DS 560 Applied Forecasting, Fall 2023

Final Project Report

Project-Title: Time series analysis and forecasting for furniture sales

Project Group Members:

1. Arun Reddy Yeduguru
2. Priyanshu Gupta
3. Shanmukh Varma Rudraraju

A. Brief Background of Enterprise

The project is situated inside the framework of an established chain of retail superstores that offer a wide variety of products, with a particular emphasis on office supplies and furnishings.

With a large market presence across multiple locations, the superstore serves a diverse range of customers with varying buying habits.

Having a thorough understanding of sales patterns is essential for both strategic planning and preserving competitive advantage, particularly when reacting to shifting consumer preferences and market conditions.

Robust sales forecasting is an essential tool for the superstore due to the complex nature of retail operations, which are influenced by a myriad of elements like seasons, economic situations, and customer trends.

B. Description of the Business Problem

Accurate and trustworthy sales forecasting is essential to addressing the superstore's strategic challenges.

Many corporate operations, such as inventory control, workforce planning, budgeting, and marketing initiatives, depend on accurate forecasting.

The sales information of office supplies and furniture, two significant revenue streams with distinct demand patterns, is the focus of this project.

The superstore can place orders for products, plan promotions, and allocate resources wisely by properly predicting sales trends in these areas.

Robust sales forecasting is an essential tool for the superstore due to the complex nature of retail operations, which are influenced by a myriad of elements like seasons, economic situations, and customer trends.

C. Collection and Analysis of Input Data

The dataset encompasses a comprehensive collection of Superstore Sales data over four years, detailing transactions in furniture and office supplies.

Rigorous data cleaning steps were taken to ensure data quality. This included the removal of extraneous information such as shipment details, customer demographics, and transactional IDs, which are peripheral to the primary objective of sales forecasting.

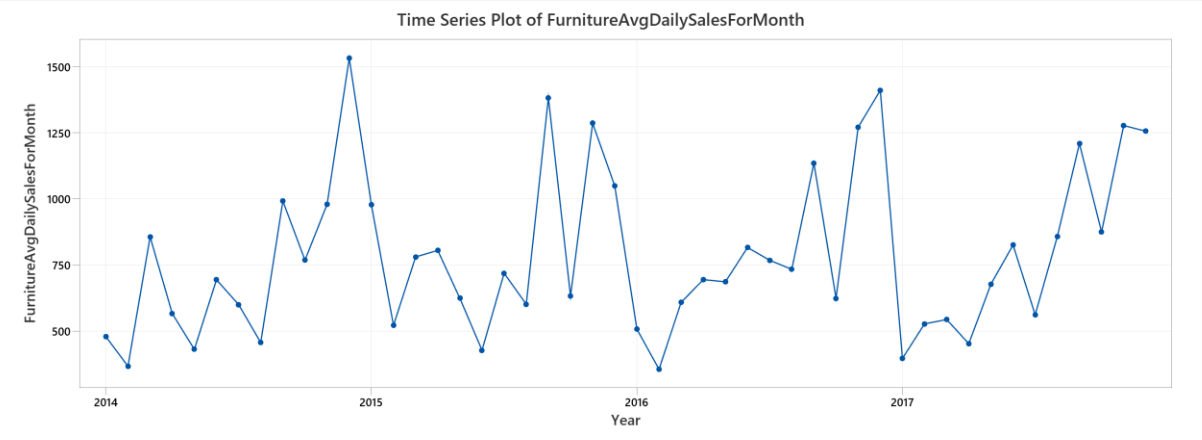
The data underwent a thorough process to address missing or inconsistent entries, ensuring a robust dataset for analysis.

Sales figures were aggregated at a daily level, focusing on the total sales in each furniture and office supplies category separately.

After the data set is collected and cleaning. We have loaded the data into Minitab for Analysis.

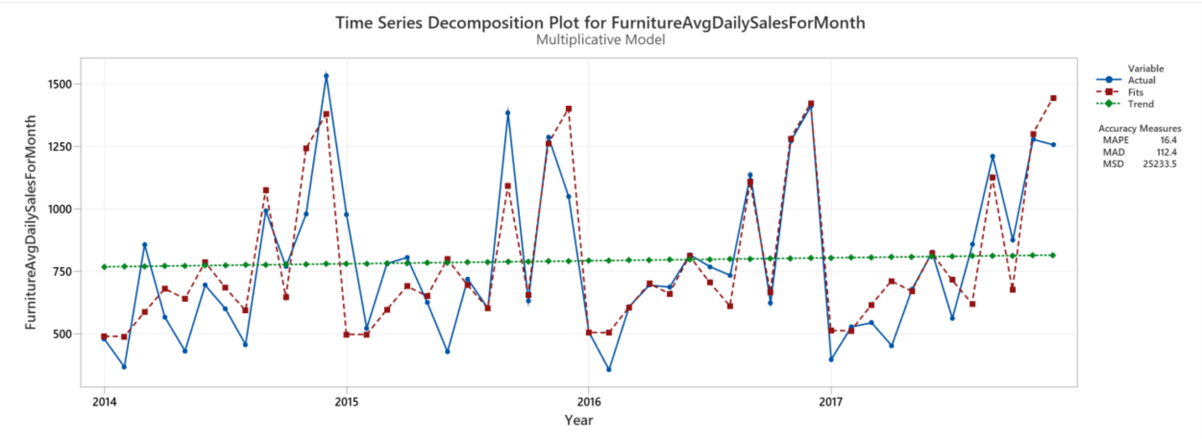
Furniture Sales Analysis:

First, we have plotted the Time Series plot of Furniture Sales to gain some insights.



