Time Series Forecasting – Sparkling Wines

**Sharmila Sengupta Chowdhry**

**Table of Contents**

Contents

[Q1. Read the data as an appropriate Time Series data and plot the data. 3](#_Toc73912181)

[Q2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. 4](#_Toc73912182)

[Q3. Split the data into training and test. The test data should start in 1991. 6](#_Toc73912183)

[Q4. 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. 7](#_Toc73912184)

[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. 13](#_Toc73912185)

[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. 15](#_Toc73912186)

[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. 18](#_Toc73912187)

[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. 21](#_Toc73912188)

[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. 22](#_Toc73912189)

[Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. 23](#_Toc73912190)

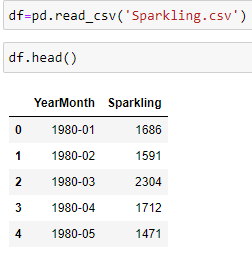
PROBLEM 1: TIME SERIES FORECASTING – SPARKLING DATA

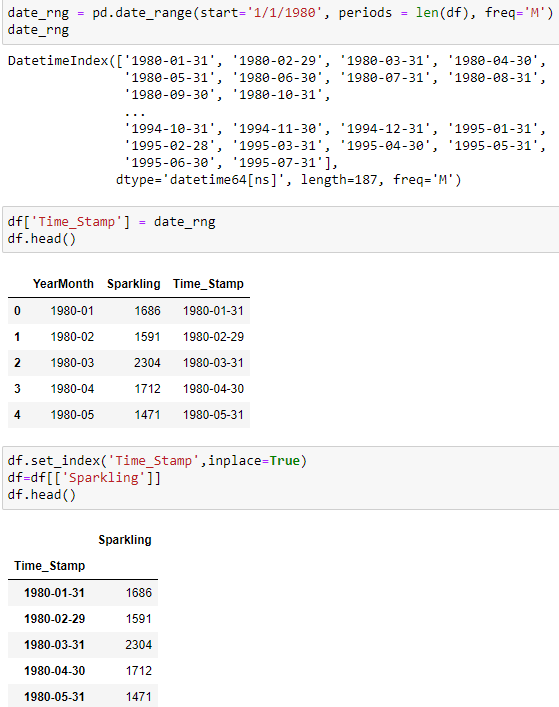
## Q1. Read the data as an appropriate Time Series data and plot the data.

Read the data, checking the header and tail data to validate that the data is uploaded correctly. We have defining time stamps as this is monthly time series, plot the time series plot.

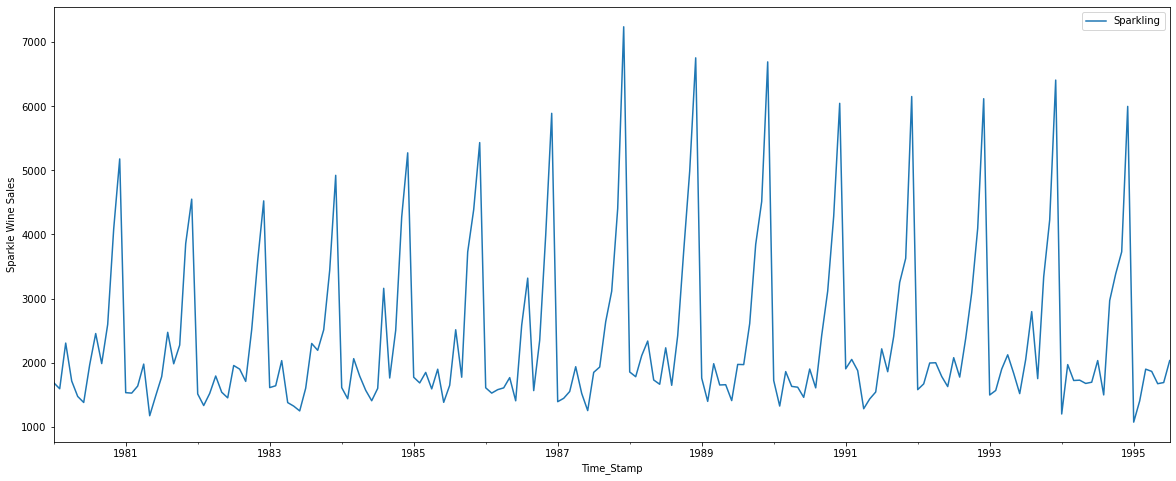
Please refer to below the primary observations on the Sparkle Wine data:

* Read the data file of Sparkling Wine data and setting the YearMonth as index in the column





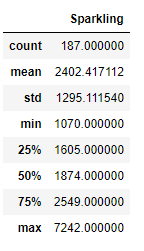
* **Observation:** Checked for initial and last data. For the year 1995 we only have data till July and hence we will keep this in mind while checking for trend in the data



* **Observation:** Plotted the time series data, this Time Series seems to have Seasonality and but no Trend.

