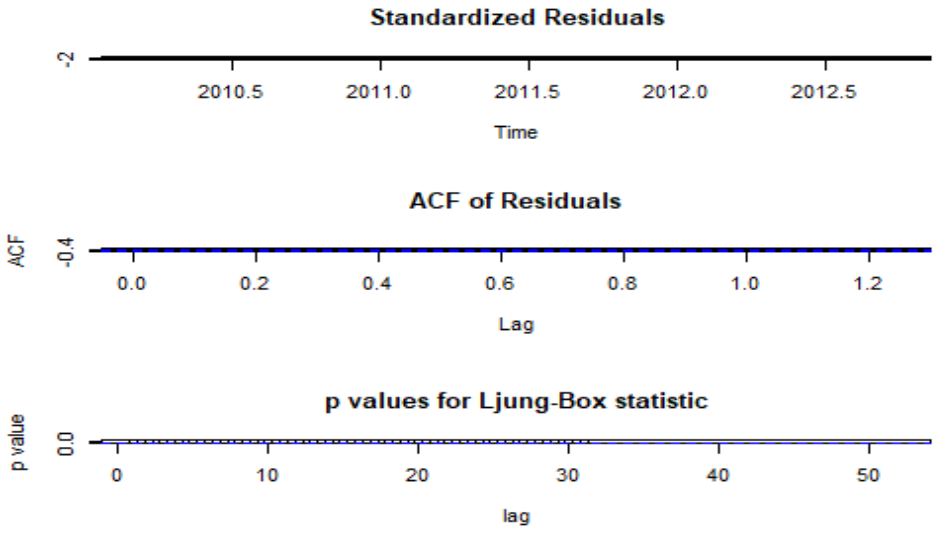
The ACF and PACF suggests and AR1 and MA2 model with a seasonality of 12 months.

Model Fitting with SARIMA (1,1,2) (0,0,1) ^12

```
> fit

Call:
arima(x = sales_d, order = c(1, 1, 2), seasonal = list(order = c(0, 0, 1), period = 12))

Coefficients:
          ar1      ma1      ma2    sma1
      -0.5091  -0.2449  -0.7551  0.1240
s.e.   0.4039   0.3069   0.3025  0.2734

sigma^2 estimated as 84413476:  log likelihood = -328.56,  aic = 667.13
>
```

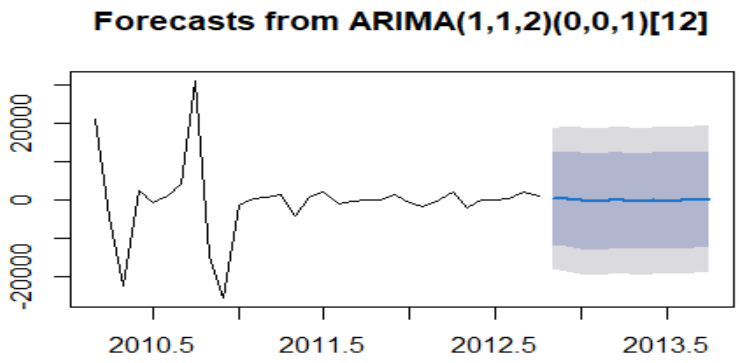


- The Standardized residuals show some pattern until end of 2010 where the sales are highest. They again decrease mid-2011 and then remain constant. Since there is no continuous pattern, it resembles white noise.
- Similarly, ACF also resembles the Standardized Residuals in terms of white noise.
- QQ plot has few outliers in the beginning but otherwise has a good fit
- The P-values are above/equal to the significance level.

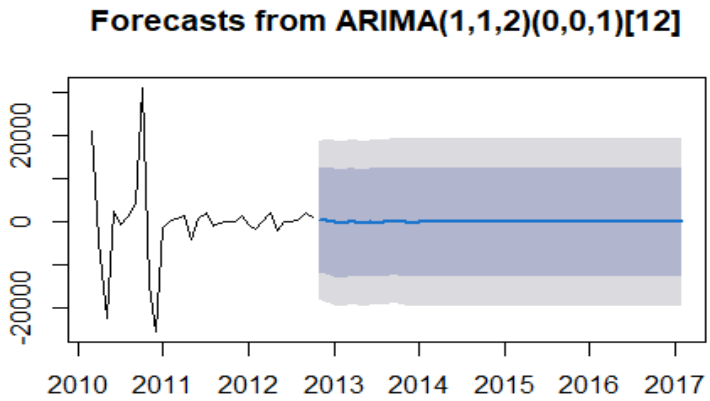


Sales Forecasting for SARIMA:

Transformed Non-stationary series: Yearly

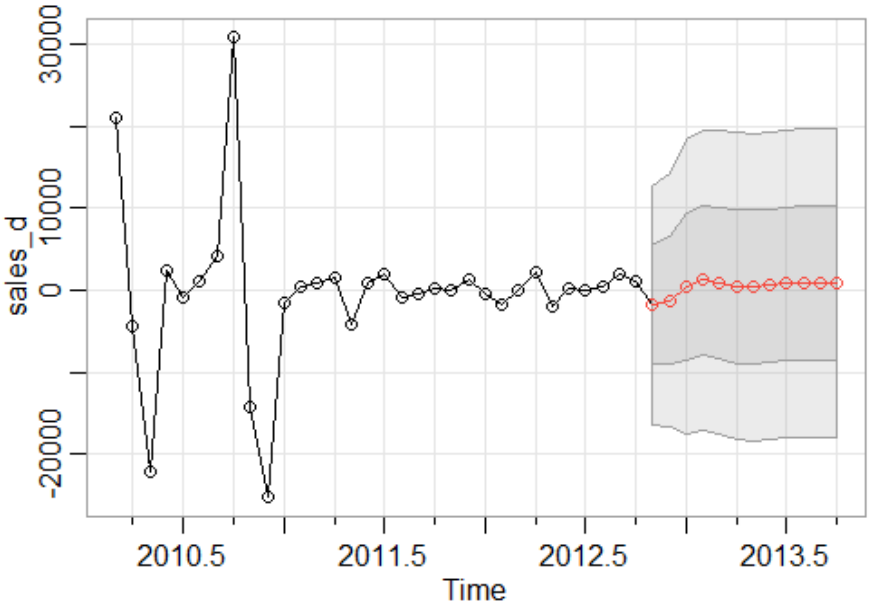


Transformed Non-stationary series: Monthly

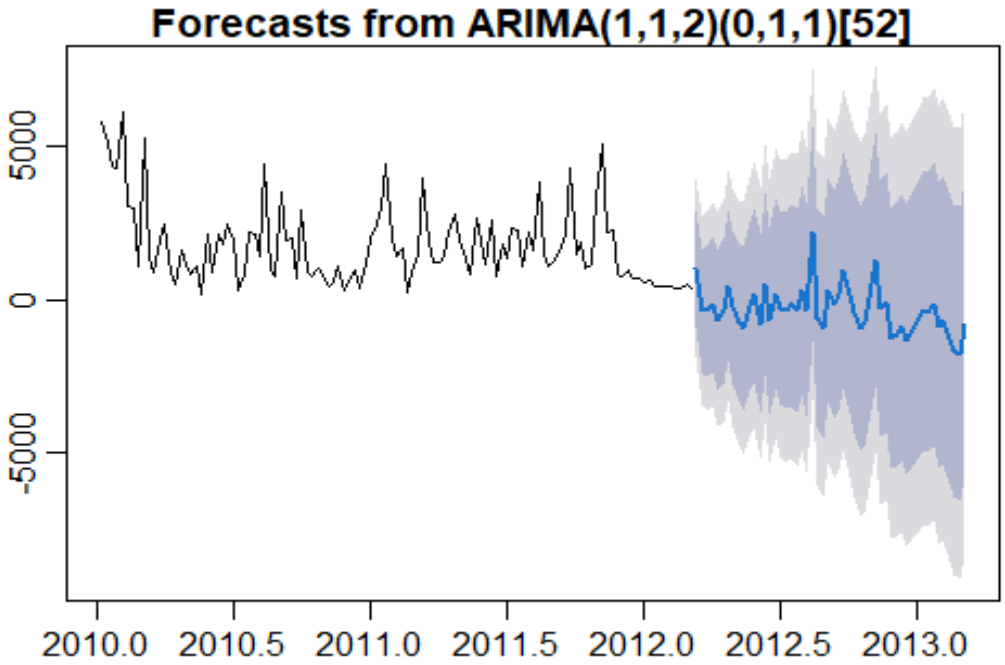


Sales Forecasting on training and test datasets:

- Slice the Walmart training data into train and test datasets by Validation set approach and an 80 – 20 % split.
- Convert the test data into a time series and fit the SARIMA (1,1,2) (0,0,1) ^12 model.
- Predict and plot the sales forecast on the Test data



- The Predicted Walmart Sales indicate that there would be a decline in sales at beginning of 2013 and sales would pick up again Mid-2013.
- However, the Sales Forecasting on SARIMA (1,1,2) indicates that the sales revenue would remain constant over the forecasted years 2013 – 2017.



The Sales forecasts on the Test dataset display both trend and seasonality. This is different from the monthly and yearly Sales forecasts that we had received on the Training dataset.

MSPE calculation:

We now have the training sales forecasts as well as the test forecasts obtained after fitting the model. Thus, it is possible for us to calculate the MSPE or one step ahead prediction error.

The below code snippet gives MSPE calculation which gives a difference of 13 % in training and test Sales forecasts.

```
actual <- c(1036.39229115592, -374.583985770334, -381.963923095491, -165.278632105435
predicted <- c(3893.79402088054, 3692.19766201801, 21079.2277023924, -8471.3860306178
MAPE <- mean(abs((actual - predicted)/actual))
MAPE
# The MAPE value suggests that on average, the forecast is off by ~13%"
```

Hence, we conclude the Walmart Weekly Sales Forecast is off by 13 %.

Alternate Method: Building Model with and without the features

Part A:

As a next step to further enhance or test a new methodology, an alternate way of analysis was evaluated to forecast sales. As we know, from the initial data visualization that the sales of the stores are correlated with the type of store or size of the store i.e Type A, B or C, three new dataset were chosen corresponding to each store type. This helped us to manage the correlation between the stores with this step. This helped us to enhance our current model based on sales aggregated by date overall

