PROJECT REPORT
Time Series Forecasting

PREEJA RAJESH
PGP – DSBA

Contents

Problem Statement 2:

The data of Rose wine sales in the 20th century is to be analysed. 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: Rose.csv

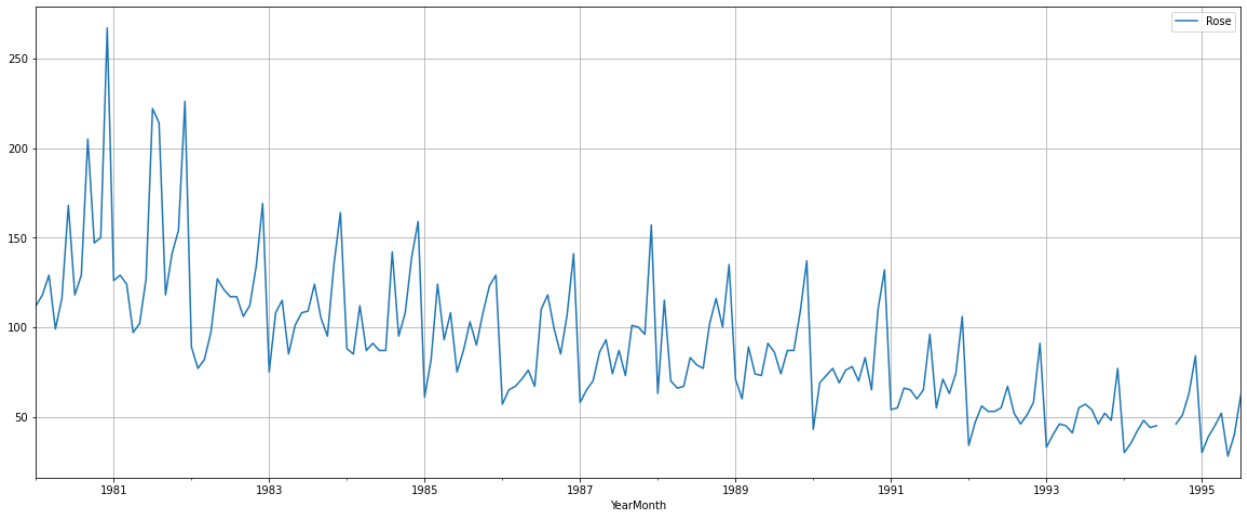
1. Read the data as an appropriate Time Series data and plot the data.

Data set:

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

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

2.1 Exploratory Data Analysis:



- **We can see there is down trend and seasonal pattern both associated with it.**
- **There are total 187 rows and 1 column 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   Rose    185 non-null    float64

dtypes: float64(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.

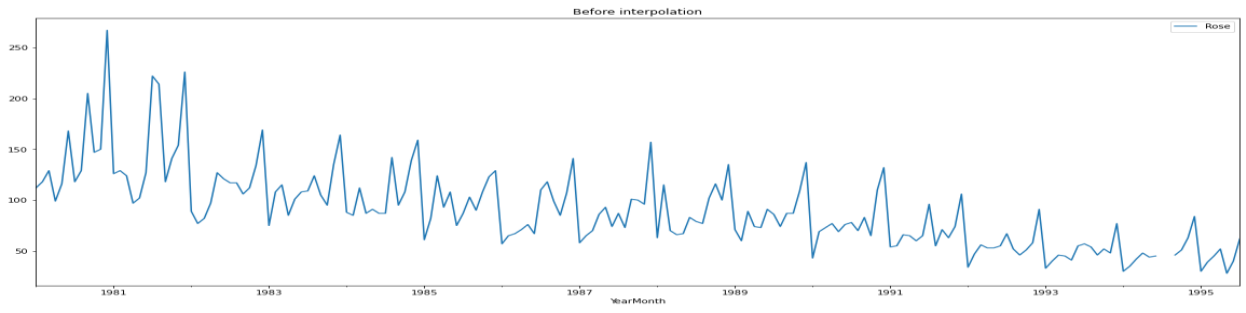
The basic measures of descriptive statistics tell us how the Sales have varied across years. But remember, for this measure of descriptive statistics we have averaged over the whole data without taking the time component into account.

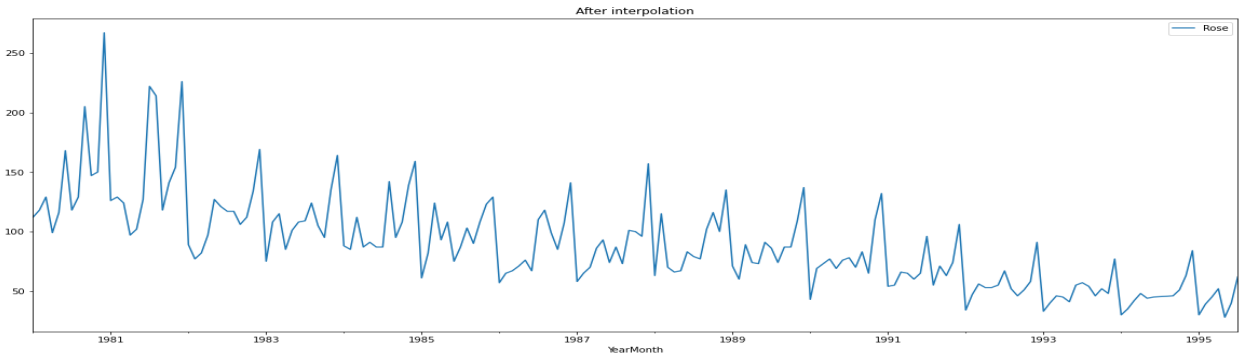
Checks if any value in the dataframe is null.

There are 2 null values present in the dataset

- No missing data is allowed in time series as data is ordered.
- It is simply not possible to shift the series to fill in the gaps.

Hence we need to do interpolation.





Imputed value

YearMonth	Rose
1994-07-01	45.333333

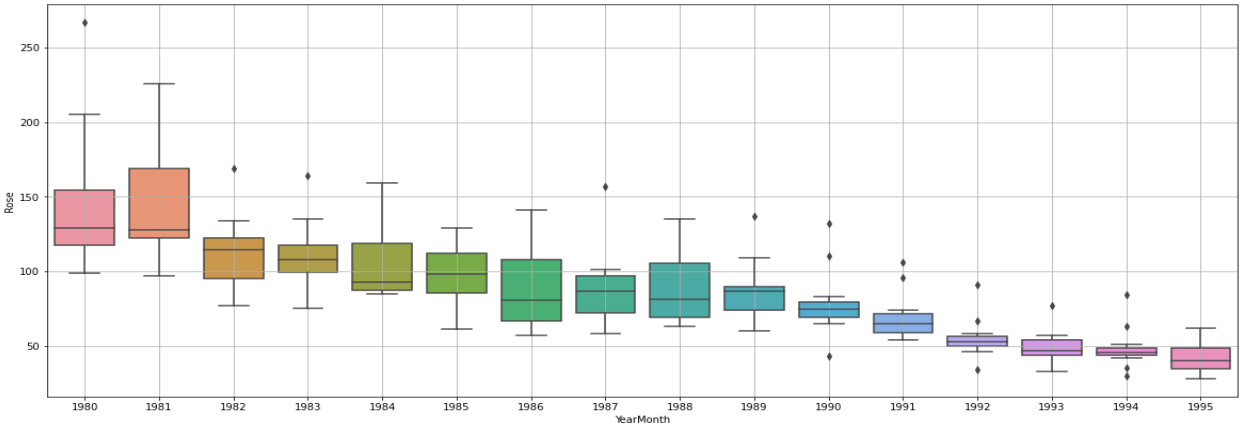
YearMonth	Rose
1994-08-01	45.666667

Original value

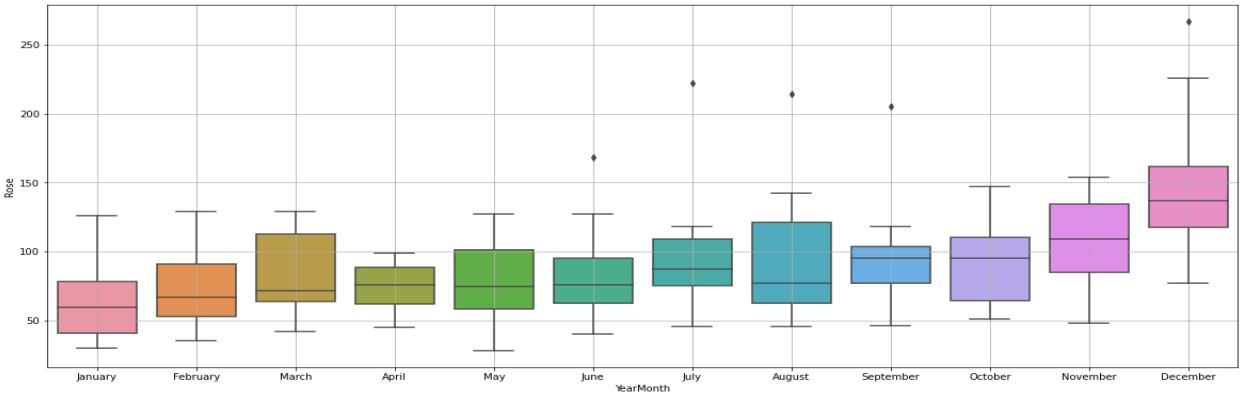
YearMonth	Rose
1994-07-01	NaN

YearMonth	Rose
1994-08-01	NaN

Yearly Boxplot:



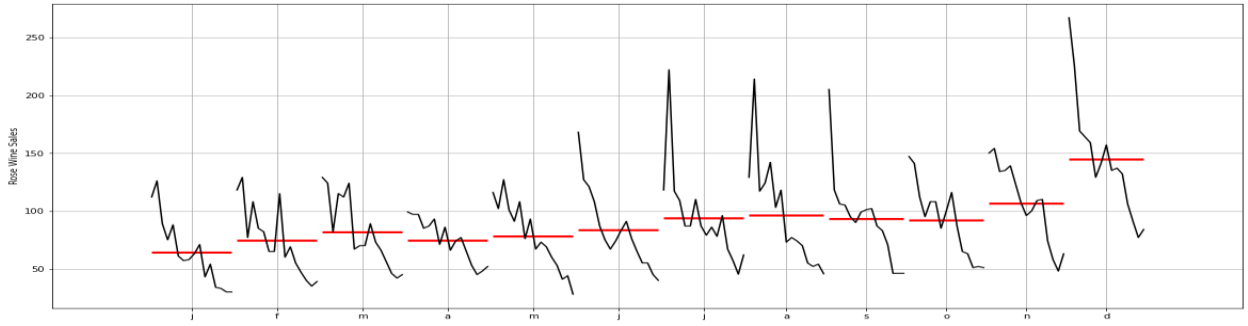
Monthly Plot:



Inference:

- There is a clear distinction of 'Rose 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.

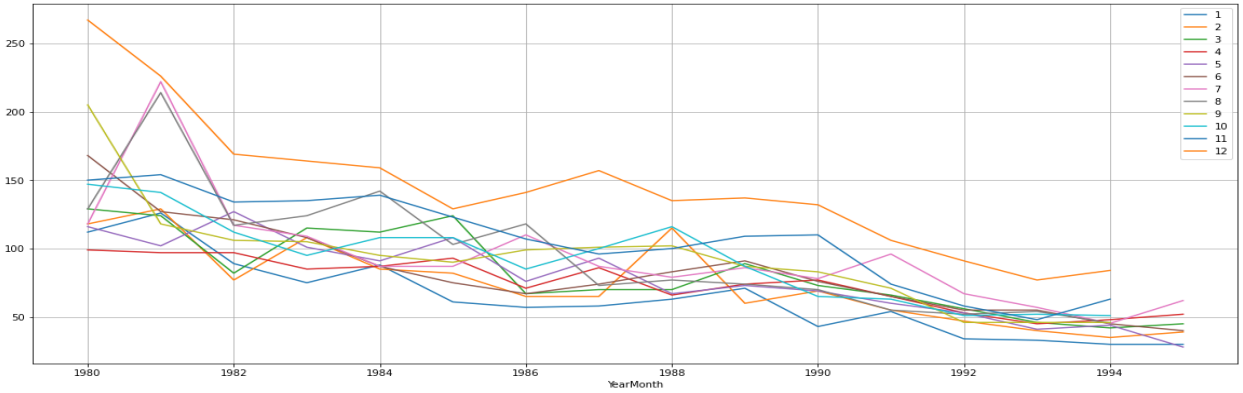


Inference:

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

Plot a graph of monthly Sparkling Wine Sales across years:

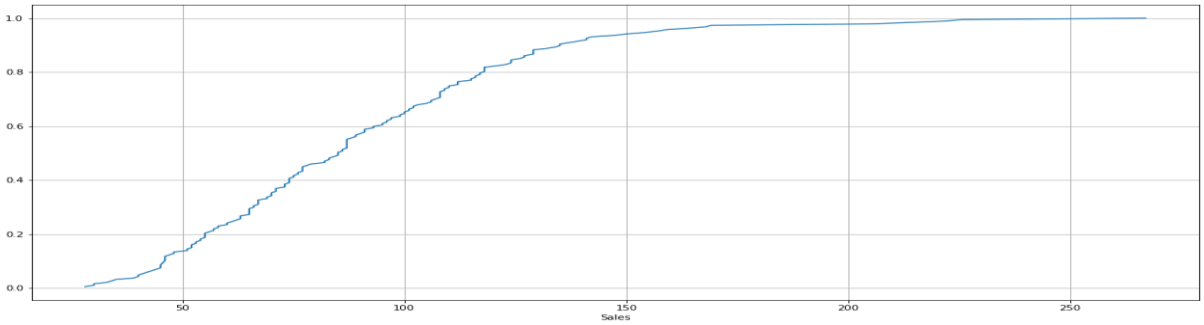
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	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



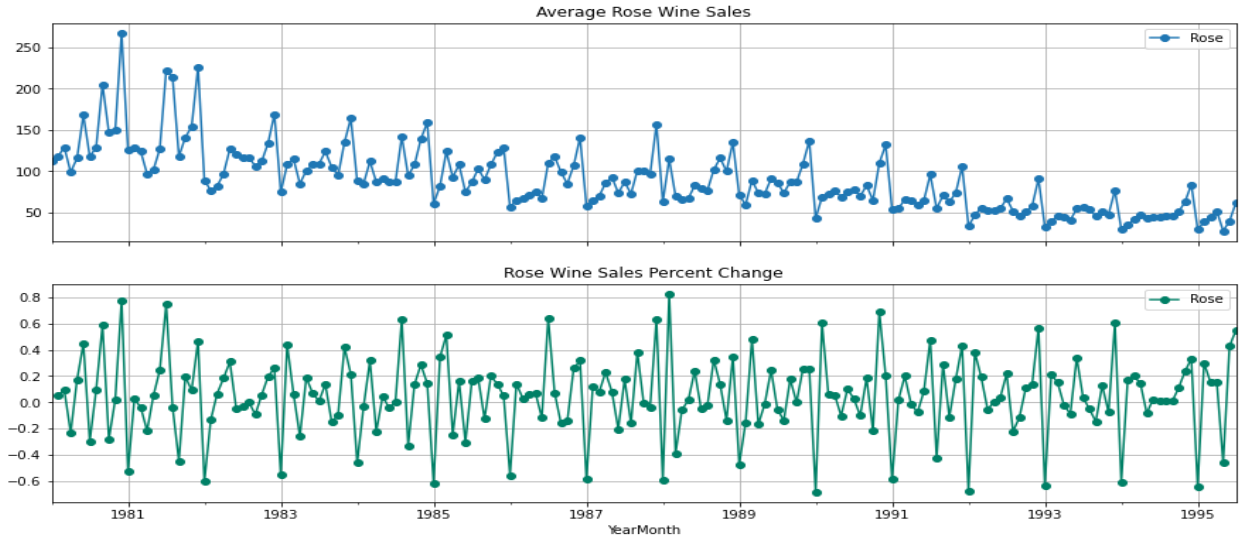
Inference:

- December month has the highest sales of Rose 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.