Distinguishable patterns emerge when visualizing the data. The time series exhibits a recurring seasonality pattern, characterized by consistently low sales at the start of each year and elevated sales toward the year's end. Moreover, there is a noticeable upward trend within each calendar year, accompanied by a few months of lower sales in the mid-year.

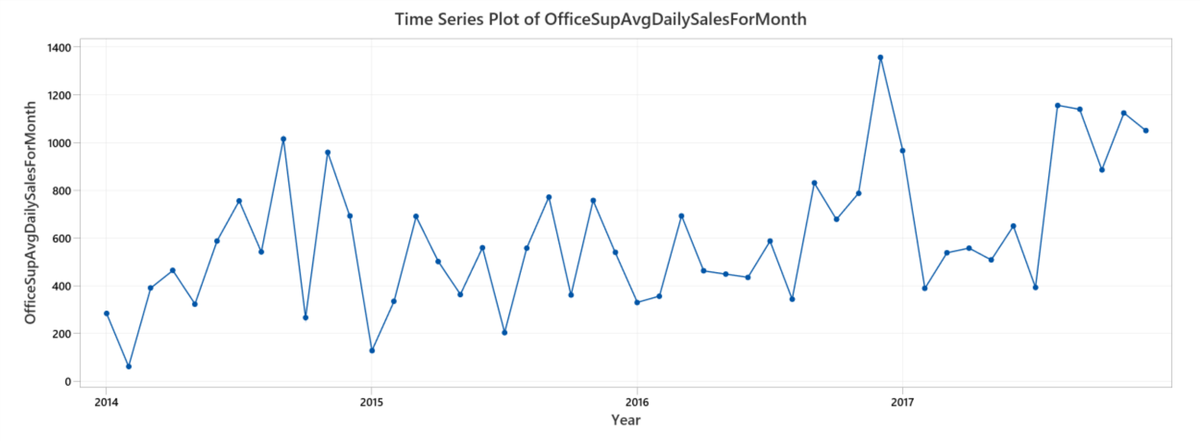
We also visualized our data using time-series decomposition which allowed us to decompose our time series into three distinct components: trend, seasonality, and noise. Based on the time series plot above, which shows fluctuations in sales with both the peaks and troughs increasing over time, a multiplicative model has chosen during time-series decomposition. In a multiplicative model, the seasonal variations are assumed to be proportional to the level of the time series, which seems to be the case here as the amplitude of the seasonal swings appears to increase as the overall sales trend upwards.



The plot above clearly shows that the sales of furniture is unstable, along with its obvious seasonality.

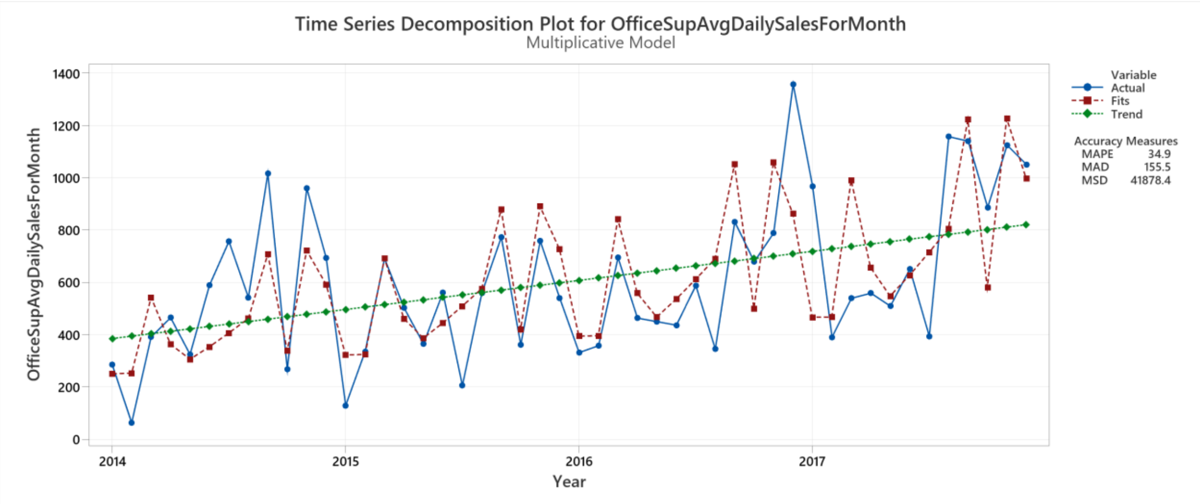
Office Supplies Sales Analysis:

First, we have plotted the Time Series plot of Office Supplies Sales to gain some insights.



There is a clear pattern that suggests seasonality in the sales data. Peaks and troughs appear regularly, indicating that sales might be higher in certain months and lower in others. This could correspond to periods of increased demand due to events like back-to-school seasons or holidays. Moreover, there is a noticeable upward trend within each calendar year, accompanied by a few months of lower sales in the mid-year.

