PGPDSBA Online FEB A 2021

Pavan Kumar R Naik

PGP-DSBA Online

Feb A 2021

08/10/2021

Table of Contents

List of Figures

List of Tables

Problem:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze 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.

Data Set: Sparkling.csv

Q1.1. Read the data as an appropriate Time Series data and plot the data.

Solution:

Sample of Dataset:

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

Table 1.1. Dataset Sample (Sparkling)

- The data set contains two columns of monthly timestamp from Jan 1980 to July 1995 and sales of Sparkling wine for the respective month

Method 1:

- Create a time stamp and adding to the data frame as index by dropping the 'YearMonth' column

```
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 1.2. Time stamp index (Sparkling)

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

Table 1.3. Dataset sample post timestamp indexing (Sparkling)

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

Table 1.4. Dataset sample using parse_dates(Sparkling)

- Sparkling dataset has no missing values

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

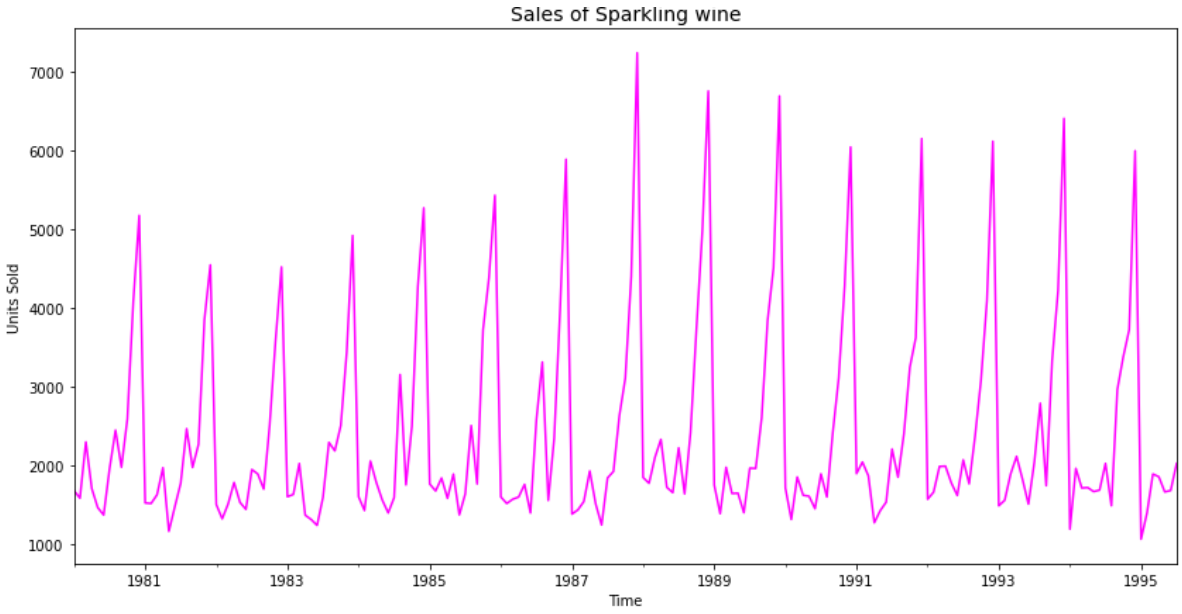


Fig 1.1. Time series plot (Sparkling)

Inference:

- Dataset shows significant seasonality and have no consistent trend however has the upward and downward slopes during the time period
- Wine is consistently favored over the years by the customers

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

Solution:

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

Table 1.5. Dataset summary (Sparkling)

- 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 for Sparkling Dataset:

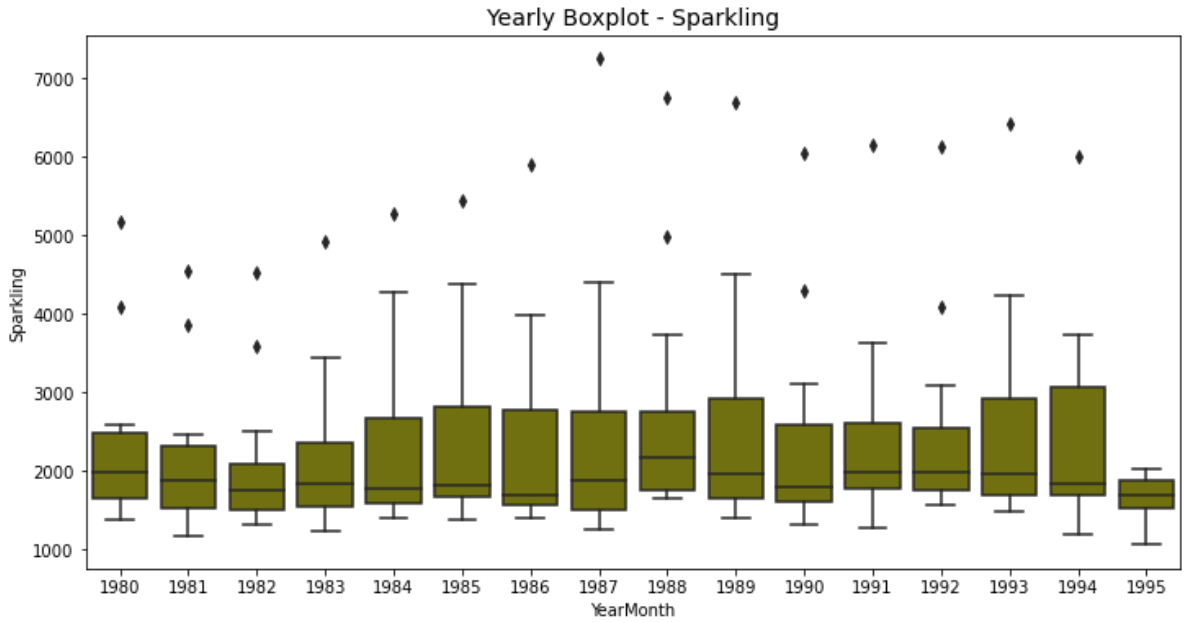


Fig 1.2. Yearly boxplot (Sparkling)

Monthly Boxplot for all the years for Sparkling Dataset:

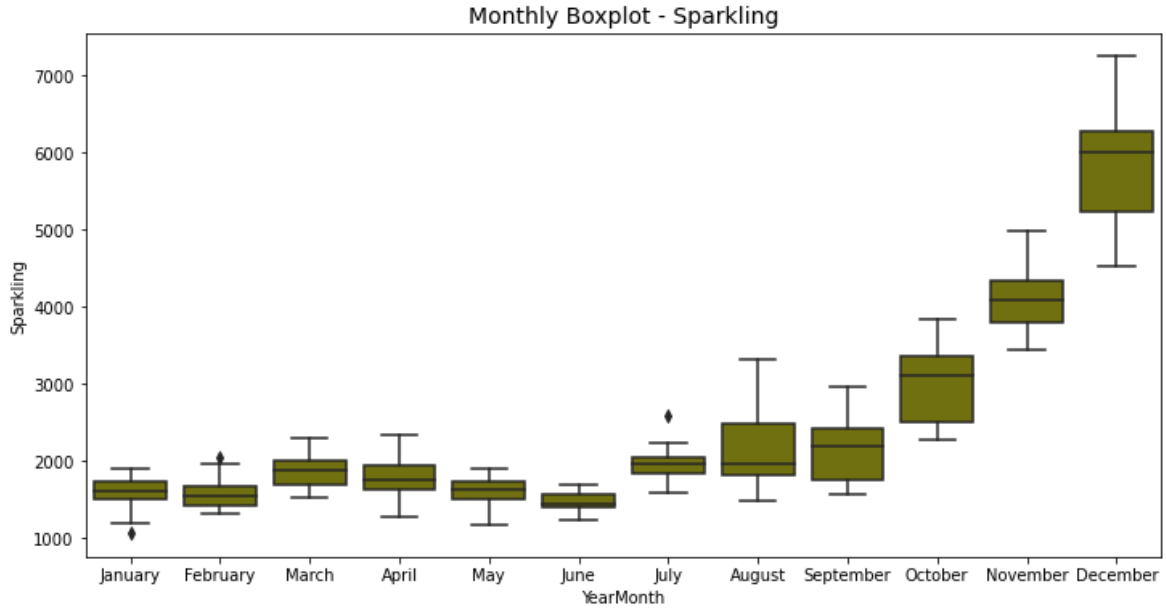


Fig 1.3. Monthly boxplot (Sparkling)

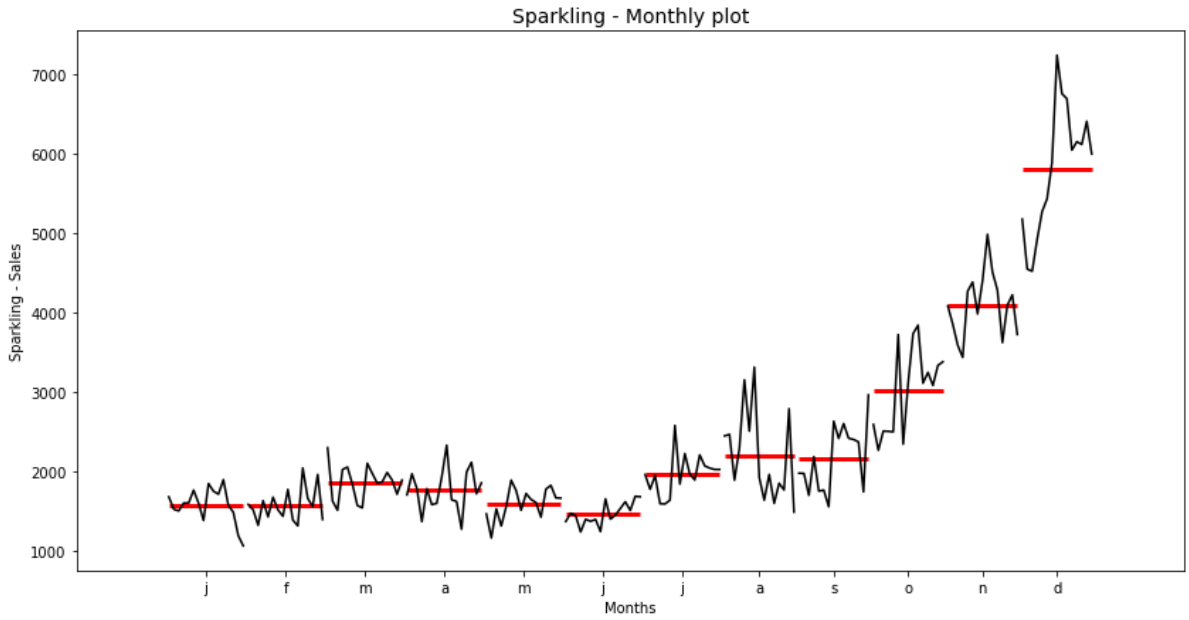


Fig 1.4. Monthly Time series plot (Sparkling)

Inference:

- 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 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
- The monthly plot for Sparkling shows means and variation of units sold each month over the years. Sales in seasonal month's shows a higher variation than in the lean months
- Sale in December with a mean few 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

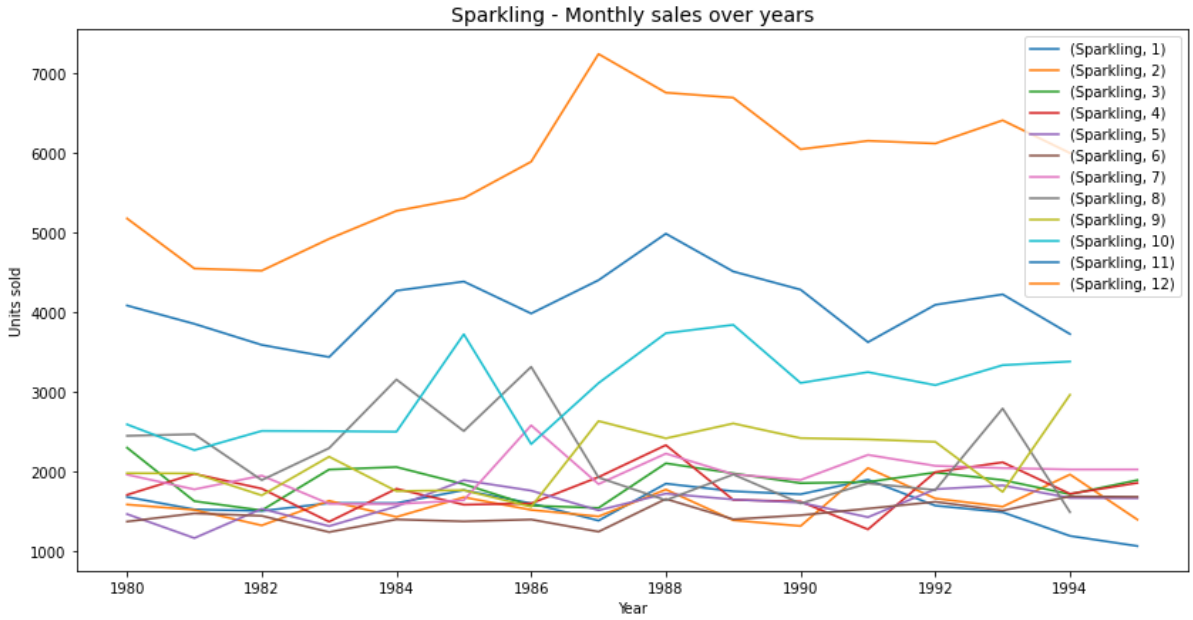


Fig 1.5. Monthly sales over the years (Sparkling)

- The highest volume of Sparkling wines 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:

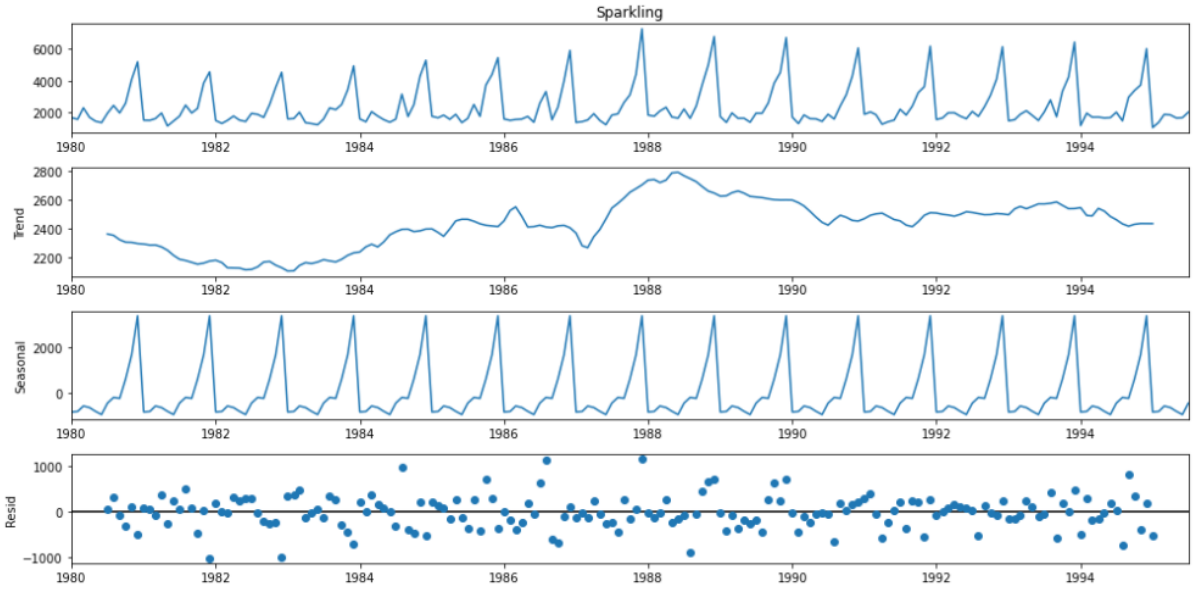


Fig 1.6. Additive model (Sparkling)

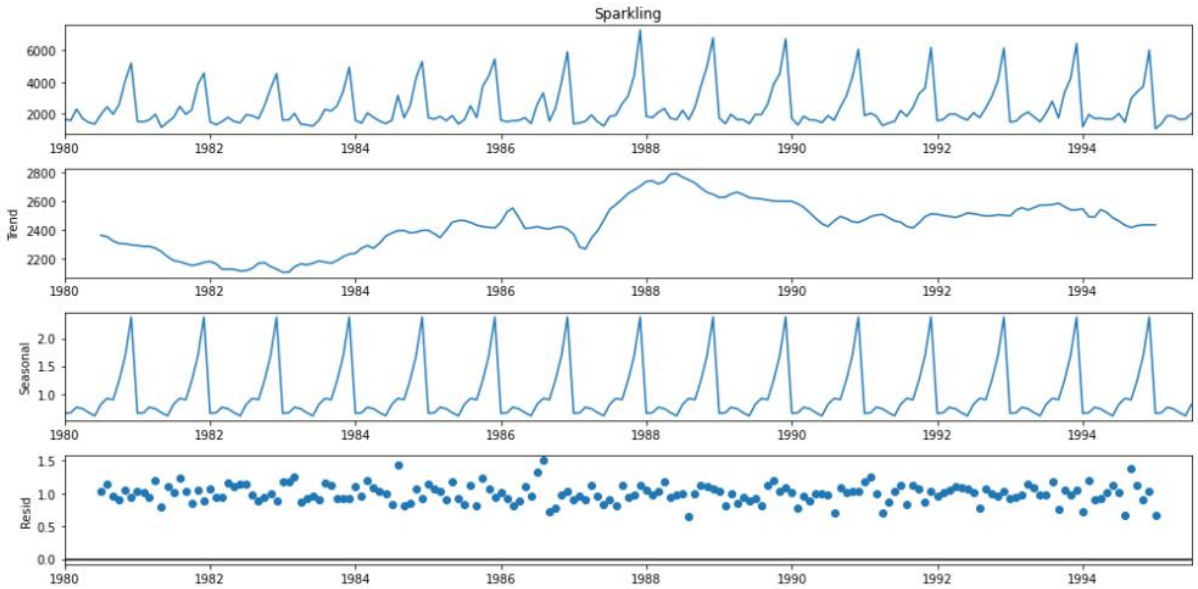


Fig 1.7. Multiplicative model (Sparkling)

Inference:

- 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 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

Q 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

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

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

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 Test Data:

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

Table 1.6. Train and Test split (Sparkling)