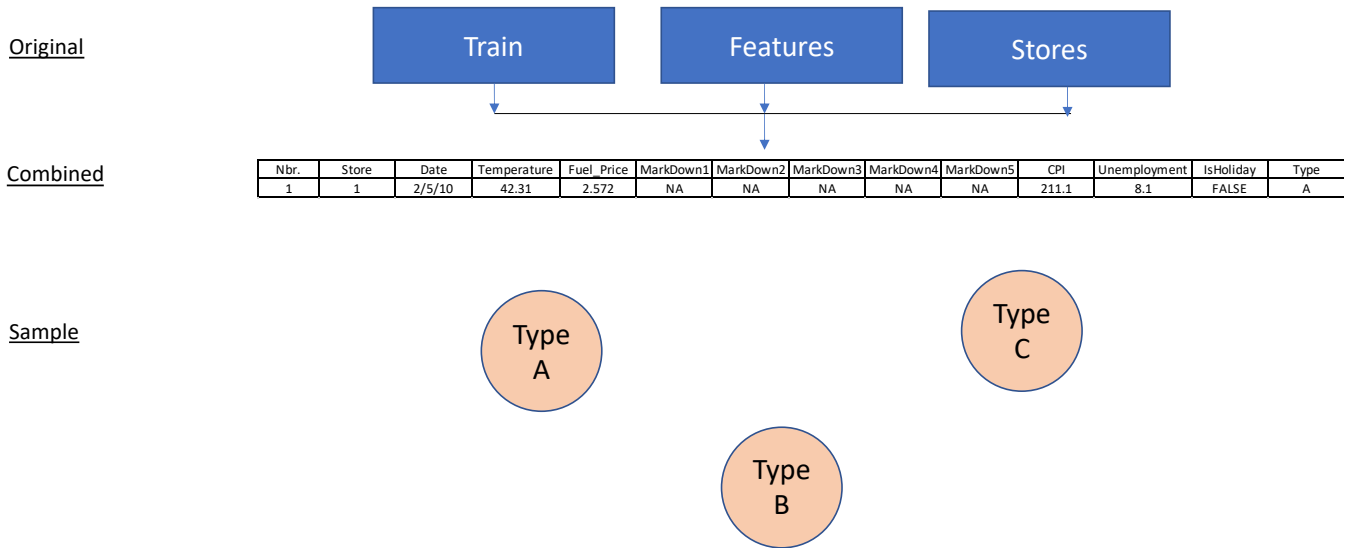
I. Dataset for Modeling:

Originally, we were provided with three datasets (Train, Features and Stores). Merged these datasets by Store and Date as key to arrive at the final merged dataset which we used as a base for modeling.

In the second step, selected one store from each of the store type randomly and used these samples for analysis of each of the three store types.

In the third step divided the dataset into 80:20 split. We ended with 143 observations for each store type.

Please see below for detailed description



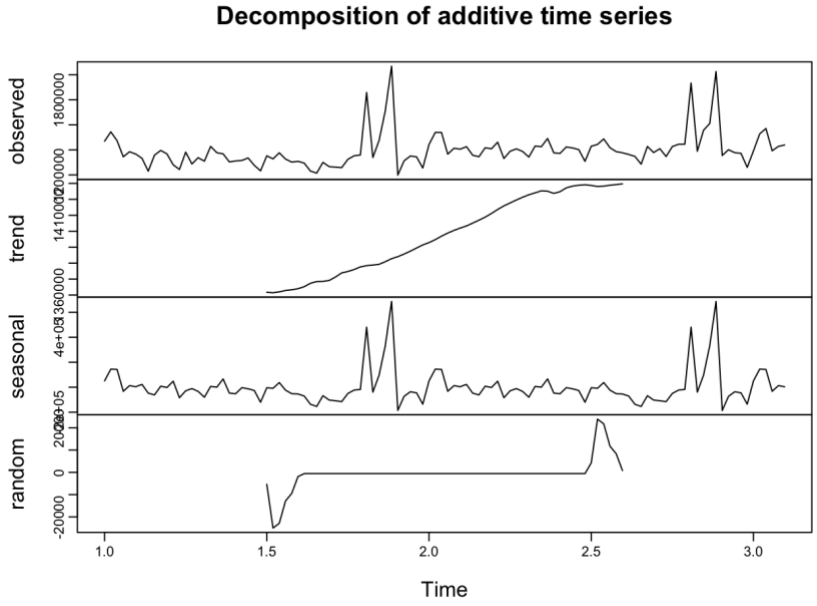
II. Model Building Process

Similar procedure as outlined in previous sections was leveraged to do the model diagnosis and the select the final mode. Model diagnosis with different parameters were implemented to analyze p values, correlations, residuals, and lag cut offs to detect the best model. We attempted different SARIMA models as shown in Figure below. The first step in building the model for all the datasets was as follows

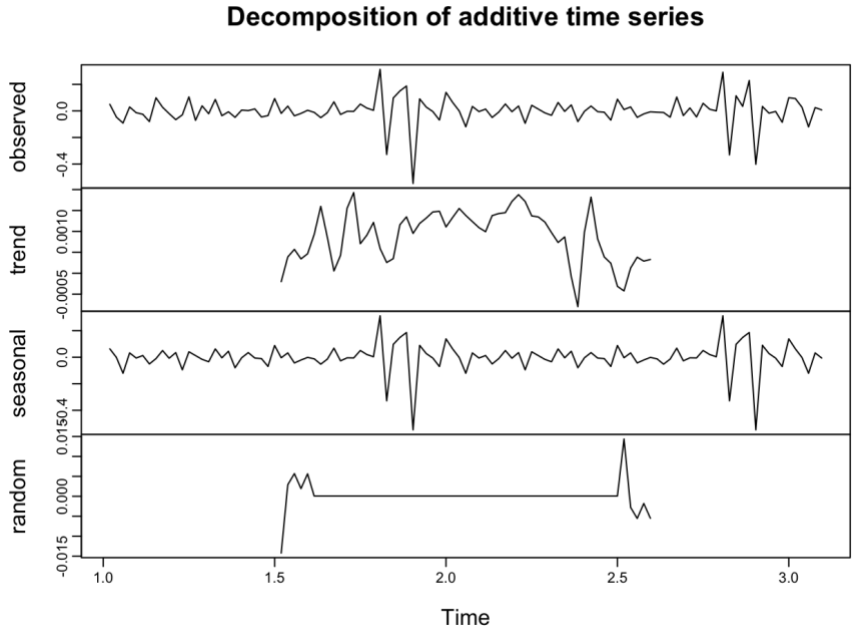
A) Split data to train and test
B) Check acf and pacf for the store data and find right models
C) Convert to time series data
D) Looked at the various components of timeseries
E) Observed some seasonality and decreasing trend
F) Removed trend in timeseries
G) Used transformed data to find appropriate model