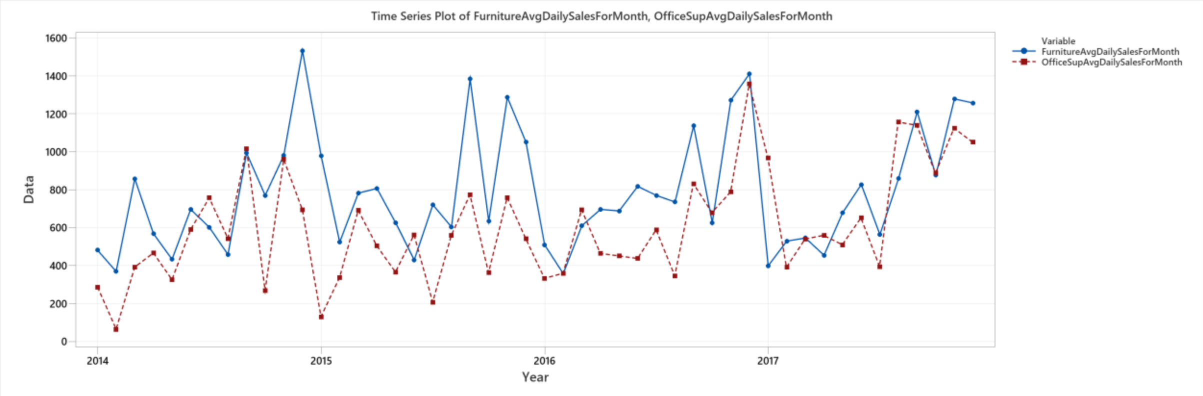
The data was analysed using time-series decomposition, which separates the time series into three components: trend, seasonality, and random noise. The plot illustrates sales data with peaks and valleys that intensify over time. Consequently, a multiplicative model was selected for the decomposition to account for the increasing swings in the sales figures.



The graph presented indicates that office supplies sales are inconsistent, with evident seasonal patterns. There is a marked month-to-month volatility in sales, with pronounced spikes and dips suggesting substantial short-term fluctuations in the sales numbers.

Comparative Analysis:

Time Series of Furniture vs. Office Supplies



We observe that sales of furniture and office supplies shared a similar seasonal pattern. Early of the year is the off season for two categories. It seems summertime is quiet for office supplies too. in addition, average daily sales for furniture are higher than those of office supplies in most of the months. It is understandable, as the value of furniture should be much higher than those of office supplies.

This completes the Analysis part. Next, we move onto the forecasting approaches.

D. Forecasting Approaches and Justifications:

Holt-Winters Method: Opted for its proficiency in addressing time series data that showcases pronounced seasonality and trends, particularly of a multiplicative nature. The Holt-Winters method, also known as Triple Exponential Smoothing, specifically accounts for level, trend, and seasonal components within a dataset, adjusting these elements in a way that can handle both additive and multiplicative seasonal effects. This method is well-suited for retail data where seasonality is a significant factor.

ARIMA Model: Selected for its robustness in modeling time series data that exhibits trends and seasonal variations, ARIMA models excel in capturing linear trends and cyclical behavior within historical data. They are a traditional choice in time series forecasting due to their established theoretical foundations and the ability to integrate differencing to manage non-stationary data.

Furniture Sales Forecasting:

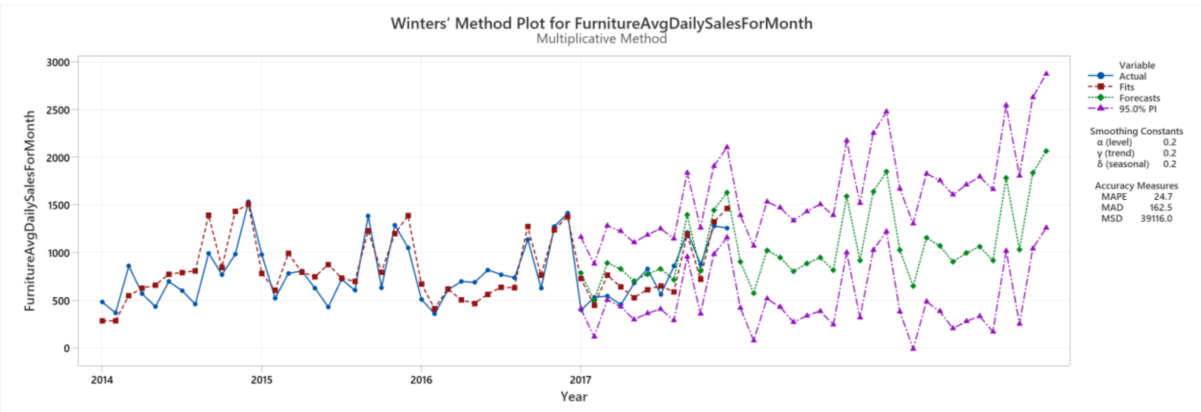
Holt-Winters Method:

We have chosen Holt-Winters method with a multiplicative model as This method is suitable for data with a seasonal pattern and a trend that is proportional to the level of the series.

We have forecasted the data from year 2017 (last year of our actual time series data) to 2019 (future forecasts). Totally 48 forecasts (out of which 36 are future forecasts).