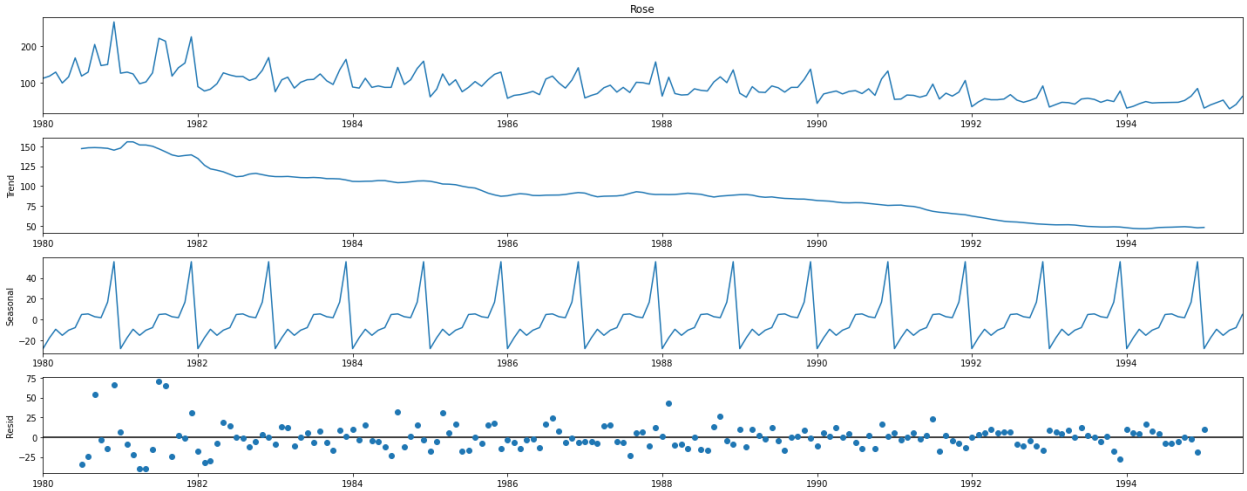


Inference:

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

2.2 Decompose the Time Series and plot the different components.

2.2.1Additive Decomposition:

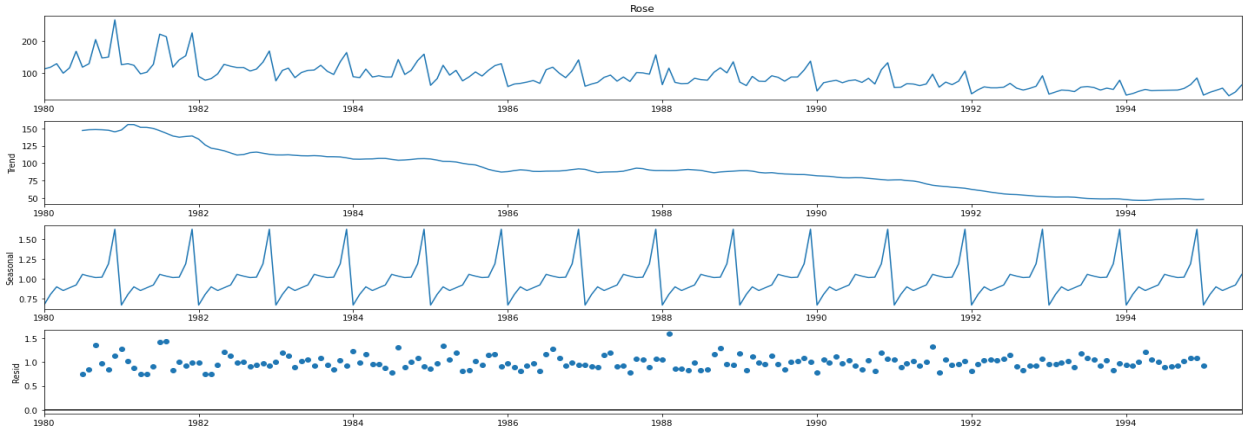


Trend		Seasonality		Residual	
YearMonth		YearMonth		YearMonth	
01-01-1980	NaN	01-01-1980	-27.908647	01-01-1980	NaN
01-02-1980	NaN	01-02-1980	-17.435632	01-02-1980	NaN
01-03-1980	NaN	01-03-1980	-9.28583	01-03-1980	NaN
01-04-1980	NaN	01-04-1980	-15.09833	01-04-1980	NaN
01-05-1980	NaN	01-05-1980	-10.196544	01-05-1980	NaN
01-06-1980	NaN	01-06-1980	-7.678687	01-06-1980	NaN
01-07-1980	147.083333	01-07-1980	4.896908	01-07-1980	-33.980241
01-08-1980	148.125	01-08-1980	5.499686	01-08-1980	-24.624686
01-09-1980	148.375	01-09-1980	2.774686	01-09-1980	53.850314
01-10-1980	148.083333	01-10-1980	1.871908	01-10-1980	-2.955241
01-11-1980	147.416667	01-11-1980	16.846908	01-11-1980	-14.263575
01-12-1980	145.125	01-12-1980	55.713575	01-12-1980	66.161425
Freq: MS,		Freq: MS,		Freq: MS,	
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:

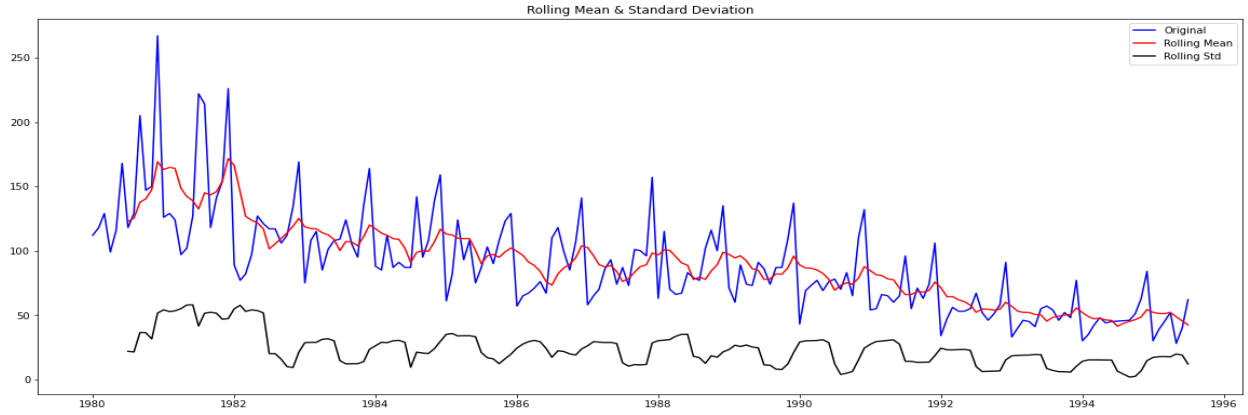


Trend		Seasonality		Residual	
YearMonth		YearMonth		YearMonth	
01-01-1980	NaN	01-01-1980	0.670111	01-01-1980	NaN
01-02-1980	NaN	01-02-1980	0.806163	01-02-1980	NaN
01-03-1980	NaN	01-03-1980	0.901164	01-03-1980	NaN
01-04-1980	NaN	01-04-1980	0.854024	01-04-1980	NaN
01-05-1980	NaN	01-05-1980	0.889415	01-05-1980	NaN
01-06-1980	NaN	01-06-1980	0.923985	01-06-1980	NaN
01-07-1980	147.083333	01-07-1980	1.058038	01-07-1980	0.758258
01-08-1980	148.125	01-08-1980	1.035881	01-08-1980	0.84072
01-09-1980	148.375	01-09-1980	1.017648	01-09-1980	1.357674
01-10-1980	148.083333	01-10-1980	1.022573	01-10-1980	0.970771
01-11-1980	147.416667	01-11-1980	1.192349	01-11-1980	0.853378
01-12-1980	145.125	01-12-1980	1.628646	01-12-1980	1.129646
Freq: MS,		Freq: MS,		Freq: MS,	
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:

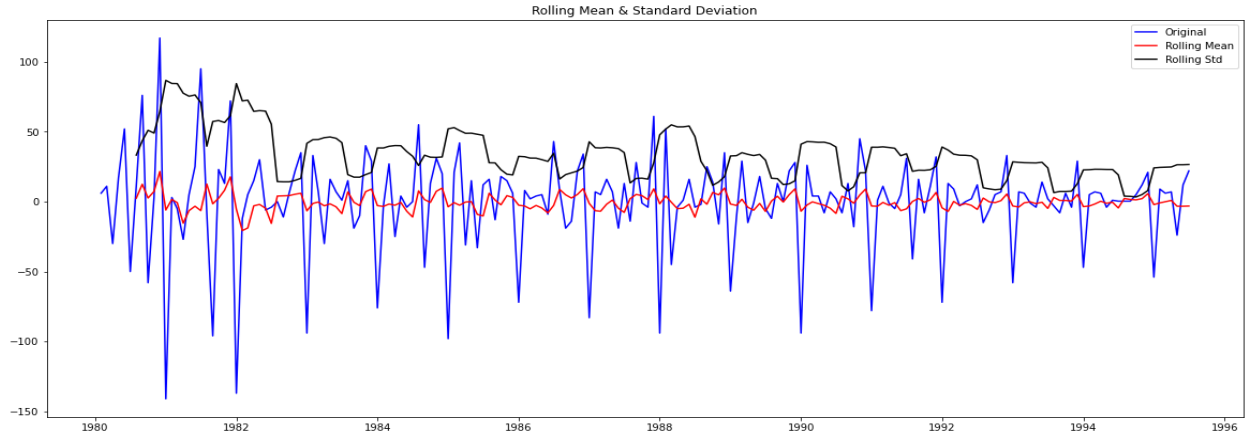
- Test Statistic -1.876699
- p-value 0.343101
- #Lags Used 13.000000
- Number of Observations Used 173.000000
- Critical Value (1%) -3.468726
- Critical Value (5%) -2.878396
- Critical Value (10%) -2.575756

dtype: float64

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:

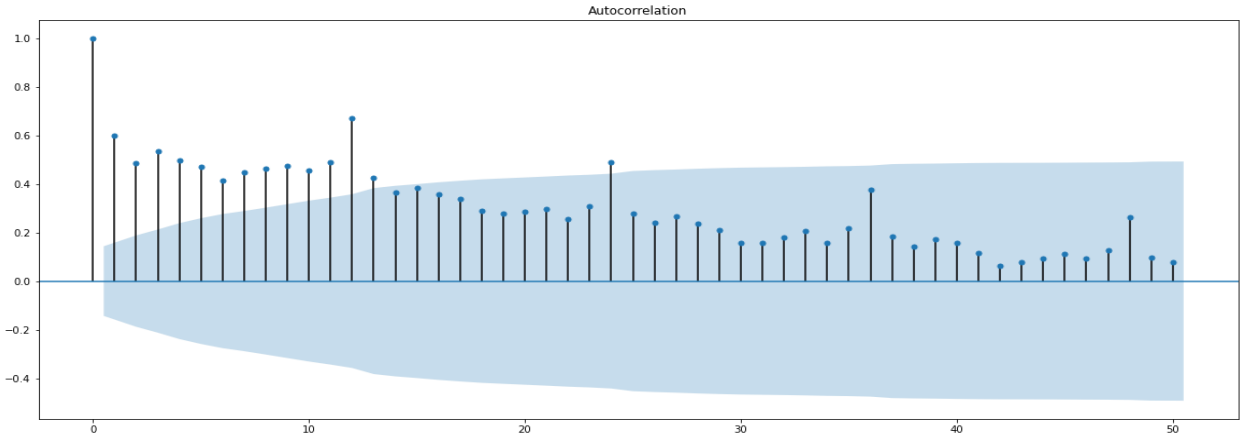
- Test Statistic -8.044392e+00
- p-value 1.810895e-12
- #Lags Used 1.200000e+01
- Number of Observations Used 1.730000e+02
- Critical Value (1%) -3.468726e+00
- Critical Value (5%) -2.878396e+00
- Critical Value (10%) -2.575756e+00

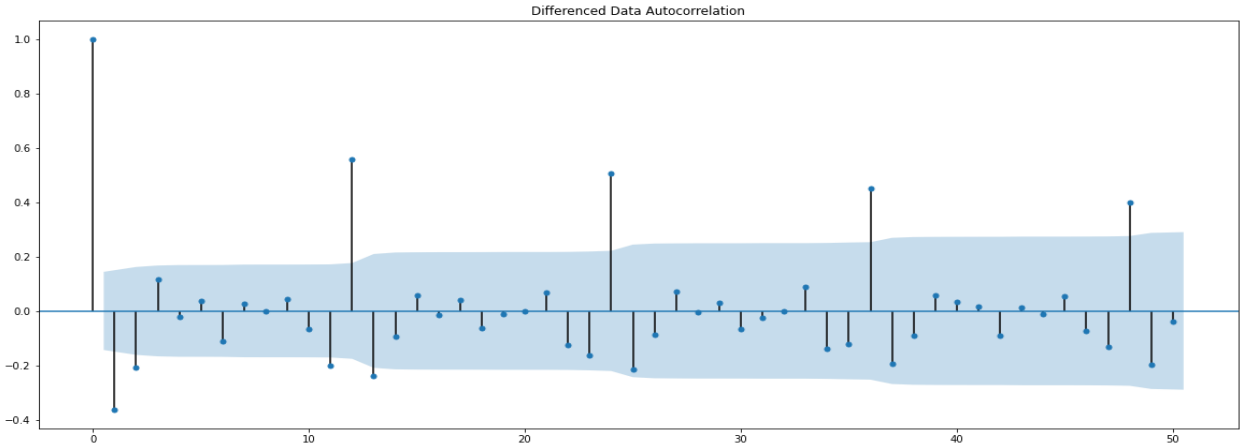
dtype: float64

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!

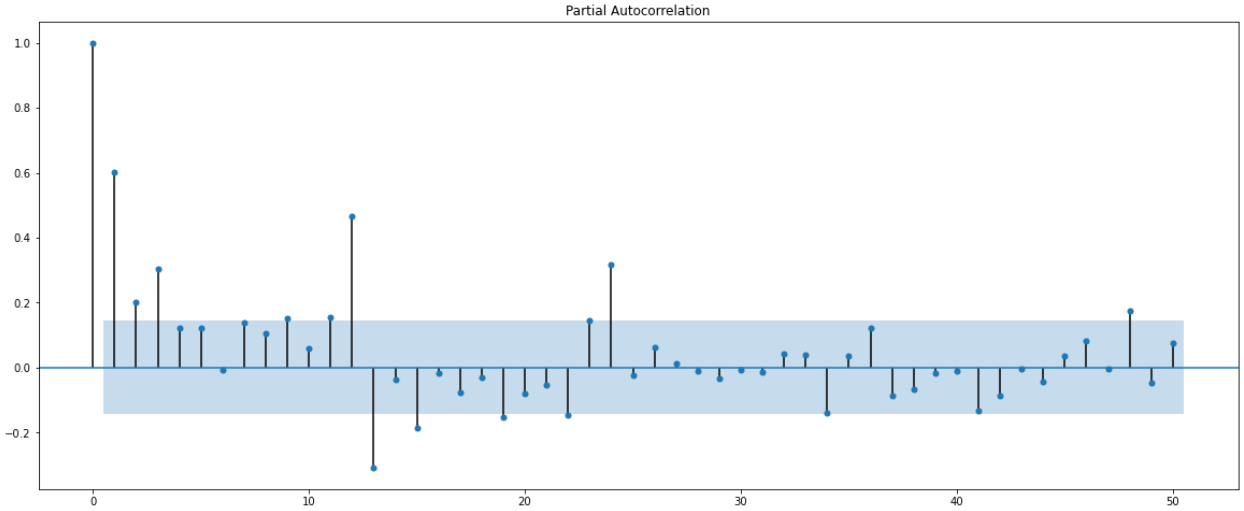
Plot the Autocorrelation function plots on the whole data.

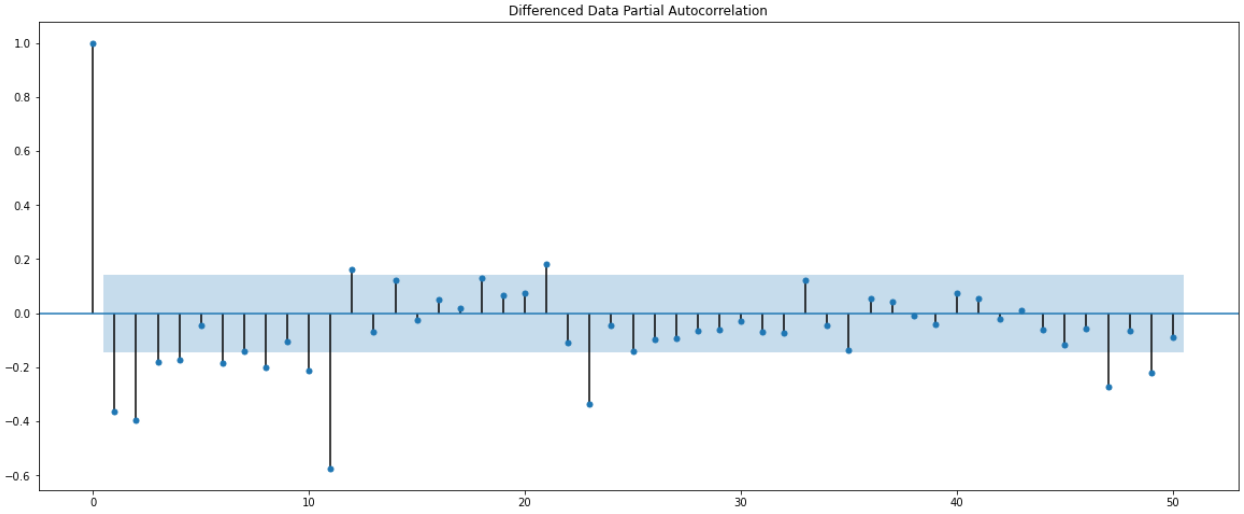




We get the order of MA term or 'q' from ACF plot. Here the order of MA term is 2 from the differenced ACF plot.

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 (4,5,6,7,8) 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	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

