OPIM 5671: Data Mining and Business Intelligence

Spring 2024, Section 712-1243

Walmart's Quarterly Sales Forecast with SAS

A time-series forecasting analysis

Team 4

Chandra Harsha Alla (3127095)

Nimish Pastaria (3134064)

Pratik More (3132440)

Sri Vinay (VN23001)

Zilong Mao (2866567)

TABLE OF CONTENTS

1. Executive Summary
2. Data Description
3. Data Exploration
4. Model Building, Evaluation and Selection using SAS

   4.1 Model for Store #9

   4.2 Model for Store #14

   4.3 Model for Store #30

   4.4 Model for Store #40
5. Inference and Conclusion
6. References

1. EXECUTIVE SUMMARY

Our problem statement is to develop a predictive model that accurately forecasts weekly sales for Walmart stores by taking into account historical sales data and economic indicators along with some external factors.

In the current supply chain industry, the cost of storing inventory is very high, given the huge demand and supply. Through accurate forecasting of the sales, we can optimize inventory management and timely delivery of stock to different store locations. The proposed sales forecasting model aims to have the capability to predict sales for Walmart stores based on attributes such as unemployment rate, CPI, fuel prices etc. This initiative also implicitly aims to enhance inventory management, staffing optimization, and strategic planning, ultimately improving operational efficiency and profitability by aligning stock levels with forecasted demand, improving staffing schedules to ensure adequate customer service during peak and off-peak periods, and informing strategic decisions such as promotions and pricing strategies based on anticipated sales trends. This targeted approach can also enhance customer satisfaction, reduce operational costs, and increase profitability by ensuring resources are efficiently allocated to meet consumer needs.

In the original dataset, we had a cumulative record of 45 stores spread across weeks from 2010 to 2012. To make our analysis distinct on stores since individual stores can exhibit nuanced behaviors, we decided to split the file into 45 files with each file containing data for a particular store spread across 2 years with a weekly period. We then segregated the 45 files into 5 pools of 9 stores each, and each team member explored their 9 stores to find the most interesting store with results that varied greatly from the rest of the stores in the respective pool. Out of the 45 stores, we found that 7 stores had a negative sales trend, 14 stores had a positive sales trend, 20 stores had a constant trend and 4 stores had irregular trends. This approach yielded Store #9, Store#14, #Store30, and Store#40 as the most interesting stores out of each pool of 9 stores varying from the other stores in a certain aspect.

As part of our data analysis, we explored numerous models, ranging from Exponential Smoothing Models, ARIMA, and ARIMAX depending on the results from initial time series exploration of each store. Based on the preliminary analysis in case significant variables were found from the cross-correlation plots, we tried to perform pre-whitening to confirm whether the variables are really affecting our target variable. And based on the pre-whitening results we then proceeded to experiment with ARIMAX, ARIMA and ESM models to find the model with the best fit but at the same time being parsimonious in nature with good accuracy. Our best models were as follows,

Store #9 – ARIMA (5,0,2)

Store #14 - ARIMA (1,0,1)(0,1,0)

Store #30 – ARIMA (5,0,3)

Store #40 - ARIMA (2,2,3)

2. DATA DESCRIPTION:

The dataset for this project is taken from Kaggle (link mentioned below), featuring historical sales data for a Walmart store. This dataset includes Weekly Sales, Holidays, and other relevant features like Temperature and Fuel Prices, affecting sales performance.

https://www.kaggle.com/datasets/varsharam/walmart-sales-dataset-of-45stores

Parameter	**Description**
Date	The Week of Sales. It is in the format of dd-mm-yyyy. The date starts from 05-02-2010
Weekly_Sales	The sales of the store in the given week
Holiday_Flag	If the week has a special Holiday or not. 1 - The week has a Holiday 0 - Fully working week
Temperature	Average Temperature of the week in the area
Fuel_Price	Price of the Fuel in the region
CPI	Customer Price Index
Unemployment	Unemployment rate of the region

3. DATA EXPLORATION:

Upon pre-processing the data using JMP, we observed that there are no missing values, no outliers, no variable conversion, no dummy variables and no binning required.

Furthermore, looking at the variable distributions, we concluded that there is no log transformation required. Fortunately, our data was clean and needed no cleaning or pre-processing to proceed with the next steps in forecasting sales.

However, the only thing that we did during the preprocessing stage was adjusting the date format using Excel, which allowed SAS to read the data accurately ensuring data integrity for our forecasting model.

4. Model Building, Evaluation and Selection using SAS

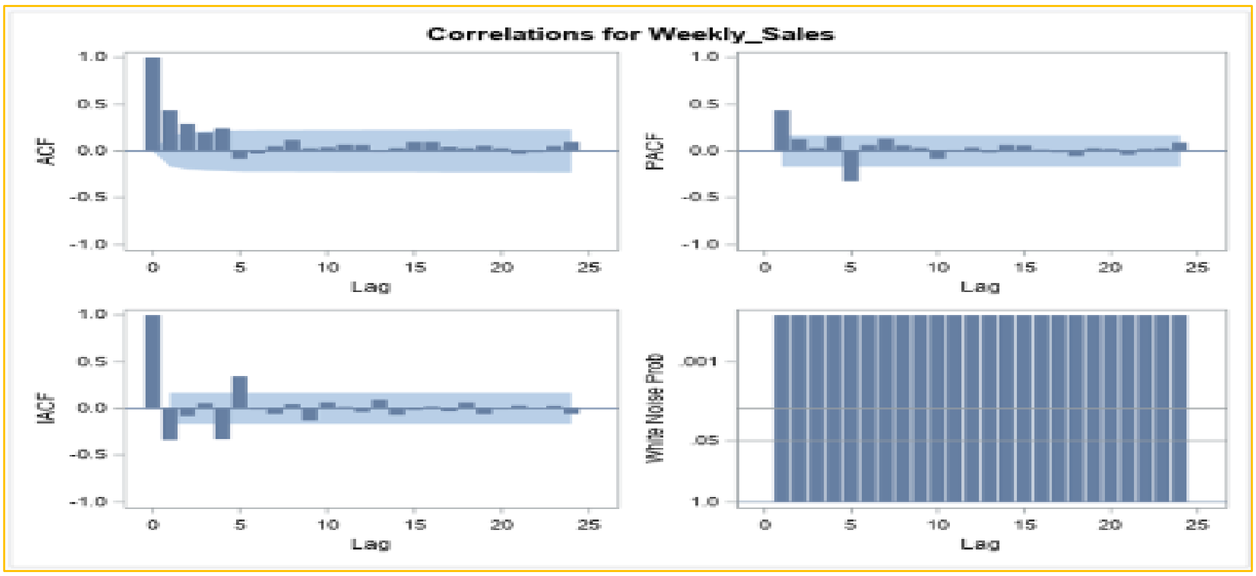
Since we're now ready with the data to build models, please find below the model building and evaluations performed for our interesting stores.

4.1 Model for Store #9

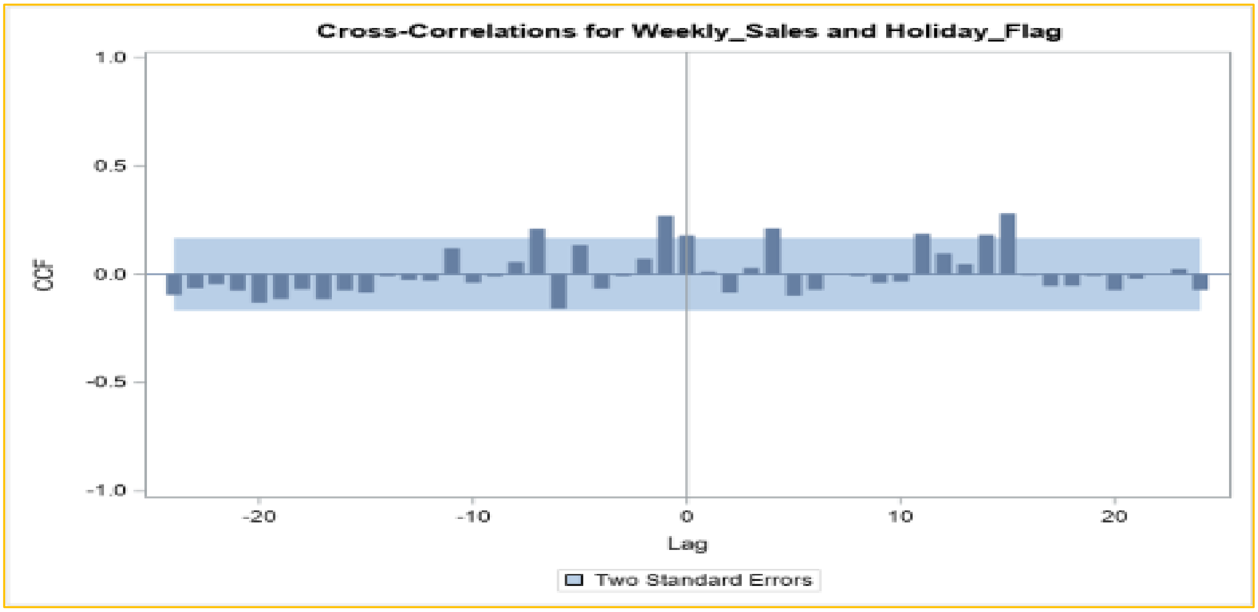
There are a total of 14 stores which display a positive weekly sales trend. Out of all the positive stores, we have selected store_9 to model as it displays initial co-relation with all the individual variables.

Data Exploration (For store 9)

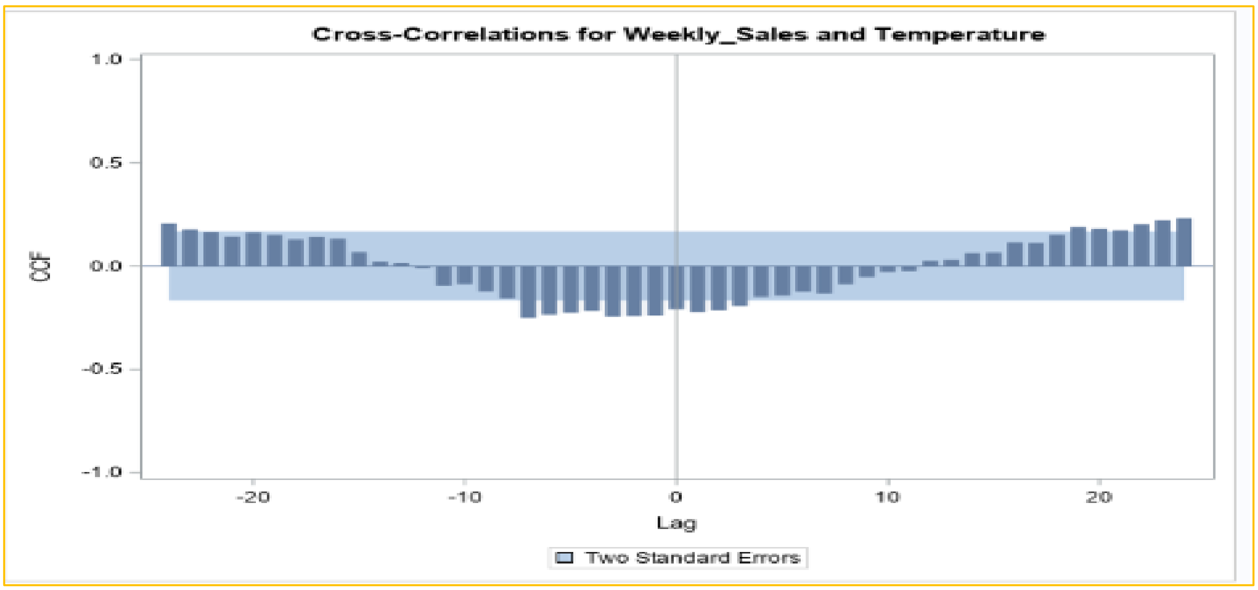
The correlations for weekly sales are as depicted below. We can see that it has failed the white noise test and there are significant spikes in different lags.



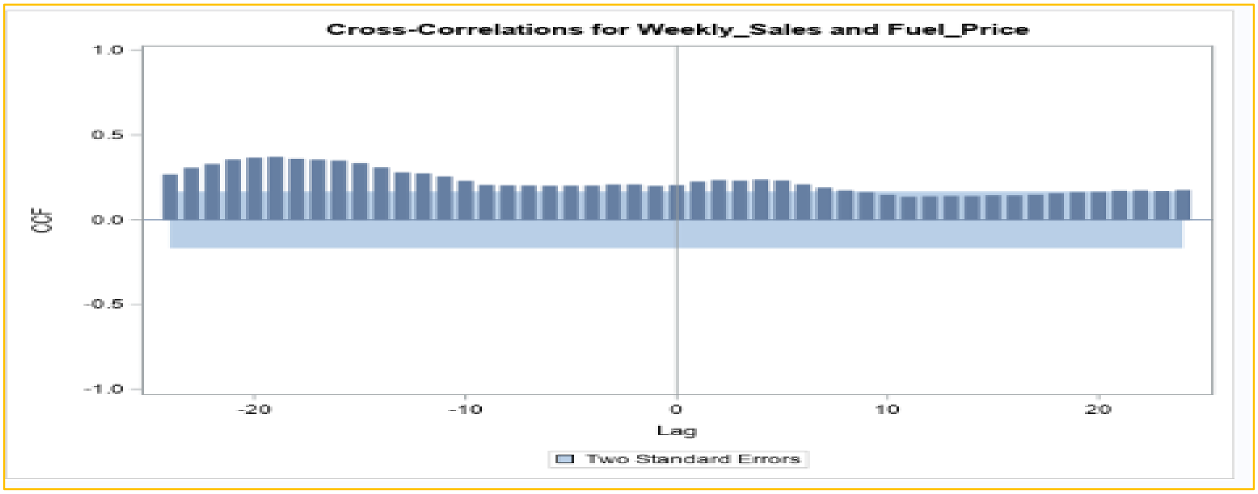
Now, let's check for correlations of individual attributes with weekly sales.

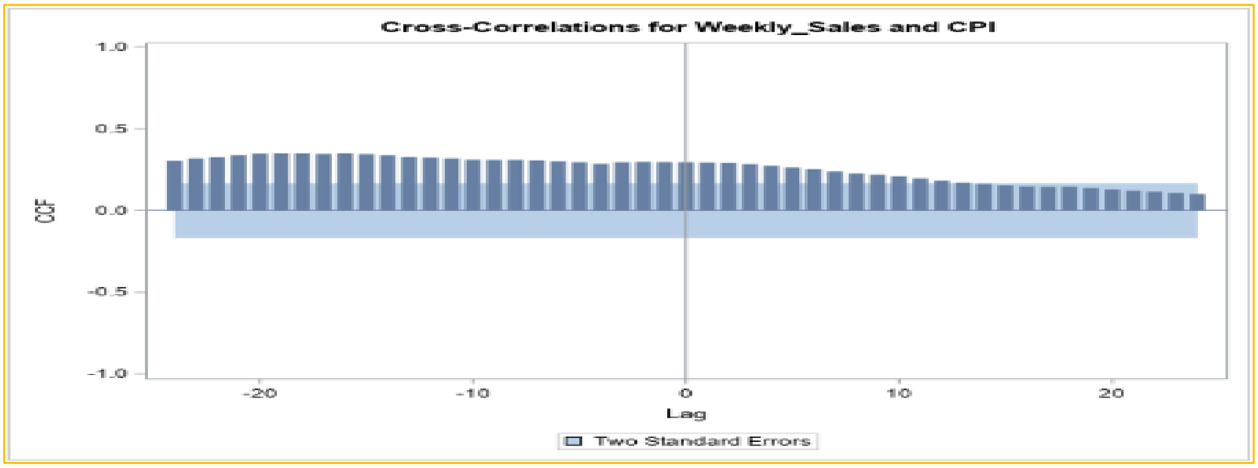


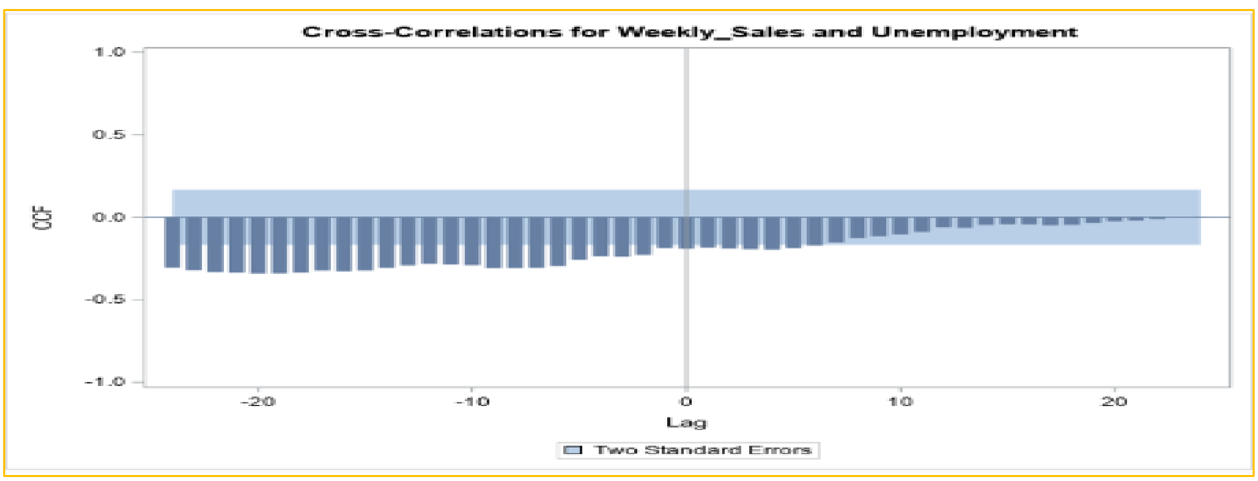
Cross-Correlations for Weekly_Sales and Holiday_Flag
1.0
0.5
0.0
-0.5
-1.0
CCF
-20
-10
0
10
20
Lag
Two Standard Errors



