



Sales Spectrum

Mining Insights for Retail Growth

— TEAM 8 —

Prof: Sudip Bhattacharjee

OPIM 5671 Data Mining and Business Intelligence

University of Connecticut

EXECUTIVE SUMMARY

The project report from Team 8, titled "Sales Spectrum: Mining Insights for Retail Growth," presented for the course OPIM 5671 Data Mining and Business Intelligence at the University of Connecticut, led by Prof. Sudip Bhattacharjee, delves into the critical areas of sales forecasting, inventory management, and regional sales prioritization within the retail sector. The goal is to enhance strategic planning, financial management, and operational efficiency through accurate month-to-month sales predictions. This executive summary encapsulates the essence of the project, highlighting its objectives, methodologies, and key findings.

Objectives:

The primary objectives outlined in the report focus on addressing the challenges of inventory management and sales optimization to navigate the fluctuations in consumer demand effectively. The project targets three main areas:

Sales Forecasting: Creating predictive models to forecast sales trends across different product categories and regions, utilizing historical sales data from 2014 to 2017.

Inventory Management: Focusing on office supplies and furniture sales to refine inventory management strategies, aiming to optimize stock levels based on predicted sales trends.

Regional Sales Prioritization: Analyzing sales forecasts by region to guide marketing efforts, store expansion strategies, and resource allocation, ensuring a tailored approach to meet local consumer preferences.

Methodology:

The report describes a comprehensive approach involving the segmentation of Superstore sales data, focusing on the categories of Furniture, Office Supplies, and Technology across various regions. The team employed time series analysis, specifically ARIMA (AutoRegressive Integrated Moving Average) models, to forecast monthly sales data, considering trends and seasonality.

Key steps in the methodology include:

Utilizing the Superstore sales dataset, consisting of 9,994 entries across 21 columns, to perform a detailed analysis.

Segmenting and aggregating sales data to monthly figures for analysis.

Conducting time series analysis to identify trends, seasonality, and stationarity in the sales data across product categories and regions.

Applying ARIMA models to predict sales, with thorough model fitting and validation processes for each product category.

Key Findings and Insights:

The analysis revealed several critical insights into sales trends, inventory management, and regional prioritization:

Sales Trends: An upward trend in sales over the years with clear seasonal patterns across all product categories, indicating the impact of cyclical events and holidays on consumer purchasing behavior.

Regional Analysis: Variations in sales trends across regions, with specific regions showing more pronounced growth and seasonality, guiding strategic regional prioritization.

Model Performance: The ARIMA models demonstrated varying degrees of effectiveness across different categories and regions, with specific models better capturing the underlying sales dynamics. The models provided a foundation for accurate sales forecasts, essential for strategic planning and operational adjustments.

Conclusion:

The project underscores the importance of leveraging data mining and business intelligence techniques to forecast sales, manage inventory, and prioritize regional strategies effectively. By harnessing predictive analytics, retailers can gain a strategic advantage, enhancing operational agility, customer satisfaction, and sustainable growth. The findings from Team 8's project offer valuable insights into the retail sector's challenges, presenting a data-driven framework for addressing these issues through accurate forecasting and strategic planning.

REPORT

1. Business Problem

The business problems our team has been focusing on addressing the challenge of inventory management and sales optimization within the retail sector. In the rapidly evolving retail landscape, the ability to accurately predict sales trends on a month-to-month basis stands as a cornerstone for strategic planning and financial management. This forecasting initiative stems from the need to navigate the fluctuations in consumer demand, which can significantly impact revenue streams and operational efficiency. Achieving accurate sales forecasts will enable our retail partners to fine-tune their marketing strategies, align production schedules, and manage cash flows more effectively. It also plays a critical role in decision-making processes related to

promotional activities, pricing adjustments, and resource allocation. In essence, understanding the month-to-month sales dynamics offers a strategic advantage, allowing retailers to proactively respond to market demands, capitalize on peak sales periods, and mitigate the risks associated with sales slumps. This forecasting effort is not just about predicting numbers; it's about crafting a data-driven approach to retail management that enhances operational agility, boosts customer satisfaction, and ultimately drives sustainable growth. We have aggregated the following business problems that we have addressed in our modeling:

1.1. Sales Forecasting:

We tackle the intricate challenge of forecasting sales across diverse product categories and regions. Utilizing historical sales data from 2014 to 2017, our mission is to create predictive models that accurately foresee sales trends. This initiative is fundamental for superstores to plan effectively, ensuring they can adapt to market demands, optimize inventory, and seize opportunities for revenue maximization. Accurate forecasting empowers businesses to align their operational strategies with future market conditions, enabling proactive management and strategic agility.

1.2. Inventory Management for Office Supplies & Furniture Sales:

Our focus narrows to refining inventory management for critical categories: office supplies and furniture. The fluctuating demand within these segments necessitates a robust forecasting approach to manage stock levels efficiently. By analyzing historical sales data, we aim to predict monthly sales trends, enabling precise inventory control. This foresight is crucial for minimizing overstock and understock situations, optimizing warehouse space, and ensuring products are available to meet consumer demand. This forecast will identify if consumer purchasing behavior is influenced by specific times of the year, which could be due to holidays, back-to-school seasons, or other cyclical events, and also focuses on inventory management. Strategic inventory management directly influences customer satisfaction, operational costs, and overall business profitability.

1.3. Targeted Region Prioritization by Regional Sales Forecasting:

The strategic allocation of marketing efforts and resources is paramount in maximizing retail success. By dissecting sales forecasts regionally, we identify areas with the highest growth potential, guiding superstores in prioritizing their investments and initiatives. This targeted approach enhances the effectiveness of marketing campaigns, store expansion strategies, and resource allocation. Understanding regional sales dynamics allows retailers to tailor their offerings and strategies to local consumer preferences, driving increased sales and fostering a competitive advantage in the marketplace.

2. Dataset Details:

The Superstore sales data utilized in our analysis was sourced from the [December Tableau User Group presentation](#). This dataset is one of the versions of Superstore Sales that were made available with the release of Tableau version 2022.1.

The original Superstore dataset consists of 9,994 entries and includes the following 21 columns:

Row ID: A unique identifier for each row.

Order ID: The ID associated with each order.

Order Date: The date when the order was placed.

Ship Date: The date when the order was shipped.

Ship Mode: The mode of shipping used for the order.

Customer ID: A unique identifier for each customer.

Customer Name: The name of the customer.

Segment: The market segment to which the customer belongs.

Country: The country of the customer.

City: The city of the customer.

State: The state of the customer.

Postal Code: The postal code of the customer.

Region: The region where the customer is located.

Product ID: A unique identifier for each product.

Category: The category of the product.

Sub-Category: The subcategory of the product.

Product Name: The name of the product.

Sales: The sales amount for the order.

Quantity: The quantity of the product ordered.

Discount: The discount applied to the order.

Profit: The profit made from the order.

We segmented the sales data using PROC SQL code and then aggregated it to monthly sales data.

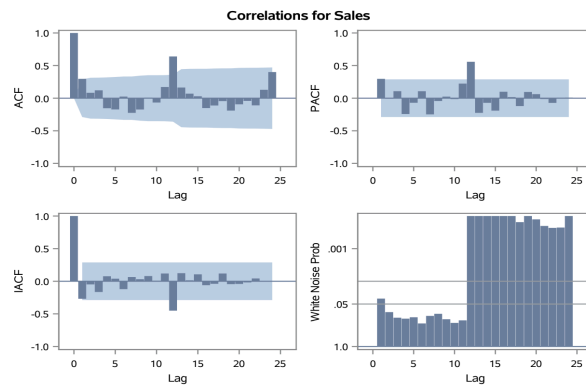
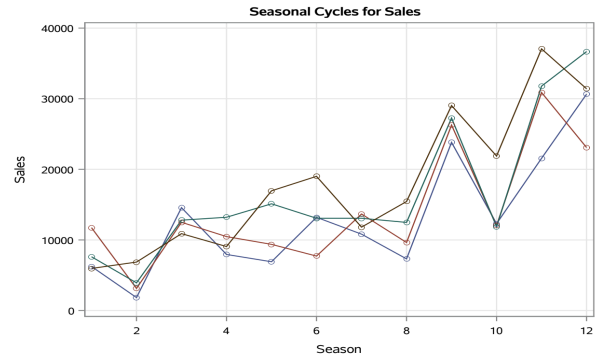
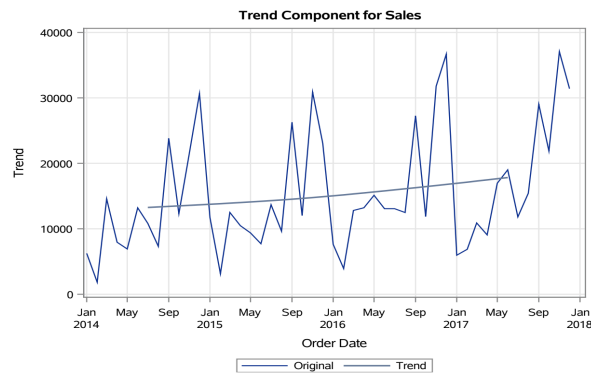
After splitting up the sales data that we needed to analyze, each team member subset the dataset based on their chosen category.

