TIME SERIES FORECASTING

ROSE WINE SALES - PROJECT REPORT

BY,

RAGAVEDHNI K R

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

- The data is read from the 'Rose.csv' file, the initial set of rows is as below.

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

- Creating a DateTimeIndex using date_range() from Pandas for the entire length of the dataset.

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

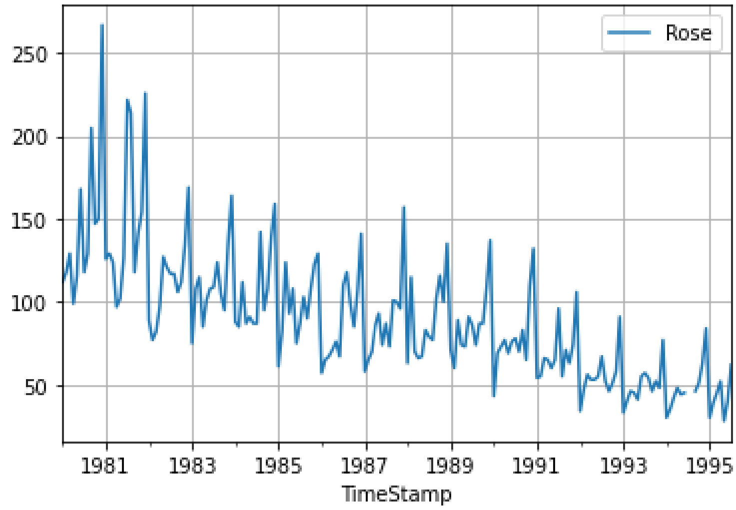
- Including that as a TimeStamp column to the dataframe, as it is required to work on Time Series data.

	YearMonth	Rose	TimeStamp
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

- Dropping the unwanted column and set the index of the dataframe as TimeStamp using the set_index(), to work with Time series data.

	Rose
TimeStamp	
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

- We can plot the dataframe as a time series data as below.



- The X-axis is the TimeStamp column and the Y-axis is the count of Rose wine sales across the years.

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

- The basic exploratory data analysis is made on the data using describe().

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

- The mean and median of the data can be seen as below.

```
Mean of the data Rose    89.909091
dtype: float64
Median of the data Rose    85.0
dtype: float64
```

- Checking the null values in the data.

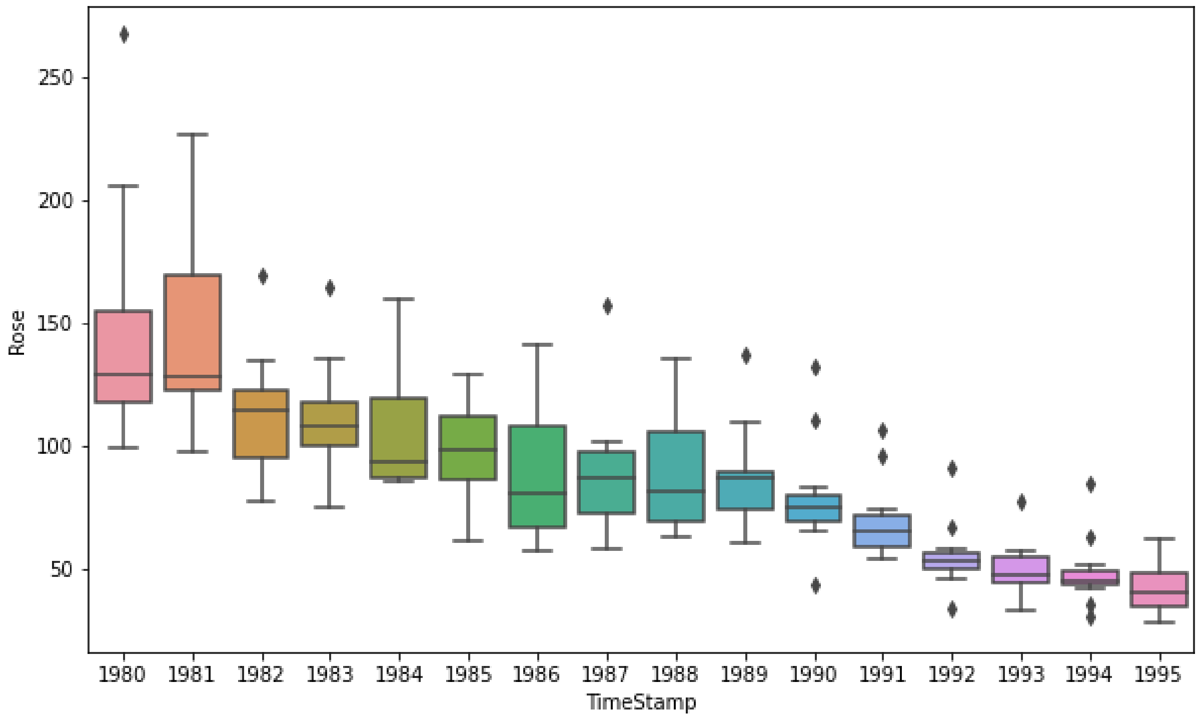
TimeStamp	Rose
1994-07-31	NaN
1994-08-31	NaN

- The null values are imputed using the interpolate method and checking the null values again.

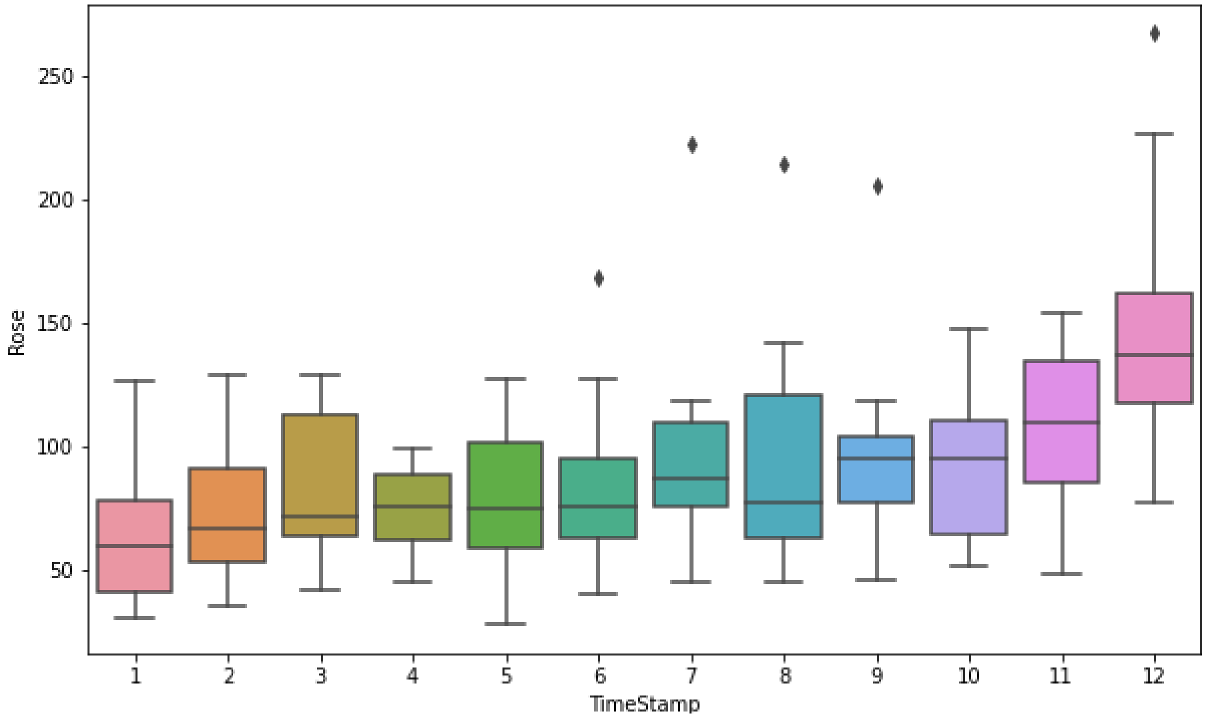
TimeStamp	Rose
1994-01-31	30.000000
1994-02-28	35.000000
1994-03-31	42.000000
1994-04-30	48.000000
1994-05-31	44.000000
1994-06-30	45.000000
1994-07-31	45.333333
1994-08-31	45.666667
1994-09-30	46.000000
1994-10-31	51.000000
1994-11-30	63.000000
1994-12-31	84.000000

```
checking Null values after imputing Rose    0
dtype: int64
```

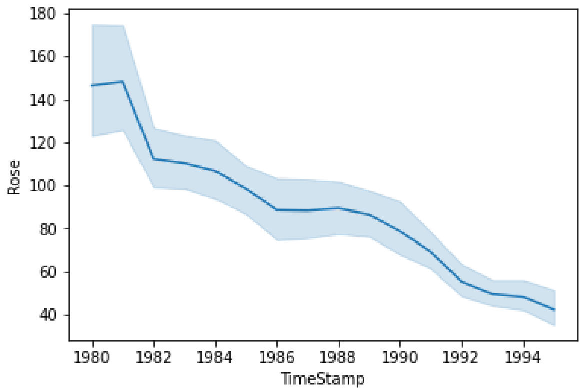
- Plotting year on year boxplot for the Rose wine sales. The sale is getting down gradually across the years.

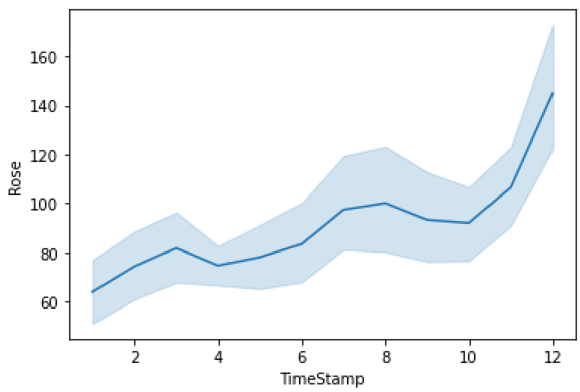


- Plotting the monthly boxplot for the Rose wine sales taking all the years into account. The sales are nearly the same except in the month of December, which is the festival season of the year.

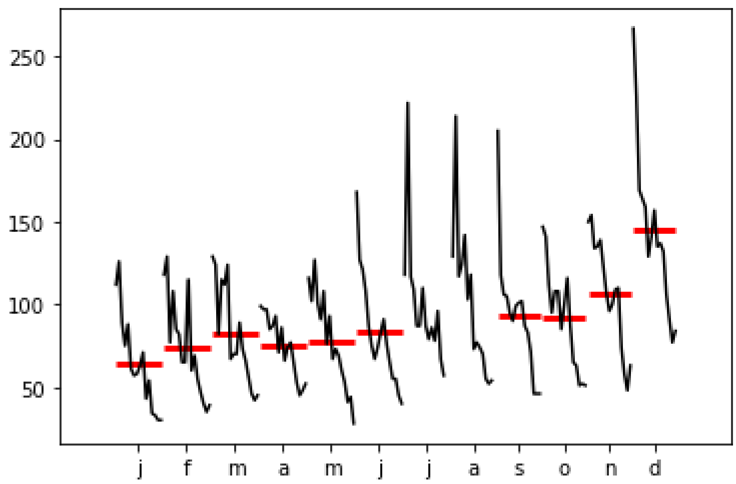


- Plotting yearly line plot across all the years and monthly line plot for all years.

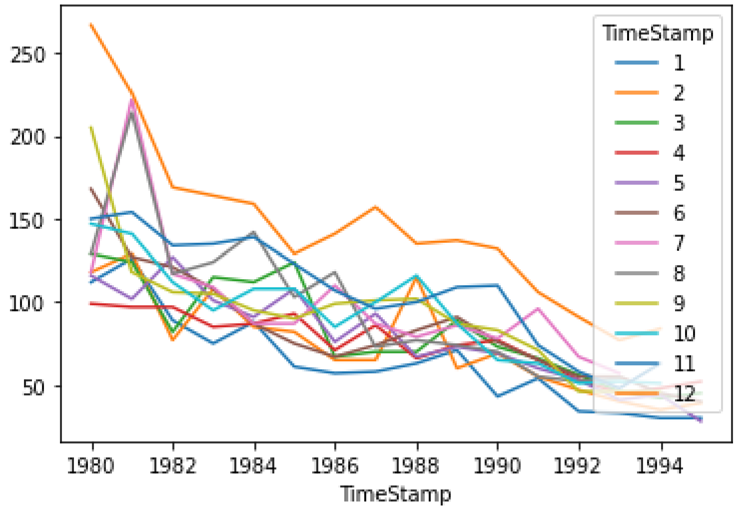




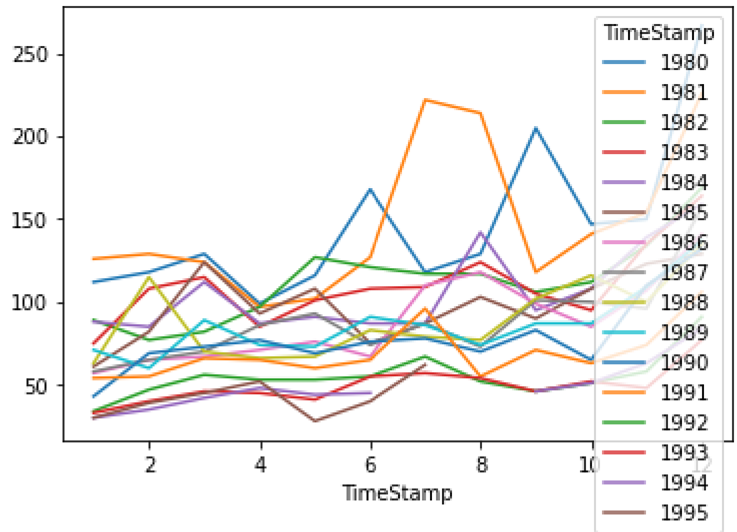
- Plotting the monthplot from statsmodels.graphics.tsaplots package.



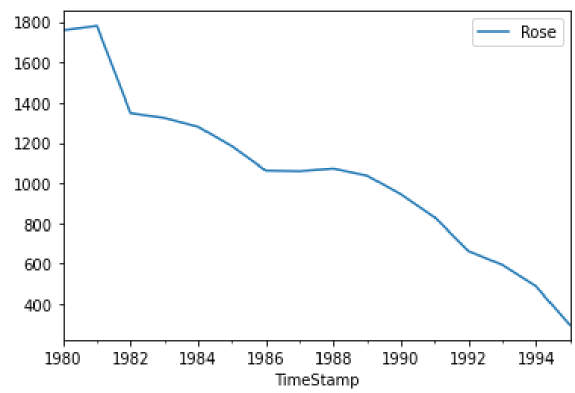
- Plotting the Time Series according to different months for across the years.