Forecasts

Period	Forecast	Lower	Upper
2017	785.13	408.79	1161.46
2017	497.15	114.92	879.38
2017	890.87	502.06	1279.67
2017	826.07	430.04	1222.09
2017	700.29	296.44	1104.14
2017	772.38	360.12	1184.63
2017	828.01	406.81	1249.20
2017	715.32	284.68	1145.97
2017	1394.60	954.03	1835.17
2017	807.51	356.57	1258.45
2017	1443.56	981.84	1905.28
2017	1629.68	1156.80	2102.57
49	904.62	420.21	1389.03
50	571.86	75.59	1068.13
51	1023.10	514.66	1531.53
52	947.18	426.29	1468.07
53	801.72	268.11	1335.34
54	882.92	336.33	1429.51
55	945.12	385.32	1504.91
56	815.32	242.10	1388.53
57	1587.30	1000.47	2174.14
58	917.82	317.17	1518.46
59	1638.54	1023.91	2253.17
60	1847.35	1218.57	2476.13
61	1024.11	381.03	1667.19
62	646.58	-10.94	1304.10
63	1155.33	483.23	1827.43
64	1068.30	381.50	1755.10
65	903.16	201.54	1604.78
66	993.46	276.92	1710.01
67	1062.22	330.65	1793.80
68	915.31	168.61	1662.01
69	1780.00	1018.08	2541.92
70	1028.13	250.90	1805.35
71	1833.51	1040.90	2626.12
72	2065.01	1256.95	2873.08



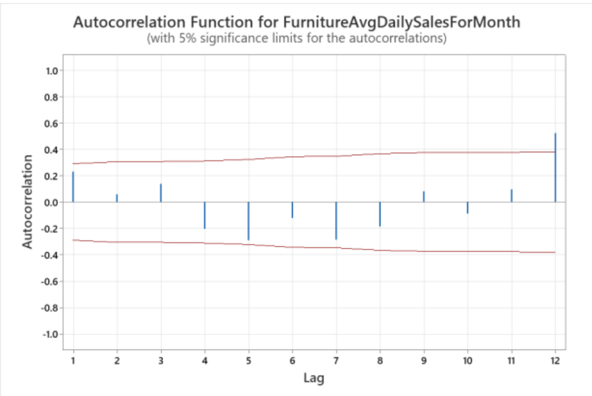
The trend component seems to vary over time, indicating changes in the sales level that are not constant, which the model has fitted with the red dashed line. There is clear seasonality in sales, as indicated by the regular patterns of peaks and troughs in the actual sales line (blue). This seasonality has been captured and reflected in the forecasts as well.

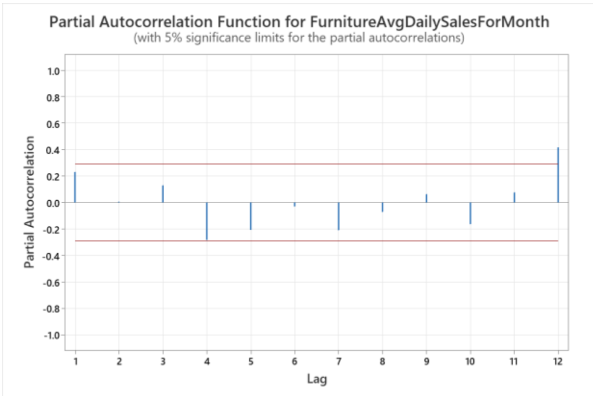
The forecasts (green line) seem to continue the pattern observed in the historical data, suggesting expectations of similar seasonal trends in the future. The 95% PI (purple dotted lines) shows the range within which future sales are expected to fall with a 95% confidence level. The width of this interval indicates the level of uncertainty in the forecasts. The actual sales data points and the model's fitted values show a good alignment, suggesting the model is capturing the underlying pattern well. However, any outliers or anomalies would need to be examined separately.

The plot suggests that while the model is doing a reasonable job of capturing the historical data's trend and seasonality, there is still variability that the model may not fully account for, as indicated by the size of the prediction intervals. This analysis assumes that the underlying patterns in the historical data will continue without significant changes.

ARIMA Model:

We are going to apply one of the most used methods for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average. ARIMA models are denoted with the notation ARIMA (p, d, q). These three parameters account for seasonality, trend, and noise in data. To find the optimal parameters we have plotted the Autocorrelation and Partial Autocorrelation plots also we have performed grid-based search to find the parameters which will produce lowest AICc value which is considered as optimal.





The data seems to exhibit significant autocorrelations at lag 12, you might start with a SARIMA model that includes a seasonal component, such as SARIMA (p, d, q) (P, D, Q)12. In order to get (p, d, q) (P, D, Q) values we performed grid search, the SARIMAX(1, 1, 0)x(1, 1, 0, 12) yields the lowest AICc value of 484.933. Therefore, we should consider this to be optimal option.

Model Selection

Model (d = 1, D = 1)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 0, P = 1, Q = 0*	-239.079	484.933	484.159	488.825
p = 0, q = 1, P = 0, Q = 0	-242.075	488.526	488.151	491.261
p = 1, q = 1, P = 0, Q = 0	-241.961	490.696	489.922	494.588
p = 1, q = 0, P = 0, Q = 1	-243.041	492.856	492.081	496.747
p = 1, q = 0, P = 0, Q = 0	-244.886	494.148	493.773	496.883
p = 1, q = 0, P = 1, Q = 1	-243.050	495.433	494.100	500.321
p = 0, q = 0, P = 0, Q = 1	-247.638	499.650	499.275	502.386
p = 1, q = 1, P = 1, Q = 0	-247.706	504.745	503.412	509.633
p = 0, q = 0, P = 1, Q = 1	-249.463	505.701	504.927	509.593
p = 0, q = 1, P = 1, Q = 0	-250.327	507.428	506.654	511.320
p = 1, q = 1, P = 1, Q = 1	-248.400	508.868	506.799	514.576
p = 0, q = 1, P = 1, Q = 1	-249.898	509.130	507.797	514.018
p = 0, q = 1, P = 0, Q = 1	-251.302	509.378	508.603	513.269
p = 1, q = 1, P = 0, Q = 1	-250.067	509.467	508.134	514.355
p = 0, q = 0, P = 1, Q = 0	-259.238	522.850	522.475	525.586