3. Model Selection (ARIMA)

The time series analysis was conducted for all the 3 major categories of products located across different regions of the superstore. This time series analysis will give a preliminary idea about how the monthly sales data has behaved over the years. In other words, is there a trend or seasonality in the series? The unit root test is also performed to check whether the series is stationary or not.

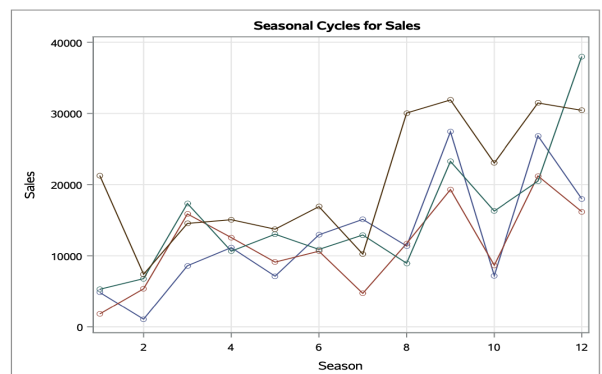
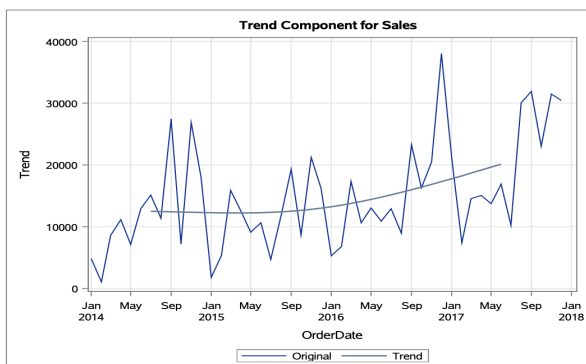
From the below plots of time series analysis of sales for Furniture, Office Supplies, and Technology, it can be seen that there is an upward trend over the years and seasonality is present as well.

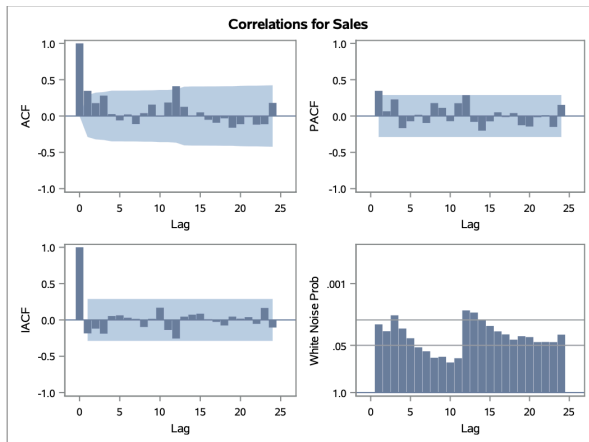
Furniture:



Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-5.7255	0.0939	-1.52	0.1198		
	1	-1.5381	0.3858	-0.57	0.4637		
	2	0.2709	0.7418	0.16	0.7269		
Single Mean	0	-29.4119	0.0004	-4.42	0.0009	9.86	0.0010
	1	-23.7401	0.0016	-3.21	0.0256	5.42	0.0350
	2	-10.9935	0.0861	-1.91	0.3271	2.23	0.5177
Trend	0	-38.7365	<.0001	-5.48	0.0002	15.04	0.0010
	1	-39.2418	<.0001	-4.13	0.0112	8.53	0.0135
	2	-22.0090	0.0227	-2.65	0.2594	3.58	0.4834

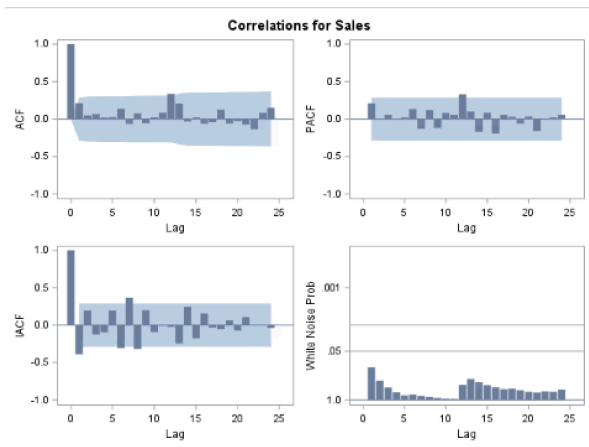
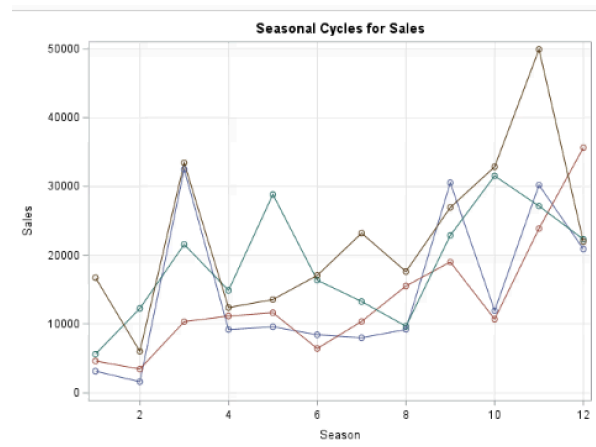
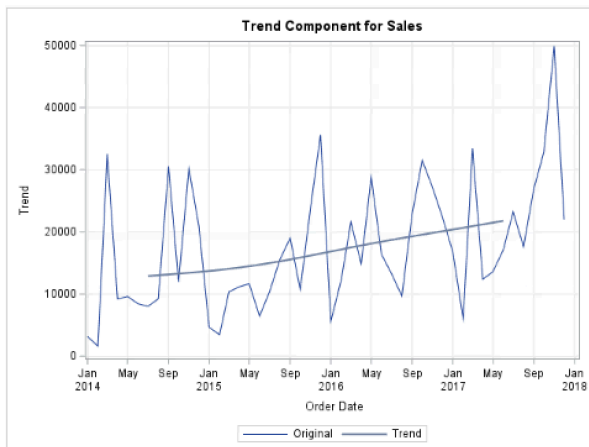
Office Supplies:





Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-5.7255	0.0939	-1.52	0.1198		
	1	-1.5381	0.3858	-0.57	0.4637		
	2	0.2709	0.7418	0.16	0.7269		
Single Mean	0	-29.4119	0.0004	-4.42	0.0009	9.86	0.0010
	1	-23.7401	0.0016	-3.21	0.0256	5.42	0.0350
	2	-10.9935	0.0861	-1.91	0.3271	2.23	0.5177
Trend	0	-38.7365	<.0001	-5.48	0.0002	15.04	0.0010
	1	-39.2418	<.0001	-4.13	0.0112	8.53	0.0135
	2	-22.0090	0.0227	-2.65	0.2594	3.58	0.4834

Technology:



Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-9.0398	0.0324	-2.14	0.0325		
	1	-2.8781	0.2397	-0.95	0.3006		
	2	-0.9525	0.4786	-0.44	0.5172		
Single Mean	0	-37.0912	0.0004	-5.52	0.0001	15.29	0.0010
	1	-35.9281	0.0004	-4.08	0.0025	8.43	0.0010
	2	-23.3209	0.0019	-2.54	0.1120	3.32	0.2511
Trend	0	-45.3808	<.0001	-6.44	<.0001	20.74	0.0010
	1	-55.7743	<.0001	-4.98	0.0011	12.29	0.0010
	2	-62.9284	<.0001	-3.89	0.0205	7.92	0.0232

REGION-WISE SALES ANALYSIS:

Time series analysis of sales for Central, South, East, and Western regions.

