TIME SERIES FORECASTING

NAME: SHOUNACK MANDAL

COURSE: PGP - DSBA Online Sep.

Date: 17/ APRIL / 2022

Contents

LIST OF FIGURE

LIST OF TABLES

PROBLEM 1: TSF SPARKLING

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

The monthly time stamp from Jan 1980 to July 1995 and the sales corresponding to the wines. Creating time stamps and adding it to the data frame to make it a time-series data.

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

Table 1 Head of the dataframe

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

Table 3 Setting data range

	YearMonth	Sparkling	Time_Stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

Table 2 Head of the time series data

Given data is not time. So we parse the date range and create a timestamp. We also notice the increasing trend in the initial years. Plotting the Sparkling Time Series to understand the behaviour of the data. The data consist of 187 data points.

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Table 4 Set Time_Stamp as an index

```
YearMonth
1980-01-01    1686
1980-02-01    1591
1980-03-01    2304
1980-04-01    1712
1980-05-01    1471
Name: Sparkling, dtype: int64
```

Table 5 View the top 5 rows of Sparkling dataset.

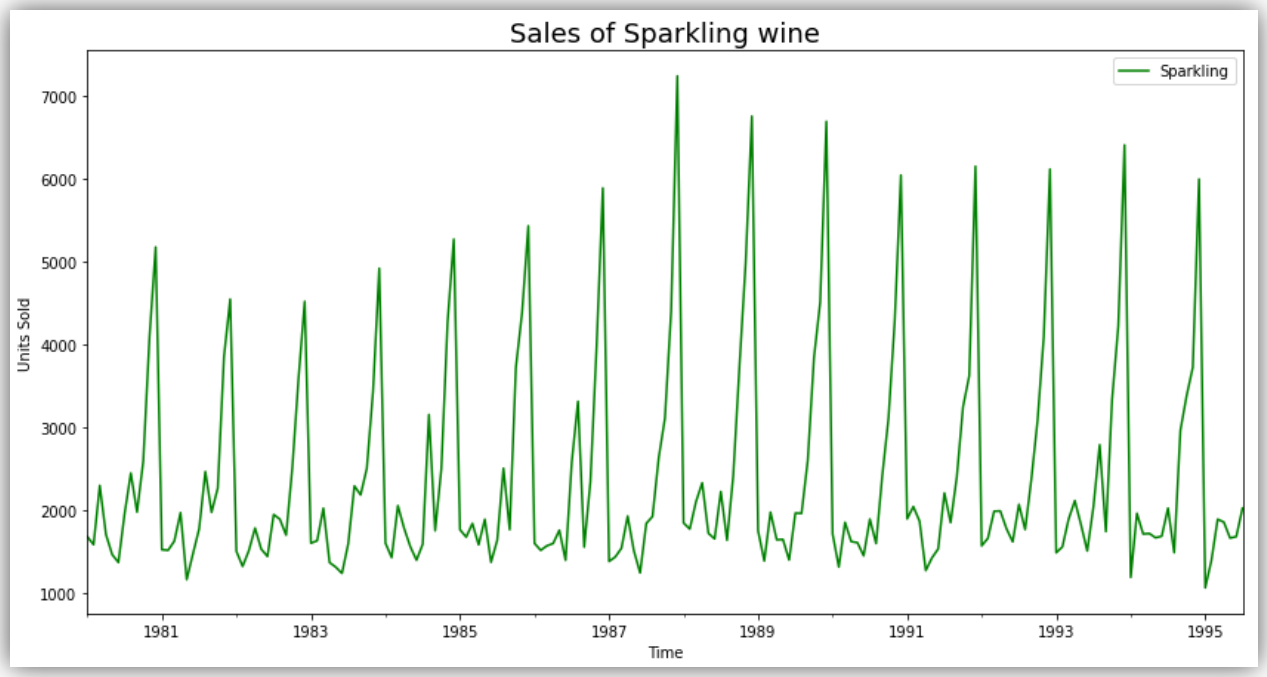


Figure 1 Plot the Sparkling time series to understand the behaviour of the data.

All the values are properly loaded for the dataset with the index as panda`s data-time format. Sparkling time series data do not contain any missing values. The Sparkling wine dataset shows significant

seasonality and doesn`t shows any consistent trend but has upward and downward slopes during the time period. Sparkling wine has been consistently favored over the years by customers.

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

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

Table 6 Data description for Sparkling Dataset

INFERENCES OF THE DATA ALSO YEARLY AND MONTHLY FIGURES: -

- Ploting a yearly boxplot to understand the spread of sales across different years and within different months across years.
- The basic measures of descriptive statistics tell us how the Sales have varied across years. But for this measure of descriptive statistics we have averaged over the whole data without taking the time component into account.
- The descriptive summary of the data shows that on an average 2402 units of Sparkling wines were sold each month on the given period of time. 50% of month's sales varied from 1605 units to 2549 units.
- Maximum sale reported in a month is 7242 units.
- The yearly-boxplot, shows that the average sale of Sparkling has been more or less consistent across the period, at or a little below 2000 units.
- The outliers in the yearly-boxplot most probably represent the seasonal sale during the seasonal months.
- The monthly-box-plot shows a clear seasonality during the festive seasonal months of October, November and December, which peaks in December. The sale lowest in the month of June.

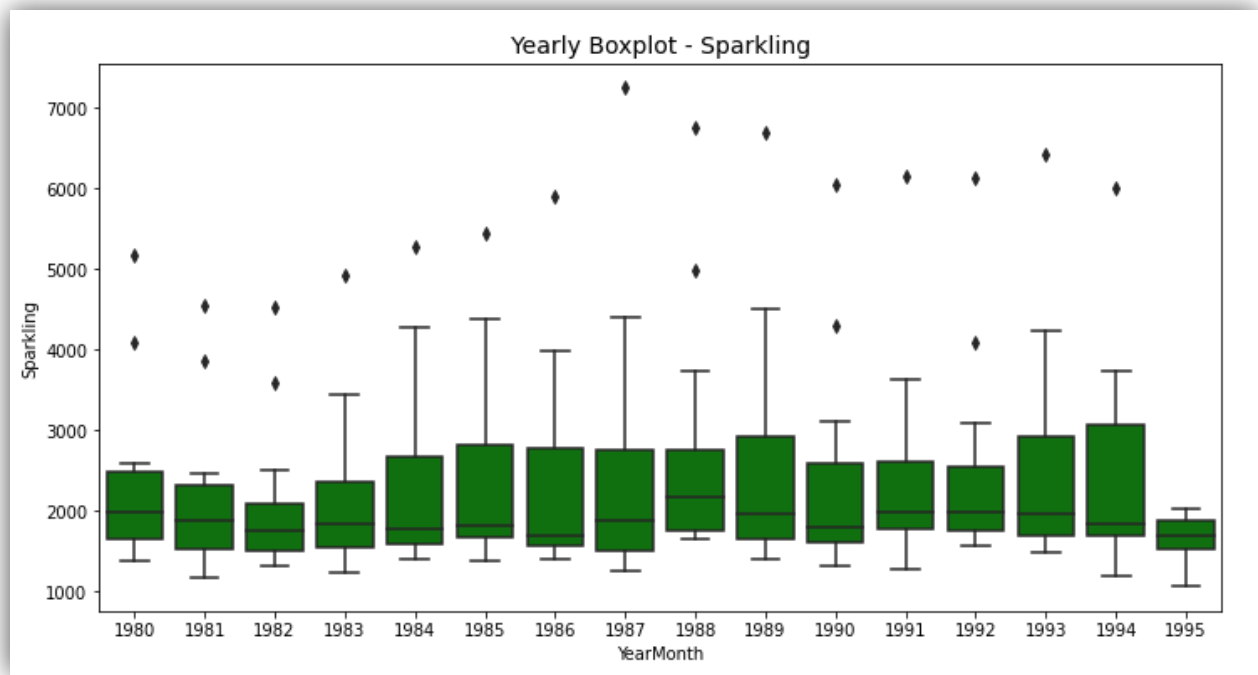


Figure 2 Yearly Boxplot

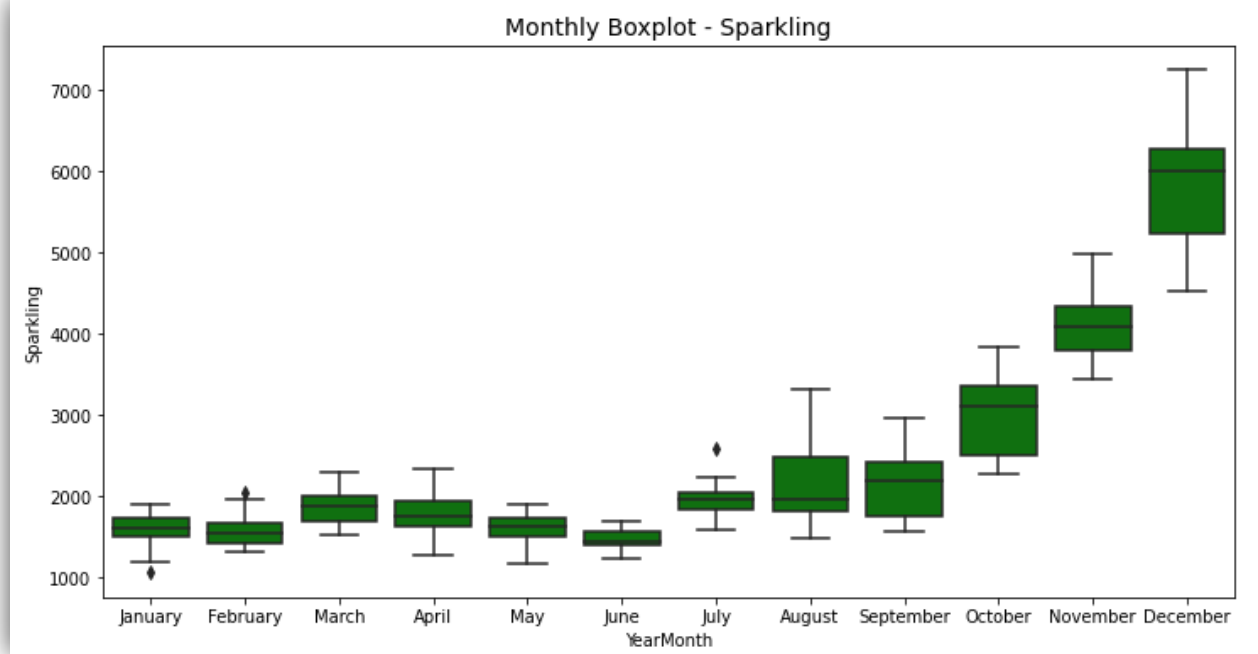


Figure 3 Monthly Boxplot for all the years for Sparkling Dataset

Plotting time series month plot to understand the spread of sales across different years and within different months across years.

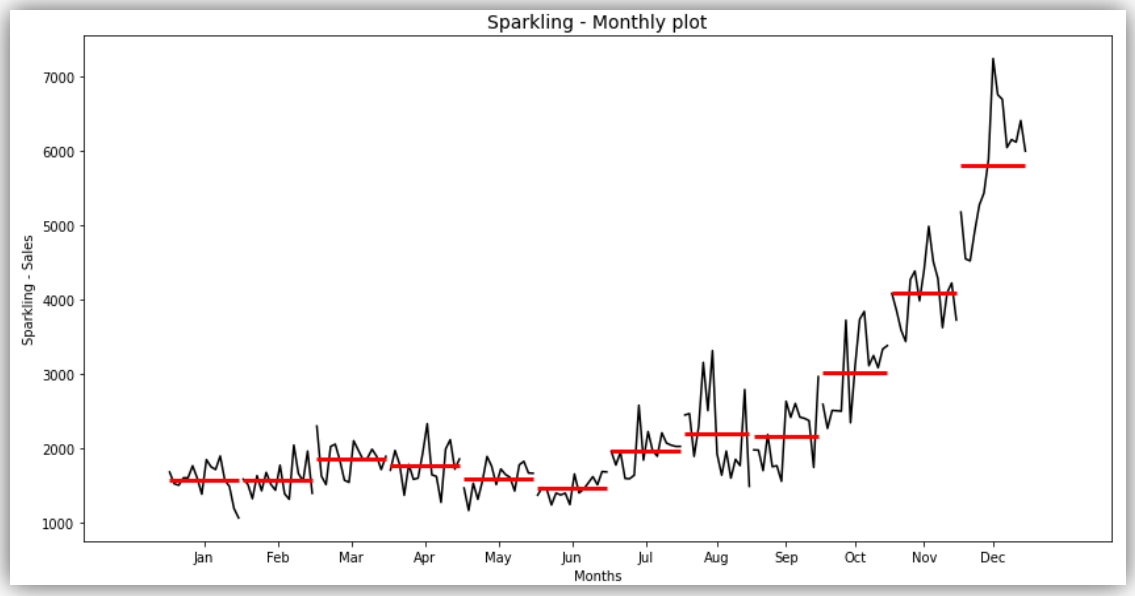


Figure 4 Monthly and yearly trends of the dataset

- The monthly plot for Sparkling shows mean and variation of units sold each month over the years. Sale's in seasonal month's shows a higher variation than in the lean months.
- Sale in December with a mean few points below 6000, varies from 7400 to 4500 units over the years. Whereas sale in November varies from 3500 units to 5000 units and sale in October varies from 2500 to 4000 units.
- The lean months from January till September shows more or less a consistent sale around 2000 units.

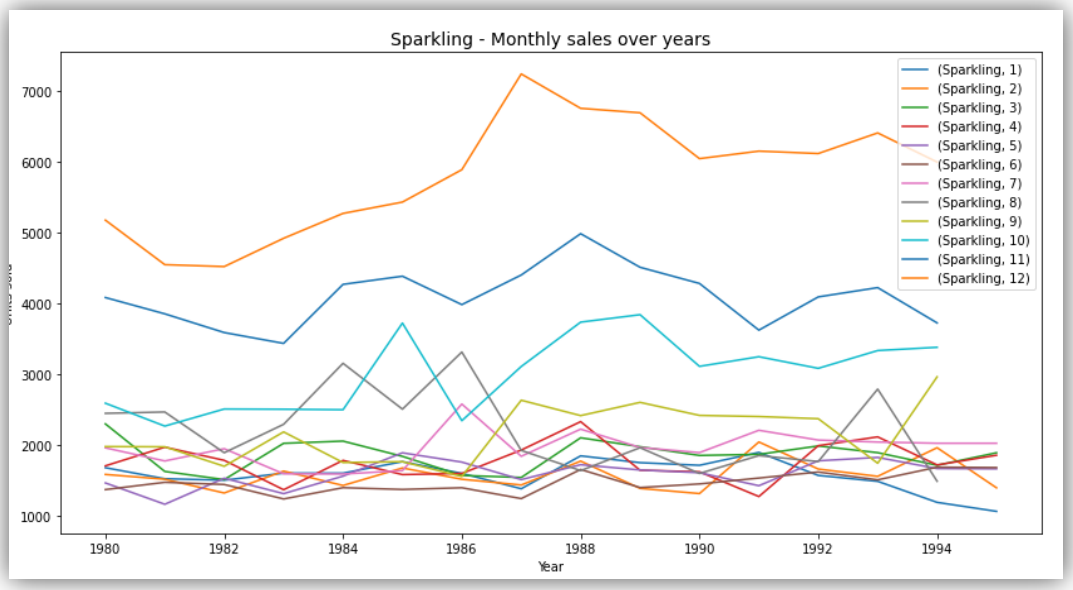


Figure 5 Monthly sales over different years

- The plot of monthly sale over the years also shows the seasonality component of the time-series, with October, November and December selling exponentially higher volumes.
- The highest volume of Sparkling wines were sold in December, 1987 and the least of December sale was in 1981. Post 1987 December sales is around an average 6500 units, which was around 5000 in early 80's.
- The seasonal sale since 1990 has been more or less consistent around 6000 units in December, 4000 units in November and 3000 units in October.
- Sales for the months from January to July is seen to be consistent across the years, compared to the rest of the months.

Decompose the Time Series and plot the different components.

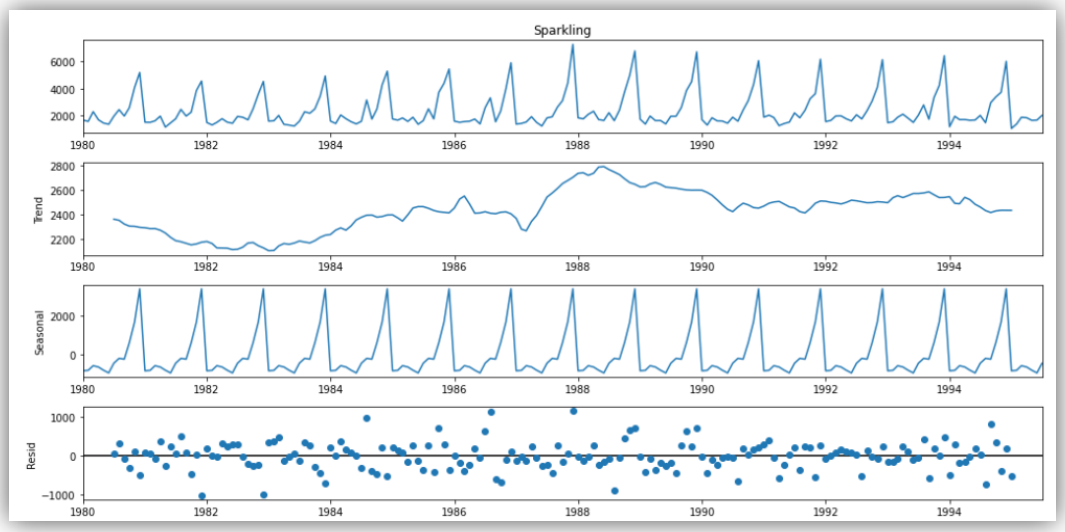


Figure 6 Decomposition of Sparkling Time Series with multiplicative Seasonality: Additive Model

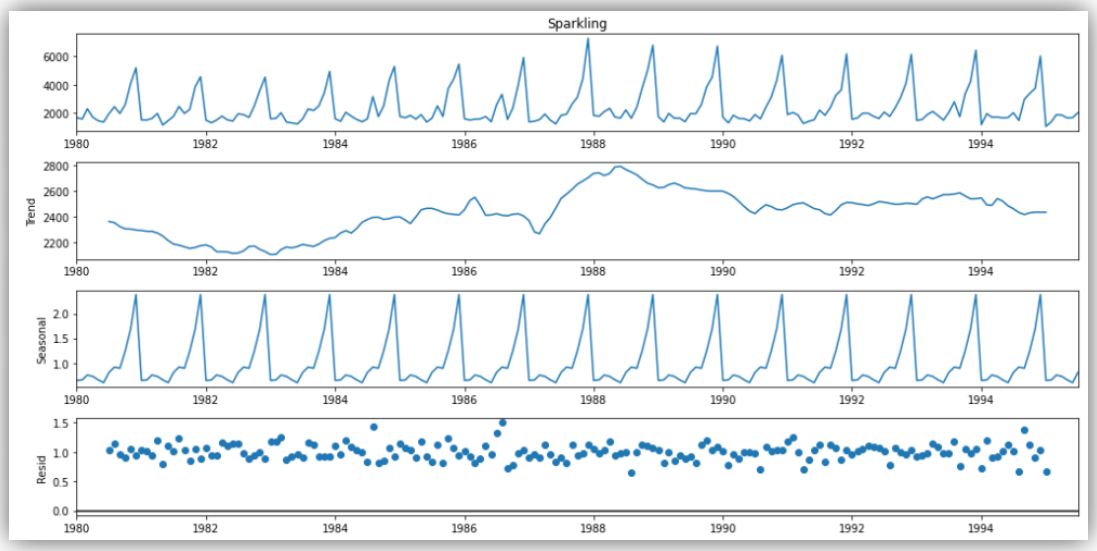


Figure 7 Decomposition of Sparkling Time Series with additive Seasonality: Multiplicative Model

- Decompose the Time Series and plot the different components:
- The takeaways from the decomposition plots of Sparkling wine sales is
- As the altitude of the seasonal peaks in the observed plot is changing according to the change in trend, the time-series is assumed to be 'multiplicative'.
- The plot of the trend component does not show a consistent trend, but an intermediary period shows an upward trend which gets consistent on the late half of time-series.
- The additive model shows the seasonality with a variance of 3000 units and the multiplicative model shows a variance of 30%.
- The residual shows a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions.
- The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 10%.
- If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then we have a multiplicative series.

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

The train and test datasets are created with year 1991 as starting year for test data :-

```
Length of Train Data: 132
Length of Test Data: 55
```

Table 7 Length of the train and test data

```
First few rows of Training Data:
            Sparkling
YearMonth
1980-01-01       1686
1980-02-01       1591
1980-03-01       2304
1980-04-01       1712
1980-05-01       1471

Last few rows of Training Data:
            Sparkling
YearMonth
1990-08-01       1605
1990-09-01       2424
1990-10-01       3116
1990-11-01       4286
1990-12-01       6047
```

```
First few rows of Test Data:
            Sparkling
YearMonth
1991-01-01       1902
1991-02-01       2049
1991-03-01       1874
1991-04-01       1279
1991-05-01       1432

Last few rows of Test Data:
            Sparkling
YearMonth
1995-03-01       1897
1995-04-01       1862
1995-05-01       1670
1995-06-01       1688
1995-07-01       2031
```

Table 8 Train and Test data

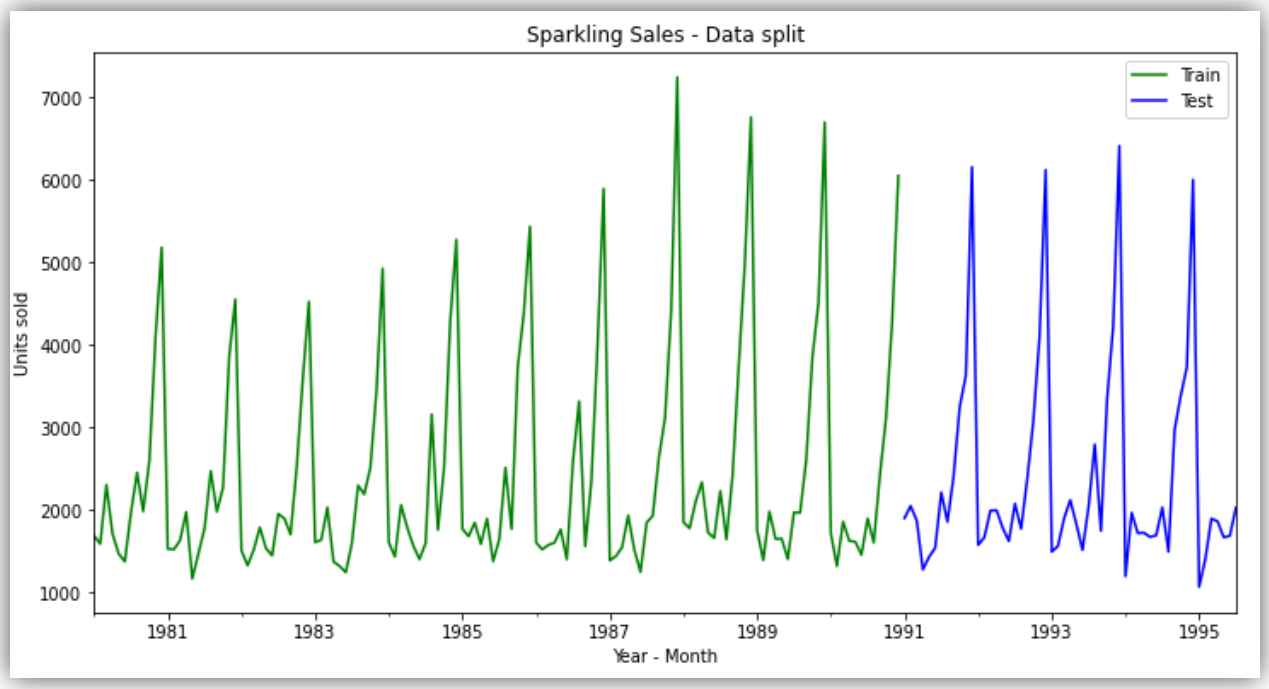


Figure 8 The Plot Sparkling Time Series as train and test

1.4 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

To regress the sale of Sparkling wines, numerical time instance order for both training and test set were generated and the values added to the respective datasets

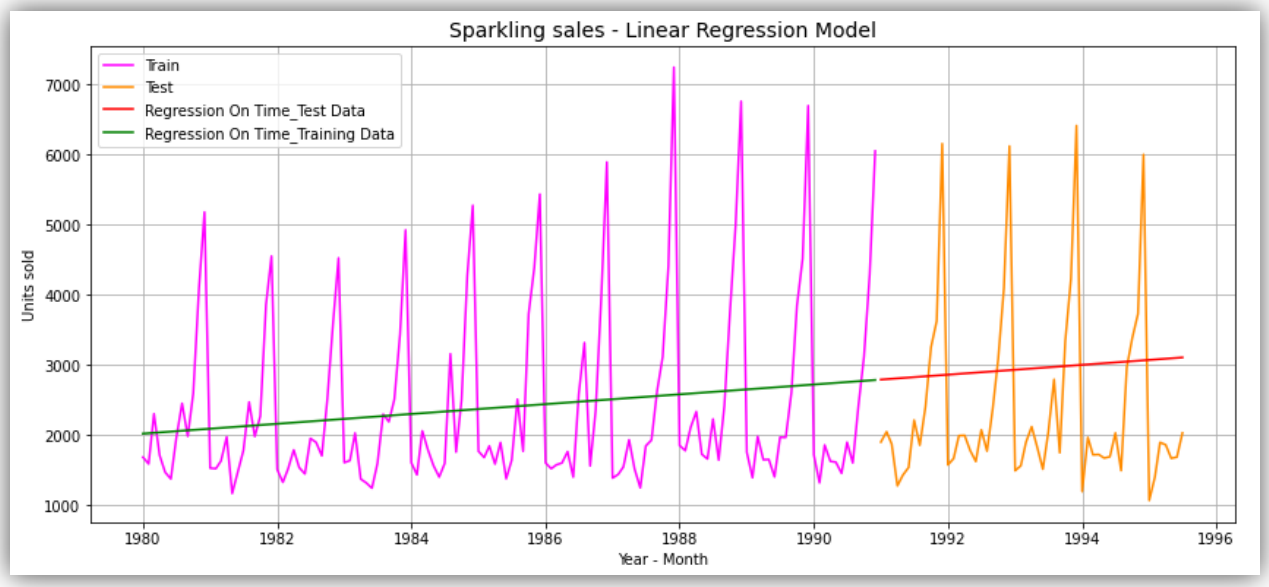


Figure 9 Linear Regression model trend of the dataset

- The linear regression plots shows a gradual upward trend in forecast of Sparkling wine, consistent with the observed trend which was not visually apparent.
- For Regression on Time forecast on the Test Data, RMSE is 1389.135.

Model 2: Naïve Forecast

- In naive model, the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.
- The model has taken the last value from the test set and fitted it on the rest of the train time period and used the same value to forecast the test set.
- For Naive forecast on the Test Data, RMSE is 3864.279
- The model do not capture the trend or seasonality for the given dataset.

