Project Report - Time Series Forecasting

Sparkling Wine

DECEMBER 10, 2023

HRITIKA VAISHNAV

INDEX

Contents	Page No.
Problem 1 : Sparkling	4
1. Read the data as an appropriate Time Series data and plot the data…..……..	4
2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition…….	6
3. Split the data into training and test. The test data should start in 1991…...……..….	12
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE…….	14
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05…	27
6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE……	33
7. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data……..……...	40
8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands…..	44
9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales……	45

LIST OF TABLES

1.	Top and Bottom 5 rows of dataset	5
2.	Information about the dataset structure and content.	5
3.	Descriptive statistics	6
4.	Separate columns for the year and month	7
5.	Top and bottom 5 values of new dataset	7
6.	pivot table of dataset	9
7.	Top and bottom rows of train dataset	13
8.	Top and bottom rows of test dataset	13
9.	Top and bottom rows of train dataset of Linear Regression	15
10.	Top and bottom rows of test dataset of Linear Regression	15
11	RMSE matrix value of Linear Regression	16
12	top 5 data of Naive Approach train dataset	17

13	RMSE matrix value of Naïve model	17
14	RMSE matrix value of Simple Average Model	18
15	top 5 data of Moving Average	19
16	RMSE matrix value of Trailing Moving Average 2-9 point	20
17	top 5 data of Simple Exponential Smoothing	21
18	RMSE matrix value alpha 0.995 Simple Exponential Smoothing	22
19	RMSE matrix value alpha 0.1 and 0.995 Simple Exponential Smoothing	23
20	top 5 data of Double Exponential Smoothing	24
21	Test RMSE values of Regression to Double Exponential Smoothing	25
22	top 5 data of Triple Exponential Smoothing	25
23	Test RMSE values of Regression to Triple Exponential Smoothing	26
24	Results of Dickey-Fuller Test	28
25	Results of Dickey-Fuller Test after differencing	29
26	Results of Dickey-Fuller Test after differencing	31
27	Results of Dickey-Fuller Test after differencing	32
28	AIC values in the ascending order	33
29	results_auto_ARIMA.summary	33
30	Test RMSE values of Regression to Auto_ARIMA	34
31	results_auto_ARIMA.summary	35
32	Test RMSE values of ARIMA(2,1,2) & ARIMA(0,1,0)	35
33	top 5 SARIMA6_AIC sort rows	36
34	results_auto_SARIMA6.summary	37
35	SARIMA 6 summary frame	38
36	Test RMSE values of ARIMA to SARIMA	38
37	top 5 SARIMA12_AIC sort rows	38
38	results_auto_SARIMA12.summary	39
39	Test RMSE values of Regression to Auto_SARIMA	40
40	results_auto_ Manual ARIMA.summary	42
41	Test RMSE values of Regression to ARIMA	43
42	results_auto_ Manual SARIMA.summary	43
43	Test RMSE values of Regression to Manual SARIMA	44
44	Test RMSE values of all models in sorted order	45
45	future_predictions rows	45
46	future_predictions with lower_ci and upper_ci	46

LIST OF FIGURES

1.	Time series plot	5
2.	Boxplot of feature list of dataset	7

3.	Yearly Boxplot of feature list of dataset	8
4.	Monthly Boxplot of feature list of dataset	8
5.	Weekly Boxplot of feature list of dataset	9
6.	Weekly Boxplot of feature list of dataset	10
7.	Math_plot of dataset	10
8.	ECDF plot of dataset	11
9.	Additive Decomposition of dataset	11
10.	Multiplicative Decomposition of dataset	12
11.	Shape of train and test dataset	12
12.	Train and test dataset upward trend with seasonality	13
13.	Numerical time series of train and test dataset	14
14.	Train and test dataset behaviour with Linear Regression	16
15.	Train and test dataset behaviour with Linear Regression	17
16.	Train and test dataset behaviour with Linear Regression	18
17.	Moving average models with rolling windows for Train dataset	19
18.	Moving average models with rolling windows for Train and test dataset	20
19.	Model Comparison Plot	21
20.	Simple Exponential Smoothing with alpha 0.995	22
21.	Simple Exponential Smoothing with alpha 0.1 to 0.995	23
22.	Double Exponential Smoothing	24
23.	Triple Exponential Smoothing	26
24.	Rolling Mean & Standard Deviation	28
25.	Rolling Mean & Standard Deviation after differencing	29
26.	Differenced Data Autocorrelation	30
27.	Differenced Data Partial Autocorrelation	30
28.	Rolling Mean & Standard Deviation after differencing	31
29.	Rolling Mean & Standard Deviation after differencing	32
30.	Differenced Data Autocorrelation	34
31.	Differenced Data Partial Autocorrelation	35
32.	Differenced Data Autocorrelation	36
33.	SARIMA diagnostics plot for seasonality as 6	37
34.	SARIMA diagnostics plot for seasonality as 12	39
35.	Sale Differenced Data Partial Autocorrelation	41
36.	Sale Differenced Data Partial Autocorrelation after drop	41
37.	Manual ARIMA diagnostics plot	42
38.	Manual SARIMA diagnostics plot	44
39.	actual and forecast along with the confidence band	46

Problem 1: Sparkling

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

The analysis of sparkling wine sales statistics over the 20th century will be the main subject of this report. I have been given the responsibility of analysing this data as an analyst for ABC Estate Wines in order to spot trends, patterns, and areas where the wine industry may expand. With this information, we can better position our items in the market, plan out our sales tactics, and predict future trends in sales.

The overall goal of this report is to offer insightful information about the wine market and strategies ABC Estate Wines may use to be successful in this fiercely competitive sector.

1. Read the data as an appropriate Time Series data and plot the data.

Import all the necessary and load our data set, Sparkling.csv and use the head() function to view the Top 5 data and the tail() function to view the bottom 5 data. Using the shape function, we can determine that there are 187 rows and 1 column. Find out the characteristics of the column using the info() method. The datatypes for the int64(1) are present and no null values.

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table 1: Top and Bottom 5 rows of dataset

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Table 2: Information about the dataset structure and content.

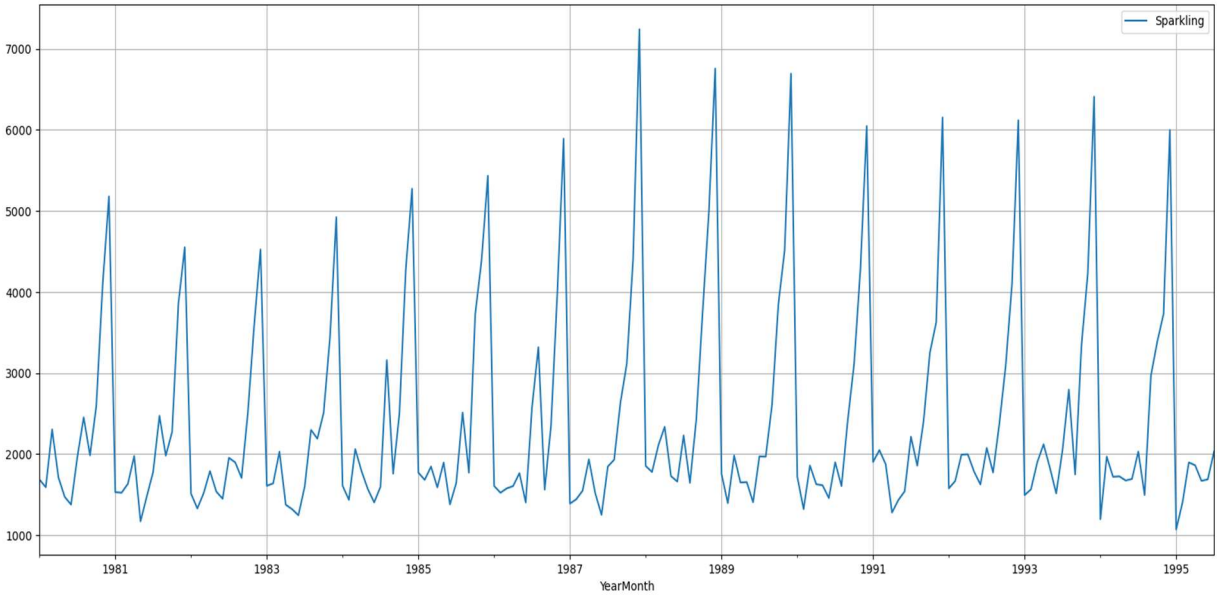


Fig 1 Time series plot

Plot the Time Series to understand the behaviour of the data, we can see that there is an upward trend from 1980 to 1988, then a slight downward trend from 1989 to 1995, with a seasonal pattern associated as well and 1988 is a peek year.

Insights:

- The dataset contains a total of 187 records.
- It consists of 1 column, which is integer data types.
- The dataset does not have null values.
- We can see that there is an upward trend from 1980 to 1988, then a slight downward trend from 1989 to 1995, with a seasonal pattern associated as well and 1988 is a peek year.

2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

Table 3 Descriptive statistics

The average sales of Sparkling Wine per month are around 2402.The maximum sale of the Wine is approx. 7242.The minimum sale of the Wine is approx. 1070.

To understand the spread of accidents across different years and within different months across years.(Create separate columns for the year and month.) for better analysis we divide the data in Year, Month columns and rename the Sparkling as Sales and these columns are integer type.

YearMonth	Sales	Year	Month
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5

Table 4 Separate columns for the year and month

YearMonth	Sales	Year	Month
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5

YearMonth	Sales	Year	Month
1995-03-01	1897	1995	3
1995-04-01	1862	1995	4
1995-05-01	1670	1995	5
1995-06-01	1688	1995	6
1995-07-01	2031	1995	7

Table 5 Top and bottom 5 values of new dataset

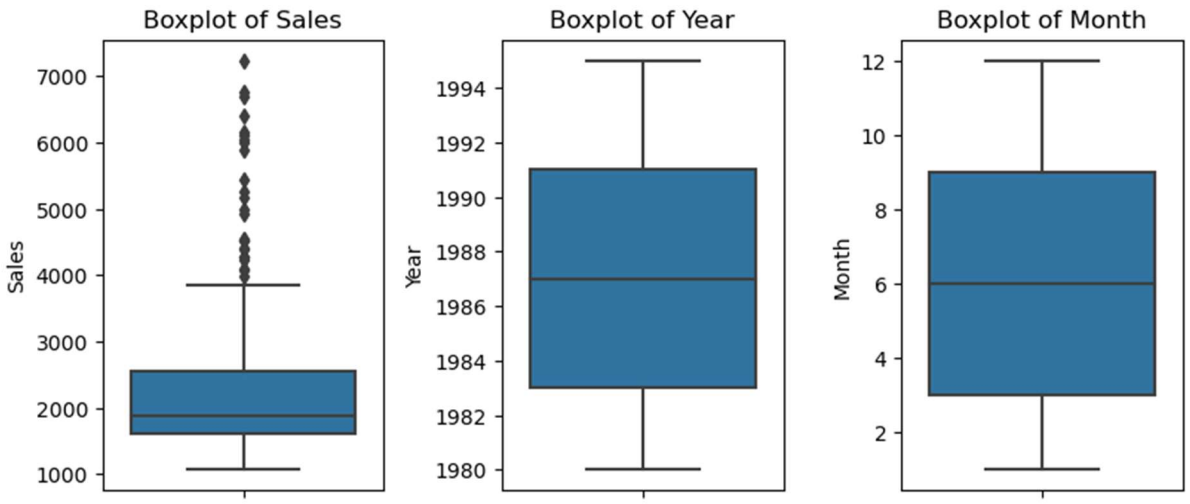


Fig 2 Boxplot of feature list of dataset

The sales boxplot has outliers, but we are choosing not to treat them as they do not have much effect on the time series model.

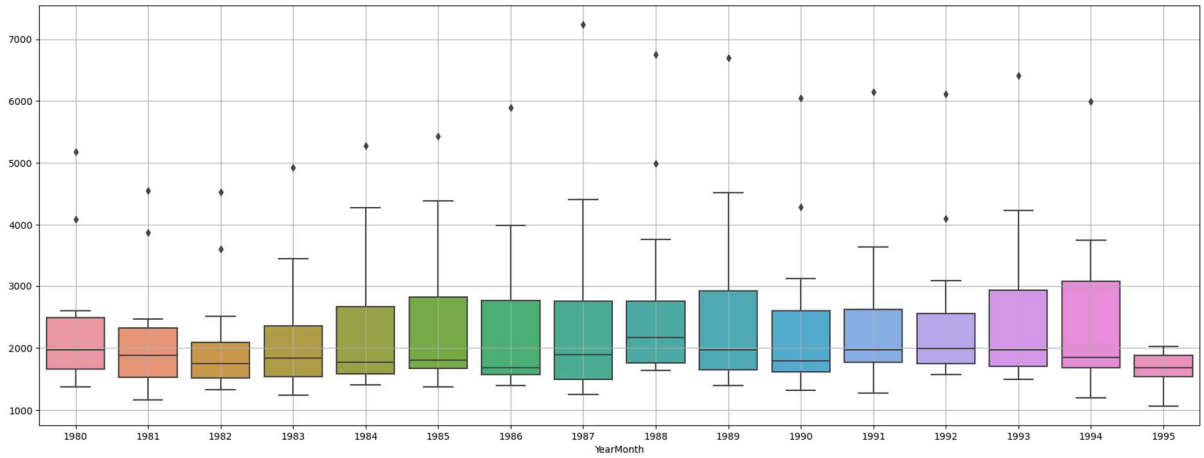


Fig 3 Yearly Boxplot of feature list of dataset

The yearly boxplots show that sales have increased from 1982 to 1985 and also increased in 1987 and 1989, but after 1989, sales decreased towards the last few years and there was a peak in 1988–1989, and that there has been stability throughout time and there are outliers in every year.

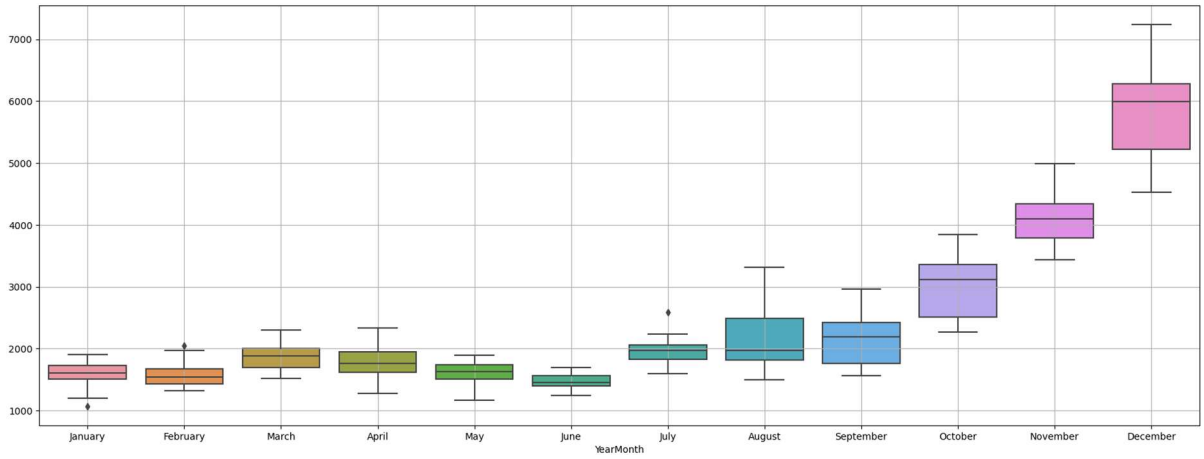


Fig 4 Monthly Boxplot of feature list of dataset

January sales are the lowest, and December sales are the largest. Sales remain steady from January through July, after which they begin to rise in August. January, February, and July months have outliers.