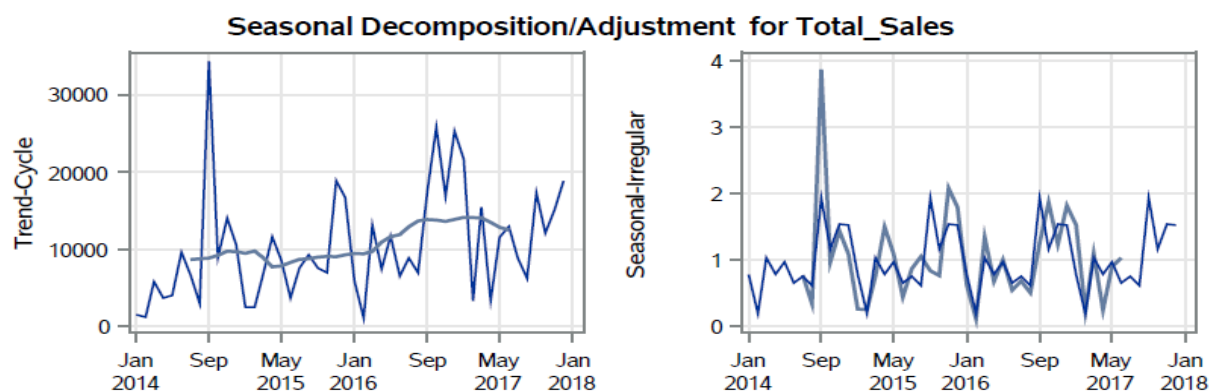
Central region: We are not modeling the central region as we observed that the residuals in the data are distributed as White noise.

Western Region: There has been an upward trend over the years and seasonality is present as well.

Southern Region: We are not modeling the central region as we observed that the residuals in the data are distributed as White noise.

Eastern Region: There has been a slightly upward trend over the years and seasonality is present.

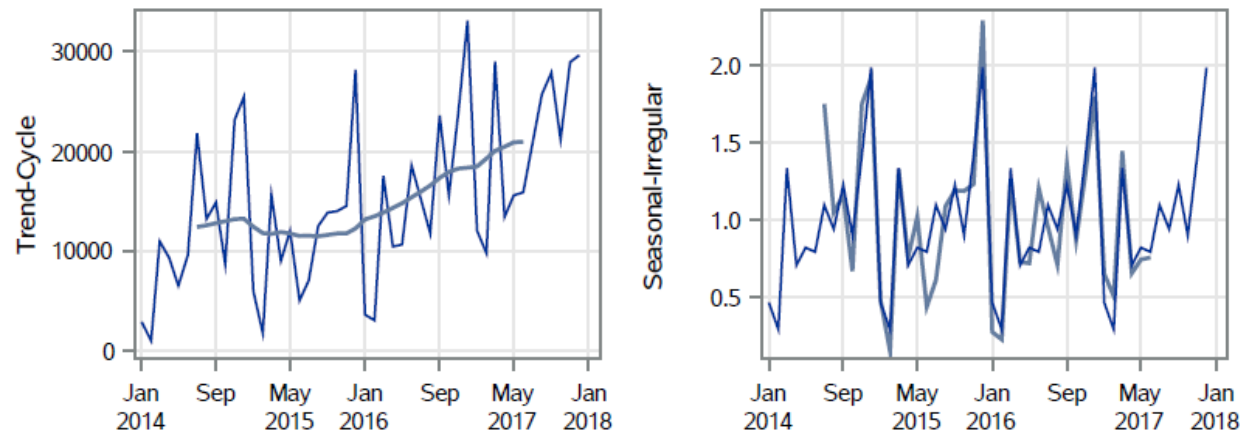
CENTRAL REGION:



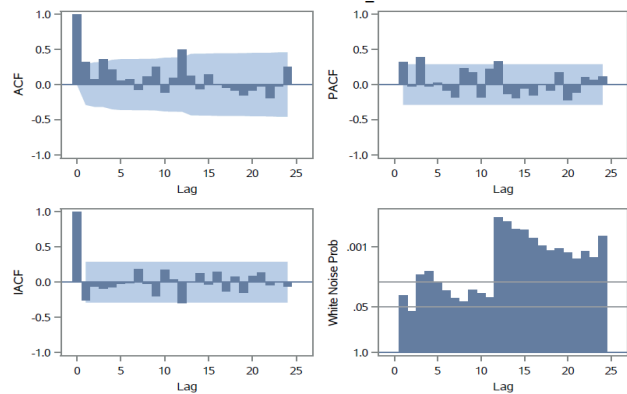
Augmented Dickey-Fuller Unit Root Tests						
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F
Zero Mean	0	-10.7282	0.0189	-2.30	0.0221	
	1	-3.6835	0.1820	-1.15	0.2252	
	2	-1.5992	0.3774	-0.66	0.4247	
Single Mean	0	-37.2610	0.0004	-5.45	0.0001	14.90
	1	-28.7552	0.0004	-3.74	0.0063	7.12
	2	-22.1724	0.0027	-2.90	0.0527	4.35
Trend	0	-41.7470	<.0001	-5.94	<.0001	17.61
	1	-35.7092	0.0003	-4.08	0.0125	8.35
	2	-30.2004	0.0018	-3.20	0.0967	5.13

WESTERN REGION

Seasonal Decomposition/Adjustment for Total_Sales



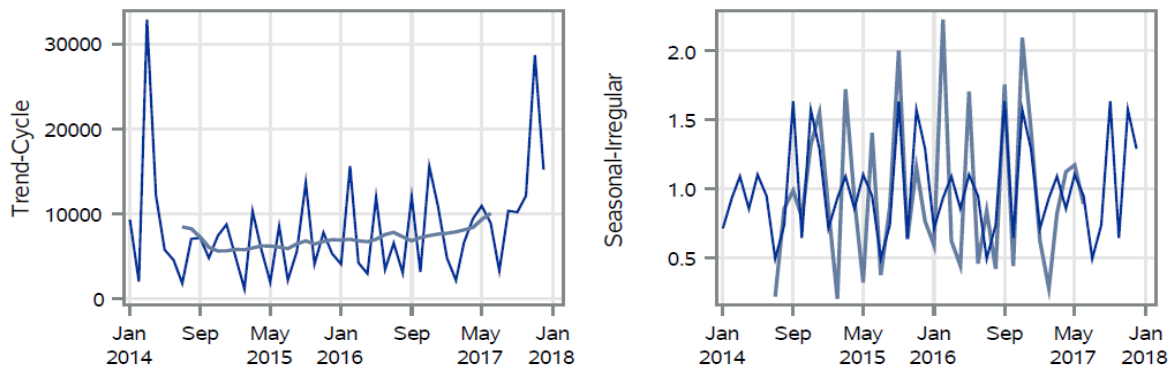
Correlations for Total_Sales

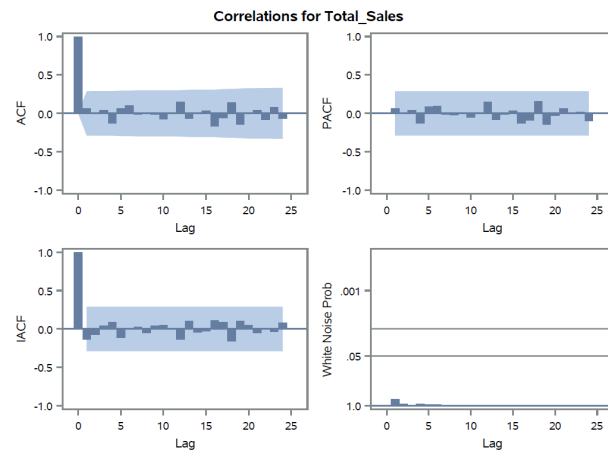


Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-5.5898	0.0981	-1.50	0.1235		
	1	-2.0766	0.3182	-0.73	0.3951		
	2	0.3782	0.7688	0.26	0.7558		
Single Mean	0	-30.9907	0.0004	-4.65	0.0005	10.93	0.0010
	1	-31.6726	0.0004	-3.81	0.0051	7.54	0.0010
	2	-9.5311	0.1289	-1.84	0.3571	2.13	0.5420
Trend	0	-43.4997	<.0001	-6.12	<.0001	18.70	0.0010
	1	-66.7196	<.0001	-5.48	0.0003	15.03	0.0010
	2	-27.0145	0.0051	-2.95	0.1575	4.38	0.3320

