TIME SERIES FORECASTING BUSINESS REPORT

PGP-DSBA

FEBRUARY 21, 2021
TSF BUSINESS REPORT
Srikanthpr.27@gmail.com

Contents

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

Data set for the Problem: Sparkling.csv and Rose.csv

Please do perform the following questions on each of these two data sets separately.

1 Problem Statement – TSF -Sparkling

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

Solution:

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

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

Figure-1: View Head of the data without panda's date-time format

The dataset 'Sparkling' contain two columns of data:
The monthly time stamp from Jan 1980 to July 1995 and the sales corresponding to the wines.

Method-1:
Create Time Stamps and adding it to the data frame to make it a Time-series data.

```
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')
```

Figure-2: Create Date-Range

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.

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

Figure-3: View Head of the Time-Series data

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]

View the top 5 rows of Sparkling dataset :

```
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
```

View the bottom 5 rows of Sparkling dataset :

```
YearMonth
1995-03-01    1897
1995-04-01    1862
1995-05-01    1670
1995-06-01    1688
1995-07-01    2031
Name: Sparkling, dtype: int64
```

Figure-4: View Head and Tail of the Time-Series data

→ All values are properly loaded for the dataset with the index as panda's date-time format.
→ Sparkling time series data do not contain any missing values.

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

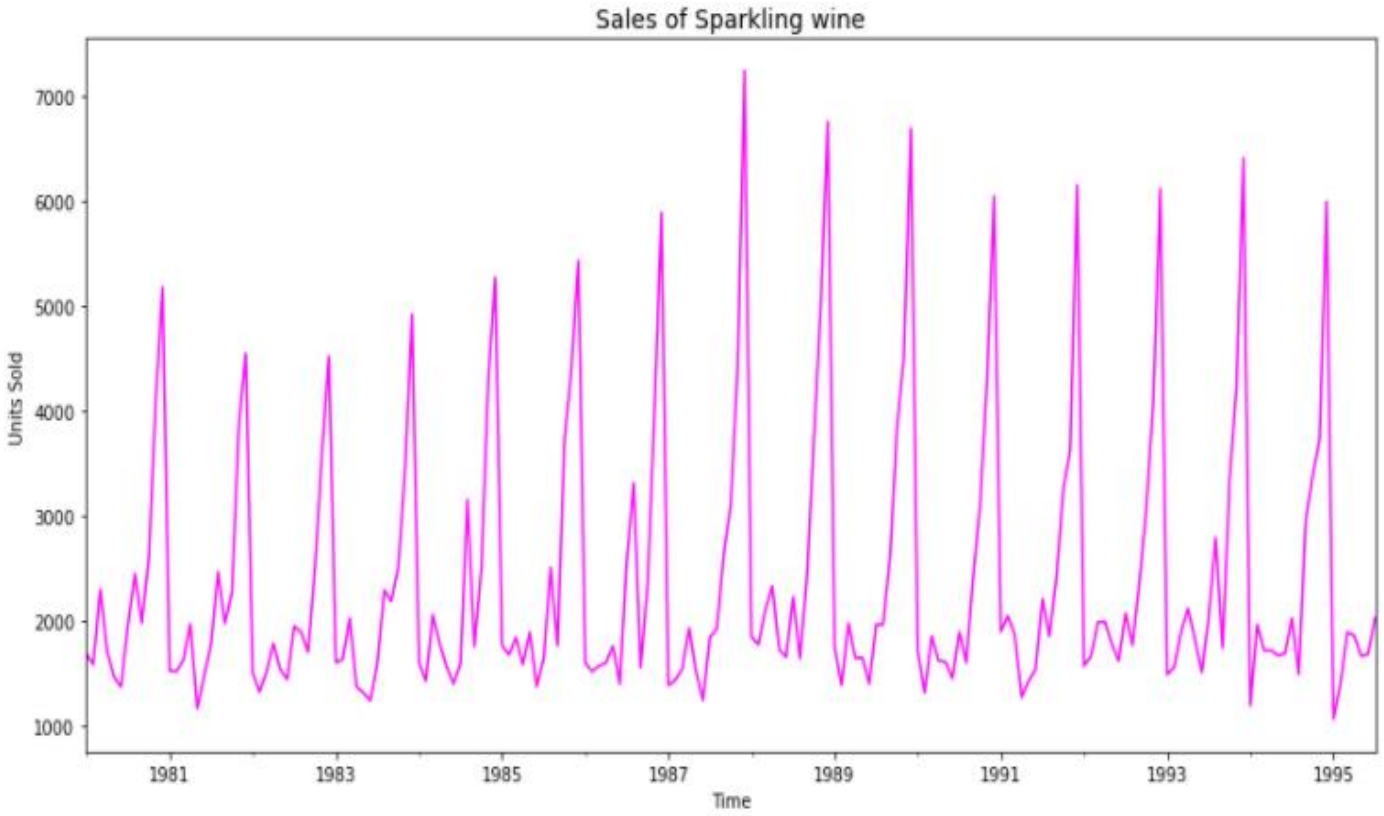


Figure-5: Plot Sparkling Time Series data

→ 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 favoured over the years by customers.

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

Solution:

Check the basic measures of descriptive statistics:

Data Description for Sparkling Dataset:

```
count     187.000000
mean     2402.417112
std      1295.111540
min      1070.000000
25%      1605.000000
50%      1874.000000
75%      2549.000000
max      7242.000000
Name: Sparkling, dtype: float64
```

Figure-6: Summary Statistics

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.

Yearly Boxplot:

Yearly Boxplot for Sparkling Dataset:

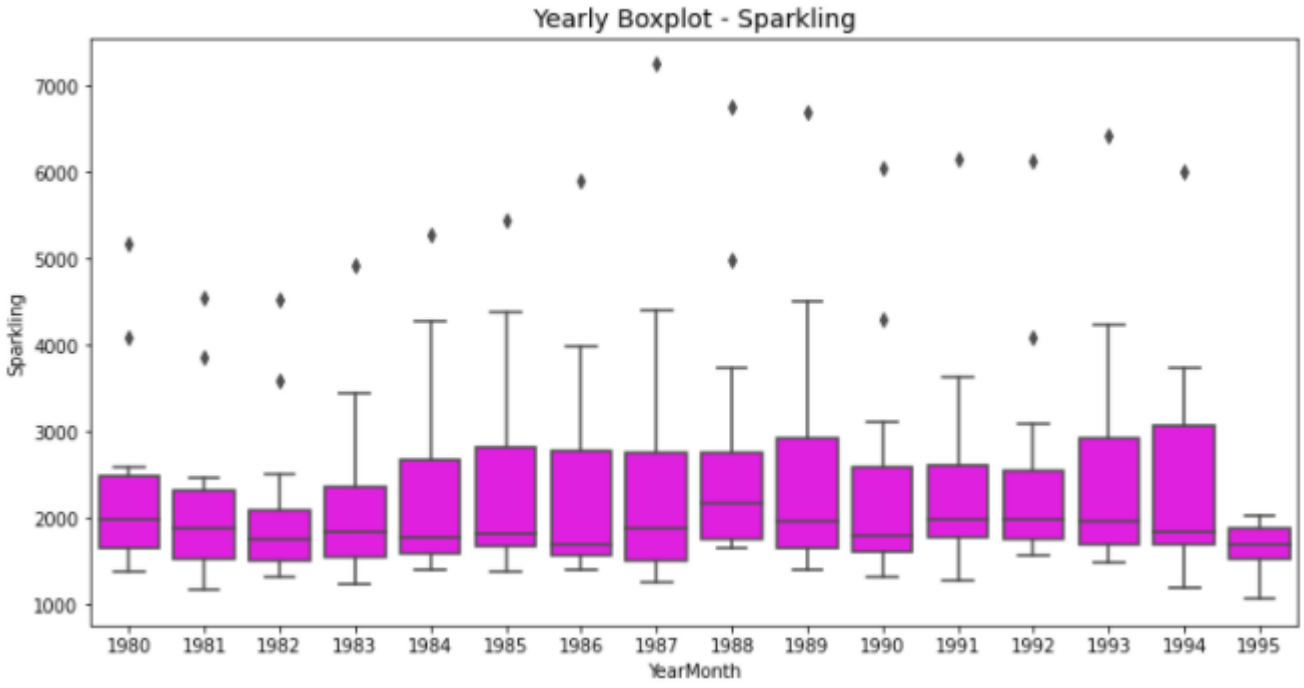


Figure-7: Yearly Boxplot

Monthly Boxplot:

Monthly Boxplot for all the years for Sparkling Dataset:

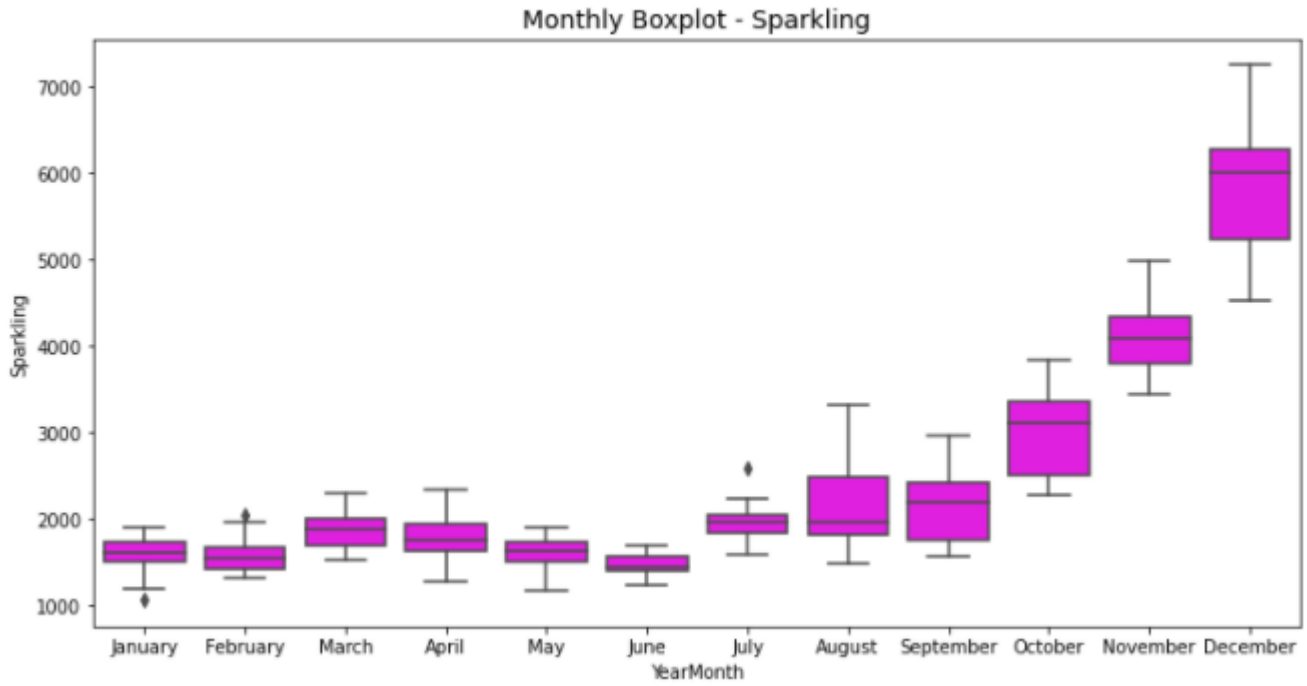


Figure-8: Monthly Boxplot

→ 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 tanks in the month of June.

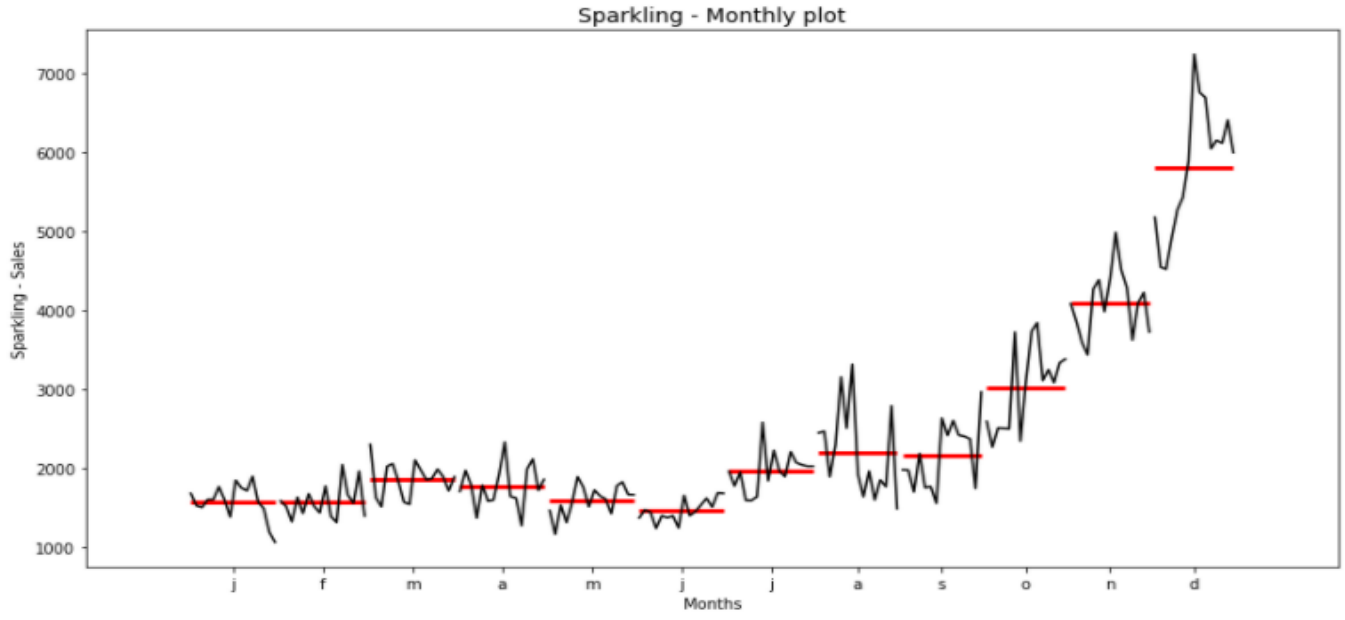


Figure-9: Monthly plot

→ 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.

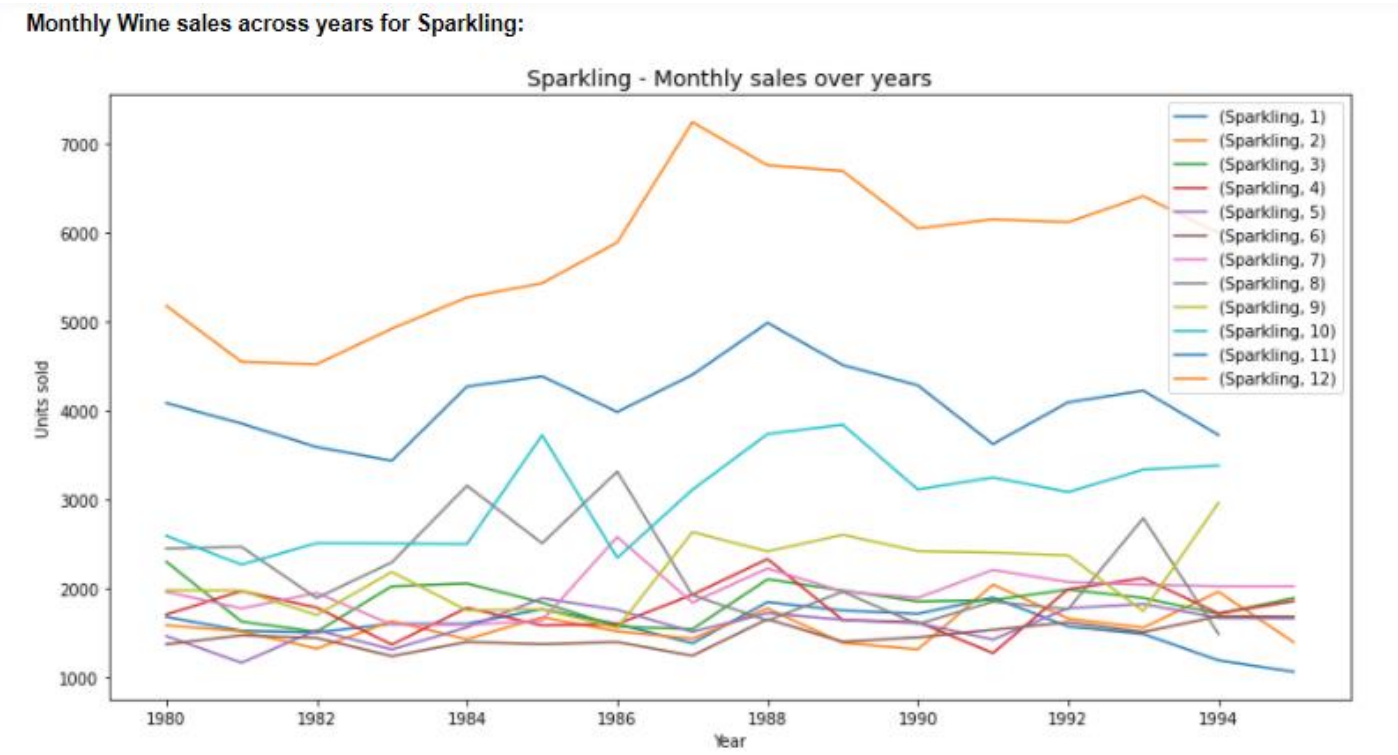


Figure-10: Monthly Sales over years

→ 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:

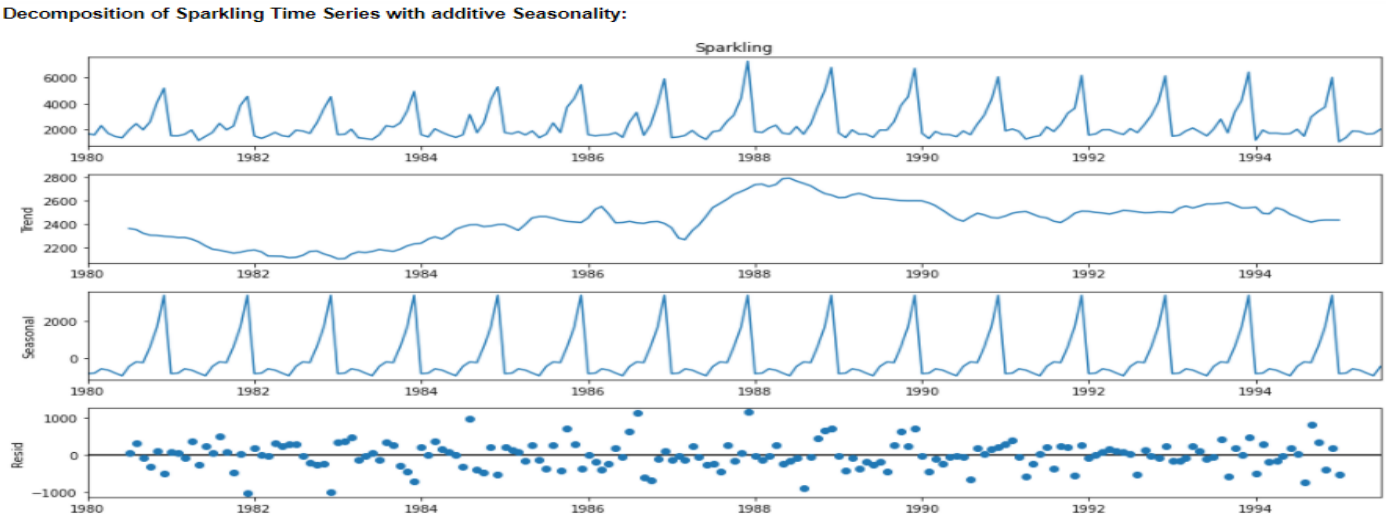


Figure-11: Additive Model

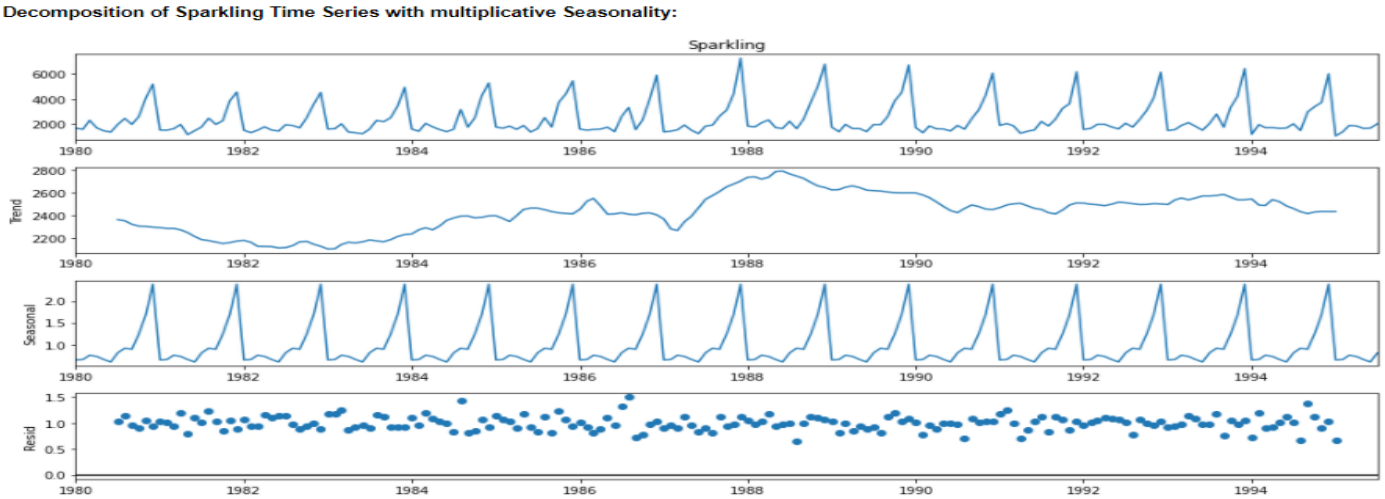


Figure-12: Multiplicative Model

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.

Solution:

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

```
train_spark = spark[spark.index.year < 1991]
test_spark = spark[spark.index.year >= 1991]
```

First few rows of Training Data:

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

First few rows of Test Data:

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

Last few rows of Training Data:

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

Last few rows of Test Data:

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

Figure-13: Train and Test Data

The Plot Sparkling Time Series as train and test

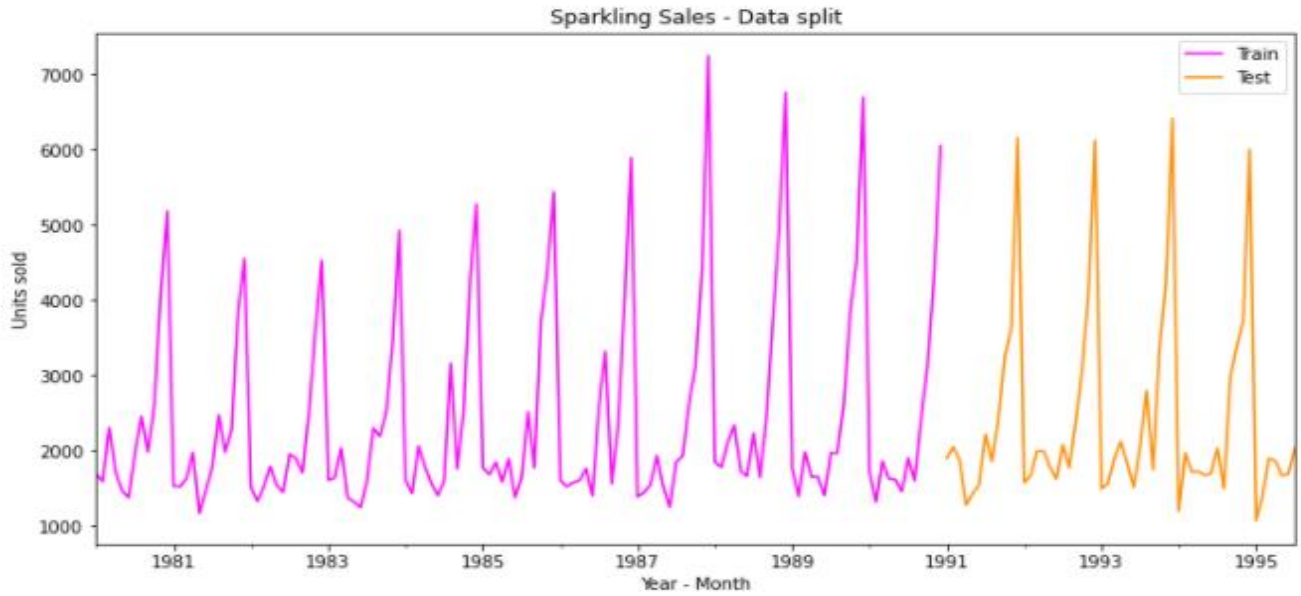


Figure-14: The Plot Sparkling Time Series as train and test

1.4 Build various 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Note: Please do try to build as many models as possible and as many iterations of models as possible with different parameters.

