BUSINESS REPORT
Time Series Forecasting
(Sparkling Wine Dataset)

Submitted By
:Ayush
Agarwal

Index

Problem: - For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze and forecast Wine Sales in the 20th century.

Dataset 1: -

(Sparkling Wines)

1.1 : - Read the data as an appropriate Time Series data and plot the data.

Converting the data into appropriate time series data our dataset will look like this:

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

Table No. 1.1.1

```
['1980-01-01', '1980-02-01', '1980-03-01', '1980-04-01',
 '1980-05-01', '1980-06-01', '1980-07-01', '1980-08-01',
 '1980-09-01', '1980-10-01',
 ...
 '1994-10-01', '1994-11-01', '1994-12-01', '1995-01-01',
 '1995-02-01', '1995-03-01', '1995-04-01', '1995-05-01',
 '1995-06-01', '1995-07-01'],
```

Fig No. 1.1.1

To understand this time series properly we will plot the data:

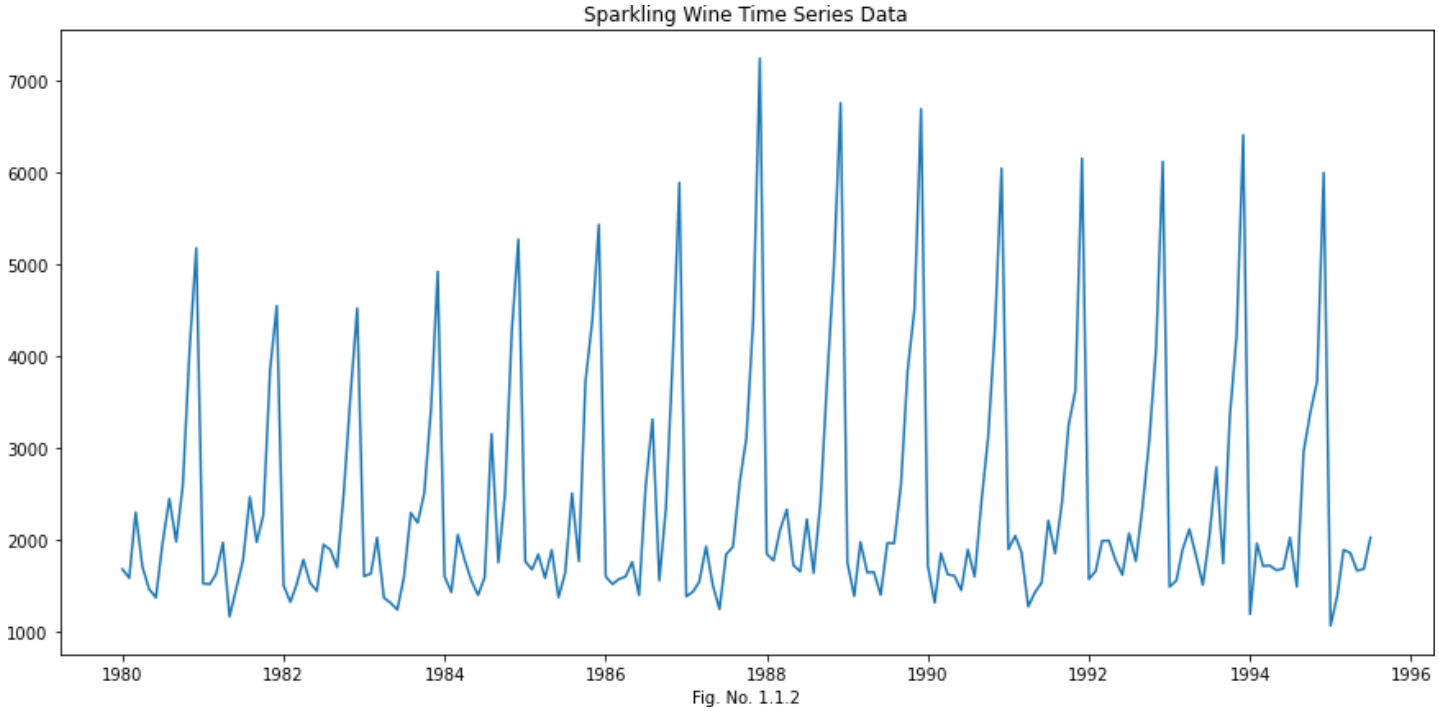


Fig. No. 1.1.2

1.2 : - Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Exploratory Data Analysis: -

a.) Shape of the dataset: -

Sparkling wine data attribute have 187 records from **01-01-1980** to **01-07-1995**.

b.) Checking for null values: -

Does not have any null values.