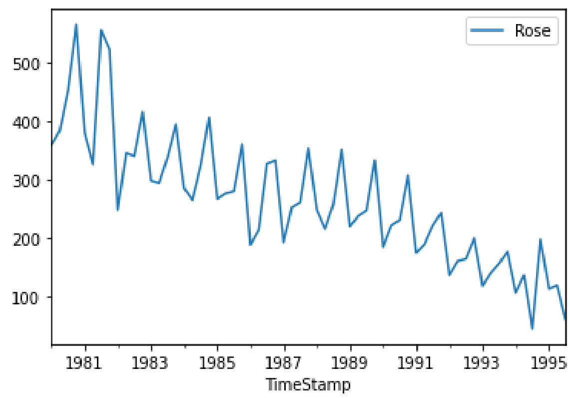


- Plotting the Time series according to across all the months for different years.



- Reading and plotting the data yearly and quarterly using the resample() and mean().



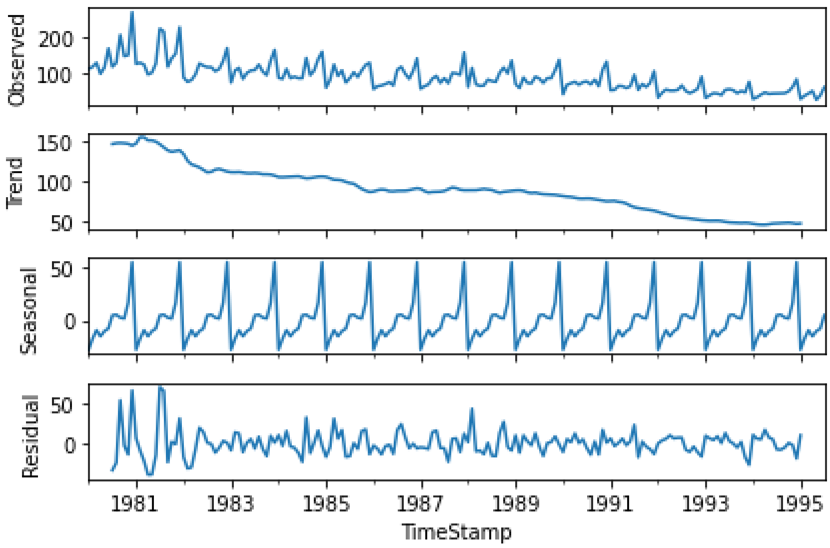


DECOMPOSITION:

- Decomposing the data using seasonal_decompose() from the statsmodels.tsa.seasonal module. Checking the model as additive or multiplicative.

ADDITIVE MODEL:

- If the seasonality and residual components are independent of the trend, then it is an additive series.



- Checking on the first 10 rows in trends, seasonal, residual components separately.

TREND

TimeStamp	Rose
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	147.083333
1980-08-31	148.125000
1980-09-30	148.375000
1980-10-31	148.083333

SEASONAL

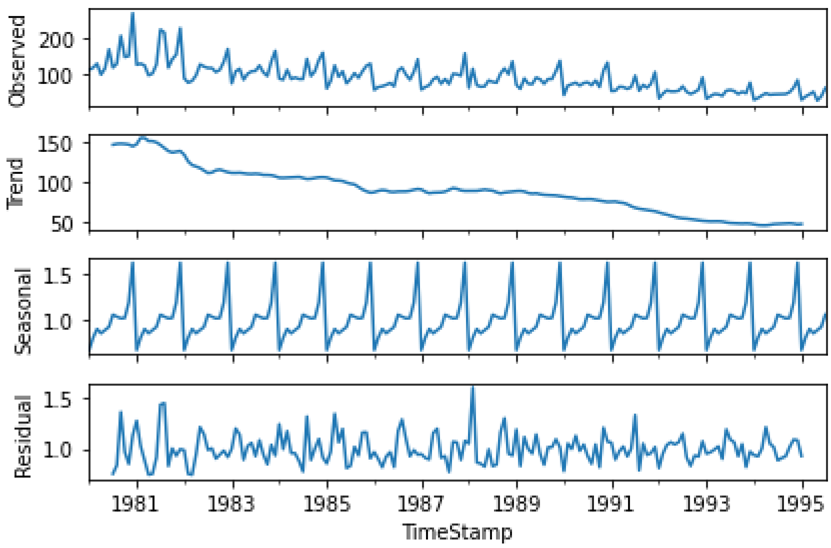
TimeStamp	Rose
1980-01-31	-27.908647
1980-02-29	-17.435632
1980-03-31	-9.285830
1980-04-30	-15.098330
1980-05-31	-10.196544
1980-06-30	-7.678687
1980-07-31	4.896908
1980-08-31	5.499686
1980-09-30	2.774686
1980-10-31	1.871908

RESIDUAL

TimeStamp	Rose
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	-33.980241
1980-08-31	-24.624686
1980-09-30	53.850314
1980-10-31	-2.955241

MULTIPLICATIVE MODEL:

- If the seasonality and residual components are in dependent, i.e., they fluctuate on trend, then it is a multiplicative series.



- Checking on the first 10 rows in trends, seasonal, residual components separately.

TREND

TimeStamp	Rose
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	147.083333
1980-08-31	148.125000
1980-09-30	148.375000
1980-10-31	148.083333

SEASONAL

TimeStamp	Rose
1980-01-31	0.670111
1980-02-29	0.806163
1980-03-31	0.901164
1980-04-30	0.854024
1980-05-31	0.889415
1980-06-30	0.923985
1980-07-31	1.058038
1980-08-31	1.035881
1980-09-30	1.017648
1980-10-31	1.022573

RESIDUAL

TimeStamp	Rose
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	0.758258
1980-08-31	0.840720
1980-09-30	1.357674
1980-10-31	0.970771

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

- The data is split into train data for training the model and test data for predicting the data using the model.
- Taking 70% of the data as train data (till the year 1990) and 30% of the data as test data (from the year 1991).
- Checking the shape of the train data and test data.

```
Shape of the train data (132, 1)
Shape of the test data (55, 1)
```

- Viewing the initial sets and end sets of rows from train and test data.

First few rows of Training Data

TimeStamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Last few rows of Training Data

TimeStamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

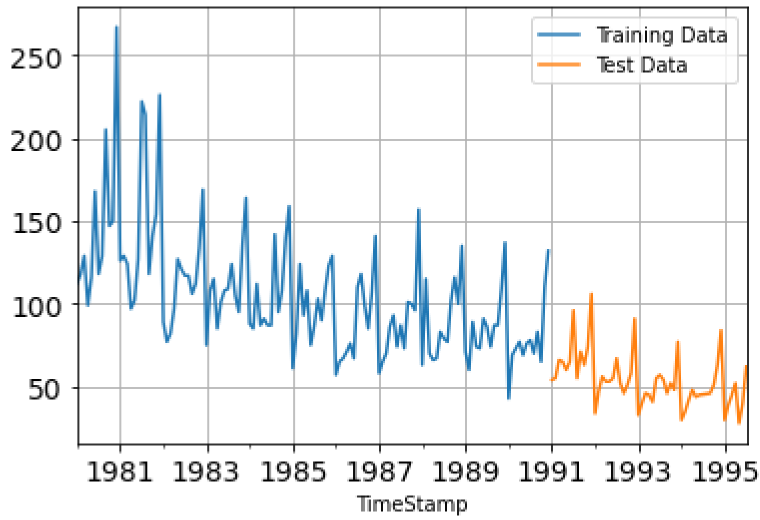
First few rows of Test Data

TimeStamp	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Last few rows of Test Data

TimeStamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

- Checking the plots of the train and test data.



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

- Building the Simple Exponential Smoothing model (SES), Double Exponential Smoothing model (DES), Triple Exponential Smoothing model (TES), Linear Regression model, Naive model, Simple Average model and Moving Average model.

SIMPLE EXPONENTIAL SMOOTHING (SES):

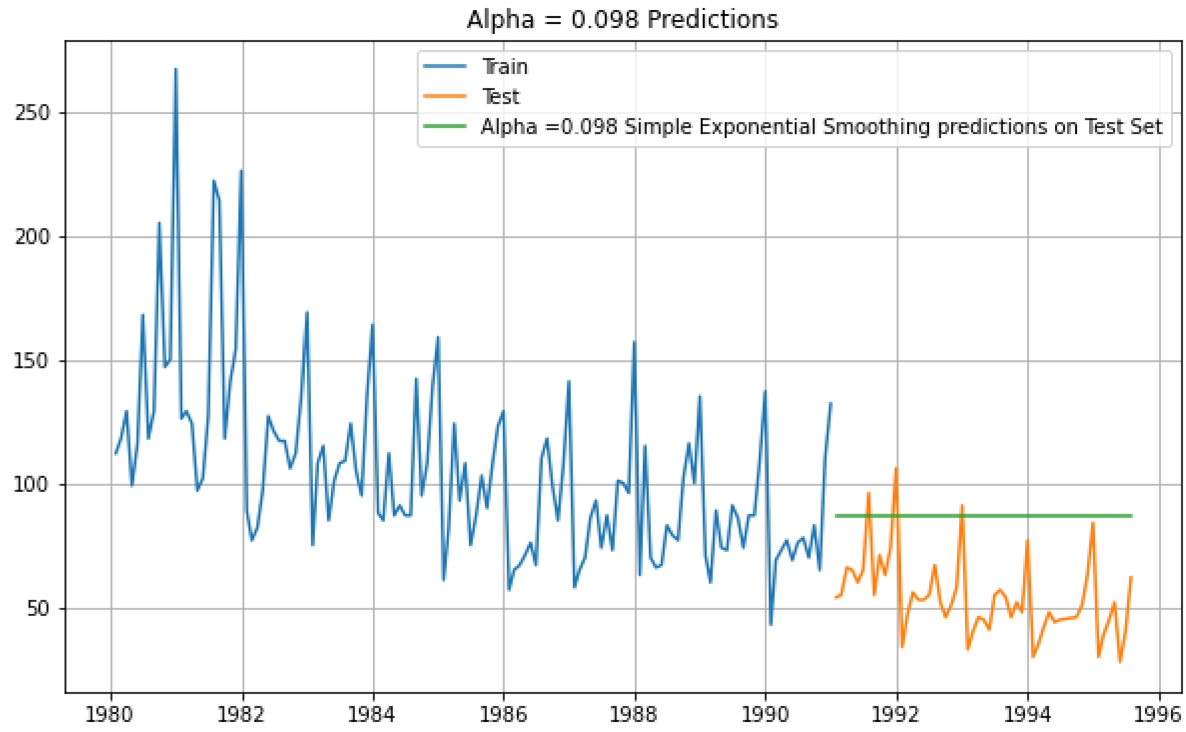
- Apply SES model on the Time Series data when there is no trend or seasonality is present in the data. Though case is almost non-available, we try this to understand how the smoothing parameter (α) controls the performance of the method.
- We create the SES model using the train data and fitting the model to maximize the log-likelihood.
- We check the optimal parameters returned by the model.

```
{'smoothing_level': 0.09874995867958046,
 'smoothing_slope': nan,
 'smoothing_seasonal': nan,
 'damping_slope': nan,
 'initial_level': 134.38699135899094,
 'initial_slope': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- Using the fitted model on the training data, we forecast on the test data.
- We set the parameter as the number of out of sample forecasts from the end of the sample (test data).

1991-01-31	87.105001
1991-02-28	87.105001
1991-03-31	87.105001
1991-04-30	87.105001
1991-05-31	87.105001
1991-06-30	87.105001
1991-07-31	87.105001
1991-08-31	87.105001
1991-09-30	87.105001
1991-10-31	87.105001
1991-11-30	87.105001
1991-12-31	87.105001

- We can check the forecasted values along with train and test values in the below plot.



- We check the RMSE of the SES model on the test data.

```
SES RMSE: 36.79624359473444
```

DOUBLE EXPONENTIAL SMOOTHING (DES):

- DES is an extension of SES, which is applicable when the data has trend but no seasonality.
- The Level and Trend component are controlled by α and β smoothing parameter respectively.
- DES model can be initialised by the train data and setting the exponential parameter accordingly and DES is fitted.
- We check the smoothing parameters for DES as below.

When Exponential= False,

```
{'smoothing_level': 0.15789473684210525,
 'smoothing_slope': 0.15789473684210525,
 'smoothing_seasonal': nan,
 'damping_slope': nan,
 'initial_level': 112.0,
 'initial_slope': 6.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

