Project On Time Series Forecasting

Sparkling Wine

Yogesh Negi

Problem Statement:

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.

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

Data Dictionary:

Column name	Details
YearMonth	Dates of sales
Sparkling	Sales of sparkling wine

Table 1: data dictionary

Data set is read using pandas library.

Rows of dataset;

Top Few Rows:

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

Last Few Rows:

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

Table 2: rows of data

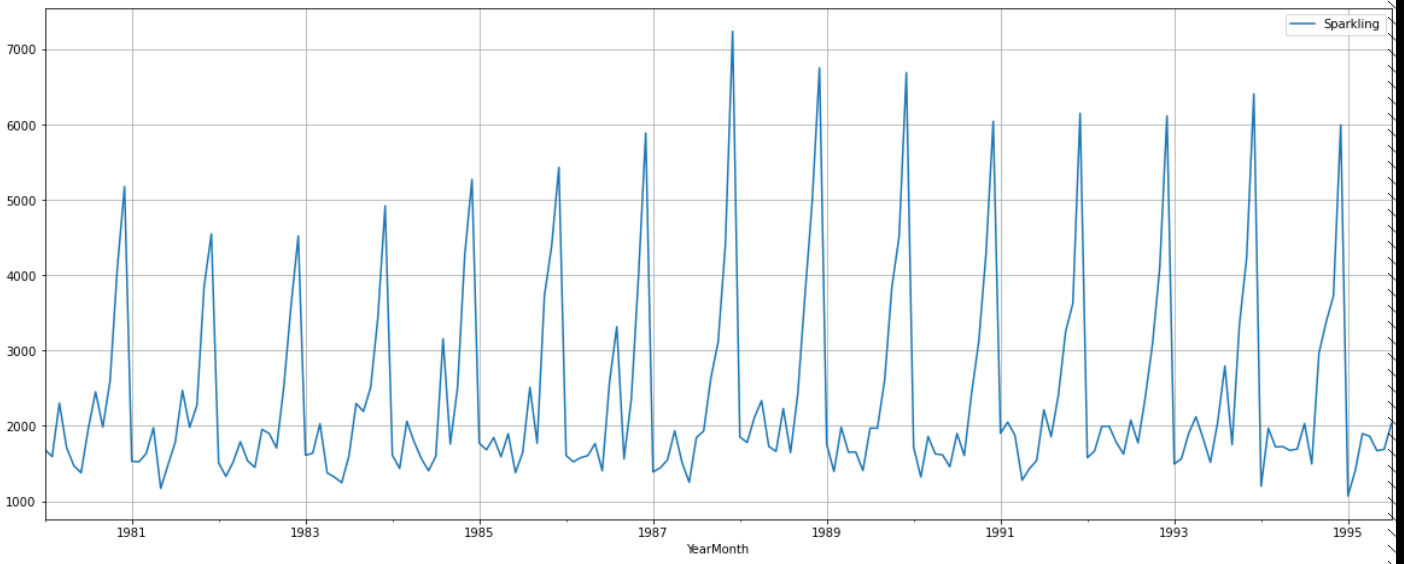
We can see data is from 1980 to 1995.

The number of Rows and Columns of the Dataset:

The dataset has 187 rows and 1 column.

Plot of the dataset:

Plot 1: dataset



Post Ingestion of Dataset:

We have divided the dataset further by extraction month and year columns from the YearMonth column and renamed the sparkling column name to Sales for better analysis of the dataset. The new dataset has 187 rows and 3 columns.

Rows of the new data set;

Table 3 : rows of new dataset

Top Few Rows:

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:

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

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

Data Type;

Index: DateTime
Sales: integer
Month: integer
Year: integer

Statistical summary:

Table 4: statistical summary of data

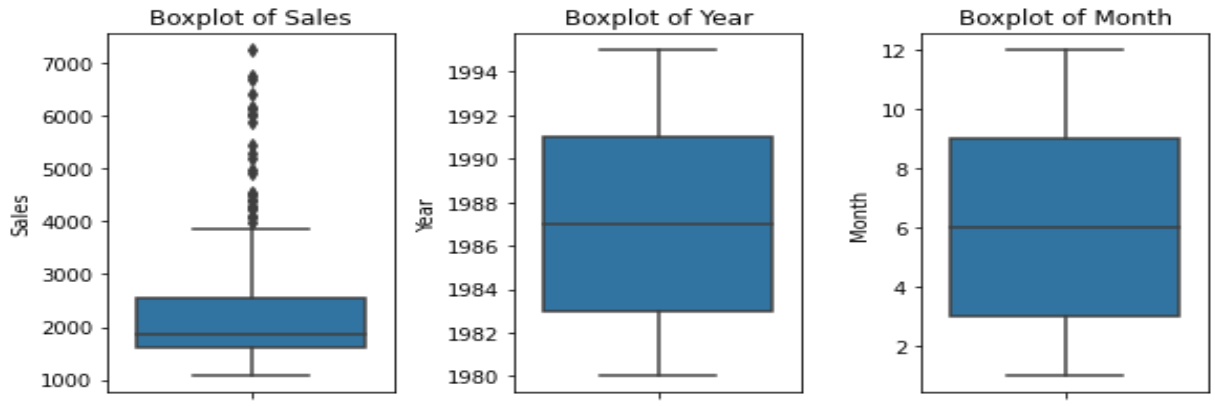
	count	mean	std	min	25%	50%	75%	max
Sales	187.0	2402.0	1295.0	1070.0	1605.0	1874.0	2549.0	7242.0
Year	187.0	1987.0	5.0	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.0	3.0	1.0	3.0	6.0	9.0	12.0

Null Value:

There are no null values present in the dataset. So we can do further analysis smoothly.

Boxplotofdataset:

plot 2: box plot of data

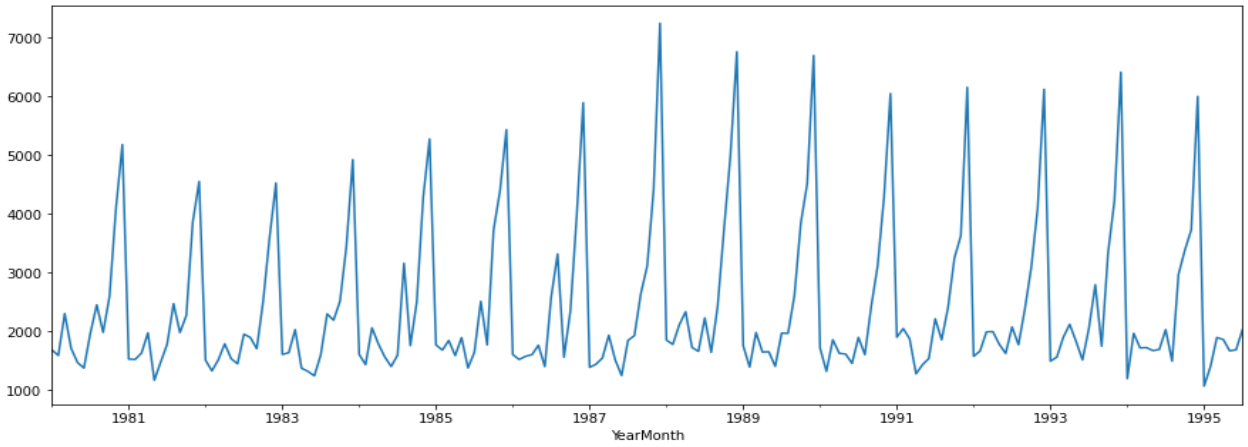


The box plot shows:

- Sales boxplot has outliers we can treat them but we are choosing not to treat them as they do not give much effect on the time series model.

Line plot of sales:

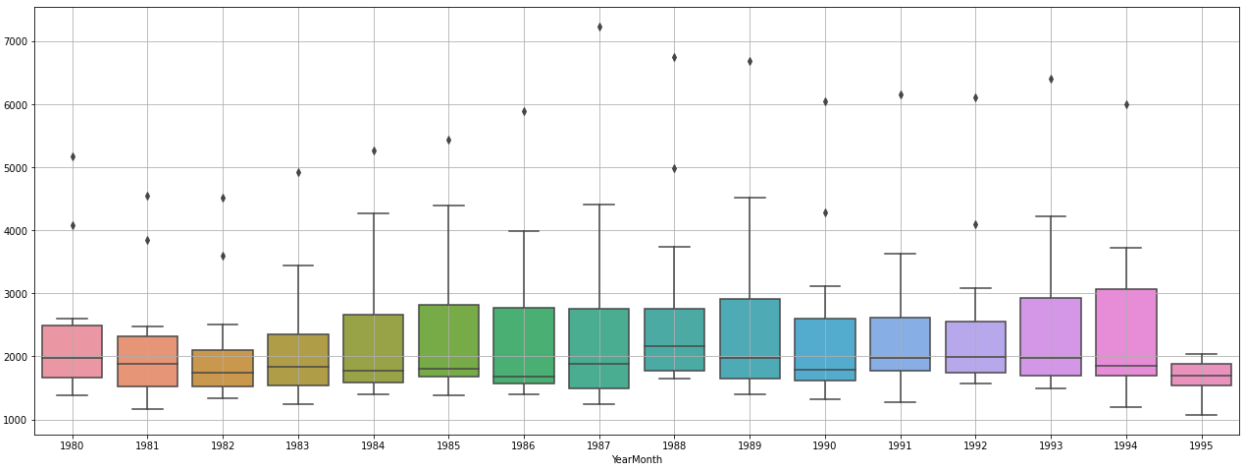
Plo 3: line plot for sales



The line plot shows the patterns of trend and seasonality and also shows that there was a peak in the year 1988.

Boxplot Yearly:

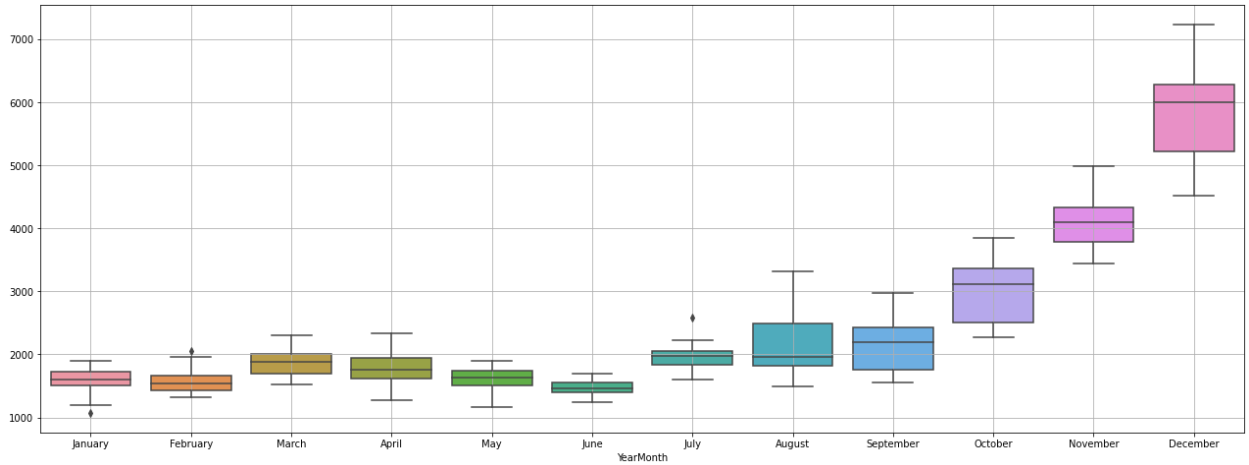
Plot 4: boxplot yearly



This yearly box plot shows there is consistency over the years and there was a peak in 1988-1989. Outliers are present in all years.

Boxplot Monthly:

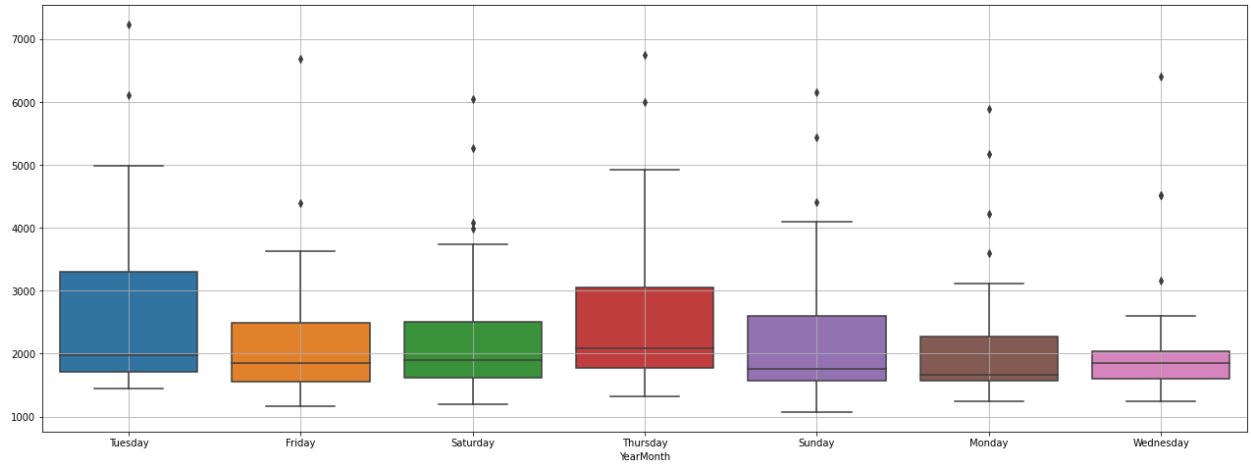
Plot 5: boxplot monthly



The plot shows that 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 January, February and July.

Boxplot Weekdaywise:

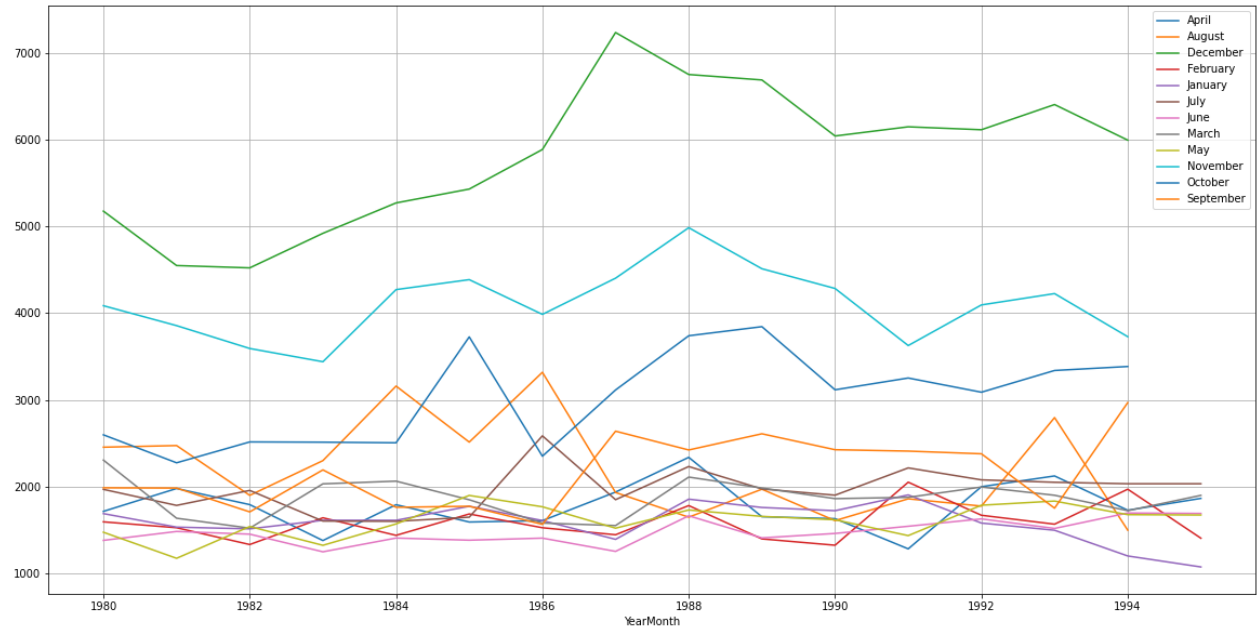
Plot 6: Boxplot weekday vise



Tuesday has more sales than other days and Wednesday has the lowest sales of the week. Outliers are present on all days which is understandable.

Graph of Monthly Sales over the years:

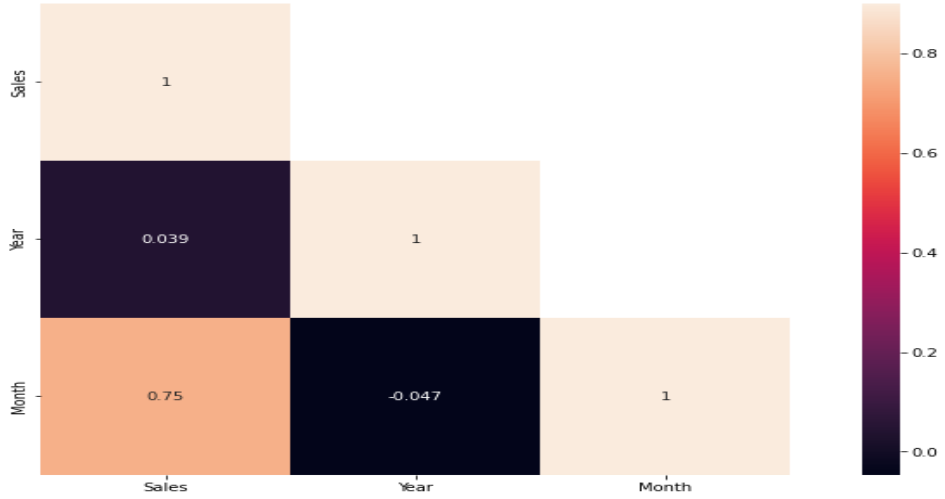
Plot 7: graph of monthly sales over the year



This plot shows that December has the highest sales over the years and the year 1988 was the year with the highest number of sales.

Correlation plot

Plot 8: correlation plot

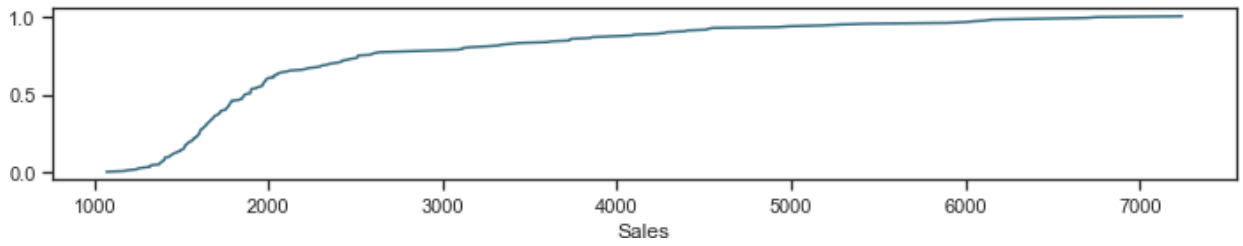


This heat map shows that there is a low correlation between sales and year. there is a more correlation between month and sales. It indicated seasonal patterns in sales

Plot ECDF: Empirical Cumulative Distribution Function

This graph shows the distribution of data.

