Project Report - Time Series Forecasting

Rose Wine

DECEMBER 10, 2023

HRITIKA VAISHNAV

INDEX

LIST OF TABLES

LIST OF FIGURES

Problem 2: Rose

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 rose 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, Rose.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 float64(1) are present and null values are there.

	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

	Rose
YearMonth	
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

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   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Table 2: Information about the dataset structure and content.

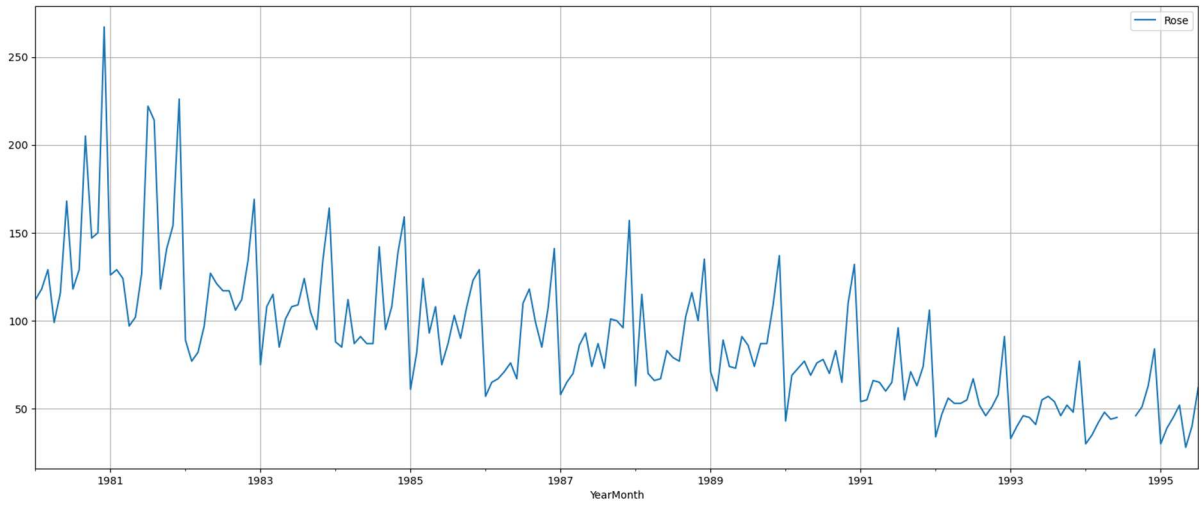


Fig 1 Time series plot

Plot the Time Series to understand the behaviour of the data, to understand the behaviour of the data, We can see that there is a downward trend a seasonal pattern associated as well.

Insights:

- The dataset contains a total of 187 records.
- It consists of 1 column, which is float data types.
- The dataset have null values.
- We can see that there is a downward trend a seasonal pattern associated as well.

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

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

Table 3 Descriptive statistics

The basic measures of descriptive statistics tell us how the Sales have varied across years. But remember, for this measure of descriptive statistics we have averaged over the whole data without taking the time component into account. So, Here The average sales of Rose Wine per month are around 90.The maximum sale of the Wine is approx 267.The minimum sale of the Wine is approx 28.

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	Rose	Year	Month
1980-01-01	112.0	1980	1
1980-02-01	118.0	1980	2
1980-03-01	129.0	1980	3
1980-04-01	99.0	1980	4
1980-05-01	116.0	1980	5

Table 4 Separate columns for the year and month

YearMonth	Sales	Year	Month
1980-01-01	112.0	1980	1
1980-02-01	118.0	1980	2
1980-03-01	129.0	1980	3
1980-04-01	99.0	1980	4
1980-05-01	116.0	1980	5

YearMonth	Sales	Year	Month
1995-03-01	45.0	1995	3
1995-04-01	52.0	1995	4
1995-05-01	28.0	1995	5
1995-06-01	40.0	1995	6
1995-07-01	62.0	1995	7

Table 5 Top and bottom 5 values of new dataset

Null Values, there are null values present in dataset. Sales column has 2 missing values. Missing value treatment we use instead of taking means for the 7th months across all the years, we just took mean of the 7th months values from a year before and a year after the missing value. Similar steps were taken for 8th month.

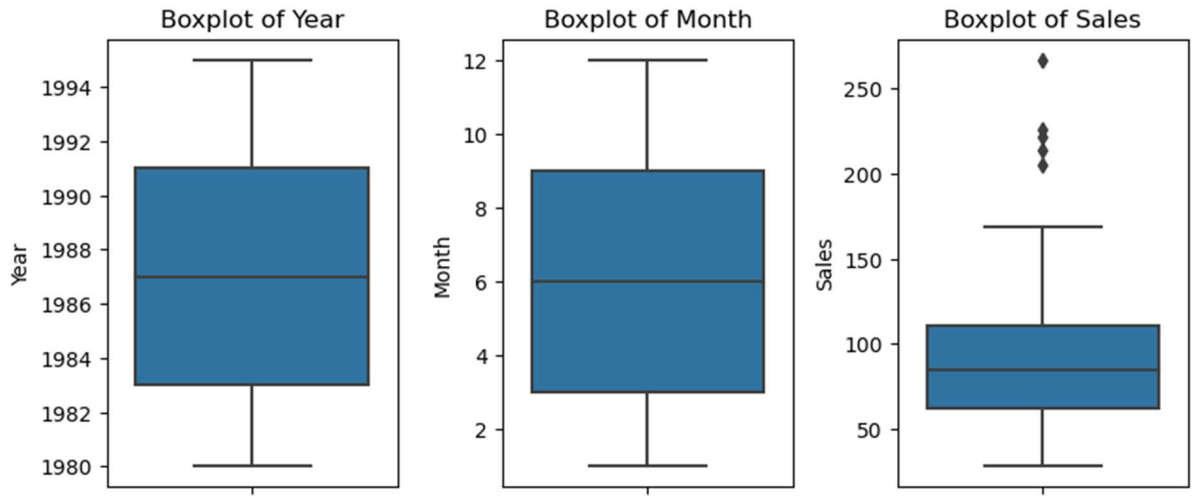


Fig 2 Boxplot of feature list of dataset

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.

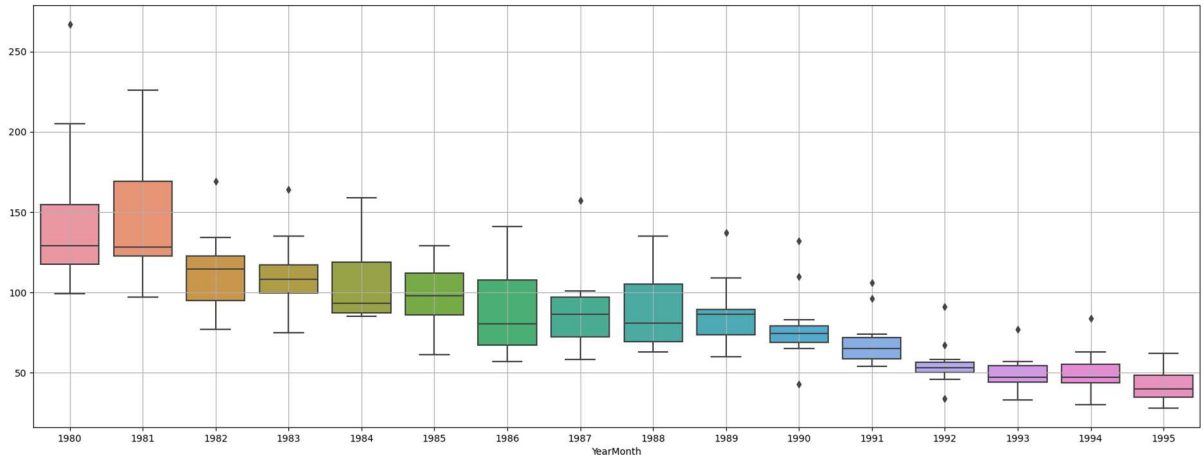


Fig 3 Yearly Boxplot of feature list of dataset

There is consistency over the years and there was a peak in 1980-1981. Outliers are present in almost all years.

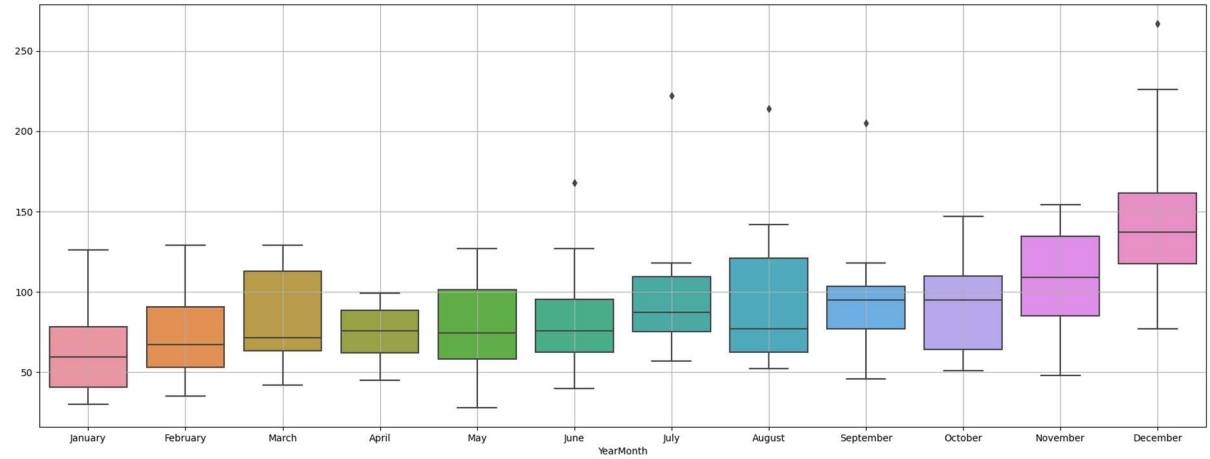


Fig 4 Monthly Boxplot of feature list of dataset

Sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to July then from august the sales start to increase. Outliers are present in June, July, august, September and December.

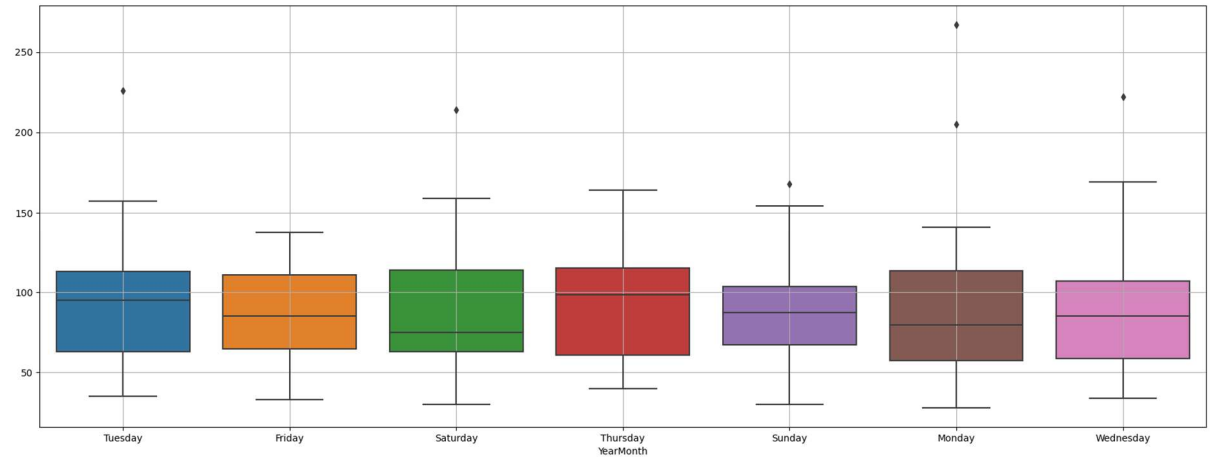


Fig 5 Weekly Boxplot of feature list of dataset

Tuesday has more sales than other days and Wednesday has the lowest sales of the week. Outliers are present on all days except Friday and Thursday.

YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	99.0	129.0	267.0	118.0	112.0	118.0	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.0	226.0	129.0	126.0	222.0	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.0	169.0	77.0	89.0	117.0	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.0	164.0	108.0	75.0	109.0	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.0	159.0	85.0	88.0	87.0	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.0	129.0	82.0	61.0	87.0	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.0	141.0	65.0	57.0	110.0	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.0	157.0	65.0	58.0	87.0	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.0	135.0	115.0	63.0	79.0	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.0	137.0	60.0	71.0	86.0	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.0	132.0	69.0	43.0	78.0	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.0	106.0	55.0	54.0	96.0	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.0	91.0	47.0	34.0	67.0	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.0	77.0	40.0	33.0	57.0	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	54.0	84.0	35.0	30.0	59.5	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.0	40.0	45.0	28.0	NaN	NaN	NaN

Table 6 pivot table of dataset

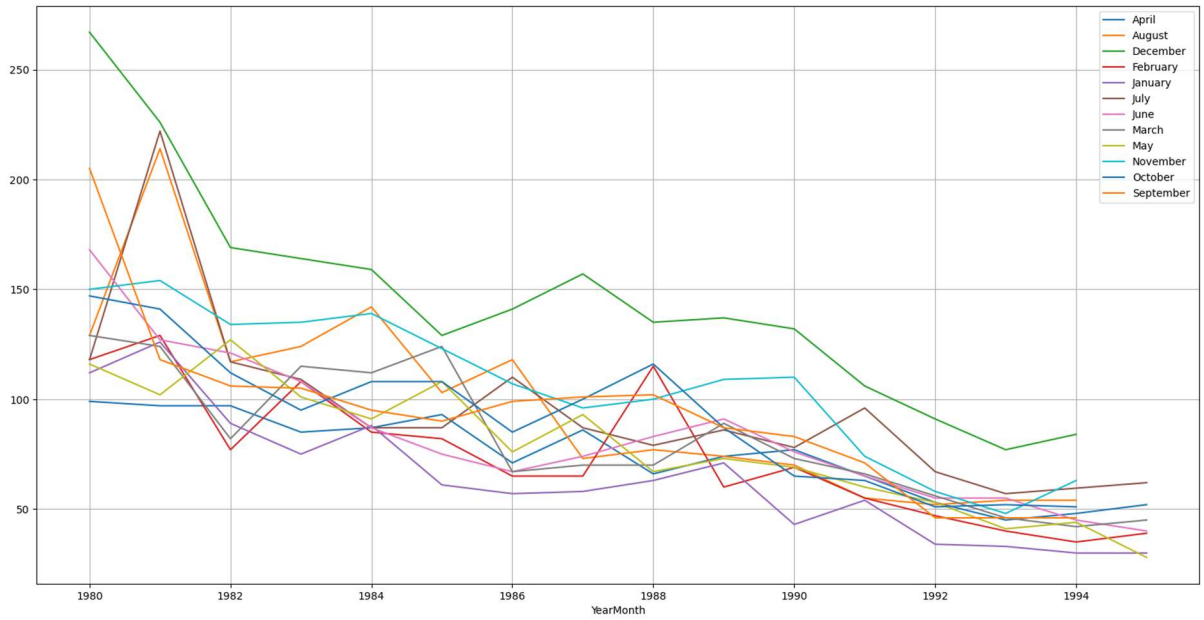


Fig 6 Weekly Boxplot of feature list of dataset

This plot shows that December has the highest sales over the years and the year 1981 was the year with the highest number of sales.

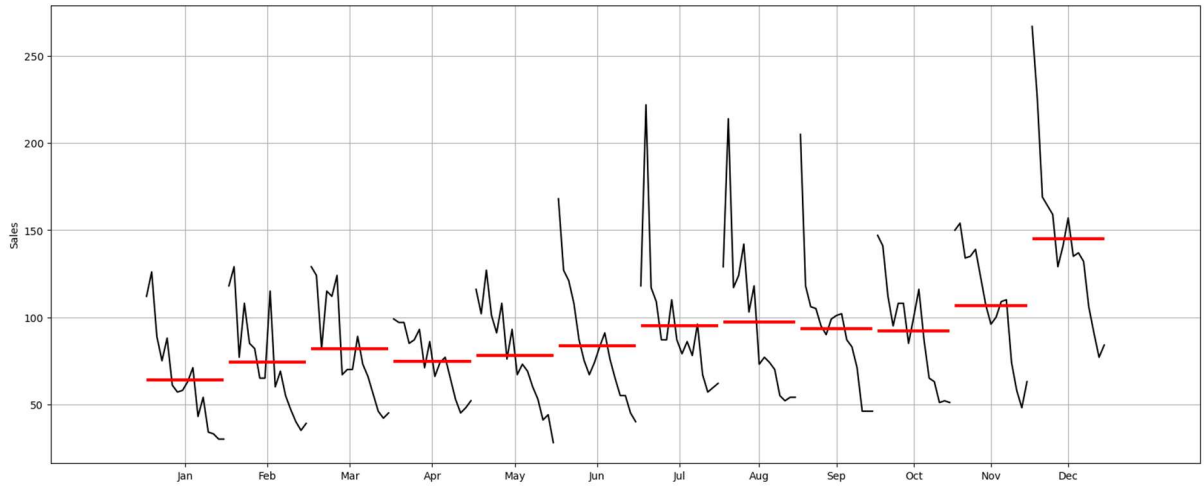


Fig 7 Math_plot of dataset

Sales are seen to increase and decrease across various months, month of June sales start increasing and in December, sales are highly increasing.

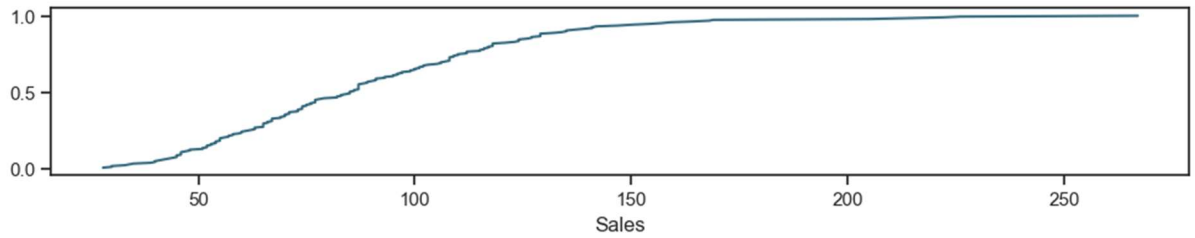


Fig 8 ECDF plot of dataset

50% of sales are below 100 sales, approx. 90% of sales are less than 150 sales, and 100% of sales are 250 sales.

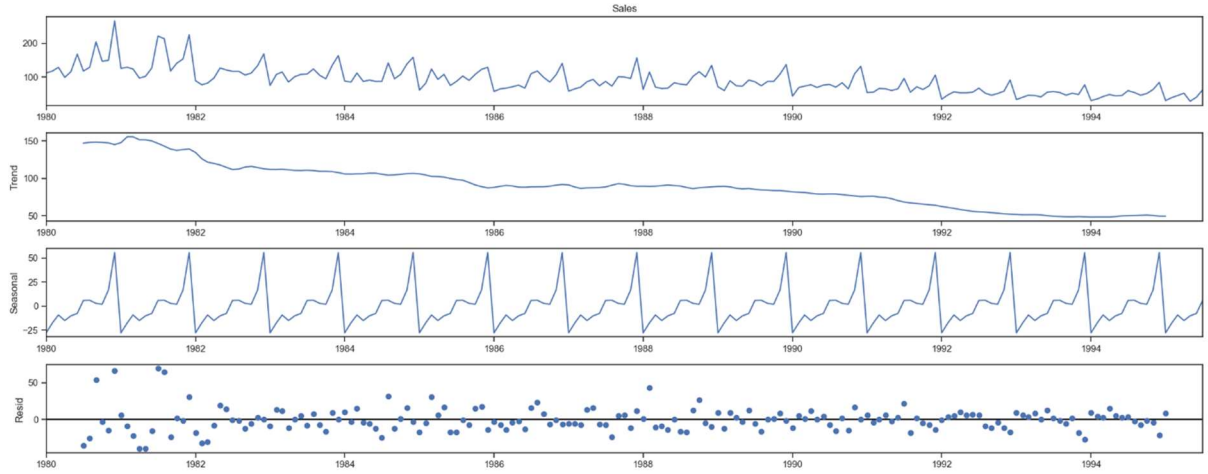


Fig 9 Additive Decomposition of dataset

Years of peak: 1981, it demonstrates that the trend has weakened from 1981. Instead of being in a straight line, residue is dispersed. Seasonality and trends are both present.

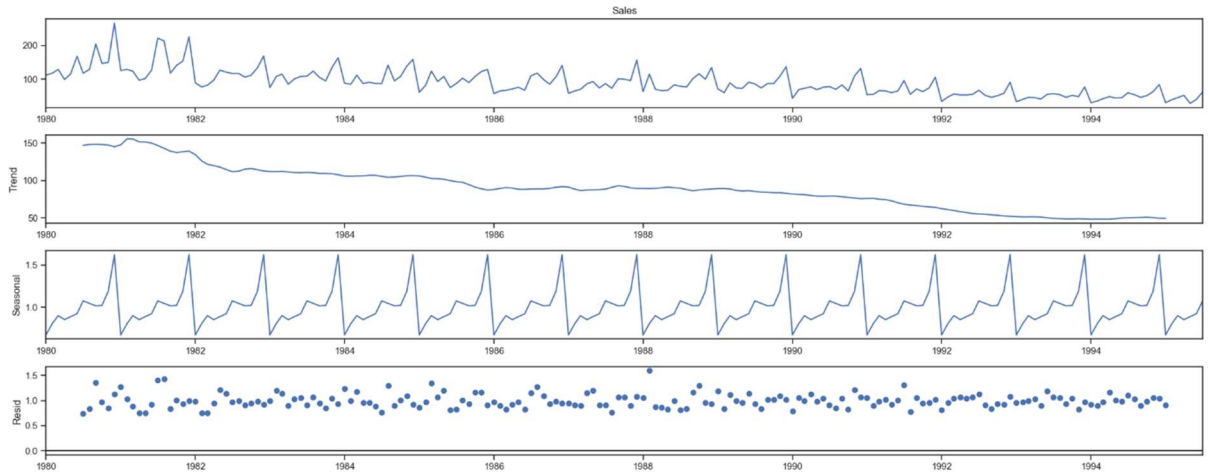


Fig 10 Multiplicative Decomposition of dataset

Peak years is 1981. It also demonstrates the trend's downward movement in the years of 1981. The residue is dispersed and roughly follows a straight path. There is seasonality as well as a trend. Additive is 0 to 50, 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 showcases consistent patterns over the years, with December consistently having the highest sales and January reporting the lowest. The identified peak year is 1981, followed by a weakening trend, indicated by dispersed residuals. The presence of seasonality and trend supports the adoption of a multiplicative model for analysis.

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.

First few rows of Training Data

YearMonth	Year	Month	Sales
1980-01-01	1980	1	112.0
1980-02-01	1980	2	118.0
1980-03-01	1980	3	129.0
1980-04-01	1980	4	99.0
1980-05-01	1980	5	116.0

Last few rows of Training Data

YearMonth	Year	Month	Sales
1990-08-01	1990	8	70.0
1990-09-01	1990	9	83.0
1990-10-01	1990	10	65.0
1990-11-01	1990	11	110.0
1990-12-01	1990	12	132.0

Table 7 Top and bottom rows of train dataset

First few rows of Test Data

YearMonth	Year	Month	Sales
1991-01-01	1991	1	54.0
1991-02-01	1991	2	55.0
1991-03-01	1991	3	66.0
1991-04-01	1991	4	65.0
1991-05-01	1991	5	60.0

Last few rows of Test Data

YearMonth	Year	Month	Sales
1995-03-01	1995	3	45.0
1995-04-01	1995	4	52.0
1995-05-01	1995	5	28.0
1995-06-01	1995	6	40.0
1995-07-01	1995	7	62.0

Table 8 Top and bottom rows of test dataset

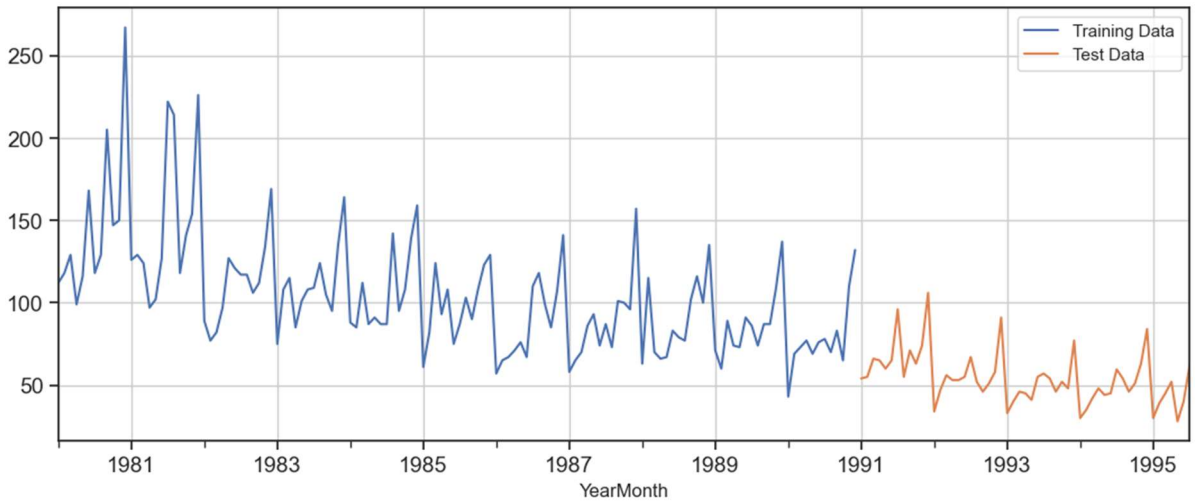


Fig 12 Train and test dataset upward trend with seasonality

Data split from 1980-1990 is training data, then 1991 to 1995 is training data. Blue curve is Training data and orange curve is test data.

