PROJECT REPORT
Time Series Forecasting

PREEJA RAJESH
PGP – DSBA

Contents

Problem Statement 1:

For this assignment, the data of different types of wine sales in the 20th century is to be analysed. Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Data set for the Problem: Sparkling.csv

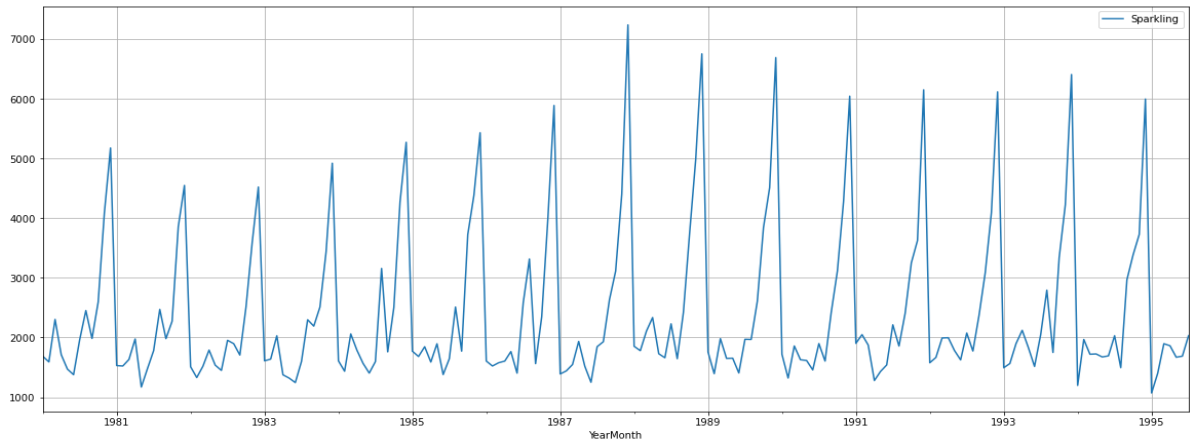
1. Read the data as an appropriate Time Series data and plot the data.

Data set:

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

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

2.1 Exploratory Data Analysis:



- **We can see there is no trend, but a seasonal pattern is associated with it.**
- **There are total 187 rows and 1 column in the dataset.**
- **There are 0 null values present in the dataset.**
- **Data types of each attribute/variables are as follows:**

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01

Data columns (total 1 columns):

#   Column     Non-Null Count  Dtype
---  ------     --------------  -----
0   Sparkling  187 non-null    int64

dtypes: int64(1)
memory usage: 2.9 KB
```

5 Point summary:

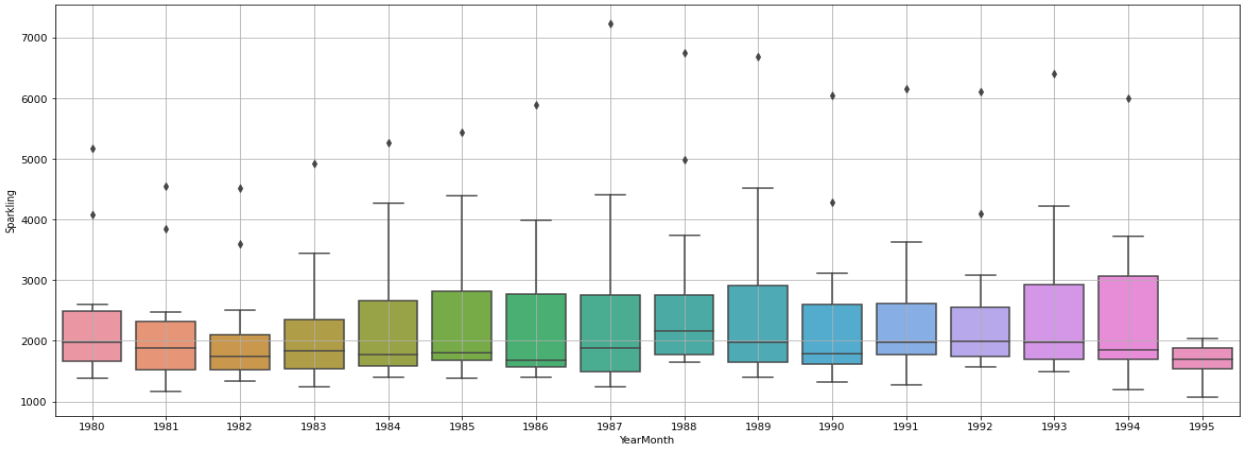
	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

Inference:

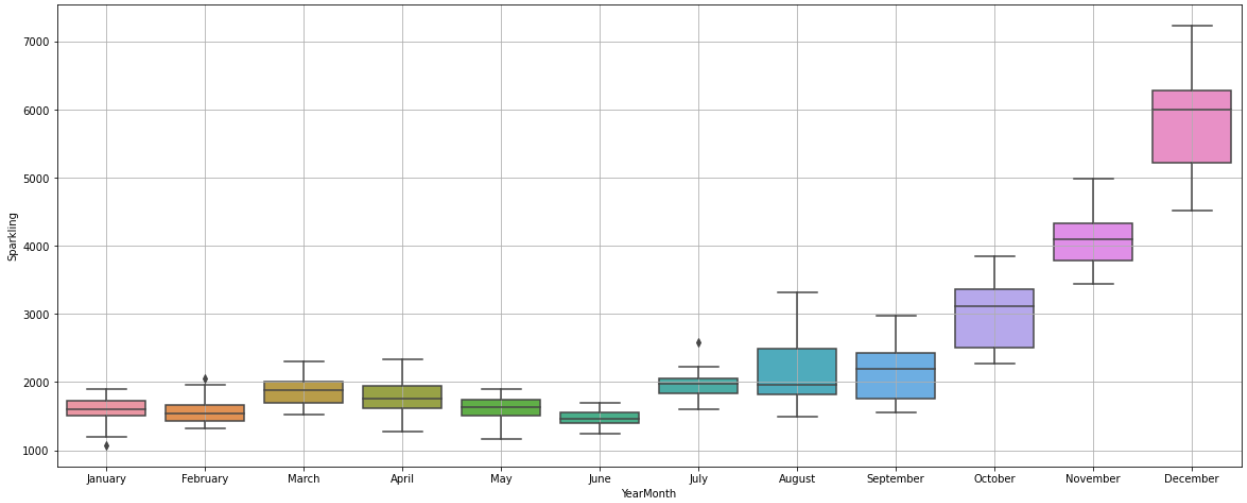
- The total sales recorded in this data is 187.
- Maximum Sparkling Wine sales is 7242.
- Here we can see 50% sales is below 1874.
- Minimum Sparkling Wine sales is 1070.

The basic measures of descriptive statistics tell us how the Sales have varied across years. But remember, for this measure of descriptive statistics we have averaged over the whole data without taking the time component into account.

Yearly Boxplot:



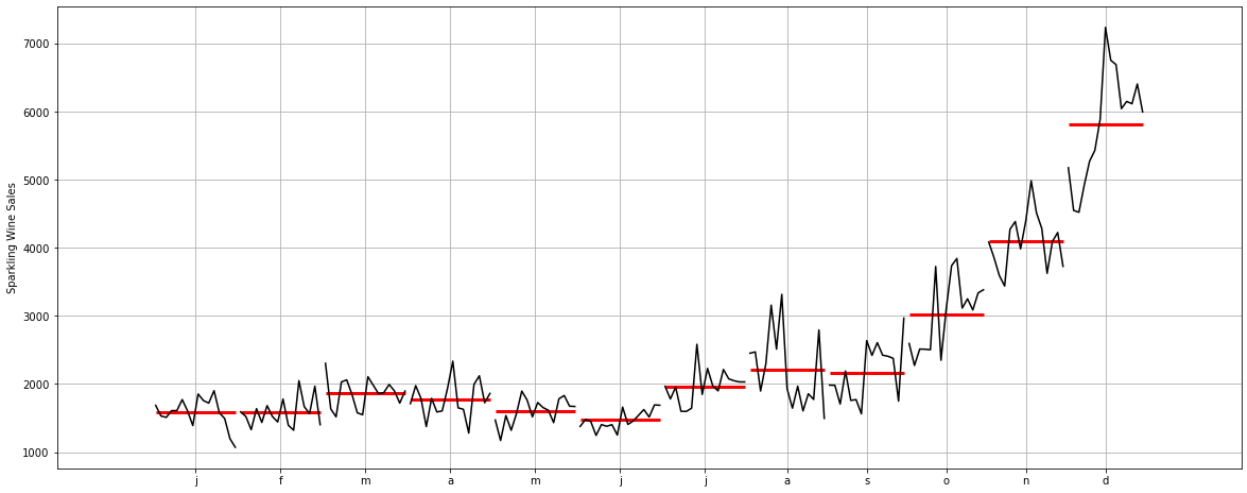
Monthly Plot:



Inference:

- There is a clear distinction of 'Sparkling wine sales' within different months spread across various years.
- The highest such numbers are being recorded in the month of November and December across various years.
- The boxplot shows that the data has few outliers, but it will not affect the modelling.

Plot a time series month plot to understand the spread of accidents across different years and within different months across years.

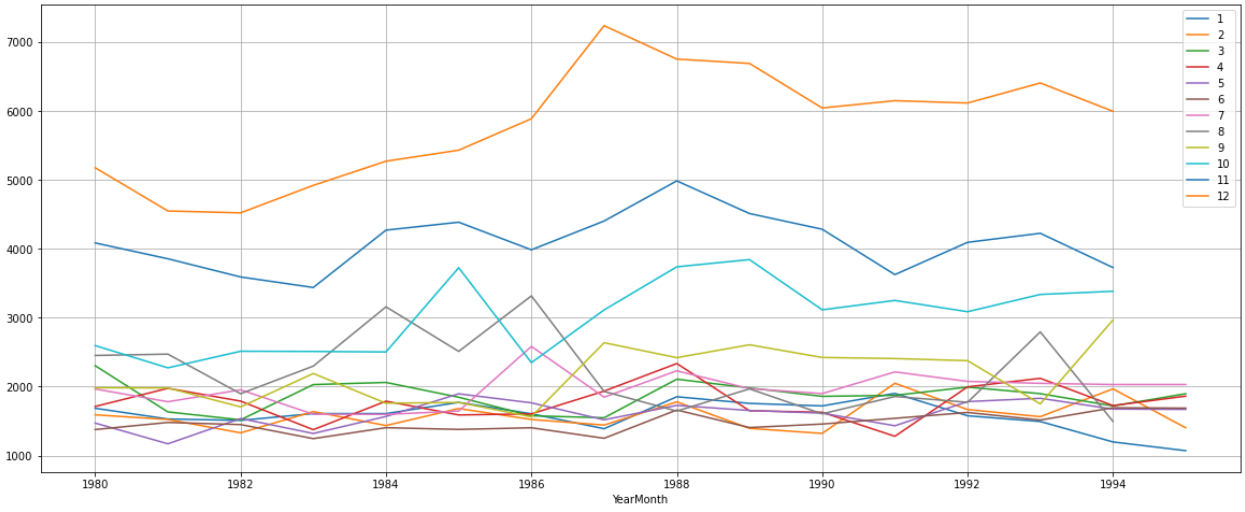


Inference:

- This plot shows us the behaviour of the Time Series ('Sparkling Wine' in this case) across various months. The red line is the median value.

Plot a graph of monthly Sparkling Wine Sales across years:

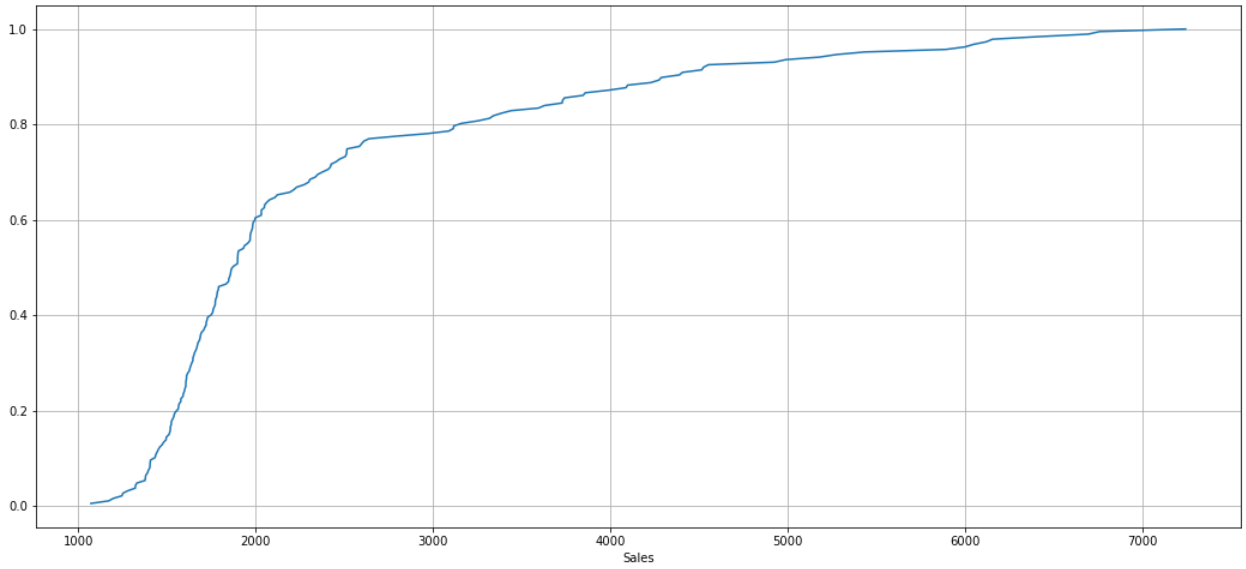
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN



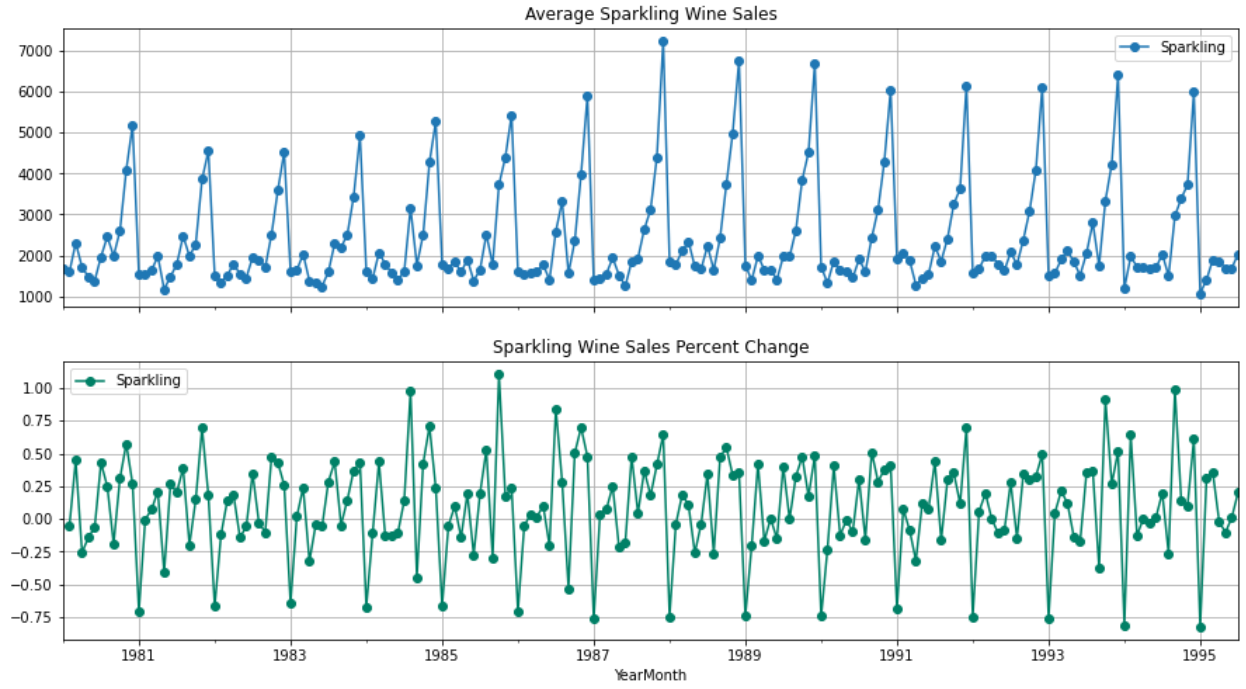
Inference:

- December month has the highest sales of sparkling wine for all the years.

Plot the Empirical Cumulative Distribution.



Plot the average Sparkling Wine Sales per month and the month on month percentage change of Sparkling Wine Sales.

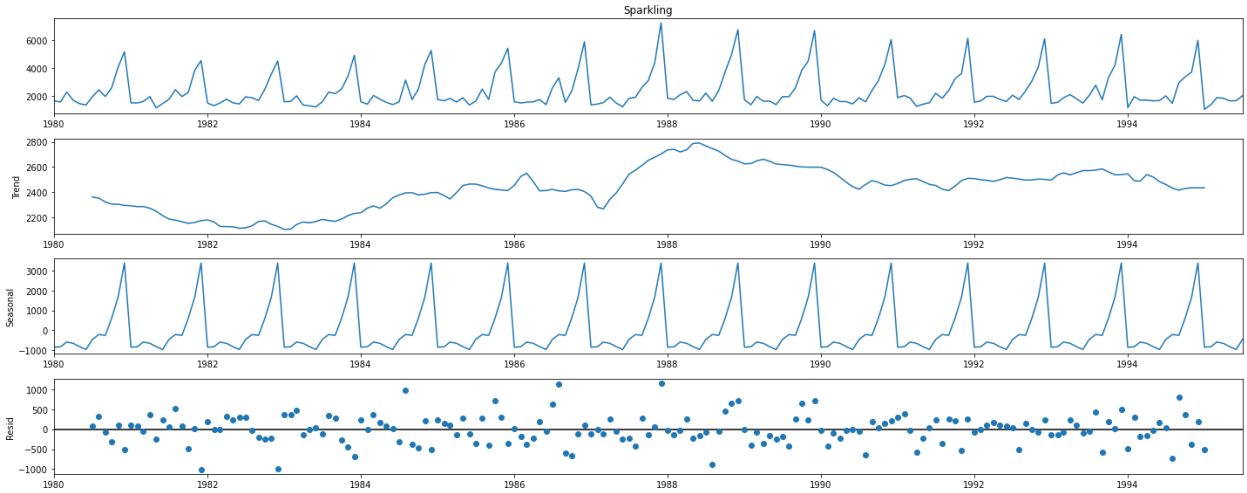


Inference:

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

2.2 Decompose the Time Series and plot the different components.

2.2.1Additive Decomposition:

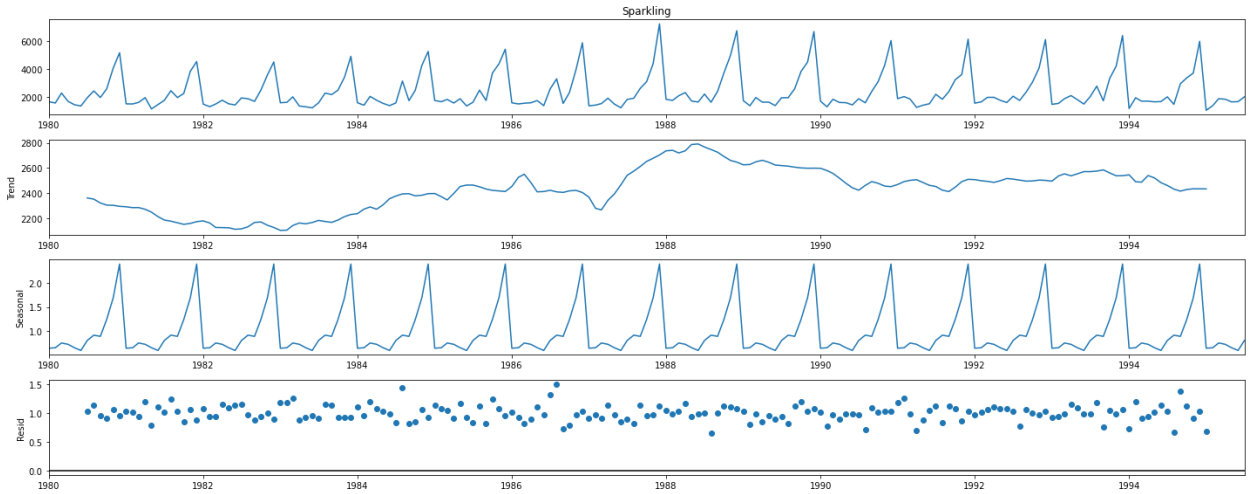


Trend		Seasonality		Residual	
YearMonth		YearMonth		YearMonth	
01-01-1980	NaN	01-01-1980	-854.260599	01-01-1980	NaN
01-02-1980	NaN	01-02-1980	-830.350678	01-02-1980	NaN
01-03-1980	NaN	01-03-1980	-592.35663	01-03-1980	NaN
01-04-1980	NaN	01-04-1980	-658.490559	01-04-1980	NaN
01-05-1980	NaN	01-05-1980	-824.416154	01-05-1980	NaN
01-06-1980	NaN	01-06-1980	-967.434011	01-06-1980	NaN
01-07-1980	2360.667	01-07-1980	-465.502265	01-07-1980	70.835599
01-08-1980	2351.333	01-08-1980	-214.332821	01-08-1980	315.999487
01-09-1980	2320.542	01-09-1980	-254.677265	01-09-1980	-81.864401
01-10-1980	2303.583	01-10-1980	599.769957	01-10-1980	-307.35329
01-11-1980	2302.042	01-11-1980	1675.067179	01-11-1980	109.891154
01-12-1980	2293.792	01-12-1980	3386.983846	01-12-1980	-501.775513
Name: trend, dtype: float64		Name: seasonal, dtype: float64		Name: resid, dtype: float64	

Inference:

- We see that the residuals are not located only around 0 from the plot of the residuals in the decomposition. Therefore, we go for multiplicative decomposition.

2.2.2 Multiplicative Decomposition:

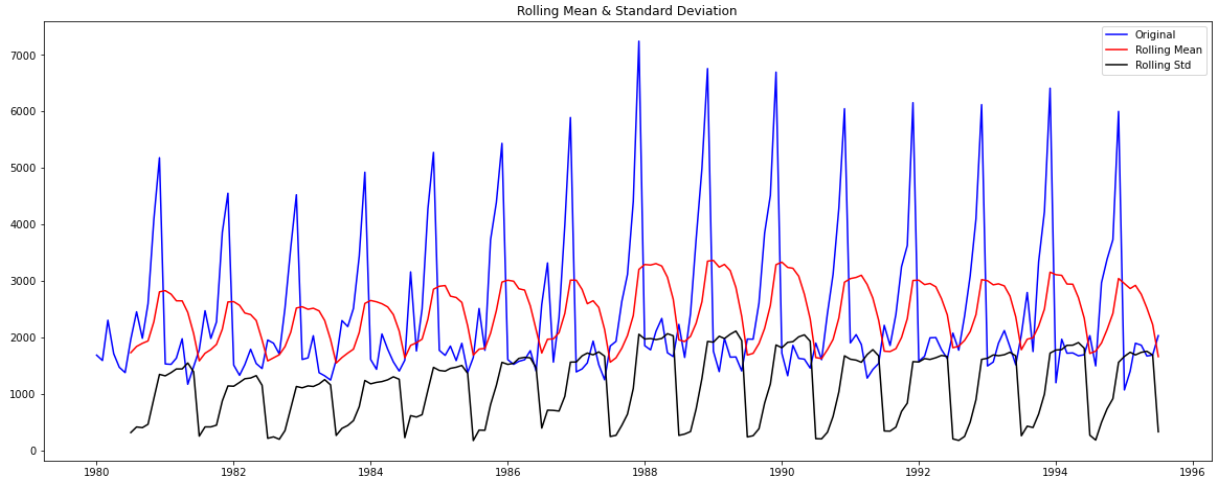


Trend		Seasonality		Residual	
YearMonth		YearMonth		YearMonth	
01-01-1980	NaN	01-01-1980	0.649843	01-01-1980	NaN
01-02-1980	NaN	01-02-1980	0.659214	01-02-1980	NaN
01-03-1980	NaN	01-03-1980	0.75744	01-03-1980	NaN
01-04-1980	NaN	01-04-1980	0.730351	01-04-1980	NaN
01-05-1980	NaN	01-05-1980	0.660609	01-05-1980	NaN
01-06-1980	NaN	01-06-1980	0.603468	01-06-1980	NaN
01-07-1980	2360.666667	01-07-1980	0.809164	01-07-1980	1.02923
01-08-1980	2351.333333	01-08-1980	0.918822	01-08-1980	1.135407
01-09-1980	2320.541667	01-09-1980	0.894367	01-09-1980	0.955954
01-10-1980	2303.583333	01-10-1980	1.241789	01-10-1980	0.907513
01-11-1980	2302.041667	01-11-1980	1.690158	01-11-1980	1.050423
01-12-1980	2293.791667	01-12-1980	2.384776	01-12-1980	0.94677
Name: trend,dtype: float64		Name: seasonal, dtype: float64		Name: resid,dtype: float64	

Inference:

- For the multiplicative series, we see that a lot of residuals are located around 1.

2.3 Check for stationarity of the whole Time Series data:

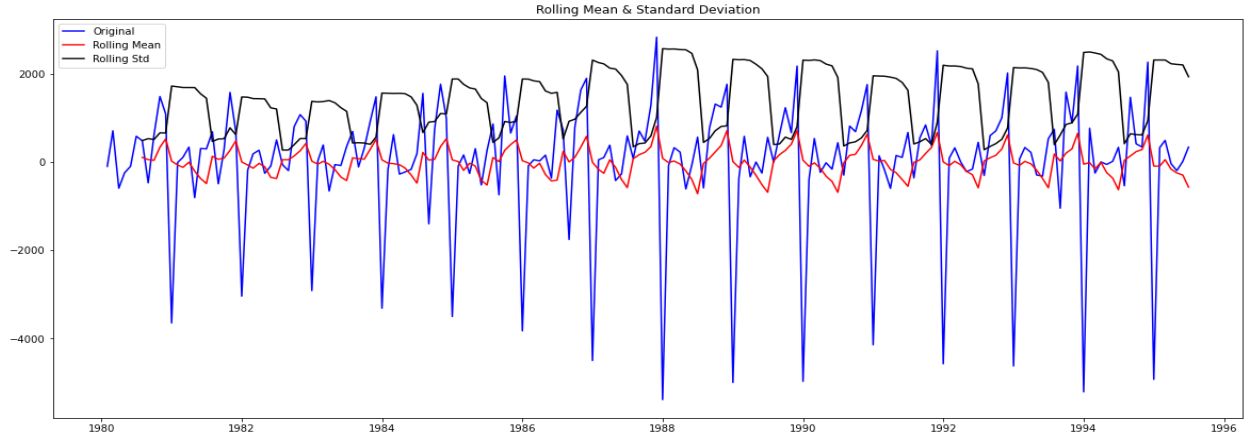



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

Inference:

- We see that at 5% significant level the Time Series is non-stationary.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

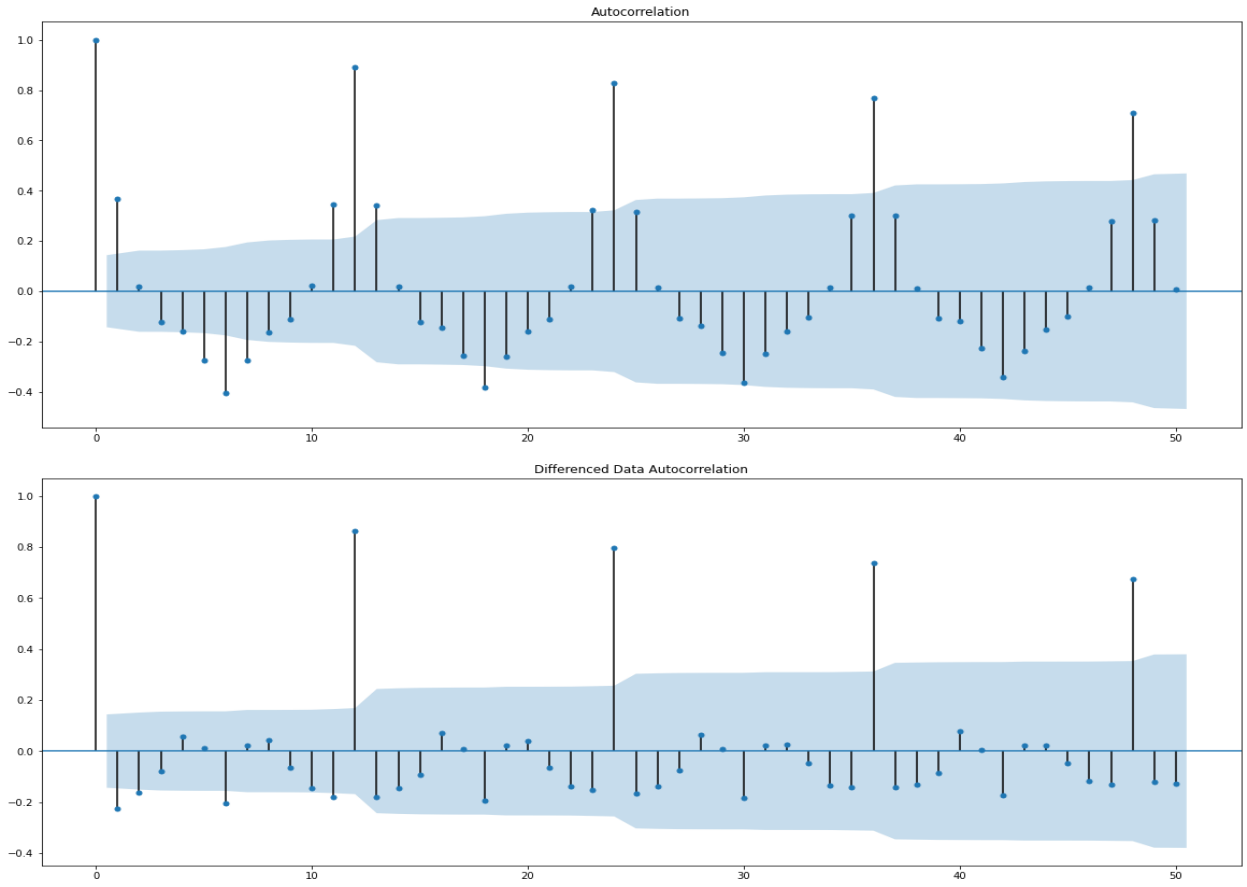