Plot 9: ECDF plot

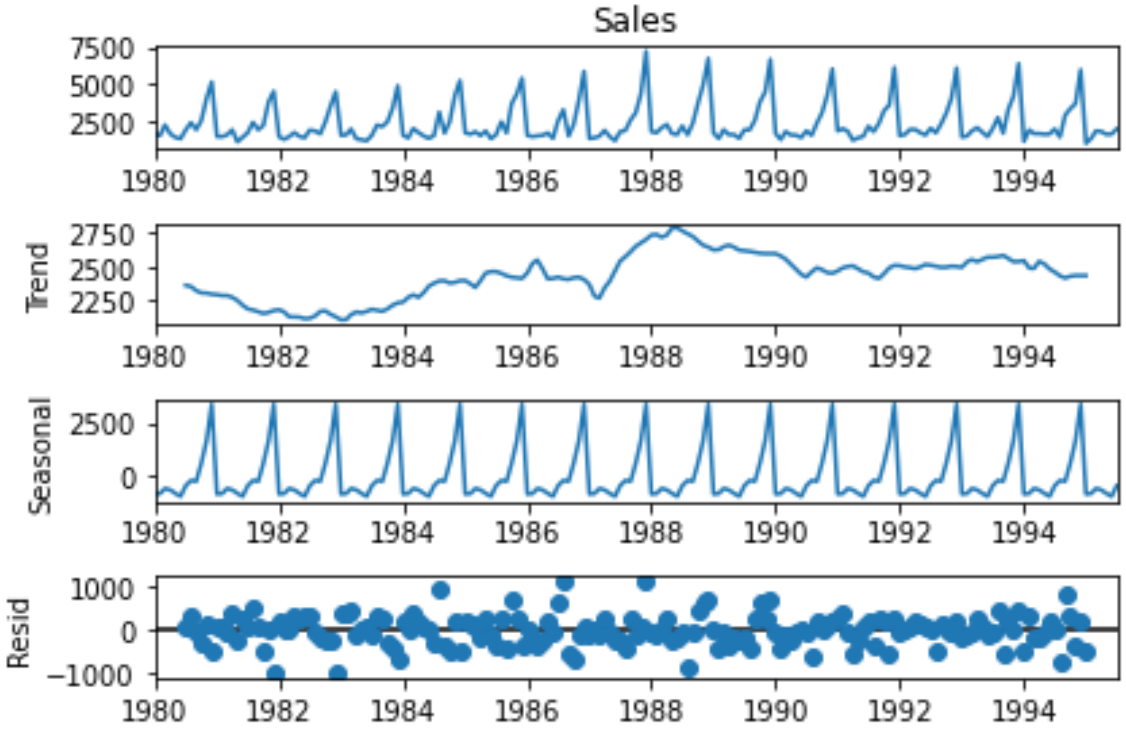


This plot shows:

- More than 50% of sales have been less than 2000
- Highest values is 7000
- Approx 80% of sales have been less than 3000

Decomposition -Additive

Plot 10: decomposition plot addictive



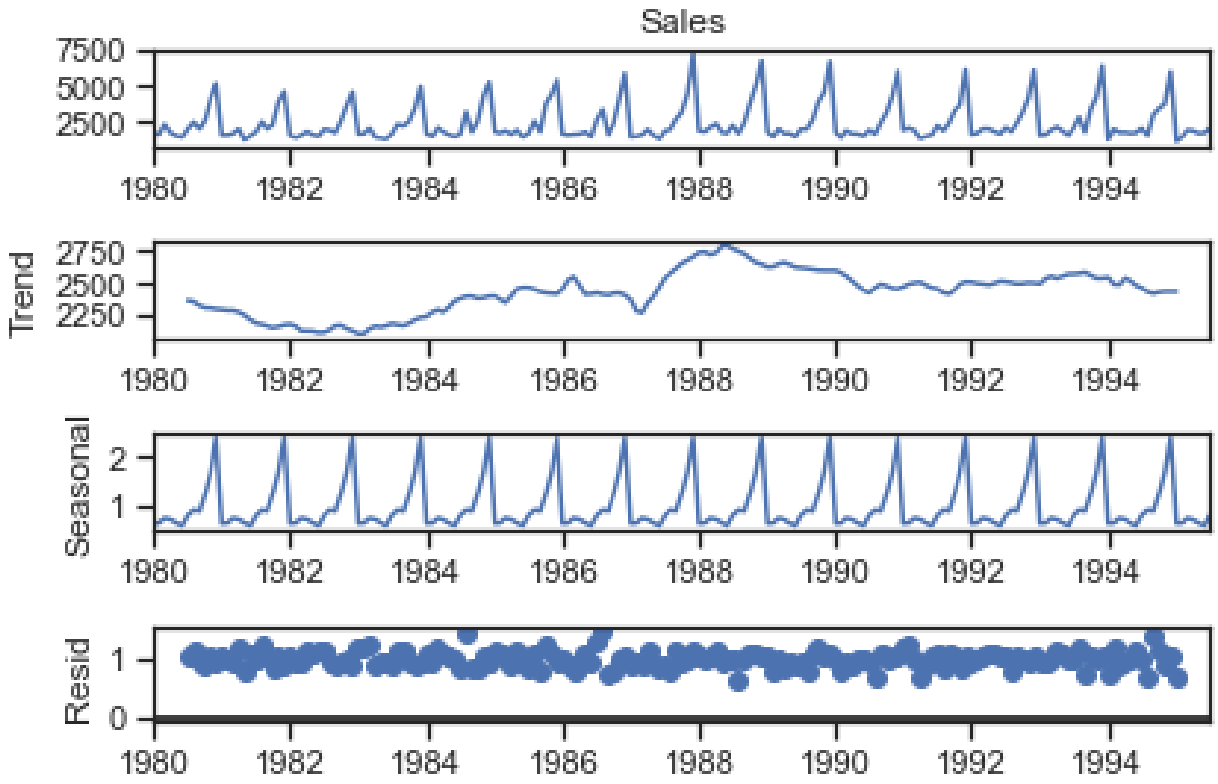
The plots show:

- Peak year 1988-1989

- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is not in a straight line.
- Both trend and seasonality are present.

Decomposition-Multiplicative

Plot 11: decomposition plot - mulatiplicative

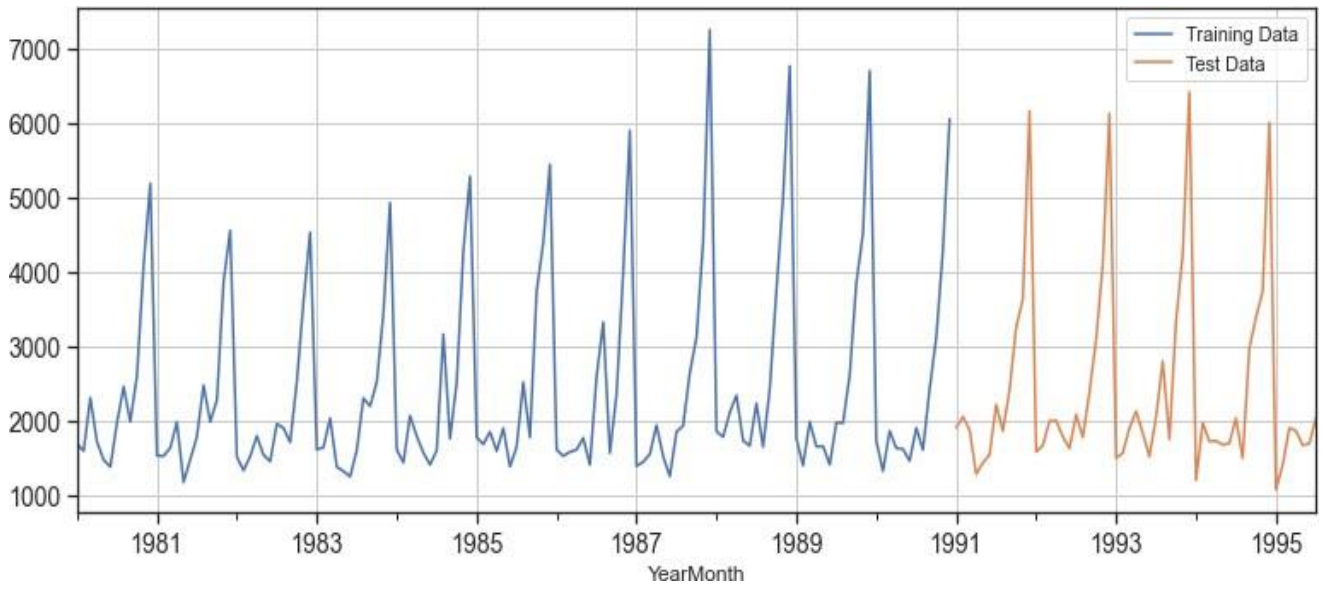


The plots show

- Peak year 1988-1989
- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is in approx a straight line.
- Both trend and seasonality are present.
- Reside is 0 to 1, while additive is 0 to 1000.
- So multiplicative model is selected owing to a more stable residual plot and lower range of residual

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

plot 12: line plot train and test dataset



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.

Rows and Columns:

train dataset has 132 rows and 3 columns. test dataset has 55 and 3 columns.

Few Rows of datasets:

Table 5: train and test dataset

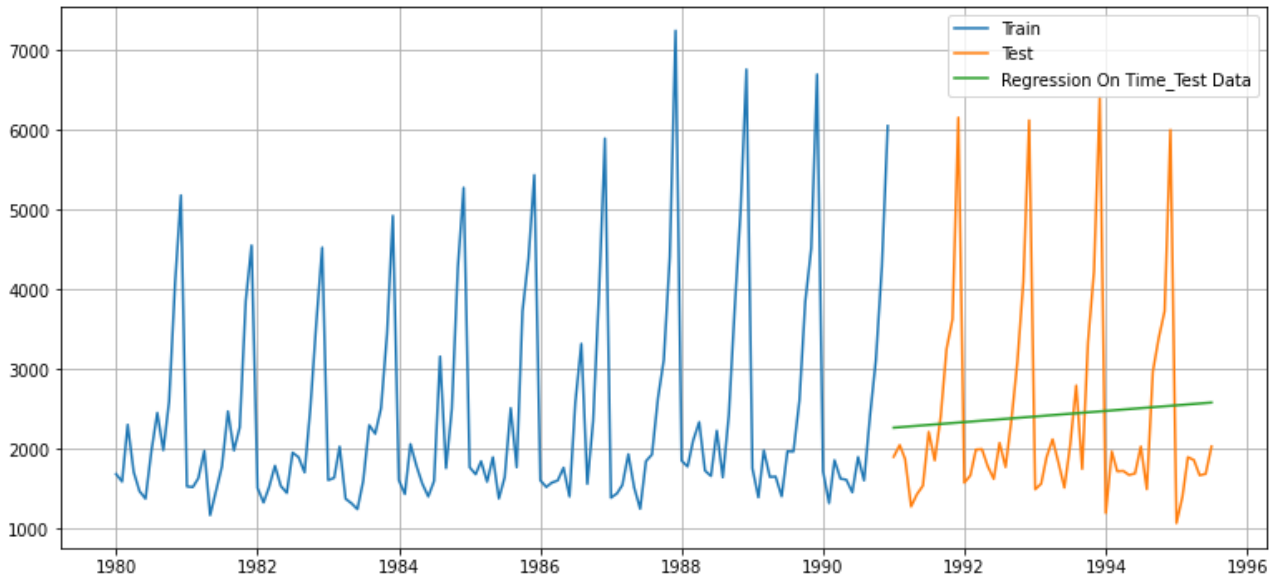
Train dataset	Test dataset
First few rows of Training Data Sales Year Month YearMonth 1980-01-01 1686 1980 1 1980-02-01 1591 1980 2 1980-03-01 2304 1980 3 1980-04-01 1712 1980 4 1980-05-01 1471 1980 5 Last few rows of Training Data Sales Year Month YearMonth 1990-08-01 1605 1990 8 1990-09-01 2424 1990 9 1990-10-01 3116 1990 10 1990-11-01 4286 1990 11 1990-12-01 6047 1990 12	First few rows of Test Data Sales Year Month YearMonth 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 Sales Year Month YearMonth 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

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, and simple 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

Plot 13: linear regression



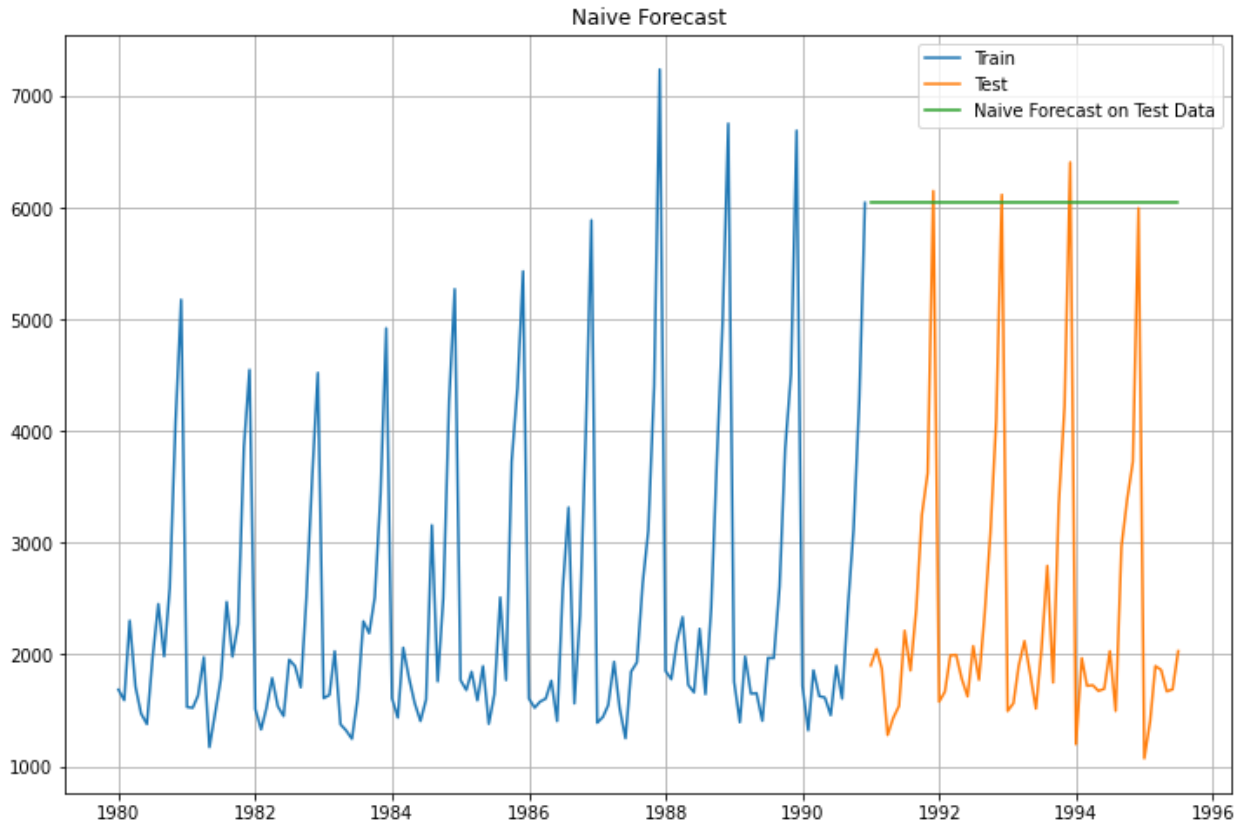
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Linear Regression	1275.867052

Model 2: Naive Approach:

Plot 14: naive approve



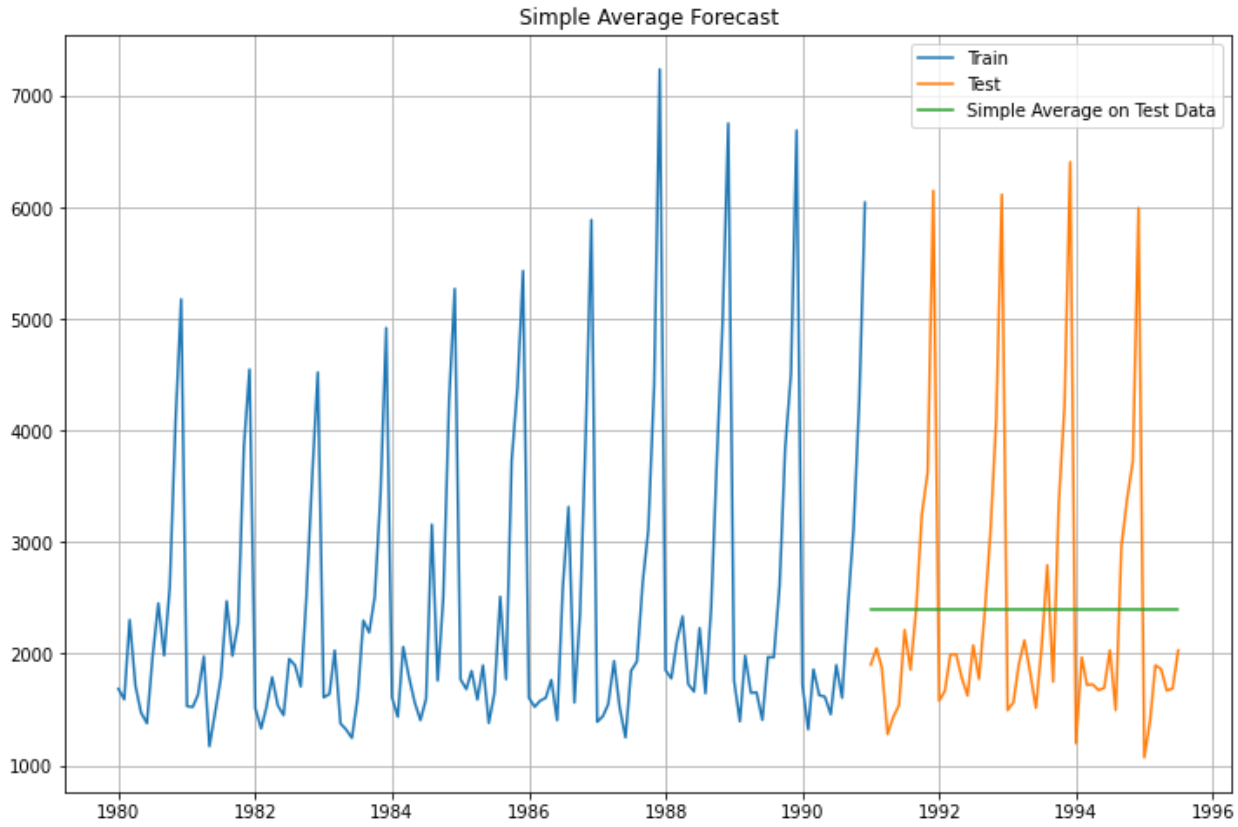
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Naive Model	3864.279352

