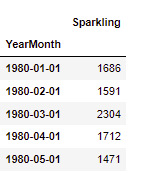
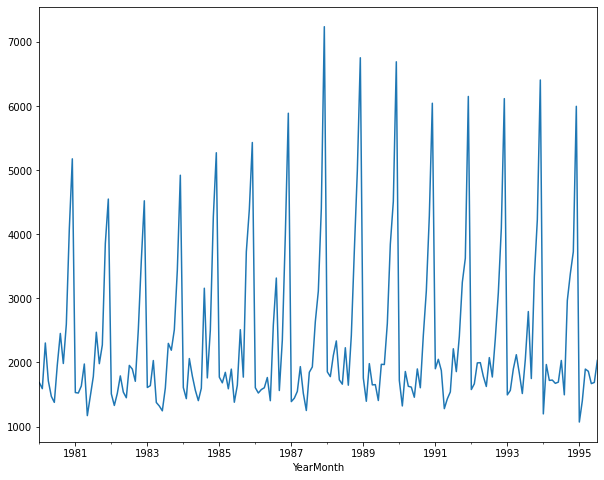
SPARKLING DATASET

Head

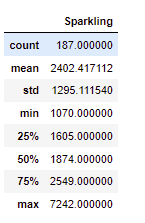


Plot



There are some irregularities at the peak of some year where we have multiple peaks but someplace we have a single peak.

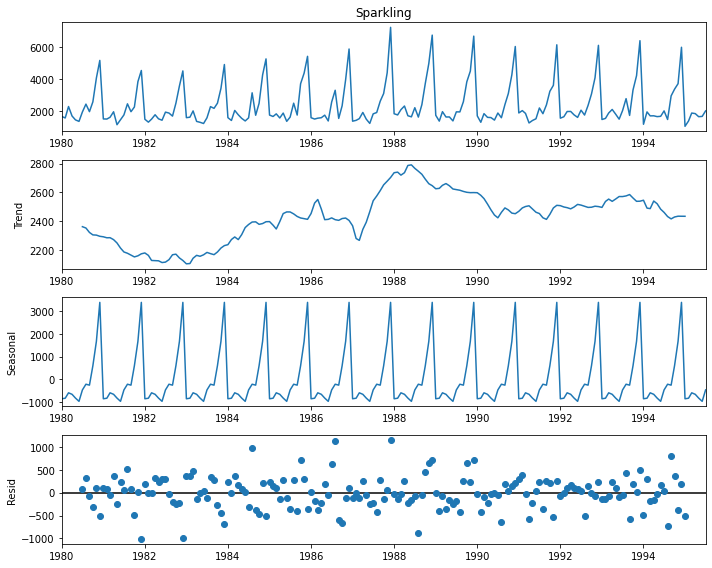
Describe



No of null entries are zero.