```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

Inference:

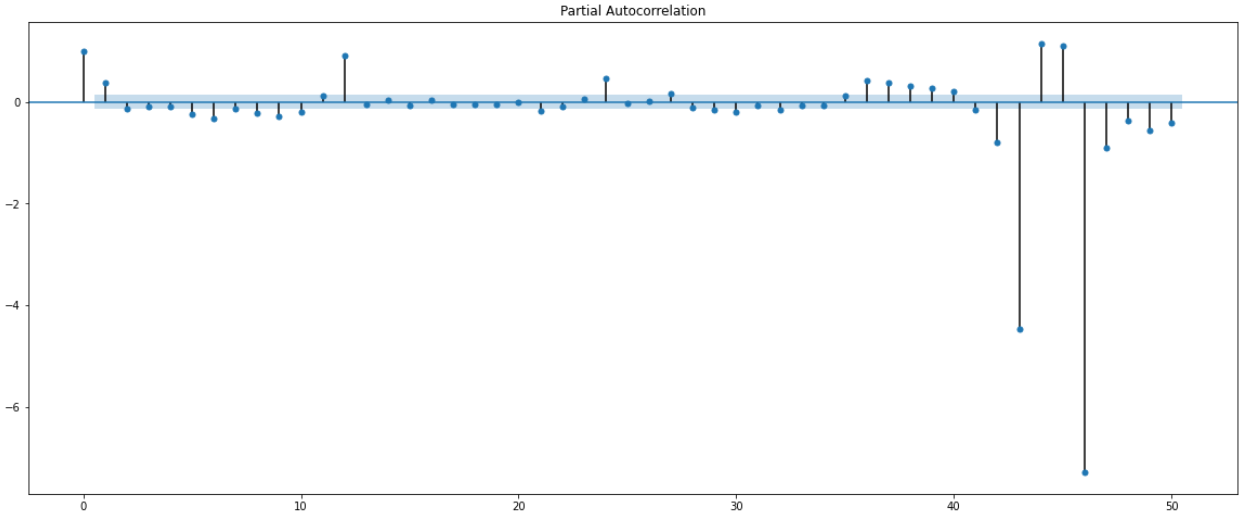
- Perfect! Our series now looks like something indescribable, oscillating around zero. The Dickey-Fuller test indicates that it is stationary, and the number of significant peaks in ACF has dropped. We can finally start modeling!

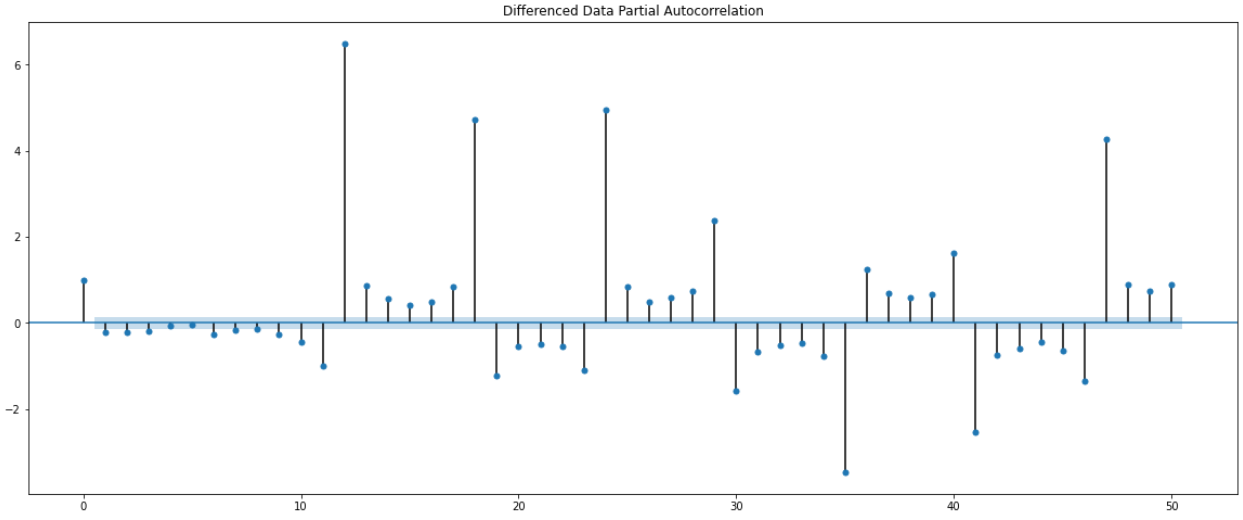
Plot the Autocorrelation function plots on the whole data.



We get the order of MA term or 'q' from ACF plot. Here the order of MA term is 2 from the differenced ACF plot.

Plot the Partial Autocorrelation function plots on the whole data.





We get the order of AR term or 'p' from PACF plot. Here the order of AR term is (2,3) from the differenced PACF plot. From the above plots, we can also say that there seems to be a seasonality in the data.