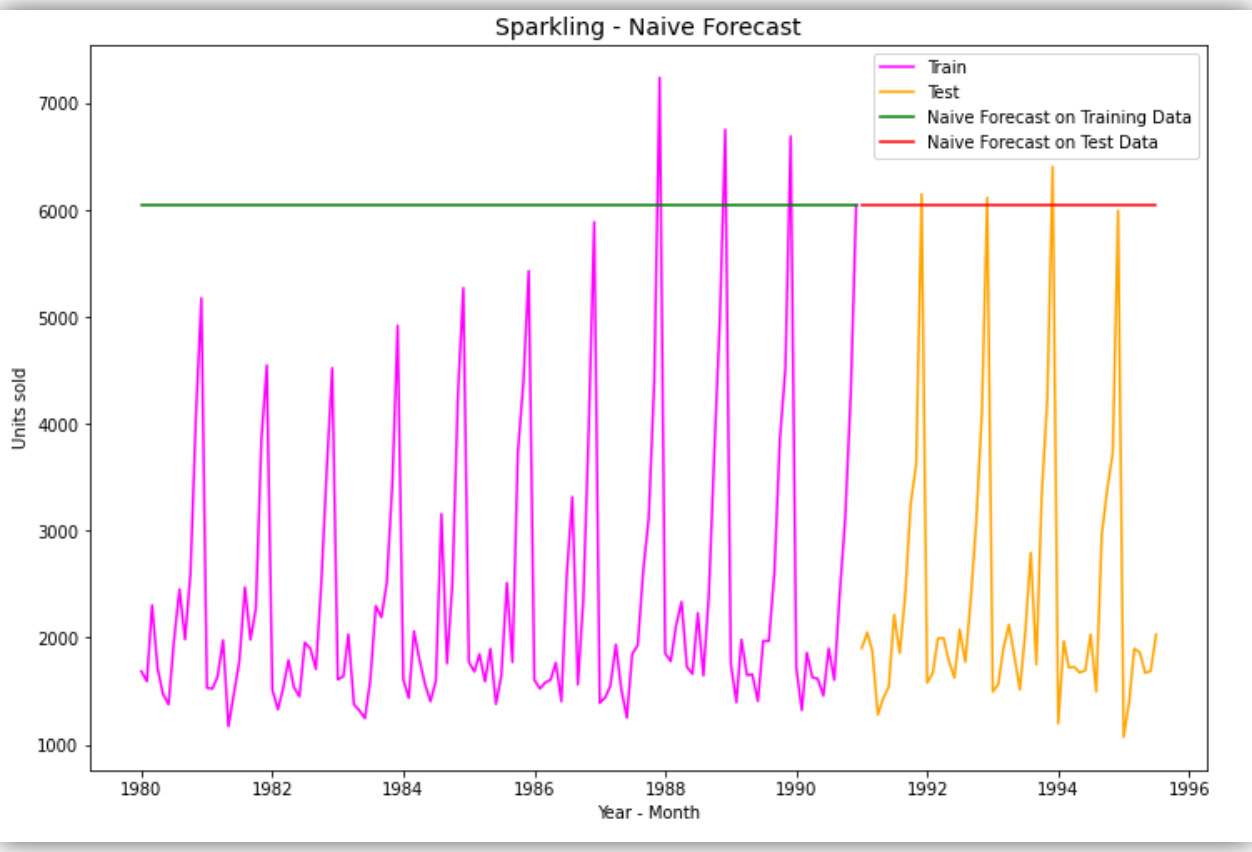


Figure 10 Naive Forecast Model

Model 3: Simple Average

The forecast is done using the mean of the time-series variable from the training set

- The model is not capable of either forecasting or able to capture the trend and seasonality present in the dataset.
- For Simple Average on the Test Data, RMSE is 1275

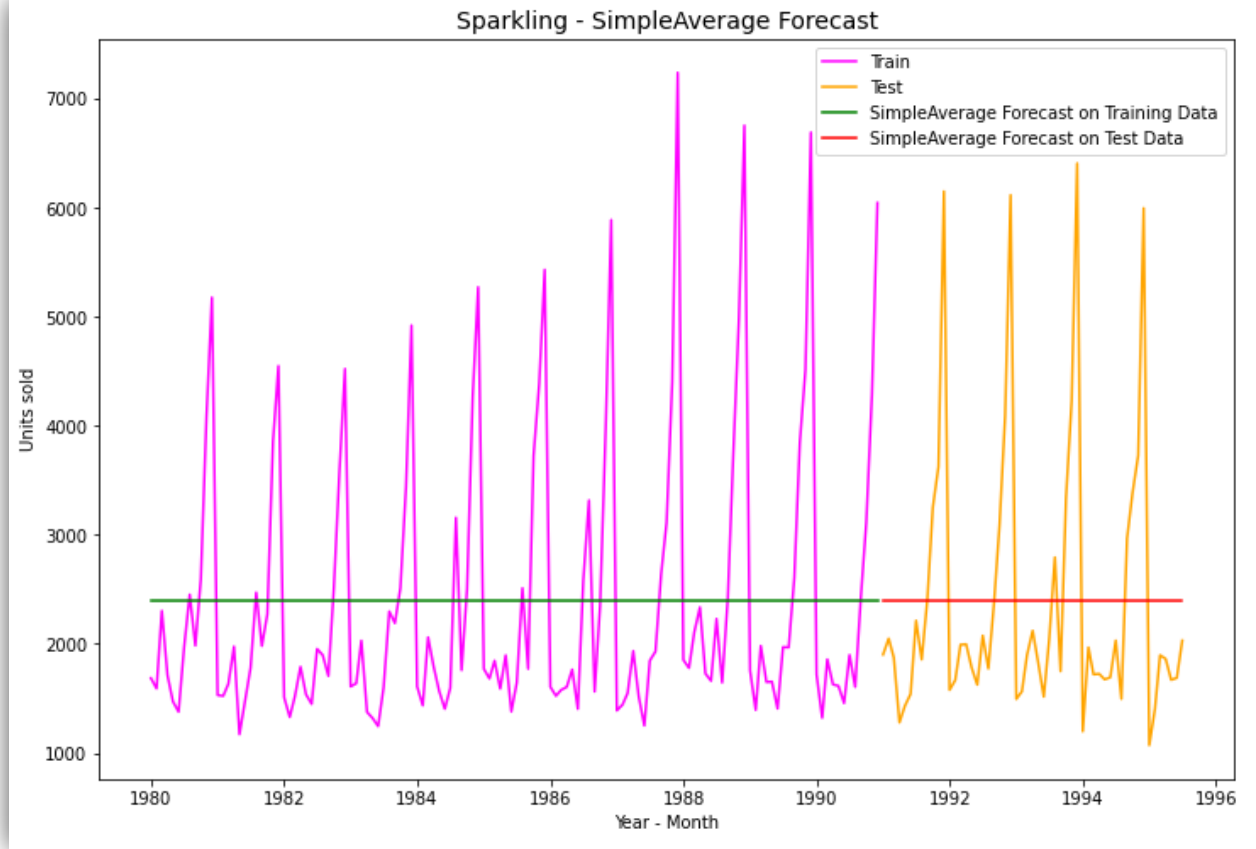


Figure 11 Simple Average model

Model 4: Moving Average Model

- For the moving average model, we will calculate rolling means (or trailing moving averages) for different intervals. The best interval can be determined by the maximum accuracy.
- The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points.
- For Sparkling dataset the accuracy is found to be higher with the lower rolling point averages.
- In moving average forecasts the values can be fitted with a delay of n number of points.
- The best interval of moving average from the model is 2 point.

YearMonth	Sparkling	Spark_Trailing_2	Spark_Trailing_4	Spark_Trailing_6	Spark_Trailing_9
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Figure 12 2,4 6 and 9 trailings for moving averages

- Calculating the rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

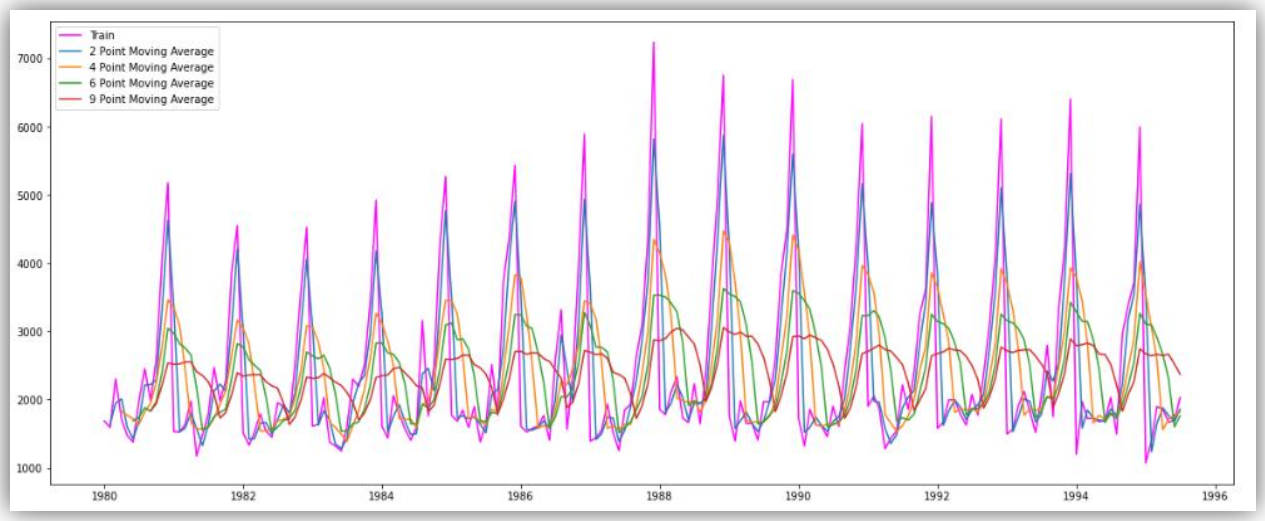


Figure 13 Simple Average Model

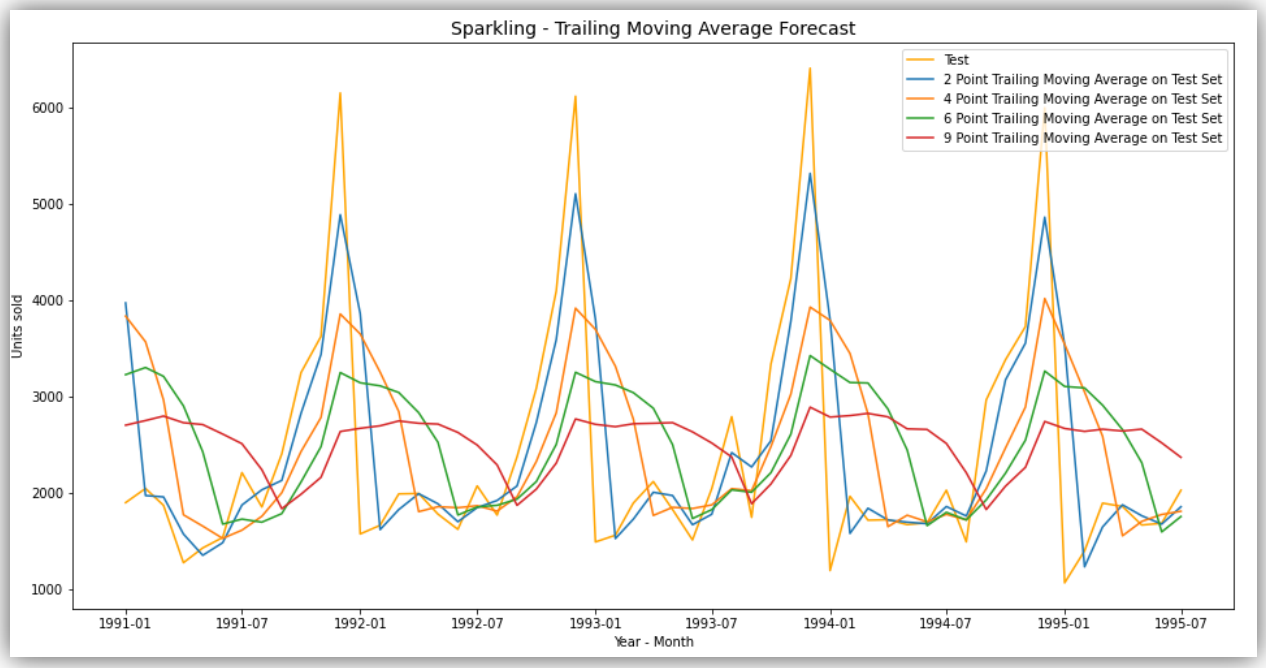


Figure 15 Trailing moving average forecast with best intervals of 2,4,6 and 9

```
For 2 point Moving Average Model forecast on the Test Data,  rmse_spark is 813.401
For 4 point Moving Average Model forecast on the Test Data,  rmse_spark is 1156.590
For 6 point Moving Average Model forecast on the Test Data,  rmse_spark is 1283.927
For 9 point Moving Average Model forecast on the Test Data,  rmse_spark is 1346.278
```

Figure 14 RMSE values of all trailing in moving average movel forecast

RMSE Values:

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverage	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315

Table 9 RMSE values

Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots.

Figure 16 Models comparison

Model 5: Simple Exponential Smoothing

- The model was ran without passing a value for alpha and used parameters: 'optimized=True, use_brute=True'.
- The auto-fit model picked up alpha = 0.0496 as the smoothing parameter.
- Simple Exponential Smoothing is applied if the time-series has neither a trend nor seasonality, which is not the case with the given data.
- The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually.
- For alpha value closer to 1, forecasts follows the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened
- For Sparkling, test RMSE is found to be higher for values closer to zero, which is same as in Simple average forecast.
- By passing manual alpha values, alpha =0.025 gives a better RMSE compared to optimized RMSE value.

YearMonth	Sparkling	predict
1991-01-01	1902	2724.932624
1991-02-01	2049	2724.932624
1991-03-01	1874	2724.932624
1991-04-01	1279	2724.932624
1991-05-01	1432	2724.932624

Table 10 predicted values

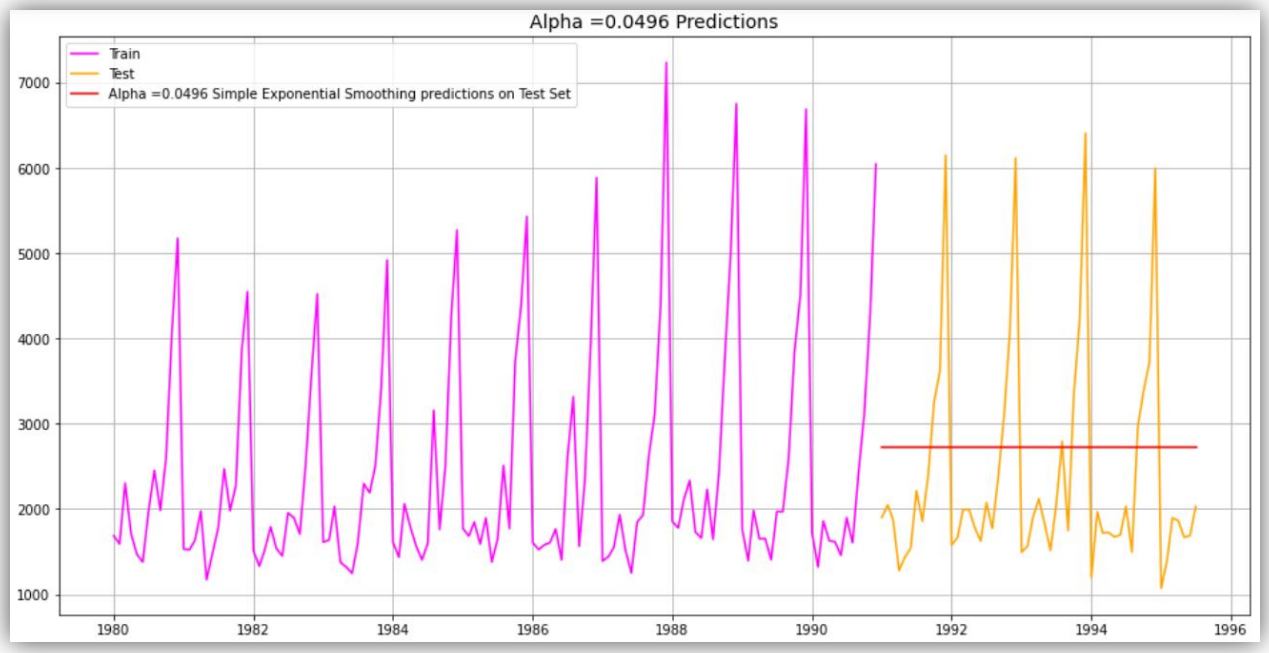


Figure 17 Alpha predicted curve

Iterative Method for Simple Exponential Smoothing

- The Double Exponential Smoothing models is applicable when data has trend, but no seasonality. Sparkling data contain slight trend component and very significant seasonality
- In first iteration, smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE values, which is as below with alpha 0.1 and beta 0.1
- On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'
- The auto-fit model retuned higher RMSE value compared to iterative alpha=0.1 and beta=0.1 RMSE value.

	Alpha Values	Train RMSE	Test RMSE
0	0.025	1322.084340	1286.248846
1	0.050	1318.429335	1316.411742
3	0.100	1333.873836	1375.393398
4	0.200	1356.042987	1595.206839
2	0.250	1359.701408	1755.488175

Figure 18 RMSE values for test and train data

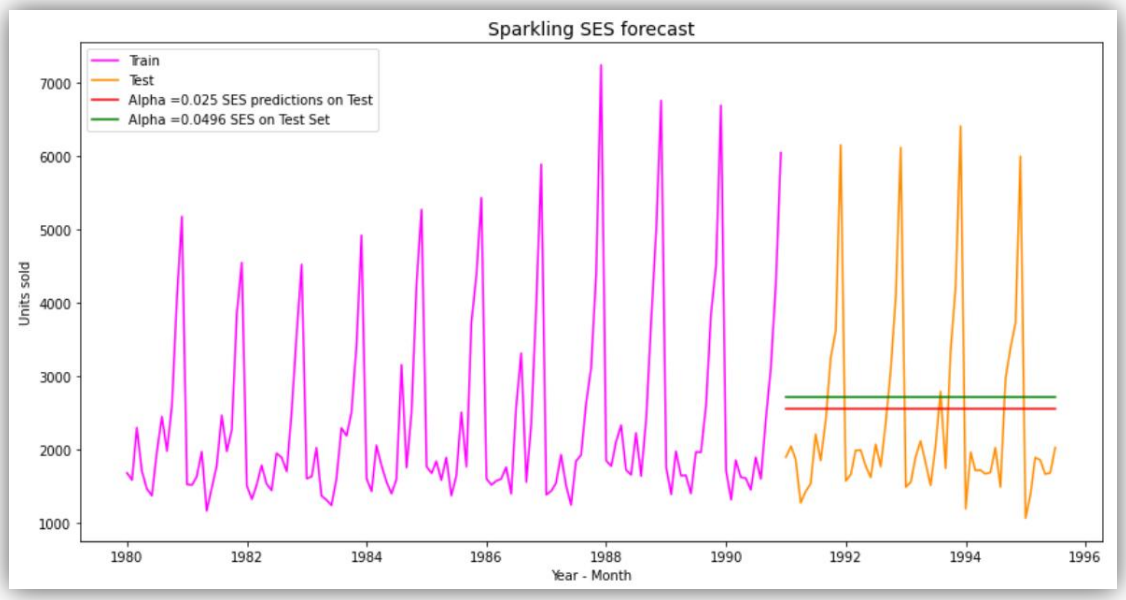


Figure 19 SES Optimised and iterative model

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

YearMonth	Sparkling	(predict_spark, 0.6885714285714285, 9.999999999999999e-05)
1991-01-01	1902	5221.278699
1991-02-01	2049	5127.886554
1991-03-01	1874	5034.494409
1991-04-01	1279	4941.102264
1991-05-01	1432	4847.710119

Table 11Viewing the first five Predictions for Test Data:

For Auto-fit Double Exponential Smoothing Model forecast on the Test Data, RMSE is 2007.23

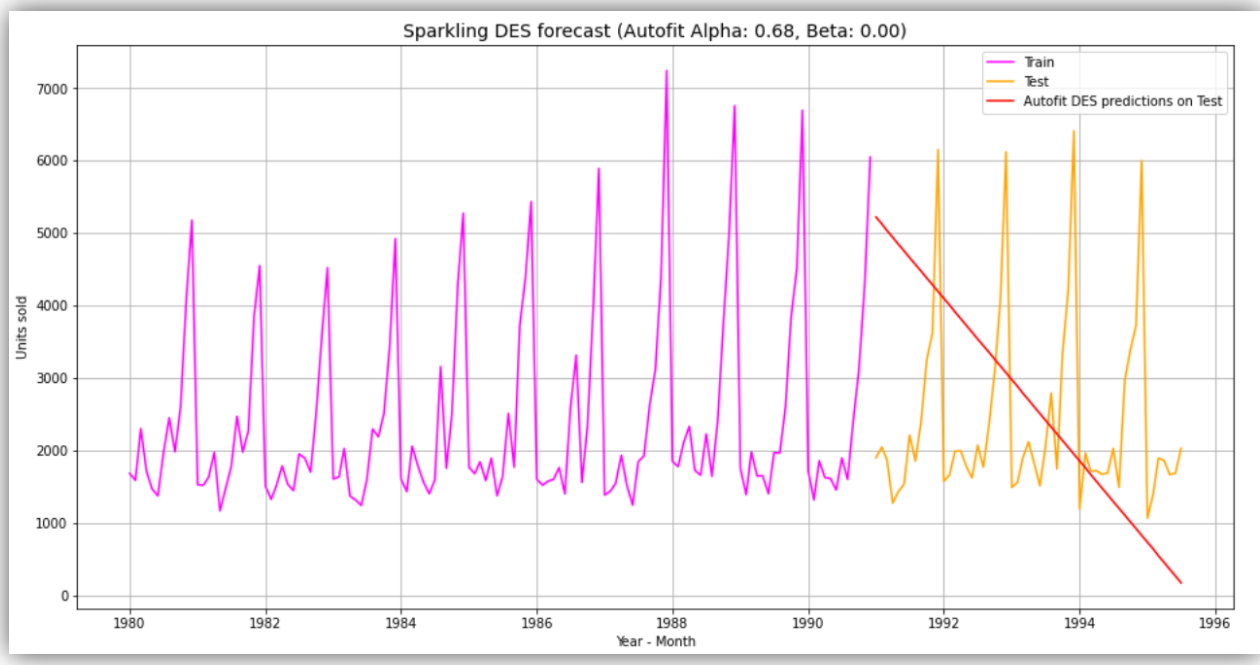


Figure 20 Plotting on both the Training and Test data for spaparkling DES fong forecasting

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

- The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Sparkling data contain slight trend and significant seasonality
- → On first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE values, which is as below with alpha 0.4, beta 0.1 and gamma 0.3
- → On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'
- → The auto-fit model retuned higher RMSE value compared to iterative alpha=0.4, beta=0.1 and gamma=0.3 RMSE value.

For Auto-fit Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 404.287 : –

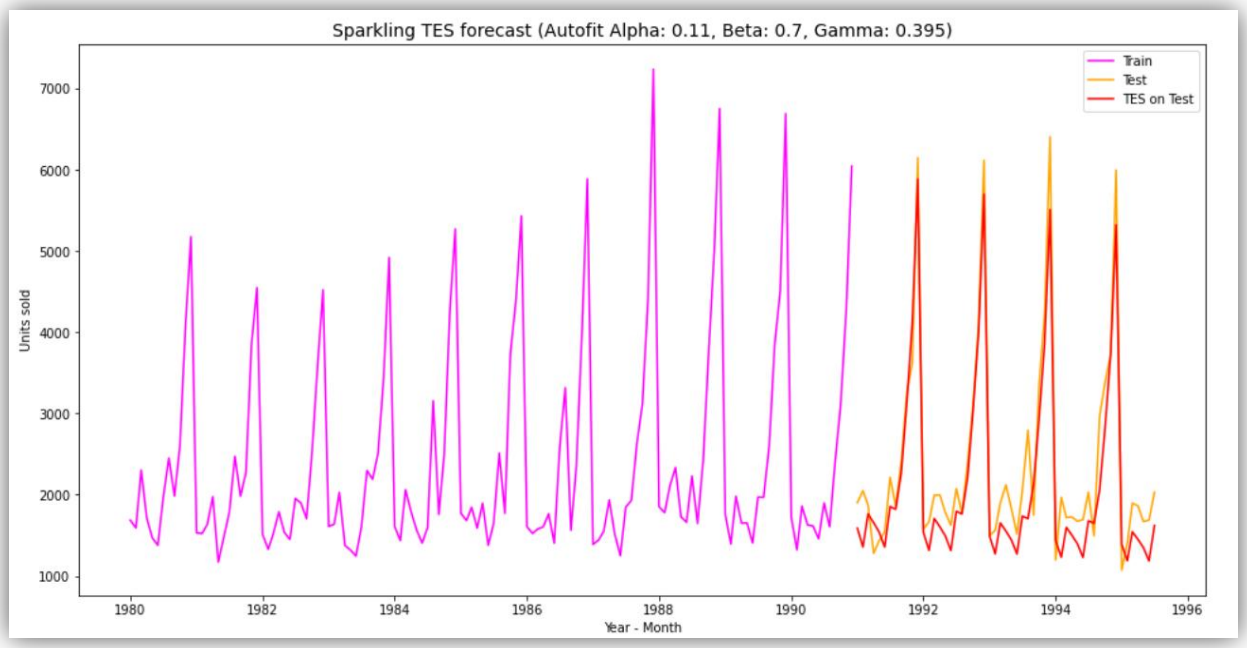


Figure 21 TES Auto-fit Model

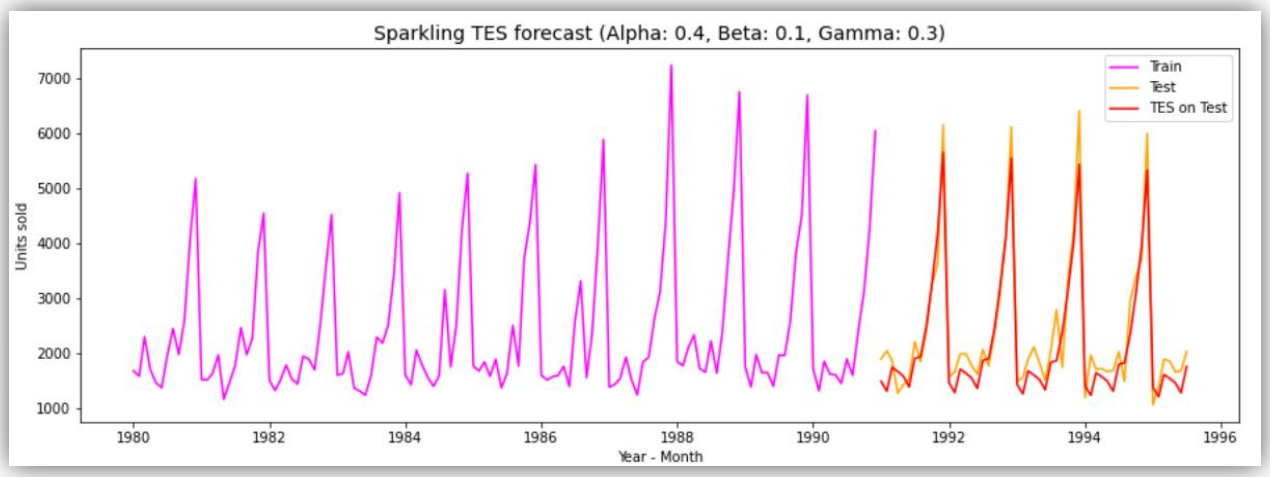


Figure 22 TES Iterstive model

Model comparison

Setting different alpha values. Higher the alpha, the more weightage is given to more recent observations:-

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverage	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha=0.0496, SES Optimized	1316.035487
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	404.286809

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverage	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha=0.0496, SES Optimized	1316.035487
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	404.286809
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	345.913415

	Alpha Values	Beta Values	Train RMSE	Test RMSE	Gamma Values
320	0.5	0.1	396.598057	345.913415	0.3
176	0.3	0.3	397.797318	361.397300	0.3
321	0.5	0.1	405.370169	365.288320	0.4
109	0.2	0.4	513.149439	368.374003	0.8
240	0.4	0.1	382.899740	376.708937	0.3

Table 12 ALPHA< BETA AND GAMMA VALUES

	Test RMSE
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	345.913415
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	404.286809
2 point TMA	813.400684
4 point TMA	1156.589694
SimpleAverage	1275.081804
6 point TMA	1283.927428
Alpha=0.025,SES iterative	1286.248846
Alpha=0.025,SES iterative	1286.248846
Alpha=0.0496, SES Optimized	1316.035487
9 point TMA	1346.278315
RegressionOnTime	1389.135175
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
NaiveModel	3864.279352

Table 13 Test RMSE values

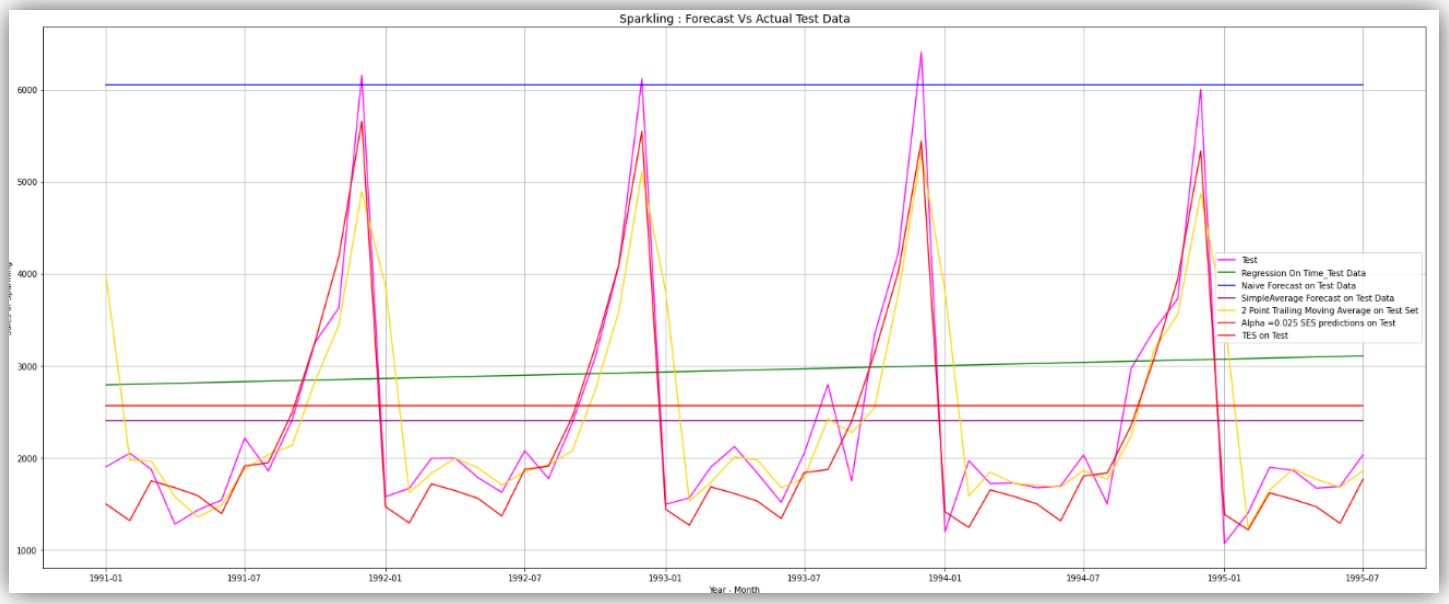


Figure 23Sparkling forecasting vs actual forecasting

- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data.
- point trailing moving average model is also found to have fit well with a slight lag in test dataset.

