# **Name: Vikram Radhakrishnan**

# **Business Report : Time Series Forecasting**

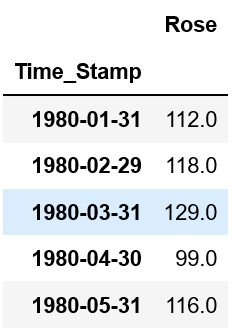
# **Date : 23rd August, 2020**

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

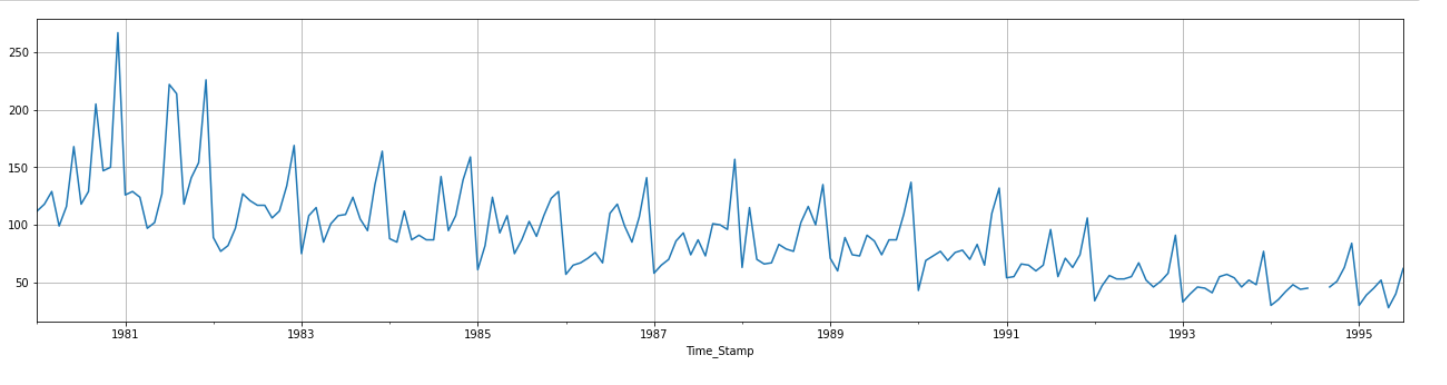
Two csv files have been provided, for which two separate jupyter notebooks have been created to work with each. In order to read them, the following procedure has been followed :

1. The files are read into a Pandas dataframe
2. A separate time series corresponding to the timeline in the data is created, which is called ‘Timestamp’ and added to the dataframe
3. The timestamp column is made into the index and the original ‘YearMonth’ column is dropped.

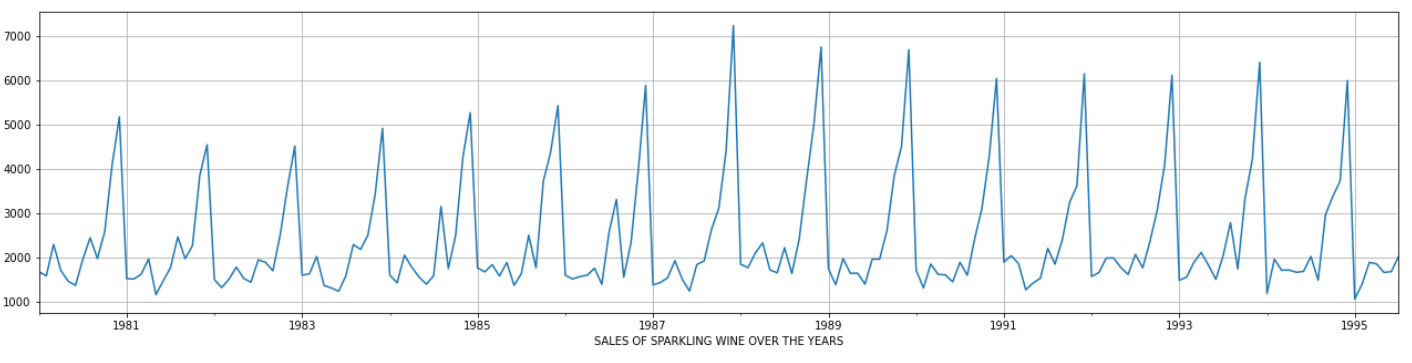
Once this procedure is followed, the format of the data will appear as follows:



**ROSE SALES OVER THE YEARS**

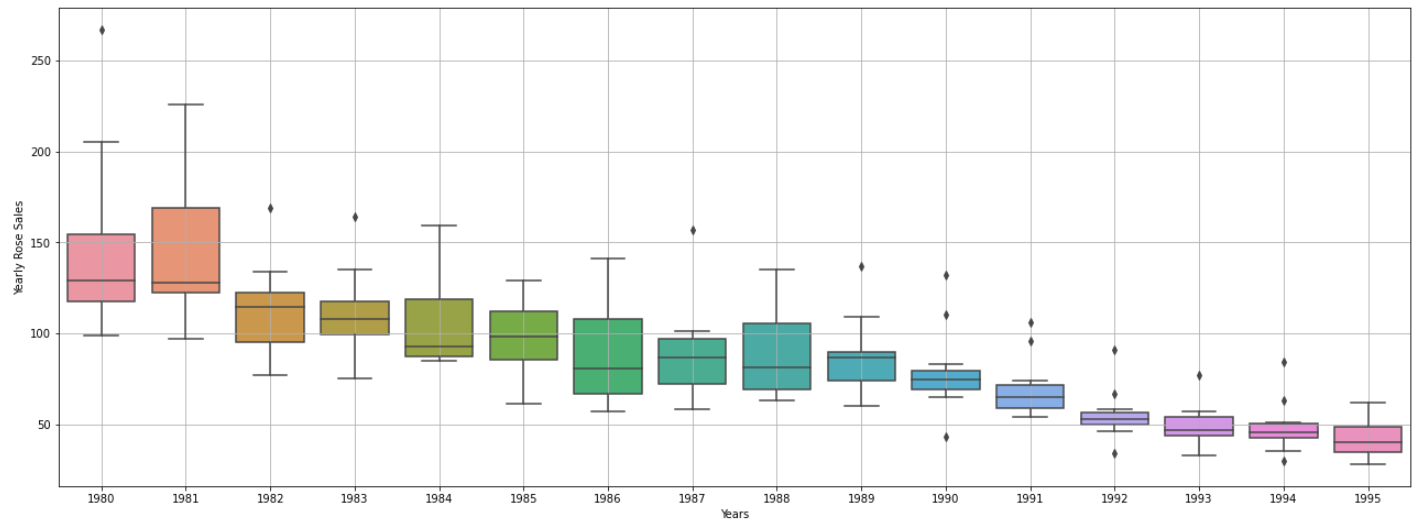


**SPARKLING WINE SALES OVER THE YEARS**

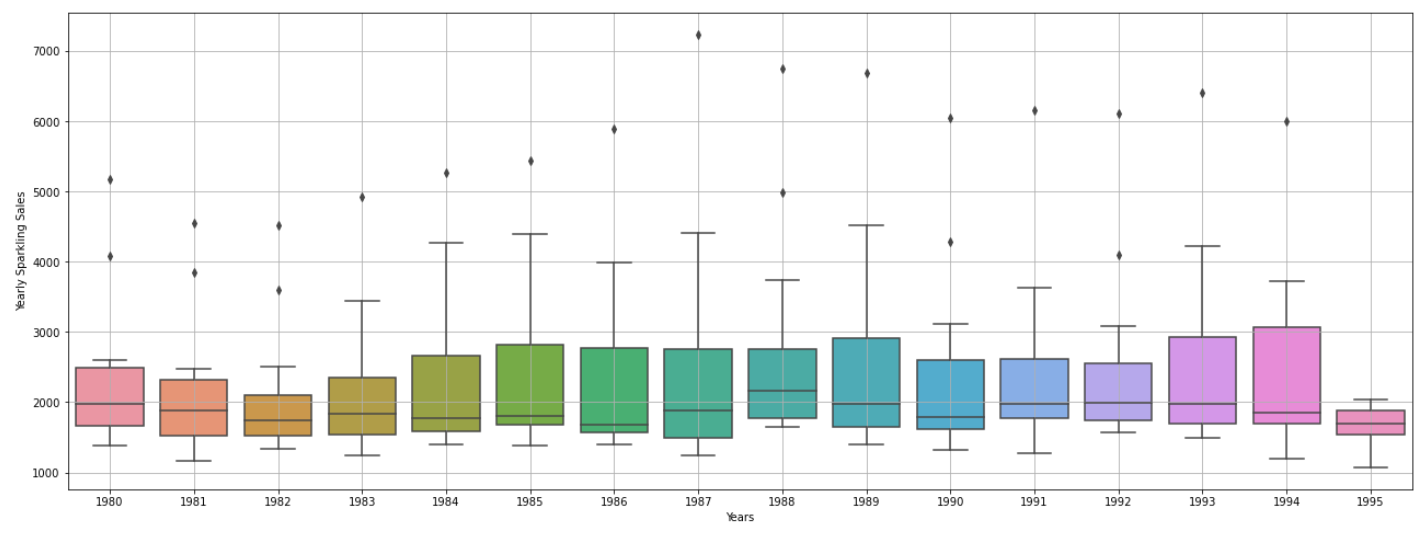


# **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

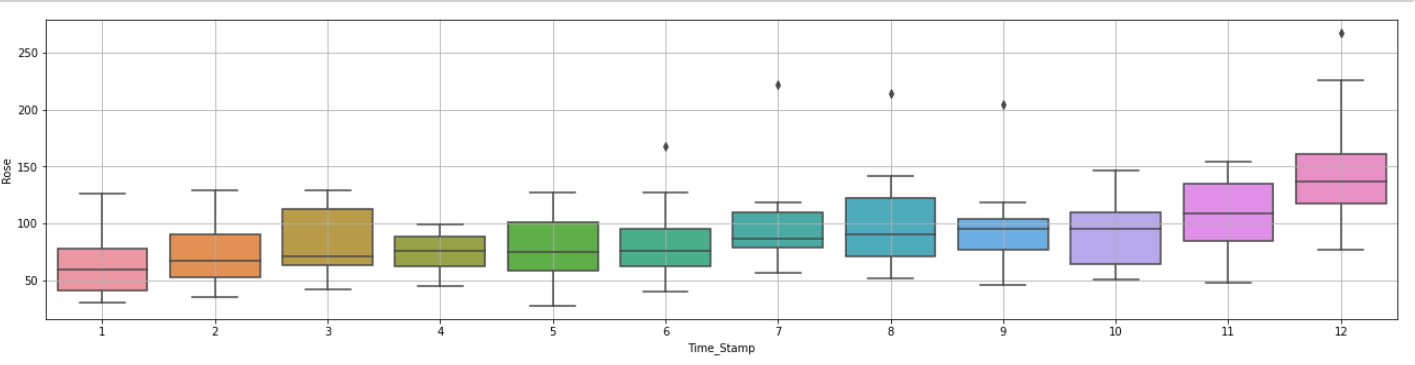
**YEARLY SALES DATA : ROSE (Box Plot)**



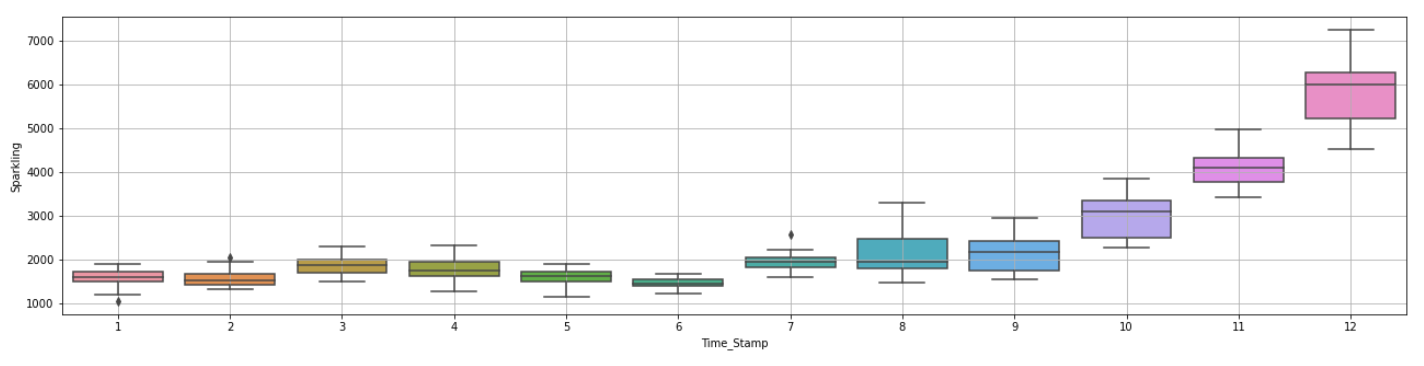
**YEARLY SALES DATA : SPARKLING (Box Plot)**



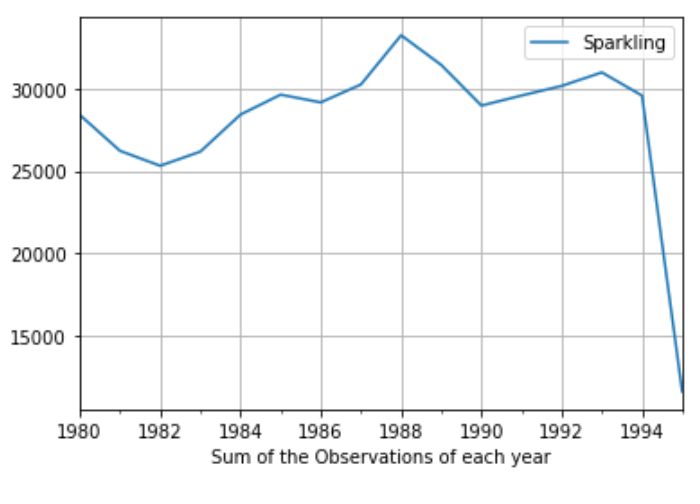
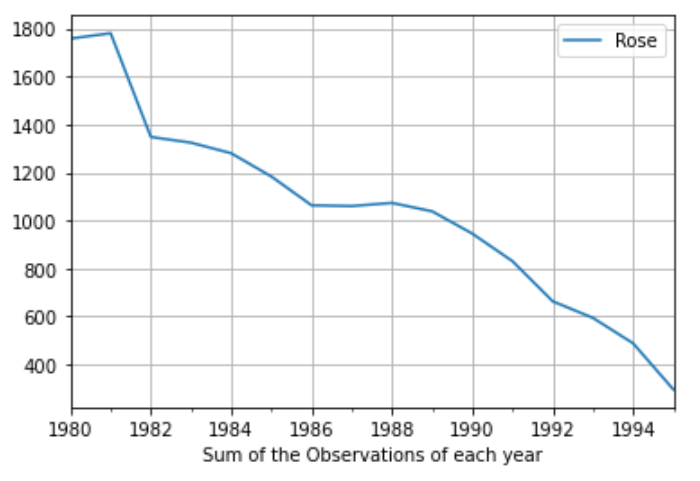
**MONTHLY SALES DATA (MEAN ACROSS YEARS) : ROSE (Box Plot)**



**MONTHLY SALES DATA (MEAN ACROSS YEARS) : SPARKLING (Box Plot)**

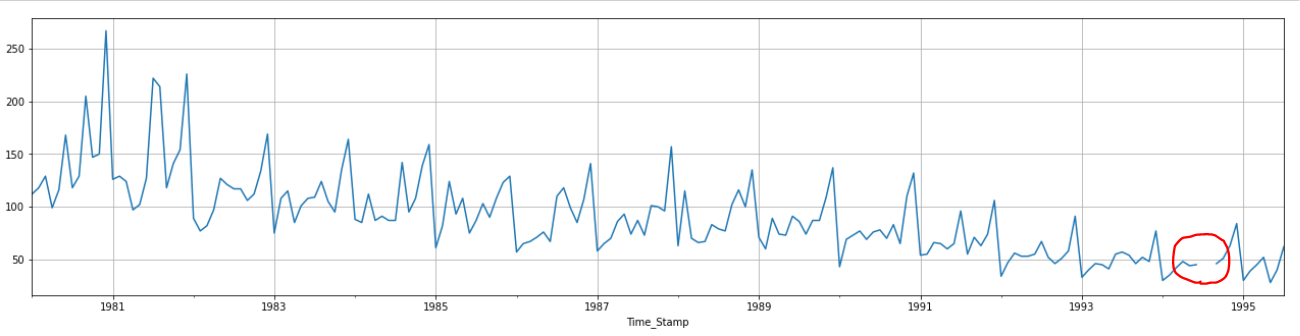


**COMPARISON OF TOTAL SALES PER YEAR – ROSE vs SPARKLING**



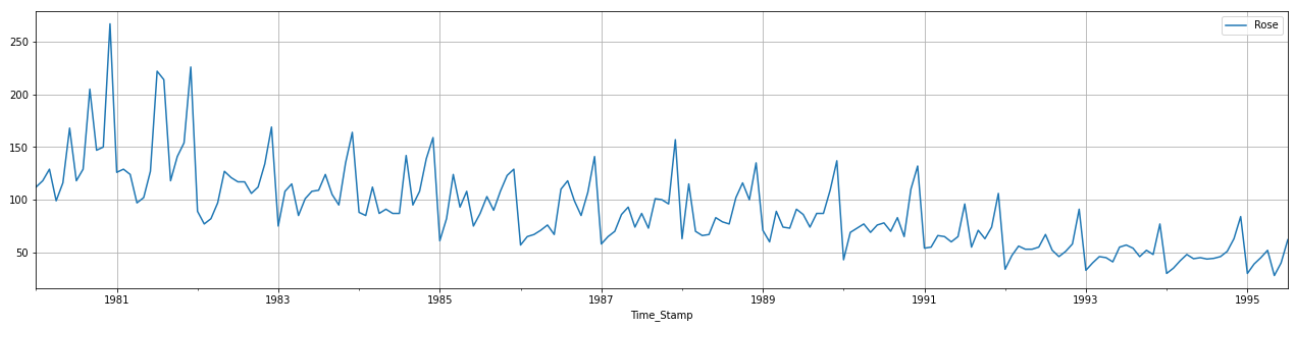
Similarly, many such plots have been checked to observe the general behavior of the data. The interested reader may go through the Jupyter Notebooks for more/different plots using the same data.