Model 3: Simple Average

Plot 15: simple average



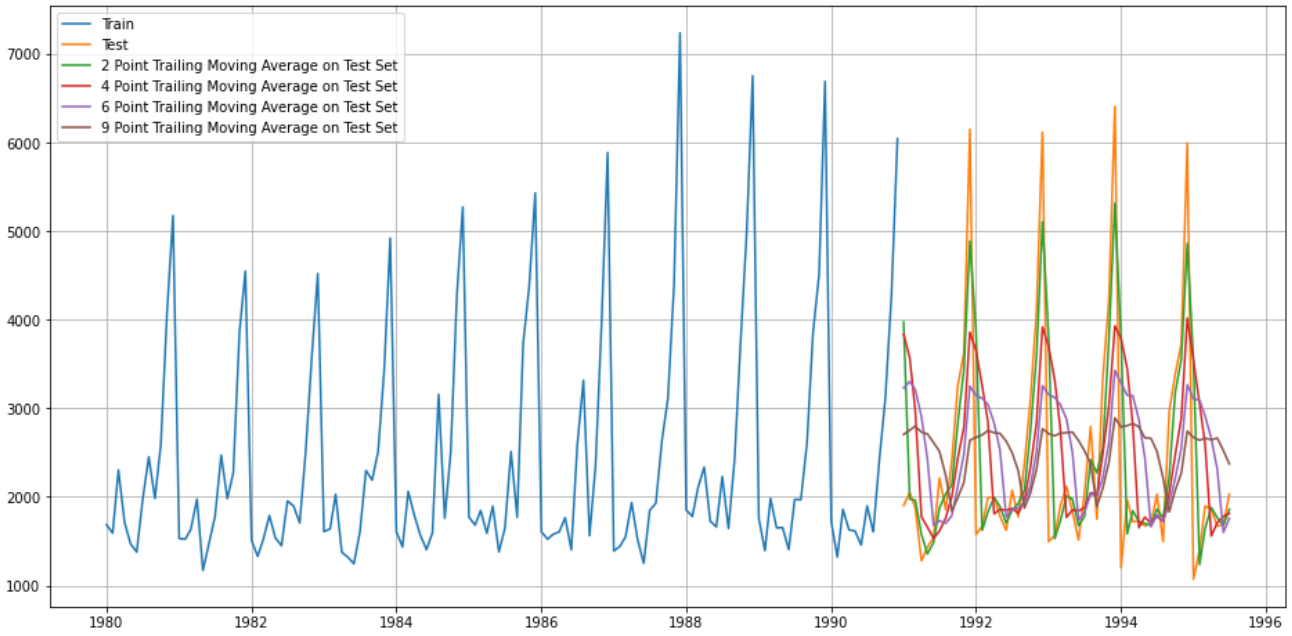
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Simple Average Model 1275.081804

Model 4: Moving Average

Plot 16: moving average



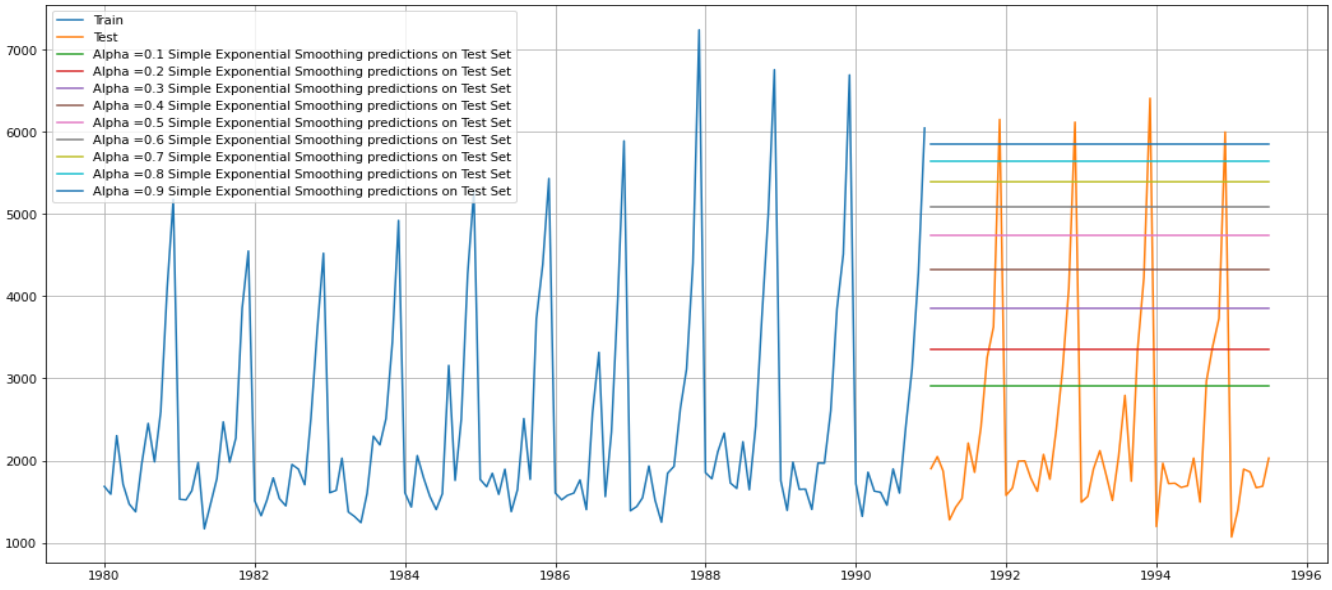
Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

We have made multiple moving average models with rolling windows varying from 2 to 9. Rolling average is a better method than simple average as it takes into account only the previous n values to make the prediction, where n is the rolling window defined. This takes into account the recent trends and is in general more accurate. The higher the rolling window, the smoother will be its curve, since more values are being taken into account.

Model 5: Simple Exponential Smoothing

Plot 17: simple exponential smothing

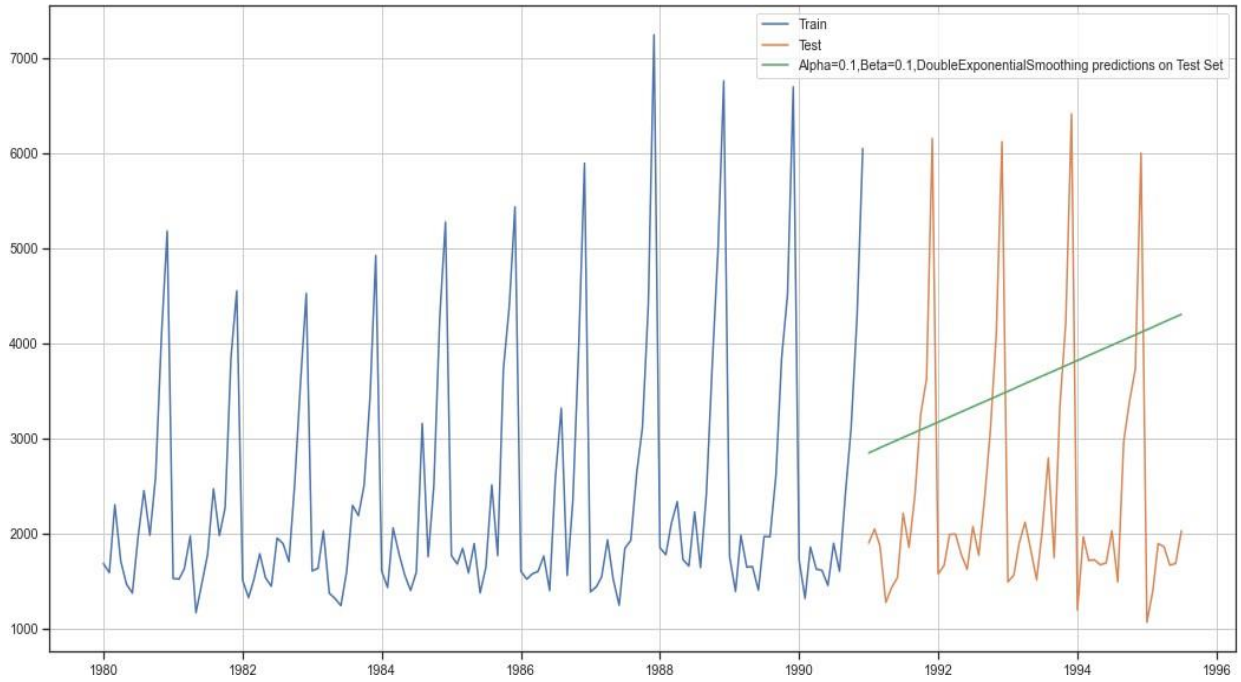


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha Values	Test RMSE
0.1	1375.393398
0.2	1595.206839
0.3	1935.507132
0.4	2311.919615
0.5	2666.351413
0.6	2979.204388
0.7	3249.944092
0.8	3483.801006

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

Plot 18: double exponential smoothing



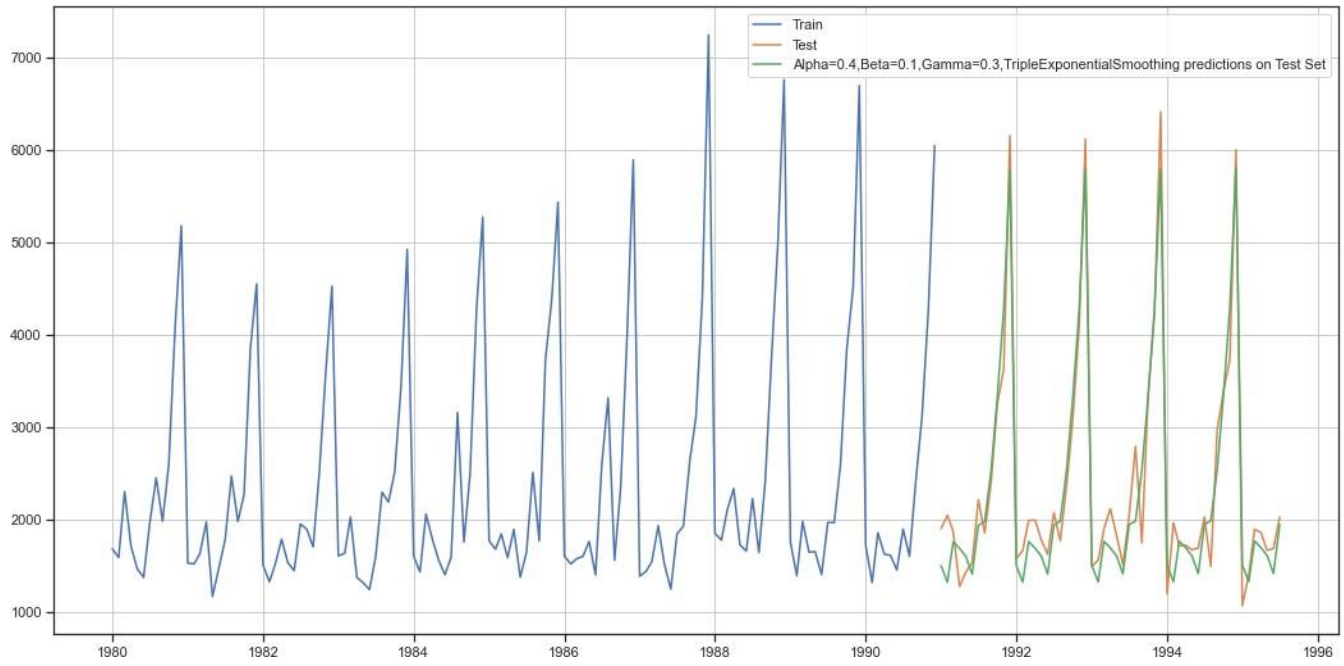
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing = 1778.564670

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

Plot 19: triple exponential smoothing



Output for a best alpha, beta, and gamma values are shown by the green color line in the above plot. The best model had both a multiplicative trends, as well as a seasonality Model, which was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing 317.434302

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

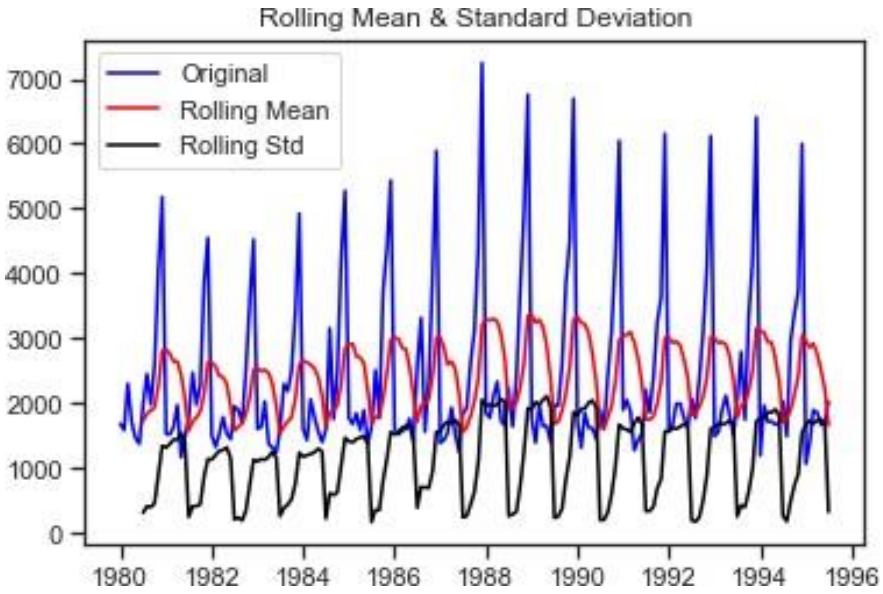
The hypothesis in a simple form for the ADF test is:

- H0 : 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 the α value.

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

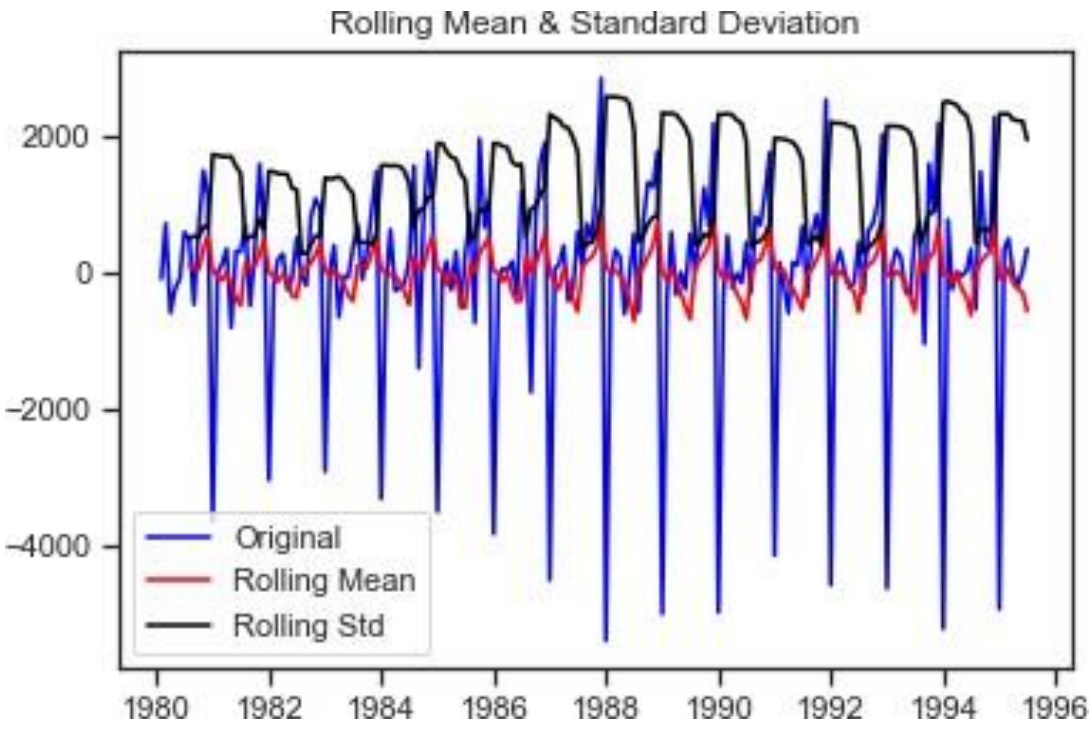
Plot 20: plot for dickey fuller test



Results of Dickey-Fuller Test:
p-value 0.601061

In order to try and make the series stationary we used the differencing approach. We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped

Plot 21: plot for dickey fuller test after differencing approch



Results of Dickey-Fuller Test:

p-value 0.000000

Dickey - Fuller test was 0.000, which is obviously less than 0.05. Hence the null hypothesis that the series is not stationary at difference = 1 was rejected, which implied that the series has indeed become stationary after we performed the differencing. Null hypothesis was rejected since the p-value was less than alpha i.e. 0.05. Also the rolling mean plot was a straight line this time around. Also the series looked more or less the same from both the directions, indicating stationarity.

We could now proceed ahead with ARIMA/ SARIMA models, since we had made the series stationary.

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

We employed a for loop for determining the optimum values of p,d,q, where p is the order of the AR (Auto-Regressive) part of the model, while q is the order of the MA (Moving Average) part of the model. d is the differencing that is required to make the series stationary. p,q values in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for d, since we had already determined d to be 1, while checking for stationarity using the ADF test.

Some parameter combinations for the Model... Model: (0, 1, 1)

Model: (0, 1, 2)

Model: (0, 1, 3)

Model: (1, 1, 0)

Model: (1, 1, 1)

Model: (1, 1, 2)

Model: (1, 1, 3)

Model: (2, 1, 0)

Model: (2, 1, 1)

Model: (2, 1, 2)

Model: (2, 1, 3)

Model: (3, 1, 0)

Model: (3, 1, 1)

Model: (3, 1, 2)
Model: (3, 1, 3)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected.

	param	AIC
10	(2, 1, 2)	2213.509213
15	(3, 1, 3)	2221.458583
14	(3, 1, 2)	2230.768028
11	(2, 1, 3)	2232.885328
9	(2, 1, 1)	2233.777626
3	(0, 1, 3)	2233.994858
2	(0, 1, 2)	2234.408323
6	(1, 1, 2)	2234.5272
13	(3, 1, 1)	2235.49868
7	(1, 1, 3)	2235.607812
5	(1, 1, 1)	2235.755095
12	(3, 1, 0)	2257.723379
8	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
4	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

the summary report for the ARIMA model with values (p=2,d=1,q=2).