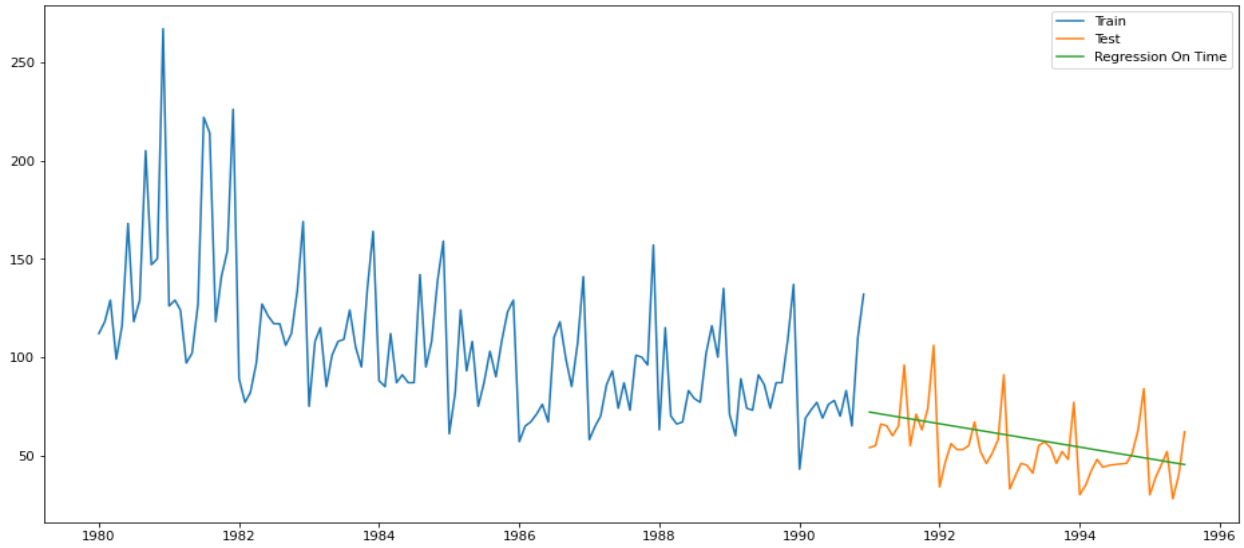
Last few rows of Test Data

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

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,

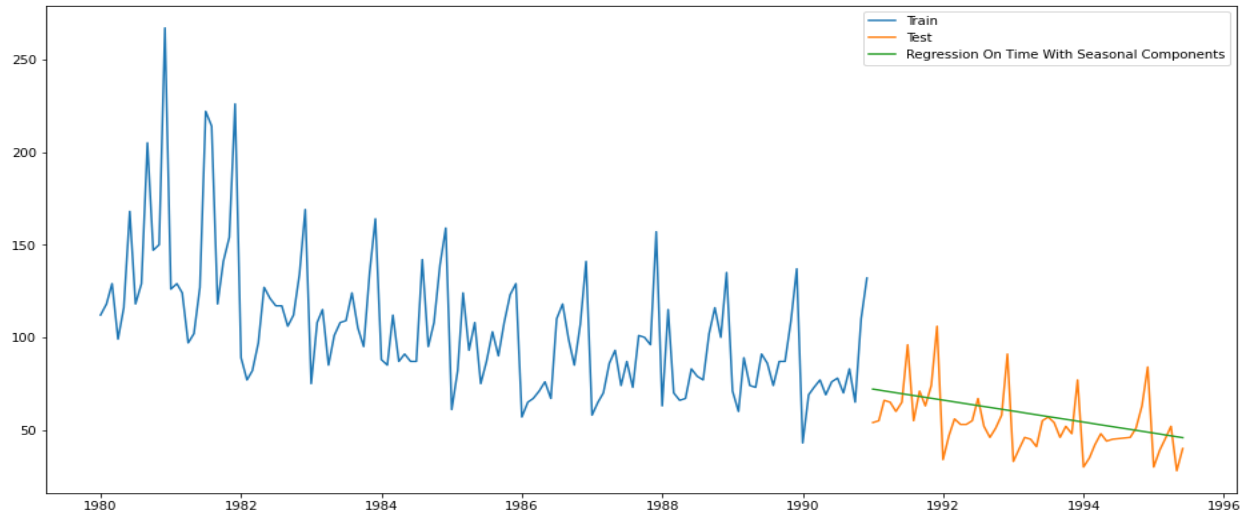
- RMSE is 15.269
- MAPE is 22.82

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



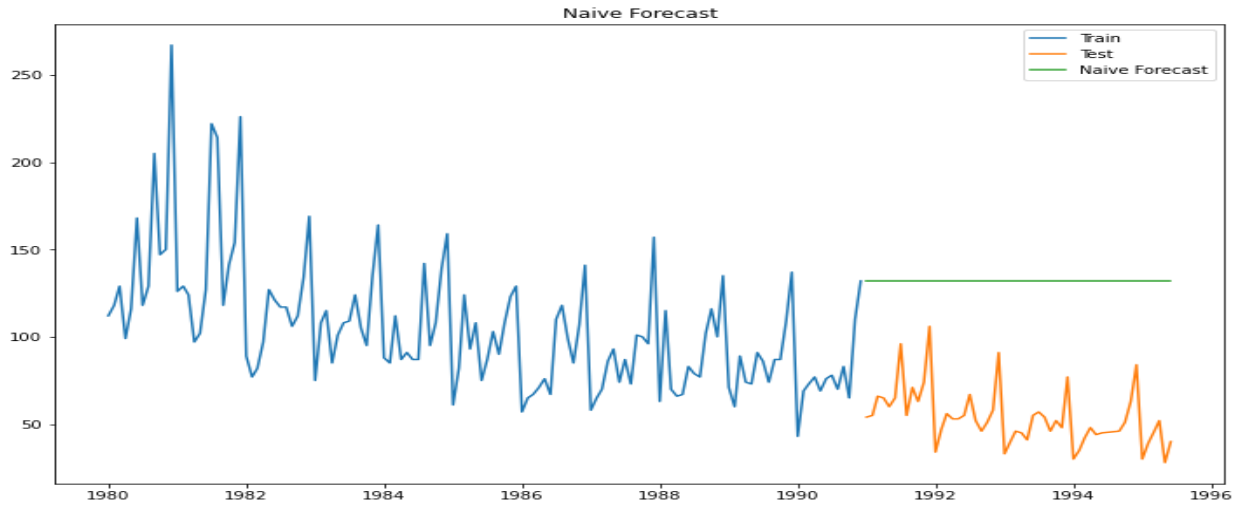
For Regression on Time Seasonal,

- RMSE is 15.243
- MAPE is 22.73

Inference:

- Output is same to the above model.

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



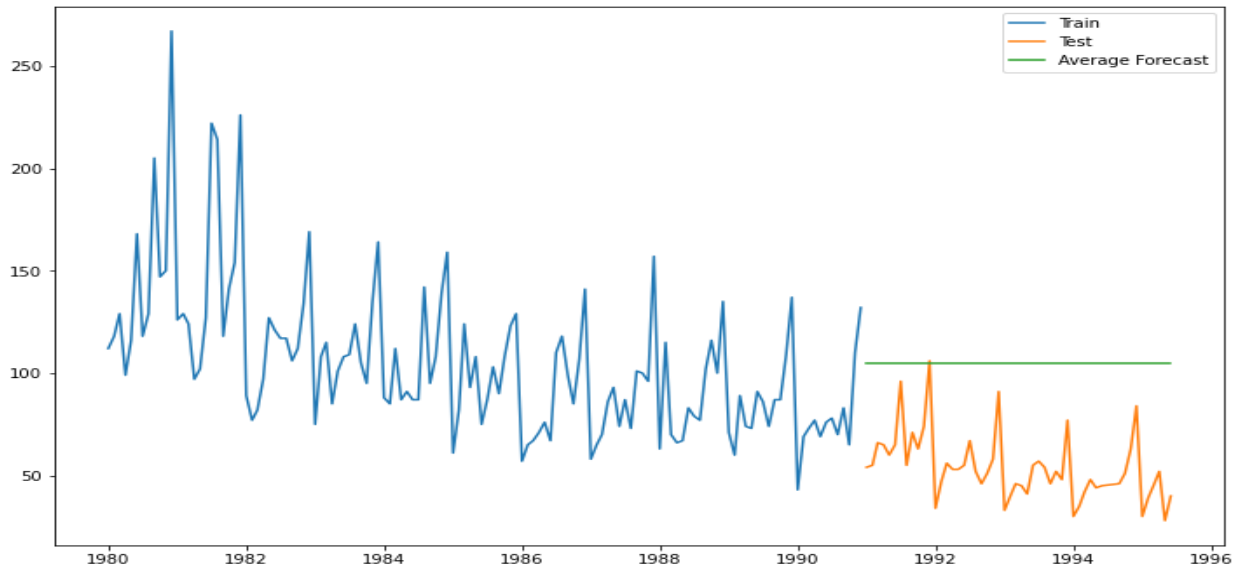
For Naive model,

- RMSE is 79.88
- MAPE is 145.79

Inference:

- We can infer from the RMSE and MAPE values and the graphs above, that Naive model is not suited for datasets with high variability.
- RegressionOnTime method is best suited for this type of 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 53.636
- MAPE is 95.48

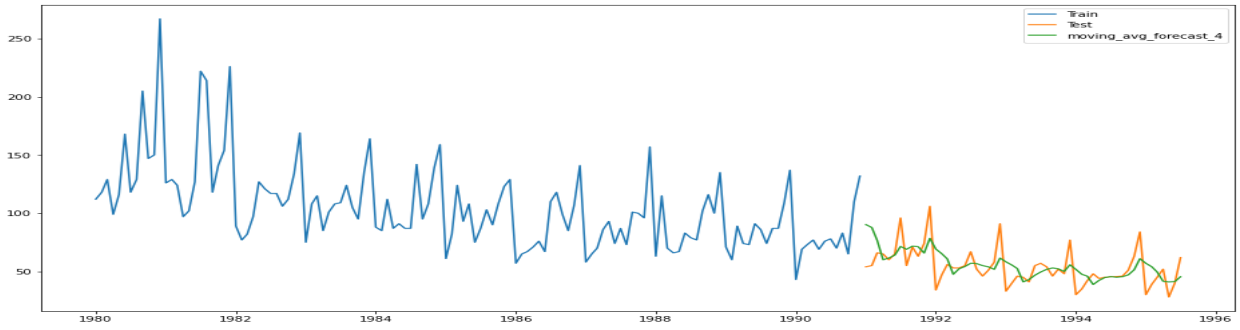
Inference:

- 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 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 or not.

Method 5: Moving Average (MA)

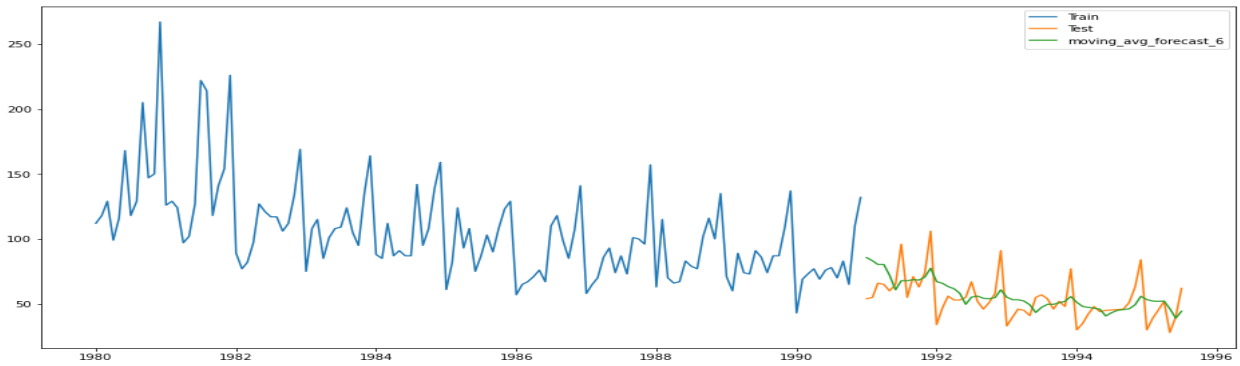
For Moving Average model, moving_avg_forecast_4

- RMSE is 14.451
- MAPE is 19.49



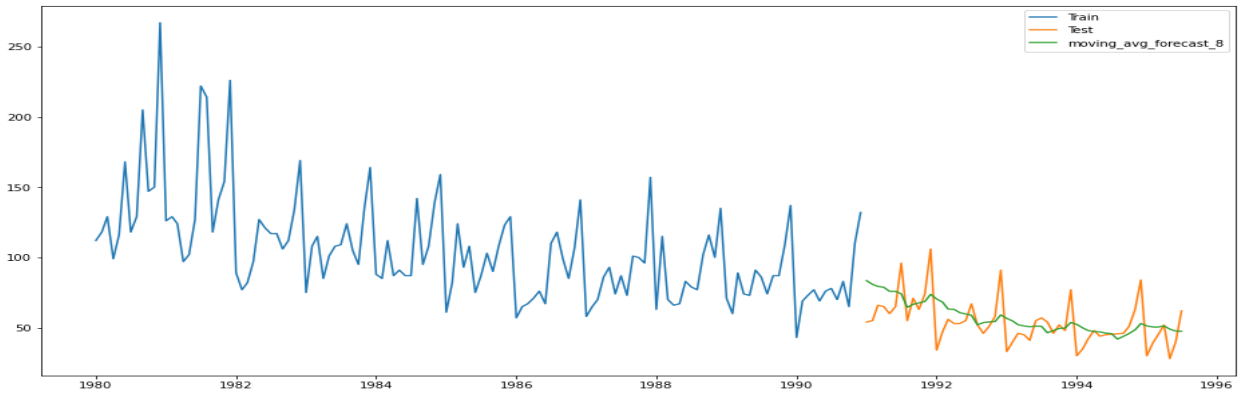
For Moving Average model, moving_avg_forecast_6

- RMSE is 14.566
- MAPE is 20.82



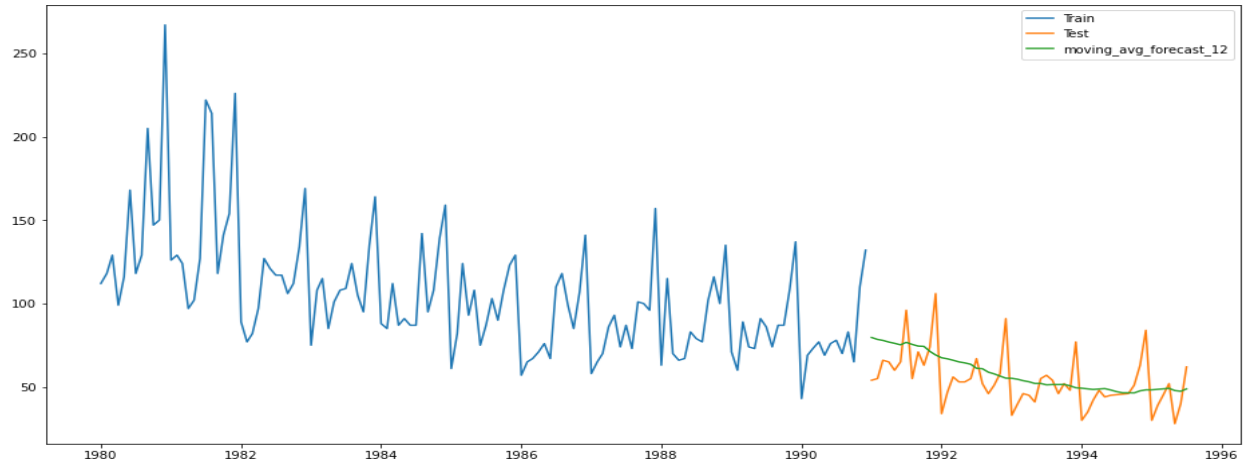
For Moving Average model, moving_avg_forecast_8

- RMSE is 14.805
- MAPE is 21.06



For Moving Average model, moving_avg_forecast_12

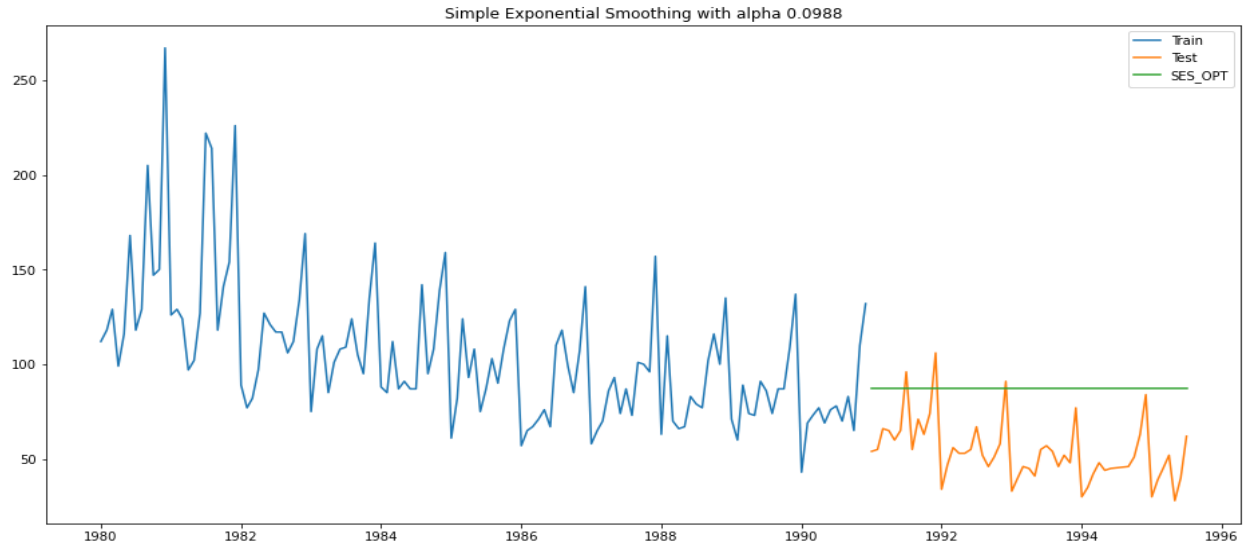
- RMSE is 15.236
- MAPE is 22.07



Method 6: Simple Exponential Smoothing

== Simple Exponential Smoothing Parameters ==

- Smoothing Level 0.1061
- Initial Level 76.6557



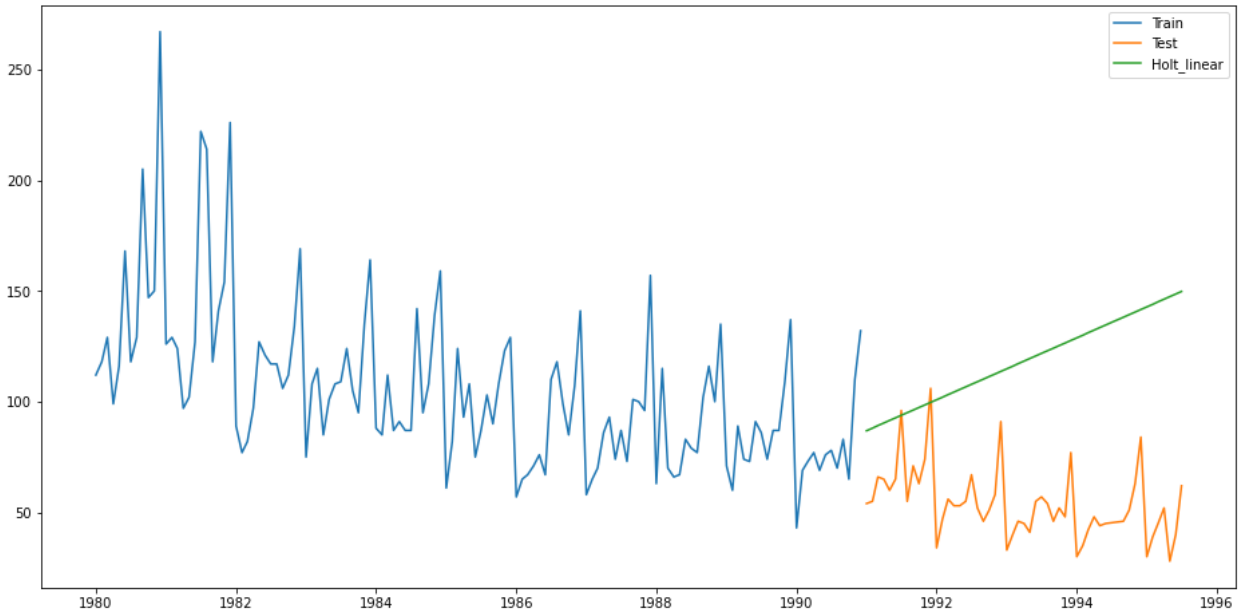
For alpha = 0.00,

- RMSE is 36.7963
- MAPE is 63.88

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

==Holt model Exponential Smoothing Parameters ==

- Smoothing Level 0.1579
- Smoothing Slope 0.1579
- Initial Level 112.0



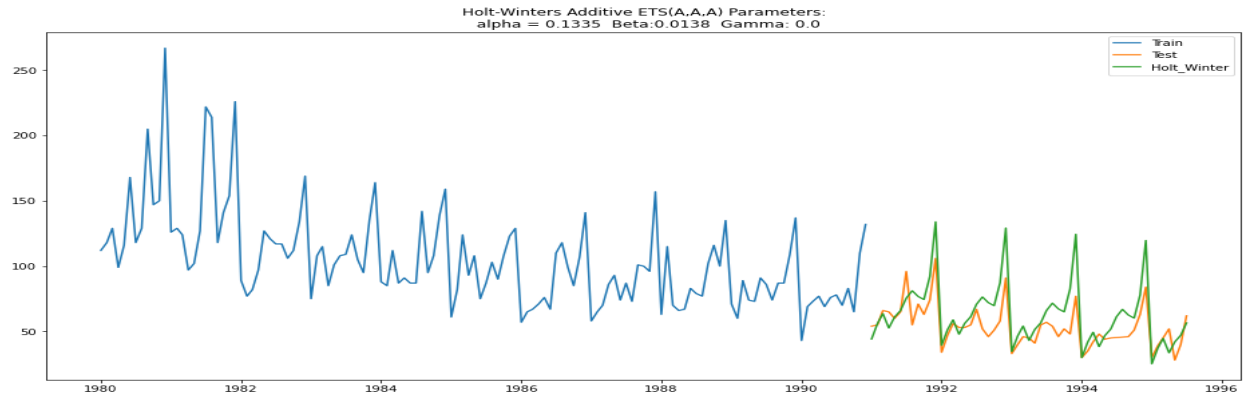
For alpha = 0.65,

- RMSE is 70.5725
- MAPE is 120.25

Method 8: Holt-Winters Method - Additive seasonality

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

- Smoothing Level: 0.1335
- Smoothing Slope: 0.0138
- Smoothing Seasonal: 0.0
- Initial Level: 76.4124
- Initial Slope: 0.0
- Initial Seasons: [38.6866 51.0194 58.9935 48.3253
  57.1198 62.5497 72.4366 78.509
  74.4762 72.5421 90.6124 132.8721]