**HANDLING MISSING DATA**



It is observed that in the year 1994, two entries are missing for Rose Wine. These are interpolated using a 3rd order spline to estimate the two missing values which are used subsequently.





After populating the missing data, the plot looks continuous as above.

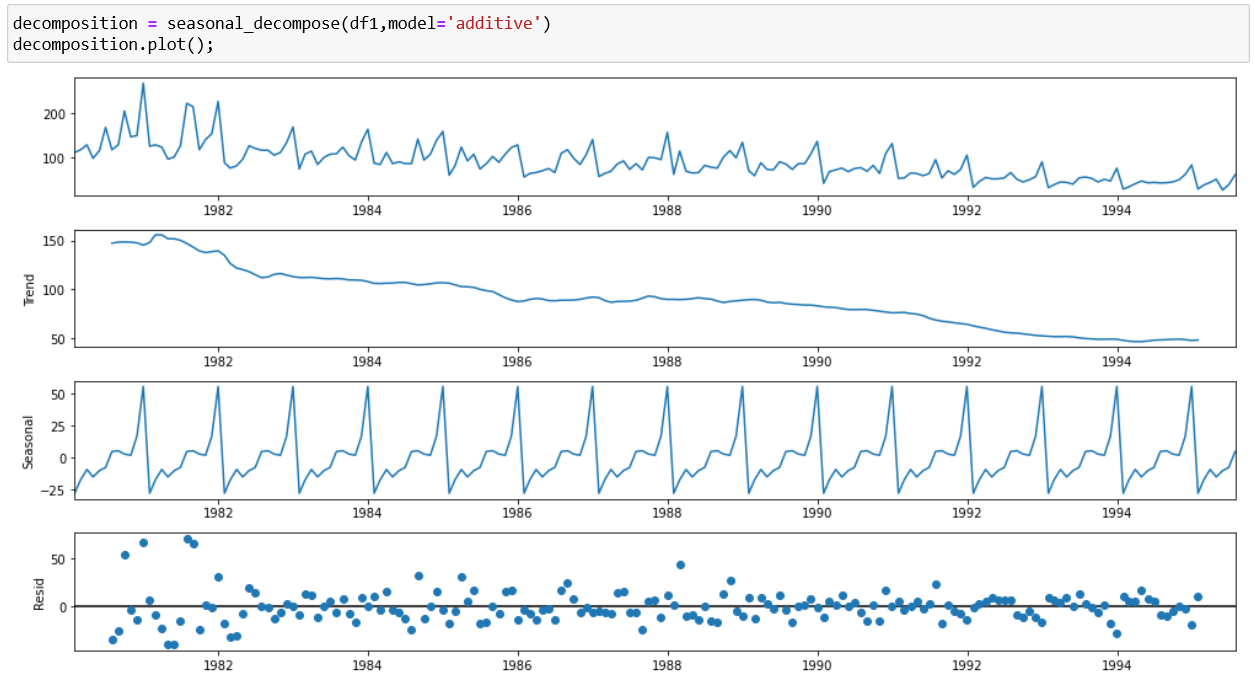
The sparkling wine data does not seem to have any such issues with missing data.

**DECOMPOSITION OF DATA:**

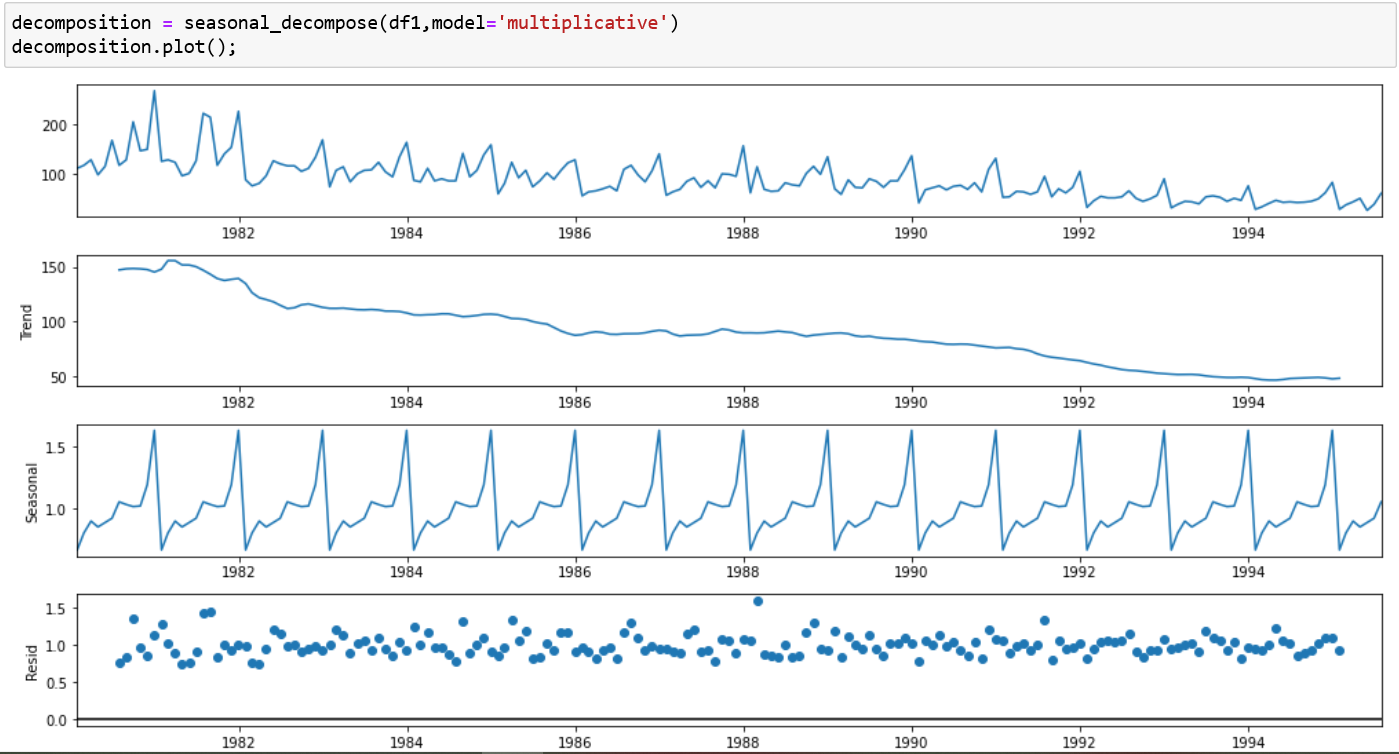
In order to understand the magnitude and nature of TREND and SEASONALITY of the data, decomposition of the data is performed for both Rose as well as Sparkling Wine.

1. **ROSE WINE :** Multiplicative decomposition is used generally when there is a strong increasing or decreasing component of seasonality that changes year-on-year (or as per corresponding observation period). In our case, we do not observe such a behavior in the magnitude of seasonality varying multiplicatively, so ADDITIVE decomposition would be the better choice.

**Additive Decomposition (Rose)**

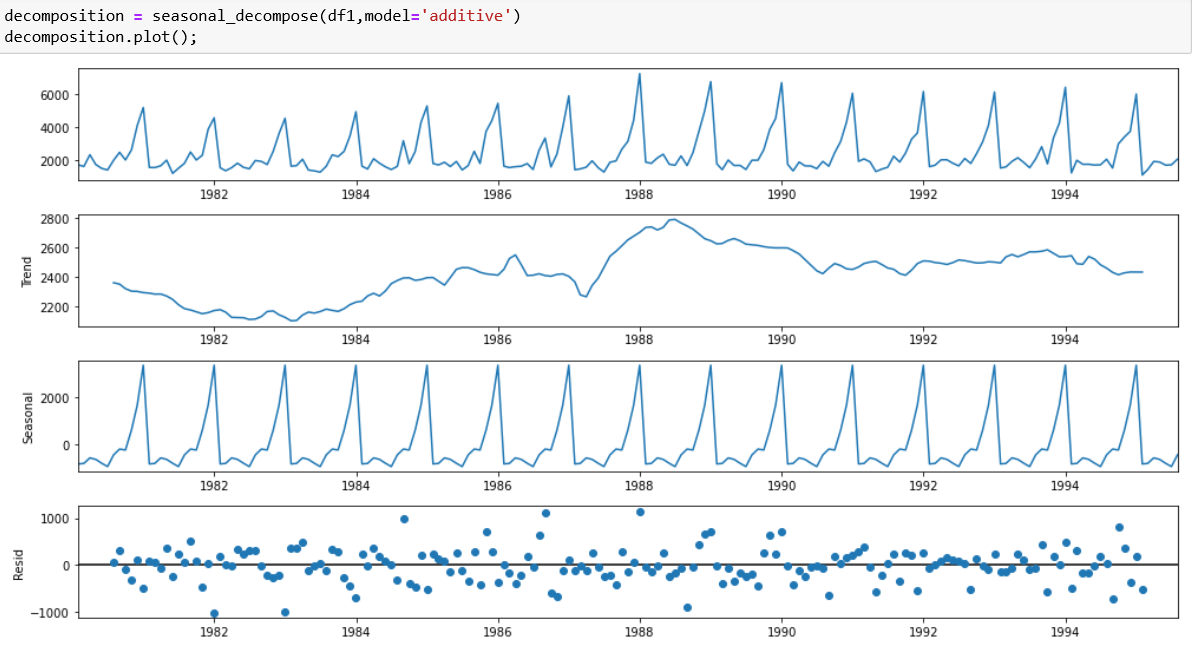


**Multiplicative Decomposition (Rose)**

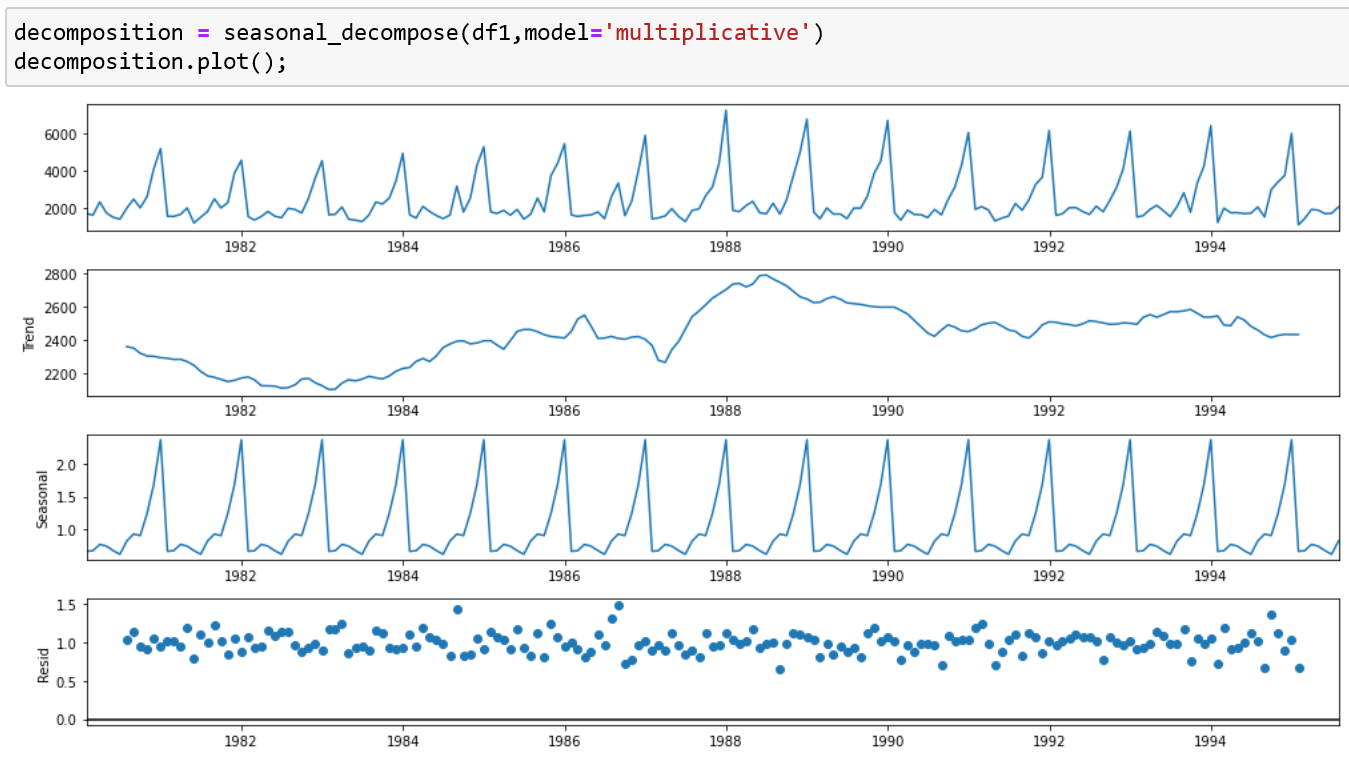


1. **SPARKLING WINE :** Multiplicative decomposition is used generally when there is a strong increasing or decreasing component of seasonality that changes year-on-year (or as per corresponding observation period). In this case, we observe a small trend in the seasonality magnitude and also observe that the residual are more or less constant at a value of approximately 1. So MULTIPLICATIVE decomposition would be the better choice.

**Additive Decomposition (Sparkling)**



**Multiplicative Decomposition (Sparkling)**

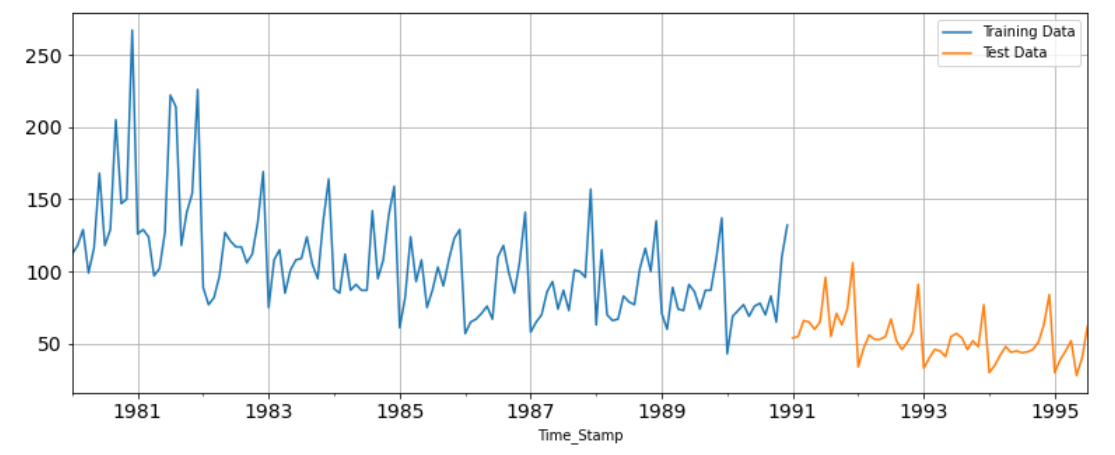


# **Split the data into training and test. The test data should start in 1991.**

The data is split into test and training data based on the year. Years before 1991 are in TRAIN and the year 1991 and hence are in TEST.



**EXAMPLE OF TEST AND TRAINING DATA PLOTTED TOGETHER FOR ROSE WINE**

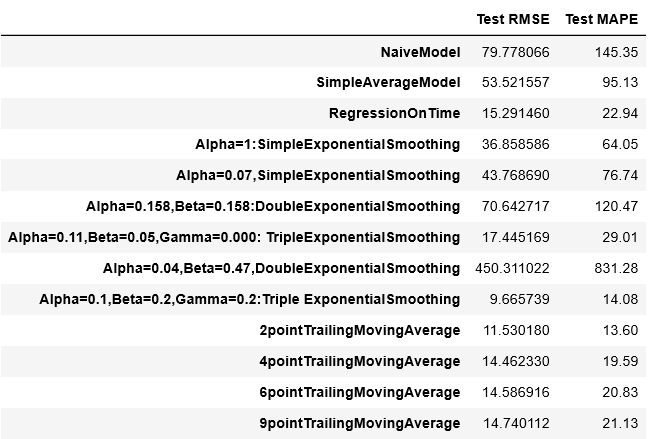


# **Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. - Please do try to build as many models as possible and as many iterations of models as possible with different parameters.**

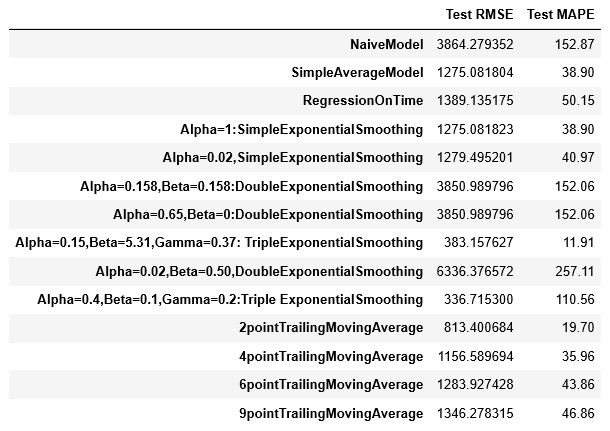
The following are the models tried :

1. Naïve Model
2. Simple Average Model
3. Linear Regression on Time
4. Simple Exponential Smoothing (with Alpha=1)
5. Simple Exponential Smoothing (Iterated for least RMSE)
6. Double Exponential Smoothing (Optimized Fit)
7. Double Exponential Smoothing (Looping through parameters for least RMSE)
8. Triple Exponential Smoothing (Optimal Fit)
9. Triple Exponential Smoothing (Looping through parameters for least RMSE)
10. Moving Averages with 2 point, 4 point, 6 point and 9 point trailing

**ROSE WINE**



**SPARKLING WINE**



# **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.**

The **Augmented Dickey Fuller Test (ADF)** determines whether a time series is non-stationary.

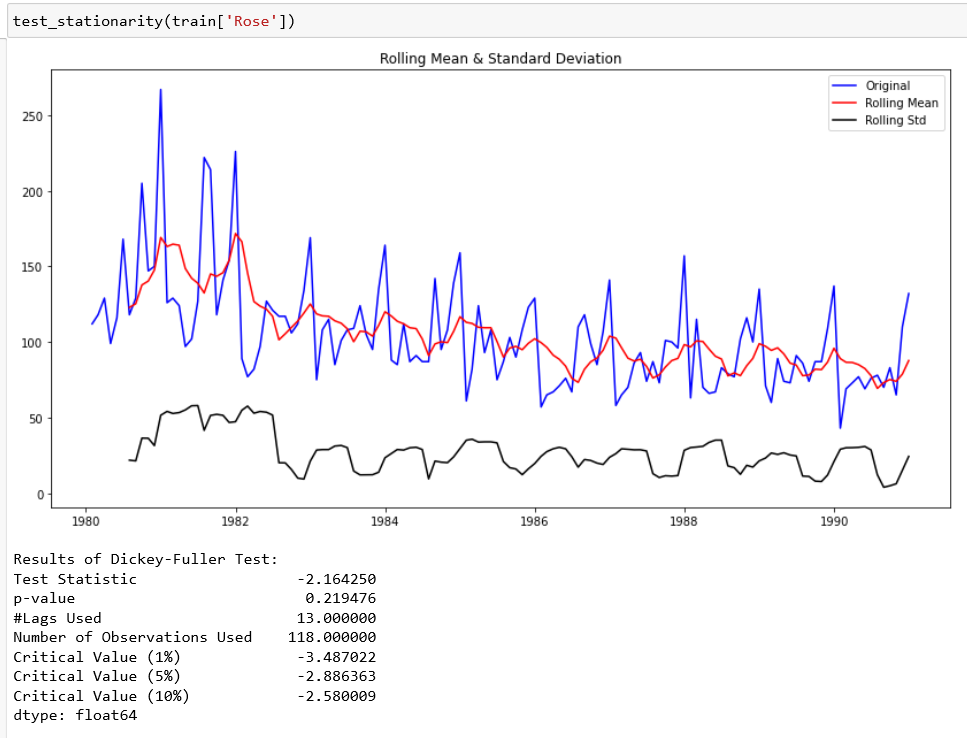
NULL HYPOTHESIS Ho : The time series is non-stationary

ALT HYPOTHESIS H1 : Time series is stationary

(Rejection of Null Hypothesis means time series is stationary)

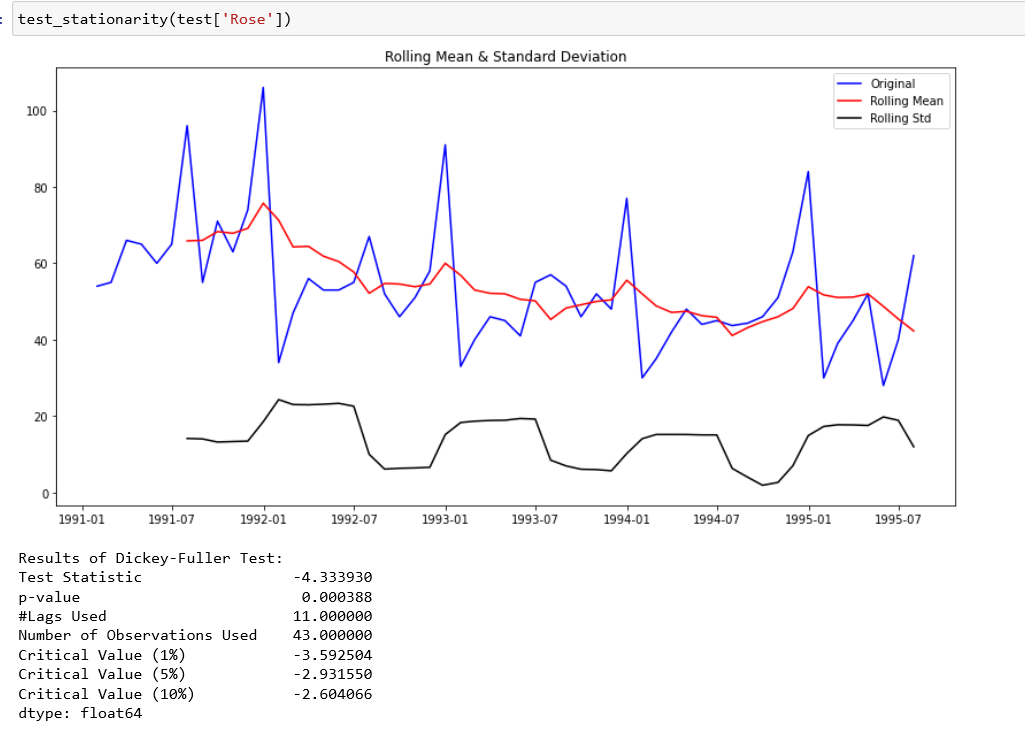
1. ROSE WINE :

TRAINING DATA :



We observe a P value greater than alpha(0.05). So the training data appears to be stationary as per the Alternate Hypothesis.

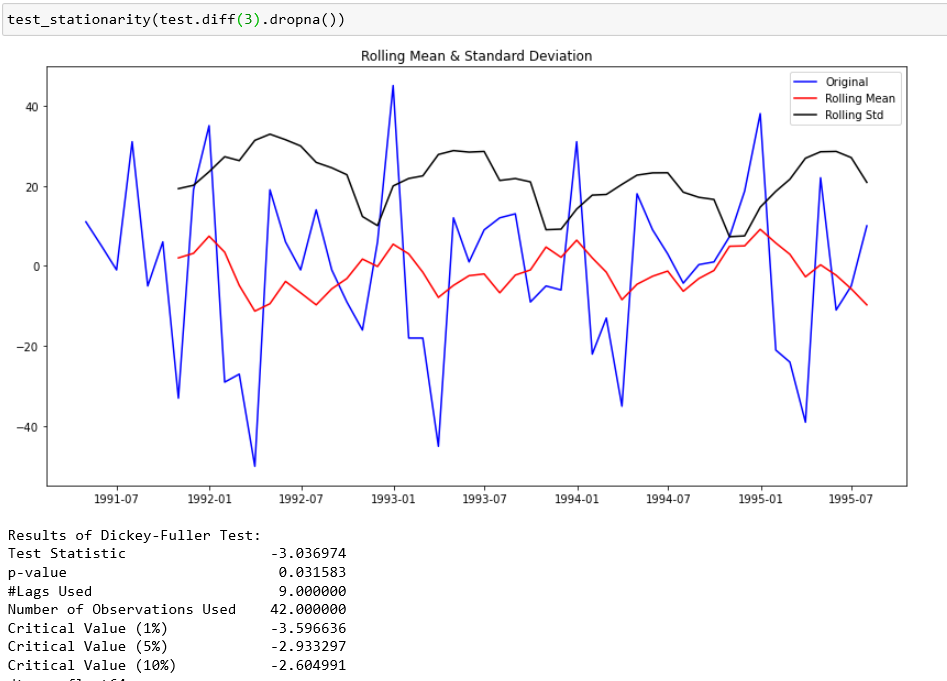
TEST DATA :



The test data appears to be non-stationary since P value is less than 0.05.