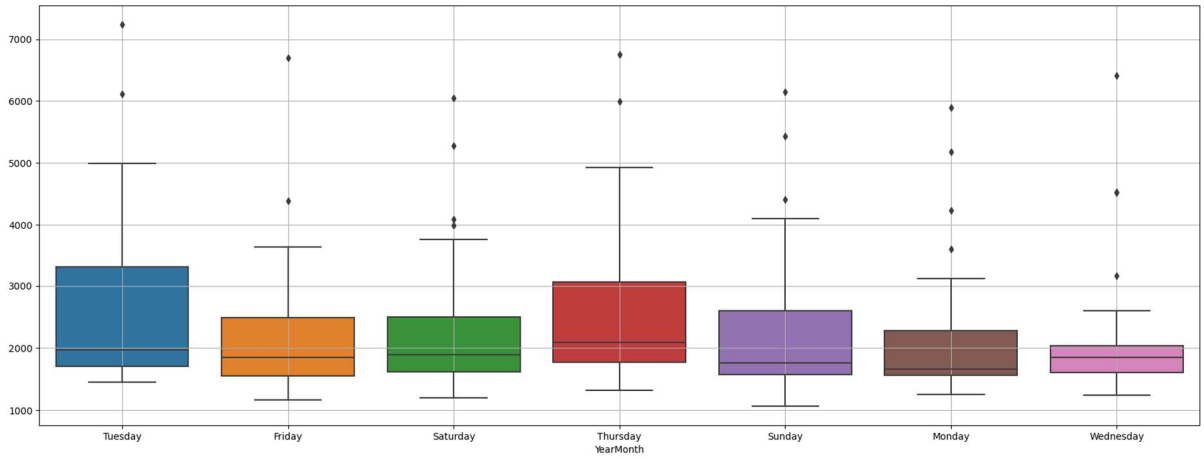


Fig 5 Weekly Boxplot of feature list of dataset

Sales are higher on Tuesday i.e. 5000 and Approx. 5000 sales on Thursday, but Sales are lowest on Wednesday i.e. approx. 2600 and outliers exist every day.

YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	1712.0	2453.0	5179.0	1591.0	1686.0	1966.0	1377.0	2304.0	1471.0	4087.0	2596.0	1984.0
1981	1976.0	2472.0	4551.0	1523.0	1530.0	1781.0	1480.0	1633.0	1170.0	3857.0	2273.0	1981.0
1982	1790.0	1897.0	4524.0	1329.0	1510.0	1954.0	1449.0	1518.0	1537.0	3593.0	2514.0	1706.0
1983	1375.0	2298.0	4923.0	1638.0	1609.0	1600.0	1245.0	2030.0	1320.0	3440.0	2511.0	2191.0
1984	1789.0	3159.0	5274.0	1435.0	1609.0	1597.0	1404.0	2061.0	1567.0	4273.0	2504.0	1759.0
1985	1589.0	2512.0	5434.0	1682.0	1771.0	1645.0	1379.0	1846.0	1896.0	4388.0	3727.0	1771.0
1986	1605.0	3318.0	5891.0	1523.0	1606.0	2584.0	1403.0	1577.0	1765.0	3987.0	2349.0	1562.0
1987	1935.0	1930.0	7242.0	1442.0	1389.0	1847.0	1250.0	1548.0	1518.0	4405.0	3114.0	2638.0
1988	2336.0	1645.0	6757.0	1779.0	1853.0	2230.0	1661.0	2108.0	1728.0	4988.0	3740.0	2421.0
1989	1650.0	1968.0	6694.0	1394.0	1757.0	1971.0	1406.0	1982.0	1654.0	4514.0	3845.0	2608.0
1990	1628.0	1605.0	6047.0	1321.0	1720.0	1899.0	1457.0	1859.0	1615.0	4286.0	3116.0	2424.0
1991	1279.0	1857.0	6153.0	2049.0	1902.0	2214.0	1540.0	1874.0	1432.0	3627.0	3252.0	2408.0
1992	1997.0	1773.0	6119.0	1667.0	1577.0	2076.0	1625.0	1993.0	1783.0	4096.0	3088.0	2377.0
1993	2121.0	2795.0	6410.0	1564.0	1494.0	2048.0	1515.0	1898.0	1831.0	4227.0	3339.0	1749.0
1994	1725.0	1495.0	5999.0	1968.0	1197.0	2031.0	1693.0	1720.0	1674.0	3729.0	3385.0	2968.0
1995	1862.0	NaN	NaN	1402.0	1070.0	2031.0	1688.0	1897.0	1670.0	NaN	NaN	NaN

Table 6 pivot table of dataset

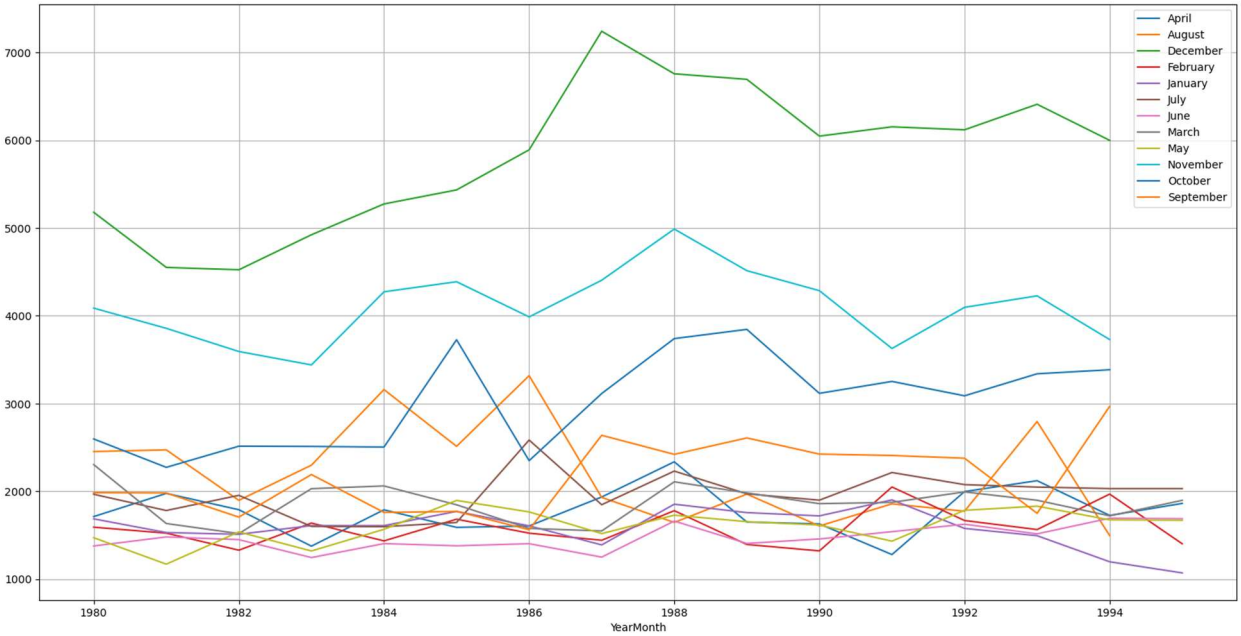


Fig 6 Weekly Boxplot of feature list of dataset

December has the highest sales across the years. According to the graph, 1987 is the peak year because sales were more than 7,000 in December.

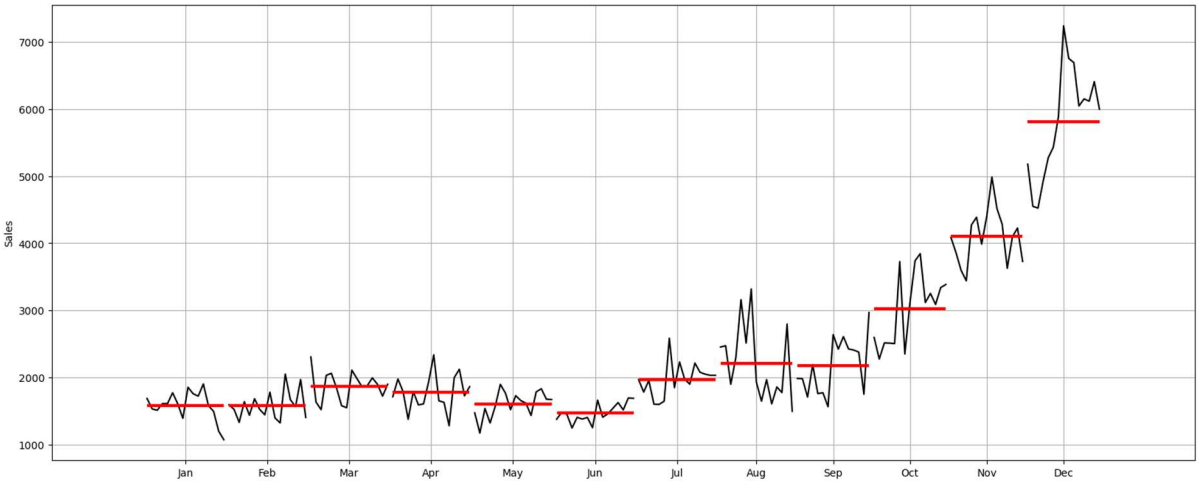


Fig 7 Math_plot of dataset

Sales are seen to increase and decrease across various months, but in December, sales are highly increasing.

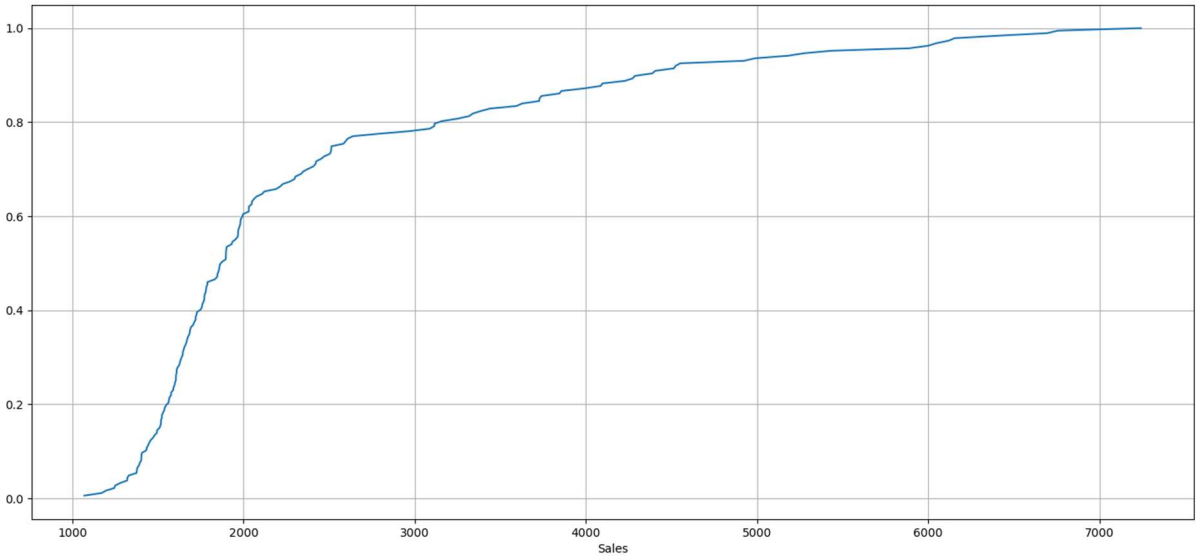


Fig 8 ECDF plot of dataset

More than 50% of sales are below 2000 sales, above 80% of sales are more than 3000 sales, and 100% of sales are 7000 sales.

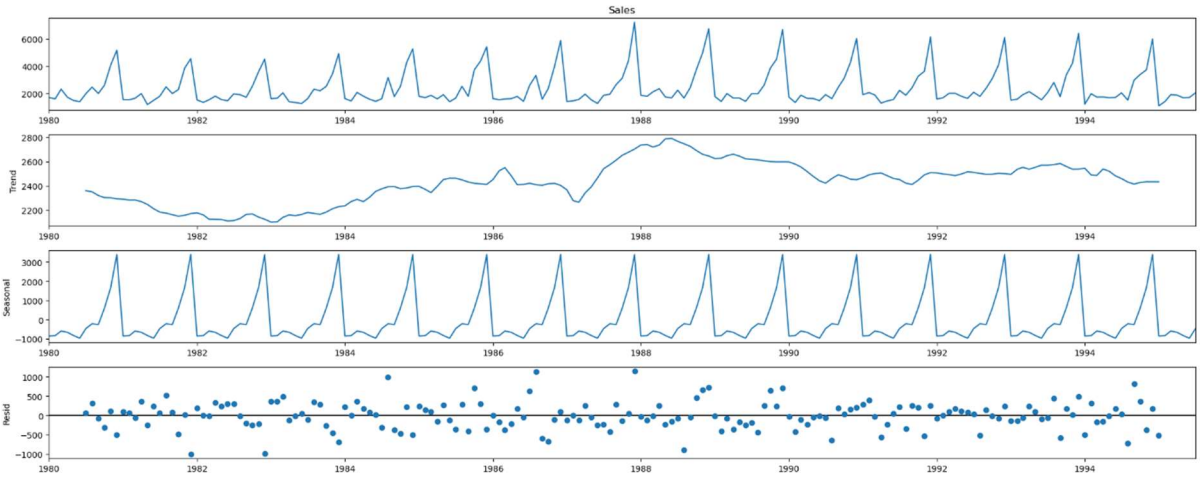


Fig 9 Additive Decomposition of dataset

Years of peak: 1988–1989, it demonstrates that the trend has weakened from 1988 and 1989. Instead of being in a straight line, residue is dispersed. Seasonality and trends are both present.