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, 34, 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, 7
1, 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, 10
4, 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, 74, 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

First few rows of Training Data

YearMonth	Year	Month	Sales	time
1980-01-01	1980	1	112.0	1
1980-02-01	1980	2	118.0	2
1980-03-01	1980	3	129.0	3
1980-04-01	1980	4	99.0	4
1980-05-01	1980	5	116.0	5

Last few rows of Training Data

YearMonth	Year	Month	Sales	time
1990-08-01	1990	8	70.0	128
1990-09-01	1990	9	83.0	129
1990-10-01	1990	10	65.0	130
1990-11-01	1990	11	110.0	131
1990-12-01	1990	12	132.0	132

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

First few rows of Test Data

YearMonth	Year	Month	Sales	time	RegOnTime
1991-01-01	1991	1	54.0	43	116.557274
1991-02-01	1991	2	55.0	44	116.062896
1991-03-01	1991	3	66.0	45	115.568518
1991-04-01	1991	4	65.0	46	115.074140
1991-05-01	1991	5	60.0	47	114.579762

Last few rows of Test Data

YearMonth	Year	Month	Sales	time	RegOnTime
1995-03-01	1995	3	45.0	93	91.838381
1995-04-01	1995	4	52.0	94	91.344003
1995-05-01	1995	5	28.0	95	90.849625
1995-06-01	1995	6	40.0	96	90.355247
1995-07-01	1995	7	62.0	97	89.860869

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

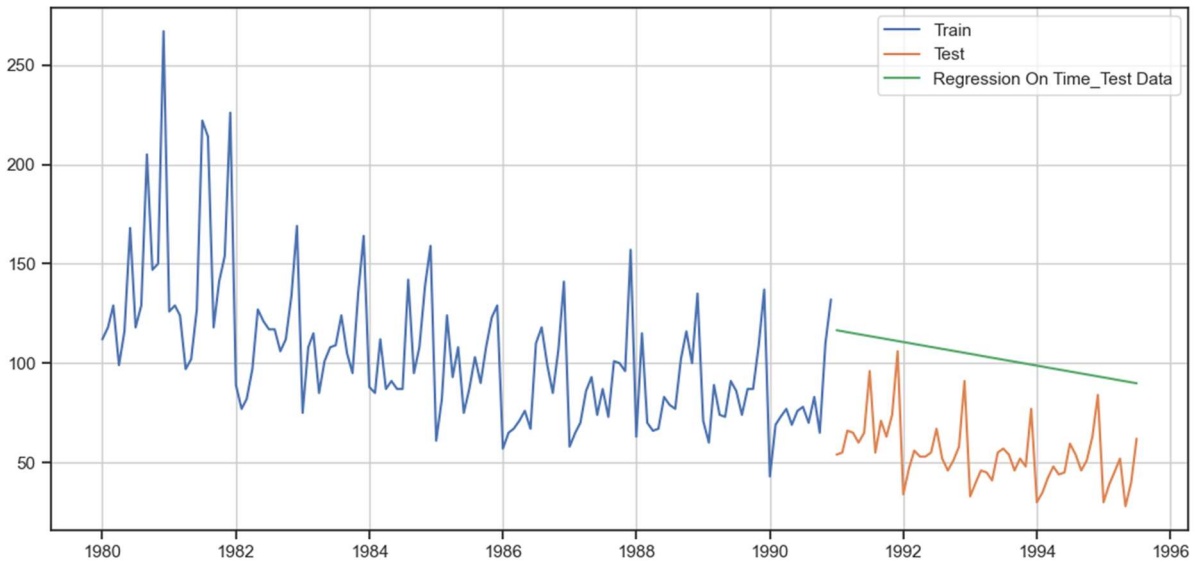


Fig 14 Train and test dataset behaviour with Linear Regression

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
Linear Regression	51.080941

