21 February 2022

TIME SERIES FORECASTING

Authored by: ANIMESH HALDER

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.

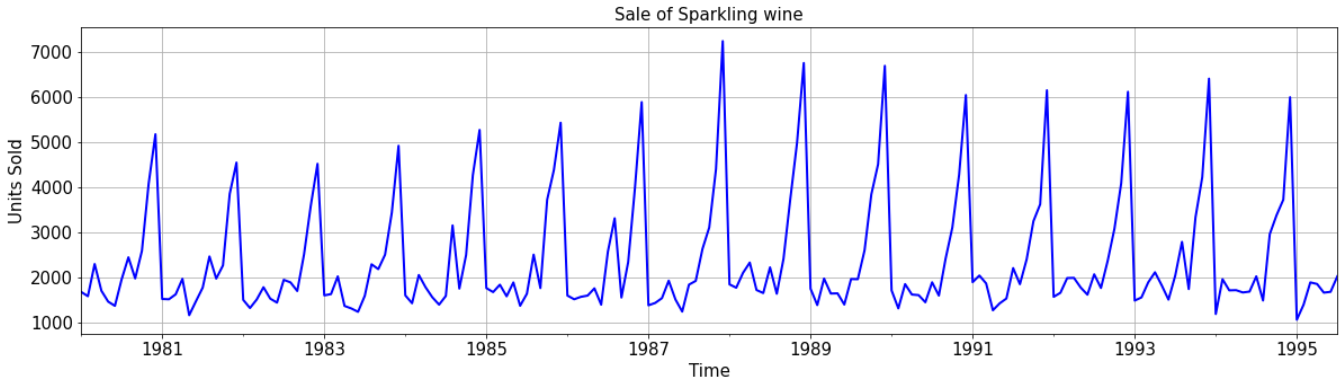
1. Data at First Glance

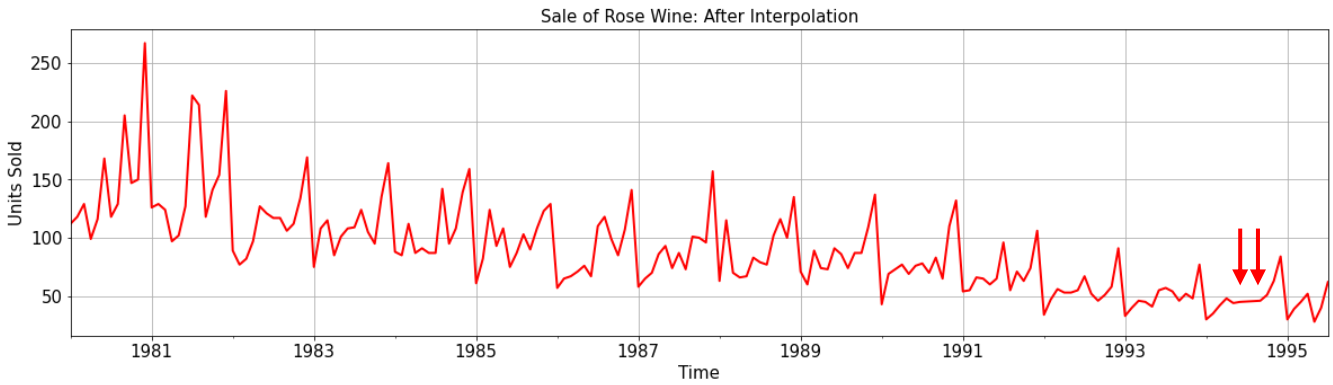
- From January, 1980 to July, 1995, data is presented on the monthly sales of two types of wines, including Sparkling and Rose.
- The given data files are read as is and a date-range has been applied on the data as index.
- In order to compare and forecast the time-series components of the two types of wines together, the given datasets have been combined to a single data frame.
- Interpolation (linear method) is used to impute missing values in the Rose time-series for two months in 1994.
- Rose data after interpolation for year 1994 is given below as well as the plot.
- There is significant seasonality in both datasets. During the time period, Rose's sale shows an obvious downward trend, while Sparkling's does not show any consistent downward trend, but has upward and downward slopes.
- Customers have consistently favored Sparkling wine over the years, but Rose has fallen out-of-favor over time.

```
INITIAL THREE ROWS
            Sparkling   Rose
YearMonth
1980-01-31       1686  112.0
1980-02-29       1591  118.0
1980-03-31       2304  129.0
====================
LAST THREE ROWS
            Sparkling  Rose
YearMonth
1995-05-31       1670  28.0
1995-06-30       1688  40.0
1995-07-31       2031  62.0
```

```
4  df.Rose['1994']

YearMonth
1994-01-31    30.000000
1994-02-28    35.000000
1994-03-31    42.000000
1994-04-30    48.000000
1994-05-31    44.000000
1994-06-30    45.000000
1994-07-31    45.336957
1994-08-31    45.673913
1994-09-30    46.000000
1994-10-31    51.000000
1994-11-30    63.000000
1994-12-31    84.000000
```

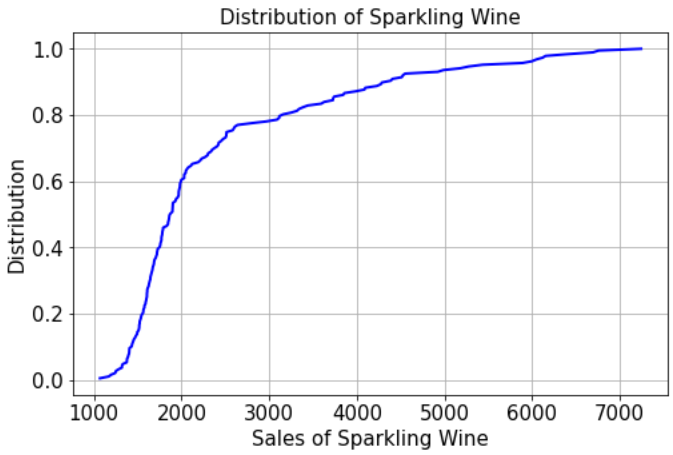




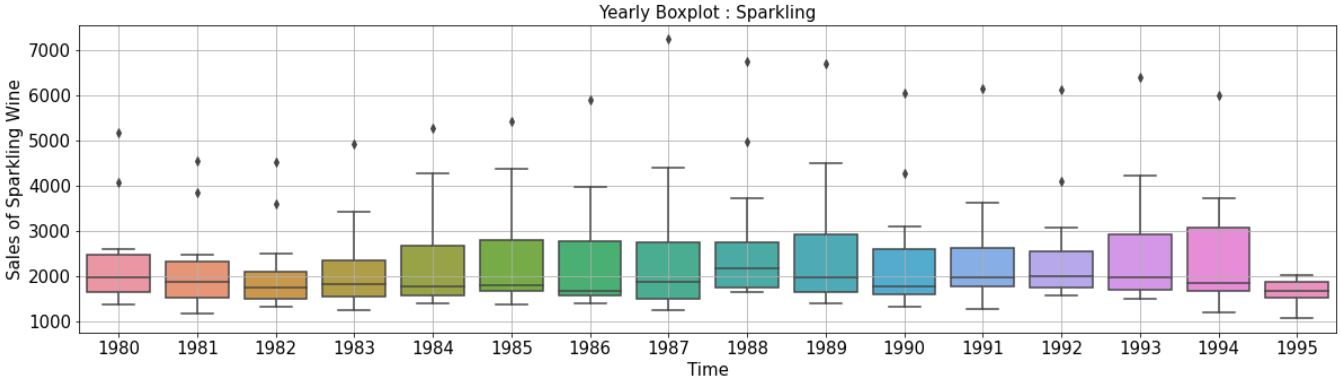
2. Exploratory Data Analysis

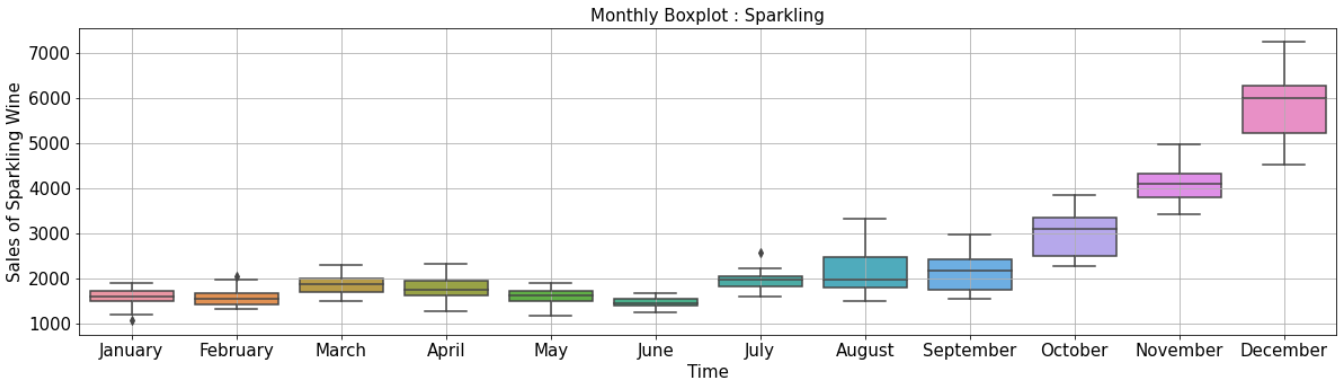
2.1. Sparkling Wine

- On average, 2402 units of sparkling wine were sold each month on the given period of time, as described in the descriptive summary of the data. The majority of month's sales ranged from 1605 to 2549 units. The highest monthly sale was 7242 units.
- Empirical CDF plots show that Sparkling wine was sold at least 3000 units in 80% of months
- Based on the yearly-boxplot, Sparkling's average sale has been near or a bit below 2000 units over the period.



- The seasonal sales during the holiday season are most likely the outliers in the yearly-boxplot.
- There is a distinct seasonal pattern in the monthly-box plot during the festive months of October, November and December, which peaks in December. From June through August, the sale declines significantly.

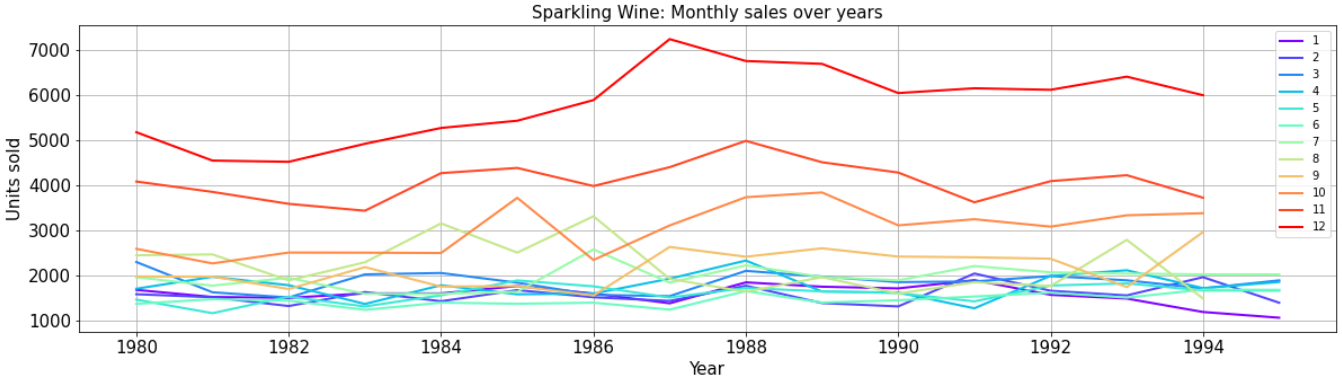




- The monthly plot for Sparkling shows mean and variation of units sold each month over the years. Sale in seasonal months shows a higher variation than in the lean months.
- Sale in December with a mean few points below 6000, varies from 7400 to 4500 units over the years. Whereas sale in November varies 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.

Statistical Description

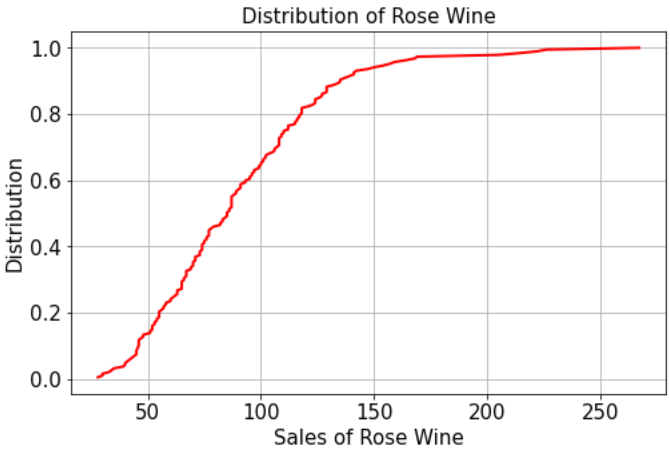
	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.111540	1070.0	1605.0	1874.0	2549.0	7242.0
Rose	187.0	89.914497	39.238259	28.0	62.5	85.0	111.0	267.0



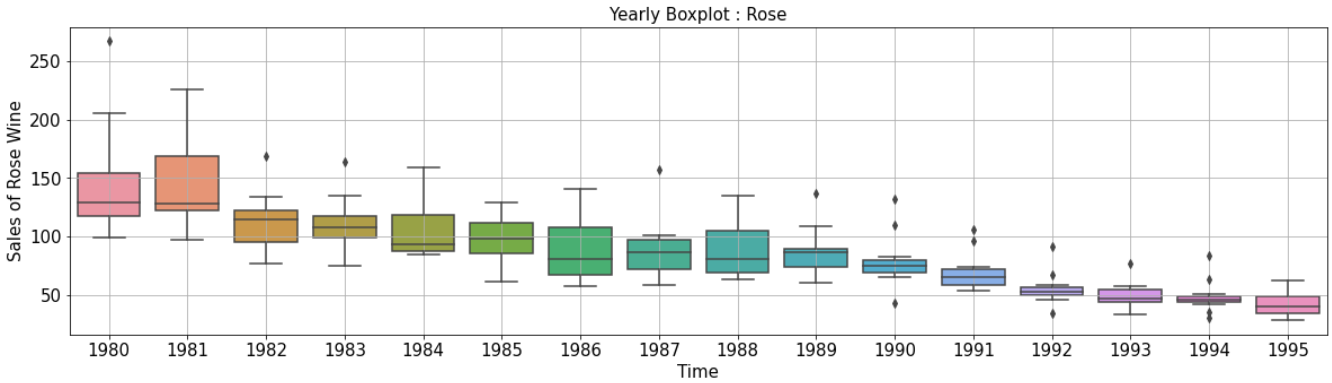
- The plot of monthly sale over the years also shows the seasonality component of the time-series, with October November and December selling exponentially higher volumes.
- The highest volume of Sparkling wines were sold in December, 1987 and the least of December sale was in 1981. Post 1987 December 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.

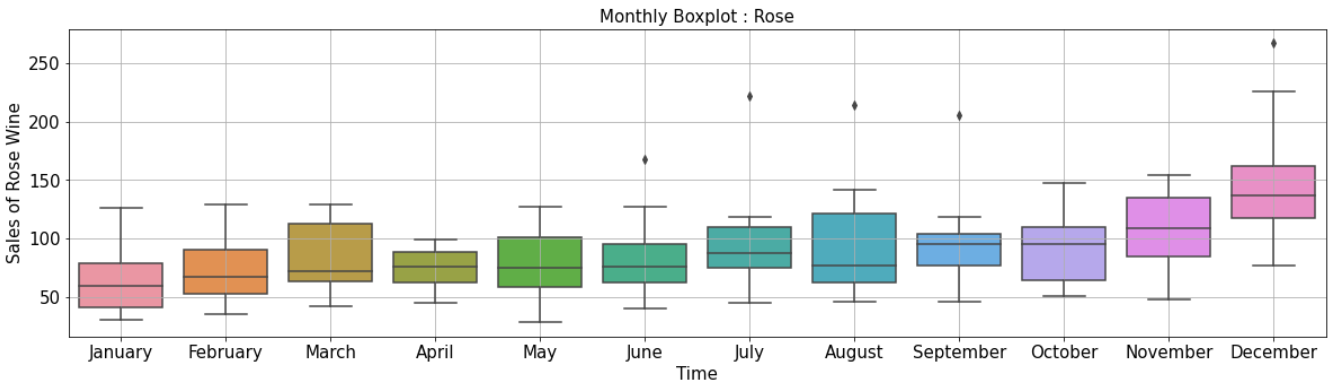
2.2. Rose Wine

- 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 Empirical CDF plot shows that, in 80% of months, at least 120 units of Rose wine were sold
- 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.





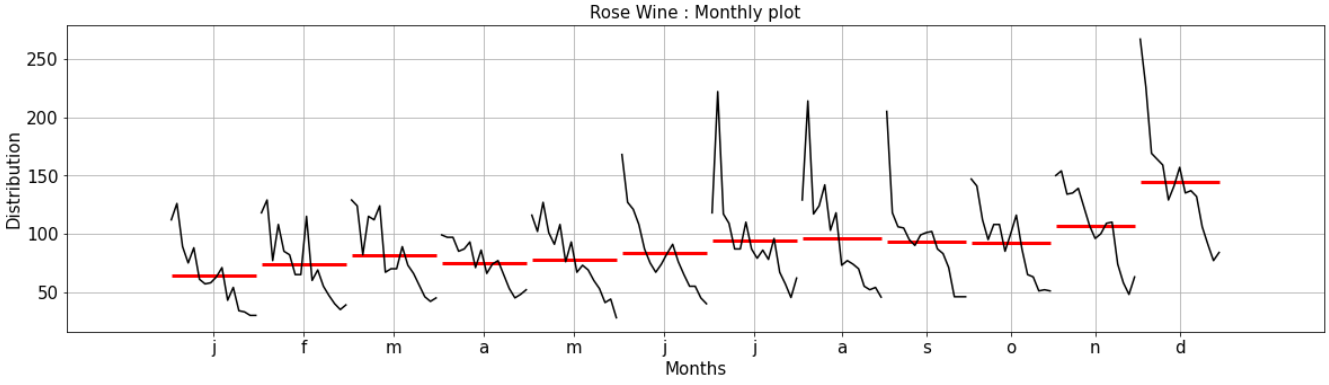
- Average sale in December is around 140 units, November is around 110 units and October is around 90 units.
- The monthly plot for Rose shows mean and variation of units sold each month over the years. Sale in months such as July, August, September and December shows a higher variation than the rest.

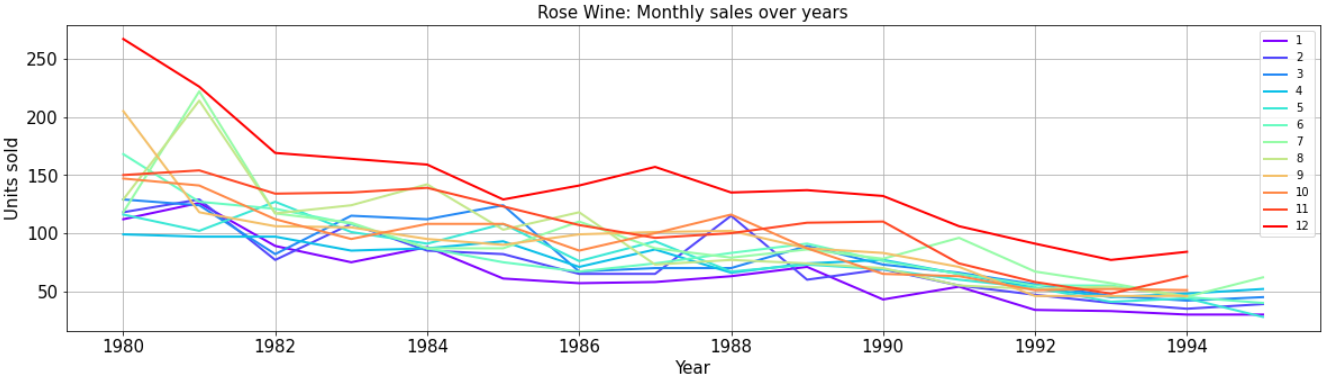
- Sale in December with a mean few points below 100, varies from 75 to 270 units over the years. Whereas the average sale is less than or closer to 100 units (above 50) for the rest of the year.

Statistical Description

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.111540	1070.0	1605.0	1874.0	2549.0	7242.0
Rose	187.0	89.914497	39.238259	28.0	62.5	85.0	111.0	267.0

- 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.

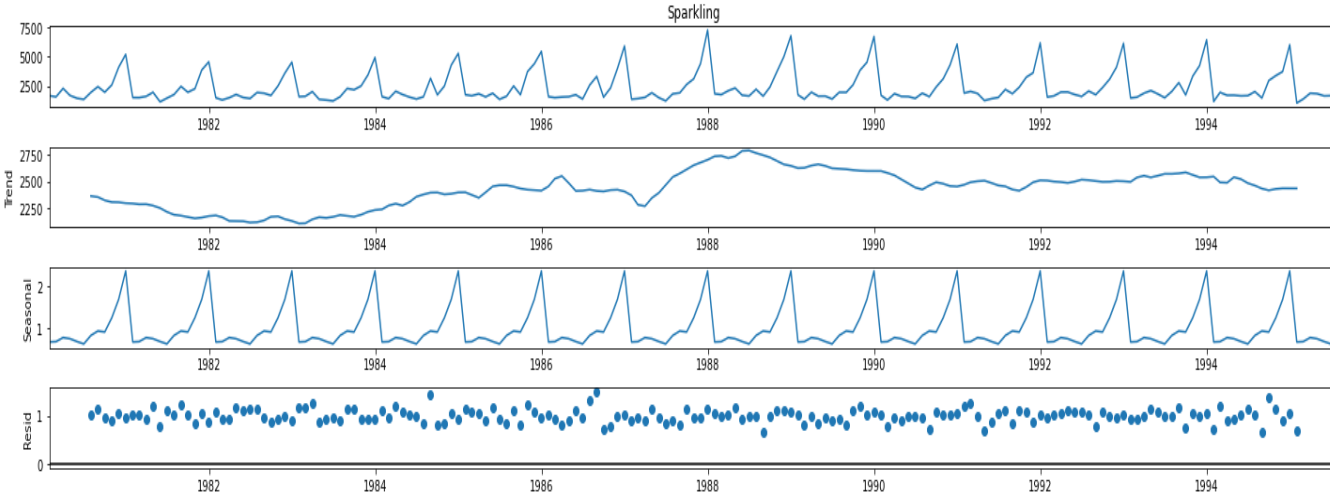
3. Time Series Decomposition

3.1. Sparkling Wine

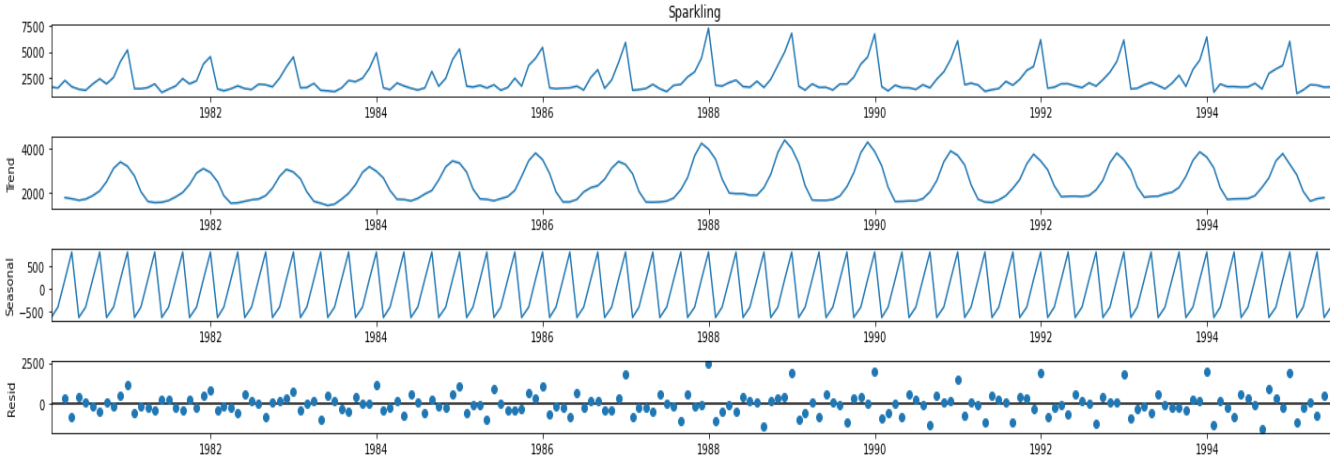
The decomposition plots of Sparkling wine sales is given here:

- 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 slope 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%.

Sparkling Wine : Multiplicative Model



Sparkling Wine : Additive Model



- 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

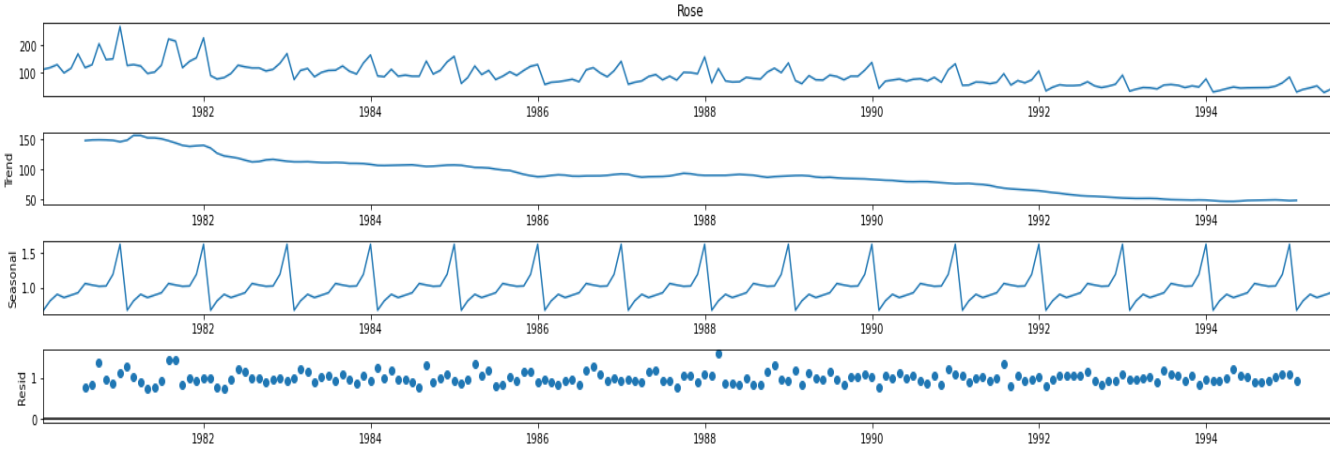
- The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 10%
- If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series.

3.1. Rose Wine

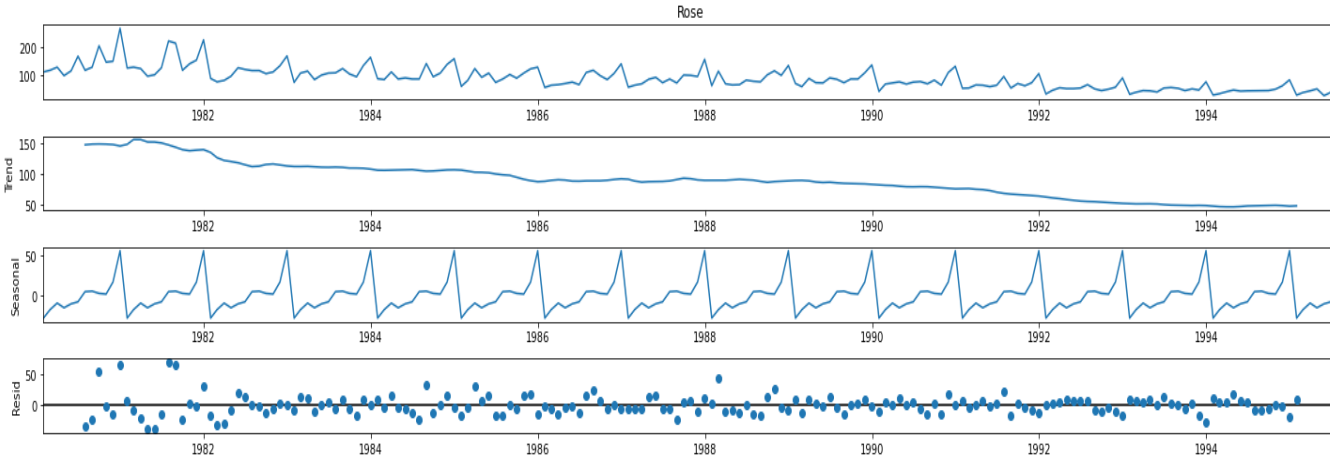
The decomposition plots of Sparkling wine sales is given here:

- The observed plot of the decomposition diagrams 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.

Sparkling Wine : Multiplicative Model



Sparkling Wine : Additive Model



- Seasonal components are quite visible and consistent in both the observed and seasonal charts of the diagrams. The additive chart shows variance in seasonality from -20 to 50 units and the multiplicative model shows variance of 16%.

- The residuals 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 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.
- The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 15%.
- As the seasonality peaks are consistently reducing its altitude in consistent with trend, the series can be treated as multiplicative in model building.

4. Splitting of the data into Training and Test

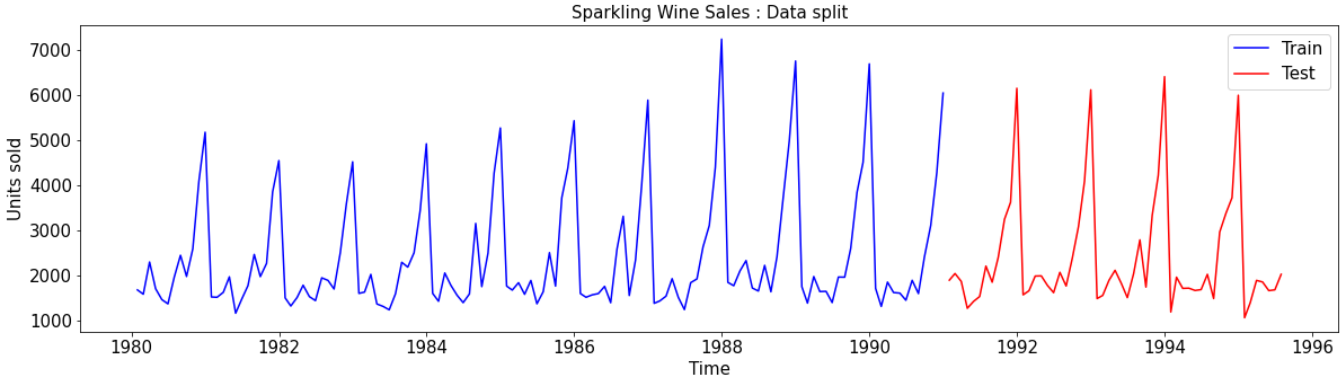
- The train and test datasets are created with year 1991 as starting year for test data, using **index.year** property of time series index.
- The plots of Sparkling and Rose time-series as train and test are given here:

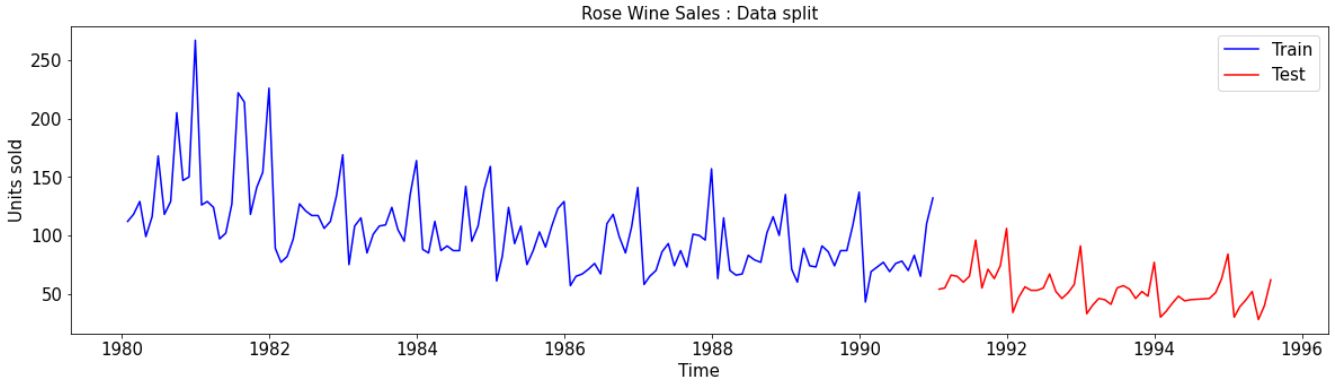
```
train=df[df.index.year < 1991]
test=df[df.index.year >= 1991]
```

Train and Test Dataset

```
================================
First Five Rows of Training Data
--------------------------------
            Sparkling   Rose
YearMonth
1980-01-31       1686  112.0
1980-02-29       1591  118.0
1980-03-31       2304  129.0
1980-04-30       1712   99.0
1980-05-31       1471  116.0
--------------------------------
Last Five Rows of Training Data
--------------------------------
            Sparkling   Rose
YearMonth
1990-08-31       1605   70.0
1990-09-30       2424   83.0
1990-10-31       3116   65.0
1990-11-30       4286  110.0
1990-12-31       6047  132.0
================================
```

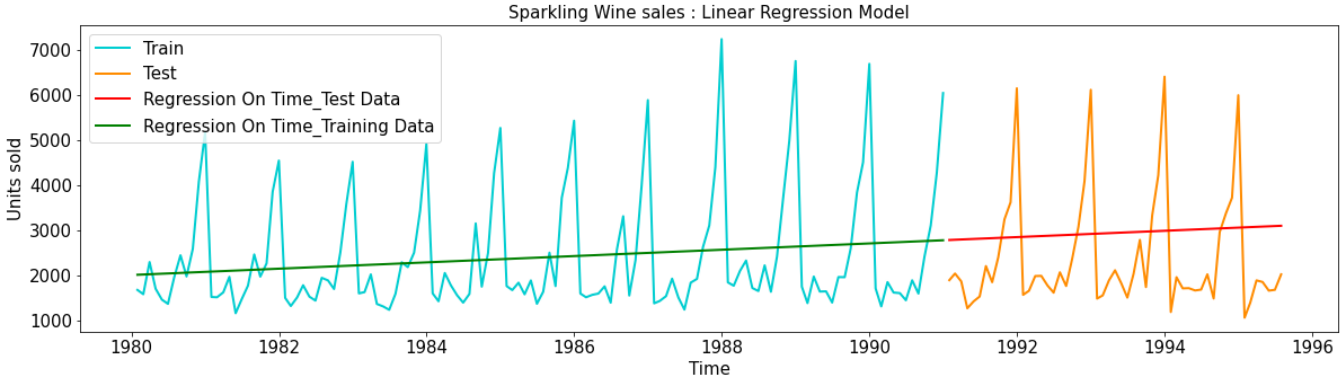
```
================================
First Five Rows of Test Data
--------------------------------
            Sparkling  Rose
YearMonth
1991-01-31       1902  54.0
1991-02-28       2049  55.0
1991-03-31       1874  66.0
1991-04-30       1279  65.0
1991-05-31       1432  60.0
--------------------------------
Last Five Rows of Test Data
--------------------------------
            Sparkling  Rose
YearMonth
1995-03-31       1897  45.0
1995-04-30       1862  52.0
1995-05-31       1670  28.0
1995-06-30       1688  40.0
1995-07-31       2031  62.0
================================
```



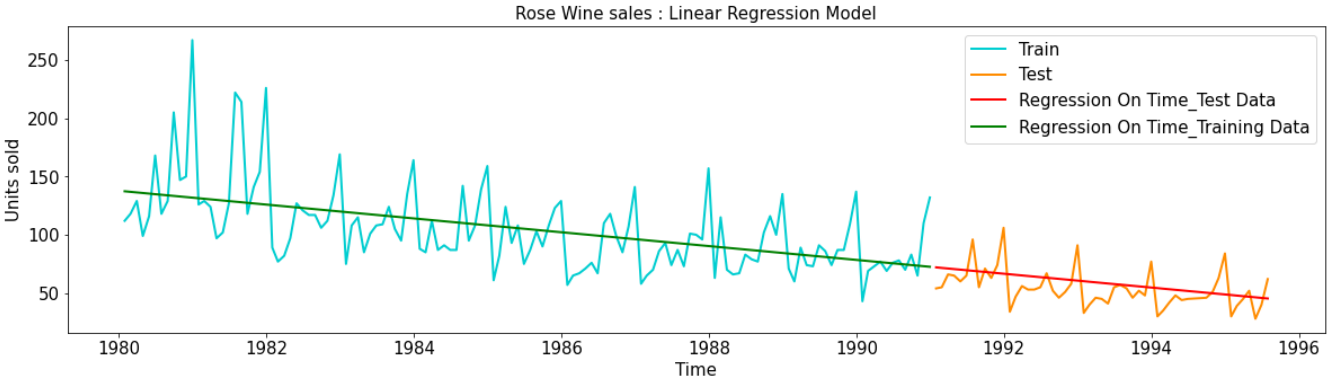


5. Linear Regression on time

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



	RMSE	MAPE
Train	1279.32	40.05
Test	1389.16	50.15

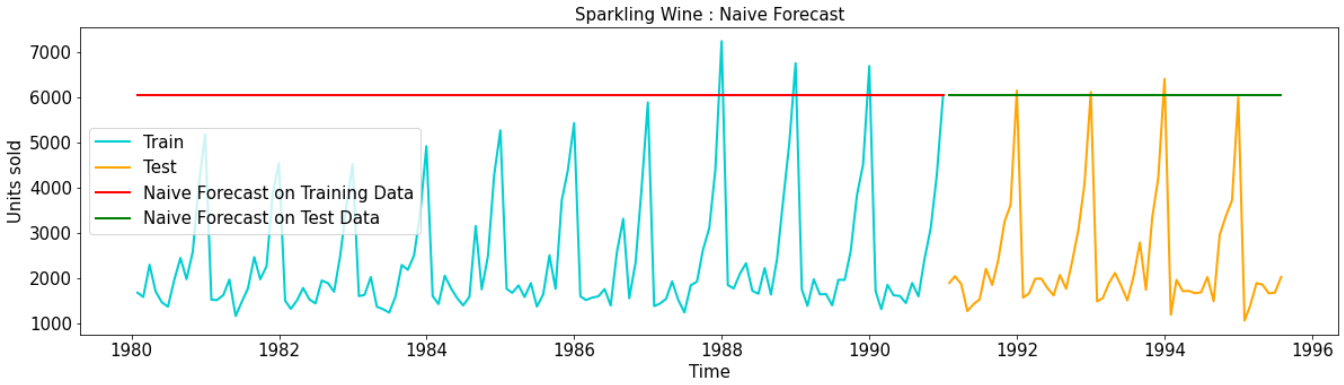


	RMSE	MAPE
Train	30.72	21.22
Test	15.27	22.82

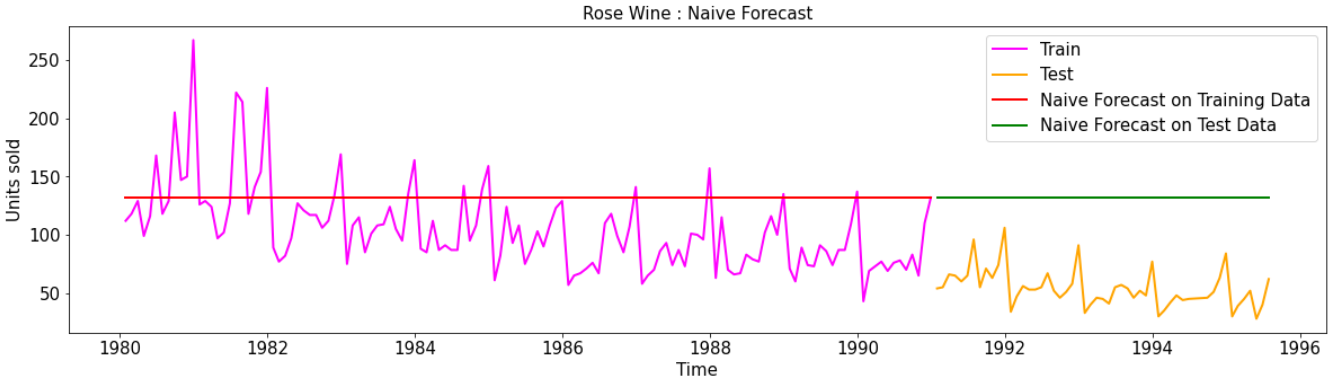
- The linear regression plots shows a gradual upward trend in forecast of Sparkling wine, consistent with the observed trend which was not visually apparent.
- The RMSE and MAPE values for Train and Test data sets are as above. 50% of forecast is erroneous.
- The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series
- The RMSE and MAPE of the forecast is given above. The model leaves a 23% error in forecast against test set.
- The model has successfully captured the trend of both the series, but does not reflects the seasonality

6. Naïve Forecasting

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.



	RMSE	MAPE
Train	3867.70	153.17
Test	3864.28	152.87

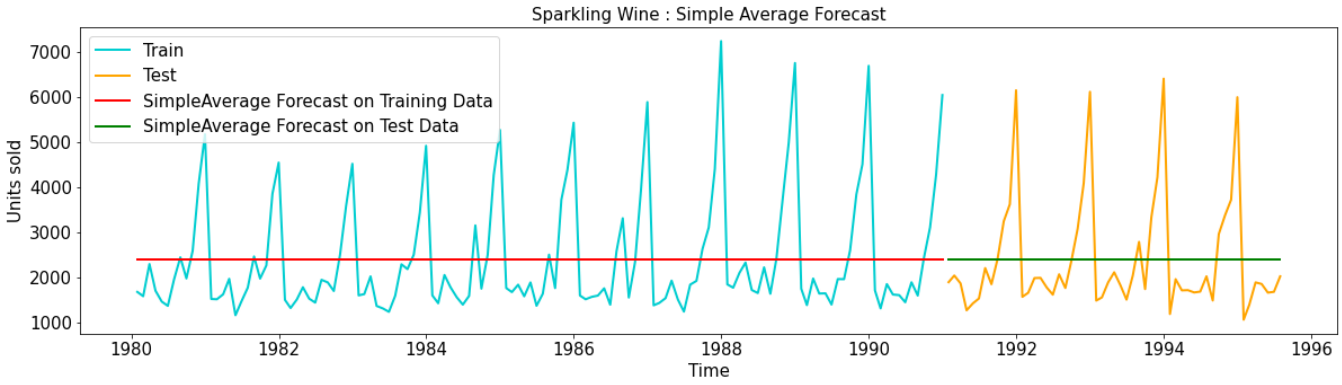


	RMSE	MAPE
Train	45.06	36.38
Test	79.72	145.10

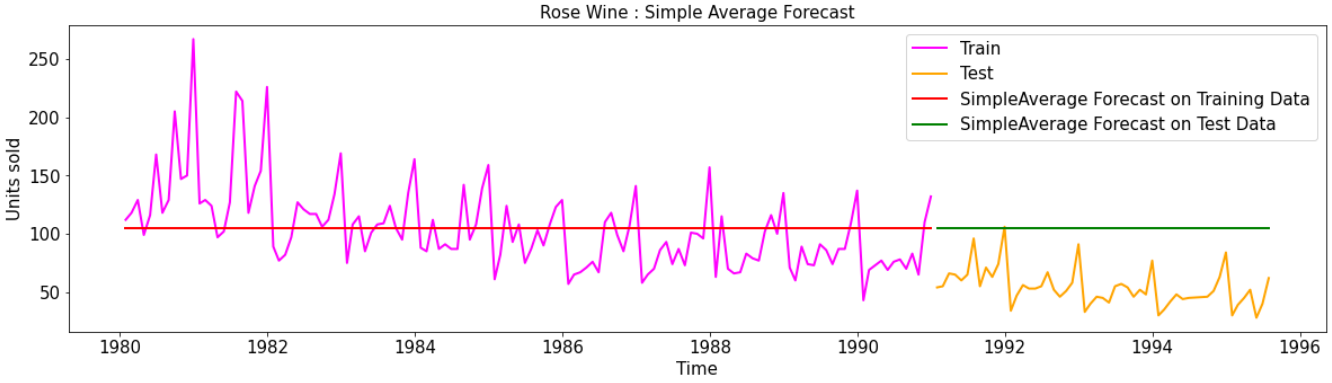
- The model has taken the last value from the test set and fitted it on the rest of the train time period and used the same value to forecast the test set.
- The performance matrices above shows a very poor fitment and high % of error.
- As Rose Wine data set has a downward trend the percentage of error in train is less very high in test.
- The model does not capture the trend nor seasonality of the given datasets.

7. Simple Average Forecasting

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



	RMSE	MAPE
Train	1298.48	40.36
Test	1275.08	38.90



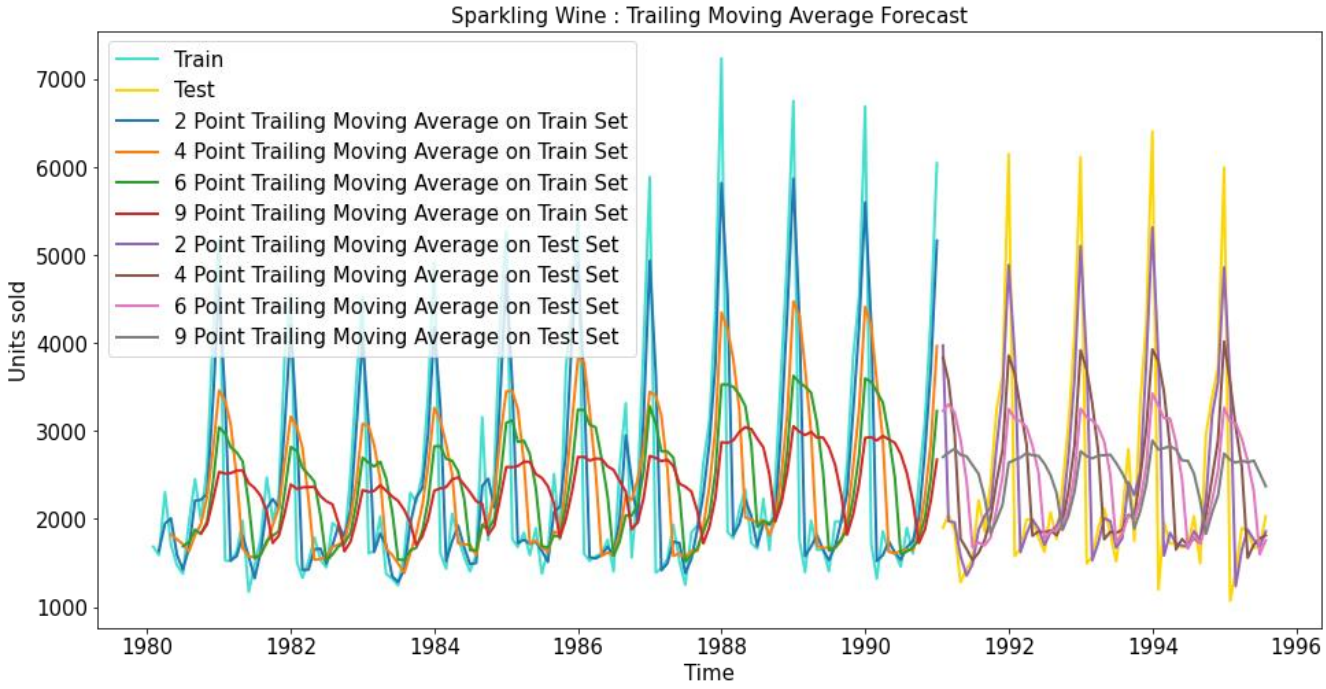
	RMSE	MAPE
Train	36.03	25.39
Test	53.46	94.93

- The model is not capable of neither forecasting nor able to capture the trend and seasonality present in the dataset.
- For Sparkling the RMSE and MAPE is consistent in both test and train datasets.
- For Rose dataset, the model forecast is almost 100% error in test data and 25% in train.
- Due to the downward trend the performance in train data set is better than the test dataset.

8. Moving Average

8.1. Sparkling Wine

- 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 (or the minimum error),
- 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 Root Mean Squared Error and Mean Absolute Percentage Error of the test set are given below
- The best interval of moving average from the model is 2 point.

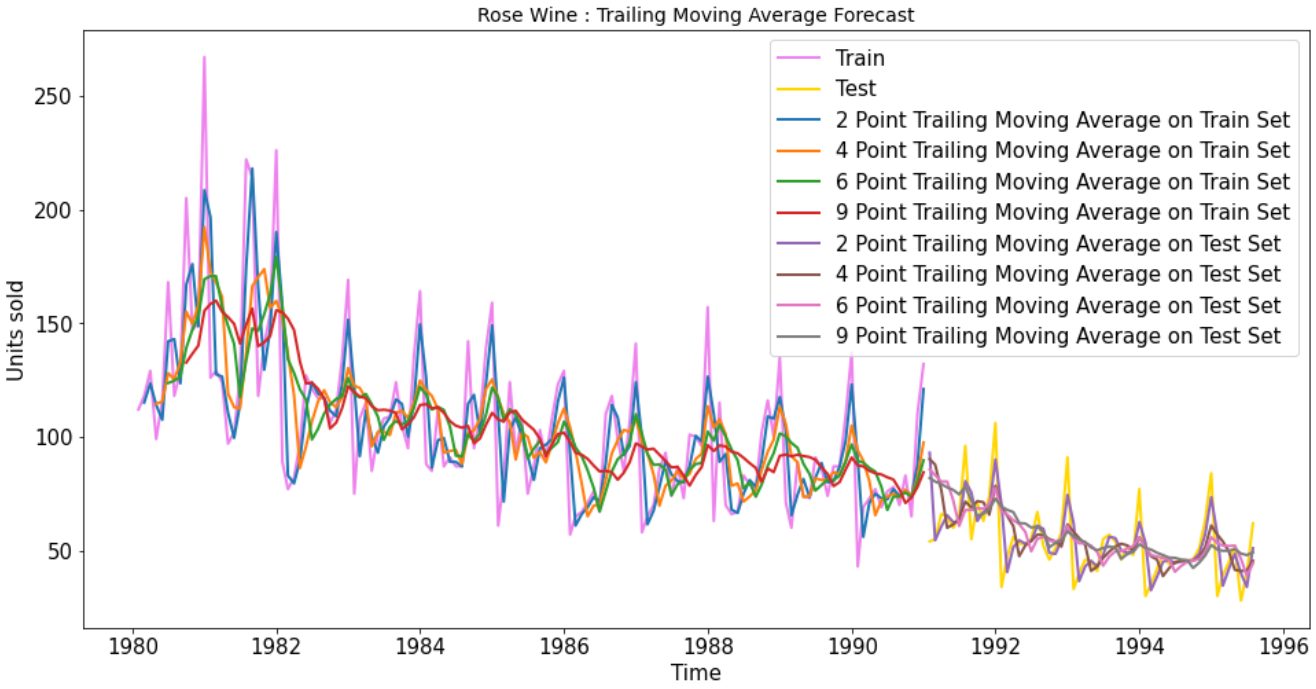


Model	RMSE	MAPE
2 point MA	813.40	19.70
4 point MA	1156.59	35.96
6 point MA	1283.93	43.86
9 point MA	1346.28	46.86

8.2. Rose Wine

- For the moving average model, we are going to calculate rolling means (or trailing moving averages) for different intervals.
- The best interval can be determined by the maximum accuracy (or the minimum error),
- 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 Root Mean Squared Error and Mean Absolute Percentage Error of the test set are given below
- The best interval of moving average from the model is 2 point.

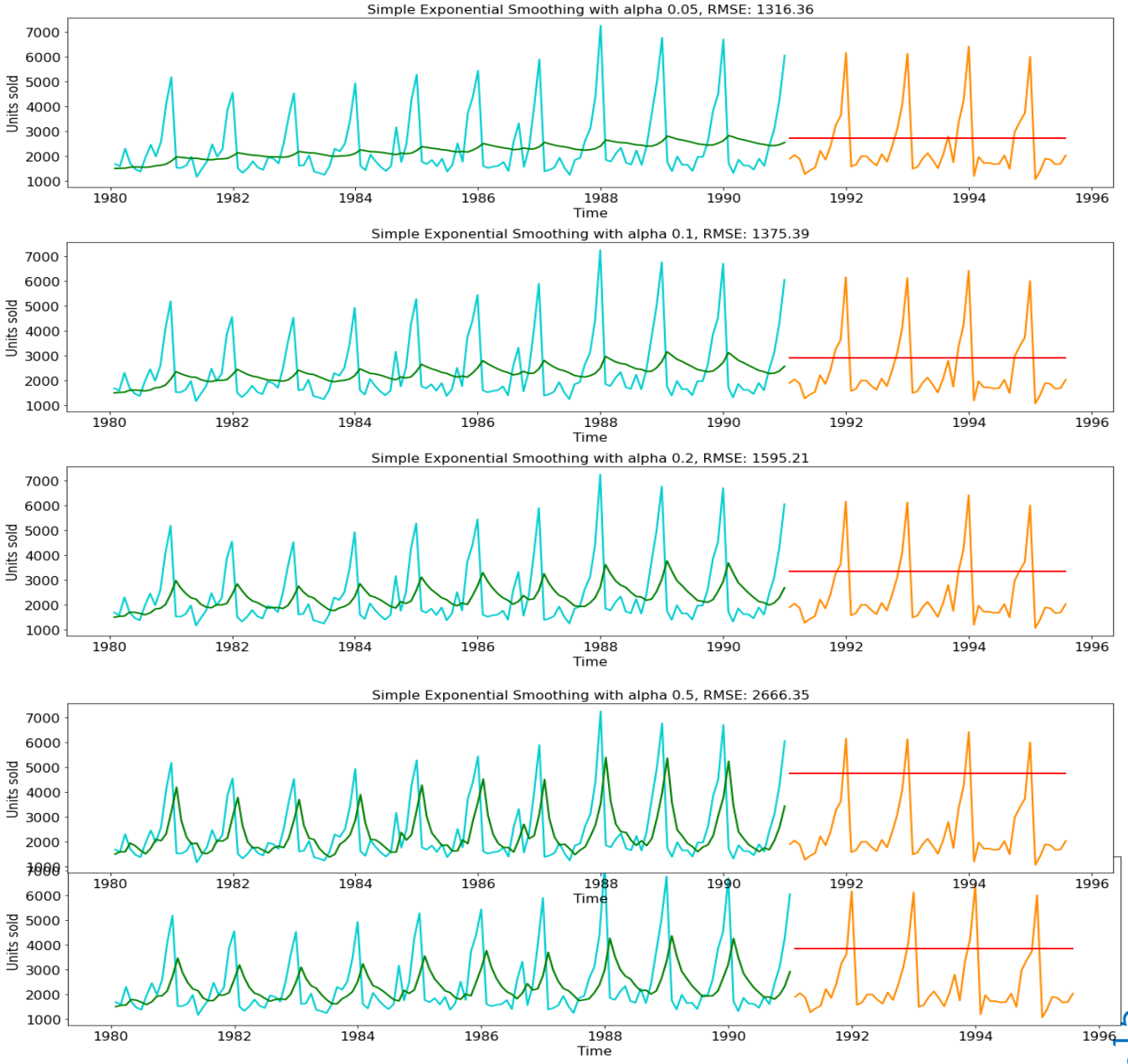


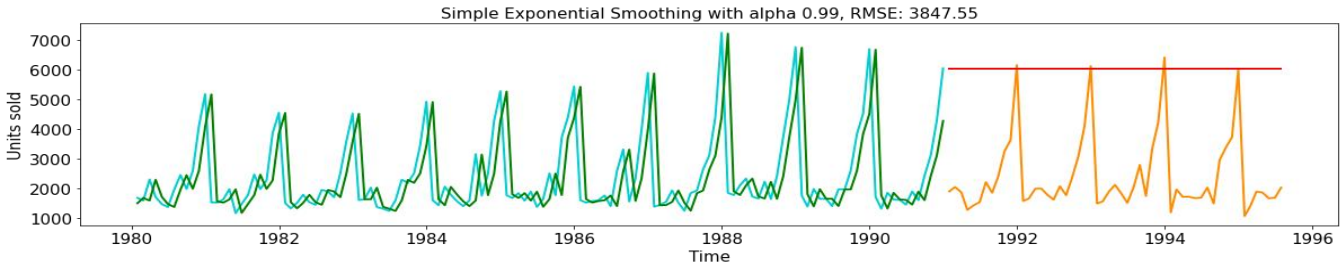
Model	RMSE	MAPE
2 point MA	11.53	13.54
4 point MA	14.45	19.49
6 point MA	14.57	20.82
9 point MA	14.73	21.01

9. Simple Exponential Smoothing

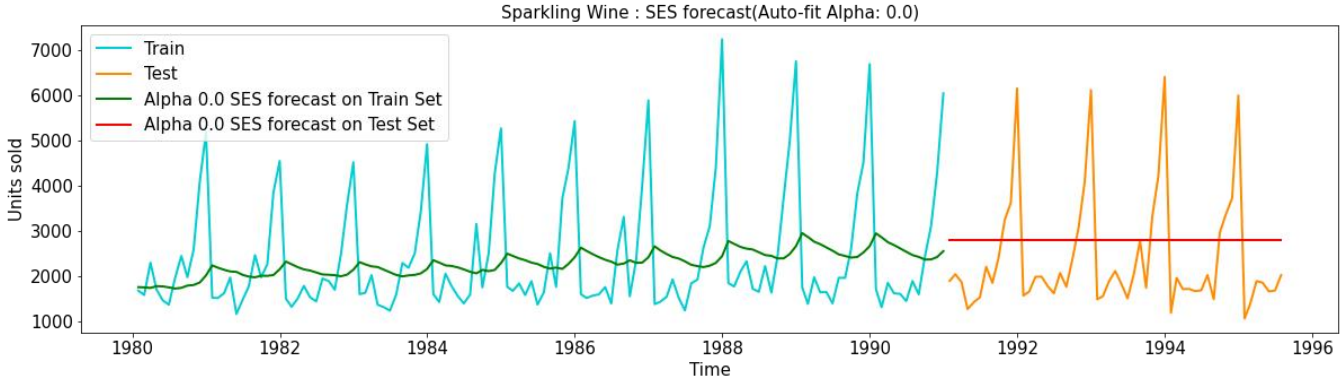
9.1. Sparkling Wine

- 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.





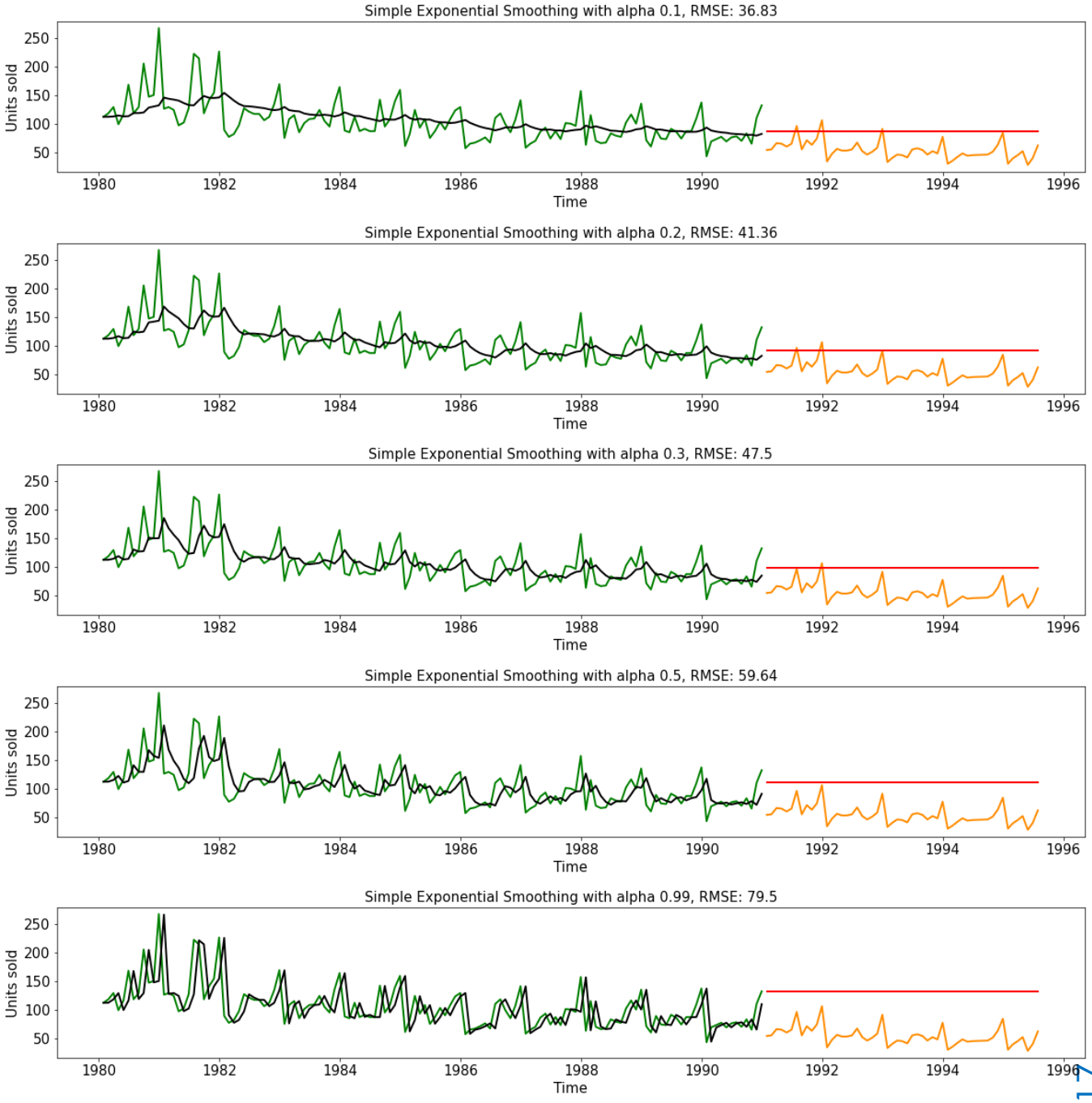
- On the second iteration, the model was ran without passing a value for alpha and used parameters '**optimized=True, use_brute=True'**
- The **autofit** model picked 0.0 as the smoothing parameter and retuned consistent RMSE values in train and test datasets, which is higher in accuracy than in first iteration
- As the smoothing level is 0.0, we got a completely smoothened out forecast with an initial value 2403.79 applied across the series.



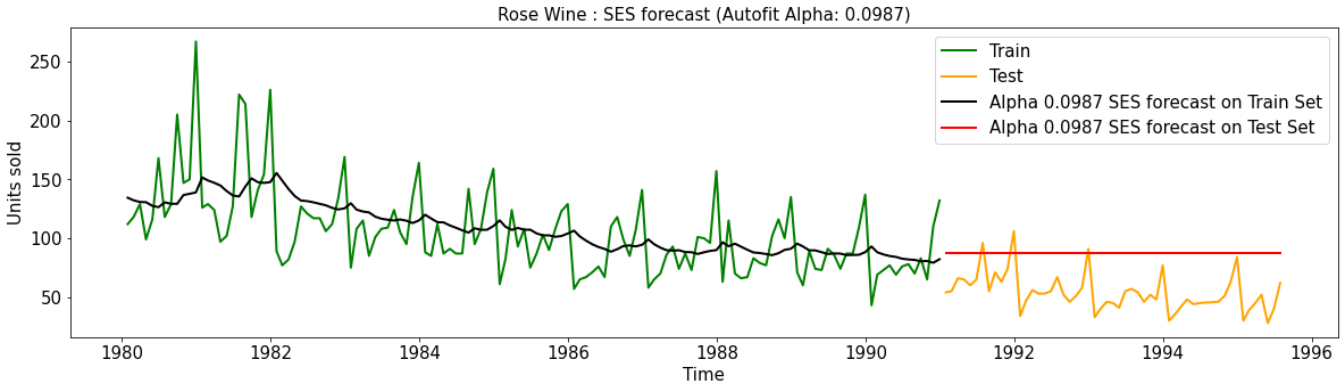
	RMSE	MAPE
Train	1298.48	40.36
Test	1275.08	38.90

9.2. Rose Wine

- Simple Exponential Smoothing is usually 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 or alpha between 0 and 1 are as below, where the values were 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.
- The test RMSE is found to be higher for values closer to zero.



- On the second iteration, the model was ran without passing a value for alpha and used parameters '**optimized=True, use_brute=True'.**
- The **autofit** model picked 0.098 as the smoothing parameter and retuned consistent RMSE values in train and test datasets, which is consistent with alpha 0.1 in first iteration.

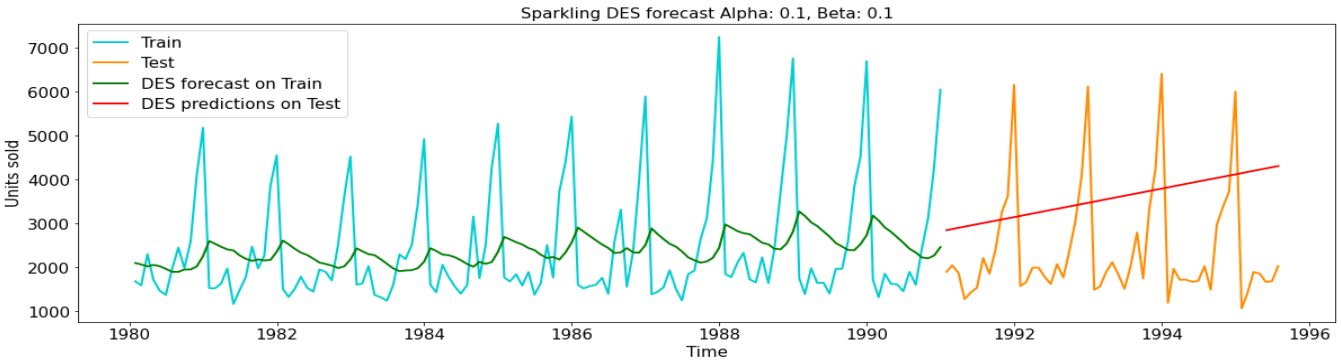


	RMSE	MAPE
Train	31.50	22.73
Test	36.80	63.88

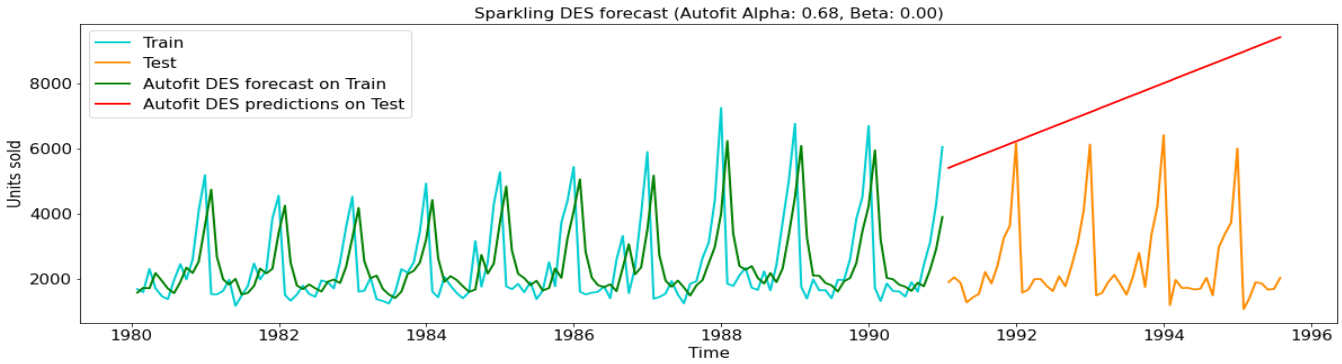
10. Double Exponential Smoothing

10.1. Sparkling Wine

- 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 and MAPE 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'**.



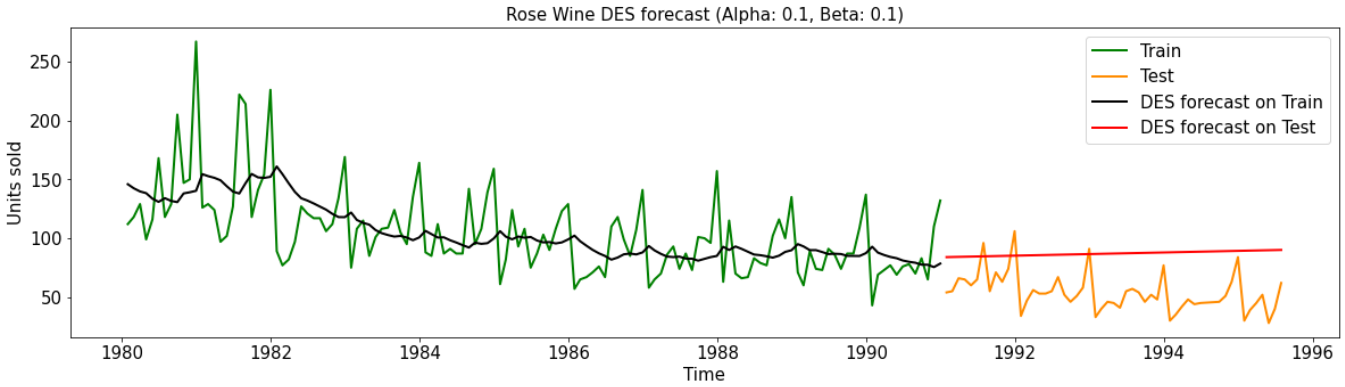
- The autofit model retuned higher accuracy in train dataset, but fared poorly in test, compared with the values in manual iteration
- The model evaluation parameters of top three models from manual iteration and the autofit models are as given above
- The best model chosen as final one is with alpha 0.1 and beta 0.1



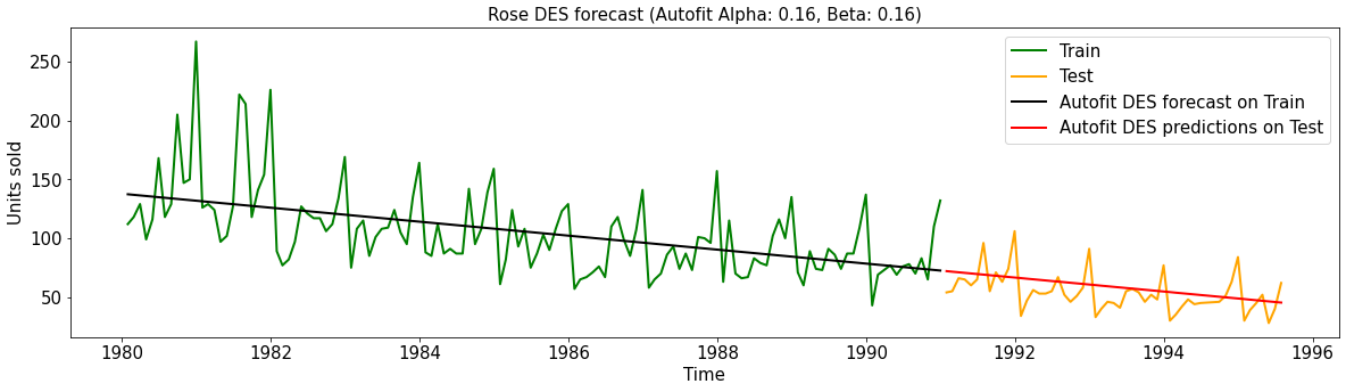
	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100	0.1000	1363.470000	44.26	1779.420000	67.23
1	0.100	0.2000	1398.190000	45.61	2601.540000	95.50
10	0.200	0.1000	1412.030000	46.62	3611.770000	135.41
2	0.100	0.3000	1431.370000	46.90	4288.430000	155.25
100	0.665	0.0001	1339.500882	38.82	5291.879833	208.74

10.2. Rose Wine

- 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 and MAPE 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'**.



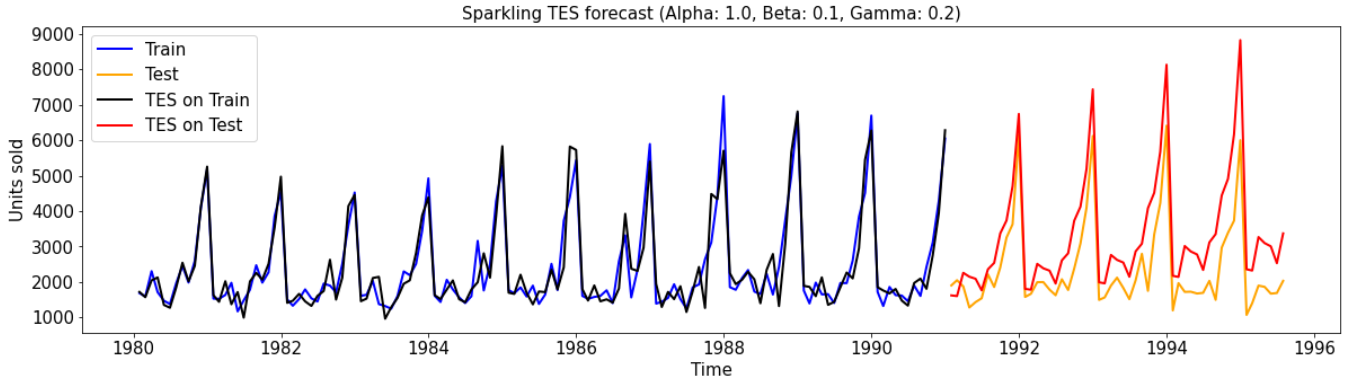
- The autofit model retuned higher accuracy in train dataset, on par with the best models from iteration 1, but faired behind in the test accuracy scores
- The model evaluation parameters of the best models are given as above
- The best model chosen as final one is the one with alpha 0.1 and beta 0.1



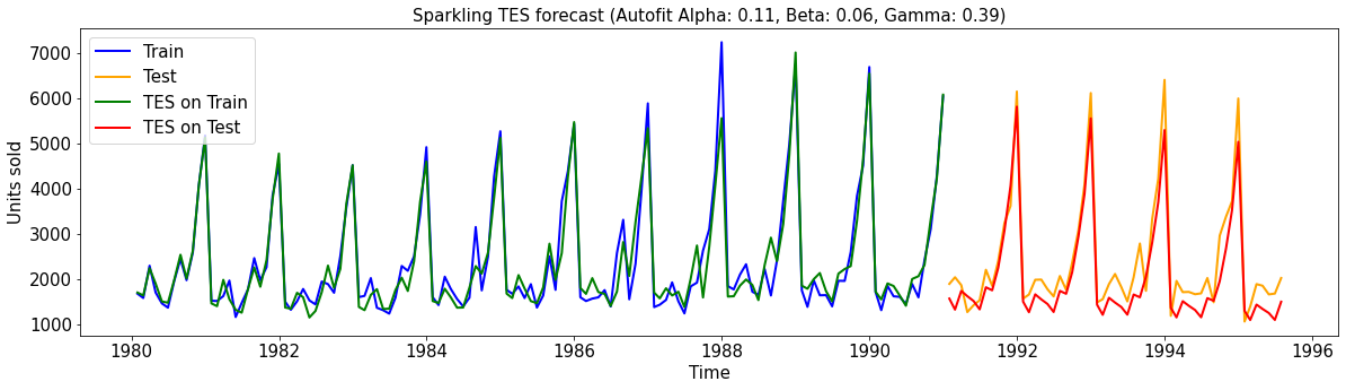
11. Triple Exponential Smoothing

11.1. Sparkling Wine

- 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.
- In 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 and MAPE values, which is as below with alpha 0.4, beta 0.1 and gamma 0.2.
- On the second iteration the model was allowed to choose the optimized values using parameters **'optimized=True, use_brute=True.**



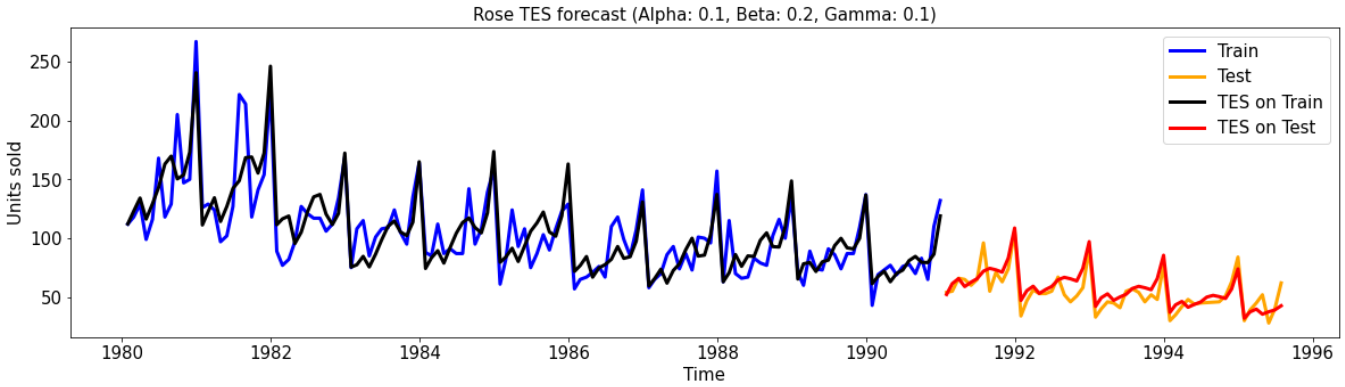
- The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but fared poorly in accuracy in test.
- The model evaluation parameters of the best models are given as above, including one from the autofit iteration.
- The best model chosen as final one is the one with alpha 0.4, beta 0.1 and gamma 0.2.



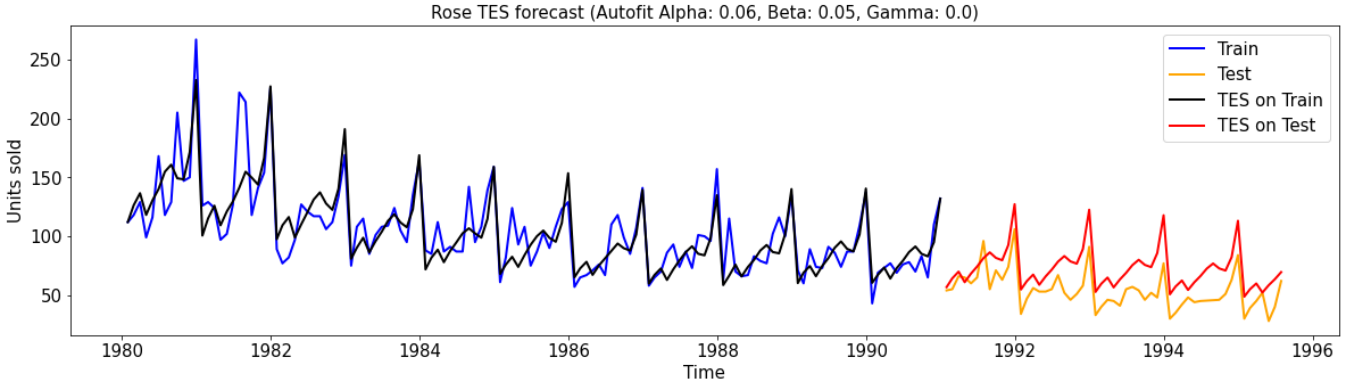
	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
301	0.4	0.1	0.2	373.247010	11.04	311.981477	10.18
211	0.3	0.2	0.2	377.305073	11.23	315.237398	10.08
300	0.4	0.1	0.1	370.612639	11.03	318.103555	10.01
402	0.5	0.1	0.3	390.175608	11.54	325.544934	9.99
30	0.1	0.4	0.1	403.885024	11.71	331.703834	10.58

11.2. Rose Wine

- The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Rose data contain both trend and seasonality significantly
- In 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 and MAPE values, which is as below with alpha 0.1, beta 0.2 and gamma 0.2.
- On the second iteration the model was allowed to choose the optimized values using parameters **'optimized=True, use_brute=True.**



- The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but fared poorly in accuracy in test
- The model evaluation parameters of the best models are given as above, including one from the autofit iteration
- The best model chosen as final one is the one with alpha 0.1, beta 0.2 and gamma 0.2



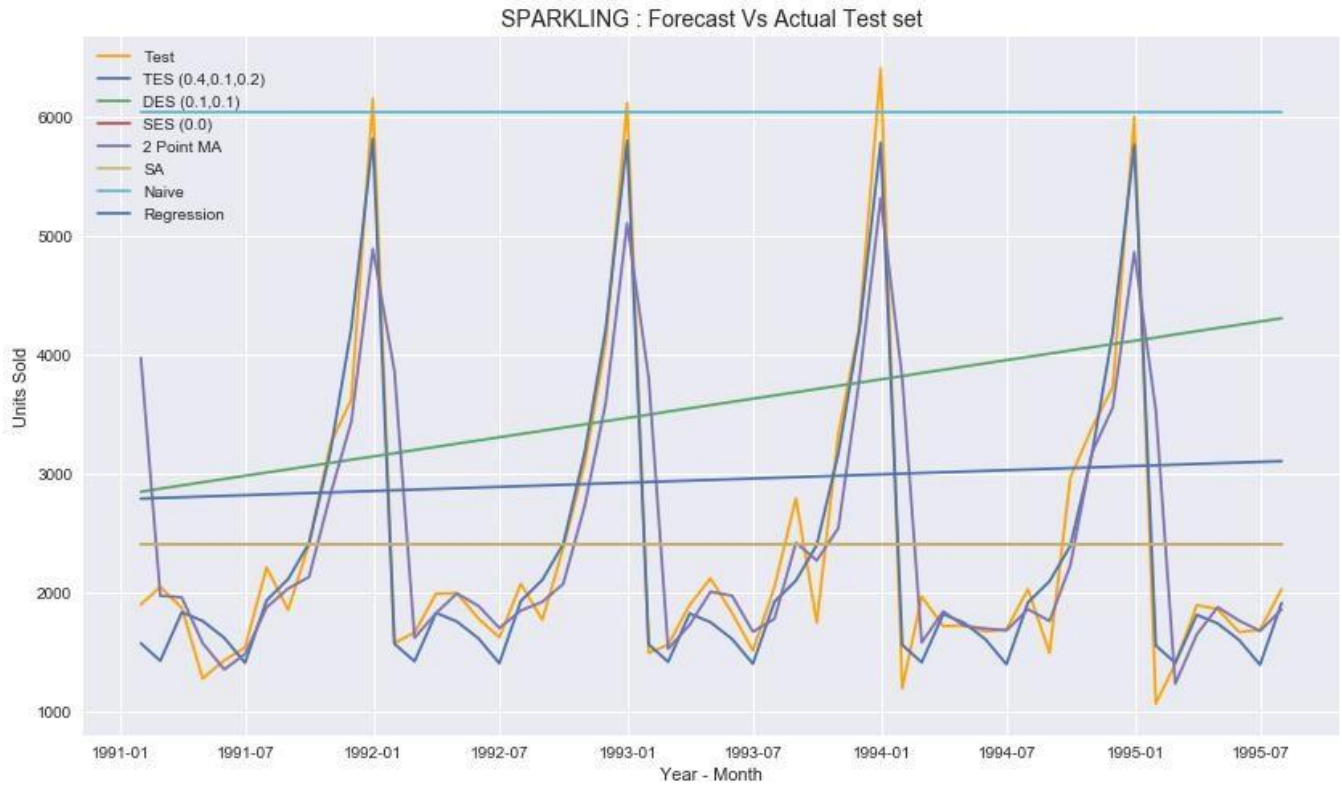
	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
10	0.1	0.2	0.1	19.651464	14.31	9.171737	13.19
11	0.1	0.2	0.2	20.140683	14.66	9.493928	13.69
151	0.2	0.6	0.2	22.793871	17.02	9.682619	13.71
142	0.2	0.5	0.3	23.300524	17.35	9.885630	14.21
12	0.1	0.2	0.3	20.725703	14.88	9.896242	14.16

12. Model Comparison

12.1. Sparkling Wine

- The accuracy of the time-series forecast models build in the previous sections of this report is as below, sorted by RMSE in test data.
- The plot of the forecasts fitted on to the test data is given as well.
- 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.

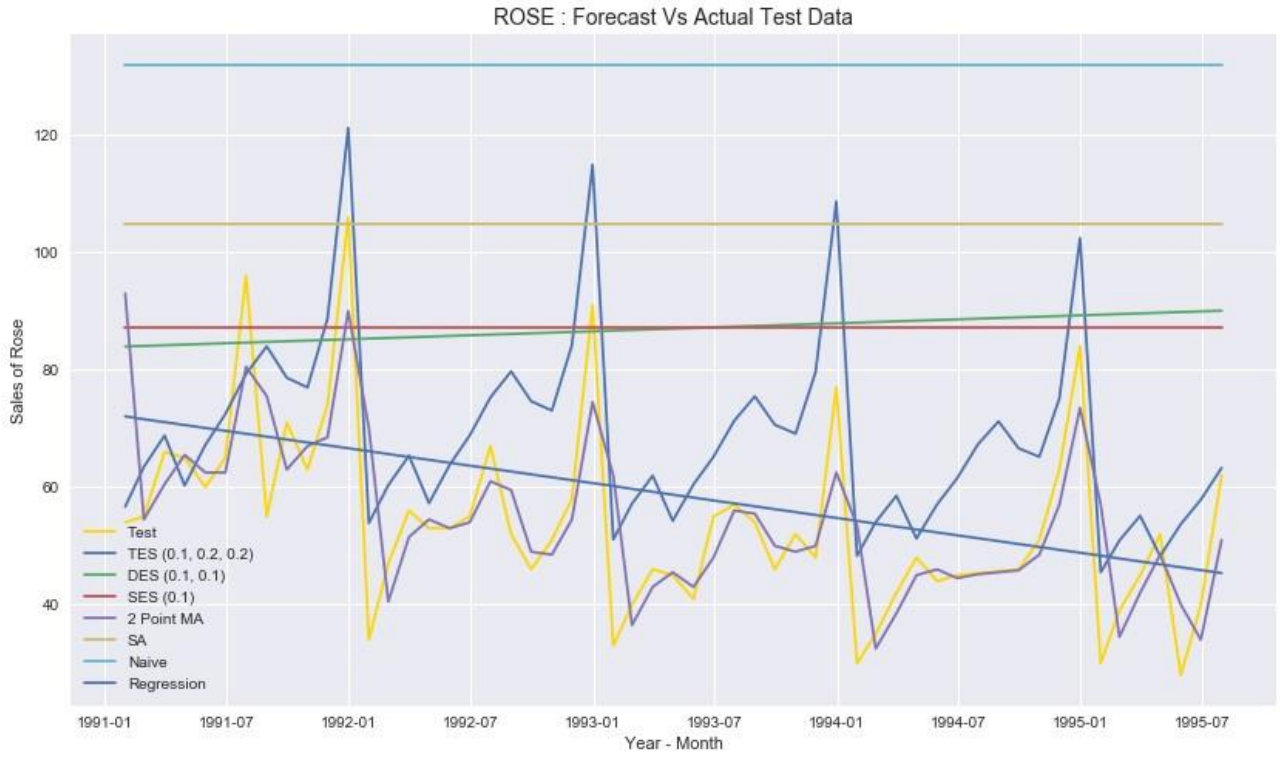
	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	311.981477	10.18
TES Alpha 0.15, Beta 0.00, Gamma 0.37	469.659106	16.39
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.00	1338.000861	47.11
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
NaiveModel	3864.279352	152.87
DES Alpha 0.6,Beta 0.0	5291.879833	208.74



12.2. Rose Wine

- The accuracy of the time-series forecast models build in the previous sections of this report is as below, sorted by RMSE in test data
- The plot of the forecasts fitted on to the test data is given as well
- 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

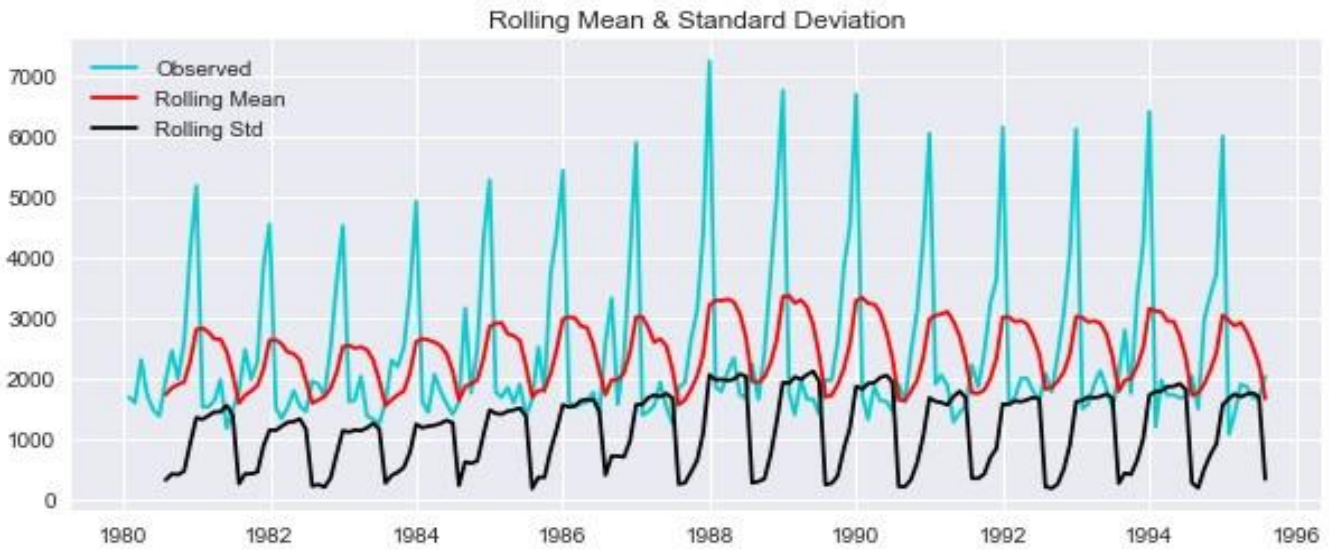
	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.640616	13.96
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
TES Alpha 0.11, Beta 0.05, Gamma 0.00	17.369210	28.88
SES Alpha 0.01	36.796019	63.88
DES Alpha 0.10, Beta 0.10	37.056912	64.02
SimpleAverage	53.460350	94.93
DES Alpha 0.16, Beta 0.16	70.572197	120.25
NaiveModel	79.718559	145.10



13. Check Stationarity of Data

13.1. Sparkling Wine

- 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

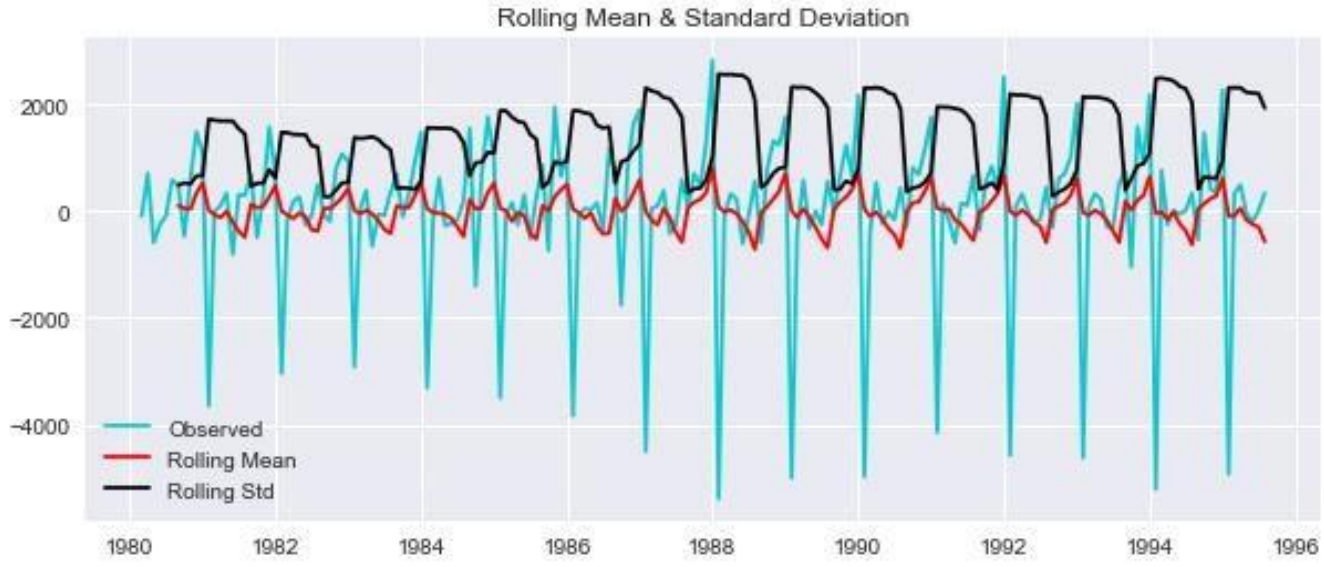
ADF on original series

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

ADF on differenced series

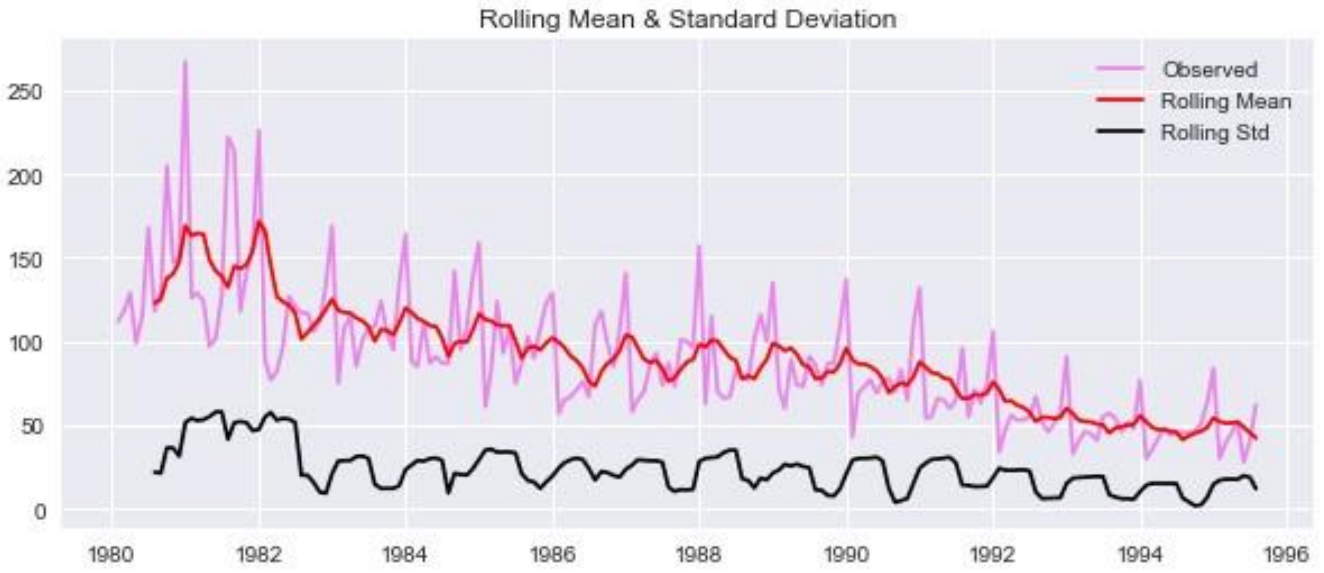
- P-Value < alpha .05
- Test statistic < Critical values
- Reject the null hypothesis
- The series is stationary

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



- Differencing of order one is applied on the Sparkling series as above and tested for stationarity. At an order of differencing 1, the series is found to be stationary as above
- The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if its multiplicative or additive in character
- The altitude of rolling mean and standard deviation 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.

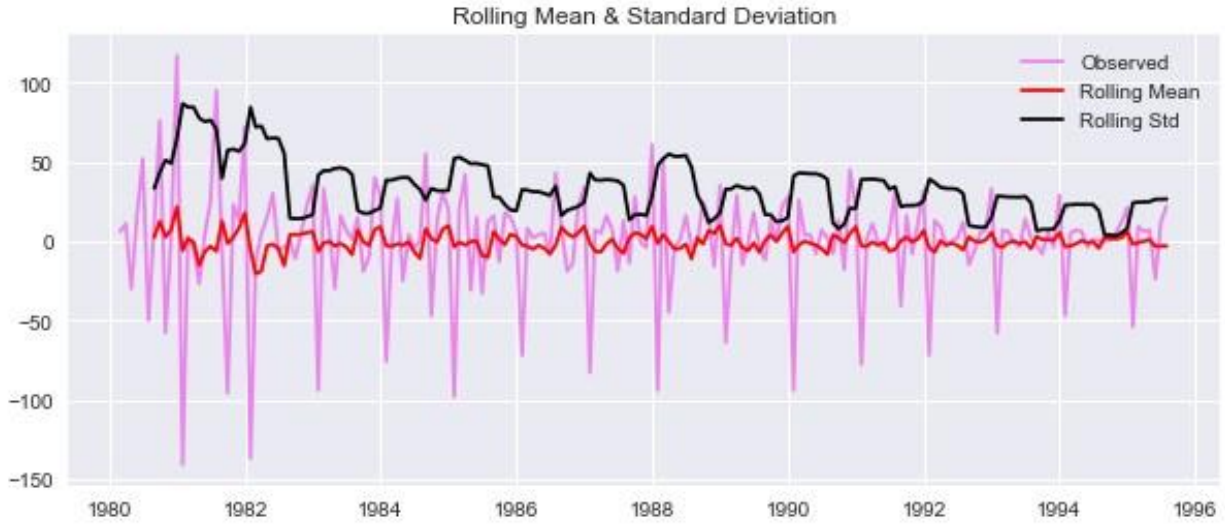
13.2. Rose Wine



Results of Dickey-Fuller Test:	
Test Statistic	-1.876719
p-value	0.343091
#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

ADF on original series

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



ADF on differenced series

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

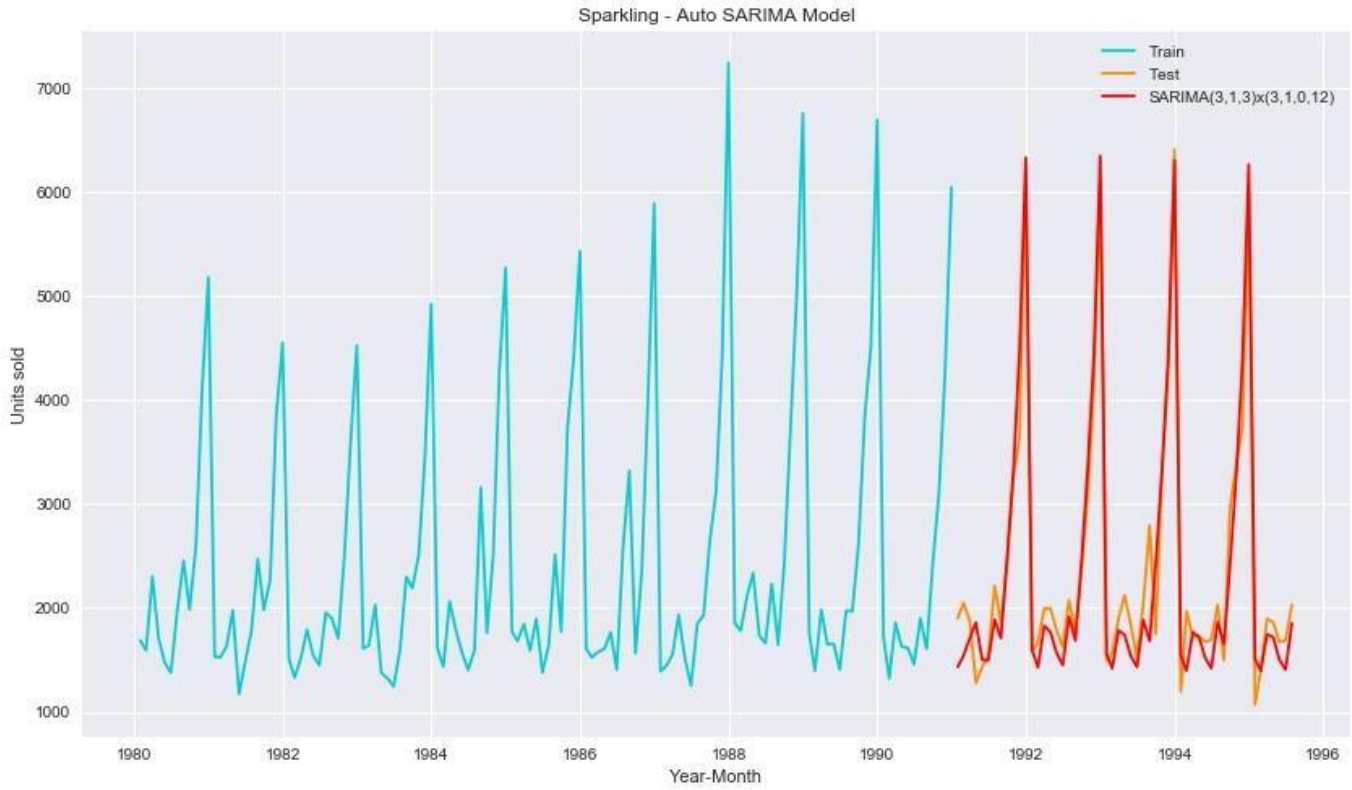
Results of Dickey-Fuller Test:	
Test Statistic	-8.044395e+00
p-value	1.810868e-12
#Lags Used	1.200000e+01
Number of Observations Used	1.730000e+02
Critical Value (1%)	-3.468726e+00
Critical Value (5%)	-2.878396e+00
Critical Value (10%)	-2.575756e+00

- Differencing of order one is applied on the Sparkling series as above and tested for stationarity
- At an order of differencing 1, the series is found to be stationary as above.
- 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 plot of rolling mean and standard deviation indicates that the seasonality is multiplicative as the altitude of plot varies with respect to trend.
- 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.

14. Auto SARIMA

14.1. Sparkling Wine

• As the Sparkling series of data contain seasonality component we will be building SARIMA model, rather than ARIMA
• Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as an element of multiplicity in seasonality is suspected
• The model built with original data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model
• The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values.



• The top three models with lowest AIC values are as given. As per the AIC criteria, the optimum values for final SARIMA model selected is (3, 1, 3)x(3, 1, 0, 12)
• The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of error, and the histogram shows the residuals follow a normal distribution

	param	seasonal	AIC
252	(3, 1, 3)	(3, 1, 0, 12)	1213.282561
253	(3, 1, 3)	(3, 1, 1, 12)	1215.213337
220	(3, 1, 1)	(3, 1, 0, 12)	1215.898777

	Test RMSE	Test MAPE
Auto SARIMA(3,1,3)x(3,1,0,12)	331.614531	10.33
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.800722	11.19

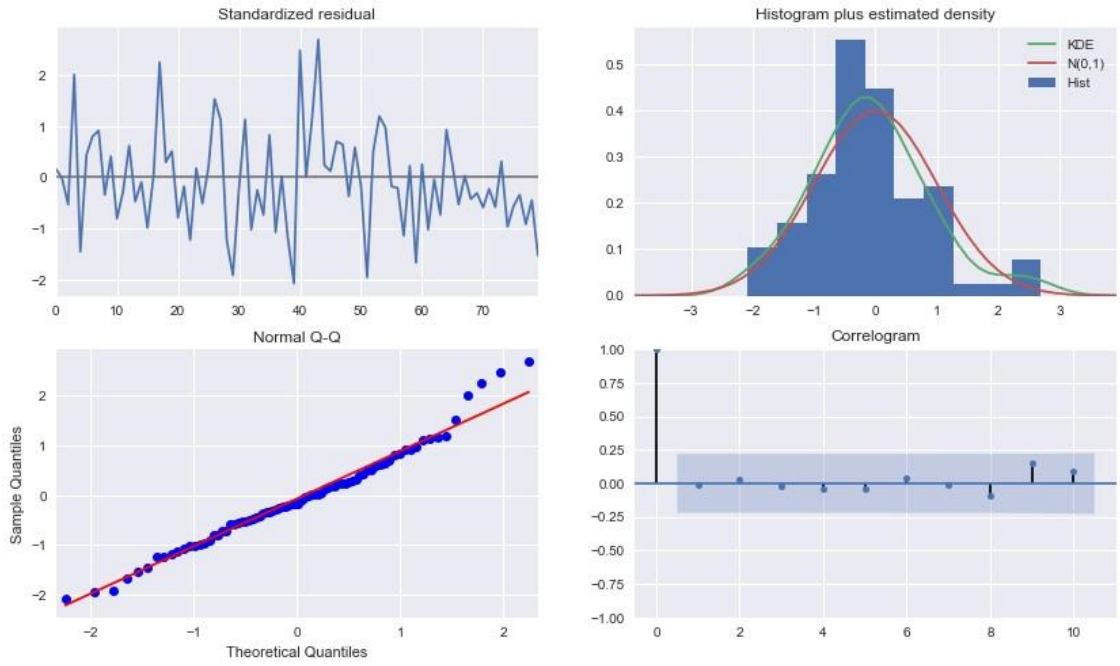
• The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point's forms roughly a straight line

• The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index.
• The RMSE and MAPE values of the automated SARIMA models built are given here.
• From the below model summary it can be inferred that AR (1), MA (1), MA (3), MA (2) terms has the highest absolute weightage.
• From the p-values it can be inferred that terms AR(1), AR(2), MA(1), MA(2), MA(3) and seasonal AR(1) are significant terms, as their values are below 0.0.

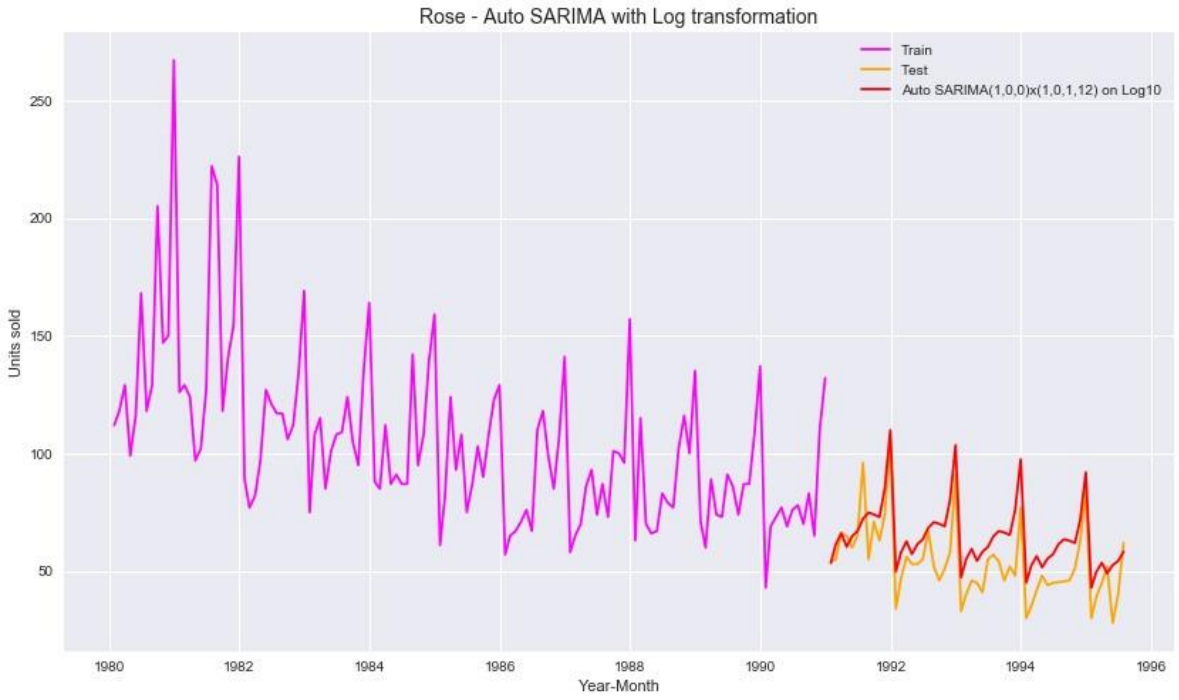
Model Summary – SARIMA (3, 1, 3)x(3, 1, 0, 12)

```
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 1, 0, 12)   Log Likelihood                -596.641
Date:                            Sun, 13 Sep 2020   AIC                           1213.283
Time:                                    19:55:24   BIC                           1237.103
Sample:                                         0   HQIC                          1222.833
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6142      0.176     -9.177      0.000      -1.959      -1.269
ar.L2         -0.6124      0.299     -2.048      0.041      -1.198      -0.026
ar.L3          0.0860      0.161      0.536      0.592      -0.229       0.401
ma.L1          0.9853      0.465      2.117      0.034       0.073       1.898
ma.L2         -0.8739      0.166     -5.268      0.000      -1.199      -0.549
ma.L3         -0.9464      0.483     -1.960      0.050      -1.893      -0.000
ar.S.L12      -0.4521      0.142     -3.193      0.001      -0.730      -0.175
ar.S.L24      -0.2345      0.144     -1.625      0.104      -0.517       0.048
ar.S.L36      -0.1007      0.122     -0.828      0.408      -0.339       0.138
sigma2     1.839e+05   8.86e+04      2.076      0.038    1.03e+04    3.57e+05
==========================================================================================
Ljung-Box (Q):                       23.21   Jarque-Bera (JB):                 4.06
Prob(Q):                              0.98   Prob(JB):                         0.13
Heteroskedasticity (H):               0.73   Skew:                             0.48
Prob(H) (two-sided):                  0.42   Kurtosis:                         3.54
==========================================================================================
```

Diagnostics plot – SARIMA (3, 1, 3)x(3, 1, 0, 12)



14.1. Rose Wine



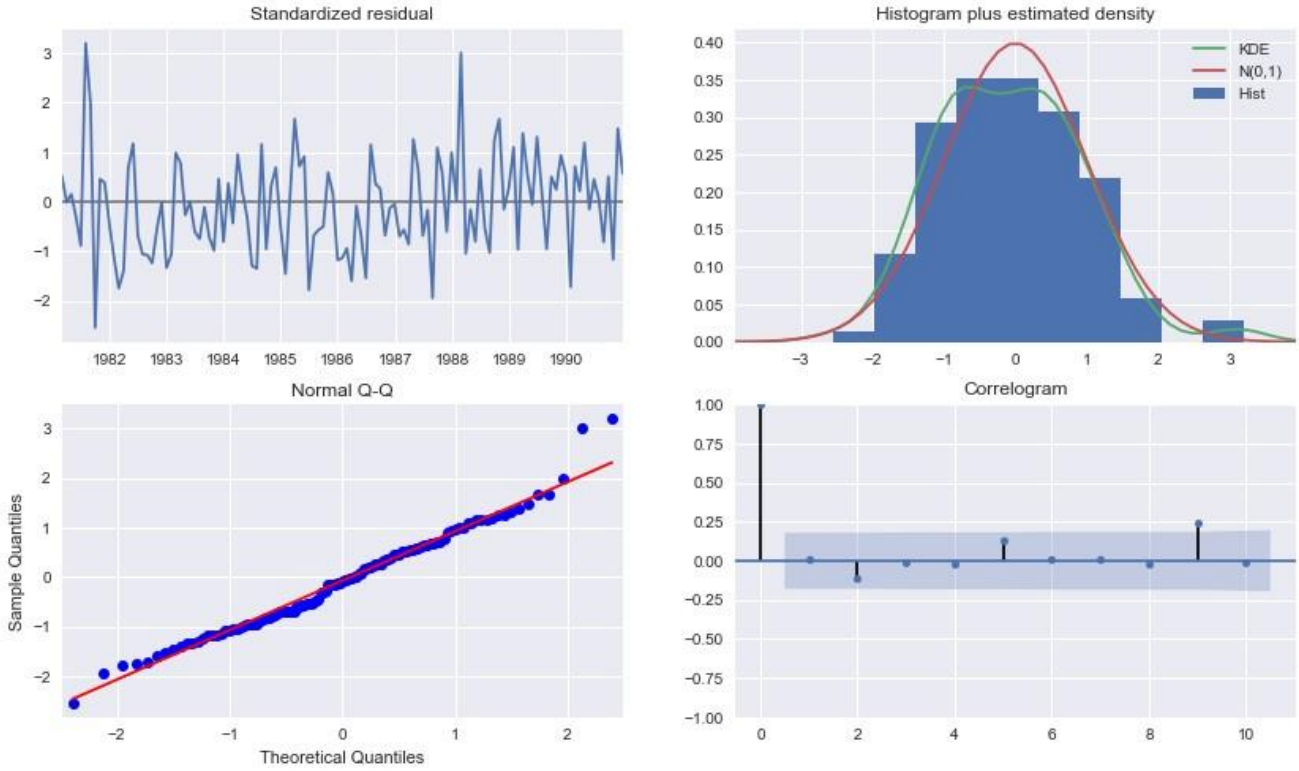
	param	seasonal	AIC
31	(1, 0, 0)	(1, 0, 1, 12)	-257.620760
4	(0, 0, 0)	(1, 0, 1, 12)	-256.170282
40	(1, 0, 1)	(1, 0, 1, 12)	-255.482062

	Test RMSE	Test MAPE
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823618	25.48
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.595882	21.93

Model Summary – SARIMA (1, 0, 0)x(1, 0, 1, 12)

```
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  132
Model:             SARIMAX(1, 0, 0)x(1, 0, 1, 12)   Log Likelihood                132.810
Date:                            Sun, 13 Sep 2020   AIC                            -257.621
Time:                                    20:24:09   BIC                            -246.504
Sample:                                01-31-1980   HQIC                           -253.107
                                     - 12-31-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1689      0.078      2.179      0.029       0.017       0.321
ar.S.L12       0.9872      0.001    751.658      0.000       0.985       0.990
ma.S.L12      -0.9411      0.351     -2.684      0.007      -1.628      -0.254
sigma2         0.0052      0.002      2.885      0.004       0.002       0.009
===================================================================================
Ljung-Box (Q):                       24.28   Jarque-Bera (JB):                 4.00
Prob(Q):                              0.98   Prob(JB):                         0.14
Heteroskedasticity (H):               0.86   Skew:                             0.40
Prob(H) (two-sided):                  0.64   Kurtosis:                         3.40
===================================================================================
```

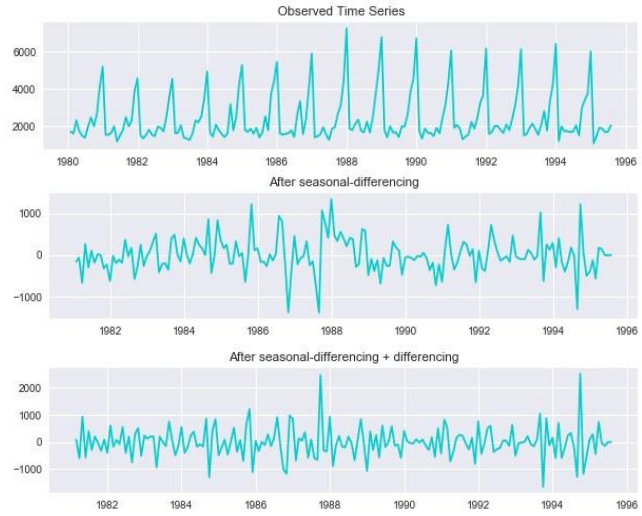
Diagnostics plot – SARIMA (1, 0, 0)x(1, 0, 1, 12)

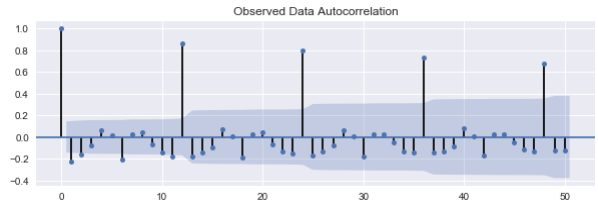


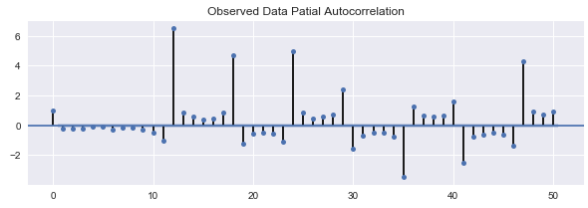
15. Manual SARIMA

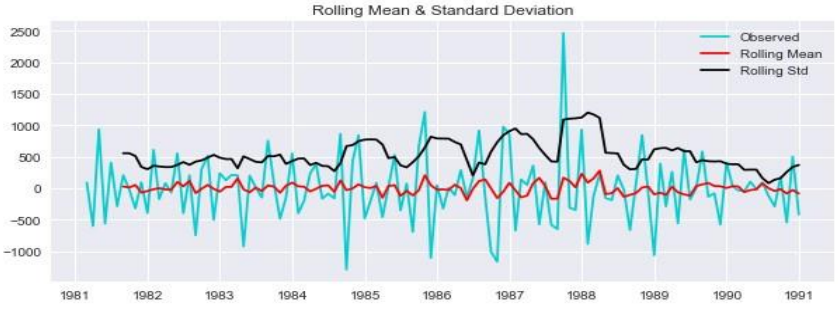
15.1. Sparkling Wine

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





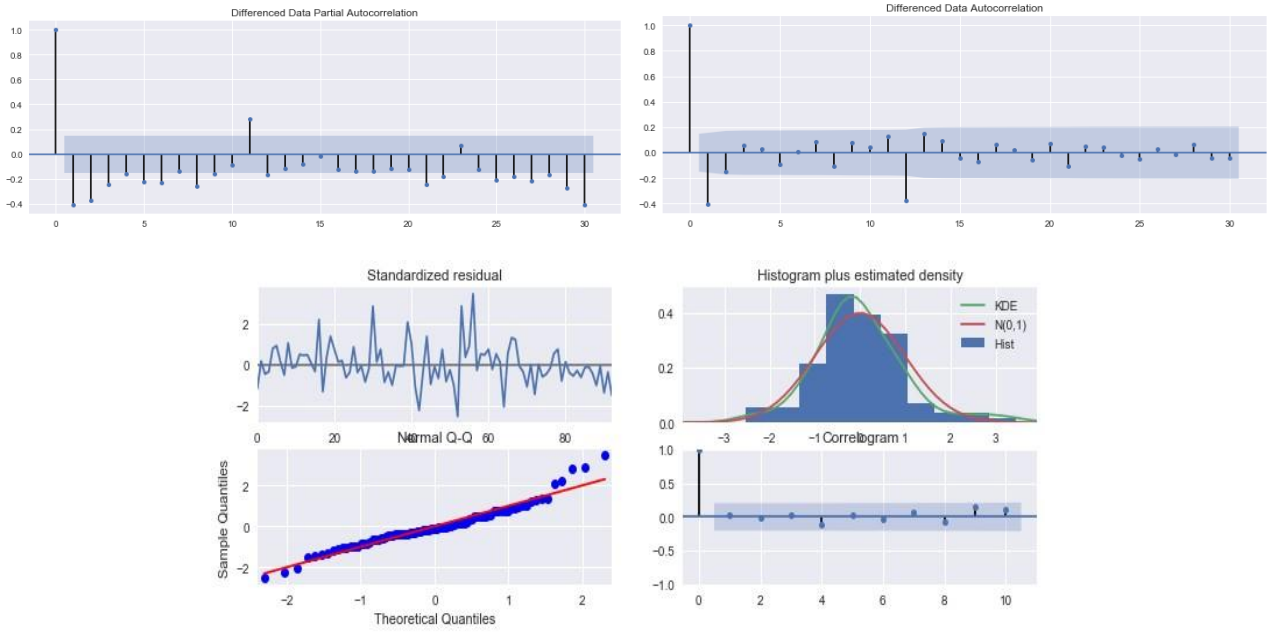





```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
#Lags Used                      10.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value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

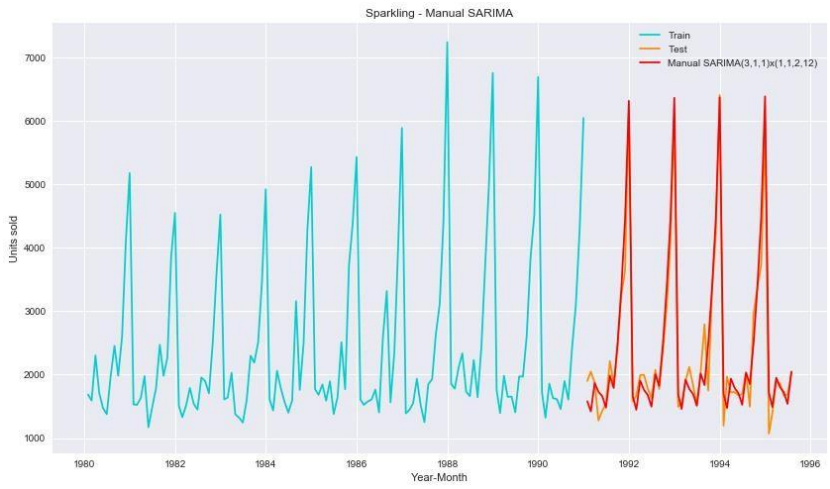
- The seasonal MA term 'Q' was later optimized to 2, by validating model performance, as the data might be under-differenced.
- The final selected terms for SARIMA model is (3, 1, 1)x(0, 1, 2, 12).

- The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point's forms roughly a straight line
- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index



	RMSE	MAPE
Test	324.108	9.48

- The model summary indicates that that only MA(1) term used in the model is significant in terms of p-values
- From the multiple iterations of SARIMA models, below is the comparison of the models in terms of its accuracy attributes of RMSE and MAPE.

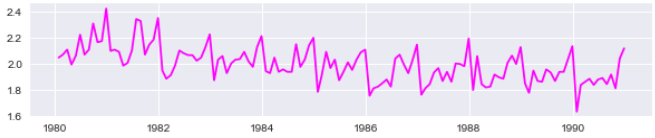


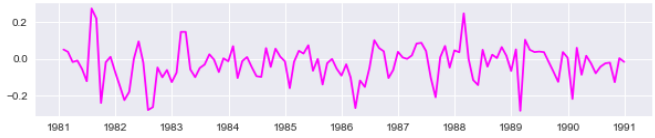
	coef	std err	z	P>|z|
ar.L1	0.2229	0.130	1.713	0.087
ar.L2	-0.0798	0.131	-0.607	0.544
ar.L3	0.0921	0.122	0.756	0.450
ma.L1	-1.0241	0.094	-10.925	0.000
ar.S.L12	-0.1992	0.866	-0.230	0.818
ma.S.L12	-0.2109	0.881	-0.239	0.811
ma.S.L24	-0.1299	0.381	-0.341	0.733

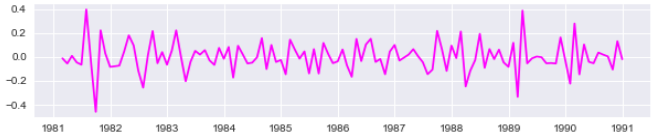
	Test RMSE	Test MAPE
Auto SARIMA(3,1,3)x(3,1,0,12)	331.614531	10.33
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.800722	11.19
Manual SARIMA(3,1,1)x(1,1,2,12)	324.108003	9.48

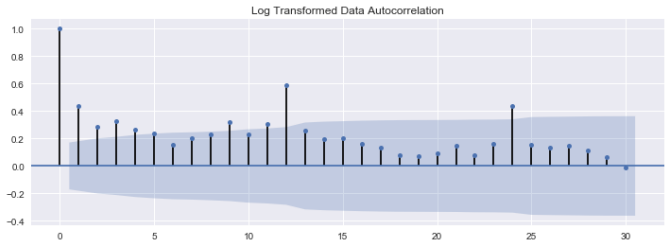
15.2. Rose Wine

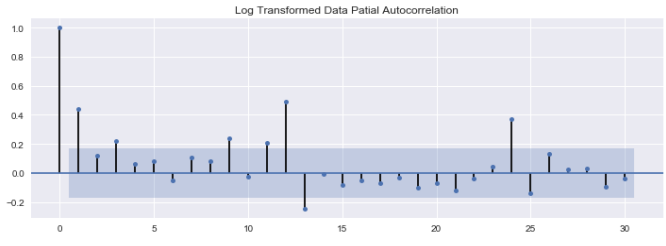
Log transformation of the Rose Wine data is done to handle multiplicity of seasonality

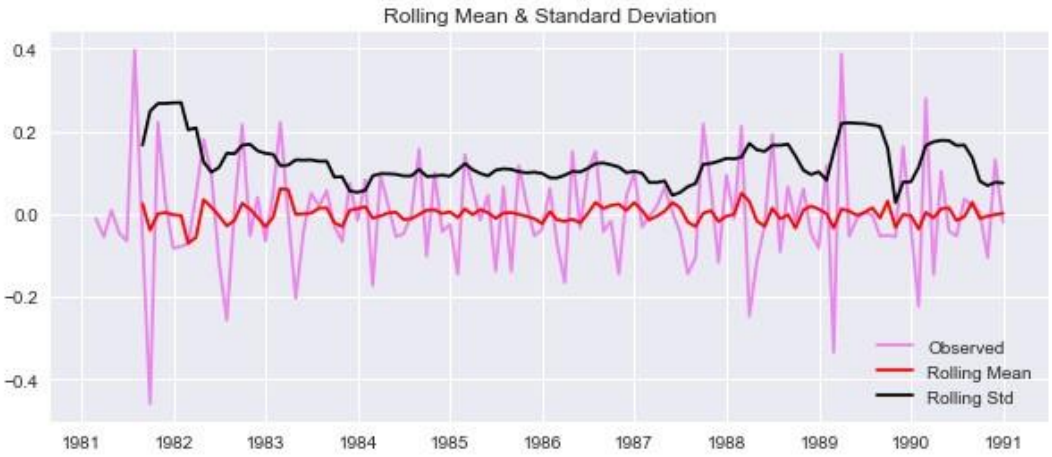








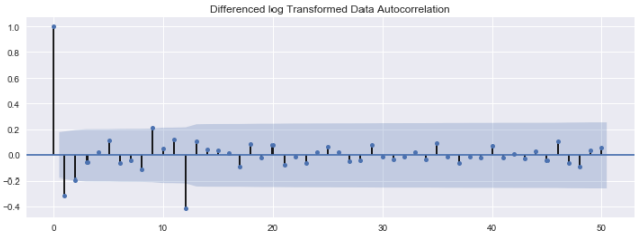


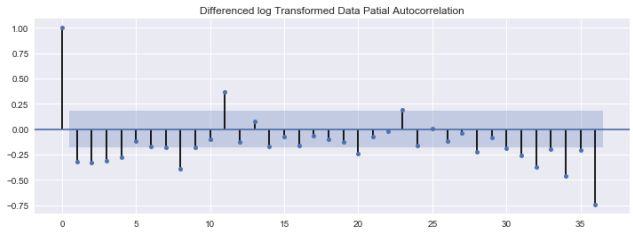


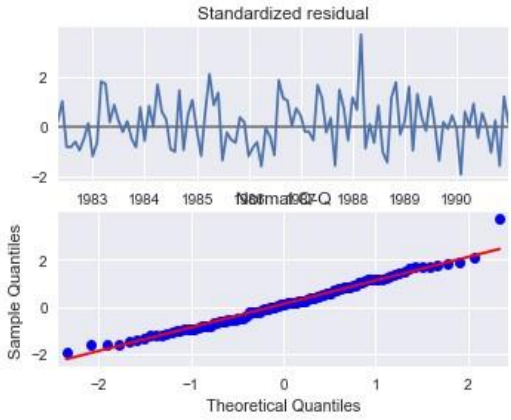
Test Statistic	-3.910109
p-value	0.001962
#Lags Used	11.000000
Number of Observations Used	107.000000
Critical Value (1%)	-3.492996
Critical Value (5%)	-2.888955
Critical Value (10%)	-2.581393

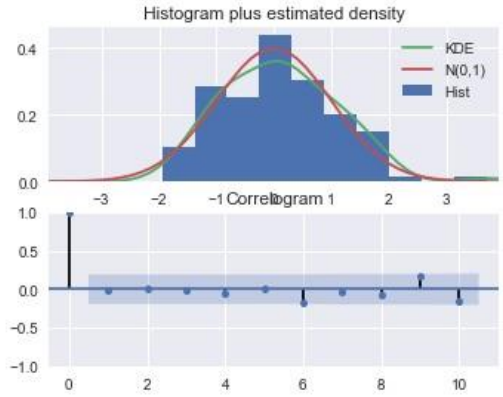
- Here we have taken alpha = 0.05 and seasonal period as 12
- The final selected terms for SARIMA model is (4, 1, 1)x(0, 1, 1, 12), as inferred from the ACF and PACF plots
- The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the point's forms roughly a straight line

- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index





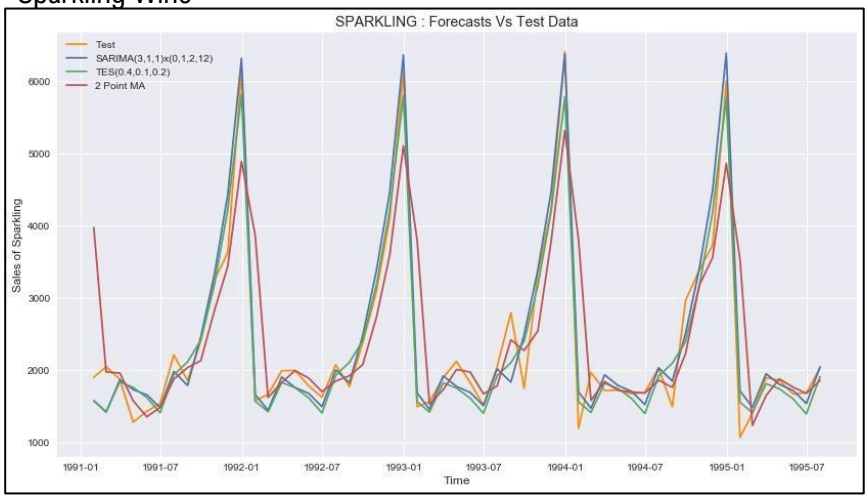




	RMSE	MAPE
Test	14.176	23.10

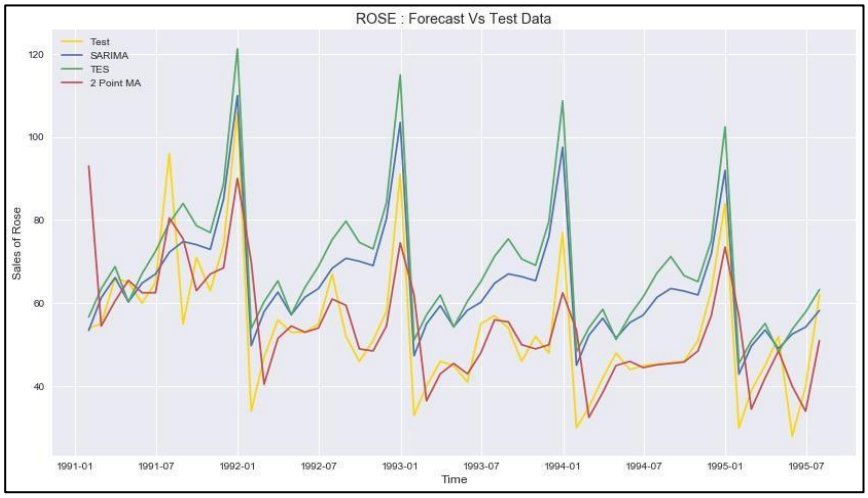
16. Model Comparison

Sparkling Wine



	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.4, Gamma 0.2	312.211095	10.20
Manual SARIMA(3,1,1)x(1,1,2,12)	324.108003	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.614531	10.33
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.800722	11.19
TES Alpha 0.15, Beta 0.00, Gamma 0.37	384.203001	11.94
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
SES Alpha 0.00	1275.081823	38.90
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.430000	67.23
DES Alpha 0.6,Beta 0.0	3851.171500	152.07
NaiveModel	3864.279352	152.87

Rose Wine



	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.640616	13.96
2 point TMA	11.529278	13.54
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.595882	21.93
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.176381	23.10
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823618	25.48
TES Alpha 0.11, Beta 0.05, Gamma 0.00	17.369210	28.88
SES Alpha 0.01	36.796019	63.88
DES Alpha 0.10, Beta 0.10	37.056912	64.02
SimpleAverage	53.460350	94.93
DES Alpha 0.16, Beta 0.16	70.572197	120.25
NaiveModel	79.718559	145.10

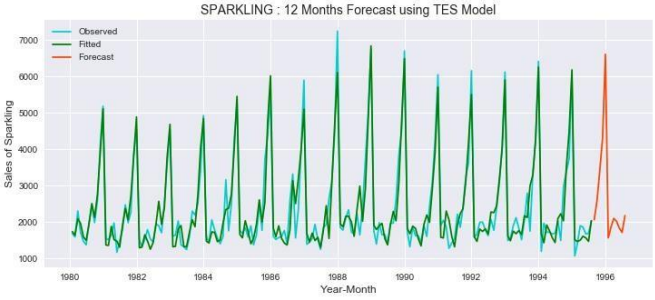
- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data
- 2 point trailing moving average is found to be having the best fitment against the test data, through with a lag of 2 and falling short at times
- Both SARIMA and TES forecasts are a bit higher than the actuals at any given point in time

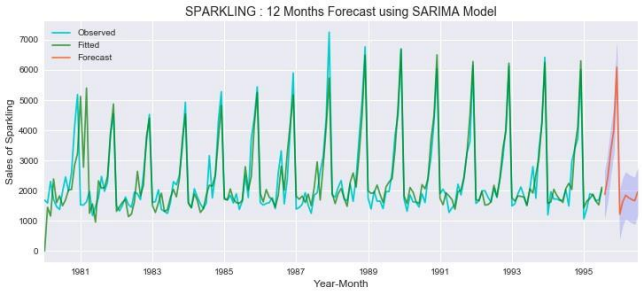
17. Predict into Future

- Based on the overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) and SARIMA are selected for final prediction into 12 months in future
- TES model **alpha: 0.4, beta: 0.1 and gamma: 0.2 & trend: 'additive', seasonal: 'multiplicative'** 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
- The 12 month prediction of the TES model is as below.

Sparkling Wine

The SARIMA model is built with parameters (3, 1, 3)x(1, 1, 2, 12), is found to be the most optimal SARIMA model

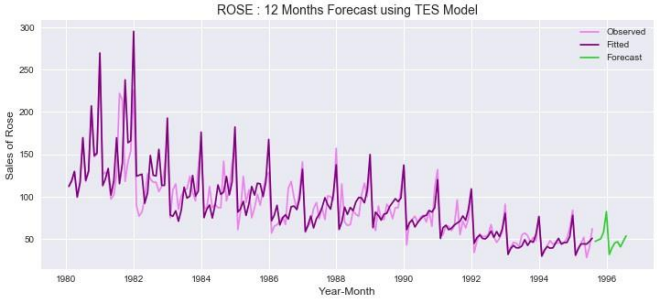


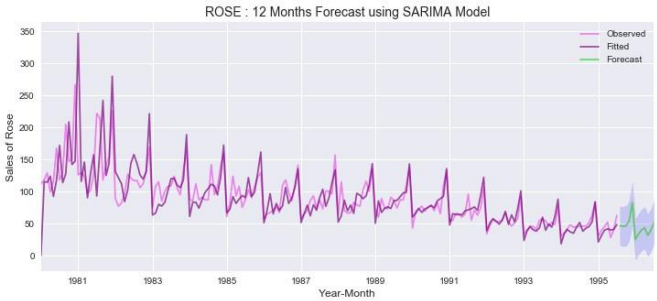


	RMSE	MAPE
TES Forecast	376.821	11.30
SARIMA Forecast	591.238	14.86

Rose Wine

The SARIMA model is built with parameters (4, 1, 1)x(0, 1, 1, 12), is found to be the most optimal SARIMA model for the complete time-series





	RMSE	MAPE
TES Forecast	20.881	14.48
SARIMA Forecast	30.676	19.40

18. Final Model

Sparkling Wine

Dep. Variable:		Sparkling	No. Observations:		187
Model:	SARIMAX(3, 1, 3)x(1, 1, 2, 12)		Log Likelihood		-1078.437
Date:	Sun, 13 Sep 2020		AIC		2176.875
Time:	20:24:33		BIC		2206.711
Sample:	01-31-1980		HQIC		2188.998
	- 07-31-1995				
Covariance Type:	opg				

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4227	0.086	-4.913	0.000	-0.591	-0.254
ar.L2	-0.9092	0.053	-17.290	0.000	-1.012	-0.806
ar.L3	0.1426	0.087	1.641	0.101	-0.028	0.313
ma.L1	-0.4114	0.078	-5.283	0.000	-0.564	-0.259
ma.L2	0.4623	0.083	5.584	0.000	0.300	0.625
ma.L3	-0.9674	0.104	-9.333	0.000	-1.171	-0.764
ar.S.L12	-0.0701	0.709	-0.099	0.921	-1.459	1.319
ma.S.L12	-0.4550	0.721	-0.631	0.528	-1.868	0.958
ma.S.L24	-0.0811	0.397	-0.205	0.838	-0.858	0.696
sigma2	1.461e+05	1.04e-06	1.4e+11	0.000	1.46e+05	1.46e+05

Ljung-Box (Q):	17.11	Jarque-Bera (JB):	35.59
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.72	Skew:	0.66
Prob(H) (two-sided):	0.26	Kurtosis:	5.03

Rose Wine

Dep. Variable:		Rose	No. Observations:		187
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)		Log Likelihood		-664.135
Date:	Sun, 13 Sep 2020		AIC		1342.270
Time:	20:24:35		BIC		1363.796
Sample:	01-31-1980		HQIC		1351.011
	- 07-31-1995				
Covariance Type:	opg				

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0914	0.084	1.093	0.274	-0.072	0.255
ar.L2	-0.1077	0.077	-1.393	0.164	-0.259	0.044
ar.L3	-0.1314	0.076	-1.729	0.084	-0.280	0.018
ar.L4	-0.1071	0.078	-1.375	0.169	-0.260	0.046
ma.L1	-0.8270	0.055	-14.901	0.000	-0.936	-0.718
ma.S.L12	-0.5963	0.059	-10.122	0.000	-0.712	-0.481
sigma2	232.4248	24.359	9.542	0.000	184.682	280.168

Ljung-Box (Q):	35.39	Jarque-Bera (JB):	5.30
Prob(Q):	0.68	Prob(JB):	0.07
Heteroskedasticity (H):	0.22	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.89

19. Recommendations

Sparkling Wine

	Sparkling
1995-08-31	1873.20
1995-09-30	2445.15
1995-10-31	3312.82
1995-11-30	3994.61
1995-12-31	6084.20
1996-01-31	1216.50
1996-02-29	1640.54
1996-03-31	1847.26
1996-04-30	1762.31
1996-05-31	1708.44
1996-06-30	1663.80
1996-07-31	1961.50

The model forecasts sale of **29510** units of Sparkling wine in 12 months into future. Which is an average sale of **2459 units per month**

The seasonal sale in December 1995 will hit a maximum of **6084 units**, before it drops to the lowest sale in January 1996; at **1216 units**.

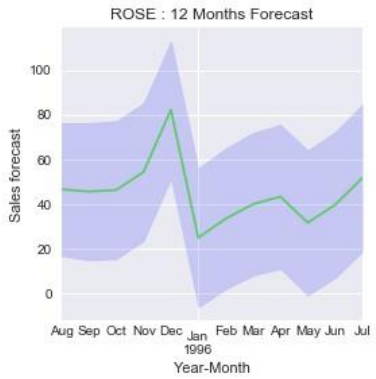
The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995 (October, November and December), which is a total of 13,392 units of sparkling wine is expected to be sold.

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

Adding more exogenous variable into the time series data can improve forecasts

Rose Wine



	ROSE
1995-08-31	46.54
1995-09-30	45.51
1995-10-31	46.23
1995-11-30	54.32
1995-12-31	82.21
1996-01-31	24.81
1996-02-29	33.35
1996-03-31	39.87
1996-04-30	43.23
1996-05-31	31.53
1996-06-30	39.56
1996-07-31	51.70

The model forecasts sale of 539 units of Rose wine in 12 months into future. Which is an average sale of 45 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 25 units.

Unlike Sparkling wine, Rose wine sells very low number of units and the standard deviation is only 14.5. Which means that higher demand does not impact procurement and production

Apart from higher sale in November and December months, Rose sales will be above average in the summer months of July and August

The winery should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions