Model Solution

Time Series Forecasting Assessment

DSBA

Contents

TSF - Rose

1. Time Series Forecast for Rose Dataset

1.1. Get the data and start analysis

Let us get started.

Load the required packages, set the working directory and load the data file.

We would be reading the file, by proving the index_col as 'YearMonth'

Dataset has 187 rows with 2 missing values and 1 feature. There is one feature which is float64 type.

Let us start the data exploration step with the head function to look at first 5 initial rows.

Let us check the head and tail of the dataset

Head

	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

	Rose
YearMonth	
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

Let us check the data types of the different variables present in the dataset .

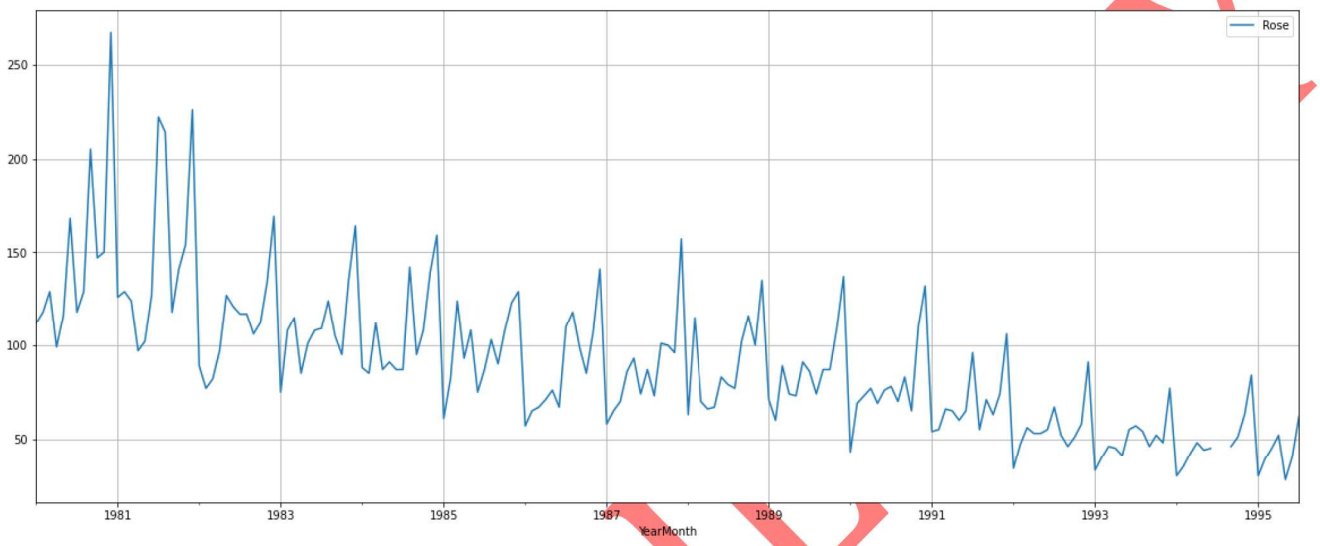
```
Datetime Index: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
```

There are 185 data points, there are two null values and the Rose dependent variable is integer type.

The number of rows: 187 with two missing values
The number of columns: 1

The data is loaded with the index as YearMonth

Plot of original data with missing values



Suppose we predict the Rose for the next few months, we will look at the past values and try to gauge and extract a pattern.

Here we observe a pattern within each year indicating a seasonal effect. Such observations will help us in predicting future values.

Note: We have used only one variable, Rose (the Rose sale of the past 15 years).

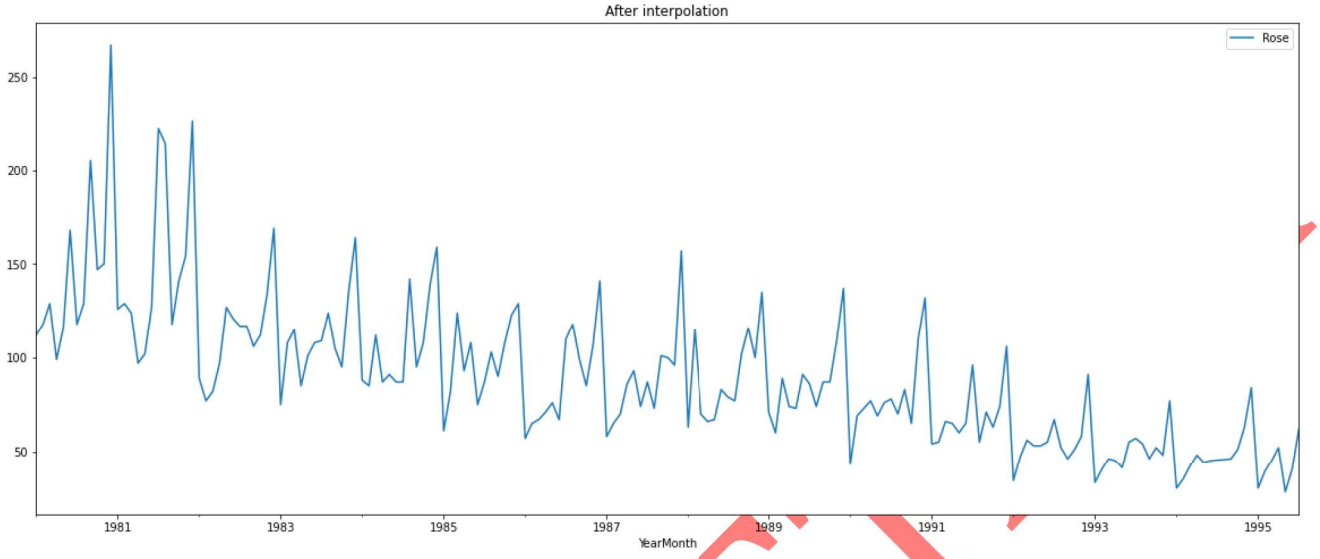
Hence, this is will be Univariate Time Series Analysis/Forecasting.

Strong downward Trend is present and Seasonality might be present.

1.2. Imputing the missing value using the interpolate function

Using the interpolate function with method as linear to impute the two missing observations.

Plot of TS post missing value imputation



Five point summary of the Rose Dataset

	Rose
count	187.00
mean	89.91
std	39.24
min	28.00
25%	62.50
50%	85.00
75%	111.00
max	267.00

The mean value is 89.91 and the median is 85. Range of the Rose is 239.

1.3. Exploratory Data Analysis

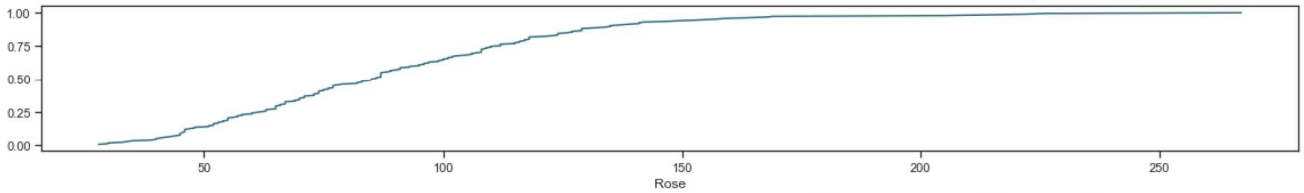
1.3.1. Univariate Time Series

A univariate time series is a series with a single time-stamped variable at time t.

Here the dataset belongs to the Rose sales from the January of 1980 to July of 1995. Here, Rose is the time-dependent variable.

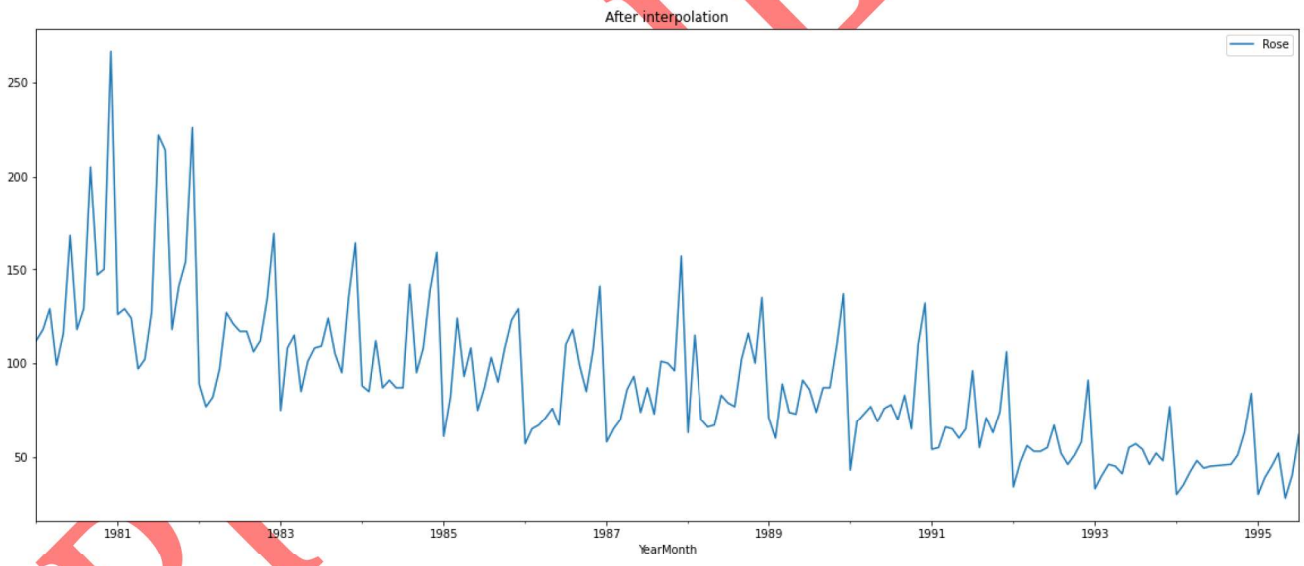
The series is a monthly series, wherein for each month between Jan-1980 and Jul-1995 a datapoint is recorded.

1.3.2. Plot ECDF: Empirical Cumulative Distribution Function



An ECDF is an estimator of the Cumulative Distribution Function. The ECDF essentially allows you to plot a feature of your data in order from least to greatest and see the whole feature as if is distributed across the data set. The data ranges from 28 to 267.

1.3.3. Plot for the Rose timeseries post missing value imputation



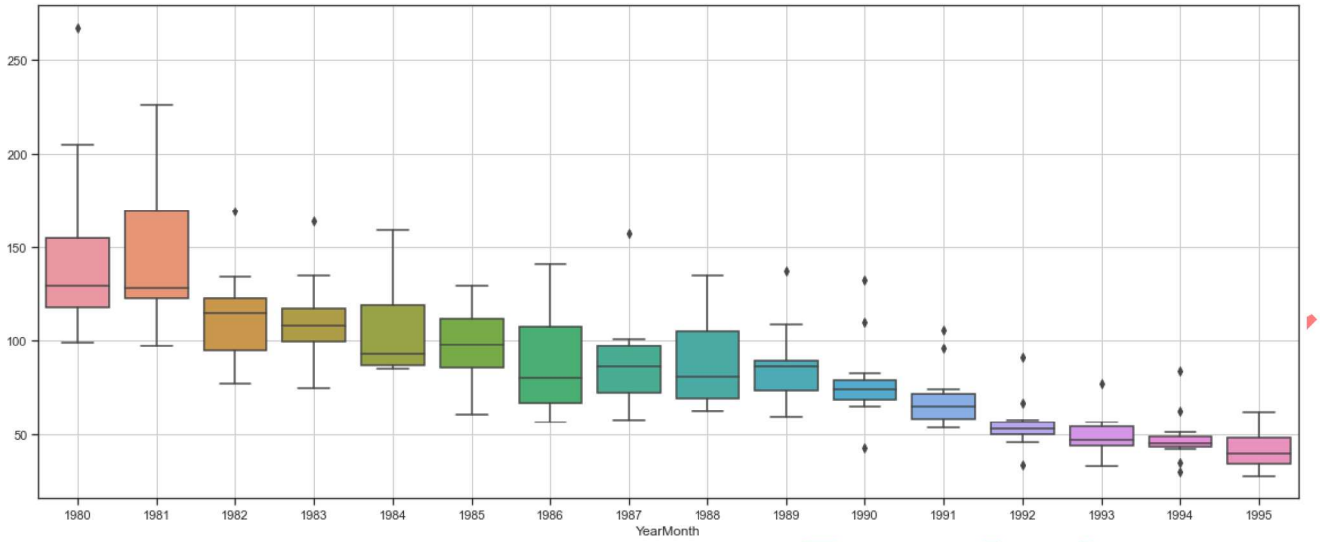
Suppose we predict the Rose for the next few months, we will look at the past values and try to gauge and extract a pattern.

Here we observe a pattern within each year indicating a seasonal effect. Such observations will help us in predicting future values.

Note: We have used only one variable, Rose (the Rose sale of the past 15 years).