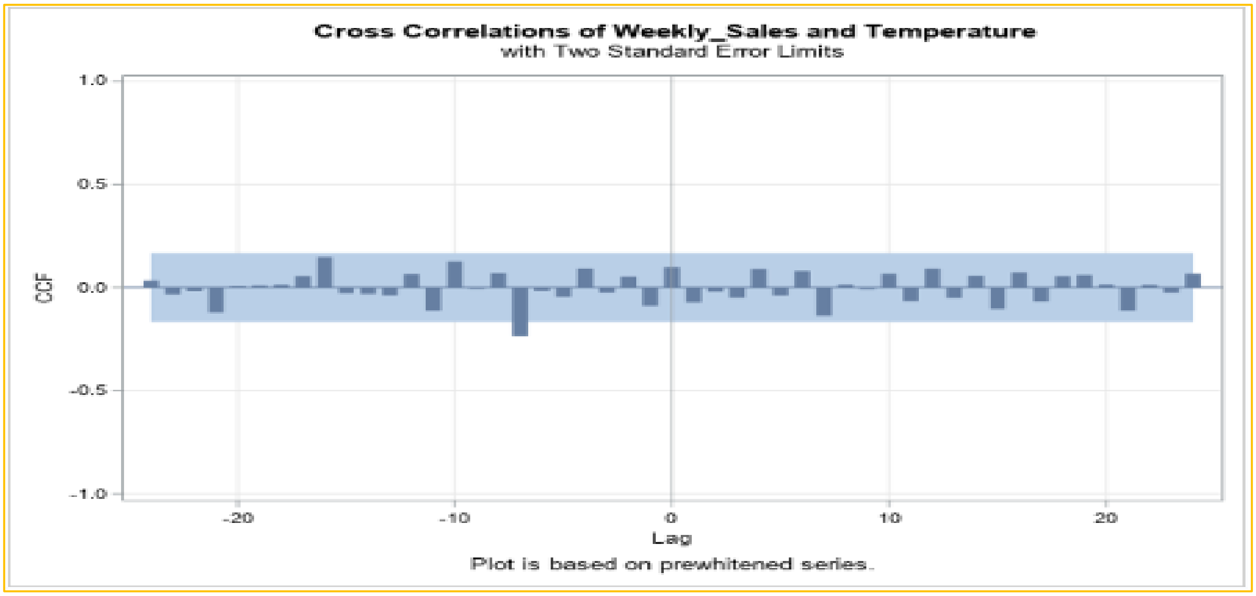
Cross-Correlations for Weekly_Sales and Temperature
1.0
0.5
0.0
-0.5
-1.0
CCF
-20
-10
0
10
20
Lag
Two Standard Errors

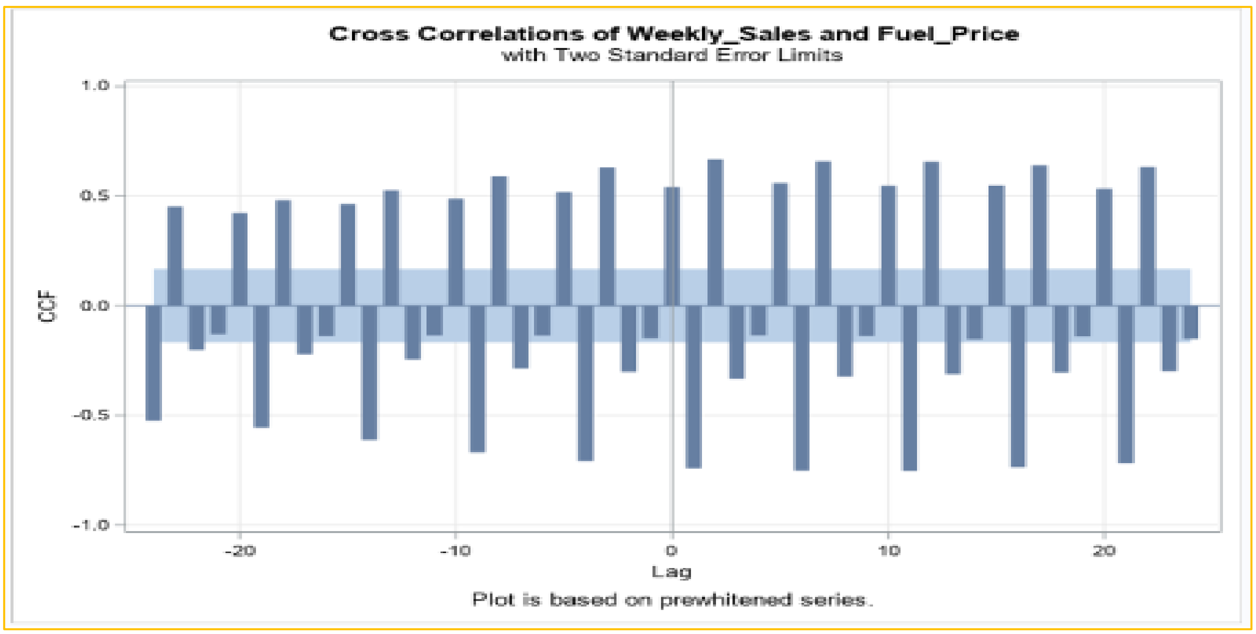

Cross-Correlations for Weekly_Sales and Fuel_Price
1.0
0.5
0.0
-0.5
-1.0
CCF
-20
-10
0
10
20
Lag
Two Standard Errors

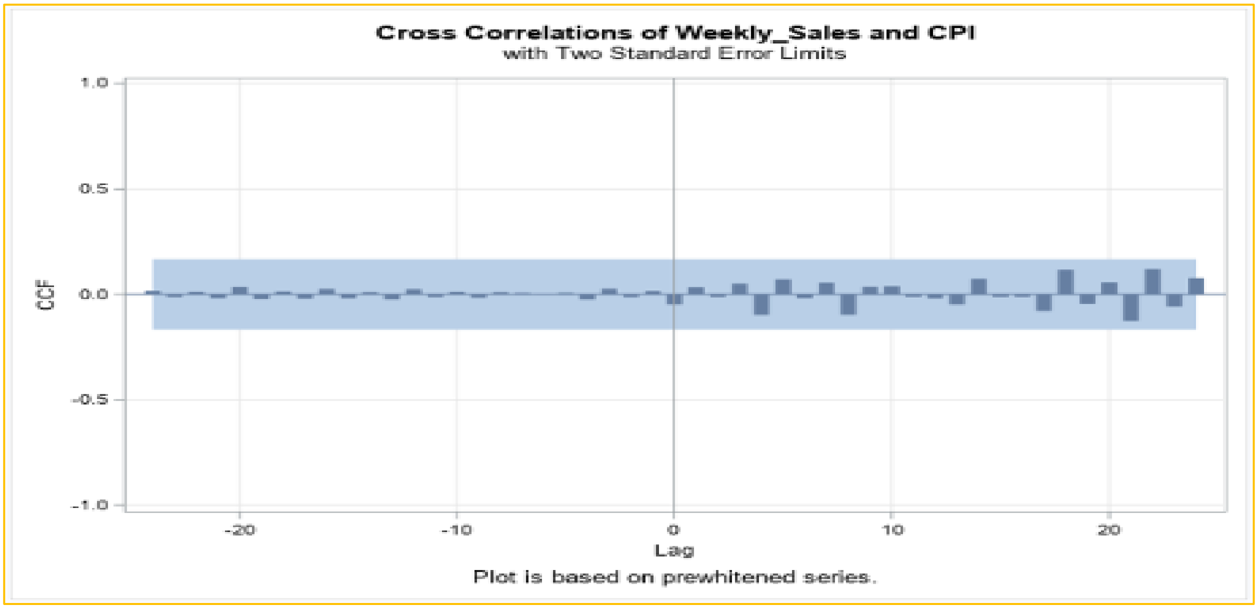

Cross-Correlations for Weekly_Sales and CPI
1.0
0.5
0.0
-0.5
-1.0
CCF
-20
-10
0
10
20
Lag
Two Standard Errors

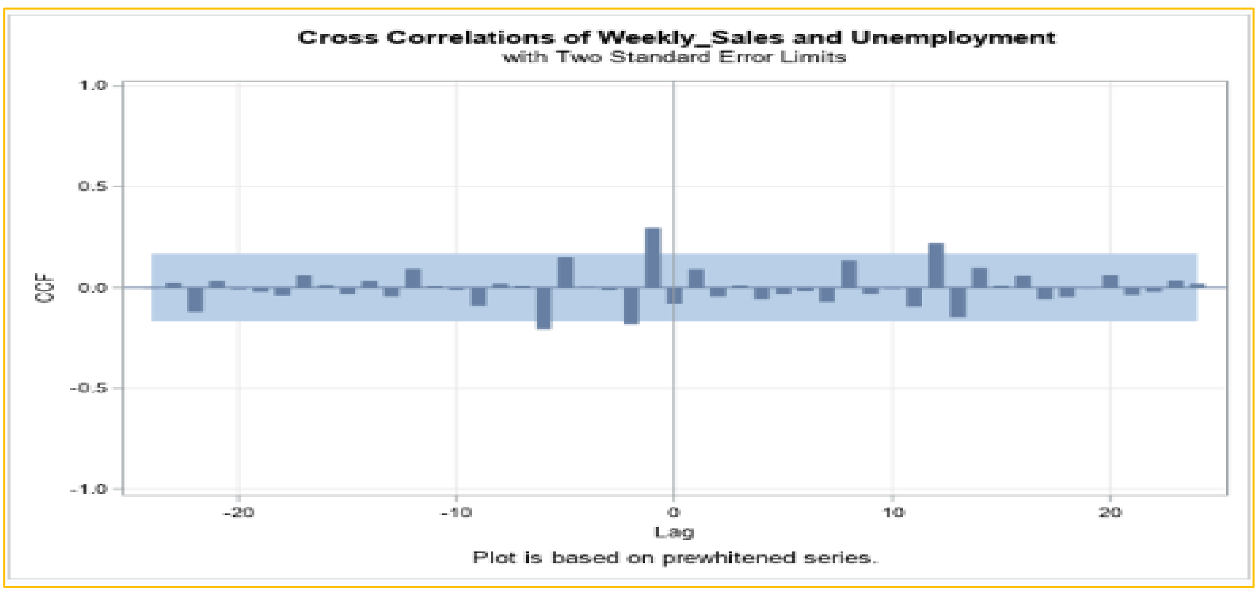

Cross-Correlations for Weekly_Sales and Unemployment
1.0
0.5
0.0
-0.5
-1.0
CCF
-20
-10
0
10
20
Lag
Two Standard Errors

We can observe from the above graphs that all the variables display cross-correlation with weekly sales. We have conducted pre-whitening to check for final cross-correlation of the attributes.







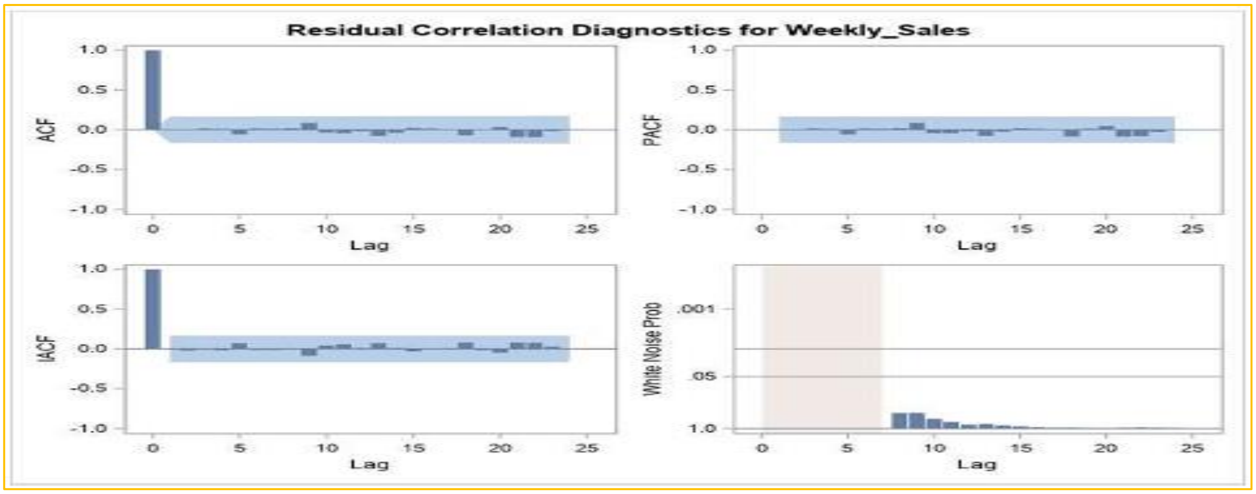


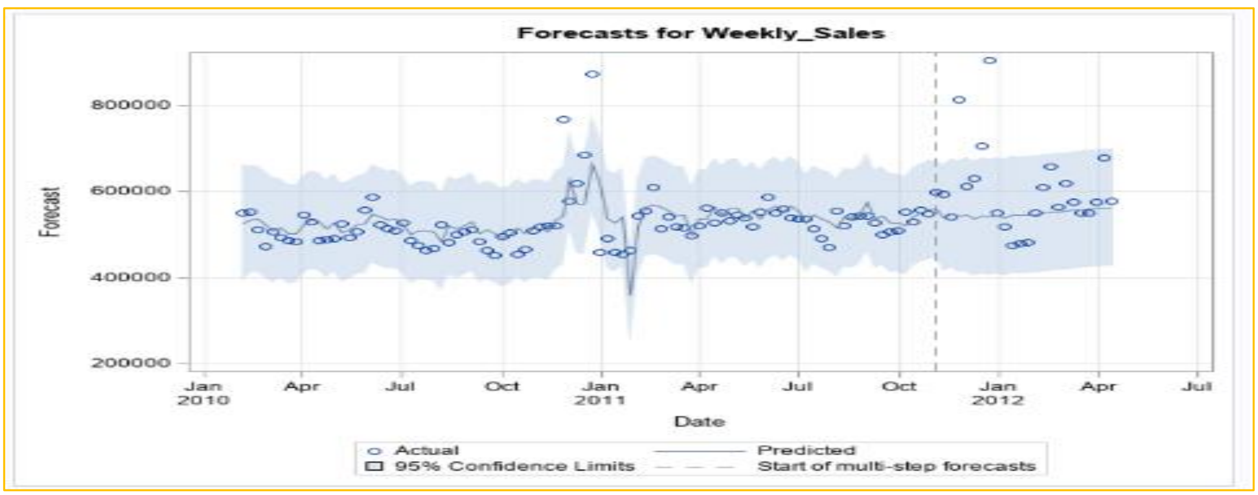
We can see from the above graphs that only Fuel price shows cross correlation with weekly sales after pre-whitening of the series. Apart from this, the holiday flag has shown a significant correlation at lag 4, which means that the weekly sales are being affected by a holiday 4 weeks back. This does not seem appropriate to us and has left out the holiday flag from the independent variables.

Model:

ARIMAX (5,0,3)

After data exploration, we have decided to use ARIMA (5,0,3) with fuel price as independent variable.

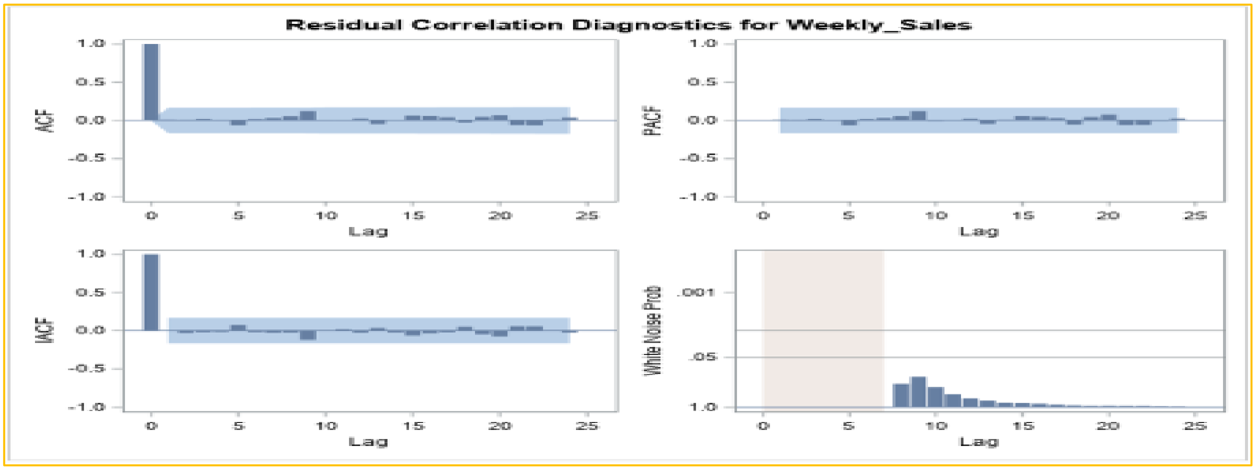


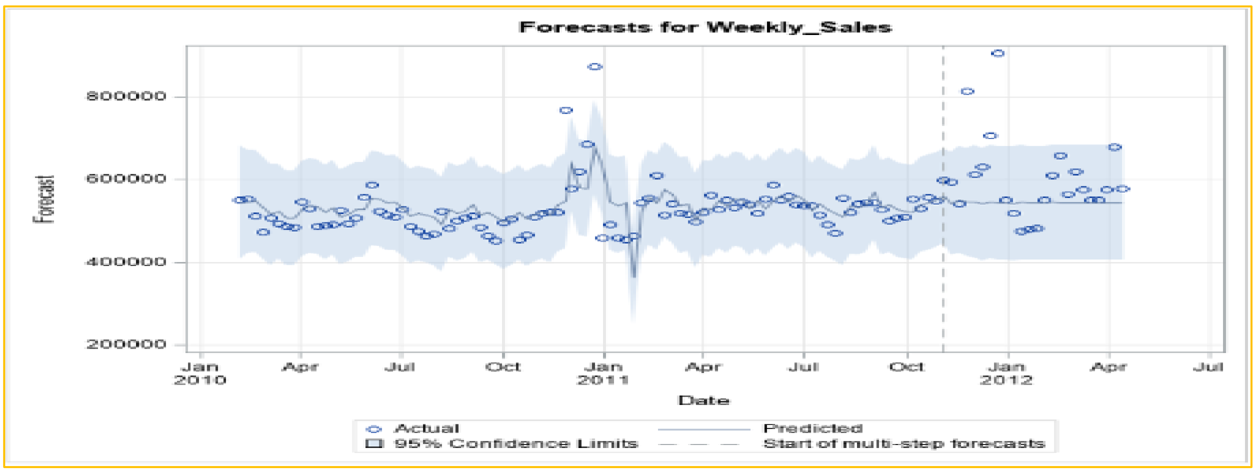


Stat	Value
AIC	3555.29
SBC	3584.9184
MAPE	93.65%

ARIMA (5,0,2)

To check if we can reduce the complexity of the model, we have decided to develop a model without fuel price as an independent model. The results of the model are as follows:





Stat	Value
AIC	3553.958969
SBC	3577.661727
MAPE	92.60%

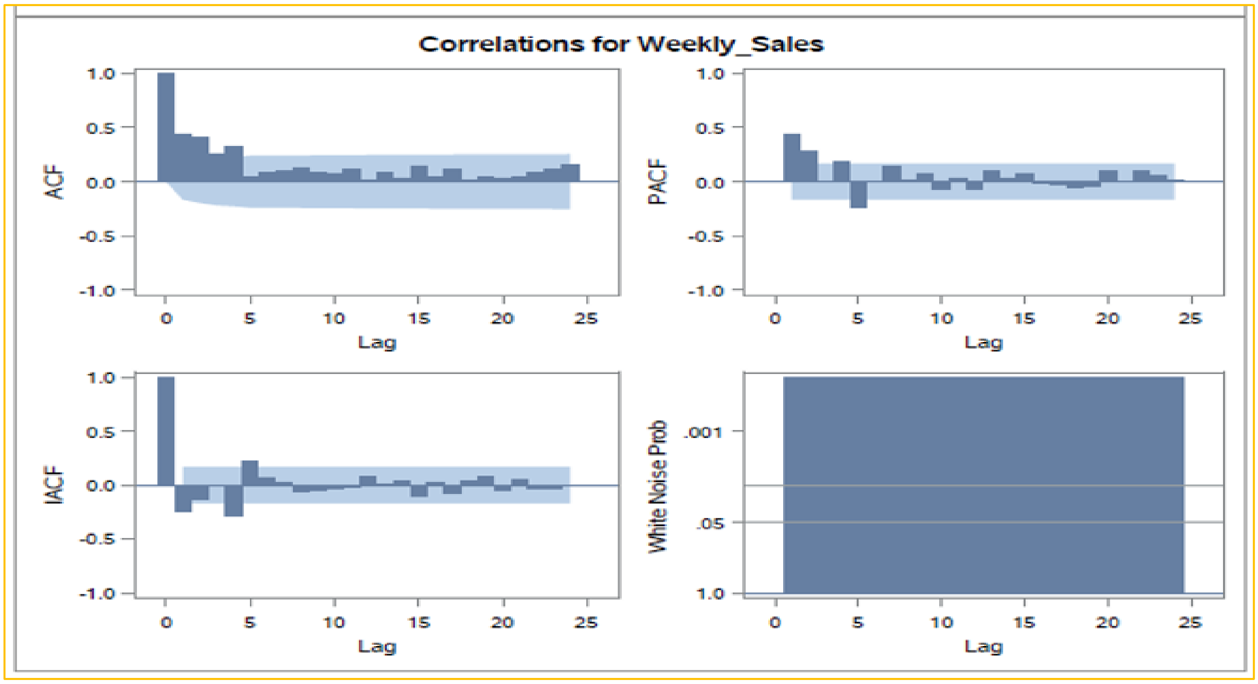
We can see from the above results that by decreasing the complexity of the model we did not have much difference in the accuracy. As a simpler model is better to use, we will recommend this model (ARIMA (5,0,2)) for final use for the stores with increasing trend.

