Time Series Forecasting

Haresh P Tayade

PGP-DSBA

Online Sept'2021

Date-17/04/2022

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.

Introduction

The intent for this project is to perform Forecasting analysis on the Sparkling dataset. I will try to analyse this dataset by using Linear Regression, Naïve Model, Simple and Moving Average models, Simple, Double and Triple Exponential Smoothing. These datasets contains 187 entries each, and I will try to build the most optimum model(s) the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Questions

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
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.
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.
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.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
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.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Description

The above problem is a case of time series forecasting where in we need to predict the future sales using historical sales data for sparkling and Rose data.

Sparkling

1.Read the data as an appropriate time series data and plot the data.

Head of dataset

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

Tail of dataset

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

Description of dataset

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

Null values

```
Sparkling    0
dtype: int64
```

Shape and Info of dataset

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

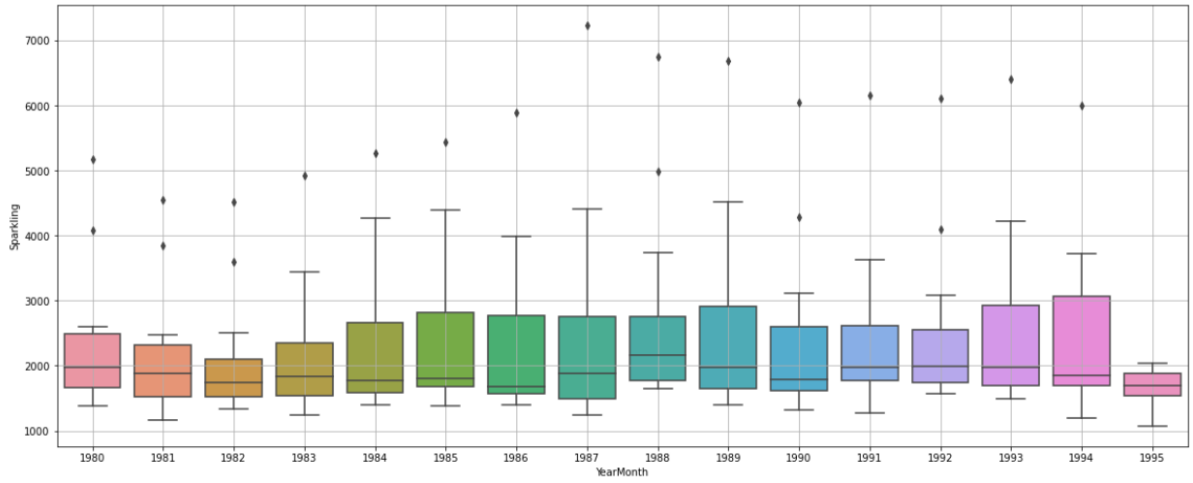
Observation

1. We can see that there are in the dataset there are 187 rows and 1 column.
2. As check found that there are no null values in the Sparkling dataset.
3. From info we can see that the values are of (int) type.
4. From description we find the mean, median, max, 25%,50%,75% values.

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

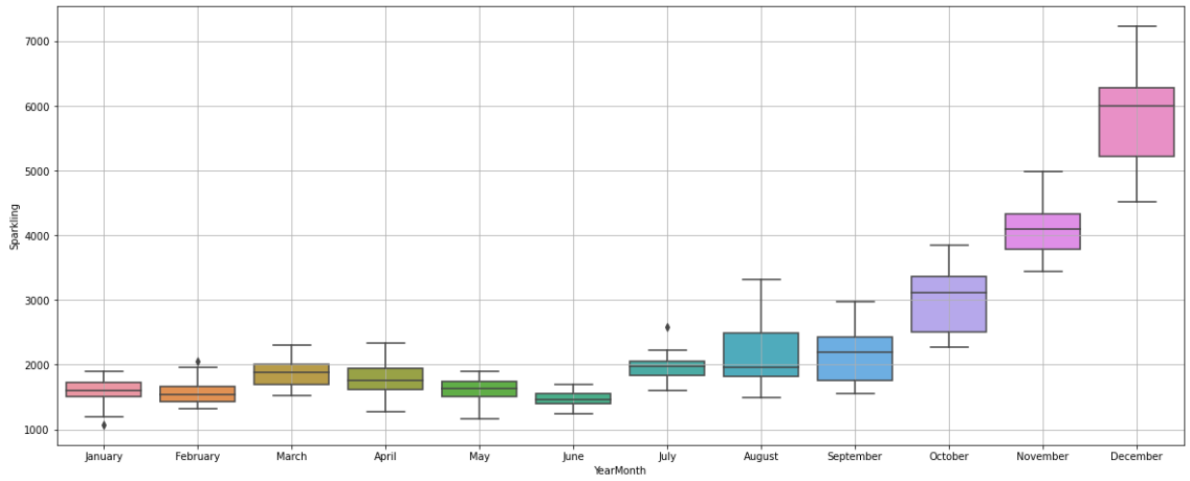
Let us check the spread of sales across different years.

Yearly plot



The yearly boxplot also shows that the sparkling wine sales have increase with the passing year.

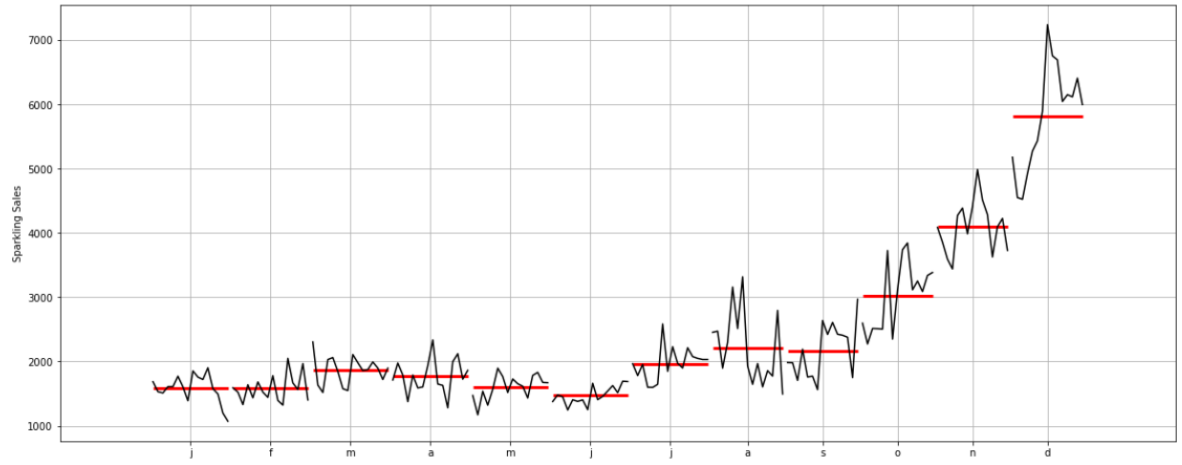
Monthly plot



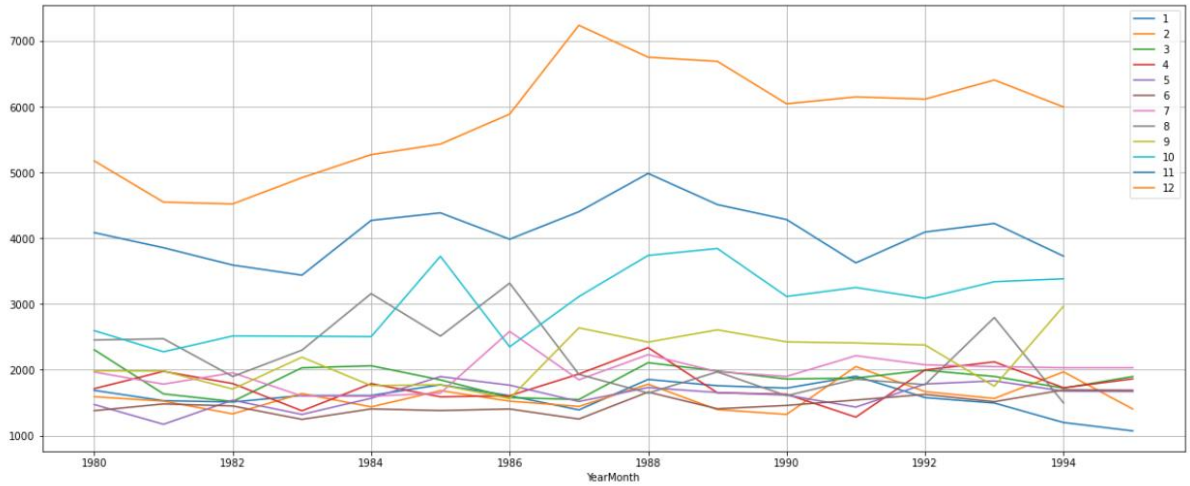
This is a clear distinct of sparkling wine sales within different months spread across various years.

The highest such number are being recorded towards the end of the year with huge spike in the month of December

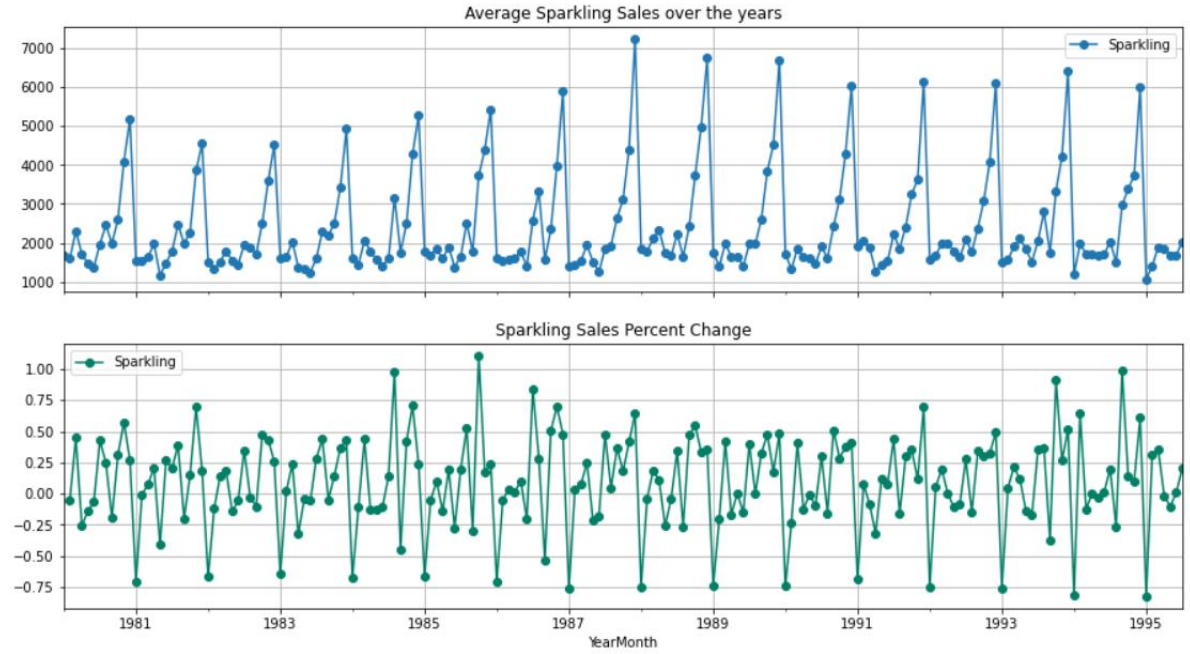
Time series month plot



Monthly sales across years

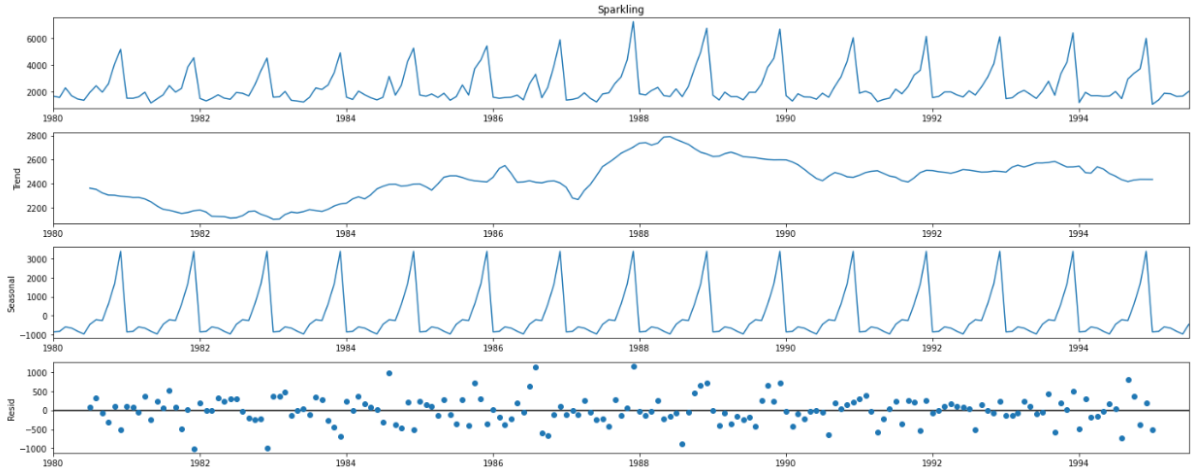


Let us check the average Rose per month and the month on month percentage change of Rose.



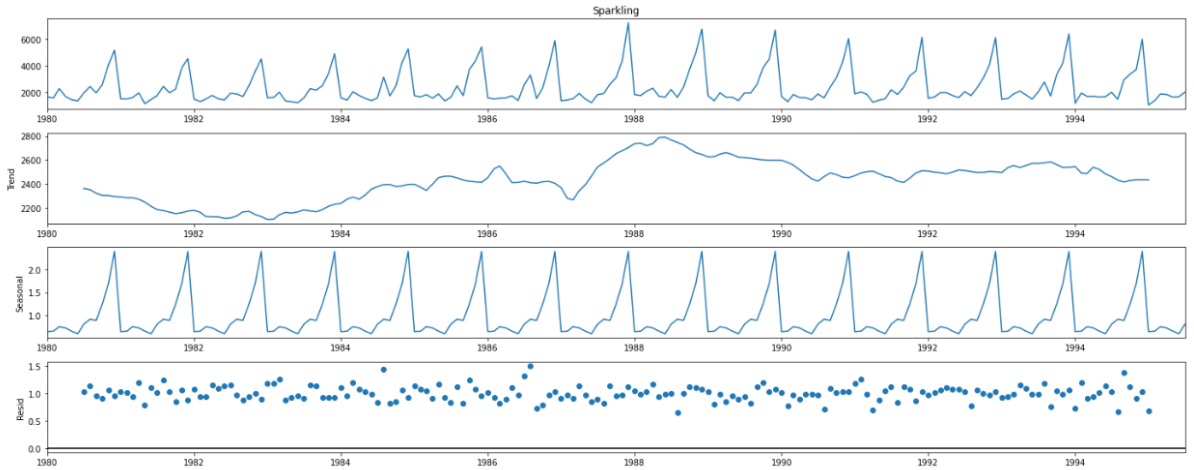
The above two graphs tells us the Average 'Sparkling' and the Percentage change of 'Sparkling' with respect to the time.

Seasonal Decomposition



Taken the above data and make then seasonal decomposition with residual ,trend ,seasonal with ADDITIVE model.

Seasonal Decomposition



Taken the above data and make then seasonal decomposition with residual ,trend ,seasonal with MULTIPLICATIVE model.

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

Training data

First few rows of Sparkling_training Data

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

Last few rows of Sparkling_training Data

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

Testing Data

First few rows of Sparkling_test Data

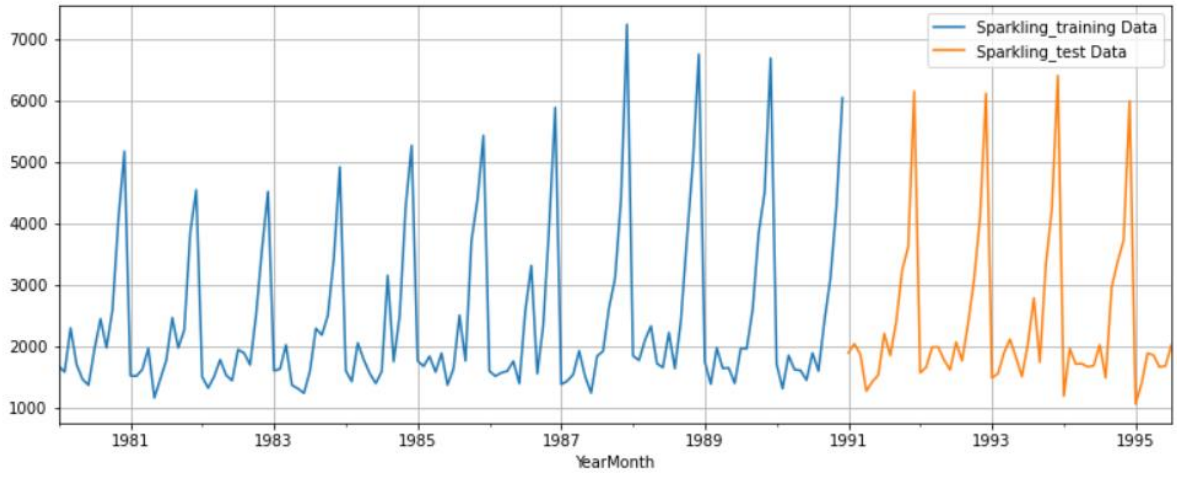
YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Last few rows of Sparkling_test Data

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

From this data till 1990 was choosen as training data and from 1991 to 1995 was choose as testing data.

Now we will see the yearly sales in training and testing



Sales plot across various years for training and testing data.

```
(132, 1)
(55, 1)
```

From above we can see that from whole data set of 187 rows and and 1 column. Almost 132 rows are in training dataset and 55 rows are in testing dataset.

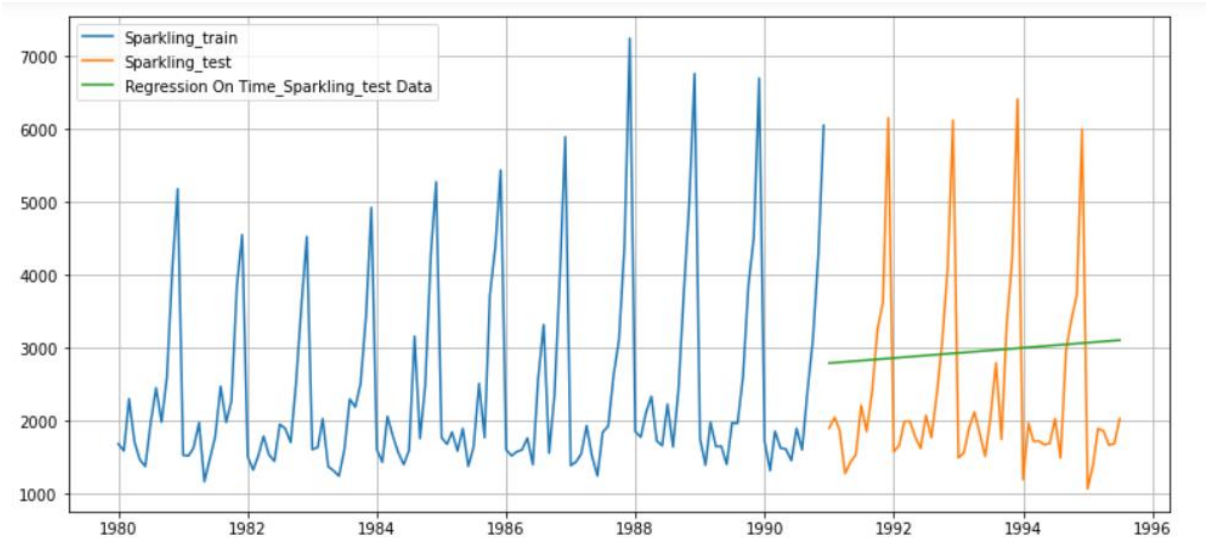
As we can see in above graph that the blue graph is of training data and orange is testing data.

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

1.Linear Regression

```
Sparkling_training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Sparkling_test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

Fitting the model on train data and test data

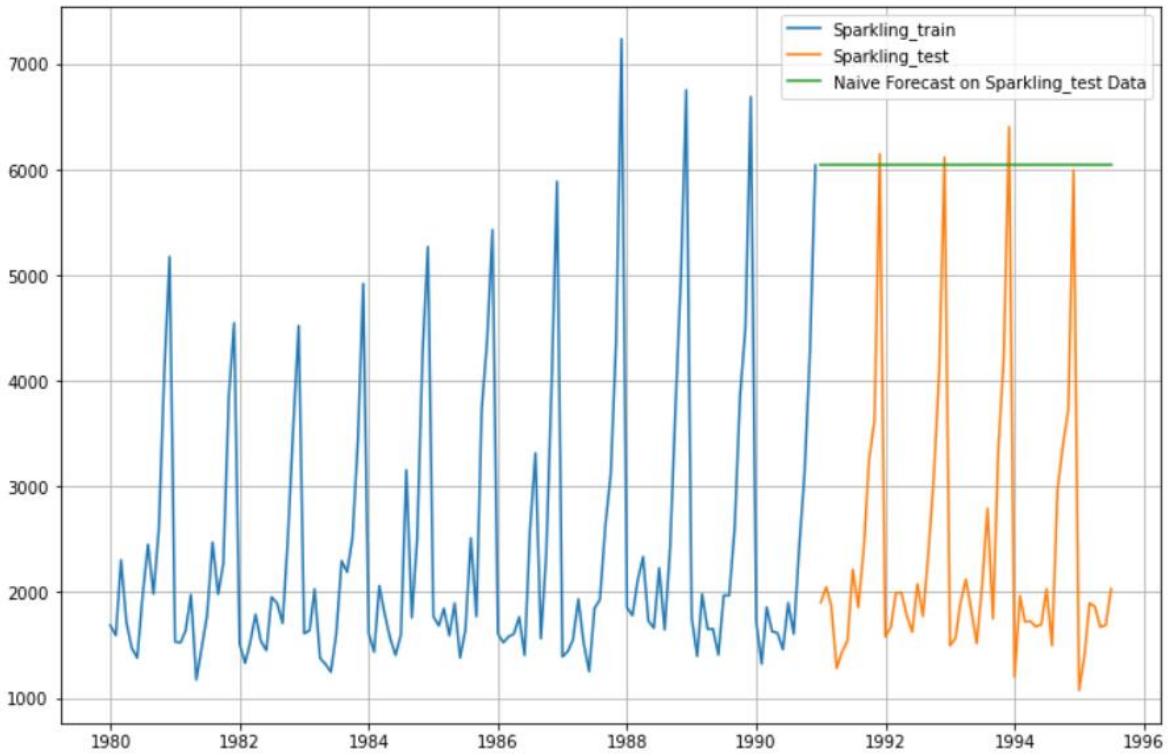


Linear Regression on Time Sparkling data

RMSE value

	Sparkling_test RMSE
RegressionOnTime	1389.135175

2.Naïve Bayes



Hence on plot it shows a straight line hence remaining all the values are same.

The RMSE value for naïve bayes is **3864.279**

But if we need to see both the result of sparkling RMSE test

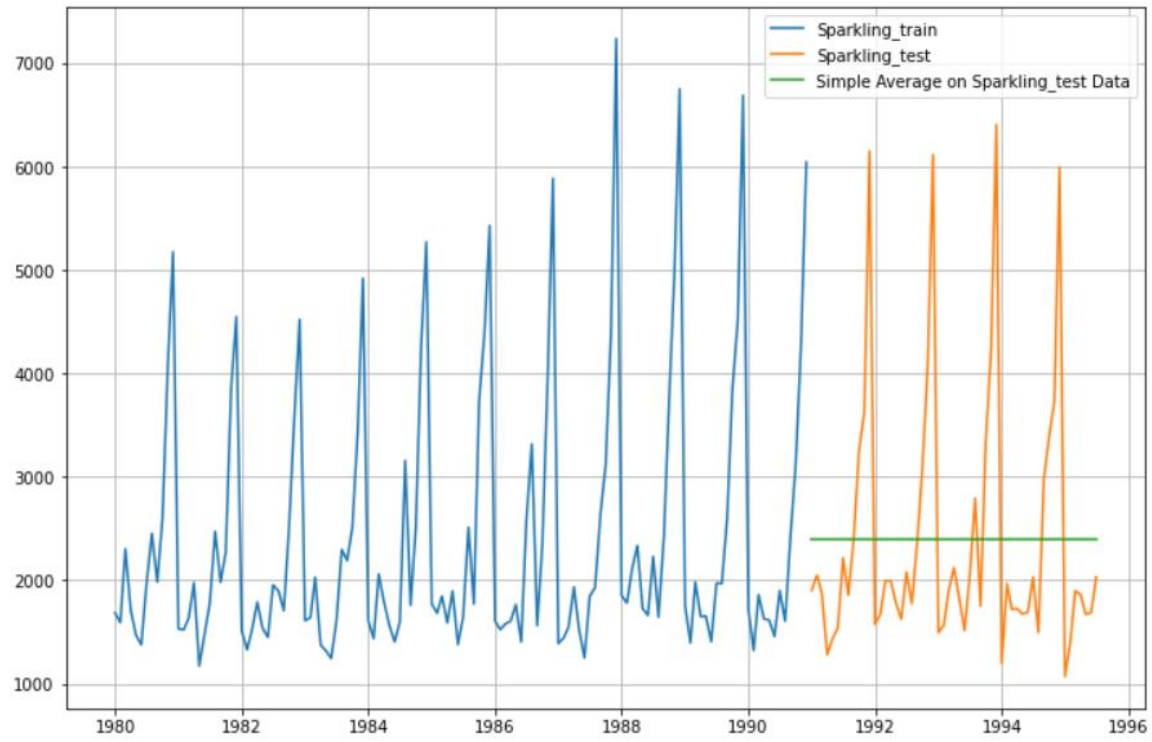
The result is

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352

As can be seen from the Naïve model performance for Sparkling wine datasets above, the Naïve model is not suitable for any of the wine datasets since the forecasts depends on the previous last observation.

3.Simple Average

In this method the predicted values are calculated basis mean value of the entire data just like naïve approach it will also show flat line.



Below is the snippet of RMSE on test data

