**PROJECT REPORT ON TIME SERIES ANALYSIS**

**Akshaya Parthasarathy**

**Batch: PGPDSBA\_online\_July E 2020**

**Problem 1:**

**Time Series Analysis for Sparkling dataset**

**Problem statement:**

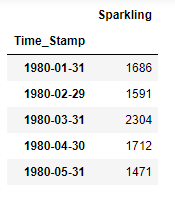
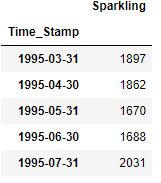
For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

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

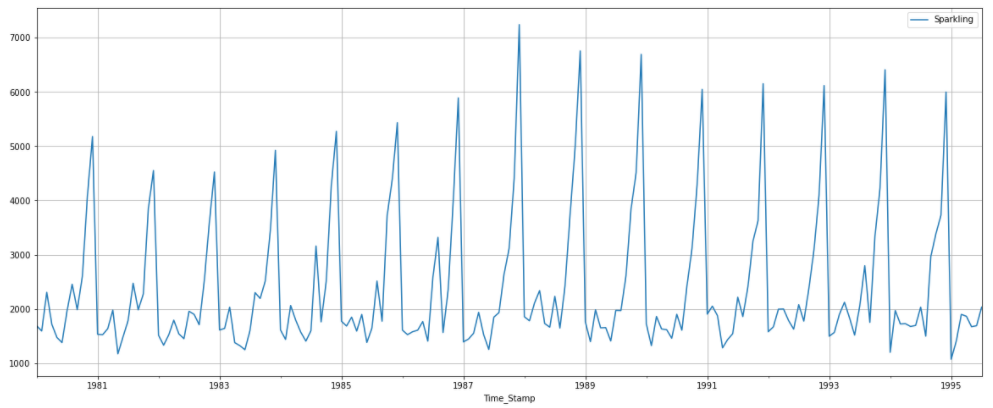
Sparkling dataset is read and converted into time series data with Time\_stamp as index and sparkling wine sales as a single column of a dataframe.

**Final Dataset:**

Head of the dataset Tail of the Dataset

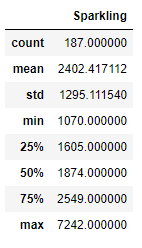
**Plot of the Dataset:**



Observations from the plot:

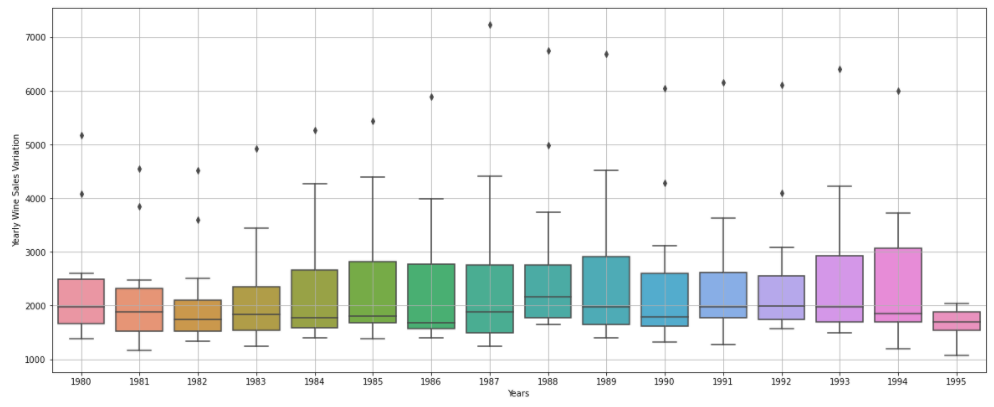
* Sparkling wine sales plot has no visible trend as such in it.
* There are very clear seasonality peaks. Small multiplicative factor is visible in beginning but gradually seasonal peaks are maintained over the later years.
* There do not appear to be any outliers.
* The seasonality suggests that the series might not be stationary.
  1. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

**Summary statistics of the dataset:**

****

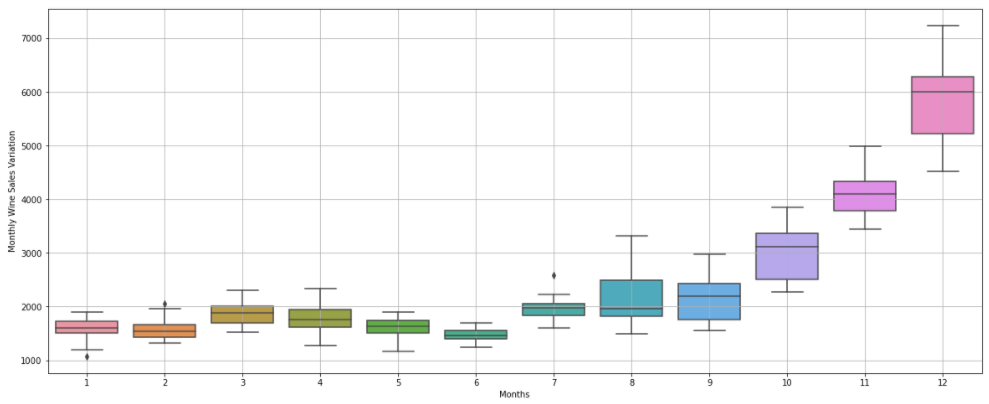
* There are no null values in the dataset.
* The mean is about 2402.42, which can be considered as level for the series.
* Standard deviation of 1295.11 is relatively large and percentiles along with standard deviation indicates the large spread of the data.

**Year Plot:**

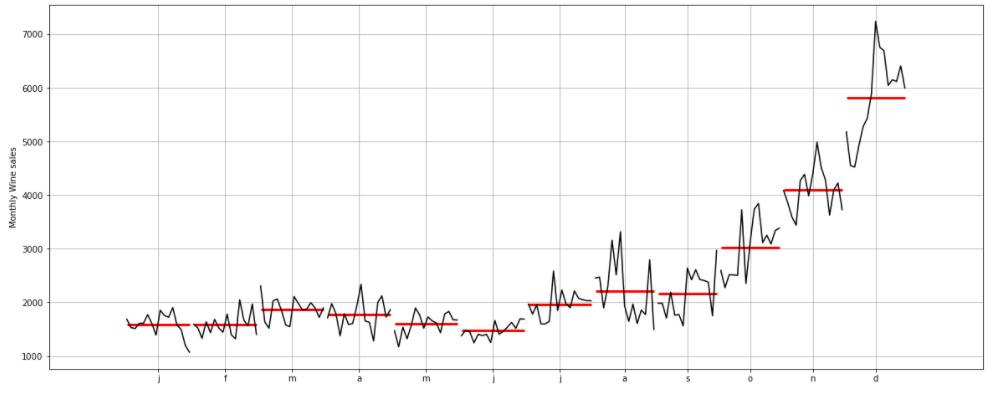
****

* The median line for each year is almost ranging in and around 2000s. Moreover, median is close to the 25th percentile which means for half of the months sales is in the lower range.
* Spread also varies for each year and we could see outliers in the all the years which maybe a part of the seasonal cycle. In a year at least one-month sale is more compared to the others.

**Month plot:**

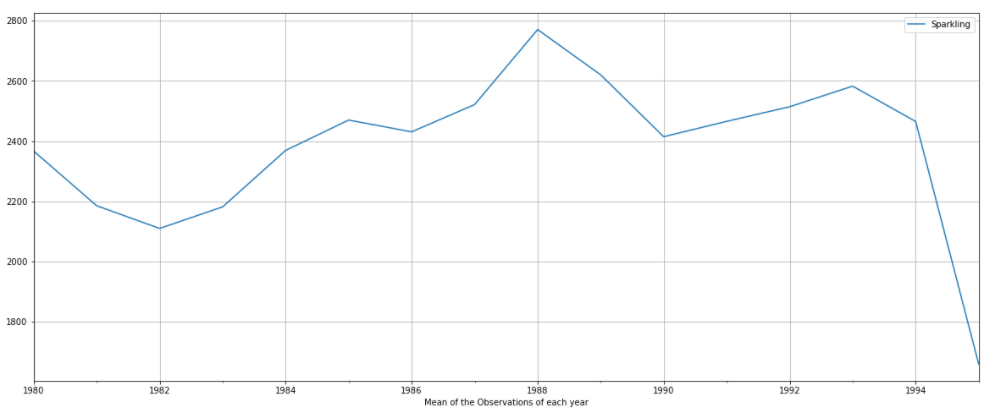
****

* Wine sales are gradually increasing over the months and peaks at December.
* December month sales has a wider spread compared to the rest of the months.

****

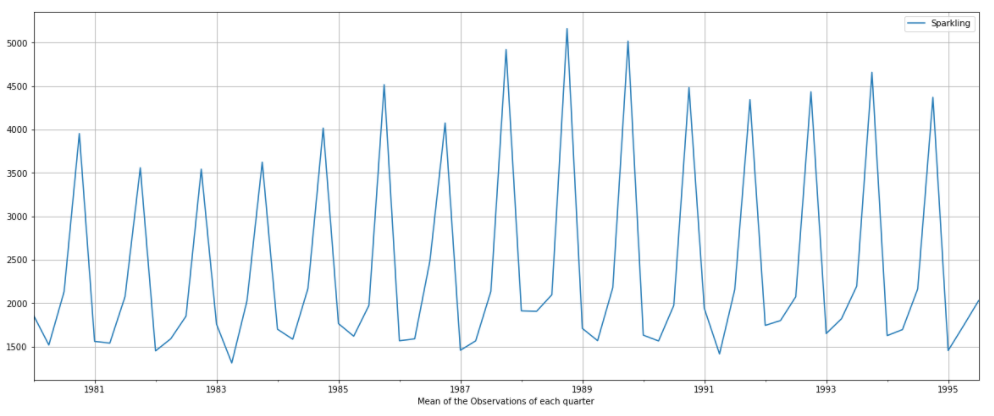
* This plot shows the yearly changes of each month.
* There seems to be different variation for each month over the years. No specific pattern is visible.

**Resample the data to obtain yearly plot:**

****

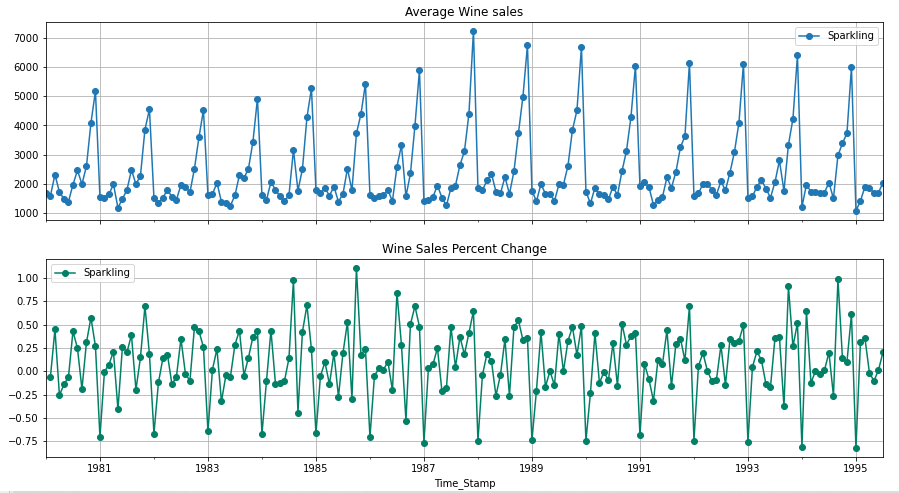
Averaging the data over the years, shows the trend present. As we can see there is an initial dip in the sales for the year 1982, which then gradually increased till 1988 and again decreased. Hence, we can conclude that there is no significant increase/decrease in the trend for the sparkling wine sales.

**Resample the data to obtain quarterly plot:**

****

This plot gives an idea about the seasonality pattern, such as sales are always low at the beginning of the year and high at the end of the year. Also, there are no multiple peaks within each year. Seasonality can be additive in nature since we are not seeing a fractional change in the seasonality over the years.

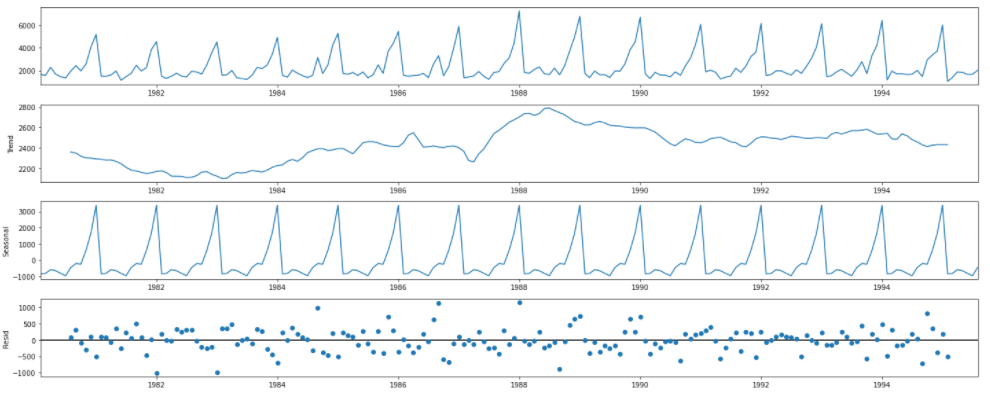
**Percentage change in wine sales:**



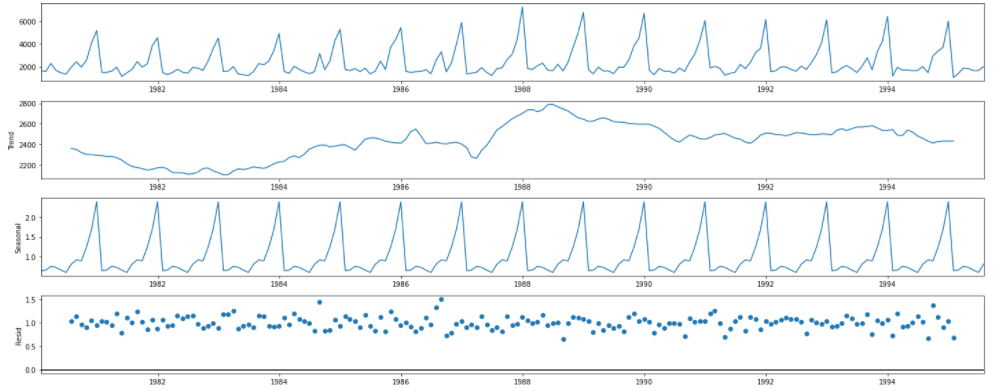
Instead of looking at the sales of wine each month, percentage change of wine sales over the years shows a great deal of variation, which indicates the non-stationarity of the time series.

**Decomposition of the time series:**

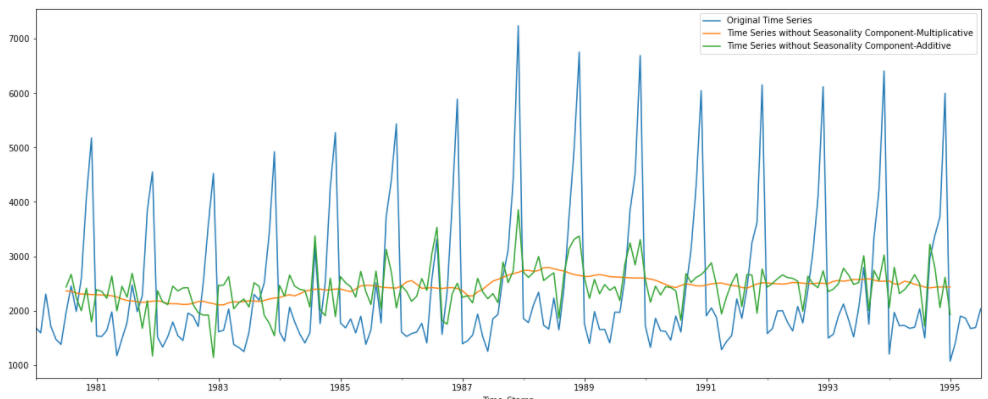
**Additive:**

****

**Multiplicative:**

****

**Plot for both the models trend+residual (without seasonality)**

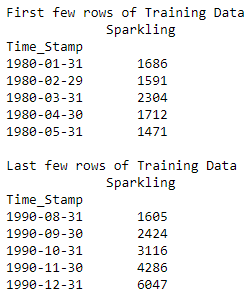
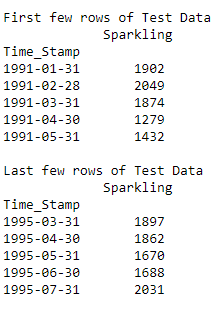


Looking at the above plot, we can infer that multiplicative model’s seasonality component is able to capture all the seasonal changes whereas additive model’s seasonality component is not capturing all the seasonal changes.

