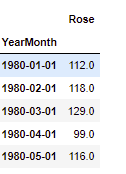
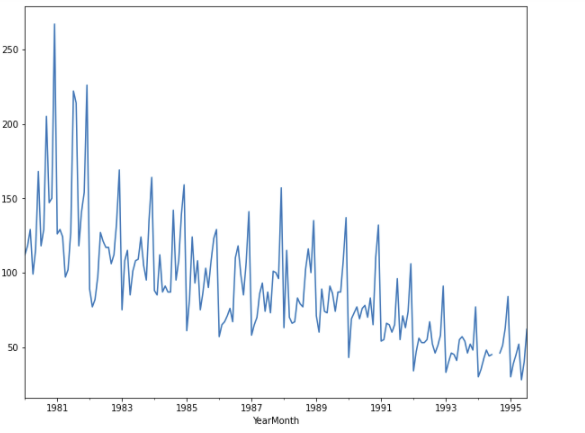
ROSE DATASET

Head

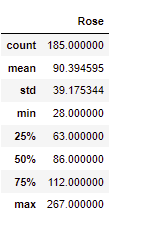


Plot



There are some irregularities at the peak of some year where we have multiple peaks but someplace we have a single peak. Also we can see a downward trend in the dataset which can be worrying for the client.

Describe

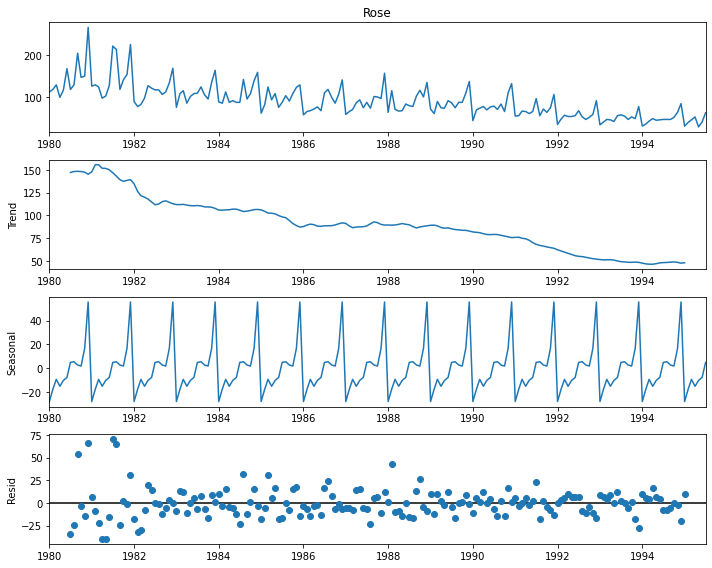


Two null entries in the dataset.

Replaced by bfill.