```
For Simple Average forecast on the Sparkling_test Data,  RMSE is 1275.082
```

But if we need to see all the 3 RMSE value on test data

The result is

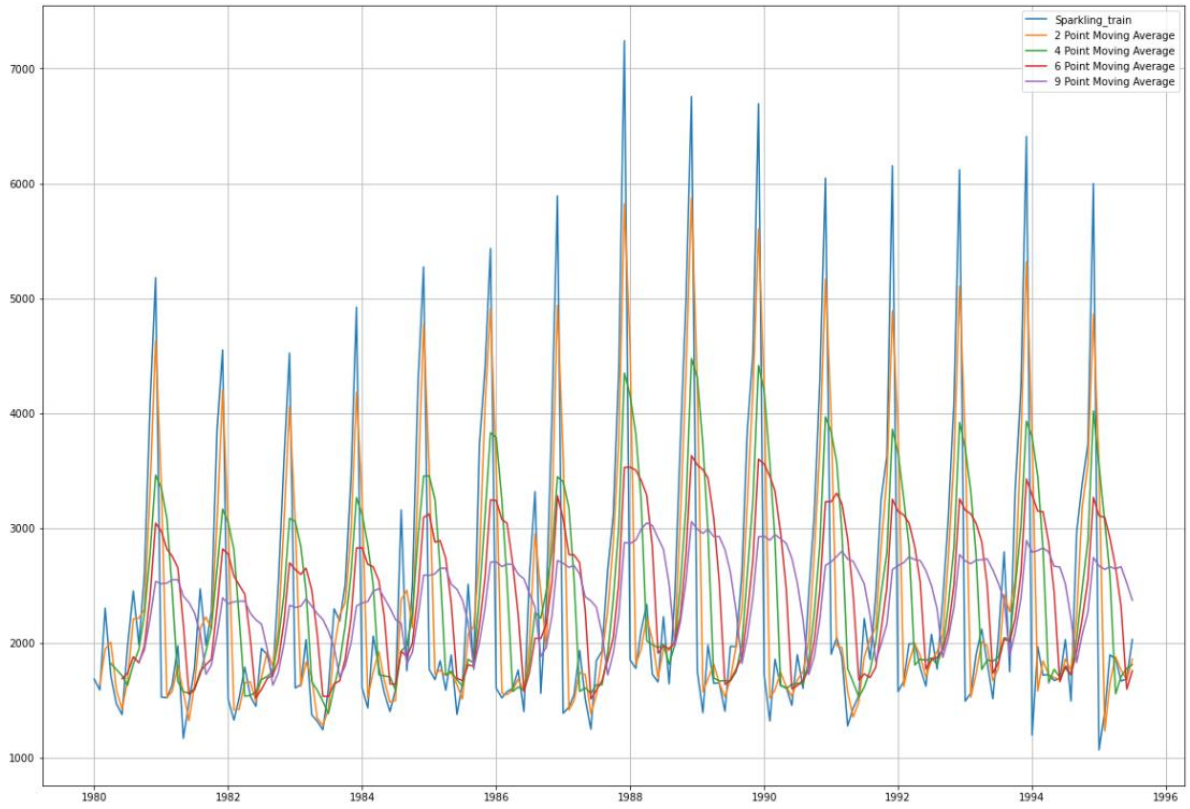
	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

As can be seen from the Simple Average model performance for Sparkling wine datasets above, the Linear Regression model has the best performance among all the three models run till now the Simple Average model shows the best performance among all the three models run till now for the Sparkling wine dataset.

4- Moving Average Model

The moving average data for sparkling wine dataset can be see below.

In this method, just like Simple average predicted values are calculated basis mean value of the previous data but on rolling basis, the number of old data to be used to predict next values can be anything, here we chose 2, 4, 6, 9 as rolling parameters. As seen from Below Plot, Moving average values of all the given parameters closely follow original data in training set, however for test set, since previous values are supposed to be not known, hence the last predicted values of train data becomes the value of entire test data or future data, hence it shows a flat line on test data.



For 2 point moving average model ,RMSE is 3046.976

For 4 point moving average model ,RMSE is 2021.856

For 6 point moving average model ,RMSE is 1521.611

For 9 point moving average model ,RMSE is 1304.619

I have applied 2, 4, 6- and 9-point trailing averages on Sparkling wine data sets.

As we can observe from the above plots, all of the trailing average plots show prediction values below the actual train and test data sets, and the 9 point trailing average plot shows the lowest prediction of all the plots.

datashown by the 2 point trailing moving average model. This observation is co rroborated by theRMSE scores for each of these moving average models.

As can be seen from the summarized performance of all the models, the 2 point moving average has shown the best performance of all the models run on the Sparkling wine datasets.

5. Simple Exponential Smoothing

Simple Exponential Smoothing is used for time series prediction when the data particularly does not follow any:

Trend: An upward or downward slope

Seasonality: Shows a particular pattern due to seasonal factors like Hours, days, Year, etc.

SES works on weighted averages i.e. the average of the previous level and current observation. Largest weights are associated with the recent observations and the smallest weights are associated with the oldest observations.

The decrease in weight is controlled by the smoothing parameter which is known as α(alpha) here.

α(alpha) value can be between 0 to 1:

α(alpha)=0: Means that forecast for future value is the average of historical data.

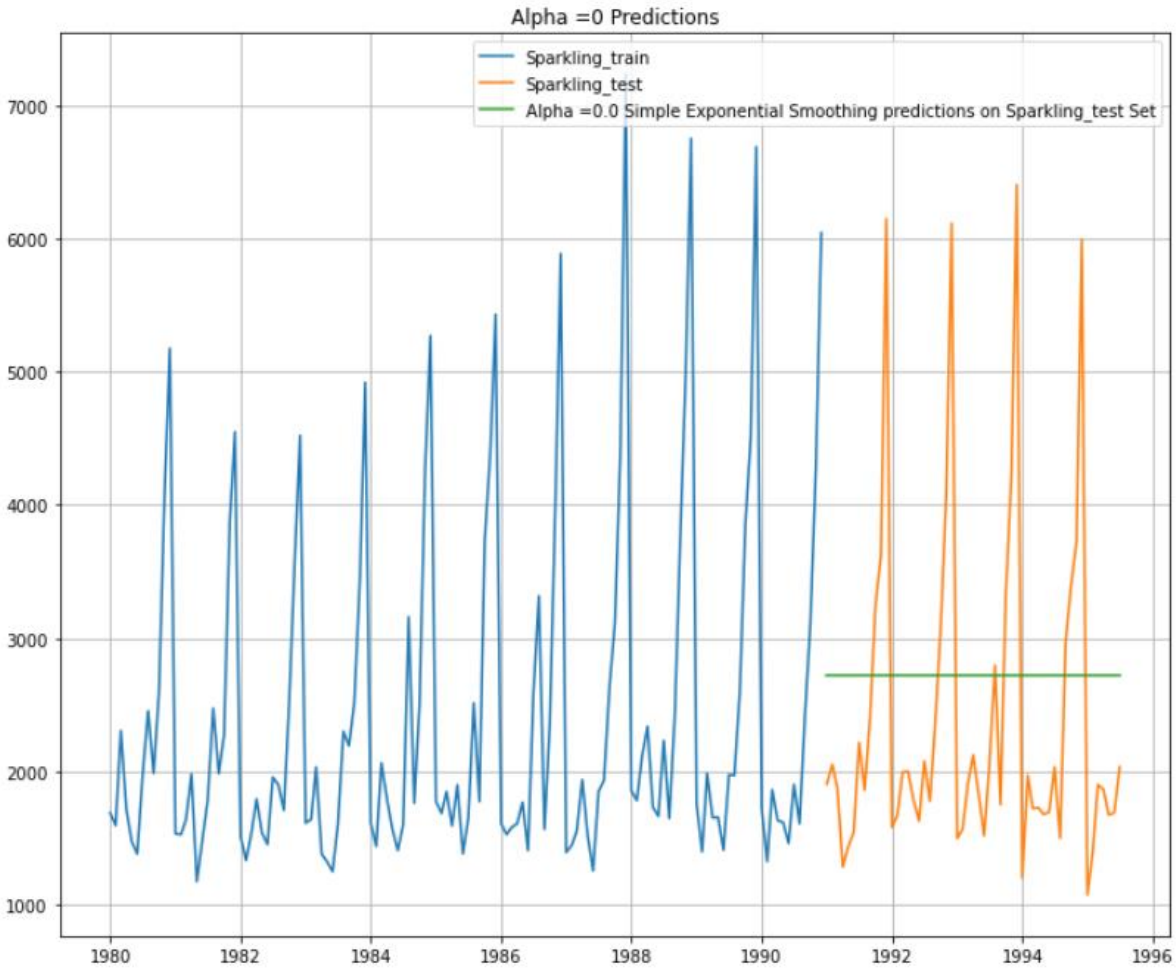
α(alpha)=1: Means that forecast for all future value is the value of the last observation.

Below is the formula for Simple Exponential Smoothing:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

After auto-fitting the model, following is the parameter obtained:

```
{'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}
```



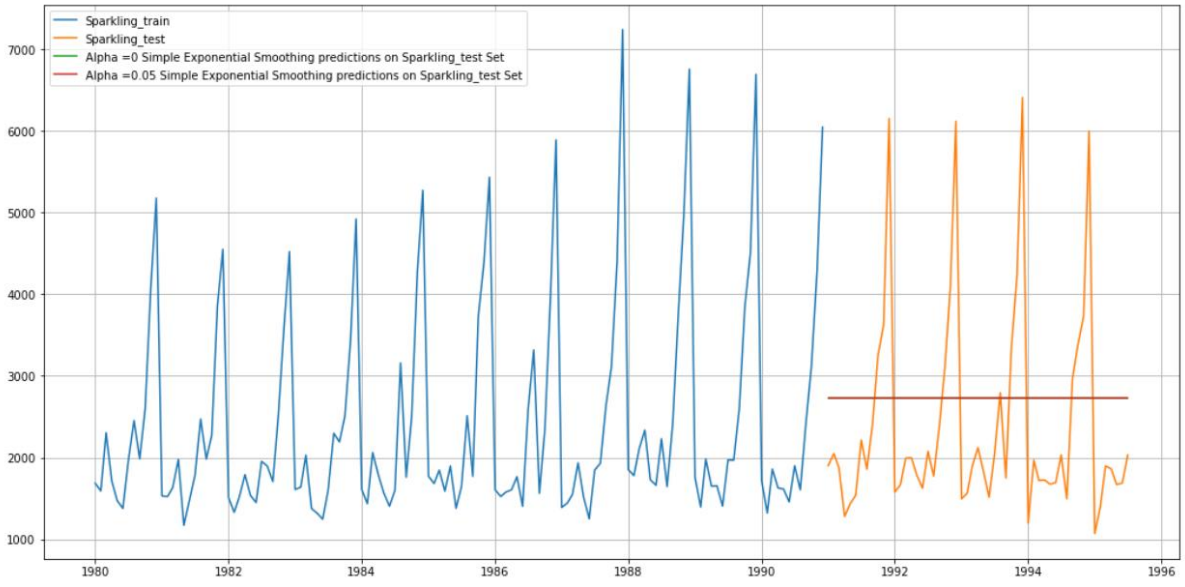
model evaluation for alpha=0.05 simple exponential smoothing

```
For Alpha =0.0005 Simple Exponential Smoothing Model forecast on the Sparkling_test Data, RMSE is 1316.035
```

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487

After trying out manually various alpha values the model, Alpha = 0.05 is the best parameter obtained with lowest TEST RMSE, below is the predicted values plotted on Time scale. As observed manual and automatic alpha values were very close, hence obtained similar forecast for both the values.

	Alpha Values	Sparkling_train RMSE	Sparkling_test RMSE
1	0.05	1318.429335	1316.411742
2	0.10	1333.873836	1375.393398
0	0.00	1483.667178	1460.954675
3	0.15	1347.521016	1466.203651
4	0.20	1356.042987	1595.206839
5	0.25	1359.701408	1755.488175
6	0.30	1359.511747	1935.507132
7	0.35	1356.733677	2123.914871
8	0.40	1352.588879	2311.919615
9	0.45	1348.095362	2493.786514
10	0.50	1344.004369	2666.351413
11	0.55	1340.811249	2828.246418
12	0.60	1338.805381	2979.204388
13	0.65	1338.131249	3119.560885
14	0.70	1338.844308	3249.944092
15	0.75	1340.955212	3371.100106
16	0.80	1344.462091	3483.801006
17	0.85	1349.373283	3588.797654
18	0.90	1355.723518	3686.794285
19	0.95	1363.586057	3778.432623



	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742

As we all know that SES model should be used on data which has no element of trend or seasonality, I still applied it on Sparkling wine data sets so as to see what's the performance of the model in this case. I used Alpha = 1 for the SES model and as expected, it did not perform well as compared to previously run models.

6. Double Exponential Smoothing (Holt's method)

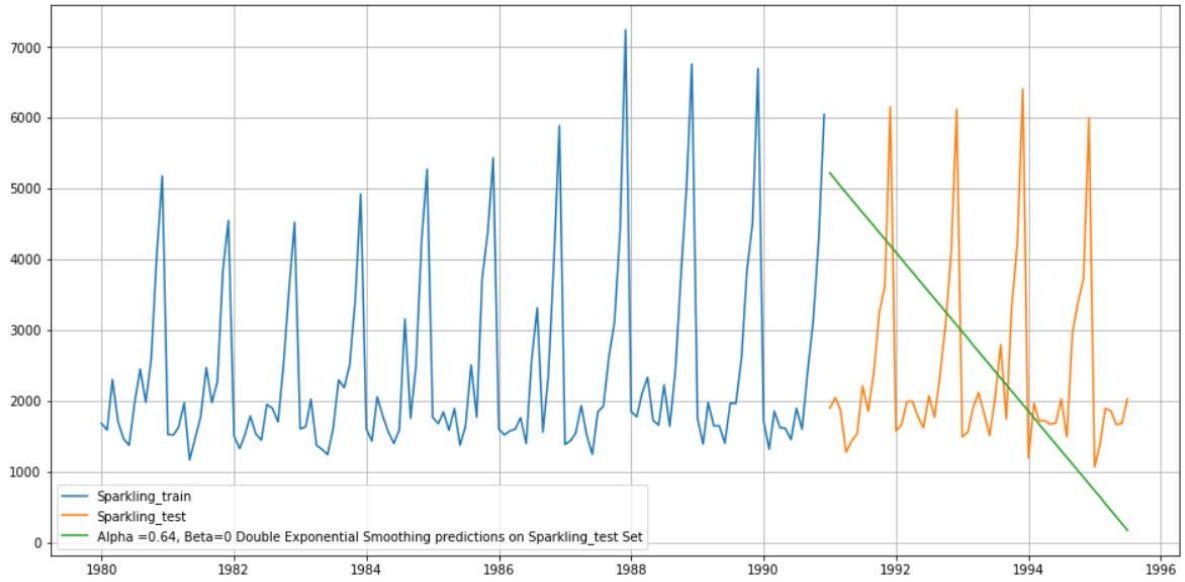
Double exponential smoothing (DWA) is also called holts exponential smoothing

Double exponential smoothing is extended from simple exponential smoothing

DWA technique is used for forecasting with trending data which has level and trend but it does not have seasonality

Level and trend are accounted for this model.

```
{'smoothing_level': 0.6885714285714285,
 'smoothing_trend': 9.999999999999999e-05,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': -95.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



model evaluation for alpha=0.64 and beta=0.005 Double exponential smoothing

```
For Alpha =0.64, Beta=0 Double Exponential Smoothing Model forecast on the Sparkling_test Data, RMSE is 2007.239
```

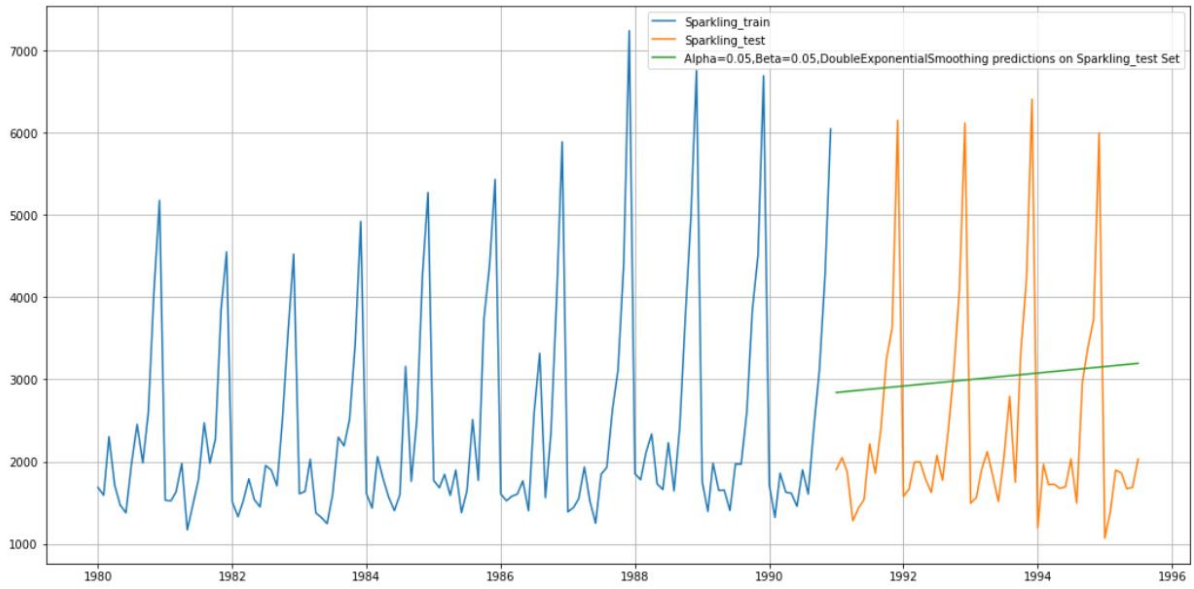
	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526

Results for all the models of sparkling test RMSE.

After manually fitting various model, below are the best parameter with lowest RMSE score on test data

	Alpha Values	Beta Values	Sparkling_train RMSE	Sparkling_test RMSE
0	0.05	0.05	1430.025526	1418.407668
3	0.05	0.20	1382.766405	1443.099273
2	0.05	0.15	1379.162520	1457.041594
1	0.05	0.10	1385.420826	1466.899629
6	0.05	0.35	1414.226231	1547.022626

Alpha=0.05 ,beta=0.05 have lowest AIC



	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668

As we all know that DES model should be used on data which has no seasonality but has levels and trends, I used the grid search to begin and we reached conclusion that Alpha =0.05 and Beta = 0.05 show the lowest RMSE Sparkling wine data sets. The DES model is the model with the worst performance so far for Sparkling wine datasets.

7.Triple Exponential Smoothing (Holt's winter model)

This is an extension of Holt's method when seasonality is found in the data.

Fore caste equation: $Yt+1=lt+bt+st-m(k+1)$

Level Equation: $lt=\alpha(Yt-st-m) + \alpha(1-\alpha)Yt-1$, $0 < \alpha < 1$

Trend Equation: $bt=\beta(lt-lt-1) + (1-\beta)bt-1$, $0 < \beta < 1$

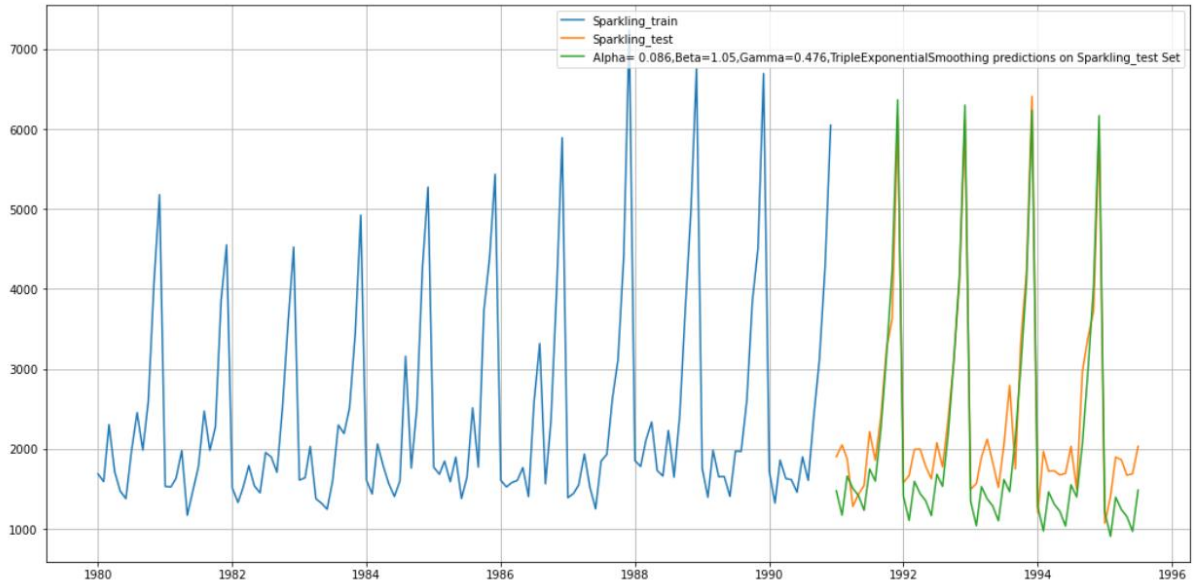
Seasonal Equation: $\gamma(Yt-lt-1-bt-1) + (1-\gamma)st-m$, $0 < \gamma < 1$

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, $\beta$and γ. This is a general method and a true multi-step ahead forecast.

The parameter obtained and predicted value plot on time series.

```
{'smoothing_level': 0.11235974440805609,
 'smoothing_trend': 0.03742154913668688,
 'smoothing_seasonal': 0.4932616459048464,
 'damping_trend': nan,
 'initial_level': 1640.2806120050896,
 'initial_trend': -3.261533670070838,
 'initial_seasons': array([  45.86595538,  -48.96808341,  662.32406973,   73.10075169,
        -168.81341007, -262.13208801,  326.10174942,  813.36401315,
         344.51476989,  956.12012048, 2446.68553948, 3538.12189099]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

TES parameter for sparkling wine dataset respectively



Triple Exponential Smoothing Outcome on the Sparkling Wine Time Series

For Alpha= 0.086,
Beta=1.05,
Gamma=0.476,
Triple Exponential Smoothing Model forecast on the Sparkling_test Data, RMSE is 473.152

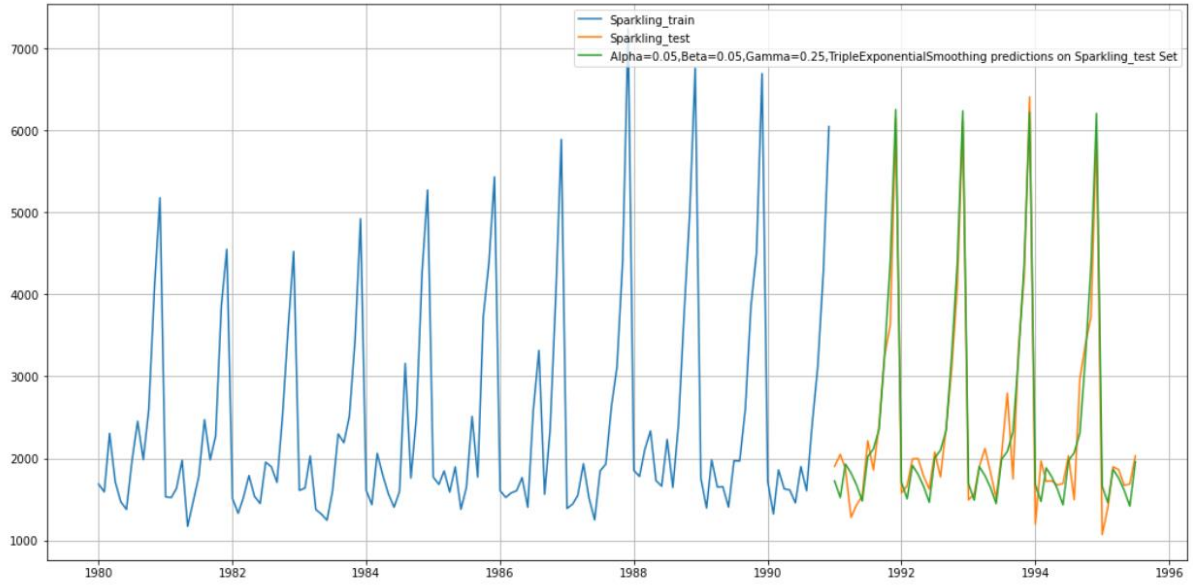
	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668
Alpha= 0.086,Beta=1.05,Gamma=0.476,TripleExponentialSmoothing	473.152417

Summarised performance of all the models

After manually fitting various model below are the best parameter with the lowest RMSE score

	Alpha Values	Beta Values	Gamma Values	Sparkling_train RMSE	Sparkling_test RMSE
0	0.05	0.05	0.05	441.393031	438.375404
1	0.05	0.05	0.10	426.200489	372.134008
2	0.05	0.05	0.15	414.012152	329.797574
3	0.05	0.05	0.20	404.312543	308.849422
4	0.05	0.05	0.25	396.732832	304.645683

Hence can be seen that alpha=0.05 and beta=0.05 and gamma = 0.20 comes out to be better parameter and below is the predicted test value on timeseries.



For the Sparkling wine dataset, the TES model offers the best RMSE and MAPE among all the models.

For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing. However, since this was a model building exercise we had gone on to build different models on the data and have compared these model with the best RMSE value on the Sparkling_test data.

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.*

Stationary process: A process is said to be stationary if its mean and variance are constant over a period of time and, the correlation between the two time periods depends only on the distance or lag between the two periods.

Mathematically, let Yt be a time series with these properties:

Mean: $E(Yt)=\mu$

Variance: $Var(Yt)=E(Yt-\mu)2=\sigma2$

Correlation: $\rho k=E[(Yt-\mu)(Yt+k-\mu)/(\sigma t\sigma t+k)]$

Where $\rho k$is the correlation (or auto-correlation) at lag $k$between the values of $Yt$and $Yt+k$

So, if mean, variance and correlation (or auto-correlation) of time series data is constant (at different lags) no matter at what point of time it is measured; i.e. if they are time invariant, the series is called a stationary time series. A series not possessing these properties is termed as a non-stationary timeseries.

We are going to use Augmented Dickey-Fuller test to test stationary of the data.

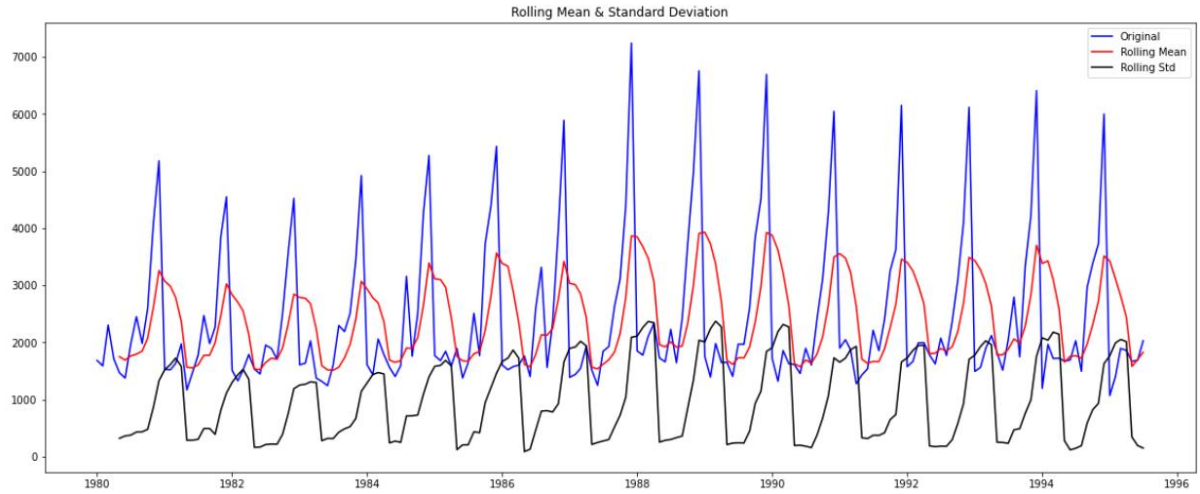
Null hypothesis – Data is non – Stationary

Alternate hypothesis – Data is stationary

Alpha = 0.05

Hence , if the value of P comes out to be less than 0.05, then we rejected null hypothesis.

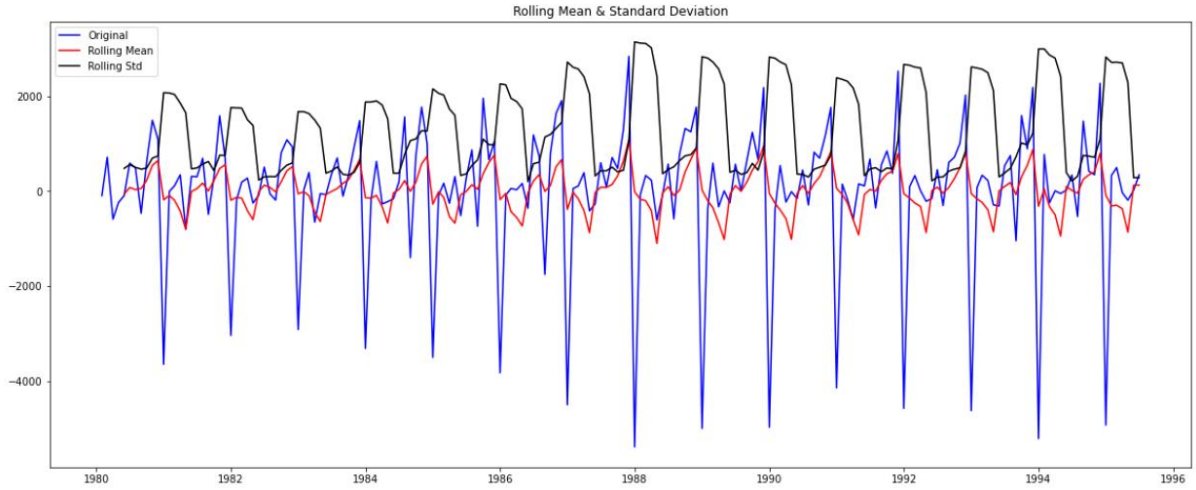
Stationary test result on original time series




```
Results of Dickey-Fuller Sparkling_test:
Sparkling_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
```

Original time series data fails the test hence it can be concluded that the time series is not stationary, hence require differentiating with respect to previous values.

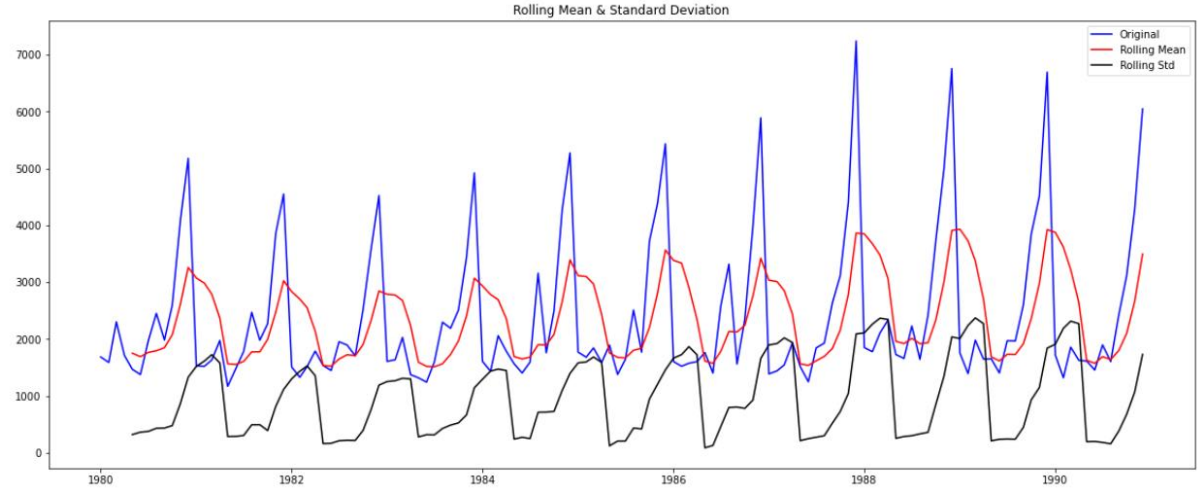
Stationarity test result on original time series after differentiating once




```
Results of Dickey-Fuller Sparkling_test:
Sparkling_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
```

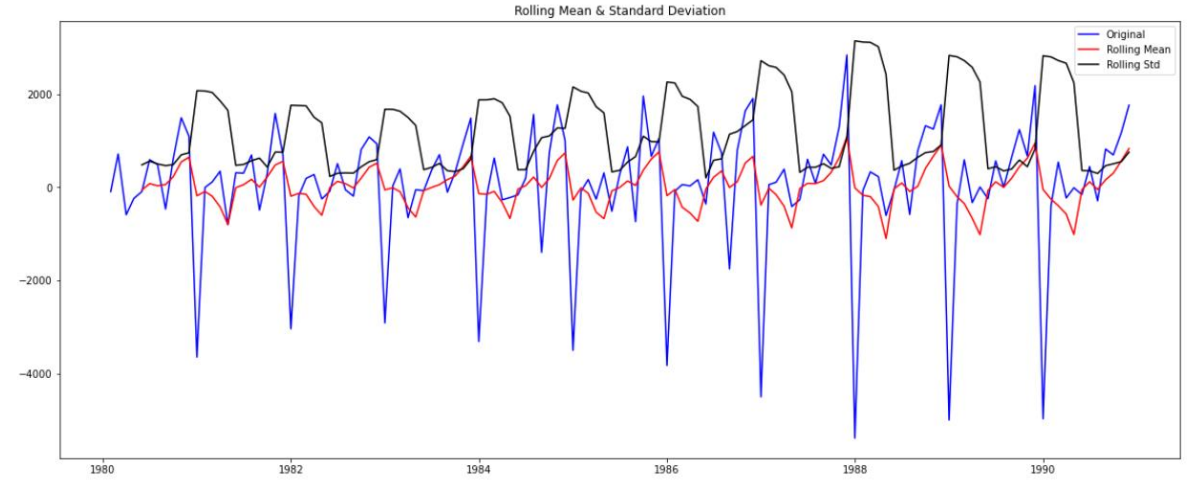
Original time series after differentiating passes the test hence it can be concluded that the time series has now become stationary and this time series can be used in ARIMA /SARIMA model.

Stationary test result on original time series training data.




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

After Differentiating



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

Original time series training data after differentiating passes the test hence it can be concluded that the time series has now become stationary and now this time series can be used in ARIMA/SARIMA model.

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.

ARIMA Model:

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q

where,

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

A pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

After fitting the model in ARIMA

	param	AIC
8	(2, 1, 2)	2210.626049
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906439
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

As we can see 2,1,2 as p, d, q values respectively gives the least RMCE Scores. Lets find the model summary and RMSE scores for test Data.

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.313
Method:	css-mle	S.D. of innovations	1013.755
Date:	Wed, 13 Apr 2022	AIC	2210.626
Time:	19:58:21	BIC	2227.877
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.636

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5845	0.519	10.753	0.000	4.567	6.602
ar.L1.D.Sparkling	1.2698	0.075	17.040	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5601	0.074	-7.617	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9957	0.043	-46.821	0.000	-2.079	-1.912
ma.L2.D.Sparkling	0.9957	0.043	23.291	0.000	0.912	1.079

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7074j	1.3361	-0.0888
AR.2	1.1335	+0.7074j	1.3361	0.0888
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	1.0043	+0.0000j	1.0043	0.0000

SARIMA model

SARIMA is Seasonal ARIMA, or simply put, ARIMA with a seasonal component.

A typical SARIMA model equation looks like the following – SARIMA(p,d,q)x(P,D,Q)lag

The parameters for these types of models are as follows:

p and seasonal P: indicate the number of AR terms (lags of the stationary series)

d and seasonal D: indicate differencing that must be done to stationary series

q and seasonal Q: indicate the number of MA terms (lags of the forecast errors)

lag: indicates the seasonal length in the data

With the AR plot, we chose two seasonality period i.e. 6 and 12 months to test effect of seasonality ad model with best AIC and RMSE will be chosen for final SARIMA model

After fitting the SARIMA Model with seasonality as 6 on the training data, below are the parameters with the lowest AIC score

	param	seasonal	AIC
50	(2, 1, 3)	(2, 1, 3, 6)	1540.893165
74	(3, 1, 3)	(1, 1, 3, 6)	1540.972904
77	(3, 1, 3)	(2, 1, 3, 6)	1544.662973
20	(1, 1, 3)	(1, 1, 3, 6)	1544.891800
47	(2, 1, 3)	(1, 1, 3, 6)	1545.465564

As we can see (3,1,3) (1,1,3,6) as p, d, q and seasonal PDQS values respectively gives the least RMCE Scores. Lets find the model summary and RMSE scores for test Data.

SARIMAX Results

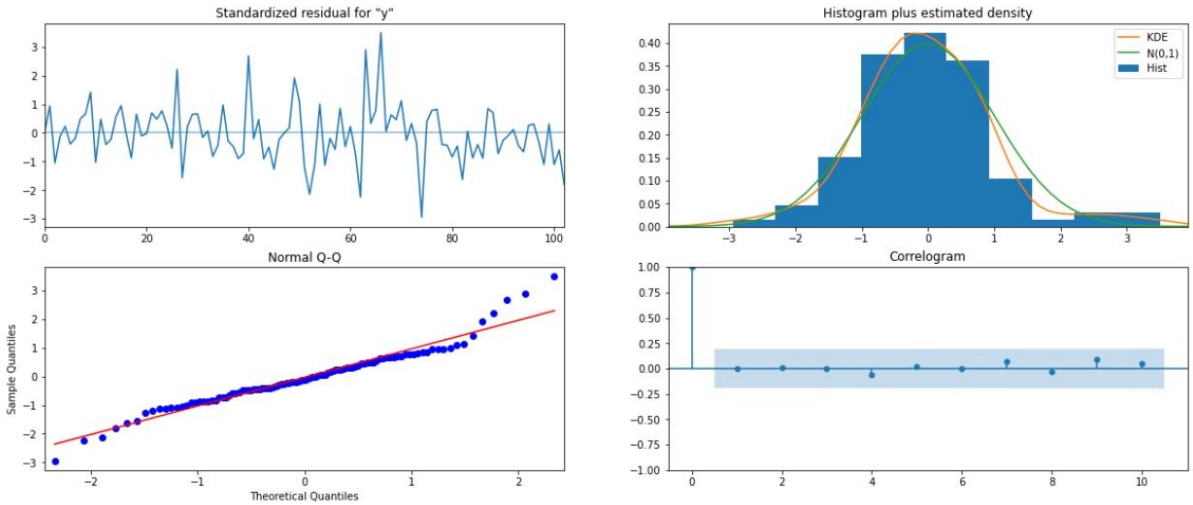
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(1, 1, 3, 6)	Log Likelihood	-762.733
Date:	Wed, 13 Apr 2022	AIC	1545.466
Time:	20:10:21	BIC	1571.813
Sample:	0	HQIC	1556.137
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0906	0.490	0.185	0.853	-0.870	1.052
ar.L2	0.4640	0.531	0.874	0.382	-0.577	1.505
ma.L1	-1.5364	0.815	-1.884	0.060	-3.134	0.062
ma.L2	-0.4980	0.676	-0.737	0.461	-1.822	0.826
ma.L3	1.1352	0.648	1.751	0.080	-0.136	2.406
ar.S.L6	-1.0247	0.008	-121.797	0.000	-1.041	-1.008
ma.S.L6	0.3332	0.290	1.149	0.251	-0.235	0.902
ma.S.L12	-0.7643	0.188	-4.075	0.000	-1.132	-0.397
ma.S.L18	0.1401	0.180	0.778	0.436	-0.213	0.493
sigma2	6.184e+04	7.51e+04	0.823	0.410	-8.54e+04	2.09e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	19.79
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	1.36	Skew:	0.49
Prob(H) (two-sided):	0.38	Kurtosis:	4.91

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

The RMSE value is



Predicting the auto Sparkling SARIMA. Summary.head()

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1338.387445	392.172885	569.742714	2107.032175
1	1055.415641	402.101820	267.310556	1843.520726
2	1637.075858	402.186730	848.804353	2425.347364
3	1515.674341	407.971787	716.064333	2315.284350
4	1118.221576	409.109582	316.381529	1920.061624

<u>The results of all models RMSE test is given below.</u>

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668
Alpha= 0.086,Beta=1.05,Gamma=0.476,TripleExponentialSmoothing	473.152417
Alpha=0.05,Beta=0.05,Gamma=0.25,TripleExponentialSmoothing	304.645683
ARIMA(2,1,2)	1374.037009
SARIMA(2, 1, 3),(1, 1, 3, 6)	949.981844

SARIMA 12:

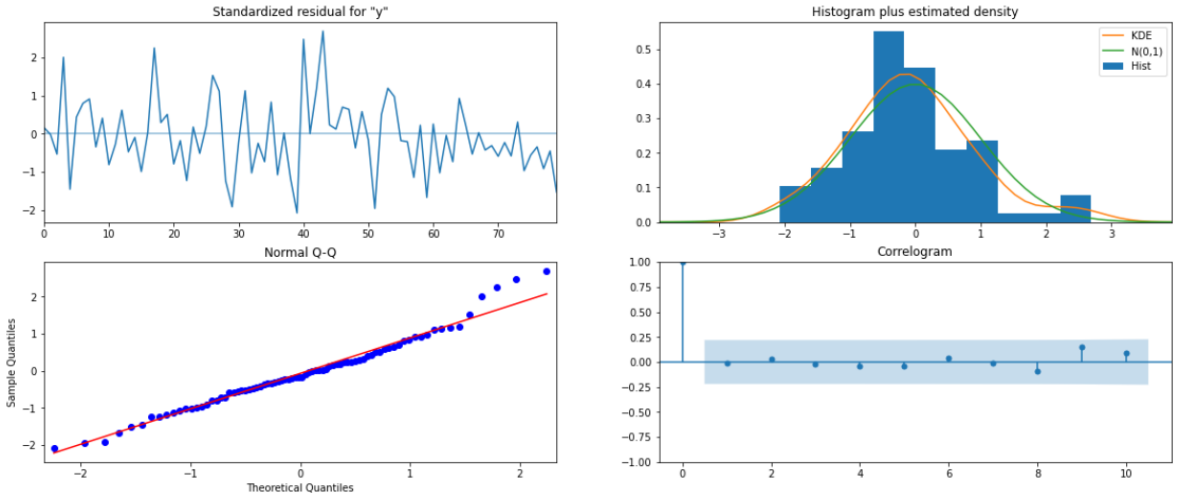
After fitting the SARIMA Model with seasonality as 6 on the training data, below are the parameters with the lowest AIC scores.

	param	seasonal	AIC
115	(1, 1, 3)	(0, 1, 3, 12)	16.000000
179	(2, 1, 3)	(0, 1, 3, 12)	18.000000
187	(2, 1, 3)	(2, 1, 3, 12)	22.000000
119	(1, 1, 3)	(1, 1, 3, 12)	870.855410
183	(2, 1, 3)	(1, 1, 3, 12)	905.599591

We see that the model built with lowest AIC comes out to be with parameters for ARIMA as (1, 1, 3) and seasonal Arima as (3, 1, 0, 12)

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 1, [], 12)   Log Likelihood                -596.641
Date:                             Wed, 13 Apr 2022   AIC                           1213.283
Time:                                     20:23:12   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      -1.199      -0.026
ar.L3          0.0861      0.161      0.536      0.592      -0.229       0.401
ma.L1          0.9855      0.472      2.089      0.037       0.061       1.910
ma.L2         -0.8738      0.166     -5.269      0.000      -1.199      -0.549
ma.L3         -0.9466      0.489     -1.936      0.053      -1.905       0.012
ar.S.L12      -0.4519      0.142     -3.191      0.001      -0.729      -0.174
ar.S.L24      -0.2341      0.144     -1.622      0.105      -0.517       0.049
ar.S.L36      -0.1008      0.122     -0.830      0.407      -0.339       0.137
sigma2      1.839e+05   8.97e+04      2.051      0.040    8136.302    3.6e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 4.06
Prob(Q):                              0.93   Prob(JB):                         0.13
Heteroskedasticity (H):               0.73   Skew:                             0.48
Prob(H) (two-sided):                  0.42   Kurtosis:                         3.54
===================================================================================

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



Predict on the Sparkling test Set using this model and evaluate the model.

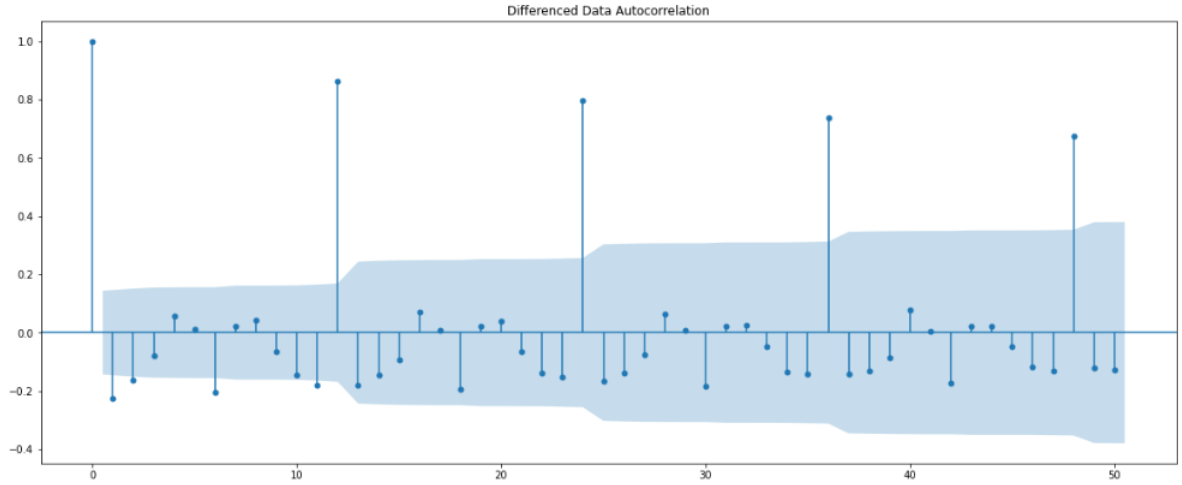
y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1430.749854	431.198896	585.615548	2275.884161
1	1540.444729	458.362560	642.070618	2438.818839
2	1707.348305	460.141436	805.487663	2609.208948
3	1858.833607	466.711604	944.095672	2773.571543
4	1501.567686	467.014275	586.236527	2416.898846

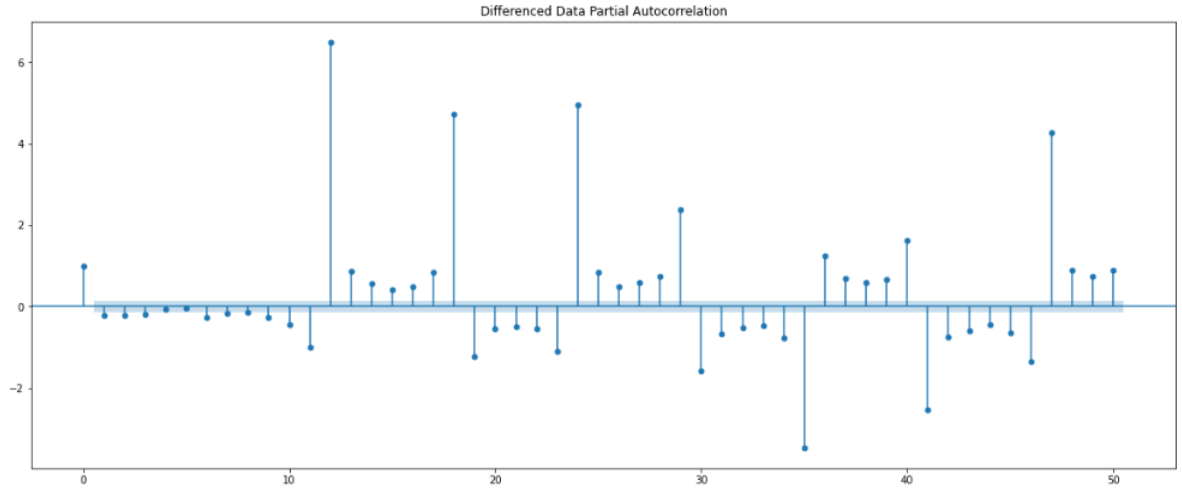
The result of all RMSE Test is as follows

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668
Alpha= 0.086,Beta=1.05,Gamma=0.476,TripleExponentialSmoothing	473.152417
Alpha=0.05,Beta=0.05,Gamma=0.25,TripleExponentialSmoothing	304.645683
ARIMA(2,1,2)	1374.037009
SARIMA(2, 1, 3),(1, 1, 3, 6)	949.981844
SARIMA(3, 1, 3)(3, 1, 0, 12)	331.642387

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.***

ARIMA MODEL





Hence we have taken alpha=0.05

The auto-regressive parameter in an ARIMA model is 'p' which comes from significant lag before which the PACF plots cuts-off 4

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

Hence building an ARIMA model basis these parameter model summary

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 3)	Log Likelihood	-1107.679
Method:	css-mle	S.D. of innovations	1093.029
Date:	Wed, 13 Apr 2022	AIC	2229.358
Time:	20:26:03	BIC	2249.484
Sample:	02-01-1980 - 12-01-1990	HQIC	2237.536

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.8833	3.397	1.732	0.083	-0.775	12.542
ar.L1.D.Sparkling	-0.8782	0.077	-11.438	0.000	-1.029	-0.728
ar.L2.D.Sparkling	-0.5714	0.077	-7.430	0.000	-0.722	-0.421
ma.L1.D.Sparkling	0.3445	0.028	12.396	0.000	0.290	0.399
ma.L2.D.Sparkling	-0.3445	0.037	-9.401	0.000	-0.416	-0.273
ma.L3.D.Sparkling	-1.0000	nan	nan	nan	nan	nan

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.7685	-1.0768j	1.3229	-0.3487
AR.2	-0.7685	+1.0768j	1.3229	0.3487
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.6723	-0.7403j	1.0000	-0.3673
MA.3	-0.6723	+0.7403j	1.0000	0.3673

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

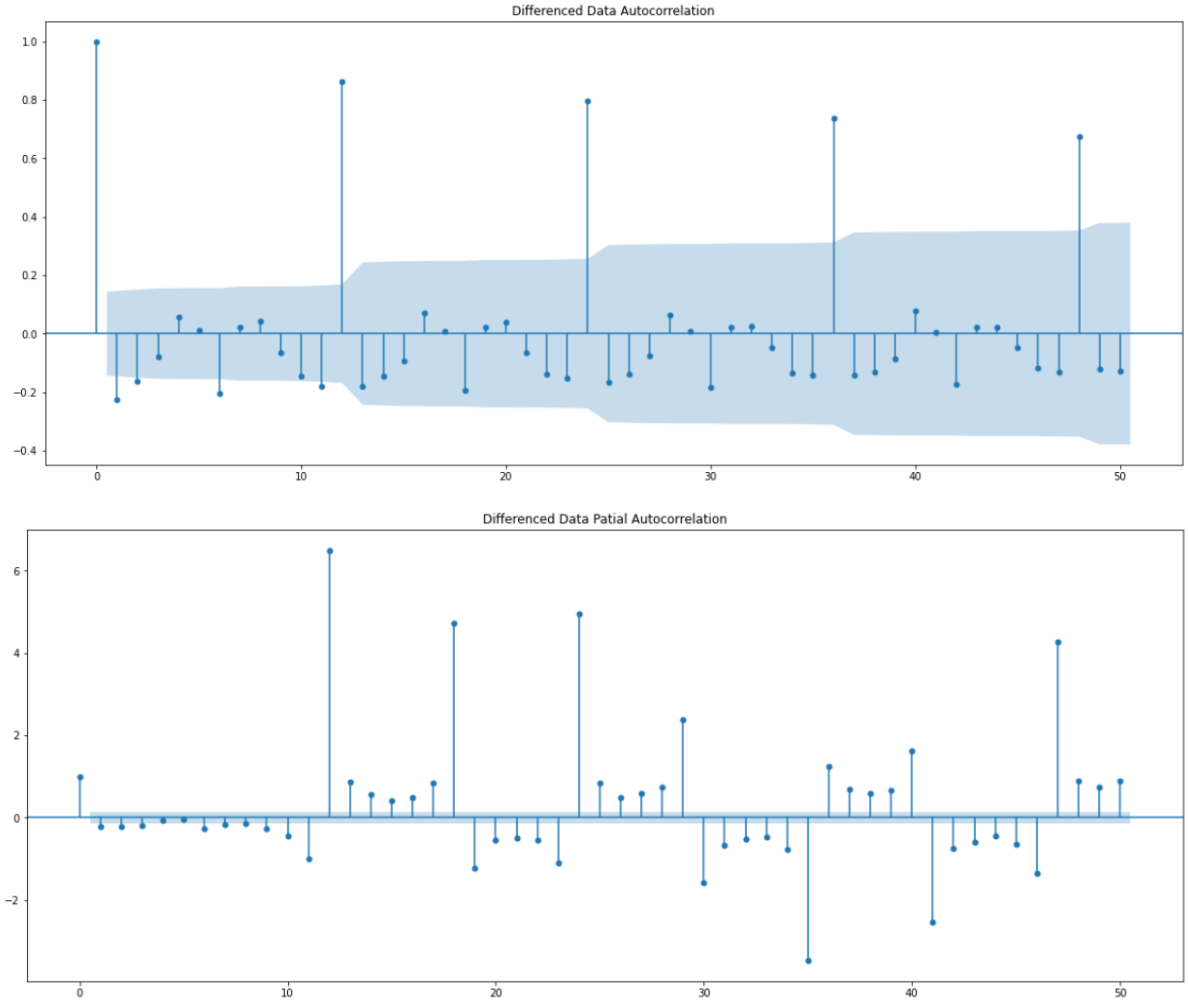
The RMSE value is