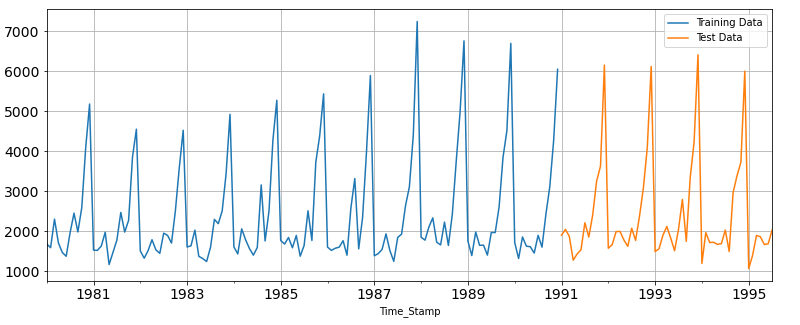
For this particular case study, both the models show more or less the same results. By looking at the previous analysis, we can conclude that the model has no distinguishing trend and seasonality does not change much with trend.

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

**Training data Test Data**

** **

**Plot of train and test data:**

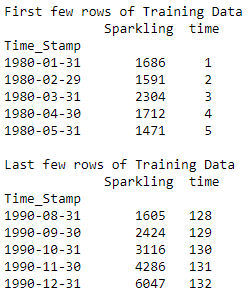
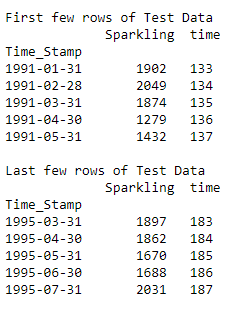
****

* 1. **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.**

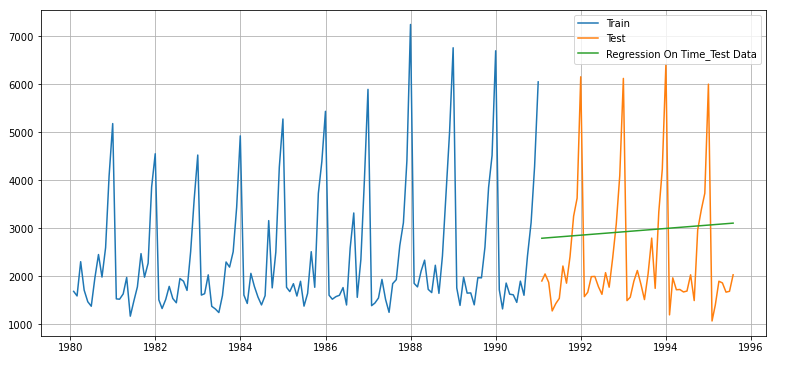
1. **Linear Regression model:**

For this model we need one independent variable and dependent variable is Sparkling sales. Creating a column for time instances in both train and test dataset and proceeding with building linear model.

Training and test dataset:

**Plot of train, test and the predicted values from the linear model:**



Linear model gives an output which captures the slight linear trend in the data but is not including the seasonality over the months.

**Performance metrics:**

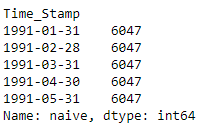


This indicates that the model was able to forecast the average monthly sparkling wine sales in the test set within 1389.135 units of the real sales.

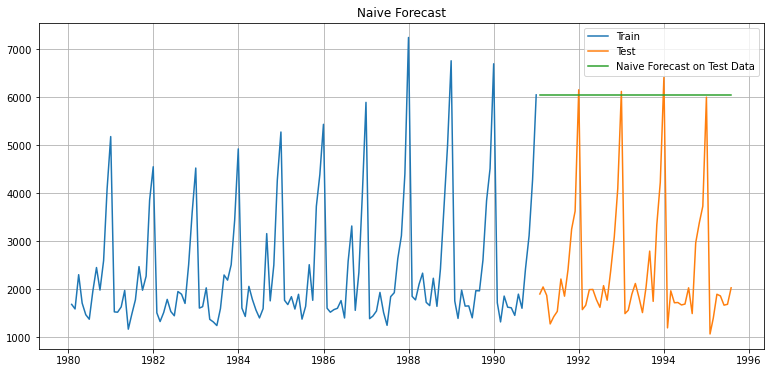
1. **Naïve Based model:**

This is the basic forecasting model, the premise of which is that the expected point value is equal to the last observed point value.

**Prediction for test data**



**Plot of train, test and the predicted values from the naïve model:**



Same like the linear model, this model gives an output which does not capture the seasonality present in the dataset.

**Performance metrics:**

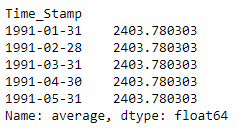


This indicates that the model was able to forecast the average monthly sparkling wine sales in the test set within 3864.279 units of the real sales. This value is pretty large, hence cannot be considered for further model building.

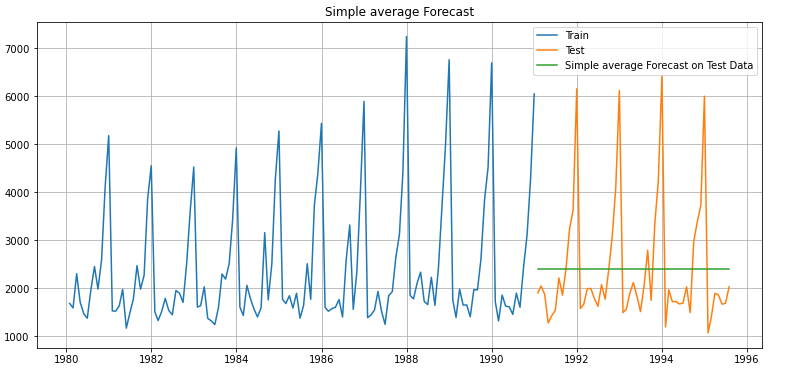
1. **Simple average model:**

In this model, the train data mean becomes the forecast of the test data.

**Prediction for test data**



**Plot of train, test and the predicted values from the simple average model:**



Simple average model gives linear output and no significant seasonality is captured.

**Performance metrics:**

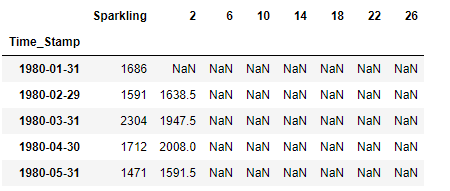


This indicates that the model was able to forecast the average monthly sparkling wine sales in the test set within 1275.082 units of the real sales.

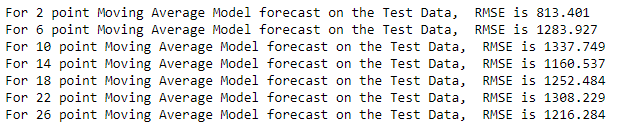
1. **Moving average model for various trailing averages:**

This model is a naïve and effective technique in time series forecasting. Window width is required that defines the number of observations that is used to calculate the moving average value. Here trailing moving average is used which takes the historical observations only in consideration for the forecast.

**Creating a loop for different window width**

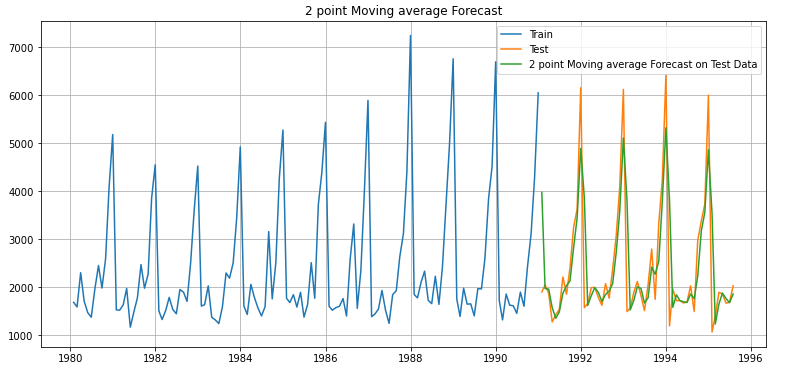
****

**Calculating the RMSE for all the moving averages:**

****

Hence, we can conclude that from among the windows chosen, 2 point moving average gives the least RMSE.

**Plot of train, test and the predicted values from the 2 point moving average model:**

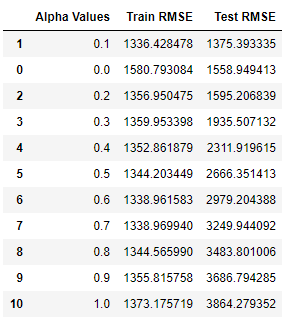


Here we can see that 2-point moving average predictions are almost matching the original test dataset values. This model is able to capture the seasonality in the dataset as well.

1. **Single exponential model:**

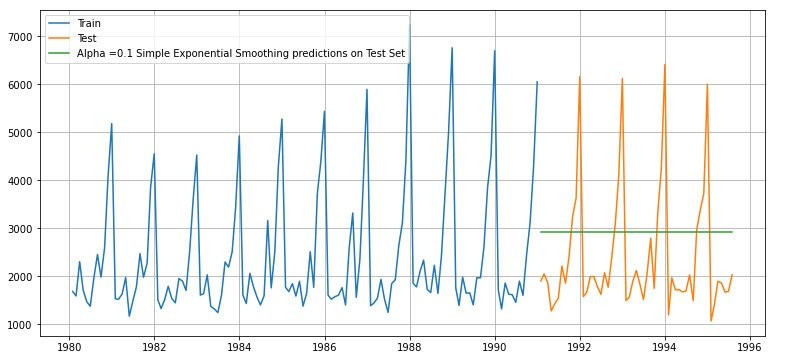
Simplest of exponential smoothing models which is suitable for data with no clear trend or seasonal pattern. It requires single parameter alpha which is the smoothing parameter for level.