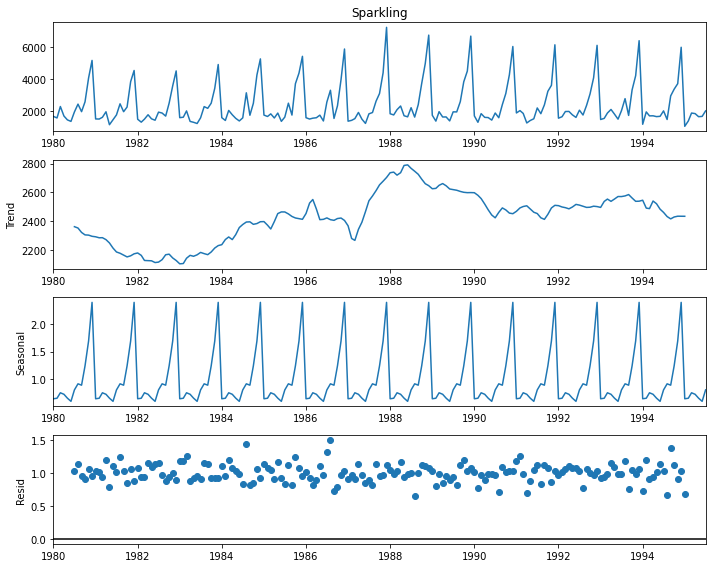
Decompose

Additive



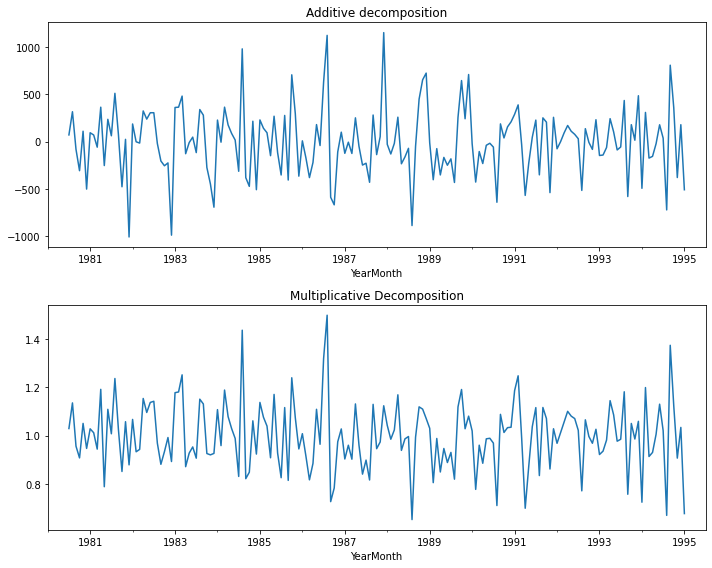
The residual values are not constant and we can observe the trend to be increasing overall. The data has yearly seasonality.

Multiplicative



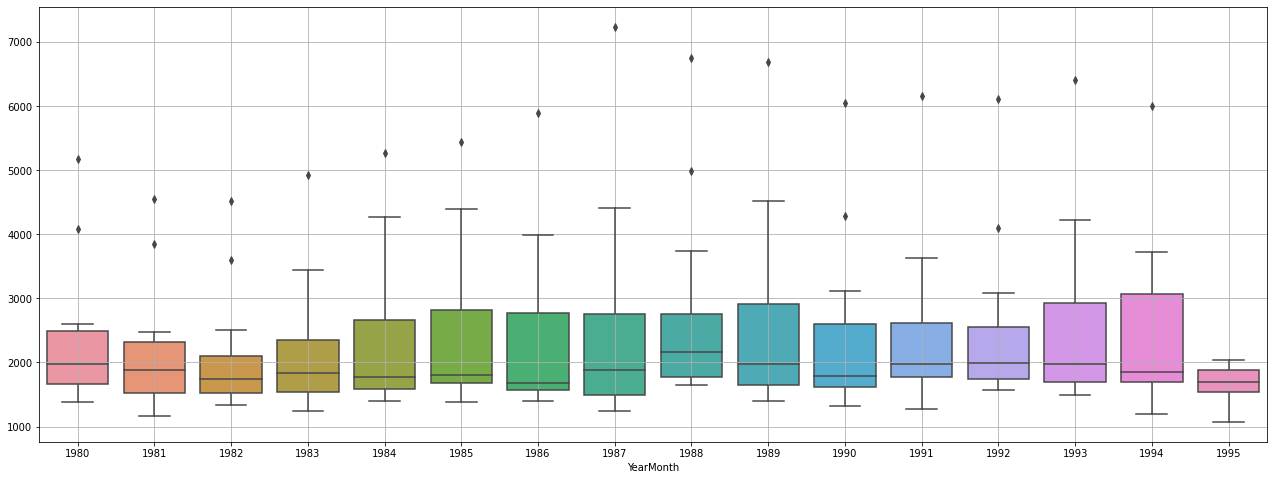
We can see that in comparison to residual of additive we have error values close by, this may be due to the fact that, here the values are multiplied instead of adding them.

