September 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Time Series Forecasting

Business Report

Thakur Arun Singh

Contents

[Problem 1: 2](#_Toc81771869)

[Problem 1.1 2](#_Toc81771870)

[Problem 1.2 4](#_Toc81771871)

[Problem 1.3 11](#_Toc81771872)

[Problem 1.4 12](#_Toc81771873)

[Problem 1.5 18](#_Toc81771874)

[Problem 1.6 25](#_Toc81771875)

[Problem 1.7 31](#_Toc81771876)

[Problem 1.8 36](#_Toc81771877)

[Problem 1.9 37](#_Toc81771878)

[Problem 1.10 40](#_Toc81771879)

## Problem 1:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze and forecast Wine Sales in the 20th century..

### Problem 1.1

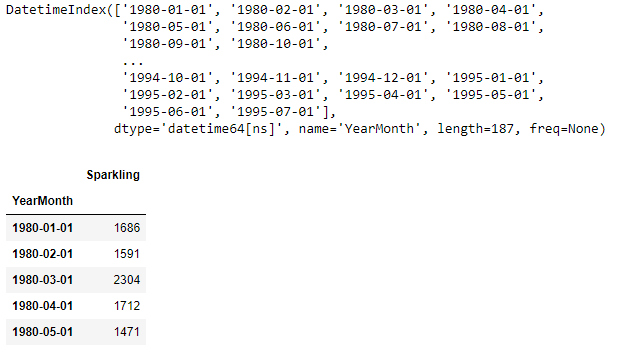
Read the data as an appropriate Time Series data and plot the data.

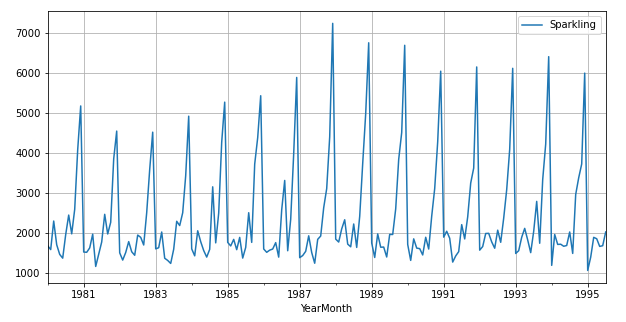
**Resolution:**

First, we import all the necessary libraries seaborn, numpy, pandas, sklearn etc to perform our analysis

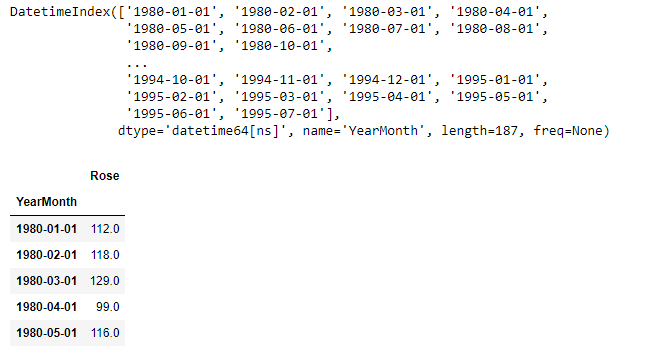
Next, we import the data set “Sparkling” and “Rose”

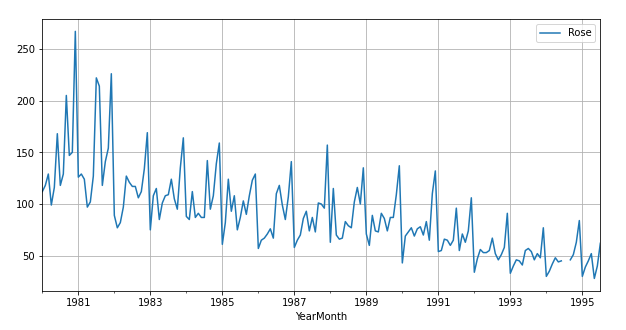
**Sparkling Dataset**

****

****

**Rose Dataset**

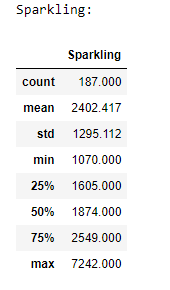
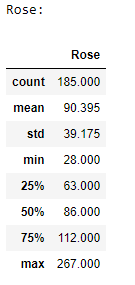


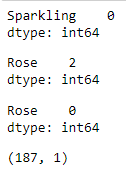


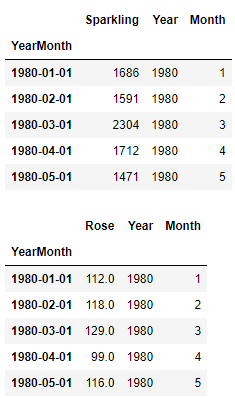
### Problem 1.2

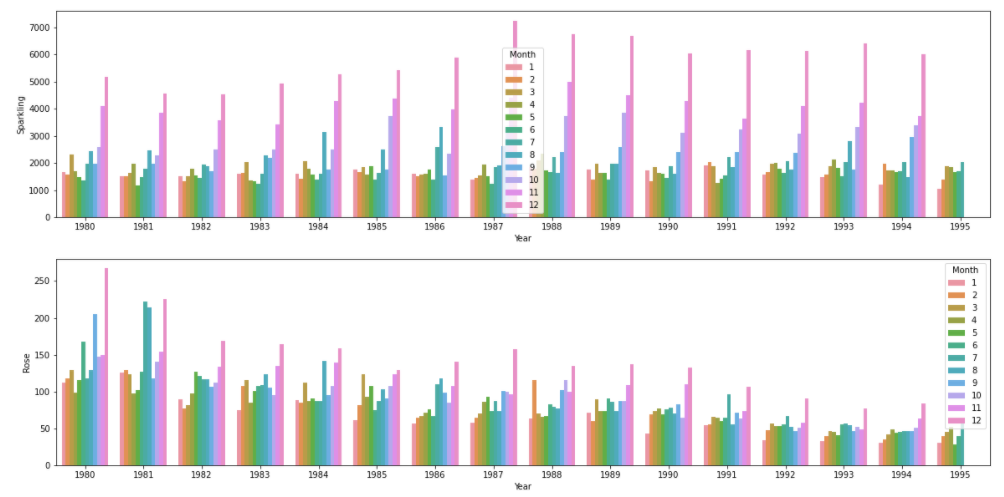
Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