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

* Data has 187 records,

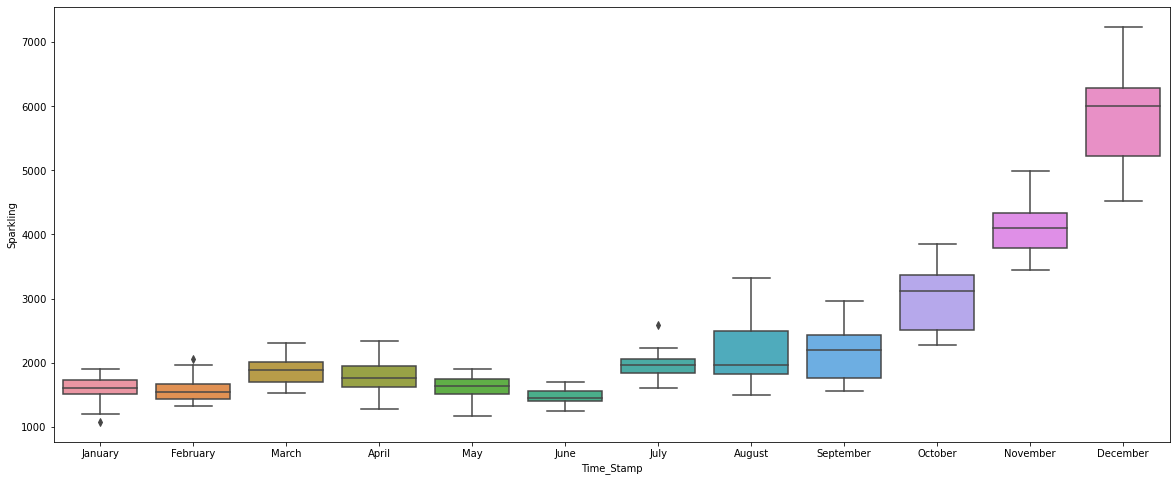


* Checking for shape indicate that there are 187 indexes/rows and 1 column

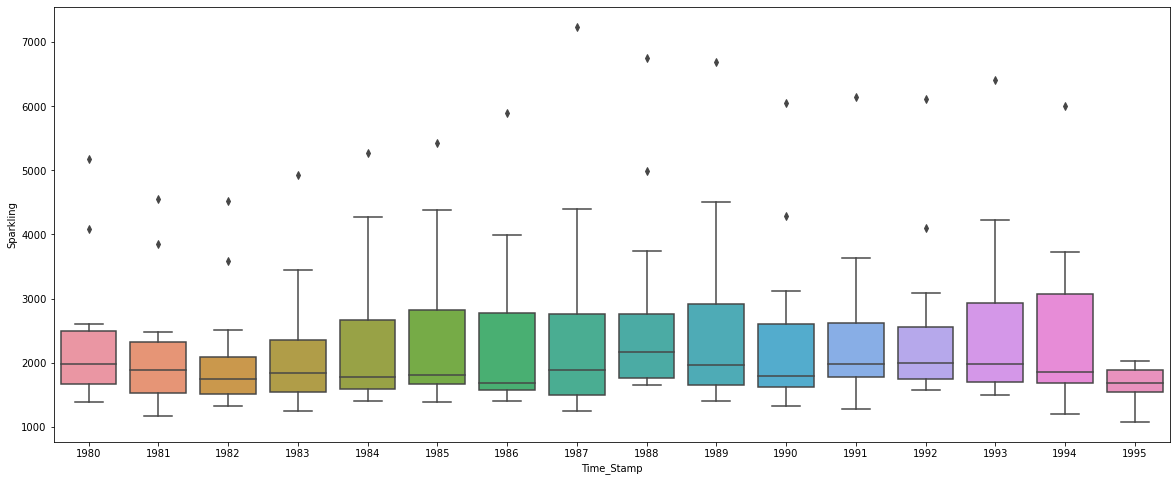


* Boxplot creation

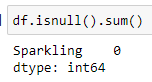
**Monthly Boxplot**

****

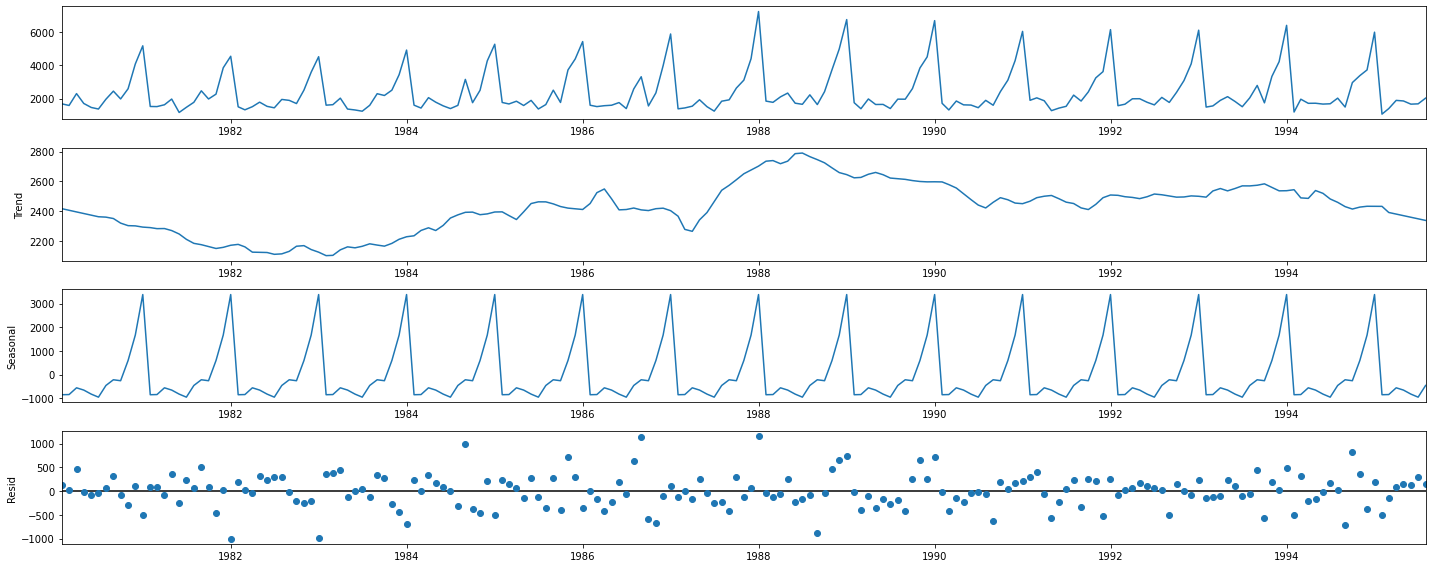
**Yearly Boxplot**

****

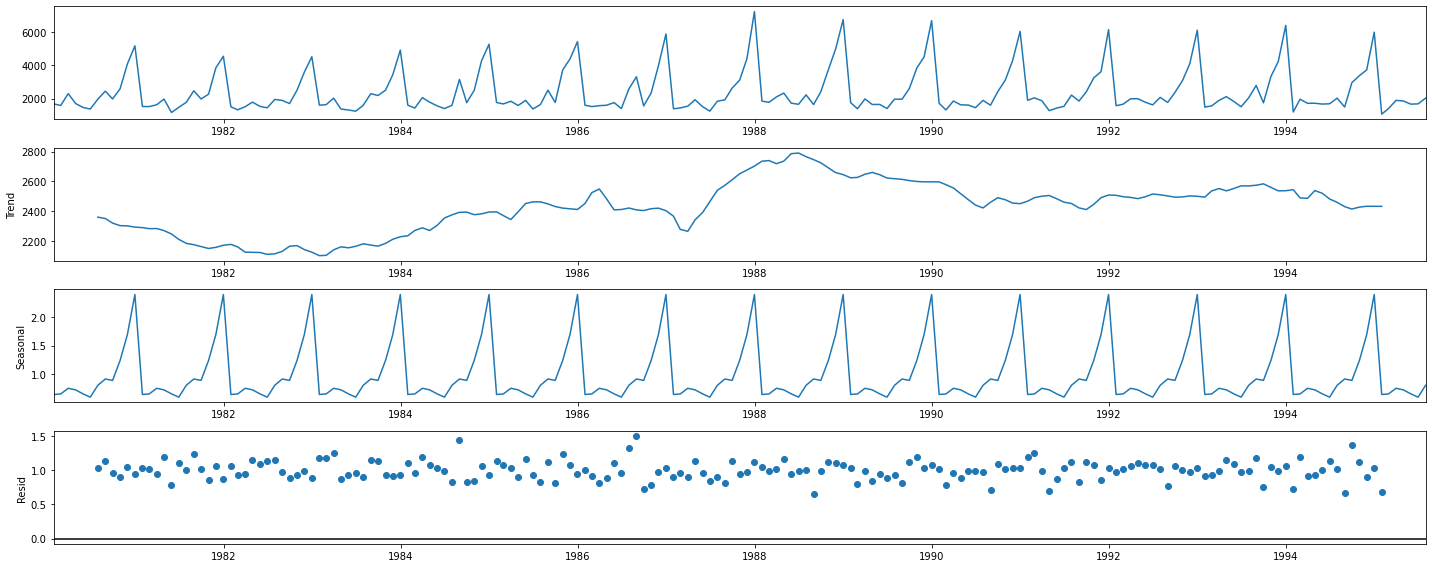
* **Observation:** Creating Boxplot to understand monthly sales contribution across all years combined. We see that the **sales tends is constant across years**. While from the monthly plot we can infer that the sales are increasing from September to December with sudden drop in January.
* Checking for null values



* **Observation:** Sparkling Wine data has no null values
* Performed Additive Decomposition on time series to understand about Trend, Seasonality and Trend



* Performed Multiplicative Decomposition on time series to understand about Trend, Seasonality and Trend



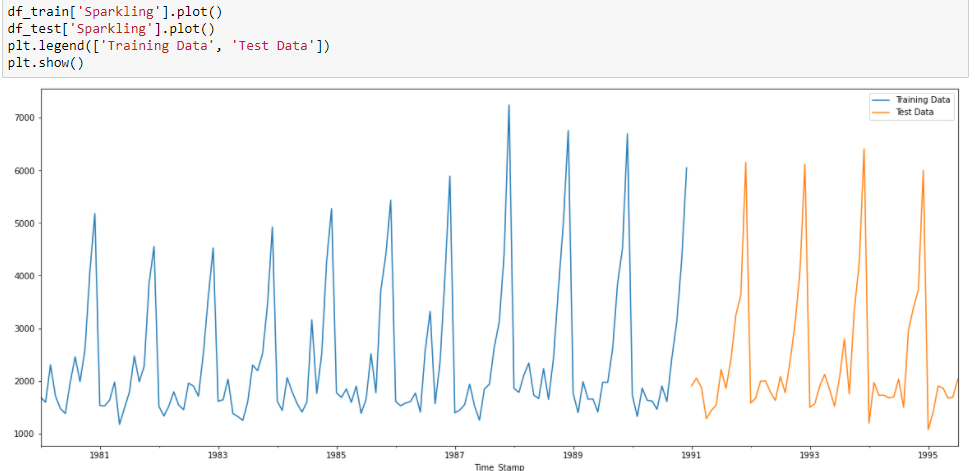
* **Observation:** A time series decomposition gives trend, seasonality and error/residual values. From the charts we can see that the Multiplicative shows residual values nearing to 1, indicating high error values while for the additive decomposition the residual values are near to 0. Hence the series is to be considered as **Additive Series.**

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

* Split the data as follows; train data has 132 records & test data has 55 records
* Data < 1991 : Train Data
* Data > 1991 : Test data



* Plotted the Training & Test data



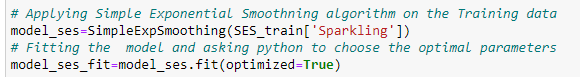
## Q4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