For the models where we can see that the p values go below the line after lag, that meant the residuals are correlated while in Models where all the p values are above the line, which means the residuals are independent. In the Model C we see the p values go below the line at lag 14 and lag 15, which means the residuals are correlated at lag 14 and lag 15.

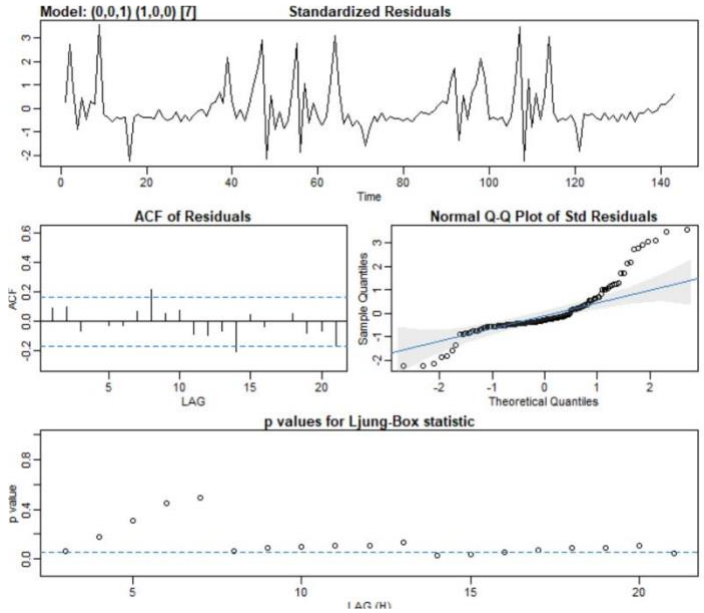
Type A : Step 1



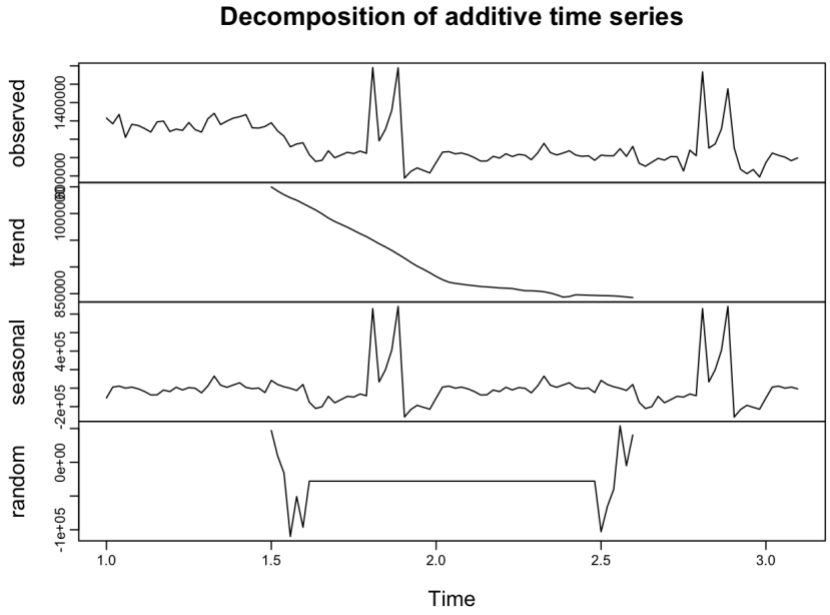
Type A: Step 2



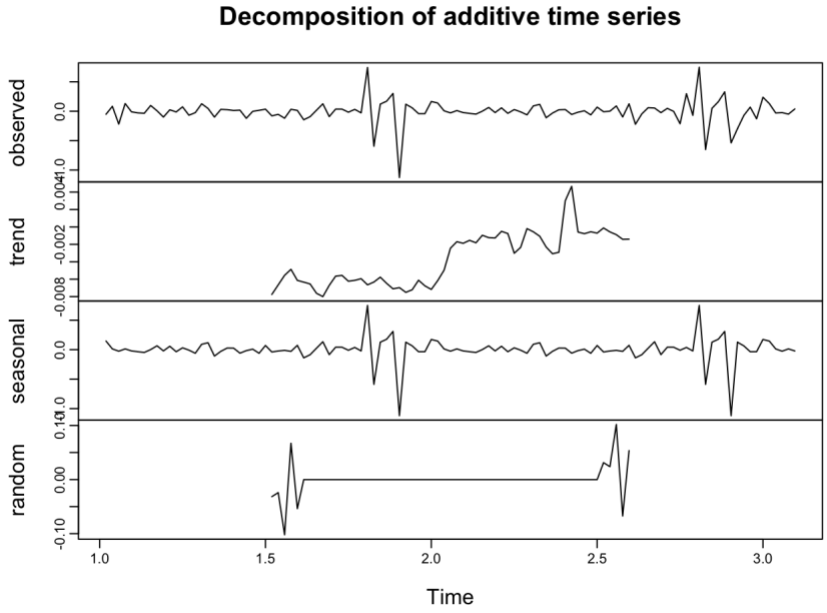
Type A: Step 3


Model: (0,0,1) (1,0,0) [7]
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value