** Best model with minimum AICc. Output for the best model follows.*

Final Estimates of Parameters

Type		Coef	SE Coef	T-Value	P-Value
AR	1	-0.543	0.148	-3.67	0.001
SAR	12	-0.948	0.126	-7.50	0.000

Differencing: 1 Regular, 1 Seasonal of order 12
Number of observations after differencing: 35

Model Summary

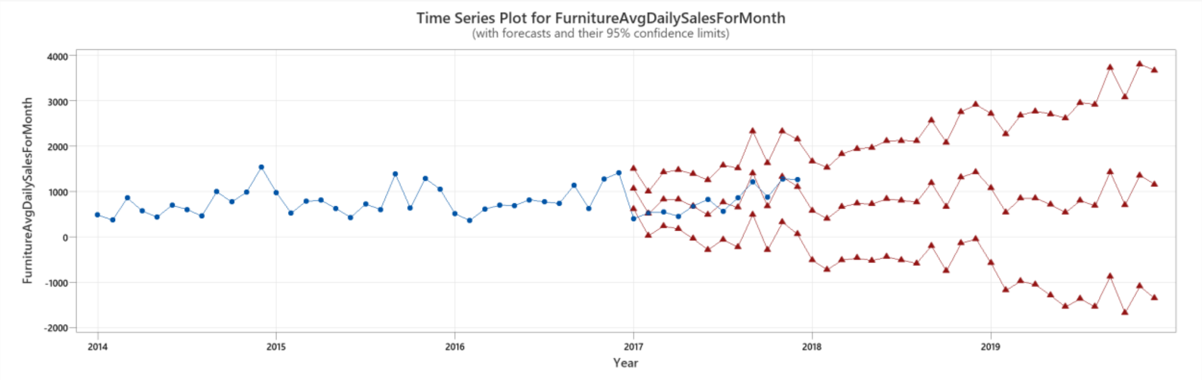
DF	SS	MS	MSD	AICc	AIC	BIC
33	1696383	51405.6	48468.1	484.933	484.159	488.825

MS = variance of the white noise series

We have forecasted the data from year 2017 (last year of our actual time series data) to 2019 (future forecasts). Totally 48 forecasts (out of which 36 are future forecasts).

Forecasts from Time Period 36

			95% Limits		
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
37	1059.33	226.73	614.86	1503.81	397.60
38	509.55	249.24	20.95	998.15	528.18
39	827.72	301.96	235.75	1419.68	544.67
40	823.08	330.39	175.39	1470.77	453.30
41	668.88	364.46	-45.60	1383.36	678.30
42	479.89	391.55	-287.70	1247.47	826.46
43	758.53	418.92	-62.72	1579.78	562.52
44	642.73	443.56	-226.82	1512.28	857.88
45	1404.89	467.44	488.52	2321.26	1209.51
46	666.70	489.88	-293.66	1627.06	875.36
47	1320.49	511.48	317.79	2323.20	1277.82
48	1102.55	532.13	59.37	2145.74	1256.30
49	571.42	555.31	-517.21	1660.05	
50	399.24	575.81	-729.57	1528.05	
51	655.40	596.52	-514.00	1824.81	
52	736.46	616.04	-471.23	1944.15	
53	720.74	635.22	-524.55	1966.04	
54	833.76	653.71	-447.77	2115.29	
55	802.64	671.76	-514.27	2119.55	
56	763.96	689.29	-587.33	2115.25	
57	1184.42	706.42	-200.44	2569.27	
58	661.49	723.12	-756.12	2079.10	
59	1308.34	739.46	-141.29	2757.97	
60	1429.08	755.44	-51.88	2910.03	
61	1068.31	841.21	-580.80	2717.42	
62	538.23	878.42	-1183.81	2260.28	
63	853.16	933.73	-977.32	2683.65	
64	853.00	975.39	-1059.15	2765.14	
65	706.02	1020.72	-1294.99	2707.03	
66	532.78	1061.28	-1547.74	2613.30	
67	795.27	1101.82	-1364.74	2955.27	
68	683.49	1140.14	-1551.64	2918.62	
69	1427.82	1177.63	-880.80	3736.44	
70	700.86	1213.74	-1678.56	3080.28	
71	1354.29	1248.93	-1094.10	3802.69	
72	1154.02	1283.09	-1361.33	3669.38	



The historical data (blue) shows a relatively stable trend with some fluctuations over time. There is no clear long-term upward or downward trend, indicating that the sales levels are consistent year over year. The forecasted sales (red) extend from the last point of historical data into the future. The forecasts seem to follow the same general trend as the historical data with a little upward trend. The confidence intervals (not distinctly shaded but can be inferred from the context) suggest there is uncertainty in the forecasts, with this uncertainty increasing the further out the forecast projects.