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -1101.755
Date:                Sat, 08 Jul 2023   AIC                           2213.509
Time:                        08:22:24   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.740      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.216      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.108      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 3.65e+28. Standard
errors may be unstable.
```

RMSE values are as below:

Auto_ARIMA	1299.978401

AUTO- SARIMA Model

A similar for loop like AUTO_ARIMA with below values was employed, resulting in the models shown below.

p = q = range(0, 4) d= range(0,2) D = range(0,2) pdq = list(itertools.product(p, d, q)) model_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, D, q))]

Examples of some parameter combinations for

Model... Model: (0, 1, 1)(0, 0, 1, 12)

Model: (0, 1, 2)(0, 0, 2, 12)

Model: (0, 1, 3)(0, 0, 3, 12)

Model: (1, 1, 0)(1, 0, 0, 12)

Model: (1, 1, 1)(1, 0, 1, 12)

Model: (1, 1, 2)(1, 0, 2, 12)

Model: (1, 1, 3)(1, 0, 3, 12)

Model: (2, 1, 0)(2, 0, 0, 12)

Model: (2, 1, 1)(2, 0, 1, 12)

Model: (2, 1, 2)(2, 0, 2, 12)

Model: (2, 1, 3)(2, 0, 3, 12)

Model: (3, 1, 0)(3, 0, 0, 12)

Model: (3, 1, 1)(3, 0, 1, 12)

Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected. Here only the top 5 models are shown.

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584248
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)	1557.340403

the summary report for the best SARIMA model with values (2,1,2)(2,0,2,12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Sat, 08 Jul 2023	AIC	1555.584
Time:	08:47:07	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

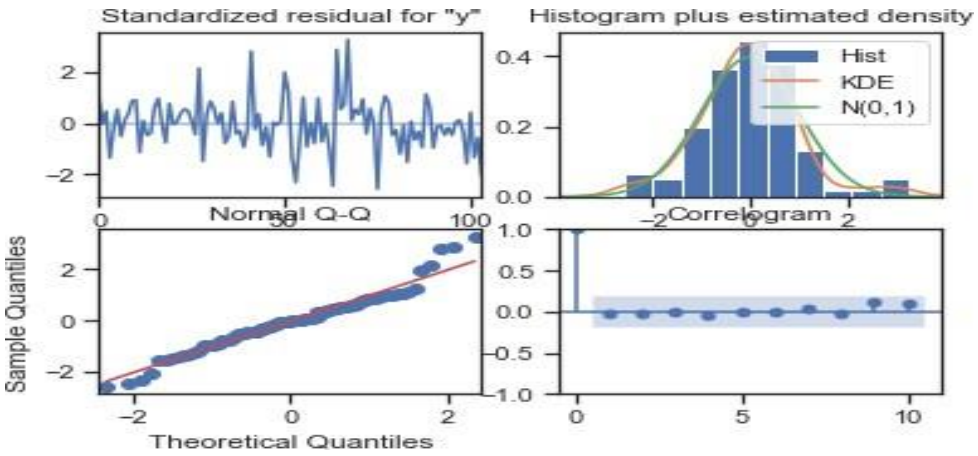
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6281	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.734	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.842	0.000	1.016	1.072
ma.S.L12	-0.5551	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1355	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.400	0.000	1.11e+05	1.9e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	11.72
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

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

We also plotted the graphs for the residual to determine if any further information can be extracted or all the usable information has already been extracted. Below were the plots for the best auto SARIMA model.

Plot 22: SARIMA plot

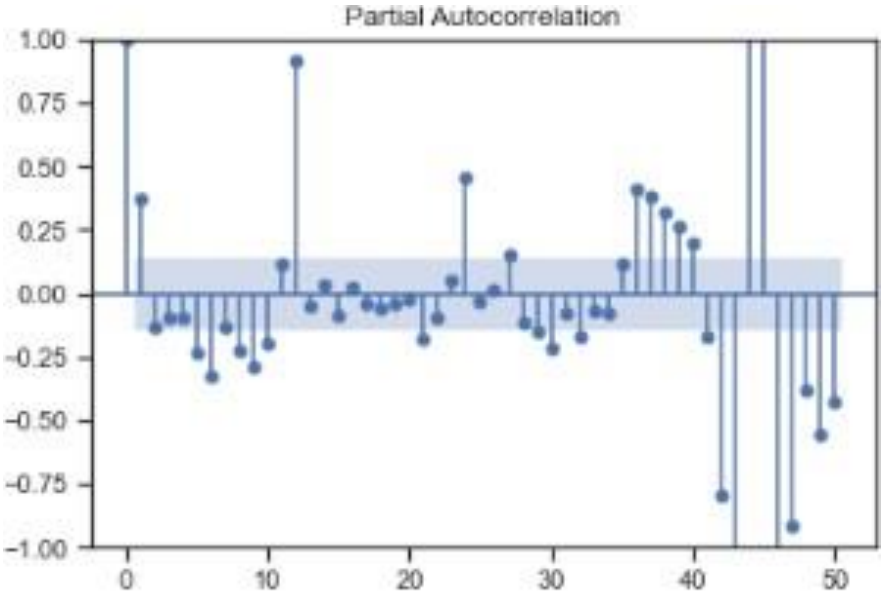


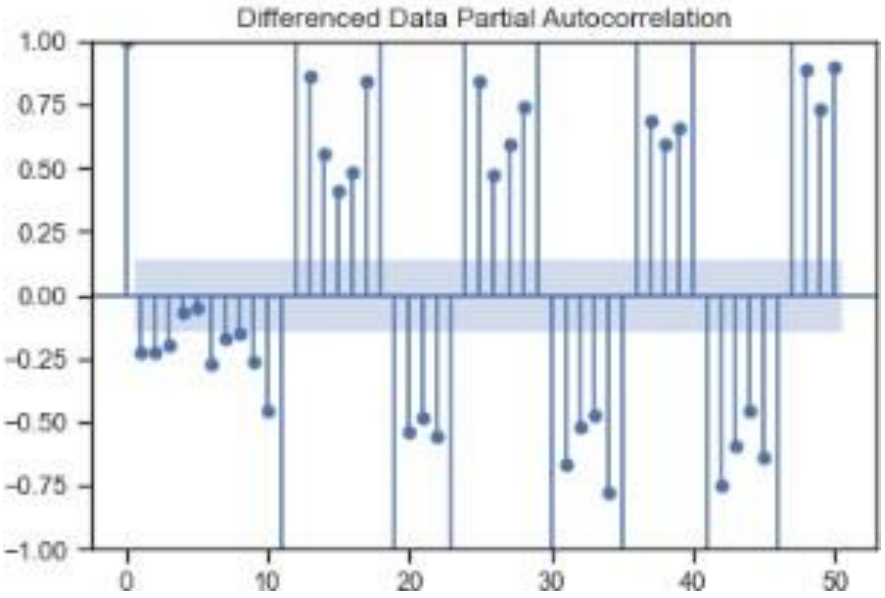
RSME of Model:

528.6069474180102

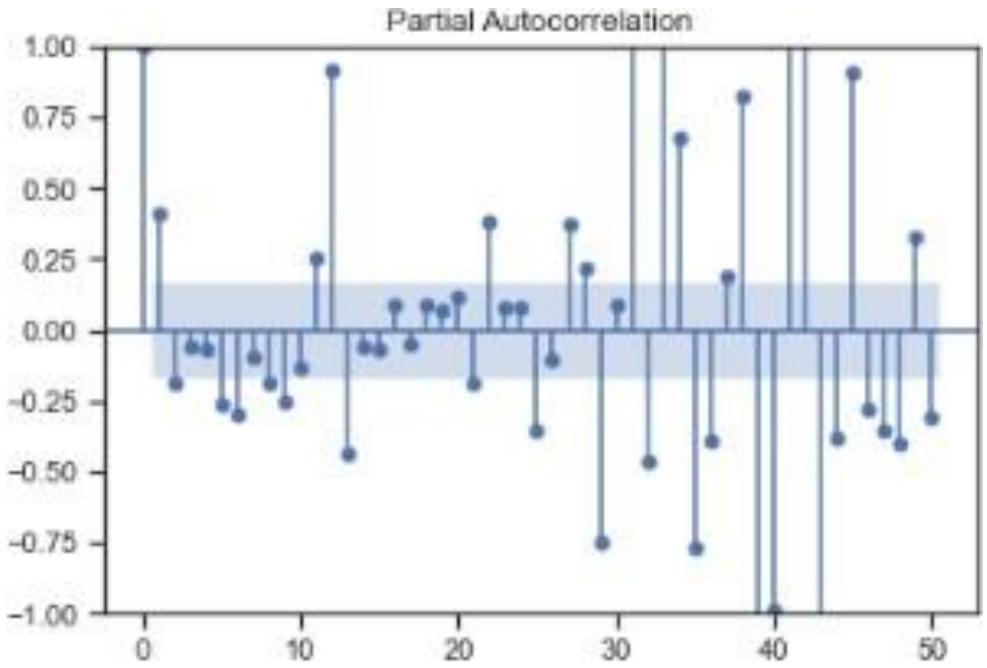
Manual- ARIMA Model

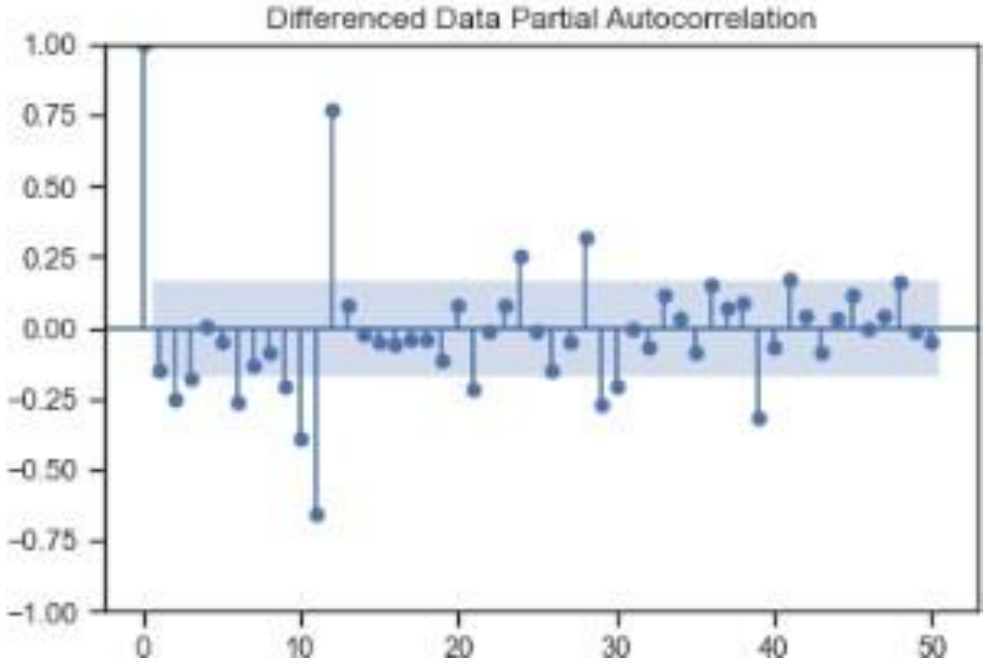
PFB the ACF plot on data





training data with diff(1):





Looking at ACF plot we can see a shard decay after lag 1 for original as well as differenced data.hence we select the q value to be 1. i.e. q=1.

Looking at PACF plot we can again see significant bars till lag 1 for differenced series which is stationary in nature, post 1 the decay is large enough. Hence we choose p value to be 1.
i.e. p=1. d values will be 1, since we had seen earlier that the series is stationary with lag1.

SARIMAX Results

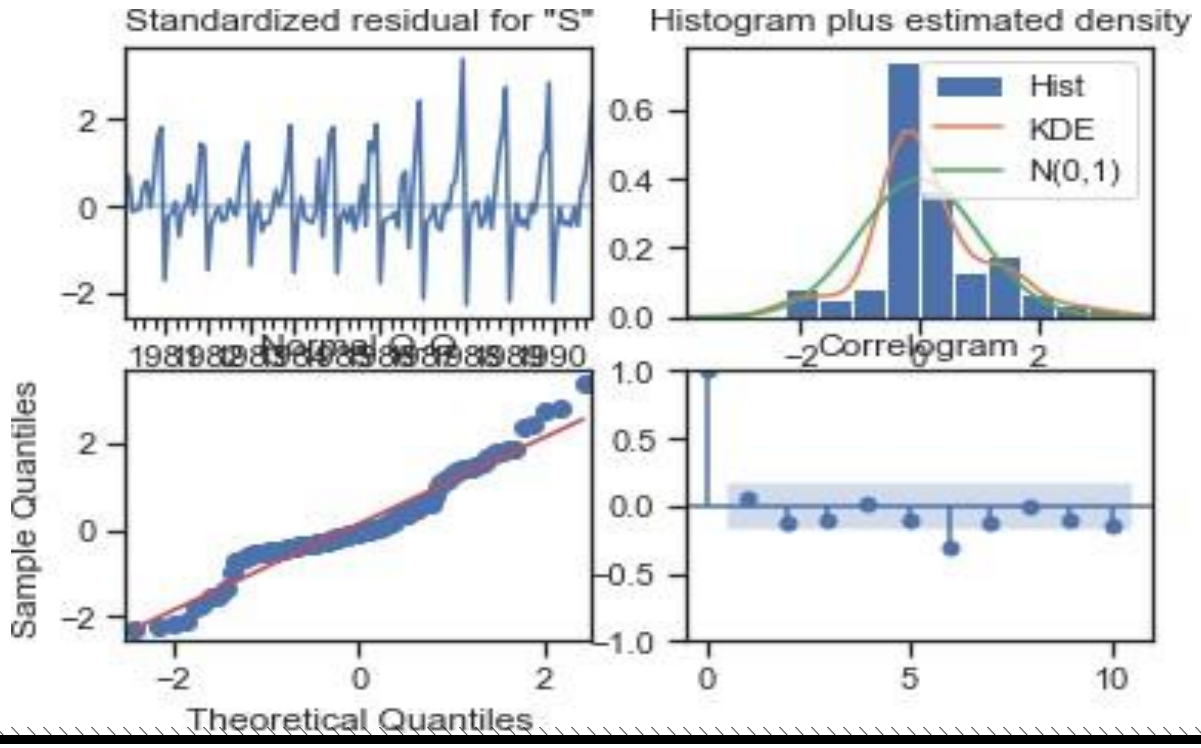
Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1114.878
Date:	Sat, 08 Jul 2023	AIC	2235.755
Time:	08:50:58	BIC	2244.381
Sample:	01-01-1980	HQIC	2239.260
	- 12-01-1990		
Covariance Type:	opg		

	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 2.92e+27. Standard errors may be unstable.

Hence the values selected for manual ARIMA:- p=1, d=1, q=1
summary from this manual ARIMA model.



Model Evaluation: RSME

1319.9367298218867

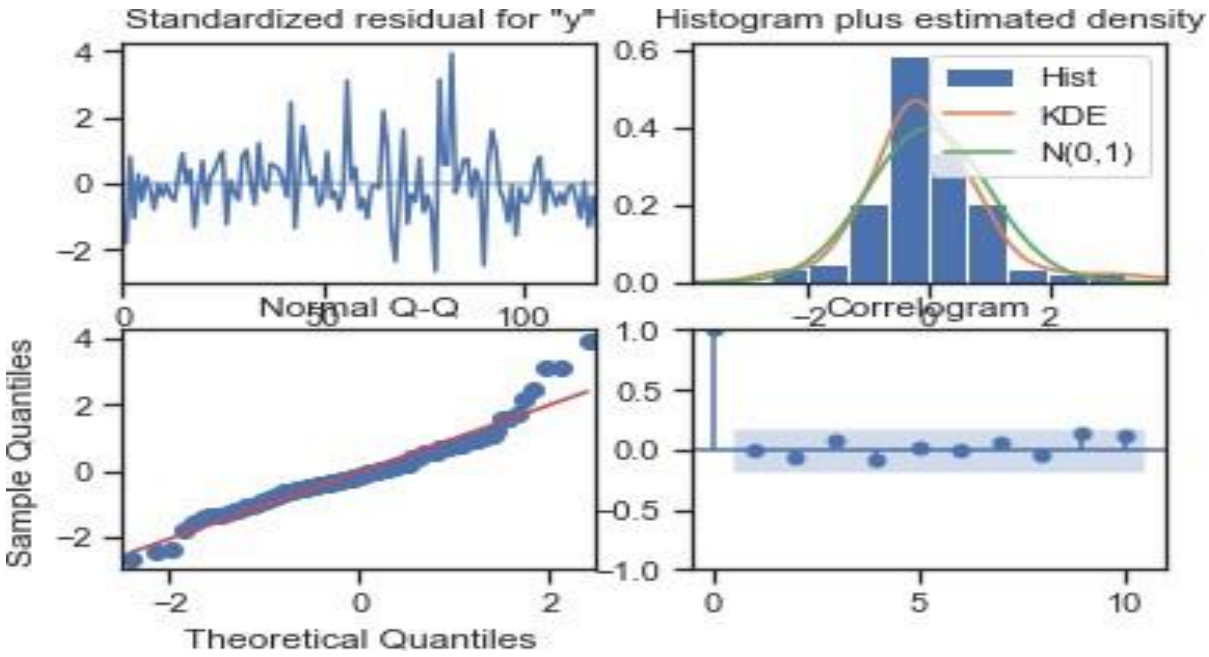
Manual SARIMA Model

SARIMAX(1, 1, 1)x(1, 1, 1, 12)

Below is the summary of the manual SARIMA model