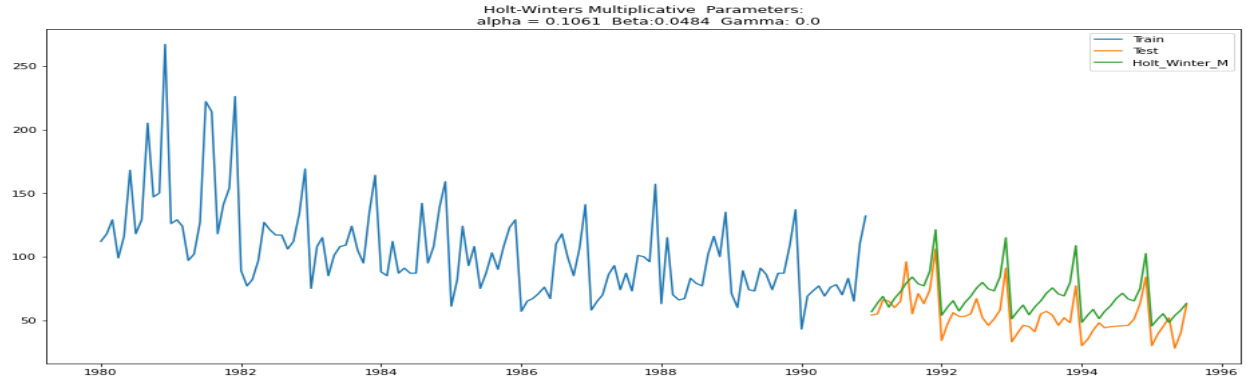
For Holt Winter alpha = 0.09, beta = 0.00, gamma = 0.48,

- 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: 0.1061
- Smoothing Slope: 0.0484
- Smoothing Seasonal: 0.0
- Initial Level: 76.6557
- Initial Slope: 0.0
- Initial Seasons:[1.4755 1.6593 1.8057 1.5889 1.7782 1.926
  2.1165 2.2514 2.1169 2.0811 2.4093 3.3045]



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

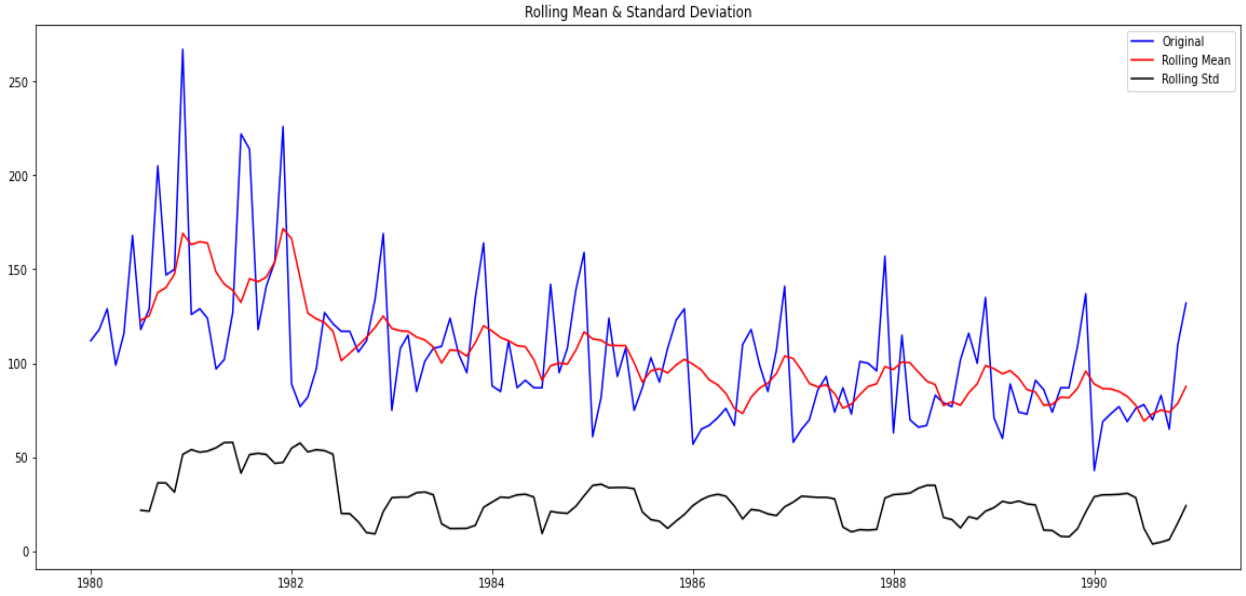
- RMSE is 383.1765
- MAPE is 11.91

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.

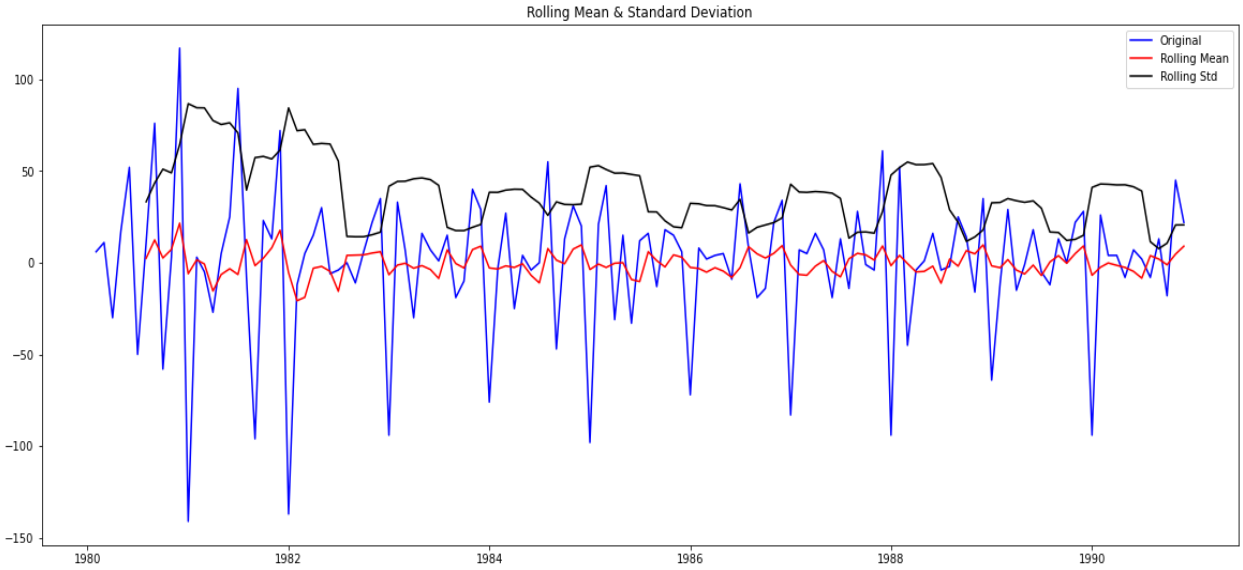
Check for stationarity of the Training Data Time Series.



Results of Dickey-Fuller Test:

- 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
  dtype: float64

We see that the series is not stationary at α = 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

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 7.061944e-09 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
18	(3, 1, 3)	1273.194169
19	(3, 1, 4)	1274.334961
2	(0, 1, 2)	1276.835375
7	(1, 1, 2)	1277.359226
6	(1, 1, 1)	1277.775747
3	(0, 1, 3)	1278.074254
4	(0, 1, 4)	1278.838364

ARIMA Model Results

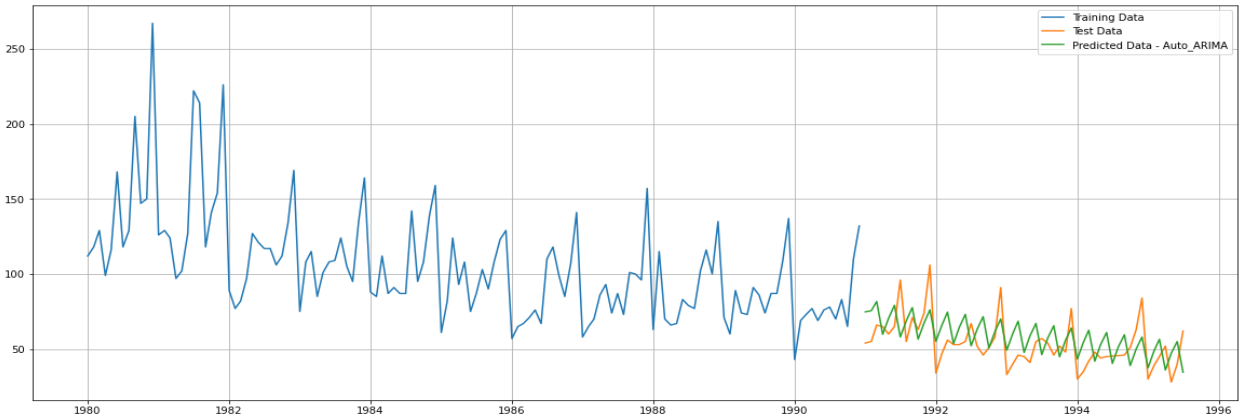
Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(3, 1, 3)	Log Likelihood	-628.597
Method:	css-mle	S.D. of innovations	28.355
Date:	Sun, 13 Sep 2020	AIC	1273.194
Time:	16:56:40	BIC	1296.196
Sample:	02-01-1980 - 12-01-1990	HQIC	1282.541

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4906	0.088	-5.549	0.000	-0.664	-0.317
ar.L1.D.Rose	-0.7245	0.086	-8.430	0.000	-0.893	-0.556
ar.L2.D.Rose	-0.7219	0.086	-8.363	0.000	-0.891	-0.553
ar.L3.D.Rose	0.2762	0.085	3.239	0.001	0.109	0.443
ma.L1.D.Rose	-0.0149	0.044	-0.336	0.737	-0.102	0.072
ma.L2.D.Rose	0.0149	0.044	0.337	0.736	-0.072	0.102
ma.L3.D.Rose	-1.0000	0.046	-21.952	0.000	-1.089	-0.911

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.5011	-0.8661j	1.0006	-0.3335
AR.2	-0.5011	+0.8661j	1.0006	0.3335
AR.3	3.6159	-0.0000j	3.6159	-0.0000
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.4925	-0.8703j	1.0000	-0.3320
MA.3	-0.4925	+0.8703j	1.0000	0.3320

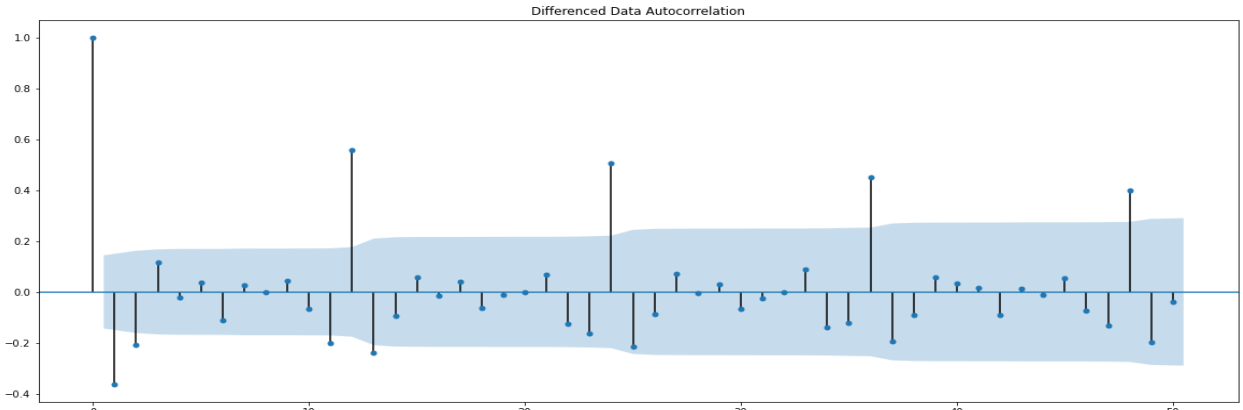
- RMSE: 15.98
- MAPE: 26.08

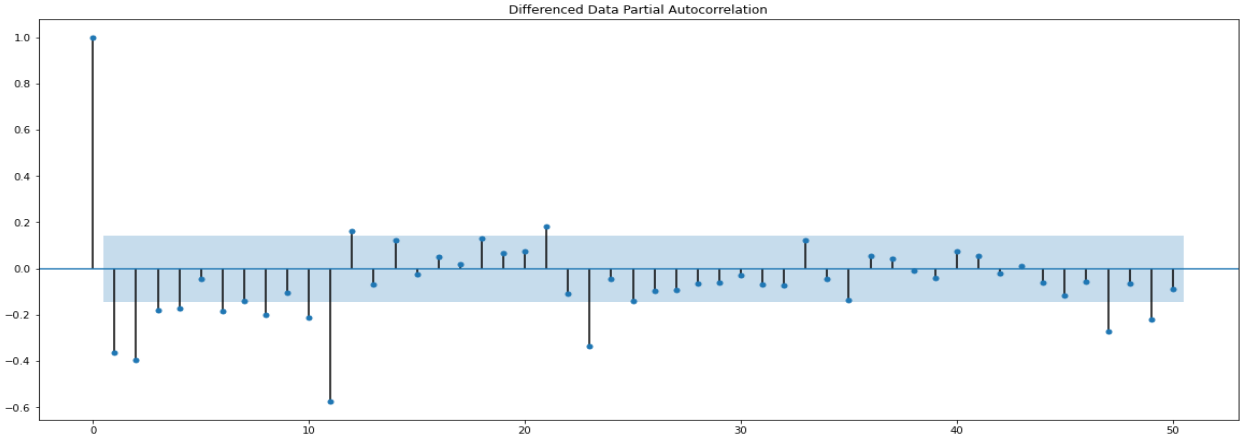


7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

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 2 and ACF plot cuts-off at lag 2.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-633.649
Method:	css-mle	S.D. of innovations	29.975
Date:	Sun, 13 Sep 2020	AIC	1279.299
Time:	16:57:00	BIC	1296.550
Sample:	02-01-1980 - 12-01-1990	HQIC	1286.309

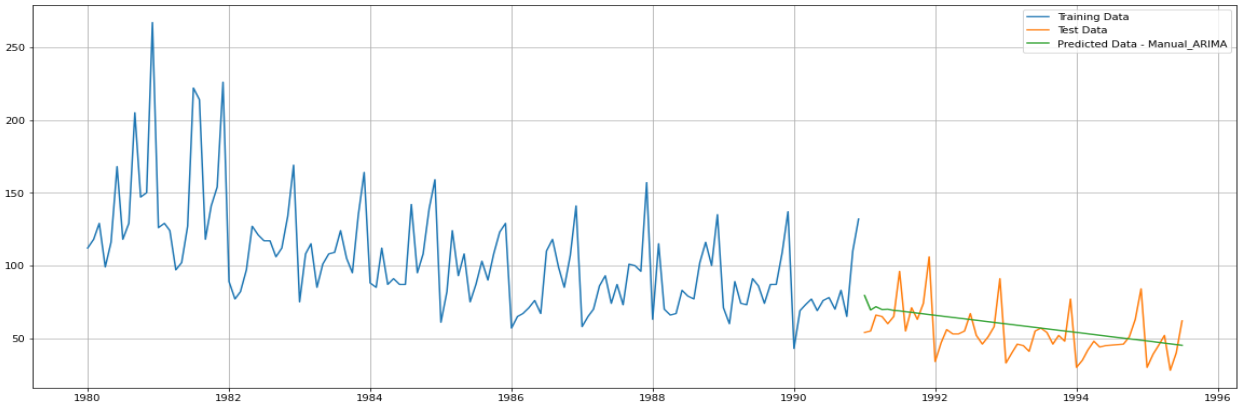
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4911	0.081	-6.076	0.000	-0.649	-0.333
ar.L1.D.Rose	-0.4383	0.218	-2.015	0.044	-0.865	-0.012
ar.L2.D.Rose	0.0269	0.109	0.246	0.806	-0.188	0.241
ma.L1.D.Rose	-0.3316	0.203	-1.633	0.102	-0.729	0.066
ma.L2.D.Rose	-0.6684	0.201	-3.332	0.001	-1.062	-0.275

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-2.0290	+0.0000j	2.0290	0.5000
AR.2	18.3387	+0.0000j	18.3387	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.4961	+0.0000j	1.4961	0.5000

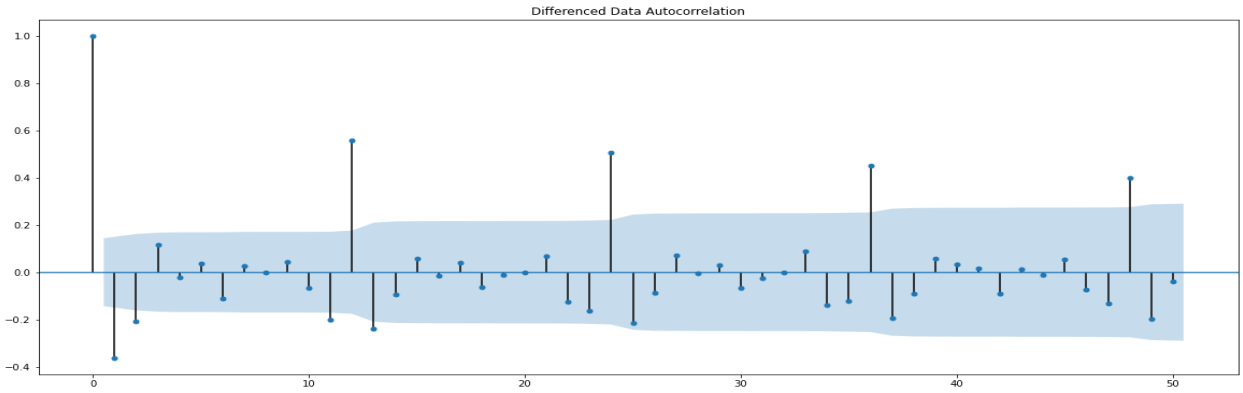
We get a comparatively simpler model by looking at the ACF and the PACF plots.