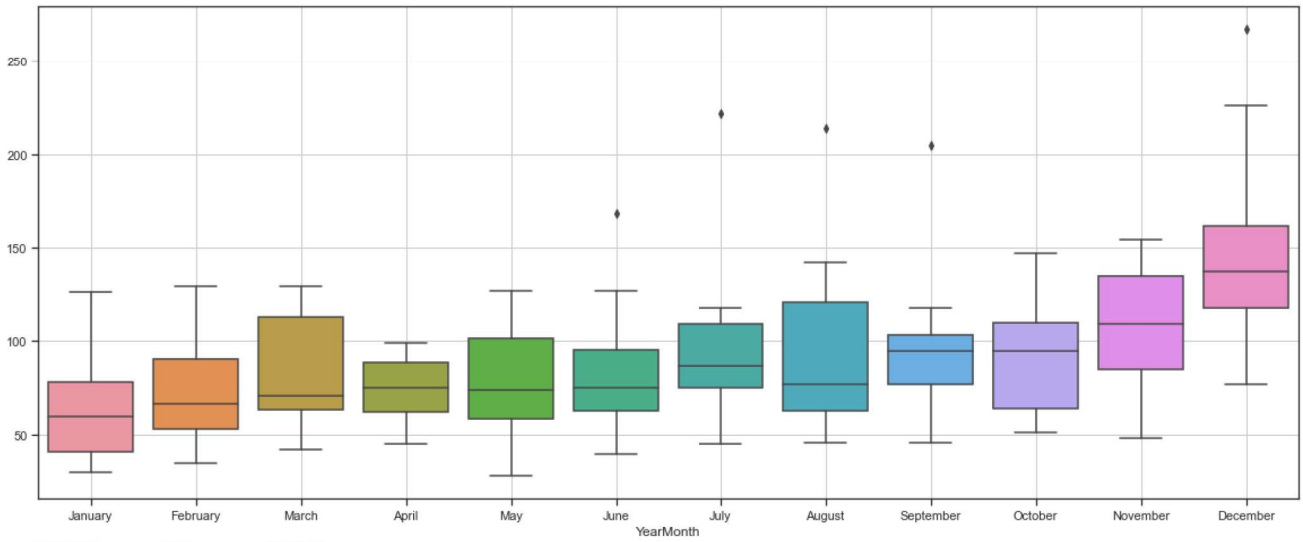
Hence, this is will be Univariate Time Series Analysis/Forecasting.

1.3.4. Box plot per year wise



Across the year there are not much noticeable difference, though there are few high ranges in initial years. Also the last year is showing less sale due to the fact that the data is recorded for 7 out of 12 months in year 1995.

1.3.5. Box plot per month wise



There is a maximum sale in month of December, followed by November and then October. These fluctuations can be attributed to the months being the holiday month or months. June shows the minimum sale.

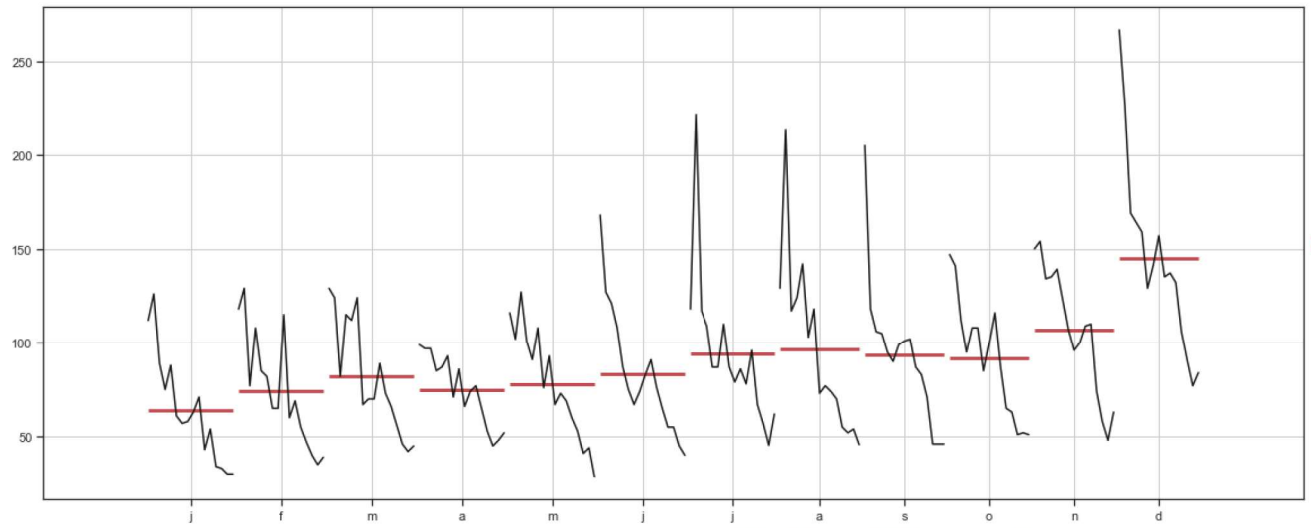
Do remember that there is one less observation for a few months in the year of 1995.

1.3.6. Year/Month Table

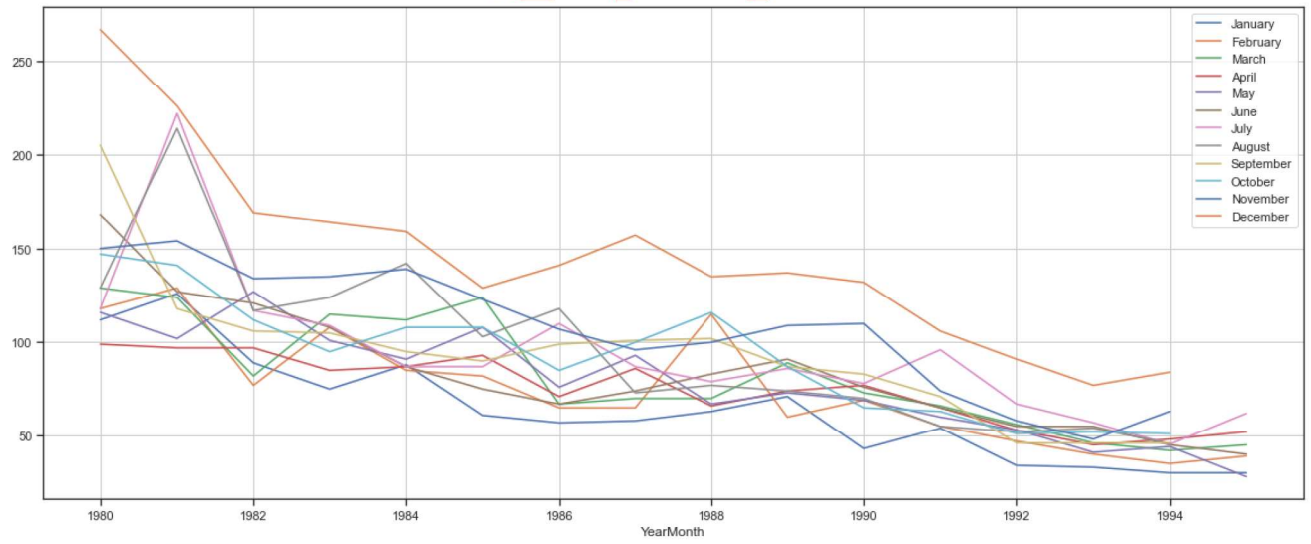
YearMonth	January	February	March	April	May	June	July	August	September	October	November	December
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	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

Post July 1995 there is no datapoint for 1995 year. This Aug to Dec are having one less datapoint than other months.

1.3.7. Month plot for each month distribution



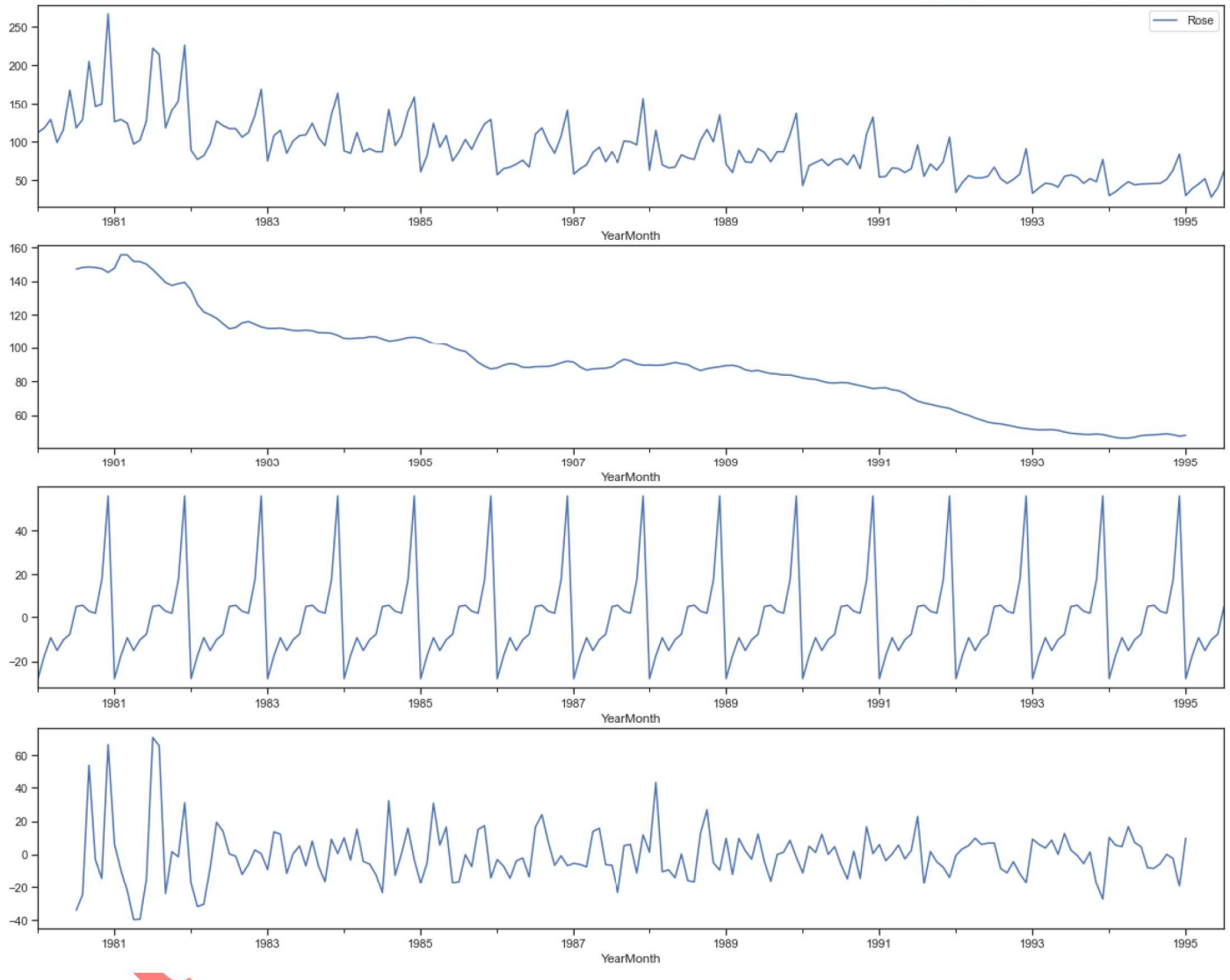
1.3.8. Line plot for comparison between each month of each year



As previously seen Month of December is outperforming for almost every year.

1.3.9. Additive Decomposition

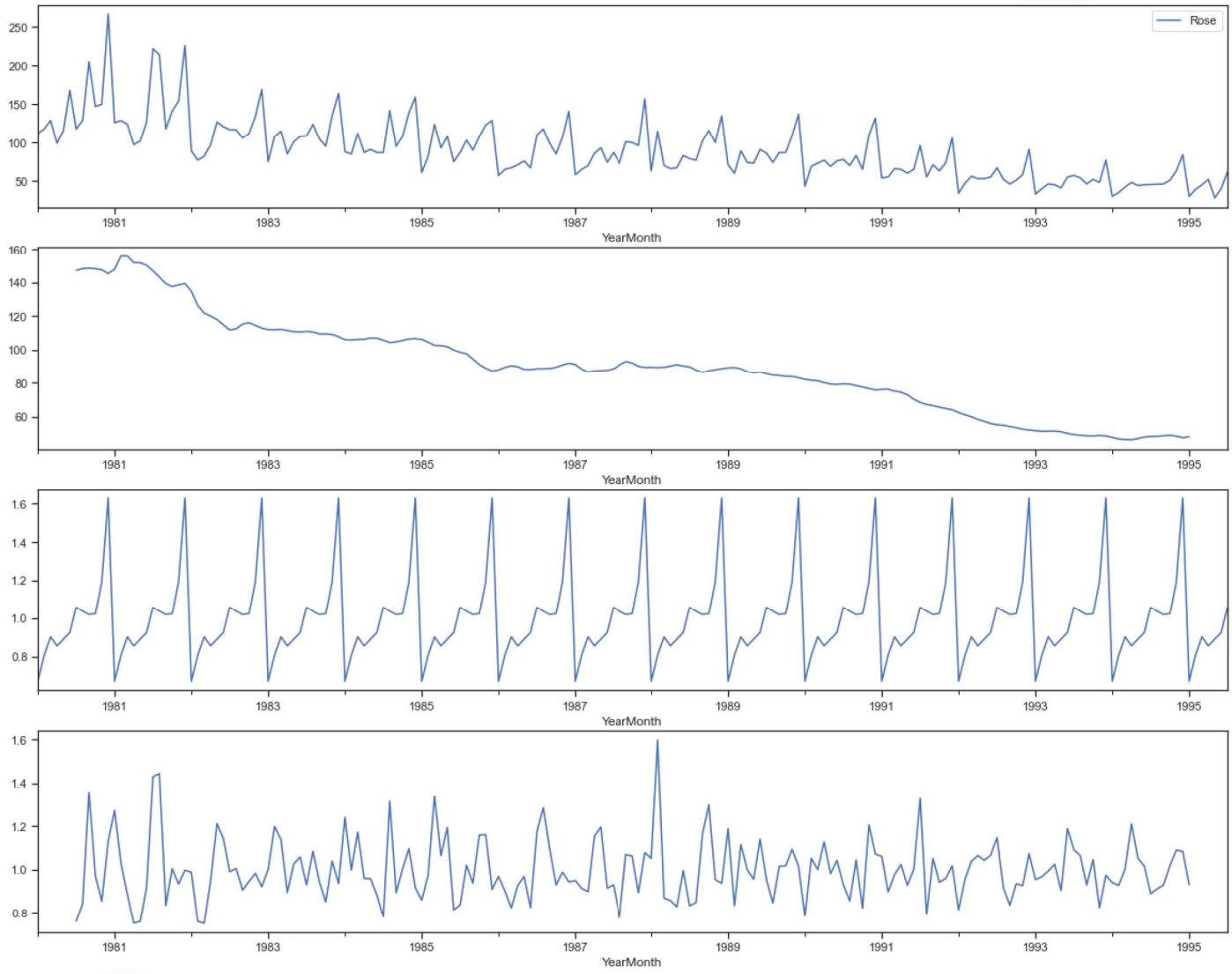
- An additive model suggests that the components are added together.
- An additive model is linear where changes over time are consistently made by the same amount. The seasonal correction is added with the trend.
- A linear seasonality has the same frequency (width of the cycles) and amplitude (height of the cycles).