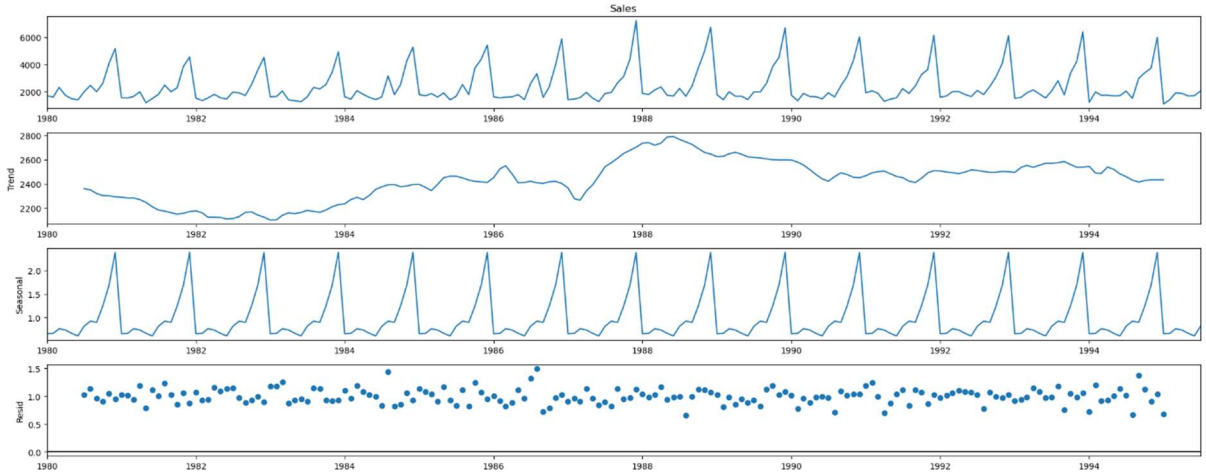


Fig 10 Multiplicative Decomposition of dataset

Peak years is 1988-1989. It also demonstrates the trend's downward movement in the years following 1988–1989. The residue is dispersed and roughly follows a straight path. There is seasonality as well as a trend. Additive is 0 to 1000, whereas live is 0 to 1. Because of the multiplicative model's shorter residual range and more stable residual plot, it is chosen.

Insights:

The sales data exhibits clear seasonal patterns with a notable increase in December sales across all years. Additionally, while there's an overall upward trend until the peak years of 1988–1989, there's a subsequent decline in sales in the years that follow. The choice of the multiplicative model is supported by the observed data patterns, showcasing both seasonality and trend effects.

3. Split the data into training and test. The test data should start in 1991.

As per the instructions given in the project we have split the data, around 1991.With training data from 1980 to 1990 December. Test data starts from the first month of January 1991 till the end.

```
Shape of datasets:
train dataset:  (132, 3)
test dataset:  (55, 3)
```

Fig 11 Shape of train and test dataset

Train dataset has 132 rows and 3 columns. Test dataset has 55 rows and 3 columns.

Rows of dataset:

First few rows of Training Data

YearMonth	Sales	Year	Month
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5

Last few rows of Training Data

YearMonth	Sales	Year	Month
1990-08-01	1605	1990	8
1990-09-01	2424	1990	9
1990-10-01	3116	1990	10
1990-11-01	4286	1990	11
1990-12-01	6047	1990	12

Table 7 Top and bottom rows of train dataset

First few rows of Test Data

YearMonth	Sales	Year	Month
1991-01-01	1902	1991	1
1991-02-01	2049	1991	2
1991-03-01	1874	1991	3
1991-04-01	1279	1991	4
1991-05-01	1432	1991	5

Last few rows of Test Data

YearMonth	Sales	Year	Month
1995-03-01	1897	1995	3
1995-04-01	1862	1995	4
1995-05-01	1670	1995	5
1995-06-01	1688	1995	6
1995-07-01	2031	1995	7

Table 8 Top and bottom rows of test dataset

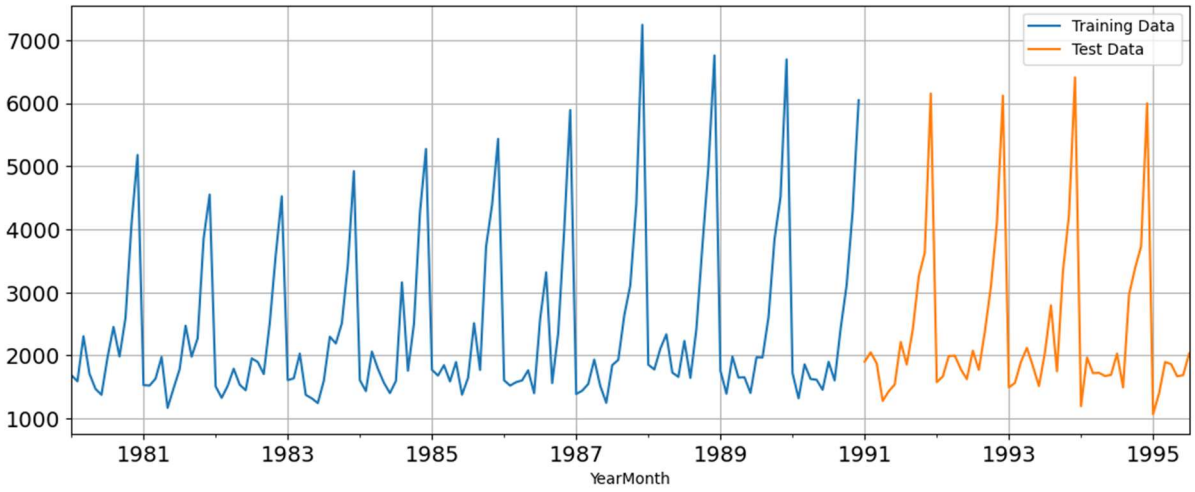


Fig 12 Train and test dataset upward trend with seasonality

The blue curve is the behaviour of a training dataset, and the orange curve is the behaviour of a test dataset.

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

- Model 1: Linear Regression
- Model 2: Naive Approach
- Model 3: Simple Average
- Model 4: Moving Average (MA)
- Model 5: Simple Exponential Smoothing
- Model 6: Double Exponential Smoothing (Holt's Model)
- Model 7: Triple Exponential Smoothing (Holt - Winter's Model)

Model 1: Linear Regression

For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 7
4, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97]
```

Fig 13 Numerical time series of train and test dataset

We see that we have successfully generated the numerical time instance order for both the training and test sets. Now we will add these values to the training and test sets.

First few rows of Training Data

YearMonth	Sales	Year	Month	time
1980-01-01	1686	1980	1	1
1980-02-01	1591	1980	2	2
1980-03-01	2304	1980	3	3
1980-04-01	1712	1980	4	4
1980-05-01	1471	1980	5	5

Last few rows of Training Data

YearMonth	Sales	Year	Month	time
1990-08-01	1605	1990	8	128
1990-09-01	2424	1990	9	129
1990-10-01	3116	1990	10	130
1990-11-01	4286	1990	11	131
1990-12-01	6047	1990	12	132

Table 8 Top and bottom rows of train dataset of Linear Regression

First few rows of Test Data

YearMonth	Sales	Year	Month	time
1991-01-01	1902	1991	1	43
1991-02-01	2049	1991	2	44
1991-03-01	1874	1991	3	45
1991-04-01	1279	1991	4	46
1991-05-01	1432	1991	5	47

Last few rows of Test Data

YearMonth	Sales	Year	Month	time
1995-03-01	1897	1995	3	93
1995-04-01	1862	1995	4	94
1995-05-01	1670	1995	5	95
1995-06-01	1688	1995	6	96
1995-07-01	2031	1995	7	97

Table 9 Top and bottom rows of test dataset of Linear Regression

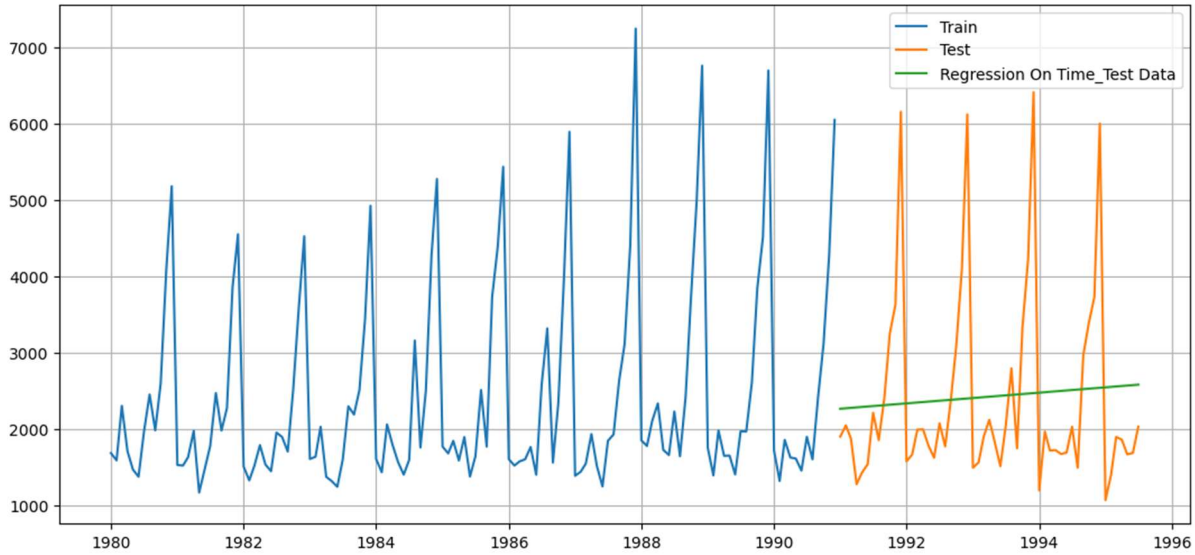


Fig 14 Train and test dataset behaviour with Linear Regression

The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values.

	Test RMSE
RegressionOnTime	1275.867052

Table 11 RMSE matrix value of Linear Regression

The value of Linear Regression with Test RMSE is 1275.867052.

Model 2: Naive Approach

$\hat{y}_{t+1} = \hat{y}_t$

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