**Resolution:**

** **





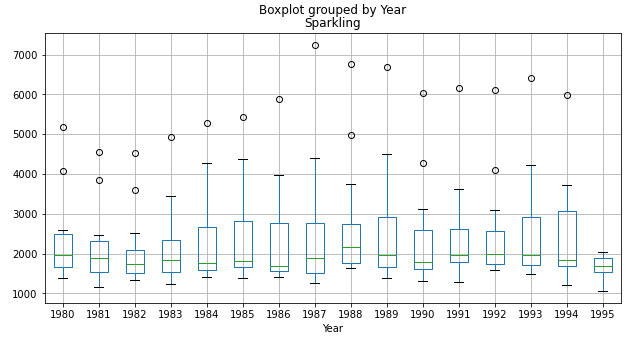


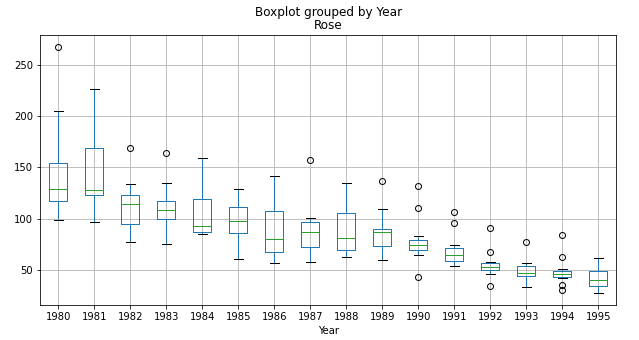
Below Pivot shows the sales made for a month in particular year:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sparkling** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Month** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **Year** |  |  |  |  |  |  |  |  |  |  |  |  |
| **1980** | 1686 | 1591 | 2304 | 1712 | 1471 | 1377 | 1966 | 2453 | 1984 | 2596 | 4087 | 5179 |
| **1981** | 1530 | 1523 | 1633 | 1976 | 1170 | 1480 | 1781 | 2472 | 1981 | 2273 | 3857 | 4551 |
| **1982** | 1510 | 1329 | 1518 | 1790 | 1537 | 1449 | 1954 | 1897 | 1706 | 2514 | 3593 | 4524 |
| **1983** | 1609 | 1638 | 2030 | 1375 | 1320 | 1245 | 1600 | 2298 | 2191 | 2511 | 3440 | 4923 |
| **1984** | 1609 | 1435 | 2061 | 1789 | 1567 | 1404 | 1597 | 3159 | 1759 | 2504 | 4273 | 5274 |
| **1985** | 1771 | 1682 | 1846 | 1589 | 1896 | 1379 | 1645 | 2512 | 1771 | 3727 | 4388 | 5434 |
| **1986** | 1606 | 1523 | 1577 | 1605 | 1765 | 1403 | 2584 | 3318 | 1562 | 2349 | 3987 | 5891 |
| **1987** | 1389 | 1442 | 1548 | 1935 | 1518 | 1250 | 1847 | 1930 | 2638 | 3114 | 4405 | 7242 |
| **1988** | 1853 | 1779 | 2108 | 2336 | 1728 | 1661 | 2230 | 1645 | 2421 | 3740 | 4988 | 6757 |
| **1989** | 1757 | 1394 | 1982 | 1650 | 1654 | 1406 | 1971 | 1968 | 2608 | 3845 | 4514 | 6694 |
| **1990** | 1720 | 1321 | 1859 | 1628 | 1615 | 1457 | 1899 | 1605 | 2424 | 3116 | 4286 | 6047 |
| **1991** | 1902 | 2049 | 1874 | 1279 | 1432 | 1540 | 2214 | 1857 | 2408 | 3252 | 3627 | 6153 |
| **1992** | 1577 | 1667 | 1993 | 1997 | 1783 | 1625 | 2076 | 1773 | 2377 | 3088 | 4096 | 6119 |
| **1993** | 1494 | 1564 | 1898 | 2121 | 1831 | 1515 | 2048 | 2795 | 1749 | 3339 | 4227 | 6410 |
| **1994** | 1197 | 1968 | 1720 | 1725 | 1674 | 1693 | 2031 | 1495 | 2968 | 3385 | 3729 | 5999 |
| **1995** | 1070 | 1402 | 1897 | 1862 | 1670 | 1688 | 2031 | NaN | NaN | NaN | NaN | NaN |

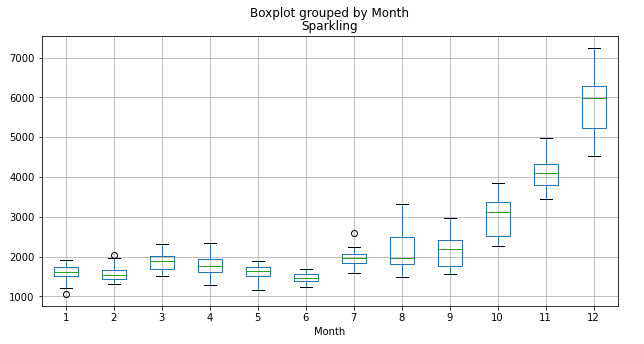
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Rose** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Month** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **Year** |  |  |  |  |  |  |  |  |  |  |  |  |
| **1980** | 112 | 118 | 129 | 99 | 116 | 168 | 118 | 129 | 205 | 147 | 150 | 267 |
| **1981** | 126 | 129 | 124 | 97 | 102 | 127 | 222 | 214 | 118 | 141 | 154 | 226 |
| **1982** | 89 | 77 | 82 | 97 | 127 | 121 | 117 | 117 | 106 | 112 | 134 | 169 |
| **1983** | 75 | 108 | 115 | 85 | 101 | 108 | 109 | 124 | 105 | 95 | 135 | 164 |
| **1984** | 88 | 85 | 112 | 87 | 91 | 87 | 87 | 142 | 95 | 108 | 139 | 159 |
| **1985** | 61 | 82 | 124 | 93 | 108 | 75 | 87 | 103 | 90 | 108 | 123 | 129 |
| **1986** | 57 | 65 | 67 | 71 | 76 | 67 | 110 | 118 | 99 | 85 | 107 | 141 |
| **1987** | 58 | 65 | 70 | 86 | 93 | 74 | 87 | 73 | 101 | 100 | 96 | 157 |
| **1988** | 63 | 115 | 70 | 66 | 67 | 83 | 79 | 77 | 102 | 116 | 100 | 135 |
| **1989** | 71 | 60 | 89 | 74 | 73 | 91 | 86 | 74 | 87 | 87 | 109 | 137 |
| **1990** | 43 | 69 | 73 | 77 | 69 | 76 | 78 | 70 | 83 | 65 | 110 | 132 |
| **1991** | 54 | 55 | 66 | 65 | 60 | 65 | 96 | 55 | 71 | 63 | 74 | 106 |
| **1992** | 34 | 47 | 56 | 53 | 53 | 55 | 67 | 52 | 46 | 51 | 58 | 91 |
| **1993** | 33 | 40 | 46 | 45 | 41 | 55 | 57 | 54 | 46 | 52 | 48 | 77 |
| **1994** | 30 | 35 | 42 | 48 | 44 | 45 | 46 | 46 | 46 | 51 | 63 | 84 |
| **1995** | 30 | 39 | 45 | 52 | 28 | 40 | 62 | NaN | NaN | NaN | NaN | NaN |