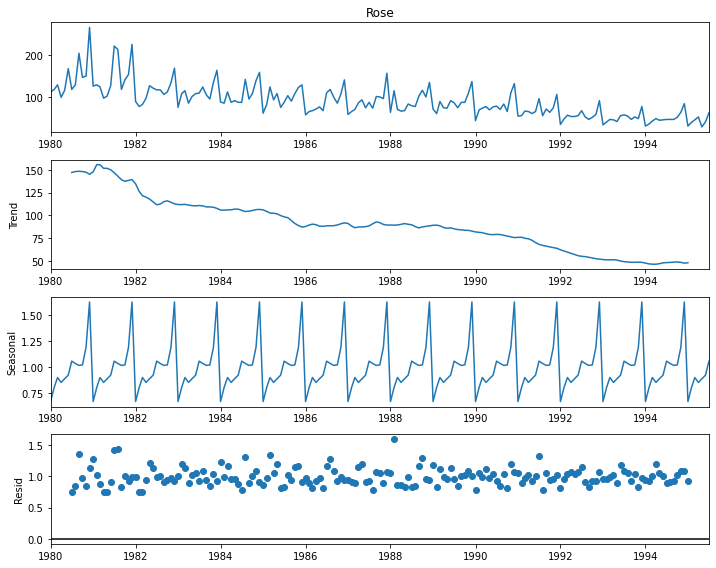
Decompose

Additive



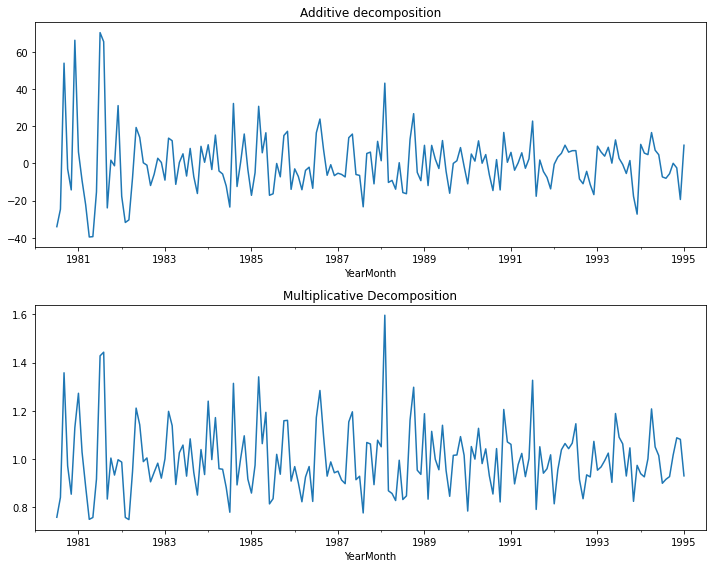
The residual values are not constant and we can observe the trend to be decreasing. The data has yearly seasonality. And the residual values are very spread out initially, and later converges which means, the decomposition for later years is pretty accurate.

Multiplicative



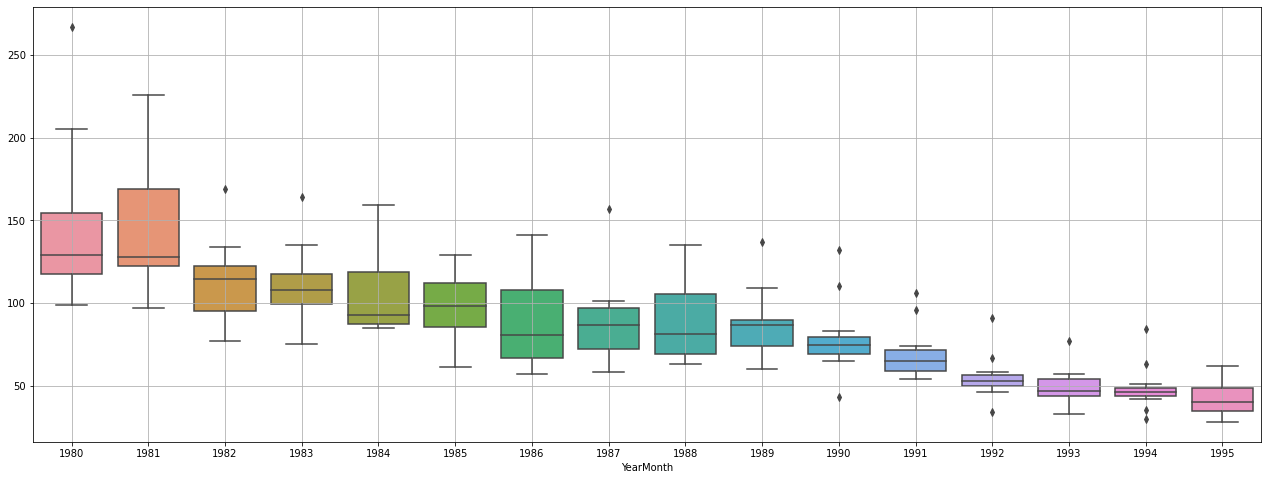
We can see that in comparison to residual of additive we have error values close by, this may be due to the fact that, here the values are multiplied instead of adding them. Also same as observed in the additive, the trend is decreasing and the seasonality is yearly.

Side – by – side comparison of residuals



The residual graph of additive decomposition has more aggressive peaks as compared to multiplicative decomposition.