SOUTHERN REGION

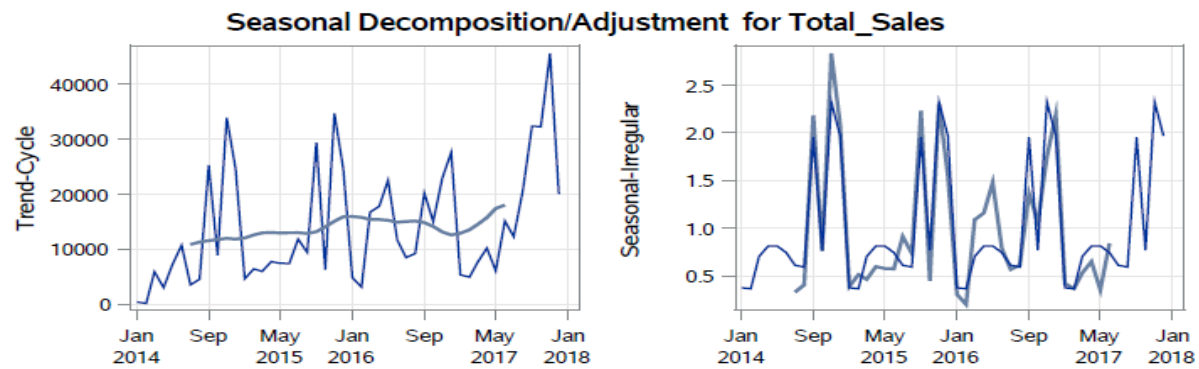
Seasonal Decomposition/Adjustment for Total_Sales





Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-15.7565	0.0038	-2.96	0.0038		
	1	-4.5465	0.1371	-1.16	0.2216		
	2	-2.7823	0.2476	-1.11	0.2386		
Single Mean	0	-44.0503	0.0004	-6.21	0.0001	19.31	0.0010
	1	-42.1842	0.0004	-3.91	0.0038	7.81	0.0010
	2	-33.9936	0.0004	-3.24	0.0236	5.26	0.0394
Trend	0	-45.0565	<.0001	-6.32	<.0001	20.05	0.0010
	1	-43.1056	<.0001	-3.97	0.0168	8.01	0.0217
	2	-34.5367	0.0004	-3.55	0.0463	9.04	0.0053

EASTERN REGION



Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-9.7183	0.0261	-2.22	0.0267		
	1	-3.4944	0.1939	-1.06	0.2577		
	2	-2.0753	0.3182	-0.69	0.4126		
Single Mean	0	-28.9597	0.0004	-4.55	0.0006	10.42	0.0010
	1	-22.1800	0.0028	-3.17	0.0281	5.20	0.0408
	2	-25.2921	0.0009	-2.80	0.0660	4.13	0.0897
Trend	0	-34.1679	0.0005	-5.04	0.0009	12.71	0.0010
	1	-28.7144	0.0031	-3.57	0.0441	6.36	0.0666
	2	-36.9316	0.0001	-3.24	0.0894	5.27	0.1641

3.1 Insights from Time Series Analysis

The following points can be summarized from the time series exploration analysis of five airports:

The trend components for sales in these categories generally show an upward trajectory, indicating growth over time.

All three categories exhibit clear seasonal patterns. This suggests that consumer purchasing behavior is influenced by specific times of the year, which could be due to holidays, back-to-school seasons, or other cyclical events.

Since the technology sales data appears to be random and does not have much autocorrelation, it might not benefit from more complex time series models.

The Augmented Dickey-Fuller (ADF) test results vary across categories but generally suggest that after accounting for trends and means, the data is stationary. This is an important characteristic of ARIMA-type models, which assume the stationarity of the data.

The ACF and PACF plots across the categories suggest different ARIMA parameters. While furniture and office supplies seem to indicate an autoregressive (AR) process, furniture may require fewer AR terms. The need for moving average (MA) terms is less clear and would need to be determined during model fitting.

Given the seasonality and trend, ARIMA / SARIMA models are suitable for all three categories. The exact parameters would need to be fine-tuned for each category based on further model diagnostics.

3.2 . MODEL FITTING

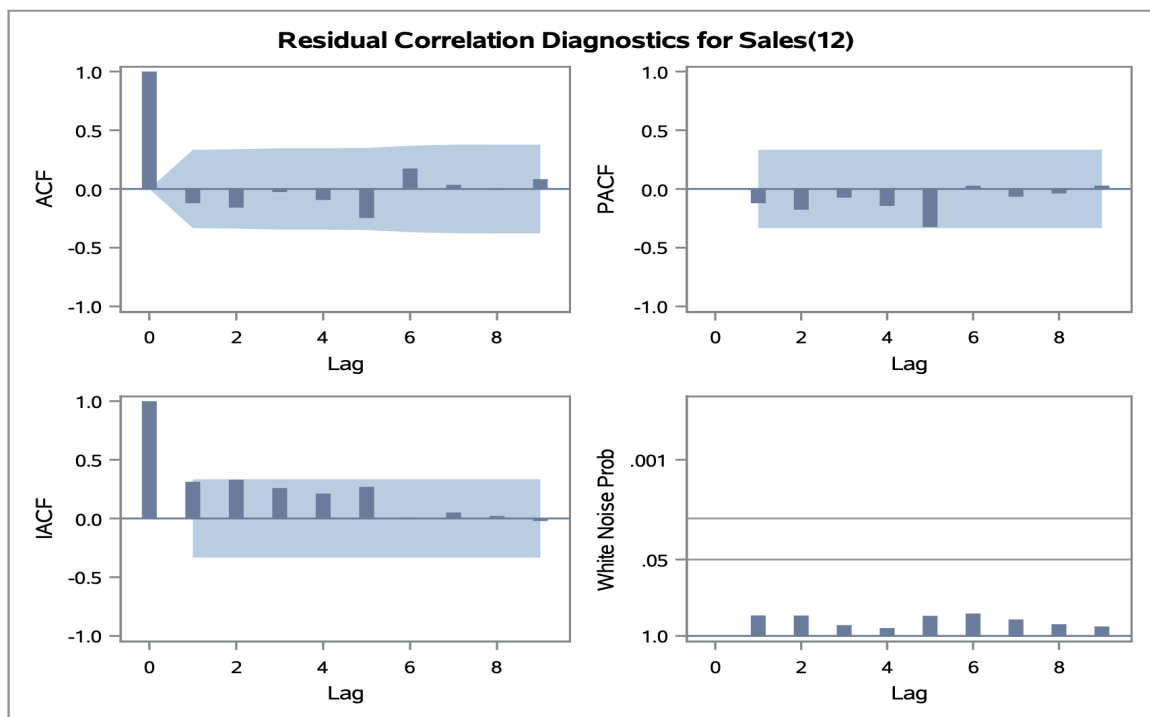
3.2.1 Product Category

3.2.1.1 Furniture

The Furniture data was modeled with ARIMA with different combinations of non-seasonal and seasonal values checking for the best-fitted model. Following are the different combinations of SARIMA(p,d,q)(P, D, Q) values tried on the dataset.

a. Case 1: SARIMA(0,0,0)(0,1,0):

As our data have seasonality, To address this, We have started fitting our model by starting with the seasonal differencing to capture and stabilize the seasonal effect. With this combination of SARIMA(0,0,0)(0,1,0), we were able to achieve the following results.



