**Time Series Forecasting Group Assignment BUSINESS ANSWER REPORT**

**Group No. 3 Assignment: Time Series Forecasting with Linear Regression, Moving Average, Single/Double/Triple Exponential Smoothing, ARIMA and SARIMA**

**Problem on Time Series Forecasting:**

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.

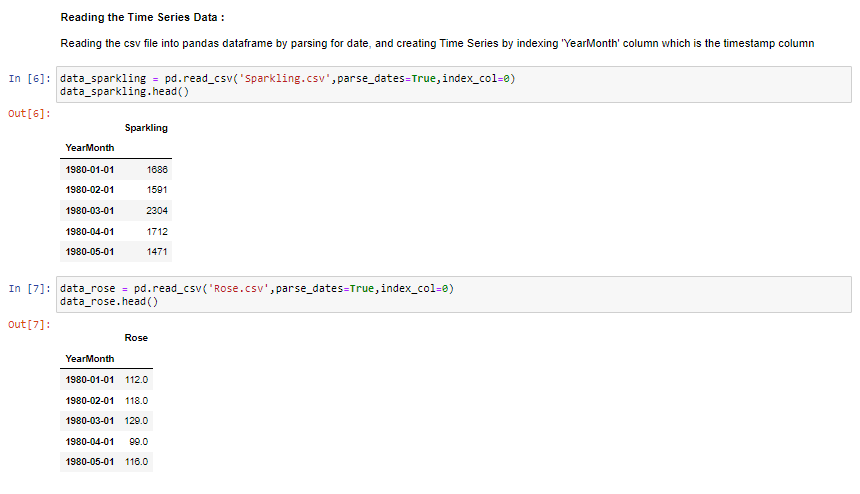
Data set for the Problem: [Sparkling.csv](https://olympus.greatlearning.in/courses/9603/files/2703523/download?verifier=pFJWYD2pSMDinscKhcPJhChjTZSVFSRVrDlpJuzX&wrap=1) and [Rose.csv](https://olympus.greatlearning.in/courses/9603/files/2703524/download?verifier=NlZnbf6uKs66ojyr1D6aaqVNsMfv07nMzNEcwHsl&wrap=1)

**Questions:**

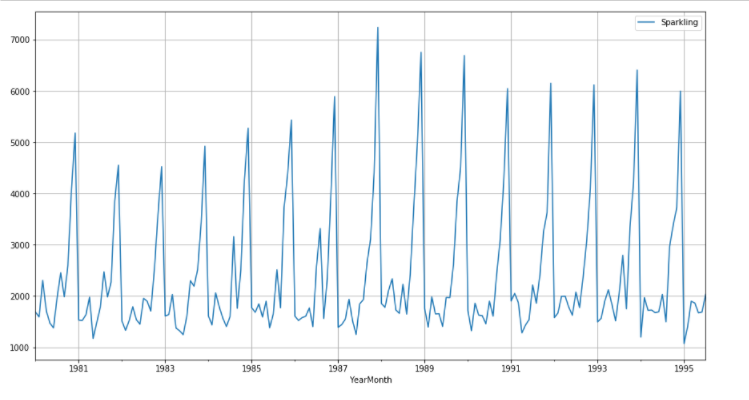
**Q.1 Read the data as an appropriate Time Series data and plot the data**

**A.1**

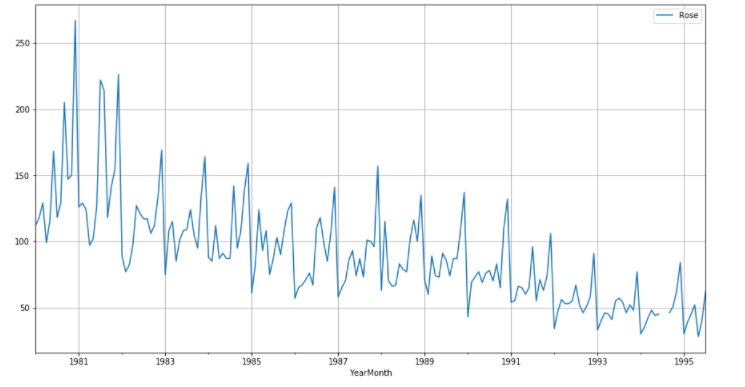
* On reading the Sparkling Sales & Rose Sales file from csv to pandas' data frame, we see we have 2 columns, 1st column is the timetamp in YYYY-MM format and 2nd column is the monthly Sparkling/Rose wine sales.
* We converted the dataframe into Time Series dataset, by indexing the timestamp column



* On plotting Sparkling Sales Time Series Data, we see the sales dont show trend. However, the fluctuation in sales shows pattern which is repetitive year after year. Hence significant seasonal component seem to be in sales data



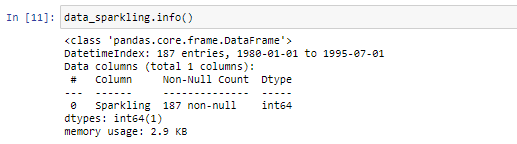
* On plotting Rose Sales Time Series Data, we see the sales seem to show significant downward trend. Also, the fluctuation in sales shows some pattern which is repetitive year after year. Hence significant seasonal component seem to be in sales data. Also, looking at the gap in monthly sales for year 1994, seems there may be missing values in sales data



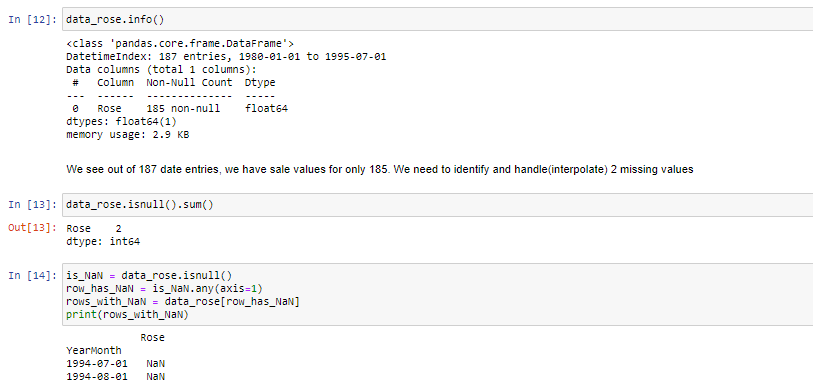
**Q.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

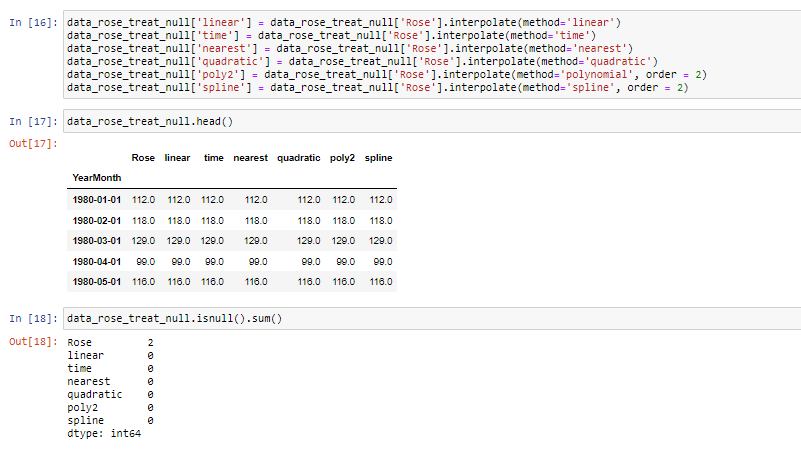
**A.2**

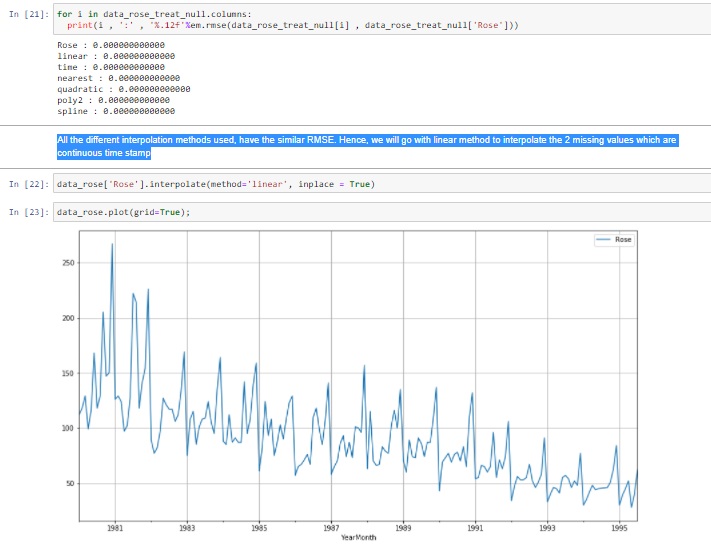
* Sparking sales TSD do not have any null data.



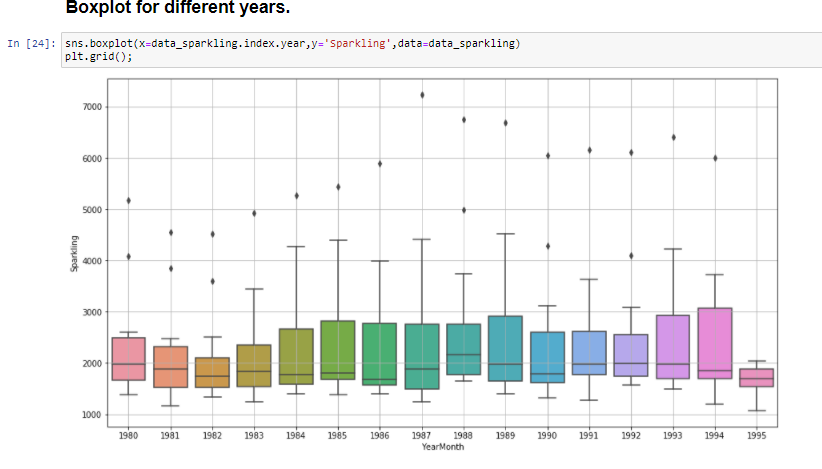
* Rose sales TSD shows having null data. Hence we tried interpolating them with different interpolated methods – linear, time, nearest, quadratic, polynomial, spline.
* All the different interpolation methods used, have the similar RMSE. Hence, we go with linear method to interpolate the 2 missing values which are continuous time stamp, as we prefer simple & straighforward method if all giving same result



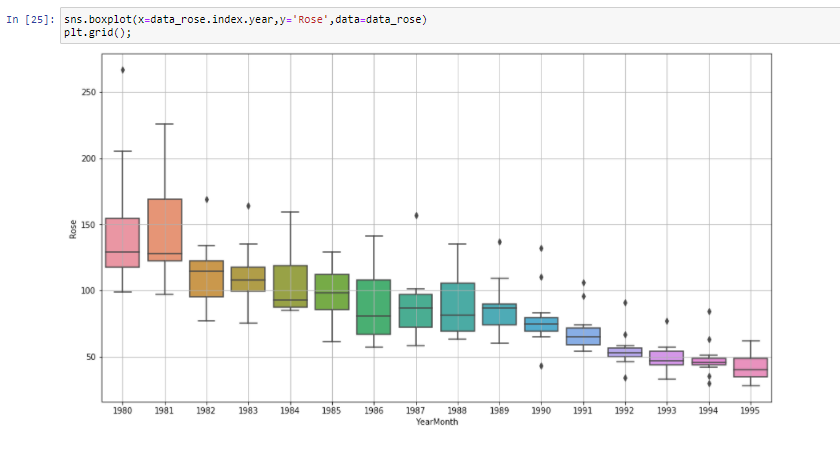




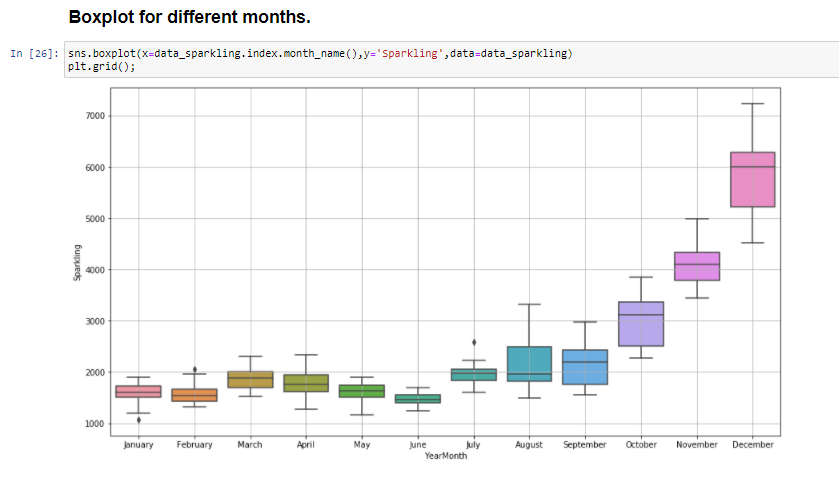
* Boxplot for different years for Sparkling sales, shows median sales around 2000, across all the given years.

