**FORECASTING SALES IN GLOBAL SUPERSTORES**

**FORECASTING  
GLOBAL SUPERSTORE   
SALES**

**FORECASTING  
GLOBAL SUPERSTORE   
TEAM-1**

**Kanchan Sharma**

**Geethasree Madiraju Nagaraju**

**Pothineni LahariPriya**

**Zhexin Zhang**

**Roshita Seelam**

**Table of Contents**

[**1. Problem Statement:**](#_41w791nf72wf) **3**

[**2. Data Description:**](#_b7wwxtjl4asu) **3**

[**3. Data Aggregation**](#_uj7i7dtf2o4d) **4**

[**4. Decomposition Analysis**](#_73cpksj53nl5) **5**

[**5. Correlation Analysis**](#_o1jm35nueoeu) **6**

[**6. Test for Stationarity**](#_wyomew2yly4v) **7**

[**7. Models**](#_a9n067sxfrmq) **7**

[Exponential Additive, Multiplicative models -](#_exqg2magncp6) 7

[a. Winters Multiplicative Exponential Smoothing:](#_i7ky8s5iev5h) 8

[b. Winters Exponential Additive:](#_3iink3mqn4fz) 9

[8. ARMA Models](#_dtypih3eq652) 11

[a. ARMA model results - (0,1)](#_wezx853fnf5m) 12

[b. ARMA model results - (1,1)](#_wdo2iqgirb2) 12

[9. ARIMA Models](#_57qodxnyvyqa) 13

[a. ARIMA Model results - (1,1,0)](#_k8azp317tyzy) 13

[b. ARIMA Model results - (1, 1, 1)](#_r00xfxpm7237) 14

[c. ARIMA Model results - (0, 1, 1)](#_c3jae8ope3qw) 15

[10. Seasonal ARIMA Models](#_qwlj2oiny0gl) 16

[a. Seasonal ARIMA model results ((1,1,0),(1,0,0))](#_mznbv12iku0s) 16

[b. Seasonal ARIMA (1,1,1)(1,0,0)](#_8byoicf8n3ft) 18

[c. Seasonal ARIMA model (1,1,1) (0,0,1)](#_l1f6y1fe7sos) 19

[d. Seasonal ARIMA model (0,1,1, 1,0,0) :](#_cdot0f99rga) 20

[**11. Modeling with Independent variables:**](#_8b8rknb4rna1) **21**

[**12. Cross Correlation Plots**](#_q56kc8sjjgo7) **21**

[a. Cross Correlation plots with isHoliday](#_mrzh6wfz0sr) 21

[b. Cross Correlation with Discount(avg)](#_r9603qypa4f4) 22

[**13. Pre Whitening**](#_w0z4dz3vgecn) **22**

[a. Pre-Whitening on Holiday data :](#_bx8n04f7bdys) 22

[b. Seasonal ARIMAX (with holiday) - (1, 12)](#_lksowwimhx6l) 24

[c. Seasonal ARIMAX (with holiday) - (1, 8)](#_9544s2z6cga) 26

[d. Seasonal ARIMAX (with holiday) - lag 12](#_8hkp942zx20u) 27

[e. PreWhitening on Discount\_avg](#_5s5hri5g19uo) 27

[**14. Model Comparison**](#_kujgzdlrxn1) **29**

[**15. Conclusion**](#_fyjosamsq0h7) **30**

#### 

#### **Problem Statement:**

In recent years, many offline retailers such as superstores and shopping malls have gradually declined in sales and the growth of visits is negative. The main reasons for this circumstance are the growth of online shopping,the economic downturn, changes in consumer preferences, and the COVID-19 epidemic. Superstores need to improve their sales and reduce costs and inventory waste to address those issues. The purpose of this project is developing a time series forecasting model to forecast product sales (demands) for the next year. We will also provide recommendations for superstores based on the results of the model and predictions.

#### **Data Description**:

**Dataset Source:** <https://www.kaggle.com/datasets/shekpaul/global-superstore?resource=download>

The superstore data set has 24 columns in the Orders table which is our primary area of focus, and the columns are listed below:

1. **Row ID:** A unique identifier for each row in the dataset, often used for reference or indexing purposes.
2. **Order ID:** A unique identifier for each customer order.
3. **Order Date:** The date when the customer's order was placed.
4. **Ship Date**:The date when the ordered products were shipped to the customer.
5. **Ship Mode**: The method or mode used for shipping the products (here it is categorized as first class, second class standard class & same day shipments).
6. **Customer ID:** A unique identifier for each customer.
7. **Customer Name:** The name of the customer who placed the order.
8. **Segment:** The segment or category to which the customer belongs (e.g., consumer, corporate, home office, etc.).
9. **City:** The city where the customer is located.
10. **State:** The state or province where the customer is located.
11. **Country:** The country where the customer is located.
12. **Postal Code:** The postal or ZIP code of the customer's location.
13. **Market:** The market in which the order was placed (e.g., US, EU, Africa etc.).
14. **Region:** The region or geographic area within the market where the order was placed.
15. **Product ID:** A unique identifier for each product in the order.
16. **Category:** The broad category to which the product belongs (e.g., office supplies, furniture, technology, etc.).
17. **Sub-Category:** A more specific sub-category within the broader category (e.g., chairs, phones, paper, etc.).
18. **Product Name:** The name or description of the specific product.
19. **Sales:** The total sales revenue generated by the product in the order.
20. **Quantity:** The quantity of the product ordered.
21. **Discount:** The discount applied to the product, if any.
22. **Profit:** The profit generated from selling the product in the order.
23. **Shipping Cost:** The cost incurred for shipping the product to the customer.
24. **Order Priority:** The priority or level of importance assigned to the order (e.g., high, medium, low & critical).

#### **Data Aggregation**

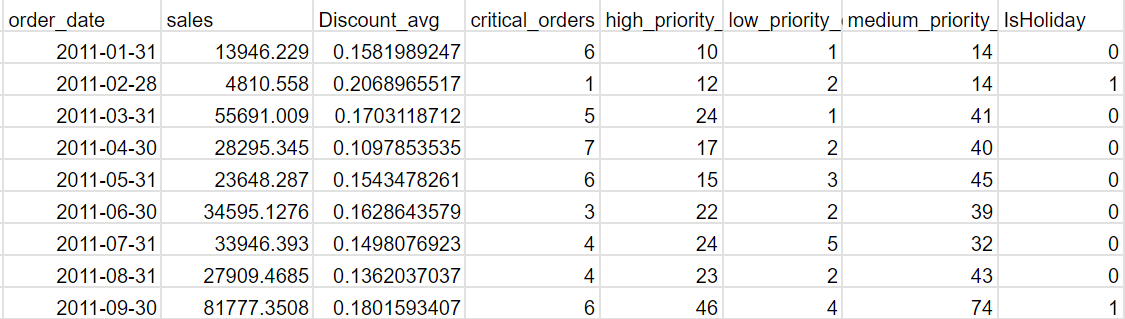
We have utilized python to explore , clean and aggregate the data. We have identified that there are no missing values in the dataset. Our dataset contains sales and other details at product level and since there could be more than one product in a particular order, we first aggregate at order level, to sum up the sales and calculate total discount for each order. Also the sales data available is at day level, we next aggregate all the sales data over the month, to get monthly aggregated sales. Since we have data corresponding to 4 years, the aggregated data resulted in a 48 point time series. Following are the independent variables.

**isHoliday**: This is an external variable that we gathered based on holidays present in a given year. It can take two values 0/1. A value of 1 indicates that there was a holiday present in that particular month and year. Since holidays have an impact on sales, we wanted to include this as an external variable along with the best model selected from ARIMA/Seasonal ARIMA experiments.

**Discount\_avg:** This indicated an average discount per order in a given month of a year. This is calculated by first calculating total discount per order and then aggregating by taking an average value over a month. Since discounts may have some impact on the sales, we have developed this variable from the data

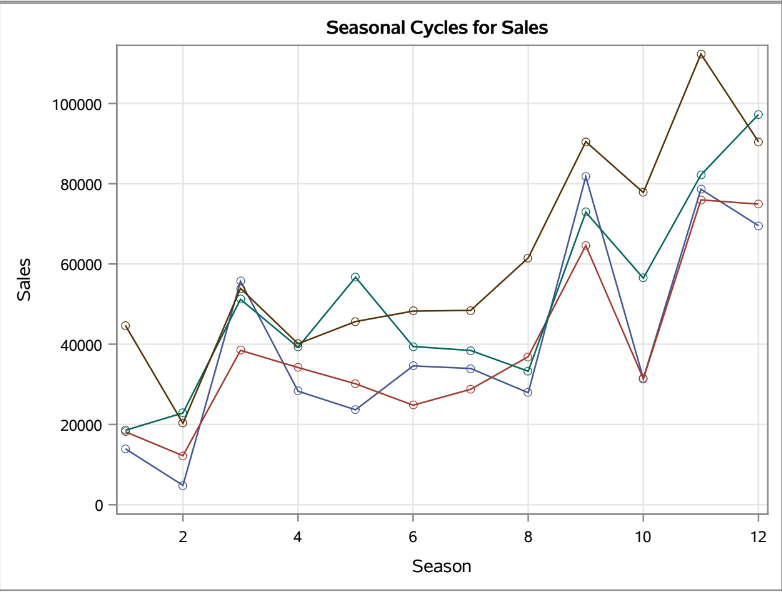
Apart from these variables, we have created variables like num\_critical\_orders, num\_high\_level\_orders etc, by taking the count of all the orders that are ordered with critical priority and high priority respectively.

