TIME SERIES FORECASTING PROJECT1

Submitted by,

Jiya Jacob

PGP-DSBA ONLINE

JULY-B 2021

DATE:20/02/2022

ROSE WINE

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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

INTRODUCTION

The purpose of this whole exercise is to perform various smoothing techniques and forecasting techniques on the given data, combine all predictions and eventually perform predictions based on past data with given trend and seasonality. The smoothing techniques are the members of time series forecasting methods or algorithms, which use the weighted average of a past observation to predict the future values or forecast the new value. These techniques are well suited for time-series data having fewer deviations with time.

DATA DICTIONARY

1. YearMonth: The time period of a certain amount of sales.
2. Rose: The amount of sales of the wine during a particular period of time.

Q1. **Read the data as an appropriate Time Series data and plot the data.**

a. Head of the Time series data

	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

Table 1: Head of the time series data

b. Tail of the Time series data

	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

Table 2: Tail of the time series data

From the above head and tail samples of data, we can see that the data has not been identified as time series. Therefore, we need to create a time index for this data. We do so by creating time stamp which starts at 01/01/1985 to 31/07/1995.Then we set this time stamp as index. Since the newly created time stamp and the column YearMonth has the same time periods, we can remove the YearMonth column which is irrelevant, thereby reaching to our final data.

c. Creating the time stamps and adding to data frame

	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

Table 3: Head of the data after creating the time stamp

d. Final data frame

	Rose
Time_Stamp	
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

Table 4: Head of the final data frame

e. Time series data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    object
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
```

From the above table we can see that there are 187 rows for this time series data, out of which one of the columns, Rose is having two null values, which has to be imputed. The data type of the Rose column is float64.

The Rose wine sales of 15 years ranging from 01/01/1980 to 31/07/1995 is recorded in the given data.

f. Checking for null values in the data.

```
YearMonth    0
Rose         2
dtype: int64
```

As mentioned above, from the above table, we can see that the column Rose is having 2 null values that has to be imputed.

g. Graphical representation of data with missing values

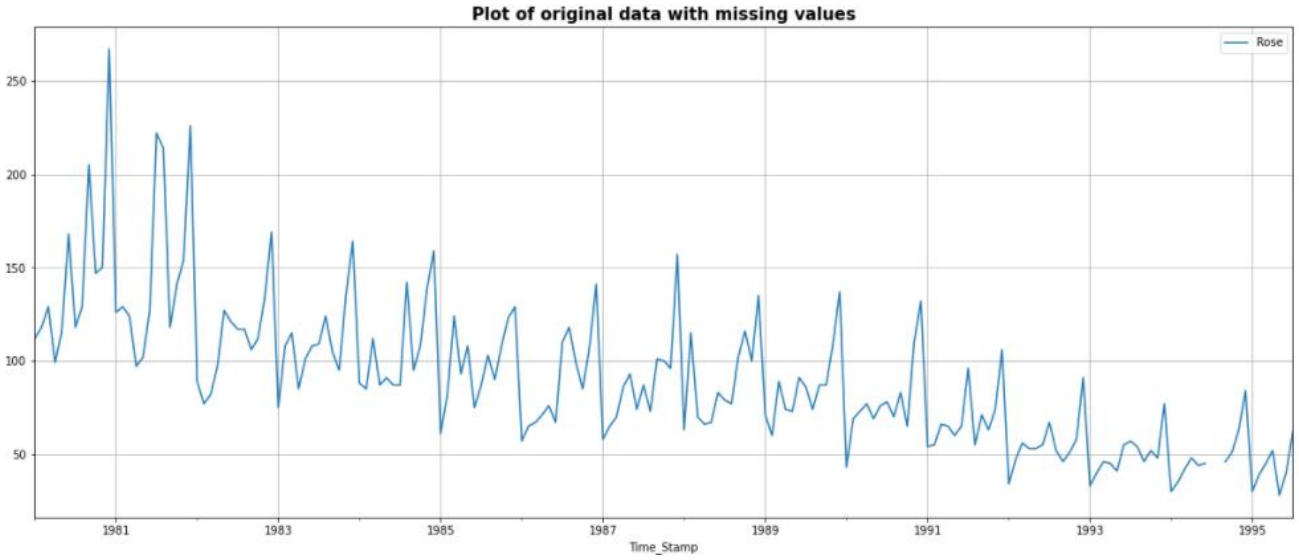


Fig 1: Graphical representation of data with missing values

From the above graph, we can see that there exists a pattern withing each year, thereby indicating a seasonal effect, that might be present. A strong downward trend is present in the above shown graph. Since we have considered only one variable (Rose column) for the plotting of the above graph, this is called the univariate time series analysis or forecasting.

Q2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

a. Treating the missing values

The missing values in the time series data frame are treated using the interpolate function.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    187 non-null    float64
dtypes: float64(1)
```

From the above table we can see that the null values in the column 'Rose' have been imputed and now the 'Rose' column is free of null values.

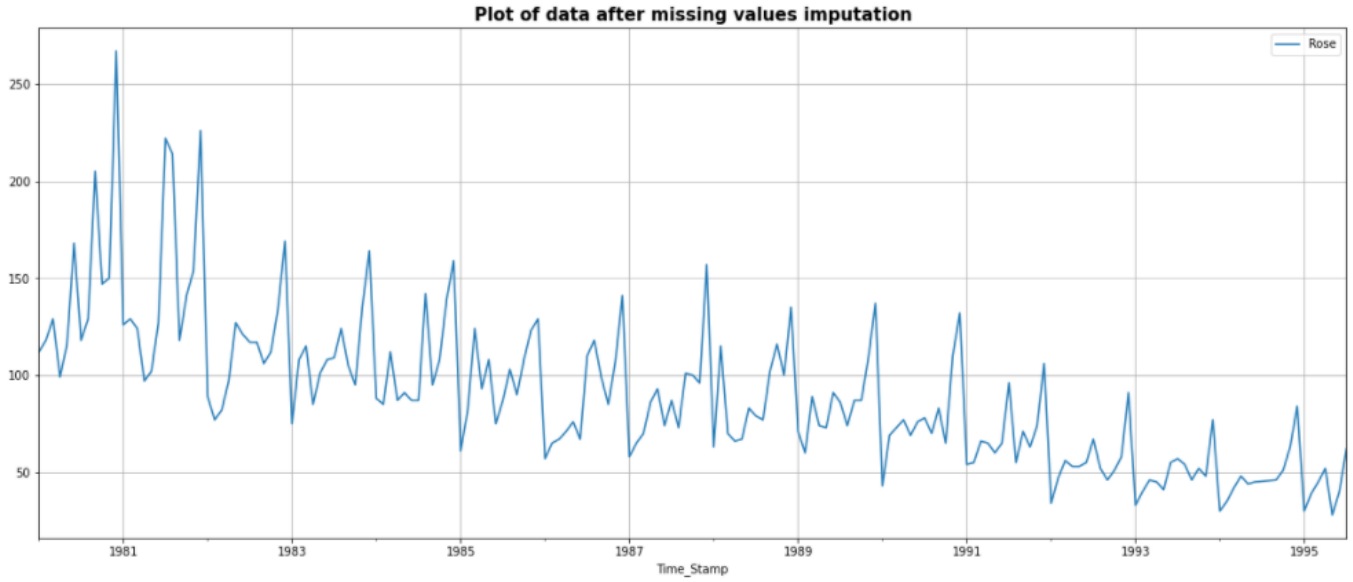


Fig 2: Graphical representation of data after missing values imputation

b. Five-point summary data

	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

From the above table we see that mean of the Rose wine sales in 20 the century is 90.39 where, the minimum sales 28 and maximum sales is 267. The median of the Rose wine sales is 86 and the standard deviation is 39.175

c. Exploratory Data Analysis

1. Univariate Time series

A univariate time series is a series with a single time-stamped variable at time 't'. Here the dataset belongs to the Rose wine sales from the January of 1980 to July of 1995. Here, Rose is the time-dependent variable. The series is a monthly series, wherein for each month between Jan-1980 and Jul-1995 a datapoint is recorded.

2. Empirical Cumulative Distribution Function (ECDF)

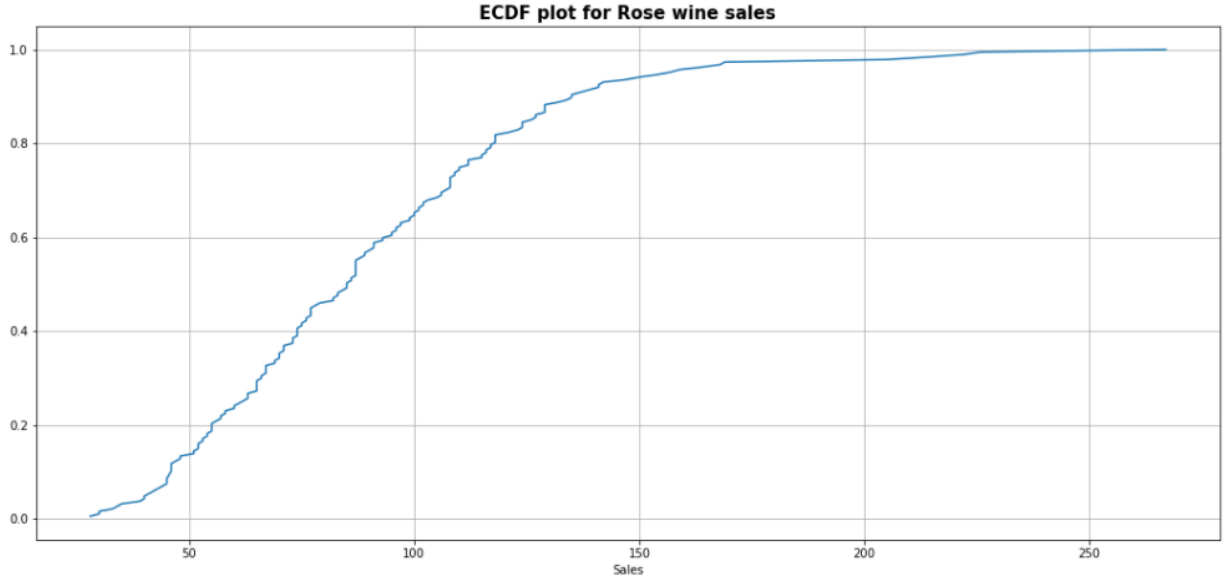
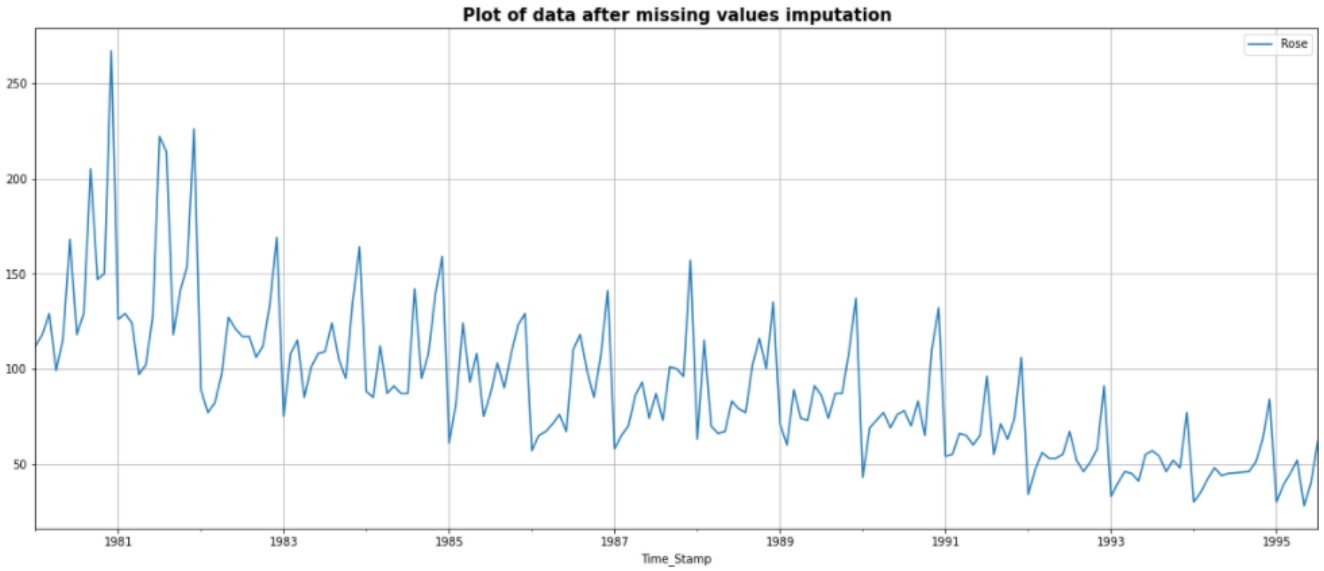


Fig 3: ECDF plot for rose wine sales

An ECDF is an estimator of the Cumulative Distribution Function. The ECDF essentially allows you to plot a feature of the data in order from the least to greatest and see the whole feature as if it is distributed across the data set. From the five-point summary statistics, we can see that the data ranges from 28 to267.

3. Plot for the Rose data after missing values imputation



The graph above shows the plot of the Rose wine data after missing values imputation. There exists a pattern within each year indicating a seasonal effect. This seasonality will help us in forecasting the future sales based on past values and the seasonality and the trend present in the data.

4. Year wise Rose wine sales assessment using Box plot

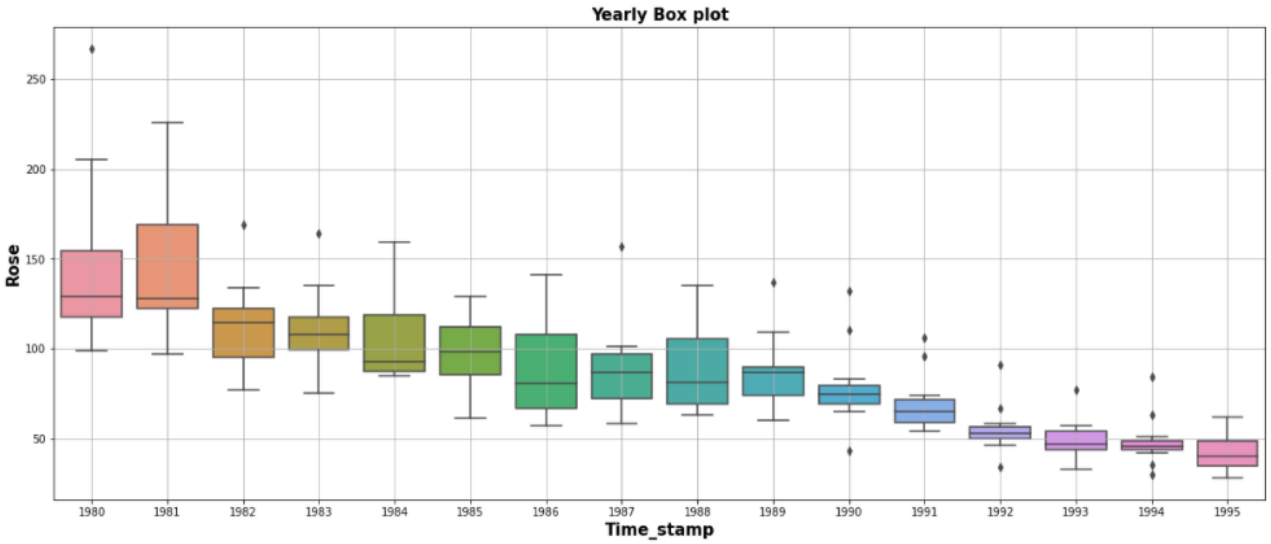


Fig 4: Yearly box plot

From the above figure we see that a clear downward trend in the sales of the Rose wine over the years. The sales during the initial years were higher. For most of the years, outliers are present. The sales for the year 1995 would be obviously low, since only the 7 month sales out of 12 months were recorded in the time series data given.

5. Monthly wise Rose wine sales assessment using box plot

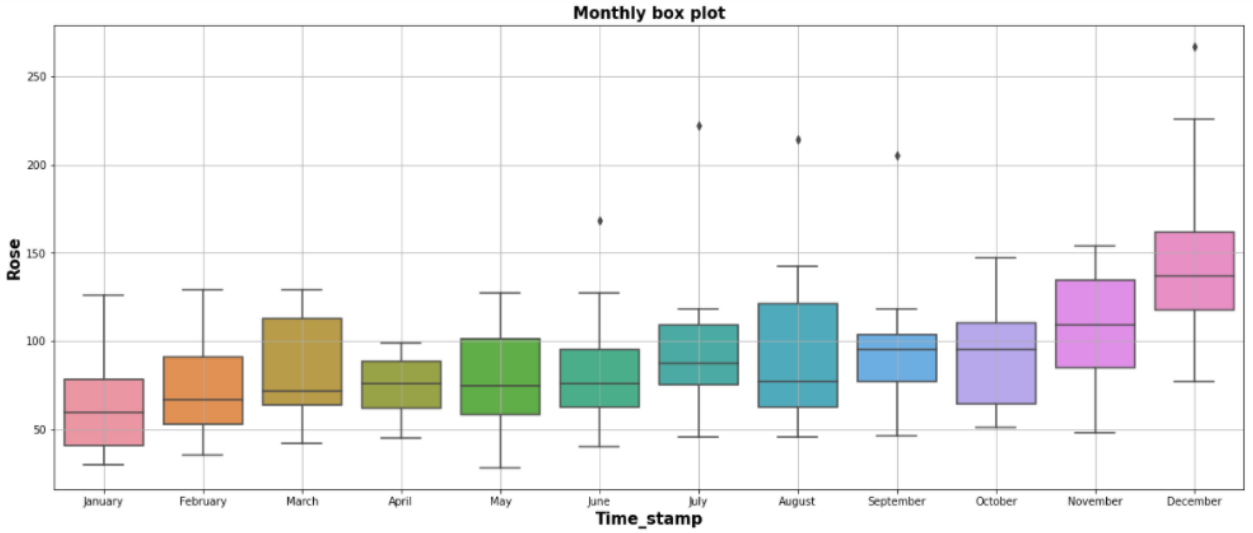


Fig 5: Monthly box plot

From the above graph, we can see that the sales of wine were highest during the month of December followed by November and October respectively. This can be due to the possible holidays and festivals happening in those months. The lowest sales of the wine happen in the month of June. The months of June, July, August, September and December has outliers.

6. Year/Month table

Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
1980	99.0	129.000000	267.0	118.0	112.0	118.000000	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.000000	226.0	129.0	126.0	222.000000	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.000000	169.0	77.0	89.0	117.000000	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.000000	164.0	108.0	75.0	109.000000	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.000000	159.0	85.0	88.0	87.000000	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.000000	129.0	82.0	61.0	87.000000	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.000000	141.0	65.0	57.0	110.000000	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.000000	157.0	65.0	58.0	87.000000	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.000000	135.0	115.0	63.0	79.000000	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.000000	137.0	60.0	71.0	86.000000	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.000000	132.0	69.0	43.0	78.000000	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.000000	106.0	55.0	54.0	96.000000	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.000000	91.0	47.0	34.0	67.000000	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.000000	77.0	40.0	33.0	57.000000	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	45.666667	84.0	35.0	30.0	45.333333	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.000000	40.0	45.0	28.0	NaN	NaN	NaN

Table 5: Year/month table

From the above table we can observe that in the year 1995 after the month of July there are no entries as we were only given data until the seventh month of 1995.

7. Monthly plot for each month sales distribution

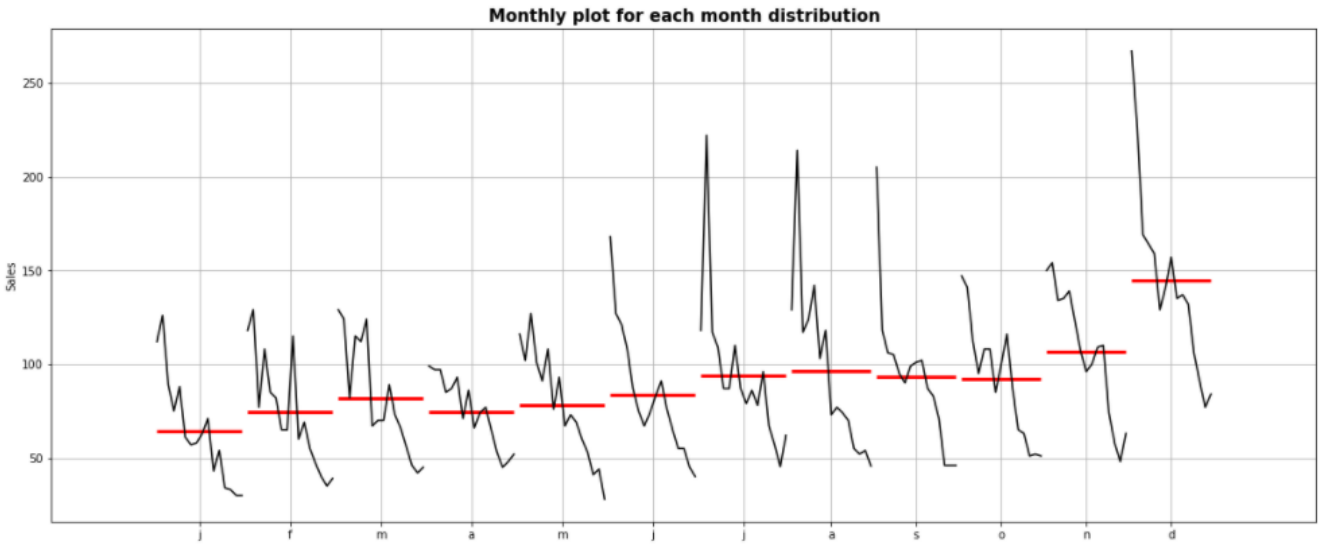


Fig 6: Monthly plot for each month distribution

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

8. Year wise Rose wine sales assessment using line plot

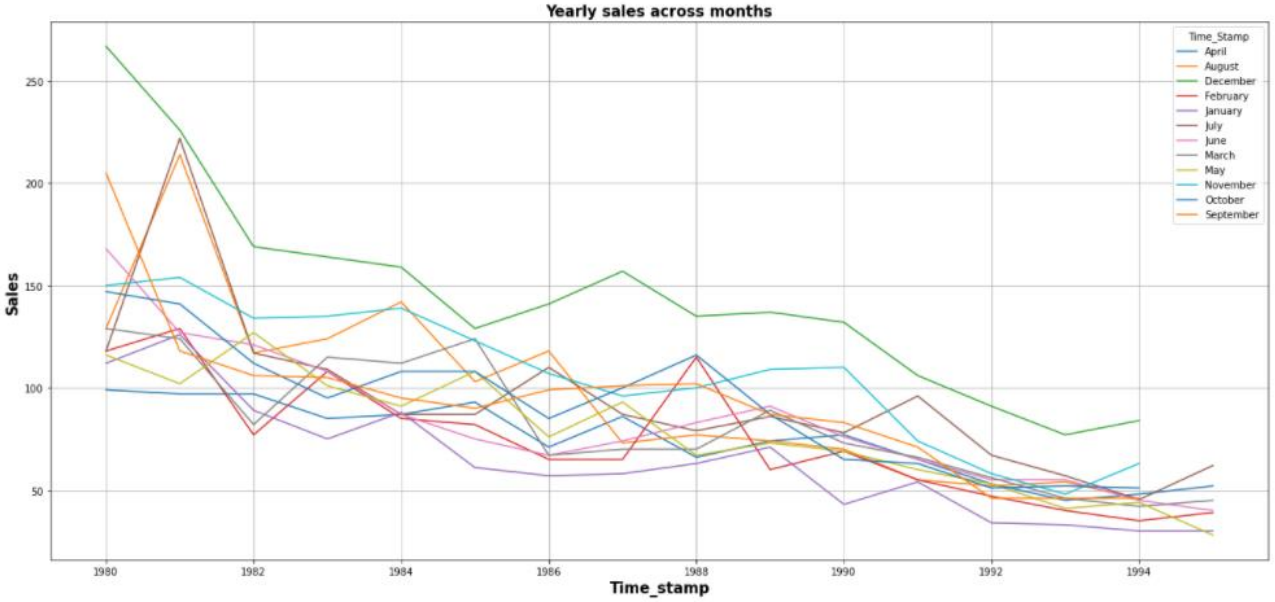


Fig 7: Yearly sales across months using line plot

In the above plot, we have done a line plot to compare the sales of Rose wine across each month over the years. From the above yearly line plot we can observe that the month of December outperforms all the other months in terms of sales, as we have seen in the boxplots above. We also observe that the month of January has the lowest number of sales.

9. Yearly sales table and plot of Rose wine

Time_Stamp	Rose
1980-12-31	1758.0
1981-12-31	1780.0
1982-12-31	1348.0
1983-12-31	1324.0
1984-12-31	1280.0
1985-12-31	1183.0
1986-12-31	1063.0
1987-12-31	1060.0
1988-12-31	1073.0
1989-12-31	1038.0
1990-12-31	945.0
1991-12-31	830.0
1992-12-31	663.0
1993-12-31	594.0
1994-12-31	579.0
1995-12-31	296.0

Table 6: Yearly sales table

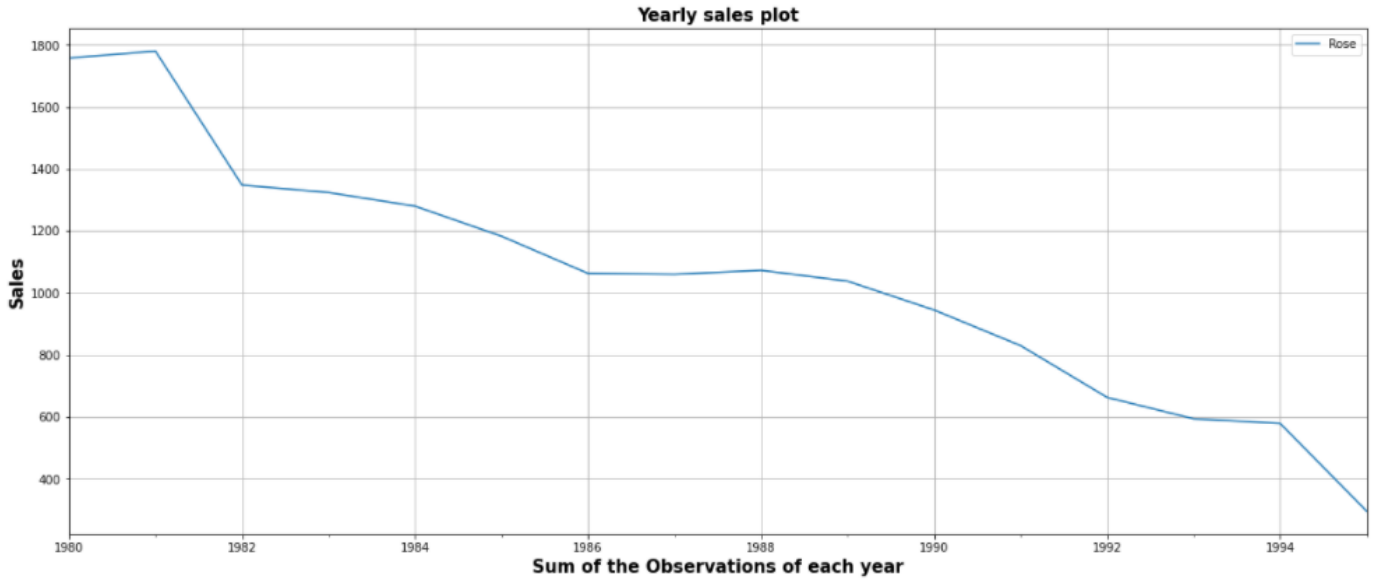


Fig 8: Yearly sales plot

From the above table and graph, we can observe that the year 1982 saw the biggest dip in sales in comparison to any years before or after. After 1982 we see a gradual decrease in sales over the years.

10. Quarterly sales table and plot of Rose wine

Time_Stamp	Rose
1980-03-31	119.666667
1980-06-30	127.666667
1980-09-30	150.666667
1980-12-31	188.000000
1981-03-31	126.333333
...	...
1994-09-30	45.666667
1994-12-31	66.000000
1995-03-31	38.000000
1995-06-30	40.000000
1995-09-30	62.000000

Table 7: Quarterly sales table

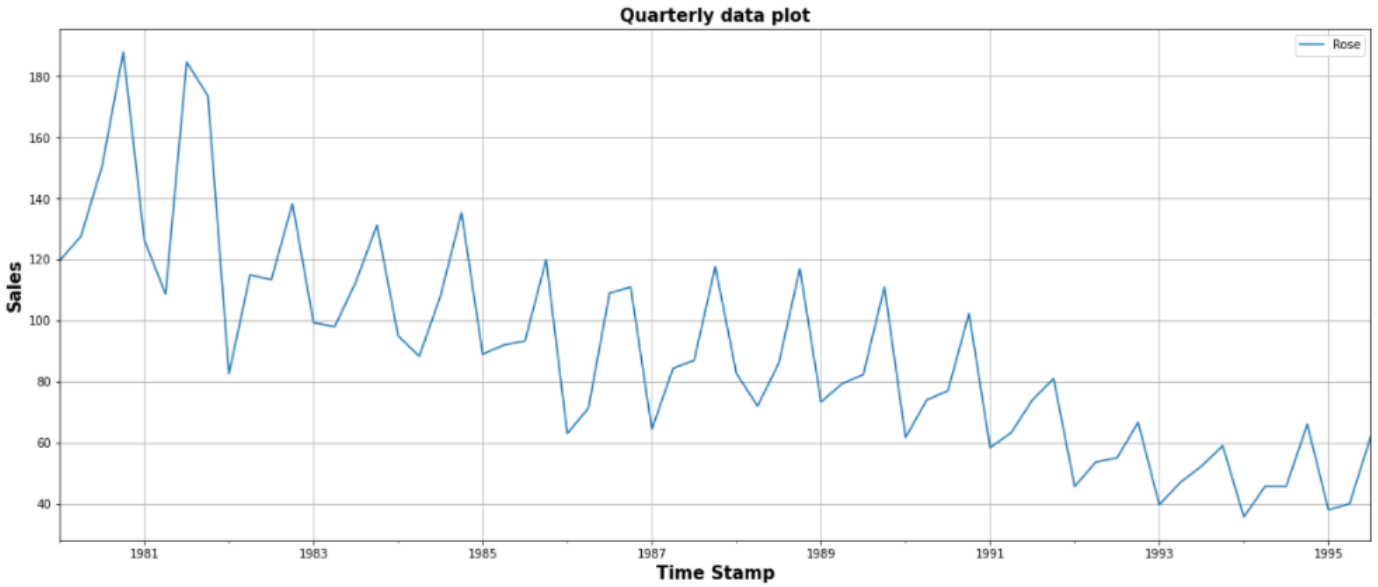


Fig 9: Quarterly sales plot

In the above table and graph, we have found out the quarterly sales of the Rose wine. The sales of 3 months make a quarter sale.

11. Daily sales table and plot of Rose wine

Time_Stamp	Rose
1980-12-31	267.0
1981-12-31	226.0
1981-07-31	222.0
1981-08-31	214.0
1980-09-30	205.0
...	...
1985-05-20	0.0
1985-05-19	0.0
1985-05-18	0.0
1985-05-17	0.0
1987-12-04	0.0

Table 8: Daily sales table

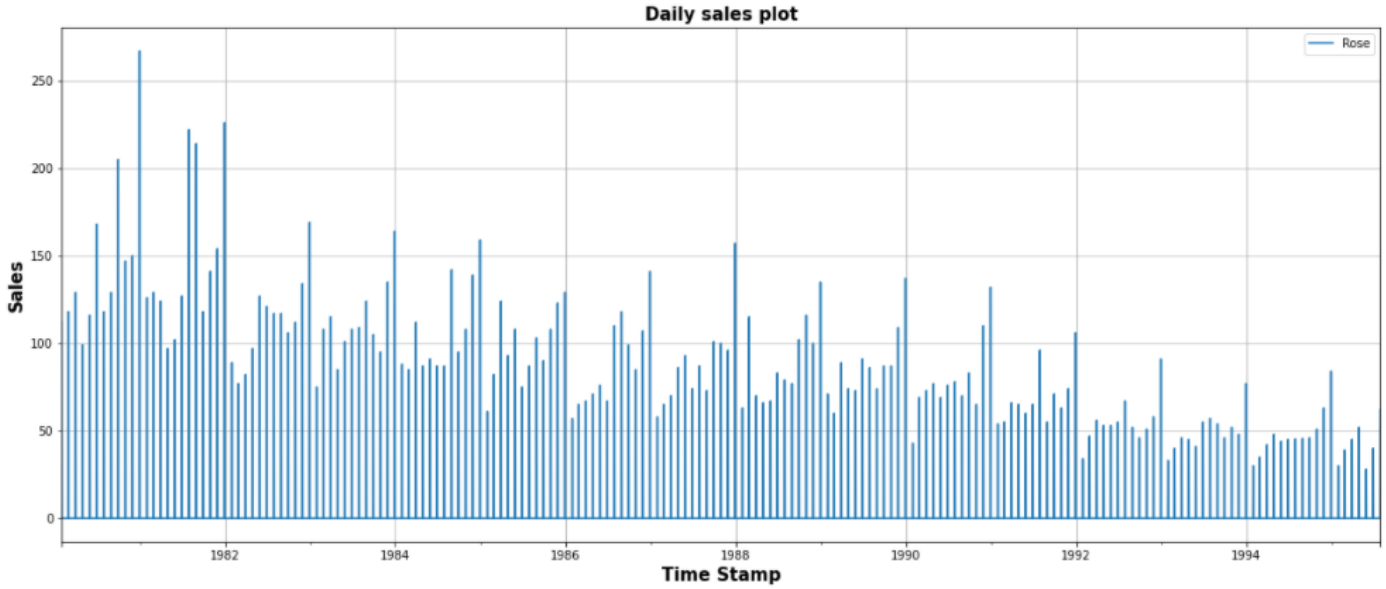


Fig 10: Daily sales plot

In the above table and graph, we have found out the daily sales of the Rose wine. We can observe that the highest number of sales was on the day 1980/12/31 and we see that there a lot of day without even a single unit sold.

D. Model Decomposition

Decomposition of a model is done:

- To understand revenue generation without the quarterly effects
  - –De-seasonalize the series
  - –Estimate and adjust by seasonality
- To compare the long-term movement of the series (Trend) vis-a-vis short-term movement (seasonality) to understand which has the higher influence
- If revenue for multiple sectors is to be compared and if the sectors show non-uniform seasonality, de-seasonalized series needs to be compared.

Mainly there are two kinds of decomposition done, one is additive and the other one is multiplicative. In simple terms, we can identify the additive or multiplicative time series by looking into the magnitude of the seasonal component. If the magnitude of the seasonal component changes with time, then the series is multiplicative. Otherwise, the series is additive.

1. Additive decomposition

- An additive model suggests that the components are added together.
- An additive model is linear where changes over time are consistently made by the same amount. The seasonal correction is added with the trend.
- A linear seasonality has the same frequency (width of the cycles) and amplitude (height of the cycles)

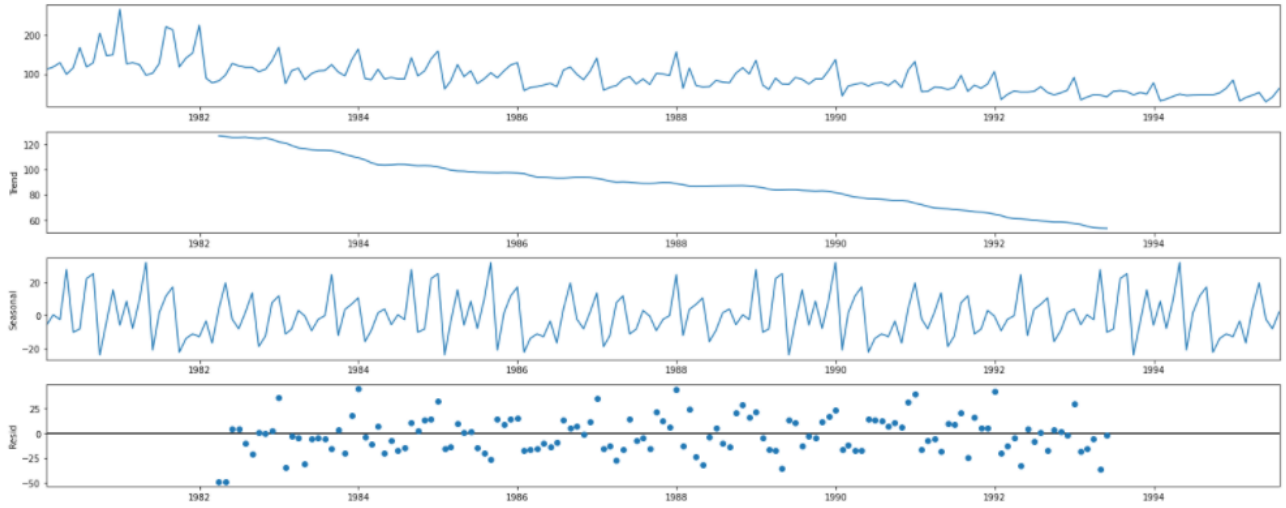


Fig 11: Additive decomposition of the data

From the above given 4 series, we can see that the trend and seasonality of the time series are clearly separated using the additive decomposition. The first series represents the given time series data, the second series represents the trend, the third one seasonality and the fourth one the residuals. It is noted that the residuals plotted in the above series does not follow a recognized pattern which ensures that the decomposition was accurate enough.

2. Multiplicative decomposition

- A multiplicative model suggests that the components are multiplied together
- A multiplicative model is non-linear.
- The seasonal correction is multiplied with the trend.
- A non-linear seasonality has an increasing or decreasing frequency (width of the cycles) and / or amplitude (height of the cycles) over time

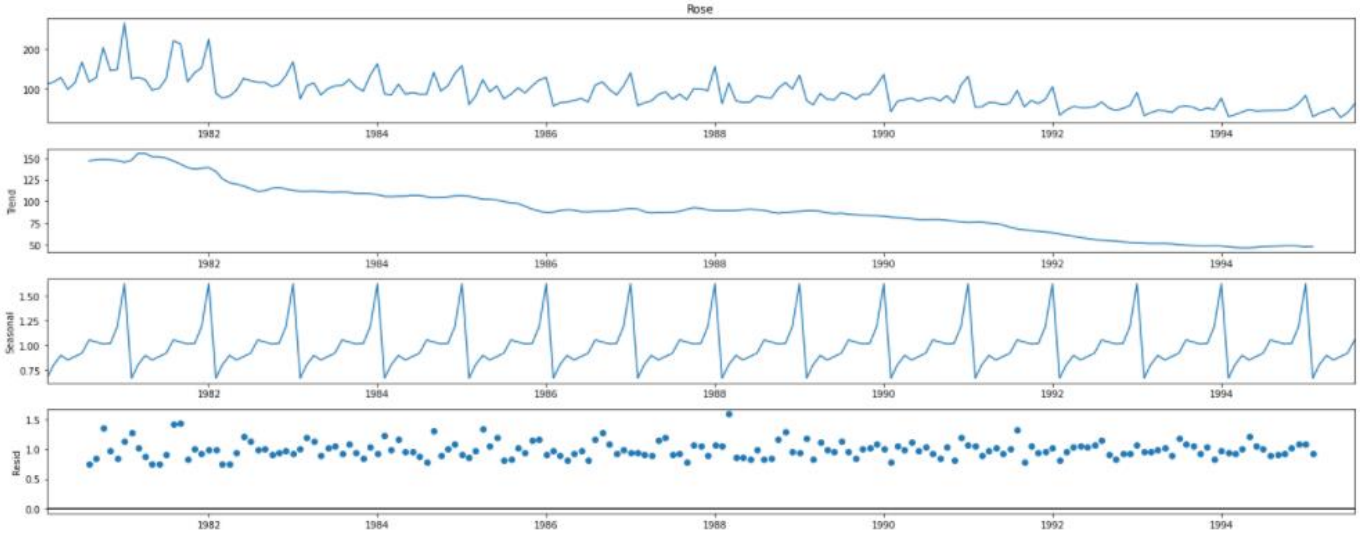


Fig 12: Multiplicative decomposition of the data

From the above given 4 series, we can see that the trend and seasonality of the time series are clearly separated

using the multiplicative decomposition. The first series represents the given time series data, the second series represents the trend, the third one seasonality and the fourth one the residuals. It is noted that the residuals plotted in the above series follows a recognized pattern which ensures that the decomposition was not the apt one.

Q3. Split the data into training and test. The test data should start in 1991.

The test data should start from 01/01/1991. Hence the data is split accordingly. The below shows the shape pf the train and test data.

```
Shape of the train data: (132, 1)
Shape of the test data: (55, 1)
```

Here we can see that train data consists of 132 rows while test data consists of 55 rows.

First few rows of Training Data

Time_Stamp	Rose	time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

Last few rows of Training Data

Time_Stamp	Rose	time
1990-08-31	70.0	128
1990-09-30	83.0	129
1990-10-31	65.0	130
1990-11-30	110.0	131
1990-12-31	132.0	132

Train data

First few rows of Test Data

Time_Stamp	Rose	time
1991-01-31	54.0	133
1991-02-28	55.0	134
1991-03-31	66.0	135
1991-04-30	65.0	136
1991-05-31	60.0	137

Last few rows of Test Data

Time_Stamp	Rose	time
1995-03-31	45.0	183
1995-04-30	52.0	184
1995-05-31	28.0	185
1995-06-30	40.0	186
1995-07-31	62.0	187

Test data

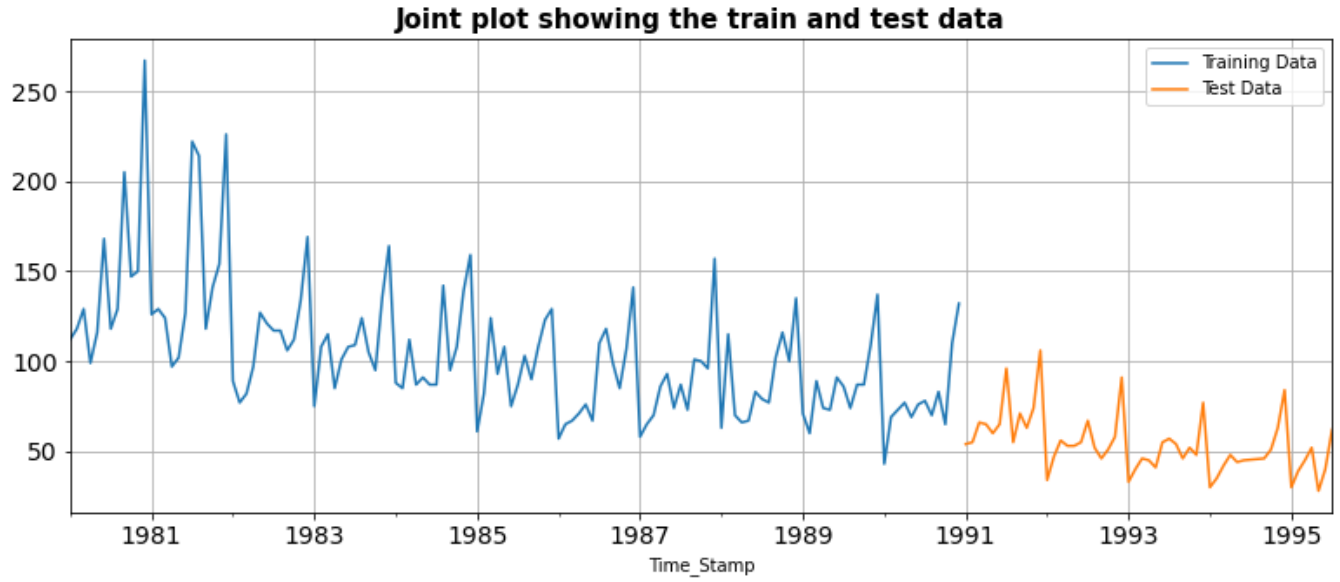


Fig 13: Joint plot showing the test and train data

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

Q4. 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.

1.Linear Regression

We use the linear regression algorithm to construct forecasting models. Linear regression is widely used in practice and adapts naturally to even complex forecasting tasks. The linear regression algorithm learns how to make a weighted sum from its input features.

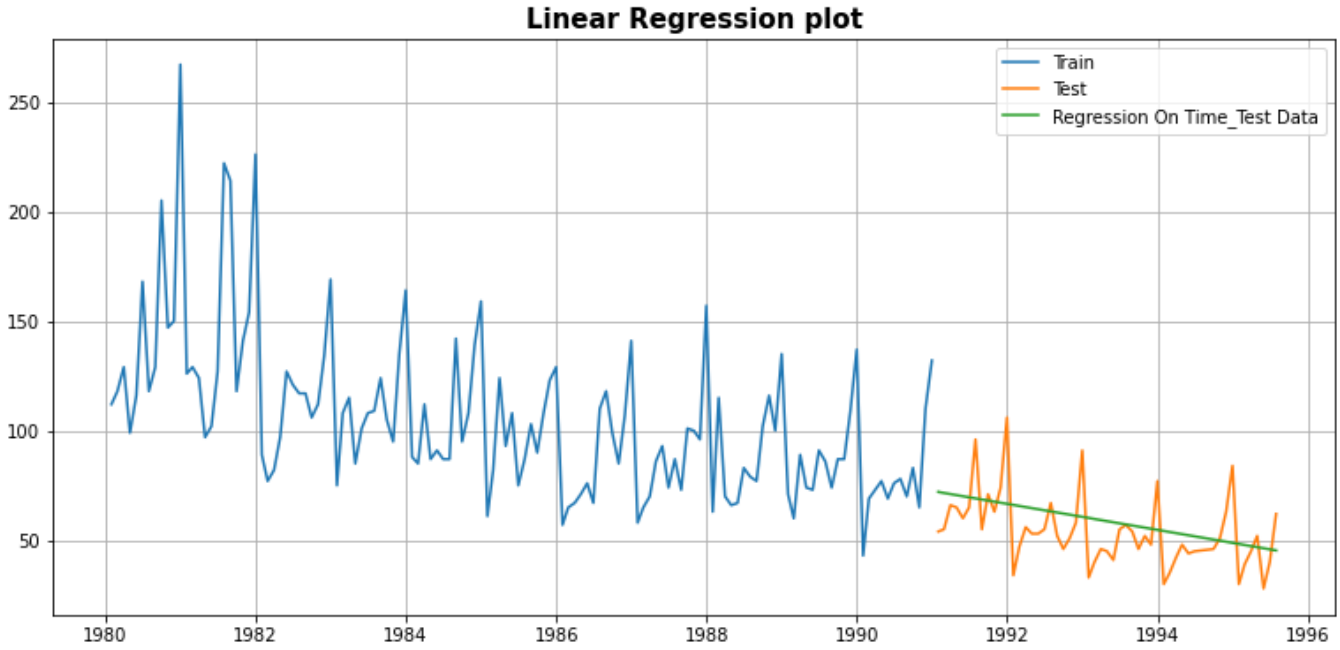


Fig 14: Linear regression plot

Model 1: Linear Regression Evaluation

```
For RegressionOnTime forecast on the Train Data,  RMSE is 30.718
For RegressionOnTime forecast on the Train Data,  MAPE is 21.220
For RegressionOnTime forecast on the Test Data,  RMSE is 15.269
For RegressionOnTime forecast on the Test Data,  MAPE is 22.820
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82

From the above results we can see that for linear regression, RMSE on Train data is 30.718 and RMSE on test data is 15.269

2.Naive Approach

Simple forecasting methods include naively using the last observation as the prediction or an average of prior observations. It is important and useful to test simple forecast strategies prior to testing more complex models. Simple forecast strategies are those that assume little or nothing about the nature of the forecast problem and are fast to implement and calculate. 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 the same as today, therefore the prediction for day after tomorrow is also today

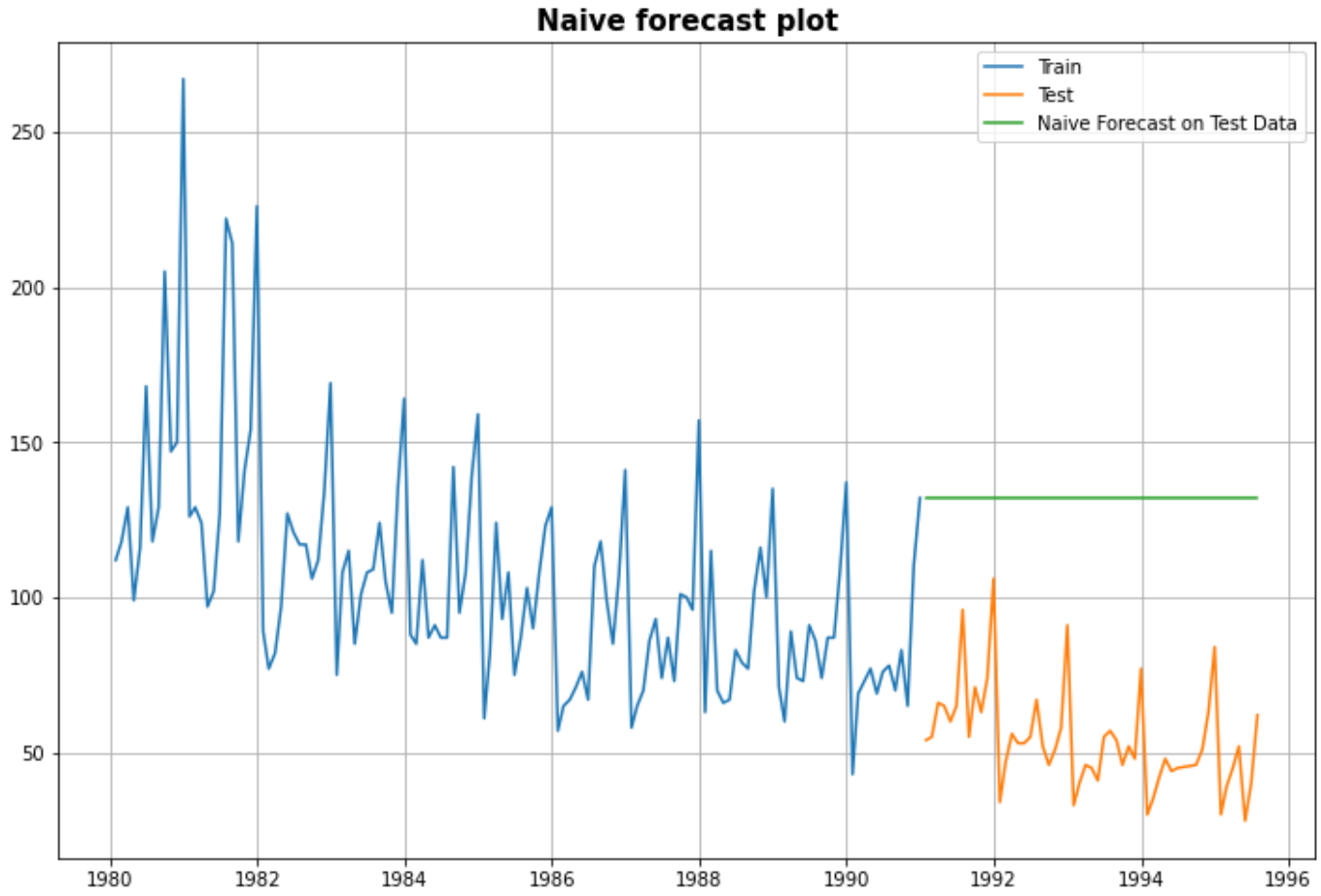


Fig 15: Naïve Forecast plot

Model 2: Naïve Forecast Evaluation

```
For Naive forecast on the Train Data,  RMSE is 2030.742
For Naive forecast on the Train Data,  MAPE is 36.380
For Naive forecast on the Test Data,  RMSE is 79.719
For Naive forecast on the Test Data,  MAPE is 145.100
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10

From the above results we can see that for linear regression, RMSE on Train data is 2030.742 and RMSE on test data is 79.719

We can infer from the RMSE values and the above graphs that the Naive method and Regression on Time models might not be suited for datasets with high variability. Naive method is best suited for stable datasets. Now we will adopt other techniques for improving the score. Now we will proceed to other techniques to improve our prediction accuracy.

3.Simple Average

This method is very simple where we average the data by months, years or quarters and then calculate the average for the period. Here, in this data, we will forecast by using the average of the training values.

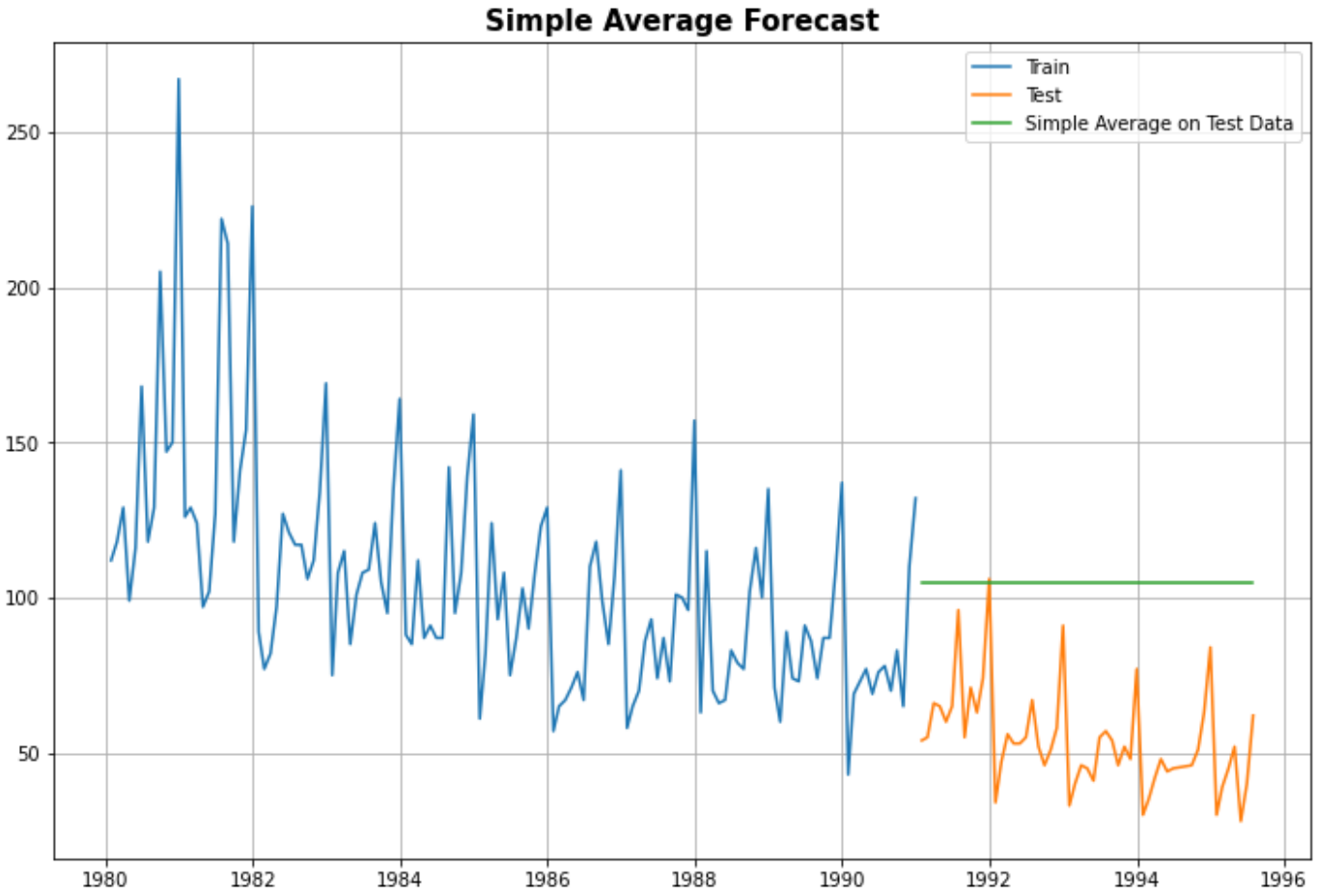


Fig 16: Simple Average Forecast plot

Model 3: Simple Average Forecast Evaluation

```
For Simple Average forecast on the Test Data,  RMSE is 53.461
For Simple Average forecast on the Test Data,  MAPE is 94.930
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93

From the above results we can see that for linear regression, RMSE on Train data is 53.461

4.Moving Average

Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting. Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series. This method is used for data where seasonal and cyclic variation is present. Here we will be calculating rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here. In this method, we are going to average over the entire data.

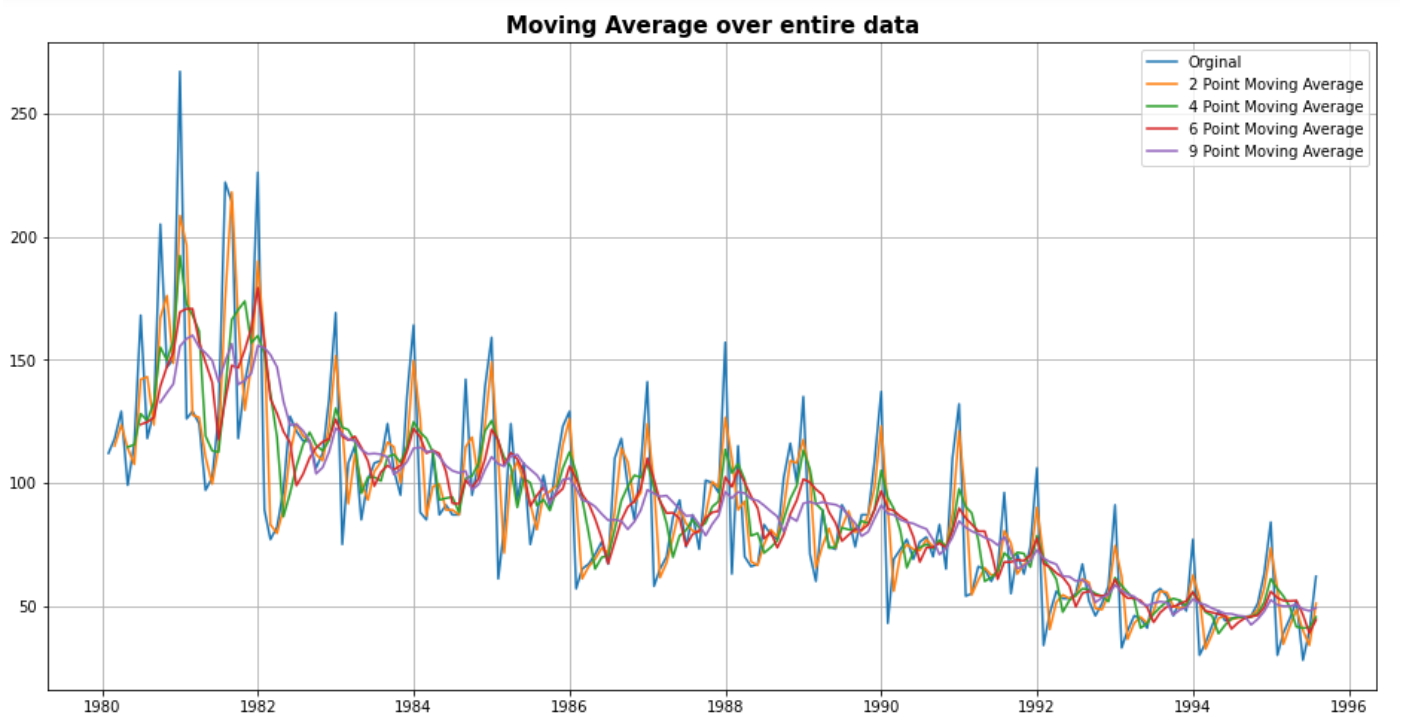


Fig 17: Moving Average over entire data

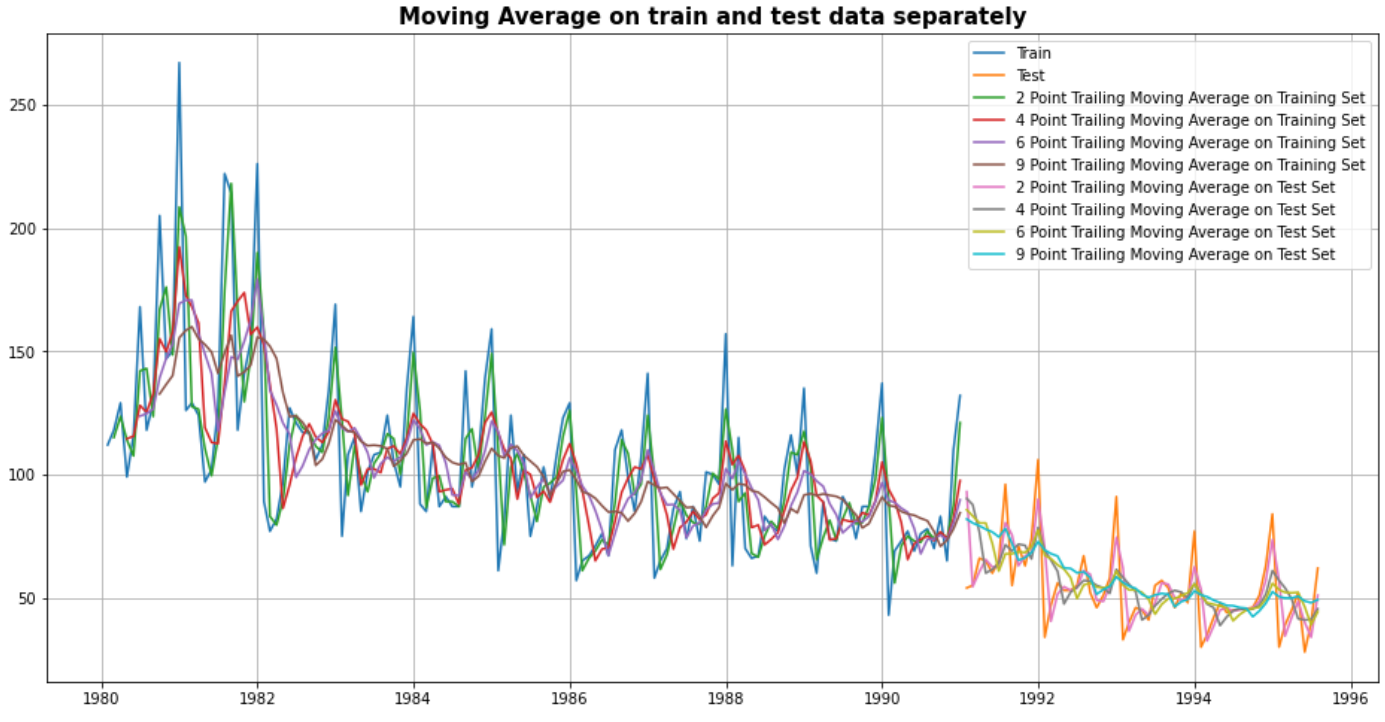


Fig 18: Moving Average over train and test data separately

Model 4: Moving Average Forecast Evaluation

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.451
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.566
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.728
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93
2pointTrailingMovingAverage	11.529278	13.54
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01

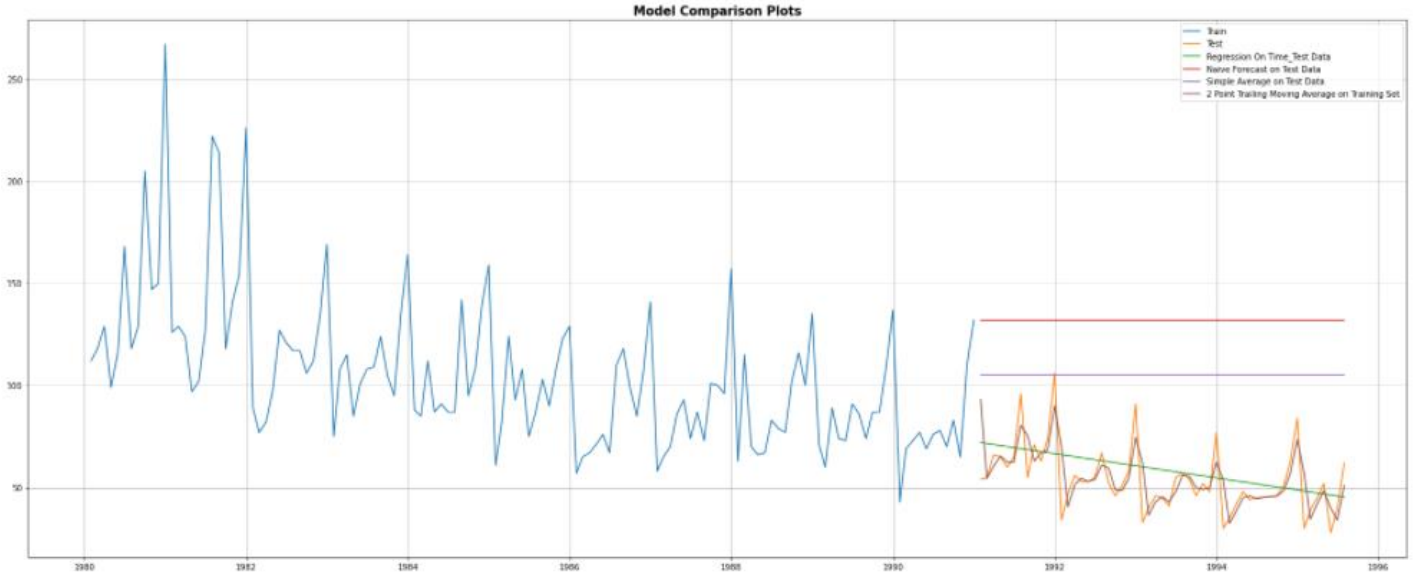


Fig 19: Model Comparison plots

Here we have plotted all the models done so far and compared the time Series plots.

5.Simple Exponential Smoothing

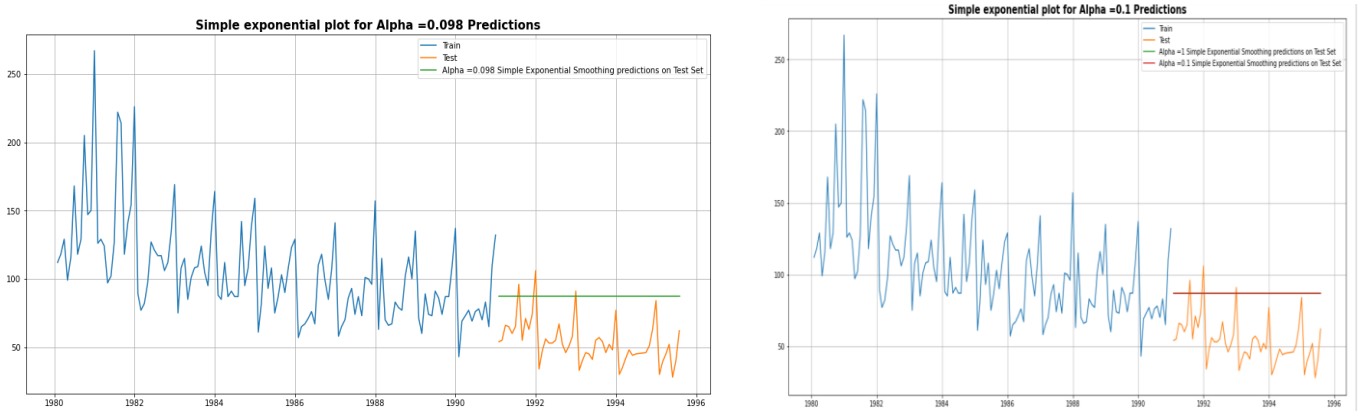


Fig 20: Simple exponential plots for alpha =0.098 and 0.1 respectively

Model 5: Simple Exponential Smoothing Evaluation

```
For Alpha =0.098 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.796

For Alpha=0.1,Simple Exponential Smoothing Model forecast on the test data,RMSE= [36.82803291069136]
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93
2pointTrailingMovingAverage	11.529278	13.54
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
Alpha=0.995,SimpleExponentialSmoothing	36.796242	63.88
Alpha=0.1,SimpleExponentialSmoothing	36.828033	NaN

6.Double Exponential Smoothing (Holt's Model)

Two parameters Alpha and Beta are estimated in this model. Level and Trend are accounted for in this model.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.923416
1	0.1	0.2	33.450729	48.688648
2	0.1	0.3	33.145789	78.156641
3	0.1	0.4	33.262191	99.583473
4	0.1	0.5	33.688415	124.269726
...	...	...	...	...
95	1.0	0.6	51.831610	801.680218
96	1.0	0.7	54.497039	841.892573
97	1.0	0.8	57.365879	853.965537
98	1.0	0.9	60.474309	834.710935
99	1.0	1.0	63.873454	780.079579

Now we will sort the data frame in the ascending order of 'Test RMSE'

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.923416
1	0.1	0.2	33.450729	48.688648
10	0.2	0.1	33.097427	65.731702
2	0.1	0.3	33.145789	78.156641
20	0.3	0.1	33.611269	98.653317

Table 9: Double exponential smoothing table with ascending order of RMSE

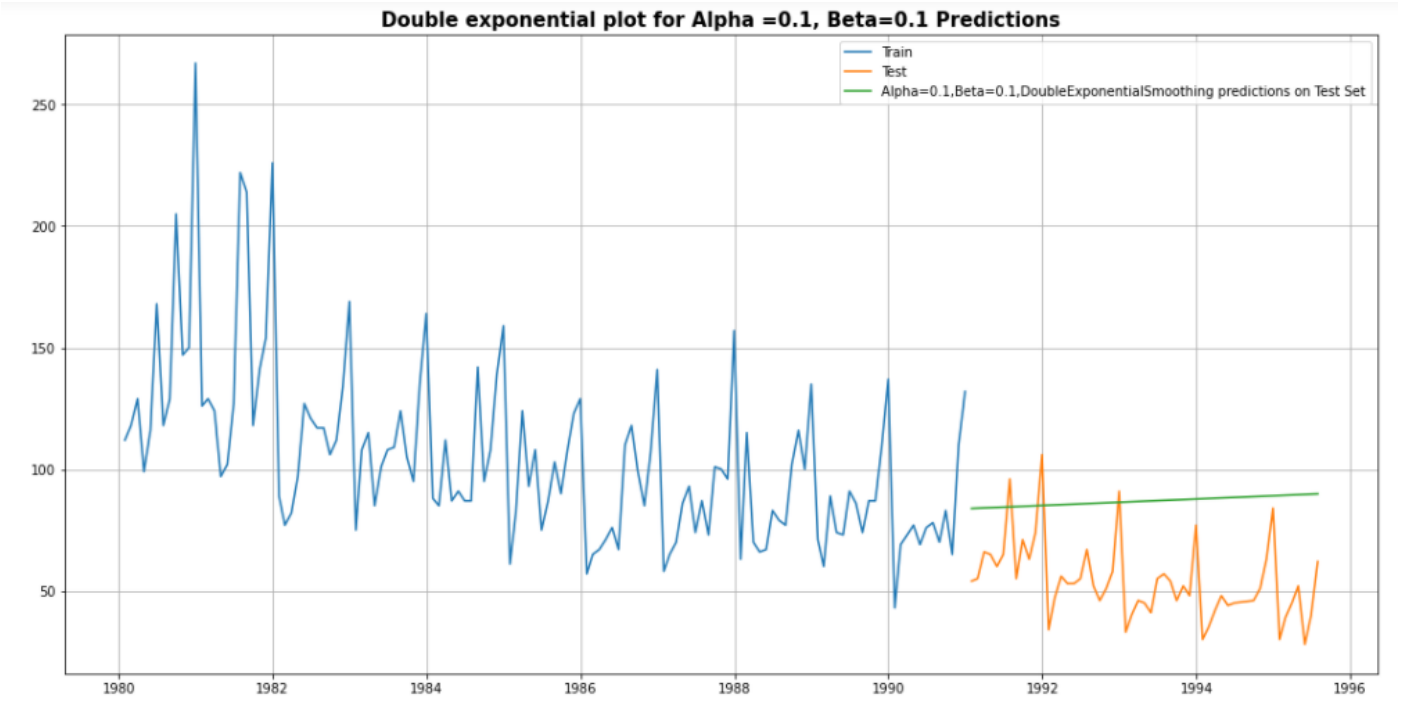


Fig 21: Double exponential smoothing plot for alpha =0.1 and beta= 0.1

Model 6: Double Exponential Smoothing Evaluation

```
For Alpha=0.1,Beta=0.1,Double Exponential Smoothing Model forecast on the test data,RMSE= [36.92341583254424]
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93
2pointTrailingMovingAverage	11.529278	13.54
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
Alpha=0.995,SimpleExponentialSmoothing	36.796242	63.88
Alpha=0.1,SimpleExponentialSmoothing	36.828033	NaN
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416	NaN

7.Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters alpha, beta and gamma are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Auto fit parameters

```
{'smoothing_level': 0.06385496671075688,
 'smoothing_trend': 0.054339412544817035,
 'smoothing_seasonal': 2.2153559567160688e-07,
 'damping_trend': nan,
 'initial_level': 52.70618108899535,
 'initial_trend': -0.3295208249140407,
 'initial_seasons': array([2.14032815, 2.42886639, 2.65328858, 2.31931542, 2.60742182,
        2.84354074, 3.12509302, 3.32298216, 3.15379088, 3.08514906,
        3.59598558, 4.96000257]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

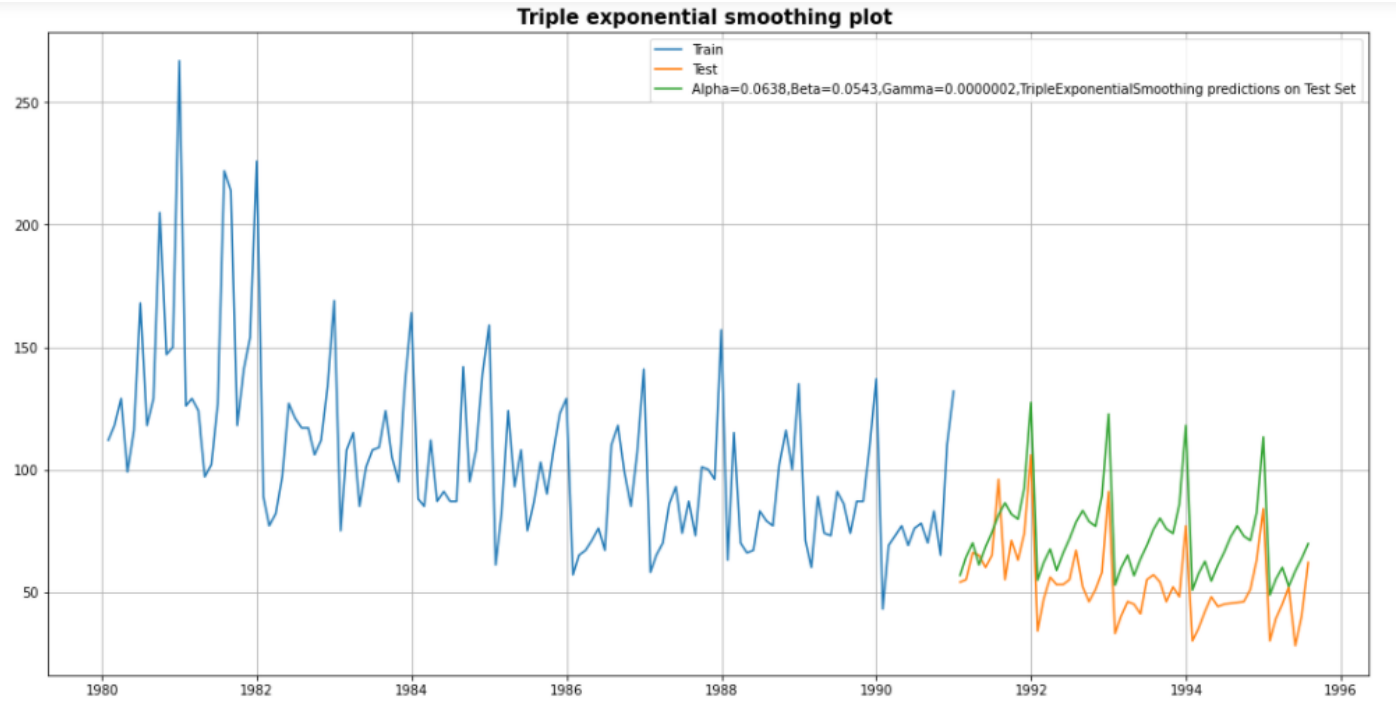


Fig 22: Triple exponential smoothing plots

Model 7: Triple Exponential Smoothing Evaluation

```
For Alpha=0.0638,Beta=0.0543,Gamma=0.0000002, Triple Exponential Smoothing Model forecast on the Test Data,  RMSE is 21.255
```

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93
2pointTrailingMovingAverage	11.529278	13.54
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
Alpha=0.995,SimpleExponentialSmoothing	36.796242	63.88
Alpha=0.1,SimpleExponentialSmoothing	36.828033	NaN
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416	NaN
Alpha=0.0638,Beta=0.0543,Gamma=0.0000002,TripleExponentialSmoothing	21.254806	NaN

8. Brute Force - Triple Exponential Smoothing

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.3	0.3	0.3	27.217969	19.057218
1	0.3	0.3	0.4	27.399095	11.201633
2	0.3	0.3	0.5	27.928512	30.565763
3	0.3	0.3	0.6	28.888611	63.623019
4	0.3	0.3	0.7	30.568635	122.472557
...	...	...	...	...	...
507	1.0	1.0	0.6	28358.458519	9603.635095
508	1.0	1.0	0.7	30724.126331	23029.955361
509	1.0	1.0	0.8	1218.755446	9626.710854
510	1.0	1.0	0.9	14150.253251	9691.905402
511	1.0	1.0	1.0	1768.254189	8138.618579

512 rows × 5 columns

Table 10: Brute force -Triple exponential smoothing table with various values for alpha, beta and gamma

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
8	0.3	0.4	0.3	28.111886	10.945435
1	0.3	0.3	0.4	27.399095	11.201633
69	0.4	0.3	0.8	32.601491	12.615607
16	0.3	0.5	0.3	29.087520	14.414604
131	0.5	0.3	0.6	32.144773	16.720720

Table 11: Brute force -Triple exponential smoothing table with increasing values of RMSE

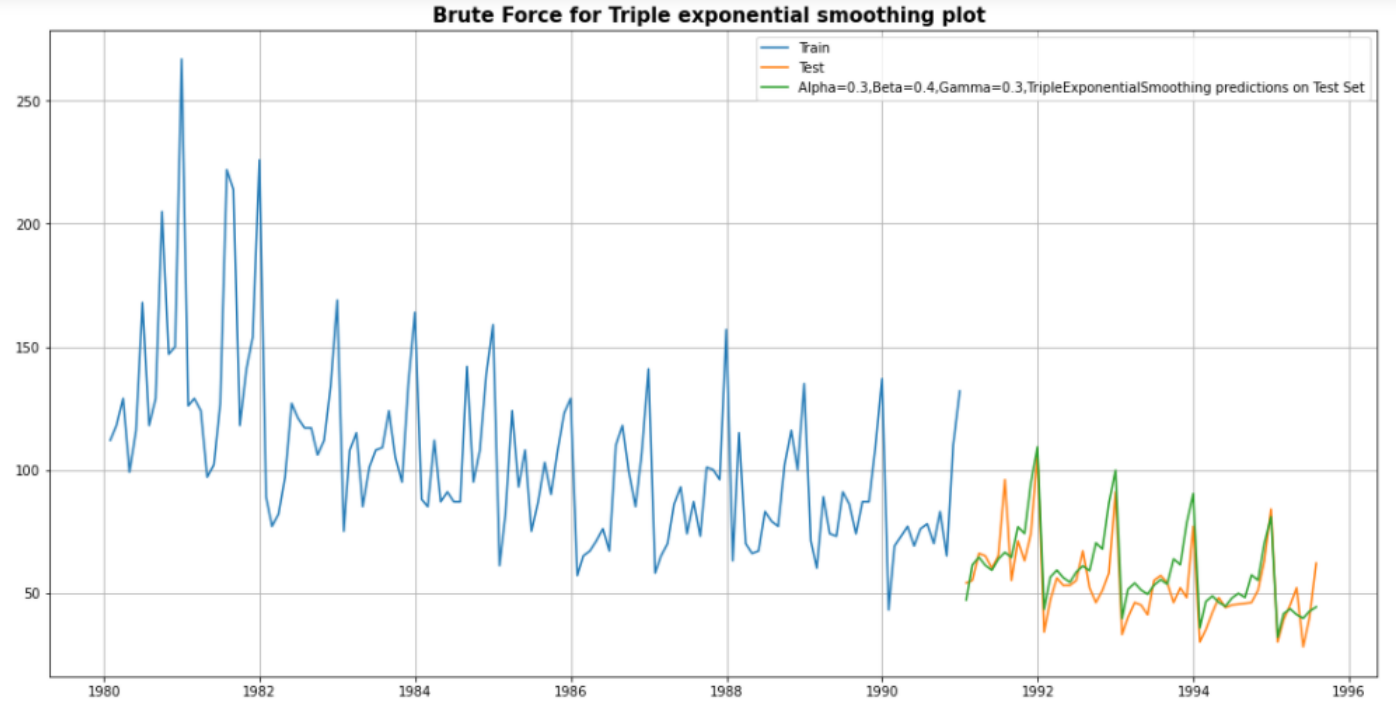


Fig 23: Brute force for triple exponential smoothing plot

Model 8: Brute Force- Triple Exponential Smoothing Evaluation

For Alpha=0.3,Beta=0.4,Gamma=0.3, Brute Force-Triple Exponential Smoothing Model forecast on the Test Data,RMSE is 10.945

	Test RMSE	Test MAPE
RegressionOnTime	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93
2pointTrailingMovingAverage	11.529278	13.54
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
Alpha=0.995,SimpleExponentialSmoothing	36.796242	63.88
Alpha=0.1,SimpleExponentialSmoothing	36.828033	NaN
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416	NaN
Alpha=0.0638,Beta=0.0543,Gamma=0.0000002,TripleExponentialSmoothing	21.254806	NaN
Alpha=0.3,Beta=0.4,Gamma=0.3,TripleExponentialSmoothing	10.945435	NaN

Smoothing models conclusion

For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing. Here we see that Brute force-triple exponential had the lowest RMSE of 10.945. However, we had gone on to build different models on the data and have compared these models with the best RMSE value on the test data. The following table shows the different smoothing models in the decreasing order of RMSE.

	Test RMSE	Test MAPE
Alpha=0.3,Beta=0.4,Gamma=0.3,TripleExponentialSmoothing	10.945435	NaN
2pointTrailingMovingAverage	11.529278	13.54
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
RegressionOnTime	15.268955	22.82
Alpha=0.0638,Beta=0.0543,Gamma=0.0000002,TripleExponentialSmoothing	21.254806	NaN
Alpha=0.995,SimpleExponentialSmoothing	36.796242	63.88
Alpha=0.1,SimpleExponentialSmoothing	36.828033	NaN
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416	NaN
SimpleAverageModel	53.460570	94.93
NaiveModel	79.718773	145.10

Table 12: Different smoothing models in the ascending order of RMSE.

Q5. 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.

- ### ACF plot

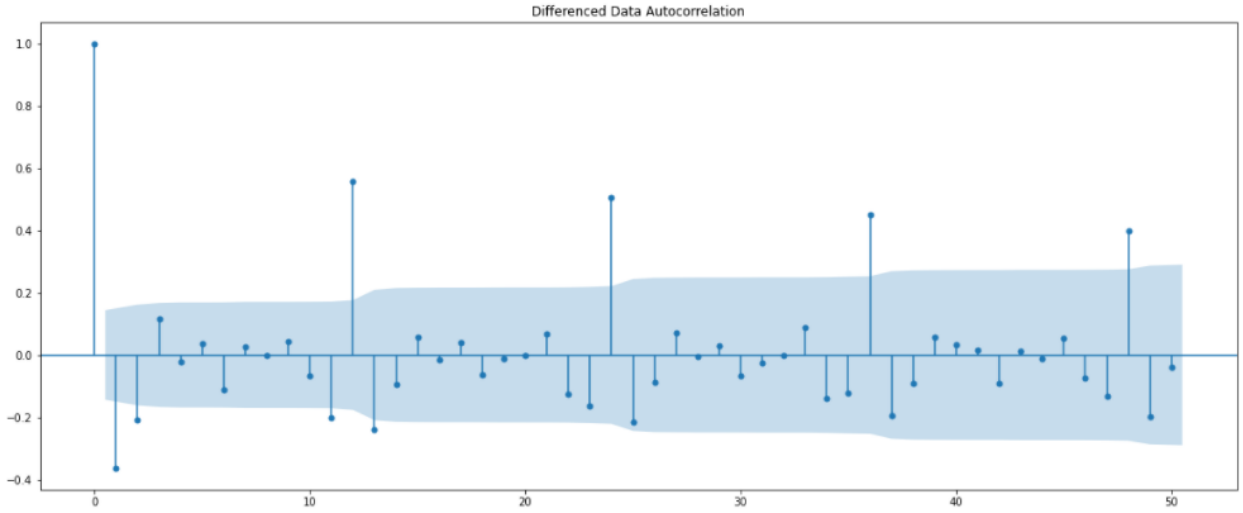


Fig 24: ACF plot

- ### PACF plot

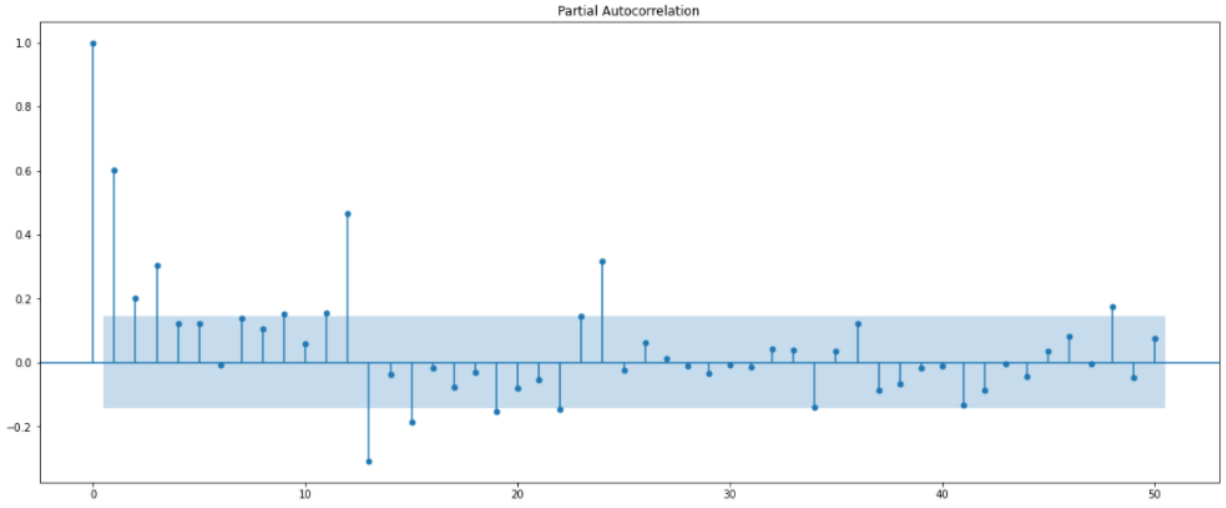


Fig 25: PACF plot

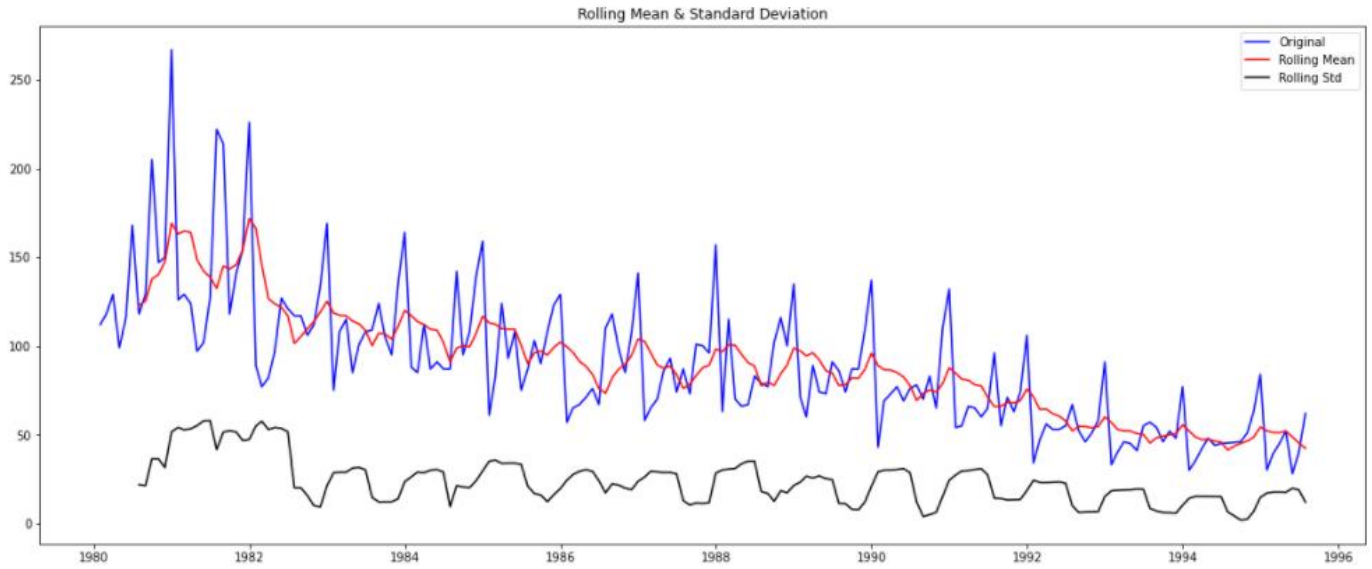
ACF and PACF plots are done with 95% confidence interval bands. We can see that seasonality after certain lags is visible is after every 12th Month. Hence, we do the stationarity test (Dickey-Fuller) for the series

Test for stationarity

Null and Alternate Hypothesis for the Augmented Dickey Fuller Test.

H0: The series is not stationary.

H1: The series is stationary.

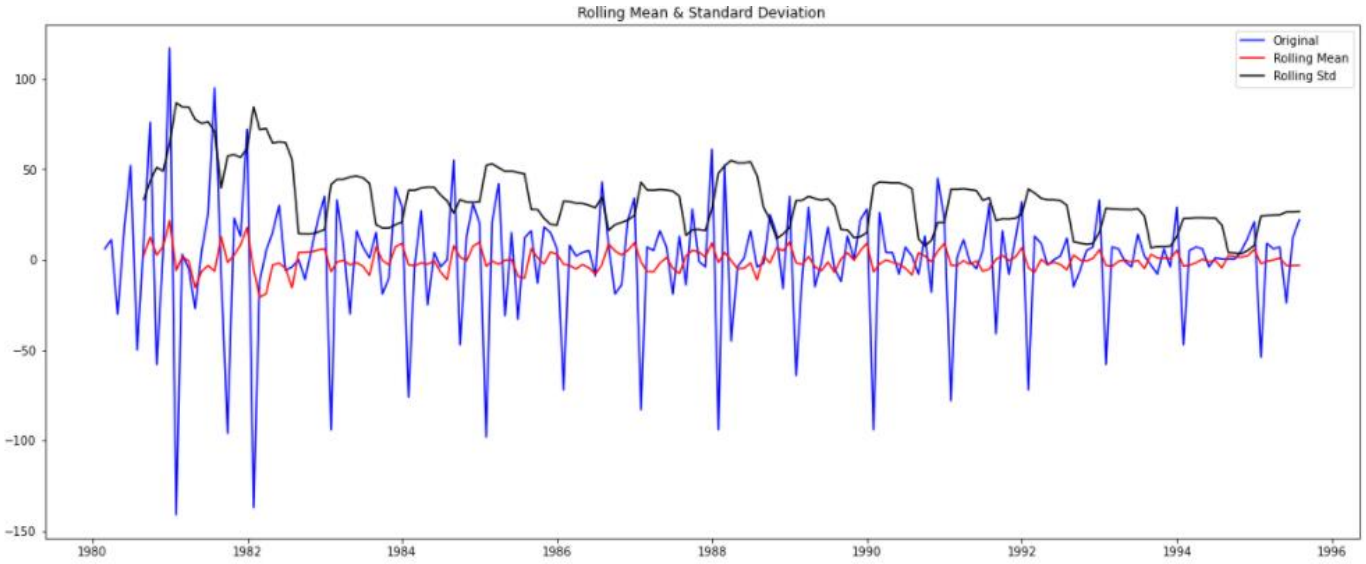


Series is not stationary with original form at alpha = 0.05.

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

From the above table, we see that p-value is not less than 0.05, therefore we fail to reject the null hypothesis.

As we have seen from the above graph and p-value, the series is not stationary. Therefore, we will check for stationarity after taking first order differencing.




```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

From the above table, it is very clear that p-value is less than 0.05, thereby we can reject the null hypothesis and we find that series is stationary post first order differencing at alpha = 0.05.

Q6. 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

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. Here the given series is not stationary and therefore differentiation is necessary. For Auto-ARIMA, we choose the best p, d and q parameters by looking at the lowest corresponding Akaike Information Criterion (AIC) values.

	param	AIC
2	(0, 1, 2)	1276.835374
5	(1, 1, 2)	1277.359224
4	(1, 1, 1)	1277.775752
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

Table 13: Different combinations of parameter values for ARIMA in the ascending order of AIC.

The best model is predicted by the lowest value of AIC. From the above table we see that p=0, d=1 and q=2 has the lowest AIC of 1276.835

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:	Fri, 18 Feb 2022	AIC	1276.835
Time:	21:24:49	BIC	1288.336
Sample:	02-29-1980 - 12-31-1990	HQIC	1281.509

	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	1.0001	+0.0000j	1.0001	0.0000
MA.2	-4.1695	+0.0000j	4.1695	0.5000

The above chart shows the Arima model results for p=0, d=1, q=2. For this particular Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1 and 2. The values of p, d, and q are calculated by giving a suitable range of values for p, d and q.

Automated ARIMA model Evaluation

```
RMSE of Automated ARIMA(0,1,2) on testing data: 15.618281486391918
```

	RMSE	MAPE
ARIMA(0,1,2)	15.618281	23.27

- ## 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. Since the given dataset possess seasonality, we can take the liberty to build the model with SARIMA. For an Auto-SARIMA, the parameters p, q, P and Q are selected based on the lowest Akaike Information Criterion (AIC).

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.901508
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

Table 14: Different combinations of parameter values for SARIMA in the ascending order of AIC.

Here we take the seasonality to be 12 since it is a yearly forecast. The best model is predicted by the lowest value of AIC. From the above table we see that p=0, d=1, q=2 and P=2, D=0, Q=2 and seasonality=12 has the lowest AIC of 887.9375

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sat, 19 Feb 2022   AIC                            887.938
Time:                                    00:01:57   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    189.947     -0.004      0.996    -373.133     371.447
ma.L2         -0.1573     29.842     -0.005      0.996     -58.646      58.332
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   4.77e+04      0.005      0.996   -9.33e+04    9.38e+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).
```

The above chart shows the Arima model results for p=0, d=1, q=2, P=2, D=0, Q=2 and seasonality=12. The values of p, d, q, P, D and Q are calculated by giving a suitable range of values for p, d and q.

Automated ARIMA model Evaluation

```
RMSE of Automated SARIMA(0,1,2)(2,0,2,12) on testing data: 26.92836200140455
```

	RMSE	MAPE
ARIMA(0,1,2)	15.618281	23.27
SARIMA(0,1,2)(2,0,2,12)	26.928362	46.60

Q7. 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 model based of cut-off points of ACF and PACF**

For this particular Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1 and 2. We are also considering the errors from the auto-regression of the first lag also. The values of p and q are calculated by looking at the ACF and the PACF plots.

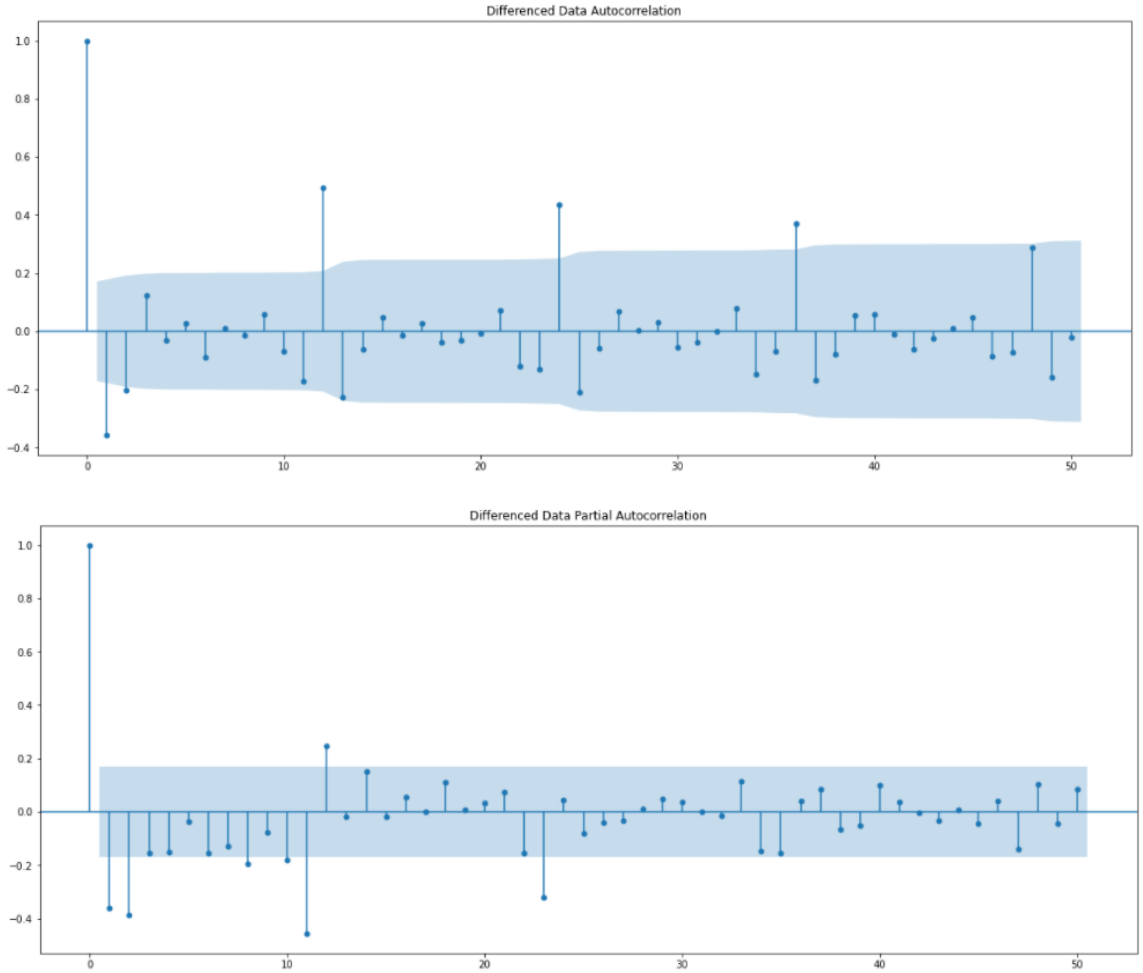


Fig 26: ACF and PACF plots for ARIMA model

From the above graphs of differenced auto-correlation and differenced partial auto correlation, we can see that p= 2, d=1, q=2. 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. Since, we have taken one difference of the series to be series, d=1.Here, we have taken alpha=0.05. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

ARIMA model based of cut-off points of ACF and PACF evaluation

```
RMSE of ARIMA model based on cut off points(2,1,2)on testing data: 15.354883182467582
```

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:	Sat, 19 Feb 2022	AIC	1279.299
Time:	00:32:59	BIC	1296.550
Sample:	02-29-1980 - 12-31-1990	HQIC	1286.309

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

Roots

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

	RMSE	MAPE
ARIMA(0,1,2)	15.618281	23.27
SARIMA(0,1,2)(2,0,2,12)	26.928362	46.60
ARIMA(2,1,2)	15.354883	22.77

- **SARIMA model based of cut-off points of ACF and PACF**

For this particular Seasonal Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1 and 2. We are also considering the errors from the auto-regression of the first lag. The values of p, q, P and Q are calculated by looking at the ACF and the PACF plots. In this particular model we have to find the p,d,q and the P,D,Q values manually by plotting the ACF and PACF plots. The p,d,q values will be same as the ARIMA model where we selected the values manually by looking at the ACF and PACF plot where we took first order differencing which are (2,1,2), but for reference purposes plots are imcluded below. Now coming to the seasonal parameter of P, D, Q we have to derive these by getting rid of trend from the data as much as possible or that which seems ideal.

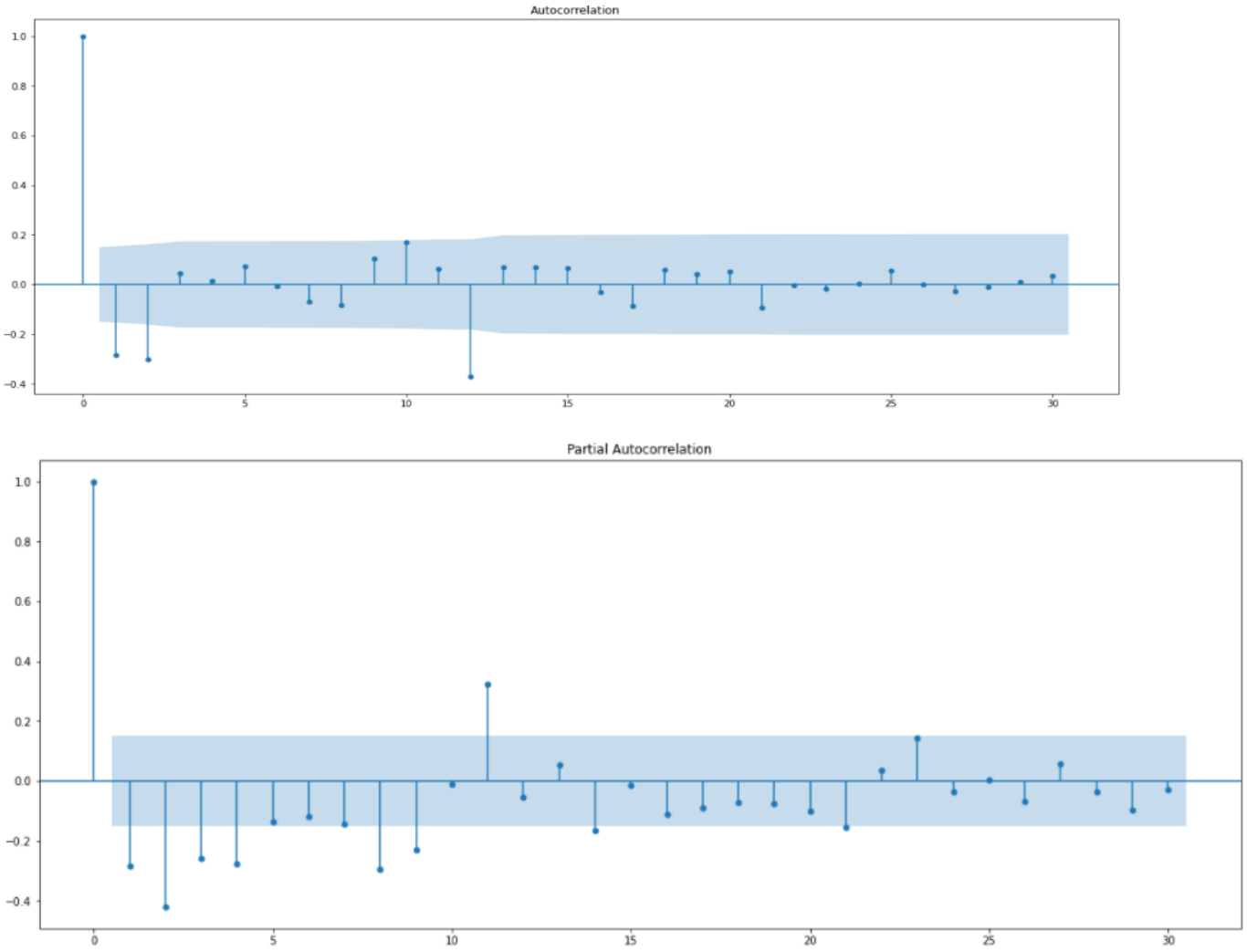


Fig 27: ACF and PACF plots for SARIMA model

From the above graphs of differenced auto-correlation and differenced partial auto correlation, we can see that P= 4, D=1, Q=2. The Auto-Regressive parameter in an ARIMA model is 'Q' 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. Since, we have taken one difference of the series to be series, D=1.Here, we have taken alpha=0.05. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

SARIMA model based of cut-off points of ACF and PACF evaluation

```
RMSE of SARIMA model based on cut off points(2,1,2)(4,1,2,12)on testing data: 17.342304252159312
```

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(4, 1, 2, 12)   Log Likelihood                -284.472
Date:                            Sun, 20 Feb 2022   AIC                            590.945
Time:                                    13:19:49   BIC                            615.520
Sample:                                         0   HQIC                           600.695
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9802      0.224     -4.377      0.000      -1.419      -0.541
ar.L2         -0.1271      0.143     -0.890      0.373      -0.407       0.153
ma.L1          0.0214      0.247      0.086      0.931      -0.463       0.506
ma.L2         -0.8832      0.193     -4.577      0.000      -1.261      -0.505
ar.S.L12      -0.7352      0.198     -3.704      0.000      -1.124      -0.346
ar.S.L24      -0.0735      0.174     -0.422      0.673      -0.415       0.268
ar.S.L36       0.0757      0.088      0.859      0.390      -0.097       0.249
ar.S.L48      -0.0064      0.021     -0.308      0.758      -0.047       0.034
ma.S.L12      -0.3539      0.695     -0.509      0.611      -1.716       1.008
ma.S.L24      -0.9033      0.557     -1.622      0.105      -1.995       0.188
sigma2       144.4813    109.809      1.316      0.188     -70.740     359.702
==========================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 6.01
Prob(Q):                              0.91   Prob(JB):                         0.05
Heteroskedasticity (H):               0.62   Skew:                             0.53
Prob(H) (two-sided):                  0.25   Kurtosis:                         3.98
==========================================================================================
```

	RMSE	MAPE
ARIMA(0,1,2)	15.618281	23.27
SARIMA(0,1,2)(2,0,2,12)	26.928362	46.60
ARIMA(2,1,2)	15.354883	22.77
SARIMA(2,1,2)(4,1,2,12)	17.342304	26.69

Q8. 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.

Model	Parameters	Test RMSE
Triple Exponential Smoothing (Brute Force-AIC)	Alpha=0.3 Beta=0.4 Gamma=0.3	10.945
Moving Average	2 point Trailing moving average	11.529
Moving Average	4 point Trailing moving average	14.451
Moving Average	6 point Trailing moving average	14.566

Moving Average	9 point Trailing moving average	14.727
Linear Regression		15.268
ARIMA (auto)	p, d, q= (0,1,2)	15.618
ARIMA (manual)	p, d, q= (2,1,2)	15.354
SARIMA (manual)	p, d, q= (2,1,2) P, D, Q= (4,1,2) Seasonality=12	17.342
Triple Exponential Smoothing	Alpha=0.063, Beta=0.054 Gamma=0.00	21.254
SARIMA (auto)	p, d, q= (0,1,2) P, D, Q= (2,0,2) Seasonality=12	26.928
Simple Exponential Smoothing	Alpha=0.098	36.796
Simple Exponential Smoothing (AIC)	Alpha=0.1	36.828
Double Exponential Smoothing	Alpha=0.3, Beta=0.3	36.923
Simple Average Model		53.460
Naïve model		79.718

Table 15: Table containing all the models along with their parameters and RMSE scores.

From the above table we can see that out of all the models we have built, triple exponential smoothing (AIC) is the best model since it has the lowest RMSE of 10.945. Therefore, we proceed to build Triple exponential model on the complete data and use it for predicting the next 12 months.

Q9. 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.

Triple exponential smoothing(Brute Force-AIC)

Let us apply Brute Force Triple Exponential Smoothing on Full Data The best-found parameters for Triple Exponential Smoothing are Alpha=0.3, Beta=0.4, Gamma=0.3

```
RMSE of Triple exponential smoothing on entire data: 20.672560612963352
```

1995-08-31	40.466299
1995-09-30	39.523150
1995-10-31	41.472534
1995-11-30	48.011557
1995-12-31	68.284700
1996-01-31	25.686714
1996-02-29	32.790356
1996-03-31	38.933982
1996-04-30	41.796358
1996-05-31	32.872603
1996-06-30	39.665795
1996-07-31	41.833692

16(a)

	lower_CI	prediction	upper_ci
1995-08-31	-0.145491	40.466299	81.078089
1995-09-30	-1.088640	39.523150	80.134940
1995-10-31	0.860744	41.472534	82.084324
1995-11-30	7.399767	48.011557	88.623347
1995-12-31	27.672910	68.284700	108.896490
1996-01-31	-14.925076	25.686714	66.298504
1996-02-29	-7.821434	32.790356	73.402146
1996-03-31	-1.677808	38.933982	79.545772
1996-04-30	1.184568	41.796358	82.408148
1996-05-31	-7.739187	32.872603	73.484393
1996-06-30	-0.945996	39.665795	80.277585
1996-07-31	1.221901	41.833692	82.445482

16(b)

Table 16(a): Table with sale predictions for next 12 months

Table 16(b): Table with sale predictions for next 12 months with lower and upper confidence intervals

In the above table, we have calculated the upper and lower confidence bands at 95% confidence level. From the above table we can see that the sales of the Rose wine which was predicted for next 12 months shows almost a low sale just like the past years' sales. This prediction follows the downward trend as in the past. This indicates that either the price of this sales was not affordable to the customers or they didn't like this wine as the initial years. Therefore, it is recommended either the price has to be brought down by the firm or the recipe of this wine has to change to make it appealing to the customers.

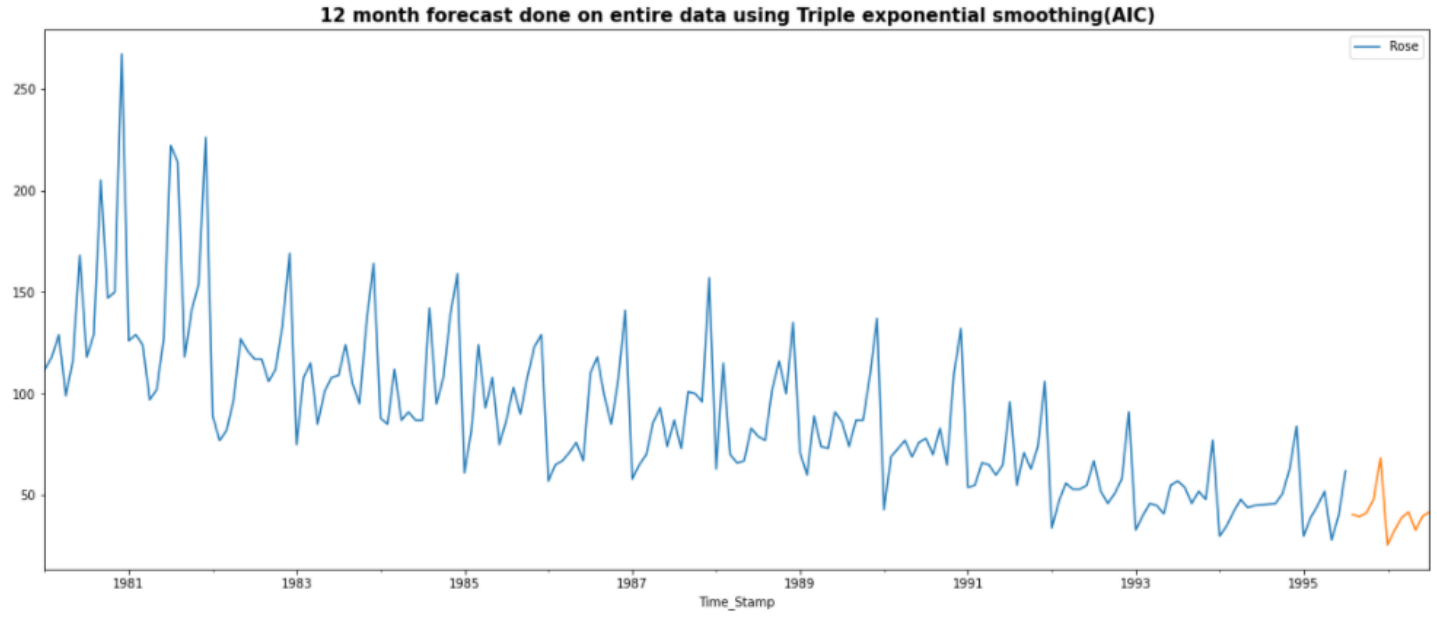


Fig 28(a)

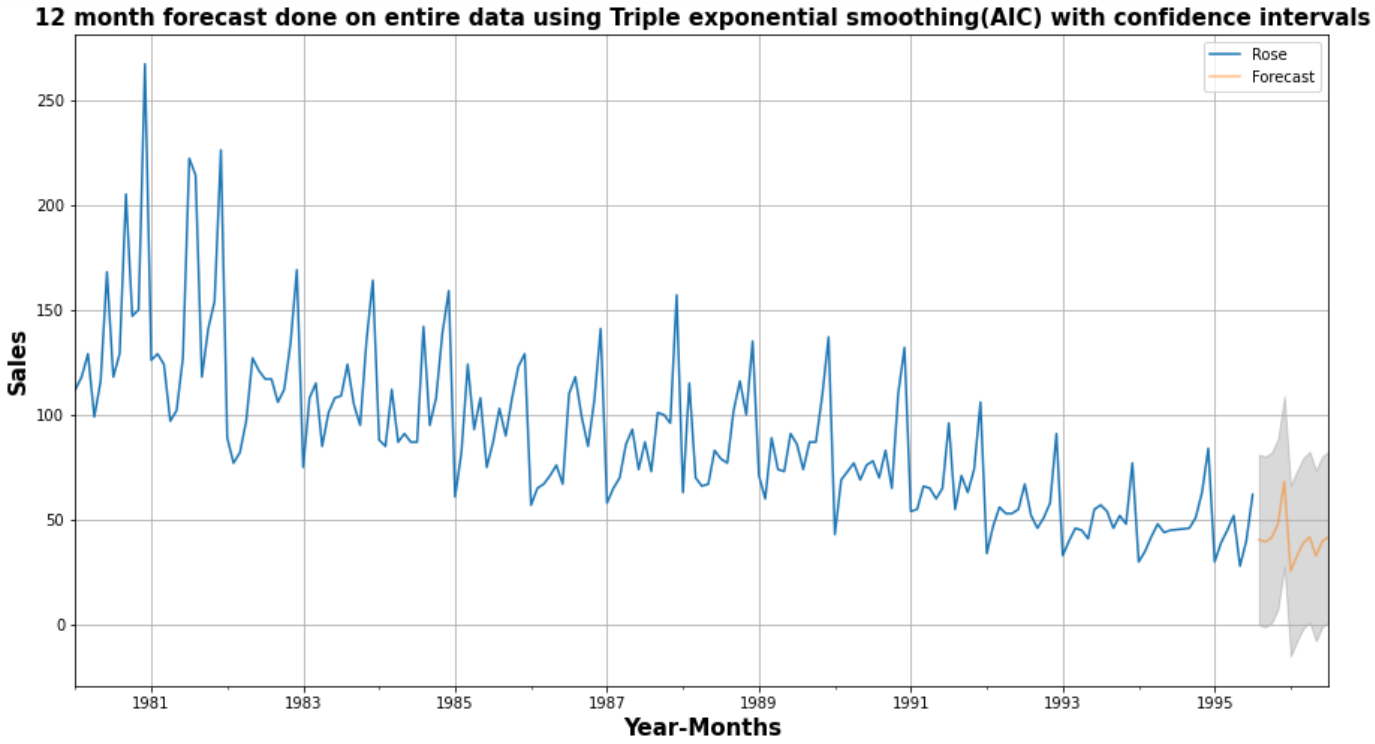


Fig 28(b)

Fig 28(a & b):12 months forecasts on entire data using Triple exponential smoothing (AIC) with and without confidence intervals

In the above plot, we have plotted the graph by taking the upper and lower confidence bands at 95% confidence level into consideration. The above plot clearly depicts a downward trend in the forecasts. The shaded region in the forecasts shows the confidence intervals of the predictions.

Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- The best Accuracy (RMSE) is observed by triple exponential smoothing (Brute force-AIC) where alpha=0.3, beta=0.4, gamma=0.3

• The second-best Accuracy (RMSE) is observed by two-point trailing moving average.

- For doing the best forecast for the Rose dataset we will choose the Triple Exponential Smoothing (brute force-AIC) as that model has the best accuracy (lowest RMSE)
- It is predicted that the predicted for next 12 months shows almost a low sale just like the past years' sales. This prediction follows the downward trend as in the past.
- In the forecasts, the month of December is having the highest sales just like the previous years. This high sales during that month can be due to the festival or holiday seasons.
- To increase the sales of Rose wine, it is recommended either the price has to be brought down by the firm or the recipe of this wine has to change to make it appealing to the customers.

TIME SERIES FORECASTING PROJECT2

Submitted by,

Jiya Jacob

PGP-DSBA ONLINE

JULY-B 2021

DATE:20/02/2022

SPARKLING WINE

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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

INTRODUCTION

The purpose of this whole exercise is to perform various smoothing techniques and forecasting techniques on the given data, combine all predictions and eventually perform predictions based on past data with given trend and seasonality. The smoothing techniques are the members of time series forecasting methods or algorithms, which use the weighted average of a past observation to predict the future values or forecast the new value. These techniques are well suited for time-series data having fewer deviations with time.

DATA DICTIONARY

1. YearMonth: The time period of a certain amount of sales.
2. Sparkling: The amount of sales of the wine during a particular period of time.

Q1. **Read the data as an appropriate Time Series data and plot the data.**

a. Head of the Time series data

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

Table 1: Head of the time series data

b. Tail of the Time series data

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

Table 2: Tail of the time series data

From the above head and tail samples of data, we can see that the data has not been identified as time series. Therefore, we need to create a time index for this data. We do so by creating time stamp which starts at 01/01/1985 to 31/07/1995.Then we set this time stamp as index. Since the newly created time stamp and the column YearMonth has the same time periods, we can remove the YearMonth column which is irrelevant, thereby reaching to our final data.

c. Creating the time stamps and adding to data frame

	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

Table 3: Head of the data after creating the time stamp

d. Final data frame

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

Table 4: Head of the final data frame

e. Time series data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   YearMonth   187 non-null    object
 1   Sparkling   187 non-null    int64 
dtypes: int64(1), object(1)
```

From the above table we can see that there are 187 rows for this time series data. The Sparkling column contains 187 non-null values which indicates that there is no presence of null values. The data type of the Sparkling column is int64.

The Sparkling wine sales of 15 years ranging from 01/01/1980 to 31/07/1995 is recorded in the given data.

f. Checking for null values in the data.

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

As mentioned above, from the above table, we can see that the column Sparkling is not having any null values.

g. Graphical representation of data

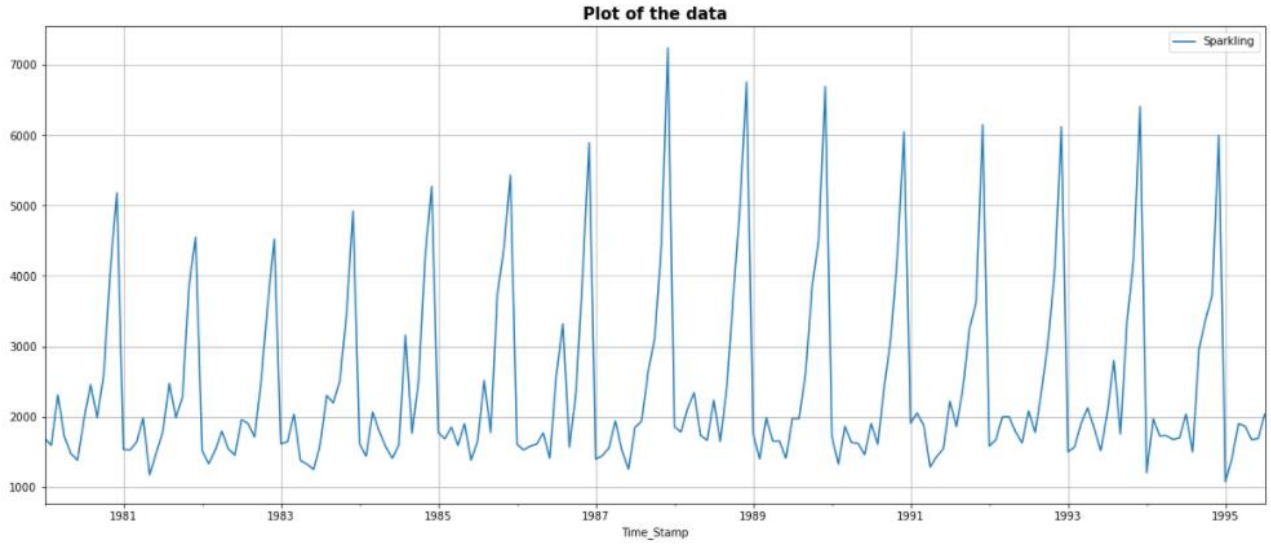


Fig 1: Graphical representation of data

From the above graph, we can see that there exists a pattern withing each year, thereby indicating a seasonal effect, that might be present. The seasonal effect can be inferred from the time series plot. This seasonal effect helps us in predicting the future values. Trend seems to be absent in above shown graph. Since we have considered only one variable (Sparkling column) for the plotting of the above graph, this is called the univariate time series analysis or forecasting. For predicting the sales of Sparkling wine for next 12 months, we need to first look into the past values, extract a pattern and then forecast the future sales.

Q2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

a. Five-point summary data

	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

From the above table we see that mean of the Sparkling wine sales in 20 the century is 2402.41 where, the minimum sales 1070 and maximum sales is 7242. The median of the Rose wine sales is 1874and the standard deviation is 1295.11. Range is 6172

b. Exploratory Data Analysis

1. Univariate Time series

A univariate time series is a series with a single time-stamped variable at time 't'. Here the dataset belongs to the Sparkling wine sales from the January of 1980 to July of 1995. Here, Sparkling is the time-dependent variable. The series is a monthly series, wherein for each month between Jan-1980 and Jul-1995 a datapoint is recorded.

2. Empirical Cumulative Distribution Function (ECDF)

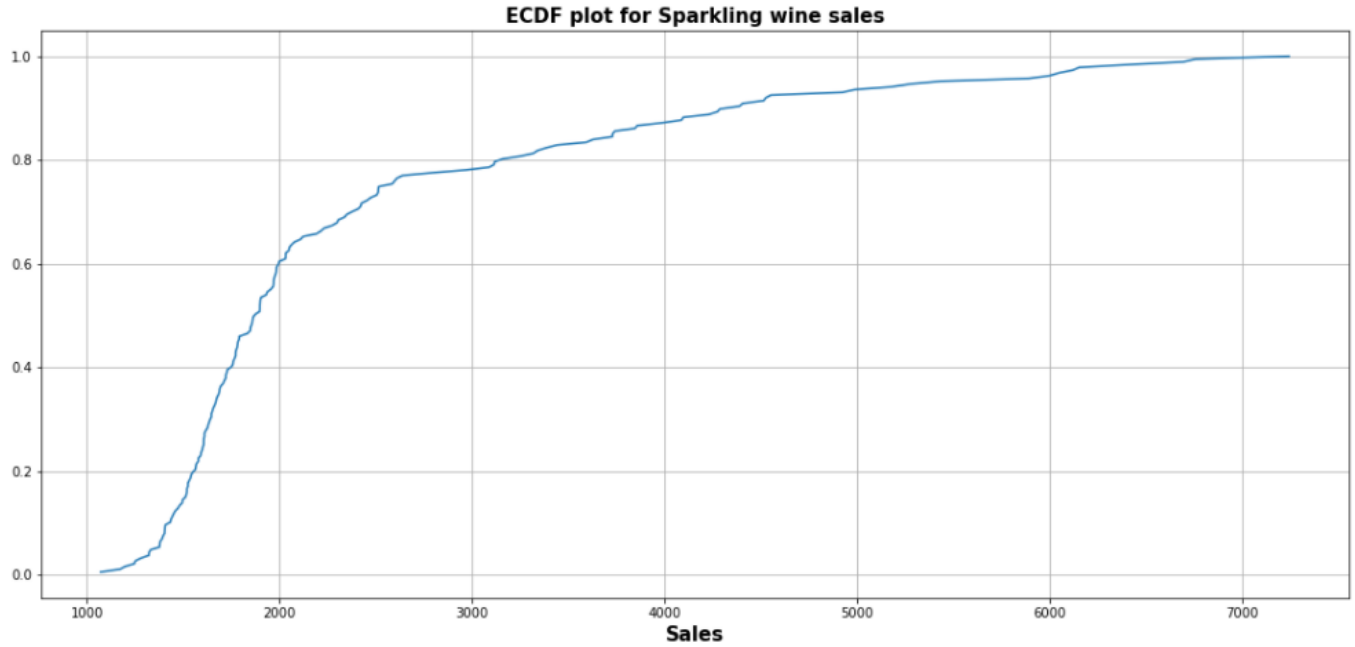
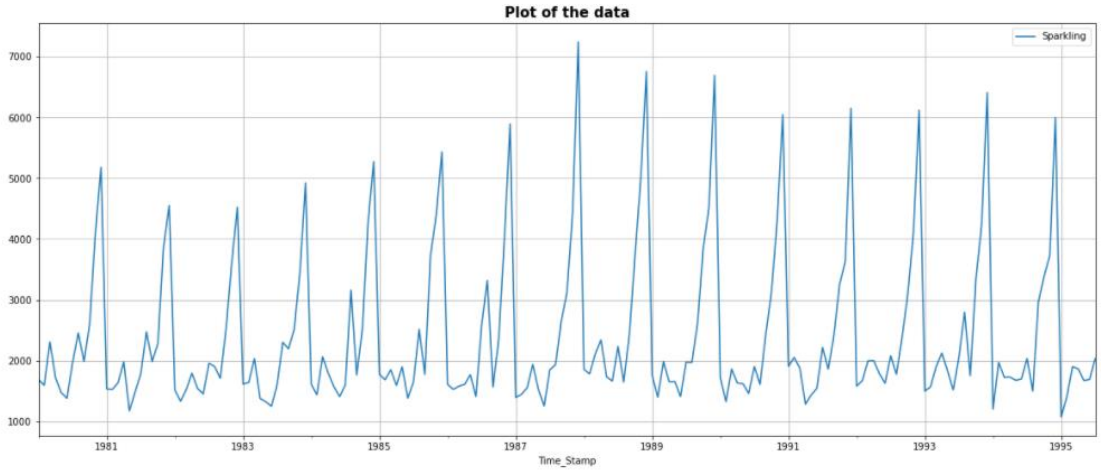


Fig 2: ECDF plot for Sparkling wine sales

An ECDF is an estimator of the Cumulative Distribution Function. The ECDF essentially allows you to plot a feature of the data in order from the least to greatest and see the whole feature as if it is distributed across the data set. From the five-point summary statistics, we can see that the data ranges from 1070 to 7242.

3. Plot for the Sparkling time series



From the above graph, we can see that there exists a pattern withing each year, thereby indicating a seasonal effect, that might be present. The seasonal effect can be inferred from the time series plot. This seasonal effect helps us in predicting the future values. Trend seems to be absent in above shown graph. Since we have considered only one variable (Sparkling column) for the plotting of the above graph, this is called the univariate time series analysis or forecasting. For predicting the sales of Sparkling wine for next 12 months, we need to first look into the past values, extract a pattern and then forecast the future sales.

4. Year wise Rose wine sales assessment using Box plot

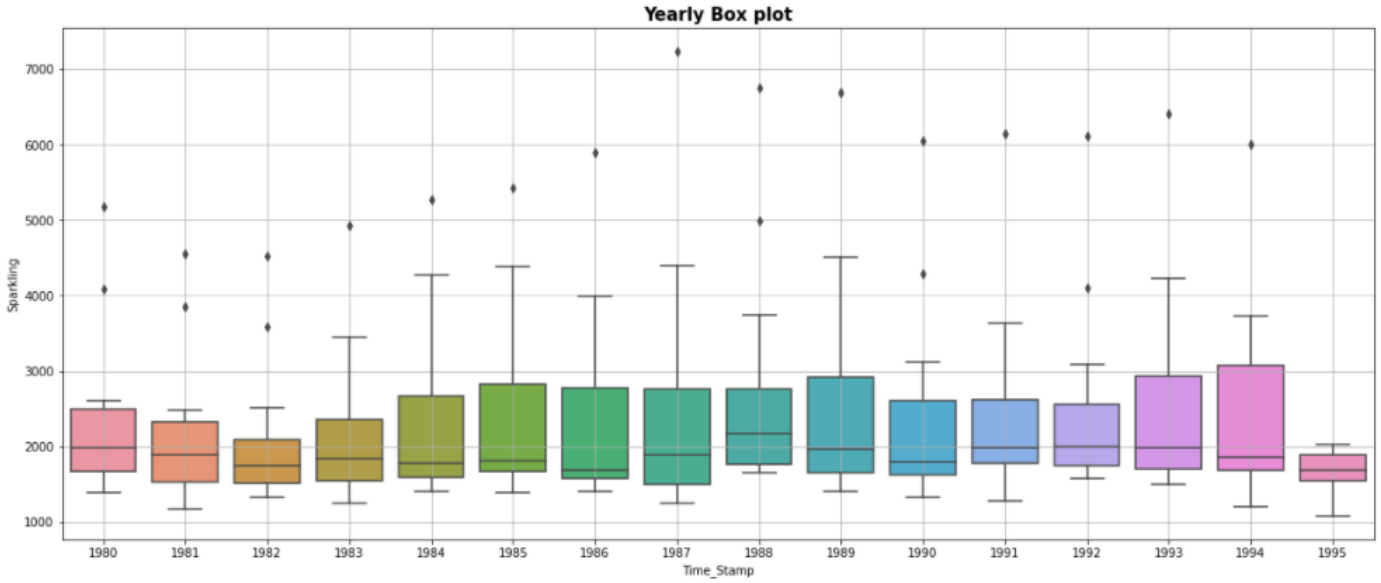


Fig 3: Yearly box plot

From the above figure we see that there isn't much noticeable difference, even though there are few high ranges in the middle years. The last year's sale is shown less because we have only 7 months sales out of 12 months in our given data. For most of the years, outliers are present.

5. Monthly wise Rose wine sales assessment using box plot

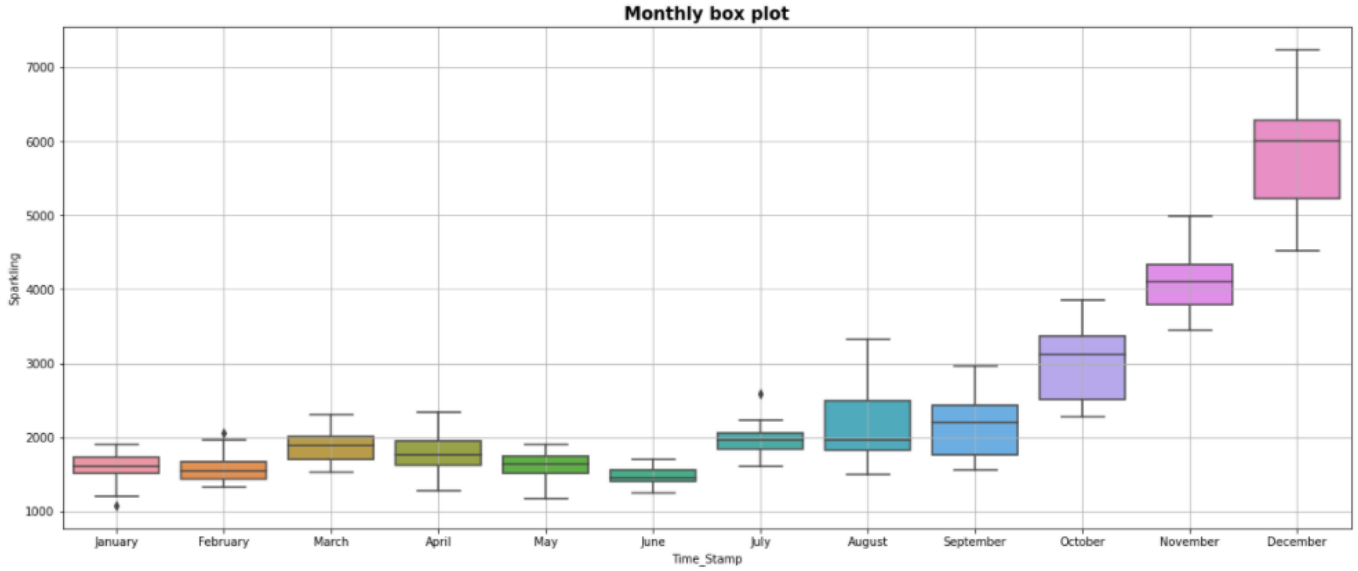


Fig 4: Monthly box plot

From the above graph, we can see that the sales of wine were highest during the month of December followed by November and October respectively. This can be due to the possible holidays and festivals happening in those months. The lowest sales of the wine happen in the month of June. The month of January, February and July has outliers. It should be kept in mind that some months will have one observation less than the other months due to the incomplete data for the year 1995.

6. Year/Month table

Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
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 5: Year/month table

From the above table we can observe that in the year 1995 after the month of July there are no entries as we were only given data until the seventh month of 1995.

7. Monthly plot for each month sales distribution

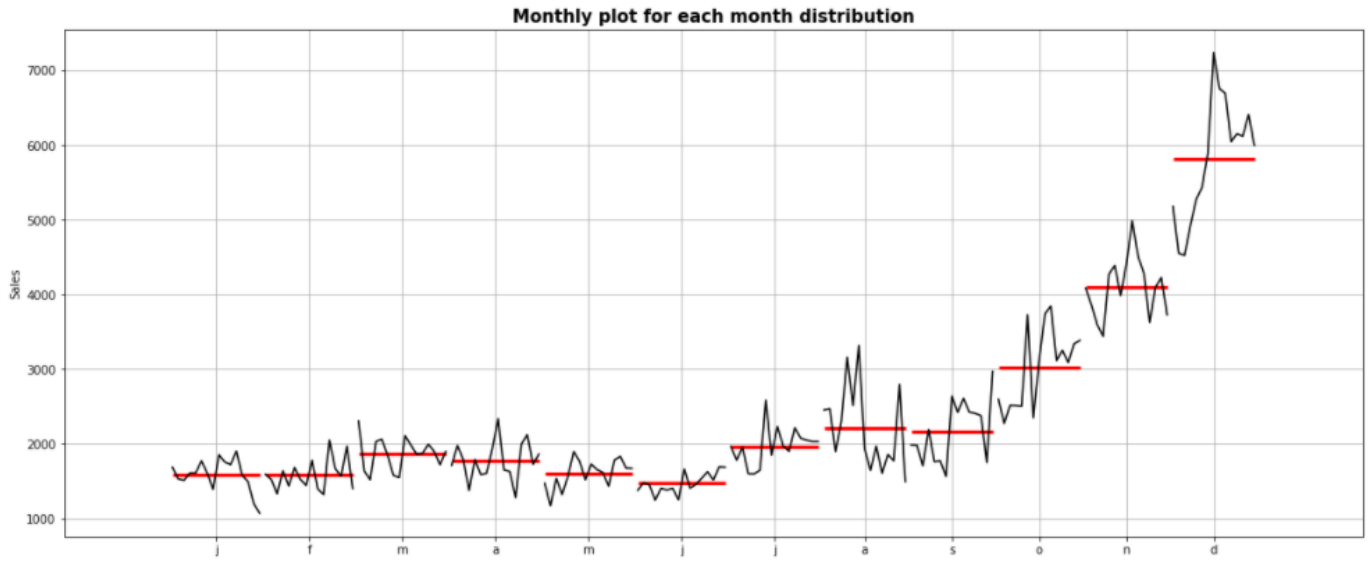


Fig 5: Monthly plot for each month distribution

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

8. Year wise Rose wine sales assessment using line plot

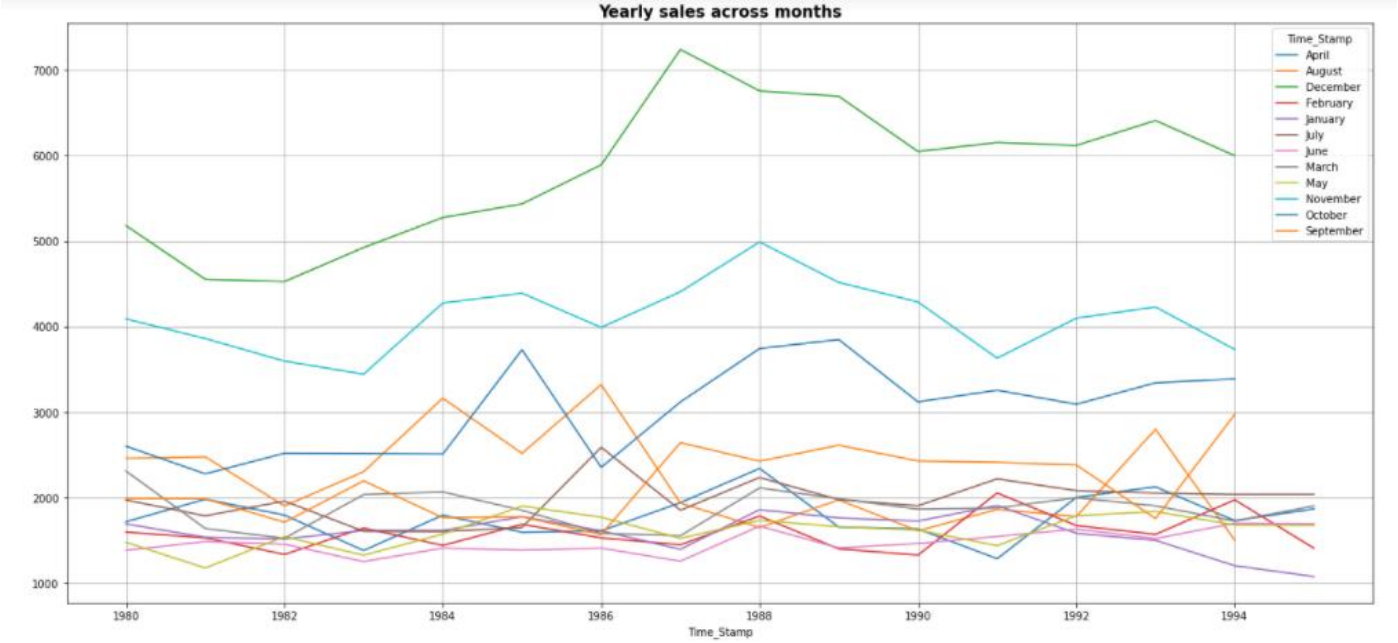


Fig 6: Yearly sales across months using line plot

In the above plot, we have done a line plot to compare the sales of Sparkling wine across each month over the years. From the above yearly line plot we can observe that the month of December outperforms all the other months in terms of sales, as we have seen in the boxplots above. We also observe that the month of January has the lowest number of sales. We can also say that the greatest sale season of the year was October-December.

9. Yearly sales table and plot of Rose wine

Time_Stamp	Sparkling
1980-12-31	28406
1981-12-31	26227
1982-12-31	25321
1983-12-31	26180
1984-12-31	28431
1985-12-31	29640
1986-12-31	29170
1987-12-31	30258
1988-12-31	33246
1989-12-31	31443
1990-12-31	28977
1991-12-31	29587
1992-12-31	30171
1993-12-31	30991
1994-12-31	29584
1995-12-31	11620

Table 6: Yearly sales table

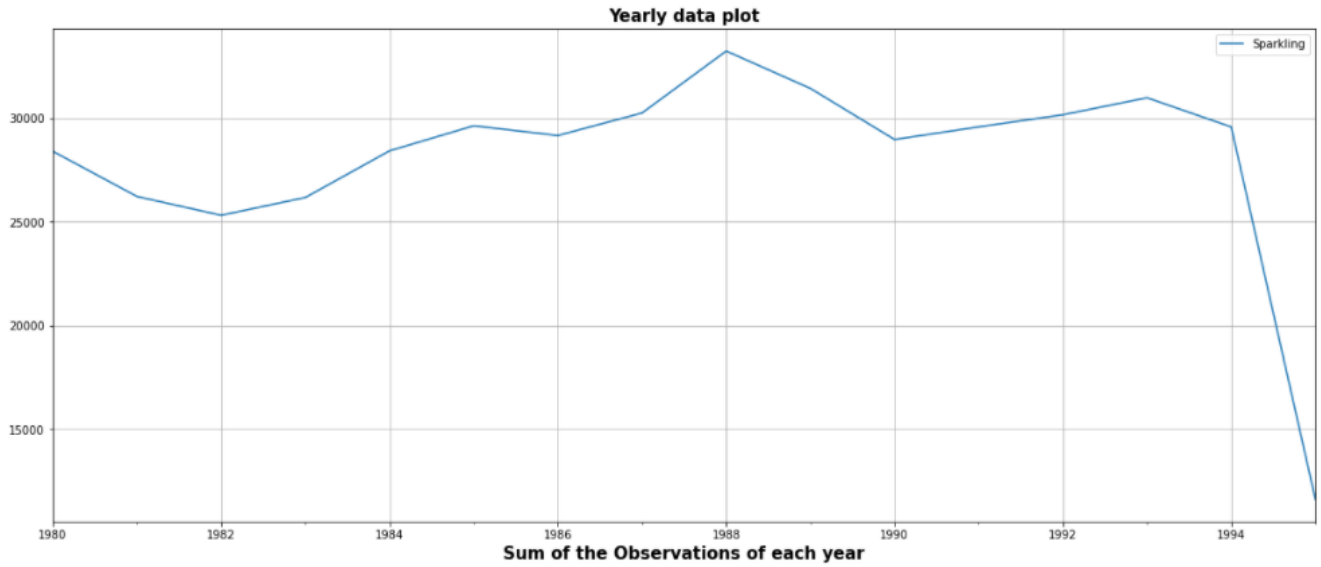


Fig 7: Yearly sales plot

From the above table and graph, we can observe that the year 1982 saw the biggest dip in sales in comparison to any years before or after. After 1982 we see a gradual increase in sales over the years. The dip in the sales after the year 1994 is due to the incomplete sales data for the year 1995.

10. Quarterly sales table and plot of Rose wine

Time_Stamp	Sparkling
1980-03-31	1860.333333
1980-06-30	1520.000000
1980-09-30	2134.333333
1980-12-31	3954.000000
1981-03-31	1562.000000
...	...
1994-09-30	2164.666667
1994-12-31	4371.000000
1995-03-31	1456.333333
1995-06-30	1740.000000
1995-09-30	2031.000000

63 rows × 1 columns

Table 7: Quarterly sales table

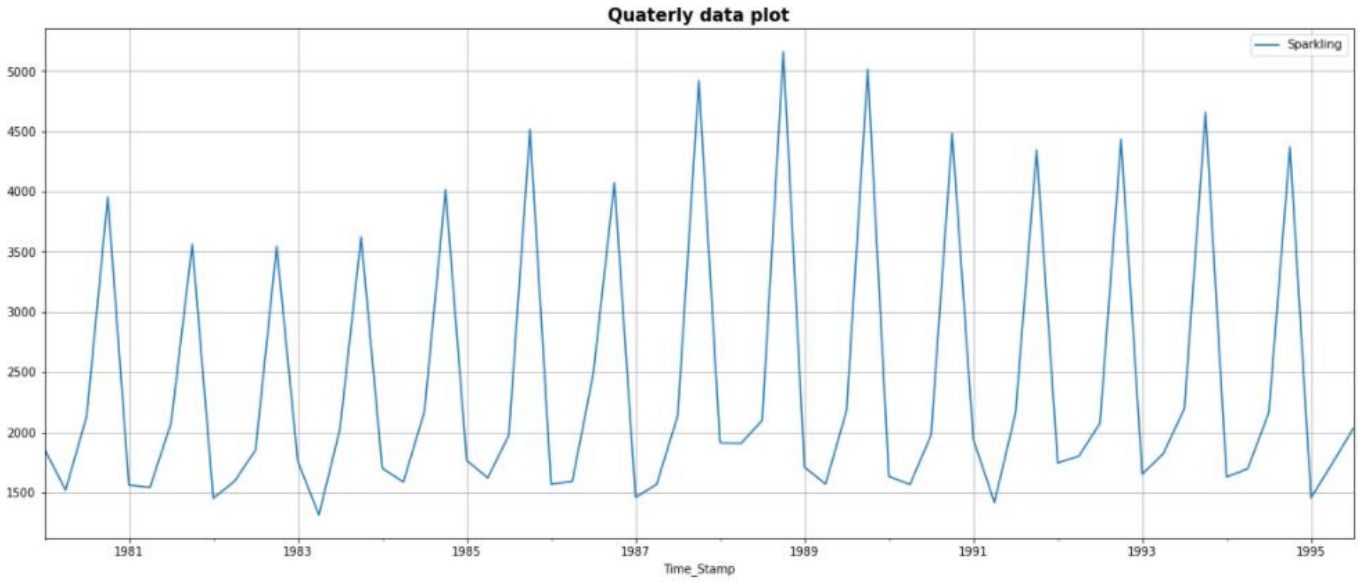


Fig 8: Quarterly sales plot

In the above table and graph, we have found out the quarterly sales of the Sparkling wine. The sales of 3 months make a quarter sale.

11. Daily sales table and plot of Rose wine

Time_Stamp	Sparkling
1987-12-31	7242
1988-12-31	6757
1989-12-31	6694
1993-12-31	6410
1991-12-31	6153
...	...
1985-06-09	0
1985-06-08	0
1985-06-07	0
1985-06-06	0
1987-12-07	0

5661 rows × 1 columns

Table 8: Daily sales table

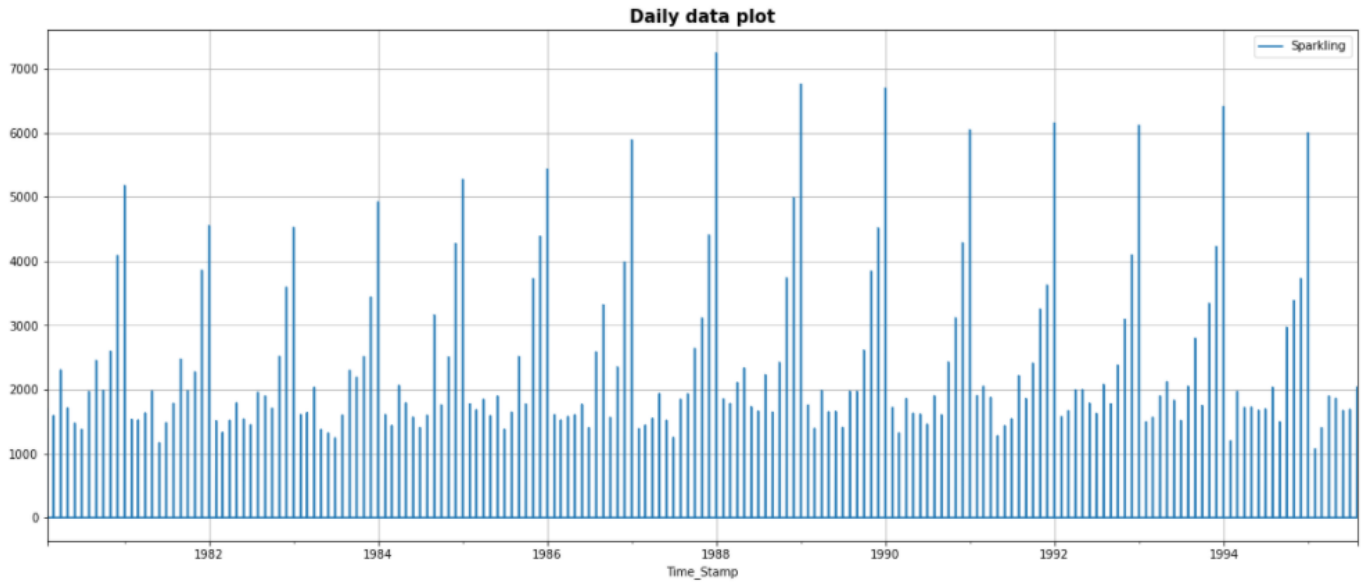


Fig 9: Daily sales plot

In the above table and graph, we have found out the daily sales of the Rose wine. We can observe that the highest number of sales was on the day 1987/12/31 and we see that there a lot of day without even a single unit sold.

c. Model Decomposition

Decomposition of a model is done:

- To understand revenue generation without the quarterly effects

  –De-seasonalize the series

–Estimate and adjust by seasonality

- To compare the long-term movement of the series (Trend) vis-a-vis short-term movement (seasonality) to understand which has the higher influence
- If revenue for multiple sectors is to be compared and if the sectors show non-uniform seasonality, de-seasonalized series needs to be compared.

Mainly there are two kinds of decomposition done, one is additive and the other one is multiplicative. In simple terms, we can identify the additive or multiplicative time series by looking into the magnitude of the seasonal component. If the magnitude of the seasonal component changes with time, then the series is multiplicative. Otherwise, the series is additive.

1. Additive decomposition

- An additive model suggests that the components are added together.
- An additive model is linear where changes over time are consistently made by the same amount. The seasonal correction is added with the trend.
- A linear seasonality has the same frequency (width of the cycles) and amplitude (height of the cycles)

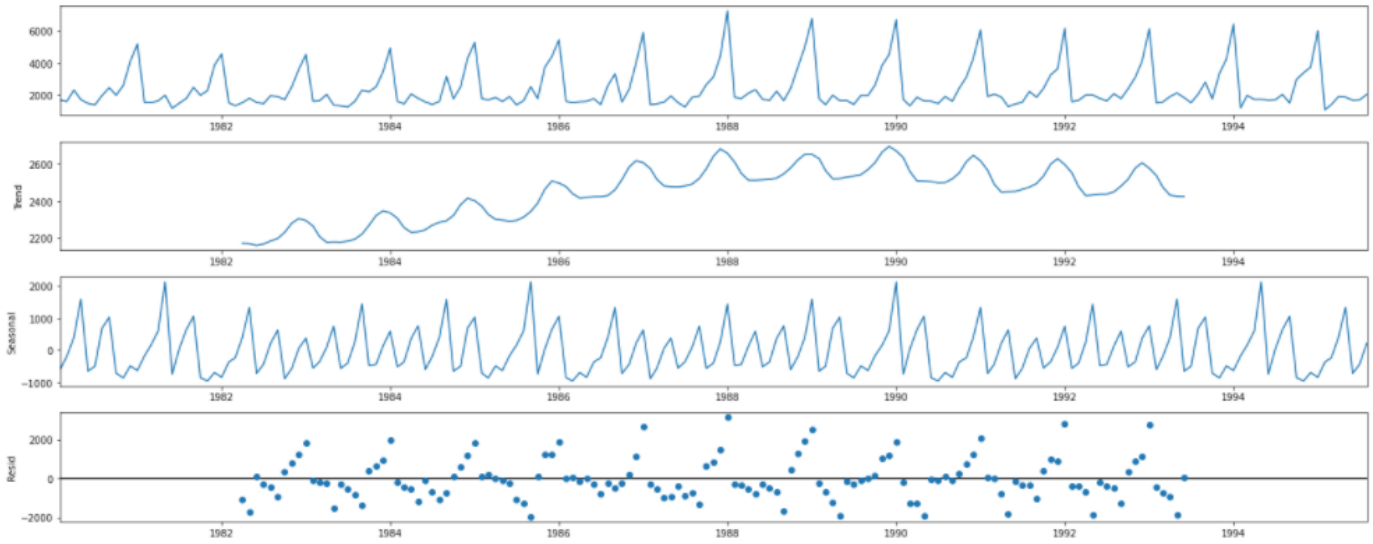


Fig 10: Additive decomposition of the data

From the above given 4 series, we can see that the trend and seasonality of the time series are clearly separated using the additive decomposition. The first series represents the given time series data, the second series represents the trend, the third one seasonality and the fourth one the residuals. It is noted that the residuals plotted in the above series follows a recognized pattern which ensures that the decomposition was not the apt one.

2. Multiplicative decomposition

- A multiplicative model suggests that the components are multiplied together
- A multiplicative model is non-linear.
- The seasonal correction is multiplied with the trend.
- A non-linear seasonality has an increasing or decreasing frequency (width of the cycles) and / or amplitude (height of the cycles) over time

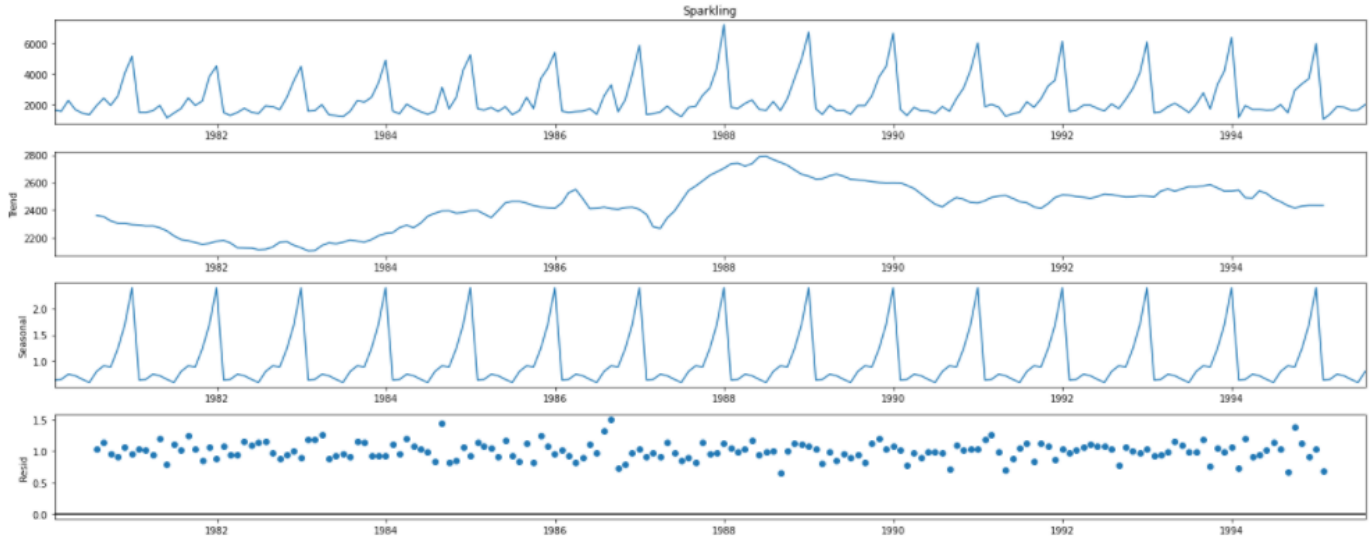
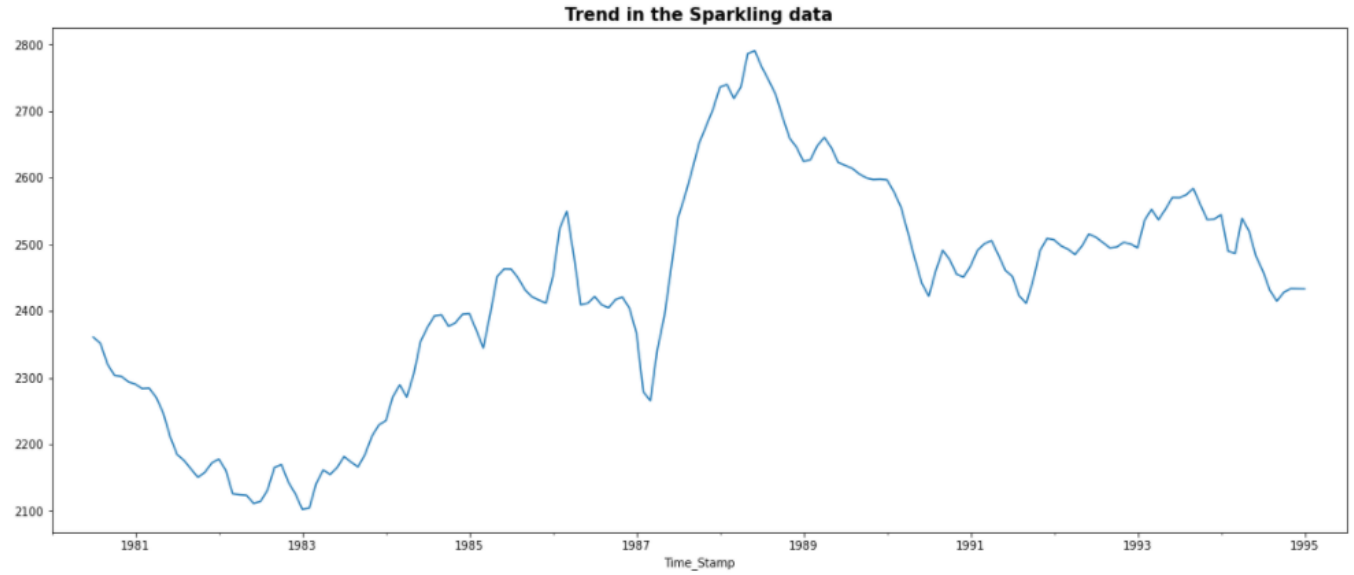


Fig 11: Multiplicative decomposition of the data

From the above given 4 series, we can see that the trend and seasonality of the time series are clearly separated using the multiplicative decomposition. The first series represents the given time series data, the second series represents the trend, the third one seasonality and the fourth one the residuals. It is noted that the residuals plotted in the above series follows a recognized pattern which ensures that the decomposition was accurate enough.

Trend in the given data set



From the above graph, we see that in the initial half, there is an upward trend which seems to reach the peak at a point and then moves in the downward direction. It appeals that in recent times, there happens to be a downward trend which helps in the future forecasts.

Q3. Split the data into training and test. The test data should start in 1991.

The test data should start from 01/01/1991. Hence the data is split accordingly. The below shows the shape of the train and test data.

```
Shape of train data: (132, 1)
Shape of test data: (55, 1)
```

Here we can see that train data consists of 132 rows while test data consists of 55 rows.

First few rows of Training Data

Time_Stamp	Sparkling	time
1980-01-31	1686	1
1980-02-29	1591	2
1980-03-31	2304	3
1980-04-30	1712	4
1980-05-31	1471	5

Last few rows of Training Data

Time_Stamp	Sparkling	time
1990-08-31	1605	128
1990-09-30	2424	129
1990-10-31	3116	130
1990-11-30	4286	131
1990-12-31	6047	132

First few rows of Test Data

Time_Stamp	Sparkling	time
1991-01-31	1902	133
1991-02-28	2049	134
1991-03-31	1874	135
1991-04-30	1279	136
1991-05-31	1432	137

Last few rows of Test Data

Time_Stamp	Sparkling	time
1995-03-31	1897	183
1995-04-30	1862	184
1995-05-31	1670	185
1995-06-30	1688	186
1995-07-31	2031	187

Train data

Test data

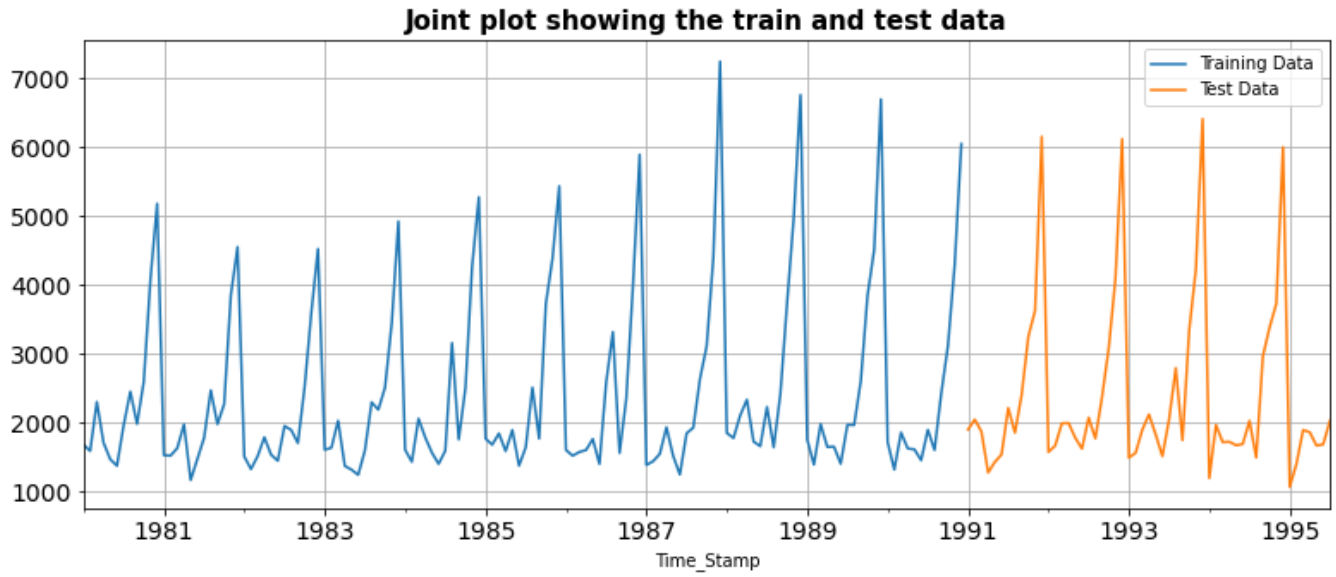


Fig 12: Joint plot showing the test and train data

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

Q4. 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.

1.Linear Regression

We use the linear regression algorithm to construct forecasting models. Linear regression is widely used in practice and adapts naturally to even complex forecasting tasks. The linear regression algorithm learns how to make a weighted sum from its input features. In this model, we are going to plot 'Sparkling' against the order of occurrence.

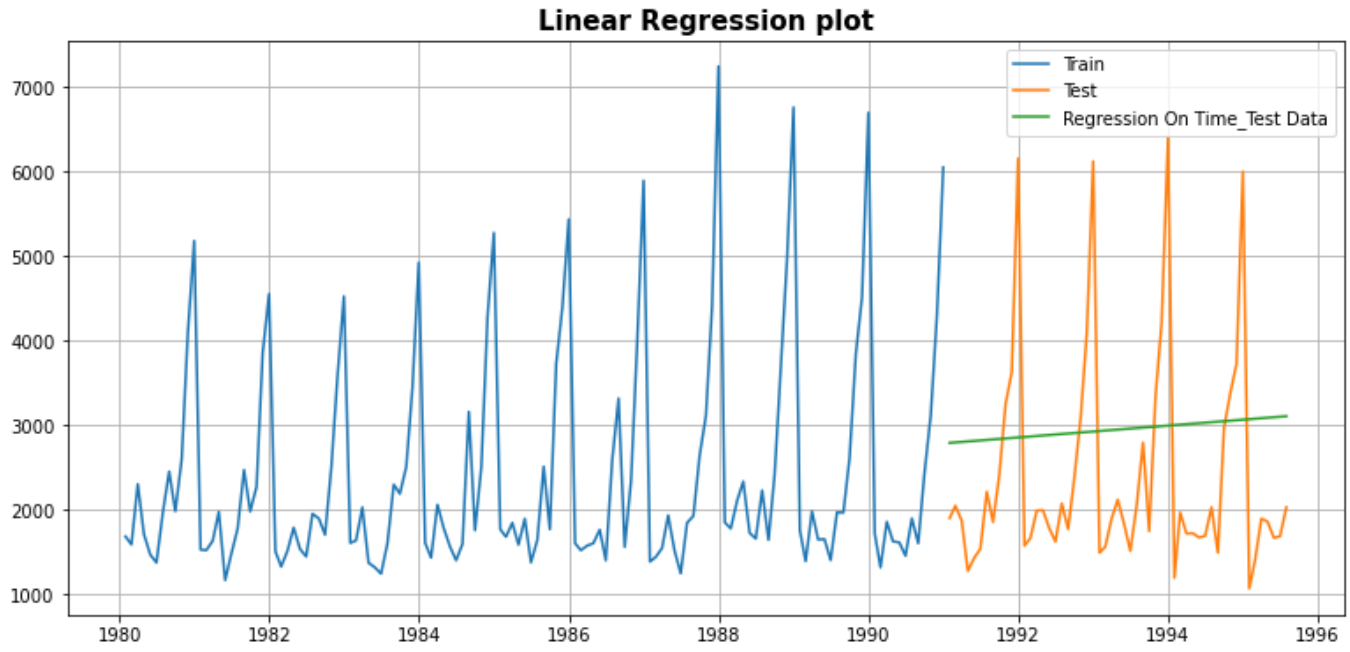


Fig 13: Linear regression plot

Model 1: Linear Regression Evaluation

```
For RegressionOnTime forecast on the Train Data,  RMSE is 1279.322
For RegressionOnTime forecast on the Train Data,  MAPE is 40.050
For RegressionOnTime forecast on the Test Data,  RMSE is 1389.135
For RegressionOnTime forecast on the Test Data,  MAPE is 50.150
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15

From the above results we can see that for linear regression, RMSE on Train data is 1279.322 and RMSE on test data is 1389.135

2.Naive Approach

Simple forecasting methods include naively using the last observation as the prediction or an average of prior observations. It is important and useful to test simple forecast strategies prior to testing more complex models. Simple forecast strategies are those that assume little or nothing about the nature of the forecast problem and are fast to implement and calculate. 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 the same as today, therefore the prediction for day after tomorrow is also today

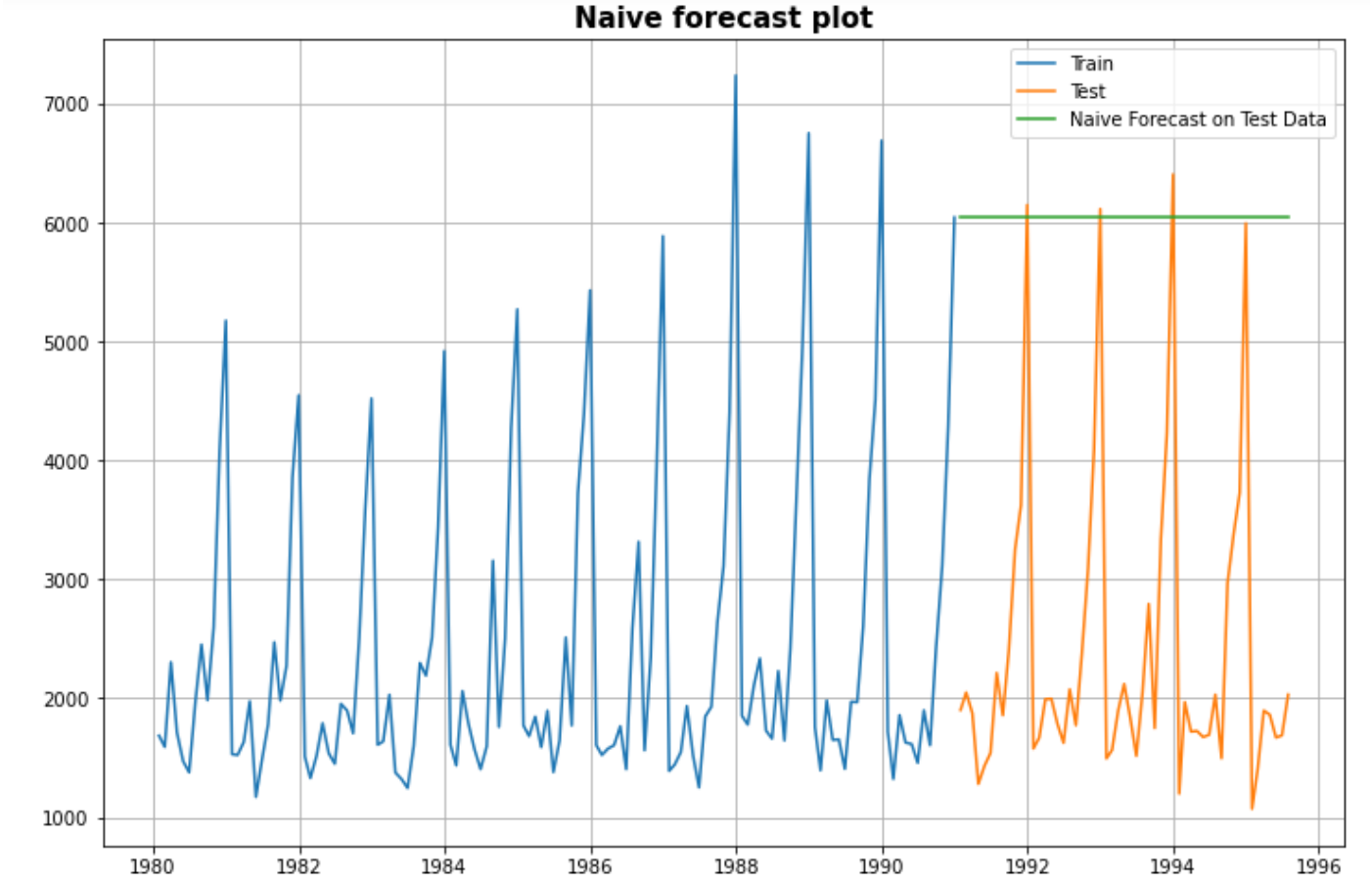


Fig 14: Naïve Forecast plot

Model 2: Naïve Forecast Evaluation

```
For Naive forecast on the Train Data,  RMSE is 14959109.492
For Naive forecast on the Train Data,  MAPE is 153.170
For Naive forecast on the Test Data,  RMSE is 3864.279
For Naive forecast on the Test Data,  MAPE is 152.870
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87

From the above results we can see that for linear regression, RMSE on Train data is 14959109.492 and RMSE on test data is 3864.279

We can infer from the RMSE values and the above graphs that the Naive method and Regression on Time models might not be suited for datasets with high variability. Naive method is best suited for stable datasets. Now we will adopt other techniques for improving the score. Now we will proceed to other techniques to improve our prediction accuracy.

3.Simple Average

This method is very simple where we average the data by months, years or quarters and then calculate the average for the period. Here, in this data, we will forecast by using the average of the training values.

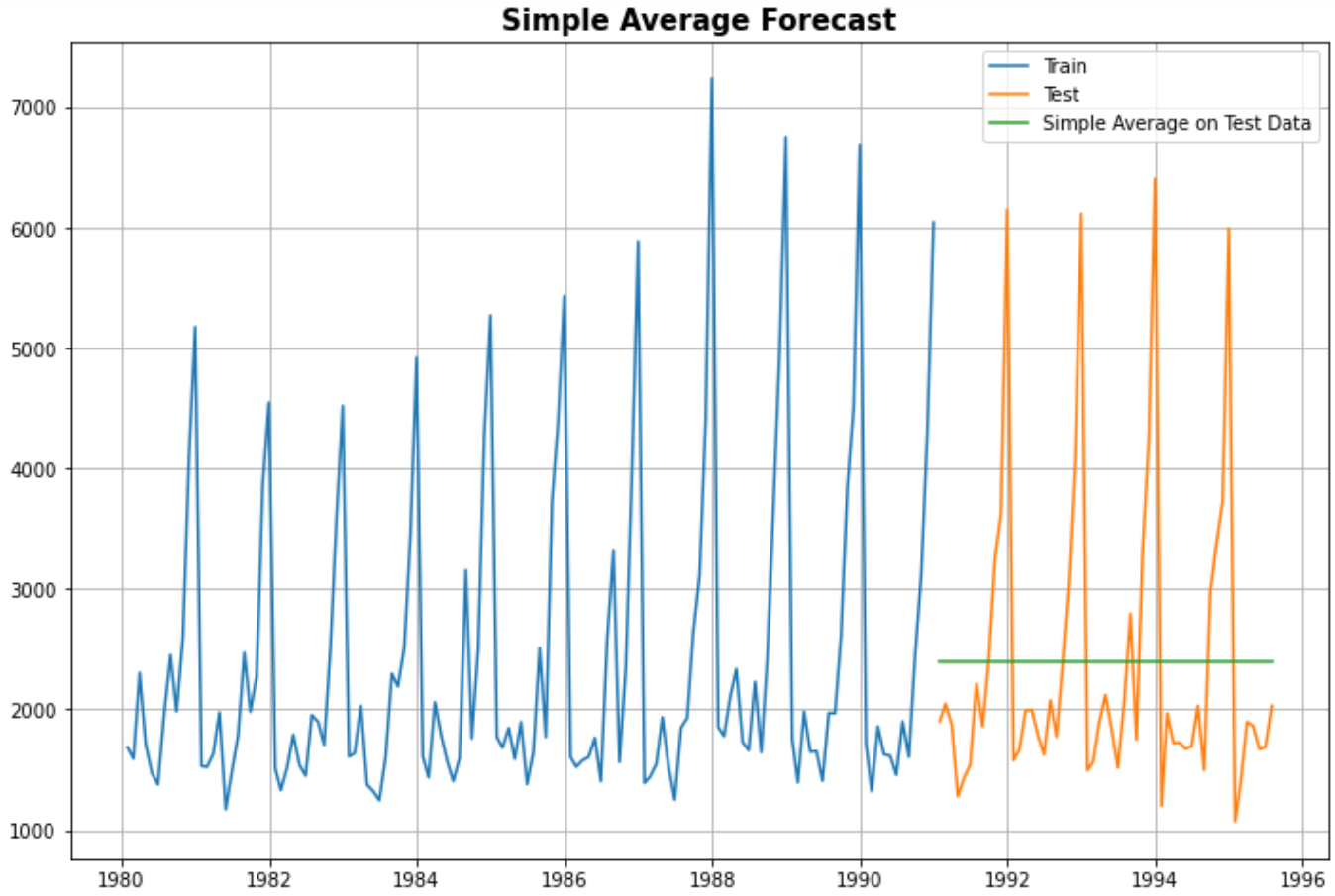


Fig 15: Simple Average Forecast plot

Model 3: Simple Average Forecast Evaluation

```
For Simple Average forecast on the Test Data,  RMSE is 1275.082
For Simple Average forecast on the Test Data,  MAPE is 38.900
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90

From the above results we can see that for linear regression, RMSE on Train data is 1275.082

4.Moving Average

Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting. Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series. This method is used for data where seasonal and cyclic variation is present. Here we will be calculating rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here. In this method, we are going to average over the entire data.

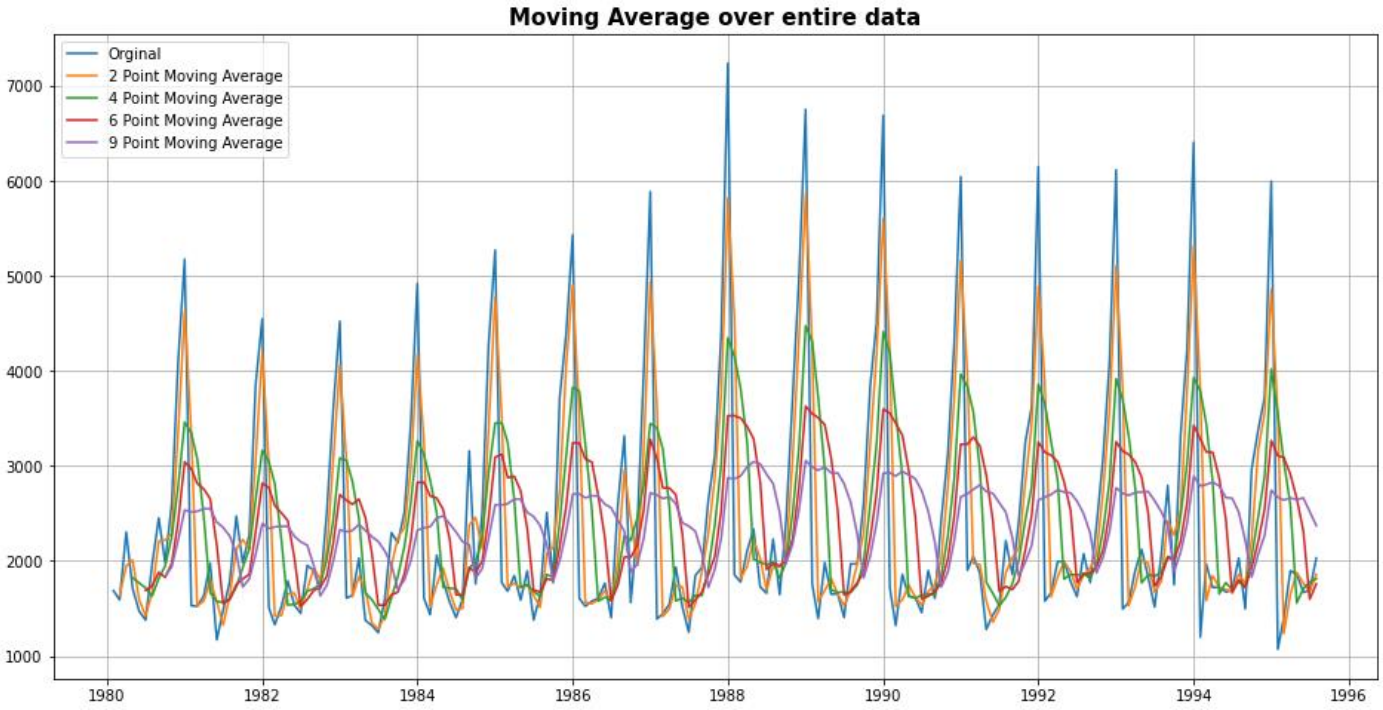


Fig 16: Moving Average over entire data

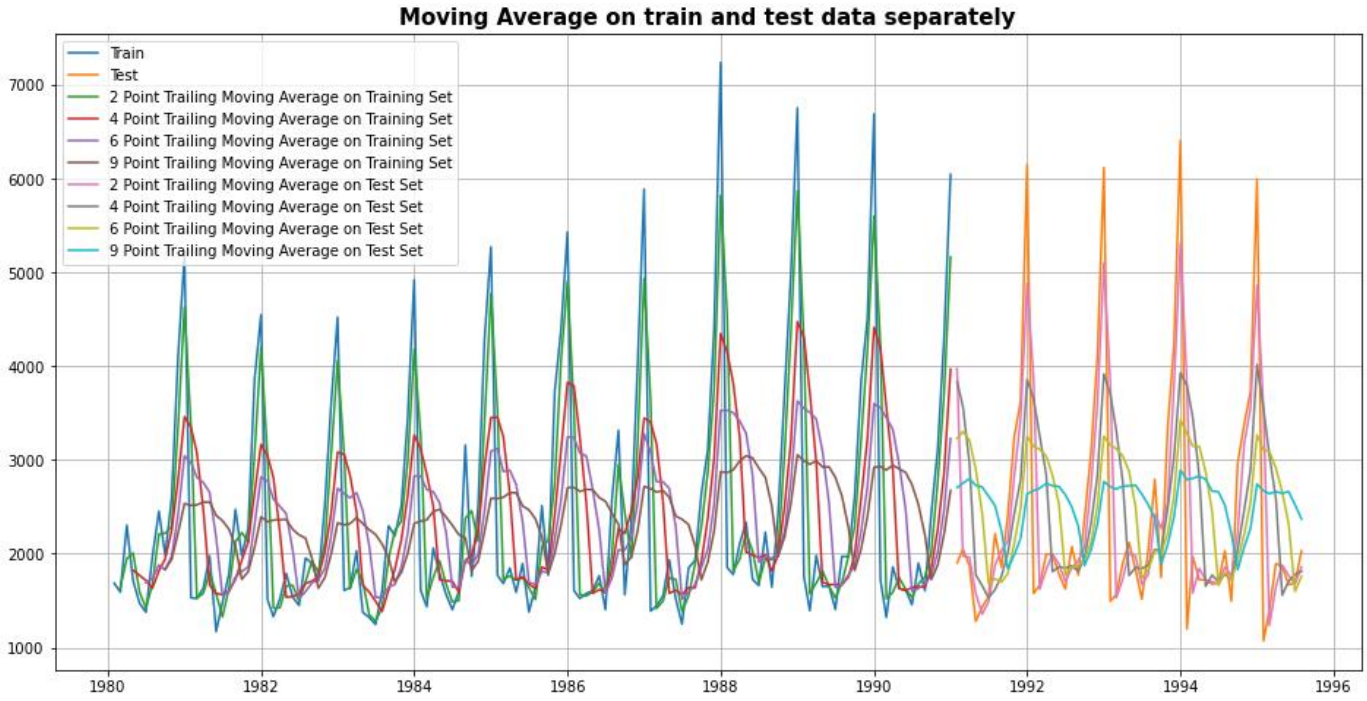


Fig 17: Moving Average over train and test data separately

Model 4: Moving Average Forecast Evaluation

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

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
6pointTrailingMovingAverage	1283.927428	43.86
9pointTrailingMovingAverage	1346.278315	46.86

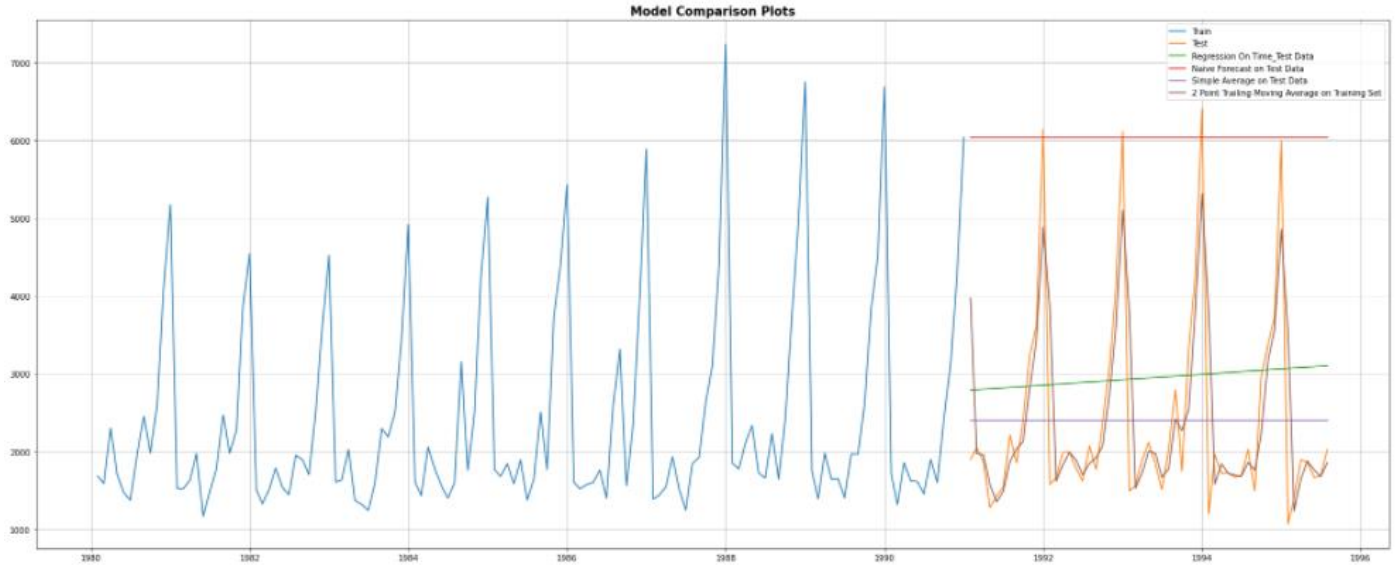
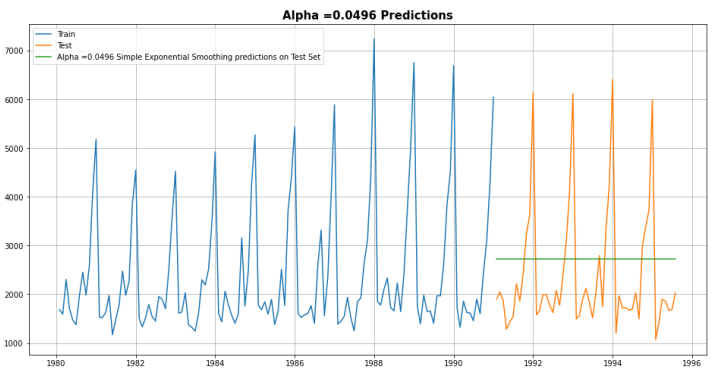


Fig 18: Model Comparison plots

Here we have plotted all the models done so far and compared the time Series plots.

5.Simple Exponential Smoothing



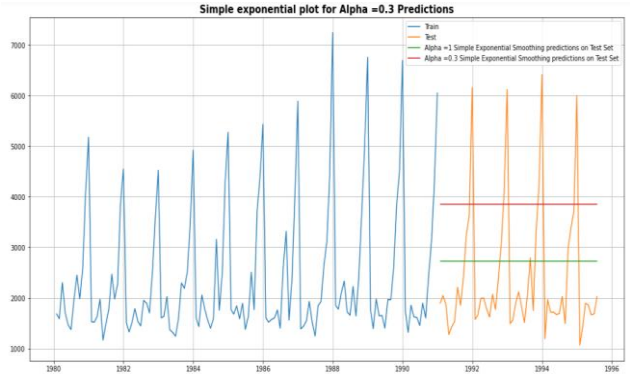


Fig 19: Simple exponential plots for alpha =0.0496 and 0.3 respectively

Model 5: Simple Exponential Smoothing Evaluation

```
For Alpha =0.0496 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.035

For Alpha=0.3,Simple Exponential Smoothing Model forecast on the test data,RMSE= [1935.5071321027176]
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
6pointTrailingMovingAverage	1283.927428	43.86
9pointTrailingMovingAverage	1346.278315	46.86
Alpha=0.0496,SimpleExponentialSmoothing	1316.034674	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	NaN

6.Double Exponential Smoothing (Holt's Model)

Two parameters Alpha and Beta are estimated in this model. Level and Trend are accounted for in this model.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1592.292788	18259.110704
1	0.3	0.4	1682.573828	26069.841401
2	0.3	0.5	1771.710791	34401.512440
3	0.3	0.6	1848.576510	42162.748095
4	0.3	0.7	1899.949006	47832.397419
...	...	...	...	...
59	1.0	0.6	1753.402326	49327.087977
60	1.0	0.7	1825.187155	52655.765663
61	1.0	0.8	1902.013709	55442.273880
62	1.0	0.9	1985.368445	57823.177011
63	1.0	1.0	2077.672157	59877.076519

64 rows × 4 columns

Now we will sort the data frame in the ascending order of 'Test RMSE'

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1592.292788	18259.110704
8	0.4	0.3	1569.338606	23878.496940
1	0.3	0.4	1682.573828	26069.841401
16	0.5	0.3	1530.575845	27095.532414
24	0.6	0.3	1506.449870	29070.722592

Table 9: Double exponential smoothing table with ascending order of RMSE

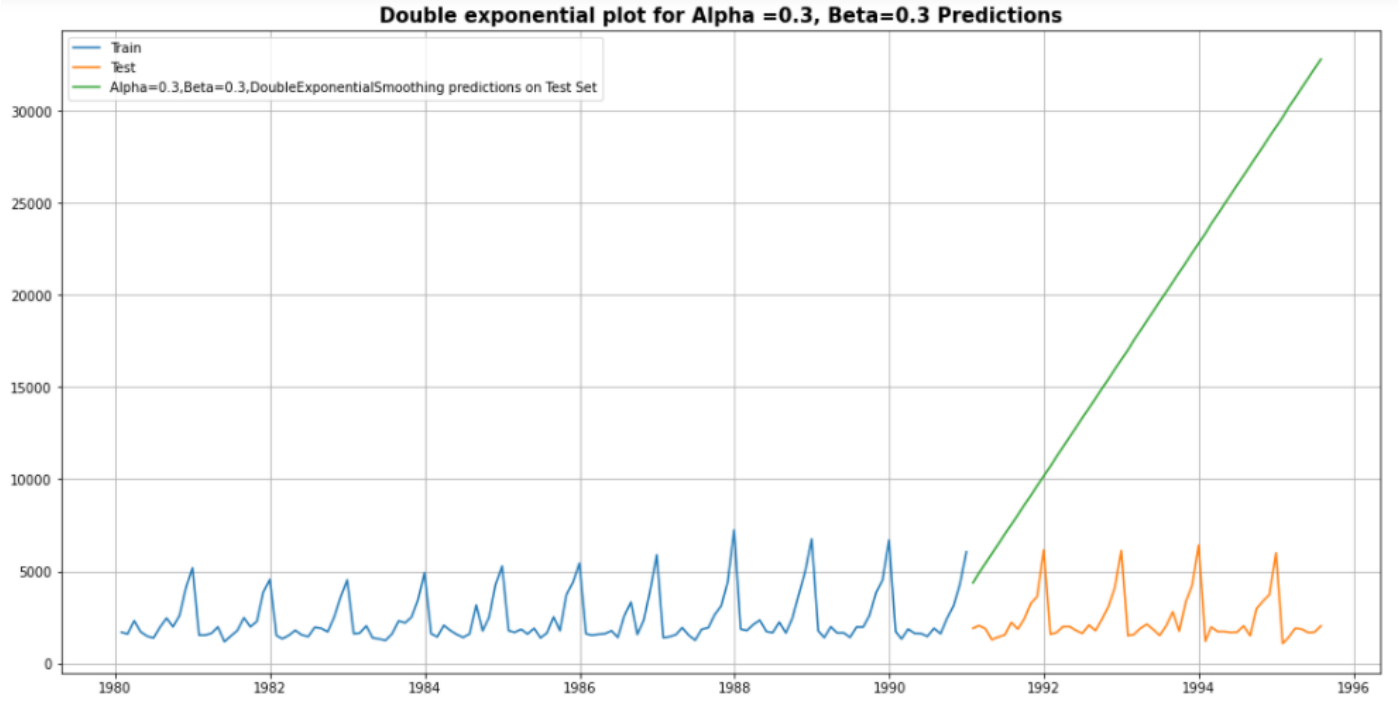


Fig 20: Double exponential smoothing plot for alpha =0.1 and beta= 0.1

Model 6: Double Exponential Smoothing Evaluation

```
For Alpha=0.3,Beta=0.3,Double Exponential Smoothing Model forecast on the test data,RMSE= [18259.11070404971]
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
6pointTrailingMovingAverage	1283.927428	43.86
9pointTrailingMovingAverage	1346.278315	46.86
Alpha=0.0496,SimpleExponentialSmoothing	1316.034674	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	NaN
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	NaN

7.Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters alpha, beta and gamma are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Auto fit parameters

```
{'smoothing_level': 0.11107308290744182,
 'smoothing_trend': 0.06167745801641925,
 'smoothing_seasonal': 0.39488777704116057,
 'damping_trend': nan,
 'initial_level': 1639.5306320456996,
 'initial_trend': -13.803739314239138,
 'initial_seasons': array([1.04411064, 1.00095858, 1.40459398, 1.20906039, 0.96413947,
        0.96754964, 1.3048211 , 1.69841076, 1.37034155, 1.81659752,
        2.84708154, 3.62462473]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

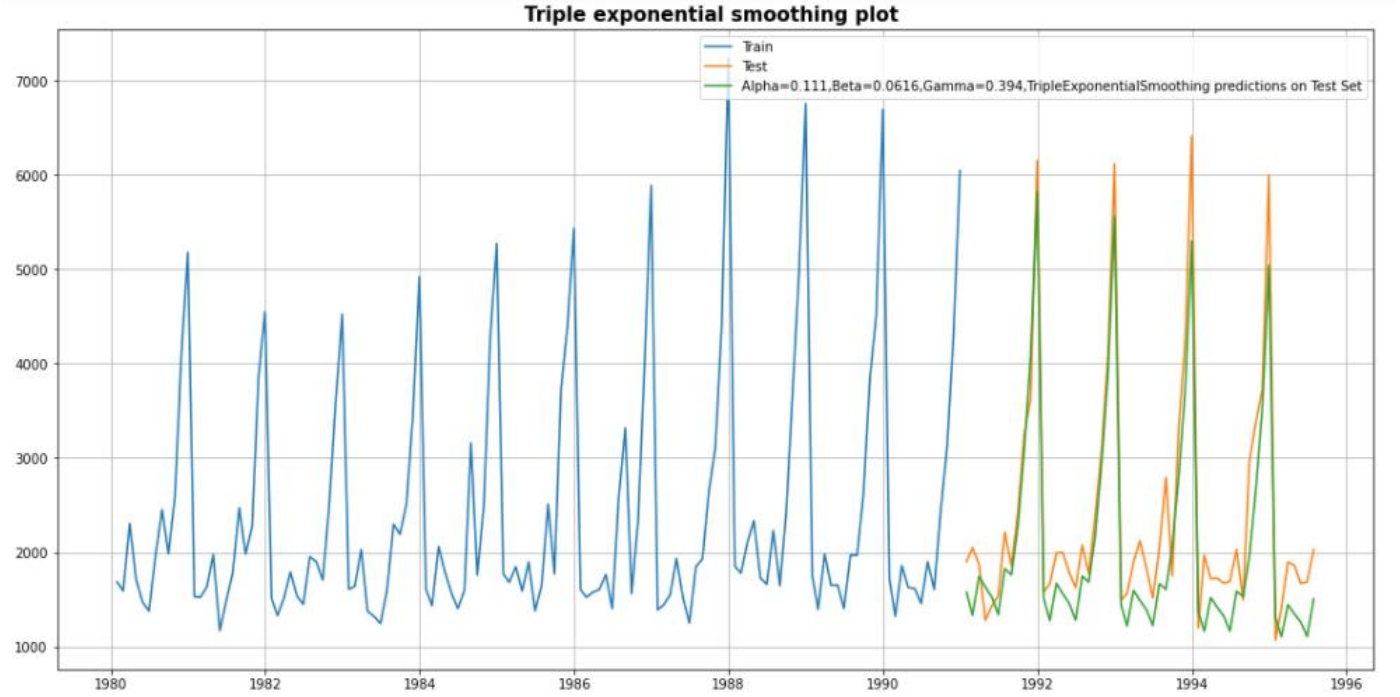


Fig 21: Triple exponential smoothing plots

Model 7: Triple Exponential Smoothing Evaluation

```
For Alpha=0.111,Beta=0.0616,Gamma=0.394, Triple Exponential Smoothing Model forecast on the Test Data,  RMSE is 469.432
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
6pointTrailingMovingAverage	1283.927428	43.86
9pointTrailingMovingAverage	1346.278315	46.86
Alpha=0.0496,SimpleExponentialSmoothing	1316.034674	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	NaN
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	NaN
Alpha=0.111,Beta=0.0616,Gamma=0.394,TripleExponentialSmoothing	469.432003	NaN

8. Brute Force - Triple Exponential Smoothing

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.3	0.3	0.3	404.513320	3.927862e+02
1	0.3	0.3	0.4	402.088628	9.513202e+02
2	0.3	0.3	0.5	408.282432	1.470487e+03
3	0.3	0.3	0.6	428.631668	2.181724e+03
4	0.3	0.3	0.7	468.958530	3.513351e+03
...	...	...	...	...	...
507	1.0	1.0	0.6	153394.791826	7.989790e+05
508	1.0	1.0	0.7	94040.964958	1.074413e+06
509	1.0	1.0	0.8	102196.953755	5.010607e+06
510	1.0	1.0	0.9	77924.294413	4.318265e+05
511	1.0	1.0	1.0	239917.432847	1.254280e+05

Table 10: Brute force -Triple exponential smoothing table with various values for alpha, beta and gamma

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.3	0.3	0.3	404.513320	392.786198
8	0.3	0.4	0.3	424.828055	410.854547
65	0.4	0.3	0.4	435.553595	421.409170
296	0.7	0.8	0.3	700.317756	518.188752
130	0.5	0.3	0.5	498.239915	542.175497

Table 11: Brute force -Triple exponential smoothing table with increasing values of RMSE

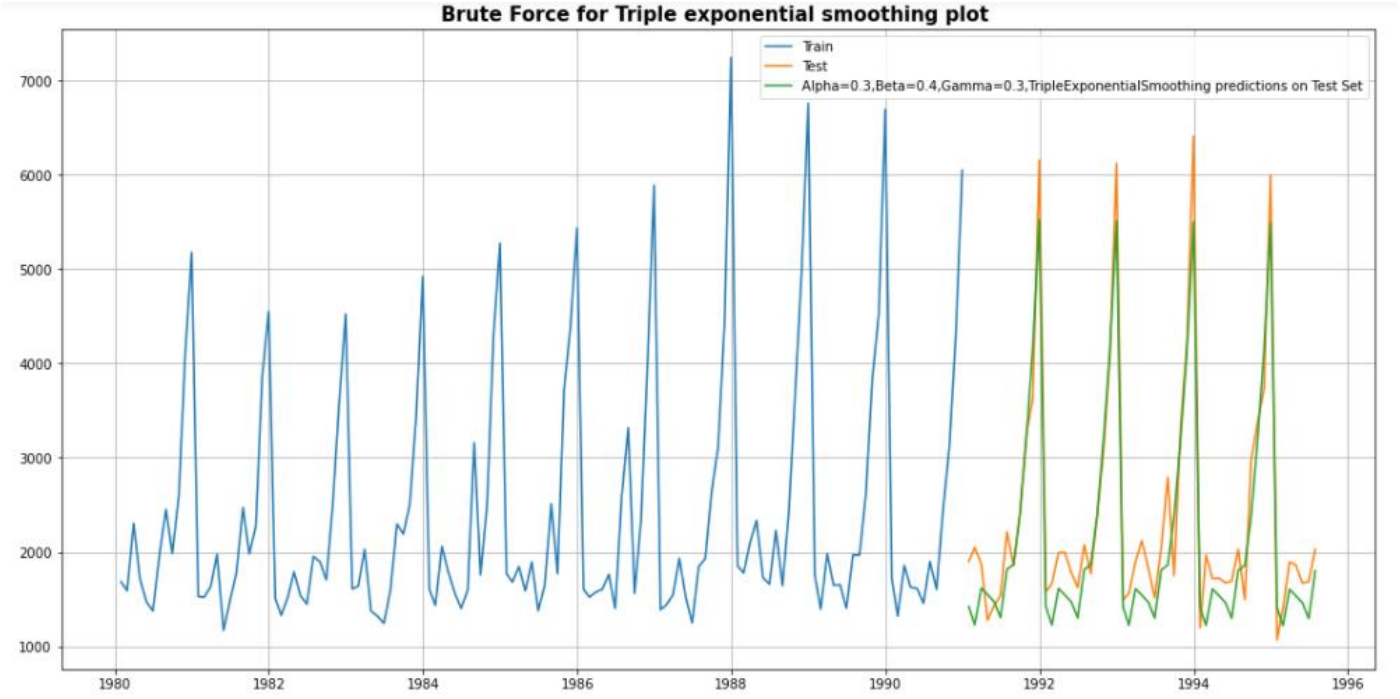


Fig 22: Brute force for triple exponential smoothing plot

Model 8: Brute Force- Triple Exponential Smoothing Evaluation

```
For Alpha=0.3,Beta=0.3,Gamma=0.3, Brute Force-Triple Exponential Smoothing Model forecast on the Test Data,RMSE is 392.786
```

	Test RMSE	Test MAPE
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
6pointTrailingMovingAverage	1283.927428	43.86
9pointTrailingMovingAverage	1346.278315	46.86
Alpha=0.0496,SimpleExponentialSmoothing	1316.034674	45.47
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	NaN
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	NaN
Alpha=0.111,Beta=0.0616,Gamma=0.394,TripleExponentialSmoothing	469.432003	NaN
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198	NaN

Table 12: Brute force -Triple exponential smoothing table with increasing values of RMSE

Smoothing models conclusion

For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double

Exponential Smoothing. Here we see that Brute force-triple exponential had the lowest RMSE of 10.945. However, we had gone on to build different models on the data and have compared these models with the best RMSE value on the test data. The following table shows the different smoothing models in the decreasing order of RMSE.

	Test RMSE	Test MAPE
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198	NaN
Alpha=0.111,Beta=0.0616,Gamma=0.394,TripleExponentialSmoothing	469.432003	NaN
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
SimpleAverageModel	1275.081804	38.90
6pointTrailingMovingAverage	1283.927428	43.86
Alpha=0.0496,SimpleExponentialSmoothing	1316.034674	45.47
9pointTrailingMovingAverage	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	NaN
NaiveModel	3864.279352	152.87
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	NaN

Table 13: Different smoothing models in the decreasing order of RMSE.

Q5. 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.

- ACF plot

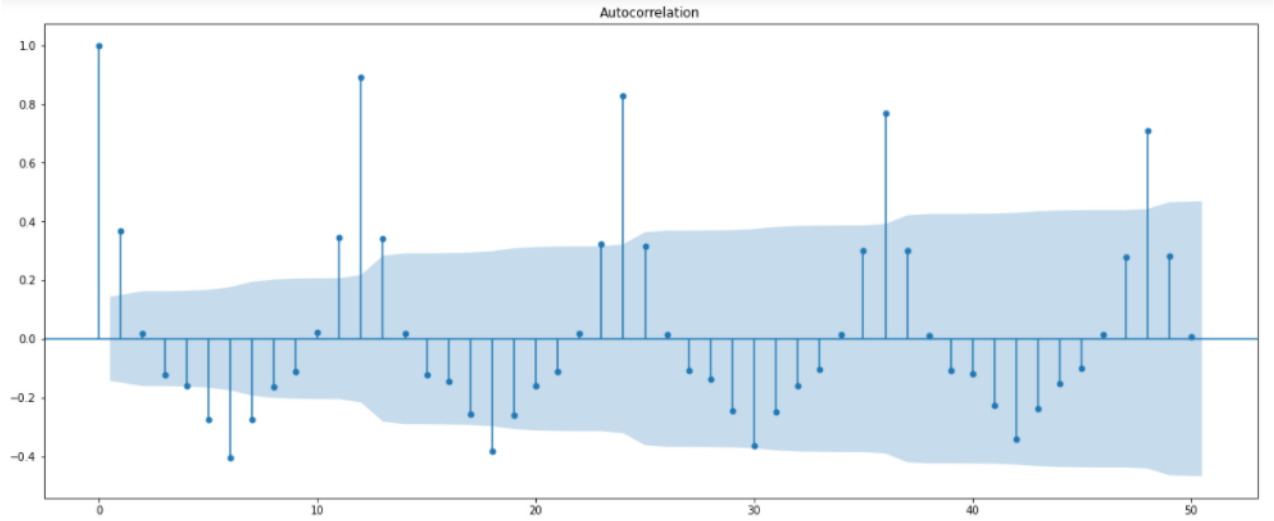


Fig 23: ACF plot

- **PACF plot**

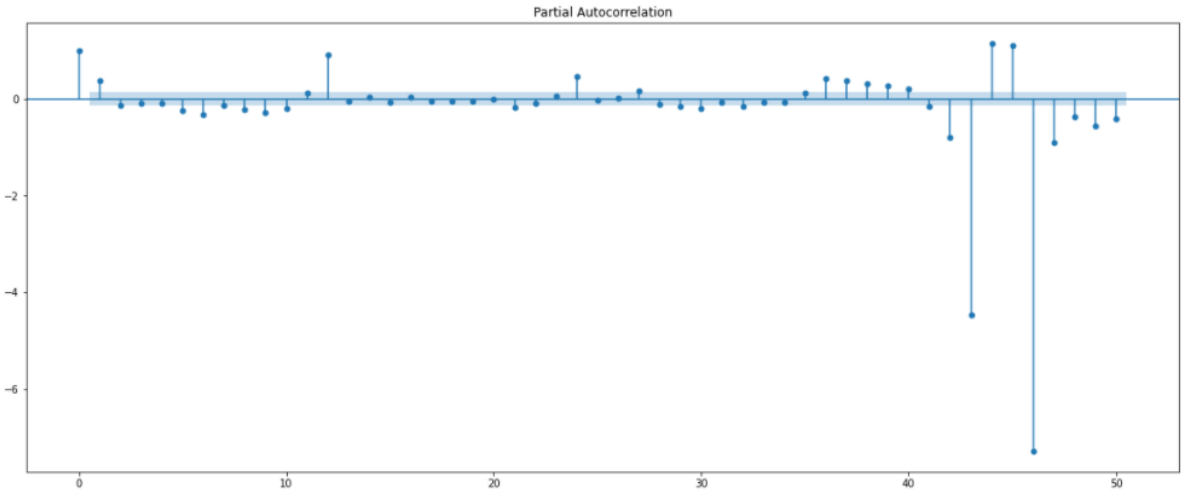


Fig 24: PACF plot

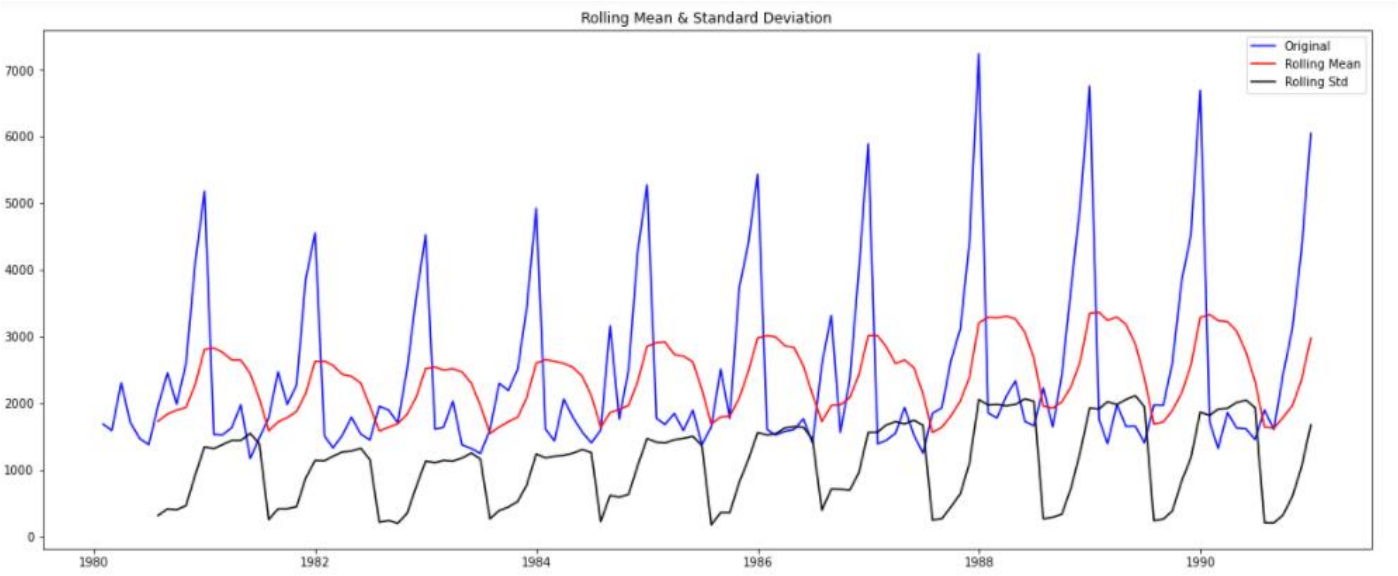
ACF and PACF plots are done with 95% confidence interval bands. We can see that seasonality after certain lags is visible is after every 12th Month. Hence, we do the stationarity test (Dickey-Fuller) for the series

Test for stationarity

Null and Alternate Hypothesis for the Augmented Dickey Fuller Test.

H0: The series is not stationary.

H1: The series is stationary.

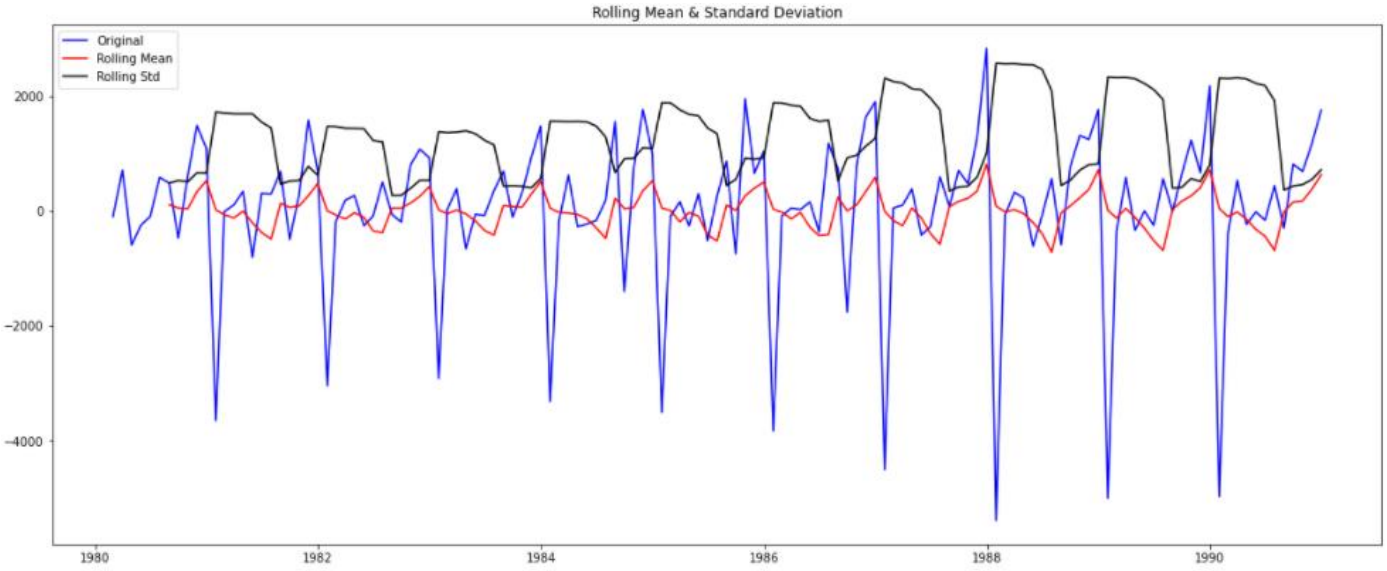


Series is not stationary with original form at alpha = 0.05.

```
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
```

From the above table, we see that p-value is not less than 0.05, therefore we fail to reject the null hypothesis.

As we have seen from the above graph and p-value, the series is not stationary. Therefore, we will check for stationarity after taking first order differencing.




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

From the above table, it is very clear that p-value is less than 0.05, thereby we can reject the null hypothesis and we find that series is stationary post first order differencing at alpha = 0.05.

Q6. 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**

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. Here the given series is not stationary and therefore differentiation is necessary. For Auto-ARIMA, we choose the best p, d and q parameters by looking at the lowest corresponding Akaike Information Criterion (AIC) values.

	param	AIC
8	(2, 1, 2)	2210.622120
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)	2264.906438
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

Table 14: Different combinations of parameter values for ARIMA in the ascending order of AIC.

The best model is predicted by the lowest value of AIC. From the above table we see that p=2, d=1 and q=2 has the lowest AIC of 2210.622120

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.311
Method:	css-mle	S.D. of innovations	1013.283
Date:	Sat, 19 Feb 2022	AIC	2210.622
Time:	17:28:04	BIC	2227.873
Sample:	02-29-1980 - 12-31-1990	HQIC	2217.632

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5852	0.518	10.788	0.000	4.570	6.600
ar.L1.D.Sparkling	1.2699	0.075	17.043	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9966	0.042	-47.013	0.000	-2.080	-1.913
ma.L2.D.Sparkling	0.9966	0.043	23.421	0.000	0.913	1.080

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1334	-0.7074j	1.3361	-0.0888
AR.2	1.1334	+0.7074j	1.3361	0.0888
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	1.0033	+0.0000j	1.0033	0.0000

The above chart shows the Arima model results for p=0, d=1, q=2. For this particular Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1 and 2. The values of p, d, and q are calculated by giving a suitable range of values for p, d and q.

Automated ARIMA model Evaluation

RMSE of Automated ARIMA(2,1,2) on testing data: 1374.2963869630776

	RMSE	MAPE
ARIMA(2,1,2)	1374.296387	48.34

- ## 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. Since the given dataset possess seasonality, we can take the liberty to build the model with SARIMA. For an Auto-SARIMA, the parameters p, q, P and Q are selected based on the lowest Akaike Information Criterion (AIC).

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934564
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)	1557.340402

Table 15: Different combinations of parameter values for SARIMA in the ascending order of AIC.

Here we take the seasonality to be 12 since it is a yearly forecast. The best model is predicted by the lowest value of AIC. From the above table we see that p=1, d=1, q=2 and P=1, D=0, Q=2 and seasonality=12 has the lowest AIC of 1555.584247

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.670
Date:                            Sat, 19 Feb 2022   AIC                           1557.340
Time:                                    17:35:57   BIC                           1578.496
Sample:                                         0   HQIC                          1565.911
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6503      0.247     -2.634      0.008      -1.134      -0.166
ar.L2         -0.0455      0.142     -0.321      0.748      -0.324       0.233
ma.L1         -0.1029      0.226     -0.455      0.649      -0.547       0.341
ma.L2         -0.7073      0.186     -3.804      0.000      -1.072      -0.343
ar.S.L12       1.0451      0.014     75.971      0.000       1.018       1.072
ma.S.L12      -1.1707      0.385     -3.041      0.002      -1.925      -0.416
ma.S.L24      -0.2373      0.142     -1.674      0.094      -0.515       0.040
sigma2      8.292e+04   3.17e+04      2.615      0.009    2.08e+04    1.45e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 10.38
Prob(Q):                              0.99   Prob(JB):                          0.01
Heteroskedasticity (H):               1.52   Skew:                              0.32
Prob(H) (two-sided):                  0.22   Kurtosis:                          4.41
===================================================================================

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

The above chart shows the Arima model results for p=1, d=1, q=2, P=2, D=0, Q=2 and seasonality=12. The values of p, d, q, P, D and Q are calculated by giving a suitable range of values for p, d and q.

Automated ARIMA model Evaluation

```
RMSE of Automated SARIMA(1,1,2)(2,0,2,12) on testing data: 555.4615117611364
```

	RMSE	MAPE
ARIMA(2,1,2)	1374.296387	48.34
SARIMA(1,1,2)(1,0,2,12)	555.461512	20.08

Q7. 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 model based of cut-off points of ACF and PACF**

For this particular Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1 and 2. We are also considering the errors from the auto-regression of the first lag also. The values of p and q are calculated by looking at the ACF and the PACF plots.

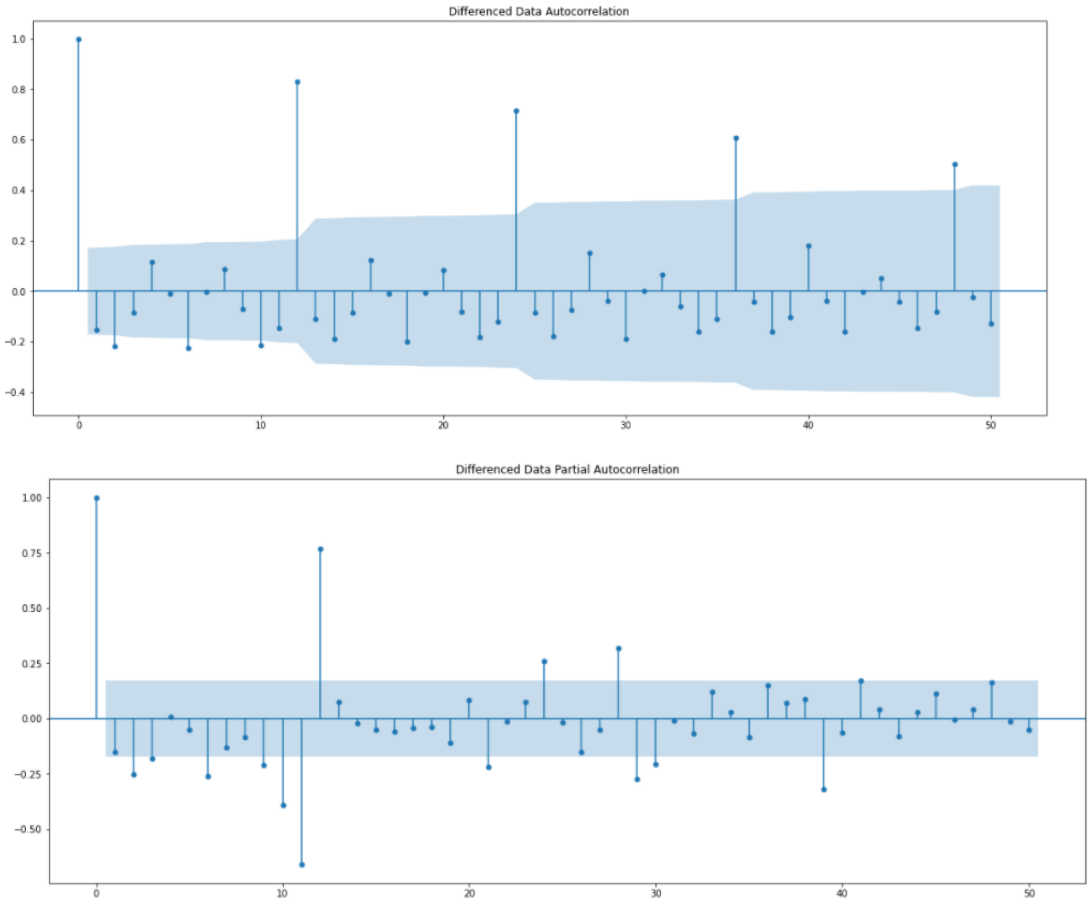


Fig 25: ACF and PACF plots for ARIMA model

From the above graphs of differenced auto-correlation and differenced partial auto correlation, we can see that p= 3, d=1, q=2. 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. Since, we have taken one difference of the series to be series, d=1.Here, we have taken alpha=0.05. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

ARIMA model based of cut-off points of ACF and PACF evaluation

```
RMSE of ARIMA model based on cut off points(3,1,2)on testing data: 1379.0006742208946
```

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.464
Method:	css-mle	S.D. of innovations	1106.121
Date:	Sun, 20 Feb 2022	AIC	2228.927
Time:	12:25:12	BIC	2249.054
Sample:	02-29-1980 - 12-31-1990	HQIC	2237.106

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9850	3.643	1.643	0.100	-1.156	13.126
ar.L1.D.Sparkling	-0.4420	5.81e-06	-7.6e+04	0.000	-0.442	-0.442
ar.L2.D.Sparkling	0.3079	1.51e-05	2.04e+04	0.000	0.308	0.308
ar.L3.D.Sparkling	-0.2501	1.31e-05	-1.92e+04	0.000	-0.250	-0.250
ma.L1.D.Sparkling	-0.0006	0.020	-0.028	0.978	-0.040	0.039
ma.L2.D.Sparkling	-0.9994	0.020	-49.246	0.000	-1.039	-0.960

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	1.1157	-1.6595j	1.9996	-0.1558
AR.3	1.1157	+1.6595j	1.9996	0.1558
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0006	+0.0000j	1.0006	0.5000

	RMSE	MAPE
ARIMA(2,1,2)	1374.296387	48.34
SARIMA(1,1,2)(1,0,2,12)	555.461512	20.08
ARIMA(3,1,2)	1379.000674	49.31

- **SARIMA model based of cut-off points of ACF and PACF**

For this particular Seasonal Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1 and 2. We are also considering the errors from the auto-regression of the first lag. The values of p, q, P and Q are calculated by looking at the ACF and the PACF plots. In this particular model we have to find the p,d,q and the P,D,Q values manually by plotting the ACF and PACF plots. The p,d,q values will be same as the ARIMA model where we selected the values manually by looking at the ACF and PACF plot where we took first order differencing which are (3,1,2), but for reference purposes plots are included below. Now coming to the seasonal parameter of P, D, Q we have to derive these by getting rid of trend from the data as much as possible or that which seems ideal.

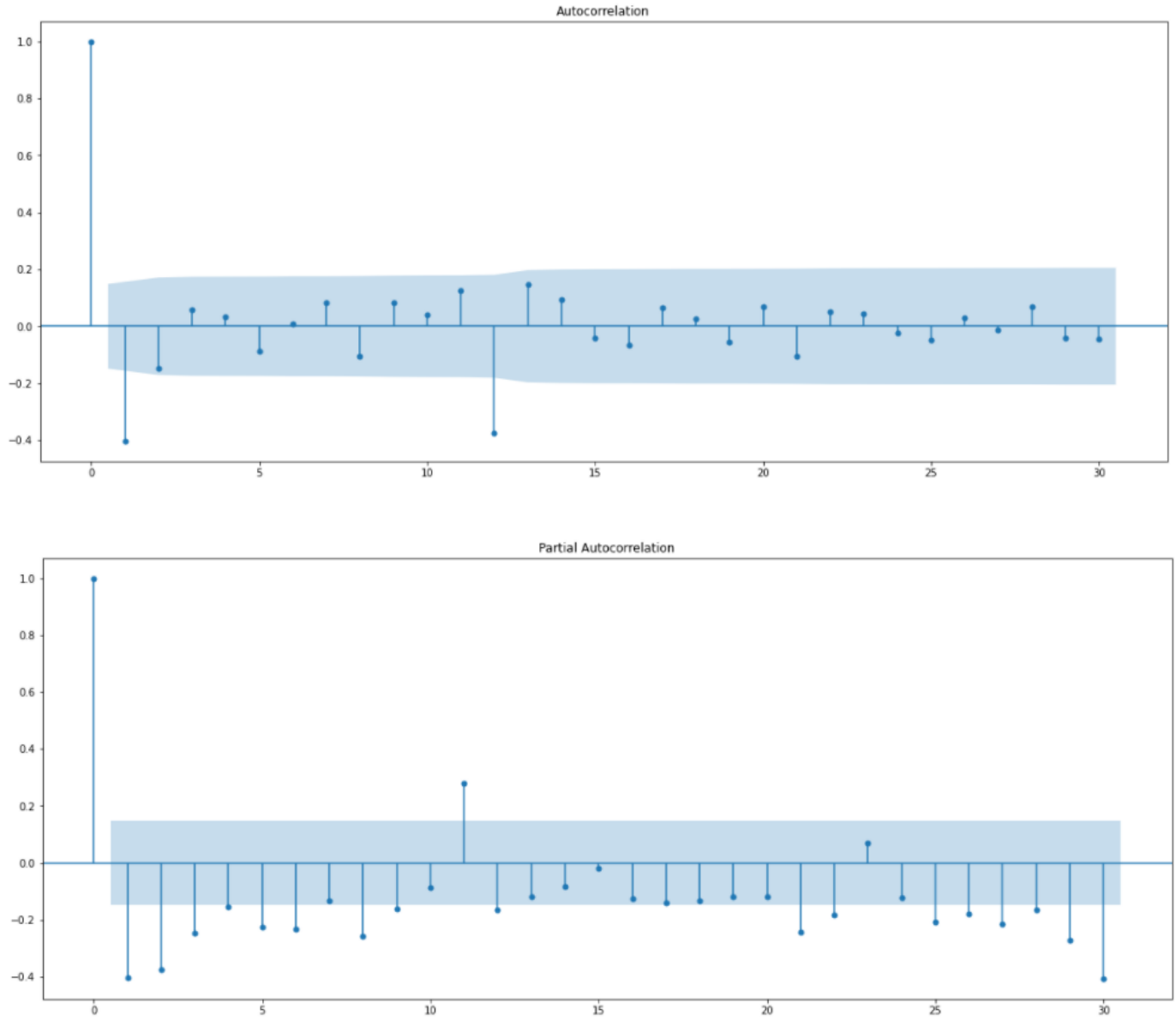


Fig 26: ACF and PACF plots for SARIMA model

From the above graphs of differenced auto-correlation and differenced partial auto correlation, we can see that P= 3, D=0, Q=1. The Auto-Regressive parameter in an ARIMA model is 'Q' 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. Since, we have taken one difference of the series to be series, D=1.Here, we have taken alpha=0.05. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

SARIMA model based of cut-off points of ACF and PACF evaluation

```
RMSE of SARIMA model based on cut off points(2,1,1)(3,1,1,12)on testing data: 347.48994993195447
```

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 1)x(3, 1, 1, 12)	Log Likelihood	-607.074
Date:	Sun, 20 Feb 2022	AIC	1230.148
Time:	13:38:37	BIC	1249.304
Sample:	0	HQIC	1237.834
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1902	0.155	1.227	0.220	-0.114	0.494
ar.L2	-0.0804	0.143	-0.561	0.575	-0.361	0.200
ma.L1	-0.9486	0.076	-12.547	0.000	-1.097	-0.800
ar.S.L12	-0.5560	0.854	-0.651	0.515	-2.230	1.118
ar.S.L24	-0.2759	0.326	-0.847	0.397	-0.914	0.363
ar.S.L36	-0.1403	0.160	-0.877	0.381	-0.454	0.173
ma.S.L12	0.1367	0.861	0.159	0.874	-1.550	1.824
sigma2	1.887e+05	2.79e+04	6.767	0.000	1.34e+05	2.43e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	24.25
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.68	Skew:	0.80
Prob(H) (two-sided):	0.32	Kurtosis:	5.15

	RMSE	MAPE
ARIMA(2,1,2)	1374.296387	48.34
SARIMA(1,1,2)(1,0,2,12)	555.461512	20.08
ARIMA(3,1,2)	1379.000674	49.31
SARIMA(2,1,1)(3,1,1,12)	347.489950	11.28

Q8. 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.

Model	Parameters	Test RMSE
SARIMA (manual)	p, d, q= (2,1,1) P, D, Q= (3,1,1) Seasonality=12	347.489
Triple Exponential Smoothing (Brute Force-AIC)	Alpha=0.3 Beta=0.3 Gamma=0.3	392.786
Triple Exponential Smoothing	Alpha=0.111, Beta=0.0616 Gamma=0.394	469.432
SARIMA (auto)	p, d, q= (1,1,2) P, D, Q= (1,0,2) Seasonality=12	555.461
Moving Average	2 point Trailing moving average	813.400
Moving Average	4 point Trailing moving average	1156.589

Simple Average Model		1275.081
Moving Average	6 point Trailing moving average	1283.927
Simple Exponential Smoothing	Alpha=0.0496	1316.034
Moving Average	9 point Trailing moving average	1346.278
ARIMA (auto)	p, d, q= (2,1,2)	1374.296
ARIMA (manual)	p, d, q= (2,1,1)	1379.000
Linear Regression		1389.135
Simple Exponential Smoothing (AIC)	Alpha=0.3	1935.507
Naïve model		3864.279
Double Exponential Smoothing	Alpha=0.3, Beta=0.3	18259.110

Table 16: Table containing all the models along with their parameters and RMSE scores.

From the above table we can see that out of all the models we have built, SARIMA (manual) is the best model since it has the lowest RMSE of 347.489. Here, we will proceed to build both SARIMA (manual) and Triple exponential model (since they both have almost near-by RMSE scores) on the complete data and use it for predicting the next 12 months.

Q9. 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.

1.SARIMA(manual)

Let us apply Brute Force SARIMA (manual) on Full Data The best-found parameters for SARIMA (manual) are (2,1,1) (3,1,1,12)

```
RMSE of SARIMA(manual) on the full Model 545.1493592269984
```

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1887.916306	400.997075	1101.976480	2673.856132
1995-09-30	2468.943552	405.418080	1674.338716	3263.548387
1995-10-31	3294.331626	405.535888	2499.495890	4089.167361
1995-11-30	3853.661757	405.651937	3058.598570	4648.724943
1995-12-31	6116.066701	406.023791	5320.274694	6911.858708
1996-01-31	1193.556118	406.365839	397.093708	1990.018527
1996-02-29	1578.027085	406.680007	780.948918	2375.105253
1996-03-31	1836.318461	406.993156	1038.626534	2634.010388
1996-04-30	1849.048333	407.308123	1050.739082	2647.357584
1996-05-31	1678.575795	407.623135	879.649130	2477.502460
1996-06-30	1633.367208	407.937773	833.823865	2432.910551
1996-07-31	2010.029934	408.252131	1209.870461	2810.189406

Table 17: Table with sale predictions for next 12 months with lower and upper confidence intervals

From the above table we can see that the sales of the Sparkling wine which was predicted for next 12 months shows a similar pattern as the previous years. We can hardly find a trend in this prediction. This indicates that the customer preference for this wine will be constant similar to the previous years. The month of December is having the highest sales just like the previous years. This high sales during that month can be due to the festival or holiday seasons. The sales can be brought even higher by giving more discounts or offers to customers and by more promotion of this wine.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(2, 1, 1)x(3, 1, 1, 12)	Log Likelihood	-1008.515
Date:	Sun, 20 Feb 2022	AIC	2033.031
Time:	13:48:06	BIC	2056.332
Sample:	01-31-1980 - 07-31-1995	HQIC	2042.500
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1102	0.090	1.229	0.219	-0.066	0.286
ar.L2	-0.0795	0.105	-0.757	0.449	-0.285	0.126
ma.L1	-0.9613	0.035	-27.184	0.000	-1.031	-0.892
ar.S.L12	-0.5605	0.514	-1.091	0.275	-1.568	0.447
ar.S.L24	-0.2755	0.251	-1.099	0.272	-0.767	0.216
ar.S.L36	-0.1573	0.121	-1.297	0.195	-0.395	0.080
ma.S.L12	0.0397	0.521	0.076	0.939	-0.981	1.060
sigma2	1.608e+05	1.55e+04	10.375	0.000	1.3e+05	1.91e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	38.67
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	0.52	Skew:	0.68
Prob(H) (two-sided):	0.03	Kurtosis:	5.23

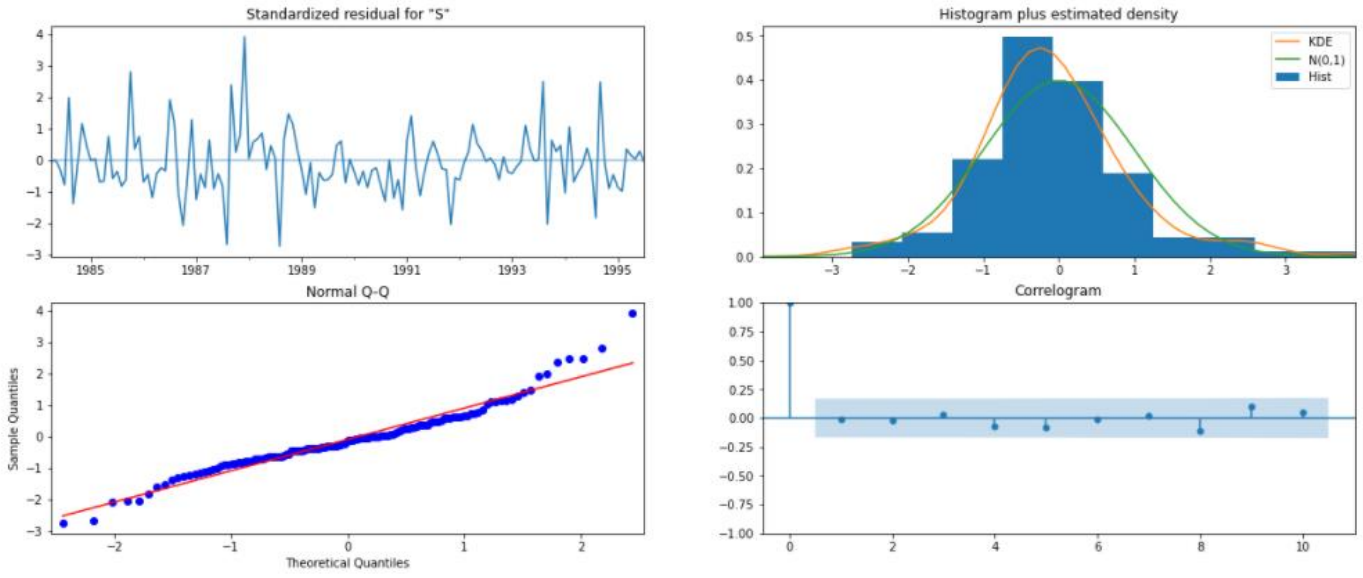


Fig 27: Full model diagnostics on entire data by SARIMA (manual) model

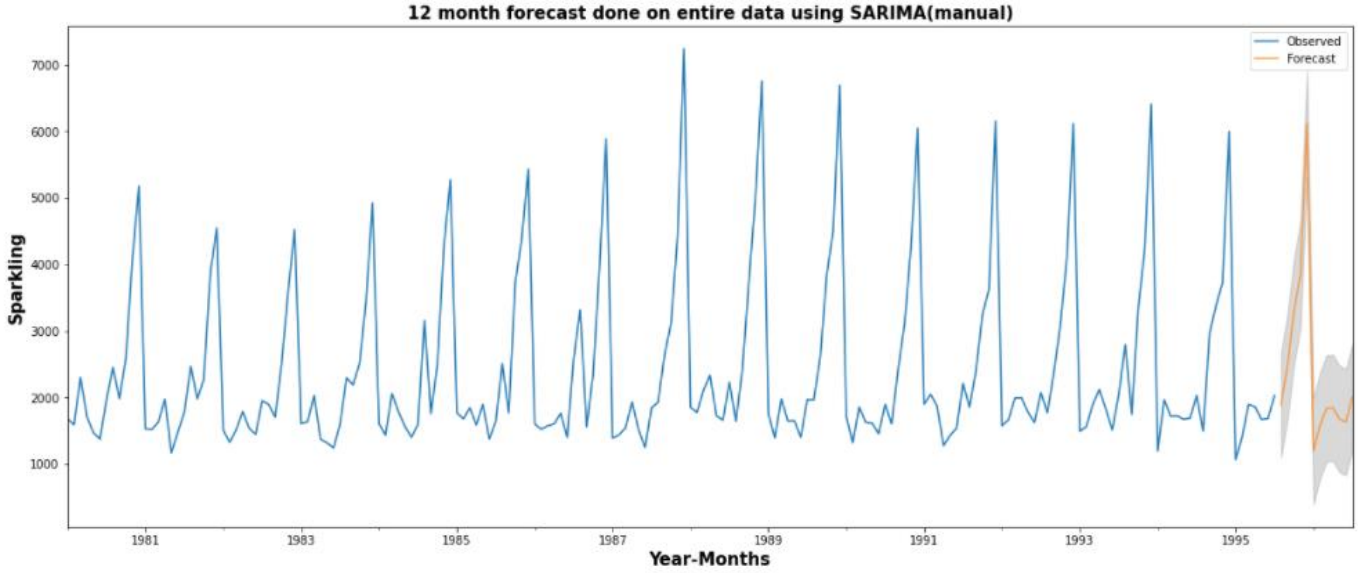


Fig 28: 12 months forecasts on entire data using SARIMA (manual)

2.Triple exponential smoothing(Brute Force-AIC)

Let us apply Brute Force Triple Exponential Smoothing on Full Data The best-found parameters for Triple Exponential Smoothing are Alpha=0.3, Beta=0.3, Gamma=0.3

```
RMSE of Triple exponential smoothing on entire data: 421.30973568581123
```

```
1995-08-31    1855.475048
1995-09-30    2487.237864
1995-10-31    3324.237414
1995-11-30    4227.091144
1995-12-31    6831.547050
1996-01-31    1585.313148
1996-02-29    2061.822043
1996-03-31    2418.525622
1996-04-30    2390.890238
1996-05-31    2158.075901
1996-06-30    2037.358404
1996-07-31    2424.415242
```

18(a)

	lower_CI	prediction	upper_ci
1995-08-31	1027.496371	1855.475048	2683.453725
1995-09-30	1659.259187	2487.237864	3315.216541
1995-10-31	2496.258736	3324.237414	4152.216091
1995-11-30	3399.112466	4227.091144	5055.069821
1995-12-31	6003.568373	6831.547050	7659.525727

18(b)

Table 18(a): Table with sale predictions for next 12 months

Table 18(b): Table with sale predictions for next 12 months with lower and upper confidence intervals

From the above table we can see that the sales of the Sparkling wine which was predicted for next 12 months shows a similar pattern as the previous years. We can hardly find a trend in this prediction. This indicates that the customer preference for this wine will be constant similar to the previous years. The month of December is having the highest sales just like the previous

years. This high sales during that month can be due to the festival or holiday seasons. The sales can be brought even higher by giving more discounts or offers to customers and by more promotion of this wine.

In the above table, we have calculated the upper and lower confidence bands at 95% confidence level.

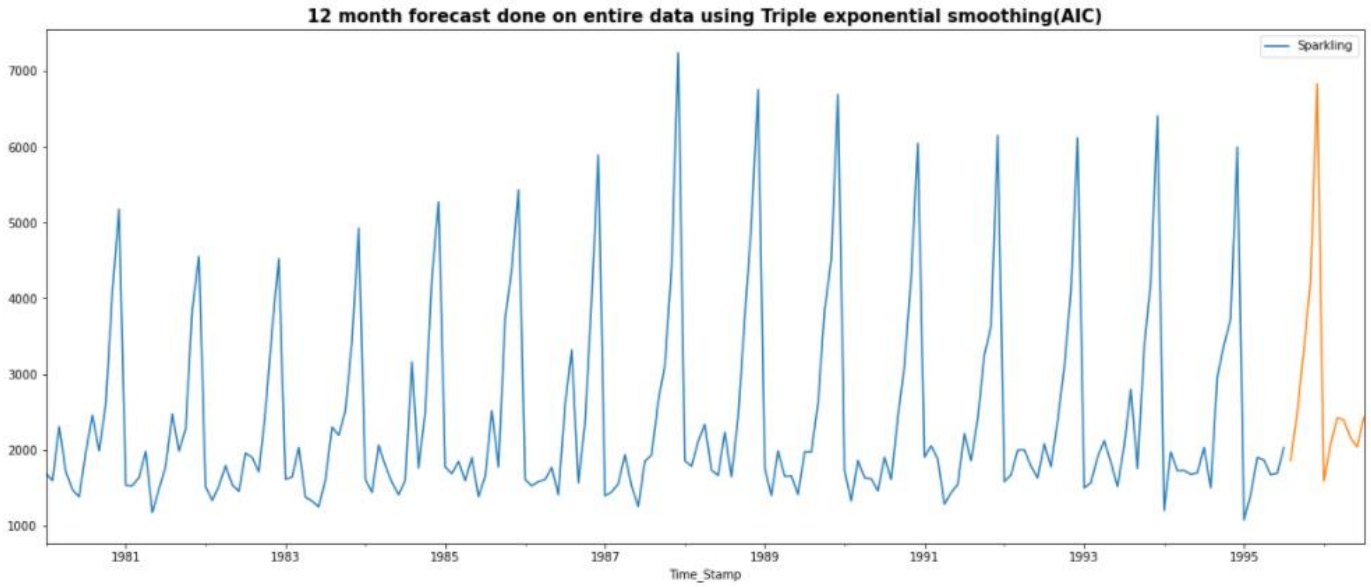


Fig 29(a)

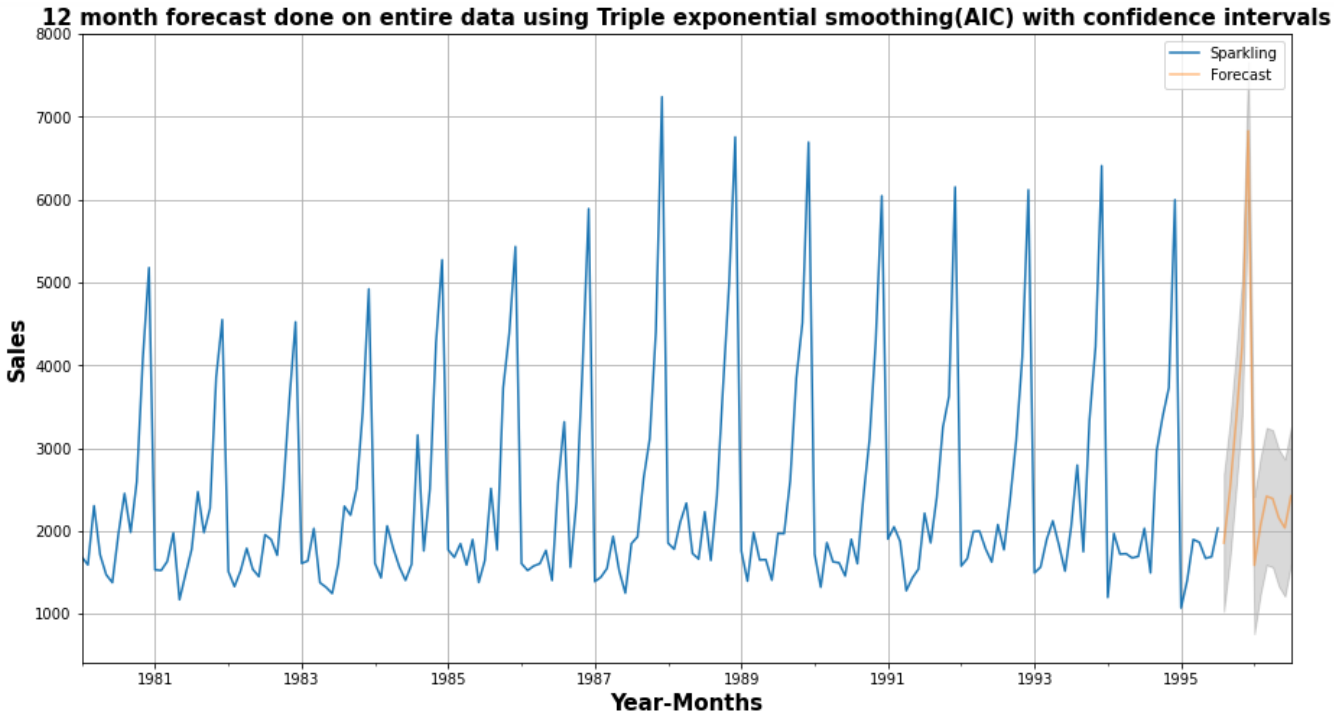


Fig 29(b)

Fig 29(a &b): 12 months forecasts on entire data using Triple exponential smoothing (AIC) with and without confidence intervals

In the above plot, we have plotted the graph by taking the upper and lower confidence bands at 95% confidence level into consideration. The above plot clearly depicts a steady sale in the forecasts. The shaded region in the forecasts shows the confidence intervals of the predictions.

From the above two forecasts, it is clearly evident that triple exponential smoothing (Brute Force-AIC) has forecasted more accurately for the next 12 months than the SARIMA model. Though the RMSE score of SARIMA model on test data was lower than the RMSE of triple exponential, when it comes to future predictions (forecast for 12 months), the triple exponential model has got lower RMSE of 421.309 on entire data compared to the RMSE score of 545.149 by SARIMA (manual) on whole data.

Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

• Best Accuracy (RMSE) is observed by SARIMA (2,1,1) (3,1,1,12), where p=2, q=1, d=1, P=3, D=1, Q=1 and seasonality=12.

• The second-best Accuracy (RMSE) is observed by triple exponential smoothing (Brute force-AIC) where alpha=0.3, beta=0.3, gamma=0.3

- For doing the best forecast for the Sparkling dataset we will choose the Triple Exponential Smoothing (brute force-AIC) as that model has the best accuracy (lowest RMSE)
- It is predicted that the sales of the Sparkling wine which was predicted for next 12 months shows a similar pattern as the previous years. Trend can be hardly found in this prediction.
- In the forecasts, the month of December is having the highest sales just like the previous years. This high sales during that month can be due to the festival or holiday seasons.