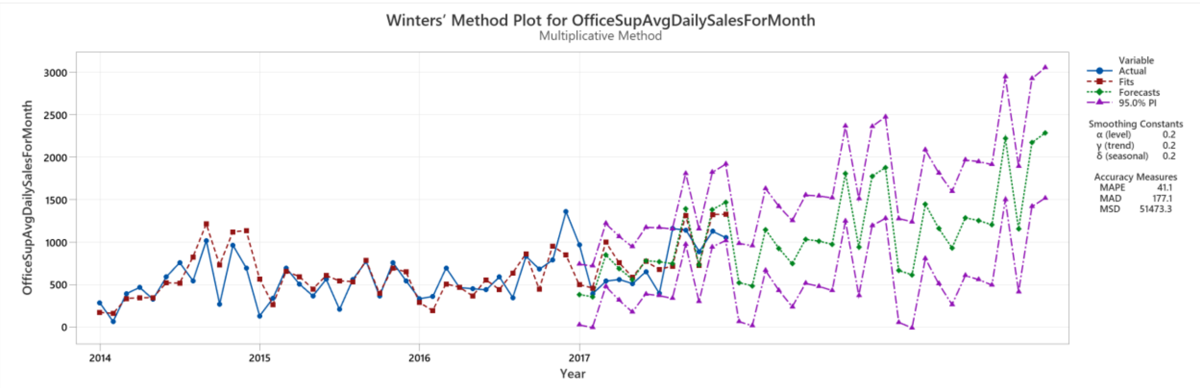
There may be seasonal patterns within the data, indicated by regular ups and downs, which the SARIMA model has considered. The forecasted values also reflect these seasonal variations, suggesting that the model expects these patterns to continue. The plot does not show extreme volatility; the sales values rise and fall in a somewhat consistent range, which the model's confidence intervals seem to capture.

Office Sales Forecasting:

Holt-Winters Method: The Holt-Winters method with a multiplicative model was selected because it is well-suited for datasets exhibiting seasonal patterns and trends that correspond proportionately to the series' magnitude. Forecasts were generated spanning from 2017, the final year of the actual time series data, through to 2019, resulting in a total of 48 forecasted data points, with 36 of these representing projections into the future.

Forecasts

Period	Forecast	Lower	Upper
2017	378.28	20.49	736.07
2017	353.54	-9.86	716.93
2017	845.17	475.52	1214.81
2017	687.16	310.66	1063.67
2017	556.57	172.62	940.52
2017	777.55	385.61	1169.49
2017	766.18	365.74	1166.63
2017	742.54	333.11	1151.96
2017	1387.55	968.69	1806.41
2017	725.62	296.90	1154.34
2017	1379.16	940.19	1818.13
2017	1465.09	1015.51	1914.68
49	520.31	59.77	980.85
50	482.25	10.44	954.07
51	1143.81	660.43	1627.20
52	923.03	427.81	1418.26
53	742.30	234.98	1249.62
54	1030.00	510.34	1549.65
55	1008.39	476.18	1540.60
56	971.24	426.27	1516.21
57	1804.23	1246.31	2362.15
58	938.20	367.15	1509.25
59	1773.58	1189.23	2357.93
60	1874.34	1276.54	2472.13
61	662.34	50.95	1273.73
62	610.97	-14.16	1236.09
63	1442.46	803.48	2081.44
64	1158.90	505.94	1811.86
65	928.03	260.98	1595.07
66	1282.45	601.21	1963.68
67	1250.60	555.07	1946.13
68	1199.95	490.04	1909.86
69	2220.91	1496.53	2945.29
70	1150.79	411.86	1889.72
71	2168.00	1414.45	2921.55
72	2283.58	1515.33	3051.83



While there is no clear long-term trend observable in the historical data, the model seems to suggest a slight increasing trend in the forecast period, although this is within the range of the confidence intervals, suggesting some uncertainty about this trend.

The periodic peaks and troughs in the actual data (blue line) suggest a strong seasonal pattern. The model seems to have fitted these fluctuations well as indicated by the red dashed line, which closely follows the blue line. The green dotted line shows the forecasted sales. It extends the seasonal pattern observed in the historical data into the future, indicating the expected behavior of the sales data based on past trends.

In summary, the plot implies that the Holt-Winters method can capture the seasonal patterns in the data and make forecasts accordingly. The presence of wide prediction intervals and a relatively high MAPE value, however, points to a moderate level of uncertainty in the model's predictions.

ARIMA Model:

We plan to employ ARIMA, a widely recognized approach for forecasting time series data. ARIMA stands for Autoregressive Integrated Moving Average, and its models are expressed using the notation ARIMA (p, d, q), with the parameters representing seasonality, trend, and noise in the dataset. To determine the best parameters, we have conducted a grid search to identify the parameter combination that results in the lowest corrected Akaike Information Criterion (AICc), which is regarded as the most suitable choice. In order to get (p, d, q)(P, D, Q) values we performed grid search, the SARIMAX(1, 1, 1)x(1, 1, 0, 12) yields the lowest AICc value of 497.054. Therefore, we should consider this to be optimal option.