Side – by – side comparison of residuals



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

Yearly box plot

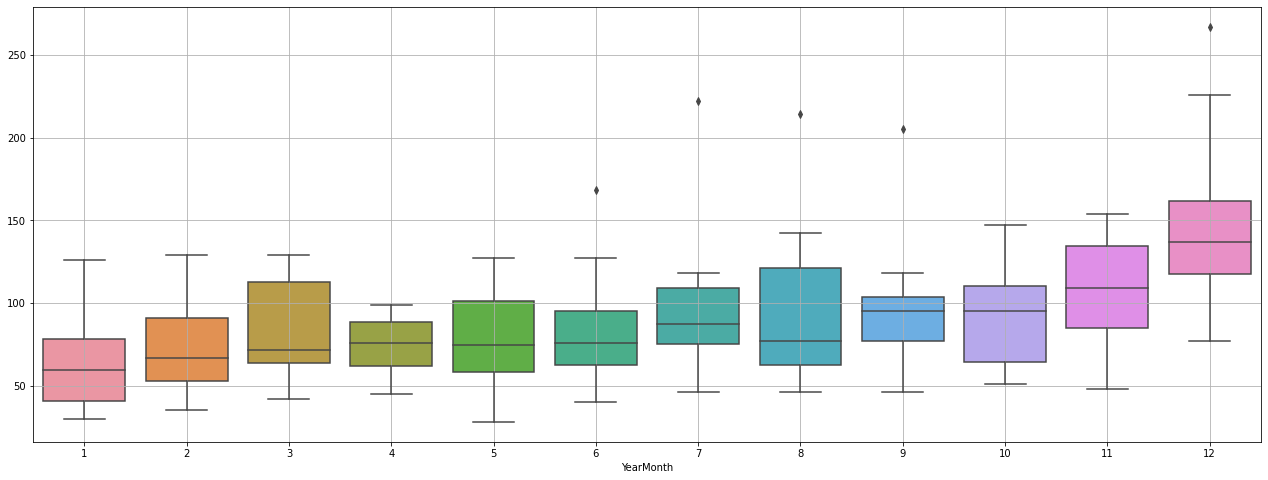


We can see outliers in the year 80,82,83,87,89,90,92,91,93,94.

The interesting thing to notice is that we have lower outliers in two years only i.e.,90 and 94 which is very strange.

Also we can see particular years to have increased distribution for e.g.: 80,81,86,88.

Monthly box plot



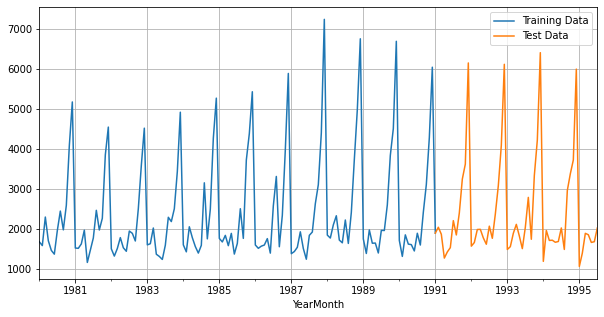
We can observe outliers in the month of June , July, august, September and December.

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

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

Train - (133, 3)

Test - (55, 3)



**Linear regression**

RMSE = 1377.73

**Simple exponential smoothing**

Autofit params:

{'smoothing\_level': 0.055569495557200844,

'smoothing\_trend': nan,

'smoothing\_seasonal': nan,

'damping\_trend': nan,

'initial\_level': 1896.819818417072,

'initial\_trend': nan,

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

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Alpha = 0.555

RMSE is 1310.611

**Double exponential smoothing**

Autofit params:

{'smoothing\_level': 0.6178571428571428,

'smoothing\_trend': 0.0001,

'smoothing\_seasonal': nan,

'damping\_trend': nan,

'initial\_level': 1686.0,

'initial\_trend': -95.0,

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

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Alpha = 0.617

Beta = 0.0001

RMSE is 2773.916

**Triple exponential smoothing**

Autofit params:

{'smoothing\_level': 0.07567236307199589,

'smoothing\_trend': 0.064826110905143,

'smoothing\_seasonal': 0.479015614919629,

'damping\_trend': nan,

'initial\_level': 1661.9198195124713,

'initial\_trend': -5.675763630399794,

'initial\_seasons': array([ 24.1196182 , -70.84039057, 641.77547213, 50.38585741,

