TIME SERIES FORECASTING PROJECT

BUSINESS REPORT

SULOCHANA

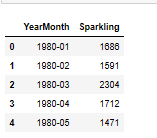
**Problem:**

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

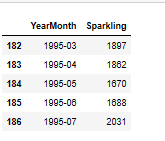
Data set for the Problem: [Sparkling.csv](https://olympus.mygreatlearning.com/courses/78183/files/7099989/download?verifier=BhcXhQpadgKXuNcZyyyt8Cm0H3iyzNwxHWCyEmuW&wrap=1)

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

**Dataset head:**

****

**Dataset tail:**

****

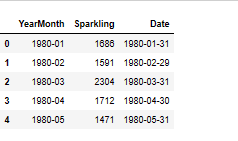
**Data shape:**

****

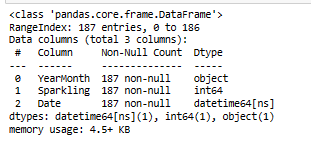
**There are total 187 entries are present in the dataset.**

**The dataset contain two columns, where the first columns contain mentioned about the month and year and second column shows sales.**

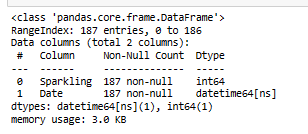
**After extracting the date from the data .Here, new column added with the name of date.**

****

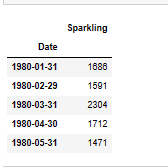
**Data info:**

****

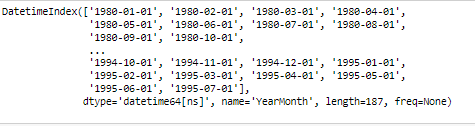
**Now above output shows three columns. Here we can remove yearmonth column after adding date column because both have same information.**

****

**Above output shows two columns. Now we convert date column into raw index. That will be useful for further analysis.**

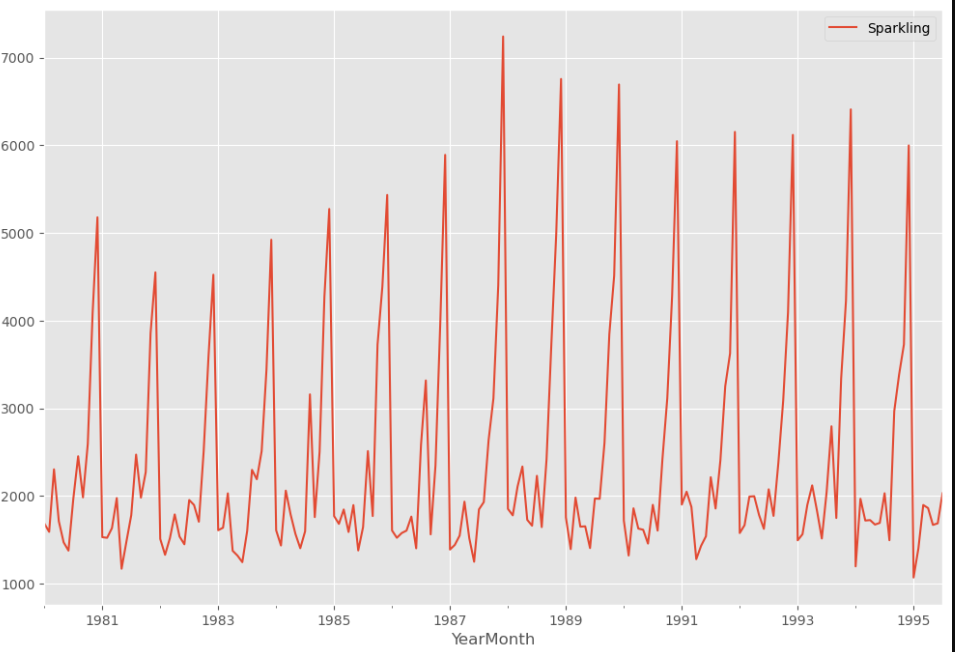
****

**Above output shows date column is converted into index. Now we have single column sparkling.**

****

**The above output shows index of the dataset. The dataset started from the January 1980 going till July 1995.**

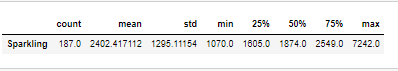
**Line Plot for basic understanding of the data:**

****

**As we can observe from the above plot the sales for sparkling wine shows slightly upward trend. There is a seasonality is visible in this dataset. So, we will explore trend and seasonality during decomposition, where we will be able to view a much detailed on these two factors.**

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

**Descriptive statistics:**

****

* **Based on above output we can say the total observations are 187**
* **The mean of the sparkling wine sale is 2402.**
* **The minimum value of sale is 1070 and the maximum value is 7242.**

**Null value checking:**

****

**As we can see there are no null values in this dataset.**

**Mean and Median of sparkling wine sale in 20th century**

****

**The above output is mean of this sale quantity.**

****

**Based on above output, the median of the sale is 1874.**

## Monthly Observations for year 1980:

## 

## Boxplot of data for each year with year labels:

## 

## As we can observe, in this data wine sales have a variation each year, the years 1985 and 1986 seems to be the years with the least variation. These year’s show some consistency in terms of sales. The highest sale seems to be happening in the 1994 and the lowest in the 1982. The sparkling wine sale appears to be going down from the year 1980 and have started increasing from the year 1985. There outliers in this data, however as it is a time series we can ignore the outlier data.

## 

## Boxplot for each Calendar month across years:

## 

## As we can observe monthly box plots, we can see the seasonality and upward trend. The sales are increasing from July month to December. The sale is relatively low in first two quarters, slowly picks up pace during the third quarter and goes on a rise till the end of the year. Monthly sales data shows skewness.

## Month plot of the Time Series:

## 

## Month sales years:

## 

## 

## 

## As can be observed from the above table and graph, the months of the December seems to be the month that derives the highest sales. The second highest sale of sparkling wine sale happened in the November. We can observe a seasonality element in the graph.

## Conversion of Data to other periodicity (Resampling)

## Data yearly sum:

## 

## 

## Df yearly mean

## 

## 

## As can observed from the above summation table and graph, sparkling wine show a dip initially with sales picking up from the year 1982 right up to the year 1988 and then observing another dip in the sales.

## Resample the data to quarterly frequency and calculate the sum for each quarter:

## 

## Plot the quarterly data as a quarter plot:

## 

## Quarterly sale mean:

## 

## 

## From the output, we can observe that the quarterly sales show an upward trend for wine sale. Also there is a slight element of seasonality in dataset.

## Data daily sale:

## 

## 

## 

### Here we don’t need to observe daily data because the frequency of observations of time series in this data is monthly.

## Data decade sale:

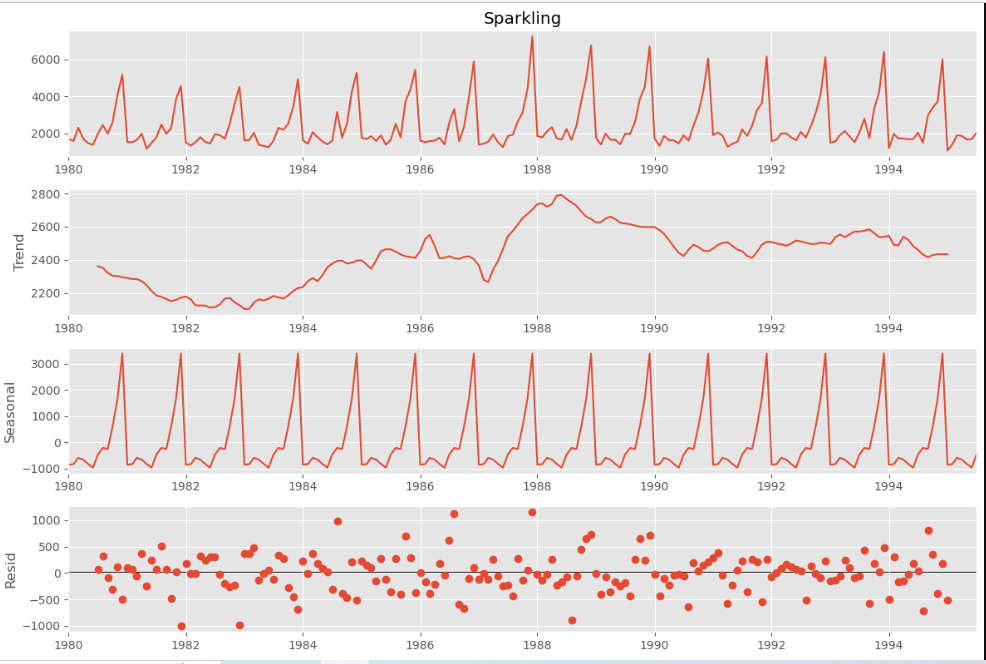
## 

## 

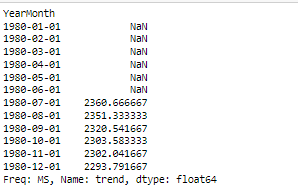
## As we can observe from the above decade sale graph, it show sale rising from 1980 to 1990 and after that sale started to decreasing.