```
1393.6371851473552
```

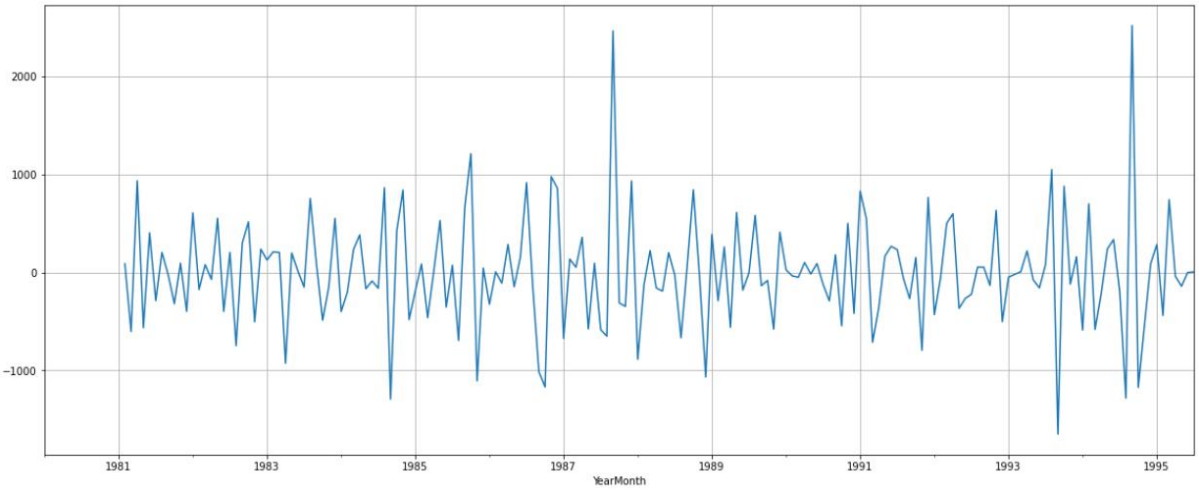
The results of all RMSE value is

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668
Alpha= 0.086,Beta=1.05,Gamma=0.476,TripleExponentialSmoothing	473.152417
Alpha=0.05,Beta=0.05,Gamma=0.25,TripleExponentialSmoothing	304.645683
ARIMA(2,1,2)	1374.037009
SARIMA(2, 1, 3),(1, 1, 3, 6)	949.981844
SARIMA(3, 1, 3)(3, 1, 0, 12)	331.642387
ARIMA(2,1,3)	1393.637185

SARIMA 12



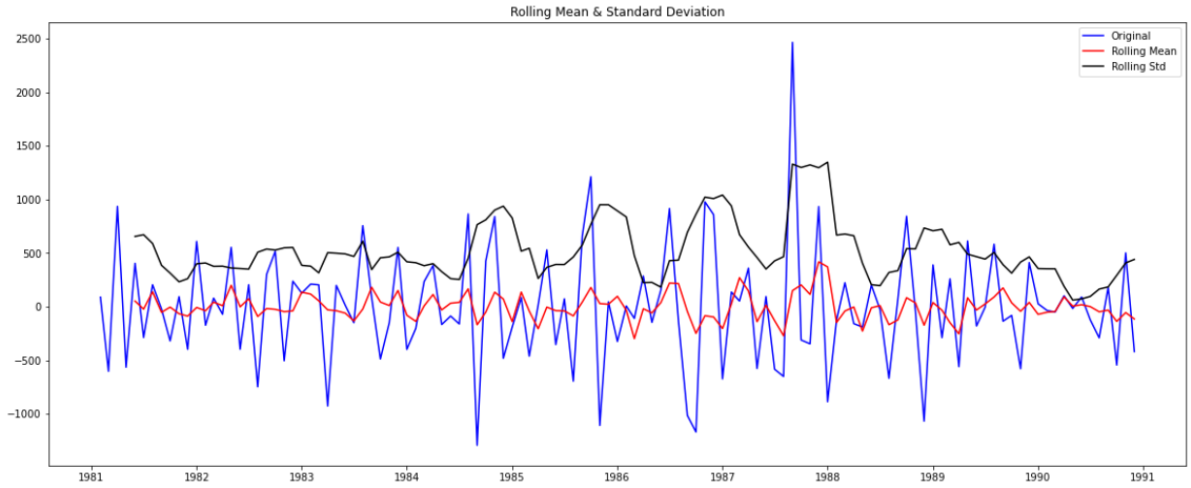
After seasonal differentiating with order of *12 and then differentiating once more*



Now see that there is almost no trend present in data .

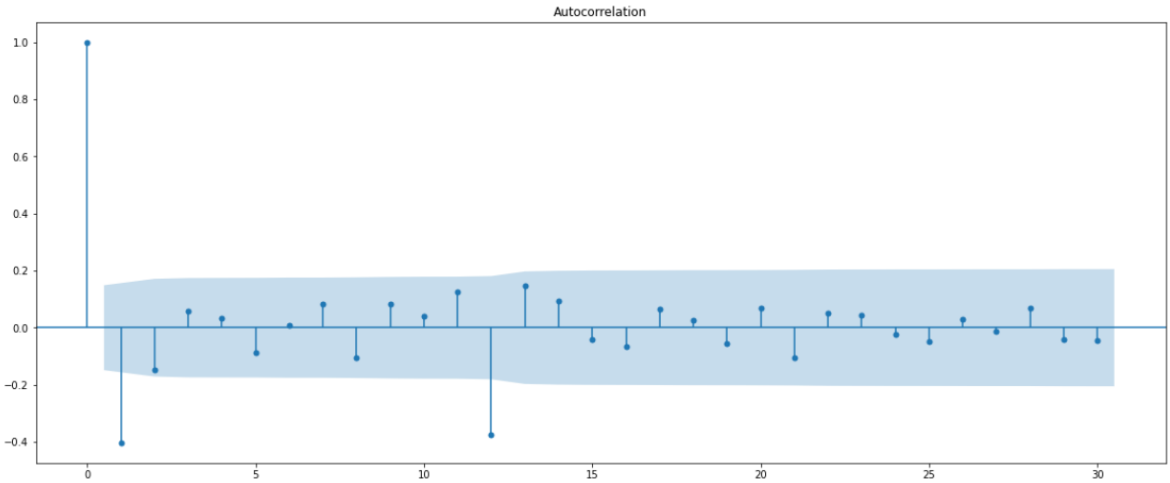
Seasonality is only present in data

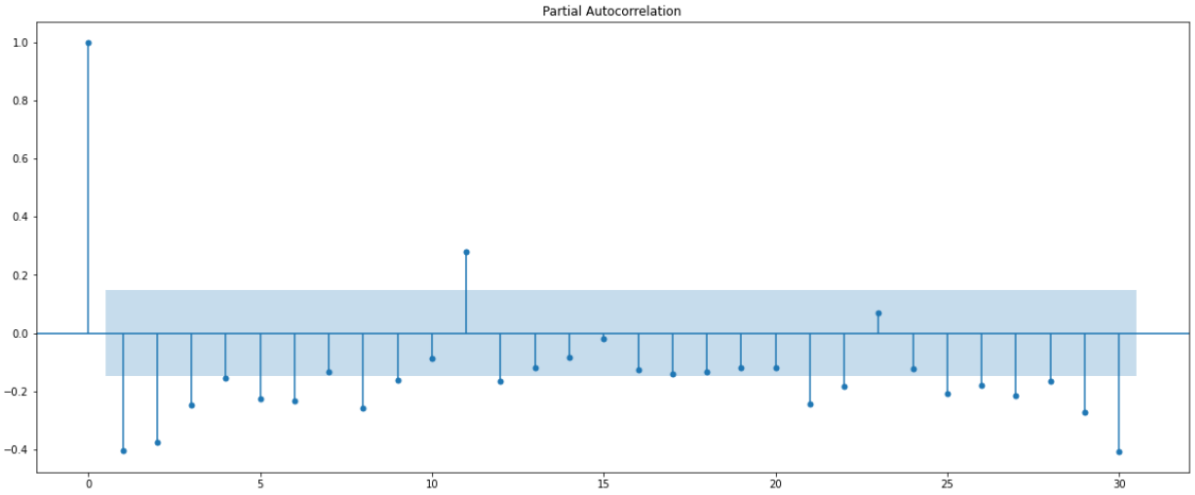
The stationary current of time series




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

The ACF and PACF plot obtained





Here we have taken alpha =0.05

Seasonality = 12

Let us build a model on basis of this parameter

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(6, 1, [1], 12)	Log Likelihood	-333.081
Date:	Wed, 13 Apr 2022	AIC	692.162
Time:	20:41:50	BIC	715.648
Sample:	0	HQIC	700.917
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.5720	0.295	-5.335	0.000	-2.150	-0.995
ar.L2	-0.7451	0.282	-2.645	0.008	-1.297	-0.193
ma.L1	1.2875	0.927	1.389	0.165	-0.529	3.104
ma.L2	-0.3665	0.494	-0.742	0.458	-1.335	0.602
ma.L3	-0.6645	0.632	-1.051	0.293	-1.904	0.574
ar.S.L12	-0.2547	0.554	-0.460	0.645	-1.340	0.830
ar.S.L24	-0.1710	0.196	-0.874	0.382	-0.554	0.212
ar.S.L36	-0.2089	0.208	-1.004	0.315	-0.616	0.199
ar.S.L48	-0.2962	0.220	-1.346	0.178	-0.728	0.135
ar.S.L60	-0.4500	0.355	-1.266	0.205	-1.146	0.246
ar.S.L72	0.0309	0.135	0.229	0.819	-0.234	0.296
ma.S.L12	-0.1172	0.525	-0.223	0.823	-1.146	0.912
sigma2	1.843e+05	1.45e+05	1.270	0.204	-1e+05	4.69e+05

Ljung-Box (L1) (Q):	0.64	Jarque-Bera (JB):	11.42
Prob(Q):	0.42	Prob(JB):	0.00
Heteroskedasticity (H):	0.17	Skew:	1.05
Prob(H) (two-sided):	0.00	Kurtosis:	4.31

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

Predict on the Sparkling_test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1297.356667	432.630273	449.416914	2145.296420
1	1433.324676	528.125500	398.217718	2468.431635
2	1530.986882	528.668186	494.816277	2567.157486
3	1434.623576	596.366740	265.766244	2603.480908
4	1457.339564	602.776194	275.919933	2638.759196

658.1533038240277

The RMSE value is

The RMSE value for all the result is

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668
Alpha= 0.086,Beta=1.05,Gamma=0.476,TripleExponentialSmoothing	473.152417
Alpha=0.05,Beta=0.05,Gamma=0.25,TripleExponentialSmoothing	304.645683
ARIMA(2,1,2)	1374.037009
SARIMA(2, 1, 3),(1, 1, 3, 6)	949.981844
SARIMA(3, 1, 3)(3, 1, 0, 12)	331.642387
ARIMA(2,1,3)	1393.637185
SARIMA(2, 1, 3)(6,1,1,12)	658.153304

From the above table we can say that the RMSE value for seasonality 12 for SARIMA model is 658.153304.

8 - ***Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.***

	Sparkling_test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2point_MovingAverage_Sparkling	3046.976092
4point_MovingAverage_Sparkling	2021.855880
6point_MovingAverage_Sparkling	1521.611250
9point_MovingAverage_Sparkling	1304.618912
Alpha=0.005,SimpleExponentialSmoothing	1316.035487
Alpha=0.05,SimpleExponentialSmoothing	1316.411742
Alpha =0.64,Beta=0,DoubleExponentialSmoothing	2007.238526
Alpha=0.05,Beta=0.05,DoubleExponentialSmoothing	1418.407668
Alpha= 0.086,Beta=1.05,Gamma=0.476,TripleExponentialSmoothing	473.152417
Alpha=0.05,Beta=0.05,Gamma=0.25,TripleExponentialSmoothing	304.645683
ARIMA(2,1,2)	1374.037009
SARIMA(2, 1, 3),(1, 1, 3, 6)	949.981844
SARIMA(3, 1, 3)(3, 1, 0, 12)	331.642387
ARIMA(2,1,3)	1393.637185
SARIMA(2, 1, 3)(6,1,1,12)	658.153304

The model to be built on the whole data

Alpha= 0.05

Beta = 0.05

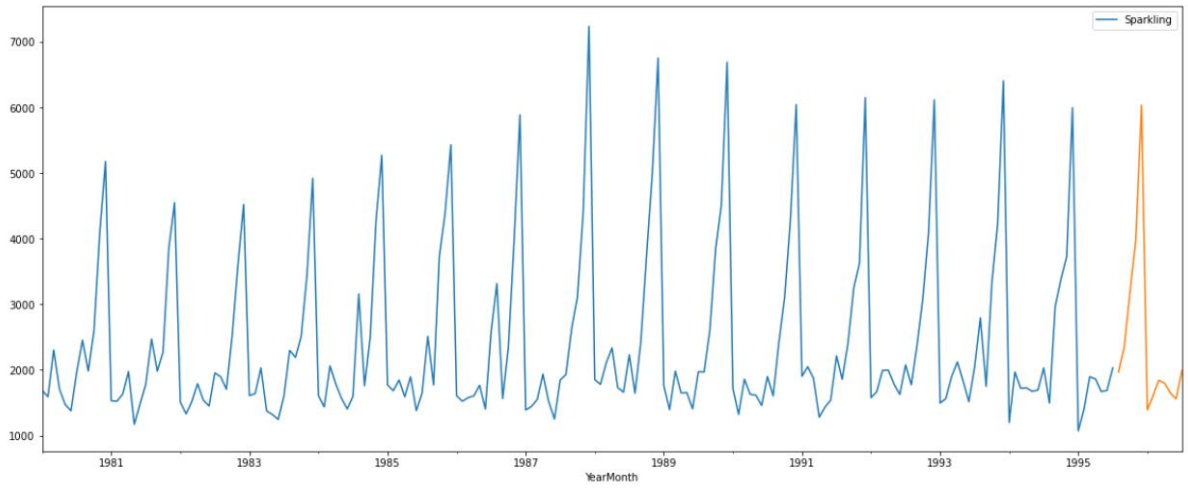
Gamma = 0.25

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.**

After building the model on the basis of the best parameter and the method obtained for training and test and apply on the whole data

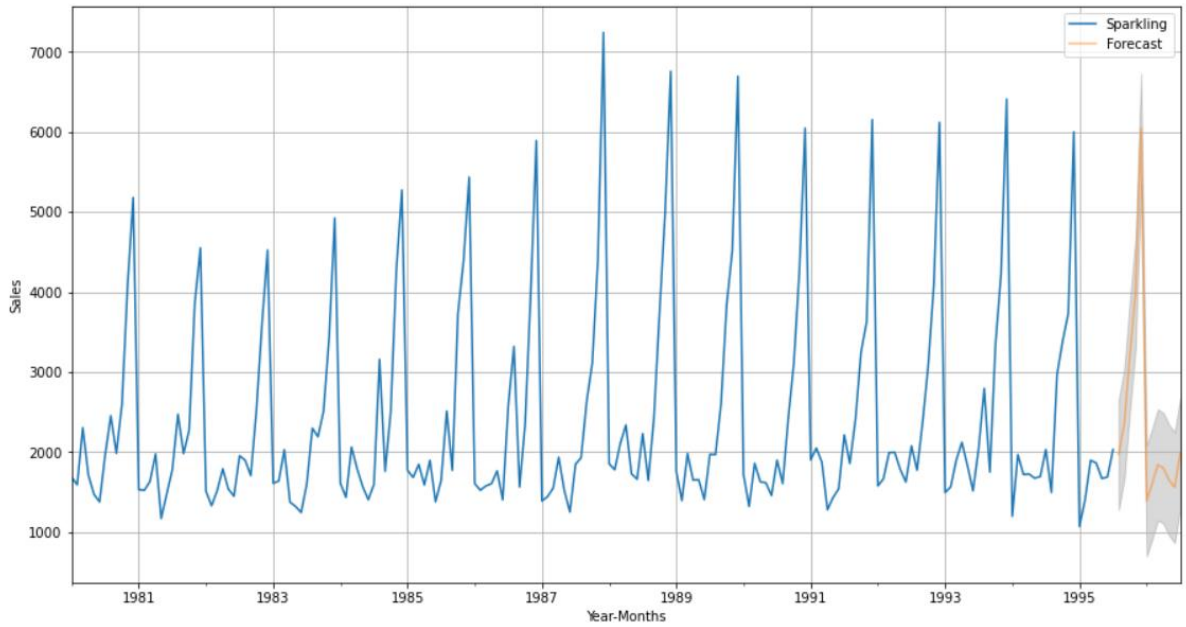
The sparkling RMSE value is

```
Sparkling RMSE: 353.11081605568245
```



With confidence interval

	lower_CI	prediction	upper_ci
1995-08-01	1275.521240	1968.500088	2661.478937
1995-09-01	1657.100572	2350.079421	3043.058269
1995-10-01	2508.324687	3201.303535	3894.282384
1995-11-01	3284.523123	3977.501972	4670.480820
1995-12-01	5346.278735	6039.257584	6732.236432



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

- *The yearly boxplots also show that the Sparkling Wine Sales have increased with the passing years.*
- *There is a clear distinction of Sparkling Wine sales within different months spread across various years.The highest such numbers are being recorded towards the end of the years with a huge spike in the month of December possibly due to festive seasons*
- *Almost 95% of sales are within 5000 units of Sparkling Wines*
- *In this dataset, after comparing various models and comparing on RMSE, we analyse that the Triple Exponential Smoothing proved to be a better model with parameters alpha = 0.05, Beta = 0.05, Gamma = 0.25*
- *Recommendations:*
- *Introduce various promotional offers across the outlets and around the year to attract customers basis price especially in the summer season where the sales seems to show low sales values.*
- *Organize testing outlets for customers to taste the Sparkling wine as few may not be aware of the taste and quality of the same*
- *Conducting training of outlets to promote these wines to customers personally based on their requirements.*

ROSE

Dataset

Questions:

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
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.
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.
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.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
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.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

1 – Read the data as ana appropriate Time Series data and plot the data.

Head of Dataset

	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Tail of dataset

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Description of dataset

	Rose
count	187.000000
mean	89.914439
std	39.238325
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Null values:

```
Rose    2
dtype: int64
```

Shape and info of dataset

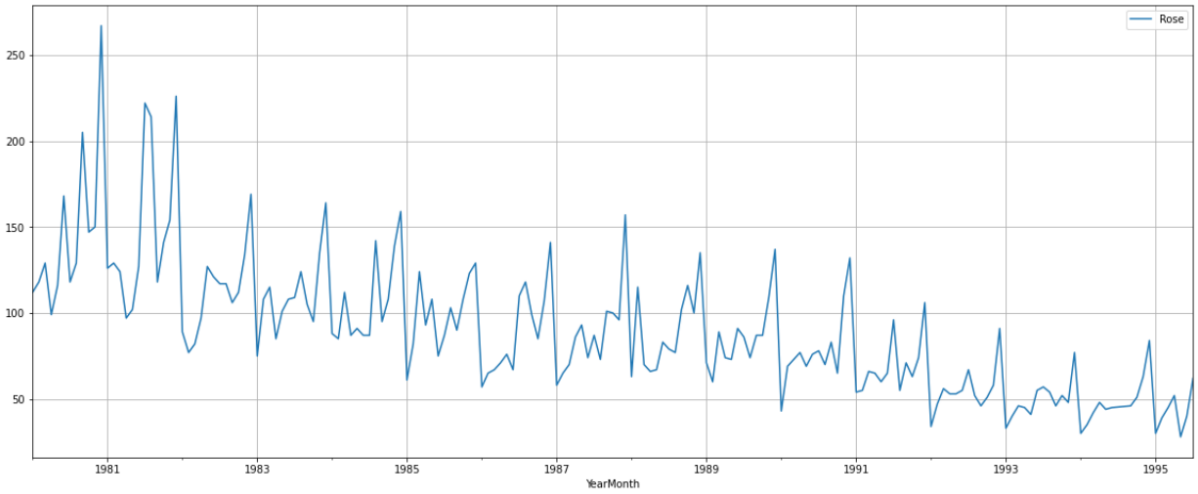
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 7.0 KB
```

	Rose
YearMonth	
1994-07-01	45.333333
1994-08-01	45.666667

Observation

1. We can see that there are in the dataset there are 187 rows and 1 column.
2. As check found that there are no null values in the Sparkling dataset.
3. From info we can see that the values are of (int) type.
4. From description we find the mean, median, max, 25%,50%,75% values.
5. By using interpolation method to impute missing data basis neighbouring data values .in above table the imputed value for the concerned time stamp.

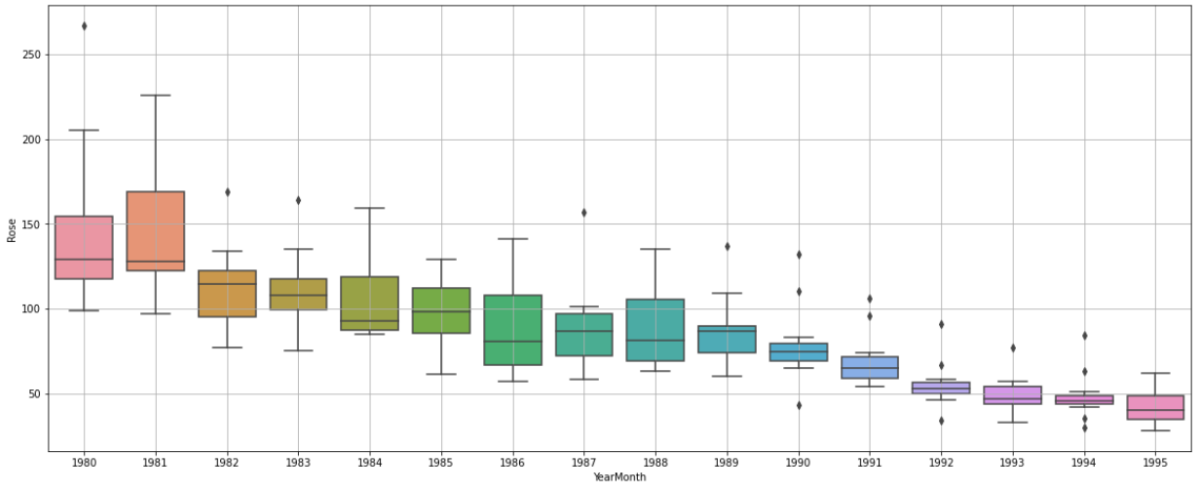
Rose Plot



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

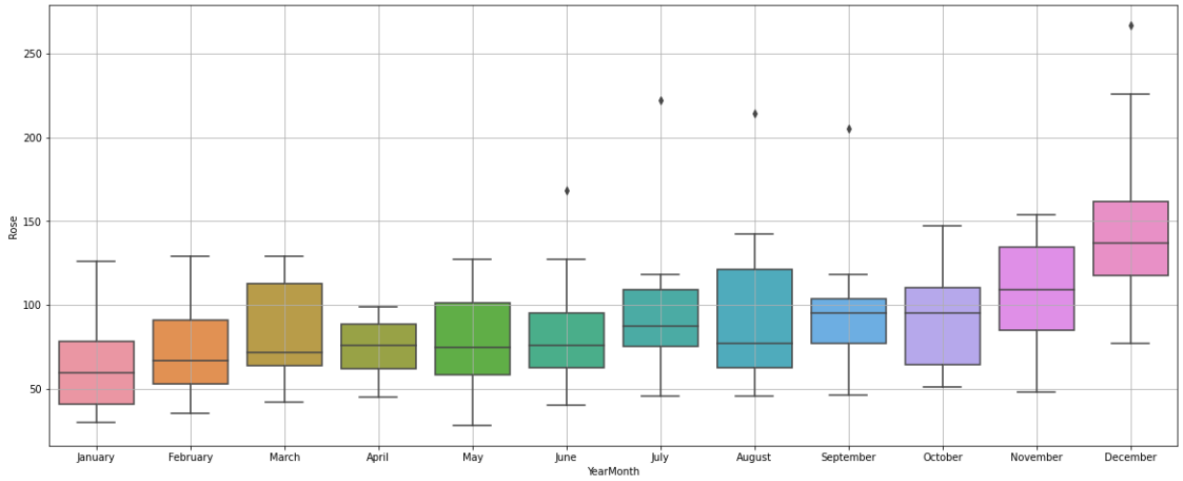
Let us check the spread of sales across different years.