```
SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                -882.088
Date:                            Sat, 08 Jul 2023   AIC                           1774.175
Time:                                    08:56:47   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).
```



Model Evaluation: RSME

359.612454

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

click to scroll output; double click to hide

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponential Smoothing	317.434302
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612454
(1,1,1),(2,0,3,12),Auto_SARIMA	528.606947
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple Average Model	1275.081804
Linear Regression	1275.867052
6pointTrailingMovingAverage	1283.927428
Auto_ARIMA	1299.978401
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponential Smoothing_Auto_Fit	1316.034674
ARIMA(3,1,3)	1319.936730
9pointTrailingMovingAverage	1346.278315
Alpha=0.1, SimpleExponential Smoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponential Smoothing	1778.564670
Naive Model	3864.279352

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.

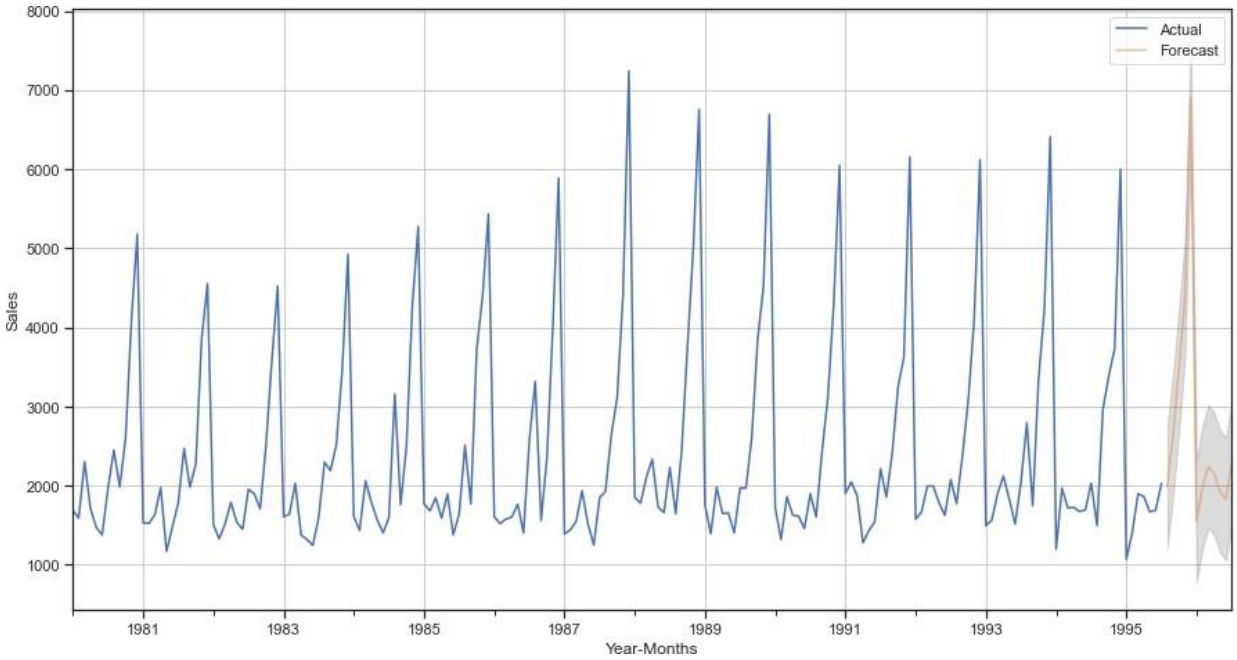
Based on the above comparison of all the various models that we had built, we can conclude that the triple exponential smoothing or the Holts-Winter model is giving us the lowest RMSE, hence it would be the most optimum model

sales predictions made by this best 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	2151.066842
1996-05-01	1929.355815
1996-06-01	1830.619260
1996-07-01	2272.156151

the sales prediction on the graph along with the confidence intervals. PFB the graph.

Plot 27: prediction plot



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

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

Sales Analysis and Recommendations for Sparkling Wine

- **Sales Projection**: The sales of Sparkling wine for the company are predicted to be at least the same as last year, if not higher. There is a potential for peak sales next year to surpass the current year's figures.

- **Consistent Popularity**: Sparkling wine has maintained its popularity among customers over the years, with only a marginal decline in sales. Despite reaching its peak in the late 1980s, it remains a sought-after wine choice for consumers.

- **Seasonality Impact**: Seasonality significantly influences the sales of Sparkling wine. Sales tend to be slower in the first half of the year, picking up from August to December.

- **Marketing Campaign Opportunities**: It is recommended for the company to focus on running marketing campaigns during the first half of the year when sales are typically slower, especially in the months of March to July.

- **Promotional Pairings**: The company could consider pairing Sparkling wine with a less popular wine, such as "Rose wine," under a special offer. This strategy aims to encourage customers to try the underperforming wine and potentially boost its sales. It provides an opportunity for cross-promotion and benefits the company by diversifying its product offerings.

By implementing these recommendations, the company can leverage the consistent popularity of Sparkling wine, optimize sales during the peak season, and strategically target the slower sales periods. The proposed promotional pairings can create additional sales opportunities and broaden customer preferences.

These measures align with the observed sales trends and provide a roadmap for the company to drive sales growth and maximize the potential of Sparkling wine in the market.

Project On Time Series Forecasting

Rose Wine

Yogesh Negi

Problem Statement:

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.

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

Data Dictionary:

Table 1: data dictionary

column	details
YearMonth	Dates of sales
Sparkling	Sales of rose wine

Data set is read using the pandas library.

Rows of dataset;

Table 2: rows of dataset

Top Few Rows :

YearMonth	Rose
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

Last Few Rows :

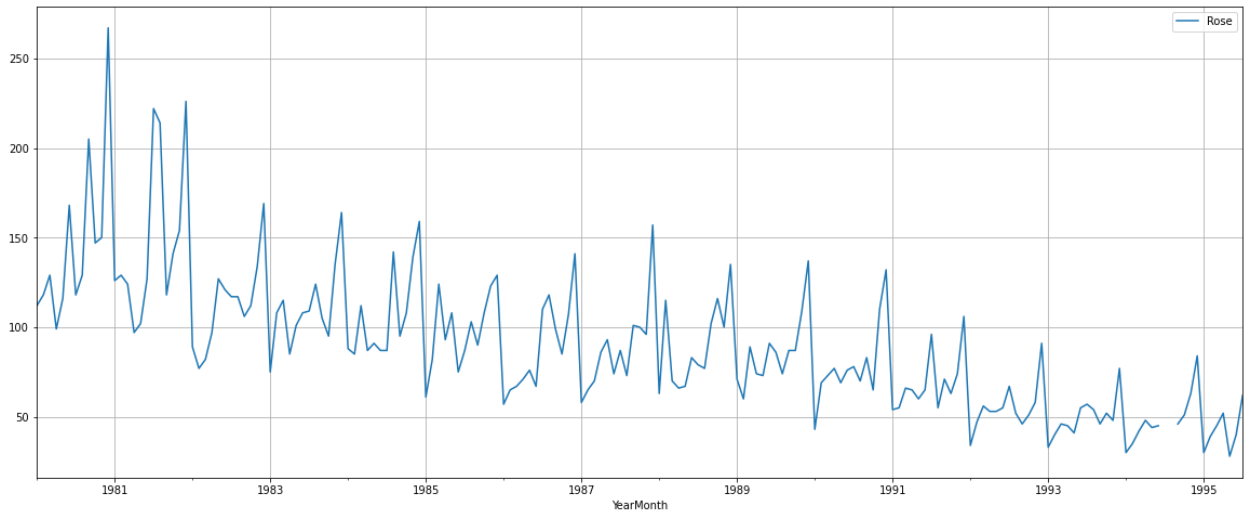
YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Number of Rows and Columns of Dataset:

The dataset has 187 rows and 1 column.

Plot of the dataset:

Plot 1 : dataset



Post Ingestion of Dataset:

We have divided the dataset further by extraction month and year columns from the YearMonth column and renamed the sparkling column name to Sales for better analysis od the dataset.

Rows of new data set;

Table 3: new rows of dataset

Top Few Rows :

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

Last Few Rows :

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

Number of Rows and Columns of Dataset: The dataset has 187 rows and 3 column.

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

Data Type;

Index: DateTime

Sales: integer Month: integer Year: integer

Statistical summary:

Table 4: statistical summary

	count	mean	std	min	25%	50%	75%	max
Sales	185.0	90.0	39.0	28.0	63.0	86.0	112.0	267.0
Year	187.0	1987.0	5.0	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.0	3.0	1.0	3.0	6.0	9.0	12.0

Null Value:

There are 2 null values present in sales the dataset.

We found the values for the months of July & August were missing for the year 1994.

Sales		Year	Month
YearMonth			
1994-07-01	NaN	1994	7