Based on the above results, this is the **best model** for the current store. We have used the same model for different stores which have an increasing weekly sales trend, but our final interpretation is that each store despite its trend should have a different model to forecast its sales

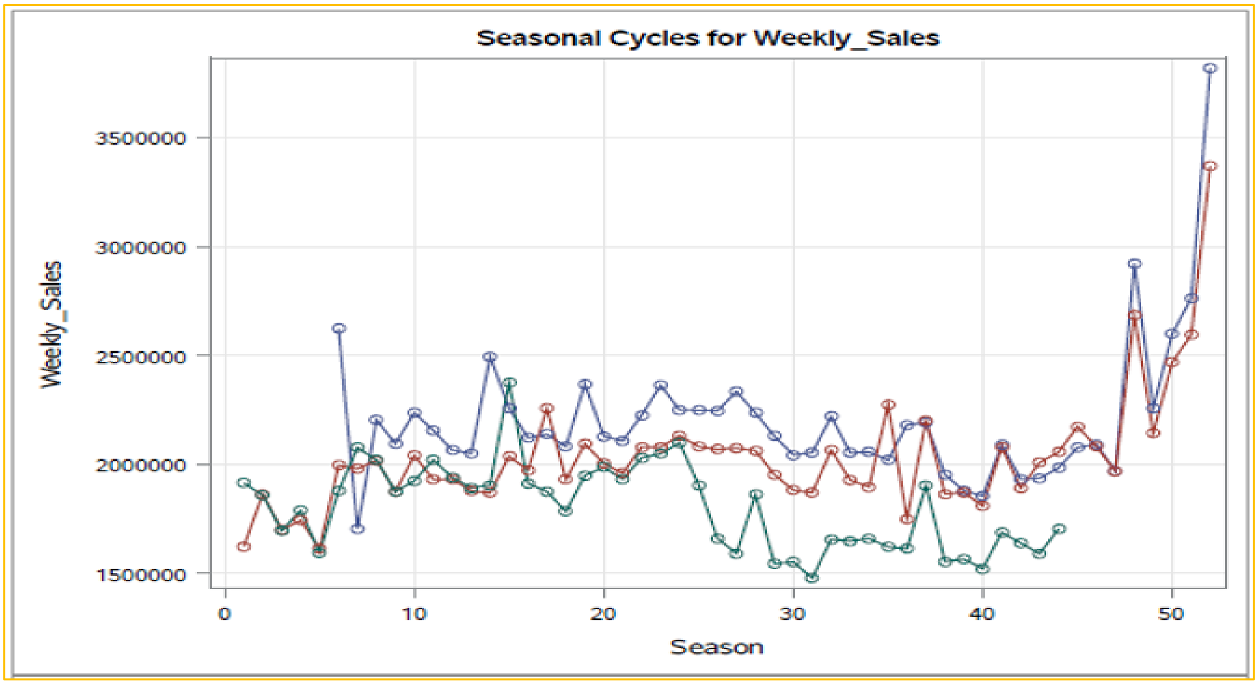
4.2 Model for Store #14

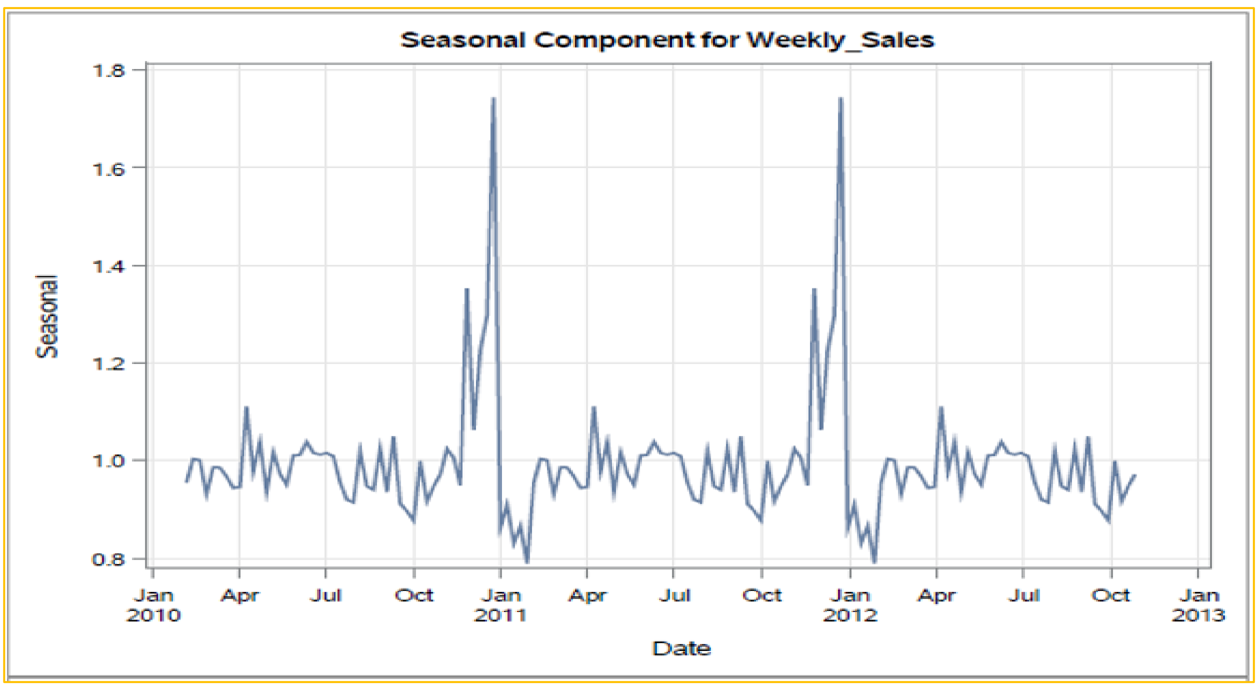
Exploration:

Weekly sales were checked to see if we can extract some information from the time series. Since the data is not white noise, we can extract information out of this:



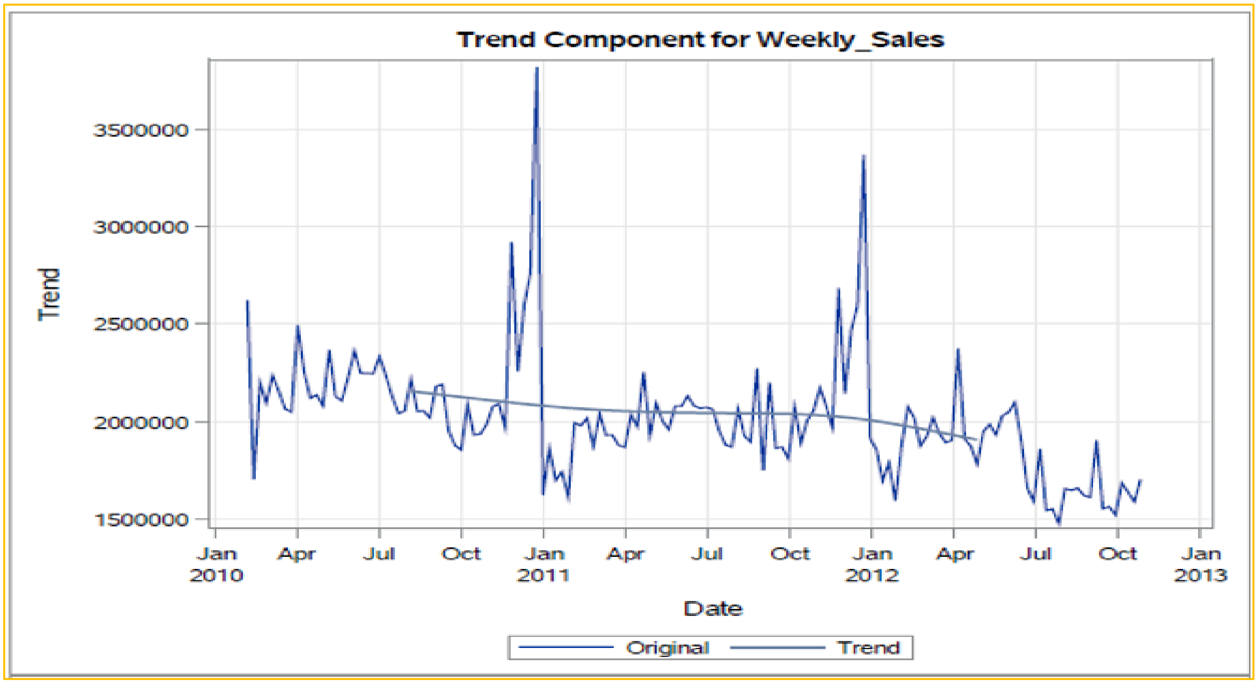
Seasonality:





From the graphs it is evident that there is some seasonality.

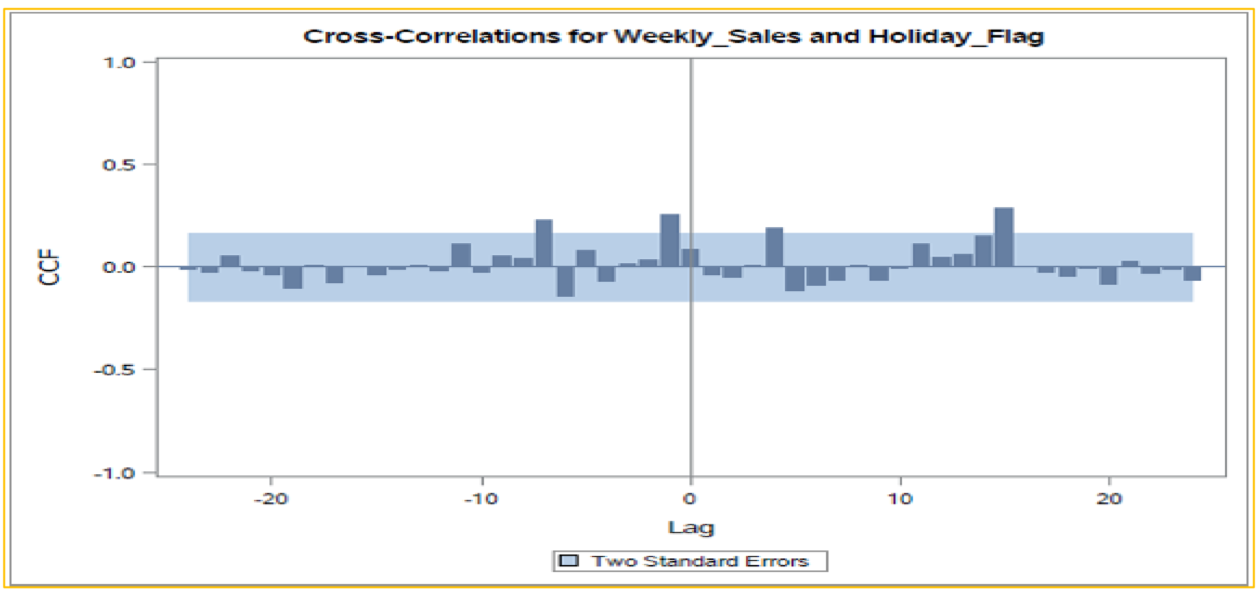
Trend:



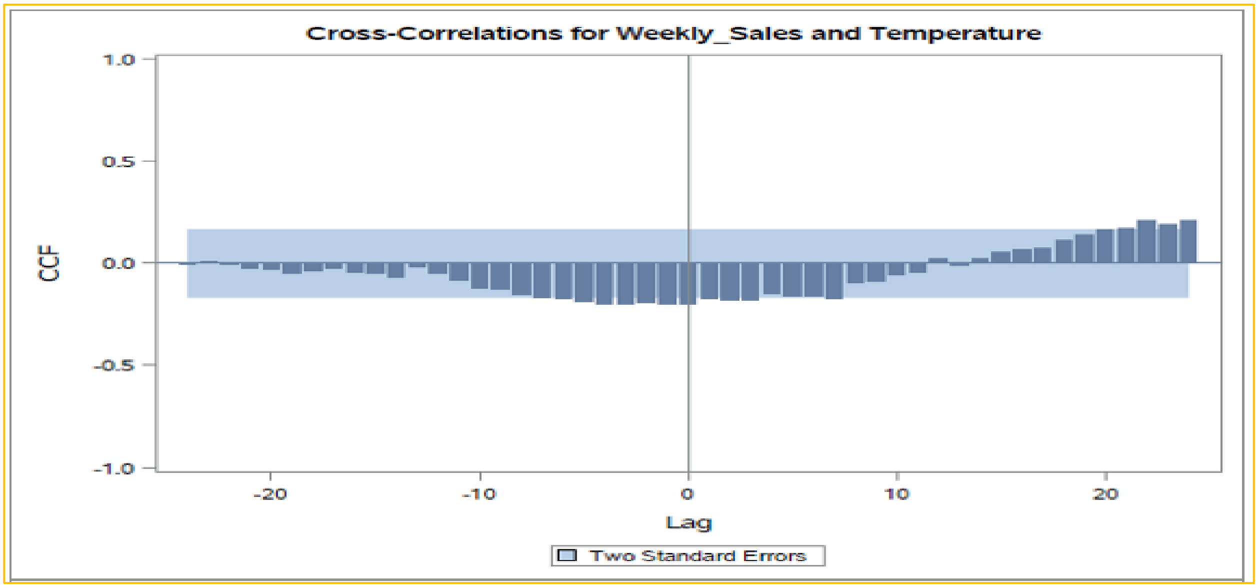
There is a negative trend of sales.

Correlations with explanatory variables:

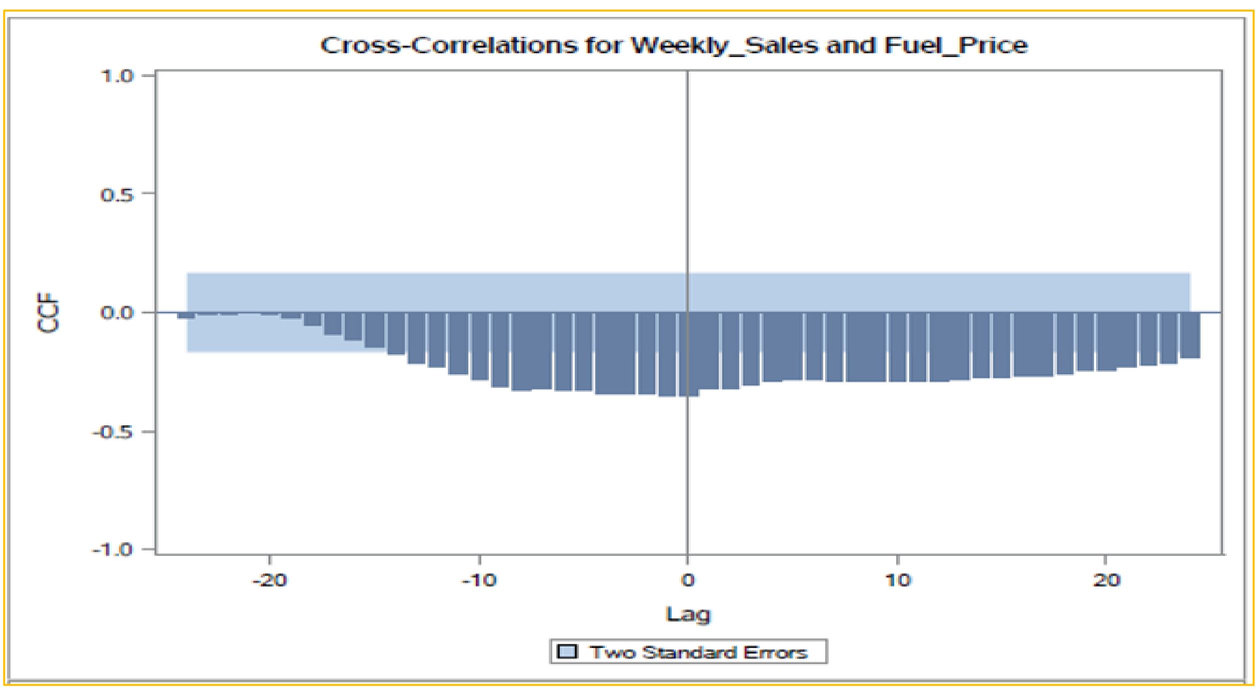
Lag 4 is Holiday affecting current week's sales. Based on reality a holiday before a month is not likely to affect sales in this week. Hence, Holiday will be considered insignificant.