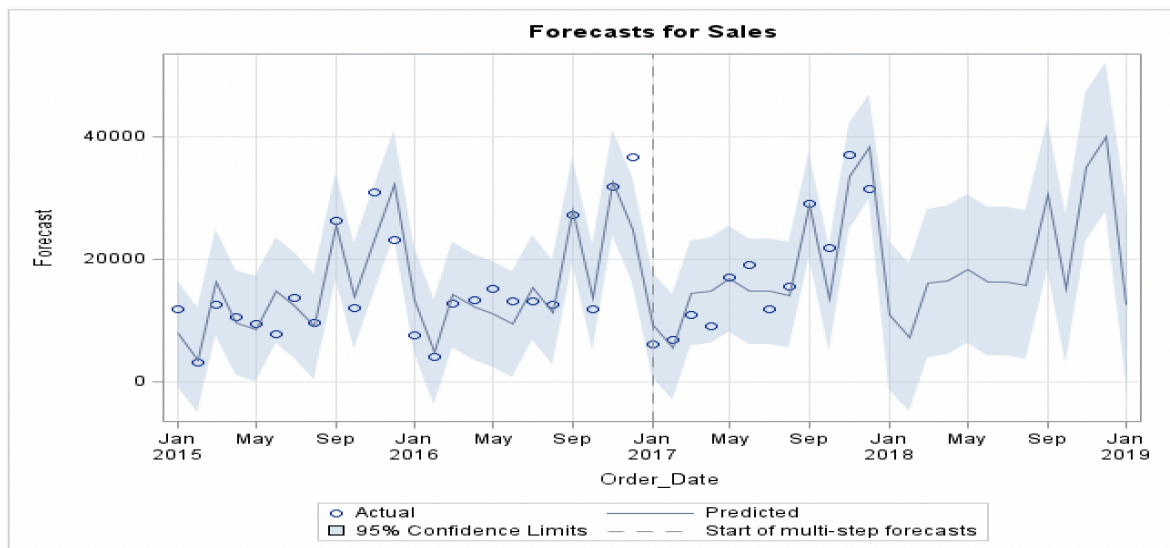
With this model, we were able to understand that ACF is suggesting that there is no autocorrelation in the residuals that the model has successfully captured the time series patterns, and the residuals are essentially random noise, which is the desired outcome in time series modeling.

	TYPE	_STAT_	_VALUE_
1	ML	AIC	706.7964

2	ML	SBC	708.38
3	ML	MAPE	24.82%

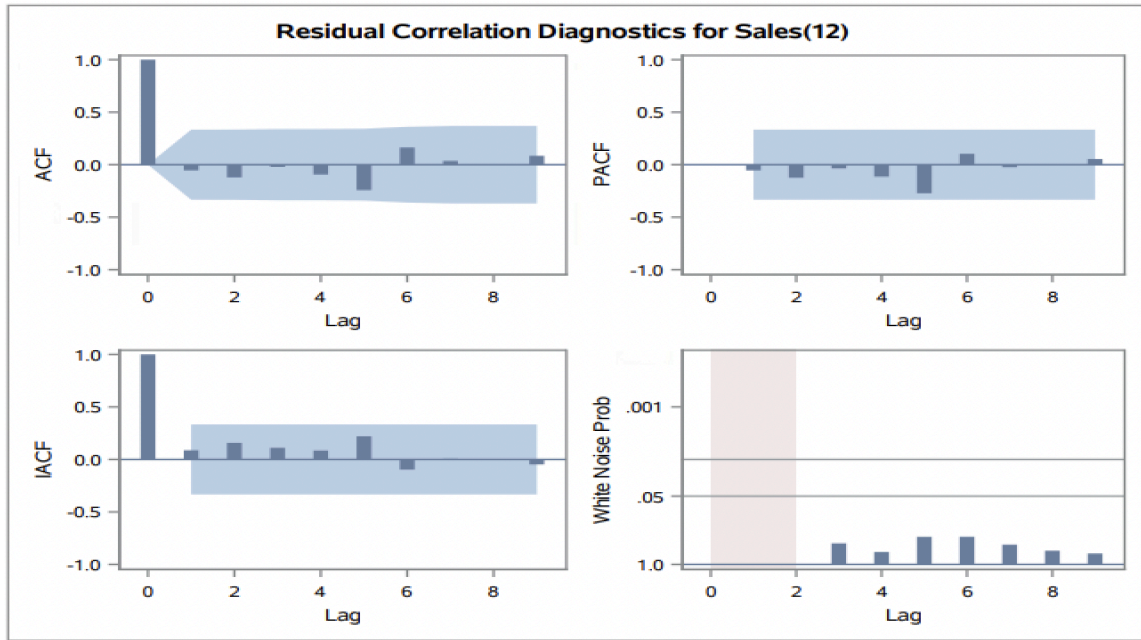
The above table gives the AIC and SBC values for this model which helps us in comparing the goodness of the model. We were able to achieve lower values of AIC and SBC with this model and a MAPE value of 24.82% which suggests it is a good model. However, we wanted to explore different combinations of non-seasonal and seasonal values.

The below plot indicates the forecast for the holdout sample.



b. Case 2: SARIMA(1,0,1)(0,1,0):

As we have observed the Autoregressive and Moving Average patterns, We have fitted our model with p and q values along with the seasonal differencing to capture and stabilize the seasonal effect. With this combination of SARIMA(1,0,1)(0,1,0), we were able to achieve the following results.

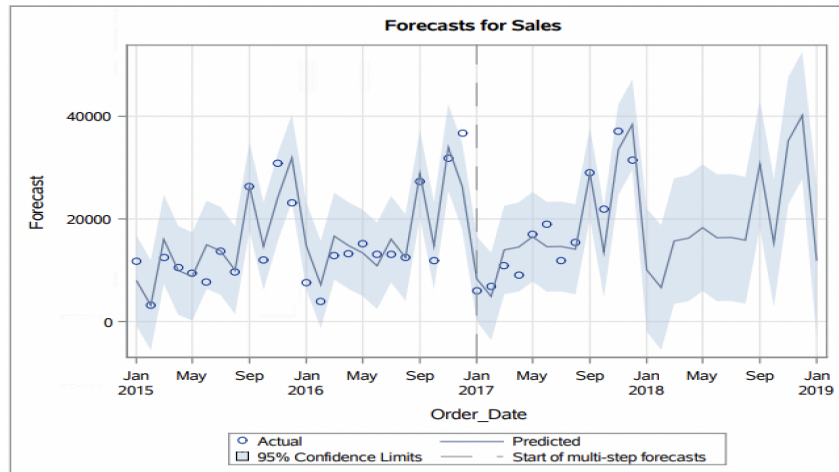


With this model, we were able to understand that ACF is suggesting that there is no autocorrelation in the residuals that the model has successfully captured the time series patterns, and the residuals are essentially random noise, which is the desired outcome in time series modeling.

	TYPE	_STAT_	_VALUE_
1	ML	AIC	707.4931
2	ML	SBC	712.2436
3	ML	MAPE	23.81%

The above table gives the AIC and SBC values for this model which helps us compare the goodness of the model. We were able to achieve the lower values of AIC and SBC with this model and a MAPE value of 23.81% which suggests it is a better model than the model we have in case 1, So this can be considered as the best model for the Furniture category.

The below plot indicates the forecast for the holdout sample.

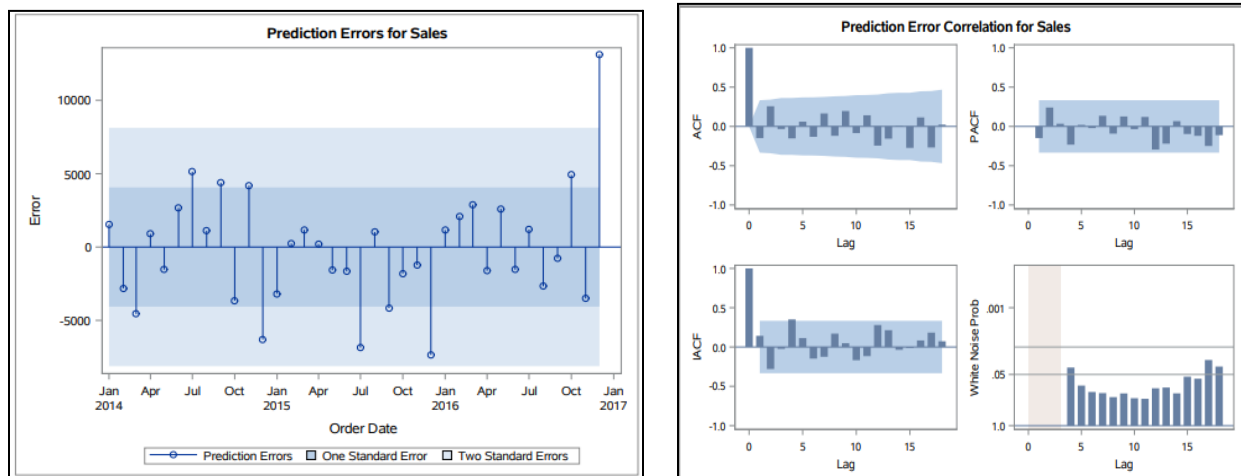


3.2.1.2 Office Supplies

The office supplies data was modeled with ARIMA and Exponential Smoothing models in order to see which performs better. Following are the test case scenarios of ARIMA vs Exponential Smoothing modeling.