## Time Series Decomposition:

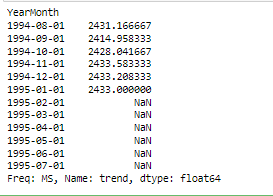
# Additive Model:



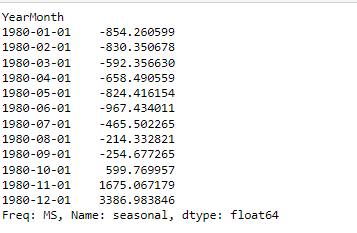
**The first 12 rows of decomposition trend:**



**Last 12 rows of the trend component:**



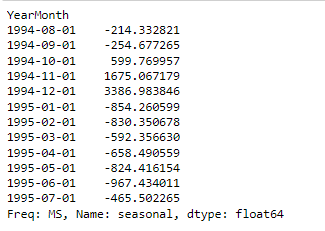
**The first 12 rows of seasonal component:**



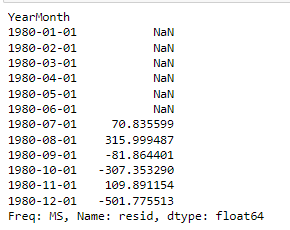
**Output the min and max seasonal component values**:

(-967.4340112433861, 3386.98384589947)

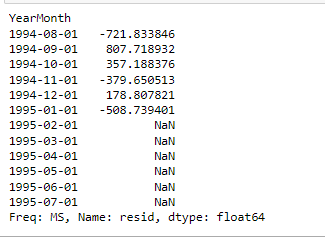
**The last 12 rows of the seasonal component:**



**The first 12 rows of the decomposition residuals:**



**The last 12 rows of the decomposition residuals:**



**Check the mean of the residual component:**

-1.2088458994707487

## De-Seasonalized Time Series:

## Time series values minus the seasonal component:

## 

## Plot the actual time series and the deseasonalized time series in different colors in the same plot:

## 

## Multiplicative Model

## Decompose time series using multiplicative model:

## 

## The first 12 rows of decomposition multiplicative trend:

## 

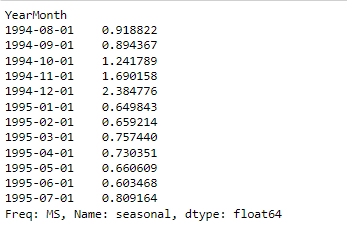
## The last 12 rows of the decomposition multiplicative trend:

## 

## The first 12 rows of the decomposition multiplicative seasonal:

## 

## The last 12 rows of the decomposition multiplicative seasonal:



## The first 12 rows of the decomposition multiplicative residual:

## 

## The last 12 rows of the decomposition multiplicative residual:

## 

## Check the mean of the residual component:

0.9997456359115033

## As we can observe from the above output, we can say that the wine time series are clearly multiplicative in nature and have a seasonal component. Here again we can see upward trend in the dataset. The plots clearly indicate that the wine sales are unstable and not uniform, and they have an apparent seasonality trend. Moreover, the seasonal variation seems to be in this dataset.

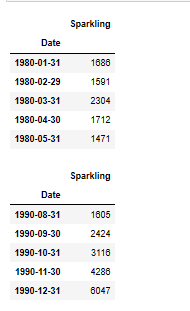
1. **Split the data into training and test. The test data should start in 1991.**

**Here the data spitted into train and test dataset. The train dataset start from the 1980 to 1991 and test data start from the year 1981 to 1995.**

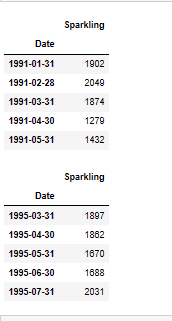
**Train and test data shape**

****

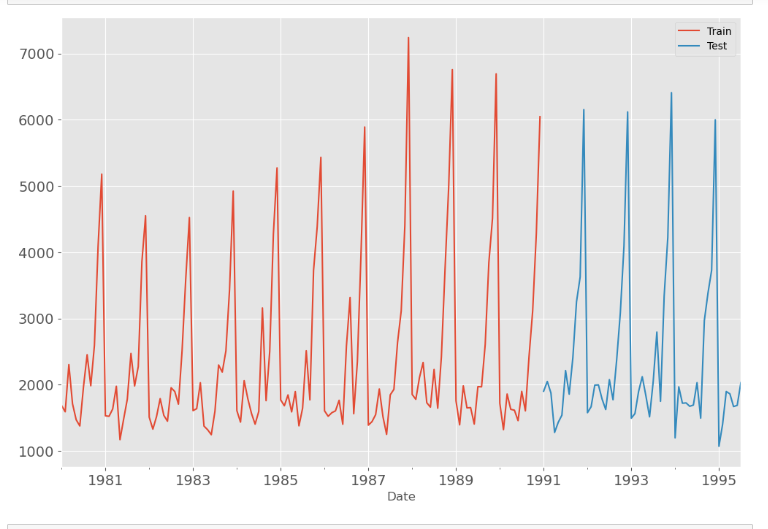
**Train data head and tail**

****

**Test data head and tail:**

****

**The plot for train and test data**

****

**As we can observe, the size of the train data frame is 132 observations and that of the test data frame is 55 observations.**

1. **Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. Should also be built on the training data and check the performance on the test data using RMSE.**

**Now we can build different models to find out which model gives better predictions for future 12 months forecasting.**

**Models**

# Model 1: Linear Regression

# Model 2: Naive Approach: 𝑦̂ 𝑡+1=𝑦𝑡

**Method 3: Simple Average**

**Method 4: Moving Average (MA)**

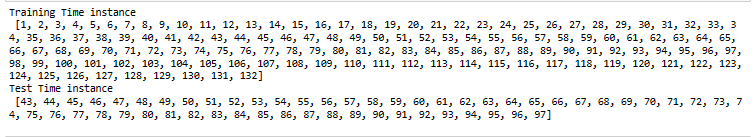
**Method 5: Simple Exponential Smoothing**

**Method 6: Double Exponential Smoothing (Holt's Model)**

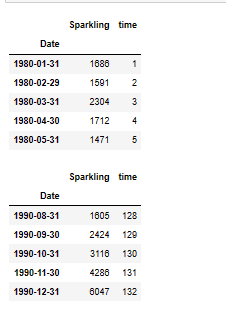
**Method 7: Triple Exponential Smoothing (Holt - winter’s Model)**

**:**

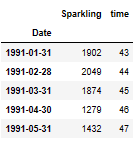
# Model 1: Linear Regression

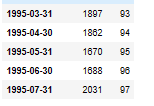
****

**Liner regression train head and tail:**

****

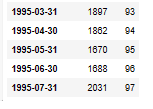
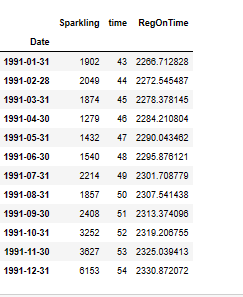
**Liner regression test head and tail:**

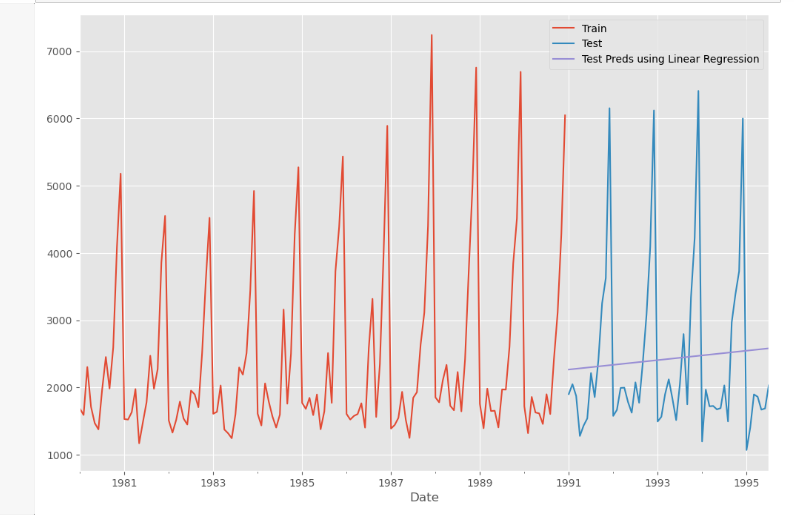
****

****

**The above output shows train and test data for the liner regression model.**

**The below output shows the results of model after fit the liner regression model.**

** **

****

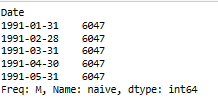
**The regression plots above depict the regression on training set as the red line and that on the test dataset as the blue line. As we can observe from the above plot sparkling wine sale shows upward trend.**