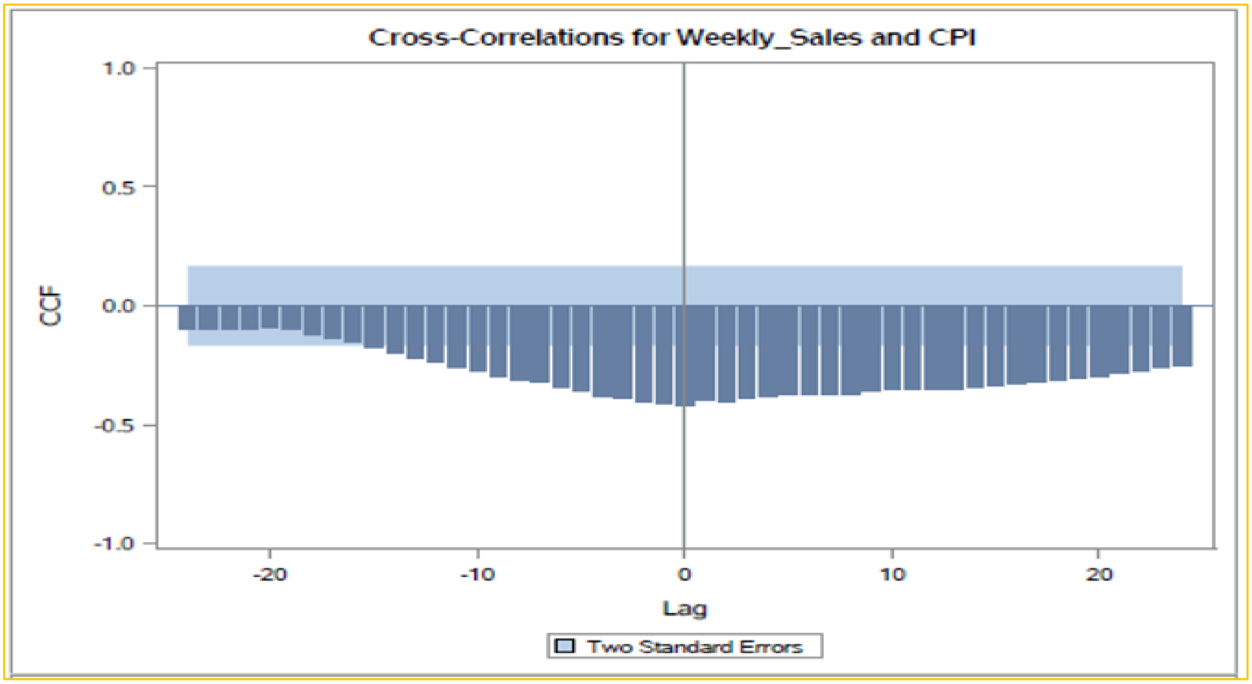
Temperature seems to influence sales. However, since temperature is a time series by itself, it needs to be further investigated by pre-whitening.



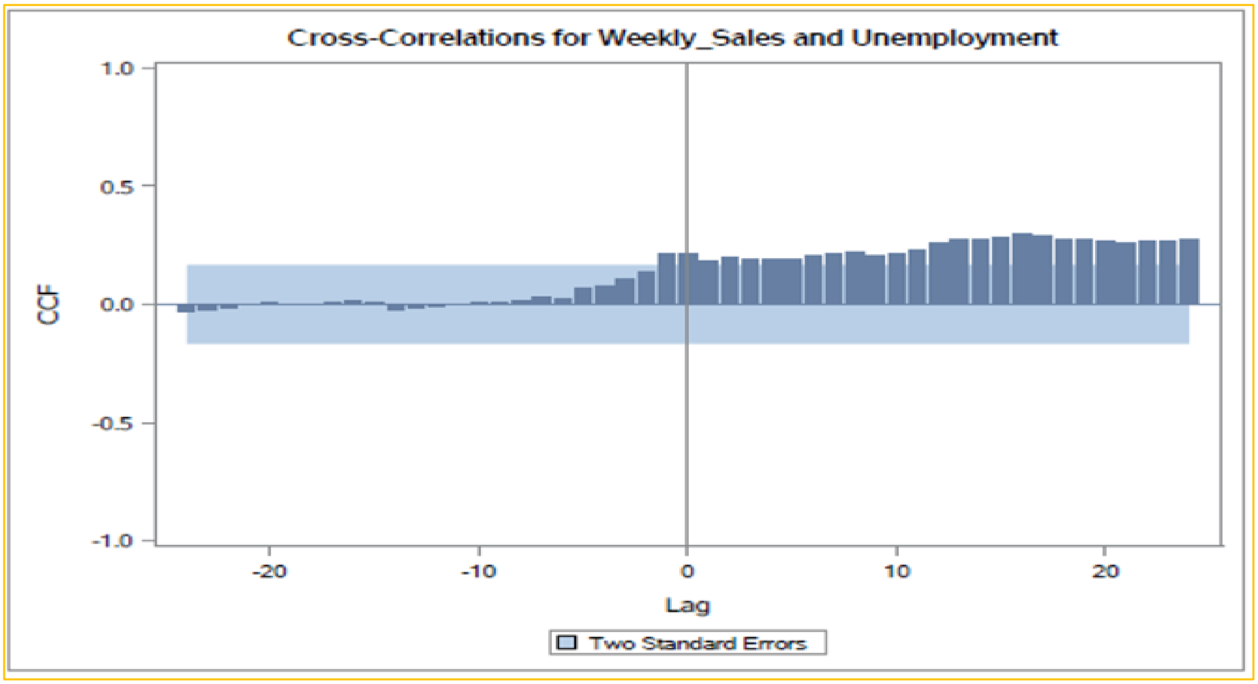
Fuel price seems to influence sales. However, since it is a time series on its own, it needs to be further investigated by pre-whitening.



CPI seems to influence sales. However, since it is a time series on its own, it needs to be further investigated by pre-whitening.



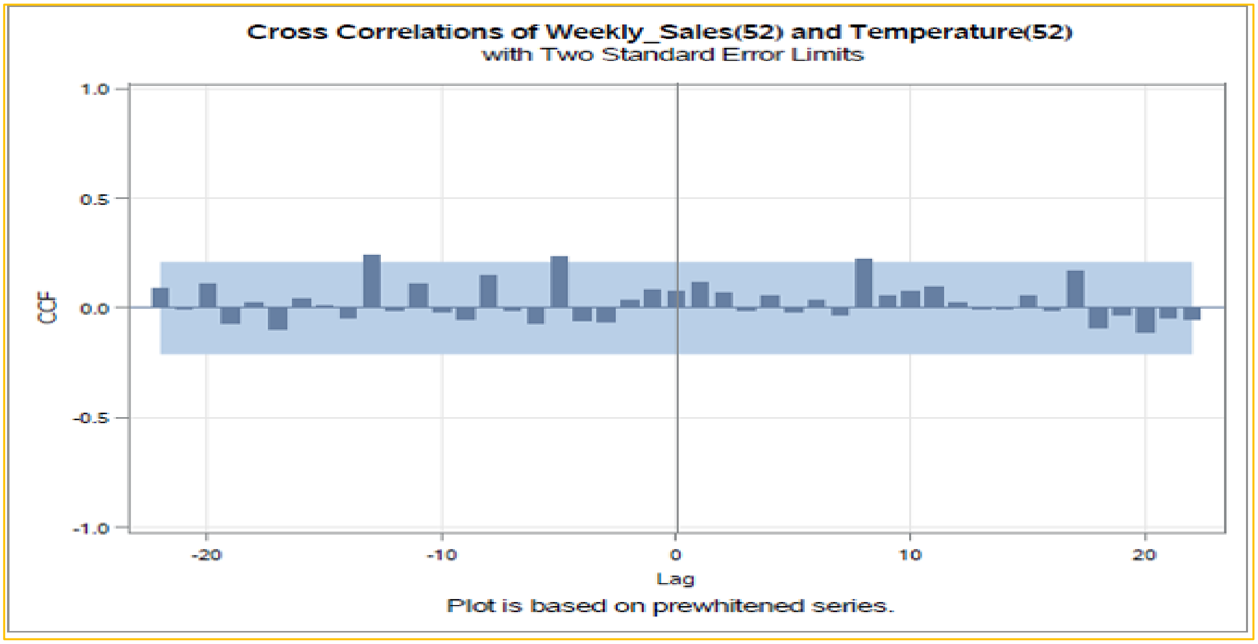
Unemployment seems to influence sales. However, since it is a time series on its own, it needs to be further investigated by pre-whitening.



Based on Dickey fuller test, data is stationary. Hence, we can use either Exponential smoothing or ARIMA models can be used.

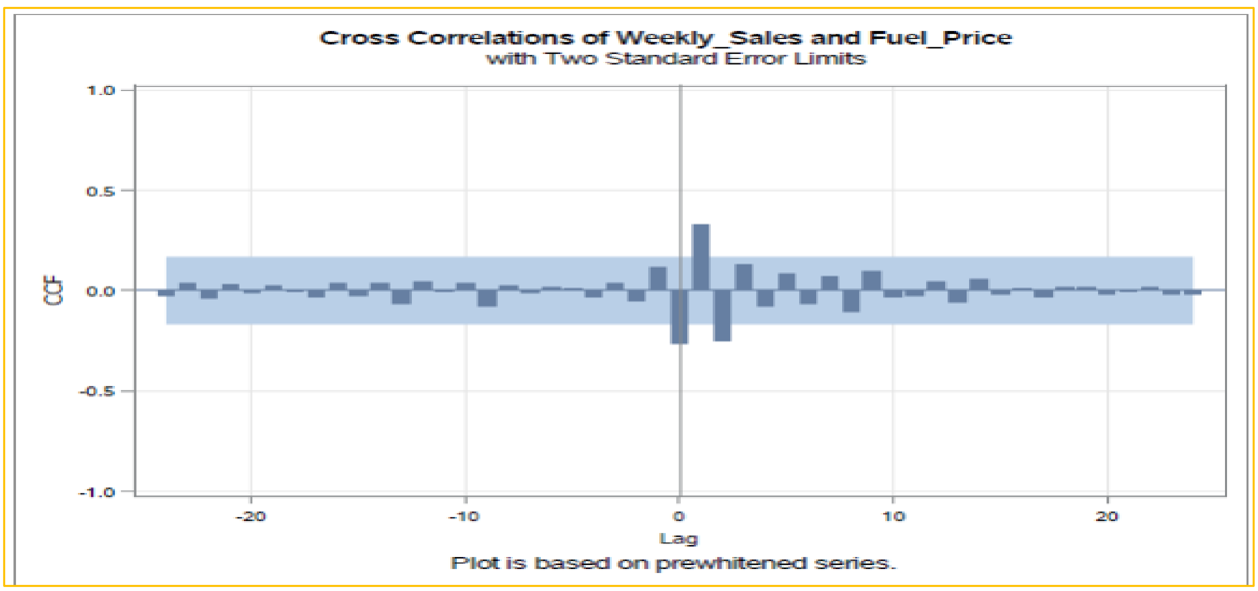
Pre-whitening test:

<u>Sales Vs Temperature</u>



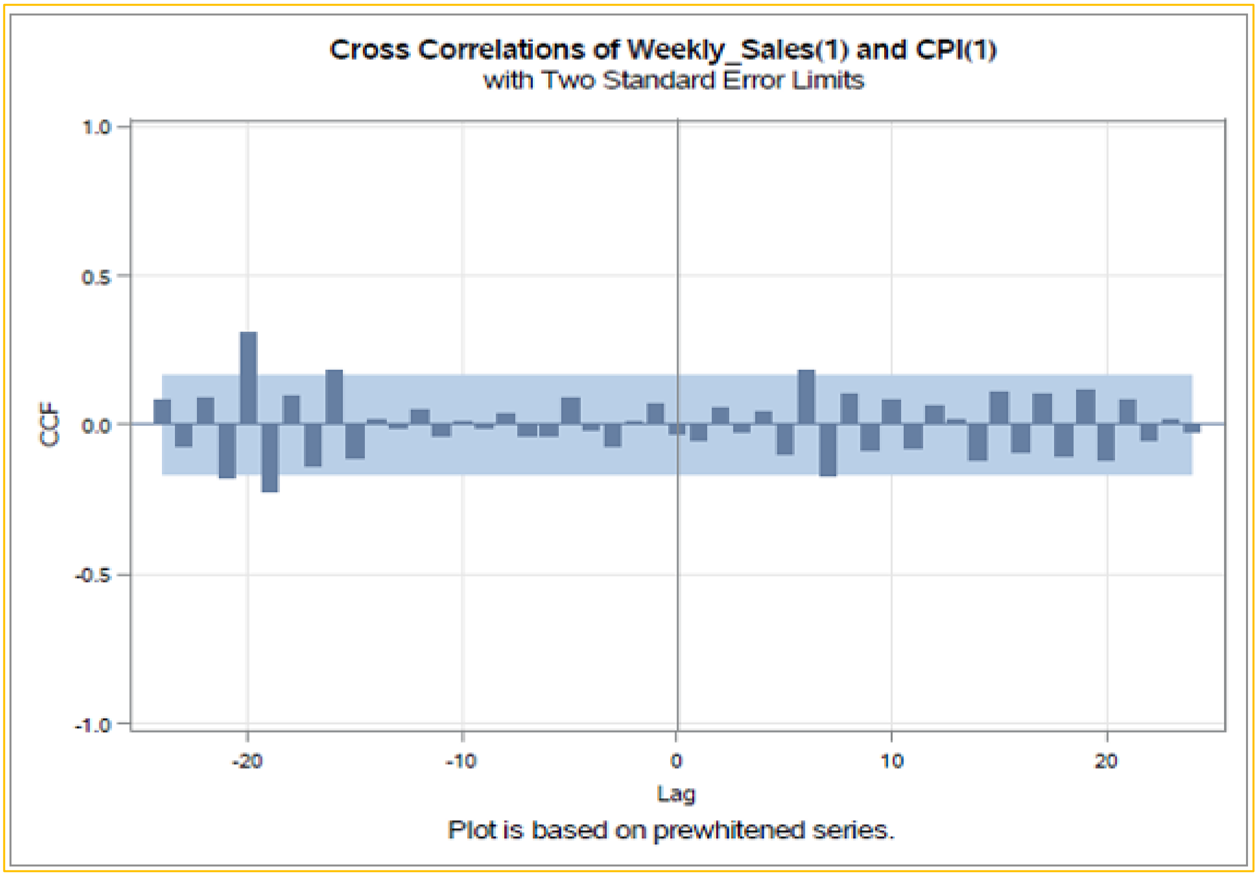
Based on this temperature of week 8 weeks behind is affect current weeks sales. This, not realistic. Hence, temperature will be considered to not affect sales.

<u>Sales Vs Fuel Price</u>



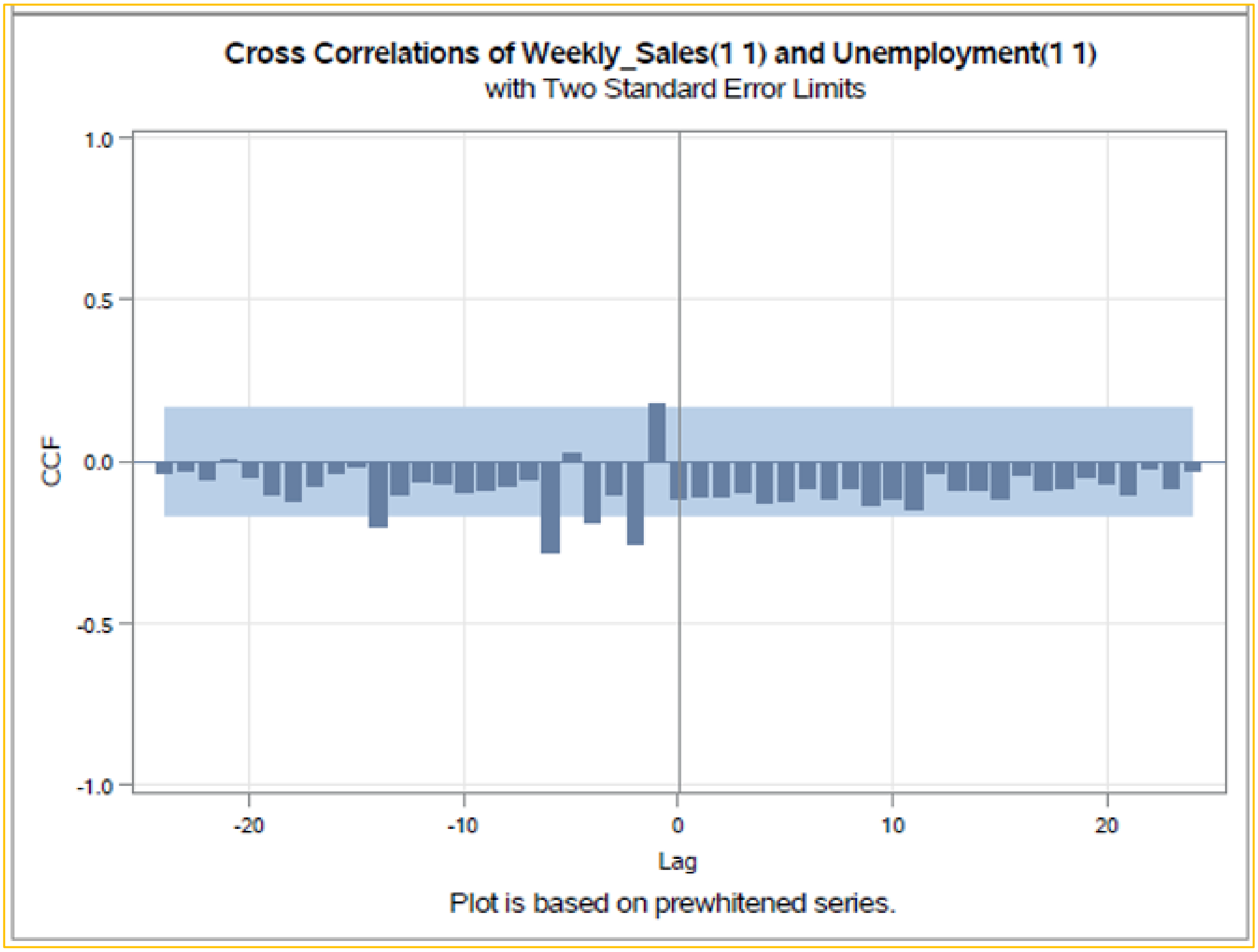
Based on this, Fuel price from 3 weeks is affecting Sales. The ARIMAX model will be developed using Fuel price as explanatory variable.

Sales Vs CPI



Based on this, CPI at lag 6 is affecting sales in the current week. Hence, it will be used to model an ARIMAX model.

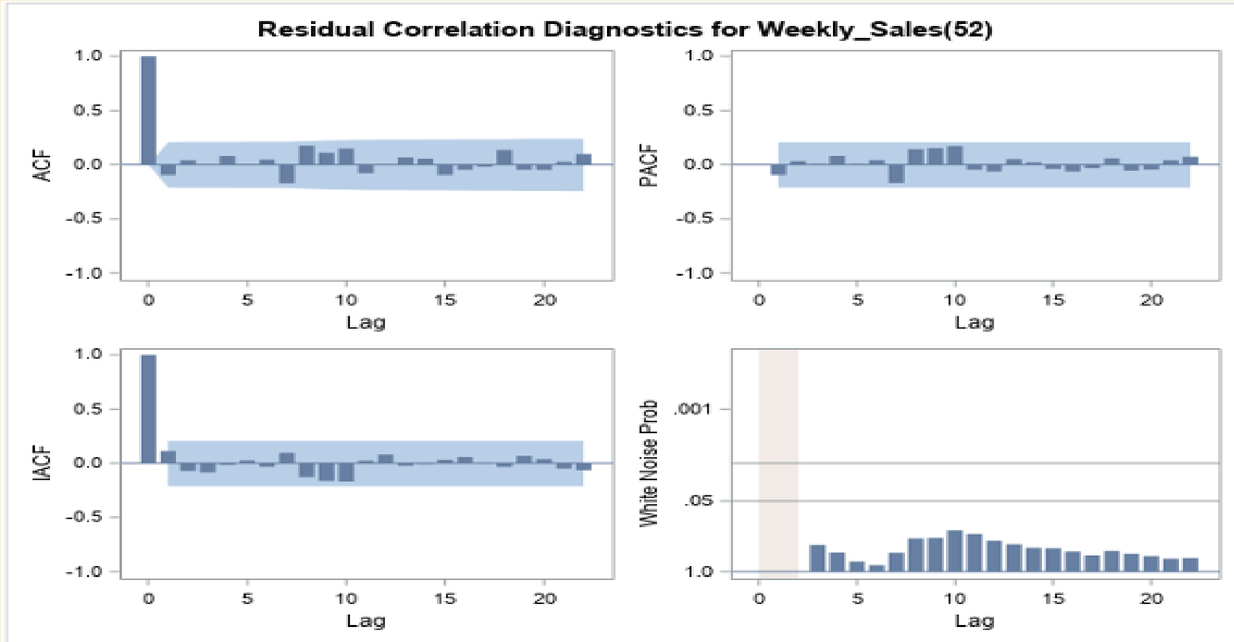
Sales VS Unemployment



Based on this, unemployment is not significant.