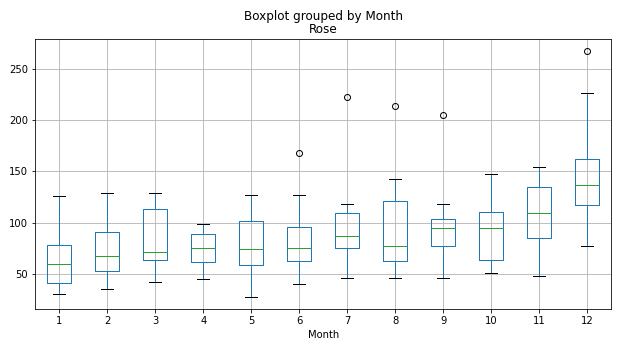
**Yearly Boxplots**

****

****

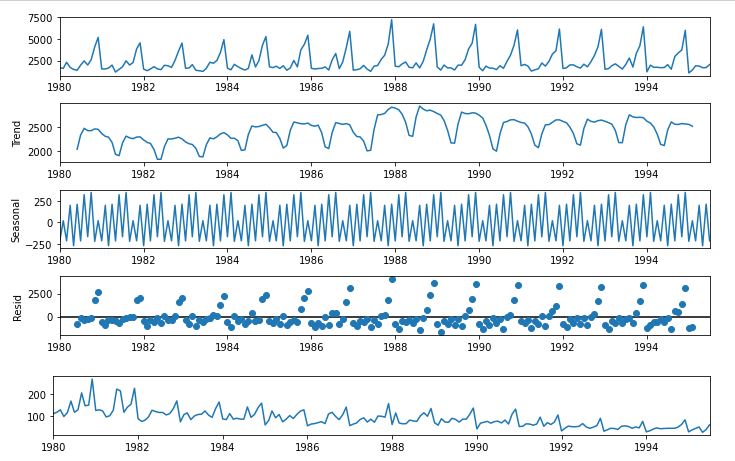
**Monthly Boxplots**

****

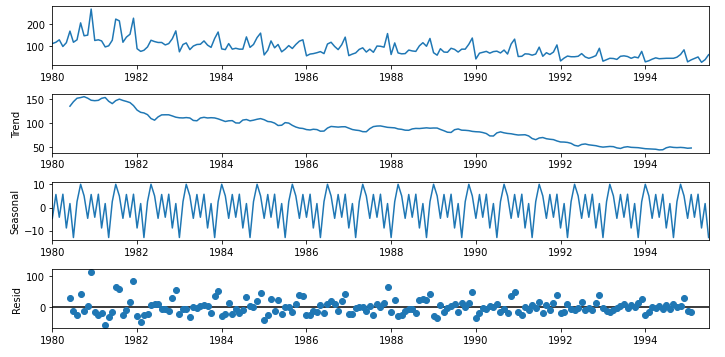
****

**Additive Decomposition:**

**Sparkling:**

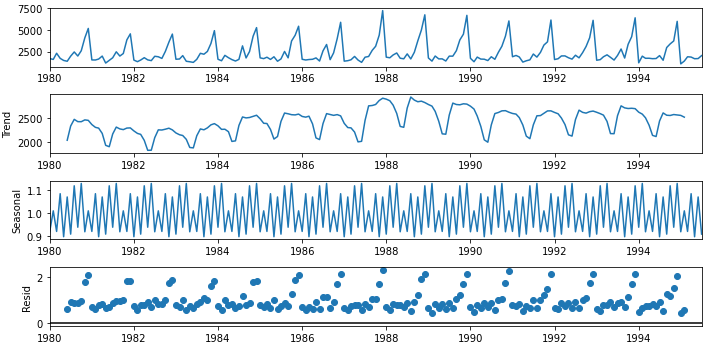
****

**Rose:**

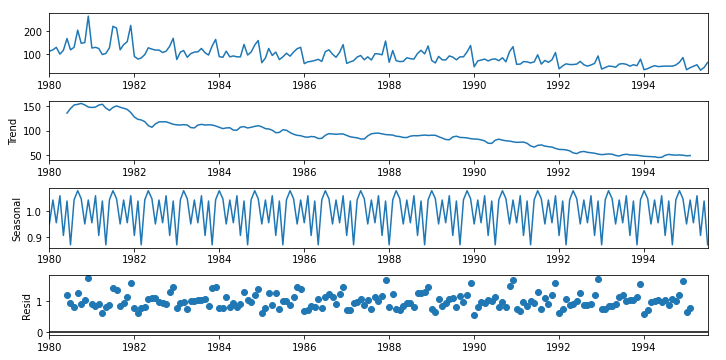
****

**Multiplicative:**

**Sparkling:**

****

**Rose:**

****

**Summary Sparkling Dataset:**

* Sparkling dataset doesn’t show a visible trend however it shows seasonality, also if observed from additive decomposition the residual is catching some pattern.
* Multiplicative decomposition on the other hand seems to dictate on the series as the scale of the residual plot had decreased considerably
* Monthly bar plots showed that the sales are higher towards the last months than the earlier.

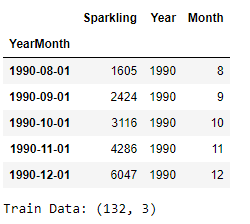
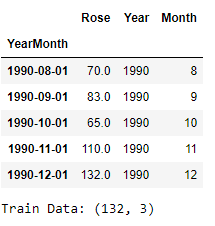
**Summary Rose Dataset:**

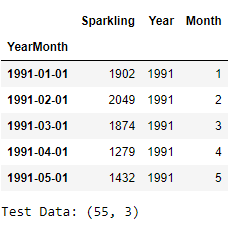
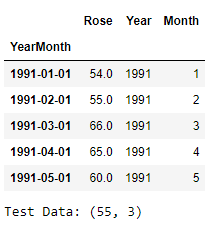
* Rose dataset show a clear decreasing trend as well as seasonality, multiplicative decomposition
* dictates the series the noise is reduced considerably in it also the seasonal patterns increase and decrease in the size across difference years
* The sales tend to go up during the July-August and also during end of the year.

### Problem 1.3

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

**Resolution:**

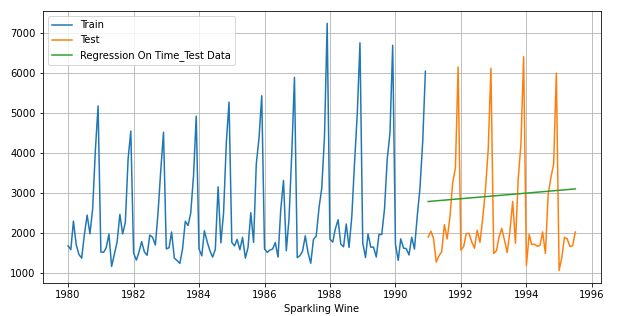
### Problem 1.4

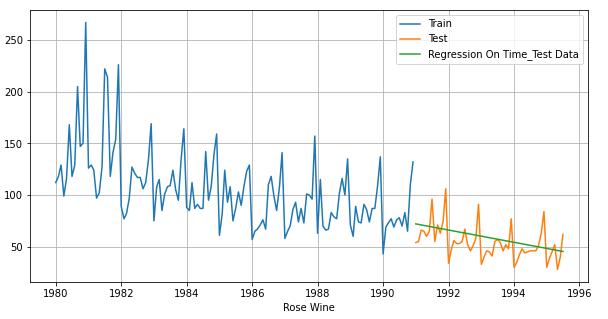
Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.

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

**Resolution:**

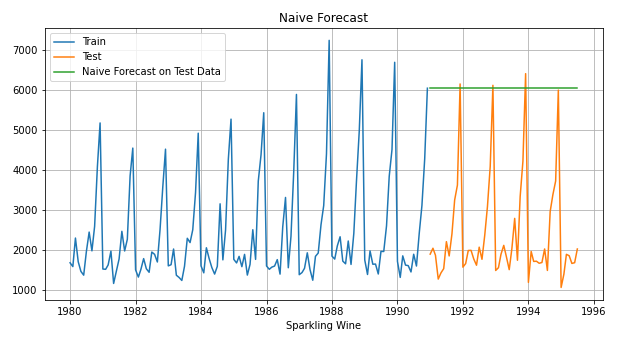
**Model 1: Linear Regression: ŷ t+1 = β y + c**

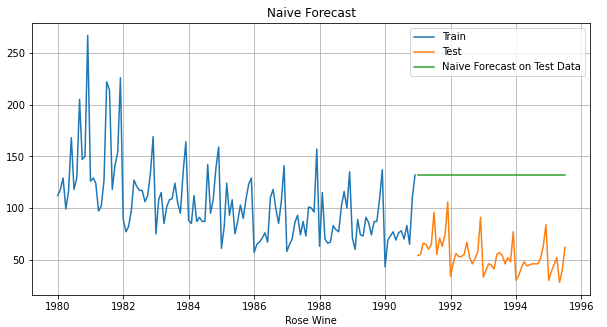
****

****

**Model 2: Naive Approach: ŷ t+1=yt**

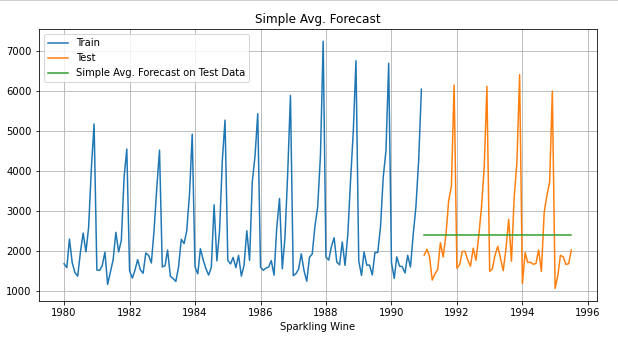
For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