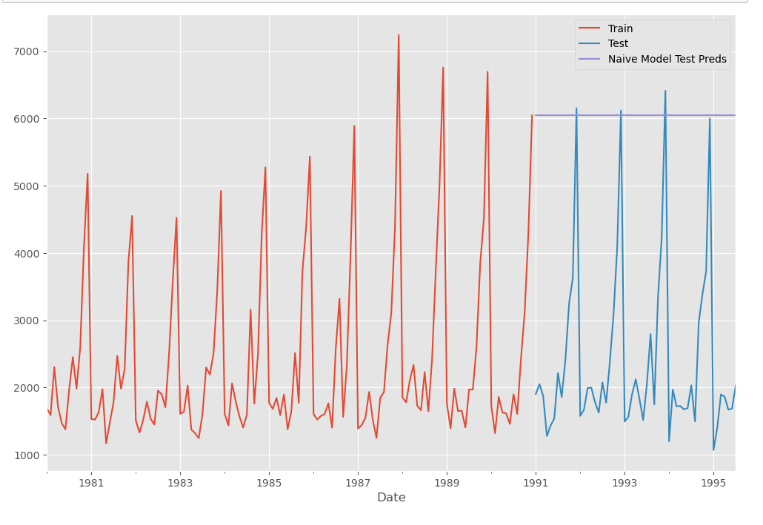
**Test Data – RMSE:**

1275.8670517560186

# Model 2: Naive Approach: 𝑦̂ 𝑡+1=𝑦𝑡

**Picking out the last value in train dataset:**

****

****

**Model Evaluation:**

**Test Data – RMSE**

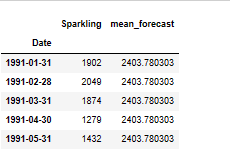
3864.2793518443914

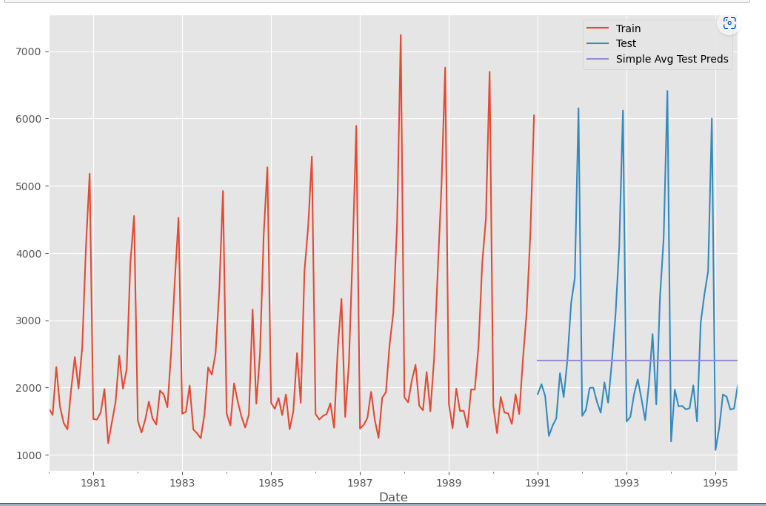
Above outputs are results of after building naïve model on data set. As can be seen from the naïve model performance from the above, we can see it is a worst performance on data. It is not suitable for our data set since the forecasts depends on the previous last observation.

# Method 3: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

The below output is the average of the dataset.

****

****

**Model Evaluation**

**Test Data – RMSE:**

1275.0818036965309

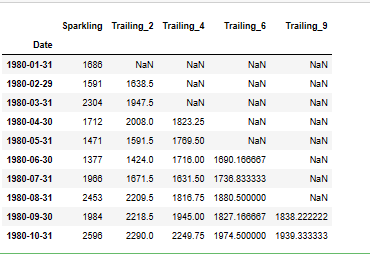
As can be seen from the simple average model performance on this dataset above, the simple average model has the best performance among all the three models run till now for this dataset.

# Method 4: Moving Average (MA)

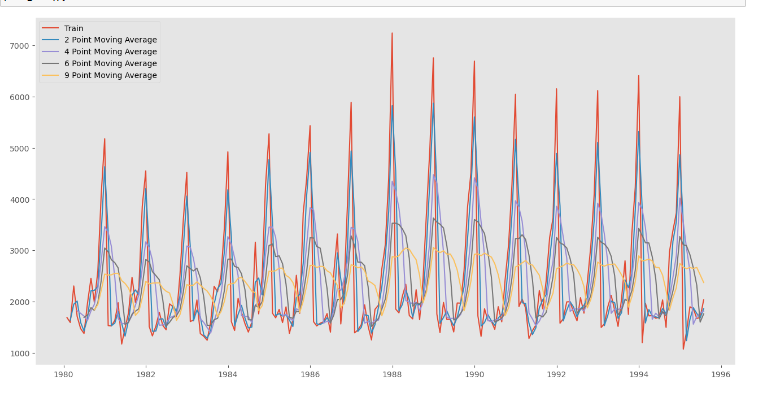
For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

**For Moving Average, we are going to average over the entire data.**

**Trailing moving averages:**

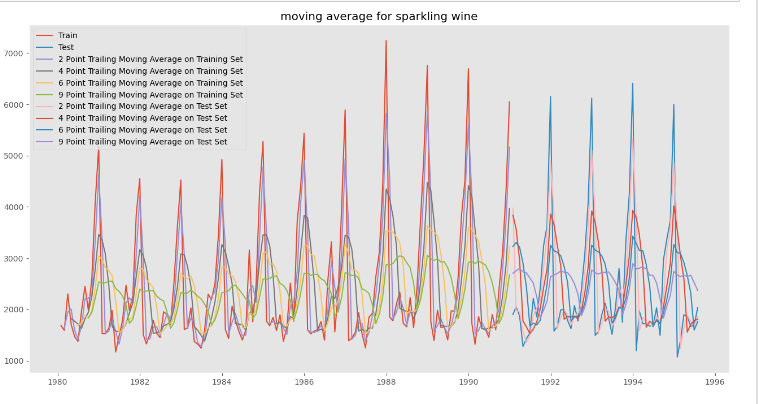
****

**Plotting on the whole data**

****

**Creating train and test set:** Here we split the data into train and test and plot this Time Series.

**Plotting on both the Training and Test data**



**Model Evaluation**

This only we have done on the test data.

Test Data - RMSE --> 2 point Trailing MA

813.4006839972983

Test Data - RMSE --> 4 point Trailing MA

1156.589694081071

Test Data - RMSE --> 6 point Trailing MA

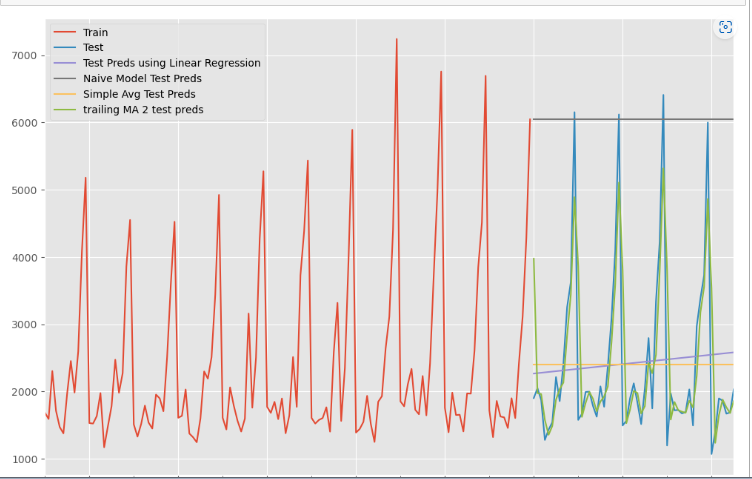
1283.9274280129855

Test Data - RMSE --> 9 point Trailing MA

1346.2783154241804

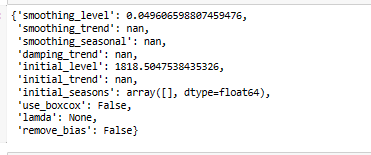
* As we can observe from the above plots, all of the trailing average plots show prediction values below the actual train and test datasets, and the 9 point trailing average plot shows the lowest prediction of all the plots.
* The closest prediction to actual data shown by the 2 point trailing moving average model
* This observation is corroborated by the RMSE scores for each of these moving average models.
* As can be seen from the summarized performance of all the models, the 2 point moving average has shown the best performance of all the models run on the Sparkling wine datasets.