- RMSE: 15.35
- MAPE: 22.77



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 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
374	(2, 1, 4)	(4, 0, 4, 6)	870.449735
499	(3, 1, 4)	(4, 0, 4, 6)	872.394839
124	(0, 1, 4)	(4, 0, 4, 6)	873.187581
624	(4, 1, 4)	(4, 0, 4, 6)	874.074545
249	(1, 1, 4)	(4, 0, 4, 6)	877.450520

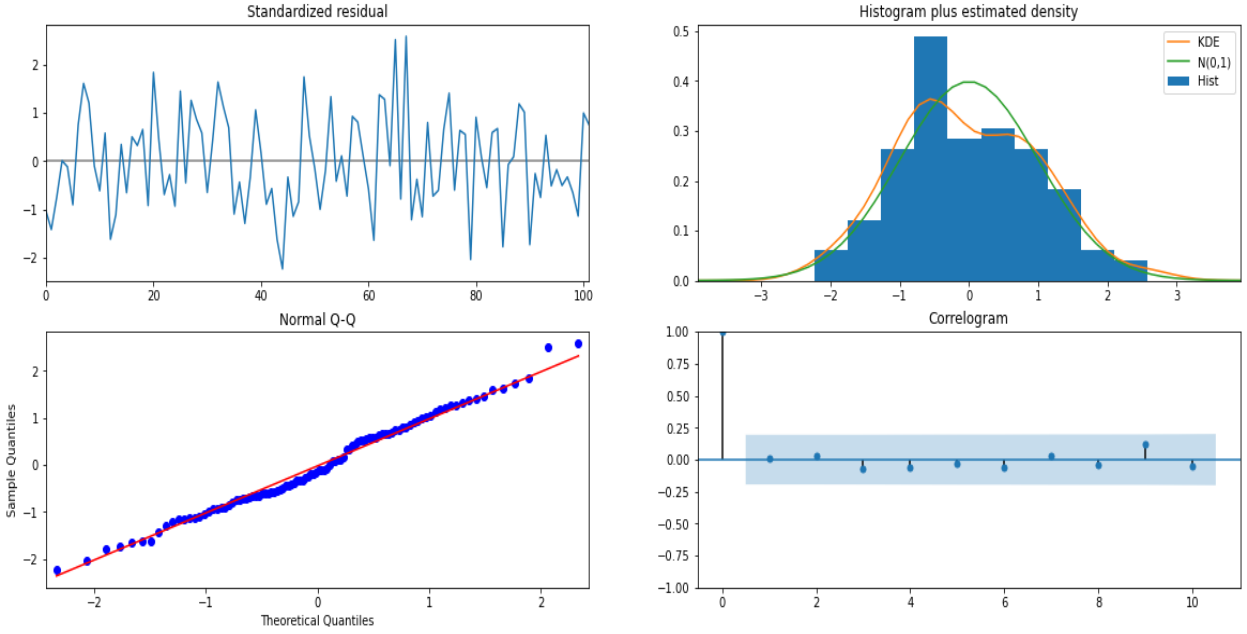
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model: SARIMAX(2, 1, 4)x(4, 0, 4, 6)		Log Likelihood	-420.225
Date:	Sun, 13 Sep 2020	AIC	870.450
Time:	17:26:10	BIC	909.824
Sample:	0	HQIC	886.394
	- 132		
Covariance Type:	opg		

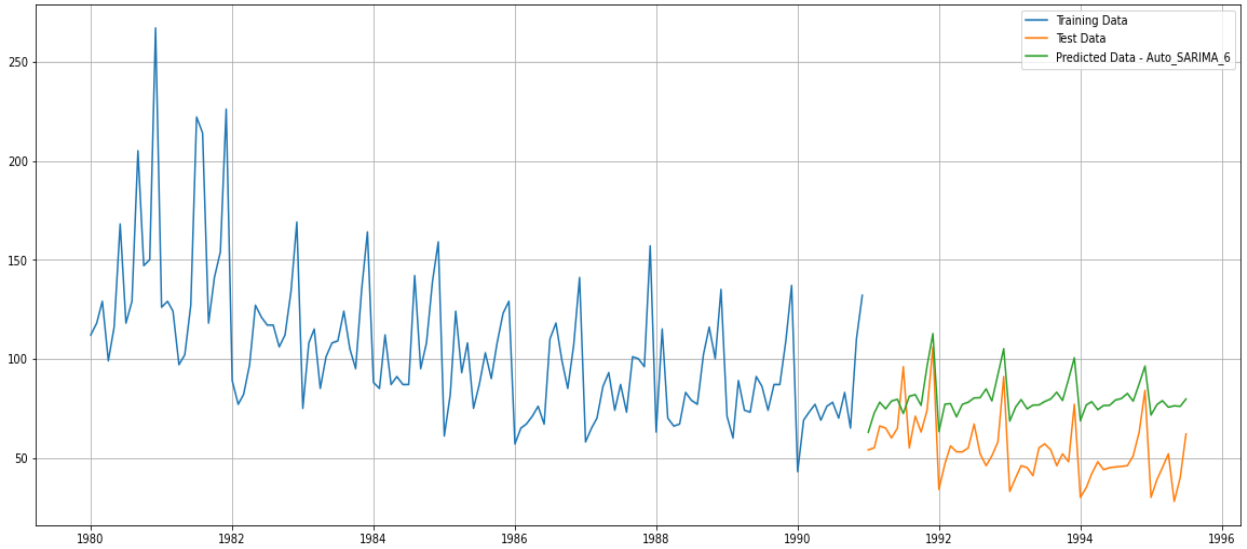
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9320	0.044	-20.954	0.000	-1.019	-0.845
ar.L2	-0.9271	0.041	-22.356	0.000	-1.008	-0.846
ma.L1	0.1058	1787.947	5.92e-05	1.000	-3504.206	3504.418
ma.L2	0.0773	1975.965	3.91e-05	1.000	-3872.744	3872.898
ma.L3	-0.9803	2139.625	-0.000	1.000	-4194.568	4192.608
ma.L4	-0.2028	368.123	-0.001	1.000	-721.711	721.305
ar.S.L6	0.2974	0.098	3.021	0.003	0.104	0.490
ar.S.L12	0.3624	0.069	5.247	0.000	0.227	0.498
ar.S.L18	-0.3066	0.073	-4.213	0.000	-0.449	-0.164
ar.S.L24	0.3590	0.058	6.171	0.000	0.245	0.473
ma.S.L6	-0.4153	0.187	-2.215	0.027	-0.783	-0.048
ma.S.L12	0.0629	0.135	0.467	0.640	-0.201	0.327
ma.S.L18	0.3813	0.141	2.712	0.007	0.106	0.657
ma.S.L24	-0.3847	0.151	-2.543	0.011	-0.681	-0.088
sigma2	186.8735	3.39e+05	0.001	1.000	-6.65e+05	6.65e+05

Ljung-Box (Q):	21.41	Jarque-Bera (JB):	1.49
Prob(Q):	0.99	Prob(JB):	0.47
Heteroskedasticity (H):	0.91	Skew:	0.21
Prob(H) (two-sided):	0.78	Kurtosis:	2.58

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (comple
x-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.6
4e+14. Standard errors may be unstable.



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: 28.44
- MAPE: 49.57

Method 13: Auto SARIMA Model_12

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

	param	seasonal	AIC
499	(3, 1, 4)	(4, 0, 4, 12)	667.307014
374	(2, 1, 4)	(4, 0, 4, 12)	667.677140
244	(1, 1, 4)	(3, 0, 4, 12)	668.899289
624	(4, 1, 4)	(4, 0, 4, 12)	669.285928
109	(0, 1, 4)	(1, 0, 4, 12)	669.898798

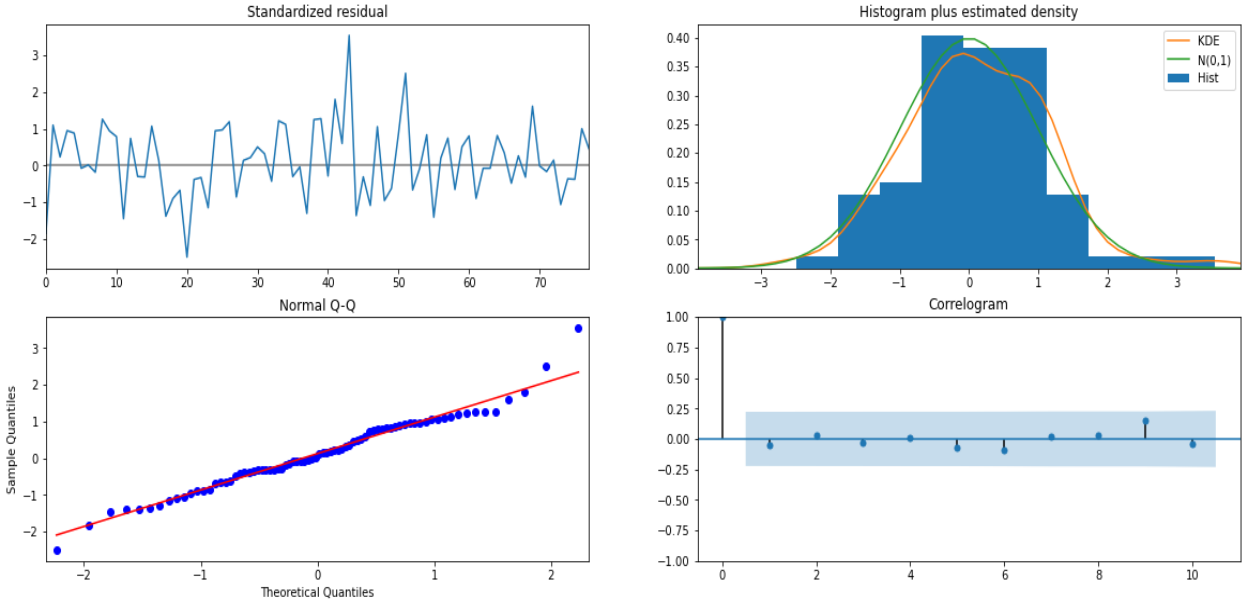
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3,1,4)x(4,0,4,12)	Log Likelihood	-317.654
Date:	Sun, 13 Sep 2020	AIC	667.307
Time:	19:15:00	BIC	705.014
Sample:	0	HQIC	682.402
	- 132		
Covariance Type:	opg		

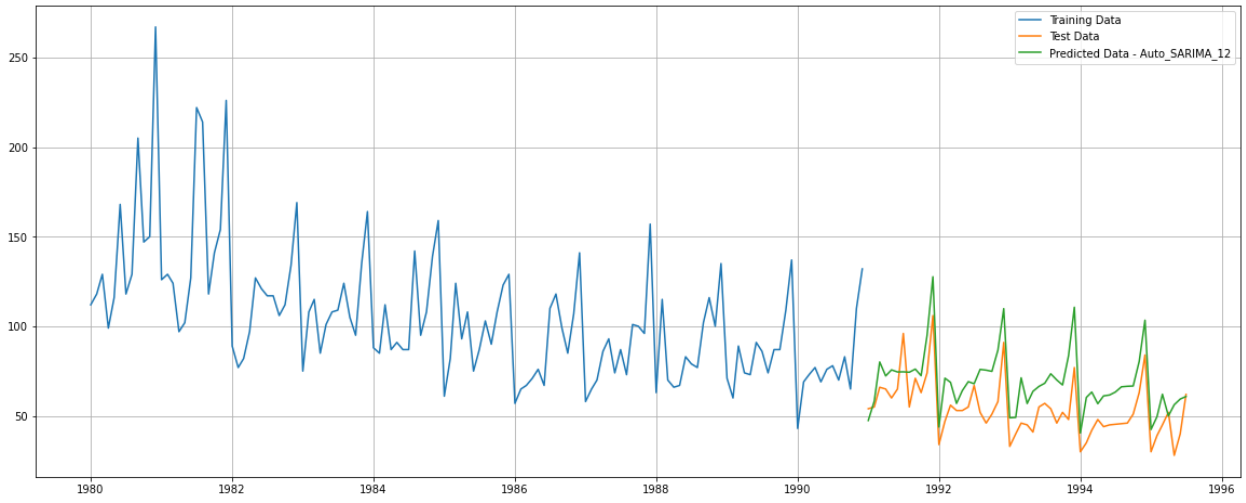
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.7519	0.256	-6.851	0.000	-2.253	-1.251
ar.L2	-1.6381	0.315	-5.204	0.000	-2.255	-1.021
ar.L3	-0.7170	0.233	-3.076	0.002	-1.174	-0.260
ma.L1	0.7808	30.727	0.025	0.980	-59.442	61.004
ma.L2	-0.0850	64.504	-0.001	0.999	-126.510	126.340
ma.L3	-1.0024	38.334	-0.026	0.979	-76.135	74.130
ma.L4	-0.9750	66.558	-0.015	0.988	-131.427	129.477
ar.S.L12	0.0838	0.200	0.419	0.675	-0.309	0.476
ar.S.L24	0.7132	0.159	4.493	0.000	0.402	1.024
ar.S.L36	0.2409	0.096	2.512	0.012	0.053	0.429
ar.S.L48	-0.1120	0.073	-1.525	0.127	-0.256	0.032
ma.S.L12	0.3152	2004.844	0.000	1.000	-3929.107	3929.737
ma.S.L24	-0.7358	1086.264	-0.001	0.999	-2129.774	2128.302
ma.S.L36	-0.8456	1993.857	-0.000	1.000	-3908.733	3907.042
ma.S.L48	0.2660	500.957	0.001	1.000	-981.592	982.124
sigma2	95.8356	1.78e+05	0.001	1.000	-3.48e+05	3.48e+05

Ljung-Box (Q):	33.29	Jarque-Bera (JB):	4.07
Prob(Q):	0.76	Prob(JB):	0.13
Heteroskedasticity (H):	0.47	Skew:	0.29
Prob(H) (two-sided):	0.06	Kurtosis:	3.96

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



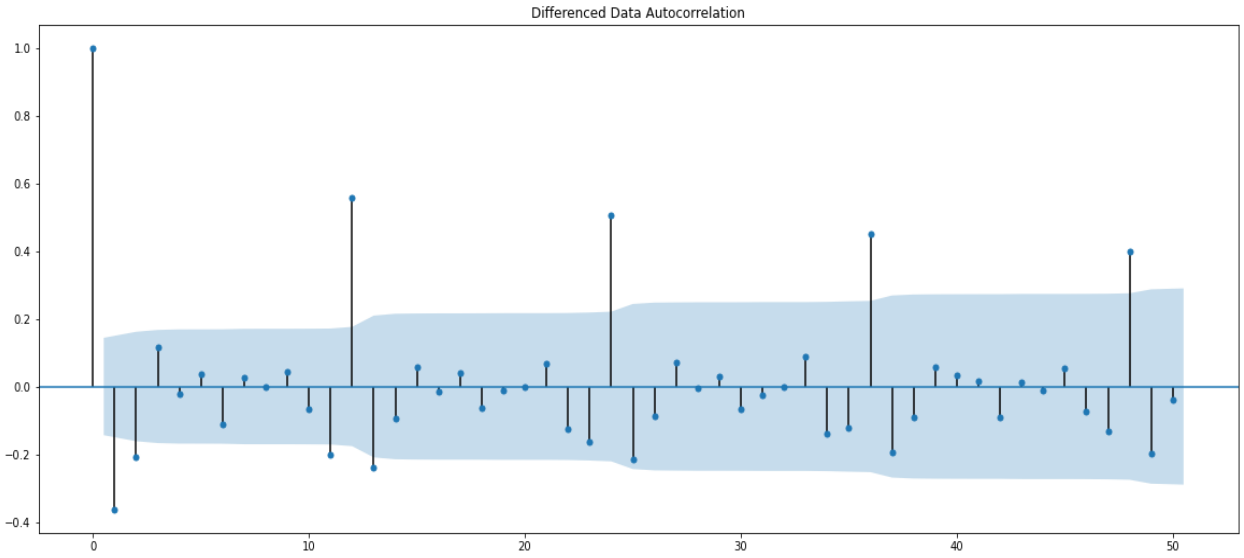
Similar to the last iteration of the model where the seasonality parameter was taken as 6, here also we see that the model diagnostics plot does not indicate any remaining information that we can get.

