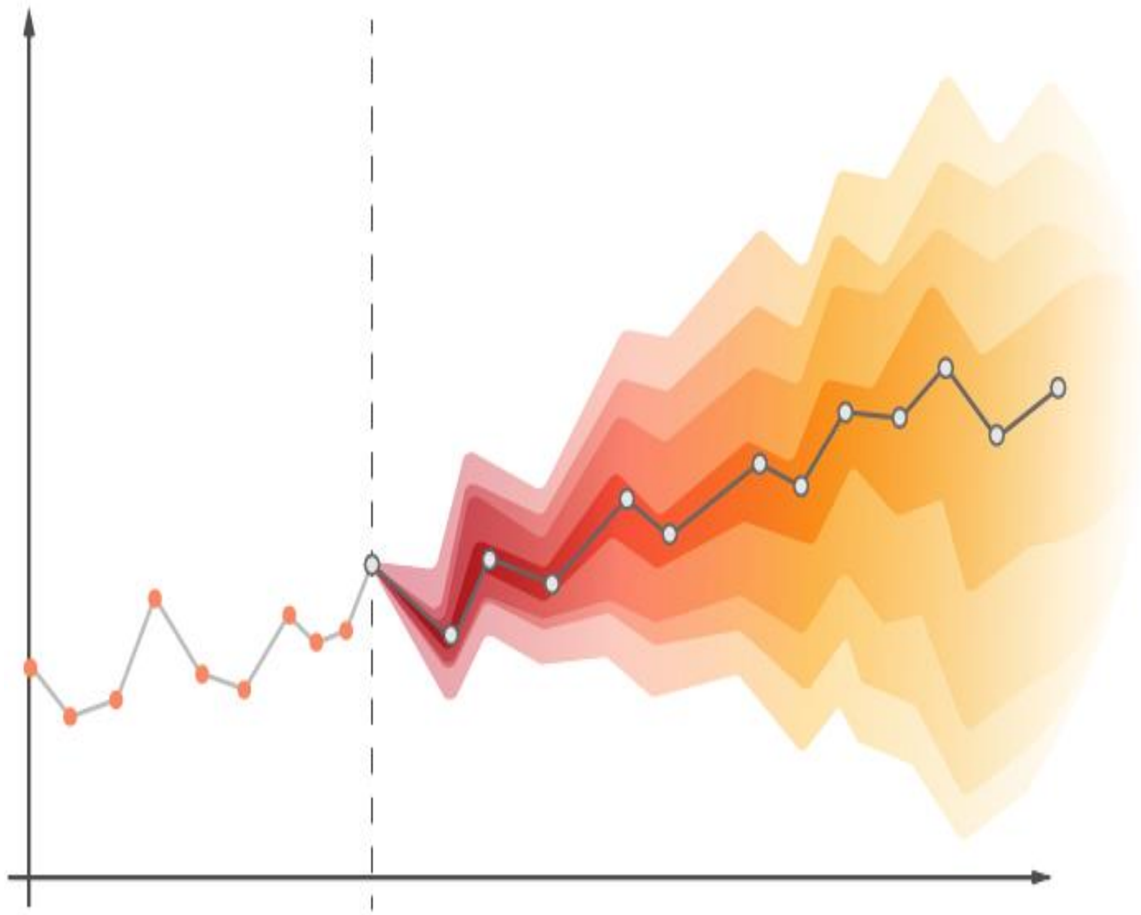


Time Series Forecasting

PROJECT

ZIANNA LYNGDOH | PGP-DSBA BATCH II

Criteria	Points
1. Read the data as an appropriate Time Series data and plot the data.	2
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.	5
3. Split the data into training and test. The test data should start in 1991.	2
4. Build all the 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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.	16
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also 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.	3
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.	8
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.	8
8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.	2
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.	3

Criteria	Points
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.	5

SPARKLING

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these 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.

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

BASIC INFORMATION OF THE DATASET

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

1. Head of data Sparkling.csv

	YearMonth	Sparkling
182	1995-03	1897
183	1995-04	1862
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

2. Tail of data Sparkling.csv

```
The number of rows:  187
The number of columns:  2
```

3. Shape of dataset

```
YearMonth    0
Sparkling    0
dtype: int64
```

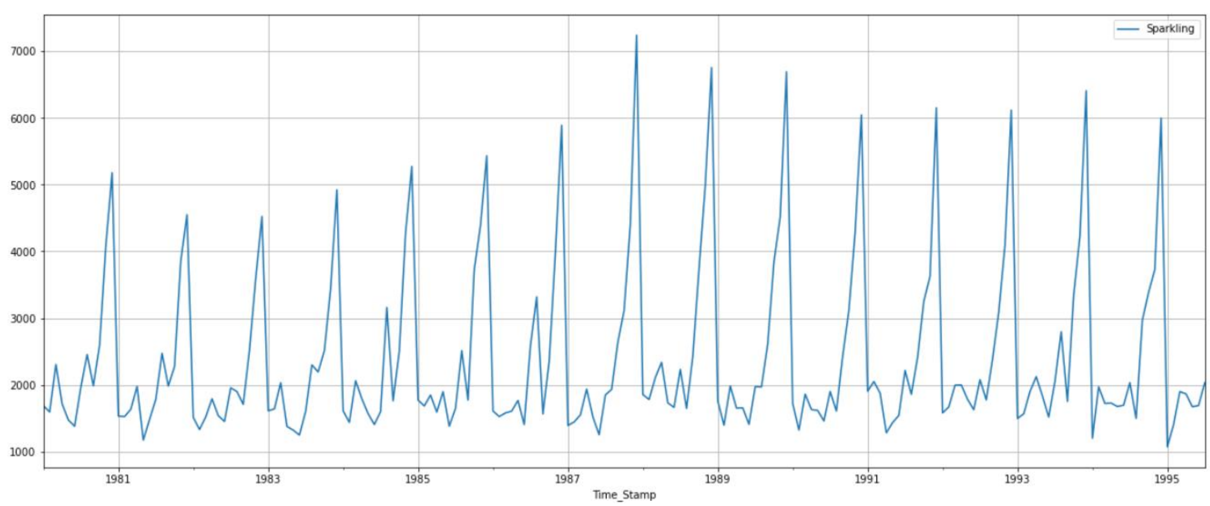
4. Null- Values in the dataset

	YearMonth	Sparkling	Time_Stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

5. Creating Time-Stamp in the dataset

	Sparkling
Time_Stamp	
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

6. Setting Time-Stamp at the Index in the dataset, after dropping 'YearMonth' Column



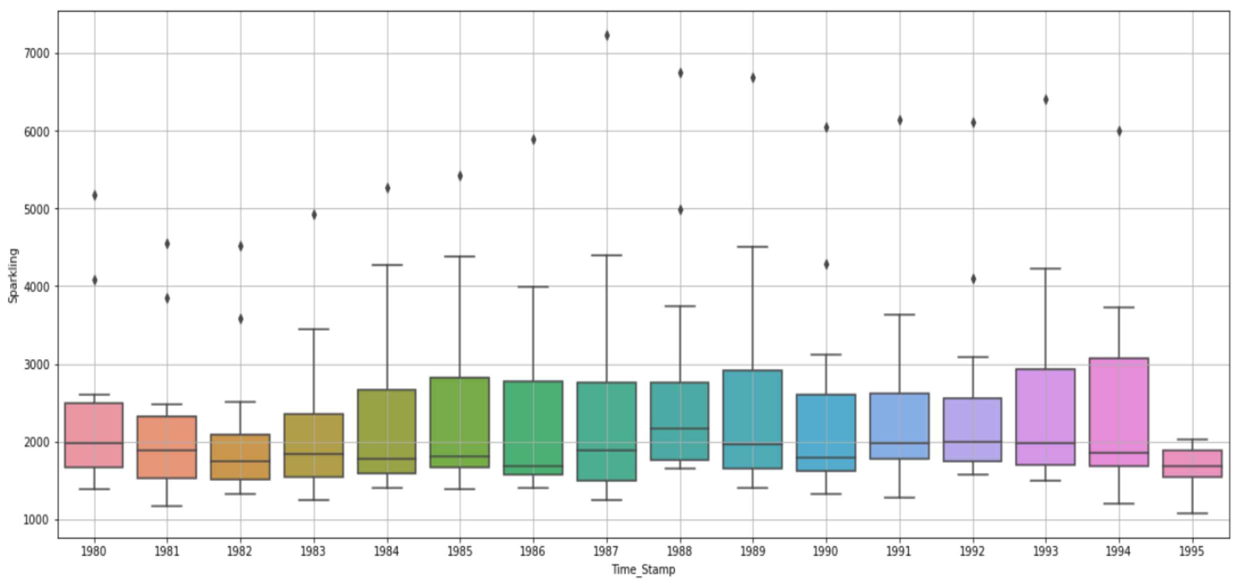
7. Plotting graph of the dataset

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

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

8. Description of the dataset

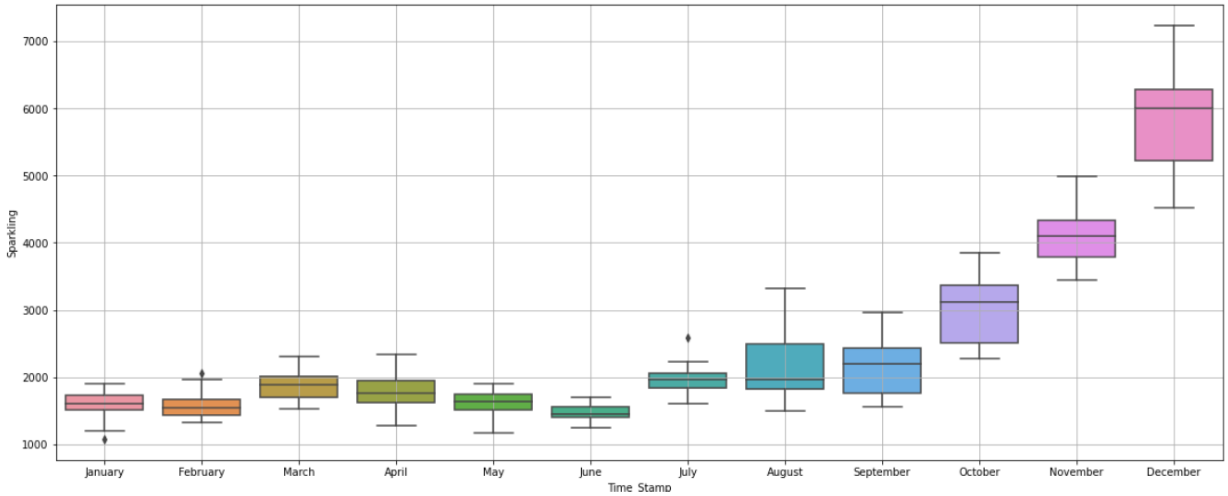
YEARLY BOXPLOT:



9. Yearly Boxplot showcasing the sale of Sparkling Wine

The yearly boxplots show us how the sale has increased and decreased over the past few years. The highest number of sales being recorded in the year 1989

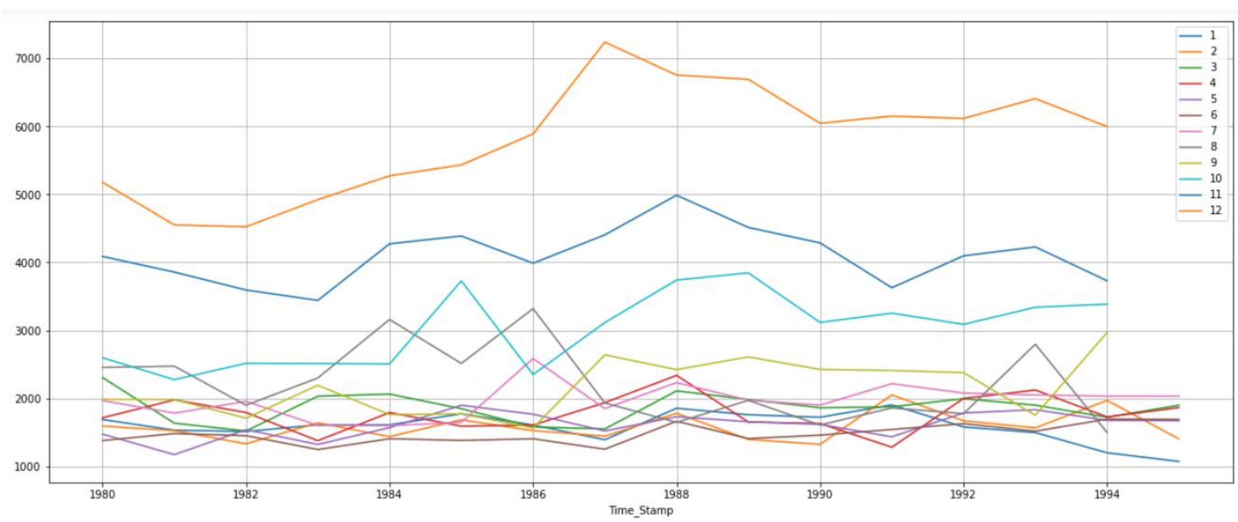
MONTHLY BOXPLOT:



10. Monthly Boxplot showcasing the sale of Sparkling Wine

From the above monthly boxplot, it is clear that the sale of Sparkling wine is the highest in the month of December.

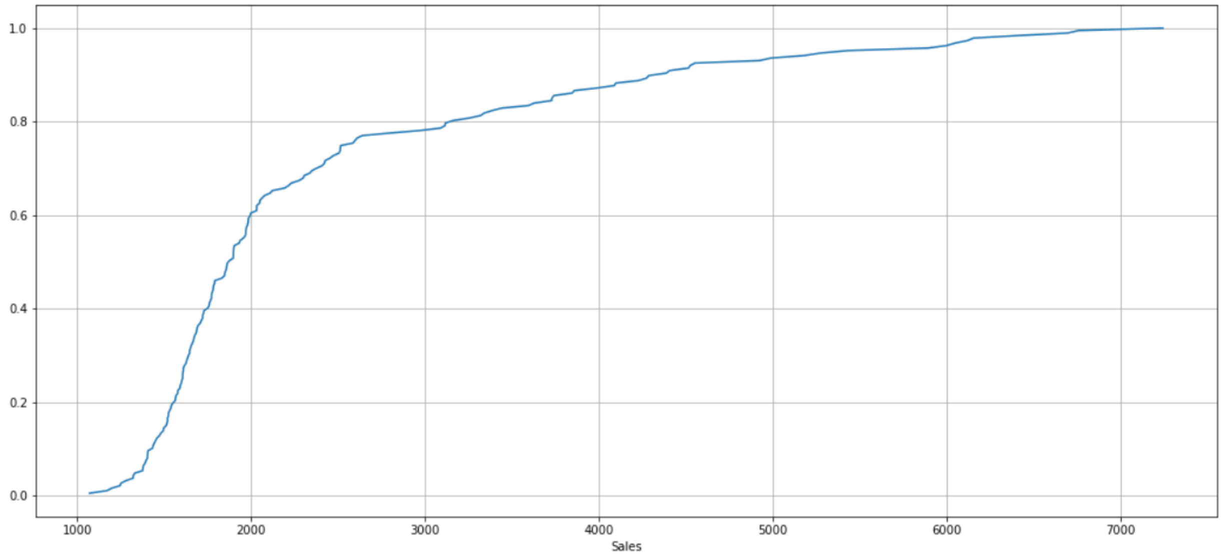
MONTHLY SALE ACROSS THE YEAR:



11. Graph showcasing the sale of Sparkling Wine across the year for every month

From the above graph it can be seen that the highest sale was in the year 1987 for the month of December.

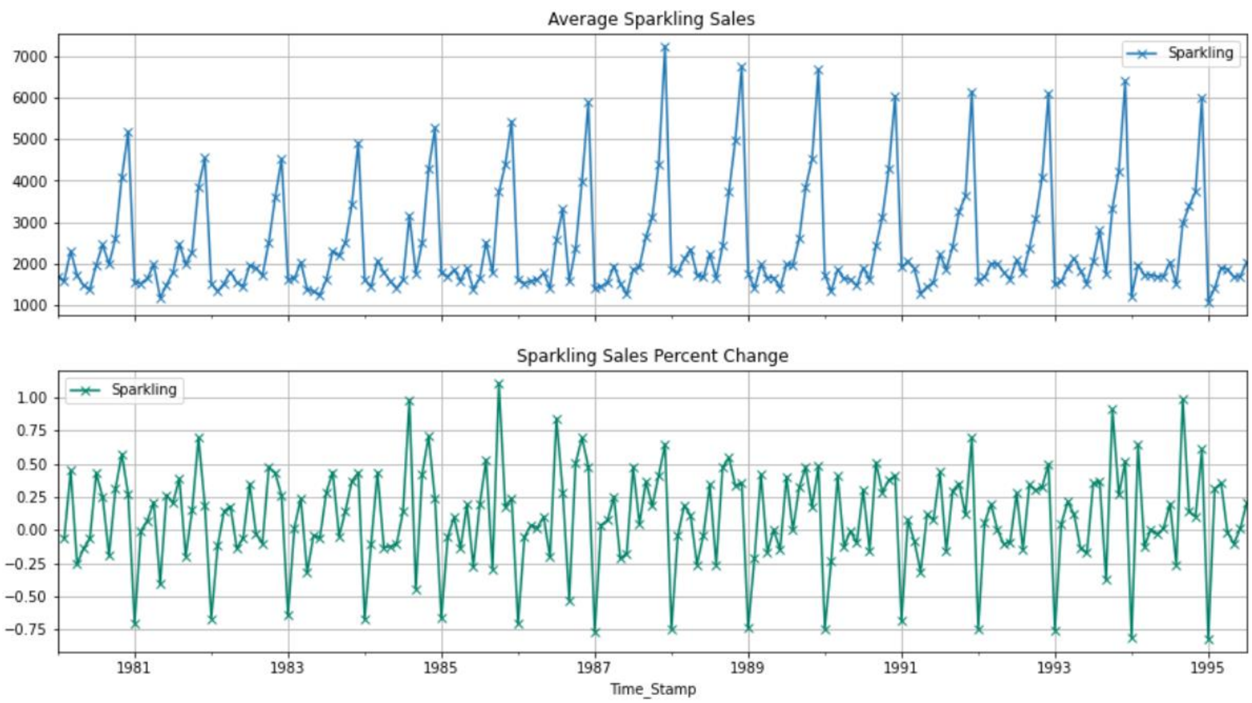
Empirical Cumulative Distribution:



12. Empirical Cumulative Distribution Graph

This particular graph tells us what percentage of data points refer to what number of Sales.

Next, we will be plotting the average Sparkling Sales per month and the month-on-month percentage change of Sparkling Sales.



13. Graph showcasing Average Sparkling sales, and its percentage change over the years.

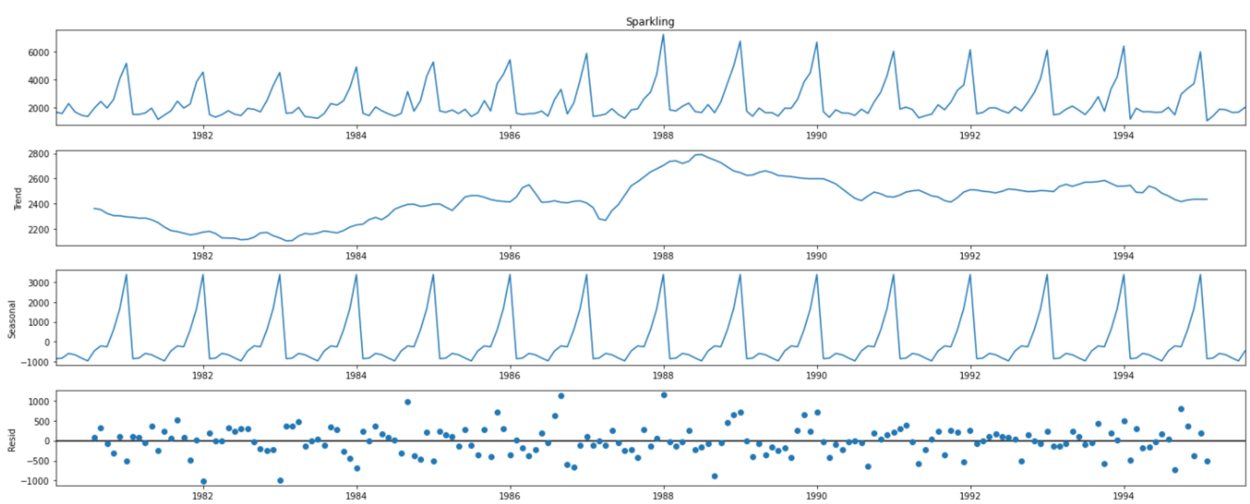
From the above graph it can be seen that the average Sparkling sale is the highest at the year 1987-1988.

From the above Exploratory data analysis it is found that:

- The data starts at 1980 and ends at 1995.
- There are 187 rows and 2 columns.
- There is no null value present in the dataset.
- We then create a time stamp and then replace it at index, while dropping the YearMonth column.

DECOMPOSITION:

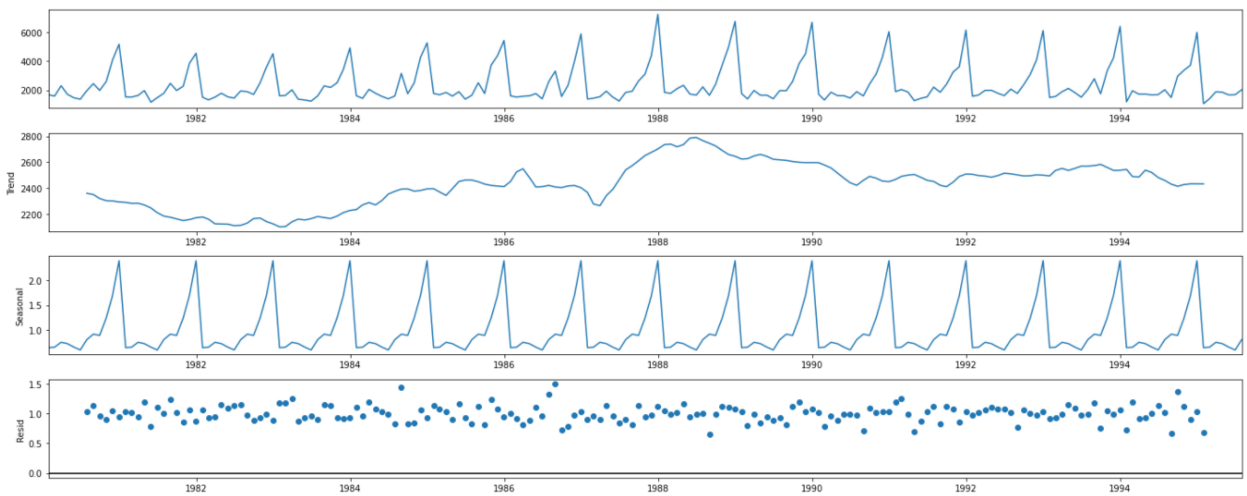
ADDITIVE DECOMPOSITION:



14. Additive Decomposition.

On performing Additive Decomposition on the dataset, it is observed that there are residuals located near 0, which gives pattern that is non-essential for model building. We then go forward with Multiplicative Decomposition.

MULTIPLICATIVE DECOMPOSITION:



14. Multiplicative Decomposition

On performing Multiplicative Decomposition on the dataset, it is observed that the residuals are now located near 1. Multiplicative Decomposition will be used for building our models.

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

The data will be split into train and test and plotting the training and test data will be done. Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

First few rows of Training Data

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

First few rows of Test Data

Time_Stamp	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Last few rows of Training Data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Last few rows of Test Data

Time_Stamp	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

15. First and last five rows of Training & Test data

4. Build all the 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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

a. <u>SIMPLE EXPONENTIAL SMOOTHING:</u>

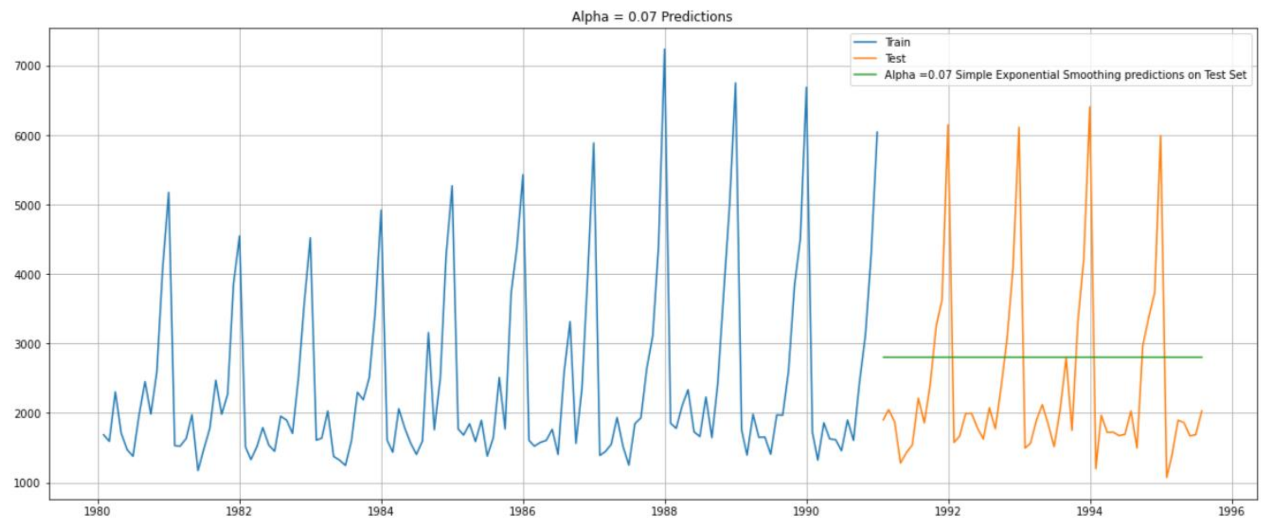
Single Exponential Smoothing, SES for short, also called Simple Exponential Smoothing, is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

To build an SES model, we first autofit the Smoothing Model and ask python to choose the optimal parameters, this is done by using *<u>autofit.params</u>* function, on passing the said function we get the following optimal parameter for SES.

```
{'smoothing_level': 0.07029459943040381,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1764.1004162520212,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

16. Optimal Parameters for Single Exponential Smoothing

It is found that python, has optimized the smoothing level to be at 0.07.



17. Graph: Forecasted Values for Simple Exponential Smoothing

The above graph showcases the forecasted values for Train and Test dataset using Simple Exponential Smoothing

We then proceed to calculate the Root Mean Square Error for the model being built using Simple Exponential Smoothing.

RMSE VALUE:

```
Simple Exponential Smoothing RMSE: 1338.0121443910186
```

	Test RMSE
Alpha= 0.07, SES	1338.012144

18. Test RMSE Value: SES

b. DOUBLE EXPONENTIAL SMOOTHING

Double exponential smoothing employs a level component and a trend component at each period. Double exponential smoothing uses two weights alpha & beta, (also called smoothing parameters), to update the components at each period.

The optimal parameters for Double Exponential Smoothing are as follows:

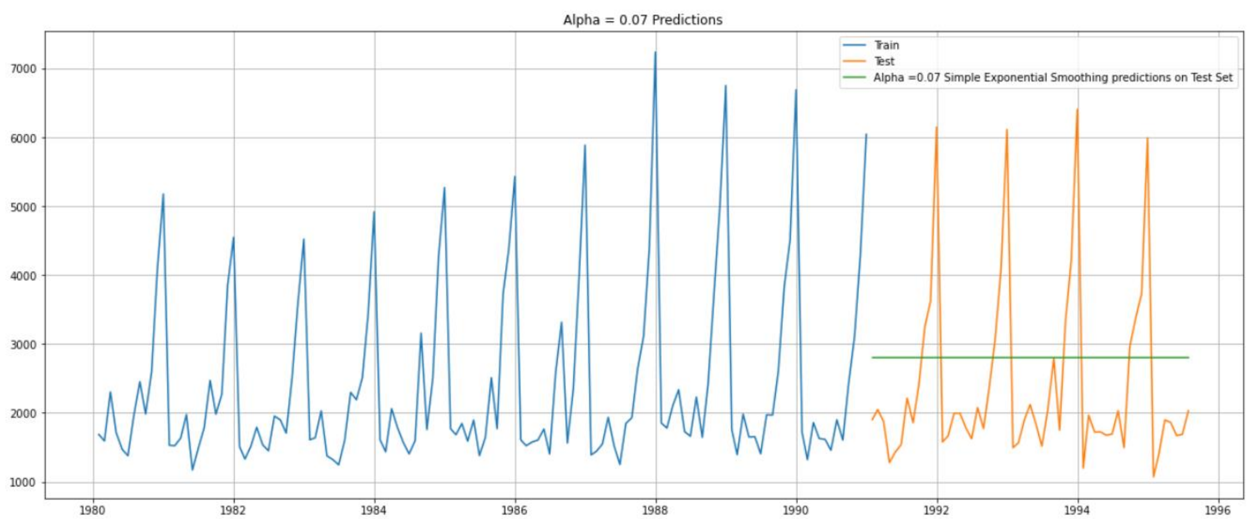
```
==Holt model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 1.9086427682180844e-08, 'smoothing_trend': 7.302464353829351e-09, 'smoothing_seasonal': nan, 'damping_tren
d': nan, 'initial_level': 137.81629861505857, 'initial_trend': -0.4943753249082896, 'initial_seasons': array([], dtype=float6
4), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

19. Optimal Parameters for Double Exponential Smoothing

It is found that python, has optimized the smoothing level to be at Alpha= 1.908, Beta= 7.302

The above data is used for forecasting on the test set.