```
YearMonth
1991-01-01    6047
1991-02-01    6047
1991-03-01    6047
1991-04-01    6047
1991-05-01    6047
Name: naive, dtype: int64
```

Table 12 top 5 data of Naive Approach train dataset

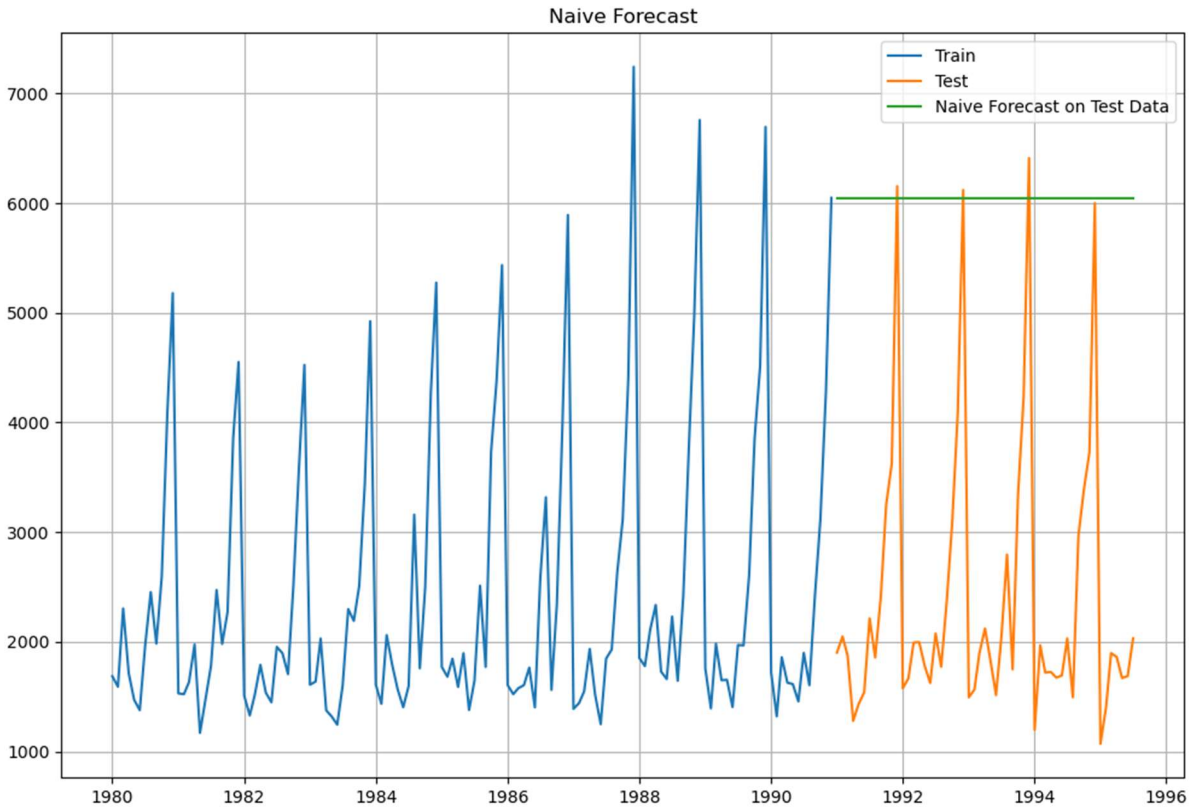


Fig 15 Train and test dataset behaviour with Linear Regression

The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352

Table 13 RMSE matrix value of Naïve model

The value of Naïve model with Test RMSE is 3864.279352.

Method 3: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

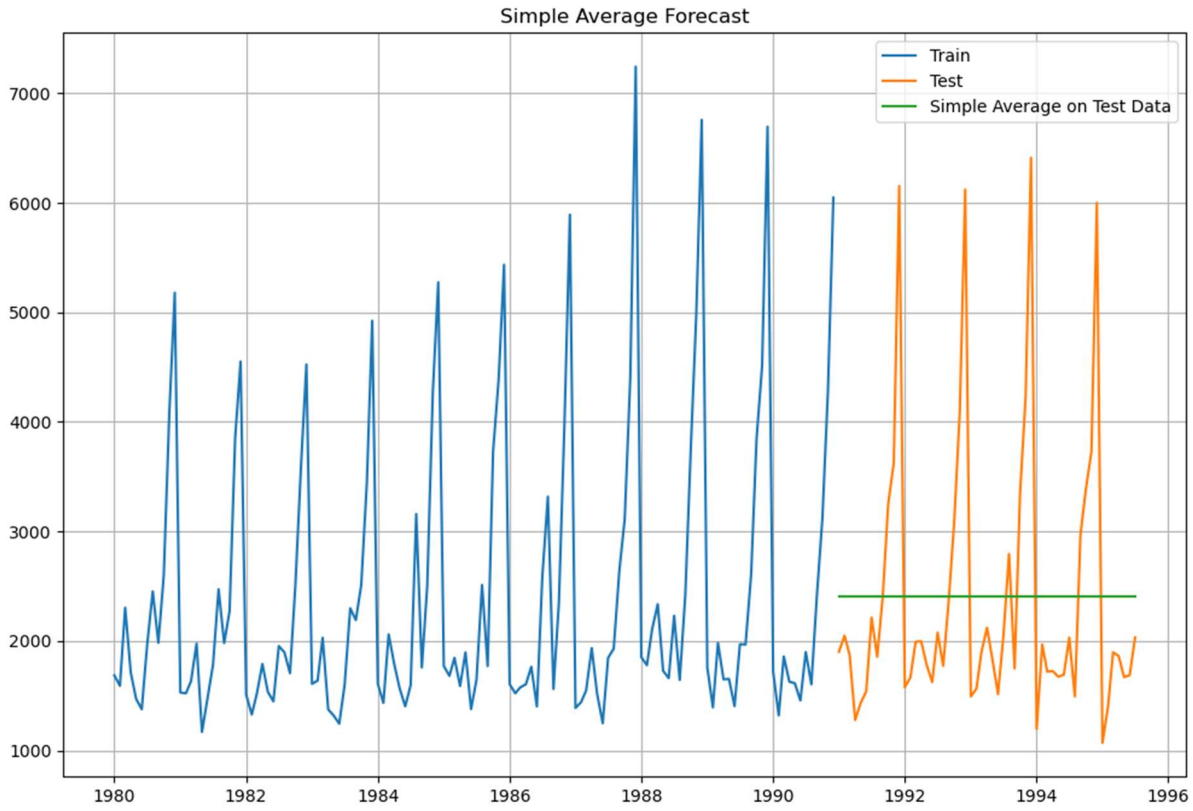


Fig 16 Train and test dataset behaviour with Linear Regression

The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

Table 14 RMSE matrix value of Simple Average Model

The value of Naïve model with Test RMSE is 1275.081804.

Method 4: Moving Average (MA)

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here. For Moving Average, we are going to average over the entire data.

YearMonth	Sales	Year	Month	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1686	1980	1	NaN	NaN	NaN	NaN
1980-02-01	1591	1980	2	1638.5	NaN	NaN	NaN
1980-03-01	2304	1980	3	1947.5	NaN	NaN	NaN
1980-04-01	1712	1980	4	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1980	5	1591.5	1769.50	NaN	NaN

Table 15 top 5 data of Moving Average

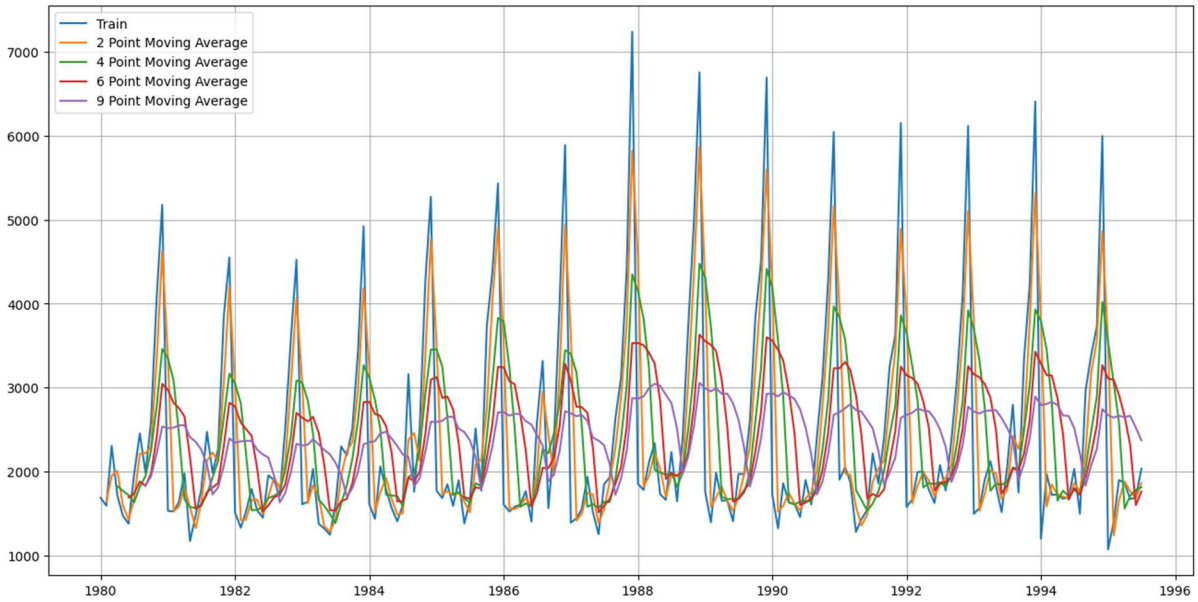


Fig 17 Moving average models with rolling windows for Train dataset

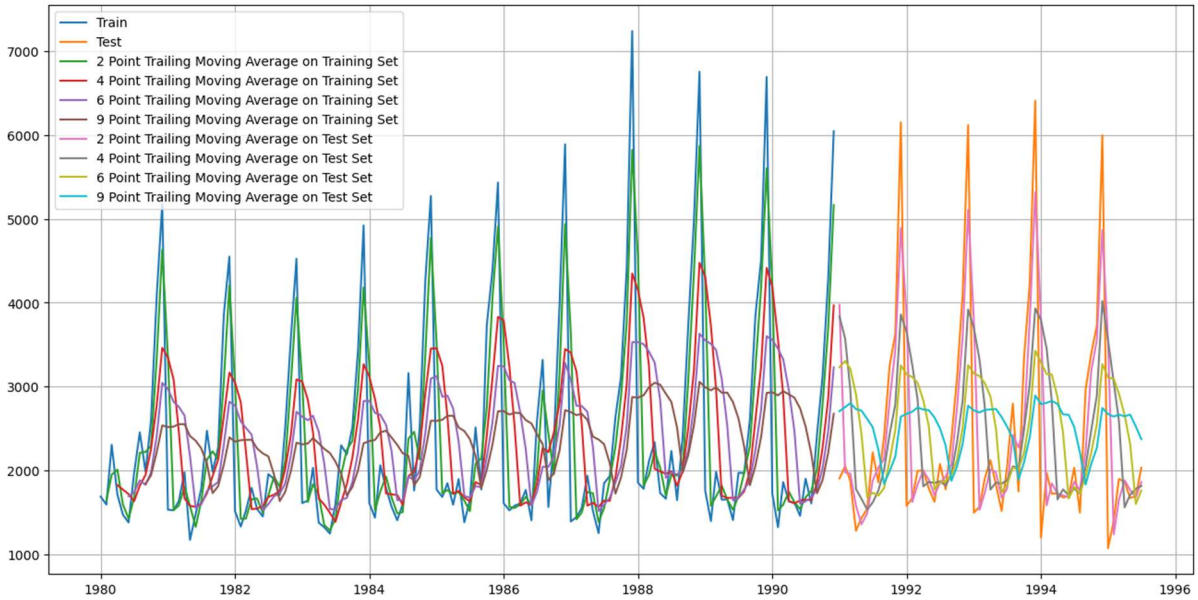


Fig 18 Moving average models with rolling windows for Train and test dataset

We have made multiple moving average models with rolling windows varying from 2 to 9.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

Table 16 RMSE matrix value of Trailing Moving Average 2-9 point

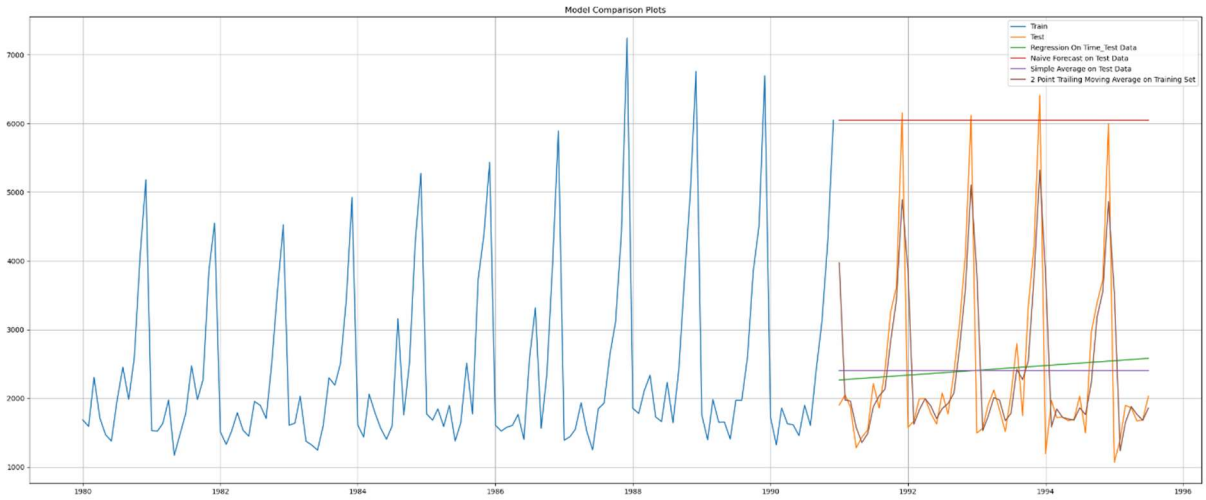


Fig 19 Model Comparison Plot

In model comparison plots, 'Naive Forecast on Test Data' is very far, and 'Regression on Time_Test Data' and 'Simple Average on Test Data' are near compare to all plots.

Method 5: Simple Exponential Smoothing

YearMonth	Sales	Year	Month	predict
1991-01-01	1902	1991	1	2724.932624
1991-02-01	2049	1991	2	2724.932624
1991-03-01	1874	1991	3	2724.932624
1991-04-01	1279	1991	4	2724.932624
1991-05-01	1432	1991	5	2724.932624

Table 17 top 5 data of Simple Exponential Smoothing

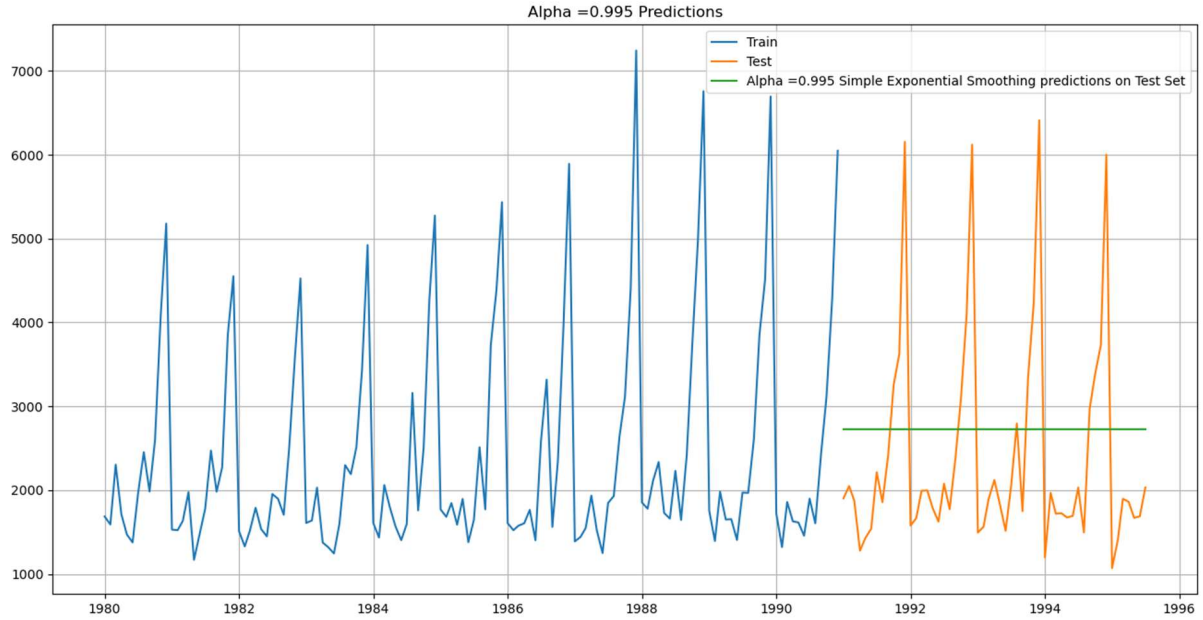


Fig 20 Simple Exponential Smoothing with alpha 0.995

The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values for 'Alpha =0.995 Simple Exponential Smoothing predictions on Test Set'.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487

Table 18 RMSE matrix value alpha 0.995 Simple Exponential Smoothing

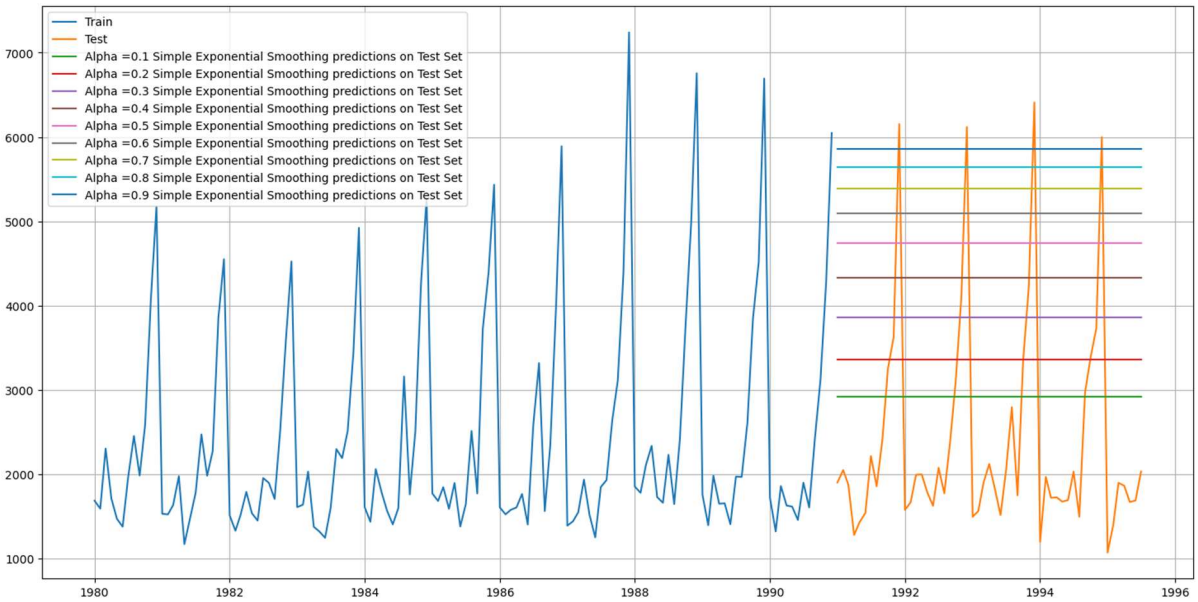


Fig 21 Simple Exponential Smoothing with alpha 0.1 to 0.995

The green line indicates the predictions made by the model, while the orange values are the actual test values. Multiple right lines show the alpha value of 0.1 to 0.9 for simple exponential smoothing predictions on the test set.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398

Table 19 RMSE matrix value alpha 0.1 and 0.995 Simple Exponential Smoothing

Method 6: Double Exponential Smoothing (Holt's Model)

Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model.

YearMonth	Sales	Year	Month	predict
1991-01-01	1902	1991	1	5221.278699
1991-02-01	2049	1991	2	5127.886554
1991-03-01	1874	1991	3	5034.494409
1991-04-01	1279	1991	4	4941.102264
1991-05-01	1432	1991	5	4847.710119

Table 20 top 5 data of Double Exponential Smoothing

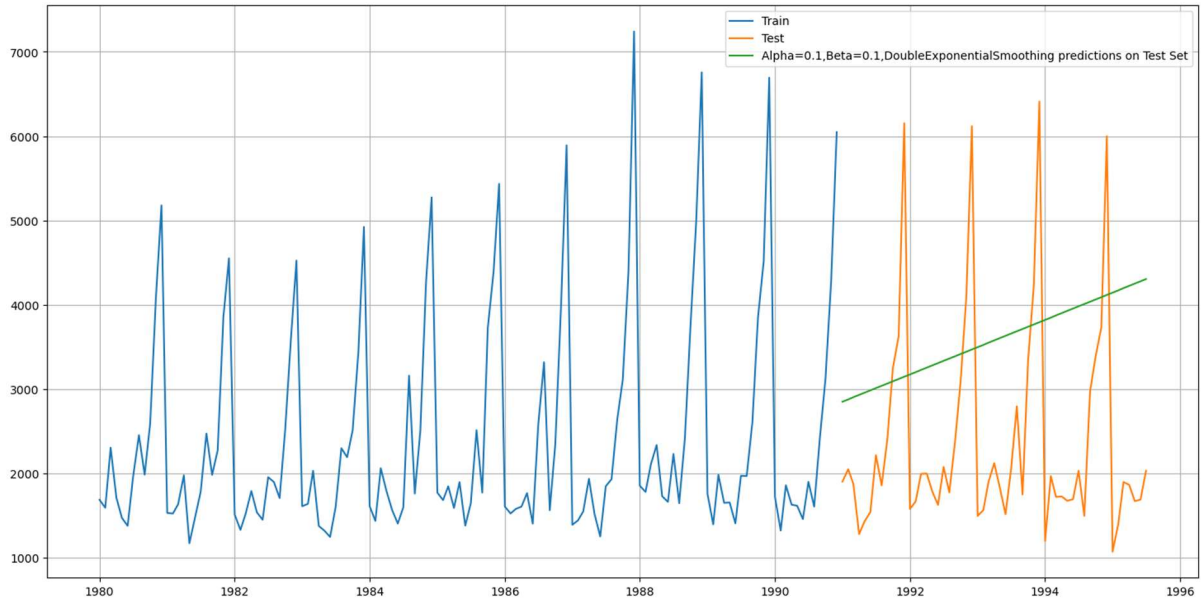


Fig 22 Double Exponential Smoothing

The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670

Table 21 Test RMSE values of Regression to Double Exponential Smoothing

Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

YearMonth	Sales	Year	Month	(predict_ta_sa, 0.1, 0.1, 0.1)	(predict_ta_sa, 0.1, 0.1, 0.2)	(predict_ta_sa, 0.1, 0.1, 0.30000000000000004)	(predict_ta_sa, 0.1, 0.1, 0.4)	(predict_ta_sa, 0.1, 0.1, 0.5)	(predict_ta_sa, 0.1, 0.1, 0.6)	(predict_ta_sa, 0.1, 0.1, 0.7000000000000001)	...