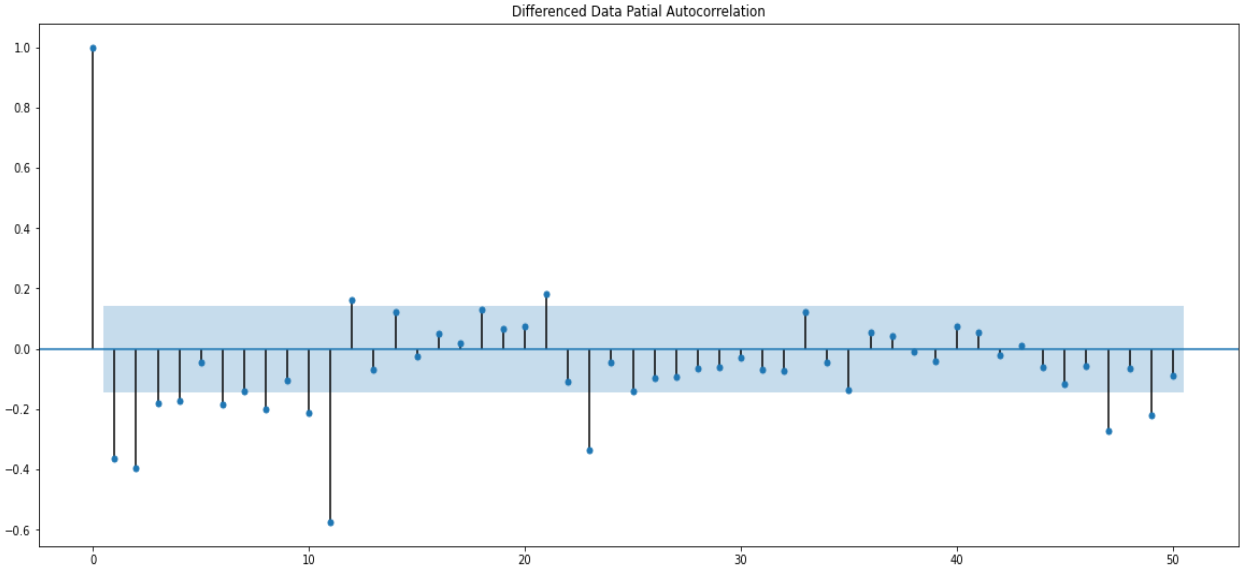


- RMSE: 18.11
- MAPE: 30.18

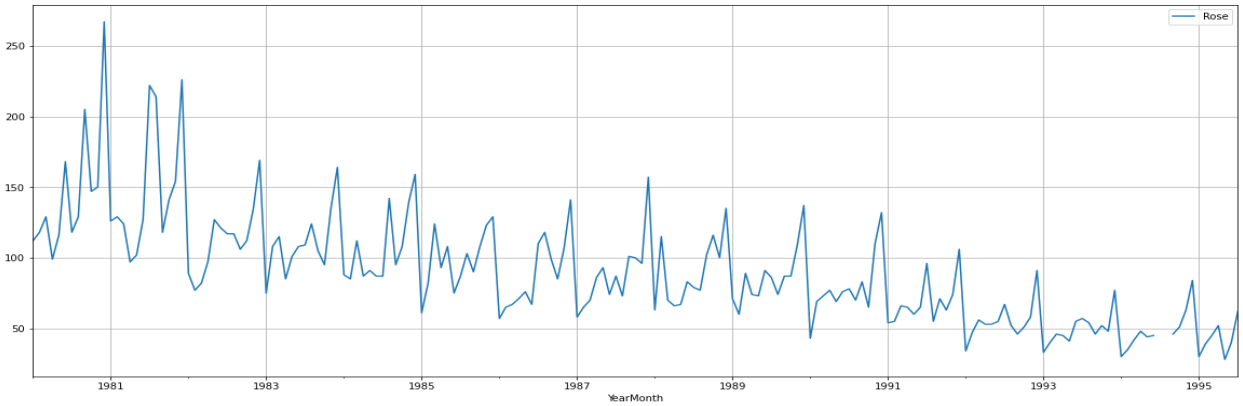
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.

Method 14: Manual SARIMA model_6

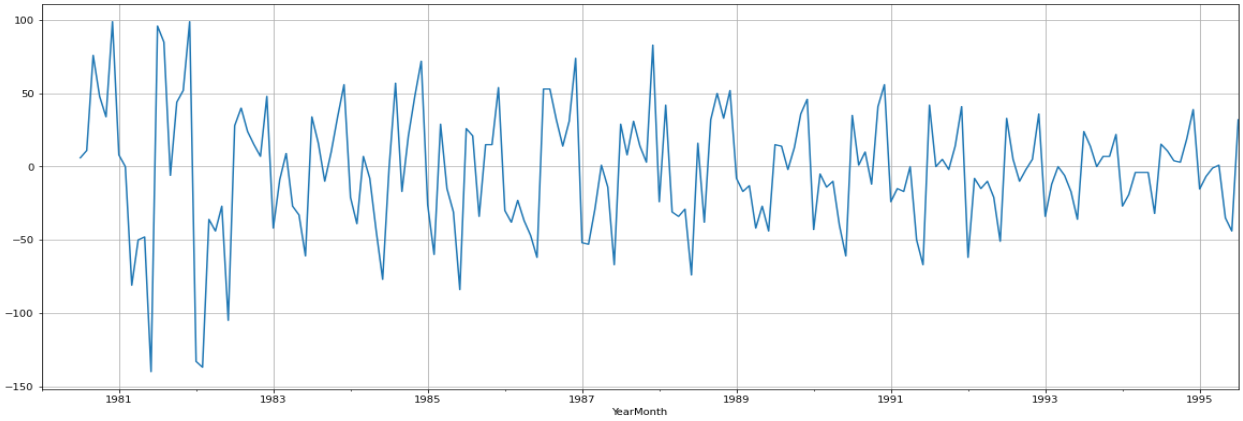




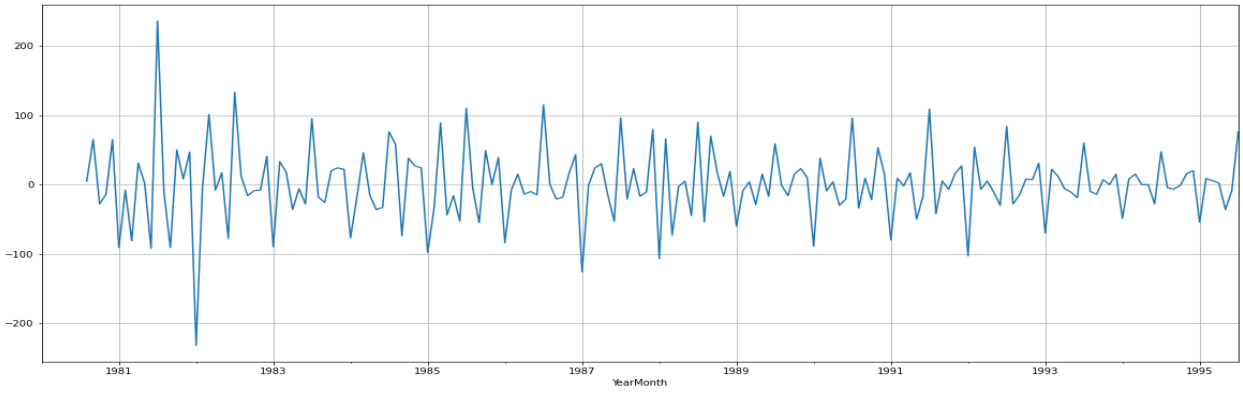
We see that our ACF plot at the seasonal interval (6) does not taper off. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.



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

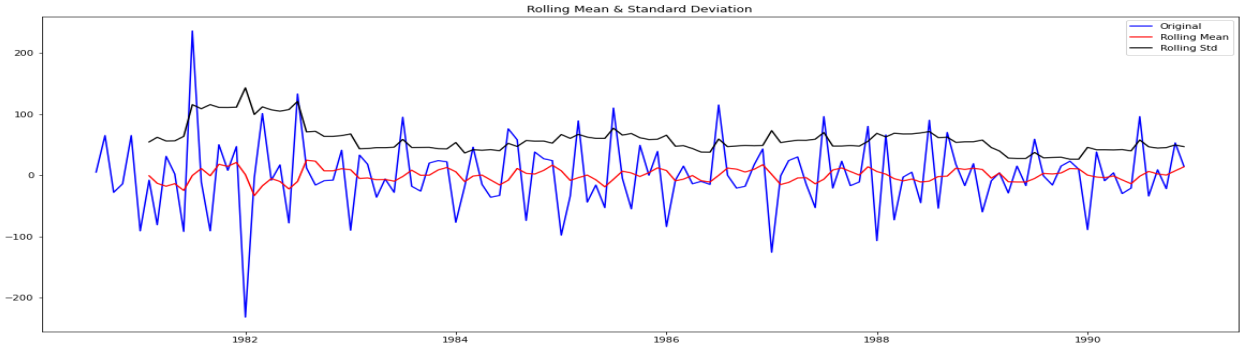


We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.



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.

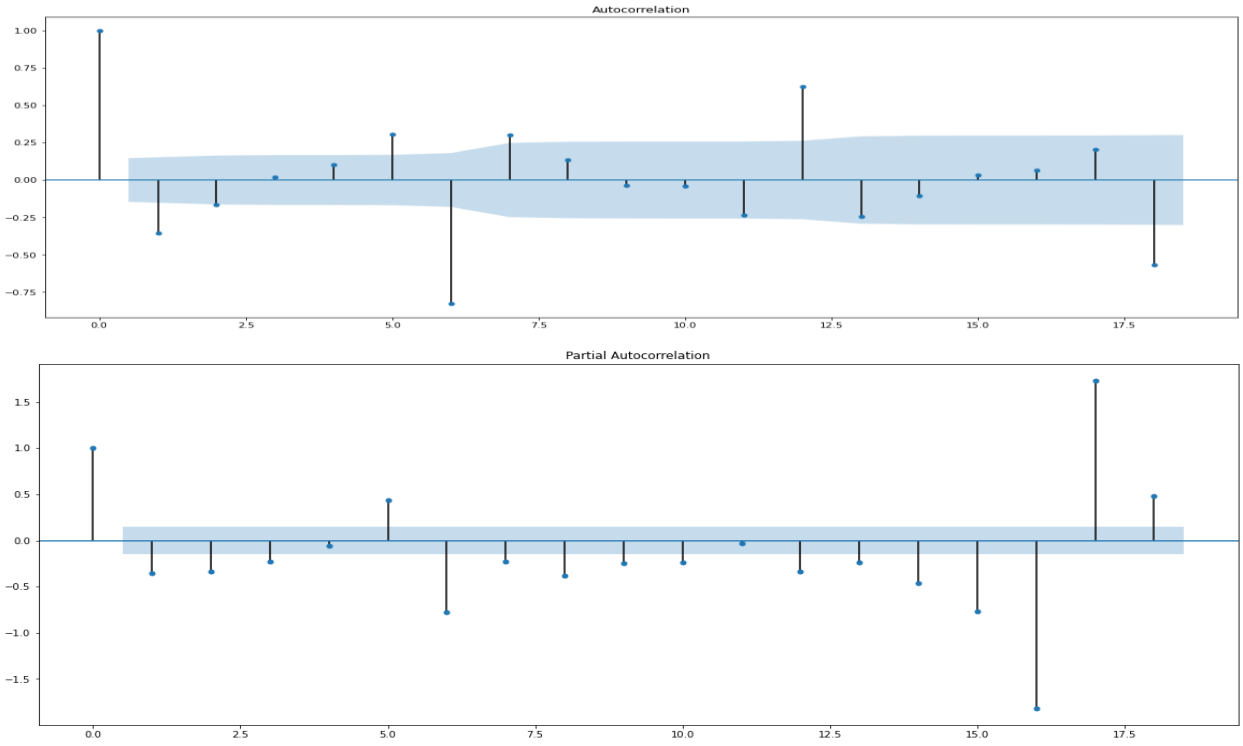


Results of Dickey-Fuller Test:

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

dtype: float64

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



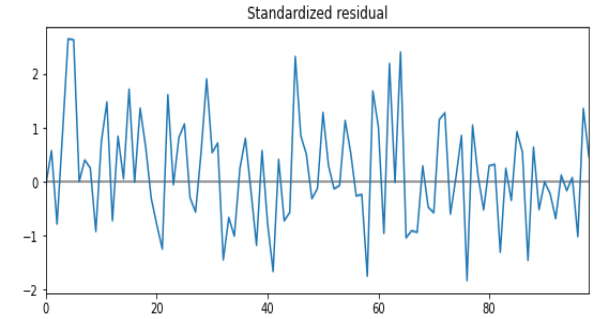
Here, we have taken alpha=0.05.

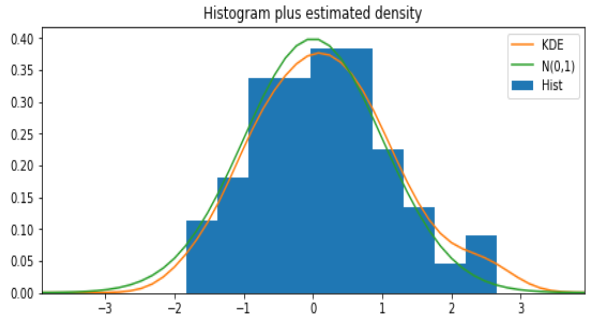
We are going to take the seasonal period as 6. We will keep the p(4) and q(2) parameters same as the ARIMA model.

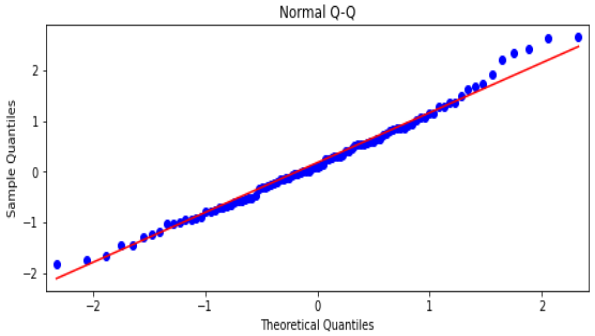
The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the lag at which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the lag at which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 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.

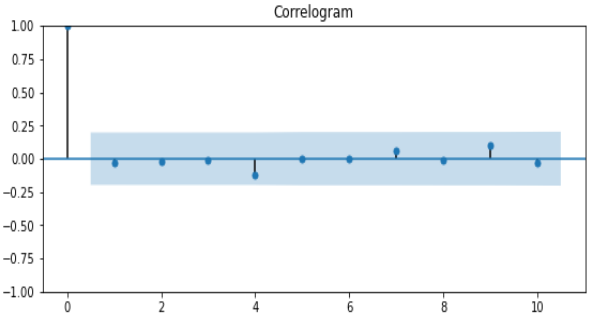
This is a common problem while building models by looking at the ACF and the PACF plots. But we are able to explain the model.