20. Graph: Forecasted Values for Double Exponential Smoothing

The above graph shows the forecast values using train and test sets, it can be noticed that the Double Exponential Smoothing has been able to pick up the trend of the data.

RMSE VALUE:

```
Double Exponential Smoothing RMSE: 3949.993290409098
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290

21. Test RMSE Value: SES & DES

c. TRIPLE EXPONENTIAL SMOOTHING

Triple exponential smoothing is used to handle the time series data containing a seasonal component. This method is based on three smoothing equations: stationary component(alpha), trend(beta), and seasonal(gamma). Both seasonal and trend can be additive or multiplicative.

The optimal parameters for Triple Exponential Smoothing are as follows:

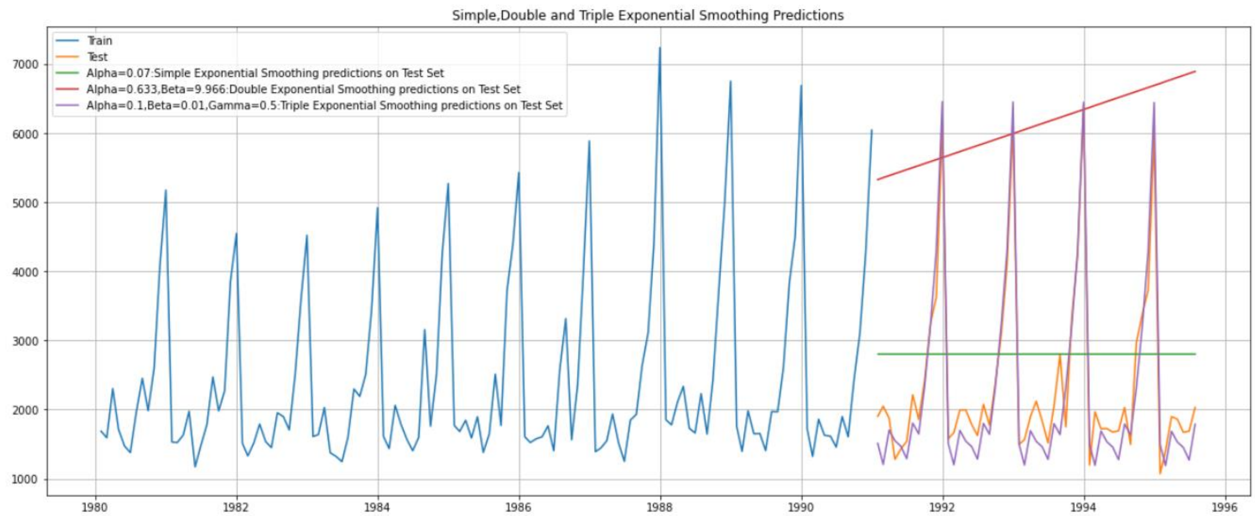
```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.08830330642635406, 'smoothing_trend': 6.730635331927582e-05, 'smoothing_seasonal': 0.004455138229351625,
'damping_trend': nan, 'initial_level': 146.88752868155674, 'initial_trend': -0.5492163940406024, 'initial_seasons': array([-31.
12207537, -18.81171138, -10.86052241, -21.52235816,
       -12.68359535,  -7.17529564,   2.7456236 ,   8.84900094,
         4.85724354,   2.9520333 ,  21.05004912,  63.29916317]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

22. Optimal Parameters for Triple Exponential Smoothing

It is found that python, has optimized the smoothing level to be at Alpha= 10.088, Beta= 6.730, Gamma= 0.004

The above data is used for forecasting on the test set.



23. Graph: Forecasted Values for Single, Double & Triple Exponential Smoothing

The above graph shows the forecast values using train and test sets, it showcases Single, Double & Triple Exponential Smoothing on the data. It can be noticed that the Triple Exponential Smoothing has been able to pick up the trend, seasonality along with the level of the data.

RMSE VALUE:

```
Triple Exponential Smoothing RMSE: 379.6956857387101
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686

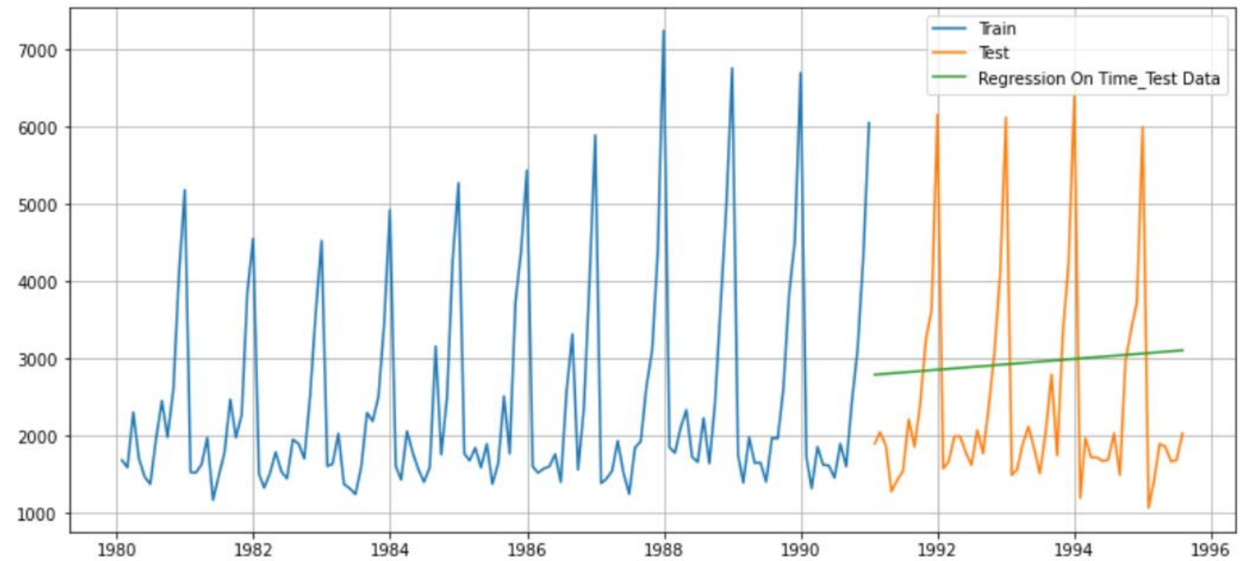
24. RMSE Values for Single, Double & Triple Exponential Smoothing

TES model shows very low RMSE value and thus is a good model that can be used for forecasting.

REGRESSION:

Time series regression is a statistical method for predicting a future response based on the response history (known as autoregressive dynamics) and the transfer of dynamics from relevant predictors.

For a regression model to be built we take the train and test data and fit it using Linear Regression.



25. Graph showcasing the regression model forecasting

From the above graph it can be seen that the regression model is not showing a very good forecasting when compared to the other models previously built such as the Triple Exponential Smoothing model.

RMSE:

```
For RegressionOnTime forecast on the Test Data,  RMSE is 1389.135
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175

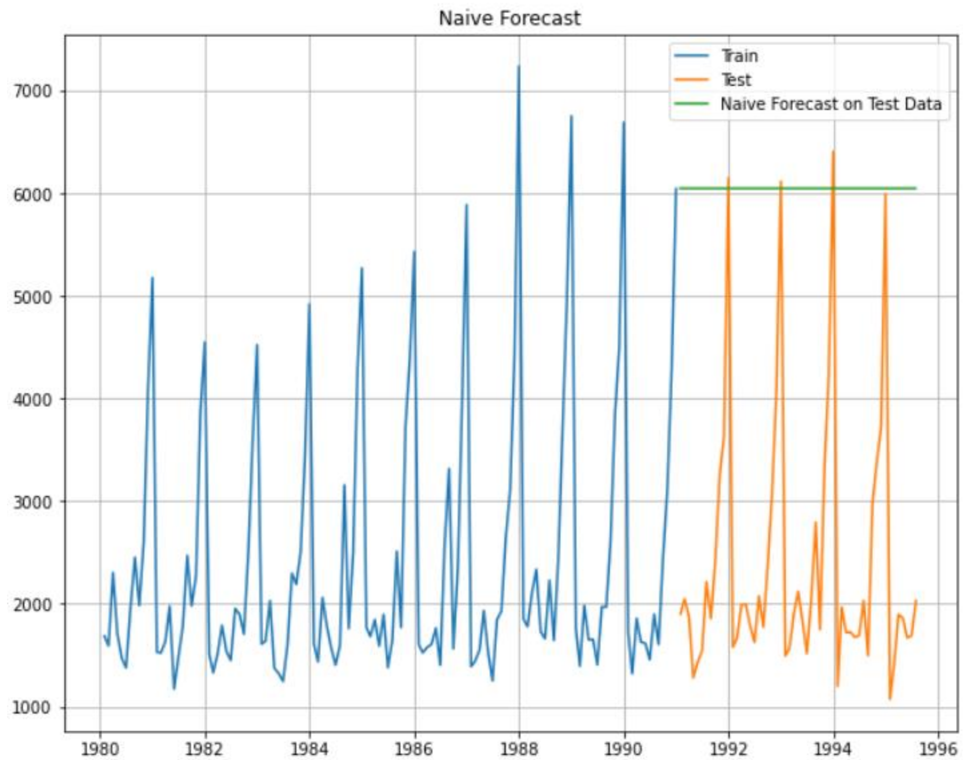
26. RMSE Values: REGRESSION along with other models built above

The RMSE value for Regression is very high as compared to Simple and Triple Exponential model built previously.

The RMSE for Regression is very low as compared to Double Exponential Smoothing, overall Regression is not the best fit model that can be used for forecasting.

NAÏVE APPROACH:

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.



27. Graph: NAÏVE APPROACH forecasting

The graph showcases the forecasting using naïve approach, in the above graph it can be seen that the Naïve forecast on test data is very high as compared to the actual data.

RMSE:

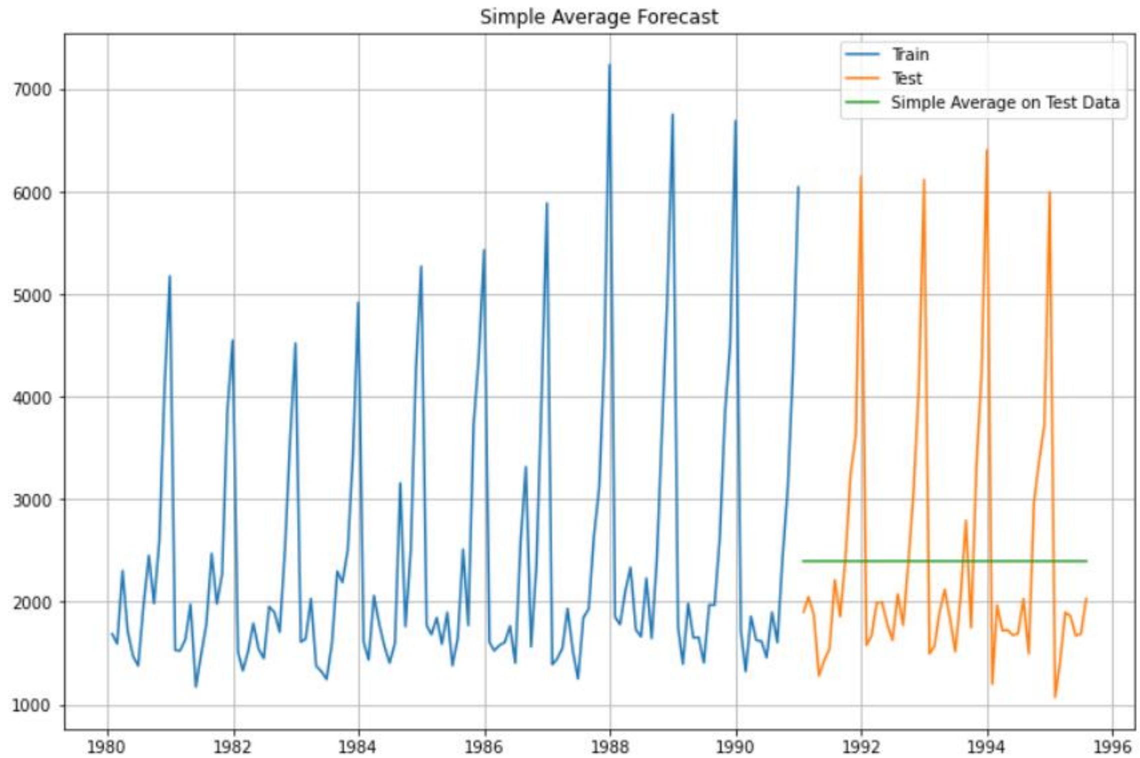
```
For RegressionOnTime forecast on the Test Data,  RMSE is 3864.279
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352

28. RMSE Value: Naïve Approach along with previously built models

SIMPLE AVERAGE:

A simple moving average (SMA) is the simplest type of technique of forecasting. Basically, a simple moving average is calculated by adding up the last 'n' period's values and then dividing that number by 'n'. So the moving average value is considering as the forecast for next period.



29. Graph: SIMPLE AVERAGE model

The graph for simple average showcases that the forecast sale is very low when compared to other models.

RMSE:

```
For Simple Average forecast on the Test Data,  RMSE is 1275.082
```

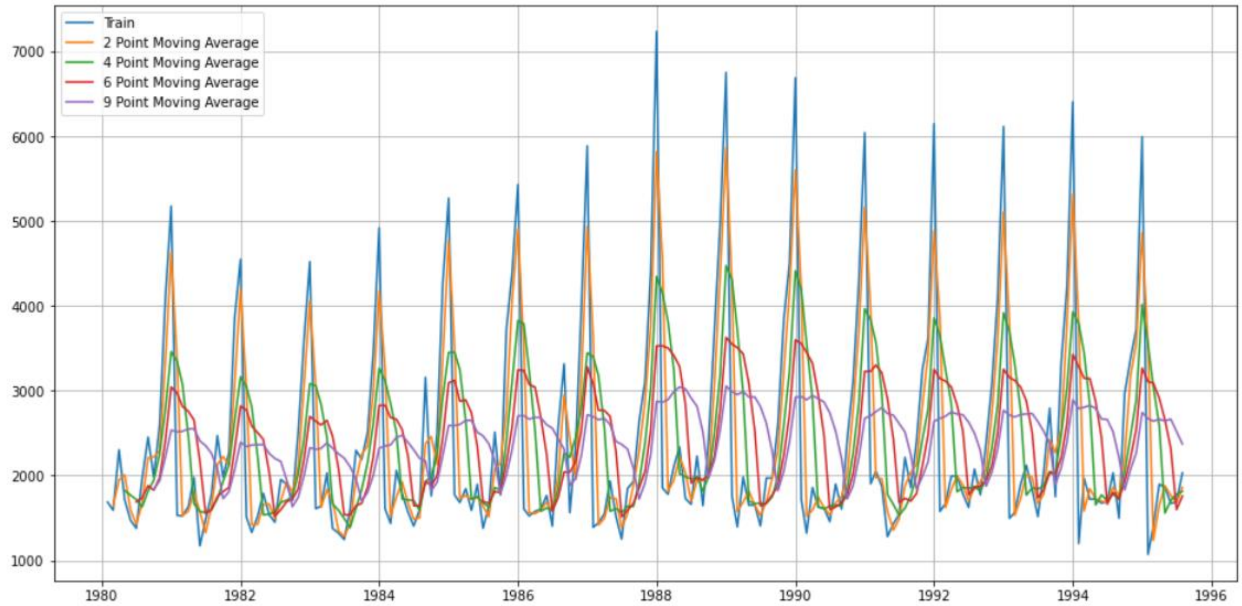
	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804

30. RMSE Values: Simple Average Model

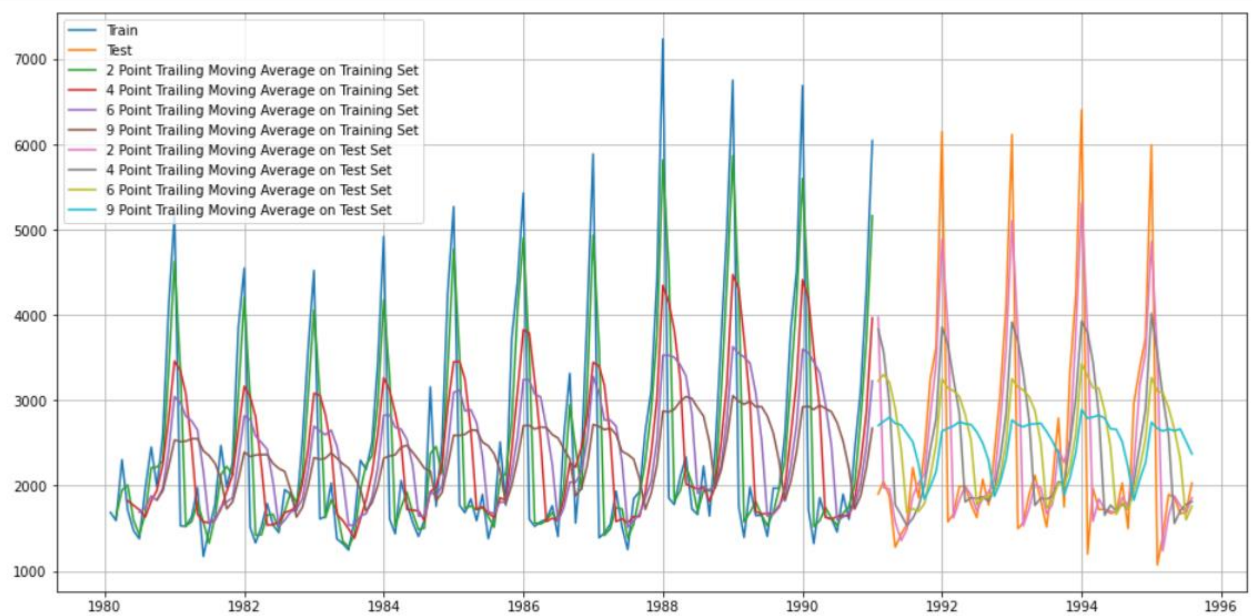
The RMSE value for Simple Average model is low as compared to Naïve and Regression models, but it is significantly higher than Triple Exponential Smoothing Model.

MOVING AVERAGE:

A moving average is a technique that calculates the overall trend in a data set. In operations management, the data set is sales volume from historical data of the company.



31. Graph: Moving Average Model Trailing points



32. Graph: Moving Average Model forecasting

From the above graph it can be noted that the 2 Point Trailing moving average is performing well on test as compared to other Trailing points.

RMSE:

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 1346.278
```

33. RMSE Values for 2, 4, 6 & 9 point Moving Average

Compared to 4,6&9 point Moving Average RMSE, the RMSE for 2 point Moving Average has a very low value.

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315

34. RMSE Values for Moving Average model, and all previously built models

Compared to previously built models the Moving Average has the second lowest RMSE value, but it is still very high as compared to the TES model.

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

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

H_0 : The Time Series has a unit root and is thus non-stationary.

H_1 : The Time Series does not have a unit root and is thus stationary. We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

```
DF test statistic is -1.798
DF test p-value is 0.7055958459932035
Number of lags used 12
```

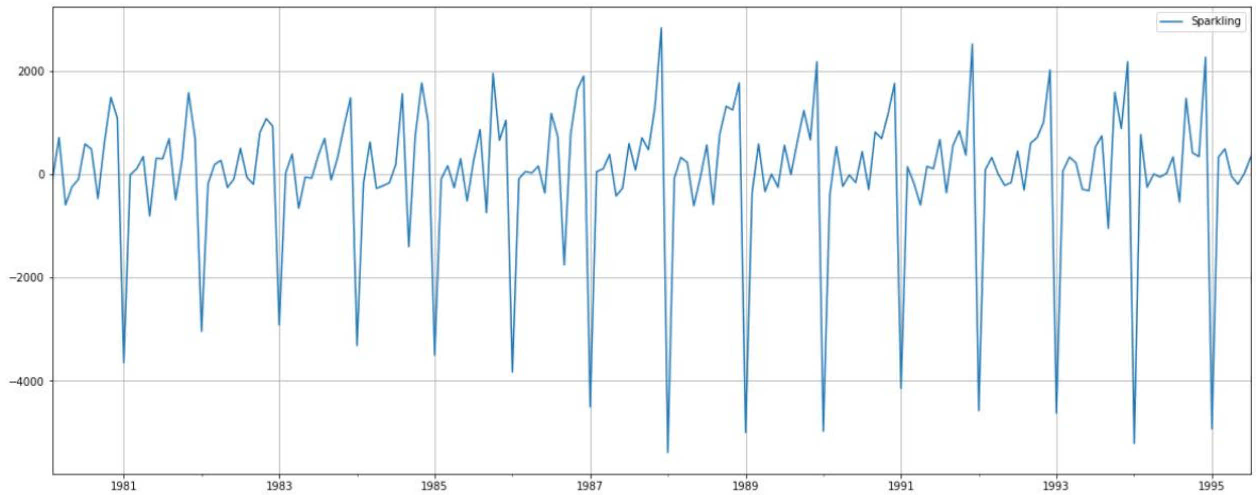
35. ADF Test using stats model

From the above test we can see that the dataset is not stationary at 5% significant level. To make the data stationary we will use one level of differencing and rerun the code.

```
DF test statistic is -44.912
DF test p-value is 0.0
Number of lags used 10
```

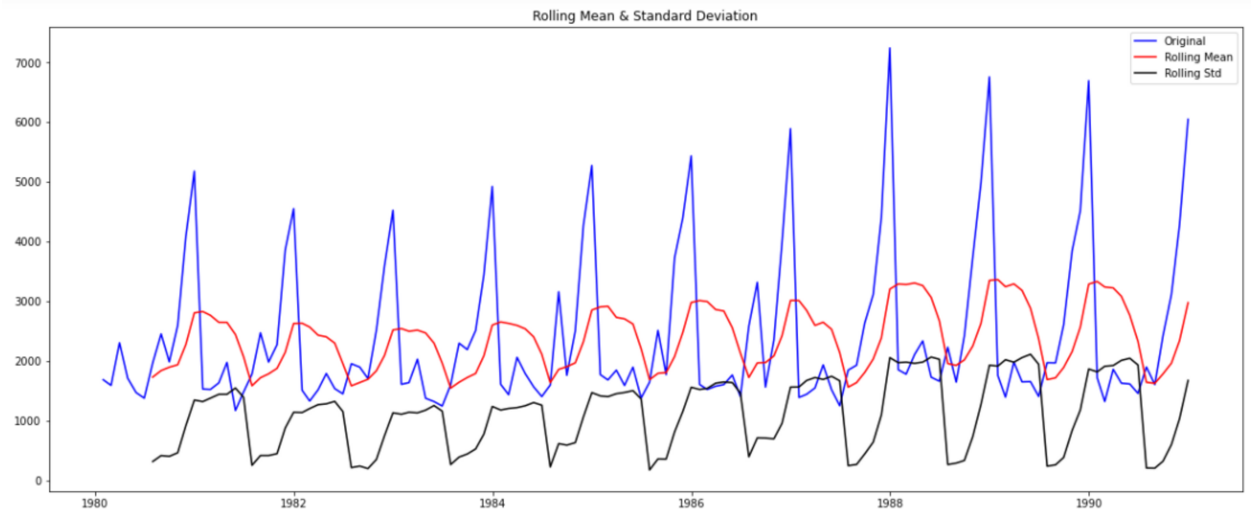
36. ADF Test with difference at one level

On using one level differencing we can see that the data is now stationary.