Solution:

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

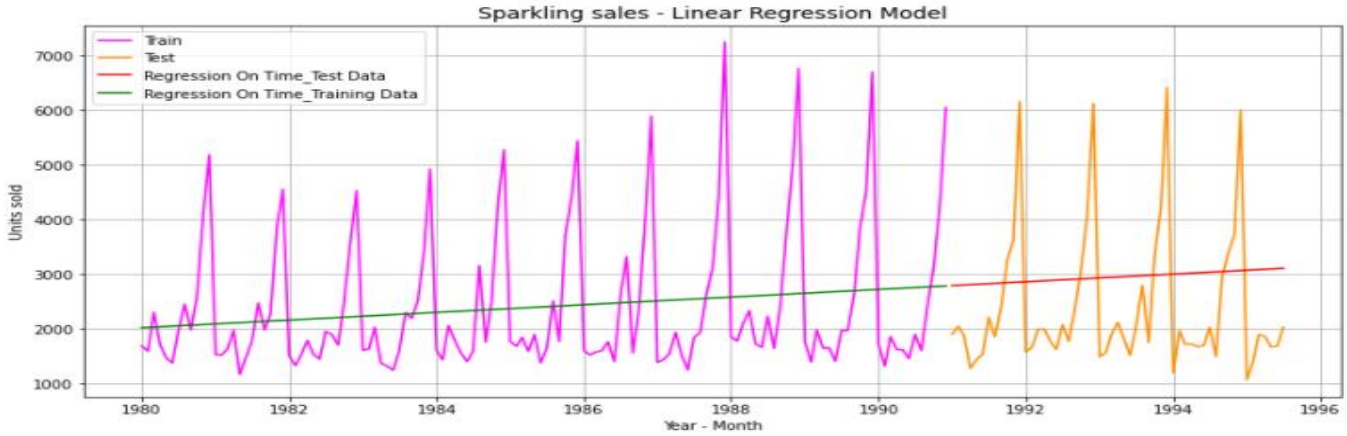


Figure-15: Linear Regression Model

→ 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

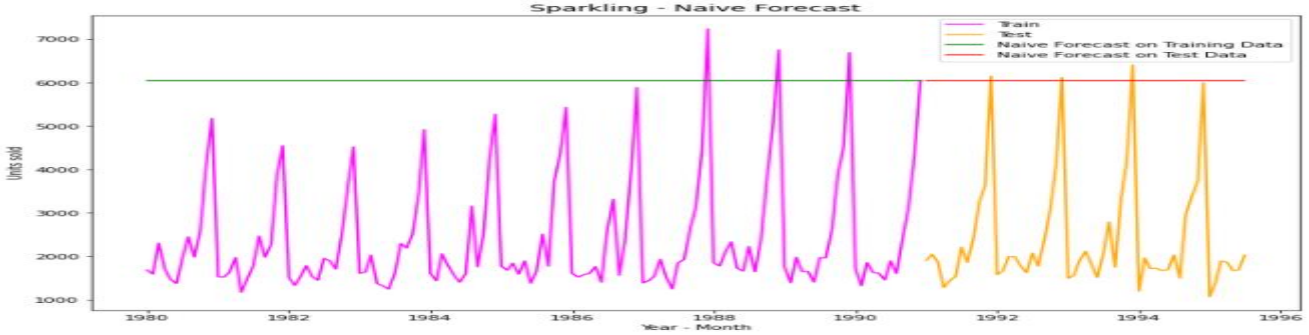


Figure-16: Naïve-Forecast Model

→ 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.

Model 3: Simple Average Model

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

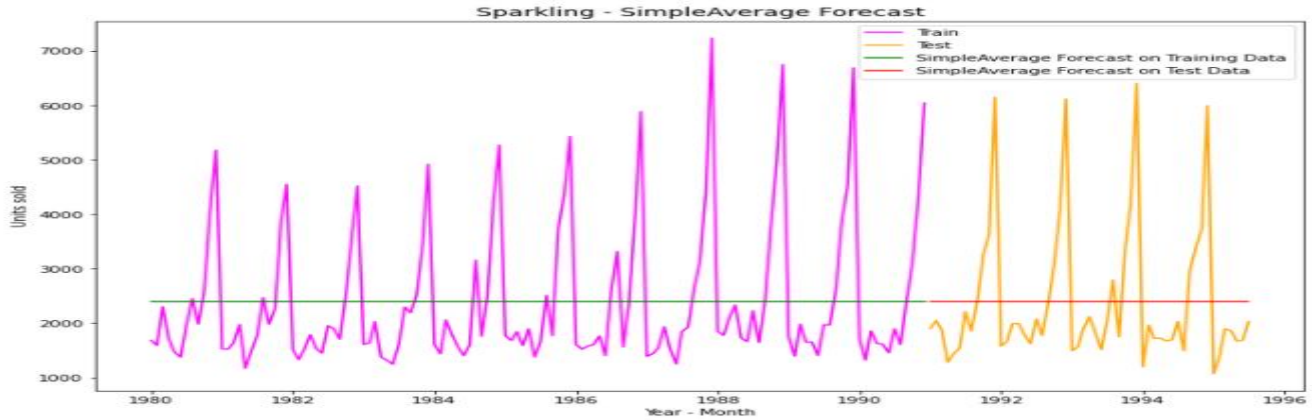


Figure-17: Simple Average Model

→ 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 1275

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

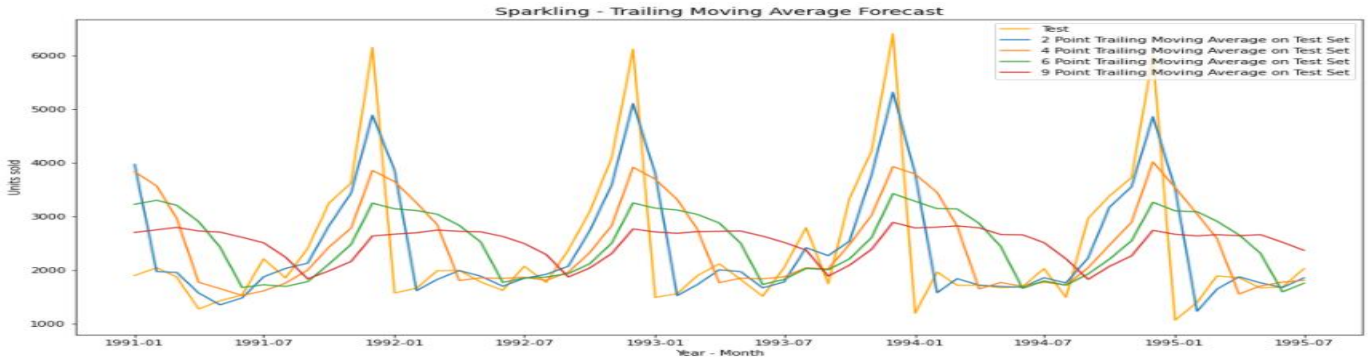


Figure-18: Trailing Moving Average Model

Model Comparison

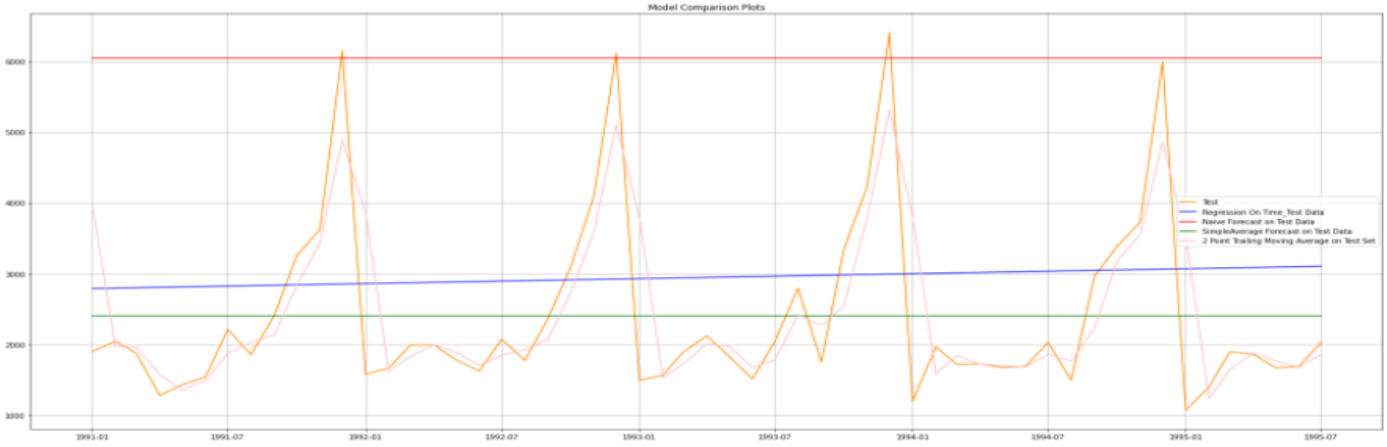


Figure-19: Model Comparison

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

Figure-20: RMSE for Test Set

Model 5: Simple Exponential Smoothing Model

- → 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.

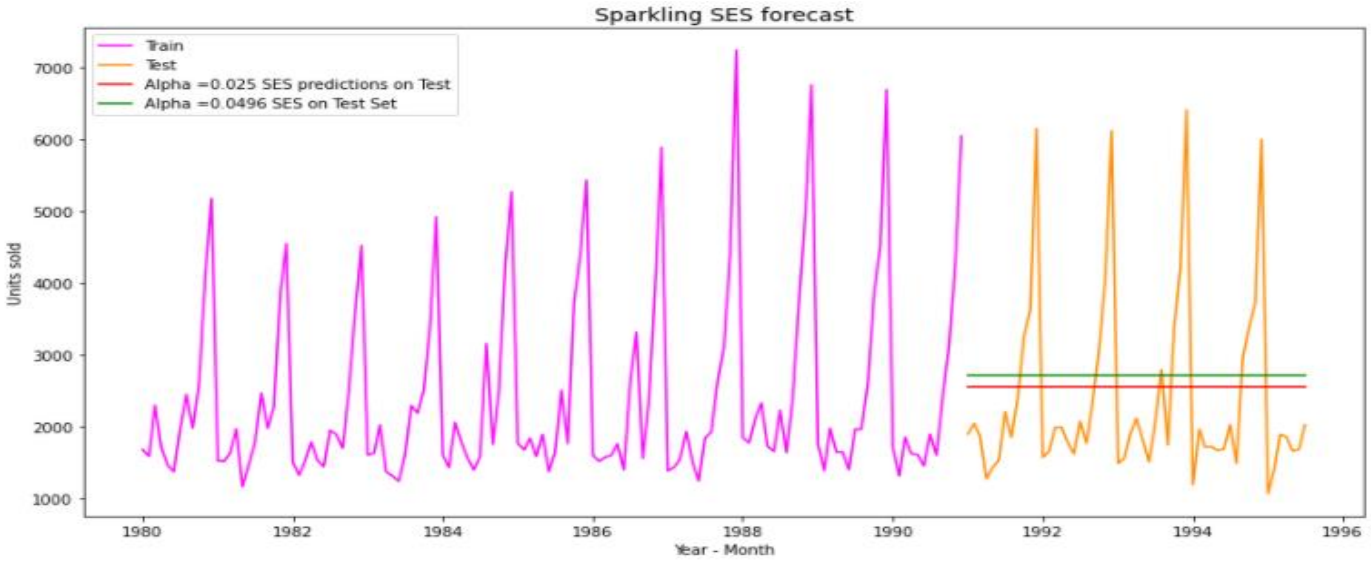


Figure-21: SES Optimised and Iterative Model

Model 6: Double Exponential Smoothing Model

→ 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.

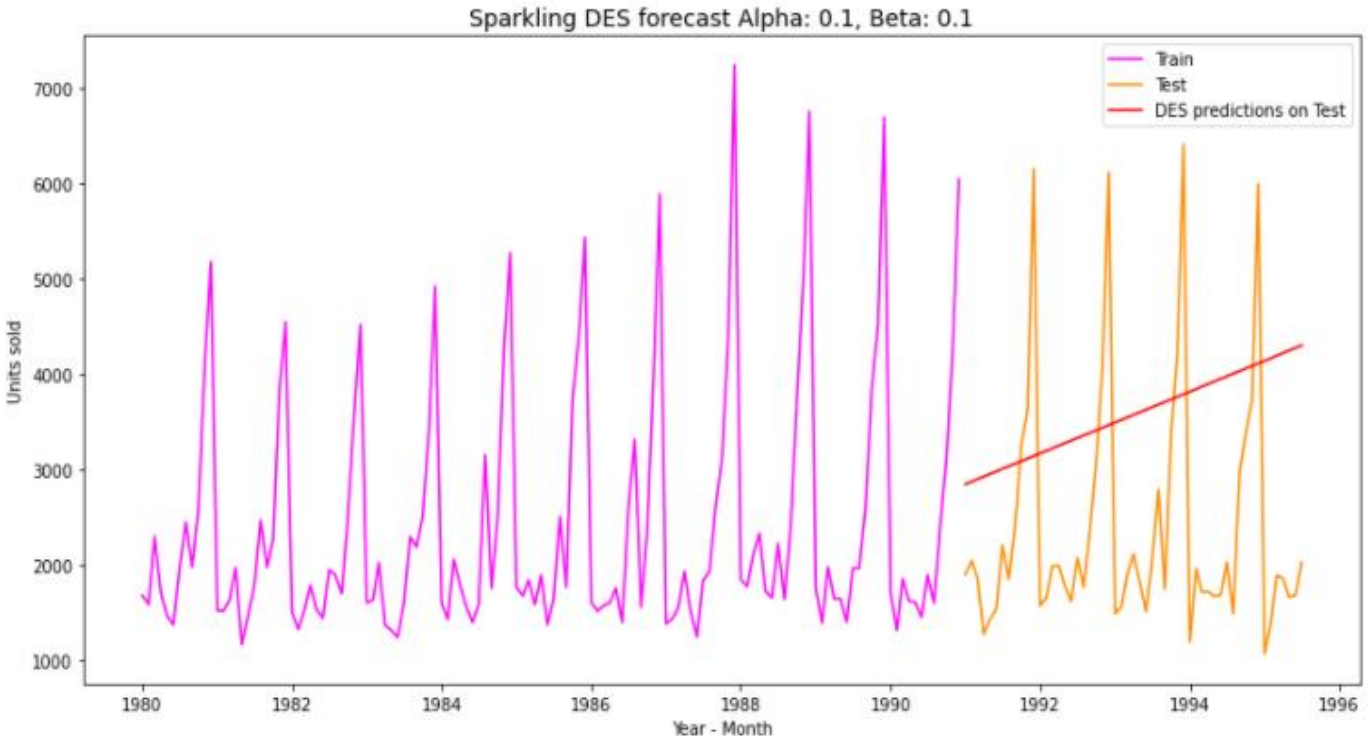


Figure-22: DES Iterative Model

Model 7: Triple Exponential Smoothing 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.

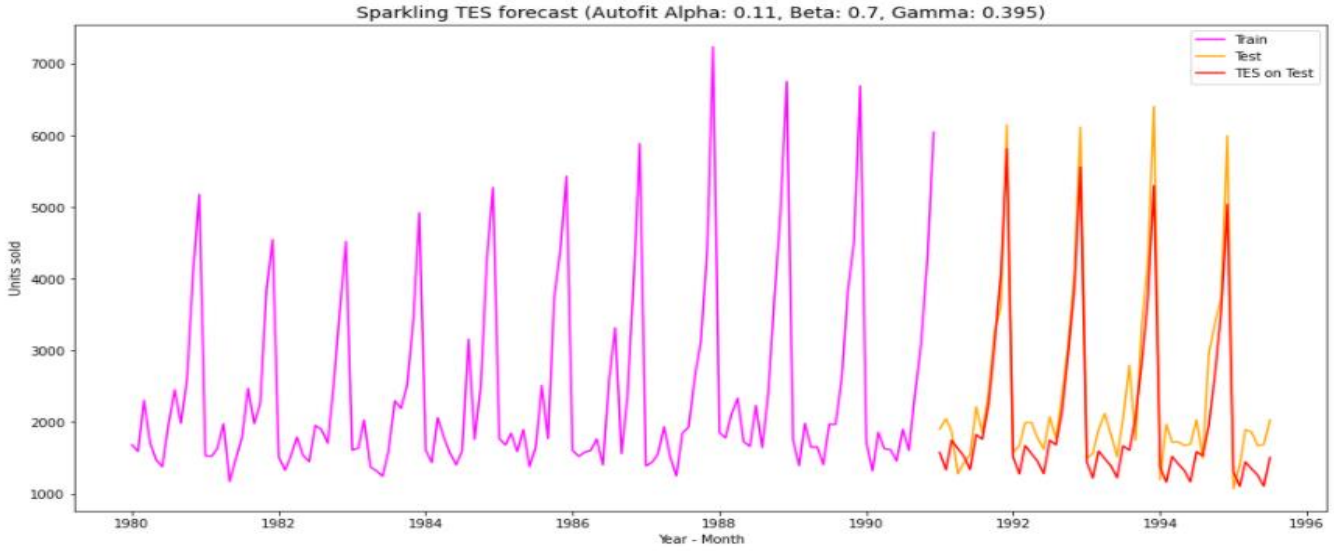


Figure-23: TES Auto-fit Model

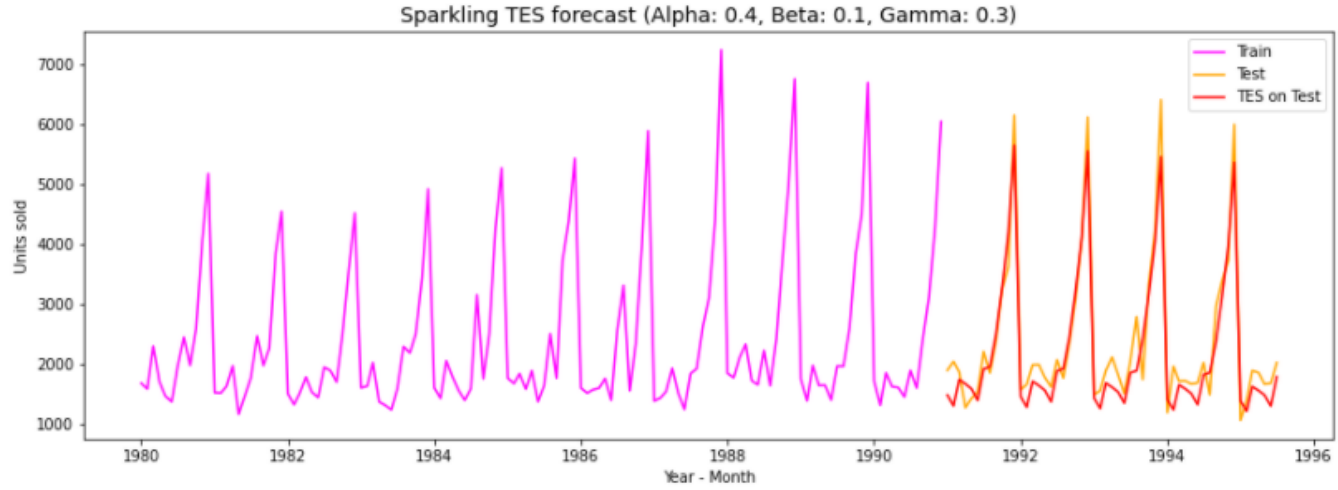


Figure-24: TES Iterative Model

Model Comparison:

	Test RMSE
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	371.367690
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	469.432003
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=0.0496, SES Optimized	1316.034674
9 point TMA	1346.278315
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DES iterative	1778.560000
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
NaiveModel	3864.279352

Figure-25: Test RMSE Values

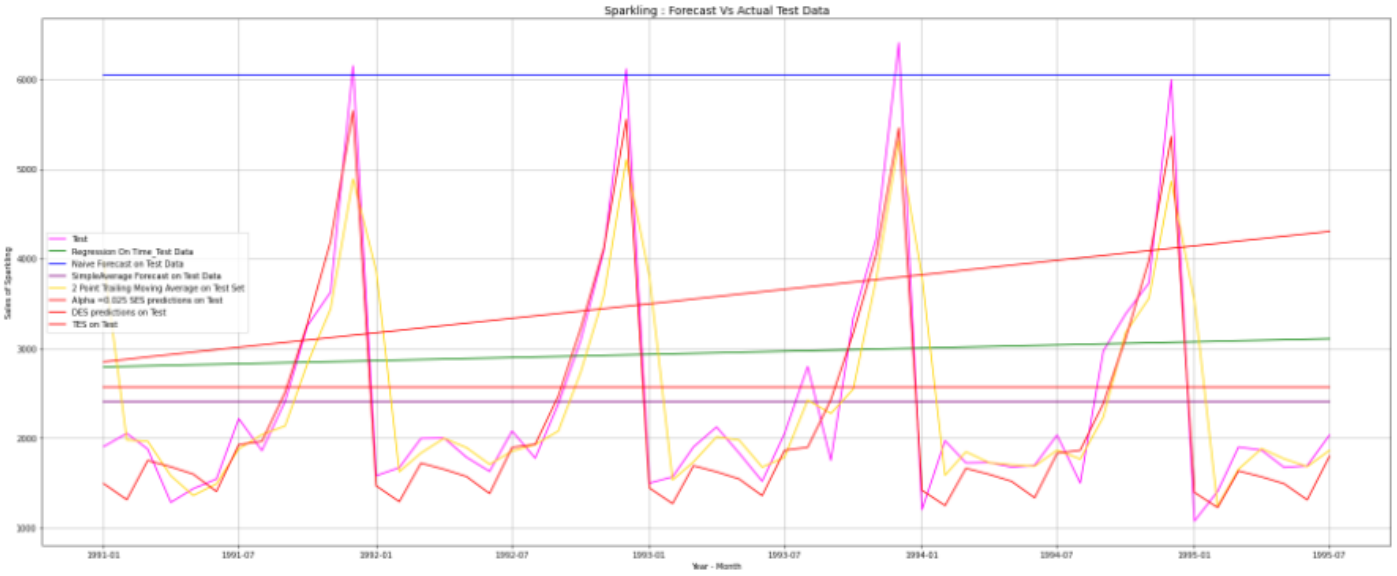


Figure-26: Sparkling Forecast v/s Actual Values

→ 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.