Running the above code performs the decomposition and plots the 4 resulting series. We observe that the trend and seasonality are clearly separated.

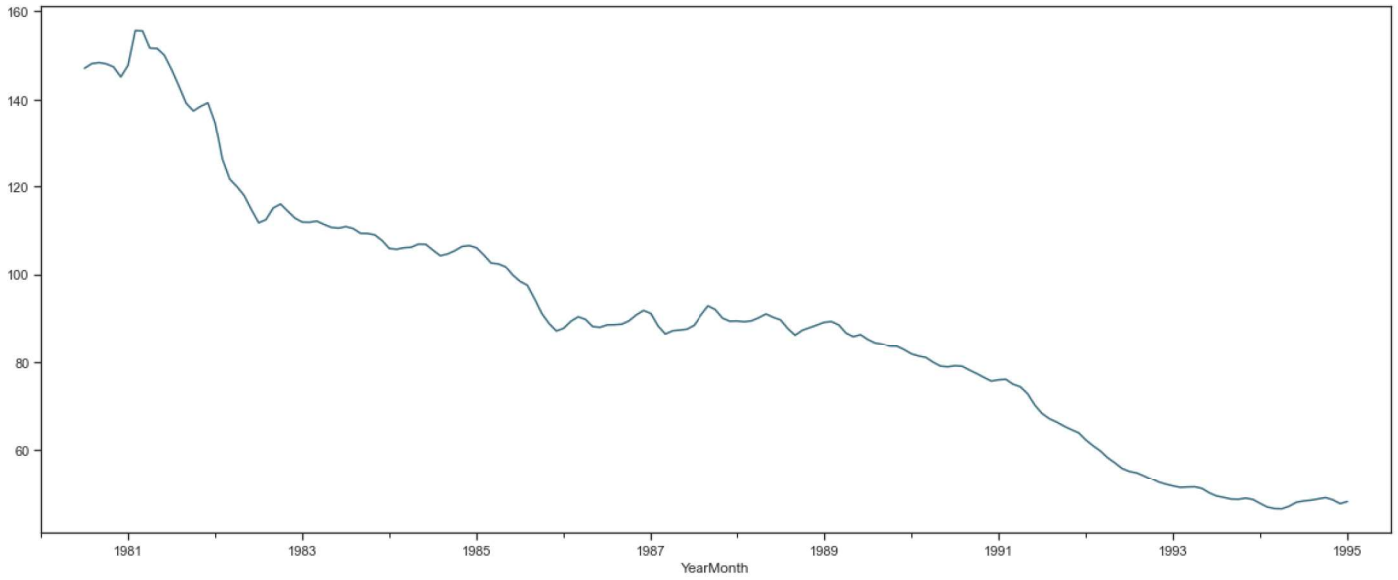
1.3.10. Multiplicative Decomposition

- A multiplicative model suggests that the components are multiplied together.
- A multiplicative model is non-linear.
- The seasonal correction is multiplied with the trend.
- A non-linear seasonality has an increasing or decreasing frequency (width of the cycles) and / or amplitude (height of the cycles) over time.



Running the above code performs the decomposition and plots the 4 resulting series. We observe that the trend and seasonality are clearly separated.

1.3.11. Closer look to the trend in the data set



There is an upward trend in the initial half which seems to be at a peak and then move in the downward direction. This will help in the forecast as in the most resent time it seems to have downward trend.

1.4. Split the data into train and test and plot the training and test data

- Split the data into train and test.
- Build different time series models on train data set and test it on test data set
- Compare the models' performance.
- The test data begins 1991 onwards. Anything before 1991 can considered in the training data so long they are contiguous.
- Size of the training dataset is (132, 1)
  Size of the test dataset is (55, 1)

First few rows of Training Data

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

Last few rows of Training Data

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

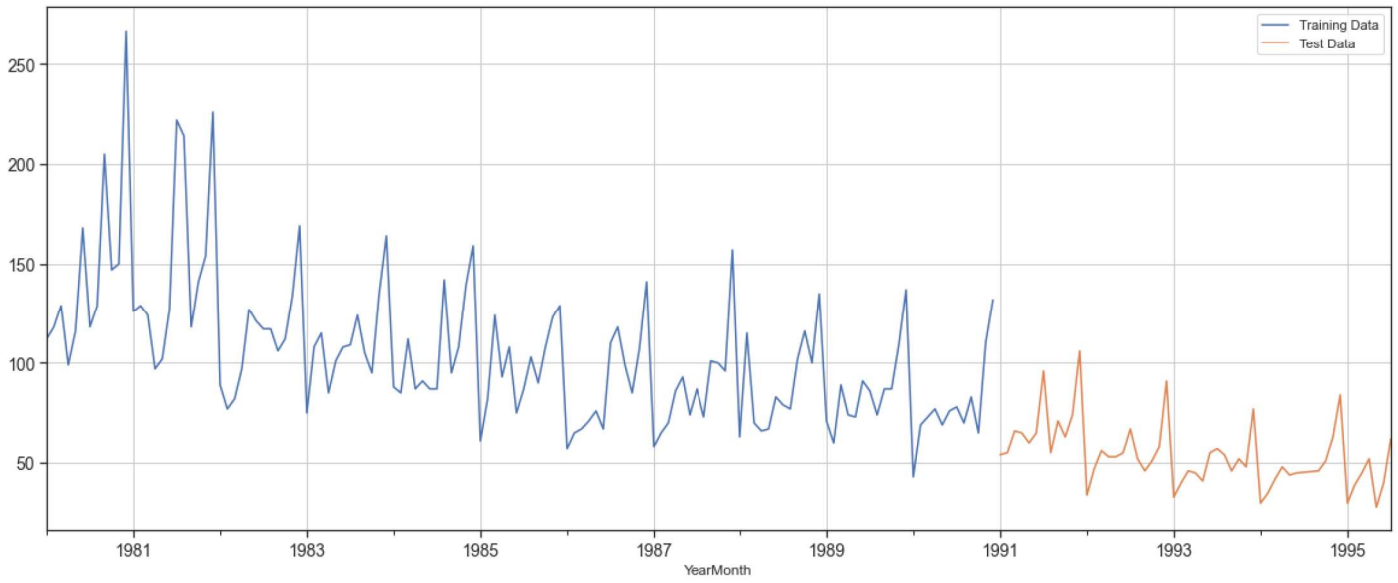
First few rows of Test Data

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

Last few rows of 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

1.4.1. Joint plot for training and test data

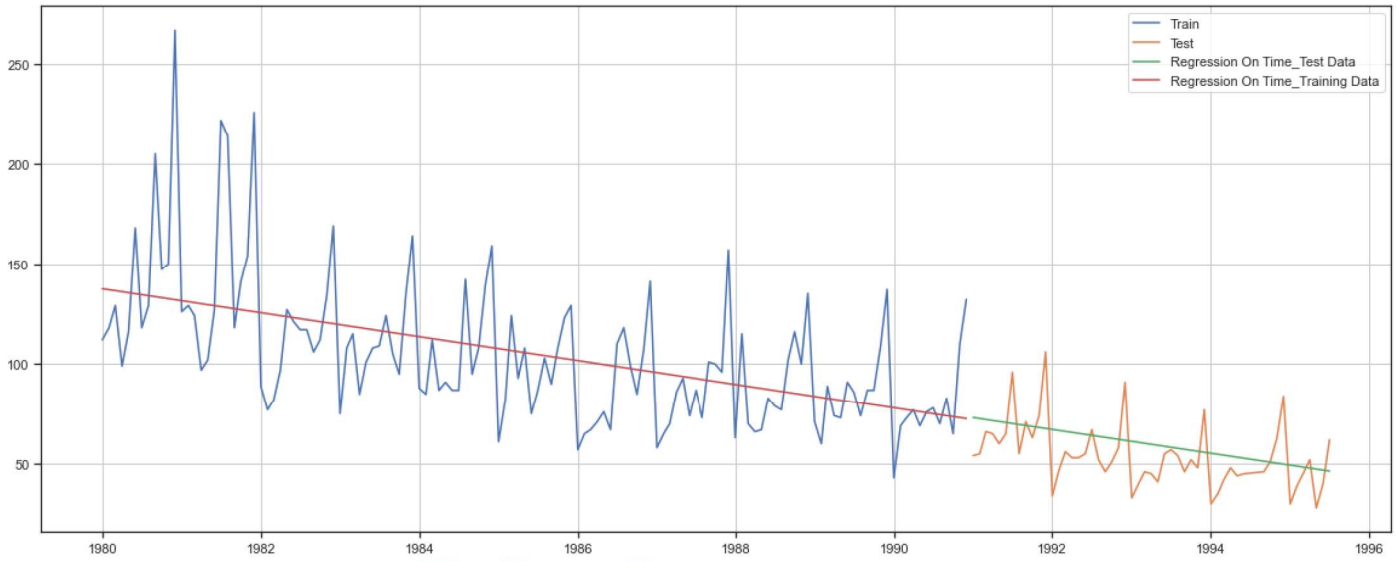


1.5. Building different models and comparing the accuracy metrics.

1.5.1. Model 1: Linear Regression

For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

Model 1 Plot - Linear Regression



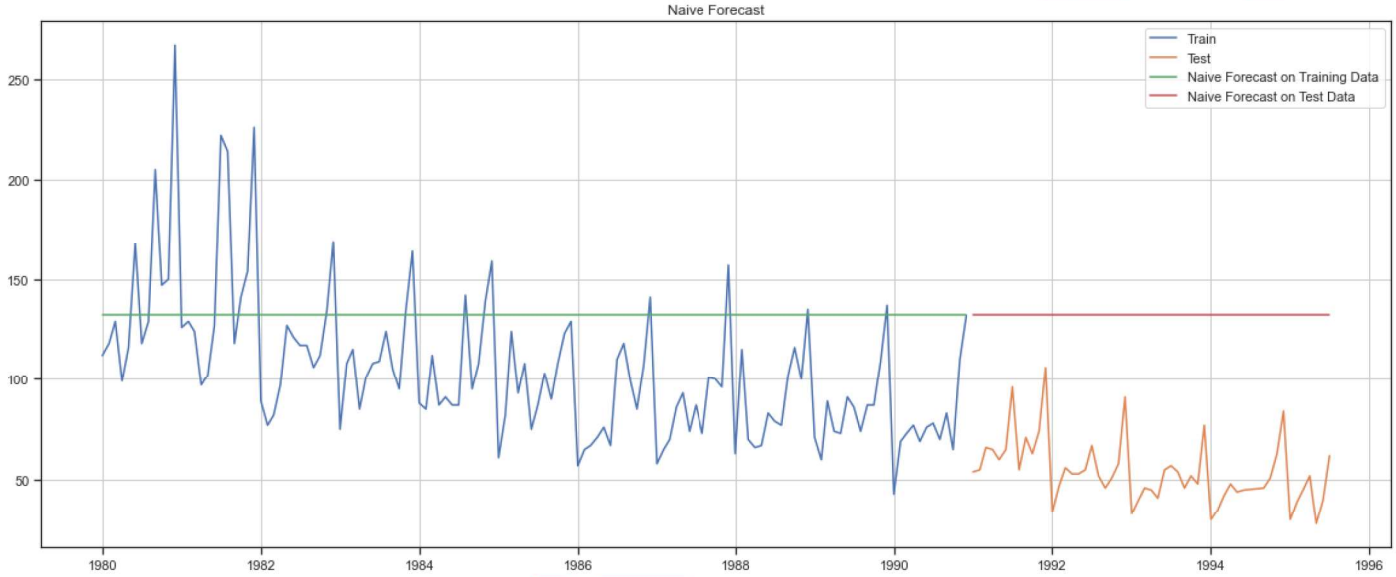
Model 1 Evaluation - Linear Regression

For RegressionOnTime forecast on the Training Data, RMSE is 30.718
For RegressionOnTime forecast on the Test Data, RMSE is 15.612

1.5.2. Model 2: 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 the same as today, therefore the prediction for day after tomorrow is also today.