After performing all the necessary aggregation, following is how our aggregated data looks like

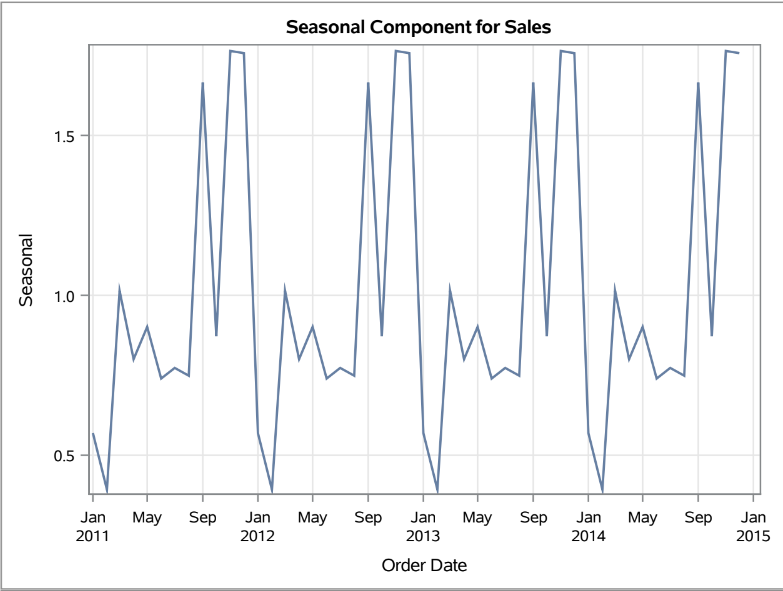


#### 

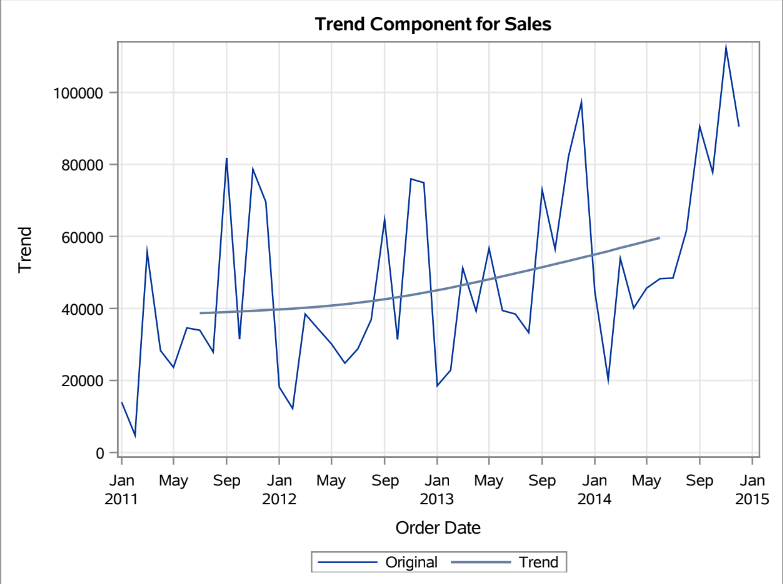
#### **Decomposition Analysis**



From the decomposition analysis, In seasonal cycles for sales graph we can see that most lines experience peaks and troughs, suggesting that sales for these products are influenced by seasonality.

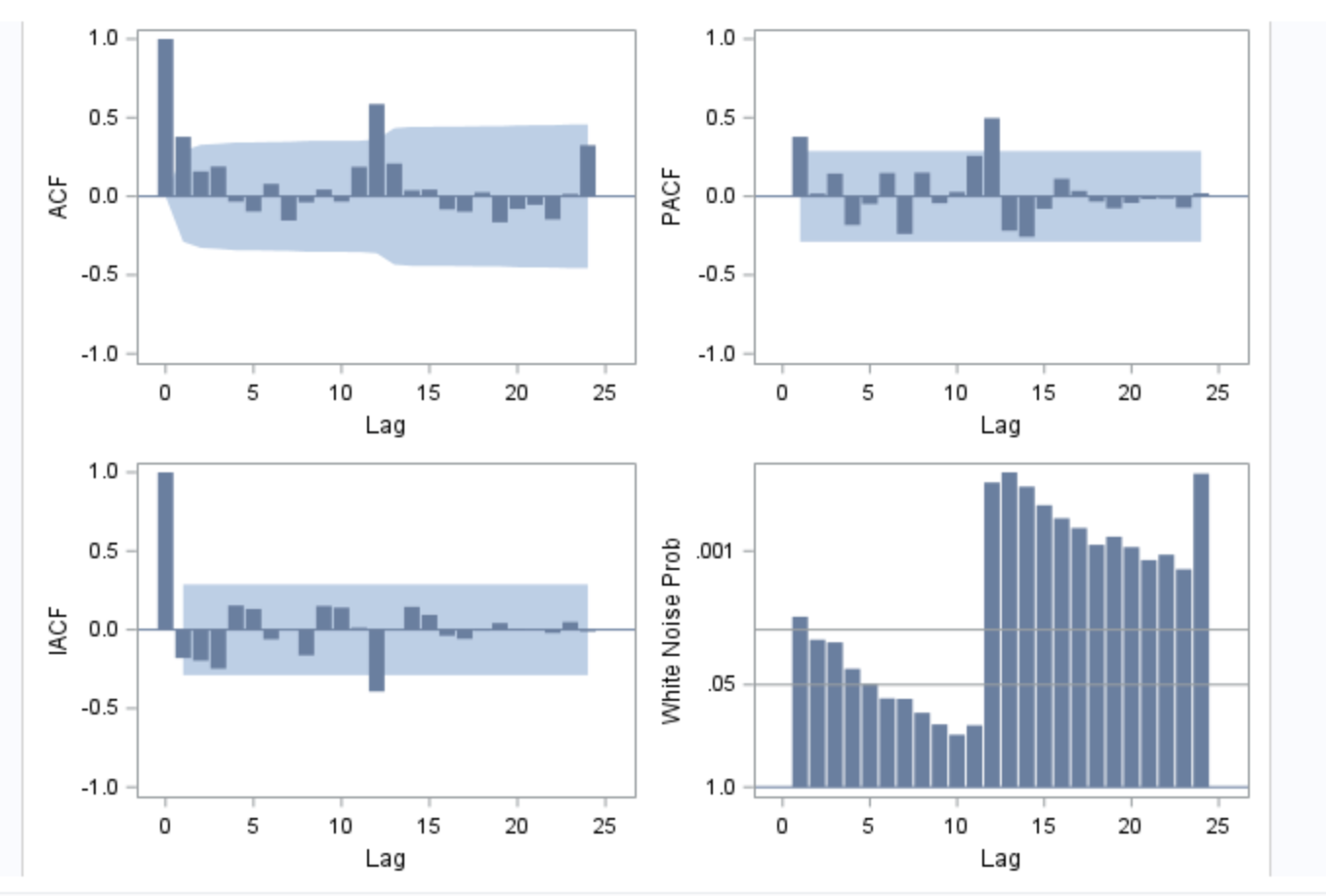


From the Seasonal components from sales graph we can clearly see that there is a seasonality with less number of sales in the month of february and peak in the month of November every year.



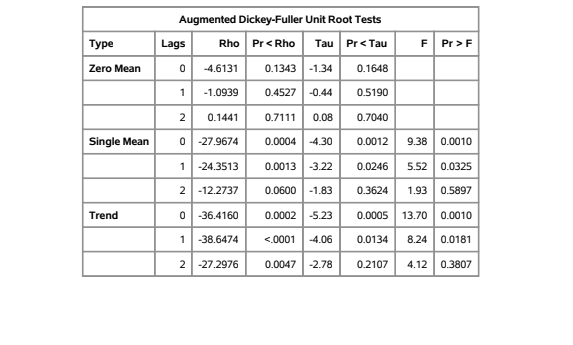
We can also see that there is an increasing trend in the data represented by the quadratic line. Since we have both trend and seasonality we will be doing winters additive model in exponential modeling.

#### **Correlation Analysis**



The ACF shows two significant spikes. It indicates that the time series is positively correlated with itself at all lags up to 12. This means that the value of a time series at any given point in time is likely to be similar to the value of a time series at a previous point in time. There is evidence of dependency across the first 12 lags. Moreover, There is an exponential decay in the ACF graph. The PACF graph shows two nonzero lags with a significant spike (lag1 and lag 12). The spike at lag 1 is significant and above the upper confidence interval. The spike at lag 12 is again significant and above the upper confidence interval. It indicates that there is a seasonality. The PACF plot shows that the time series is significantly positively correlated with itself at a lag of 1 and 12. Moreover, There is a sharp cut-off in the PACF graph. It indicates that there is an autoregressive series. This might suggest that an AR(1) model may be suitable for this time series. The correlation panel shows that it is not completely white noise since many of the bars are above the threshold line.

#### **Test for Stationarity**



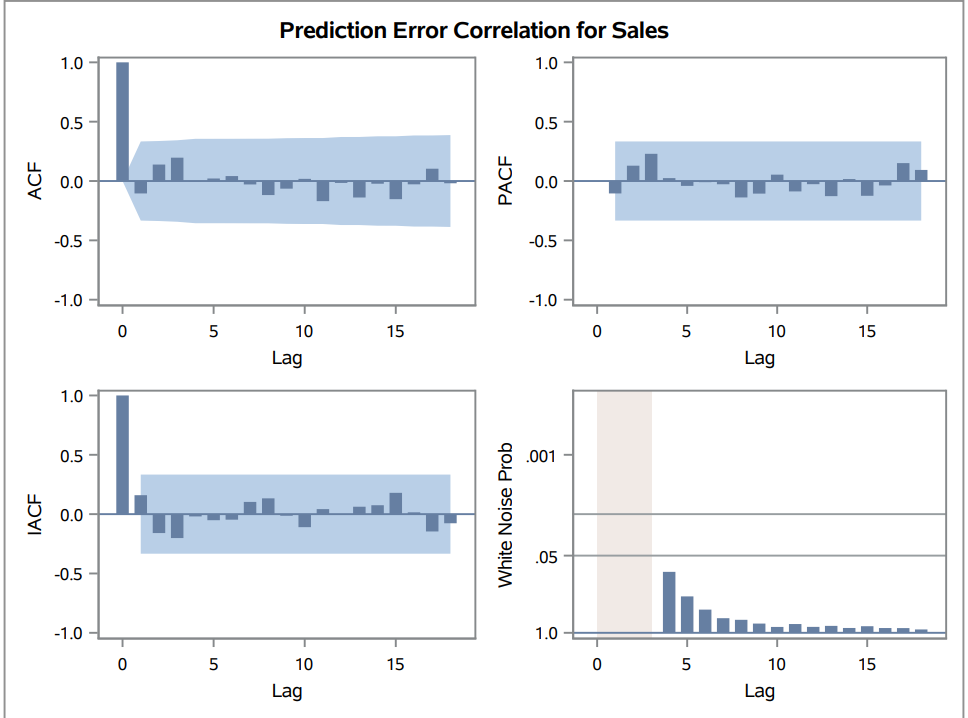
Based on the unit root test result, since our model has a trend, it can be observed that the probabilities at lag 0, lag 1 are significant, while at lag 2 is insignificant. Probably, we can conclude that the series is stationary. Hence we first tried out ARMA models (without any differencing order). However, since we have a quadratic trend and insignificant probability at lag 2 from the ADF test, we also experimented with ARIMA (differencing order of 1) and compared the results of ARMA and ARIMA models to choose the best model.