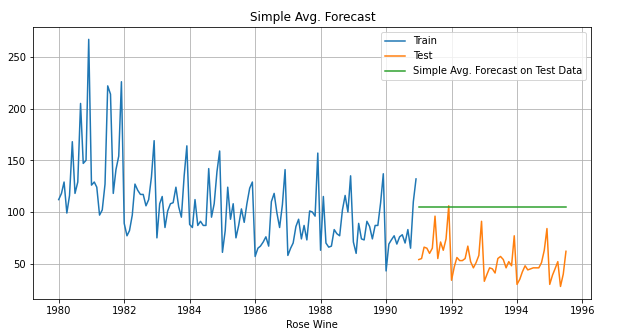




**Method 3: Simple Average:**

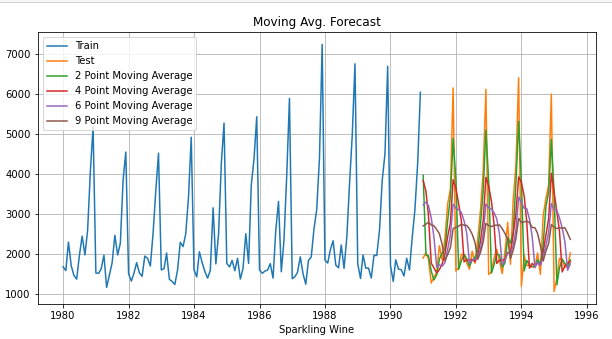
For this particular simple average method, we will forecast by using the average of the training values.

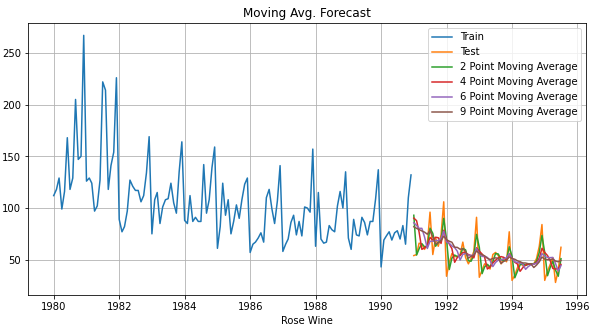




**Method 4: Moving Average(MA)**

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 minimum error. The below plot shows the forecast for different rolling means:





**Method 5: Exponential Smoothing methods**

Exponential smoothing methods consist of flattening time series data. Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations.

**Simple Exponential Smoothing (SES):** The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES). This method is suitable for forecasting data with no clear trend or seasonal pattern. In Single ES, the forecast at time (t + 1) is given by Winters, 1960

ŷt+1=αYt+(1−α)ŷt Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

Sparkling data doesn't show visible trend however it shows seasonality, Rose data on the other hand shows both trend and seasonality, all the Exponential models will still be built on both the datasets.

**Double Exponential Smoothing(DES):** One of the drawbacks of the simple exponential smoothing is that the model does not do well in the presence of the trend. This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters. Applicable when data has Trend but no seasonality. Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series A second smoothing parameter β corresponds to the trend series. Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short term average value or level and the other for capturing the trend.

Intercept or Level equation, ŷt is given by: ŷt=αyt+(1−α)ŷt Trend equation is given by

Tt=β(ŷt−ŷt−1)+(1−β)Tt−1 Here, α and β are the smoothing constants for level and trend, respectively,

0 <α < 1 and 0 < β < 1.

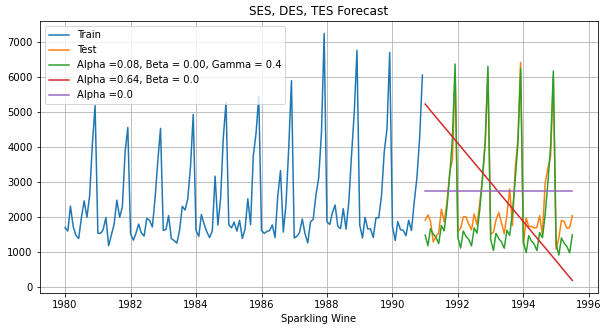
The forecast at time t + 1 is given by

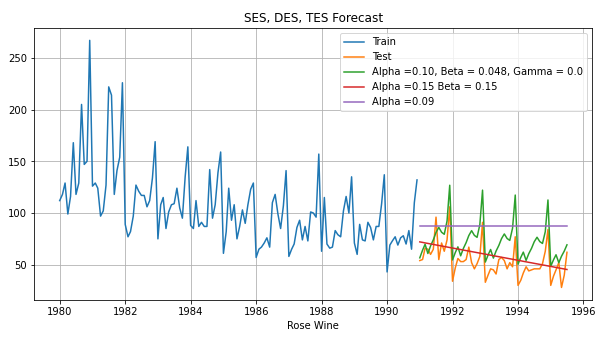
Ft+1=ŷt+Tt Ft+n=ŷt+nTt

Though our Sparkling data doesn't seem to have a visible trend we are still going to build this model for the project. Rose data has a clear trend from the plot above

**Inference**

* Here, we see that the Double Exponential Smoothing model has picked up the trend component as well (see the below fig.)
* Our data has seasonality too so we will include one more smoothing parameter for seasonality which is gamma.
* We will use ETS (A, A, A) Holt Winter's linear method with additive trend and seasonality for Sparkling data and ETS (A, A, M) Holt Winter's linear method with additive trend and multiplicative seasonality for Rose wine data. We will call it Triple Exponential Smoothing (TES)





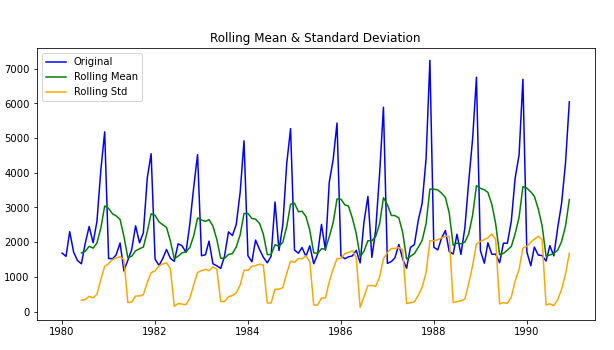
### Problem 1.5

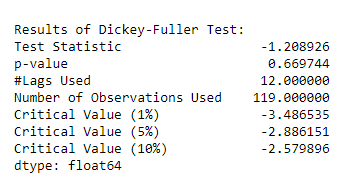
Check for the stationary of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationary and comment.

Note: Stationary should be checked at alpha = 0.05.

