Business Report

Time Series Forecasting

(Sparkling Wine Analysis)

Arnab Ghosal

19 September 2021

Executive Summary

For this assignment, the data of Sparkling wine sales in the 20th century is to be analyzed. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century. The dataset consists of the Sparkling wine sale. In this problem statement we will explore the different sale of the wine based on different months in different year. Depending on this data analysis we need to forecast the future sales.

Introduction

Forecast is a statistical method to predict an attribute using historical patterns in the data. Every business and organization apply different methods of forecasting in different situations. Therefore, it is imperative to identify what to forecast and which method of forecasting should be utilized in different scenarios so that the risk of forecasting is minimized.

The purpose of this whole exercise is to explore the dataset and various forecasting methods on this data, to understand which forecasting method is more suitable in this scenario. This assignment should help in exploring the time series, its components i.e., trend, seasonality & its effects on forecasting the sales in future.

Data Description

1. Year Month: It represent the month and year of the sale.
2. Sparkling: It represent the sale in the corresponding year & it's month. The data is from year 1980 Jan to 1995 July.

Sample of the dataset

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

Table 1 Dataset Sample

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

The Y axis represent the Sale of the Sparkling Wine

Though the above plot looks like a Time Series plot, notice that the X-Axis is not time.

In order to make the X-Axis as a Time Series we need to pass the date range manually and add a column name Time stamp in the dataframe

The plot of the data from the given csv file appears Figure 1.
But we need to convert this data into Time series object to get a proper plot.

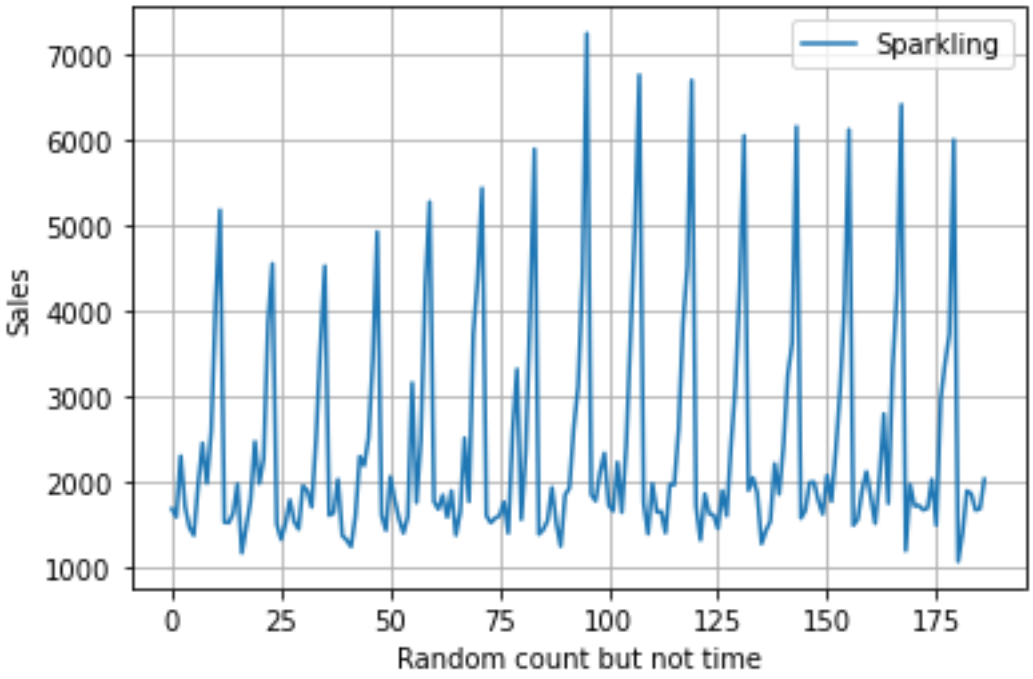


Figure 1 Dataset Time Series Plot

	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 2 Dataset with time stamp

Now we can plot the proper Time series Plot which will have Year's value on X axis & Sales value on Y axis. F

From the below plot we can observe that there is no trend in the dataset, but there is seasonality in the dataset.

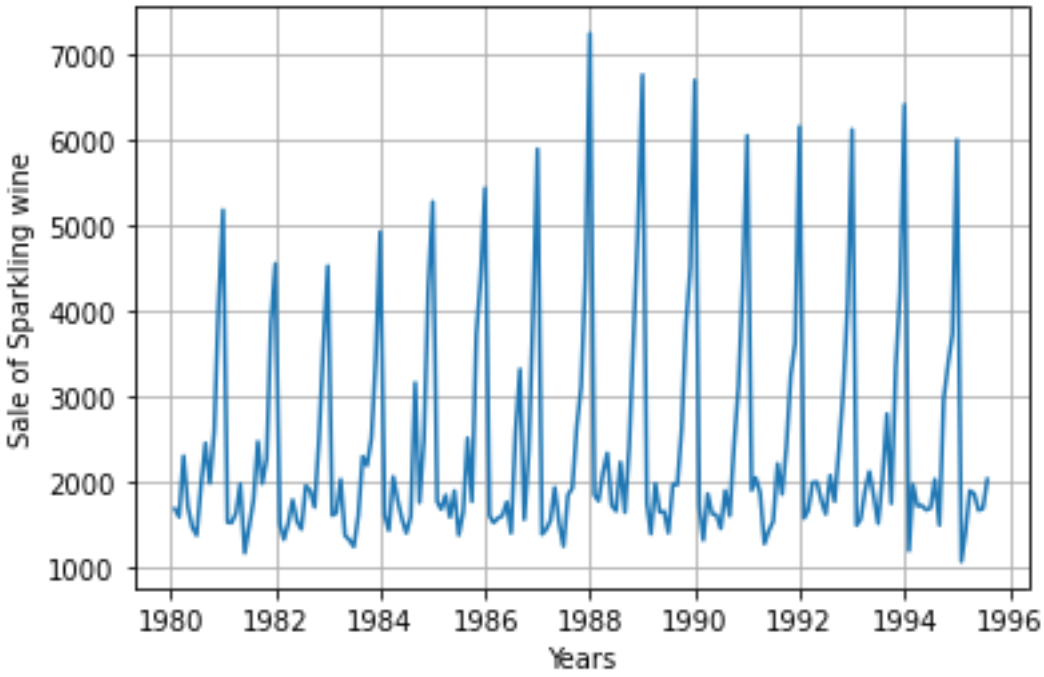


Figure 2 Time Series Plot with years

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

Exploratory Data Analysis:

1. Shape of the dataset:

The dataset has 187 rows and 2 columns.

2. Dataset Information:

The dataset has 2 column name YearMonth and Sparkling respectively.
The YearMonth has datatype as object. The Sparkling has datatype as int64.

3. Null Value Check:

There are no null values in the dataset.

4. Descriptive Statistics:

```
df.shape
```

```
(187, 2)
```

```
df.info()
```

```
<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)
memory usage: 3.0+ KB
```

```
df.isnull().sum()
```

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

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

From year 1980 Jan to 1995 July
On an average the sale of sparkling wine is 2402.417112.
The 25% sale of sparkling wine is 1605.0.
The 50% sale of sparkling wine is 1874.0.
The 75% sale of sparkling wine is 2549.0.
The min sale of sparkling is 1070.0.
The max sale of sparkling is 7242.0.
As the max value is too large than the 75% so outlier is present in the data.

Yearly Plot

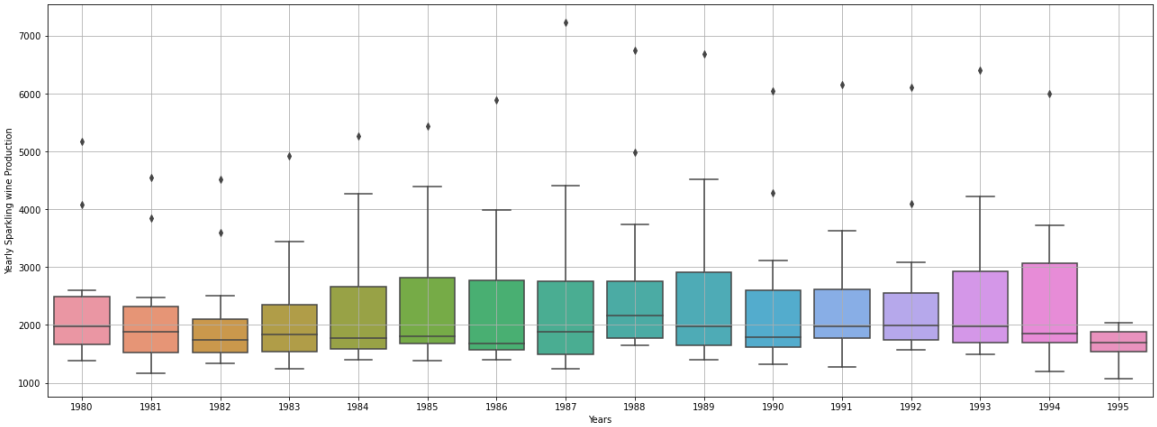


Figure 3 Yearly Plot of Sparkling Wine Sale

The lowest sale of the Sparkling wine is in the year1995.
The highest sale of the Sparkling wine was in year 1985,1987,1989.

There is no extreme increase or decrease in the sale of the Sparkling wine, that's why we don't have trend in this data.

We can clearly see that there are too many outliers in sale in each year.

Monthly Plot

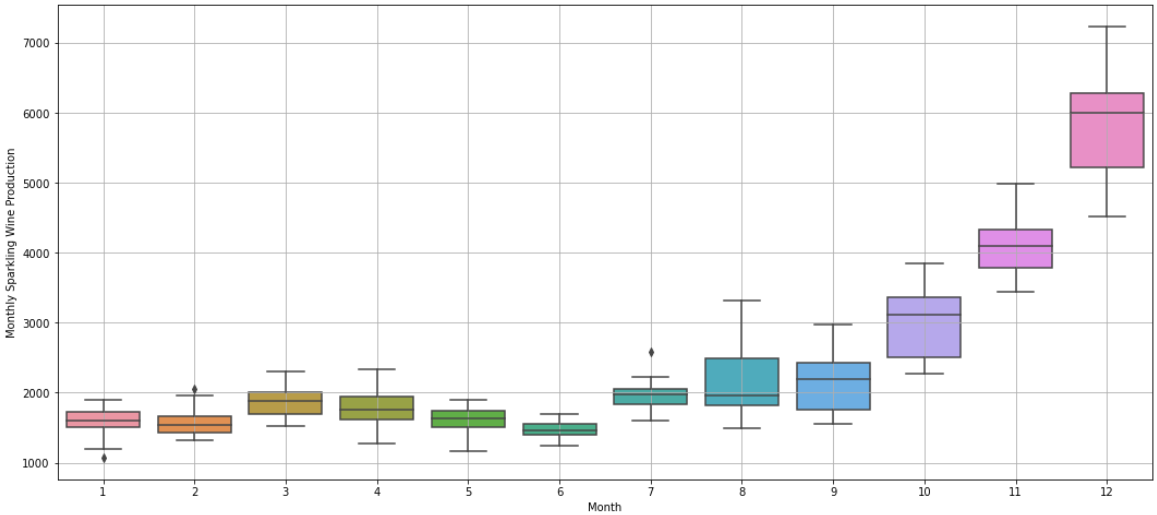


Figure 4 Month Plot (Box Plot)

The sale of sparkling wine is lowest in month of June & highest in the month of December. We can see that after July the Sparkling wine sales goes on increasing every month.

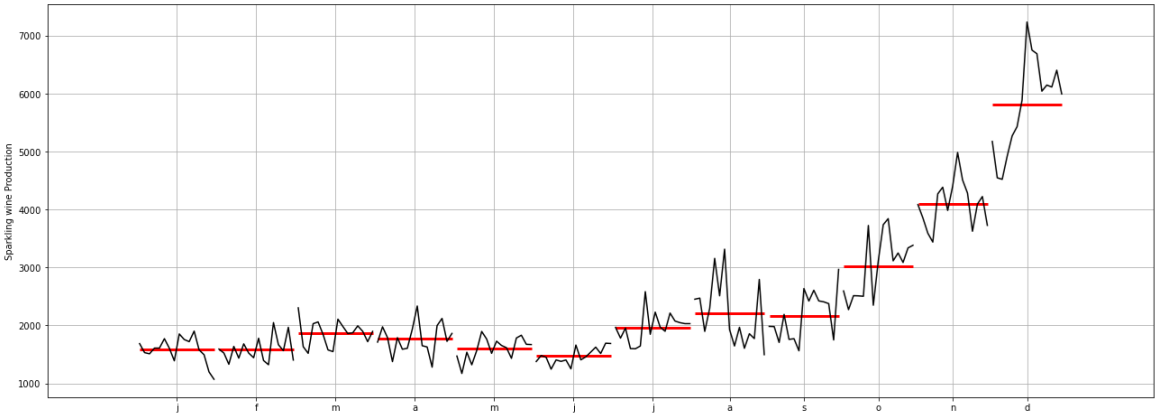


Figure 5 Month Plot with Average Sales

The vertical lines represent monthly sales, and the horizontal lines represent average sales of the given month.

- Here, it can be observed that average sales are higher in December as compared to other months.

- Average sales are lower in July as compared to other months. 9
- After month of July the Sale start increasing. From January to July the sales are not much, that means there is seasonality over here.
- In a year during January to July sale will be less and July to December is the season for good sale. **Empirical Cumulative Distribution**

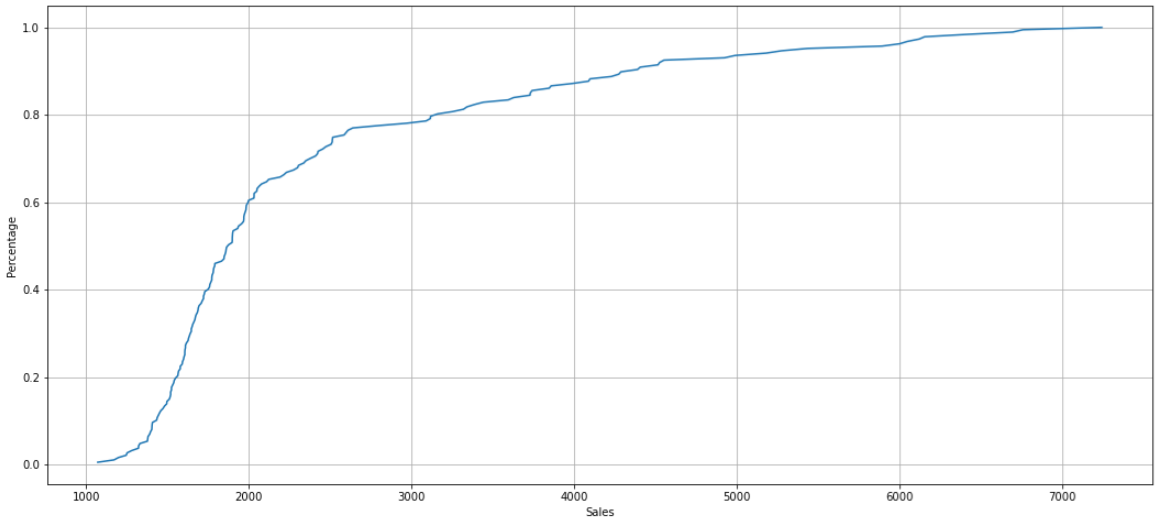


Figure 6 Empirical Cumulative Distribution Plot

- This plot y axis shows the percentage of sale.
- Among the total sale from year 1980 to 1985, 90% of total sale are under 5000.
- Sale of 6000 to 7000 is of only 10% of total sale.