When Exponential= True,

```
{'smoothing_level': 0.01920533831449077,
 'smoothing_slope': 0.012083532090751897,
 'smoothing_seasonal': nan,
 'damping_slope': nan,
 'initial_level': 150.14127529701187,
 'initial_slope': 0.9936717876881109,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- We forecast using DES model for the duration of the test data.

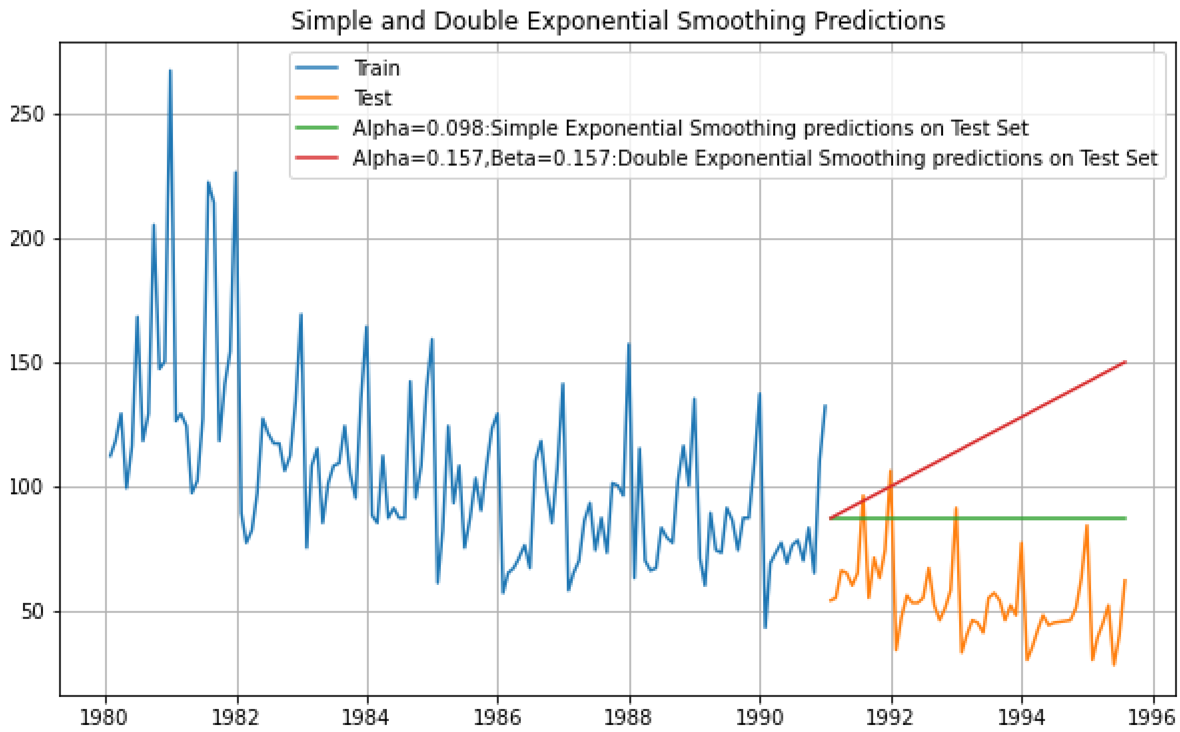
When Exponential= False,

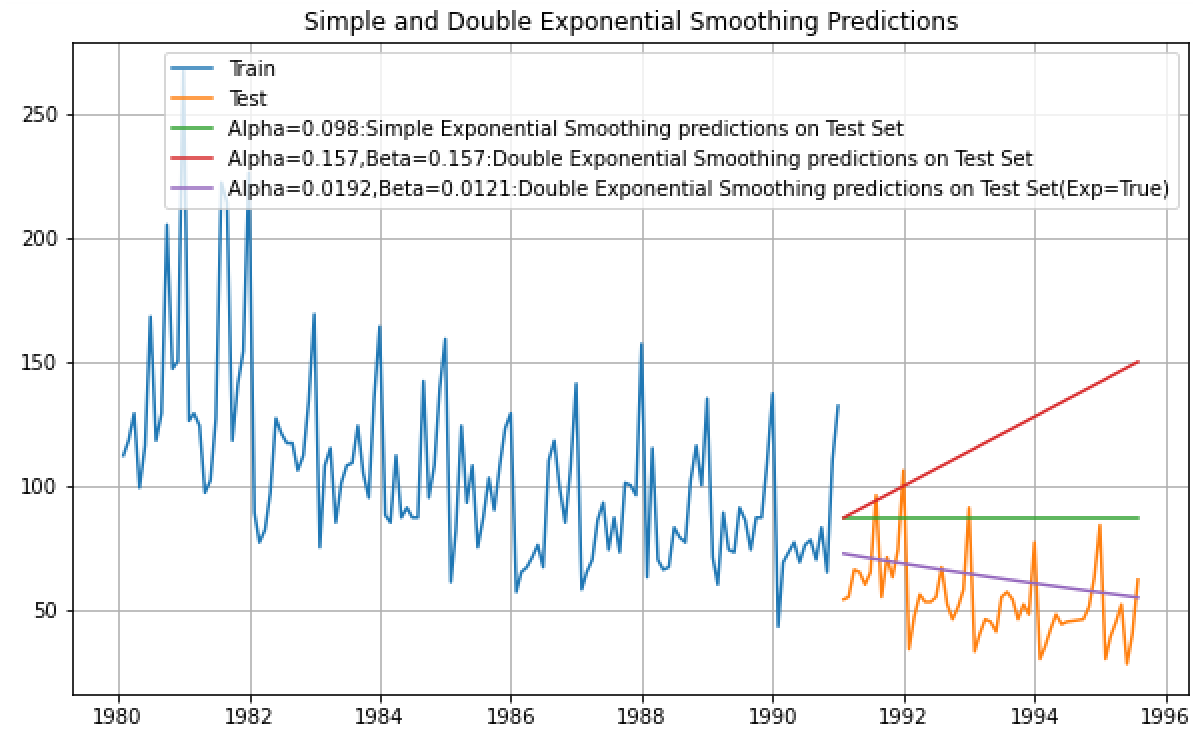
1991-01-31	86.863579
1991-02-28	88.028056
1991-03-31	89.192534
1991-04-30	90.357011
1991-05-31	91.521488
1991-06-30	92.685966
1991-07-31	93.850443
1991-08-31	95.014921
1991-09-30	96.179398
1991-10-31	97.343876
1991-11-30	98.508353
1991-12-31	99.672831

When Exponential= True,

1991-01-31	72.461546
1991-02-28	72.088847
1991-03-31	71.718065
1991-04-30	71.349190
1991-05-31	70.982213
1991-06-30	70.617123
1991-07-31	70.253911
1991-08-31	69.892566
1991-09-30	69.533081
1991-10-31	69.175444
1991-11-30	68.819647
1991-12-31	68.465680

- We can view the SES and DES forecast values along with the train data and test data from the plot below.





- We can see that the DES performs better than SES as the trend is considered for forecasting.
- We check the RMSE of the test data using both the DES models as below.
- **When Exponential= False,**

```
DES RMSE: 70.57245196981661
```

- **When Exponential= True,**

```
DES RMSE: 17.223483135193124
```

- We can see that the RMSE is much lower when the Exponential parameter is set as True in DES model.

TRIPLE EXPONENTIAL MODEL (TES) or HOLT-WINTER'S METHOD:

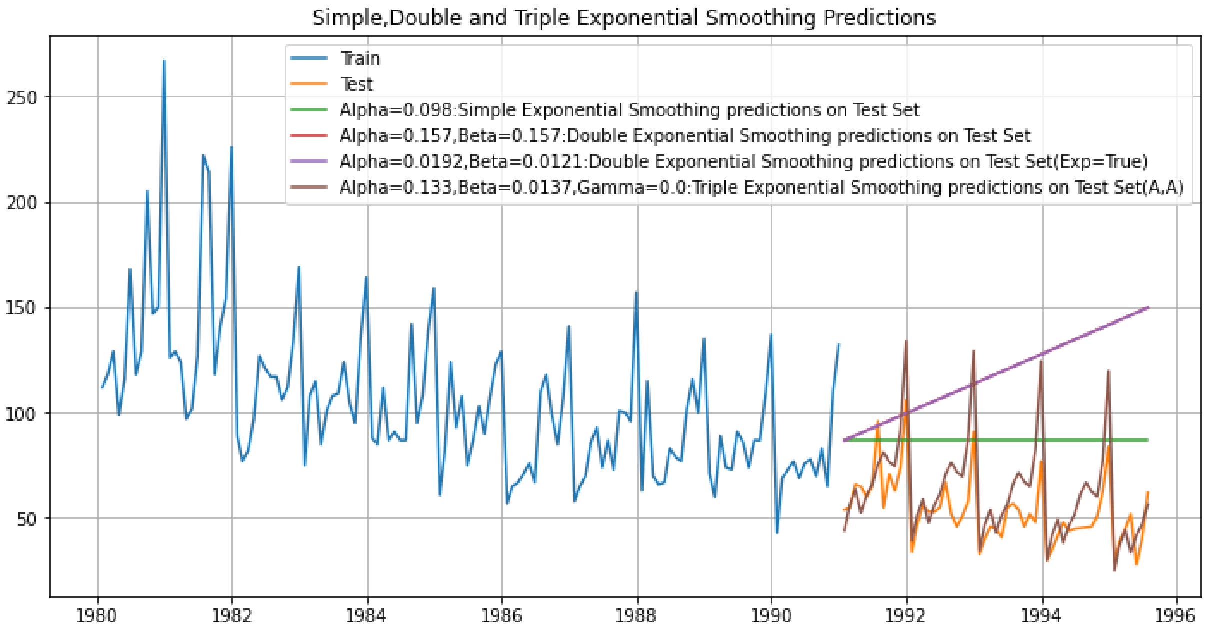
- Considering the Level, Trend, Seasonality components.
- As the seasonality can be additive or multiplicative, the TES model can be additive or multiplicative.
- We initialize the TES model using the train data and setting both the trend and seasonality as additive.
- We fit the TES model and check on the parameters as below.

```
{'smoothing_level': 0.13346905584155852,
 'smoothing_slope': 0.013798044930131528,
 'smoothing_seasonal': 0.0,
 'damping_slope': nan,
 'initial_level': 77.90998273991845,
 'initial_slope': 0.0,
 'initial_seasons': array([ 37.19347871,  49.53447903,  57.45342246,  46.82461047,
        55.5675085 ,  60.9978818 ,  70.94829431,  76.95581437,
        72.98548228,  71.11492918,  89.18261025, 131.38117683]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- We forecast the model for the duration of the test data.

```
1991-01-31     44.127161
1991-02-28     56.072510
1991-03-31     63.595803
1991-04-30     52.571339
1991-05-31     60.918586
1991-06-30     65.953308
1991-07-31     75.508069
1991-08-31     81.119938
1991-09-30     76.753955
1991-10-31     74.487750
1991-11-30     92.159780
1991-12-31    133.962695
```

- We visualize the SES, DES, and TES models together in the plot below.



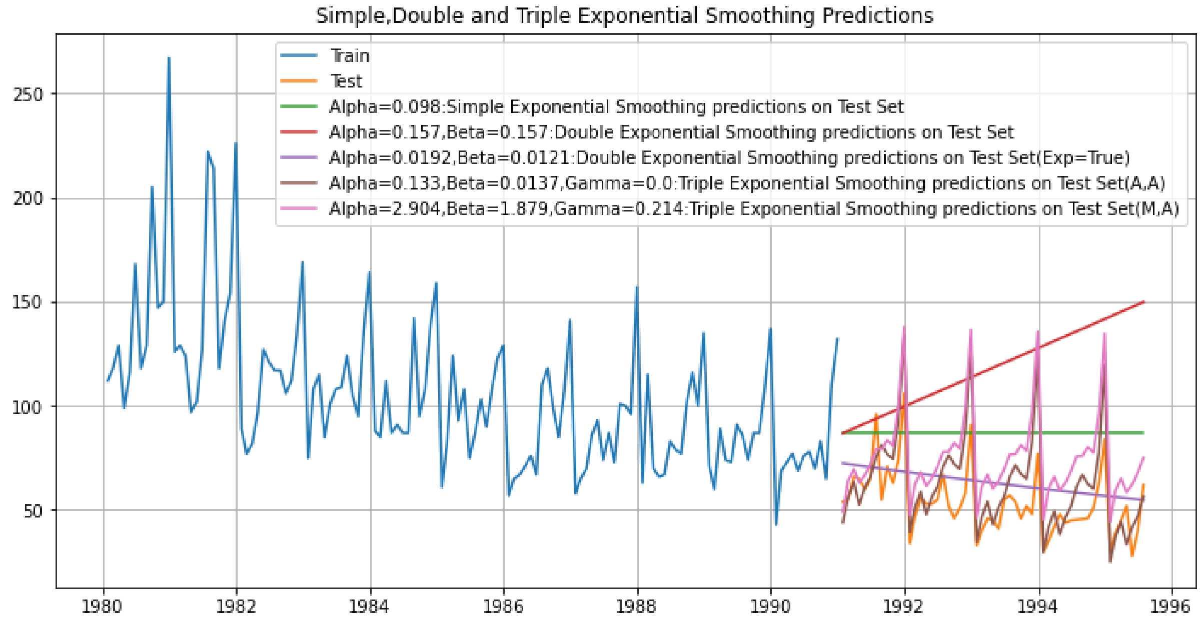
- We check on the RMSE of the test data using the TES model.

```
TES RMSE: 16.443203233657176
```

- The TES is initialized using the train data and setting the Trend as multiplicative and Seasonality as additive.
- The model is fitted and checked for the parameters.

```
{'smoothing_level': 2.9044572268848643e-10,
 'smoothing_slope': 1.8799060886954682e-13,
 'smoothing_seasonal': 0.21410032095527884,
 'damping_slope': nan,
 'initial_level': 92.76189360232776,
 'initial_slope': 0.9817396755317827,
 'initial_seasons': array([ 41.77470181,  46.52453706,  52.34729782,  30.59523424,
         43.88038643,  70.9624501 ,  83.57371289,  87.14233988,
         81.44104686,  70.31190074,  83.58551956, 161.71541608]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- The TES model is forecasted for the duration of the test data.
- We can visualize the forecast from the below plot.



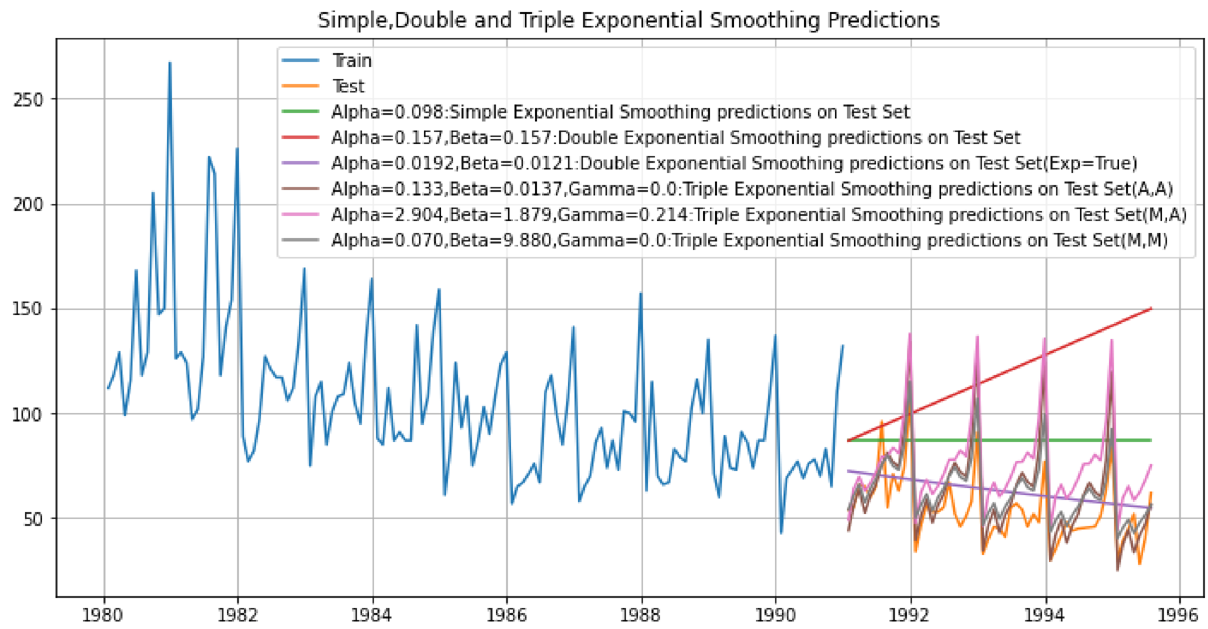
- The RMSE value of the test data using this TES model is as below.

```
TES RMSE: 25.14680702105763
```

- We initialize the TES model using the train data and setting the Trend as multiplicative and Seasonality as multiplicative.
- We fit the model and check on the parameters.