Table 11 RMSE matrix value of Linear Regression

The value of Linear Regression with Test RMSE is 51.080941

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    132.0
1991-02-01    132.0
1991-03-01    132.0
1991-04-01    132.0
1991-05-01    132.0
Name: naive, dtype: float64
```

Table 12 top 5 data of Naive Approach train dataset

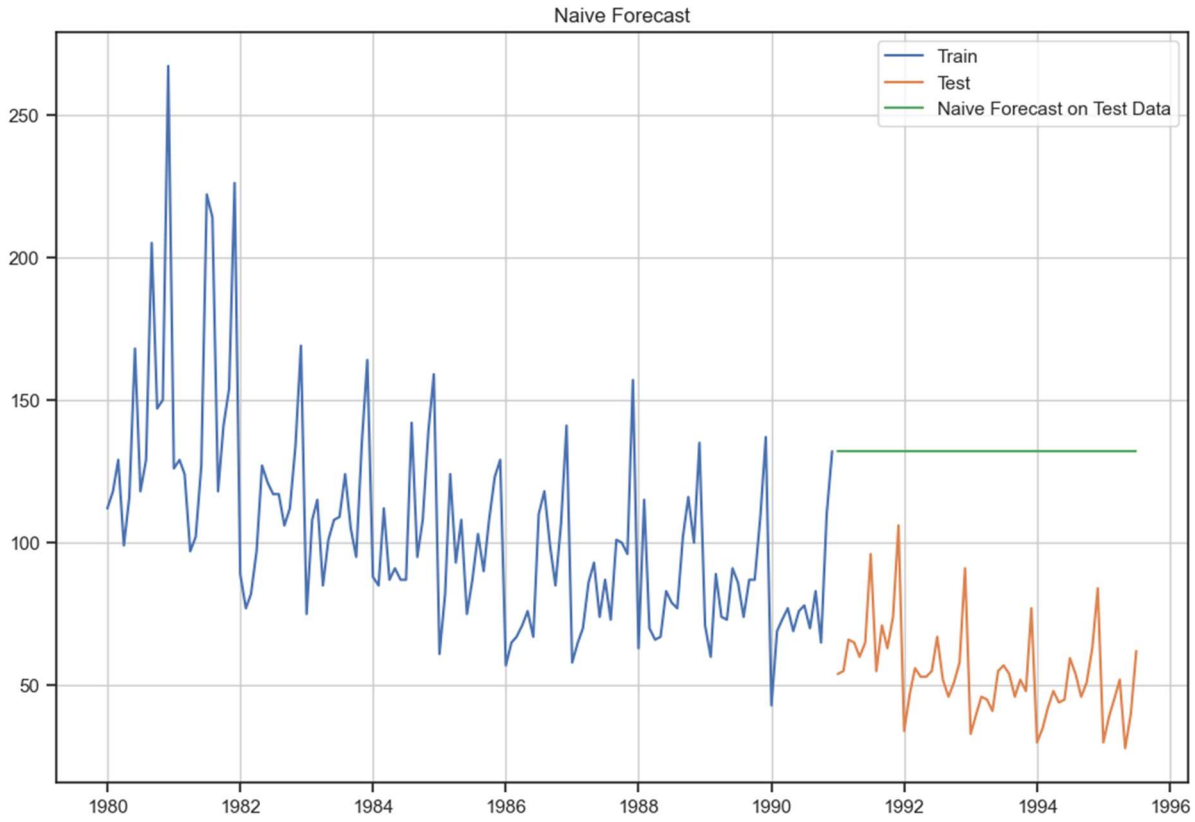


Fig 15 Train and test dataset behaviour with Linear Regression

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
Linear Regression	51.080941
Naive Model	79.304391

Table 13 RMSE matrix value of Naïve model

The value of Naïve model with Test RMSE is 79.304391

Method 3: Simple Average

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

YearMonth	Year	Month	Sales	mean_forecast
1991-01-01	1991	1	54.0	104.939394
1991-02-01	1991	2	55.0	104.939394
1991-03-01	1991	3	66.0	104.939394
1991-04-01	1991	4	65.0	104.939394
1991-05-01	1991	5	60.0	104.939394

Table 14 top 5 data of Simple Moving Average

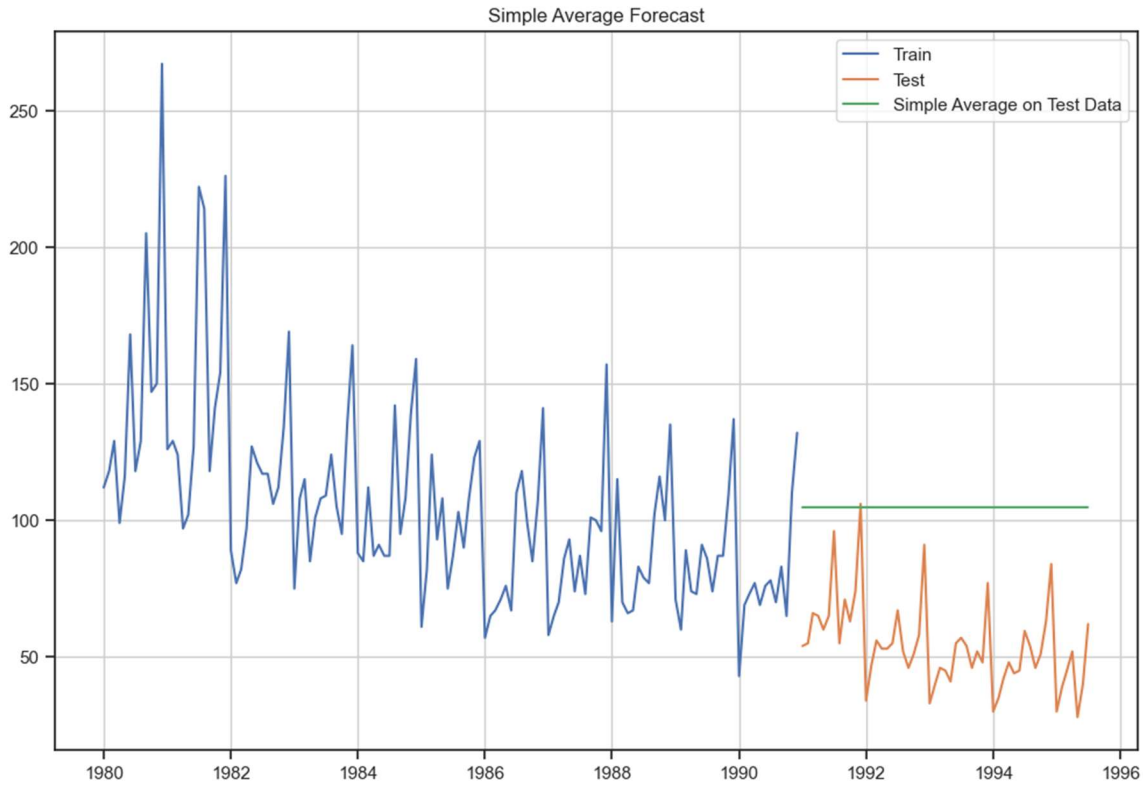


Fig 17 Simple Moving average models

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
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755

Table 15 RMSE matrix value of Simple Average Model

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	Year	Month	Sales	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1980	1	112.0	NaN	NaN	NaN	NaN
1980-02-01	1980	2	118.0	115.0	NaN	NaN	NaN
1980-03-01	1980	3	129.0	123.5	NaN	NaN	NaN
1980-04-01	1980	4	99.0	114.0	114.5	NaN	NaN
1980-05-01	1980	5	116.0	107.5	115.5	NaN	NaN

Table 16 top 5 data of Moving Average

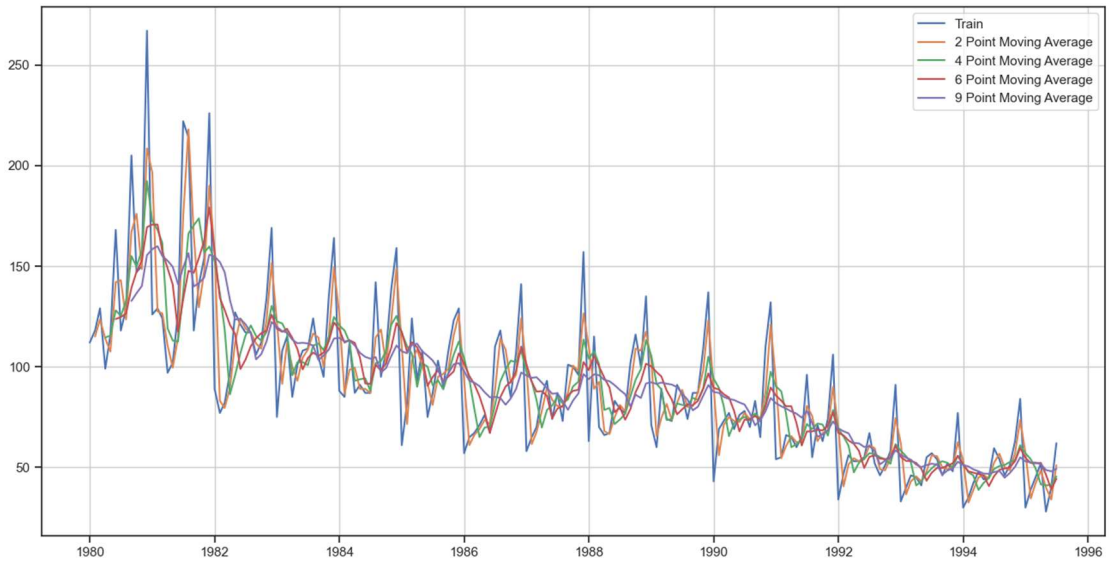


Fig 18 Moving average models with rolling windows for Train dataset

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

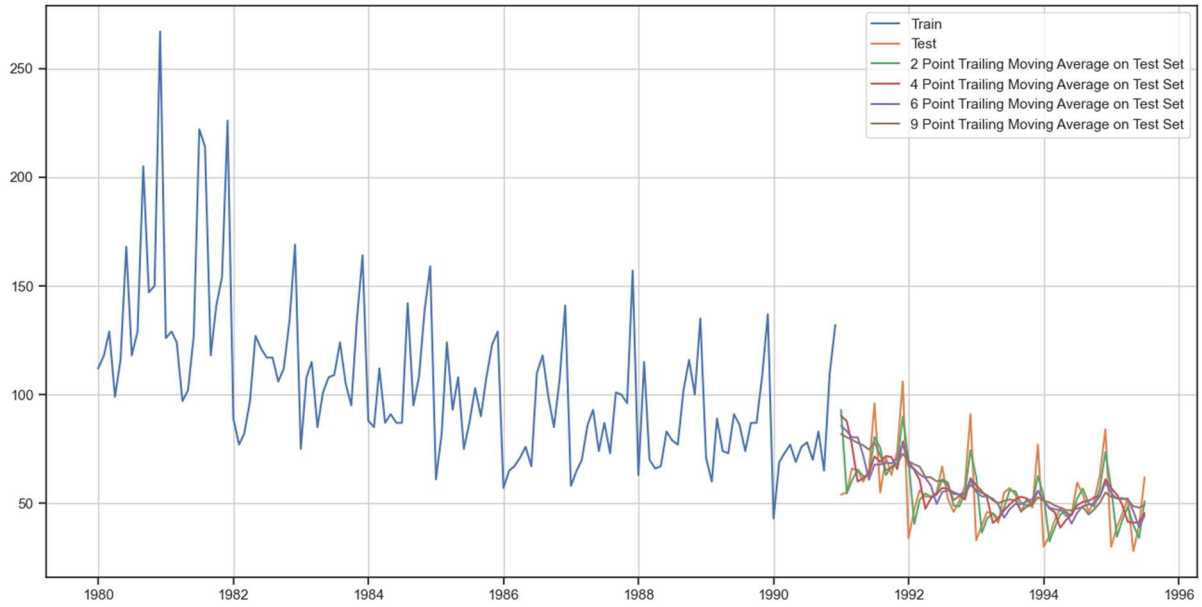


Fig 19 Moving average models with rolling windows for test dataset

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

	Test RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139

Table 17 RMSE matrix value of Trailing Moving Average 2-9 point

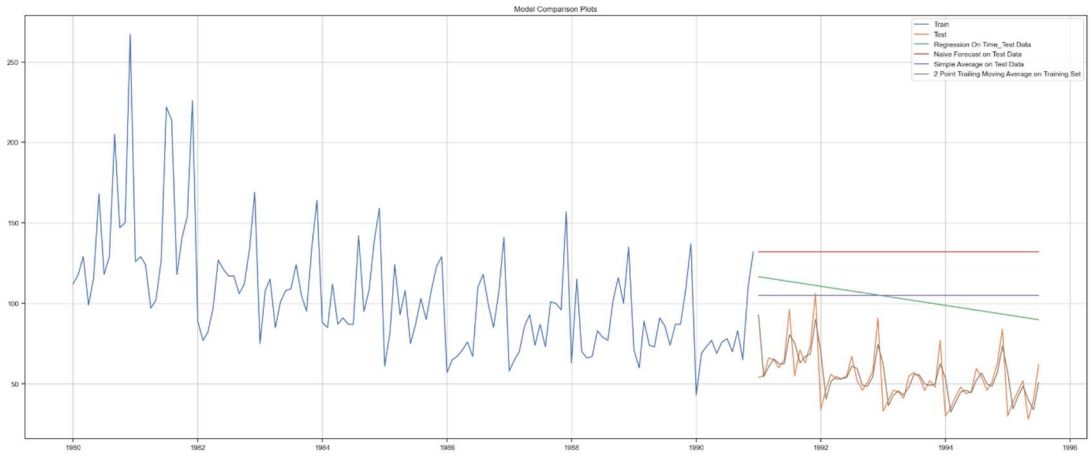


Fig 20 Model Comparison Plot

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	Year	Month	Sales	predict
1991-01-01	1991	1	54.0	87.104983
1991-02-01	1991	2	55.0	87.104983
1991-03-01	1991	3	66.0	87.104983
1991-04-01	1991	4	65.0	87.104983
1991-05-01	1991	5	60.0	87.104983

Table 18 top 5 data of Simple Exponential Smoothing

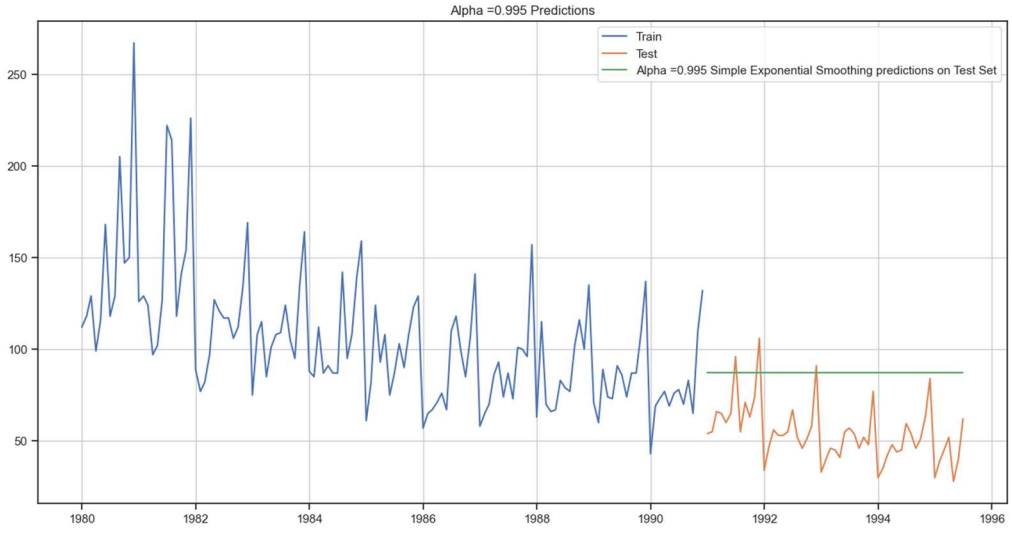


Fig 21 Simple Exponential Smoothing with alpha 0.995

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
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777

Table 19 RMSE matrix value alpha 0.995 Simple Exponential Smoothing

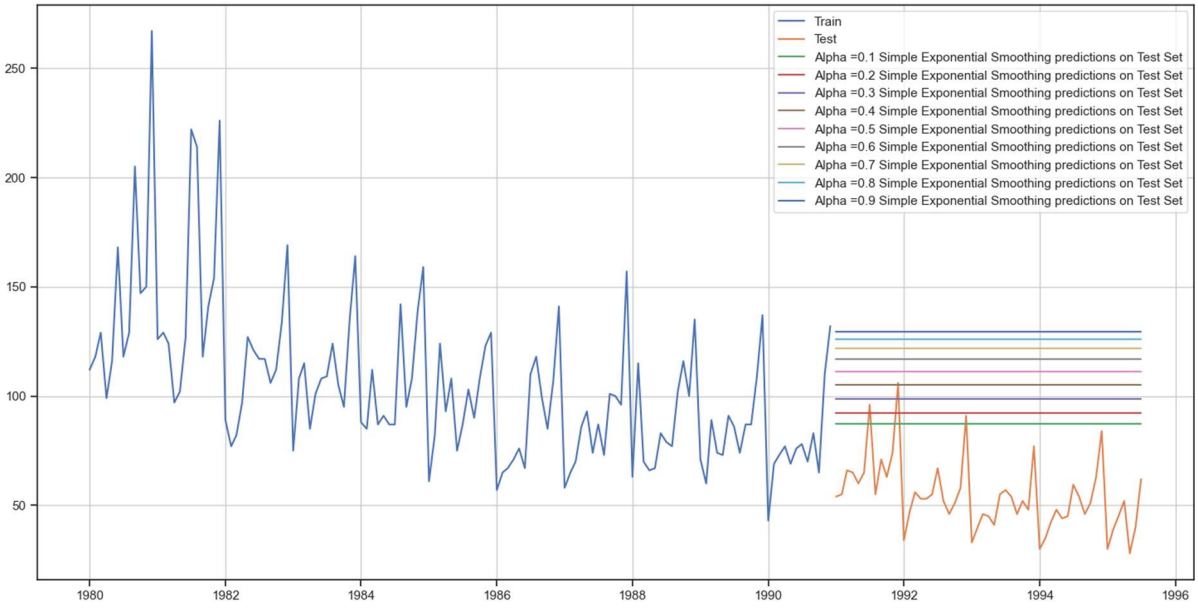


Fig 22 Simple Exponential Smoothing with alpha 0.1 to 0.995

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
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535

Table 20 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	Year	Month	Sales	predict
1991-01-01	1991	1	54.0	73.259732
1991-02-01	1991	2	55.0	72.767150
1991-03-01	1991	3	66.0	72.274569
1991-04-01	1991	4	65.0	71.781987
1991-05-01	1991	5	60.0	71.289405

Table 21 top 5 data of Double Exponential Smoothing

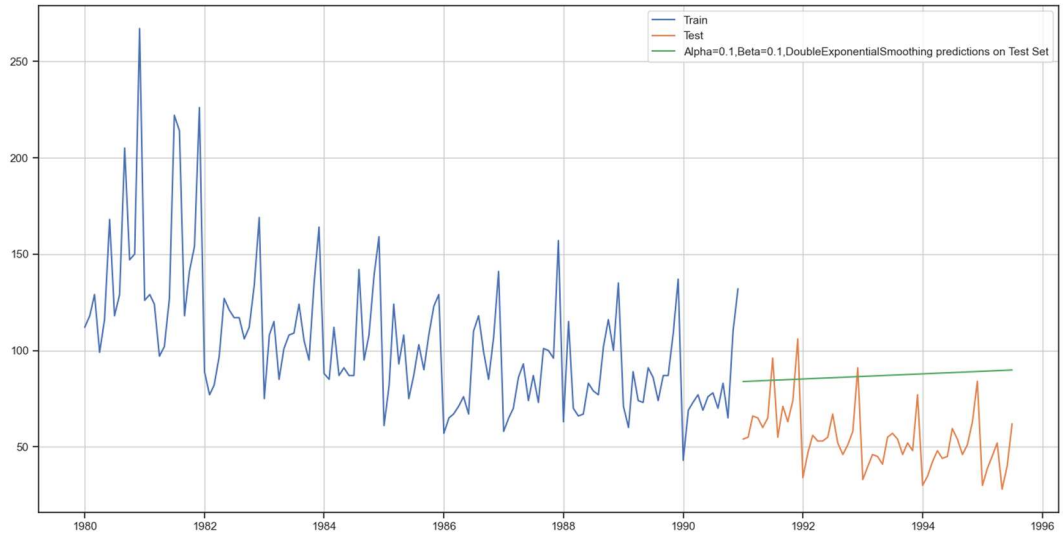


Fig 22 Double Exponential Smoothing

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