**Model 1: Simple Exponential Smoothening Model**

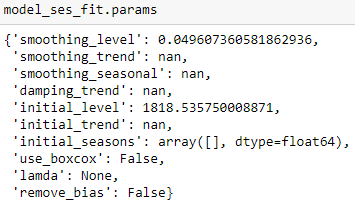
* Loading the relevant libraries



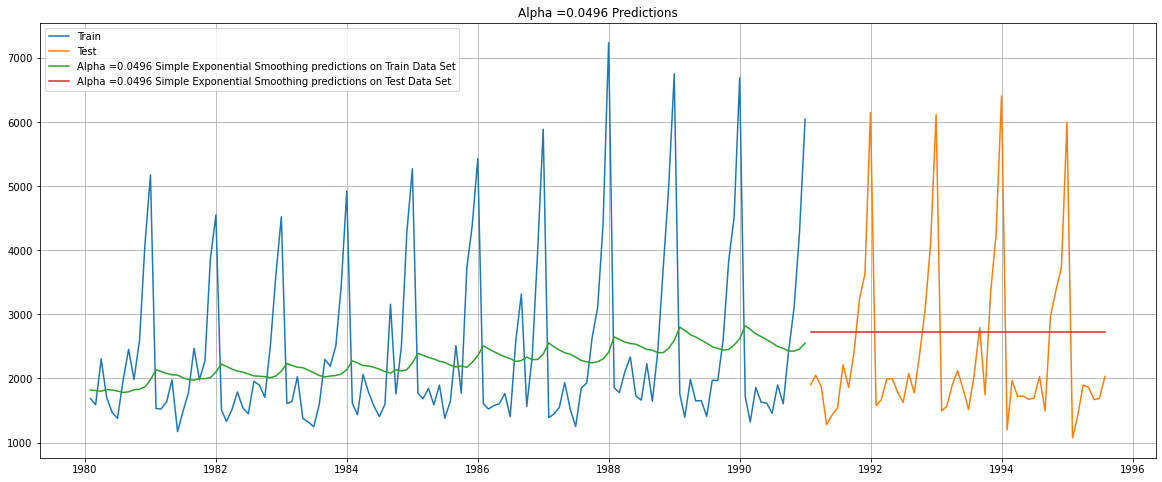
* Model creation



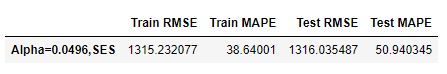
* Parameter Extraction



* **Observation:** The smoothing level is close to 0, hence series is using past values for smoothening
* Plotting the training and test data along with respective prediction values at Alpha =0.049



* **Observation:** Here, we see that this Simple Exponential Smoothing (in which the parameters are automatically estimated by Python) is working like a Simple Average model. This is happening because the value 𝛼 is close to 0. Lesser the value of 𝛼 , more weight is given to older value
* Model Evaluation

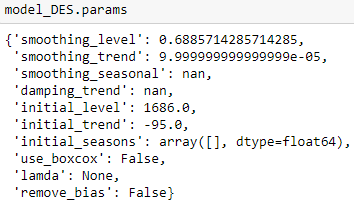


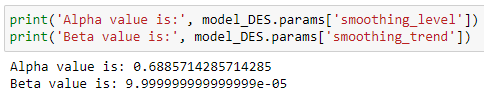
**Model 2: Double Exponential Smoothening model - Holt**

* Model creation

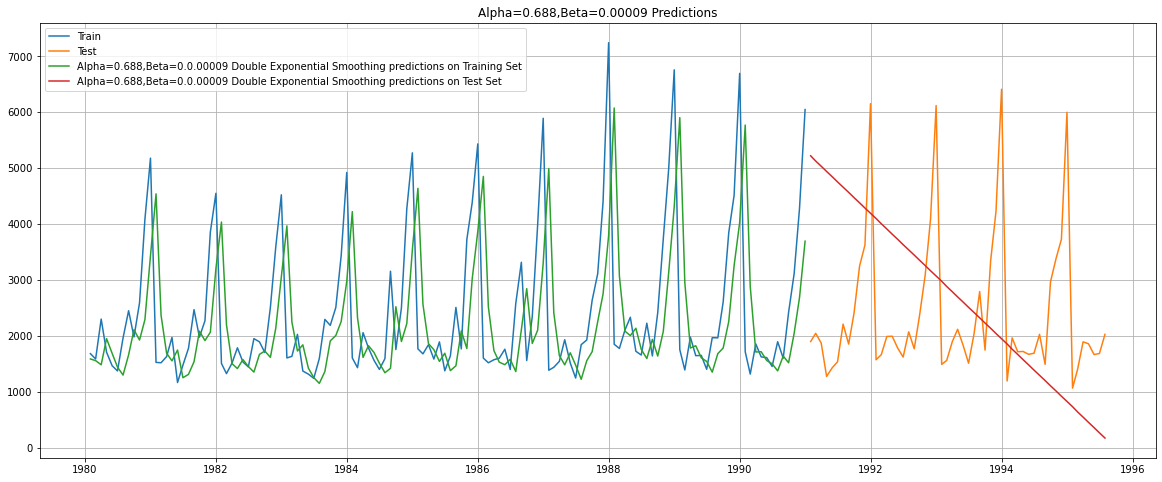


* Parameter Extraction

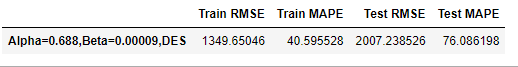




* **Observation:** - Here, we see that Python has optimized the Level co-efficient (alpha value) to 0.688 but the Beta Value is extremely low. Indicative of no presence of trend in the series
* Plotting the training and test data along with respective prediction values at Alpha=0.688, Beta=0.00009



* Model Evaluation



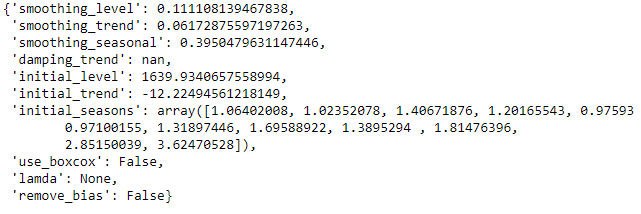
**Observation:** - Double Exponential Smoothening is not performing better as RMSE has increased as compare to the Simple Exponential Smoothening models

**Model 3: Triple Exponential Smoothening Model – Holt Winter**

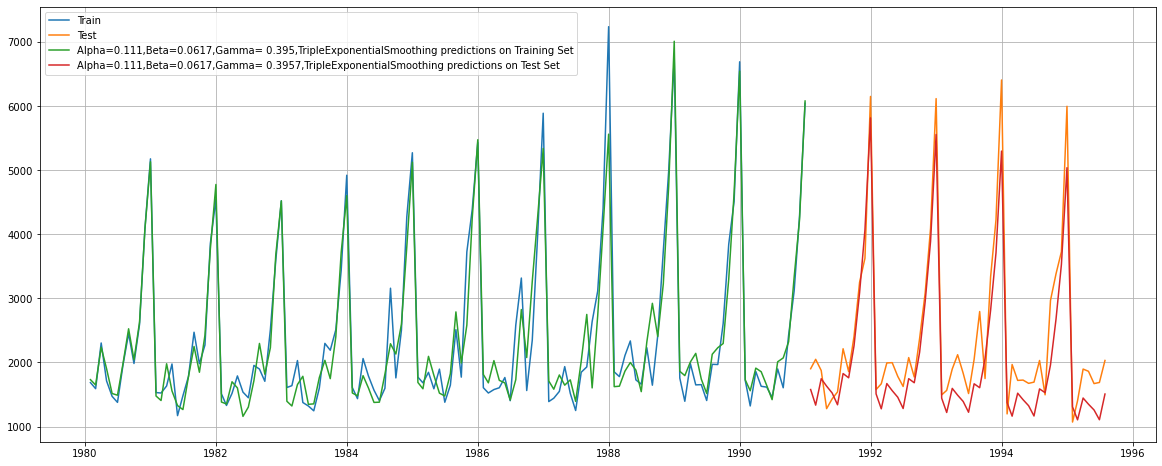
* Model creation



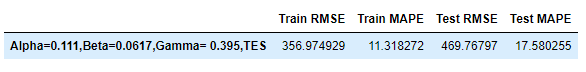
* Parameter Extraction



* **Observation:** Triple Exponential Smoothening model is giving the seasonal value
* Plotting the training and test data along with respective prediction values at Alpha=0.111,Beta=0.0617,Gamma= 0.395