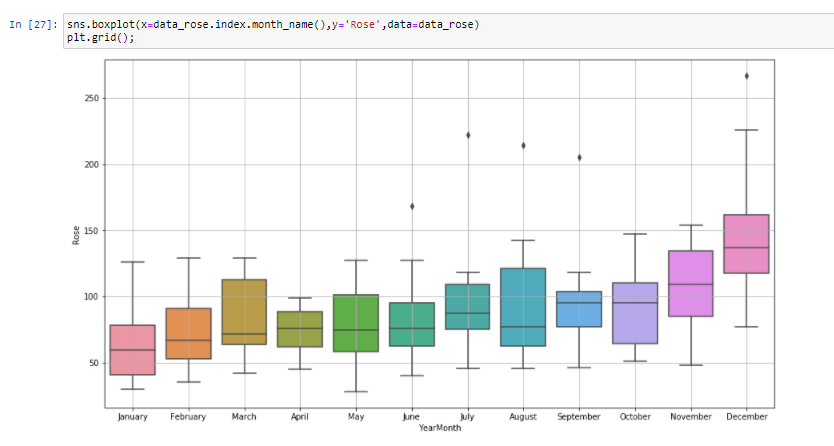


* Boxplot for different years for Sparkling sales, shows median sales trend going down from median 100 to below median 50 across years from 1980 to 1995.

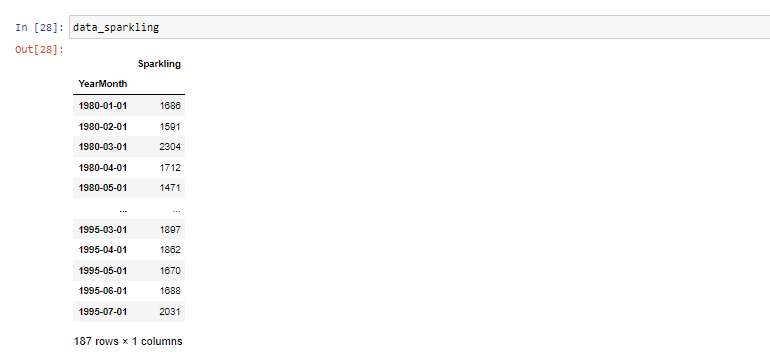


* Boxplot of different months across years for Sparkling & Rose sales, shows increase in sales during the month of November-December every year.

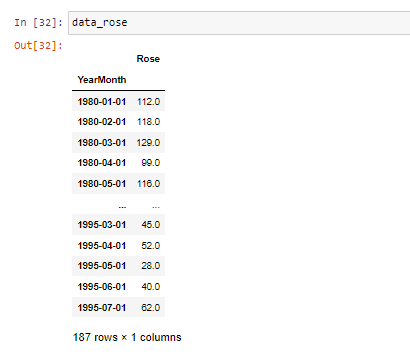




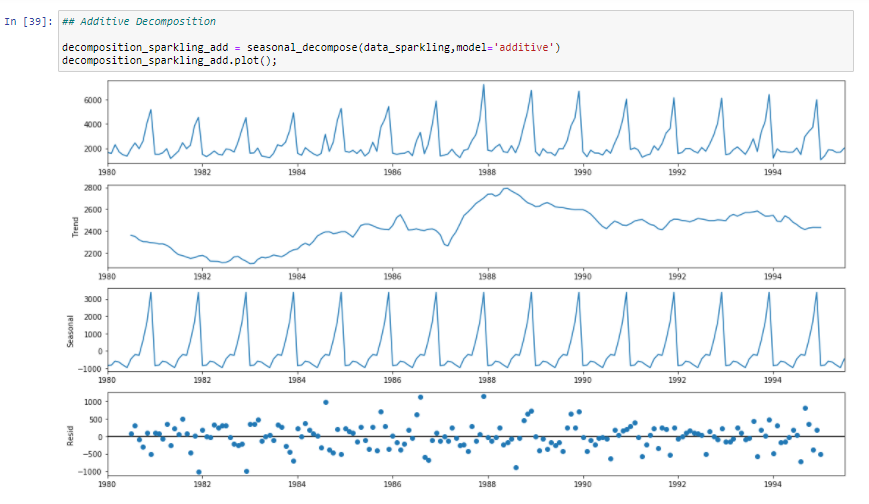
* We have 187 monthly sales data for Sparkling wines starting from Jan-1980 to July-1995.



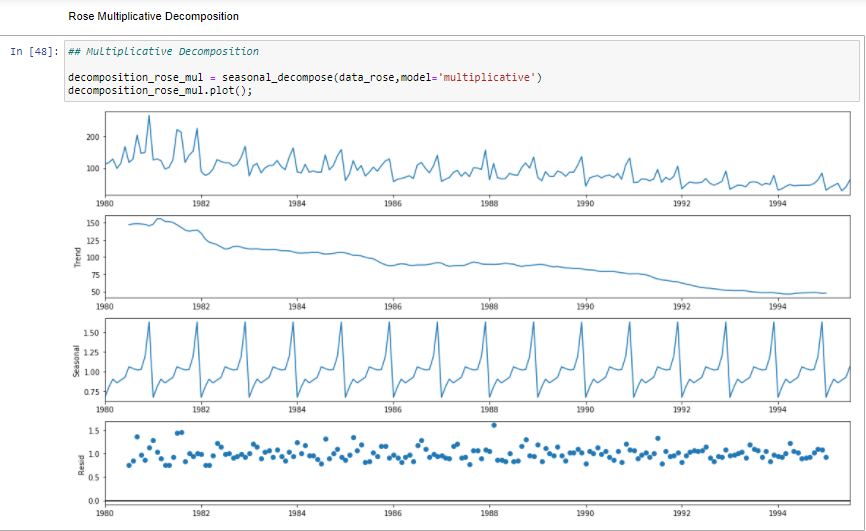
* We have 187 monthly sales data for Rose wines starting from Jan-1980 to July-1995.



* For Sparkling sales, additive seasonal decomposition seems appropraite as the residual shows more random pattern around 0.



* For Rose sales, multiplicative seasonal decomposition seems appropraite as the residual shows more random pattern around 1.



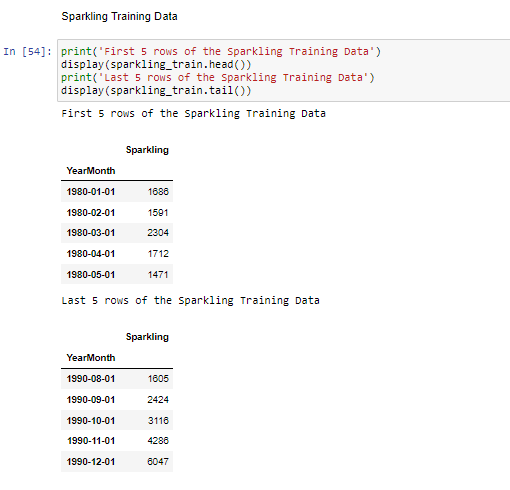
**Q.3 Split the data into training and test. The test data should start in 1991.**

**A.3**

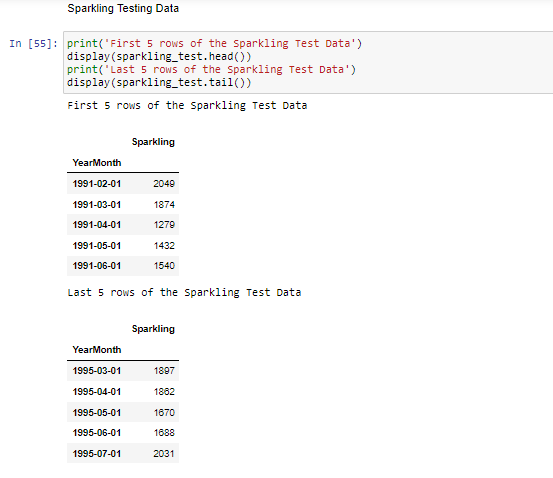
* Splitting the Sparkling Sales TSD into training and test. The test set starts from 1991



* Sparkling Sales Train TSD



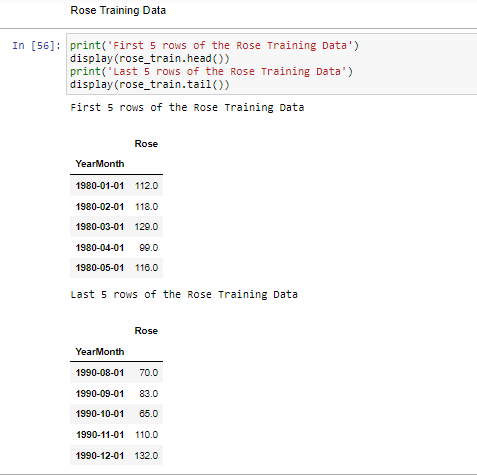
* Sparkling Sales Test TSD



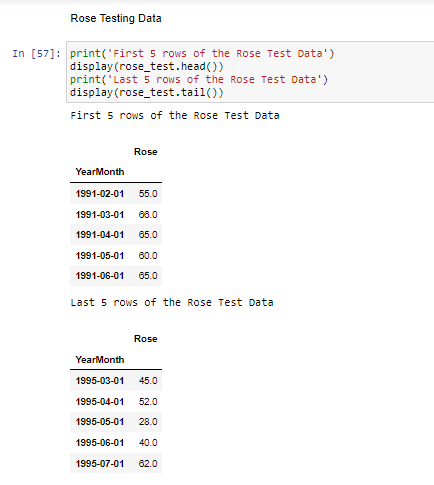
* Splitting the Rose Sales TSD into training and test. The test set starts from 1991



* Rose Sales Train TSD



* Rose Sales Test TSD



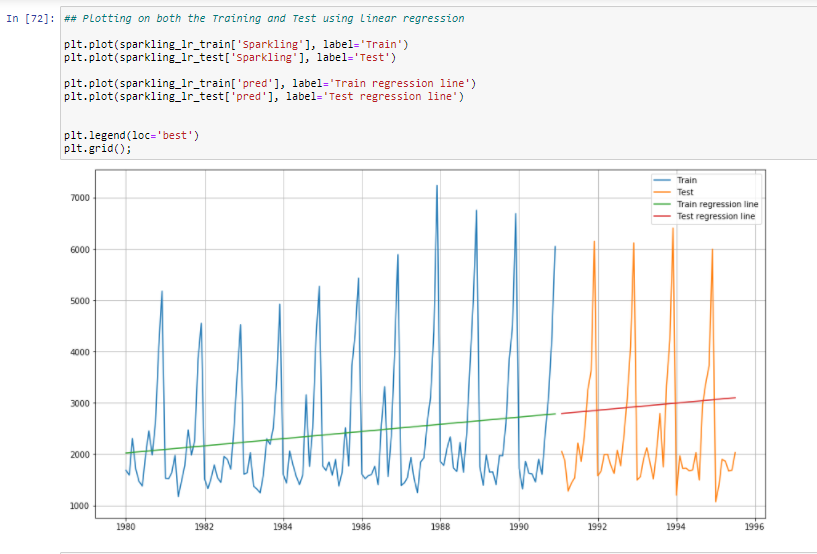
**Q.4 Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.**  
**Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using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