Loop is created for different alpha values from 0 to 1 and corresponding RMSE values are calculated.



We can see that for alpha = 0.1, test RMSE is the lowest, which indicates that more of history is taken into account when making a prediction.

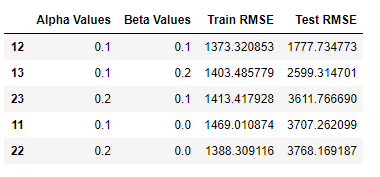
**Plot of train, test and the predicted values from the Alpha=0.1, SES model:**



1. **Double Exponential smoothing (Holt method):**

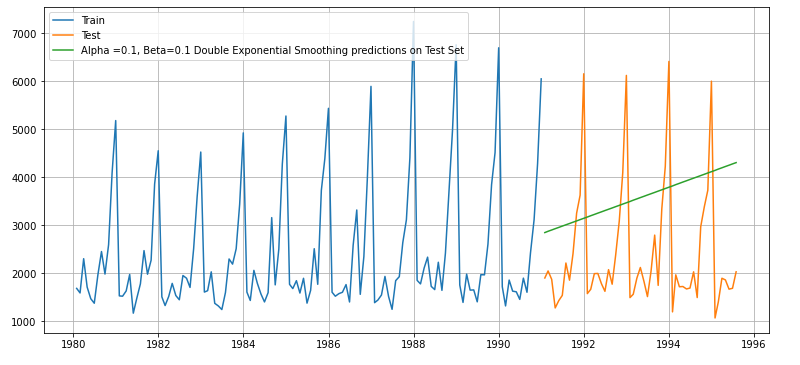
Double Exponential Smoothing is an extension to Exponential Smoothing that explicitly adds support for trends in the univariate time series. In addition to the alpha parameter for controlling smoothing factor for the level, an additional smoothing factor is added to control the decay of the influence of the change in trend called beta.

Loop is created for different alpha and beta values from 0 to 1 and corresponding RMSE values are calculated.

****

For alpha = 0.1 and beta = 0.1, test RMSE is the lowest. Smaller values of beta indicate we are extrapolating only the recent trend.

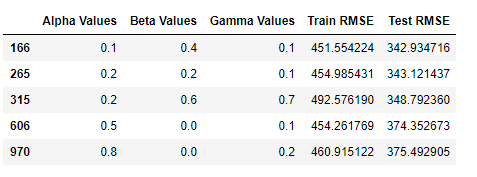
**Plot of train, test and the predicted values from the Alpha=0.1, Beta=0.1, DES model:**



1. **Triple Exponential model (Holt-Winters method):**

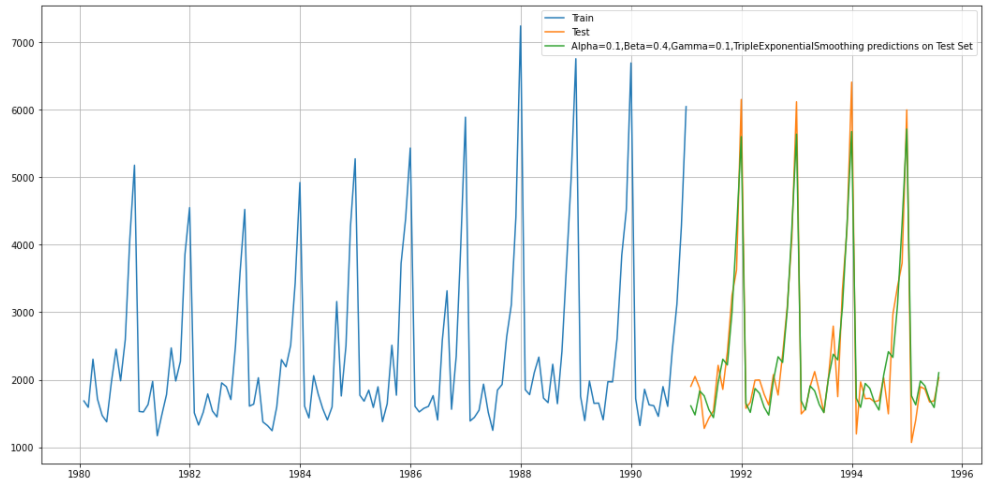
Triple Exponential Smoothing is an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series. In addition to the alpha and beta smoothing factors, a new parameter is added called gamma that controls the influence on the seasonal component. As with the trend, the seasonality may be modelled as either an additive or multiplicative process for a linear or exponential change in the seasonality.

Loop is created for different alpha, beta and gamma values from 0 to 1 and corresponding RMSE values are calculated.



For alpha = 0.1, beta = 0.4 and gamma = 0.1, test RMSE is the lowest.

**Plot of train, test and the predicted values from the Alpha=0.1, Beta=0.4, Gamma=0.1 TES model:**



From all the models built till now, Triple exponential model gives the least RMSE comparatively. This model was able to forecast the average monthly sparkling wine sales in the test set within 342.934 units of the real sales. And the graph of predicted values follows the original values to an extent.

* 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.**

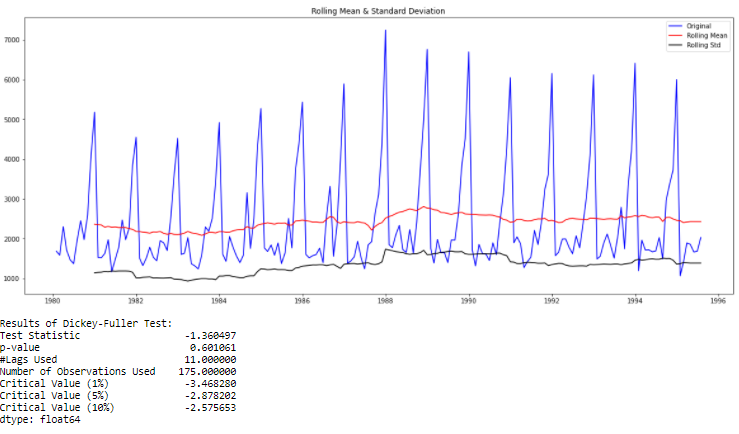
**Stationarity** is the property of exhibiting constant statistical properties (mean, variance and autocorrelation). If there is an increase in time series, then it is not stationary. Seasonal components can be removed by subtracting periodical values.

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

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

* Ho: 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.

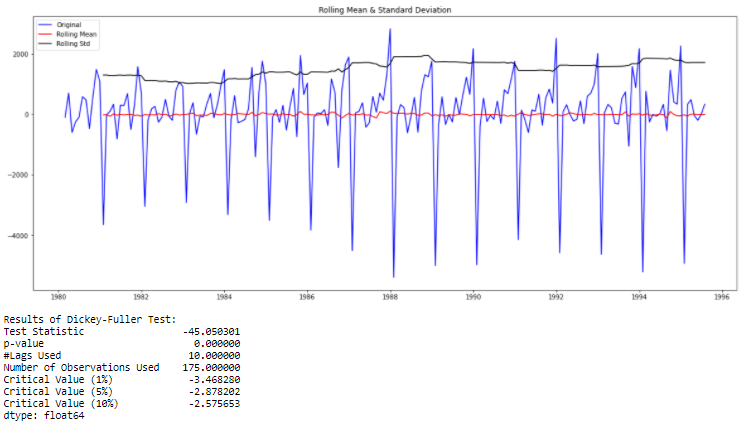
**Stationarity of the original time series:**

****

p-value (0.60) is greater than the alpha (0.05). Hence at 95% confidence, we are not rejecting the Null hypothesis. Hence the time series has a unit root and is thus not stationary.

In order to make the series stationary, first order differencing is taken when removes the linear trend in the series.

**Stationarity of the first order differenced time series:**

****

Here the p-value (0.0) is less than alpha (0.05). Therefore, we are rejecting the null hypothesis and accepting the alternative hypothesis which states that the series is stationary if we take first order differencing.