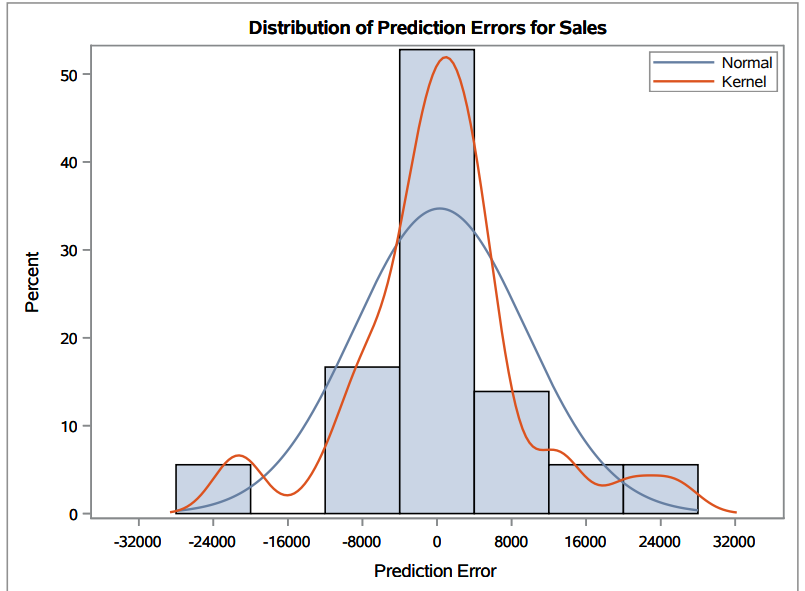
#### **Models**

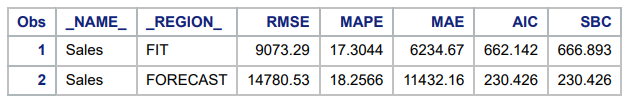
##### **Exponential Additive, Multiplicative models -**

From the seasonal cycles plot, it can be observed that although seasonal cycles look stable for most of the time, we can see that it is unstable at some points. Hence we have tried both additive and Multiplicative models and chosen the best one based on performance. We have chosen a 25% holdback period which equals 12 months of data and we evaluated the models performance on holdout data.

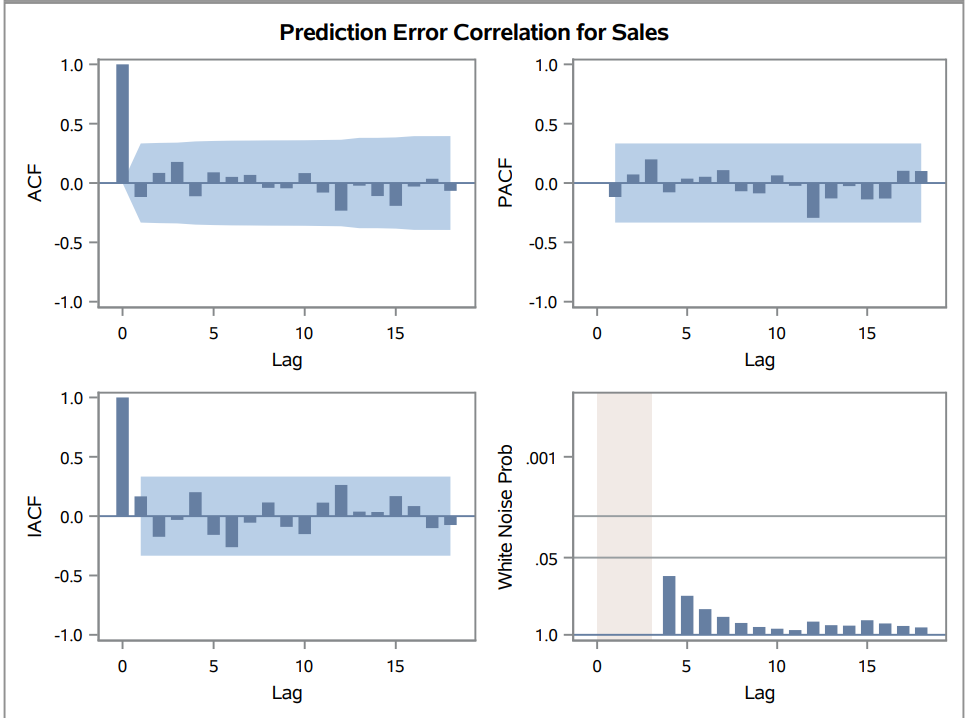
###### **Winters Multiplicative Exponential Smoothing:**

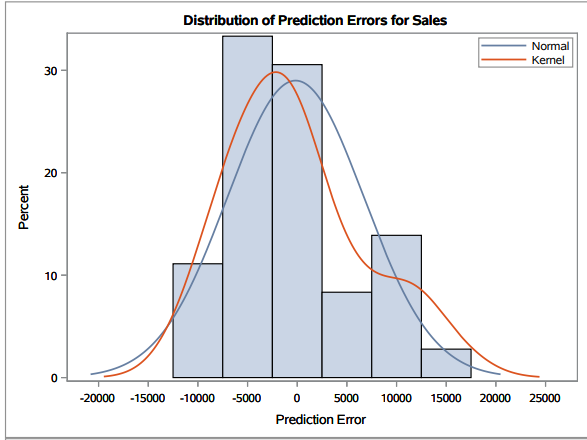




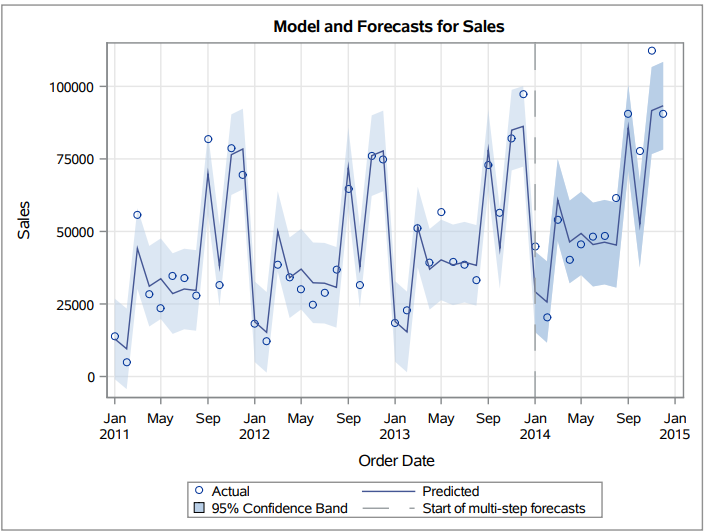


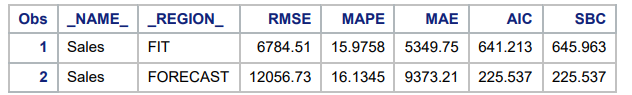
###### **Winters Exponential Additive**:





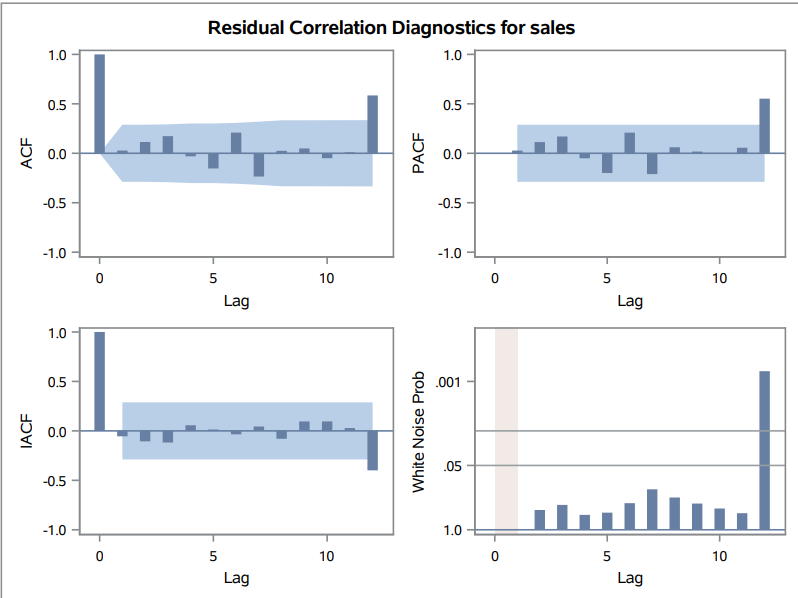
Upon fitting the Winters Additive Model, residuals are white noise and there is no autocorrelation left out in the residual. Hence the model fit the data well and there is no signal left in the residuals.





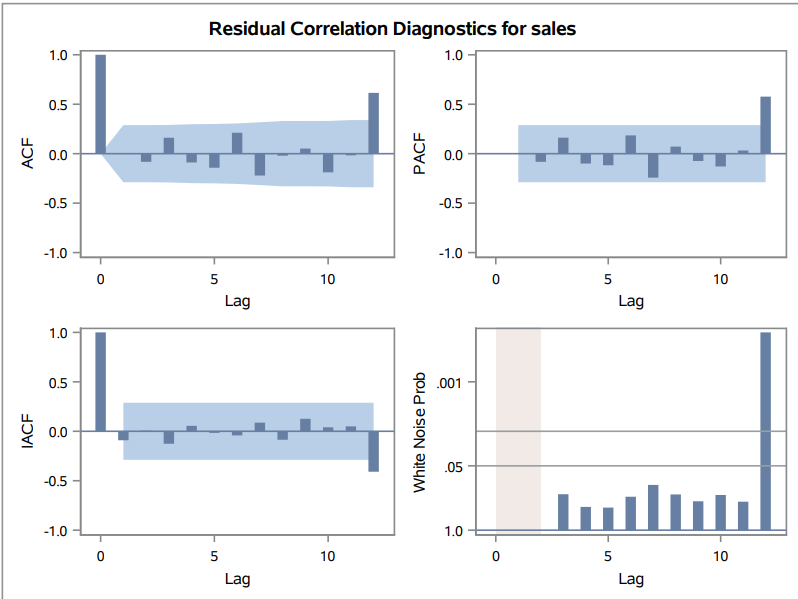
##### **ARMA Models**

###### **ARMA model results - (0,1)**



Interpretation:Based on the ACF and PACF, we found that there is a spike outside the confidence interval. It indicates that there is a correlation. In addition, It is not completely white noise. In this trial we can also see that there is a spike at lag 12 in ACF and PACF and also we can also see that residuals are not distributed as white noise. So we will not be considering this model.