Model 2 Plot - Naive



Model 2 Evaluation - Naive

For Naive Model forecast on the Training Data, RMSE is 45.064
For Naive Model forecast on the Test Data , RMSE is 79.719

We can infer from the RMSE values and the above graphs that the Naive method and Regression on Time models might not be suited for datasets with high variability.

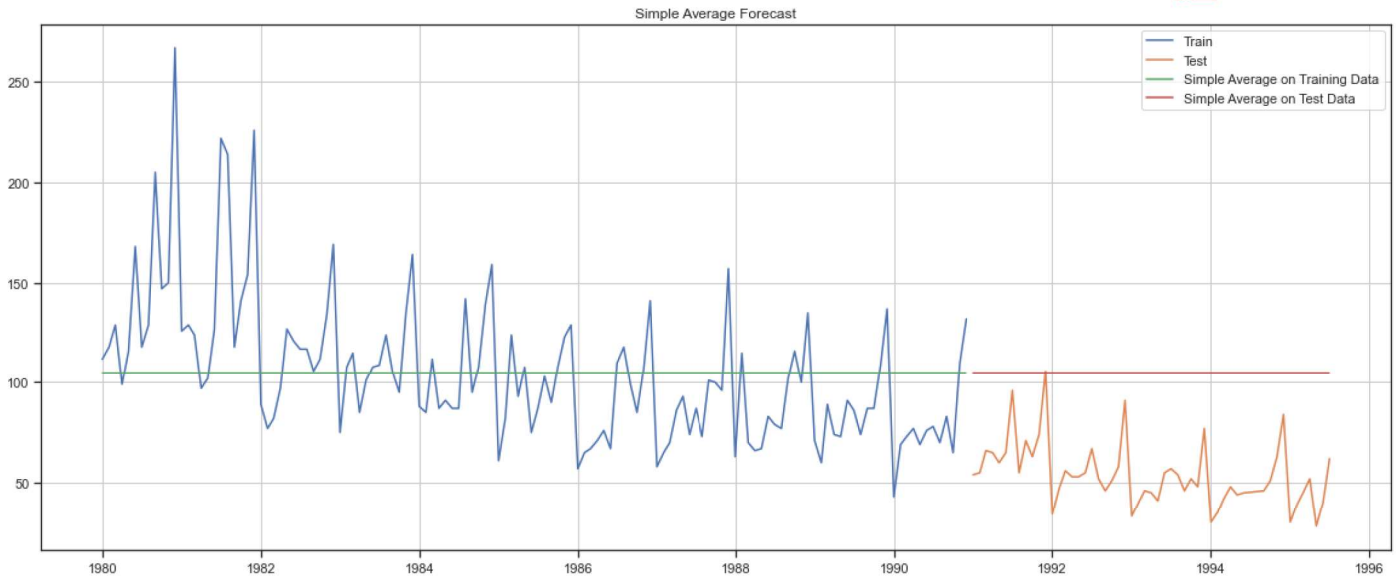
Naive method is best suited for stable datasets. We can still improve our score by adopting different techniques.

Now we will look at another technique and try to improve our prediction accuracy.

1.5.3. Model 3: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

Model 3 Plot - Simple Average



Model 3 Evaluation - Simple Average

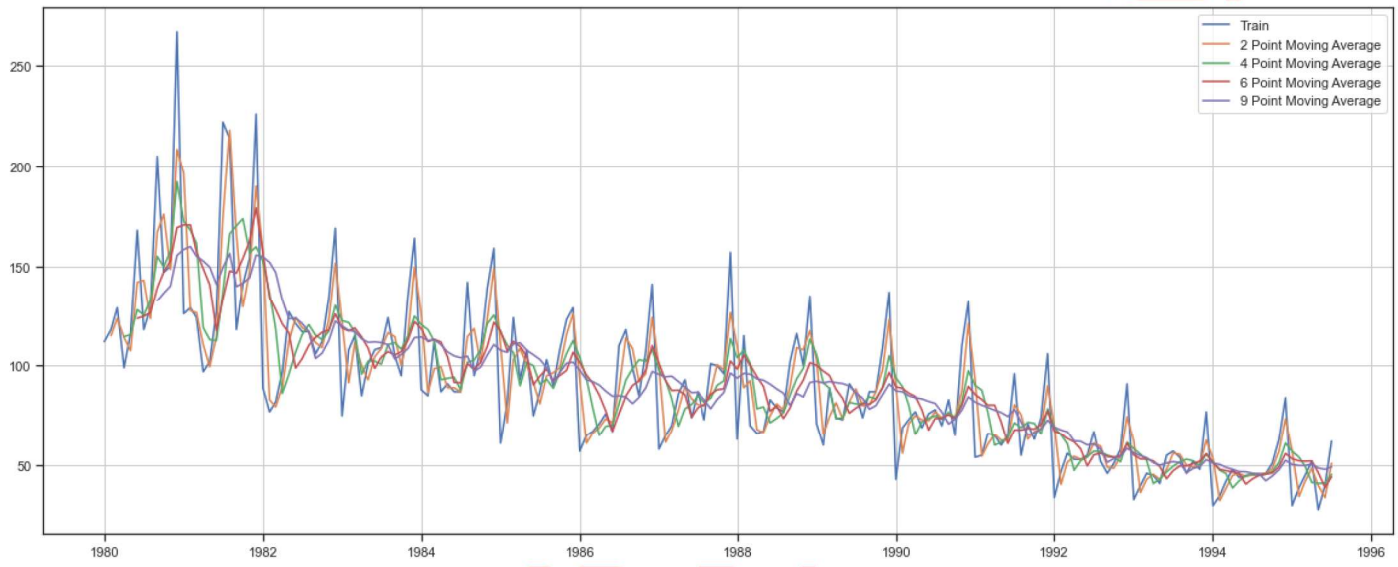
For Simple Average Model forecast on the Training Data, RMSE is 36.034
For Simple Average forecast on the Test Data, RMSE is 53.461

1.5.4. Model 4: 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.

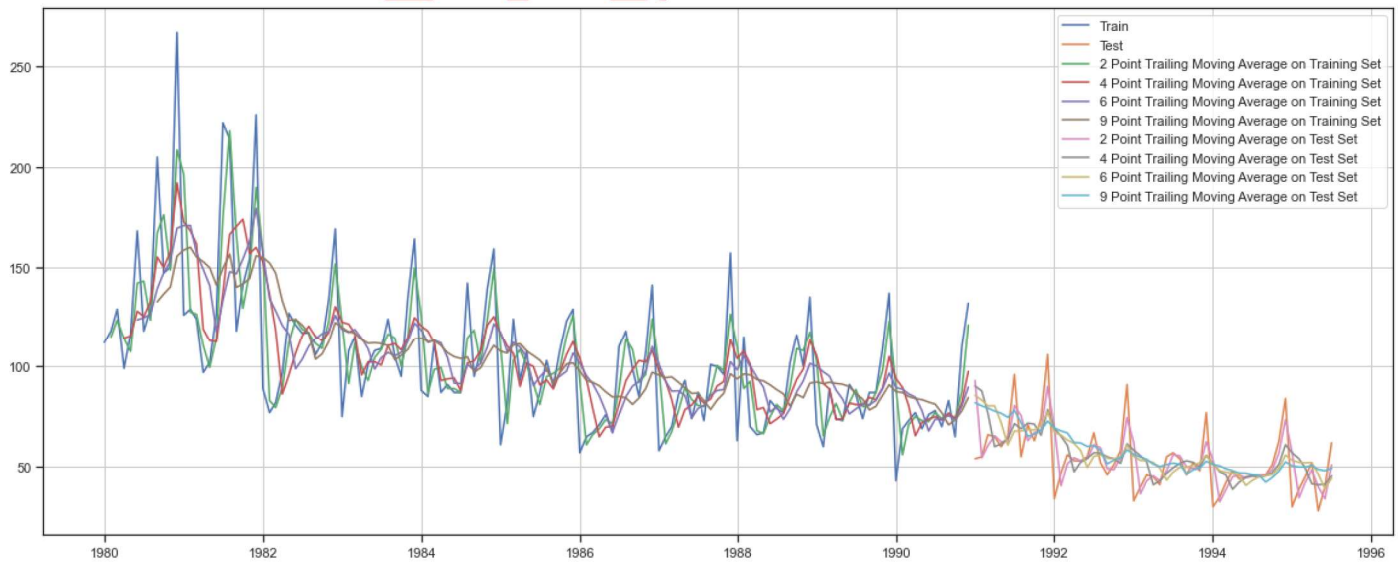
For Moving Average, we are going to average over the entire data.

Model 4 Plot - Moving Average on entire data



For Moving Average, we are going to average over the training data.

Model 4 Plot - Moving Average on Train and test Data separately

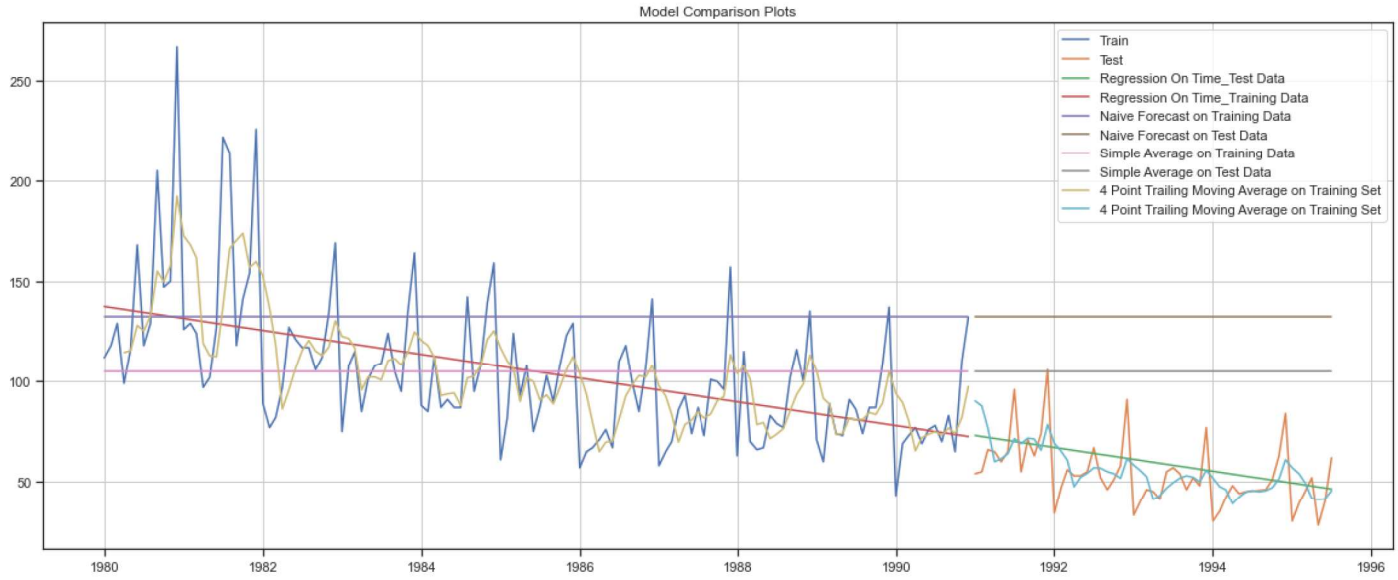


Model 4 Evaluation - Moving Average

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

Let us plot all the models done so far and compare the time Series plots.