-190.79735505, -284.6450009 , 304.19074442, 791.25633558,

322.29425114, 934.22492376, 2425.12210891, 3517.00434923]),

'use\_boxcox': False,

'lamda': None,

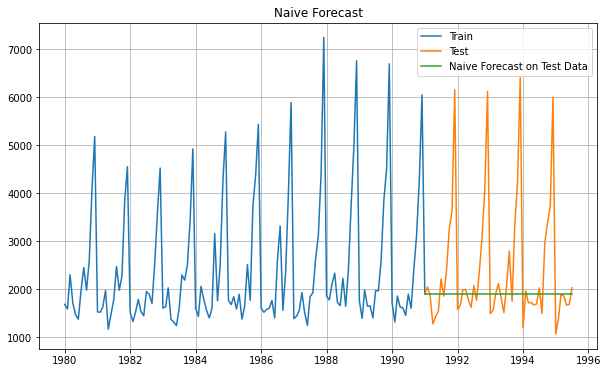
'remove\_bias': False}

Alpha = 0.055

Beta = 0.065

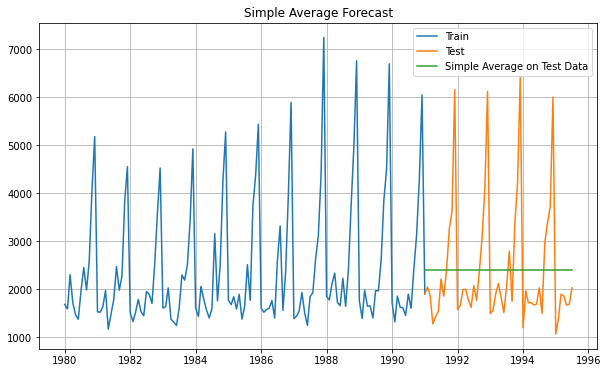
Gamma = 0.478

**Naïve model**



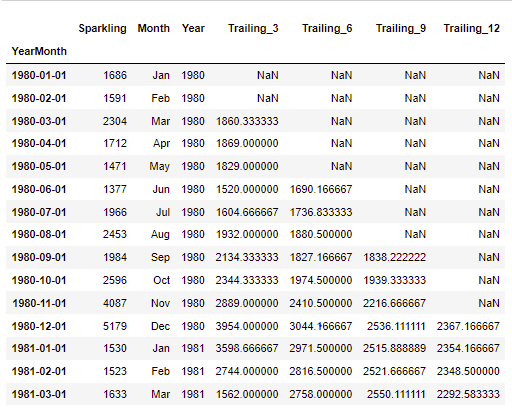
RMSE is 1368.563

**Simple average**

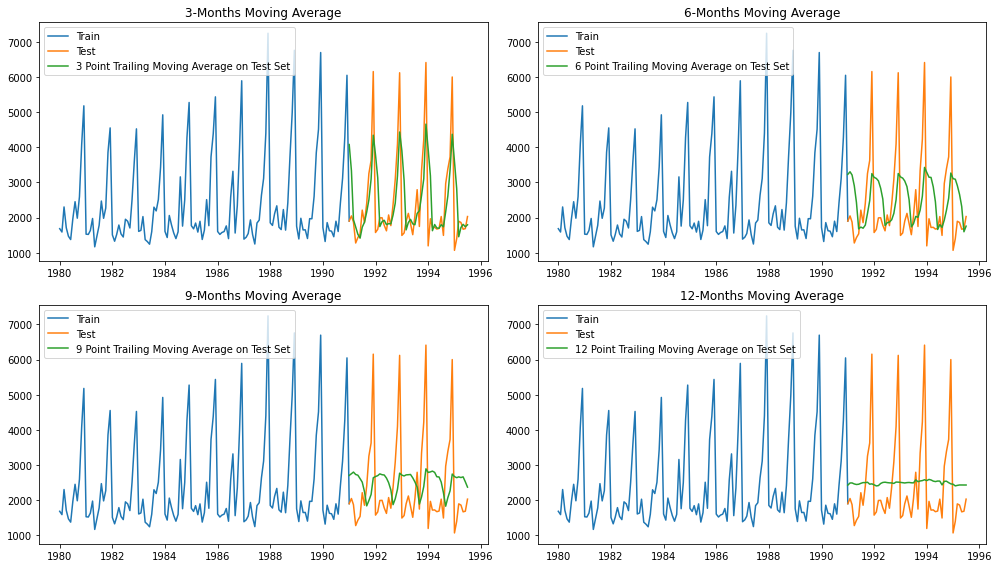
****

RMSE is 1275.074

**Moving average**



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



Three trailing is giving us the closest result.