###### **ARMA model results - (1,1)**

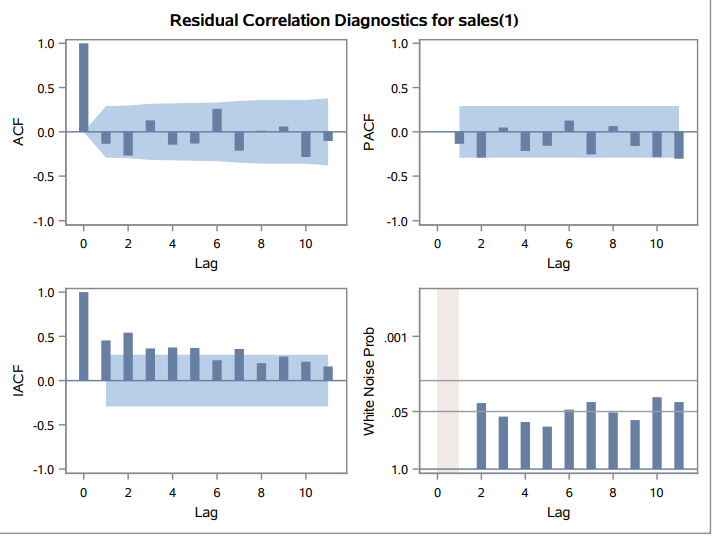


Interpretation: Even in ARMA(1,1) we can see that there is a spike at lag 12 and we can also see that residuals are not white noise. So we will not be considering this model.

##### **ARIMA Models**

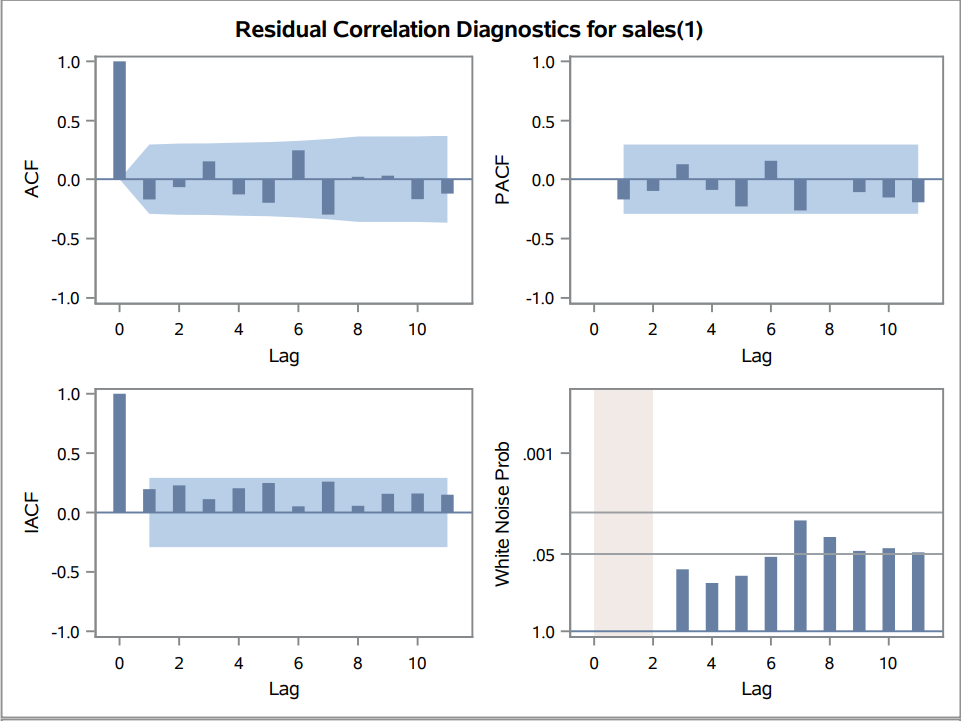
We tried ARIMA models with a differencing order of 1 as none of the ARMA models were able to fit the data well such that residuals resulted in white noise.

###### **ARIMA Model results - (1,1,0)**



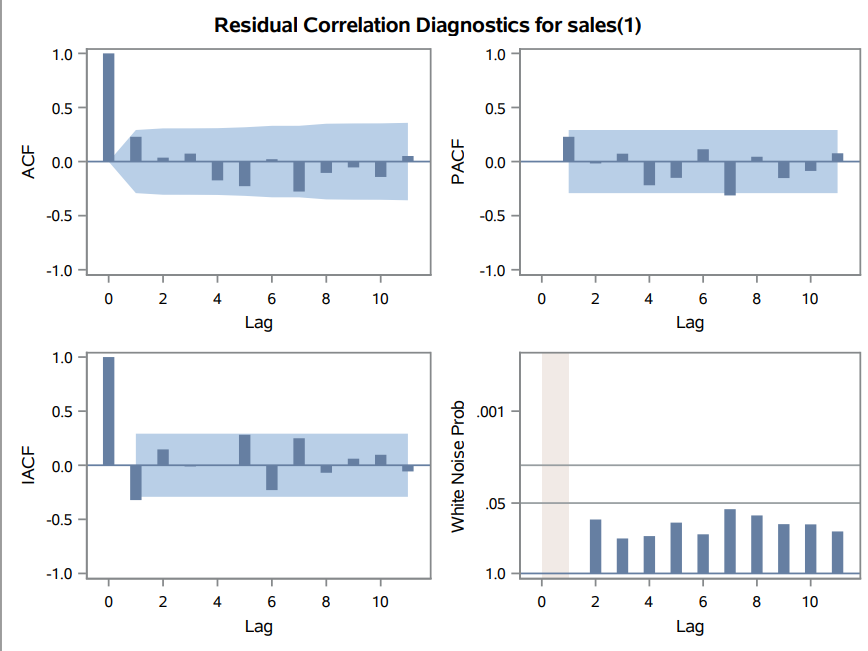
Interpretation: In the ARIMA (1,1,0), there are no significant spikes in the ACF and PACF graph which is a good thing. However, residuals are not completely white noise. Therefore, we will also not consider this model.

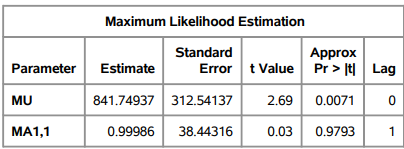
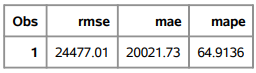
###### **ARIMA Model results - (1, 1, 1)**

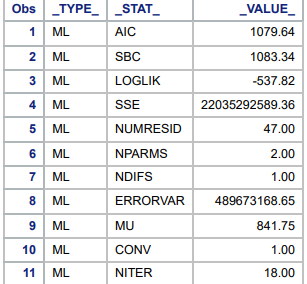


Interpretation: When we consider ARIMA(1,1,1) , Even though there are no spikes in ACF and PACF, residuals are not white noise. Hence the model did not fit the series completely and we will not be able to consider this model further.

###### **ARIMA Model results - (0, 1, 1)**



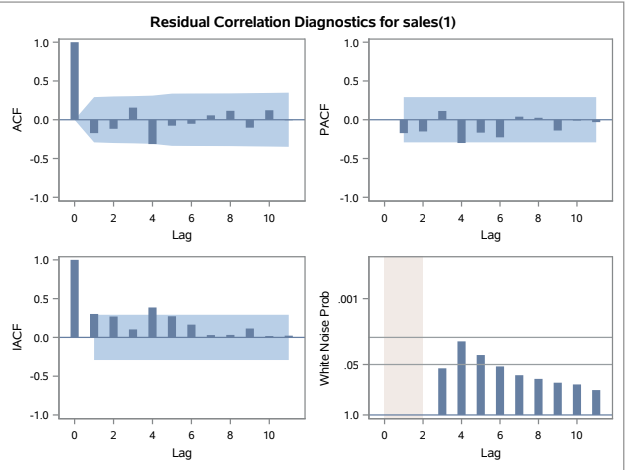


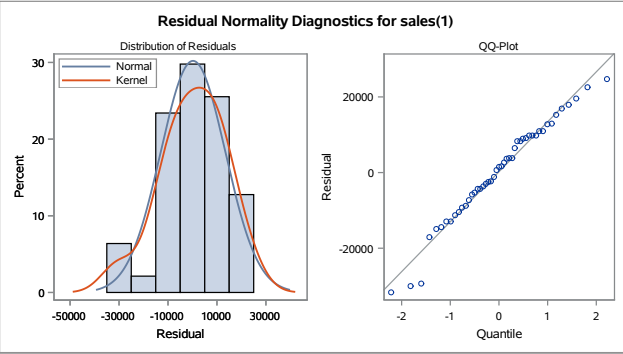
Interpretation: We can see that with this combination of p,d,q, the model was able to fit the data well and the resultant residuals are white noise. There is no autocorrelation left in the series. We can consider this model for comparison against other models to choose the best performing model

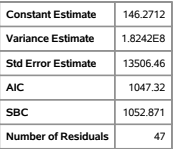
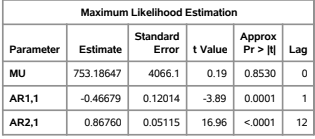
From ACF and PACF plots, we found that there is seasonality in this time series data, we decided to try to use the seasonal ARIMA model

