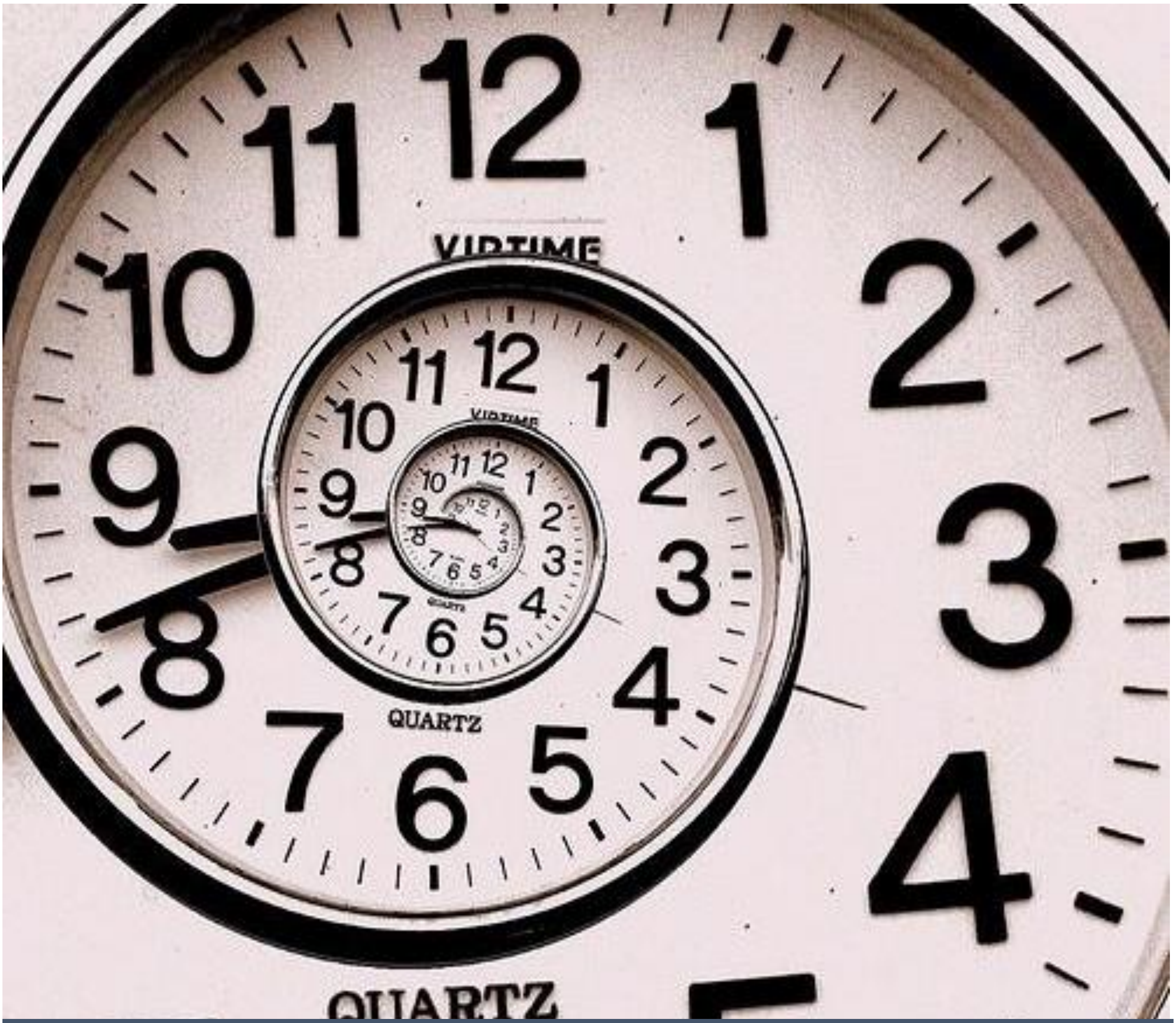


Time series forecasting

Project:

Nitin kumar 9/25/22 Dsba project

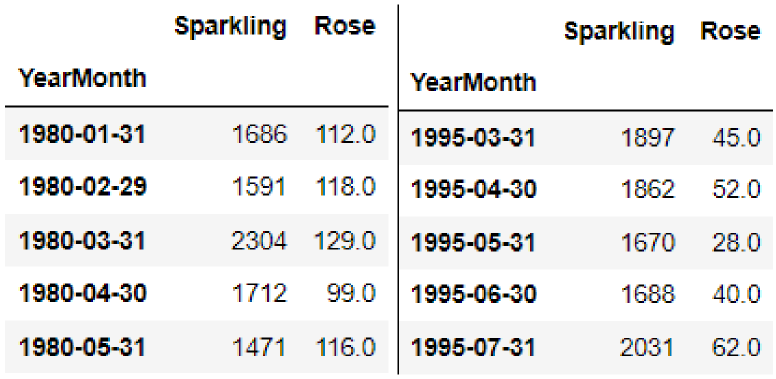
Contents

Problem:

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

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

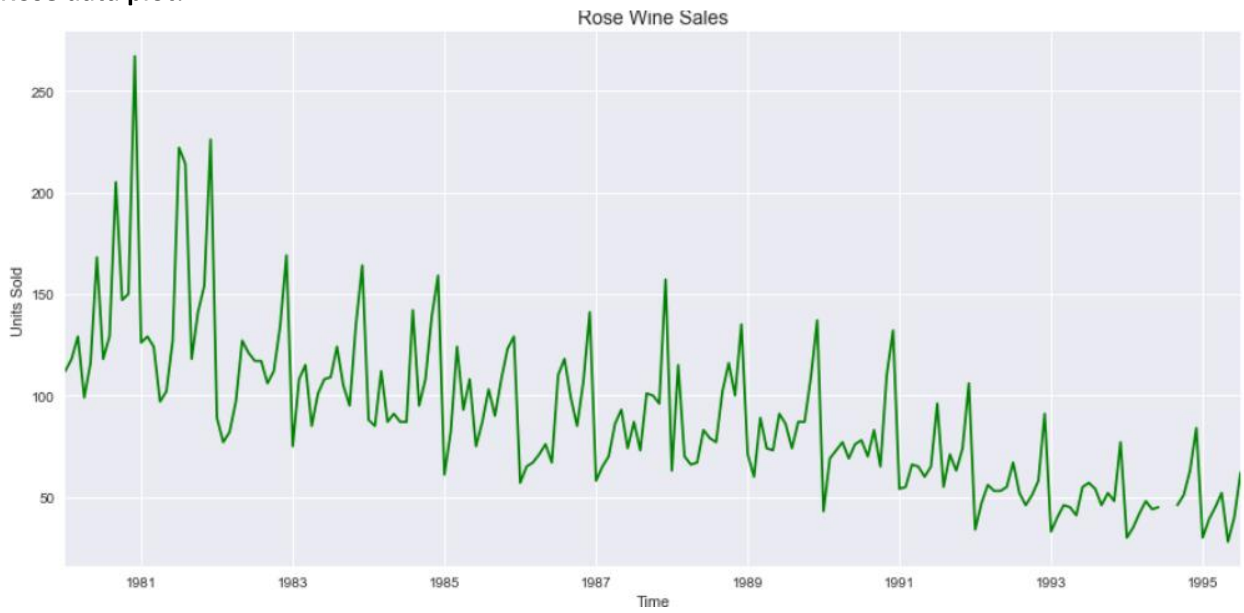
- Data set are read as time series data suing parse_date=True & index_col='YearMonth'
- Frist 5 rows and bottom 5 rows of both the data are given below



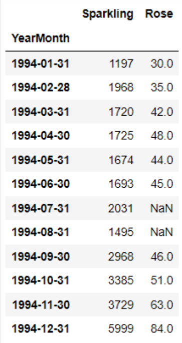
YearMonth	Sparkling	Rose
1980-01-31	1686	112.0
1980-02-29	1591	118.0
1980-03-31	2304	129.0
1980-04-30	1712	99.0
1980-05-31	1471	116.0

YearMonth	Sparkling	Rose
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

Rose data plot: -



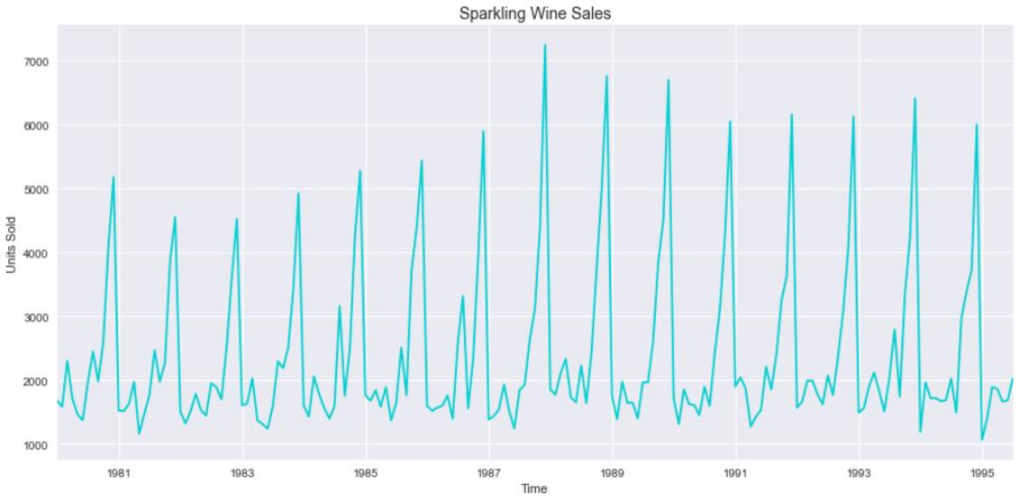
- From the plot we can see that there are missing values in the data set



YearMonth	Sparkling	Rose
1994-01-31	1197	30.0
1994-02-28	1968	35.0
1994-03-31	1720	42.0
1994-04-30	1725	48.0
1994-05-31	1674	44.0
1994-06-30	1693	45.0
1994-07-31	2031	NaN
1994-08-31	1495	NaN
1994-09-30	2968	46.0
1994-10-31	3385	51.0
1994-11-30	3729	63.0
1994-12-31	5999	84.0

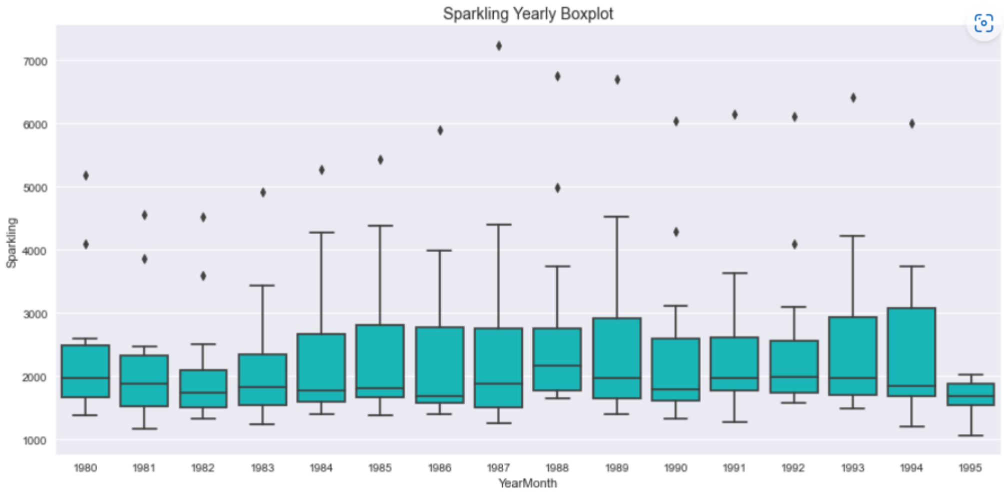
there are two missing values in the data set

