Time Series Forecasting-RoseWine

KUMAR ANKIT
PGP-DSBA
(PGPDSBA.O.SEP23.C)

INDEX

List of Tables:

List of Plots:

Problem Statement:

ABC Estate Wines has been a leader in the rose wine industry for many years, offering high-quality wines to consumers all around the world. As the company continues to expand its reach and grow its customer base, it is essential to analyze market trends and forecast future sales to ensure continued success.

In this report, we will focus on analyzing the sales data for rose wine in the 20th century. Asan analyst for ABC Estate Wines, I have been tasked with reviewing this data to identify patterns, trends, and opportunities for growth in the wine market.This knowledge will help us to make informed decisions about how to position our products in the market, optimize our sales strategies, and forecast future sales trends.

Overall, this report aims to provide valuable insights into the wine market and how ABCEstate Wines can continue to succeed in this highly competitive industry.

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 data set;

Table 2: rows of dataset

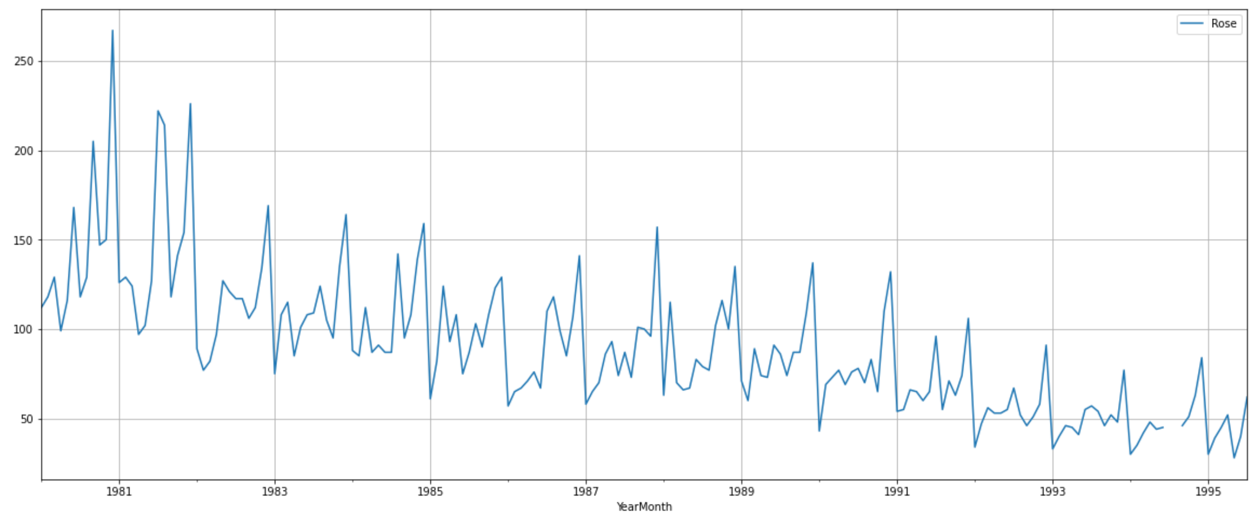
Top Few Rows :		Last Few Rows :	
	Rose		Rose
YearMonth		YearMonth	
1980-01-01	112.0	1995-03-01	45.0
1980-02-01	118.0	1995-04-01	52.0
1980-03-01	129.0	1995-05-01	28.0
1980-04-01	99.0	1995-06-01	40.0
1980-05-01	116.0	1995-07-01	62.0

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 odthe 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.

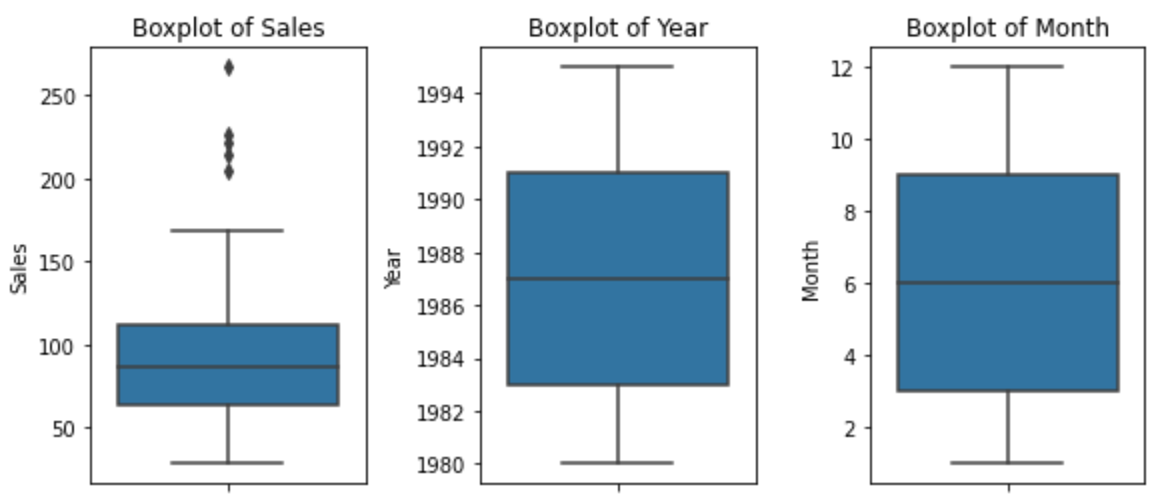
Mean - Before & After

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

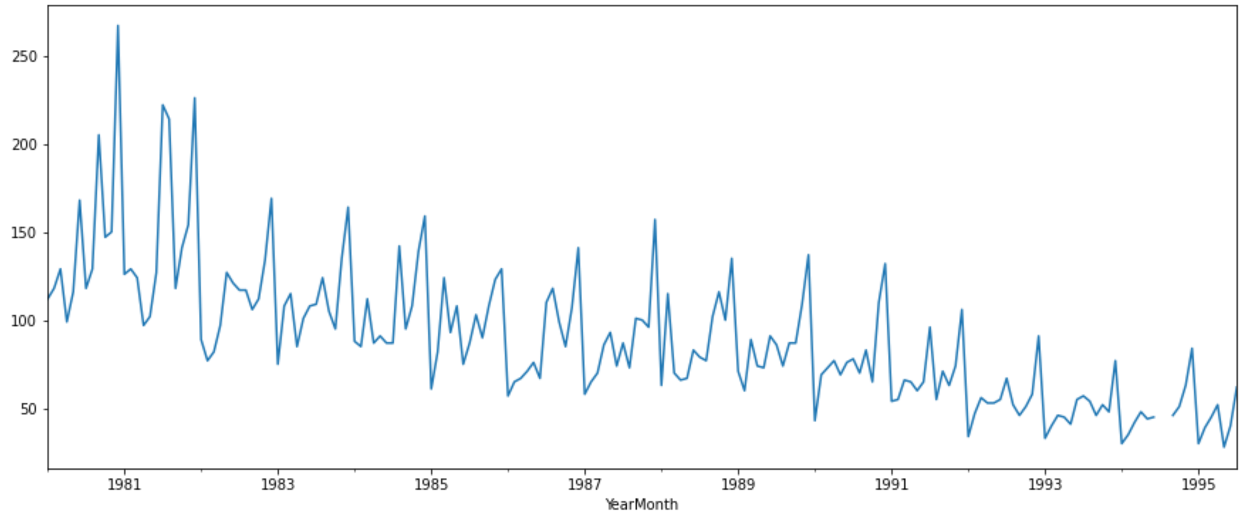


The box plot shows:

- 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:

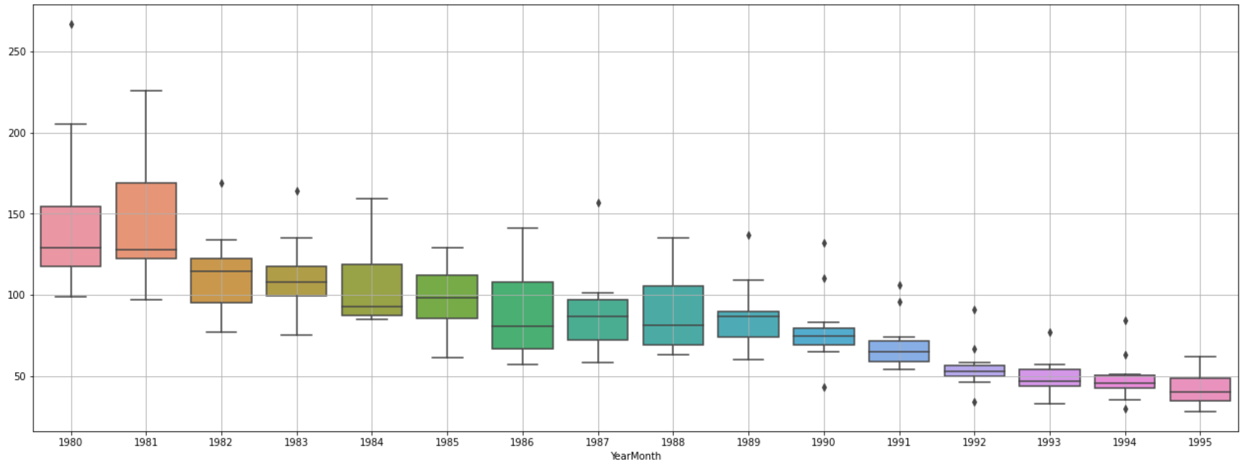
Plot 3: line plot of sales



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

Boxplot Yearly:

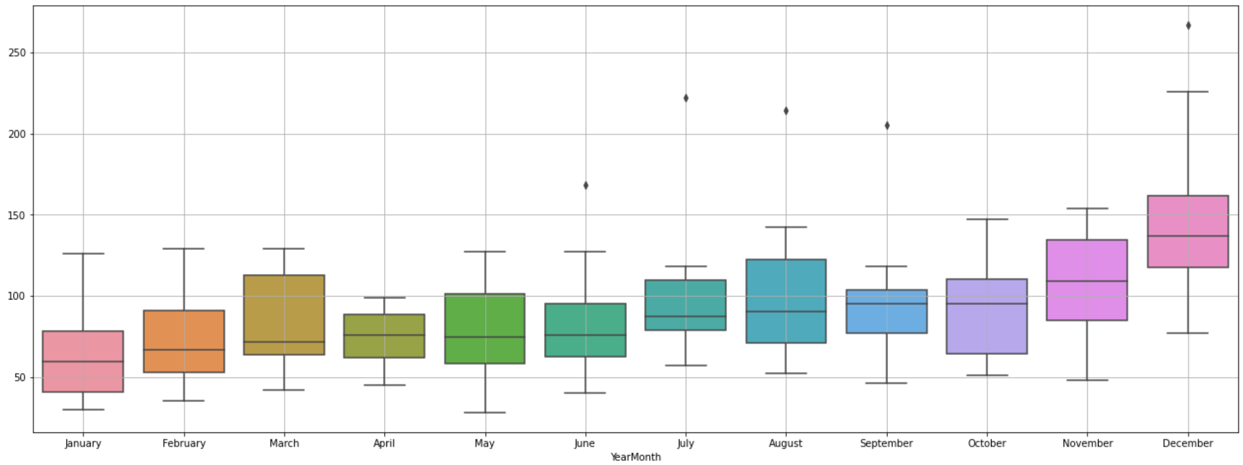
Plot 4: boxplot yearly



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:

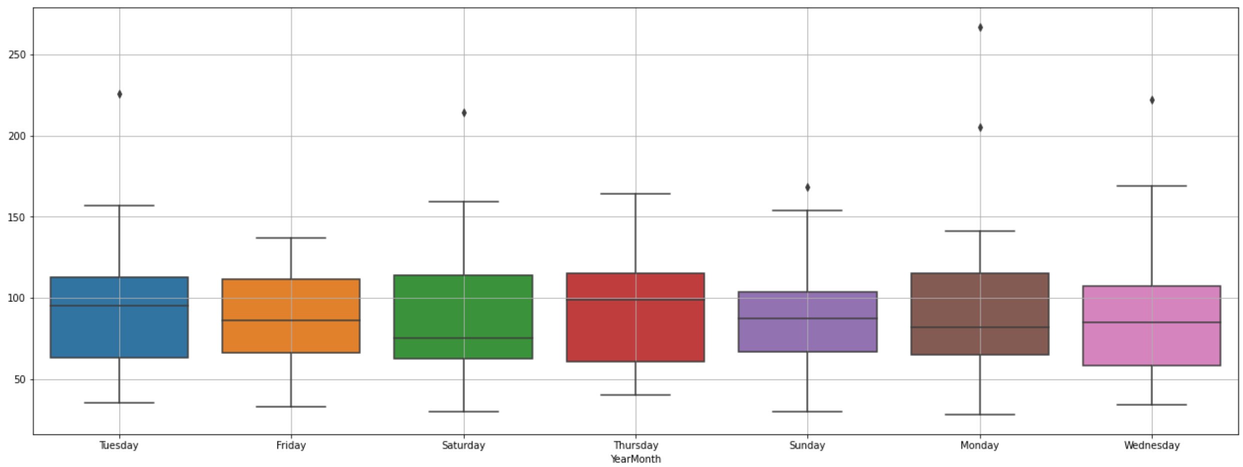
Plot 5: boxplot monthly



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 Weekday vise:

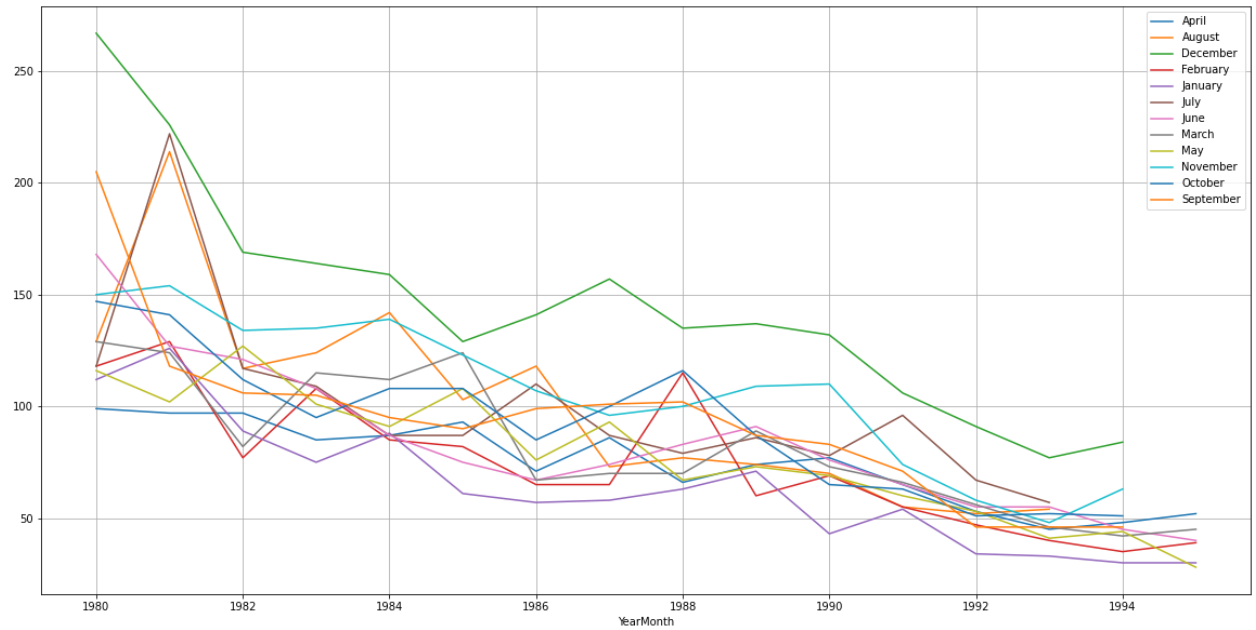
Plot 6: boxplot weekday vise



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:

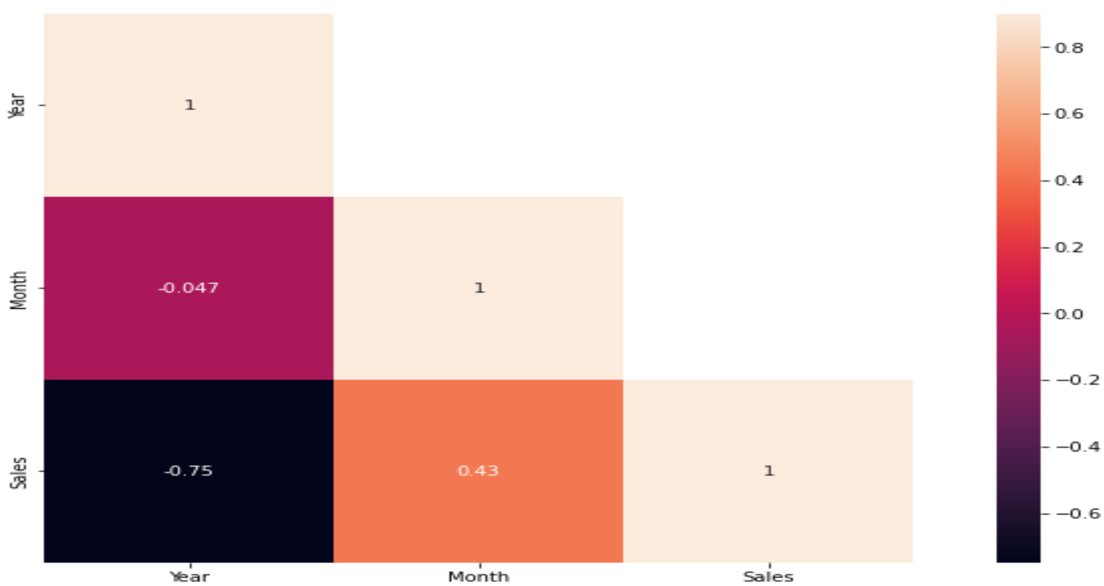
Plot 7: graph of monthly sales over the years



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

Correlation plot

Plot 8:correlatation plot

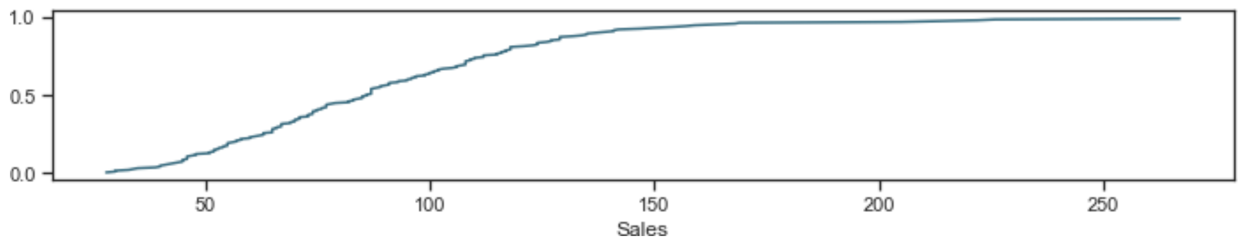


This heat map shows that there was little correlation between Sales and the Years data, there significantly more correlation between the month and Sales columns. Clearly indicating a seasonal pattern in our Sales data. Certain months have higher sales, while certain months have lesser.

Plot ECDF: Empirical Cumulative Distribution Function

This graph shows the distribution of data.Plot

9: ECDF plot

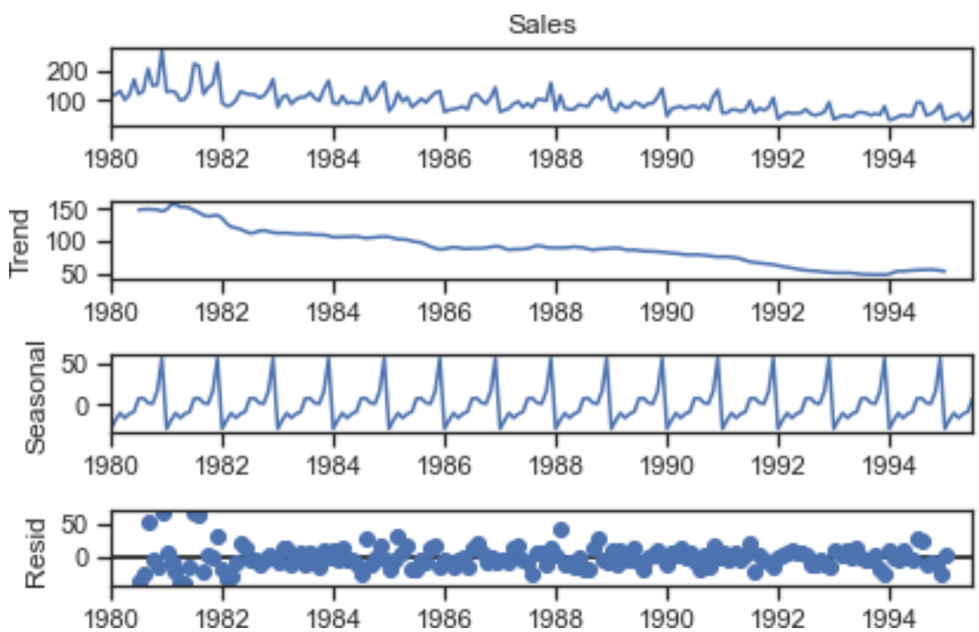


This plot shows:

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

Decomposition -Additive

Plot 10 : decomposition additive

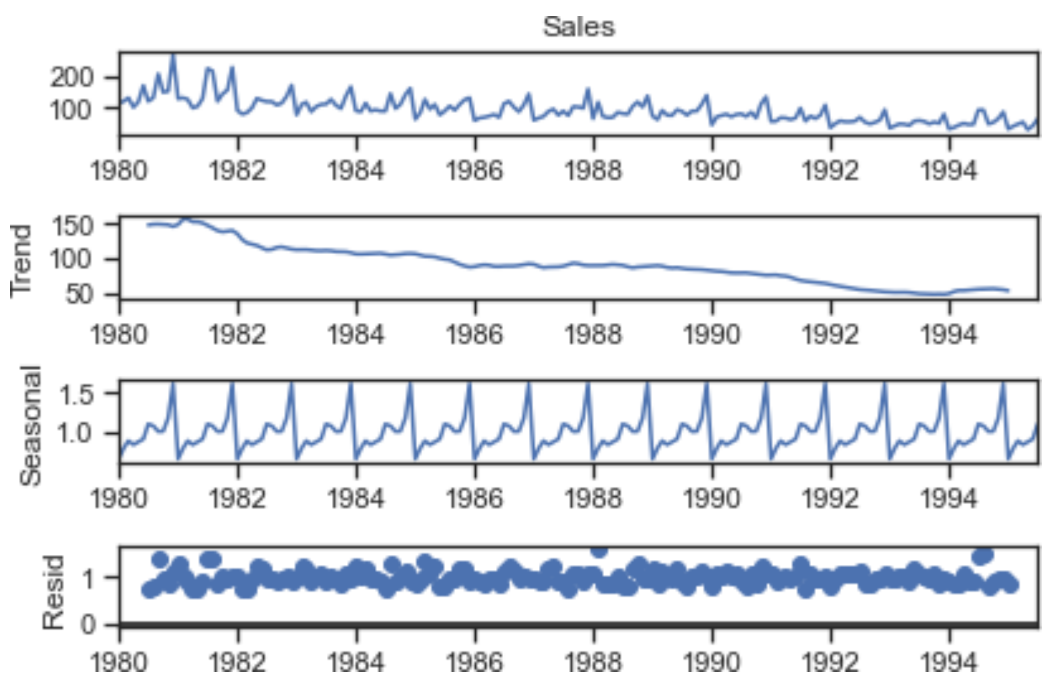


The plots show:

- 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

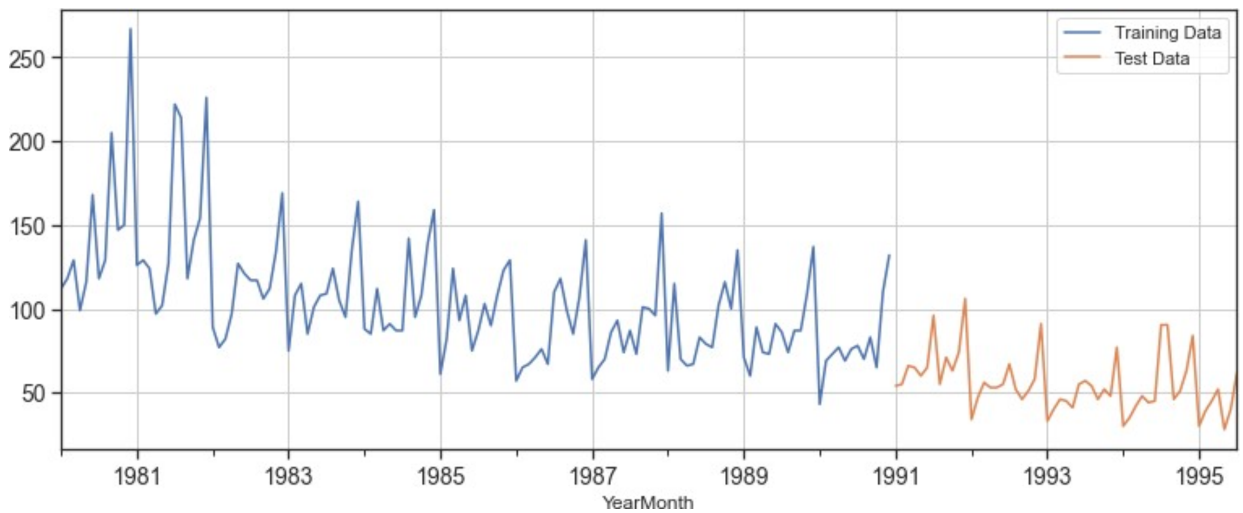


The plots show:

- 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

<table>
<tr><th>Train dataset</th><th>Test dataset</th></tr>
<tr><td><pre>
First few rows of Training Data
             Sales Year Month
YearMonth
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
             Sales Year Month
YearMonth
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
</pre></td><td><pre>
First few rows of Test Data
             Sales Year Month
YearMonth
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
             Sales Year Month
YearMonth
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
</pre></td></tr>
</table>

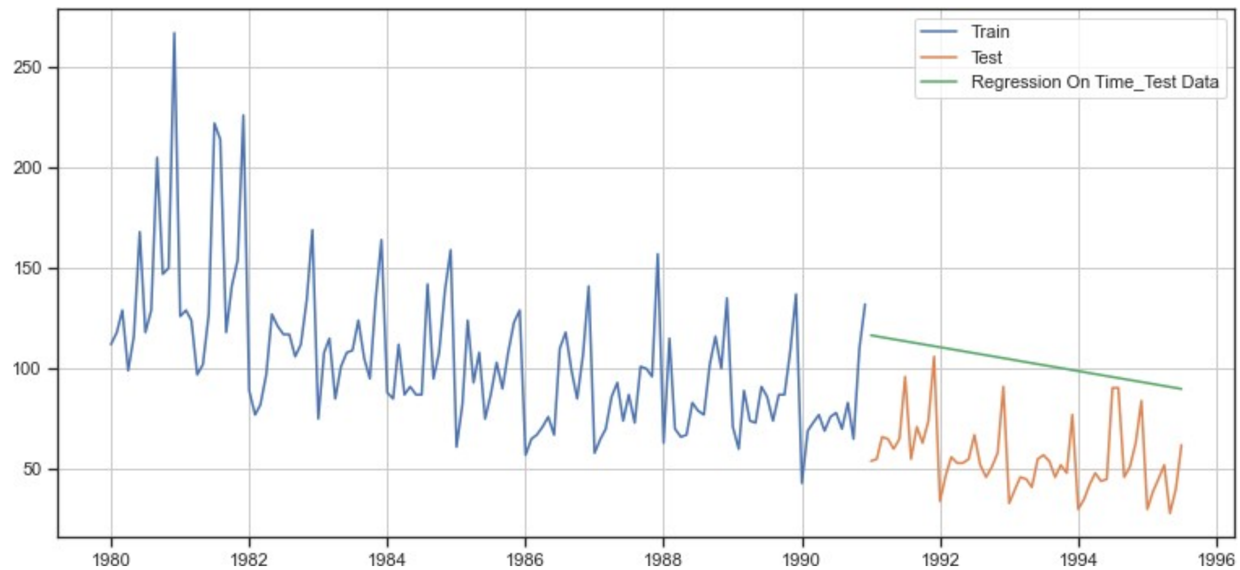
4.Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

- 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)

Linear Regression	51.080941

Model 1: Linear Regression

Plot 13: linear regression



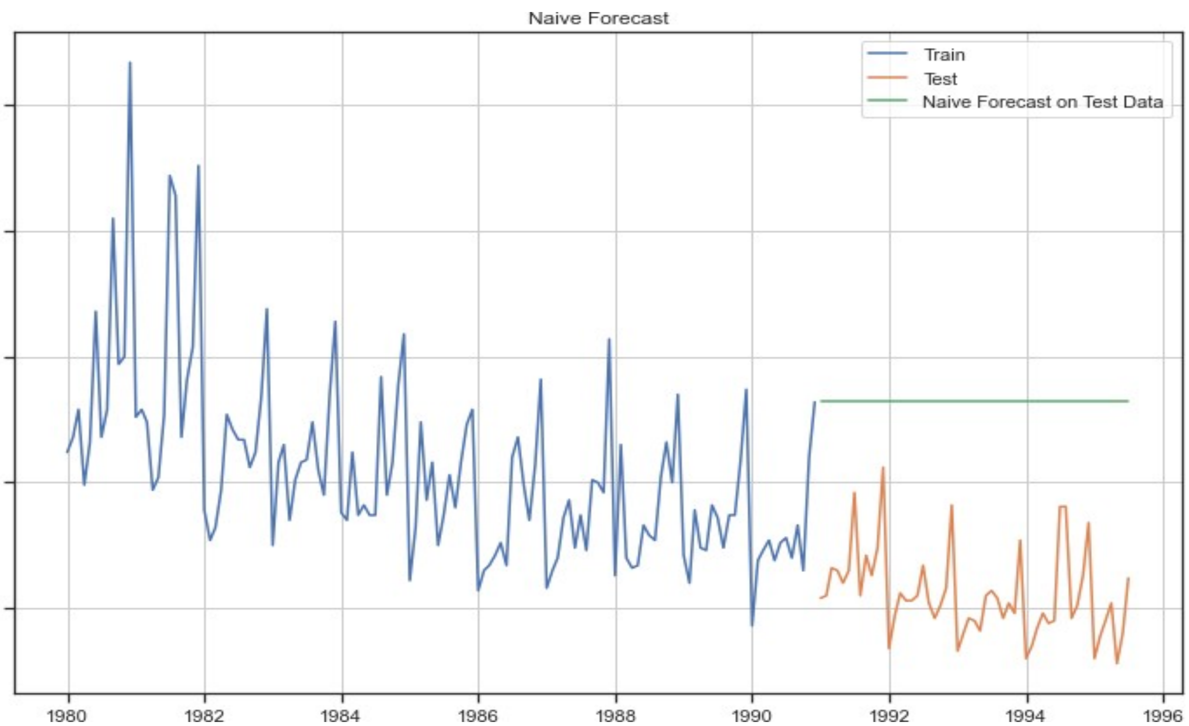
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

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

Naive Model 79.304391

Model 2: Naive Approach:

Plot 14: naive approach



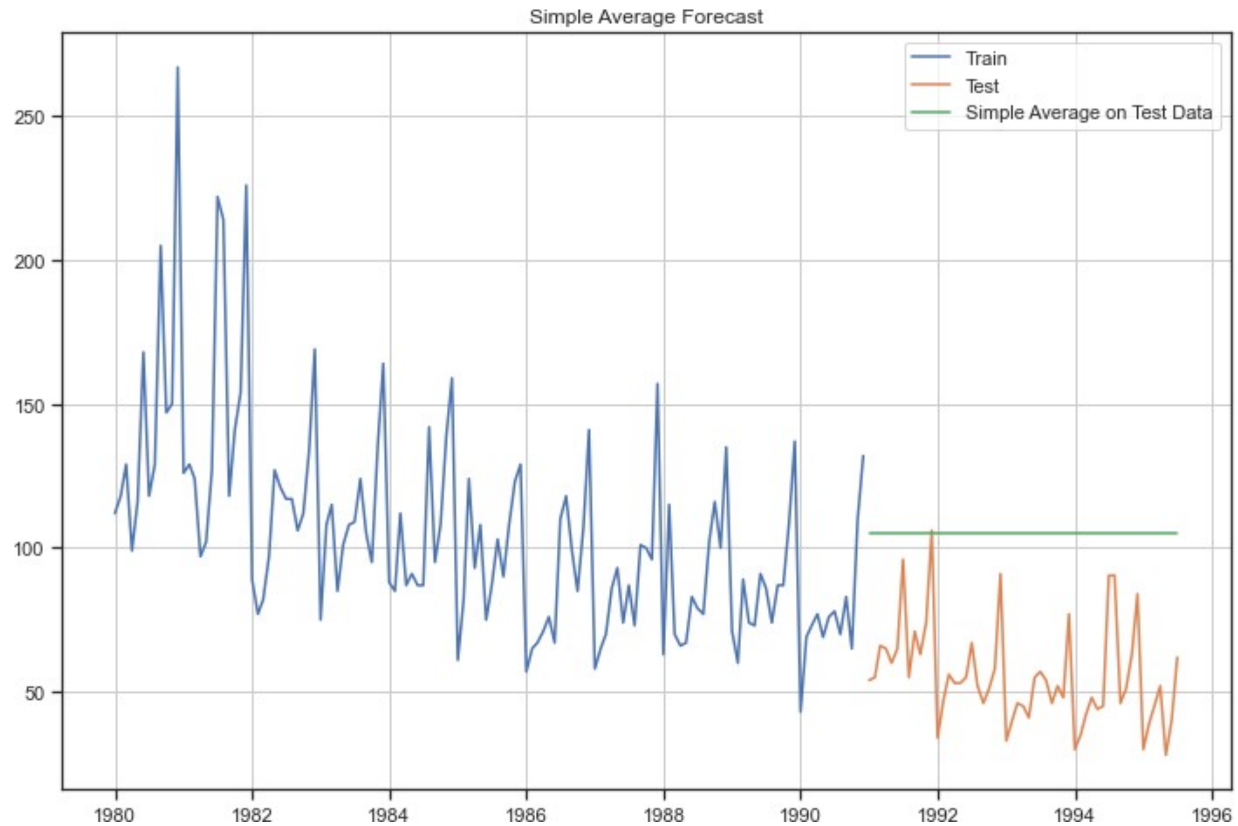
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

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

Simple Average Model	53.049755

Method 3: Simple Average

Plot 15: simple average

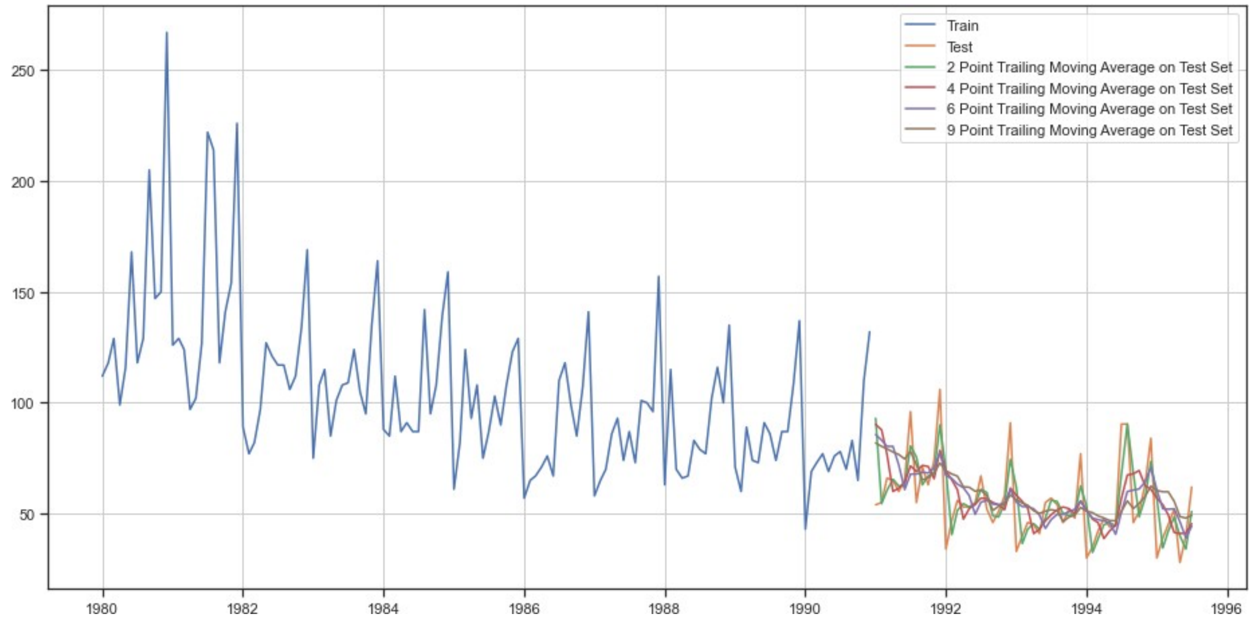


The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

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

Method 4: Moving Average(MA)

Plot 16: moving average



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

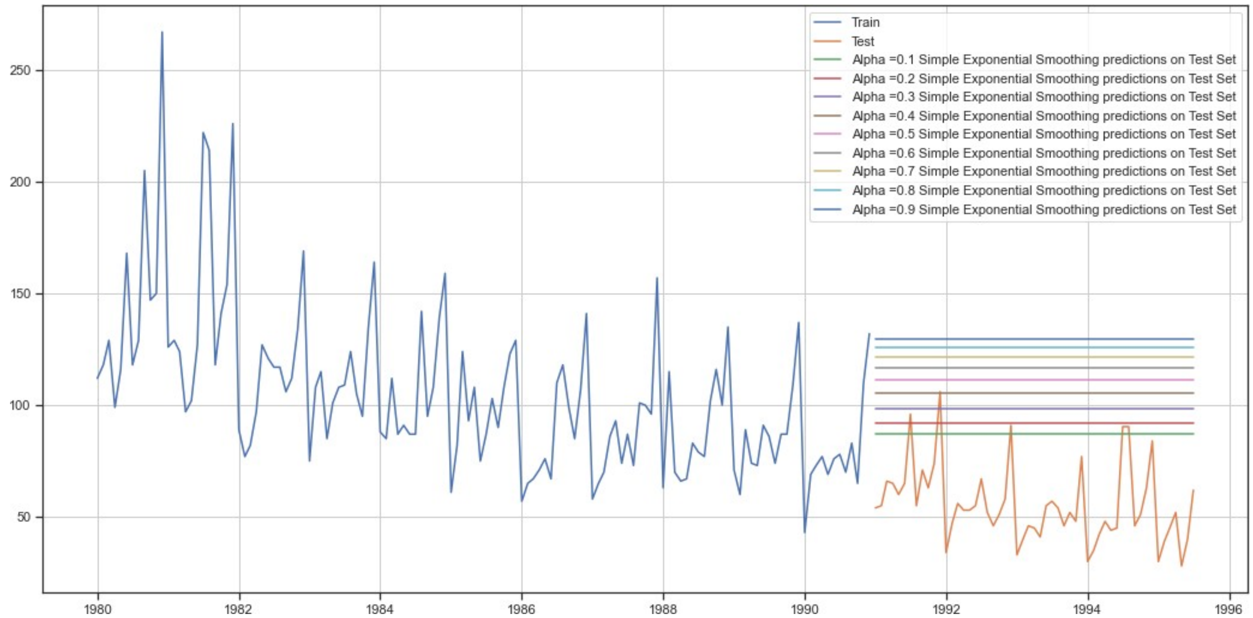
2	pointTrailingMovingAverage	11.589082
4	pointTrailingMovingAverage	14.506190
6	pointTrailingMovingAverage	14.558008
9	pointTrailingMovingAverage	14.797139

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.