```
{'smoothing_level': 0.0700104887428228,
 'smoothing_slope': 9.88074207901599e-19,
 'smoothing_seasonal': 0.0,
 'damping_slope': nan,
 'initial_level': 76.65441503982137,
 'initial_slope': 0.993905330571457,
 'initial_seasons': array([1.45242339, 1.64428963, 1.79921093, 1.57525049, 1.76978283,
        1.90993608, 2.10105283, 2.24524946, 2.11409058, 2.07326895,
        2.41579163, 3.31143032]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- We forecast the model for the duration of the test data.
- We can visualize the model forecast ranges from the below plot.



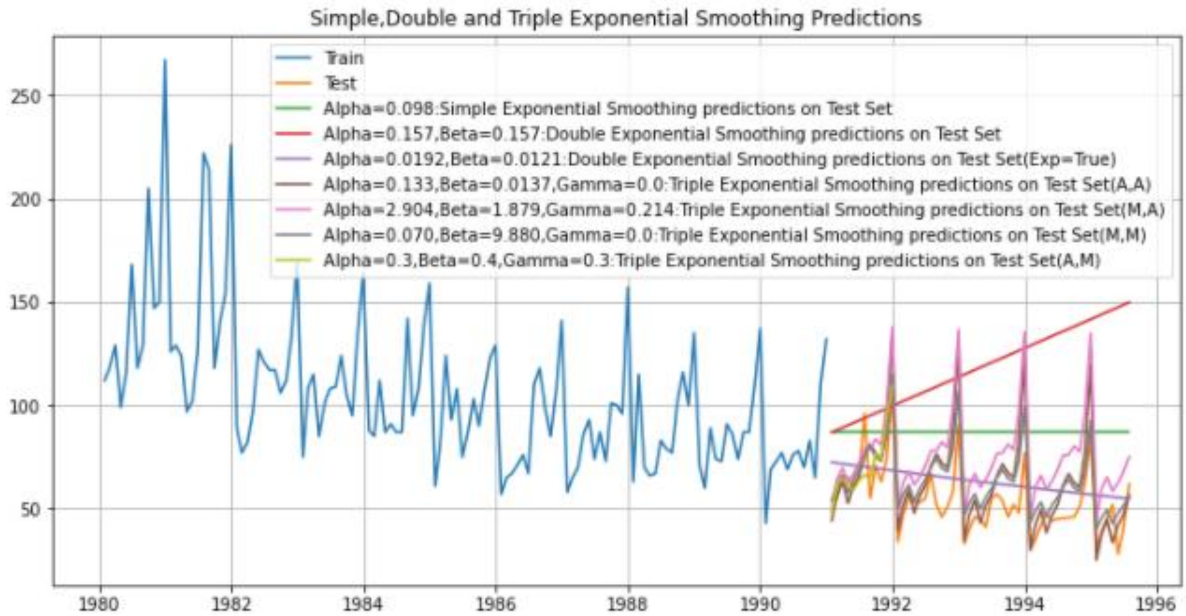
- The RMSE of the Test data using TES model is checked.

```
TES RMSE: 12.795795972028388
```

- We initialize the TES model using the train data and setting the Trend as additive and Seasonality as multiplicative.
- We fit the model with the smoothing parameters and check on the parameters.

```
{'smoothing_level': 0.3,
 'smoothing_slope': 0.4,
 'smoothing_seasonal': 0.3,
 'damping_slope': nan,
 'initial_level': 76.63636363636364,
 'initial_slope': 0.15277777777777787,
 'initial_seasons': array([1.46144721, 1.53973903, 1.68327402, 1.29181495, 1.51364176,
        2.19217082, 1.53973903, 1.68327402, 2.67497034, 1.91814947,
        1.95729537, 3.48398577]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- We forecast the model for the duration of the test data.
- We can visualize the model forecast ranges from the below plot.



- The RMSE of the Test data using TES model is checked.

Alpha=0.3,Beta=0.4,Gamma=0.3:TES	10.945435

LINEAR REGRESSION:

- For this particular linear regression, we are going to train the 'Rose' variable against the order of the occurrence. For this we need to modify our training and testing data before fitting it into a linear regression.
- We create 2 series of data called train_time and test_time for the independent variable and Rose as the dependent variable.
- Checking the head and tail rows of the train and test data.

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

TimeStamp	Rose	time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

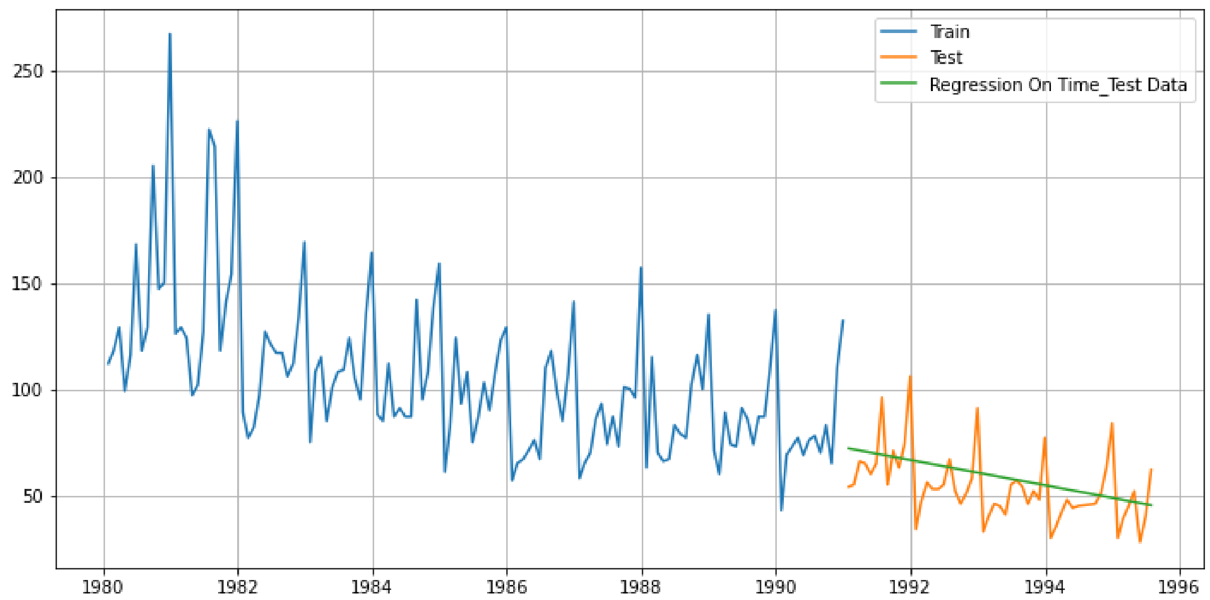
```
First few rows of Test Data
```

TimeStamp	Rose	time
1991-01-31	54.0	133
1991-02-28	55.0	134
1991-03-31	66.0	135
1991-04-30	65.0	136
1991-05-31	60.0	137

- The Linear Regression model is created and fitted on the train data.

```
lr.fit(X, y, sample_weight=None)
```

- The forecast values are visualized in the plot below.



- The RMSE of the test data is calculated using the LR model.

```
Regression on test data:  15.268955197146555
```

Naive Approach:

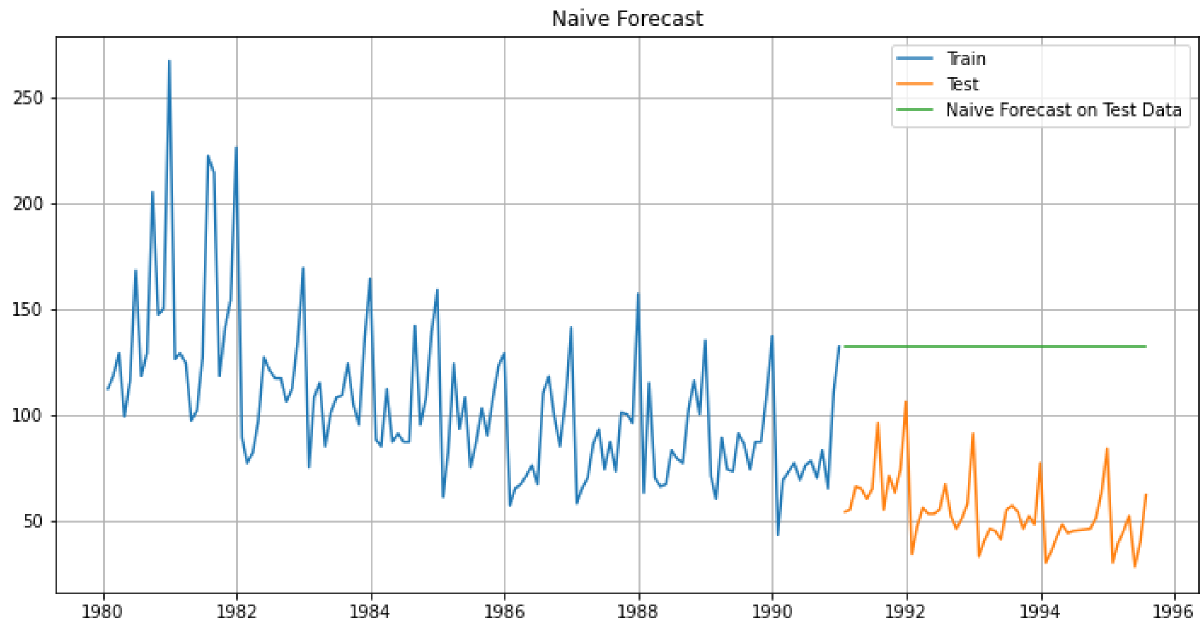
- For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.
- Checking the tail of the train data for the Naive model.

TimeStamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

- The Naive model uses the last value of the train data as the forecast value for the entire test data. We can see the same for the test data as below.