Sparkling Wine data plot: -

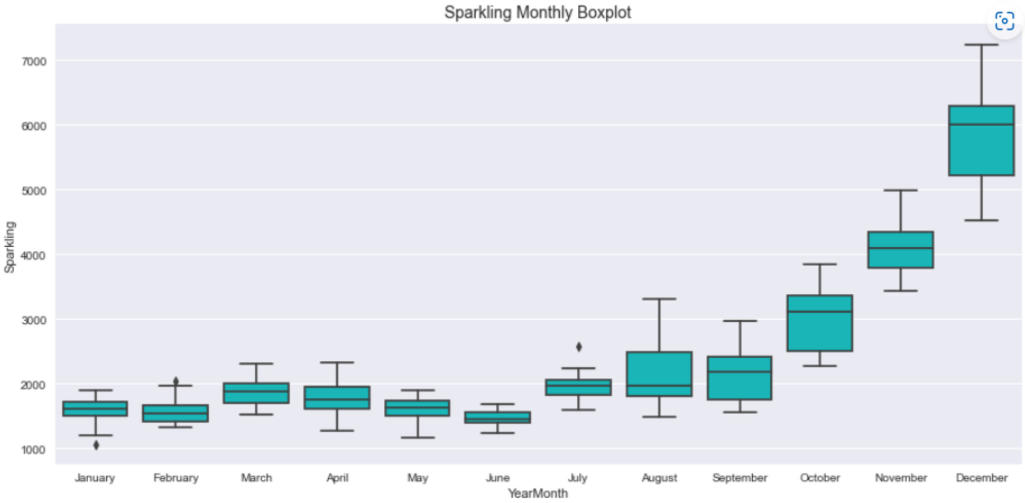


2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Sparkling Yearly Boxplot



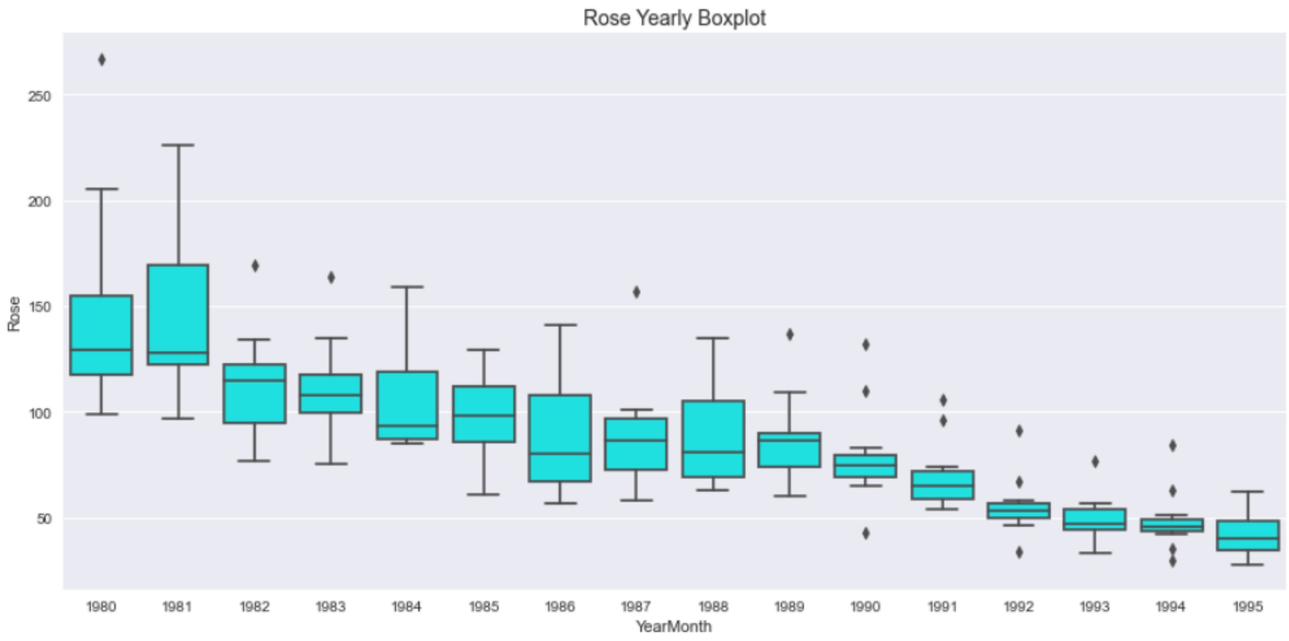
Sparkling Monthly Boxplot



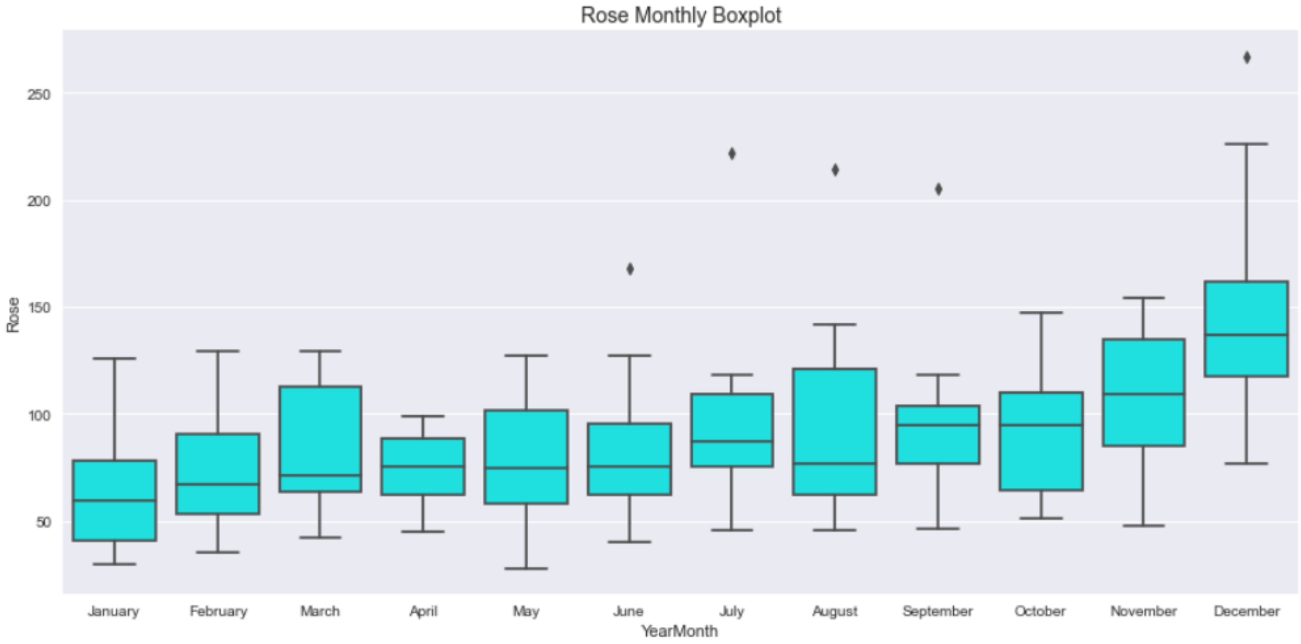
- From the above plot we Can see that there are more sales from October to December

- This spike is more due to holiday season in starting from October

Rose Yearly Boxplot

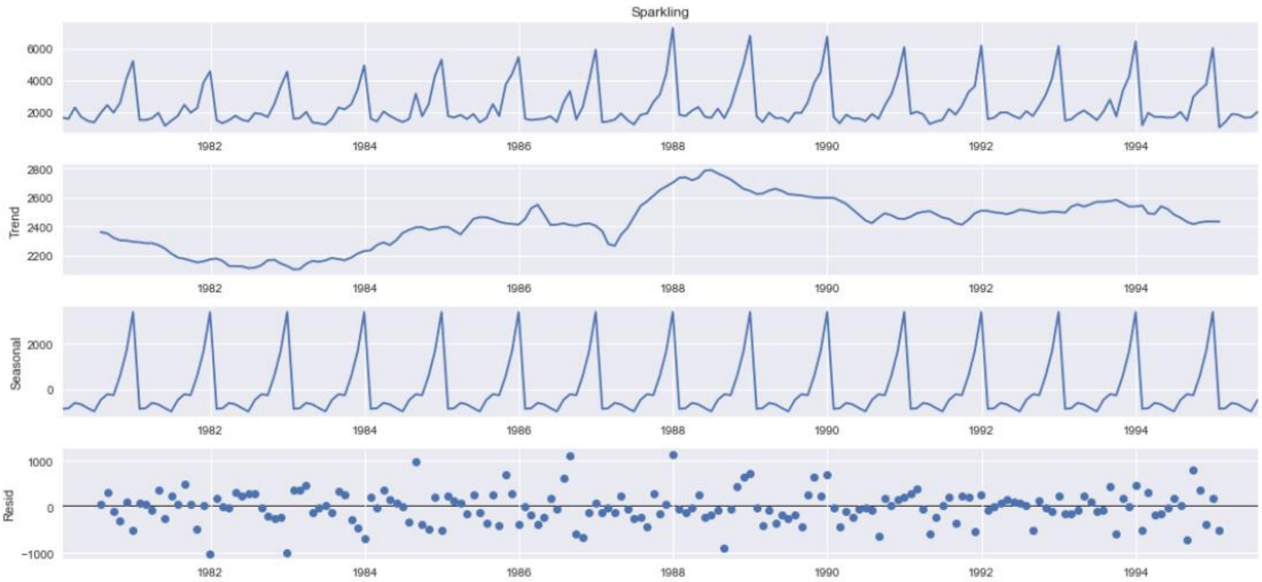


Rose Monthly Boxplot

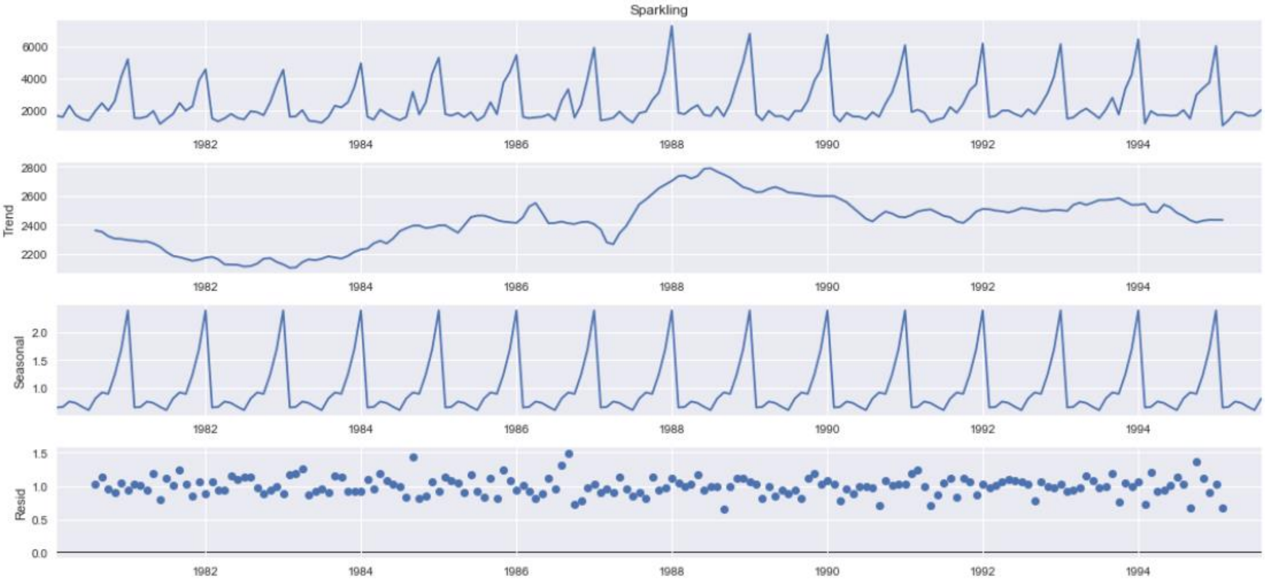


- From the above plot we Can see that there are more sales from October to December
- This spike is more due to holiday season in starting from October

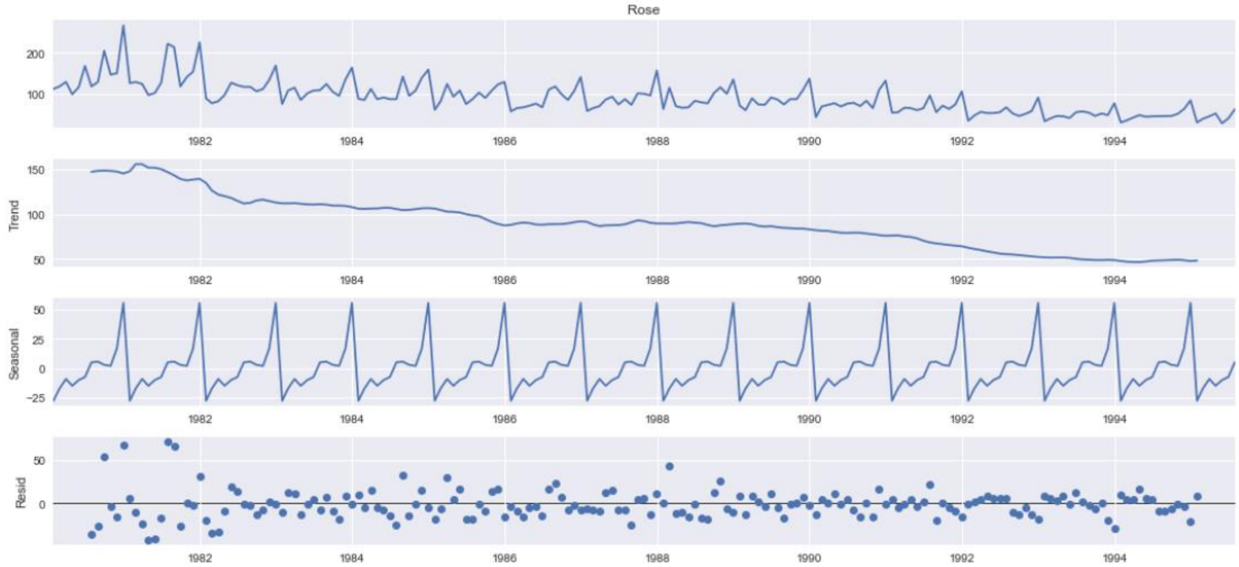
Addictive decomposition of sparkling: -



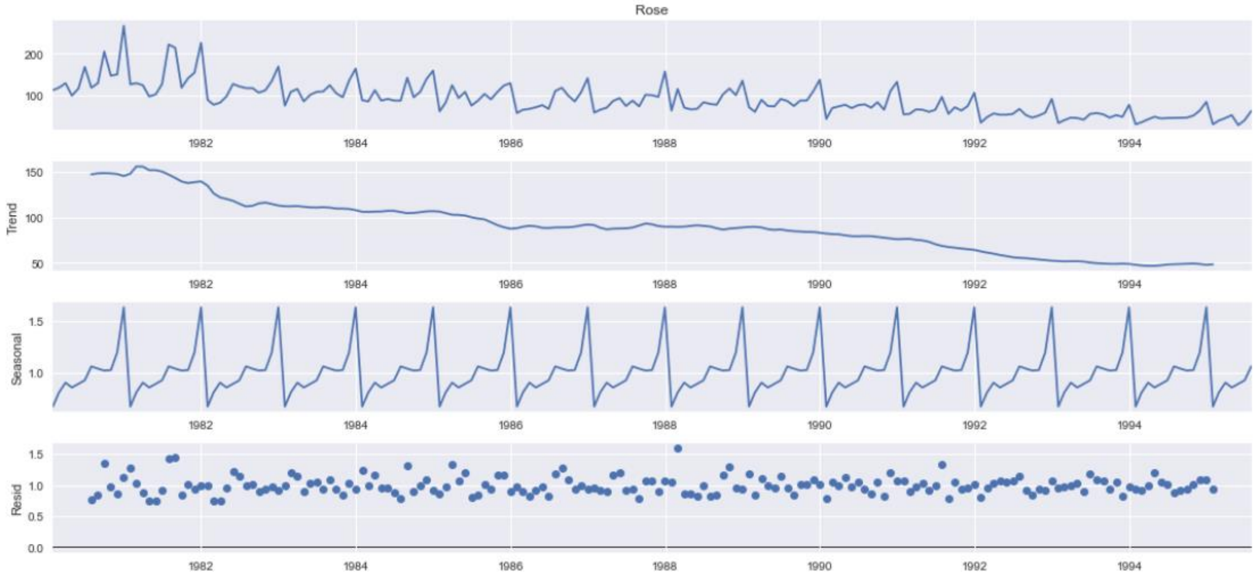
Multiplicative decomposition of sparkling: -



Addictive decomposition of Rose -



Multiplicative decomposition of rose: -



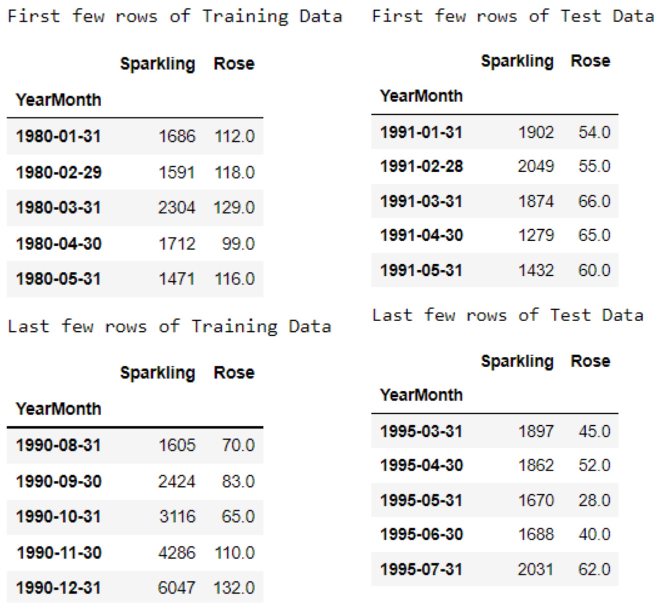
Hear by observing the residual patterns of addictive and multiplicative models od rose and sparkling data set it seems that

Rose is multiplicative

Sparkling is addictive

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

- Both datasets of rose and sparkling data set are split at the year 1991
- Test data set starts at 1991



```
First few rows of Training Data
```