Yearly plot



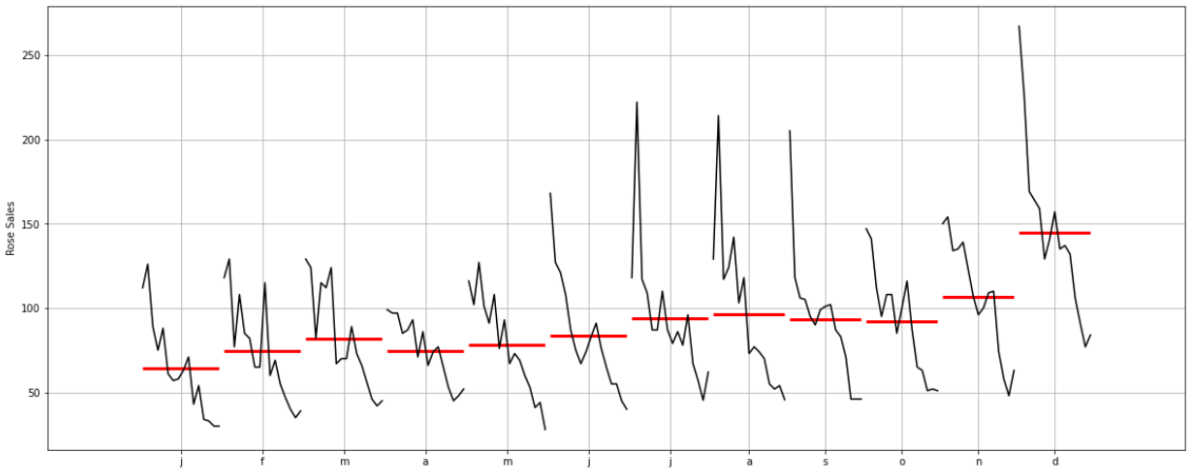
The yearly boxplot also shows that the sparkling wine sales have increase with the passing year.

Monthly plot



There is a clear distinction of Rose Wine sales within different months spread across various years. The highest such numbers are being recorded towards the end of the years with a huge spike in the month of December.

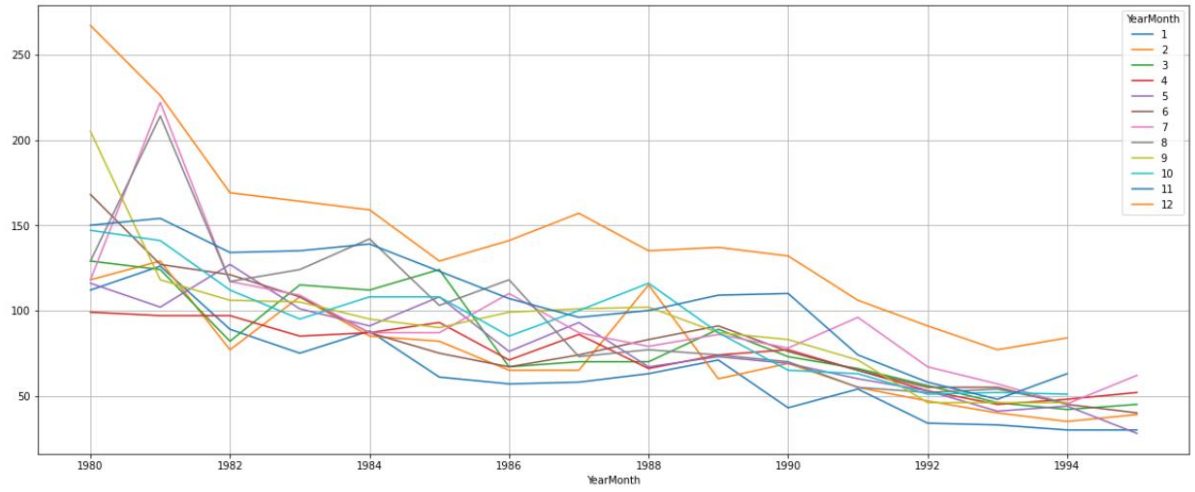
Let us check Time series month plot to understand the spread of Sales across different years and within different months across years.



As also observed from the monthly plot, sales for November – December month have been the highest across the years, but have started decreasing with passing years.

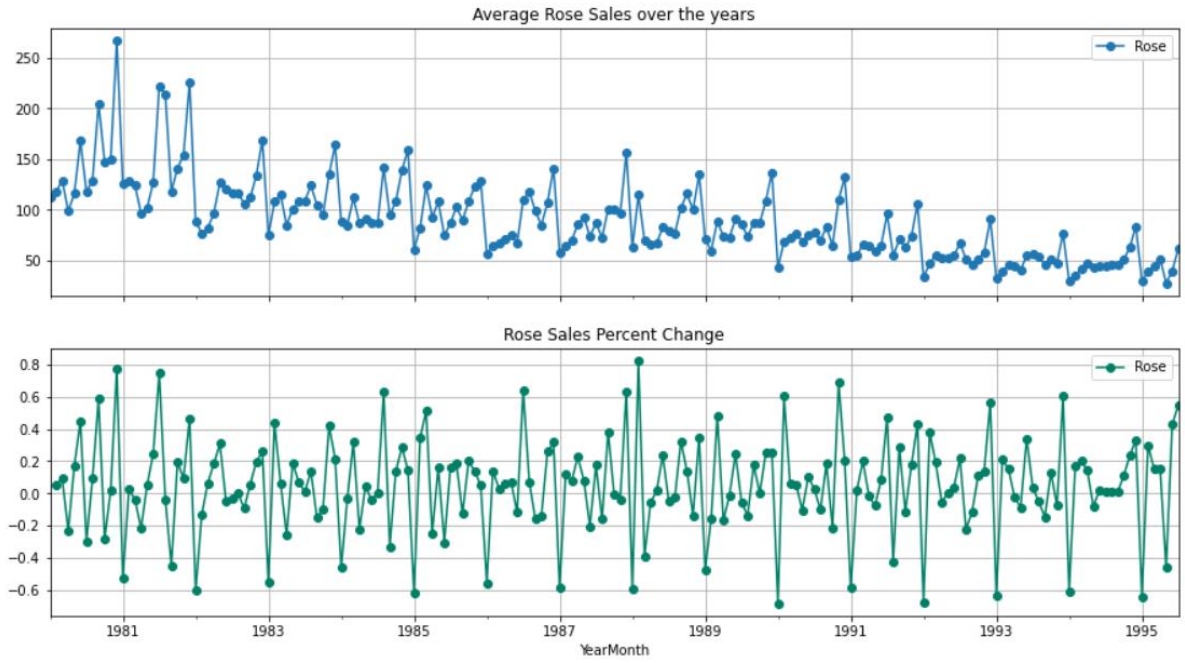
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.333333	45.666667	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN

Monthly sales across years



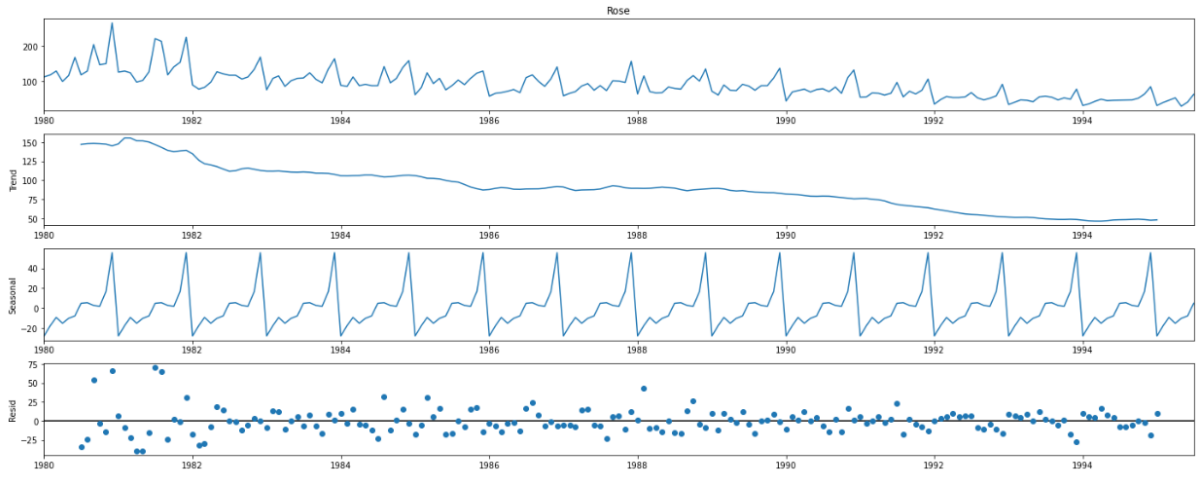
Sales plot across various years in different months

Let us check the average Rose per month and the month on month percentage change of Rose.

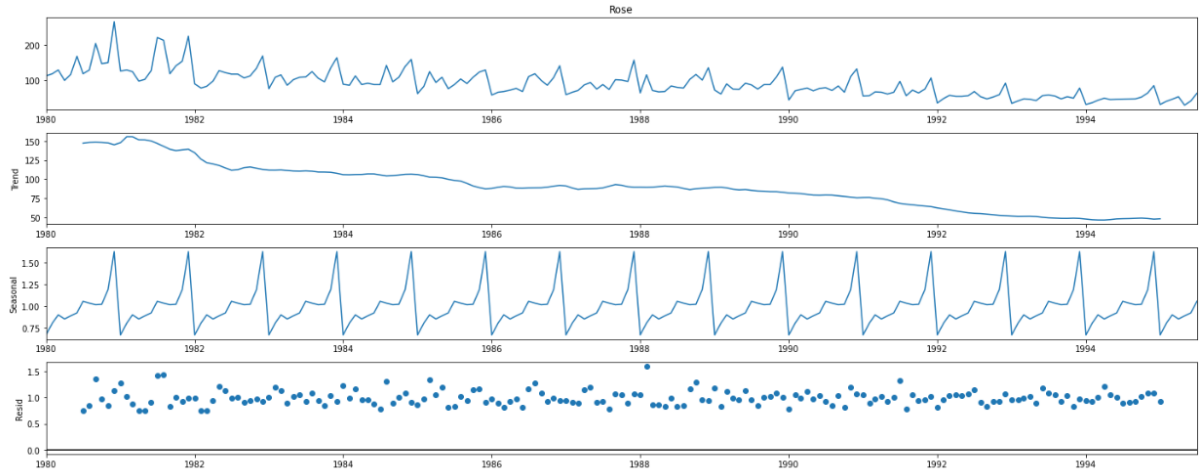


The above two graphs tells us the Average 'Rose' and the Percentage change of 'Rose' with respect to the time.

Seasonal Decomposition: Additive model



Seasonal Decomposition: Multiplicative model



From above the additive model suggest the forecast future data.

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

First few rows of Rose_training Data

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of Rose_training Data

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

Data for training before 1990

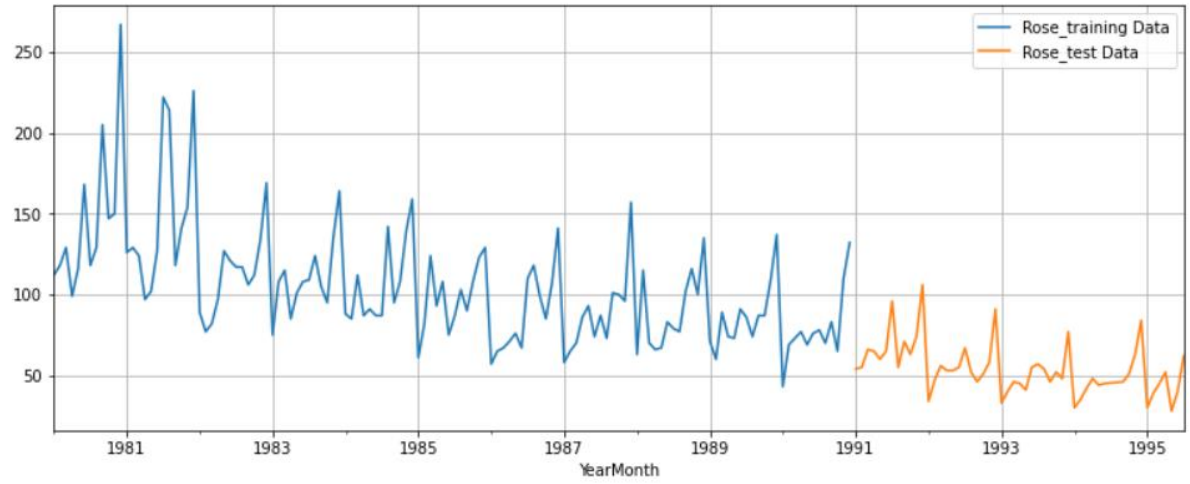
First few rows of Rose_test Data

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

Last few rows of Rose_test Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Data for testing after 1991



As per the graph and the data there are 132 rows and 1 column in training dataset

And 55 rows and 1 column in testing dataset.

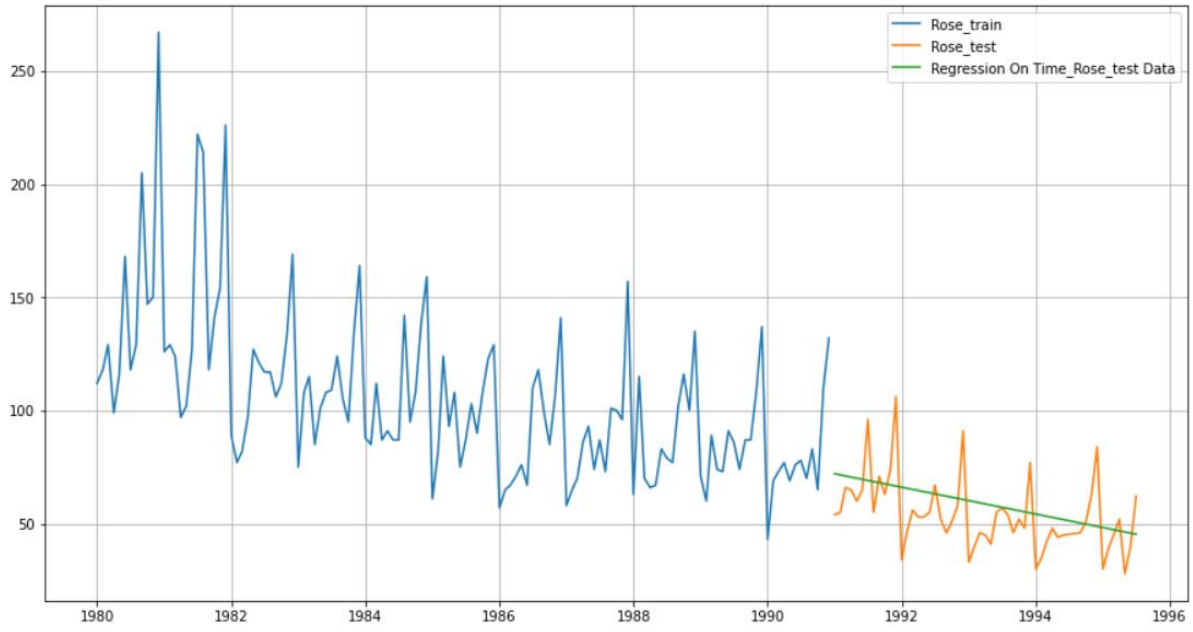
4 - Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other model such as regression ,naive forecast models and simple average models. Should also be built on the training data and check the performance on the test data using RMSE.

Model 1 – Linear Regression

```
Rose_training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Rose_test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

Fitting the model on train and Test data with parameters as below:

Following is the RMSE values on test data after fitting the model basis training data



Linear Regression on time sparkling data

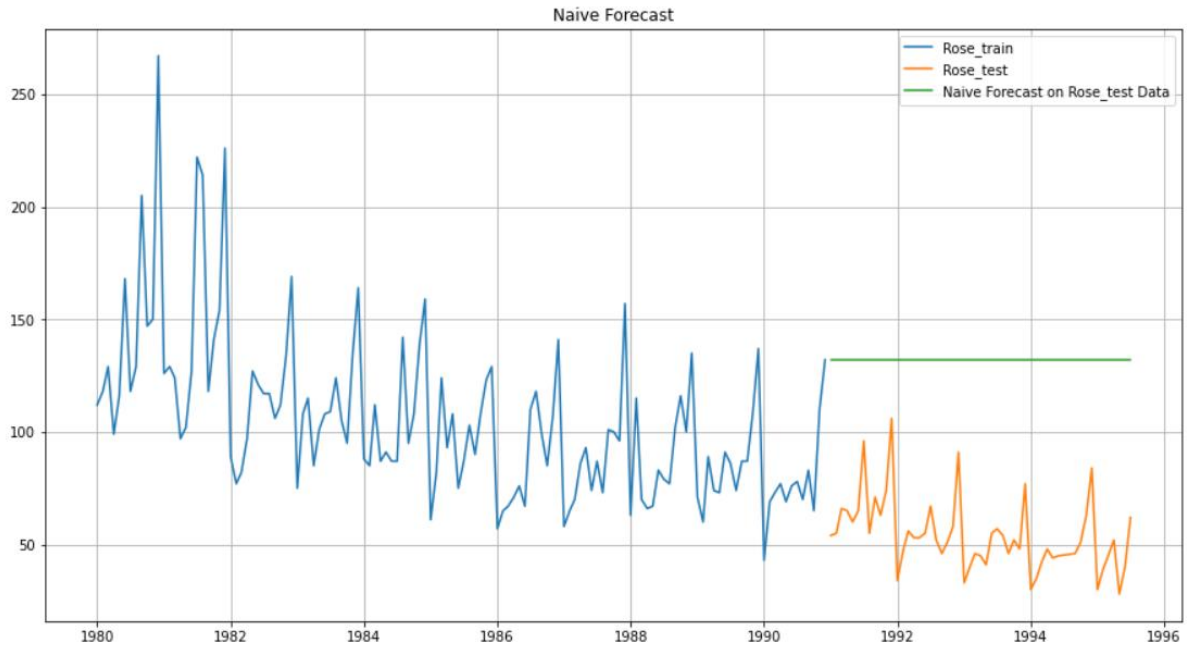
For RegressionOnTime forecast on the Rose_test Data, RMSE is 15.269

	Rose_test RMSE
RegressionOnTime	15.268955

The Linear Regression result of RMSE

Model – 2 – Naïve Bayes

As depicted in the formula predicted values is same as last value, hence on plot it shows a flat line since all the values are same



The Naïve Bayes result for RMSE

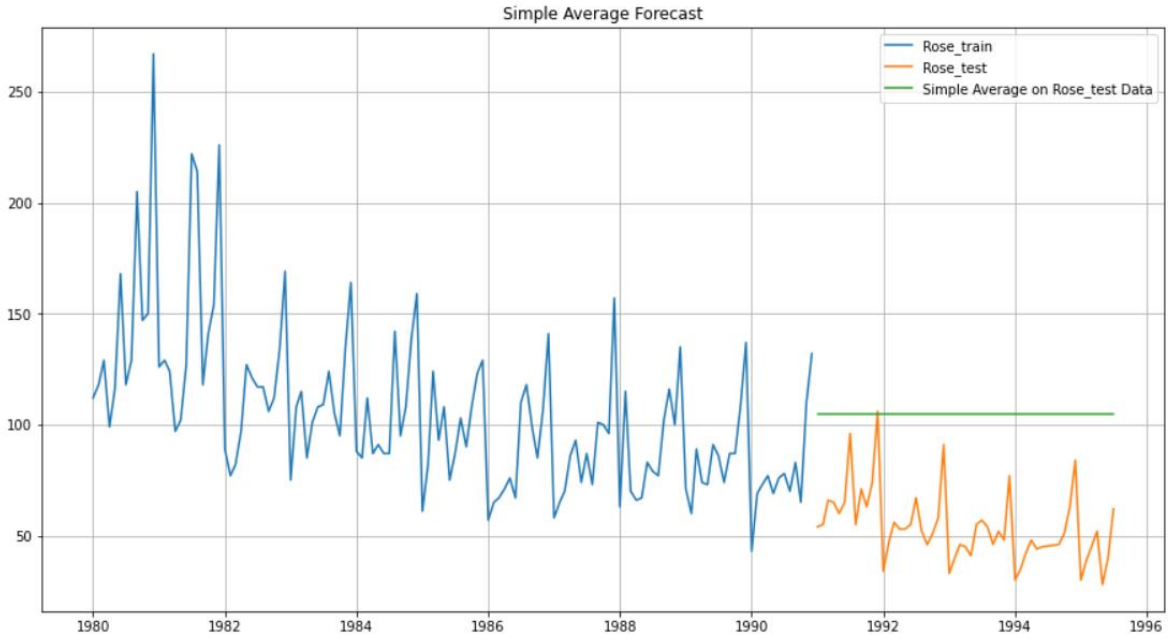
```
For NAive Based forecast on the Rose_test Data,  RMSE is 79.719
```

Comparing Results for all RMSE is

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773

3.Simple Average

In this method predicted values are calculated basis mean value of the entire data, just like Naïve approach it will also show a flat line



The RMSE result for simple average dataset is

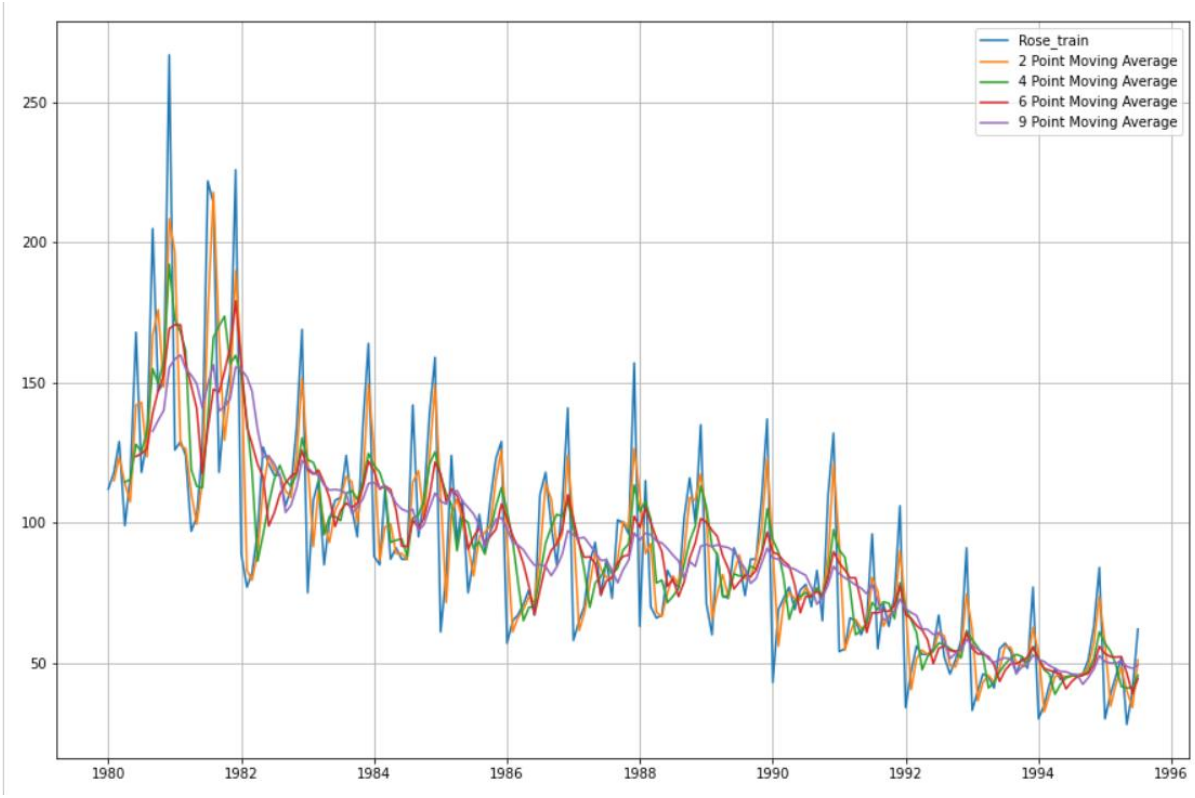
```
For Simple Average forecast on the Rose_test Data,  RMSE is 53.461
```

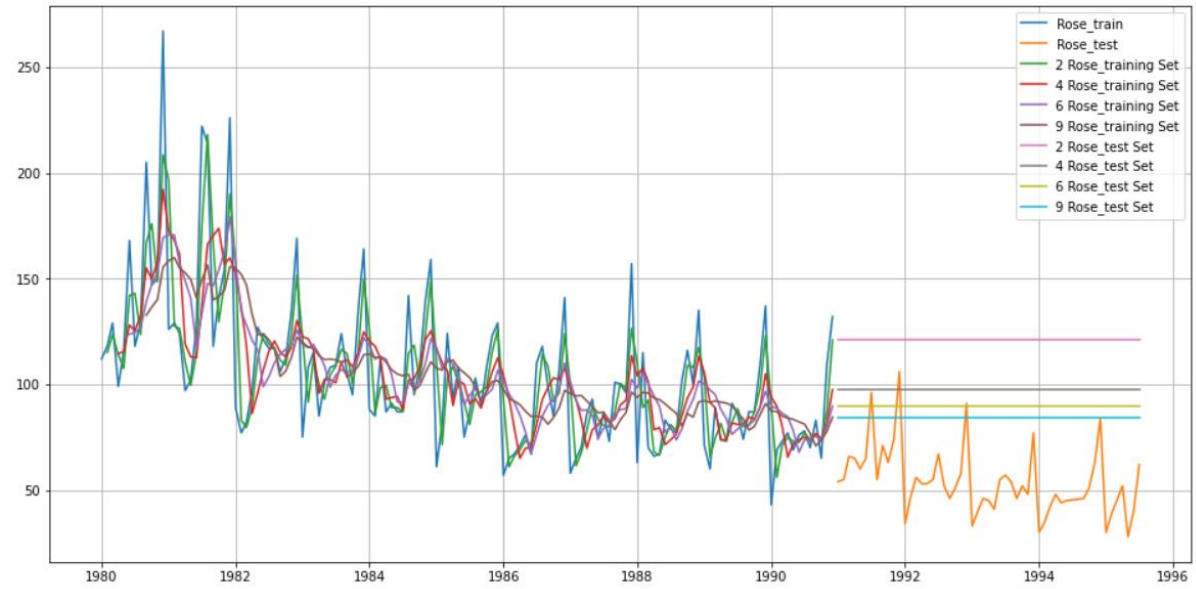
Comparing results for all RMSE is

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570

4 – Moving Average

In this method, just like Simple average predicted values are calculated basis mean value of the previous data but on rolling basis, the number of old data to be used to predict next values can be anything, here we chose 2, 4, 6, 9 as rolling parameters. As seen from Below Plot, Moving average values of all the given parameters closely follow original data in training set, however for test set, since previous values are supposed to be not known, hence the last predicted values of train data becomes the value of entire test data or future data, hence it shows a flat line on test data.



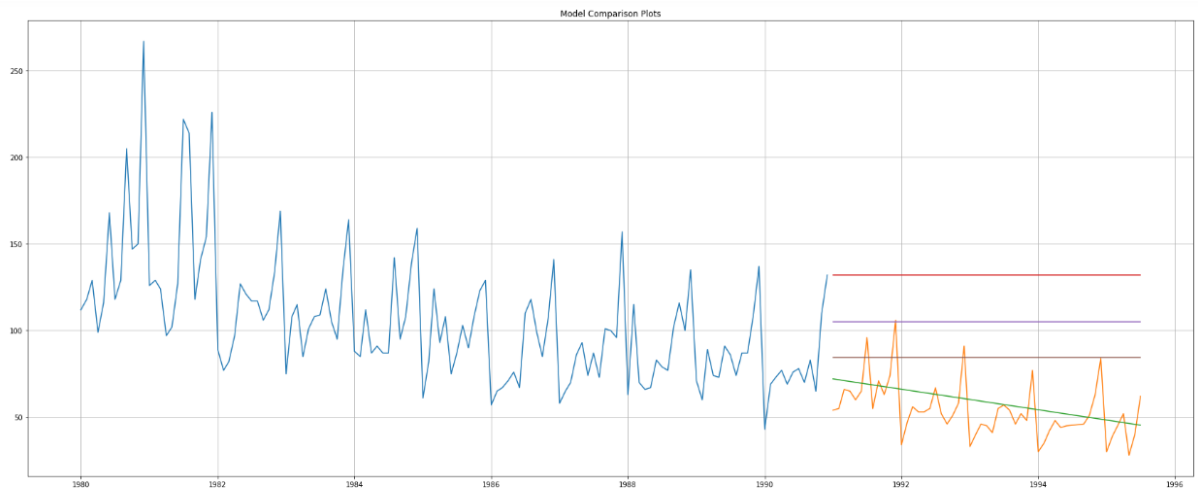


The RMSE results for points 2,4,6,9 is