1.5 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.

- Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determine the presence of unit root in the series to understand if the series is stationary or not
- → Null Hypothesis: The series has a unit root, that is series is non-stationary
- → Alternate Hypothesis: The series has no unit root, that is series is stationary
- → If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary
- → The ADF test on the original Sparkling series retuned the below values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis.
- → Differencing of order one is applied on the Sparkling series as below and tested for stationarity. At an order of differencing 1, the series is found to be stationary as below
- → The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if it's multiplicative or additive in character.
- → The altitude of rolling mean and std dev is seen changing according to change in slope, which indicates multiplicity.

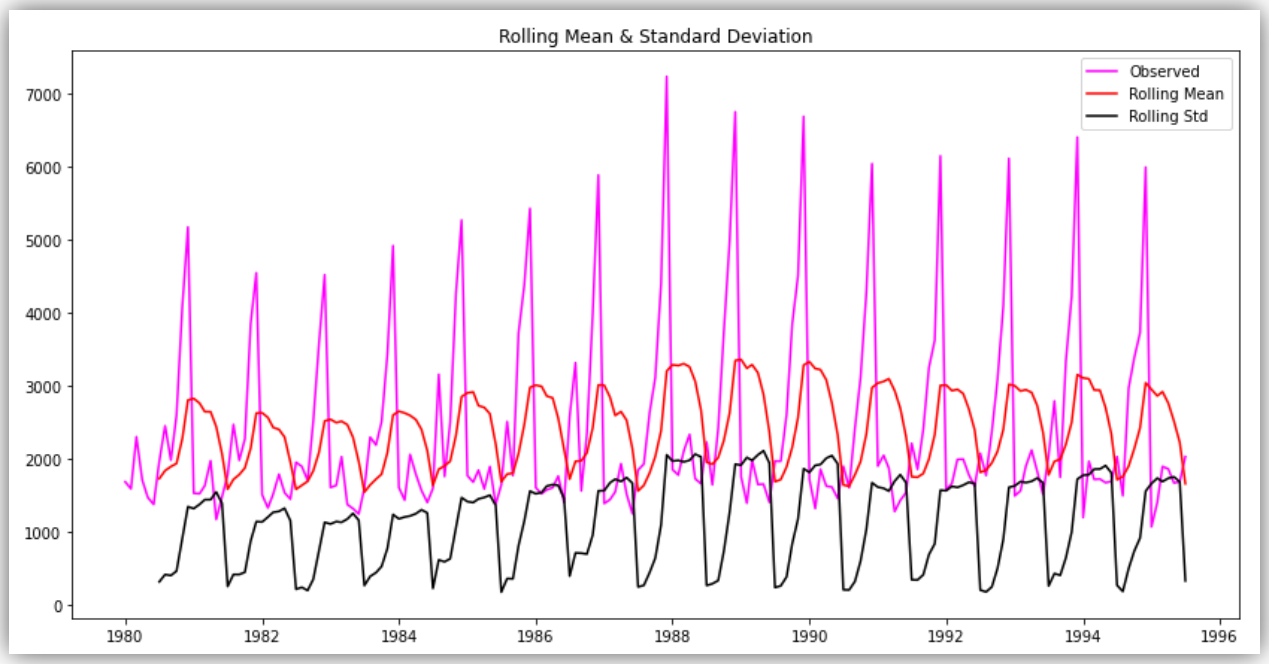


Figure 24 ADF test on original series

We see that at 5% significant level the Time Series is non-stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not. Difference of original series Seasonality is multiplicative as the Std deviation and mean varies according to the change in trend.

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

Table 14 Dickey fuller test results

- Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determine the presence of unit root in the series to understand if the series is stationary or not
- Null Hypothesis: The series has a unit root, that is series is non-stationary
- Alternate Hypothesis: The series has no unit root, that is series is stationary
- If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary
- The ADF test on the original Sparkling series retuned the below values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis.
- We see that at alpha= 0.05 the Time Series is indeed stationary. d=1

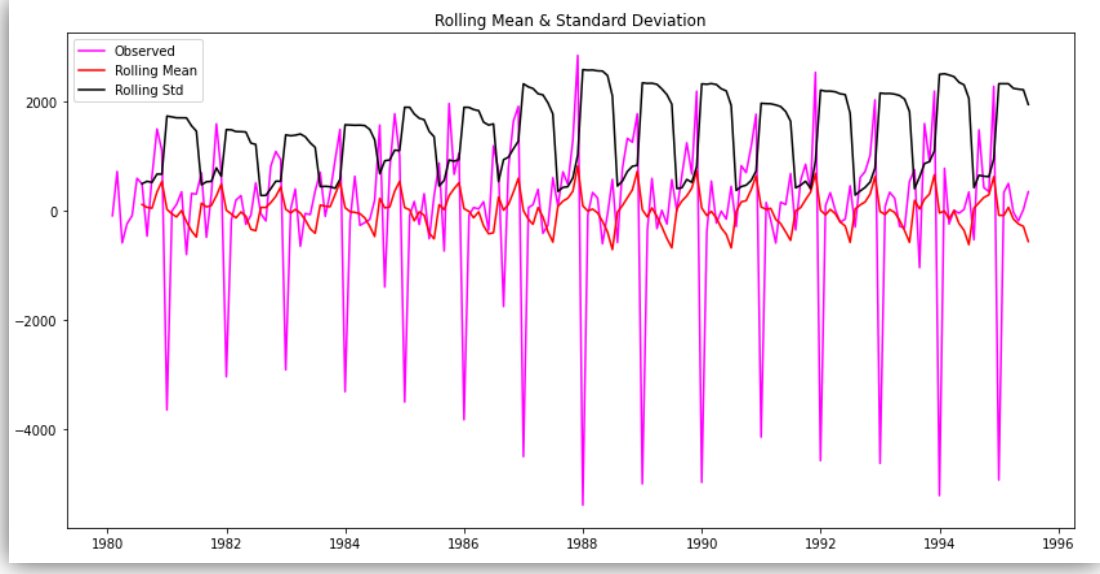


Figure 25 ADF Test after differencing d=1

If the series is non-stationary, stationarize the Time Series by taking a difference of the Time Series. Then we can use this particular differenced series to train the ARIMA/SARIMA models. We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there.Also we can look at other kinds of transformations as part of making the time series stationary like taking logarithms.

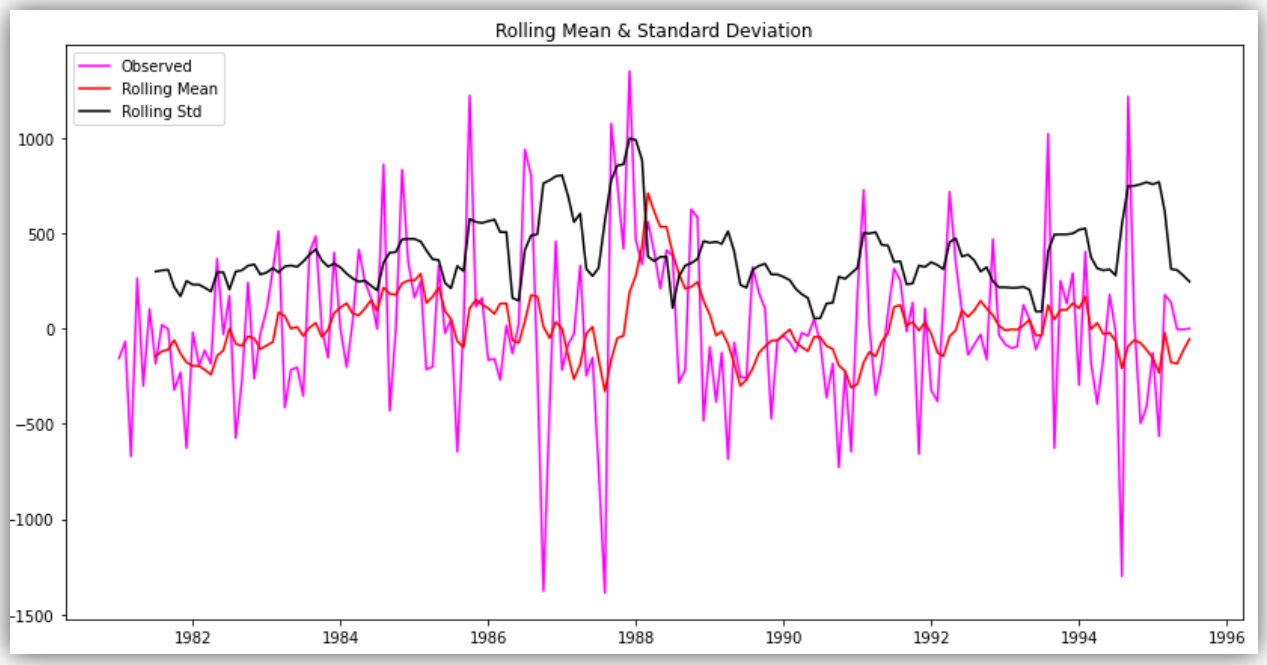


Figure 26 ADF test on log series after differencing

```
Results of Dickey-Fuller Test:
Test Statistic                  -4.460165
p-value                          0.000232
#Lags Used                      11.000000
Number of Observations Used    163.000000
Critical Value (1%)             -3.471119
Critical Value (5%)             -2.879441
Critical Value (10%)            -2.576314
dtype: float64
```

Figure 27 Dickey fuller test result after differencing

We see that at = 0.05 the Time Series is indeed stationary. But seasonality is multiplicative

Checking for stationarity of the Training Data Time Series

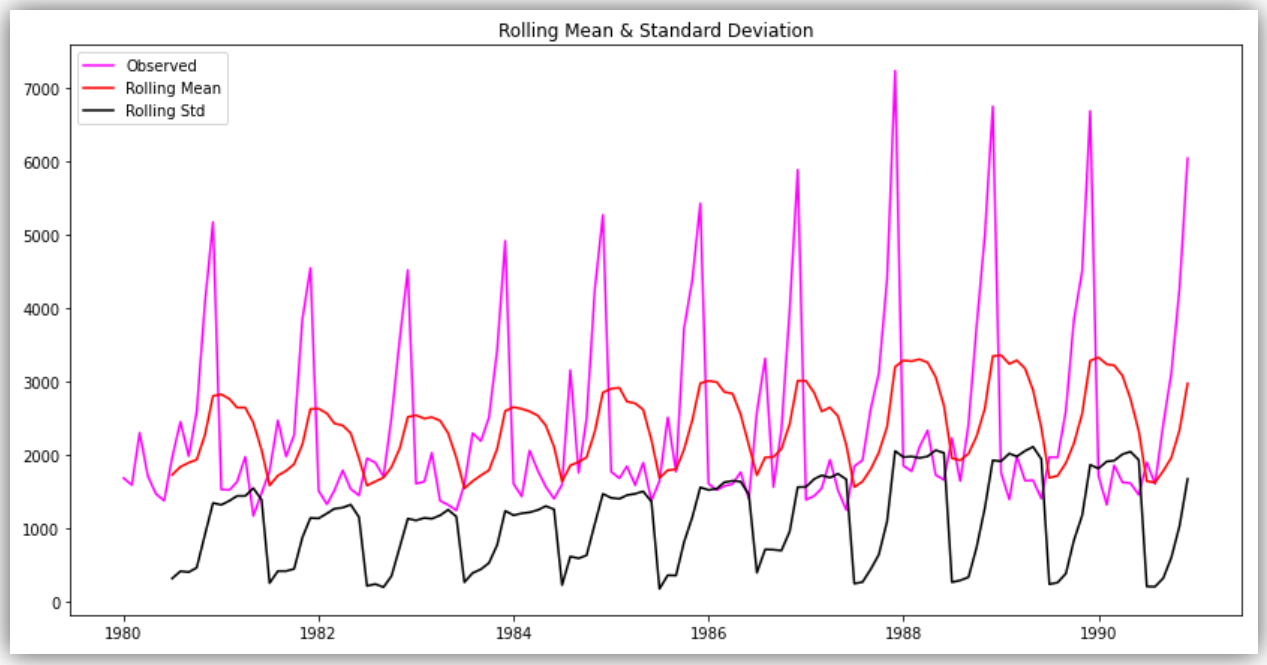


Figure 29 Graph with no stationarity

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

Figure 28 Dickey fuller test result for stationarity graph

- We see that at 5% significant level the Time Series is non-stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not.
- We see that at alpha = 0.05 the Time Series is indeed stationary. d=1

1.6 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.

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sun, 17 Apr 2022	AIC	2213.509
Time:	13:37:31	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	28.781	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.741	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.218	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-07	5.51e+12	0.000	1.1e+06	1.1e+06

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

Table 15 Auto-ARIMA Model

For Auto-ARIMA Model forecast accuracy_score on the Test Data, RMSE is 1299.980

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverage	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha=0.0496, SES Optimized	1316.035487
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	404.286809
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	345.913415
Auto_ARIMA(2,1,2)	1299.979640

Table 16 Auto - ARIMA Model results

- ARIMA model was built with optimised model and found the least AIC value =2210.62 at (2, 1, 2).
- As the Sparkling series of data contain seasonality component, ARIMA model do not perform well. The RMSE value for this Auto- ARIMA model is 1375.

Dep. Variable:		y	No. Observations:		132
Model:	SARIMAX(1, 1, 2)x(0, 1, 2, 12)		Log Likelihood		-685.174
Date:	Sun, 17 Apr 2022		AIC		1382.348
Time:	13:39:54		BIC		1397.479
Sample:	0		HQIC		1388.455
	- 132				
Covariance Type:	opg				

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5507	0.287	-1.922	0.055	-1.112	0.011
ma.L1	-0.1612	0.235	-0.687	0.492	-0.621	0.299
ma.L2	-0.7218	0.175	-4.132	0.000	-1.064	-0.379
ma.S.L12	-0.4062	0.092	-4.401	0.000	-0.587	-0.225
ma.S.L24	-0.0274	0.138	-0.198	0.843	-0.298	0.243
sigma2	1.705e+05	2.45e+04	6.956	0.000	1.22e+05	2.19e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	13.48
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	0.89	Skew:	0.60
Prob(H) (two-sided):	0.75	Kurtosis:	4.44

Table 17Auto SARIMA Model

- The model was built on train data with seasonality 12 and with different optimal parameters (p, d, q)x(P, D, Q) parameters, the lowest AIC is 1382.35 was obtained at (1, 1, 2)x(0, 1, 2, 12).
- The model was built with the above parameters.
- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point forms roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
- The RMSE values of the automated SARIMA model is 382.58

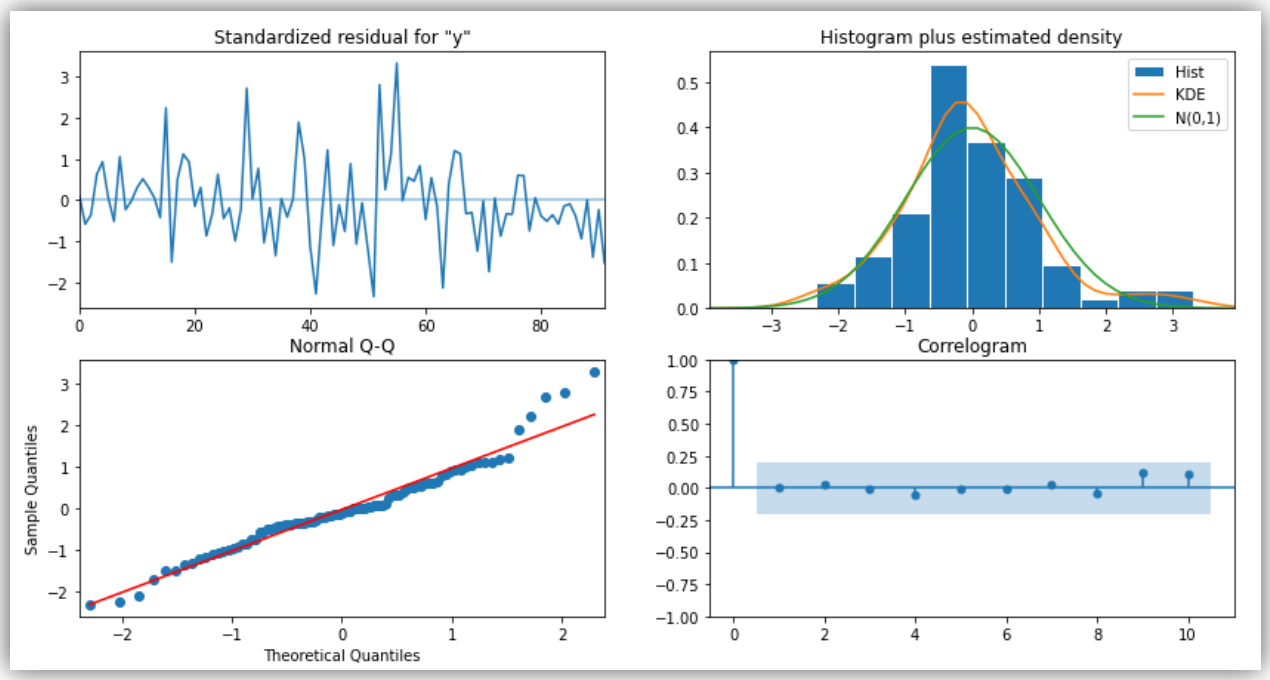


Figure 30 Diagnostic plot

YearMonth	Sparkling	spark_forecasted
1991-01-01	1902	1460.244621
1991-02-01	2049	1392.437156
1991-03-01	1874	1743.201695
1991-04-01	1279	1650.066918
1991-05-01	1432	1522.656020

Table 18 Forecasted results on test data

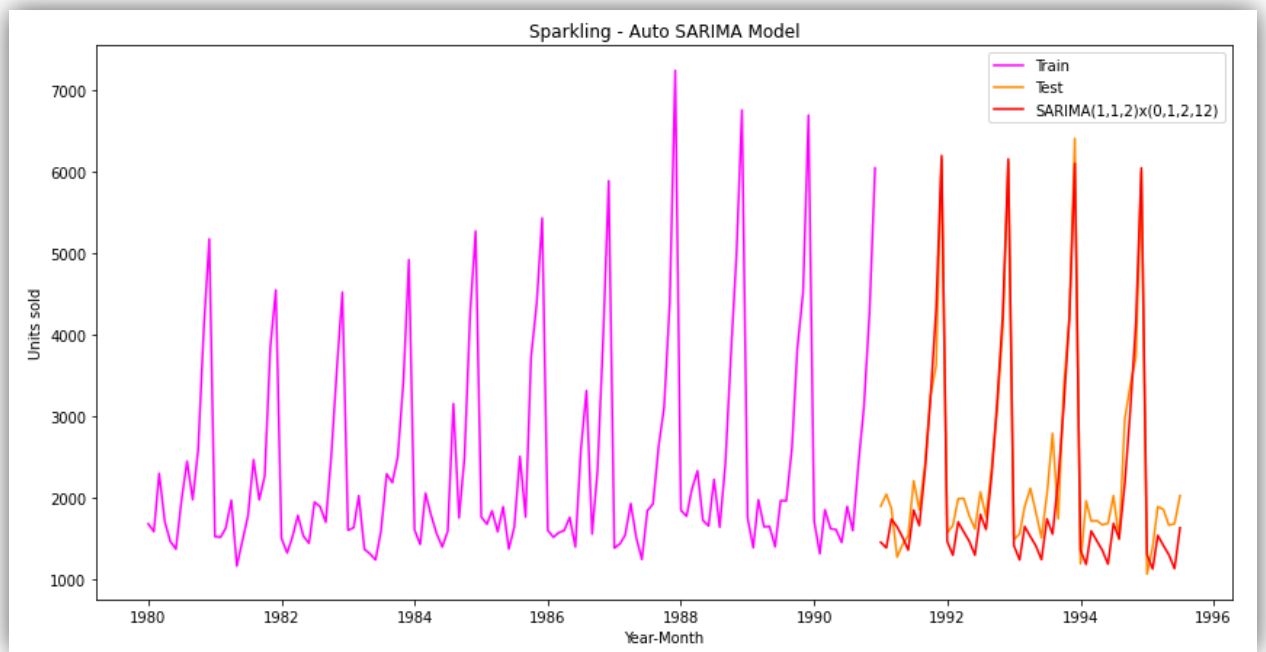


Figure 31 Plot of actual vs forecasted results on the test data on auto sarima model

Auto – Sarima model on log series data

Dep. Variable:			Sparkling	No. Observations:		132
Model:	SARIMAX(0, 1, 1)x(1, 0, 1, 12)			Log Likelihood		146.236
Date:			Sun, 17 Apr 2022	AIC		-284.472
Time:			13:44:10	BIC		-273.423
Sample:			01-01-1980	HQIC		-279.986
			- 12-01-1990			
Covariance Type:			opg			
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8966	0.045	-19.863	0.000	-0.985	-0.808
ar.S.L12	1.0112	0.020	49.871	0.000	0.971	1.051
ma.S.L12	-0.6489	0.075	-8.629	0.000	-0.796	-0.502
sigma2	0.0045	0.001	7.842	0.000	0.003	0.006

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	5.26
Prob(Q):	0.74	Prob(JB):	0.07
Heteroskedasticity (H):	1.43	Skew:	-0.00
Prob(H) (two-sided):	0.27	Kurtosis:	4.04

Figure 32 Log series SARIMA model results

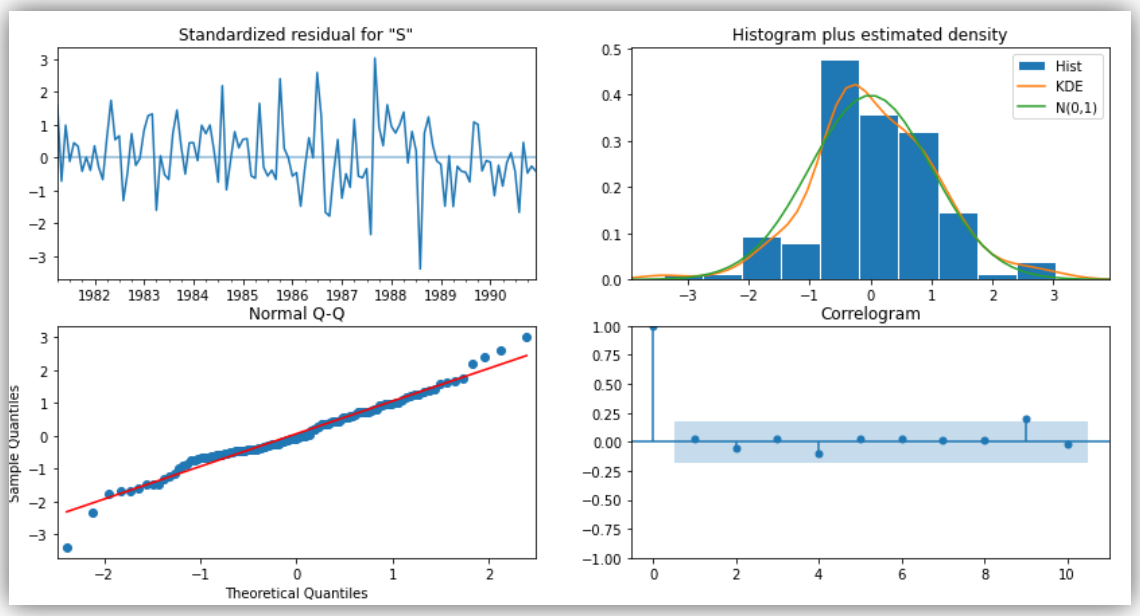


Figure 33 diagnostic plot

The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.

- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point forms roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
- From the above model summary it can be inferred that MA.L1, AR.L.S12, MA.L.S12 terms has the highest absolute weightage.
- From the p-values it can be inferred that terms MA.L1, AR.L.S12, MA.L.S12 are significant terms, as their values are below 0.05.
- The RMSE values of the automated SARIMA of log series model is 336.58

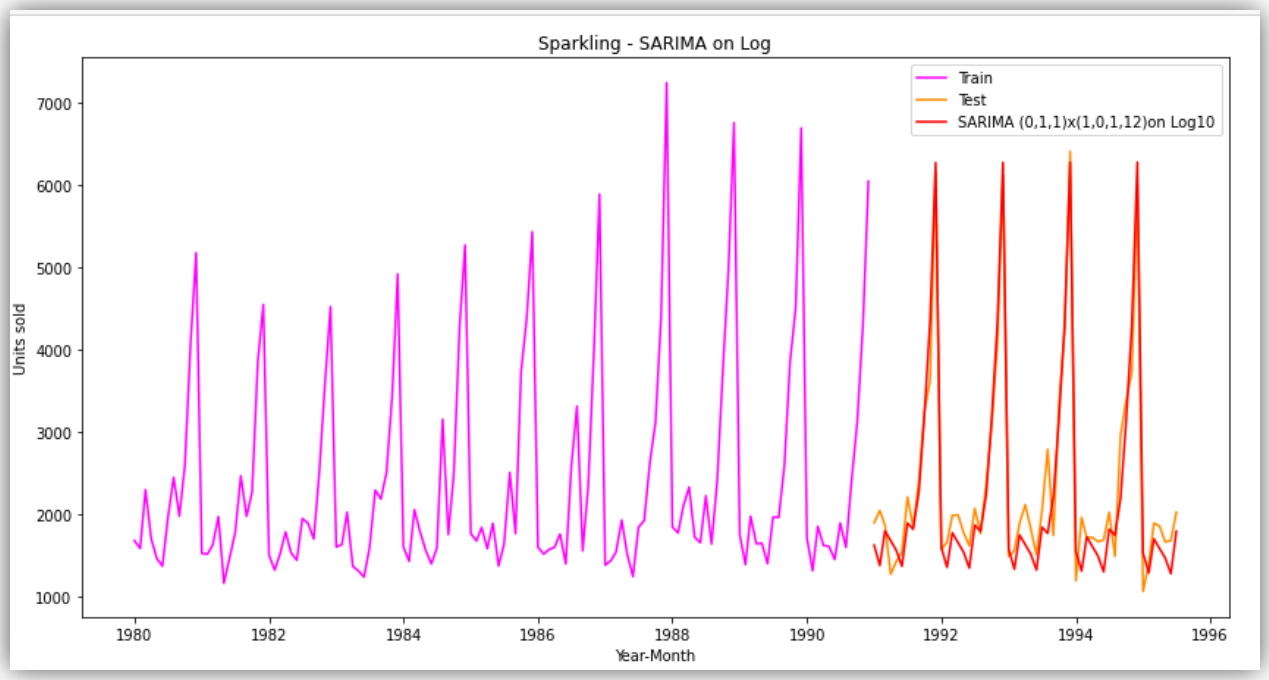


Figure 34 Plot of actual vs forecasted results on the test data on sarima model

The model built with log series data has a lower RMSE value when compared to original train data. For Auto-SARIMA_log Model forecast on the Test Data, RMSE is 336.799

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

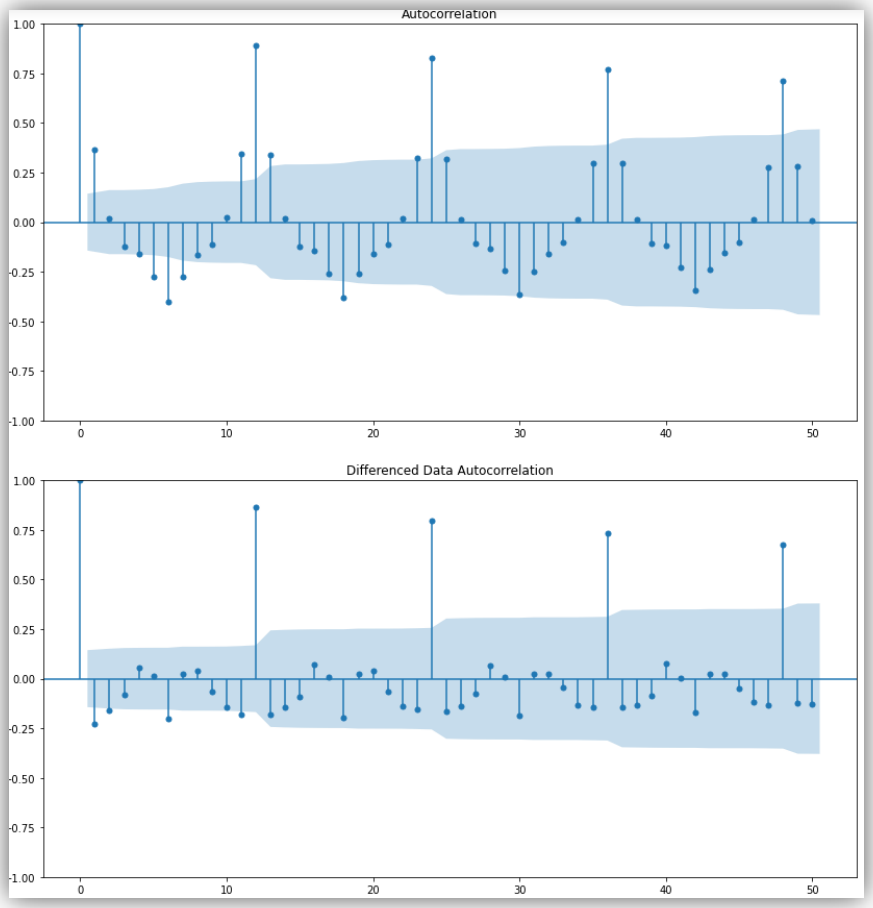
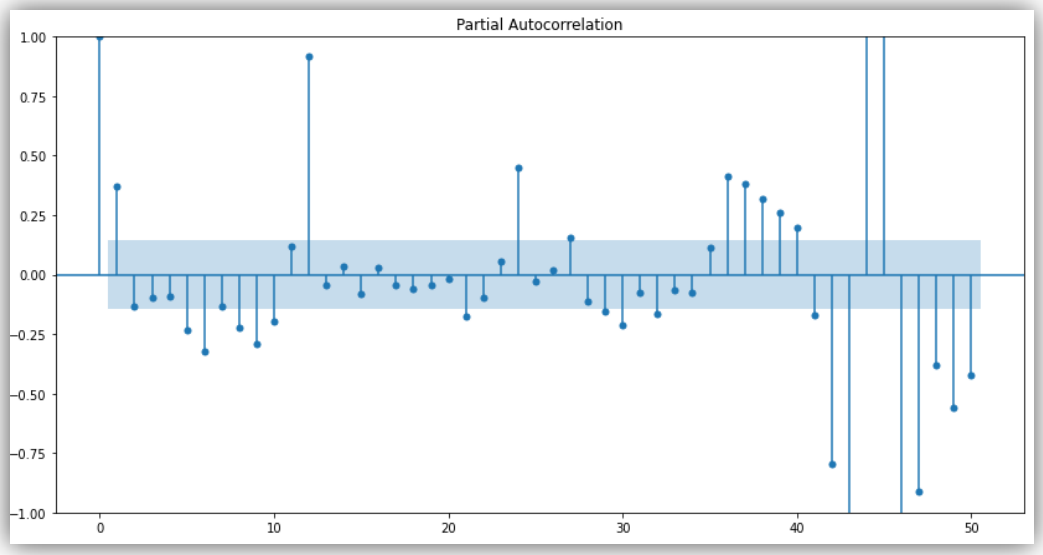


Figure 35 ACF PLOTS

- Here, we have taken alpha=0.05.
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
- By looking at above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

Check for stationarity of the Training Data Time Series.