YearMonth	Sparkling	Rose
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 few rows of Training Data
```

YearMonth	Sparkling	Rose
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

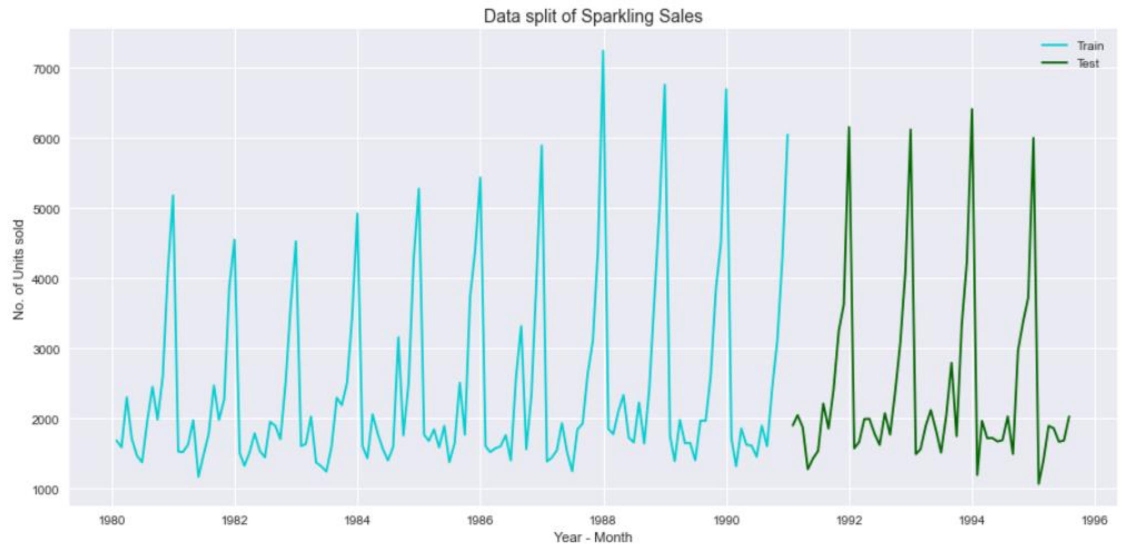
```
First few rows of Test Data
```

YearMonth	Sparkling	Rose
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

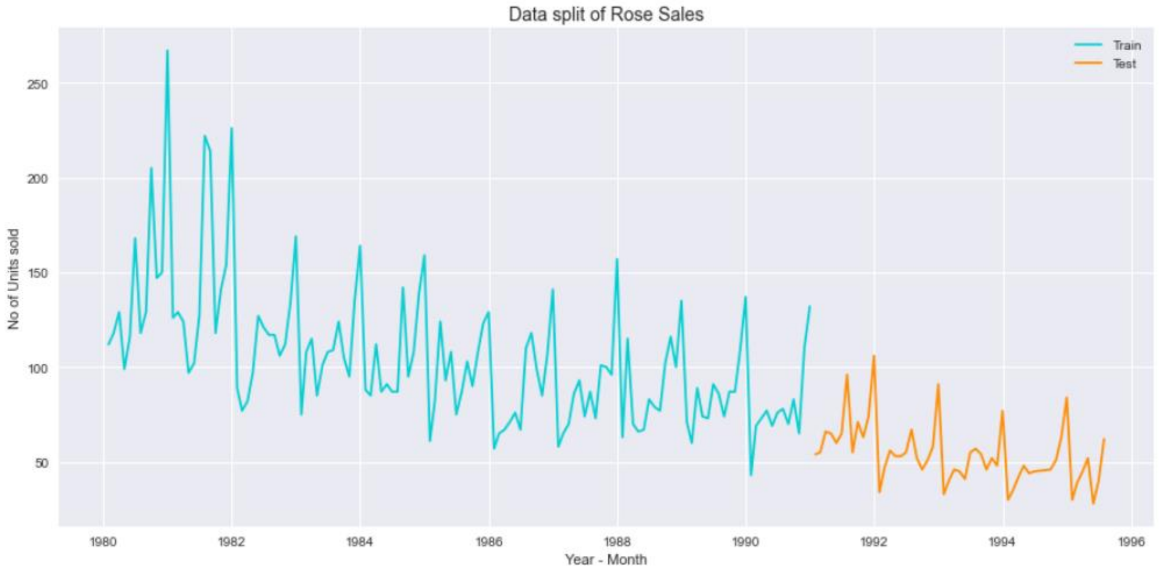
```
Last few rows of Test Data
```

YearMonth	Sparkling	Rose
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

- Data split for sparkling sales

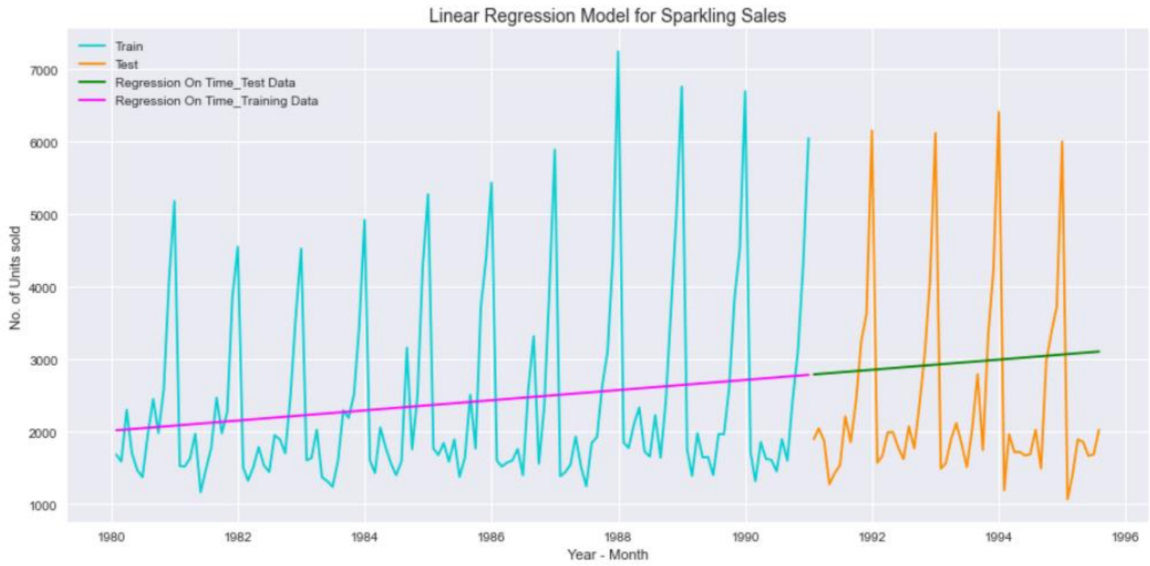


- Data split for rose

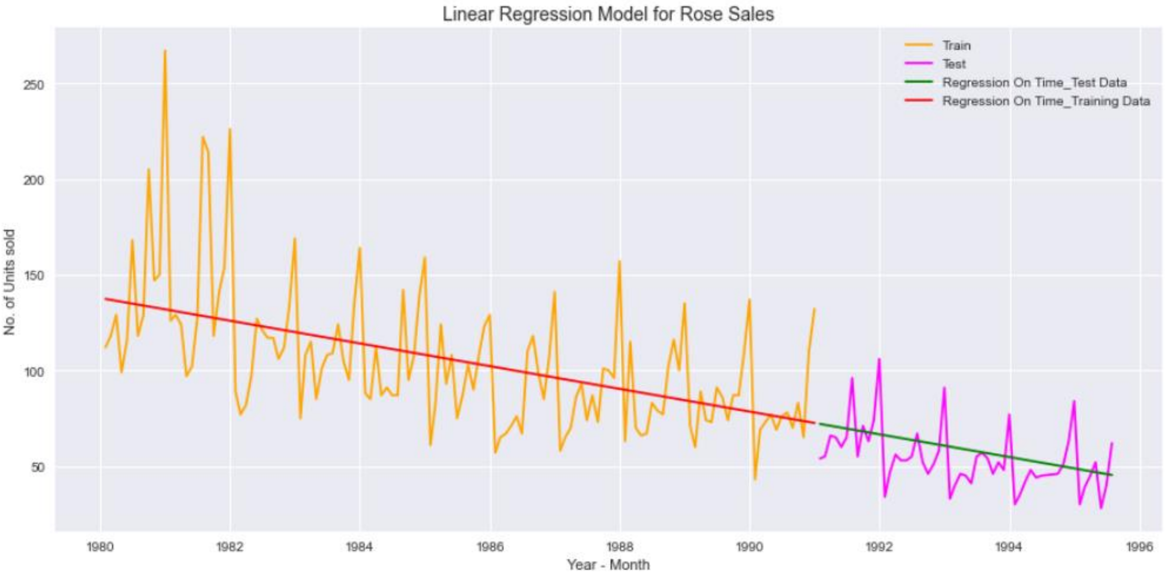


4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on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

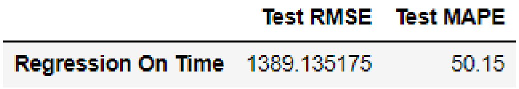
Linear regression model for sparkling sales



Linear regression model for rose sales

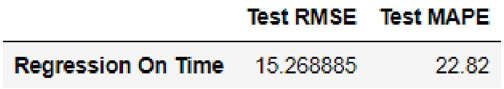


RMSE & MAPE for testing sparkling dataset



	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15

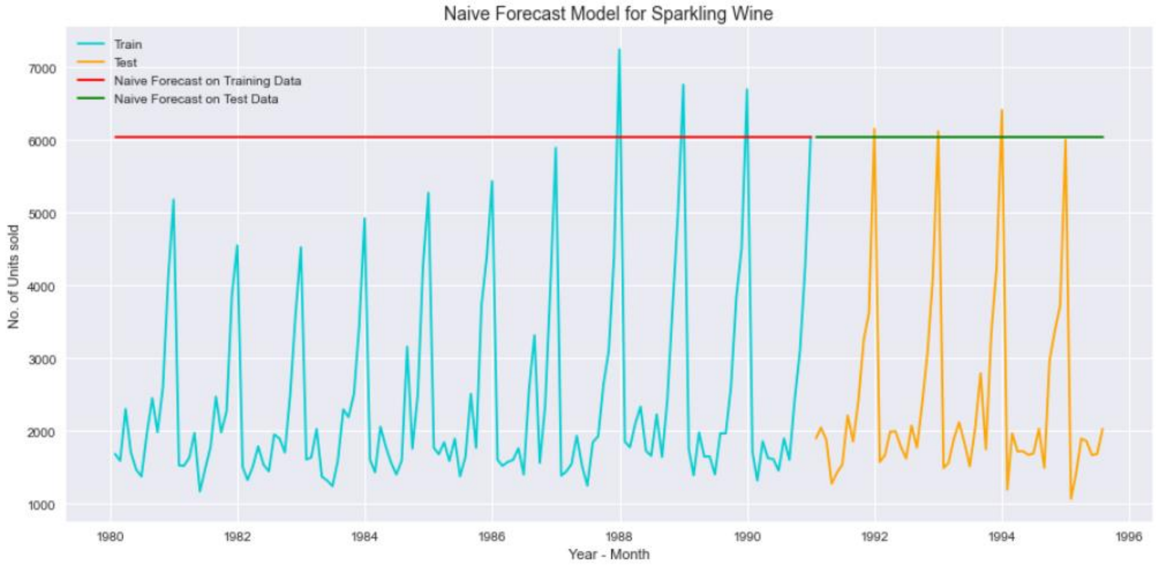
RMSE & MAPE for testing Rose dataset



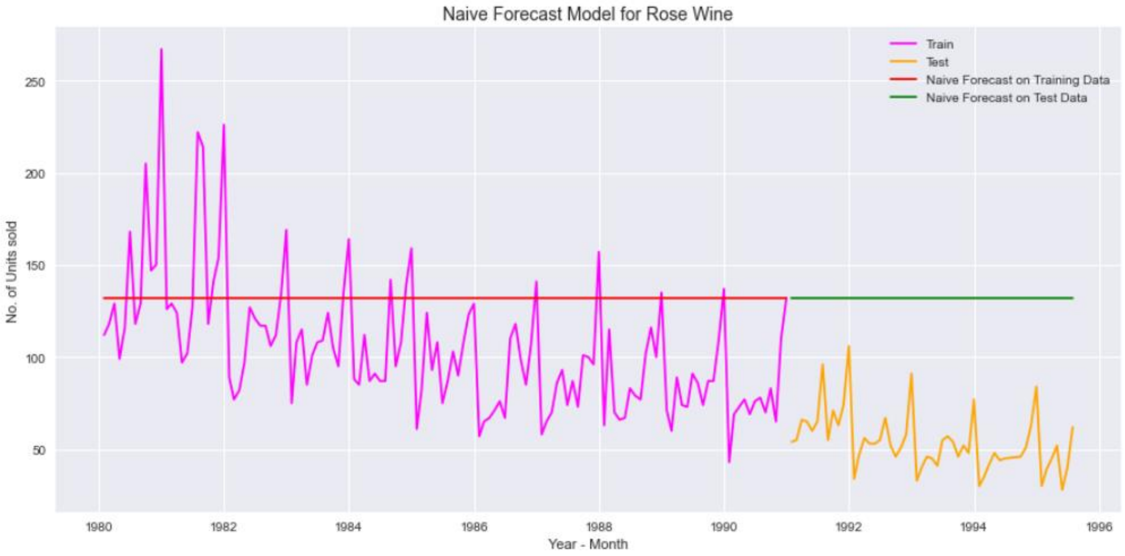
	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82

Model 3:-Naïve

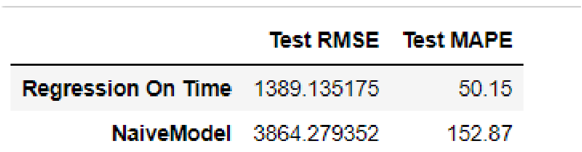
Naive Forecast Model for Sparkling Wine



Naive Forecast Model for Rose Wine

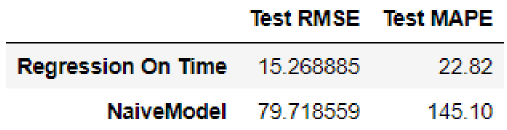