1.5.5. Model 5: Double Exponential Smoothing (Holt's Model)

Two parameters Alpha and Beta are estimated in this model. Level and Trend are accounted for in this model.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.0	0.0	476.514666	1023.308161
1	0.0	0.1	476.514666	1023.308161
2	0.0	0.2	476.514666	1023.308161
3	0.0	0.3	476.514666	1023.308161
4	0.0	0.4	476.514666	1023.308161
...	...	...	...	...
116	1.0	0.6	51.831610	801.680218
117	1.0	0.7	54.497039	841.892573
118	1.0	0.8	57.365879	853.965537
119	1.0	0.9	60.474309	834.710935
120	1.0	1.0	63.873454	780.079579

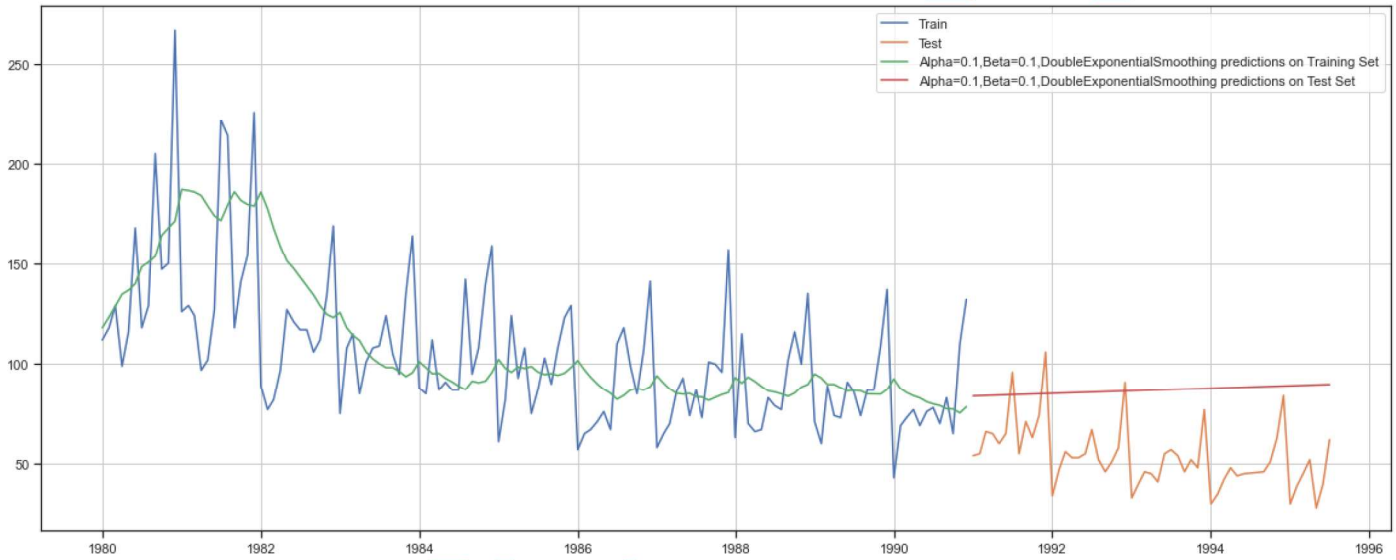
121 rows × 4 columns

Let us sort the data frame in the ascending ordering of the 'Test RMSE'

	Alpha Values	Beta Values	Train RMSE	Test RMSE
12	0.1	0.1	34.439111	36.923416
13	0.1	0.2	33.450729	48.688648
23	0.2	0.1	33.097427	65.731702
14	0.1	0.3	33.145789	78.156641
34	0.3	0.1	33.611269	98.653317