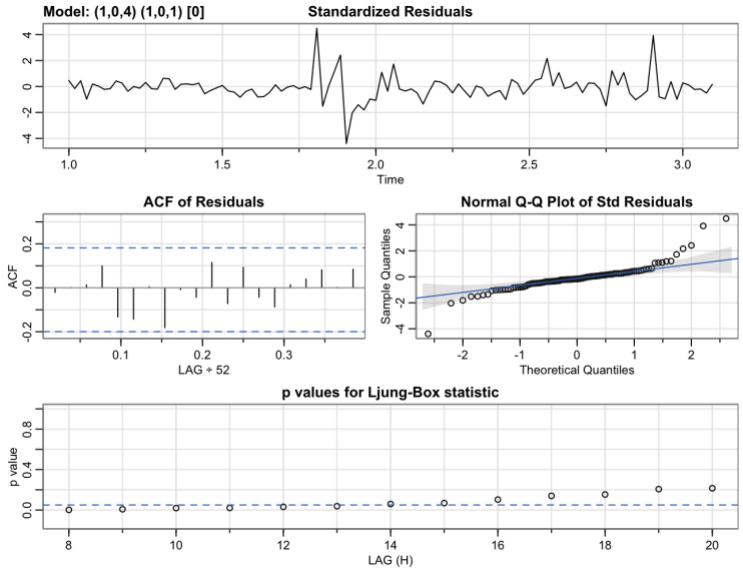
Type B: Step 1


Decomposition of additive time series
observed
trend
seasonal
random
Time

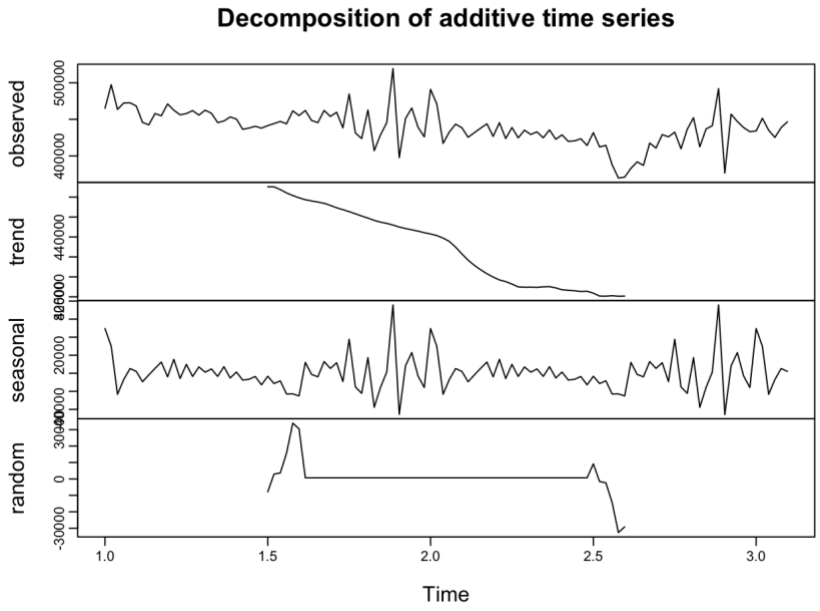
Type B: Step 2



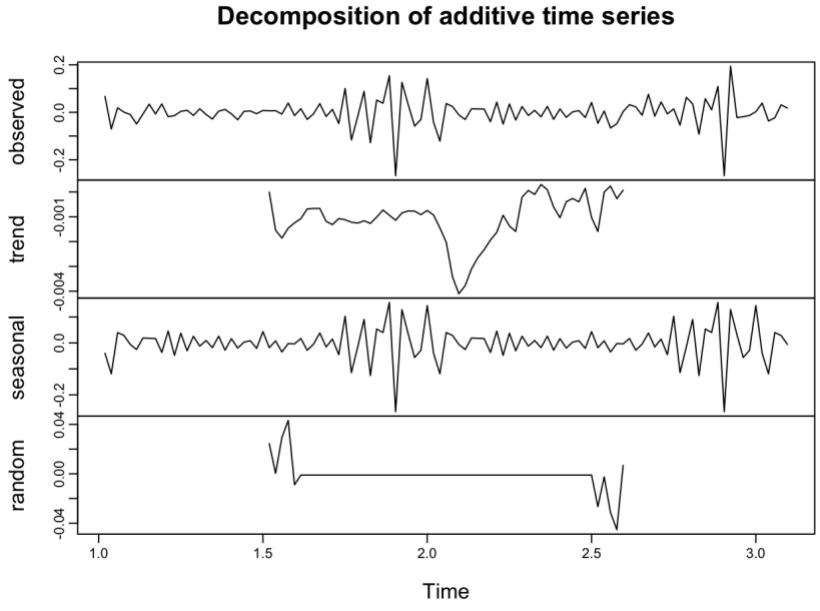
Type B: Step 3



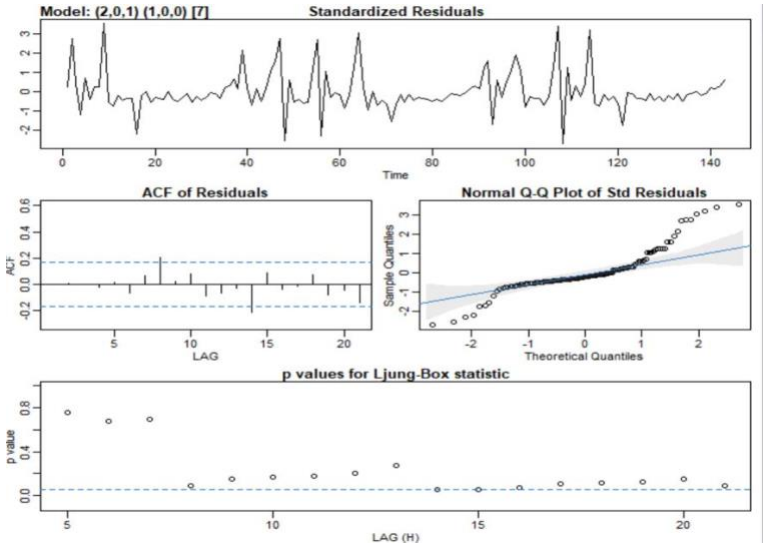
Type C: Step 1


Decomposition of additive time series
observed
trend
seasonal
random
Time

Type C: Step 2


Decomposition of additive time series
observed
trend
seasonal
random
Time

Type C: Step 3



Part B) SARIMA with additional variables

The second model that was implemented was a SARIMA with additional variables. This dataset was used to incorporate explanatory variables such as unemployment, gasoline prices, and CPI in the initial time series. **As, shown in the previous analysis and also based on the p values in the linear regression, we saw that the trend of temperature over time has a slight increase but with a slightly different variance, which supports no need for logistic** transformation or differencing. The trend for sales, as analyzed before, showed no distinct trend with seasonality.

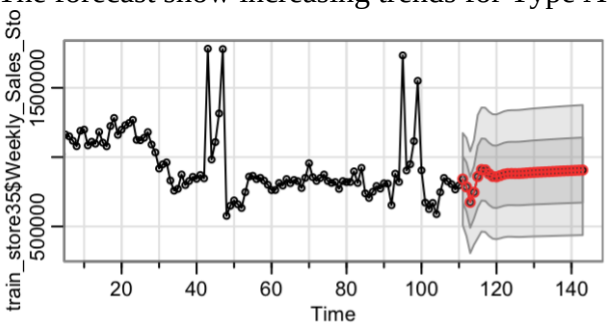