Sparkling Testing Data - RMSE and MAPE



	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15
NaiveModel	3864.279352	152.87

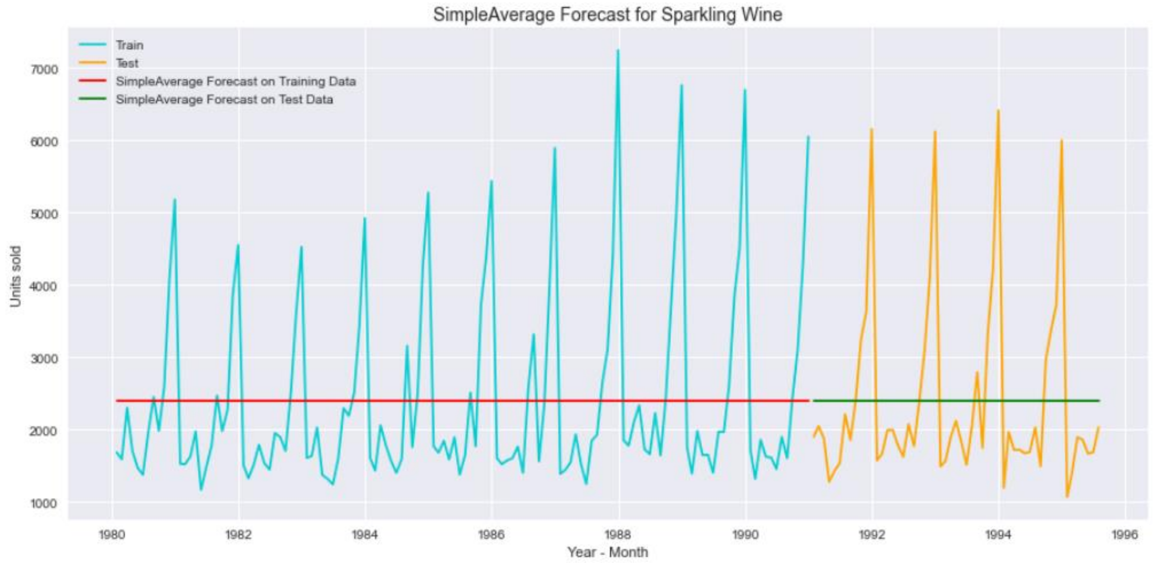
Rose Testing Data - RMSE and MAPE



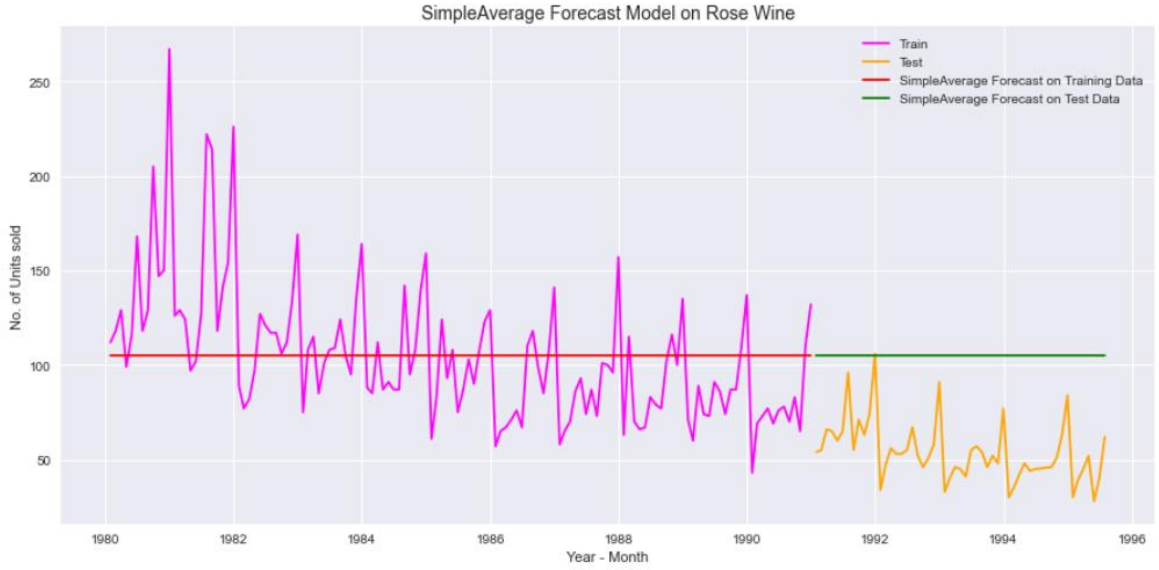
	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10

Model 3: Simple Average

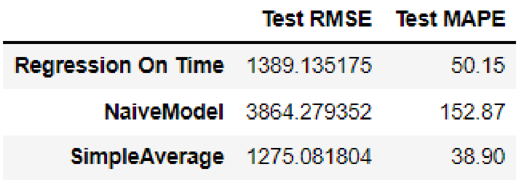
SimpleAverage Forecast for Sparkling Wine



SimpleAverage Forecast Model on Rose Wine

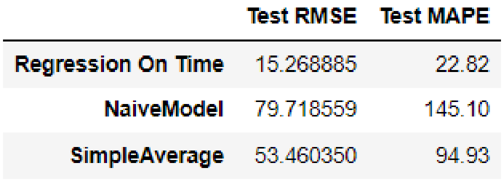


sparkling Testing Data - RMSE and MAPE



	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90

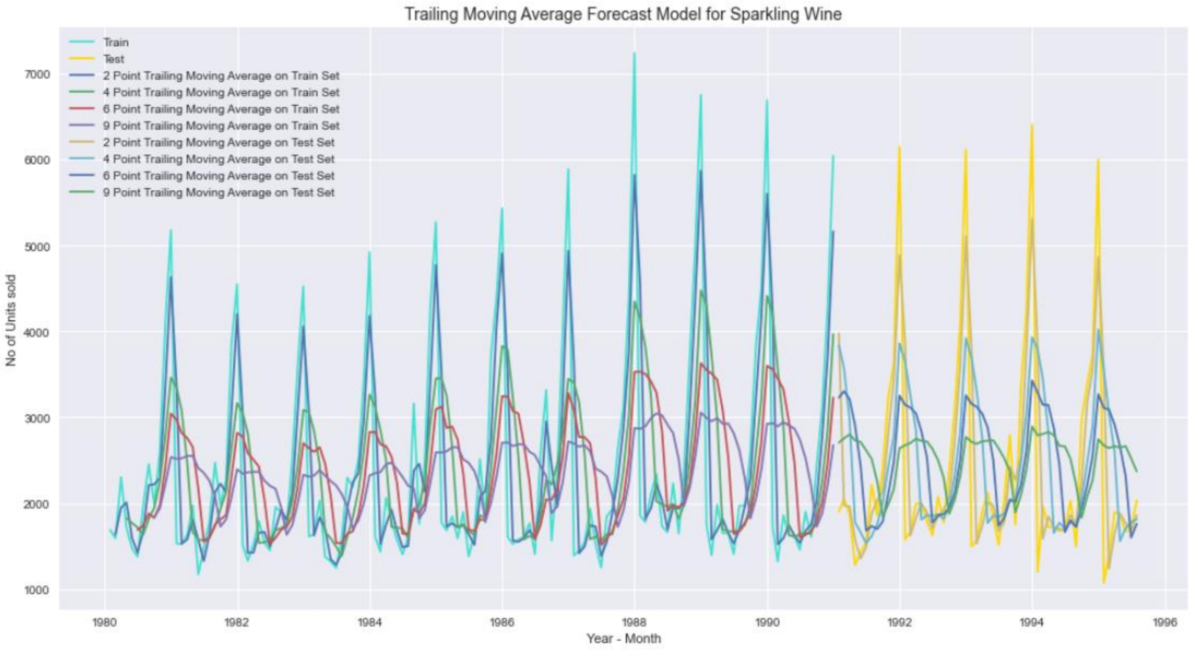
Rose Testing Data - RMSE and MAPE



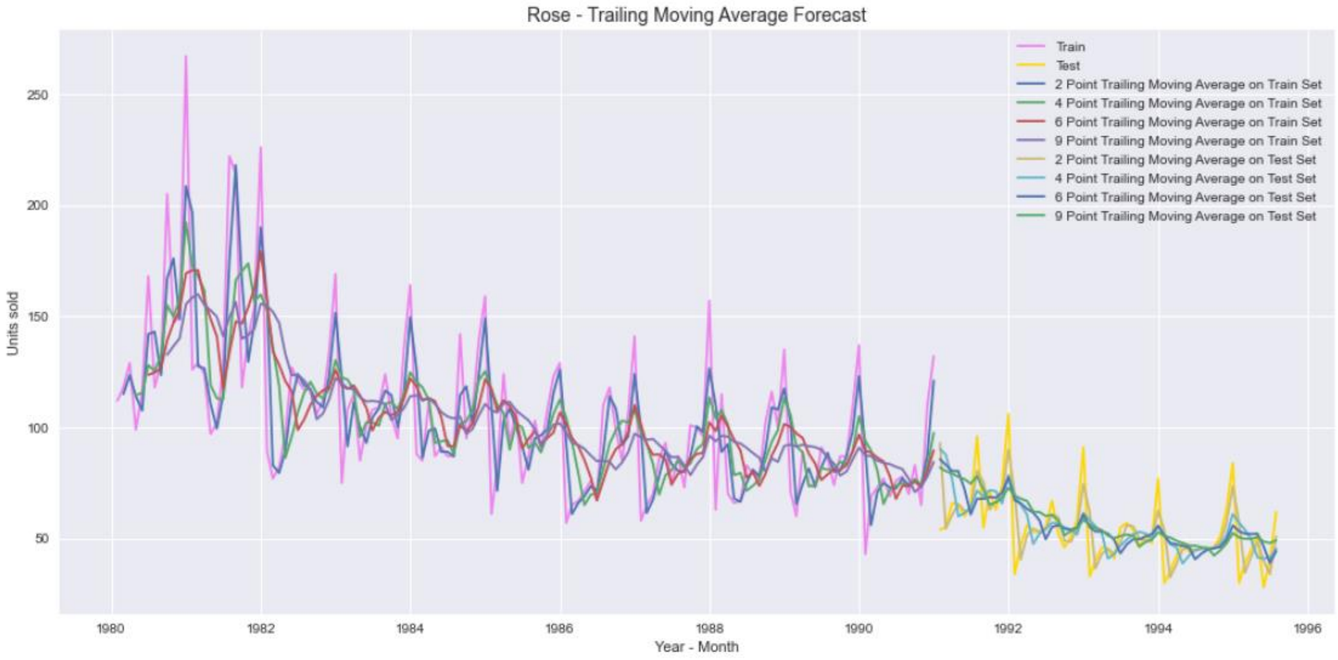
	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93

Model 4: Moving Average

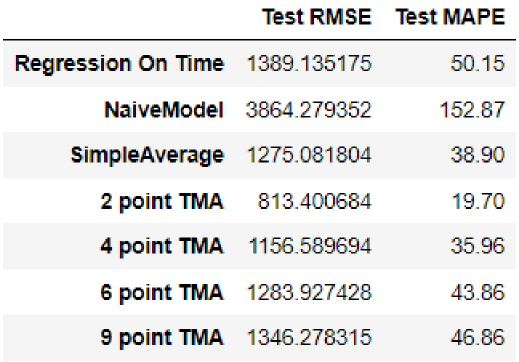
Trailing Moving Average Forecast Model for Sparkling Wine



Rose - Trailing Moving Average Forecast

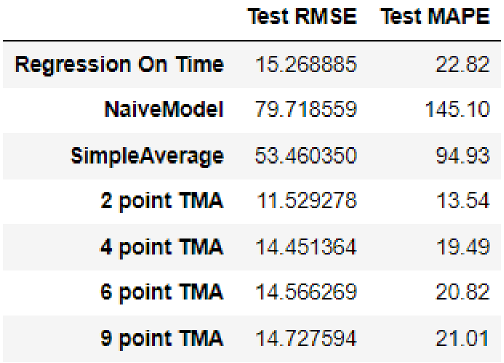


For Sparkling Wine Test Data



	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86

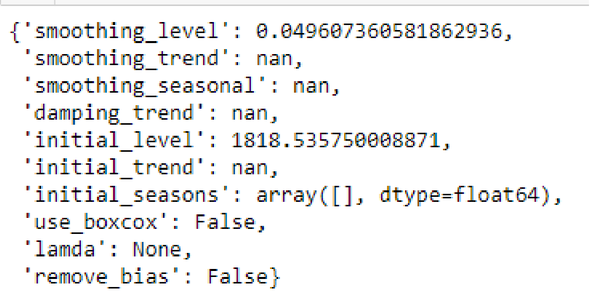
For Sparkling Wine Test Data



	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
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

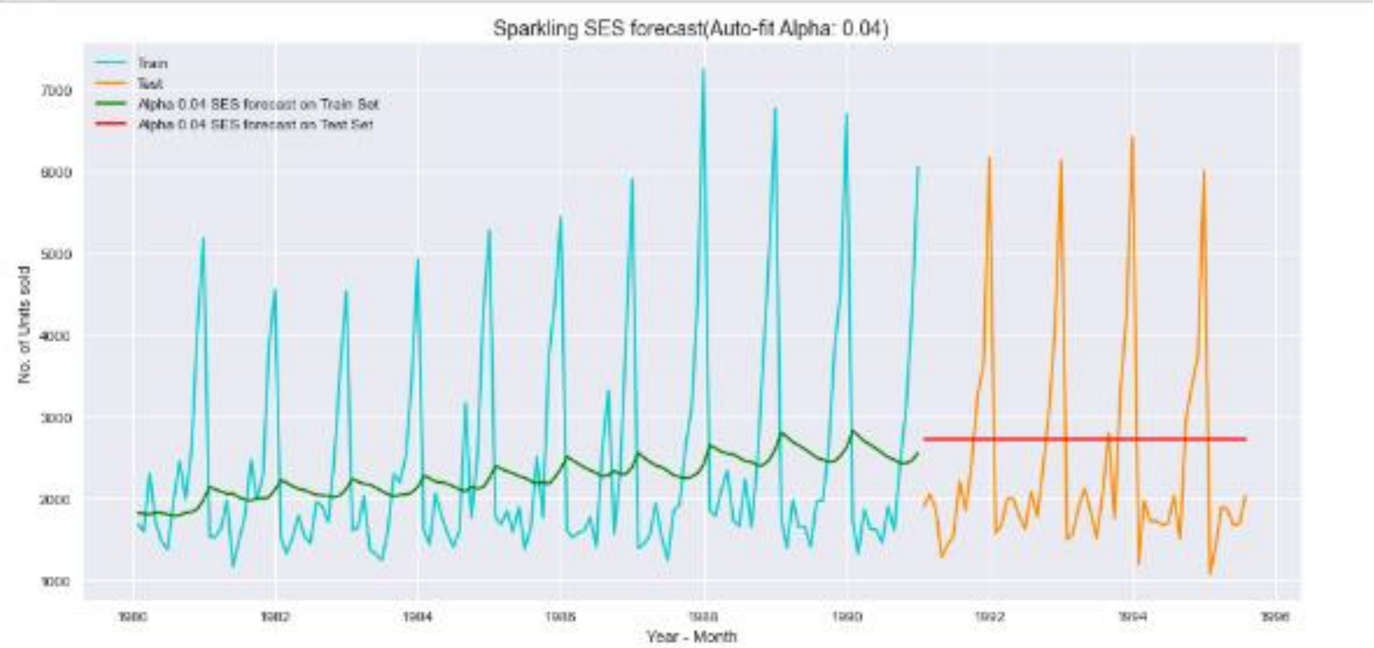
Model 5: Simple Exponential Smoothing

model_SES_autofit.params

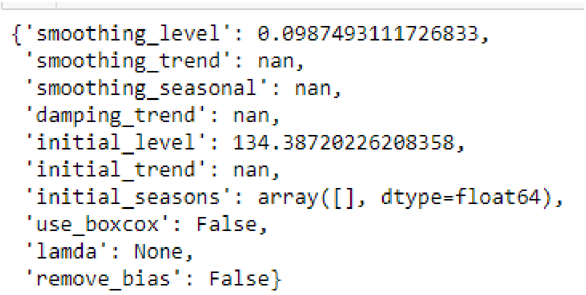