**A.4**

# **Sparkling wine monthly sales - Model building**

1. **Model 1 - Linear Regression for Sparkling wine sales**

**It picks the trend in given training TSD, and forecast for the test set**



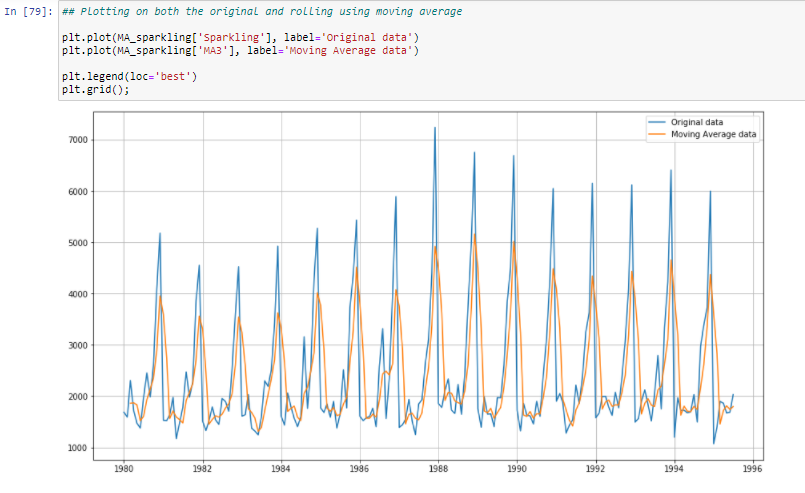


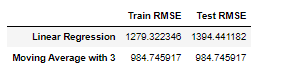
1. **Model 2 - Moving Average for Sparkling wine sales**

**We selected Moving average of 3 as it was giving the least error. Hence, our Moving Average model picks the 3 time period lag behaviour.**

**More the lags we consider for moving average, smoother is the forecast made.**

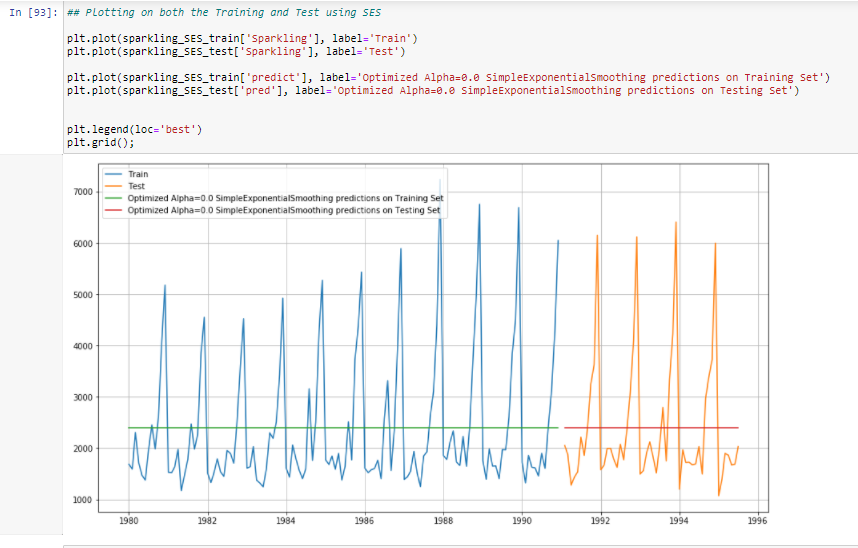
**Less the lags we consider for moving average, it picks the small fluctuation & forecast also shows the fluctuation.**

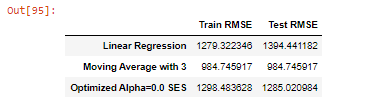




1. **Model 3 - SES(optimized parameter) for Sparkling wine sales**

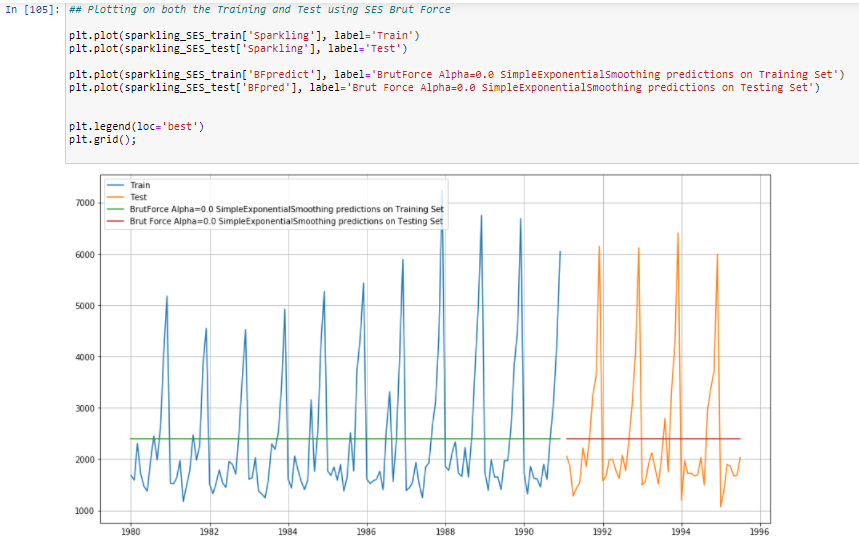
**Simple Exponential Smoothing forecast for the level in the TSD. While building SES model optimized is set to True, to let model automatically pick the optimized best smoothing\_level, which came smoothing\_level = 0.0**

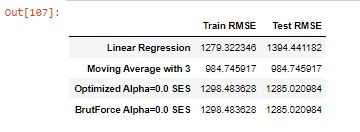




1. **Model 3 - SES(BrutForce parameter) for Sparkling wine sales**

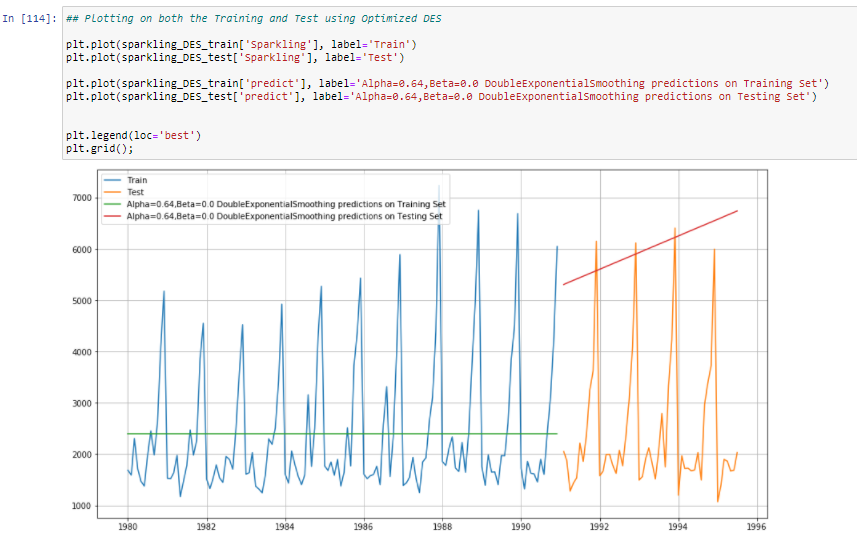
**Simple Exponential Smoothing forecast for the level in the TSD. While building SES model manually, we iterated thru range(0,10) for different smoothing\_level & build SES with capturing RMSE for each of them. Smoothing\_level = 0.0, gave the least RMSE manuaaly as well.**

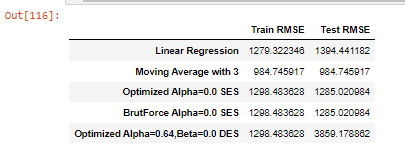




1. **Model 4 - DES Holt's Model for Sparkling wine sales**

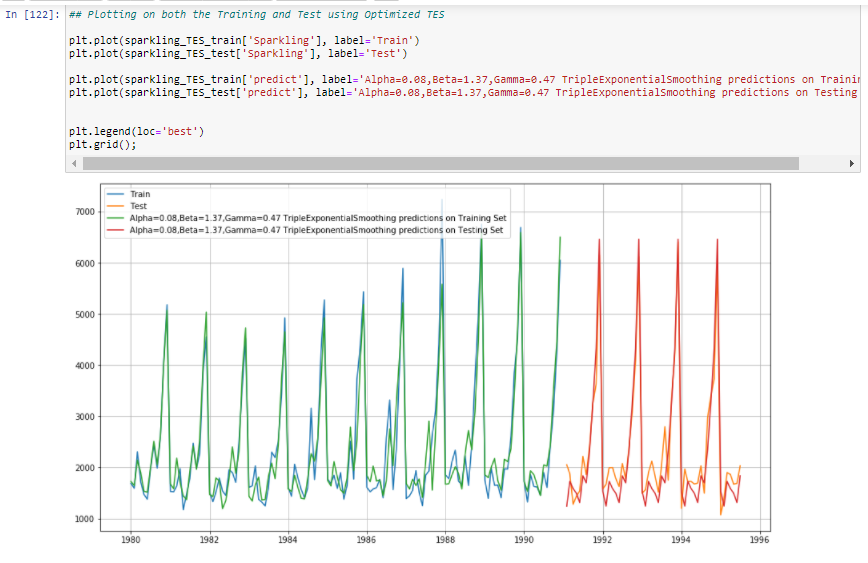
**Double Exponential Smoothing forecast for the level & trendcomponent in the TSD. While building DES model optimized is set to True, to let model automatically pick the optimized best smoothing\_level & smoothing\_slope which came to be** 0.64 & 0.0 **respectively**

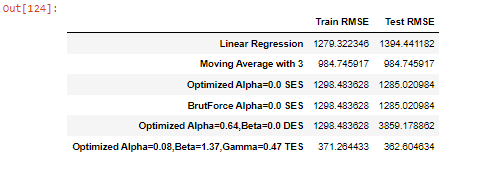




1. **MODEL 5 TRIPLE EXPONENTIAL SMOOTHING - HOLT WINTER's Model for Sparkling wine sales**

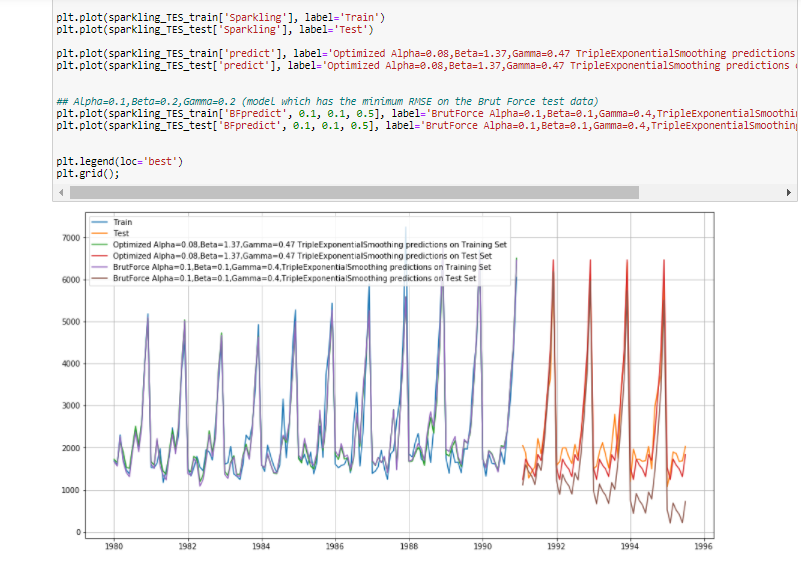
**Triple Exponential Smoothing forecast for the level, trend and seasonal component in the TSD. While building TES model optimized is set to True, to let model automatically pick the optimized best smoothing\_level, smoothing\_slope & smoothing\_seasonal which came to be** 0.08, 1.37 & 0.47 **respectively**