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

3.1 Train-Test Split

- Training Dataset: The sample of data used to fit the model.
- Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.
- Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.
- Train (132, 1)
- Test (55, 1)

First few rows of Training Data

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

Last few rows of Training Data

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

First few rows of Test Data

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

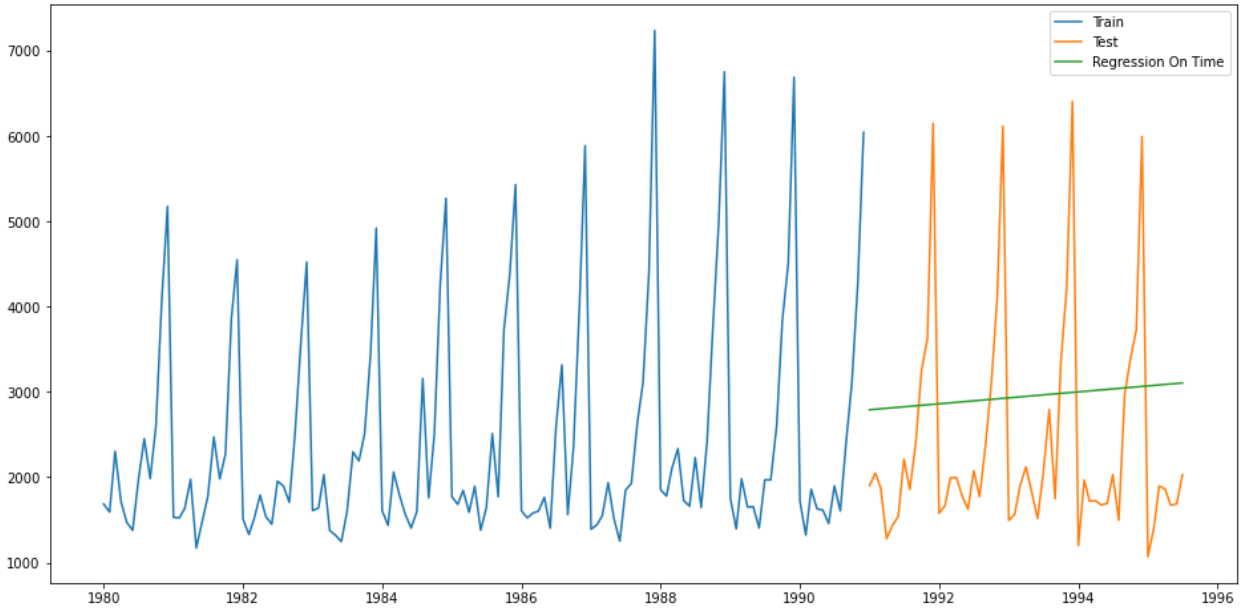
Last few rows of Test Data

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

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.

Modelling:

Method 1: Regression on Time



For RegressionOnTime,

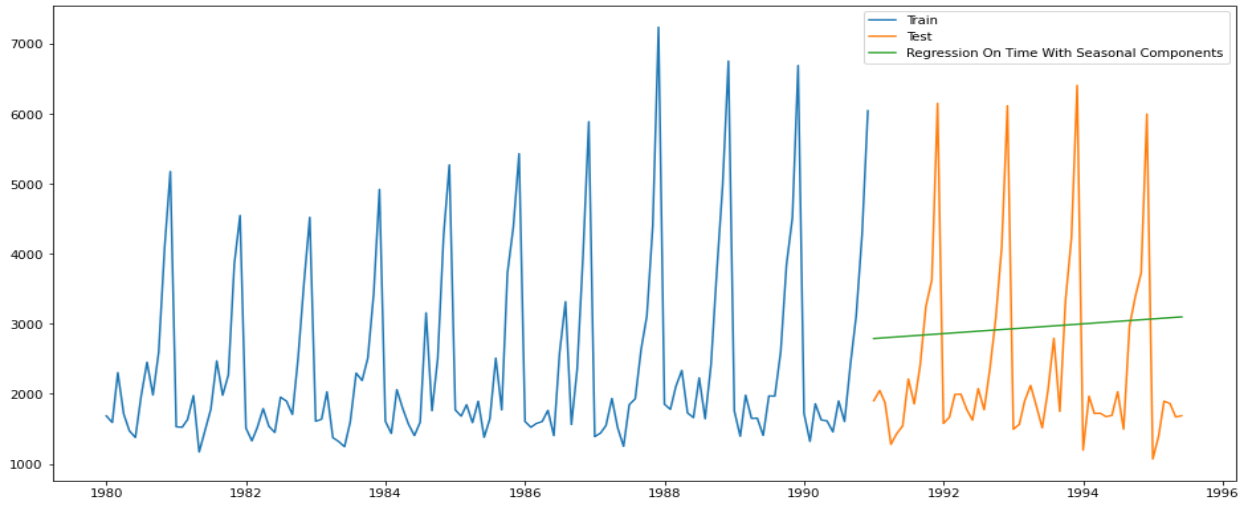
- RMSE is 1389.135
- MAPE is 50.15

Inference:

- Linear regression is a statistical tool used to help predict future values from past values. It is commonly used as a quantitative way to determine the underlying trend and when price s are overextended.

- This linear regression indicator plots the trendline value for each data point.

Method 2: Regression on Time with Seasonal Components



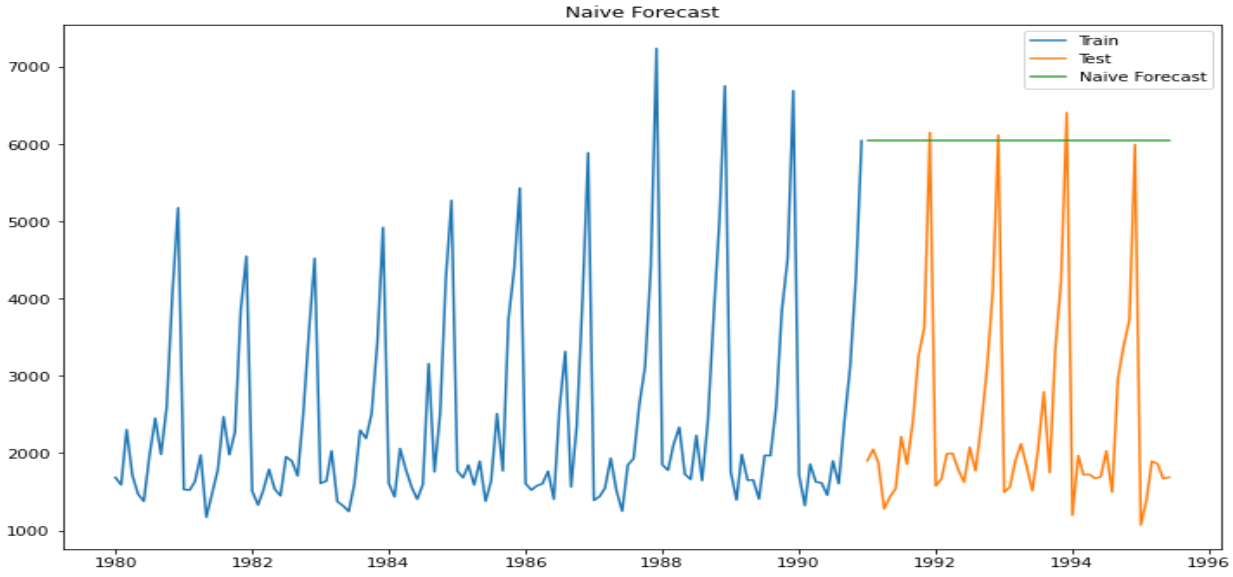
For Regression on Time Seasonal,

- RMSE is 1394.276
- MAPE is 50.11

Inference:

- Output is same to the above model.

Method 3: Naive Approach: $\hat{y}_{t+1} = y_t$



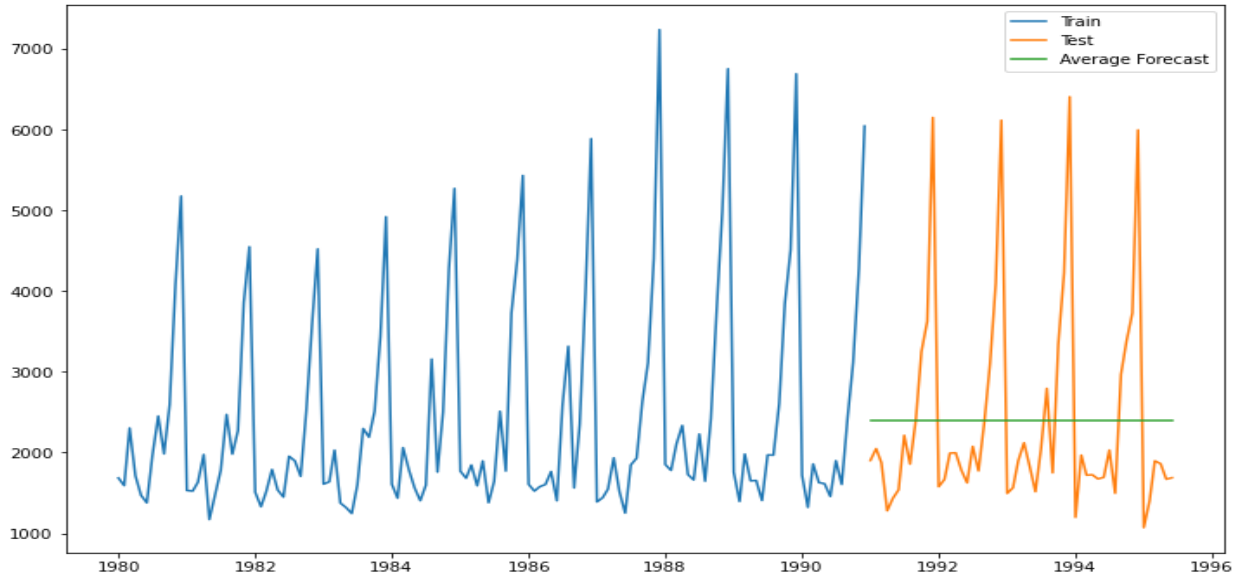
For Naive model,

- RMSE is 3861.413
- MAPE is 152.17

Inference:

- We can infer from the RMSE and MAPE values and the graphs above, that Naive method and Regression on Time with Seasonal Components model are not 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 score.

Method 4: Simple Average



For Simple Average Model,

- RMSE is 1285.834
- MAPE is 39.22

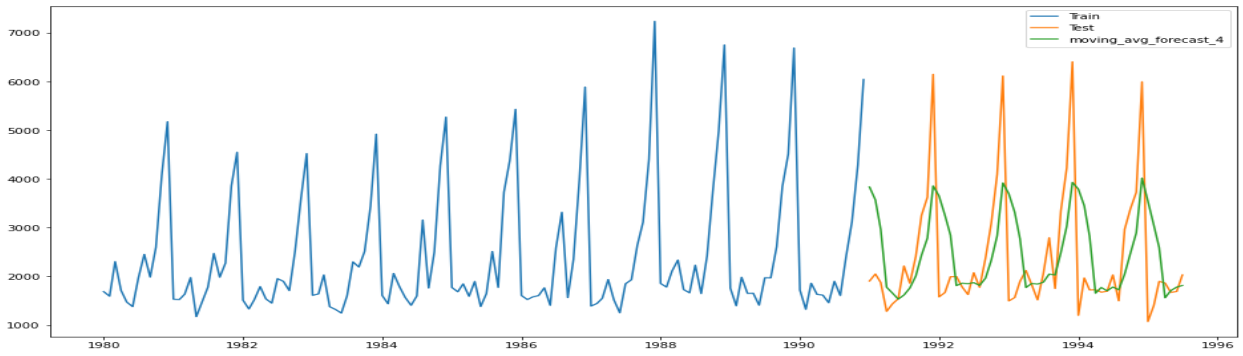
Inference:

- We can see that this model has improved our score.
- Hence, we can infer from the score that this method works best when the average at each time period remains constant. The score of Average method is better than Naive method. We should move step by step to each model and confirm whether it improves our model o r not.

Method 5: Moving Average (MA)

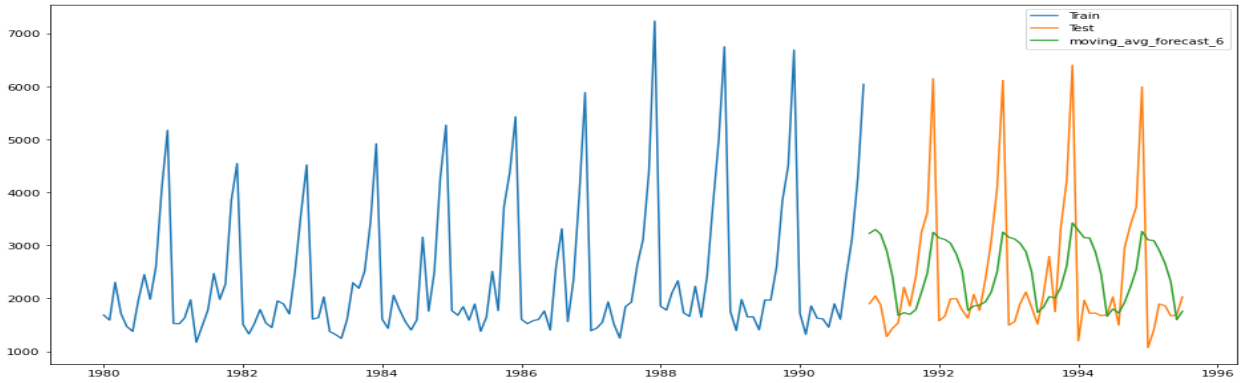
For Moving Average model, moving_avg_forecast_4

- RMSE is 1156.590
- MAPE is 35.96



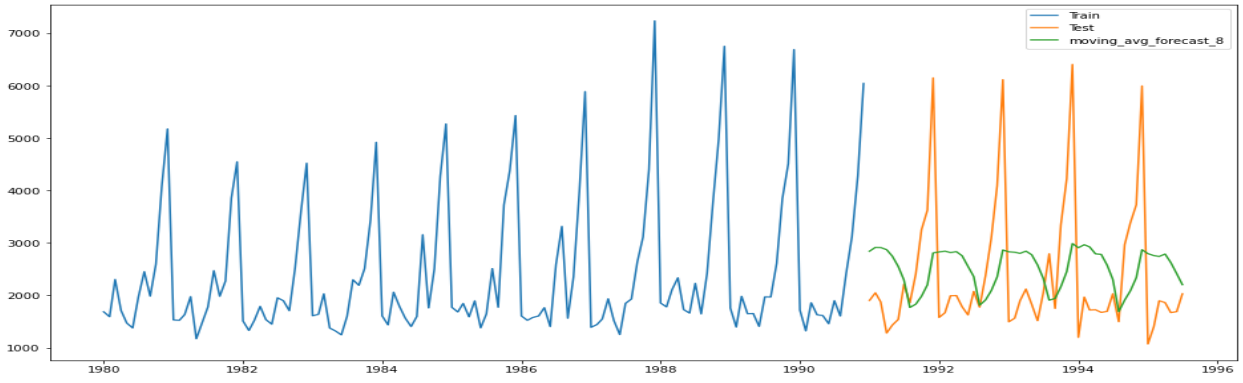
For Moving Average model, moving_avg_forecast_6

- RMSE is 1283.927
- MAPE is 43.86



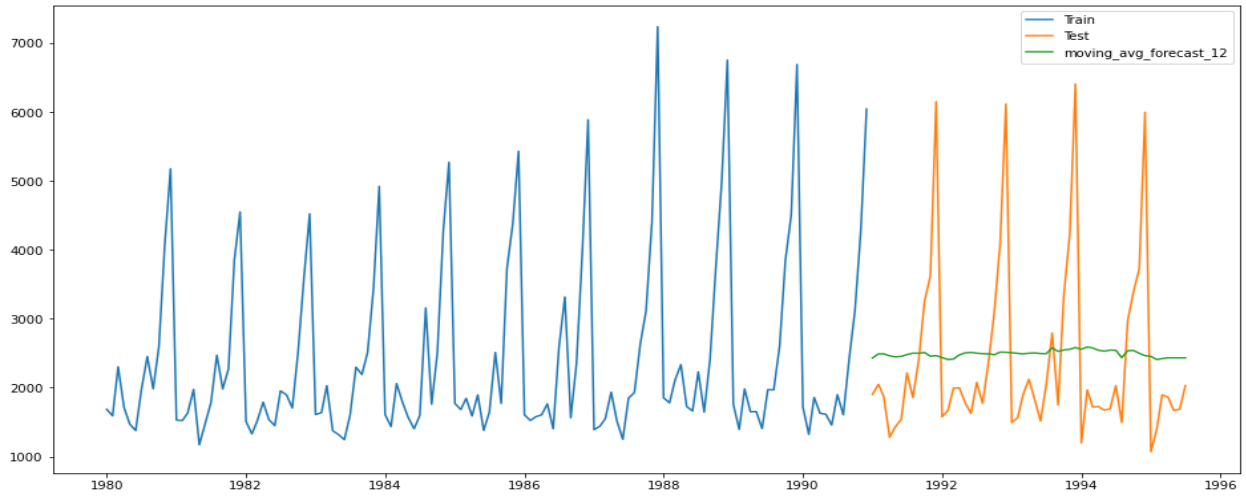
For Moving Average model, moving_avg_forecast_8