c.) Checking Descriptive Statistics of the Dataset: -

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

Table No. 1.2.1

d.) Checking mean and median value comparison along with dataset plot: -

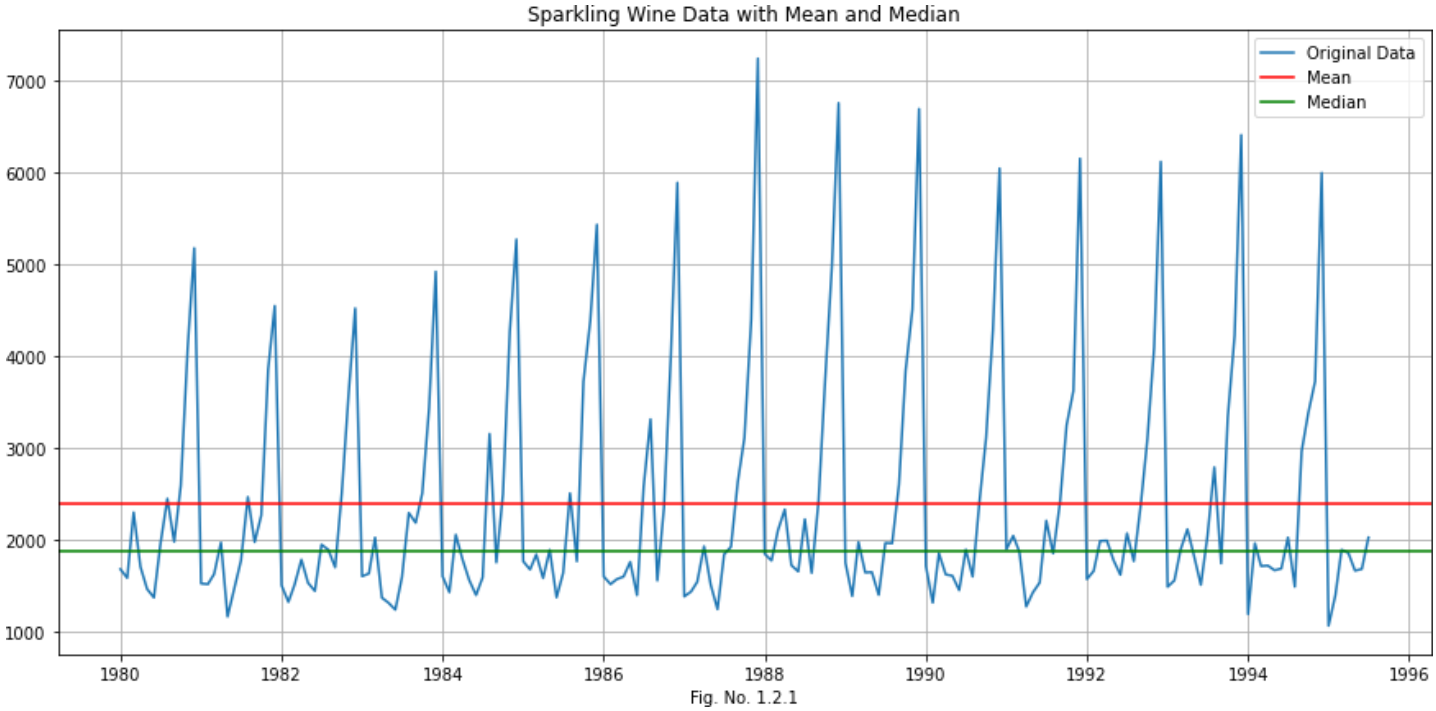


Fig. No. 1.2.1

e.) Checking yearly boxplot of the dataset: -

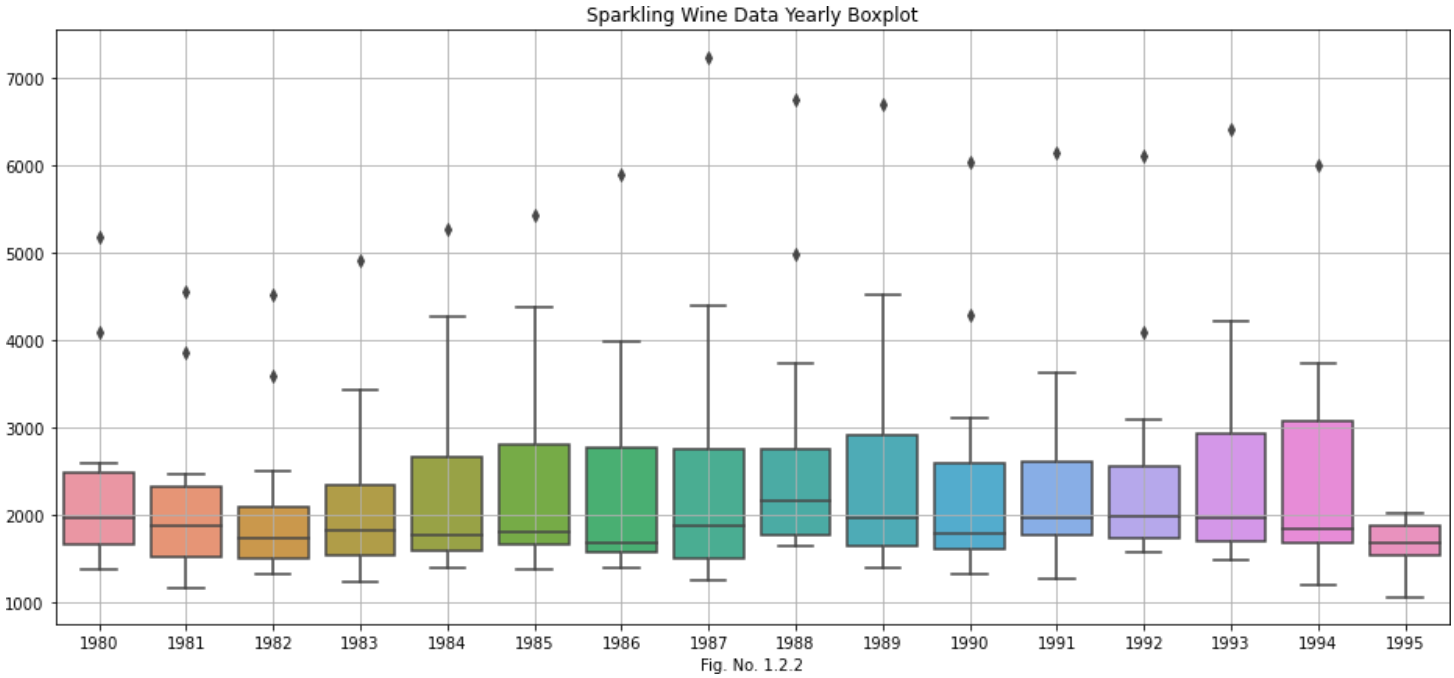


Fig. No. 1.2.2

f.) Checking monthly boxplot of the dataset: -

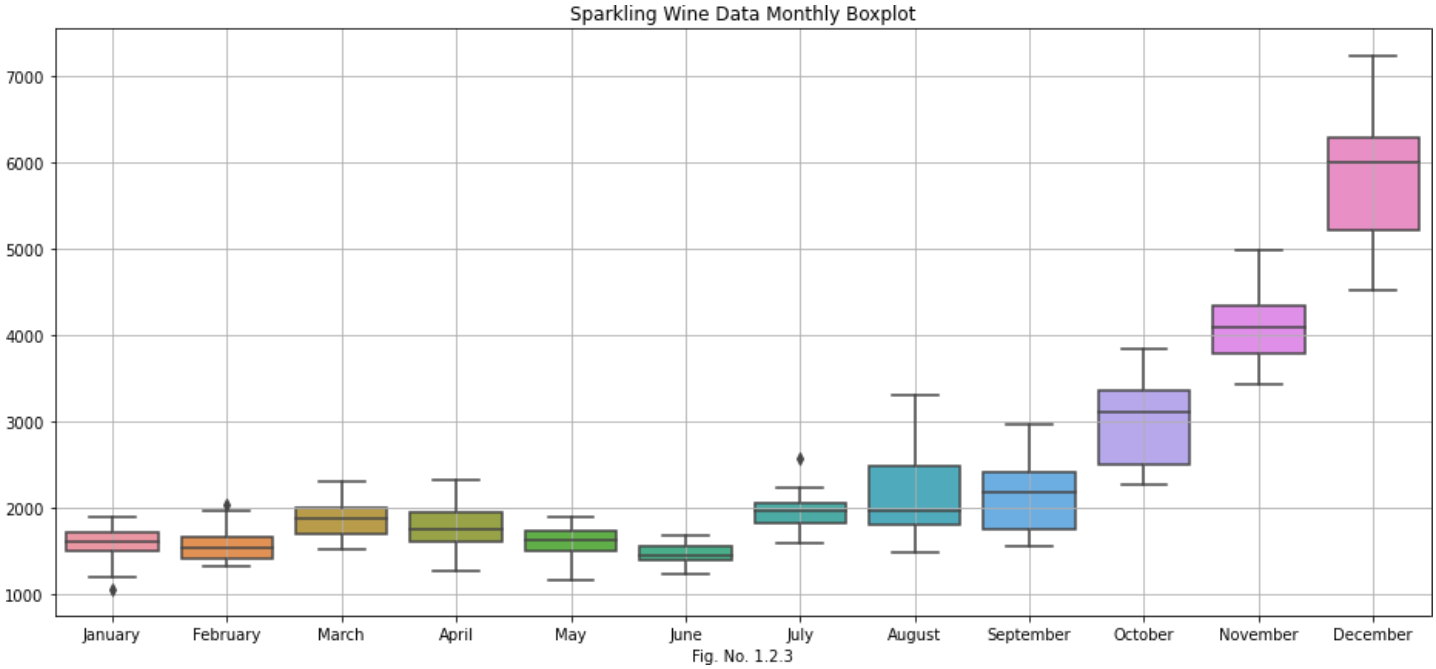


Fig. No. 1.2.3

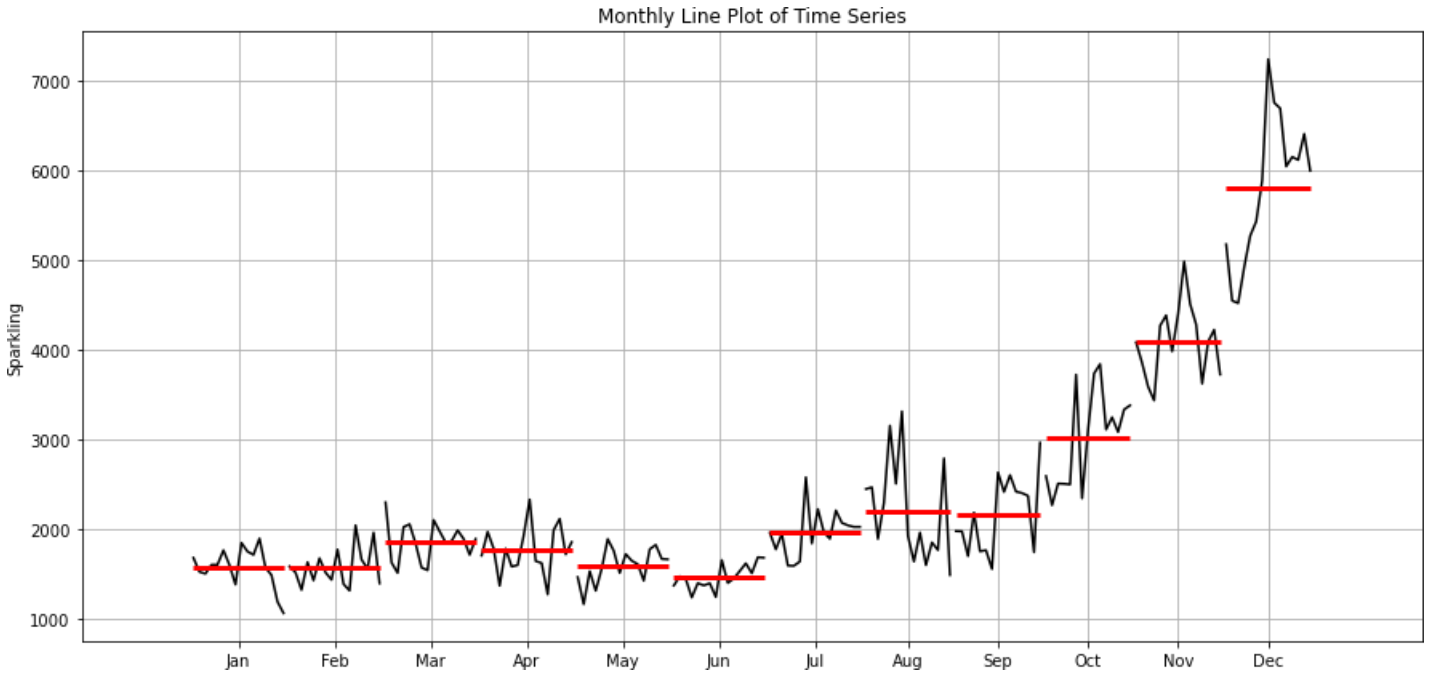
g.) Comparison between monthly and yearly data using line plot: -

I.) Monthly and Yearly Table: -

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

Table No. 1.2.2

II.) Monthly Line Plot with Respect to Every Year: -



III.) Yearly Line Plot With Respect to Every Month: -

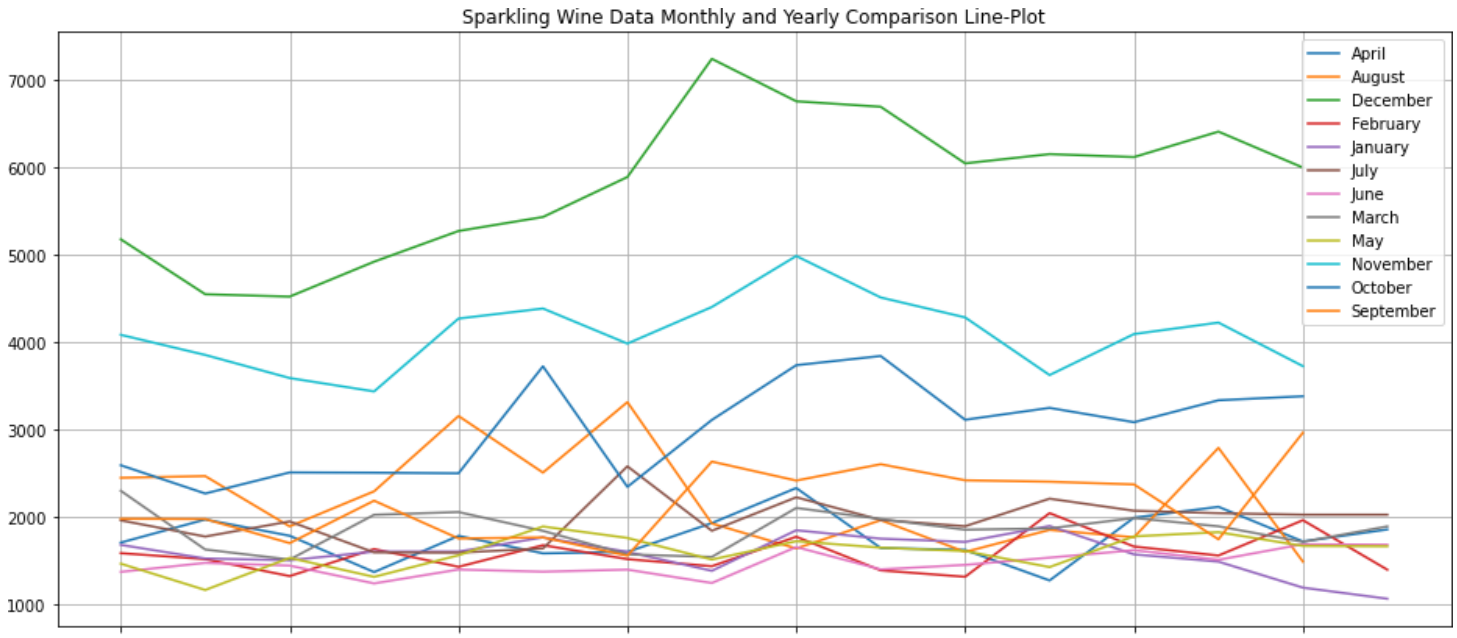


Fig. No. 1.2.4

Insights from the Exploratory Data Analysis: -

- Sparkling wine production data is present from Jan-1980 to July-1995.
- Data does not have any null values.
- Descriptive statistics shows that it has outliers present in the dataset, it represents the huge variation in wine production within the months or years.
- Outlier present in every individual boxplot in yearly boxplot which means any specific month is having high wine production volume.
- Last quarter of the year produce high volume of wine in which December month having high production followed by November, October and September.

Decomposition: -

a.) Additive: -

After decomposing dataset using Additive model we get trend, seasonal and residual(error) plot,

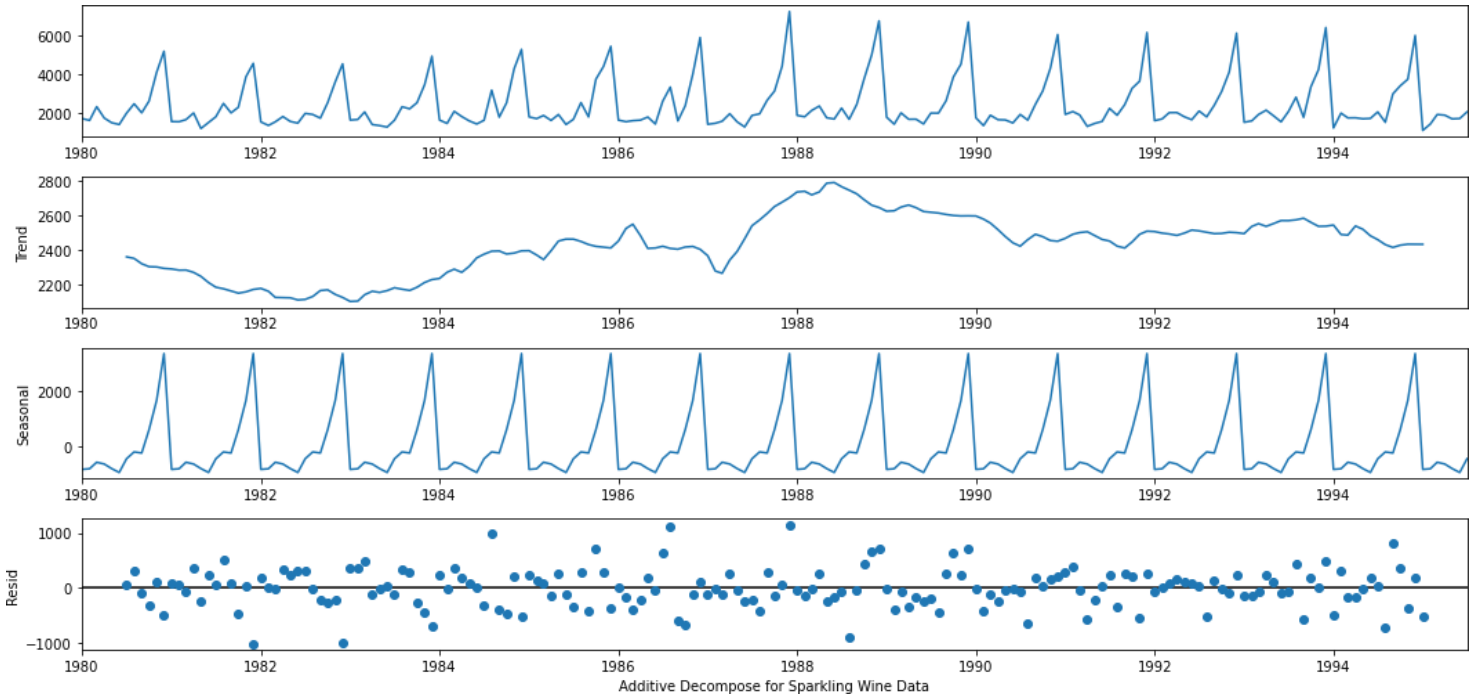


Fig No. 1.2.5

Insights from Additive Decomposition: -

- Trend is not observed.
- Seasonality is observed.
- Residual (Error) lying within the range of -1000 and 1000.

b.) Multiplicative: -

After decomposing dataset using Multiplicative model we get trend, seasonal and residual(error) plot,

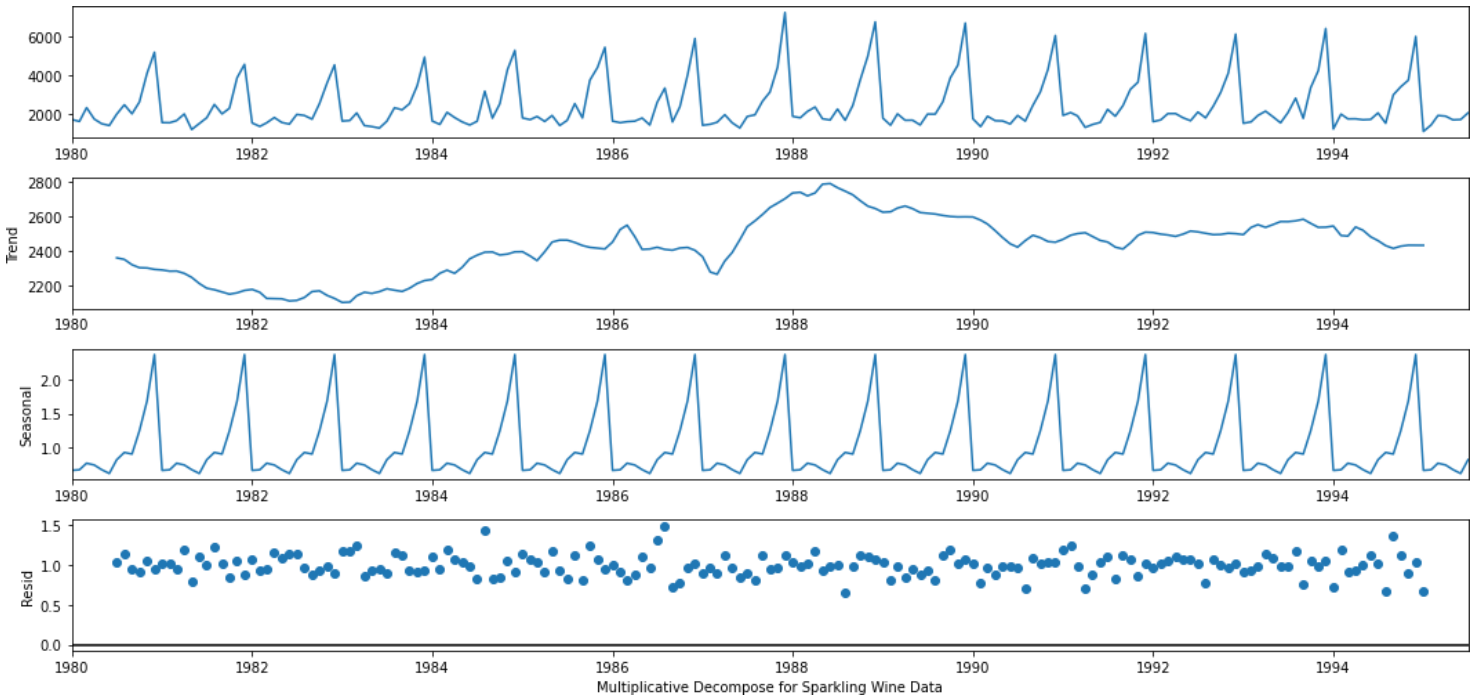


Fig No. 1.2.6

Insights from Multiplicative Decomposition: -

- Trend is not observed.
- Seasonality is observed.
- Residual (Error) lying within the range of 0.5 and 1.5, here this is percentage error.

1.3 : - Split the data into training and test. The test data should start in 1991.

After splitting the dataset into train and test set our new dataset will look like this,

Top 5 Rows of Train Data

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

Top 5 Rows of Test Data

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

- Train Dataset having range from **Jan-1980 to Dec-1990** i.e., 132 records.
- Test Dataset having range from **Jan-1991 to Jul-1995** i.e., 55 records.

Plot of train and test dataset,

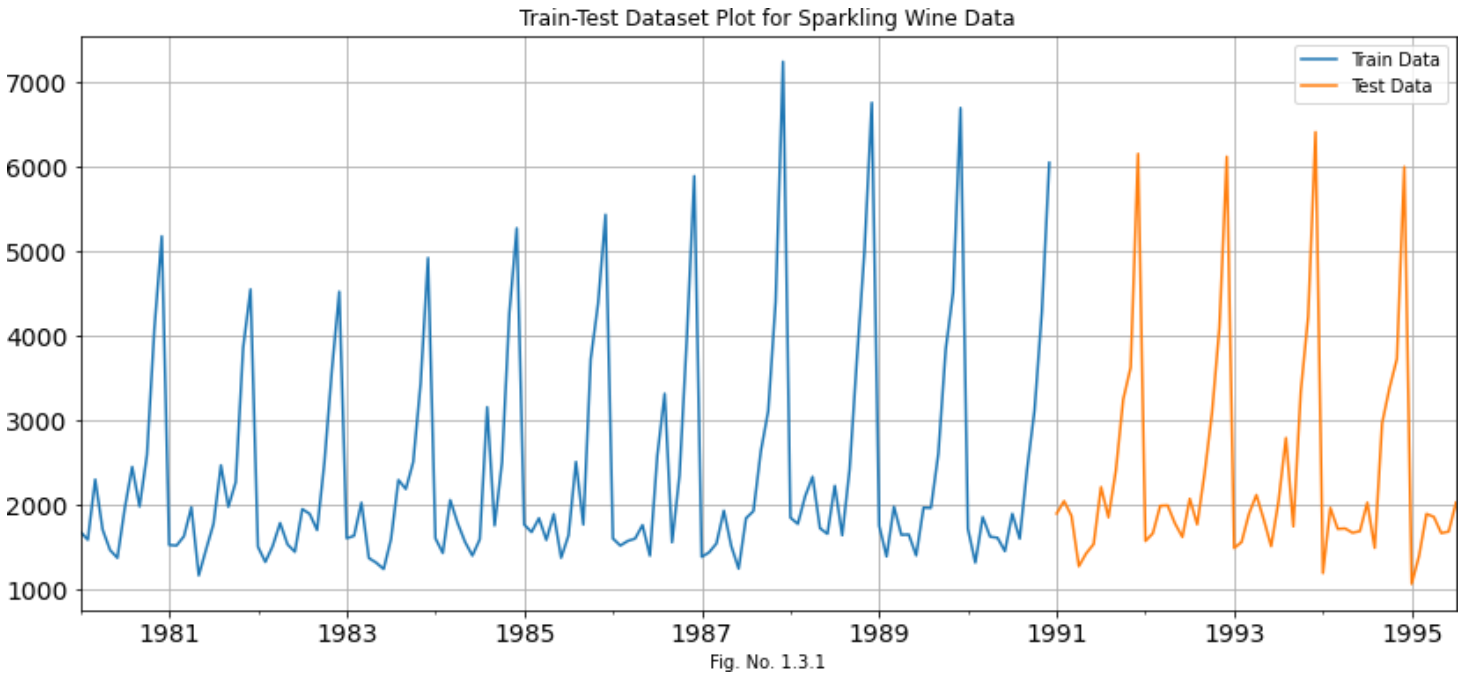


Fig. No. 1.3.1

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

First, we will evaluate on **Linear Regression, Naïve Model, Simple Average, Moving Average** and then on **exponentialsmoothing.**

Linear Regression

After adding date range column in an ordinal format to the regression model as independent variable we can forecast accordingly, dataset after adding date range it will look like,

Top 5 Rows for Linear Regression Train

YearMonth	Sparkling	time
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

Top 5 Rows for Linear Regression Test

YearMonth	Sparkling	time
1991-01-01	1902	133
1991-02-01	2049	134
1991-03-01	1874	135
1991-04-01	1279	136
1991-05-01	1432	137

After training the dataset on train dataset and predicting it on test dataset we got our predicted values which can be visualizevia this plot,

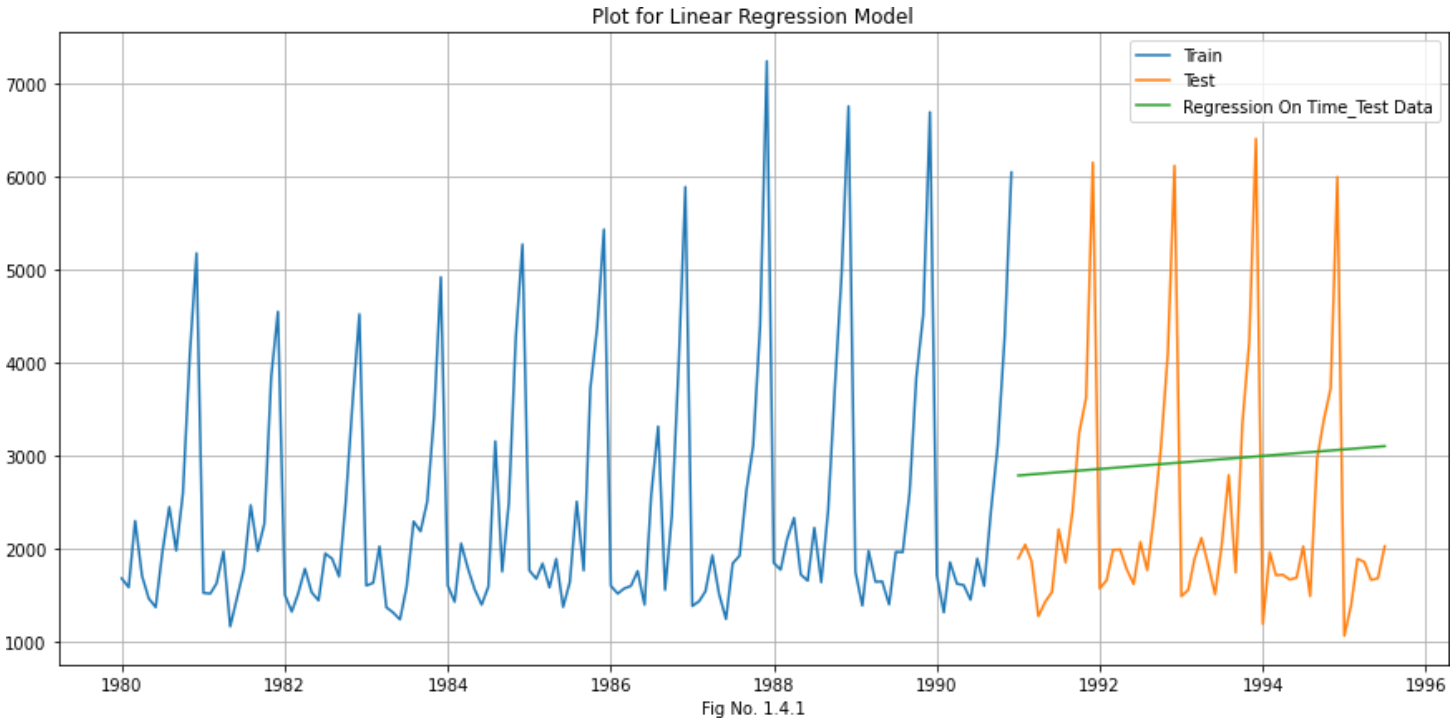


Fig No. 1.4.1

Model Evaluation: -

We can evaluate the model by calculating the RSME (Root Mean Square Error) on Test Data, minimum the RSME better the model and for this model RSME would be,

	Test RMSE
RegressionOnTime	1389.135175

Root Mean Square Error of Linear Regression Model for Test Data is 1389.135

Naïve Model

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

In naïve model, predicted values are of the train dataset last value. for this model forecast values would be 6047.

After getting the predicted values we can visualize our data,

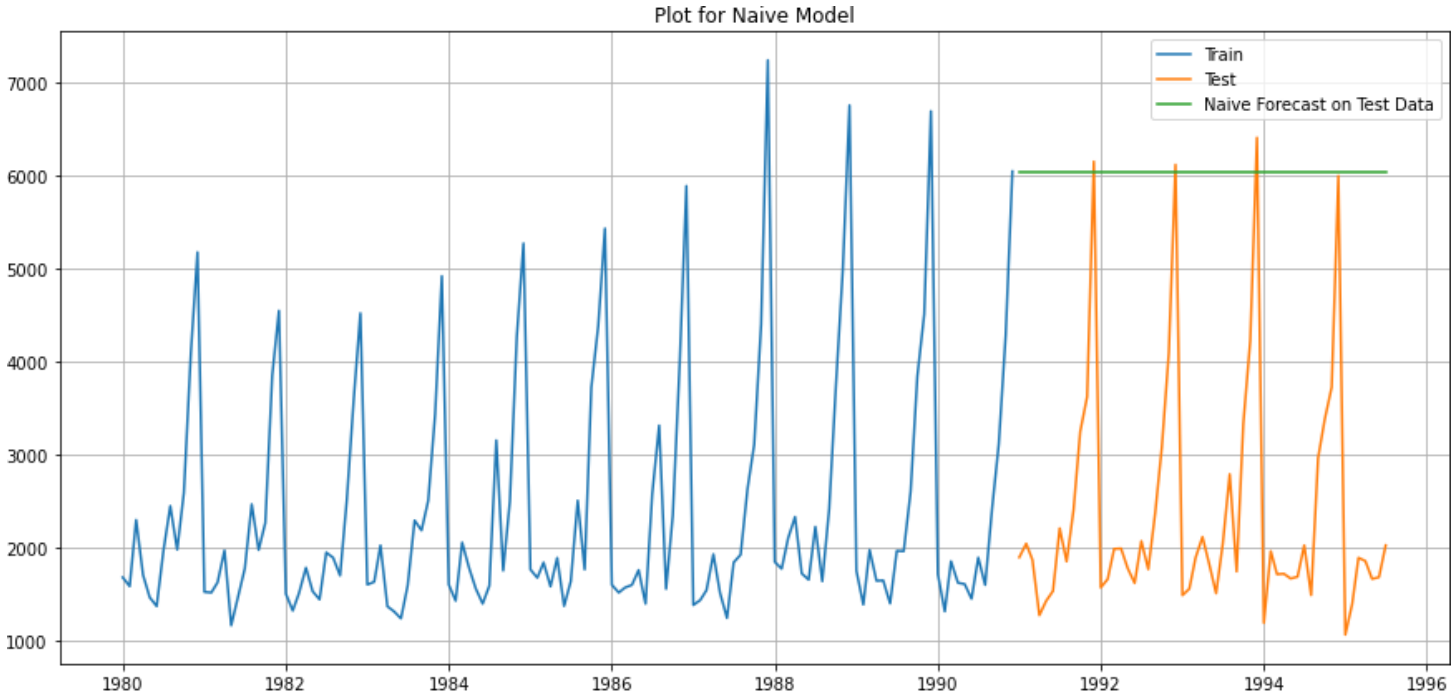


Fig No. 1.4.2

Model Evaluation: -

For this model RSME would be,

	Test RMSE
RegressionOnTime	1389.135175
NaiveOnTime	3864.279352

Root Mean Square Error of Naive Model for Test Data is 3864.279.

Simple Average Model

YearMonth	Sparkling	Avg
1991-01-01	1902	2403.780303
1991-02-01	2049	2403.780303
1991-03-01	1874	2403.780303
1991-04-01	1279	2403.780303
1991-05-01	1432	2403.780303

For simple average model, we will predict through mean of train dataset, and here mean of train dataset is 2403.7803

After getting the predicted values we can visualize our data,

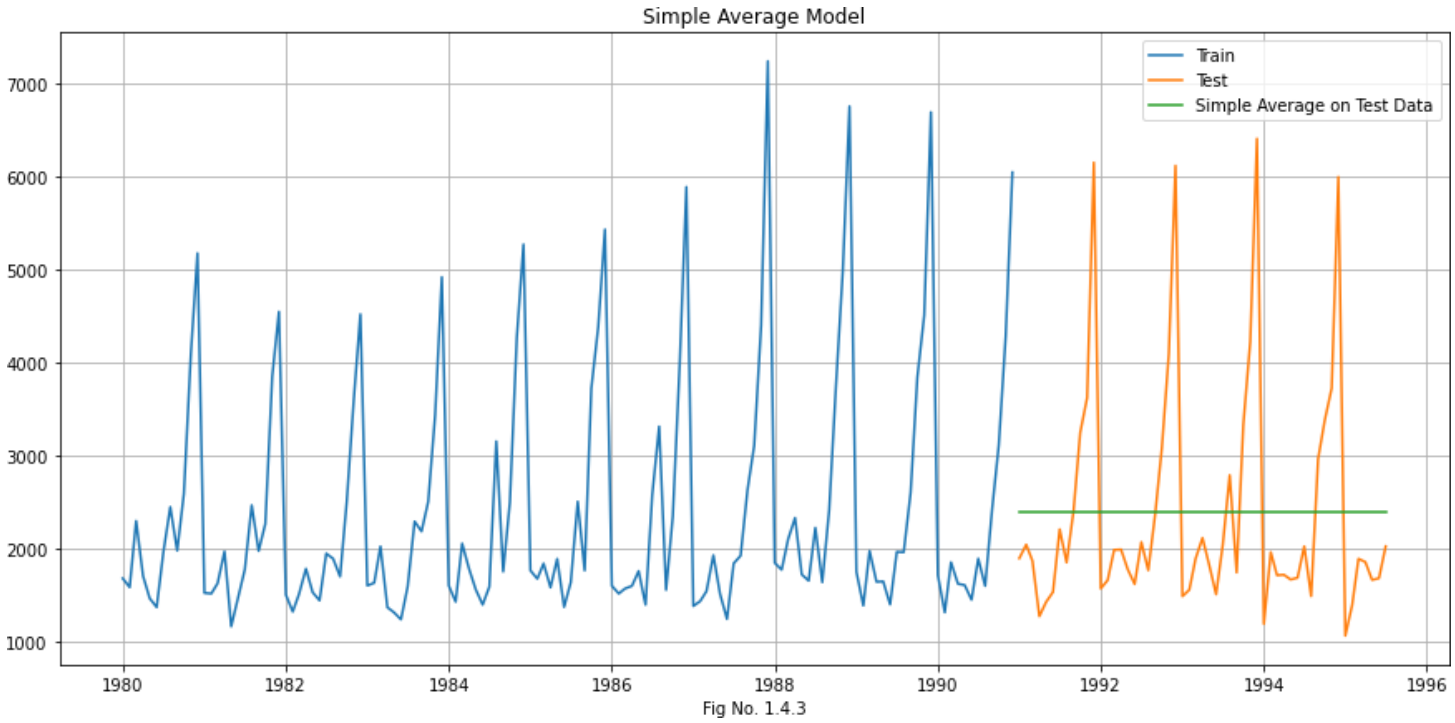


Fig No. 1.4.3

Model Evaluation: -

For this model RSME would be,

	Test RMSE
RegressionOnTime	1389.135175
NaiveOnTime	3864.279352
SimpleAverage	1275.081804

Root Mean Square Error of Simple Average Model for Test Data is 1275.0818.

Moving Average Model

YearMonth	Sparkling	Trailing2	Trailing4	Trailing6	Trailing9
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

A moving average is defined as an average of fixed number of items in the time series which move through the series by dropping the top items of the previous averaged group and adding the next in each successive average.
Here we took averages of 2,4,6 and 9 successive records.

After getting the predicted values on each moving average rolling parameter we can visualize our data,

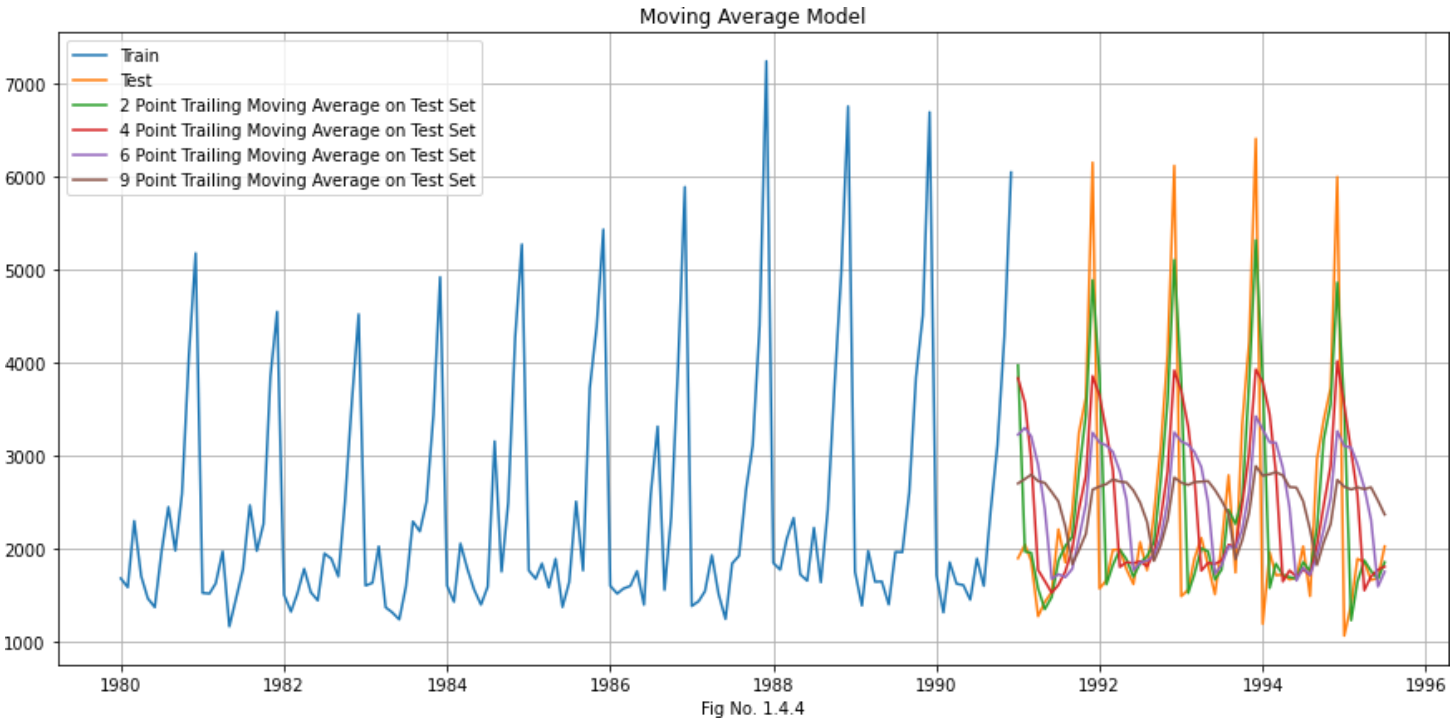


Fig No. 1.4.4

Here it is difficult to say that which rolling parameter is best to analyze, we will check RSME values for each parameter.

Model Evaluation: -

For this model RSME would be,

	Test RMSE
RegressionOnTime	1389.135175
NaiveOnTime	3864.279352
SimpleAverage	1275.081804
2 Point Trailing on Test Data	813.400684
4 Point Trailing on Test Data	1156.589694
6 Point Trailing on Test Data	1283.927428
9 Point Trailing on Test Data	1346.278315

Root Mean Square Error of Moving Average Model for Test Data for 2 Point Trailing Average is 813.4006.

We can also compare all forecasting parameter using plot,

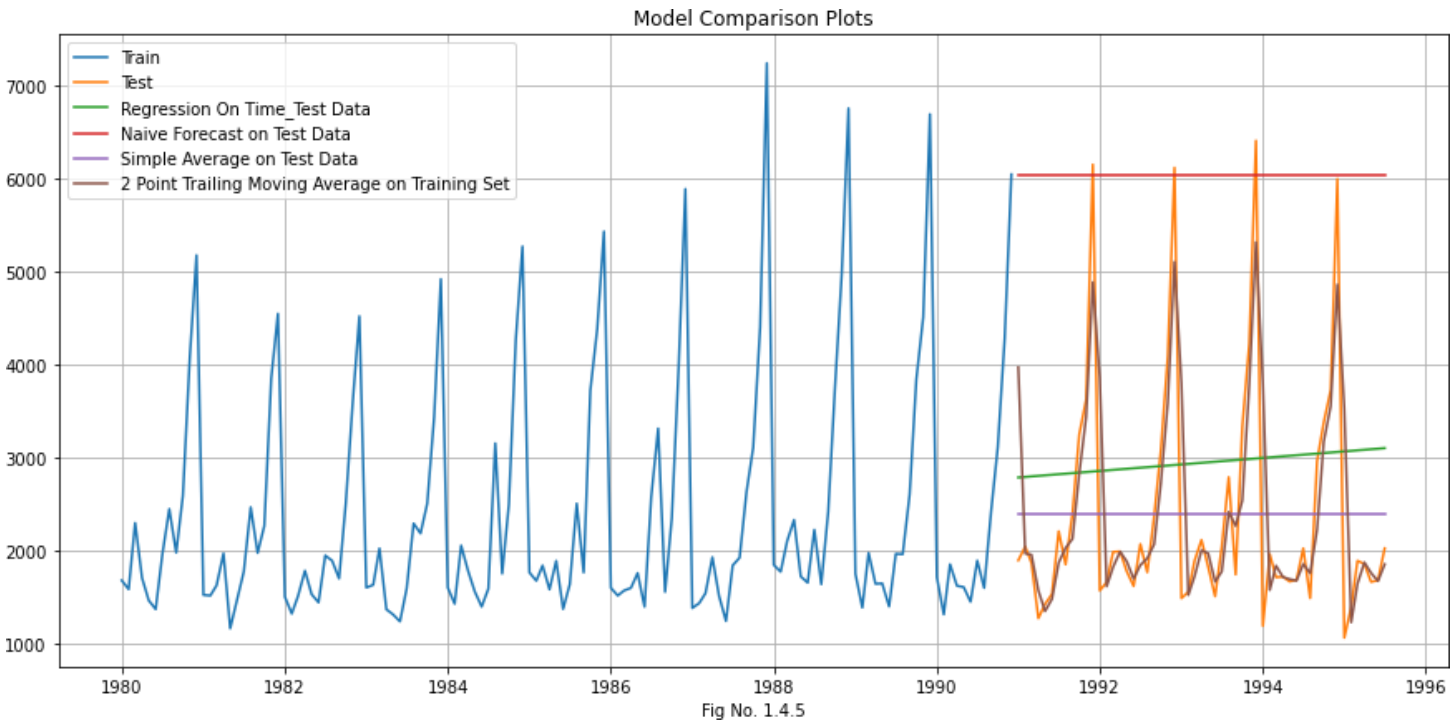


Fig No. 1.4.5

From the above plot it is clearly seen that **2-point trailing moving average** have best fitted line to test dataset, and for more better forecast result and RSME value we will go for Exponential Smoothing Method.

Simple Exponential Smoothing

For more better model we will evaluate our model using exponential smoothing, and simple exponential smoothing is oneof them, here we consider smoothing level only and after fitting our model we will get best parameter to analyze further,

```
{'smoothing_level': 0.07029120765764557,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1764.0137060346985,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Best Parameters for Simple Exponential Smoothing

```
1991-01-01    2804.675124
1991-02-01    2804.675124
1991-03-01    2804.675124
1991-04-01    2804.675124
1991-05-01    2804.675124
Freq: MS, dtype: float64
```

Top 5 Rows of Predicted Values on Test Data for Simple Exponential Smoothing

After prediction on best parameters, we will plot the data for better understanding,

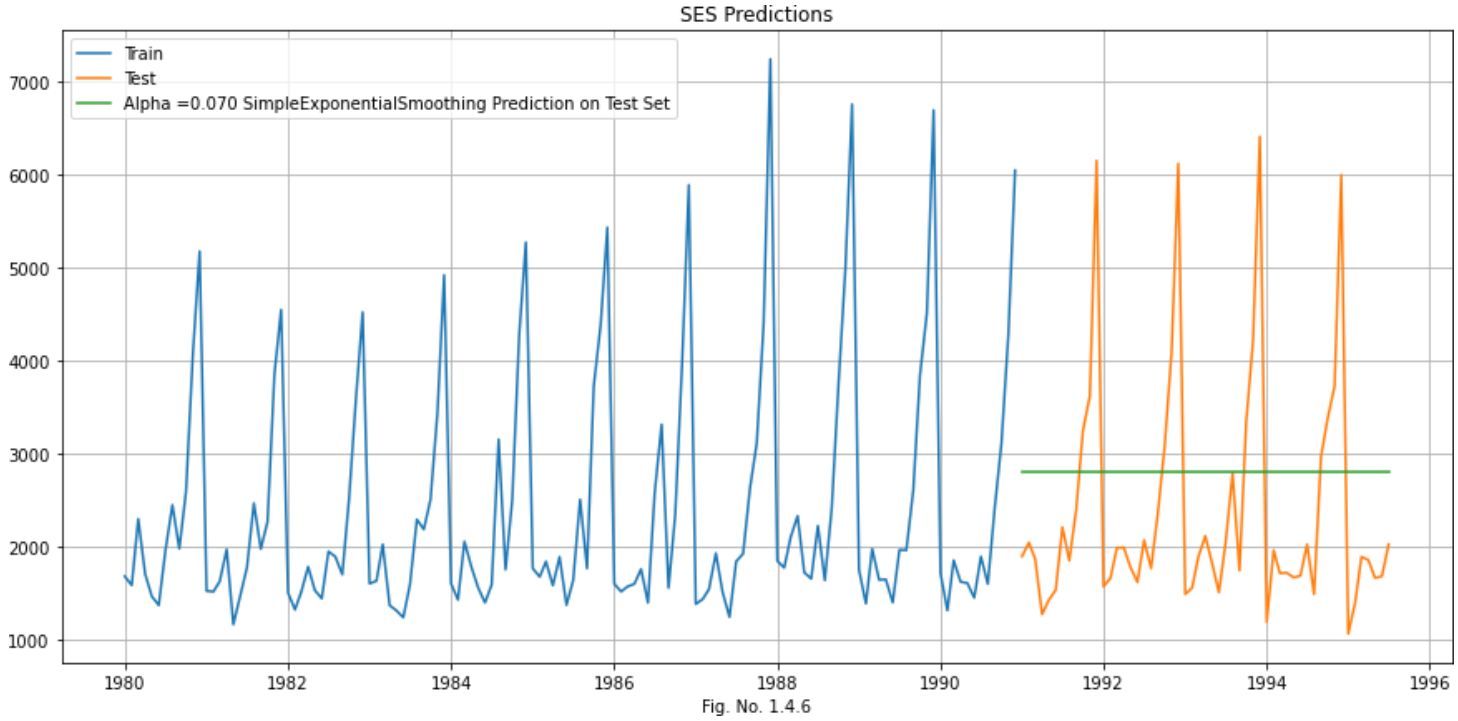


Fig. No. 1.4.6

Model Evaluation: -

For this model RSME would be,

	Test RMSE
Alpha=0.070 SimpleExponentialSmoothing	1338.008384

Root Mean Square Error of Simple Exponential SmoothingModel for Test Data is 1338.0083

Double Exponential Smoothing

Here we consider smoothing level and smoothing trend, after fitting our model we will get best parameter to analyzefurther,

```
{'smoothing_level': 0.6649999999999999,
 'smoothing_trend': 0.0001,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1502.1999999999991,
 'initial_trend': 74.87272727272739,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Best Parameters for Double Exponential Smoothing

```
1991-01-01    5401.733026
1991-02-01    5476.005230
1991-03-01    5550.277433
1991-04-01    5624.549637
1991-05-01    5698.821840
Freq: MS, dtype: float64
```

Top 5 Rows of Predicted Values on Test Data for Double Exponential Smoothing

After prediction on best parameters, we will plot the data for better understanding,

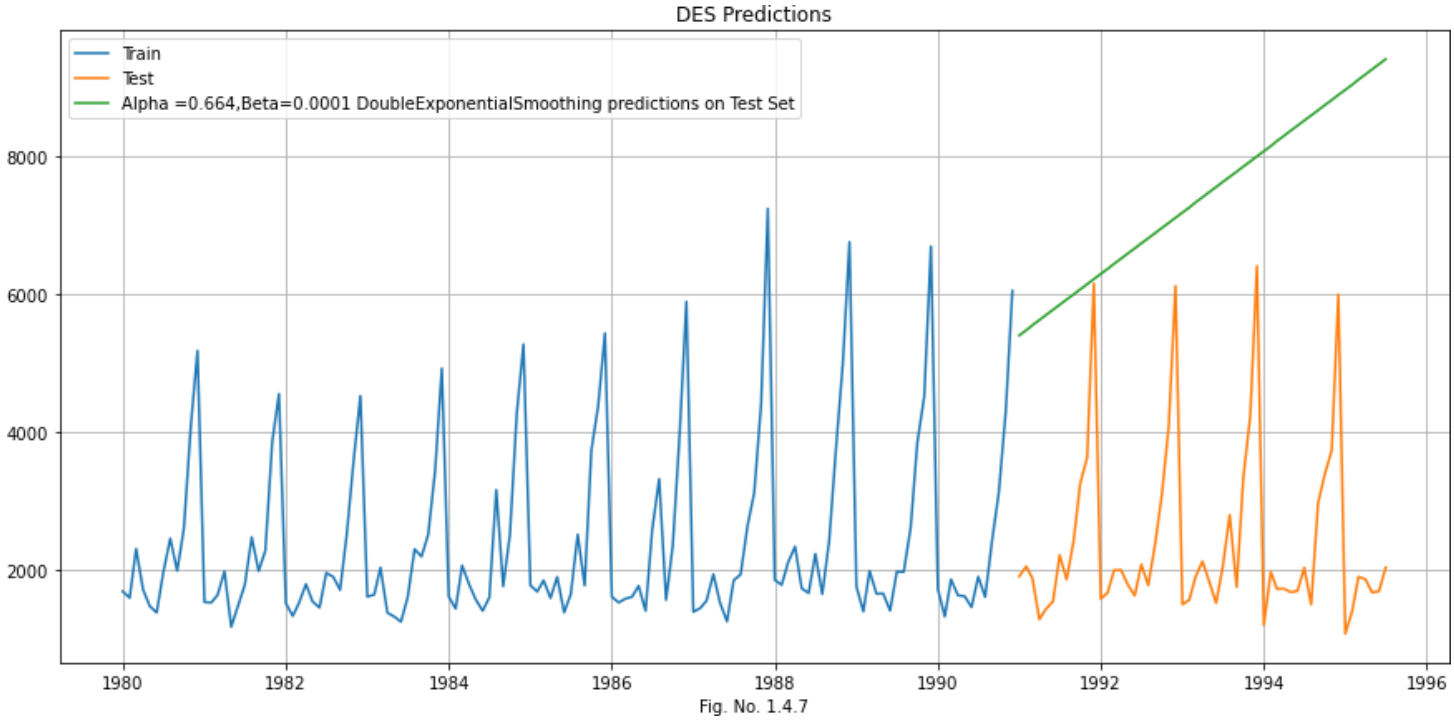


Fig. No. 1.4.7

Model Evaluation: -
For this model RSME would be,

	Test RMSE
Alpha=0.070 SimpleExponentialSmoothing	1338.008384
Alpha=0.664,B=0.0001 DoubleExponentialSmoothing	5291.879833

Root Mean Square Error of Double Exponential Smoothing for Test Data is 5291.8798, it is clearly seenthat RSME is not improved for this model.

Triple Exponential Smoothing

Here we consider smoothing level, smoothing trend and as well as smoothing seasonality, after fitting our model we will get best parameter to analyze further,

```
{'smoothing_level': 0.11133818361298699,
 'smoothing_trend': 0.049505131019509915,
 'smoothing_seasonal': 0.3620795793580111,
 'damping_trend': nan,
 'initial_level': 2356.4967888704355,
 'initial_trend': -10.187944726007238,
 'initial_seasons': array([0.71296382, 0.68242226, 0.90755008, 0.80515228, 0.65597218,
        0.65414505, 0.88617935, 1.13345121, 0.92046306, 1.21337874,
        1.87340336, 2.37811768]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Best Parameters for Triple Exponential Smoothing

After prediction on best parameters, we will plot the data for better understanding,

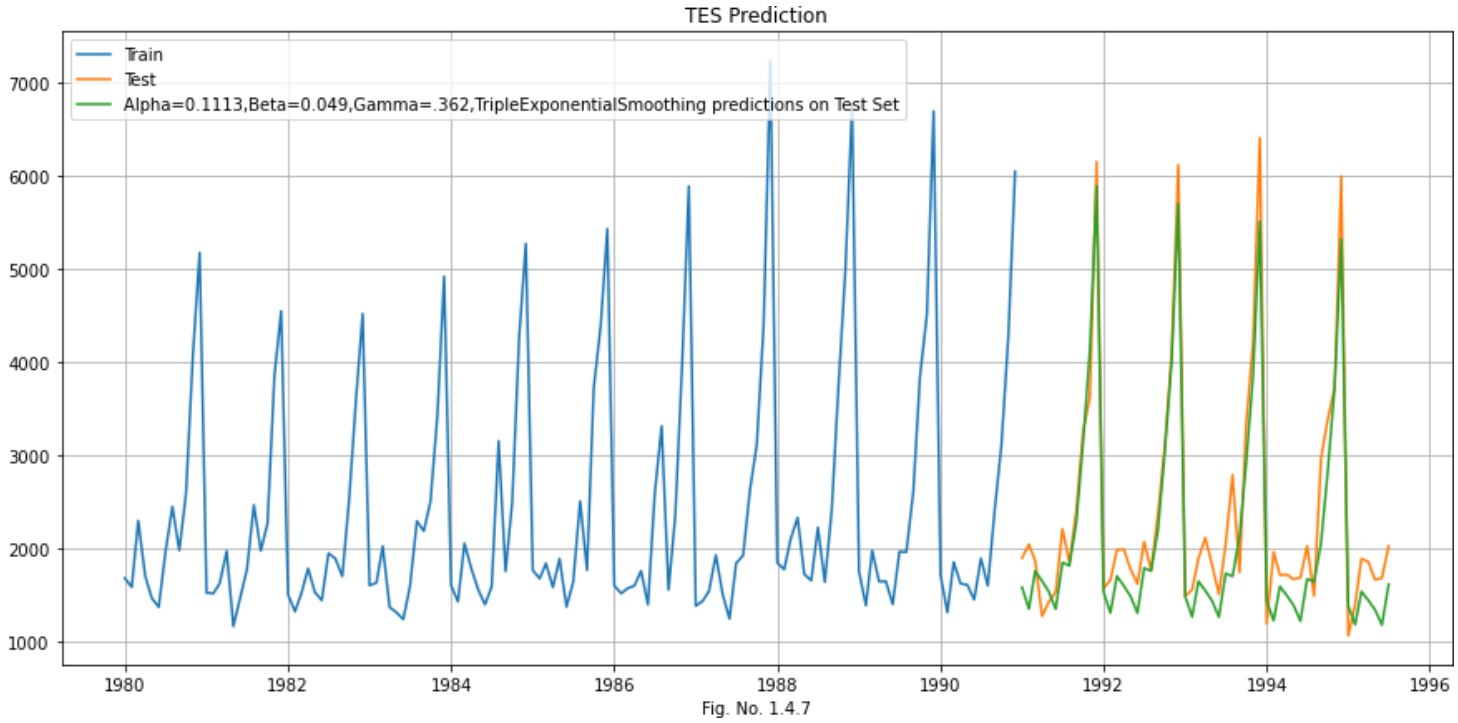


Fig. No. 1.4.7

Model Evaluation: -
For this model RSME would be,

	Test RMSE
Alpha=0.070 SimpleExponentialSmoothing	1338.008384
Alpha=0.664,B=0.0001 DoubleExponentialSmoothing	5291.879833
Alpha=0.1113,B=0.048,Y=.362 TripleExponentialSmoothing	404.286809

Root Mean Square Error of Triple Exponential Smoothing for Test Data is 404.2868.For better understanding we will visualize all three smoothing model along with train and test dataset,

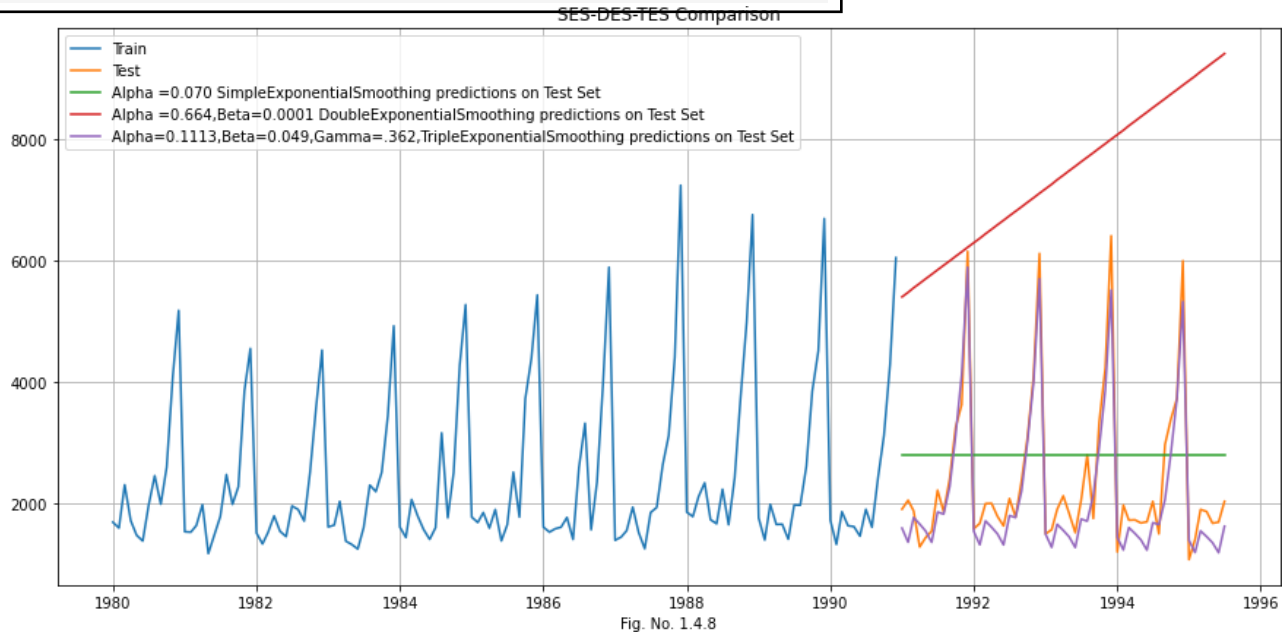


Fig. No. 1.4.8

We are end with all model like Linear Regression, Naïve Model, Simple Average, Moving Average and all Exponential smoothing models now compare all RSME together and stat which would be the best model for the prediction,

	Test RMSE
RegressionOnTime	1389.135175
NaiveOnTime	3864.279352
SimpleAverage	1275.081804
2 Point Trailing on Test Data	813.400684
4 Point Trailing on Test Data	1156.589694
6 Point Trailing on Test Data	1283.927428
9 Point Trailing on Test Data	1346.278315
Alpha=0.070 SimpleExponentialSmoothing	1338.008384
Alpha=0.664,B=0.0001 DoubleExponentialSmoothing	5291.879833
Alpha=0.1113,B=0.048,Y=.362 TripleExponentialSmoothing	404.286809

Table No. 1.4.1

Conclusion: -

It can be concluded that Triple Exponential Smoothing is best model for forecasting with Alpha=0.1113, Beta=0.048 and Gama=0.362 having RSME Value=404.2868.

1. 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 a unit root test which determines whether there is a unit root and subsequently whetherthe series is non-stationary.

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

- $H0$: The Time Series has a unit root and is thus non-stationary.
- $H1$: 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 i.e., 0.05.

If we found p-value greater than α value than we don't have enough evidence to reject the null hypothesis and its stats thatdata is non stationary.
After applying Augmented Dickey-Fuller we found certain results i.e.,

```
Sparkling Data test statistic is -1.650
Sparkling Data test p-value is 0.7721785247132078
Number of lags used 11
Number of Observation Used 175
Critical Values {'1%': -4.011455293061225, '5%': -3.4358815193469385, '10%': -3.141957196268222}
```

We observed p-value for the dataset is **0.7721** which is greater than significance value (0.05) so in that case we can say that our dataset is non stationary.
To make dataset stationary we will take first order differencing and apply Dickey-Fuller test again,
After applying Dickey-Fuller again result we got,

```
Sparkling Data test statistic is -44.912
Sparkling Data test p-value is 0.0
Number of lags used 10
Number of Observation Used 175
Critical Values {'1%': -4.011455293061225, '5%': -3.4358815193469385, '10%': -3.141957196268222}
```

Here p-value is less than the level of significance hence we can say that our dataset become stationary after first order differencing.

YearMonth	Sparkling	Differencing
1980-01-01	1686	NaN
1980-02-01	1591	-95.0
1980-03-01	2304	713.0
1980-04-01	1712	-592.0
1980-05-01	1471	-241.0

We can see new dataset here after first order differencing,
Here is a plot of stationary dataset,

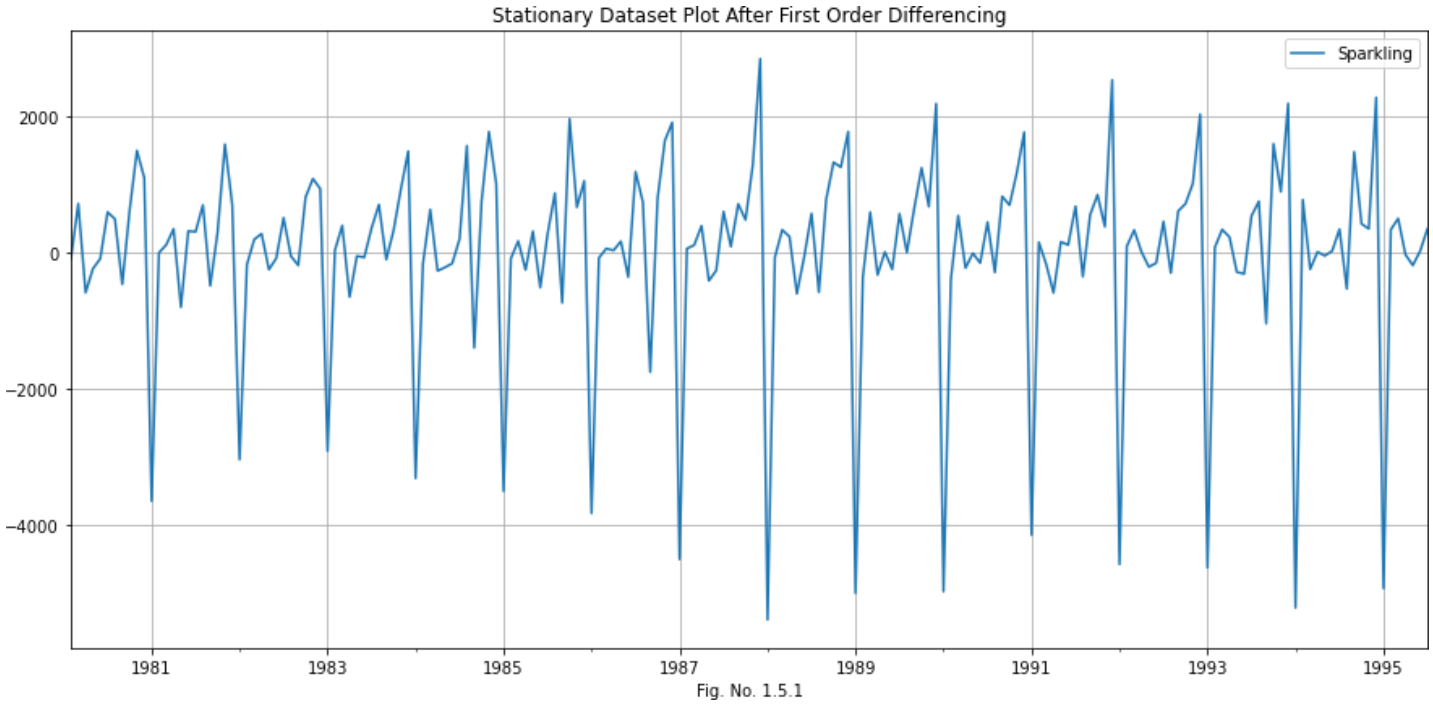


Fig. No. 1.5.1

1.6 : - 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.

For building ARIMA/SARIMA model on train dataset first we have to check stationarity, so after applying dickey-fuller on train dataset parameter we got,

```
Train Dataset test statistic is -2.062
Train Dataset test p-value is 0.5674110388593658
Number of lags used 12
```

The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationaries the Time Series.

Apply dickey-fuller after first level differencing on test data,

```
Train Dataset test statistic is -7.968
Train Dataset test p-value is 8.479210655515133e-11
Number of lags used 11
```

Now, let us go ahead and plot the differenced training data.

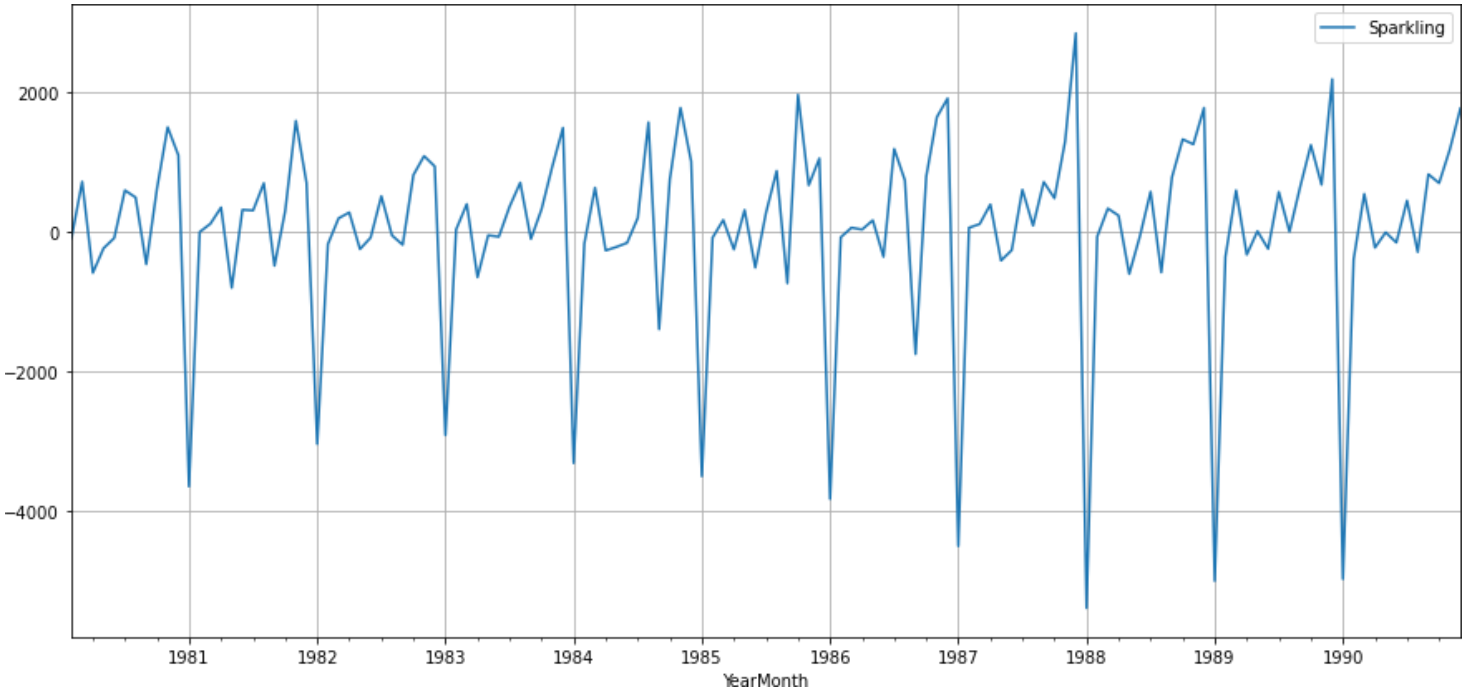


Fig No. 1.6.1

Automated Version of the ARIMA Model

For automated version of ARIMA model we fixed ranges for defined parameter that is p, d, q.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACFplot = range (0,4), i.e., (0,1,2,3)
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot =range (0,4), i.e., (0,1,2,3)
- The differencing parameter in an ARIMA model is ‘d’ which comes from making dataset stationary=range (1,2), i.e.,1.

After applying itertools to make different combination and fetch Akaike Information Criteria (AIC) for train data,

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.458954
14	(3, 1, 2)	2230.952333
11	(2, 1, 3)	2232.937076
9	(2, 1, 1)	2233.777626

ARIMA Model with p=2, d=1, q=2 having the lowest AIC value i.e., **2213.509.**
Let's check the summary of the model with these parameters.

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Sparkling   No. Observations:                  132
Model:                     ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                    Sun, 13 Nov 2022   AIC                           2213.509
Time:                            14:45:29   BIC                           2227.885
Sample:                        01-01-1980   HQIC                          2219.351
                             - 12-01-1990
Covariance Type:                      opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.781      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.218      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.46e+27. Standard errors may be unstable.
```

Fig No. 1.6.2

Diagnostics Plot: -

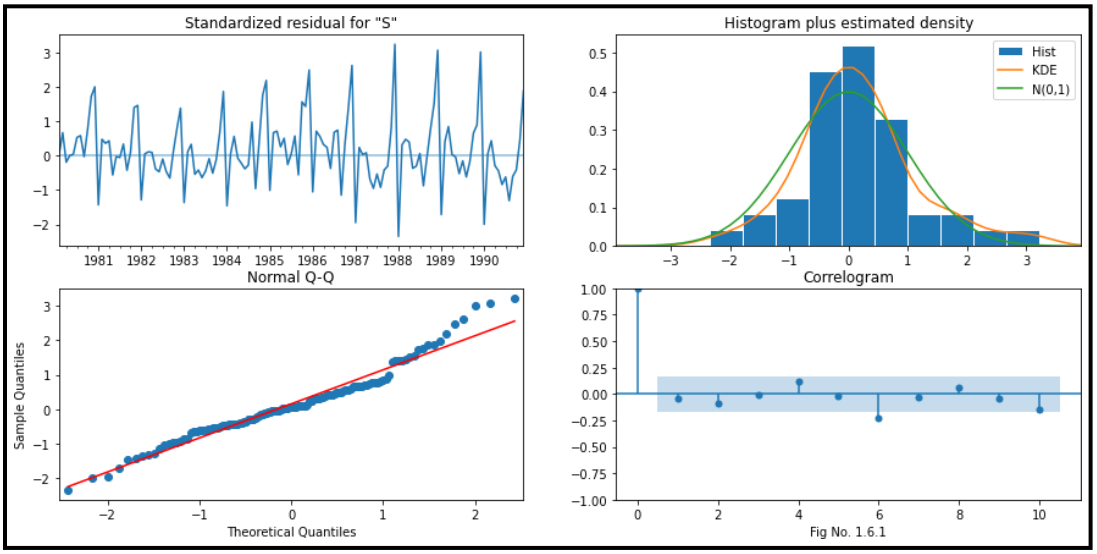


Fig No. 1.6.3

Model Evaluation: -

After getting lowest AIC value of train dataset, we will evaluate our model on test dataset,

	RMSE	MAPE
ARIMA(2,1,2)	1299.97964	47.099986

RMSE for automated ARIMA model is 1299.97 and MAPE is 47.099.

Automated Version of the SARIMA Model

For automated version of SARIMA model, we fixed ranges for defined parameter that is p, d, q and P, D, Q.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACFplot = range (0,3), i.e., (0,1,2)
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot =range (0,3), i.e., (0,1,2)
- The differencing parameter in an ARIMA model is 'd' which comes from making dataset stationary=range (1,2), i.e.,1.
- The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag before which the PACFplot = range (0,3), i.e., (0,1,2)
- The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag before the ACF plot =range (0,3), i.e., (0,1,2)
- The differencing parameter in a SARIMA model is 'D' which comes from making dataset stationary=range (0,1), i.e.,0.
- With seasonal factor=12

After applying itertools to make different combination and fetch Akaike Information Criteria (AIC) for train data,

	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)	1557.121564
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340403

SARIMA Model with p=1, d=1, q=2 and P=1, D=0, Q=2 and seasonality of 12having the lowest AIC value i.e., **1555.5842.**
Let's check the summary of the model with these parameters.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-769.967
Date:	Sun, 13 Nov 2022	AIC	1555.935
Time:	15:13:02	BIC	1577.090
Sample:	01-01-1980 - 12-01-1990	HQIC	1564.505
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6381	0.287	-2.226	0.026	-1.200	-0.076
ma.L1	-0.3049	0.185	-1.645	0.100	-0.668	0.058
ma.L2	-0.8914	0.275	-3.246	0.001	-1.430	-0.353
ar.S.L12	0.7612	0.567	1.343	0.179	-0.350	1.872
ar.S.L24	0.2951	0.590	0.500	0.617	-0.861	1.451
ma.S.L12	1.8837	3.336	0.565	0.572	-4.655	8.422
ma.S.L24	-1.8034	2.474	-0.729	0.466	-6.652	3.045
sigma2	1.857e+04	4.87e+04	0.382	0.703	-7.68e+04	1.14e+05

Ljung-Box (L1) (Q):	0.08	Jarque-Bera (JB):	12.54
Prob(Q):	0.78	Prob(JB):	0.00
Heteroskedasticity (H):	1.55	Skew:	0.35
Prob(H) (two-sided):	0.20	Kurtosis:	4.55

Fig No. 1.6.4

Diagnostics Plot: -

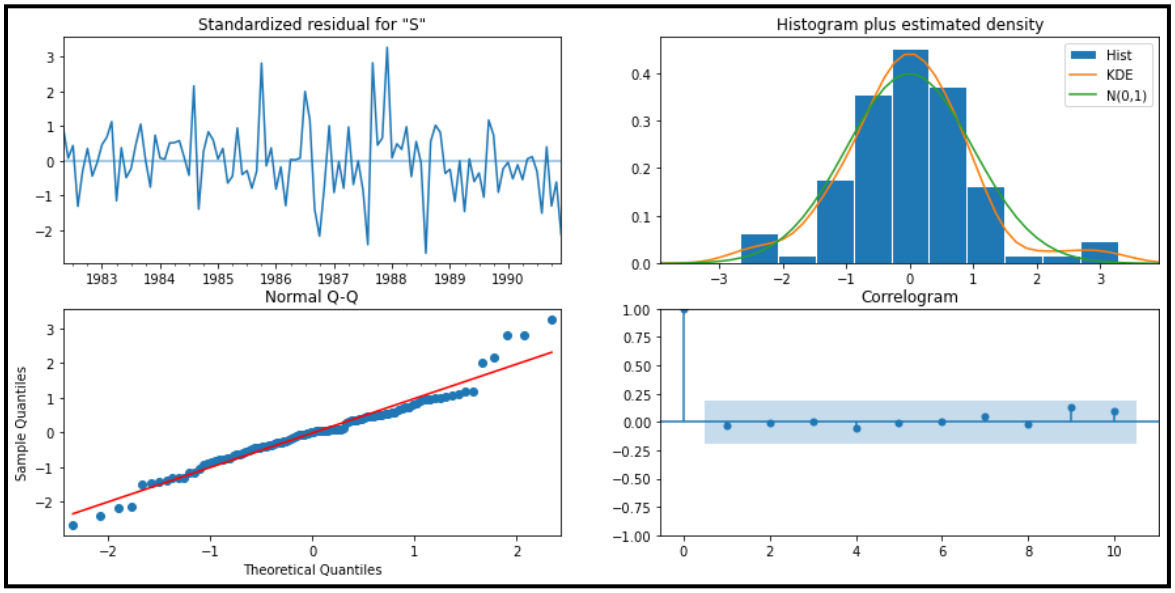


Fig No. 1.6.5

Model Evaluation: -

After getting lowest AIC value of train dataset, we will evaluate our model on test dataset,

	RMSE	MAPE
SARIMA(1,1,2)(2,0,2,12)	546.559779	21.950344

RMSE for automated SARIMA model is 546.5597 and MAPE is 21.9503

1.7 : - Build a table (create a data frame) with all the models built along with their correspondingparameters and the respective RMSE values on the test data.

	Test RMSE
RegressionOnTime	1389.135175
NaiveOnTime	3864.279352
SimpleAverage	1275.081804
2 Point Trailing on Test Data	813.400684
4 Point Trailing on Test Data	1156.589694
6 Point Trailing on Test Data	1283.927428
9 Point Trailing on Test Data	1346.278315
Alpha=0.070 SimpleExponentialSmoothing	1338.008384
Alpha=0.664,B=0.0001 DoubleExponentialSmoothing	5291.879833
Alpha=0.1113,B=0.048,Y=.362 TripleExponentialSmoothing	404.286809

	RMSE	MAPE
ARIMA Auto(2,1,2)	1299.979640	47.099986
ARIMA Manual(3,1,2)	1286.234646	45.206901
SARIMA Auto(1,1,2)(2,0,2,12)	546.559779	21.950344
SARIMA Manual(3,1,2)(1,0,1,12)	609.167441	26.281391

Table No. 1.8.1

1.8 : - Based on the model-building exercise, build the most optimum model(s) on the complete data andpredict 12 months into the future with appropriate confidence intervals/bands.

Based on model building exercise the best model was SARIMA Automated model, now imposing the same parameters on complete dataset.
After imposing the most optimum model on the complete dataset, summary would be look like,

Fig No. 1.9.1

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-1172.665
Date:	Sun, 13 Nov 2022	AIC	2361.330
Time:	12:21:08	BIC	2385.881
Sample:	01-01-1980 - 07-01-1995	HQIC	2371.300
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6528	0.269	-2.427	0.015	-1.180	-0.126
ma.L1	-0.2799	0.222	-1.258	0.208	-0.716	0.156
ma.L2	-0.8056	0.251	-3.205	0.001	-1.298	-0.313
ar.S.L12	0.7076	0.587	1.206	0.228	-0.442	1.857
ar.S.L24	0.3130	0.598	0.523	0.601	-0.860	1.486
ma.S.L12	-1.2498	0.660	-1.893	0.058	-2.544	0.044
ma.S.L24	-0.5398	0.924	-0.584	0.559	-2.351	1.272
sigma2	5.229e+04	1.7e+04	3.079	0.002	1.9e+04	8.56e+04

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	27.55
Prob(Q):	0.94	Prob(JB):	0.00
Heteroskedasticity (H):	1.00	Skew:	0.51
Prob(H) (two-sided):	1.00	Kurtosis:	4.77

Diagnostics Plot: -

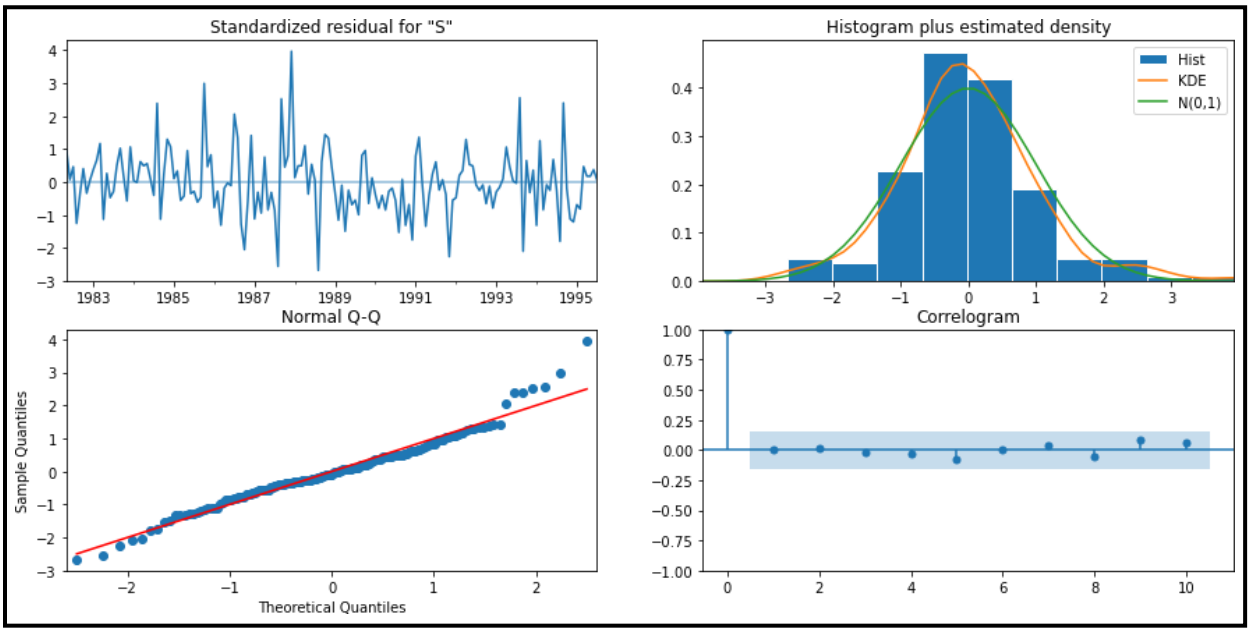


Fig No. 1.9.2

Predicting 12 months into the future with appropriate confidence intervals/bands, and dataset for next forecasted 12 months would be,

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1858.649008	381.027638	1111.848560	2605.449455
1995-09-01	2456.724202	385.980057	1700.217192	3213.231213
1995-10-01	3317.869146	386.088840	2561.148924	4074.589368
1995-11-01	4018.303130	387.862601	3258.106401	4778.499859
1995-12-01	6290.638030	387.923503	5530.321936	7050.954124
1996-01-01	1236.206702	388.832735	474.108545	1998.304858
1996-02-01	1547.889337	389.074601	785.317130	2310.461543
1996-03-01	1813.160185	389.702411	1049.357494	2576.962876
1996-04-01	1787.672935	390.056987	1023.175290	2552.170581
1996-05-01	1628.564435	390.580115	863.041476	2394.087394
1996-06-01	1564.556621	390.988755	798.232743	2330.880498
1996-07-01	1997.996568	391.469601	1230.730249	2765.262887

Table No. 1.9.1

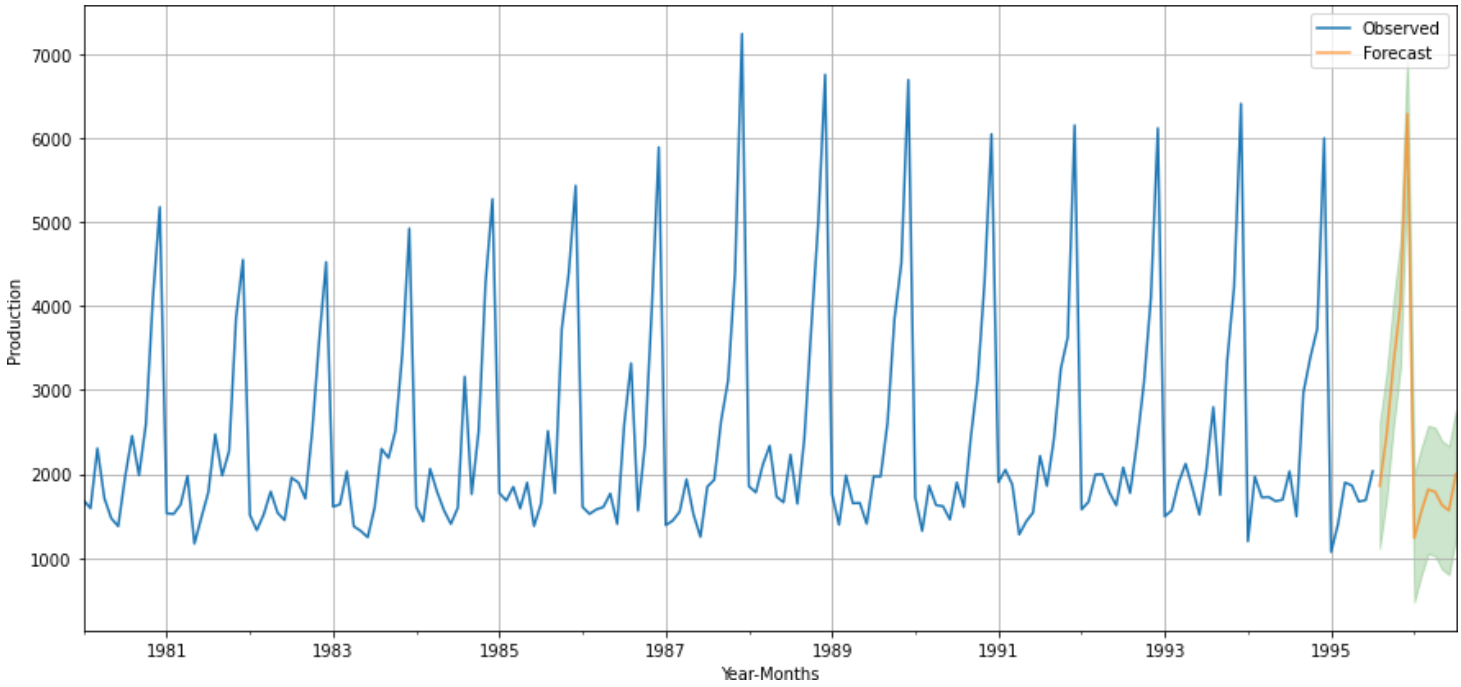


Fig No. 1.9.3

Here is the dataset and plot for next 12 month forecasted values with appropriate interval band.

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

- Sparkling wine production was high in winters i.e., in last quarter and less in summers so company should take step up the production in summers.
- Triple Exponential Smoothing is the best forecasting model or this dataset.
- 2nd best forecasting model for this dataset is SARIMA Automated model.
- There is seasonality in the dataset but trend is not observed.
- KDE plot is almost same for all model.