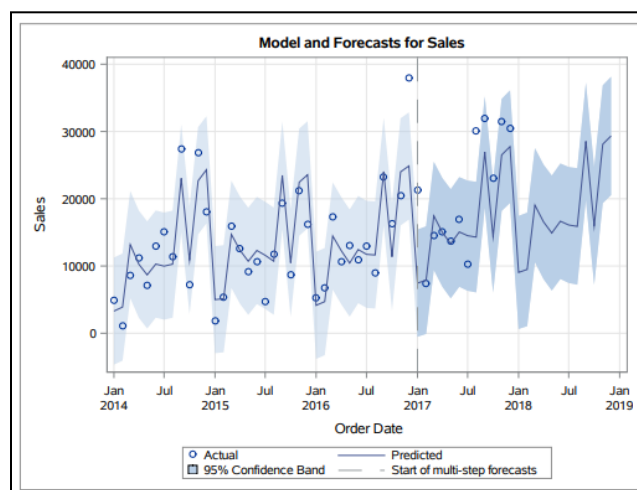
a. Test Case 1- Winters Additive Exponential Smoothing

The following are the output graphs obtained upon modeling:

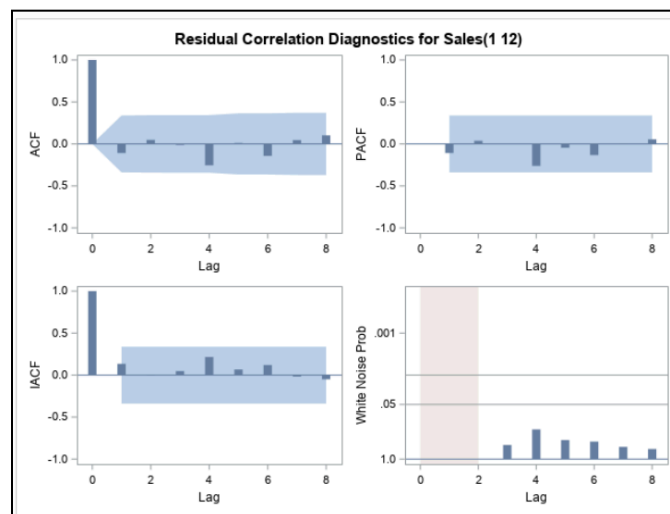


Obs	_NAME_	_REGION_	NPARMS	TSS	SST	RMSE	MAPE	MAE	RSQUARE	AIC	SBC
1	Sales	FIT	3	8303470330.4	2090093462.74	3888.40	35.3594	2984.42	0.73958	601.134	605.885
2	Sales	FORECAST	0	5894528285.3	847543323.38	7080.99	23.1693	5074.00	0.29008	212.764	212.764

The Exponential Smoothing Model shown demonstrates moderate explanatory power with an R-square of 0.73958, suggesting it accounts for approximately 74% of the variance in sales. However, the high RMSE and MAPE values indicate the model's predictions are not very accurate. The large AIC and SBC values also imply potential overfitting or the need for model improvement.

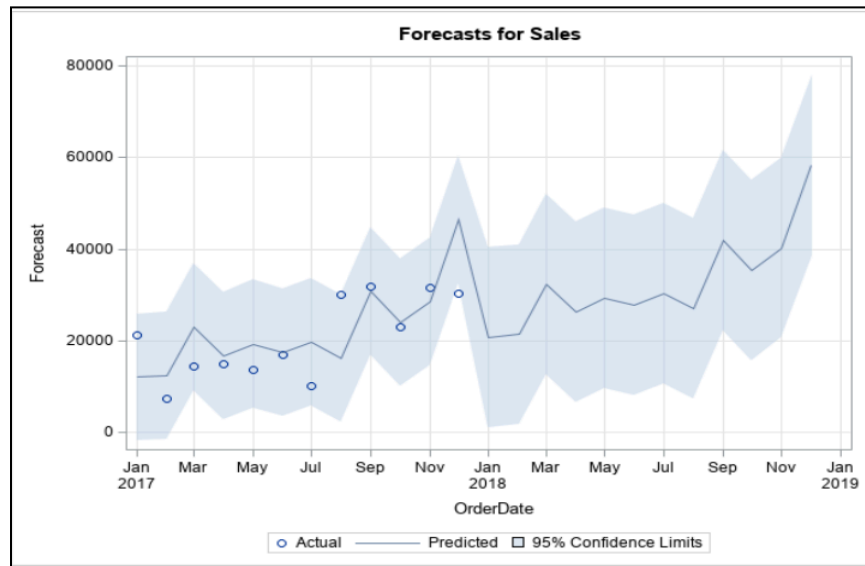


b. Test Case 2- SARIMA(1,1,1)(1,0,1)

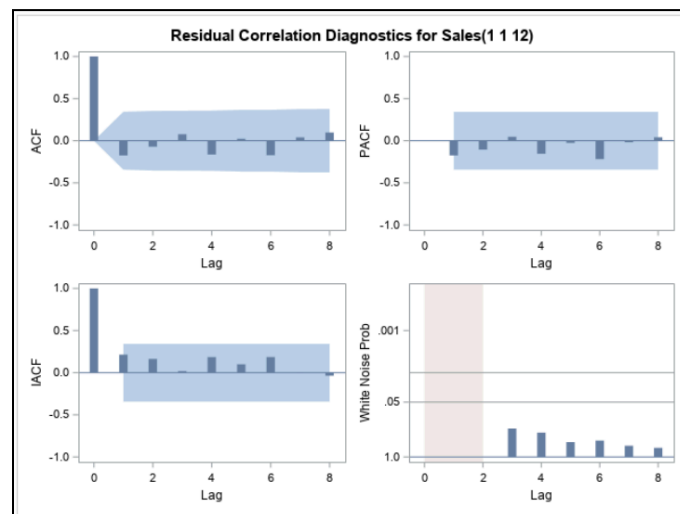


	TYPE	_STAT_	_VALUE_
1	ML	AIC	725.8322805
2	ML	SBC	730.4983247
3	ML	MAPE	14.15

The SARIMA (1,1,1)(0,1,0) model presents a significantly lower MAPE of 14.15% compared to the previously discussed model, suggesting a more accurate fit and better forecasting ability.

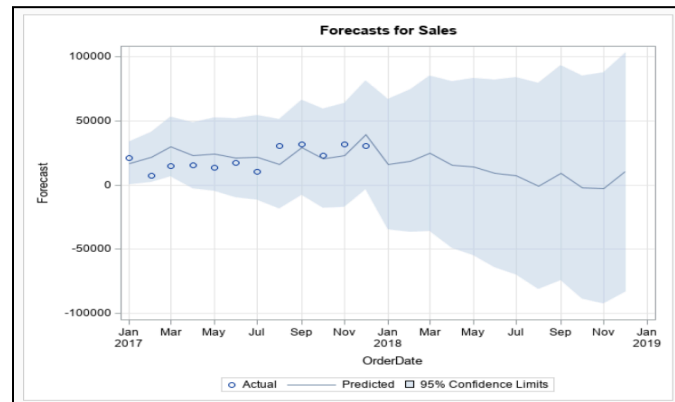


c. Test Case 3 SARIMA(1,2,1)(1,0,1)-



TYPE	_STAT_	_VALUE_
ML	AIC	719.31
ML	SBC	723.889
ML	MAPE	55.36%

The SARIMA model with the configuration (1,2,1)(0,1,0) appears to be underperforming, with a high Mean Absolute Percentage Error (MAPE) of 55.36%. This indicates that the average forecast error is more than half of the actual value, which is quite imprecise. Additionally, the AIC and SBC values are high, suggesting that the model might not be the best fit for the data and that there could be room for improvement either by tweaking the model parameters or considering a different modeling approach.