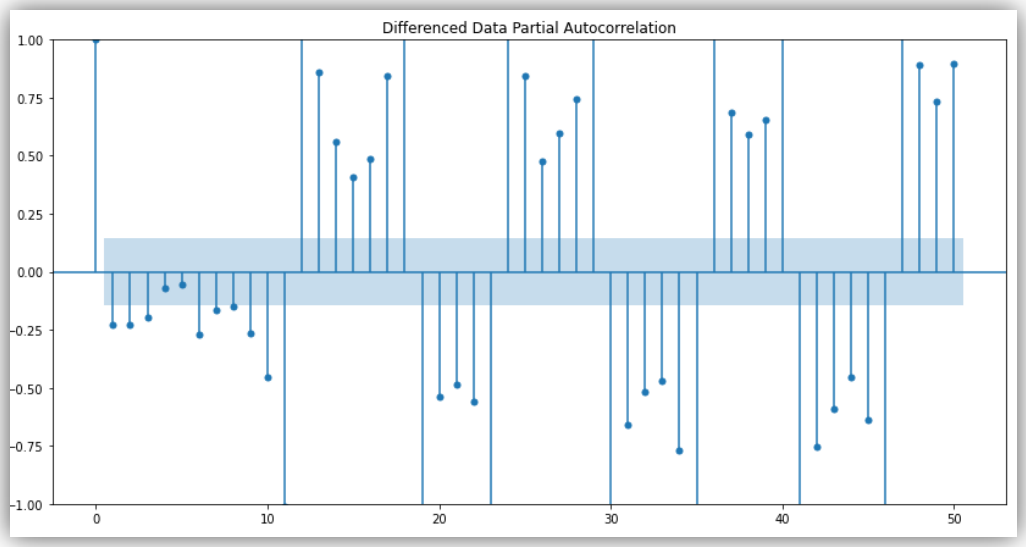


Figure 36 Time series plots

Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

We get a comparatively simpler model by looking at the ACF and the PACF plots.

Note: When we see that both the AR(p) and the MA(q) model are of order 0, we have to convert the input variable into a 'float64' type variable else Python might throw an error.

SARIMAX Results

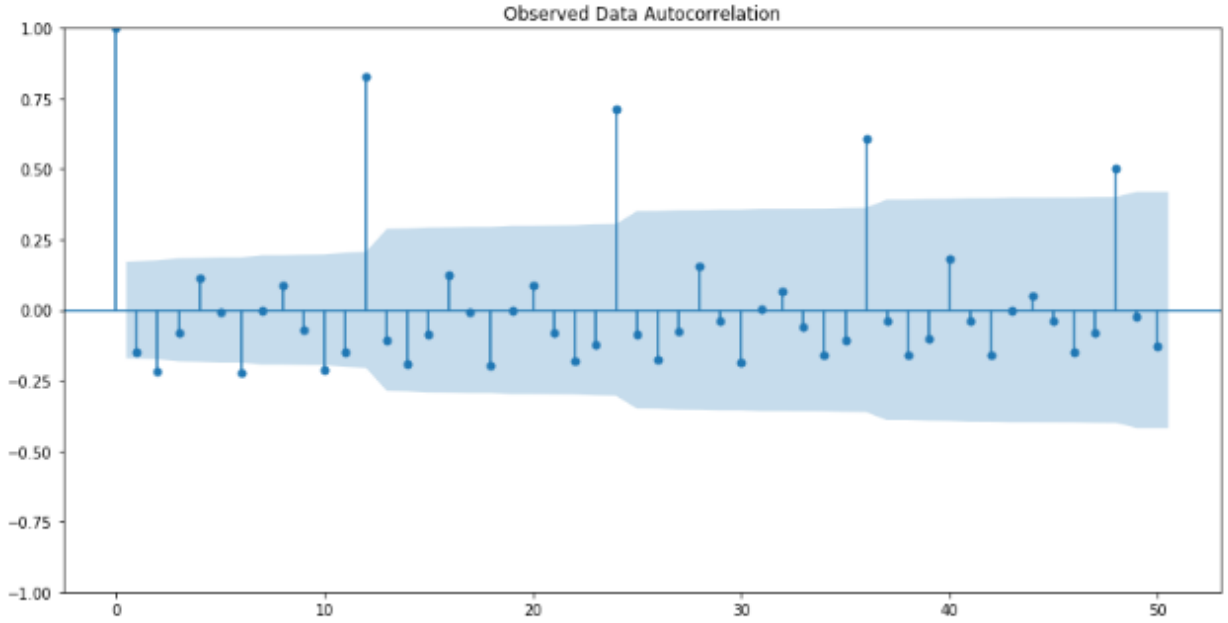
Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.832
Date:	Sun, 17 Apr 2022	AIC	2267.663
Time:	13:44:13	BIC	2270.538
Sample:	01-01-1980	HQIC	2268.831
	- 12-01-1990		
Covariance Type:	opg		

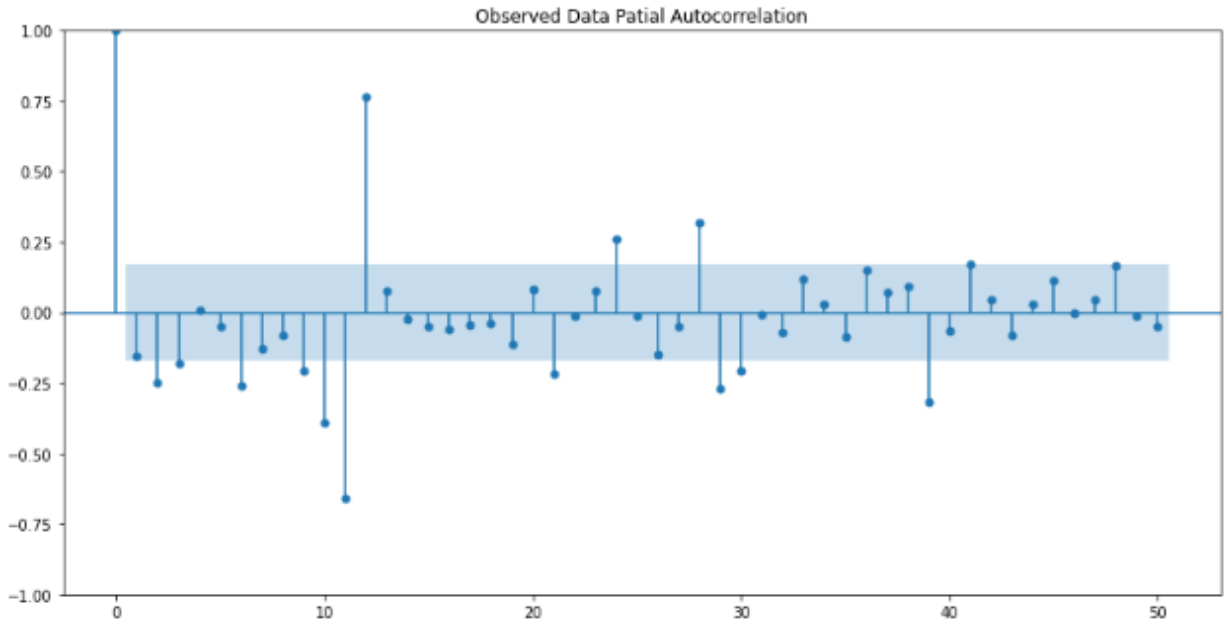
	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.885e+06	1.29e+05	14.658	0.000	1.63e+06	2.14e+06

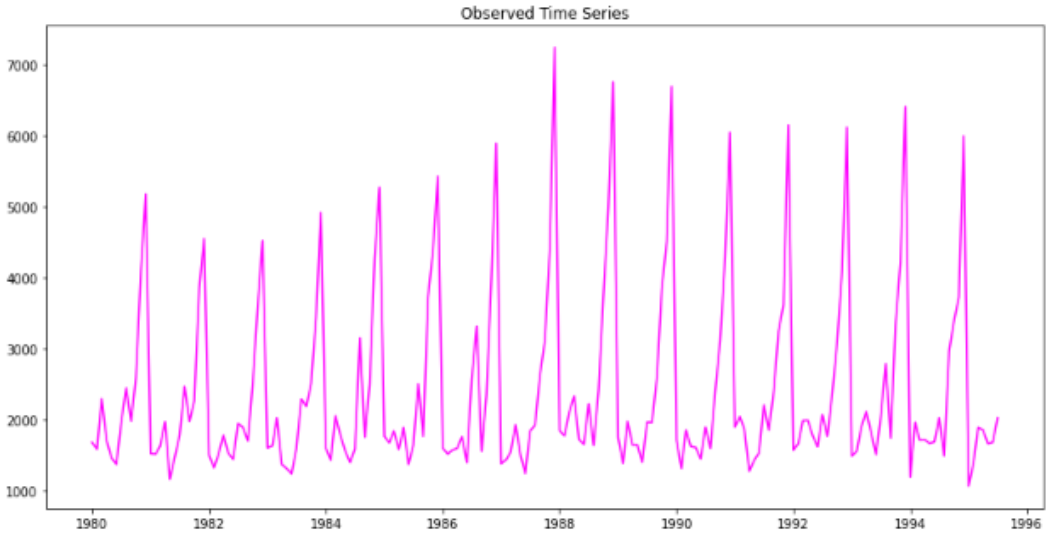
Ljung-Box (L1) (Q):	3.07	Jarque-Bera (JB):	198.83
Prob(Q):	0.08	Prob(JB):	0.00
Heteroskedasticity (H):	2.46	Skew:	-1.92
Prob(H) (two-sided):	0.00	Kurtosis:	7.65

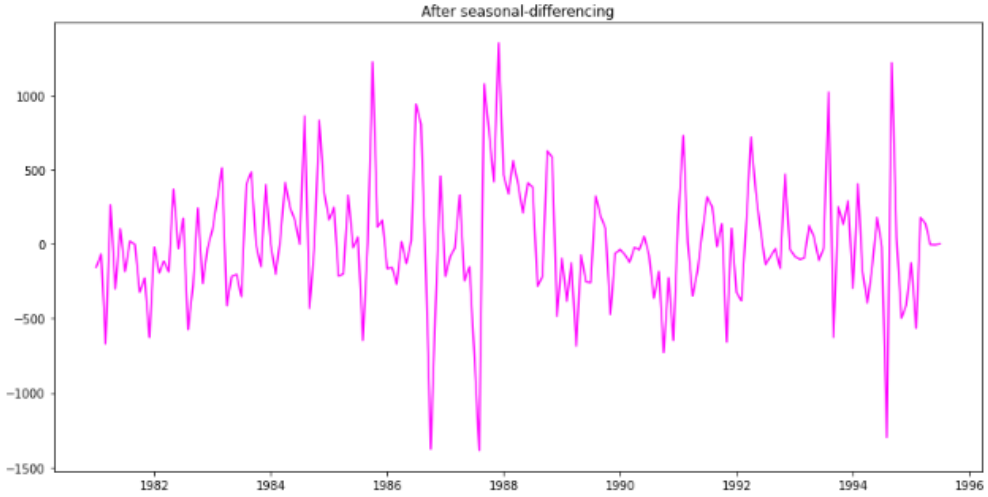
For Manual-ARIMA Model forecast on the Test Data, RMSE is 3864.279

Manual ARIMA









The marginal trend in the data is still seen

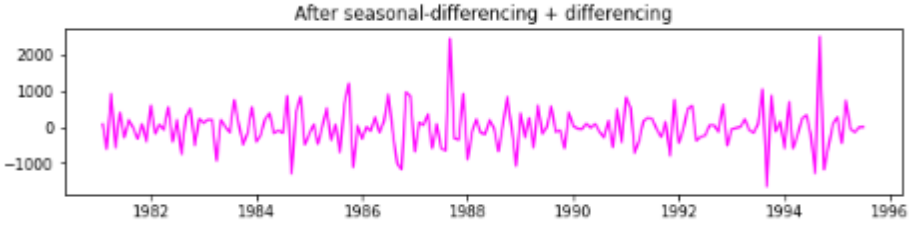


Figure 37 Time series Plots

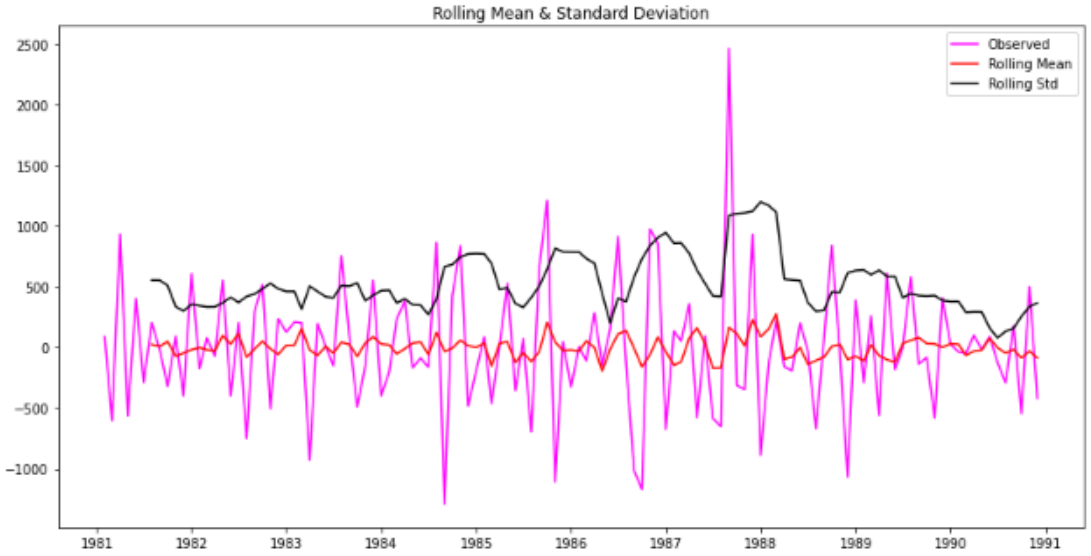


Figure 38 ADF TEST

- From the ACF plot of the observed/ train data, it can be inferred that at seasonal interval of 12, the plot is not quickly tapering off. So a seasonal differencing of 12 has to be taken
- From the plots above an apparent slight trend is still existing after differencing of seasonal order of 12. With a further differencing of order one, no trend is present.
- An ADF test need to be done to check the stationarity after the above differencing. With a p-value below alpha 0.05 and test statistic below critical values, it can be confirmed that the data is stationary.
- ACF and PACF plots of the seasonal-differenced + one order differenced data is created to find the values for (p,d,q)x(P,D,Q).

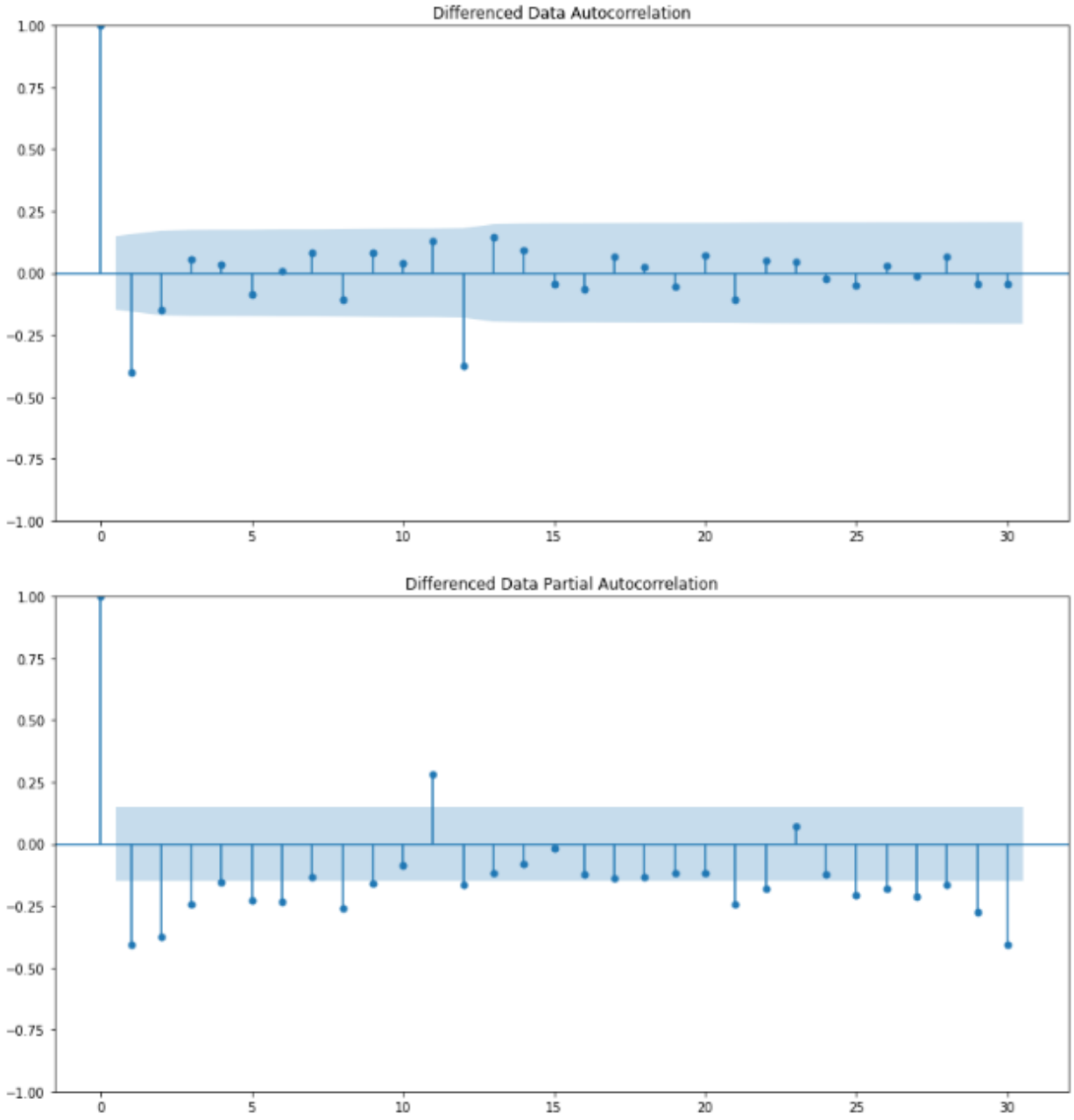


Figure 39 ACF AND PACF PLOTS

SARIMAX Results

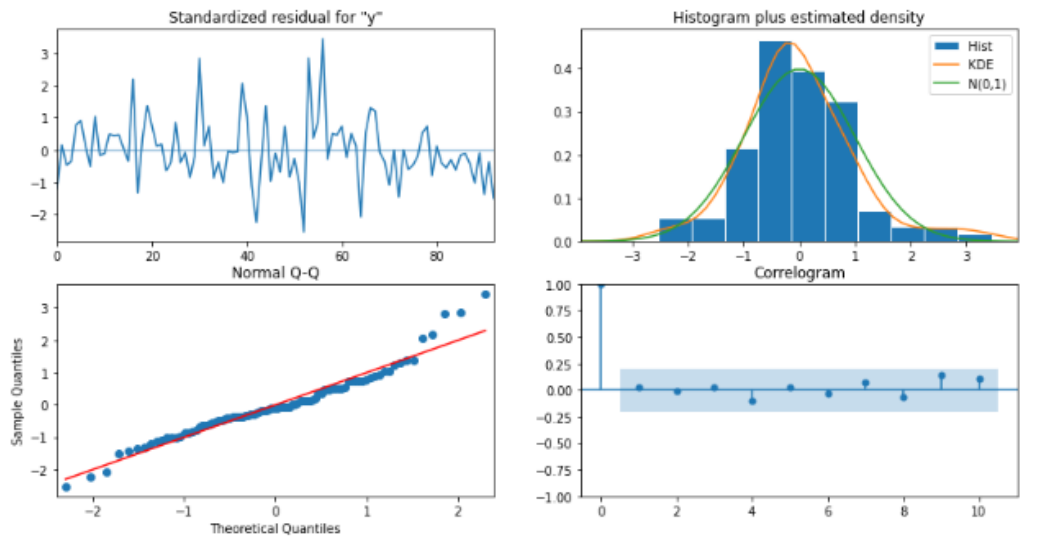
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-693.697
Date:	Sun, 17 Apr 2022	AIC	1403.394
Time:	14:04:08	BIC	1423.654
Sample:	0	HQIC	1411.574
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2229	0.130	1.713	0.087	-0.032	0.478
ar.L2	-0.0798	0.131	-0.607	0.544	-0.337	0.178
ar.L3	0.0921	0.122	0.756	0.450	-0.147	0.331
ma.L1	-1.0241	0.094	-10.925	0.000	-1.208	-0.840
ar.S.L12	-0.1992	0.866	-0.230	0.818	-1.897	1.499
ma.S.L12	-0.2109	0.881	-0.239	0.811	-1.938	1.516
ma.S.L24	-0.1299	0.381	-0.341	0.733	-0.877	0.617
sigma2	1.654e+05	2.62e+04	6.302	0.000	1.14e+05	2.17e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	19.66
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.81	Skew:	0.69
Prob(H) (two-sided):	0.56	Kurtosis:	4.78

Figure 40 SARIMAX RESULTS

- Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till 3rd lag it's significant before cut-off, so AR term 'p = 3' is chosen. At seasonal lag of 12, it almost cuts off, so seasonal AR 'P = 1'
- From ACF plot it can be seen that lag 1 is significant before it cuts off, so MA term 'q =1' is selected and at seasonal lag of 12, a significant lag is apparent, so kept seasonal MA term 'Q = 1' initially.
- The seasonal MA term 'Q' was later optimized to 2, by validating model performance, as the data might be under-differenced.
- The final selected terms for SARIMA model is (3, 1, 1)*(1,1,2,12).
- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point forms roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
- The RMSE values of the automated SARIMA model is 324.10



y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1579.909525	416.594338	763.399626	2396.419423
1	1419.153441	429.113891	578.106061	2260.200820
2	1868.142987	429.104332	1027.113950	2709.172025
3	1731.471376	430.973026	886.779766	2576.162986
4	1659.821743	431.906018	813.301503	2506.341984

YearMonth	Sparkling	spark_forecasted	spark_forecasted_log	manual_spark_forecasted
1991-01-01	1902	1460.244621	1629.418684	1579.909525
1991-02-01	2049	1392.437156	1384.549093	1419.153441
1991-03-01	1874	1743.201695	1804.208809	1868.142987
1991-04-01	1279	1650.066918	1685.516569	1731.471376
1991-05-01	1432	1522.656020	1569.599978	1659.821743

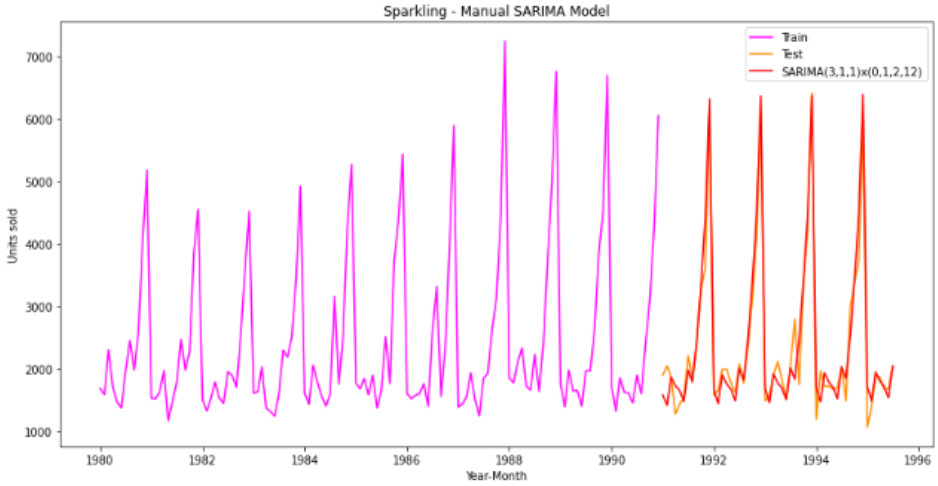


Figure 41 Actual plot vs forecast results on test data

1.8 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
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverage	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha=0.0496, SES Optimized	1316.035487
Alpha=0.025,SES iterative	1286.248848
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.025,SES iterative	1286.248846
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	404.286809
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	345.913415
Auto_ARIMA(2,1,2)	1299.979640
Auto_SARIMA(1, 1, 2)*(0, 1, 2, 12)	382.576734
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	336.799059
Manual_ARIMA(0,1,0)	3864.279352
Manual_SARIMA#(3,1,1)*(1,1,2,12)	324.106510

Figure 42 RMSE Values

Manual SARIMA (3,1,1)*(1,1,2,12) is found to be the best model, followed by Auto_SARIMA model

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

	Test RMSE
Manual_SARIMA#(3,1,1)*(1,1,2,12)	324.106510
Manual_SARIMA#(3,1,1)*(1,1,2,12)	324.106510
Manual_SARIMA#(3,1,1)*(1,1,2,12)	324.106510
Manual_SARIMA#(3,1,1)*(1,1,2,12)	324.106510
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	336.799059
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	345.913415
Auto_SARIMA(1, 1, 2)*(0, 1, 2, 12)	382.576734
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	404.286809
2 point TMA	813.400684
4 point TMA	1156.589694
SimpleAverage	1275.081804
6 point TMA	1283.927428
Alpha=0.025,SES iterative	1286.248846
Alpha=0.025,SES iterative	1286.248846
Auto_ARIMA(2,1,2)	1299.979640
Alpha=0.0496, SES Optimized	1316.035487
9 point TMA	1346.278315
RegressionOnTime	1389.135175
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Manual_ARIMA(0,1,0)	3864.279352
NaiveModel	3864.279352

Table 19 RMSE VALUES

SARIMAX Results

Dep. Variable:	y	No. Observations:	187
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-1094.342
Date:	Sun, 17 Apr 2022	AIC	2204.685
Time:	20:33:33	BIC	2228.662
Sample:	0	HQIC	2214.427
	- 187		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1159	0.086	1.349	0.177	-0.052	0.284
ar.L2	-0.0639	0.100	-0.636	0.525	-0.261	0.133
ar.L3	0.0473	0.091	0.521	0.603	-0.131	0.225
ma.L1	-0.9658	0.036	-26.792	0.000	-1.036	-0.895
ar.S.L12	-0.1973	0.706	-0.279	0.780	-1.581	1.186
ma.S.L12	-0.3455	0.717	-0.482	0.630	-1.751	1.060
ma.S.L24	-0.1219	0.398	-0.306	0.759	-0.902	0.658
sigma2	1.528e+05	1.53e+04	10.019	0.000	1.23e+05	1.83e+05

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	42.29
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	0.77	Skew:	0.71
Prob(H) (two-sided):	0.37	Kurtosis:	5.20

Table 20 SARIMA MODEL RESULTS

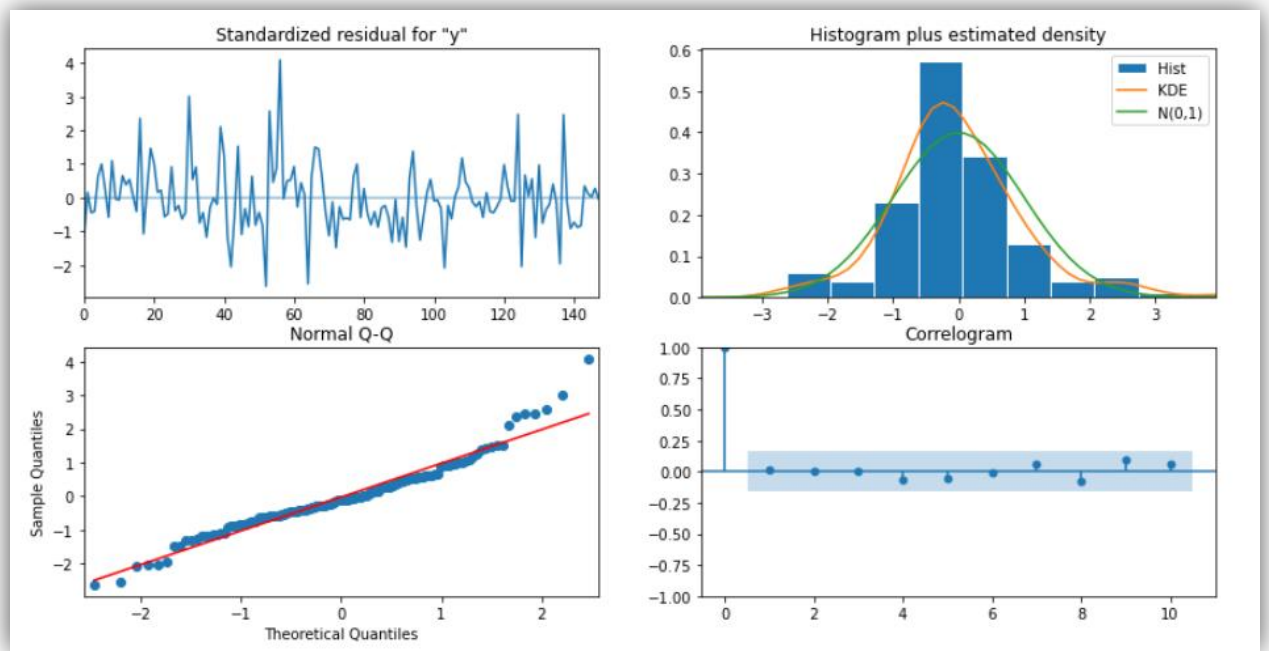


Table 21 DIAGONOSTIC PLOT

For Manual-SARIMA Model forecast on the Entire Data, RMSE is 547.591

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1870.888580	390.915192	1104.708884	2637.068277
1	2489.623611	395.293846	1714.861909	3264.385313
2	3299.650018	395.322875	2524.831422	4074.468615
3	3934.056615	396.282365	3157.357451	4710.755779
4	6135.396032	396.768714	5357.743642	6913.048421

- Based on the overall model evaluation and comparison, Maual SARIMA is selected for final prediction into 12 months in future.
- Manual SARIMA model with optimal parameters (3,1,1)*(1,1,2,12) is found to be the best model in terms of accuracy scored against the full data.
- The model predicts an upward trend and continuation of the seasonal surge in sales in the upcoming 12 months. According to the model the seasonal sale will be more than that of the previous year.