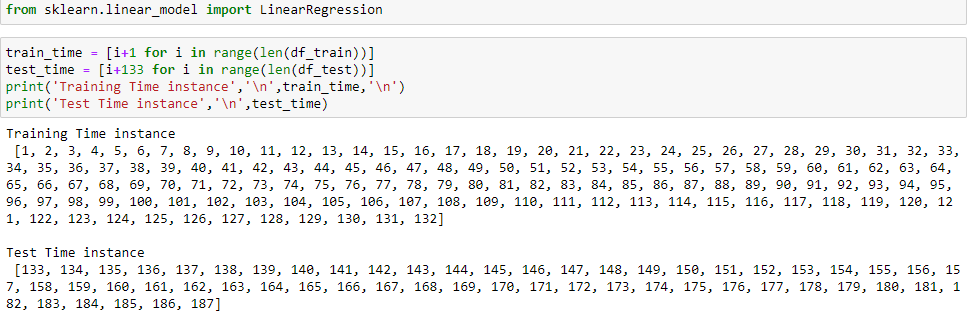
* Model Evaluation

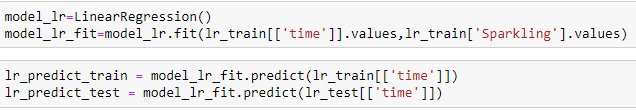


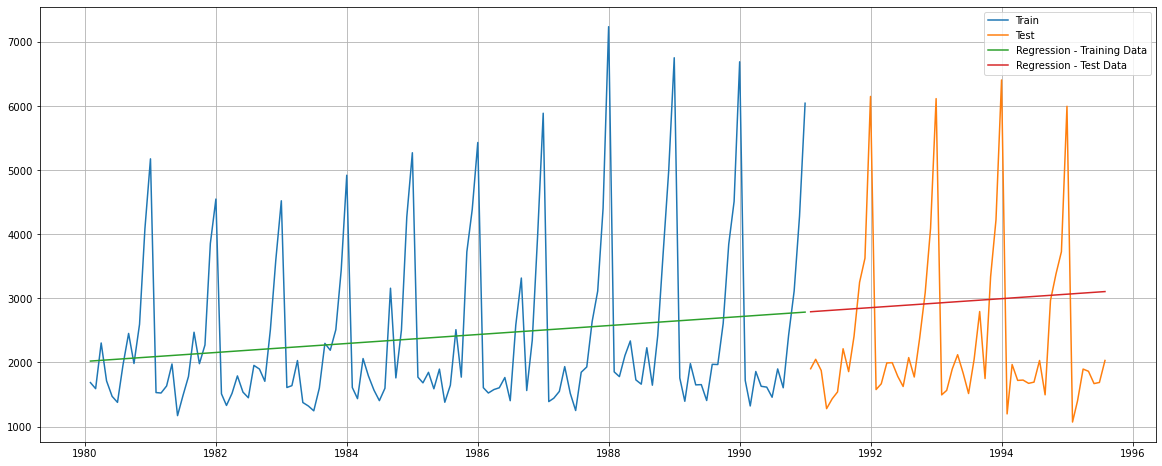
* **Observation:** We have seen all the Smoothening models and on the basis of RMSE and MAPE. The best model till now appears to be Triple Exponential Smoothing as it is able to correctly estimate the trend and seasonality data. Also the difference of Train and Test RMSE values are very less.