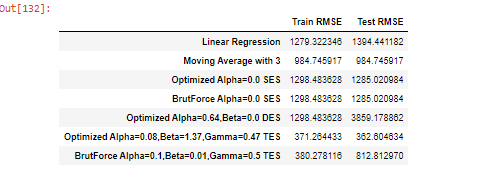




1. **MODEL 5 - BrutForce TRIPLE EXPONENTIAL SMOOTHING - HOLT WINTER's Model for Sparkling wine sales**

**Triple Exponential Smoothing forecast for the level, trend and seasonal component in the TSD. While building TES model manually, we iterated thru range(0,10) for different smoothing\_level, smoothing\_slope, smoothing\_seasonal & build TES with capturing RMSE for each of them. Smoothing\_level = 0.1, smoothing\_slope = 0.01, smoothing\_seasonal = 0.5 gave the least RMSE manually as well.**

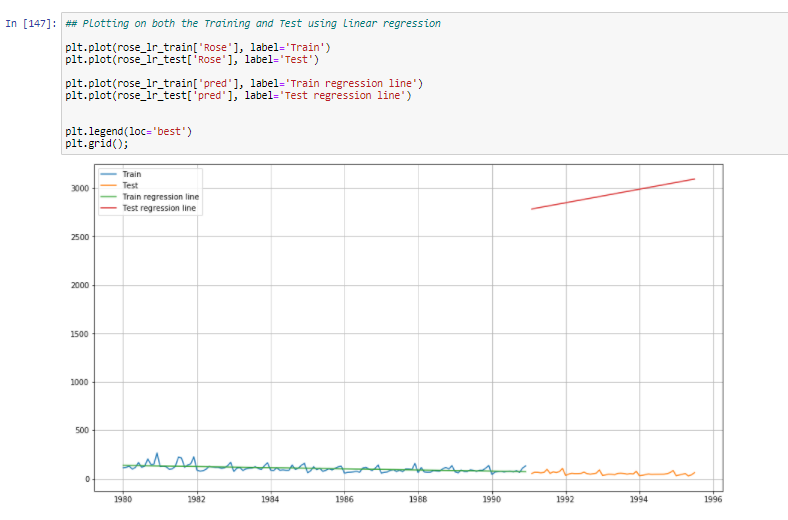




# **Rose wine monthly sales - Model building**

1. **Model 1 - Linear Regression for Rose wine sales**

**It picks the trend in TSD, and forecast for the test set**



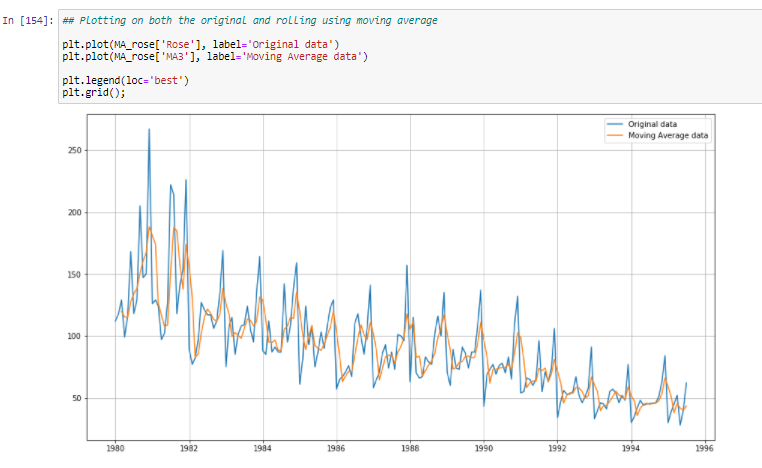


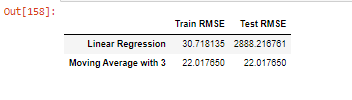
1. **Model 2 - Moving Average for Rose wine sales**

**We selected Moving average of 3 as it was giving the least error. Hence, our Moving Average model picks the 3 time period lag behaviour.**

**More the lags we consider for moving average, smoother is the forecast made.**

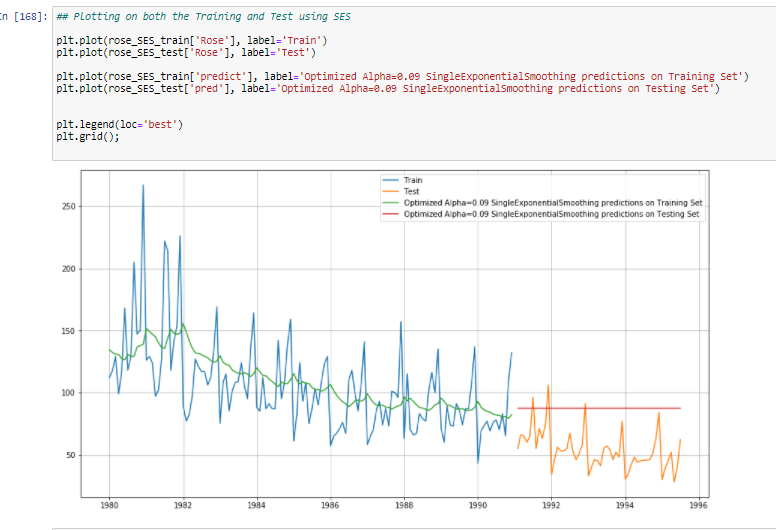
**Less the lags we consider for moving average, it picks the small fluctuation & forecast also shows the fluctuation.**





1. **Model 3 - SES(optimized parameter) for Rose wine sales**

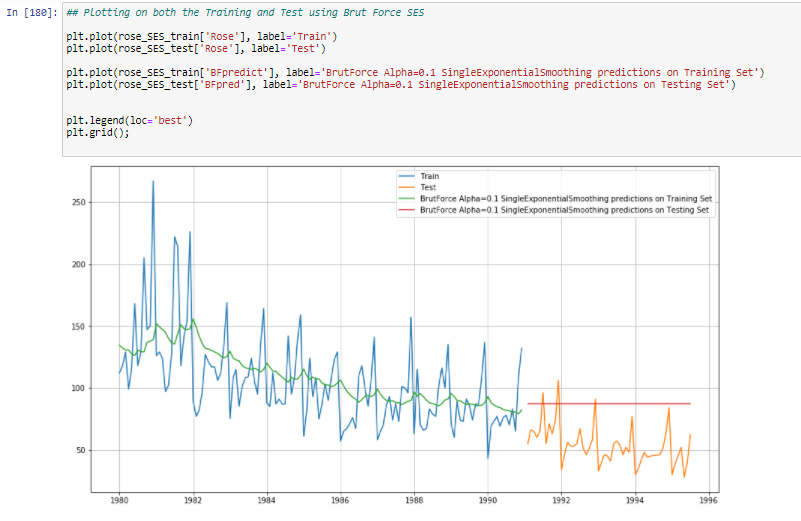
**Simple Exponential Smoothing forecast for the level in the TSD. While building SES model optimized is set to True, to let model automatically pick the optimized best smoothing\_level, which came smoothing\_level = 0.09**

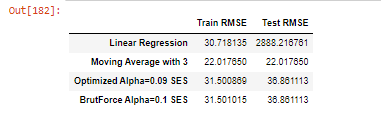




1. **Model 3 - SES(BrutForce parameter) for Rose wine sales**

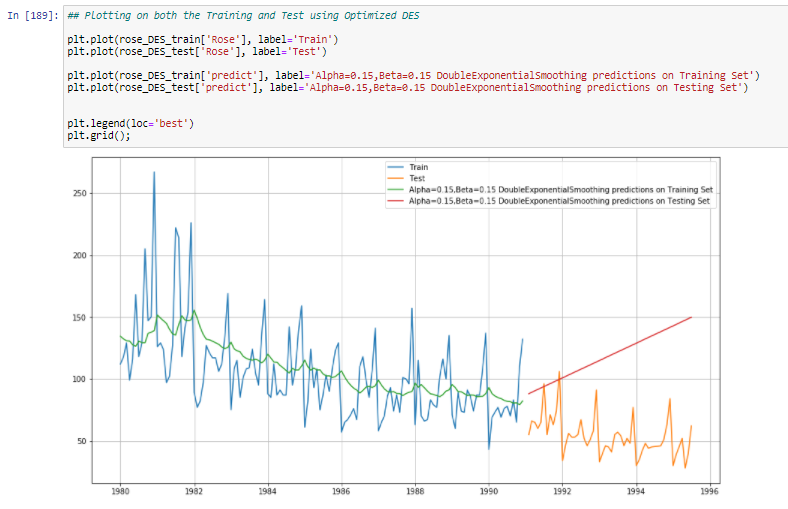
**Simple Exponential Smoothing forecast for the level in the TSD. While building SES model manually, we iterated thru range(0,10) for different smoothing\_level & build SES with capturing RMSE for each of them. Smoothing\_level = 0.1 gave the least RMSE manuaaly as well.**

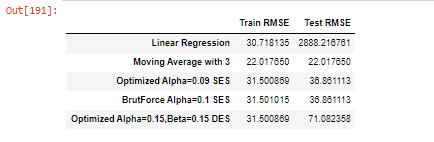




1. **Model 4 - DES Holt's Model for Rose wine sales**

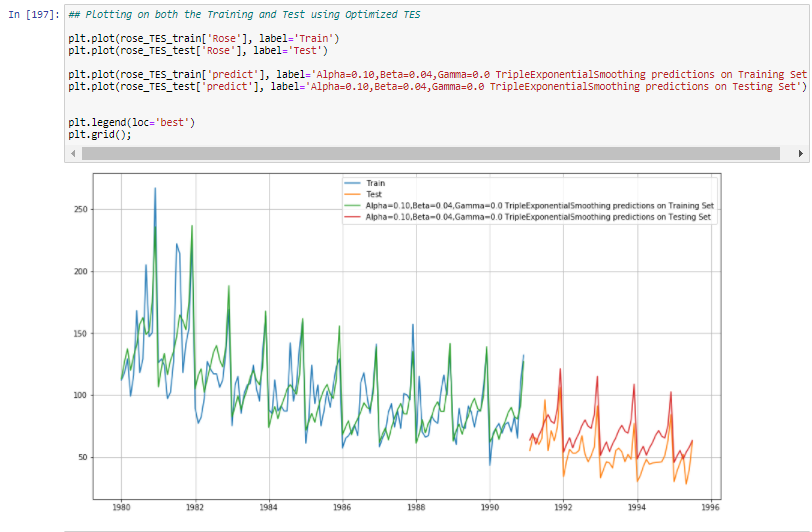
**Double Exponential Smoothing forecast for the level & trendcomponent in the TSD. While building DES model optimized is set to True, to let model automatically pick the optimized best smoothing\_level & smoothing\_slope which came to be** 0.15 & 0.15 **respectively**

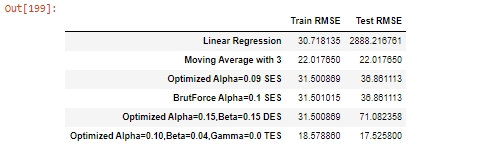




1. **MODEL 5 TRIPLE EXPONENTIAL SMOOTHING - HOLT WINTER's Model for Rose wine sales**

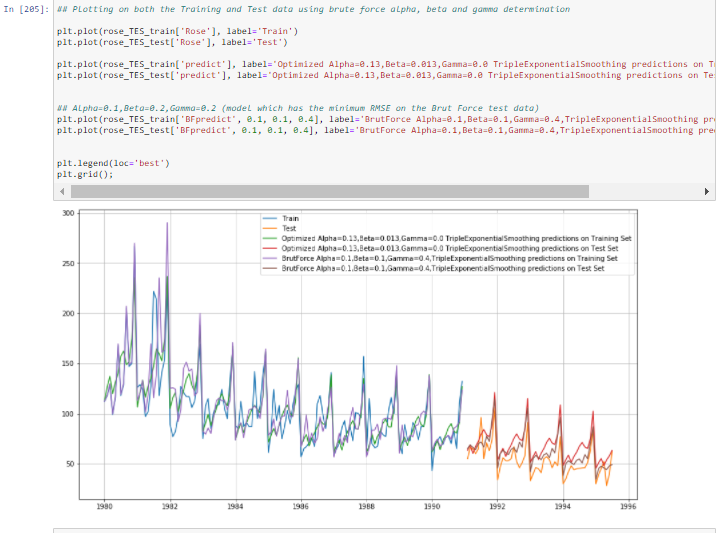
**Triple Exponential Smoothing forecast for the level, trend and seasonal component in the TSD. While building TES model optimized is set to True, to let model automatically pick the optimized best smoothing\_level, smoothing\_slope & smoothing\_seasonal which came to be** 0.10, 0.04 & 0.0 **respectively**