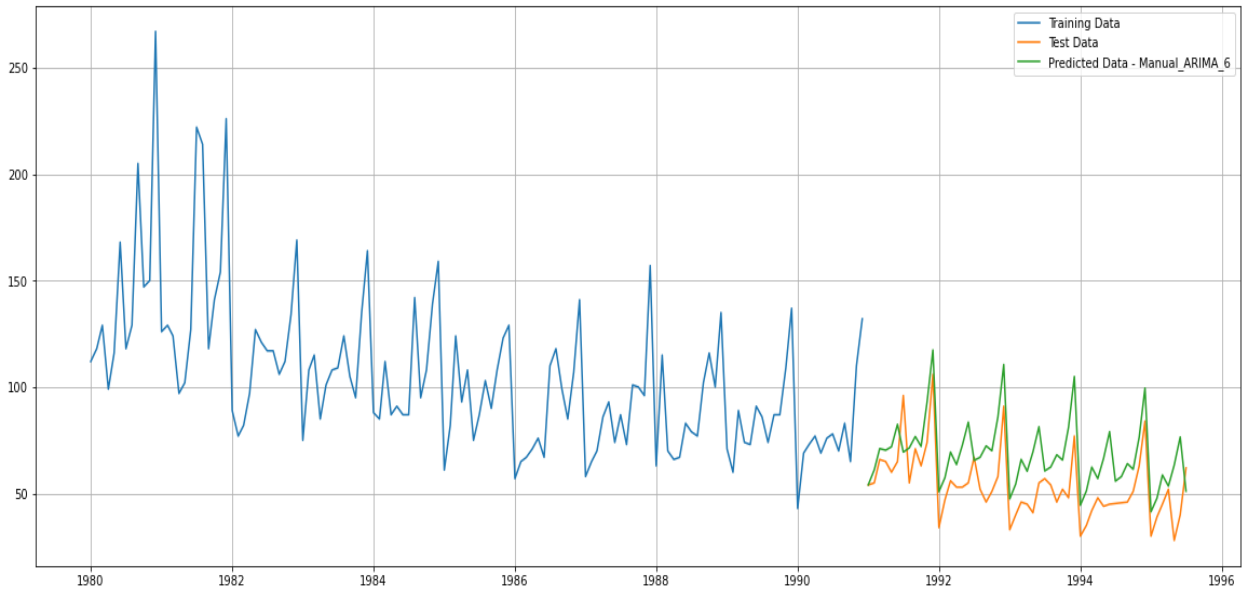
```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                             y        No. Observations:                  132
Model: SARIMAX(3,1,1)x(2,1,[1,2,3,4],6)  Log Likeli                -417.602
Date:                     Sun, 13 Sep 2020   AIC                              857.205
Time:                             19:15:34   BIC                              885.751
Sample:                                  0   HQIC                             868.755
                                       - 132
Covariance Type:                       opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
----
ar.L1          0.1304      0.130      1.000      0.318      -0.125       0.386
ar.L2          0.0212      0.140      0.151      0.880      -0.254       0.296
ar.L3         -0.1585      0.127     -1.248      0.212      -0.407       0.090
ma.L1         -0.8298      0.105     -7.884      0.000      -1.036      -0.624
ar.S.L6       -0.8837      0.182     -4.856      0.000      -1.240      -0.527
ar.S.L12       0.0503      0.174      0.289      0.772      -0.290       0.391
ma.S.L6       -0.1336      0.195     -0.687      0.492      -0.515       0.248
ma.S.L12      -0.5123      0.123     -4.173      0.000      -0.753      -0.272
ma.S.L18       0.2112      0.125      1.683      0.092      -0.035       0.457
ma.S.L24      -0.1425      0.126     -1.134      0.257      -0.389       0.104
sigma2       257.8214     45.489      5.668      0.000     168.665     346.977
==========================================================================================
Ljung-Box (Q):                31.35   Jarque-Bera (JB):                 2.01
Prob(Q):                       0.83   Prob(JB):                         0.37
Heteroskedasticity (H):        0.52   Skew:                             0.34
Prob(H) (two-sided):           0.06   Kurtosis:                         2.85
==========================================================================================

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







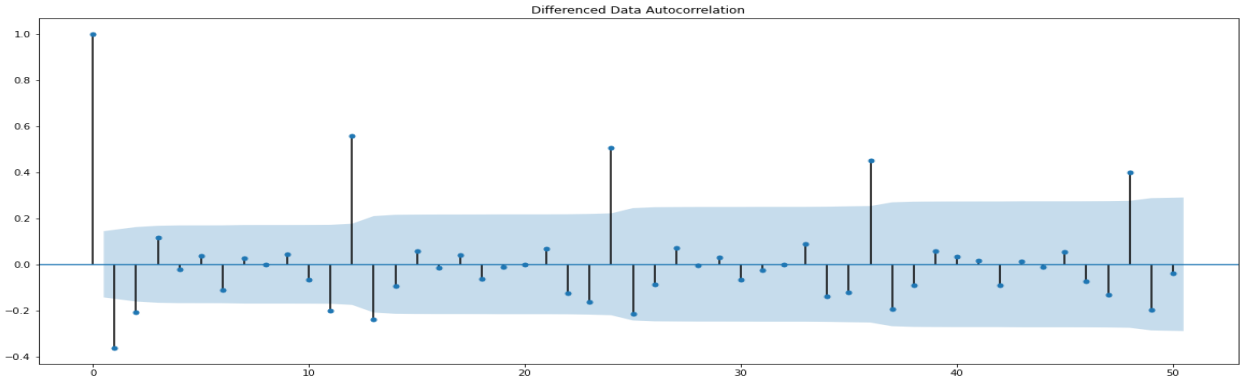


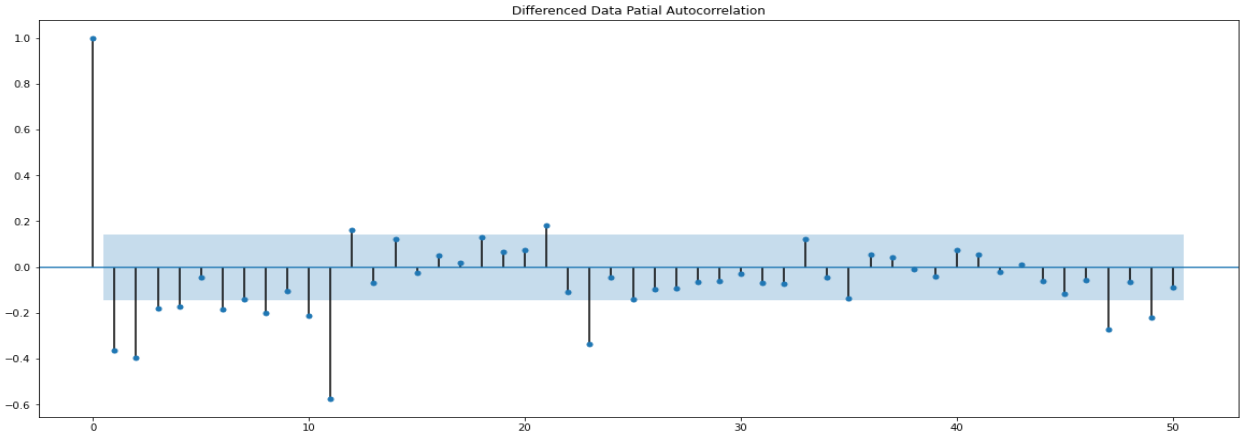


- RMSE: 18.35
- MAPE: 29.95

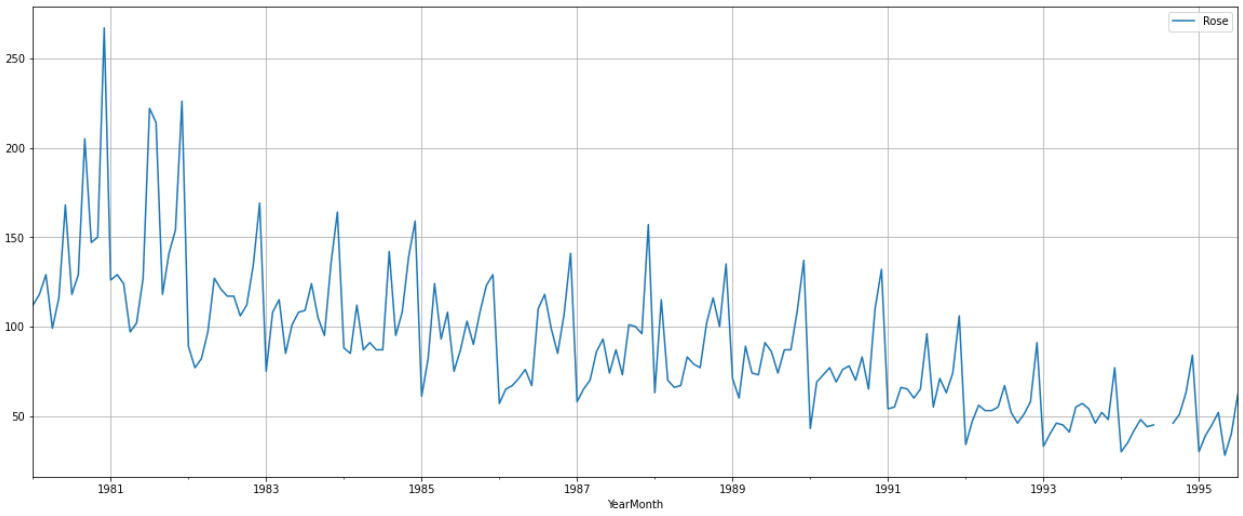
Method 15: Manual SARIMA model_12

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

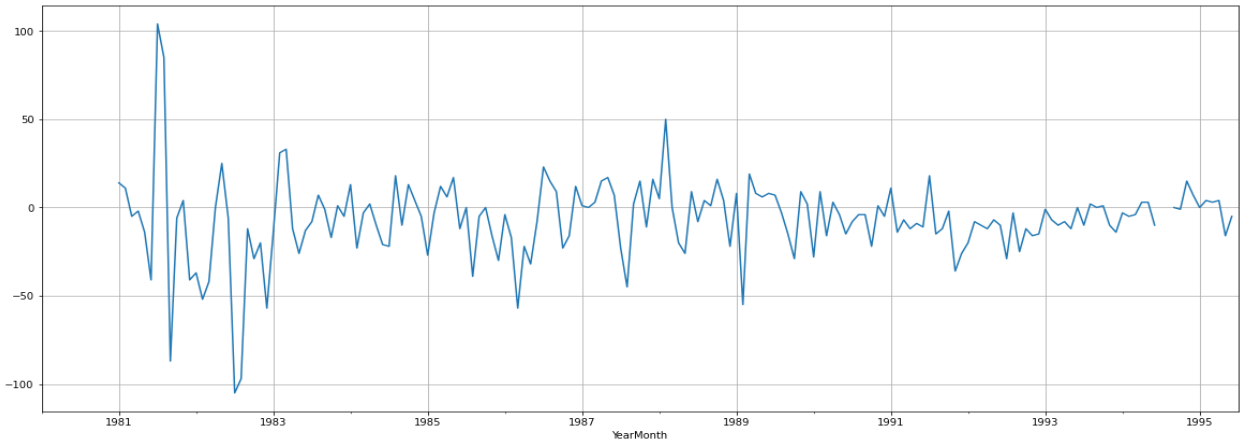


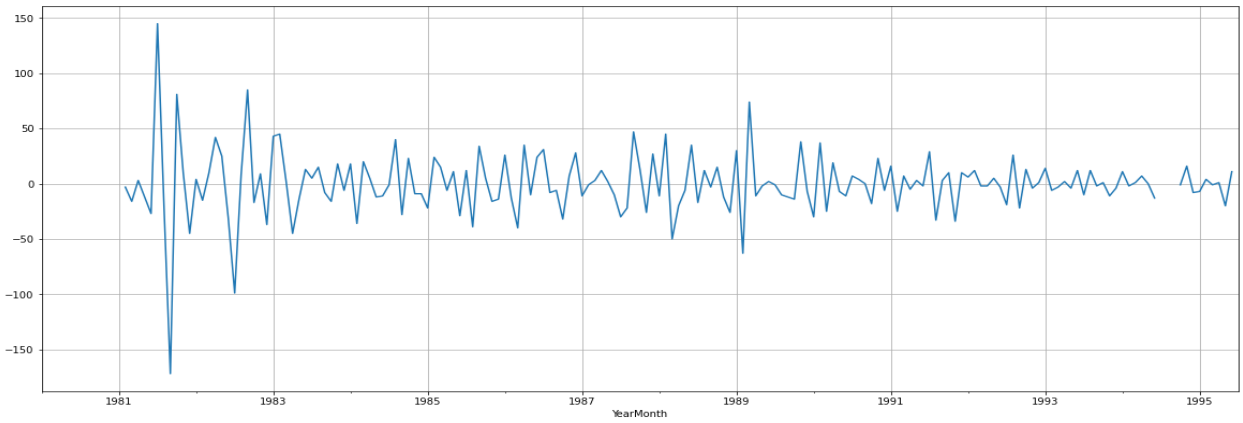


We see that our ACF plot at the seasonal interval (12) does not taper off. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.

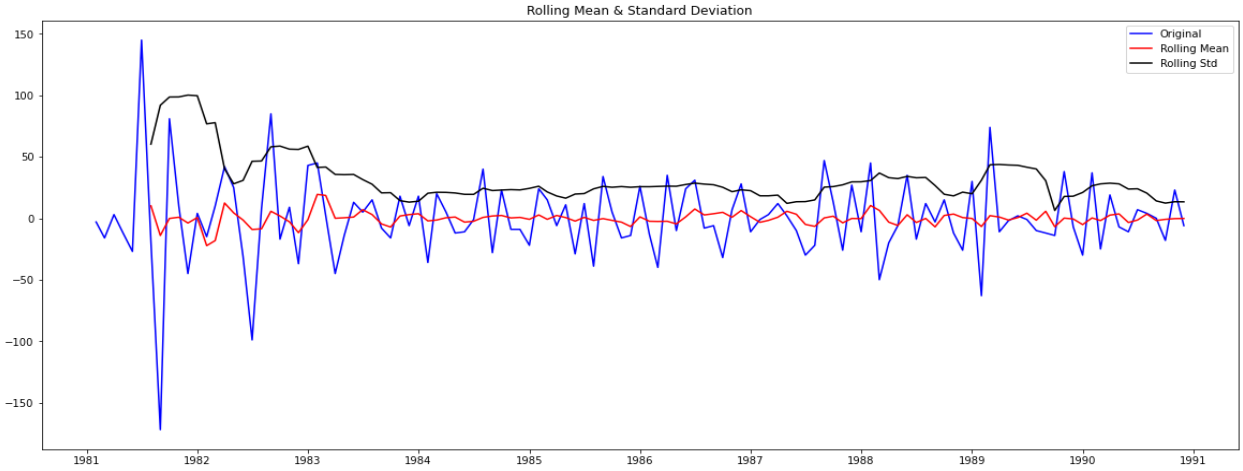


We see that there is a slight trend and a seasonality. So, now we take a seasonal differencing and check the series.





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.

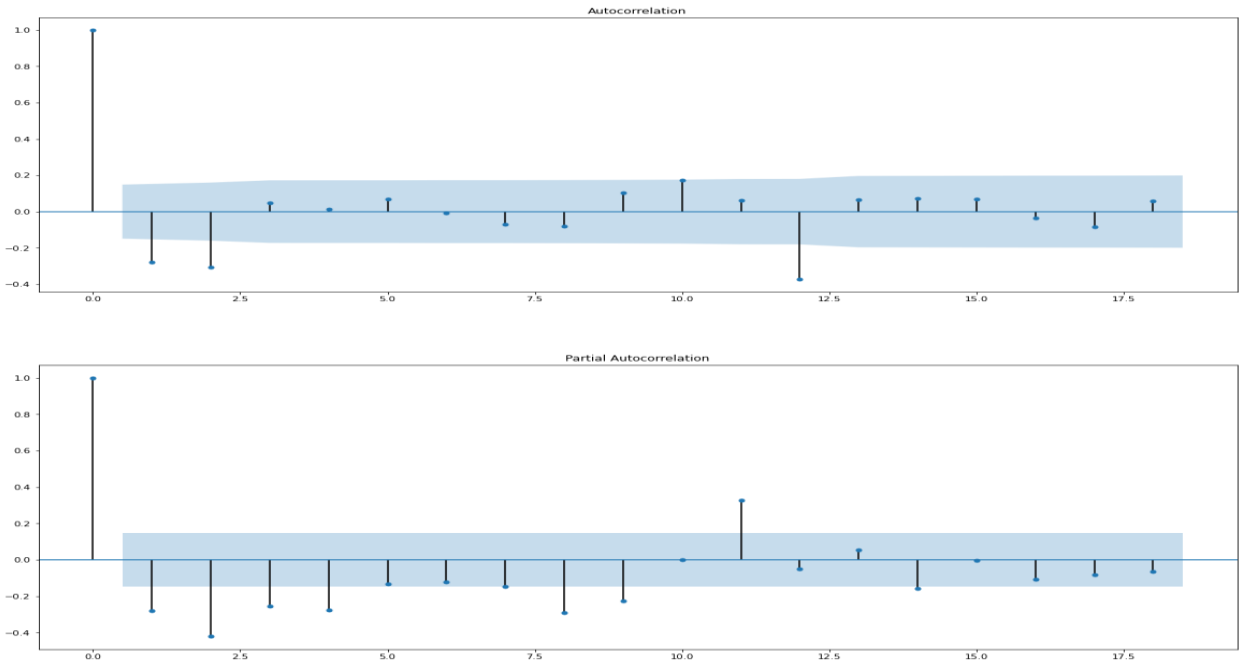


Results of Dickey-Fuller Test:

- Test Statistic -3.692348
- p-value 0.004222
- #Lags Used 11.000000
- Number of Observations Used 107.000000
- Critical Value (1%) -3.492996
- Critical Value (5%) -2.888955
- Critical Value (10%) -2.581393

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(4) and q(2) 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.

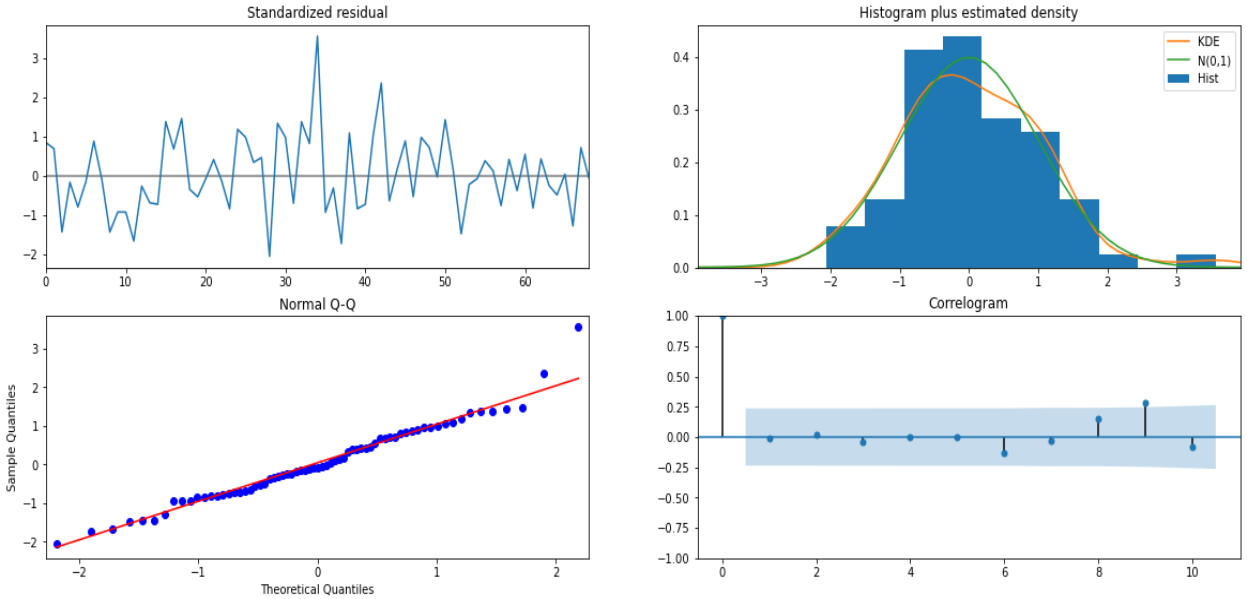
SARIMAX Results

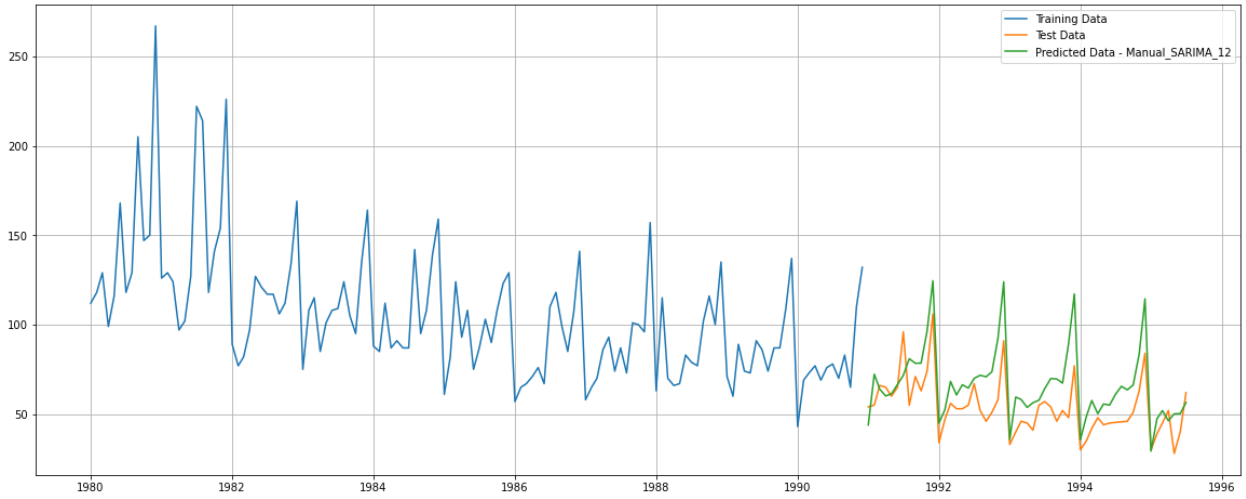
Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2,1,2)x(4,1,2,12)	Log Likelihood	-284.472
Date:	Sun, 13 Sep 2020	AIC	590.945
Time:	19:16:23	BIC	615.520
Sample:	0 - 132	HQIC	600.695
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9797	0.225	-4.364	0.000	-1.420	-0.540
ar.L2	-0.1273	0.143	-0.890	0.373	-0.408	0.153
ma.L1	0.0206	0.247	0.084	0.933	-0.463	0.504
ma.L2	-0.8824	0.193	-4.573	0.000	-1.261	-0.504
ar.S.L12	-0.7350	0.199	-3.702	0.000	-1.124	-0.346
ar.S.L24	-0.0734	0.174	-0.422	0.673	-0.414	0.268
ar.S.L36	0.0758	0.088	0.859	0.390	-0.097	0.249
ar.S.L48	-0.0064	0.021	-0.308	0.758	-0.047	0.034
ma.S.L12	-0.3538	0.696	-0.509	0.611	-1.717	1.009
ma.S.L24	-0.9030	0.558	-1.620	0.105	-1.996	0.190
sigma2	144.5941	109.858	1.316	0.188	-70.723	359.911

Ljung-Box (Q):	51.74	Jarque-Bera (JB):	6.01
Prob(Q):	0.10	Prob(JB):	0.05
Heteroskedasticity (H):	0.62	Skew:	0.53
Prob(H) (two-sided):	0.25	Kurtosis:	3.98

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





- RMSE: 17.34
- MAPE: 26.69

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	15.269000	22.82
0	RegressionOnTimeSeasonal	15.243000	22.73
0	Naive_model	79.888000	145.79
0	Simple Average	53.636000	95.48
0	moving_avg_forecast_4	14.451000	19.49
0	moving_avg_forecast_6	14.566000	20.82
0	moving_avg_forecast_8	14.805000	21.06
0	moving_avg_forecast_12	15.236000	22.07
0	SES	36.796250	63.88
0	Holt_linear	70.572452	120.25
0	Holt_Winter	16.447061	22.88
0	Holt_Winter M	17.369490	28.88
0	Auto_ARIMA(3,1,3)	15.985092	26.08
0	Manual_ARIMA(2,1,2)	15.354879	22.77
0	Auto_SARIMA(2,1,4)(4,0,4,6)	28.449744	49.57
0	Auto_SARIMA(3,1,4)(4,0,4,12)	18.110420	30.18
0	Manual_SARIMA(3,1,1)(2,1,4,6)	18.359456	29.95
0	Manual_SARIMA(2,1,2)(4,1,2,12)	17.341983	26.69

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.

Model 1

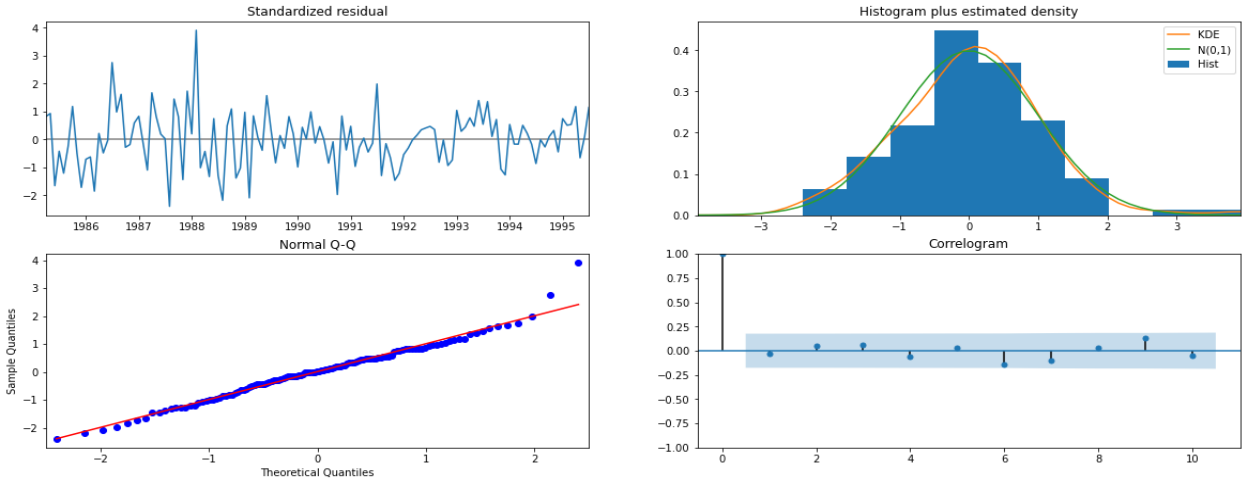
SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(2,1,2)x(4,1,2,12)	Log Likelihood	-492.373
Date:	Sun, 13 Sep 2020	AIC	1006.747
Time:	19:17:40	BIC	1037.770
Sample:	01-01-1980	HQIC	1019.349
	- 07-01-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0122	0.104	9.702	0.000	0.808	1.217
ar.L2	-0.1787	0.110	-1.620	0.105	-0.395	0.038
ma.L1	-1.9516	77.583	-0.025	0.980	-154.011	150.107
ma.L2	1.0000	79.509	0.013	0.990	-154.836	156.836
ar.S.L12	-0.7531	0.127	-5.910	0.000	-1.003	-0.503