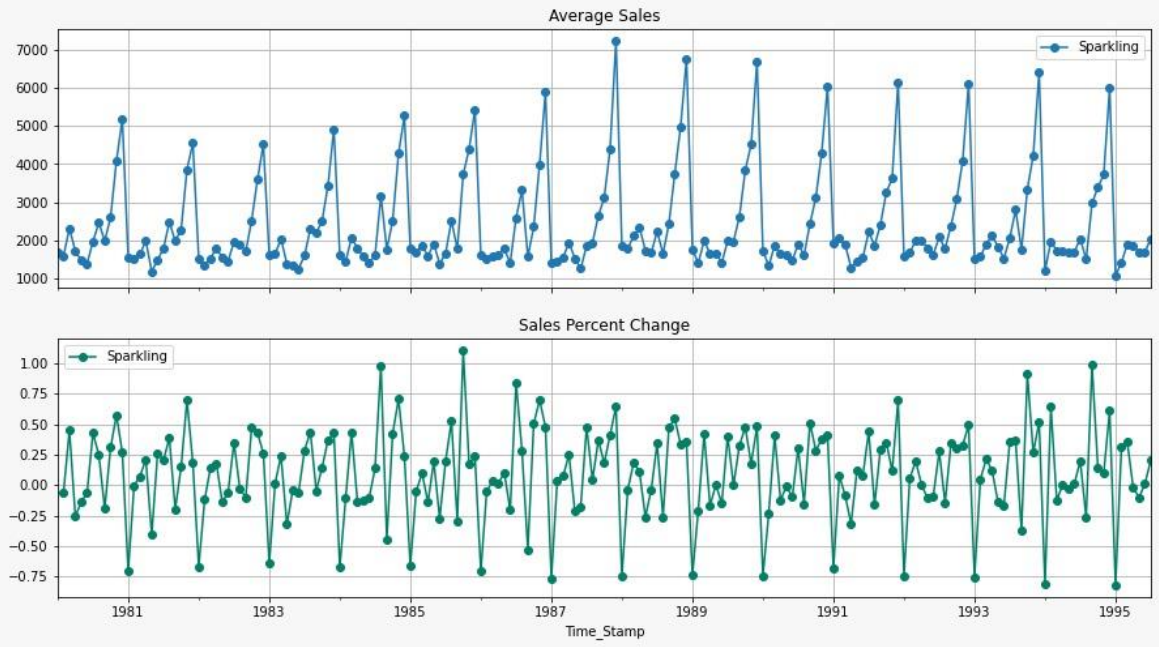


Figure 7 Percent Change in Sales

- From the average Sale we came to known there is no trend in the time series.

- The below percent change graph is flat over the whole series , this shows that the amount of the
change is constant over whole series.

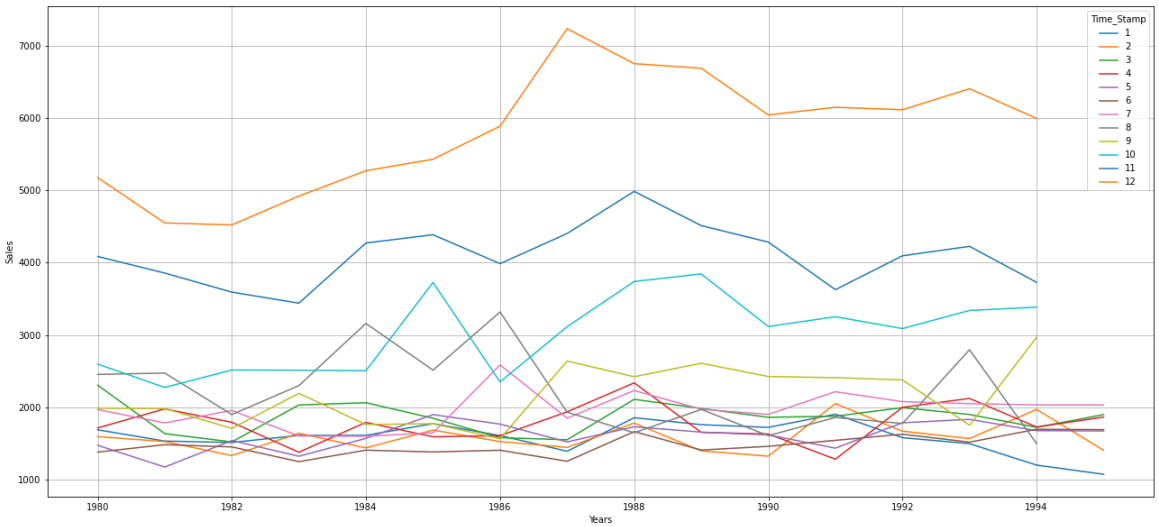


Figure 8 Yearly sales across months

In every year the sales starts increasing fron the month of July to December. From January to July the sales are less

Decomposition of Time Series

The time series have three components.

1. Trend (Long term movement)

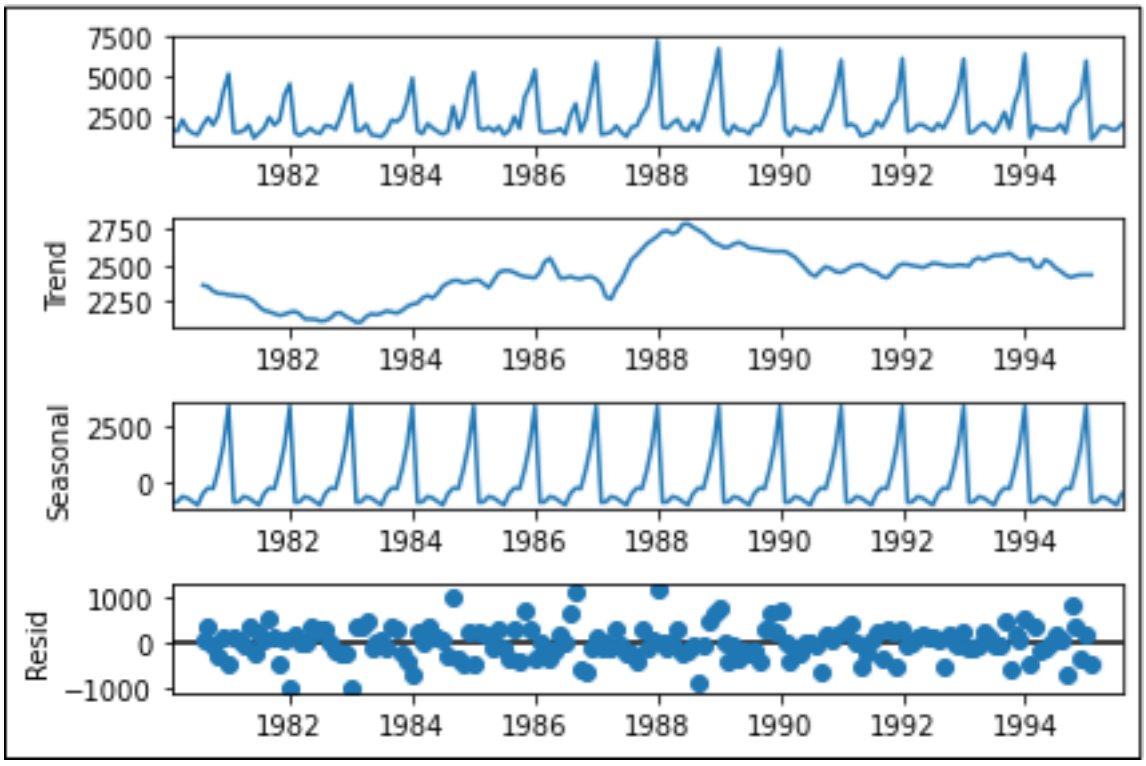
2. Seasonal component: Intra-year stable fluctuations repeatable over the entire length of the
series

3. Irregular component (Random movements)/Residuals/Noise.
In decomposition of the time series, we came to known clearly on a broader platform regarding these three components.

- From the Decomposition plot we can clearly see that the Sparkling wine sales data do not have trend component in it. As we can see that there is no continuous increasing, decreasing behavior in the data.

- There is Multiplicative Seasonality in the data. **Additive Model**

Figure 9 Additive Decomposition Model



For Additive Model we see that the residuals are located around 0 from the plot of the residuals

in the decomposition. But they are very much disperse not totally concentrated on 0 line.

Multiplicative Model

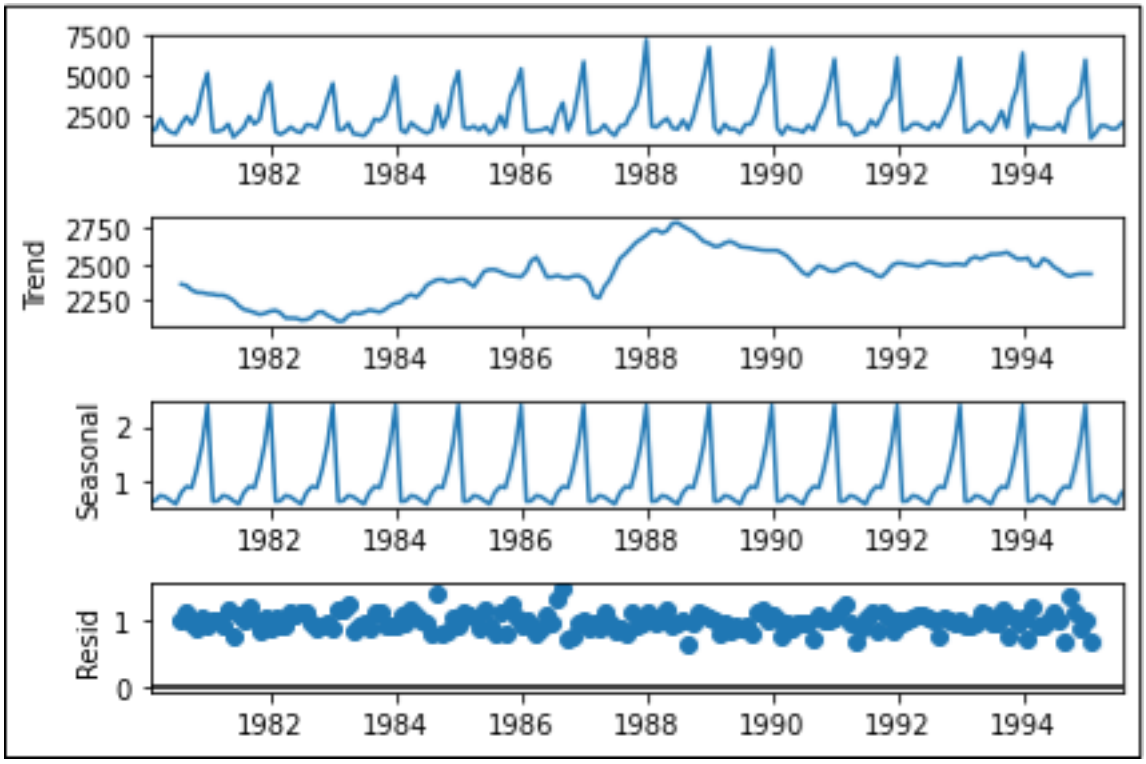


Figure 10 Multiplicative Decomposition Model

For the multiplicative series, we can see that lot of residuals are located around 1. So here we can consider the seasonality in the dataset is Multiplicative seasonality.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	-854.26	-830.35	-592.36	-658.49	-824.42	-967.43	-465.5	-214.33	-254.68	599.77	1675.07	3386.98

Table 3 Seasonal Indices

Since this is monthly data, there are 12 seasonal indices.

In December Sparkling wine sales is the highest among all months in the same year, as borne by the highest value of the seasonal component whereas in January (lowest value of the seasonality) sales is the lowest.

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

Before a forecast method is proposed, the method needs to be validated. For that purpose, data must be split into two sets i.e., training and testing. Training data helps in identifying and fitting right model(s) and test data is used to validate the same.

In case of time series data, the test data is the most recent part of the series so that the ordering in the data is preserved.

• As asked in the question for Sparkling wine sale series, the first 10 years of data from 1980 to 1990 is used for training purpose and last 5 years of data from 1991 to 1995 is used for testing purpose.

Training Dataset (1980 to 1990)

Training Set starting few points and end few points.

First few rows of Training Data

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

Last few rows of Training Data

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

Testing Dataset (1991 to 1995)

Testing Set starting few points and end few points.

Last few rows of Test Data

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

First few rows of Test Data

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

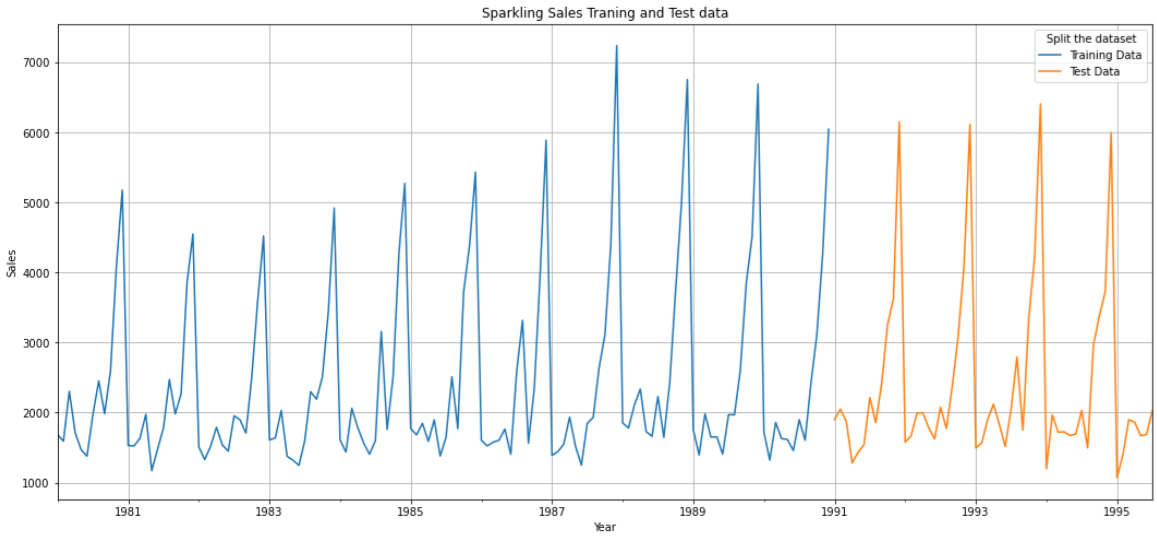


Figure 11 Plot of Training and test data division

So, from the graph above we can see that the training and testing data are successfully got split as per the requirement.

Shape of the Train & Test dataset

```
print(train.shape)
print(test.shape)
```

```
(132, 1)
(55, 1)
```

Out of 181 Rows after splitting data into train & test dataset. Now train dataset have 132 rows and test dataset have 55 rows.

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

For Forecasting using Decomposition approach there are various models which can be used.

List of the models are:

1. **Linear Regression**
2. **Naive Forecast**
3. **Simple Average**
4. **Moving Average**
5. **Simple smoothing method**
6. **Holt's Method (Double Exponential Smoothing)**
7. **Holt-Winter's method (Triple Exponential Smoothing)**

Performance Evaluation of the Model:

Forecasting accuracy measures compare the predicted values against the observed values to quantify the predictive power of the proposed model. Mathematically, it can be defined as

Forecast error et for period t is given by: et = Ŷt - Yt Where

Ŷt = forecast value for time period t

Yt = actual value in time period t

n = No. of observations **Measure of Forecasting error**

Root Mean Square Error (RMSE): $RMSE$=

Lower the RMSE value better is the model.

For Forecasting there are various models which can be used. List of the models are:

1. **Linear Regression**
2. **Naive Forecast**
3. **Simple Average**
4. **Moving Average**

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

1. Linear Regression

The linear regression model is built on the train dataset and forecasting is done on the test dataset. The Forecast line does not follow the trend or seasonality in the test data.
This forecast is not useful does not work well.
The plot of the Forecast is as given below:

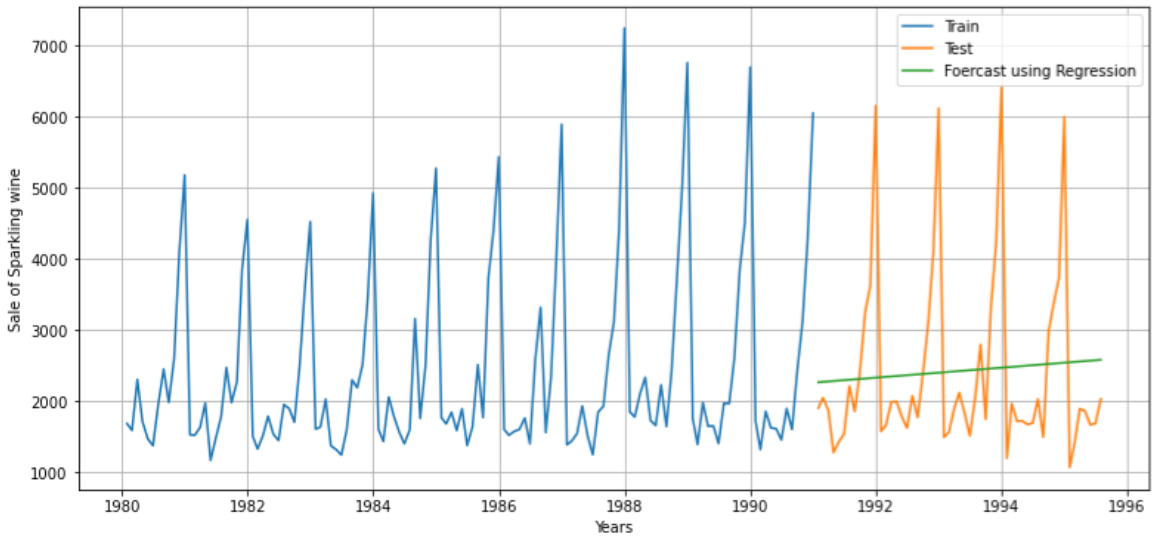


Figure 12 Linear Regression Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.
• RMSE value for Linear Regression is **1275.867052.**

2. Navie Forecast

Naïve Forecast uses the last observed value for forecasting
The Navie model is built on the train dataset and forecasting is done on the test dataset. The Forecast line does not follow the trend or seasonality in the test data.
This forecast is not useful does not work well.
The plot of the Forecast is as given below:

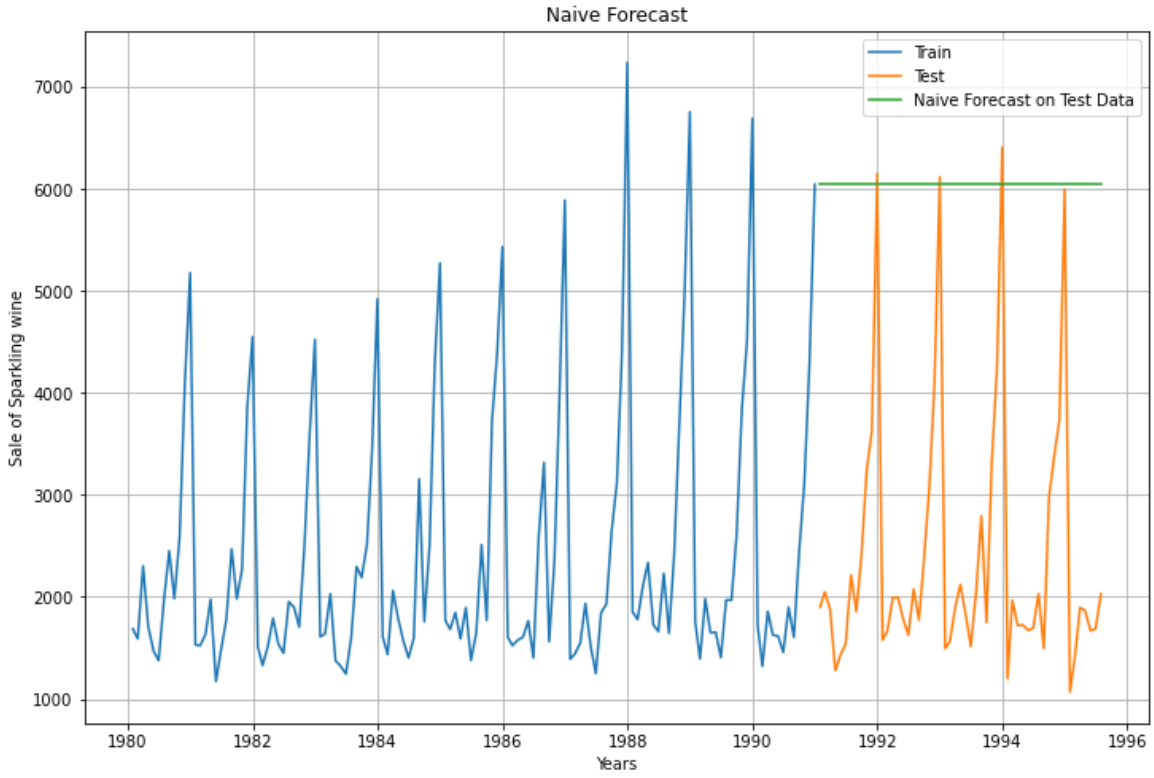


Figure 13 Naive Model Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.
• RMSE value for Naïve Model is **3864.279352.**

3. Simple Average

The Simple Average model is built on the train dataset and forecasting is done on the test dataset. The Forecast line does not follow the trend or seasonality in the test data.

This forecast is not useful.
The plot of the Forecast is as given below:

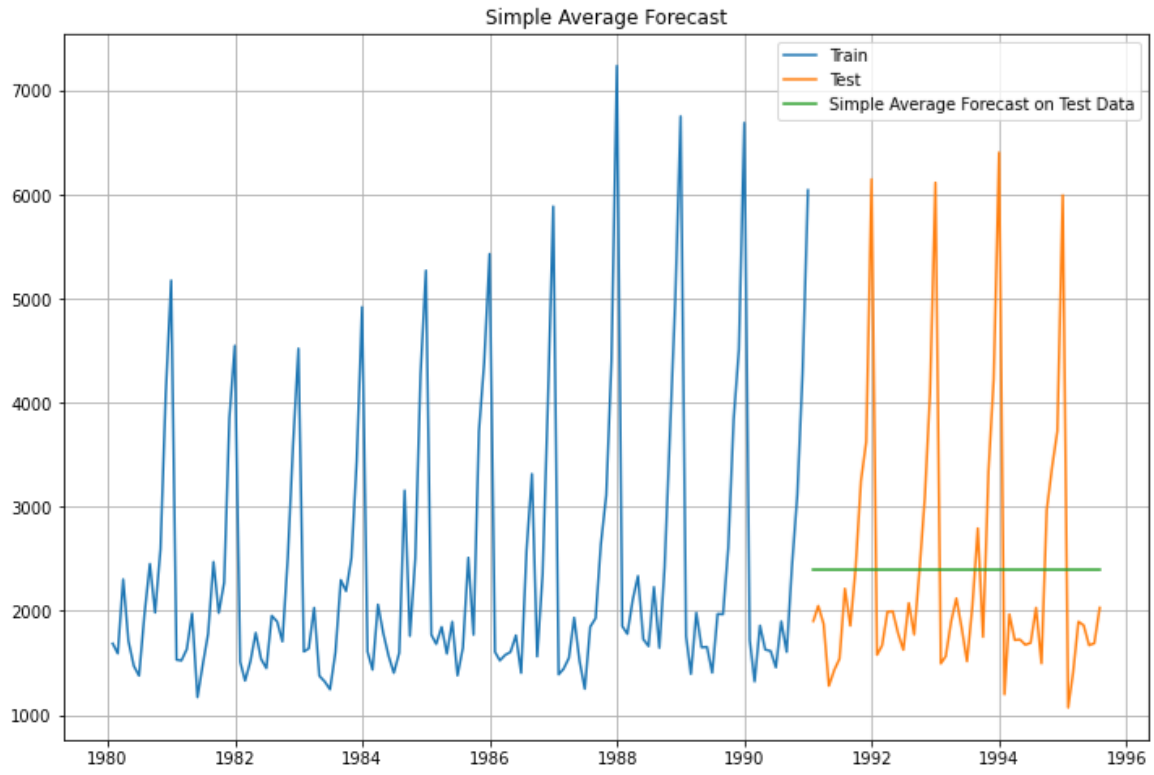


Figure 14 Simple Average Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.
- RMSE value for Simple Average Model is **1275.081804.**

4. Moving Average

For forecasting take average over a window of certain width & move the window.

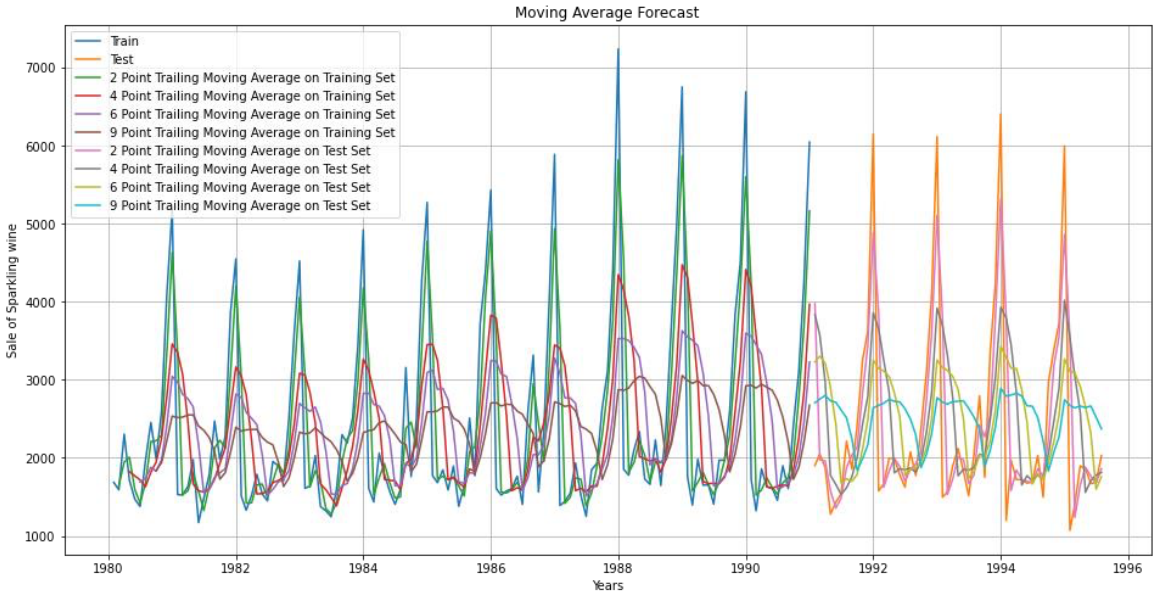


Figure 15 Moving average Forecasting Plot

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for 2point Training Moving Average is **813.400684.**
- RMSE value for 4point Training Moving Average is **1156.589694.**
- RMSE value for 6point Training Moving Average is **1283.927428.**
- RMSE value for 9point Training Moving Average is **1346.278315.**
  2-point Training Moving average model is best from the all the moving average points model.

5. Simple Exponential Smoothing

SES or one-parameter exponential smoothing is applicable to time series which do not contain either of trend or seasonality. Forecast by SES is given by:

$\hat{Y}t+1 = \alpha Yt+\alpha(1-\alpha)\ Yt-1+\alpha(1-\alpha)2Yt-1+\cdots,\ 0<\alpha<1$

where, α is the smoothing parameter for the level. Such a series is hard to find. This is a one- step-ahead forecast where all the forecast values are identical.

We do the Simple Exponential modelling in two ways.

- By setting the default values of the parameters.

- By taking the best value of the parameter from range 0.3 to 1

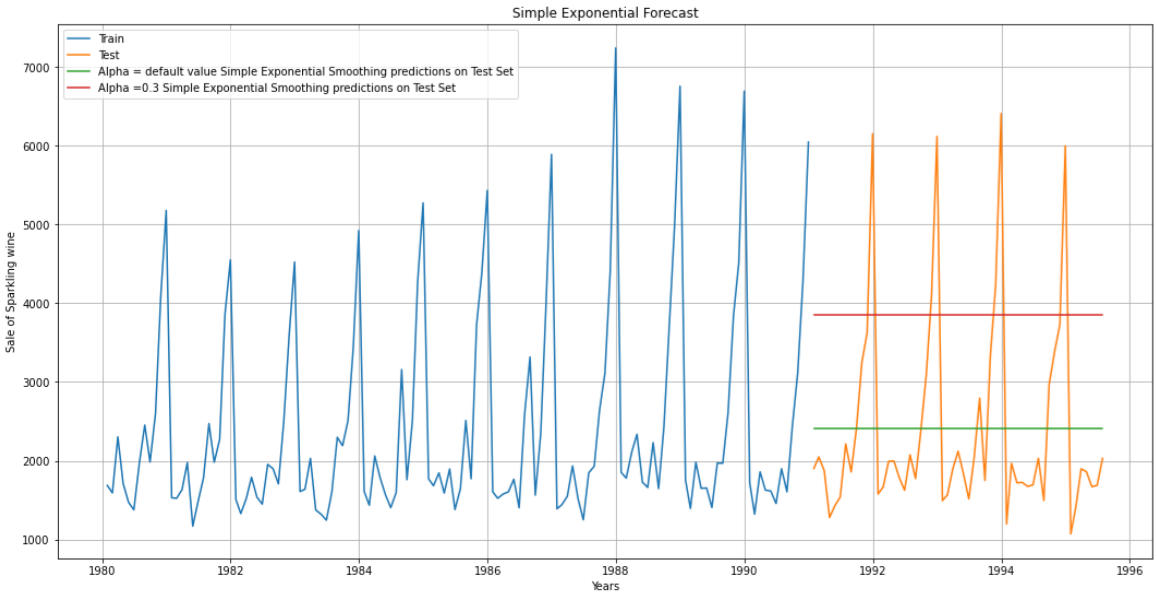


Figure 16 Simple Exponential Smoothening Forecasting Plot

We apply the Simple Exponential Smoothing at the default alpha value and at alpha = 0.3.

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for SES at default alpha value is **1275.082.**

- RMSE value for SES at alpha value = 0.3 is **1935.507132.**
  From the above two the result for default alpha value is better than the alpha value 0.3.
  In Simple Exponential Smoothing determines the level parameter only which is indicated in the graph above i.e., it forecast only the level parameter. But no trend and seasonality are observed in forecasting, so for our case this simple Exponential Smoothing is not useful.

6. Double Holt's Method (Double Exponential Smoothing)

This method is an extension of SES method, proposed by Holt in 1957. This method is applicable where trend is present in the data but no seasonality.

α is the smoothing parameter for the level and β is the smoothing parameter for trend.
But in our data trend is not present, so this modelling type will not

appropriate for our case. We do the Double Exponential modelling in two ways.

- By setting the default values of the parameters.
- By taking the best value of the parameter from range 0.3 to 1.

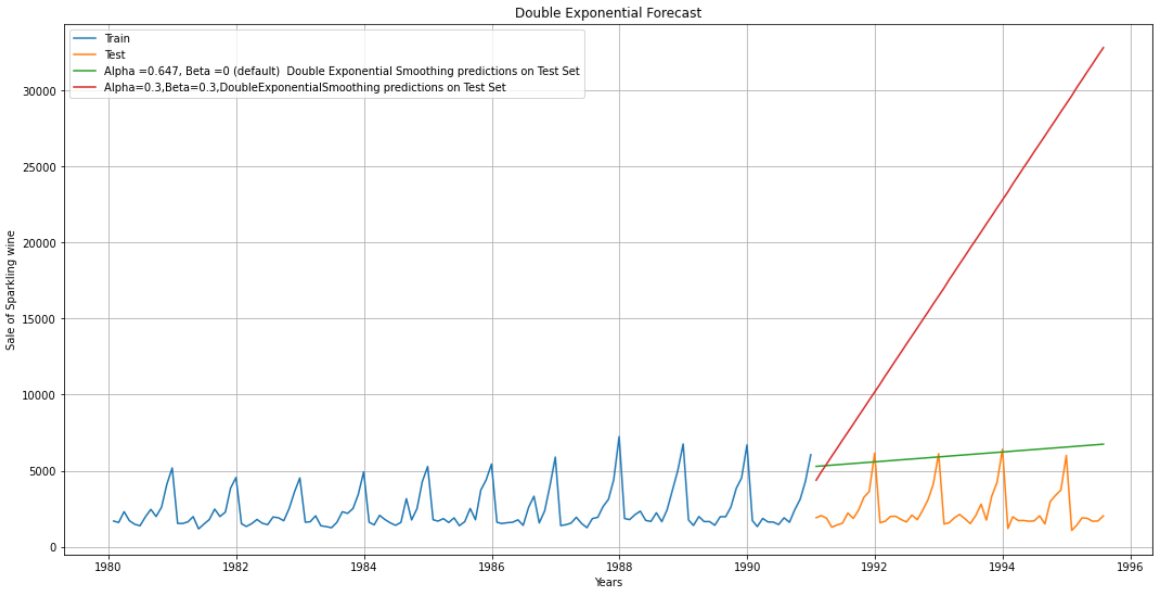


Figure 17 Double Exponential Smoothening Forecasting Plot

We apply the DES at the default alpha & beta value and at alpha 0.3, Beta=0.3

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for SES at default alpha value is **3851.331290.**
- RMSE value for SES at alpha & beta value = 0.3 is **18259.110704.**
  From the above two the result for alpha & beta 0.3 value is better than the default values.
  In Double Exponential Smoothing determines the level & trend parameter only which is indicated in the graph above i.e., it forecast only the level & trend parameter. But no seasonality is observed in forecasting, so for our case this Double Exponential Smoothing is not useful.

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

This is an extension of Holt's method where along with trend seasonality is also found in the data.

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, β and γ. This is a general method and a true multi-step ahead forecast.

But in our data trend is not present, but seasonality is there so let's implement and check the TES model performance in this case.
We do the Triple Exponential modelling in two ways.

- By setting the default values of the parameters.
- By taking the best value of the parameter from range 0.3 to 1.

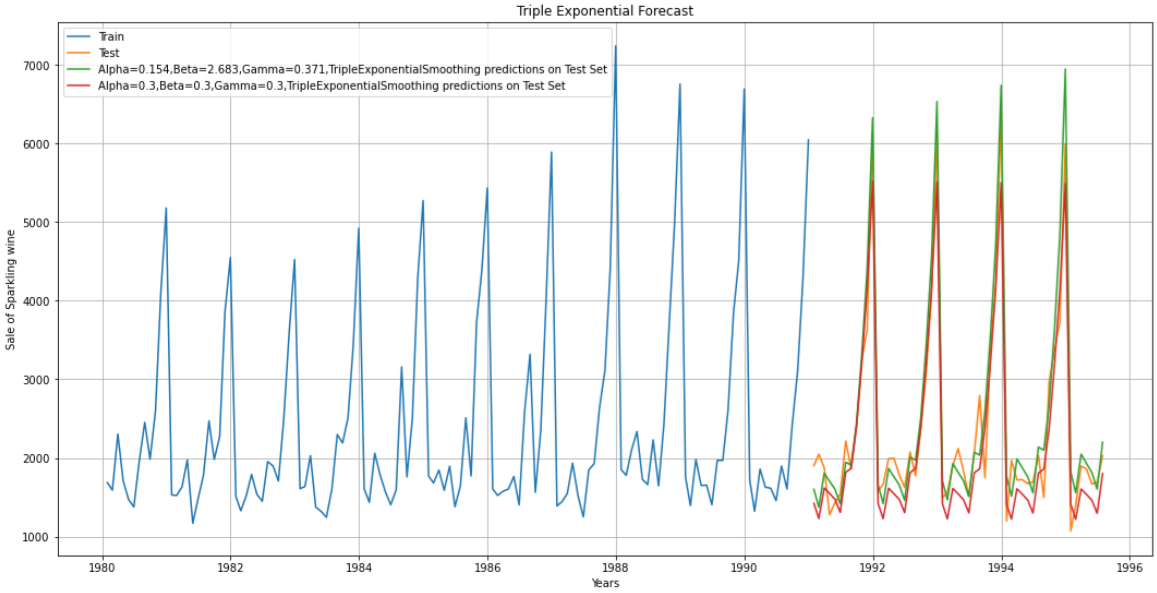


Figure 18 Triple Exponential Smoothening Forecasting Plot

We apply the TES at the default alpha & beta value and at alpha 0.3,Beta=0.3,Gamma = 0.3

Model Evaluation:

Model Evaluation by calculating the RMSE value.

- RMSE value for SES at default alpha value is **383.192343.**
- RMSE value for SES at alpha & beta value = 0.3 is **392.786198.**
  From the above two the result for alpha, beta, gamma default value is better than the 0.3 values.
  As compare to other exponential model Triple Exponential model has forecasted quite well, as we can see this from graph above.

Sorting of the RMSE values of all the models

	Test RMSE
Alpha=0.154,Beta=2.683,Gamma=0.371,TripleExponentialSmoothing	383.192343
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
Alpha=default value,SimpleExponentialSmoothing	1275.081839
RegressionOnTime	1275.867052
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.3,SimpleExponentialSmoothing	1935.507132
Alpha=0.647,Beta=0,default DoubleExponentialSmoothing	3851.331290
NaiveModel	3864.279352
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704

Table 4 Sorting of RMSE values for different models

Among all the values the TES (default values) model RMSE is less, so this model is best till now.

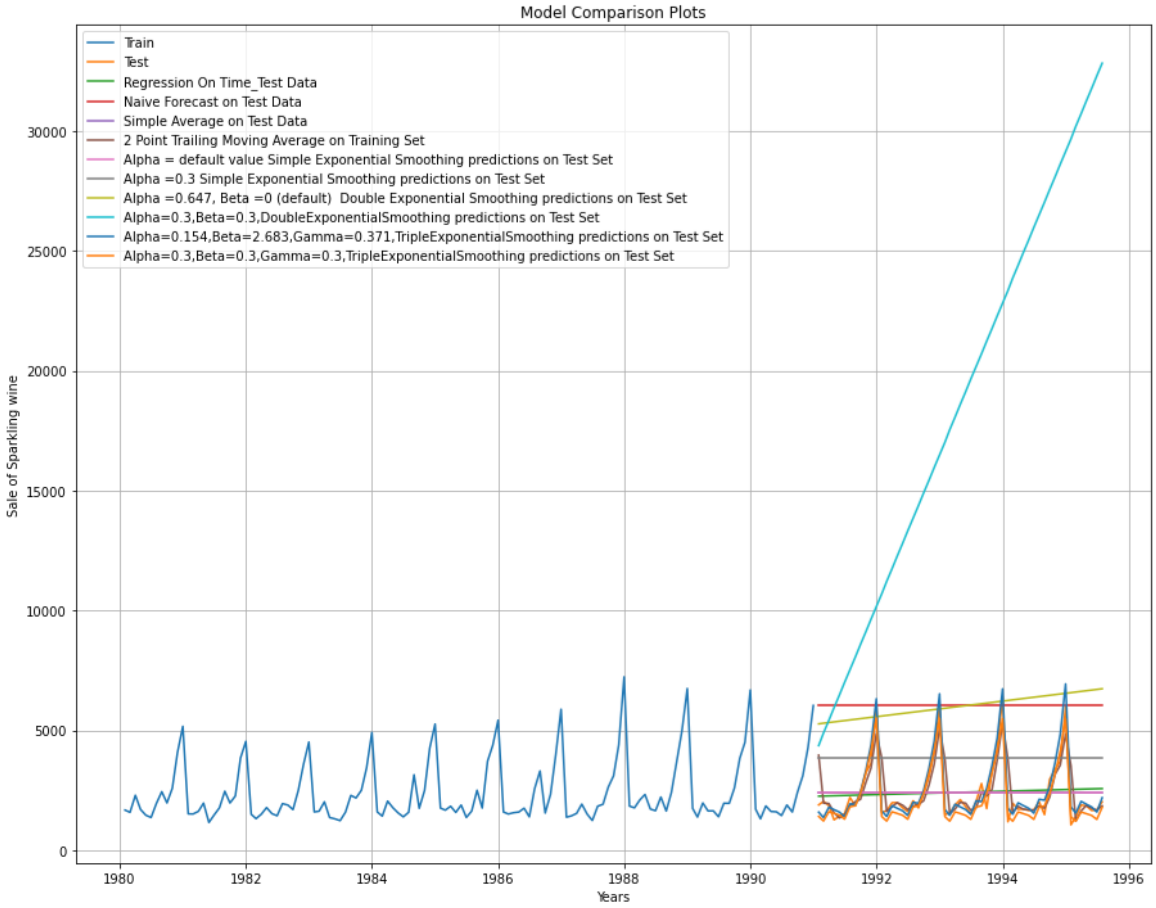


Figure 19 All model Comparison Plot

Among all the Model we can see that the Triple Exponential model forecast is closer to the test data while other models are directly showing the tangent line which is not at a relevant forecast.

Q.5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.

Note: Stationarity should be checked at alpha = 0.05.

- To build the ARIMA/ SARIMA model the time series must be Stationary.
- So first we need to check whether the series is stationary or not.
- If not, then make the series stationary to apply the ARIMA/SARIMA model on time series for forecasting.

- Since ARIMA model requires a stationary series, a formal stationarity test needs to be applied to the time series under consideration.

- Augmented Dickey-Fuller Test: A formal test to check whether time series data follows stationary process.

H0: Time series is non-stationary.
H1: Time series is stationary.
If p>0.05 we fail to reject the Null Hypothesis i.e., our Time Series will be non-Stationary. If p<0.05 we fail to reject the Alternate Hypothesis i.e., our Time Series will be Stationary.

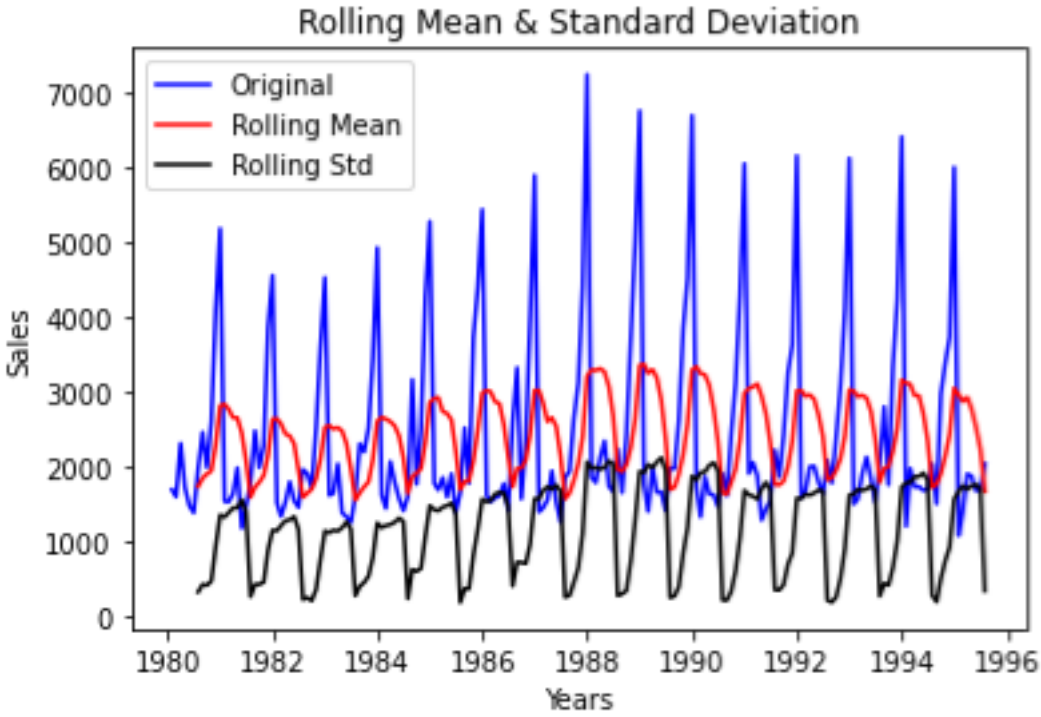


Figure 20 Time Series Stationarity Check on whole data

From the graph we can see that the mean and standard deviation are not constant, so we can predict that the series is non-stationary.
But to be more accurate we will execute Augmented Dickey-Fuller Test, this test gives us the p value. This p value tells us whether the series is stationary or not.

Auto Regression equation is $X_t = \Phi X_{t-1} + \varepsilon_t$

$X_t = X_{t-1} + \varepsilon_t$ (if $\Phi=1$)
$X_t - X_{t-1} = \varepsilon_t$ (Stationary Series)

So, for Time Series to be Stationary the Φ must be 1, the ADF test finds what is the probability that the Φ is 1. This probability is nothing but the p-value.

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

Figure 21 ADF Test Report for whole data

From the above Result we see that the p value is 0.601061 which is greater than 0.05. So, the Time Series is not stationary.
Often differencing a non-stationary time series leads to a stationary series.
So just differencing the series by 1 the plot is as below:

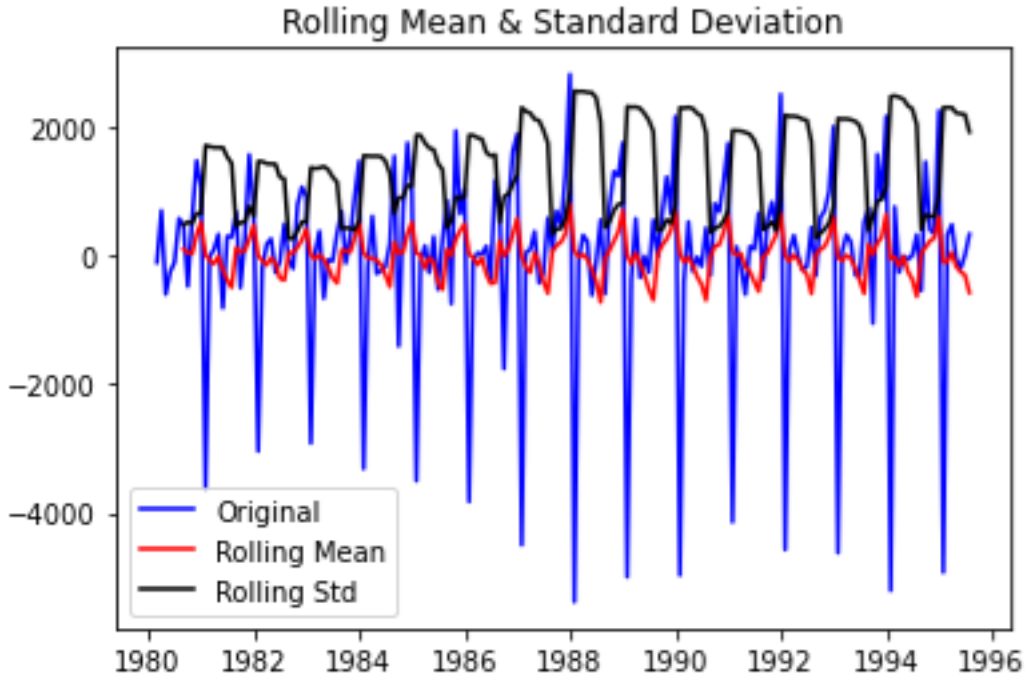


Figure 22 Time Series Stationarity Check after differentiation by 1

The mean and std dev of the series is constant all over the series, this shows that the series is now stationary.

The ADF test Result is

Now the p-value is less than 0.05, so the series is now become Stationary.

Now we must split the data into the train and test for the further modelling process, so we need to check the stationary of the train time series.

Train time series Stationary Check

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

Figure 23 ADF Test report after differentiation by 1

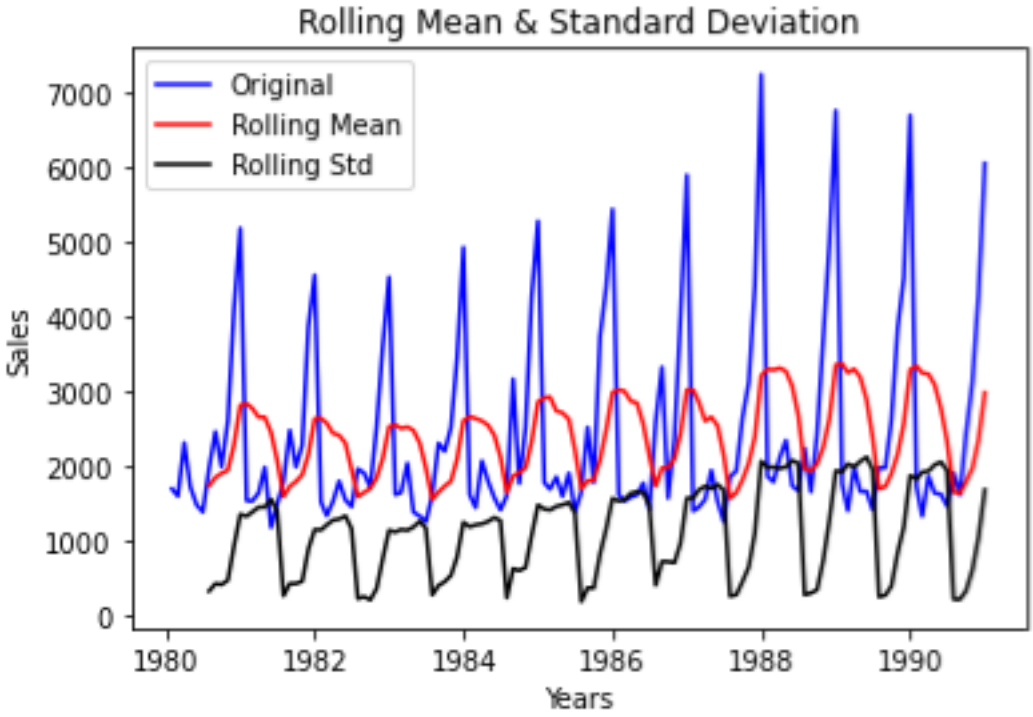


Figure 24 Time Series Stationary Check on Train dataset

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

Figure 25 ADF test report on Train dataset

From the above Result we see that the p value is 0.669744 which is greater than 0.05. So, the Time Series is not stationary.
Often differencing a non-stationary time series leads to a stationary series.
So just differencing the series by 1 the plot is as below:

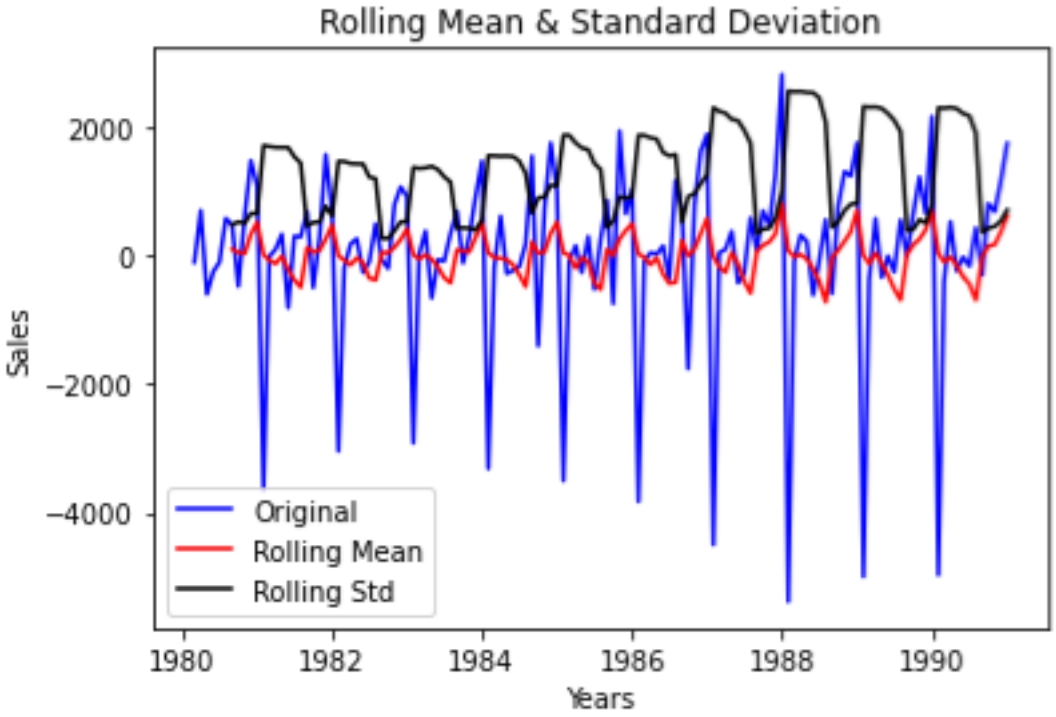


Figure 26 Time Series Stationary Check on Train dataset differentiation by 1

```
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
```

Figure 27 ADF test report on Train dataset after differentiation by 1

Now the p-value is less than 0.05, so the Train Time series is now become Stationary.

Q.6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

$ARIMA$ **(p, d, q) Model**: ARIMA is defined by 3 parameters. p: No of autoregressive terms

d: No of differencing to stationaries the series

q: No of moving average terms.

For building model, we will consider values as p = q = range (0, 3), d= range (1,2)

So, the Model for all combination of p, d, q will be taken under consideration & the one with least AIC score will be choose as best fit. Some parameter combinations for the Model are as follow:

Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
For our case study Sorted AIC for respective p, d, q values are as follows:

	param	AIC
8	(2, 1, 2)	2210.623720
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

Table 5 Sorted AIC values for ARIMA

So, the best fit is (2,1,2) with lowest AIC value of 2210.623720.
Now we will use (2,1,2) values as p, d, q value and build the ARIMA model.
ARIMA model summary is as follow:

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.312
Method:	css-mle	S.D. of innovations	1013.589
Date:	Fri, 21 May 2021	AIC	2210.624
Time:	20:24:11	BIC	2227.875
Sample:	02-29-1980	HQIC	2217.634
	- 12-31-1990		

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5845	0.519	10.767	0.000	4.568	6.601
ar.L1.D.Sparkling	1.2699	0.075	17.041	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	-1.9960	0.043	-46.887	0.000	-2.079	-1.913
ma.L2.D.Sparkling	0.9960	0.043	23.336	0.000	0.912	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	1.0003	+0.0000j	1.0003	0.0000
MA.2	1.0037	+0.0000j	1.0037	0.0000

Figure 28 Summary test report of AIC ARIMA model

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value.

RMSE value for ARIMA (2,1,2) is **1374.10845.**

S$ARIMA$ (p, d, q) (P, D, Q, F) **Model**

As we can see while doing the EDA that there is Seasonality in the data. So SARIMA model is specially use for such kind of cases.

- Seasonal ARIMA models are more complex models with seasonal adjustments.
- These models are used when time series data has significant seasonality.
- The most general form of seasonal ARIMA is $ARIMA(p,d,q)*ARIMA(P,D,Q)$[m],
- where P, D, Q are defined as seasonal AR component, seasonal difference and seasonal MA component respectively. And 'm' represents the frequency (time interval) at which the data is observed.
- We will build SARIMA model for seasonality 6 & seasonality 12 and check which is best model.
- We will build the SARIMA model by using the least AIC terms as we build for ARIMA.

For building model, we will consider values as p = q = range (0, 4), d= range (0,2),
D = range (0,2),
So, the Model for all combination given below will be taken under consideration & the one with least AIC score will be choose as best fit.

SARIMA AIC model for Seasonality 6:

There will be total 1030 combination few are listed below:

Model: (0, 0, 1) (0, 0, 1, 6)
Model: (0, 0, 2) (0, 0, 2, 6)
Model: (0, 0, 3) (0, 0, 3, 6)
Model: (0, 1, 0) (0, 1, 0, 6)
Model: (0, 1, 1) (0, 1, 1, 6)
Model: (0, 1, 2) (0, 1, 2, 6)
Model: (0, 1, 3) (0, 1, 3, 6)
Model: (1, 0, 0) (1, 0, 0, 6)
Model: (1, 0, 1) (1, 0, 1, 6)
Model: (1, 0, 2) (1, 0, 2, 6)

Sorted AIC for respective (p, d, q) (P, D, Q) values are as follows:

	param	seasonal	AIC
751	(2, 1, 3)	(1, 1, 3, 6)	1538.983321
239	(0, 1, 3)	(1, 1, 3, 6)	1543.955385
1015	(3, 1, 3)	(2, 1, 3, 6)	1545.078763
495	(1, 1, 3)	(1, 1, 3, 6)	1545.960957
247	(0, 1, 3)	(2, 1, 3, 6)	1547.096308

Table 6 Sorted AIC model for Seasonality 6

So, the best fit is (2,1,3) (1,1,3,6) with lowest AIC value of 1538.983321.

Now we will use (2,1,3) values as p, d, q value and (1,1,3,6) values as P, D, Q, F and build the SARIMA model.

SARIMA model summary is as follow:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(1, 1, 3, 6)	Log Likelihood	-759.492
Date:	Sun, 23 May 2021	AIC	1538.983
Time:	10:57:51	BIC	1565.331
Sample:	0	HQIC	1549.655
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.7445	0.066	-26.360	0.000	-1.874	-1.615
ar.L2	-0.7875	0.071	-11.086	0.000	-0.927	-0.648
ma.L1	0.8598	0.337	2.548	0.011	0.199	1.521
ma.L2	-1.1975	0.128	-9.354	0.000	-1.448	-0.947
ma.L3	-1.1136	0.330	-3.371	0.001	-1.761	-0.466
ar.S.L6	-1.0251	0.008	-133.133	0.000	-1.040	-1.010
ma.S.L6	0.3662	0.275	1.330	0.183	-0.173	0.906
ma.S.L12	-0.7113	0.183	-3.895	0.000	-1.069	-0.353
ma.S.L18	0.1208	0.154	0.783	0.433	-0.181	0.423
sigma2	9.084e+04	7.54e-06	1.2e+10	0.000	9.08e+04	9.08e+04

Ljung-Box (Q):	22.58	Jarque-Bera (JB):	12.03
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.38	Skew:	0.41
Prob(H) (two-sided):	0.36	Kurtosis:	4.46

Figure 30 Summary test report of AIC SARIMA seasonality 6 model

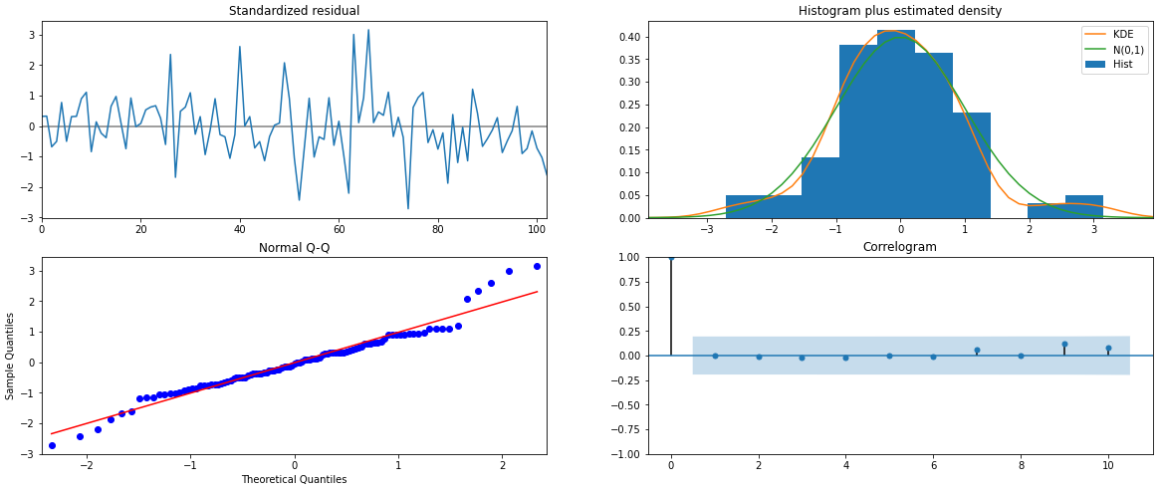


Figure 31 Model Performance check dist plot (SARIMA Seasonality 6)

From the above plot we can see that the Residual are around the zero line.

The histogram is normalized.
The near about all points in the Q-Q plot lies on the line. We can see there are no significant lags in the series.
So this model is quite good enough.

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (2,1,3) (1,1,3,6) is **790.481.**

This is less than the ARIMA model. So, we can say performance of AIC SARIMA model with seasonality 6 is better than the AIC ARIMA model.

SARIMA AIC model for Seasonality 12:

There will be total 1030 combination few are listed below:

Model: (0, 0, 1) (0, 0, 1, 12)
Model: (0, 0, 2) (0, 0, 2, 12)
Model: (0, 0, 3) (0, 0, 3, 12)
Model: (0, 1, 0) (0, 1, 0, 12)
Model: (0, 1, 1) (0, 1, 1, 12)
Model: (0, 1, 2) (0, 1, 2, 12)
Model: (0, 1, 3) (0, 1, 3, 12)
Model: (1, 0, 0) (1, 0, 0, 12)
Model: (1, 0, 1) (1, 0, 1, 12)
Model: (1, 0, 2) (1, 0, 2, 12)

Sorted AIC for respective (p, d, q) (P, D, Q) values are as follows:

	param	seasonal	AIC
1020	(3, 1, 3)	(3, 1, 0, 12)	1213.282554
1021	(3, 1, 3)	(3, 1, 1, 12)	1215.213334
956	(3, 1, 1)	(3, 1, 0, 12)	1215.898777
1022	(3, 1, 3)	(3, 1, 2, 12)	1216.479955
988	(3, 1, 2)	(3, 1, 0, 12)	1216.859172

Table 7 Sorted AIC model for Seasonality 12

So, the best fit is (3,1,3) (3,1,0,12) with lowest AIC value of 1213.282554

Now we will use (3,1,3) values as p, d, q value and (3,1,0,12) values as P, D, Q, F and build the SARIMA model.

SARIMA model summary is as follow:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 3)x(3, 1, [], 12)	Log Likelihood	-596.641
Date:	Sun, 23 May 2021	AIC	1213.283
Time:	12:26:46	BIC	1237.103
Sample:	0	HQIC	1222.833
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6142	0.176	-9.169	0.000	-1.959	-1.269
ar.L2	-0.6125	0.299	-2.047	0.041	-1.199	-0.026
ar.L3	0.0859	0.161	0.535	0.593	-0.229	0.401
ma.L1	0.9858	0.464	2.122	0.034	0.075	1.896
ma.L2	-0.8730	0.166	-5.261	0.000	-1.198	-0.548
ma.L3	-0.9459	0.482	-1.964	0.050	-1.890	-0.002
ar.S.L12	-0.4520	0.142	-3.193	0.001	-0.730	-0.175
ar.S.L24	-0.2339	0.144	-1.620	0.105	-0.517	0.049
ar.S.L36	-0.1005	0.122	-0.827	0.408	-0.339	0.138
sigma2	1.839e+05	8.84e+04	2.080	0.037	1.06e+04	3.57e+05

Ljung-Box (Q):	23.19	Jarque-Bera (JB):	4.07
Prob(Q):	0.98	Prob(JB):	0.13
Heteroskedasticity (H):	0.73	Skew:	0.48
Prob(H) (two-sided):	0.41	Kurtosis:	3.54

Figure 32 Summary test report of AIC SARIMA seasonality 12 model

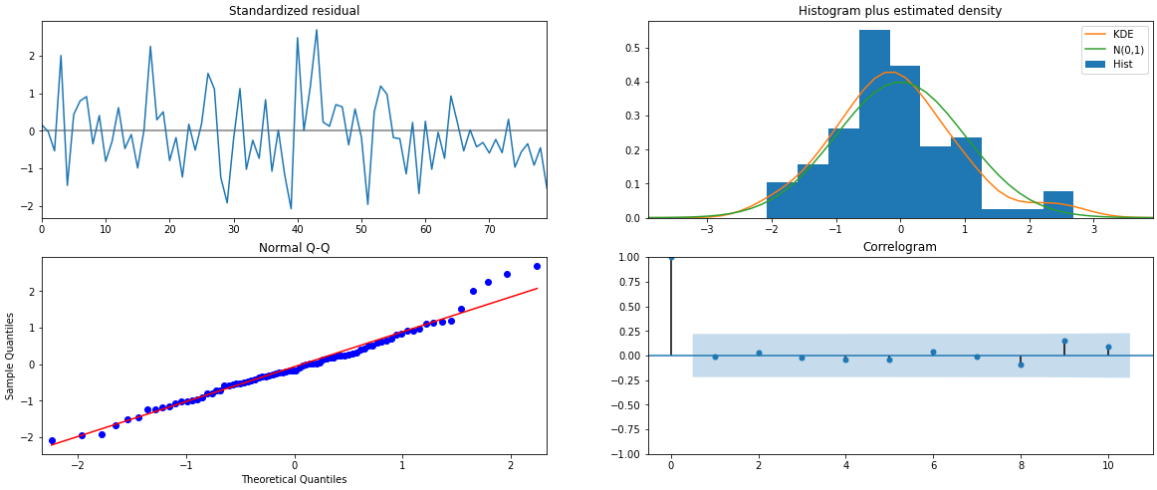


Figure 33 Model Performance check dist plot (SARIMA Seasonality 12)

From the above plot we can see that the Residual are around the zero line.
The histogram is normalized.
The near about all points in the Q-Q plot lies on the line.
We can see there are no significant lags in the series.

So, this model is quite good enough.

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (3,1,3) (3,1,0,12) is **332.1511.**

This is less than the ARIMA & SARIMA seasonality 6 model. So, we can say performance of AIC SARIMA model with seasonality 12 is better than the AIC ARIMA & SARIMA seasonality 6 model.

Sr No	Model	RMSE
1	AIC ARIMA (2,1,2)	1374.10845
2	AIC SARMA (Seasonality 6) (2,1,3) (1,1,3,6)	790.481
3	AIC SARMA (Seasonality 12) (3,1,3) (3,1,0,12)	332.1511

Table 8 Comparison of AIC models RMSE values

So, among the models build using the lowest Akaike Information Criteria (AIC) SARIMA model with seasonality 12 is performing best for our case study.

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

ARIMA Model based on the cut-off points of ACF and PACF.

Another way to build the model is by taking the (p, d, q) (P, D, Q) values manually by observing the ACF and PACF plot.
As we know that to make train data stationary, we have taken the difference of 1.
So now plotting the ACF and PACF plot of this differenced train data time series

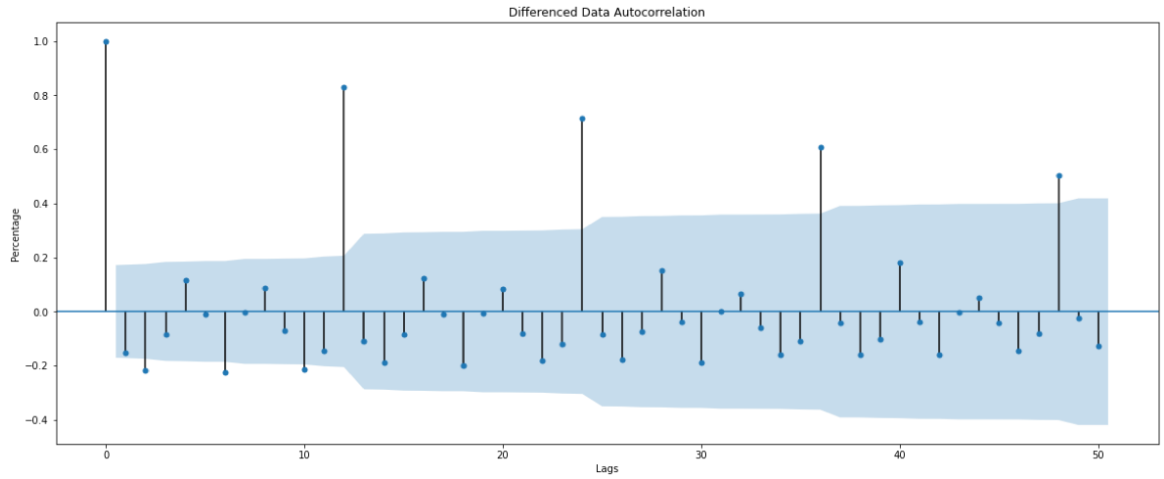


Figure 33 Differential ACF Plot of Training data

From the ACF plot we can see the significant lag count is 0 ,so we can take **q value as 0**.

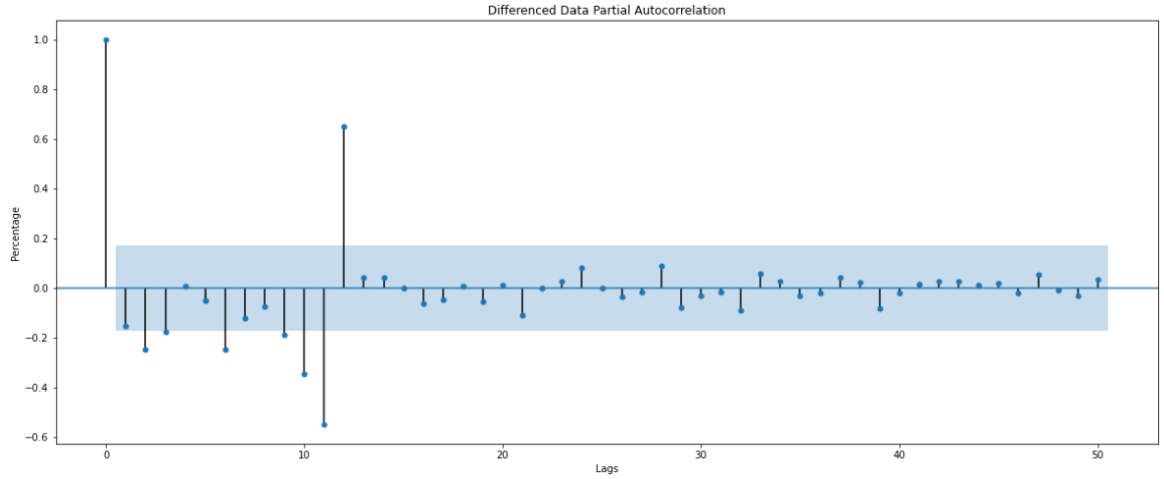


Figure 34 Differential PACF Plot of Training data

From the PACF plot we can see the significant lag count is 0, so we can take **p value as 0.** So, the order will be (0,1,0)

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.791
Method:	css	S.D. of innovations	1377.911
Date:	Sat, 22 May 2021	AIC	2269.583
Time:	23:11:22	BIC	2275.333
Sample:	02-29-1980 - 12-31-1990	HQIC	2271.919

	coef	std err	z	P>|z|	[0.025	0.975]
const	33.2901	120.389	0.277	0.782	-202.667	269.248

Figure 35 Summary test report of ARIMA model based on the cutoff points of ACF and PCF plot

MODEL EVALUATION

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for ARIMA (0,1,0) is **4779.154299.**

SARIMA Model with Seasonality 6 based on the cut-off points of ACF and PACF.

We have taken differentiation of 6 for Seasonality of 6 so diff (6) is D=1
Time Series plot after taking diff (6)

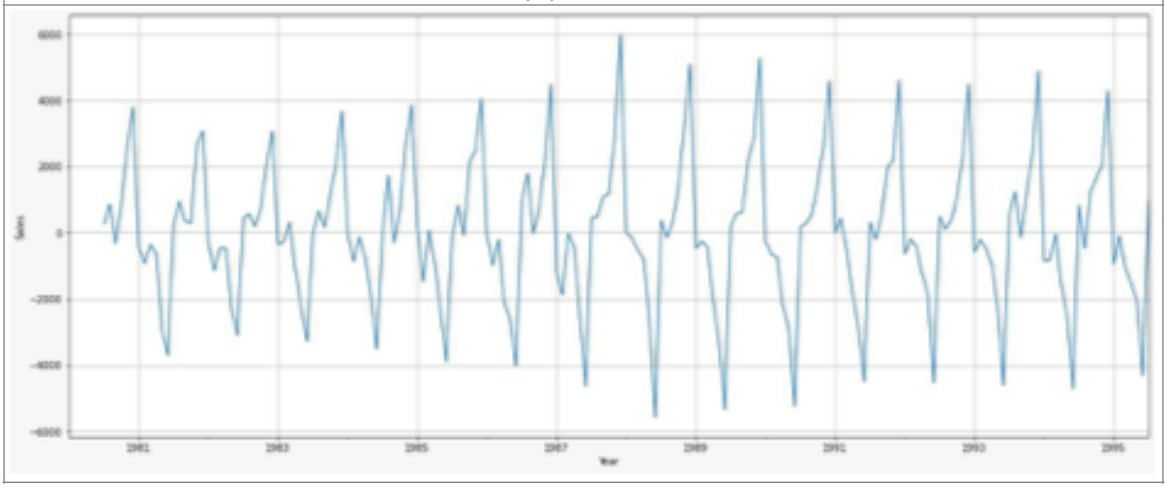


Figure 36 Time Series Plot after Differentiation of 6

Applying the diff (6) on Train series and checking stationary property

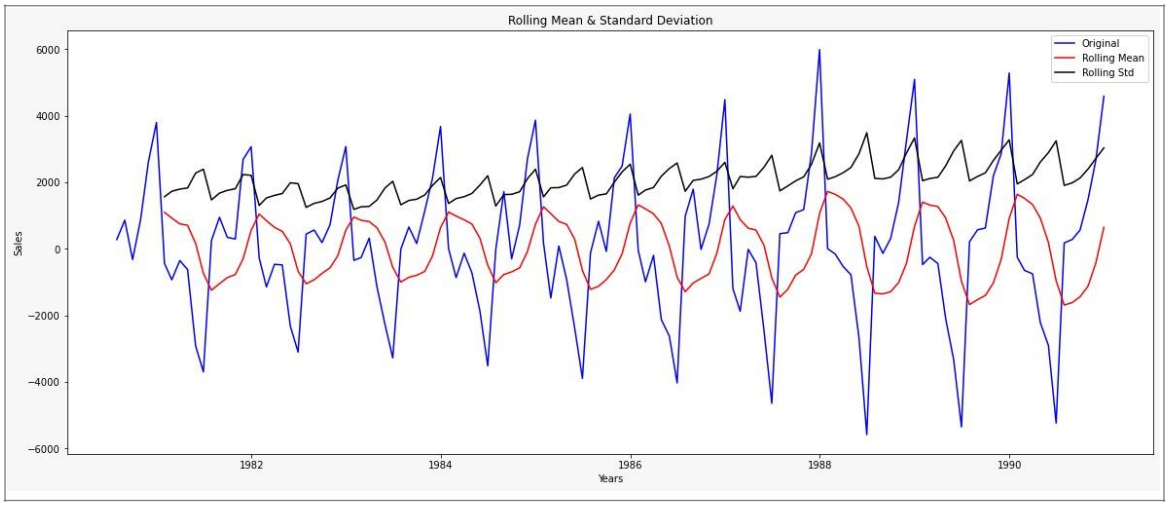


Figure 37 Time Series Stationary check after Differentiation of 6

```
Results of Dickey-Fuller Test:
Test Statistic                  -8.181919e+00
p-value                          8.088278e-13
#Lags Used                       6.000000e+00
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
```

Figure 38 ADF test report SARIMA model seasonality 6 based on the cutoff points of ACF and PCF plot

So we can say from p value that train series is stationary after taking seasonal diff (6)

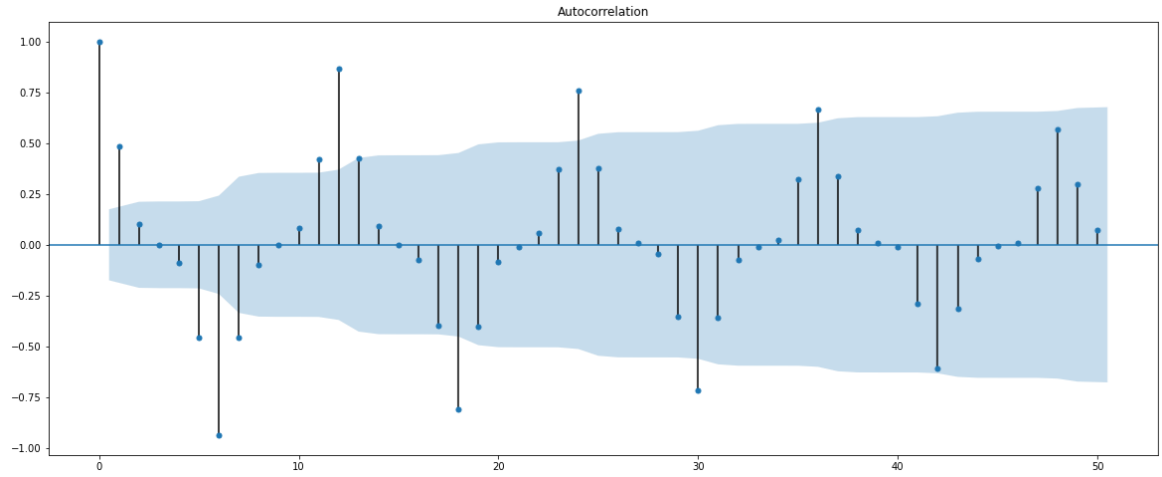
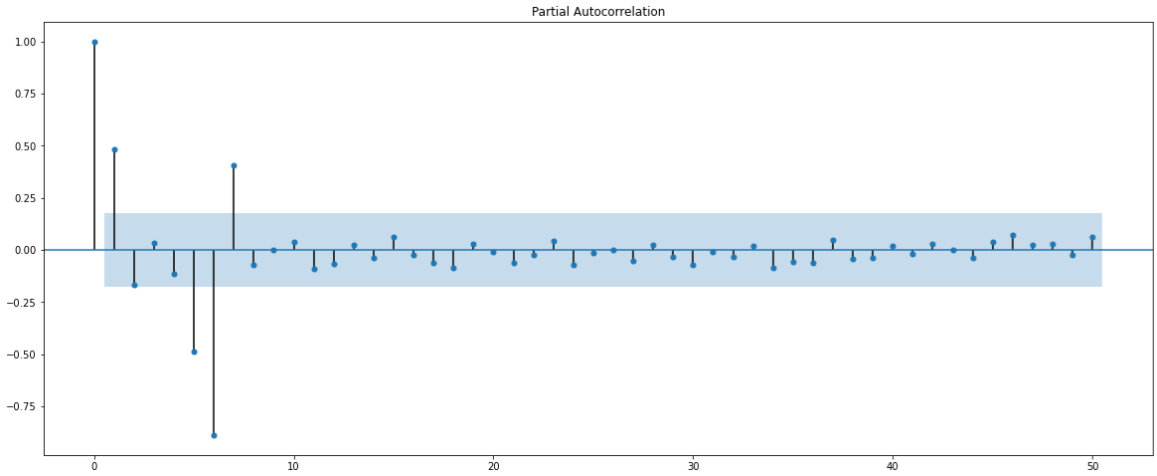


Figure 39 ACF Plot after differentiation of 6

From the ACF plot we can see the significant lag count is 1,so we can take **Q value as 1.**



MODEL EVALUATION

Figure 40 PACF Plot after differentiation of 6

From the PACF plot we can see the significant lag count is 1, so we can take **P value as 1.** So, the order will be (0,1,0) (1,1,1,6)
SARIMA Model summary is as follows: -

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 0)x(1, 1, [1], 6)	Log Likelihood	-909.770
Date:	Sat, 22 May 2021	AIC	1825.540
Time:	23:41:02	BIC	1833.853
Sample:	0	HQIC	1828.915
	132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L6	-0.9688	0.028	-34.608	0.000	-1.024	-0.914
ma.S.L6	-0.1158	0.117	-0.991	0.322	-0.345	0.113
sigma2	3.031e+05	2.87e+04	10.577	0.000	2.47e+05	3.59e+05

Ljung-Box (Q):	59.84	Jarque-Bera (JB):	39.20
Prob(Q):	0.02	Prob(JB):	0.00
Heteroskedasticity (H):	1.13	Skew:	0.69
Prob(H) (two-sided):	0.70	Kurtosis:	5.46

Figure 41 Summary test report of SARIMA seasonality 6 model based on the cutoff points of ACF and PCF plot

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (0,1,0) (1,1,1,6) is **1472.075509.**

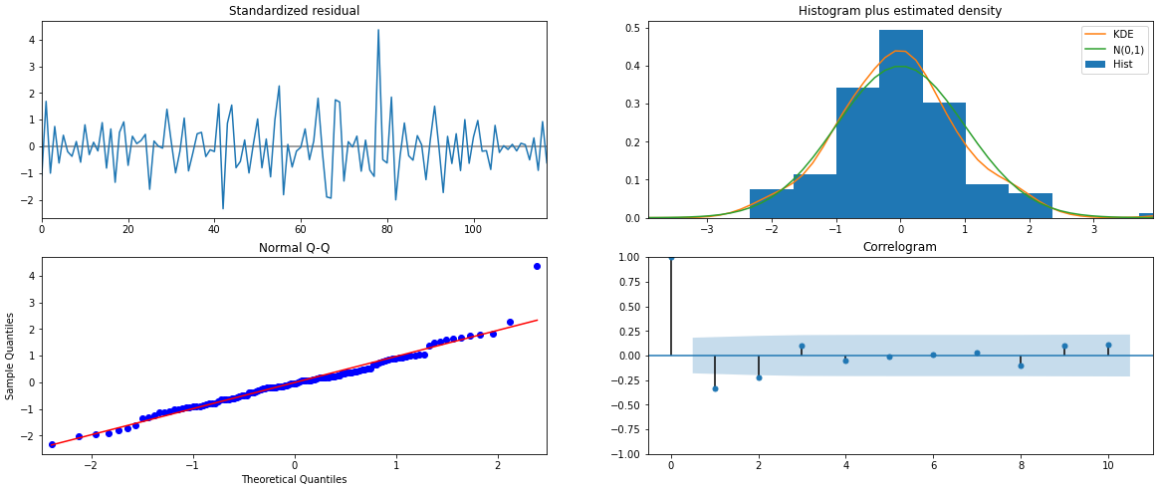


Figure 42 Model Performance check dist plot (SARIMA Seasonality 6) based on cutoff point of ACF and PACF plot

From the above plot we can see that the Residual are around the zero line. The histogram is normalized.

The near about all points in the Q-Q plot lies on the line. We can see there is one significant lag in the series.
So, this model is not good enough.

SARIMA Model with Seasonality 12 based on the cut-off points of ACF and PACF.

We have taken differentiation of 12 for Seasonality of 12 so diff (12) is D=2
Time Series plot after taking diff (12)

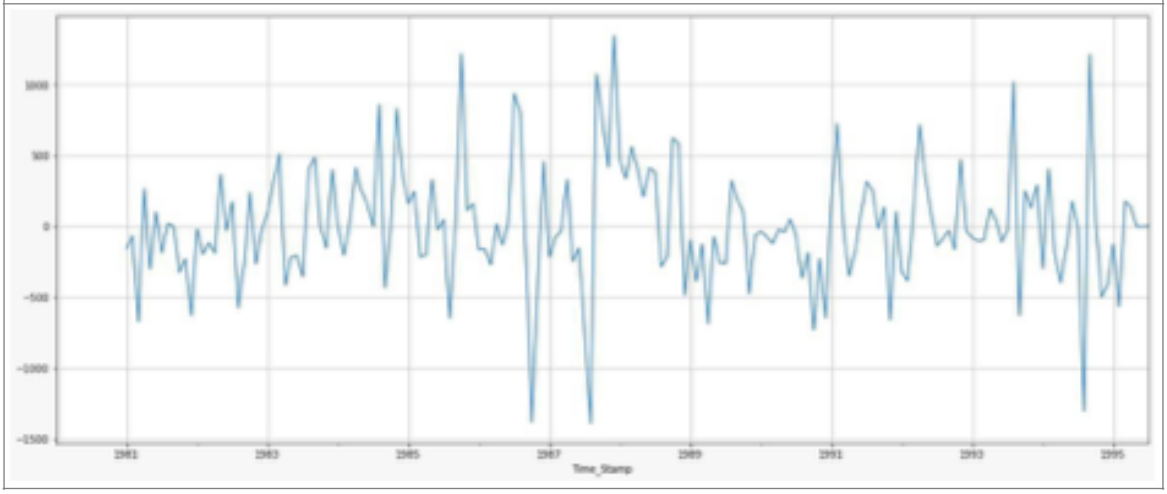
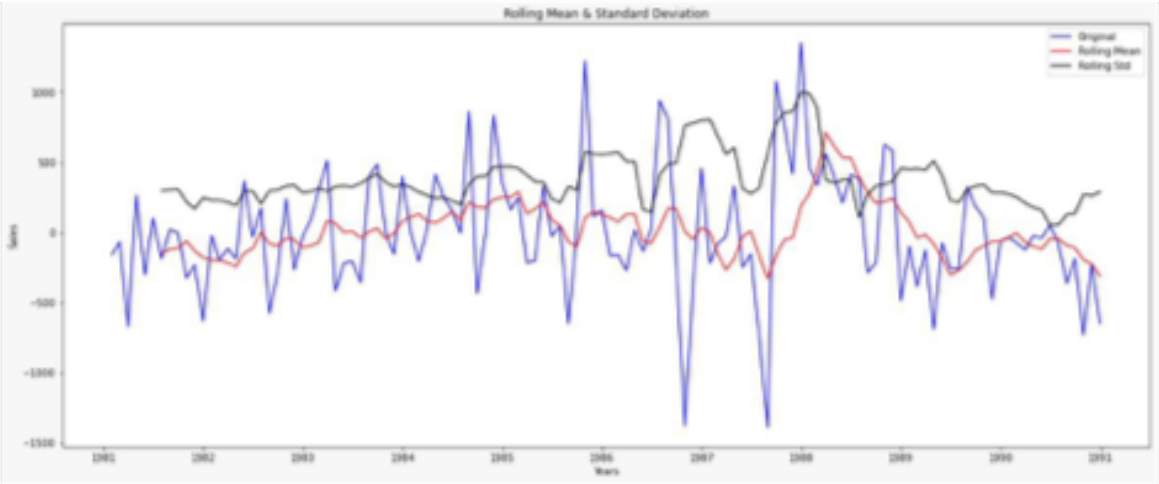


Figure 43 Time Series Plot after Differentiation of 12



Applying the diff (12) on Train series and checking stationary property

Figure 44 Time Series Stationary check after Differentiation of 12

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.136812
p-value                          0.023946
#Lags Used                      11.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

Figure 45 ADF test report SARIMA model seasonality 12 based on the cutoff points of ACF and PCF plot

So we can say from p value that train series is stationary after taking seasonal diff (12)

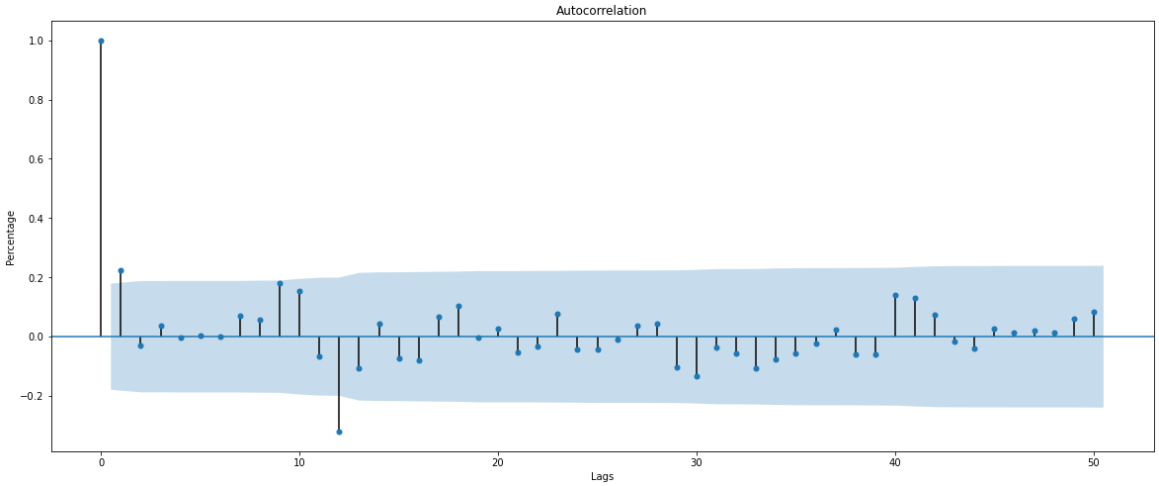


Figure 46 ACF Plot after differentiation of 12

From the ACF plot we can see the significant lag count is 1 ,so we can take **Q value as 1.**

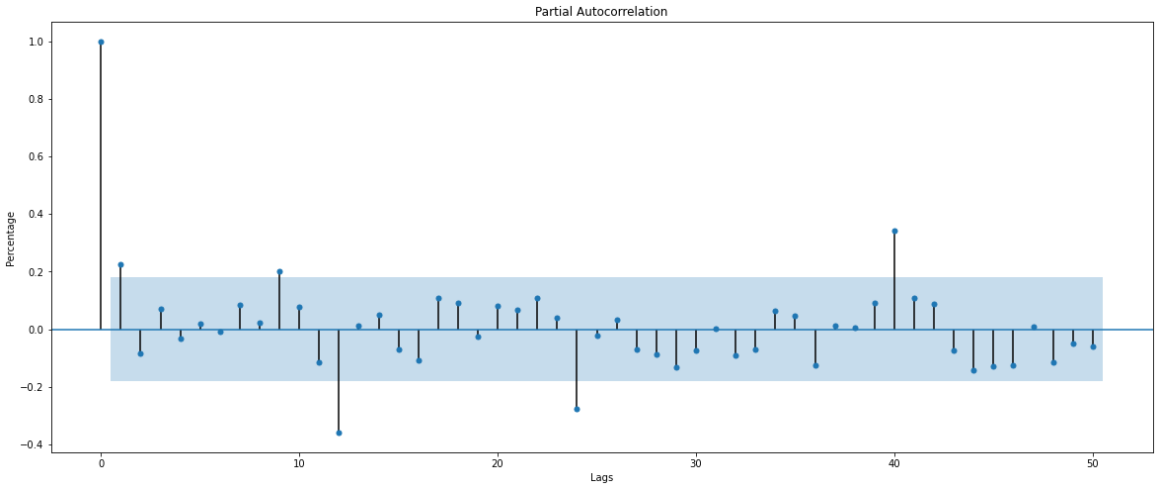


Figure 47 PACF Plot after differentiation of 12

MODEL EVALUATION

From the PACF plot we can see the significant lag count is 1, so we can take **P value as 1.**

So, the order will be (0,1,0) (1, 2, 1, 12)

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 0)x(1, 2, [1], 12)	Log Likelihood	-733.205
Date:	Sat, 22 May 2021	AIC	1472.410
Time:	23:54:37	BIC	1480.039
Sample:	0	HQIC	1475.491
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L12	-0.4021	0.087	-4.643	0.001	-0.474	-0.141
ma.S.L12	-1.0004	0.114	-8.797	0.000	-1.223	-0.778
sigma2	2.876e+05	3.95e-07	7.28e+11	0.000	2.88e+05	2.88e+05

Ljung-Box (Q):	66.07	Jarque-Bera (JB):	23.75
Prob(Q):	0.01	Prob(JB):	0.00
Heteroskedasticity (H):	0.89	Skew:	0.60
Prob(H) (two-sided):	0.76	Kurtosis:	5.15

Figure 48 Summary test report of SARIMA seasonality 12 model based on the cutoff points of ACF and PCF plot

To evaluate the model performance, we need to calculate the RMSE value. RMSE value for SARIMA (0,1,0) (1,1,1,6) is **3592.015916641514.**

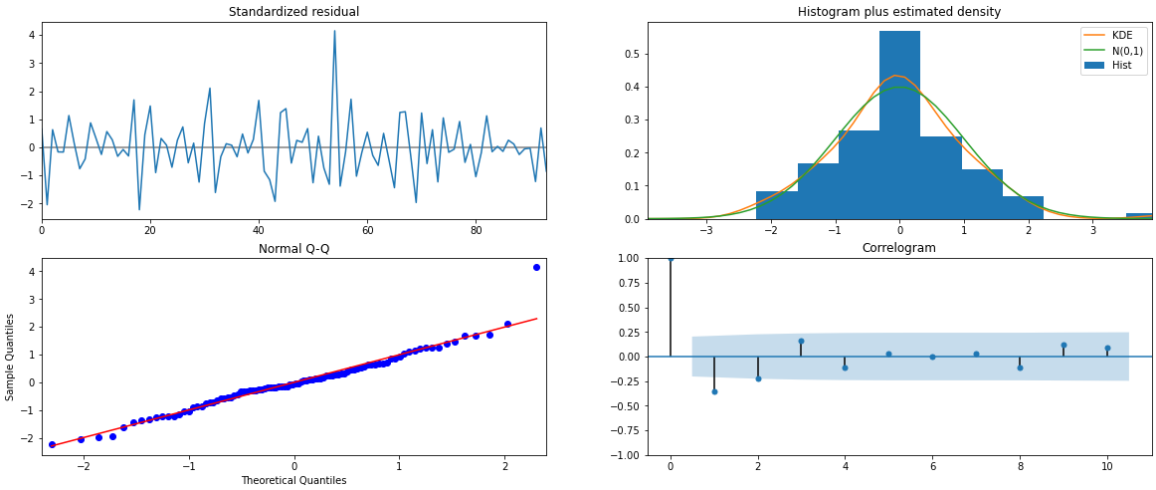


Figure 49 Model Performance check dist plot (SARIMA Seasonality 12) based on cutoff point of ACF and PACF plot

From the above plot we can see that the Residual are around the zero line.
The histogram is normalized.
The near about all points in the Q-Q plot lies on the line.
But we can see there is one significant lag in the series.

So, this model is not good enough.

Sr No	Model	RMSE
1	Manual ARIMA	4779.154299
2	Manual SARIMA at seasonality 6	1472.075509
3	Manual SARIMA at seasonality 12	3592.015916

Table 9 RMSE values comparison based on cutoff points of ACF and PACF

So, among ARIMA/SARIMA models based on the cut-off points of ACF and PACF Manual SARIMA at seasonality 6 is better among all, but as we have seen it above plot the series have one significant lag. so, this model is not as best fit for our case study.

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

Sr. No	Model	RMSE
1	auto SARIMA(3,1,3)(3,1,0,12)	332.1512
2	Alpha=0.154,Beta=2.683,Gamma=0.371,TripleExponentialSmoothing	383.1923
3	Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.7862
4	auto SARIMA(2,1,3)(1,1,3,6)	790.481
5	2pointTrailingMovingAverage	813.4007
6	4pointTrailingMovingAverage	1156.59
7	SimpleAverageModel	1275.082
8	Alpha=default value,SimpleExponentialSmoothing	1275.082
9	RegressionOnTime	1275.867
10	6pointTrailingMovingAverage	1283.927
11	9pointTrailingMovingAverage	1346.278
12	ARIMA(2,1,2)	1374.108
13	manual SARIMA(0,1,0)(1,1,1,6)	1472.076
14	Alpha=0.3,SimpleExponentialSmoothing	1935.507
15	manual SARIMA(0,1,0)(1,2,1,12)	3592.016
16	Alpha=0.647,Beta=0,default DoubleExponentialSmoothing	3851.331
17	NaiveModel	3864.279
18	ARIMA(0,1,0)	4779.154
19	Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11

Table 10 All models RMSE results

Among all the model the AIC SARIMA (3,1,3)(3,1,0,12) model RMSE value is least. So for this case study the Triple Exponential model is the best fit.

Q.9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

- The most optimum model among all is AIC SARIMA (3,1,3)(3,1,0,12).
- The SARIMA model is built on whole data for future forecast.
- While forecasting gives the step of 12 as we have to forecast for next 12 months.
- After building the model, the summary report is as follows: -

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(3, 1, 3)x(3, 1, [], 12)	Log Likelihood	-998.042
Date:	Sun, 23 May 2021	AIC	2016.083
Time:	12:52:14	BIC	2045.136
Sample:	01-31-1980	HQIC	2027.890
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.0049	0.104	-9.705	0.000	-1.208	-0.802
ar.L2	-0.8247	0.102	-8.057	0.000	-1.025	-0.624
ar.L3	0.1034	0.088	1.171	0.242	-0.070	0.277
ma.L1	0.2015	0.105	1.923	0.054	-0.004	0.407
ma.L2	-0.1297	0.090	-1.445	0.148	-0.306	0.046
ma.L3	-0.9668	0.090	-10.785	0.000	-1.143	-0.791
ar.S.L12	-0.5570	0.074	-7.523	0.000	-0.702	-0.412
ar.S.L24	-0.2808	0.118	-2.372	0.018	-0.513	-0.049
ar.S.L36	-0.1629	0.088	-1.859	0.063	-0.335	0.009
sigma2	1.488e+05	9.65e-07	1.54e+11	0.000	1.49e+05	1.49e+05

Ljung-Box (Q):	20.72	Jarque-Bera (JB):	42.48
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	0.55	Skew:	0.73
Prob(H) (two-sided):	0.05	Kurtosis:	5.33

Figure 50 Summary result for full model forecasting

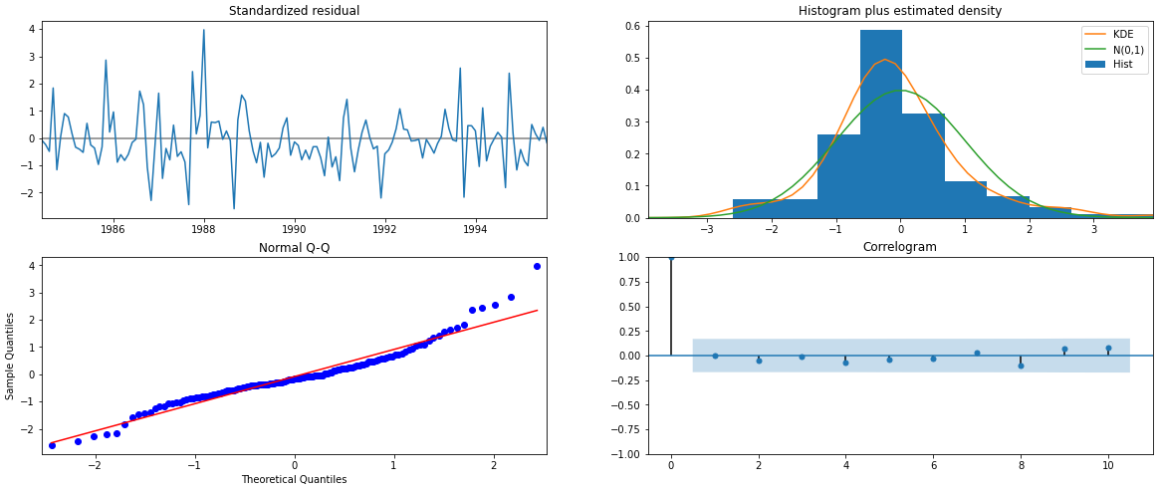


Figure 51 Model Performance check dist plot for full model forecasting

- From the above plot we can see that the Residual are around the zero line.
- The histogram is normalized.
- The near about all points in the Q-Q plot lies on the line.
- But we can see there is no significant lag in the series.
- So, this model is good enough.

Confidence Band/Interval:

- When we predict certain time points into the future, we might need to have a concept of Confid ence Band for our predictions. This gives us range of values in which our predictions will be lying in thefuture or for the future time stamps.
- Hence the upper and lower confidence bands at 95% confidence level are calculated.

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1930.727126	388.361180	1169.553201	2691.901050
1995-09-30	2399.790645	395.206473	1625.200190	3174.381099
1995-10-31	3332.383341	395.556028	2557.107772	4107.658910
1995-11-30	3870.532438	395.566886	3095.235589	4645.829288
1995-12-31	6090.890063	396.763842	5313.247222	6868.532904

Table 11 Upper and Lower Confidence band

Model Evaluation:

Calculating the RMSE value for the full model build RMSE of the Full Model 612.7458479527716

Forecast Graph with the confidence band

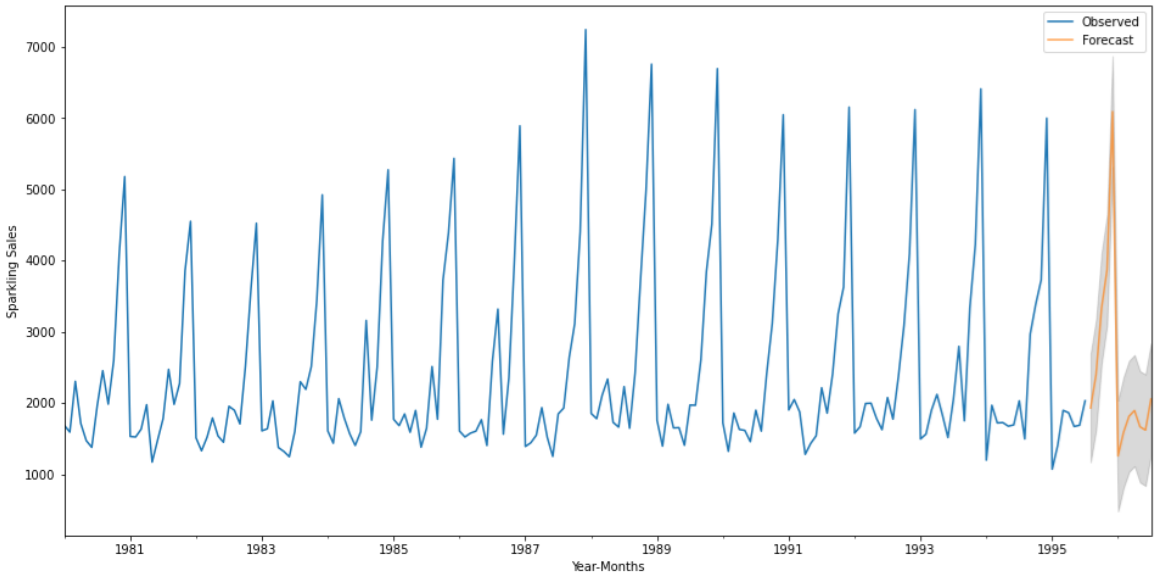


Figure 52 Forecast Graph with the confidence band

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

Summary the various steps performed in this project:

1. Do EDA, if any missing value present do the missing value treatment.
2. Plot the time series.

3. There are two approaches to proceed do Decomposition or ARIMA.

Decomposition Approach

1. See if trend & seasonality is present in the data.

2. If trend & seasonality is not in the data then apply linear regression, naive, simple
average, moving averages, simple exponential smoothing.

3. Check Accuracy of the model by calculating RMSE.

4. If only trend is available in the data, then apply Double Exponential Smoothing
model & check accuracy.

5. If both trend and seasonality is present, then apply Triple Exponential Smoothing
model & check accuracy.

ARIMA approach:

1. If data is not stationary, then do differencing and make series stationary.

2. If data is stationary, then plot the ACF &PACF plot

3. If the data is seasonal then apply ARIMA (p, d, q) (P, D, Q) F model & check
accuracy.

4. If the data is not seasonal then apply ARIMA (p, d, q) model & check accuracy.

Comment on the model

- We have built different model, among all the model the AIC SARIMA model for seasonality of 12 RMSE value is least.

- So, for this case study the AIC SARIMA model for seasonality of 12 is the best fit.

- The forecasting the wine sale using this model is good enough, the confidence band is not much

away from the actual forecast.

Findings

- There is no trend in the data, that means we can say that there is no extreme increase or decrease in the sales over the years.
- The sale of sparkling wine is lowest in month of June & highest in the month of December.
- We can see that after July the Sparkling wine sales goes on increasing every month.
- After month of July the Sale start increasing. From January to July the sales are not much, that means there is seasonality over here.
- In a year during January to July sale will be less and July to December is the season for good sale.

Measures that the company should be taking for future sales.

- So, as we come to known that the December is peak point for sale so accordingly production and stock of the Sparkling wine must be maintained in the month of the December.
- Company must be able to grab this December season and increase the sale and earn profits.
- Again, we can see that from January to June Sale is low among all the years, so some strategy must be planned to increase the sale, such as some discounts/ offers, so that the sale will get increase in those respective months.
- As we can see year by year the sale is near about same, it's not increasing but as it is also not decreasing is a good thing. But to increase the sale we need to do the study of the Market Place where the wine is sale. Need to find out the reason why the demand is not increasing or vice versa need to think what can be done so that the demand will be increase.
- To increase demand many aspects can be considered such as Taste of the wine, cost, Market area, crowd in the locality. If we found out the proper liking of the people & if we could do any work around it then surely, we can increase the sale of Sparkling Wine of ABC Estate in future.

THE END