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

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 Sparkling 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                  -1.360497
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
```

ADF on original series
- P-Value > alpha .05
- Test statistic > Critical values
- Fail to reject the null hypothesis
- The series is non-stationary

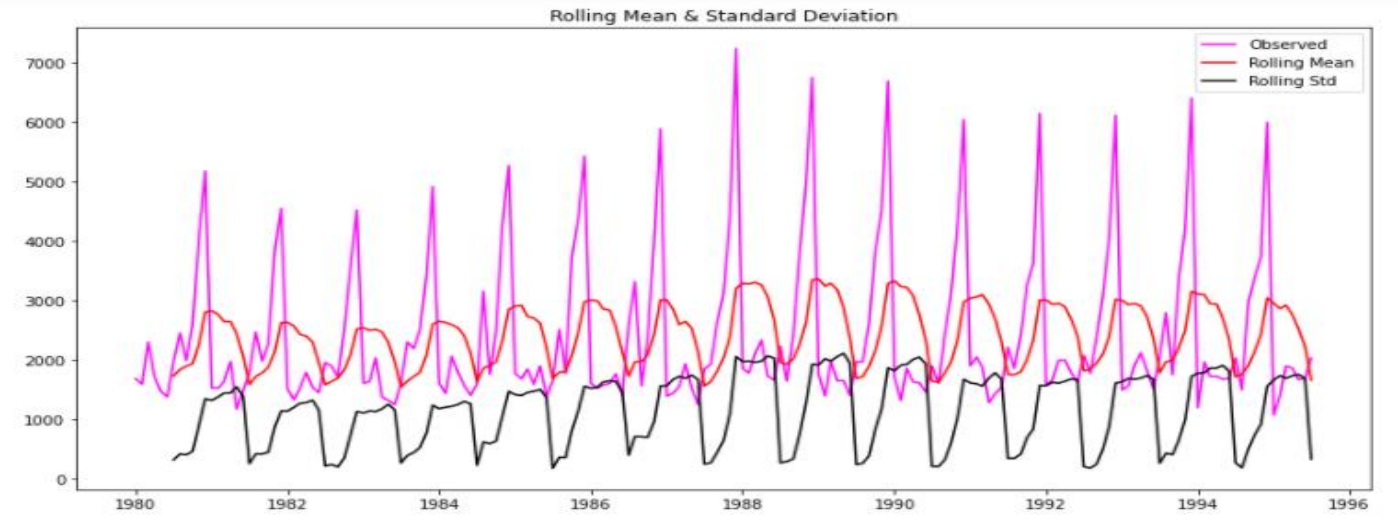


Figure-27: ADF test on Original Series

- → 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.

→ 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                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

ADF on differenced series

- P-Value < alpha .05
- Test statistic < Critical values
- Reject the null hypothesis
- The series is stationary

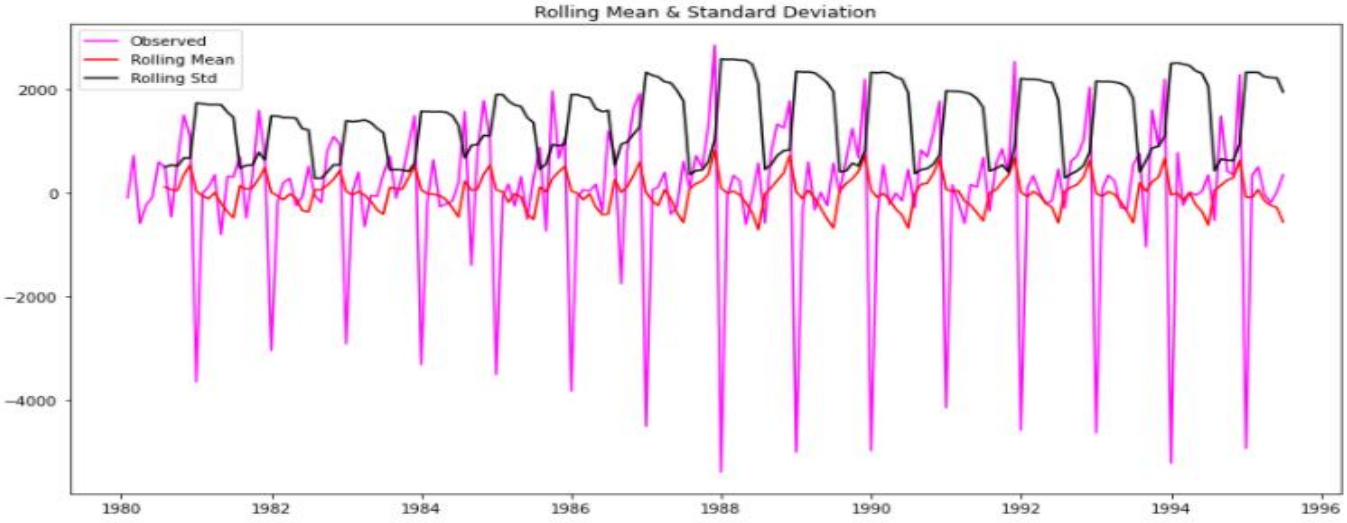
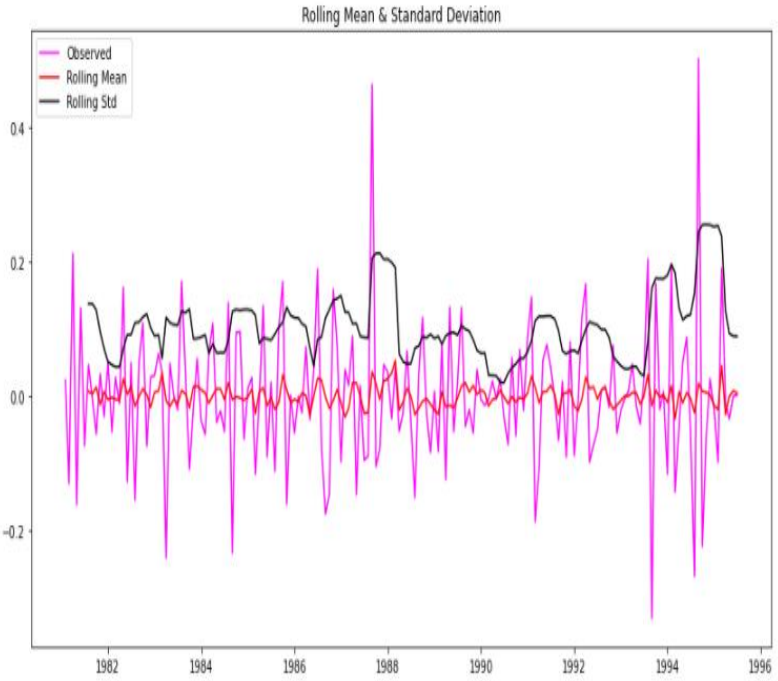
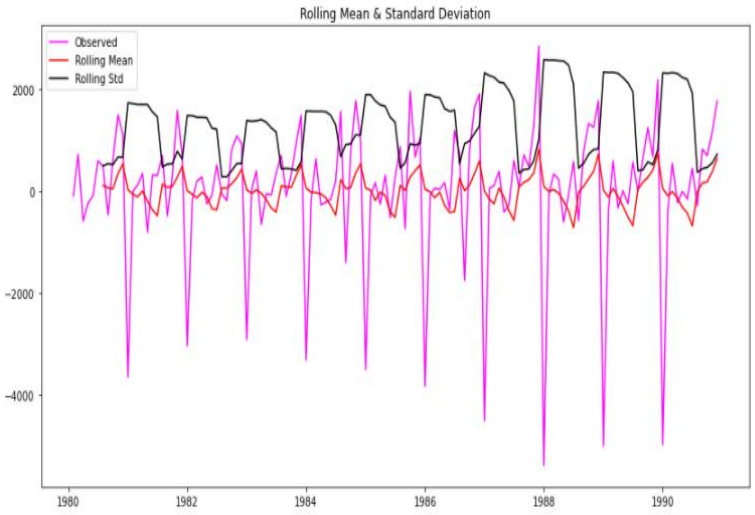


Figure-28: ADF test after differencing d=1




```
Results of Dickey-Fuller Test:
Test Statistic                  -5.254601
p-value                          0.000007
#Lags Used                      12.000000
Number of Observations Used    161.000000
Critical Value (1%)             -3.471633
Critical Value (5%)             -2.879665
Critical Value (10%)            -2.576434
dtype: float64
```

Figure-29: ADF test on log series after differencing




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

Figure-30: ADF test on train data after differencing 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.

Solution:

Model 8: Auto-ARIMA Model

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.311
Method:                       css-mle   S.D. of innovations           1013.283
Date:                Mon, 22 Feb 2021   AIC                           2210.621
Time:                        17:35:07   BIC                           2227.873
Sample:                    02-01-1980   HQIC                          2217.631
                         - 12-01-1990
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              5.5852      0.518     10.789      0.000       4.571       6.600
ar.L1.D.Sparkling  1.2699      0.075     17.043      0.000       1.124       1.416
ar.L2.D.Sparkling -0.5601      0.074     -7.617      0.000      -0.704      -0.416
ma.L1.D.Sparkling -1.9966      0.042    -47.026      0.000      -2.080      -1.913
ma.L2.D.Sparkling  0.9966      0.043     23.428      0.000       0.913       1.080
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1335           -0.7074j            1.3361           -0.0888
AR.2            1.1335           +0.7074j            1.3361            0.0888
MA.1            1.0003           +0.0000j            1.0003            0.0000
MA.2            1.0030           +0.0000j            1.0030            0.0000
-----------------------------------------------------------------------------
```

Figure-31: ARIMA Model Result

→ 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.

Model 9a: Auto-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.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-685.174
Date:	Mon, 22 Feb 2021	AIC	1382.348
Time:	17:36:47	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

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

Figure-32: SARIMA Model Result

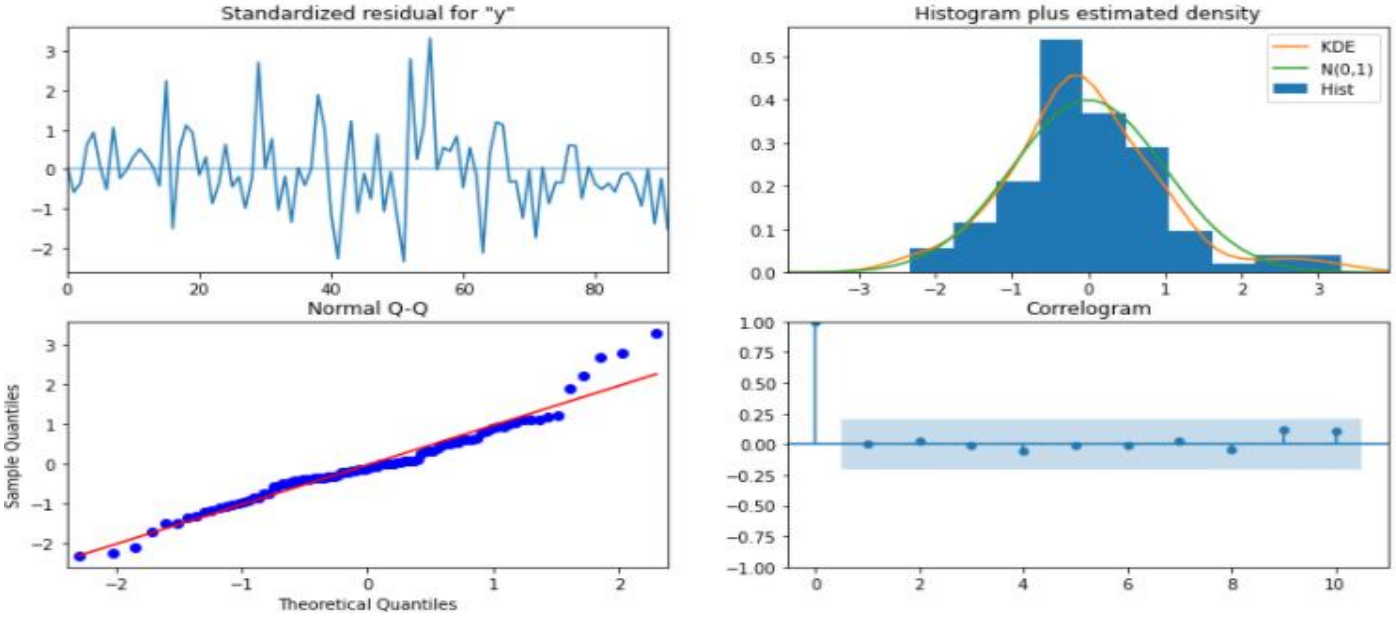


Figure-33: Diagnostic-plot

→ 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.
→ The RMSE values of the automated SARIMA model is 382.58

YearMonth	Sparkling	spark_forecasted
1991-01-01	1902	1460.244631
1991-02-01	2049	1392.437177
1991-03-01	1874	1743.201708
1991-04-01	1279	1650.066938
1991-05-01	1432	1522.656035

Figure-34: Forecasted Result on test data

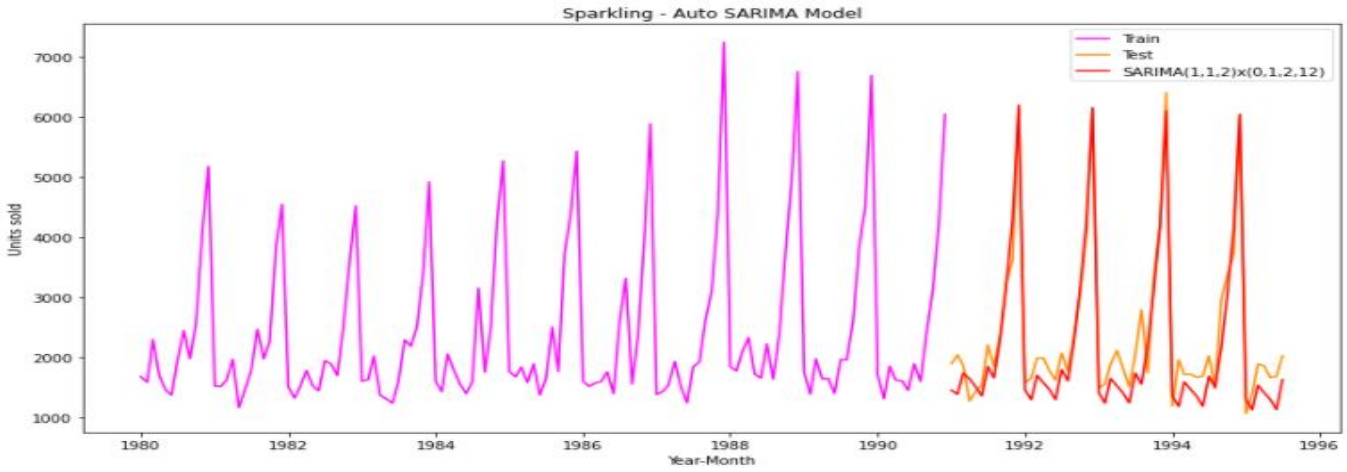


Figure-35: Plot of Actual v/s Forecasted Result on test data

Model 9b: Auto-SARIMA Model on log series data

→ 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 284.48 was obtained at (0, 1, 1)*(1, 0, 1, 12).

→ The model was built with the above parameters.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  132
Model:             SARIMAX(0, 1, 1)x(1, 0, 1, 12)   Log Likelihood                 146.236
Date:                            Mon, 22 Feb 2021   AIC                           -284.472
Time:                                    17:40:11   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
===================================================================================

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

Figure-36: Log Series SARIMA Model Result

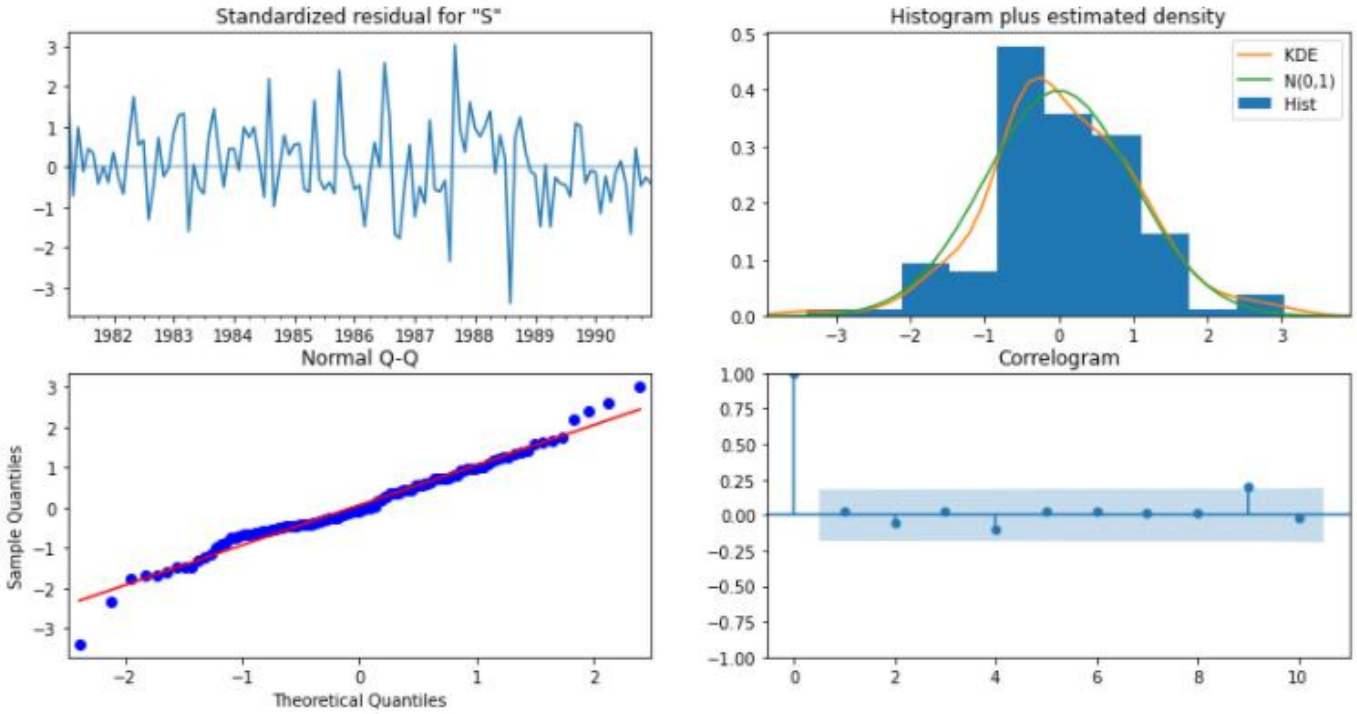


Figure-37: Diagnostic-plot

- → 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

YearMonth	Sparkling	spark_forecasted	spark_forecasted_log
1991-01-01	1902	1460.244631	1629.416485
1991-02-01	2049	1392.437177	1384.547008
1991-03-01	1874	1743.201708	1804.206479
1991-04-01	1279	1650.066938	1685.514276
1991-05-01	1432	1522.656035	1569.597811

Figure-38: Forecasted Result on test data

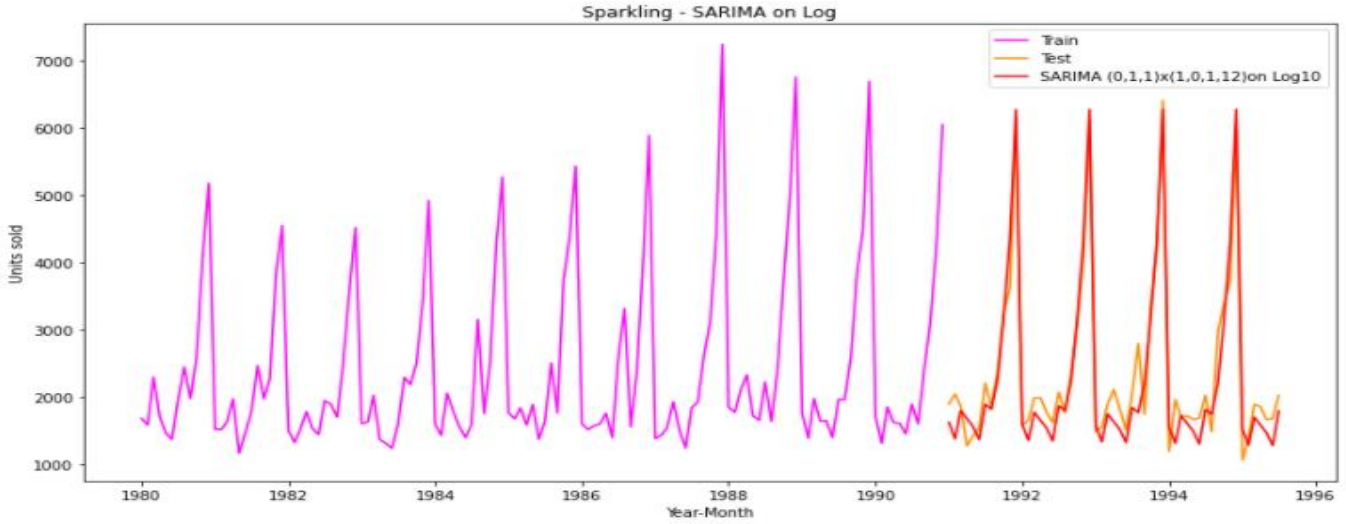


Figure-39: Plot of Actual v/s Forecasted Result on test data

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

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.

Solution:

Model 10: Manual ARIMA

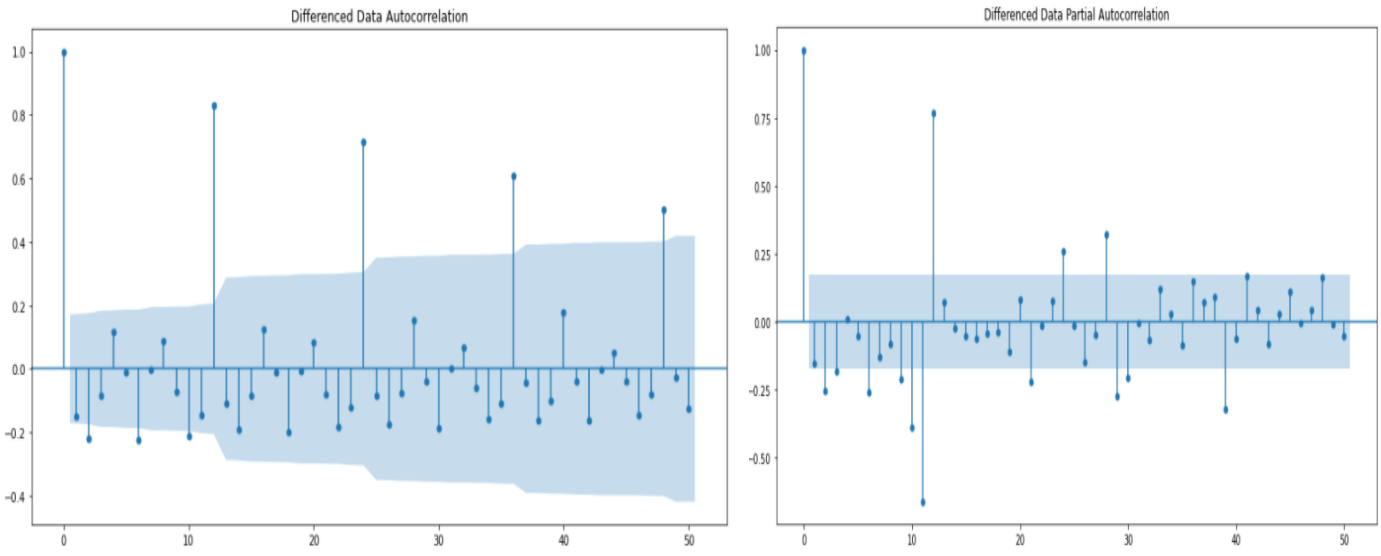


Figure-40: ACF and PACF Plots

→ 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.

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.791
Method:	css	S.D. of innovations	1377.911
Date:	Mon, 22 Feb 2021	AIC	2269.583
Time:	23:25:08	BIC	2275.333
Sample:	02-01-1980 - 12-01-1990	HQIC	2271.919

	coef	std err	z	P>|z|	[0.025	0.975]
const	33.2901	120.389	0.277	0.782	-202.667	269.248

Figure-41: Manual ARIMA Summary Results

→ The RMSE value of manual ARIMA model is 4780. Since the ARIMA model do not capture the seasonality, this model do not perform well.