Model 5 Plot – Double Exponential Smoothing

Plotting on both the Training and Test data



Model 5 Evaluation – Double Exponential Smoothing

	Test RMSE
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416

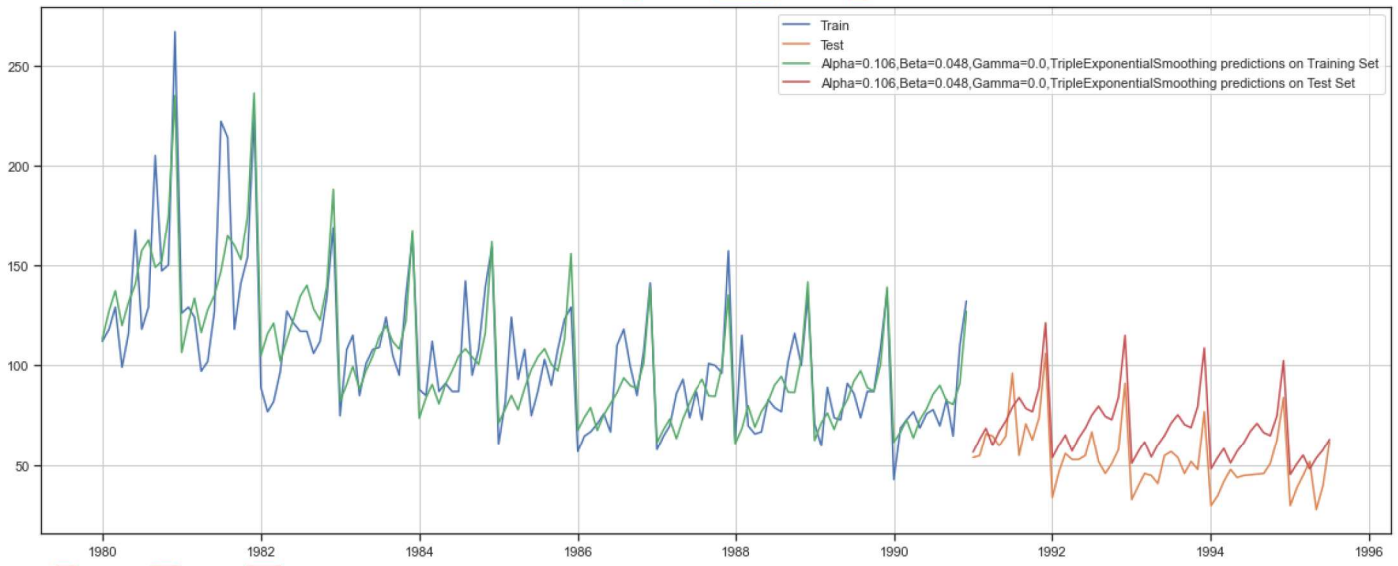
1.5.6. Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters alpha, beta and gamma are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Model 6 - Auto-Fit Parameters

```
{'smoothing_level': 0.10609618064451229,
 'smoothing_slope': 0.04843862120441897,
 'smoothing_seasonal': 0.0,
 'damping_slope': nan,
 'initial_level': 76.65565260938652,
 'initial_slope': 0.0,
 'initial_seasons': array([1.47550324, 1.65927197, 1.805727  , 1.58888878, 1.77822773,
        1.92604444, 2.11649538, 2.25135281, 2.11690672, 2.08112911,
        2.40927381, 3.30448271]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Model 6 Plot – Triple Exponential Smoothing



Model 6 Evaluation – Triple Exponential Smoothing

For Alpha=0.106,Beta=0.048,Gamma=0.0, Triple Exponential Smoothing Model forecast on the Training Data, RMSE is 18.579

For Alpha=0.106,Beta=0.048,Gamma=0.0, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 17.369

1.5.7. Model 7: Brute Force - Triple Exponential Smoothing

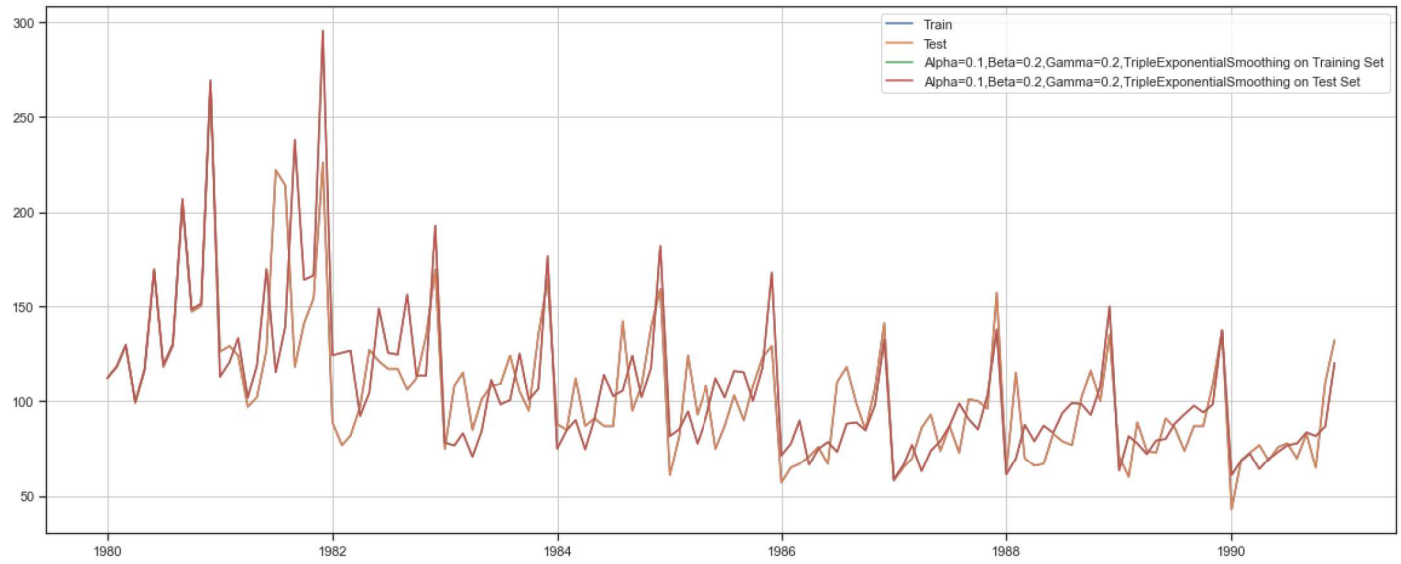
	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.0	0.0	0.0	80.561220	145.018833
1	0.0	0.0	0.1	56.145363	87.888477
2	0.0	0.0	0.2	42.908943	60.882008
3	0.0	0.0	0.3	35.596435	48.171374
4	0.0	0.0	0.4	31.442303	42.054597
...	...	...	...	...	...
1326	1.0	1.0	0.6	28358.458519	9603.635095
1327	1.0	1.0	0.7	30724.126331	23029.955359
1328	1.0	1.0	0.8	1218.755446	9626.710854
1329	1.0	1.0	0.9	14150.253251	9691.905402
1330	1.0	1.0	1.0	1768.254189	8138.618579

1331 rows × 5 columns

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
145	0.1	0.2	0.2	24.365597	9.640687
146	0.1	0.2	0.3	23.969166	9.935740
144	0.1	0.2	0.1	25.529854	9.943539
300	0.2	0.5	0.3	27.631767	10.026210
310	0.2	0.6	0.2	28.289836	10.031639

With Brute force a better RMSE value is found.

Model 7 Plot – Brute Force for Triple Exponential Smoothing

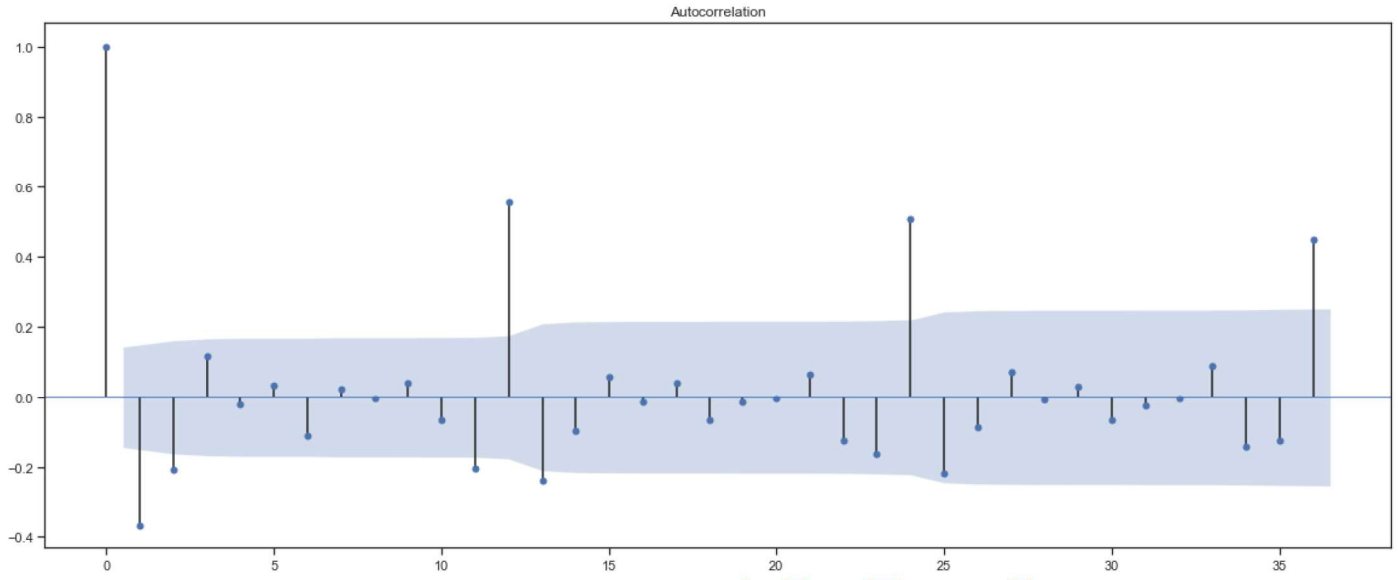


Model 7 Evaluation – Brute Force for Triple Exponential Smoothing

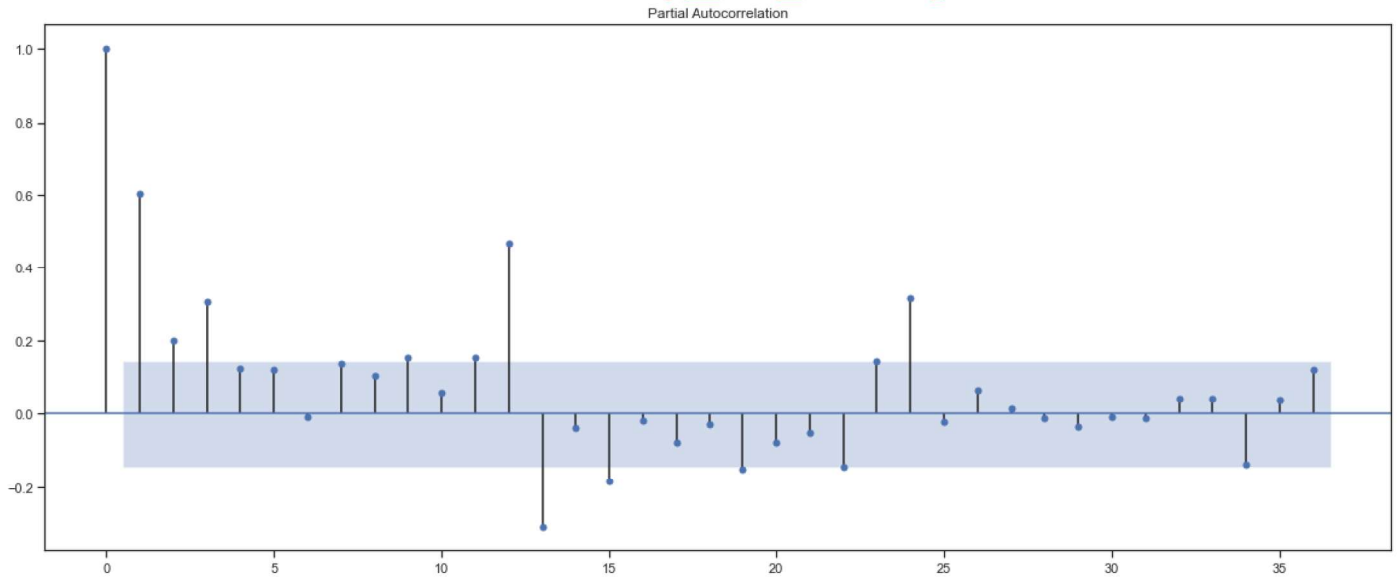
	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,BruteForce Triple Exponential Smoothing	9.640687

1.5.8. ACF/ PCF Plots

ACF Plot



PACF Plot

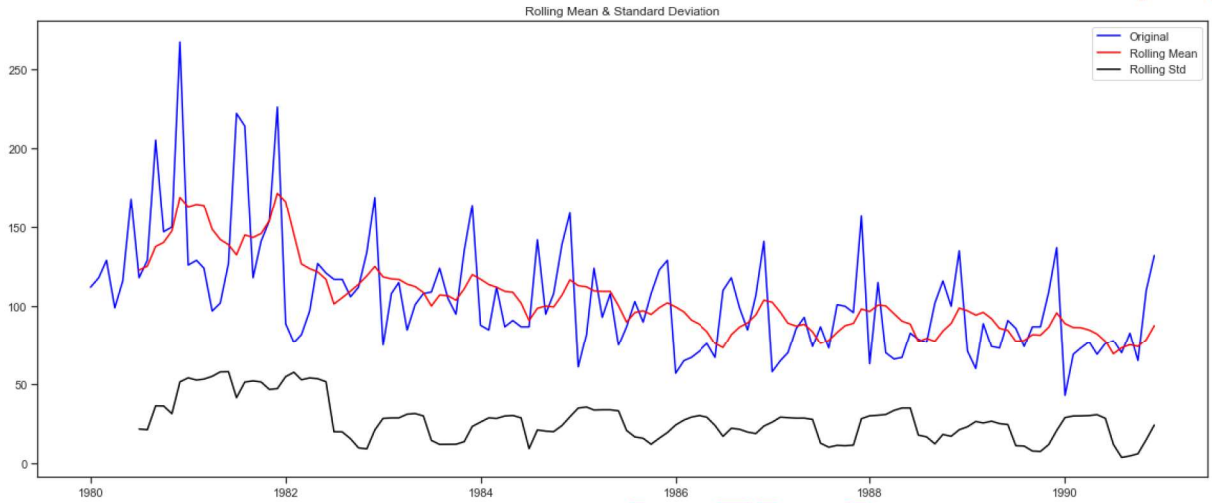


Seasonality after certain lags is visible. Every 12th Month. ACF and PACF plots are done with 95% confidence interval bands.

Test for stationarity of the series - Dicky Fuller test

Null and Alternate Hypothesis for the Augmented Dickey Fuller Test.

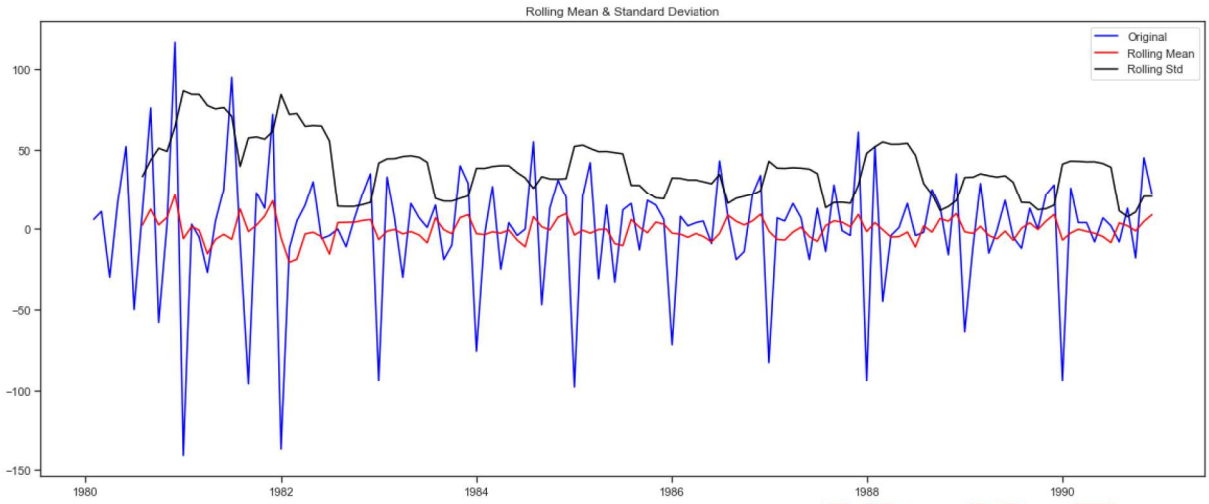
H0: The series is not stationary.
H1: The series is stationary.



Series is not stationary with original form at alpha = 0.05.

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

Let us try check for stationarity after taking first order differencing.



Series is stationary post first order differencing at alpha = 0.05

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

1.5.9. Model 8: ARIMA Model by picking the pdq values from the ACF/ PACF plot

- p = 4
- d = 1
- q = 3

Summary of ARIMA (4, 1, 3) Model

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(4, 1, 3)	Log Likelihood	-633.476
Method:	css-mle	S.D. of innovations	29.584
Date:	Mon, 20 Jul 2020	AIC	1284.953
Time:	14:10:05	BIC	1310.829
Sample:	02-01-1980 - 12-01-1990	HQIC	1295.468

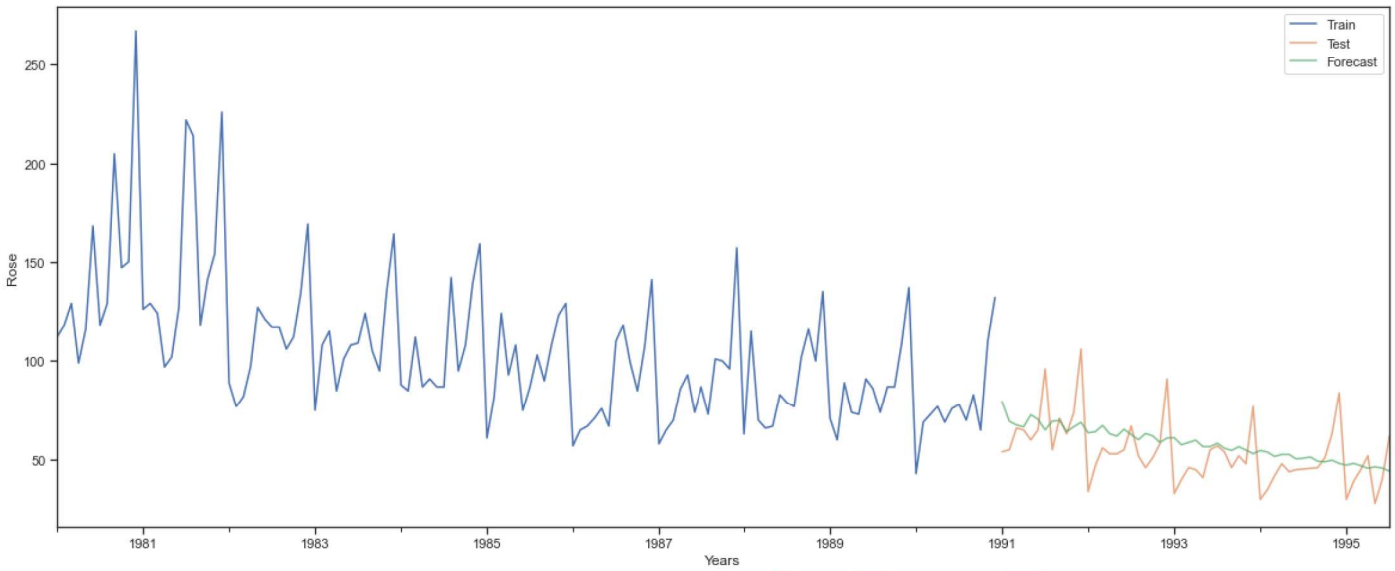
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4922	0.078	-6.343	0.000	-0.644	-0.340
ar.L1.D.Rose	-0.4884	0.092	-5.293	0.000	-0.669	-0.308
ar.L2.D.Rose	-0.8867	0.100	-8.836	0.000	-1.083	-0.690
ar.L3.D.Rose	0.1323	0.102	1.292	0.196	-0.068	0.333
ar.L4.D.Rose	-0.1239	0.094	-1.315	0.189	-0.309	0.061
ma.L1.D.Rose	-0.2908	0.041	-7.164	0.000	-0.370	-0.211
ma.L2.D.Rose	0.2908	nan	nan	nan	nan	nan
ma.L3.D.Rose	-1.0000	nan	nan	nan	nan	nan

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.3928	-0.9570j	1.0345	-0.3120
AR.2	-0.3928	+0.9570j	1.0345	0.3120
AR.3	0.9265	-2.5852j	2.7462	-0.1952
AR.4	0.9265	+2.5852j	2.7462	0.1952
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.3546	-0.9350j	1.0000	-0.3077
MA.3	-0.3546	+0.9350j	1.0000	0.3077

For this particular Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1,2 3 and 4. We are also considering the errors from the auto-regression of the first three lags. The values of p and q are calculated by looking at the ACF and the PACF plots.

Model 8 Plot – ARIMA Model by picking the pdq values from the ACF/ PACF plot



Model 8 Evaluation – ARIMA Model by picking the pdq values from the ACF/ PACF plot

	Test RMSE
ARIMA(4, 1, 3) looking at ACF/PACF	15.337615

1.5.10. Model 9: SARIMA Model by picking the pdq and PDQ values from the ACF/ PACF plot

- p = 4
- d = 1
- q = 3
- P = 3
- D = 0
- Q = 2

Summary of SARIMA(4, 1, 3)(3, 0, 2) 12

SARIMAX Results

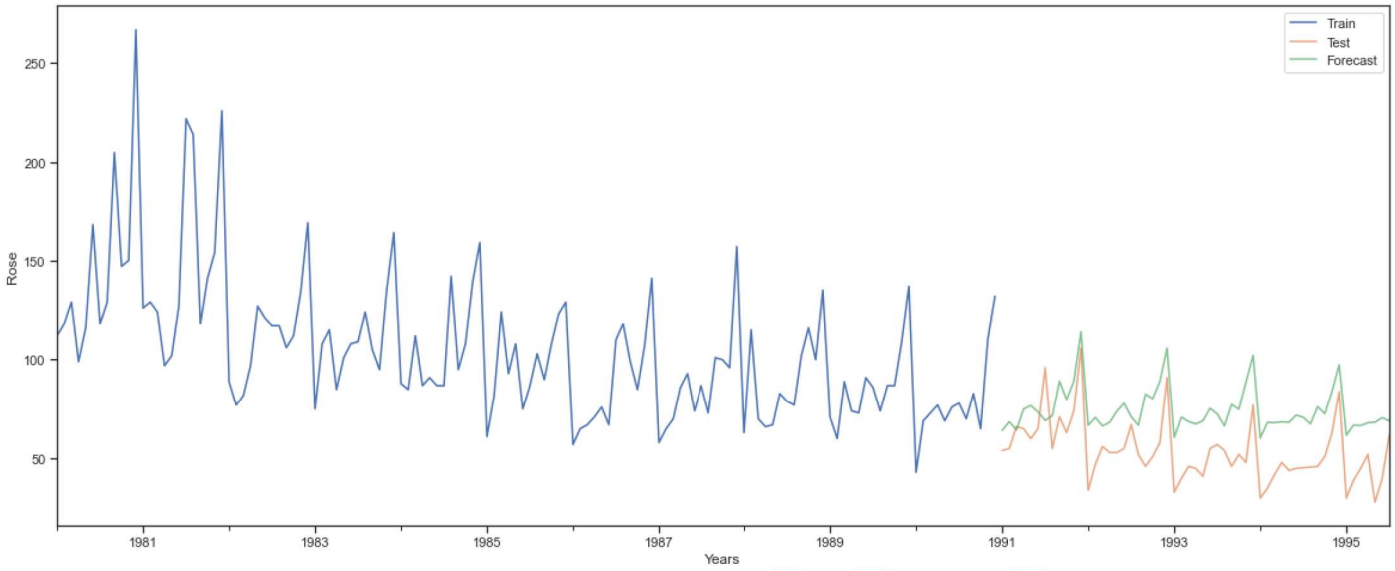
Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(4, 1, 3)x(3, 0, [1, 2], 12)	Log Likelihood	-373.520
Date:	Mon, 20 Jul 2020	AIC	773.041
Time:	14:10:49	BIC	805.682
Sample:	01-01-1980	HQIC	786.210
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1069	0.158	0.678	0.498	-0.202	0.416
ar.L2	-0.8551	0.140	-6.108	0.000	-1.129	-0.581
ar.L3	-0.0886	0.146	-0.606	0.544	-0.375	0.198
ar.L4	0.0257	0.133	0.193	0.847	-0.235	0.287
ma.L1	-1.0421	5.496	-0.190	0.850	-11.813	9.729
ma.L2	1.0392	2.909	0.357	0.721	-4.662	6.740
ma.L3	-0.9977	6.243	-0.160	0.873	-13.233	11.238
ar.S.L12	0.6151	0.209	2.938	0.003	0.205	1.025
ar.S.L24	0.0660	0.138	0.480	0.631	-0.204	0.336
ar.S.L36	0.1316	0.077	1.718	0.086	0.019	0.282
ma.S.L12	-0.2701	0.284	-0.950	0.342	-0.827	0.287
ma.S.L24	-0.0337	0.162	-0.208	0.835	-0.351	0.283
sigma2	191.5897	1197.096	0.160	0.873	-2154.675	2537.854

Ljung-Box (Q):	28.25	Jarque-Bera (JB):	1.73
Prob(Q):	0.92	Prob(JB):	0.42
Heteroskedasticity (H):	0.84	Skew:	0.34
Prob(H) (two-sided):	0.64	Kurtosis:	3.09

For this particular Seasonal Auto-Regressive Integrated Moving Average we are regressing the original series on itself at t he lags of 1,2 3 and 4. We are also considering the errors from the auto-regression of the first three lags. For the seasonal parameters, we are considering the regression of the series on itself one with a lag of three years (or 36 months) and considering the errors from the first 2 auto-regressions. The values of p, q, P and Q are calculated by looking at the ACF and the PACF plots.

Model 9 Plot – SARIMA Model by picking the pdq and PDQ values from the ACF/ PACF plot



Model 9 Evaluation – SARIMA Model by picking the pdq and PDQ values from the ACF/ PACF plot

	Test RMSE
SARIMA(4, 1, 3)(3, 0, 2, 12) looking at ACF/PACF	23.351738

1.5.11. Model 10 Auto ARIMA

Series is not stationary and hence differentiation would be required. For an Auto-ARIMA, we calculate the best p and q p arameters by looking at the lowest corresponding Akaike Information Criterion (AIC) values.

	param	AIC
2	(0, 1, 2)	1276.835374
5	(1, 1, 2)	1277.359236
4	(1, 1, 1)	1277.775755
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

ARIMA(0,1,2) has the lowest AIC

Summary of ARIMA(0, 1, 2)

ARIMA Model Results

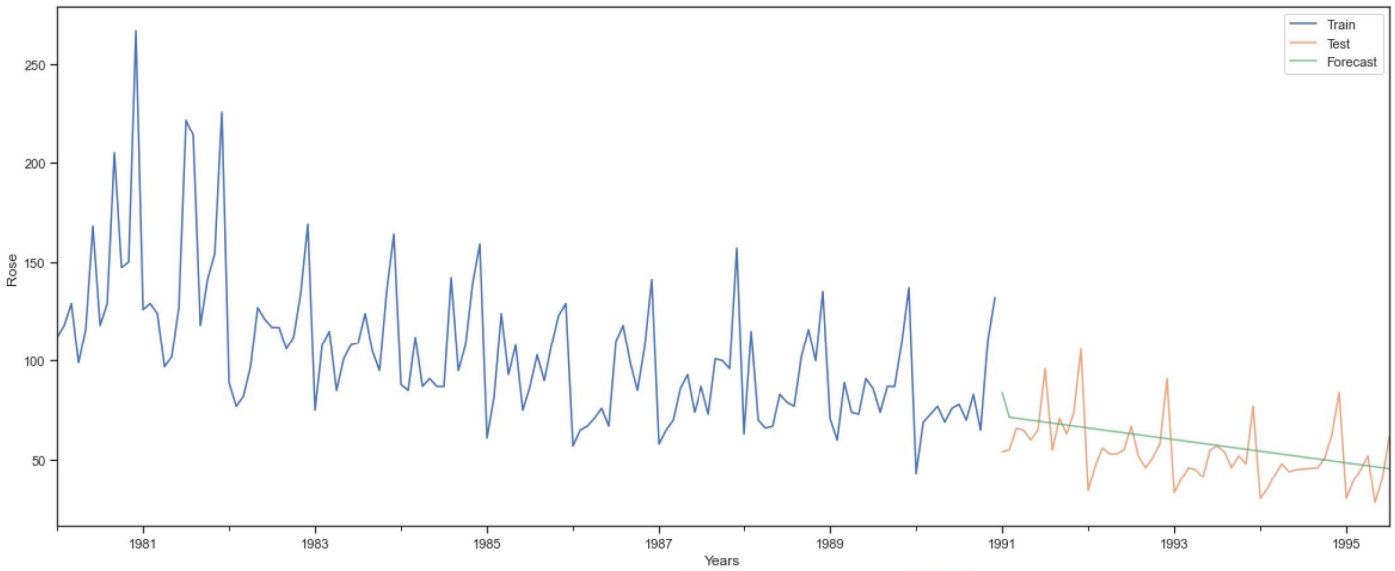
Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418
Method:	css-mle	S.D. of innovations	30.167
Date:	Mon, 20 Jul 2020	AIC	1276.835
Time:	14:11:12	BIC	1288.336
Sample:	02-01-1980 - 12-01-1990	HQIC	1281.509

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4886	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	-2.517	0.012	-0.427	-0.053

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	-4.1697	+0.0000j	4.1697	0.5000

Model 10 Plot – Auto ARIMA



Model 10 Evaluation – Auto ARIMA

	Test RMSE
ARIMA(0, 1, 2) Best AIC	15.618094

1.5.12. Model 11: Auto SARIMA

As the dataset has seasonality.. Let's build the model with SARIMA. For an Auto-SARIMA, the parameters p, q, P and Q are selected based on the lowest Akaike Information Criterion (AIC).

Top 5 best AIC values for Auto SRIMA

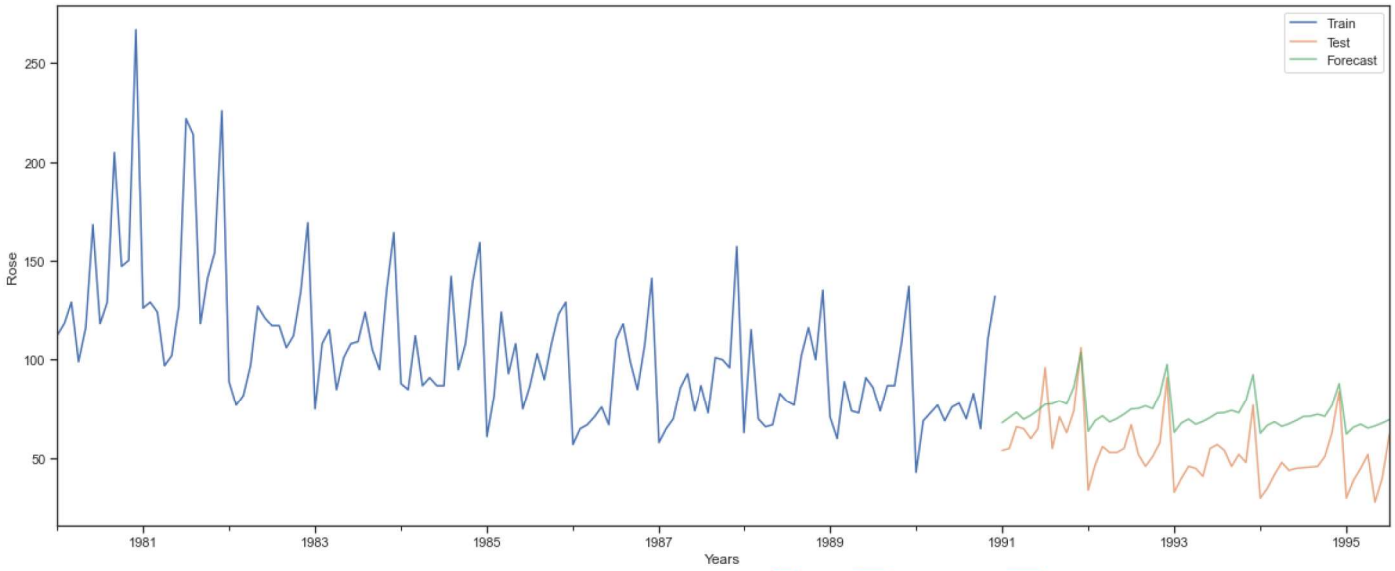
	param	seasonal	AIC
26	(0, 1, 2)	(1, 0, 1, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668849
53	(1, 1, 2)	(2, 0, 2, 12)	896.466818
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346498

Lowest AIC value is for SARIMA(0, 1, 2)(1, 0, 1, 12)

Summary of SARIMA(0, 1, 2)x(1, 0, 1) 12

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(1, 0, [1], 12)   Log Likelihood                -517.770
Date:                              Mon, 20 Jul 2020   AIC                           1045.540
Time:                                      14:12:14   BIC                           1059.308
Sample:                                  01-01-1980   HQIC                          1051.129
                                       - 12-01-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8010      0.128     -6.275      0.000      -1.051      -0.551
ma.L2         -0.2348      0.087     -2.700      0.007      -0.405      -0.064
ar.S.L12       0.8648      0.033     26.202      0.000       0.800       0.929
ma.S.L12      -0.7729      0.156     -4.966      0.000      -1.078      -0.468
sigma2       369.2159     72.145      5.118      0.000     227.814     510.618
===================================================================================
Ljung-Box (Q):                       22.44   Jarque-Bera (JB):                 87.24
Prob(Q):                              0.99   Prob(JB):                          0.00
Heteroskedasticity (H):               0.41   Skew:                              0.32
Prob(H) (two-sided):                  0.01   Kurtosis:                          7.20
===================================================================================
```