```
{'smoothing_level': 0.049607360581862936,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1818.535750008871,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Sparkling SES forecast(Auto-fit Alpha: 0.04

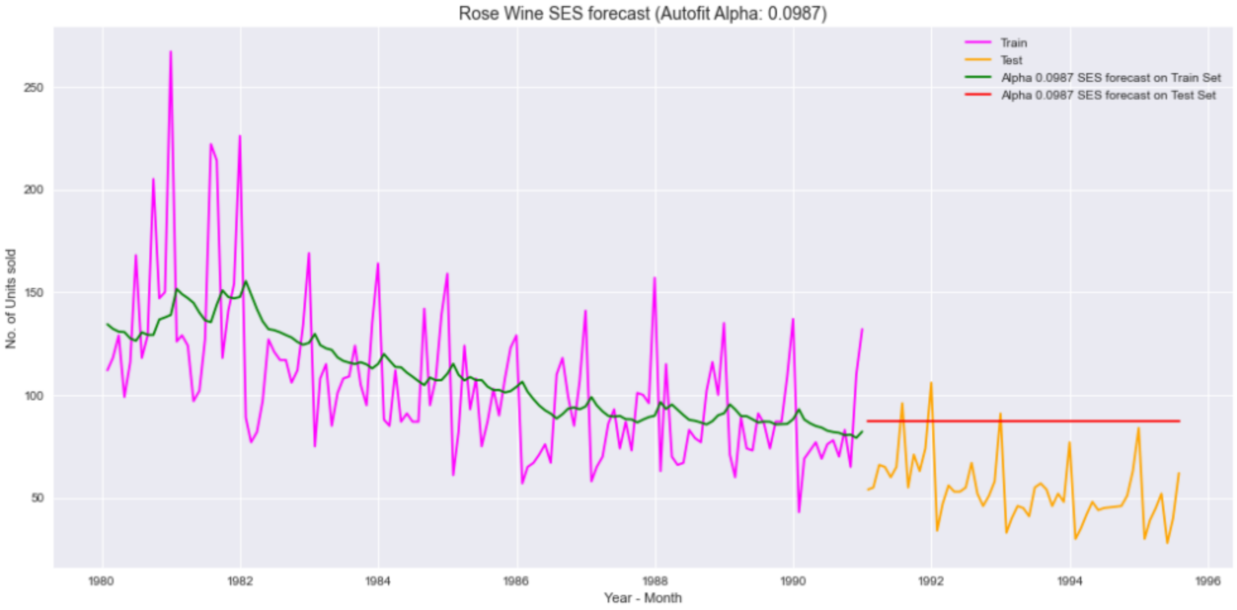


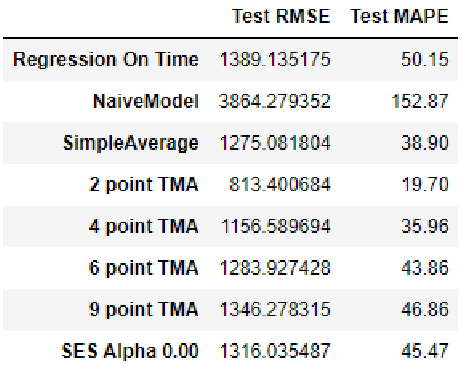
For Rose Wine



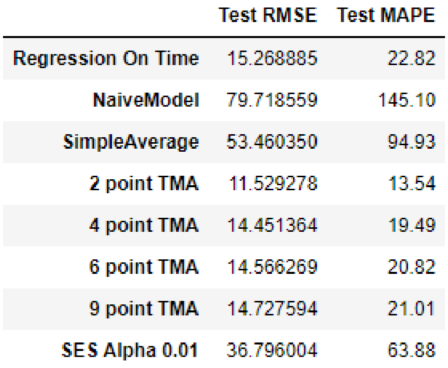
```
{'smoothing_level': 0.0987493111726833,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38720226208358,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Rose Wine SES forecast (Autofit Alpha: 0.0987)





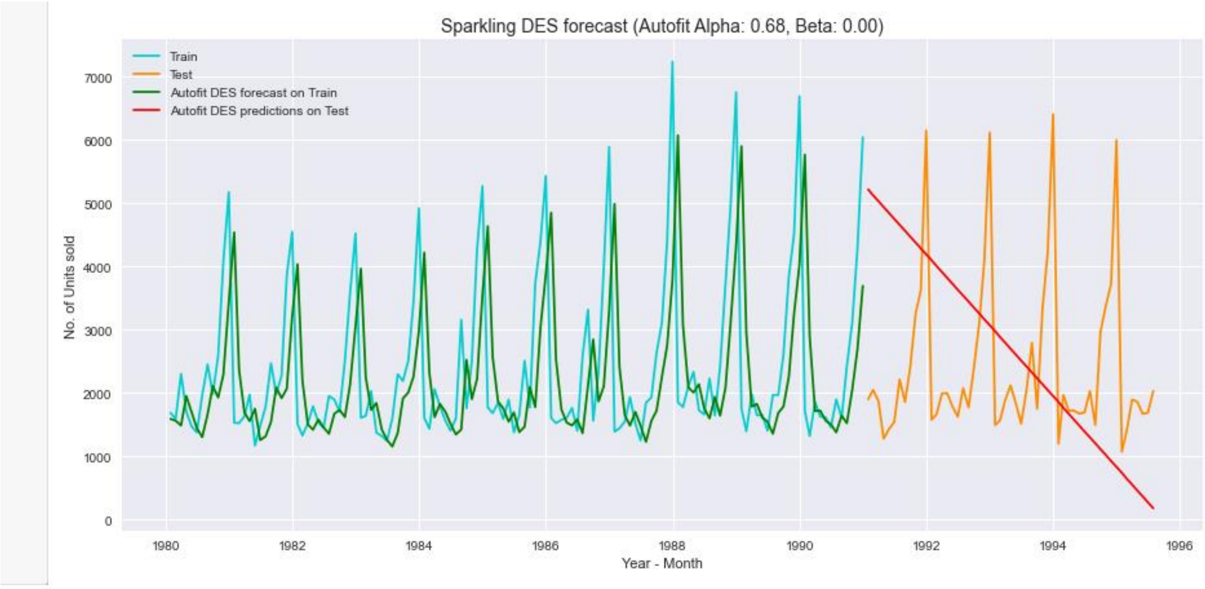
	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.00	1316.035487	45.47



	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
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
SES Alpha 0.01	36.796004	63.88

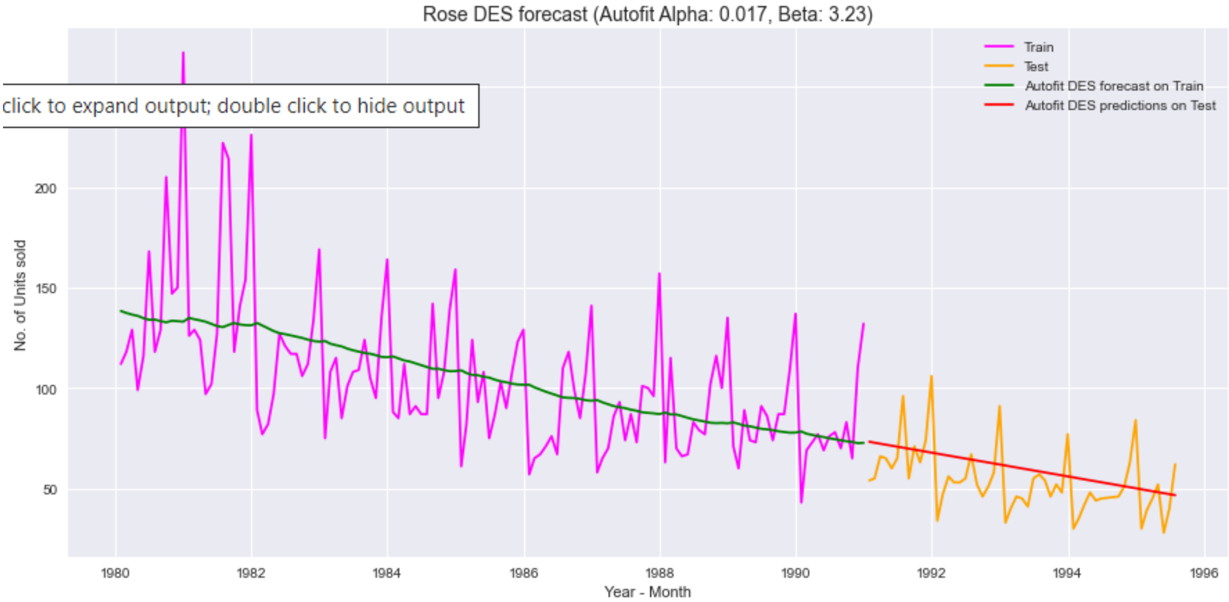
Model 6: Double Exponential Smoothing (Holt's Model)

Sparkling DES forecast (Autofit Alpha: 0.68, Beta: 0.00

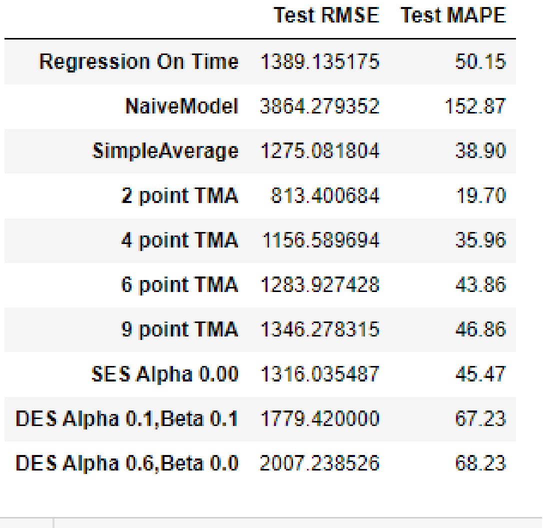


For Rose Wine

Rose DES forecast (Autofit Alpha: 0.017, Beta: 3.23)

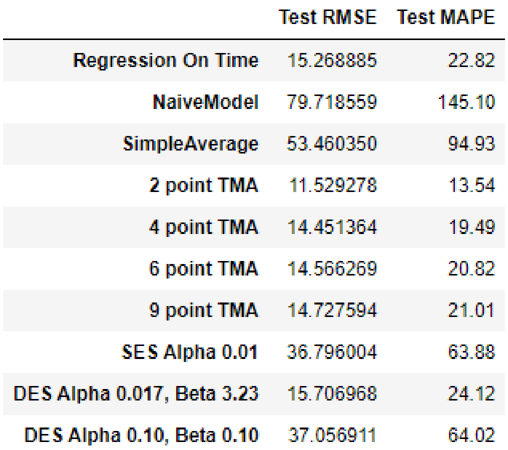


Model Evaluation for sparking data set



	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.00	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23

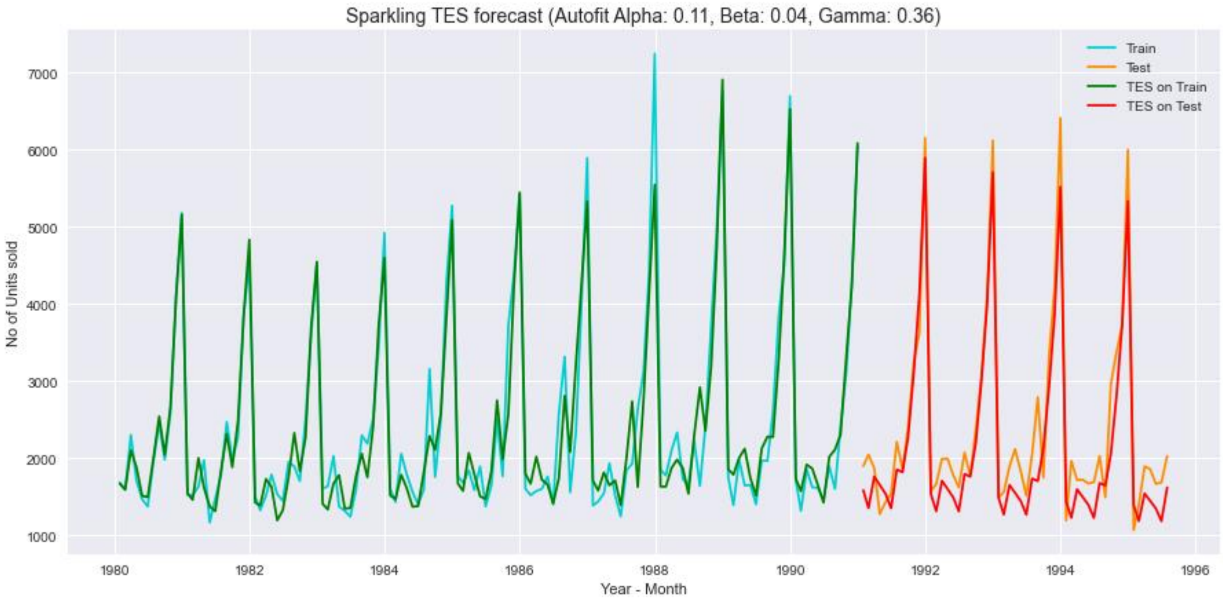
Model Evaluation for rose data set



	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
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
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.017, Beta 3.23	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02

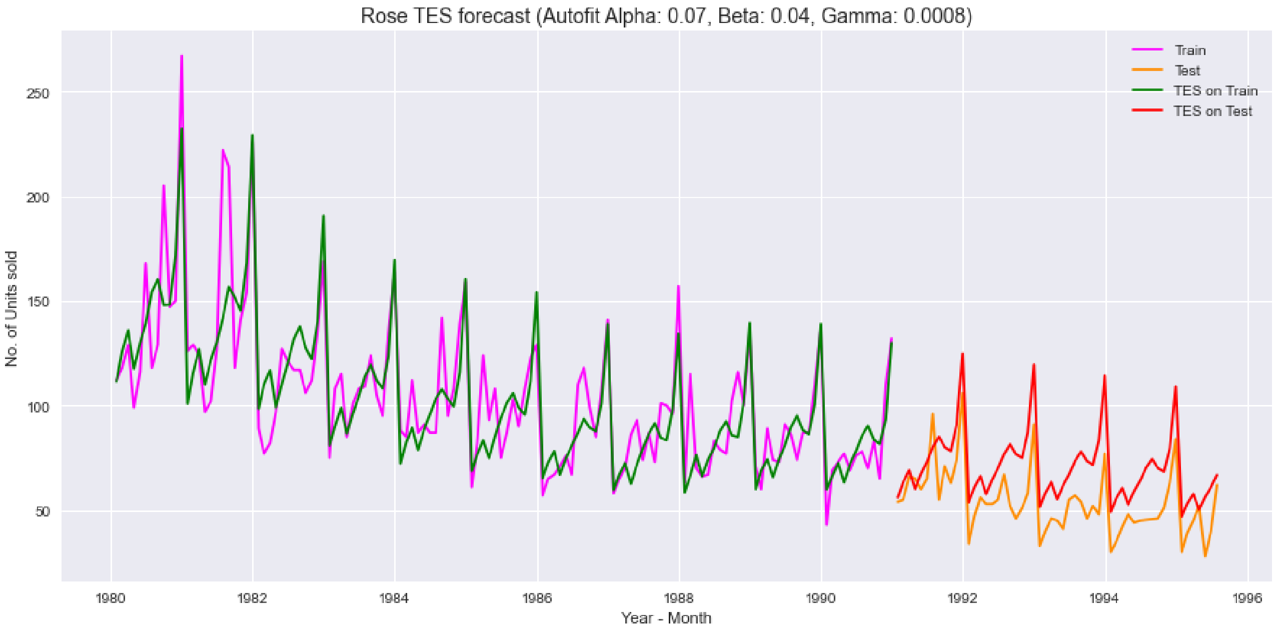
Model 7: Triple Exponential Smoothing (Holt - Winter's Model)

Sparkling TES forecast (Autofit Alpha: 0.11, Beta: 0.04, Gamma: 0.36)

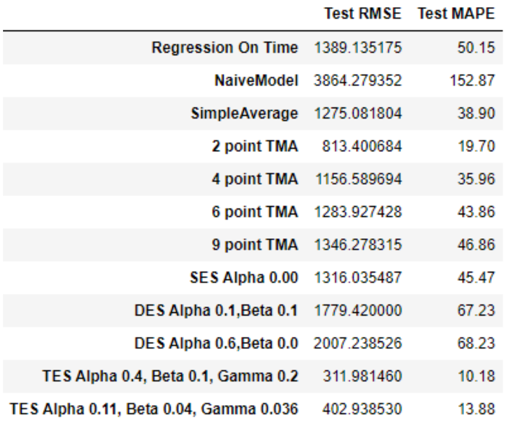


For Rose Wine

Rose TES forecast (Autofit Alpha: 0.07, Beta: 0.04, Gamma: 0.0008

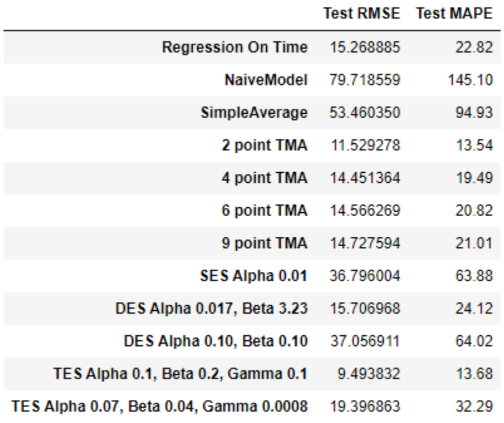


Model evaluation for sparkling data set



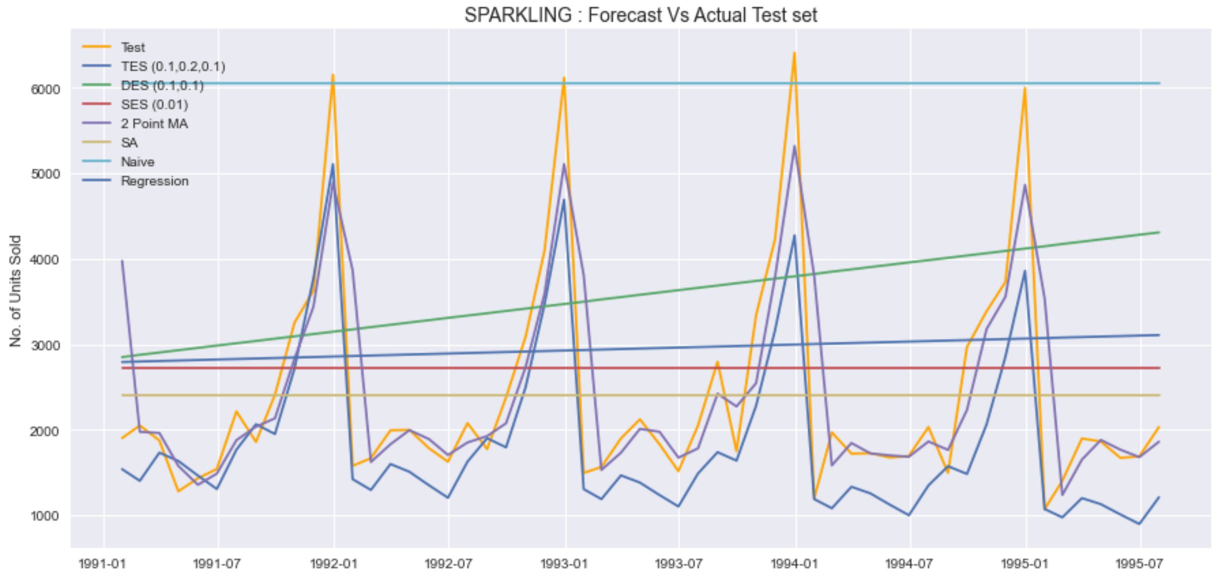
	Test RMSE	Test MAPE
Regression On Time	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverage	1275.081804	38.90
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
SES Alpha 0.00	1316.035487	45.47
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 0.0	2007.238526	68.23
TES Alpha 0.4, Beta 0.1, Gamma 0.2	311.981460	10.18
TES Alpha 0.11, Beta 0.04, Gamma 0.036	402.938530	13.88

Model evaluation for rose data set

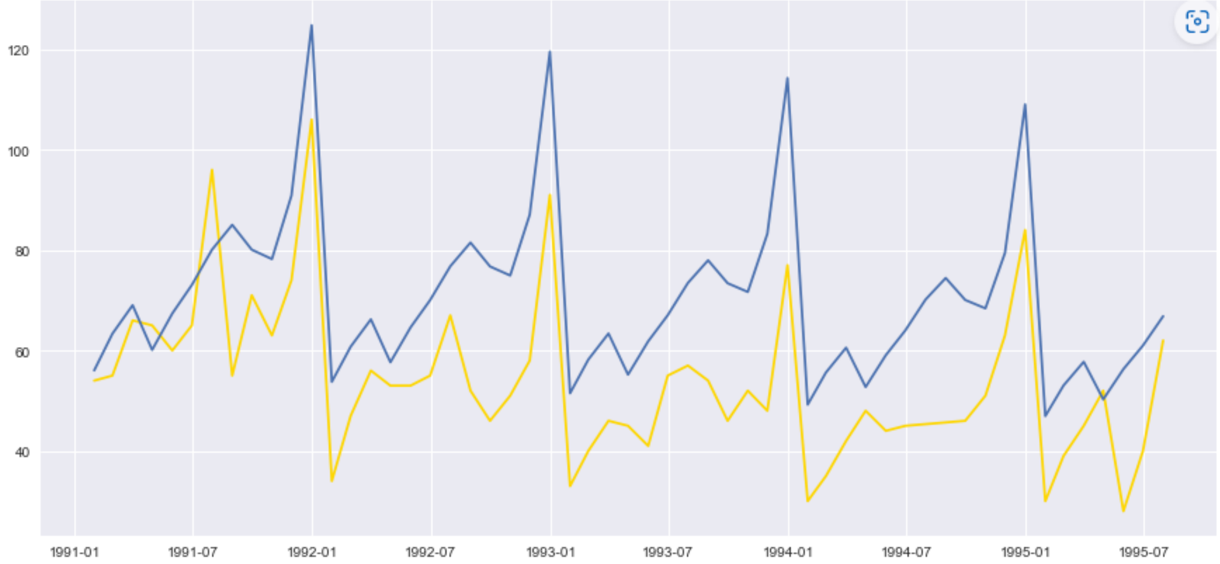


	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.017, Beta 3.23	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02
TES Alpha 0.1, Beta 0.2, Gamma 0.1	9.493832	13.68
TES Alpha 0.07, Beta 0.04, Gamma 0.0008	19.396863	32.29

SPARKLING : Forecast Vs Actual Test set



ROSE : Forecast Vs Actual Test Data



We fitted various models to the Train split and tested it on Test split. Accuracy metrics used is Root Mean Squared Error (RMSE) on Test data

Model 1 - Linear Regression (= B + B, X, + E)

We regressed variables 'Rose' and 'Sparkling against their individual time instances

We modified the datasets and tagged individual sales to their time instances

TEST RMSE ROSE = 15.27 1389.14 TEST RMSE SPARKLING =

Model 2. Naive Approach

Naive approach says that prediction for tomorrow is same as today

And, prediction for day-after is same as tomorrow

So, effectively all future predictions are going to be same as today

TEST RMSE ROSE = 79.72 3864.28

Model 3 - Simple Average 07 +1 = 1+2 = ...- = Means

All future predictions are the same as the simple average of all data till today

TEST RMSE ROSE = 53.46 1275.08 TEST RMSE SPARKLING

Model 4 - Moving Average (MA)

We calculate rolling means (Moving averages) over different intervals for the whole train data

2 Pt MA ====> means, we find average of 1st and 2nd to predict 3rd dhe similarly, average of 2nd and 3rd to predict 4th and so

4 Pt MA ====> means, we find average of 1st, 2nd, 3rd & 4th to predict 5th Time Series Project 2 PT MA ====> TEST RMSE ROSE = 11.53 TEST RMSE SPARKLING = 813.40

4 PT MA ====> TEST RMSE SPARKLING = 1156.59 TEST RMSE ROSE = 14.45

6 PT MA ====> | TEST RMSE SPARKLING = 1283.93 TEST RMSE ROSE = 14.57

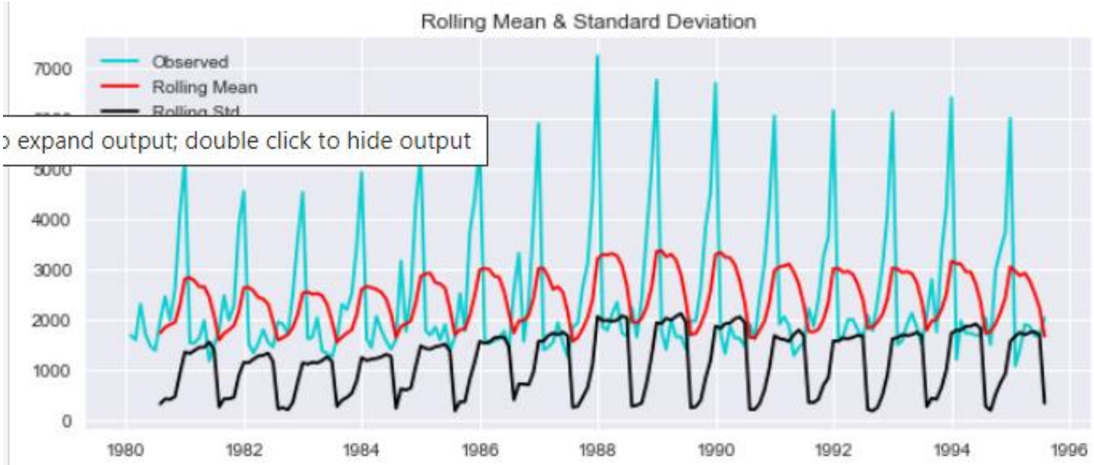
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.

Note: Stationarity should be checked at alpha = 0.05.

- To check for stationary data set we use augmentel dicky-fuller test to check the stationary of the data
- Hypothesis of ADF test:
  HO-time series is not stationary
  H1-time series is stationary
- Alpha =0.05
  If p value is leser then alpha we consider that time series is stationary
  If p value is greater then alpha we consider that the time series is not stationary

  Sparkling data set

  Original data

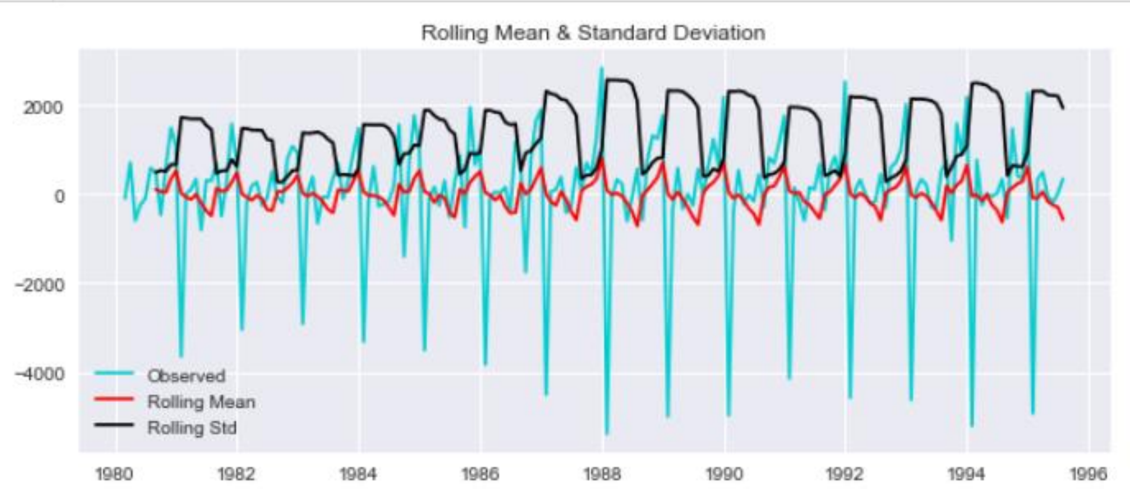