Model 11: Manual SARIMA

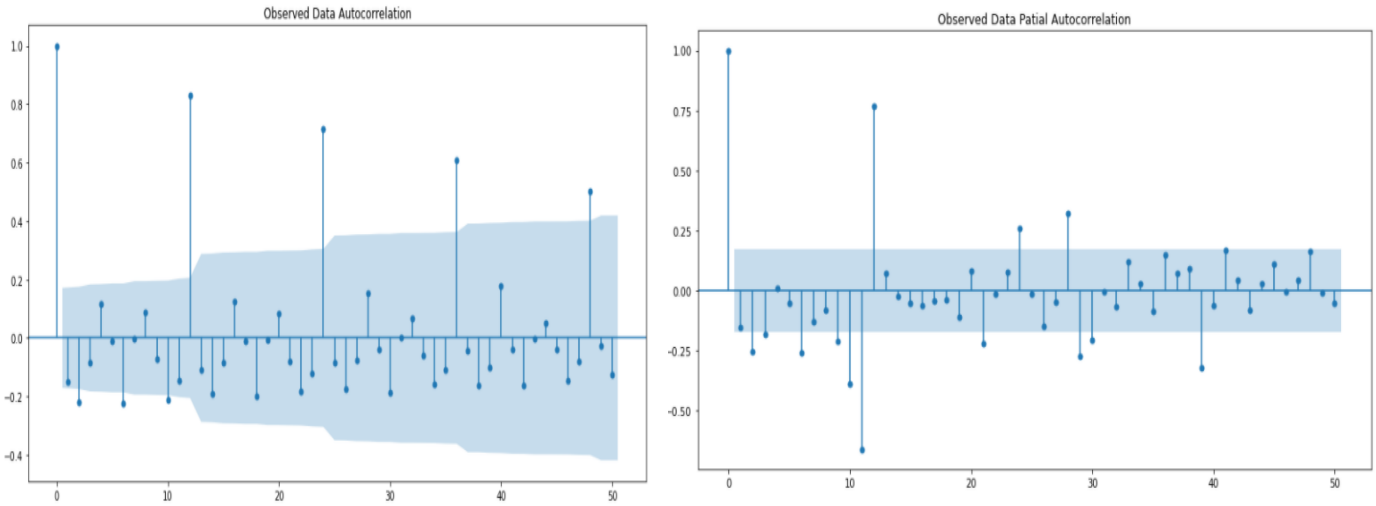
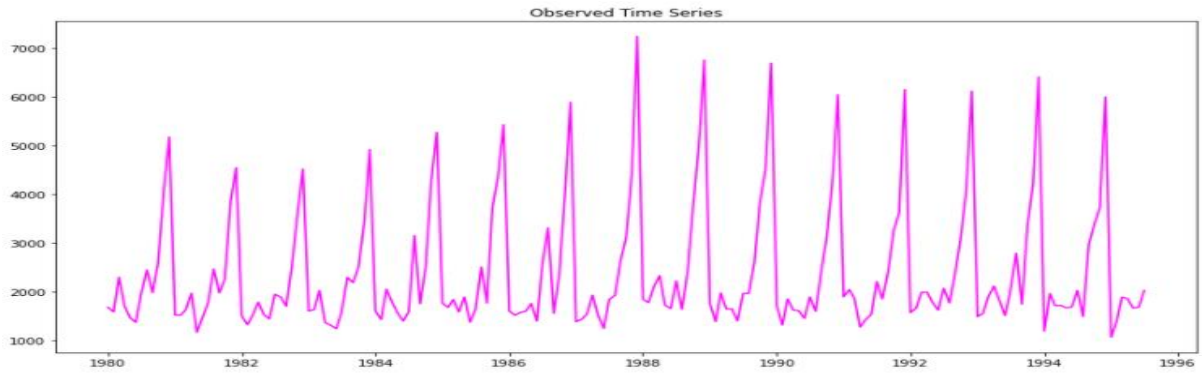


Figure-42: ACF and PACF Plots

→ 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



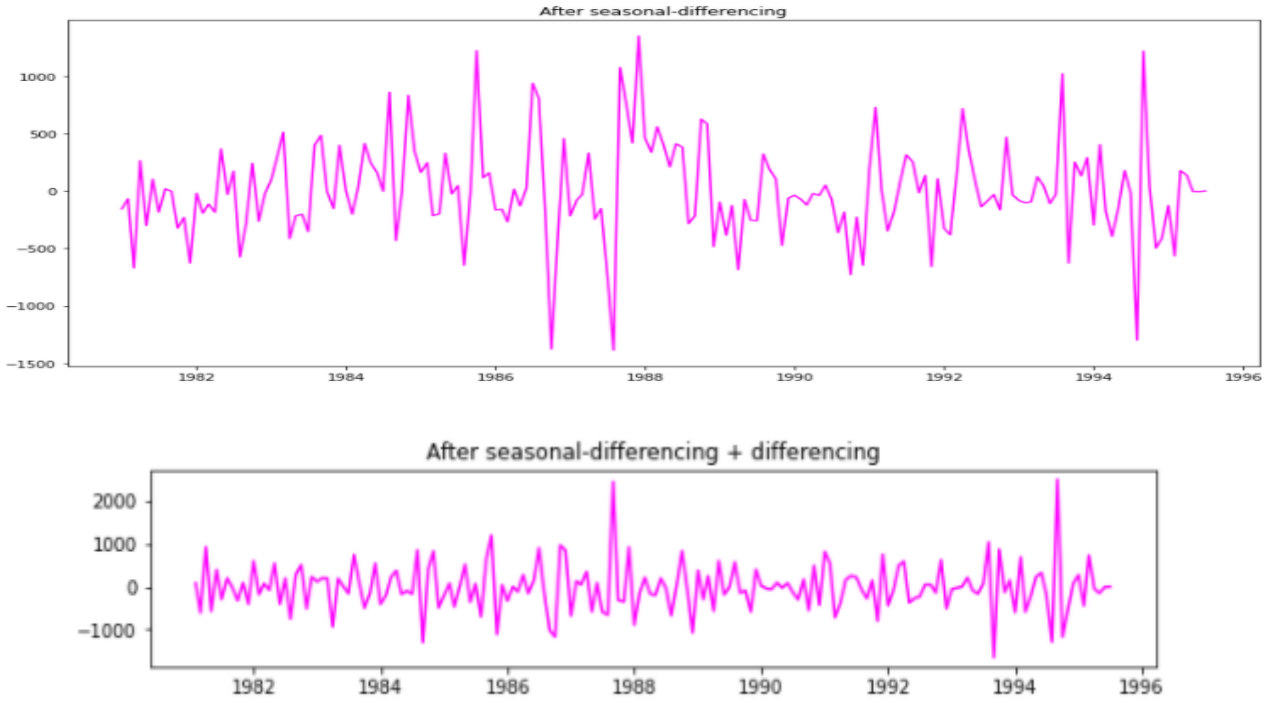
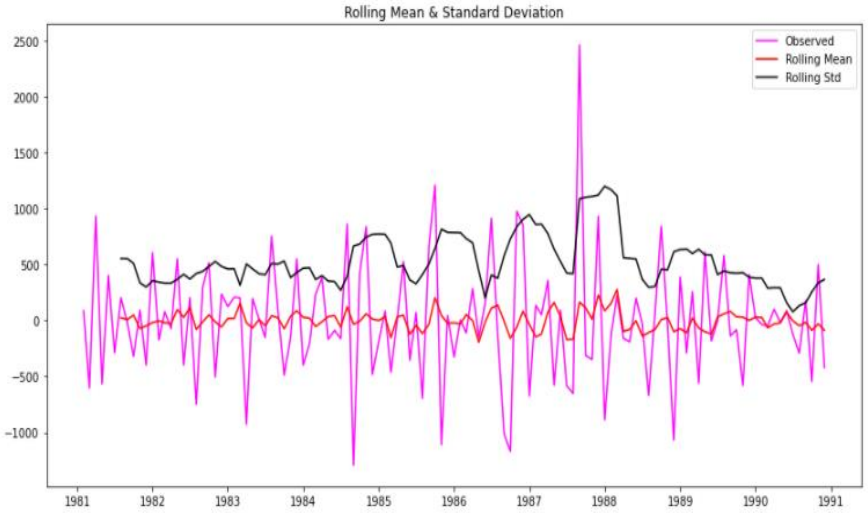


Figure-43: Time series plots

→ 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.




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
#Lags Used                      10.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

Figure-44: 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).

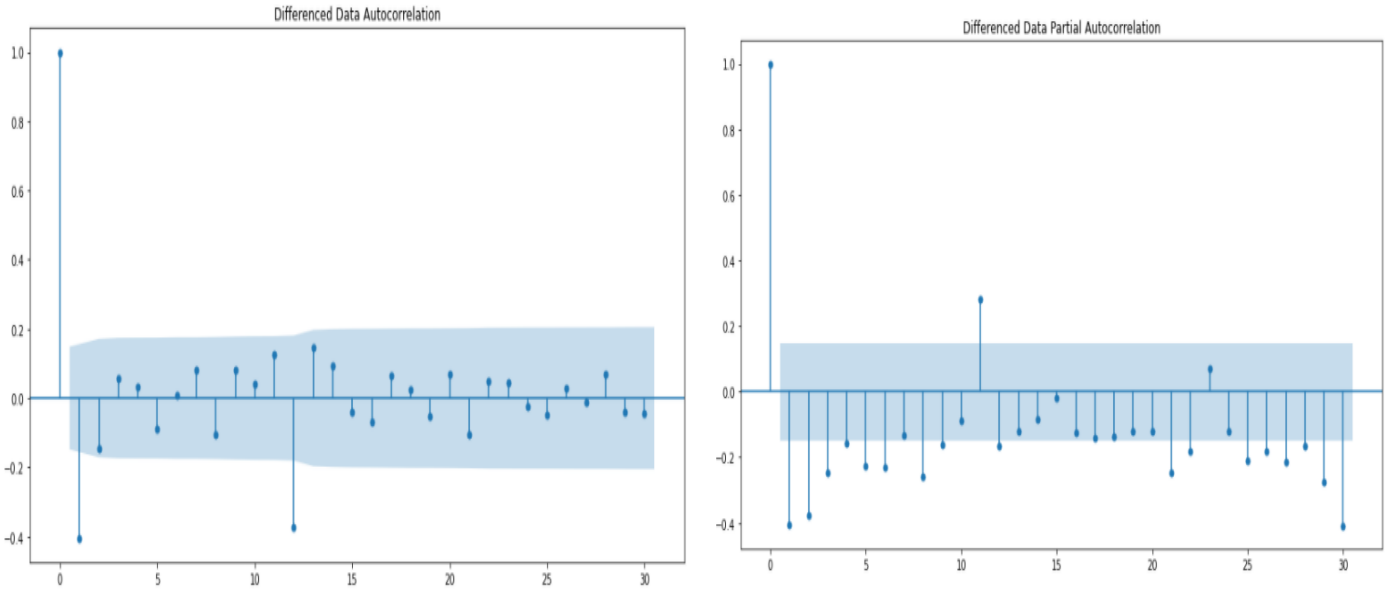


Figure-44: ACF and PACF plots

- → Here we have taken alpha = 0.05 and seasonal period as 12.
- → From the PACF plot it can be seen that till 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

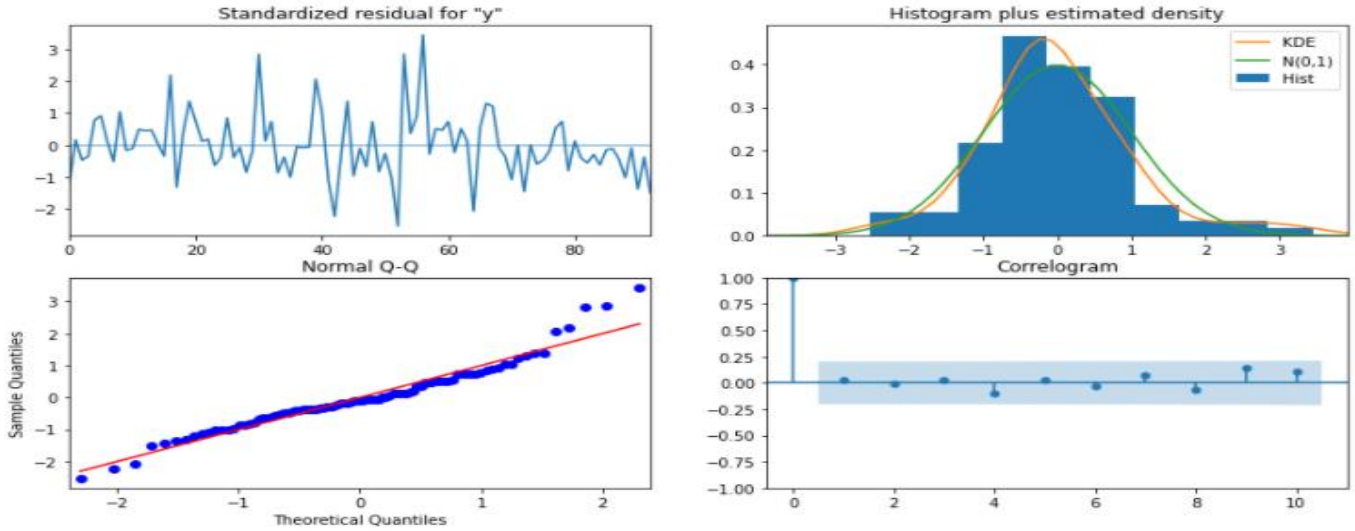


Figure45: Diagnostic-plot

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-693.697
Date:	Mon, 22 Feb 2021	AIC	1403.394
Time:	23:14:52	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

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

Figure-46: Manual SARIMA Model

YearMonth	Sparkling	spark_forecasted	spark_forecasted_log	manual_spark_forecasted
1991-01-01	1902	1460.244631	1629.416485	1579.910093
1991-02-01	2049	1392.437177	1384.547008	1419.154320
1991-03-01	1874	1743.201708	1804.206479	1868.143938
1991-04-01	1279	1650.066938	1685.514276	1731.472337
1991-05-01	1432	1522.656035	1569.597811	1659.822745

Figure-47: Manual SARIMA Forecasted Values

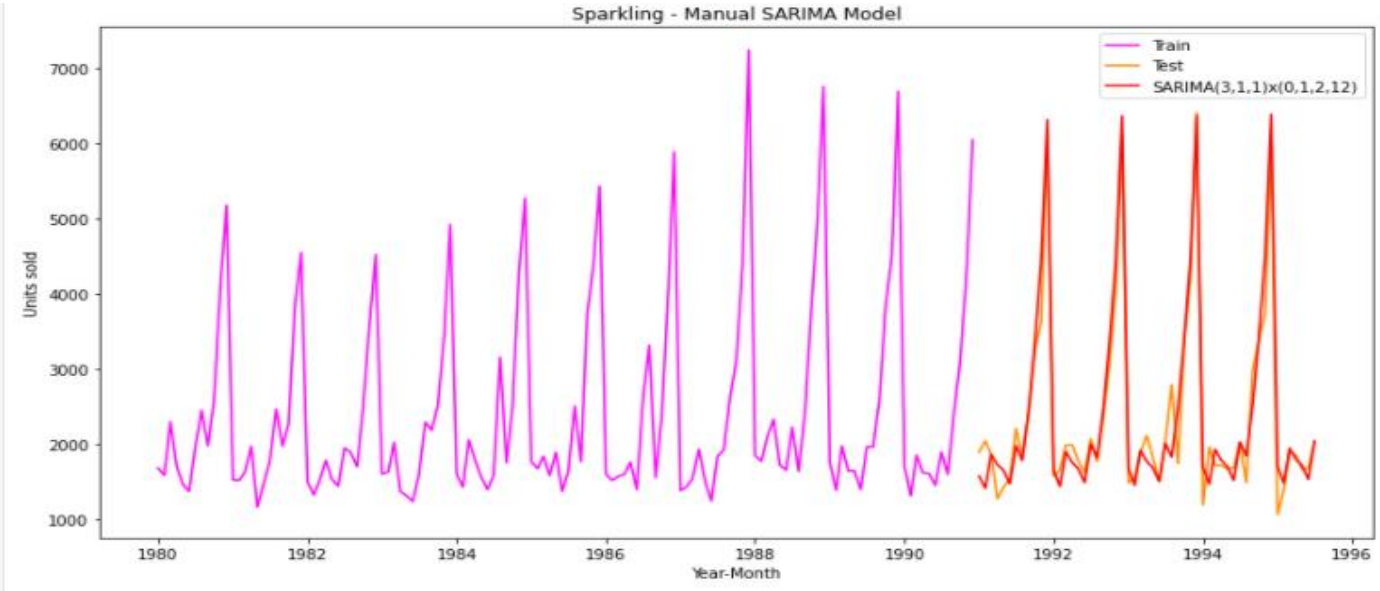


Figure-48: Plot Actual v/s Forecast Result on test data

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

Solution:

	Test RMSE
Manual_SARIMA#(3,1,1)*(1,1,2,12)	324.106824
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	336.800144
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	371.367690
Auto_SARIMA(1, 1, 2)*(0, 1, 2, 12)	382.576712
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	469.432003
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=0.0496, SES Optimized	1316.034674
9 point TMA	1346.278315
Auto_ARIMA(2,1,2)	1374.297411
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DES iterative	1778.560000
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
NaiveModel	3864.279352
Manual_ARIMA(0,1,0)	4779.154299

Figure-49: 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.

Solution:

→ 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.

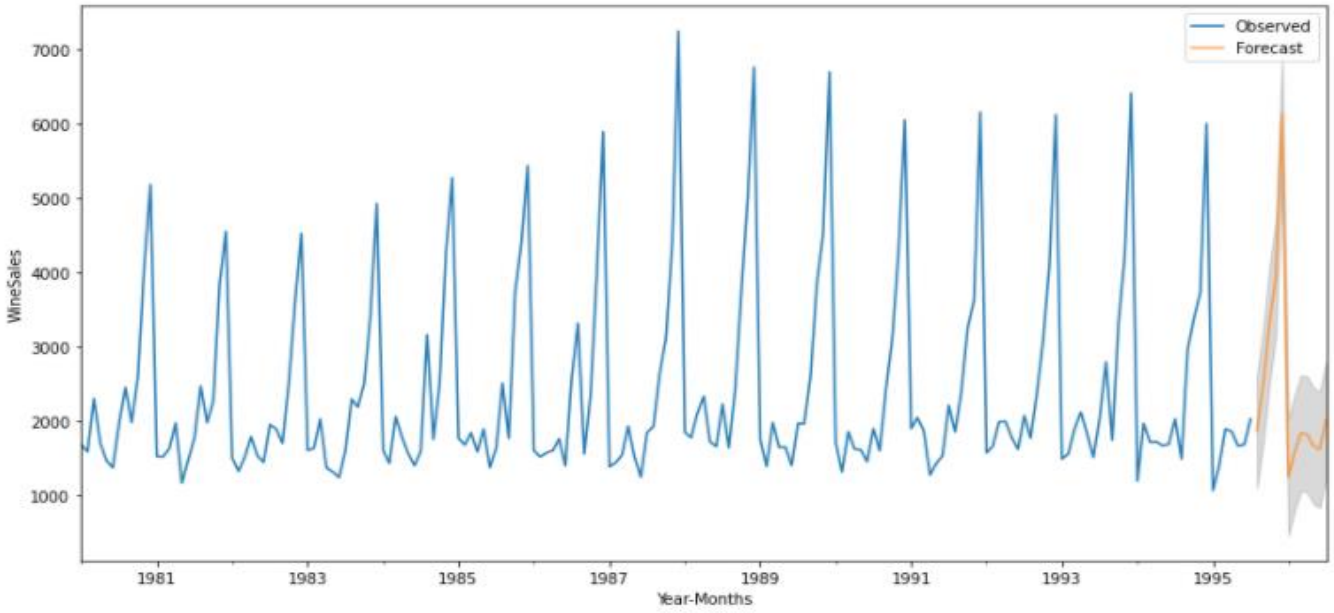


Figure-50: Plot Actual and Future Forecast Result

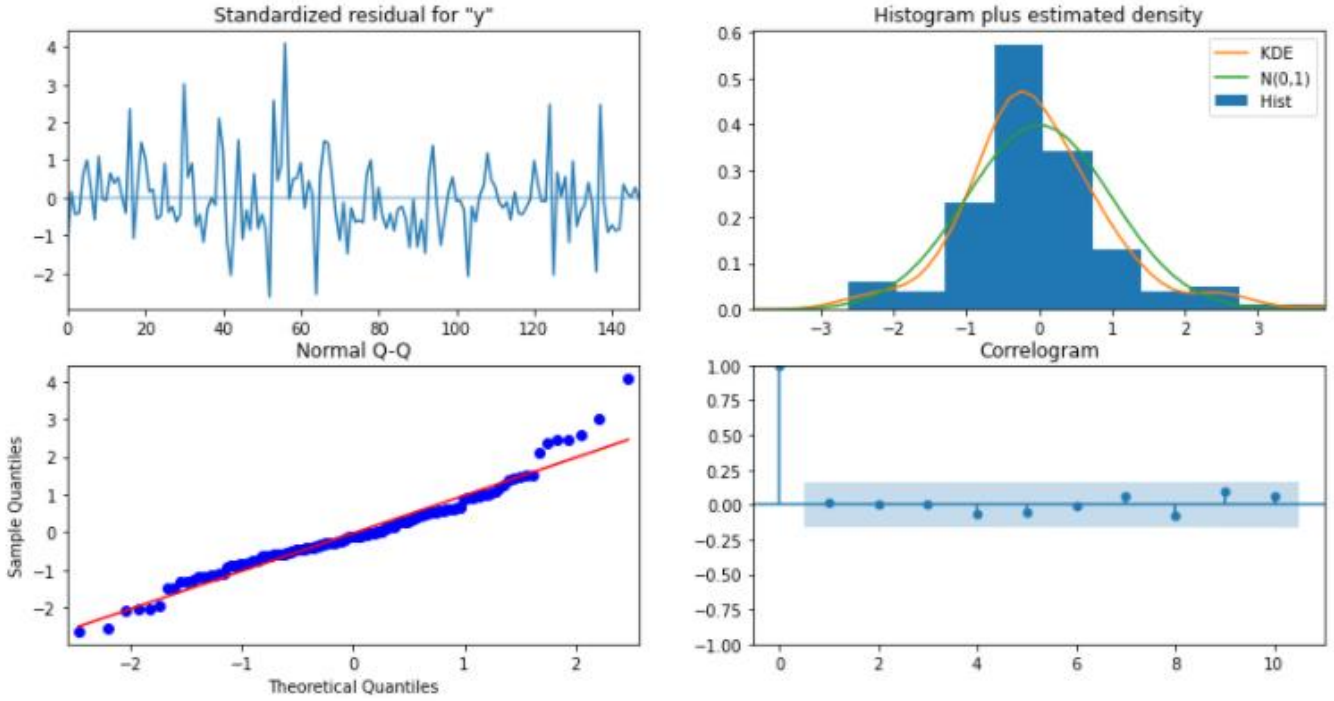


Figure-51: Diagnostic plot

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.

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

Solution:

→ 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.

```
1995-08-31    1870.888610
1995-09-30    2489.623600
1995-10-31    3299.650019
1995-11-30    3934.056643
1995-12-31    6135.396048
1996-01-31    1245.727222
1996-02-29    1584.643764
1996-03-31    1840.705268
1996-04-30    1823.847833
1996-05-31    1668.706106
1996-06-30    1620.472487
1996-07-31    2020.534859
Freq: M, Name: mean, dtype: float64
```

Figure-52: Forecasted Result

```
count      12.000000
mean     2461.187705
std      1391.118211
min      1245.727222
25%      1656.647701
50%      1855.796939
75%      2692.130205
max      6135.396048
Name: mean, dtype: float64
```

Figure-53: Summary statistics

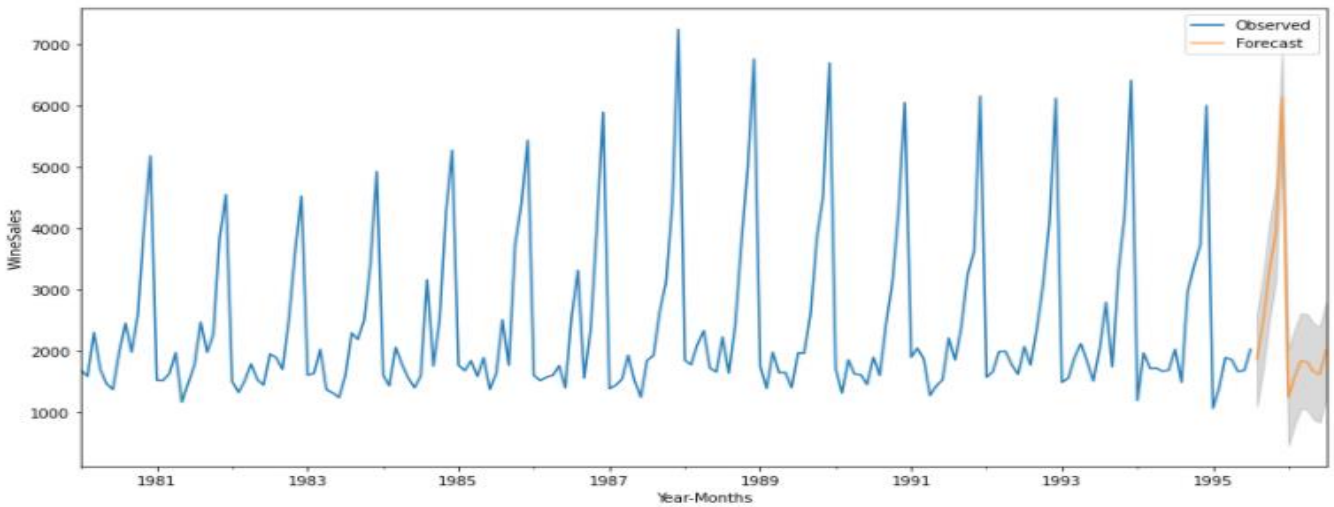


Figure-54: Plot Actual and Future Forecast Result

2 Problem Statement – TSF – Rose Dataset

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-1: View Head of the data without panda's date-time format

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:

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

```
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')
```

Figure-2: Create Date-Range

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.

	Rose
Time_Stamp	
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

Figure-3: View Head of the Time-Series data

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]

View the top 5 rows of Rose dataset :

```
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
```

View the bottom 5 rows of Rose dataset :

```
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
```

Figure-4: View Head and Tail of the Time-Series data

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].

The number of Null values in Rose dataset:

```
2
```

The datetime stamps for which the Time Series Data in Rose is not present:

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

The new interpolated values of the previously missing values:

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

Figure-5: Handling Missing Values

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

Sales Data of Rose Wines:

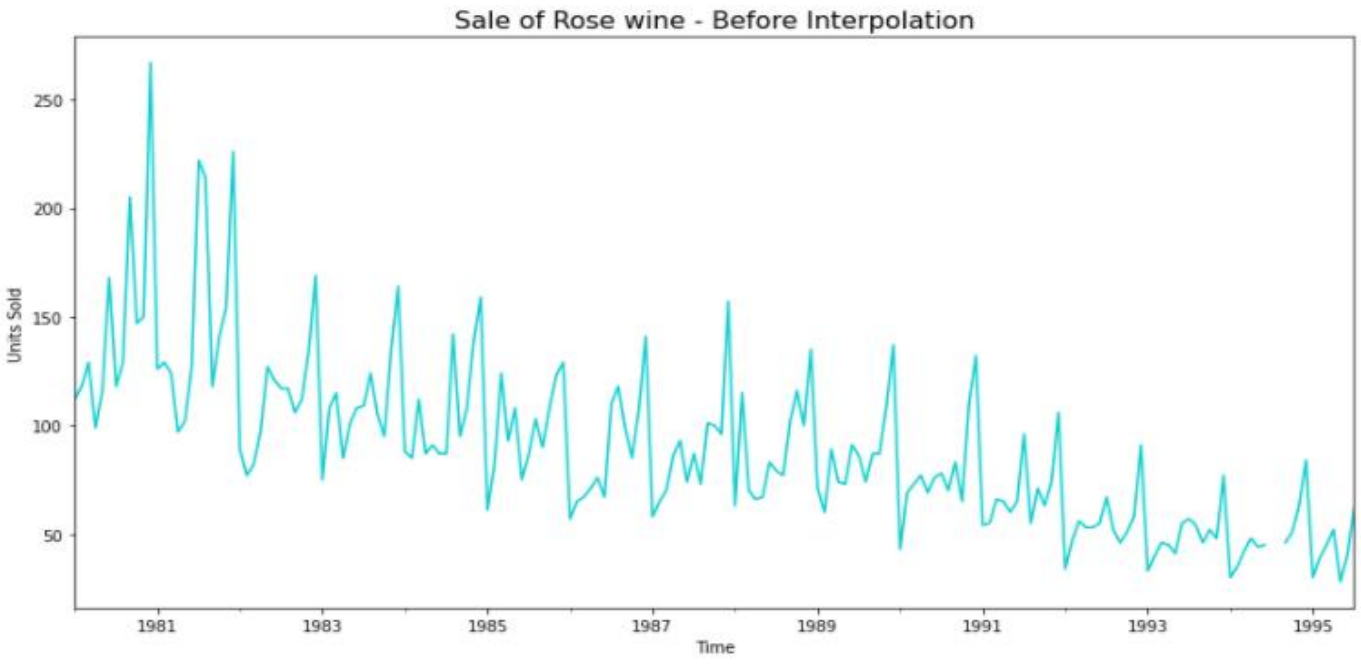


Figure-6a: Plot Rose Time Series data – Before Interpolation

Sales Data of Rose Wines after missing value treatment:

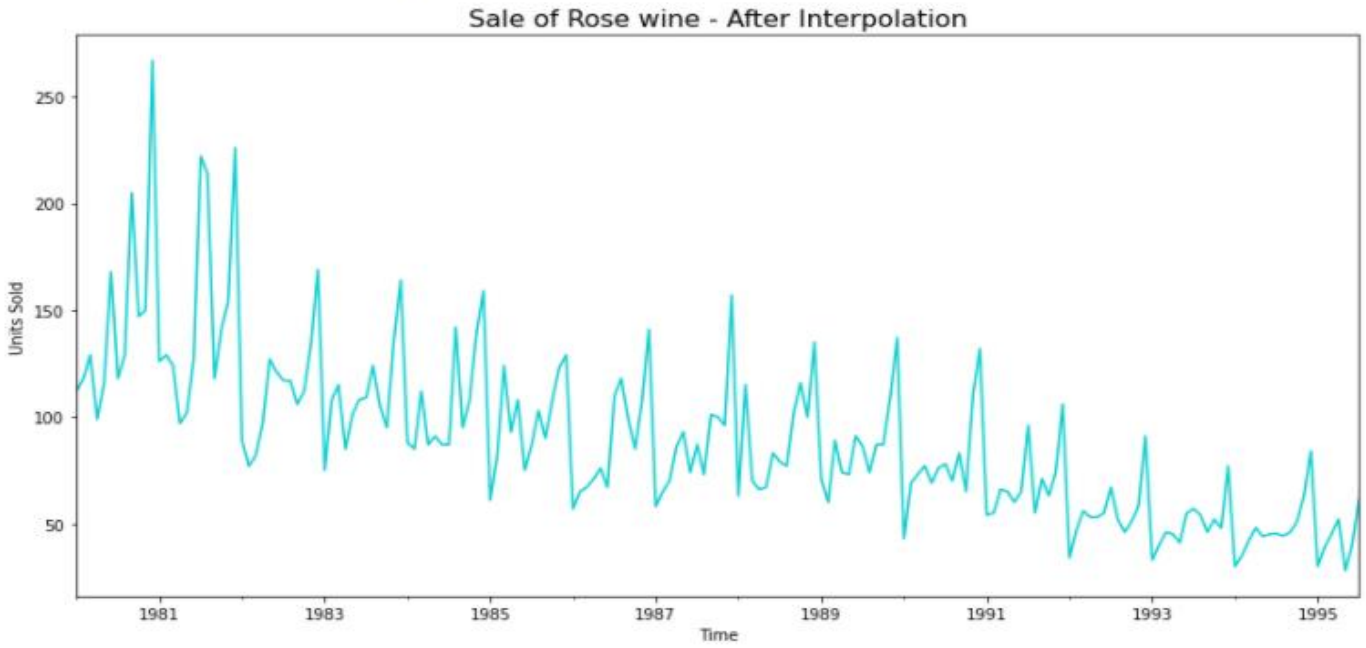


Figure-6b: Plot Rose Time Series data – After Interpolation

→ 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.

Solution:

Check the basic measures of descriptive statistics:

Data Description for Rose Dataset:

```
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
```

Figure-7: Summary 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.

Yearly Boxplot:

Yearly Boxplot for Rose Dataset:

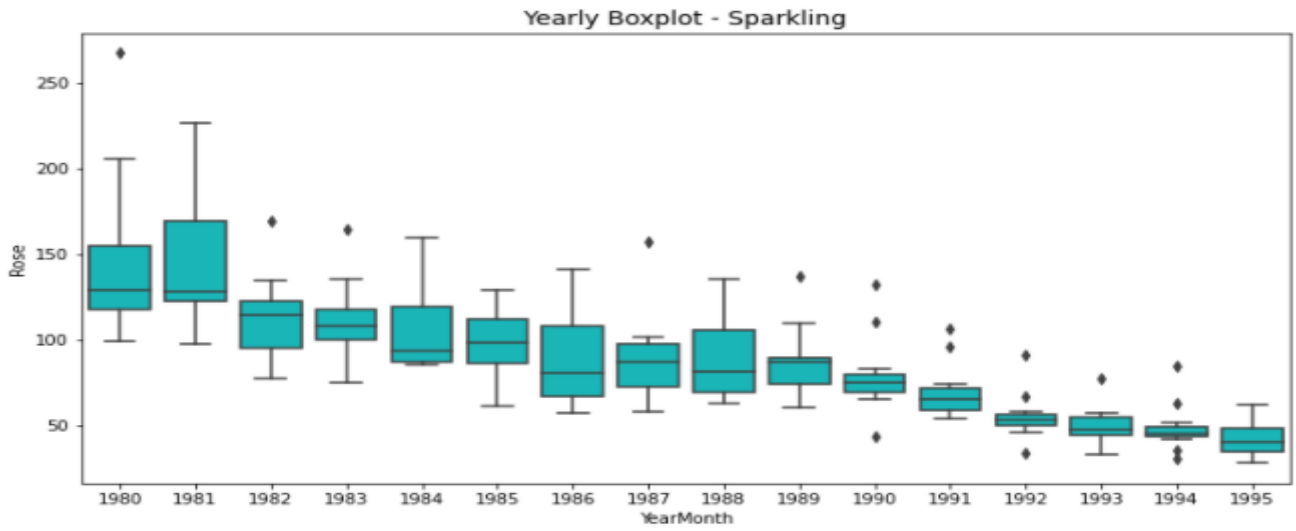


Figure-8: Yearly Boxplot

Monthly Boxplot for all the years for Rose Dataset:

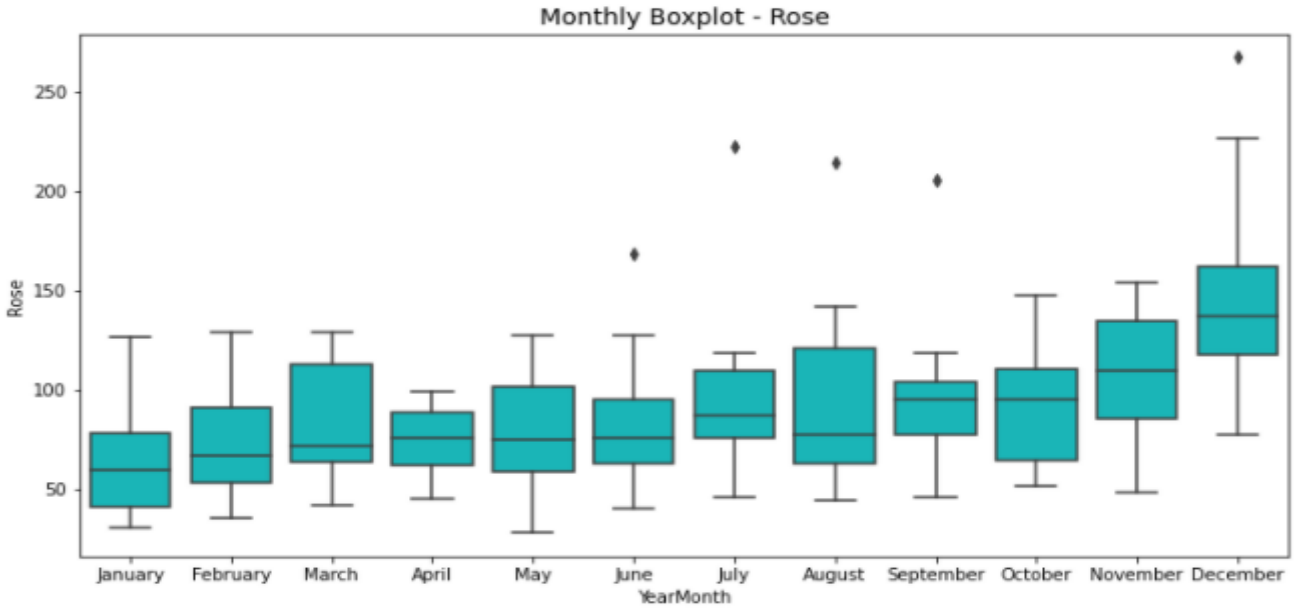


Figure-9: Monthly Boxplot

→ 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.

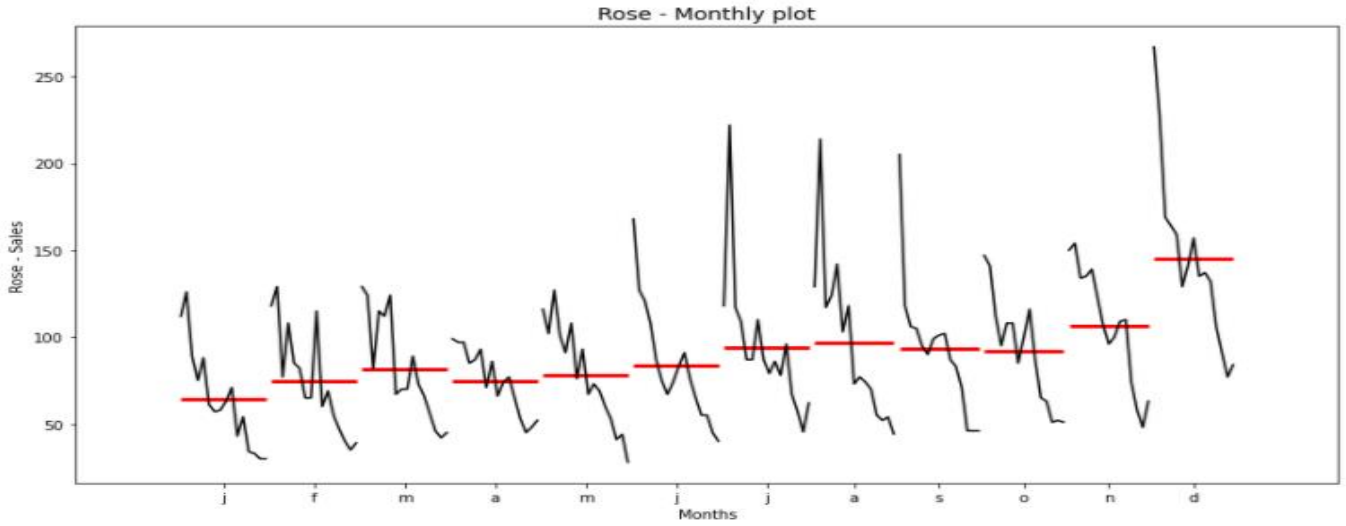


Figure-10: Monthly plot

→ 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.

Monthly Wine sales across years for Rose:

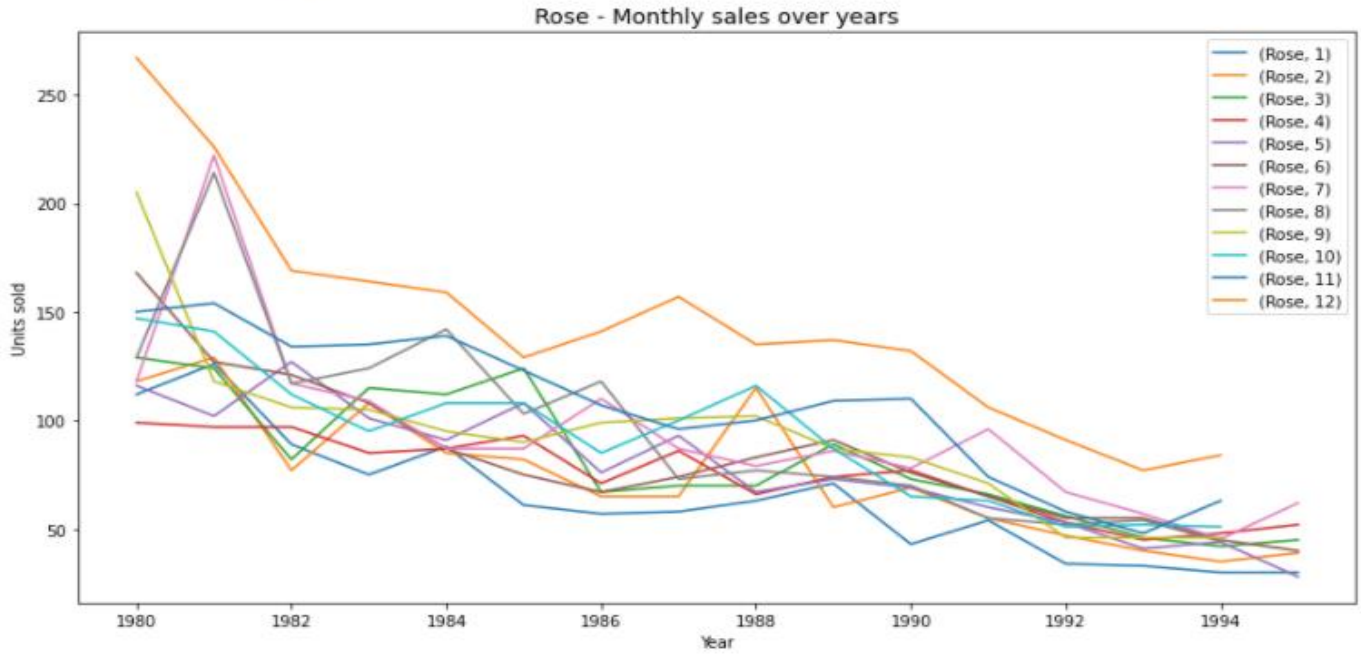


Figure-10: Monthly Sales over years

→ 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.

Decompose the Time Series and plot the different components:

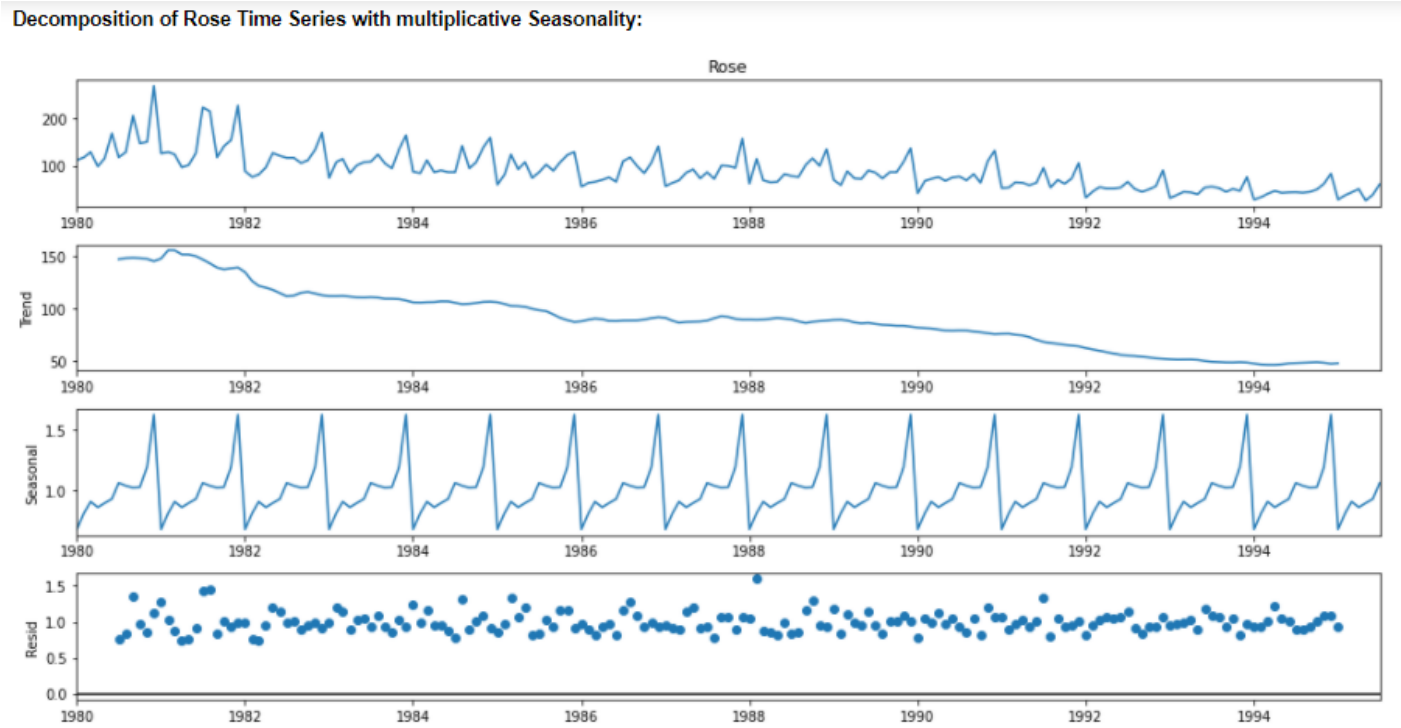


Figure-11: Multiplicative Model

- → The observed plot of the decomposition diagram shows visible annual seasonality and a downward trend. The early period of the plot shows higher variation than in the later periods
- → The trend diagram shows a downward trend overall. Exponential dips can be seen between 1981 and 1983 and later from 1991 to 1993
- → Seasonal components are quite visible and consistent in both the observed and seasonal charts of the diagrams. The multiplicative model shows variance in seasonality of 16%
- → The residuals shows a pattern of high variability across the period of time-series, which is more or less consistent.
- → The variance in residuals shows higher variance in the early period of the series, which explains the higher variance in observed plot at same time period.
- → As the seasonality peaks are consistently reducing its altitude in consistent with trend, the series can be treated as multiplicative in model building

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

Solution:

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

```
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-12: Train and Test Data

The Plot Rose Time Series as train and test

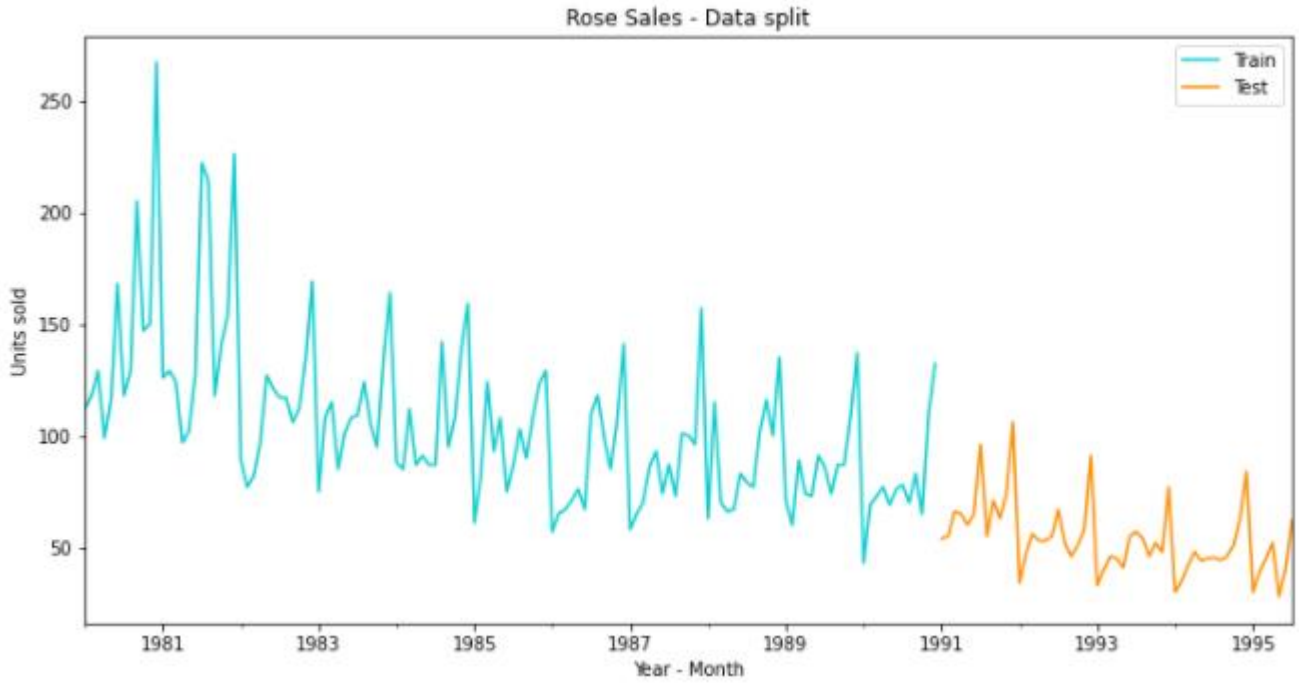


Figure-13: The Plot Rose Time Series as train and test

2.4 Build various 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Note: Please do try to build as many models as possible and as many iterations of models as possible with different parameters.

Solution:

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

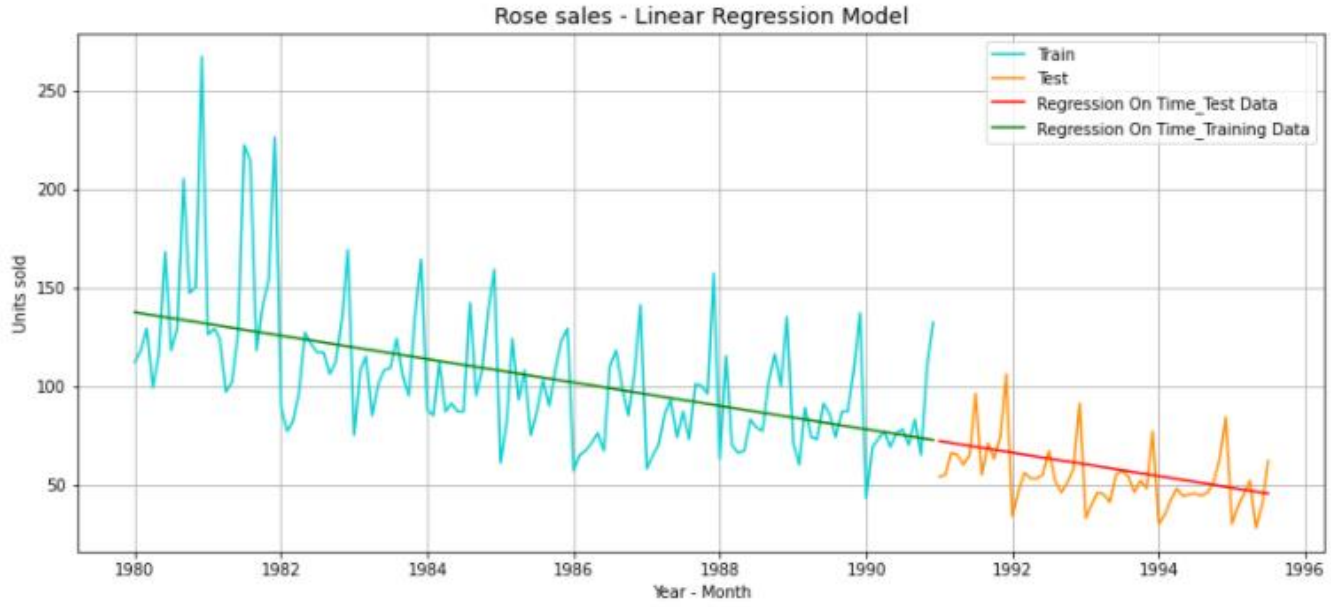


Figure-14: Linear Regression Model

- → 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 : 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 79.75.
- → The model do not capture the trend or seasonality for the given dataset.

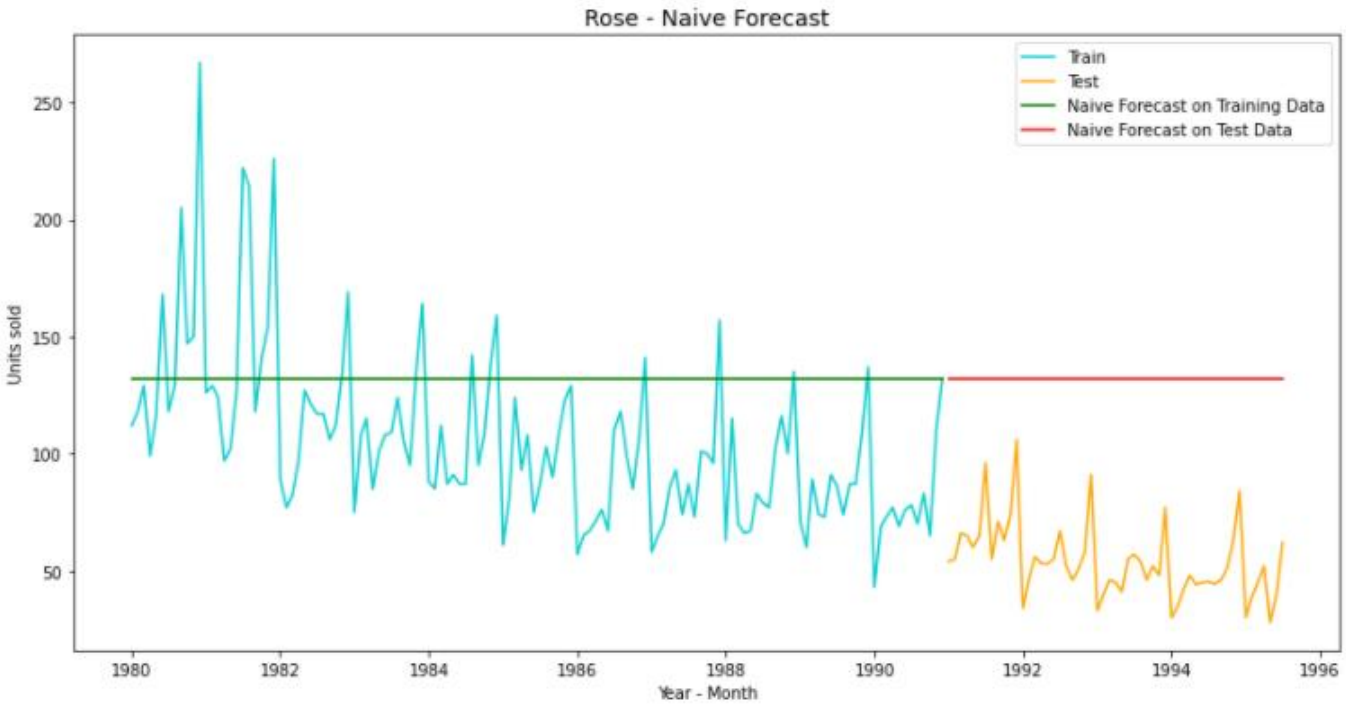


Figure-15: Naïve Forecast Model

Model 3: Simple Average Model

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

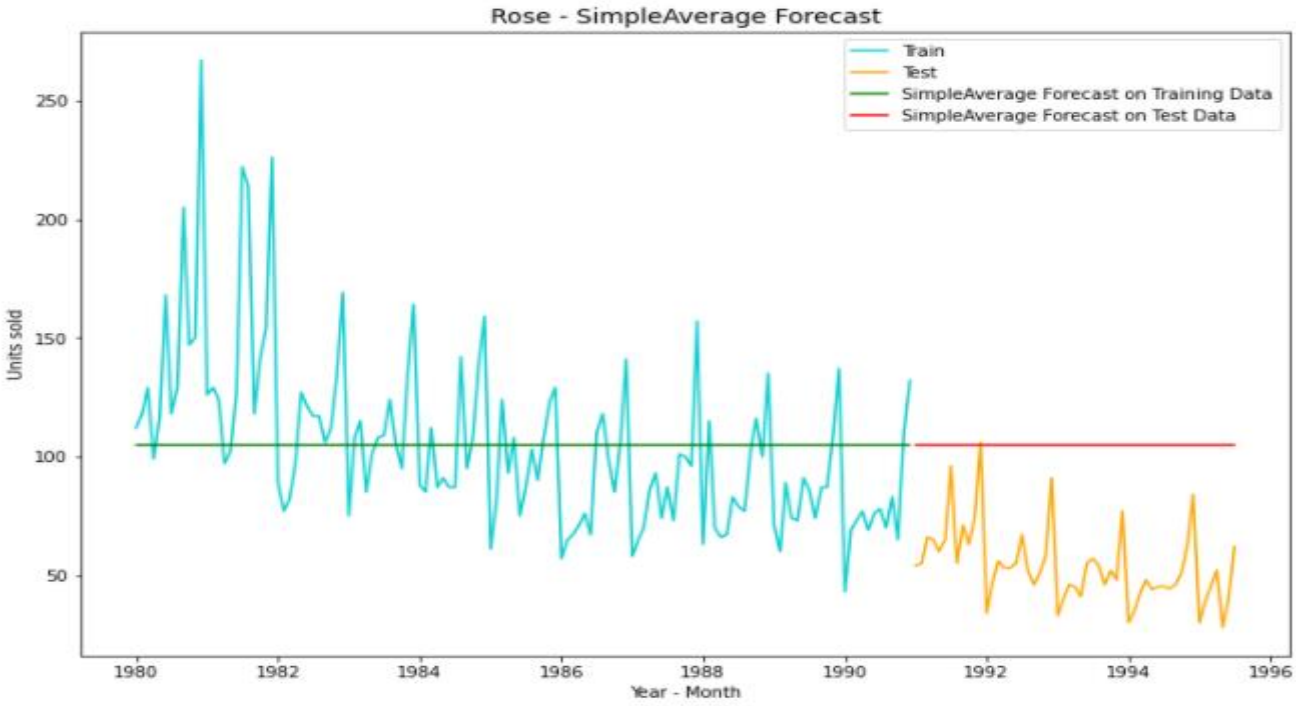


Figure-16: Simple Average Model

→ 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 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 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.

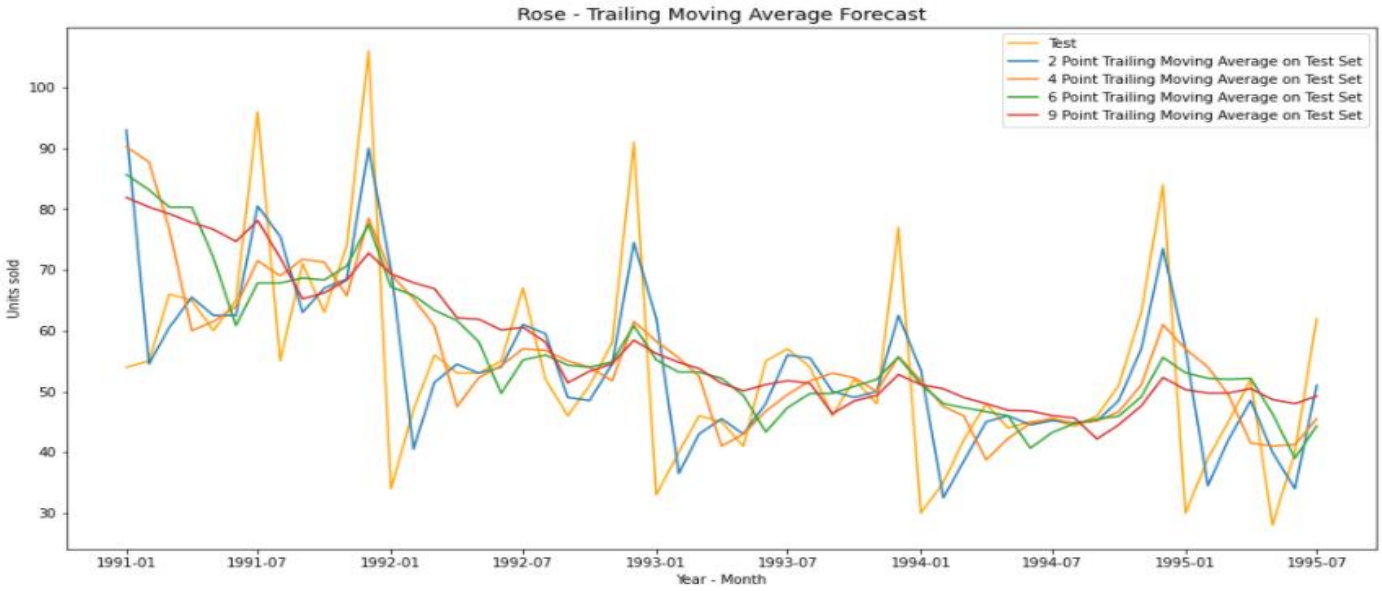
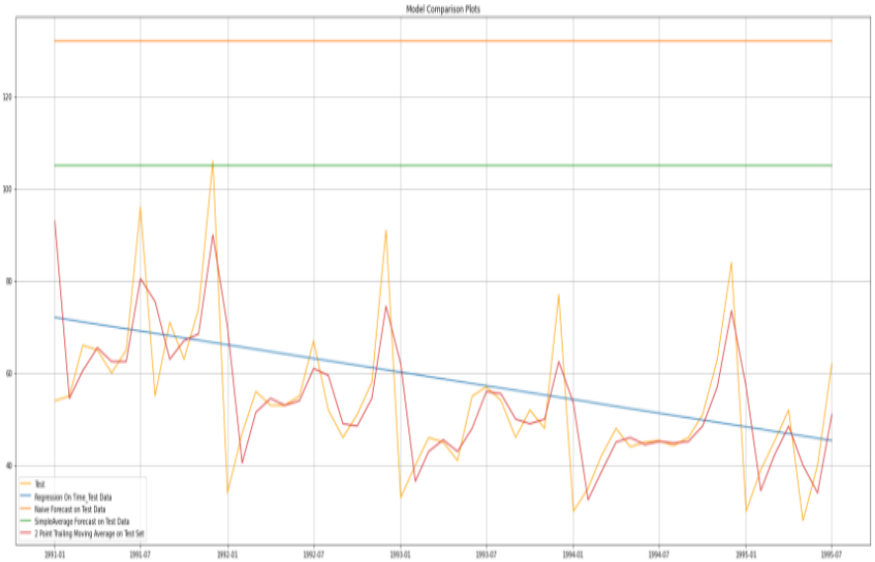


Figure-17: Moving Average Model

Model Comparison and RMSE Values



	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

Figure-18: Model comparison and RMSE on Test data

Model 5: Simple Exponential Smoothing Model

- → 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.

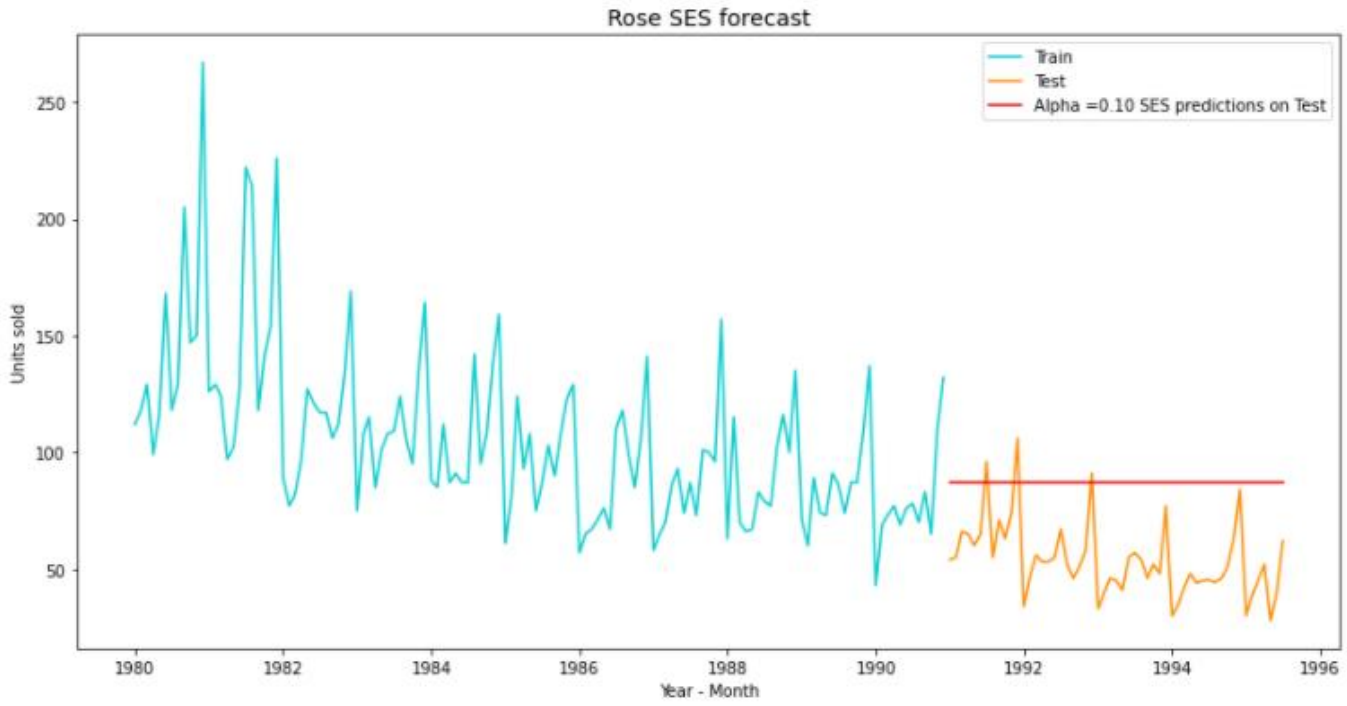


Figure-19: SES Iterative Model

Model 6: Double Exponential Smoothing 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.

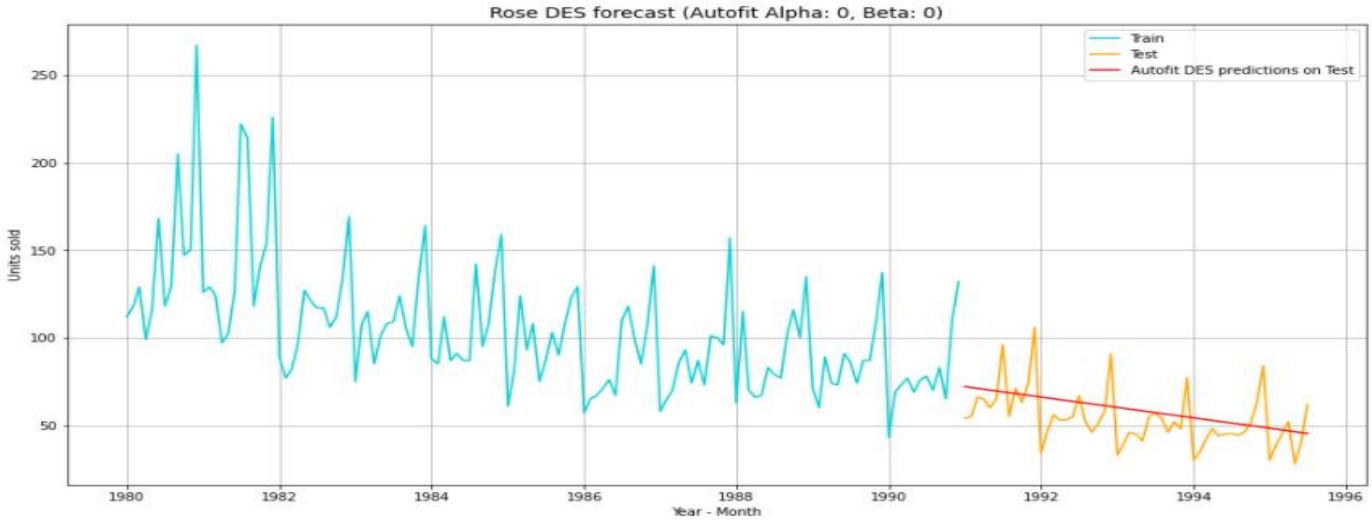


Figure-20: DES Optimised Model

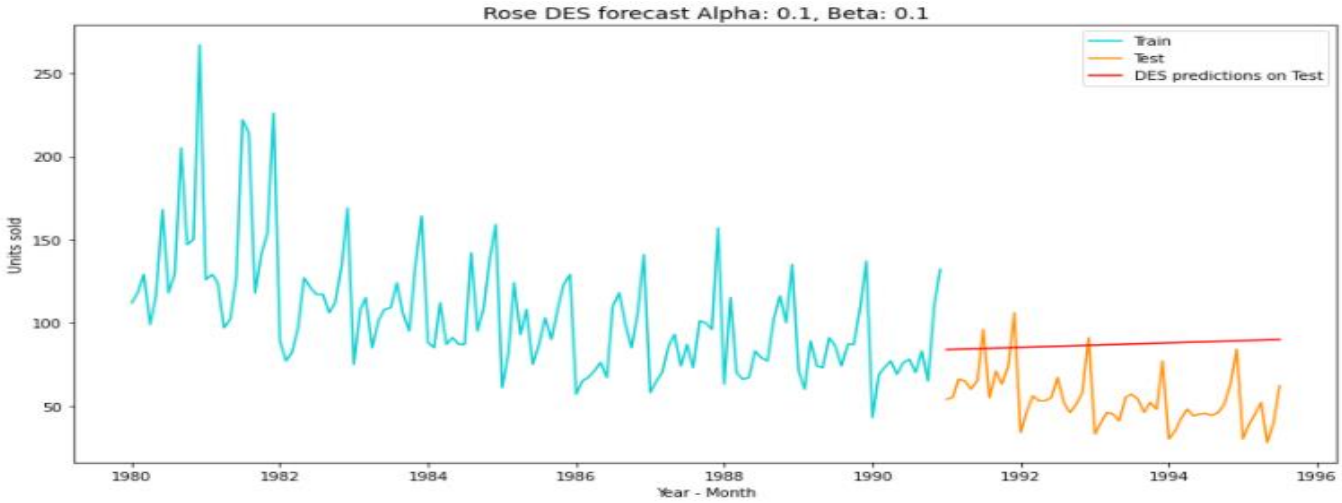


Figure-21: DES Iterative Model

Model 7: Triple Exponential Smoothing 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.

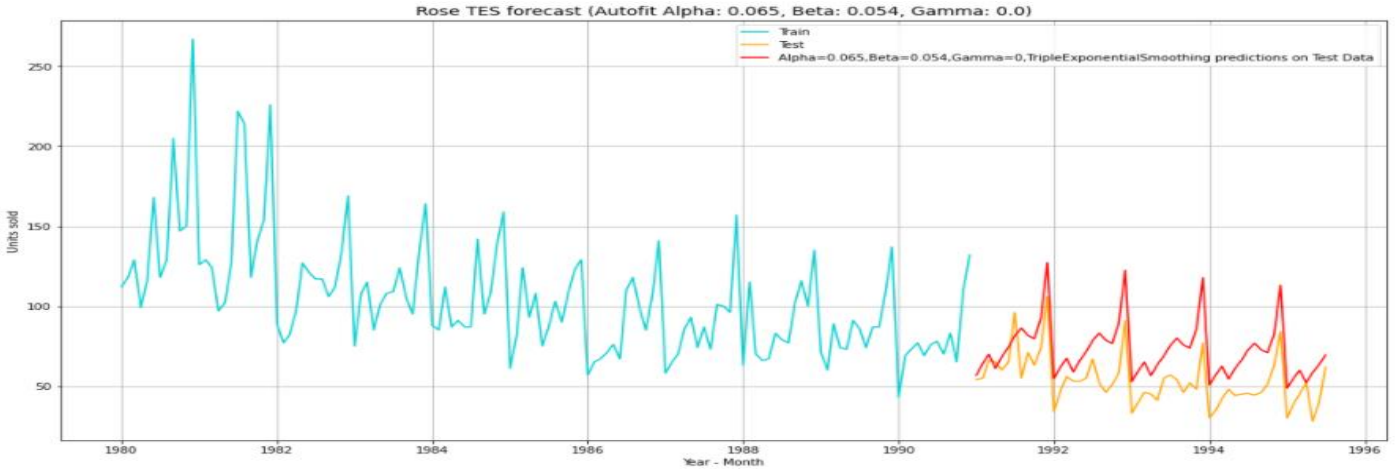


Figure-22: TES Optimised Model

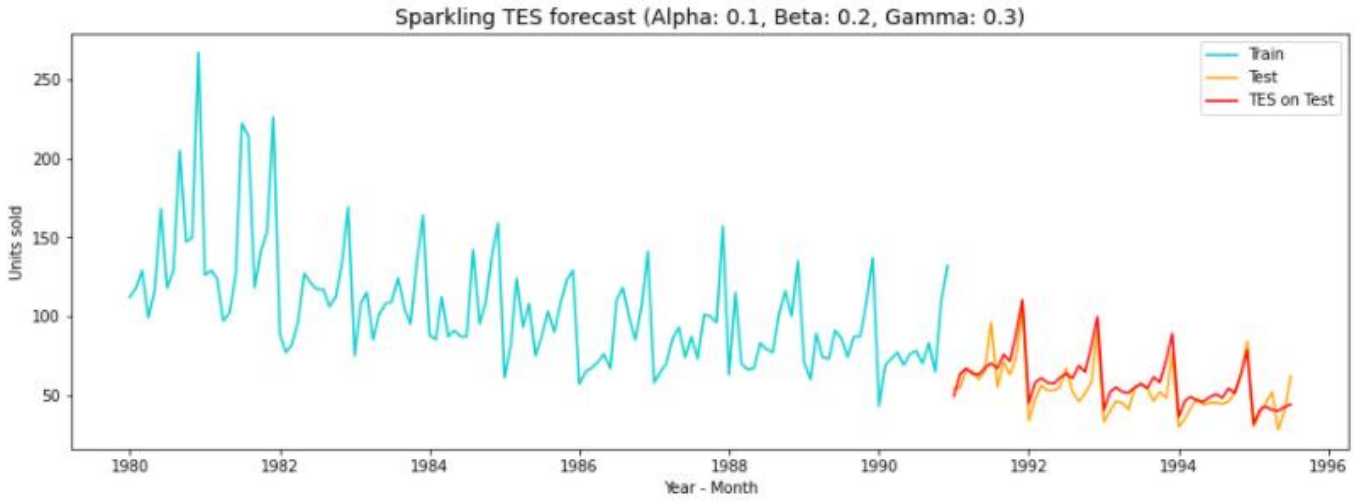


Figure-23: TES Iterative Model

Model Comparison:

	Test RMSE
Alpha=0.1,Beta=0.2,gamma=0.3,TES	9.943563
2 point TMA	11.530054
4 point TMA	14.458402
6 point TMA	14.572976
9 point TMA	14.732918
Alpha=0.0,Beta=0.0, DES Optimized	15.278359
RegressionOnTime	15.278369
Alpha=0.065,Beta=0.054,gamma=0.0 TES Optimized	21.197161
Alpha=0.0987, SES Optimized	36.824478
Alpha=0.10,SES	36.856268
Alpha=0.1,Beta=0.1,DES	36.950000
SimpleAverage	53.488233
NaiveModel	79.745697

Figure-24: TEST RMSE Values

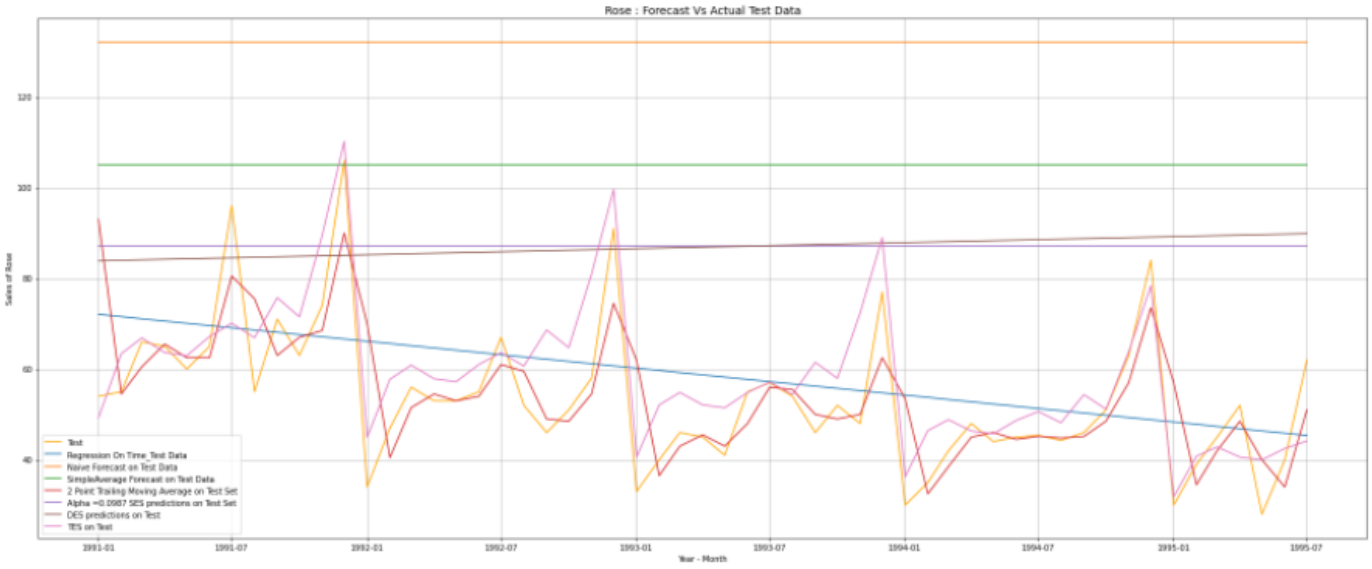


Figure-25: Sparkling Forecast v/s Actual Values

→ 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
```