- RMSE is 1342.568
- MAPE is 46.46



For Moving Average model, moving_avg_forecast_12

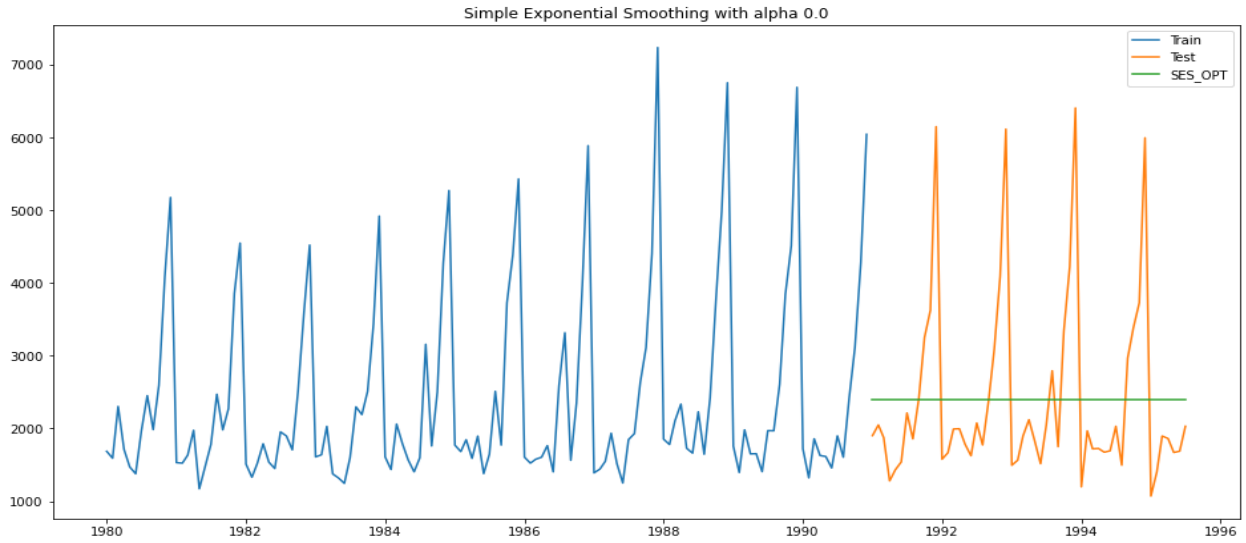
- RMSE is 1267.925
- MAPE is 40.19



Method 6: Simple Exponential Smoothing

== Simple Exponential Smoothing Parameters ==

- Smoothing Level 0.0
- Initial Level 2403.7936



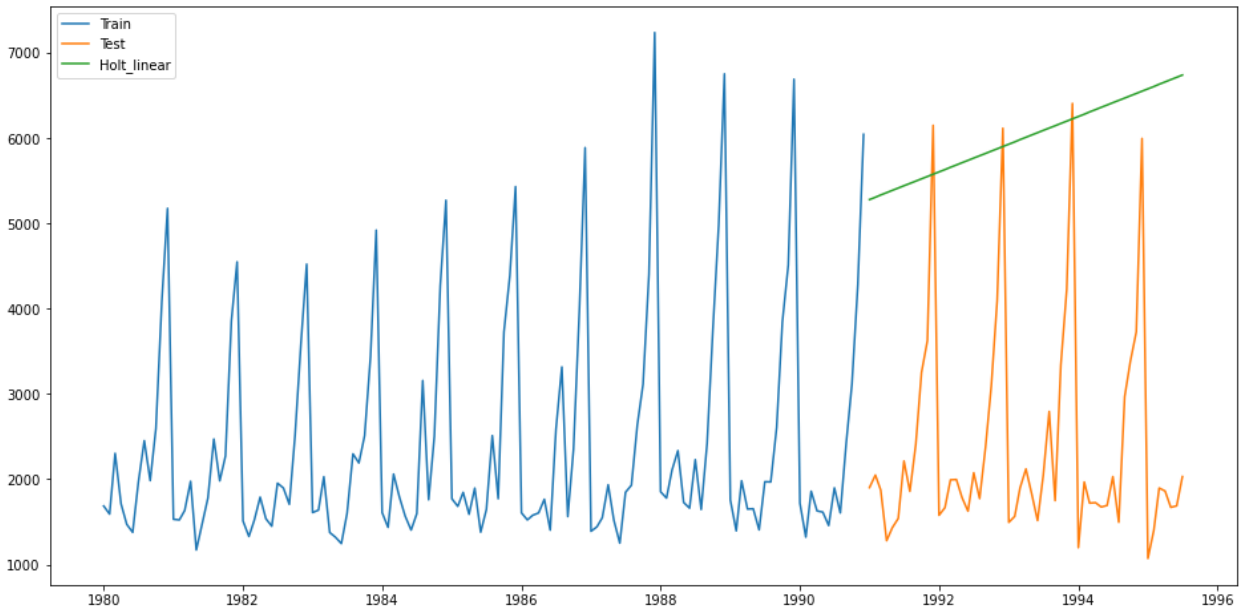
For alpha = 0.00,

- RMSE is 1275.0819
- MAPE is 38.90

Method 7: Holt's Linear Trend Method (Double Exponential Smoothing)

==Holt model Exponential Smoothing Parameters ==

- Smoothing Level 0.6478
- Smoothing Slope 0.0
- Initial Level 1686.0838



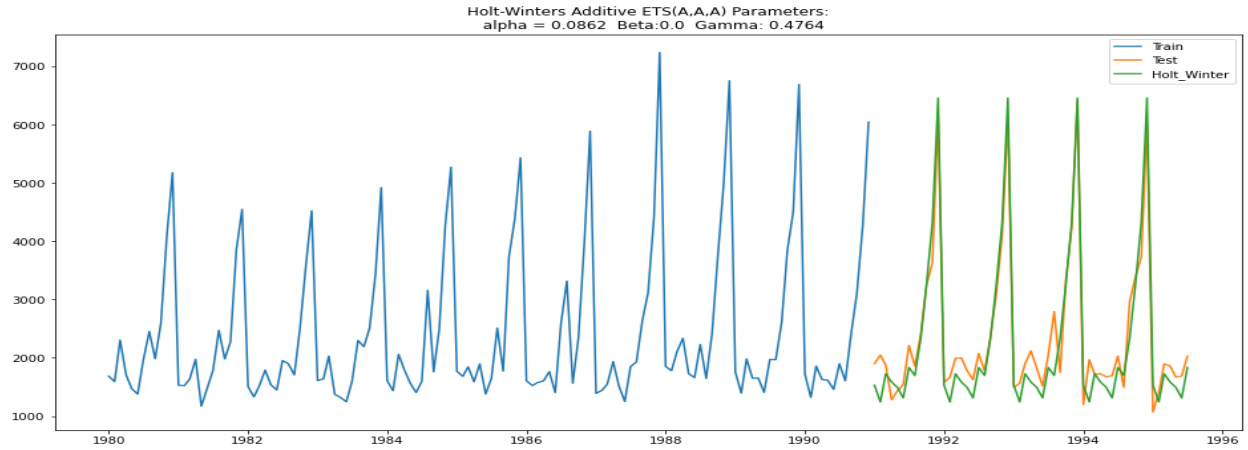
For alpha = 0.65,

- RMSE is 3851.3012
- MAPE is 152.07

Method 8: Holt-Winters Method - Additive seasonality

== Holt-Winters Additive ETS(A,A,A) Parameters ==

- Smoothing Level: 0.0862
- Smoothing Slope: 0.0
- Smoothing Seasonal: 0.4764
- Initial Level: 1684.7981
- Initial Slope: 0.0066
- Initial Seasons: [39.1924 -37.2514 464.9323 205.9561 -1 40.6716 -156.8233 338.078 856.8102 403.5116 971.2622 24 01.6525 3426.7843]



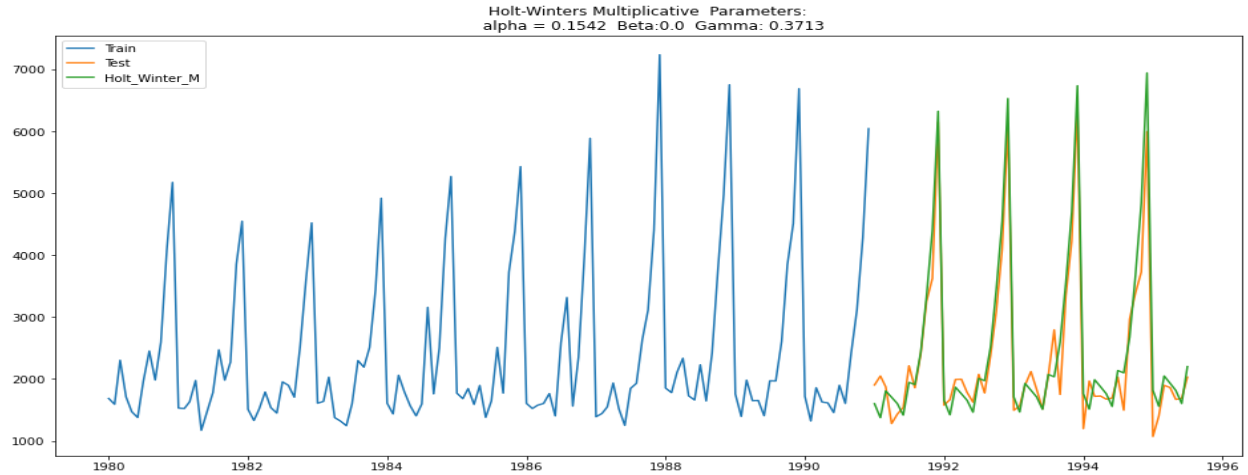
For Holt Winter alpha = 0.09, beta = 0.00, gamma = 0.48,

- RMSE is 362.7422
- MAPE is 12.08

Method 9: Holt-Winters Method - Multiplicative Model

```
== Holt-Winters Multiplicative ETS (A, A, M) Parameters ==

Smoothing Level:    0.1542
Smoothing Slope:    0.0
Smoothing Seasonal:0.3713
Initial Level:      1639.9993
Initial Slope:      4.8484
Initial Seasons: [1.0084 0.969  1.2418 1.1321 0.9398 0.9381
                  1.2246 1.5443 1.2734 1.632 2.4829 3.1186]
```



For alpha = 0.15, beta = 0.00, gamma = 0.37,

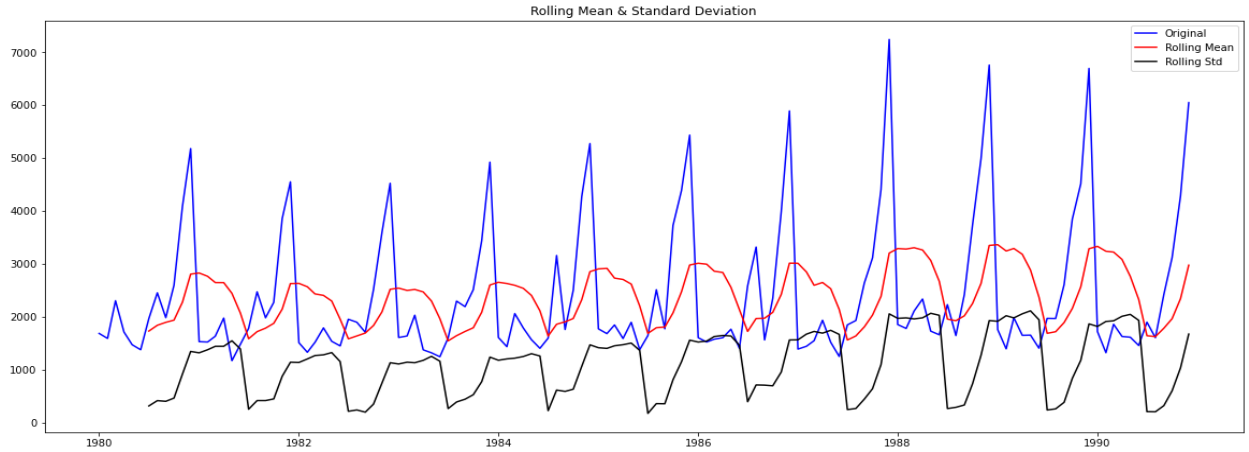
- RMSE is 383.1765
- MAPE is 11.91

Inference:

As of now, we observe that Moving average of window width of 4 seems to be a good fit for the data.

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and 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.

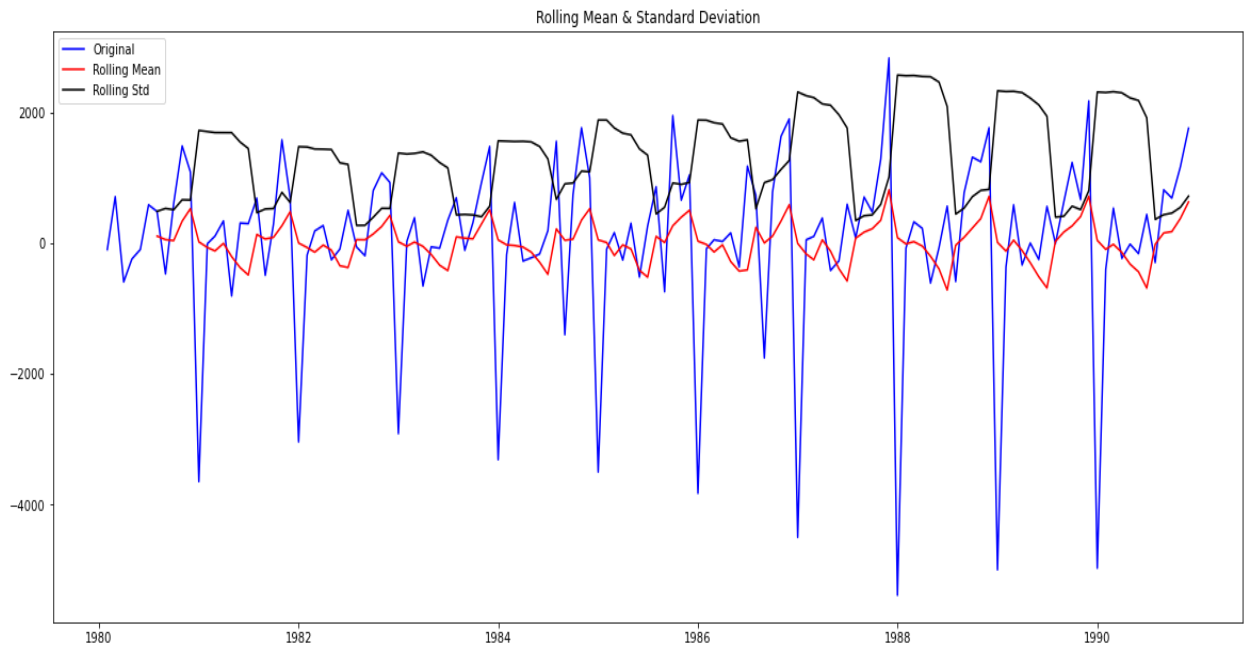
Check for stationarity of the Training Data Time Series.



Results of Dickey-Fuller Test:

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

We see that the series is not stationary at α = 0.05.



Results of Dickey-Fuller Test:

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

Inference:

We see that after taking a difference of order 1 the series have become stationary at $\alpha = 0.05$.

- The results show that the test statistic i.e. the p-value is 2.280104e-12 which is <0.05, therefore, we reject the null hypothesis and hence time series is stationary. This suggests that we can reject the null hypothesis with a significance level of less than 1% (i.e. a low probability that the result is a statistical fluke).
- Rejecting the null hypothesis means that the process has no unit root, and in turn that the time series is stationary or does not have time-dependent structure.

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.

Method 10: Auto ARIMA Model

	param	AIC
10	(2, 1, 2)	2210.620488
15	(3, 1, 3)	2225.661559
14	(3, 1, 2)	2228.927152
11	(2, 1, 3)	2229.358094
9	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098

ARIMA Model Results

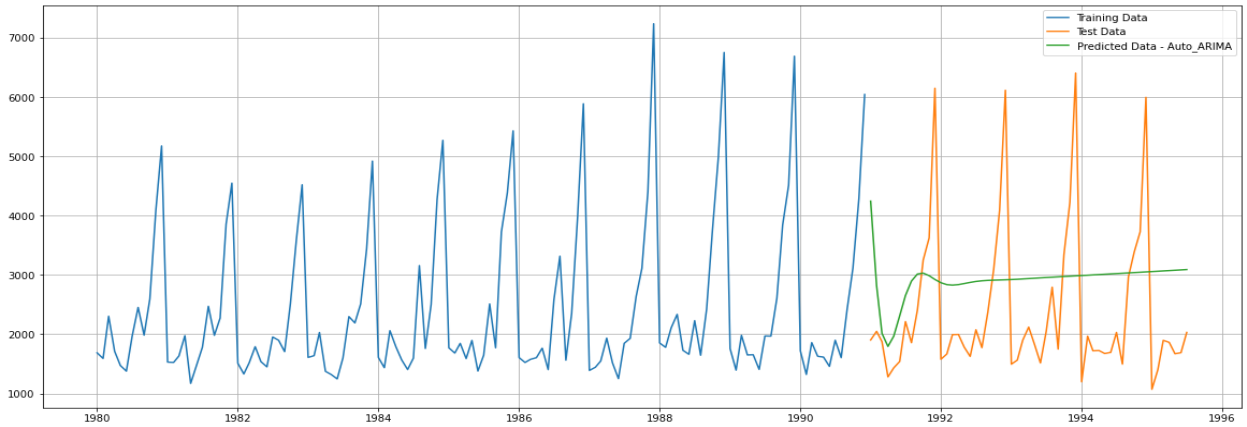
Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-109.310
Method:	css-mle	S.D. of innovations	1013.051
Date:	Fri, 11 Sep 2020	AIC	2210.620
Time:	19:27:43	BIC	2227.872
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.630

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5844	0.518	10.781	0.000	4.569	6.600
ar.L1.D.Sparkling	1.2698	0.075	17.044	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5601	0.074	-7.617	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9971	0.042	-47.045	0.000	-2.080	-1.914
ma.L2.D.Sparkli	0.9971	0.043	23.454	0.000	0.914	1.080

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7074j	1.3362	-0.0888
AR.2	1.1335	+0.7074j	1.3362	0.0888
MA.1	1.0002	+0.0000j	1.0002	0.0000
MA.2	1.0027	+0.0000j	1.0027	0.0000

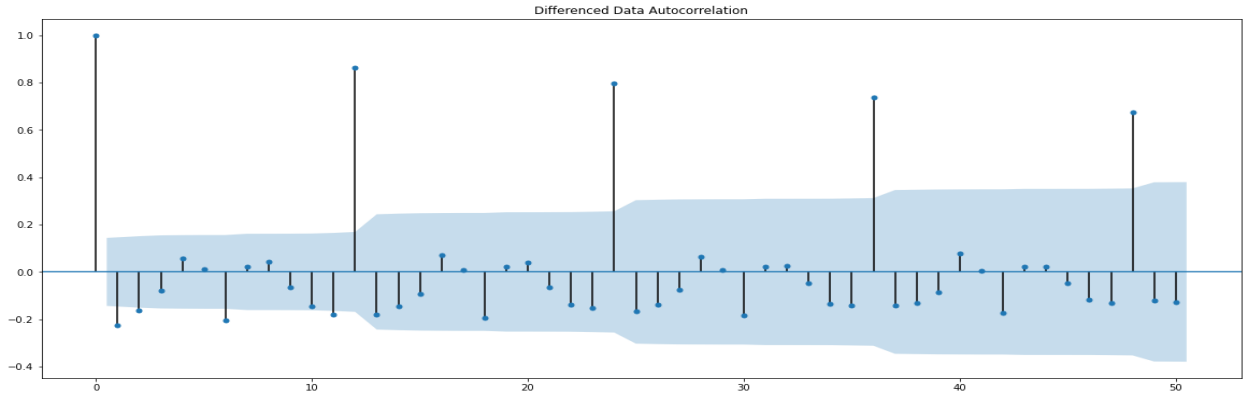
- RMSE: 1374.3863564598623
- MAPE: 48.35

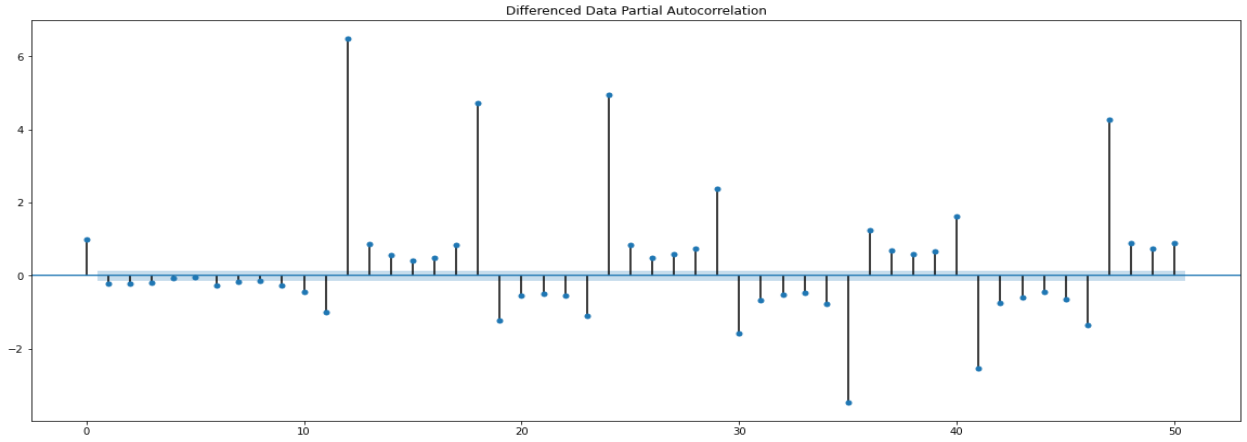


7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Method 11: Manual ARIMA Model

Let us look at the ACF and the PACF plots once more.