ARIMA

CPI and Fuel price were considered to arrive at an ARIMAX model (0,0,2,0,1,0).



Based on the white noise probability graph, we have extracted sufficient signal from the model.

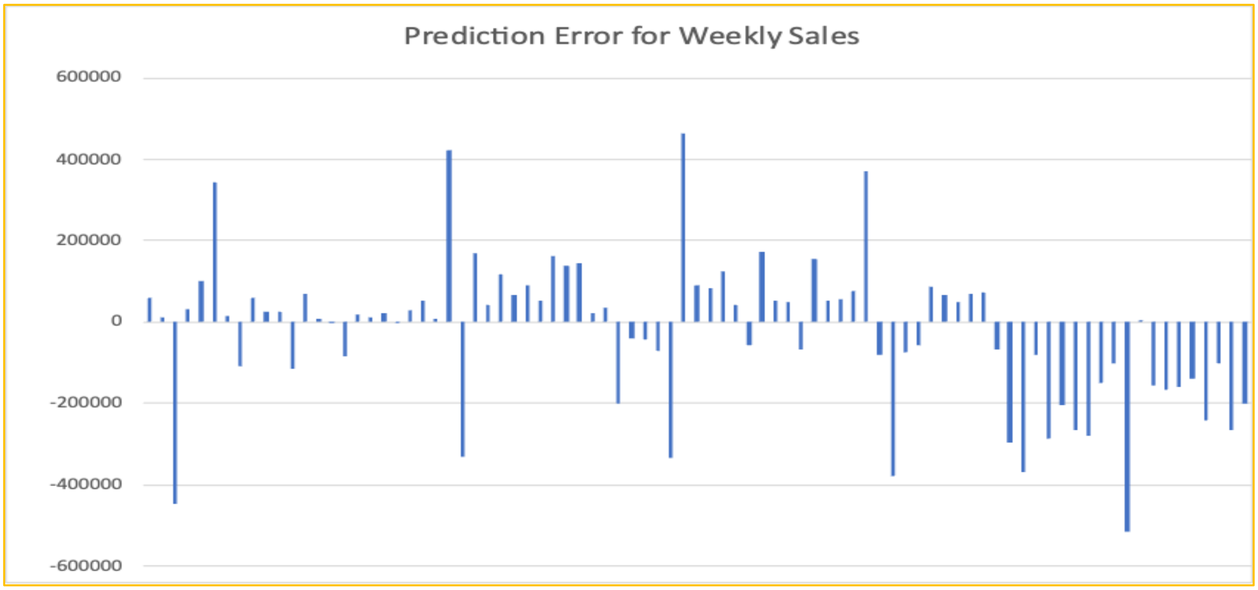
Model metrics:

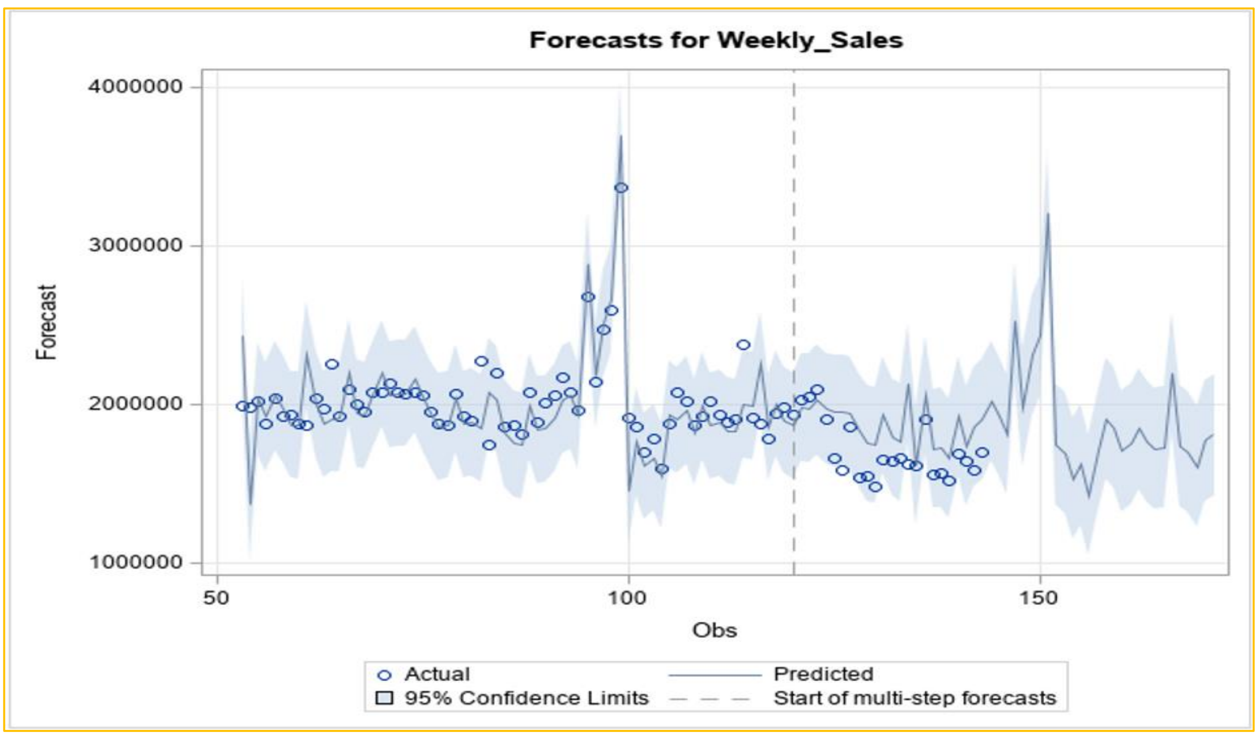
Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	-183122.3	73446.5	-2.49	0.0127	0
MA1,1	0.79080	0.11017	7.18	<.0001	1
AR1,1	0.95383	0.05486	17.39	<.0001	1

Constant Estimate	-8454.89
Variance Estimate	2.957E10
Std Error Estimate	171973.3
AIC	2455.804
SBC	2463.336
Number of Residuals	91

MAPE= 6.85

RMSE= 177740.2

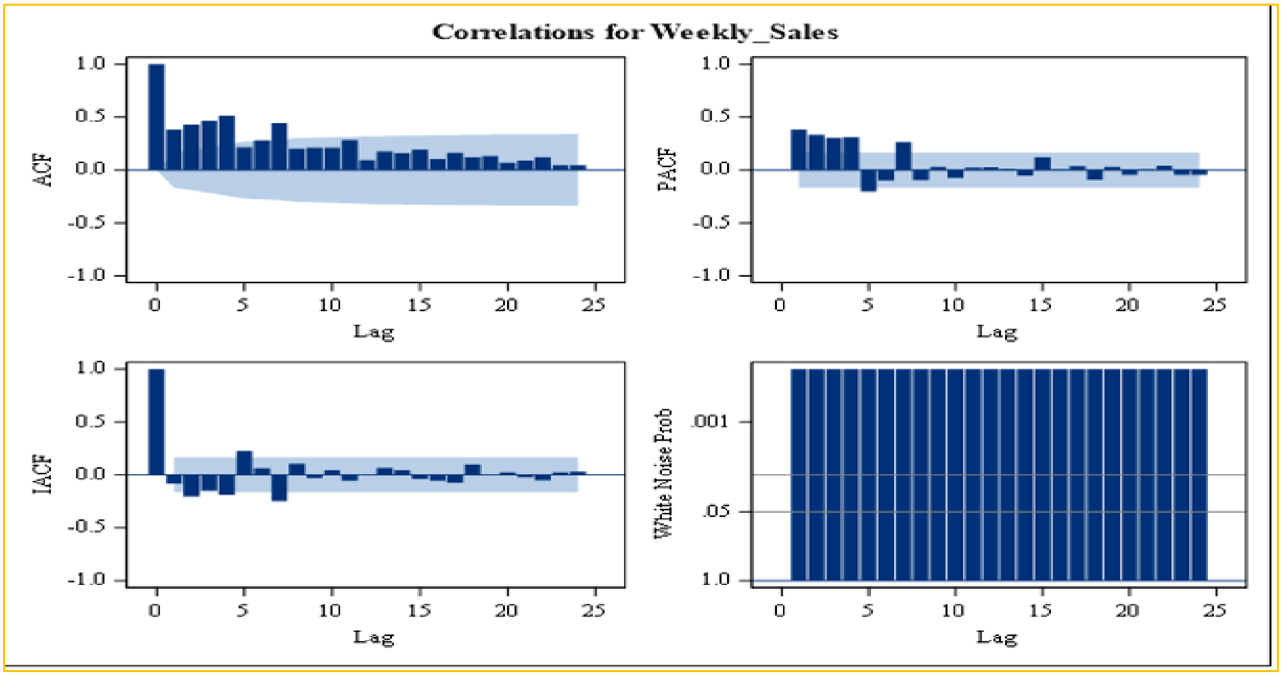




As can be seen from the metrics and graph, the ARIMAX model was not a good model when compared to the simple ARIMA model. Hence, the final model was ARIMA (1,0,1)(0,1,0).

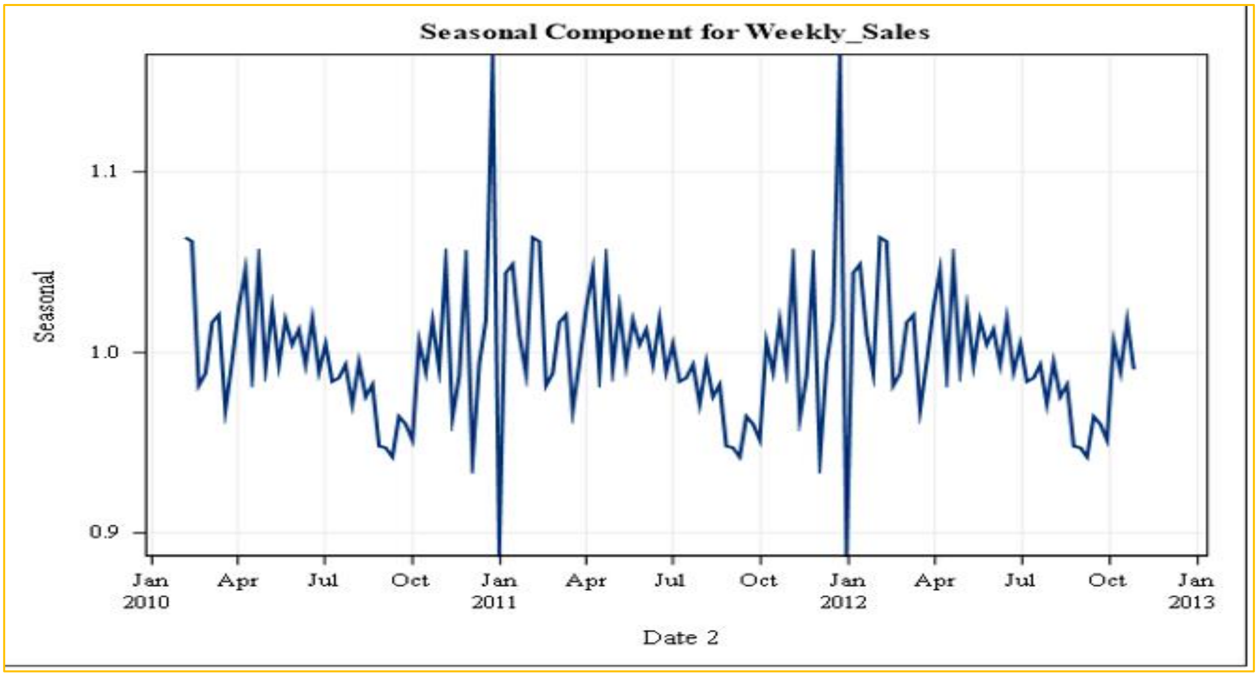
4.3 Model for Store #30

Exploration:



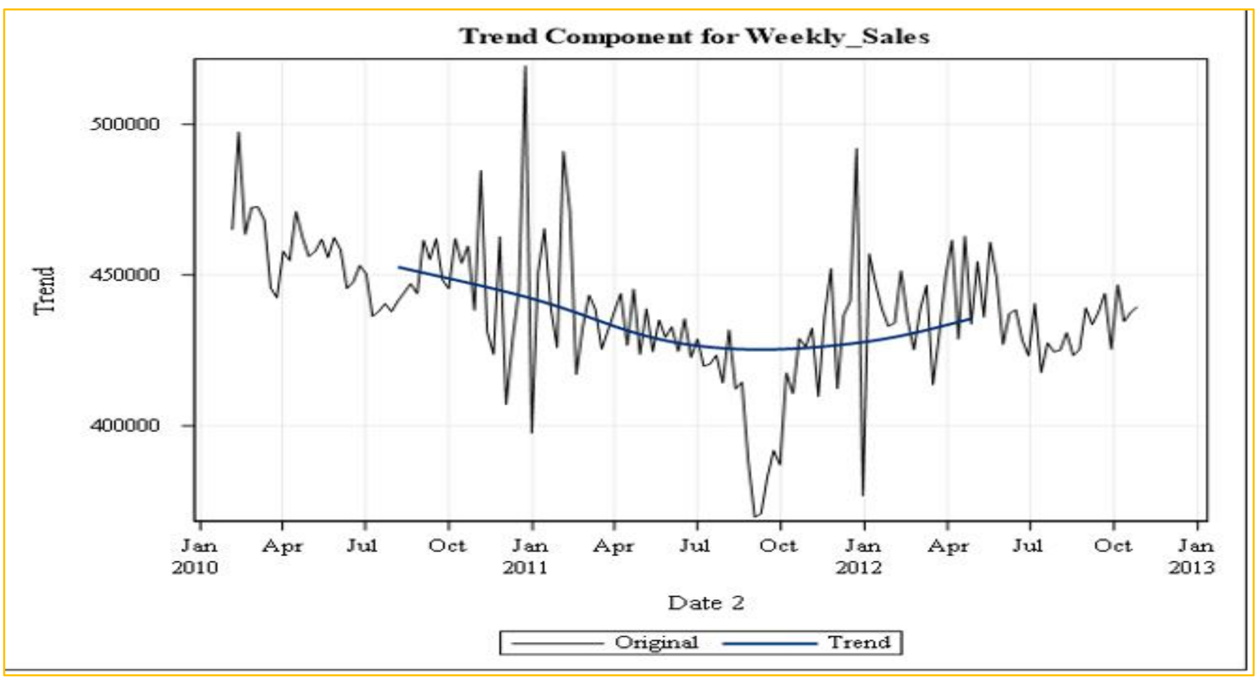
Weekly sales were checked to see if we can extract some information from the time series. Since, the data is not white noise, we can extract information out of this.

Seasonality:



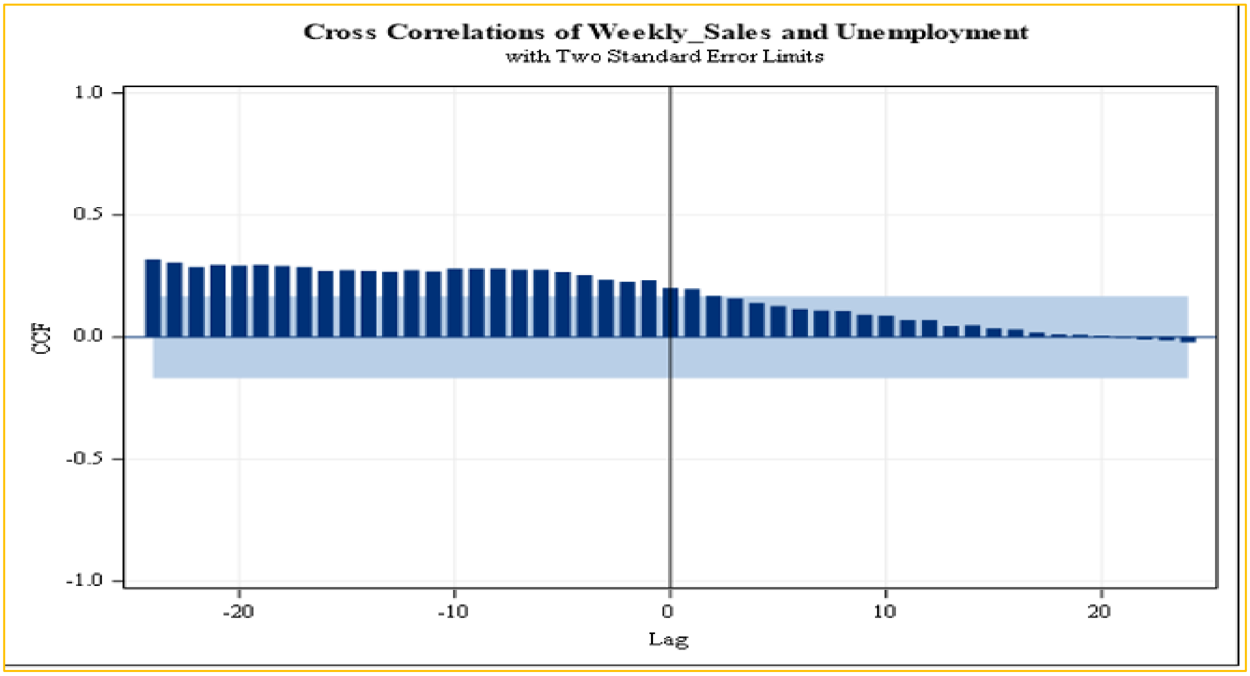
From the graphs it is evident that there is some seasonality.