**Before we build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots.**

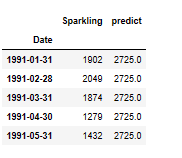


# Method 5: Simple Exponential Smoothing

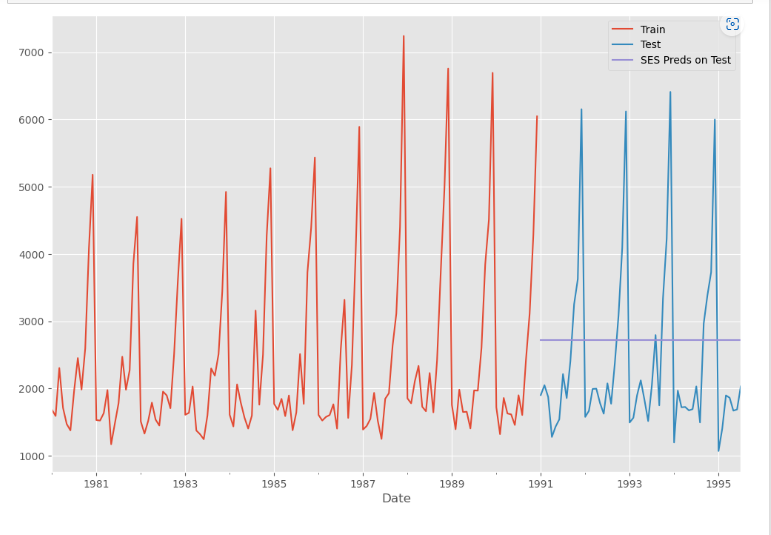
**Simple exponential smoothing autofit params:**



**SES test data head:**



**Plot on SES train, test and predicted data**

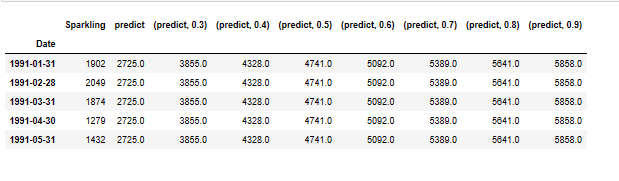


**Model Evaluation for Simple Exponential Smoothing**

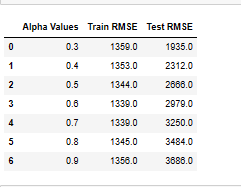
**Test Data Rmse:**

1316.0521680734807

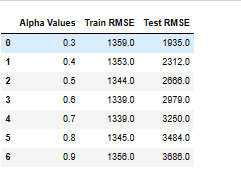
**Below is a ouptput after applying different alpha values:**

****

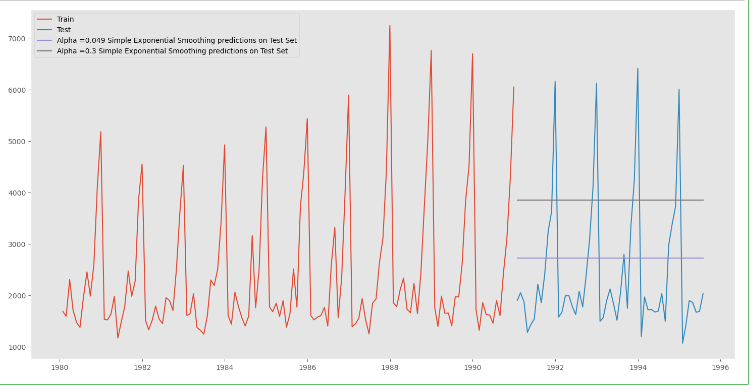
**Model Evaluation**

****

**After sorting the test Rmse values:**

****

**Plotting on both the Training and Test data:**

****

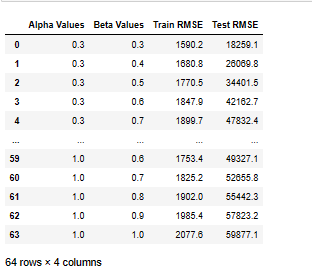
**As we all know that SES model should be used on data which has no element of trend or seasonality. So, on this dataset it did not perform well as compared to the previous models.**

# Method 6: Double Exponential Smoothing (Holt's Model)

##### *Two parameters***𝛼***and***𝛽***are estimated in this model. Level and Trend are accounted for in this model.*

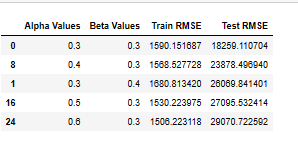
**Identifying Optimum Values of 𝛼 & 𝛽**

**Below is a ouptput after applying different alpha and beta values:**

****

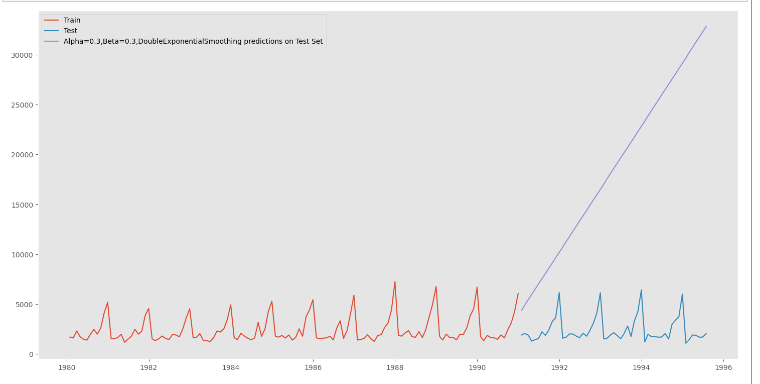
**Following are the results from after running the DES model on this dataset.**

**Sorting test Rmse values for finding best alpha and beta values:**

****

For Alpha = 0.3, Beta = 0.3 Simple Exponential Smoothening Model forecast on the Test data RMSE = 18259.110704

**Plotting on both the Training and Test data:**

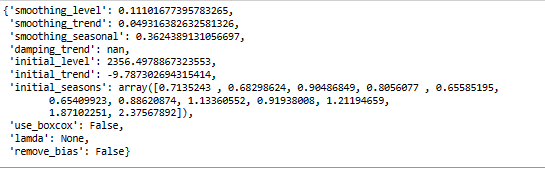
****

**As we all know that DES model should be used on data which has no seasonality but has level and trends. So this gives a worst performance on dataset as compared to other models.**

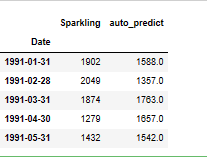
# Method 7: Triple Exponential Smoothing (Holt - winter’s Model)