1991-01-01	1902	1991	1	1671.894991	1540.529588	1472.827405	1444.947521	1440.100315	1446.456719	1456.102126	...
1991-02-01	2049	1991	2	1535.938082	1354.094081	1236.723426	1163.127303	1118.381068	1091.681321	1075.381082	...
1991-03-01	1874	1991	3	1882.992874	1728.658127	1644.294990	1605.772780	1593.658780	1593.602194	1595.700721	...
1991-04-01	1279	1991	4	1798.243923	1638.281580	1535.922824	1469.062420	1424.230588	1393.229741	1370.975056	...
1991-05-01	1432	1991	5	1576.572747	1470.697707	1394.544409	1347.223962	1324.218679	1318.006765	1319.843928	...

5 rows × 3461 columns

Table 22 top 5 data of Triple Exponential Smoothing

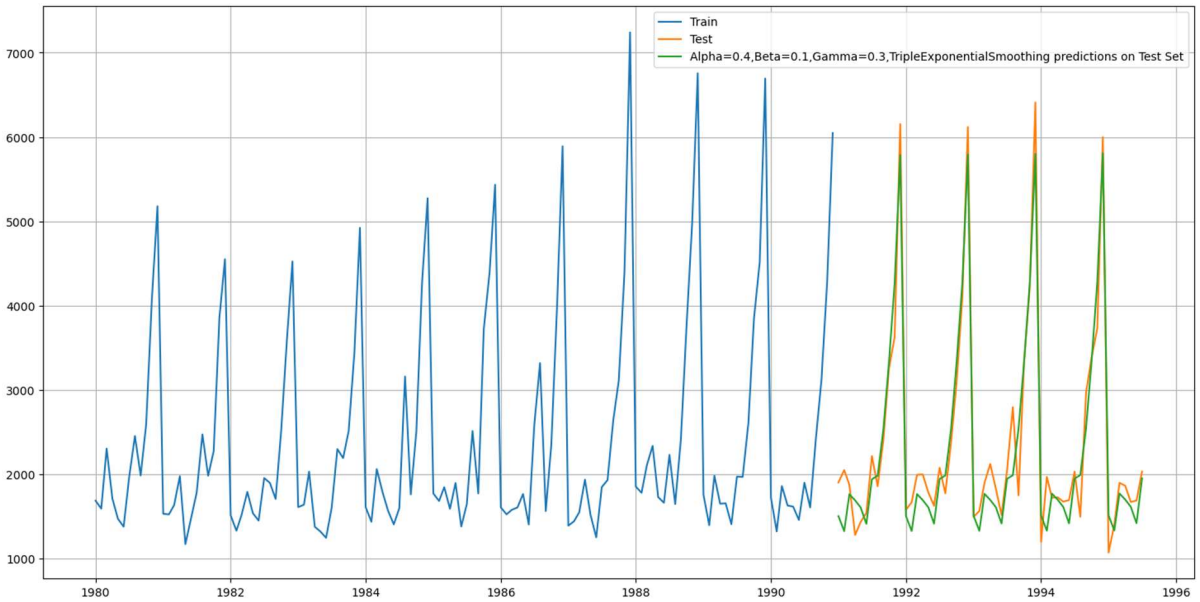


Fig 23 Triple Exponential Smoothing

A best alpha, beta, and gamma values are shown by the green colour line in the above plot. The best model had both a multiplicative trends, as well as a seasonality Model.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.035487
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525

Table 23 Test RMSE values of Regression to Triple Exponential Smoothing

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.
   Note: Stationarity should be checked at alpha = 0.05.

Check for stationarity of the whole Time Series data

The Augmented Dickey-Fuller test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:
H0 : The Time Series has a unit root and is thus non-stationary.
H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than α value.
We see that at 5% significant level the Time Series is non-stationary.

We utilized the differencing strategy to attempt to make the series stationary. We employed the diff() function without any arguments on the current series, assuming a default diff value of 1. We also eliminated the NaN values since order 1 differencing would produce a first value that would need to be deleted as NaN.

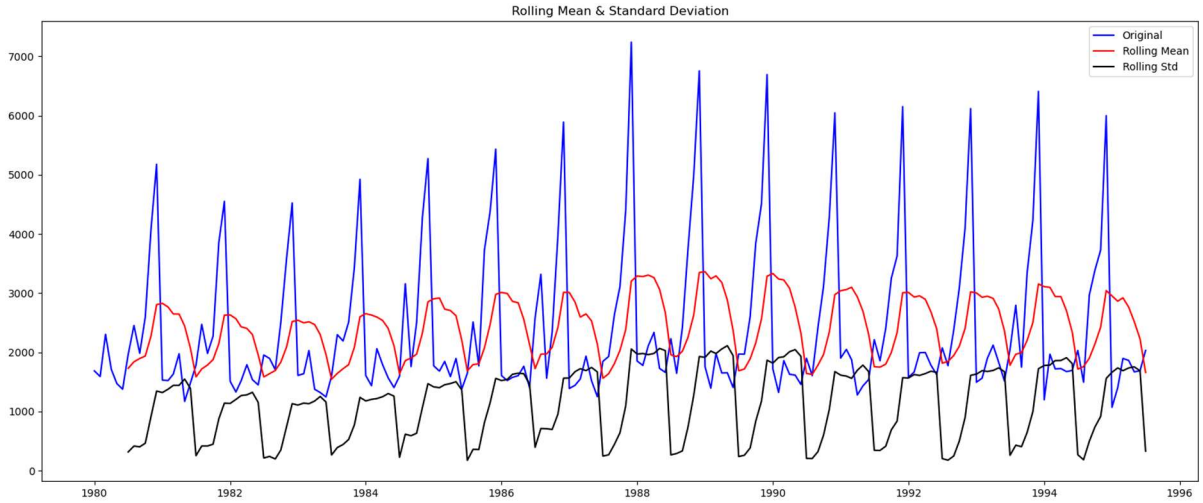


Fig 24 Rolling Mean & Standard Deviation

Results of Dickey-Fuller Test:	
Test Statistic	-1.360497
p-value	0.601061
#Lags Used	11.000000
Number of Observations Used	175.000000
Critical Value (1%)	-3.468280
Critical Value (5%)	-2.878202
Critical Value (10%)	-2.575653
dtype: float64	

Table 24 Results of Dickey-Fuller Test

We see that at 5% significant level the Time Series is non-stationary.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

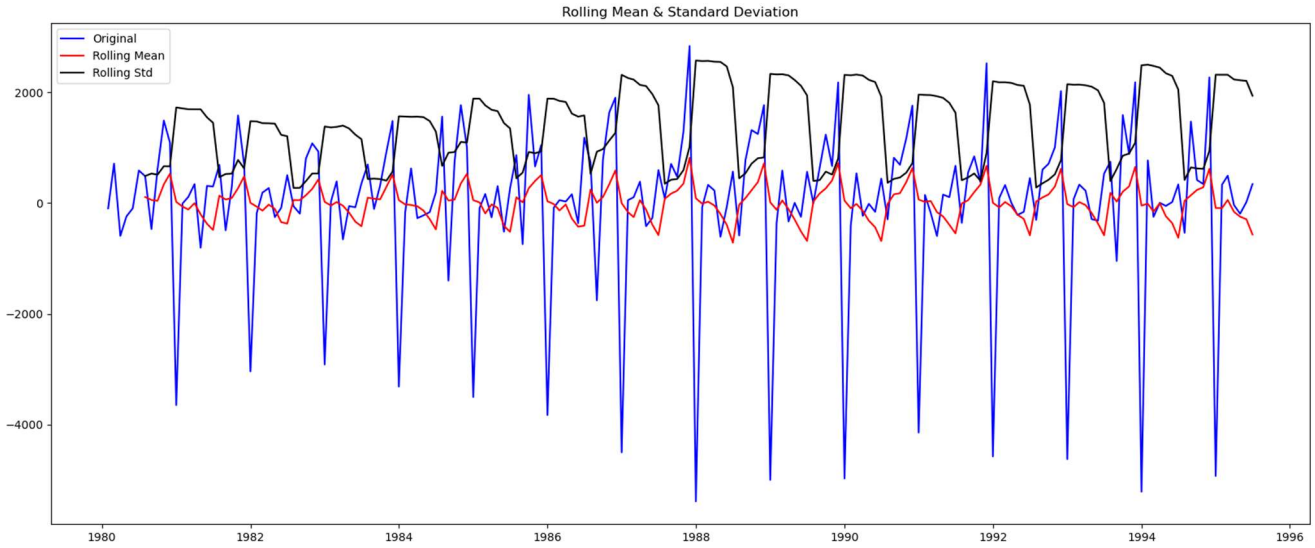


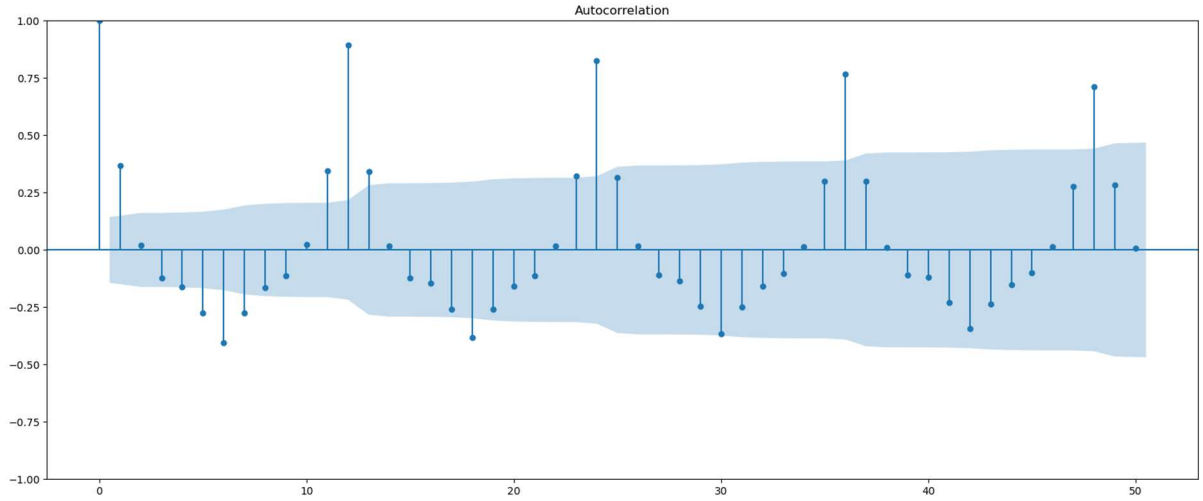
Fig 25 Rolling Mean & Standard Deviation after differencing

Results of Dickey-Fuller Test:	
Test Statistic	-45.050301
p-value	0.000000
#Lags Used	10.000000
Number of Observations Used	175.000000
Critical Value (1%)	-3.468280
Critical Value (5%)	-2.878202
Critical Value (10%)	-2.575653
dtype: float64	

Table 25 Results of Dickey-Fuller Test after differencing

- We see that at $\alpha = 0.05$ the Time Series is indeed stationary.

Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.



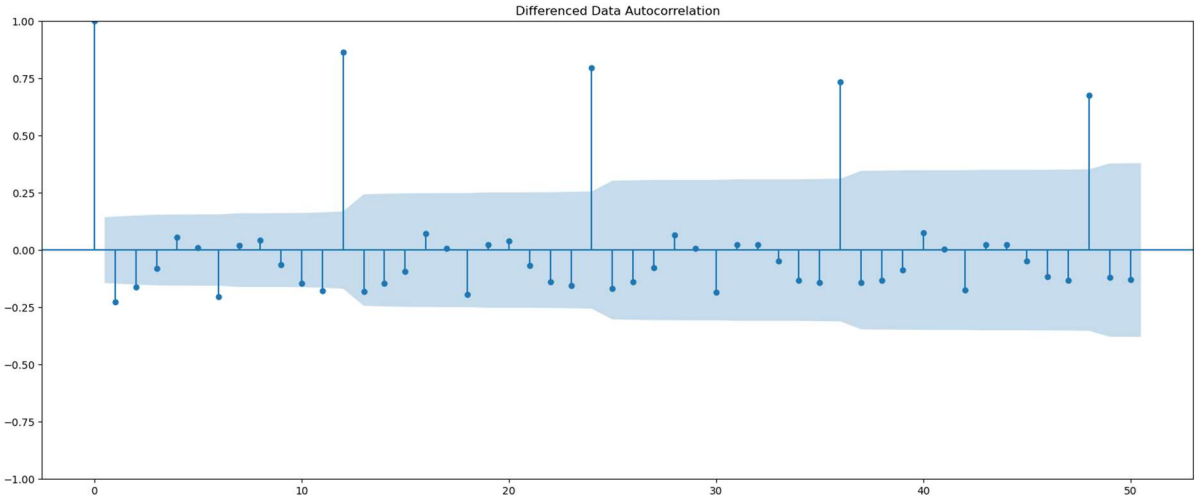


Fig 26 Differenced Data Autocorrelation

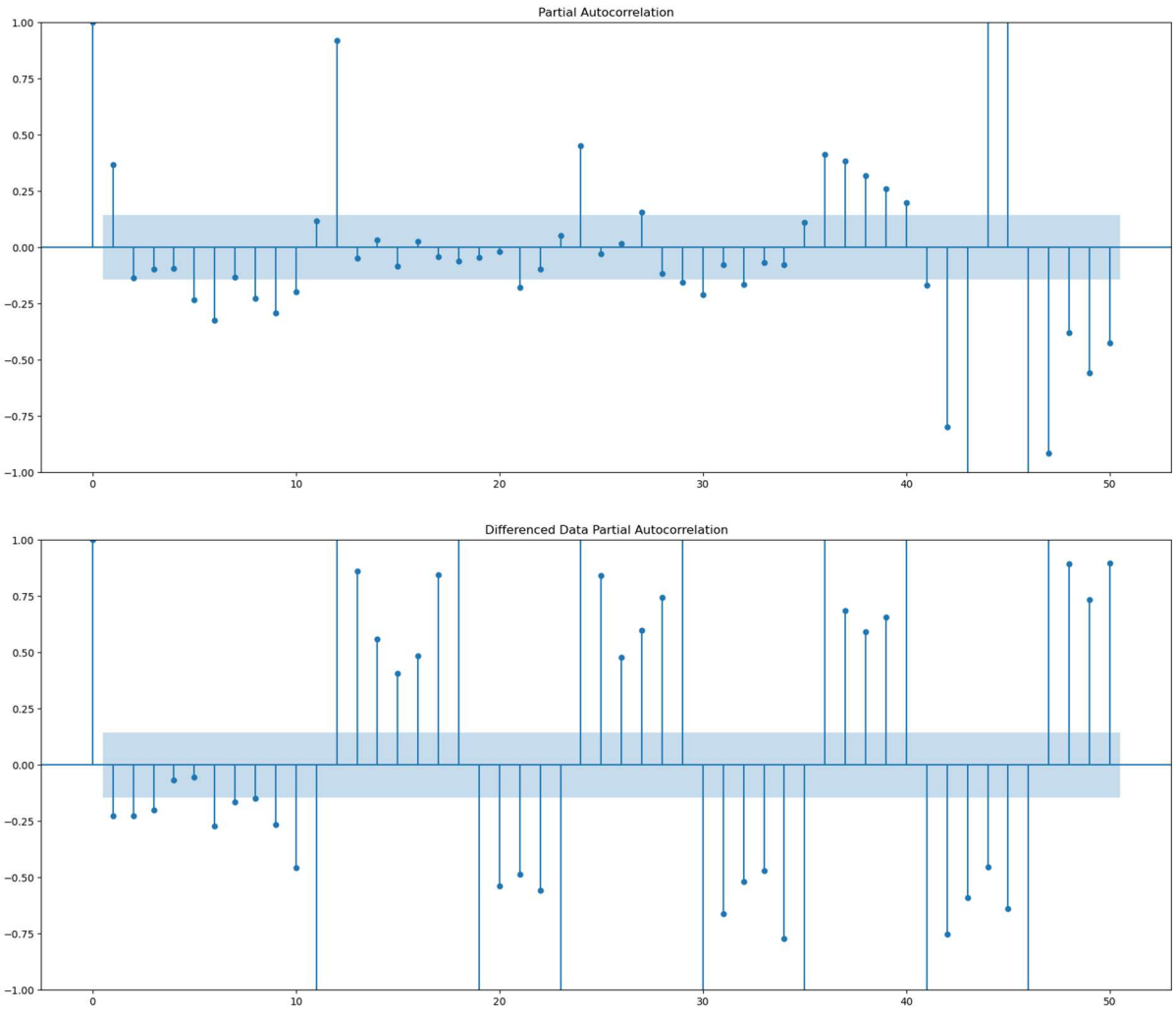


Fig 27 Differenced Data Partial Autocorrelation

Check for stationarity of the Train Time Series data.

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

H0 : Test Time Series has a unit root and is thus non-stationary.

H1 : Test Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

We see that at 5% significant level the Time Series is non-stationary.

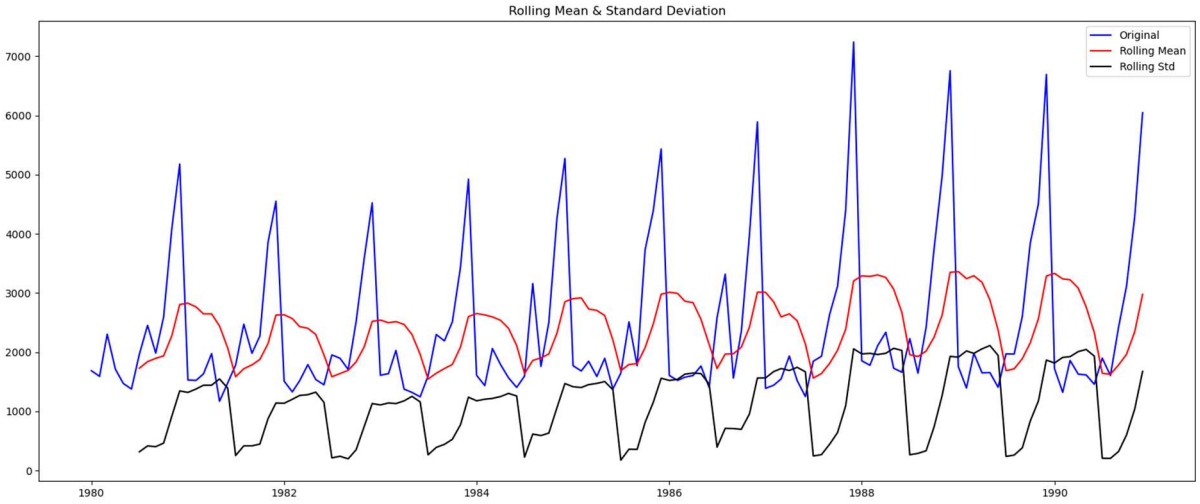


Fig 28 Rolling Mean & Standard Deviation after differencing

Results of Dickey-Fuller Test:	
Test Statistic	-1.208926
p-value	0.669744
#Lags Used	12.000000
Number of Observations Used	119.000000
Critical Value (1%)	-3.486535
Critical Value (5%)	-2.886151
Critical Value (10%)	-2.579896
dtype: float64	

Table 26 Results of Dickey-Fuller Test after differencing