Inference:

Here, we have taken alpha = 0.05.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the lag at which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the lag at which the ACF plot cuts-off to 0.
- By looking at the above plots, we can say that the PACF plot cuts-off at lag 3 and ACF plot cuts-off at lag 2.

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.464
Method:	css-mle	S.D. of innovations	1106.010
Date:	Fri, 11 Sep 2020	AIC	2228.927
Time:	19:28:18	BIC	2249.054
Sample:	02-01-1980 - 12-01-1990	HQIC	2237.105

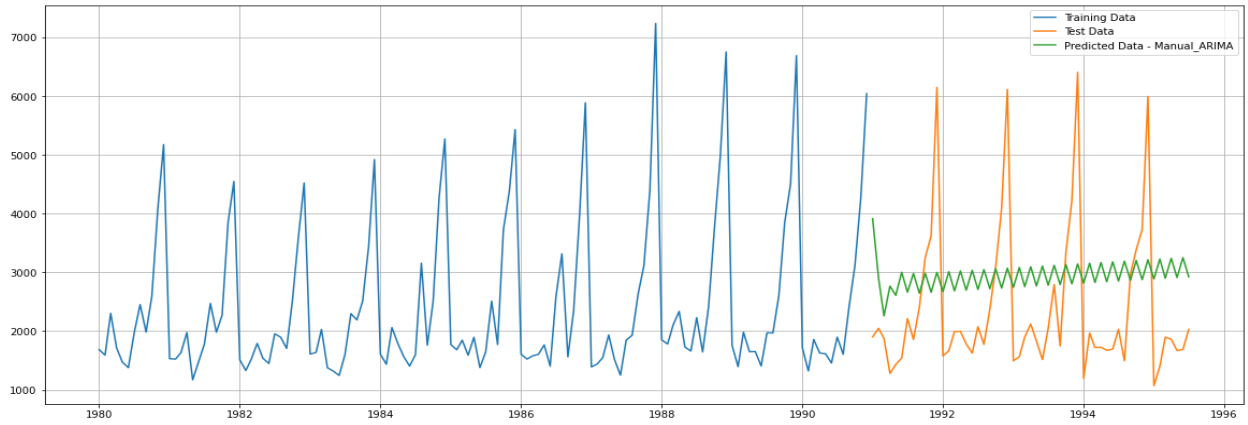
	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9850	nan	nan	nan	nan	nan
ar.L1.D.Sparkling	-0.4419	nan	nan	nan	nan	nan
ar.L2.D.Sparkling	0.3079	7.74e-06	3.98e+04	0.000	0.308	0.308
ar.L3.D.Sparkling	-0.2501	nan	nan	nan	nan	nan
ma.L1.D.Sparkling	-0.0004	0.028	-0.013	0.990	-0.055	0.055
ma.L2.D.Sparkling	-0.9996	0.028	-35.603	0.000	-1.055	-0.945

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	1.1156	-1.6594j	1.9995	-0.1558
AR.3	1.1156	+1.6594j	1.9995	0.1558
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0004	+0.0000j	1.0004	0.5000

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

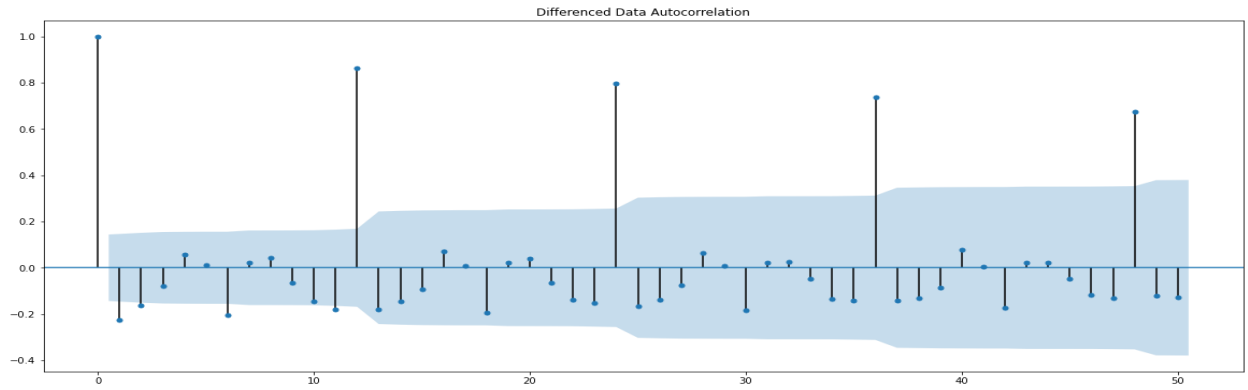
- RMSE: 1378.9279400921973
- MAPE: 49.31



We see that the difference in RMSE values is about 4 with a manual model built.

Method 12: Auto SARIMA Model_6

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model.



Inference:

- We see that there can be a seasonality of 6 as well as 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

Setting the seasonality as 6 for the first iteration of the auto SARIMA model.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.670869
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888804
80	(2, 1, 2)	(2, 0, 2, 6)	1729.363553
17	(0, 1, 1)	(2, 0, 2, 6)	1741.696452
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379779

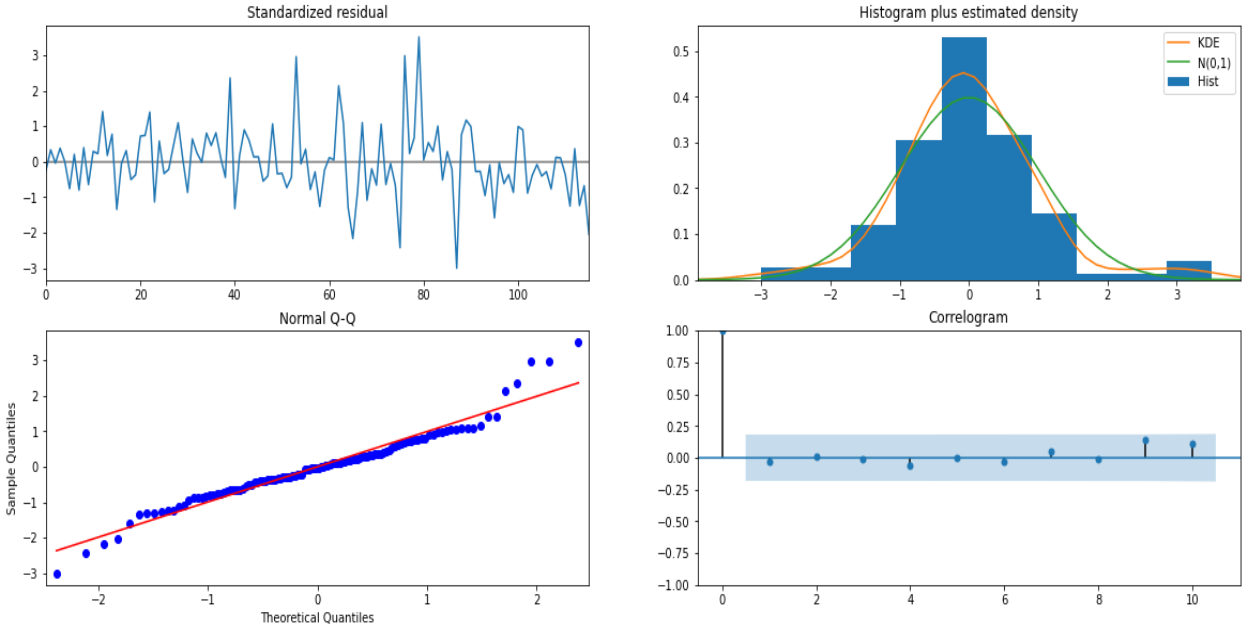
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood		-855.835
Date:	Fri, 11 Sep 2020	AIC	1727.671
Time:	19:29:26	BIC	1749.700
Sample:	0	HQIC	1736.613
	- 132		
Covariance Type:	opg		

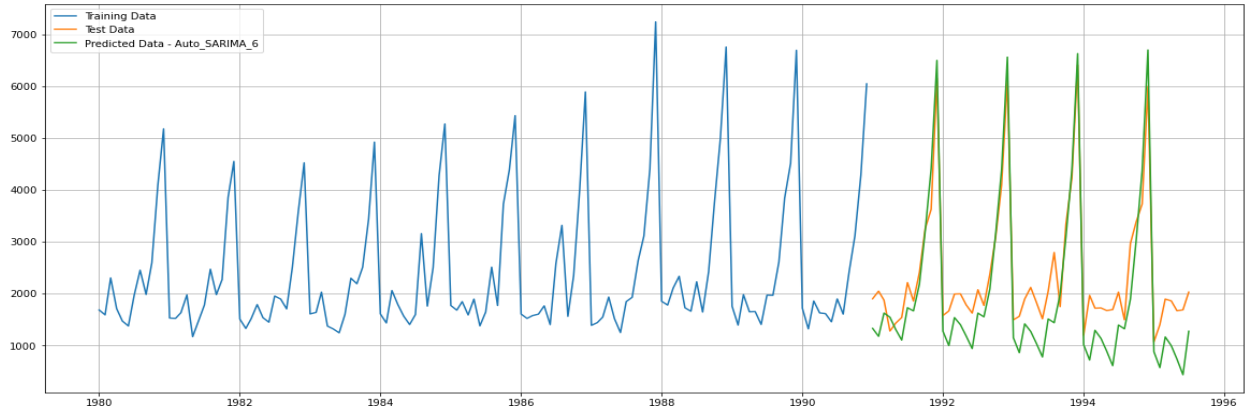
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6453	0.286	-2.256	0.024	-1.206	-0.085
ma.L	-0.3355	0.228	-1.474	0.140	-0.781	0.111
ma.L2	-0.8805	0.277	-3.178	0.001	-1.423	-0.338
ar.S.L6	-0.0045	0.027	-0.165	0.869	-0.057	0.049
ar.S.L12	1.0361	0.018	56.106	0.000	1.000	1.072
ma.S.L6	0.0676	0.152	0.444	0.657	-0.231	0.366
ma.S.L12	-0.6125	0.093	-6.593	0.000	-0.795	-0.430
sigma2	1.152e+05	1.78e+04	6.458	0.000	8.03e+04	1.5e+05

Ljung-Box (Q):	28.94	Jarque-Bera (JB):	25.27
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	2.63	Skew:	0.47
Prob(H) (two-sided):	0.00	Kurtosis:	5.09

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



From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.



- RMSE: 626.8772381455102
- MAPE: 22.55

Method 13: Auto SARIMA Model_12

Setting the seasonality as 12 for the first iteration of the auto SARIMA model.

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.929655
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121563
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