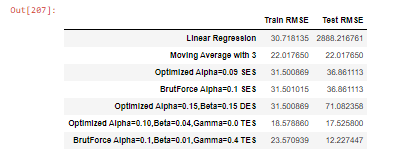




1. **MODEL 5 - BrutForce TRIPLE EXPONENTIAL SMOOTHING - HOLT WINTER's Model for Rose wine sales**

**Triple Exponential Smoothing forecast for the level, trend and seasonal component in the TSD. While building TES model manually, we iterated thru range(0,10) for different smoothing\_level, smoothing\_slope, smoothing\_seasonal & build TES with capturing RMSE for each of them. Smoothing\_level = 0.1, smoothing\_slope = 0.01, smoothing\_seasonal = 0.4 gave the least RMSE manually as well.**





**Q.5 Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.**  
**Note: Stationarity should be checked at alpha = 0.05.**

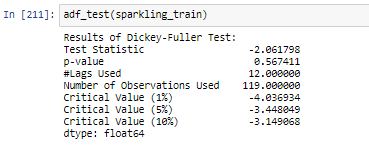
**A.5**

* We ran Augmented Dickey-Fuller test on Sparkling & Rose monthly sales TSD , to check for the stationarity of data.

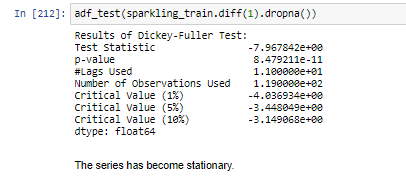
**Null Hypothesis for ADF test is** : Data is non-stationary

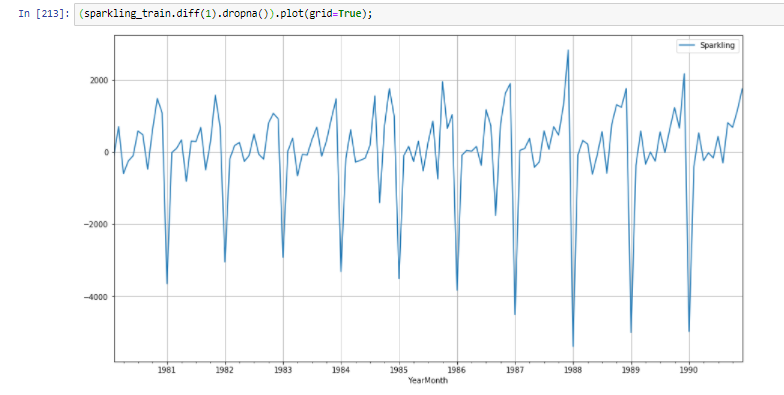
**Alternate Hypothesis for ADF test is** : Data is stationary

* **ADF test on Sparkling wine sales**

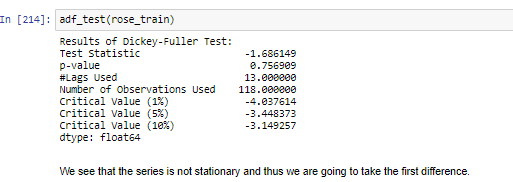


* We see p-value os 0.56, which is not less than significant p-value 0.05.Hence, data is non-stationary and thus we are going to take the first difference.
* ADF test on 1st difference of Sparkling wine sales, shows data has become stationary as p-val is 8.47\*10-11

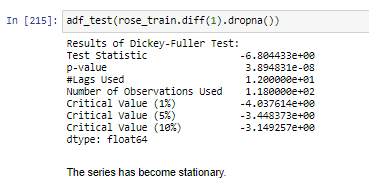


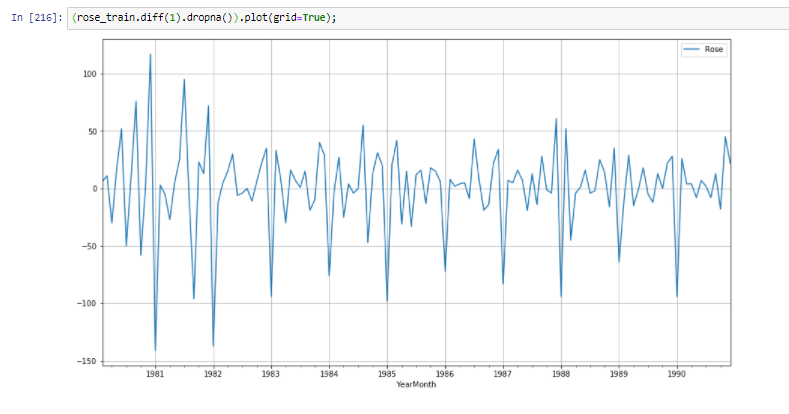


* **ADF test on Rose wine sales**



* We see p-value os 0.75, which is not less than significant p-value 0.05.Hence, data is non-stationary and thus we are going to take the first difference.
* ADF test on 1st difference of Rose wine sales, shows data has become stationary as p-val is 83.89\*10-08

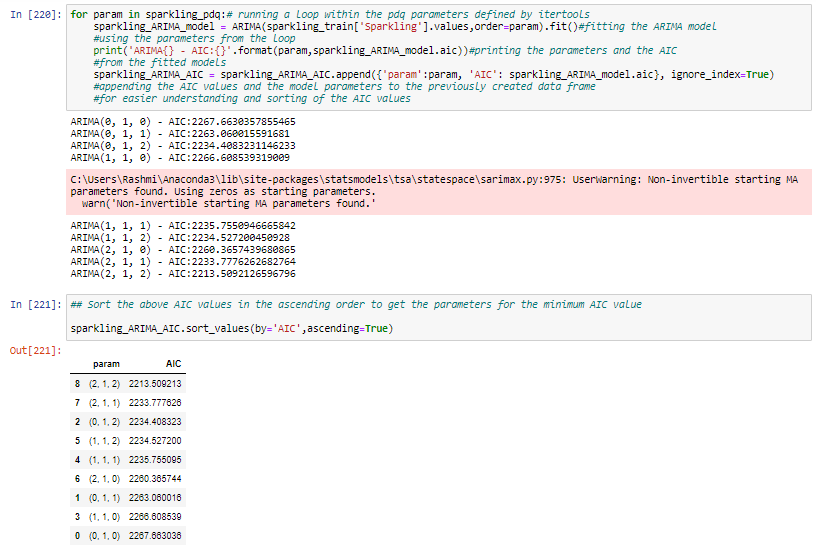




**Q.6 Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

**A.6**

* **Sparkling wine monthly sales – Automated version of ARIMA/SARIMA**
* Ran **ARIMA** on Sparkling wine sales TSD , by looping through p(AutoRegressor order) ,d(Difference order for stationarity),q(Moving Average order) parameters.
* For each and every combination of p,d,q ARIMA model is run and AIC is calculated.
* ‘p=2, d=1, q=2 gives the lowest ARIMA AIC for Sparkling wine sales



* Below is the model summary of ARIMA with lowest AIC for Sparking wine sales TSD

We have L1 & L2 component for AR & MA model for ARIMA in model summary. All the p-val is significant with 95% confidence.

AIC is Akaike Information Criteria captures the loss of information. Lower the vale of AIC, better is the model. Here for ARIMA model, AIC is 2213.509.

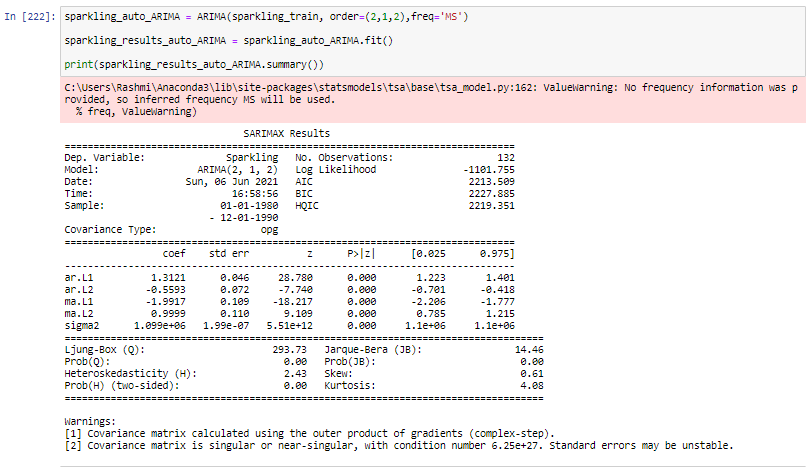
BIC stands for Bayesian Information Criteria

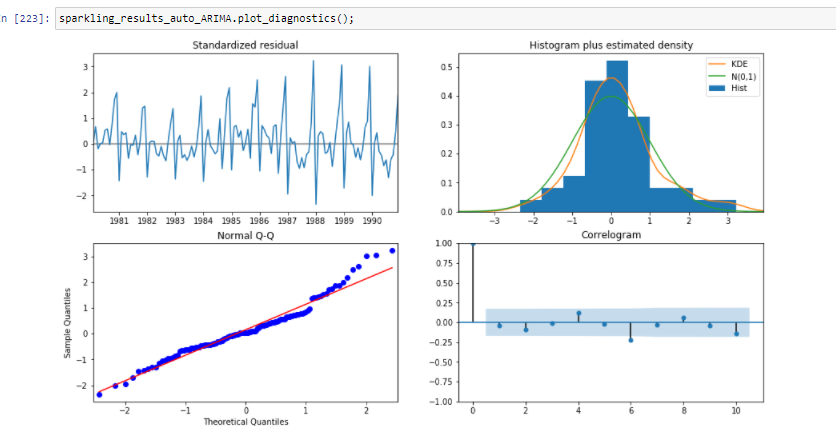
HQIC stands for Hannan Quinn Information Criteria

Jarque-Bera test is used for finding the Normality of residual. Looking at JB test p-val, we reject the Null Hypothesis, that means residual is not Normally distributed, which is reflected in Skew=0.61( normal range is between –0.5to0.5) & Kurtosis=4.08(usually it should be less than 3). The same is reflected in the histogram& Q-Q plot of the model diagnostic plot

Lung box test is for independence of residuals. Looking at Lung box test p-val, we reject the Null Hypothesis, that means residual are not independent(correlation of residual is present). The same is reflected in the correlogram of the model diagnostic plot, which shows lag point=6 is outside the confidence region in blue

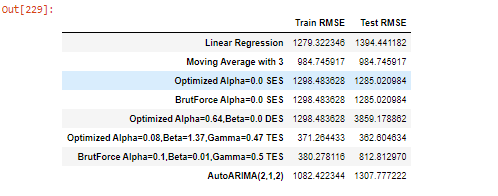
Heteroscedastic test shows the residuals are heteroscedastic.