Model Selection

Model (d = 1, D = 1)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 1, P = 1, Q = 0*	-243.860	497.054	495.721	501.942
p = 0, q = 1, P = 0, Q = 0	-249.377	503.129	502.754	505.864
p = 1, q = 0, P = 0, Q = 1	-249.145	505.065	504.290	508.957
p = 1, q = 1, P = 0, Q = 0	-249.403	505.580	504.806	509.472
p = 1, q = 0, P = 1, Q = 0	-249.902	506.579	505.804	510.470
p = 1, q = 0, P = 0, Q = 0	-252.954	510.282	509.907	513.018
p = 0, q = 0, P = 0, Q = 1	-253.514	511.402	511.027	514.138
p = 0, q = 1, P = 1, Q = 1	-251.829	512.991	511.658	517.880
p = 0, q = 1, P = 0, Q = 1	-254.714	516.202	515.428	520.094
p = 1, q = 1, P = 0, Q = 1	-254.739	518.812	517.479	523.700
p = 1, q = 0, P = 1, Q = 1	-255.542	520.417	519.084	525.305
p = 1, q = 1, P = 1, Q = 1	-254.608	521.284	519.215	526.992
p = 0, q = 1, P = 1, Q = 0	-257.507	521.789	521.015	525.681
p = 0, q = 0, P = 1, Q = 1	-263.565	533.904	533.130	537.796
p = 0, q = 0, P = 1, Q = 0	-265.269	534.912	534.537	537.648

** Best model with minimum AICc. Output for the best model follows.*

Final Estimates of Parameters

Type		Coef	SE Coef	T-Value	P-Value
AR	1	0.391	0.173	2.25	0.031
SAR	12	-0.957	0.150	-6.40	0.000
MA	1	0.99066	0.00707	140.17	0.000

Differencing: 1 Regular, 1 Seasonal of order 12
Number of observations after differencing: 35

Model Summary

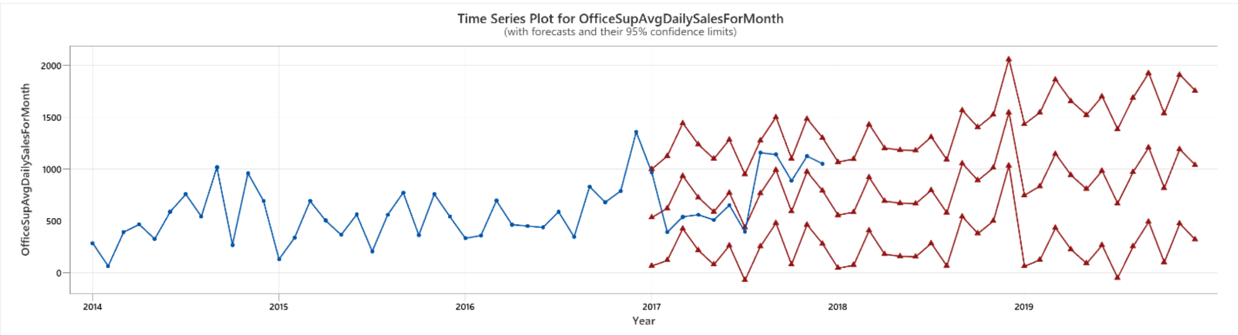
DF	SS	MS	MSD	AICc	AIC	BIC
32	1818476	56827.4	51956.5	497.054	495.721	501.942

MS = variance of the white noise series

We have forecasted the data from year 2017 (last year of our actual time series data) to 2019 (future forecasts). Totally 48 forecasts (out of which 36 are future forecasts).

Forecasts from Time Period 36

			95% Limits		
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
37	530.98	238.385	63.65	998.31	967.01
38	621.01	256.757	117.67	1124.36	389.88
39	932.82	259.777	423.56	1442.09	538.90
40	726.21	260.377	215.77	1236.66	558.23
41	587.26	260.537	76.50	1098.01	508.78
42	771.64	260.601	260.76	1282.52	650.46
43	437.11	260.640	-73.85	948.07	393.90
44	764.83	260.671	253.81	1275.85	1156.15
45	989.76	260.698	478.68	1500.83	1139.14
46	589.90	260.724	78.78	1101.03	886.05
47	974.02	260.750	462.84	1485.19	1124.01
48	789.91	260.776	278.69	1301.13	1049.55
49	554.69	261.143	42.75	1066.64	
50	583.78	261.258	71.61	1095.95	
51	919.01	261.313	406.74	1431.29	
52	689.11	261.350	176.76	1201.46	
53	670.30	261.381	157.89	1182.71	
54	665.34	261.410	152.87	1177.81	
55	795.97	261.439	283.45	1308.49	
56	577.49	261.467	64.91	1090.07	
57	1052.56	261.494	539.93	1565.20	
58	889.55	261.522	376.86	1402.23	
59	1010.85	261.550	498.11	1523.59	
60	1547.73	261.578	1034.94	2060.53	
61	746.92	349.932	60.91	1432.92	
62	834.34	362.673	123.36	1545.33	
63	1147.16	365.062	431.49	1862.82	
64	939.55	365.651	222.73	1656.37	
65	805.73	365.867	88.49	1522.97	
66	982.02	365.987	264.54	1699.50	
67	667.38	366.077	-50.28	1385.03	
68	971.74	366.157	253.93	1689.55	
69	1207.36	366.232	489.40	1925.32	
70	817.64	366.306	99.53	1535.74	
71	1190.51	366.380	472.27	1908.76	
72	1037.23	366.453	318.83	1755.62	



The historical sales data (blue line) fluctuates with a series of peaks and troughs. These fluctuations do not show a clear directional trend over the years leading up to 2017. Instead, there is a cyclical pattern that repeats over time.