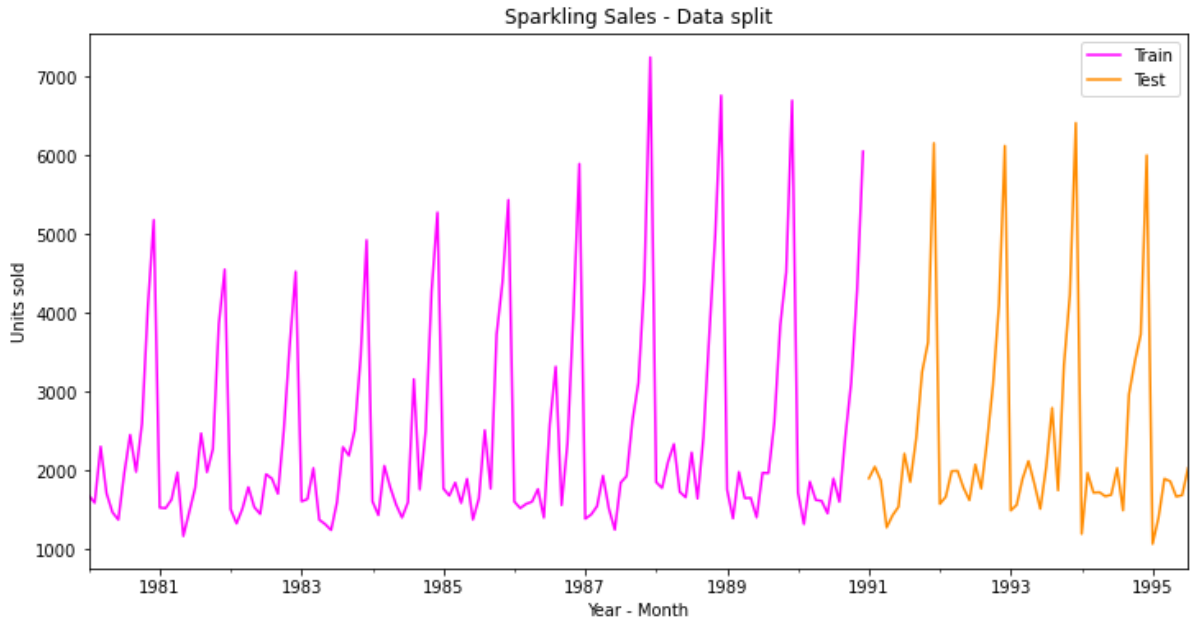


Fig 1.8. Train and Test plot (Sparkling)

Q 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.

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

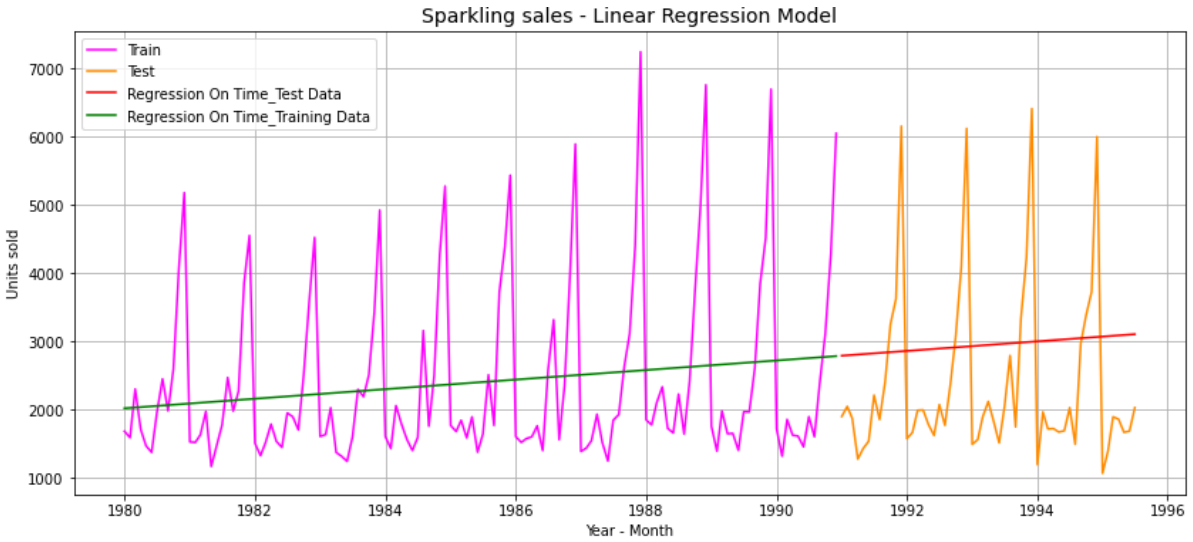


Fig 1.9. Linear Regression model (Sparkling)

- The linear regression plot shows a gradual upward trend in forecast of Sparkling wine, consistent with the observed trend which was not visually apparent - 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

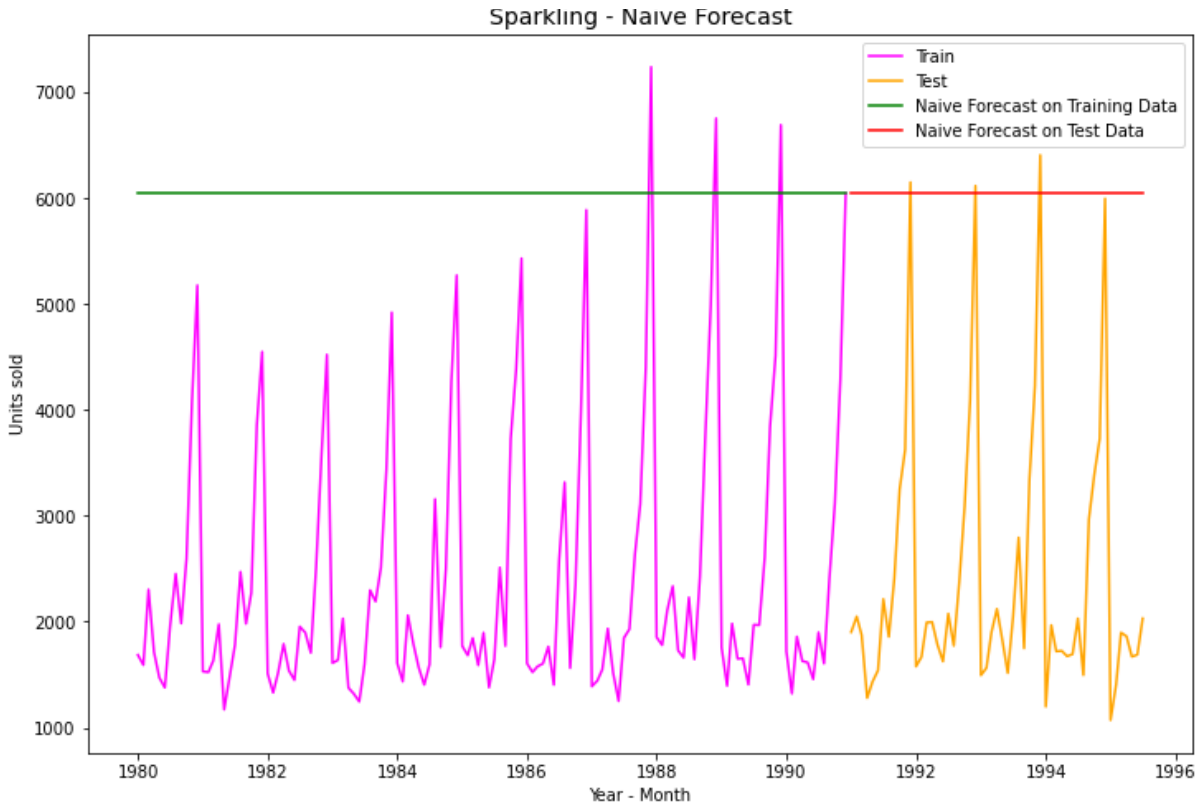


Fig 1.10. Naive model (Sparkling)

- 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 - RMSE is 3864.279

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
- The model is not capable of either forecasting or able to capture the trend and seasonality present in the dataset - RMSE is 1275

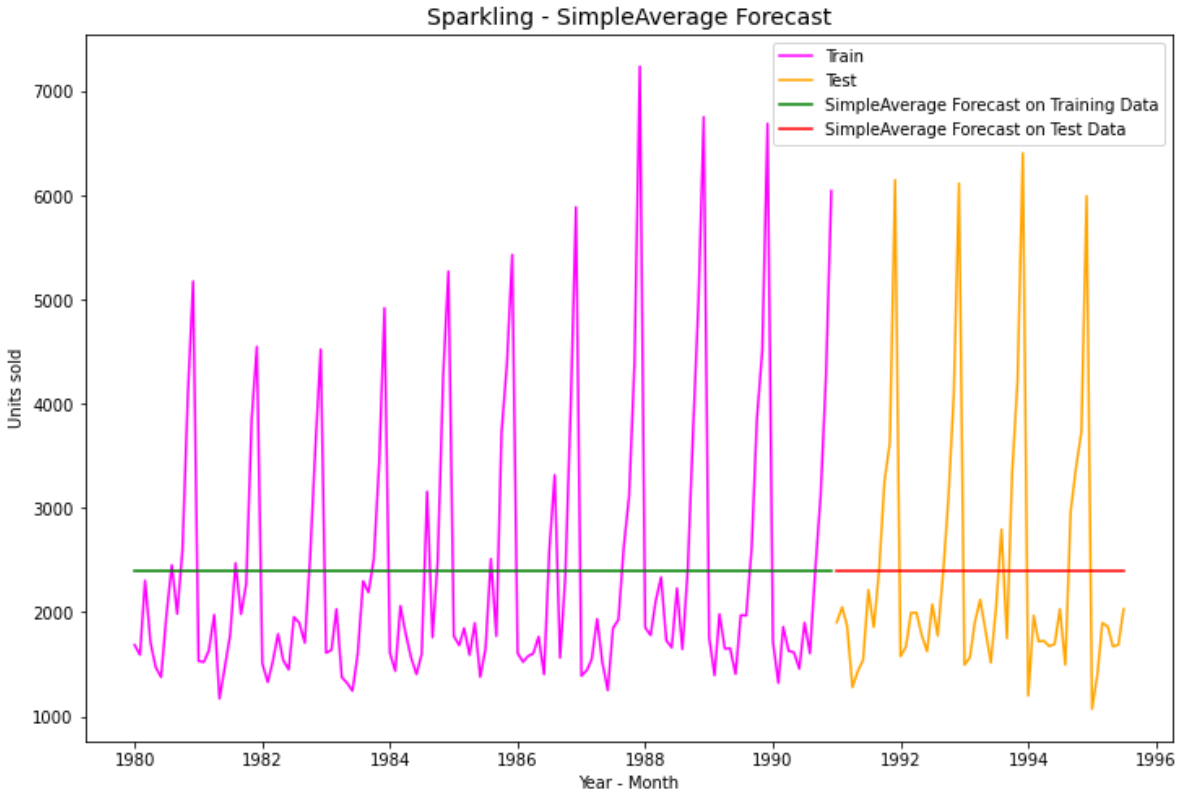


Fig 1.11. Simple average model (Sparkling)

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
- 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

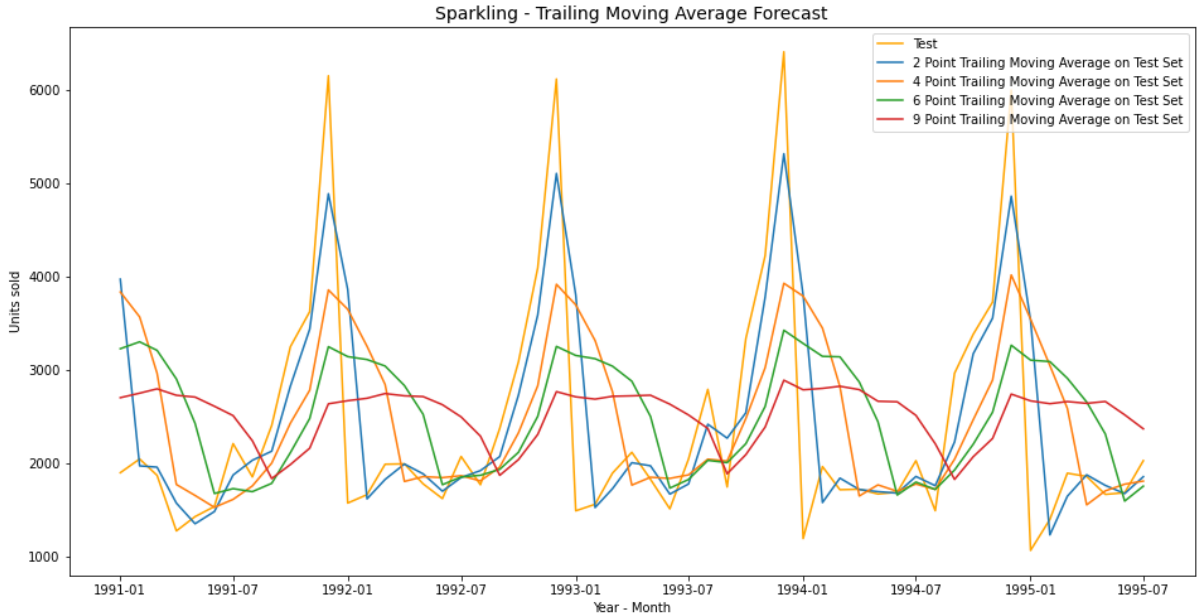


Fig 1.12. Trailing moving avg. model (Sparkling)

- Model Comparison:

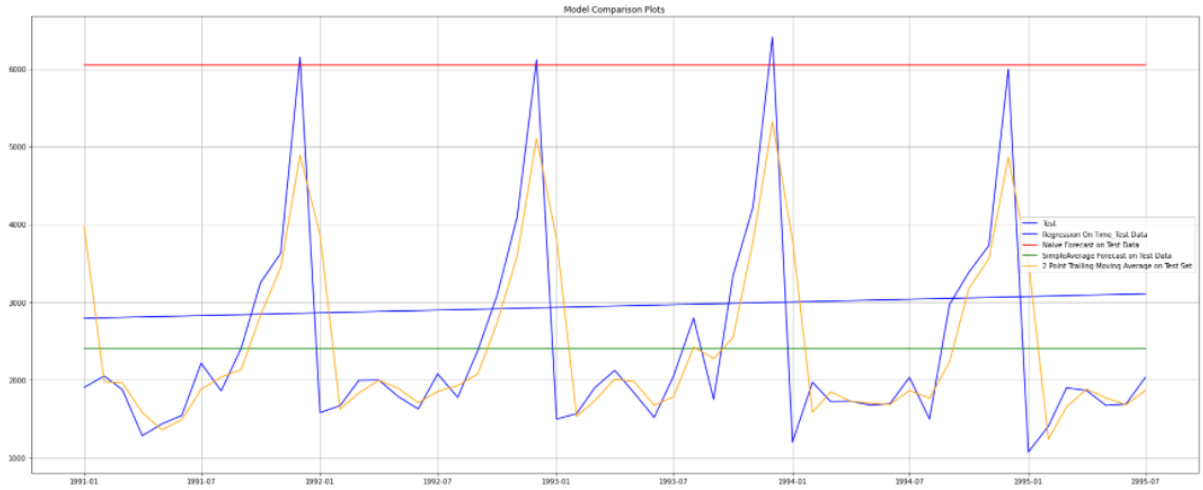


Fig 1.13. Model comparison plot (Sparkling)

- 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 1.7. RMSE on Test (Sparkling)

Model 5: Simple Exponential Smoothing

- The model is run 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.

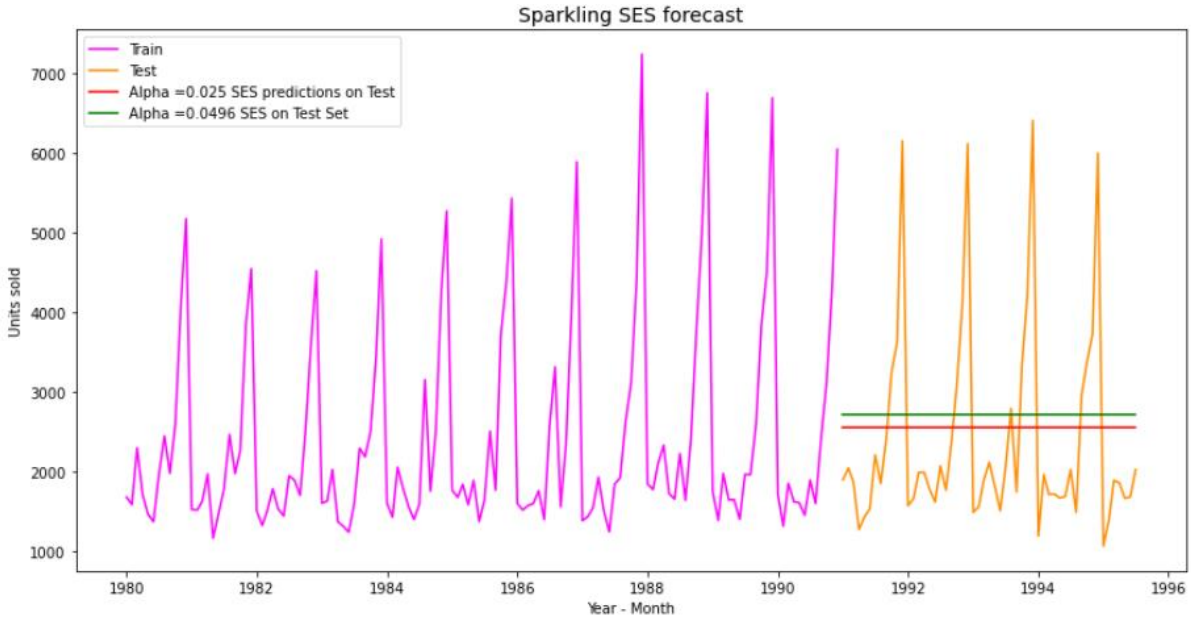


Fig 1.14. SES optimised and iterative model (Sparkling)

Model 6: Double 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

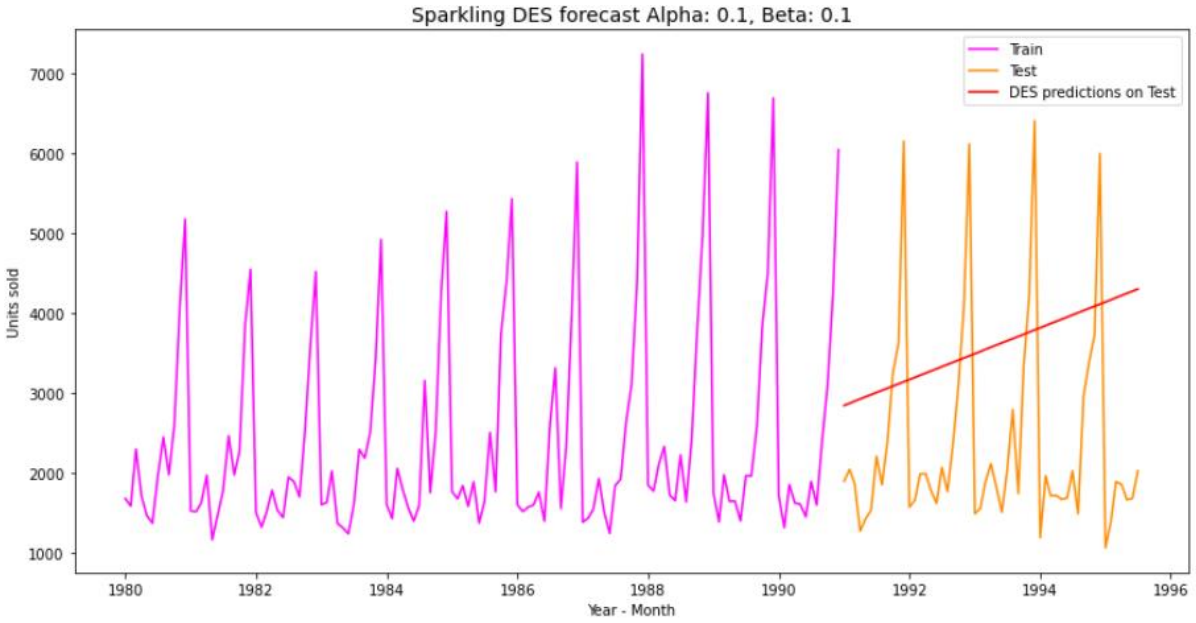


Fig 1.15. DES Iterative model (Sparkling)

Model 7: Triple Exponential Smoothing

- 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

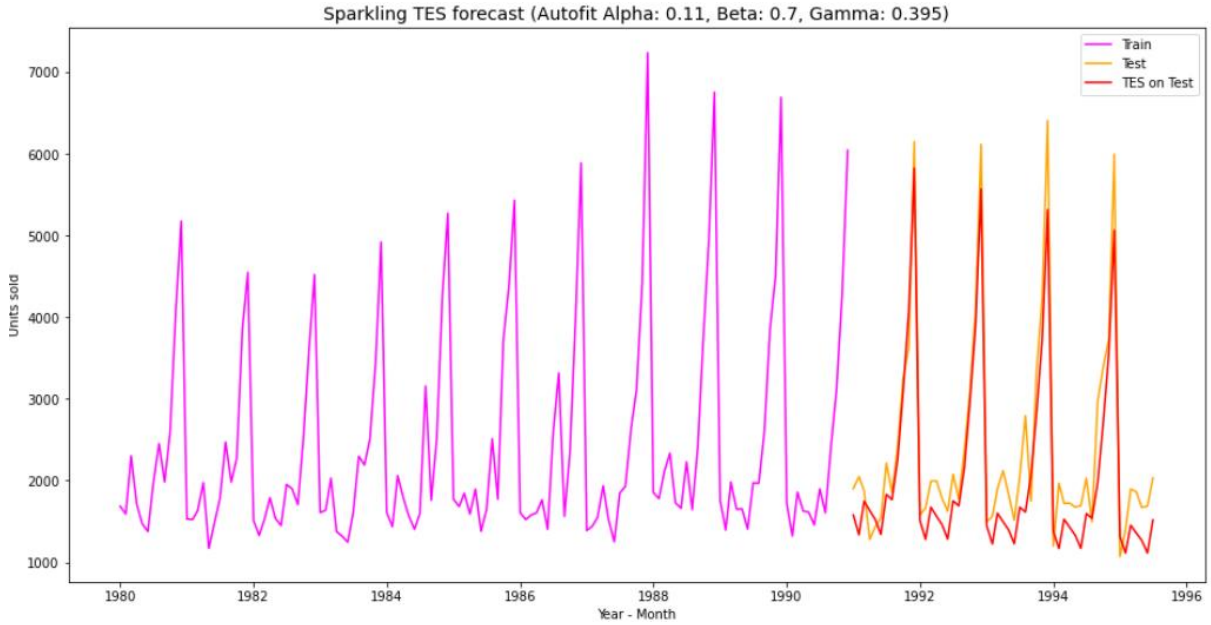


Fig 1.16. TES Autofit model (Sparkling)

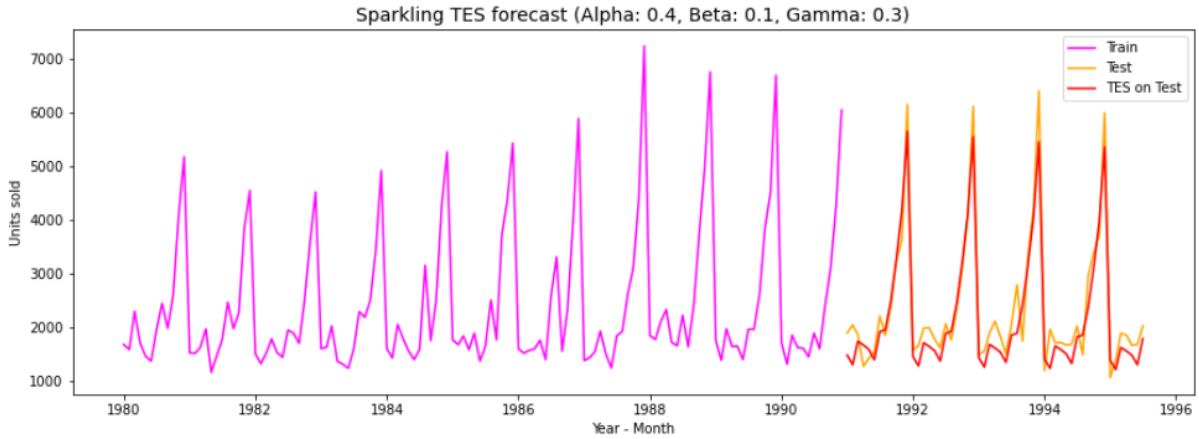


Fig 1.17. TES Iterative model (Sparkling)

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	463.501976
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

Table 1.8. RMSE Values (Sparkling)

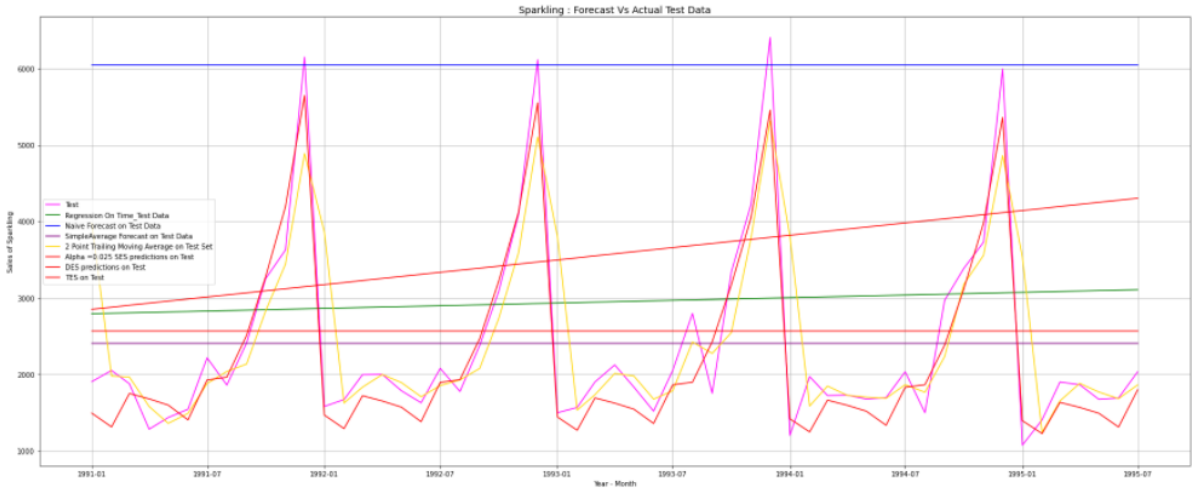


Fig 1.18. Forecast vs Actual (Sparkling)

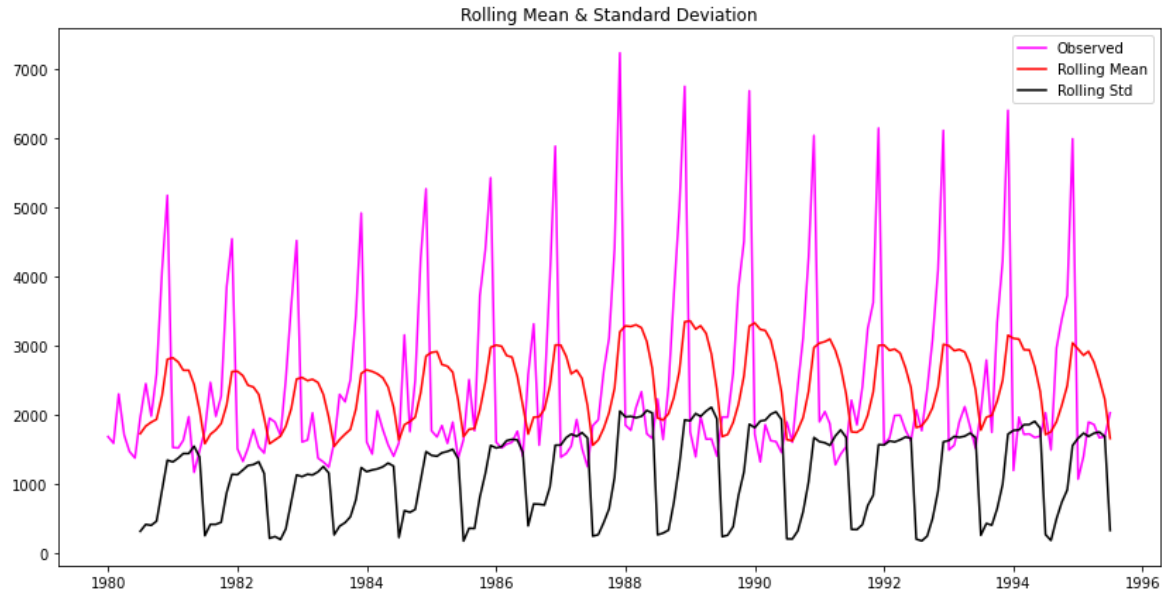
- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data
- 2-point trailing moving average model is also found to have fit well with a slight lag in test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

Q 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 determines 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 0.05 so we fail to reject the null hypothesis

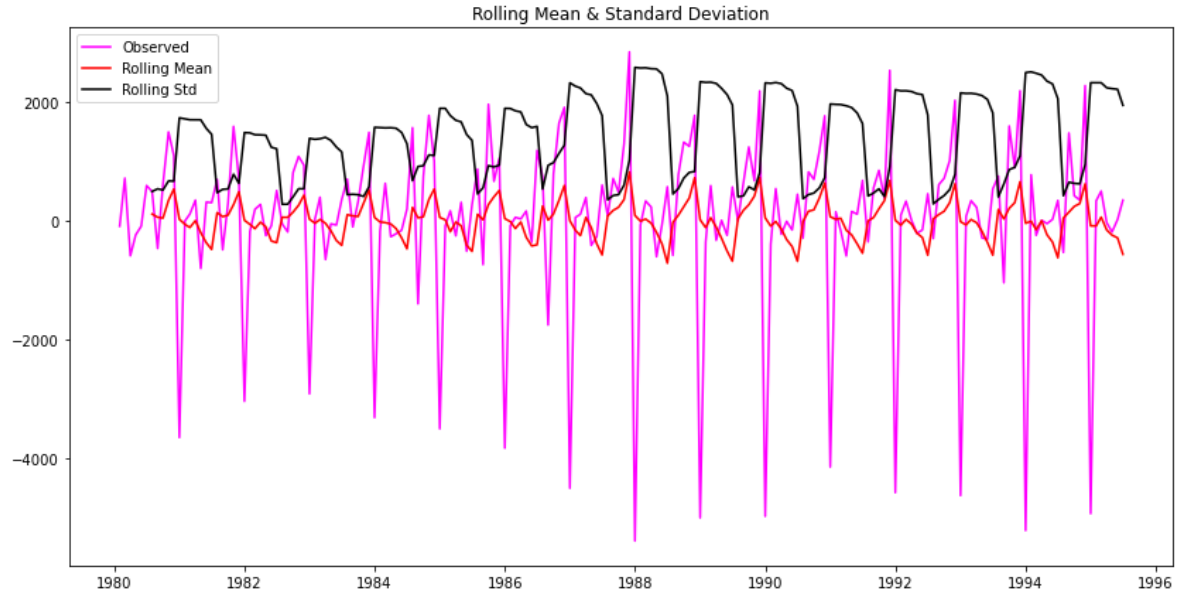



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

Fig 1.19. ADF test on original dataset (Sparkling)

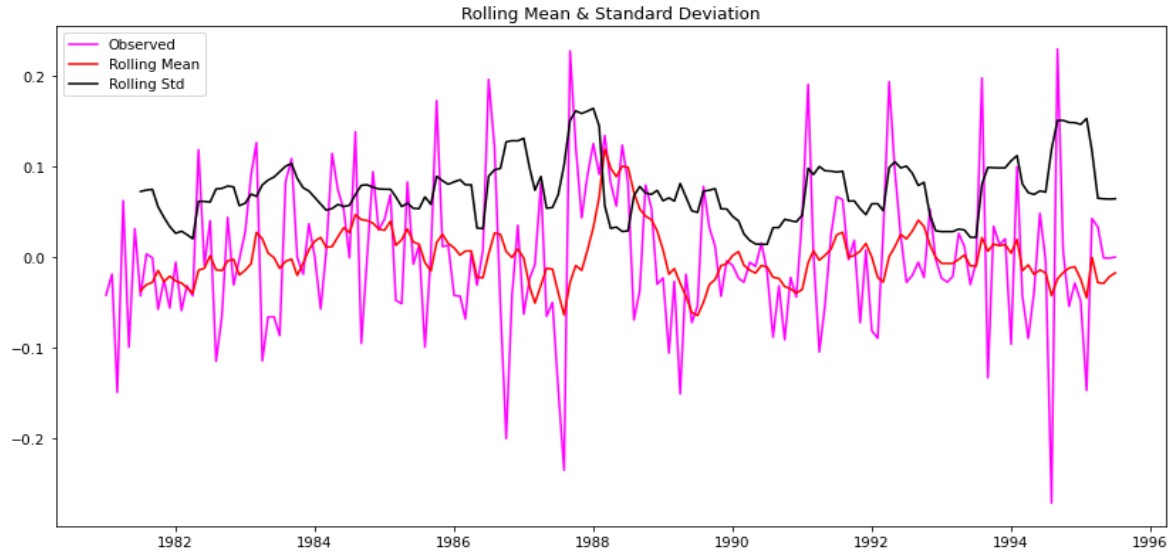
- Differencing of order one is applied on the Sparkling series as below and tested for stationarity. At an order of differencing 1, the series is found to be stationary as below
- The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if it's 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



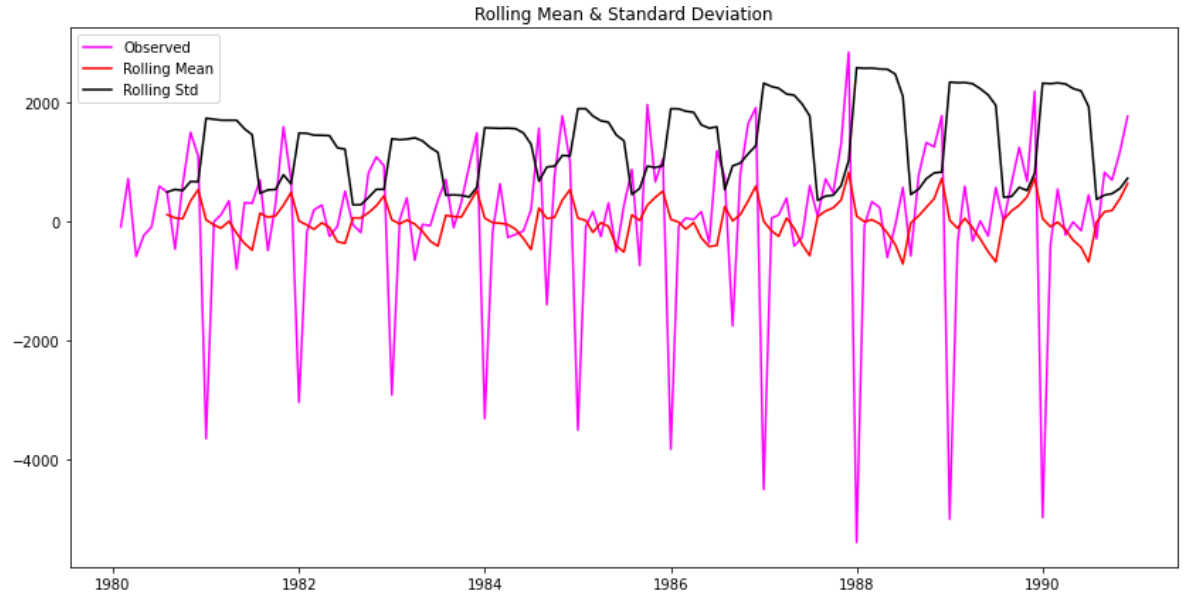

```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical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

Fig 1.20. ADF test with degree 1(Sparkling)




```
Results of Dickey-Fuller Test:
Test Statistic                  -5.183811
p-value                          0.000009
#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
```

Fig 1.21. ADF test on log series (Sparkling)




```
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
```

Fig 1.22. ADF test on train series with degree 1 (Sparkling)

Q 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

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.309
Method:	css-mle	S.D. of innovations	1012.640
Date:	Fri, 01 Oct 2021	AIC	2210.618
Time:	20:39:47	BIC	2227.869
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.628

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5855	0.517	10.811	0.000	4.573	6.598
ar.L1.D.Sparkling	1.2699	0.074	17.046	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	-1.9980	0.042	-47.154	0.000	-2.081	-1.915
ma.L2.D.Sparkling	0.9980	0.042	23.539	0.000	0.915	1.081

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	1.0005	+0.0000j	1.0005	0.0000
MA.2	1.0016	+0.0000j	1.0016	0.0000

Table 1.9. Auto ARIMA (Sparkling)

- ARIMA model was built using iterative function and found the least AIC value =2210.62 at (2, 1, 2) p, d, q
- 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 9: Auto SARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-685.174
Date:	Fri, 01 Oct 2021	AIC	1382.348
Time:	20:40:27	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 1.10. Auto SARIMA (Sparkling)

- SARIMA 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 RMSE value for this Auto- SARIMA model is 382.58

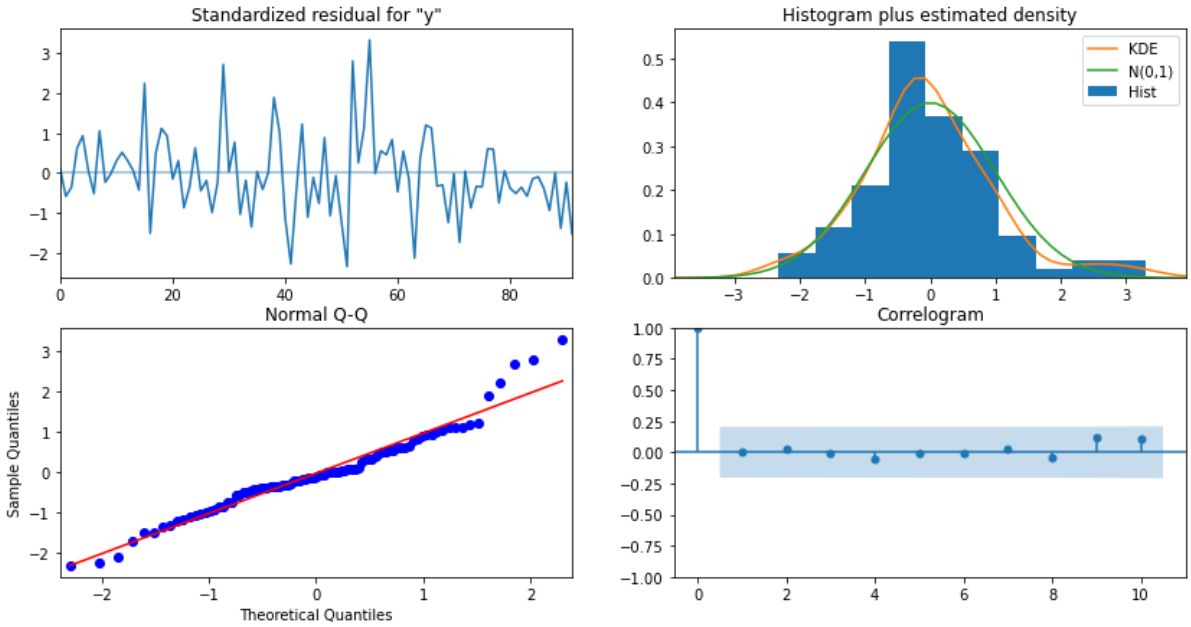


Fig 1.23. Diagnostic plot (Sparkling)

- 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 autocorrelation of the residuals and there are no significant lags above the confidence index

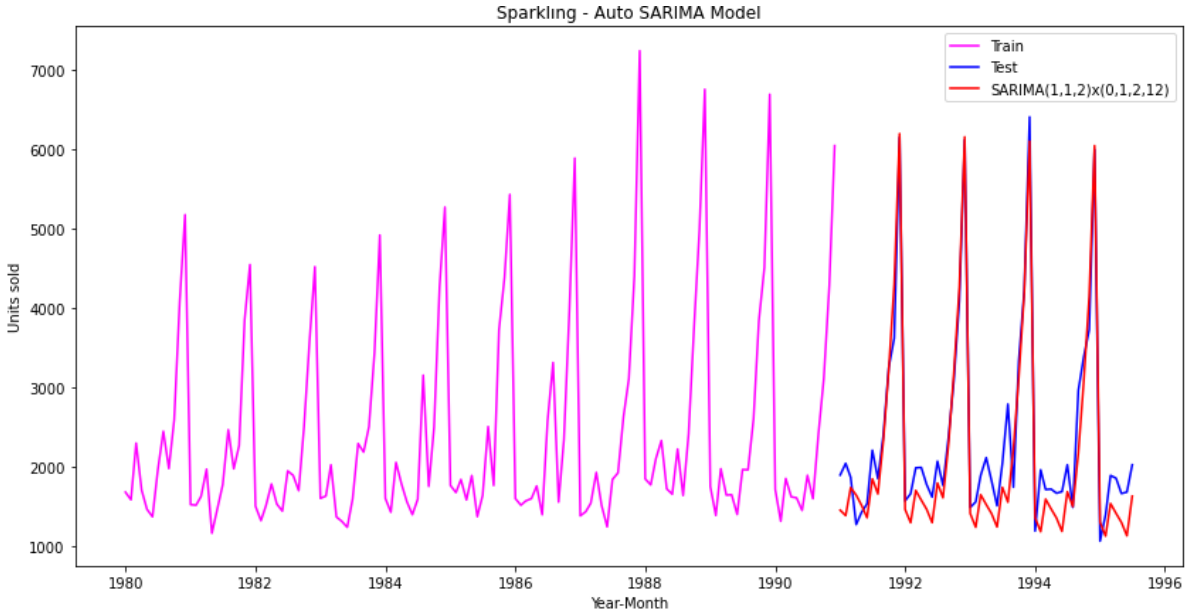


Fig 1.24. Actual vs forecast using SARIMA(Sparkling)

Model 10: Auto SARIMA on Log series

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Log Likelihood	146.236
Date:	Fri, 01 Oct 2021	AIC	-284.472
Time:	20:41:53	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	-19.860	0.000	-0.985	-0.808
ar.S.L12	1.0112	0.020	49.864	0.000	0.971	1.051
ma.S.L12	-0.6489	0.075	-8.627	0.000	-0.796	-0.501
sigma2	0.0045	0.001	7.841	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

Table 1.11. Auto SARIMA Log (Sparkling)

- 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 RMSE value for this Auto- SARIMA Log model is 336.58

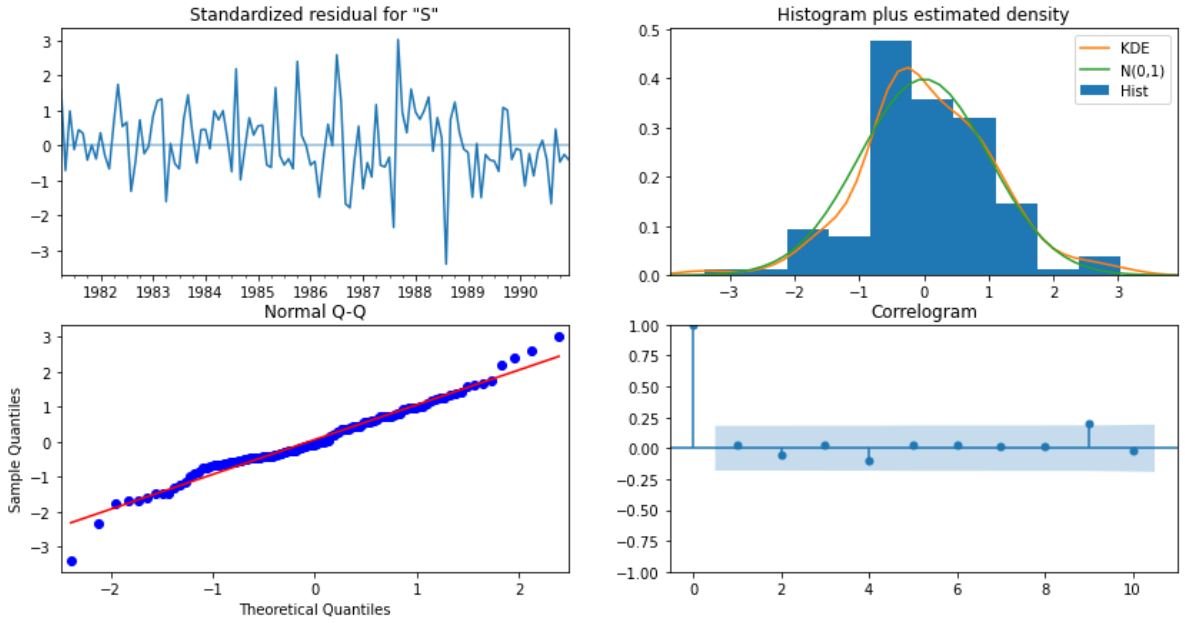


Fig 1.25. Diagnostic plot (Sparkling)

- The he 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 autocorrelation of the residuals and there are no significant lags above the confidence index.
- 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

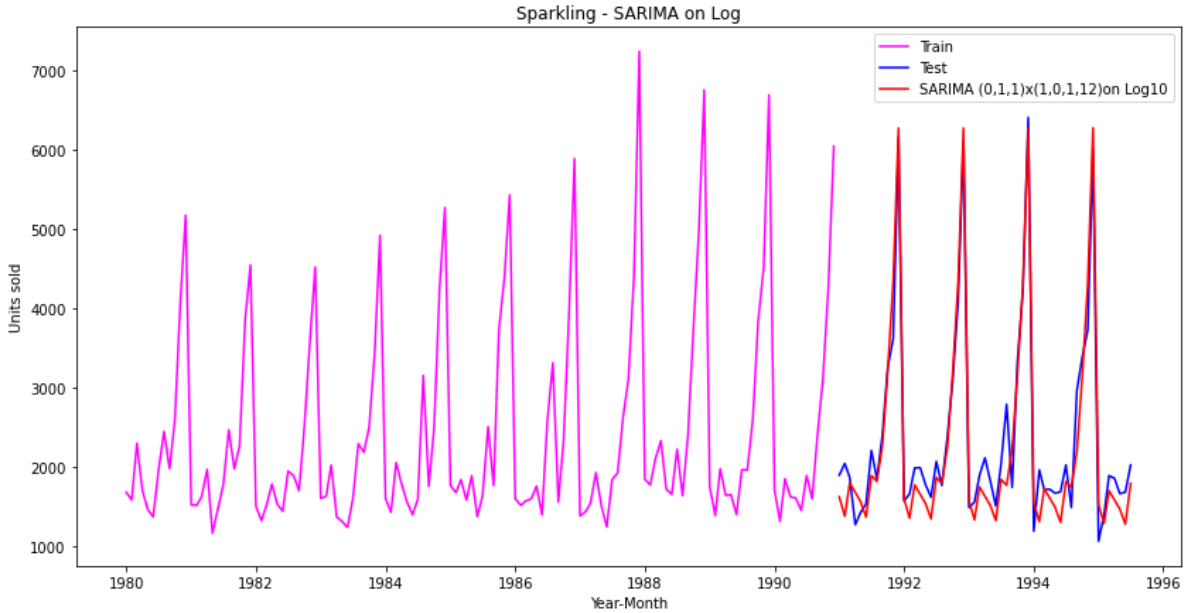


Fig 1.26. Actual vs forecast using SARIMA Log (Sparkling)

Q 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 11: Manual ARIMA

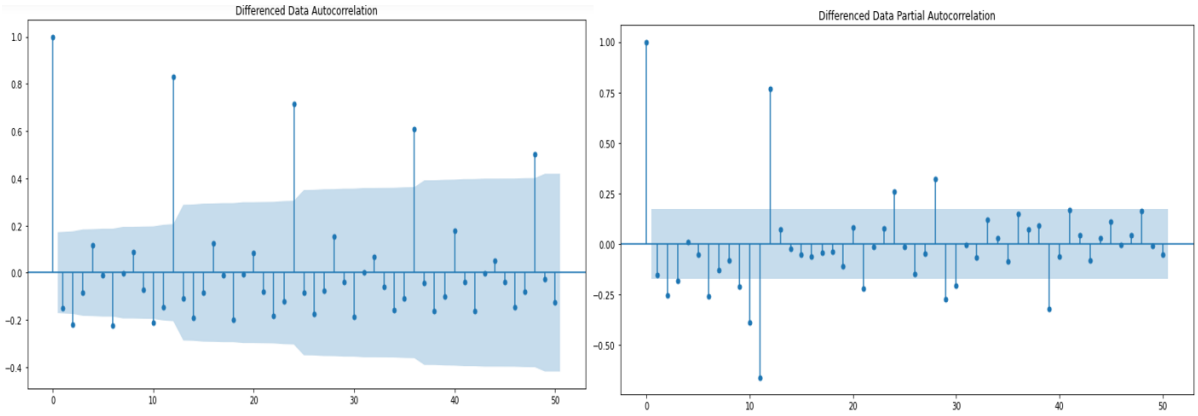


Fig 1.27. ACF and PCF plot (Sparkling)

- 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:	Fri, 01 Oct 2021	AIC	2269.583
Time:	20:41:55	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

Table 1.12. Manual Arima summary (Sparkling)

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

Model 12: Manual SARIMA

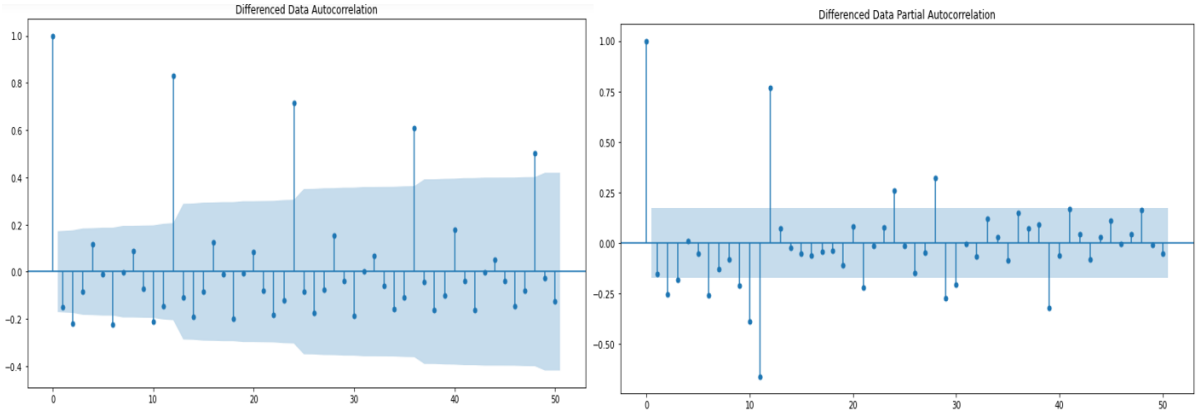


Fig 1.28. ACF and PCF plot (Sparkling)

- it can be inferred that at seasonal interval of 12, the plot is not quickly tapering off. Hence a seasonal differencing of 12 has to be taken

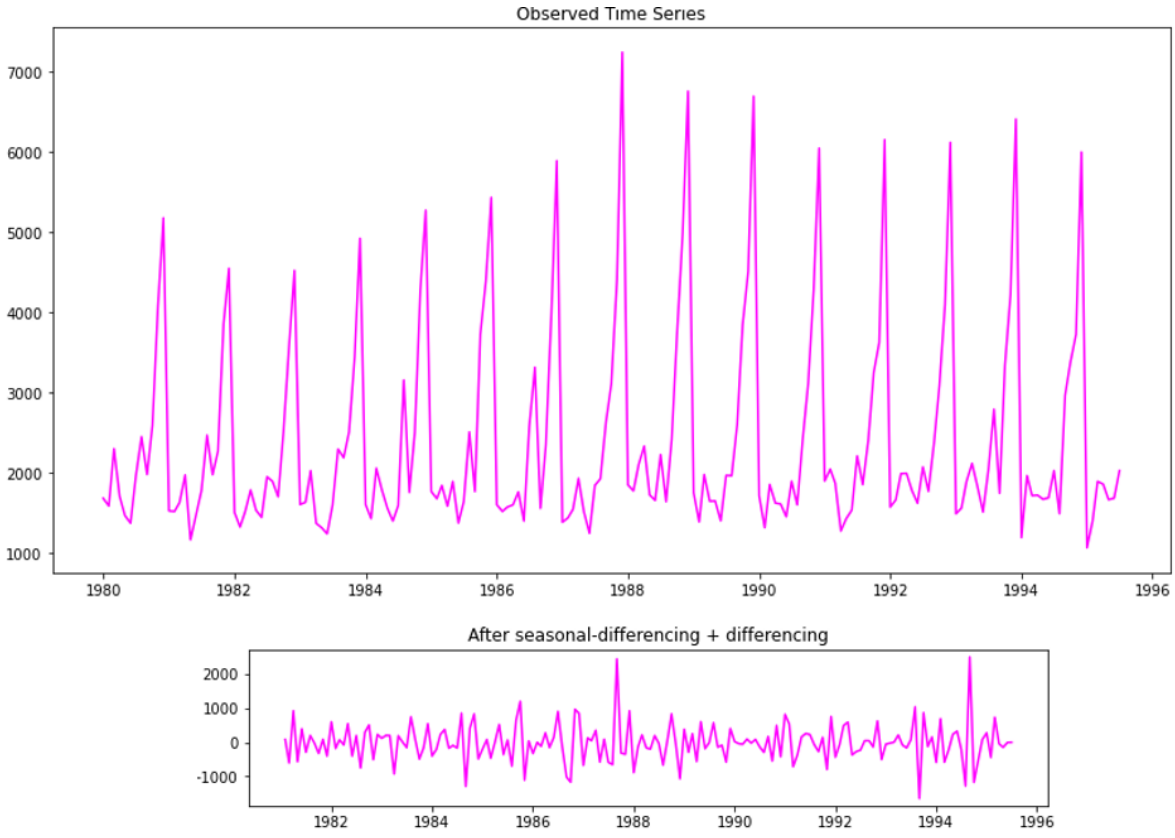
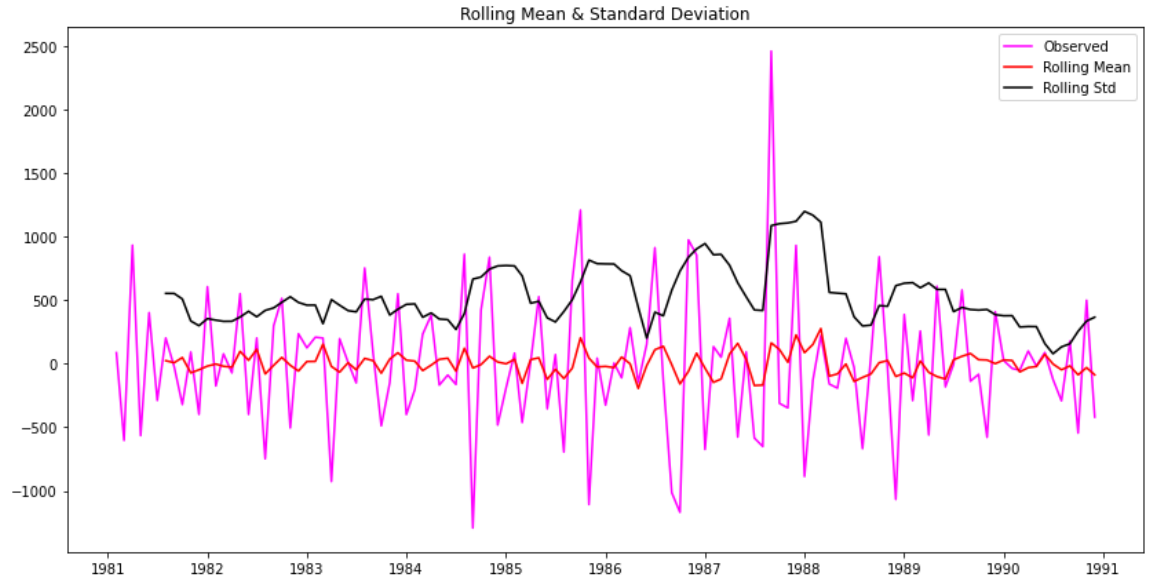


Fig 1.29. Time series plot (Sparkling)

- 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
```

Fig 1.30. ADF Test (Sparkling)

- 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)

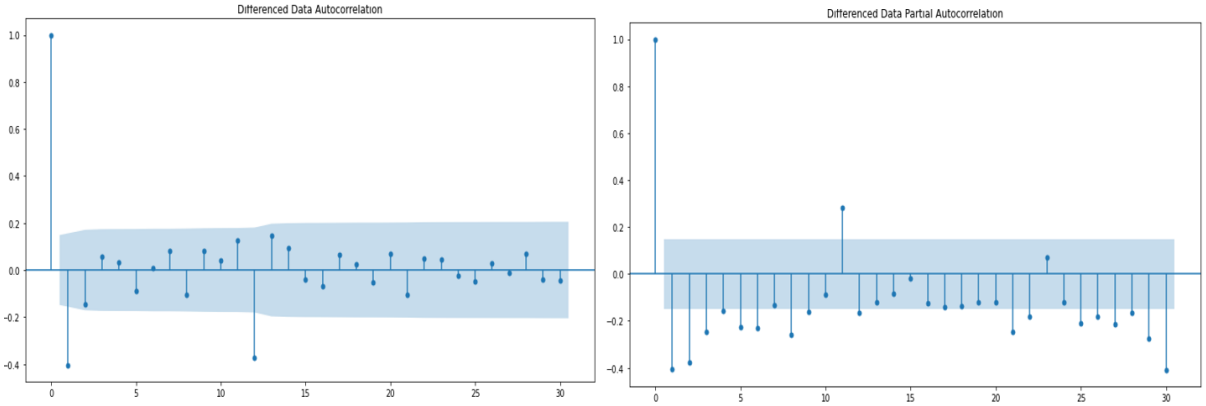


Fig 1.31. ACF and PACF plot of differenced (Sparkling)

- 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 (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 autocorrelation of the residuals and there are no significant lags above the confidence index.
  → The RMSE values of the automated SARIMA model is 324.1

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-693.697
Date:	Fri, 01 Oct 2021	AIC	1403.394
Time:	20:41:58	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

Table 1.13. Manual SARIMA summary (Sparkling)

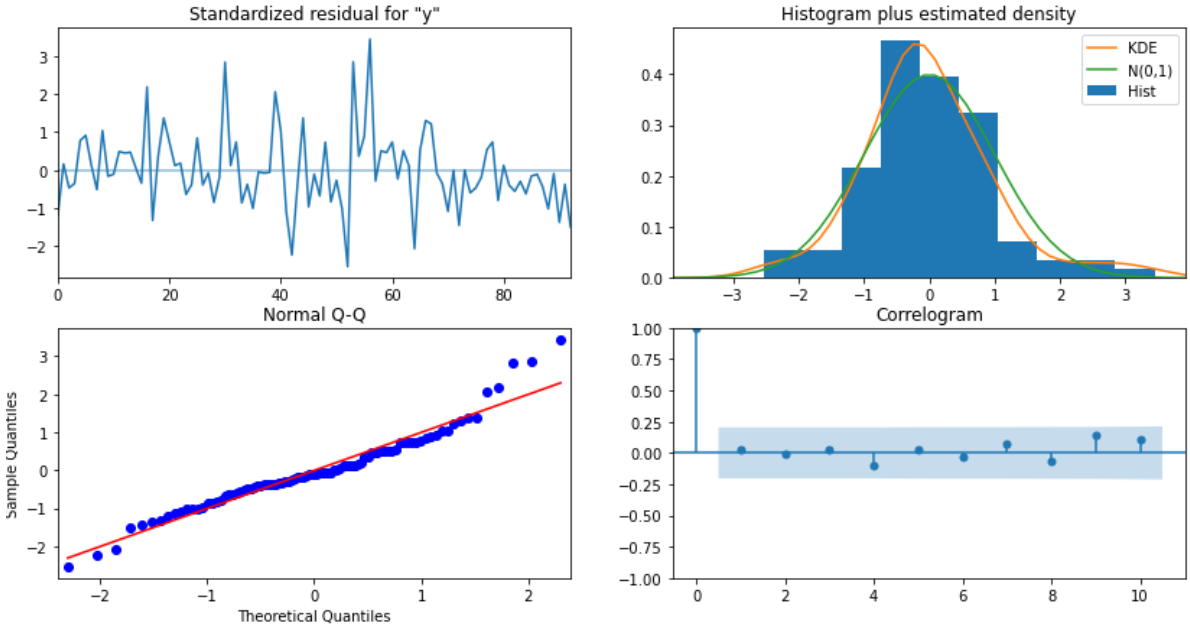


Fig 1.32. Diagnostic plot Manual SARIMA (Sparkling)

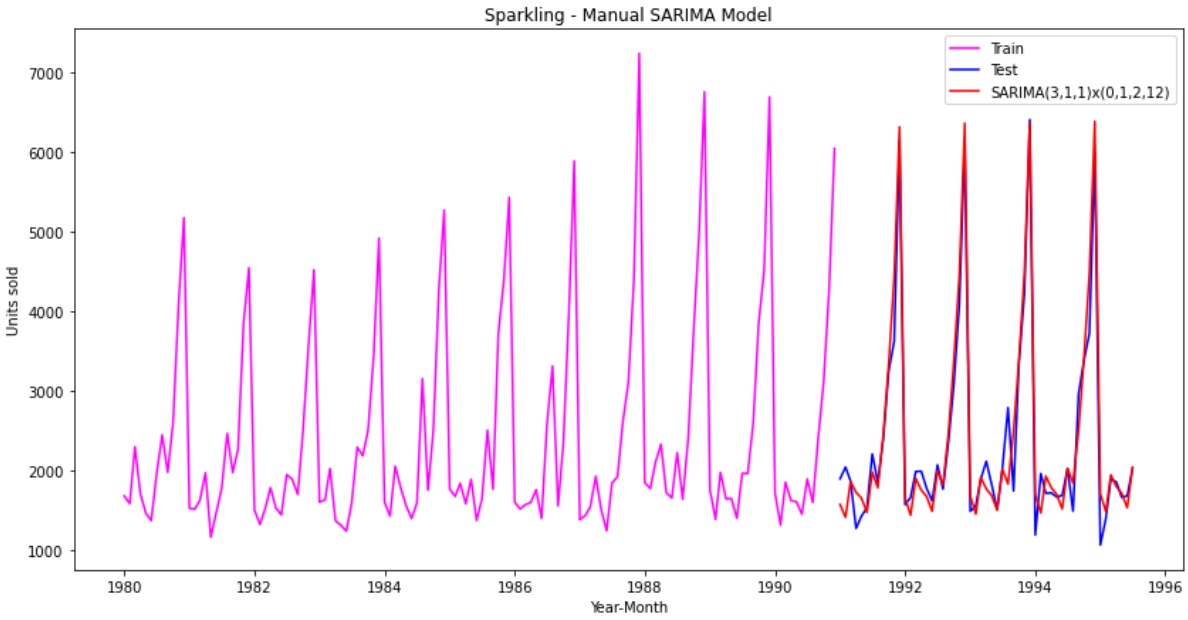


Fig 1.33. Actual vs forecast using Manual SARIMA(Sparkling)

Q 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)	324.107744
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	336.786145
Alpha=0.4,Beta=0.1,gamma=0.3,TES iterative	371.367690
Auto_SARIMA(1, 1, 2)*(0, 1, 2, 12)	382.576744
Alpha=0.11,Beta=0.7,gamma=0.395 TES Optimized	463.501976
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)	1374.649208
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

Table 1.14. RMSE values of all models (Sparkling)

Q 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, Manual 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 - RMSE is 547.591
- 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

SARIMAX Results

Dep. Variable:	y	No. Observations:	187
Model:	SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-1094.342
Date:	Fri, 01 Oct 2021	AIC	2204.685
Time:	20:42:01	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 1.15. Manual SARIMA on full dataset (Sparkling)

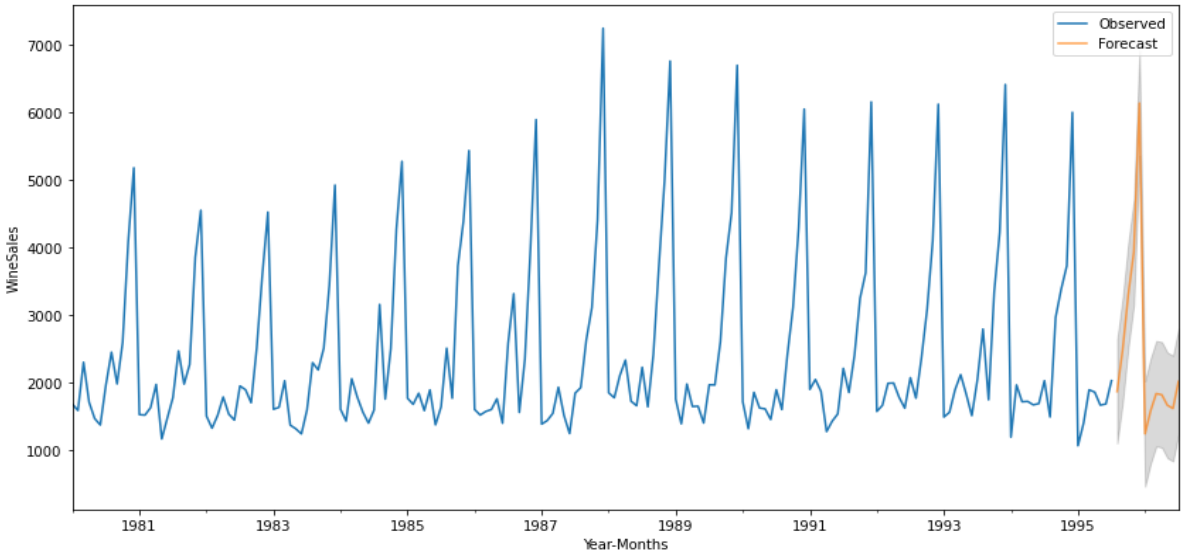


Fig 1.34. Actual vs Future 12 months forecast (Sparkling)

Q 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.

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    1870.888561
1995-09-30    2489.623620
1995-10-31    3299.650018
1995-11-30    3934.056600
1995-12-31    6135.396022
1996-01-31    1245.727169
1996-02-29    1584.643741
1996-03-31    1840.705251
1996-04-30    1823.847822
1996-05-31    1668.706091
1996-06-30    1620.472484
1996-07-31    2020.534842
Freq: M, Name: mean, dtype: float64
```

Table 1.16. Forecast 12 months (Sparkling)

```
count      12.000000
mean     2461.187685
std      1391.118211
min      1245.727169
25%      1656.647689
50%      1855.796906
75%      2692.130219
max      6135.396022
Name: mean, dtype: float64
```

Table 1.17. Forecast 12 months summary (Sparkling)

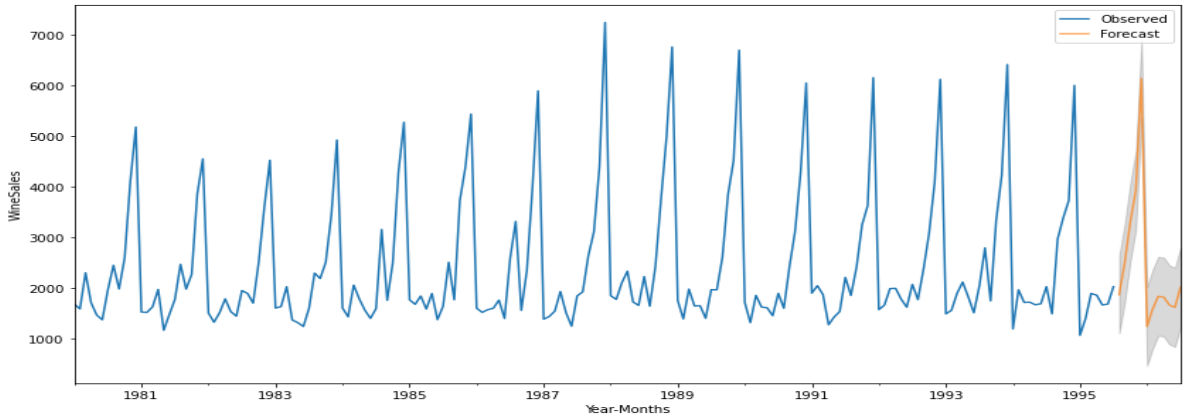


Fig 1.34. Actual vs Future 12 months forecast (Sparkling)

Data Set: Rose.csv

Q2.1. Read the data as an appropriate Time Series data and plot the data.

Solution:

Sample of Dataset:

	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

Table 2.1. Dataset Sample (Rose)

- The data set contains two columns of monthly time stamp from Jan 1980 to July 1995 and the sales corresponding to the wine

Method 1:

- Create a time stamp and adding to the data frame as index by dropping the 'YearMonth' column

```
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 2.2. Time stamp index (Rose)

	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

Table 2.3. Dataset sample post timestamp indexing (Rose)

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
```

Table 2.4. Dataset sample using parse_dates (Rose)

- Rose dataset has 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    45.364189
1994-08-01    44.279246
Name: Rose, dtype: float64
```

Table 2.5. Imputed values for missing data (Rose)

Plot the Rose Time Series to understand the behavior of the data:

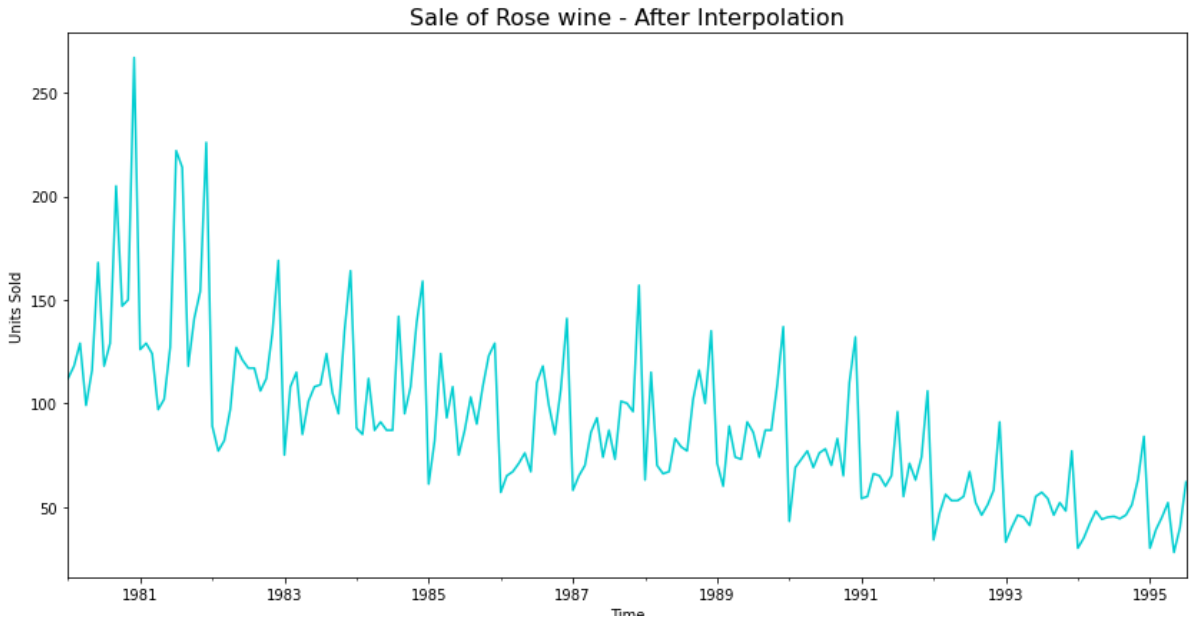


Fig 2.1. Time series plot (Rose)

Inference:

- 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-favor over the years

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

Solution:

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 2.6. Dataset summary (Rose)

- 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 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

Yearly Boxplot for Rose Dataset:

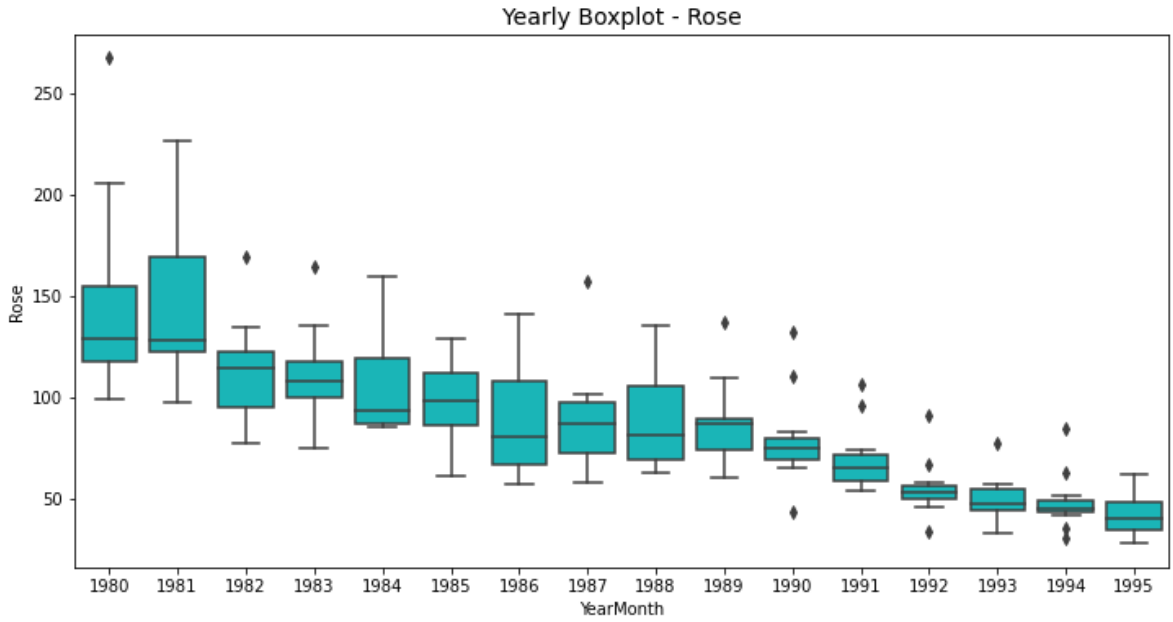


Fig 2.2. Yearly boxplot (Rose)

Monthly Boxplot for all the years for Rose Dataset:

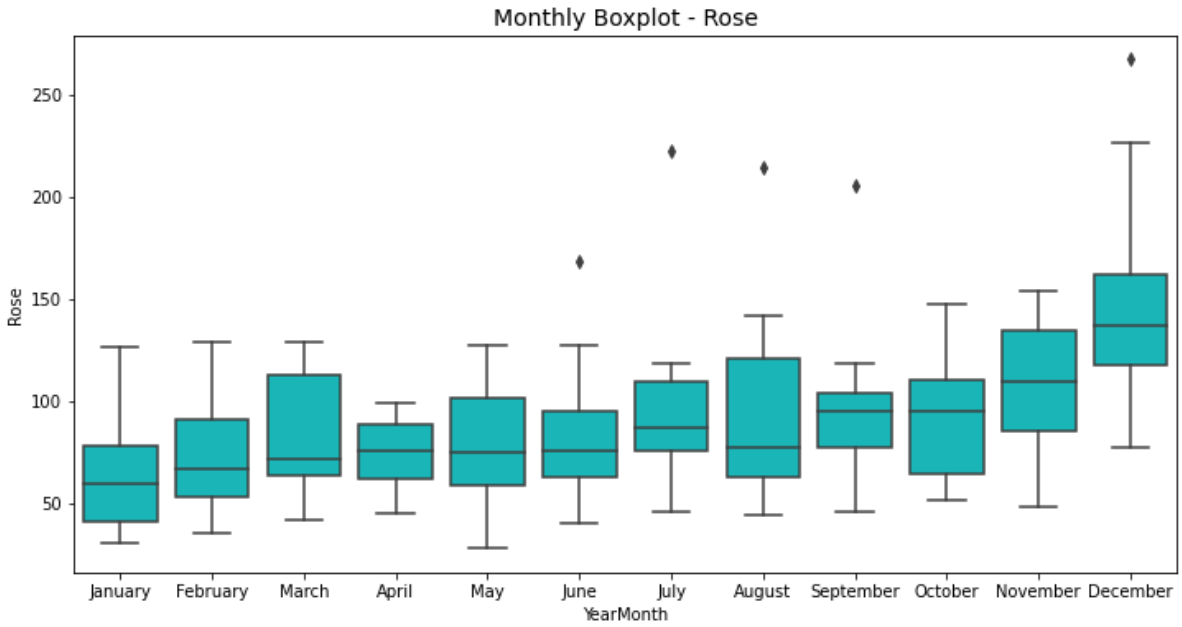


Fig 2.3. Monthly boxplot (Rose)

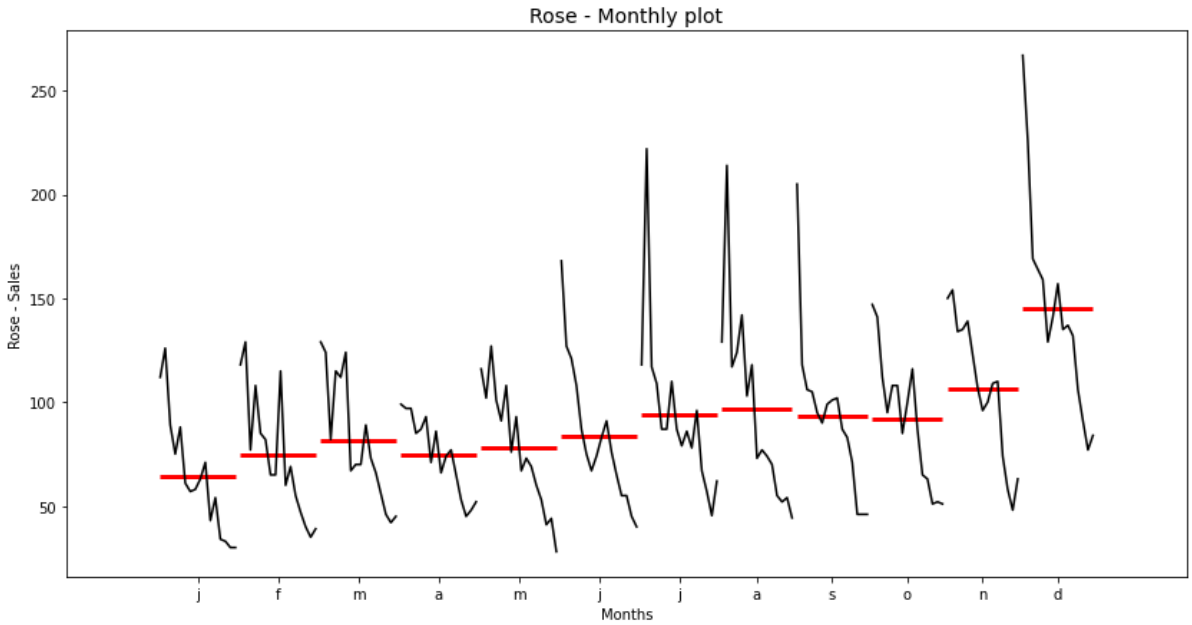


Fig 2.4. Monthly Time series plot (Rose)

Inference:

- 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.

- 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 show a higher variation than the rest
- Sale in December with a mean few 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:

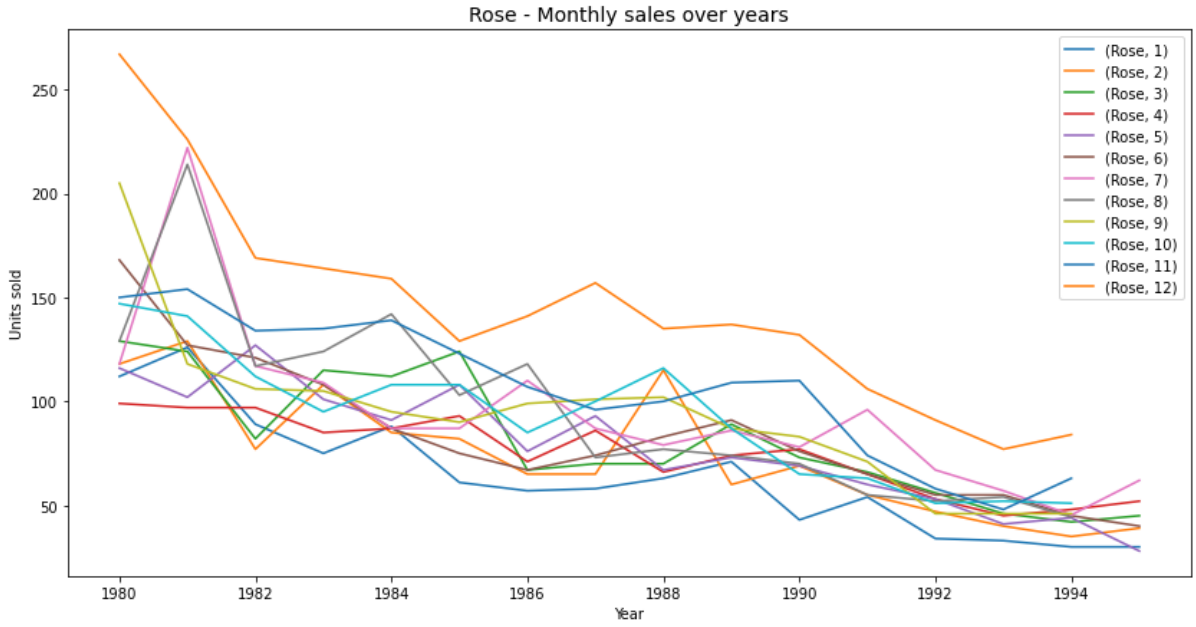


Fig 2.5. Monthly sales over the years (Rose)

- 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 was 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:

Decomposition of Rose Time Series with multiplicative Seasonality:

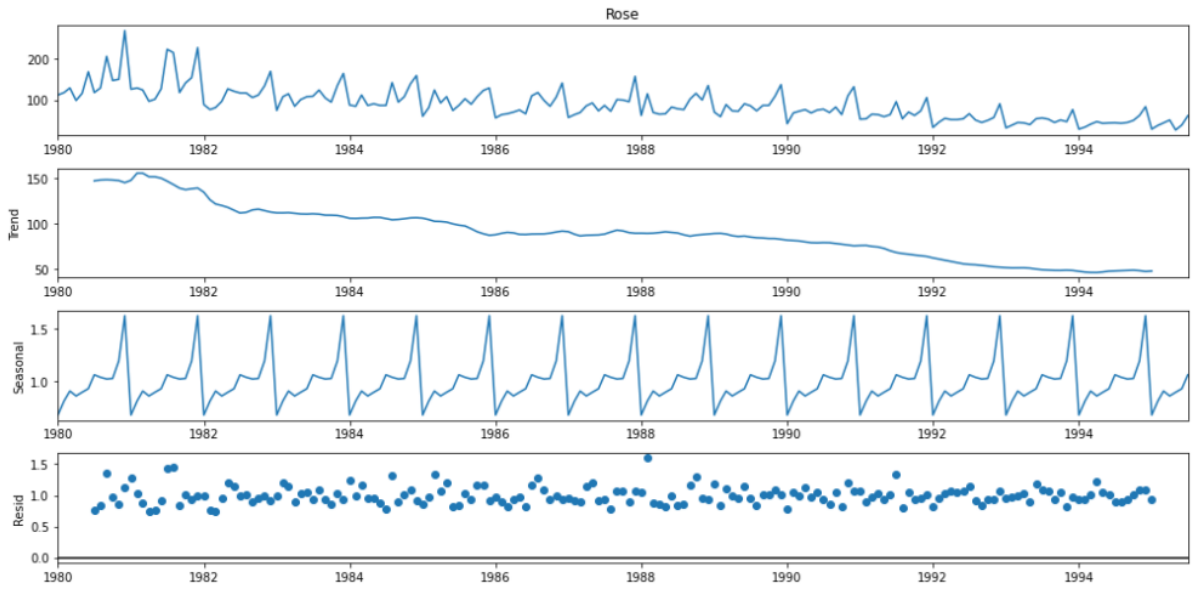


Fig 2.6. Multiplicative model (Rose)

Inference:

- 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 show 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

Q 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

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

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 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

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

Table 2.7. Train and Test split (Rose)

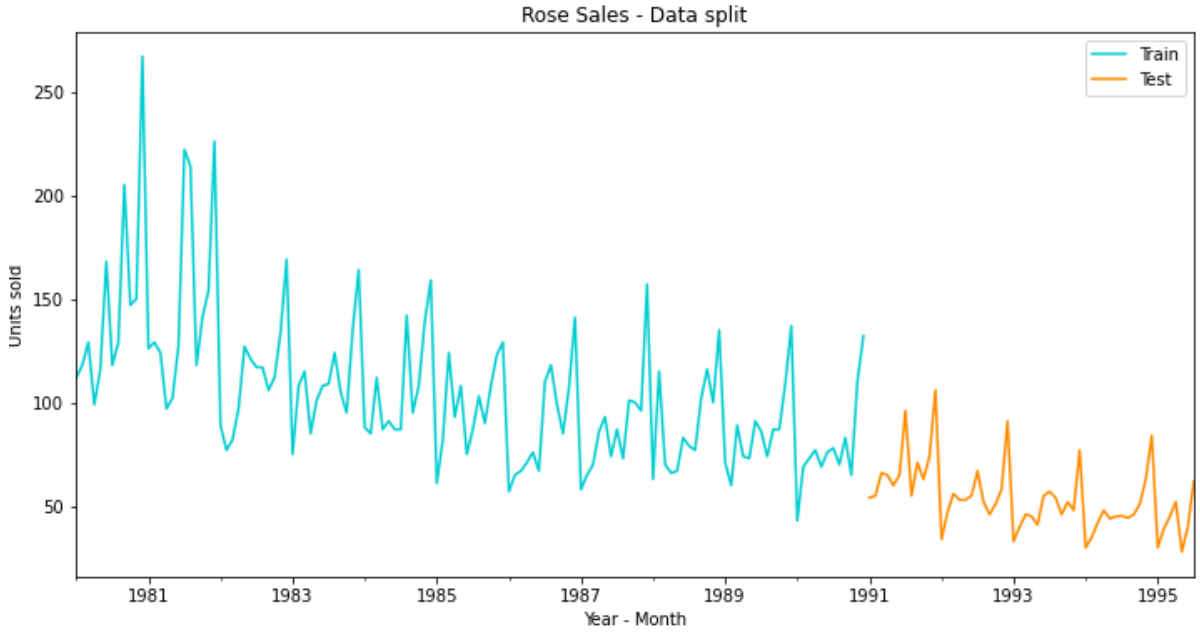


Fig 2.7. Train and Test plot (Rose)

Q 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.

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

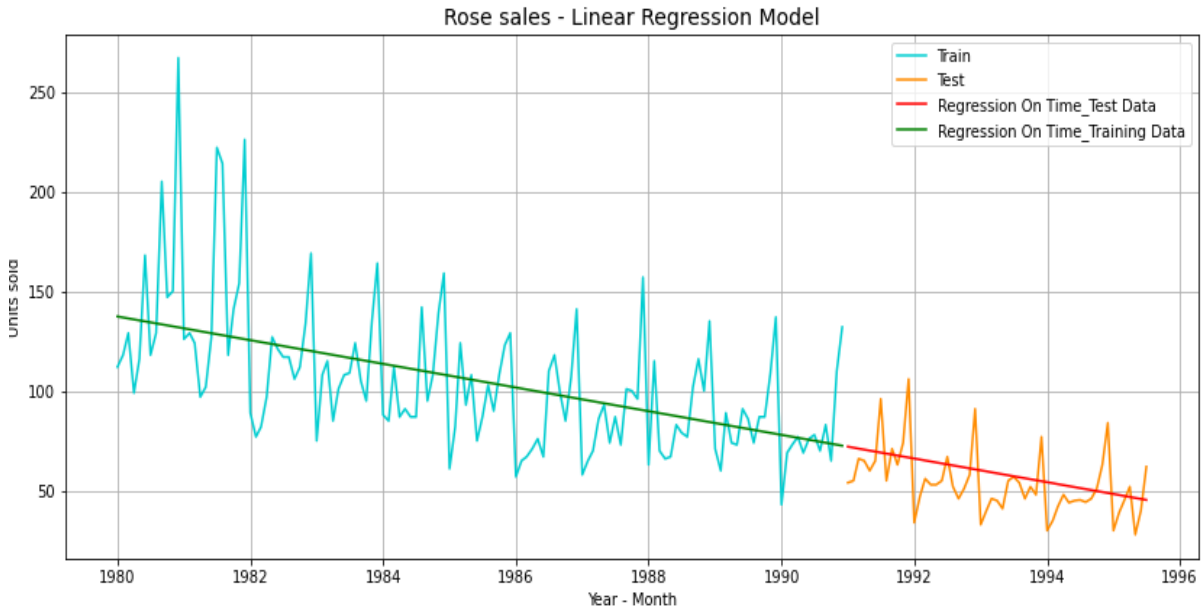


Fig 2.8. Linear Regression model (Rose)

- The linear regression plot shows a downward trend in forecast of Rose wine, consistent with the observed trend which was not visually apparent - RMSE is 15.278
- he linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series

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 does not capture the trend or seasonality for the given dataset

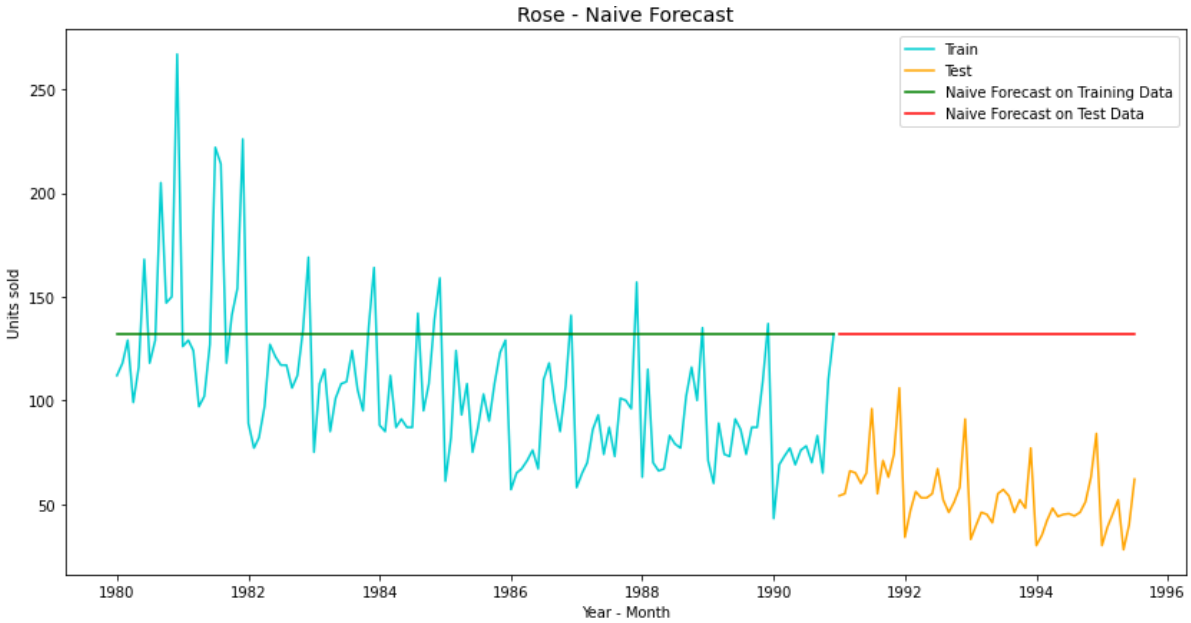


Fig 2.9. Naive model (Rose)

- 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 - RMSE is 79.745

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
- The model is not capable of either forecasting or able to capture the trend and seasonality present in the dataset - RMSE is 53.488

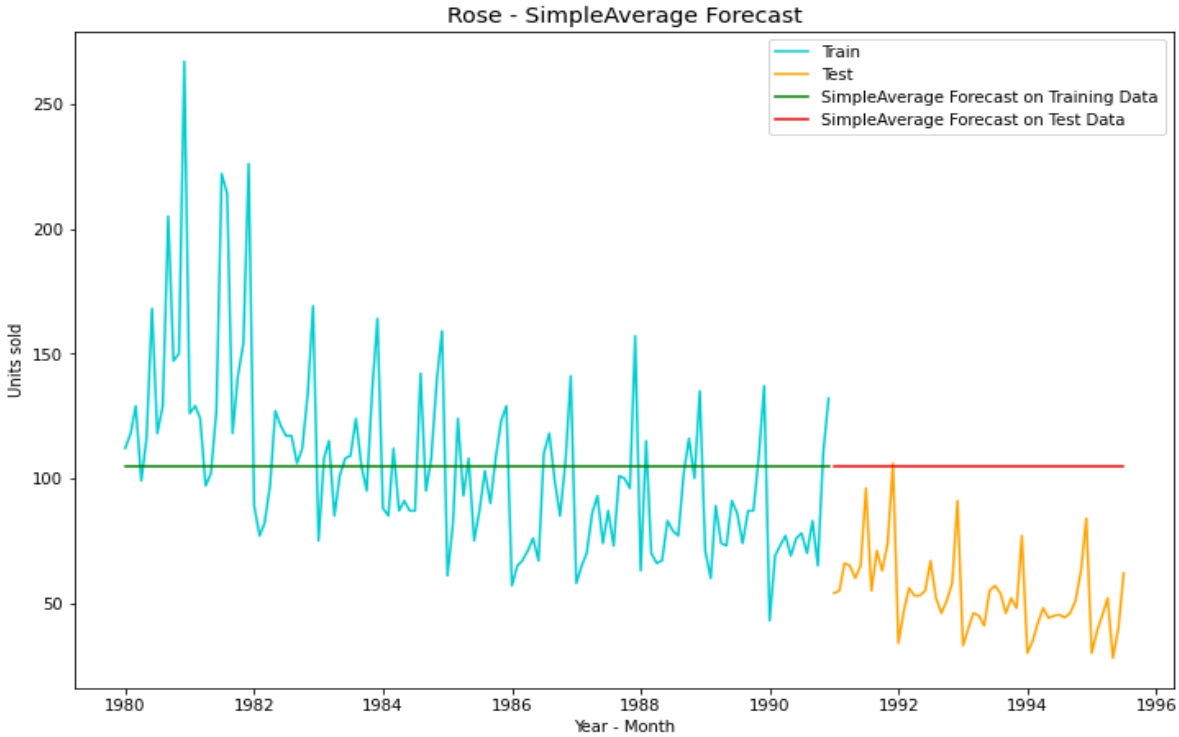


Fig 2.10. Simple average model (Rose)

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
- 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

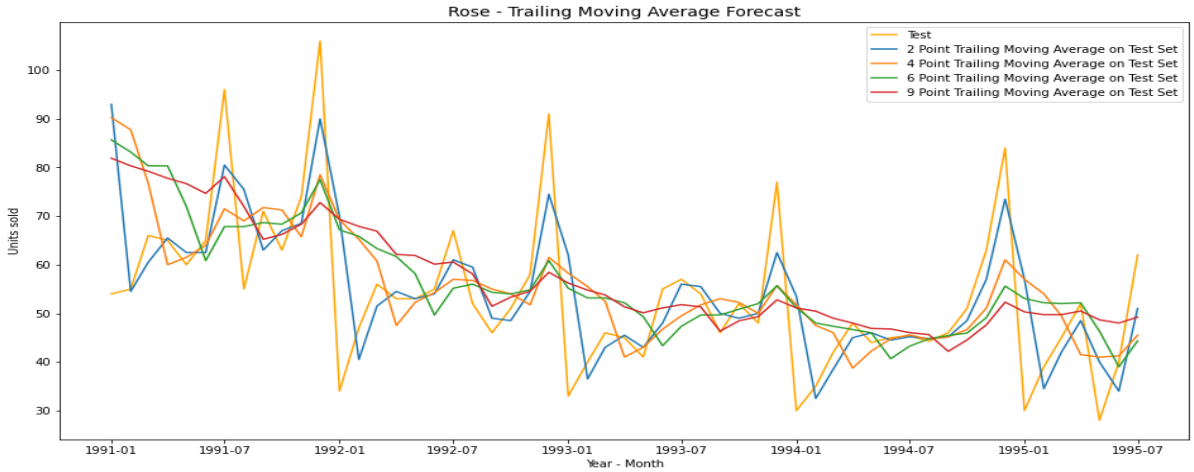


Fig 2.11. Trailing moving avg. model (RMSE)

- Model Comparison:

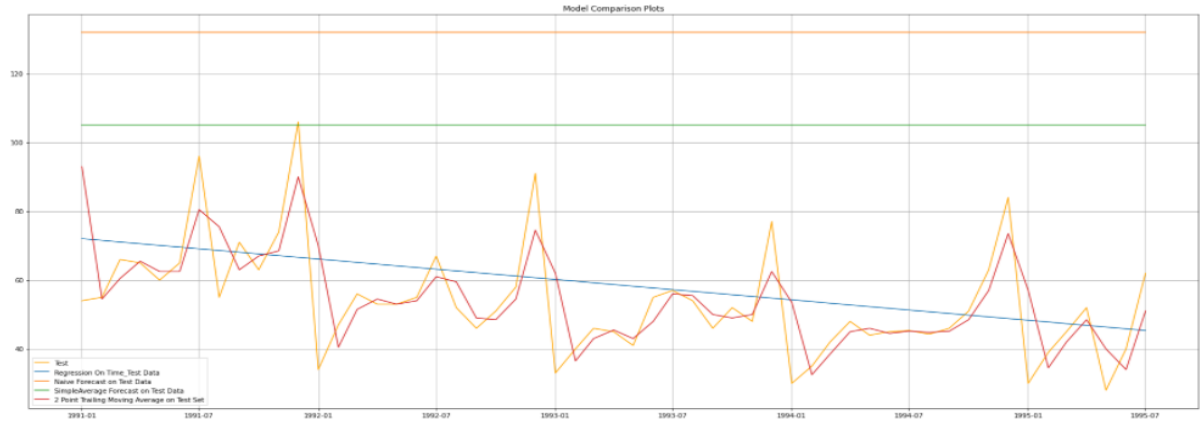


Fig 2.12. Model comparison plot (Rose)

- 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

Table 2.8. RMSE on Test (Rose)

Model 5: Simple Exponential Smoothing

- The model is run 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 follow 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

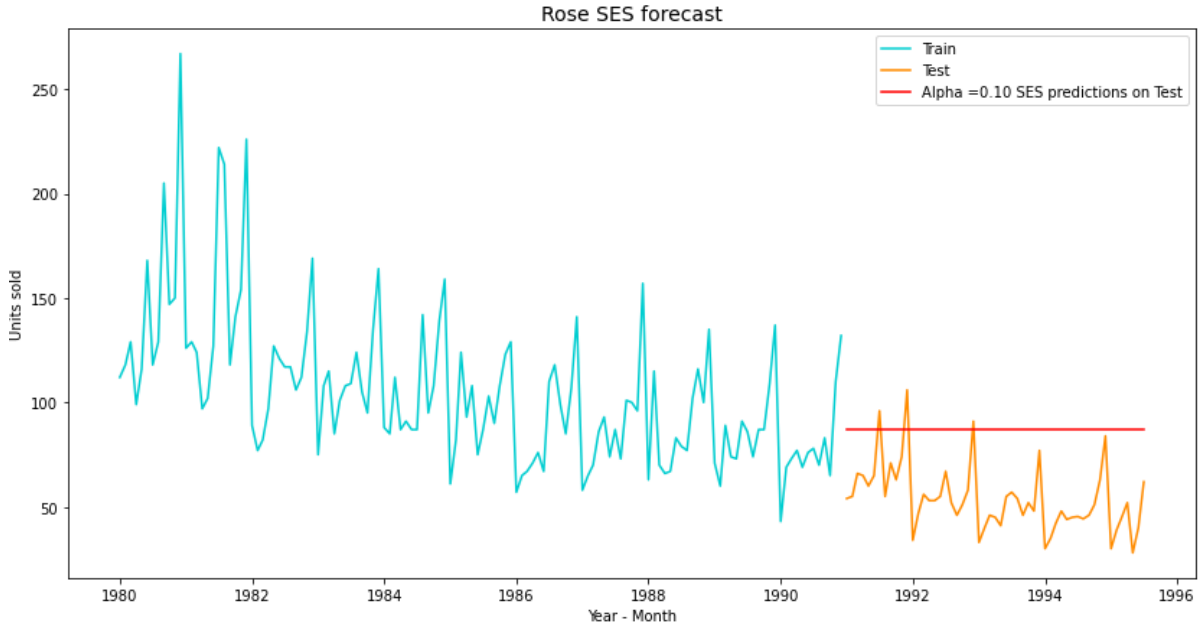


Fig 2.13. SES iterative model (Rose)

Model 6: Double Exponential Smoothing

- 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 retuned higher RMSE value compared to iterative alpha=0.1 and beta=0.1 RMSE value

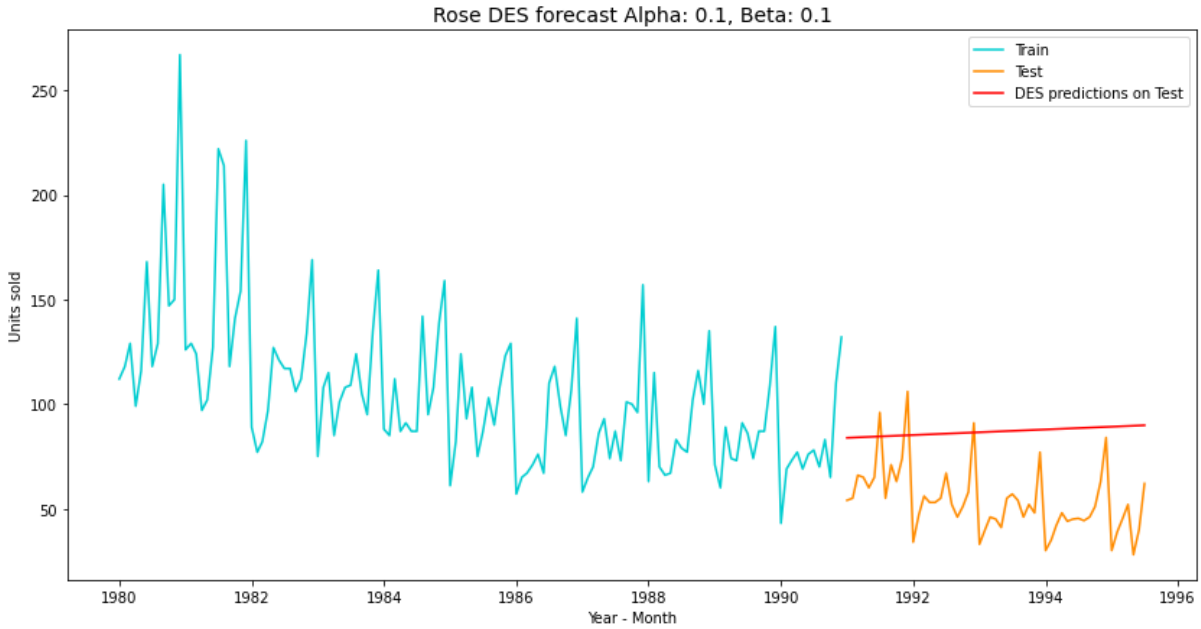


Fig 2.14. DES Iterative model (Rose)

Model 7: Triple Exponential Smoothing

- 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=0.4, beta=0.1 and gamma=0.3 RMSE value

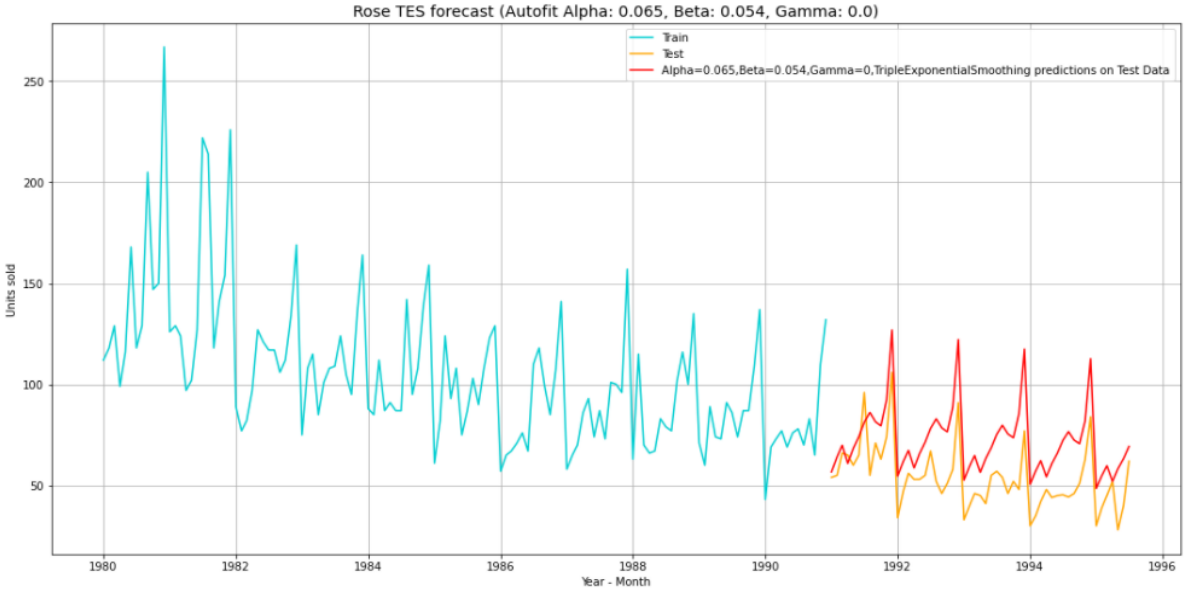


Fig 2.15. TES Autofit model (Rose)

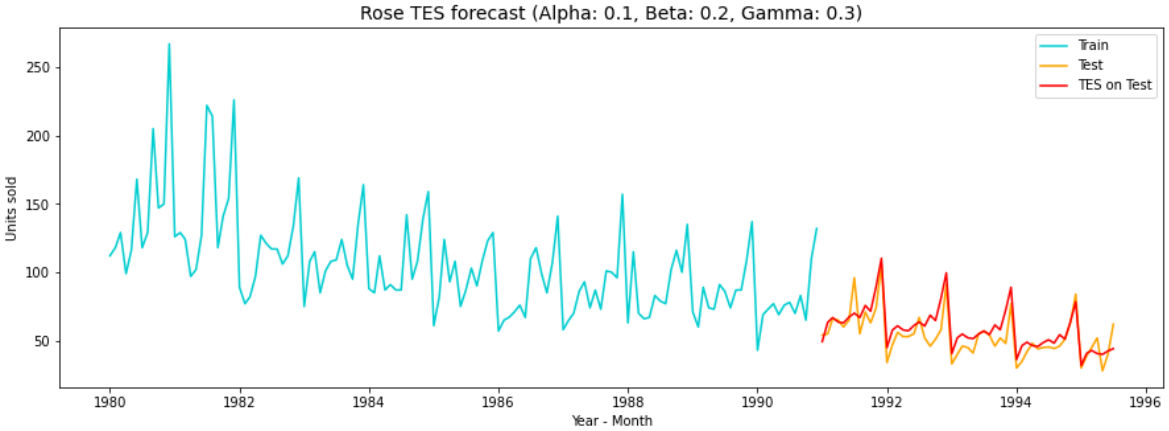


Fig 2.16. TES Iterative model (Rose)

Model Comparison:

	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
RegressionOnTime	15.278369
Alpha=0.065,Beta=0.054,gamma=0.0 TES Optimized	21.027528
Alpha=0.0987, SES Optimized	36.824480
Alpha=0.10,SES_Iterative	36.856268
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
Alpha=0.0,Beta=0.0, DES Optimized	38.310445
SimpleAverage	53.488233
NaiveModel	79.745697

Table 2.9. RMSE Values (Rose)

Fig 2.17. Forecast vs Actual (Rose)

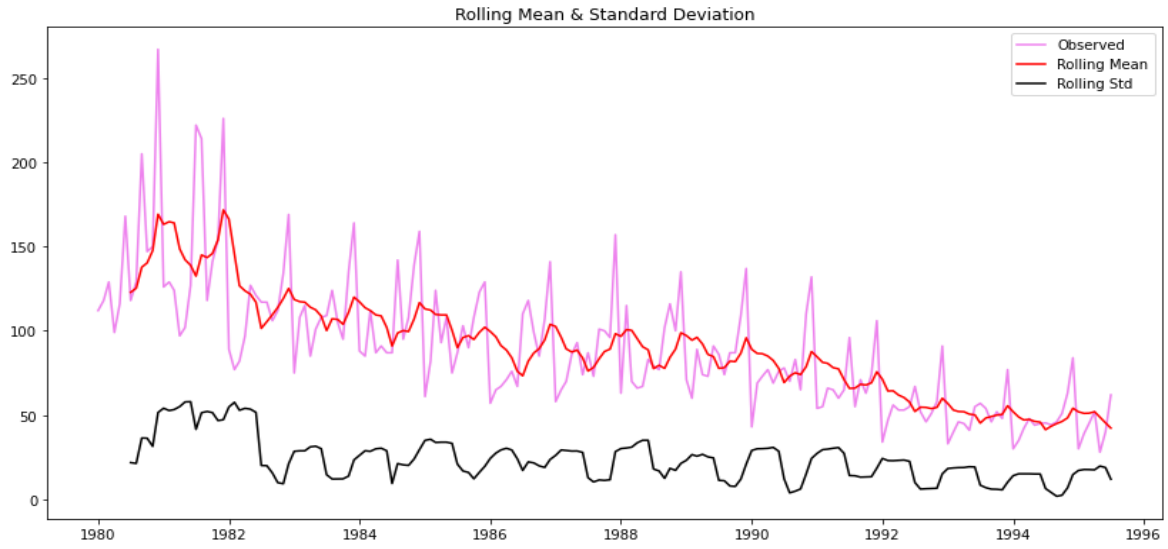
- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data
- 2-point trailing moving average model is also found to have fit well with a slight lag in test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

Q 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:

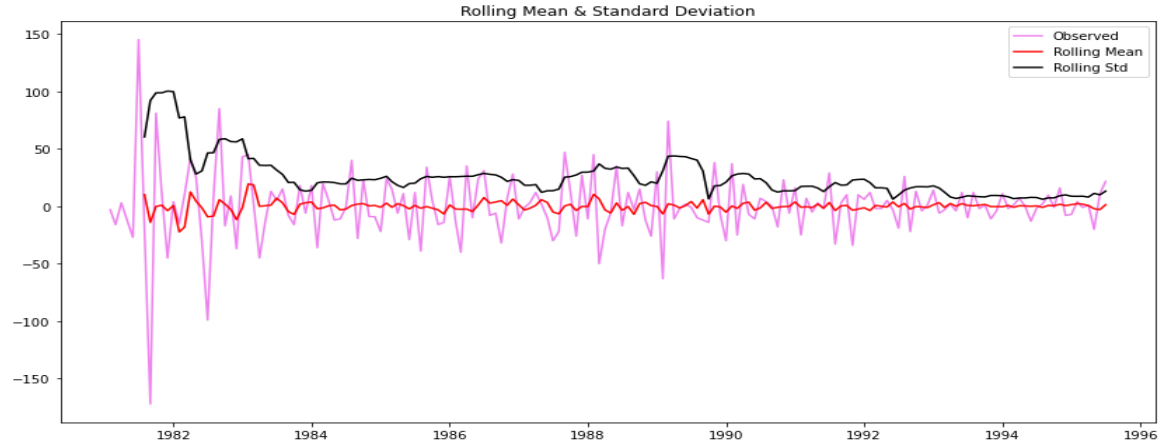
- Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determines the presence of unit root in the series to understand if the series is stationary or not
- Null Hypothesis: The series has a unit root, that is series is non-stationary
  Alternate Hypothesis: 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 0.05 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
```

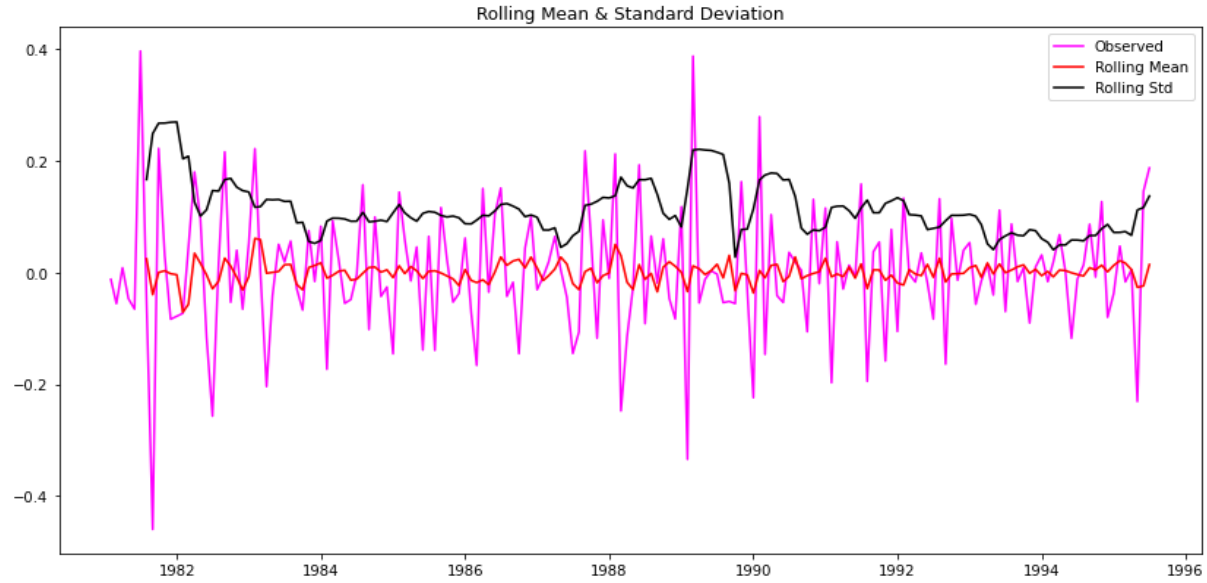
Fig 2.18. ADF test on original dataset (Rose)

- 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 mode



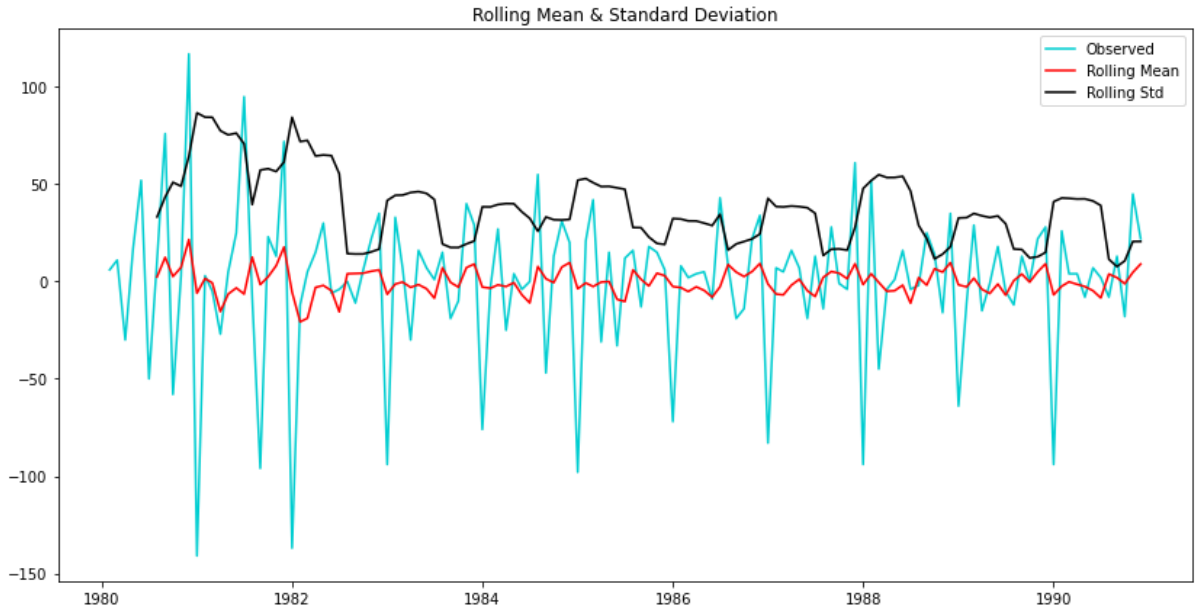

```
Results of Dickey-Fuller Test:
Test Statistic                  -4.605791
p-value                          0.000126
#Lags Used                      11.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value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

Fig 2.19. ADF test with degree 1(Rose)




```
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
```

Fig 2.20. ADF test on log series (Rose)




```
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
```

Fig 2.21. ADF test on train series with degree 1 (Rose)

Q 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

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	30.167
Date:	Sun, 10 Oct 2021	AIC	1276.835
Time:	16:58:18	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	-0.7601	0.101	-7.499	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	1.0000	+0.0000j	1.0000	0.0000
MA.2	-4.1695	+0.0000j	4.1695	0.5000

Table 2.10. Auto ARIMA (Rose)

- ARIMA model was built using iterative function and found the least AIC value =1276 at (0, 1, 2) p, d, q
- 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

Model 9: Auto SARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-380.485
Date:	Sun, 10 Oct 2021	AIC	774.969
Time:	16:58:52	BIC	792.622
Sample:	0 - 132	HQIC	782.094
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9524	0.184	-5.166	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	187.8652	45.276	4.149	0.000	99.127	276.604

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 2.11. Auto SARIMA (Rose)

- SARIMA 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 RMSE value for this Auto- SARIMA model is 16.53

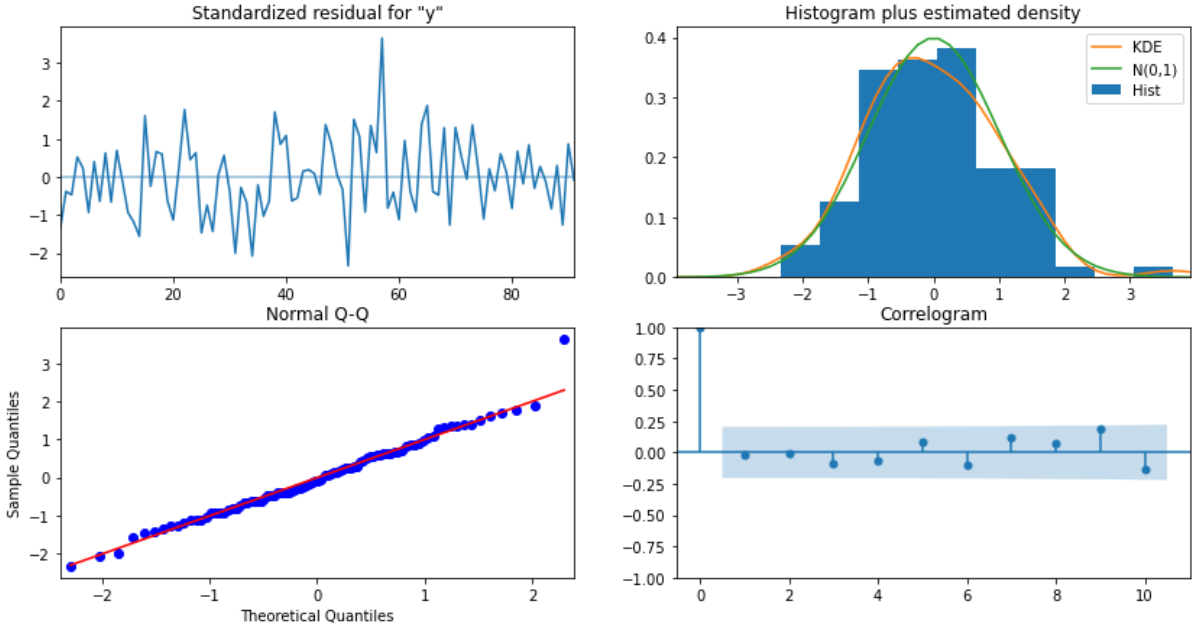


Fig 2.22. Diagnostic plot (Rose)

- 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 autocorrelation of the residuals and there are no significant lags above the confidence index

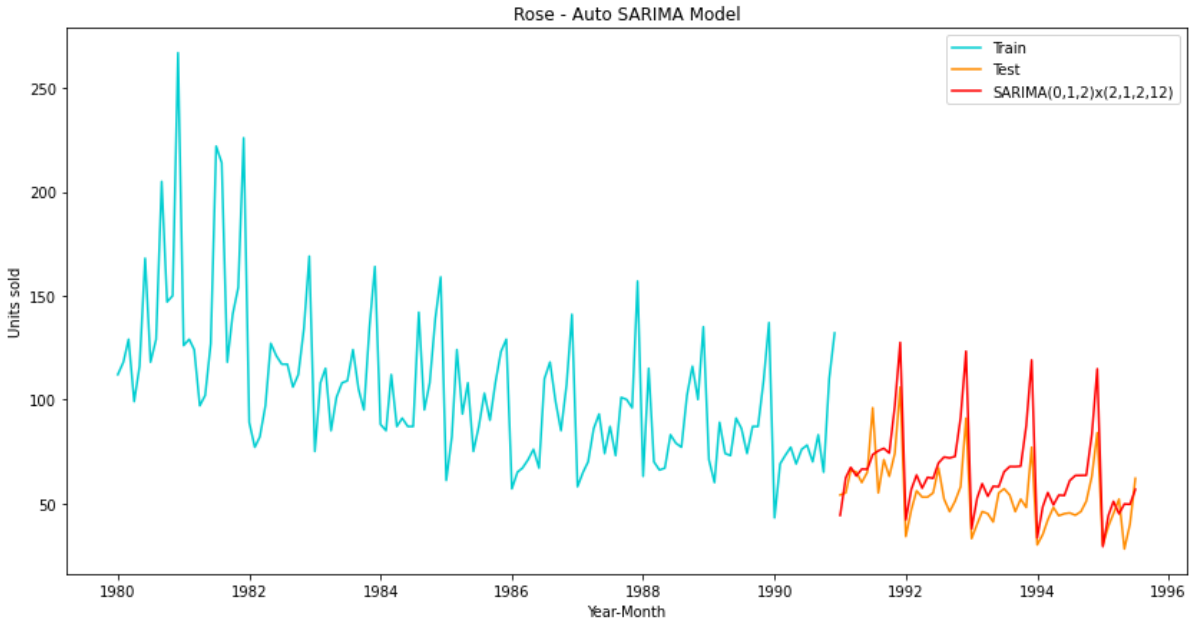


Fig 2.23. Actual vs forecast using SARIMA(Rose)

Model 10: Auto SARIMA on Log series

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Log Likelihood	127.538
Date:	Sun, 10 Oct 2021	AIC	-247.076
Time:	17:00:06	BIC	-236.028
Sample:	01-01-1980	HQIC	-242.591
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.0653	0.058	-18.390	0.000	-1.179	-0.952
ar.S.L12	0.9555	0.028	33.786	0.000	0.900	1.011
ma.S.L12	-0.8305	0.151	-5.497	0.000	-1.127	-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

Table 2.12. Auto SARIMA Log (Rose)

- 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 RMSE value for this Auto- SARIMA Log model is 17.92

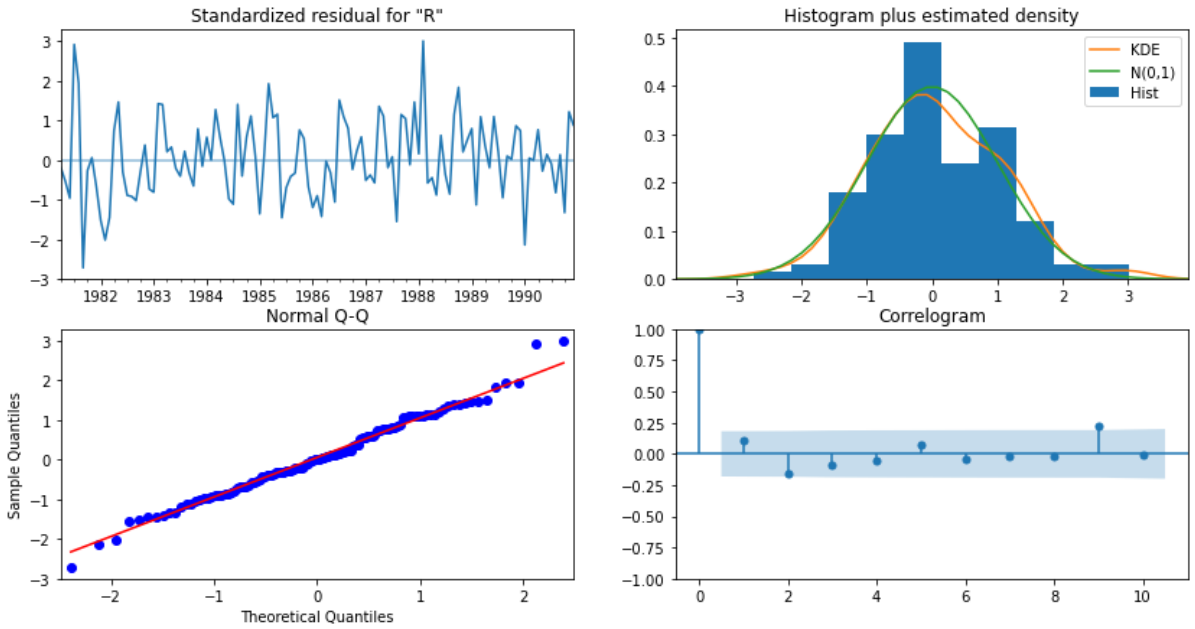


Fig 2.24. Diagnostic plot (Rose)

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

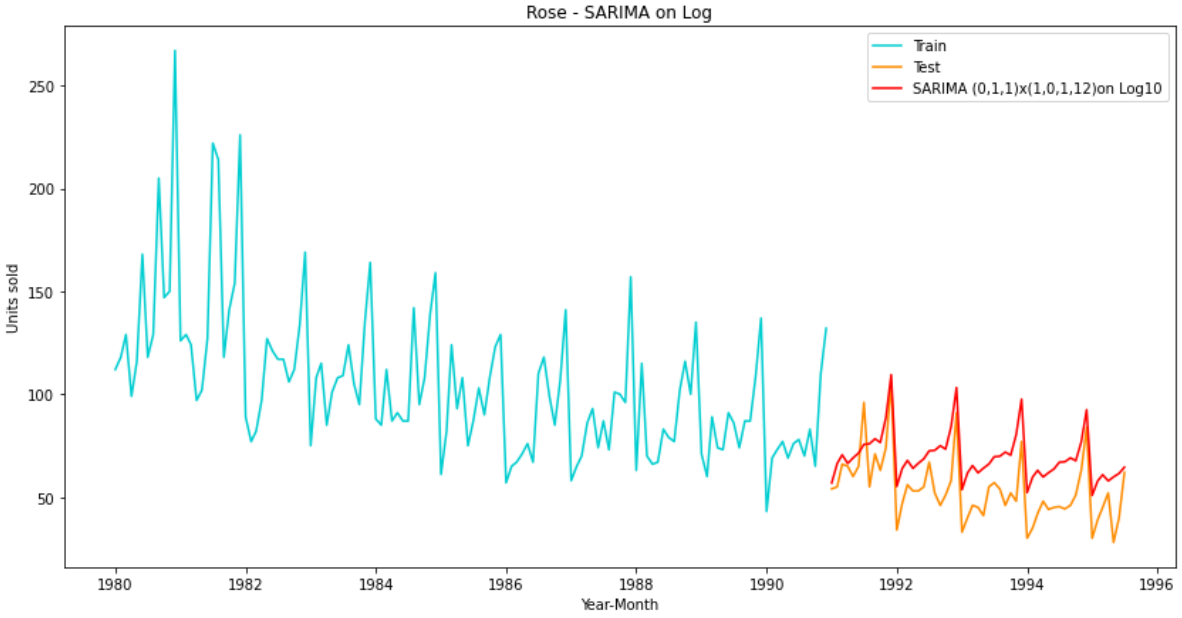


Fig 2.25. Actual vs forecast using SARIMA Log (Rose)

Q 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 11: Manual ARIMA

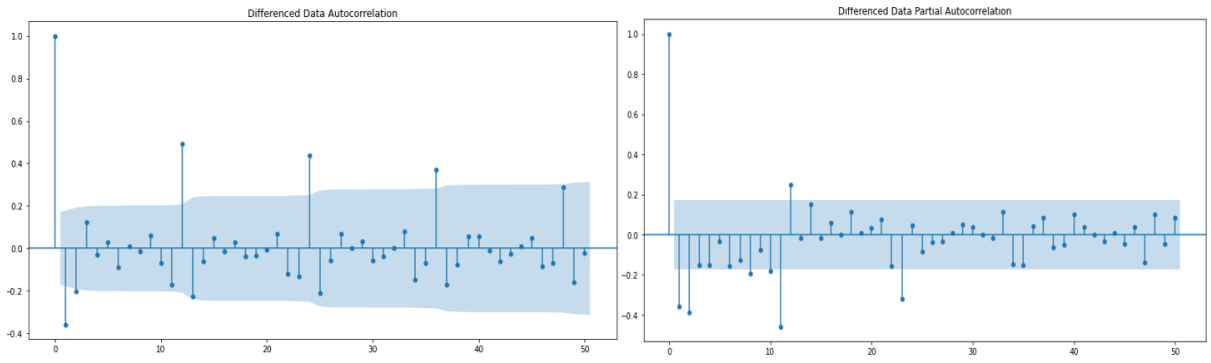


Fig 2.26. ACF and PCF plot (Rose)

- 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 2
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2

- By looking at above plots, we can say that both the PACF and ACF plot cuts-off at lag 2

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-633.649
Method:	css-mle	S.D. of innovations	29.975
Date:	Sun, 10 Oct 2021	AIC	1279.299
Time:	17:54:50	BIC	1296.550
Sample:	02-01-1980 - 12-01-1990	HQIC	1286.309

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4911	0.081	-6.076	0.000	-0.649	-0.333
ar.L1.D.Rose	-0.4383	0.218	-2.015	0.044	-0.865	-0.012
ar.L2.D.Rose	0.0269	0.109	0.246	0.806	-0.188	0.241
ma.L1.D.Rose	-0.3316	0.203	-1.633	0.102	-0.729	0.066
ma.L2.D.Rose	-0.6684	0.201	-3.332	0.001	-1.062	-0.275

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-2.0290	+0.0000j	2.0290	0.5000
AR.2	18.3387	+0.0000j	18.3387	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.4961	+0.0000j	1.4961	0.5000

Table 2.13. Manual Arima summary (Rose)

- The RMSE value of manual ARIMA model is 15.36. Since the ARIMA model do not capture the seasonality, this model does not perform well

Model 12: Manual SARIMA

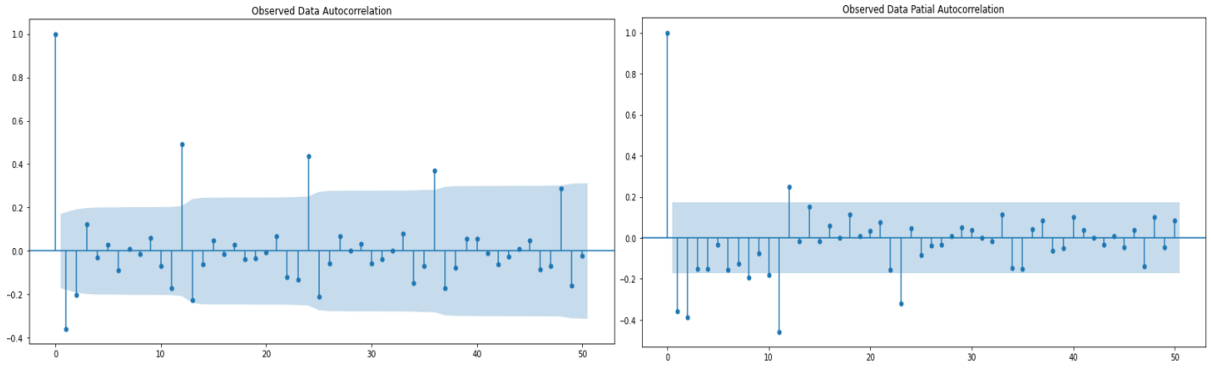


Fig 2.27. ACF and PCF plot (Rose)

- 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

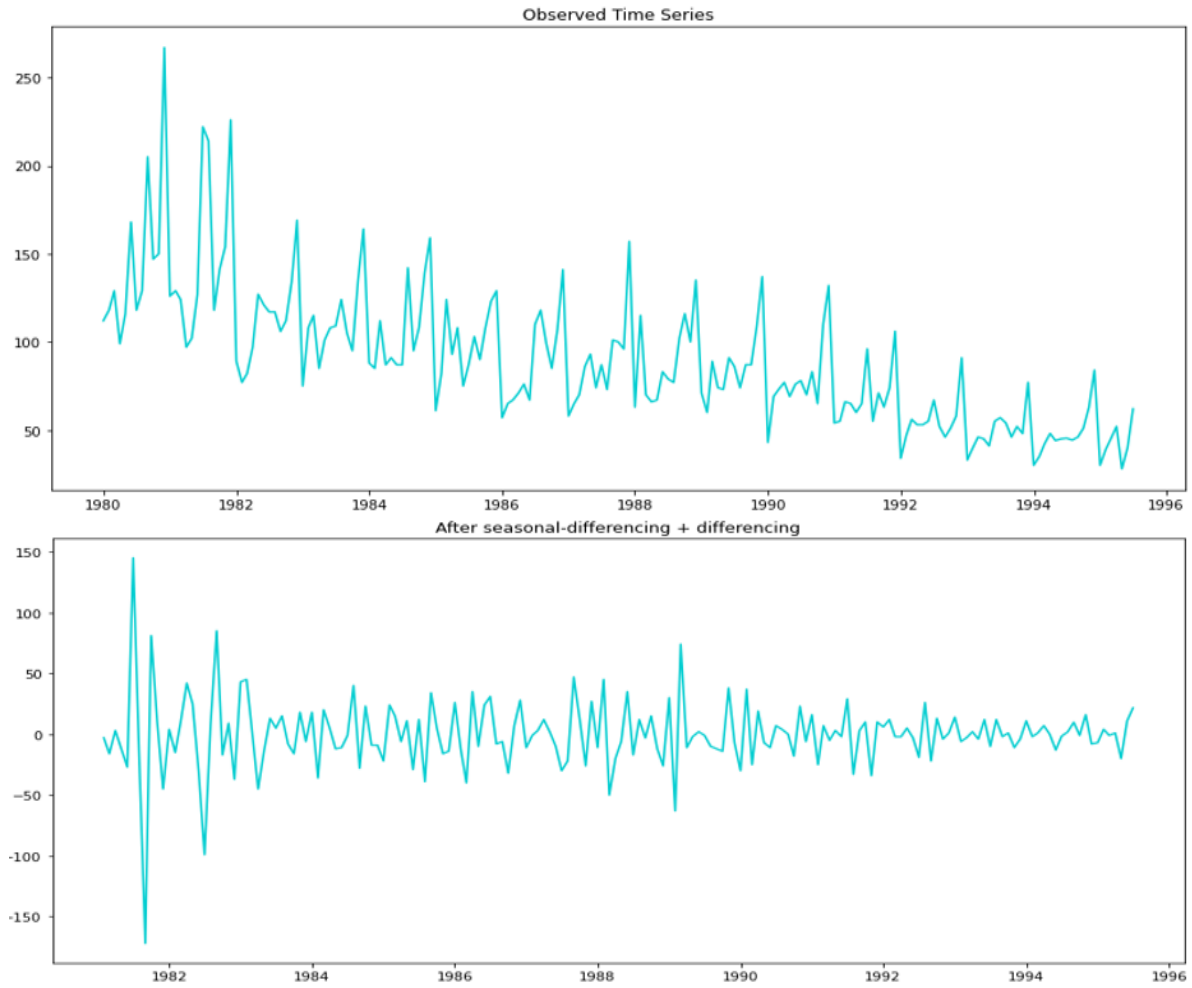
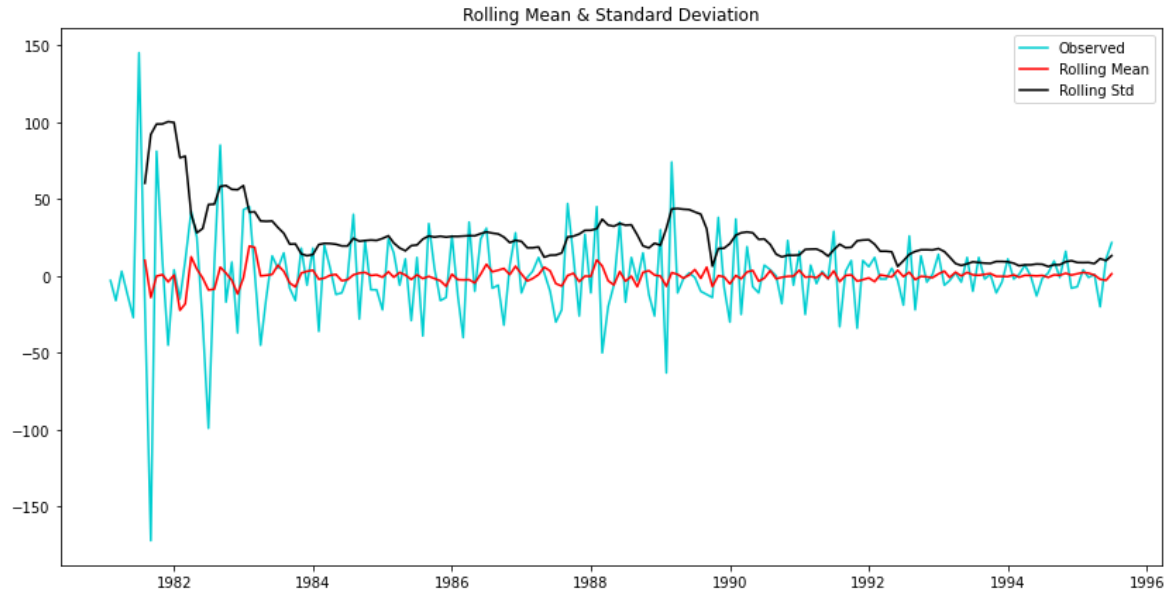


Fig 2.28. Time series plot (Rose)

- 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                  -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
```

Fig 2.29. ADF Test (Rose)

- 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)

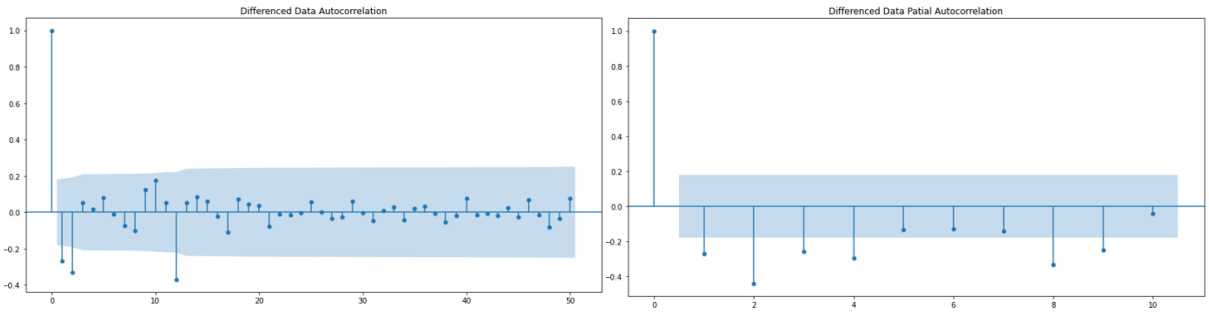


Fig 2.30. ACF and PACF plot of differenced (Rose)

- 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

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-384.369
Date:	Sun, 10 Oct 2021	AIC	786.737
Time:	17:54:52	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

Table 2.14. Manual SARIMA summary (Rose)

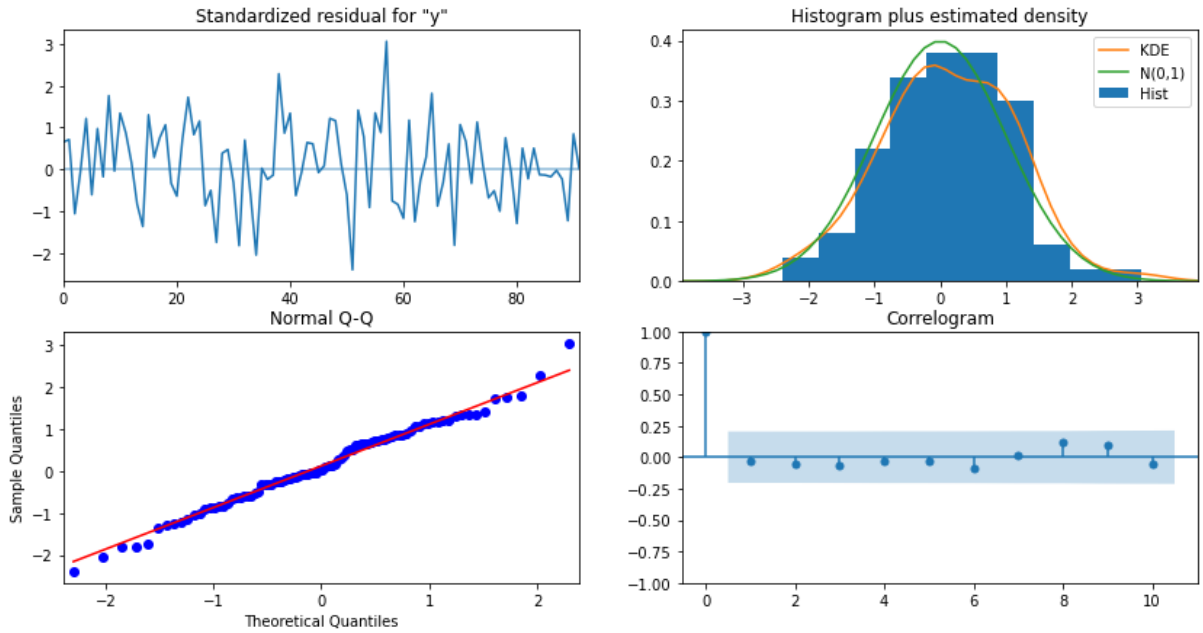


Fig 2.31. Diagnostic plot Manual SARIMA (Rose)

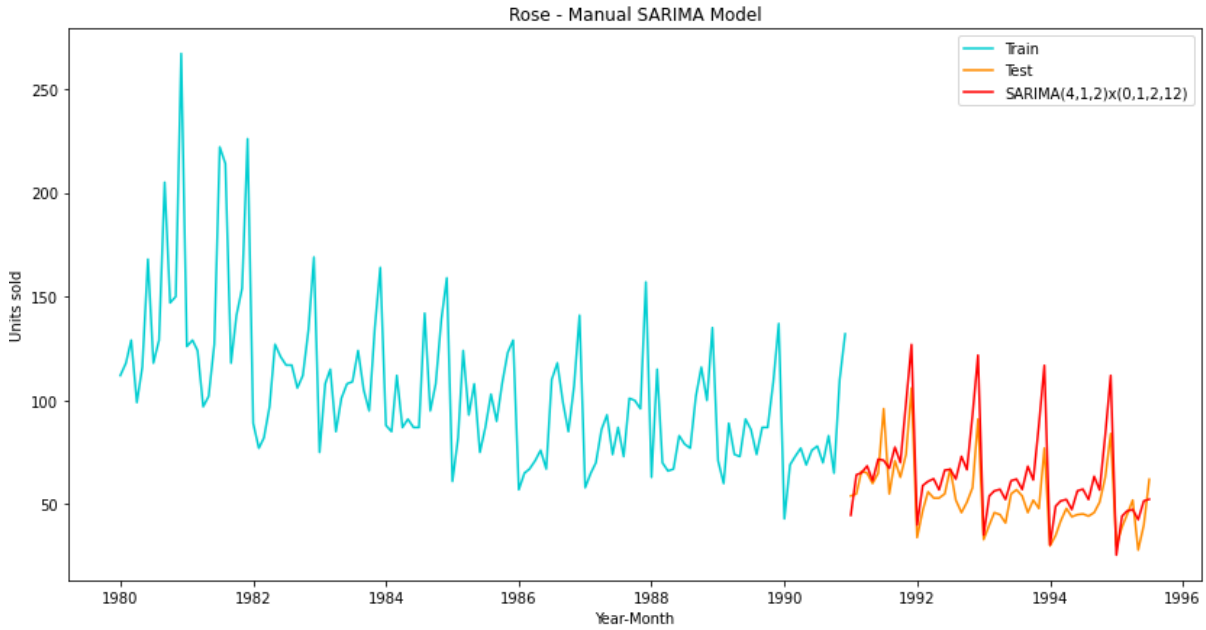


Fig 2.32. Actual vs forecast using Manual SARIMA(Rose)

Q 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
RegressionOnTime	15.278369
Manual_ARIMA(2,1,2)	15.363923
Manual_SARIMA(4, 1, 2)*(0, 1, 2, 12)	15.388806
Auto_ARIMA(0, 1, 2)	15.627457
Auto_SARIMA(0, 1, 2)*(2, 1, 2, 12)	16.527732
Auto_SARIMA_log(0, 1, 1)*(1, 0, 1, 12)	17.917600
Alpha=0.065,Beta=0.054,gamma=0.0 TES Optimized	21.027528
Alpha=0.0987, SES Optimized	36.824480
Alpha=0.10,SES_Iterative	36.856268
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
Alpha=0.0,Beta=0.0, DES Optimized	38.310445
SimpleAverage	53.488233
NaiveModel	79.745697

Table 2.15. RMSE values of all models (Rose)

- 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

Q 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

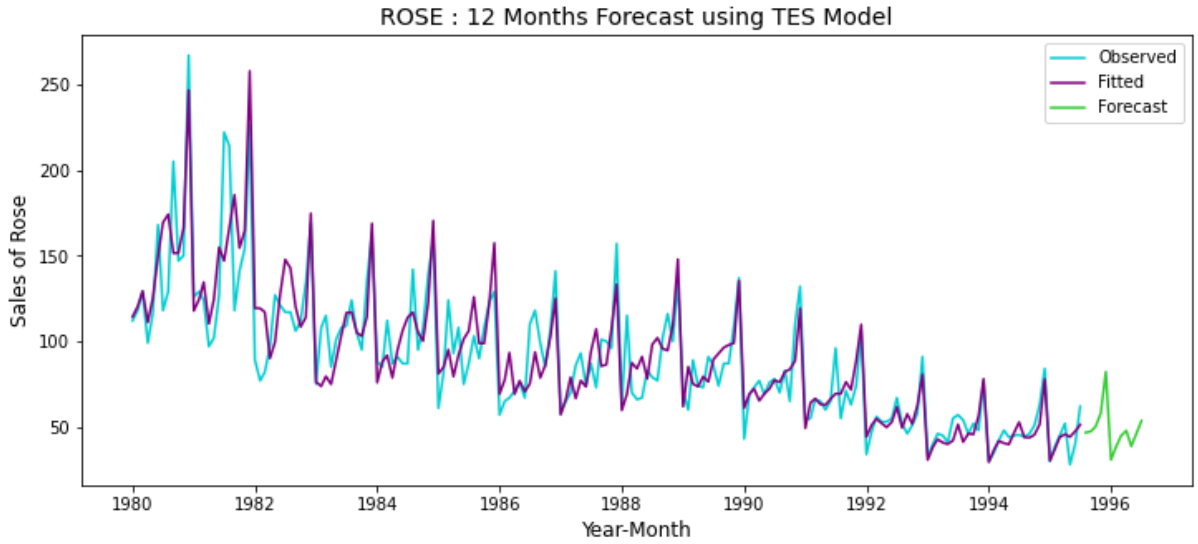


Fig 2.33. Actual vs Future 12 months forecast (Rose)

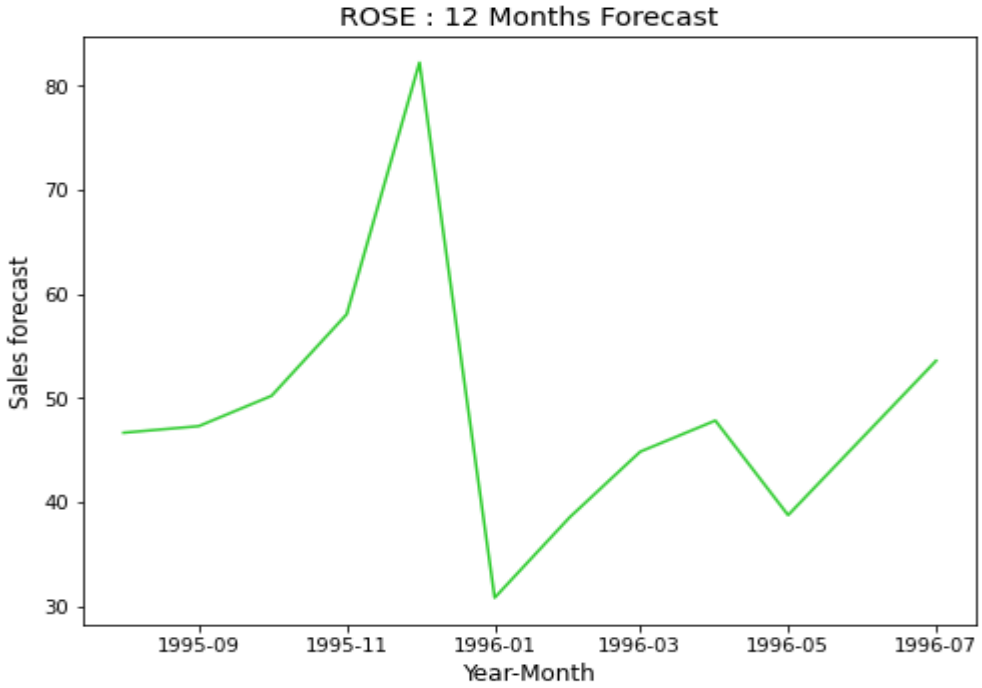


Fig 2.34. Future 12 months forecast (Rose)

Q 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.

Solution:

- 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

1995-08-01	46.645790
1995-09-01	47.277864
1995-10-01	50.192392
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

Table 2.16. Forecast 12 months (Rose)

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

Table 2.17. Forecast 12 months summary (Rose)

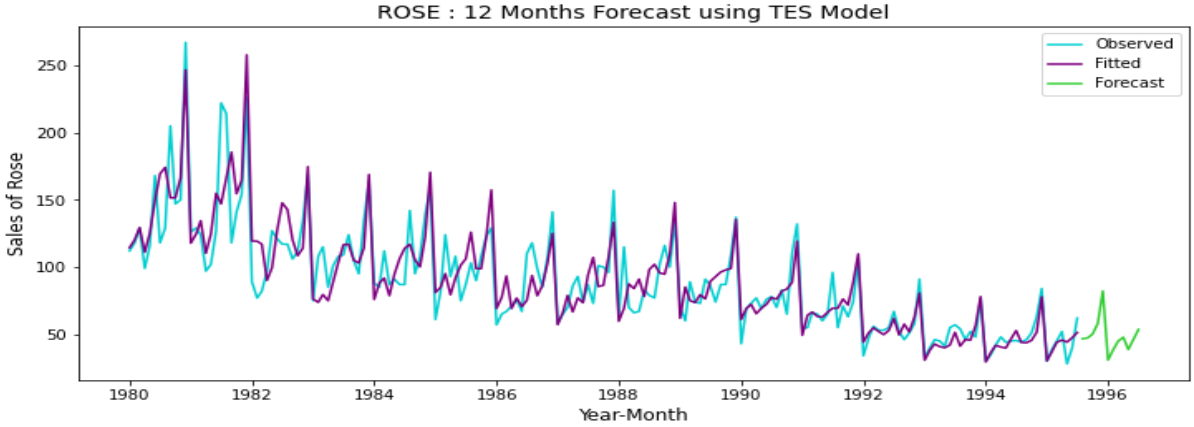


Fig 2.33. Actual vs Future 12 months forecast (Rose)

Thanks & regards,
Pavan Kumar R Naik