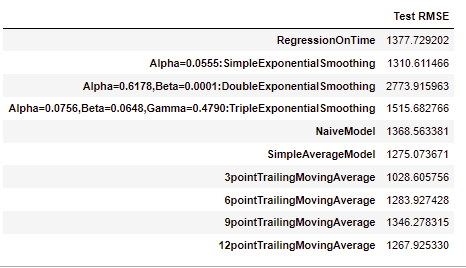
For 3 point Moving Average Model forecast on the Training Data, RMSE is 1028.606

For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

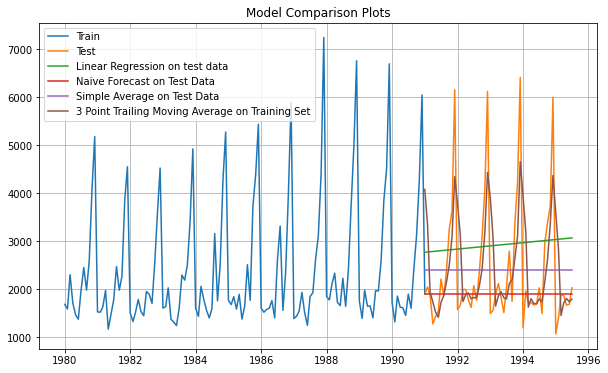
For 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278

For 12 point Moving Average Model forecast on the Training Data, RMSE is 1267.925

Table:

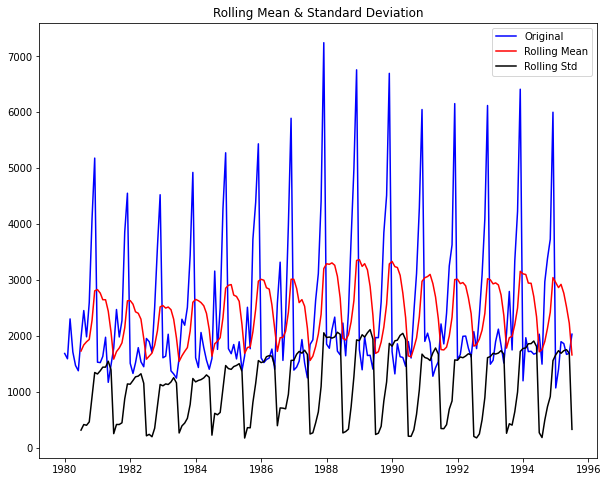


Model comparison plot



Arima

We first need to perform the test for stationarity.

Normal dataset

Results of Dickey-Fuller Test:

Test Statistic -1.360497

p-value 0.601061

#Lags Used 11.000000

Number of Observations Used 175.000000

Critical Value (1%) -3.468280

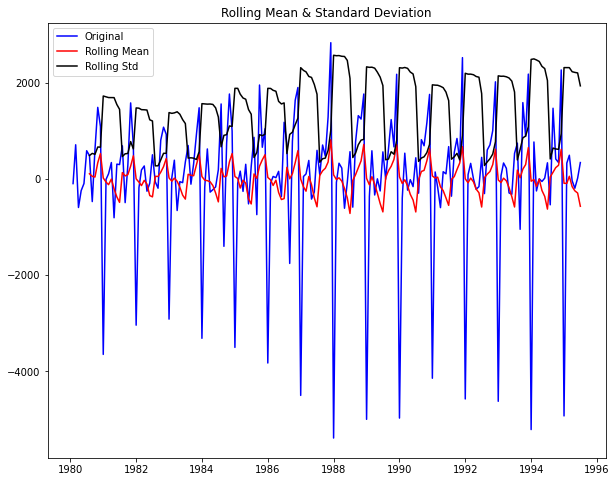
Critical Value (5%) -2.878202

Critical Value (10%) -2.575653

dtype: float64

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

Diff dataset



Results of Dickey-Fuller Test:

Test Statistic -45.050301

p-value 0.000000

#Lags Used 10.000000

Number of Observations Used 175.000000

Critical Value (1%) -3.468280

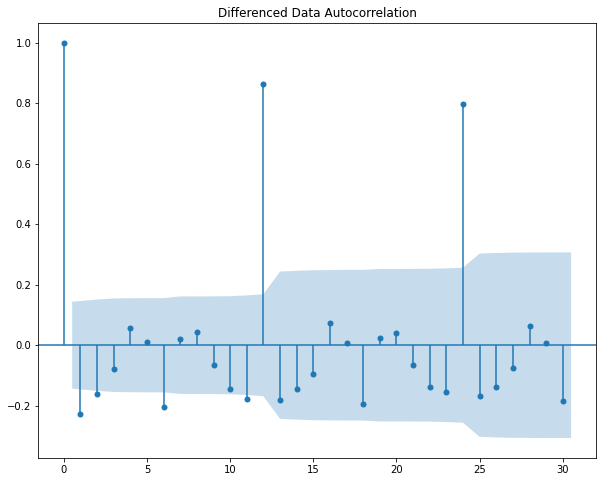
Critical Value (5%) -2.878202

Critical Value (10%) -2.575653

dtype: float64

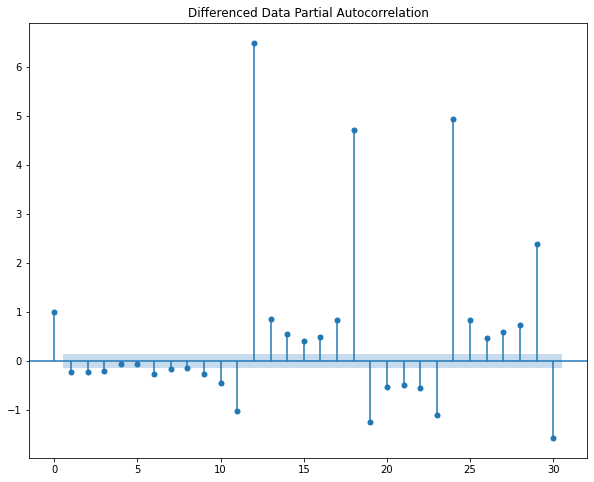
This is fine as the value of p is under the desired 5% .

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

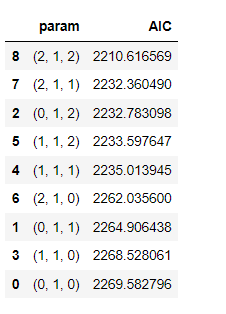
train = df[df.index < '1991-01-01']

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 =2 , d = 1, q = 2.

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

RMSE - 1375.1911127976161

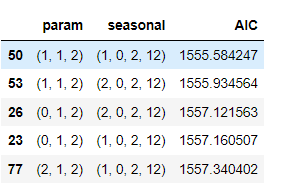
AIC - 2210.617

We will now try to improve on this.

Automated SARIMA MODEL.

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

Sorted below:

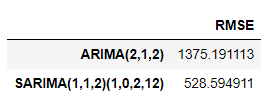


Now, we apply the values of the top parameter to the model and train it. Then after predicting on the test data. ( alpha = 0.05)