```
For 2 point Moving Average Model forecast on the Rose_training Data,  RMSE is 68.970
For 4 point Moving Average Model forecast on the Rose_test Data,  RMSE is 46.404
For 6 point Moving Average Model forecast on the Rose_test Data,  RMSE is 39.126
For 9 point Moving Average Model forecast on the Rose_test Data,  RMSE is 34.411
```

Comparing the results for all RMSE is

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938



Model comparison PLots

Model – 5 – Simple Exponential Smoothing

Simple Exponential Smoothing is used for time series prediction when the data particularly does not follow any:

Trend: An upward or downward slope

Seasonality: Shows a particular pattern due to seasonal factors like Hours, days, Year, etc.

SES works on weighted averages i.e. the average of the previous level and current observation. Largest weights are associated with the recent observations and the smallest weights are associated with the oldest observations.

The decrease in weight is controlled by the smoothing parameter which is known as α(alpha) here. α(alpha) value can be between 0 to 1:

o α(alpha)=0: Means that forecast for future value is the average of historical data.

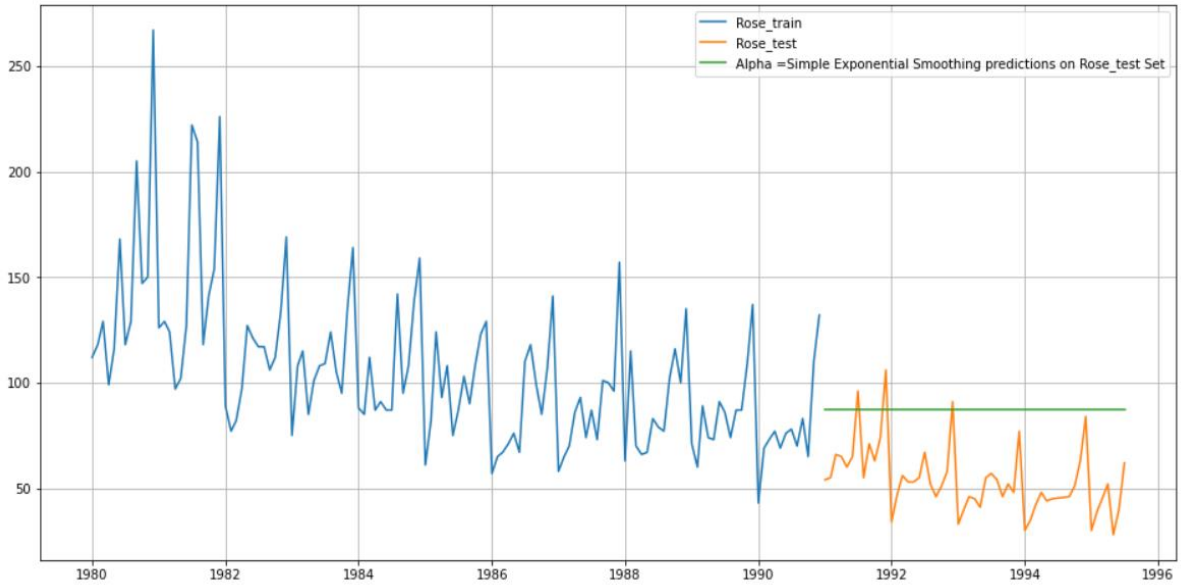
o α(alpha)=1: Means that forecast for all future value is the value of the last observation

Below is the formula for Simple Exponential Smoothing:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

After auto-fitting the model, following is the parameter obtained:

```
{'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}
```



model evaluation for alpha=0.987 simple exponential smoothing

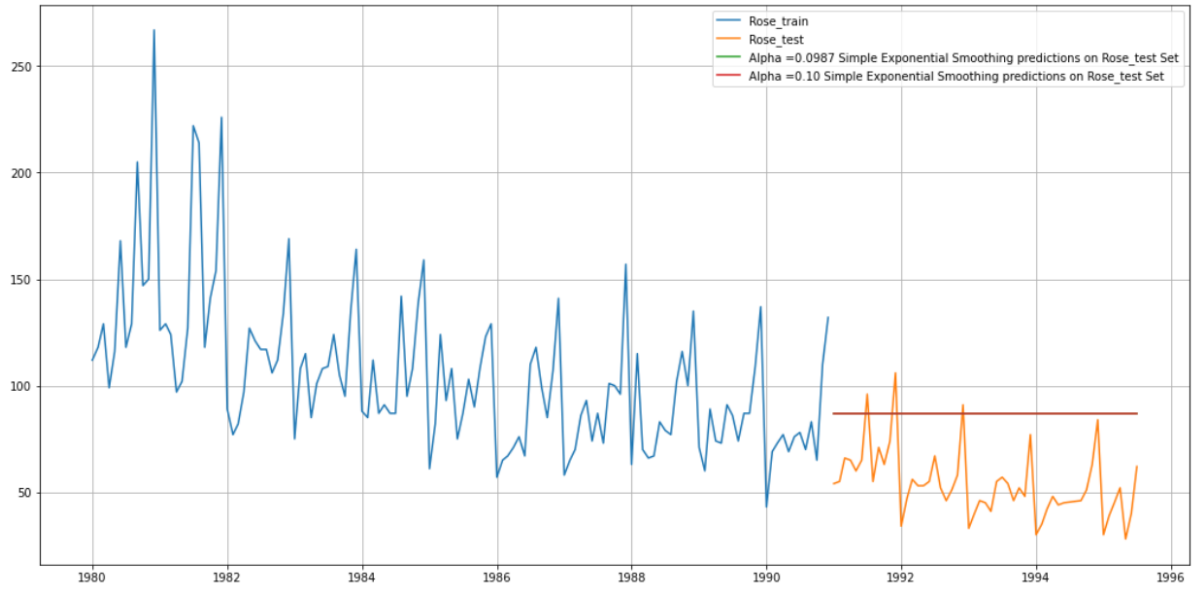
```
For Alpha =0.0987 Simple Exponential Smoothing Model forecast on the Rose_test Data, RMSE is 36.796
```

Comparing the results for all RMSE is

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227

After trying out manually various alpha values the model, Alpha = 0.10 is the best parameter obtained with lowest TEST RMSE, below is the predicted values plotted on Time scale. As observed manual and automatic alpha values were very close, hence obtained similar forecast for both the values.

	Alpha Values	Rose_train RMSE	Rose_test RMSE
2	0.10	31.815610	36.828033
1	0.05	32.449102	37.011448
3	0.15	31.809845	38.722125
4	0.20	31.979391	41.361876
5	0.25	32.211871	44.360796
6	0.30	32.470164	47.504821
7	0.35	32.744341	50.665672
8	0.40	33.035130	53.767406
9	0.45	33.346578	56.767133
10	0.50	33.682839	59.641786
0	0.00	36.719452	60.243378
11	0.55	34.047042	62.378989
12	0.60	34.441171	64.971288
13	0.65	34.866356	67.412903
14	0.70	35.323261	69.698162
15	0.75	35.812435	71.820852
16	0.80	36.334596	73.773992
17	0.85	36.890835	75.549736
18	0.90	37.482782	77.139276
19	0.95	38.112735	78.532696



As per alpha = 0.10 the RMSE value will be little change .

So the new value will be

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033

Method 6 – Double Exponential Smoothing

Double Exponential Smoothing (DWA) is also called as Holts Double Exponential Smoothing.

Double Exponential Smoothing is extended form of Simple Exponential Smoothing.

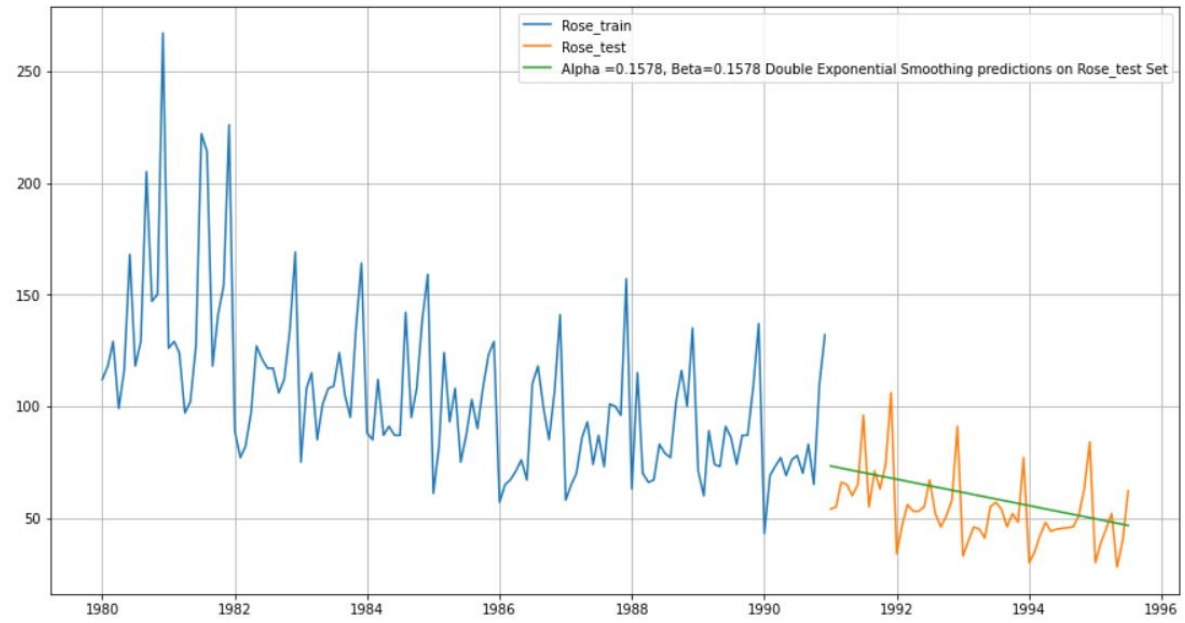
Double Exponential Smoothing technique is used for forecasting with trending data which has level and trend but it does not have seasonality.

Two parameters α and β are estimated in this model.

Level and Trend are accounted for in this model.

After auto-fitting the model on the training data, following is the parameters obtained:

```
{'smoothing_level': 0.017549790270679714,
 'smoothing_trend': 3.236153800377395e-05,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 138.82081494774005,
 'initial_trend': -0.492580228245491,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



For Alpha =0.1578, Beta=0.1578 Double Exponential Smoothing Model forecast on the Rose_test Data, RMSE is 15.707

```
For Alpha =0.1578, Beta=0.1578 Double Exponential Smoothing Model forecast on the Rose_test Data, RMSE is 15.707
```

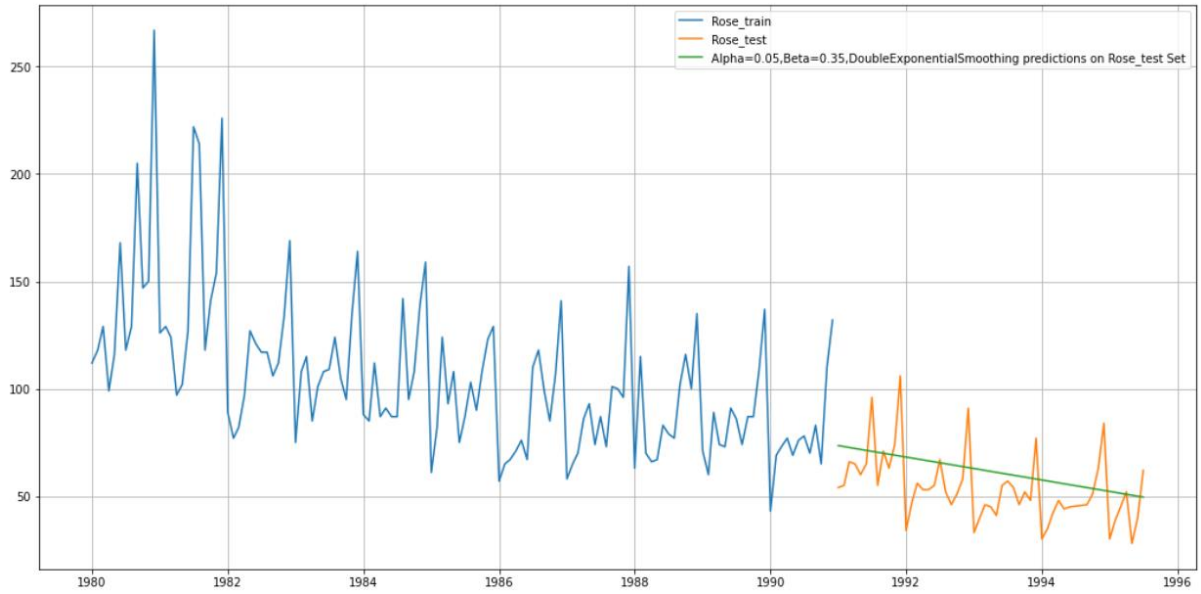
Comparing the result of all the RMSE result

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033
Alpha =0.1578,Beta=0.1578,DoubleExponentialSmoothing	15.707052

After manually fitting various model, below are the best parameters with lowest RMSE Scores on test Data:

	Alpha Values	Beta Values	Rose_train RMSE	Rose_test RMSE
6	0.05	0.35	36.233997	16.329097
5	0.05	0.30	36.616877	18.624638
2	0.05	0.15	39.106563	23.716964
0	0.05	0.05	49.734056	31.526909
7	0.05	0.40	35.783737	31.578177

Hence can be seen that Alpha =005, Beta=0.35, comes out to be better parameters and below is predicted test values on time series scale.



Comparing all results of RMSE

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033
Alpha =0.1578,Beta=0.1578,DoubleExponentialSmoothing	15.707052
Alpha=0.05,Beta=0.35,DoubleExponentialSmoothing	16.329097

Method – 7 - Triple Exponential Smoothing

This is an extension of Holt's method when seasonality is found in the data.

Fore caste equation: $Yt+1=lt+bt+st-m(k+1)$

Level Equation: $lt=\alpha(Yt-st-m) + \alpha(1-\alpha)Yt-1$, $0 < \alpha < 1$

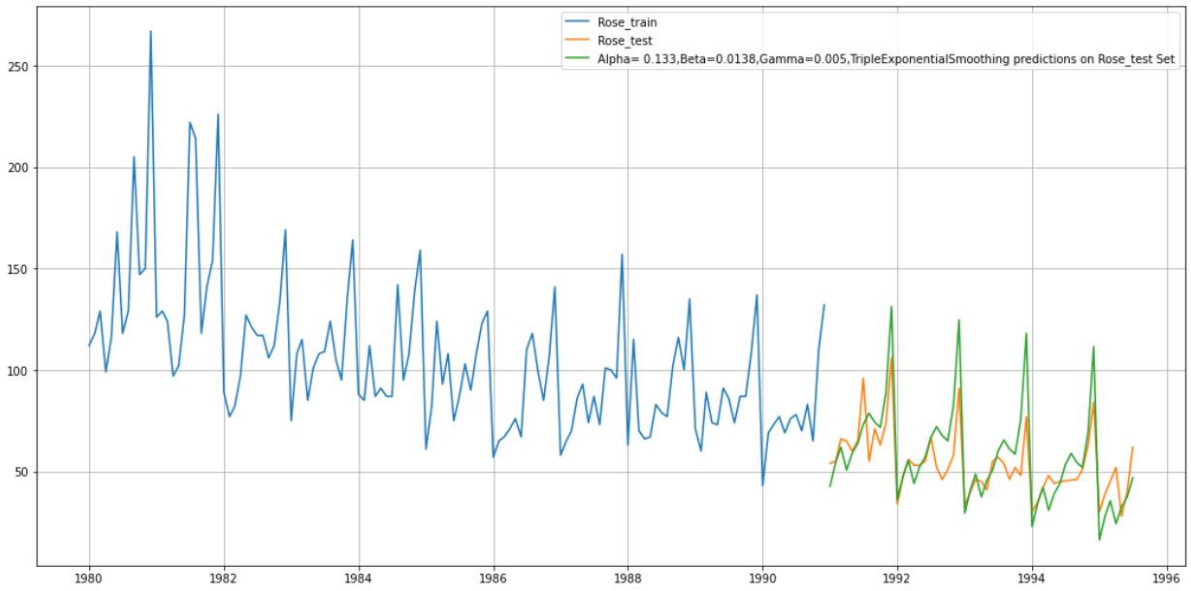
Trend Equation: $bt=\beta(lt-lt-1) + (1-\beta)bt-1$, $0 < \beta < 1$

Seasonal Equation: $\gamma(Yt-lt-1-bt-1) + (1-\gamma)st-m$, $0 < \gamma < 1$

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, $\beta$and γ. This is a general method and a true multi-step ahead forecast.

After auto-fitting the model on the training data, following is the parameters obtained and predicted value plot on time series:

```
{'smoothing_level': 0.08485622209289158,
 'smoothing_trend': 0.0005280630369796539,
 'smoothing_seasonal': 0.006764526794519119,
 'damping_trend': nan,
 'initial_level': 77.31958915163194,
 'initial_trend': -0.5501794952033382,
 'initial_seasons': array([ 38.63205024,  50.94041582,  59.06227683,  48.26306762,
        57.05566819,  62.6629802 ,  72.54958097,  78.56552615,
        74.62626287,  72.65336605,  90.54607132, 133.36580632]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```




