PROJECT – TIME SERIES FORECASTING

Submitted by

PRIYADHARSHINI K

On 25 JULY 2021

Table of Contents
PROBLEM STATEMENT
DATA INGESTION
1. Read the data as an appropriate Time Series data and plot the data.
DATA PREPARATION
2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
MODELING
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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.
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.
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.
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.
INFERENCE
8. 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.
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.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

PROBLEM STATEMENT

The data of different types of wine sales in the 20th century is to be analyzed. Both 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.

DATA INGESTION

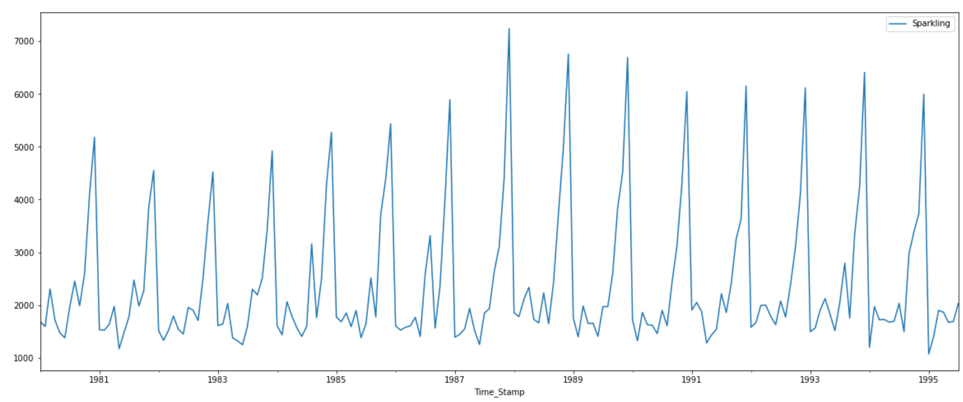
1. Read the data as an appropriate Time Series data and plot the data.

2 datasets 'Sparkling' and 'Rose' are given for analysis.

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Each of the dataset has 2 columns – one column with Period and other with column with sales of wine in that period. The data has object and float64 data types. The data type must be converted into DateTime data type for analysis using the ***to_datetime()*** function.