37. Dataset Plot to check stationarity

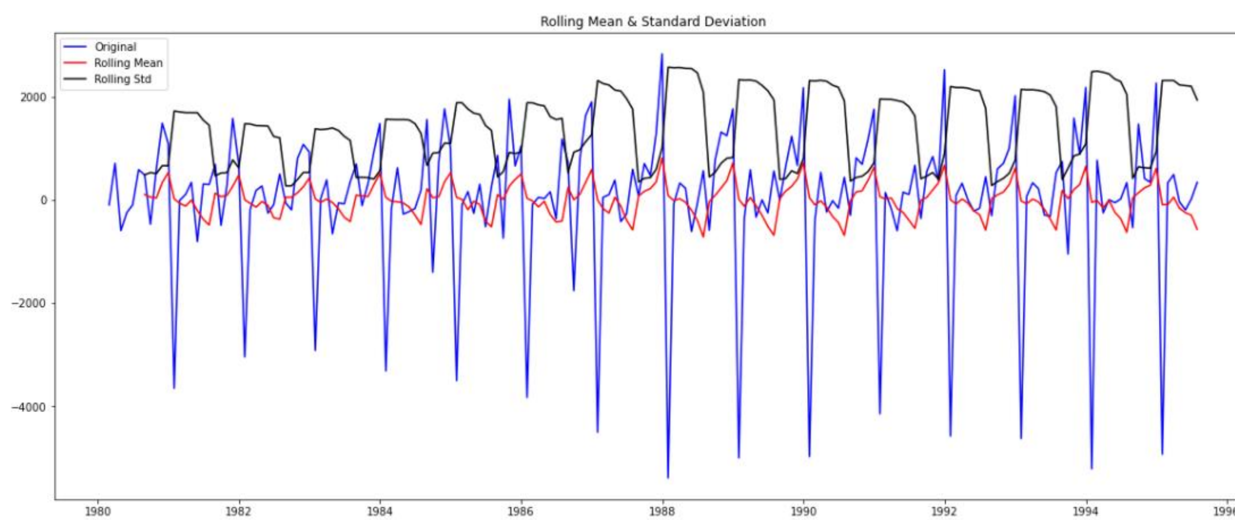
We will now check the stationarity of the Training dataset.




```
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
```

38. Training Dataset Stationarity check

We see that the Training dataset is not stationarity, we will use the difference order 1 to make the data stationary.




```
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
```

39. Training Dataset Stationarity using difference order 1

We see that at α = 0.05 the Time Series Training dataset is now indeed stationary.

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.

AUTOMATED ARIMA:

Auto Regressive Integrated Moving Average (ARIMA) model is among one of the more popular and widely used statistical methods for time-series forecasting. It is a class of statistical algorithms that captures the standard temporal dependencies that is unique to a time series data.

We use different parameter combinations to find the value of p,d,q to build our model. In automated ARIMA python does the permutation and combination to arrive at the best values for p,d,q.

```
ARIMA(0, 1, 0) - AIC:2269.582796371201
ARIMA(0, 1, 1) - AIC:2264.9064368172944
ARIMA(0, 1, 2) - AIC:2232.783097684661
ARIMA(1, 1, 0) - AIC:2268.5280606648653
ARIMA(1, 1, 1) - AIC:2235.0139453510988
ARIMA(1, 1, 2) - AIC:2233.597647119537
ARIMA(2, 1, 0) - AIC:2262.035600095461
ARIMA(2, 1, 1) - AIC:2232.3604898848293
ARIMA(2, 1, 2) - AIC:2210.616429928409
```

40. Various ARIMA combination

	param	AIC
8	(2, 1, 2)	2210.616430
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

41. Sorting the ARIMA values based on lowest AIC

We see that the automated ARIMA model can be built using (2,1,2) as it has the lowest AIC.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.308
Method:                       css-mle   S.D. of innovations           1012.036
Date:                Sun, 20 Mar 2022   AIC                           2210.616
Time:                        18:21:42   BIC                           2227.868
Sample:                    02-29-1980   HQIC                          2217.626
                         - 12-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5847      0.517     10.803      0.000       4.571       6.598
ar.L1.D.Sparkling     1.2702      0.074     17.052      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5604      0.074     -7.621      0.000      -0.705      -0.416
ma.L1.D.Sparkling    -1.9992      0.042    -47.131      0.000      -2.082      -1.916
ma.L2.D.Sparkling     0.9992      0.042     23.572      0.000       0.916       1.082
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1333           -0.7071j            1.3358           -0.0888
AR.2            1.1333           +0.7071j            1.3358            0.0888
MA.1            1.0004           -0.0007j            1.0004           -0.0001
MA.2            1.0004           +0.0007j            1.0004            0.0001
-----------------------------------------------------------------------------
```

42. Automated ARIMA Result

RMSE:

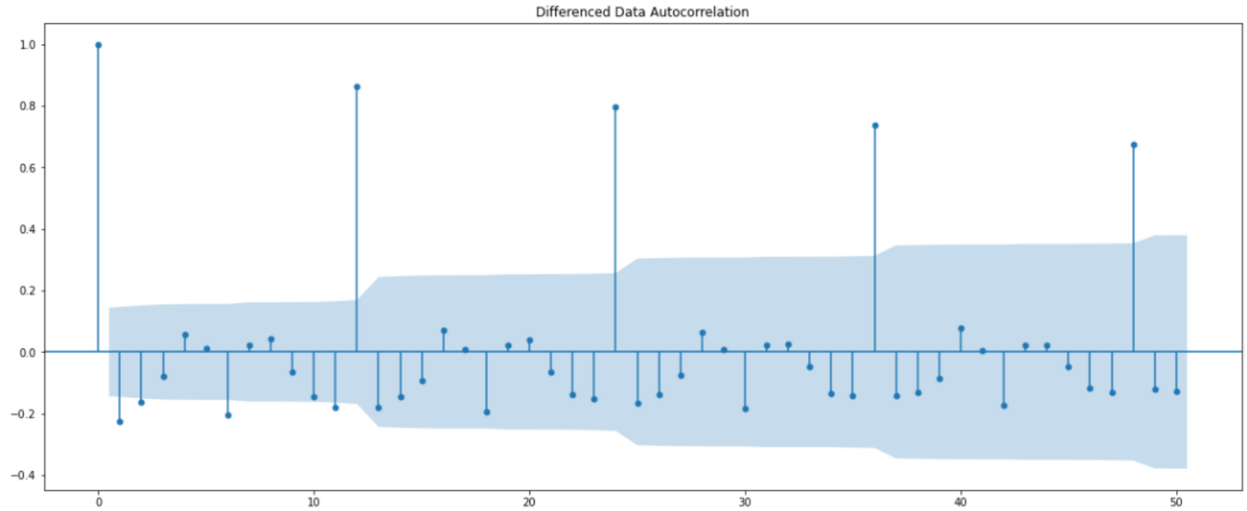
For Automated Arima in Test Data, RMSE is `1374.9580479969213`

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048

43. RMSE Values for Automated ARIMA & all the other models built

AUTOMATED SARIMA:

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.



44. ACF Plot showcasing seasonality

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.

SARIMA at 6

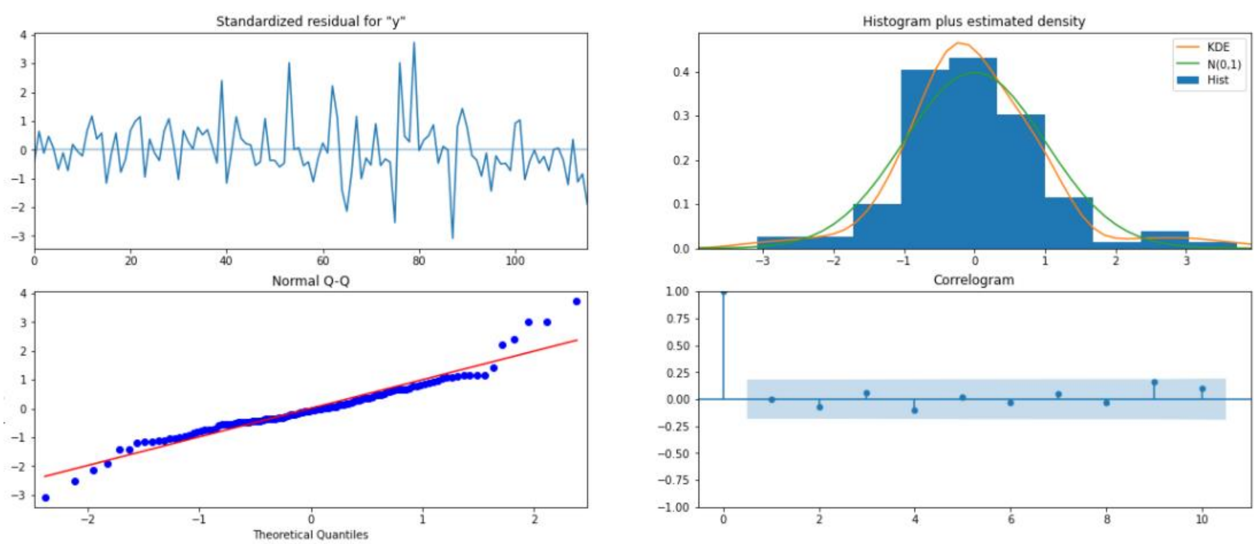
	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888803
80	(2, 1, 2)	(2, 0, 2, 6)	1729.178594
53	(1, 1, 2)	(2, 0, 2, 6)	1729.934495
17	(0, 1, 1)	(2, 0, 2, 6)	1741.696451
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379778

45. Automated SARIMA AIC values in ascending order

```
                                      SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -856.944
Date:                            Sun, 20 Mar 2022   AIC                           1727.889
Time:                                    18:22:04   BIC                           1747.164
Sample:                                         0   HQIC                          1735.713
                                            - 132                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7851      0.103     -7.655      0.000      -0.986      -0.584
ma.L2         -0.0975      0.112     -0.870      0.384      -0.317       0.122
ar.S.L6        0.0022      0.026      0.084      0.933      -0.048       0.053
ar.S.L12       1.0396      0.018     58.260      0.000       1.005       1.075
ma.S.L6        0.0428      0.143      0.298      0.765      -0.238       0.324
ma.S.L12      -0.6203      0.090     -6.878      0.000      -0.797      -0.444
sigma2      1.475e+05   1.42e+04     10.372      0.000     1.2e+05    1.75e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                38.96
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               2.85   Skew:                             0.58
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.59
===================================================================================

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

46. SARIMAX Result



47. Plot Diagnostic for SARIMA

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.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1375.580323	384.072914	622.811244	2128.349403
1	1116.607983	392.839126	346.657445	1886.558521
2	1667.556841	395.413558	892.560508	2442.553174
3	1528.328132	397.974441	748.312561	2308.343702
4	1372.199160	400.519004	587.196337	2157.201983

48. SARIMA predicted summary frame

RMSE:

```
For SARIMA on the Test Data,  RMSE is 601.353
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048
SARIMA(0,1,2)(2,0,2,6)	601.353424

49. RMSE Values for Automated SARIMA & all the other models built

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

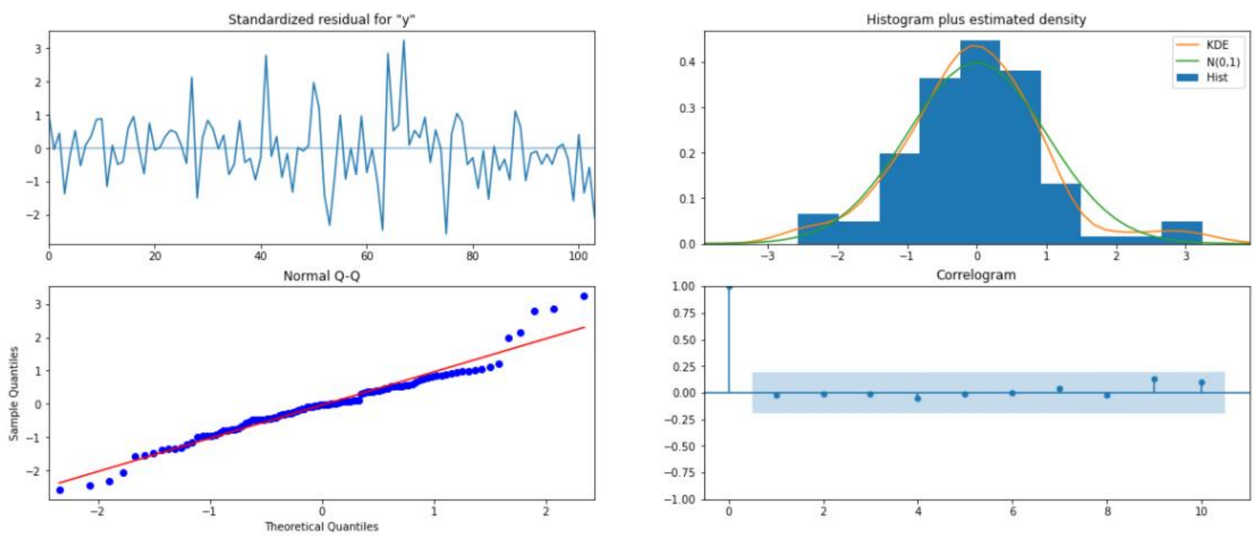
	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934563
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121563
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

50. SARIMA values based on lowest AIC values

```
                                     SARIMAX Results                                     
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 20 Mar 2022   AIC                           1555.584
Time:                                    18:22:32   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     72.836      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   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 (complex-step).
```

51. SARIMA at 12 Result



52. Plot Diagnostic for SARIMA at 12

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:

```
For SARIMA on the Test Data,  RMSE is 528.588
```

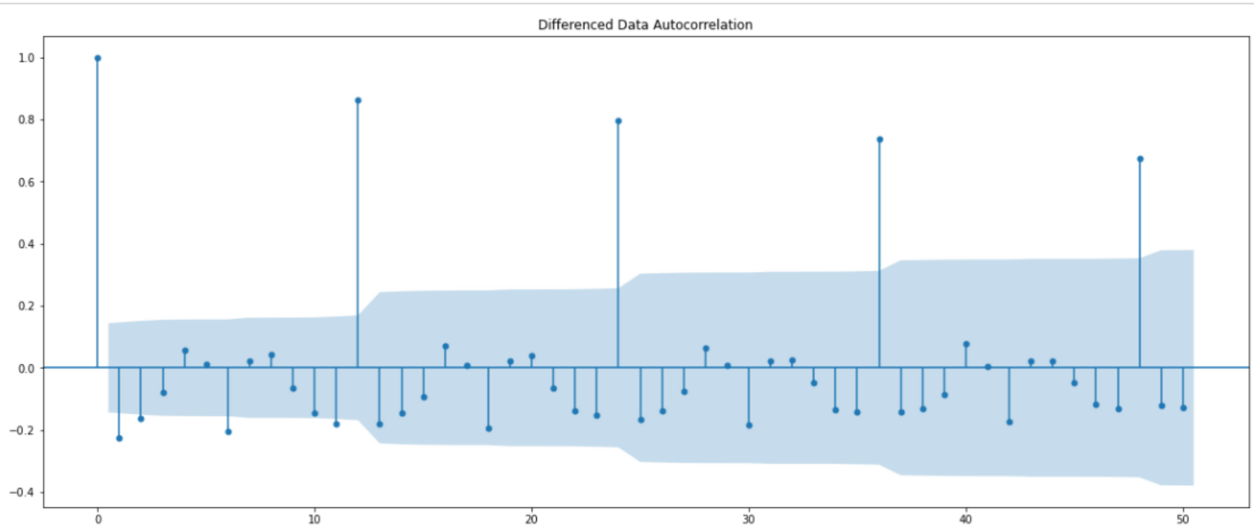
	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048
SARIMA(0,1,2)(2,0,2,6)	601.353424
SARIMA(1,1,2)(1,0,2,12)	528.588034

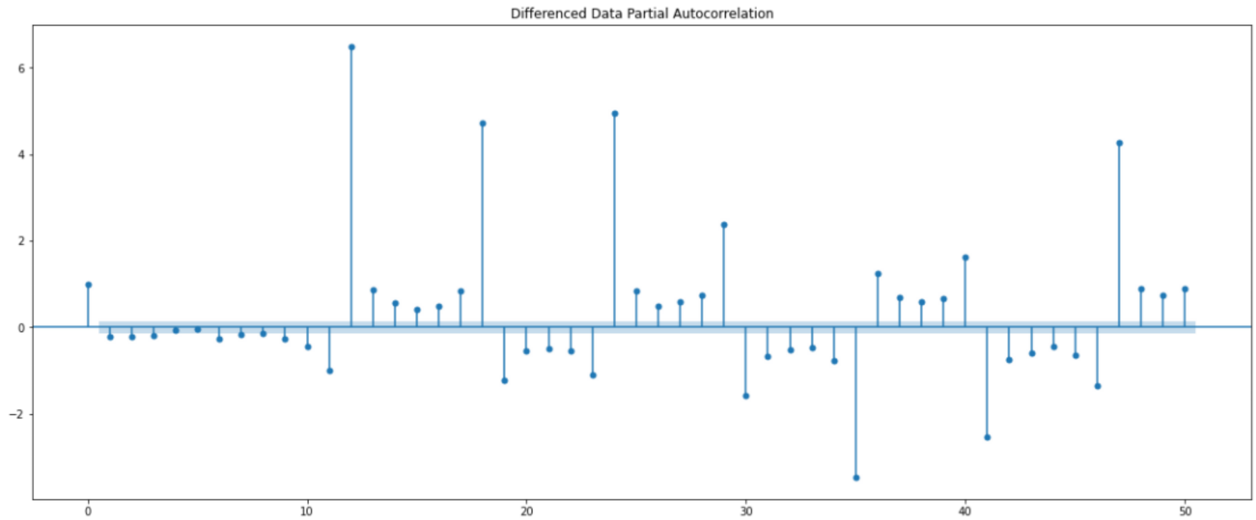
53. RMSE Values for Automated SARIMA at 12 & all the other models built

We see that the RMSE value has only reduced further by a margin when the seasonality parameter was changed to 12, and there is not much difference in the model diagnostics of the two models.

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.

ARIMA:





54. ACF & PACF Plot

Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0. By looking at the above plots, we will take the value of p and q to be 0 and 0 respectively.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(0, 1, 2)   Log Likelihood               -1112.392
Method:                       css-mle   S.D. of innovations           1159.696
Date:                Sun, 20 Mar 2022   AIC                           2232.783
Time:                        18:22:33   BIC                           2244.284
Sample:                    02-29-1980   HQIC                          2237.456
                         - 12-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 6.2472      3.800      1.644      0.100      -1.201      13.696
ma.L1.D.Sparkling    -0.5555      0.073     -7.583      0.000      -0.699      -0.412
ma.L2.D.Sparkling    -0.4445      0.071     -6.247      0.000      -0.584      -0.305
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -2.2495           +0.0000j            2.2495            0.5000
-----------------------------------------------------------------------------
```

55. ARIMA Model Result

RMSE:

```
For MANUAL ARIMA on the Test Data,  RMSE is 1417.502
```

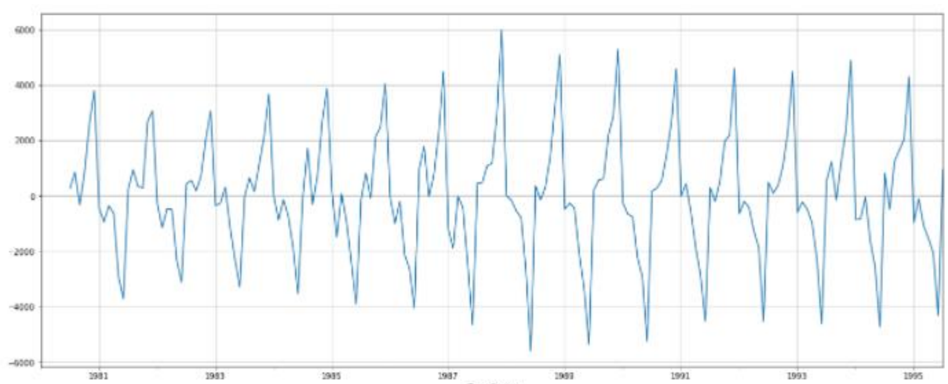
	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048
SARIMA(0,1,2)(2,0,2,6)	601.353424
SARIMA(1,1,2)(1,0,2,12)	528.588034
MANUAL ARIMA(0,1,2)	1417.502239

56. RMSE Values for Manual ARIMA & all the other models built

SARIMA:

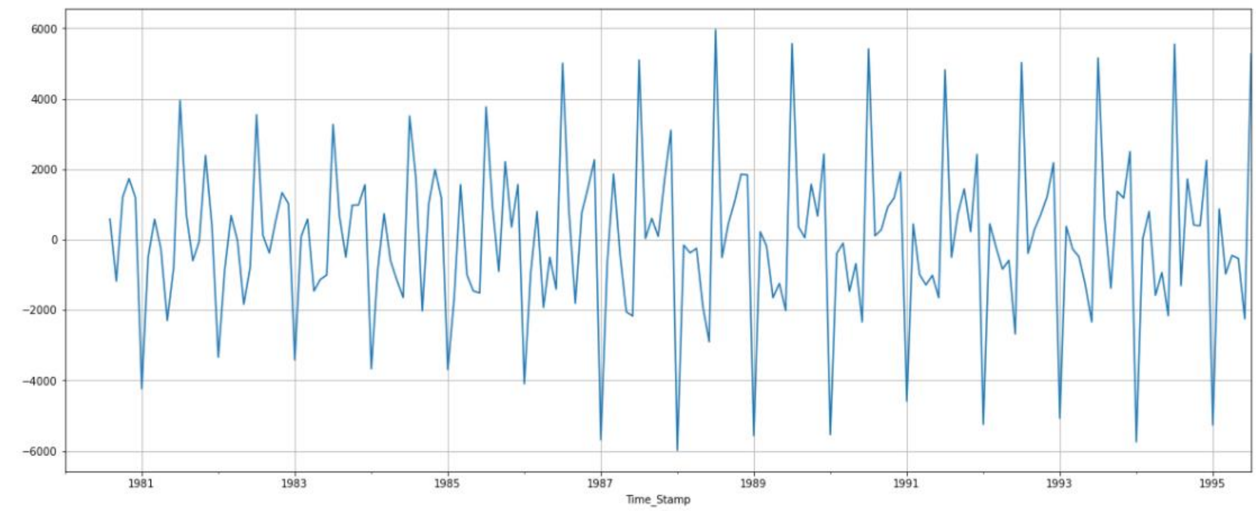
On obseriving the ACF & PACF plots above we notice that there is a trend and a seasonality. So, now we take a seasonal differencing and check the series.

Difference at 6:



57. Graph for SARIMA at 6: data frame

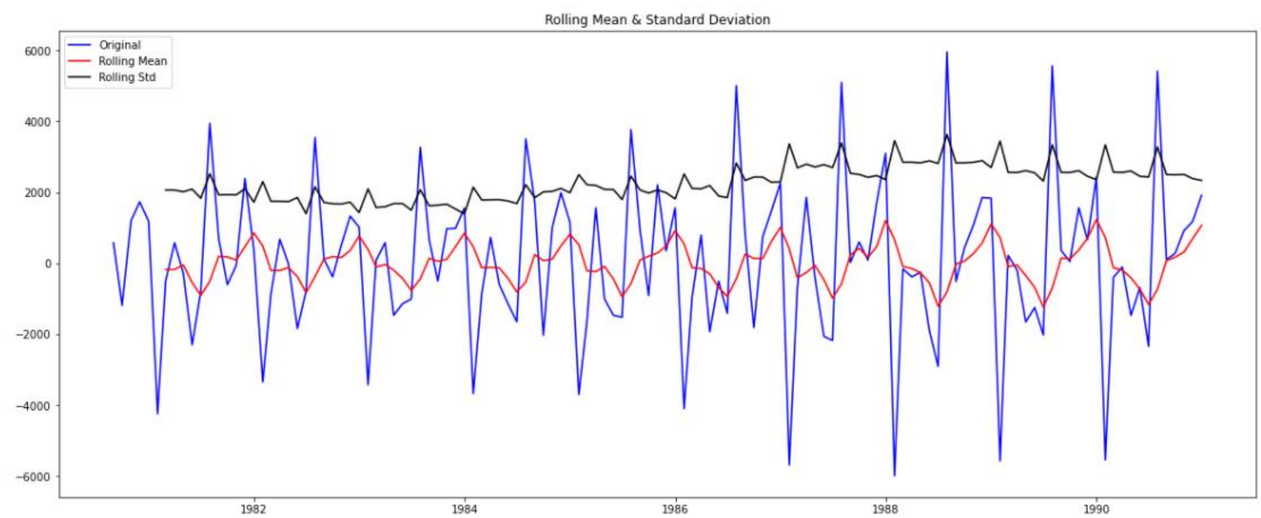
We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.



58. Graph for data set: SARIMA at 6 1st order difference on seasonally differenced series

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

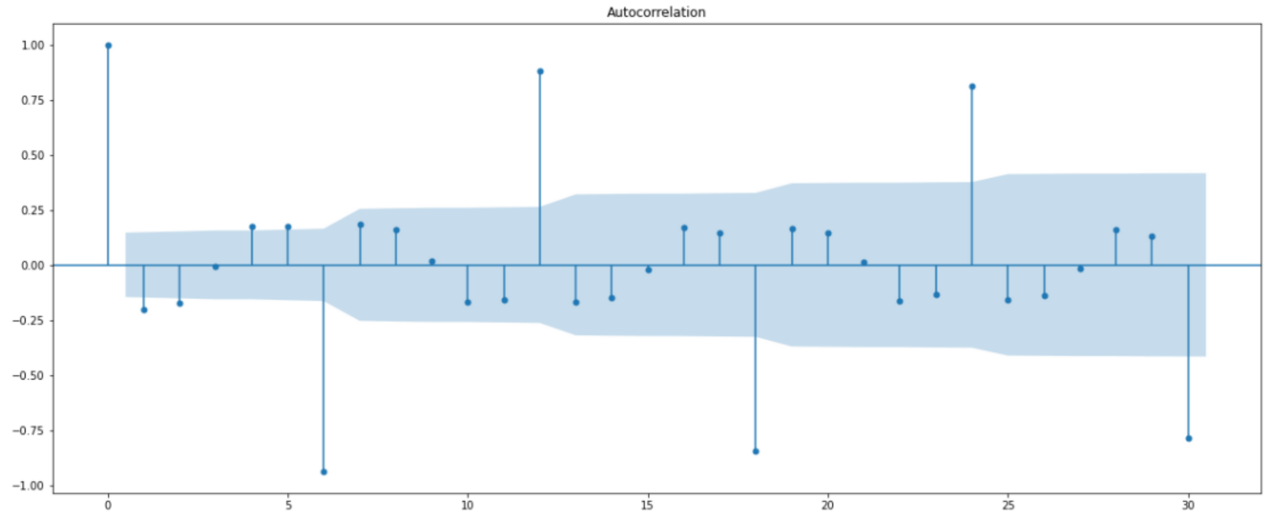
Let us go ahead and check the stationarity of the above series before fitting the SARIMA model

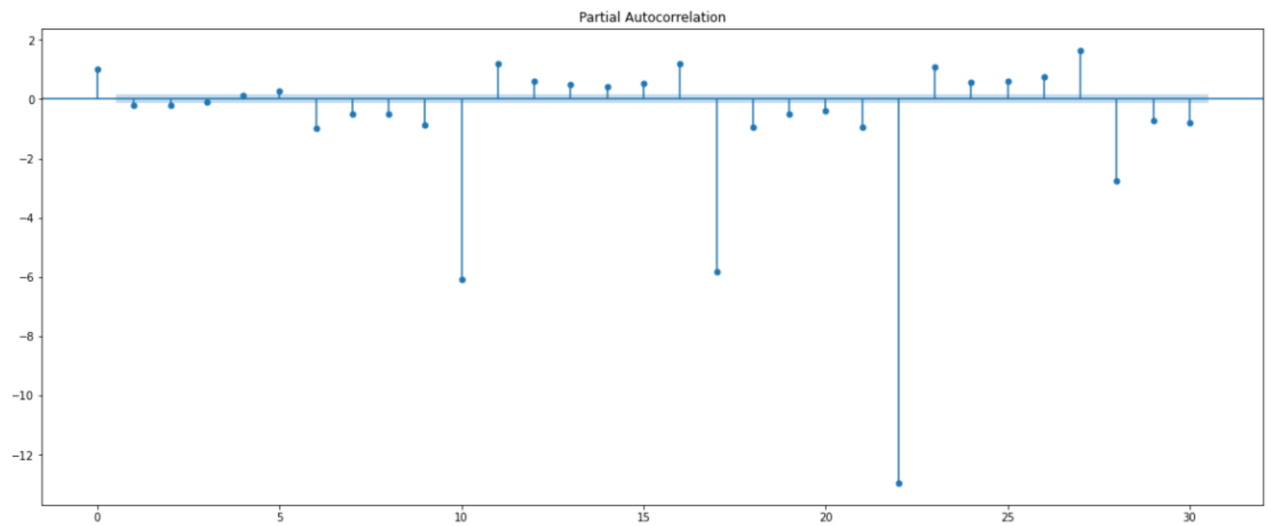



```
Results of Dickey-Fuller Test:
Test Statistic                 -7.017242e+00
p-value                         6.683657e-10
#Lags Used                      1.300000e+01
Number of Observations Used     1.110000e+02
Critical Value (1%)            -3.490683e+00
Critical Value (5%)            -2.887952e+00
Critical Value (10%)           -2.580857e+00
dtype: float64
```

59. Stationarity of dataset for building SARIMA model

We can see that the data is stationary. We will now plot the new ACF & PACF plots for the new modified time series.