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

- The model forecasts sale of 29535 units of Sparkling wine in 12 months into future. Which is an average sale of 2462 units per month.
- The seasonal sale in December 1995 will hit a maximum of 6136 units, before it drops to the lowest sale in January 1996; at 1246 units
- The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995 (October, November and December), which is a total of 13,370 units of sparkling wine is expected to be sold.
- The forecast also indicates that the year-on-year sale of sparkling wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve the sale compared to previous years.

```
count       12.000000
mean      2461.187692
std       1391.118211
min       1245.727187
25%       1656.647694
50%       1855.796919
75%       2692.130213
max       6135.396032
Name: mean, dtype: float64
```

Table 22 Summary statistics

1995-08-31	1870.888580
1995-09-30	2489.623611
1995-10-31	3299.650018
1995-11-30	3934.056615
1995-12-31	6135.396032
1996-01-31	1245.727187
1996-02-29	1584.643750
1996-03-31	1840.705257
1996-04-30	1823.847826
1996-05-31	1668.706097
1996-06-30	1620.472485
1996-07-31	2020.534848

Freq: M, Name: mean, dtype: float64

Table 23 Forecasted results

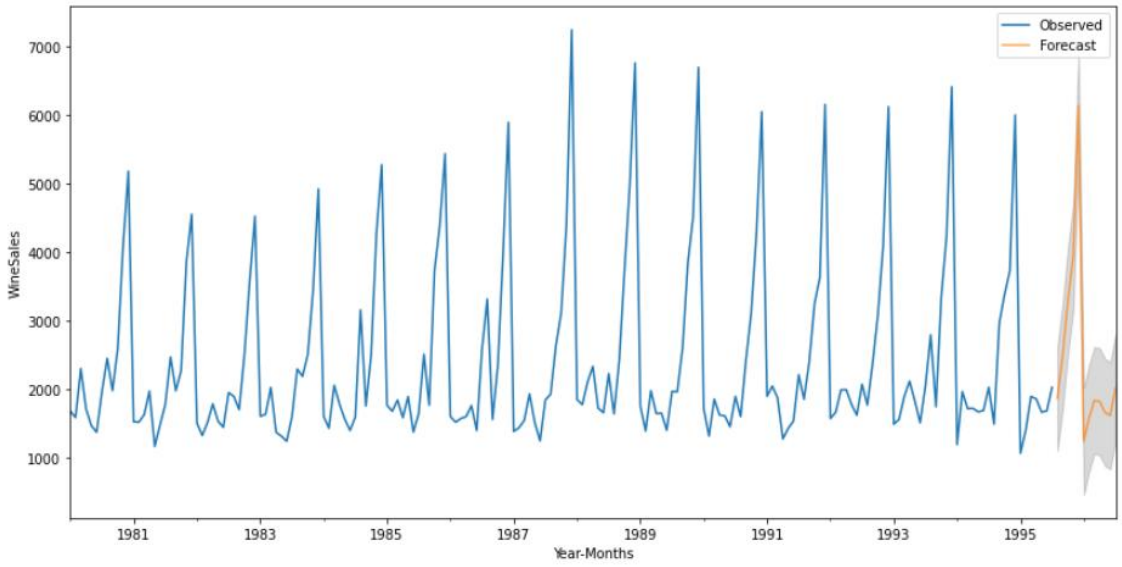


Figure 43 Plot Actual and FutureForecasat Results

- The model forecasts sale of 29535 units of Sparkling wine in 12 months into future. Which is an average sale of 2462 units per month. The seasonal sale in December 1995 will hit a maximum of 6136 units, before it drops to the lowest sale in January 1996; at 1246 units. The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995 (October, November and December), which is a total of 13,370 units of sparkling wine is expected to be sold. The forecast also indicates that the year-on-year sale of sparkling wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve the sale compared to previous years

PROBLEM 2 : TSF - ROSE

2.1 Read the data as an appropriate Time Series data and plot the data

Solution:

Loaded required packages and read Monthly sales of Rose wine dataset without using panda's date-time format.

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Figure 44 View head of the dataset

The dataset Rose contain two columns of data:

The monthly time stamp from Jan 1980 to July 1995 and the sales corresponding to the wines.

Method-1:

Table 24 Create Date-Range

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

- Create Time Stamps and adding it to the data frame to make it a Time-series data.

- Add the time stamp to the original data-frame and set the time stamp as an index, also drop the YearMonth column from the dataset.

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Table 25 View head of the time series

Method-2:

Alternate way to read the original data-frame has a Time series data is by using panda's functions. [parse_dates=True, squeeze=True, index_col=0]

```
YearMonth
1980-01-01    112.0
1980-02-01    118.0
1980-03-01    129.0
1980-04-01     99.0
1980-05-01    116.0
Name: Rose, dtype: float64
```

```
YearMonth
1995-03-01    45.0
1995-04-01    52.0
1995-05-01    28.0
1995-06-01    40.0
1995-07-01    62.0
Name: Rose, dtype: float64
```

Table 26 View head and tail of the time series

- All values are properly loaded for the dataset with the index as panda's date-time format. The Rose Time series has values in float64 data-type format.
- Rose time series contain 2 missing values, they are for the time stamp '1994-07-01' and '1994-08-01'
- Impute the null values by using interpolation [polynomial of order 2].

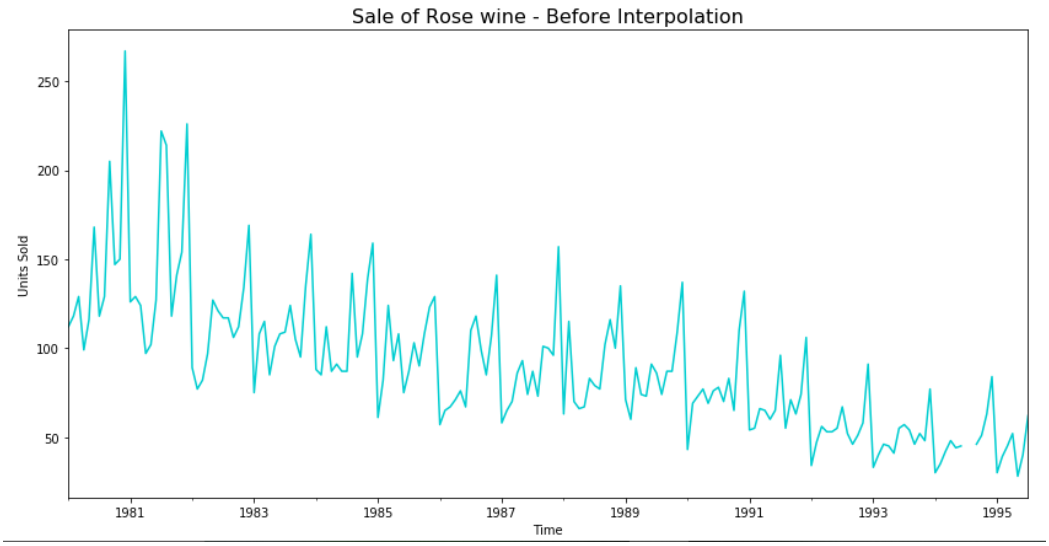
```
YearMonth
1994-07-01   NaN
1994-08-01   NaN
Name: Rose, dtype: float64
```

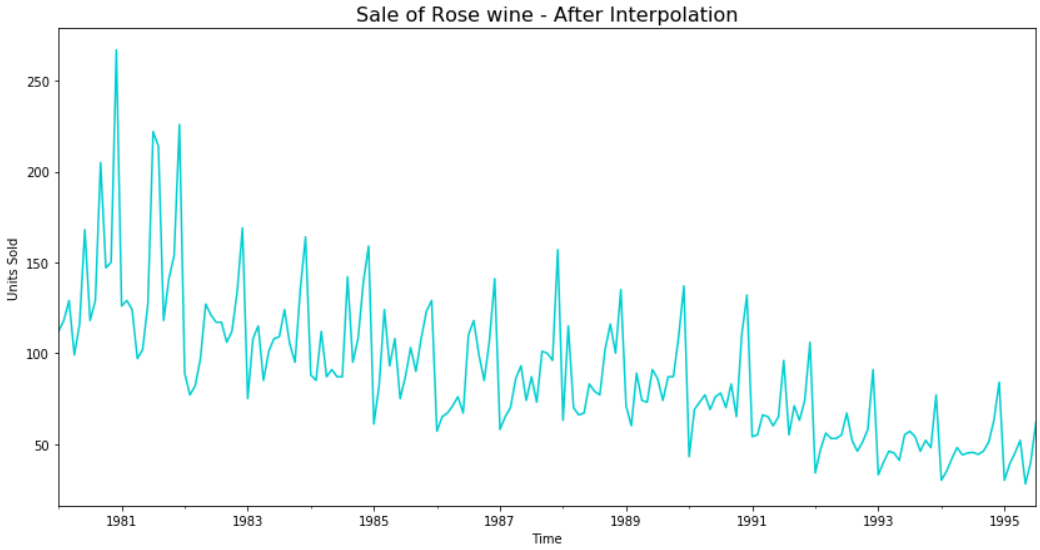
```
YearMonth
1994-07-01    45.364189
1994-08-01    44.279246
Name: Rose, dtype: float64
```

Table 27 Handling missing values

Plot the Sparkling Time Series to understand the behaviour of the data:

Sales Data of Rose Wines:





- The Rose wine dataset shows significant seasonality and decreasing Trend could be observed with a multiplicative seasonality present.
- The demand for Rose had been fell out-of-favour over the years.

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

Check the basic measures of descriptive statistics:

```
count    187.000000
mean      89.907184
std       39.246679
min       28.000000
25%       62.500000
50%       85.000000
75%      111.000000
max      267.000000
Name: Rose, dtype: float64
```

Table 28 measures of descriptive statistics

- The mean value of the Time Series is nearly same as the median values. As a time series data it may signify presence of decreasing trend and multiplicative seasonality.
- The descriptive summary of the data shows that on an average 90 units of Rose wines were sold each month on the given period of time. 50% of months sales varied from 63 units to 112 units. Maximum sale reported in a month is 267 units and minimum of 28 units
- The basic measures of descriptive statistics tell us how the Sales have varied across years. But for this measure of descriptive statistics we have averaged over the whole data without taking the time component into account.

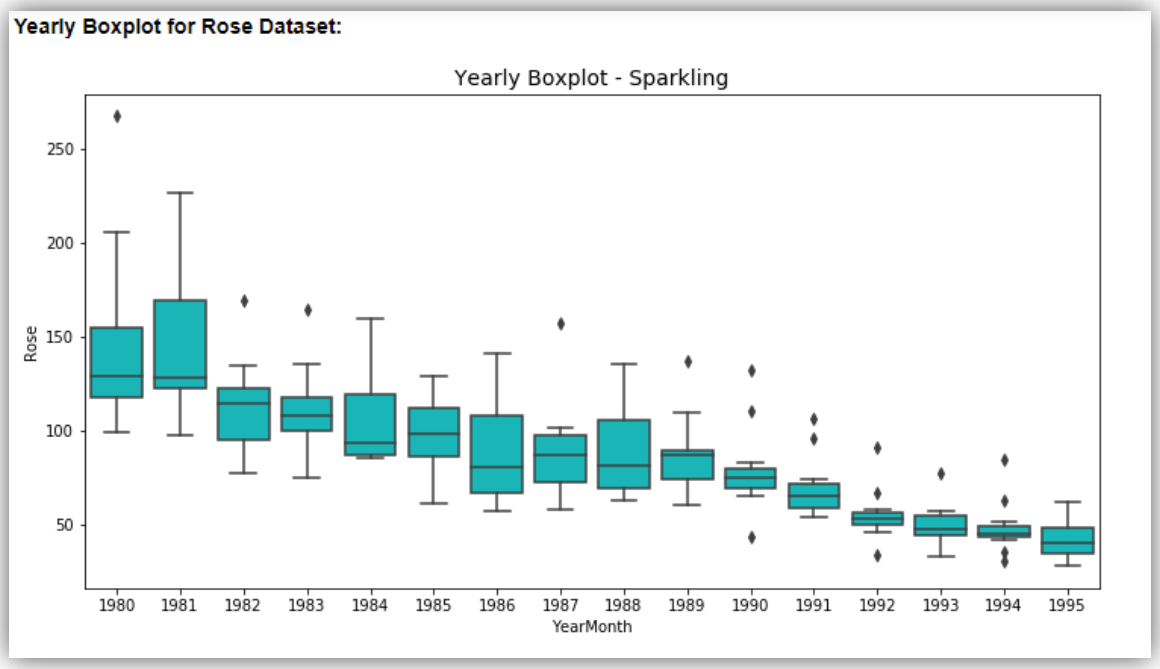


Figure 45 Yearly Boxplot

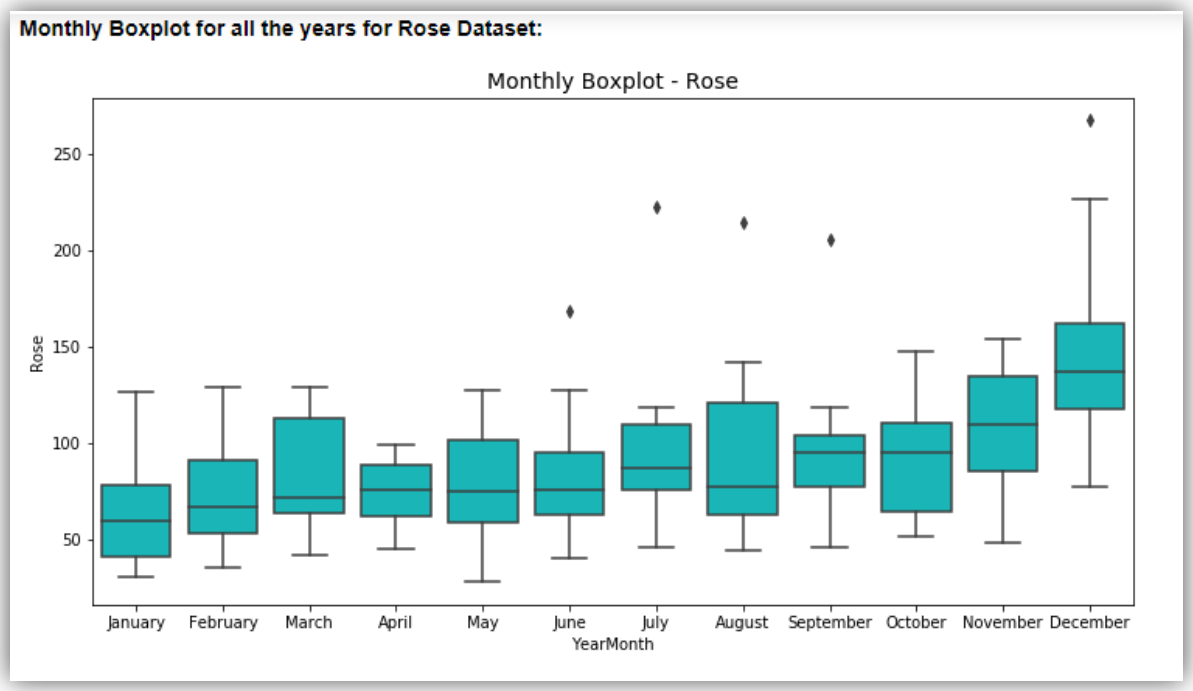


Figure 46 Monthly Boxplot

- The yearly-boxplot, shows that the average sale of Rose wine moving according to the downward trend in sales over the years. The outliers over upper bound in the yearly-boxplot most probably represent the seasonal sale during the seasonal months.
- The monthly-box-plot shows a clear seasonality during the seasonal months of November and December. Though the sale tanks in the month of January, it picks up in the due course of the year.
- Average sale in December is around 140 units, November is around 110 units and October is around 90 units.

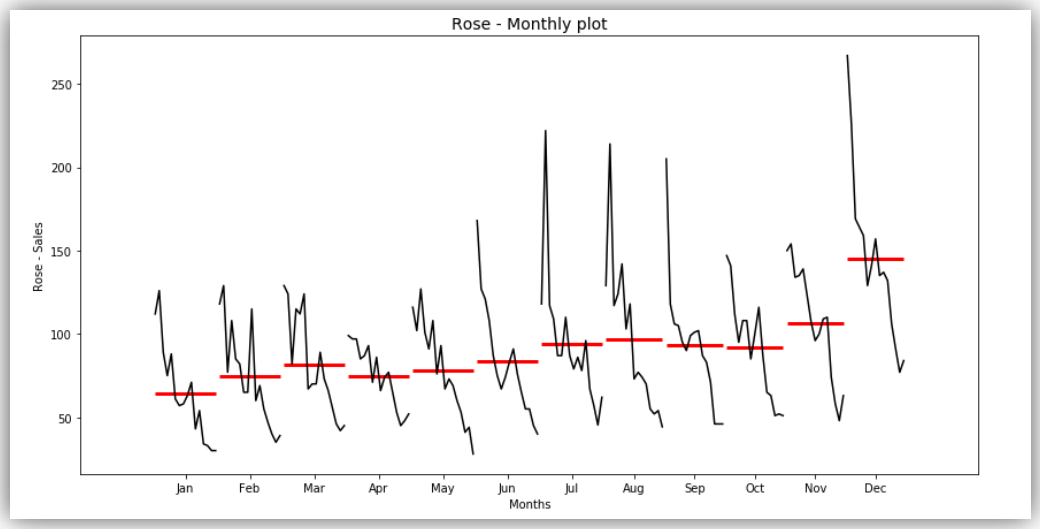


Figure 47 Monthly Plot

- The monthly plot for Rose shows mean and variation of units sold each month over the years. Sale in months such as July, August, September and December shows a higher variation than the rest
- Sale in December with a mean few points below 100, varies from 75 to 270 units over the years. Whereas the average sale is less than or closer to 100 units (above50) for the rest of the year.

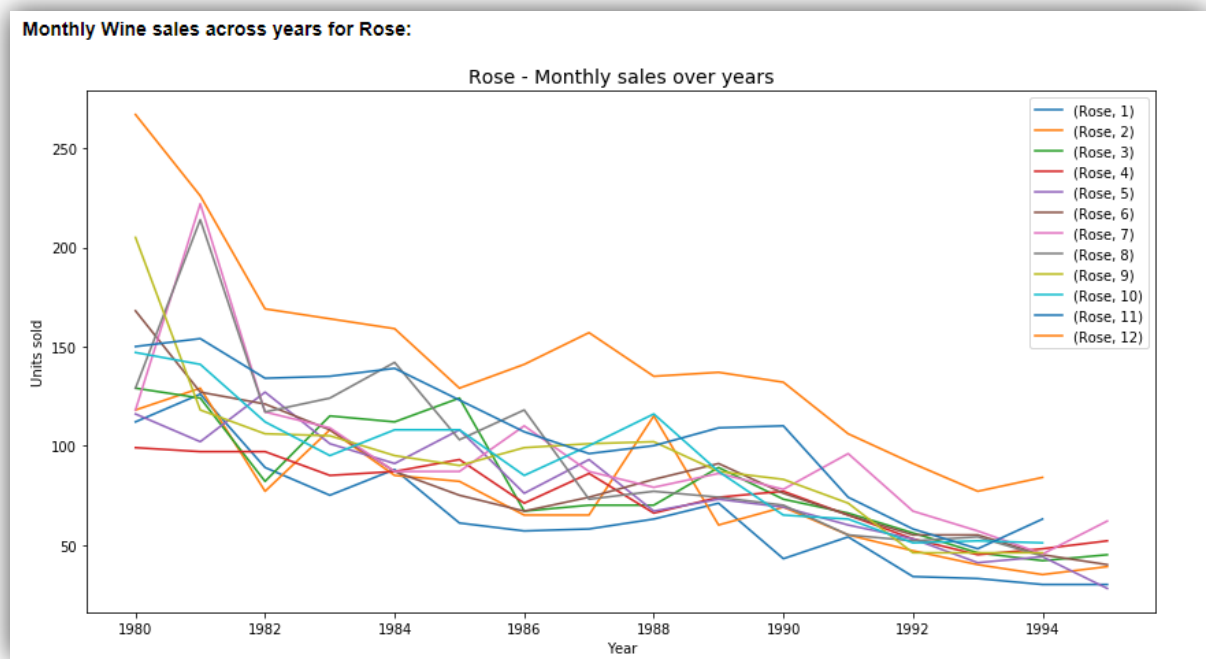


Figure 48 Monthly Sales Over Years

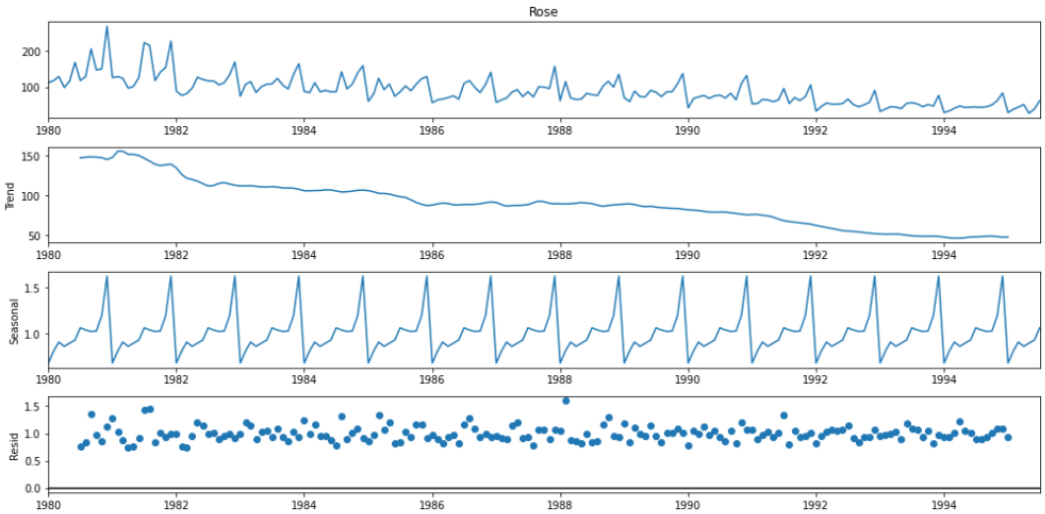


Figure 49 Decomposition of Rose Time Series with multiplicative Seasonality

- The plot of monthly sale over the years also shows the seasonality component of the time-series, with November and December selling exponentially higher volumes than other months.
- The highest volume of Rose wines were sold in December, 1980 and the least of December sale was in 1993. Though December sale picked after 1983, it consistently dipped after 1987.
- The time series is decomposed and a decreasing trend is observed. Also the residuals look random and in general not with high deviations.

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

The train and test datasets are created with year 1991 as starting year for test data :-

```
rose = pd.DataFrame(rose)
train_rose = rose[rose.index.year < 1991]
test_rose = rose[rose.index.year >= 1991]
```

First few rows of Training Data:

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 of Training Data:

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

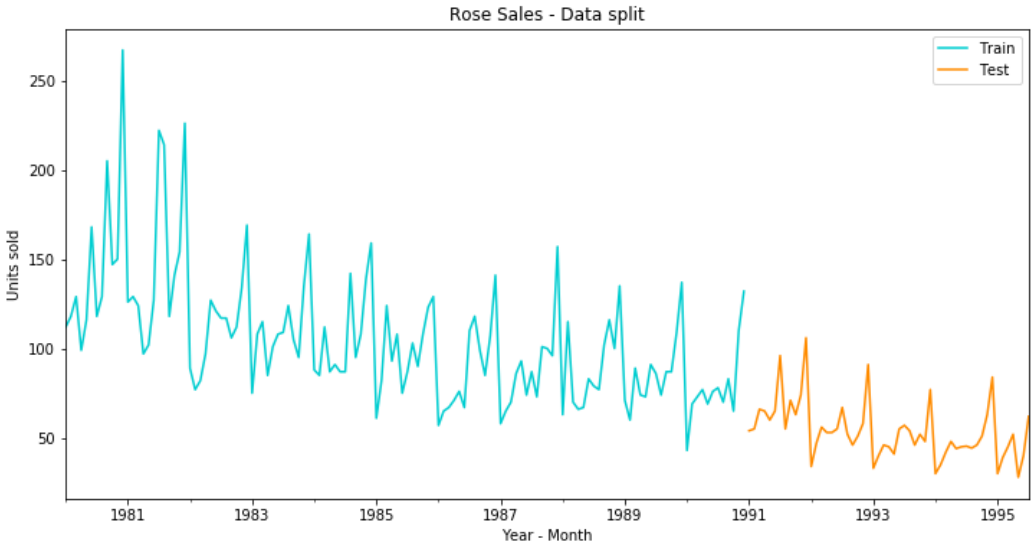
First few rows of Test Data:

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Test Data:

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

Figure 50 Train and test dataset

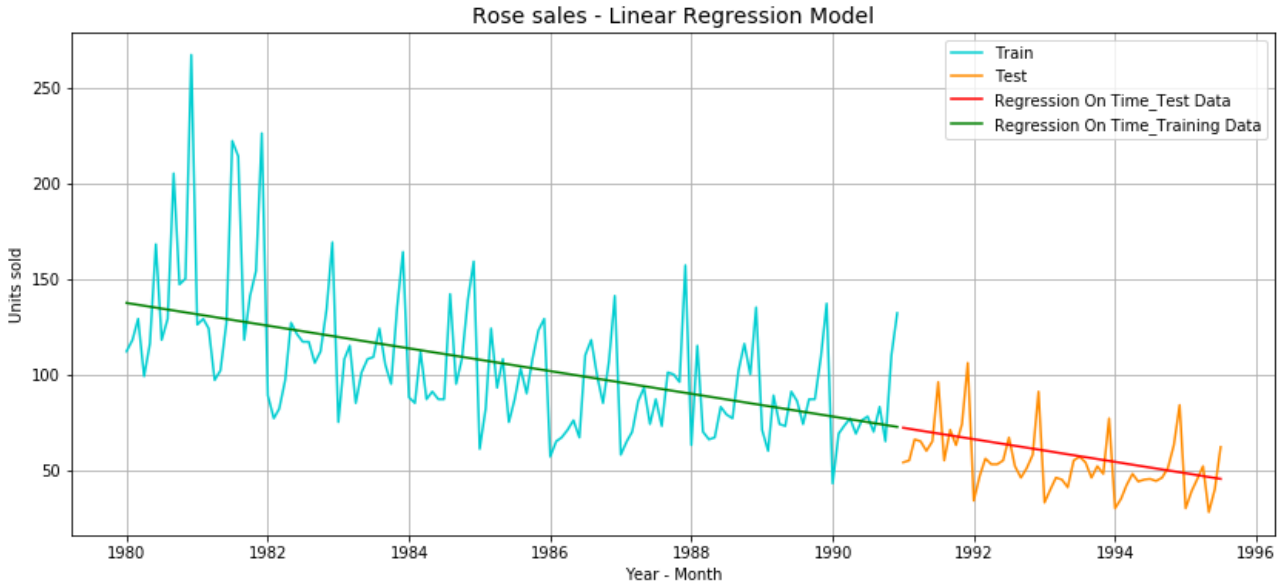


51 The Plot Rose Time Series as train and test

2.4 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

To regress the sale of Rose wines, numerical time instance order for both training and test set were generated and the values added to the respective datasets



52Linear Regression Model

- The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series.
- For Regression on Time forecast on the Test Data, RMSE is 15.278
- The model has successfully captured the trend of the series, but does not reflect the seasonality.