ar.S.L24	-0.0257	0.155	-0.166	0.868	-0.330	0.278
ar.S.L36	-0.0261	0.096	-0.272	0.785	-0.214	0.162
ar.S.L48	-0.0175	0.030	-0.582	0.560	-0.076	0.041
ma.S.L12	0.0474	0.180	0.263	0.793	-0.306	0.401
ma.S.L24	-0.5474	0.184	-2.981	0.003	-0.907	-0.188
sigma2	149.1970	1.19e+04	0.013	0.990	-2.31e+04	2.34e+04

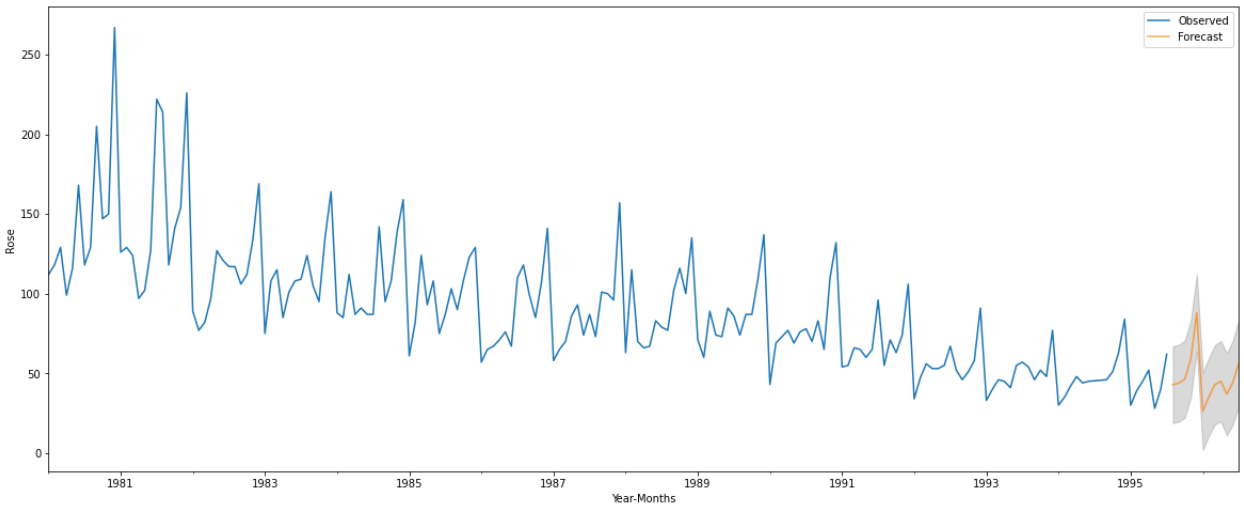
Ljung-Box (Q):	41.03	Jarque-Bera (JB):	9.97
Prob(Q):	0.43	Prob(JB):	0.01
Heteroskedasticity (H):	0.26	Skew:	0.32
Prob(H) (two-sided):	0.00	Kurtosis:	4.23

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



Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	42.882449	12.307073	18.761030	67.003868
1995-09-01	43.830453	12.339931	19.644632	68.016274
1995-10-01	46.343647	12.356903	22.124563	70.562732
1995-11-01	59.105613	12.358431	34.883534	83.327693
1995-12-01	87.922373	12.366062	63.685337	112.159408
1996-01-01	26.195641	12.414435	1.863795	50.527488
1996-02-01	35.192859	12.518705	10.656647	59.729070
1996-03-01	42.855579	12.682835	17.997679	67.713479
1996-04-01	45.031205	12.901380	19.744966	70.317445
1996-05-01	36.887220	13.165315	11.083677	62.690762
1996-06-01	44.290399	13.465051	17.899384	70.681414
1996-07-01	56.466726	13.791683	29.435524	83.497928

- RMSE of the Full Model 44.36
- MAPE of manual_SARIMA_12_full_data: 27.74



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	15.269000	22.82
0	RegressionOnTimeSeasonal	15.243000	22.73
0	Naive_model	79.888000	145.79
0	Simple Average	53.636000	95.48
0	moving_avg_forecast_4	14.451000	19.49
0	moving_avg_forecast_6	14.566000	20.82
0	moving_avg_forecast_8	14.805000	21.06
0	moving_avg_forecast_12	15.236000	22.07
0	SES	36.796250	63.88
0	Holt_linear	70.572452	120.25
0	Holt_Winter	16.447061	22.88
0	Holt_Winter M	17.369490	28.88
0	Auto_ARIMA(3,1,3)	15.985092	26.08
0	Manual_ARIMA(2,1,2)	15.354879	22.77
0	Auto_SARIMA(2,1,4)(4,0,4,6)	28.449744	49.57
0	Auto_SARIMA(3,1,4)(4,0,4,12)	18.110420	30.18
0	Manual_SARIMA(3,1,1)(2,1,4,6)	18.359456	29.95
0	Manual_SARIMA(2,1,2)(4,1,2,12)	17.341983	26.69
0	Fulldata_Manual_SARIMA(2,1,2)(4,1,2,12)	44.366065	27.74

Inference:

As of now, we observe that Manual_SARIMA **(2,1,2)(4,1,2,12)** seems to be a good fit for the data, since the RMSE (17.34) value and MAPE (26.69) respectively is low compared to other models. From the graph we can see the prediction is also like the test data.

Model 2

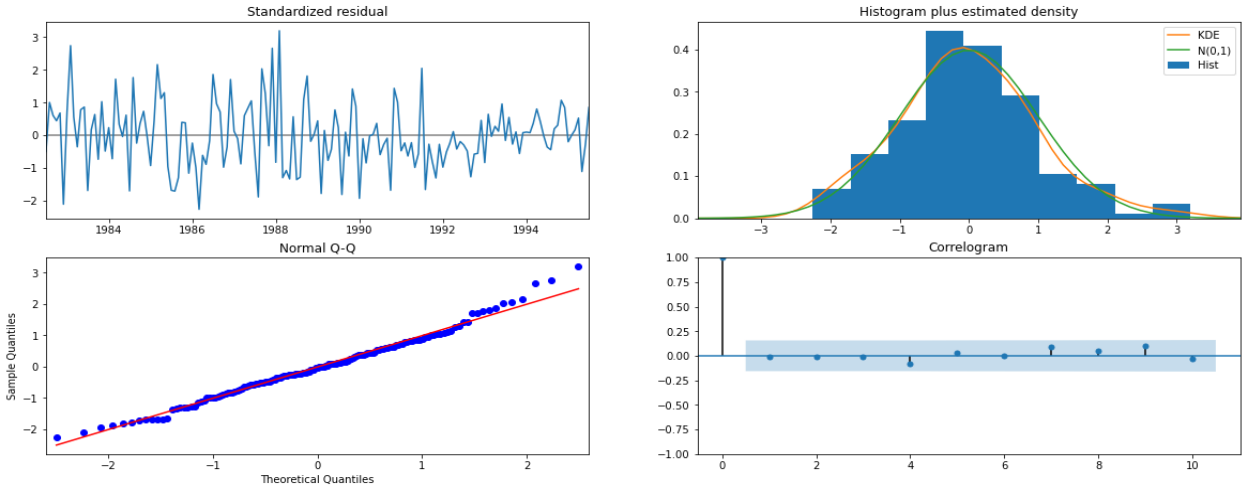
SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(2,1,4)x(4,0,4,6)	Log Likelihood	-623.724
Date:	Sun, 13 Sep 2020	AIC	1277.448
Time:	21:32:04	BIC	1323.291
Sample:	01-01-1980	HQIC	1296.066
	- 07-01-1995		
Covari	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.5563	0.019	82.962	0.000	1.520	1.593
ar.L2	-0.9345	0.018	-51.226	0.000	-0.970	-0.899
ma.L1	-2.4906	11.164	-0.223	0.823	-24.372	19.391
ma.L2	2.4097	23.554	0.102	0.919	-43.755	48.574
ma.L3	-0.8690	12.044	-0.072	0.942	-24.474	22.736
ma.L4	-0.0025	0.086	-0.029	0.976	-0.171	0.166
ar.S.L6	0.3256	0.071	4.560	0.000	0.186	0.466
ar.S.L12	0.4109	0.074	5.542	0.000	0.266	0.556
ar.S.L18	-0.3753	0.062	-6.065	0.000	-0.497	-0.254
ar.S.L24	0.3846	0.062	6.168	0.000	0.262	0.507
ma.S.L6	-0.3259	0.111	-2.928	0.003	-0.544	-0.108
ma.S.L12	-0.1115	0.118	-0.949	0.343	-0.342	0.119
ma.S.L18	0.5841	0.134	4.343	0.000	0.321	0.848
ma.S.L24	-0.3566	0.107	-3.328	0.001	-0.567	-0.147
sigma2	147.4871	2033.442	0.073	0.942	-3837.985	4132.960

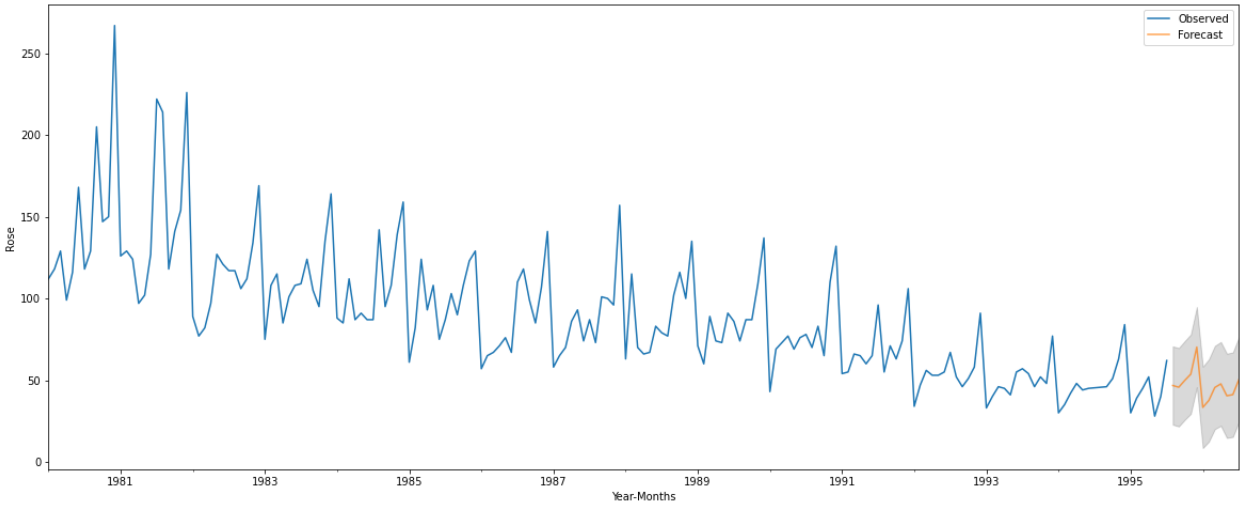
Ljung-Box (Q):	27.18	Jarque-Bera (JB):	3.76
Prob(Q):	0.94	Prob(JB):	0.15
Heteroskedasticity (H):	0.37	Skew:	0.34
Prob(H) (two-sided):	0.00	Kurtosis:	3.35

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



Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	46.744334	12.221744	22.790156	70.698511
1995-09-01	45.687710	12.253168	21.671941	69.703479
1995-10-01	49.858260	12.303255	25.744325	73.972196
1995-11-01	53.695494	12.392146	29.407334	77.983654
1995-12-01	70.237812	12.529089	45.681249	94.794374
1996-01-01	33.353954	12.698892	8.464584	58.243325
1996-02-01	37.567127	12.863591	12.354952	62.779302
1996-03-01	45.540354	12.997603	20.065520	71.015187
1996-04-01	47.728126	13.089603	22.072975	73.383277
1996-05-01	40.443890	13.148067	14.674151	66.213628
1996-06-01	41.160399	13.190582	15.307332	67.013465
1996-07-01	50.317624	13.233754	24.379942	76.255306

- RMSE of the Full Model 54.64
- MAPE of manual_SARIMA_12_full_data: 28.42



	Method	RMSE	MAPE
0	RegressionOnTime	15.269000	22.82
0	RegressionOnTimeSeasonal	15.243000	22.73
0	Naive_model	79.888000	145.79
0	Simple Average	53.636000	95.48
0	moving_avg_forecast_4	14.451000	19.49
0	moving_avg_forecast_6	14.566000	20.82
0	moving_avg_forecast_8	14.805000	21.06
0	moving_avg_forecast_12	15.236000	22.07
0	SES	36.796250	63.88
0	Holt_linear	70.572452	120.25
0	Holt_Winter	16.447061	22.88
0	Holt_Winter M	17.369490	28.88
0	Auto_ARIMA(3,1,3)	15.985092	26.08
0	Manual_ARIMA(2,1,2)	15.354879	22.77
0	Auto_SARIMA(2,1,4)(4,0,4,6)	28.449744	49.57
0	Auto_SARIMA(3,1,4)(4,0,4,12)	18.110420	30.18
0	Manual_SARIMA(3,1,1)(2,1,4,6)	18.359456	29.95
0	Manual_SARIMA(2,1,2)(4,1,2,12)	17.341983	26.69
0	Fulldata_Manual_SARIMA(2,1,2)(4,1,2,12)	44.366065	27.74
0	Fulldata_Manual_SARIMA(2,1,4)(4,0,4,6)	54.641890	28.42

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(2,1,4)(4,0,4,6))

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, each 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 the trend for future is down and the predicted future sales is also showing downward trend. The company needs to take some good measures.

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