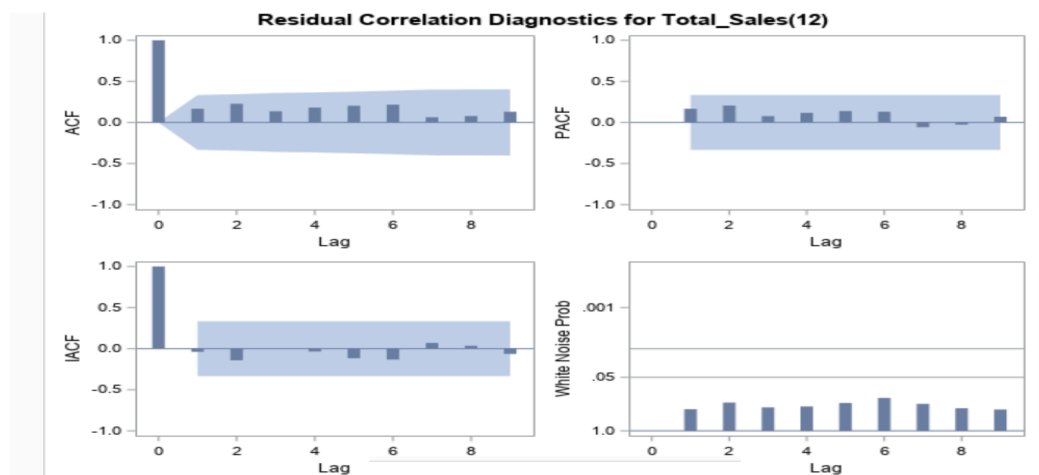


3.2.1 Region-wise sales analysis

3.2.1.1. Western Region Analysis

The western region sales data was modeled with ARIMA to see which performs better. Following are the test case scenarios of ARIMA vs Exponential Smoothing modeling.

a. Test Case 1 SARIMA(0,0,0)(0,1,0):



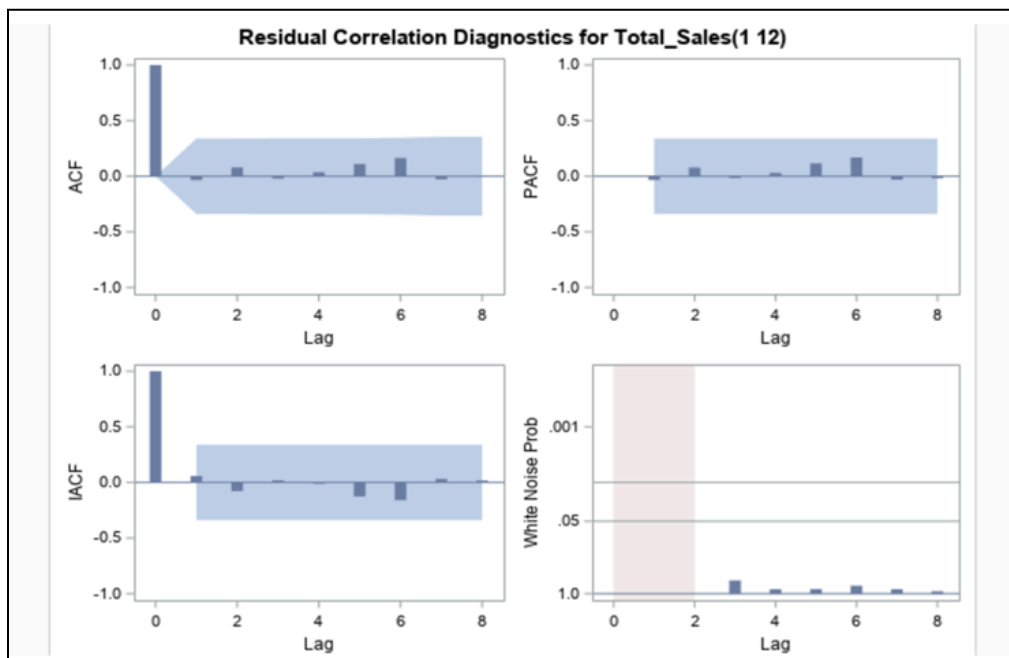
	TYPE	_STAT_	_VALUE_
1	ML	AIC	727.54
2	ML	SBC	729.1
3	ML	MAPE	22.26%

We have

observed lower

AIC and SBC Values for this model and have a MAPE value of 22.26% which suggests this is a good model. However, we wanted to explore different combinations of non-seasonal and seasonal values, so we then tried running the SARIMA (1,1,1)(0,1,0) combination believing it would be the best-fit model for the western region.

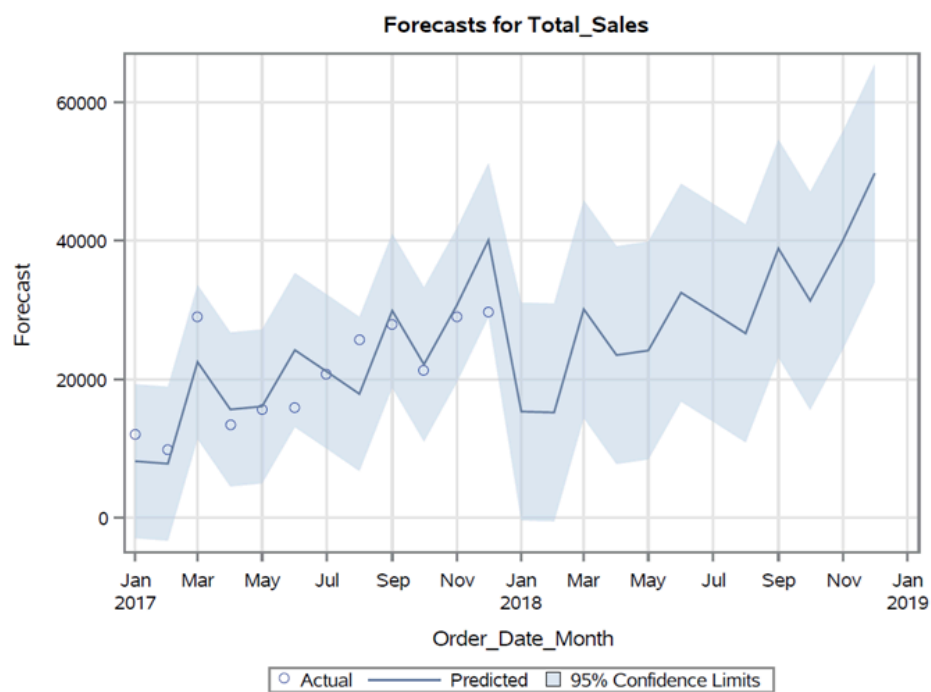
b. Test Case 2 SARIMA(1,1,1)(0,1,0):



	TYPE	_STAT_	_VALUE_
1	ML	AIC	710.6
2	ML	SBC	715.3
3	ML	MAPE	19.26

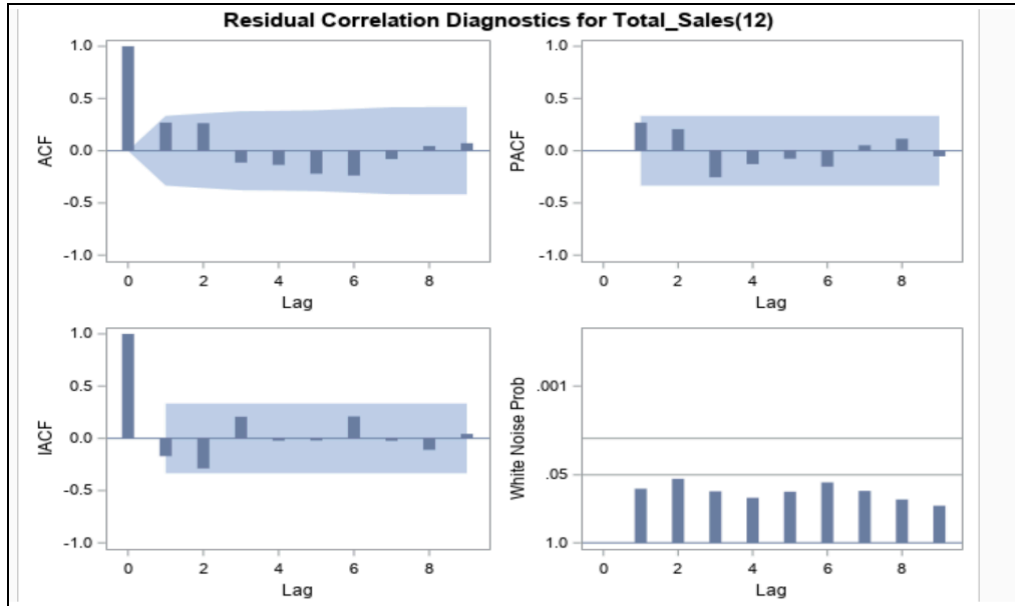
The above table gives the AIC and SBC values for this model which helps us compare the goodness of the model. We were able to achieve the lower values of AIC and SBC with this model and a MAPE value of 19.26% which suggests it is a better model than the model we have in case 1, So this can be considered as the best model for the Western region sales analysis.