1994-08-01	NaN	1994	8

We tried following approaches to impute the data, these were as below.

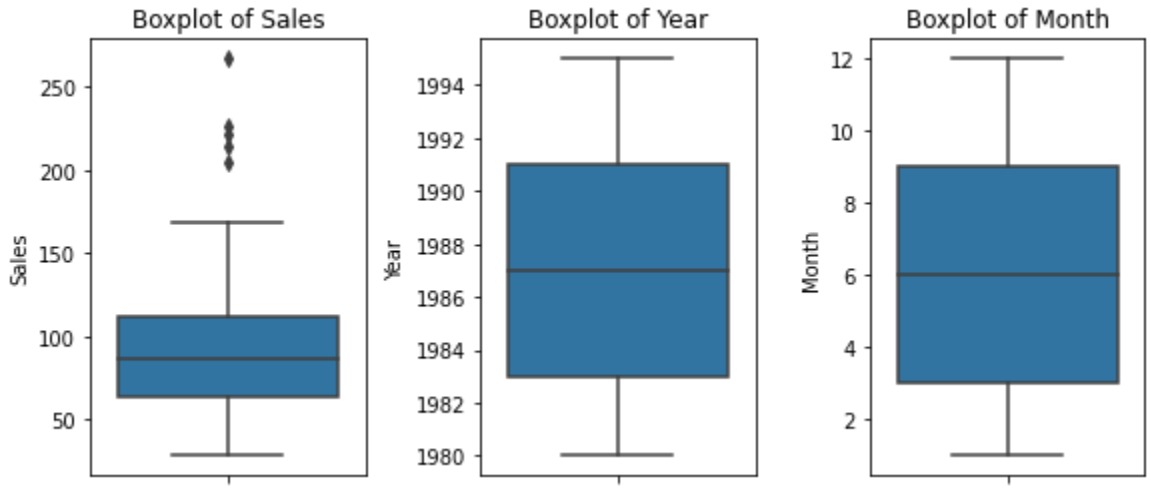
Mean - Before & After

Treating null values is very important to do further analysis.

In this approach, 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.

Boxplot of dataset:

Plot 2: boxplot of data

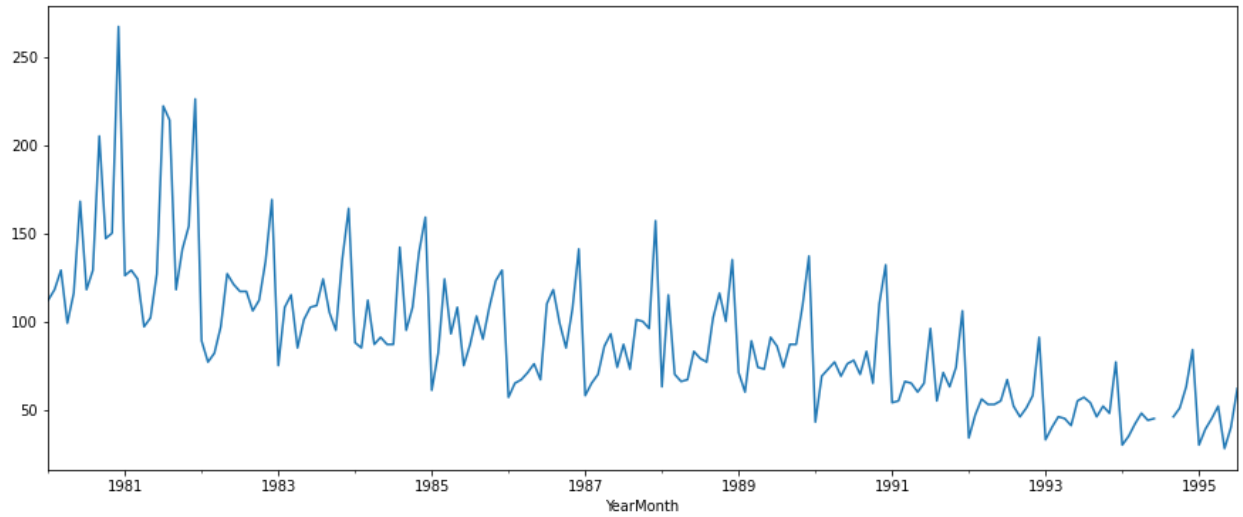


The box plot shows:

- Sales boxplot has outliers we can treat them but we are choosing not to treat them as they do not give much effect on the time series model.

Line plot of sales:

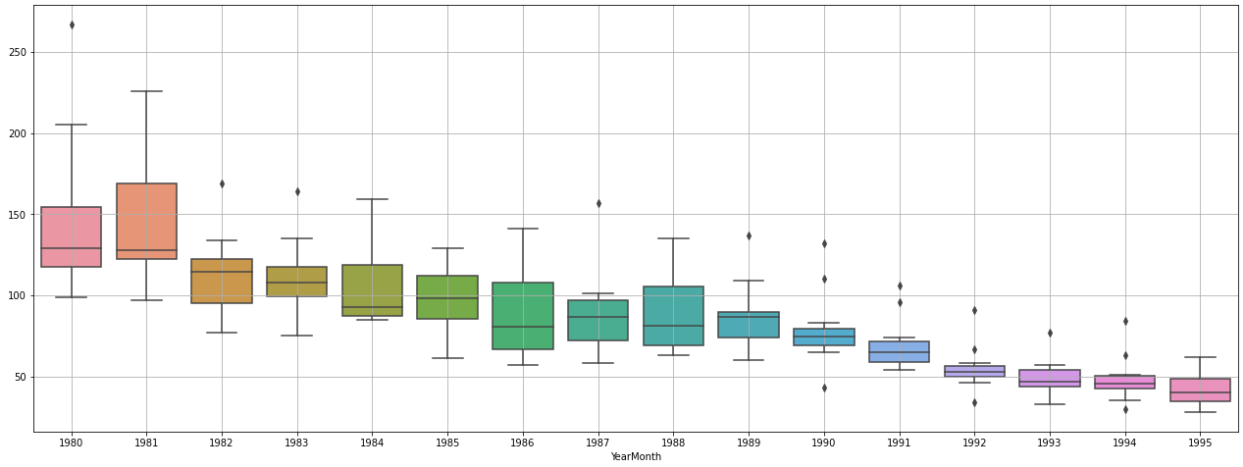
Plot 3: line plot of sales



The line plot shows the patterns of trend and seasonality and also shows that there was a peak in the year 1981.

Boxplot Yearly:

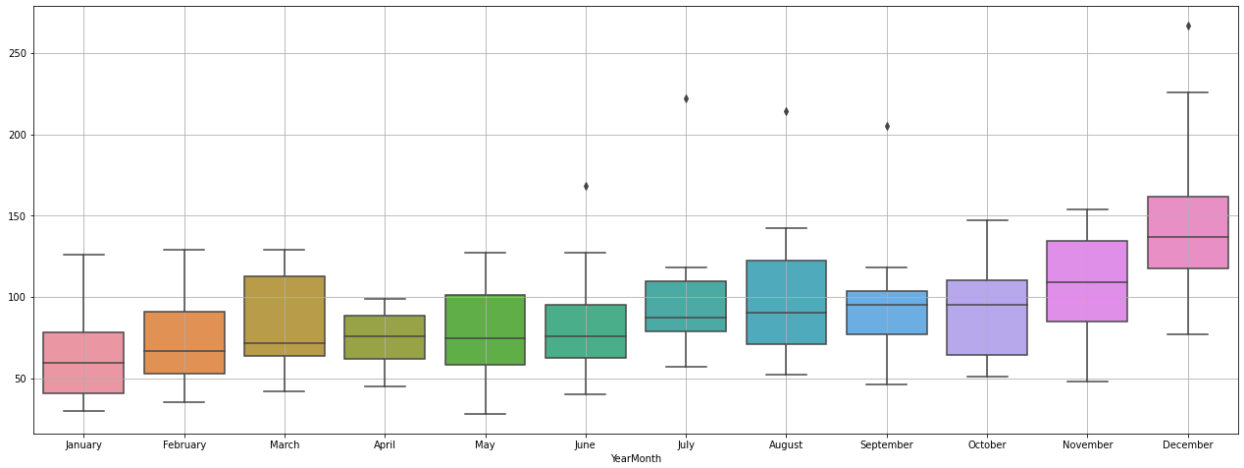
Plot 4: boxplot yearly



This yearly box plot shows there is consistency over the years and there was a peak in 1980-1981. Outliers are present in almost all years.

Boxplot Monthly:

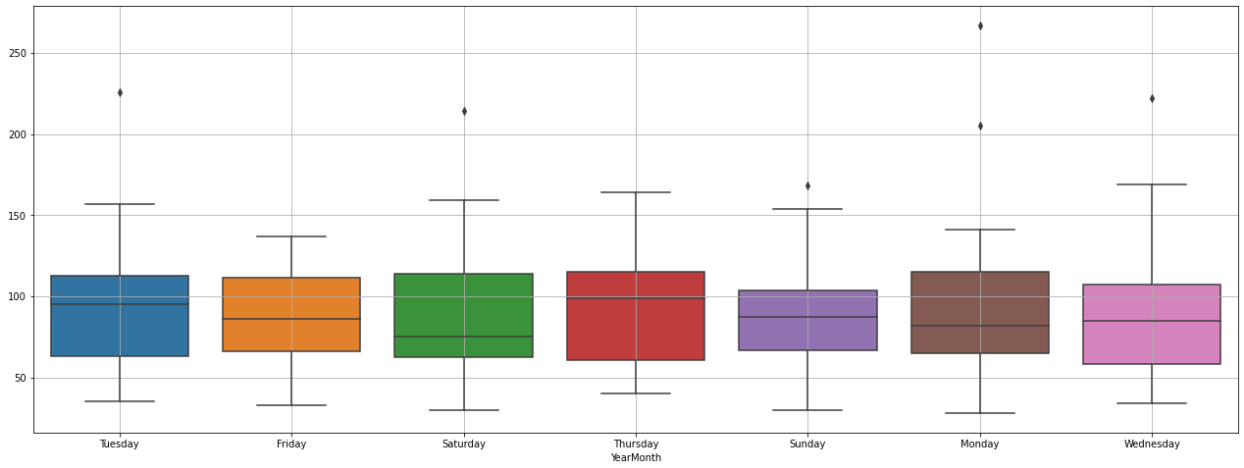
Plot 5: boxplot monthly



The plot shows that 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.

Boxplot Weekdayvise:

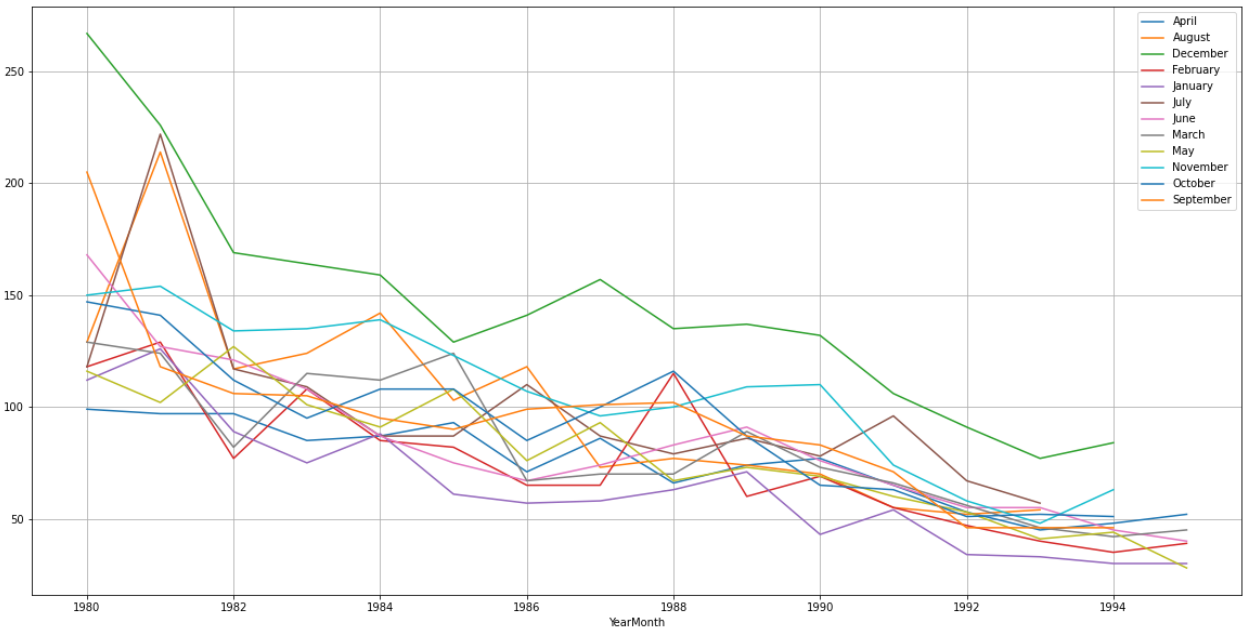
Plot 6: boxplot weekday vise



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.

Graph of Monthly Sales over the years:

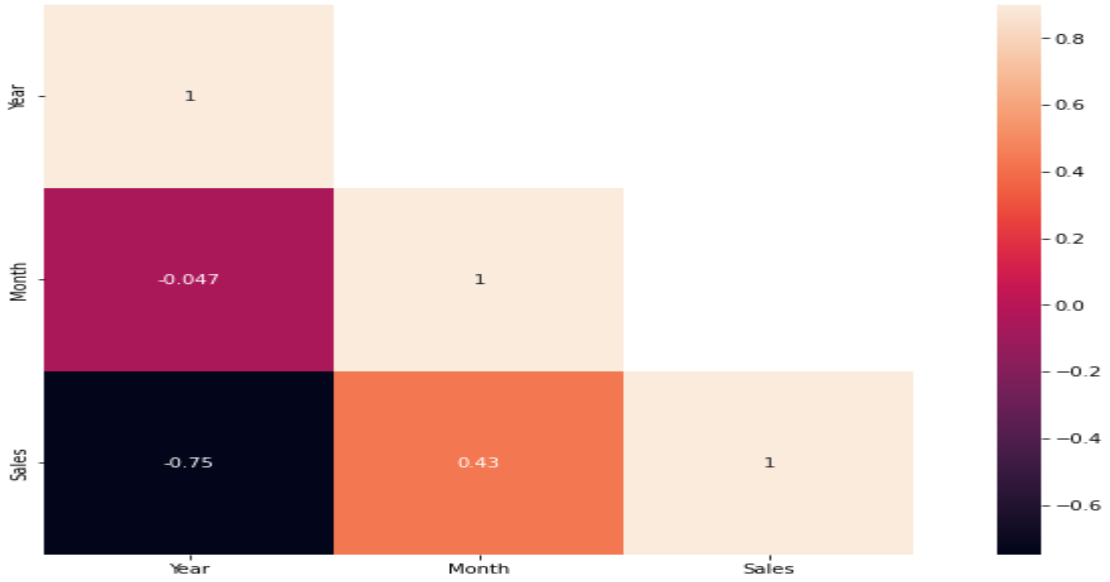
Plot 7: graph of monthly sales over the years



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.

Correlation plot

Plot 8:correlatation plot

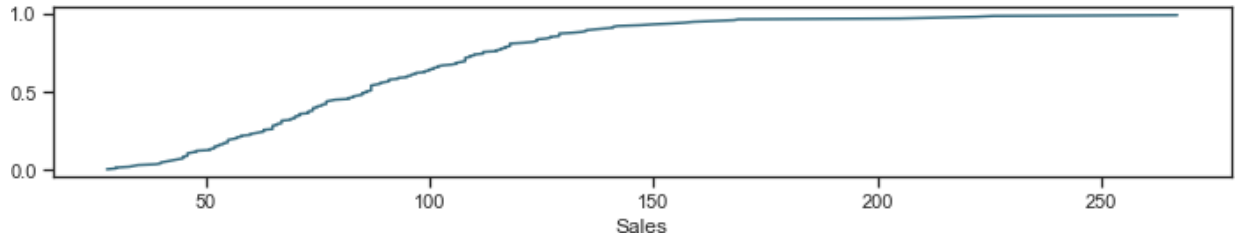


This heat map shows that there was little correlation between Sales and the Years data, there significantly more correlation between the month and Sales columns. Clearly indicating a seasonal pattern in our Sales data. Certain months have higher sales, while certain months have lesser.

Plot ECDF: Empirical Cumulative Distribution Function

This graph shows the distribution of data. Plot

9: ECDF plot

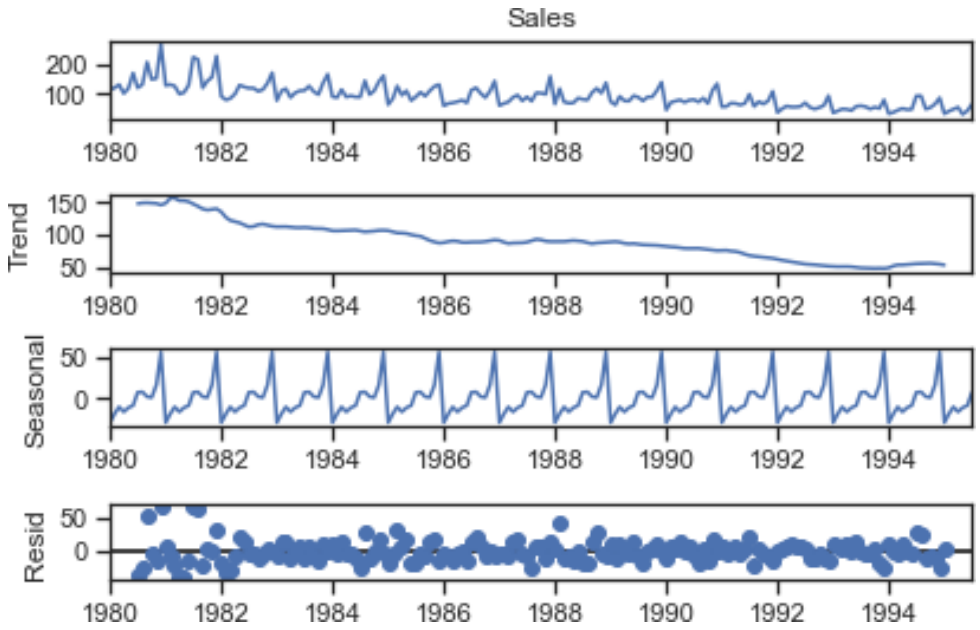


This plot shows:

- 50% sales has been less 100
- Highest vales is 250
- Aprox 90% sales has been less than 150

Decomposition -Additive

Plot 10 : decomposition additive

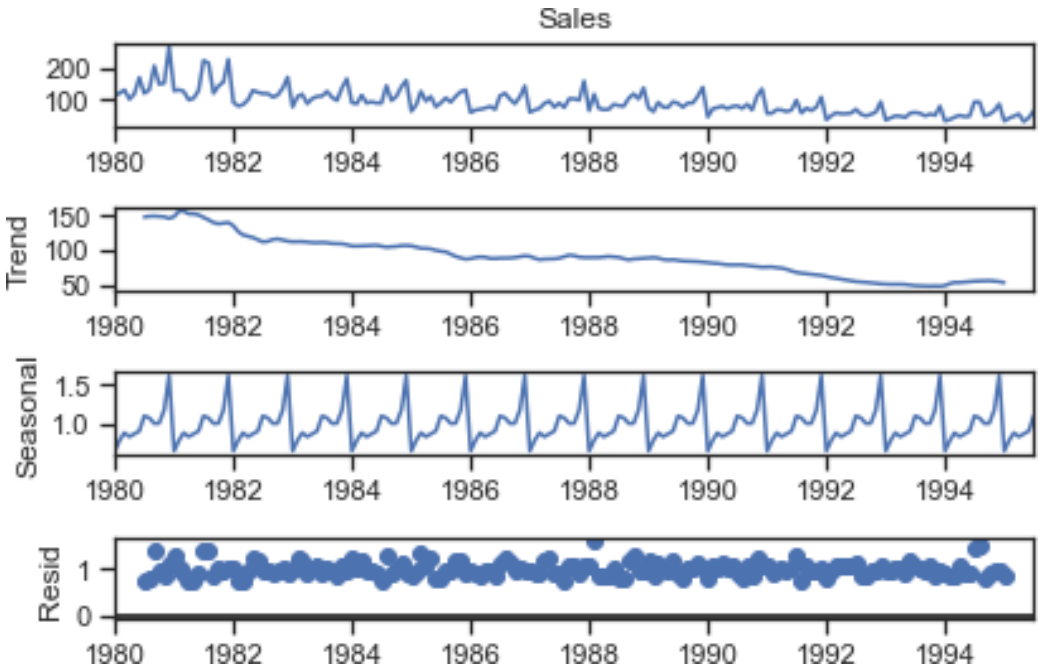


The plots show:

- Peak year 1981
- It also shows that the trend has declined over the year after 1981
- Residue is spread and is not in a straight line.
- Both trend and seasonality are present.

Decomposition-Multiplicative

Plot 11: decomposition multiplicative

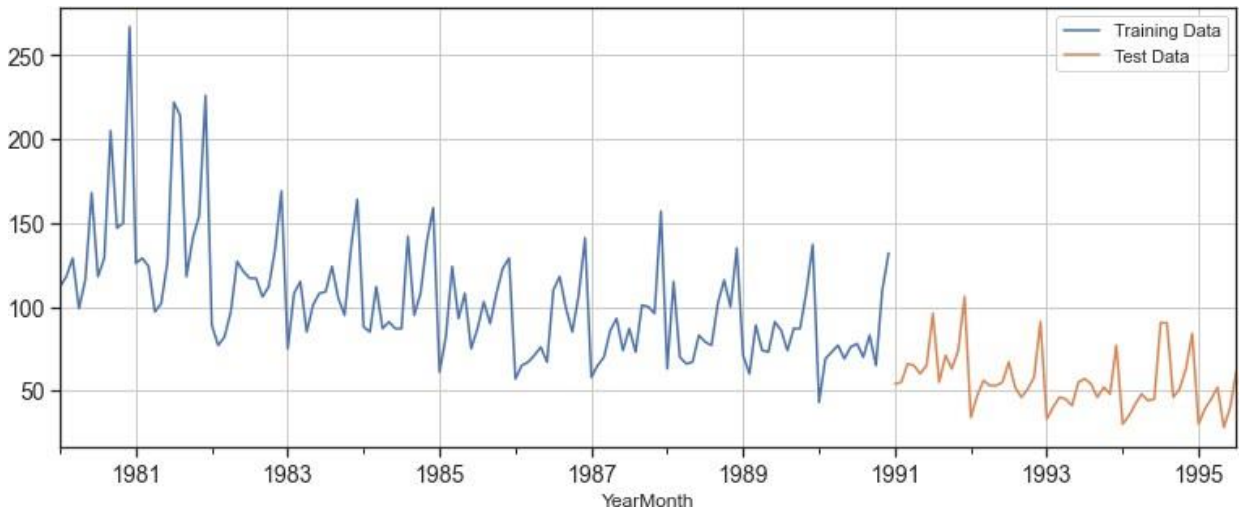


The plots show:

- Peak year 1981
- It also shows that the trend has declined over the year after 1981.
- Residue is spread and is in approx a straight line.
- Both trend and seasonality are present.
- Reside is 0 to 1, while for additive is 0 to 50.
- So multiplicative model is selected owing to a more stable residual plot and lower range of residuals.

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

Plot 12: training and test dataset



Data split from 1980-1990 is training data, then 1991 to 1995 is training data.

Rows and Columns:

train dataset has 132 rows and 3 columns.
test dataset has 55 and 3 columns.

Few Rows of datasets:

Table 5: train and test dataset rows

Train dataset	Test dataset

First few rows of Training Data

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

Last few rows of Training Data

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

First few rows of Test Data

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

Last few rows of Test Data

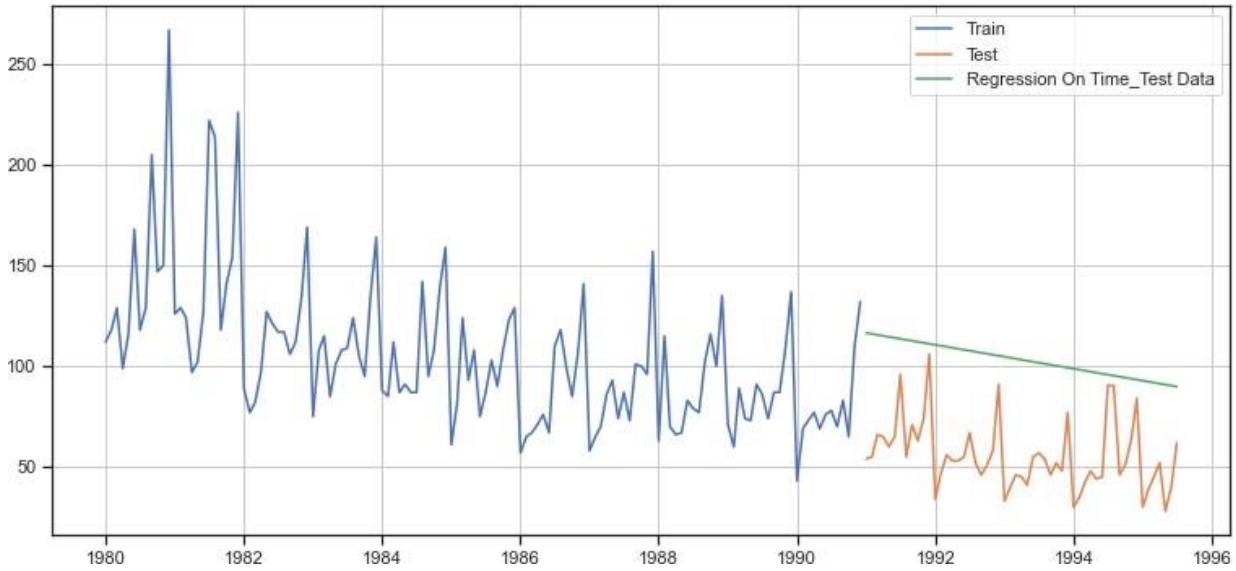
YearMonth	Sales	Year	Month
1995-03-01	45.0	1995	3
1995-04-01	52.0	1995	4
1995-05-01	28.0	1995	5
1995-06-01	40.0	1995	6
1995-07-01	62.0	1995	7

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, and simple 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

Plot 13: linear regression



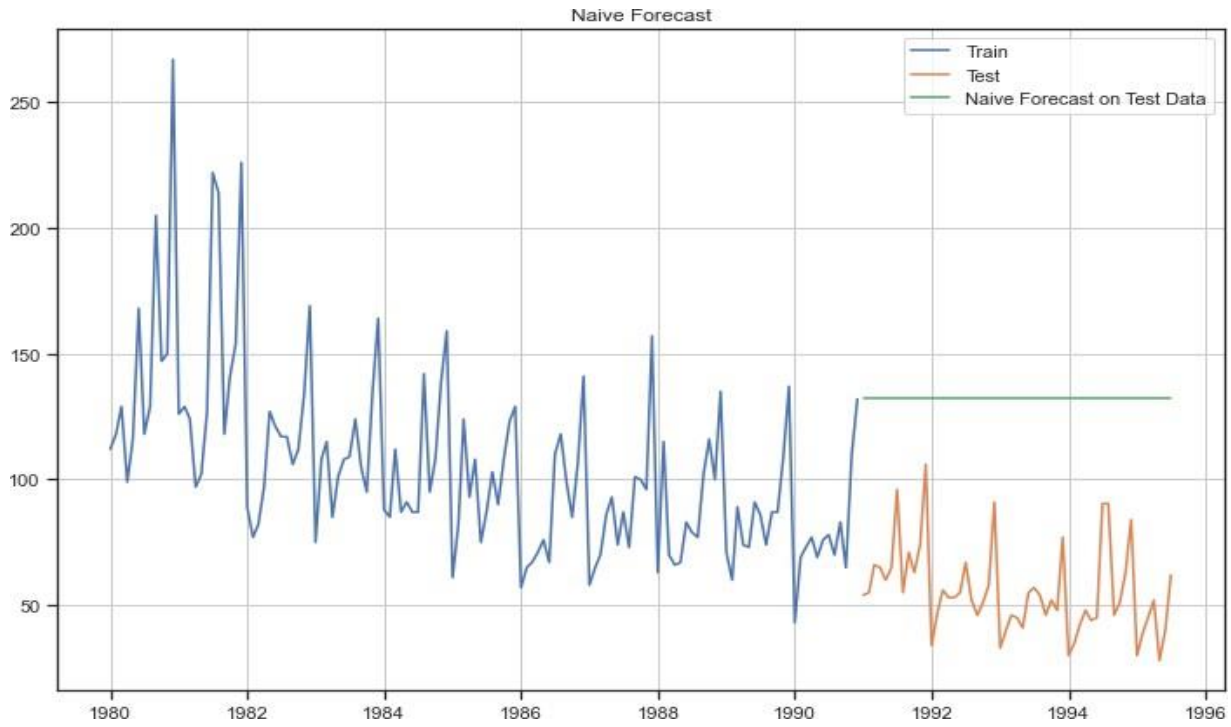
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Linear Regression	51.080941

Model 2: Naive Approach:

Plot 14: naive approach



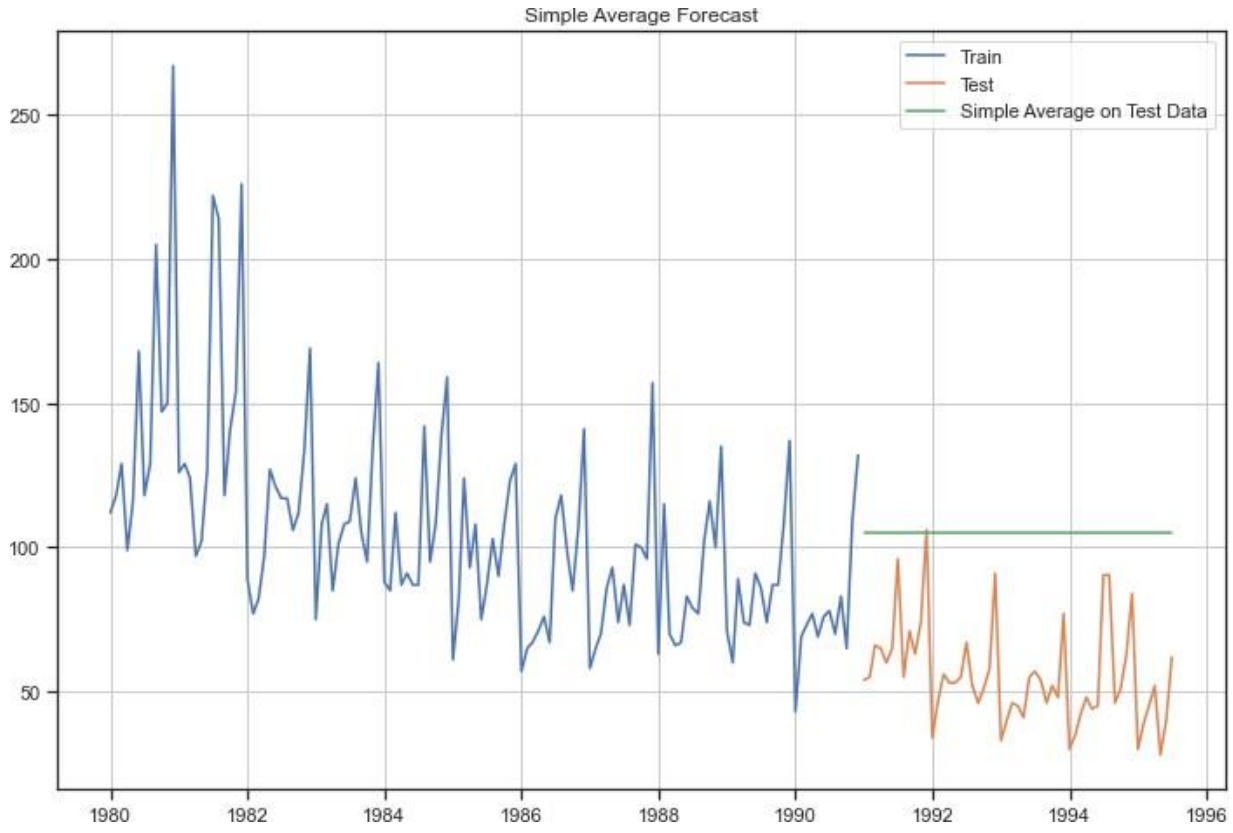
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Naive Model 79.304391

Model 3: Simple Average

Plot 15: simple average



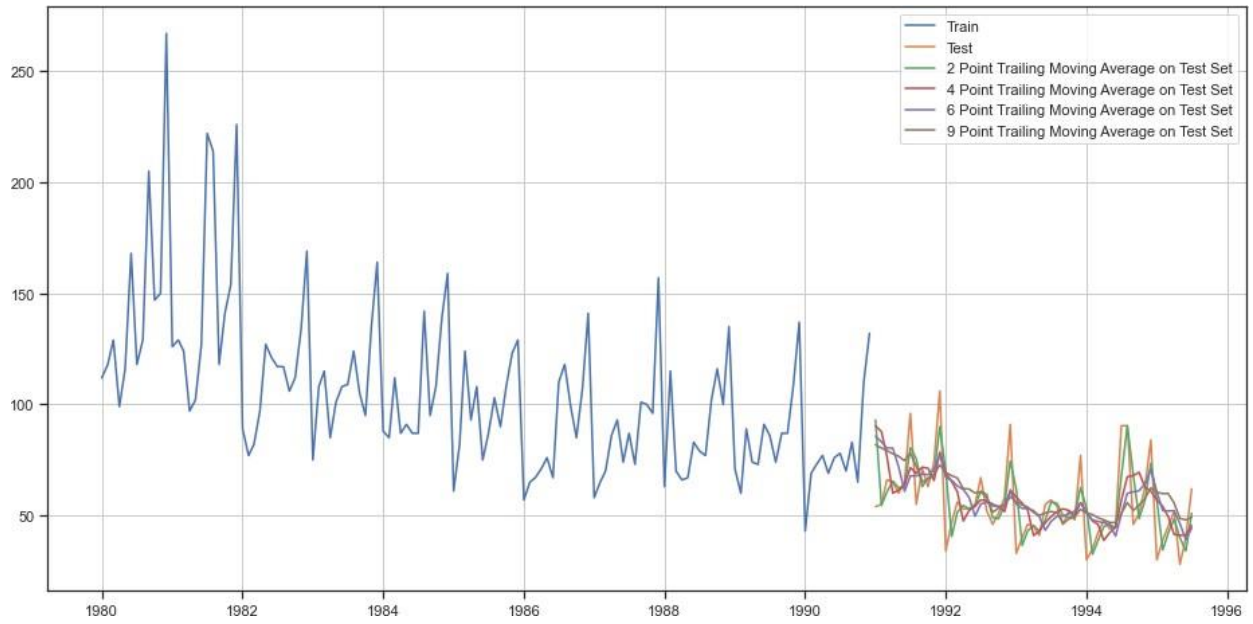
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Simple Average Model	53.049755

Model 4: Moving Average

Plot 16: moving average



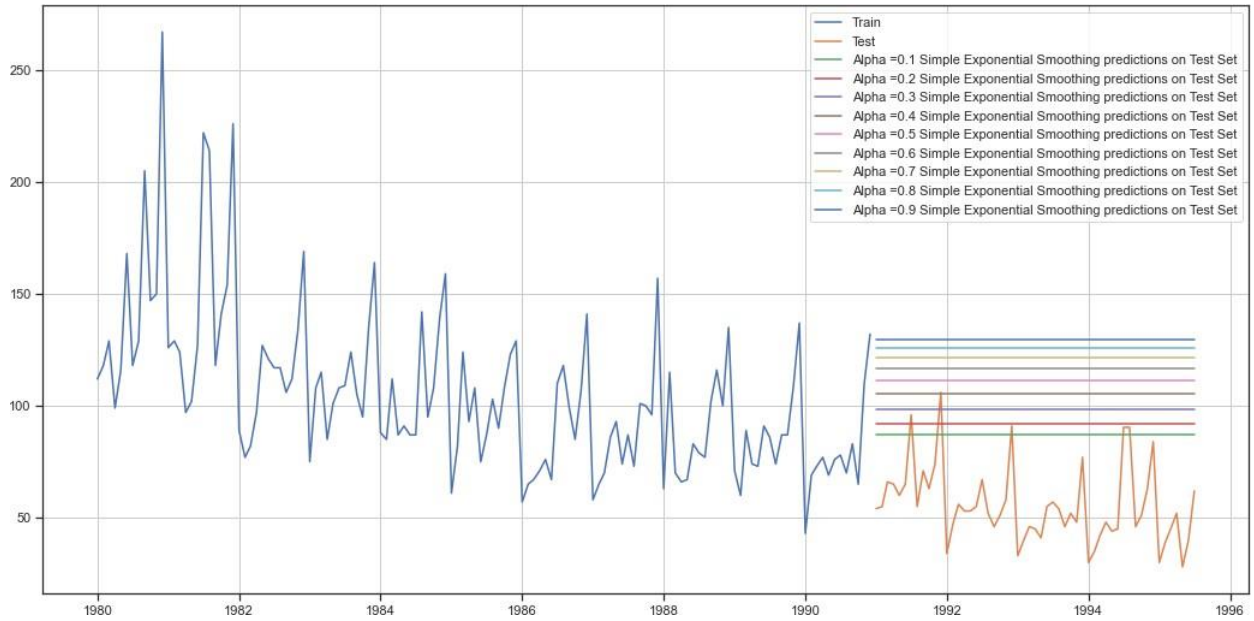
Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

2	pointTrailingMovingAverage	11.589082
4	pointTrailingMovingAverage	14.506190
6	pointTrailingMovingAverage	14.558008
9	pointTrailingMovingAverage	14.797139

We created multiple moving average models with rolling windows varying from 2 to 9. Rolling average is a better method than simple average as it takes into account only the previous n values to make the prediction, where n is the rolling window defined. This takes into account the recent trends and is in general more accurate. Higher the rolling window, smoother will be its curve, since more values are being taken into account.

Model 5: Simple Exponential Smoothing

Plot 17: simple exponential smoothing

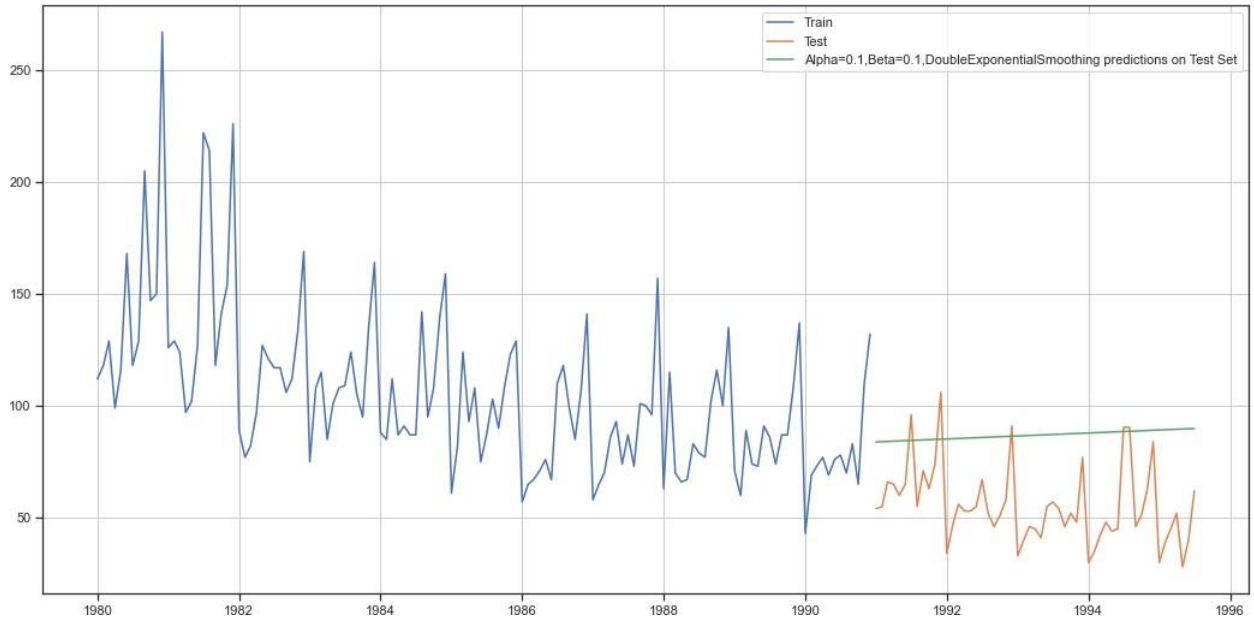


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.1,SimpleExponentialSmoothing	36.429535

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