**Model 4: Linear Regression**

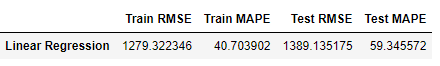
* Model creation







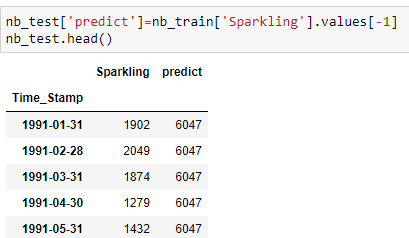
* Model Evaluation

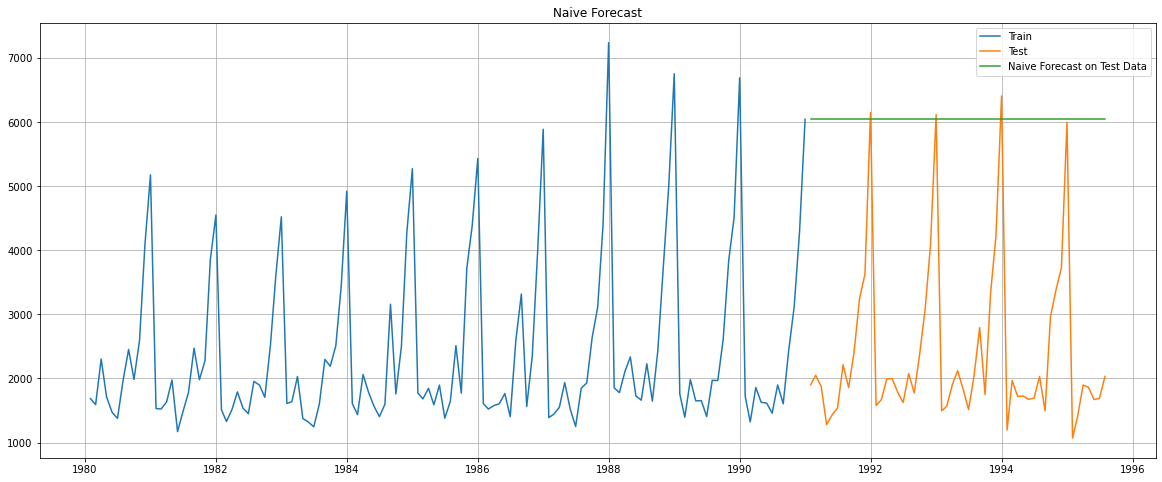


**Model 5: Naïve Bayes**

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

* Model creation



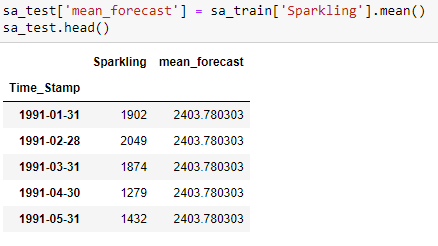


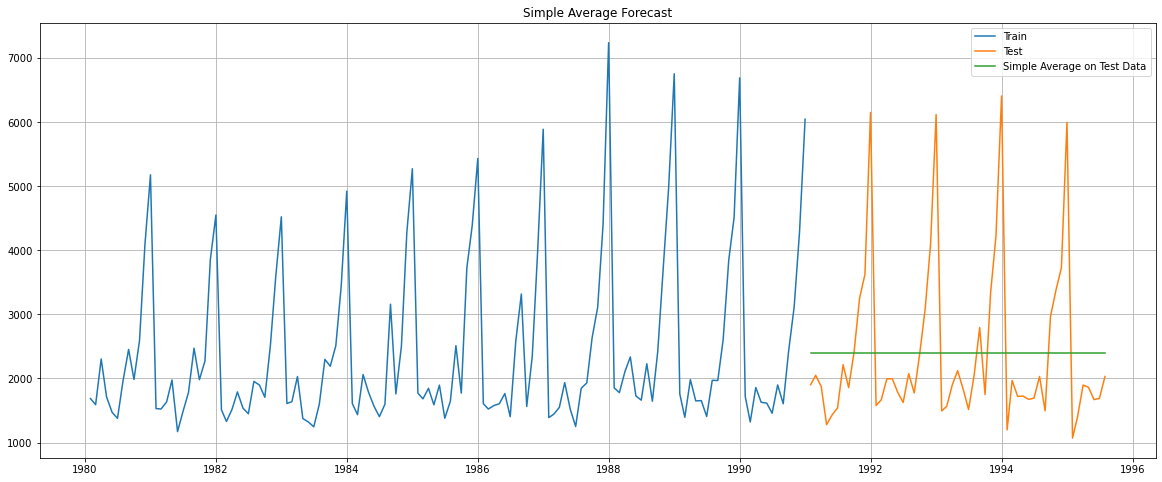
* Model Evaluation



**Model 6: Simple Average Method**

* Model Creation



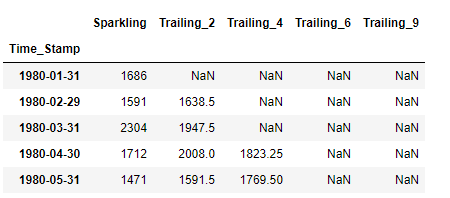
****

* Model Evaluation

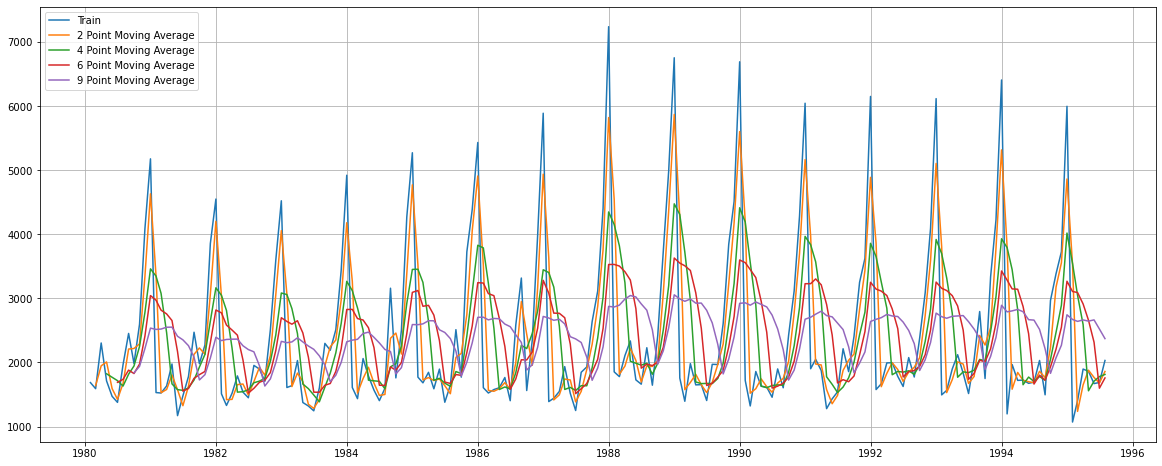


**Method 7: Moving Average Method**

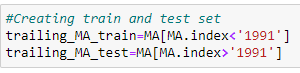
* In moving average model, calculated the rolling means of few intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.
* Generated the predictions on Training and Test data as follows



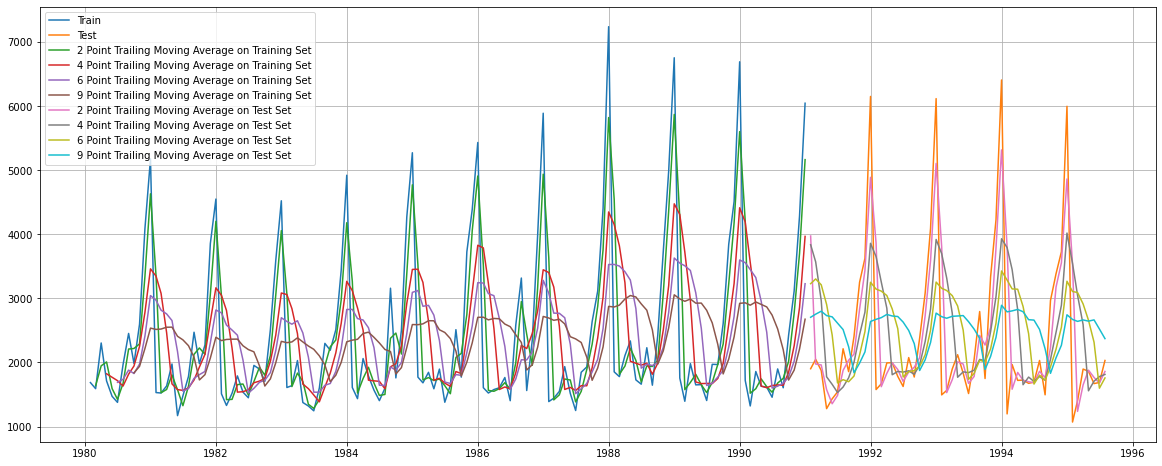
* Plotting the whole data set



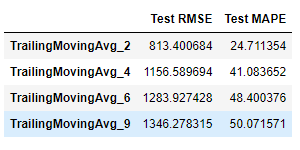
* Splitting into the Train and Test



* Plotting the Train and Test



* Model Evaluation



* **Observation:** By noticing the Plot and the RMSE on test data, it can be concluded that the 2 point trailing Moving average gives more accurate model performance as compare to the other trailing models

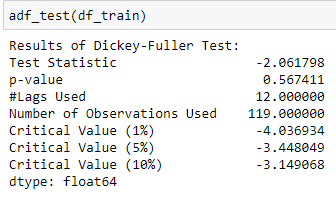
## Q5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

**Stationarity of the Data series;**

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

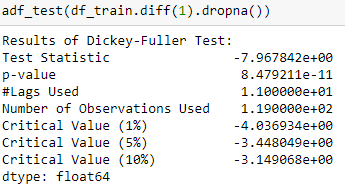
The hypothesis in a simple form for the ADF test is:

* + H0 : The Time Series has a unit root and is thus non-stationary.
  + H1 : The Time Series does not have a unit root and is thus stationary.
* We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the alpha value.
* Apllied ADF Test, we see that at 5% significant level,the Time Series is non-stationary.



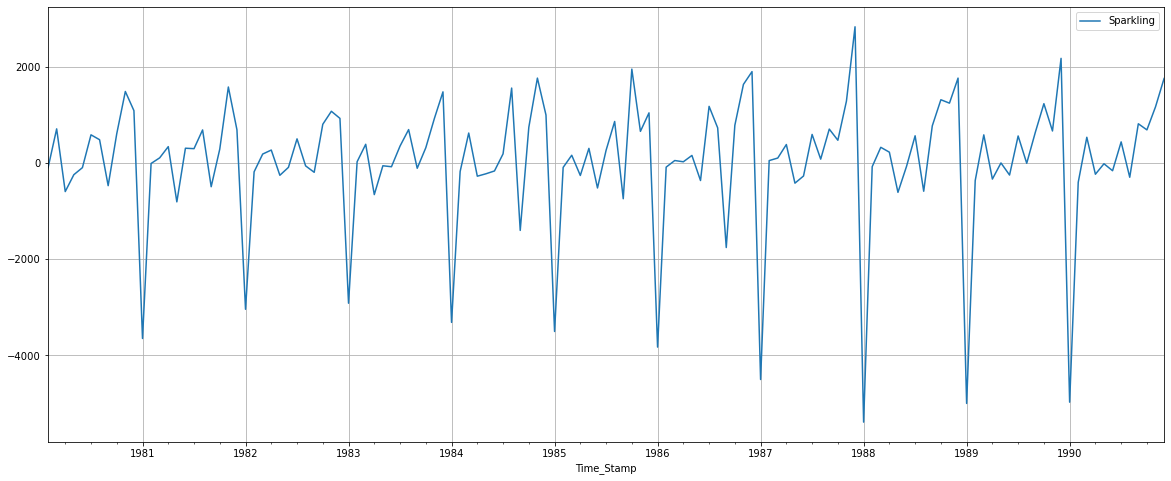
**Observation:** P\_value >0.05, so the series is not stationary. Therefore we should make it stationary by taking the difference of the series

* Let us take the difference of the series and then check for stationarity. We see that the series has become stationary after a taking a difference of first order.



**Observation:** P\_value <0.05, so the series becomes stationary.

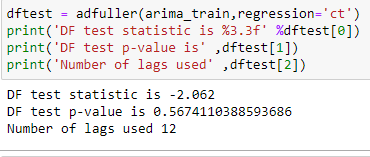
* Let us now check how the differenced series looks. We see that the trend has been arrested but there is some seasonality.