We see that at 5% significant level the Train Time Series is non-stationary.
Let us take a difference of order 1 and check whether the Time Series is stationary or not.

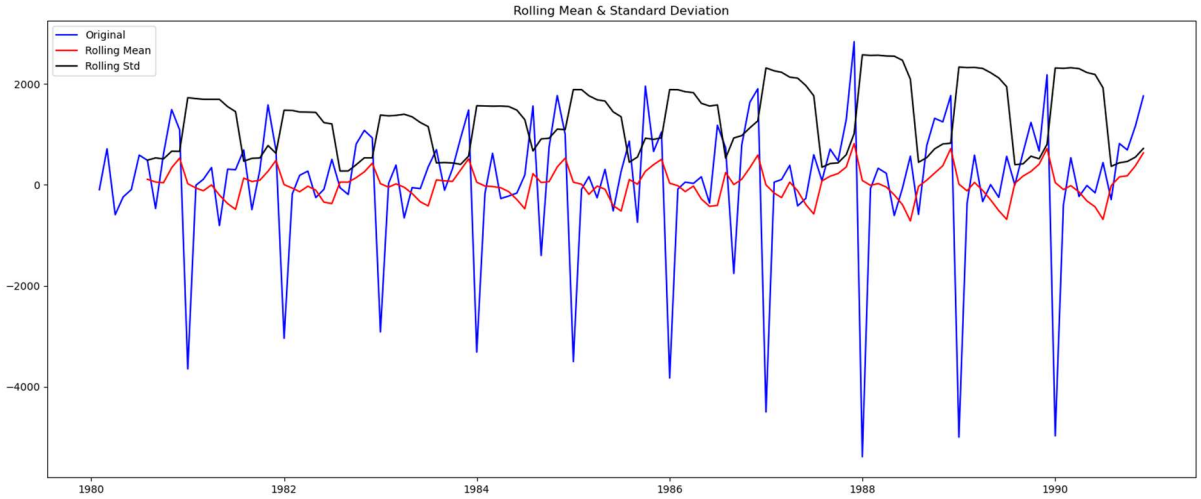


Fig 29 Rolling Mean & Standard Deviation after differencing

Results of Dickey-Fuller Test:	
Test Statistic	-8.005007e+00
p-value	2.280104e-12
#Lags Used	1.100000e+01
Number of Observations Used	1.190000e+02
Critical Value (1%)	-3.486535e+00
Critical Value (5%)	-2.886151e+00
Critical Value (10%)	-2.579896e+00
dtype: float64	

Table 28 Results of Dickey-Fuller Test after differencing

We see that at $\alpha = 0.05$ Train Time Series is indeed stationary.

We see that after taking a difference of order 1 the series have become stationary = 0.05.

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Auto - Arima Model

	param	AIC
8	(2, 1, 2)	2213.509212
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095
6	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
3	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

Table 28 AIC values in the ascending order

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sat, 09 Dec 2023	AIC	2213.509
Time:	15:52:59	BIC	2227.885
Sample:	01-01-1980 - 12-01-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.781	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.741	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.218	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-07	5.51e+12	0.000	1.1e+06	1.1e+06

Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	14.46
Prob(Q):	0.67	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.61
Prob(H) (two-sided):	0.00	Kurtosis:	4.08

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 7.7e+28. Standard errors may be unstable.

Table 29 results_auto_ARIMA.summary

Predict on the Test Set using this model and evaluate the model.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.035487
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525
Auto_ARIMA	1299.979665

Table 30 Test RMSE values of Regression to Auto_ARIMA

Build a version of the ARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.

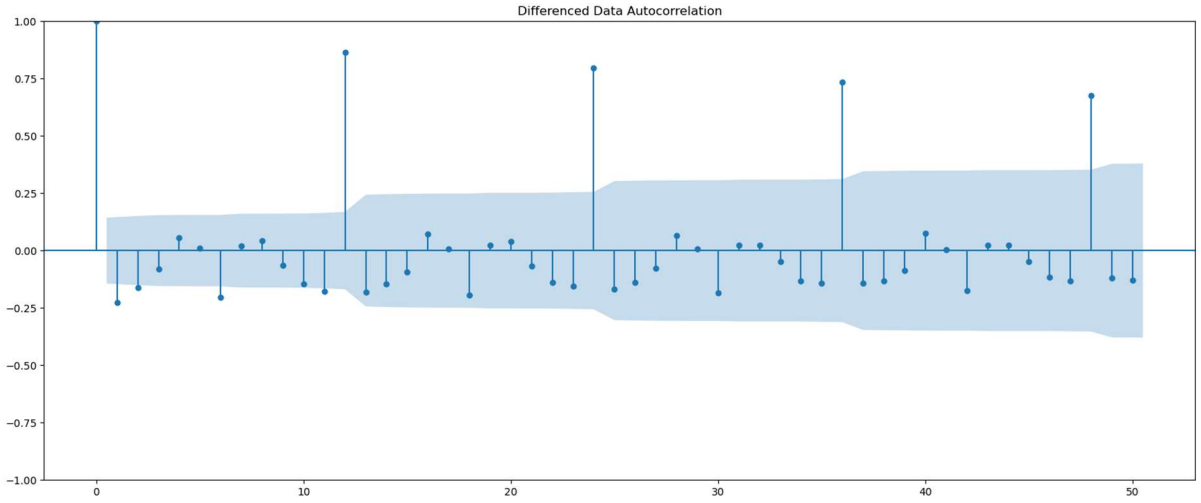


Fig 30 Differenced Data Autocorrelation

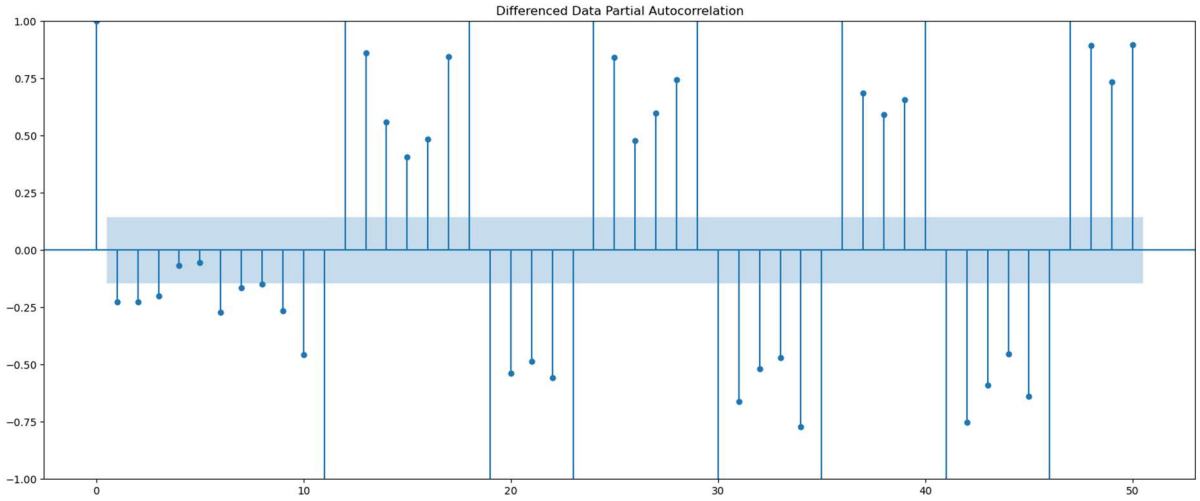


Fig 31 Differenced Data Partial Autocorrelation

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.832
Date:	Sat, 09 Dec 2023	AIC	2267.663
Time:	15:53:00	BIC	2270.538
Sample:	01-31-1980 - 12-31-1990	HQIC	2268.831
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.885e+06	1.29e+05	14.658	0.000	1.63e+06	2.14e+06

Ljung-Box (L1) (Q):	3.07	Jarque-Bera (JB):	198.83
Prob(Q):	0.08	Prob(JB):	0.00
Heteroskedasticity (H):	2.46	Skew:	-1.92
Prob(H) (two-sided):	0.00	Kurtosis:	7.65

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table 31 results_auto_ARIMA.summary

Predict on the Test Set using this model and evaluate the model.

	RMSE
ARIMA(2,1,2)	1299.979665
ARIMA(0,1,0)	3864.279352

Table 32 Test RMSE values of ARIMA(2,1,2) & ARIMA(0,1,0)

Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

AUTO- SARIMA

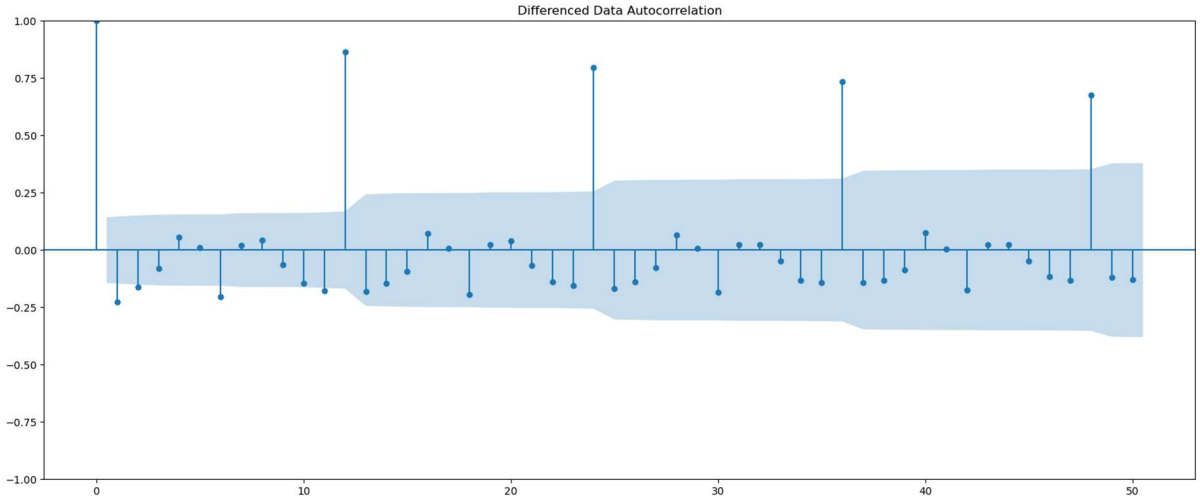


Fig 32 Differenced Data Autocorrelation

Setting the seasonality as 6 for the first iteration of the auto SARIMA model.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678701
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888803
80	(2, 1, 2)	(2, 0, 2, 6)	1729.192621
17	(0, 1, 1)	(2, 0, 2, 6)	1741.703672
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379778

Table 33 top 5 SARIMA6_AIC sort rows

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-856.944
Date:	Sun, 10 Dec 2023	AIC	1727.889
Time:	07:15:09	BIC	1747.164
Sample:	0	HQIC	1735.713
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.7852	0.103	-7.655	0.000	-0.986	-0.584
ma.L2	-0.0975	0.112	-0.870	0.384	-0.317	0.122
ar.S.L6	0.0022	0.026	0.084	0.933	-0.048	0.053
ar.S.L12	1.0396	0.018	58.256	0.000	1.005	1.075