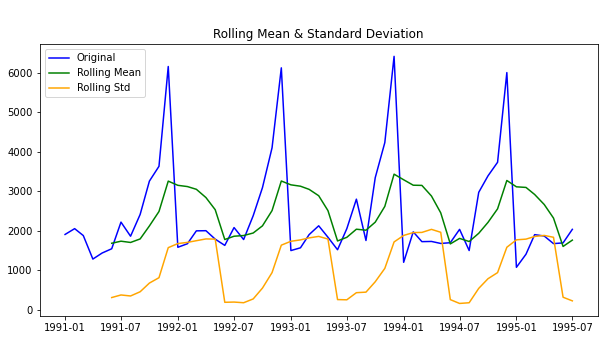
**Resolution:**

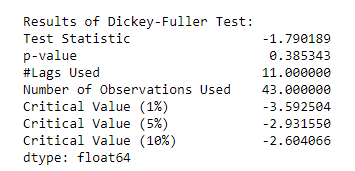
**Sparkling Train set:**

****

****

**Sparkling Test set:**

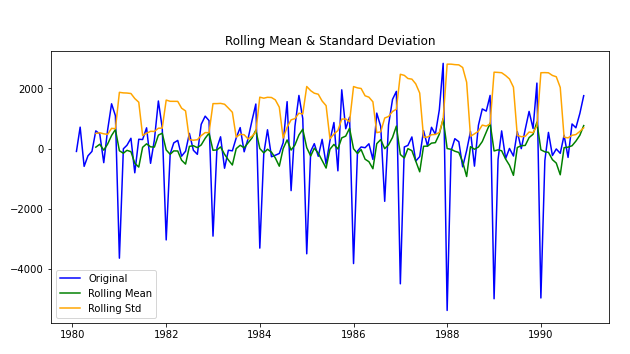
****

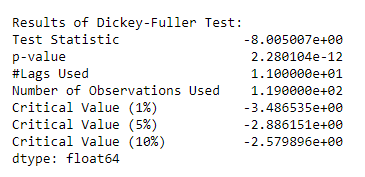
****

Since the Null Hypothesis H0 : The series is non-stationary Alternate Hypothesis H1: The series is stationary

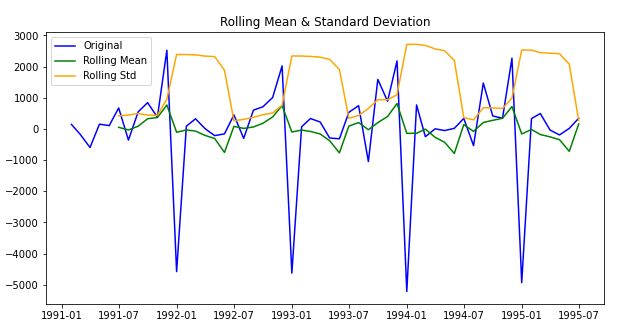
We cannot reject the null as the p values for both of series is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above

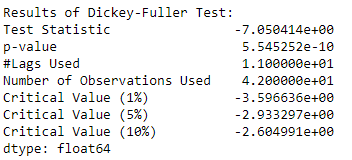
**Differenced Sparkling Train set:**

****

****

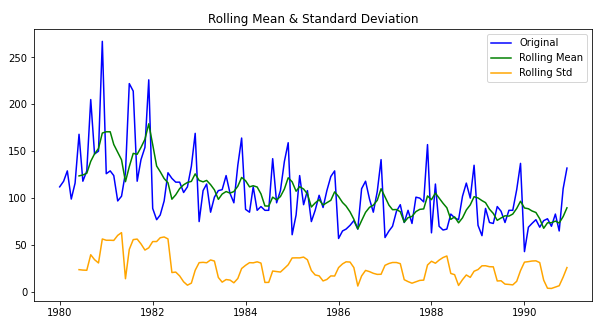
**Differenced Sparkling Test set:**

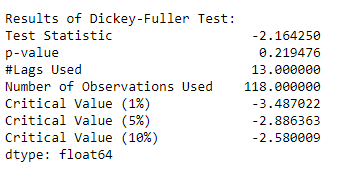
****

****

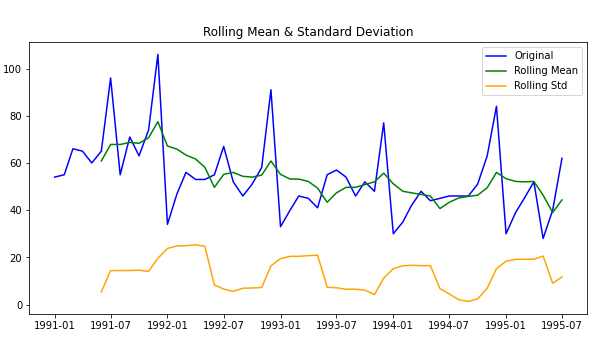
We can now see that the p –value < than 0.05 so we can reject the null-hypothesis and accept the alternate. So we say the series is stationary.

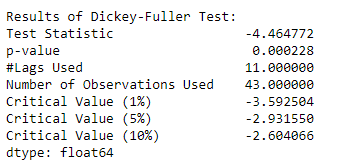
**Rose Train Set:**

****

****

**Rose Test Set:**

****

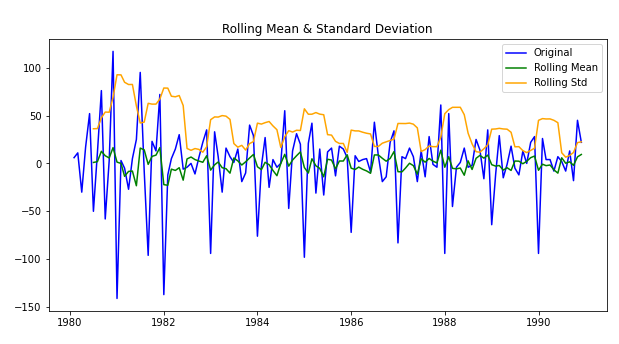
****

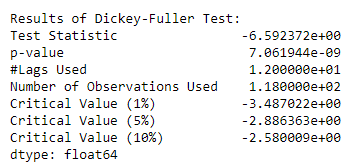
Since the Null Hypothesis H0: The series is non-stationary Alternate Hypothesis H1: The series is stationary we cannot reject the null as the p values is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above Train set of Rose Wine dataset, on the contrary we can reject the null as the p values is less than 0.05 (significance level) from the Augmented Dickey Fuller test above Test set of Rose Wine dataset

We can correct the non-stationary by using multiple methods like taking differences at various level, using logged transformed series etc.

Here we will take difference of level 1 of the original train series and we will use the train dataset as is.

**Differenced Rose Train set:**

****

****

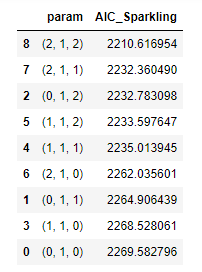
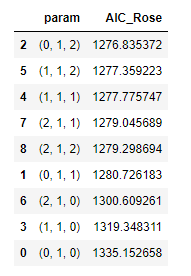
### Problem 1.6

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.