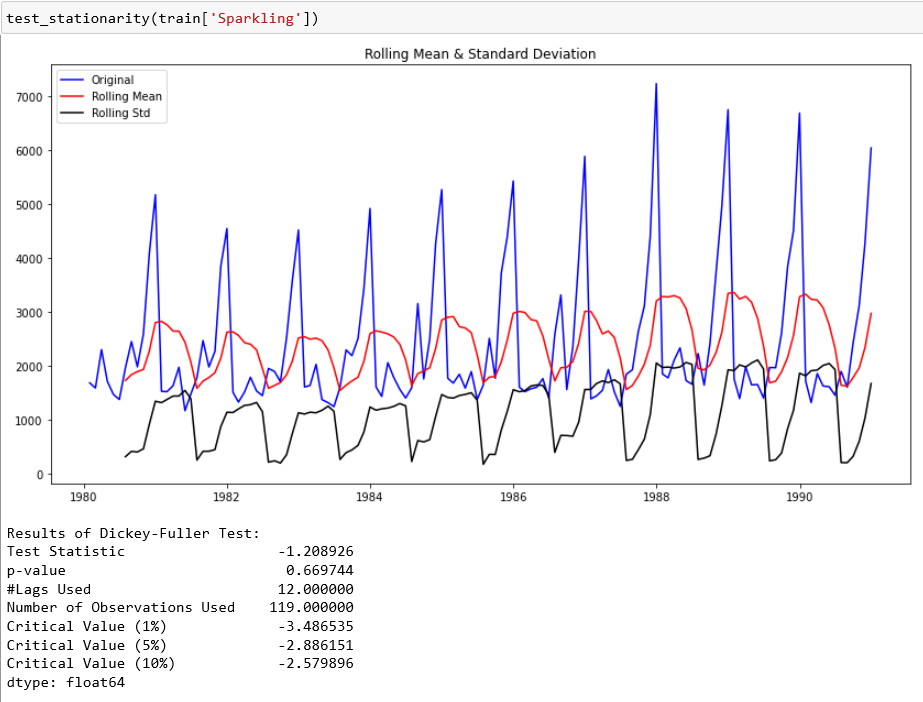
**DIFFERENCING** is one way to make a non-stationary time series stationary — compute the differences between consecutive observations.

In this case, we apply differencing until the P value becomes less than 0.05, which in our case takes three times.

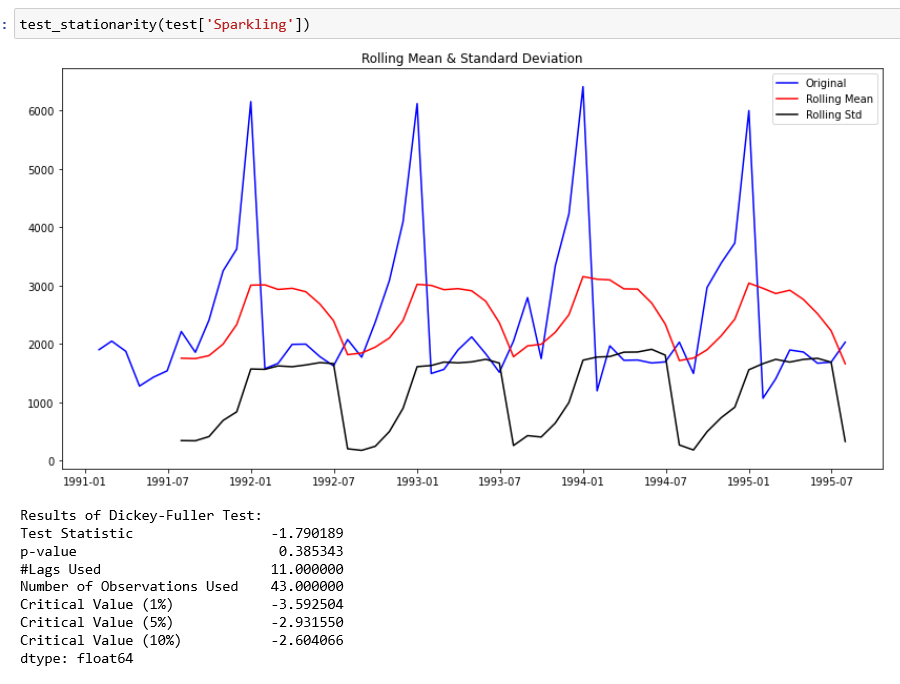


Here we apply differencing to make the series stationary.

1. SPARKLING WINE :



We observe a P value greater than alpha(0.05). So the training data appears to be stationary as per the Alternate Hypothesis.

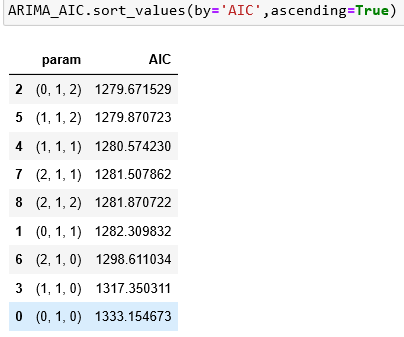


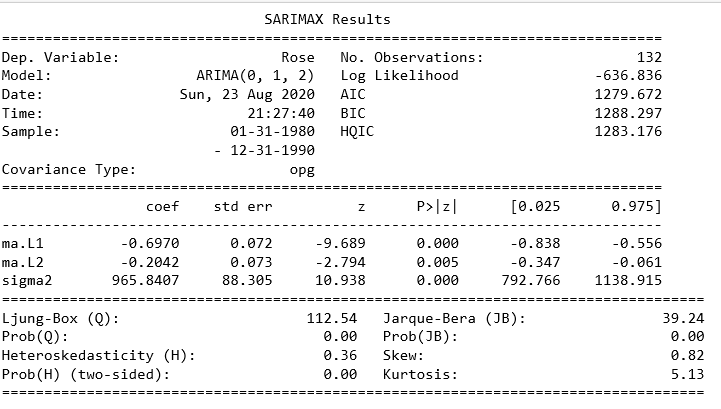
Test Data also shows a P value greater than 0.05. Therefore the test data exhibits stationarity.

# **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.**

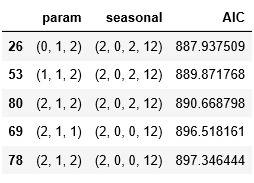
The p,d,q values of the ARIMA are varied between 0 and 2 to perform a design of experiments (DOE) through itertools. The values are then sorted to reveal the combination that provides the lowest AIC value. This value is used to construct the best ARIMA model in this case.

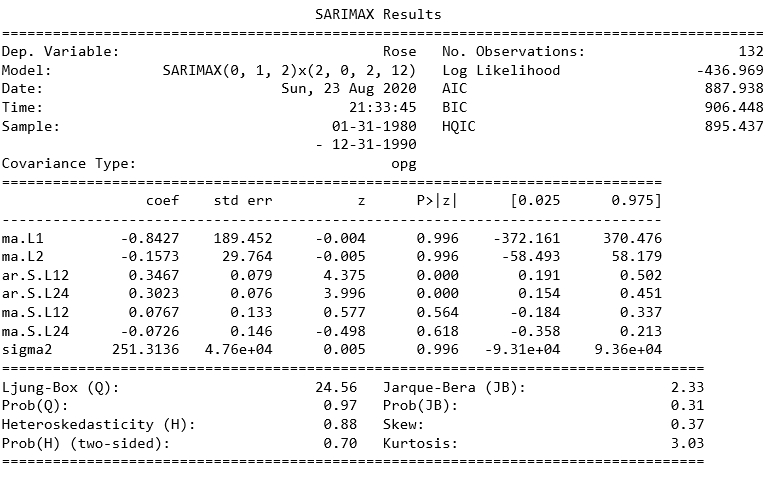
1. **ROSE :**

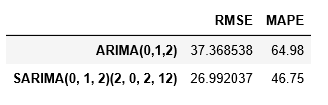




Similarly the SARIMA model is also tried out by varying both the seasonal and non-seasonal variables in an iterative fashion as above and checked for least AIC.

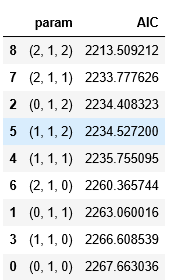


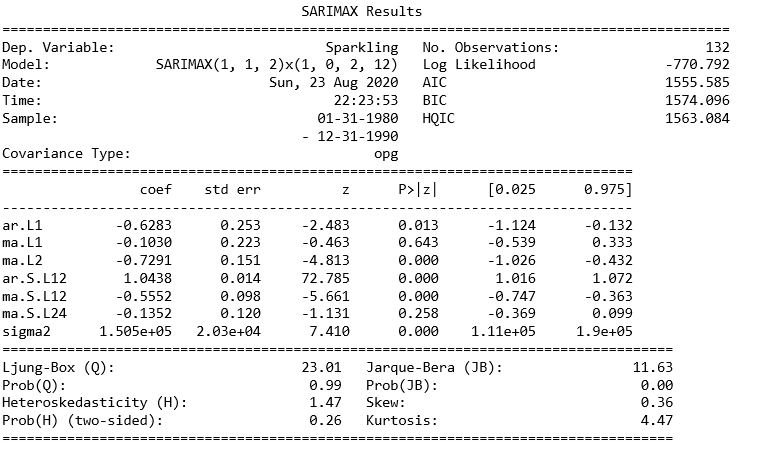




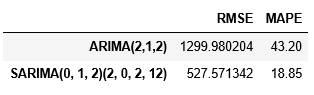
Finally, its found that the SARIMA model performs better in our case with a lower RMSE value, and hence can be chosen.

1. **SPARKLING :**



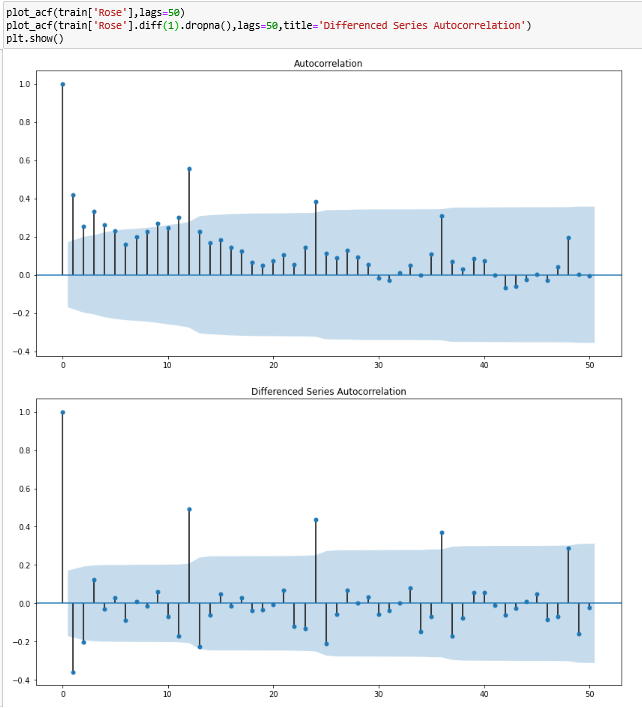


The ARIMA/SARIMA values are plotted and it is found that the SARIMA model performs much better in terms of having a lower RMSE value, compared to ARIMA.

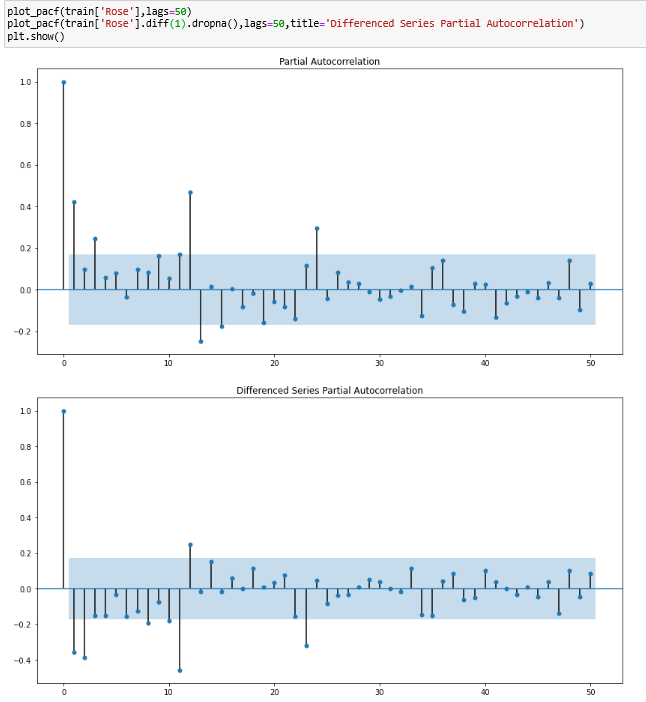


# **Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

**ROSE: ACF and Differenced ACF**

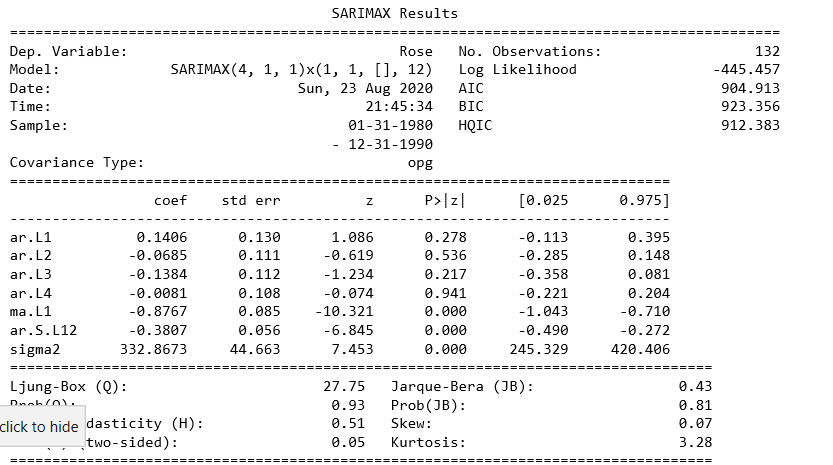


**ROSE: PACF and Differenced PACF**

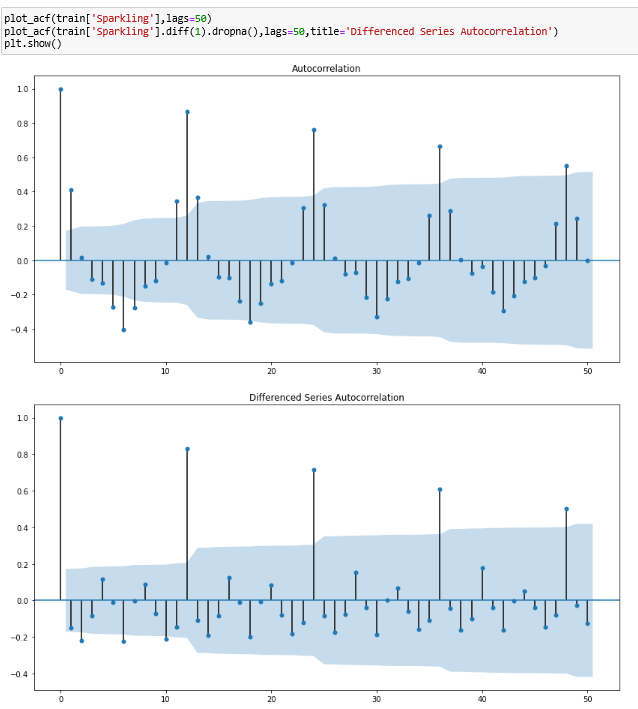


The PACF provides an estimate of a good ‘p’ value and the ACF for the ‘q’ value. Differencing to make the data stationary gives an estimate of the value of ‘d’. Similarly, through differenced measures, we can get the values of P,D,Q.

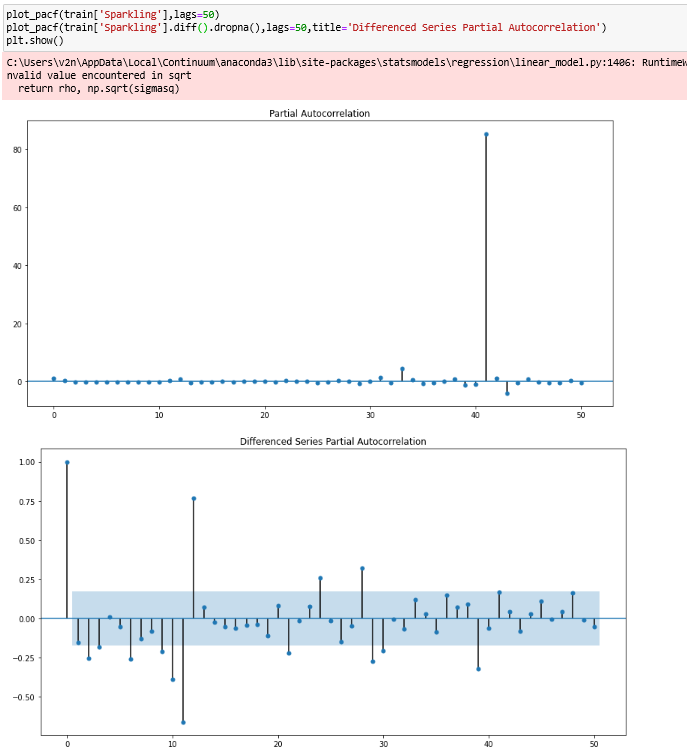
In our case, we go with p=4, d=1,q=1 for non-seasonal values, and P=1,D=1,Q=0 for seasonal values and estimate the value of RMSE for this condition. It is found to be lower than the earlier SARIMA and ARIMA as per comparative table below.

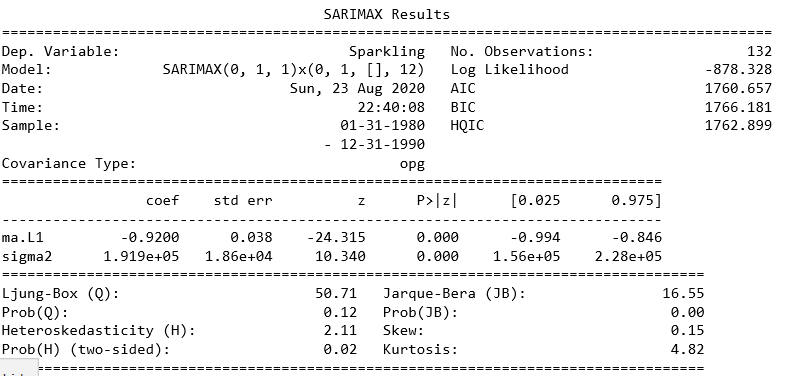


**SPARKLING : ACF and Differenced ACF**



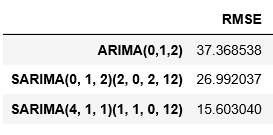
**SPARKLING : PACF and Differenced PACF**



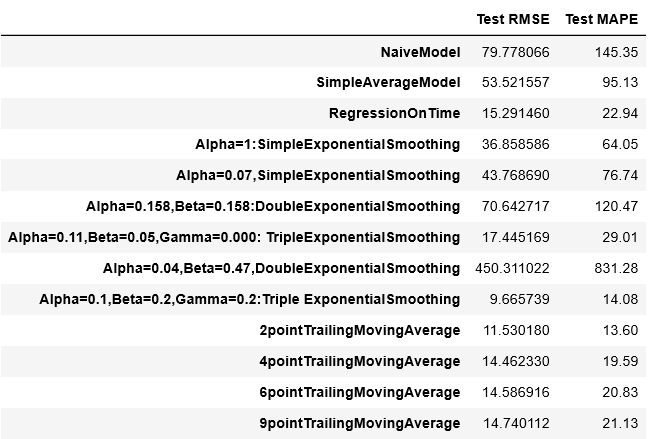


# **Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

**ROSE (ARIMA/SARIMA) :**

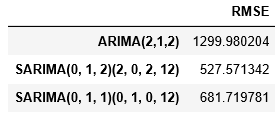


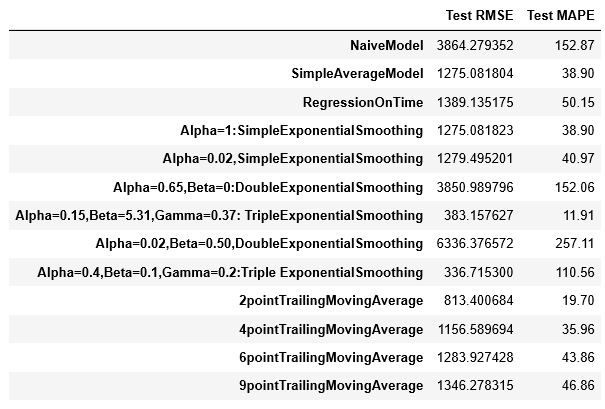
**ROSE (Other Models):**



The SARIMA and the moving average models provide a low RMSE value and can be used for future predictions on sales for the next 12 months.

**SPARKLING (ARIMA/SARIMA) :**





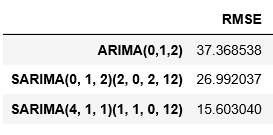
The Triple Exponential model with an RMSE of 337 units appears to be the best model, but provides unexpected results when a prediction is done. For ARIMA/SARIMA, the SARIMA model with an RMSE of 528 units appears suitable with a prediction on expected lines.

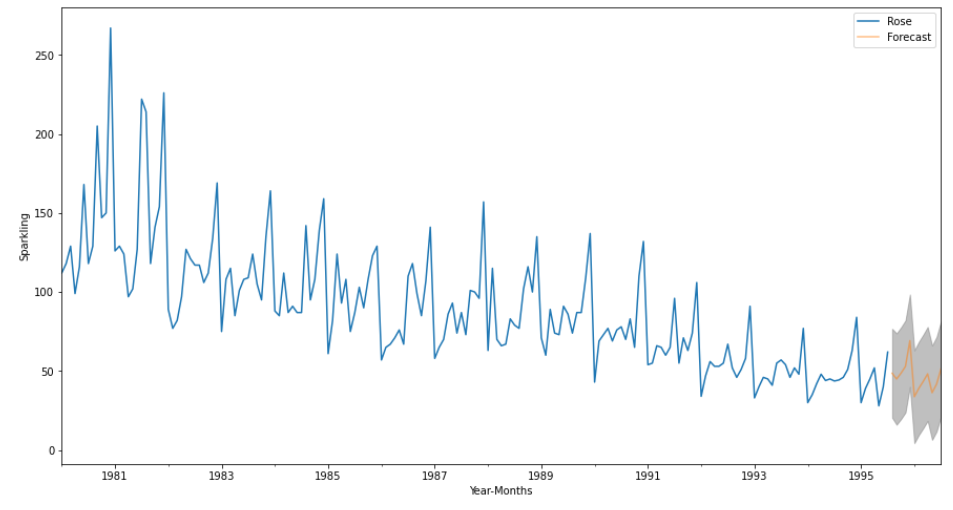
# **Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.**

**ROSE WINE (PREDICTION) :**

Orange line shows prediction, gray line shows the confidence interval.

The SARIMA model shown below is utilized for prediction owing to its low RMSE score of ~15 units and owing to the ability of the model to capture both seasonal and non-seasonal effects.



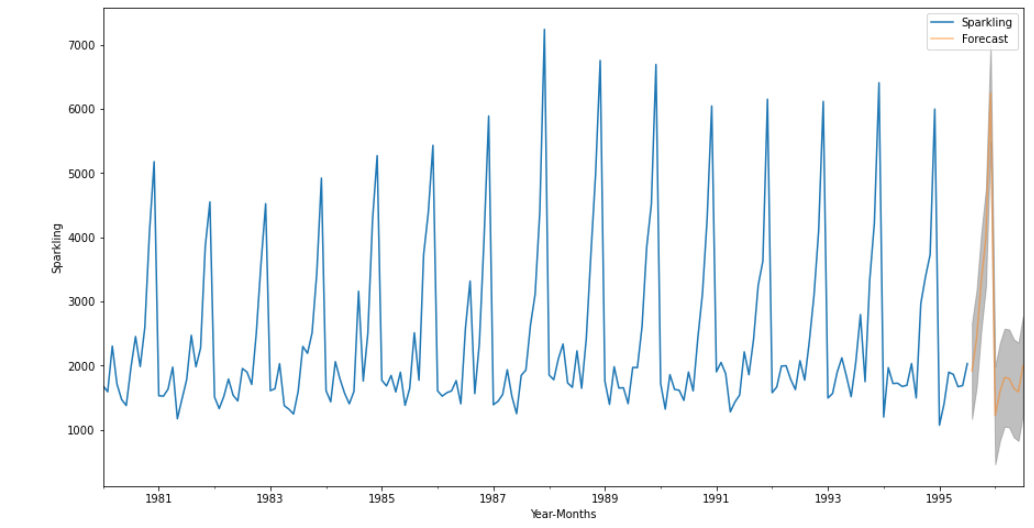


**SPARLKLING WINE (PREDICTION) :**

Orange line shows prediction, gray line shows the confidence interval of 95%.

The SARIMA model shown below is utilized for prediction owing to its low RMSE score of ~19 units and owing to the ability of the model to capture both seasonal and non-seasonal effects.





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

Based on the forecasting graphs shown in the previous pages for both varieties of Wine, it is clear that

1. The sales of Rose wine has been decreasing over the years and is predicted to decrease over the next 12 months.
2. The sales of Sparkling wine has been steady with promise of increasing and appears to follow a similar trend over the next twelve months.
3. The sales of Wine appears to be highest during the holiday season.

Keeping these observations in mind, its recommended that ABC company look into the reasons for decrease in the popularity of their Rose wine, and undertake any sales or marketing campaigns to improve its sales. Its possible that other drinks have been introduced into the market and are eating into total market that Rose wine is a part of.

With Sparkling wine on the other hand, its observed that the sales is steady or even improving over the years. Its therefore necessary to keep the good work and start campaigns that build the popularity of sparkling wine exponentially and improve its sales.