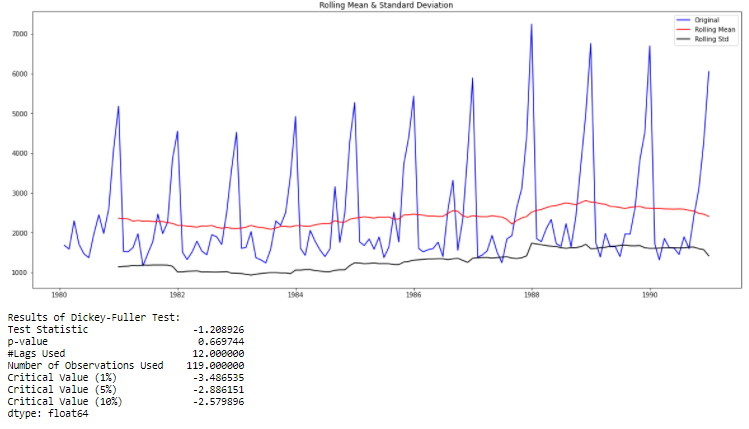
* 1. **Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

1. **ARIMA**

**ARIMA** is the most commonly used method for time-series forecasting, which stands for Autoregressive Integrated Moving Average model. They are denoted with the notation ARIMA (p, d, q) parameters which account for the seasonality, trend and noise in the data.

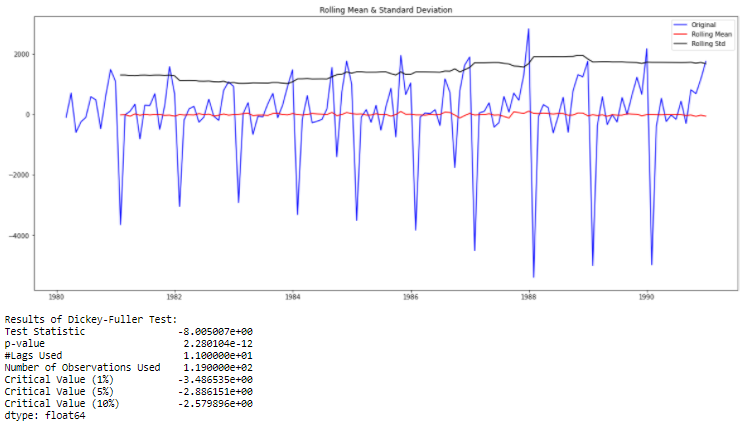
* **AR**: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations. This is used to determine the order to p which is the number of lag observations included in the model.
* **I**: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary. This is denoted by d which is the number of times the raw observations are difference, also called degree of differencing.
* **MA**: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. This is used to determine q which is the size of the moving average window.

**Testing the stationarity of the train dataset:**

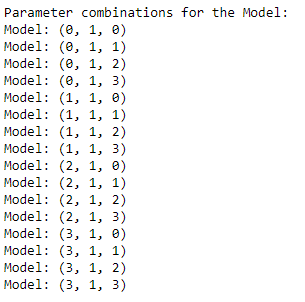
****

Train series is not stationary since the p-value is greater than alpha. Hence taking the first order differencing which means for further building the model, we will take the value of d as 1.

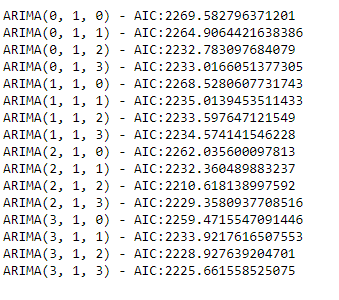
**First order differenced train time series:**

****

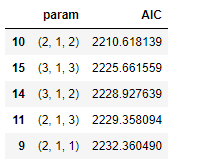
Created a loop to take different combinations of p and q values in the range (0, 4) and d as 1.



Building the model for the above parameter combinations and AIC values of each are compared. Akaike Information Criterion (AIC) measures the loss of information. Lower the value of AIC, better the model.

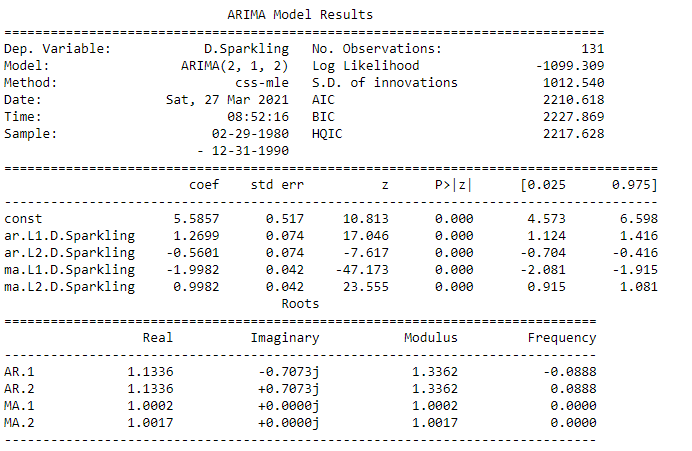


Sorting the AIC in ascending order, to get the optimum model.



Model is built with (p, d, q) parameters as (2, 1, 2).

**Model Summary:**

****

We can see that all the terms have a p-value of 0.0 which indicates the significance of AR and MA lags used. AR.L1, MA.L1 and MA.L2 are the more significant terms based on the test-statistic.

**Performance Metric:**

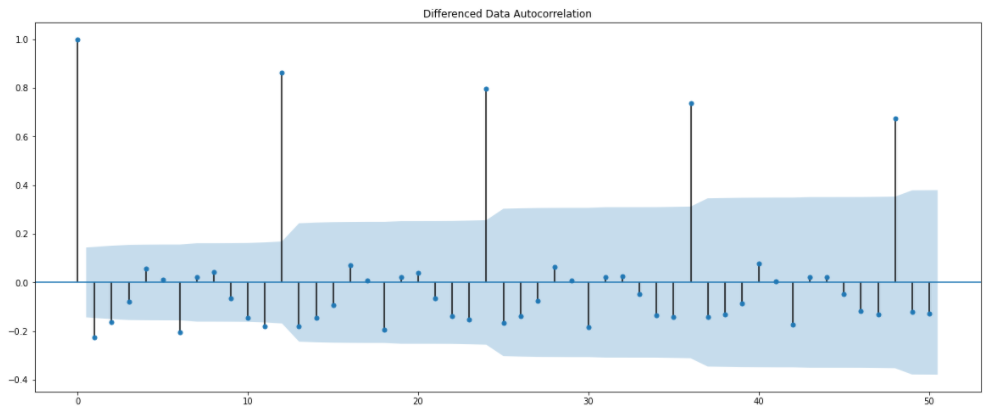
****

1. **SARIMA:**

Since the original dataset has a seasonal component, ARIMA model can be extended to SARIMA model to include the seasonality. We have extra parameters (P, D, Q, s) for the seasonal components.

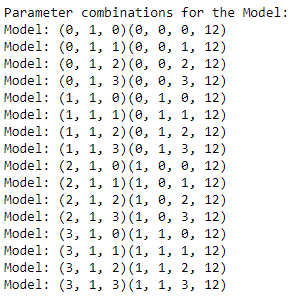
* Seasonal P indicates number of autoregressive terms, lags of the stationarized series.
* Seasonal D indicates differencing that must be done to stationarize series for the seasonality.
* Seasonal Q indicates number of moving average terms, lags of the forecast errors.
* S indicates the seasonal length of the data

**Check for seasonality in the difference auto correlation plot:**

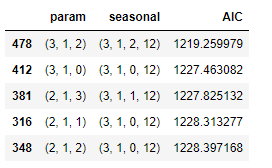
****

We can see regular positive spikes in an interval of 12, indicating monthly seasonality.

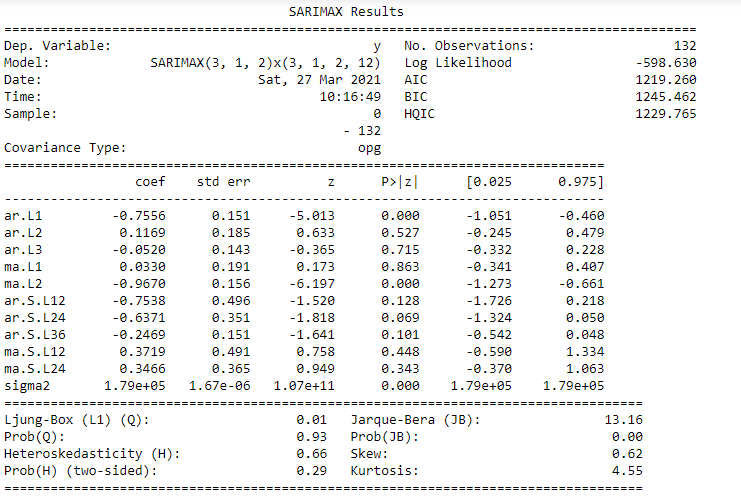
Created a loop to take different combinations of p (P) and q (Q) values in the range (0, 4) and d (D) as 1.