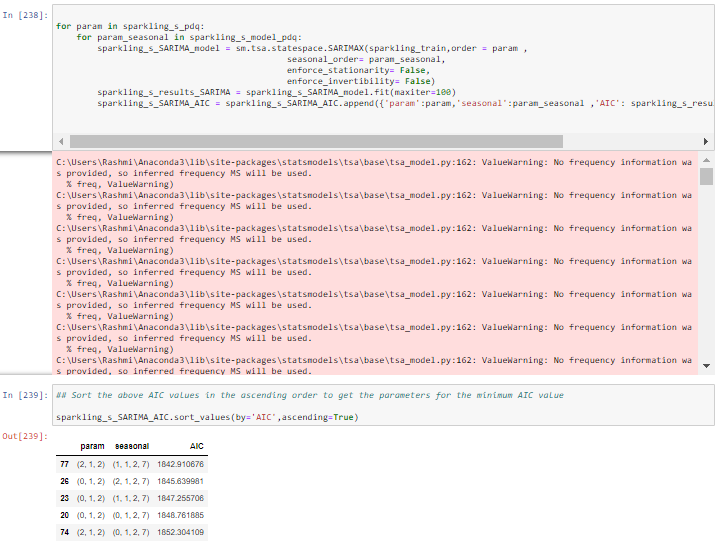


* Forecasting of Auto ARIMA for Sparkling wine sales TSD, with upper & lower confidence interval range is plotted as below

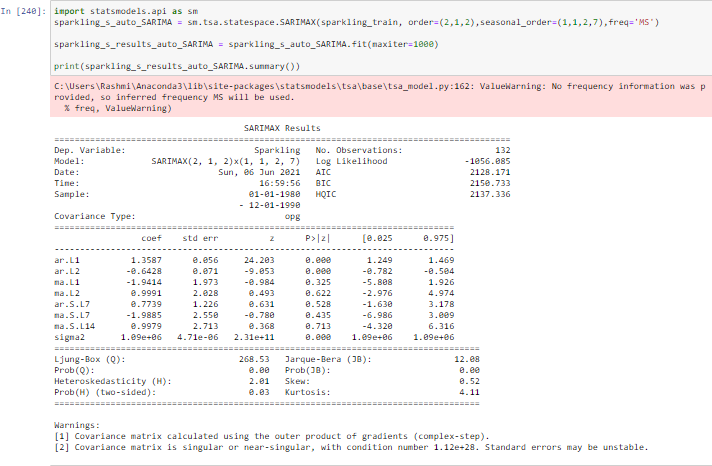




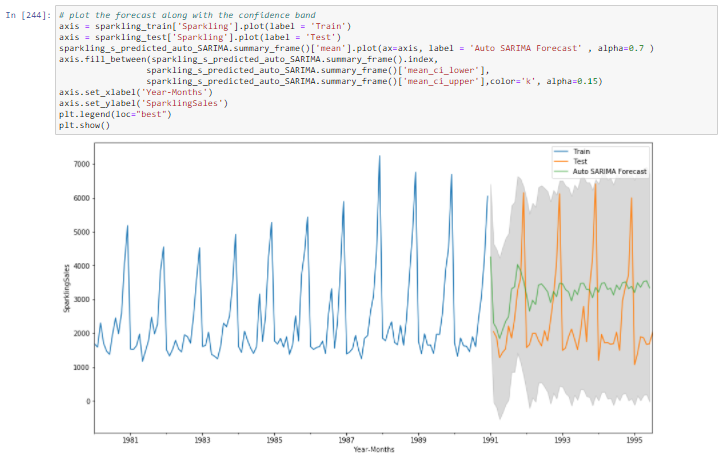
* Ran **SARIMA** on Sparkling wine sales TSD , by looping through Seasonal parameter – P,D,Q,M and p(AutoRegressor order) ,d(Difference order for stationarity),q(Moving Average order) parameters.
* For each and every combination of (p,d,q)(P,D,Q,M) SARIMA model is run and AIC is calculated.
* ‘p=2, d=1, q=2 & P=1, D=1, Q=2, M=7 gives the lowest SARIMA AIC for Sparkling wine sales

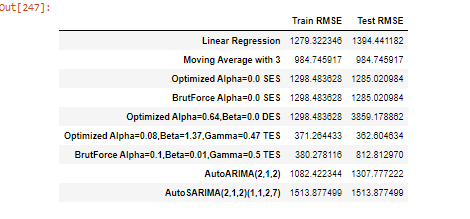


* Below is the model summary of SARIMA with lowest AIC for Sparking wine sales TSD

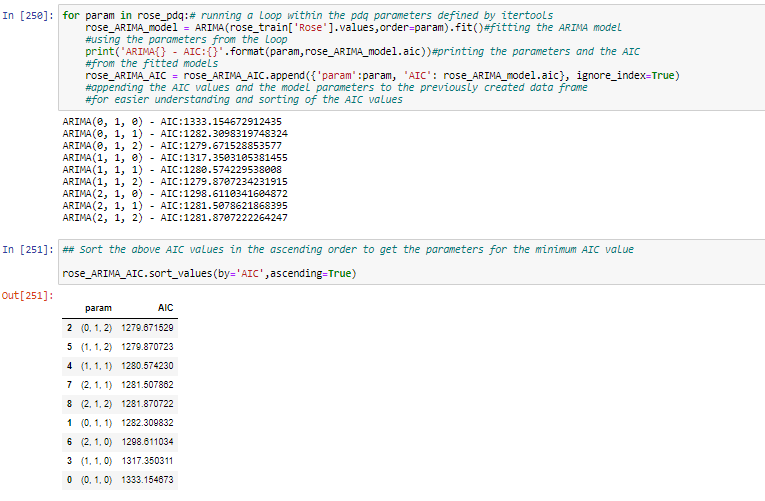


* Forecasting of Auto SARIMA for Sparkling wine sales TSD, with upper & lower confidence interval range is plotted as below





* **Rose wine monthly sales – Automated version of ARIMA/SARIMA**
* Ran **ARIMA** on Rose wine sales TSD , by looping through p(AutoRegressor order) ,d(Difference order for stationarity),q(Moving Average order) parameters.
* For each and every combination of p,d,q ARIMA model is run and AIC is calculated.
* ‘p=0, d=1, q=2 gives the lowest ARIMA AIC for Rose wine sales



* Below is the model summary of ARIMA with lowest AIC for Rose wine sales TSD

We have L1 & L2 component for MA model for ARIMA in model summary. All the p-val is significant with 95% confidence.

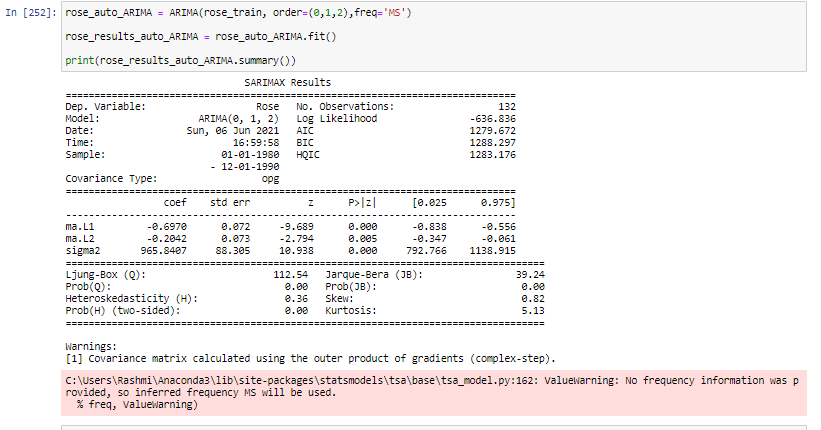
AIC is Akaike Information Criteria captures the loss of information. Lower the vale of AIC, better is the model. Here for ARIMA model, AIC is 1279.672.

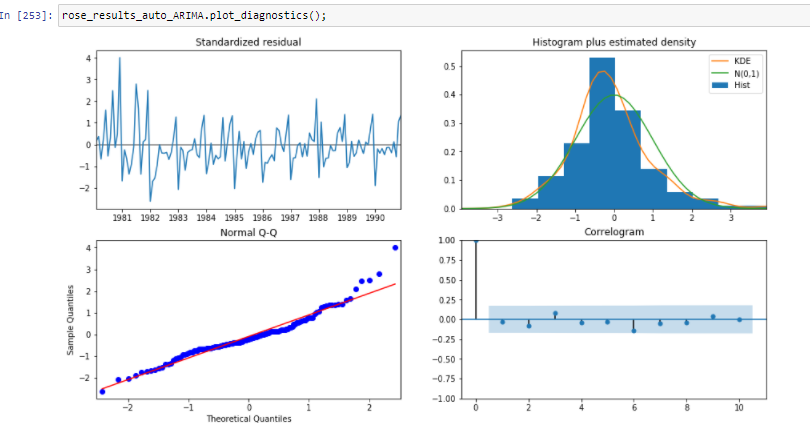
BIC stands for Bayesian Information Criteria

HQIC stands for Hannan Quinn Information Criteria

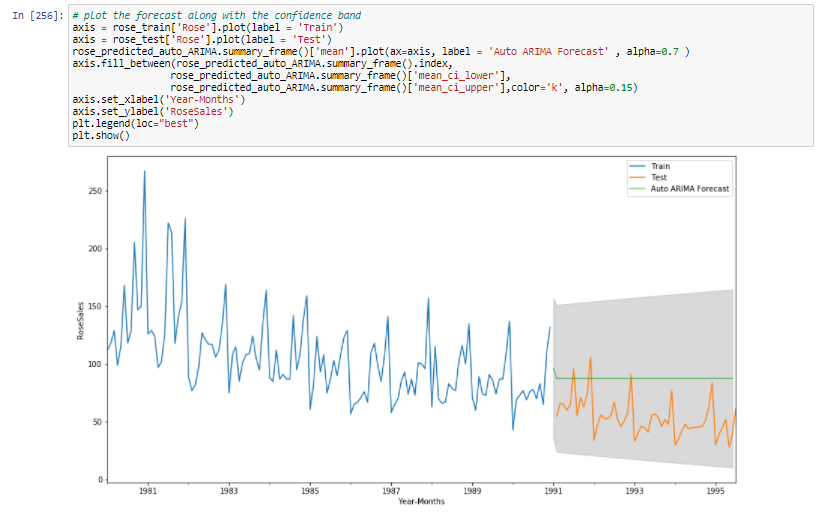
Jarque-Bera test is used for finding the Normality of residual. Looking at JB test p-val, we reject the Null Hypothesis, that means residual is not Normally distributed, which is reflected in Skew=0.82( normal range is between –0.5to0.5) & Kurtosis=5.13(usually it should be less than 3). The same is reflected in the histogram& Q-Q plot of the model diagnostic plot

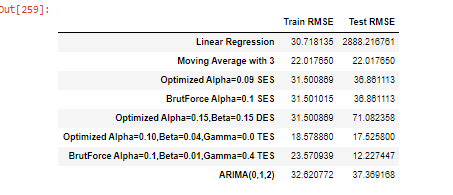
Heteroscedastic test shows the residuals are heteroscedastic.