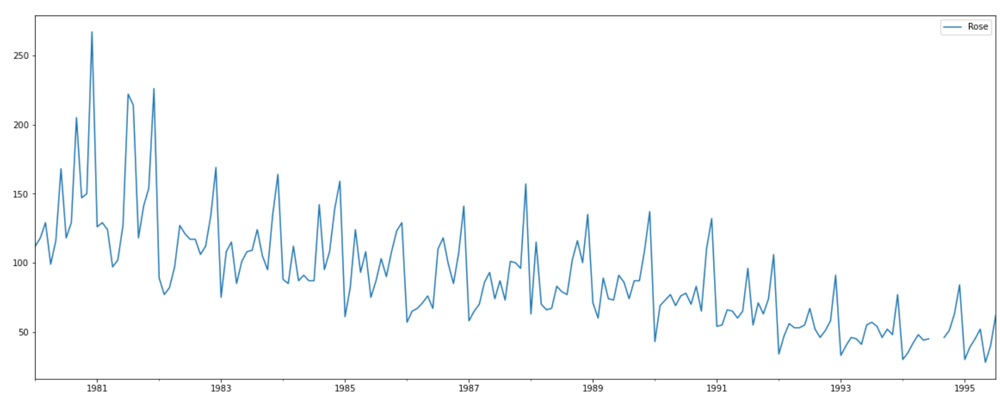


Figure 1. Time series plot of 'Sparkling' and 'Rose dataset

DATA PREPARATION

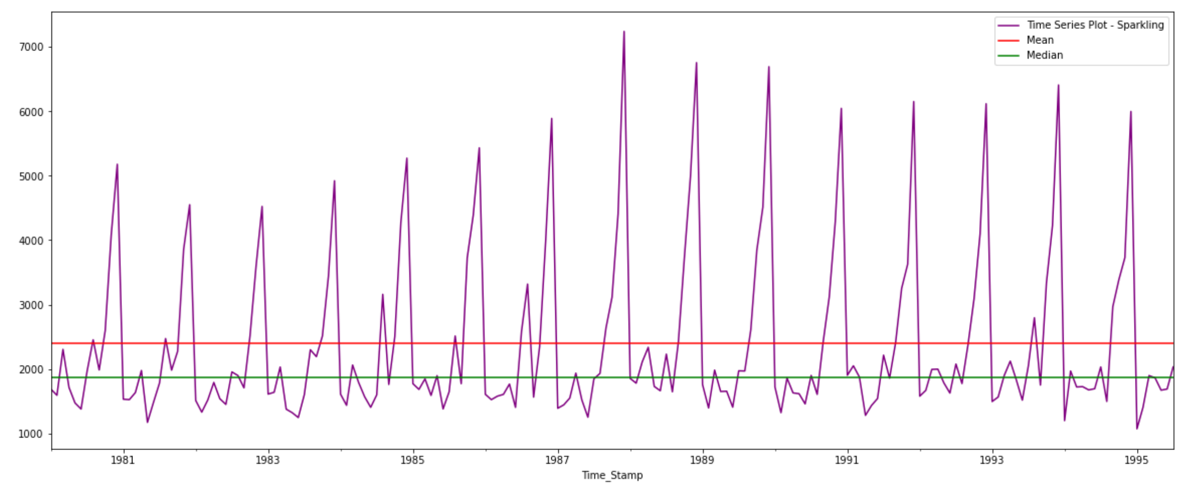
2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.

Use df.describe() function to perform basic data exploration – mean, standard deviation, count values.

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

There are a total of 187 line items in Sparkling and 185 line items in Rose data set. Considering the timeframe to be the same for both the datasets, Rose line item must have 2 missing items.



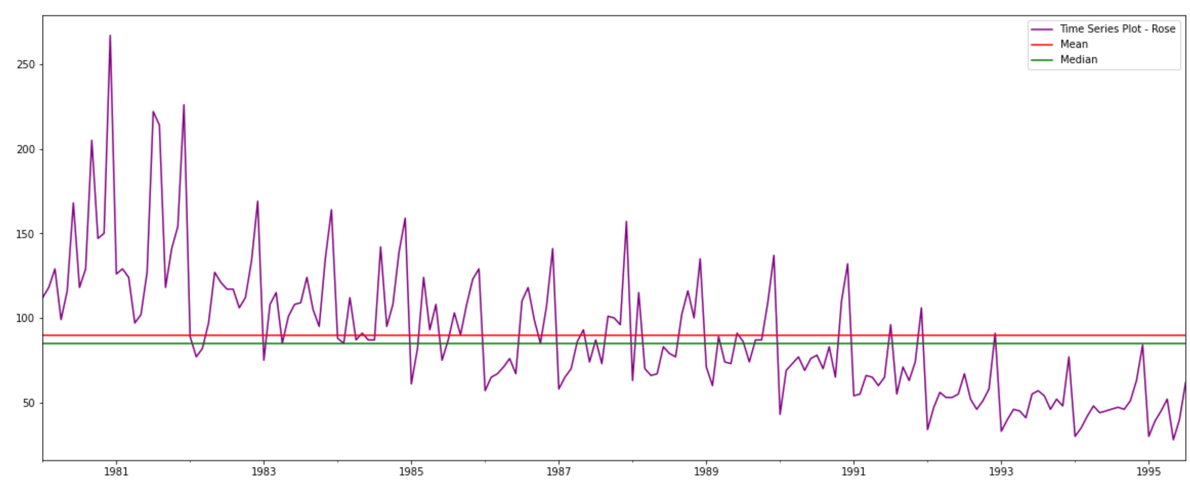


Figure 2. Time series plot of ‘Sparkling’ and ‘Rose dataset with mean and median values

Year on year analysis:

The year on year box plot for ‘Sparkling’ and ‘Rose’ shows the presence of outliers in the data.

The wine sales for ‘Sparkling’ type is almost consistent over the years compared to the abruptly decreasing sales of the ‘Rose’ wine type.