****

**Built the model with parameters which had the least AIC:**

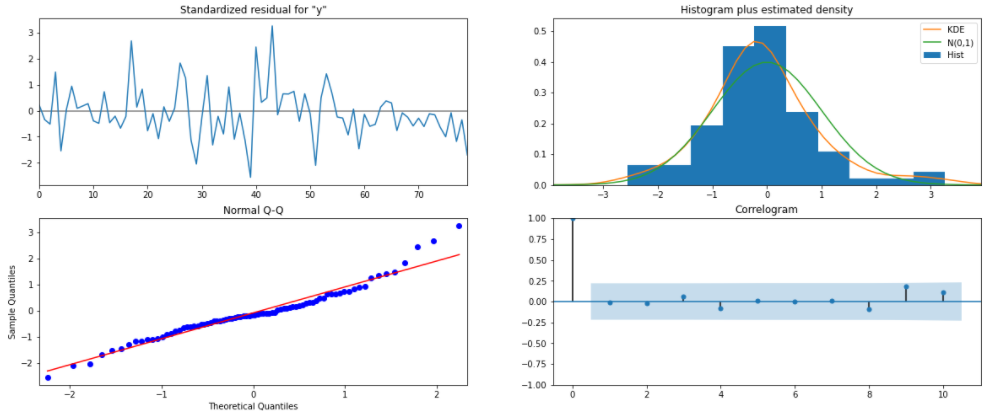
****

**Summary of SARIMA model**

****

There are some AR and MA lag points which are not significant because the p-value is greater than alpha.

**Plot Diagnostics:**

****

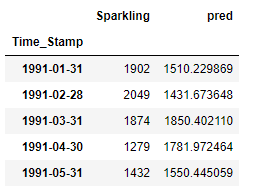
* Jarque-Bera test for the normality of the data. Prob(JB) is less than alpha, therefore rejecting the null hypothesis, residuals are normally distributed.
* Skewness and Kurtosis are little outside the standard range.
* LB test has a probability of 0.93 which is greater than alpha. Residuals are independent and homoscedastic in nature.

**Performance metrics**

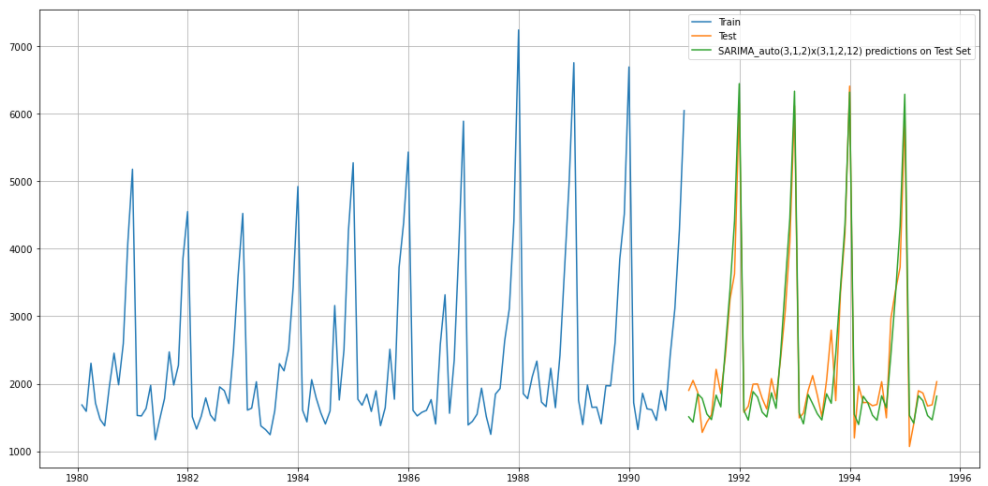
****

**Of the two models built,** SARIMA model has the least RMSE since it includes the parameters for seasonality in the data.

Observed and predicted values:



**Plot of train, test and the predicted values from the SARIMA (3, 1, 2) x (3, 1, 2, 12) model:**



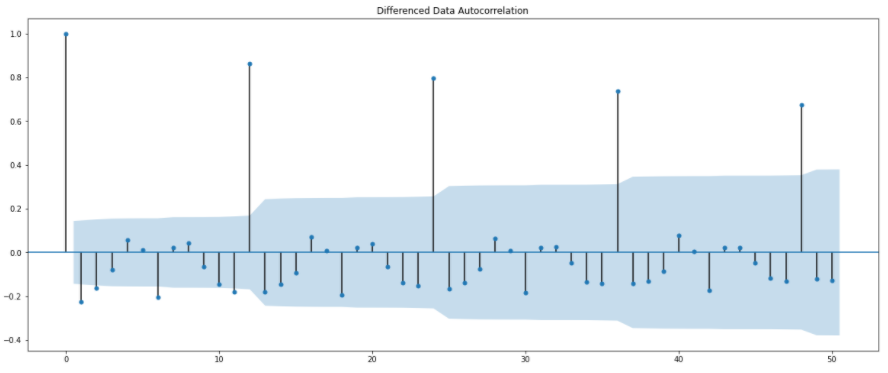
Predicted values are able to closely follow the observed value in the SARIMA model.

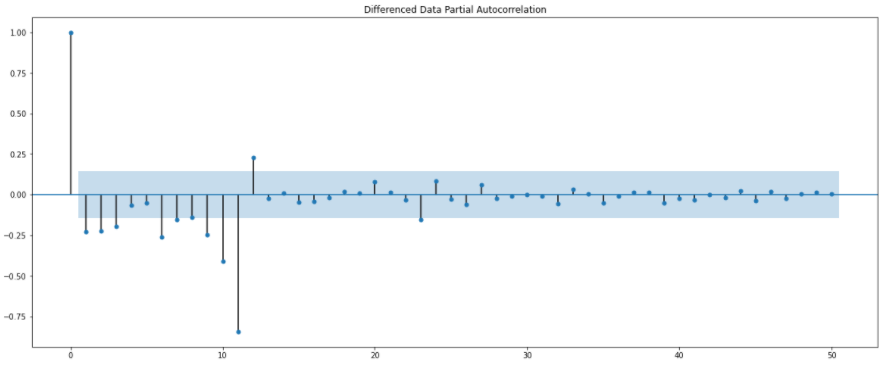
* 1. **Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

1. **ARIMA**

As we have already seen that data is stationary only after first order differencing, we take d as 1. For identifying p and q parameters plotting ACF and PACF plot for the differenced time series.

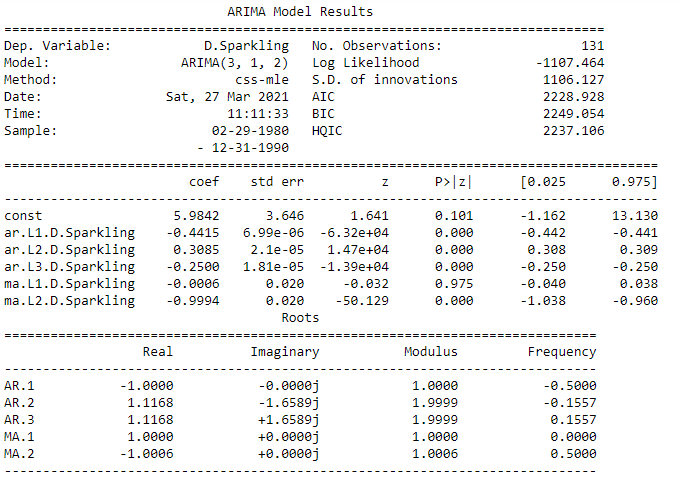
* p is the point in PACF plot beyond which the partial autocorrelation dies down.
* q is the point in ACF plot beyond which the correlation dies down.





From the above plots, we chose p as 3 and q as 2 for further model building.

**Model Summary:**

****

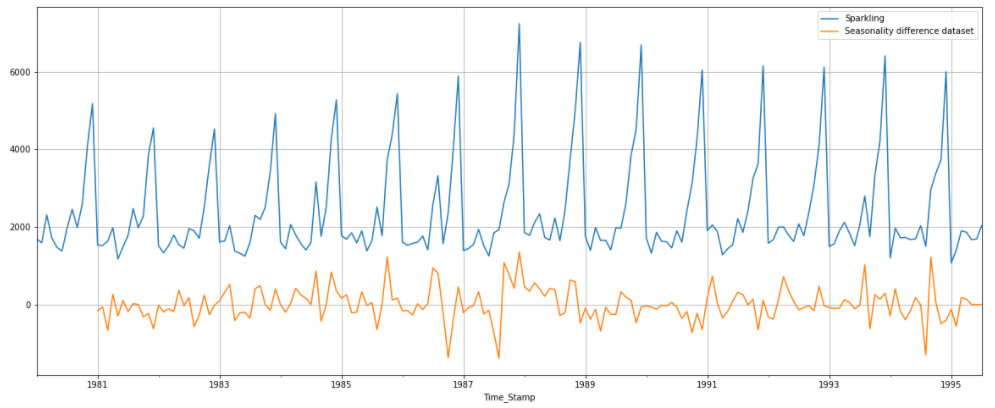
Except MA.L1 term others are significant.