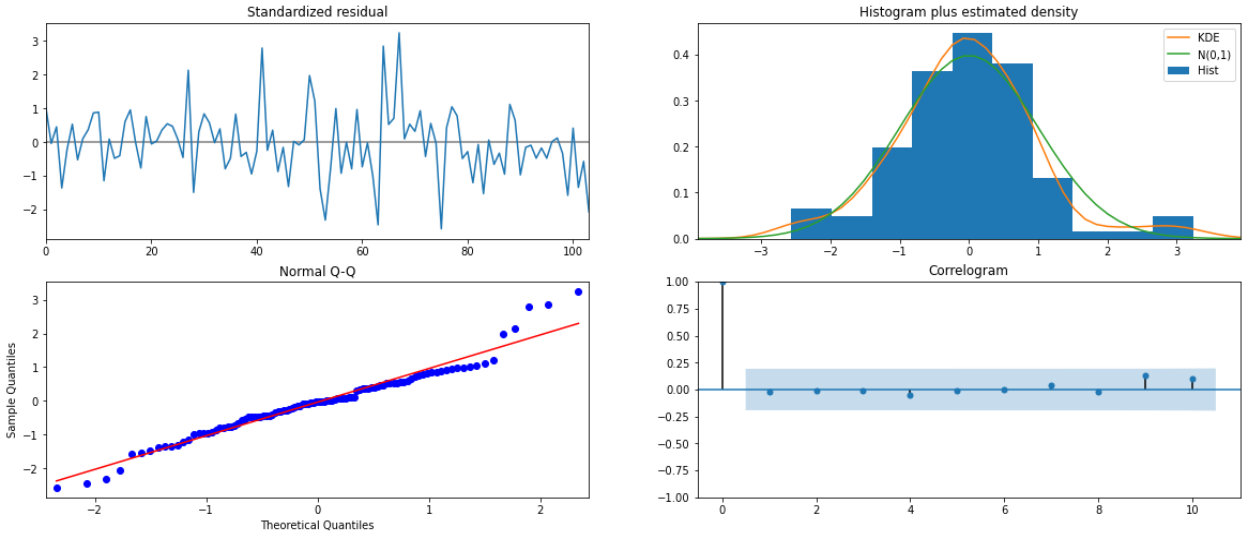
SARIMAX Results

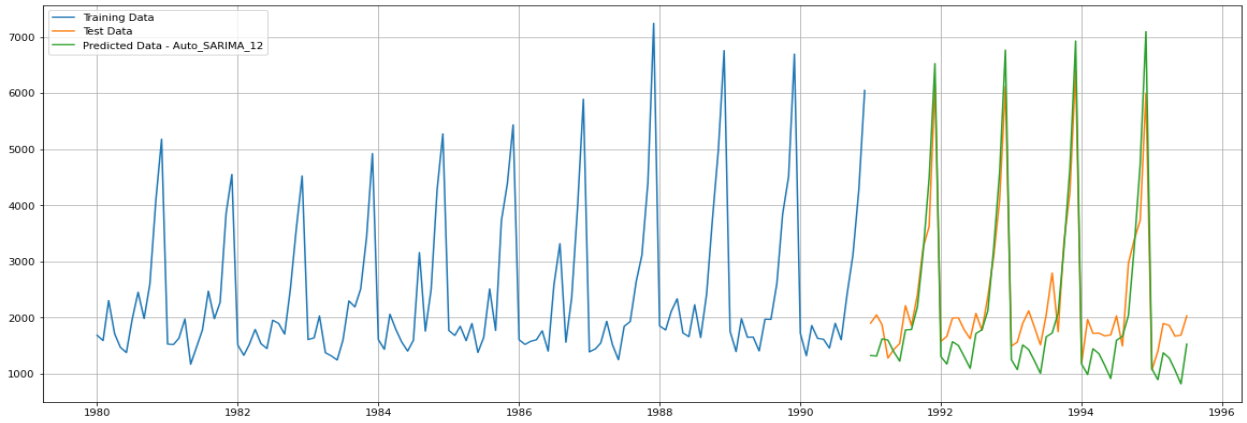
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Fri, 11 Sep 2020	AIC	1555.584
Time:	19:31:51	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6282	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.735	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.837	0.000	1.016	1.072
ma.S.L12	-0.5550	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1355	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.400	0.000	1.11e+05	1.9e+05

```
================================================================================
Ljung-Box (Q):                    23.02   Jarque-Bera (JB):                 11.72
Prob(Q):                           0.99   Prob(JB):                         0.00
Heteroskedasticity (H):            1.47   Skew:                             0.36
Prob(H) (two-sided):               0.26   Kurtosis:                         4.48
================================================================================

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

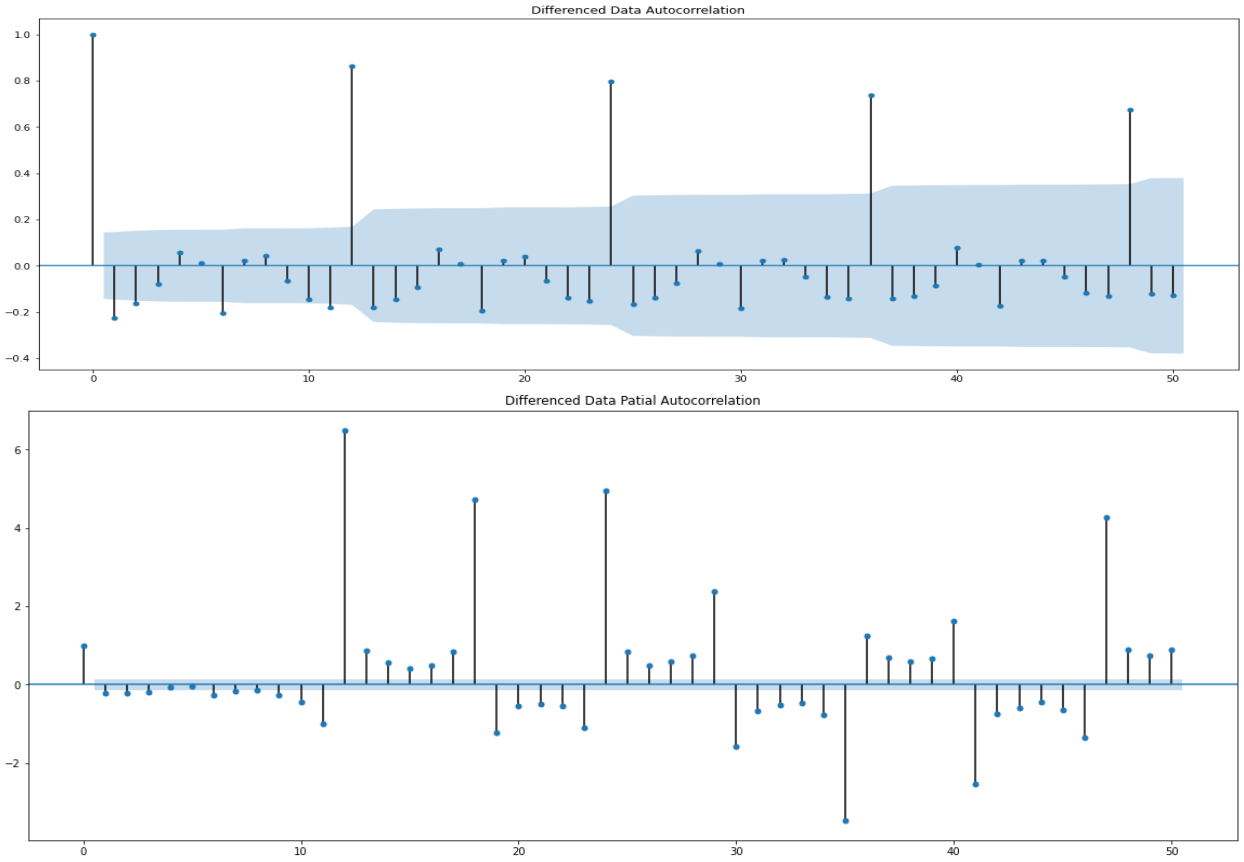




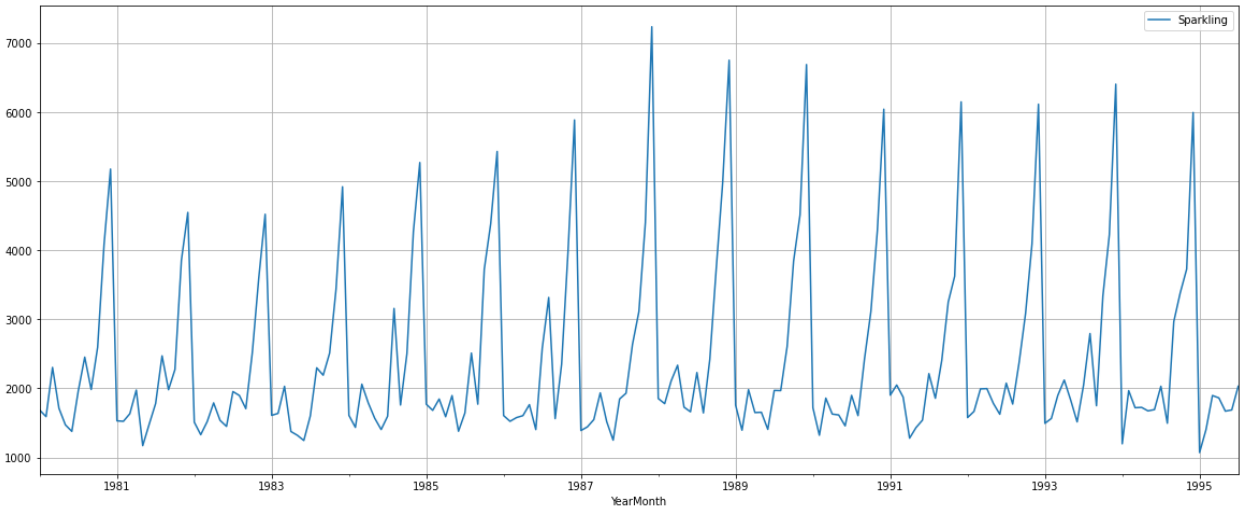
- RMSE: 528.6299166392437
- MAPE: 18.89

Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.

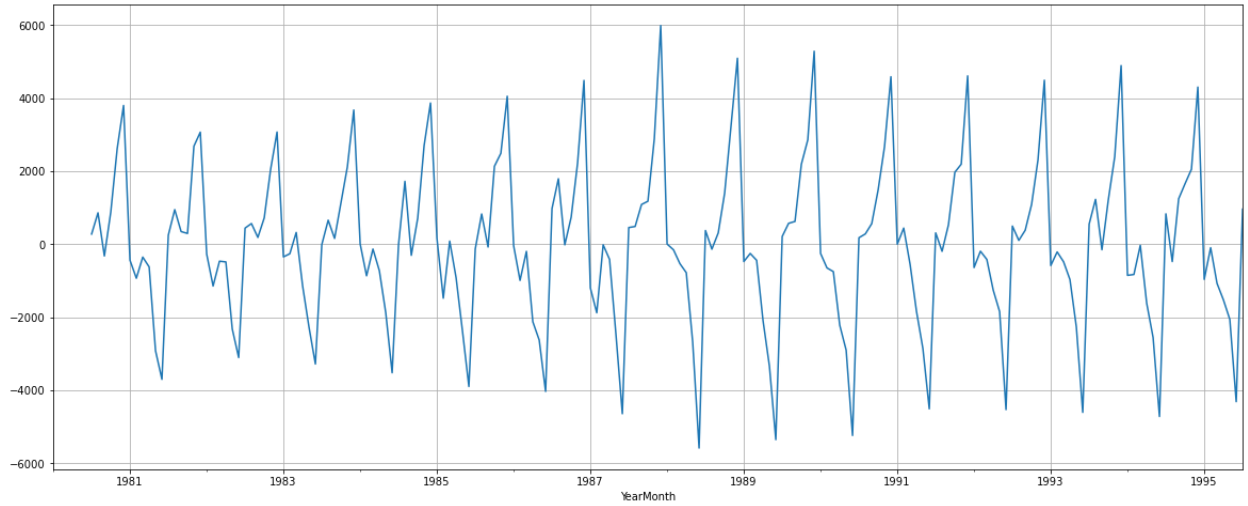
Method 14: Manual SARIMA model_6



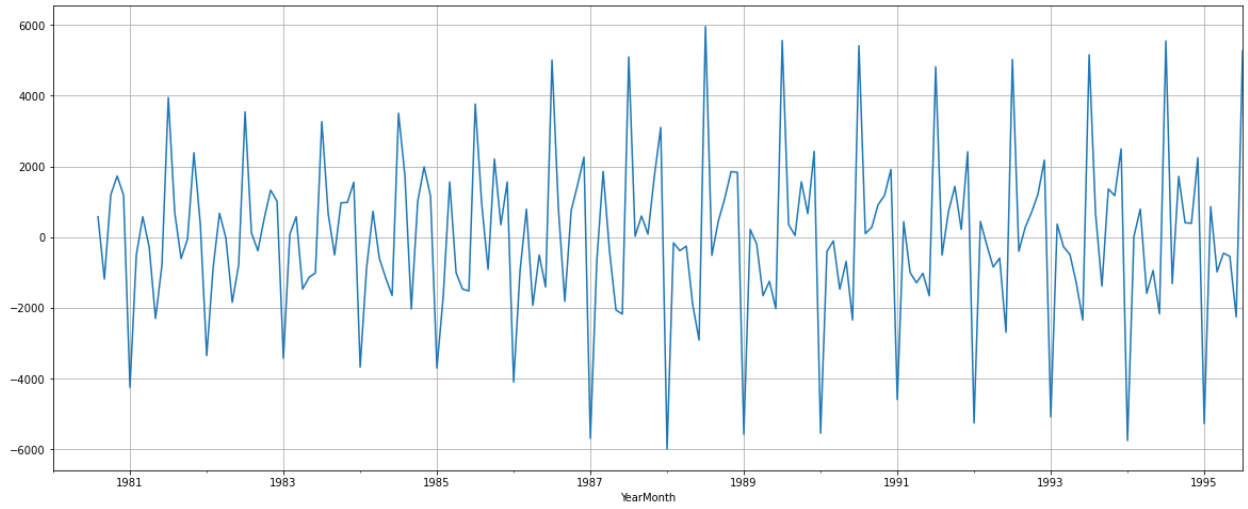
We see that our ACF plot at the seasonal interval (6) does not taper off. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.



We see that there is a slight trend and a seasonality. So, now we take a seasonal differencing and check the series.

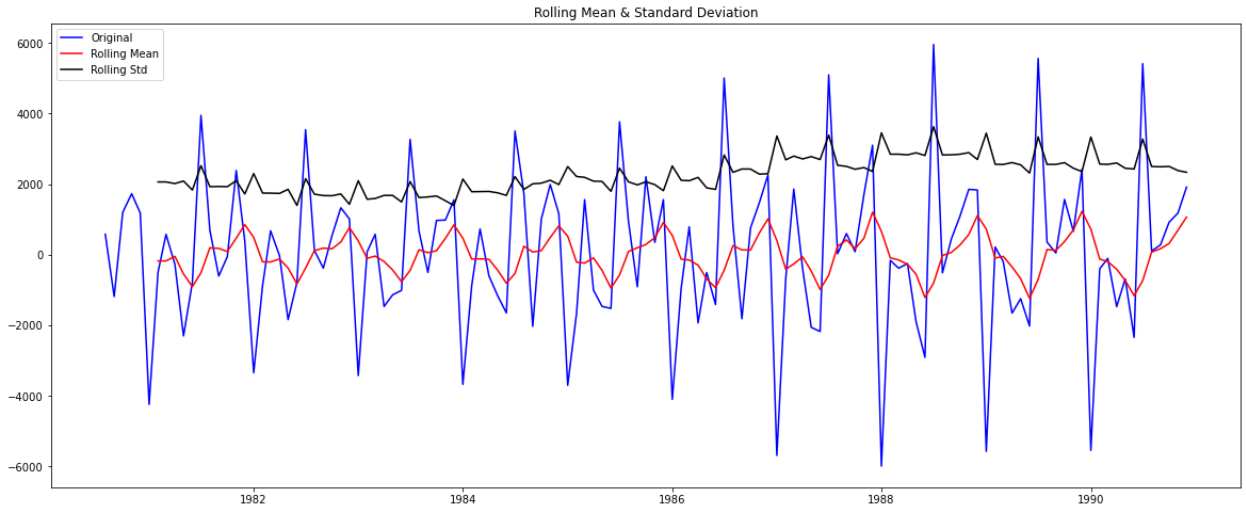


We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.



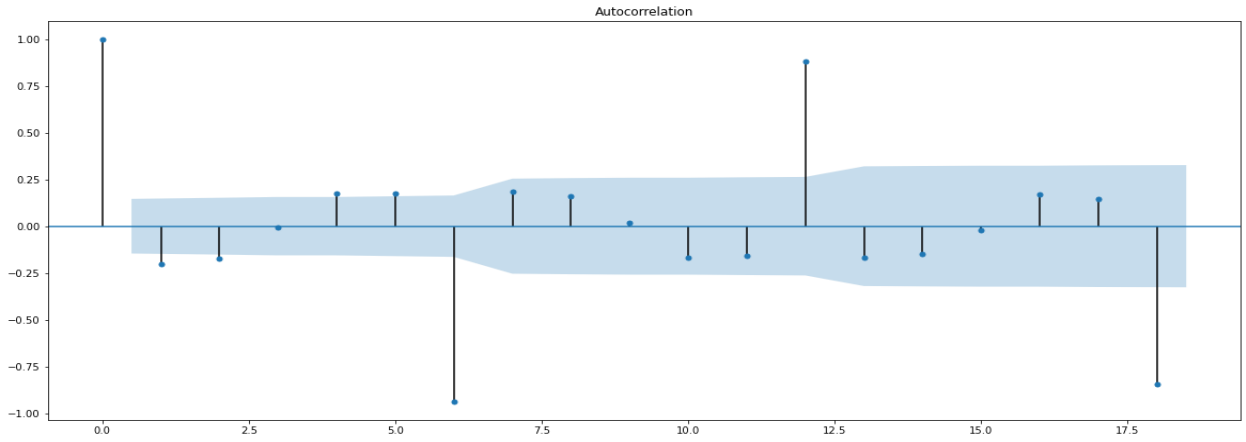
Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

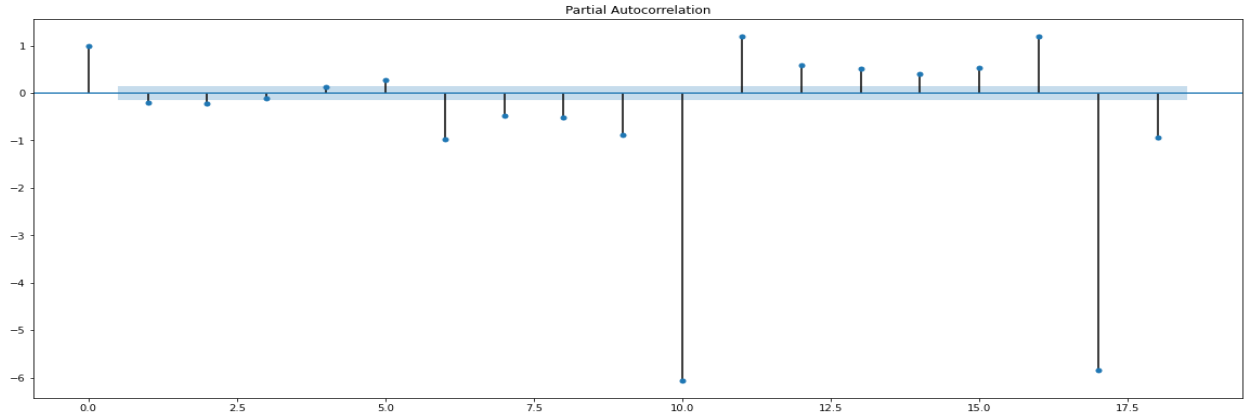
Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.




```
Results of Dickey-Fuller Test:
Test Statistic                  -7.017242e+00
p-value                          6.683657e-10
#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
```

Checking the ACF and the PACF plots for the new modified Time Series.





Here, we have taken alpha=0.05.

We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the lag at which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the lag at 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). By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0.

This is a common problem while building models by looking at the ACF and the PACF plots. But we are able to explain the model.

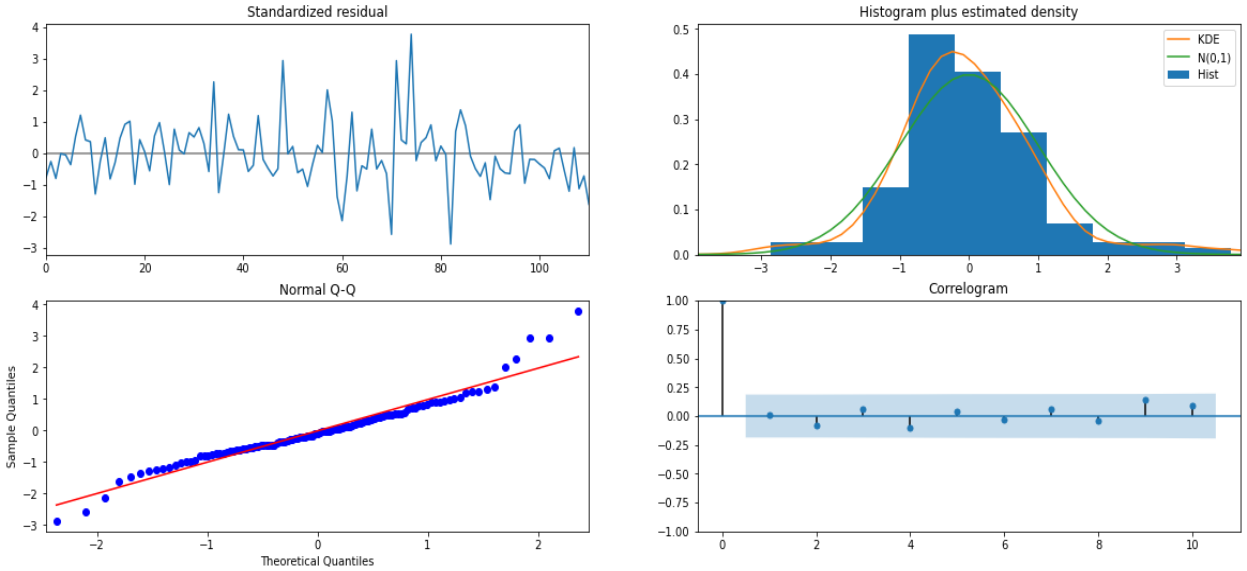
SARIMAX Results

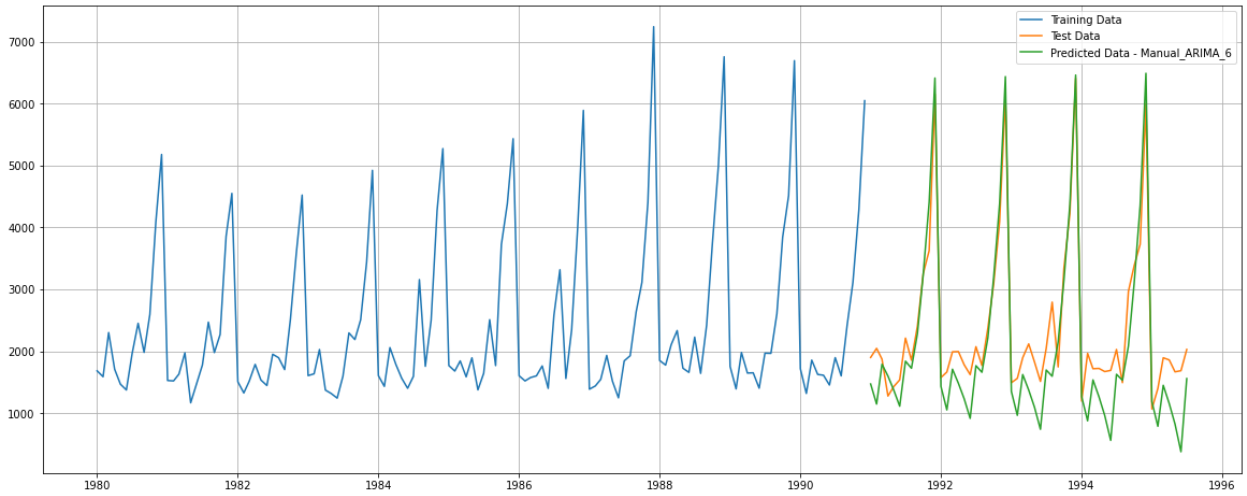
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1,1,1)x(2,1,[1,2],6)	Log Likelihood	-821.646
Date:	Fri, 11 Sep 2020	AIC	1657.293
Time:	19:32:42	BIC	1676.260
Sample:	0 - 132	HQIC	1664.987
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1365	0.116	1.177	0.239	-0.091	0.364
ma.L1	-0.9224	0.059	-15.734	0.000	-1.037	-0.808
ar.S.L6	-1.0876	0.196	-5.562	0.000	-1.471	-0.704
ar.S.L12	-0.0709	0.199	-0.356	0.722	-0.462	0.320
ma.S.L6	-0.6650	0.219	-3.038	0.002	-1.094	-0.236
ma.S.L12	-1.2388	0.348	-3.556	0.000	-1.922	-0.556
sigma2	6.918e+04	2.68e+04	2.578	0.010	1.66e+04	1.22e+05

Ljung-Box (Q):	26.78	Jarque-Bera (JB):	36.09
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	1.91	Skew:	0.66
Prob(H) (two-sided):	0.05	Kurtosis:	5.47

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

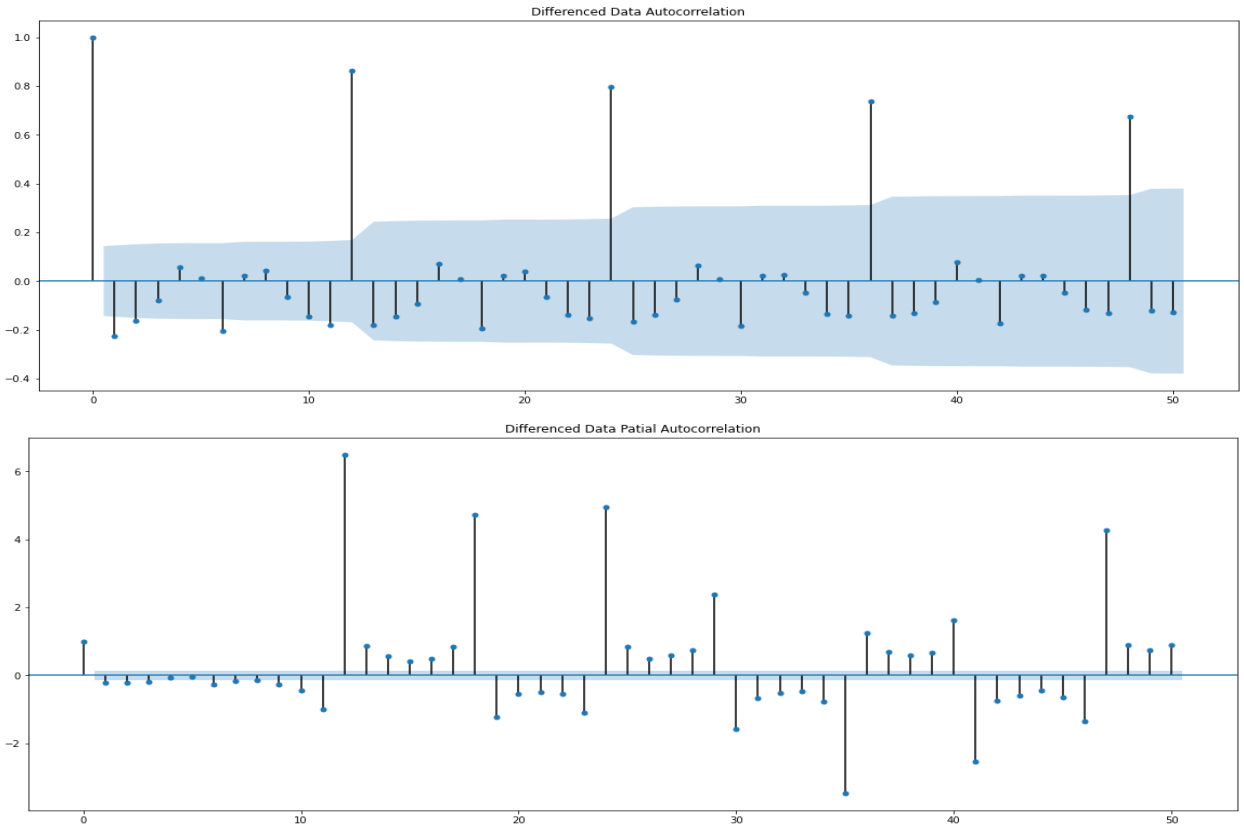




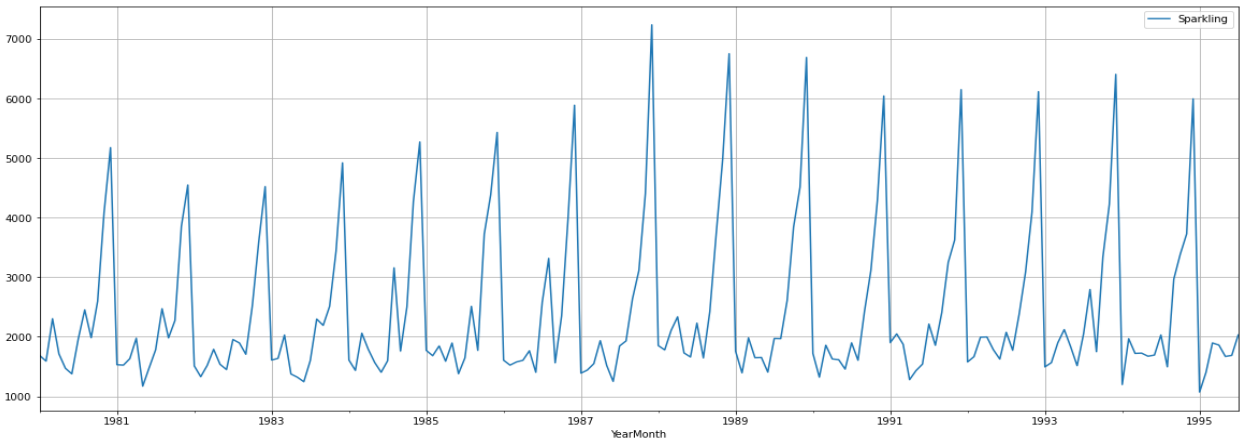
- RMSE: 547.745552622476
- MAPE: 18.52

Method 15: Manual SARIMA model_12

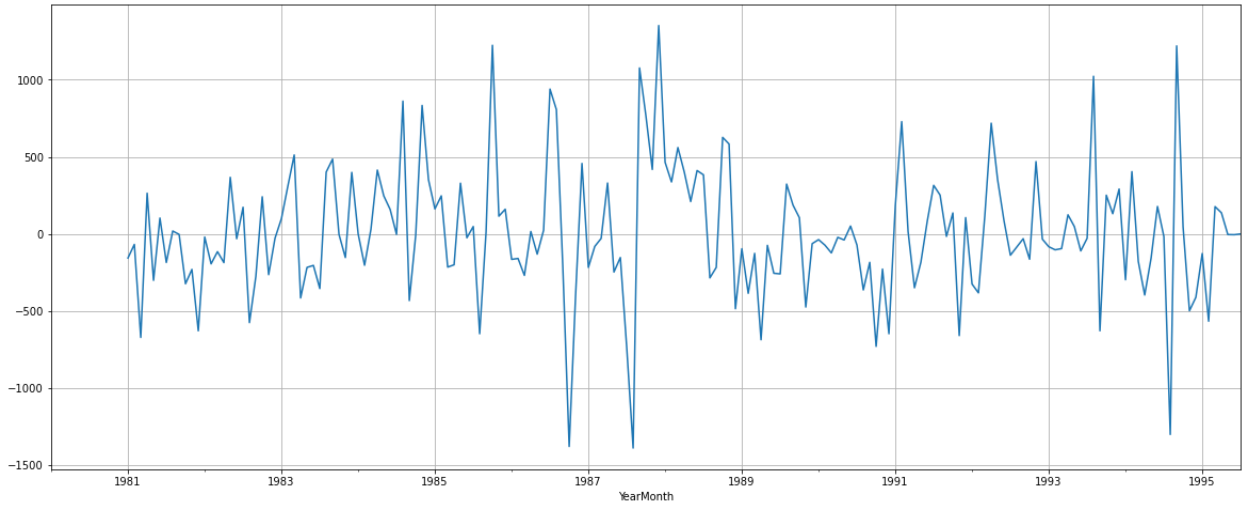
Let us look at the ACF and the PACF plots once more.

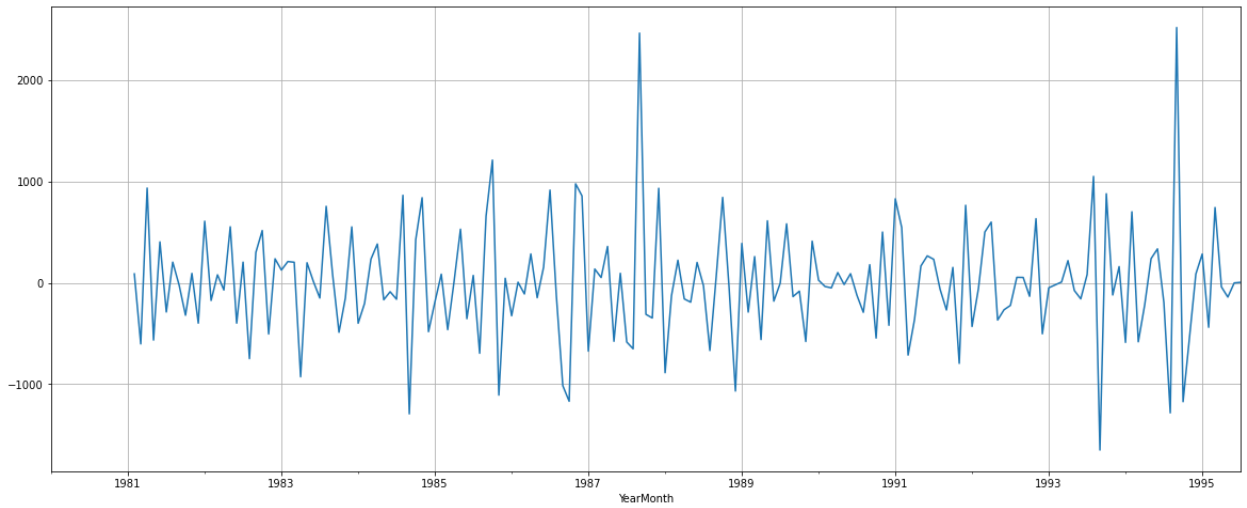


We see that our ACF plot at the seasonal interval (12) does not taper off. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.

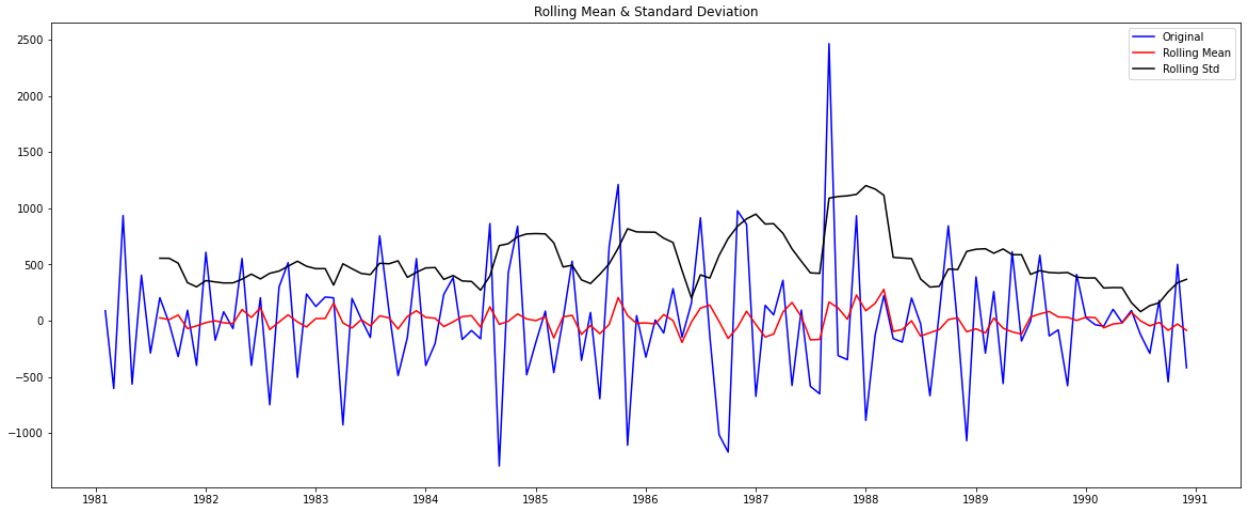


We see that there is a slight trend and a seasonality. So, now we take a seasonal differencing and check the series.



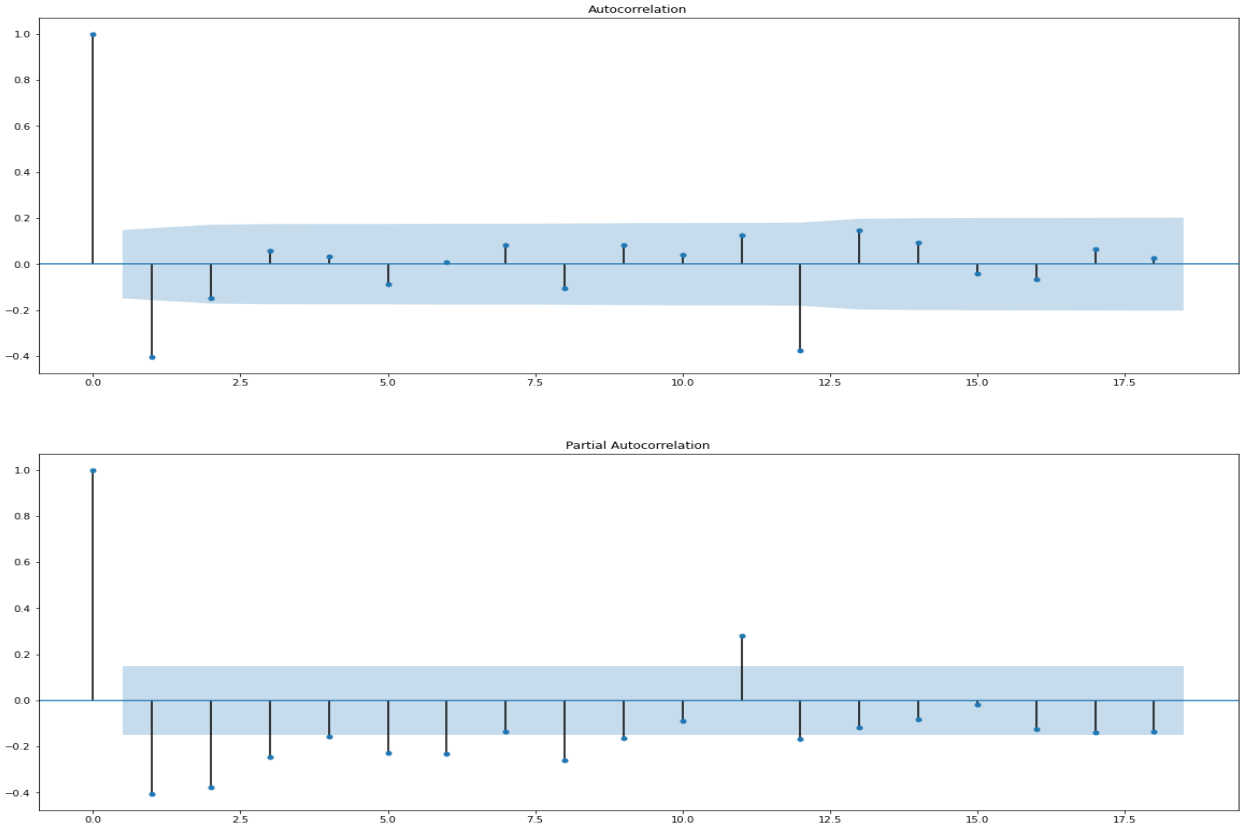


Now we see that there is almost no trend present in the data. Seasonality is only present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.



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

Checking the ACF and the PACF plots for the new modified Time Series.



Here, we have taken alpha=0.05.

- We are going to take the seasonal period as 12. We will keep the p(1) and q(1) parameters same as the ARIMA model.
- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the lag at which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the lag at 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). By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0. This is a common problem while building models by looking at the ACF and the PACF plots. But we can explain the model.

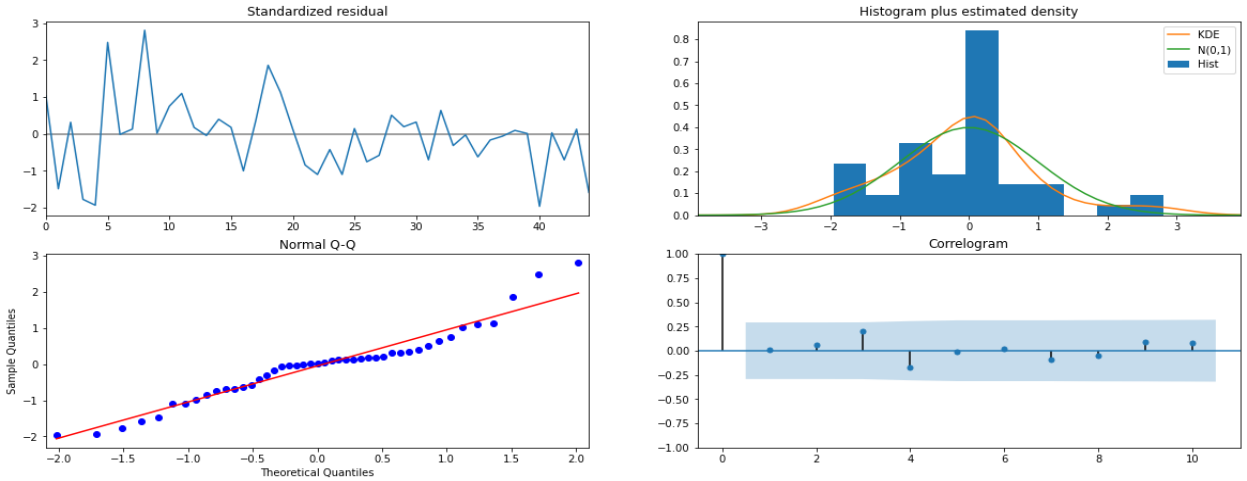
SARIMAX Results

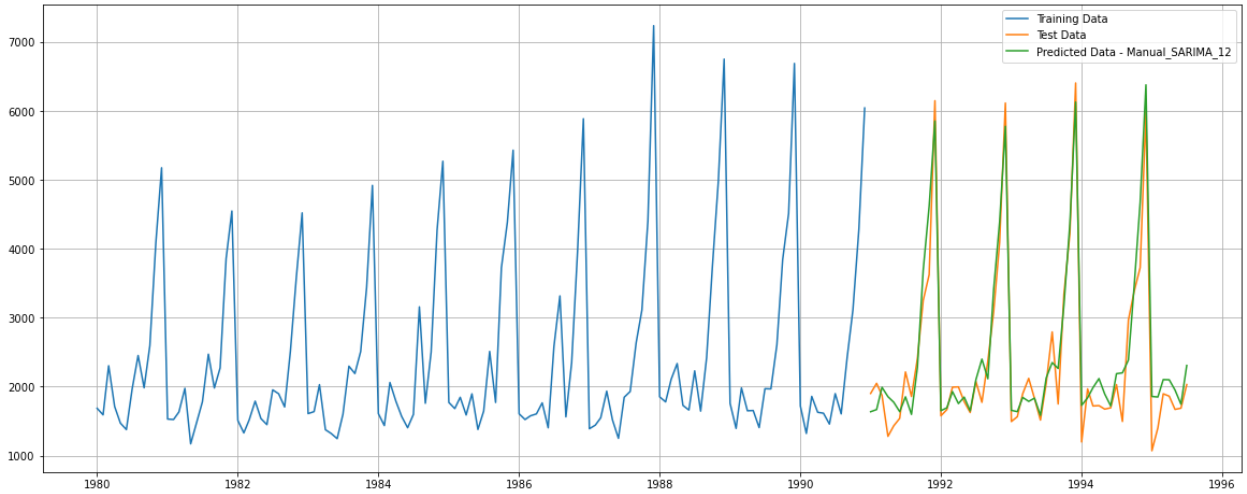
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2,1,2)x(6,1,[1],12)	Log Likelihood	-332.541
Date:	Fri, 11 Sep 2020	AIC	689.082
Time:	19:33:59	BIC	710.762
Sample:	0 - 132	HQIC	697.164
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5748	0.251	-2.292	0.022	-1.066	-0.083
ar.L2	0.1666	0.226	0.737	0.461	-0.276	0.609
ma.L1	-0.0002	84.856	-2.07e-06	1.000	-166.314	166.314
ma.L2	-1.0002	50.928	-0.020	0.984	-100.817	98.816
ar.S.L12	-0.9089	0.202	-4.489	0.000	-1.306	-0.512
ar.S.L24	-0.4304	0.280	-1.538	0.124	-0.979	0.118
ar.S.L36	-0.2761	0.297	-0.930	0.352	-0.858	0.306
ar.S.L48	-0.2867	0.222	-1.290	0.197	-0.722	0.149
ar.S.L60	-0.5158	0.326	-1.582	0.114	-1.155	0.123
ar.S.L72	-0.2801	0.265	-1.056	0.291	-0.800	0.240
ma.S.L12	0.9977	51.134	0.020	0.984	-99.222	101.218
sigma2	1.115e+05	0.001	1.02e+08	0.000	1.11e+05	1.11e+05

Ljung-Box (Q):	23.69	Jarque-Bera (JB):	3.34
Prob(Q):	0.98	Prob(JB):	0.19
Heteroskedasticity (H):	0.32	Skew:	0.49
Prob(H) (two-sided):	0.03	Kurtosis:	3.91

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 9.76e+25. Standard errors may be unstable.





- RMSE: 364.76598481914493
- MAPE: 11.86

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

Mean Absolute Percentage Error (MAPE):

- This is the same as MAE but is computed as a percentage, which is very convenient when you want to explain the quality of the model to management, [0, +∞]
- Mean absolute percentage error is a relative error measure that uses absolute values to keep the positive and negative errors from cancelling one another out and uses relative errors to enable you to compare forecast accuracy between time-series models.

  The formula for calculating the MAPE:

$$MAPE = \frac{\sum_{t=1}^{n} | \hat{Y}_t - Y_t |}{\sum_{t=1}^{n} (| \hat{Y}_t | + | Y_t |)/2}$$

  where *Yt* is the actual value of a point for a given time *t*, *n* is the total number of fitted points, and

$$\hat{Y}_t$$

  is the forecast value for the time period *t*.

Root Mean Square Error (RMSE):

- Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
- Root mean squared error is an absolute error measure that squares the deviations to keep the positive and negative deviations from canceling one another out. This measure also tends to exaggerate large errors, which can help when comparing methods.

The formula for calculating RMSE:

$$\sqrt{\frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n}}$$

Where. *Yt* is the actual value of a point for a given time *t*, *n* is the total number of fitted points, and $\hat{Y}_t$ is the fitted forecast value for the time period *t*.

	Method	RMSE	MAPE
0	RegressionOnTime	1389.135000	50.15
0	RegressionOnTimeSeasonal	1394.276000	50.11
0	Naive_model	3861.413000	152.17
0	Simple Average	1285.834000	39.22
0	moving_avg_forecast_4	1156.590000	35.96
0	moving_avg_forecast_6	1283.927000	43.86
0	moving_avg_forecast_8	1342.568000	46.46
0	moving_avg_forecast_12	1267.925000	40.19
0	SES	1275.081852	38.90
0	Holt_linear	3851.301168	152.07
0	Holt_Winter	362.742174	12.08
0	Holt_Winter M	383.176452	11.91
0	Auto_ARIMA(2,1,2)	1374.386356	48.35
0	Manual_ARIMA(3,1,2)	1378.927940	49.31
0	Auto_SARIMA(1,1,2)(2,0,2,6)	626.877238	22.55
0	Auto_SARIMA(1,1,2)(1,0,2,12)	528.629917	18.89
0	Manual_SARIMA(1,1,1)(2,1,2,6)	547.745553	18.52
0	Manual_SARIMA(2,1,2)(6,1,1,12)	364.765985	11.86

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.

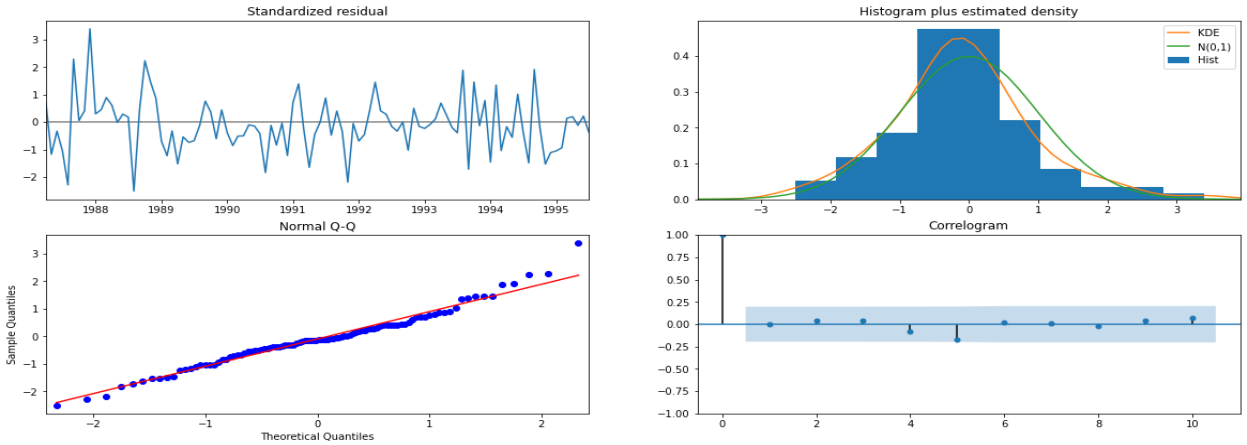
SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(2,1,2)x(6,1,[1],12)	Log Likelihood	-735.450
Date:	Fri, 11 Sep 2020	AIC	1494.900
Time:	19:35:50	BIC	1526.162
Sample:	01-01-1980 - 07-01-1995	HQIC	1507.552
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8517	0.125	-6.814	0.000	-1.097	-0.607
ar.L2	0.0593	0.128	0.463	0.643	-0.192	0.310
ma.L1	0.0433	0.162	0.267	0.789	-0.275	0.361
ma.L2	-0.9566	0.142	-6.740	0.000	-1.235	-0.678
ar.S.L12	-1.0264	0.191	-5.379	0.000	-1.400	-0.652
ar.S.L24	-0.6080	0.198	-3.072	0.002	-0.996	-0.220
ar.S.L36	-0.4319	0.168	-2.575	0.010	-0.761	-0.103
ar.S.L48	-0.2875	0.163	-1.767	0.077	-0.606	0.031
ar.S.L60	-0.2479	0.148	-1.673	0.094	-0.538	0.042
ar.S.L72	-0.2299	0.088	-2.616	0.009	-0.402	-0.058
ma.S.L12	0.5934	0.195	3.040	0.002	0.211	0.976
sigma2	1.387e+05	2.03e-06	6.82e+10	0.000	1.39e+05	1.39e+05

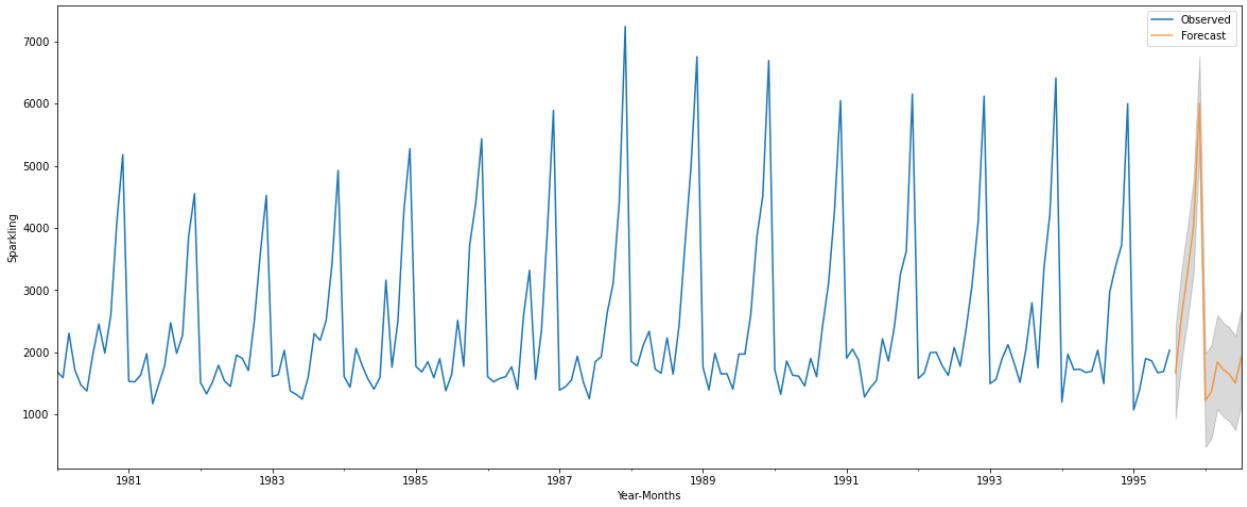
Ljung-Box (Q):	20.11	Jarque-Bera (JB):	7.93
Prob(Q):	1.00	Prob(JB):	0.02
Heteroskedasticity (H):	0.59	Skew:	0.42
Prob(H) (two-sided):	0.14	Kurtosis:	4.10

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.97e+25. Standard errors may be unstable.



Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1665.945498	373.962947	932.991591	2398.899406
1995-09-01	2581.675661	380.212526	1836.472803	3326.878519
1995-10-01	3264.898007	380.224968	2519.670764	4010.125250
1995-11-01	4005.926882	382.297884	3256.636799	4755.216966
1995-12-01	6005.683287	382.295878	5256.397135	6754.969439
1996-01-01	1222.132077	384.006522	469.493123	1974.771031
1996-02-01	1364.587584	384.005163	611.951294	2117.223874
1996-03-01	1839.780611	385.442382	1084.327423	2595.233798
1996-04-01	1716.966754	385.461019	961.477038	2472.456469
1996-05-01	1644.865104	386.691381	886.963925	2402.766283
1996-06-01	1503.601572	386.740497	745.604126	2261.599017
1996-07-01	1915.795053	387.809647	1155.702111	2675.887995

- RMSE of the Full Model 618.9984872246131
- MAPE of manual_SARIMA_12_full_data: 16.63



We see that we have certainly been able to take advantage of seasonality to get a better prediction with thinner confidence intervals. We saw that differencing on the seasonal scale helped make the model more accurate on the test data.

	Method	RMSE	MAPE
0	RegressionOnTime	1389.135000	50.15
0	RegressionOnTimeSeasonal	1394.276000	50.11
0	Naive_model	3861.413000	152.17
0	Simple Average	1285.834000	39.22
0	moving_avg_forecast_4	1156.590000	35.96
0	moving_avg_forecast_6	1283.927000	43.86
0	moving_avg_forecast_8	1342.568000	46.46
0	moving_avg_forecast_12	1267.925000	40.19
0	SES	1275.081852	38.90
0	Holt_linear	3851.301168	152.07
0	Holt_Winter	362.742174	12.08
0	Holt_Winter M	383.176452	11.91
0	Auto_ARIMA(2,1,2)	1374.386356	48.35
0	Manual_ARIMA(3,1,2)	1378.927940	49.31
0	Auto_SARIMA(1,1,2)(2,0,2,6)	626.877238	22.55
0	Auto_SARIMA(1,1,2)(1,0,2,12)	528.629917	18.89
0	Manual_SARIMA(1,1,1)(2,1,2,6)	547.745553	18.52
0	Manual_SARIMA(2,1,2)(6,1,1,12)	364.765985	11.86
0	Fulldata_Manual_SARIMA(2,1,2)(6,1,1,12)	618.998487	16.63

Inference:

As of now, we observe that Manual_SARIMA(2,1,2)(6,1,1,12)and Holt_Winter M seems to be a good fit for the data, since the RMSE (364.7, 362.7) value and MAPE (11.86,11.91) respectively is low compared to other models.

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

Comment on our Final Model (Manual_SARIMA(2,1,2)(6,1,1,12))

The summary attribute that results from the output of SARIMAX returns a significant amount of information, but we'll focus our attention on the table of coefficients. The coef column shows the weight (i.e. importance) of each feature and how each one impacts the time series. The P>|z| column informs us of the significance of each feature weight. Here, approx. 8 weight has a p-value lower or close to 0.05, so it is reasonable to retain all of them in our model.

When fitting seasonal ARIMA models (and any other models for that matter), it is important to run model diagnostics to ensure that none of the assumptions made by the model have been violated. The plot_diagnostics object allows us to quickly generate model diagnostics and investigate for any unusual behaviour. Our primary concern is to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If the seasonal ARIMA model does not satisfy these properties, it is a good indication that it can be further improved.

In this case, our model diagnostics suggests that the model residuals are normally distributed based on the following:

- In the top right plot, we see that the red KDE line follows closely with the N(0,1) line (where N(0,1)) is the standard notation for a normal distribution with mean 0 and standard deviation of 1). This is a good indication that the residuals are normally distributed.
- The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N (0, 1). Again, this is a strong indication that the residuals are normally distributed.
- The residuals over time (top left plot) do not display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

These observations lead us to conclude that our model produces a satisfactory fit that could help us understand our time series data and forecast future values.

Although we have a satisfactory fit, some parameters of our seasonal ARIMA model could be changed to improve our model fit. For example, our grid search only considered a restricted set of parameter combinations, so we may find better models if we widened the grid search.

Measures that the company should be taking for future sales.

Since there is no trend but seasonality. The company's wine sales would be high in November and December in future years according to the model prediction.

The company needs to improve their sales in rest of the months.

Wine selections on the menu should include more than the region of origin, the type, and the year. Select words like fruity, bold, earthy, light, sweet, dry and dessert to describe the actual taste. It will help customers narrow the options and increase sales.

Try giving your customers opportunities to try selections with these simple strategies:

- Open the bar for tasting events. You do not have to offer samples of every wine, but occasionally opening the bar for a wine tasting or wine pairing event can bring in customers on a slow night.
- Bring in a few bottles of something new every month. Promote these selections to your email list. Invite them in for a special glass of your featured wine.
- Always have a featured wine. Pair it with a signature or special dish and make it a special for the week or month. Do not forget to share it with your customers on Facebook.

The End