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010

Table 21 Test RMSE values of Regression to Double Exponential Smoothing

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

YearMonth	Year	Month	Sales	(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)	...	(predict_tm_s 0.9, 0.8, 0.
1991-01-01	1991	1	54.0	45.711834	46.537302	46.559436	46.071952	45.225493	44.012323	42.371394	...	3.735952e+
1991-02-01	1991	2	55.0	56.369270	60.659980	62.645947	63.356292	63.447907	63.255359	62.959944	...	7.148786e+
1991-03-01	1991	3	66.0	63.004762	65.794341	66.979401	67.649481	68.330584	69.025477	69.443834	...	1.357303e+
1991-04-01	1991	4	65.0	51.663022	58.369250	62.190538	64.397061	65.989684	67.481555	68.978770	...	2.576482e+
1991-05-01	1991	5	60.0	58.931424	61.246579	62.077338	62.127541	62.025325	62.111373	62.439476	...	-1.230813e+

5 rows × 3461 columns

Table 23 top 5 data of Triple Exponential Smoothing

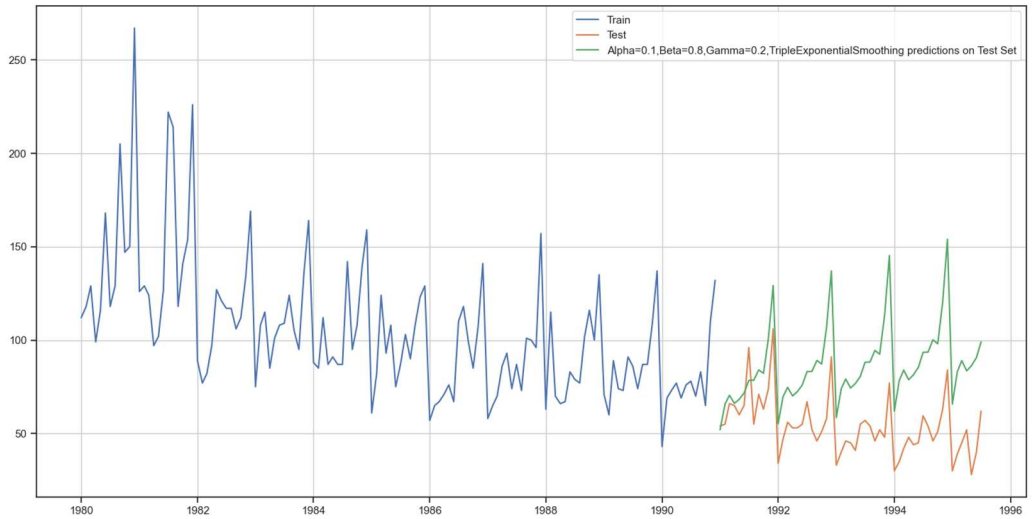


Fig 24 Triple Exponential Smoothing

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
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Alpha=0.1,Beta=0.8,Gamma=0.2,TripleExponentialSmoothing	8.992350

Table 24 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.

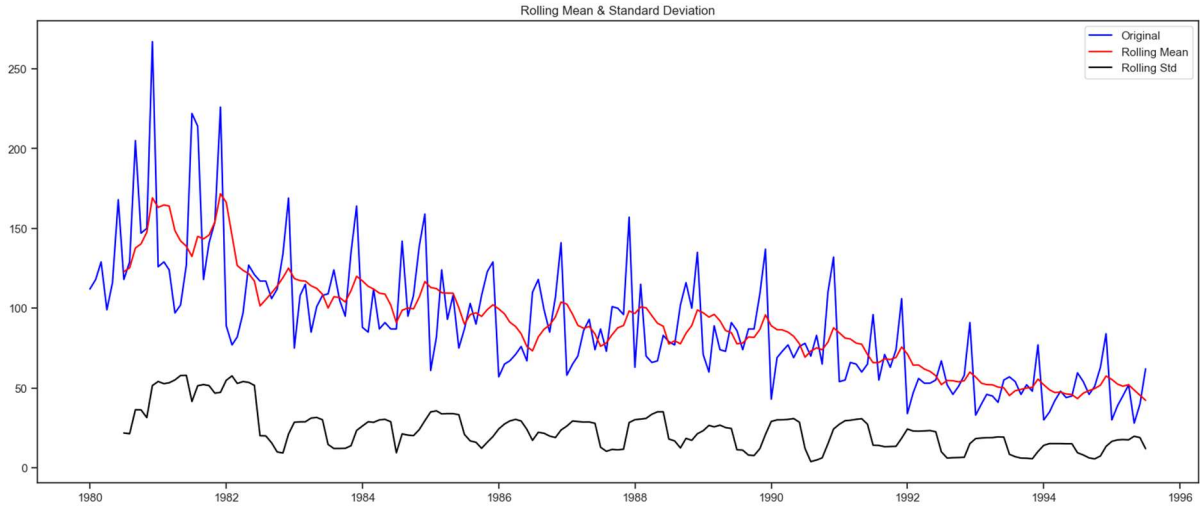


Fig 25 Rolling Mean & Standard Deviation

Results of Dickey-Fuller Test:	
Test Statistic	-1.892338
p-value	0.335674
#Lags Used	13.000000
Number of Observations Used	173.000000
Critical Value (1%)	-3.468726
Critical Value (5%)	-2.878396
Critical Value (10%)	-2.575756
dtype: float64	

Table 25 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.

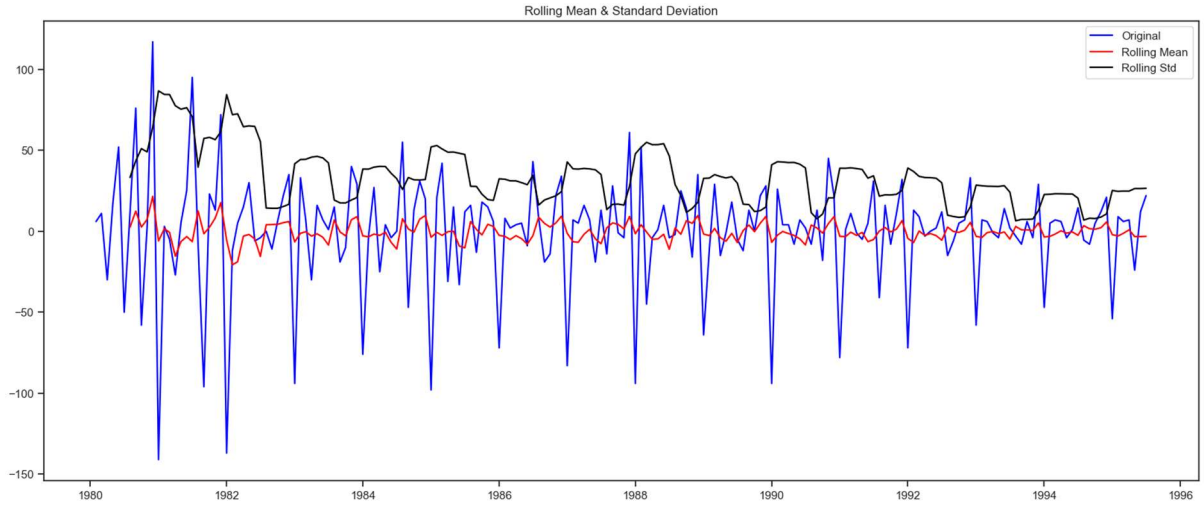


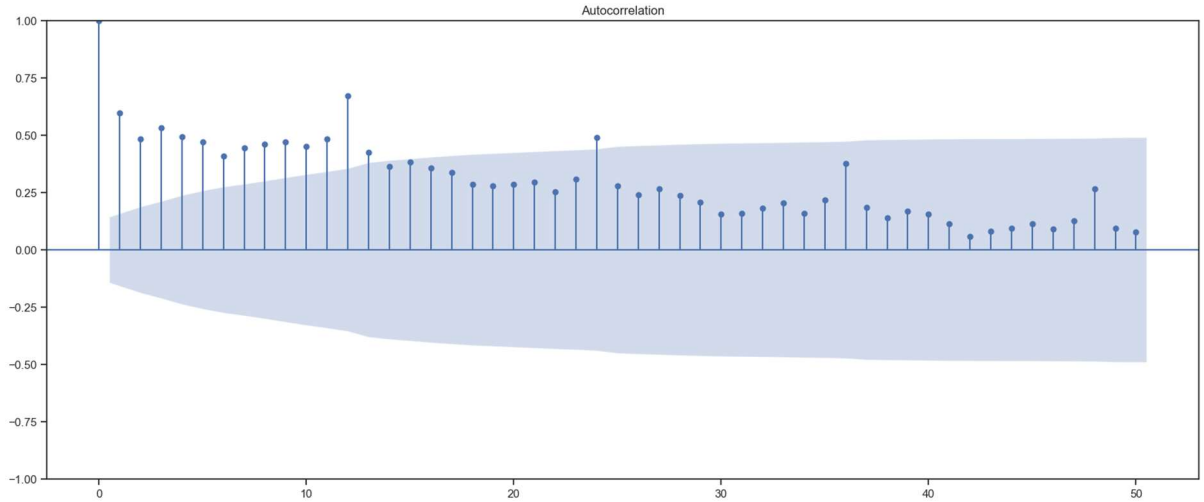
Fig 25 Rolling Mean & Standard Deviation after differencing

Results of Dickey-Fuller Test:	
Test Statistic	-8.032729e+00
p-value	1.938803e-12
#Lags Used	1.200000e+01
Number of Observations Used	1.730000e+02
Critical Value (1%)	-3.468726e+00
Critical Value (5%)	-2.878396e+00
Critical Value (10%)	-2.575756e+00
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.



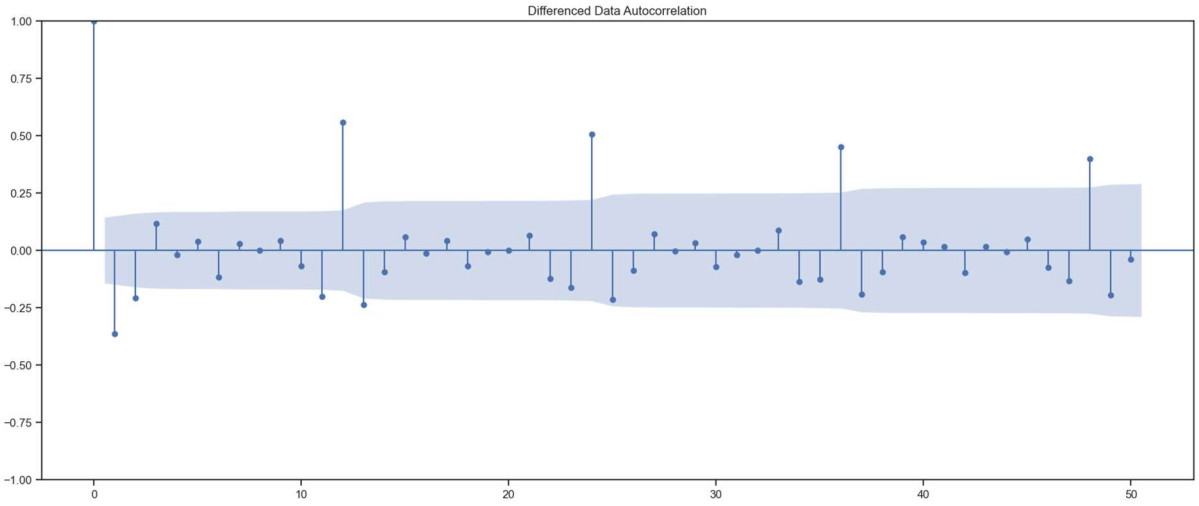


Fig 27 Differenced Data Autocorrelation

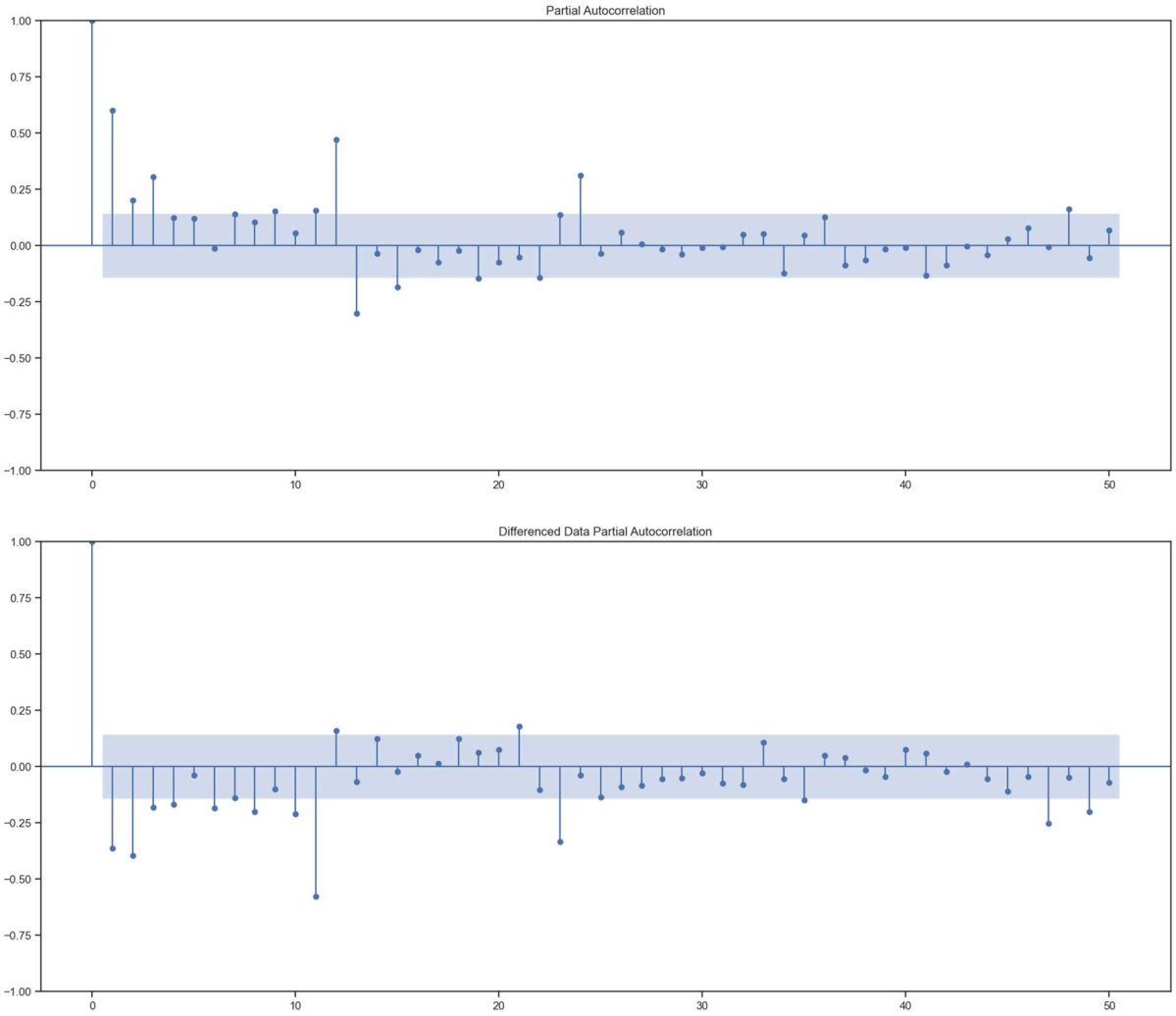


Fig 28 Differenced Data Partial Autocorrelation

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.

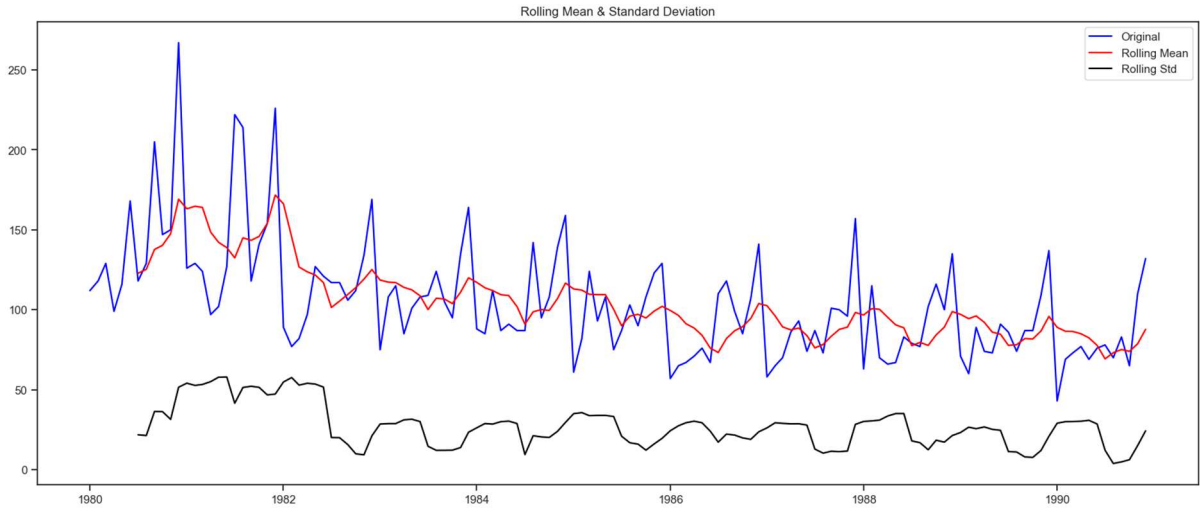


Fig 29 Rolling Mean & Standard Deviation after differencing

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

Table 27 Results of Dickey-Fuller Test after differencing

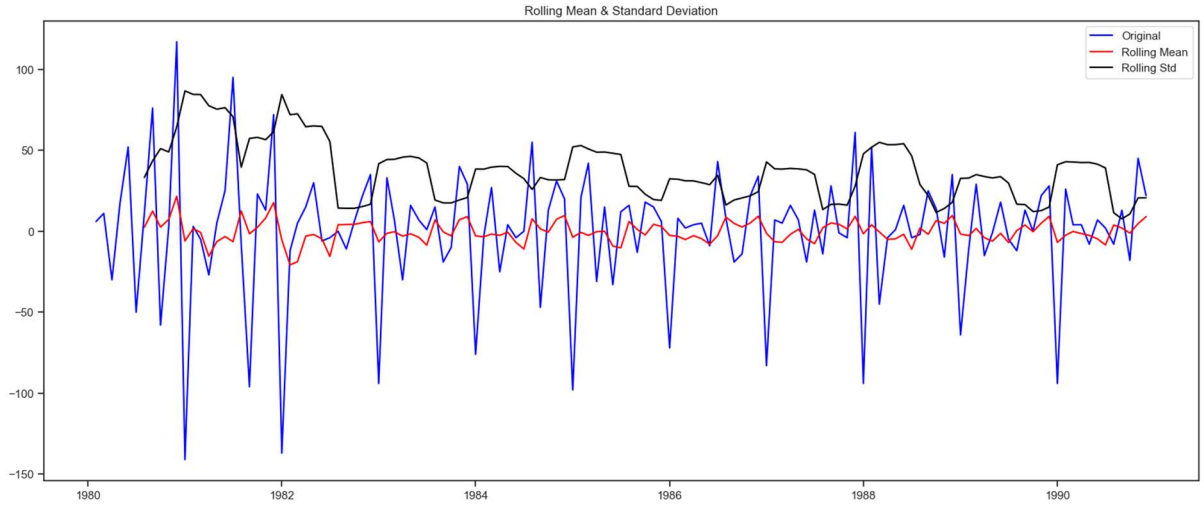


Fig 30 Rolling Mean & Standard Deviation after differencing

Results of Dickey-Fuller Test:	
Test Statistic	-6.592372e+00
p-value	7.061944e-09
#Lags Used	1.200000e+01
Number of Observations Used	1.180000e+02
Critical Value (1%)	-3.487022e+00
Critical Value (5%)	-2.886363e+00
Critical Value (10%)	-2.580009e+00
dtype: float64	

Table 28 Results of Dickey-Fuller Test after differencing

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
11	(2, 1, 3)	1274.695412
15	(3, 1, 3)	1278.667917
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.574230
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

Table 29 AIC values in the ascending order

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.348
Date:	Sun, 10 Dec 2023	AIC	1274.695
Time:	20:41:43	BIC	1291.947
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.705