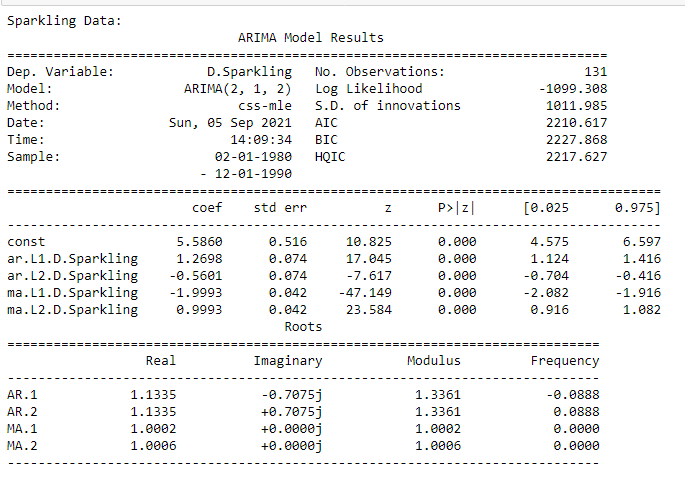
**Resolution:**

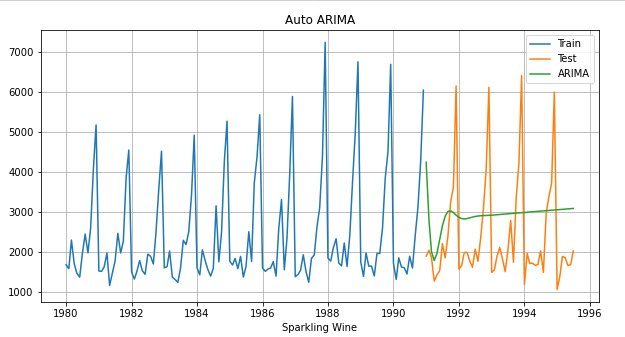
**ARIMA**

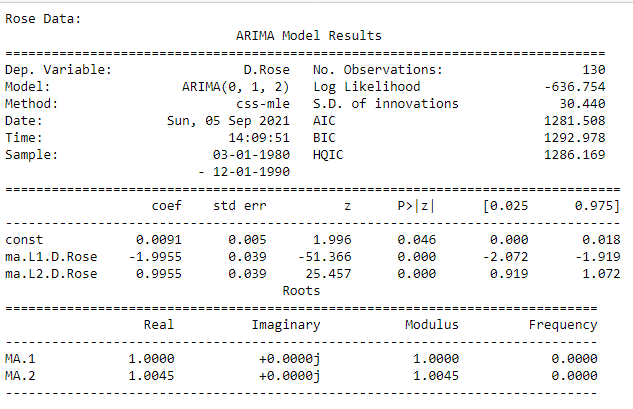
AIC score for both Sparkling and Rose wine dataset for different models is below:

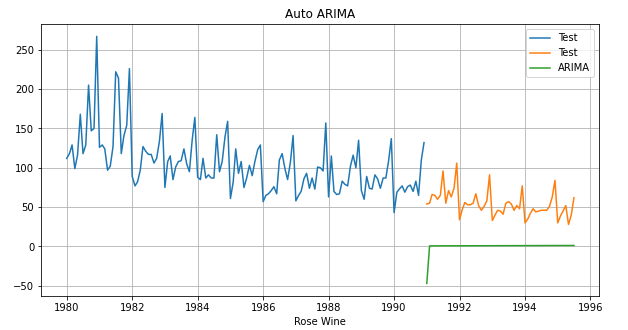
 

An automated model of (2,1,2) will be built on sparkling wine data and (0,1,2) on rose wine data. Both are of difference order 1.



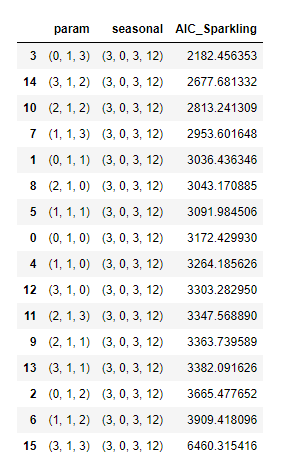
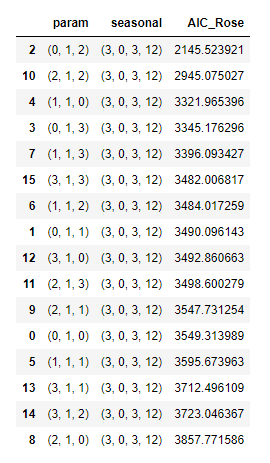






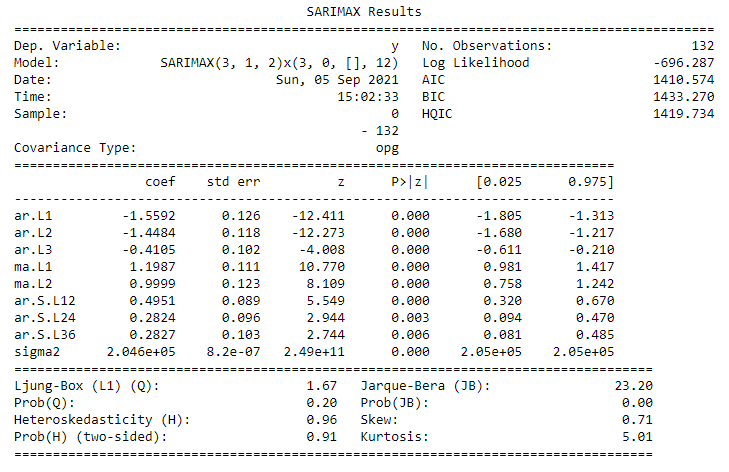
From the ACF plot we see a significant seasonal correlation after every 11th interval Setting the seasonality as 12 for the first iteration of the auto SARIMA model.

AIC scores for SARIMAX model

**** ****

An automated SARIMA model of (3,1,2) will be built on sparkling wine data and (3,1,1) on rose wine data. both are of difference order 1 and seasonality 12.

**Sparkling Data:**

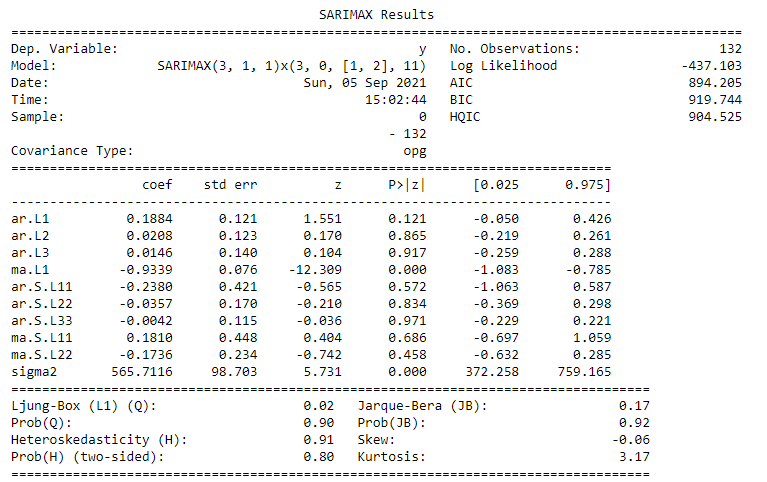
****

Note:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

[2] Covariance matrix is singular or near-singular, with condition number 2.3e+26. Standard errors may be unstable.

**Rose Data:**

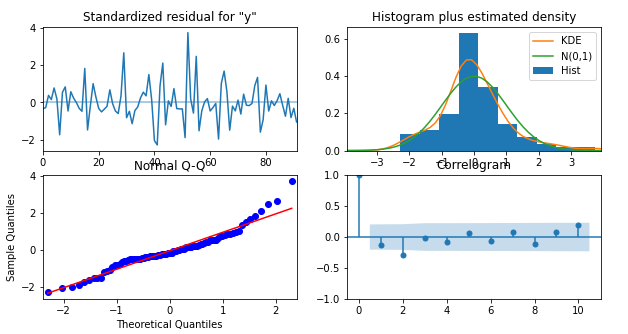
****

Note:

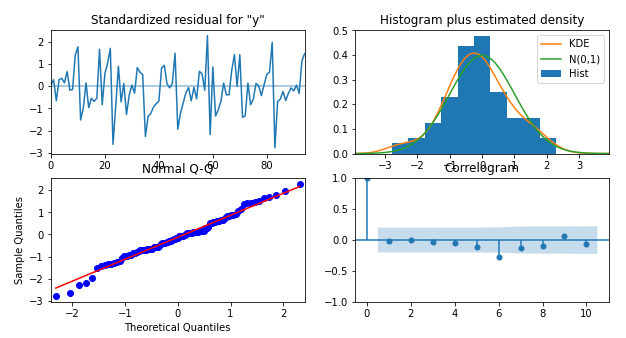
[1] Covariance matrix calculated using the outer product of gradients (complex-step)

**Diagnostic plots for Auto SARIMA model are as below:**

**Sparkling Data:**

****

**Rose Data:**

****

**Sparkling Dataset Diagnostic:**

From the diagnostic plots we see that the assumptions of Normality, heteroscedasticity as seems to be getting satisfied as well the series show randomness and no auto correlation between the residuals

**Rose Dataset Diagnostic:**

The plot shows randomness of the residual also the assumption of normality and heteroscedasticity is satisfied, it shows no auto correlation until lag 5, then shows a rise in significance at 6.