```
Alpha= 0.133,Beta=0.0138,Gamma=0.005, Triple Exponential Smoothing Model forecast on the Rose_test Data,  RMSE is 14.257
```

For the above graph the RMSE value is given ***14.257***

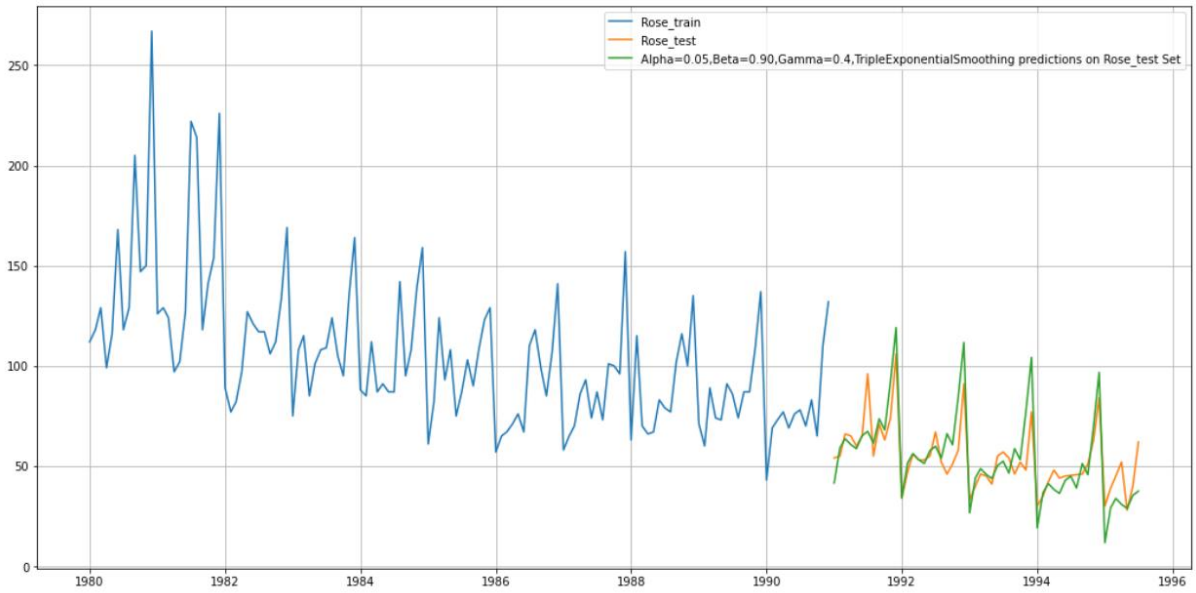
Now comparing all the results of RMSE

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033
Alpha =0.1578,Beta=0.1578,DoubleExponentialSmoothing	15.707052
Alpha=0.05,Beta=0.35,DoubleExponentialSmoothing	16.329097
Alpha= 0.133,Beta=0.0138,Gamma=0.005,TripleExponentialSmoothing	14.257122

After manually fitting various model, below are the best parameters with lowest RMSE Scores on test Data:

	Alpha Values	Beta Values	Gamma Values	Rose_train RMSE	Rose_test RMSE
347	0.05	0.90	0.40	26.036441	11.652209
366	0.05	0.95	0.35	26.257775	11.689432
328	0.05	0.85	0.45	25.896403	11.704586
309	0.05	0.80	0.50	25.820490	11.811688
146	0.05	0.40	0.35	24.528346	11.865056

Hence can be seen that Alpha =005, Beta=0.90 and Gamma = 0.40, comes out to be better parameters and below is predicted test values on time series scale.



The RMSE value is 11.6522

As by comparing all the results

We see the best model is Triple Exponential Smoothing with multiplicative seasonality with the parameter α = 0.05, β = 0.90 and γ = 0.40.

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.

Stationary process: A process is said to be stationary if its mean and variance are constant over a period of time and, the correlation between the two time periods depends only on the distance or lag between the two periods.

Mathematically, let Yt *be a time series with these properties:*

Mean: $E(Yt)=\mu$

Variance: $Var(Yt)=E(Yt-\mu)2=\sigma2$

Correlation: $\rho k=E[(Yt-\mu)(Yt+k-\mu)/(\sigma t\sigma t+k)]$

Where ρk*is the correlation (or auto-correlation) at lag* k*between the values of* Yt*and* $Yt+k$

So, if mean, variance and correlation (or auto-correlation) of time series data is constant (at different lags) no matter at what point of time it is measured; i.e. if they are time invariant, the series is called a stationary time series. A series not possessing these properties is termed as a non-stationary timeseries.

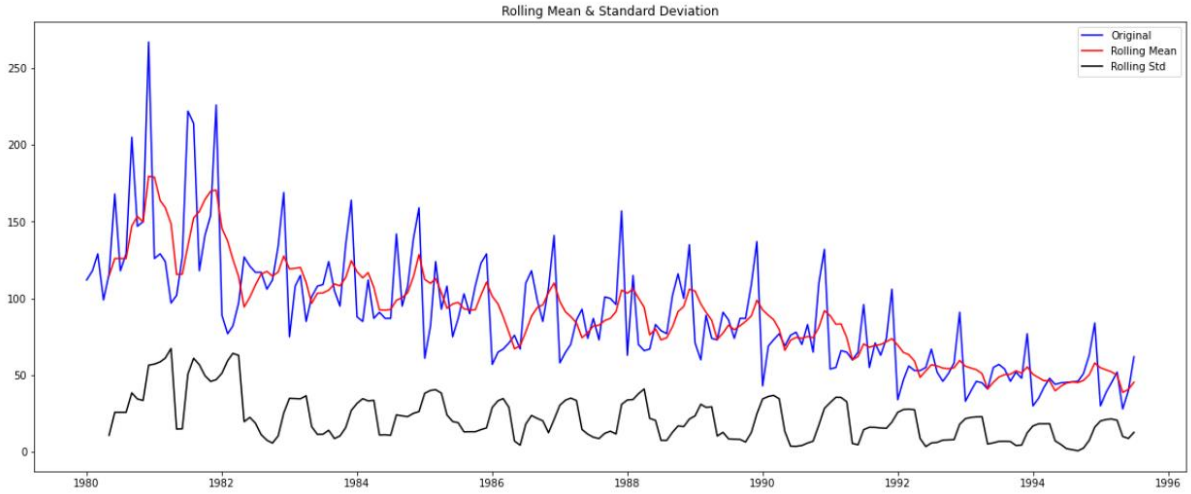
We are going to use Augmented Dickey – Fuller test to test stationary of the data

Null hypothesis – data is non – stationary

Alternate hypothesis – data is stationary

Alpha = 0.05

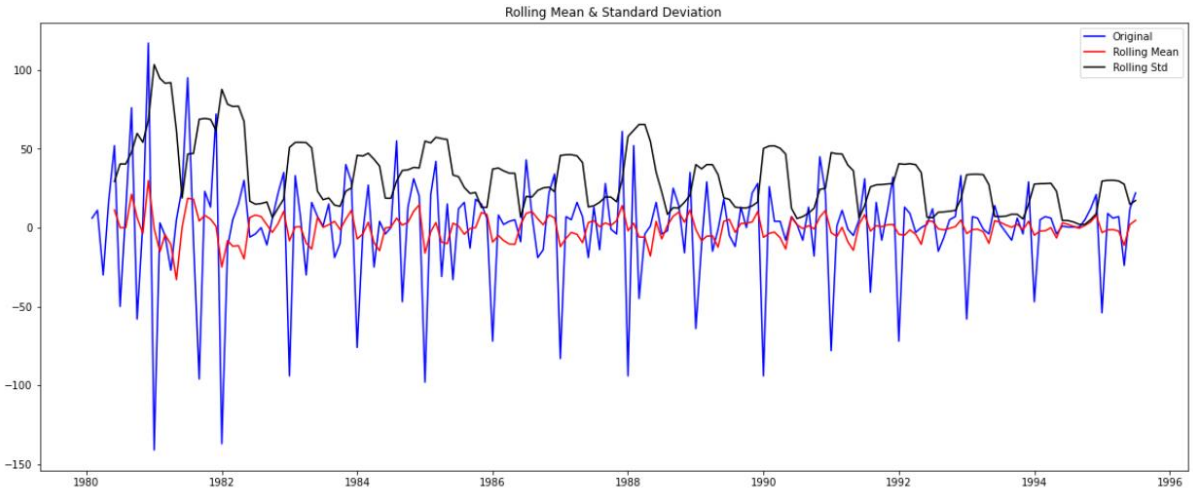
Here if the value of p comes out to be less than 0.05 , then we reject the null hypothesis




```
Results of Dickey-Fuller Rose_test:
Rose_test Statistic             -1.876699
p-value                          0.343101
#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
```

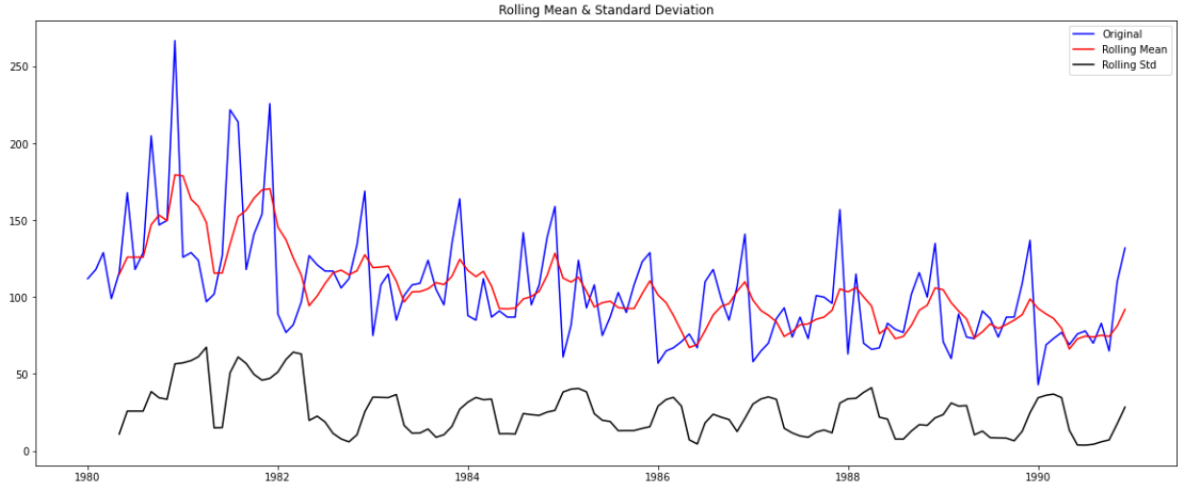
Stationary test result on original time series

Original Time Series, fails the test, hence it can be concluded that the time series is not stationary, hence requires differentiating with respect to previous values.




```
Results of Dickey-Fuller Rose_test:
Rose_test Statistic            -8.044392e+00
p-value                         1.810895e-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
```

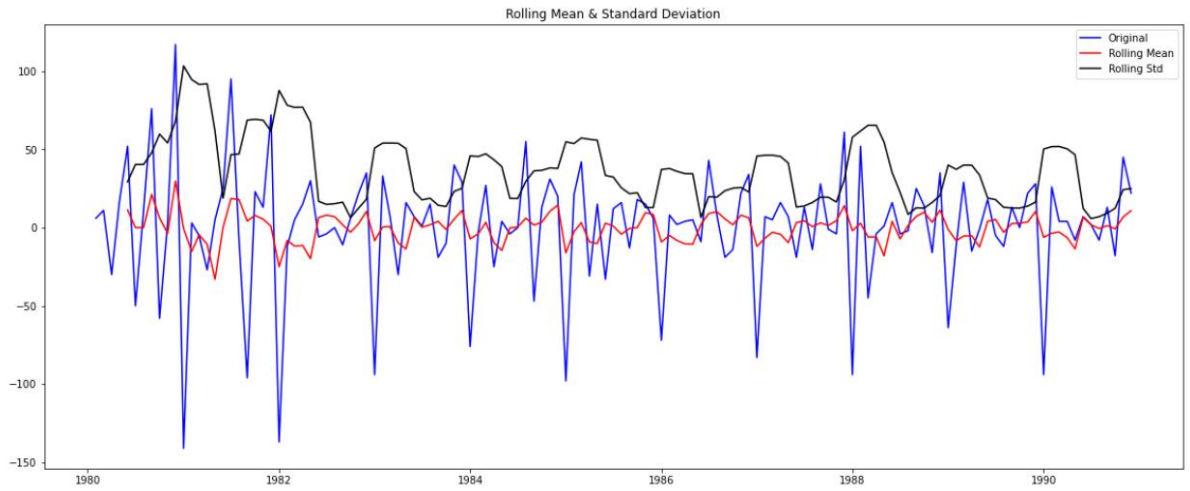
Original Time Series after differentiating passes the test, hence it can be concluded that the time series has now become stationary and now this time series can be used in AIRMA/SARIMA model.



Stationary test results on original time series training data

```
Results of Dickey-Fuller Rose_test:
Rose_test Statistic             -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

Original Time Series Training Data, fails the test, hence it can be concluded that the time series is not stationary, hence requires differentiating with respect to previous values



Stationarity Test Result on Original Time Series Training Data after first order Differentiating:

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

Original Time Series Training Data after differentiating passes the test, hence it can be concluded that the time series has now become stationary and now this time series can be used in AIRMA/SARIMA model.

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***.

ARIMA Model:

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modelled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q

where, p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

Likewise a pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

After fitting the ARIMA Model on the training data, below are the parameters with the lowest AIC scores

	param	AIC
15	(3, 1, 3)	1273.194097
2	(0, 1, 2)	1276.835373
6	(1, 1, 2)	1277.359228
5	(1, 1, 1)	1277.775754
3	(0, 1, 3)	1278.074260
9	(2, 1, 1)	1279.045689
10	(2, 1, 2)	1279.298694
7	(1, 1, 3)	1279.312640
13	(3, 1, 1)	1279.605962
1	(0, 1, 1)	1280.726183
14	(3, 1, 2)	1280.969247
11	(2, 1, 3)	1281.196226
12	(3, 1, 0)	1299.478739
8	(2, 1, 0)	1300.609261
4	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

As we can see 3,1,3 as p, d, q values respectively gives the least RMCE Scores. Lets find the model summary and RMSE scores for test Data.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(3, 1, 3)	Log Likelihood	-628.597
Method:	css-mle	S.D. of innovations	28.356
Date:	Sun, 17 Apr 2022	AIC	1273.194
Time:	16:09:27	BIC	1296.196
Sample:	02-01-1980 - 12-01-1990	HQIC	1282.541

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4906	0.088	-5.548	0.000	-0.664	-0.317
ar.L1.D.Rose	-0.7243	0.086	-8.411	0.000	-0.893	-0.556
ar.L2.D.Rose	-0.7218	0.087	-8.342	0.000	-0.891	-0.552
ar.L3.D.Rose	0.2763	0.085	3.234	0.001	0.109	0.444
ma.L1.D.Rose	-0.0151	0.045	-0.339	0.735	-0.102	0.072
ma.L2.D.Rose	0.0151	0.044	0.340	0.734	-0.072	0.102
ma.L3.D.Rose	-1.0000	0.046	-21.901	0.000	-1.089	-0.911

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.5011	-0.8661j	1.0006	-0.3335
AR.2	-0.5011	+0.8661j	1.0006	0.3335
AR.3	3.6142	-0.0000j	3.6142	-0.0000
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.4925	-0.8703j	1.0000	-0.3320
MA.3	-0.4925	+0.8703j	1.0000	0.3320

	Rose_test RMSE
ARIMA(3,1,3)	15.986441

.

The RMSE for (3,1,3) for rose test is 15.986441.

SARIMA Model:

SARIMA is Seasonal ARIMA, or simply put, ARIMA with a seasonal component.

A typical SARIMA model equation looks like the following – SARIMA(p,d,q)x(P,D,Q)lag

The parameters for these types of models are as follows:

p and seasonal P: indicate the number of AR terms (lags of the stationary series)

d and seasonal D: indicate differencing that must be done to stationary series

q and seasonal Q: indicate the number of MA terms (lags of the forecast errors)

lag: indicates the seasonal length in the data

With the AR plot, we chose two seasonality period i.e. 6 and 12 months to test effect of seasonality ad model with best AIC and RMSE will be chosen for final SARIMA model

After fitting the SARIMA Model with seasonality as 6 on the training data, below are the parameters with the lowest AIC scores.

	param	seasonal	AIC
375	(2, 1, 3)	(2, 1, 3, 6)	889.189817
503	(3, 1, 3)	(2, 1, 3, 6)	891.125985
511	(3, 1, 3)	(3, 1, 3, 6)	893.125640
367	(2, 1, 3)	(1, 1, 3, 6)	894.757072
127	(0, 1, 3)	(3, 1, 3, 6)	894.905687

As we can see (2,1,3) (2,1,3,6) as p, d, q and seasonal PDQS values respectively gives the least RMCE Scores. Lets find the model summary and RMSE scores for test Data.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 1, 3, 6)	Log Likelihood	-433.595
Date:	Sun, 17 Apr 2022	AIC	889.190
Time:	19:18:15	BIC	918.172
Sample:	0	HQIC	900.929
	- 132		
Covariance Type:	opg		

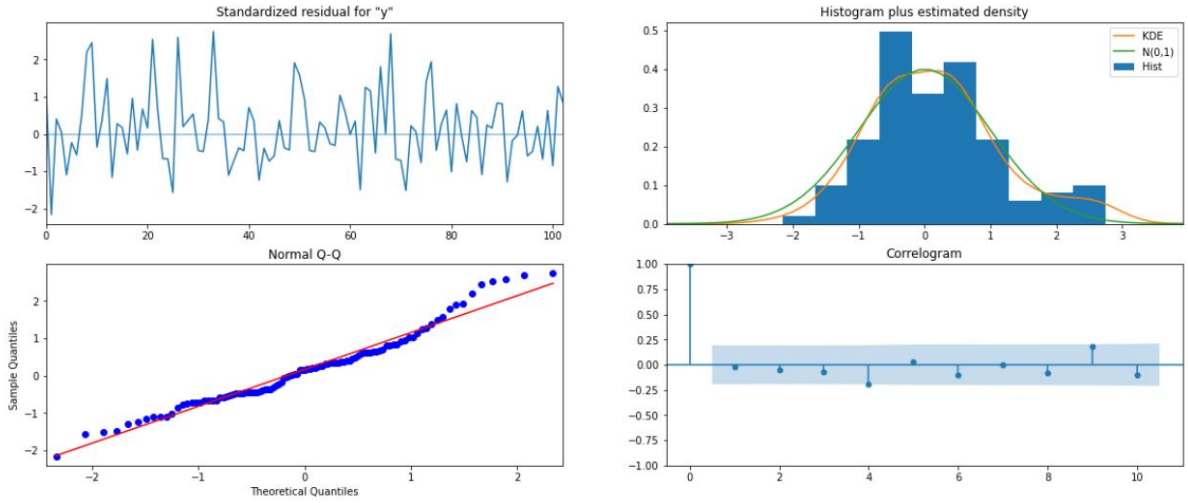
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5746	0.023	25.067	0.000	0.530	0.620
ar.L2	-0.9162	0.021	-43.468	0.000	-0.958	-0.875
ma.L1	-1.4571	31.539	-0.046	0.963	-63.273	60.359
ma.L2	1.5182	128.954	0.012	0.991	-251.228	254.264
ma.L3	-0.8409	86.128	-0.010	0.992	-169.648	167.966
ar.S.L6	-0.4347	0.106	-4.087	0.000	-0.643	-0.226
ar.S.L12	0.4834	0.102	4.734	0.000	0.283	0.684
ma.S.L6	-1.6588	3.472	-0.478	0.633	-8.464	5.147
ma.S.L12	-1.0800	9.161	-0.118	0.906	-19.036	16.876
ma.S.L18	1.5946	5.609	0.284	0.776	-9.398	12.588
sigma2	68.3643	7026.947	0.010	0.992	-1.37e+04	1.38e+04

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	5.80
Prob(Q):	0.86	Prob(JB):	0.05
Heteroskedasticity (H):	0.45	Skew:	0.56
Prob(H) (two-sided):	0.02	Kurtosis:	3.29

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

lets find the RMSE Values and nature of residuals.

The result of RMSE value is 16.74614792145421



predict on the rose test set using this model and evaluate the model

	y	mean	mean_se	mean_ci_lower	mean_ci_upper
0		54.884819	15.046420	25.394377	84.375260
1		63.978125	15.158671	34.267676	93.688575
2		71.844534	15.456135	41.551066	102.138002
3		69.056538	15.808567	38.072317	100.040759
4		77.042959	15.994144	45.695012	108.390905

Comparing the results of all RMSE value is

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033
Alpha =0.1578,Beta=0.1578,DoubleExponentialSmoothing	15.707052
Alpha=0.05,Beta=0.35,DoubleExponentialSmoothing	16.329097
Alpha= 0.133,Beta=0.0138,Gamma=0.005,TripleExponentialSmoothing	14.257122
Alpha=0.05,Beta=0.90,Gamma=0.4,TripleExponentialSmoothing	11.652209
ARIMA(3,1,3)	15.986441
SARIMA(2, 1, 3),(2, 1, 3, 6)	16.746148

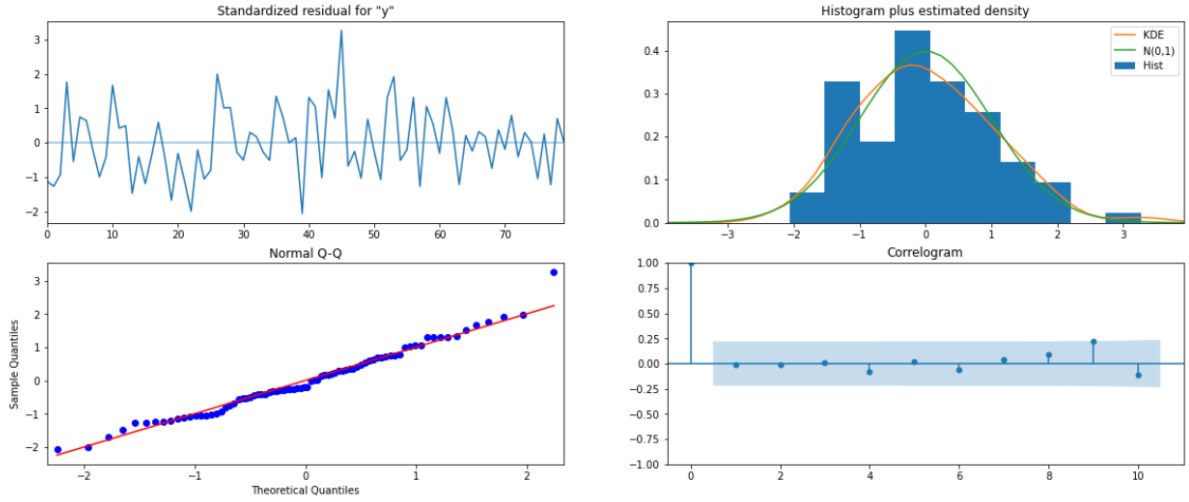
SARIMA 12:

After fitting the SARIMA Model with seasonality as 6 on the training data, below are the parameters with the lowest AIC scores

	param	seasonal	AIC
215	(1, 1, 2)	(2, 1, 3, 12)	18.000000
343	(2, 1, 2)	(2, 1, 3, 12)	20.000000
255	(1, 1, 3)	(3, 1, 3, 12)	22.000000
191	(1, 1, 1)	(3, 1, 3, 12)	83.687767
351	(2, 1, 2)	(3, 1, 3, 12)	106.383825

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 1, 1, 12)   Log Likelihood                -331.681
Date:                            Sun, 17 Apr 2022   AIC                            681.363
Time:                                    19:47:05   BIC                            702.801
Sample:                                         0   HQIC                           689.958
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0173      0.151      0.114      0.909      -0.279       0.314
ar.L2         -0.0426      0.141     -0.302      0.762      -0.319       0.234
ar.L3         -0.0575      0.119     -0.485      0.628      -0.290       0.175
ma.L1         -0.9388      0.085    -11.107      0.000      -1.104      -0.773
ar.S.L12       0.0907      0.126      0.720      0.471      -0.156       0.337
ar.S.L24      -0.0436      0.108     -0.406      0.685      -0.254       0.167
ar.S.L36   -3.594e-05      0.053     -0.001      0.999      -0.103       0.103
ma.S.L12      -0.9997    183.450     -0.005      0.996    -360.555     358.555
sigma2       185.4302    3.4e+04      0.005      0.996   -6.65e+04    6.68e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 2.56
Prob(Q):                              0.91   Prob(JB):                         0.28
Heteroskedasticity (H):               0.56   Skew:                             0.42
Prob(H) (two-sided):                  0.13   Kurtosis:                         3.22
===================================================================================

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



y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	45.230490	14.458298	16.892746	73.568234
1	63.053692	14.503247	34.627850	91.479533
2	68.119687	14.453297	39.791745	96.447628
3	61.826875	14.449534	33.506309	90.147441
4	68.437081	14.470038	40.076328	96.797835

predict on the rose test set and evaluate the model

The result for RMSE value for

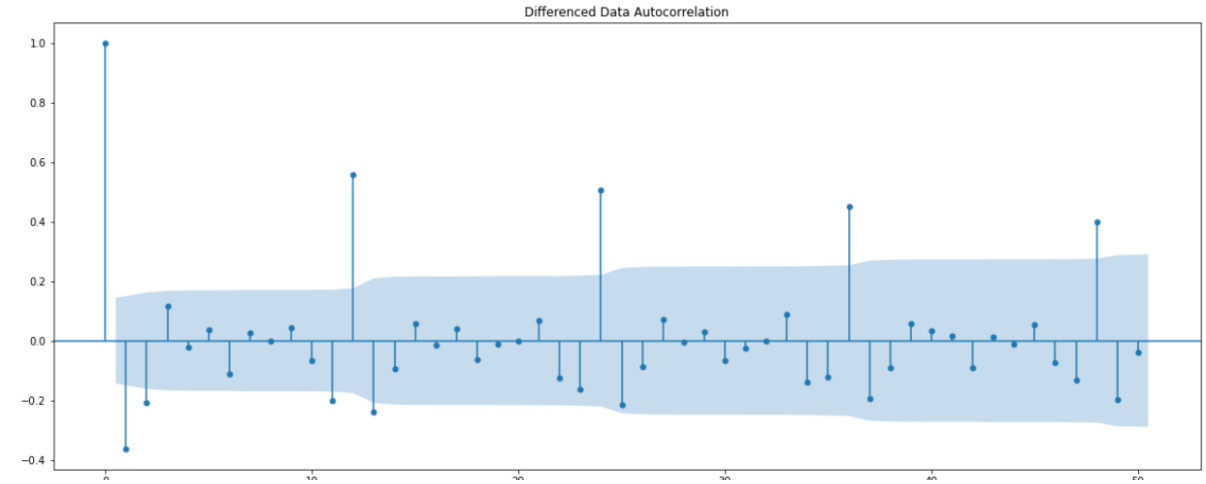
```
16.824029096257128
```

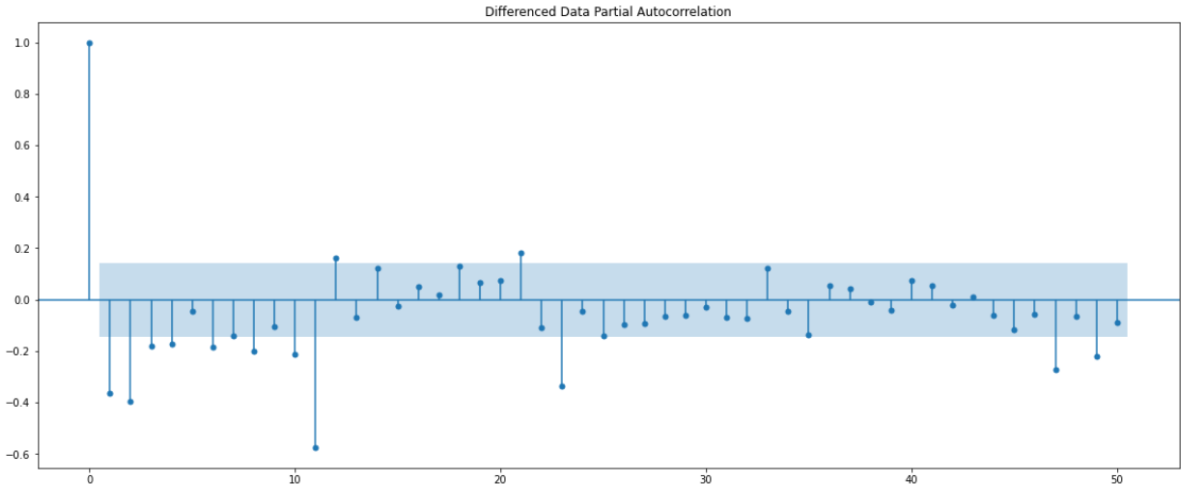
Now comparing all the RMSE values

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033
Alpha =0.1578,Beta=0.1578,DoubleExponentialSmoothing	15.707052
Alpha=0.05,Beta=0.35,DoubleExponentialSmoothing	16.329097
Alpha= 0.133,Beta=0.0138,Gamma=0.005,TripleExponentialSmoothing	14.257122
Alpha=0.05,Beta=0.90,Gamma=0.4,TripleExponentialSmoothing	11.652209
ARIMA(3,1,3)	15.986441
SARIMA(2, 1, 3),(2, 1, 3, 6)	16.746148
SARIMA(3,1,1)(3,1,1,12)	16.824029

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.