##### Three parameters **𝛼**, **𝛽** and **𝛾** are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

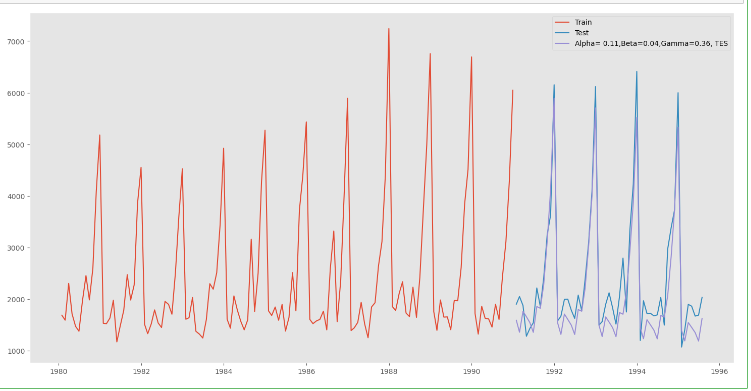
**TES model autofit params:**

****

**Prediction on the test data:**

****

**Plotting on both the Training and Test using autofit:**

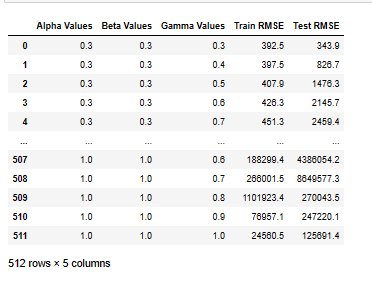
****

**Test Data Rmse:**

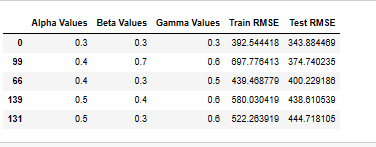
402.923001893098

**Identifying Optimum Values of , 𝛽 and 𝛾**

**Below is a ouptput after applying different alpha, beta and gama values:**

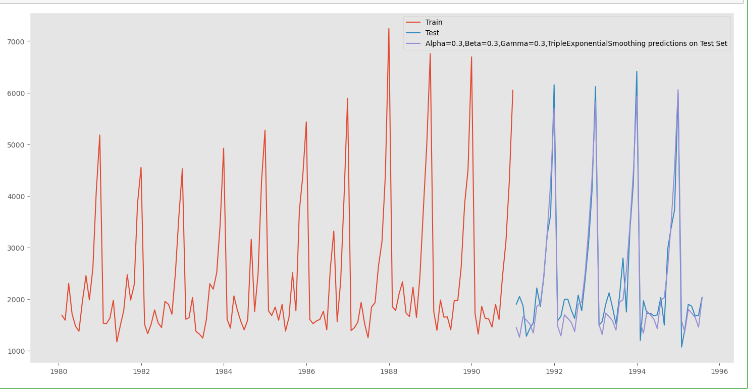
****

**After sorting the values:**

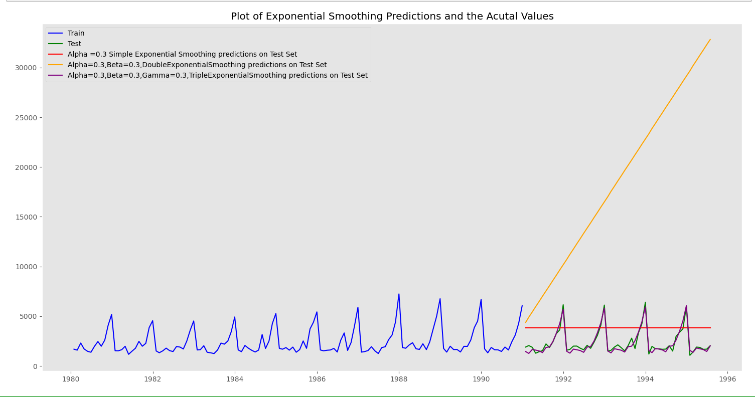


For Alpha = 0.3, Beta = 0.3, gamma = 0.3 triple Exponential Smoothening Model forecast on the Test data RMSE = 343.884469.

**Plotting on both the Training and Test data using brute force alpha, beta and gamma determination:**



**Plotting on both the Training and Test data and comparing the 3 exponential models:**



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

We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters 𝛼 = 0.3, 𝛽 = 0.3 and 𝛾 = 0.3.

1. **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.**

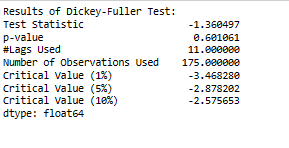
**Check for stationarity of the whole Time Series data:**

**Augmented dicky fuller test is used on this data for checking for the stationarity of the data.**

**Test for stationarity of the series - Dicky Fuller test:**

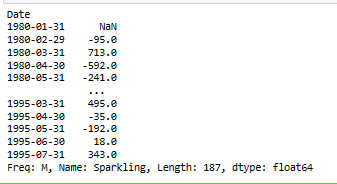
**P- value:**

0.6010608871634863

****

We see that at 5% significance level the Time Series is non-stationary as at high p-value we are unable to reject the null hypothesis.

**Let us take a difference of order 1 and check whether the Time Series is stationary or not.**

****

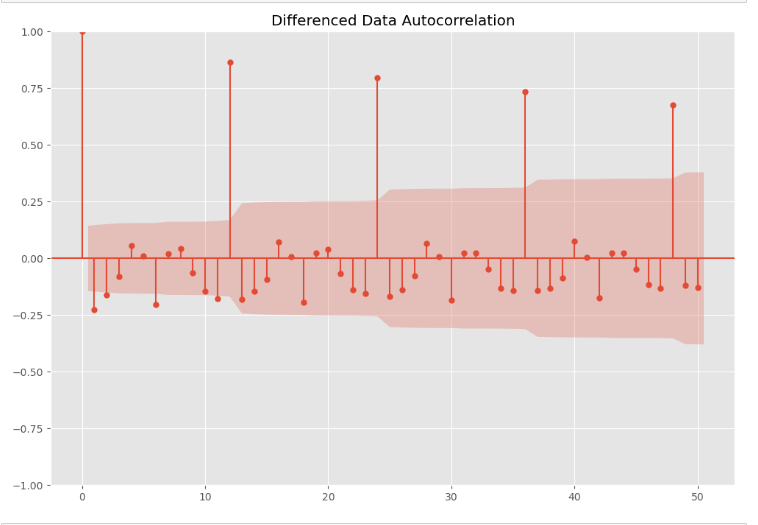
**After differencing the data p value:**

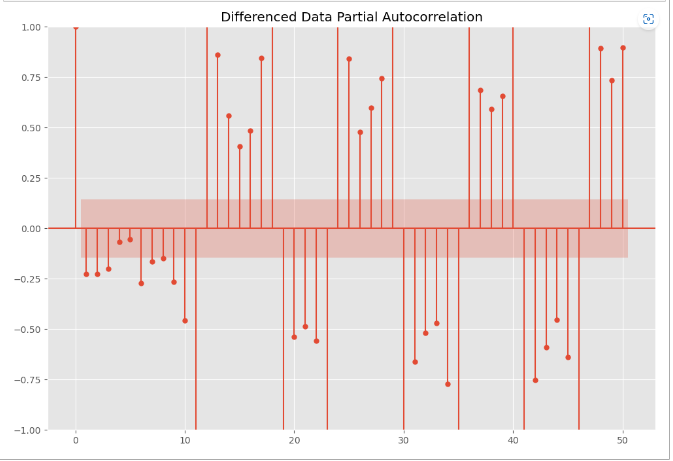
0.0

After difference of order 1 we see that at 𝛼 = 0.05 the Time Series is indeed stationary as the p-value is lower than 0.05 and hence we can reject the null hypothesis which says that the time series is not stationary. So differentiation by 1 makes the time series stationary.

1. **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.**

**Autocorrelation & Partial Autocorrelation function**

****

****

From the above plots, we can say that there seems to be seasonality in the data as there are significant correlations at multiples of 6 and 12 lags.

**Check for stationarity of the Training Data Time Series.**

**After performing augmented dicky fuller test on the train data this is the P value**

0.6697444263523331

**We see that the series is not stationary at 𝛼 = 0.05.**

**After differencing the data p value:**

2.2801043558263994e-12

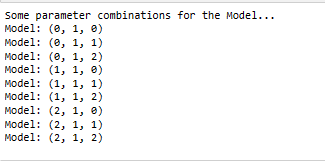
After difference of order 1 we see that at 𝛼 = 0.05 the Time Series is indeed stationary as the p-value is lower than 0.05 and hence we can reject the null hypothesis which says that the time series is not stationary. So differentiation by 1 makes the time series stationary.

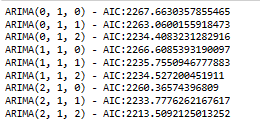
# Automated version of an ARIMA model: Based on lowest Akaike Information Criteria (AIC).

The data has some seasonality so ideally we should build a SARIMA model. But here we are building an ARIMA model both automatically and manually, by looking at the minimum AIC criterion and manually, by looking at the ACF and the PACF plots.

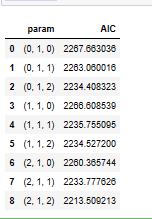
**ARIMA MODEL:**

**Following are the results of ARIMA model on data:**

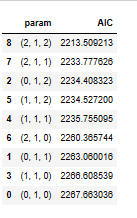


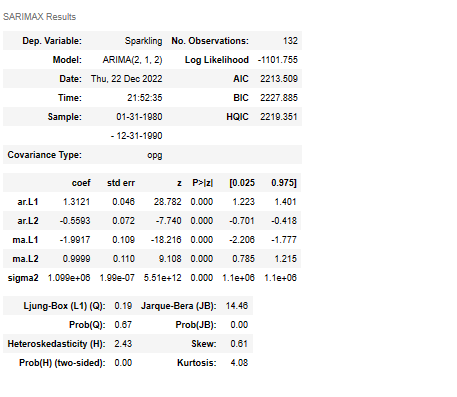
****

**ARIMA\_AIC**

****

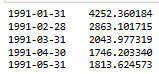
**Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value:**

****

****

**Predict on the Test Set & Evaluation - Auto Arima Model**

**Predicted auto ARIMA head:**

****

**Predicted auto ARIMA [0] contains the predictions (mean values):**

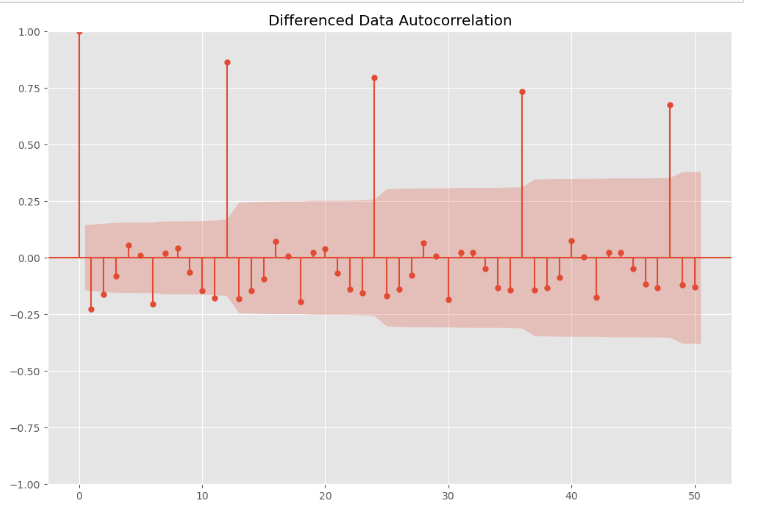
**RMSE autoarima:**

2249.492599772577

The lowest AIC of sparkling data is 2249.492600 for p,d,q values of 2,1,2 respectively.