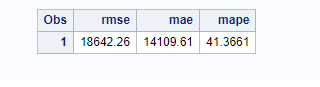
##### **Seasonal ARIMA Models**

###### **Seasonal ARIMA model results ((1,1,0),(1,0,0))**



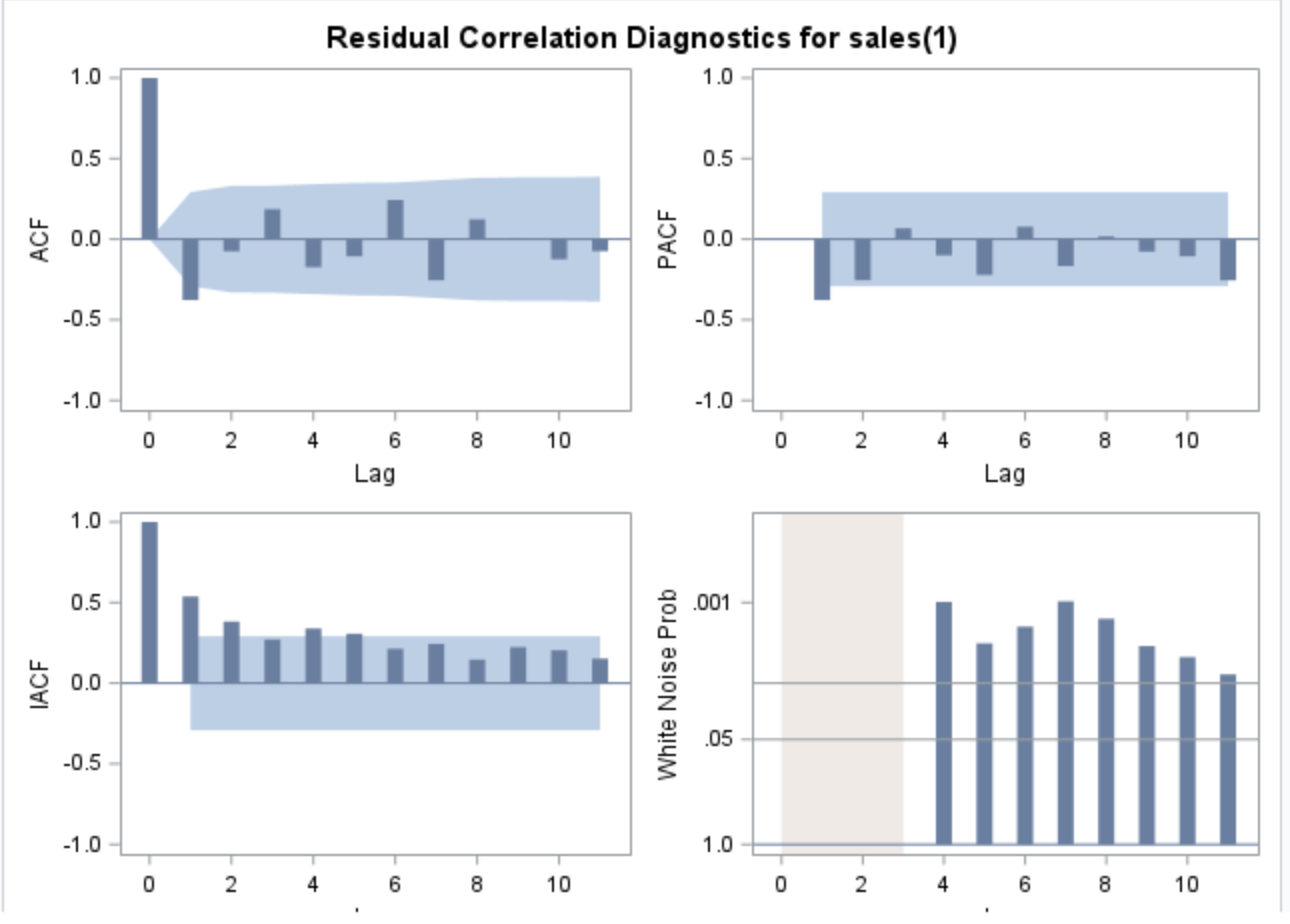






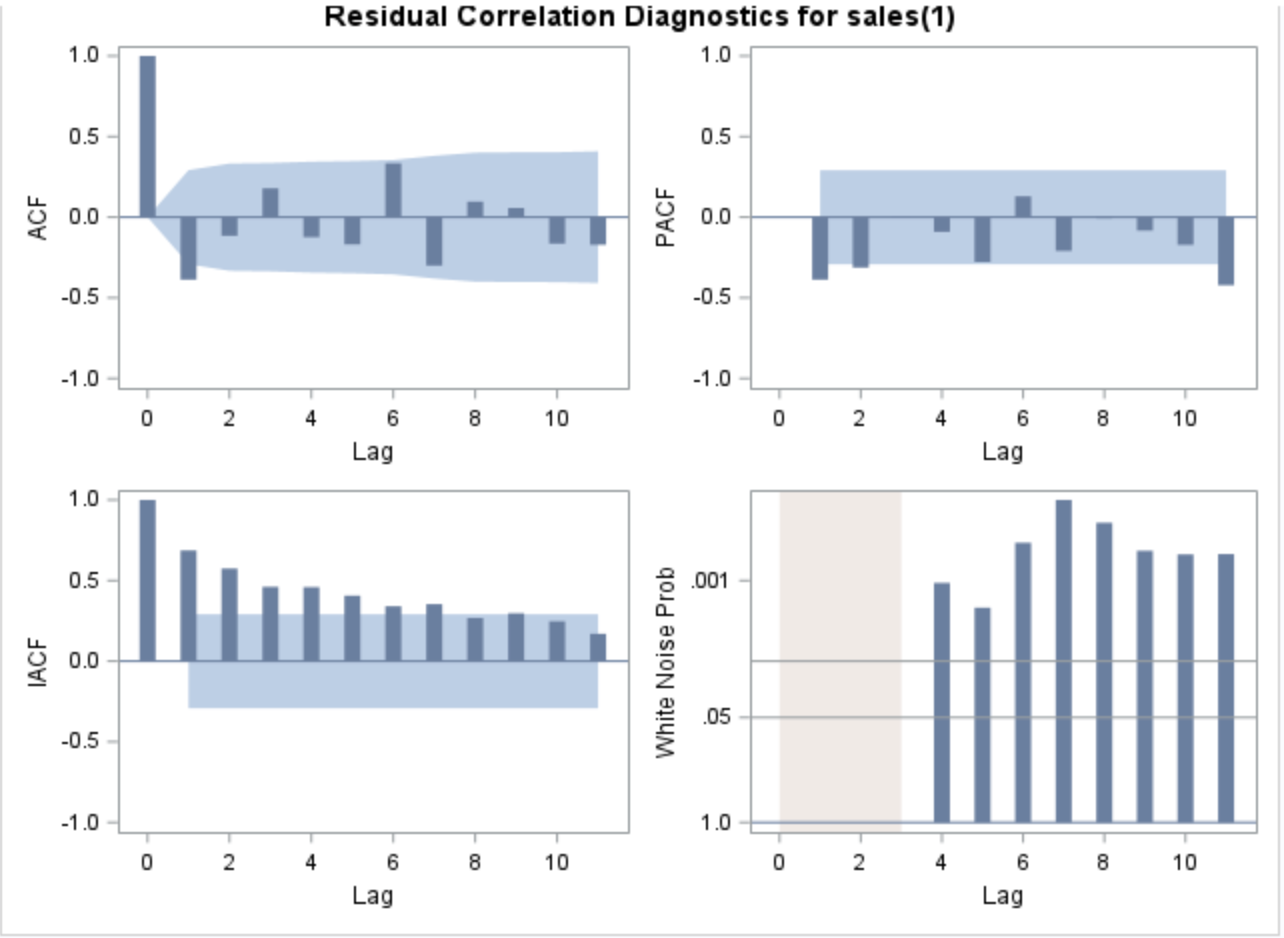
**Interpretation:**. From the residual correlation diagnostics for sales, we can find that there are no significant spikes beyond the confidence interval which means that there is no correlation. Some of the bars in the white noise probability graph exceed the threshold. This means that the residuals are not perfectly random. The RMSE of this model is 18642.26 which is higher than we expected. This indicates that this model still has room for improvement.

###### **Seasonal ARIMA (1,1,1)(1,0,0)**



Interpretation: Based on the residual correlation diagnostics, there are some significant spikes beyond the confidence interval which means that there is a correlation. Moreover, all of the bars in the white noise probability graph exceed the threshold. It indicates that there are non-random patterns in residuals. We will not use this model since it didn’t pass the white noise test.

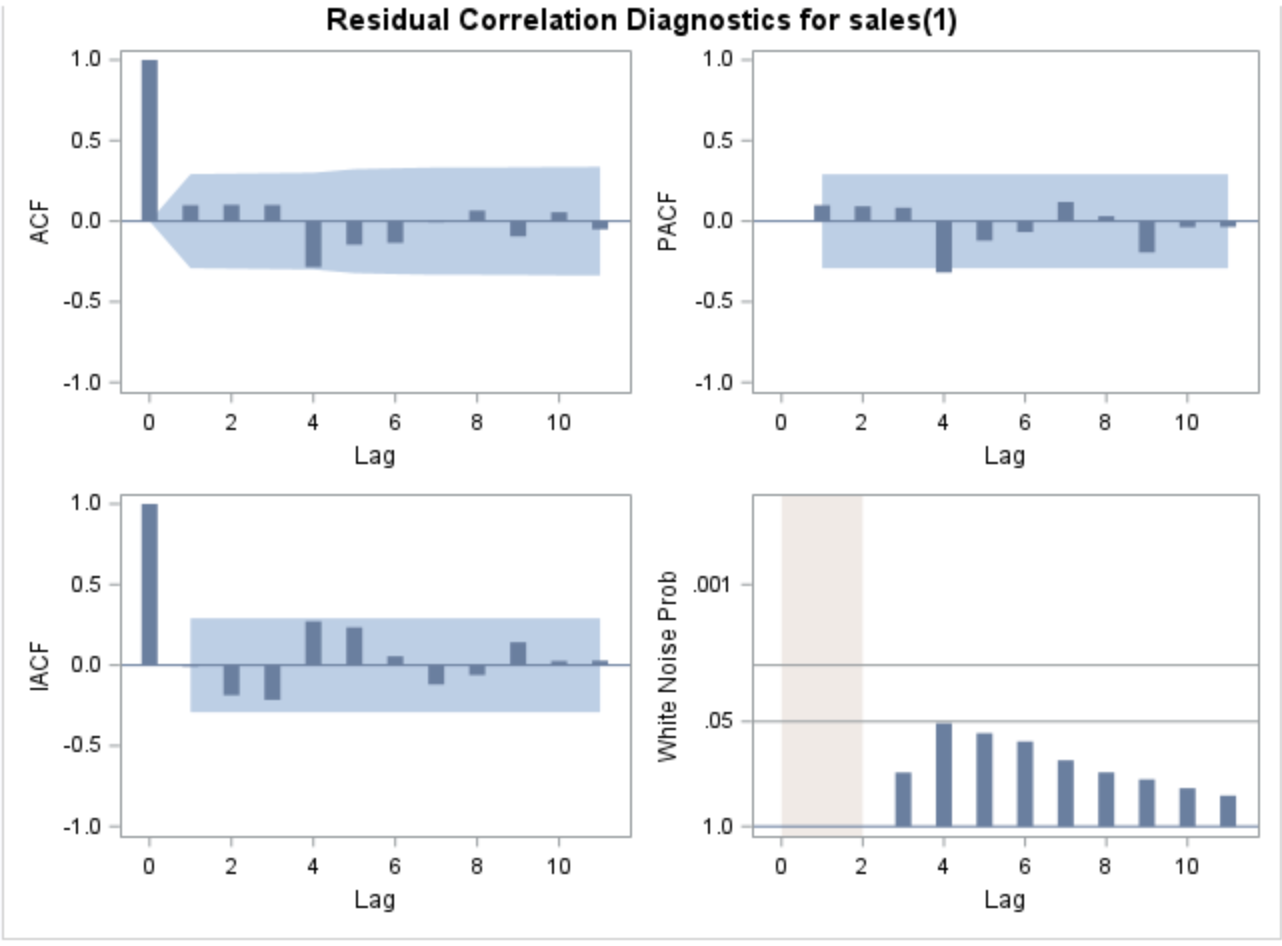
###### **Seasonal ARIMA model (1,1,1) (0,0,1)**