```
TimeStamp
1991-01-31    132.0
1991-02-28    132.0
1991-03-31    132.0
1991-04-30    132.0
1991-05-31    132.0
Name: naive, dtype: float64
```

- The forecast value with train and test data can be seen in the below plot.



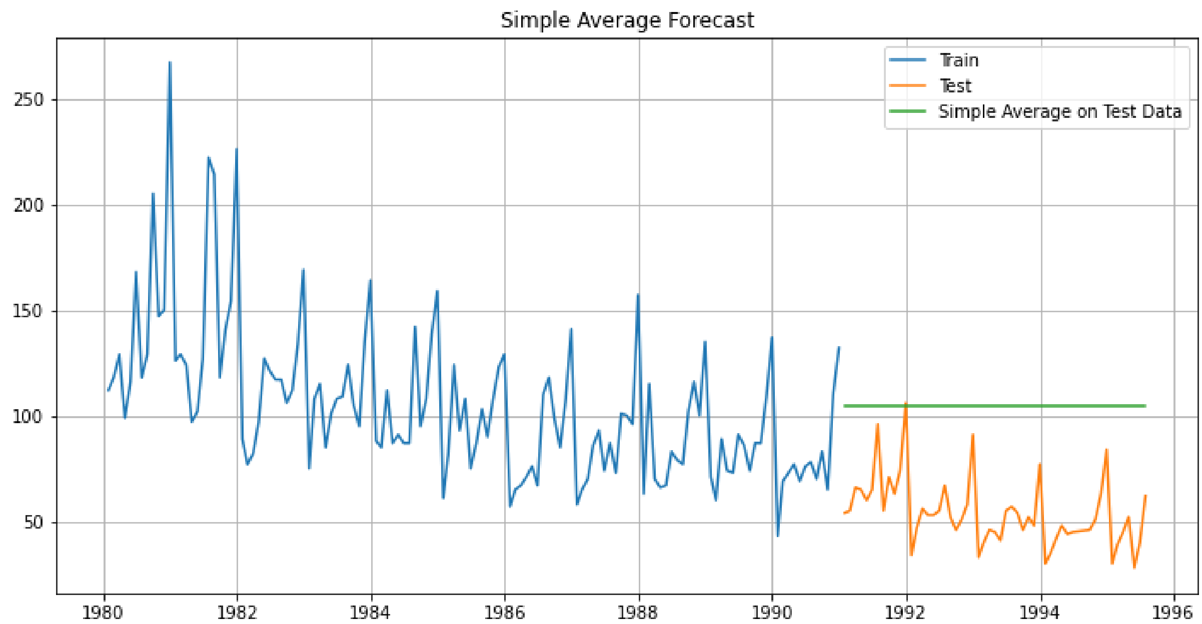
- The RMSE value for the test data can be checked.
  `For Naive forecast on the Test Data,  RMSE is 79.719`

SIMPLE AVERAGE MODEL:

- For this simple average method, we will forecast by using the average of the training values.
- The test is the mean or average of the train data. We can see the test data.

TimeStamp	Rose	mean_forecast
1991-01-31	54.0	104.939394
1991-02-28	55.0	104.939394
1991-03-31	66.0	104.939394
1991-04-30	65.0	104.939394
1991-05-31	60.0	104.939394

- We can visualize the forecast of the model along with train and test values.



- The RMSE of the test data is calculated and seen as below.

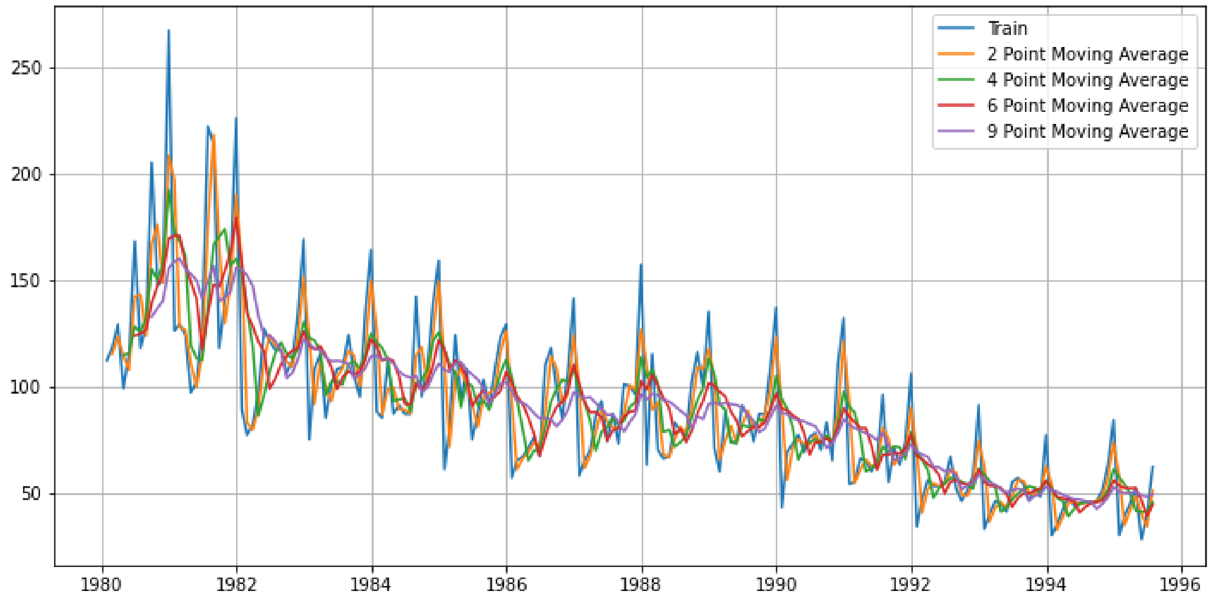
```
For Simple Average forecast on the Test Data,  RMSE is 53.461
```

MOVING AVERAGE (MA):

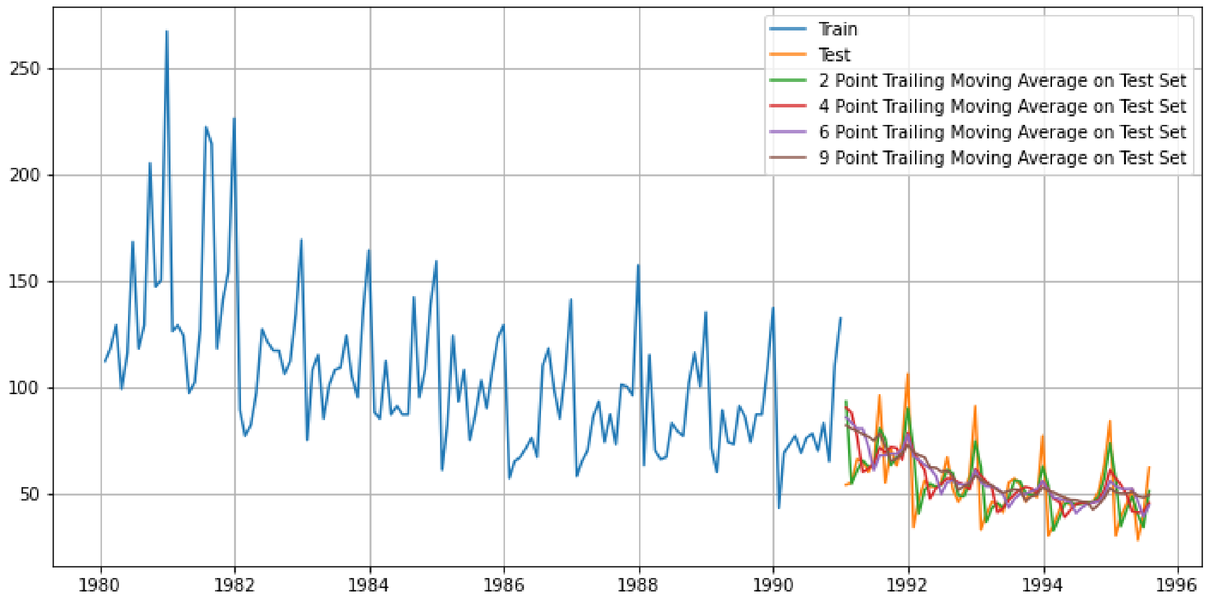
- For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals.
- The best interval can be determined by the maximum accuracy (or the minimum error) over here.
- For Moving Average, we are going to average over the entire data.
- We apply rolling mean of 2,4,6,9 on the data.
- We can see below the original data and rolling mean values.
- NaN is set where there are no values present.

TimeStamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN

- We can visualize the plot below with original data and the rolling means.



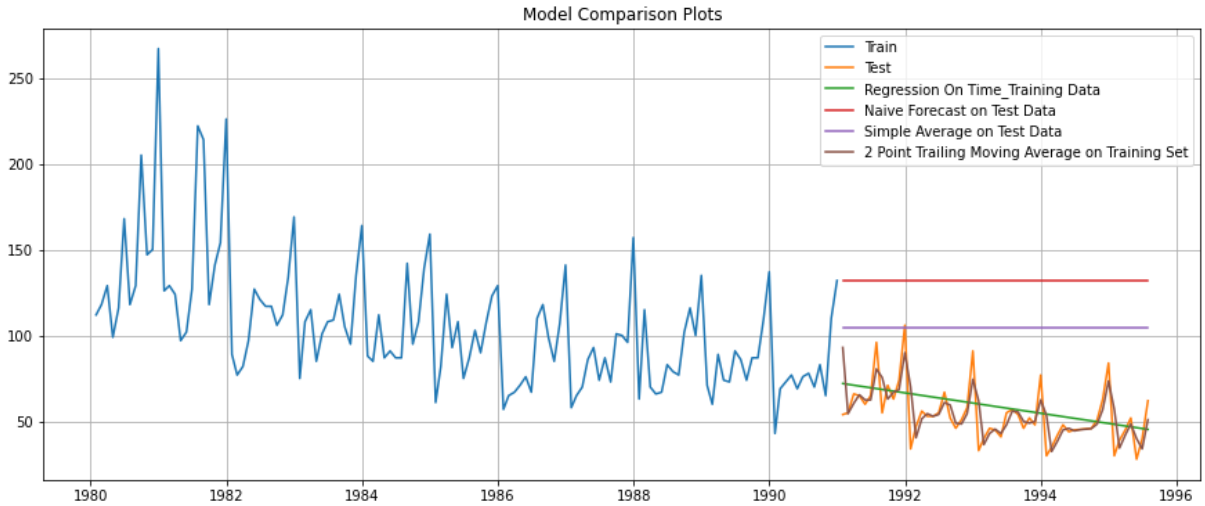
- Splitting the data as train and test data (from 1991 year). We apply the Moving average model on the test data for different trailing points.
- We can visualize the trailing values along with train and test data.



- Checking the RMSE values for the test data on the Moving average model.

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.451
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.566
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.728
```

- We can see that the 2 point Moving Average model forecast has least RMSE value comparatively.
- We plot all the models together and visualize them.



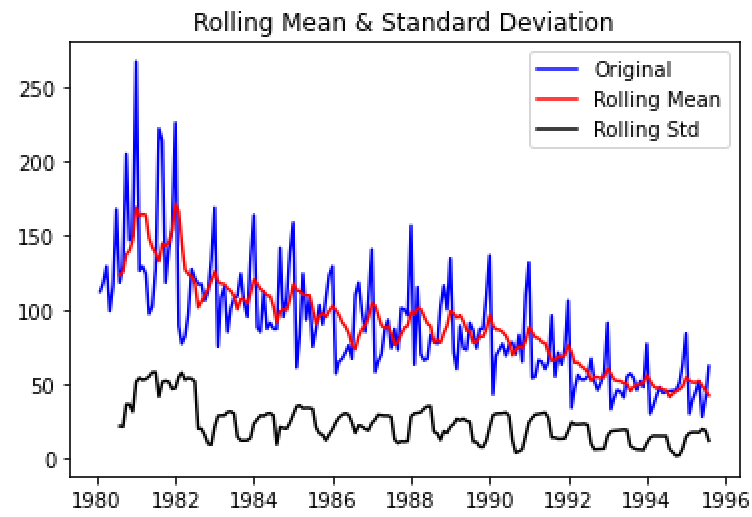
5. **Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.**
   - The stationarity of the data is checked using the Augmented Dickey Fuller (ADF) test.
   - The hypothesis for the ADFuller test is as follows:

```
STEP 1: H0 : Time Series is non-stationary
        H1 : Time Series is stationary

STEP 2: Consider the level of significanace (α) as 0.05

STEP 3: Using the Augmented Dickey Fuller (ADF) test we test for stationarity.
```

- Applying the original data on the ADFuller test and we can see the results below.

```
Results of Dickey-Fuller Test:
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
```

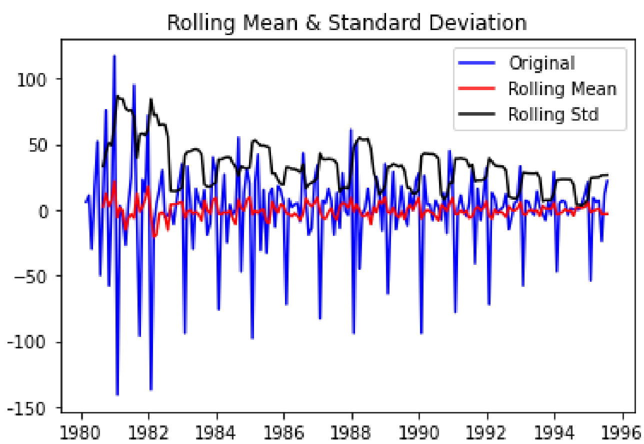
- We can check for the mean and standard deviation of the data for stationarity.



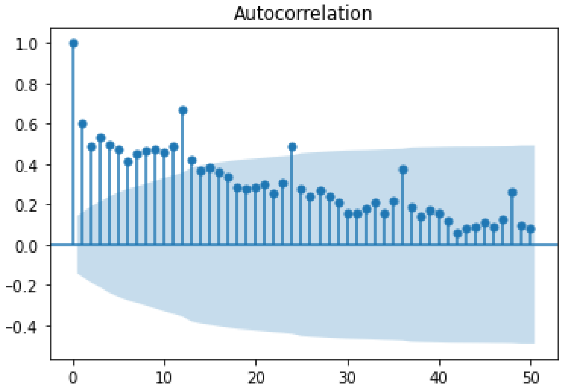
- We can see that the p-value is greater than 0.05 (Level of Significance) hence we cannot reject the Null Hypothesis.
- The given model is non-stationary and we have to make the model as stationary.
- To make a model as stationary we can take appropriate levels of differencing or apply mathematical transformations.
- Here we apply differencing with various levels and initial level with period of 1.
- Applying the differenced data to the ADFuller test and checking the results.

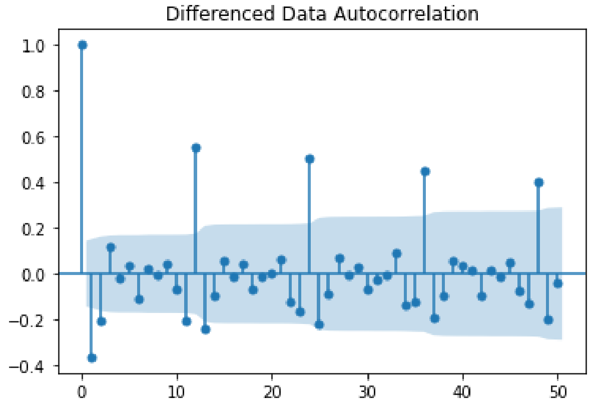
```
Results of Dickey-Fuller Test:
Test Statistic                  -8.044392e+00
p-value                          1.810895e-12
#Lags Used                       1.200000e+01
Number of Observations Used      1.730000e+02
Critical Value (1%)             -3.468726e+00
Critical Value (5%)             -2.878396e+00
Critical Value (10%)            -2.575756e+00
dtype: float64
```

- We get the p-value less than 0.05, hence rejecting the Null hypothesis and the model stationary now.
- We see that after taking a difference of order 1 the series have become stationary at $\alpha = 0.05$.
- We can plot the mean and standard deviation of the data after differencing.

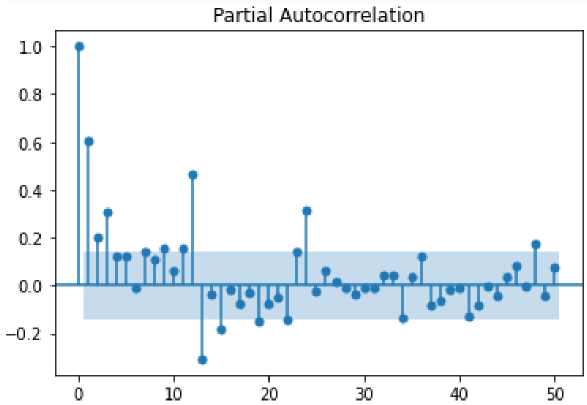


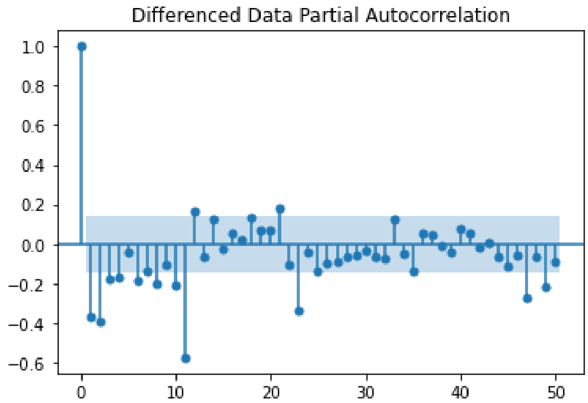
- We built the ACF plots for the original data and the data after differencing.





- We built the PACF plots for the original data and the data after differencing.





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.**

- The data has some seasonality so ideally we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model by looking at the minimum AIC criterion.

AUTOMATED ARIMA MODEL:

- We create a loop to get different combination of parameters of p (for AR) and q (for MA) in the range of 0 and 2, while the order of differencing is kept as 1 for stationarity.

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
```

- Creating a dataframe for the parameter combinations and their respective AIC values and sort them to get minimum AIC value.

	param	AIC
2	(0, 1, 2)	1276.835377
5	(1, 1, 2)	1277.359224
4	(1, 1, 1)	1277.775758
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

- Building the automated ARIMA model with the train data and order of (0, 1, 2) where AIC is low for the series.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -634.418
Method:                       css-mle   S.D. of innovations             30.167
Date:                Sun, 21 Feb 2021   AIC                           1276.835
Time:                        09:42:20   BIC                           1288.336
Sample:                    02-29-1980   HQIC                          1281.509
                         - 12-31-1990
================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -0.4885      0.085     -5.742      0.000      -0.655      -0.322
ma.L1.D.Rose   -0.7601      0.101     -7.499      0.000      -0.959      -0.561
ma.L2.D.Rose   -0.2398      0.095     -2.518      0.013      -0.427      -0.053
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2           -4.1695           +0.0000j            4.1695            0.5000
-----------------------------------------------------------------------------
```

- We forecast on the duration of the test data.

- Checking the RMSE for the test data using the ARIMA (0, 1, 2) model.

ARIMA(0,1,2)	15.618896

AUTOMATED SARIMA MODEL:

- We see that there can be a seasonality of 6 and 12 from the ACF plot. We will run our auto SARIMA models by setting seasonality as 6 and 12.