Model 11 Plot – Auto SARIMA



Model 11 Evaluation – Auto SARIMA

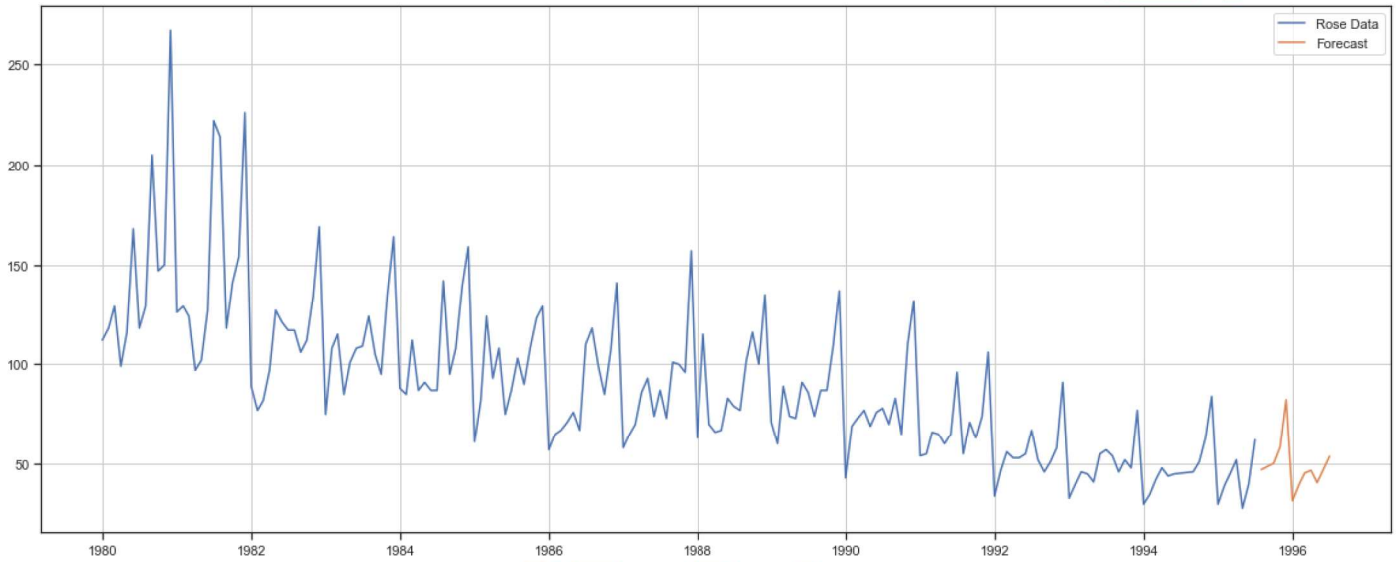
	Test RMSE
SARIMA(0, 1, 2)(1, 0, 1, 12) Best AIC	21.642126

2. Forecast Rose using the best fit BruteForce TripleExponentialSmoothing

Let us apply Brute Force Triple Exponential Smoothing on Full Data

The best found parameters for Triple Exponential Smoothing are Alpha=0.1, Beta=0.2, Gamma=0.2

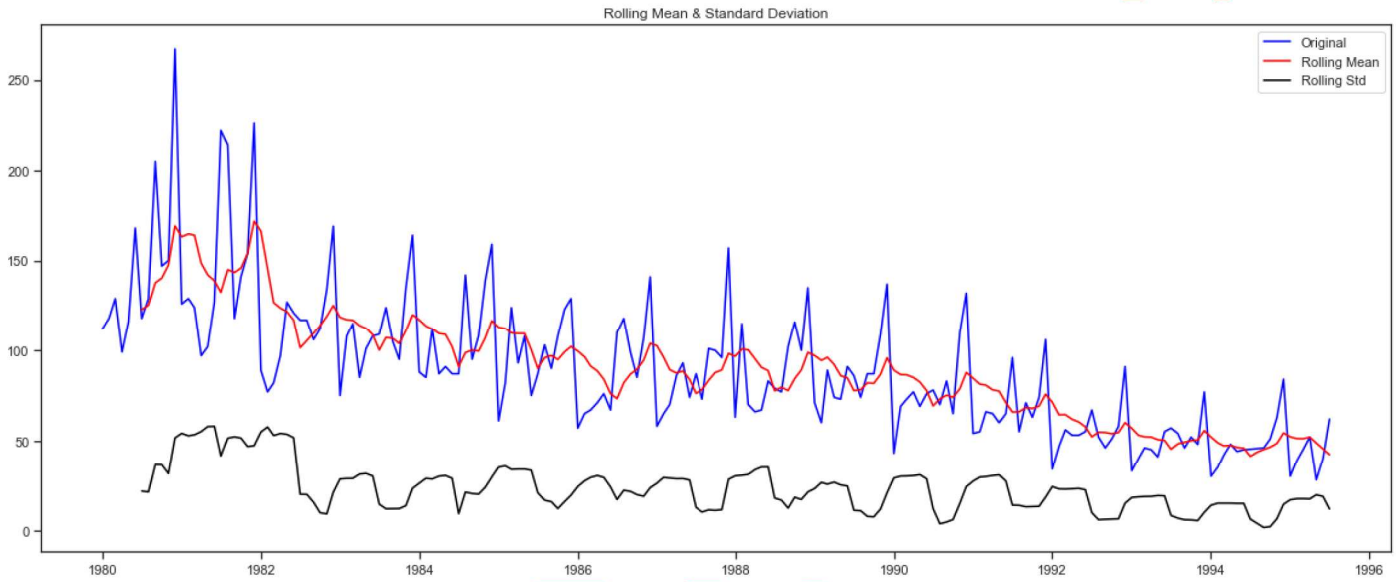
Plot for Best Next 12 month forecast



3. Forecast Rose using the best fit ARIMA model

Once we have found the parameters that best fit to the data, we are going to use these parameter to train of full Rose dataset and then forecast for the 12 months in future.