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6783	0.084	-19.999	0.000	-1.843	-1.514
ar.L2	-0.7291	0.084	-8.687	0.000	-0.894	-0.565
ma.L1	1.0446	0.618	1.691	0.091	-0.166	2.255
ma.L2	-0.7720	0.132	-5.858	0.000	-1.030	-0.514
ma.L3	-0.9045	0.560	-1.616	0.106	-2.002	0.192
sigma2	860.3101	519.823	1.655	0.098	-158.525	1879.145

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.51
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

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

Table 30 results_auto_ARIMA.summary

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

	Test RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Alpha=0.1,Beta=0.8,Gamma=0.2,TripleExponentialSmoothing	8.992350
Auto_ARIMA	36.416372

Table 31 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.

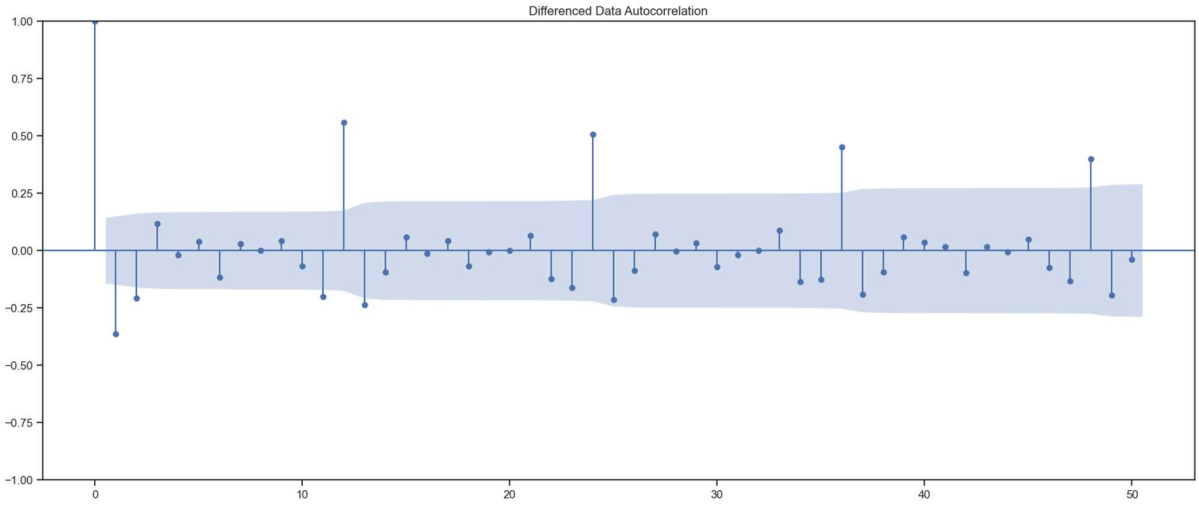


Fig 31 Differenced Data Autocorrelation

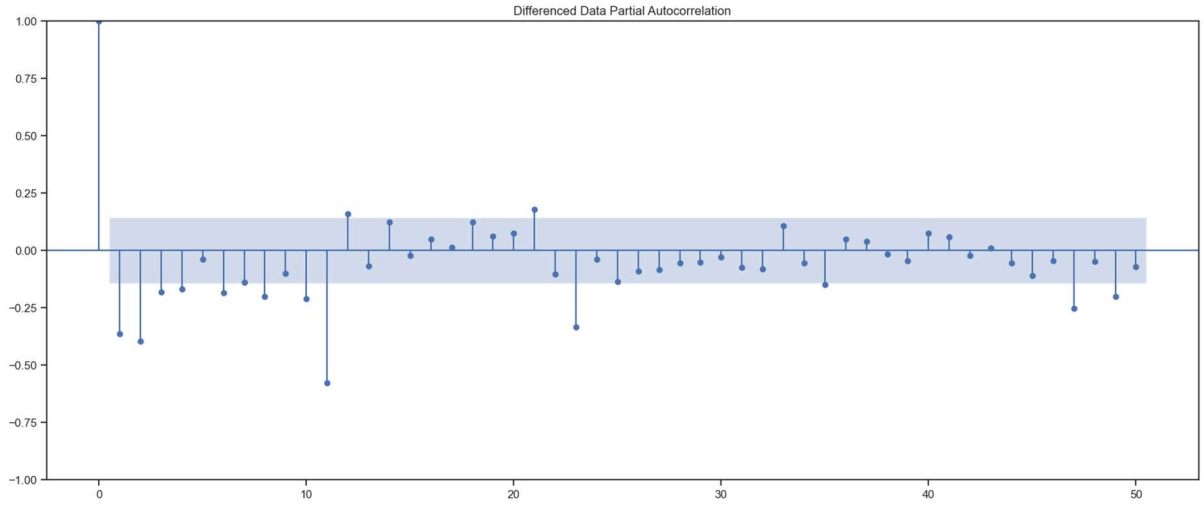


Fig 32 Differenced Data Partial Autocorrelation

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(0, 1, 0)	Log Likelihood	-665.577
Date:	Sun, 10 Dec 2023	AIC	1333.155
Time:	20:47:33	BIC	1336.030
Sample:	01-31-1980	HQIC	1334.323
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1515.6738	122.418	12.381	0.000	1275.740	1755.608

Ljung-Box (L1) (Q):	17.11	Jarque-Bera (JB):	59.55
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.38	Skew:	-0.95
Prob(H) (two-sided):	0.00	Kurtosis:	5.70

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

Table 32 results_auto_ARIMA.summary

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

	RMSE
ARIMA(2,1,2)	36.416372
ARIMA(0,1,2)(2,0,2,6)	79.304391

Table 33 Test RMSE values of ARIMA(2,1,2) & ARIMA(0,1,2)

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

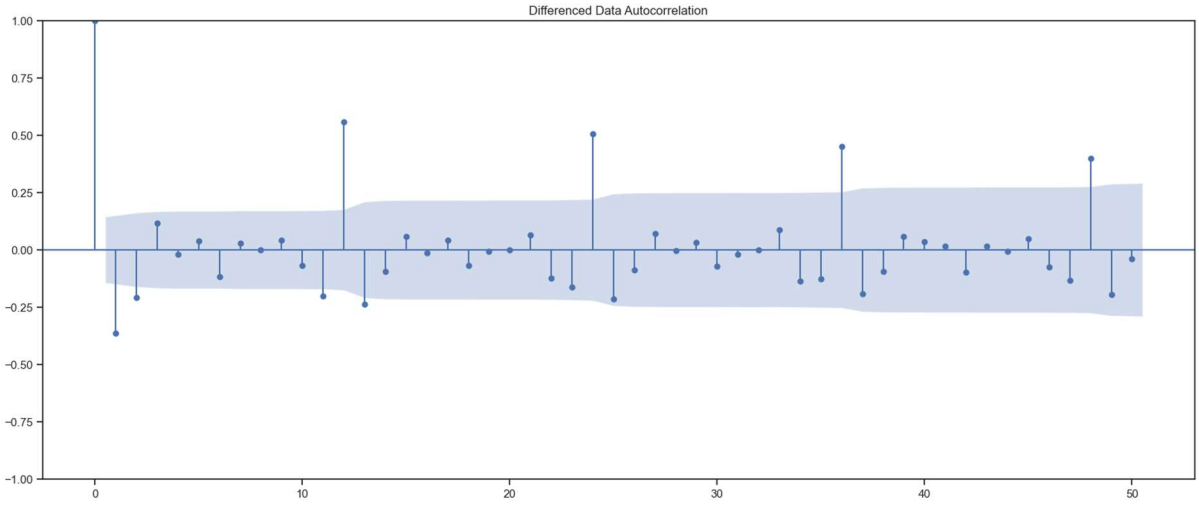


Fig 33 Differenced Data Autocorrelation

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

	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400286
238	(3, 1, 2)	(3, 0, 2, 12)	774.880934
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.495330
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018

Table 34 top 5 SARIMA6_AIC sort rows

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-377.200
Date:	Sun, 10 Dec 2023	AIC	774.400
Time:	22:44:14	BIC	799.618
Sample:	0 - 132	HQIC	784.578
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0464	0.126	0.367	0.714	-0.202	0.294
ar.L2	-0.0060	0.120	-0.050	0.960	-0.241	0.229
ar.L3	-0.1808	0.098	-1.837	0.066	-0.374	0.012
ma.L1	-0.9370	0.067	-13.905	0.000	-1.069	-0.805
ar.S.L12	0.7639	0.165	4.639	0.000	0.441	1.087
ar.S.L24	0.0840	0.159	0.527	0.598	-0.229	0.397
ar.S.L36	0.0727	0.095	0.764	0.445	-0.114	0.259
ma.S.L12	-0.4969	0.250	-1.988	0.047	-0.987	-0.007
ma.S.L24	-0.2191	0.210	-1.044	0.296	-0.630	0.192
sigma2	192.1546	39.628	4.849	0.000	114.486	269.823

Ljung-Box (L1) (Q):	0.30	Jarque-Bera (JB):	1.64
Prob(Q):	0.58	Prob(JB):	0.44
Heteroskedasticity (H):	1.11	Skew:	0.33
Prob(H) (two-sided):	0.77	Kurtosis:	3.03

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

Table 35 results_auto_SARIMA6.summary

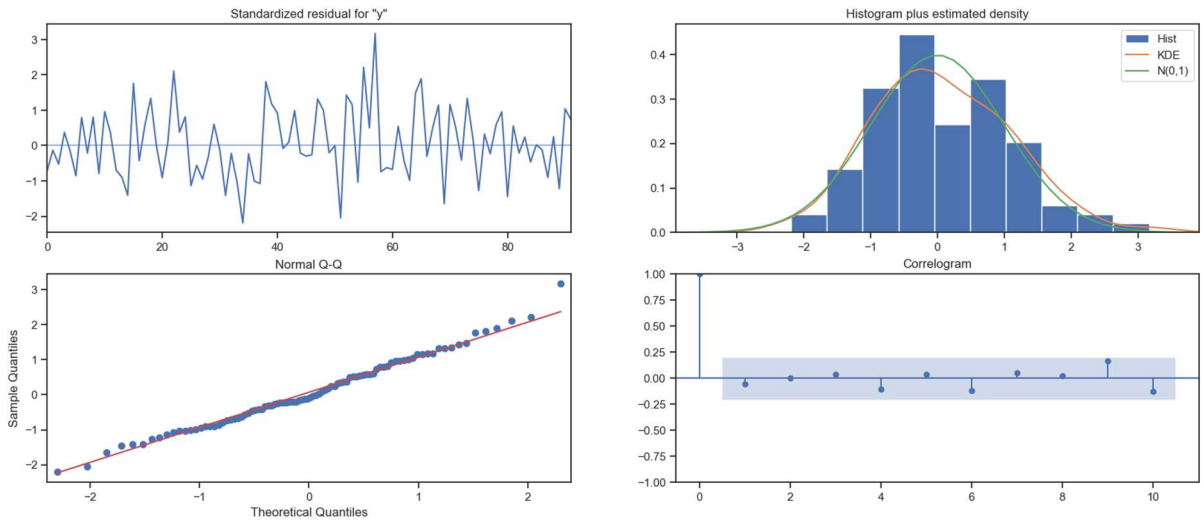


Fig 33 SARIMA diagnostics plot for seasonality as 6

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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	55.235804	13.907559	27.977489	82.494118
1	68.122716	13.990997	40.700866	95.544565
2	67.908796	14.012048	40.445687	95.371906