**Automated version of a SARIMA model -Parameter Selection with lowest Akaike Information Criteria (AIC).**

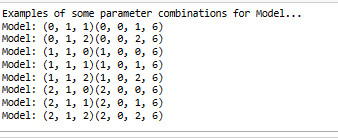
Let us look at the ACF plot once more to understand the seasonal parameters PDQ for the SARIMA model.

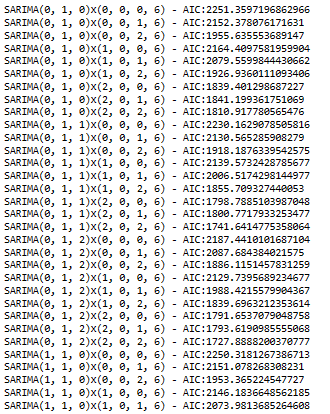
****

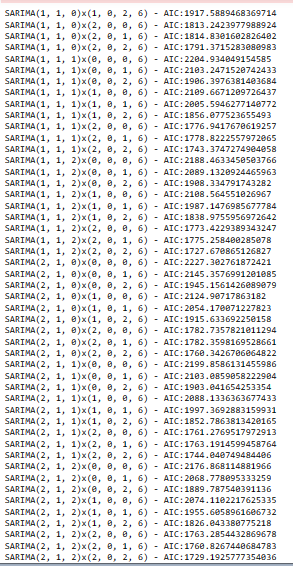
We see that there can be a seasonality of 6 and 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

**Auto SARIMA model - With Seasonality as 6**

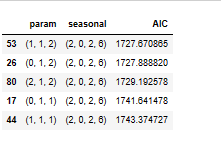
**The following are results of SARIMA model on sparkling data.**

****

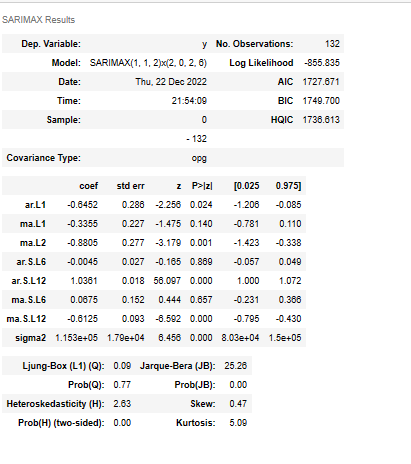
****

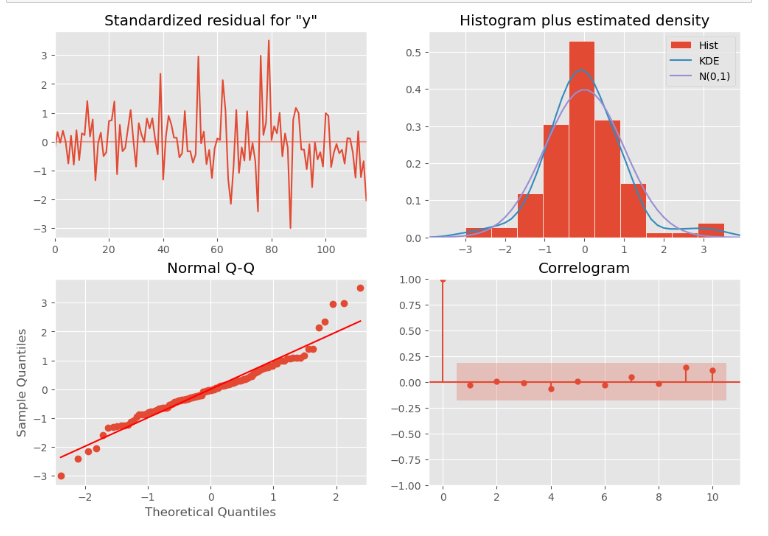
****

**Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value:**

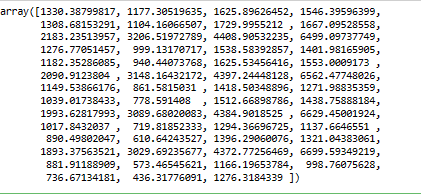
****

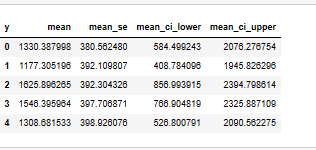
**build a SARIMA model using the pdq and PDQ values identified above. Following results is after run the model on data.**

****

****

**Prediction on the Test Set & Evaluation:**

****

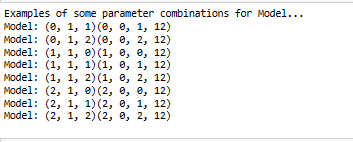
****

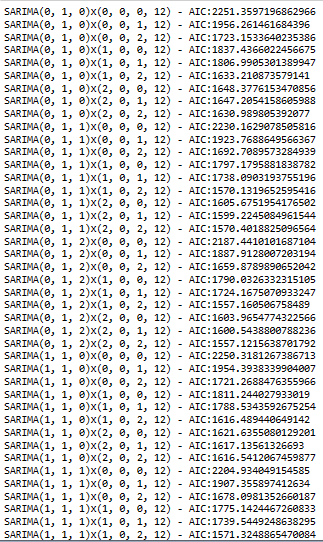
**Rmse of auto SARIMA as 6 month seasonality:**

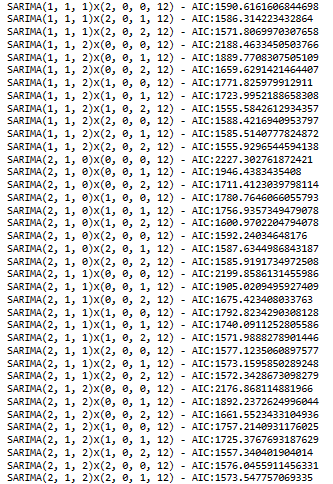
626.854441818761

The lowest AIC of SARIMA model on sparkling data is 1727.670865 for P, D, Q values of 1, 1, 2 respectively.

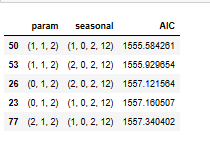
**Setting the seasonality as 12 for the second iteration of the auto SARIMA model.**

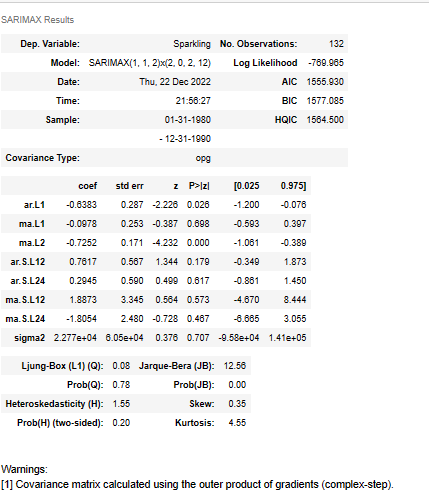
****

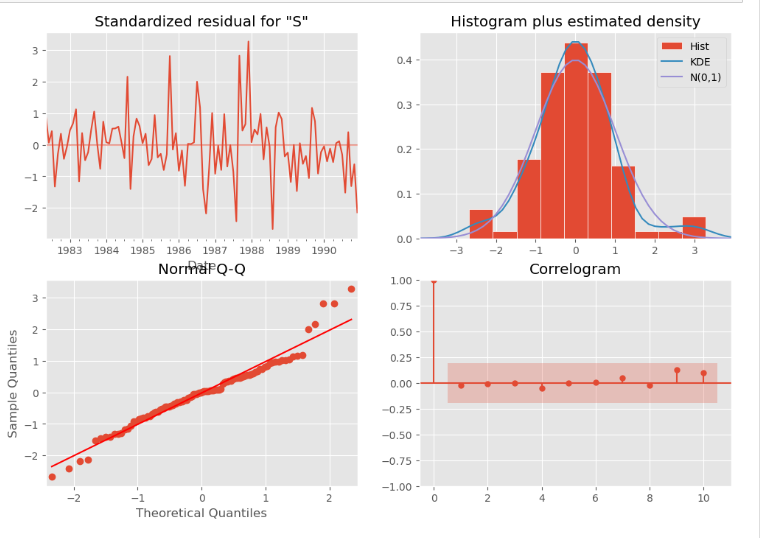
****

****

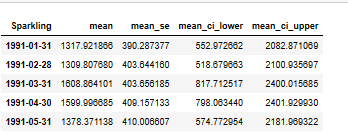
**Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value:**