Alpha=0.1,SimpleExponentialSmoothing	36.429535

Method 5: Simple Exponential Smoothing

Plot 17: simple exponential smoothing

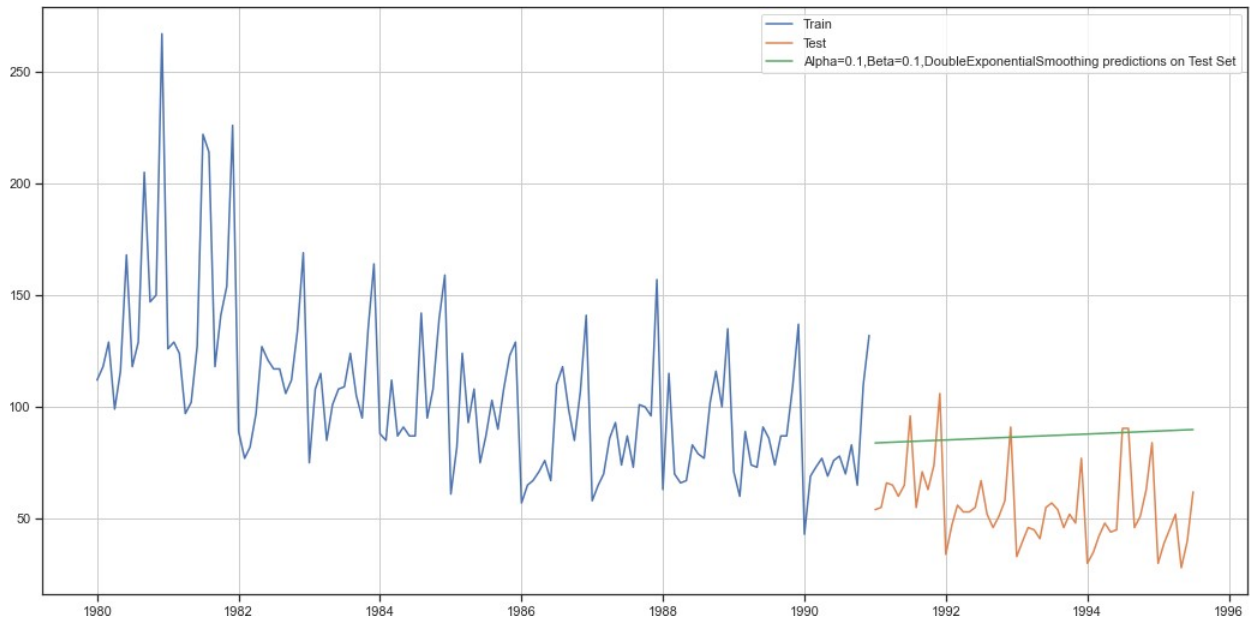


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

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

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

Plot 18: double exponential smoothing

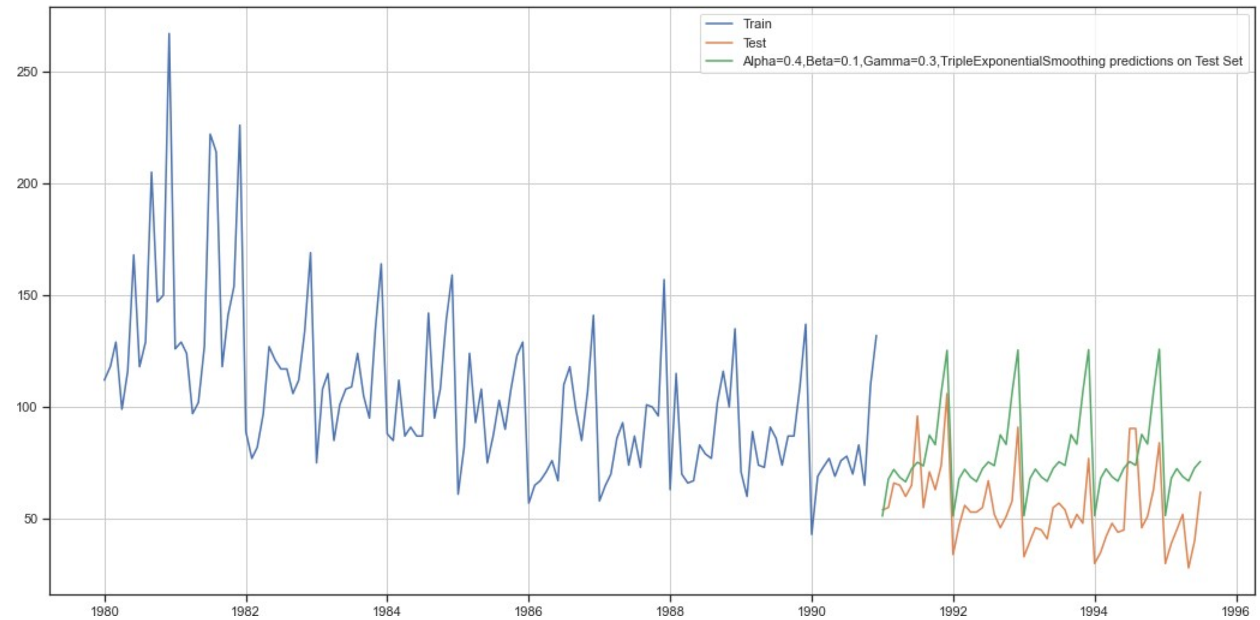


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

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

Method 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.

5.Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and alsomention 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.

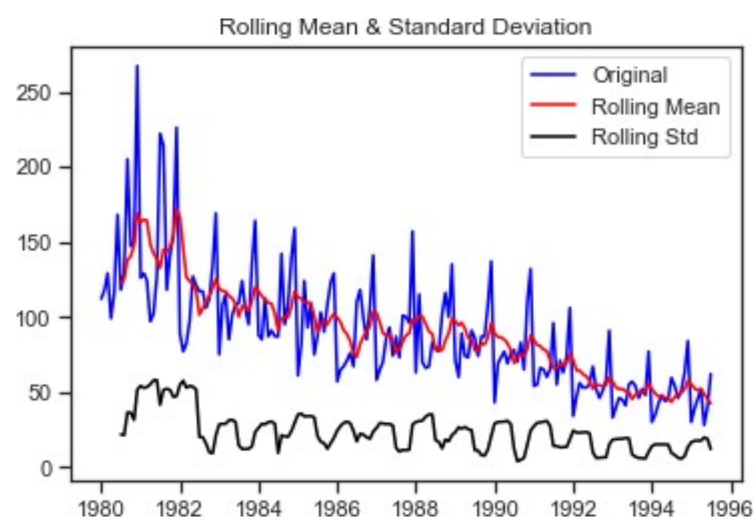
The hypothesis in a simple form for the ADF test is:

- 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 thep-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:

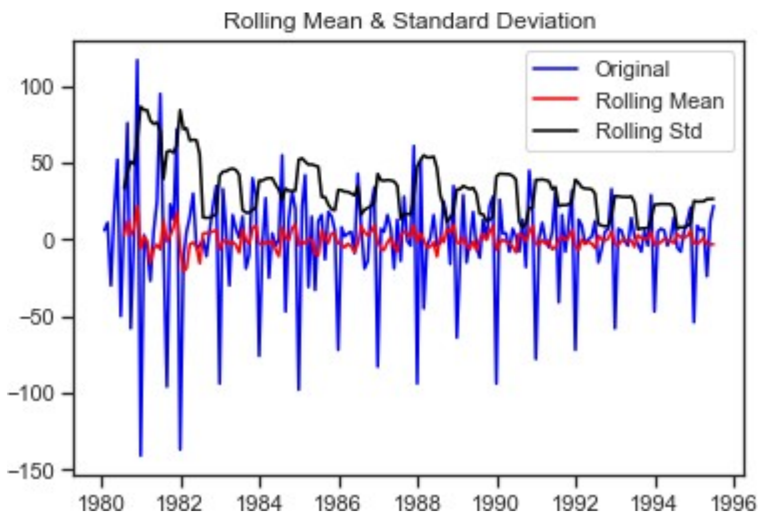
Test Statistic	-1.892338
p-value	0.335674

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 modelin 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.

36.42079120523518

	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:                Thu, 16 Feb 2023   AIC                           1274.695
Time:                        01:28:51   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         -1.6774      0.084    -19.998      0.000      -1.842      -1.513
ar.L2         -0.7282      0.084     -8.679      0.000      -0.893      -0.564
ma.L1          1.0448      0.631      1.656      0.098      -0.192       2.281
ma.L2         -0.7716      0.133     -5.799      0.000      -1.032      -0.511
ma.L3         -0.9044      0.572     -1.582      0.114      -2.025       0.216
sigma2       860.2717    531.354      1.619      0.105    -181.163    1901.706
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.35
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:

AUTO- SARIMA Model

A similar for loop like AUTO_ARIMA with below values was employed, resulting in themodels 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

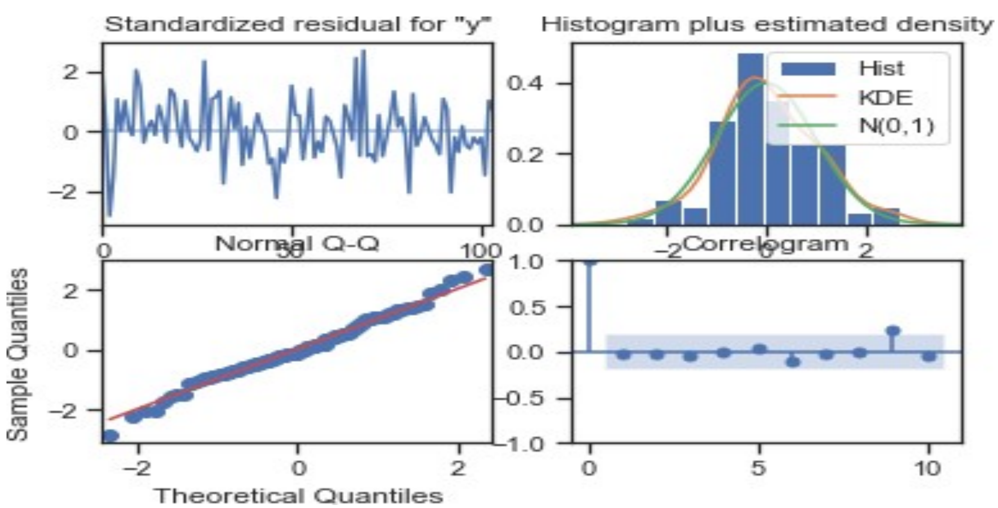
the summary report for the best SARIMA model with values (3,1,1)(3,0,2,12)

18.53502803217281


```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -446.366
Date:                            Fri, 17 Feb 2023   AIC                            906.732
Time:                                    02:07:40   BIC                            925.243
Sample:                                         0   HQIC                           914.231
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1142      0.369     -0.310      0.757      -0.837       0.608
ma.L1         -0.6699    365.024     -0.002      0.999    -716.104     714.764
ma.L2         -0.3301    120.553     -0.003      0.998    -236.610     235.950
ar.S.L12       0.6255      0.059     10.547      0.000       0.509       0.742
ma.S.L12      -0.1613      0.126     -1.285      0.199      -0.407       0.085
ma.S.L24       0.1133      0.134      0.844      0.399      -0.150       0.376
sigma2       299.9889    1.1e+05      0.003      0.998   -2.14e+05    2.15e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):                 0.59
Prob(Q):                              0.78   Prob(JB):                         0.74
Heteroskedasticity (H):               0.71   Skew:                             0.08
Prob(H) (two-sided):                  0.31   Kurtosis:                         3.33
===================================================================================
```