3	66.786275	14.099369	39.152019	94.420531
4	69.760397	14.108730	42.107795	97.412999

Table 35 SARIMA 6 summary frame

	RMSE
ARIMA(2,1,2)	36.416372
ARIMA(0,1,2)(2,0,2,6)	79.304391
SARIMA(0,1,2)(2,0,2,6)	18.535655

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
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686900
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

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	-440.343
Date:	Sun, 10 Dec 2023	AIC	896.687
Time:	22:36:53	BIC	917.842
Sample:	0	HQIC	905.257
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8742	0.066	13.236	0.000	0.745	1.004
ma.L1	-1.9996	0.161	-12.384	0.000	-2.316	-1.683
ma.L2	1.0000	0.162	6.169	0.000	0.682	1.318
ar.S.L12	0.3728	0.059	6.334	0.000	0.257	0.488
ar.S.L24	0.3210	0.054	5.900	0.000	0.214	0.428
ma.S.L12	0.0295	0.117	0.253	0.800	-0.199	0.258
ma.S.L24	-0.1678	0.133	-1.266	0.206	-0.428	0.092
sigma2	241.8909	0.001	1.81e+05	0.000	241.888	241.893

Ljung-Box (L1) (Q):	4.35	Jarque-Bera (JB):	0.25
Prob(Q):	0.04	Prob(JB):	0.88
Heteroskedasticity (H):	0.78	Skew:	0.11
Prob(H) (two-sided):	0.46	Kurtosis:	2.92

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1e+21. Standard errors may be unstable.

Table 38 results_auto_SARIMA12.summary

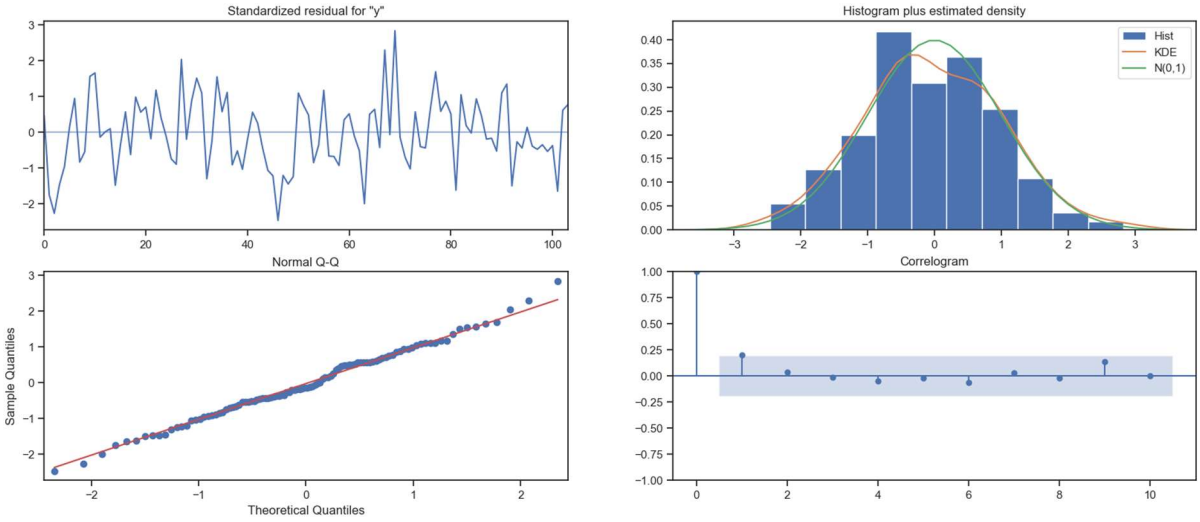


Fig 34 SARIMA diagnostics plot for seasonality as 12

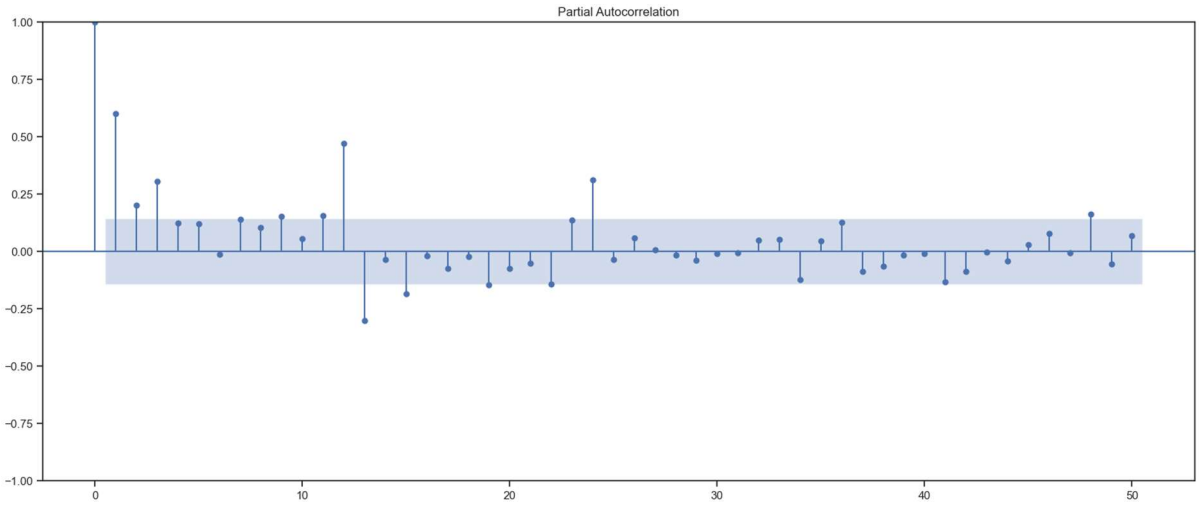
Predict on the Test Set using this model and evaluate the model.

	Test RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Alpha=0.1,Beta=0.8,Gamma=0.2,TripleExponentialSmoothing	8.992350
Auto_ARIMA	36.416372
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535655

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



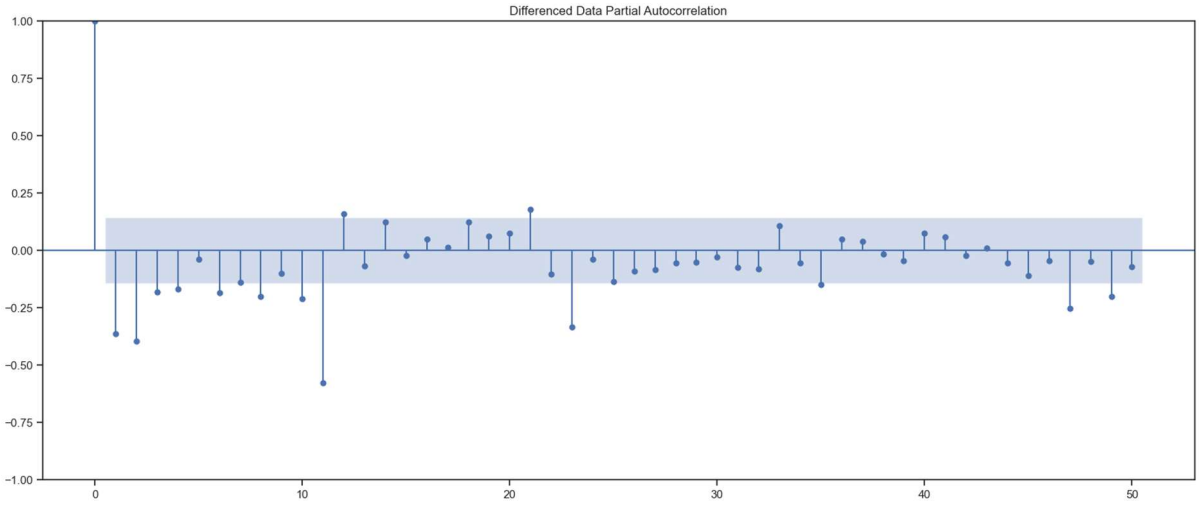


Fig 35 Sale Differenced Data Partial Autocorrelation

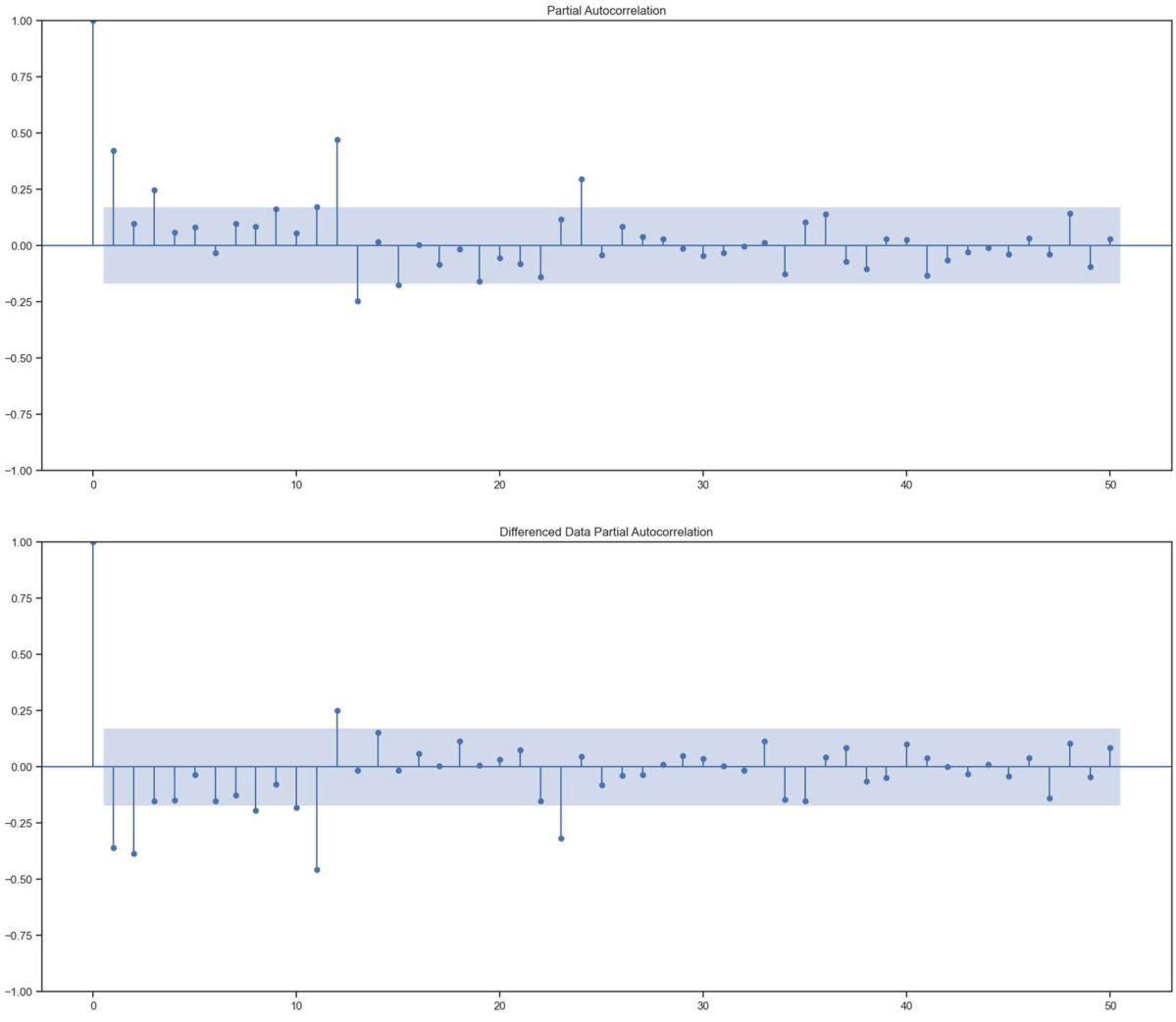


Fig 36 Sale Differenced Data Partial Autocorrelation after drop

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Sun, 10 Dec 2023	AIC	1281.871
Time:	22:54:44	BIC	1296.247
Sample:	01-01-1980 - 12-01-1990	HQIC	1287.712
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4540	0.469	-0.969	0.333	-1.372	0.464
ar.L2	0.0001	0.170	0.001	0.999	-0.334	0.334
ma.L1	-0.2541	0.459	-0.554	0.580	-1.154	0.646
ma.L2	-0.5984	0.430	-1.390	0.164	-1.442	0.245
sigma2	952.1601	91.424	10.415	0.000	772.973	1131.347

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	34.16
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.79
Prob(H) (two-sided):	0.00	Kurtosis:	4.94

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

Table 40 results_auto_ Manual ARIMA.summary

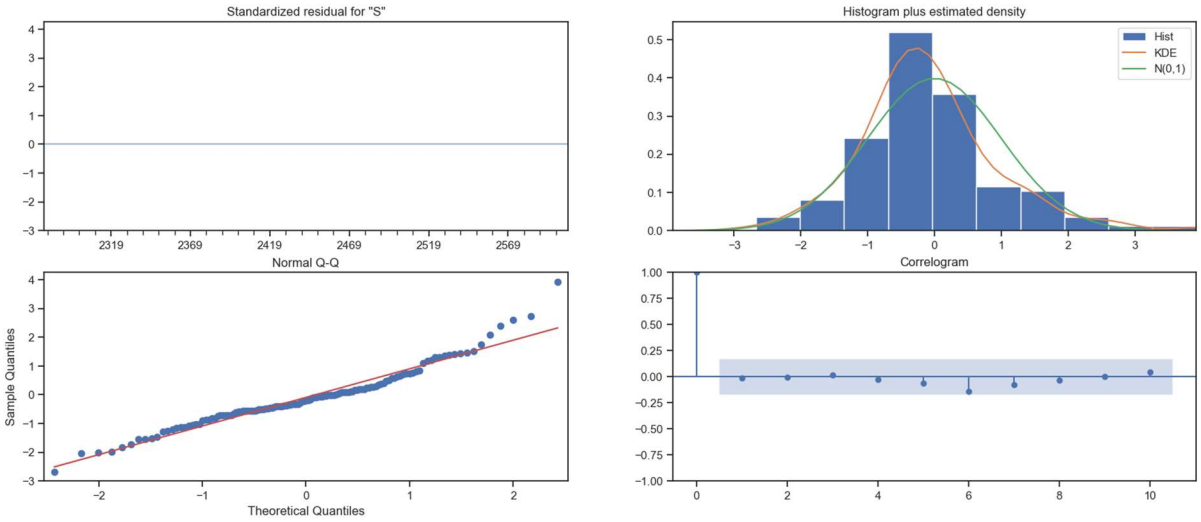


Fig 37 Manual ARIMA diagnostics plot

	Test RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Alpha=0.1,Beta=0.8,Gamma=0.2,TripleExponentialSmoothing	8.992350
Auto_ARIMA	36.416372
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535655
ARIMA(3,1,3)	36.473225

Table 41 Test RMSE values of Regression to ARIMA

Model 12 : Manual SARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-538.016
Date:	Sun, 10 Dec 2023	AIC	1094.031
Time:	22:57:36	BIC	1119.044
Sample:	0	HQIC	1104.188
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5492	0.228	-2.410	0.016	-0.996	-0.103
ar.L2	-0.0744	0.099	-0.752	0.452	-0.268	0.119
ma.L1	-0.1701	0.216	-0.787	0.431	-0.594	0.254
ma.L2	-0.6696	0.228	-2.939	0.003	-1.116	-0.223