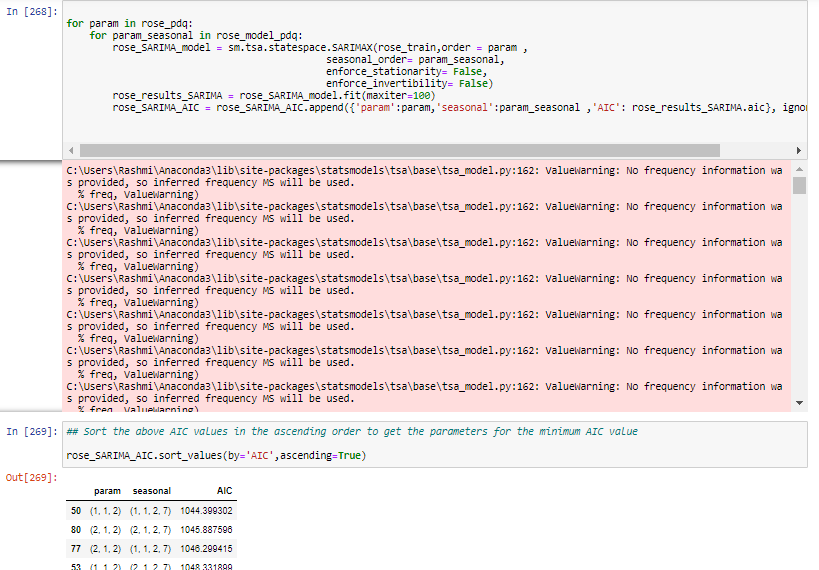


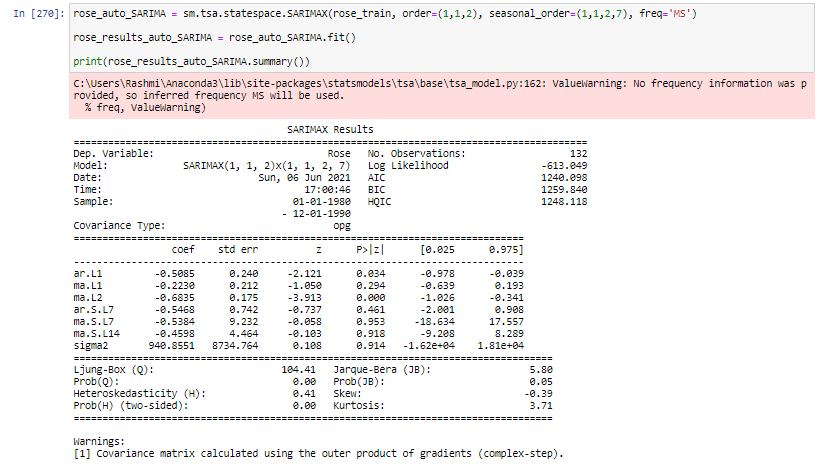
* Forecasting of Auto ARIMA for Rose wine sales TSD, with upper & lower confidence interval range is plotted as below

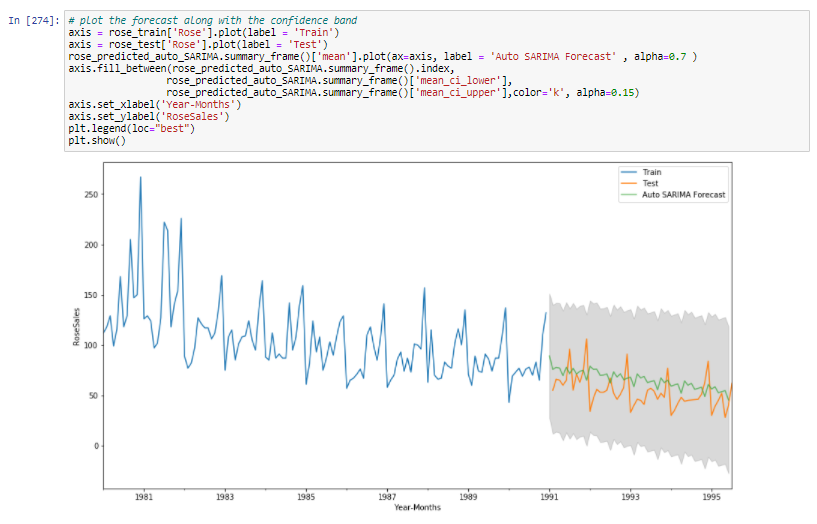


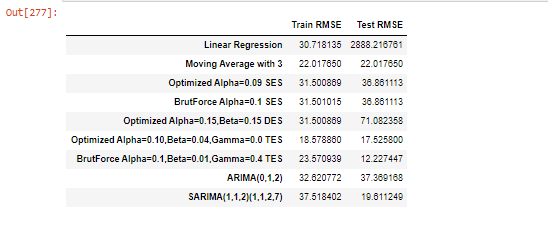


* Ran **SARIMA** on Rose wine sales TSD , by looping through Seasonal parameter – P,D,Q,M and p(AutoRegressor order) ,d(Difference order for stationarity),q(Moving Average order) parameters.
* For each and every combination of (p,d,q)(P,D,Q,M) SARIMA model is run and AIC is calculated.
* ‘p=1, d=1, q=2 & P=1, D=1, Q=2, M=7 gives the lowest SARIMA AIC for Rose wine sales



* Below is the model summary of SARIMA with lowest AIC for Rose wine sales TSD
* 
* Forecasting of Auto SARIMA for Rose wine sales TSD, with upper & lower confidence interval range is plotted as below

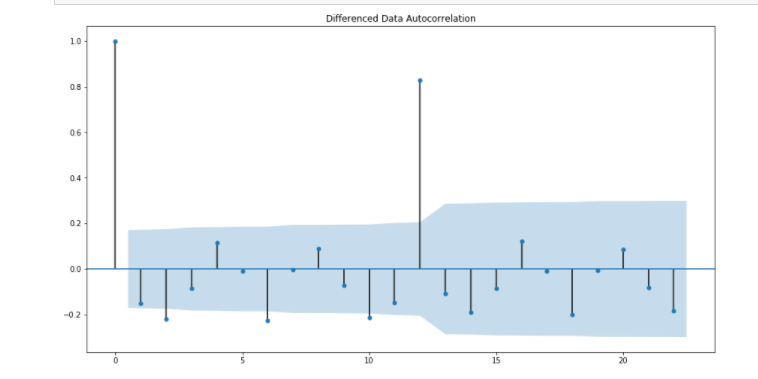




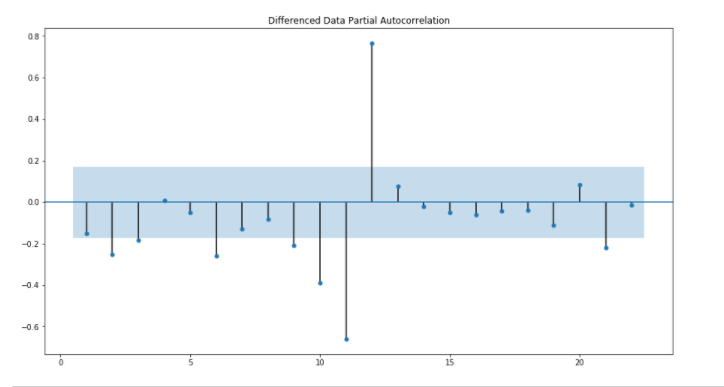
**Q.7** **Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

**A.7**

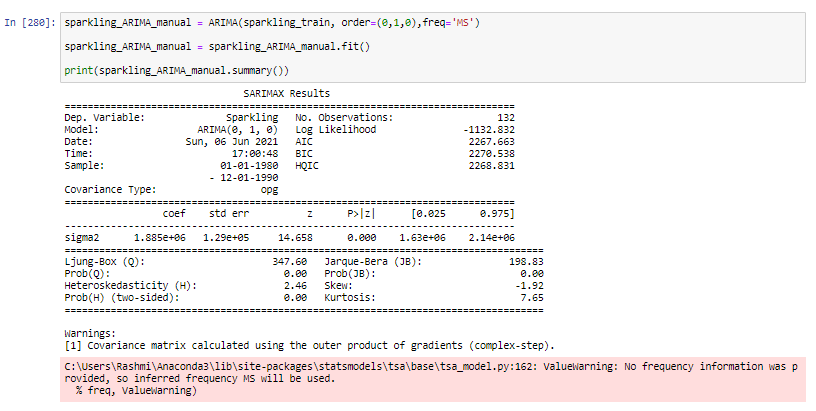
* **Sparkling wine monthly sales - Model building**
* Auto-Correlation plot for 1 differenced Sparkling wine sales TSD, which shows q=0 order is significant, as sequentially after order =0 , the next lag order id within the confidence blue region



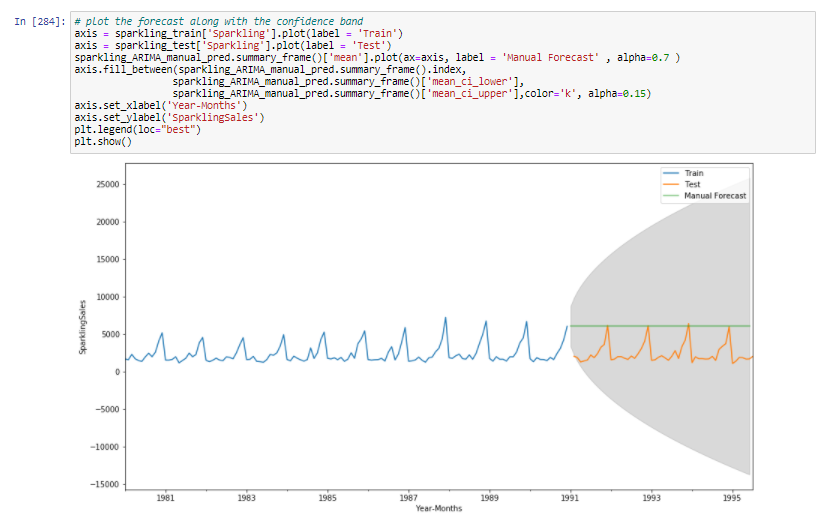
* Partial Auto-Correlation plot for 1 differenced Sparkling wine sale TSD, which shows p=0 order is significant, as sequentially after order =0 , the next lag order id within the confidence blue region

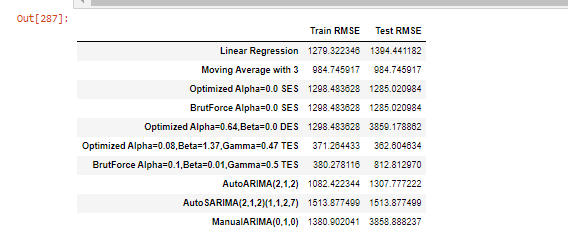


* Manual ARIMA model bulid summary based on ACF and PACF plot cut-off pts .

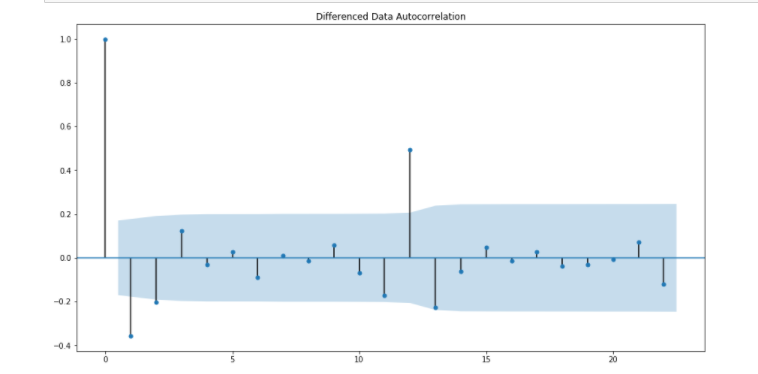


* Forecasting with the ARIMA model thus built, with upper and lower confidence interval is plotted as below

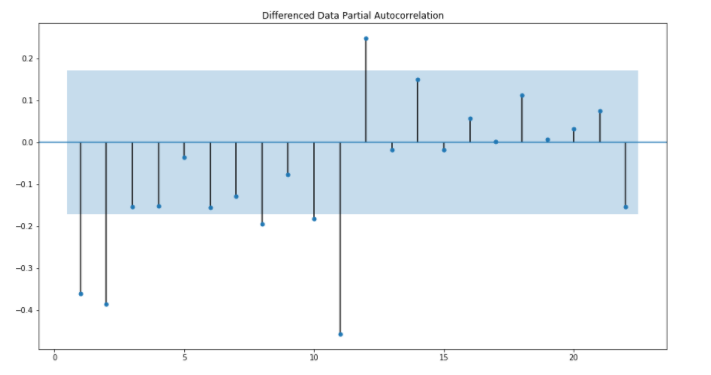




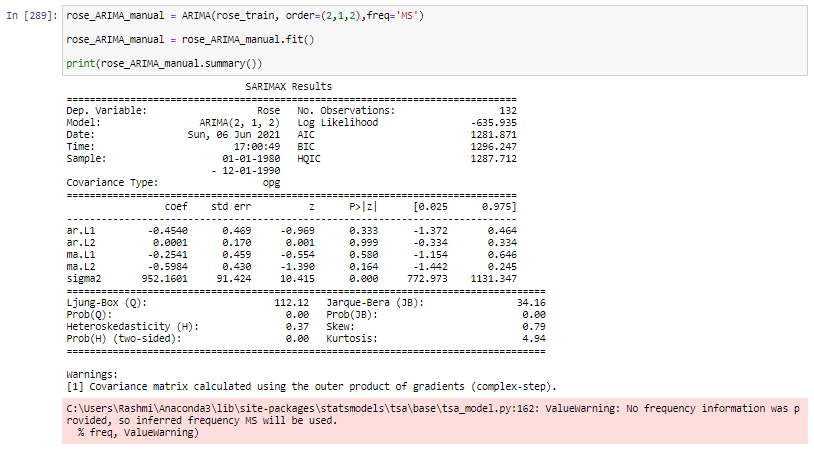
* **Rose wine monthly sales - Model building**
* Auto-Correlation plot for 1 differenced Rose wine sales TSD, which shows q=2 order is significant, as sequentially after order = 2, the next lag order id within the confidence blue region