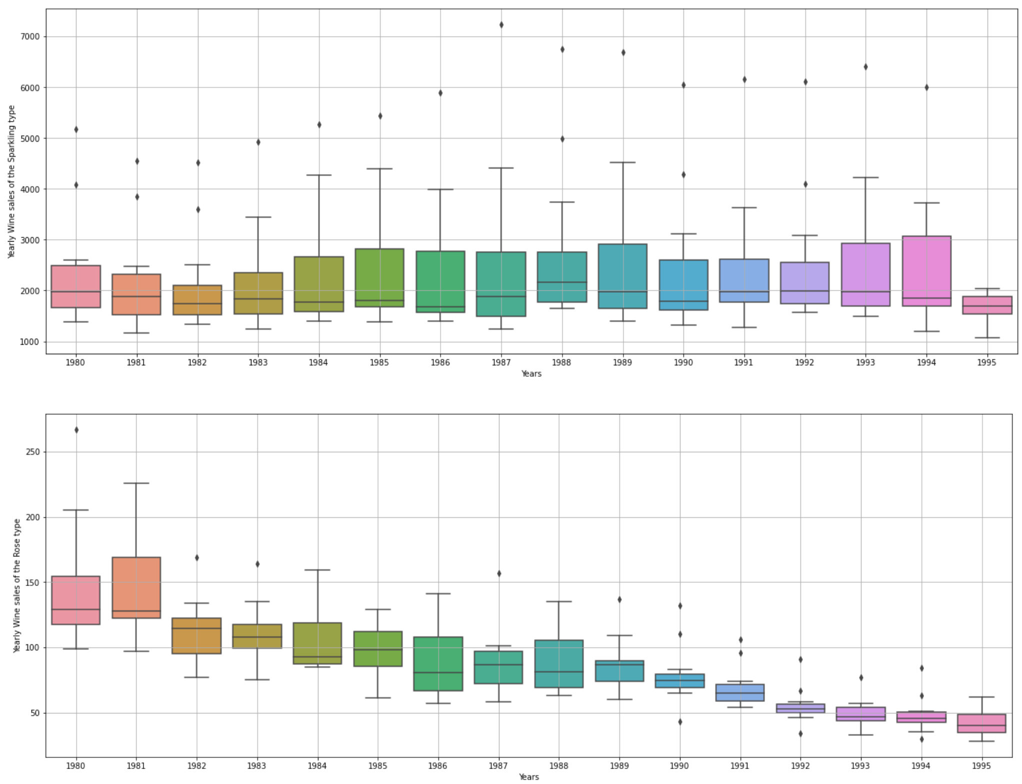


Figure 3. YoY Box plot of ‘Sparkling’ and ‘Rose wine sales

Quarterly analysis:

The quarterly analysis proves that the sale of wine is predominant towards the last quarter of the year, with an increasing trend from the beginning of the year.

‘Sparkling’ sales is almost getting doubled in the last quarter compared to the other 3 quarters.

‘Rose’ type is having a gradual increase in the sales in all the quarters, reaching the maximum by 4th quarter.

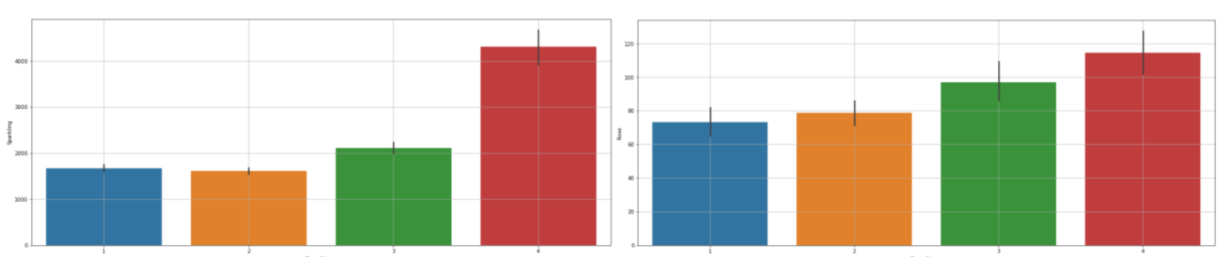


Figure 4. Quarterly Box plot of ‘Sparkling’ and ‘Rose wine sales

Monthly analysis:

The range of sales (between 25% to 75% value) of ‘Sparkling type’ is very minimum across the months whereas for the ‘Rose’ wine type, the value is almost consistent. The wine sale in the last 3 months of the year reach new heights for the Sparkling type compared to a not-so-great increase in the Rose type. The monthly plot shows that the outliers are not commonly present in ‘Sparkling’ type whereas ‘Rose’ type has the outlier presence in about 5 months of the year.