Seasonality as 6 of the Auto SARIMA Model:

- We create a loop to get different combination of parameters of p (for AR) and q (for MA) in the range of 0 and 2, while the order of differencing d is kept as 1 for stationarity and the order of differencing D is kept as 0 for the seasonal stationarity.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

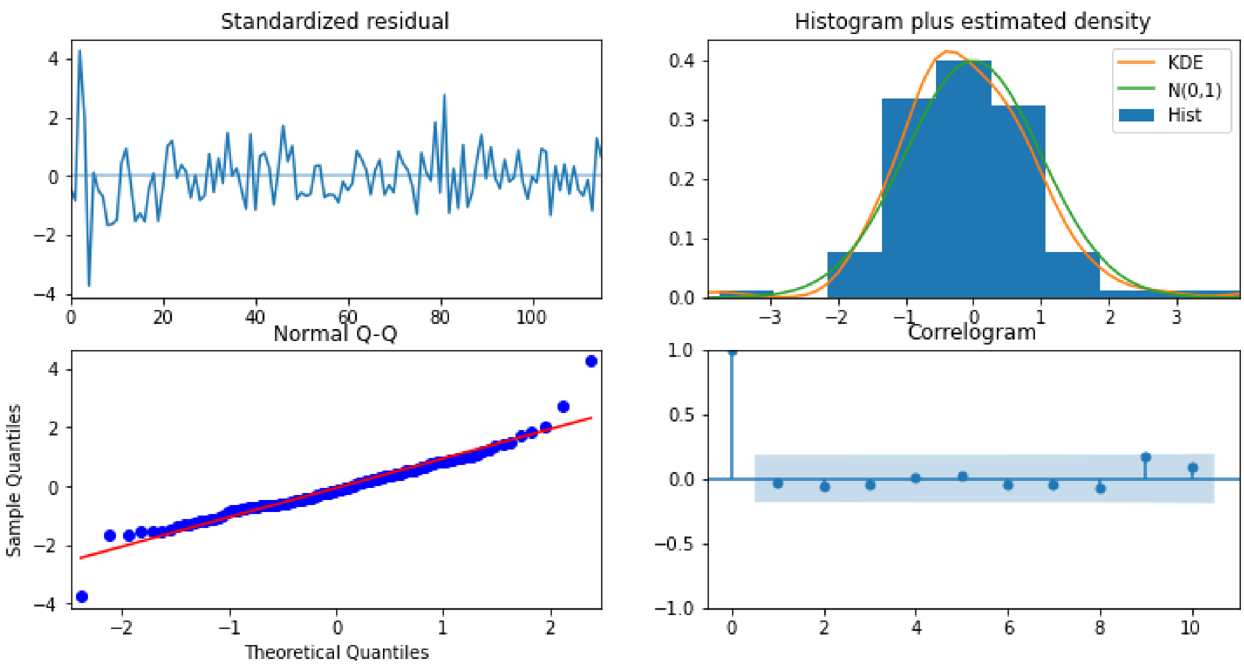
- Running the SARIMA model with all the possible parameter combinations and select the minimum AIC value.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.220888
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

- Taking the order (1, 1, 2) (2, 0, 2, 6) as the parameter and building the automated SARIMA model for the series on the train data.

```
                              Statespace Model Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -512.828
Date:                             Sun, 21 Feb 2021   AIC                           1041.656
Time:                                     09:43:25   BIC                           1063.685
Sample:                                          0   HQIC                          1050.598
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5939      0.152     -3.918      0.000      -0.891      -0.297
ma.L1         -0.1954    162.555     -0.001      0.999    -318.798     318.407
ma.L2         -0.8046    130.836     -0.006      0.995    -257.238     255.628
ar.S.L6       -0.0625      0.035     -1.789      0.074      -0.131       0.006
ar.S.L12       0.8451      0.039     21.906      0.000       0.769       0.921
ma.S.L6        0.2226    162.623      0.001      0.999    -318.513     318.958
ma.S.L12      -0.7774    126.379     -0.006      0.995    -248.477     246.922
sigma2       335.1796      0.779    430.424      0.000     333.653     336.706
===================================================================================
Ljung-Box (Q):                       15.89   Jarque-Bera (JB):                56.68
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.47   Skew:                             0.52
Prob(H) (two-sided):                  0.02   Kurtosis:                         6.26
===================================================================================
```

- Visualizing the Diagnostic plots for standardized residuals of one endogenous variable.



- We forecast on the duration of the test data.
- Checking the RMSE for the test data using the SARIMA (1, 1, 2) (2, 0, 2, 6) model.

SARIMA(0,1,2)(2,0,2,6)	26.132376

Seasonality as 12 of the Auto SARIMA Model:

- We create a loop to get different combination of parameters of p (for AR) and q (for MA) in the range of 0 and 2, while the order of differencing d is kept as 1 for stationarity and the order of differencing D is kept as 0 for the seasonal stationarity.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

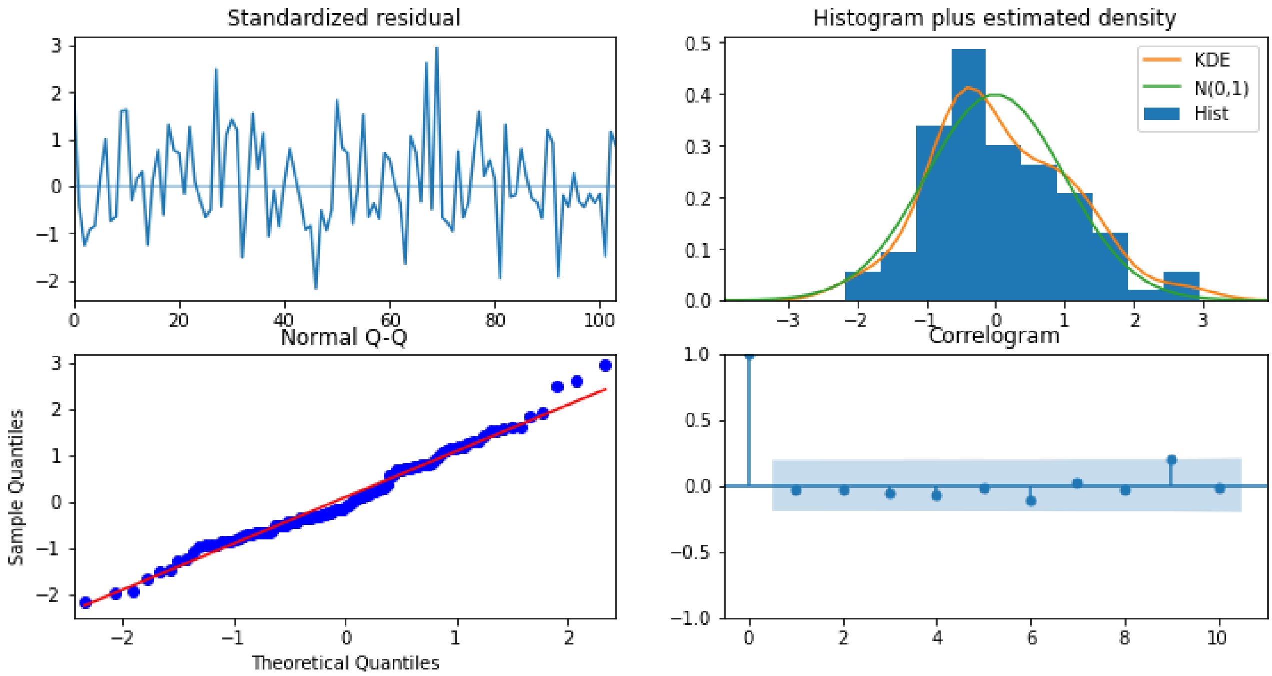
- Running the SARIMA model with all the possible parameter combinations and select the minimum AIC value.

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686906
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

- Taking the order (0, 1, 2) (2, 0, 2, 12) as the parameter and building the automated SARIMA model for the series on the train data.

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                             Sun, 21 Feb 2021   AIC                            887.938
Time:                                     09:45:13   BIC                            906.448
Sample:                                          0   HQIC                           895.437
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.842     -0.004      0.996    -372.926     371.240
ma.L2         -0.1573     29.825     -0.005      0.996     -58.614      58.299
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.77e+04      0.005      0.996   -9.33e+04    9.38e+04
===================================================================================
Ljung-Box (Q):                       24.56   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.97   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================
```

- Visualizing the Diagnostic plots for standardized residuals of one endogenous variable.



- We forecast on the duration of the test data.
- Checking the RMSE for the test data using the SARIMA (0, 1, 2) (2, 0, 2, 12) model.

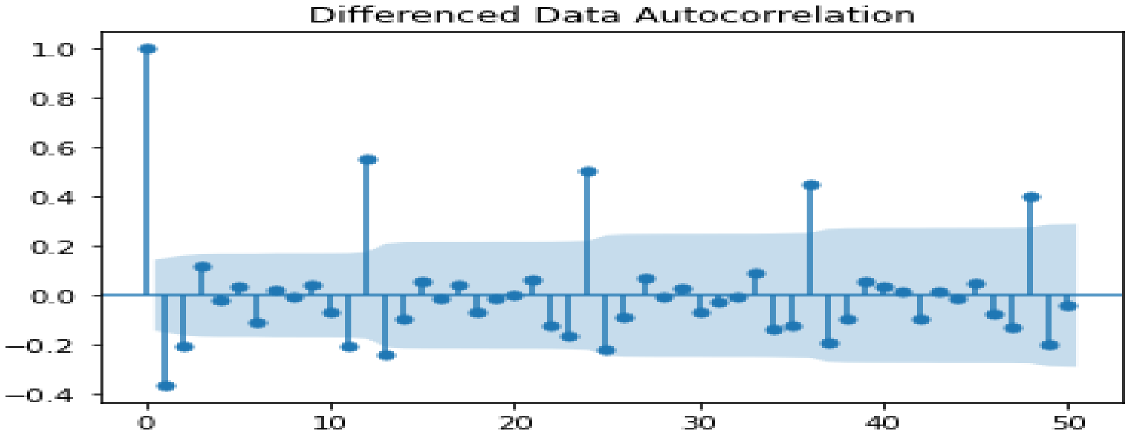
SARIMA(0,1,2)(2,0,2,12)	26.928362

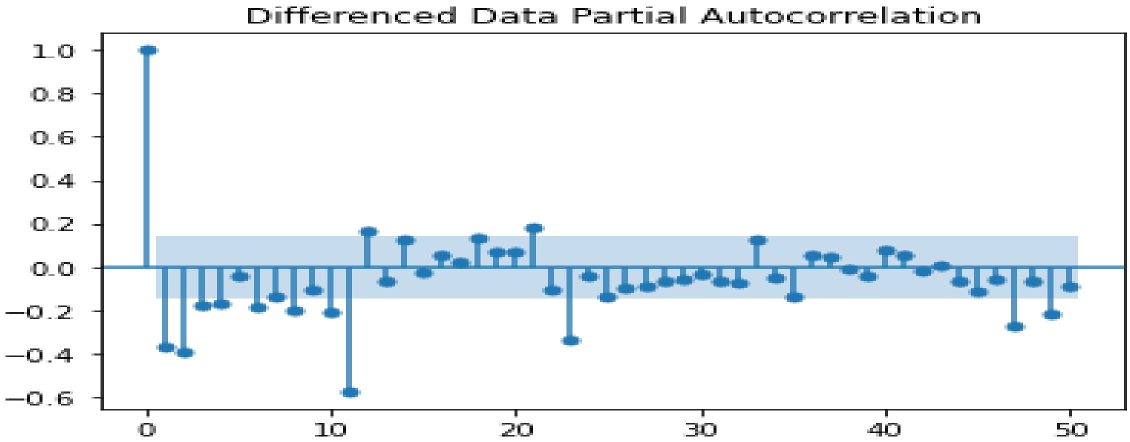
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

- The data has some seasonality so ideally we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model by looking at the ACF and the PACF plots.

ARIMA USING CUT-OFF POINTS FROM ACF AND PACF PLOTS:

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
- Checking the ACF and PACF plots for the data differencing of level 1.





- By looking at the above plots, we can say that both the PACF has 4 AR terms (p) and ACF plot has 2 MA terms(q).
- Taking the summary of ARIMA (4,1,2) as seen below.
- Check the AIC score and the p-value in the summary. We get some unexplainable results.
- We try to reduce the AR component and check the model again.

```
                              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, 21 Feb 2021   AIC                           1283.753
Time:                        09:45:20   BIC                           1306.754
Sample:                    02-29-1980   HQIC                          1293.099
                         - 12-31-1990
================================================================================
                   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     -2.693      0.008      -0.616      -0.097
ar.L3.D.Rose     0.1855      0.132      1.402      0.163      -0.074       0.445
ar.L4.D.Rose    -0.2228      0.091     -2.443      0.016      -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           -0.4116j            1.1770           -0.0569
AR.2            1.1027           +0.4116j            1.1770            0.0569
AR.3           -0.6862           -1.6643j            1.8002           -0.3122
AR.4           -0.6862           +1.6643j            1.8002            0.3122
MA.1            0.9753           -0.2209j            1.0000           -0.0355
MA.2            0.9753           +0.2209j            1.0000            0.0355
-----------------------------------------------------------------------------
```

- Hence we select the model of order (3, 1, 2) at be the best model for ARIMA using the cut-off points from the ACF and PACF plots.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(3, 1, 2)   Log Likelihood                -633.485
Method:                       css-mle   S.D. of innovations             29.950
Date:                Mon, 22 Feb 2021   AIC                           1280.969
Time:                        20:09:39   BIC                           1301.096
Sample:                    02-29-1980   HQIC                          1289.147
                         - 12-31-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4883      0.085     -5.723      0.000      -0.655      -0.321
ar.L1.D.Rose    -0.3558      0.332     -1.071      0.286      -1.007       0.295
ar.L2.D.Rose     0.0279      0.120      0.232      0.817      -0.208       0.264
ar.L3.D.Rose     0.0597      0.104      0.577      0.565      -0.143       0.263
ma.L1.D.Rose    -0.4141      0.325     -1.275      0.205      -1.051       0.223
ma.L2.D.Rose    -0.5858      0.323     -1.811      0.073      -1.220       0.048
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.8011           -1.4472j            2.3105           -0.3923
AR.2           -1.8011           +1.4472j            2.3105            0.3923
AR.3            3.1352           -0.0000j            3.1352           -0.0000
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2           -1.7070           +0.0000j            1.7070            0.5000
-----------------------------------------------------------------------------
```

- We forecast on the duration of the test data.
- Checking the RMSE for the test data using the ARIMA (3, 1, 2) model.