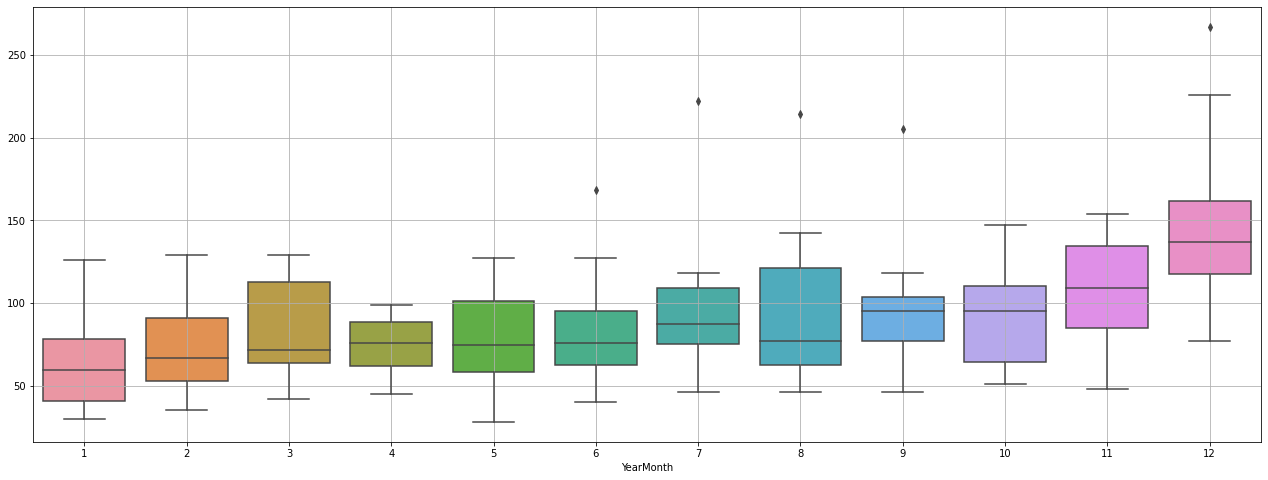
Yearly box plot



We can see outliers in the year 80,82,83,87,89,90,92,91,93,94.

We can see the decreasing size of the boxplot which is a redflag for the client.

Monthly box plot



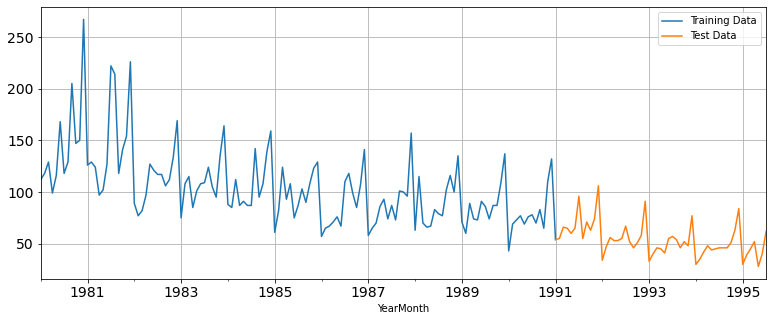
We can observe outliers in the month of June , July, august, September and December.

We have high outliers in months of July august as well as wide whisker length.

--------------------------------------------------------------------------------------------------------------------------------------data was split in from the year 1991.

Train - (133, 3)

Test - (55, 3)



**Linear regression**

RMSE = 14.935

**Simple exponential smoothing**

Autofit params:

{'smoothing\_level': 0.09764588980510863,

'smoothing\_trend': nan,

'smoothing\_seasonal': nan,

'damping\_trend': nan,

'initial\_level': 134.41619408114488,

'initial\_trend': nan,

'initial\_seasons': array([], dtype=float64),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Alpha = 0.097

RMSE is 33.859

**Double exponential smoothing**

**Auto fit param:**

{'smoothing\_level': 9.098504051752187e-07,

'smoothing\_trend': 1.8412431182457466e-08,

'smoothing\_seasonal': nan,

'damping\_trend': nan,

'initial\_level': 138.08633631739735,

'initial\_trend': -0.5003917258354766,

'initial\_seasons': array([], dtype=float64),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Alpha = 9.098

Beta = 1.841

RMSE is 14.937

**Triple exponential smoothing**

Autofit params:

{'smoothing\_level': 0.075736443908668,

'smoothing\_trend': 0.054095230994916865,

'smoothing\_seasonal': 0.41076802008127705,

'damping\_trend': nan,

'initial\_level': 74.76302257024037,

'initial\_trend': 1.0063884869742872,

'initial\_seasons': array([1.64424724, 1.71000839, 1.80391887, 1.56482856, 1.74024994,

2.24741506, 2.19948312, 2.28667059, 2.74243282, 2.03538837,

2.19031921, 3.57179474]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

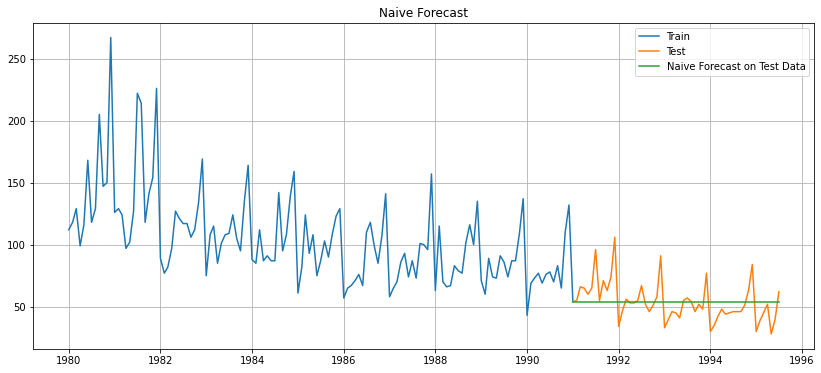
Alpha = 0.076

Beta = 0.054

Gamma = 0.410

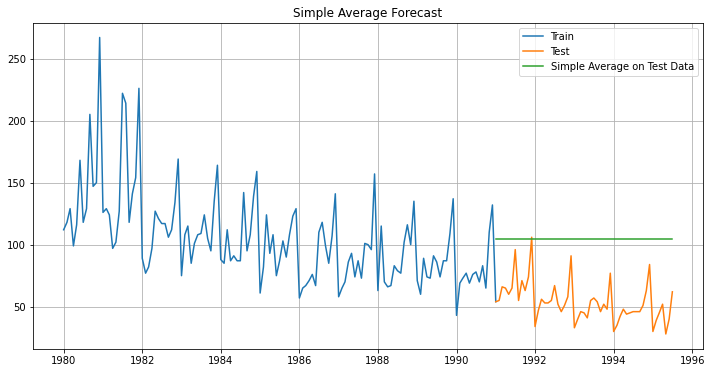
RMSE is 27.814

**Naïve model**



RMSE is 15.751

**Simple average**

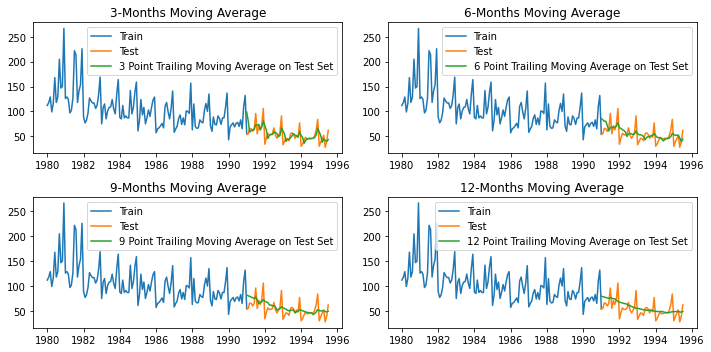
****

RMSE is 53.075

**Moving average**



No of trailing – 1 is the number of na values at the beginning.



Three trailing is giving us the closest result.

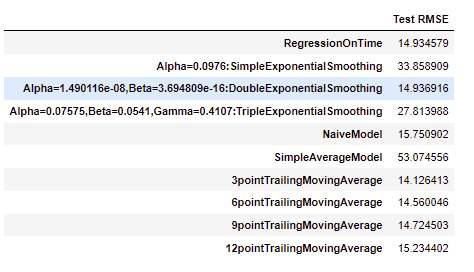
For 3 point Moving Average Model forecast on the Training Data, RMSE is 14.126