ar.S.L12	-1.0133	0.524	-1.934	0.053	-2.040	0.014
ar.S.L24	-0.1000	0.175	-0.570	0.568	-0.443	0.244
ma.S.L12	0.2912	68.873	0.004	0.997	-134.698	135.281
ma.S.L24	-0.7083	48.888	-0.014	0.988	-96.528	95.111
sigma2	430.2067	2.94e+04	0.015	0.988	-5.72e+04	5.81e+04

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	27.17
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	0.26
Prob(H) (two-sided):	0.00	Kurtosis:	5.28

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

Table 42 results_auto_ Manual SARIMA.summary

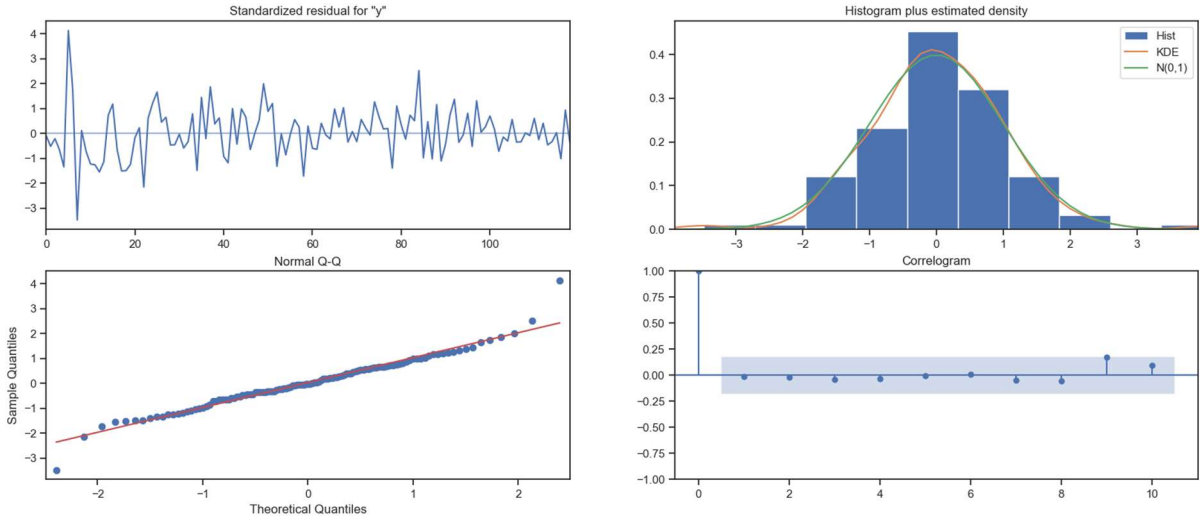


Fig 38 Manual SARIMA diagnostics plot

	Test RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Alpha=0.1,Beta=0.8,Gamma=0.2,TripleExponentialSmoothing	8.992350
Auto_ARIMA	36.416372
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535655
(2,1,2)(2,1,2,12),Manual_SARIMA	14.972976

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.

We can clearly see that triple exponential smoothing model with alpha 0.1, beta 0.8 and gamma 0.2 is the best as it he the lowest RSME score.

	Test RMSE
Alpha=0.1,Beta=0.8,Gamma=0.2,TripleExponentialSmoothing	8.992350
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
(2,1,2)(2,1,2,12),Manual_SARIMA	14.972976
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535655
Alpha=0.995,SimpleExponentialSmoothing	36.397777
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Auto_ARIMA	36.416372
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Linear Regression	51.080941
Simple Average Model	53.049755
Naive Model	79.304391

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	44.510965
1995-09-01	41.726134
1995-10-01	45.505262
1995-11-01	53.516557
1995-12-01	76.944425
1996-01-01	28.806916
1996-02-01	36.473494
1996-03-01	42.771692
1996-04-01	45.418290
1996-05-01	35.490857
1996-06-01	43.504695
1996-07-01	51.516141

Table 45 future_predictions rows

	lower_CI	prediction	upper_ci
1995-08-01	5.303162	44.510965	83.718767
1995-09-01	2.518332	41.726134	80.933937
1995-10-01	6.297460	45.505262	84.713065
1995-11-01	14.308754	53.516557	92.724359
1995-12-01	37.736623	76.944425	116.152227

Table 46 future_predictions with lower_ci and upper_ci

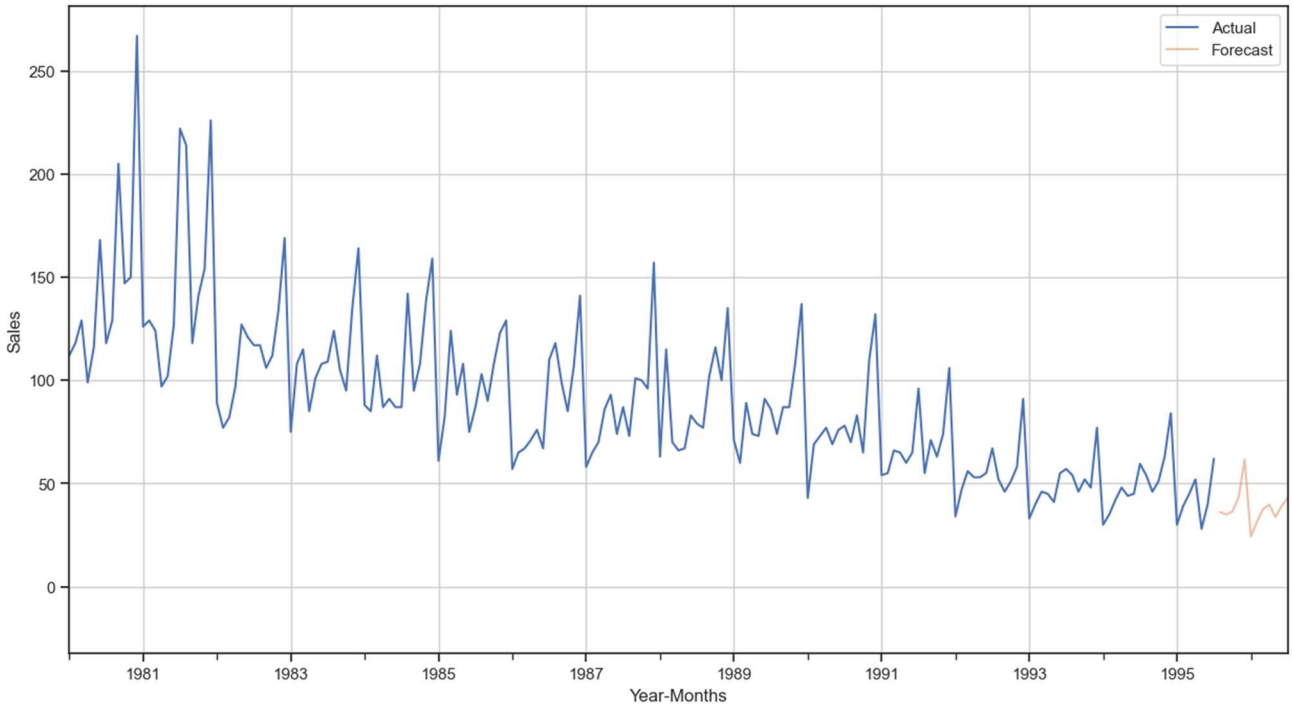


Fig 39 actual and forecast along with the confidence band

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

Insights and recommendations:

- For more than ten years, the Rose wine sales have been going down steadily. This happens especially during winter (January) and goes up during festive times.
- It's better to advertise the wine when sales are low, like from April to June. Campaigns during big festivals might not make much difference because sales are already high then.
- People tend to buy less wine in January due to the cold weather. So, advertising during that time might not change their minds.
- Spend more on advertising during quieter months to boost sales when they're usually low.
- Figure out why people aren't buying Rose wine as much. Then, change how it's made or how it's advertised to get more people interested again.

- Try new things with the wine or talk more about what makes it special. This could bring back people's interest.
- By advertising cleverly during quieter times and finding out why people aren't buying as much, the company can try to turn around the falling sales of Rose wine and get more people excited about it again.