Trend:

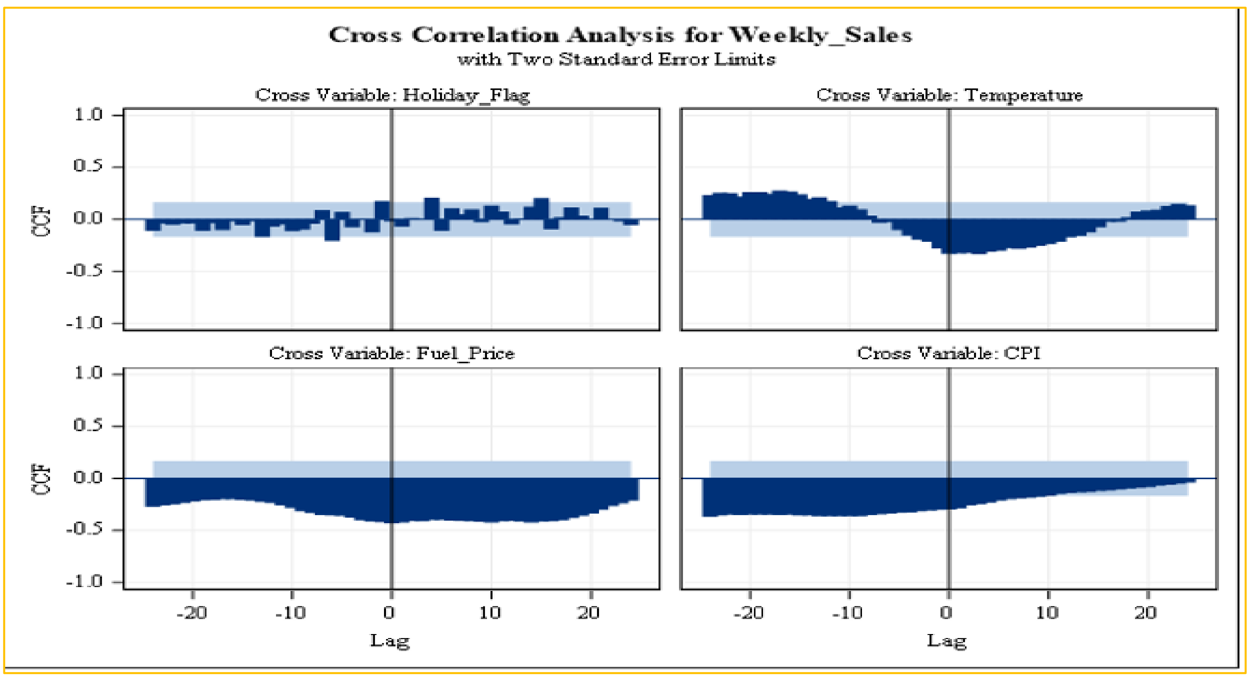


There is no significant trend of sales by going down first then going up.

Correlations with explanatory variables:



Lag 1 is Unemployment affecting current week's sales. Based on the graph, unemployment rate one month before seems to have same influence on the weekly sales with the current month. Hence, unemployment will be considered insignificant.



For Holiday_Flag, as there is no obvious decay, and not too many lags are significant, it may not be considered as a significant variable. For Temperature, Fuel_Price, and CPI, although there are decays, those lags' effects seem to be not stronger than the current weeks, so they may not be considered as significant variables.

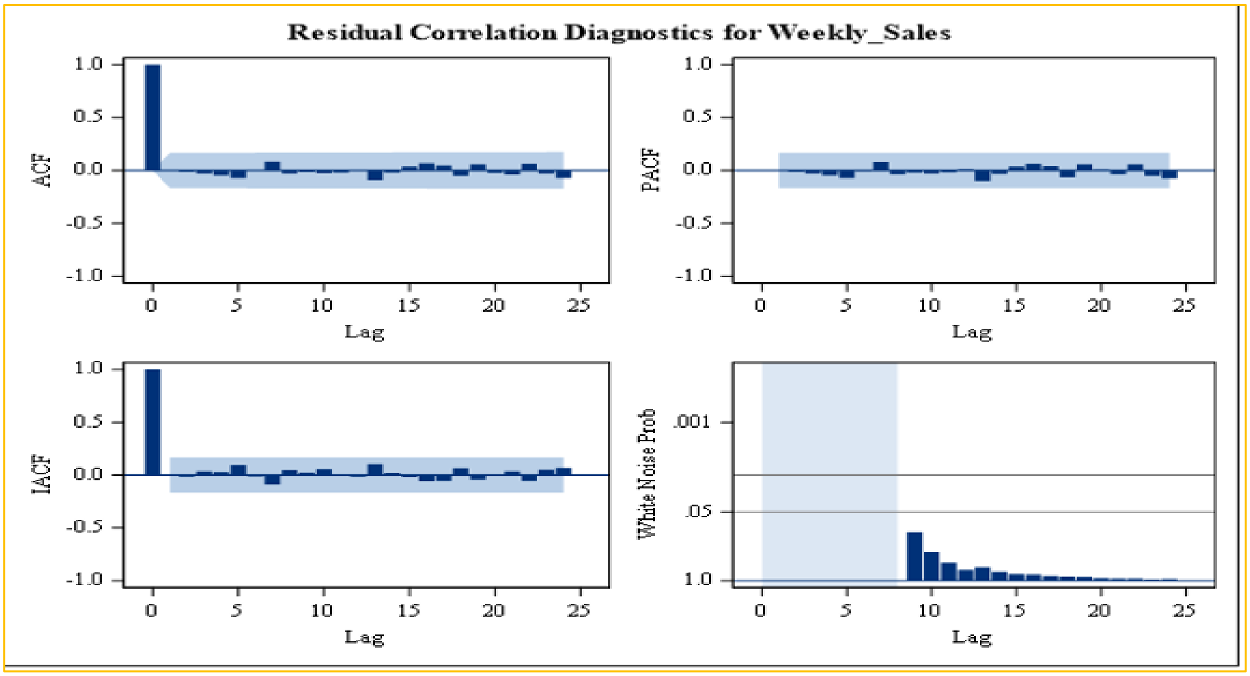
Unit Root Test

Augmented Dickey-Fuller Unit Root Tests							
Type	**Lags**	**Rho**	**Pr < Rho**	**Tau**	**Pr < Tau**	**F**	**Pr > F**
Zero Mean	0	-0.2946	0.6147	-0.43	0.5260		
	1	-0.1797	0.6408	-0.49	0.5016		
Single Mean	0	-87.5621	0.0012	-7.94	<.0001	31.56	0.0010
	1	-44.6287	0.0012	-4.91	0.0001	12.12	0.0010
Trend	0	-101.849	0.0001	-8.81	<.0001	38.80	0.0010
	1	-55.4554	0.0005	-5.30	0.0001	14.18	0.0010

According to the Dickey-Fuller Unit Root Tests, there is no unit roots here.

ARIMA

As no variables are significant enough, ARIMA model shall be suitable for this store.



Based on the white noise probability graph, we have extracted sufficient signal from the model.

Model metrics:

Constant Estimate	296927.1
Variance Estimate	2.8127E8
Std Error Estimate	16771.02
AIC	3198.994
SBC	3225.659
Number of Residuals	143

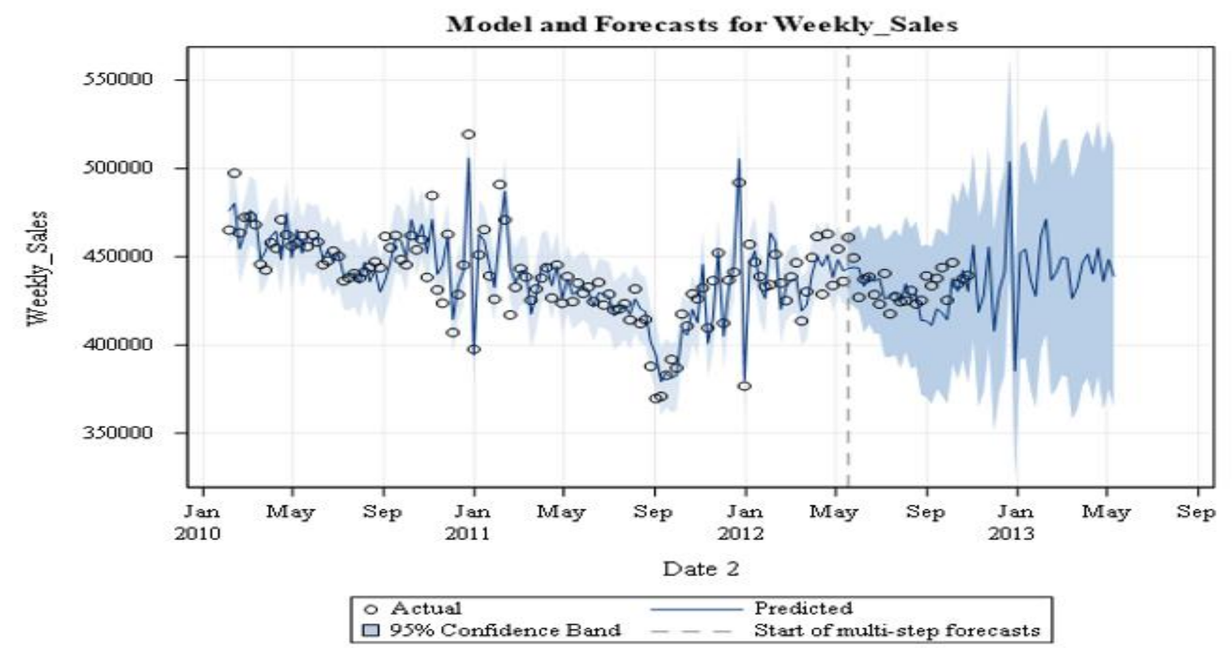
MAPE= -0.55%

RMSE= 16884.7809

Additive seasonal exponential smoothing model

Model Metrics:

Obs	_NAME_	_REGION_	N	RMSE	MAPE	AIC	SBC
1	Weekly_Sales	FIT	119	9866.33	1.76589	2192.86	2198.42
2	Weekly_Sales	FORECAST	24	12744.98	2.42471	453.74	453.74



Based on the above models it was concluded that ARIMA(5,0,3) model is better.

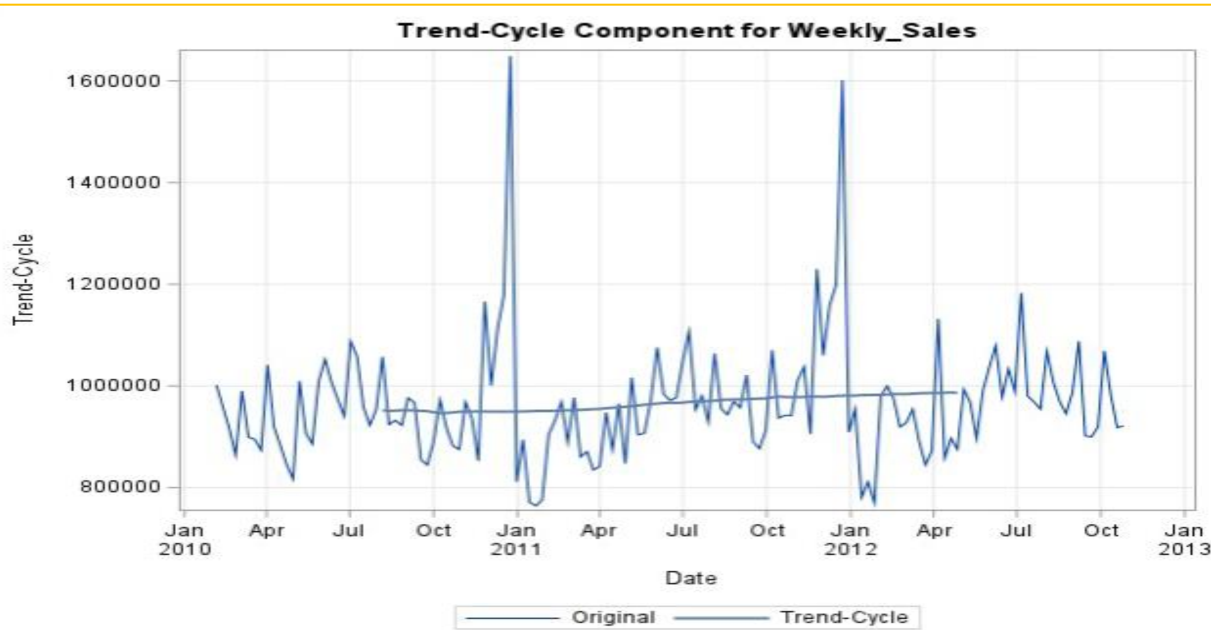
4.4 Model for Store #40

Time Series Exploration:

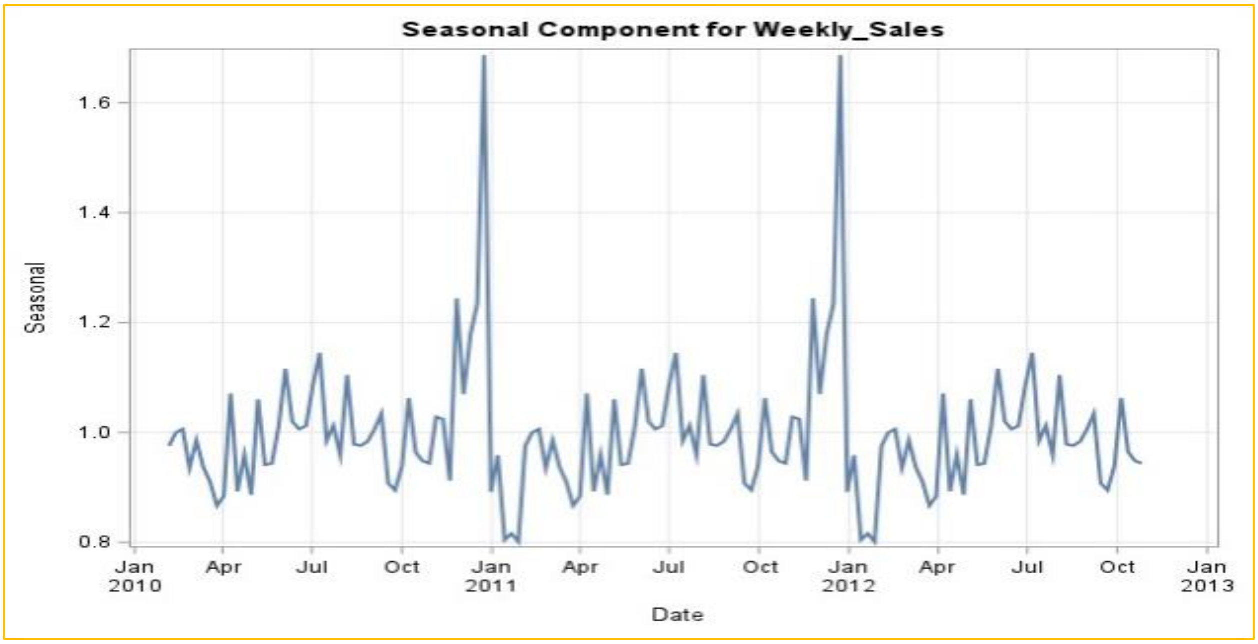
In ADF test, we found that p-value is less than 0.0001 and thus less than 0.05. So, we concluded that it is a stationary series.

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-1.6194	0.3775	-0.92	0.3151		
	1	-0.6023	0.5470	-0.56	0.4743		
	2	-0.4118	0.5883	-0.43	0.5249		
Single Mean	0	-101.884	0.0001	-8.87	<.0001	39.31	0.0010
	1	-75.2470	0.0012	-6.06	<.0001	18.38	0.0010
	2	-98.0822	0.0012	-5.95	<.0001	17.71	0.0010
Trend	0	-103.424	0.0001	-8.94	<.0001	39.95	0.0010
	1	-77.4697	0.0005	-6.12	<.0001	18.71	0.0010
	2	-102.631	0.0001	-6.00	<.0001	18.04	0.0010

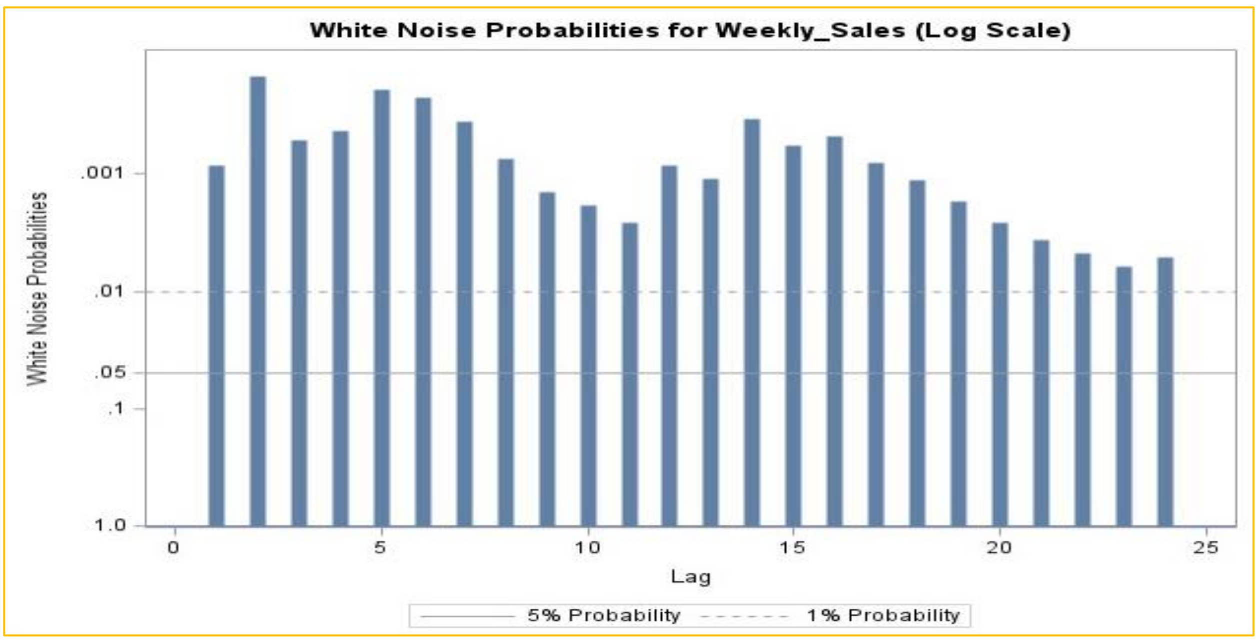
From the exploration graph of the trend, it is observed that the trend is flat, and it is a positive trend as it is not oscillating directions.



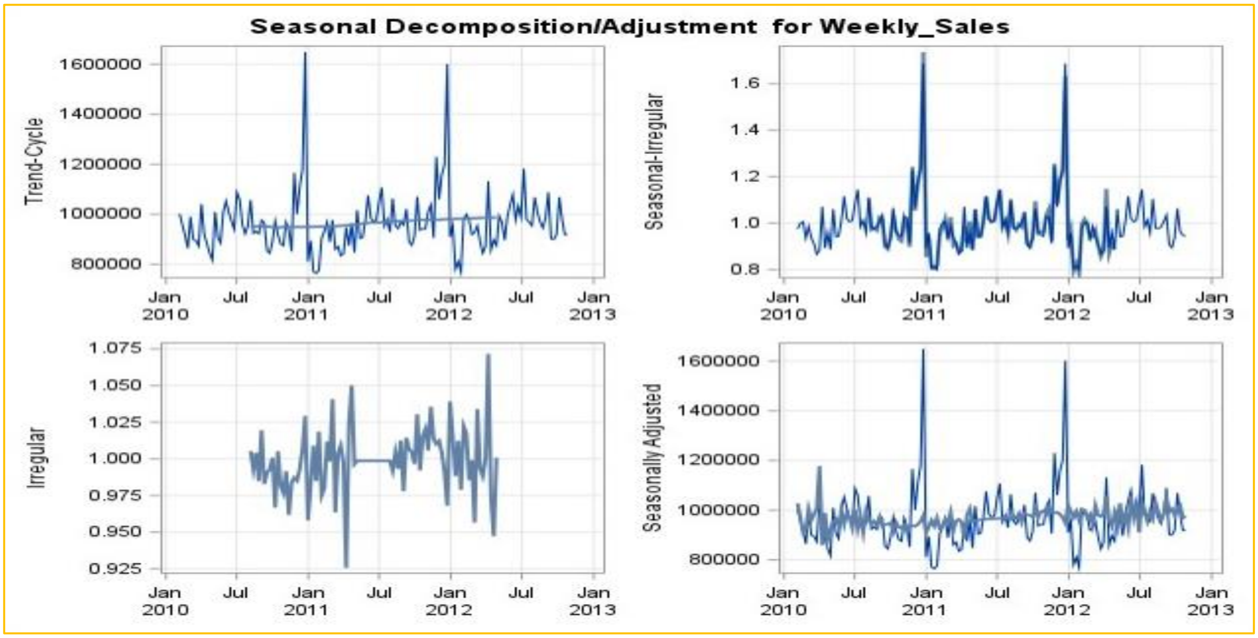
Seasonality significance is visible, and it is more than 20%. Sales going up from Nov till Dec end and then again going down from Jan onwards.