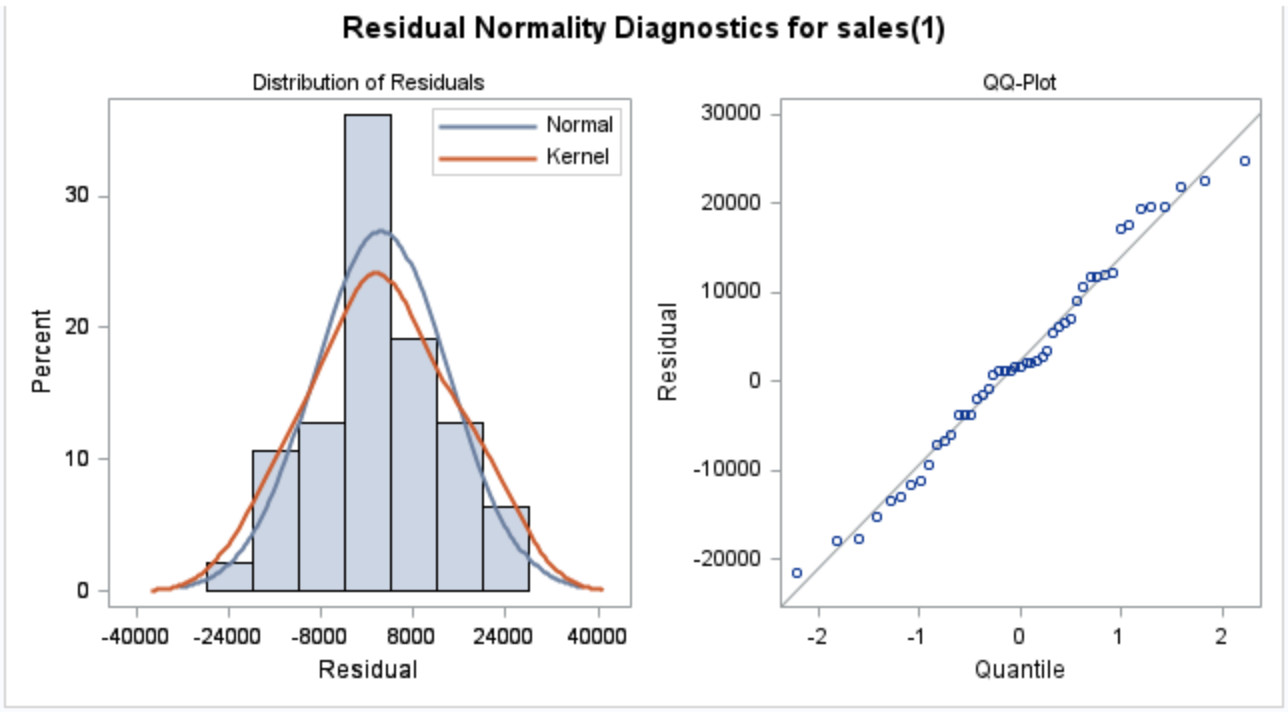


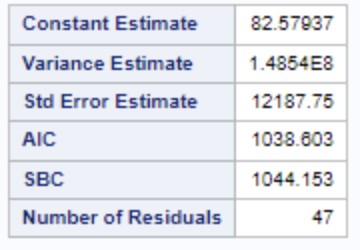
Interpretation: Based on the residual correlation diagnostics, there are some significant spikes beyond the confidence interval which means that there is a correlation. Moreover, all of the bars in the white noise probability graph exceed the threshold. It indicates that there are non-random patterns in residuals. We will not use this model since it didn’t pass the white noise test.

###### **Seasonal ARIMA model (0,1,1, 1,0,0) :**

##### 

****

****

****

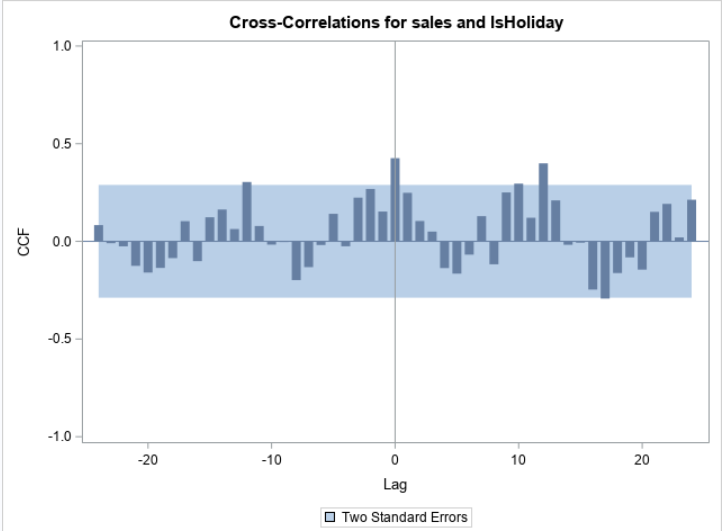
**Interpretation:** From the residual correlation diagnostics for sales, we can find that there are no significant spikes beyond the confidence interval which means that there is no correlation. All of the bars in the white noise probability graph below the threshold. It indicates that there are random patterns in residuals. The distribution of residuals is very close to a normal distribution. Moreover, The RMSE of this model is 16023.74 and the MAPE is 29.23%. Comparing the RMSE, MAE and MAPE, Among all the experiments tried with Seasonal ARIMA, this model is the one that fit the series well and resulted in white noise residuals.Hence we decided to move forward with this model while considering independent variables like isHoliday, average discount per order.

#### **Modeling with Independent variables:**

Since our independent variables are isHoliday and discount\_avg, we first observed the cross correlation between independent variable and target variables sales. We also conducted prewhitening by fitting a model on an independent variable first and then using the same model to fit on our target variable. If the resultant residuals are white noise, then the independent has no impact on the target variable. If they are not completely white, then independent variable has an impact and we can include it in the modeling to improve the performance of model fit.

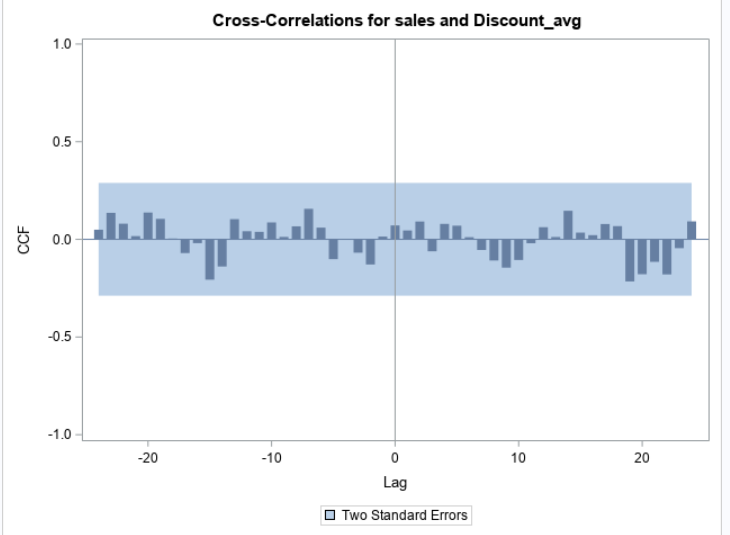
#### **Cross Correlation Plots**

##### **Cross Correlation plots with isHoliday**



We can see cross correlation between holiday and sales at lag 0 and lag 12.

##### **Cross Correlation with Discount(avg)**



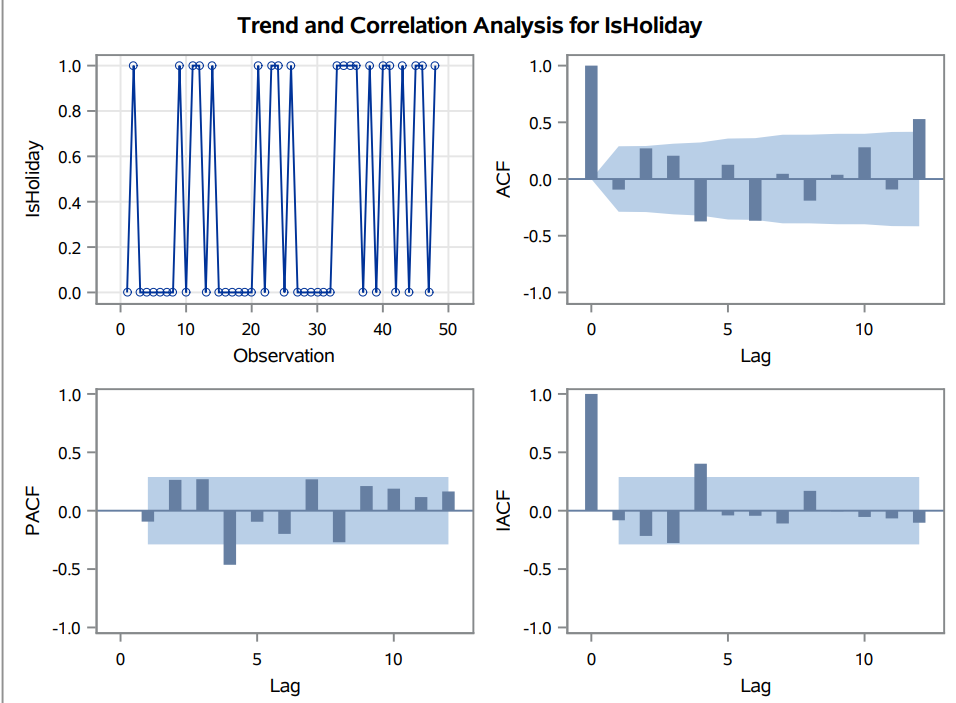
We do not see any significant cross correlation between sales and average discount per order.