We leveraged the plotted correlation plot earlier o visualize the relationship between temperature and sales. Additionally, the ACF plot (right) showed the correlation function when both these variables are combined. We can see from the correlation plot (left) that there is no clear relationship between the two variables. The correlation implies a weak to moderate, negative relationship. When we analyzed the ACF plot, we saw a seasonality and a high negative correlation at lag 0. Since there was a cut off around the eighth lag, a model with an AR parameter of seven or eight seemed like a potential fit. In addition, since the PACF plot in Figure 8 shows a cut of at 2, we used a MA parameter of 2 for our model.

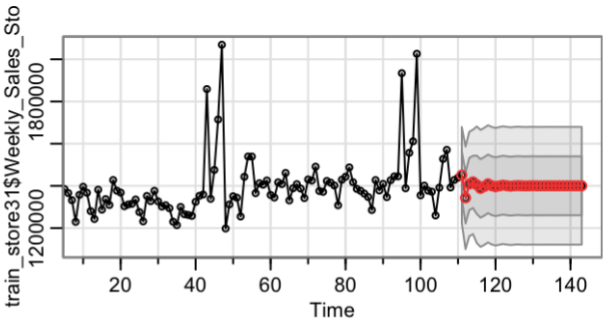
We used linear regression along with the sarima model and came up with the final MSPE. The model with additional variables along with Temperature showed better results especially in Type A which had higher sales.

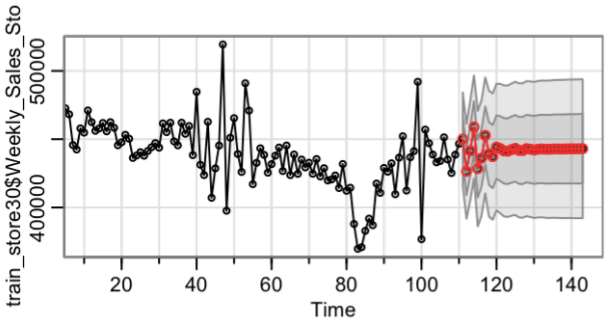
	Sarima Without Features	Sarima With Variables (with Linear regression)
Type A	0.118	0.105
Type B	0.56	0.45
Type C	0.075	0.11

- Temperature and Unemployment were significant features for the model
- Tried Linear regression