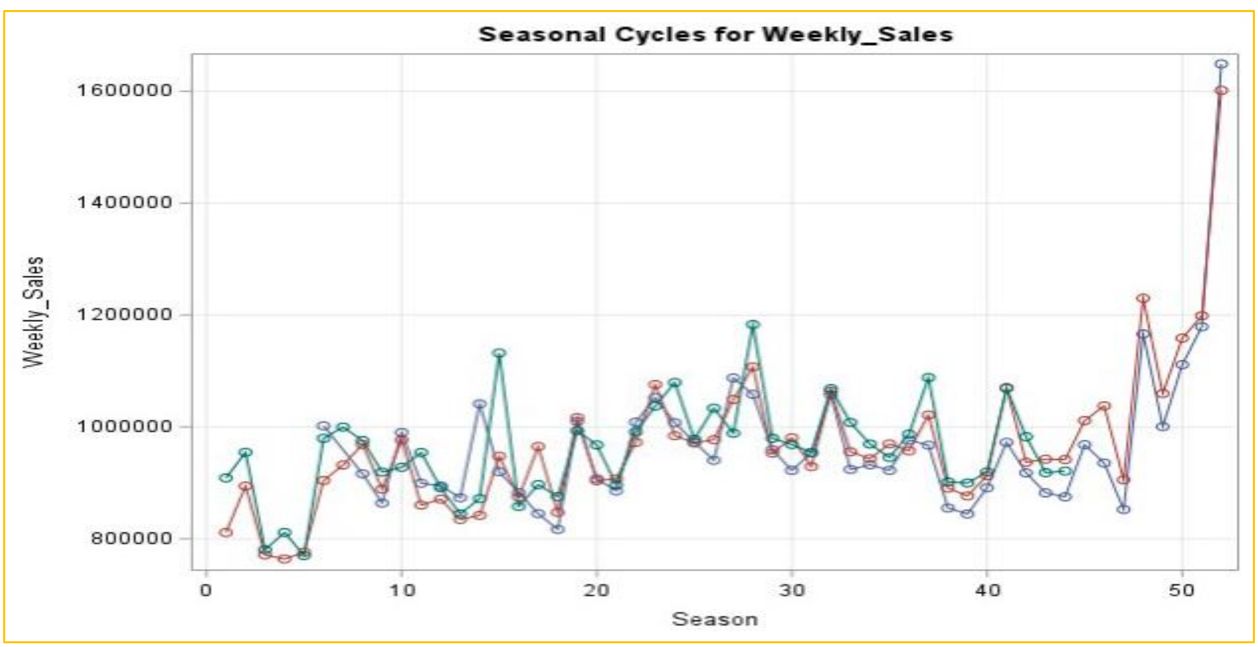
In Ljung-Box chi square test it is evident that the signals are significant in the time series and cannot be discarded as just white noise.



Decomposition of components is shown below,



From the seasonal cycle graph, it is visible that all years follow a similar seasonal cycle. Thus, seasonality is stable and strong in the time series.



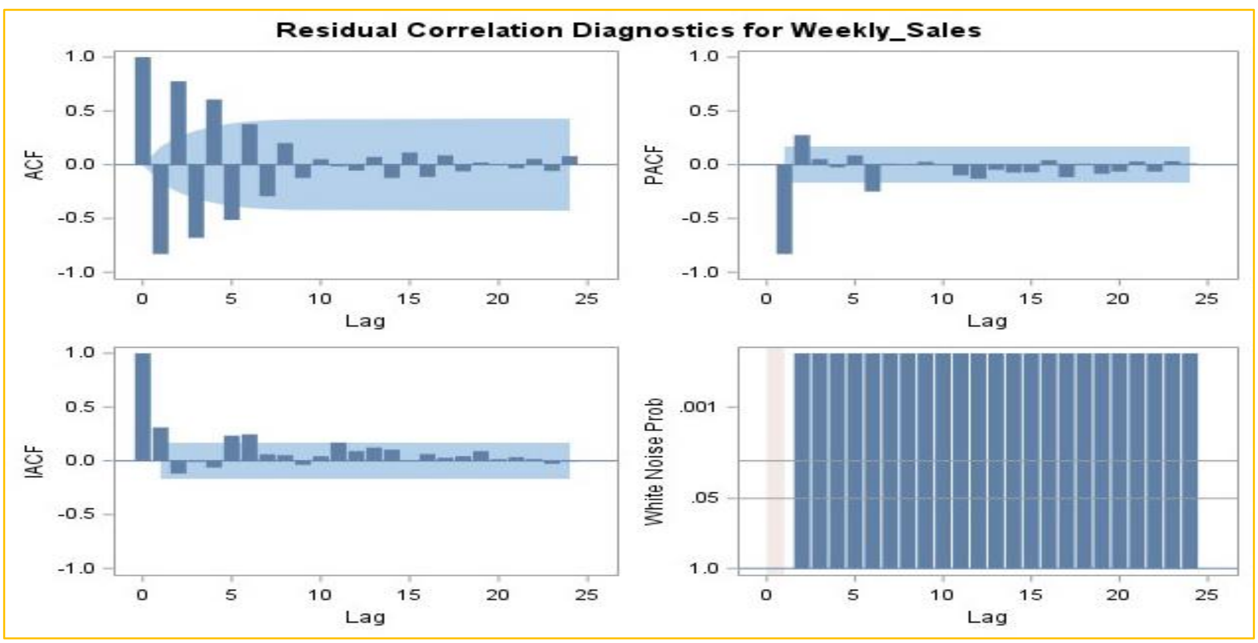
Modeling:

Moving Average Model

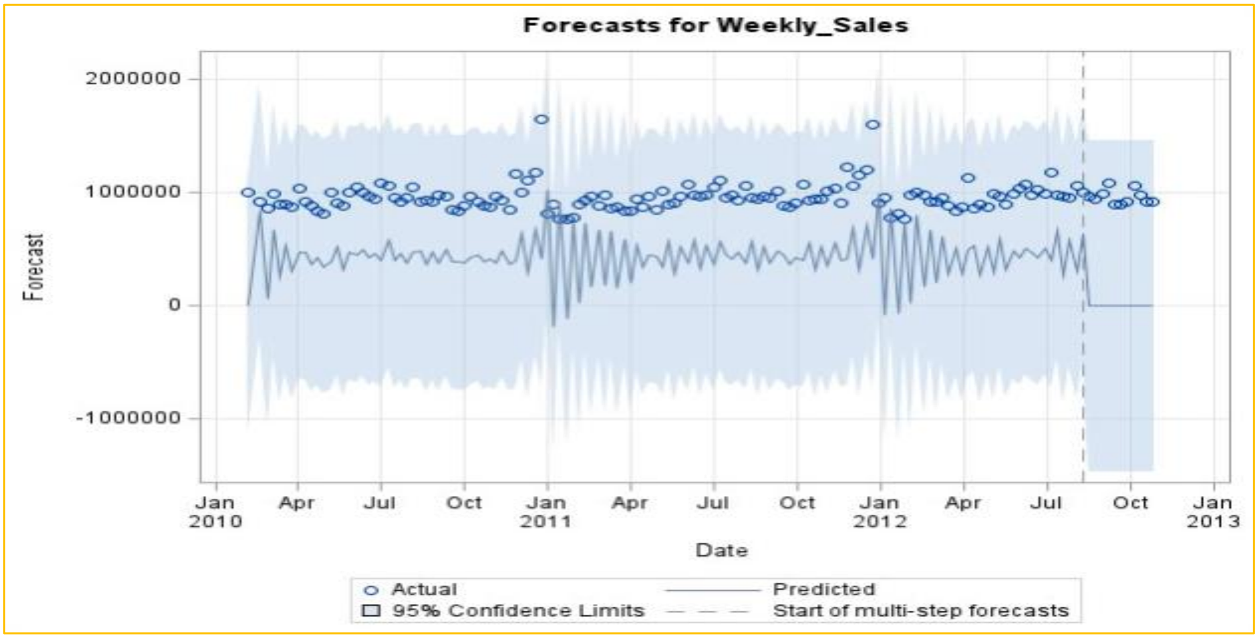
First, we ran the moving average model wherein based on all the statistics we concluded that it was not the best fit for the time series.

Variance Estimate	3.294E11
Std Error Estimate	573898.5
AIC	4169.874
SBC	4172.83
Number of Residuals	142

In the residual diagnostics reports we can see that signal is fully visible in thus not a good fit model.



Below is the forecast graph generated through moving average model

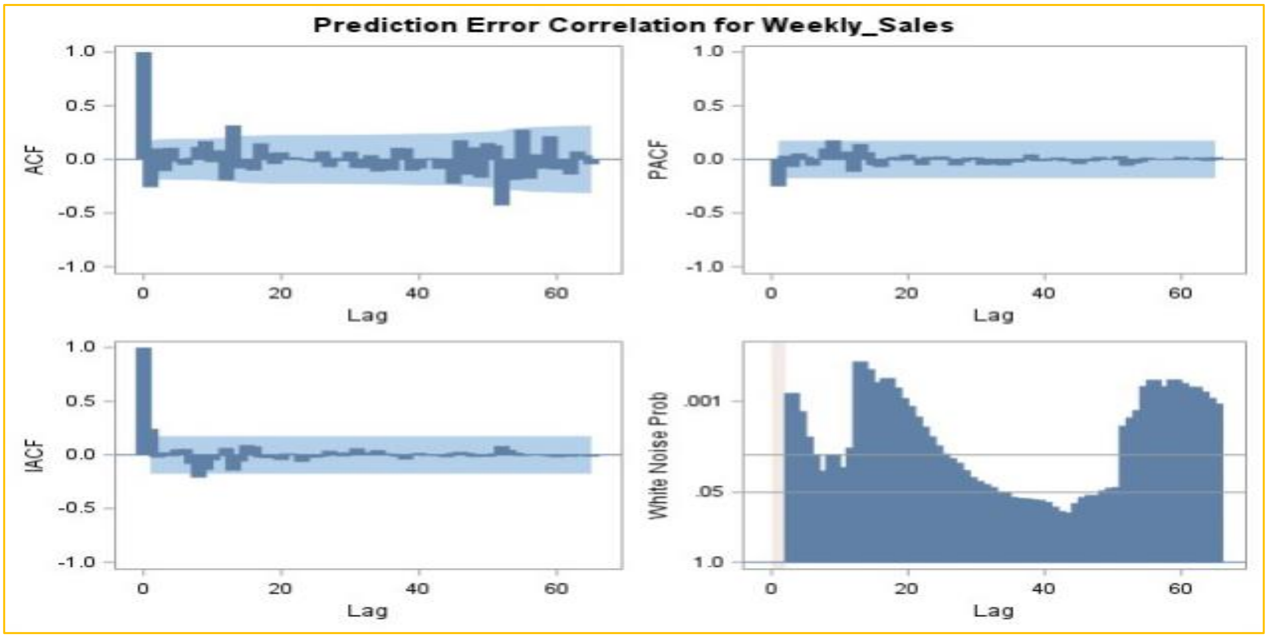


The equation generated through the moving average model.

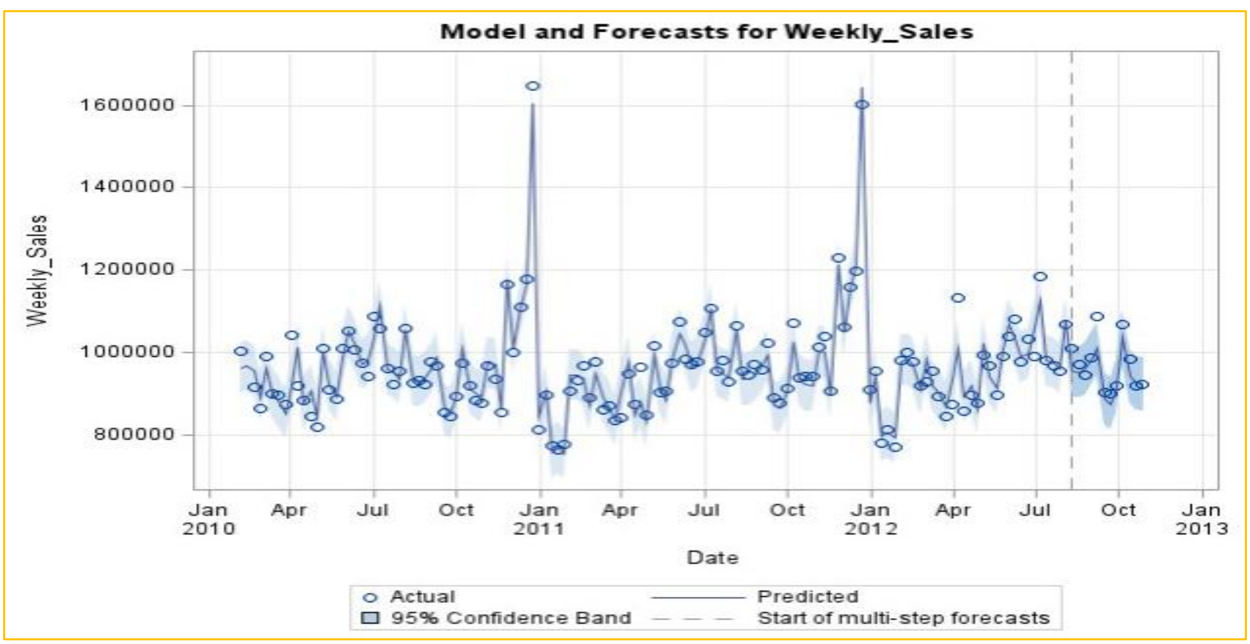
Moving Average Factors	
Factor 1:	1 + 0.84 B**(1)

Exponential Additive Model

As the seasonality is present in the time series and there is no trend thus we went ahead with exponential additive model to cater the seasonality.



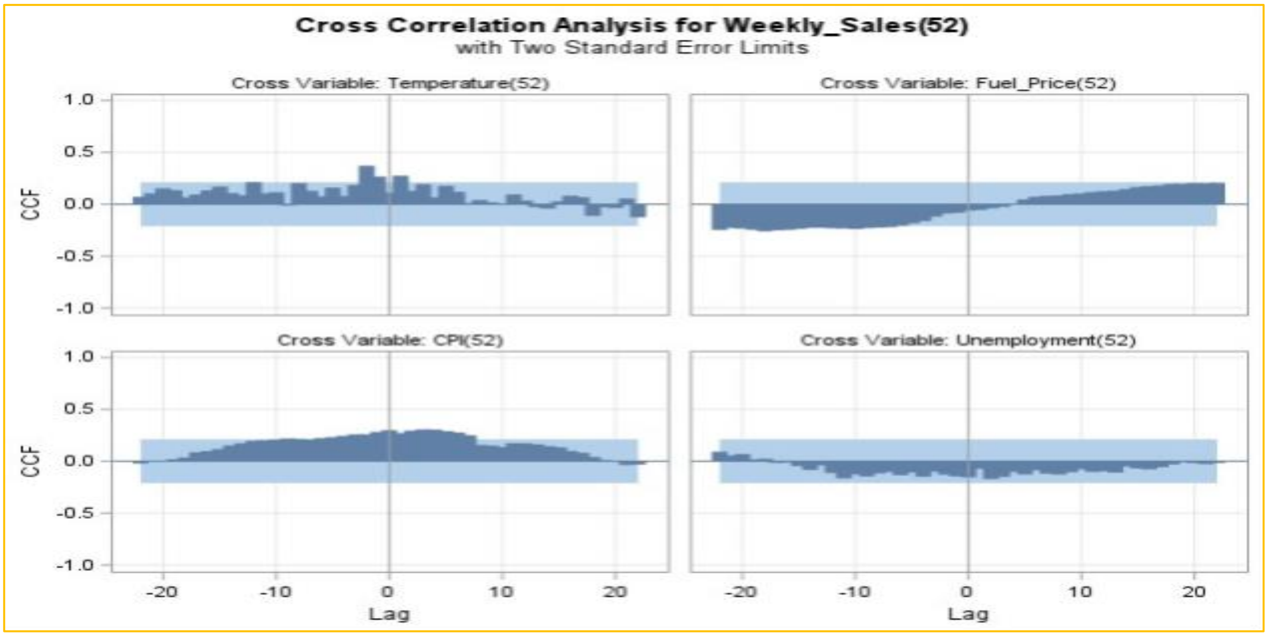
We rejected the exponential additive model as it failed to utilize all the signal and also AIC & SBC was too high.



ARIMAX (1,1,3)

Below we can see the cross-Correlation analysis of Dependent variable i.e. Weekly Sales with other independent variables such as Temperature, Fuel_Price, CPI and Unemployment.

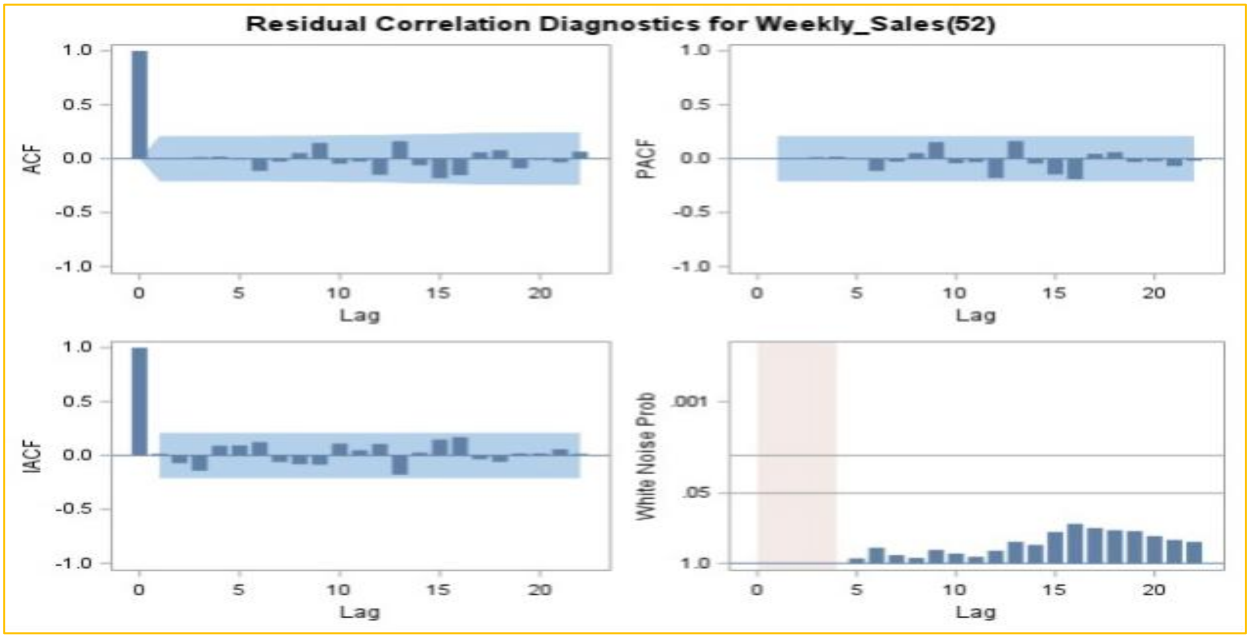
The cross correlation did not show any significant relationship between Weekly Sales and other independent variables.