Rmse - 528.5949106740883

AIC - 1555.584

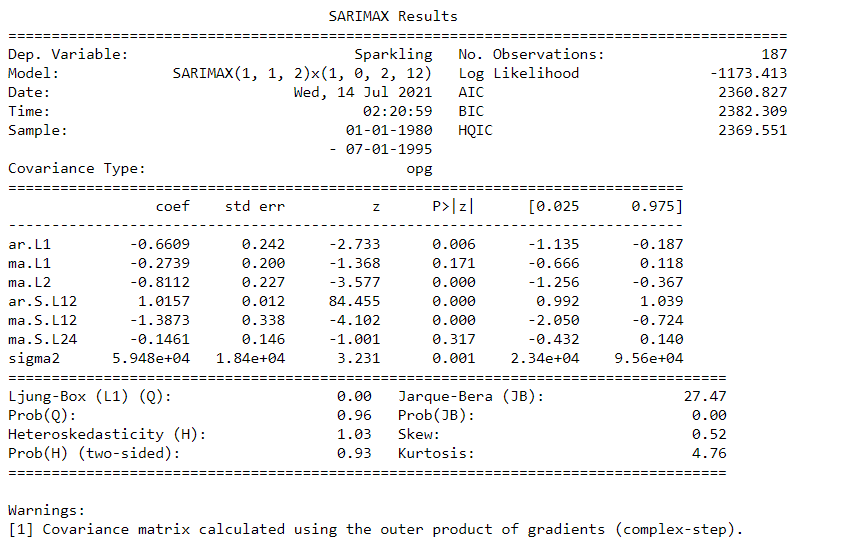
So, this is better than the previous arima model.



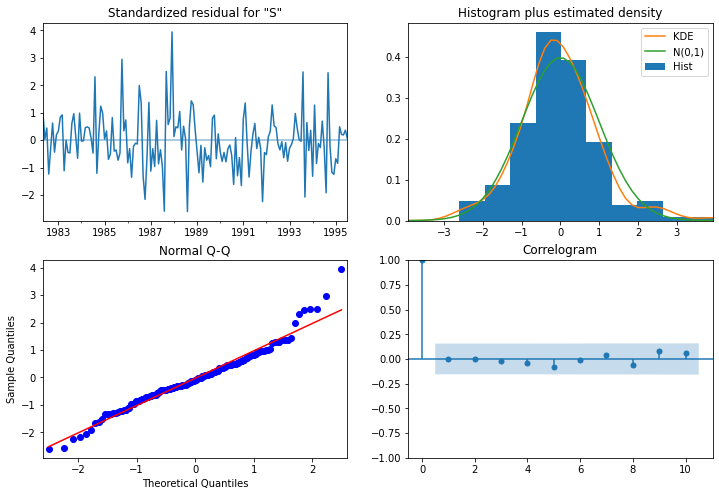
Order = (1,1,2)

seasonal\_order = (1, 0, 2, 12)

FULL DATA MODEL



Diagnostic of our model.



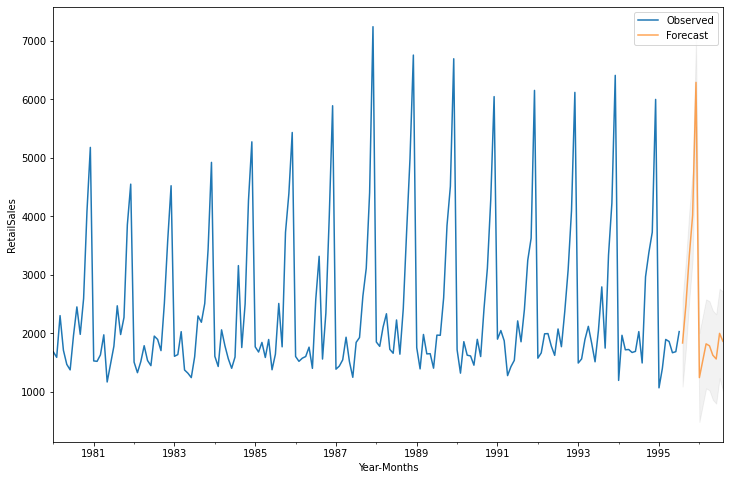
TOP LEFT: The residual error seems to fluctuate around a mean zero and has almost uniform variance.

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

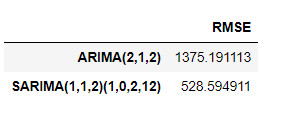
BOTTOM LEFT: The dots fall on the red line indicating normal distribution.

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:

1. We can see a steady trend in our data.
2. We are in a stable business but growth is always appreciated.
3. We can provide coupons to our loyal customers and also give out exclusive schemes for new customers, so we can increase our sales.
4. We have a loyal customer base thus, the steady trend, We need to acquire new customers and work on them.
5. Our product is doing reasonably well, we can increase the sales around the peak time by having a special discount period to capitalize on the trend.