## 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.

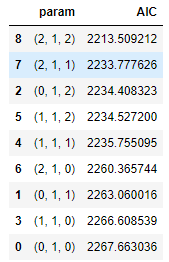
**Model 8: Auto ARIMA Model**

* Checking the stationarity of the train data



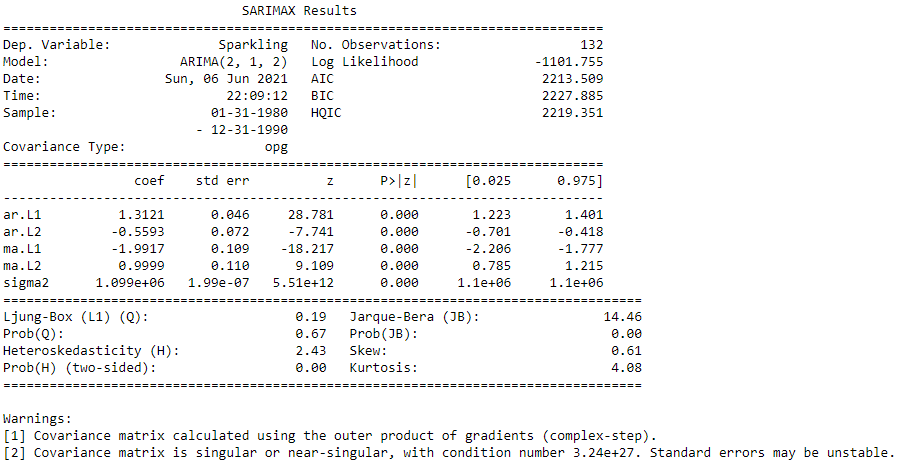
**Observation:** As the P\_value >0.05, we need to make the series stationary by taking the first order difference.

* Looking at minimum AIC by providing combination of p,d,q

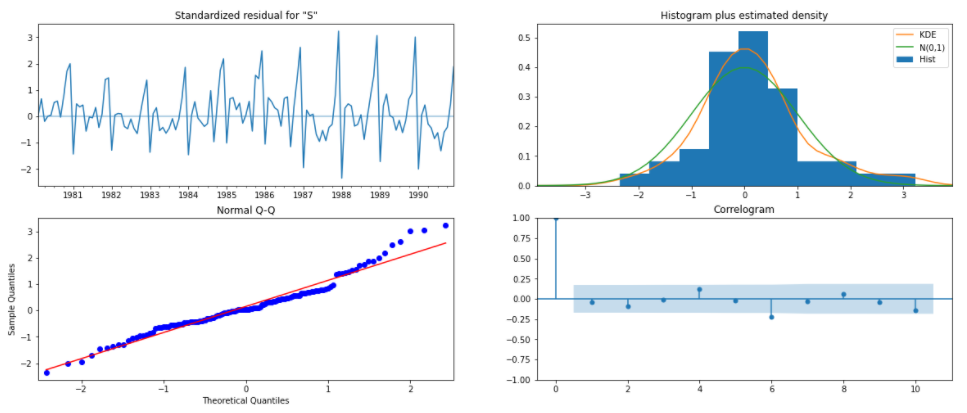


**Observation:** Order (2,1,2) has the minimum AIC, hence putting the same order in the model

* Model creation



* Diagnostic Plot

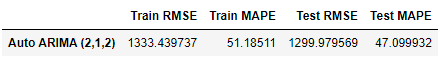


**Observation:** - The diagnostics plot looks like that it tallies with the theoretical values. i.e. Histogram is normal and Blue dots are closer to best fit line

**Predict using Test data using ARIMA model**

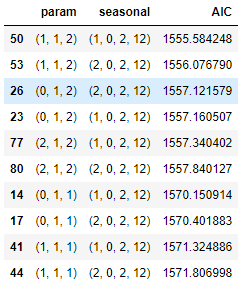


* Model Evaluation



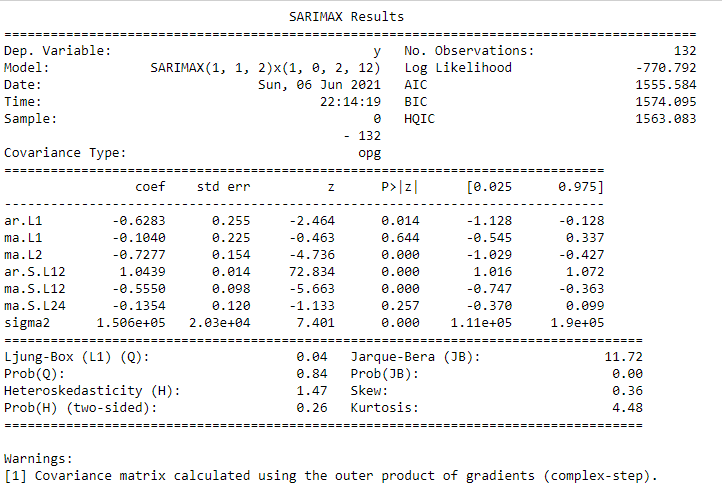
**Model 9: Auto SARIMA Model**

* Looking at minimum AIC value

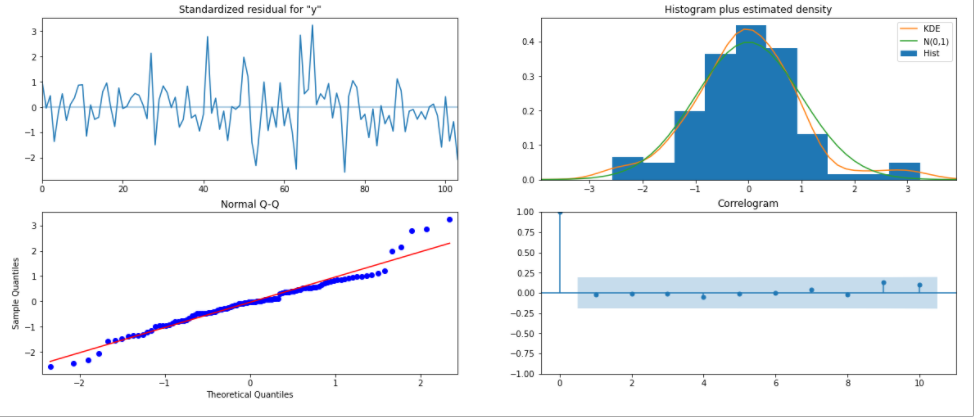


**Observation:** Order (1,1,2)(1,0,2,12) has the minimum AIC, hence putting the same order in the model

* Model creation



* Diagnostics Plot



**Observation:** - The diagnostics plot looks like that it tallies with the theoretical values. i.e. Histogram is normal and Blue dots are closer to best fit line