```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

hear we can see that alpha is greater then p value so we say that the time series for

original data set made stationary after adding lag 1

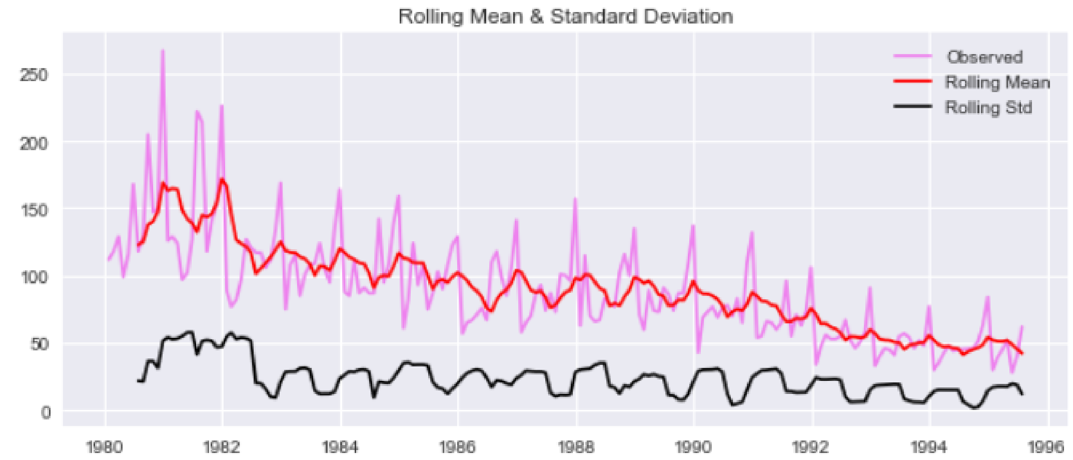
stationary data set



```
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
```

Rose data Set: -

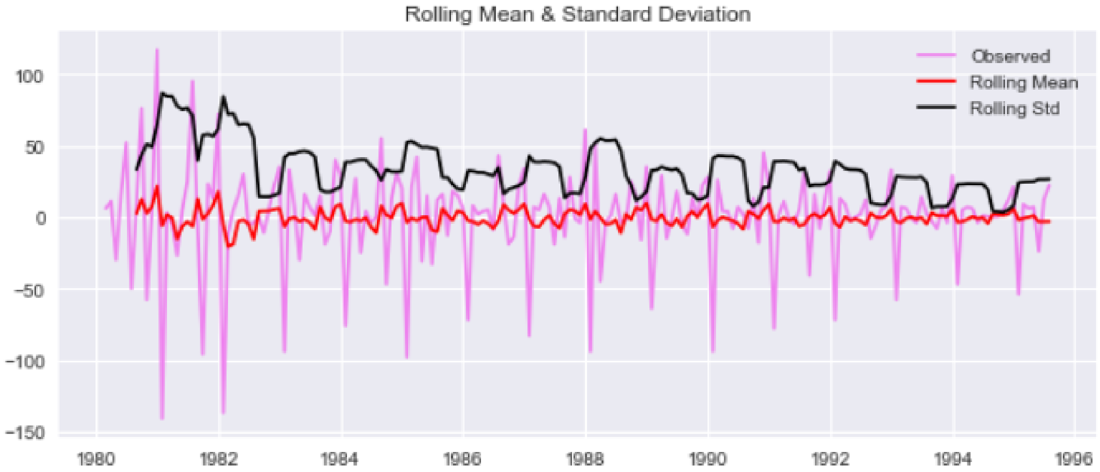
Original data set



```
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
dtype: float64
```

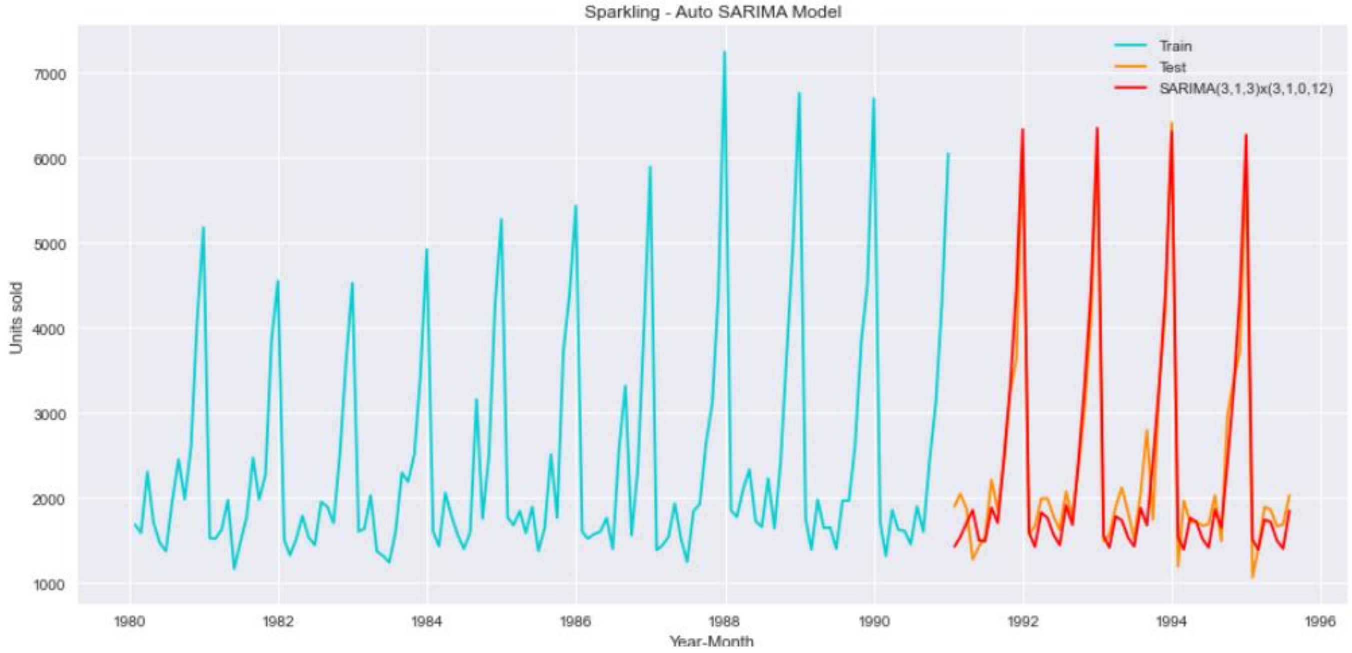
Original data set is not stationary

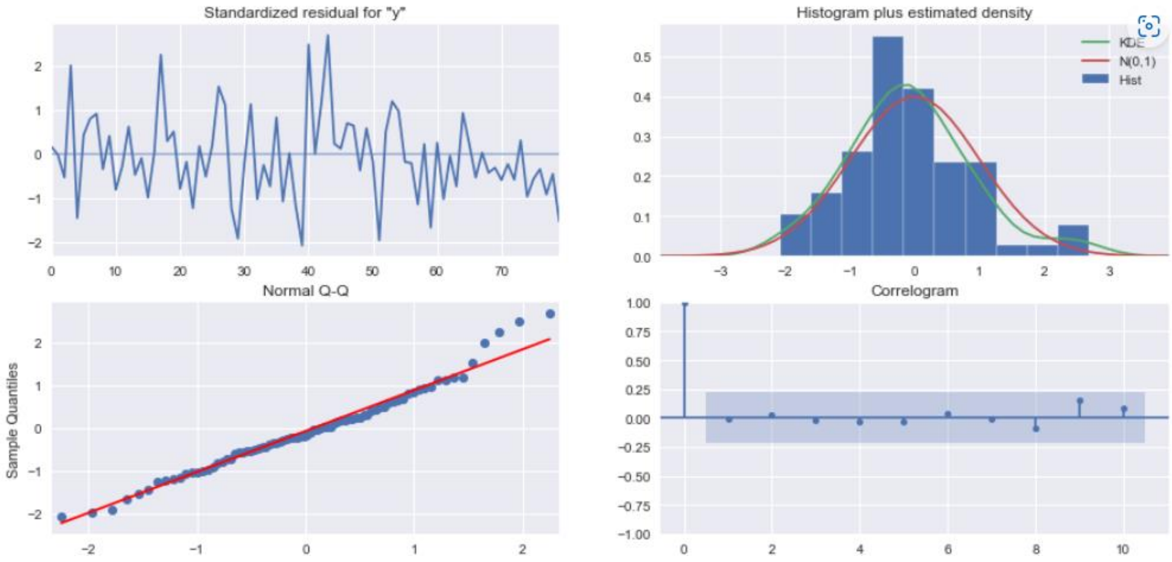
Stationary data set of rose



```
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
dtype: float64
```