For the ARIMA model





The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 4

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

Hence building an ARIMA model basis these parameters, following model summary is obtained

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(4, 1, 2)	Log Likelihood	-633.876
Method:	css-mle	S.D. of innovations	29.793
Date:	Sun, 17 Apr 2022	AIC	1283.753
Time:	19:47:08	BIC	1306.754
Sample:	02-01-1980 - 12-01-1990	HQIC	1293.099

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1905	0.576	-0.331	0.741	-1.319	0.938
ar.L1.D.Rose	1.1685	0.087	13.391	0.000	0.997	1.340
ar.L2.D.Rose	-0.3562	0.132	-2.692	0.007	-0.616	-0.097
ar.L3.D.Rose	0.1855	0.132	1.402	0.161	-0.074	0.445
ar.L4.D.Rose	-0.2227	0.091	-2.443	0.015	-0.401	-0.044
ma.L1.D.Rose	-1.9506	nan	nan	nan	nan	nan
ma.L2.D.Rose	1.0000	nan	nan	nan	nan	nan

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1027	-0.4115j	1.1770	-0.0569
AR.2	1.1027	+0.4115j	1.1770	0.0569
AR.3	-0.6863	-1.6644j	1.8003	-0.3122
AR.4	-0.6863	+1.6644j	1.8003	0.3122
MA.1	0.9753	-0.2209j	1.0000	-0.0355
MA.2	0.9753	+0.2209j	1.0000	0.0355

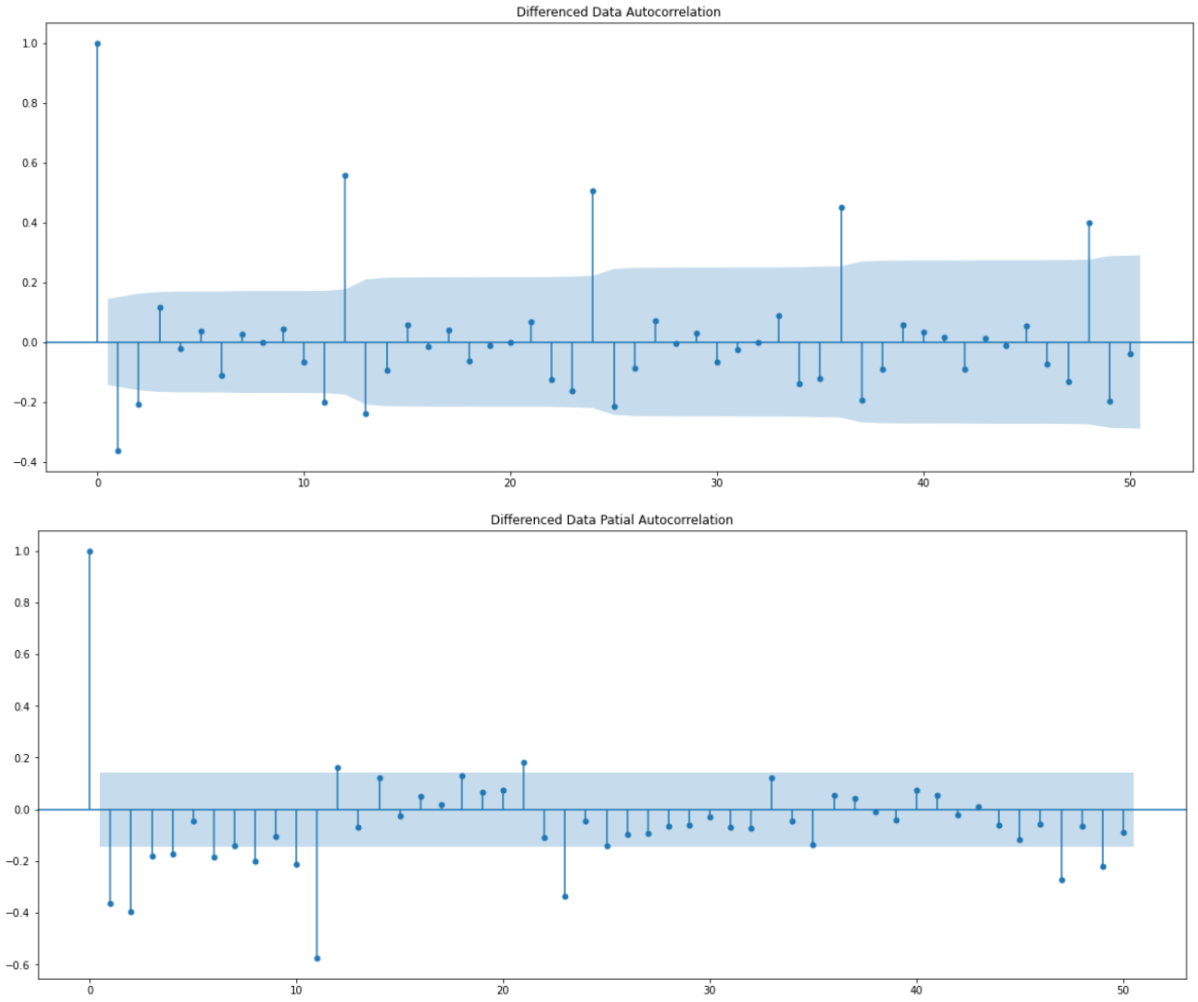
We can see that both the AR(p) and the MA(q) model are of order 3 and 2 respectively

Predict on the Rose_test Set using this model and evaluate the model.

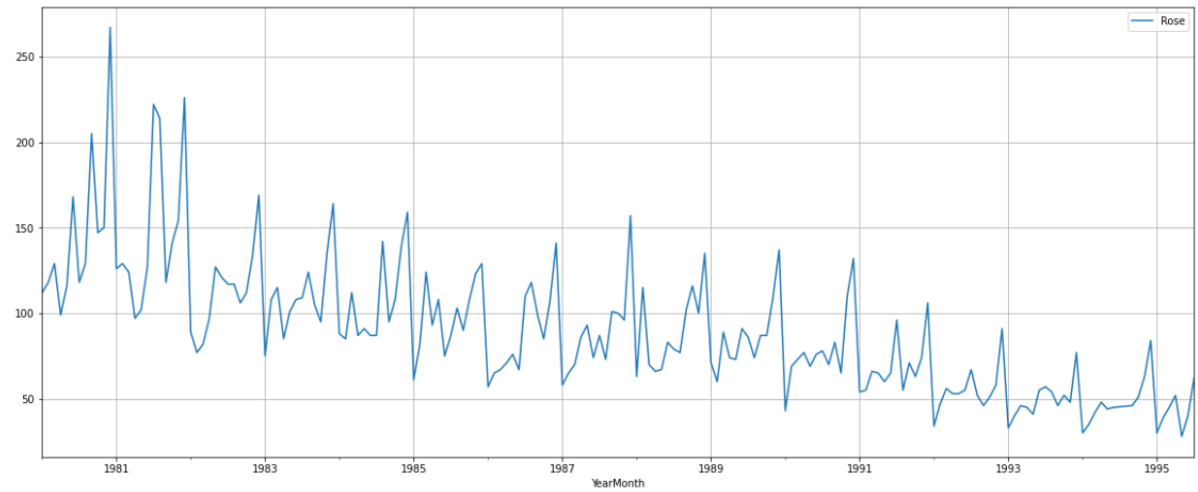
The RMSE value is

```
33.950456932235916
```

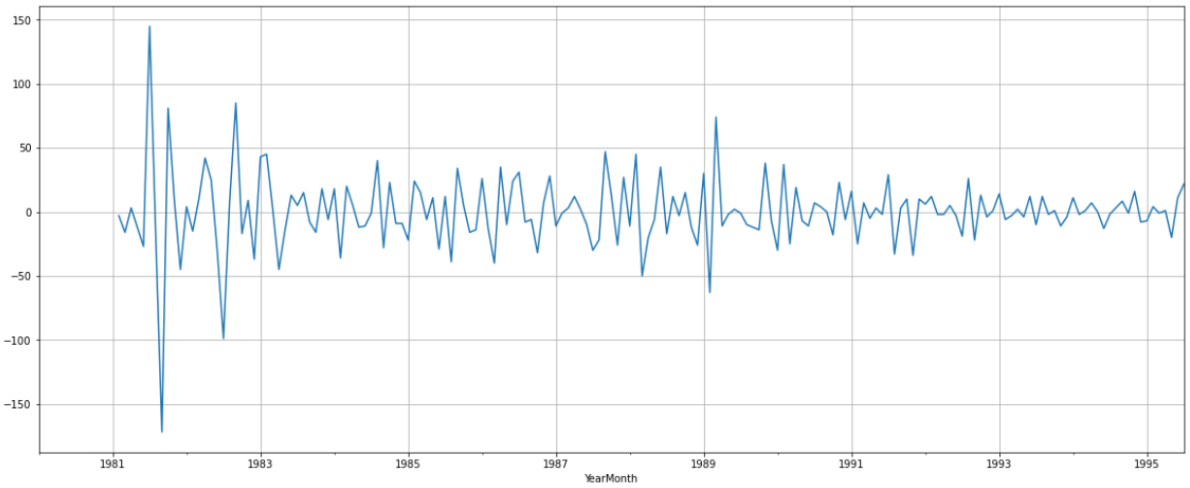
SARIMA – 12



After Seasonal differentiating with order of 12 and then differentiating once more, following plot was obtained



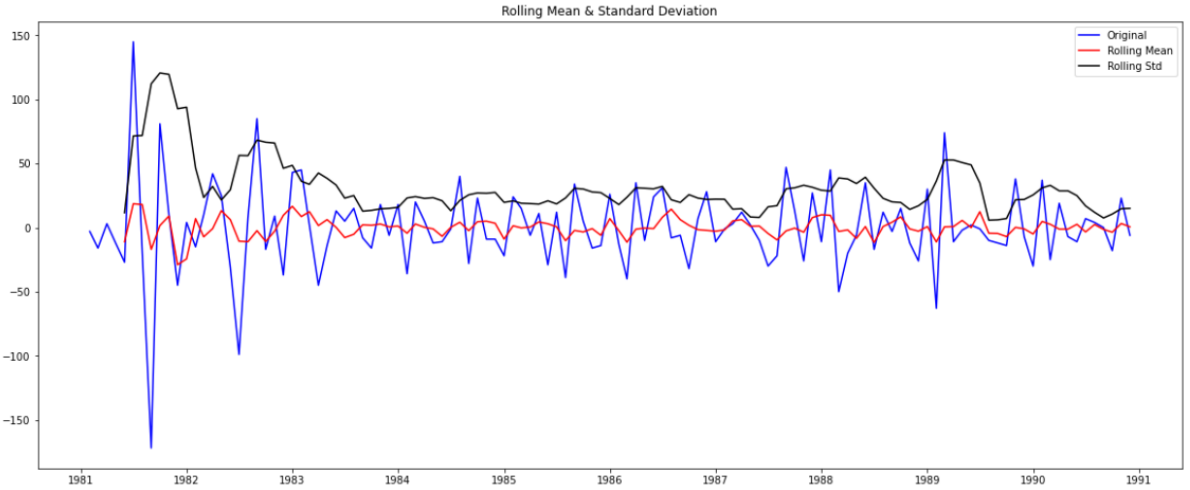
Normal plot before differentiating



After Differentiating

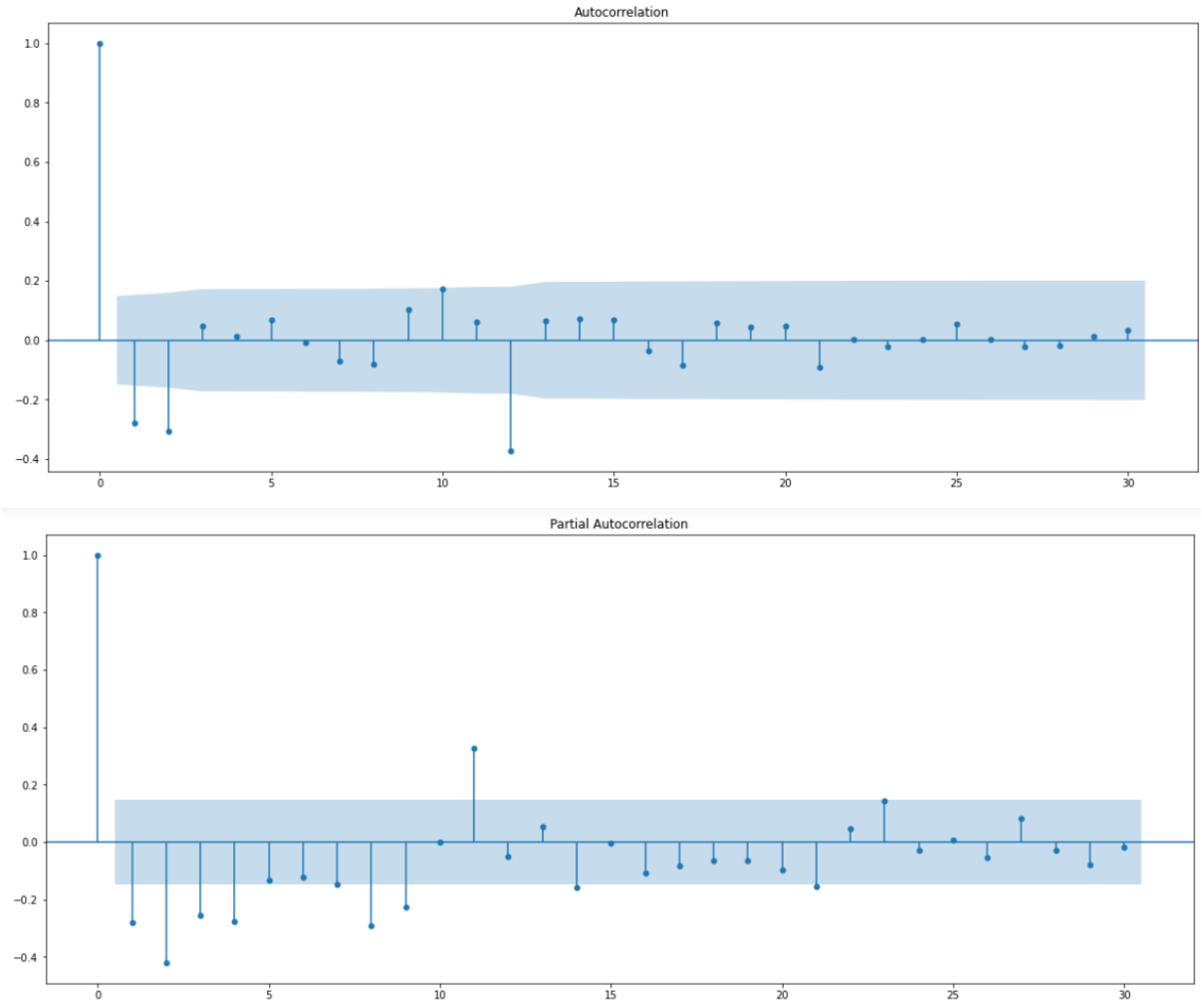
Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

Let us check the stationarity of the current time series,




```
Results of Dickey-Fuller Rose_test:
Rose_test Statistic            -3.692348
p-value                         0.004222
#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
dtype: float64
```

It can be observed, the current series passes the Stationarity Test. Following is the ACF and PACF plot obtained.



We are going to take the seasonal period as 12.

We will keep the p(4) and q(2) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 4.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 2.

Checking the model Summary

SARIMAX Results

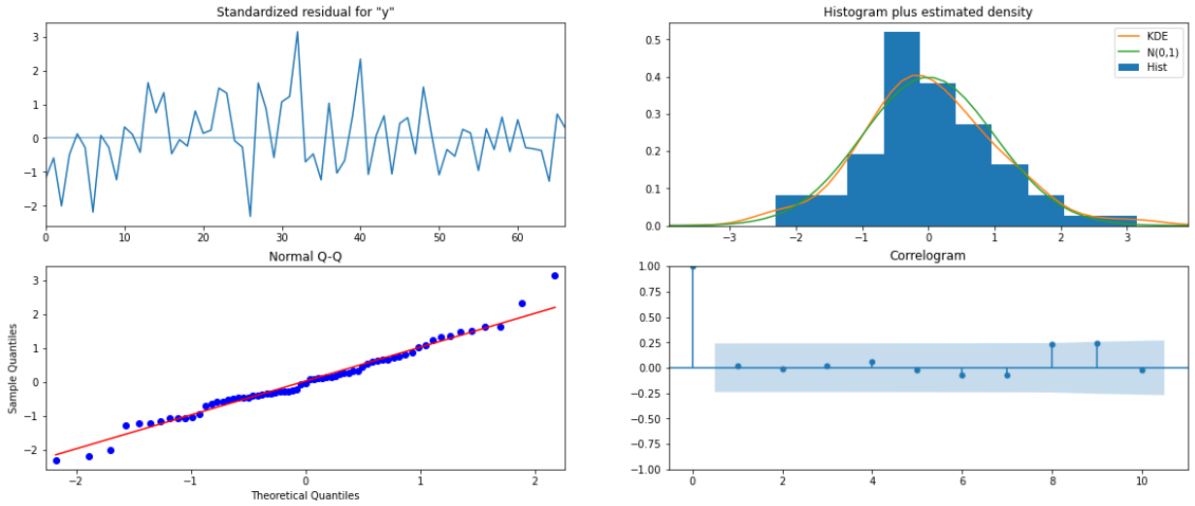
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(4, 1, 2, 12)	Log Likelihood	-277.661
Date:	Sun, 17 Apr 2022	AIC	581.322
Time:	19:47:32	BIC	609.983
Sample:	0	HQIC	592.663
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9743	0.189	-5.161	0.000	-1.344	-0.604
ar.L2	-0.1123	0.279	-0.402	0.688	-0.660	0.435
ar.L3	-0.1044	0.270	-0.387	0.698	-0.633	0.424
ar.L4	-0.1285	0.151	-0.849	0.396	-0.425	0.168
ma.L1	0.1605	0.211	0.761	0.446	-0.253	0.574
ma.L2	-0.8395	0.234	-3.584	0.000	-1.299	-0.380
ar.S.L12	-0.1443	0.364	-0.396	0.692	-0.858	0.569
ar.S.L24	-0.3597	0.213	-1.692	0.091	-0.776	0.057
ar.S.L36	-0.2153	0.102	-2.102	0.036	-0.416	-0.015
ar.S.L48	-0.1195	0.090	-1.333	0.182	-0.295	0.056
ma.S.L12	-0.5157	0.343	-1.505	0.132	-1.188	0.156
ma.S.L24	0.2085	0.372	0.561	0.575	-0.520	0.937
sigma2	215.3729	0.002	1.34e+05	0.000	215.370	215.376

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	2.41
Prob(Q):	0.86	Prob(JB):	0.30
Heteroskedasticity (H):	0.49	Skew:	0.32
Prob(H) (two-sided):	0.10	Kurtosis:	3.68

Covariance matrix calculated using the outer product of gradients (co mplex-step).

Covariance matrix is singular or near-singular, with condition number 5.84e+20. Standard errors may be unstable.



y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	46.386385	14.771252	17.435263	75.337507
1	62.934275	14.990255	33.553914	92.314635
2	63.527464	14.999940	34.128123	92.926806
3	66.472883	15.180132	36.720371	96.225395
4	63.540686	15.181010	33.786454	93.294918

The result of RMSE value is

```
17.52937513564139
```

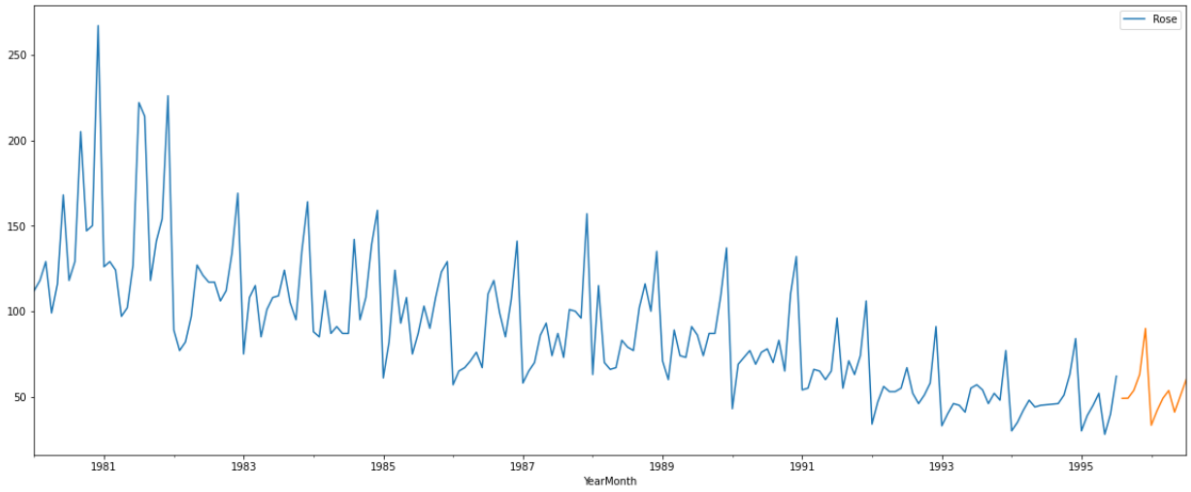
8 - *Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.*

	Rose_test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2point_MovingAverage_Rose	68.970159
4point_MovingAverage_Rose	46.403626
6point_MovingAverage_Rose	39.126446
9point_MovingAverage_Rose	34.410938
Alpha=0.0987,SimpleExponentialSmoothing	36.796227
Alpha=0.10,SimpleExponentialSmoothing	36.828033
Alpha =0.1578,Beta=0.1578,DoubleExponentialSmoothing	15.707052
Alpha=0.05,Beta=0.35,DoubleExponentialSmoothing	16.329097
Alpha= 0.133,Beta=0.0138,Gamma=0.005,TripleExponentialSmoothing	14.257122
Alpha=0.05,Beta=0.90,Gamma=0.4,TripleExponentialSmoothing	11.652209
ARIMA(3,1,3)	15.986441
SARIMA(2, 1, 3),(2, 1, 3, 6)	16.746148
SARIMA(3,1,1)(3,1,1,12)	16.824029
ARIMA(4,1,2)	33.950457
SARIMA(4, 1, 2)(4,1,2,12)	17.529375

Alpha=0.05,Beta=0.90,Gamma=0.40,Triple Exponential Smoothing

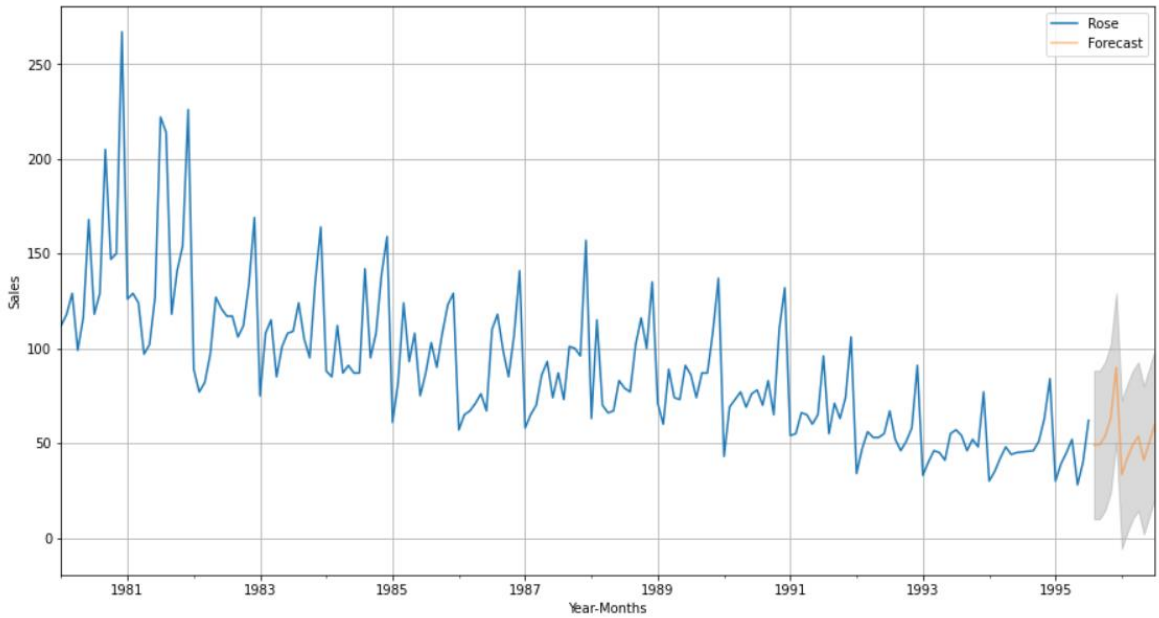
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

After building the model basis the best parameter and method obtained for the Training and Test and applying it on whole, following is the predicted value plot obtained for next 12 months on Time Series



With confidence interval

	lower_CI	prediction	upper_ci
1995-08-01	9.848444	49.039517	88.230589
1995-09-01	9.935943	49.127016	88.318089
1995-10-01	14.519307	53.710380	92.901452
1995-11-01	23.825898	63.016970	102.208043
1995-12-01	50.717596	89.908668	129.099741



The RMSE value for the whole model is

```
Rose RMSE: 19.971968883744363
```

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

Insights:

- *The yearly boxplots also shows that the Rose Wine Sales have decreased with the passing years.*
- *There is a clear distinction of Rose Wine sales within different months spread across various years.*
- *The highest such numbers are being recorded towards the end of the years with a huge spike in the month of December possibly due to festive seasons*.
- *Almost 95% of sales are within 150 units of Rose Wines*.
- *In this dataset, after comparing various models and comparing on RMSE, we analyse that the Triple Exponential Smoothing proved to be a better model with parameters alpha = 0.05, Beta = 0.90, Gamma = 0.4*

Recommendations:

- *Introduce various promotional offers across the outlets and around the year to attract customers basis price.*
- *Organize testing outlets for customers to taste the ROSE wine as few may not be aware of the taste and quality of the same*.
- *Conducting training of outlets to promote these wines to customers personally based on their requirements*