**Performance metrics:**

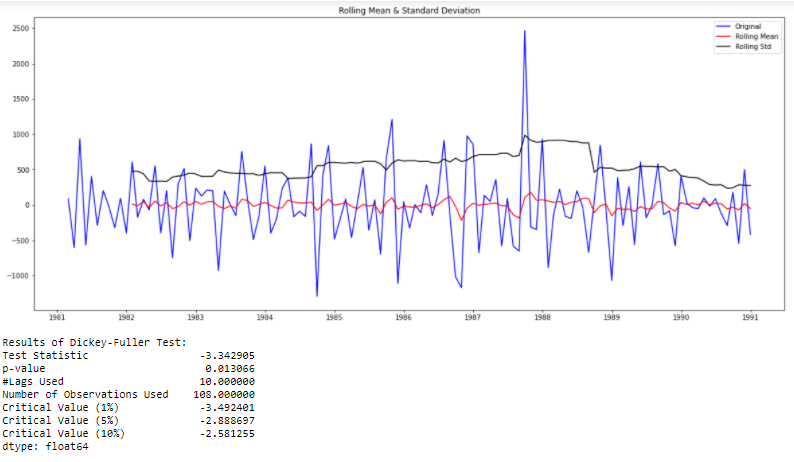
****

1. **SARIMA**

In order derive the seasonal P, D, Q, s for the SARIMA model, first we are checking the plot of original data and seasonal differenced data (12 in this case)

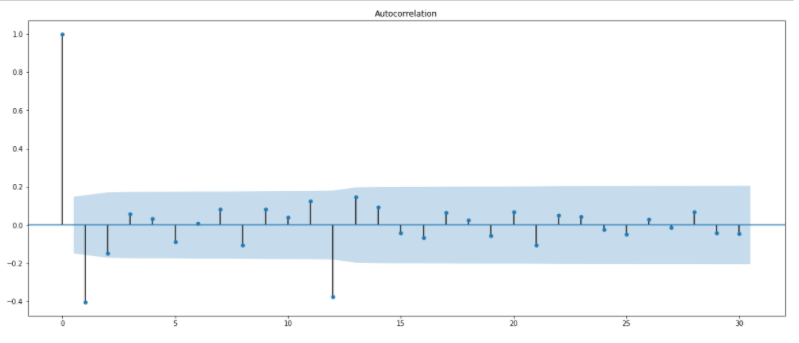


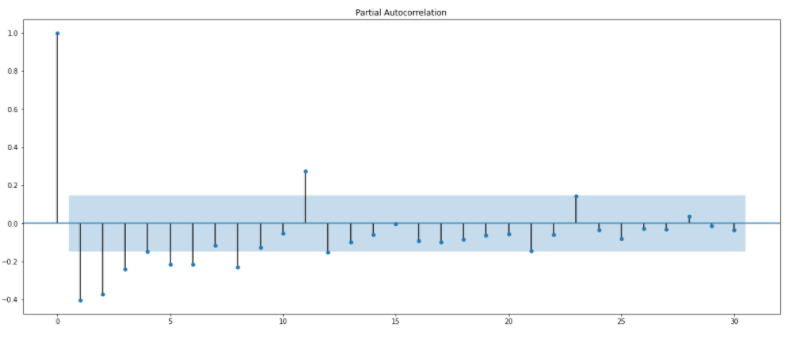
There is still some trend and seasonality shown by the seasonality differenced dataset. Hence taking the first order differencing for that and checking for the stationarity.



p-value is less than alpha, hence the series is stationary for first order differenced and seasonality differenced dataset.

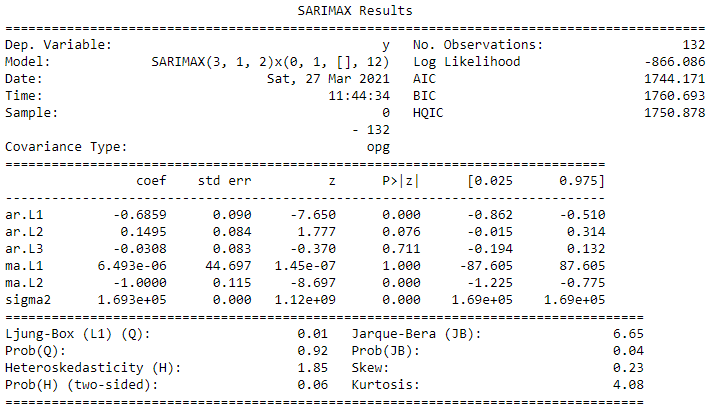
**ACF and PACF plot:**

****

****

Since in both the plots, there are no significant positive spikes and both plots decay is more gradual. We are taking a value of (0, 0) for (P, Q).

**Building the SARIMA model with (3, 1, 2) x (0, 1, 0, 12) as parameters:**



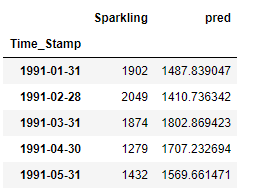
Significant terms are AR.L1 (lag 1) and MA.L2 (lag 2).

**Performance metrics:**

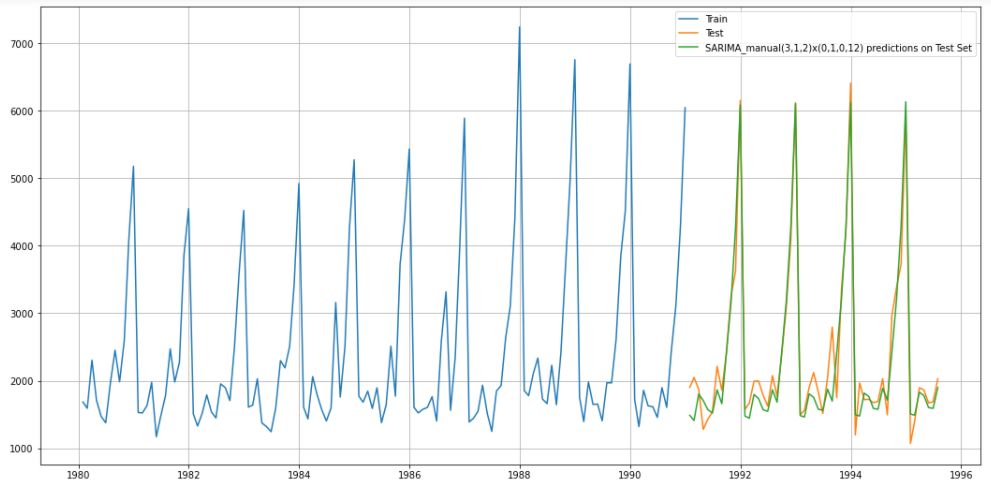
****

**Comparing the above two models, SARIMA with (3, 1, 2) x (0, 1, 0, 12) as parameters gives the optimum model.**

Observed and predicted values:

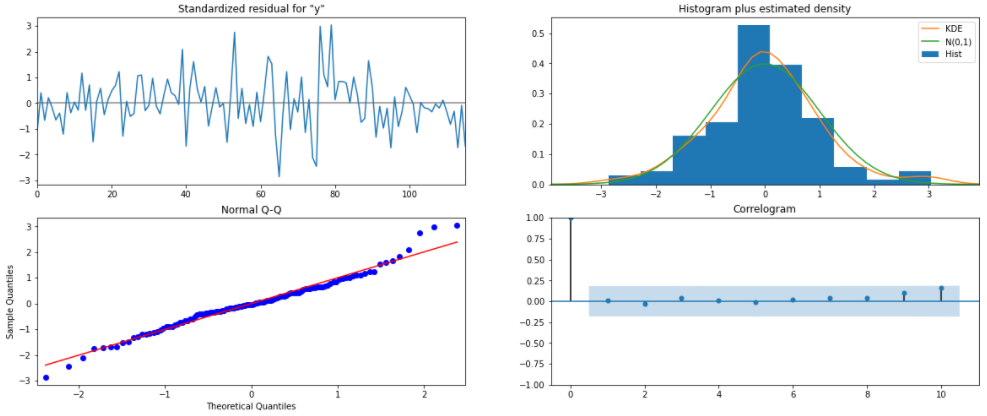


**Plot of train, test and the predicted values from the SARIMA (3, 1, 2) x (0, 1, 0, 12) model:**

****

This model predicted value is able to closely follow all the peaks of the observed value.