The forecast show increasing trends for Type A



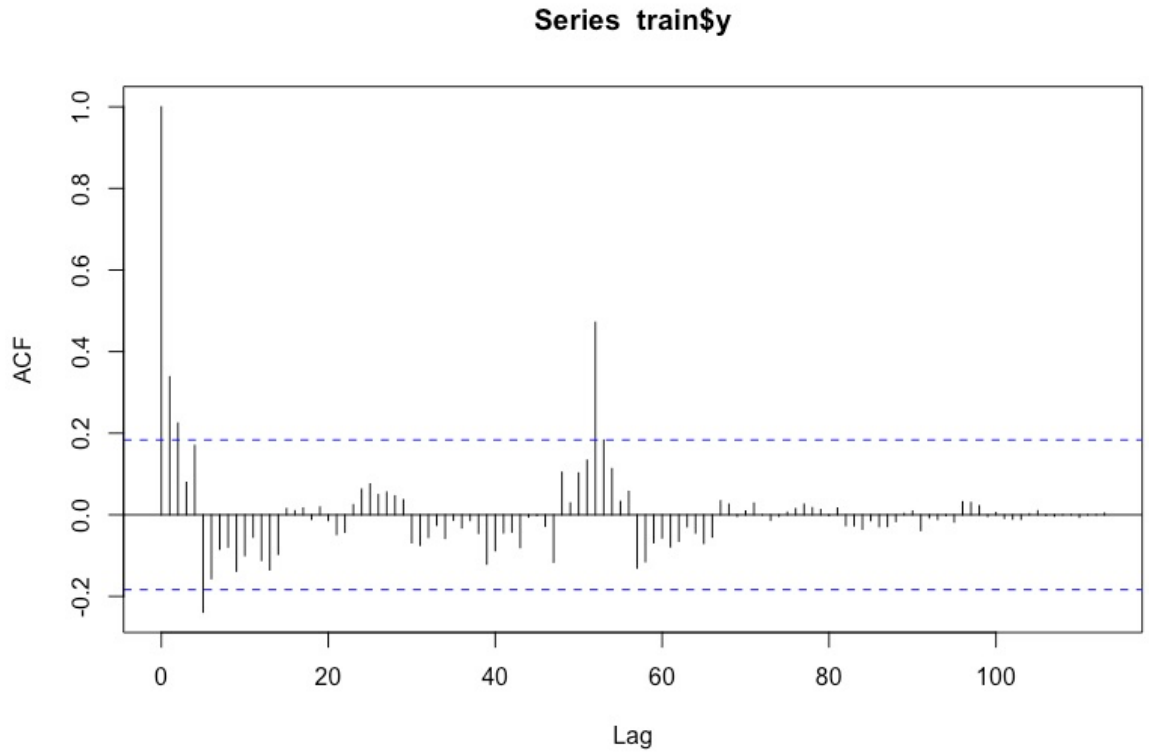




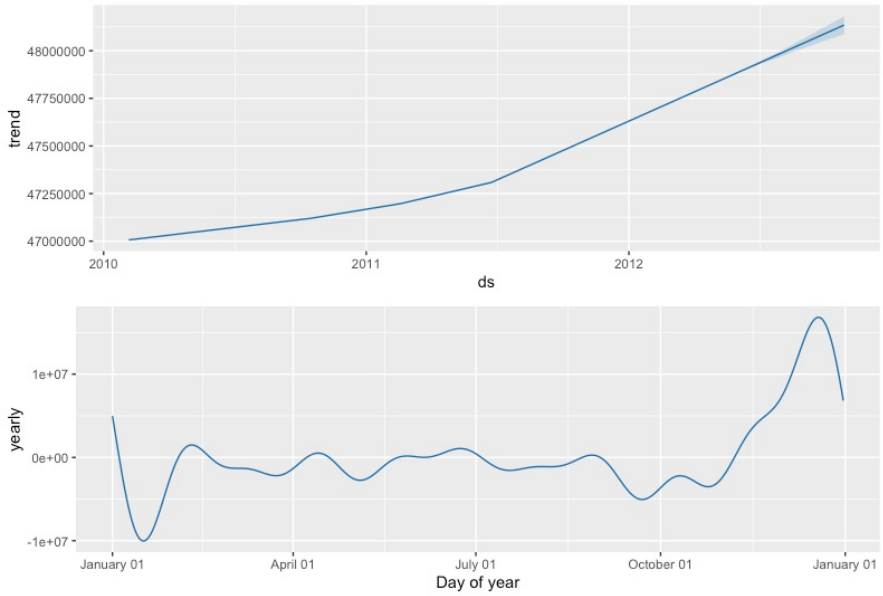
Prophet Model and Sales Forecasting

The last model was used with Walmart Sales Data is the Prophet model. There are several reasons for using this model. First, the Prophet model is suitable for data with solid seasonality, and the Sales data has extreme seasonality based on the analysis from the previous part. Second, the Prophet model helps identify and state change points in the time series. These changepoints can either be manually configured or automatically detected. The experiment found that manually defining the number of changing points helped increase the model's accuracy. Otherwise, the model might be overfitting.