ADF on original series
- P-Value > alpha .05
- Test statistic > Critical values
- Fail to reject the null hypothesis
- The series is non-stationary

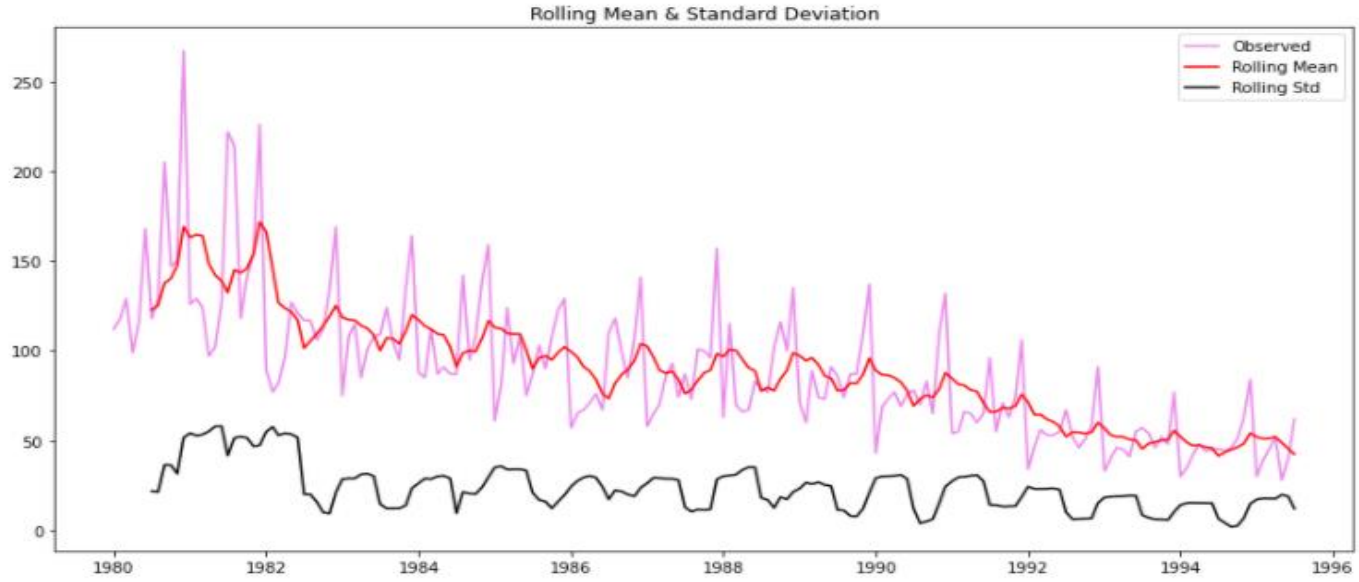


Figure-24: ADF test on Original Series

- → 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
```

ADF on differenced series

- P-Value < alpha .05
- Test statistic < Critical values
- Reject the null hypothesis
- The series is stationary

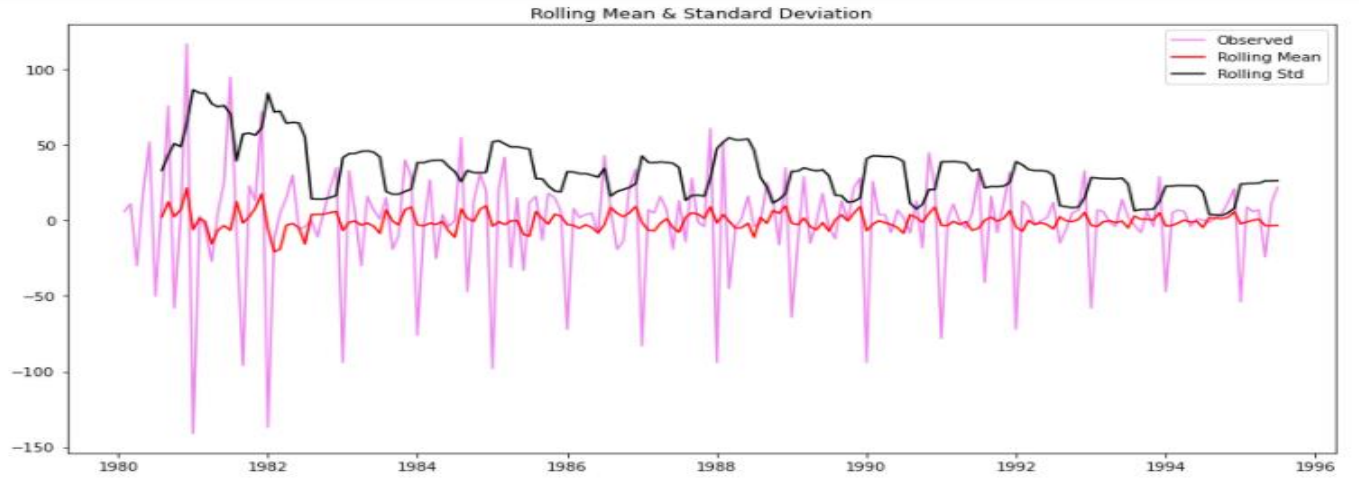
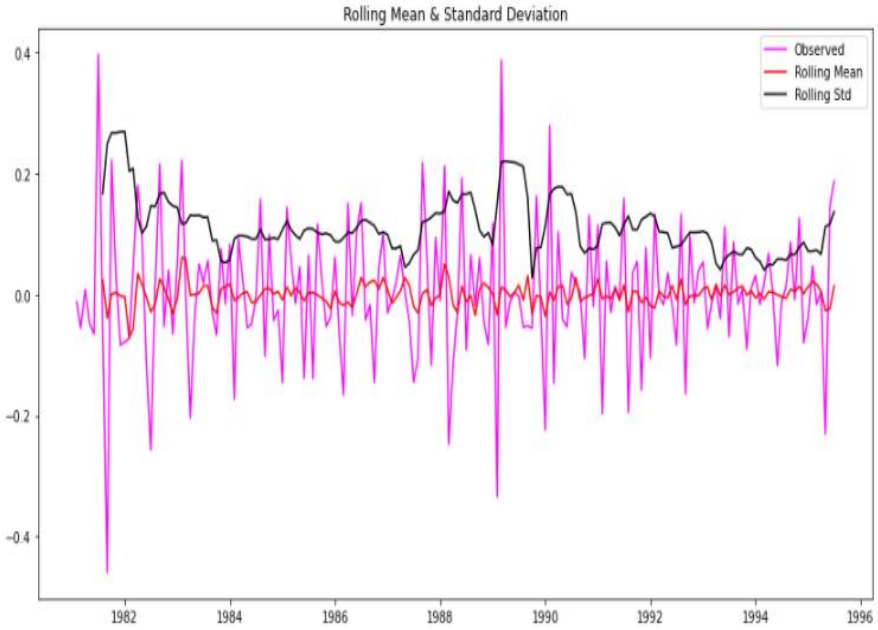
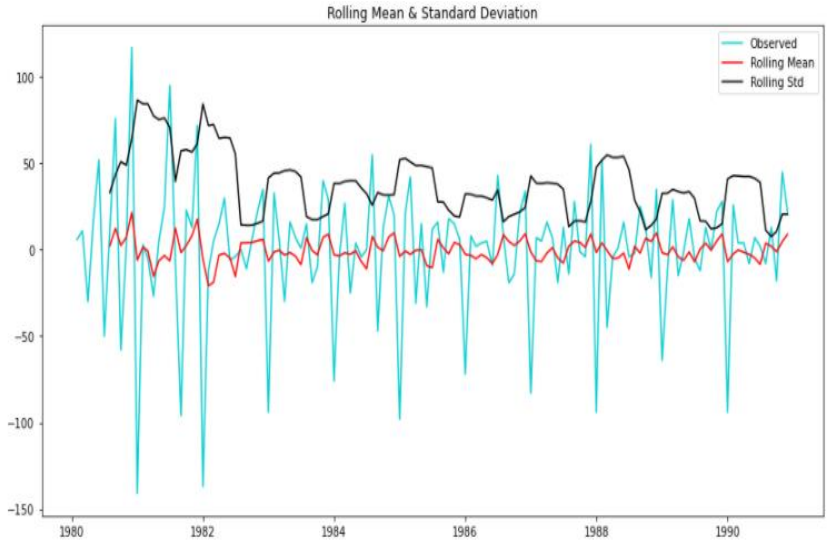


Figure-25: ADF test after differencing d=1




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.628179
p-value                          0.000115
#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-26: 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-27: ADF test on train data after differencing 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.

Solution:

Model 8: Auto-ARIMA Model

→ 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 15.63.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418
Method:	css-mle	S.D. of innovations	30.168
Date:	Mon, 22 Feb 2021	AIC	1276.835
Time:	19:09:53	BIC	1288.336
Sample:	02-01-1980 - 12-01-1990	HQIC	1281.509

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4885	0.085	-5.742	0.000	-0.655	-0.322
ma.L1.D.Rose	-0.7600	0.101	-7.500	0.000	-0.959	-0.561
ma.L2.D.Rose	-0.2398	0.095	-2.518	0.012	-0.427	-0.053

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	-4.1694	+0.0000j	4.1694	0.5000

Figure-28: Auto ARIMA Model

Model 9a: Auto-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 774.97 was obtained at (0, 1, 2)x(2, 1, 2, 12).

→ The model was built with the above parameters.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-380.485
Date:	Mon, 22 Feb 2021	AIC	774.969
Time:	19:12:00	BIC	792.622
Sample:	0 - 132	HQIC	782.094
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9523	0.184	-5.165	0.000	-1.314	-0.591
ma.L2	-0.0764	0.126	-0.606	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	187.8695	45.281	4.149	0.000	99.121	276.619

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

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

Figure-29: SARIMA Model Result

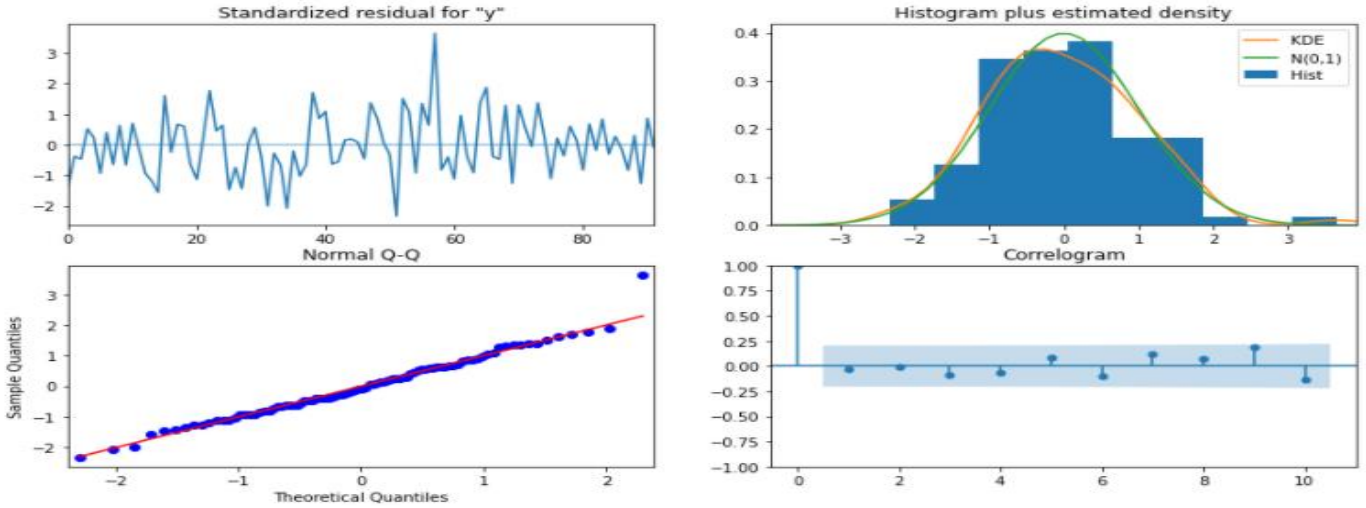


Figure-30: Diagnostic-plot

- → 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	44.213640
1991-02-01	55.0	62.326561
1991-03-01	66.0	67.313577
1991-04-01	65.0	63.160986
1991-05-01	60.0	66.474665

Figure-31: Forecasted Result on test data

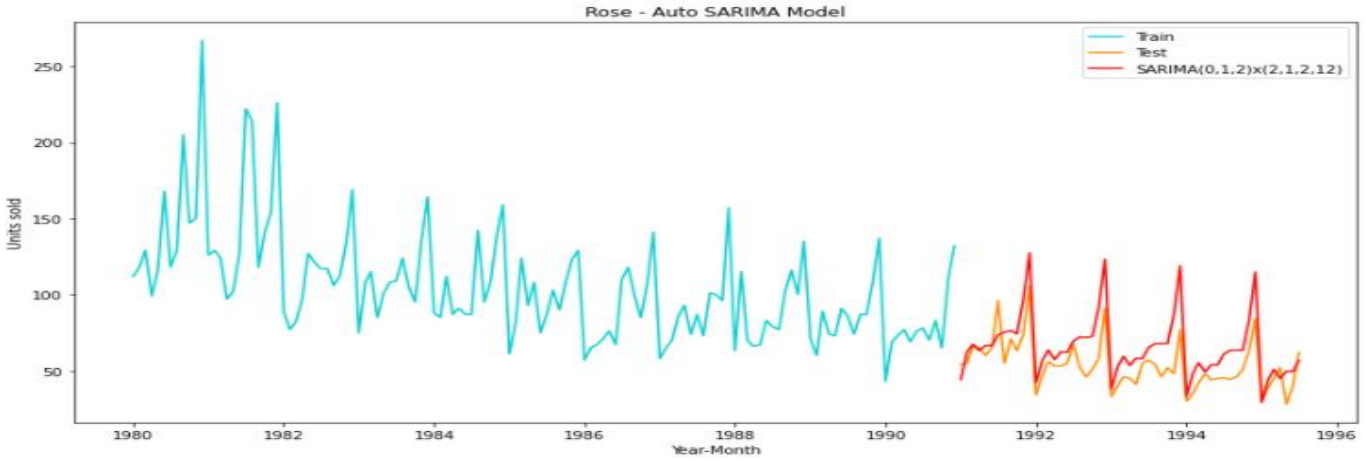


Figure-32: Plot of Actual v/s Forecasted Result on test data

Model 9b: Auto-SARIMA Model on log series data

→ 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:	Mon, 22 Feb 2021	AIC	-247.076
Time:	19:15:52	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	-18.389	0.000	-1.179	-0.952
ar.S.L12	0.9555	0.028	33.770	0.000	0.900	1.011
ma.S.L12	-0.8303	0.151	-5.499	0.000	-1.126	-0.534
sigma2	0.0051	0.001	5.147	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-33: Log Series SARIMA Model Result

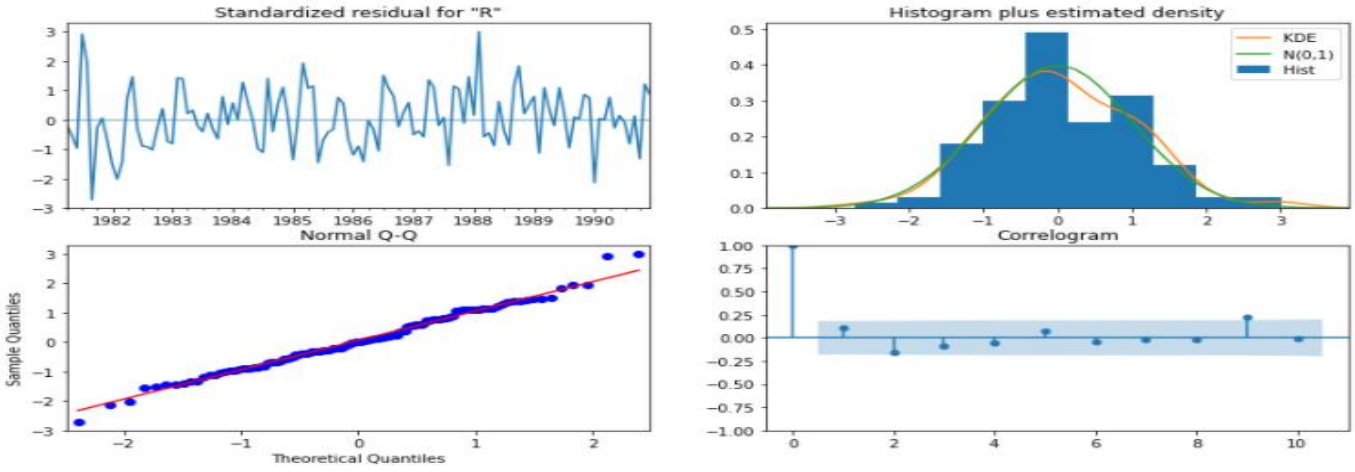


Figure-34: Diagnostic-plot

→ 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 17.93

YearMonth	Rose	rose_forecasted	rose_forecasted_log
1991-01-01	54.0	44.213640	56.850665
1991-02-01	55.0	62.326561	66.338168
1991-03-01	66.0	67.313577	70.531810
1991-04-01	65.0	63.160986	66.372950
1991-05-01	60.0	66.474665	69.031777

Figure-35: Forecasted Result on test data

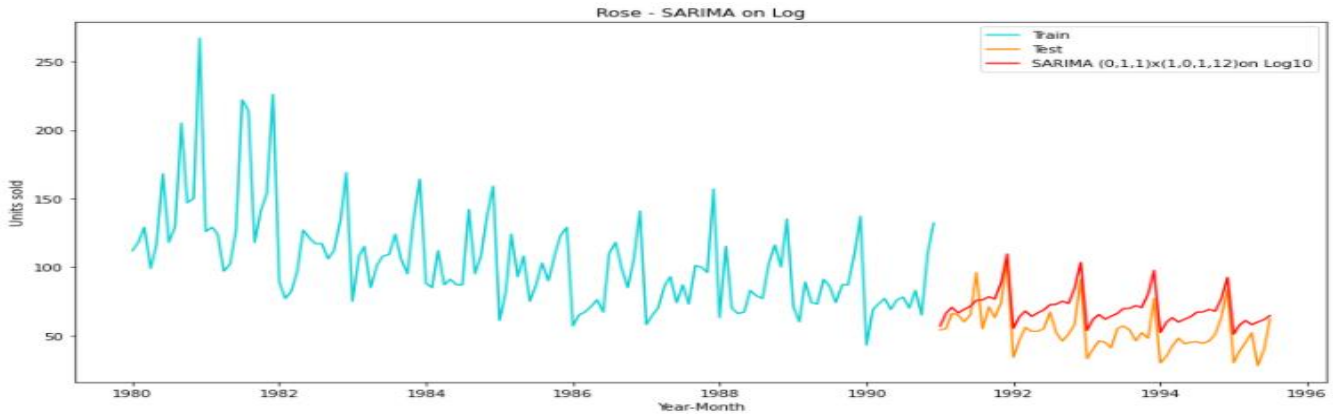


Figure-36: Plot of Actual v/s Forecasted Result on test data

→ 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.

Solution:

Model 10: Manual ARIMA

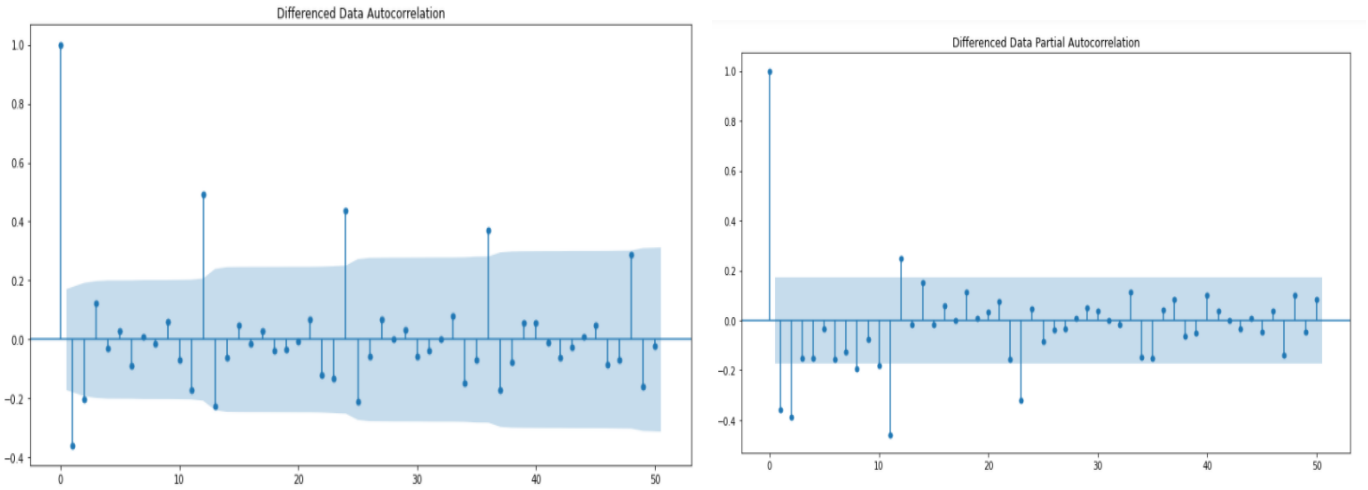


Figure-37: ACF and PACF Plots

- → 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.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 0)	Log Likelihood	-665.576
Method:	css	S.D. of innovations	38.931
Date:	Tue, 23 Feb 2021	AIC	1335.153
Time:	06:24:48	BIC	1340.903
Sample:	02-01-1980 - 12-01-1990	HQIC	1337.489

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1527	3.401	0.045	0.964	-6.514	6.819

Figure-38: Manual ARIMA Summary Results

- → The RMSE value of manual ARIMA model is 84.16. Since the ARIMA model do not capture the seasonality, this model do not perform well.

Model 11: Manual SARIMA

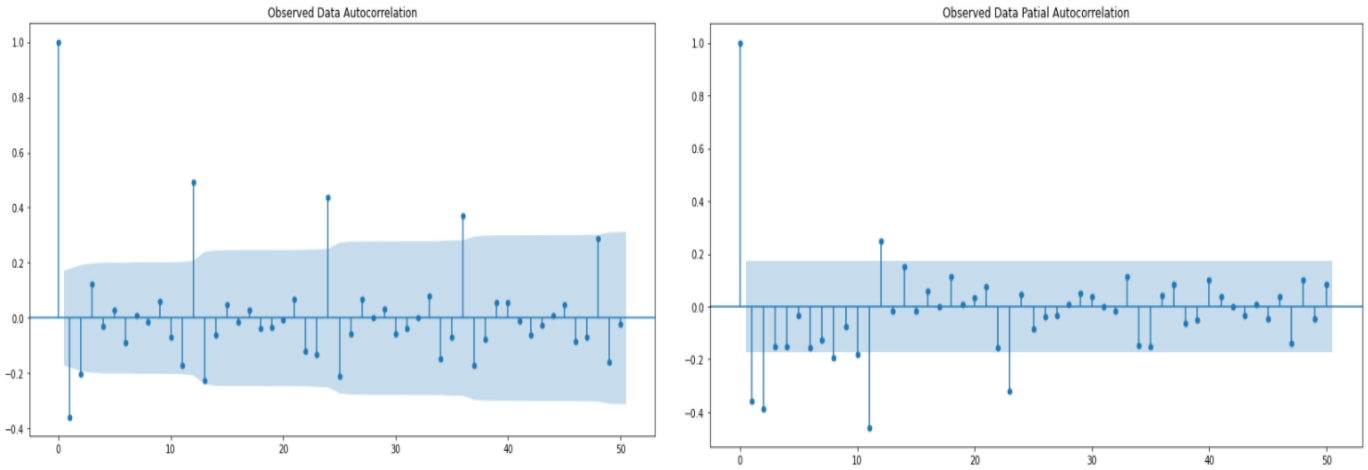


Figure-39: ACF and PACF Plots

- → 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

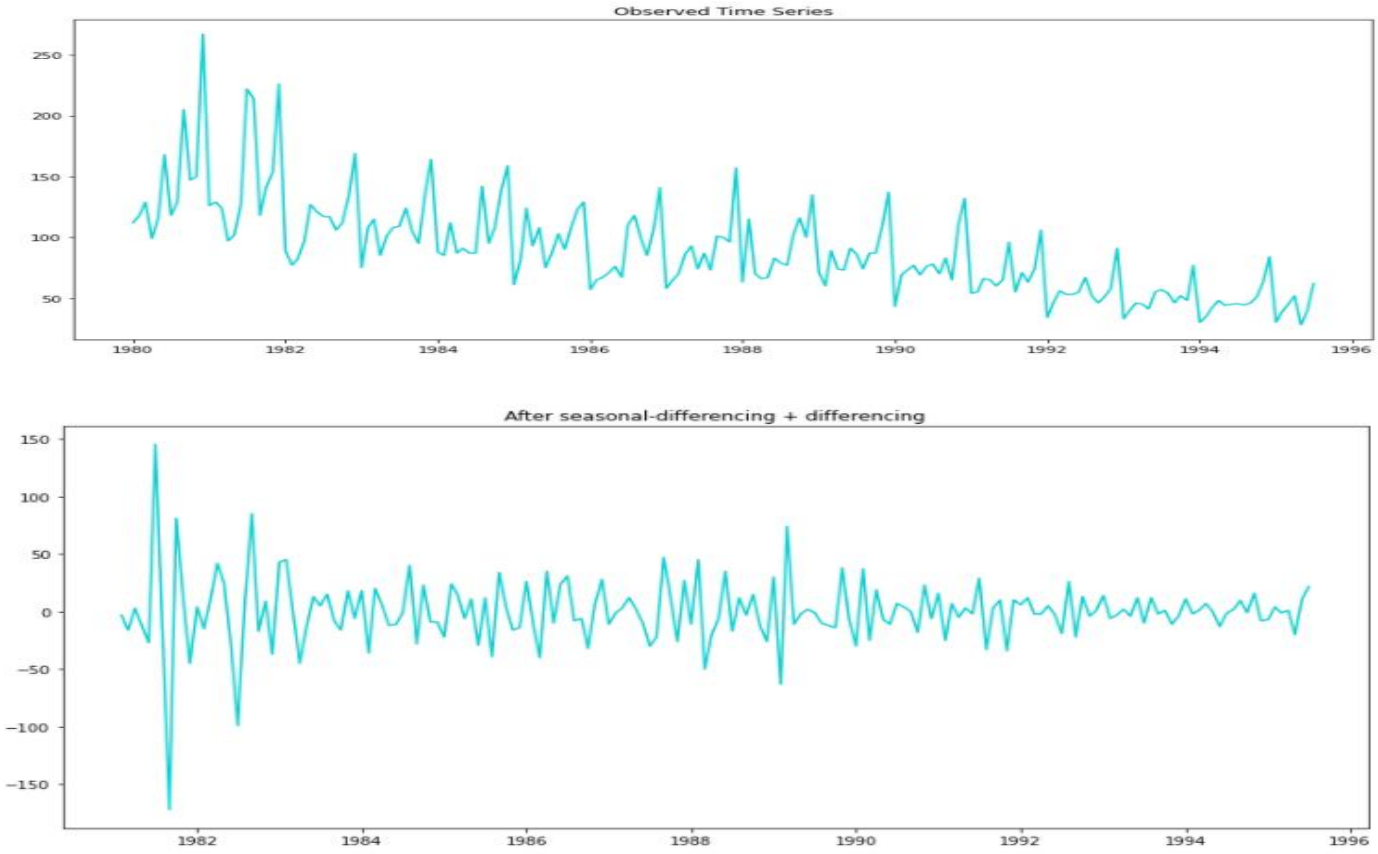
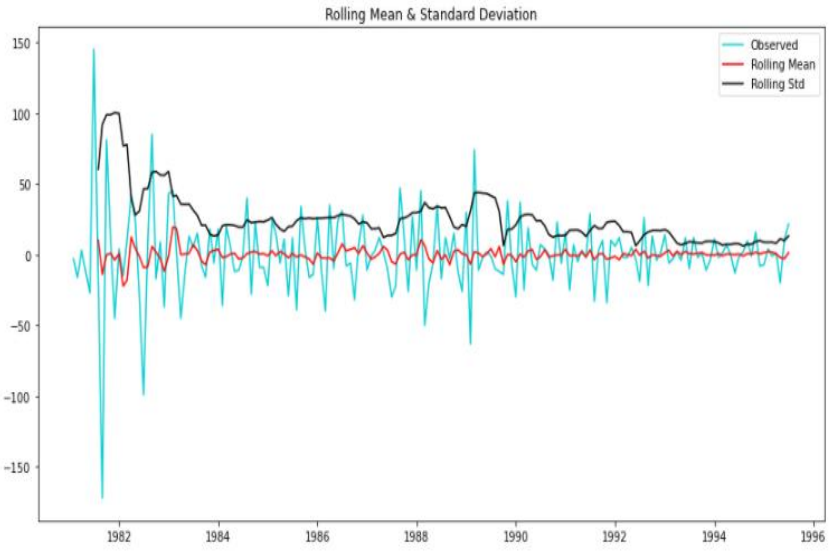


Figure-40: Time series plots

→ 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-41: 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).

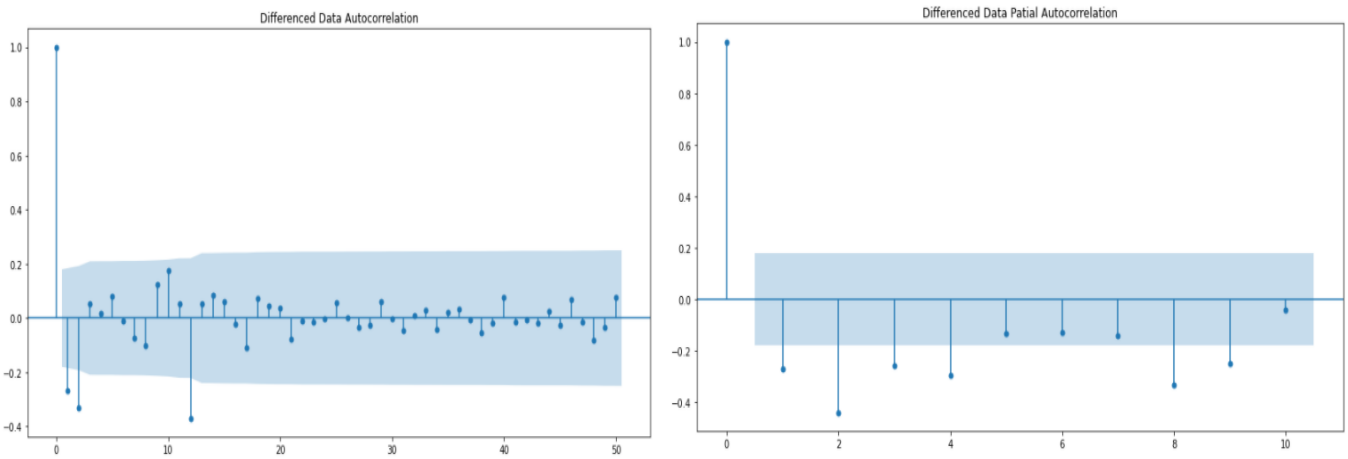


Figure-42: ACF and PACF plots

→ 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.

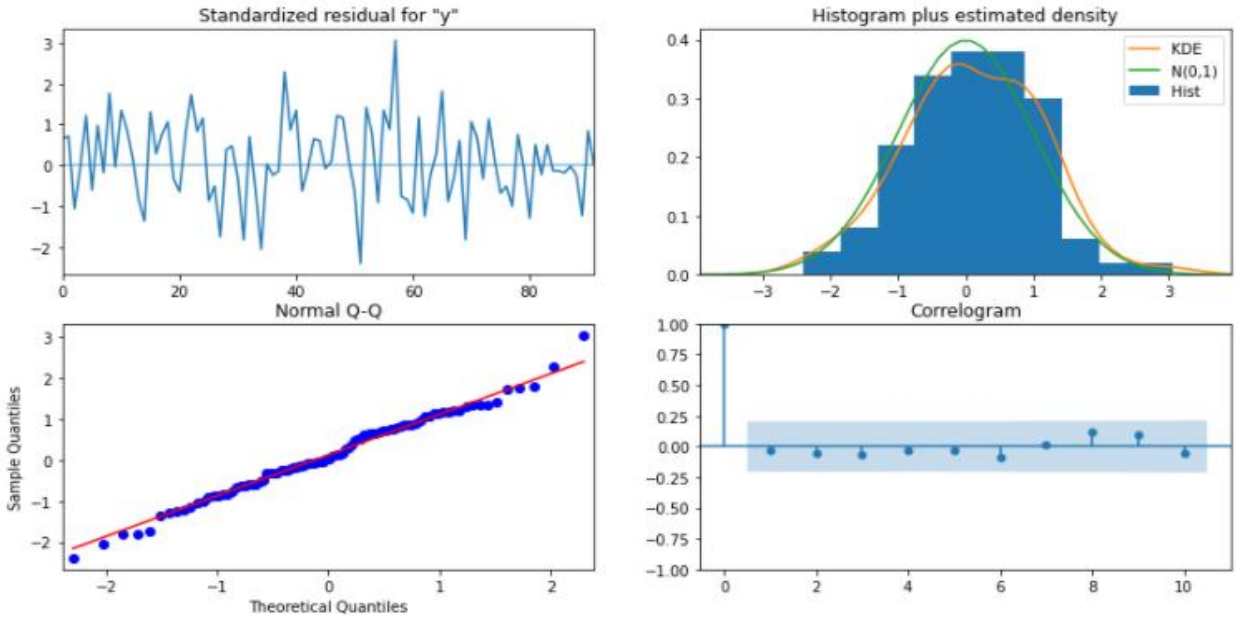


Figure43: Diagnostic-plot

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(4, 1, 2)x(0, 1, 2, 12)   Log Likelihood                -384.369
Date:                            Mon, 22 Feb 2021   AIC                            786.737
Time:                                    19:16:00   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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.36e+20. Standard errors may be unstable.
```