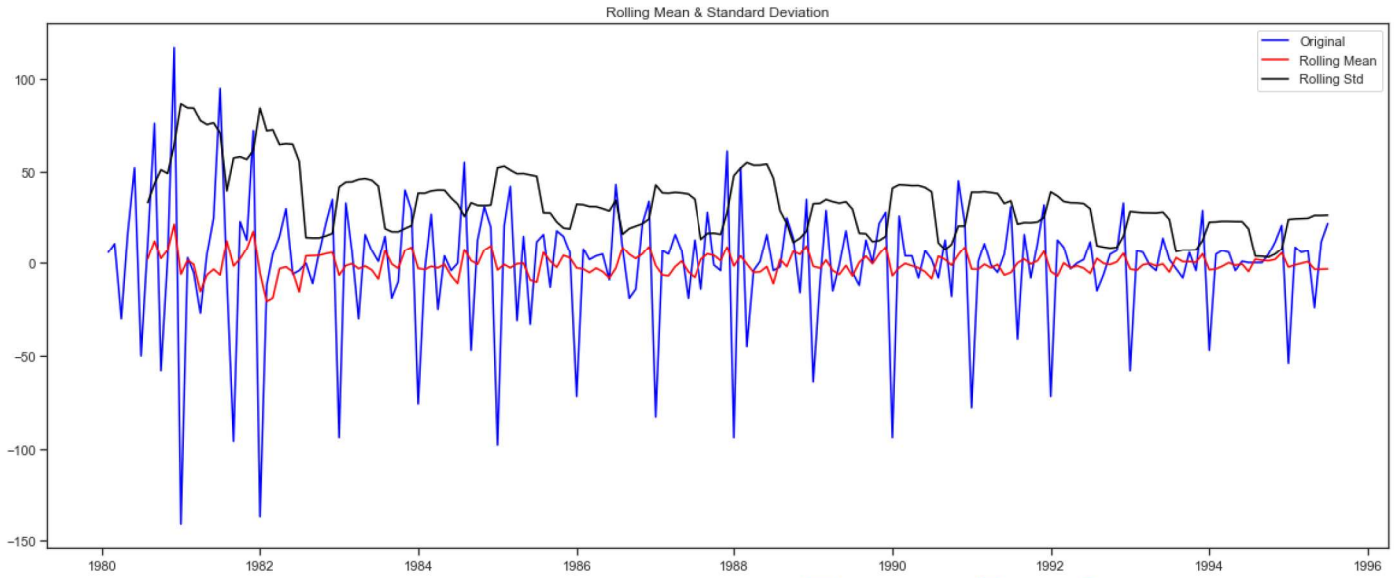
Stationarity test on original data is failed



The whole series is non-stationary at alpha = 0.05 using the Augmented Dickey Fuller test.

```
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
```

Stationarity test post differencing on first order is passed



The whole series is stationary at alpha = 0.05 using the Augmented Dickey Fuller test.

```
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
```

Summary of ARIMA(4, 1, 3) on full data

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	186
Model:	ARIMA(4, 1, 3)	Log Likelihood	-874.742
Method:	css-mle	S.D. of innovations	26.056
Date:	Mon, 20 Jul 2020	AIC	1767.485
Time:	14:12:31	BIC	1796.517
Sample:	02-01-1980 - 07-01-1995	HQIC	1779.250

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5234	0.039	-13.580	0.000	-0.599	-0.448
ar.L1.D.Rose	-0.5108	0.075	-6.841	0.000	-0.657	-0.364
ar.L2.D.Rose	-0.9194	0.083	-11.133	0.000	-1.081	-0.758
ar.L3.D.Rose	0.0901	0.085	1.063	0.288	-0.076	0.256
ar.L4.D.Rose	-0.1389	0.077	-1.806	0.071	-0.290	0.012
ma.L1.D.Rose	-0.2978	0.027	-11.097	0.000	-0.350	-0.245
ma.L2.D.Rose	0.2980	0.029	10.450	0.000	0.242	0.354
ma.L3.D.Rose	-0.9999	0.035	-28.297	0.000	-1.069	-0.931

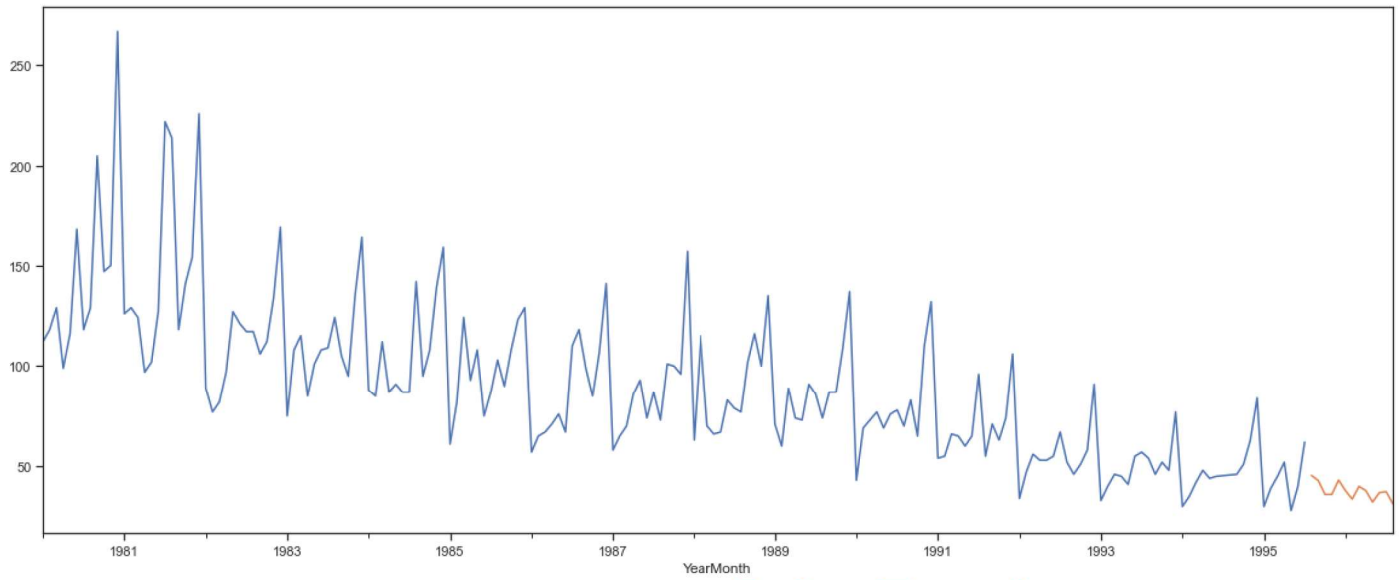
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.3922	-0.9627j	1.0395	-0.3116
AR.2	0.3922	+0.9627j	1.0395	0.3116
AR.3	0.7163	-2.4794j	2.5808	-0.2052
AR.4	0.7163	+2.4794j	2.5808	0.2052
MA.1	1.0001	-0.0000j	1.0001	-0.0000
MA.2	-0.3510	-0.9364j	1.0000	-0.3071
MA.3	-0.3510	+0.9364j	1.0000	0.3071

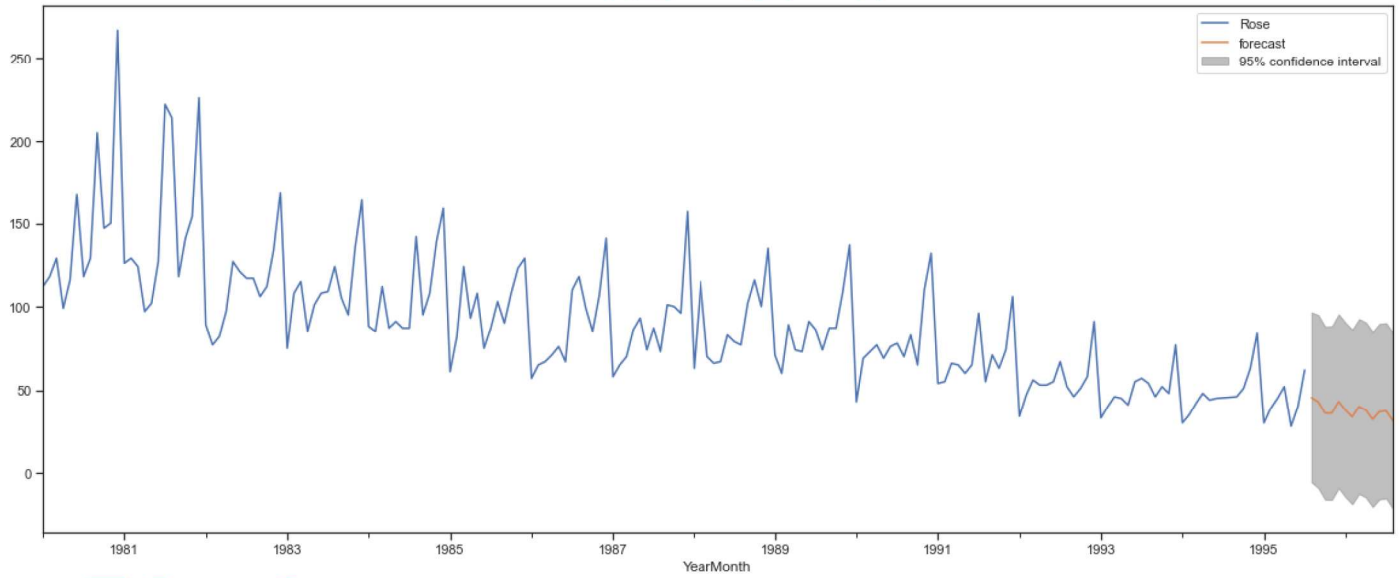
Forecast for next 12 months

1995-08-01	45.438985
1995-09-01	42.912551
1995-10-01	35.973316
1995-11-01	35.995472
1995-12-01	43.139922
1996-01-01	37.898441
1996-02-01	33.675915
1996-03-01	39.994688
1996-04-01	37.886816
1996-05-01	32.204531
1996-06-01	36.903346
1996-07-01	37.362017
1996-08-01	31.291208

Plotting the actual Time Series and the forecasted Time



Plotting the actual Time Series with the confidence interval



4. Inference

- Lowest AIC is obtained using ARIMA(0,1,2), it is as expected as we are using the simple method for ARIMA. Though the Accuracy parameters are not satisfactory.

- Best Accuracy (RMSE) is observed by following Brute Force method on Triple Exponential Smoothing. Best parameter for the same are
  - Alpha=0.1
  - Beta=0.2
  - Gamma=0.2

- ARIMA model build using the p,d,q value looking at the ACF/ PACF plot is second best and the model is ARIMA(4, 1, 3)

For doing the best forecast for the Rose dataset we will choose the Triple Exponential Smoothing.

5. Appendix - Summary of Testing RMSE

	Test RMSE
RegressionOnTime	15.611866
NaiveModel	79.718773
SimpleAverageModel	53.460570
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
Alpha=0.106,Beta=0.048,Gamma=0.0,TripleExponentialSmoothing	17.369489
Alpha=0.1,Beta=0.2,Gamma=0.2,BruteForce TripleExponentialSmoothing	9.640687
ARIMA(4, 1, 3) looking at ACF/PACF	15.337615
SARIMA(4, 1, 3)(3, 0, 2, 12) looking at ACF/PACF	23.351738
ARIMA(0, 1, 2) Best AIC	15.618094
SARIMA(0, 1, 2)(1, 0, 1, 12) Best AIC	21.642126