**Predict using Test data using SARIMA model**



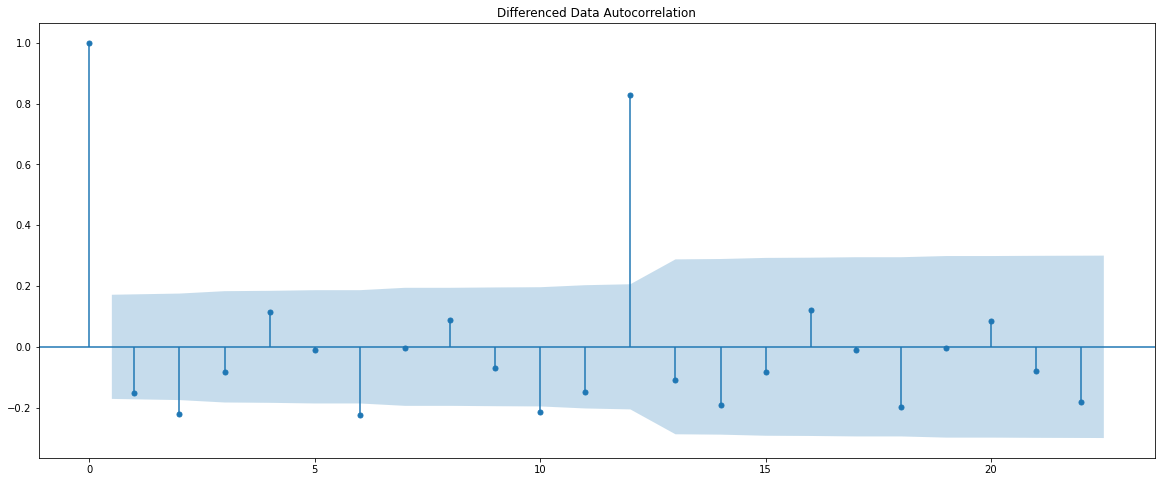
* Model Evaluation

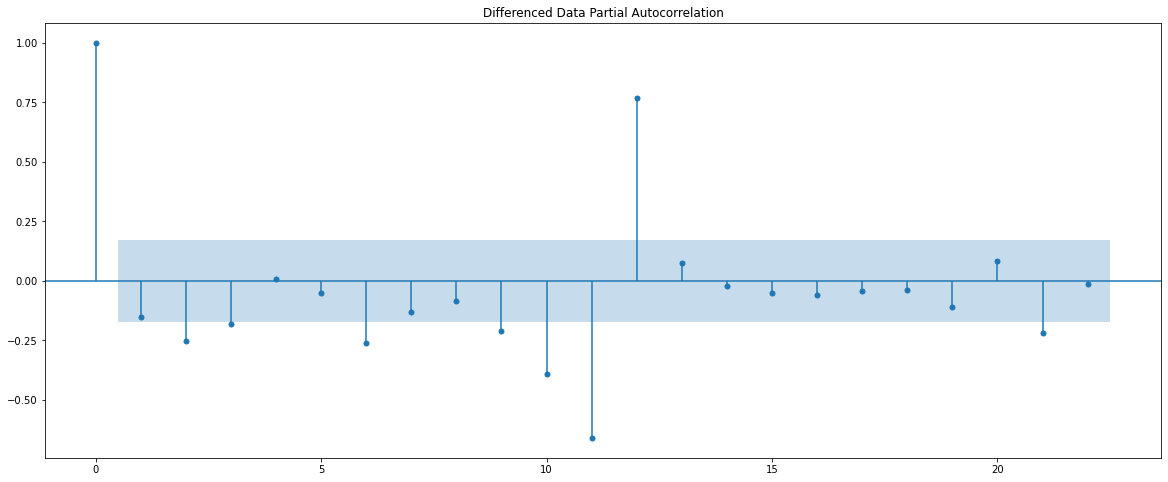


## Q7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

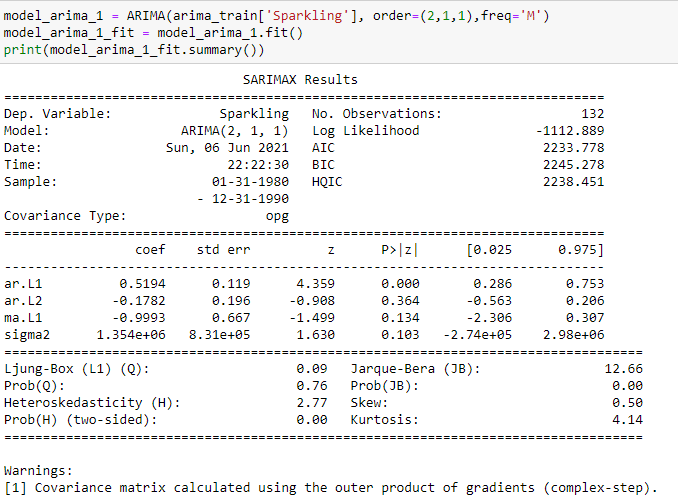
**Model 10: ARIMA model based on cut-off points from ACF and PCF plots**

* Let us look at the ACF and the PACF plots of the train dataset

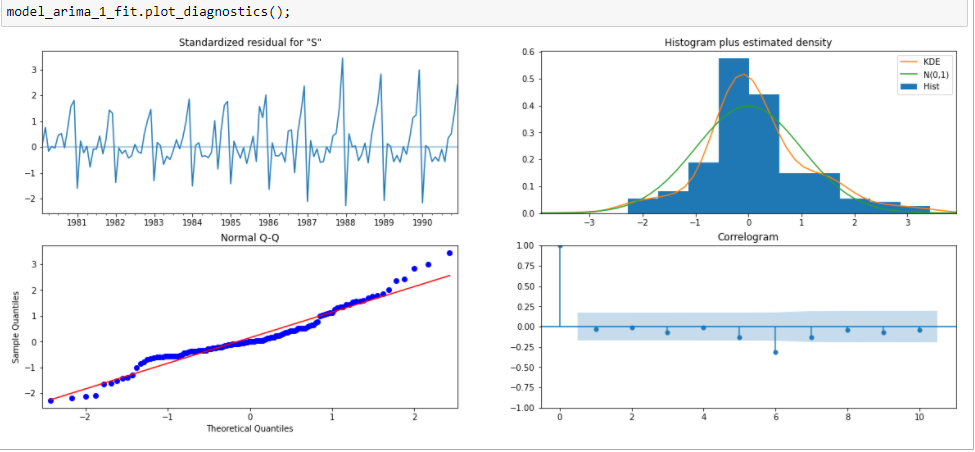




* **Observation:** 
  + Above taken alpha=0.05
  + From the ACF plot it is clear that the lag 1,2 and 12 are significant in a year and lag=12 shows a peak
  + The Auto-Regressive (AR) parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. Value for p=2
  + The Moving-Average (MA) parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0. Value for q=1
  + P=0, Q=0 and D=1
* Model Creation



* Diagnostics Plot



**Observation:** - In this case, our model diagnostics suggests that the model residuals are normally distributed based on the following:

1. The KDE plot of the residuals on the top right is almost similar with the normal distribution.

2. The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken and is a strong indication that the residuals are normally distributed.

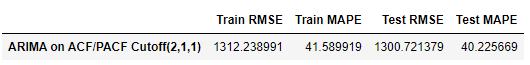
3. The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

Thus it appears that our ARIMA model is working fine.

**Predict using Test data using ARIMA model**

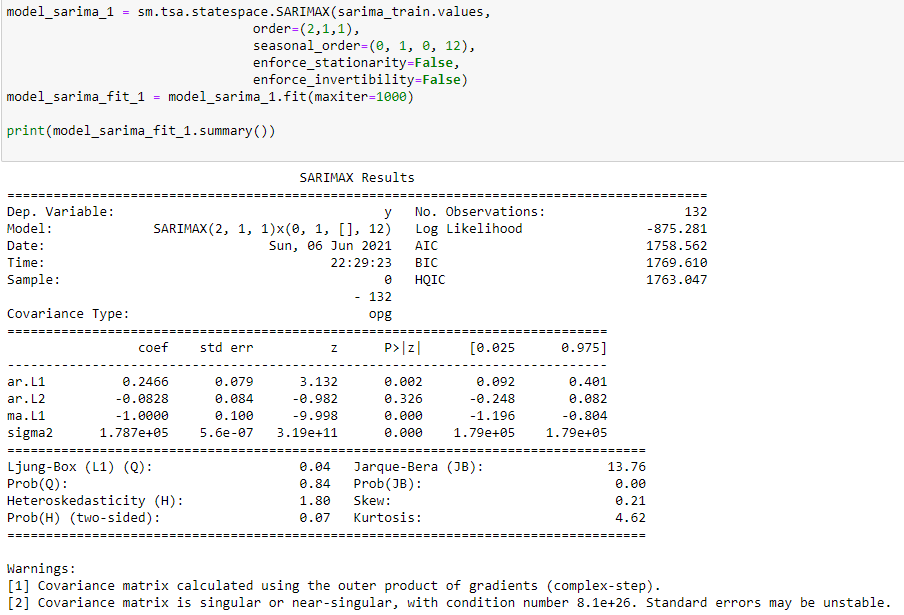


* Model Evaluation

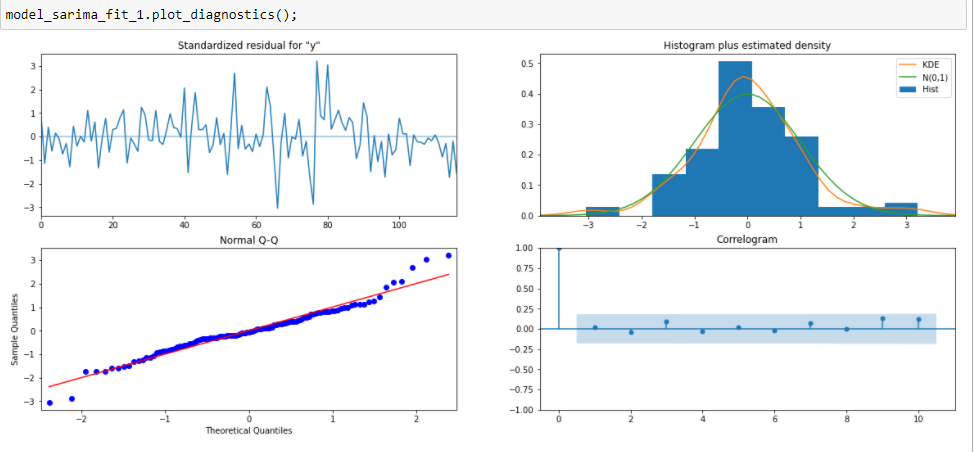


**Model 11: SARIMA model based on cut-off points from ACF and PCF plots**

* From the ACF and PACF graphs above we have Order (2,1,1) Seasonality (0,1,0,12).
* Model Creation



* Diagnostics Plot



**Observation:** - In this case, our model diagnostics suggests that the model residuals are normally distributed based on the following:

1. The KDE plot of the residuals on the top right is almost similar with the normal distribution.

2. The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken and is a strong indication that the residuals are normally distributed.