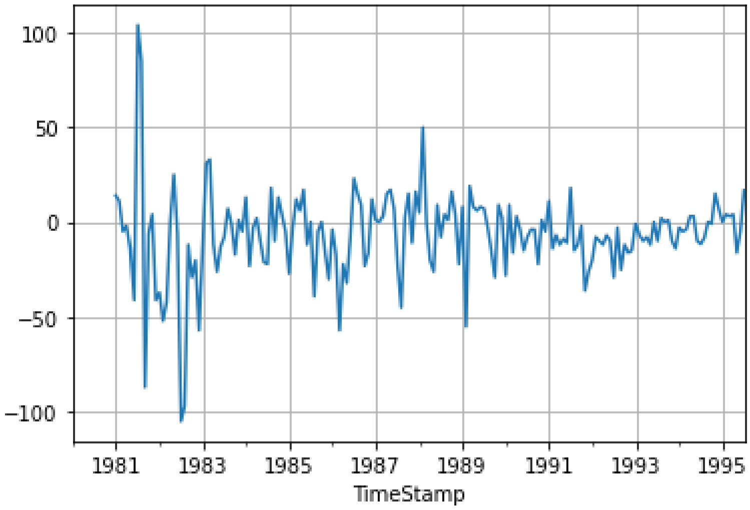
ARIMA(3,1,2)	30.689740

SARIMA USING CUT-OFF POINTS FROM ACF AND PACF PLOTS:

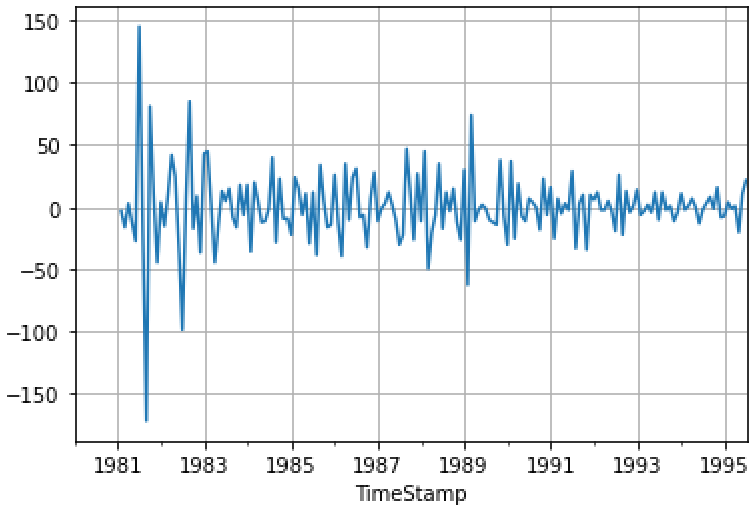
- We see that there can be a seasonality of 12 from the ACF plot. We will run our auto SARIMA models by setting seasonality as 12.

Seasonality as 12 of the SARIMA Model:

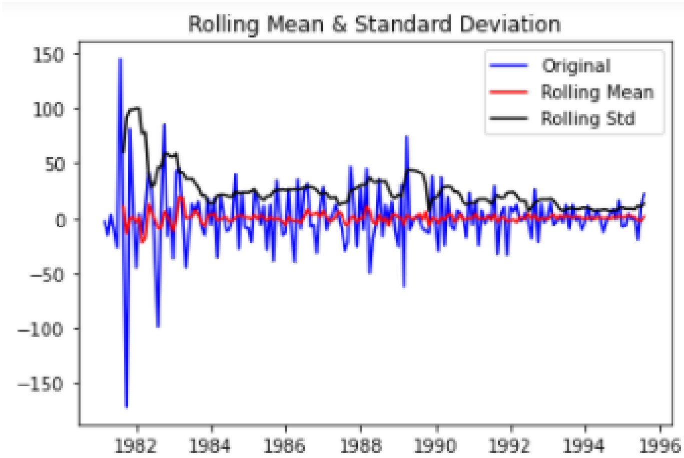
- Taking the seasonality as 12 and checking the plot of the data.



- We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

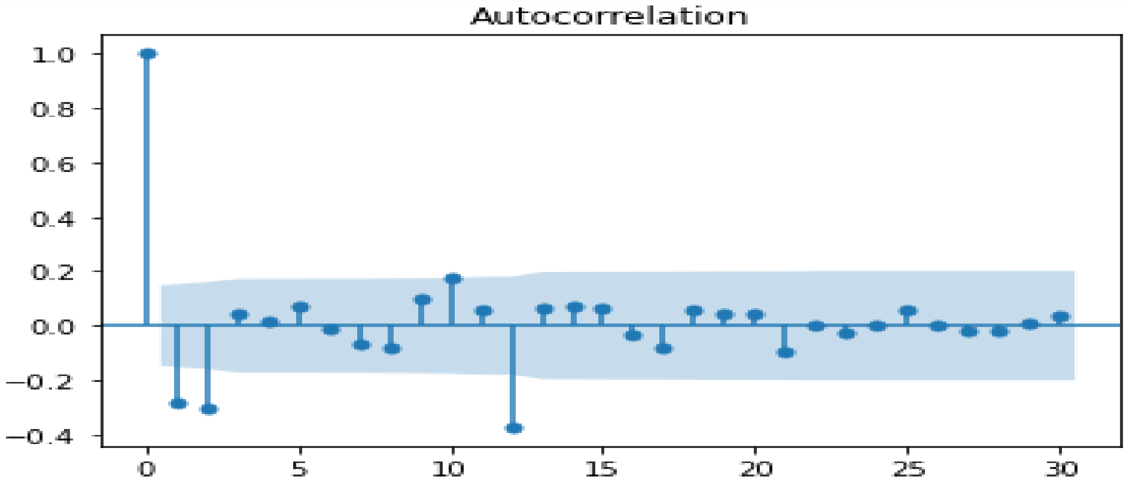


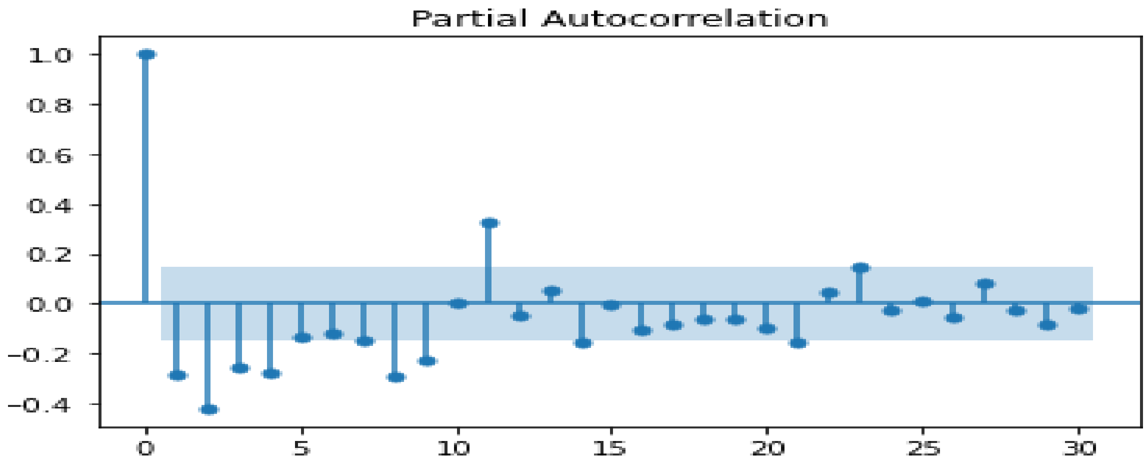
- Checking the stationarity of the seasonally first order differencing series.




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.605725
p-value                          0.000126
#Lags Used                      11.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value (10%)            -2.576373
dtype: float64
```

- The ACF and PACF plots for the seasonality is plotted and checked for the order.





- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0.
- Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).
- The below is summary of the SARIMA with order (3, 1, 2) (4, 1, 2, 12) with lowest AIC possible.

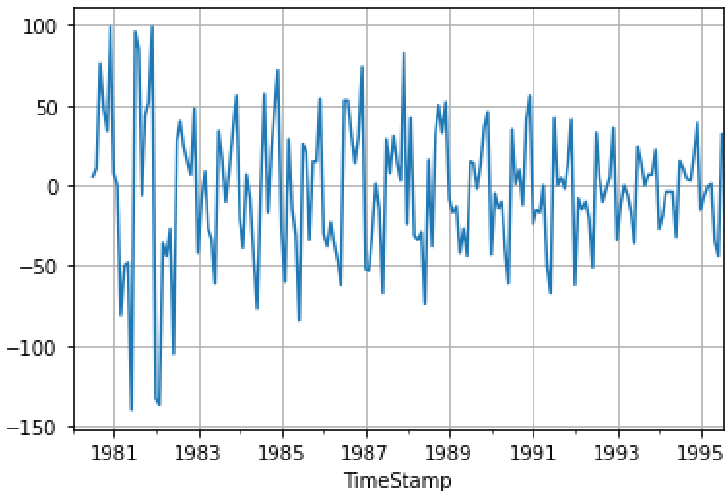
```
                                  Statespace Model Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(4, 1, 2, 12)   Log Likelihood                -281.243
Date:                            Mon, 22 Feb 2021   AIC                            586.486
Time:                                    22:35:59   BIC                            613.120
Sample:                                         0   HQIC                           597.040
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.0405      0.202     -5.163      0.000      -1.435      -0.645
ar.L2         -0.1393      0.275     -0.507      0.612      -0.678       0.399
ar.L3          0.0501      0.152      0.330      0.741      -0.247       0.347
ma.L1          0.1351      0.198      0.683      0.495      -0.252       0.523
ma.L2         -0.9142      0.212     -4.309      0.000      -1.330      -0.498
ar.S.L12      -0.7903      0.242     -3.259      0.001      -1.266      -0.315
ar.S.L24      -0.1390      0.240     -0.580      0.562      -0.609       0.331
ar.S.L36      -0.0539      0.117     -0.459      0.646      -0.284       0.176
ar.S.L48      -0.1216      0.082     -1.476      0.140      -0.283       0.040
ma.S.L12       0.2759    167.962      0.002      0.999    -328.924     329.476
ma.S.L24      -0.7250    121.640     -0.006      0.995    -239.136     237.686
sigma2       162.3103   2.73e+04      0.006      0.995   -5.33e+04    5.36e+04
===================================================================================
Ljung-Box (Q):                       46.89   Jarque-Bera (JB):                 3.32
Prob(Q):                              0.21   Prob(JB):                         0.19
Heteroskedasticity (H):               0.51   Skew:                             0.41
Prob(H) (two-sided):                  0.11   Kurtosis:                         3.70
===================================================================================
```

- We forecast on the duration of the test data.
- Checking the RMSE for the test data using the SARIMA (3, 1, 2) (4, 1, 2, 12) model.

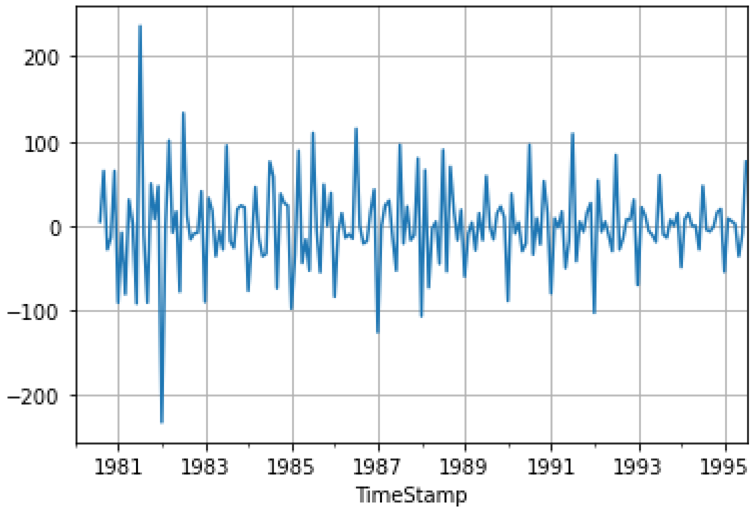
SARIMA(3,1,2)(4,1,2,12)	16.010039

Seasonality as 6 of Auto SARIMA Model:

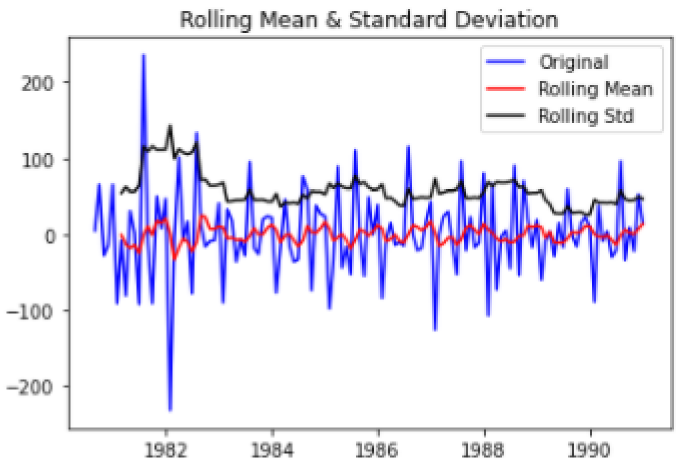
- Taking the seasonality as 6 and checking the plot of the data.



- We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

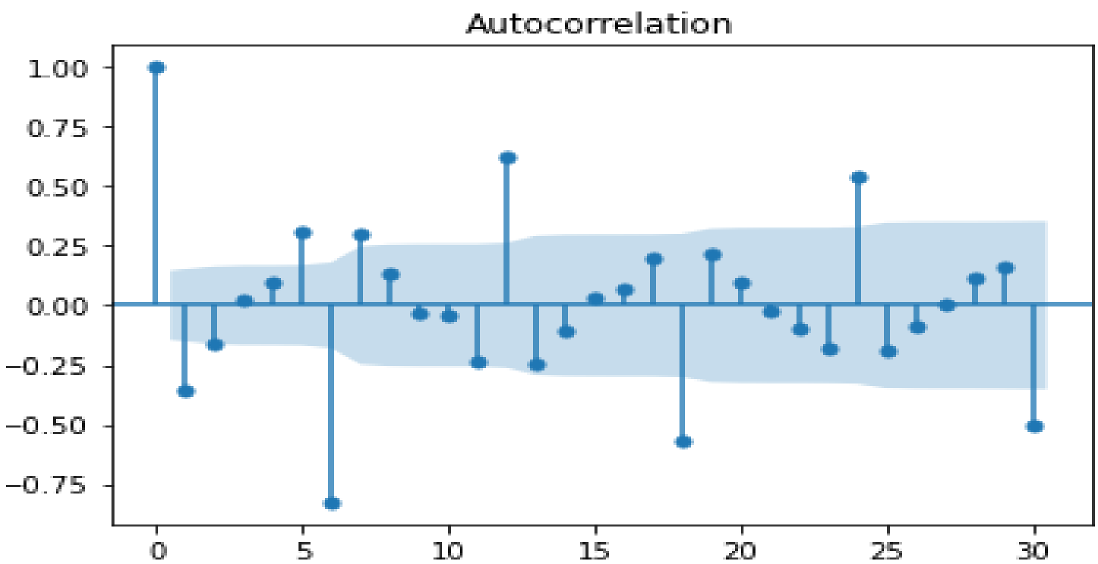


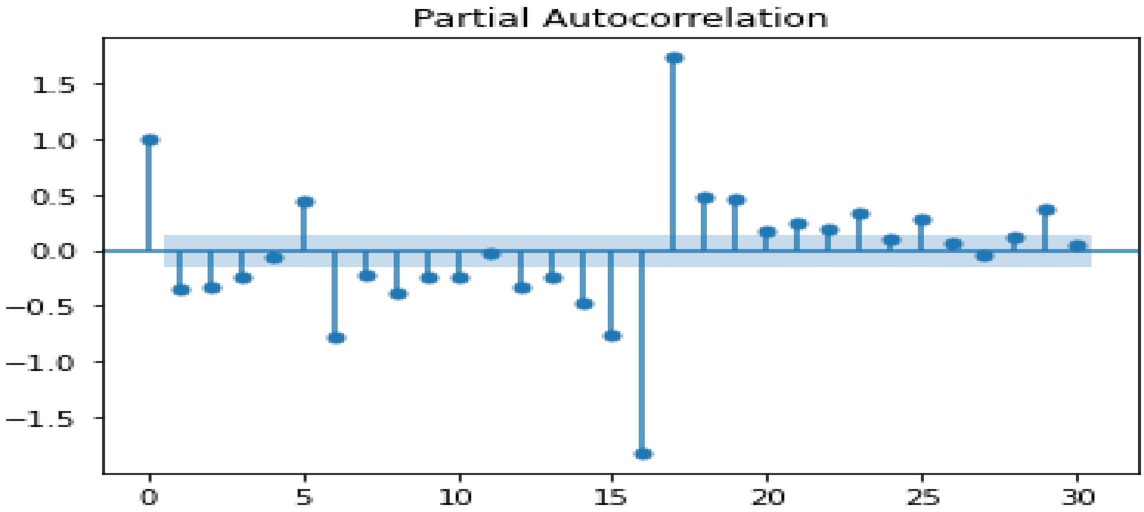
- Checking the stationarity of the seasonally first order differencing series.