Model 2: Naive forecast

- In naive model, the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.
- The model has taken the last value from the test set and fitted it on the rest of the train time period and used the same value to forecast the test set.
- For Naive forecast on the Test Data, RMSE is 79.75.
- The model do not capture the trend or seasonality for the given dataset.

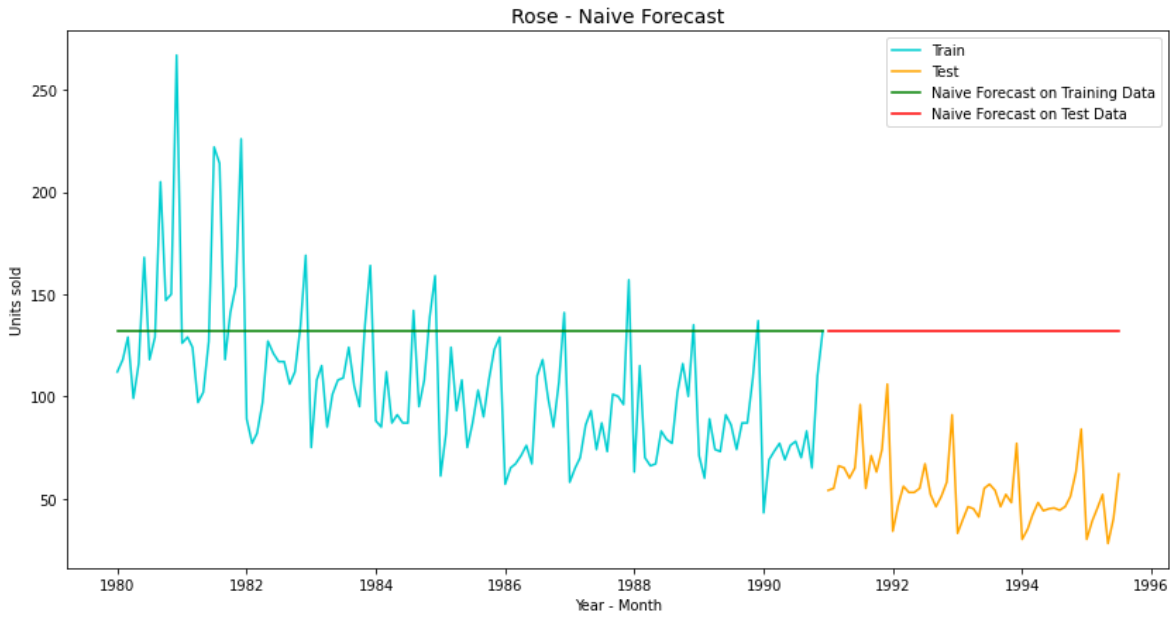


Figure 53 Naive Forecast Model

Model 3: Simple Average

In the Simple Average model, the forecast is done using the mean of the time-series variable from the training set.

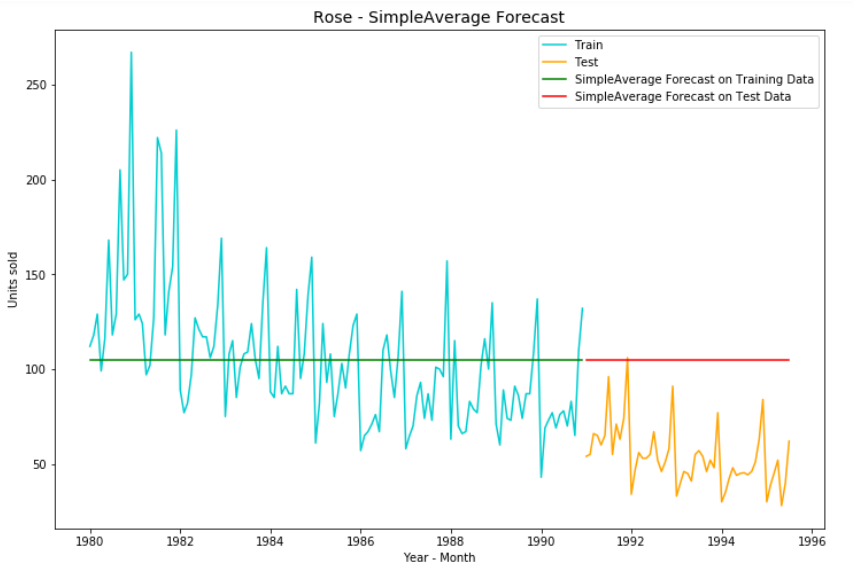


Figure 54 Simple forecast Model

- The model is not capable of either forecasting or able to capture the trend and seasonality present in the dataset.
- For Simple Average on the Test Data, RMSE is 53.48

Model 4: Moving Average

- For the moving average model, we will calculate rolling means (or trailing moving averages) for different intervals. The best interval can be determined by the maximum accuracy.
- The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points.
- For Rose dataset the accuracy is found to be higher with the lower rolling point averages.
- In moving average forecasts the values can be fitted with a delay of n number of points.
- The best interval of moving average from the model is 2 point.

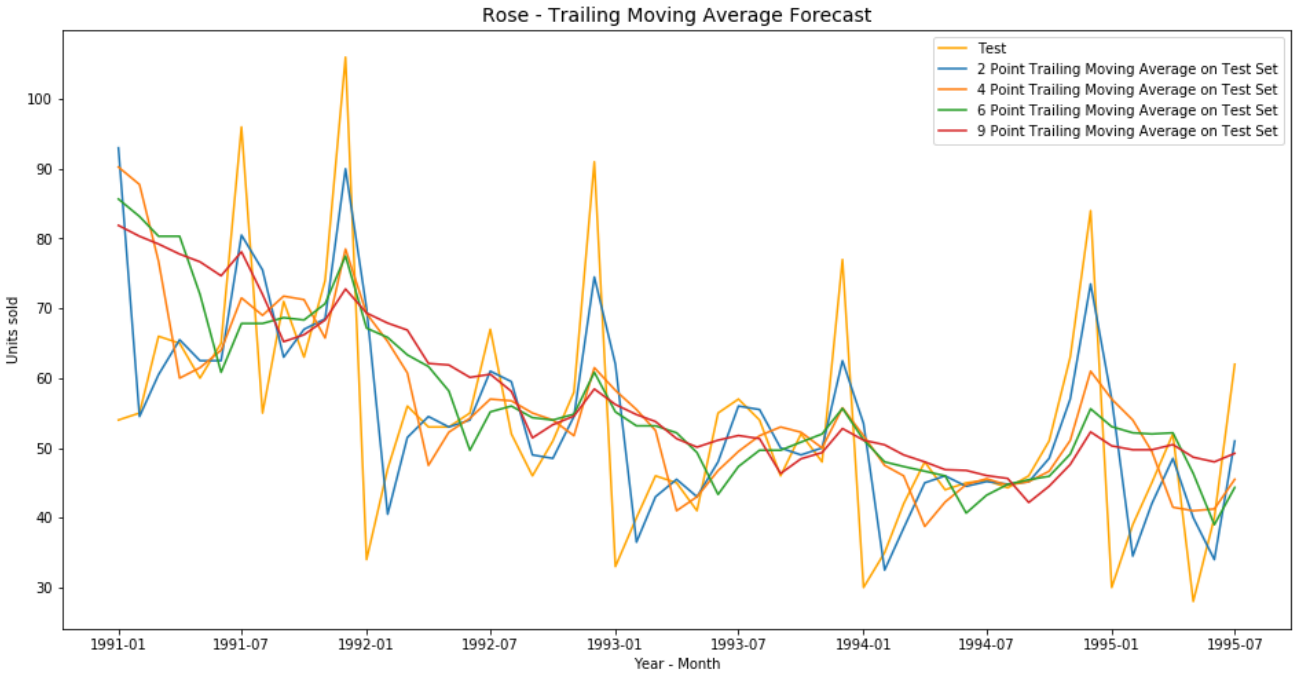


Figure 55 Moving Average Model

	Test RMSE
RegressionOnTime	15.278369
2 point TMA	11.530054
4 point TMA	14.458402
6 point TMA	14.572976
9 point TMA	14.732918

Table 29 RMSE VALUE

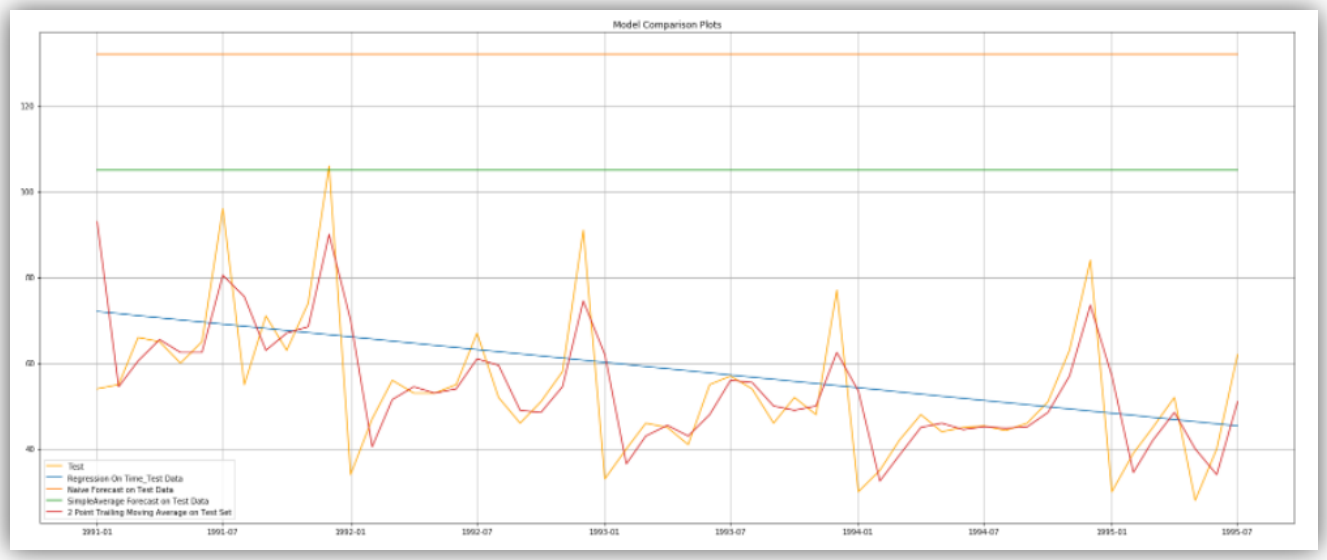


Figure 56 Model comparison and RMSE on test data

Model 5: Simple Exponential Smoothing

- The model was ran without passing a value for alpha and used parameters: 'optimized=True, use_brute=True'.
- The auto-fit model picked up alpha = 0.0987 as the smoothing parameter.
- Simple Exponential Smoothing is applied if the time-series has neither a trend nor seasonality, which is not the case with the given data.
- The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually.
- For alpha value closer to 1, forecasts follows the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened
- For Rose, test RMSE is found to be higher for values closer to zero, which is same as in Simple average forecast.
- Both manual alpha =0.10 and optimized alpha value are having similar RMSE value.

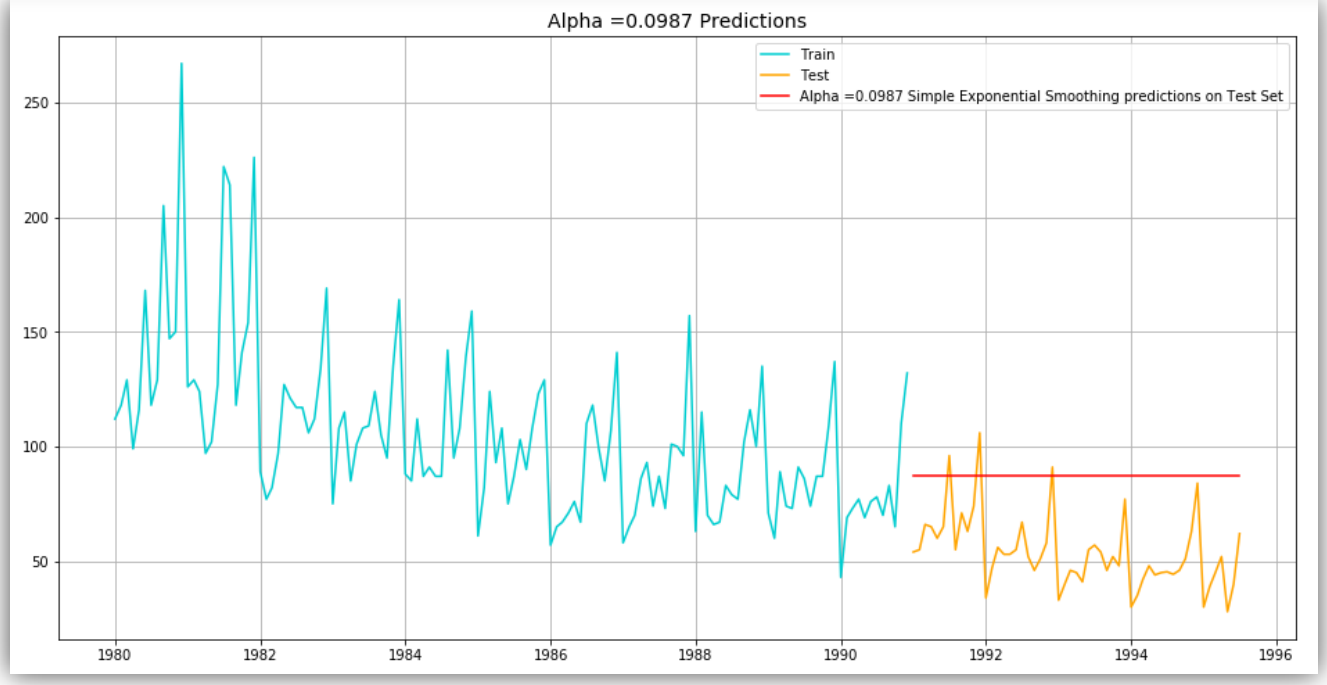


Figure 57 SES Iterative Model

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

- The Double Exponential Smoothing models is applicable when data has trend, but no seasonality. Rose data contain significant trend component and seasonality.
- In first iteration, smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE values, which is as below with alpha 0.1 and beta 0.1
- On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'
- The auto-fit model has lower RMSE value compared to iterative alpha=0.1 and beta=0.1 RMSE value.

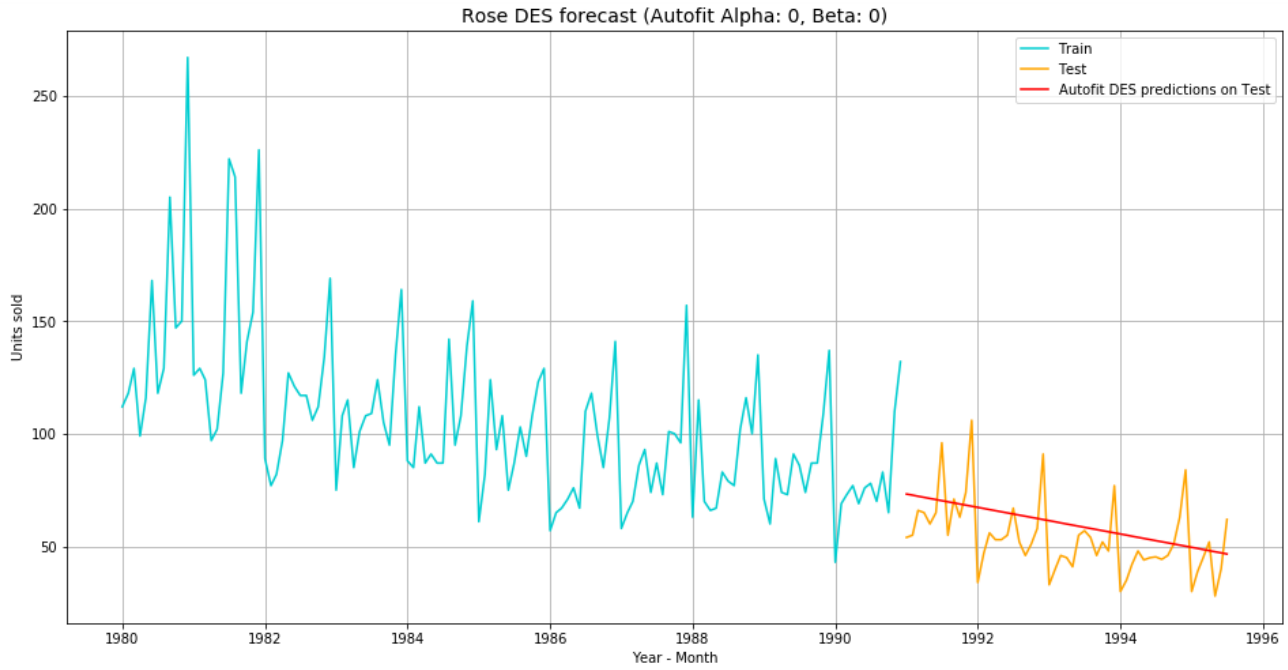


Figure 58 DES Optimised Model

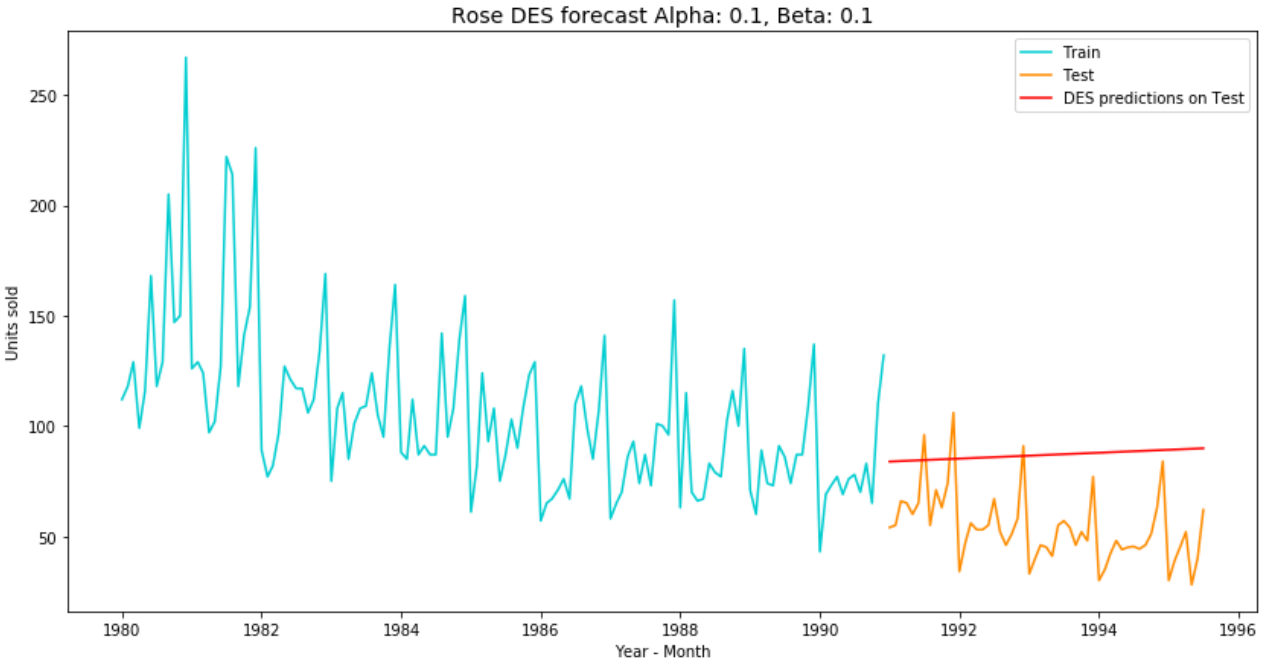


Figure 59 DES Iterative Model

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

- The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Rose data contain significant trend and seasonality.
- On first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE values, which is as below with alpha 0.4, beta 0.1 and gamma 0.3
- On the second iteration the model was allowed to choose the optimized values using parameters 'optimized=True, use_brute=True'
- The auto-fit model retuned higher RMSE value compared to iterative alpha=0.1, beta=0.2 and gamma=0.3 RMSE value.

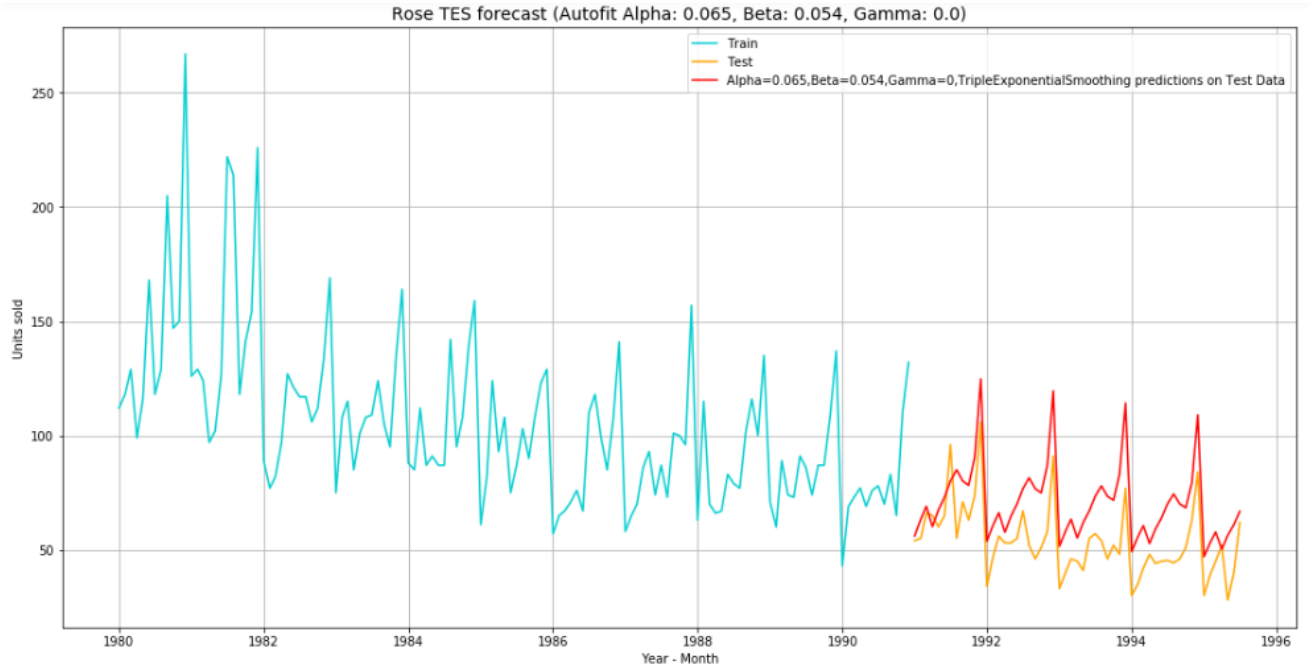


Figure 60 TES Optimised Model.

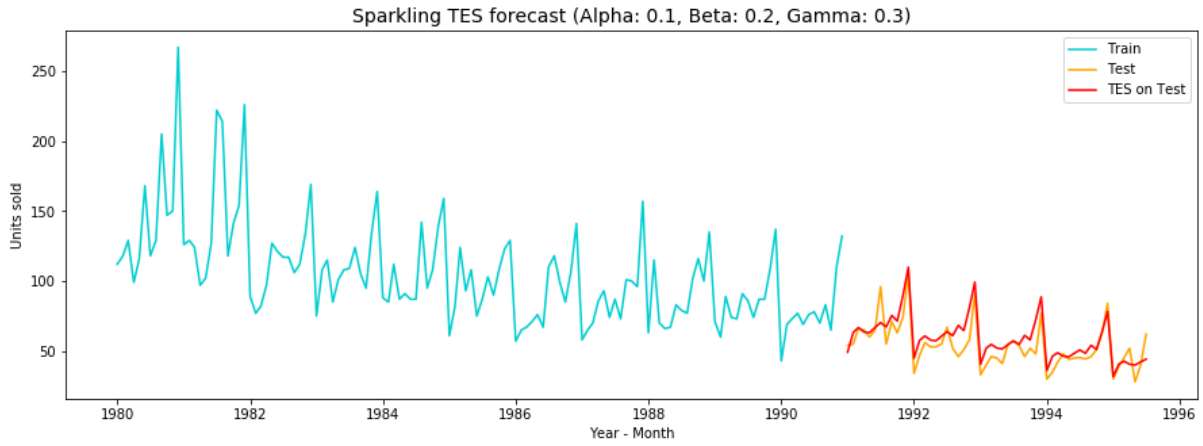


Figure 61 TES Iterative Model

Model Comparison:

	Test RMSE
Alpha=0.1,Beta=0.2,gamma=0.3,TES_Iterative	9.880143
2 point TMA	11.530054
4 point TMA	14.458402
6 point TMA	14.572976
9 point TMA	14.732918
RegressionOnTime	15.278369
Alpha=0.0,Beta=0.0, DES Optimized	15.718202
Alpha=0.065,Beta=0.054,gamma=0.0 TES Optimized	19.434699
Alpha=0.0987, SES Optimized	36.824464
Alpha=0.10,SES_Iterative	36.856268
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
SimpleAverage	53.488233
NaiveModel	79.745697

Table 30 Test RMSE Value

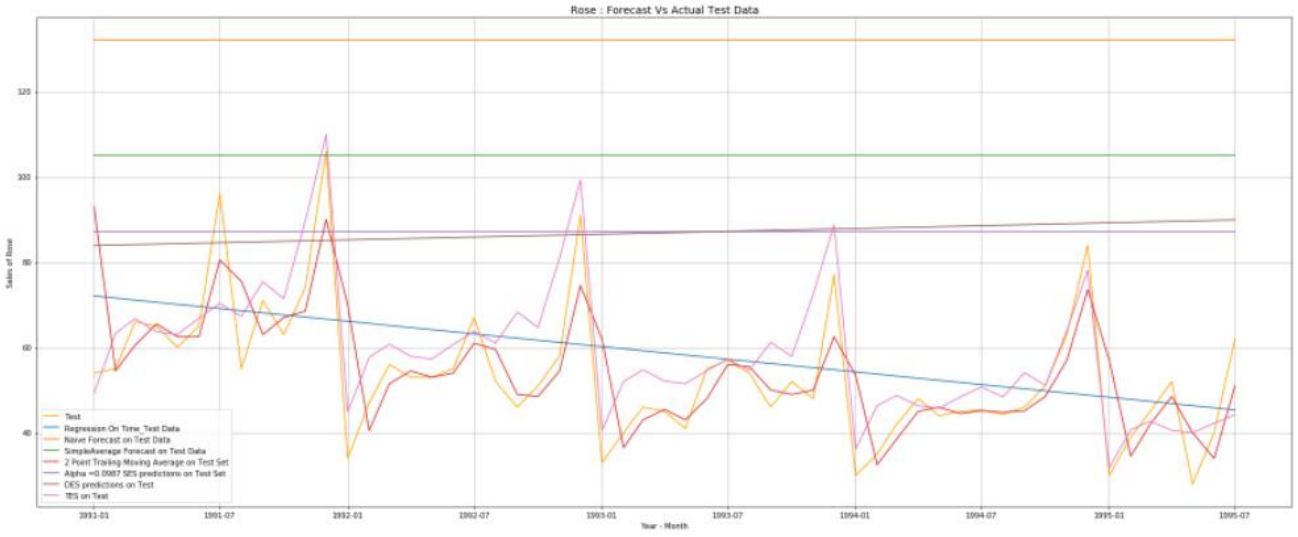


Figure 62 Combination of diferent forecasts

- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data.
- 2 point trailing moving average model is also found to have fit well with a slight lag in test dataset.

2.5 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.

Solution:

- Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determine the presence of unit root in the series to understand if the series is stationary or not
- Null Hypothesis: The series has a unit root, that is series is non-stationary
- Alternate Hypothesis: The series has no unit root, that is series is stationary
- If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary
- The ADF test on the original Rose series retuned the below values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis.