From the cross correlation plots, it can be observed that sales has cross correlation with isHoliday but not with Discount\_avg. We further do prewhitening to see which lags have significant cross correlation.

#### **Pre Whitening**

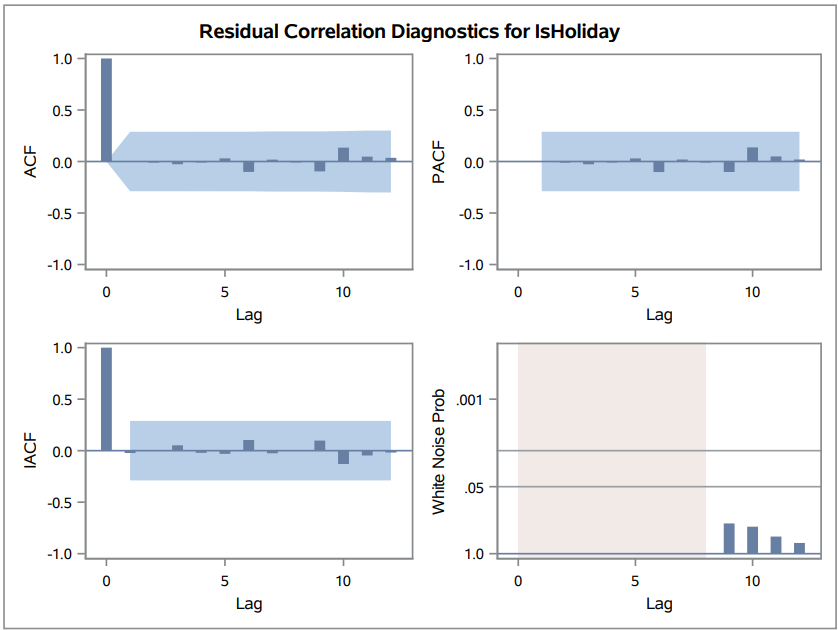
##### **Pre-Whitening on Holiday data :**

Following are the ACF, PACF, IACF plots of isHoliday Series.

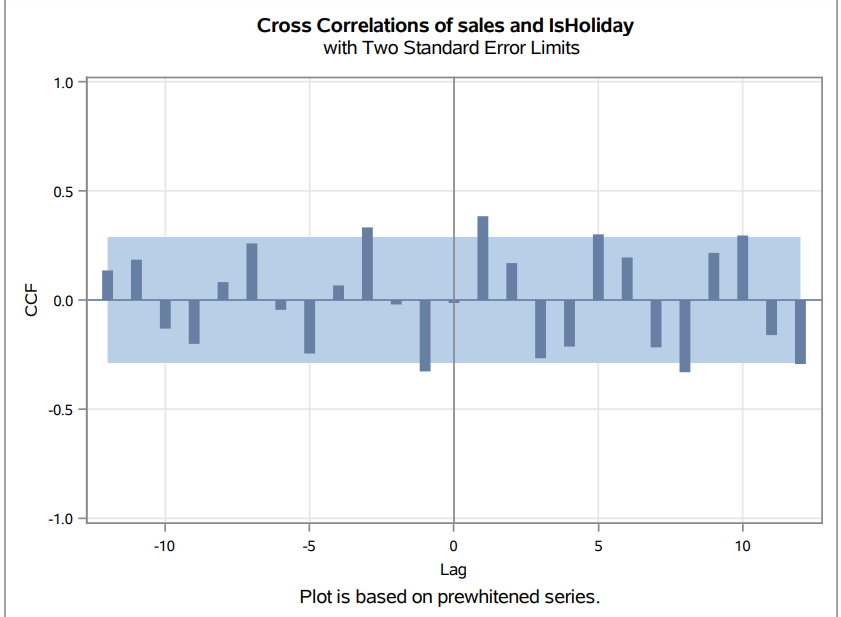


Fit a Model on Holiday data

Upon trying different experiments, the best model that could fit isHoliday well is ARIMA. Following are the residual plots after fitting the best model on isHoliday. We can observe that residuals are white noise and have no autocorrelation



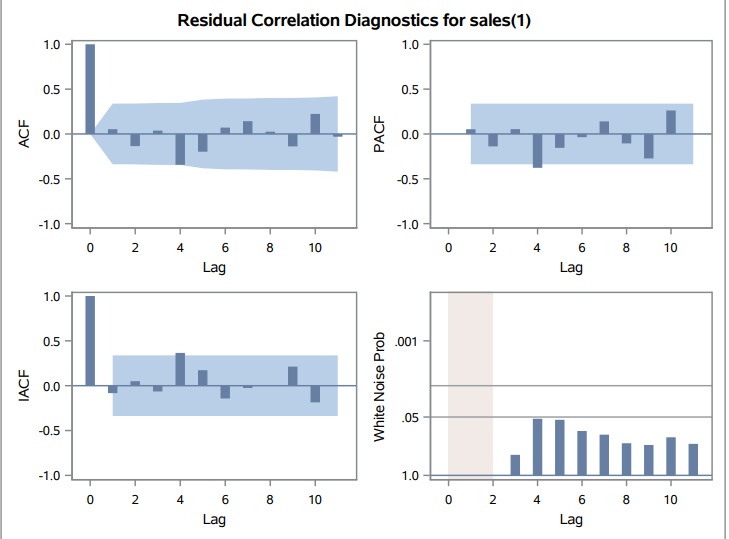
Now , we can fit the same model on sales data and observe the residuals. Below is the plot of cross correlation of sales on the fit model, we can see that there is some significant spike at lag 1 and it indicates that holiday has some impact on sales. Hence we can include Holiday as an independent variable along with the best Seasonal ARIMA Model.



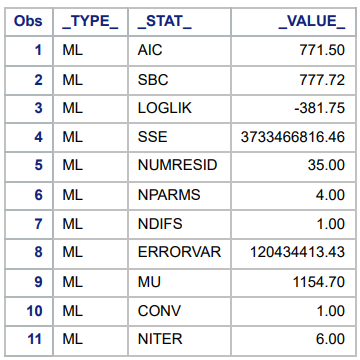
Since isholiday has significant lag effect at lag 1 and lag 12, we included is\_holiday as external variable at these two lags and build the following Seasonal ARIMAX model

##### **Seasonal ARIMAX (with holiday) - (1, 12)**

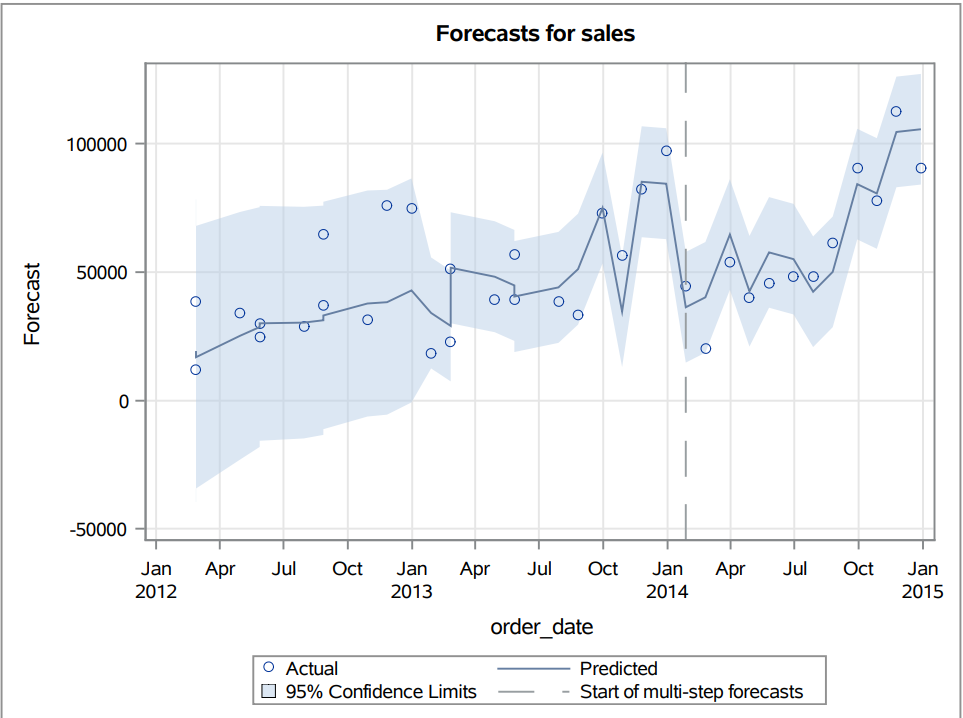
Take Holiday as independent and consider lag 1, 12





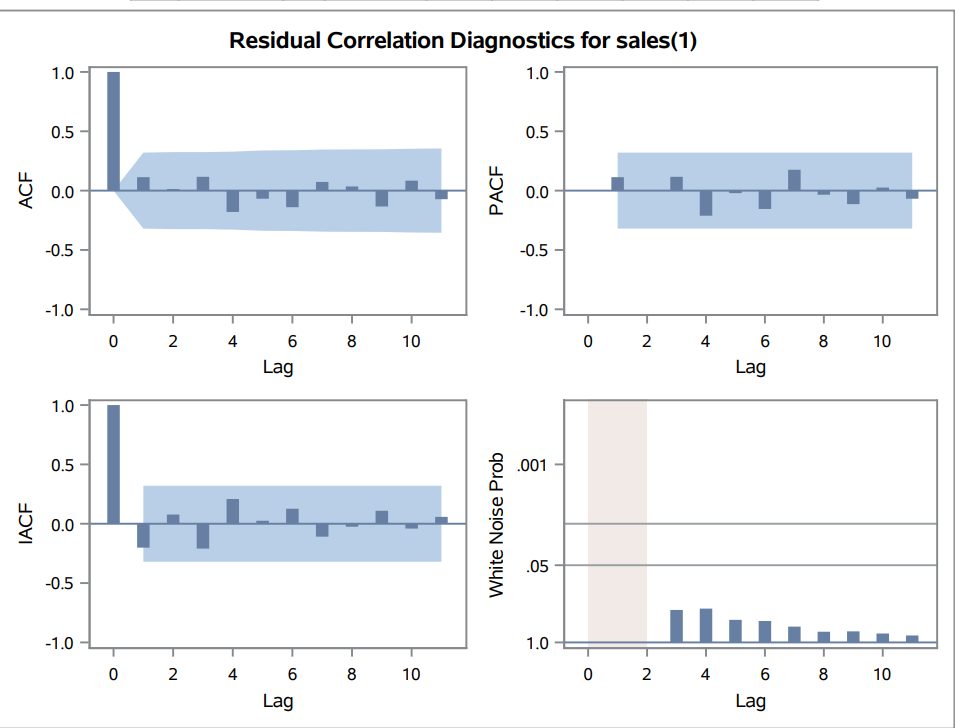


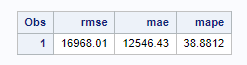
We can see that the resultant residuals are white noise and there is no significant autocorrelation. This model has an MAPE of 25%. Below is the plot to visualize the holdout sample forecasted by Seasonal ARIMAX model with holiday data considered at lag 1 and lag 12.