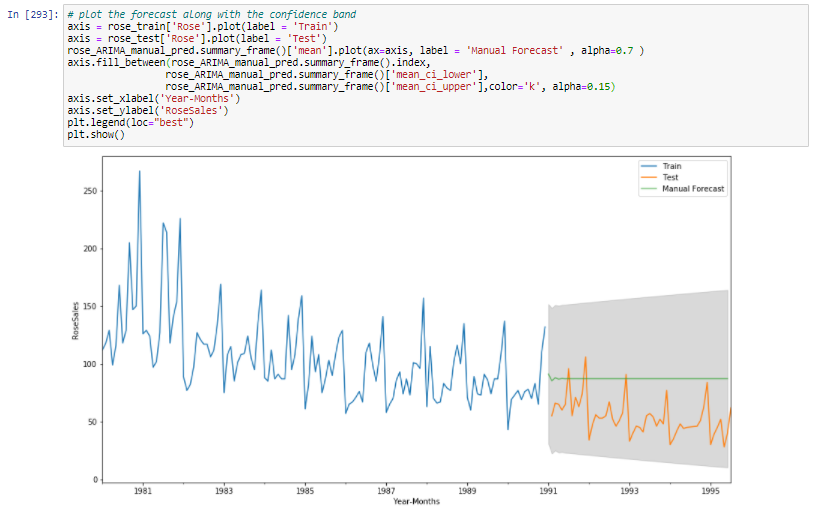
* Partial Auto-Correlation plot for 1 differenced Rose wine sale TSD, which shows p=2 order is significant, as sequentially after order = 2 , the next lag order id within the confidence blue region

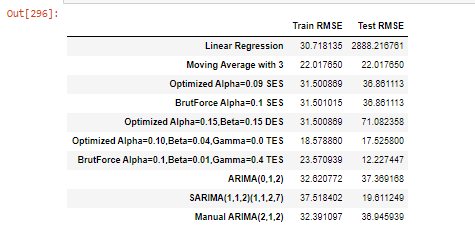


* Manual ARIMA model summary based on ACF and PACF plot cut-off pts .



* Forecasting with the ARIMA model thus built, with upper & lower confidence interval is plotted as below

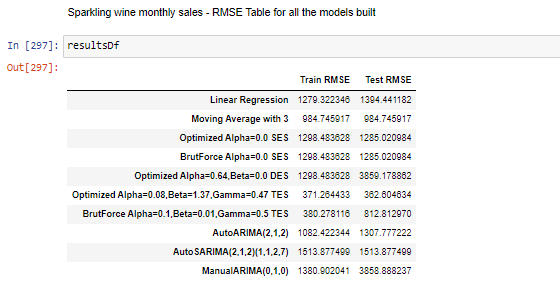




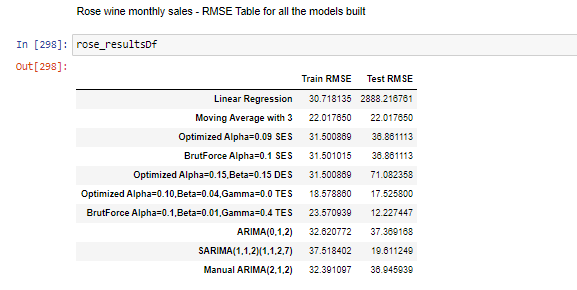
**Q.8 Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

**A.8**

* **Sparkling wine monthly sales - Model with RMSE of each**
* Below table reflects all the model build with their corresponding hyperparameters and train-test RMSE while model building for Sparkling wine dataset



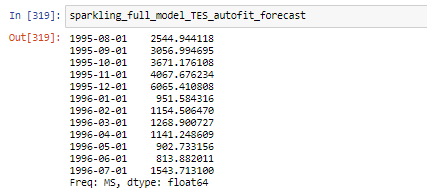
* **Rose wine monthly sales - Model with RMSE of each**
* Below table reflects all the model build with their corresponding hyperparameters and train-test RMSE while model building for Sparkling wine dataset

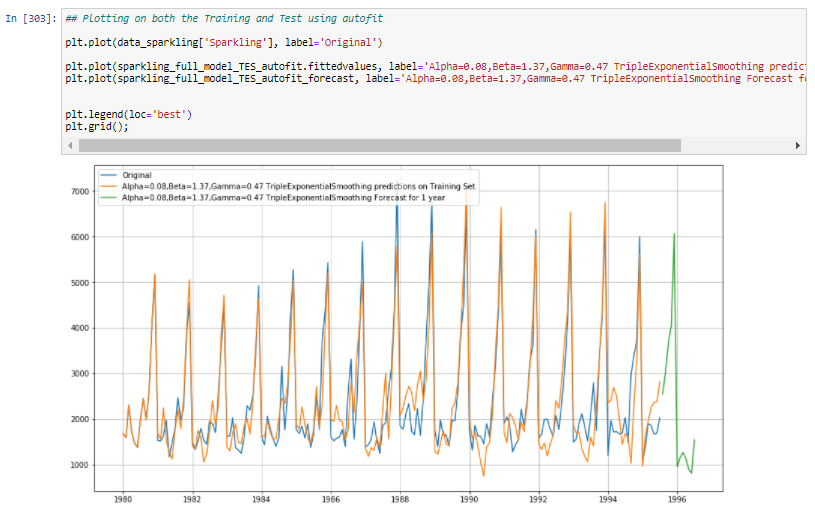


**Q.9 Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.**

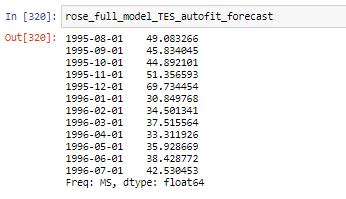
**A.9**

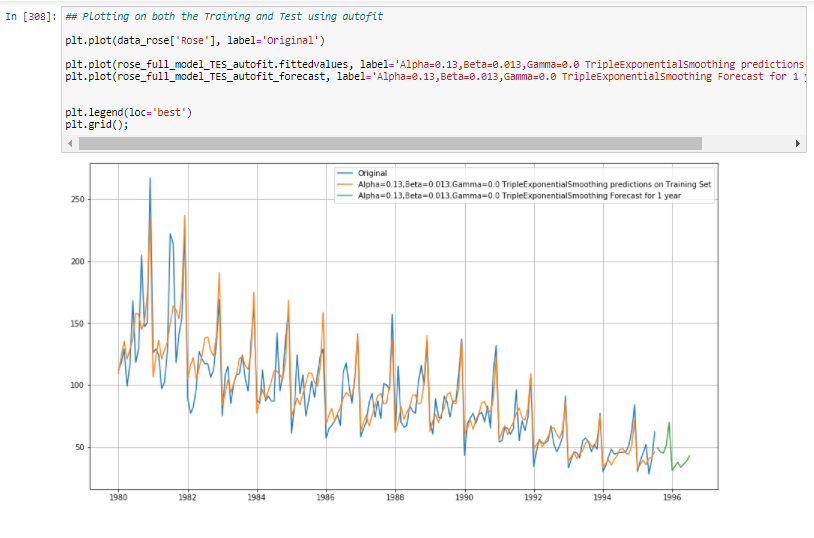
* **Sparkling wine monthly sales – Forecasting for next 12 months**
* Comparison of all the build model for Sparkling wine dataset, based on RMSE shows Optimized Triple Exponential smoothing gives the least RMSE of all & thus is used for final model building with all the dataset records & forecast for next 12 months is made.





* **Rose wine monthly sales – Forecasting for next 12 months**
* Comparison of all the build model for Rose wine dataset, based on RMSE shows Optimized Triple Exponential smoothing gives the least RMSE of all & thus is used for final model building with all the dataset records & forecast for next 12 months is made.

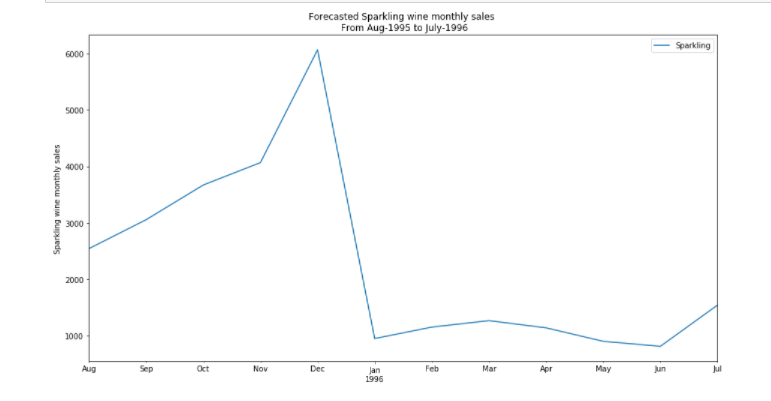




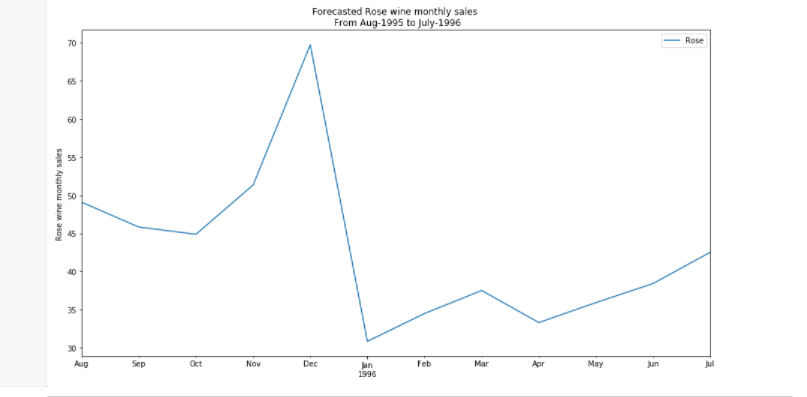
**Q.10 Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**

**A.10**

* **Sparkling wine monthly sales – Findings**
* Finally, we build the optimized TES model on the given Sparkling wines dataset to forecast for the next 12 months in future.
* The sales of Sparkling wines is in the thousands range, which is showing flat trend in the given 15 yrs
* The highest peak sale of the wine is in the month of December every year.
* The seasonal component is additive in nature here
* Below is the forecast of sale for next 12 months



* **Rose wine monthly sales – Findings**
* Finally, we build the optimized TES model on the given Rose wines dataset to forecast for the next 12 months in future.
* The sales of Rose wines is in the hundreds range, which is showing downward trend in the given 15 yrs.
* The highest peak sale of the wine is in the month of December every year
* The seasonal component is multiplicative in nature here
* Below is the forecast of sale for next 12 months



* **Sparkling - Rose wine sales – Relative forcasted monthly sale**
* Both Sparkling & Rose wine sales show large peak in the month Nov – Dec.
* Both Sparkling & Rose wine sales show slight peak in the month March.
* Sparkling monthly sales is in the range of thousands while Rose monthly sales is in the range of hundreds.
* Rose wine seems to be more expensive hence the sales seem to be fair for rose wine, as common people might not be buying expensive stuff much. Also the trend for rose sale shows downward sale, that means the sale is going down across the year.
* Company might to come up with bundle sell of rose wine with Sparkling wine(or other heavy selling wine product), for people to attract towards Rose wine, which might increase potential buyers of Rose wine & sale of Rose wine can increase.
* Sparkling on the other hand have heavy sales which is constant across the years. Although, company also might needs to look at increasing the sparkling sales to upward the sale trend & hence profit.