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

Plot 22: sarima plots



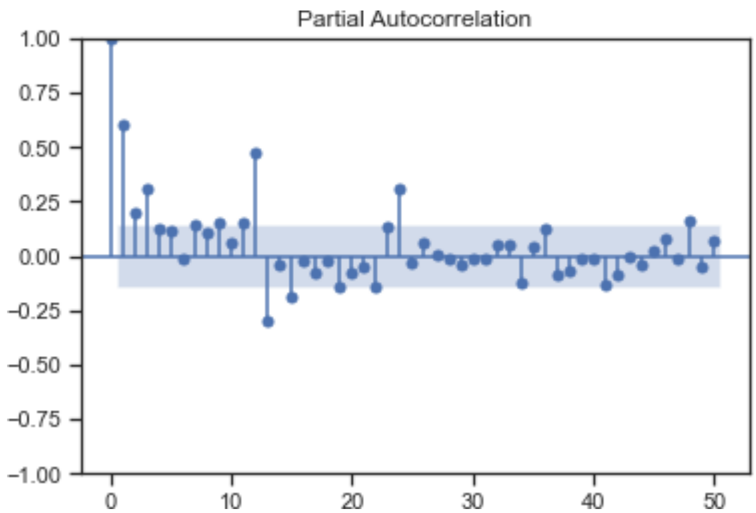
RSME of Model:

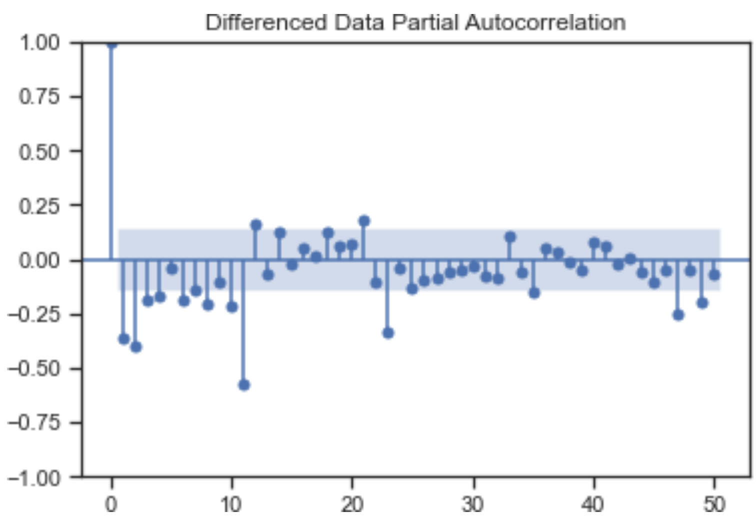
7.Build ARIMA/SARIMA models based on the cut-off pointsof ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Manual- ARIMA Model

PACF the ACF plot on data :

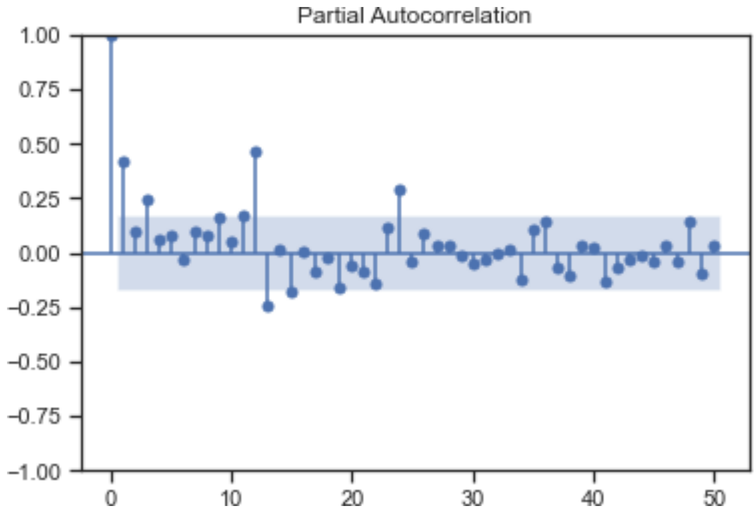
Plot 23: PACF and ACF plots

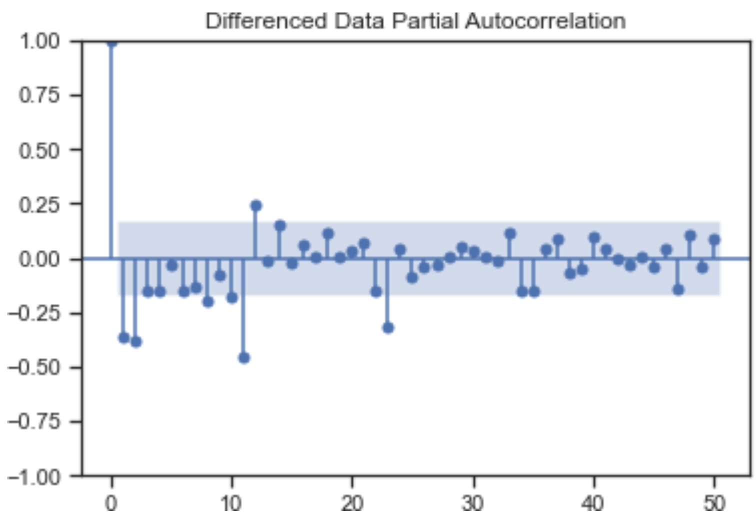




Following is plotting the PACF and ACFgraph for the training data.

Plot 24: PACF and ACF plota of train date





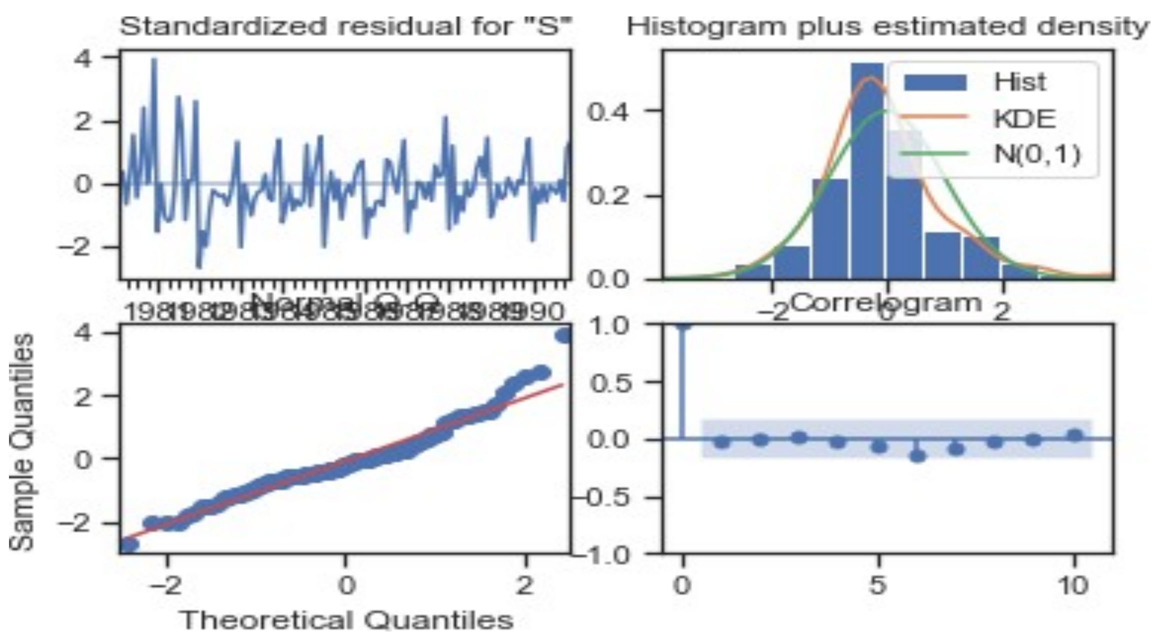
Hence the values selected for manual ARIMA:- p=2, d=1, q=2

summary from this manual ARIMA model.