ma.S.L6	0.0428	0.143	0.298	0.766	-0.238	0.324
ma.S.L12	-0.6203	0.090	-6.878	0.000	-0.797	-0.444
sigma2	1.475e+05	1.42e+04	10.371	0.000	1.2e+05	1.75e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	38.96
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	2.85	Skew:	0.58
Prob(H) (two-sided):	0.00	Kurtosis:	5.59

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table 34 results_auto_SARIMA6.summary

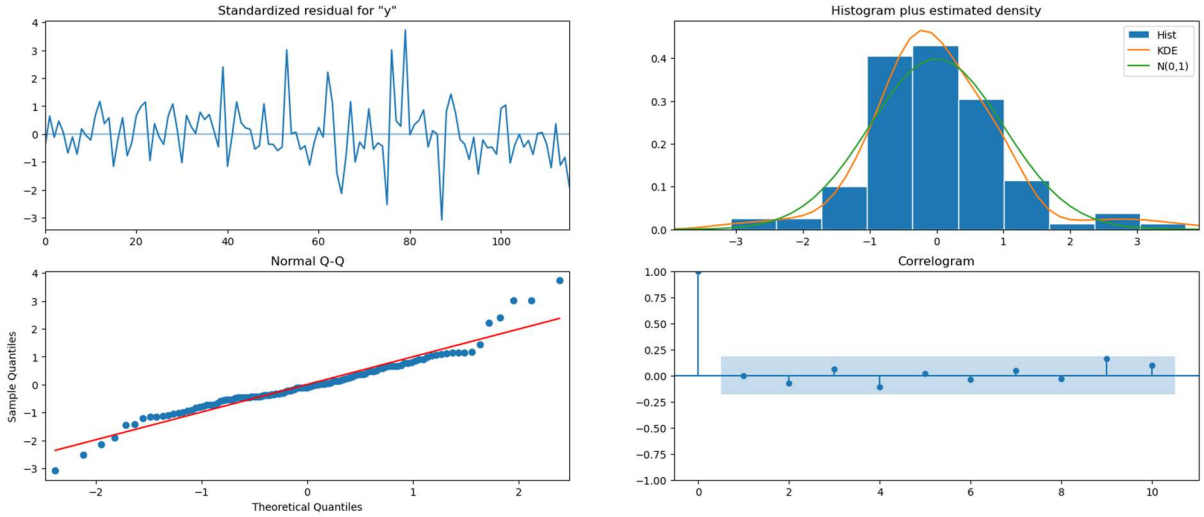


Fig 33 SARIMA diagnostics plot for seasonality as 6

Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1375.637068	384.082746	622.848718	2128.425418
1	1116.695584	392.847249	346.729125	1886.662044
2	1667.581669	395.420856	892.571033	2442.592306
3	1528.342301	397.980917	748.314036	2308.370565
4	1372.244943	400.524670	587.231015	2157.258871

Table 35 SARIMA 6 summary frame

	RMSE
ARIMA(2,1,2)	1299.979665
ARIMA(0,1,0)	3864.279352
SARIMA(0,1,2)(2,0,2,6)	601.293594

Table 36 Test RMSE values of ARIMA to SARIMA

Setting the seasonality as 12 for the second iteration of the auto SARIMA model.

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934563
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121564
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

Table 37 top 5 SARIMA12_AIC sort rows

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-769.967
Date:	Sun, 10 Dec 2023	AIC	1555.935
Time:	07:29:08	BIC	1577.090
Sample:	0 - 132	HQIC	1564.505
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6380	0.287	-2.225	0.026	-1.200	-0.076
ma.L1	-0.3050	0.185	-1.645	0.100	-0.668	0.058
ma.L2	-0.8914	0.275	-3.245	0.001	-1.430	-0.353
ar.S.L12	0.7611	0.567	1.342	0.179	-0.350	1.872
ar.S.L24	0.2952	0.590	0.500	0.617	-0.861	1.451
ma.S.L12	1.8829	3.334	0.565	0.572	-4.651	8.417
ma.S.L24	-1.8029	2.472	-0.729	0.466	-6.649	3.043
sigma2	1.858e+04	4.87e+04	0.382	0.703	-7.68e+04	1.14e+05

Ljung-Box (L1) (Q):	0.08	Jarque-Bera (JB):	12.54
Prob(Q):	0.78	Prob(JB):	0.00
Heteroskedasticity (H):	1.55	Skew:	0.35
Prob(H) (two-sided):	0.20	Kurtosis:	4.55

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table 38 results_auto_SARIMA12.summary

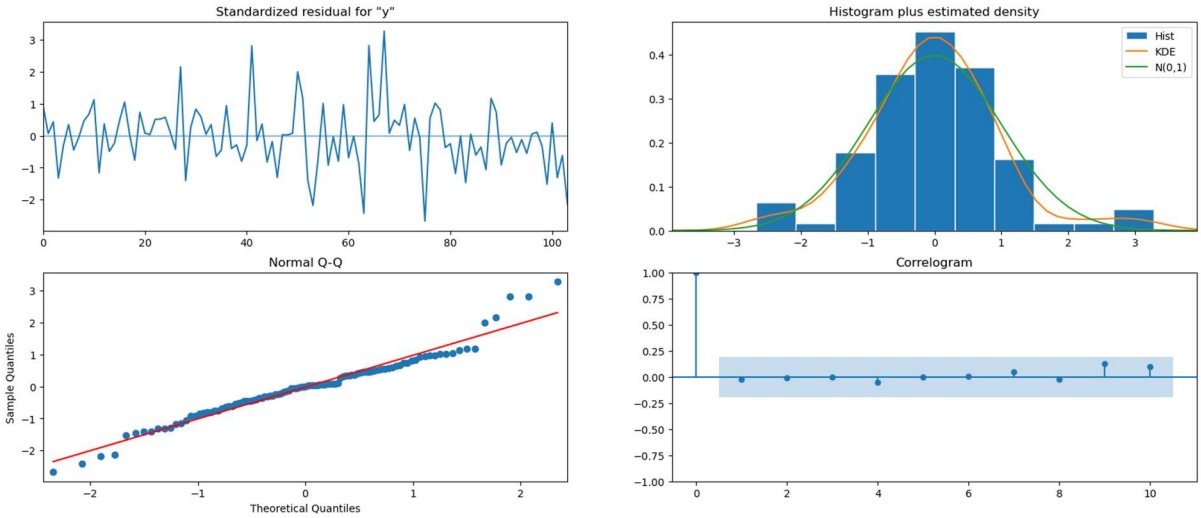


Fig 34 SARIMA diagnostics plot for seasonality as 12

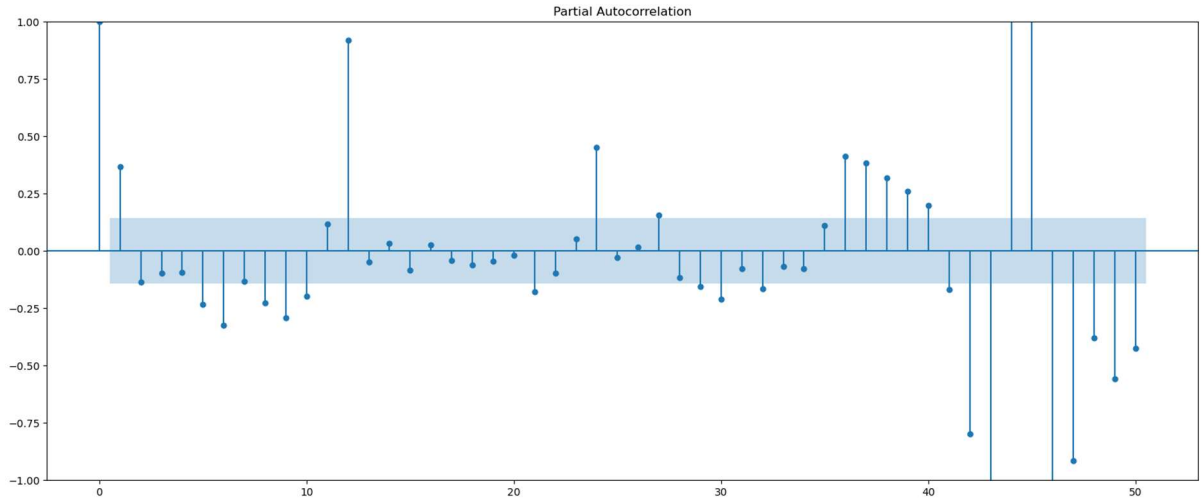
Predict on the Test Set using this model and evaluate the model.

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.035487
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525
Auto_ARIMA	1299.979665
(1,1,1),(2,0,3,12),Auto_SARIMA	546.541774

Table 39 Test RMSE values of Regression to Auto_SARIMA

7. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Model 11 : Manual ARIMA



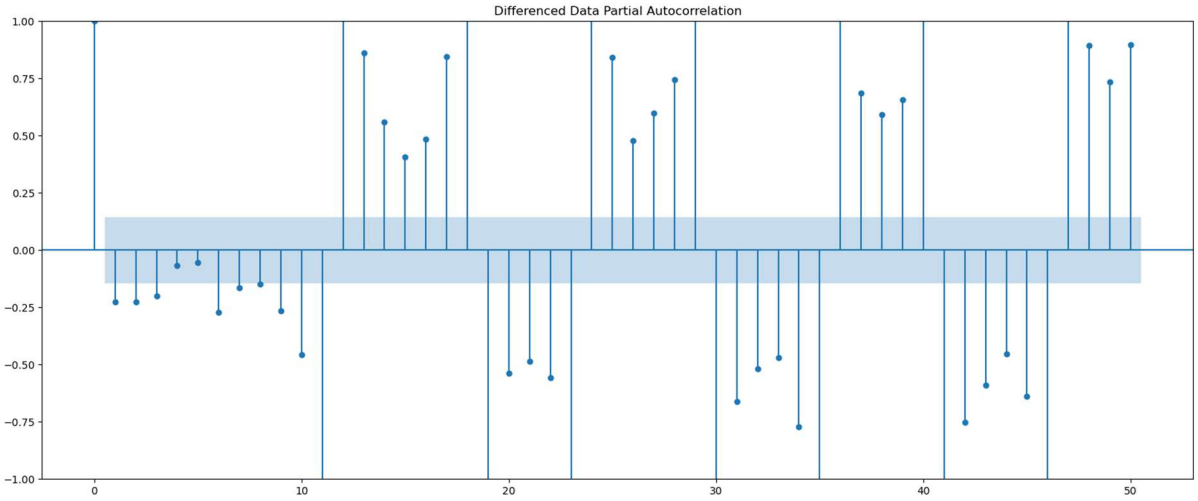


Fig 35 Sale Differenced Data Partial Autocorrelation

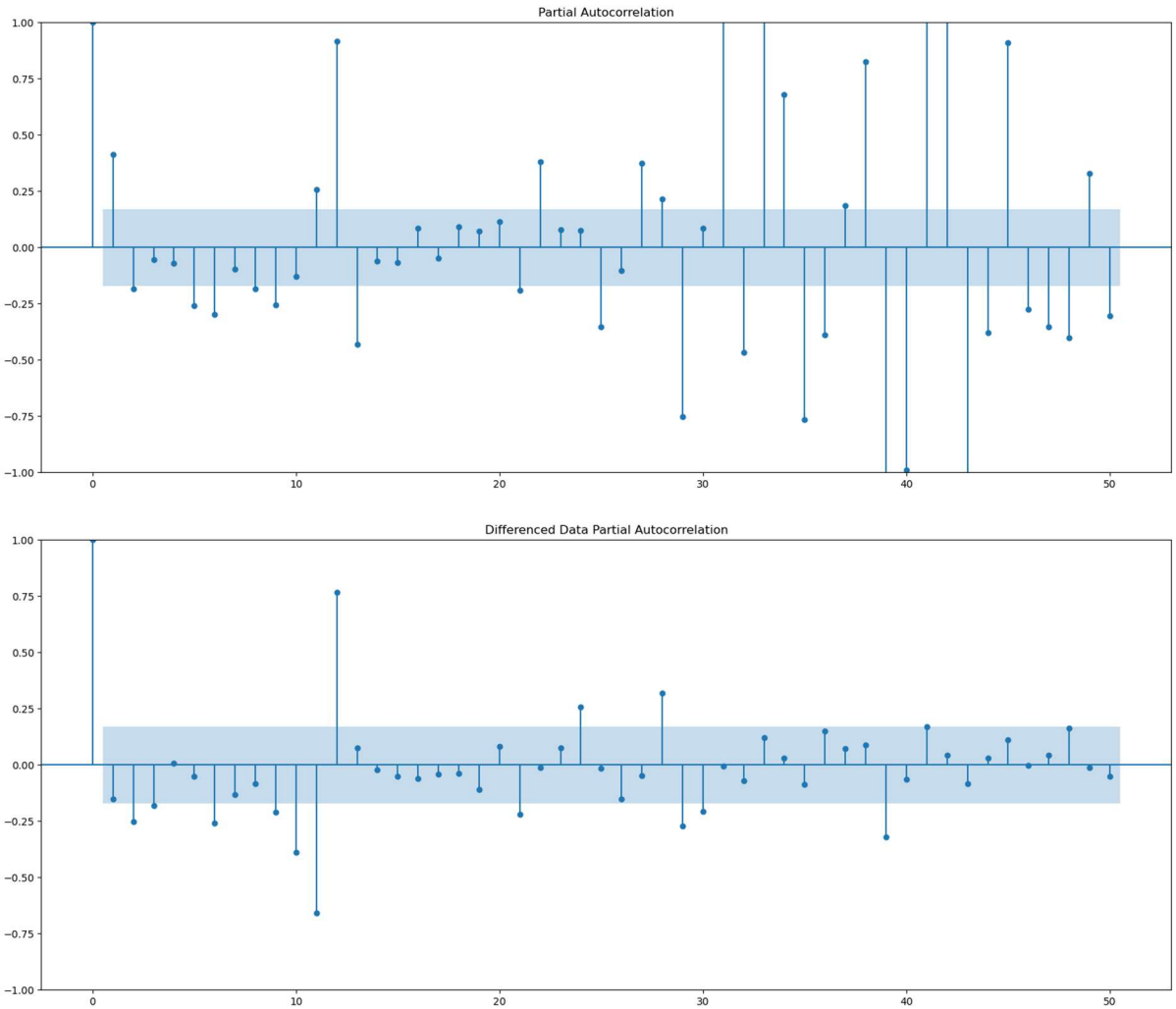


Fig 36 Sale Differenced Data Partial Autocorrelation after drop

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1114.878
Date:	Sun, 10 Dec 2023	AIC	2235.755
Time:	07:57:15	BIC	2244.381
Sample:	01-01-1980 - 12-01-1990	HQIC	2239.260
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4494	0.043	10.366	0.000	0.364	0.534
ma.L1	-0.9996	0.102	-9.811	0.000	-1.199	-0.800
sigma2	1.401e+06	7.57e-08	1.85e+13	0.000	1.4e+06	1.4e+06

Ljung-Box (L1) (Q):	0.50	Jarque-Bera (JB):	10.42
Prob(Q):	0.48	Prob(JB):	0.01
Heteroskedasticity (H):	2.64	Skew:	0.46
Prob(H) (two-sided):	0.00	Kurtosis:	4.03

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.66e+28. Standard errors may be unstable.

Table 40 results_auto_ Manual ARIMA.summary

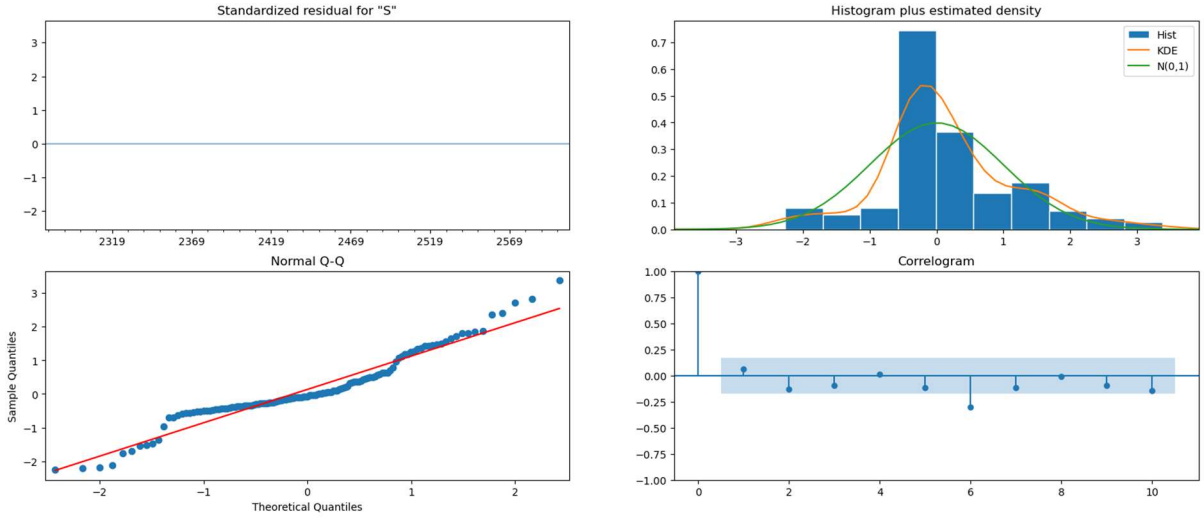


Fig 37 Manual ARIMA diagnostics plot

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.035487
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525
Auto_ARIMA	1299.979665
(1,1,1),(2,0,3,12),Auto_SARIMA	546.541774
ARIMA(3,1,3)	1319.936735

Table 41 Test RMSE values of Regression to ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-882.088
Date:	Sun, 10 Dec 2023	AIC	1774.175
Time:	07:59:55	BIC	1788.071