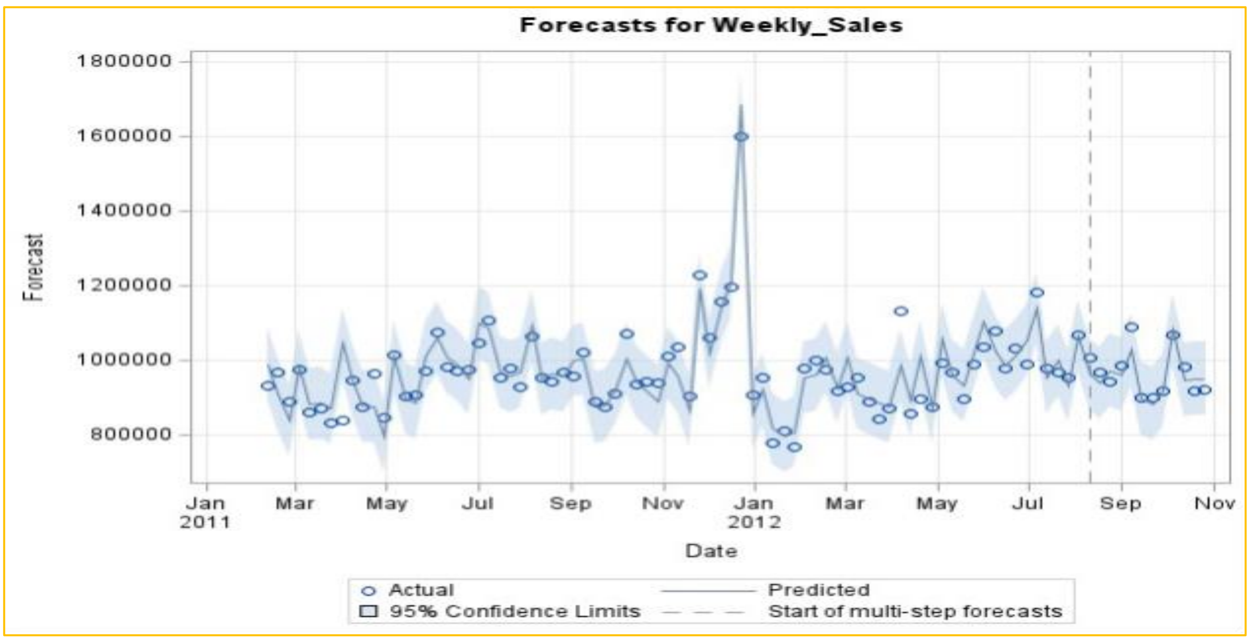
From the value of AIC and SBC, we conclude that ARIMAX was not the best fit model.

Constant Estimate	-49328
Variance Estimate	2.3881E9
Std Error Estimate	48868.48
AIC	2207.453
SBC	2229.951
Number of Residuals	90

Below is the screenshots of Residual diagnostic which show that ARIMAX did a good job, but it cannot be considered since there is no cross correlation.



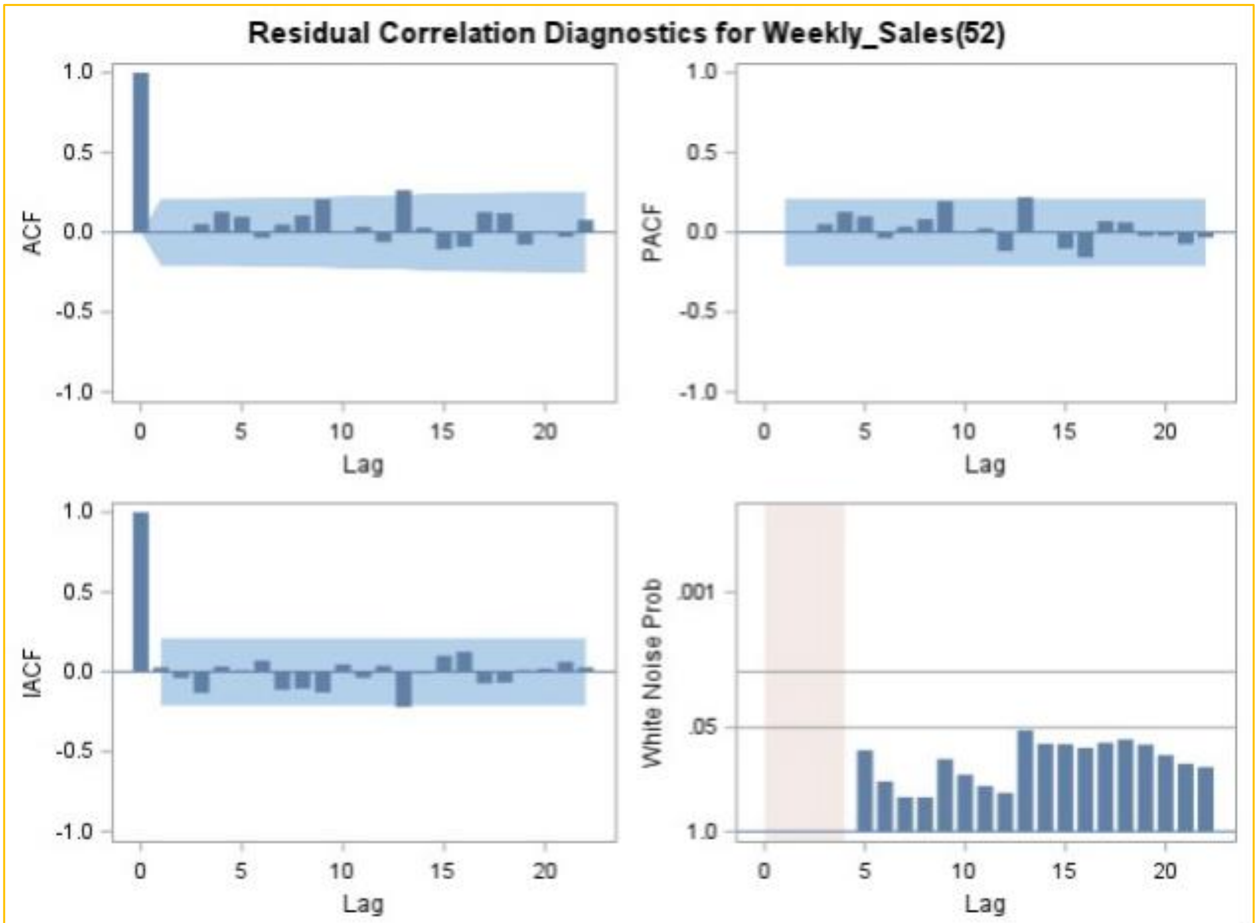
Below is the screenshot of forecast of weekly sales using ARIMAX.



ARIMA (1,1,3)

In the ARIMA first we ran ARIMA for p=1, d= 1 and q=3. But it did not provide the results as expected.

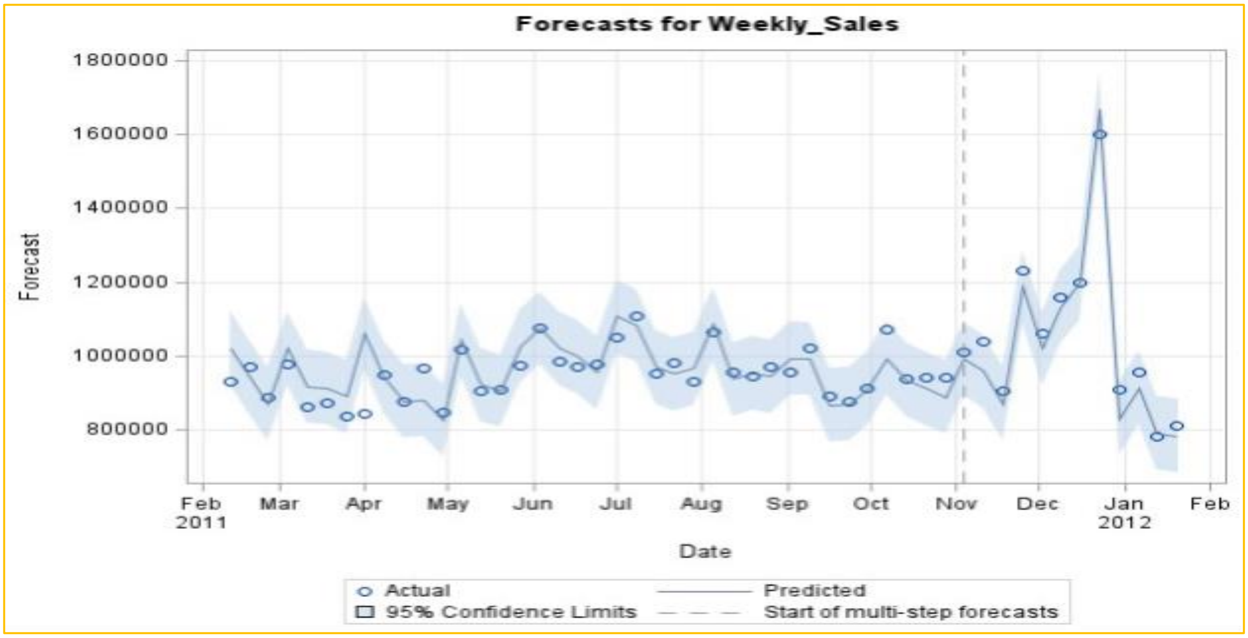
By looking at residual diagnostic we can see that signals are still visible in the residual.



Even more, the AIC and SBC values are not any better than the other models.

Constant Estimate	29297.65
Variance Estimate	2.5565E9
Std Error Estimate	50561.7
AIC	2209.884
SBC	2222.383
Number of Residuals	90

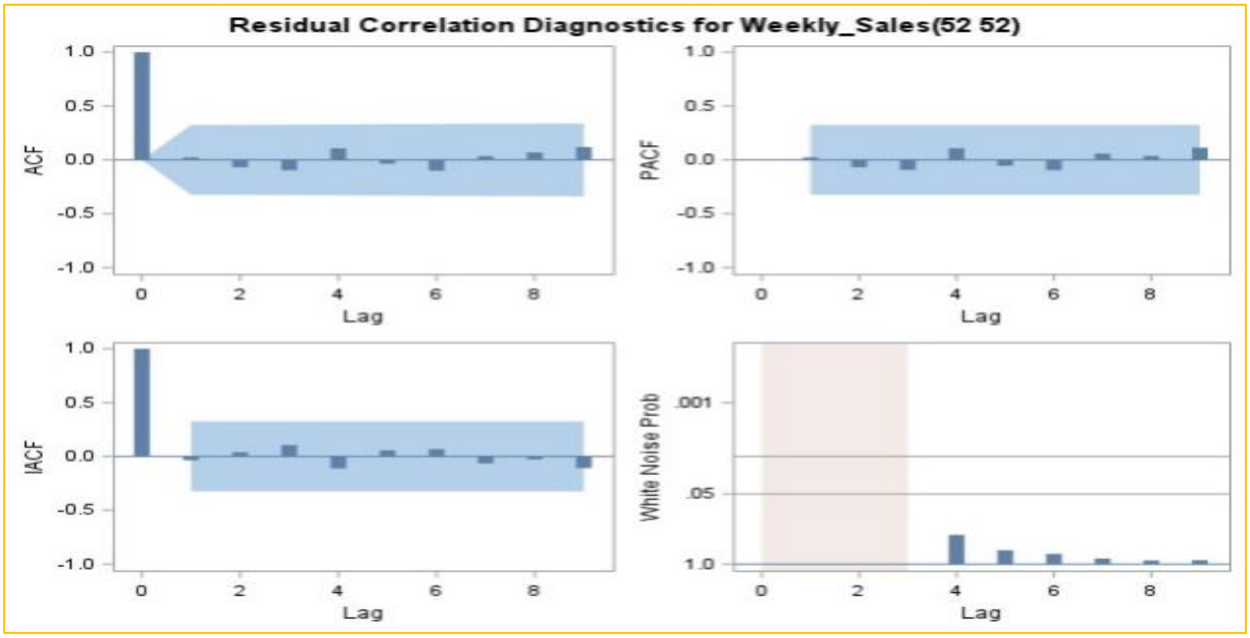
Below is the screenshot of the Forecasting graph using ARIMA (1,1,3)



ARIMA (2,2,3)

In this model we can ARIMA using p= 2, D= 2 and q= 3. We achieved best results using the aforementioned parameters.

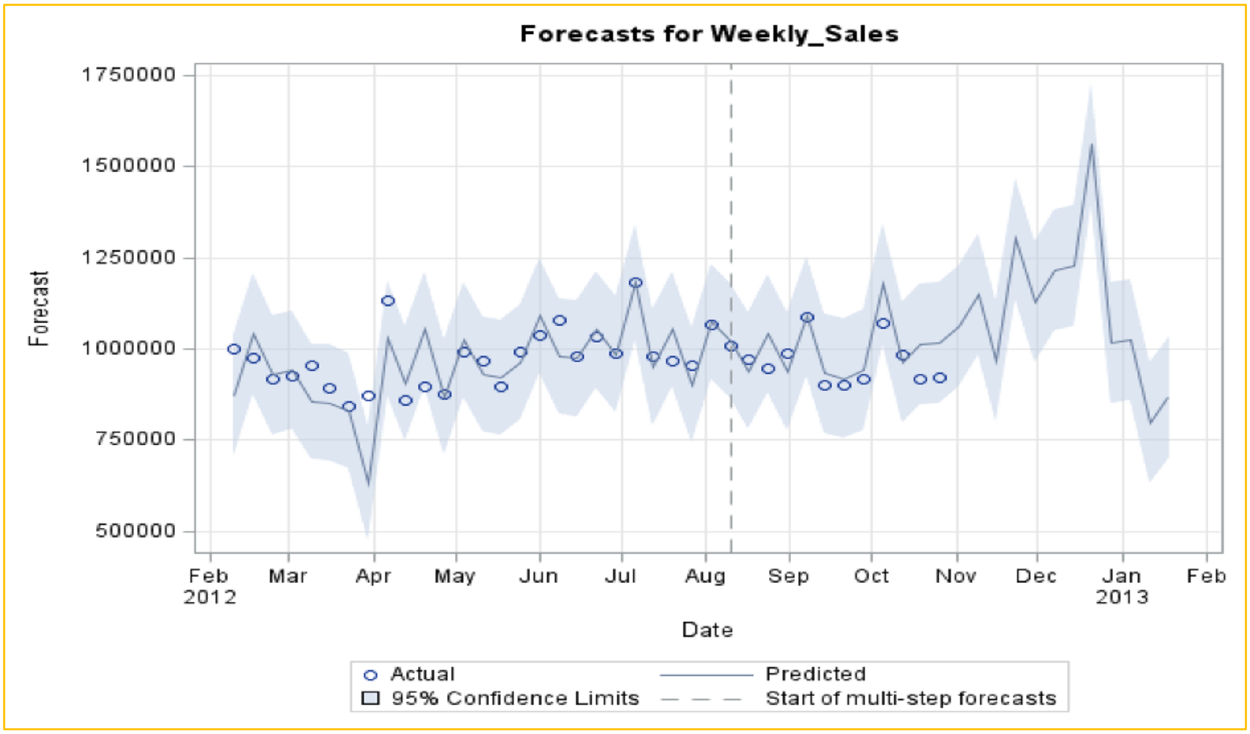
As you can see in the below attached picture, there are negligible signals in the white noise test and the white noise is more than 0.05 and significantly visible. Thus we could assure that we had extracted most of the signals using this model. Furthermore, ACF, PACF and IACF are insignificant in residual diagnostics.



Below we can see the results of AIC and SBC wherein it is visible that the values are significantly lower than other models and are in the range of 900.

Constant Estimate	12088.82
Variance Estimate	6.3818E9
Std Error Estimate	79886.39
AIC	969.62
SBC	976.1704
Number of Residuals	38

In the graph below we can see the forecasting achieved through ARIMA (2,2,3) model.


Forecasts for Weekly_Sales
Forecast
1750000
1500000
1250000
1000000
750000
500000
Feb 2012
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan 2013
Feb
Date
Actual
Predicted
95% Confidence Limits
Start of multi-step forecasts

5. Inference and Conclusion

The project's nuanced approach, segmenting data into individual store files to capture unique sales trends, underscores the complexity and variability inherent in retail operations across different locations.

Our exploration revealed diverse sales trends among the stores: seven exhibiting negative trends, fourteen positive, twenty constant, and four irregulars, showcasing the varied landscape of retail performance. This variability necessitated a tailored analytical approach, leading to the selection of ARIMA and ARIMAX models for their robustness in accommodating the data's nuances. The models' specifications—ranging from ARIMA (5,0,2) for Store #9 to ARIMA (2,2,3) for Store #40—were chosen based on their ability to accurately reflect the underlying sales patterns, validated through rigorous cross-correlation and pre-whitening analyses. These methodologies not only confirmed the impact of selected variables on sales but also ensured the parsimony and accuracy of the predictive models.

The practical applications of our findings are manifold. By accurately forecasting weekly demand, Walmart can proactively manage its workforce, ensuring optimal staffing levels to meet customer needs efficiently. This foresight extends to supply chain management, where predictive insights enable the optimization of shipping routes and inventory distribution, significantly reducing operational costs and enhancing timely product delivery. Furthermore, our analysis informs strategic inventory decisions, allowing for a dynamic adjustment of product assortment in response to evolving customer preferences.

Moreover, if we further explore and forecast sales using additional parameters that were not in the scope of this project, using the same approach followed in this project, the results could help in personalizing the shopping experience representing a forward-thinking approach to retail management. By tailoring offerings and interactions to individual customer preferences, Walmart can foster a more engaging and satisfying shopping journey, enhancing loyalty, and driving sales.

In essence, this project exemplifies the transformative potential of time series forecasting in the retail industry. As we move forward, the methodologies and findings of this study could serve as a valuable resource for retailers aiming to leverage predictive analytics for business optimization.

6. REFERENCES

- https://communities.sas.com/t5/SAS-Communities-Library/SAS-Visual-Forecasting-8-4-Interpreting-Results-and-Diagnostic/ta-p/581294
- https://www.kaggle.com/datasets/varsharam/walmart-sales-dataset-of-45stores
- https://documentation.sas.com/doc/en/etscdc/14.2/etsug/etsug_arima_details08.html
- Time Series Modeling Essentials Course Notes - George Fernandez, Marc Huber (2019)