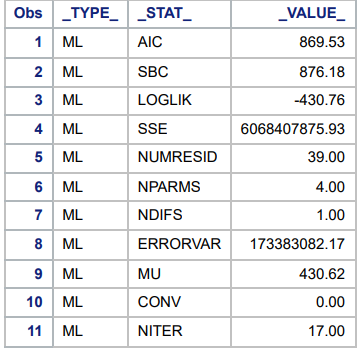


###### **c. Seasonal ARIMAX (with holiday) - (1, 8)**

Use isHoliday as an independent variable at lags 1 & 8 and build a Seasonal ARIMAX model.

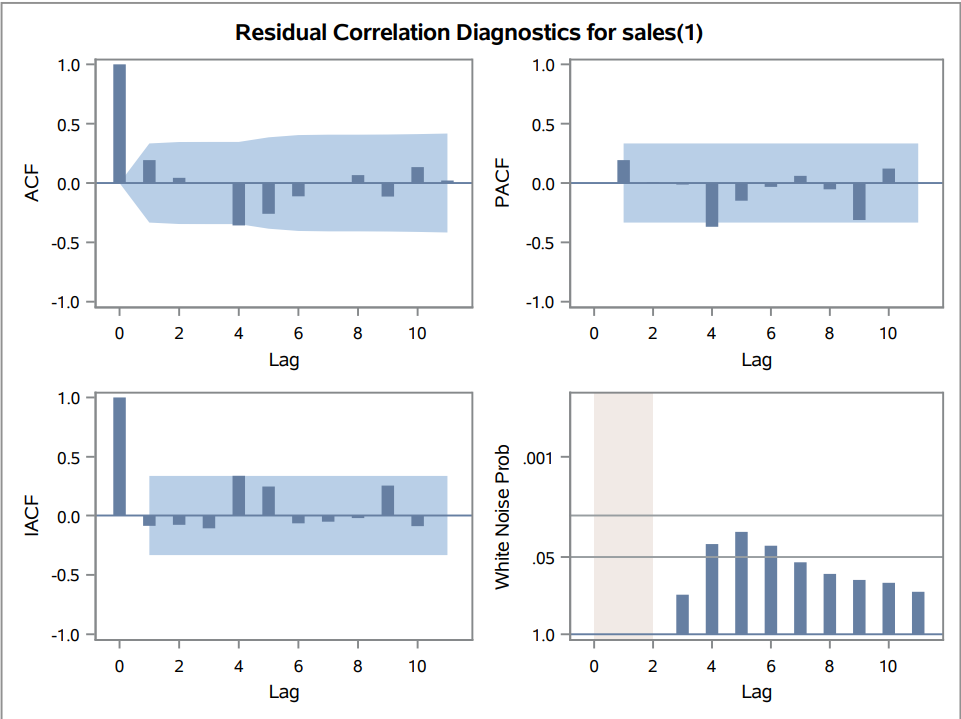






Although Seasonal ARIMAX with isholiday lagged at 1, 8 resulted in white noise residuals, the MAPE of this model is a bit higher and probably this is not the best choice out of Seasonal ARIMAX models experimented.

###### **d. Seasonal ARIMAX (with holiday) - lag 12**

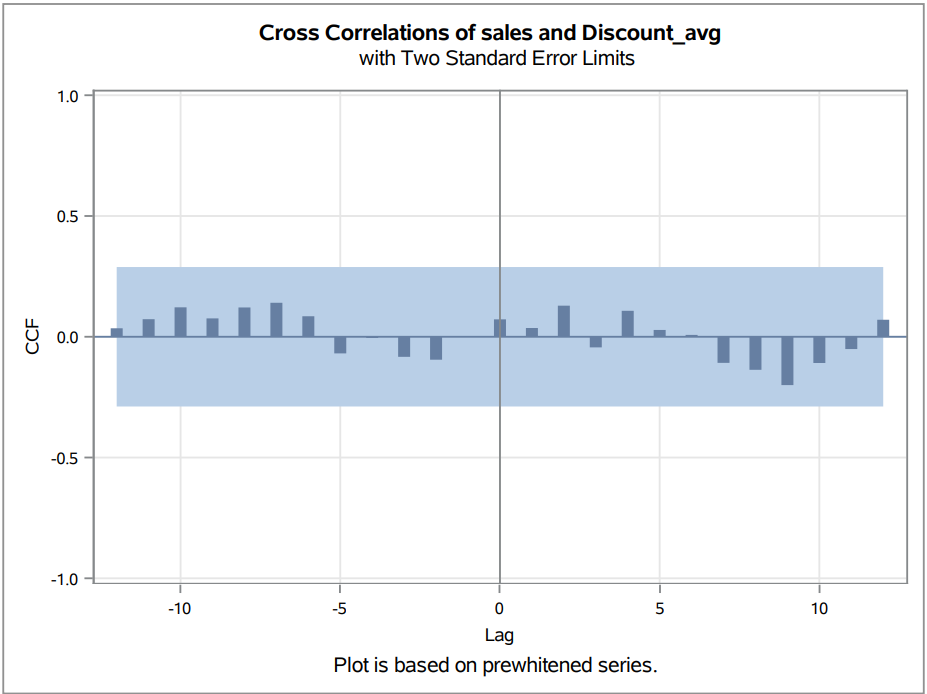


Upon only including isHoliday at lag 12, we can observe that the model didn’t fit the data well and the residuals show some level of autocorrelation at few lags, Hence this model will not be considered further.

Upon comparing all the Seasonal ARIMAX experiments, Seasonal ARIMAX with isHoliday variable considered at lag 1, 12 has best performance in terms of MAPE, AIC, SBC.

##### **e. PreWhitening on Discount\_avg**

We first fit an ARIMA Model on discount\_avg such that residuals are white noise.

We then use the same model and apply it on sales. Below is the plot of cross correlation between sales and discount\_avg after fitting the same model that fits discount\_avg. 

Prewhitening result shows that there is no significant impact of discount average on sales.

Below is a summary of all the models experimented and their corresponding performance comparison. All those models where residuals did not pass white noise test are not considered further in best model selection.

#### **Model Comparison**

| No. | Model | White Noise Test | RMSE | MAE | MAPE | AIC | SBC |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Winters Multiplicative Exponential | Passed | 9073.29 | 6234.67 | 17.3044 | 662.142 | 666.893 |
| 2 | Winters Exponential Additive | Passed | 6784.51 | 5349.75 | 15.9758 | 641.213 | 645.963 |
| 3 | ARMA(1,0) | Failed | 24788.83 | 20210.24 | 65.2594 | 1103.52 | 1107.26 |
| 4 | ARMA(0,1) | Failed | 24477.01 | 20021.73 | 64.9136 | 1104.51 | 1108.25 |
| 5 | ARMA(1,1) | Failed | 24938.03 | 20323.73 | 66.241 | 1104.96 | 1110.58 |
| 6 | ARIMA(1,1,0) | Failed | 32428.03 | 25648.48 | 74.2618 | 1087.69 | 1091.39 |
| 7 | ARIMA(1,1,1) | Failed | 23532.44 | 19395.19 | 53.6151 | 1080.13 | 1085.68 |
| 8 | ARIMA(0,1,1) | Passed | 24477.01 | 20021.73 | 64.9136 | 1079.64 | 1083.34 |
| 9 | Seasonal ARIMA(1,1,0),(1,0,0) | Failed | 18642.26 | 14109.61 | 41.3661 | 1047.32 | 1052.87 |
| 10 | Seasonal ARIMA((1,1,1),(1,0,0)) | Failed | 22536.59 | 16898.68 | 48.4435 | 1072.73 | 1080.13 |
| 11 | Seasonal ARIMA((1,1,1),(0,0,1)) | Failed | 17456.05 | 13456.32 | 44.9624 | 1090.91 | 1098.31 |
| 12 | Seasonal ARIMAX(1,12) | Passed | 14064.26 | 10749.82 | 25.2272 | 771.50 | 777.72 |
| 13 | Seasonal ARIMAX(1,8) | Passed | 16968.01 | 12546.43 | 38.8812 | 869.53 | 876.18 |
| 14 | Seasonal ARIMAX(12) with holiday | Failed | 13996.86 | 8802.36 | 27.14 | 797.32 | 803.66 |

#### **Conclusion**

**Model:** Our best model in terms of MAPE and other accuracy metrics is winter's additive exponential smoothing model which slightly outperformed Seasonal ARIMAX with holiday variables considered at lag 1 and lag 12. However, Seasonal ARIMAX is quite a complex model, it has higher AIC and SBC when compared to simple winters additive exponential model. But considering that the SARIMAX model has taken independent variables into consideration, it might be more robust in making future forecasts, as the exponential smoothing model would only consider the target variable to generate a forecast . Hence we conclude that seasonal ARIMAX is our best final model. Initially, our group believed that average discount per order would have some impact on the sales, but upon prewhitening, we found that average discount per order has no significant impact on the sales.

**Business insights & Recommendations:**

Our forecast is projecting growth in the business of the Global Superstore in the coming years. Over the years, It also depicts a constant spike around the month of December-January and a constant dip around the month of February. So based on these insights, we would make the following recommendations to the store:

1. Optimize their inventory to avoid under stocking during the month of December and understocking during the month of February.
2. The marketing team can work on re-strategizing the marketing around the month of february and find a way to take their products to the customers in those months where they don't intend on spending as much.
3. The investors can get a hint on where to invest through the forecast and also, there can be plans around expansion.