Sample:	0	HQIC	1779.818
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1957	0.104	1.878	0.060	-0.009	0.400
ma.L1	-0.9404	0.053	-17.897	0.000	-1.043	-0.837
ar.S.L12	0.0711	0.242	0.294	0.769	-0.404	0.546
ma.S.L12	-0.5035	0.221	-2.277	0.023	-0.937	-0.070
sigma2	1.51e+05	1.33e+04	11.371	0.000	1.25e+05	1.77e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	45.66
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	2.61	Skew:	0.82
Prob(H) (two-sided):	0.00	Kurtosis:	5.56

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table 42 results_auto_ Manual SARIMA.summary

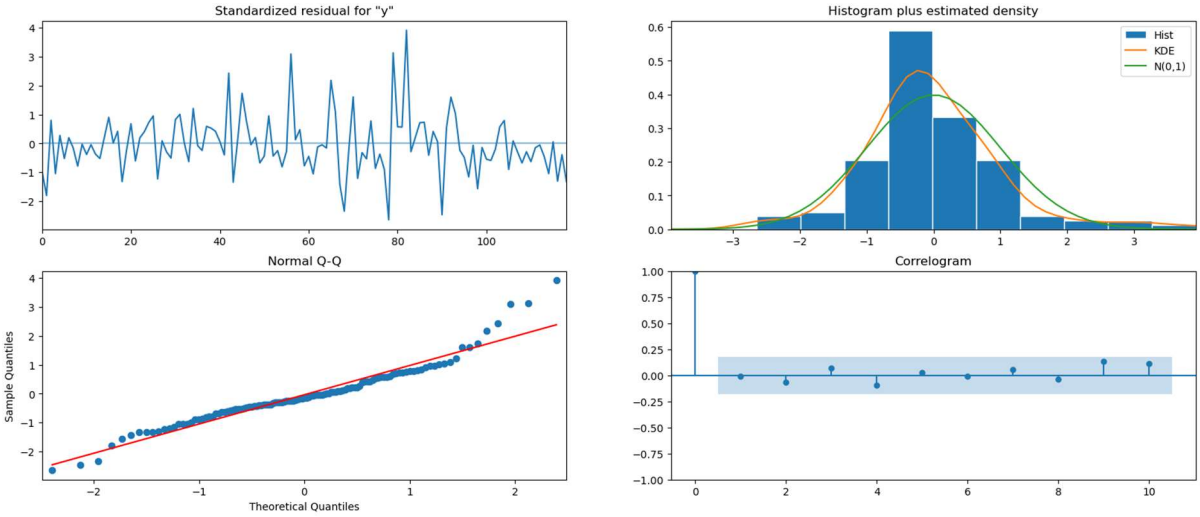


Fig 38 Manual SARIMA diagnostics plot

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.035487
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525
Auto_ARIMA	1299.979665
(1,1,1),(2,0,3,12),Auto_SARIMA	546.541774
ARIMA(3,1,3)	1319.936735
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612447

Table 43 Test RMSE values of Regression to Manual SARIMA

8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612447
(1,1,1),(2,0,3,12),Auto_SARIMA	546.541774
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
RegressionOnTime	1275.867052
6pointTrailingMovingAverage	1283.927428
Auto_ARIMA	1299.979665
Alpha=0.995,SimpleExponentialSmoothing	1316.035487
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.035487
ARIMA(3,1,3)	1319.936735
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
NaiveModel	3864.279352

Table 44 Test RMSE values of all models in sorted order

9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Based on the above comparison of all the various models that we had built, we can conclude that the triple exponential smoothing or the Holts-Winter model is giving us the lowest RMSE, hence it would be the most optimum model.

	Sales_Predictions
1995-08-01	1988.782193
1995-09-01	2652.762887
1995-10-01	3483.872246
1995-11-01	4354.989747
1995-12-01	6900.103171
1996-01-01	1546.800546
1996-02-01	1981.361768
1996-03-01	2245.459724
1996-04-01	2151.066942
1996-05-01	1929.355815
1996-06-01	1830.619260
1996-07-01	2272.156151

Table 45 future_predictions rows

	lower_CI	prediction	upper_ci
1995-08-01	1213.490105	1988.782193	2764.074282
1995-09-01	1877.470798	2652.762887	3428.054975
1995-10-01	2708.580157	3483.872246	4259.164335
1995-11-01	3579.697659	4354.989747	5130.281836
1995-12-01	6124.811083	6900.103171	7675.395260

Table 46 future_predictions with lower_ci and upper_ci

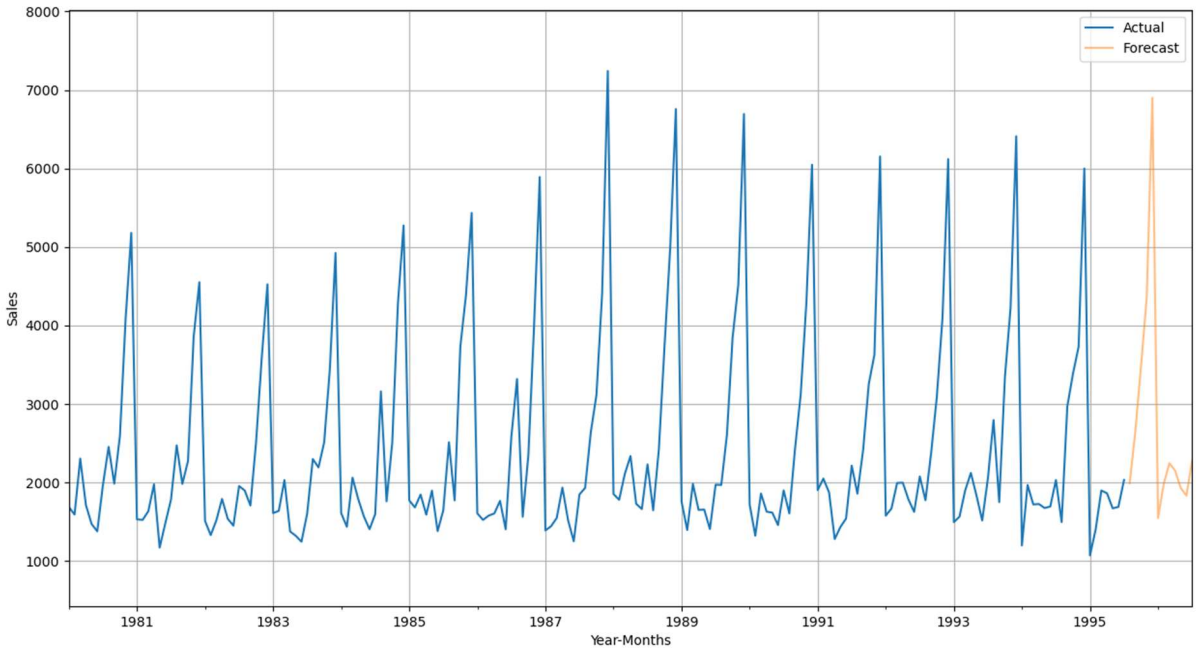


Fig 39 actual and forecast along with the confidence band

1 year into the future are shown in orange colour, while the confidence interval has been shown in grey colour.

Insights and recommendations:

- People tend to buy more sparkling wine in December. That's a great time to make sure we have enough stock and to advertise more.

- Some years, October and November sales stay the same. But after certain years, things change. It's important to understand why to prepare better for those changes.
- When we hold events or tastings in our store, more people come in and buy. So, doing more events could help sell more.
- Some years had big changes after a certain time. Keeping an eye on these years might help us spot things we can do to keep sales up.
- Sometimes, the easiest way to guess sales is the best. We should use simple methods that still give us good guesses.
- By using both old information and our guesses about the future, we can change our plans quickly. This means we're ready for whatever happens and can do better business.
- The sales data exhibits clear seasonal patterns with a notable increase in December sales across all years. Additionally, while there's an overall upward trend until the peak years of 1988–1989, there's a subsequent decline in sales in the years that follow. The choice of the multiplicative model is supported by the observed data patterns, showcasing both seasonality and trend effects.