```
Results of Dickey-Fuller Test:
Test Statistic                  -6.882869e+00
p-value                          1.418693e-09
#Lags Used                       1.300000e+01
Number of Observations Used      1.110000e+02
Critical Value (1%)             -3.490683e+00
Critical Value (5%)             -2.887952e+00
Critical Value (10%)            -2.580857e+00
dtype: float64
```

- The ACF and PACF plots for the seasonality is plotted and checked for the order.





- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0.
- Remember to check the ACF and the PACF plots only at multiples of 6 (since 6 is the seasonal period).
- The below is summary of the SARIMA with order (3, 1, 2) (3, 1, 2, 6) with lowest AIC possible.

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(3, 1, 2, 6)   Log Likelihood                -444.189
Date:                           Mon, 22 Feb 2021   AIC                            910.378
Time:                                   20:35:49   BIC                            939.466
Sample:                                        0   HQIC                           922.162
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7725      0.141     -5.495      0.000      -1.048      -0.497
ar.L2          0.0499      0.173      0.288      0.773      -0.290       0.390
ar.L3         -0.0821      0.115     -0.716      0.474      -0.307       0.143
ma.L1          0.1166   2129.926   5.48e-05      1.000   -4174.462    4174.695
ma.L2         -0.8833   1881.476     -0.000      1.000   -3688.508    3686.741
ar.S.L6       -0.9376      0.125     -7.516      0.000      -1.182      -0.693
ar.S.L12      -0.4482      0.147     -3.050      0.002      -0.736      -0.160
ar.S.L18      -0.3474      0.102     -3.418      0.001      -0.547      -0.148
ma.S.L6       -0.0285      0.173     -0.165      0.869      -0.368       0.311
ma.S.L12       0.0359      0.170      0.211      0.833      -0.297       0.368
sigma2       288.5420   6.15e+05      0.000      1.000    -1.2e+06     1.2e+06
===================================================================================
Ljung-Box (Q):                       24.29   Jarque-Bera (JB):                 4.58
Prob(Q):                              0.98   Prob(JB):                         0.10
Heteroskedasticity (H):               0.72   Skew:                             0.51
Prob(H) (two-sided):                  0.34   Kurtosis:                         3.09
===================================================================================
```

- We forecast on the duration of the test data.
- Checking the RMSE for the test data using the SARIMA (3, 1, 2) (3, 1, 2, 6) model.

SARIMA(3,1,2)(3,1,2,6)	18.035140

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

```
In Rose the sorted order of Test RMSE values:
```

	Test RMSE
Alpha=0.3,Beta=0.4,Gamma=0.3:TES	10.945435
2pointTrailingMovingAverage	11.529278
Alpha=0.070,Beta=9.880,Gamma=0.0:TES	12.795796
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
ARIMA(3,1,2)	15.522264
ARIMA(0,1,2)	15.618896
SARIMA(3,1,2)(4,1,2,12)	16.010039
Alpha=0.133,Beta=0.0137,Gamma=0.0:DES	16.443203
Alpha=0.0192,Beta=0.0121:DES	17.223483
SARIMA(3,1,2)(3,1,2,6)	18.035140
Alpha=2.904,Beta=1.879,Gamma=0.214:TES	25.146807
SARIMA(0,1,2)(2,0,2,6)	26.132376
SARIMA(0,1,2)(2,0,2,12)	26.928362
Alpha=0.098,SES	36.796244
SimpleAverageModel	53.460570
Alpha=0.157,Beta=0.157:DES	70.572452
NaiveModel	79.718773

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

- The most optimum model on the complete data is Triple Exponential Smoothing or Holt - Winter's Method as seen from the above table in the sorted order from the least RMSE value.
- Checking on the TES parameters.

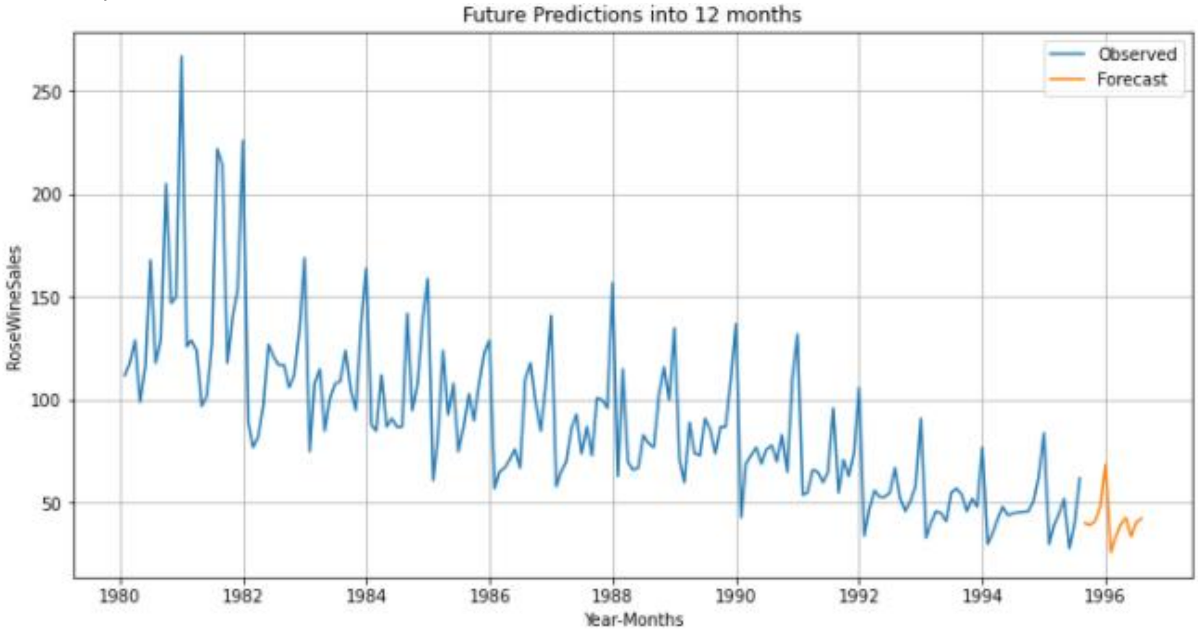
```
{'smoothing_level': 0.3,
 'smoothing_slope': 0.4,
 'smoothing_seasonal': 0.3,
 'damping_slope': nan,
 'initial_level': 64.0,
 'initial_slope': 0.15277777777777787,
 'initial_seasons': array([1.75     , 1.84375 , 2.015625, 1.546875, 1.8125  , 2.625   ,
       1.84375 , 2.015625, 3.203125, 2.296875, 2.34375 , 4.171875]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

- Using the TES model, forecasting for the duration of 12 months into the future.

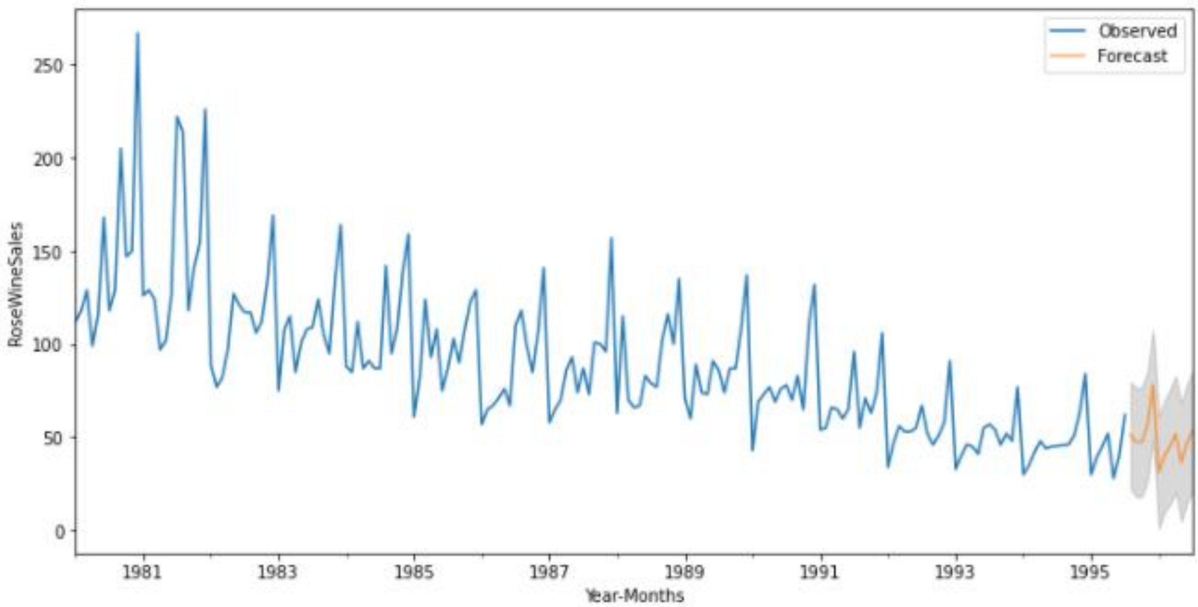
- The predicted values for 12 months into the future are as follows.

```
1995-08-31    40.074648
1995-09-30    39.245730
1995-10-31    41.312369
1995-11-30    48.067469
1995-12-31    68.735454
1996-01-31    25.994292
1996-02-29    33.337602
1996-03-31    39.734801
1996-04-30    42.746984
1996-05-31    33.573536
1996-06-30    40.505265
1996-07-31    42.410526
Freq: M, dtype: float64
```

- The plot with the observed data and forecasted data is seen below.

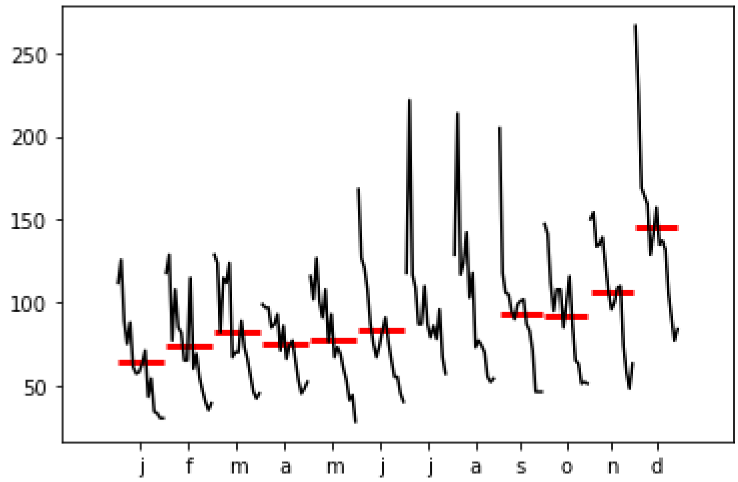


- Using the SARIMA model, we get the below forecast with confidence interval.

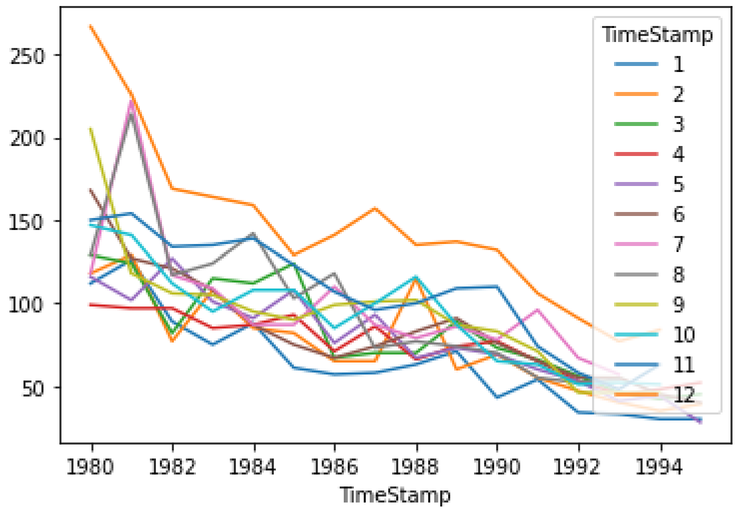


10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- The series seems to be multiplicative while decomposition and the best model (least RMSE) is when the seasonality of the model is considered as multiplicative.



- Every year in the month of December, the sale is high. It might be due to the festive season of the year.
- The sale seems to be fairly good at the start of the series (1980) and gradually decreases across the years.



- The sale is decreasing across the years, this might be because the customer's preference is changing or the quantity of the production is reduced in the company.
- In the forecast, we can see that the sale is further reduced for the year 1995-1996.
- From this we can say that in near future the sales might get reduce further hence immediate steps to be taken by the company to raise the sale.
- The company can try some methods to directly meet the customers and get feedback and try changing them in the production side.
- The company can use its Marketing & Sales team to pitch in new ideas that are different from their competitors.
- The company can try for discount sale or free sale occasionally and grab the attention of the customers.