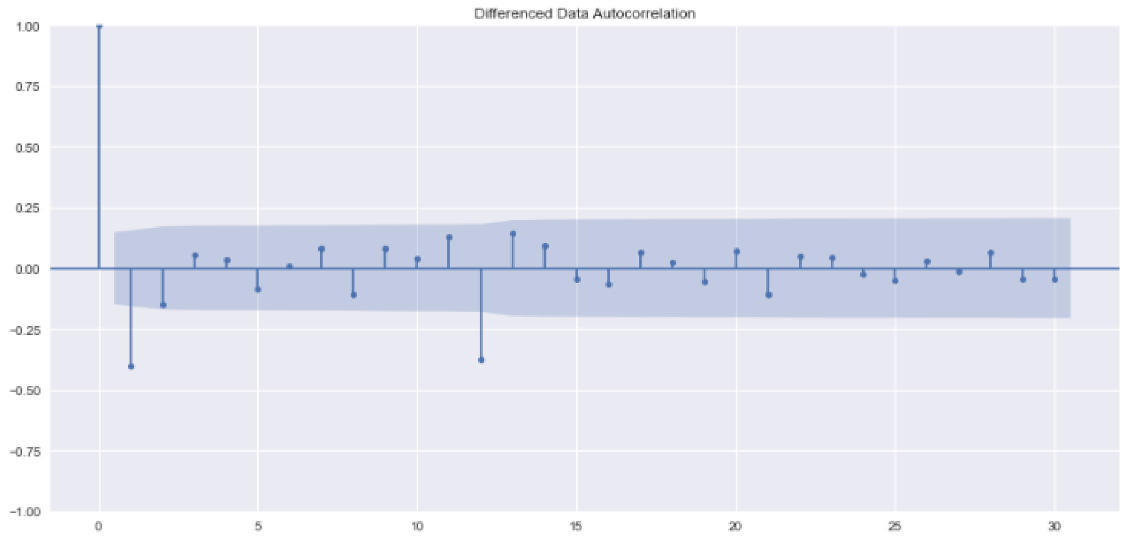
6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

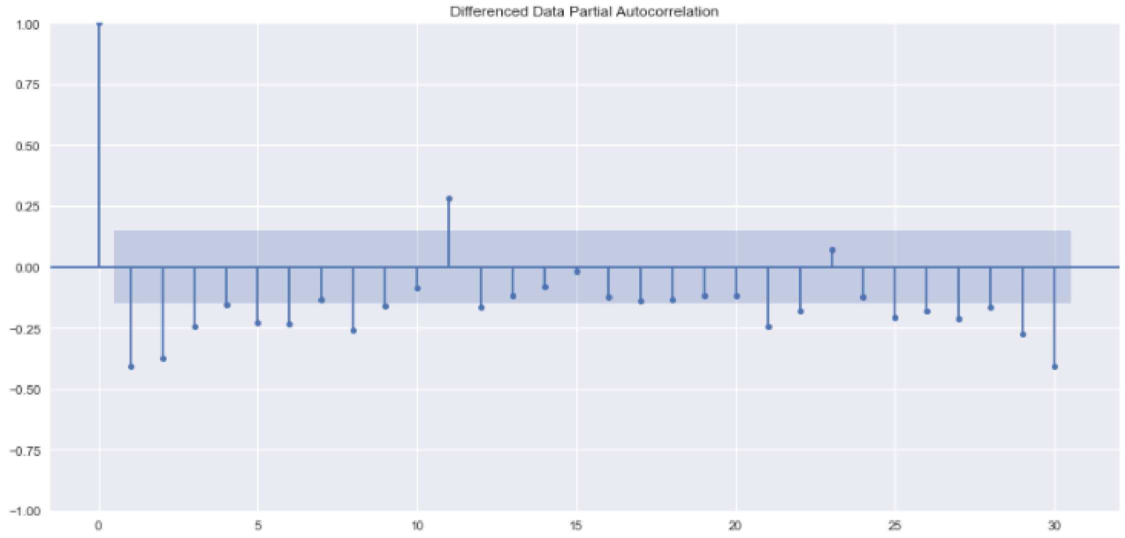




7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

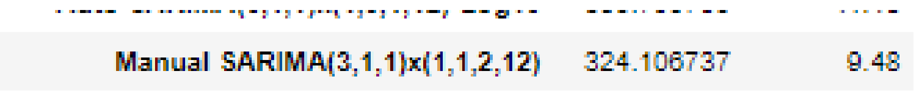
Manual spark arima on rose:-





For manual sparl arima cut off points are p=3,d=1,q=1

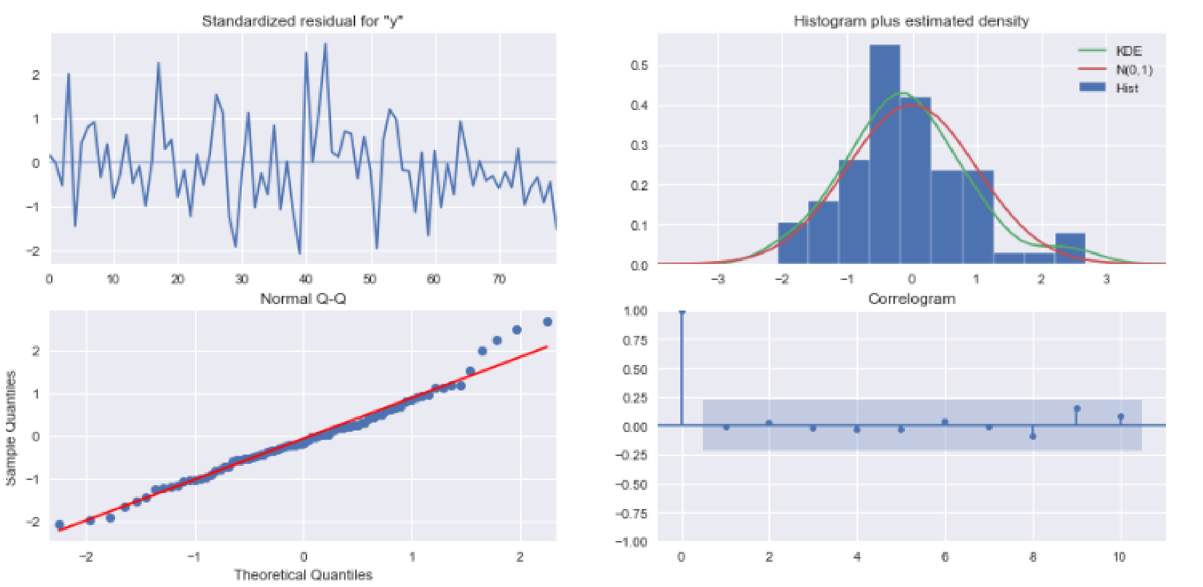
Test rmse is



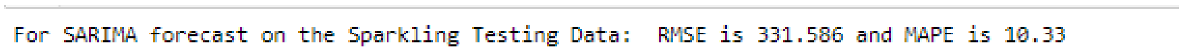
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106737	9.48

Auto sarima on sparkling:-

For auto sarima model we take p=3,d=1,q=3

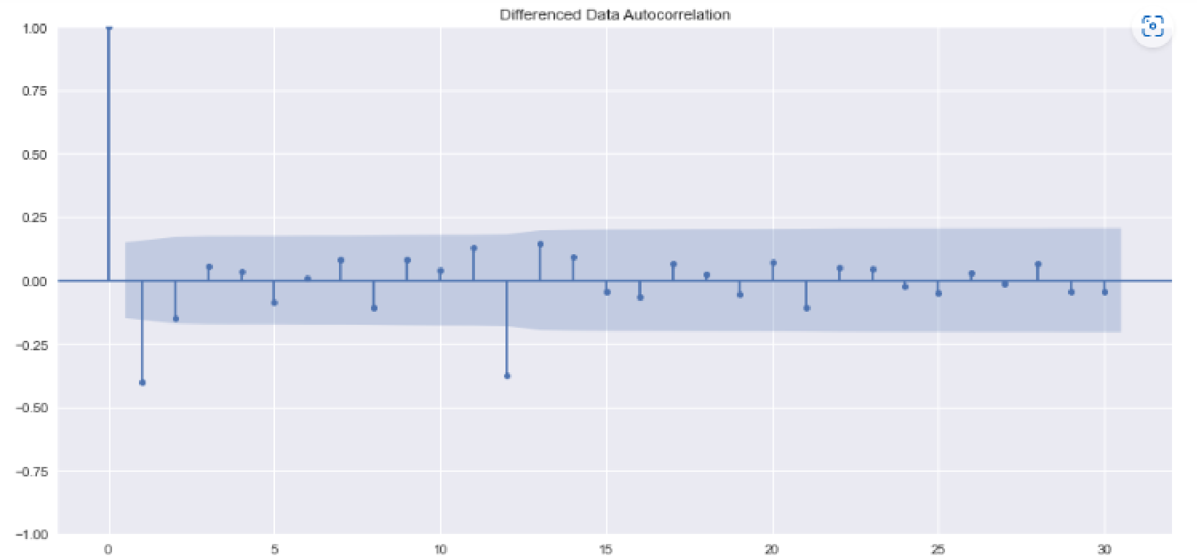


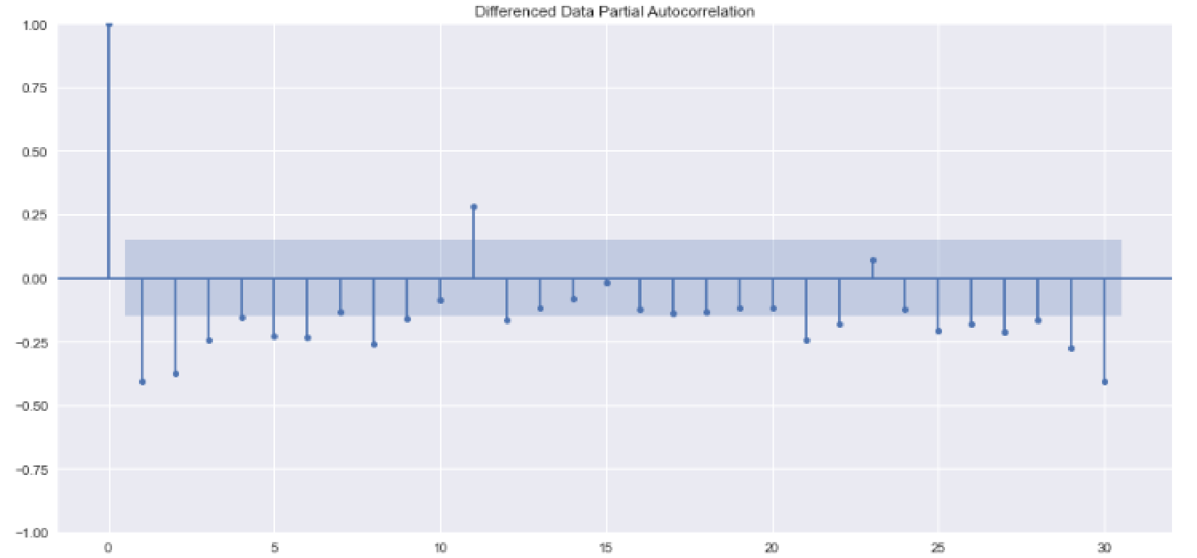
Rmse value :-



```
For SARIMA forecast on the Sparkling Testing Data:  RMSE is 331.586 and MAPE is 10.33
```

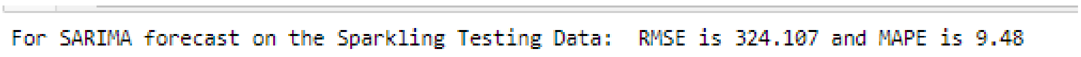
Manual spark arima on sparkling:-





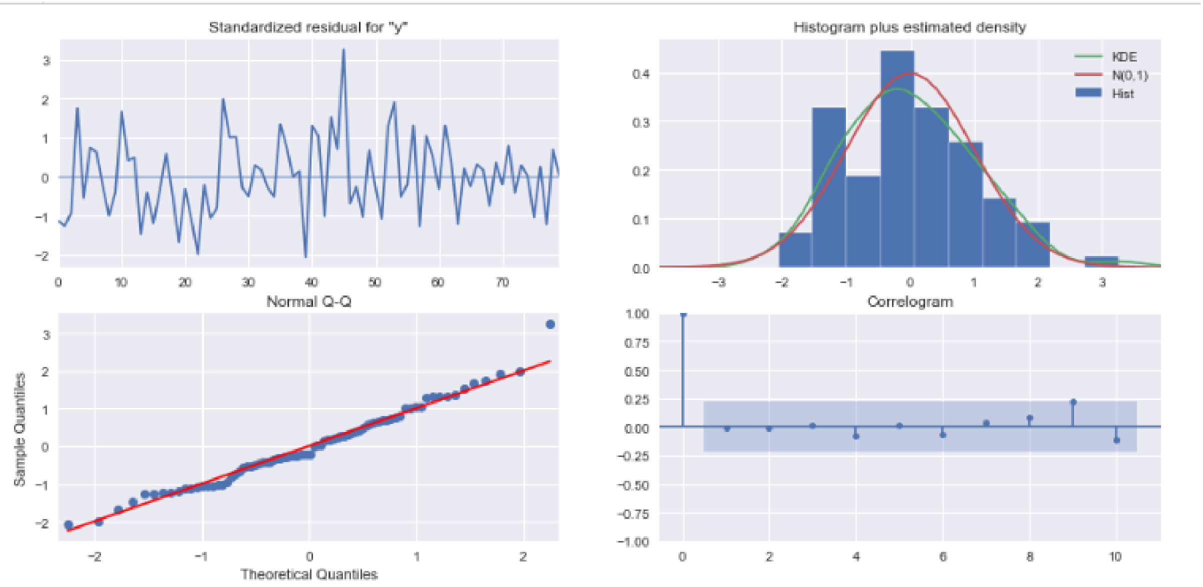
From the above plot we take p=3,d=1,q=1

Test rmse:-



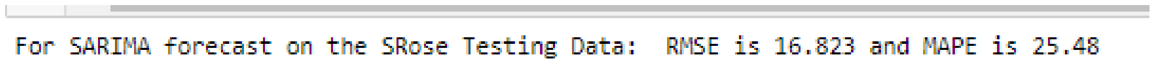
```
For SARIMA forecast on the Sparkling Testing Data:  RMSE is 324.107 and MAPE is 9.48
```