60. ACF & PACF plots for the newly modifies 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.

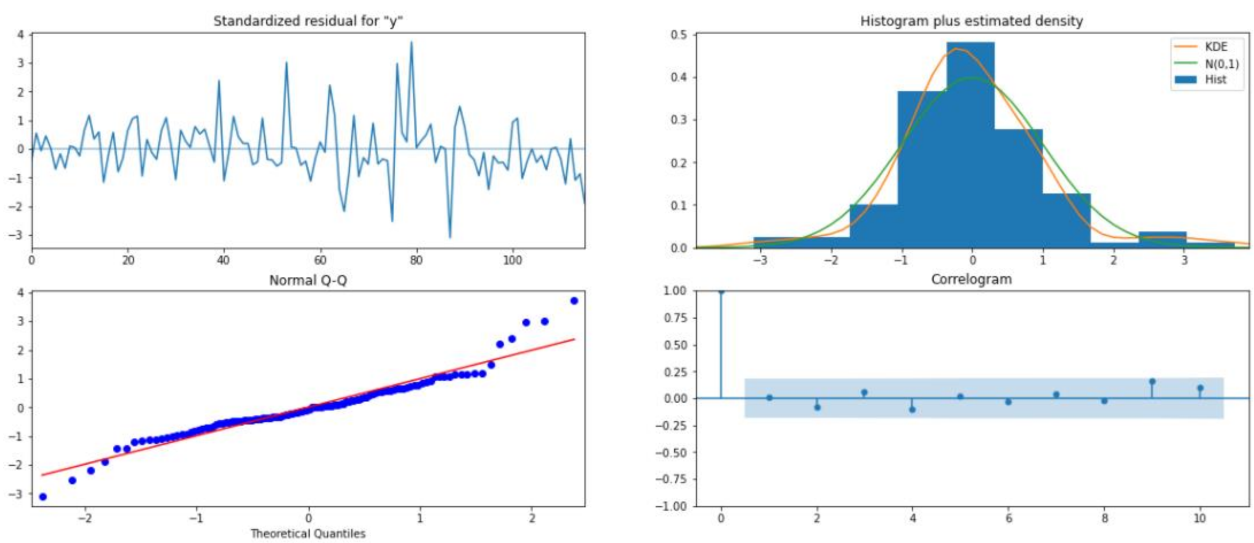
The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF 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, 2)x(2, 0, 2, 6)   Log Likelihood                -856.967
Date:                           Sun, 20 Mar 2022   AIC                           1729.934
Time:                                   18:22:37   BIC                           1751.963
Sample:                                        0   HQIC                          1738.877
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0913      1.069      0.085      0.932      -2.005       2.187
ma.L1       -210.0726    4.6e+04     -0.005      0.996   -9.03e+04    8.99e+04
ma.L2        234.6570   5.16e+04      0.005      0.996   -1.01e+05    1.01e+05
ar.S.L6        0.0029      0.025      0.118      0.906      -0.046       0.052
ar.S.L12       1.0403      0.017     59.901      0.000       1.006       1.074
ma.S.L6        0.4641      0.145      3.210      0.001       0.181       0.747
ma.S.L12      -0.9520      0.171     -5.571      0.000      -1.287      -0.617
sigma2         1.7587    773.721      0.002      0.998   -1514.707    1518.225
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                38.16
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               2.88   Skew:                             0.57
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.57
===================================================================================

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

61. Manual SARIMA Result



62. Plot Diagnostic for SARIMA at 6

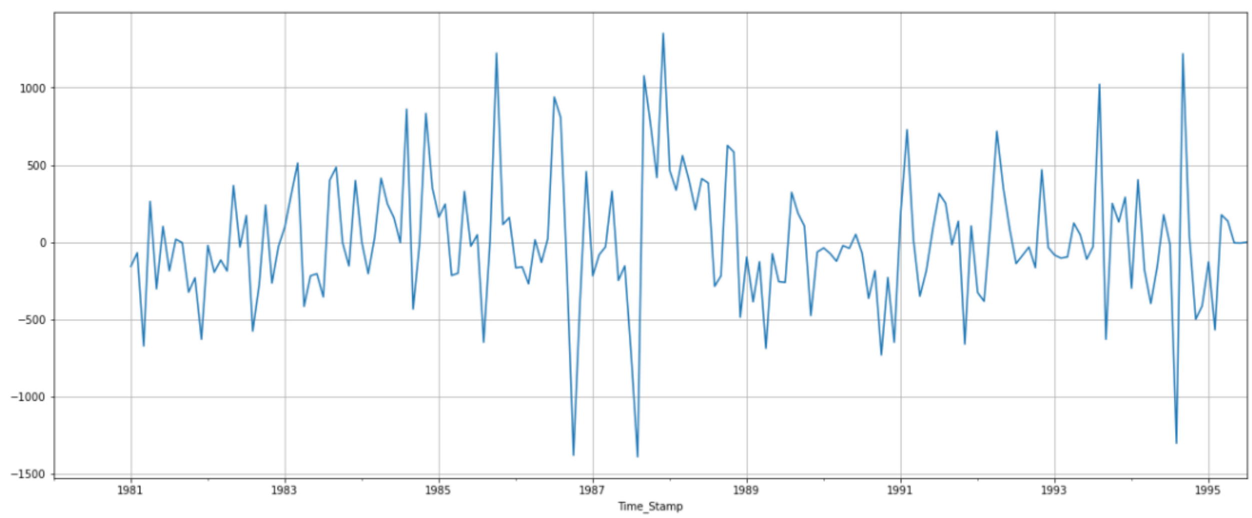
RMSE:

```
For MANUAL SARIMA on the Test Data,  RMSE is 607.940
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048
SARIMA(0,1,2)(2,0,2,6)	601.353424
SARIMA(1,1,2)(1,0,2,12)	528.588034
MANUAL ARIMA(0,1,2)	1417.502239
MANUAL SARIMA(1,1,2)(2,0,2,6)	607.940086

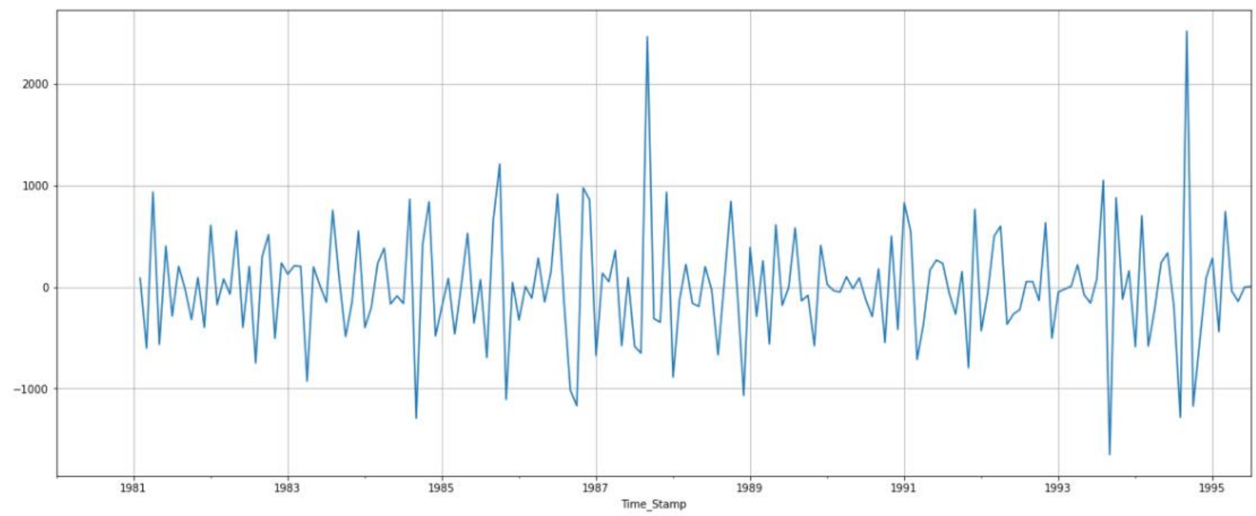
63. RMSE Values for Manual SARIMA at 6 & all the other models built

SARIMA AT 12:



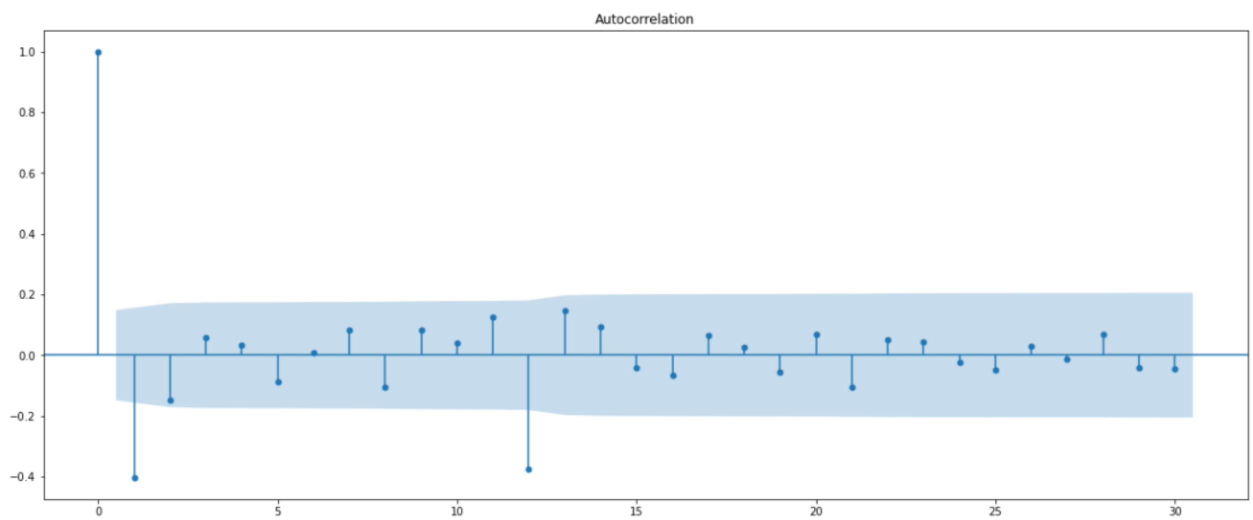
64. Graph for SARIMA at 12

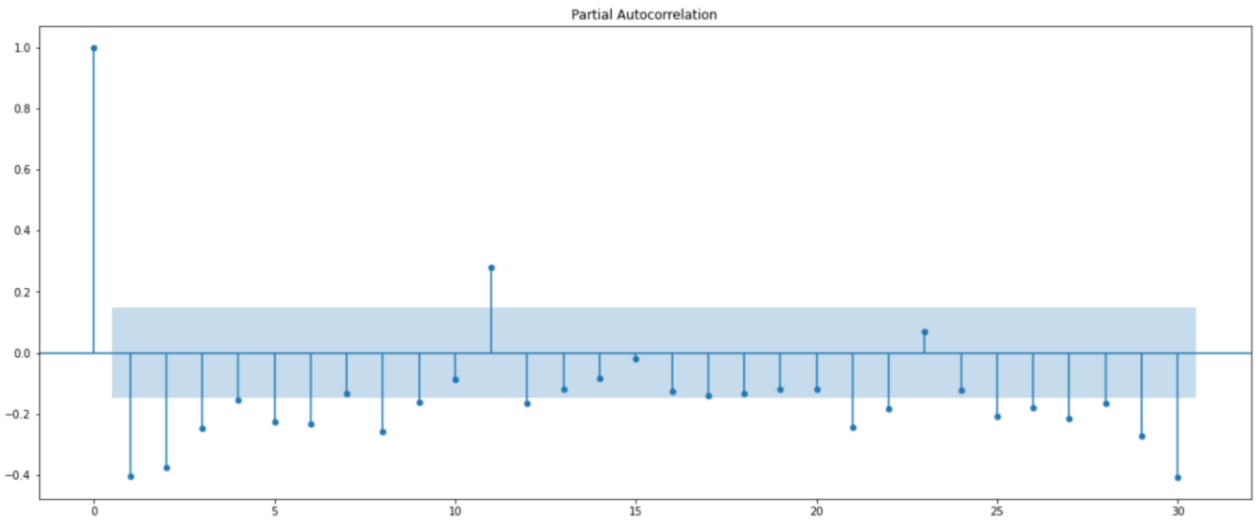
We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.



65. Graph for data set: SARIMA at 12 1st order difference on seasonally differenced series

Now we see that there is almost no trend present in the data. Seasonality is only present in the data. Lets check the ACF & PACF plot for this new data.



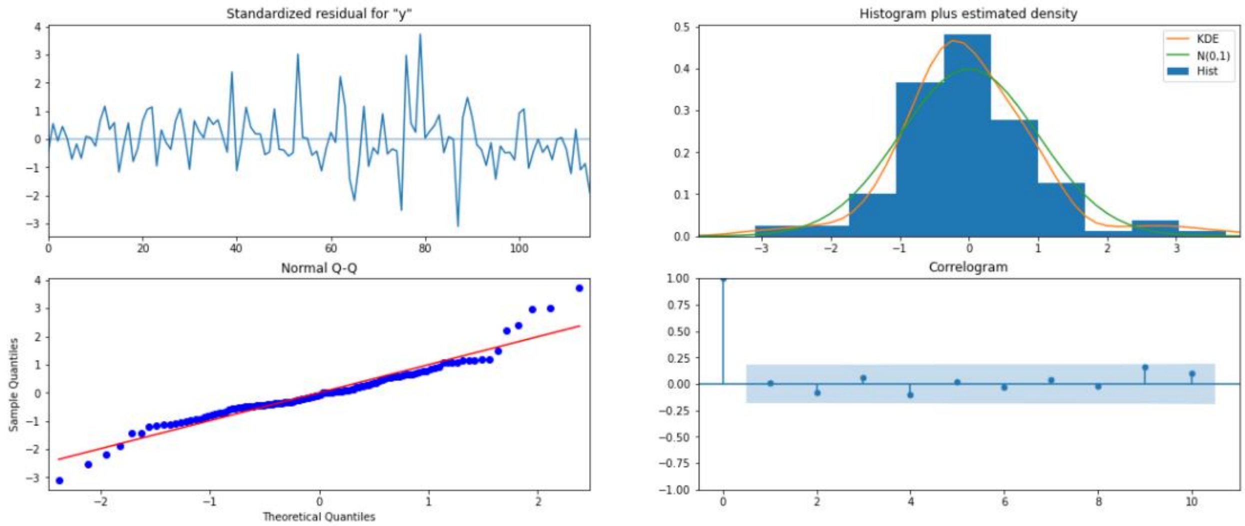


66. ACF & PACF Plots for newly modified Time Series: SARIMA at 12

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -856.967
Date:                           Sun, 20 Mar 2022   AIC                           1729.934
Time:                                   18:22:41   BIC                           1751.963
Sample:                                        0   HQIC                          1738.877
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0913      1.069      0.085      0.932      -2.005       2.187
ma.L1       -210.0726    4.6e+04     -0.005      0.996   -9.03e+04    8.99e+04
ma.L2        234.6570   5.16e+04      0.005      0.996   -1.01e+05    1.01e+05
ar.S.L6        0.0029      0.025      0.118      0.906      -0.046       0.052
ar.S.L12       1.0403      0.017     59.901      0.000       1.006       1.074
ma.S.L6        0.4641      0.145      3.210      0.001       0.181       0.747
ma.S.L12      -0.9520      0.171     -5.571      0.000      -1.287      -0.617
sigma2         1.7587    773.721      0.002      0.998   -1514.707    1518.225
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                38.16
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               2.88   Skew:                             0.57
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.57
===================================================================================

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

67. SARIMA at 12 Result



68. Plot Diagnostics: SARIMA at 12

RMSE:

```
For MANUAL SARIMA on the Test Data,  RMSE is 607.940
```

	Test RMSE
Alpha= 0.07, SES	1338.012144
Alpha= 0.663, Beta= 9.966, DES	3949.993290
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
Regression	1389.135175
Naive Model	3864.279352
Simple Average Model	1275.081804
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048
SARIMA(0,1,2)(2,0,2,6)	601.353424
SARIMA(1,1,2)(1,0,2,12)	528.588034
MANUAL ARIMA(0,1,2)	1417.502239
MANUAL SARIMA(1,1,2)(2,0,2,6)	607.940086
MANUAL SARIMA(1,1,2)(2,0,2,12)	607.940086

69. RMSE Values for Manual SARIMA at 12 & all the other models built

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

	Test RMSE
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	379.695686
SARIMA(1,1,2)(1,0,2,12)	528.588034
SARIMA(0,1,2)(2,0,2,6)	601.353424
MANUAL SARIMA(1,1,2)(2,0,2,6)	607.940086
MANUAL SARIMA(1,1,2)(2,0,2,12)	607.940086
2point Trailing Moving Average	813.400684
4 point Trailing Moving Average	1156.589694
Simple Average Model	1275.081804
6point Trailing Moving Average	1283.927428
Alpha= 0.07, SES	1338.012144
9 point Trailing Moving Average	1346.278315
ARIMA(2,1,2)	1374.958048
Regression	1389.135175
MANUAL ARIMA(0,1,2)	1417.502239
Naive Model	3864.279352
Alpha= 0.663, Beta= 9.966, DES	3949.993290

70. Data Frame: Models Built along with their RMSE

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.

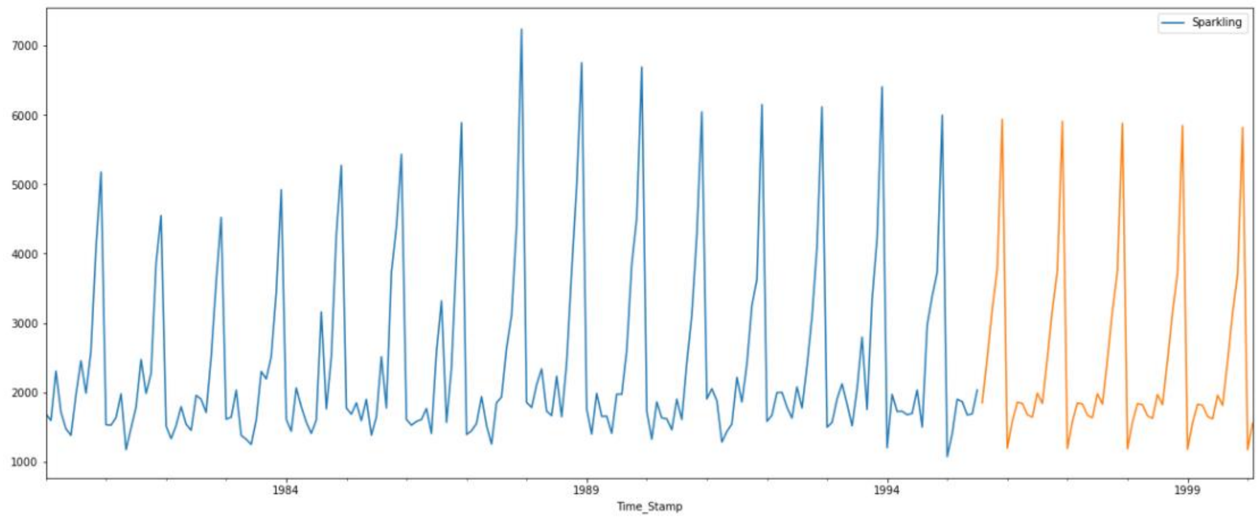
Building the most optimum model on the Full Data

We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters α = 0.25, β = 0.0 and γ = 0.74.

The same is put into a dataframe for building the most optimum model on the complete data.

The RMSE score for the full model is 363.446

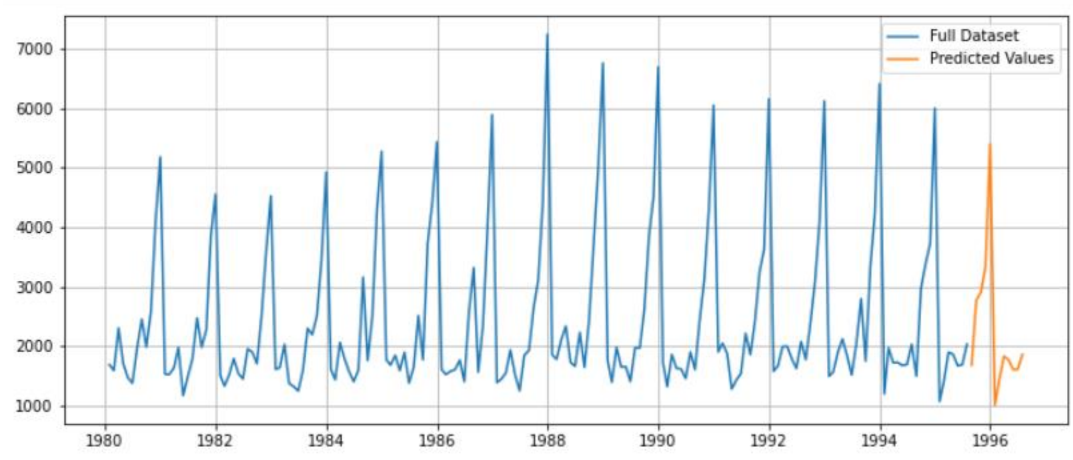
Then the prediction on full model is done. The graph for the same is below:



71. Prediction for full dataset