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.872615
p-value                          0.345051
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

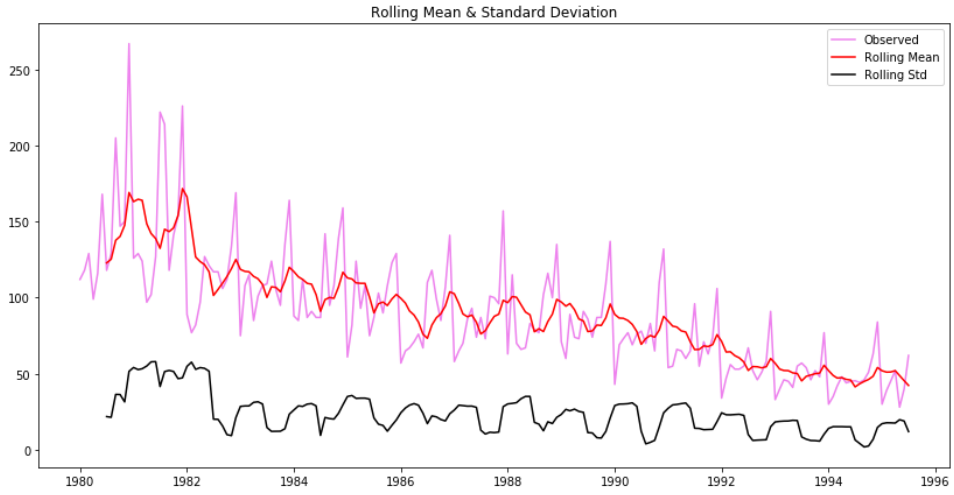


Figure 63 ADF test on Original Series.

Differencing of order one is applied on the Rose series as below and tested for stationarity. At an order of differencing 1, the series is found to be stationary as below

The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if it's multiplicative or additive in character.

The altitude of rolling mean and std dev is seen changing according to change in slope, which indicates multiplicity.

The ADF test is also done in this exercise with logarithmic transformation of the train data and differencing of seasonal order (12), to understand if removing the multiplicity of the seasonal component will have an impact on the accuracy of model.

```
Results of Dickey-Fuller Test:
Test Statistic                  -8.044081e+00
p-value                          1.814191e-12
#Lags Used                       1.200000e+01
Number of Observations Used      1.730000e+02
Critical Value (1%)             -3.468726e+00
Critical Value (5%)             -2.878396e+00
Critical Value (10%)            -2.575756e+00
dtype: float64
```

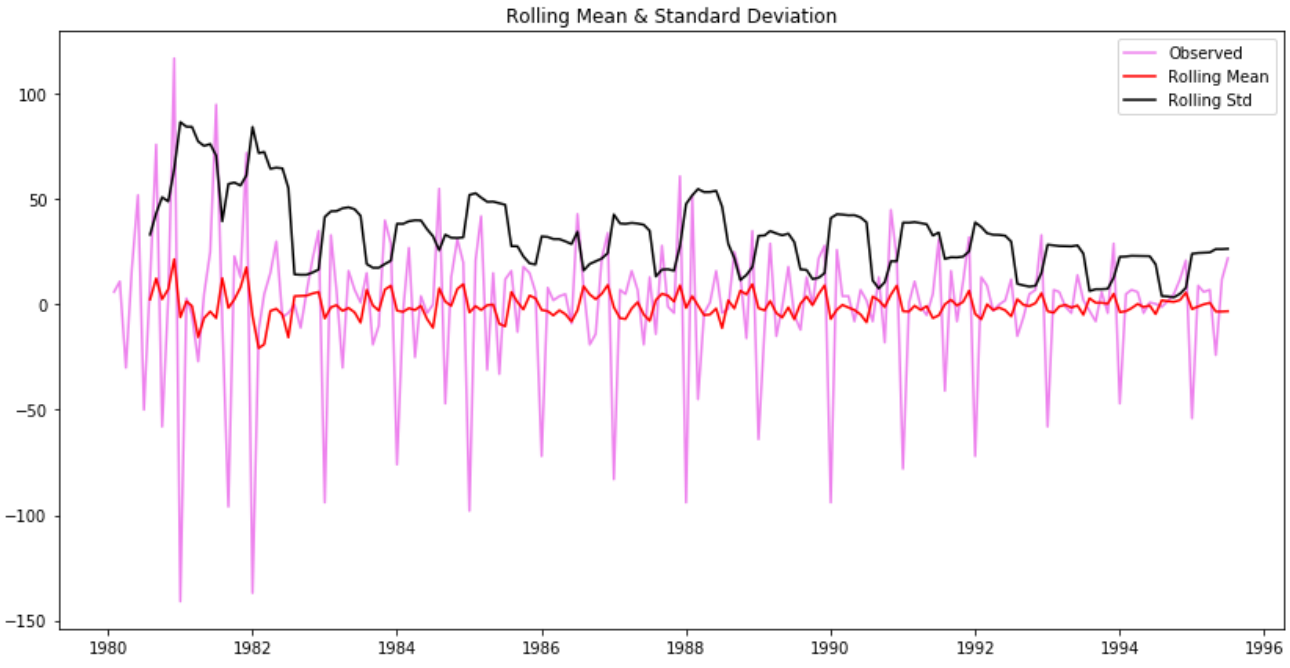


Figure 64 ADF test after differencing d=1

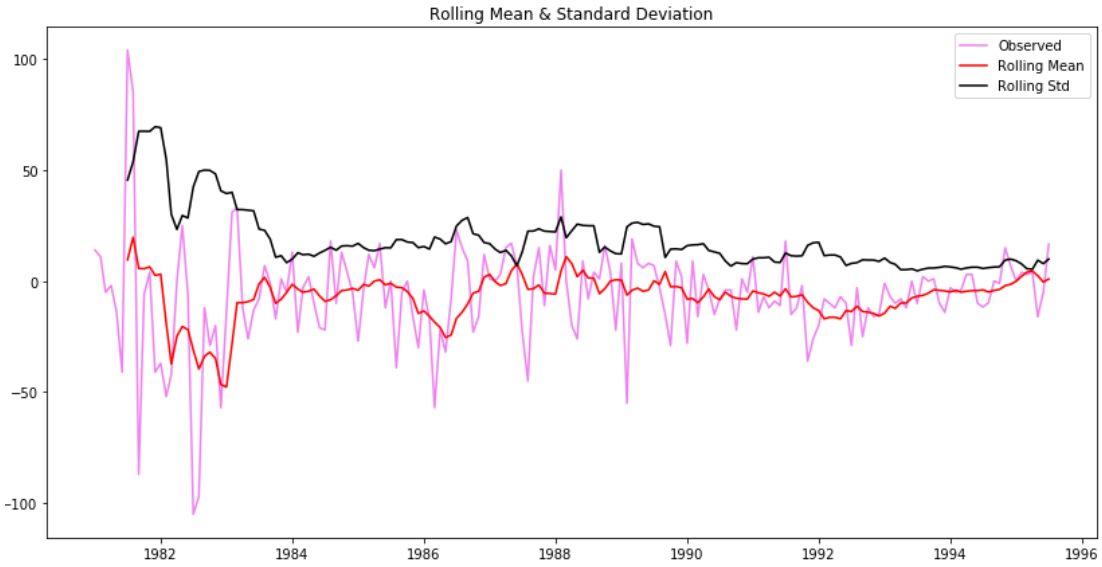


Figure 65 ADF test on log series after differencing

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

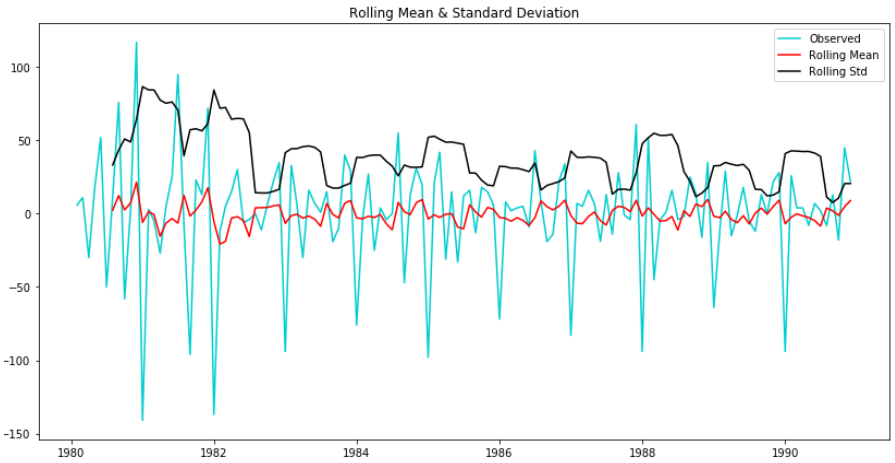


Figure 66 ADF test on train data after dofferencing d-1

- We see that at 5% significant level the Time Series is non-stationary.
- Let us take a difference of order 1 and check whether the Time Series is stationary or not.
- **We see that at $\alpha\alpha$ = 0.05 the Time Series is indeed stationary. d=1**

2.6 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.

Model 8: Auto-ARIMA

- ARIMA model was built with optimised model and found the least AIC value =1276 at (0, 1, 2).
- As the Rose series of data contain seasonality component, ARIMA model do not perform well. The RMSE value for this Auto- ARIMA model is 45.04.

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(0, 1, 2)	Log Likelihood	-636.836
Date:	Sun, 17 Apr 2022	AIC	1279.672
Time:	20:01:13	BIC	1288.297
Sample:	01-01-1980	HQIC	1283.176
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.6970	0.072	-9.689	0.000	-0.838	-0.556
ma.L2	-0.2042	0.073	-2.794	0.005	-0.347	-0.061
sigma2	965.8407	88.305	10.938	0.000	792.766	1138.915

Ljung-Box (L1) (Q):	0.14	Jarque-Bera (JB):	39.24
Prob(Q):	0.71	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	0.82
Prob(H) (two-sided):	0.00	Kurtosis:	5.13

Table 31 AUTO ARIMA Model

Model 9A: Auto-SARIMA

- The model was built on train data with seasonality 12 and with different optimal parameters (p, d, q)x(P, D, Q) parameters, the lowest AIC is 774.97 was obtained at (0, 1, 2)x(2, 1, 2, 12).
- The model was built with the above parameters.

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-380.485
Date:	Sun, 17 Apr 2022	AIC	774.969
Time:	20:05:35	BIC	792.622
Sample:	0	HQIC	782.094
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9524	0.184	-5.167	0.000	-1.314	-0.591
ma.L2	-0.0764	0.126	-0.605	0.545	-0.324	0.171
ar.S.L12	0.0480	0.177	0.271	0.786	-0.299	0.395
ar.S.L24	-0.0419	0.028	-1.513	0.130	-0.096	0.012
ma.S.L12	-0.7526	0.301	-2.503	0.012	-1.342	-0.163
ma.S.L24	-0.0721	0.204	-0.354	0.723	-0.472	0.327
sigma2	187.8634	45.274	4.149	0.000	99.128	276.599

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	4.86
Prob(Q):	0.81	Prob(JB):	0.09
Heteroskedasticity (H):	0.91	Skew:	0.41
Prob(H) (two-sided):	0.79	Kurtosis:	3.77

Table 32 SARIMA Model Result

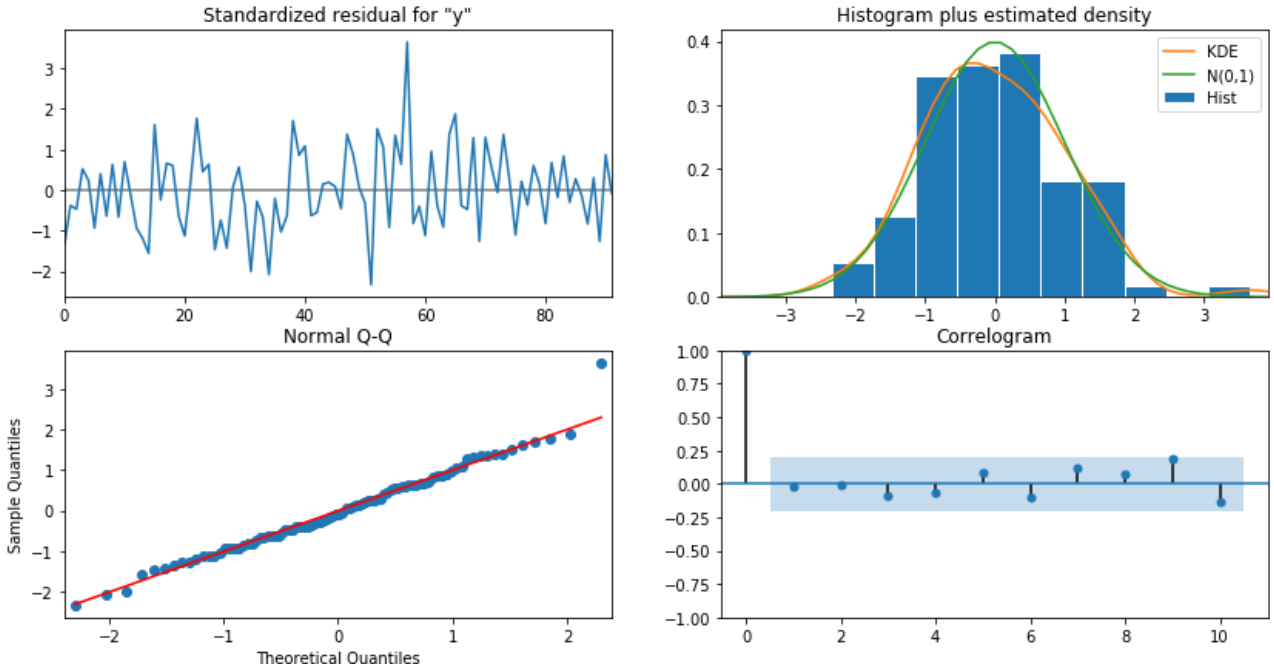


Figure 67 Diagnostic Plot

- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point forms roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
- The RMSE values of the automated SARIMA model is 16.53

YearMonth	Rose	rose_forecasted
1991-01-01	54.0	44.213661
1991-02-01	55.0	62.326882
1991-03-01	66.0	67.313375
1991-04-01	65.0	63.161025
1991-05-01	60.0	66.474493

Table 33 Forecasted result on test data

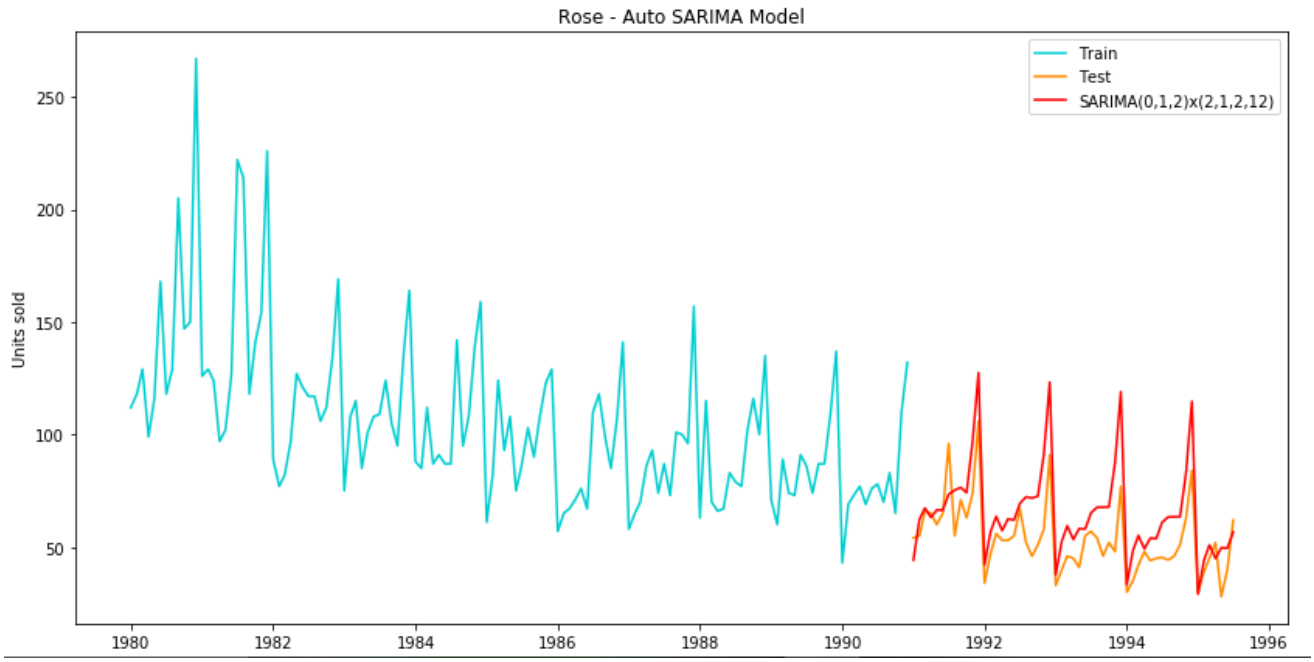


Figure 68 Plot of actual v/s Forecasted result on test data.

Model-9B AUTO SARIMA on Log Series

- The model was built on log transformed train data and with seasonality 12 and with different optimal parameters (p, d, q)x(P, D, Q) parameters, the lowest AIC is -247.08 was obtained at (0, 1, 1)*(1, 0, 1, 12).
- The model was built with the above parameters.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Log Likelihood	127.538
Date:	Sun, 17 Apr 2022	AIC	-247.076
Time:	19:47:51	BIC	-236.028
Sample:	01-01-1980 - 12-01-1990	HQIC	-242.591
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.0652	0.058	-18.391	0.000	-1.179	-0.952
ar.S.L12	0.9555	0.028	33.779	0.000	0.900	1.011
ma.S.L12	-0.8304	0.151	-5.498	0.000	-1.126	-0.534
sigma2	0.0051	0.001	5.146	0.000	0.003	0.007

Ljung-Box (L1) (Q):	1.31	Jarque-Bera (JB):	0.98
Prob(Q):	0.25	Prob(JB):	0.61
Heteroskedasticity (H):	0.80	Skew:	0.18
Prob(H) (two-sided):	0.50	Kurtosis:	3.26

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

Figure: Log series SARIMA Model result

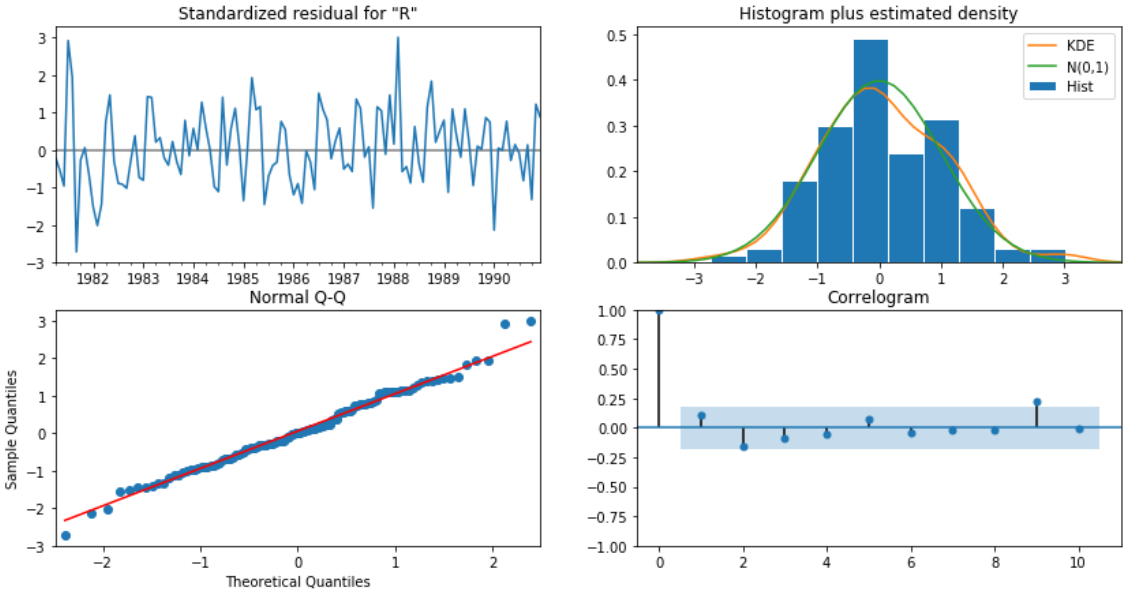


Figure 69 Diagnostic Plot

- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point forms roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
- From the above model summary it can be inferred that MA.L1, AR.L.S12, MA.L.S12 terms has the highest absolute weightage.
- From the p-values it can be inferred that terms MA.L1, AR.L.S12, MA.L.S12 are significant terms, as their values are below 0.05.
- The RMSE values of the automated SARIMA of log series model is 17.91

YearMonth	Rose	rose_forecasted	rose_forecasted_log
1991-01-01	54.0	44.213661	56.850589
1991-02-01	55.0	62.326882	66.336719
1991-03-01	66.0	67.313375	70.530760
1991-04-01	65.0	63.161025	66.370372
1991-05-01	60.0	66.474493	69.031105

Table 34 Forecasted result on test data.

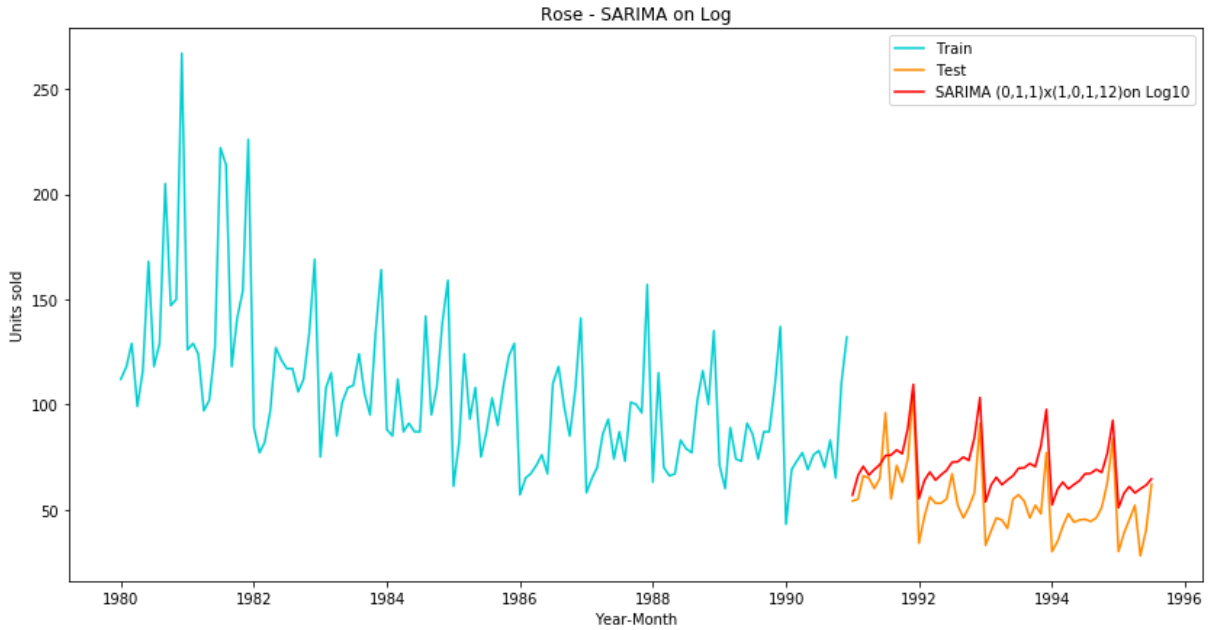


Figure 70 Above shows Plot of Actual v/s Forecasted result on test data.

- The model built with log series data has a higher RMSE value when compared to original train data.

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

Model-10 Manual ARIMA

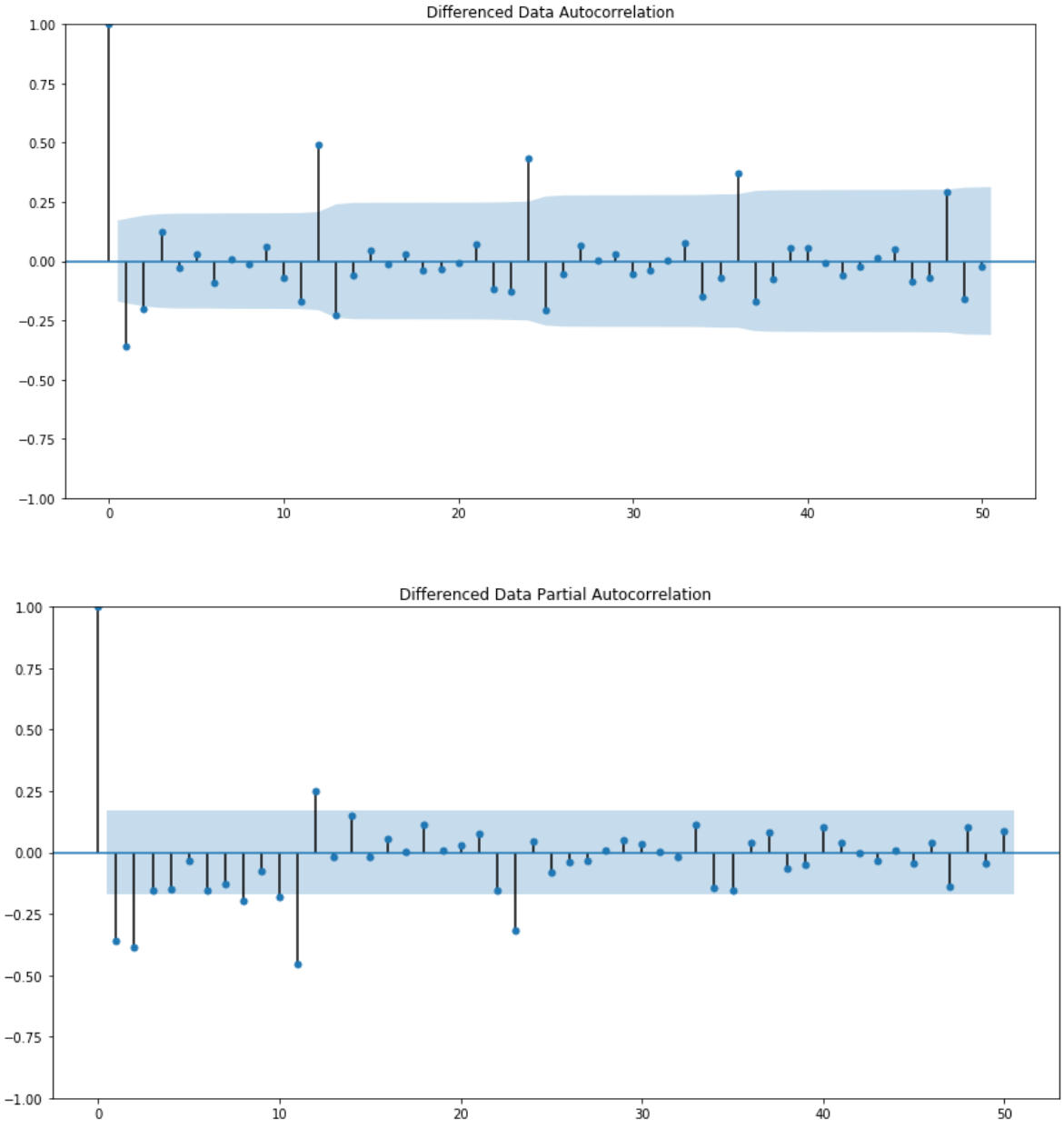


Figure 71: ACF and PACF Plots

- Here, we have taken alpha=0.05.
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
- By looking at above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

```
==========================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -665.577
Date:                Sun, 17 Apr 2022   AIC                           1333.155
Time:                        19:47:53   BIC                           1336.030
Sample:                    01-01-1980   HQIC                          1334.323
                         - 12-01-1990
Covariance Type:                 opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
sigma2      1515.6738    122.418     12.381      0.000    1275.740    1755.608
==========================================================================
Ljung-Box (L1) (Q):                  17.11   Jarque-Bera (JB):                59.55
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.38   Skew:                            -0.95
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.70
==========================================================================
```

Figure 72: Manual ARIMA Summary result.

- The RMSE value of manual ARIMA model is 79.74. Since the ARIMA model do not capture the seasonality, this model do not perform well.