The below plot indicates the forecast for the holdout sample.



3.2.1.2. Eastern Region Analysis

a. Test Case 1 SARIMA(0,0,0)(0,1,0):



	TYPE	_STAT_	_VALUE_
1	ML	AIC	752.594
2	ML	SBC	754.178
3	ML	MAPE	69%

The above table gives the AIC and SBC values for this model which helps us compare the goodness of the model. We were able to achieve a MAPE value of 69% which suggests the model is of very low accuracy and is hence not desirable.

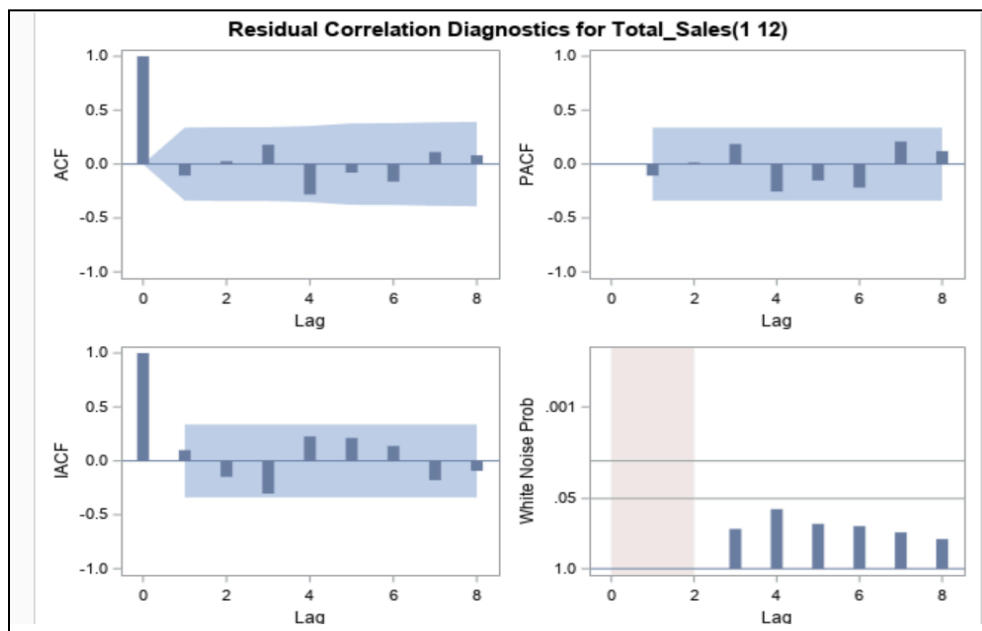
a.Test Case 2 - ESM(Additive Seasonal):

Obs	_NAME_	NOBS	RMSE	MAPE	AIC	SBC
1	Total_Sales	36	3989.84	31.7711	600.989	604.156
2	Total_Sales	12	9230.17	39.3183	219.126	219.126

3.2.1.3. Total Sales Analysis

a. Test Case 1 - SARIMA(1,1,1)(0,1,0):

As we have observed the Autoregressive and Moving Average patterns, We have fitted our model with p and q values along with the seasonal differencing to capture and stabilize the seasonal effect. With this combination of SARIMA(1,1,1)(0,1,0), we were able to achieve the following results.



	TYPE	_STAT_	_VALUE_
1	ML	AIC	769.6
2	ML	SBC	774.2
3	ML	MAPE	25.5%

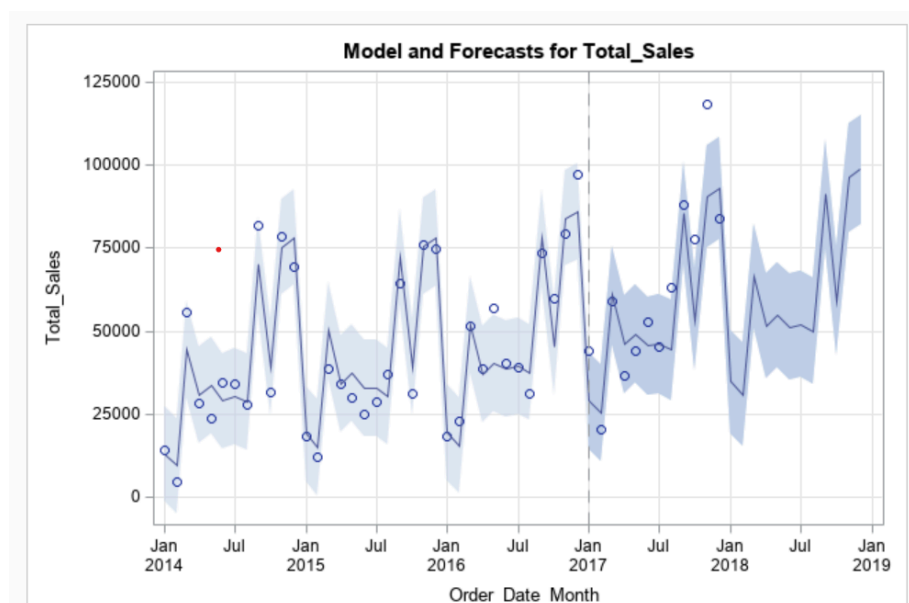
We were able to achieve lower values of AIC and SBC with this model and a MAPE value of 25.5% which suggests it is a good model. However, we wanted to explore different models to get a better model.

b. Test case 2 - ESM (Winters Additive):

The following are the output graphs obtained upon modeling:



Obs	_NAME_	_REGION_	N	NPARMS	TSS	RMSE	MAPE	MAE	RSQUARE	AIC	SBC
1	Total_Sales	FIT	36	3	87005734844	7053.74	16.9712	5622.04	0.90602	644.015	648.765
2	Total_Sales	FORECAST	12	0	52771657518	13673.47	17.7877	10611.95	0.71854	228.557	228.557



The Exponential Smoothing Model with Winters Additive demonstrates moderate explanatory power with an R-square of 0.906. Therefore, the RMSE and low MAPE values indicate the model's predictions are very accurate and can be considered a best-fit model.

5. CONCLUSION

In addressing the challenges of sales forecasting, inventory management, and targeted region prioritization, our report outlines strategic initiatives aimed at enhancing operational efficiency and driving sales growth.

1. Implement marketing campaigns and promotions in June, September, and December to enhance overall sales, leveraging seasonal buying behaviors.
2. During slower sales months, introduce customer loyalty programs to encourage repeat purchases, maintaining revenue flow and customer engagement.
3. Utilize category-specific predicted values and confidence interval ranges to manage inventory efficiently, reducing storage costs while ensuring product availability.
4. Align marketing campaigns and stock levels to peak seasonal periods to maximize sales of the furniture category. Consider long-term investments in inventory and supply chain enhancements to support the growing demand.
5. Diversify product offerings in the office supplies category to cater to peak buying times, such as back-to-school or end-of-financial-year sales.

6. REFERENCES

1. Software SAS
2. Time Series Modeling Essential Course Notes, 2019 SAS Institute Inc. Cary, NC, USA, ISBN 978-1-64295-144-8