For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.560

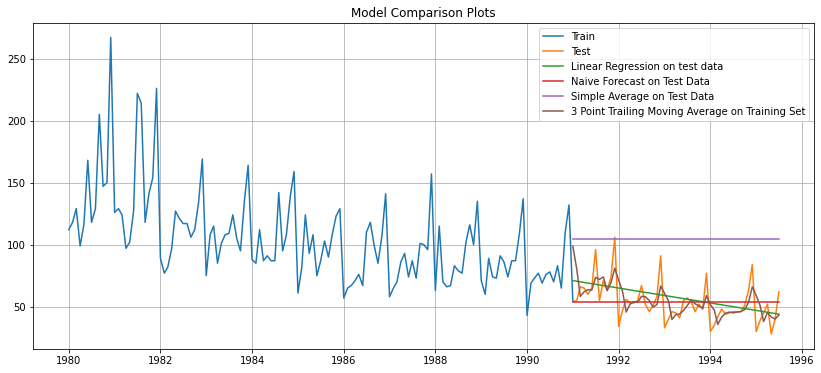
For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.725

For 12 point Moving Average Model forecast on the Training Data, RMSE is 15.234

Table:



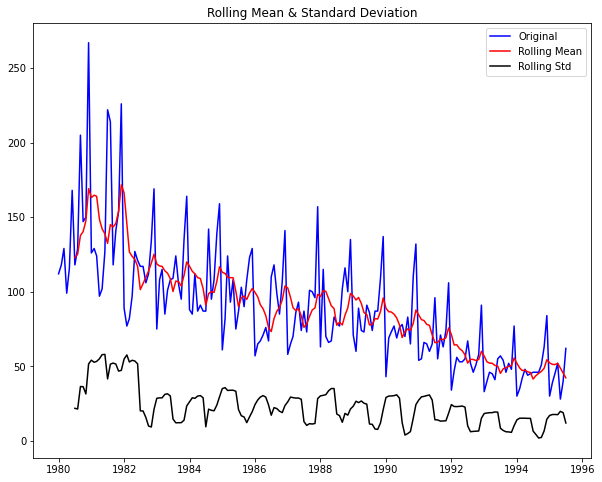
Model comparison plot



We can see the best model is three-point trailing moving average.

Arima

We first need to perform the test for stationarity.

Normal dataset

Results of Dickey-Fuller Test:

Test Statistic -1.877440

p-value 0.342747

#Lags Used 13.000000

Number of Observations Used 173.000000

Critical Value (1%) -3.468726

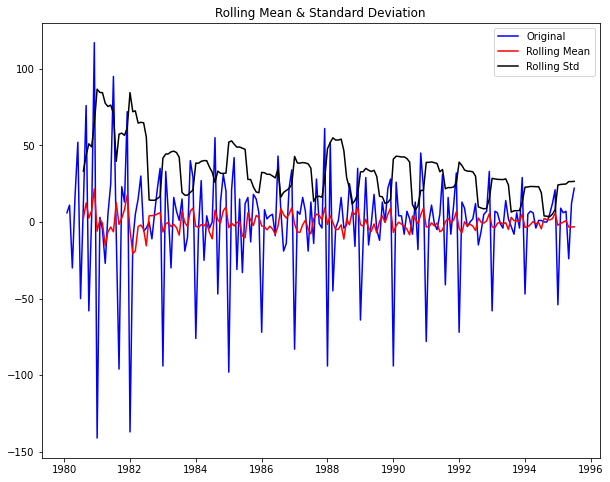
Critical Value (5%) -2.878396

Critical Value (10%) -2.575756

dtype: float64

We can see that here it is not stationary. So, we need to apply diff.

Diff dataset



Results of Dickey-Fuller Test:

Test Statistic -8.044614e+00

p-value 1.808550e-12

#Lags Used 1.200000e+01

Number of Observations Used 1.730000e+02

Critical Value (1%) -3.468726e+00

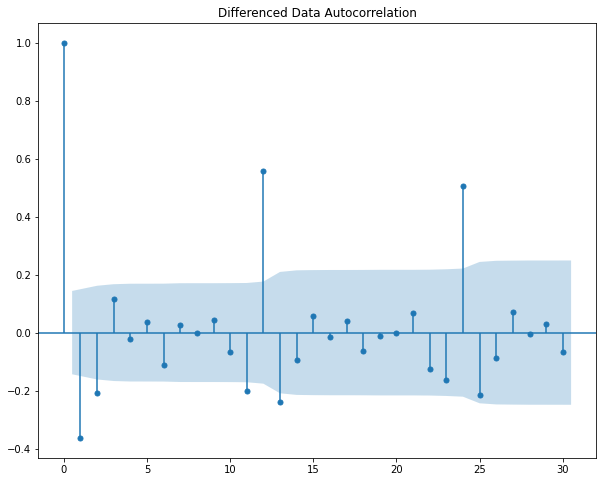
Critical Value (5%) -2.878396e+00

Critical Value (10%) -2.575756e+00

dtype: float64

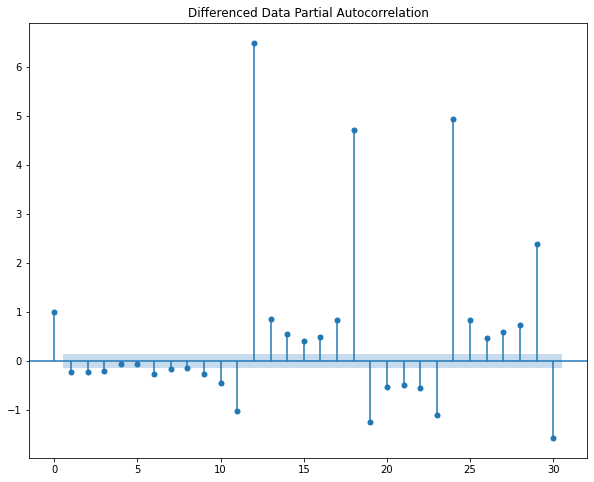
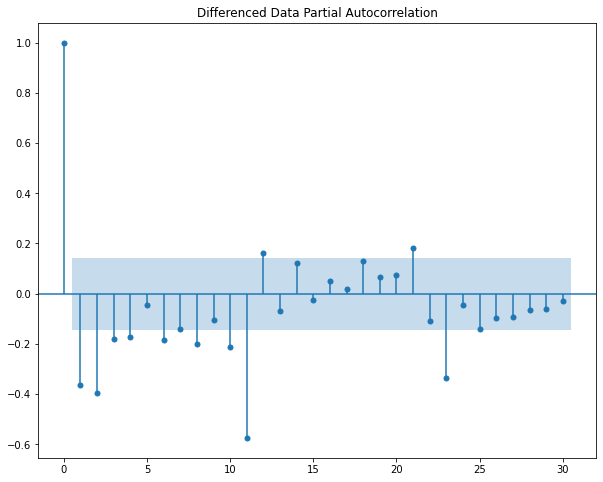
This is fine as the value of p is under the desired 5% .

ACF



We can see a patten at 12 years and also the peak is decreasing as the previous value has more autocorrelation.

PACF



We can analyse the seasonality in the dataset.

Splitting the dataset

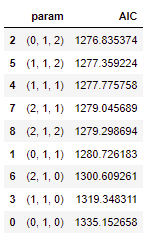
train = df[df.index < '1991-01-01']

test = df[df.index >= '1991-01-01']

Train - (132, 3)

Test - (55, 3)

ARIMA model is run using range of values of p, q and d. Trying to find the param with the least AIC value.



As we can see, p =0 , d = 1, q = 2.

Applying this to ARIMA and testing on the test data, obtaining the result.

RMSE – 15.611

AIC – 1276.835

We will now try to improve on this.

Automated SARIMA MODEL.

Same parameter was applied to SARIMA for a period of 12.

Sorted below:

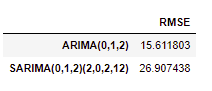


Now, we apply the values of the top parameter to the model and train it. Then after predicting on the test data. ( alpha = 0.05)