Though visual plots satisfy most assumptions the test proves it wrong seen from the summary of SARIMAX model for both the dataset.

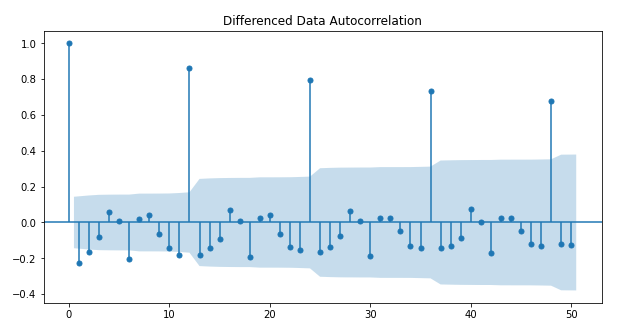
### Problem 1.7

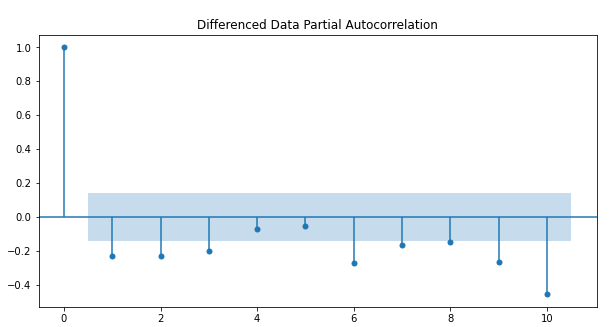
Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

**Resolution:**

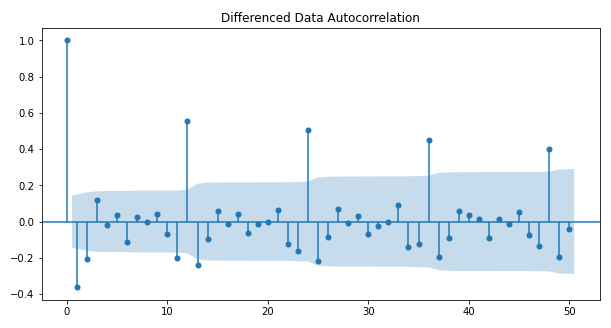
**ARIMA**

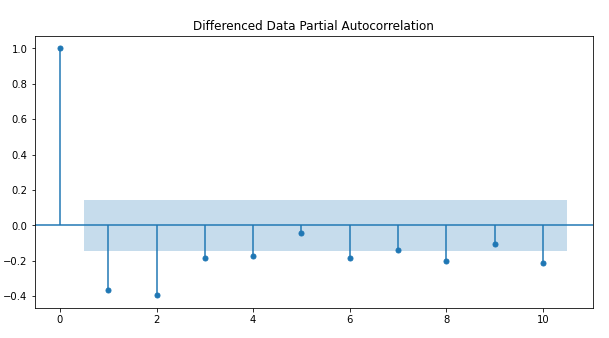
**Sparkling Dataset:**

****

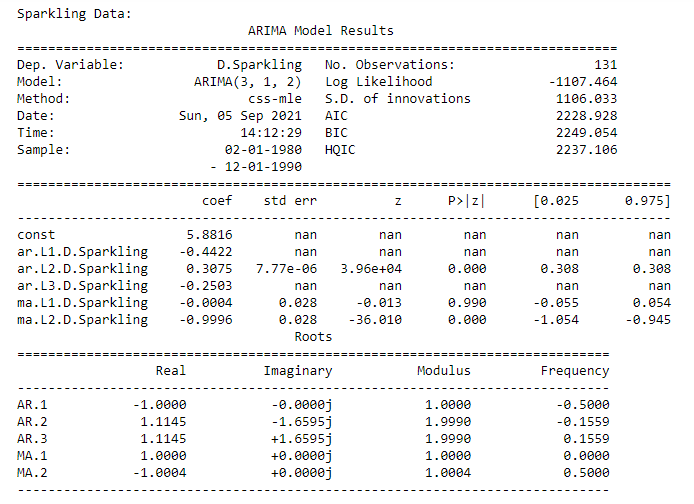
****

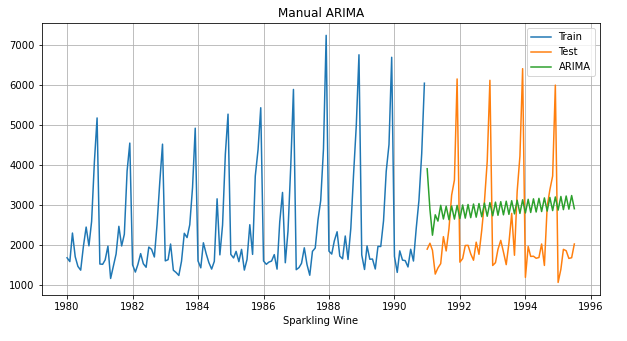
**Rose Dataset:**

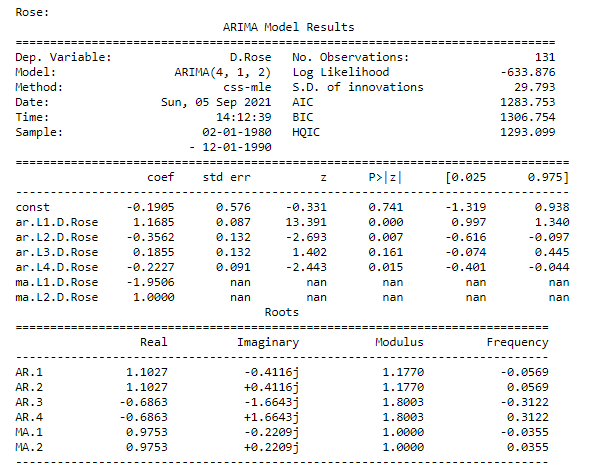
****

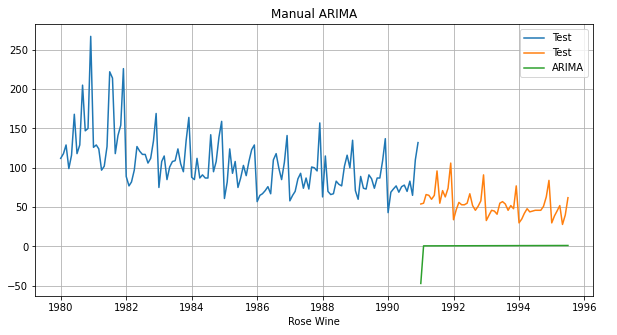
****

* 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 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
* By looking at the above plots for Sparkling data, we can say that both the PACF cuts off at 3 and ACF plot cuts-off at lag 2.
* By looking at the above plots for Rose data, we can say that PACF cuts off at 4 and ACF plot cuts-off at lag 2.