Model-11 Manual SARIMA

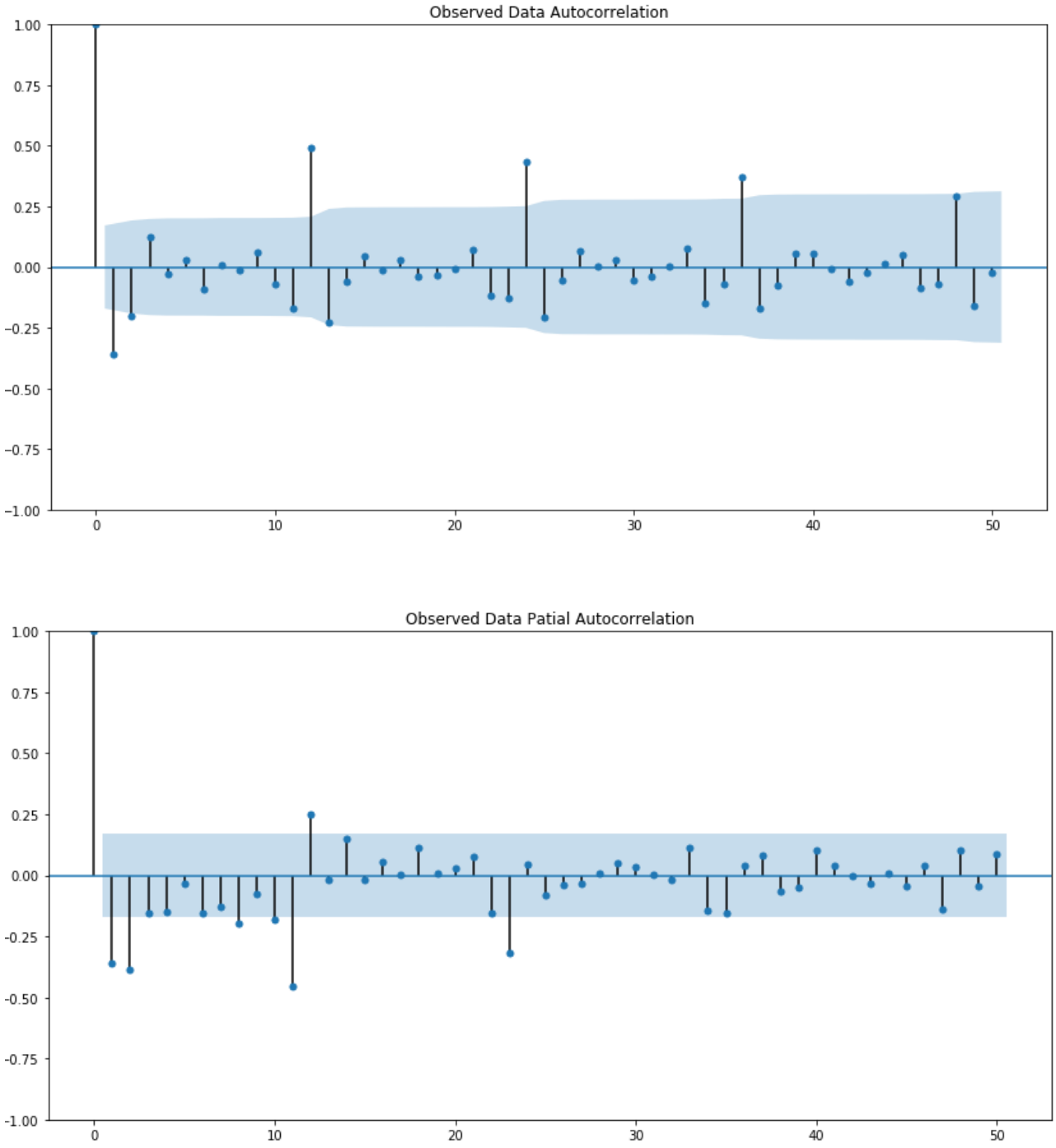


Figure 73: ACF and PACF Plots

- From the ACF plot of the observed/ train data, it can be inferred that at seasonal interval of 12, the plot is not quickly tapering off. So a seasonal differencing of 12 has to be taken

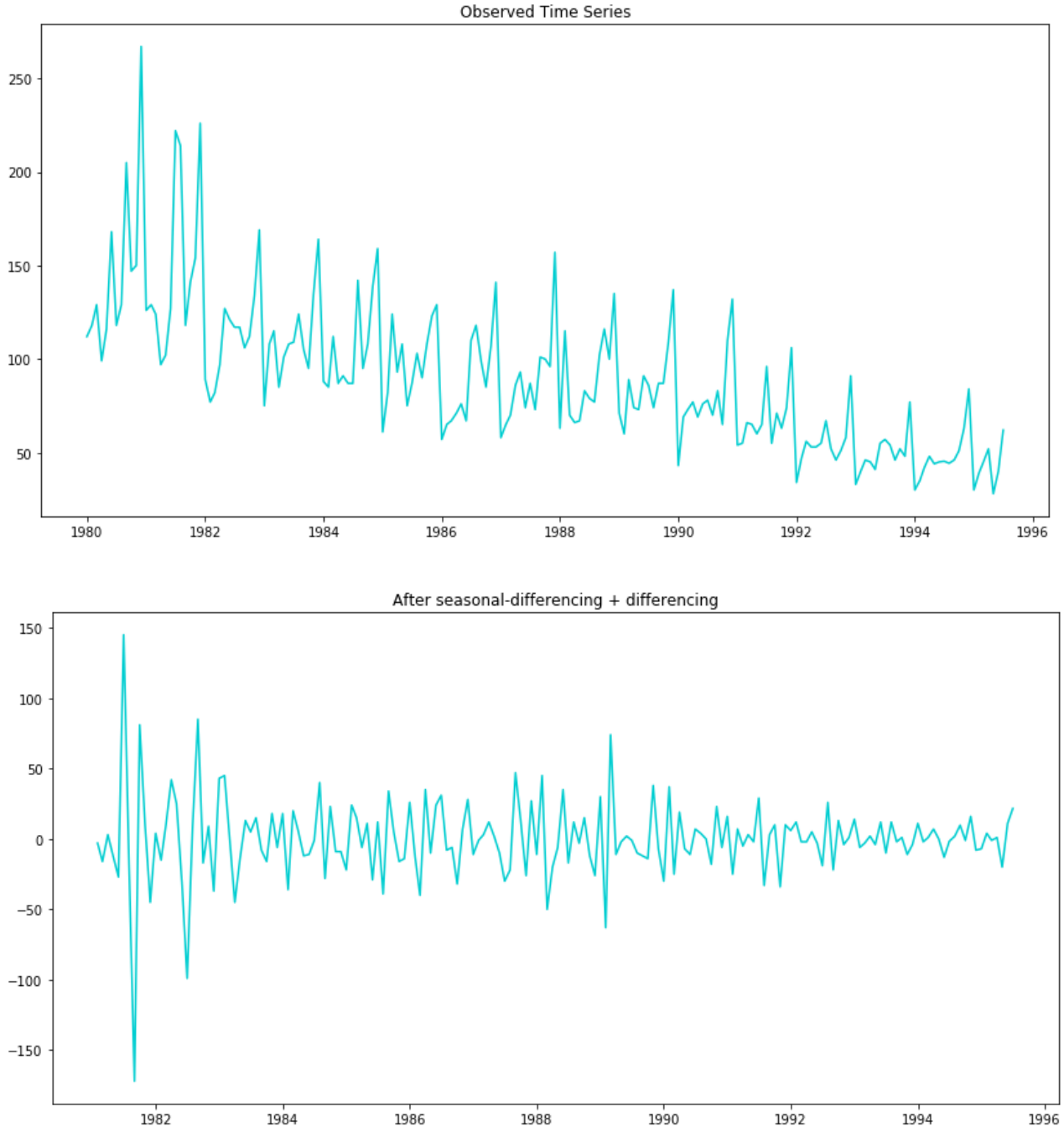
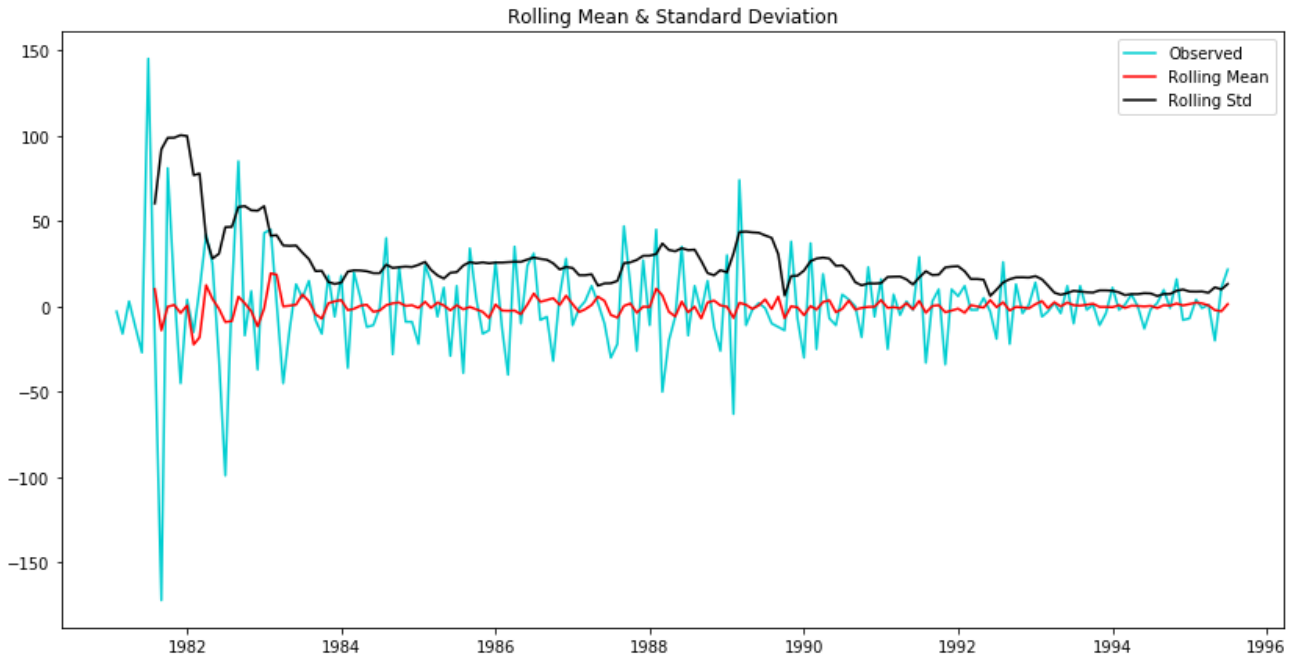


Figure 74: Time Series Plots

- An ADF test need to be done to check the stationarity after the above differencing. With a p-value below alpha 0.05 and test statistic below critical values, it can be confirmed that the data is stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.605791
p-value                          0.000126
#Lags Used                      11.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value (10%)            -2.576373
dtype: float64
```

Figure 75 :ADF test

ACF and PACF plots of the seasonal-differenced + one order differenced data is created to find the values for (p,d,q)x(P,D,Q).

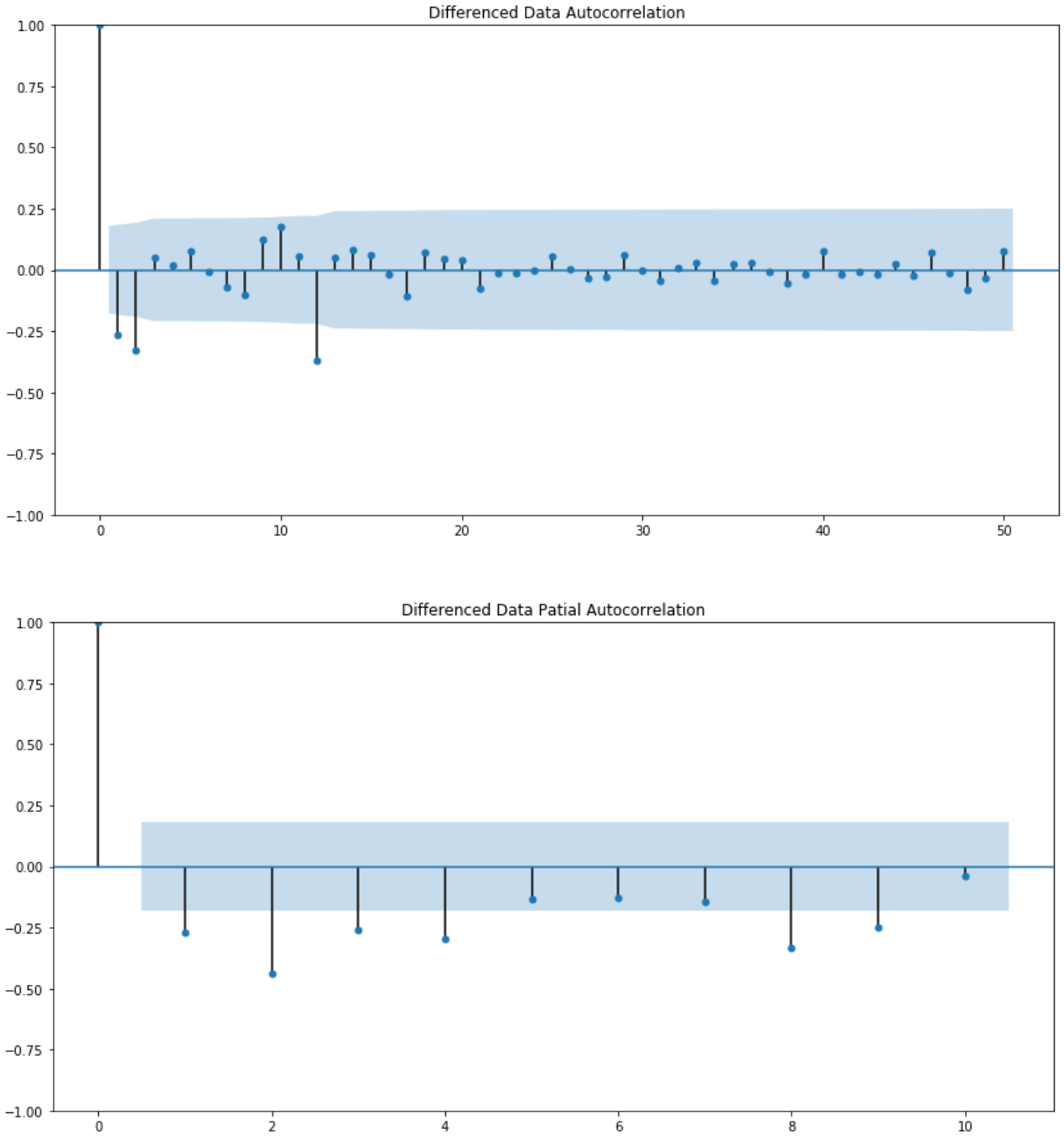


Figure 71 ACF and PACF Plots

- Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till 4th lag it's significant before cut-off, so AR term 'p = 4' is chosen. At seasonal lag of 12, seasonal AR 'P = 0'.

- From ACF plot it can be seen that till lag 2nd is significant before it cuts off, so MA term 'q =2' is selected and at seasonal lag of 12, a significant lag is apparent, so kept seasonal MA term 'Q = 1' initially.
- The seasonal MA term 'Q' was later optimized to 2, by validating model performance, as the data might be under-differenced.
- The final selected terms for SARIMA model is (4, 1, 2)*(0,1,2,12).
- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point forms roughly a straight line.
- The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
- The RMSE values of the automated SARIMA model is 15.38.

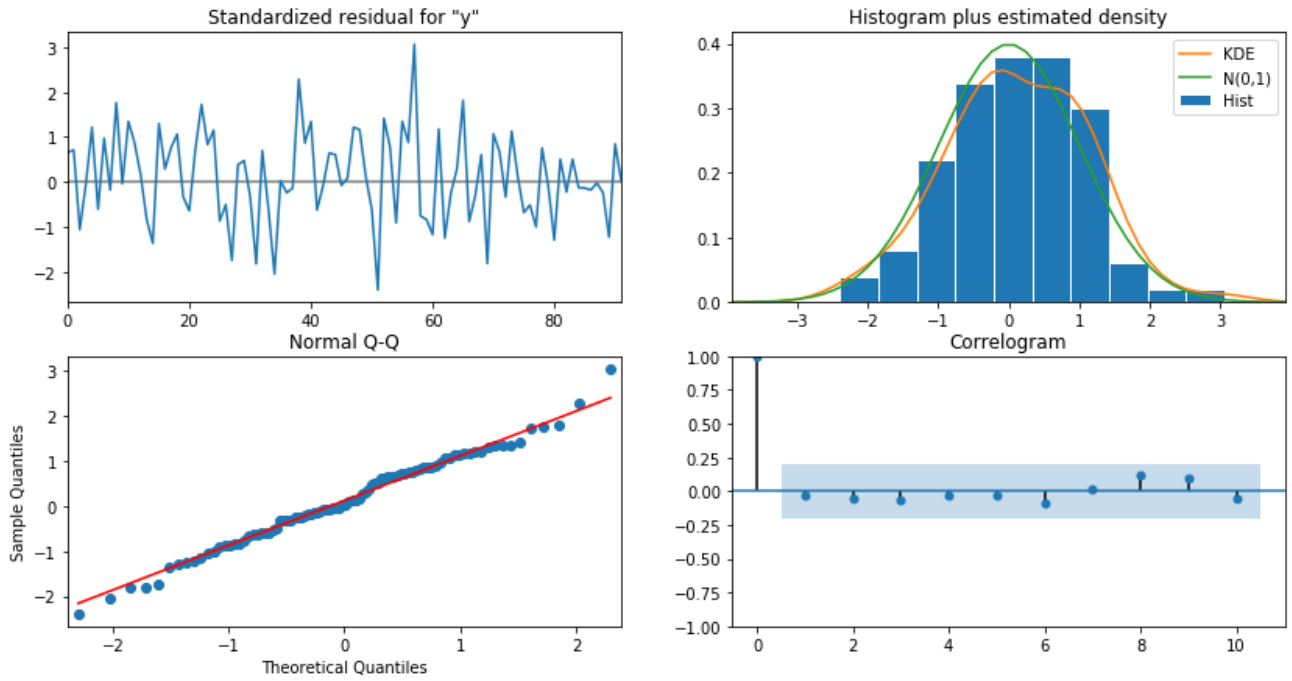


Figure 72 Diagnostic Plot

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-384.369
Date:	Sun, 17 Apr 2022	AIC	786.737
Time:	19:47:58	BIC	809.433
Sample:	0	HQIC	795.898
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8967	0.132	-6.814	0.000	-1.155	-0.639
ar.L2	0.0165	0.171	0.097	0.923	-0.319	0.352
ar.L3	-0.1132	0.174	-0.650	0.515	-0.454	0.228
ar.L4	-0.1598	0.116	-1.380	0.168	-0.387	0.067
ma.L1	0.1508	0.174	0.866	0.387	-0.191	0.492
ma.L2	-0.8492	0.164	-5.166	0.000	-1.171	-0.527
ma.S.L12	-0.3907	0.102	-3.848	0.000	-0.590	-0.192
ma.S.L24	-0.0887	0.091	-0.977	0.329	-0.267	0.089
sigma2	238.9649	0.001	2.02e+05	0.000	238.963	238.967

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	0.01
Prob(Q):	0.80	Prob(JB):	0.99
Heteroskedasticity (H):	0.76	Skew:	-0.01
Prob(H) (two-sided):	0.46	Kurtosis:	3.06

Figure 73 Manual SARIMA Model

YearMonth	Rose	rose_forecasted	rose_forecasted_log	manual_rose_forecasted
1991-01-01	54.0	44.213661	56.850589	44.733041
1991-02-01	55.0	62.326882	66.336719	64.208694
1991-03-01	66.0	67.313375	70.530760	65.110689
1991-04-01	65.0	63.161025	66.370372	68.453063
1991-05-01	60.0	66.474493	69.031105	61.423433

Figure 74 Manual SARIMA Forecasted Values

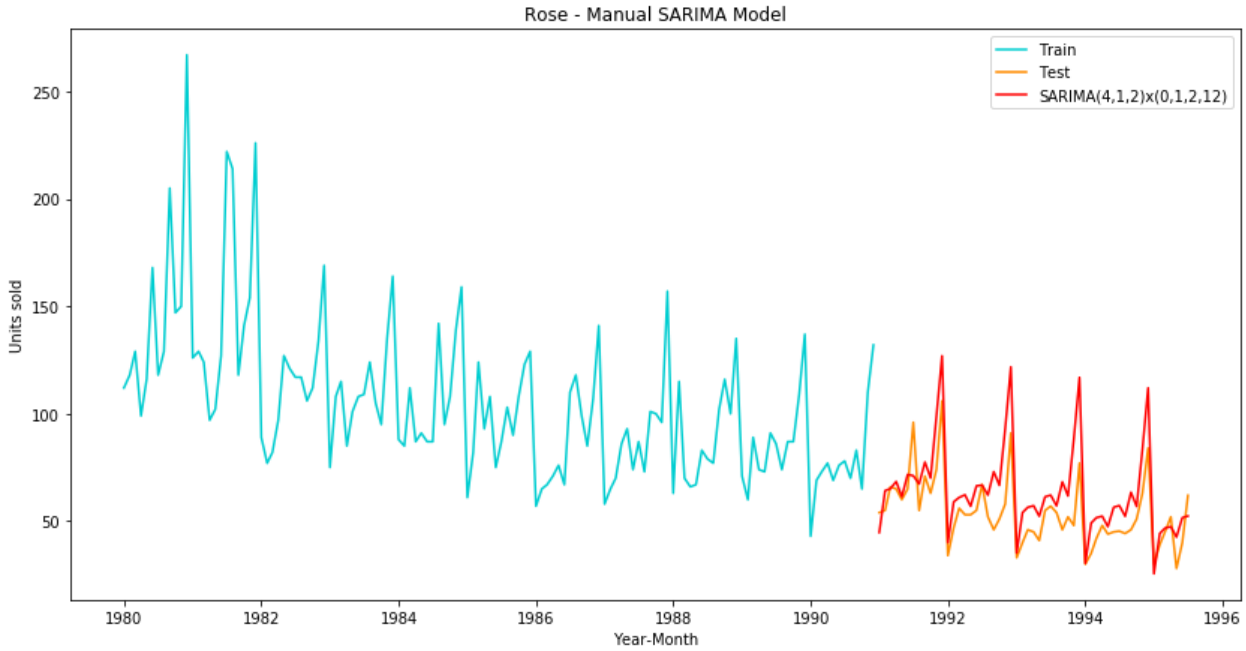


Figure 75 Plot Actual v/s Forecasted Result on test data

2.8 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.

Solution:

	Test RMSE
RegressionOnTime	15.278369
NaiveModel	79.745697
SimpleAverage	53.488233
2 point TMA	11.530054
4 point TMA	14.458402
6 point TMA	14.572976
9 point TMA	14.732918
Alpha=0.0987, SES Optimized	36.824464
Alpha=0.10,SES_Iterative	36.856268
Alpha=0.0,Beta=0.0, DES Optimized	15.718202
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
Alpha=0.065,Beta=0.054,gamma=0.0 TES Optimized	19.434699
Alpha=0.1,Beta=0.2,gamma=0.3,TES_Iterative	9.880143
Auto_ARIMA(0, 1, 2)	45.048584
Auto_SARIMA(0, 1, 2)*(2, 1, 2, 12)	16.527809
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	17.919401
Manual_ARIMA(0,1,0)	79.745697
Manual_SARIMA(4, 1, 2)*(0, 1, 2, 12)	15.388806
Auto_ARIMA(0, 1, 2)	45.048584
Auto_SARIMA(0, 1, 2)*(2, 1, 2, 12)	16.527809
Manual_ARIMA(0,1,0)	79.745697
Manual_SARIMA(4, 1, 2)*(0, 1, 2, 12)	15.388806

Figure 76 RMSE Value

- Triple Exponential Smoothing (Holt Winter's) with alpha: 0.1, beta: 0.2 and gamma: 0.3 is found to be the best model, followed by 2-point trailing moving average model.

2.9 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 overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) is selected for final prediction into 12 months in future.
- TES model alpha: 0.1, beta: 0.2 and gamma: 0.3 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data.
- The model predicts continuation of the trend in sales and seasonality in year-end sales. The prediction shows a stabilization of downward trend, as the sales will be almost same as previous observed year.
- The RMSE value of TES obtained for the entire dataset is 17.88

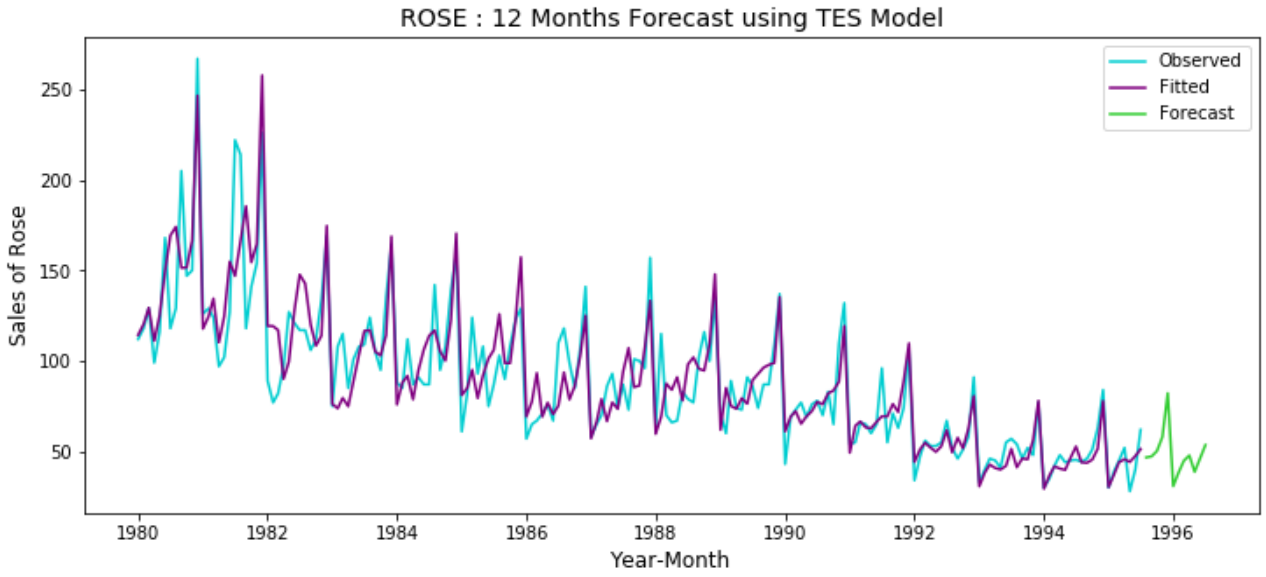


Figure 77 Plot Actual and Future Forecast result.

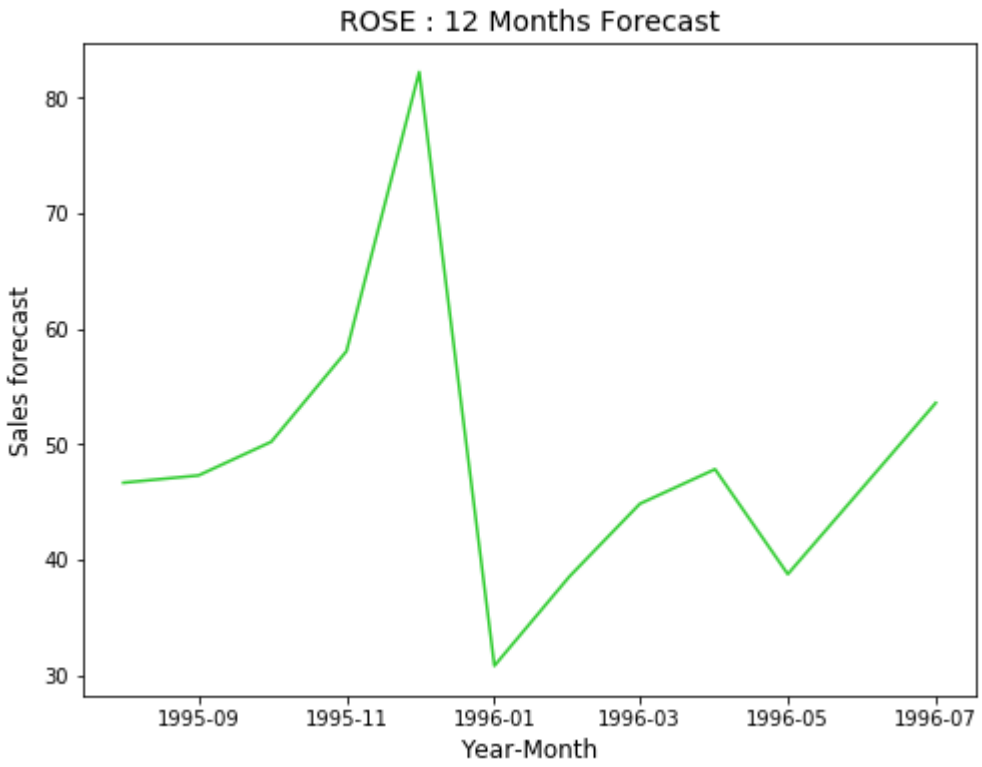


Figure 78 Future Forecast Plot

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

1995-08-01	46.645790
1995-09-01	47.277865
1995-10-01	50.192393
1995-11-01	58.032966
1995-12-01	82.211767
1996-01-01	30.793144
1996-02-01	38.536058
1996-03-01	44.822234
1996-04-01	47.814473
1996-05-01	38.727986
1996-06-01	46.255070
1996-07-01	53.559025

Freq: MS, dtype: float64

count	12.000000
mean	48.739064
std	12.747211
min	30.793144
25%	43.298672
50%	46.961827
75%	51.034051
max	82.211767

dtype: float64

Table 35 Future Forecast Result and summary statistics

- The model forecasts sale of 585 units of Rose wine in 12 months into future. Which is an average sale of 48 units per month.
- The seasonal sale in December 1995 will reach a maximum of 82 units, before it drops to the lowest sale in January 1996; at 30 units.
- Unlike Sparkling wine, Rose wine sells very low number of units and the standard deviation is only 12.75. Which means that higher demand does not impact procurement and production.
- The ABC estate wine should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions.