2022

Time Series Forecasting Business Report

Yogender Singh
Great Learning
10/9/2022

Contents

INTRODUCTION

This report consists of Time Series analysis and forecasting of 1 dataset -

- DATASET 2 - Sales data of Sparkling Wine

Please find the Jupyter Code Notebook. Analysis code is in Python.

PROBLEM – Sparkling Wine Sales

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

Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century.

Datasets used - Sales of Sparkling Wine

SYNOPSIS

1. Total No. Of Sparkling Data Entries = 187

   No. Of Missing Values in Sparkling data = 0

   No. Of Duplicate entries in Sparkling data = 0

2. Both datasets are split in Train : Test at year 1991 - Test data starts at 1991

3. Various forecasting models applied are -
   1. Linear Regression
   2. Naive Bayes
   3. Simple Average
   4. 2-pt Moving Average

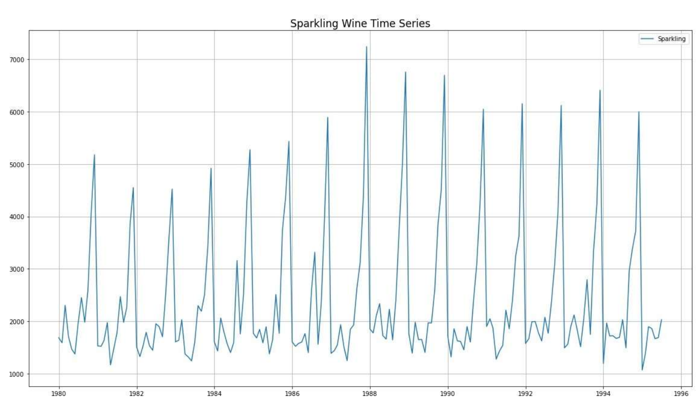
5. 4-pt Moving Average
6. 6-pt Moving Average
7. 9-pt Moving Average
8. Single Exponential Smoothing
9. Double Exponential Smoothing (Holt's Model)
10. Triple Exponential Smoothing (Holt-Winter Model)
11. ARIMA / SARIMA (Auto fitted)
12. ARIMA / SARIMA (Manually fitted)

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

- Both Datasets are read and stored as Pandas Data Frames for analysis
- Datasets are read as Time Series data using parse_dates=True& index_col='YearMonth'
- First 5 rows of Sparkling data are given below -

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

- Sparkling Data plot -



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

✦ Exploratory Data Analysis -

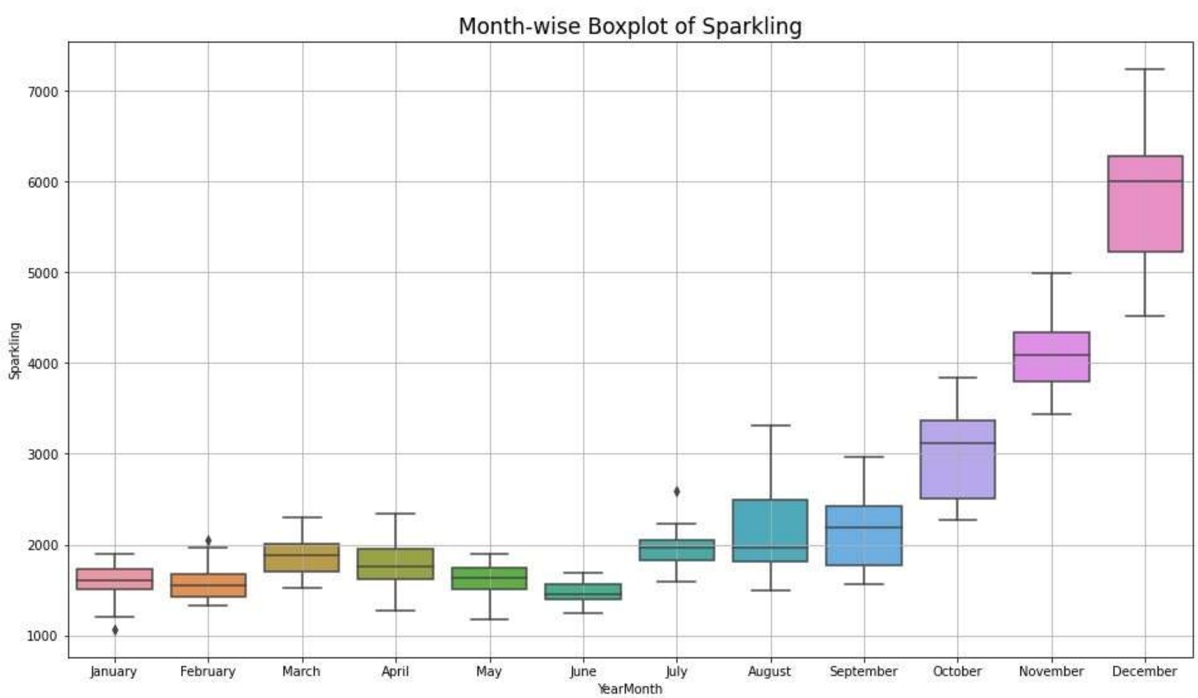
	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417	1295.112	1070.0	1605.0	1874.0	2549.0	7242.0

Descriptive Stats Sparkling datasets

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries,
1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #    Column       Non-Null Count     Dtype
---   -------      ---------------    -----
 0    Sparkling    187 non-null       int64
dtypes: int64(1)
memory usage: 2.9 KB
```

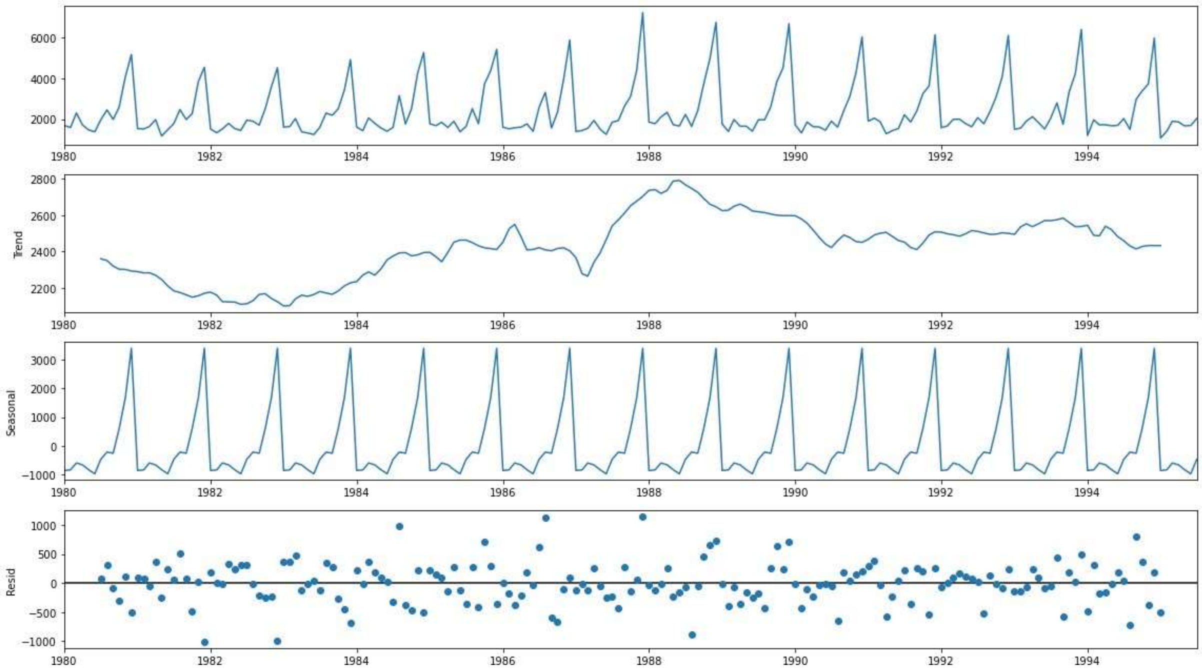
Info - Sparkling data

- Month-wise Boxplot of Sparkling -

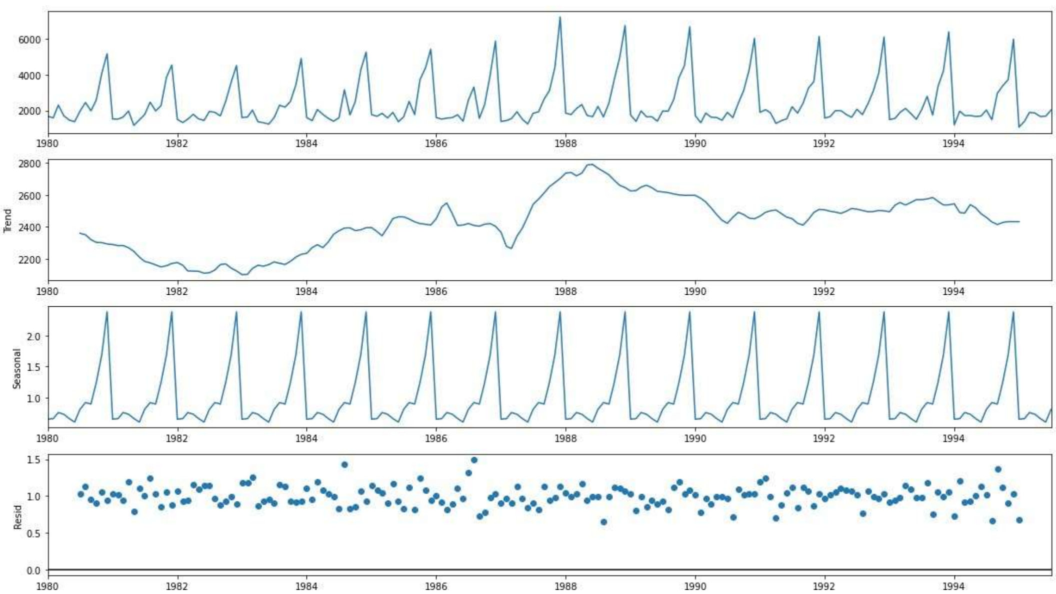


- Sales of Sparkling, show a spike in the last quarter of Oct to Dec
- This spike may be due to the Holiday season starting in Oct.

✦ Additive Decomposition of Sparkling -



✦ Multiplicative Decomposition of Sparkling -



YearMonth	trend
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	2360.67
1980-08-01	2351.33
1980-09-01	2320.54
1980-10-01	2303.58
1980-11-01	2302.04
1980-12-01	2293.79

Sparking Trend

YearMonth	seasonal
1980-01-01	0.65
1980-02-01	0.66
1980-03-01	0.76
1980-04-01	0.73
1980-05-01	0.66
1980-06-01	0.60
1980-07-01	0.81
1980-08-01	0.92
1980-09-01	0.89
1980-10-01	1.24
1980-11-01	1.69
1980-12-01	2.38

Sparking Seasonality

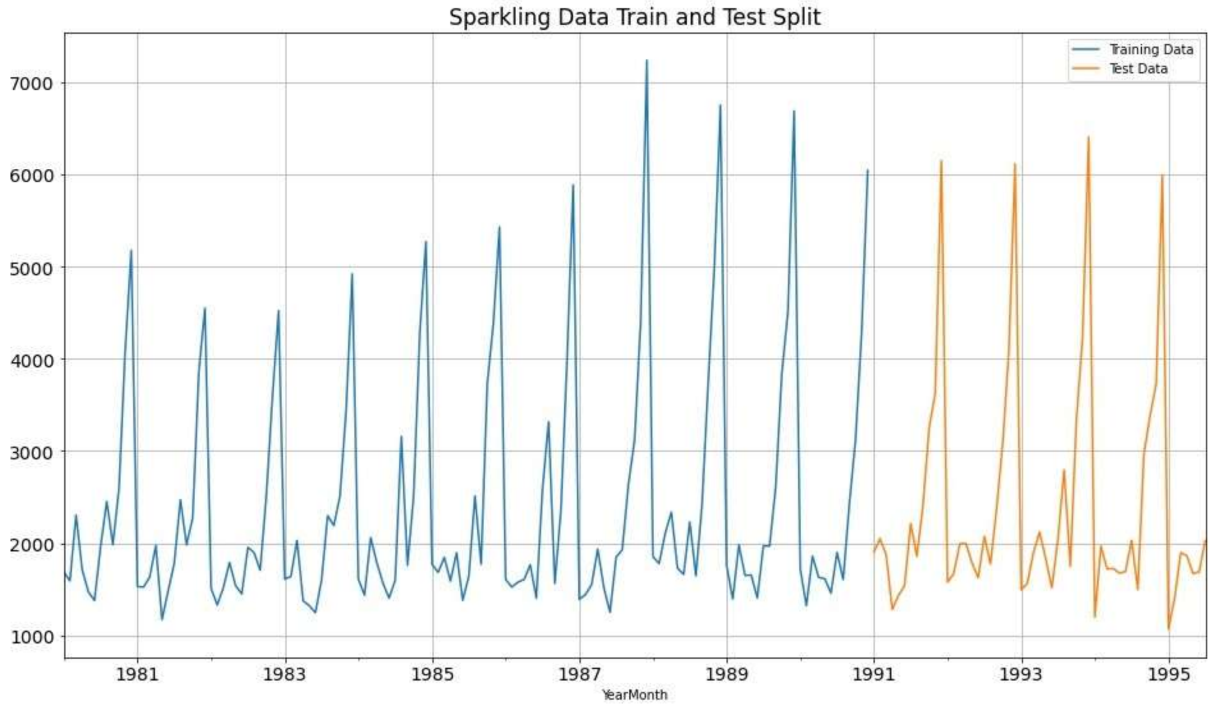
YearMonth	resid
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	1.03
1980-08-01	1.14
1980-09-01	0.96
1980-10-01	0.91
1980-11-01	1.05
1980-12-01	0.95

Sparking Residue

- Additive Models -
    - The seasonality is relatively constant over time
    - $y_t = Trend + Seasonality + Residual$
- Multiplicative Models -
    - The seasonality increases or decreases over time. It is proportionate to the trend
    - $y_t = Trend * Seasonality * Residual$
- Here by just observing the patterns of Additive and Multiplicative models of Sparkling datasets. It seems that -
    - Sparkling is Additive

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

- datasets of Rose and Sparkling are split at the year 1991
- Test datasets start at 1991



• Sparkling dataset - TRAIN

YearMonth	Sparkling
1980-01-01	1686.00
1980-02-01	1591.00
1980-03-01	2304.00
1980-04-01	1712.00
1980-05-01	1471.00

Sparkling Train - First 5

YearMonth	Sparkling
1990-08-01	1605.00
1990-09-01	2424.00
1990-10-01	3116.00
1990-11-01	4286.00
1990-12-01	6047.00

Sparkling Train - Last 5

• Sparkling dataset - TEST

YearMonth	Sparkling
1991-01-01	1902.00
1991-02-01	2049.00
1991-03-01	1874.00
1991-04-01	1279.00
1991-05-01	1432.00

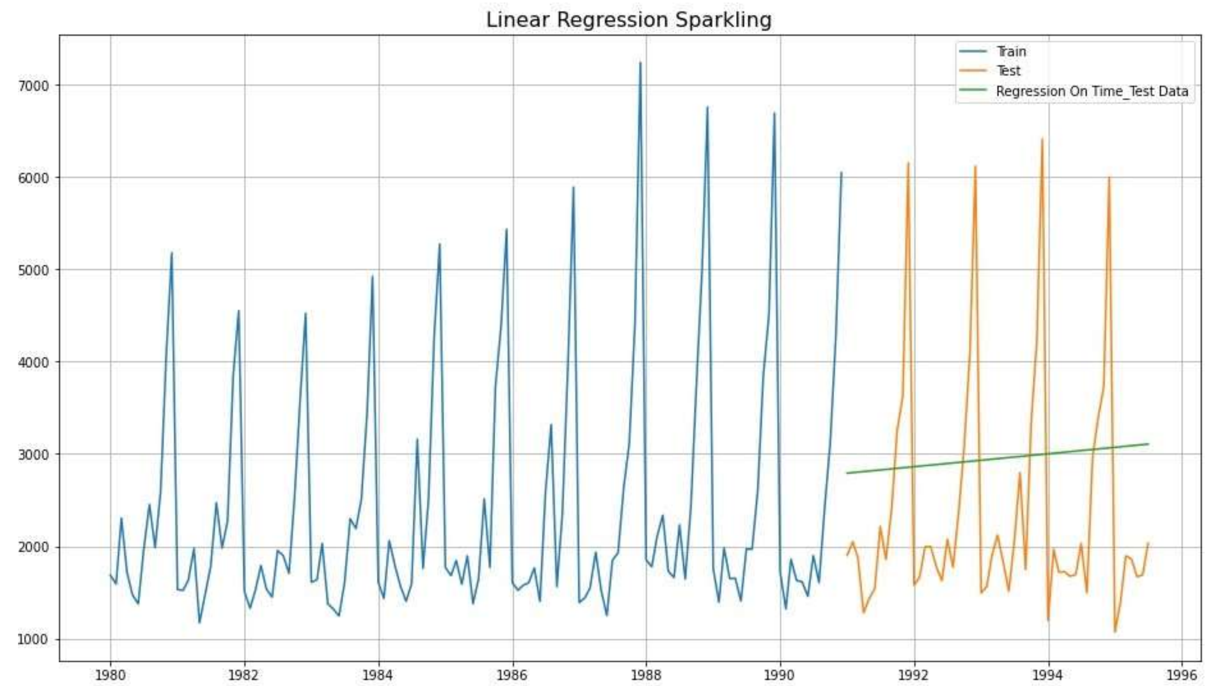
Sparkling Test - First 5

YearMonth	Sparkling
1995-03-01	1897.00
1995-04-01	1862.00
1995-05-01	1670.00
1995-06-01	1688.00
1995-07-01	2031.00

Sparkling Test - Last 5

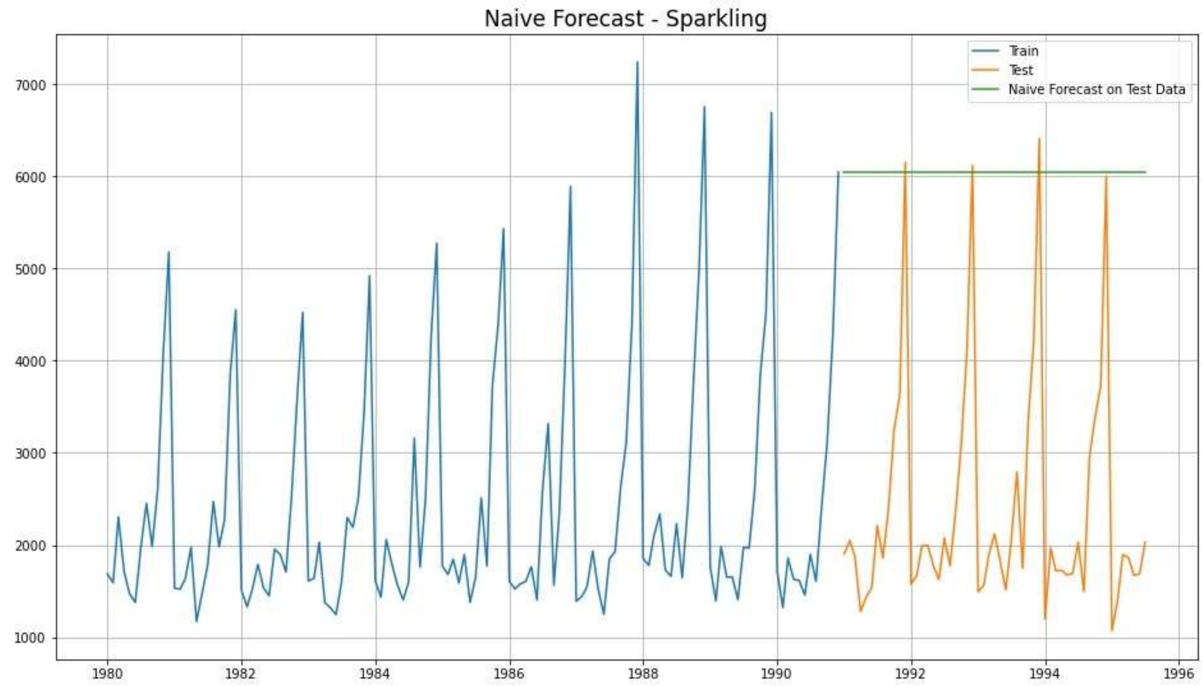
[Q 4] Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

✦ Model 1 - Linear Regression



	Test RMSE Sparkling
RegressionOnTime	1389.14

✦ Model 2 - Naive Bayes

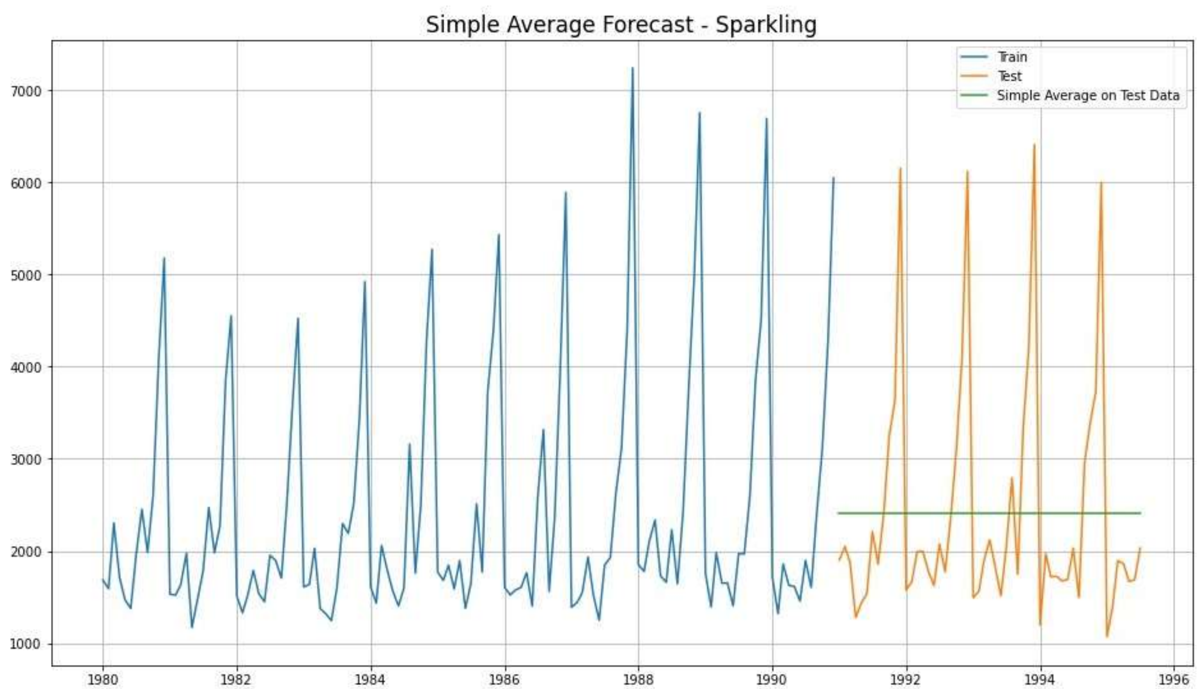


	Test RMSE Sparkling
RegressionOnTime	1389.14
Naive Model	3864.28

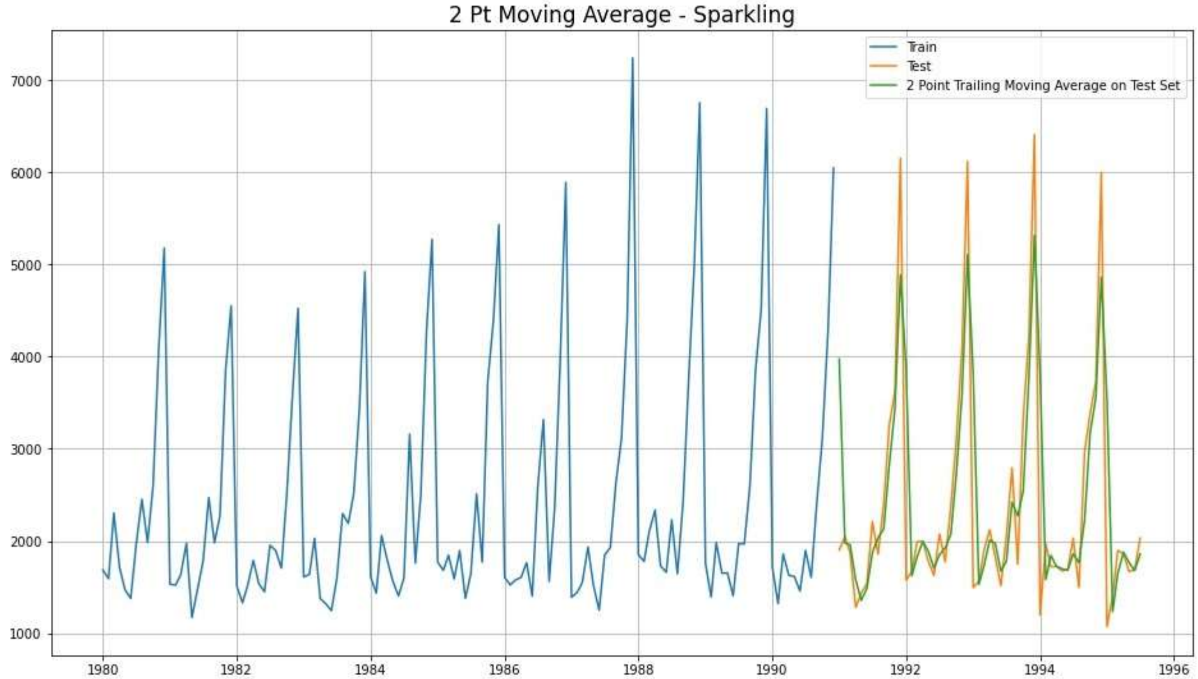
✦ Model 3 - Simple Average

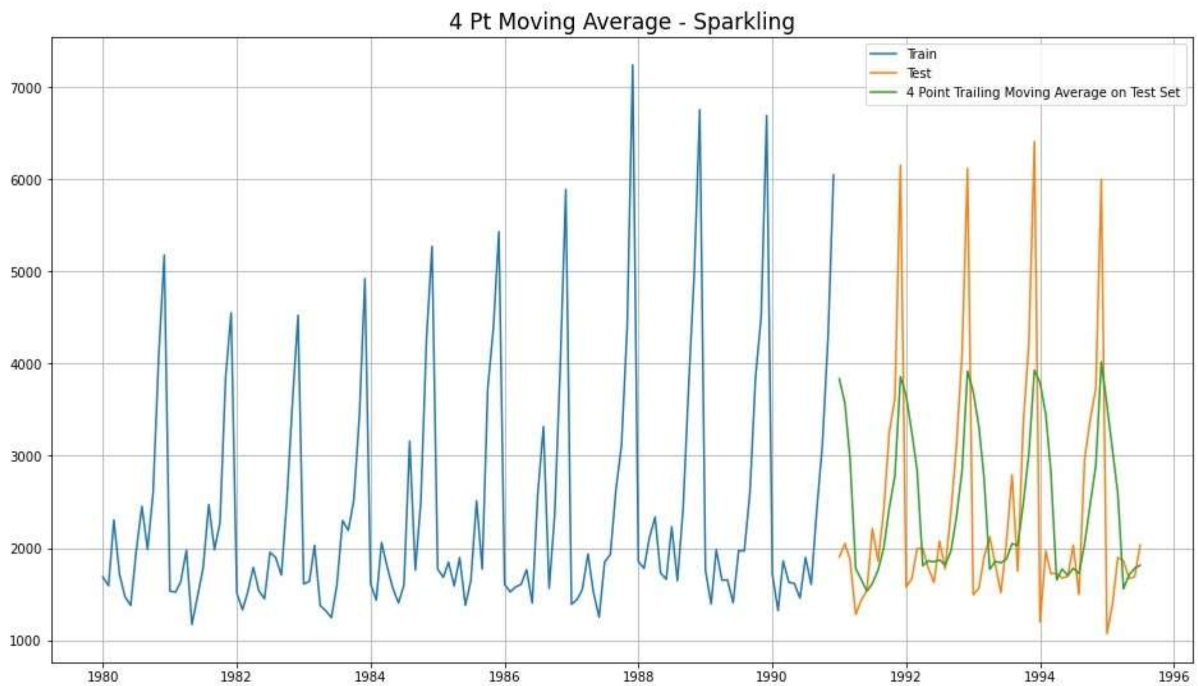
	Test RMSE Sparkling
RegressionOnTime	1389.14
Naive Model	3864.28
SimpleAverageModel	1275.08

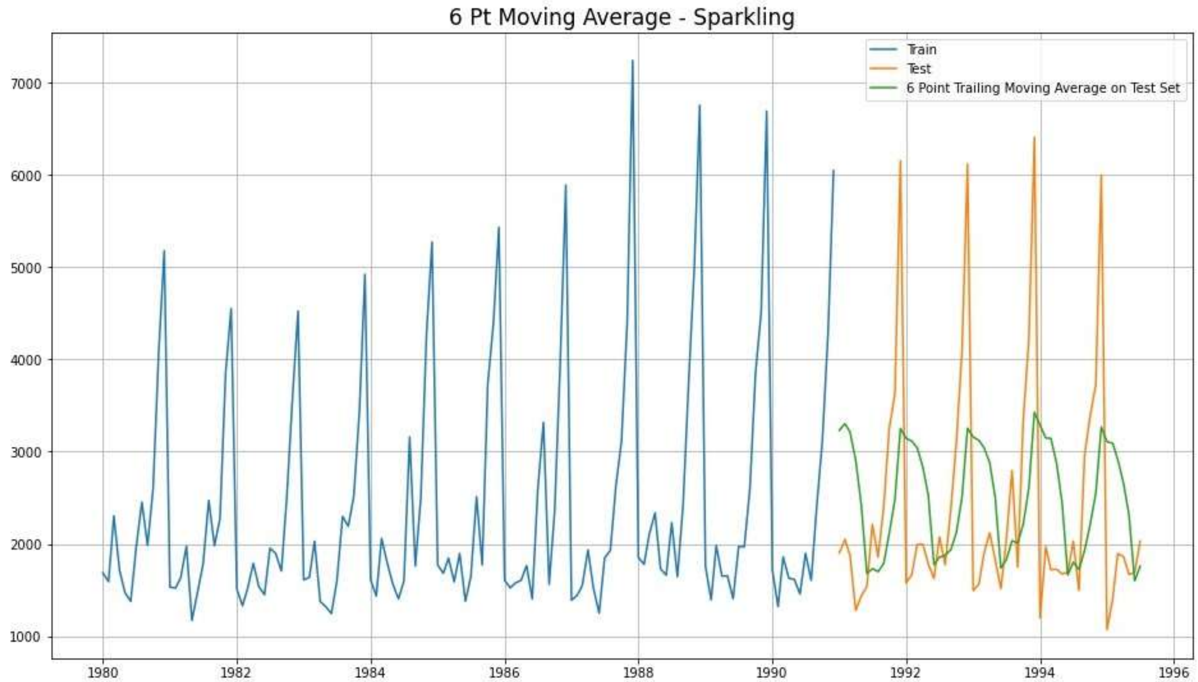
✦ Model 4.A - Moving Average (Sparkling)

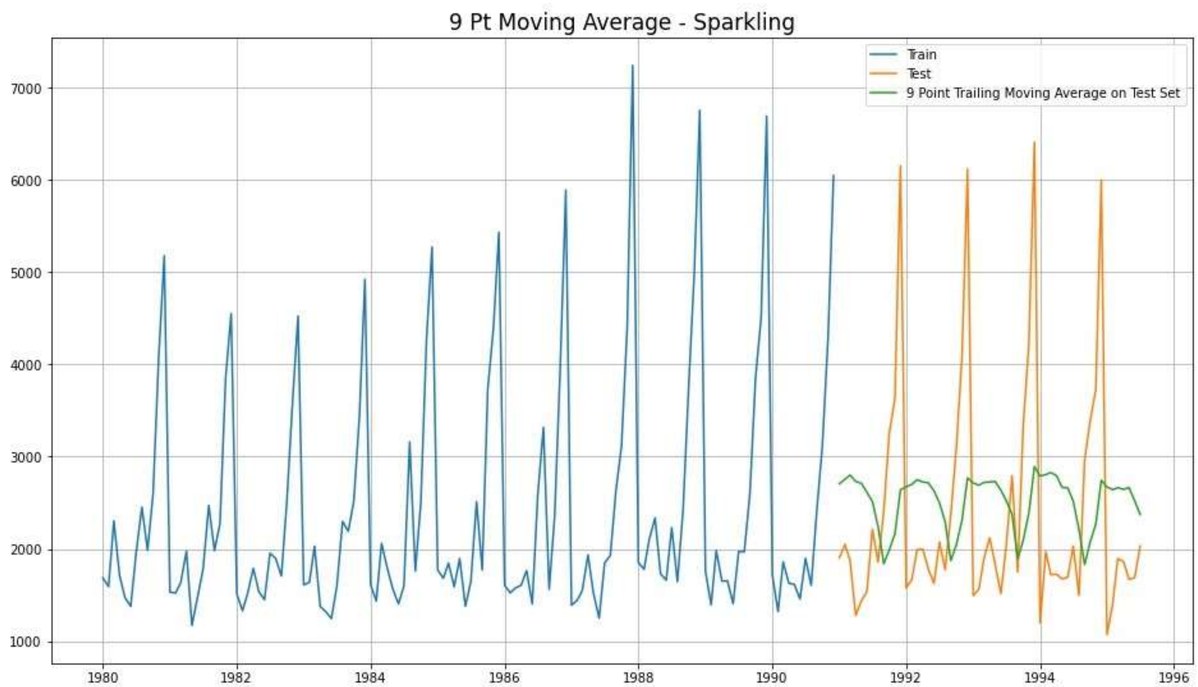


✦ Model 4.B - Moving Average (Sparkling)



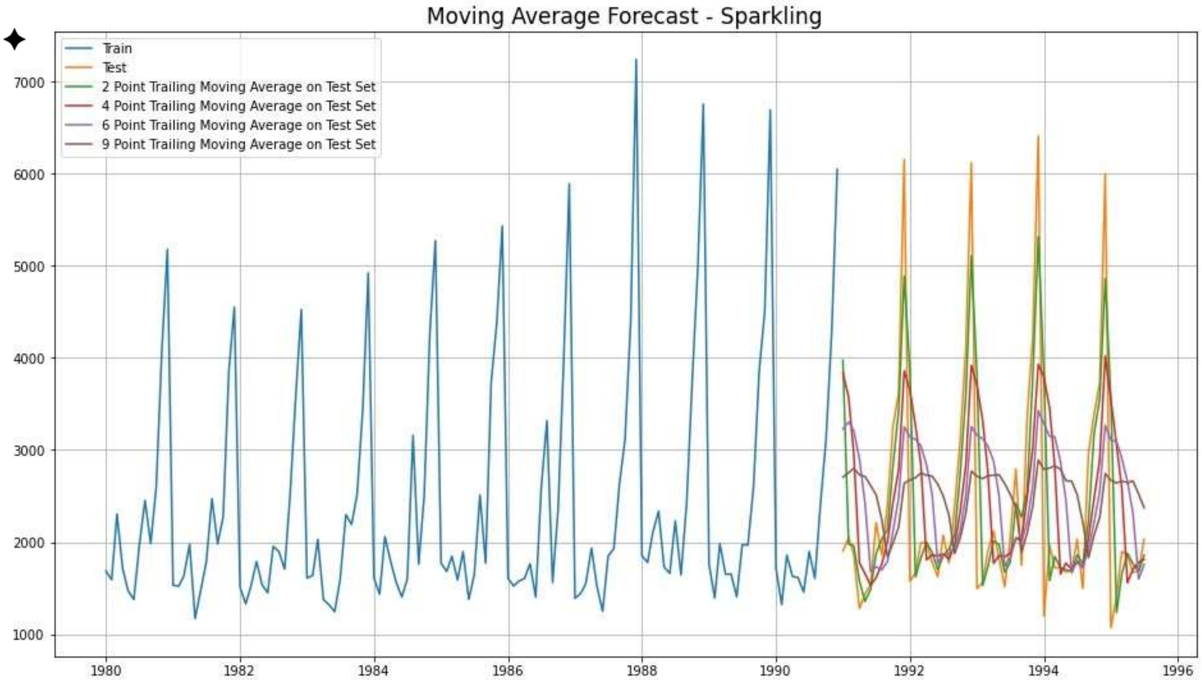






	Test RMSE Sparkling
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28

✦ Consolidated Moving Average Forecasts (Sparkling)



NOTE -

- We have built 4 models till now for Sparkling Wine dataset
- We fitted various models to the Train split and Tested it on Test split. Accuracy metrics used is Root Mean Squared Error (RMSE) on Test data
- Model 1 - Linear Regression ($y_t = \beta_0 + \beta_1 X_t + \epsilon_t$)
  - We regressed variables 'Sparkling' against their individual time instances
  - We modified the datasets and tagged individual sales to their time instances
  - TEST RMSE SPARKLING = 1389.14
- Model 2 - Naive Approach ($\hat{y}_{t+1} = y_t$)
  - Naive approach says that prediction for tomorrow is same as today
  - And, prediction for day-after is same as tomorrow
  - So, effectively all future predictions are going to be same as today
  - TEST RMSE SPARKLING = 3864.28

- <u>Model 3 - Simple Average</u> ($\hat{y}_{t+1} = \hat{y}_{t+2} = \ldots = \hat{y}_{t+n} = Mean(y_1, y_2, \ldots, y_t)$)
    - All future predictions are the same as the simple average of all data till today
    - TEST RMSE SPARKLING = 1275.08
- <u>Model 4 - Moving Average (MA)</u>
    - We calculate rolling means (Moving averages) over different intervals for the whole train data
    - 2 Pt MA ====> means, we find average of 1st and 2nd to predict 3rd
      similarly, average of 2nd and 3rd to predict 4th and so on
    - 4 Pt MA ====> means, we find average of 1st, 2nd, 3rd & 4th to predict 5th
      also, average of 2nd, 3rd, 4th & 5th to predict 6th and so on.

2 PT MA ====>

TEST RMSE SPARKLING = 813.40

- 4 PT MA ====>

  TEST RMSE SPARKLING = 1156.59

- 6 PT MA ====>

  TEST RMSE SPARKLING = 1283.93

- 9 PT MA ====>

  TEST RMSE SPARKLING = 1346.28

	Test RMSE Sparkling
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	**813.40**
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28

Consolidated Scores of Regression, Naive, Simple Average & Moving Average

- **Till now, Best Model which gives lowest RMSE score for Sparkling is ——> 2 Pt Moving Average Model**

- We'll continue to forecast using Exponential Smoothing Models for both datasets of Sparkling Wine Sales

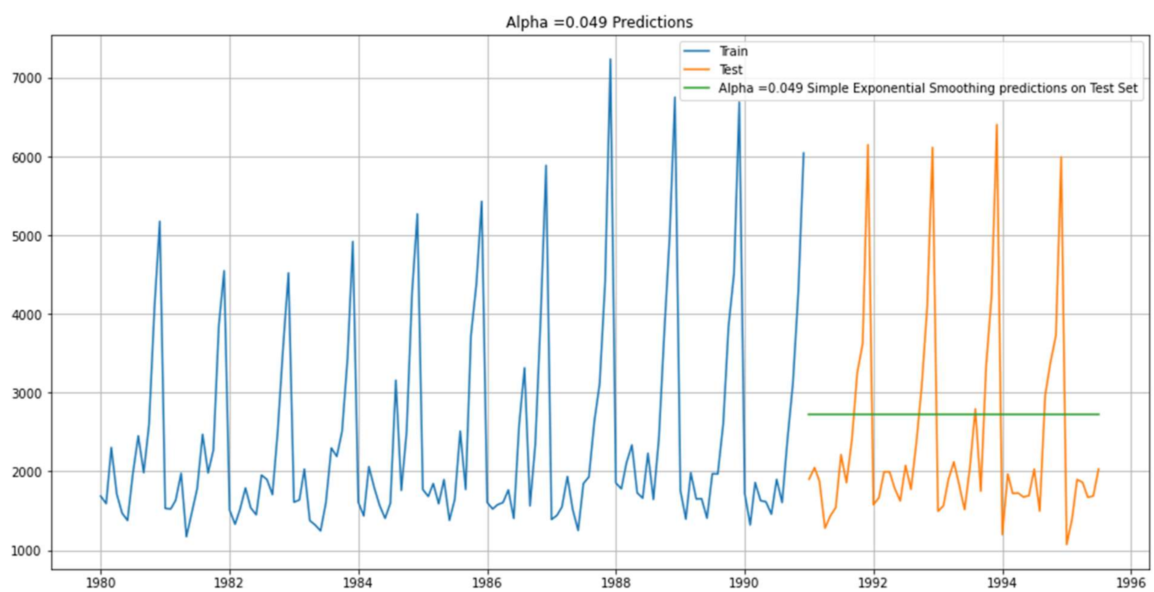
- Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations

- Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be None(N), Additive (N), Additive damped (Ad), Multiplicative (M) or Multiplicative damped (Md)

- One or more parameters control how fast the weights decay. The values ofthe parameters lie between 0 and 1.

- We'll build following Exponential Smoothing Models -

✦ Single Exponential Smoothing with Errors



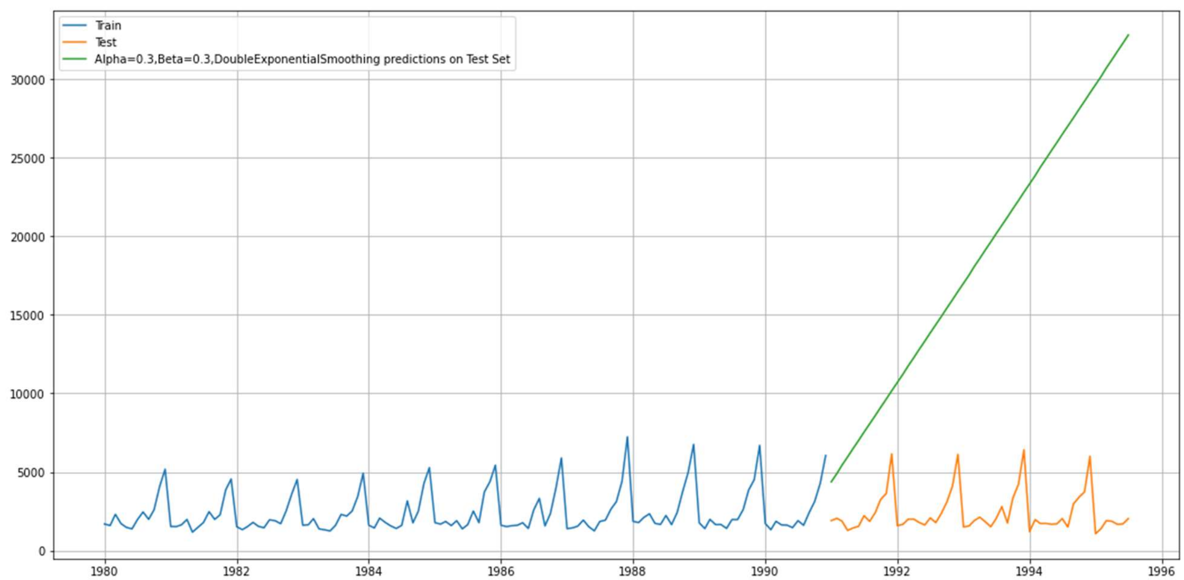
- For Sparkling - Level Parameter, Alpha = 0.049

	Test RMSE Sparkling

RegressionOnTime	1389.14
Naive Model	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	**813.40**
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
Simple Exponential Smoothing	1316.03

- Best Model till now for Sparkling ——- > 2 Pt Moving Average Model.

✦ Double Exponential Smoothing with Additive Errors, Trends -



- In Sparkling - DES shows a non-existent trend. DES does not perform well here.

- Sparkling - Level parameter, Alpha = 0.3

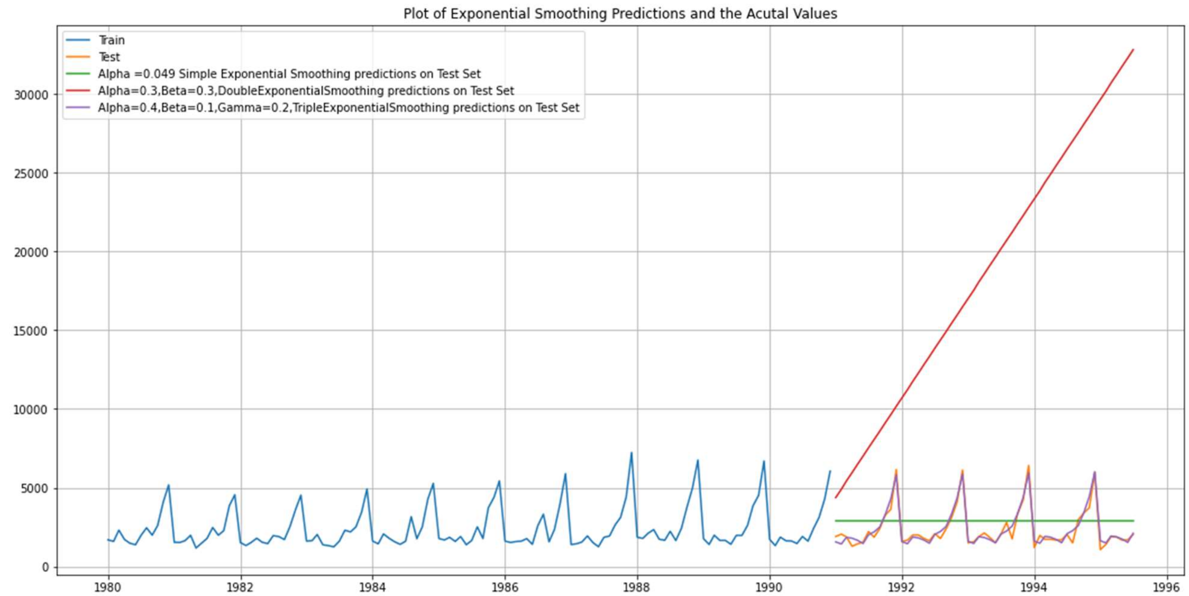
  Trend parameter, Beta = 0.3

	Test RMSE Sparkling
RegressionOnTime	1389.14

NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	**813.40**
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
Simple Exponential Smoothing	1338.00
Double Exponential Smoothing	1375.39

- Best Model till now for Rose and Sparkling ——- > 2 Pt Moving Average Model

✦ Triple Exponential Smoothing with Errors, Seasonality -



- Sparkling - TES has picked up the trend and seasonality very well

- Sparkling - Level parameter, Alpha = 0.4

  Trend parameter, Beta = 0.1

  Seasonality parameter, Gamma = 0.2

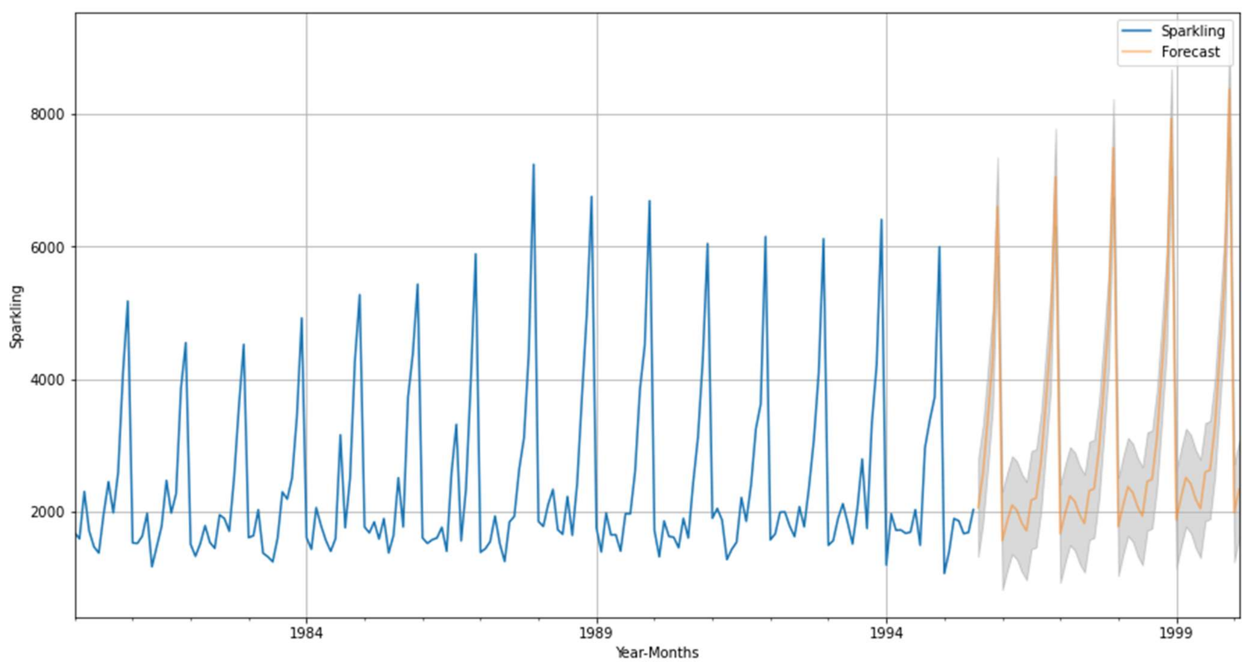
	Test RMSE Sparkling
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08

2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
Simple Exponential Smoothing	1338.00
Double Exponential Smoothing	1375.39
Triple Exponential Smoothing	**317.43**

- Best Model till now for Sparkling ——- > **Triple Exponential Smoothing with**
- Alpha = 0.4
- Trend parameter, Beta =0.1
- Seasonality parameter, Gamma = 0.2

✦ Best Models for Sparkling -

- Best Model till now for Sparkling ——- > **Triple Exponential Smoothing with**
- Alpha = 0.4
- Trend parameter, Beta =0.1
- Seasonality parameter, Gamma = 0.2



[Q 5] Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test.

If the data is found to be non-stationary, take appropriate steps to make it stationary.

Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

✦ <u>To Check Stationarity of Data</u> -

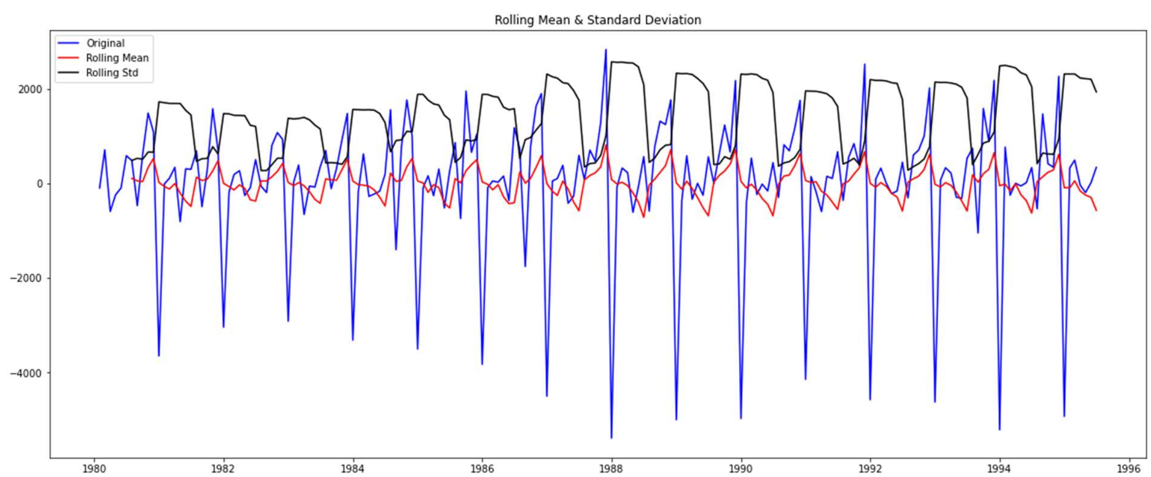
- We use Augmented Dicky - Fuller (ADF) Test to check the Stationarity of Data
- Hypotheses of ADF Test :

  H_0 *Time Series is not Stationary*

  H_a *Time Series is Stationary*
- So for Industry standard (also given for this problem), the Confidence Interval is 95%
- Hence, alpha = 0.05
- So in ADF Test, if p-value < alpha ===> We reject the Null Hypothesis and hence conclude that given Time Series is Stationary
- So in ADF Test, if p-value > alpha ===> We fail to reject the Null Hypothesis and hence conclude that given Time Series is Not Stationary
- If Time Series is not Stationary, then we apply one level of differencing and check for Stationarity again.
- Again, if the Time Series is still not Stationary, we apply one more level of differencing and check for Stationarity again
- Generally, with max 2 levels of differencing, Time Series becomes Stationary
- Once the Time Series is Stationary then we are ready to apply ARIMA / SARIMA models

✦ **Stationarity of Sparkling Wine Dataset** -

- Augmented Dicky-Fuller Test was applied to the whole Sparkling dataset
- We found, p-value = 0.60106
- Here, p-value > alpha=0.05
- We fail to reject the Null Hypothesis and hence conclude that Sparkling Wine Time Series is Not Stationary
- We take 1 level of differencing and check again for Stationarity
- Now, p-value = 0.00
- Now, p-value < alpha=0.05
- Now, we reject the Null Hypothesis and conclude that Sparkling Time Series is Stationary with a lag of 1.



[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 / SARIMA Models** -

- ARIMA is an acronym for Auto-Regressive Integrated Moving Average

- SARIMA stands for Seasonal ARIMA, when the TS has seasonality. ARIMA / SARIMA are forecasting models on Stationary Time Series

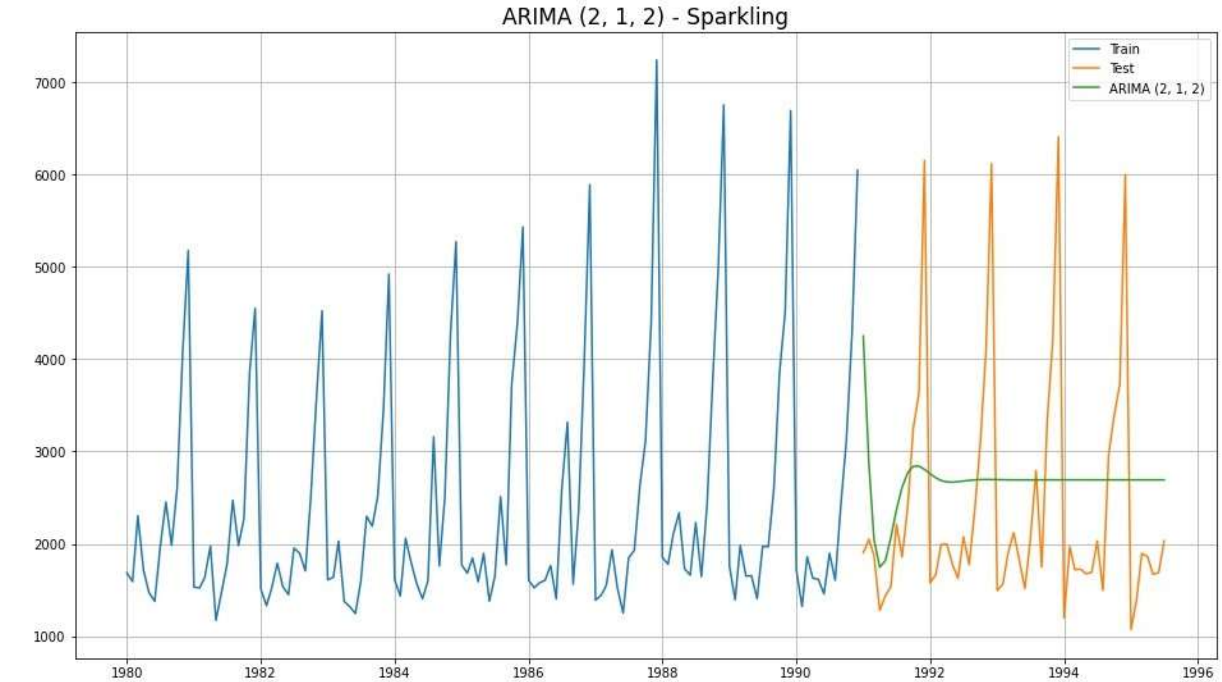
✦ ARIMA / SARIMA Modelling on Train Sparkling Data -

- We check for stationarity of Train Rose & Sparkling data by using Augmented Dicky Fuller Test
- We take a difference of 1 and make both these datasets Stationary
- We apply the following iterations to both these datasets -
    1. ARIMA Automated
    2. SARIMA Automated

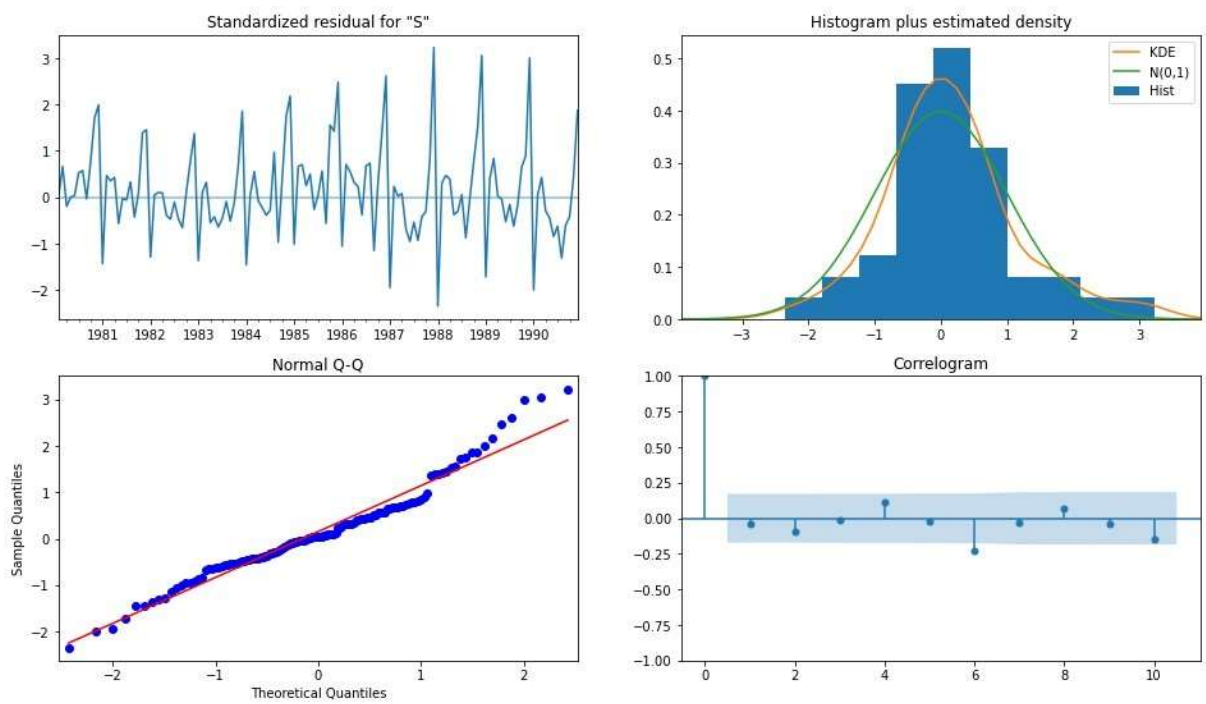
1. ARIMA Automated -

- We create a grid of all possible combinations of (p, d, q)
- Range of p = Range of q = 0 to 2, Constant d = 1
- Some parameter combinations for the Model...
- 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)
- We fit ARIMA models to each of these combinations for both datasets
- We choose the combination with the least Akaike Information Criteria (AIC)
- We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
- Finally, we check the accuracy of this model by checking RMSE of Test set

- For **Sparkling**, Best Combination with **Least AIC** is - **(p, d, q) —-> (2, 1, 2)**



ARIMA (2, 1, 2) Diagnostic Plot - Sparkling



	Test RMSE Sparkling
ARIMA(2,1,2)	1299.97

2. SARIMA Automated -

- We create a grid of all possible combinations of (p, d, q) along with Seasonal (P, D, Q) & Seasonality of 12

- Range of p = Range of q = 0 to 3, Constant d = 1

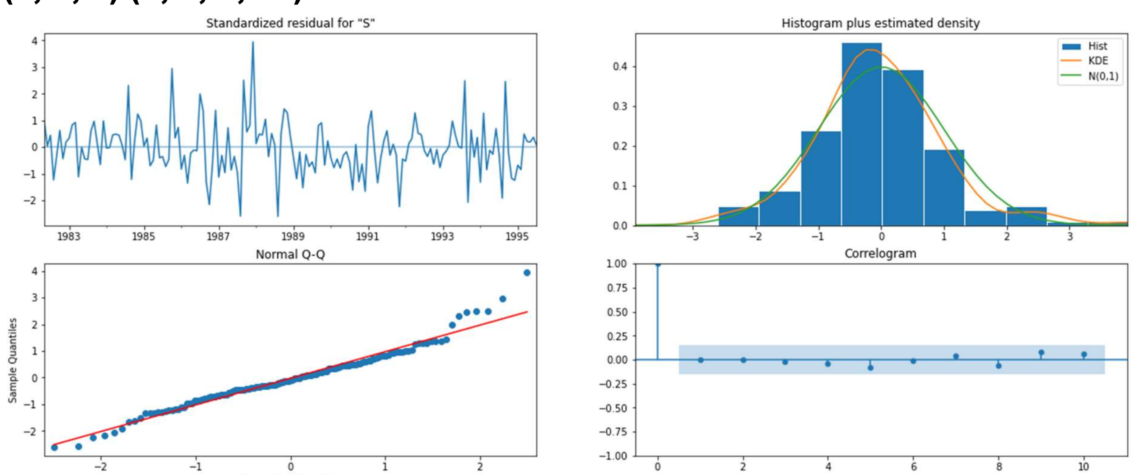
- Range of Seasonal P = Range of Seasonal Q = 0 to 3, Constant D = 1, Seasonality m = 12

- Few Examples of the grid (p, d, q) (P, D, Q, m) -

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

- We fit SARIMA models to each of these combinations and select with least AIC
- We fit SARIMA to this best combination of (p, d, q) (P, D, Q, m) to the Train set and forecast on the Test set. Then, we check accuracy using RMSE on Test set

- For **Sparkling**, Best Combination with **low AIC** and low Test RMSE is - **(1, 1, 2) (1, 0, 2, 12)**



	Test RMSE Sparkling
ARIMA(2,1,2)	1299.98
SARIMA (1, 1, 2) (1, 0, 2, 12)	528.59

- **Till Now, Best Model for Sparkling with Least RMSE ->SARIMA (1, 1, 2) (1, 0, 2, 12)**

[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.

✦ Auto-Correlation Function (ACF) -

- Autocorrelation refers to how correlated a time series is with its past values. e.g. y_t with y_{t-1} also y_{t+1} with y_t and so on.
- 'Auto' part of Autocorrelation refers to Correlation of any time instance with its previous time instance in the SAME Time Series
- ACF is the plot used to see the correlation between the points, up to and including the lag unit
- ACF indicates the value of 'q' - which is the Moving Average parameter in ARIMA / SARIMA models

✦ Partial Auto-Correlation Function (PACF) -

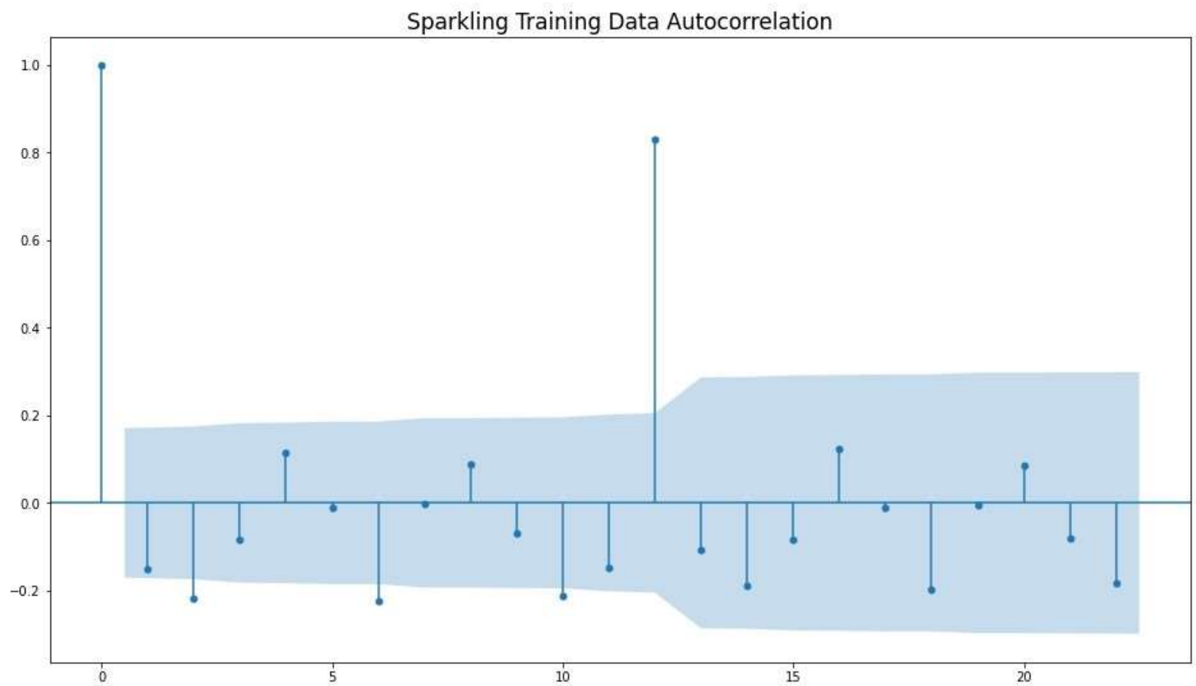
- Partial Autocorrelation refers to how correlated a time series is with its past lag values.
- For example, let lag=k, then Partial Autocorrelation is Correlation of y_t with y_{t-k}, ignoring the effects of all the instances between y_t and y_{t-k}
- PACF is the plot used to see the correlation between the lag points
- PACF indicates the value of 'p' - which is the Auto-Regressive parameter in ARIMA / SARIMA models

✦ ACF & PACF of Sparkling -

- Observing the cutoffs in ACF and PACF plots for Sparkling dataset, we get -

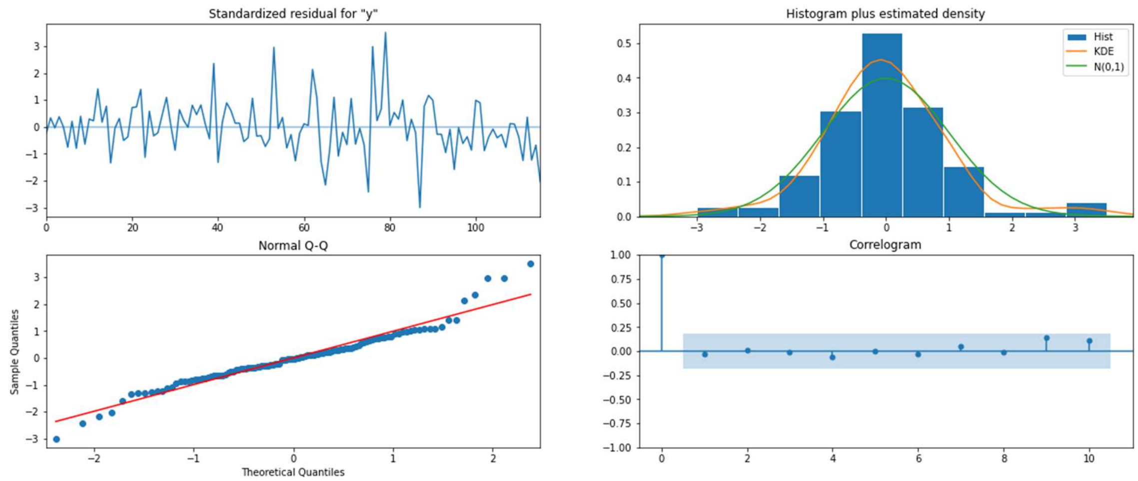
    FOR ARIMA —-> p = 0, q = 0 and difference d = 1

    FOR SARIMA —-> p = 0, q = 0, d = 1 and P = 0, 1, 2, 3 | D = 0, Q = 1, 2, 3



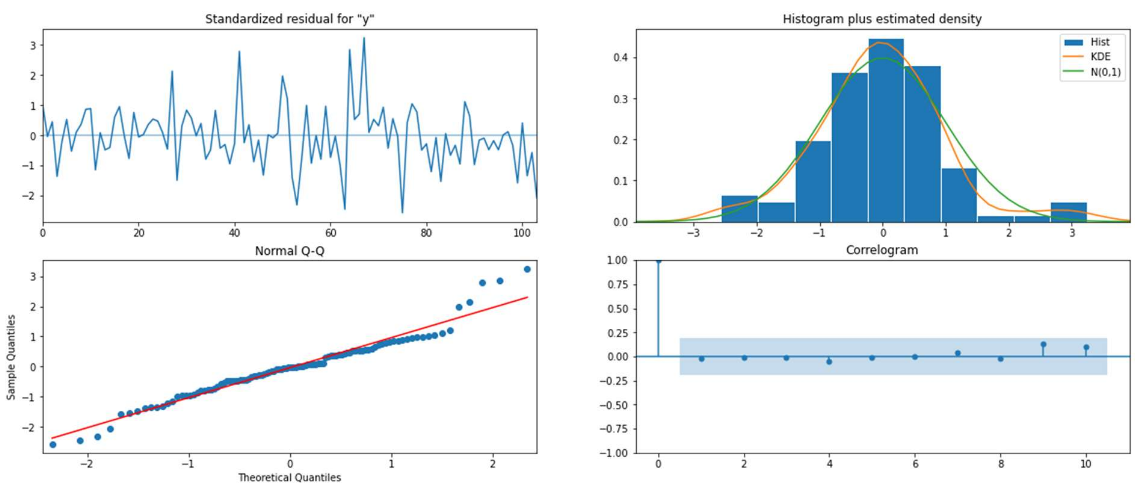
SARIMA (0, 1, 0) (1, 1, 1, 12 Diagnostic Plot - SPARKLING

4. **SARIMA - Sparkling - (1, 1, 2) (2, 0, 2, 6)**



	Test RMSE Sparkling
ARIMA(3, 1, 2)	1281.48
SARIMA (1, 1, 2) (2, 0, 2, 6)	626.89

5. **SARIMA Manual - Sparkling - (1, 1, 2) (1, 0, 2, 12)**



SARIMA (1, 1, 2) (1, 0, 2, 12 Diagnostic Plot - SPARKLING

	Test RMSE Sparkling
ARIMA(2,1,2)	1299.979
ARIMA(3,1,2)	1281.482
SARIMA(1,1,2)(2,0,2,6)	626.8982
SARIMA(1,1,2)(1,0,2,12)	528.5944

- **Best Model for Sparkling with Least RMSE**

  —-> SARIMA (1, 1, 2) (1, 0, 2, 12)

- Seasonal P and Q - it was difficult to gauge the correct values here as the data was not enough and cutoffs were not visible

- Hence, we tried multiple combinations of Seasonal P and Q as given above

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

✦ All Models built with SPARKLING (sorted by RMSE) -

Models With Parameters	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.049,SimpleExponentialSmoothing	1316.034674
Alpha=0.3,Beta =0.3 DoubleExponentialSmoothing	1375.393398
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	402.946854
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
ARIMA(2,1,2)	1299.979832
ARIMA(3,1,2)	1281.482077
SARIMA(1,1,2)(2,0,2,6)	626.898233
SARIMA(1,1,2)(1,0,2,12)	528.594459

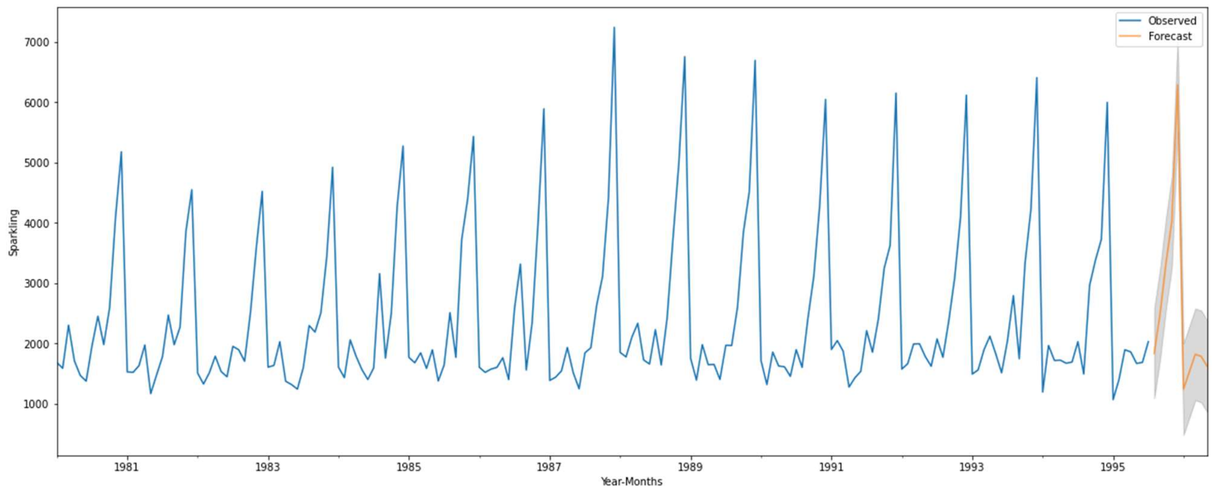
[Q 9] Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

- **Best Model as per the Least RMSE on SPARKLING Test set ——>**

Triple Exponential Smoothing

Alpha 0.4, β = 0.1, γ = 0.2

✦ **<u>Sparkling Forecast Next 10 months - Triple Exponential Smoothing Trend, Seasonality</u>**



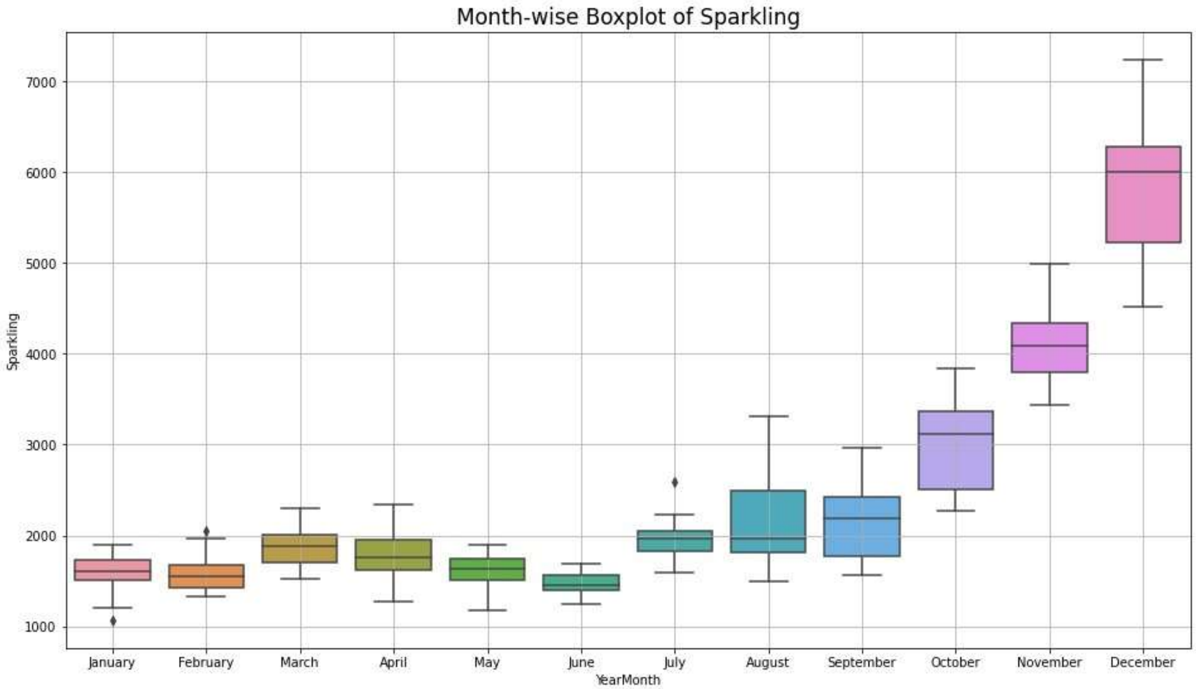
[Q 10] Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

✦ **<u>Sparkling Wine Sales - Comments :</u>**

- Sparkling wine sales don't show any upward or downward trend
  - This shows flat sales over long term range
- Also, there is very high spike in sales seen in the last quarter of every year from Oct to Dec
  - This might be due to the Holiday season in this period

- Highest peak in sales is seen in Dec every year

- Dec sales are almost 3 times of Sep sales



- Sparkling sales, an instant crashing slump is seen in the first quarter of every year from Jan

  - This might be due to the after effect or hangover of Holidays

- Sales slowly pick up only from Jul-Aug

✦ Sparkling Wine Sales - Forecast Models :

- Triple Exponential Smoothing - Holt-Winters Models perform the best on Sparkling datasets, considering the least RMSE on Test data

- There has been incremental improvements in Test RMSE with each tuning of parameters

- Finally, **for forecast of Sparkling Wine Sales - we choose Holt-Winters with Seasonality and Trend**

✦ Sparkling Wine Sales - Suggestions :

- Even for Sparkling, Holiday season is around the corner and forecast shows increasing sales and sharp peak in Dec. Hence, Company should stock up

- Sparkling wine has great holiday sales, so this shows popularity.

- So no need to introduce any offers here but hammering Ads are suggested in these times of Oct-Dec. This will drive sales even further.

- Sparkling wines are generally associated with celebrations and mainly to burst open.

- A special designer bottle can be introduced at a cheaper price just for bursting. This will maximize profits

- Year on Year sales do not show any significant increase or decrease

- Though, Holiday spikes are extreme, but general Year on Year sales need to be investigated more. Early period from Jan should be used to do this deep dive

—————— END OF PROJECT ——————