Rmse - 26.907

AIC - 887.938

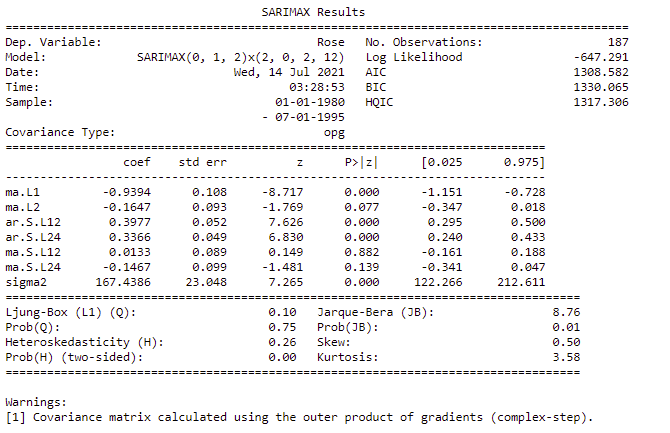
So, this is better than the previous arima model in terms of AIC but RMSE is a bit more but the change is AIC is more inclining. So, in general this model is better than prev one.



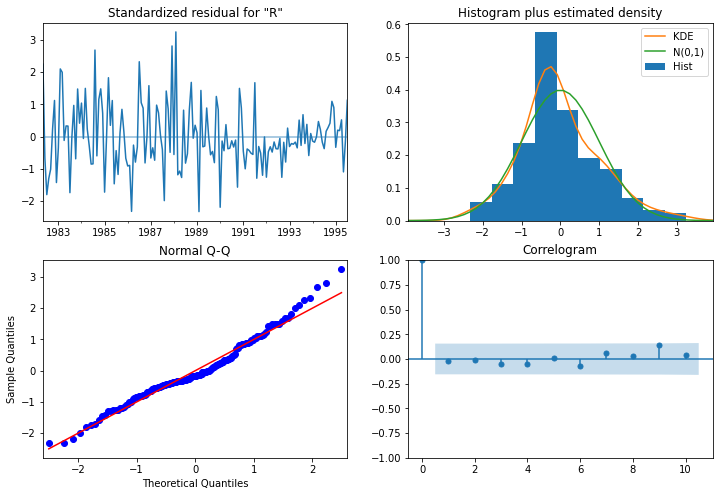
Order = (0,1,2)

seasonal\_order = (2, 0, 2, 12)

FULL DATA MODEL



Diagnostic of our model.



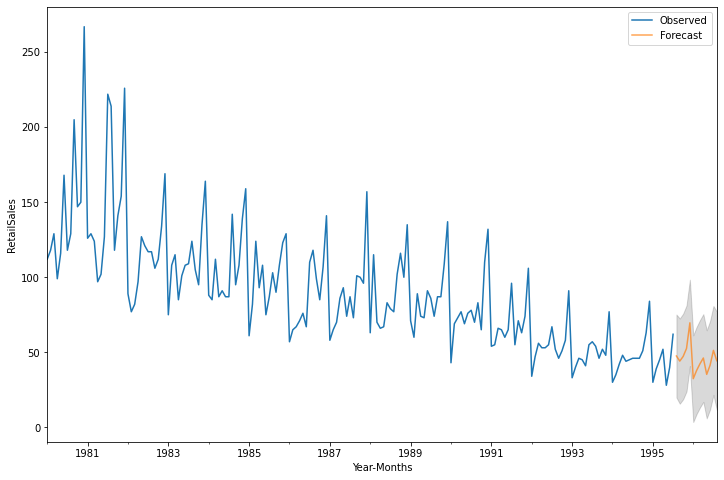
TOP LEFT: The residual error seems to fluctuate around a mean zero and has lower variance in the latest years.

TOP RIGHT: The density plot suggests normal distribution with mean zero.

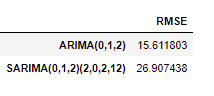
BOTTOM LEFT: The dots fall on the red line indicating normal distribution.

BOTTOM RIGHT: The correlogram, aka ACF plot shows the residual errors are not auto correlated. Any autocorrelation would imply that there is some pattern in the residual errors which are not explained in the model.

It’s overall a good fit. We are good to forecast.



We can see are forecast looks pretty accurate. Following the trend and seasonality.



SUGGESTIONS:

* We can definitely increase our sales in the popular month by using discount coupons.
* We can also use techniques like up selling and cross selling to increase our sales.
* We know the fact, that our business is declining at a steady rate, we need to inject sales with the help of marketing or realise if there is a problem with the product itself.
* We also, need to ask why some months are doing well and try to replicate the conditions in different months.