****

****

****

**Predict on the Test Set & Evaluation**

****

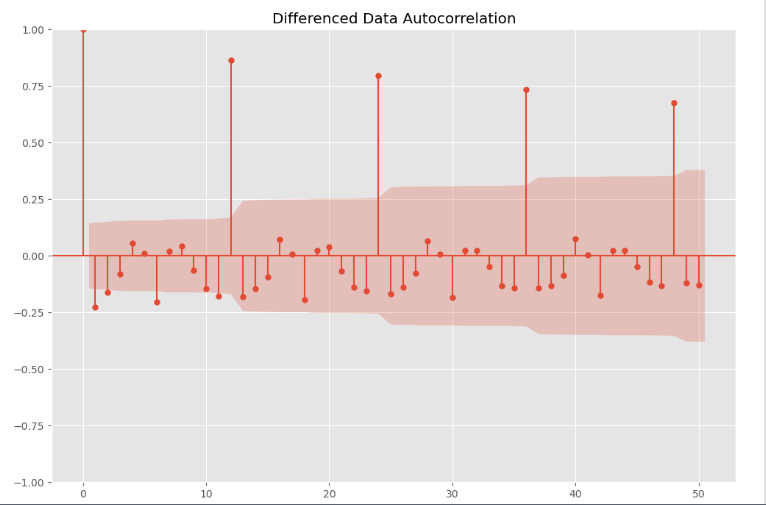
**Rmse auto sarima as 12 month seasonality:**

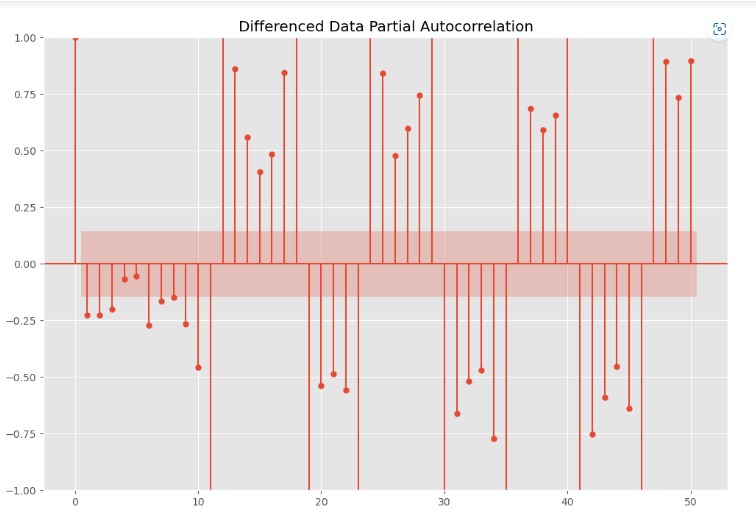
546.383792177467

As can be observed, the lowest AIC of SARIMA model on sparkling data is 1555.584261for P, D, Q values of 1, 1, 2 respectively.

1. **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.**

**Manual ARIMA model - Using ACF & PACF plots:**

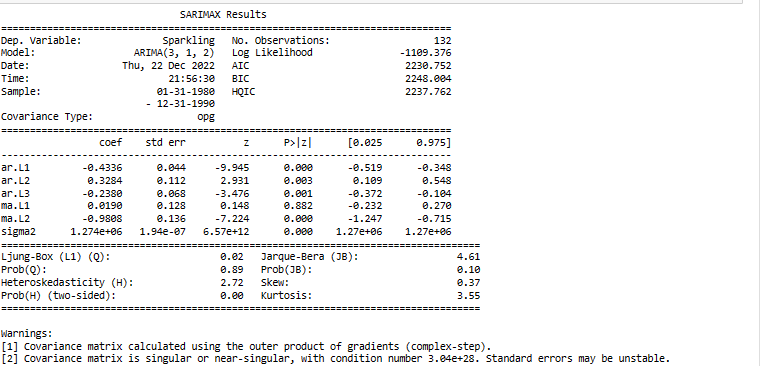
****

****

### Here, we have taken alpha=0.05.

The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 3. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2. By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 2 and 3. So, our pdq values are 3, 1, 2.

The following are the results of manual ARIMA model on sparkling dataset:

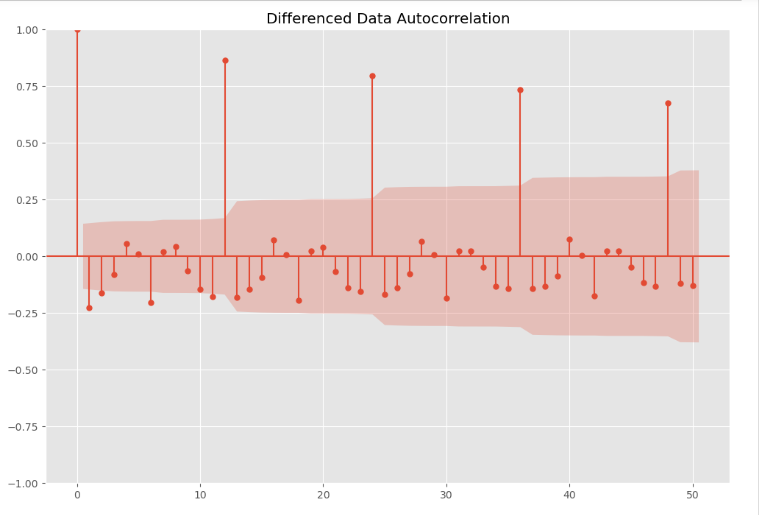


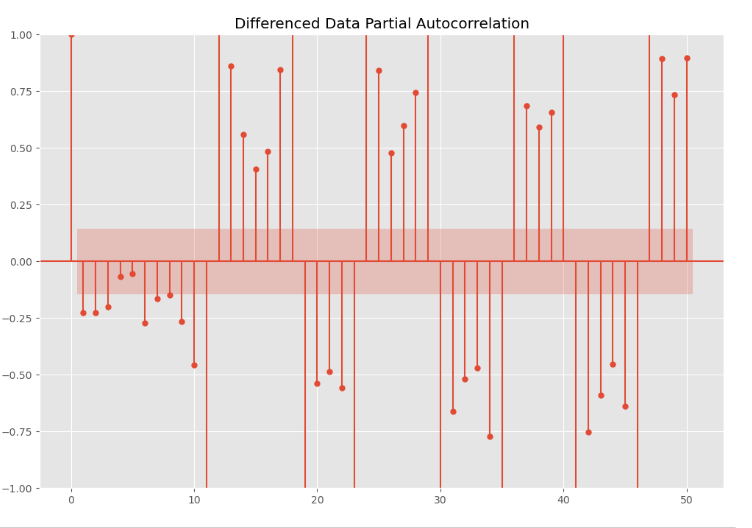
**Predict on the Test Set & Evaluation**