RMSE: 36.47322487814613


```
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Fri, 17 Feb 2023   AIC                           1281.871
Time:                        02:22:45   BIC                           1296.247
Sample:                    01-01-1980   HQIC                          1287.712
                         - 12-01-1990
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
===================================================================================
```

Plot 25: manual arima model plots



Model Evaluation: RSME

Manual SARIMA Model

Looking at the ACF and PACF plots for training data, we can clearly see significant spikes atlags 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)

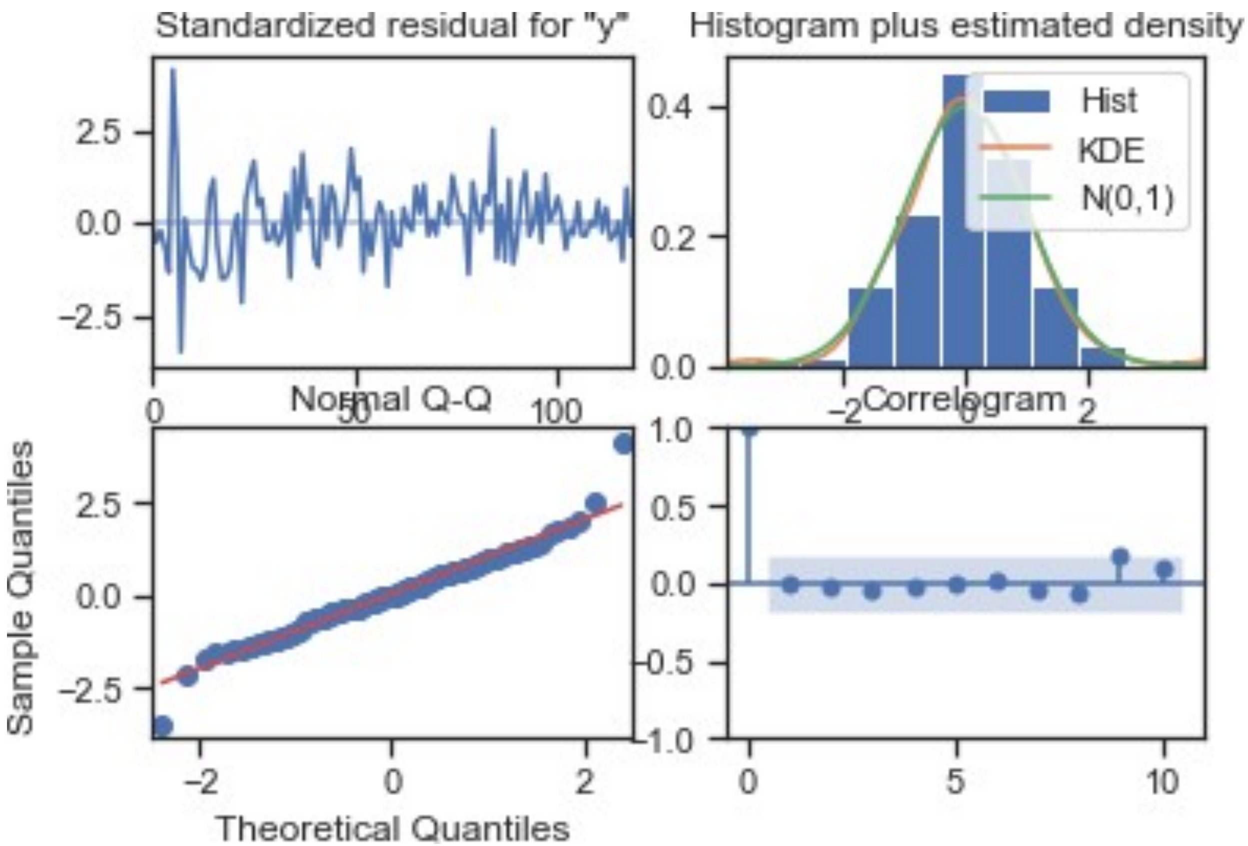
Below is the summary of the manual SARIMA model

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 132
Model:             SARIMAX(2, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -538.016
Date:                             Fri, 17 Feb 2023   AIC                           1094.031
Time:                                     02:30:44   BIC                           1119.044
Sample:                                          0   HQIC                          1104.188
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5489      0.228     -2.406      0.016      -0.996      -0.102
ar.L2         -0.0745      0.099     -0.753      0.452      -0.268       0.119
ma.L1         -0.1705      0.217     -0.787      0.431      -0.595       0.254
ma.L2         -0.6692      0.228     -2.934      0.003      -1.116      -0.222
ar.S.L12      -1.0135      0.524     -1.934      0.053      -2.041       0.014
ar.S.L24      -0.1002      0.175     -0.572      0.568      -0.444       0.243
ma.S.L12       0.2902     15.207      0.019      0.985     -29.515      30.095
ma.S.L24      -0.7073     10.863     -0.065      0.948     -21.998      20.583
sigma2       430.7381   6348.468      0.068      0.946    -1.2e+04    1.29e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                27.16
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).
```

14.975041301618377

Plot 26: manula sarima plots



Model Evaluation: RSME

8. Build a table (create a data frame) with all the models built along withtheir 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 andgamma 0.2 is the best as it he the lowest RSME score.

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

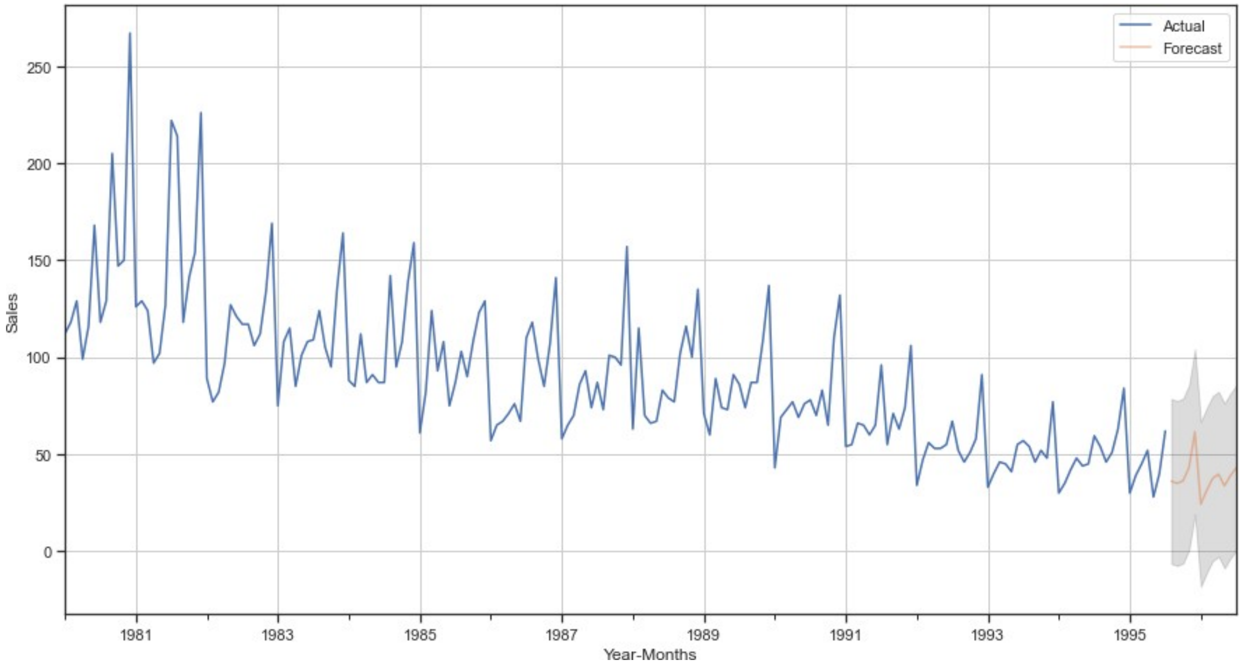
Based on the above comparison of all the various models that we had built, we can conclude that the triple exponential smoothing or the Holts-Winter model is giving us the lowest RMSE, hence it would be the most optimum model

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.

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

- The analysis of the wine sales data indicates a clear downward trend for the Rose wine variety for the company, which has been declining in popularity for more than a decade.
- This trend is expected to continue in the future as well, based on the predictions of the most optimal model.
- Wine sales are highly influenced by seasonal changes, with sales increasing during festival season and dropping during peak winter time i.e. January.
- The company should consider running campaigns to boost the consumption of the wine during the rest of the year, as sales are subdued during this period.
- Campaigns during the lean period (April to June) might yield maximum results for the company, as sales are low during this period, and boosting them would increase the overall performance of the wine in the market across the year.
- Running campaigns during peak periods (such as during festivals) might not generate significant impact on sales, as they are already high during this time of the year.
- Campaigns during peak winter time (January) are not recommended as people are less likely to purchase wine due to climatic reasons, and running campaigns during this period may not change people's opinion.
- The company should also consider exploring reasons behind the decline in popularity of the Rose wine variety, and if needed, revamp its production and marketing strategies to regain the market share.