```
1995-08-31    1679.266722
1995-09-30    2768.527065
1995-10-31    2905.649091
1995-11-30    3327.644442
1995-12-31    5399.749252
1996-01-31    1010.934121
1996-02-29    1433.117417
1996-03-31    1830.576460
1996-04-30    1772.339841
1996-05-31    1604.364674
1996-06-30    1610.179234
1996-07-31    1859.791410
Freq: M, dtype: float64
```

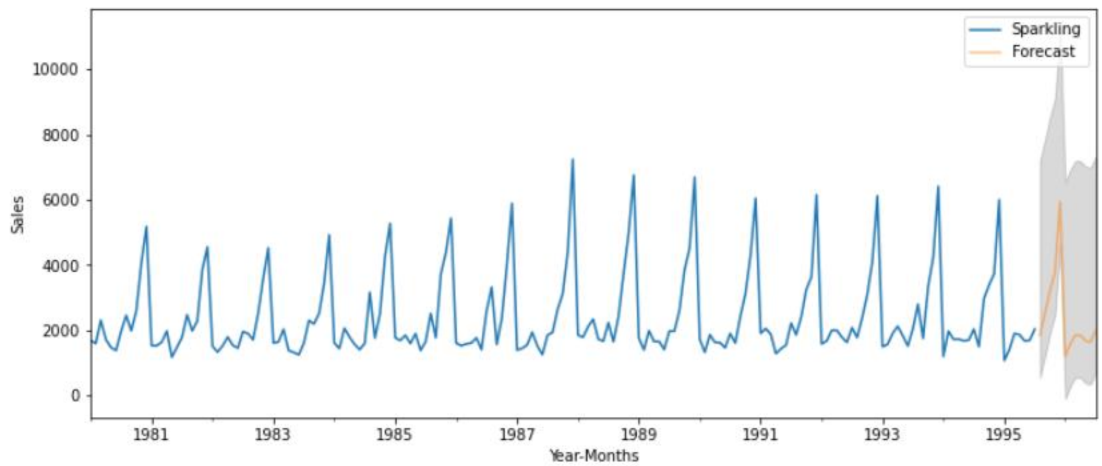
72. List of Prediction for next 12 months



73. Predicted values for Sparkling Wine Sale

	lower_CI	prediction	upper_ci
1995-08-31	552.232194	1679.266722	6509.187152
1995-09-30	1641.492537	2768.527065	7598.447494
1995-10-31	1778.614563	2905.649091	7735.569520
1995-11-30	2200.609915	3327.644442	8157.564872
1995-12-31	4272.714725	5399.749252	10229.669682

74. Class Intervals



75. Graph for class intervals and next 12 months prediction

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

- On visualization the above data, it can be noted that Sparkling has the highest sale around the year 1987, and in the month of December.
- Following December the month of November, October & September have also had number of Sparkling wine sale over the years to be more than other months.
- The month of January, February, June and May have the lowest number of Sparkling Wine sales across the years.
- The highest sale of Sparkling wine is in the month of December with its mean being around 5800 with respect to sales, and the lowest is in the month of June at which the mean sale is around 600.
- The decomposition of the data was first done using additive decomposition, but it showed residuals to have pattern around 0, then on using multiplicative method the data did not show any pattern for the residuals and the components were located around 1.
- Several models were built: Simple Exponential, Double Exponential, Triple Exponential Smoothing, Simple Average, moving average, linear regression.
- The models were built on training data and performance was checked on test data using RMSE.
- The lowest RMSE value is considered to be optimal for creating the best model for overall predictions.
- Based on the above models built, Triple Exponential Smoothing (with parameters, alpha=0.25, beta=0.0, gamma=0.74) had the lowest AIC value at 379.

- With the use of TES as the best fit model, future sales predictions for next 12 months are done. Plotting and predicting the same along with confidence intervals.
- Using the predictions from the models, the company now has an insight on the profit, loss, production quantity, materials needed for the production of Sparkling Wine.
- With these insights the company can forecast which month will have higher or lower production with respect to the demand and supply for the particular month.
- To make their business more profitable, the company can use various strategies, such as advertisements in Radio, Television, Local Newspapers, pamphlets etc to grow their consumer base, and boost sale, and production for the time period where the sale is low.
- Various discounts and offers can be given both to the sale team as well as people who purchase the products to boost sales.
- In future, the company can have low production for months with lowest sale to avoid wastage which in turn helps with profit.
- Special offers at certain events can be discussed to uplift the production at low seasons.

PROBLEM 2

ROSE:

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

BASIC INFORMATION OF THE DATASET

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

76. Head of data Rose.csv

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

77. Tail of data Rose.csv

	YearMonth	Rose	Time_Stamp
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

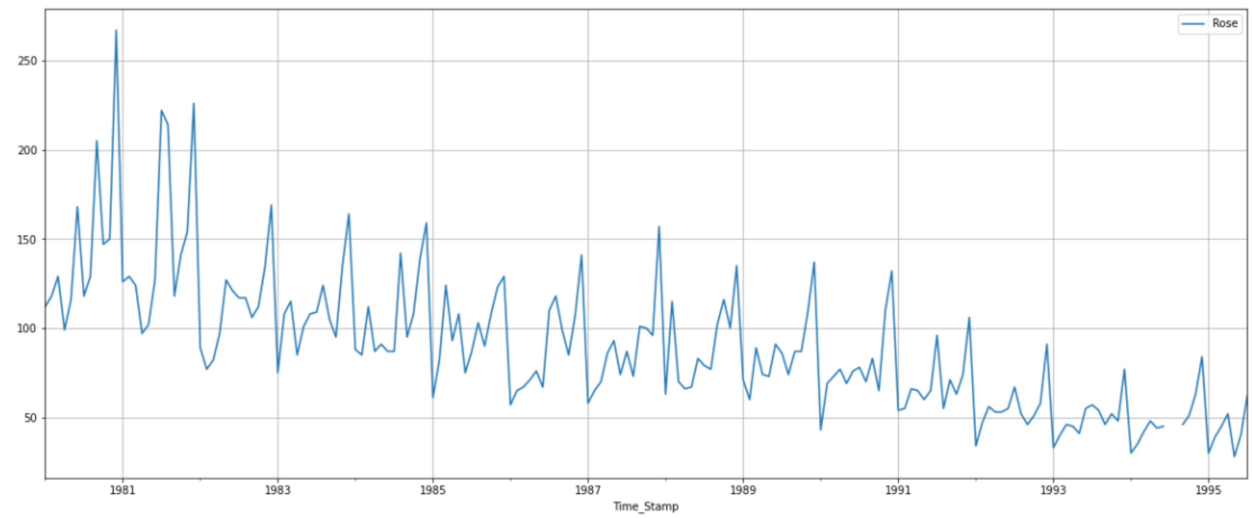
78. Creating Time-Stamp in the dataset

	Rose
Time_Stamp	
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

79. Setting Time-Stamp at the Index in the dataset, after dropping 'YearMonth' Column

```
The number of rows:  187
The number of columns:  1
```

80. Shape of dataset



81. Plotting graph of the dataset

4. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

82. Description of dataset

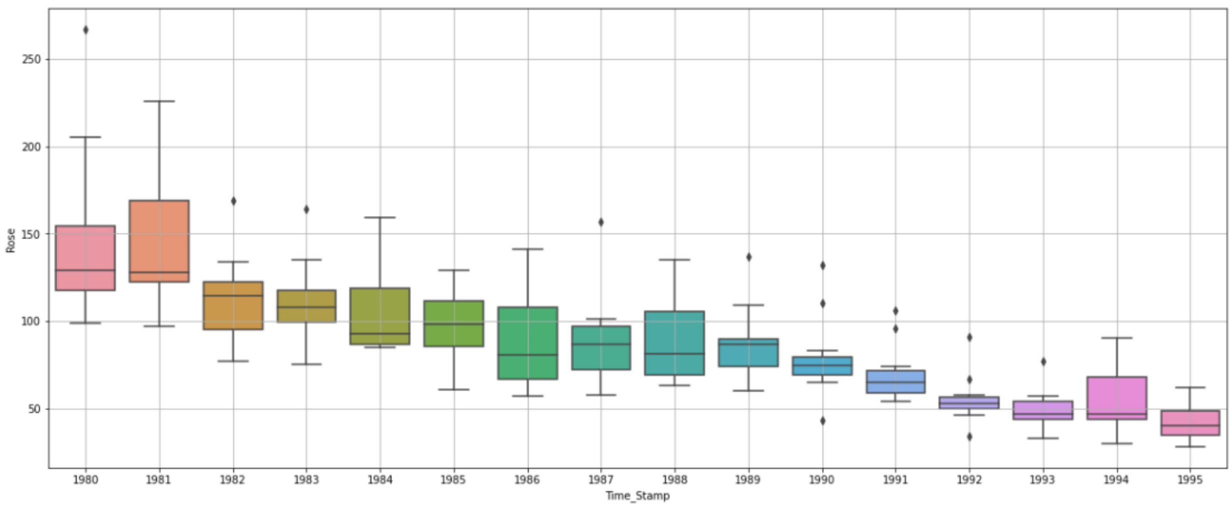
```
Rose     2
dtype: int64
```

83. Null-Values in dataset

```
Rose     0
dtype: int64
```

84. Null values in dataset after imputation

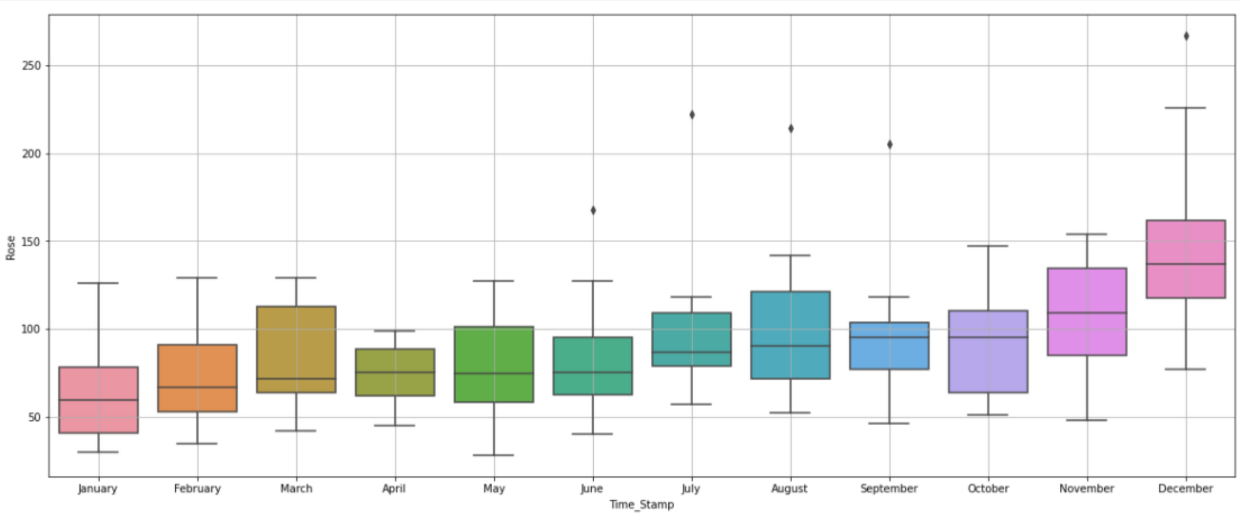
YEARLY BOXPLOT:



85. Yearly Boxplot showcasing the sale of Rose Wine

The yearly boxplots show us how the sale has increased and decreased over the past few years. The highest number of sales being recorded in the year 1981.

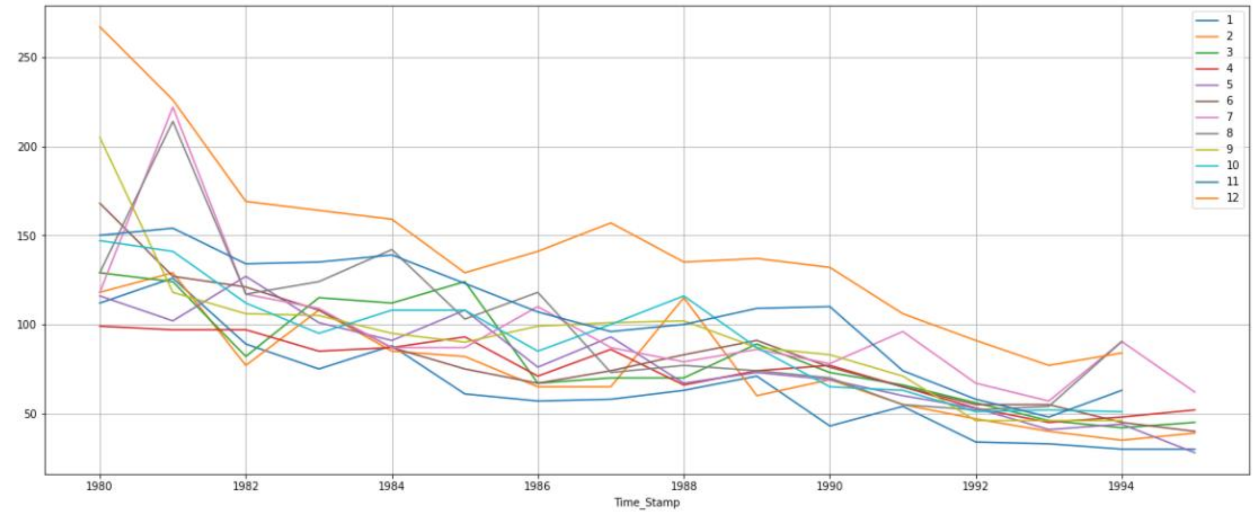
MONTHLY BOXPLOT:



86. Monthly Boxplot showcasing the sale of Rose Wine

From the above monthly boxplot, it can be seen that the highest sale of Rose wine happened in the month of December followed by November, October & August.

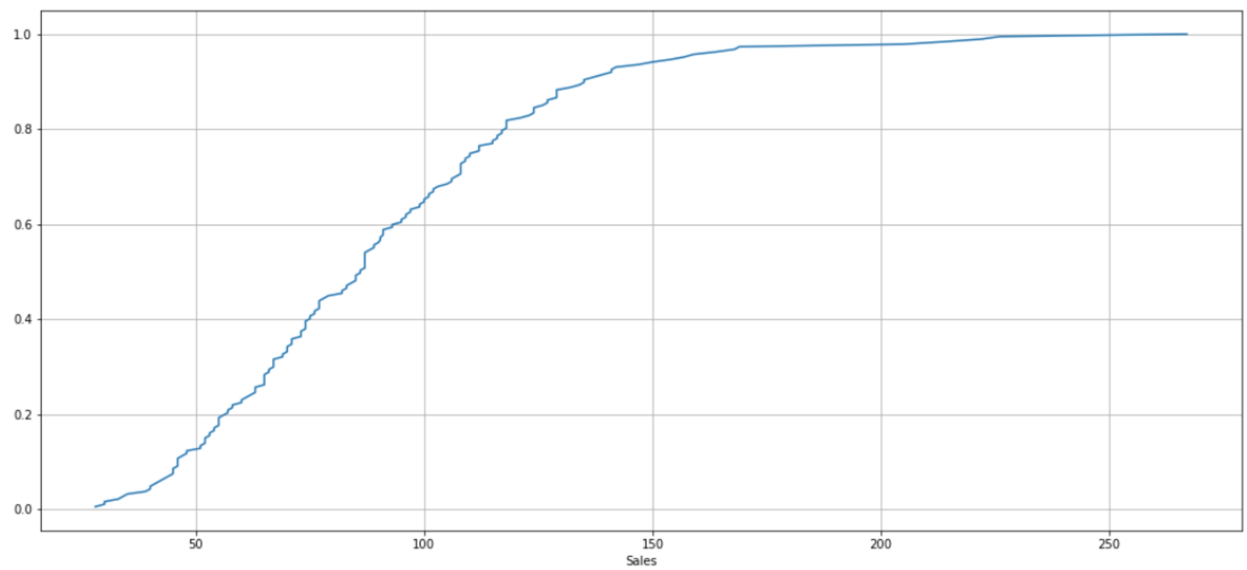
MONTHLY SALE ACROSS THE YEAR:



87. Graph showcasing the sale of Rose Wine across the year for every month

From the above graph it can be seen that the highest sale for Rose Wine was in the year of 1980 and it has been declining ever since.

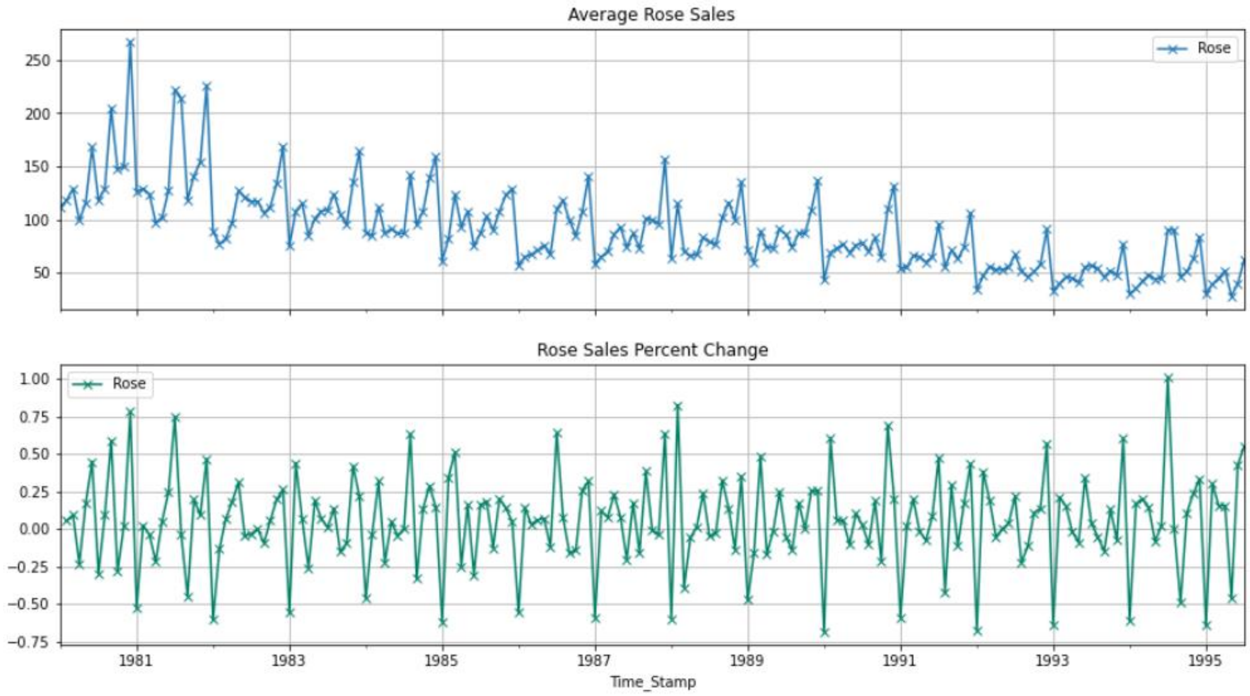
Empirical Cumulative Distribution:



88. Empirical Cumulative Distribution Graph

This particular graph tells us what percentage of data points refer to what number of Sales.

Next, we will be plotting the average Rose Sales per month and the month-on-month percentage change of Rose Sales.



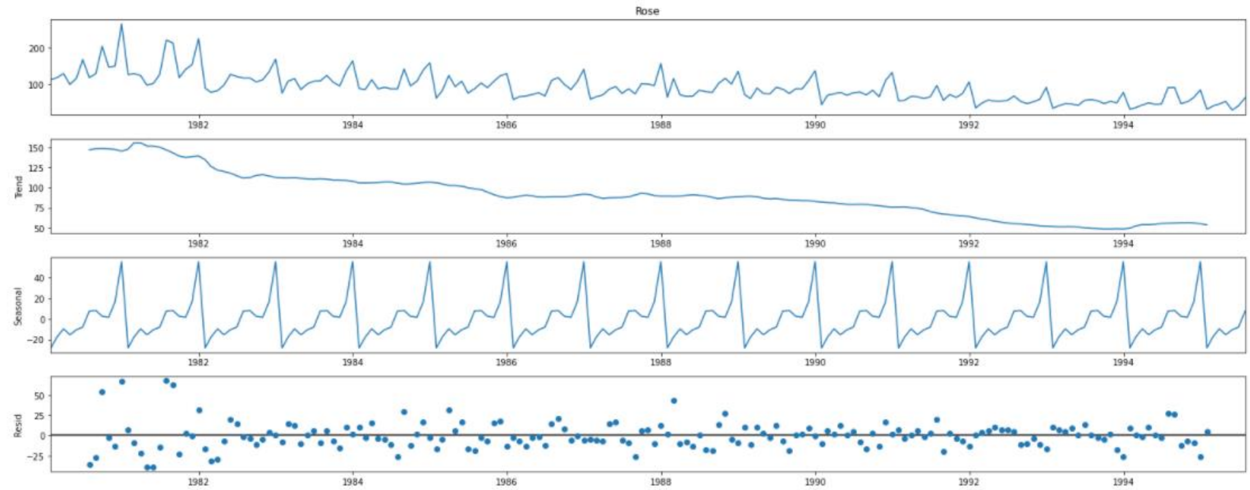
89. Graph showcasing Average Rose sales, and its percentage change over the years

From the above Exploratory data analysis it is found that:

- The data starts at 1980 and ends at 1995.
- There are 187 rows and 2 columns.
- There two null value present in the dataset which were then imputed using mean.
- We then create a time stamp and then replace it at index, while dropping the YearMonth column.
- The graph for average Rose sales shows that the sale for Rose has only been declining over the past few years.
- The highest sale was recorded in the year 1980, while the month in which the sale is highest is December followed by November & October
- The lowest sale is recorded in the year 1995.

DECOMPOSITION:

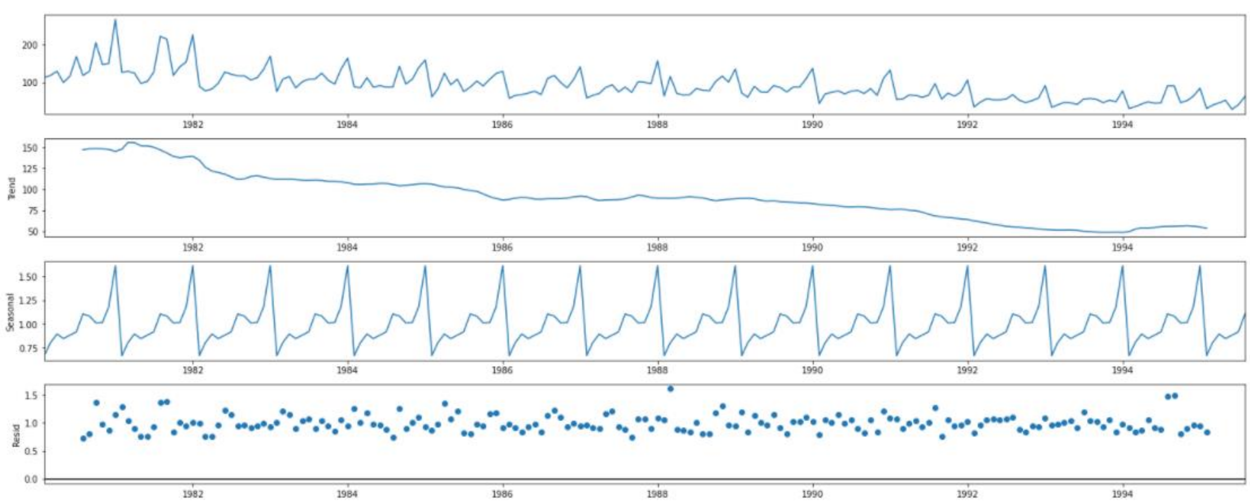
ADDITIVE DECOMPOSITION:



90. Additive Decomposition

On performing Additive Decomposition on the dataset, it is observed that there are residuals located near 0, which gives pattern that is non-essential for model building. We then go forward with Multiplicative Decomposition.

MULTIPLICATIVE DECOMPOSITION:



91. Multiplicative Decomposition

On performing Multiplicative Decomposition on the dataset, it is observed that the residuals are now located near 1. Multiplicative Decomposition will be used for building our models

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

The data will be split into train and test and plotting the training and test data will be done. Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

First few rows of Training Data

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

First few rows of Test Data

Time_Stamp	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Last few rows of Training Data

Time_Stamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

Last few rows of Test Data

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

92. First and last five rows of Training & Test data

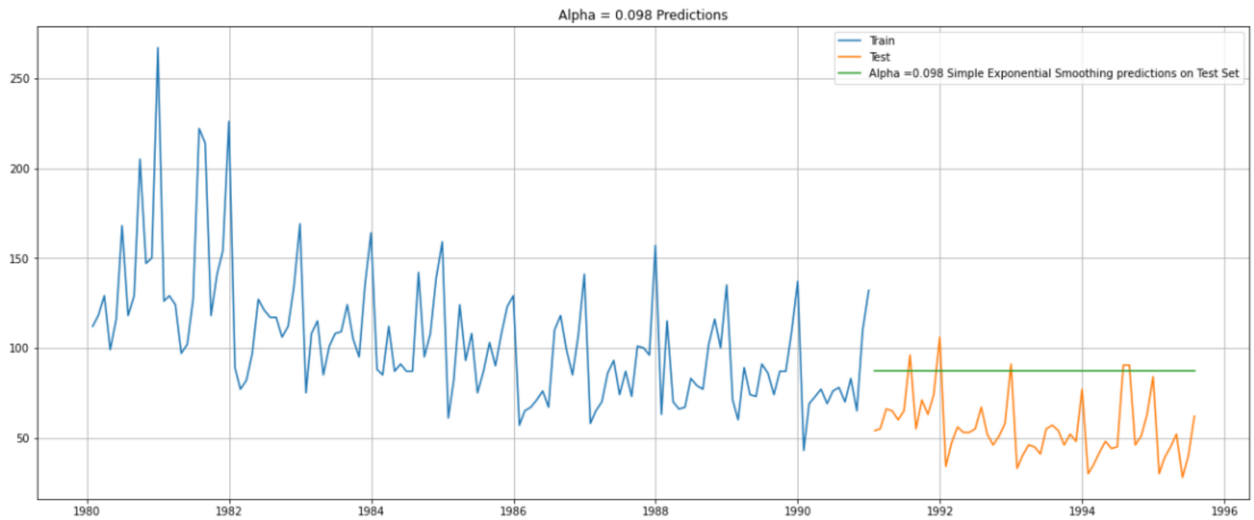
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

a. <u>SIMPLE EXPONENTIAL SMOOTHING:</u>

```
{'smoothing_level': 0.09874933517484011,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38703609891138,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

93. Optimal Parameters for Single Exponential Smoothing

It is found that the optimized smoothing level is at 0.0987



94. Graph: Forecasted Values for Simple Exponential Smoothing

The above graph showcases the forecasted values for Train and Test dataset using Simple Exponential Smoothing

We then proceed to calculate the Root Mean Square Error for the model being built using Simple Exponential Smoothing.

RSME:

```
Simple Exponential Smoothing RMSE: 35.936198483436414
```

	Test RMSE
Alpha= 0.07, SES	35.936198

95. RMSE Values

b. <u>DOUBLE EXPONENTIAL SMOOTHING</u>

The optimal parameters for Double Exponential Smoothing are as follows:

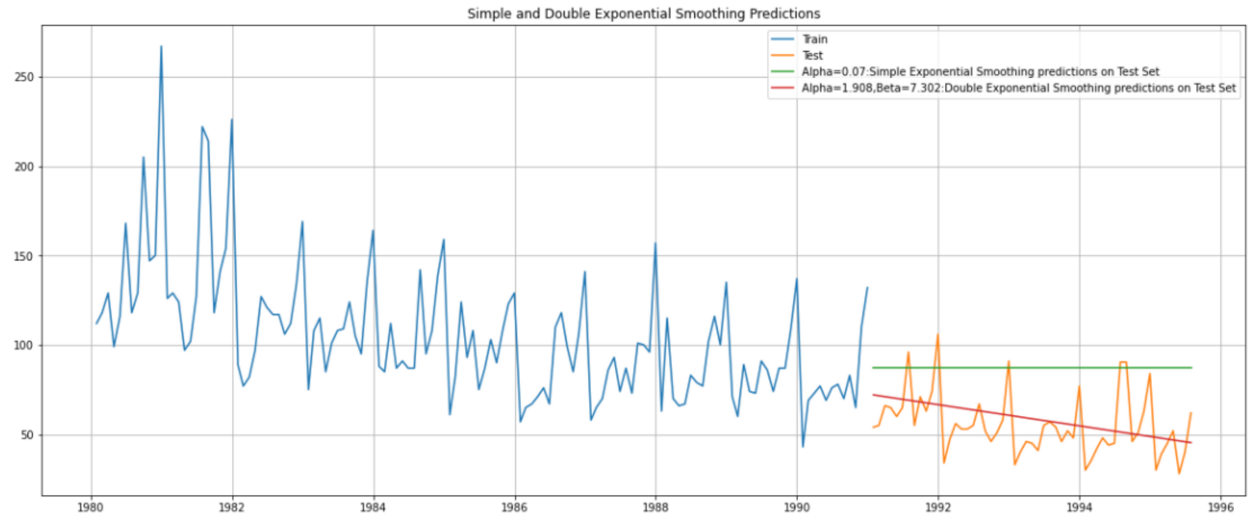
```
==Holt model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 1.9086427682180844e-08, 'smoothing_trend': 7.302464353829351e-09, 'smoothing_seasonal': nan, 'damping_tren
d': nan, 'initial_level': 137.81629861505857, 'initial_trend': -0.4943753249082896, 'initial_seasons': array([], dtype=float6
4), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

96. Optimal Parameters for Double Exponential Smoothing

It is found that the optimal smoothing level for Alpha=1.908, Beta= 7.302.

The above data is used for forecasting on the test set.



97. Graph: Forecasted Values for Double Exponential Smoothing

RMSE:

```
Double Exponential Smoothing RMSE: 16.979631467541886
```

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631

98. Test RMSE Value: SES & DES

c. TRIPLE EXPONENTIAL SMOOTHING

The optimal parameters for Triple Exponential Smoothing are as follows:

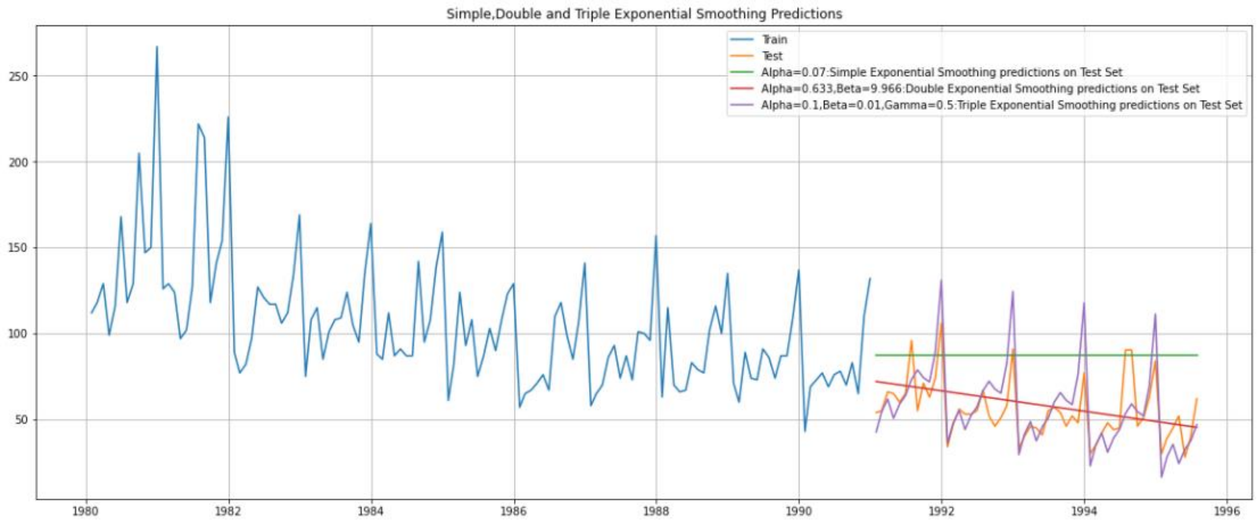
```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.08830330642635406, 'smoothing_trend': 6.730635331927582e-05, 'smoothing_seasonal': 0.004455138229351625,
'damping_trend': nan, 'initial_level': 146.88752868155674, 'initial_trend': -0.5492163940406024, 'initial_seasons': array([-31.
12207537, -18.81171138, -10.86052241, -21.52235816,
       -12.68359535,  -7.17529564,   2.7456236 ,   8.84900094,
         4.85724354,   2.9520333 ,  21.05004912,  63.29916317]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

99. Optimal Parameters for Triple Exponential Smoothing

It is found that python, has optimized the smoothing level to be at Alpha= 0.088, Beta= 6.730, Gamma= 0.004

The above data is used for forecasting on the test set.



100. Graph: Forecasted Values for Single, Double & Triple Exponential Smoothing

The above graph shows the forecast values using train and test sets, it showcases Single, Double & Triple Exponential Smoothing on the data. It can be noticed that the Triple Exponential Smoothing has been able to pick up the trend, seasonality along with the level of the data.

RMSE:

```
Triple Exponential Smoothing RMSE: 15.548489757559484
```

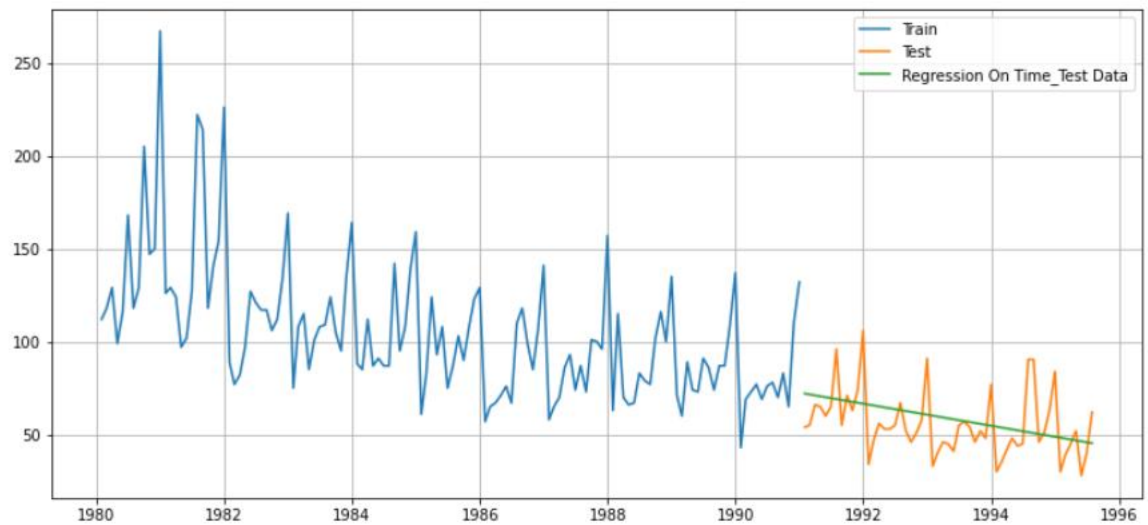
	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490

101. RMSE Values for Single, Double & Triple Exponential Smoothing

TES model shows very low RMSE value and thus is a good model that can be used for forecasting.

REGRESSION:

For a regression model to be built we take the train and test data and fit it using Linear Regression.



102. Graph showcasing the regression model forecasting

RMSE:

```
For RegressionOnTime forecast on the Test Data,  RMSE is 16.979
```

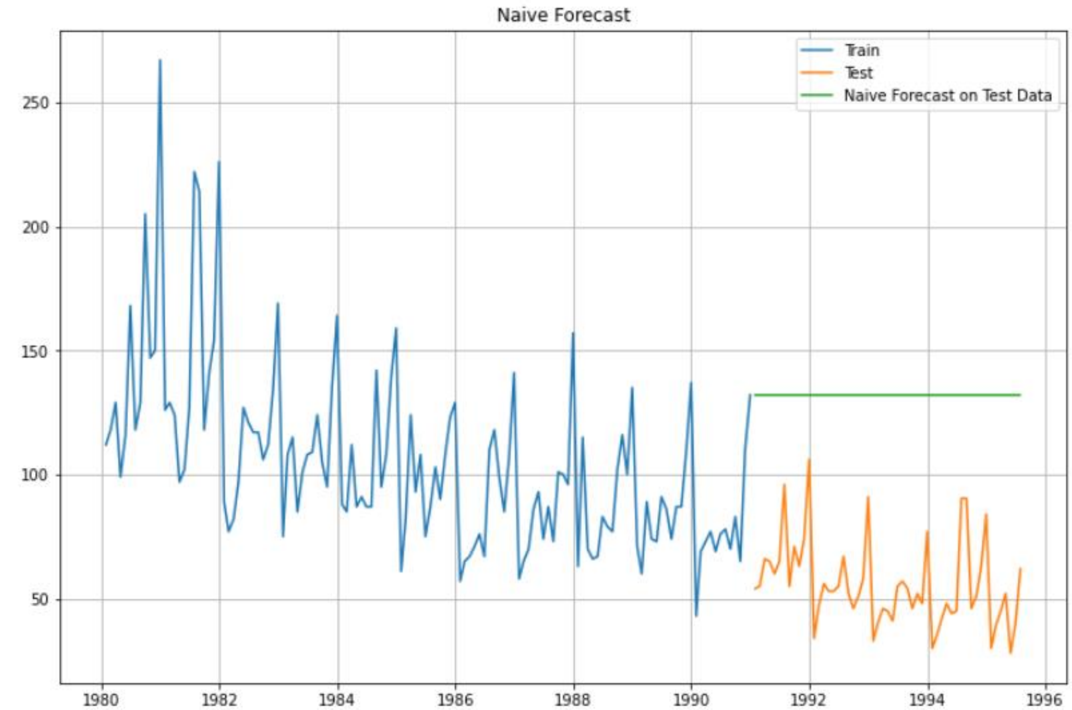
	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414

103. RMSE Values: REGRESSION along with other models built above

The RMSE value for Regression is high as compared to Triple Exponential model built previously.

The RMSE for Regression is very low as compared to Single & Double Exponential Smoothing, overall Regression is not the best fit model that can be used for forecasting.

NAÏVE APPROACH:



104. Graph: NAÏVE APPROACH forecasting

RMSE:

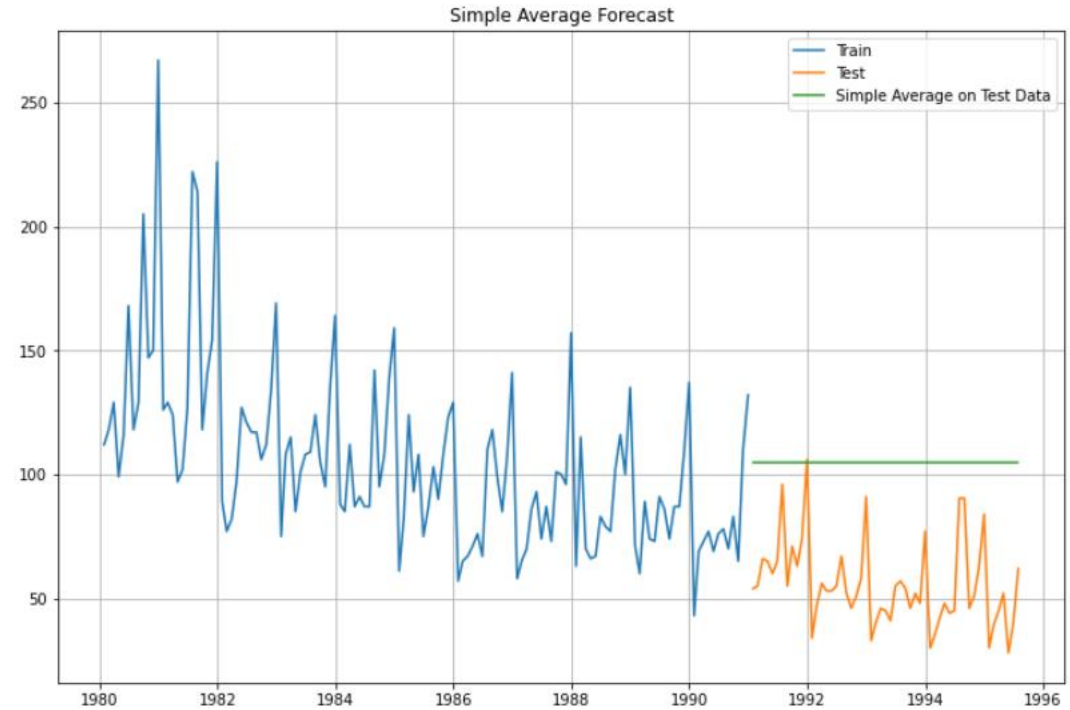
```
For RegressionOnTime forecast on the Test Data,  RMSE is 78.396
```

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083

105. *RMSE Value: Naïve Approach along with previously built models*

The RMSE for Naïve Bayes is so far the highest as compared to other models built. Hence, this cannot be used for best fit model for overall predictions.

SIMPLE AVERAGE:



106. Graph: SIMPLE AVERAGE model

RMSE:

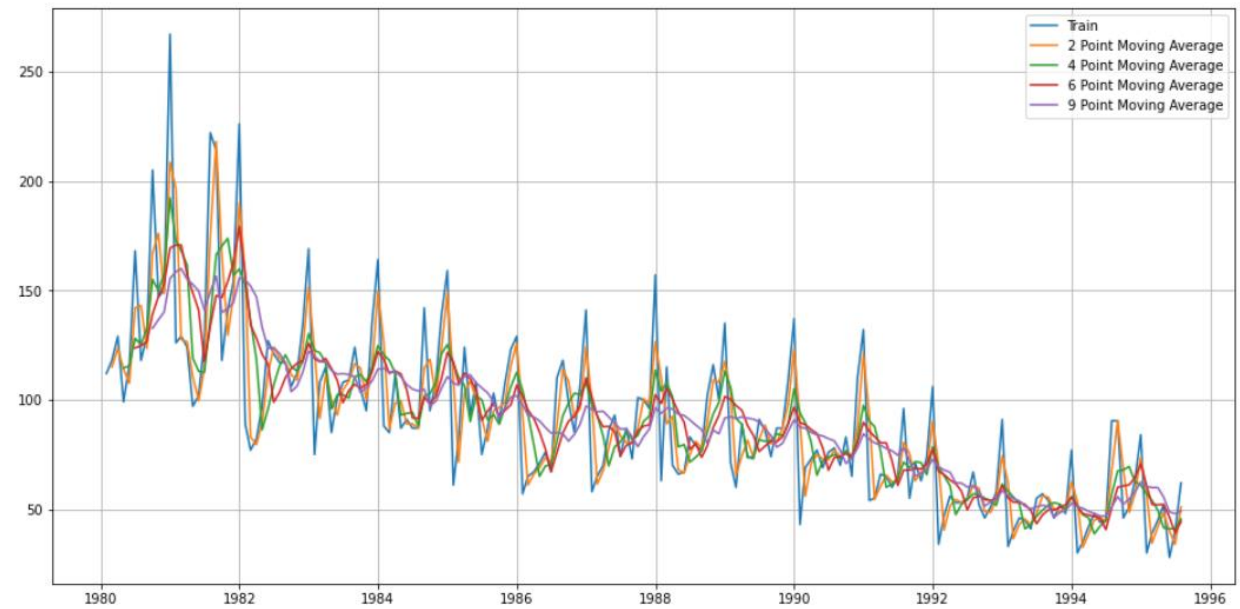
```
For Simple Average forecast on the Test Data,  RMSE is 52.319
```

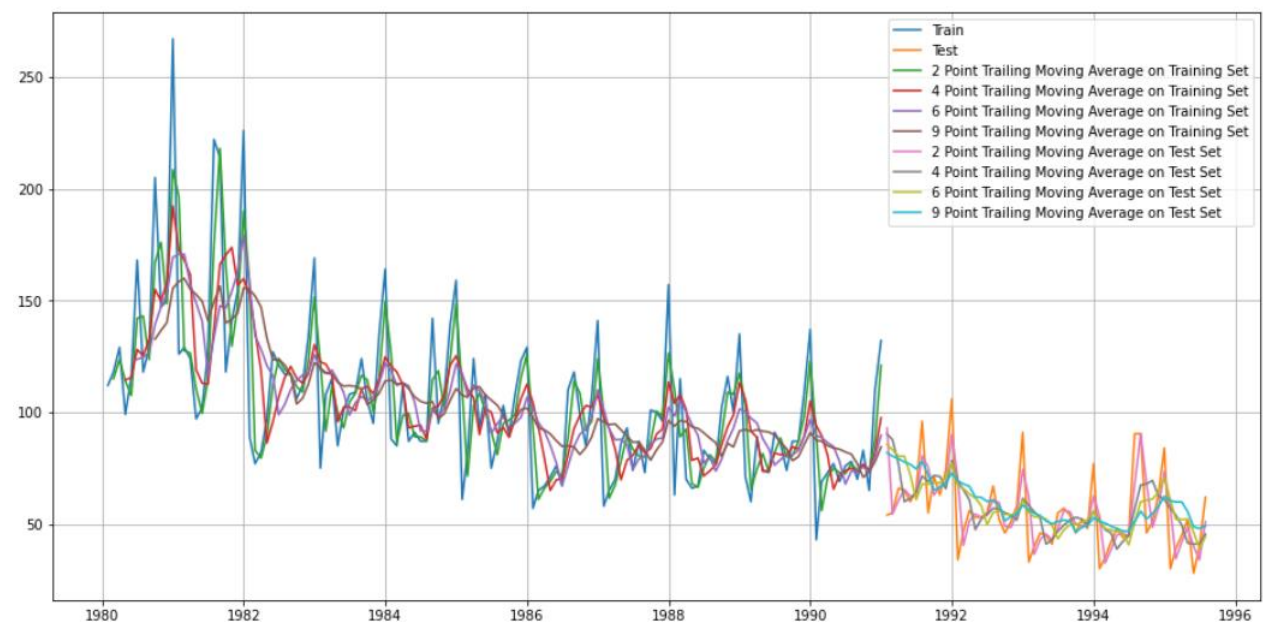
	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735

107. *RMSE Values: Simple Average Model*

The RMSE value for Simple Average model is low as compared to Naïve, but it is significantly higher than Triple Exponential Smoothing Model.

MOVING AVERAGE:





108. Graph: Moving Average Model forecasting

From the above graph it can be noted that the 2 Point Trailing moving average is performing well on test as compared to other Trailing points.

We can further calculate the RMSE for each point to compare the model.

RMSE:

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 12.298
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 15.846
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 15.986
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 16.501
```

109. RMSE Values for 2, 4, 6 & 9 point Moving Average

Compared to 4,6&9 point Moving Average RMSE, the RMSE for 2 point Moving Average has a very low value.

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823

110. RMSE Values for Moving Average model, and all previously built models

Compared to previously built models the Moving Average has the lowest RMSE value, but it is still an unstable method for forecasting.

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

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

H_0 : The Time Series has a unit root and is thus non-stationary.

H_1 : The Time Series does not have a unit root and is thus stationary. We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

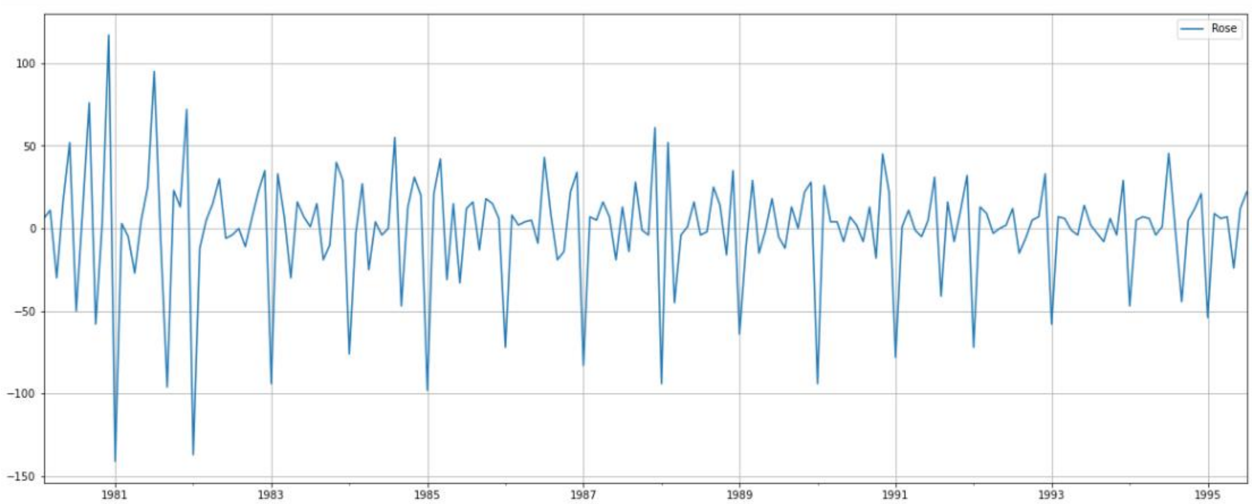
```
DF test statistic is -2.443
DF test p-value is 0.35712504484559315
Number of lags used 12
```

111. . ADF Test using stats model

```
DF test statistic is -7.988
DF test p-value is 7.599609649114865e-11
Number of lags used 12
```

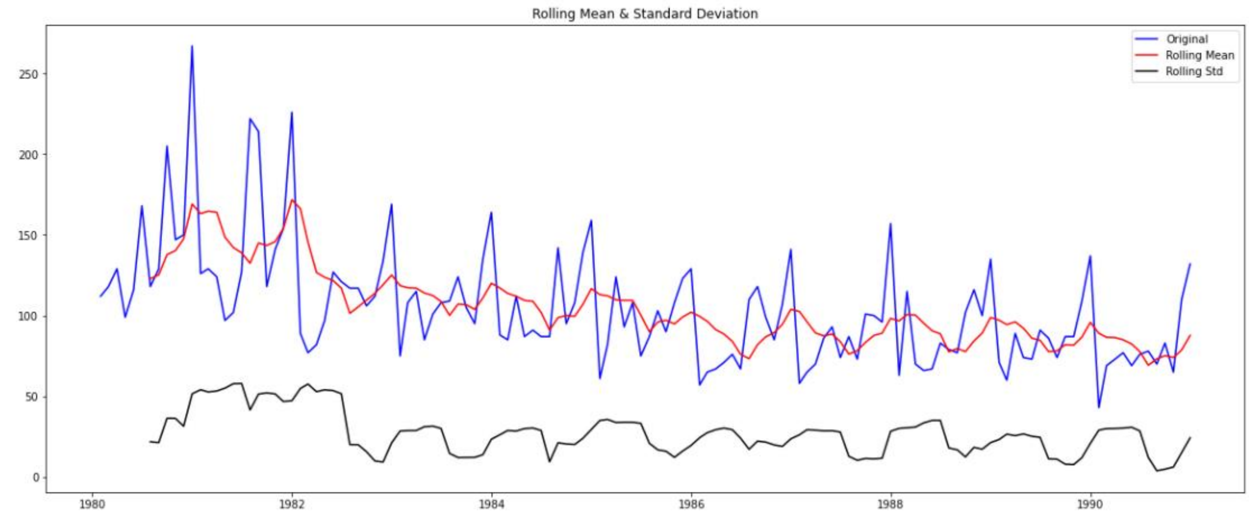
112. ADF Test with difference at one level

On using one level differencing we can see that the data is now stationary.



113. Dataset Plot to check stationarity

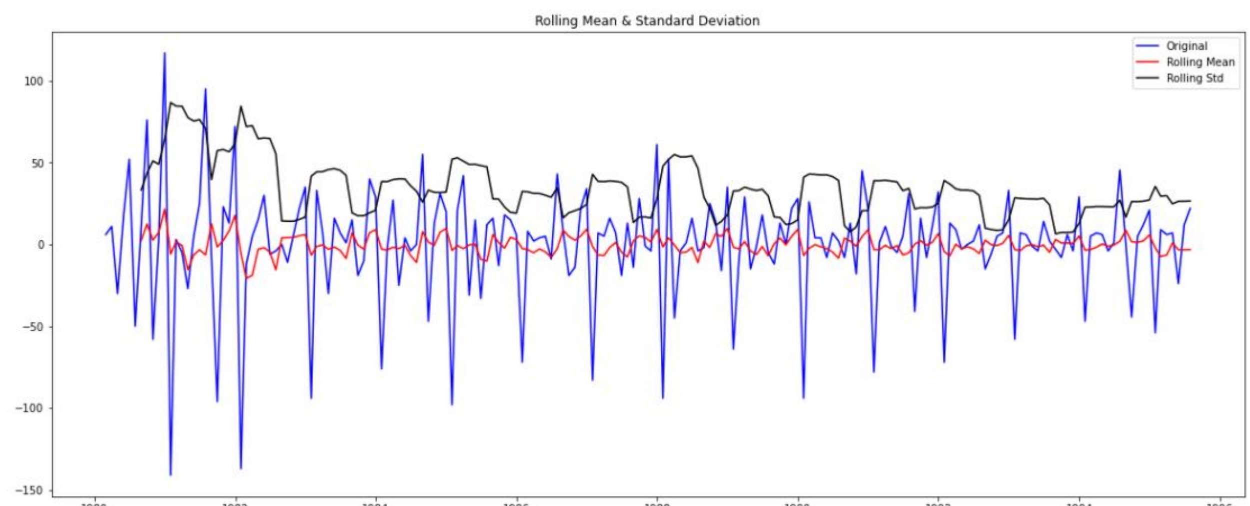
We will now check the stationarity of the Training dataset.




```
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
```

114. Training Dataset Stationarity check

We see that the Training dataset is not stationarity, we will use the difference order 1 to make the data stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                -7.855944e+00
p-value                        5.442646e-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
```

115. Training Dataset Stationarity using difference order 1

We see that at α = 0.05 the Time Series Training dataset is now indeed stationary.

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.

AUTOMATED ARIMA:

We use different parameter combinations to find the value of p,d,q to build our model. In automated ARIMA python does the permutation and combination to arrive at the best values for p,d,q.

```
ARIMA(0, 1, 0) - AIC:1335.1526583086775
ARIMA(0, 1, 1) - AIC:1280.7261830464035
ARIMA(0, 1, 2) - AIC:1276.8353726229147
ARIMA(1, 1, 0) - AIC:1319.348310580781
ARIMA(1, 1, 1) - AIC:1277.7757468404075
ARIMA(1, 1, 2) - AIC:1277.359225603172
ARIMA(2, 1, 0) - AIC:1300.6092611745498
ARIMA(2, 1, 1) - AIC:1279.0456894093106
ARIMA(2, 1, 2) - AIC:1279.2986939364973
```

116. Various ARIMA combination

	param	AIC
2	(0, 1, 2)	1276.835373
5	(1, 1, 2)	1277.359226
4	(1, 1, 1)	1277.775747
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

117. Sorting the ARIMA values based on lowest AIC

We see that the automated ARIMA model can be built using (0,1,2) as it has the lowest AIC.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -634.418
Method:                       css-mle   S.D. of innovations             30.167
Date:                Sun, 20 Mar 2022   AIC                           1276.835
Time:                        19:56:00   BIC                           1288.336
Sample:                    02-29-1980   HQIC                          1281.509
                         - 12-31-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4885      0.085     -5.742      0.000      -0.655      -0.322
ma.L1.D.Rose    -0.7601      0.101     -7.499      0.000      -0.959      -0.561
ma.L2.D.Rose    -0.2398      0.095     -2.518      0.012      -0.427      -0.053
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -4.1695           +0.0000j            4.1695            0.5000
-----------------------------------------------------------------------------
```

118. Automated ARIMA result

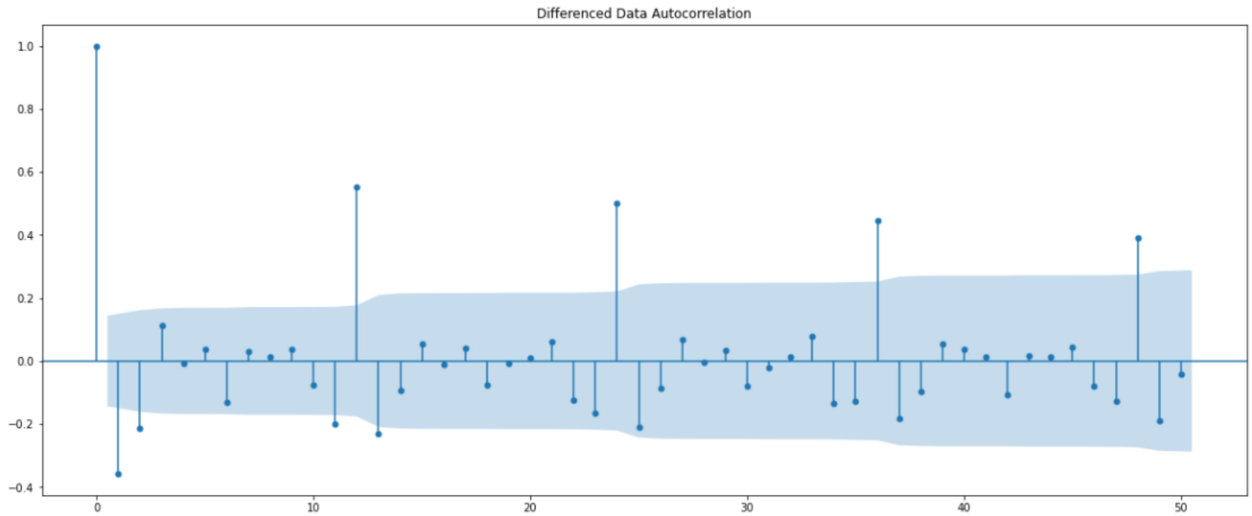
RMSE:

```
For ARIMA on the Test Data,  RMSE is 17.280
```

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
ARIMA(0,1,2)	17.279653

119. RMSE Values for Automated ARIMA & all the other models built

AUTOMATED SARIMA:



120. ACF plot showing seasonality

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.

SARIMA at 6

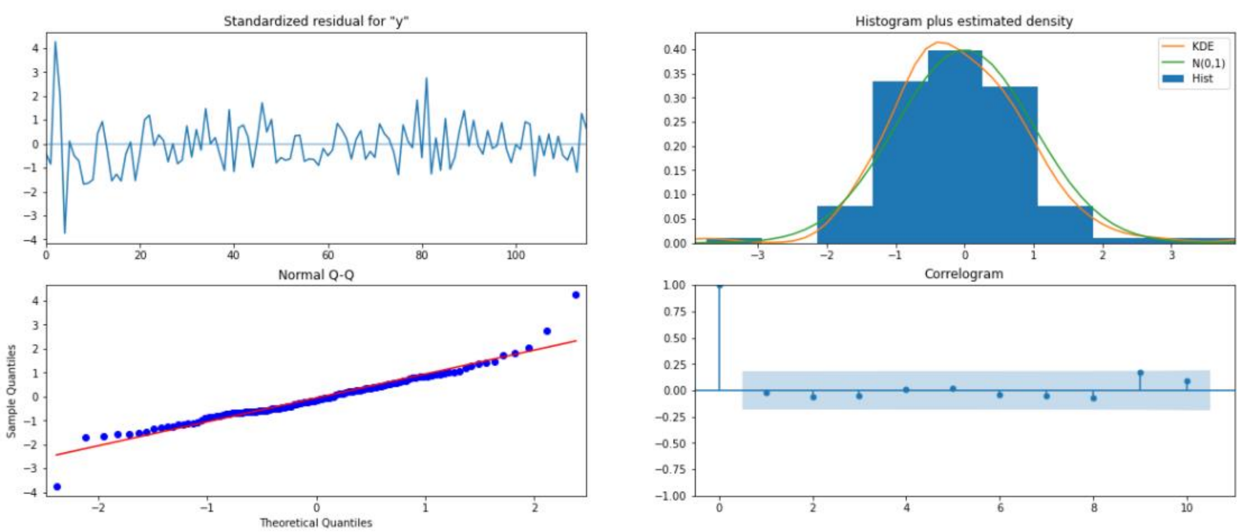
	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655821
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.220397
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

121. Automated SARIMA AIC values in ascending order

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -512.828
Date:                            Sun, 20 Mar 2022   AIC                           1041.656
Time:                                    19:56:17   BIC                           1063.685
Sample:                                         0   HQIC                          1050.598
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5941      0.152     -3.901      0.000      -0.893      -0.296
ma.L1         -0.1954    142.307     -0.001      0.999    -279.113     278.722
ma.L2         -0.8047    114.551     -0.007      0.994    -225.320     223.711
ar.S.L6       -0.0626      0.035     -1.764      0.078      -0.132       0.007
ar.S.L12       0.8450      0.039     21.886      0.000       0.769       0.921
ma.S.L6        0.2226    748.180      0.000      1.000   -1466.184    1466.629
ma.S.L12      -0.7774    581.671     -0.001      0.999   -1140.831    1139.277
sigma2       335.1706   2.59e+05      0.001      0.999   -5.08e+05    5.09e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):                56.67
Prob(Q):                              0.79   Prob(JB):                         0.00
Heteroskedasticity (H):               0.47   Skew:                             0.52
Prob(H) (two-sided):                  0.02   Kurtosis:                         6.26
===================================================================================

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

122. SARIMAX AT 6 RESULT



123. Plot Diagnostic for SARIMA

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:

For SARIMA at 6 on the Test Dataa, RMSE is 17.27965342842907

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
ARIMA(0,1,2)	17.279653
SARIMA(1,1,2)(2,0,2,6)	17.279653

124. RMSE Values for Automated SARIMA & all the other models built

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

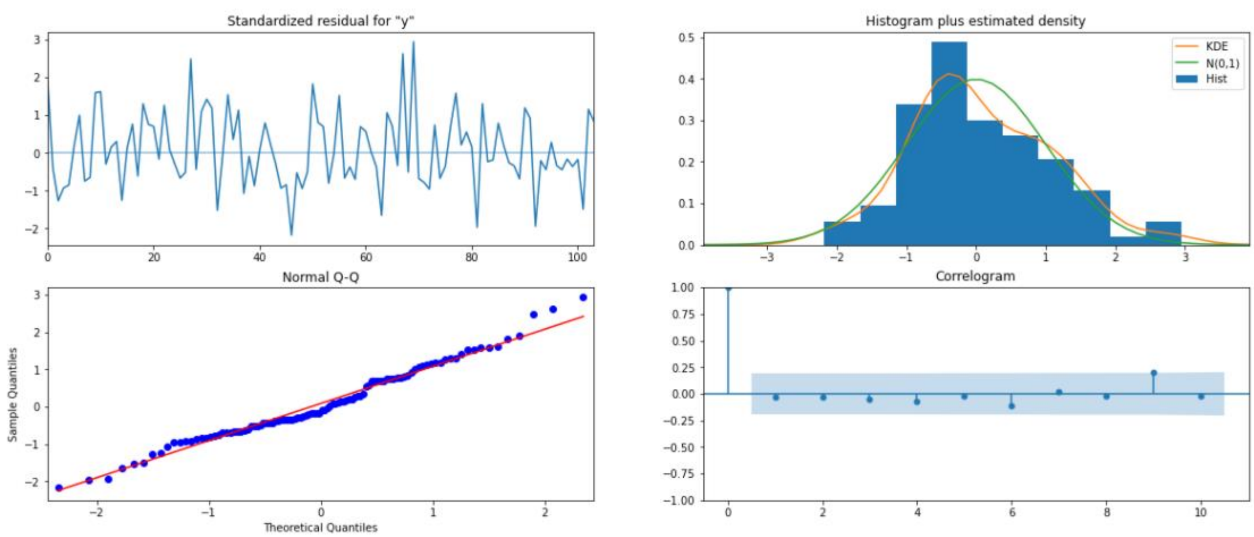
	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686964
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

125. SARIMA values based on lowest AIC

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 20 Mar 2022   AIC                            887.938
Time:                                    19:56:39   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    190.073     -0.004      0.996    -373.380     371.694
ma.L2         -0.1573     29.862     -0.005      0.996     -58.685      58.370
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.78e+04      0.005      0.996   -9.34e+04    9.39e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================

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

126. SARIMA at 12 RESULT



127. Plot Diagnostic for SARIMA at 12

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:

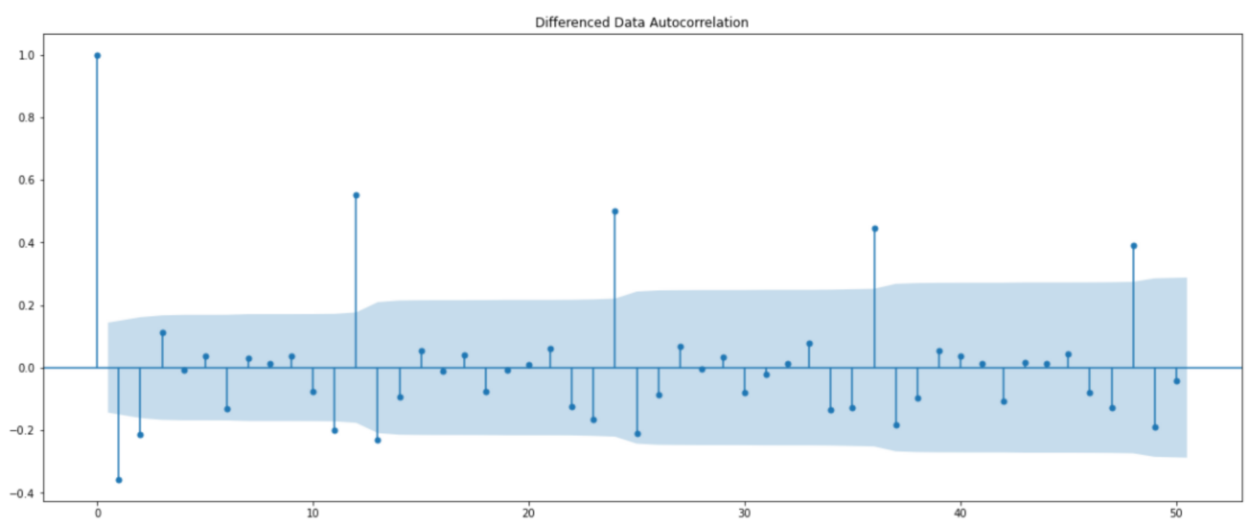
```
For SARIMA on the Test Data,  RMSE is 26.417
```

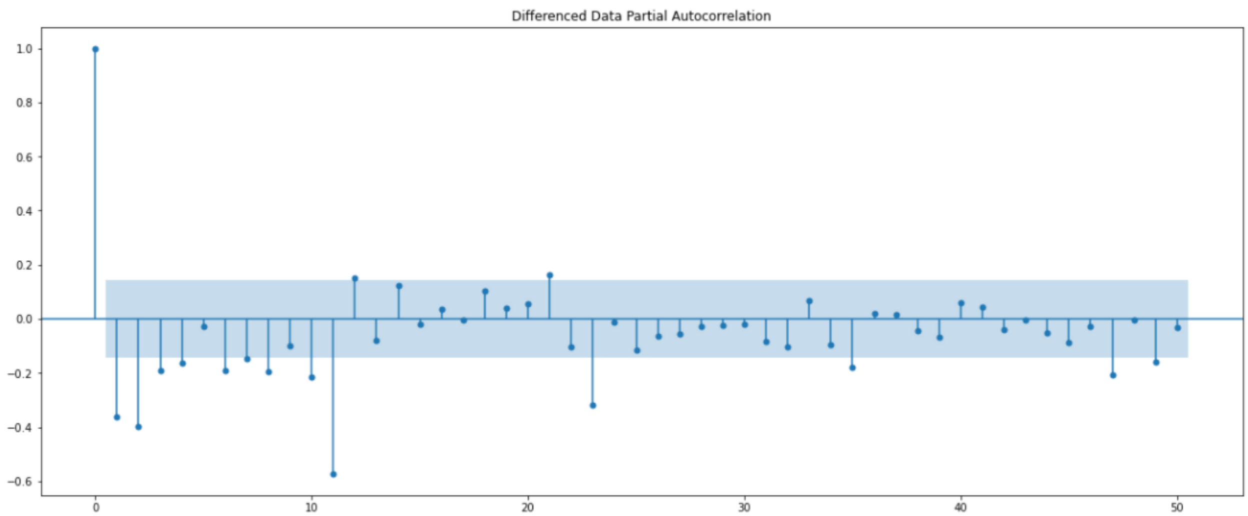
	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
ARIMA(0,1,2)	17.279653
SARIMA(1,1,2)(2,0,2,6)	17.279653
SARIMA(0,1,2)(2,0,2,12)	26.417374

128. RMSE Values for Automated SARIMA at 12 & all the other models built

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.

ARIMA:





129. ACF & PACF Plot

Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 2. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2. By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.

```
                             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, 20 Mar 2022   AIC                           1279.299
Time:                        19:56:40   BIC                           1296.550
Sample:                    02-29-1980   HQIC                          1286.309
                         - 12-31-1990
================================================================================
                   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           18.3389           +0.0000j           18.3389            0.0000
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.4961           +0.0000j            1.4961            0.5000
-----------------------------------------------------------------------------
```

130. ARIMA model Result

RMSE:

```
For MANUAL ARIMA on the Test Data,  RMSE is 17.076
```

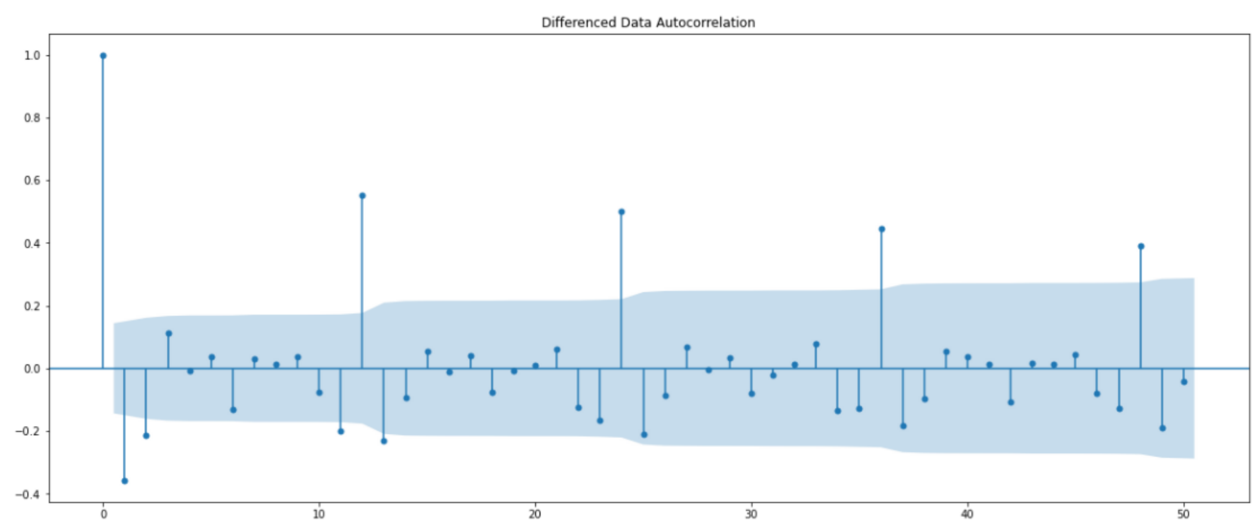
	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
ARIMA(0,1,2)	17.279653
SARIMA(1,1,2)(2,0,2,6)	17.279653
SARIMA(0,1,2)(2,0,2,12)	26.417374
MANUAL ARIMA(2,1,2)	17.075734

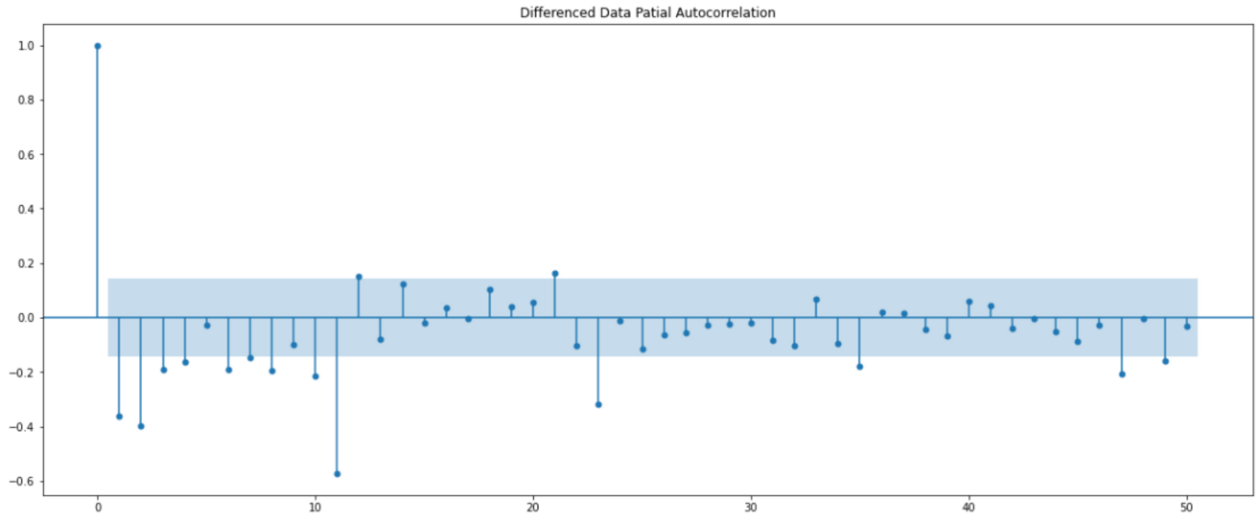
131. RMSE Values for Manual ARIMA & all the other models built

SARIMA

On obseriving the ACF & PACF plots above we notice that there is a trend and a seasonality. So, now we take a seasonal differencing and check the series.

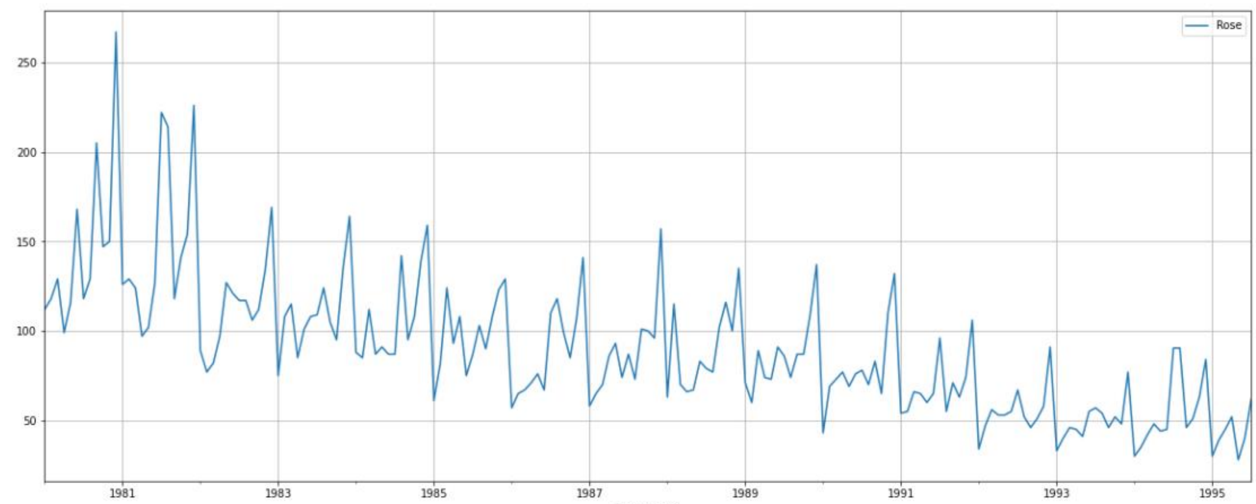
Difference at 6:





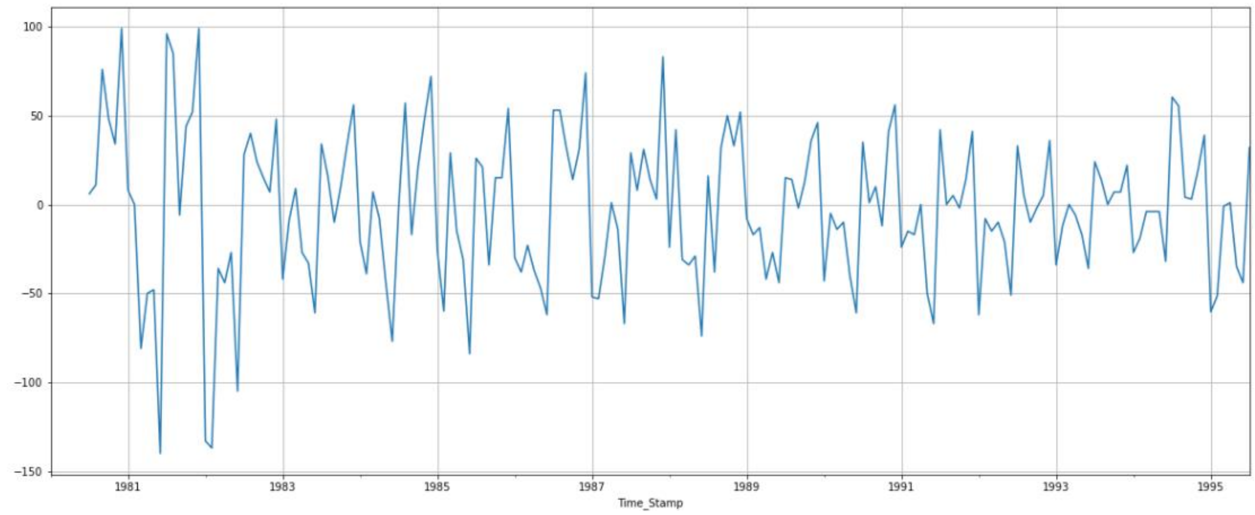
132. Differenced Data: ACF & PACF Plot

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.



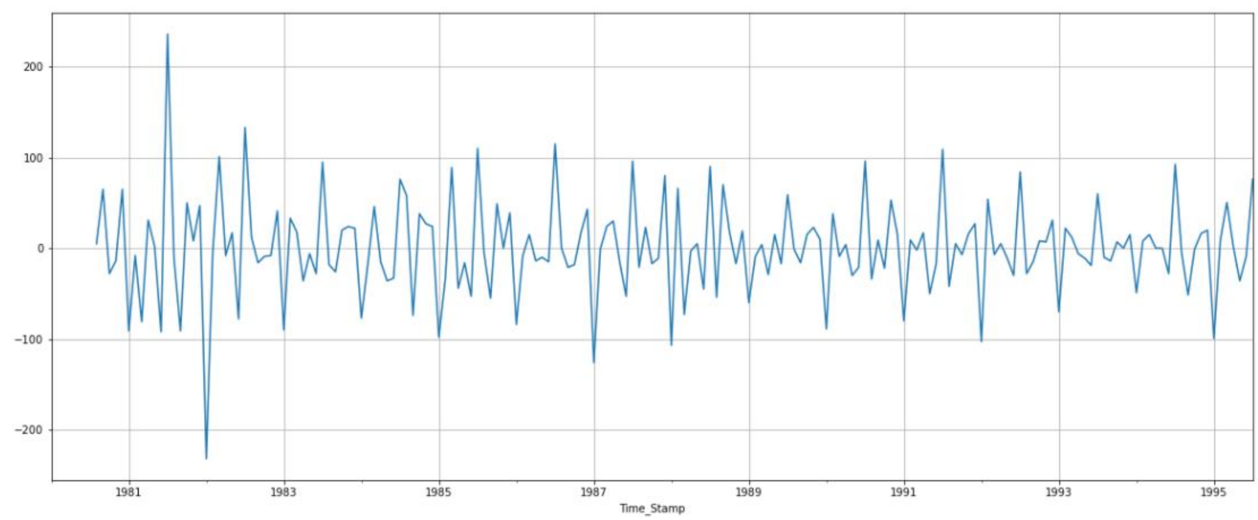
133. Series Graph

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



134. Series Graph : with difference at 6

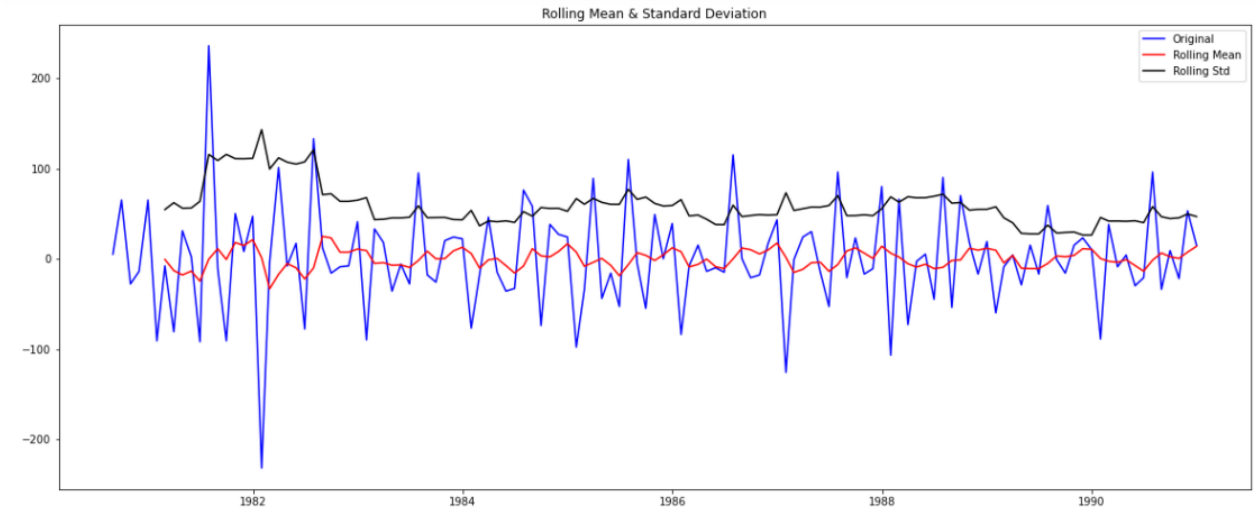
We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series



135. Series Graph: with order 1 difference for seasonality at 6 differenced series

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

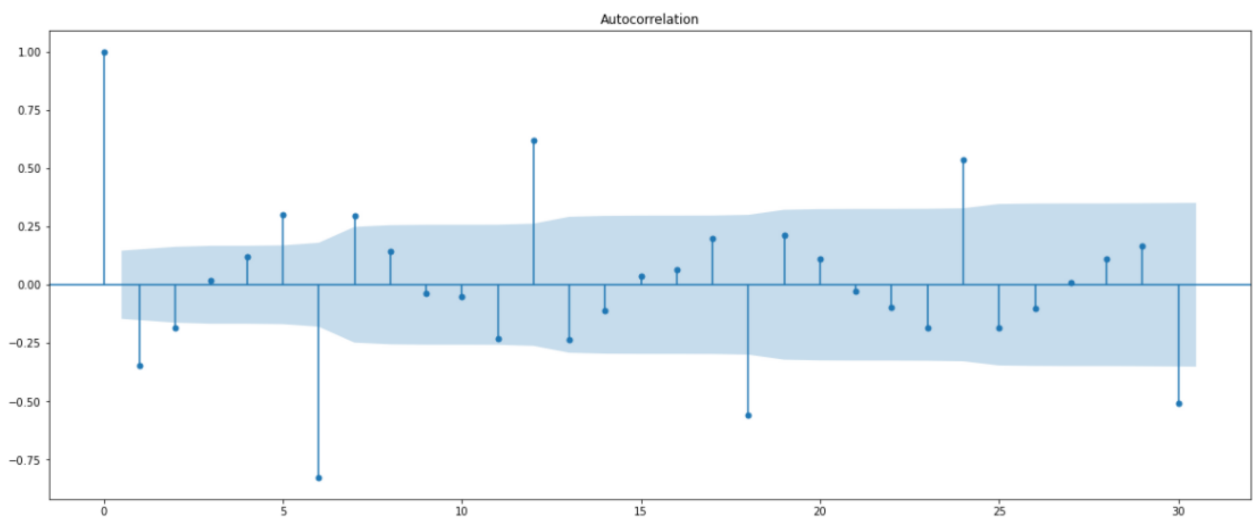
Let us go ahead and check the stationarity of the above series before fitting the SARIMA model

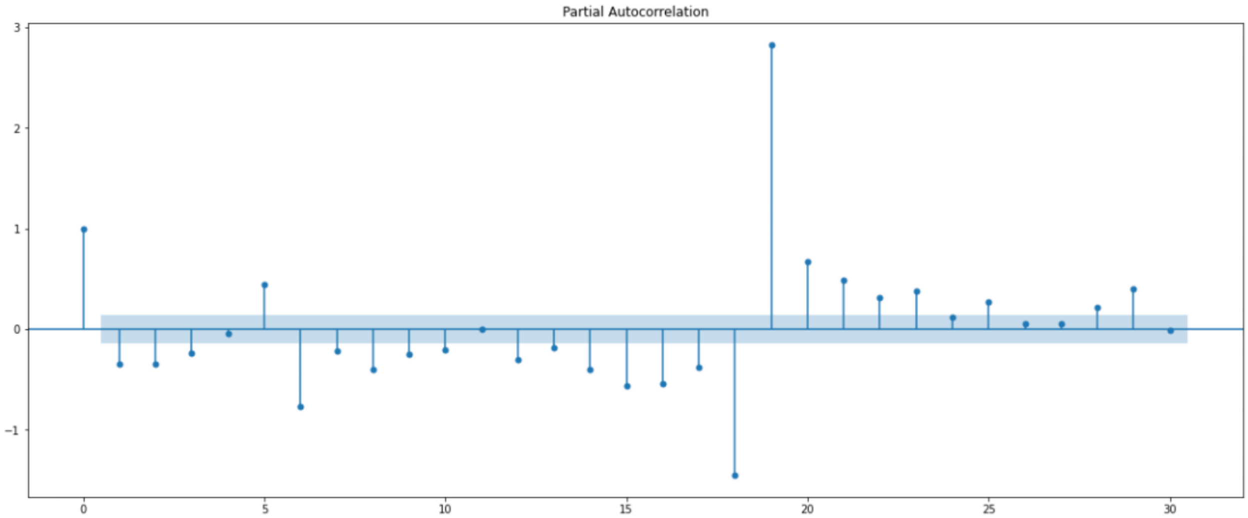



```
Results of Dickey-Fuller Test:
Test Statistic                -6.882869e+00
p-value                        1.418693e-09
#Lags Used                     1.300000e+01
Number of Observations Used    1.110000e+02
Critical Value (1%)           -3.490683e+00
Critical Value (5%)           -2.887952e+00
Critical Value (10%)          -2.580857e+00
dtype: float64
```

136. Stationarity of dataset for building SARIMA model

We can see that the data is stationary. We will now plot the new ACF & PACF plots for the new modified time series.





137. ACF & PACF Plot for SARIMA at 6: newly modified Time Series

Here, we have taken alpha=0.05.

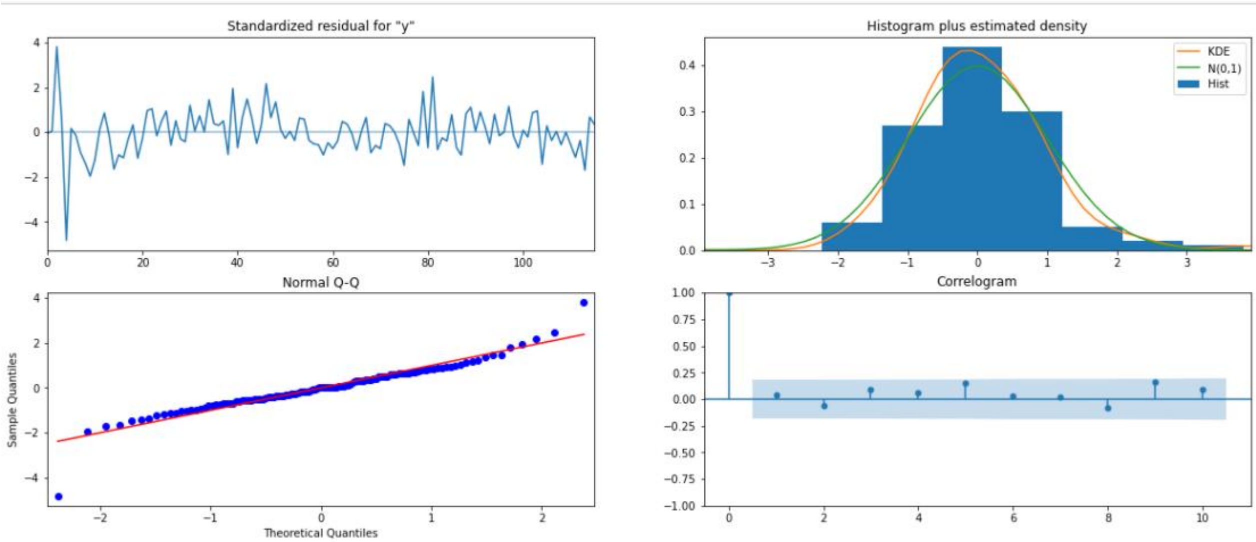
We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2. The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 2. 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.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -513.610
Date:                           Sun, 20 Mar 2022   AIC                           1045.220
Time:                                   19:56:43   BIC                           1070.003
Sample:                                        0   HQIC                          1055.281
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0477      0.120      8.745      0.000       0.813       1.283
ar.L2         -0.2224      0.134     -1.659      0.097      -0.485       0.040
ma.L1         -1.9995     37.091     -0.054      0.957     -74.697      70.698
ma.L2          1.0003     37.121      0.027      0.979     -71.755      73.756
ar.S.L6       -0.1126      0.025     -4.519      0.000      -0.162      -0.064
ar.S.L12       0.7999      0.025     32.477      0.000       0.752       0.848
ma.S.L6        0.2936     37.098      0.008      0.994     -72.418      73.005
ma.S.L12      -0.7063     26.200     -0.027      0.978     -52.057      50.644
sigma2       315.1612      0.178   1773.411      0.000     314.813     315.509
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):               120.33
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               0.44   Skew:                            -0.31
Prob(H) (two-sided):                  0.01   Kurtosis:                         7.95
===================================================================================

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+20. Standard errors may be unstable.
```

139. Manual SARIMA Result



140. Plot Diagnostic for SARIMA at 6

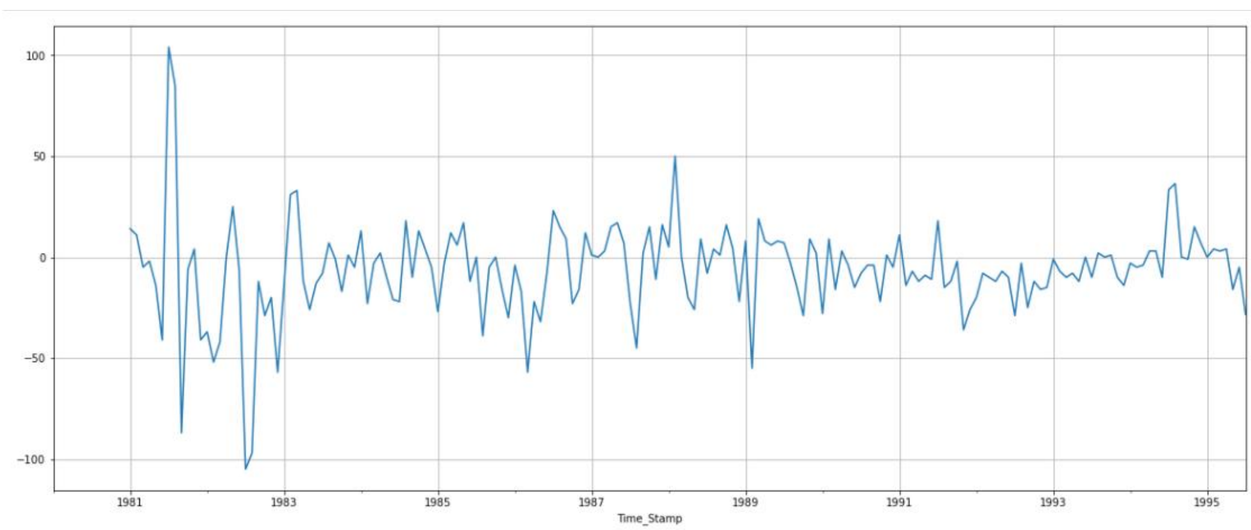
RMSE:

```
For MANUAL SARIMA on the Test Data,  RMSE is 29.600
```

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
ARIMA(0,1,2)	17.279653
SARIMA(1,1,2)(2,0,2,6)	17.279653
SARIMA(0,1,2)(2,0,2,12)	26.417374
MANUAL ARIMA(2,1,2)	17.075734
MANUAL SARIMA(2,1,2)(2,0,2,6)	29.599792

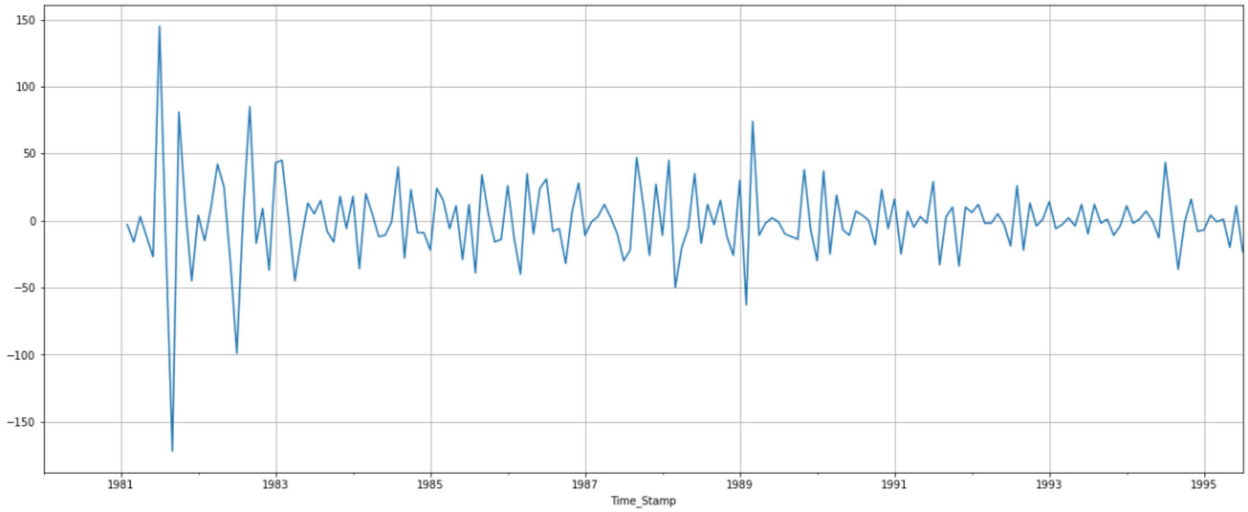
141. RMSE Values for Manual SARIMA at 6 & all the other models built

SARIMA AT 12:

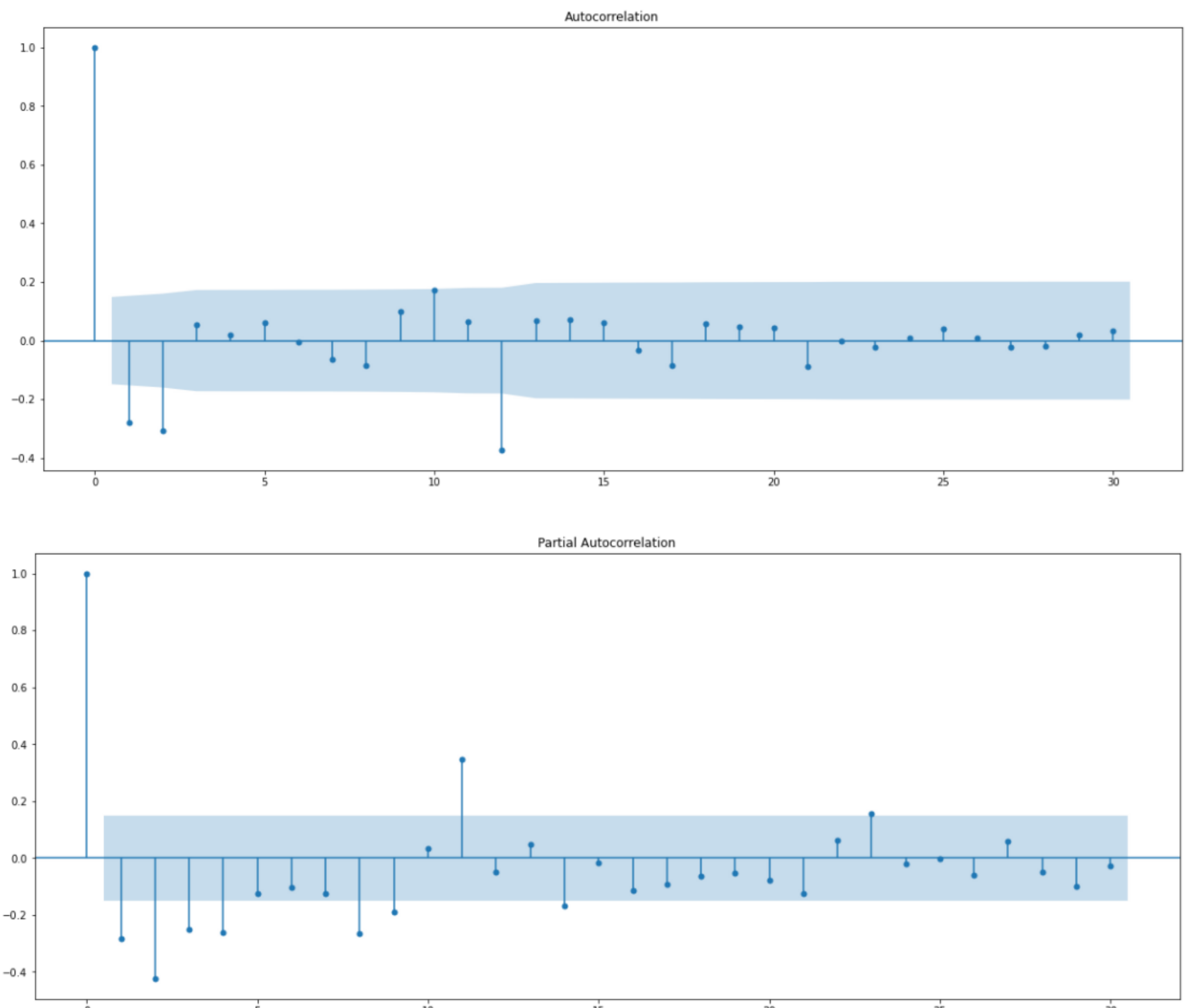


142. Graph for SARIMA at 12

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.



143. Graph for data set: SARIMA at 12 1st order difference on seasonally differenced series

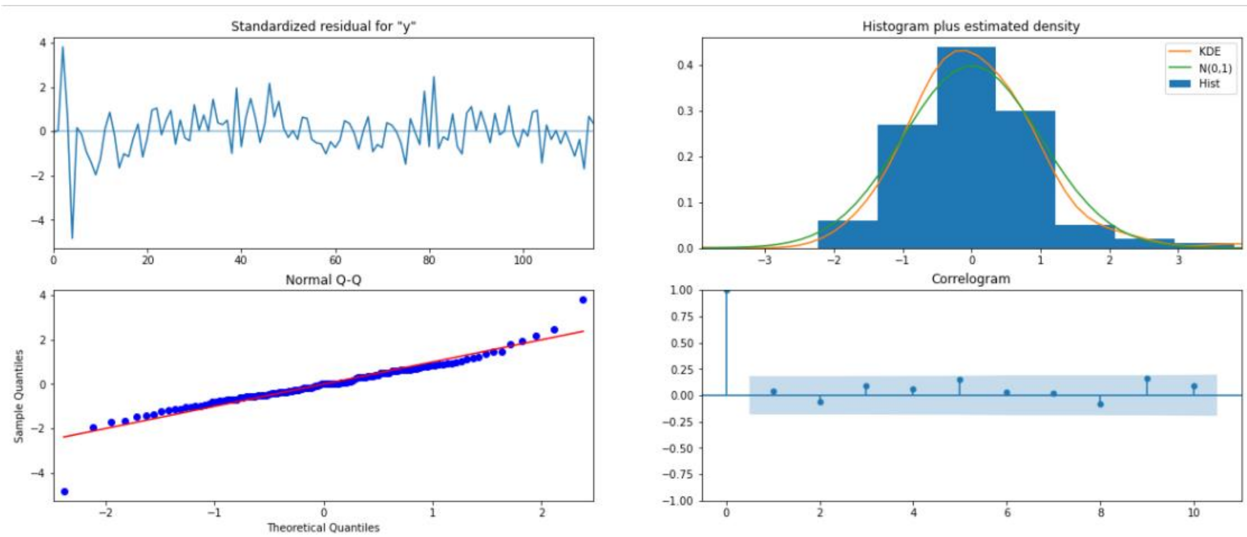


144. ACF & PACF Plots for newly modified Time Series: SARIMA at 12

```
                                    SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -513.610
Date:                           Sun, 20 Mar 2022   AIC                           1045.220
Time:                                   19:56:46   BIC                           1070.003
Sample:                                        0   HQIC                          1055.281
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0477      0.120      8.745      0.000       0.813       1.283
ar.L2         -0.2224      0.134     -1.659      0.097      -0.485       0.040
ma.L1         -1.9995     37.091     -0.054      0.957     -74.697      70.698
ma.L2          1.0003     37.121      0.027      0.979     -71.755      73.756
ar.S.L6       -0.1126      0.025     -4.519      0.000      -0.162      -0.064
ar.S.L12       0.7999      0.025     32.477      0.000       0.752       0.848
ma.S.L6        0.2936     37.098      0.008      0.994     -72.418      73.005
ma.S.L12      -0.7063     26.200     -0.027      0.978     -52.057      50.644
sigma2       315.1612      0.178   1773.411      0.000     314.813     315.509
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):               120.33
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               0.44   Skew:                            -0.31
Prob(H) (two-sided):                  0.01   Kurtosis:                         7.95
===================================================================================

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+20. Standard errors may be unstable.
```

145. SARIMA at 12 Result



146. Plot Diagnostics: SARIMA at 12

RMSE:

```
For MANUAL SARIMA on the Test Data,  RMSE is 29.600
```

	Test RMSE
Alpha= 0.07, SES	35.936198
Alpha= 1.908, Beta= 7.302, DES	16.979631
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
Regression	16.979414
Naive Model	78.396083
Simple Average Model	52.318735
2point Trailing Moving Average	12.298291
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
ARIMA(0,1,2)	17.279653
SARIMA(1,1,2)(2,0,2,6)	17.279653
SARIMA(0,1,2)(2,0,2,12)	26.417374
MANUAL ARIMA(2,1,2)	17.075734
MANUAL SARIMA(2,1,2)(2,0,2,6)	29.599792
MANUAL SARIMA(2,1,2)(2,0,2,12)	29.599792

147. RMSE Values for Manual SARIMA at 12 & all the other models built

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

	Test RMSE
2point Trailing Moving Average	12.298291
Alpha=0.1, Beta=0.01, Gamma=0.5, TES	15.548490
4 point Trailing Moving Average	15.845558
6point Trailing Moving Average	15.986163
9 point Trailing Moving Average	16.500823
Regression	16.979414
Alpha= 1.908, Beta= 7.302, DES	16.979631
MANUAL ARIMA(2,1,2)	17.075734
ARIMA(0,1,2)	17.279653
SARIMA(1,1,2)(2,0,2,6)	17.279653
SARIMA(0,1,2)(2,0,2,12)	26.417374
MANUAL SARIMA(2,1,2)(2,0,2,6)	29.599792
MANUAL SARIMA(2,1,2)(2,0,2,12)	29.599792
Alpha= 0.07, SES	35.936198
Simple Average Model	52.318735
Naive Model	78.396083

148. Data Frame: Models Built along with their RMSE

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.

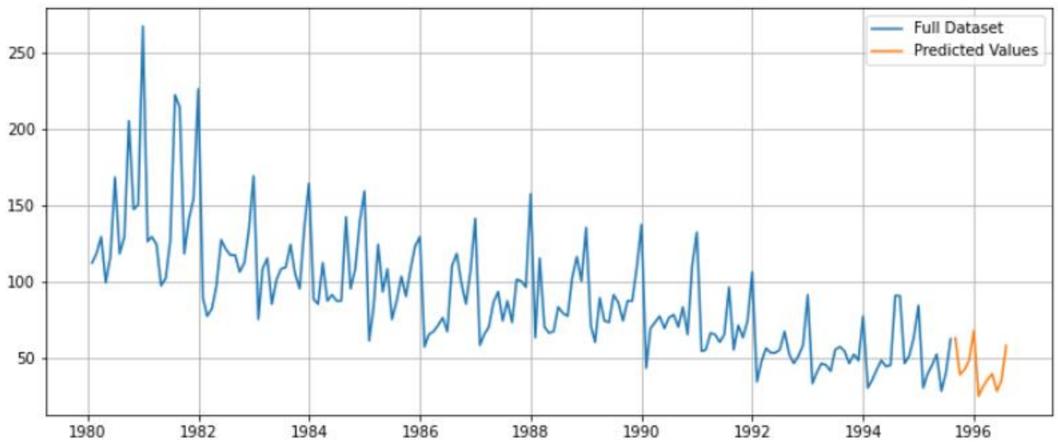
Building the most optimum model on the Full Data

We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters $\alpha = 0.1$, $\beta = 0.01$ and $\gamma = 0.05$.

The same is put into a dataframe for building the most optimum model on the complete data.

The RMSE score for the full model is 18.970

Then the prediction on full model is done. The graph for the same is below:



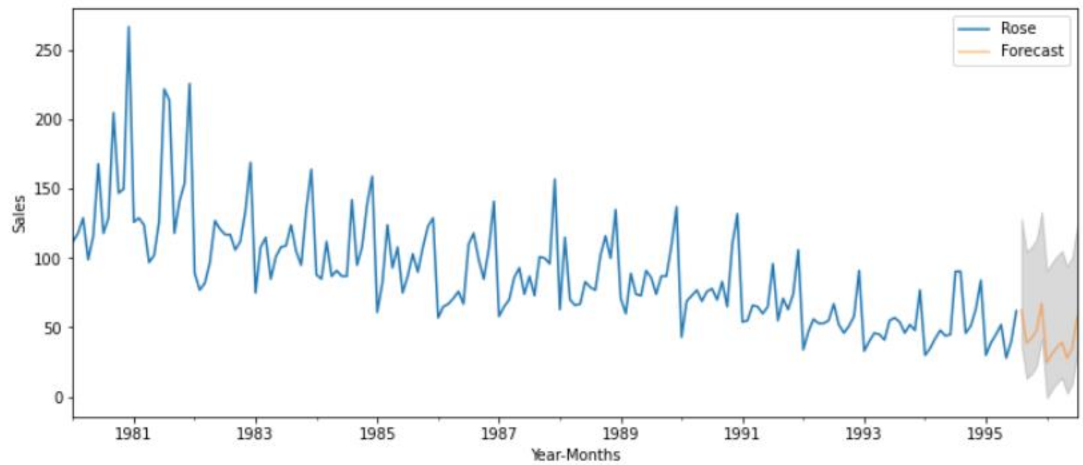
149. Graph: Prediction on Full Dataset

```
1995-08-31    62.590882
1995-09-30    38.767694
1995-10-31    42.086701
1995-11-30    47.946761
1995-12-31    67.441246
1996-01-31    24.610148
1996-02-29    30.686046
1996-03-31    35.595213
1996-04-30    39.225245
1996-05-31    27.918623
1996-06-30    34.727766
1996-07-31    57.674372
Freq: M, dtype: float64
```

150. List of Prediction for next 12 months

	lower_CI	prediction	upper_ci
1995-08-31	37.070903	62.590882	128.698277
1995-09-30	13.247715	38.767694	104.875090
1995-10-31	16.566722	42.086701	108.194097
1995-11-30	22.426782	47.946761	114.054156
1995-12-31	41.921267	67.441246	133.548642

151. Class Intervals



152. Graph for class intervals and next 12 months prediction

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

- On visualization the above data, it can be noted that Rose has the highest sale around the year 1980, and in the month of December.
- Following December the month of November & October have also had number of Rose wine sale over the years to be more than other months.
- The month of January, February, June and May have the lowest number of Rose Wine sales across the years.
- The decomposition of the data was first done using additive decomposition, but it showed residuals to have pattern around 0, then on using multiplicative method the data did not show any pattern for the residuals and the components were located around 1.
- Several models were built: Simple Exponential, Double Exponential, Triple Exponential Smoothing, Simple Average, moving average, linear regression.
- The models were built on training data and performance was checked on test data using RMSE.
- The lowest RMSE value is considered to be optimal for creating the best model for overall predictions.
- Based on the above models built, Triple Exponential Smoothing (with parameters, alpha=0.1, beta=0.01, gamma=0.5) had the lowest AIC value at 18.970.
- With the use of TES as the best fit model, future sales predictions for next 12 months are done. Plotting and predicting the same along with confidence intervals.
- Using the predictions from the models, the company now has an insight on the profit, loss, production quantity, materials needed for the production of Sparkling Wine.
- With these insights the company can forecast which month will have higher or lower production with respect to the demand and supply for the particular month.
- To make their business more profitable, the company can use various strategies, such as advertisements in Radio, Television, Local Newspapers, pamphlets etc to grow their consumer base, and boost sale, and production for the time period where the sale is low.
- Various discounts and offers can be given both to the sale team as well as people who purchase the products to boost sales.

- In future, the company can have low production for months with lowest sale to avoid wastage which in turn helps with profit.
- Special offers at certain events can be discussed to uplift the production at low seasons.