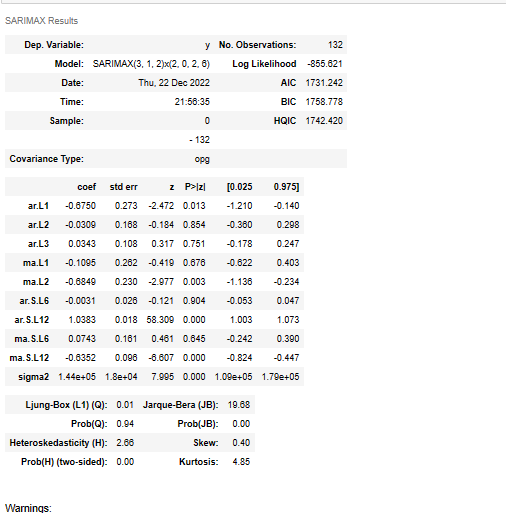
**RMSE manual ARIMA:**

1283.1274142895754

**Manual SARIMA model setting the seasonality as 6- Using ACF & PACF plots:**

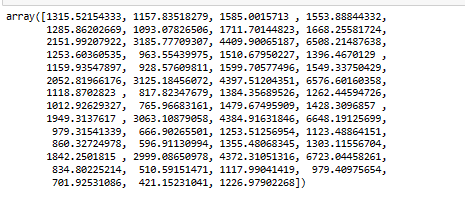
****

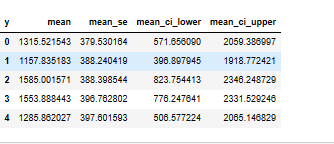
****

****

**Predict on the Test Set & Evaluation:**

**Predicted manual SARIMA 6month**

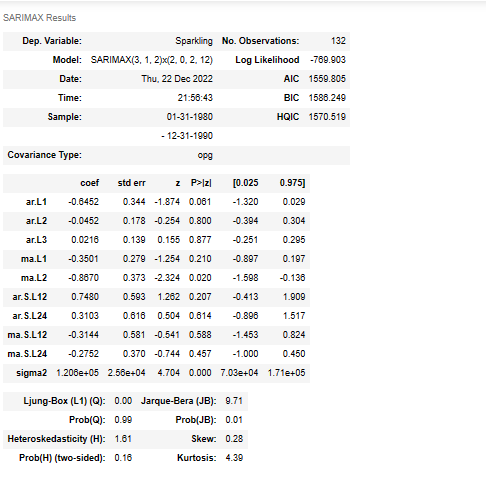
****

****

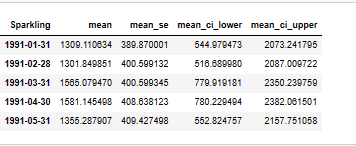
**Rmse manual sarima 6 month**

648.59610108432

**Manual SARIMA model setting the seasonality as 12- Using ACF & PACF plots**

****

**Predict on the Test Set & Evaluation**

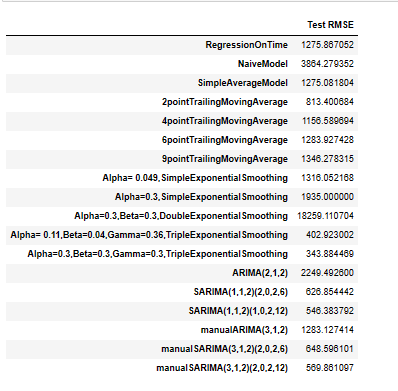
****

**Rmse manual sarima12**

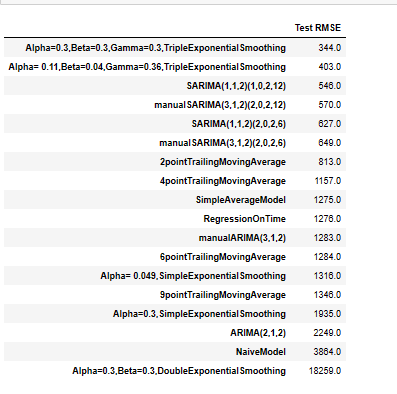
569.8610972484652

1. **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.**

**These are the results after performing different models on this data set.**

****

**Sort the above test rmse values in the ascending order to find out the best optimum model on the sparkling wine sale for prediction of future 12 month forecasting.**

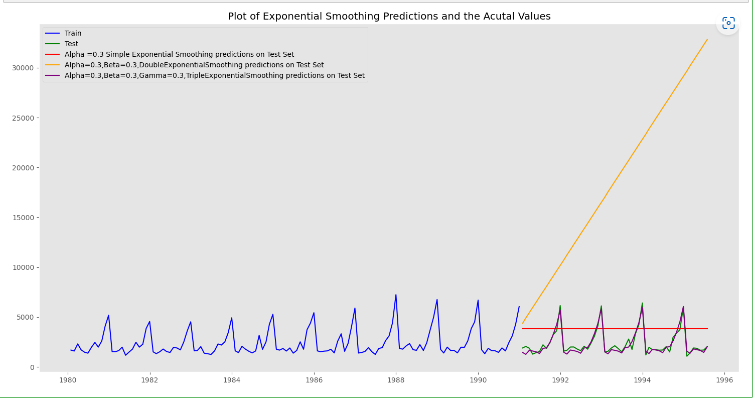
****

**After building various models on this dataset we can say the triple exponential model is the best optimum model with the parameters α = 0.3, β = 0.3, ϒ = 0.3 as compared to other models. This model gives lowest test rmse value when compared to other models.**

1. **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.**

We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters 𝛼 = 0.3, 𝛽 = 0.3 and 𝛾 = 0.3.

**Plotting on both the Training and Test data and comparing the 3 exponential models**

****

**From above output we can clearly observe the prediction of triple exponential model is replicating the original test data. So, we can say it performs very well.**

**Evaluate the model on the whole and predict 12 months into the future.**

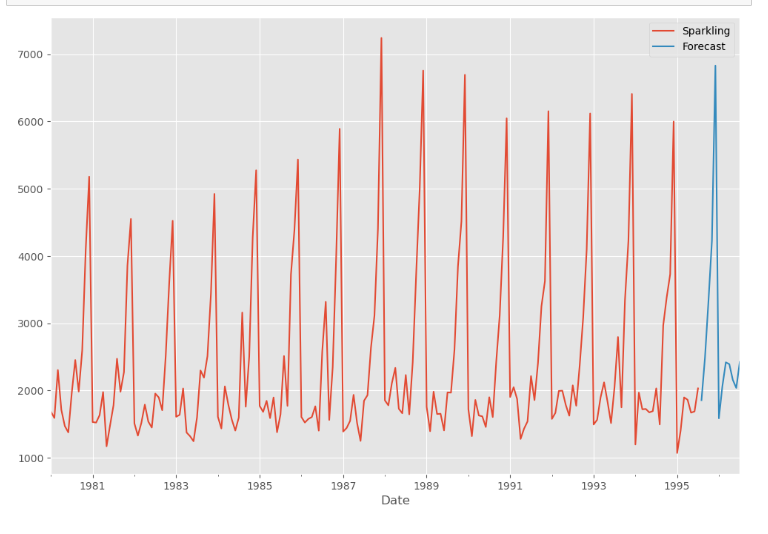
**These following are the results of after run the model on whole data.**

**After predicting 12 months into future the below is the Rmse value.**

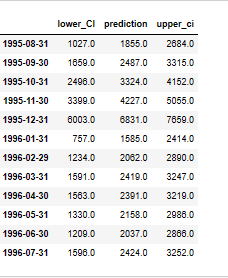
**RMSE on fullmodel:**

421.39248511567126

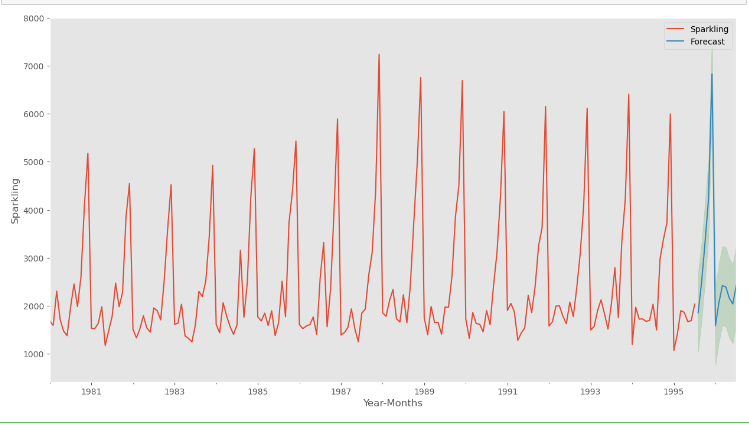
**Plot the forecast**

****

**Prediction Confidence Intervals: Margin of Error**

****

**Plot the forecast along with the confidence band**

****

**Finally the above output is the forecasting future 12 months by using triple exponential model with parameters α = 0.3, β = 0.3, ϒ = 0.3.**

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

* Sales in Sparking does not have uniform trend but increased in some years and decreased later.
* Business study may be done to find why sales are not increasing and what the contributing factors.
* Study can also include seeing which wine product has substituted/ had higher sales in the years of low sales of Sparkling.
* With promotion and focused effort with micro detailing it may be feasible to increase the sales.
* Sales of Sparkling wine higher in the later part of the year. This may be due to climatic condition of the geography under study.
* The company should come up with discount offers in the months of January to June as sales are low in this month.
* When we observe as decade plot the sale is gradually increasing from 1980 to 1990 and started to decrease after 1990. This may be due issue of quality or competition of other wine products in the market.
* Hence, the company should focus on quality of the wine and should promote the product with different offers. Once it started to pick up the sale then we can continue normal sale.