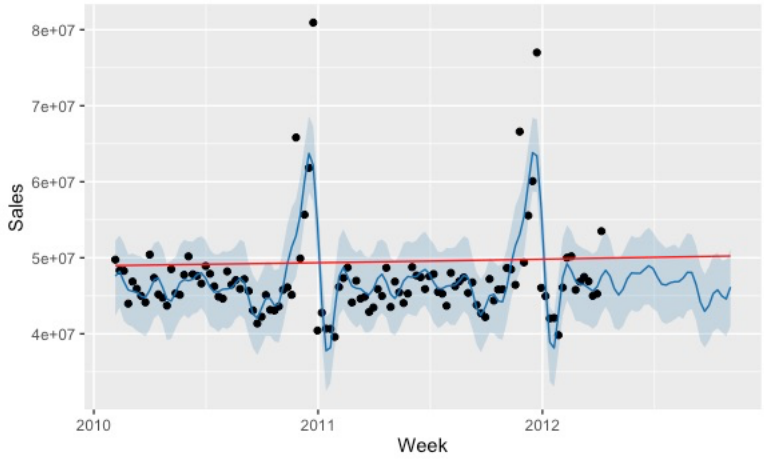
Determine the original Prophet Model:



From the ACF plot, there are five significant changepoints that should be included in the Prophet model.



The above image shows the components of the Prophet model when the prediction is applied. A clear linear trend could be seen from the data. The yearly residual trends also show no definite pattern, which confirms the linearity of the time series.

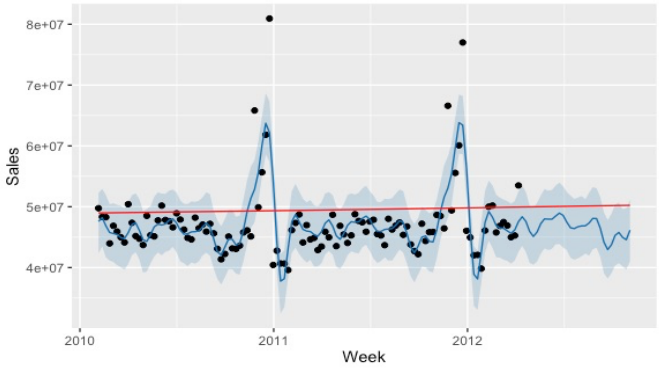


As shown by the plot above, the sales predictions for the last 30 weeks are within the confidence bands. The trend is also consistent with the trend implied in the training data. The MSPE for the Prophet model is 2.357449%. This result is impressive for fitting and predicting a model.

Prophet Model with different parameters:

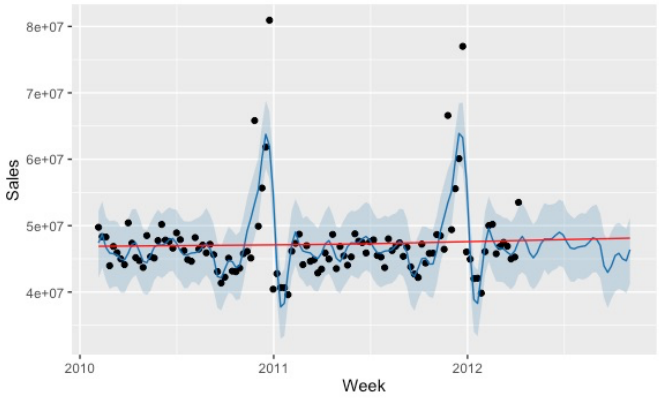
Moreover, many different parameters were experienced to illustrate the influences of those parameters to the Prophet model, determined by the change in MSPE.

Prophet model with weekly seasonality



- Introduced weekly seasonality from the original model
- Experience the change in MSPE. For example, the model MSPE is 2.383013%, which slightly increased from the original model.
- The increasing of MSPE result in rejecting the weekly seasonality as a parameter for Prophet model.

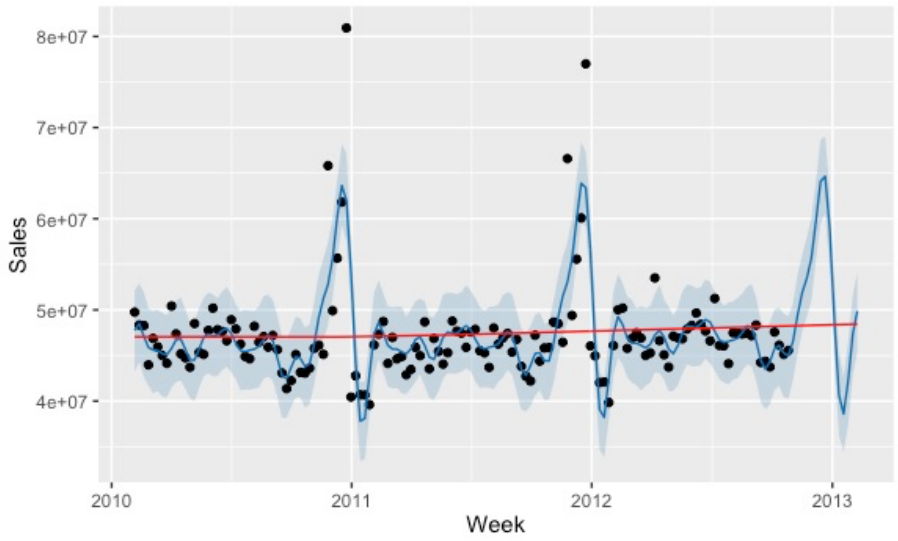
Prophet Model with Holiday



- Introduced holiday for the training model.
- This model MSPE is 2.357862%, which increased from the original model.
- The increasing of MSPE suggested that the Prophet model should not include holiday as a parameter.

Forecasting with Prophet Model:

Prophet model was also used to forecast the additional 14 weeks from the original data.



In conclusion, the Prophet model did a great job fitting and forecasting Walmart's Weekly Sales. The MSPEs are low for all experienced models, and the model potentially can capture the trend and seasonality of the data.