Plot 18: double exponential smoothing

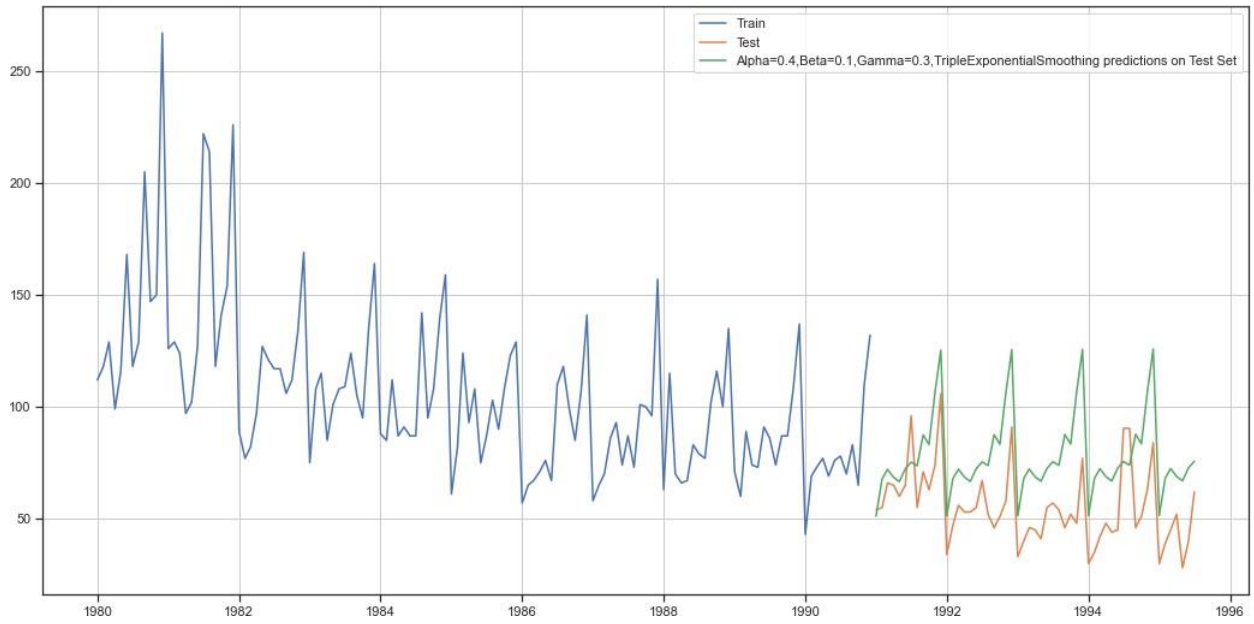


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing36.510010

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

Plot19 : plot triple exponential smoothing



Output for best alpha, beta and gamma values is shown by the green color line in the above plot. Best model had both multiplicative trend as well as seasonality.

So far this is the best model

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing 8.992350

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

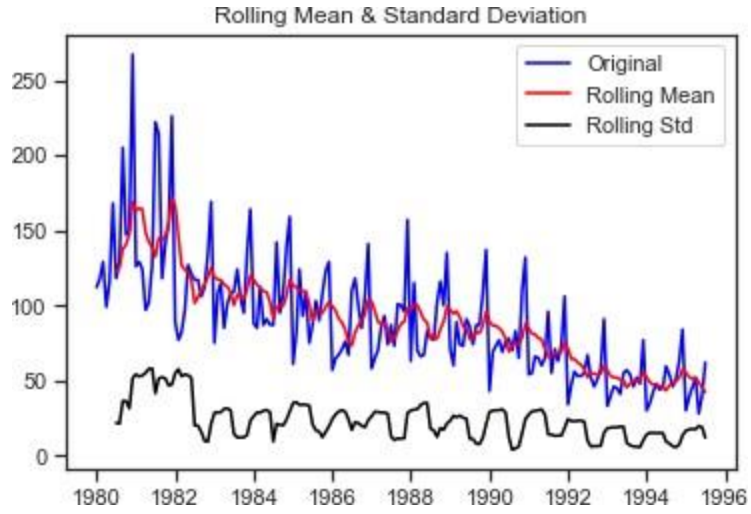
The hypothesis in a simple form for the ADF test is:

- H0 : 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 the α value.

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

20: dickey fuller test



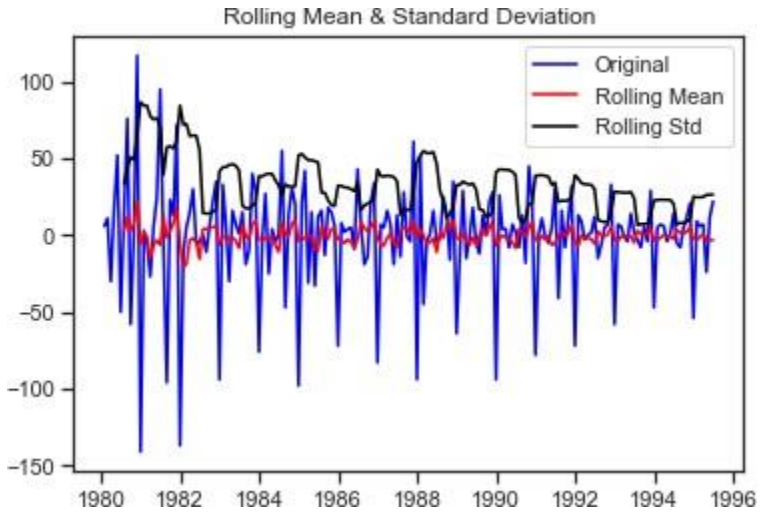
Results of Dickey-Fuller Test:

Test Statistic	-1.892338
p-value	0.335674

we failed to reject the null hypothesis, which implies the Series is not stationary in nature. In

order to try and make the series stationary we used the differencing approach. We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped

Plot 21: dickey fuller test after diff



Results of Dickey-Fuller Test:

Test Statistic	-8.032729e+00
p-value	1.938803e-12

the null hypothesis that the series is not stationary at difference = 1 was rejected, which implied that the series has indeed become stationary after we performed the differencing.

We could now proceed ahead with ARIMA/ SARIMA models, since we had made the series stationary.

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

We employed a for loop for determining the optimum values of p,d,q, where p is the order of the AR (Auto-Regressive) part of the model, while q is the order of the MA (Moving Average) part of the model. d is the differencing that is required to make the series stationary. p,q values in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for d, since we had already determined d to be 1, while checking for stationarity using the ADF test.

Some parameter combinations for the Model...

Model: (0, 1, 1)

Model: (0, 1, 2)

Model: (0, 1, 3)

Model: (1, 1, 0)

Model: (1, 1, 1)

Model: (1, 1, 2)

Model: (1, 1, 3)

Model: (2, 1, 0)

Model: (2, 1, 1)

Model: (2, 1, 2)

Model: (2, 1, 3)

Model: (3, 1, 0)

Model: (3, 1, 1)

Model: (3, 1, 2)
Model: (3, 1, 3)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected.

	param	AIC
11	(2, 1, 3)	1274.695273
15	(3, 1, 3)	1278.658803
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.57423
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

the summary report for the ARIMA model with values (p=2,d=1,q=3).

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.348
Date:	Sat, 08 Jul 2023	AIC	1274.695
Time:	09:32:32	BIC	1291.946
Sample:	01-01-1980	HQIC	1281.705
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6780	0.084	-20.018	0.000	-1.842	-1.514
ar.L2	-0.7288	0.084	-8.694	0.000	-0.893	-0.564
ma.L1	1.0448	0.649	1.609	0.108	-0.228	2.317
ma.L2	-0.7717	0.134	-5.751	0.000	-1.035	-0.509
ma.L3	-0.9045	0.589	-1.536	0.125	-2.059	0.250
sigma2	859.1032	546.561	1.572	0.116	-212.137	1930.343

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.45
Prob(Q):	0.88	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).

RMSE values are as below:

36.41939808200016

AUTO- SARIMA Model

A similar for loop like AUTO_ARIMA with below values was employed, resulting in the models shown below.

p = q = range(0, 4) d= range(0,2) D = range(0,2) pdq = list(itertools.product(p, d, q))
model_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, D, q))]

Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected. Here only the top 5 models are shown.

	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400287
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)	775.49533
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018

the summary report for the best SARIMA model with values (3,1,1)(3,0,2,12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-377.200
Date:	Sat, 08 Jul 2023	AIC	774.400
Time:	09:45:58	BIC	799.618
Sample:	0	HQIC	784.578
	- 132		
Covariance Type:	opg		

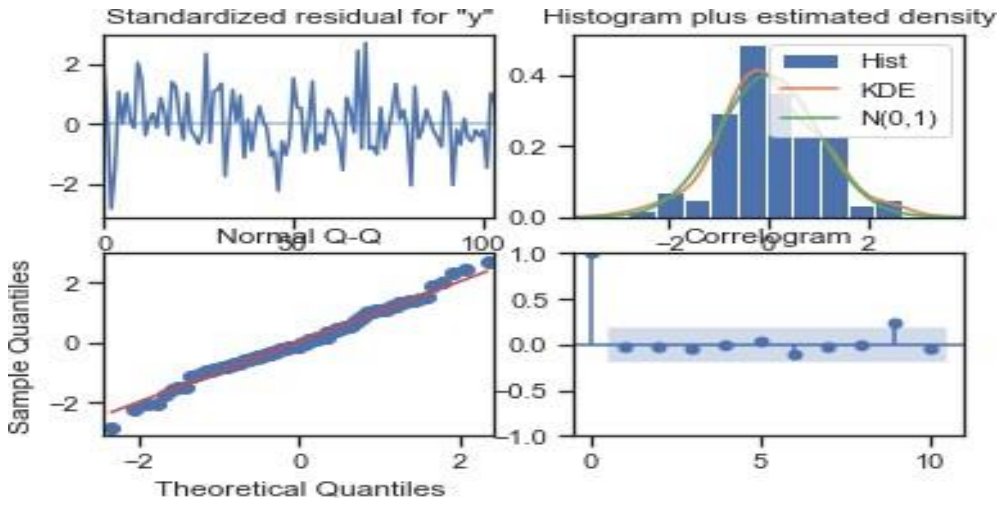
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0464	0.126	0.367	0.714	-0.201	0.294
ar.L2	-0.0060	0.120	-0.050	0.960	-0.241	0.229
ar.L3	-0.1808	0.098	-1.837	0.066	-0.374	0.012
ma.L1	-0.9370	0.067	-13.905	0.000	-1.069	-0.805
ar.S.L12	0.7639	0.165	4.639	0.000	0.441	1.087
ar.S.L24	0.0840	0.159	0.527	0.598	-0.229	0.397
ar.S.L36	0.0727	0.095	0.764	0.445	-0.114	0.259
ma.S.L12	-0.4968	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	192.1578	39.629	4.849	0.000	114.486	269.830

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).

We also plotted the graphs for the residual to determine if any further information can be extracted or all the usable information has already been extracted. Below were the plots for the best auto SARIMA model.

Plot 22: sarima plots



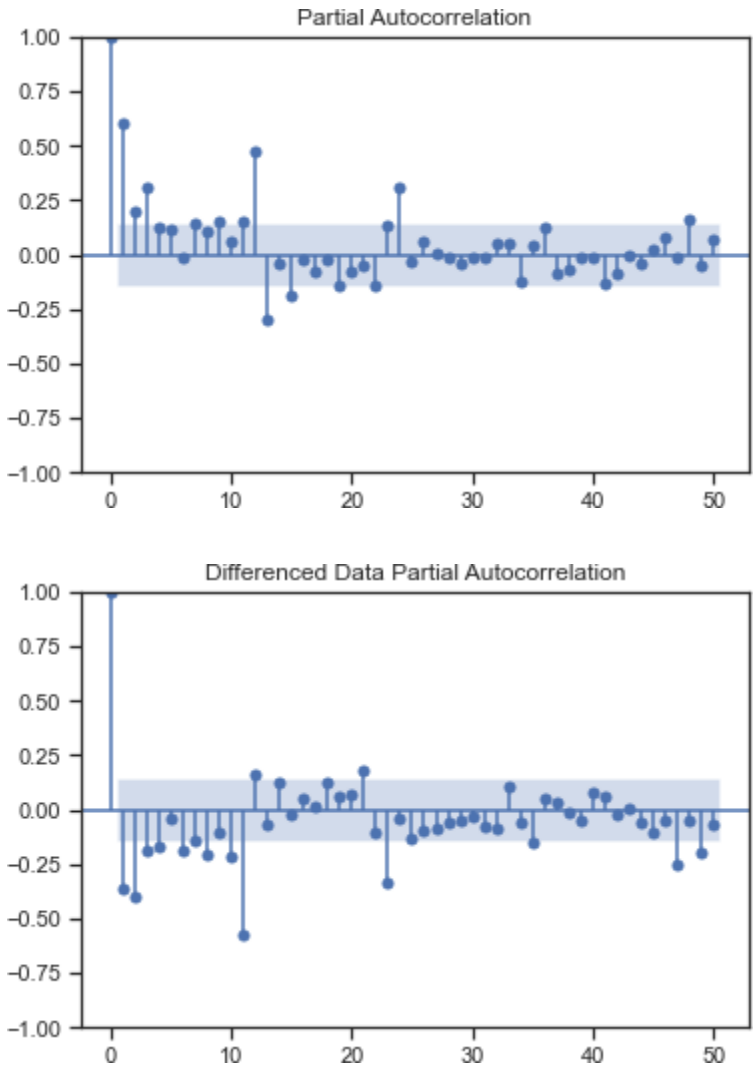
RSME of Model:

18.53550451433853

Manual- ARIMA Model

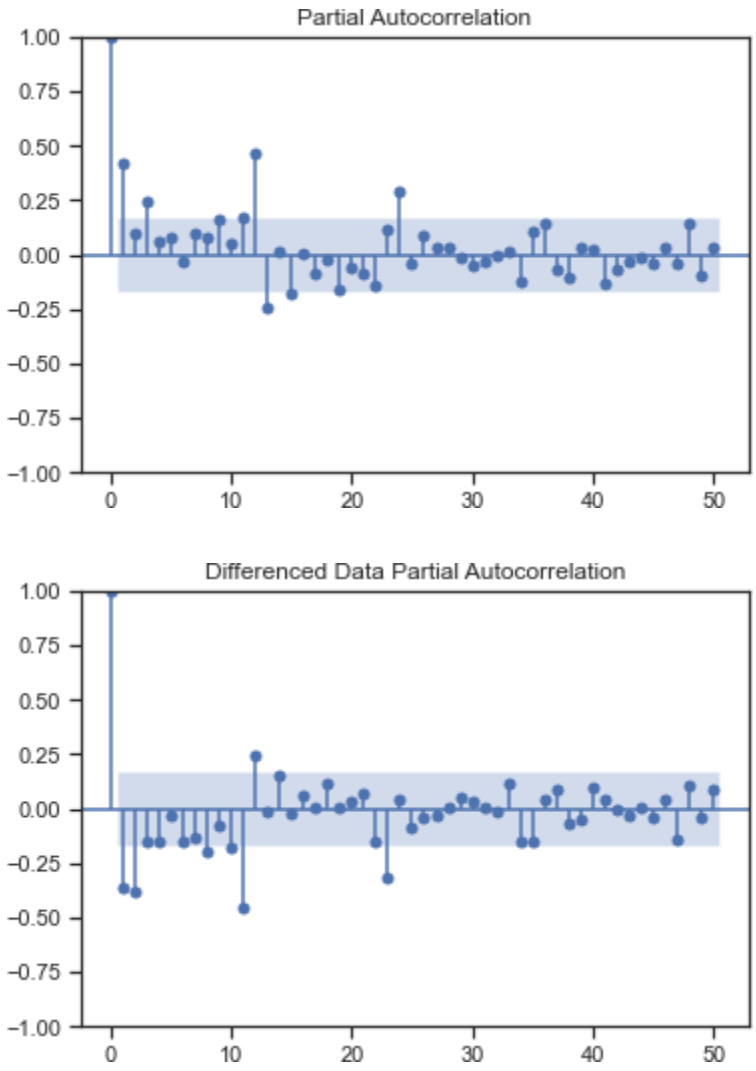
PACF the ACF plot on data :

Plot 23: PACF and ACF plots



Following is plotting the PACF and ACFgraph for the training data.

Plot 24: PACF and ACF plota of train date