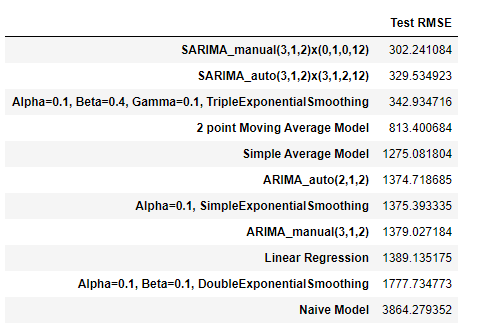
**Plot Diagnostics:**

****

* From the quantiles plot, we can see some deviations at the beginning and end of the data.
* Residuals follow normal distribution to an extent.
* There is almost no skewness in the residuals plot.
  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.**

In order to understand the performance of each model, RMSE is used which is the root mean squared error. The benefit of RMSE is that it penalizes large errors and the scores are in the same unit as the forecast values.

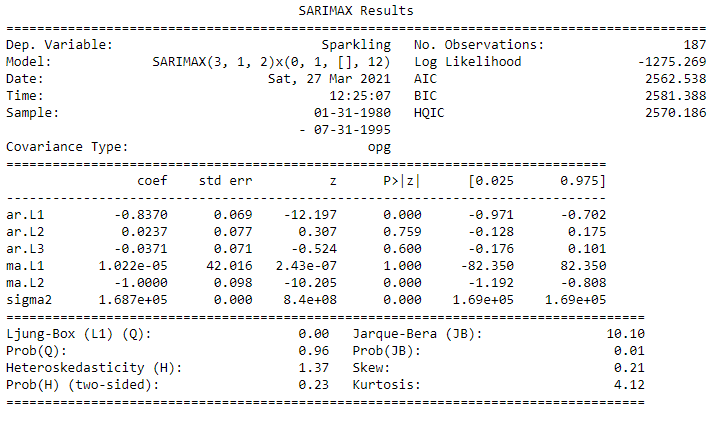
Dataframe is created at the end of each model and their corresponding RMSE values for the test data are stored.



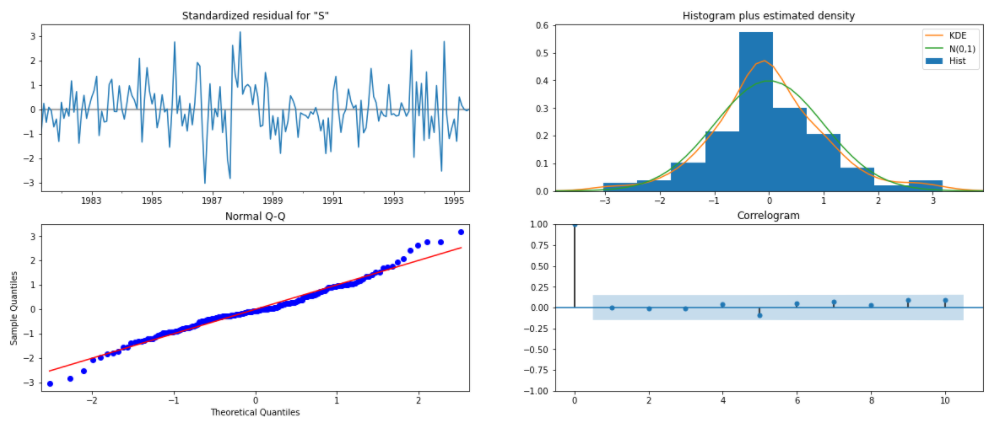
Best optimum model obtained is **SARIMA\_manual (3, 1, 2) x (0, 1, 0, 12)** because of the lowest test RMSE value compared to other models.

* 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.**

From the best optimum model obtained, building the model on the whole dataset.

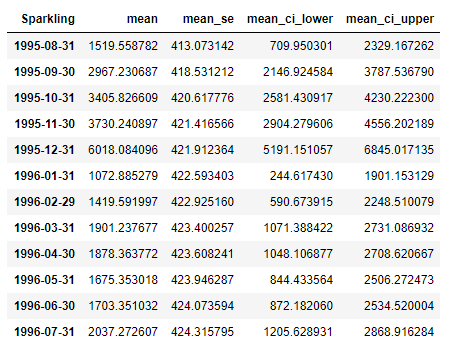


**Plot diagnostics for the full model:**



**Prediction for the next 12 months:**

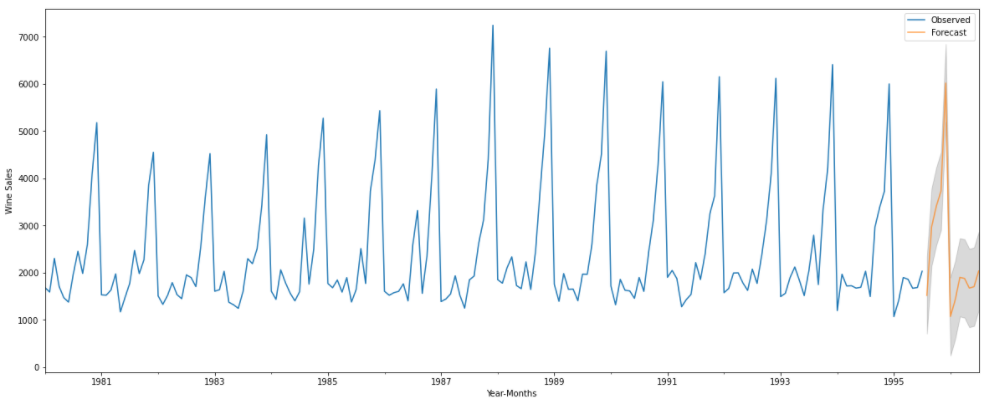
* We are getting the predicted mean for the next 12 months with the standard error.
* Upper and lower confidence bands at 95% confidence interval (+/- 1.96 standard deviations)
* Residuals has a behaviour closer to the normal distribution.

****

**Performance Metrics:**



**Plot of the observed and forecasted values:**

****

* Orange line indicates the forecasted values and the grey region surrounding is confidence interval within which the final values can range.
* We can say that SARIMA model was able to capture both trend and seasonality correctly. Since the forecasted values follow the same pattern as the given dataset with peaks at year end.

Hence, we can conclude that **SARIMA\_manual (3, 1, 2) x (0, 1, 0, 12)** is the best optimum model with an assumption of RMSE value 575.94.

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

The **objective** of this case study is to analyse and forecast wine sales for next 12 months from the end date in the dataset.

**Initial Analysis of the data:**

* We can observe several characteristics of the series: first, that it maintains a slightly similar behaviour, without radical changes in the trend; Secondly, it presents a strong seasonal behaviour, which explains these ups and downs, far from the average of the series.
* No outliers as the peaks are fully described by the seasonality present in the dataset.
* Additive model decomposition is preferred because the seasonality does not change over time, this can be confirmed by the quarterly plot as seen before.
* There was a dip in wines sales till the year 1982 which then took an incremental change till 1988.
* First order differencing was needed to make the series stationary.

**Model building findings:**

Taking into consideration the initial analysis done,several models were built like Linear regression, Naïve model, Simple average, moving average with different rolling windows, various exponential models and ARIMA/SARIMA.

Out of the models built, Naïve forecast is the worst in terms of seasonal forecasting followed by double exponential smoothing and linear regression models. SARIMA outperforms all other models, in that manually selected parameters for SARIMA model gave the best results in terms of RMSE. The baseline prediction for the dataset resulted in an RMSE of 302.24. The chosen model forecasted the next 12 months with an RMSE of 575. At 95% confidence, range of the sales is also predicted.

We have considered only sales of the wine for that particular month, if there were other variables like location, outlet type and customers age could also be some important factors affecting the sales which can help us better in forecasting.

**Business recommendations:**

From the insights gathered during the analysis, wine sales see its highest in the autumn with demand increasing till December and drops off after holidays. This can be due to the weather conditions; high temperature can cause spoilage. Therefore, selling or shipping on wine in such conditions can shut its flavour development down. Taking advantage of the weather in the second half of the year would be the best way going forward to increase the company’s wine sales.

More than 50% of the sales is in the third and fourth quarter. Since there is no decrease in the trend forecasted for the future months, ABC Estate wines can concentrate on boosting their wine sales in the months of demand. In order to implement this, ABC estate wines can:

* Offering a sale in a smart way like digital shopping offers with free shipping, sale in places which are more often visited by the people.
* Promoting wine gift sets and innovative wine bags for specific occasions since the second half of the year is having more holidays.
* Introducing winery visits.
* Host or join events which can include wine tasting. So that people attending the event can be offered a deal in the wine package.
* Digital media marketing campaigns would definitely improve the sales.

Since, Sparkling wine is known for as a gift during the holiday season especially the last days before New year is a great opportunity for ABC estate wines. By implementing a couple or some of the ideas together from above would be benefitable for the company.