The forecasted trend (red line) suggests that the cyclical pattern observed in the historical data is expected to continue. The model seems to predict an ongoing series of rises and falls in sales figures. The forecast indicates that the average daily sales will oscillate around a certain level, without a pronounced long-term upward or downward trajectory. The confidence intervals (not explicitly marked but can be inferred from the 'whiskers' or lines emanating from forecasted points) suggest a range of uncertainty around the forecasts. These intervals appear to increase over time, indicating less certainty in the model's predictions the further out we go. The continuation of a similar pattern of peaks and troughs in the forecast suggests that the model has identified and is accounting for seasonality in the data.

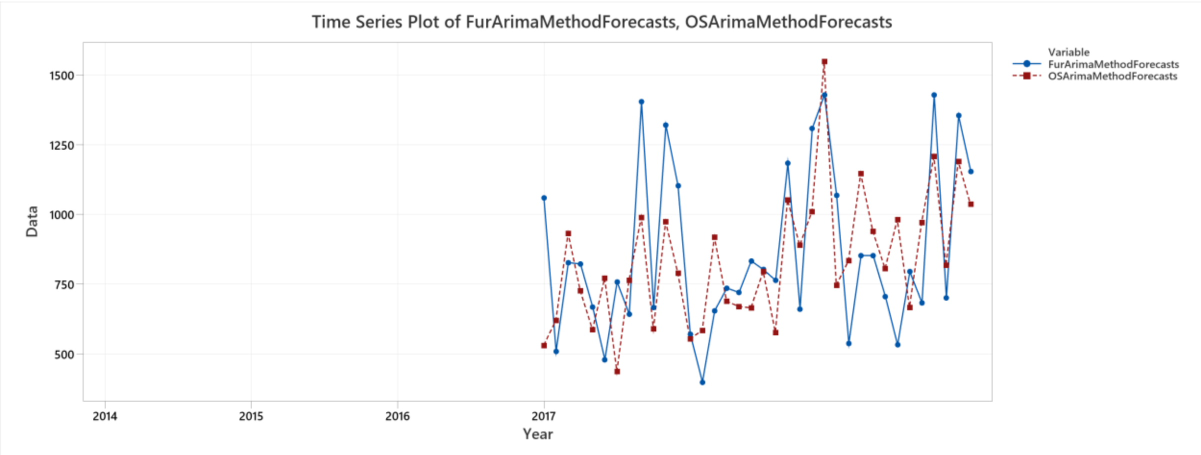
Based on the above forecasting results we think that the ARIMA model performed well over Holt-Winters Method because the forecasts values of ARIMA MODEL are closer to the actual value for historical data (i.e., for year 2017) when compared that of over Holt-Winters forecasts values.

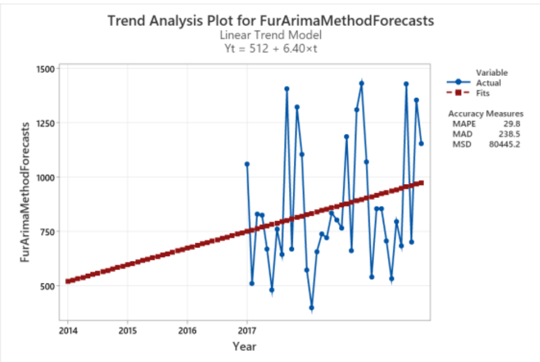
This completes the Forecasting Approaches part.

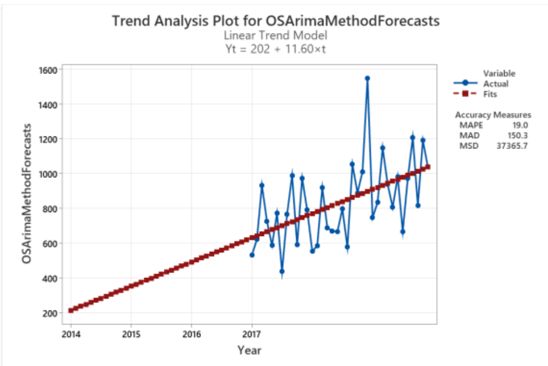
E. Analysis of your output results:

Along with the above analysis results in Sections C&D some additional results include:

Trend and Forecast Visualization:







It's apparent from the trend analysis that both furniture and office supplies sales have been experiencing a steady upward trajectory over the observed period, indicating a consistent rise in sales. The office supplies sector shows a somewhat more pronounced rate of increase, suggesting a more robust growth pattern when compared to furniture sales. This sustained growth for both categories is likely to persist into the future, with office supplies potentially leading in terms of sales expansion.

F. Recommendations to Business Management

Strategic Resource Allocation: Make better use of the forecast data to plan the distribution of resources by matching projected sales patterns with personnel and inventory levels.

Targeted Marketing Initiatives: Create and carry out marketing efforts that are suited to periods of predicted low demand with the goal of piquing consumer interest and increasing sales.

Inventory optimization involves eliminating instances of excess stock and preventing stock shortages by using the predictions' insights to improve inventory management techniques. This improves overall operational efficiency and customer happiness.

G. Suggestions for Future Work

Data Enrichment: By adding other datasets—like online engagement measurements, consumer reviews, and economic indicators—the forecasting models' relevance and accuracy could be greatly increased.

Advanced Modelling approaches: More complex and precise forecasts may be produced by experimenting with state-of-the-art machine learning and artificial intelligence approaches, such as ensemble models and deep learning.

Extending the Analytical Purview: A more comprehensive understanding of the superstore's performance and position in the market would be possible by expanding the research to include more product categories, regional sales variances, and long-term trend analysis.

This expanded report dives further into the study's techniques, analytical insights, and strategic implications, offering a thorough and extensive review of every facet of the research.