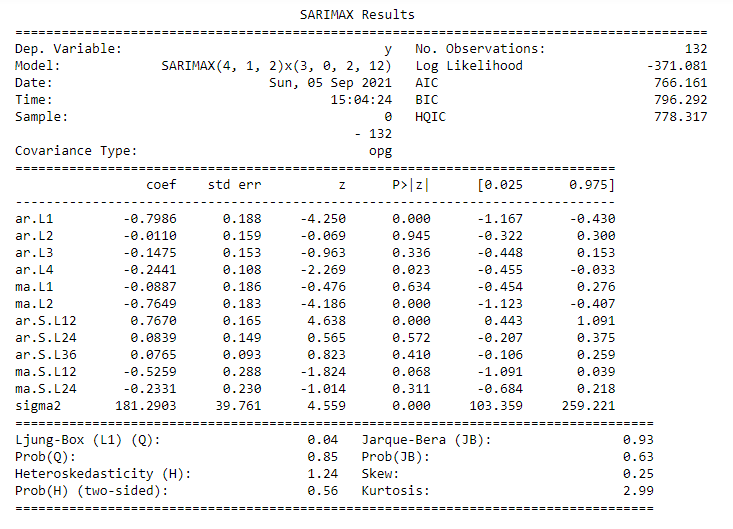


AIC for sparkling data is the lowest for the model (3,1,2), also we saw the from ACF and PACG plots that the cut off of p and q are at 3 and 2 resp. so we conclude that the auto SARIMAX and the manual SARIMAX models are the same.

**SARIMA**

For Rose data let's build a model at the p and q cut off at 4, 2 respectively.

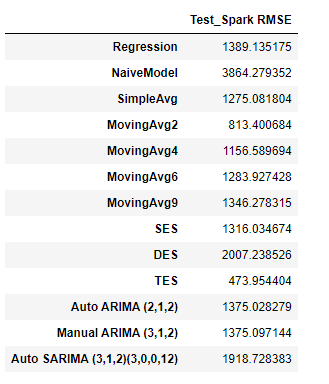
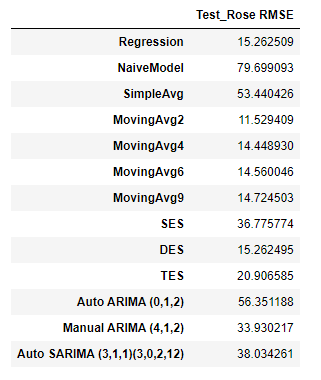
**Manual SARIMAX Summary on Rose data:**



### Problem 1.8

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

**Resolution:**

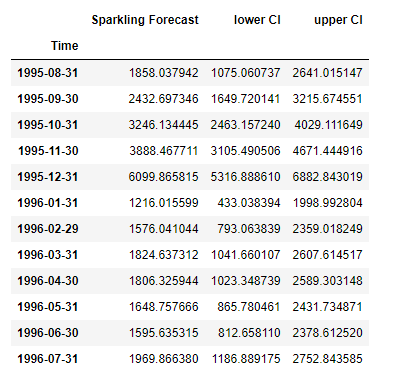
 

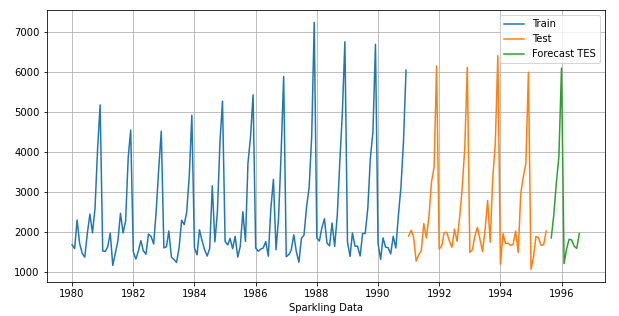
### Problem 1.9

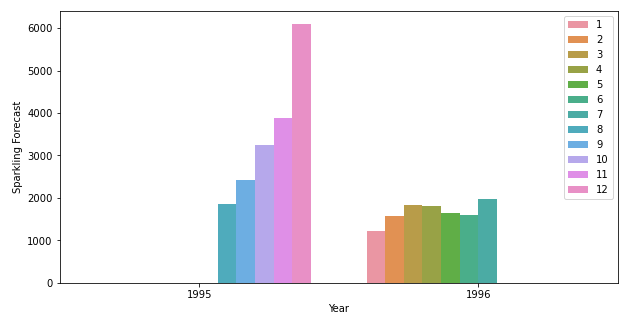
Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

**Resolution:**

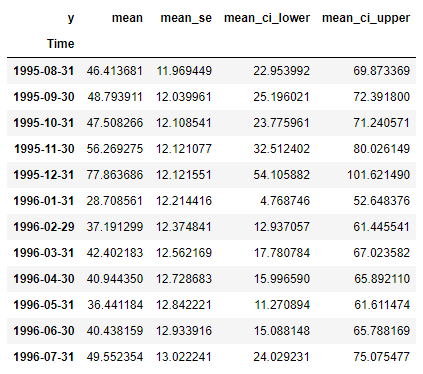
For Sparkling dataset, we see that Triple Exponential smoothing gives the best forecast, so we will move forward with that for forecasting

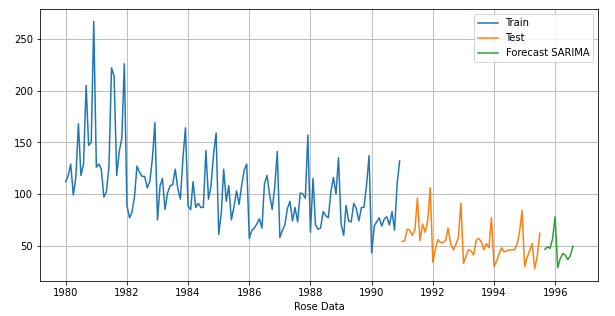


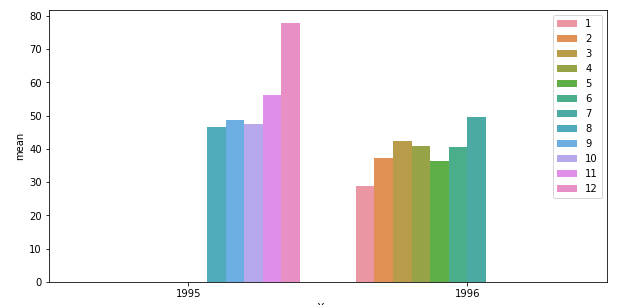




For Rose dataset rolling avg shows the best RMSE, however since the window chosen was very small(2,4,6,9) it was natural it was going to work well on Test set. The other model which gave the best RMSE was TES and Manual SARIMAX (4,1,2)(3,0,2,12). We will built a final model on the entire Rose dataset using SARIMAX.







### Problem 1.10

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

**Resolution:**

**Sparkling Wine data:**

* TES (Triple Exponential Smoothing) has worked the best for the forecast with lowest RMSE on test data
* You can see from the above chart that the forecast for next 12 months is slightly over the sales of the previous 12 months however, there isn't a considerable increase.
* Observed from the month wise bar plots previously, we can say that the sales of Sparkling wine tend to go up in last two months probably because it's a holiday season than the rest and its lowest around Jun and July
* ABC can take various measures to increase the sales towards the beginning and mid of the year, it can introduce promotional activities or discounts during the low sales period.
* ABC can tie up with events like concerts, weddings etc. and do some sponsorships to boost sales during the slack

**Rose Wine data:**

* We chose manual SARIMAX model to predict for the Rose wine data. The model was passed the cut offs found through ACF and PACF plots of q and p respectively and seasonality of 12 as the plots showed a patterned significance after 11 lags.
* You can see from the above plot for Rose wine data the forecast for 1996 is more or less same as of for 1995.
* Observed from the monthly bar plot sales shows an increasing trend from August towards December, it’s on the lower side beginning of the year
* ABC can take sought promotional activities and implement some discounts during the first half of the year

The End

Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*