Figure-44: Manual SARIMA Model

YearMonth	Rose	rose_forecasted	rose_forecasted_log	manual_rose_forecasted
1991-01-01	54.0	44.213640	56.850665	44.733041
1991-02-01	55.0	62.326561	66.338168	64.208694
1991-03-01	66.0	67.313577	70.531810	65.110689
1991-04-01	65.0	63.160986	66.372950	68.453063
1991-05-01	60.0	66.474665	69.031777	61.423433

Figure-45: Manual SARIMA Forecasted Values

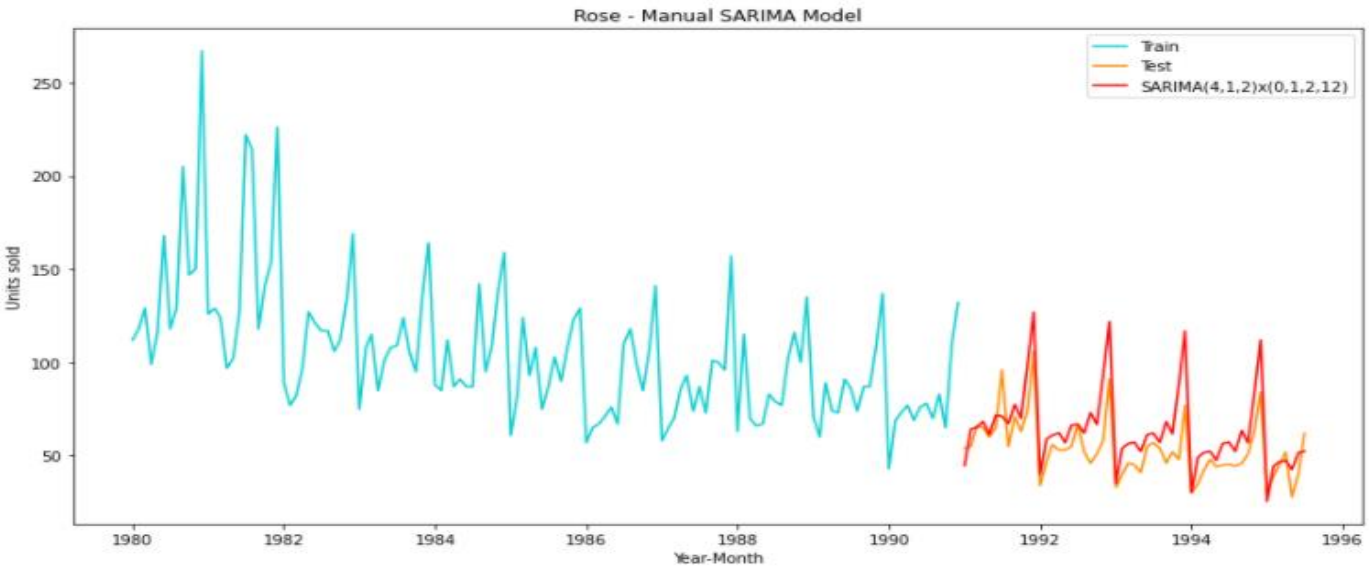


Figure-46: Plot Actual v/s Forecast Result on test data

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

Solution:

	Test RMSE
Alpha=0.1,Beta=0.2,gamma=0.3,TES_Iterative	9.943563
2 point TMA	11.530054
4 point TMA	14.458402
6 point TMA	14.572976
9 point TMA	14.732918
Alpha=0.0,Beta=0.0, DES Optimized	15.278359
RegressionOnTime	15.278369
Manual_SARIMA(4, 1, 2)*(0, 1, 2, 12)	15.388806
Auto_ARIMA(0, 1, 2)	15.629114
Auto_SARIMA(0, 1, 2)*(2, 1, 2, 12)	16.527864
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	17.921012
Alpha=0.065,Beta=0.054,gamma=0.0 TES Optimized	21.197161
Alpha=0.0987, SES Optimized	36.824478
Alpha=0.10,SES_Iterative	36.856268
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
SimpleAverage	53.488233
NaiveModel	79.745697
Manual_ARIMA(0,1,0)	84.160493

Figure-47: RMSE Values

→ 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.

Solution:

→ 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

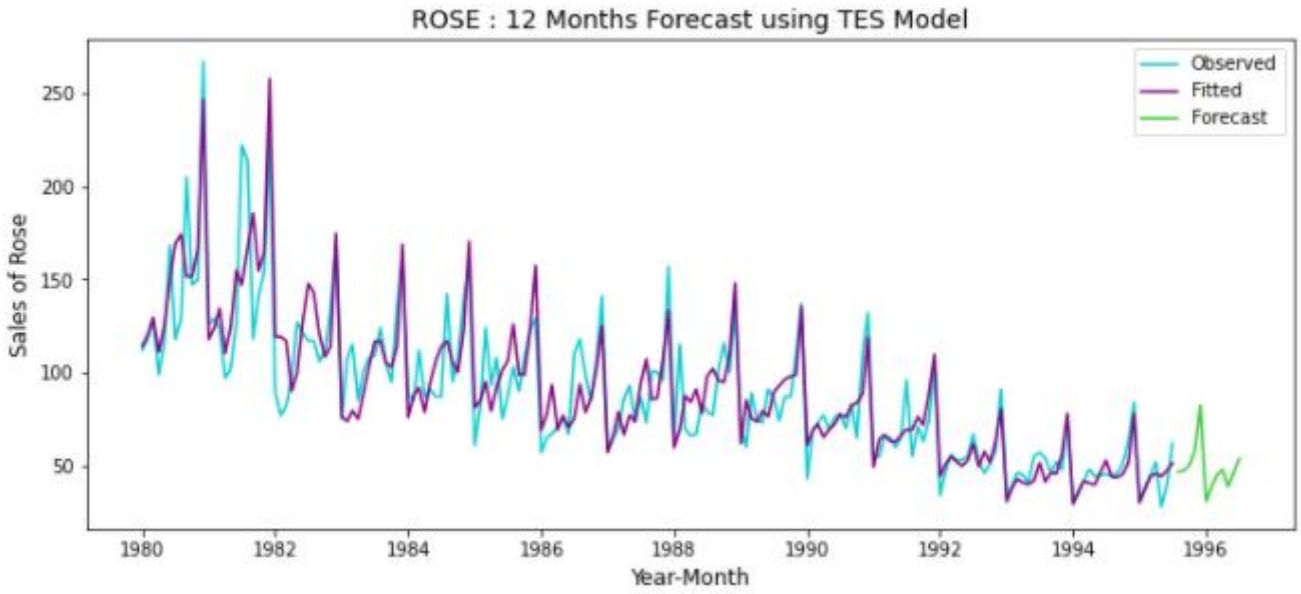


Figure-48: Plot Actual and Future Forecast Result

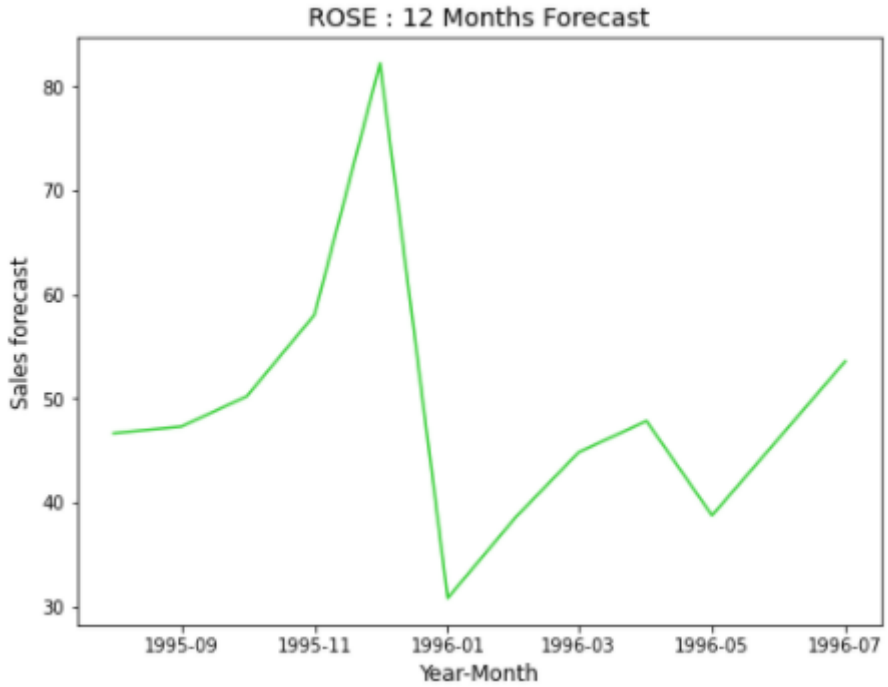


Figure-49: Future Forecast Result

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.

Solution:

```
1995-08-01    46.645790
1995-09-01    47.277864
1995-10-01    50.192393
1995-11-01    58.032965
1995-12-01    82.211766
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      82.211766
dtype: float64
```

Figure-50: 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.