Sarima model on rose:-



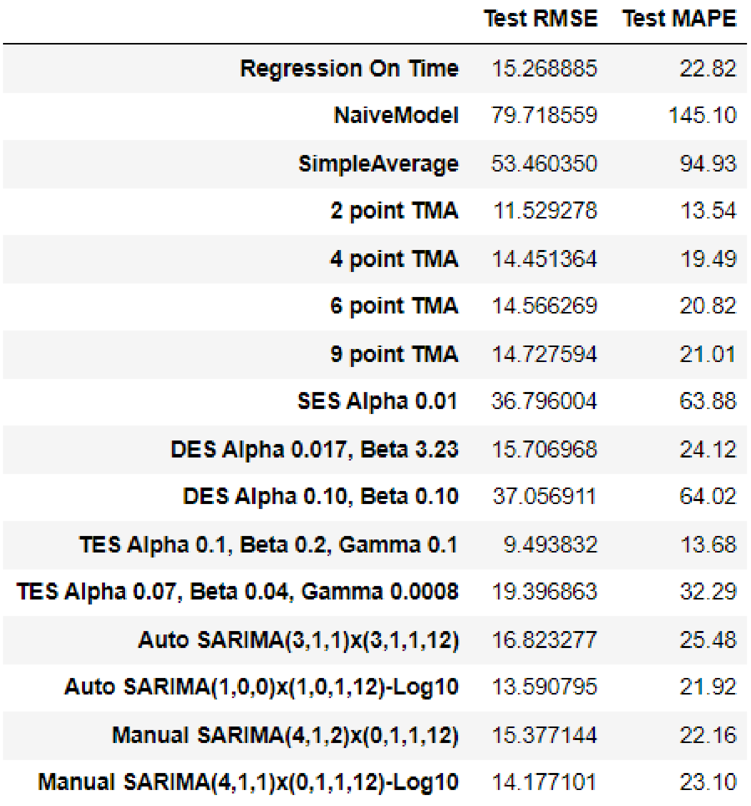
For the sarima model on rose we take p=3,d=1,q=1

Test rmse:-



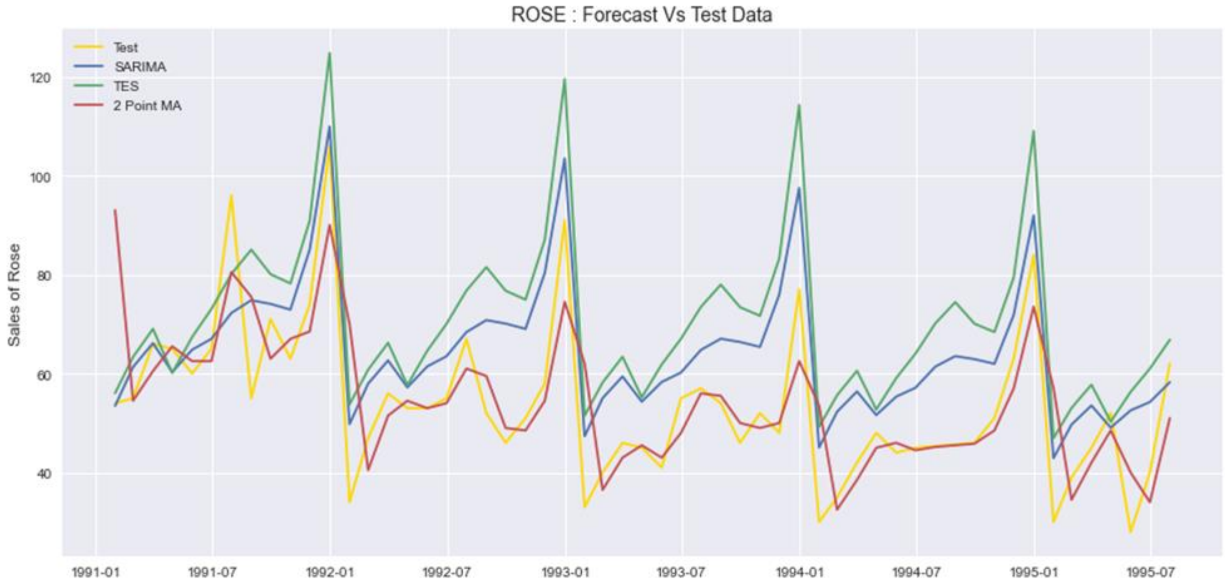
```
For SARIMA forecast on the SRose Testing Data:  RMSE is 16.823 and MAPE is 25.48
```

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



	Test RMSE	Test MAPE
Regression On Time	15.268885	22.82
NaiveModel	79.718559	145.10
SimpleAverage	53.460350	94.93
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
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.017, Beta 3.23	15.706968	24.12
DES Alpha 0.10, Beta 0.10	37.056911	64.02
TES Alpha 0.1, Beta 0.2, Gamma 0.1	9.493832	13.68
TES Alpha 0.07, Beta 0.04, Gamma 0.0008	19.396863	32.29
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823277	25.48
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.590795	21.92
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.177101	23.10

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.

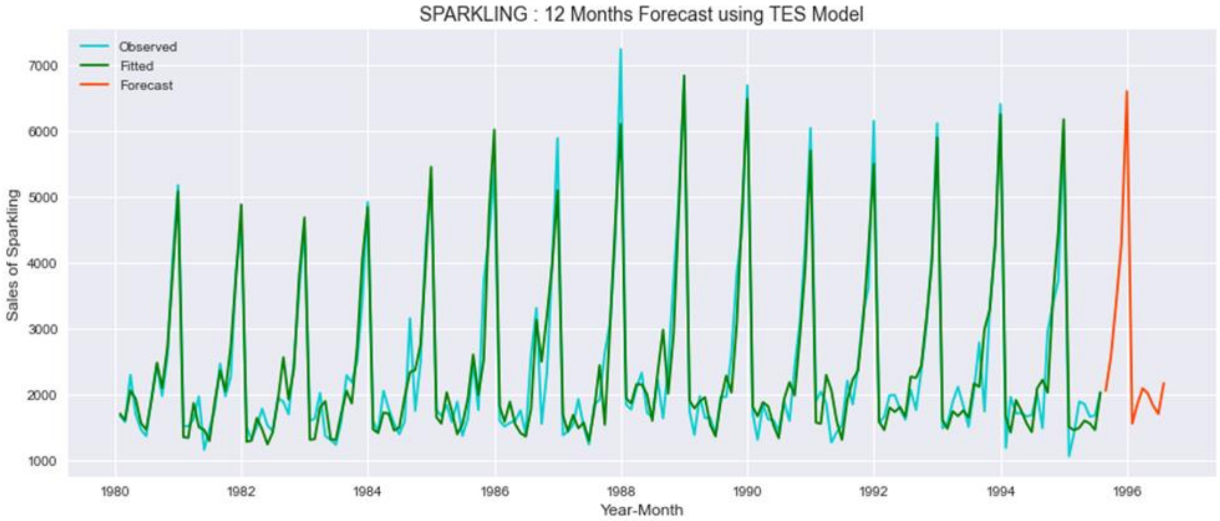


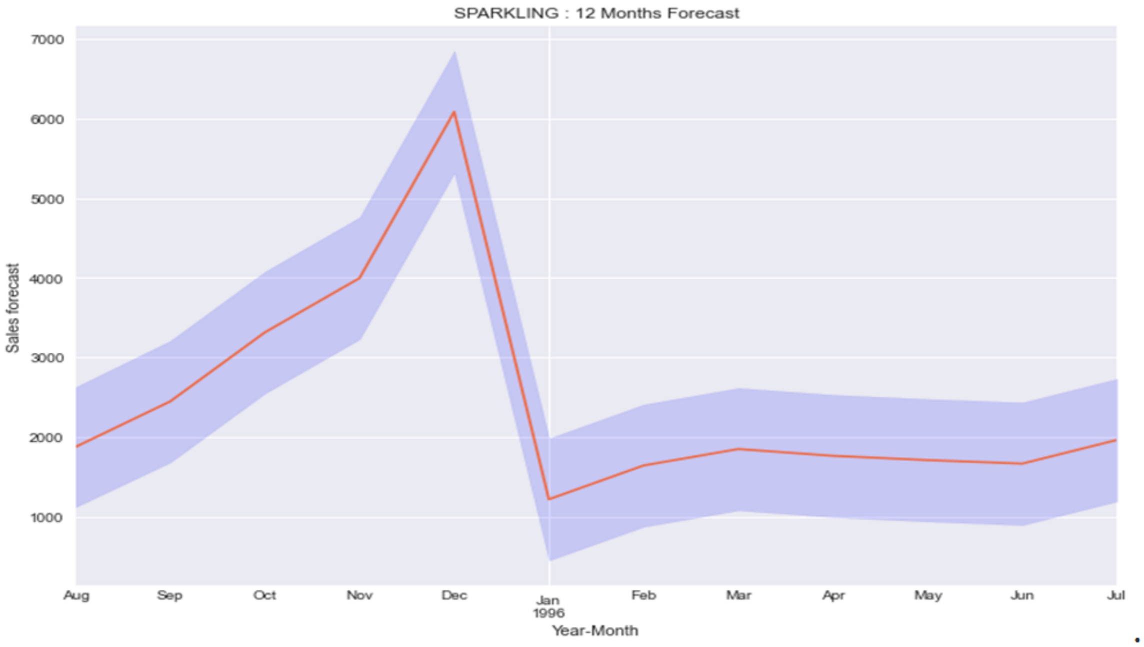
- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data.

- 2 point trailing moving average is found to be having the best fitment against the test data, through with lag of 2 and falling short at times.

- Both SARIMA and Triple Exponential Smoothing are found a bit higher than actuals at any given point in time

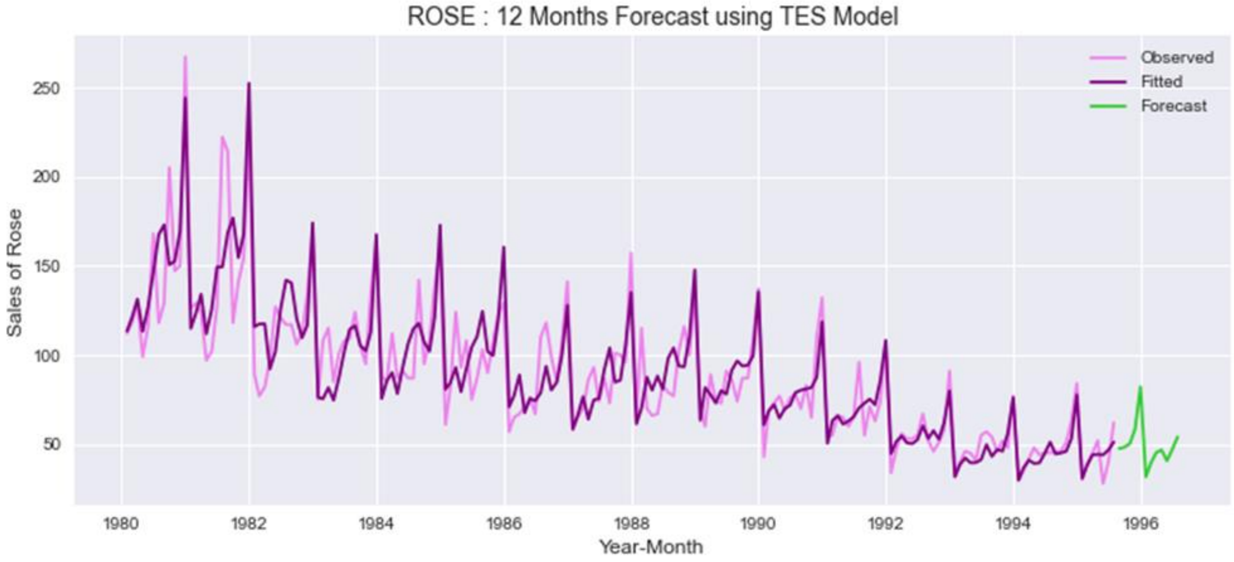
Line Plot for 12 month forecast on Sparkling dataset

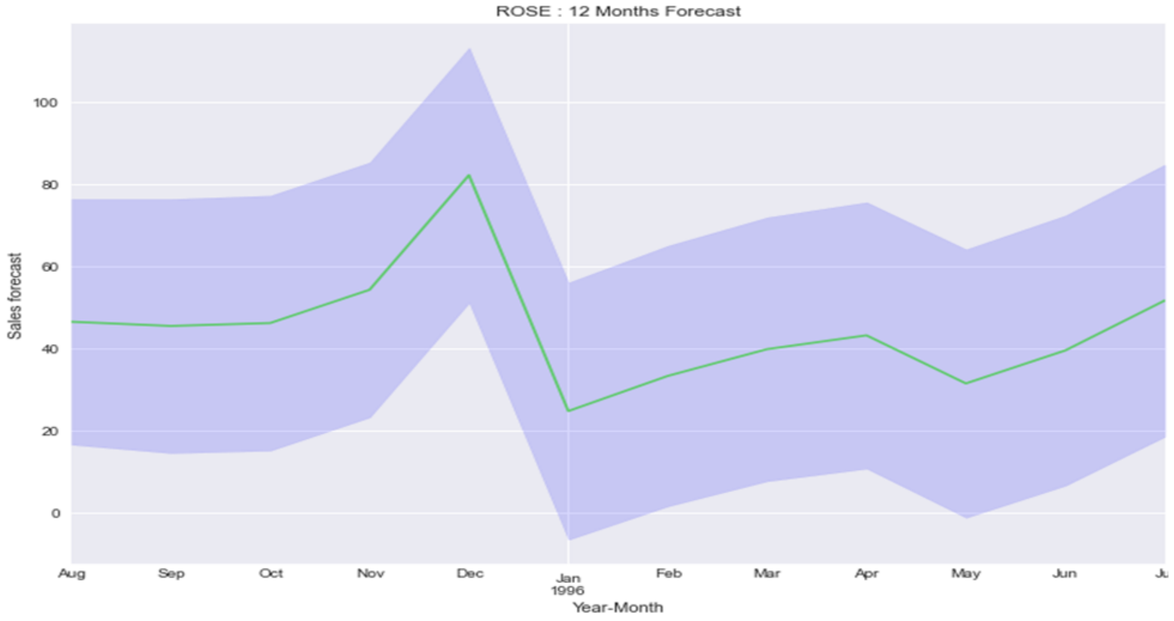




The seasonal sale in December 1995 will hit a maximum of 6084 units before it drops to the lowest sale in January 1996 at 1215 units.

Line Plot for 12 month forecast on Rose dataset





The seasonal sale in December 1995 will hit a maximum of 82 units before it drops to the lowest sale in January 1996 at 24 units.

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

ROSE wine sales

- Rose wine shows a clear trend of decline sales since 1980
- This shows a decline in popularity of variant of wine
- Also, there is a clear spike in sales in oct to dec
  This is due to holiday season in this period
  Highest sales in the December we can see
- There is also a crash in sales in sales in the first quate in the year from Jan
- Sales slowly pick up from Jan

Suggestions: -

- Holiday season is around the corner and forecast shows increasing sales and sharp peak in dec hence company should stock up
- Company can rebrand its rose variant along with a new wine master
- Company should take advantage of the oncoming spike from aug-oct by introduction of offers and ad campaigns
- If there is no significant upward trend in sales by this dec then company has 2 options discontinuing this variant and come up with new ventures

Sparkling-wine sales

- **Triple exponential smoothing performs the best on sparkling dataset considering the least RMSE with each tuning of parameter**
- **Even for sparkling holiday season is around the dec hence company should stock up**
- **Sparkling wine has a great holiday sale so this shows popularity**
- **So, no need to introduce any offers ads are suggested in these season**
- **Year on year sales do not show any significant increase or decrease**
- **Though holiday spikes are extreme but general year on year sales need to be investigated more. Early period from Jan should be used to do this deep dive**