3. The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

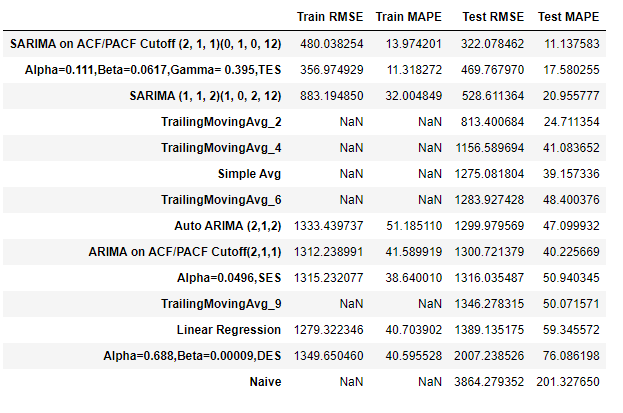
**Predict using Test data using SARIMA model**



* Model Evaluation



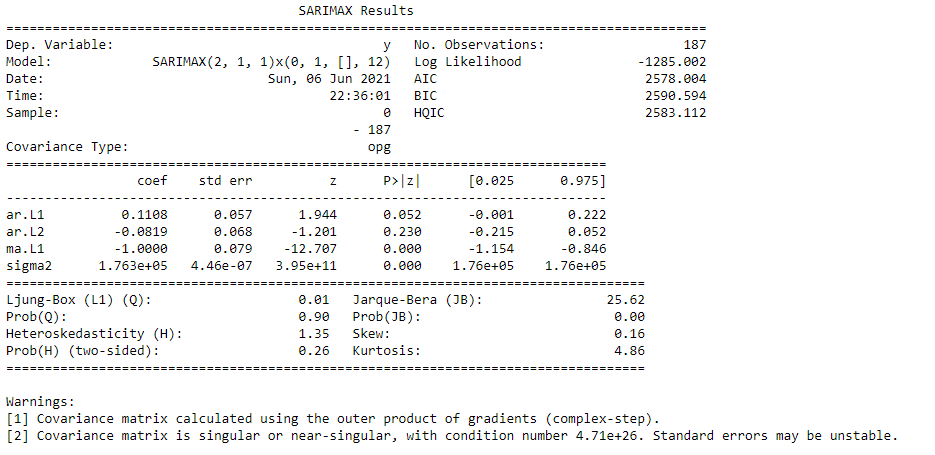
## Q8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

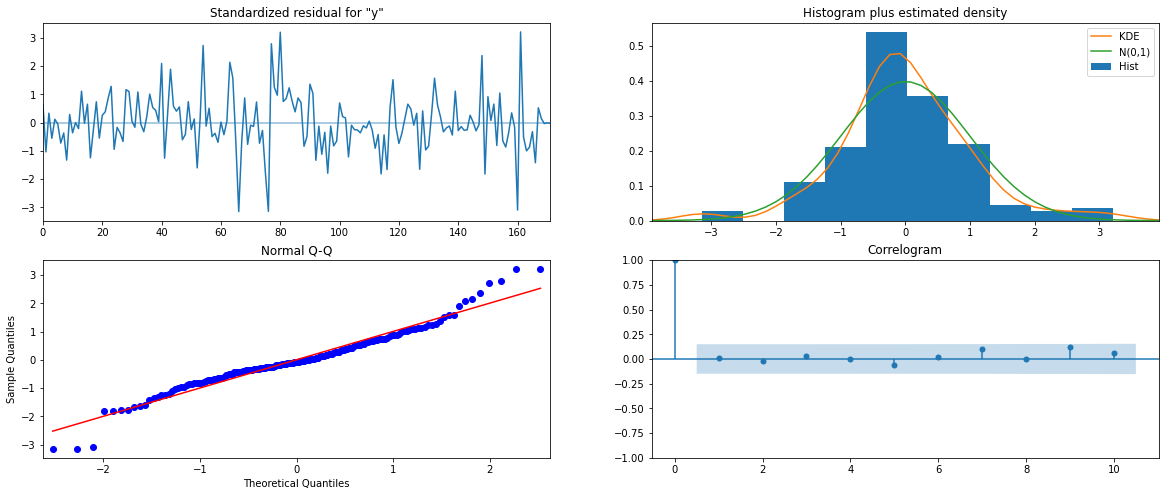


**Observation:** - Amongst all the models SARIMA on ACF/PACF Cutoff (2,1,1)(0,1,0,12) is the best model to forecast as its RMSE and MAPE values are smaller than the other model which means that the error in prediction is much smaller

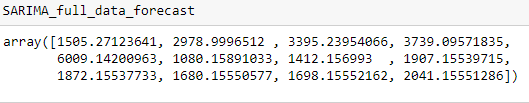
## Q9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

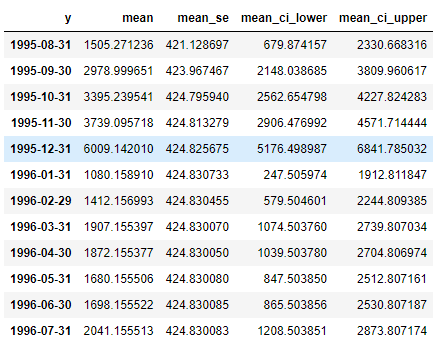
**Final Model - Building SARIMA model on Full Dataset with (2,1,1)(0,1,0,12)**



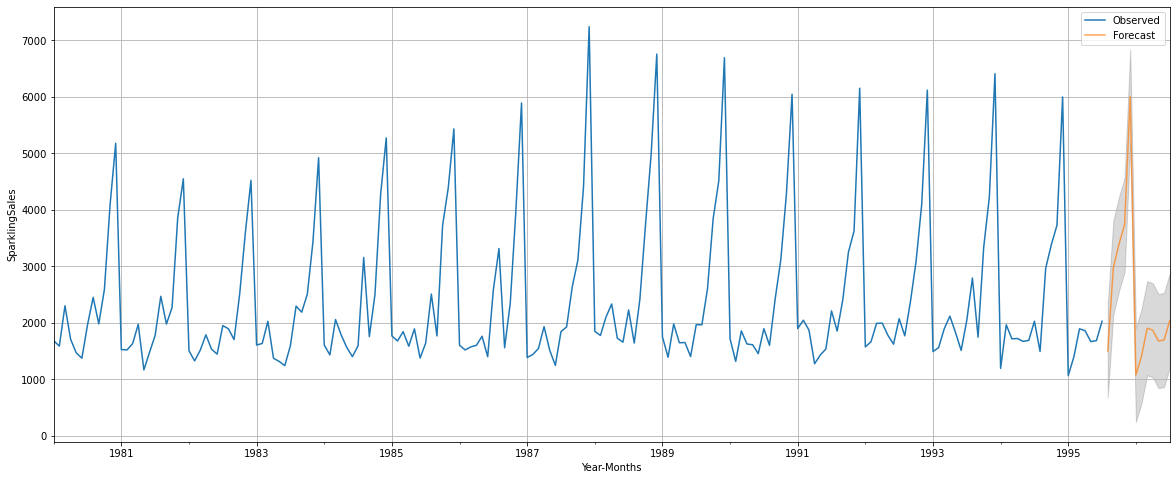


Evaluate the Final model on the whole and predict 12 months into the future





* Plotted the forecast along with the confidence band



## 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 Sparkling wine sales are expected to follow the constant trend seasonal pattern in next 12 months
* As per the forecast the Sparkling Wine sales are expected to be rising till the end of year i.e. December. With December being the highest Sales month. This could be due to the demand generated from festivals like Christmas and vacation period.
* Hence it is recommended to the business that they should stock up the supplies of Sparling Wine for the December month as the sales are nearly 6X as compare to the month of January
* There is steep fall in the sales level for month of January and hence business should think ramp down the production for the months of January to March
* Sharp fall in sales for Jan-1996 is forecasted which is much lower than any historical years. Company should invest or circle around a strategy to counter the decline.
* Overall the sales of Sparkling Wine are constant. While Business should think of
  + Enhance the marketing budget to improve the sales in other months
  + Company can think of giving discounts or can think of cross-sell options with other wine categories like – Red Wine
  + Should focus on improving the quality of the product