Hence the values selected for manual ARIMA:- p=2, d=1, q=2 summary from this manual ARIMA model.

SARIMAX Results

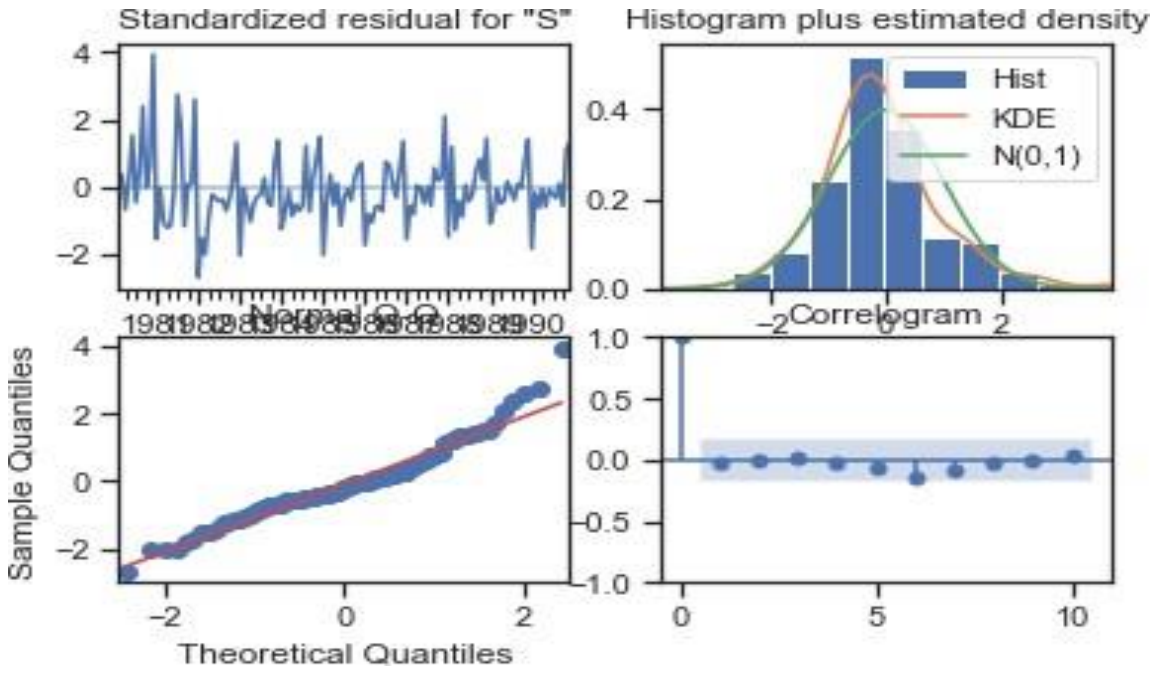
Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Sat, 08 Jul 2023	AIC	1281.871
Time:	11:06:38	BIC	1296.247
Sample:	01-01-1980 - 12-01-1990	HQIC	1287.712
Covariance Type:	opg		

	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).

Plot 25: manual arima model plots



Model Evaluation: RSME

RMSE: 36.47322487814613

Manual SARIMA Model

Looking at the ACF and PACF plots for training data, we can clearly see significant spikes at lags 12,24,36,48 etc, indicating a seasonality of 12. The parameters used for manual SARIMA model are as below.

SARIMAX(2, 1, 2)x(2, 1, 2, 12)

Below is the summary of the manual SARIMA model

SARIMAX Results

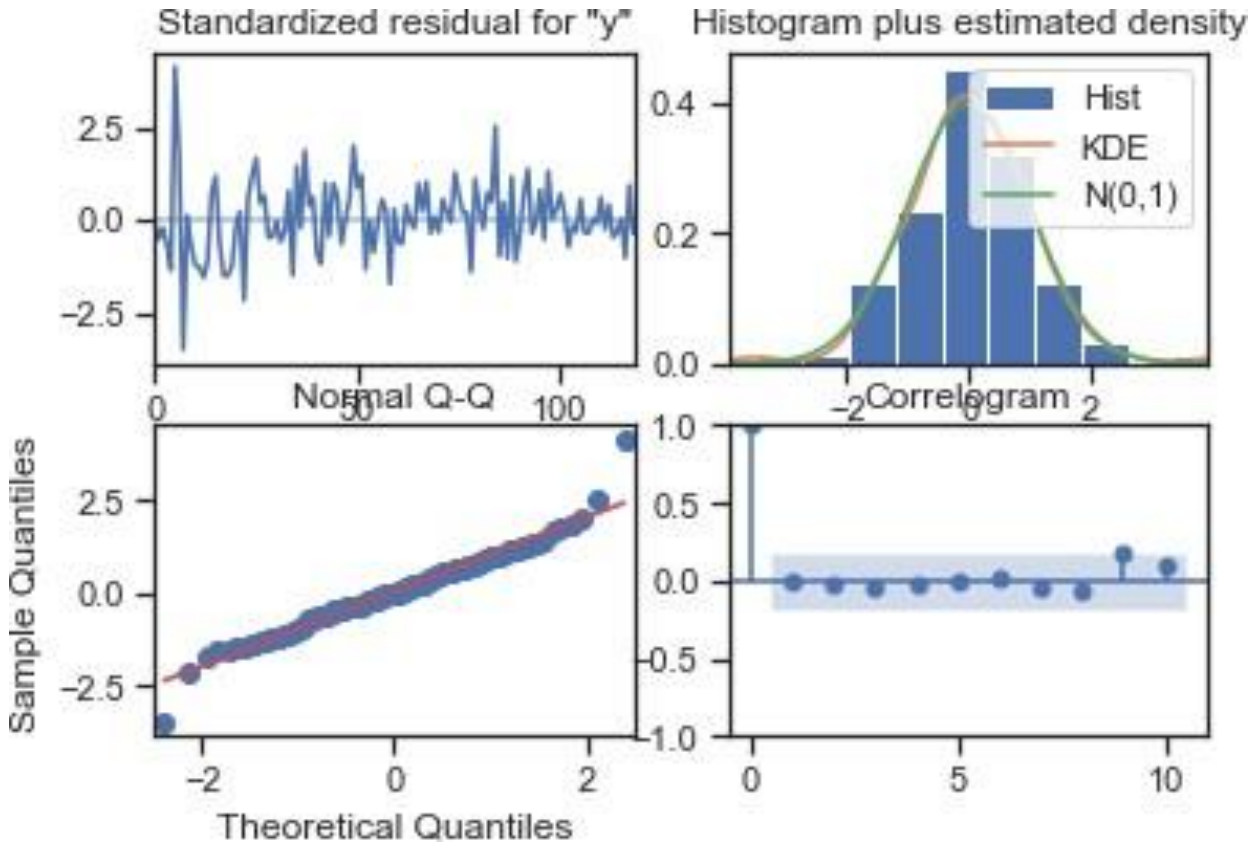
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-538.016
Date:	Sat, 08 Jul 2023	AIC	1094.031
Time:	11:06:49	BIC	1119.044
Sample:	0	HQIC	1104.188
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5492	0.228	-2.409	0.016	-0.996	-0.102
ar.L2	-0.0744	0.099	-0.753	0.451	-0.268	0.119
ma.L1	-0.1703	0.216	-0.787	0.431	-0.594	0.254
ma.L2	-0.6694	0.228	-2.937	0.003	-1.116	-0.223
ar.S.L12	-1.0135	0.524	-1.936	0.053	-2.040	0.013
ar.S.L24	-0.1003	0.175	-0.572	0.567	-0.444	0.243
ma.S.L12	0.2908	25.246	0.012	0.991	-49.190	49.771
ma.S.L24	-0.7076	17.972	-0.039	0.969	-35.932	34.517
sigma2	430.4507	1.07e+04	0.040	0.968	-2.05e+04	2.13e+04

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	27.15
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).

Plot 26: manula sarima plots



Model Evaluation: RSME

14.975746352466942

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

	Test RMSE
Alpha=0.2,Beta=0.7,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	14.975041
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535028
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397777
Auto_ARIMA	36.420791
Alpha=0.1,SimpleExponentialSmoothing	36.429535
ARIMA(3,1,3)	36.473225
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Linear Regression	51.080941
Simple Average Model	53.049755
Naive Model	79.304391

We can clearly see that triple exponential smoothing model with alpha 0.1, beta 0.7 and gamma 0.2 is the best as it he the lowest RSME score.

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.

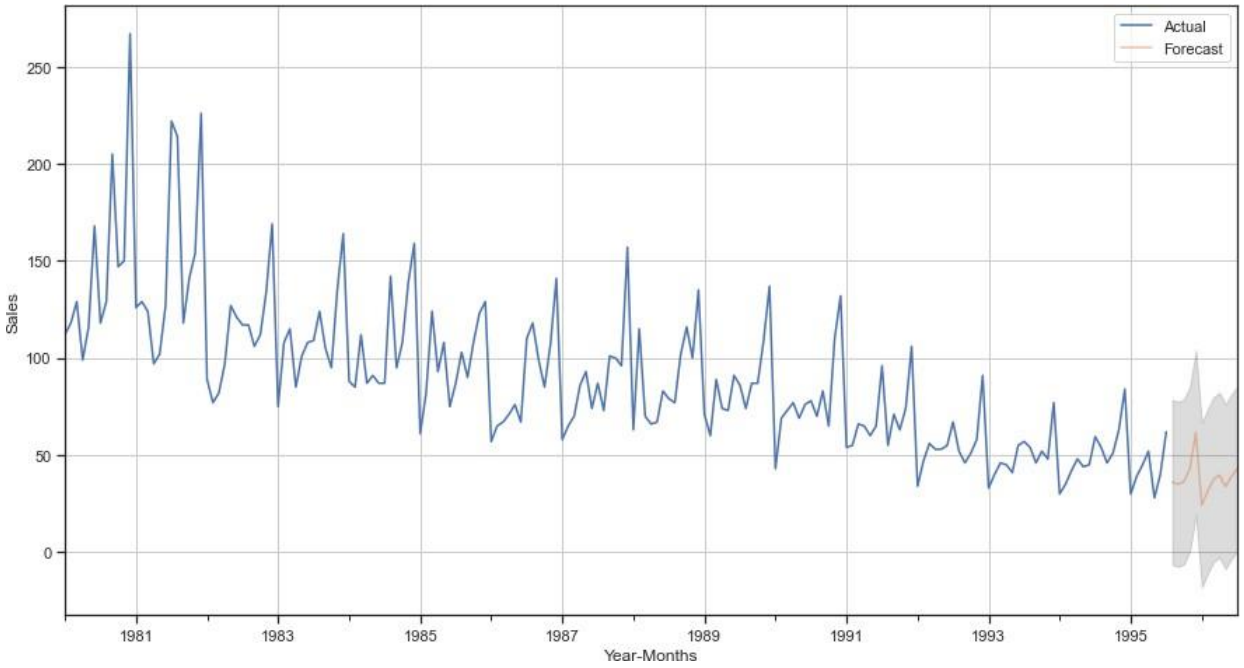
Based on the above comparison of all the various models that we had built, we can conclude that the triple exponential smoothing or the Holts-Winter model is giving us the lowest RMSE, hence it would be the most optimum model

sales predictions made by this best optimum model.

	Sales_Predictions
1995-08-01	36.096841
1995-09-01	34.999961
1995-10-01	36.289937
1995-11-01	43.126839
1995-12-01	61.593978
1996-01-01	24.293852
1996-02-01	31.406019
1996-03-01	37.545514
1996-04-01	39.735393
1996-05-01	33.753457
1996-06-01	38.868148
1996-07-01	43.093112

the sales prediction on the graph along with the confidence intervals. PFB the graph.

Plot 27: prediction plot



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

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

Analysis and Recommendations for Rose Wine Sales

- **Declining Popularity**: Analysis of the wine sales data reveals a clear downward trend for the Rose wine variety for the company. The popularity of Rose wine has been declining for over a decade.
- **Future Trend Prediction**: Based on the predictions of the most optimal model, this downward trend is expected to continue in the future.
- **Seasonal Influence**: Wine sales are significantly influenced by seasonal changes. Sales tend to increase during the festival season, while they decline during peak winter months, particularly in January.
- **Campaign Opportunities**: To counteract the subdued sales during the rest of the year, the company should consider running campaigns to boost the consumption of Rose wine. Specifically, campaigns during the lean period from April to June are likely to yield maximum results, as sales are typically low during this period.
- **Peak Period Considerations**: Running campaigns during peak periods, such as festivals, may not generate a significant impact on sales, as they are already high during this time of the year. Therefore, it is advisable to focus marketing efforts on other periods.
- **Avoiding Peak Winter Campaigns**: It is not recommended to run campaigns during peak winter time, especially in January, as people are less likely to purchase wine due to climatic reasons. Campaigns during this period may not effectively change people's purchasing behavior.
- **Revamp Strategies**: The company should also consider exploring the reasons behind the decline in the popularity of Rose wine. It may be necessary to revamp production and marketing strategies to regain market share and enhance the overall performance of the wine in the market.

By implementing these recommendations, the company can address the declining trend in Rose wine sales and optimize its marketing efforts. By focusing on the lean period and avoiding peak winter campaigns, the company can effectively boost sales and regain market share. Additionally, understanding the underlying reasons for the decline and making necessary strategic adjustments will position the company for improved performance in the future.