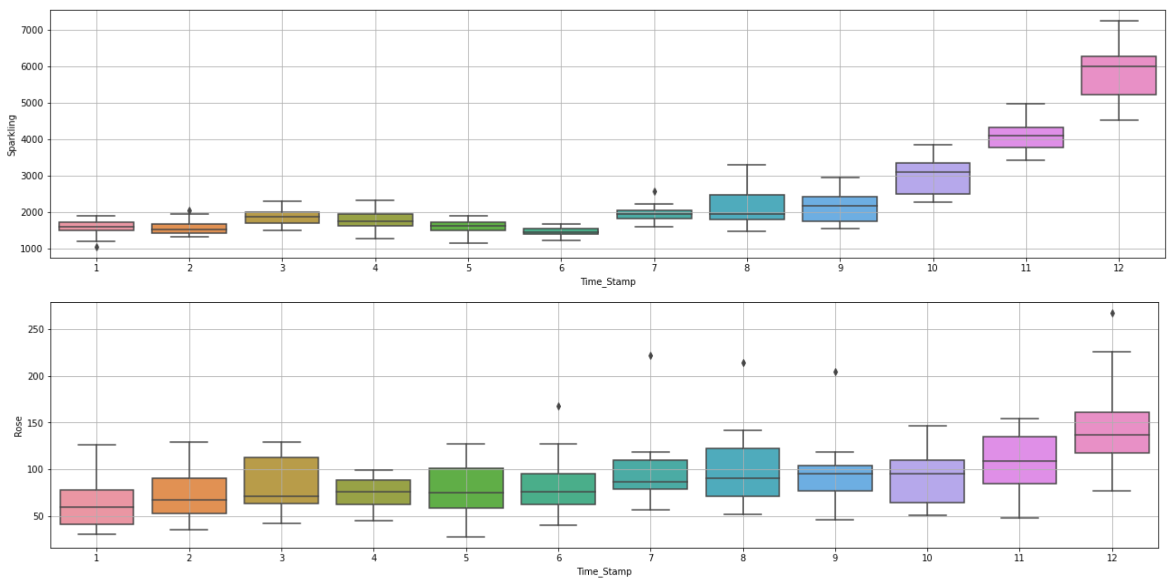


Figure 5. Monthly Box plot of ‘Sparkling’ and ‘Rose wine sales

The plot shows to understand level, trend, and seasonality in the data sets

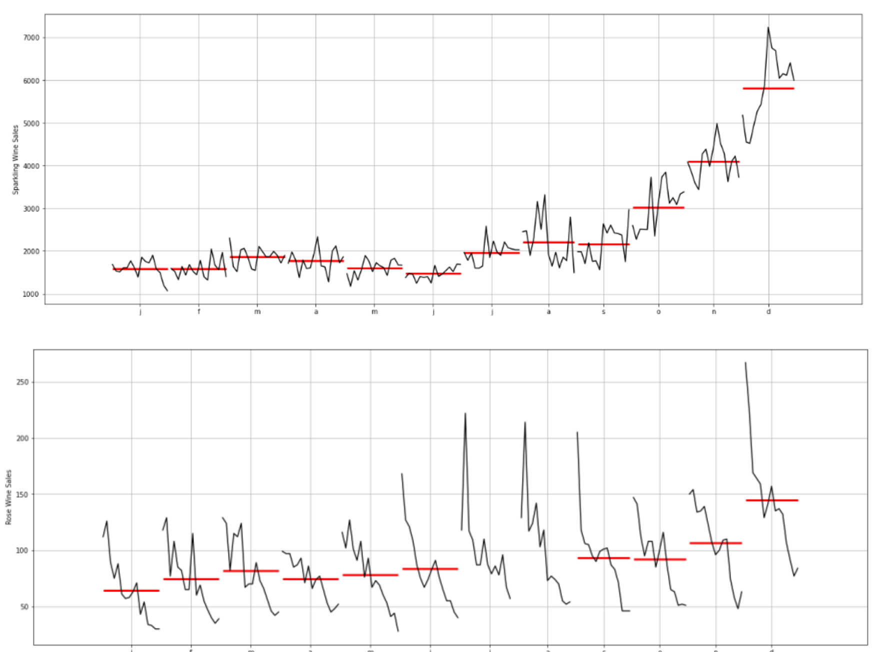
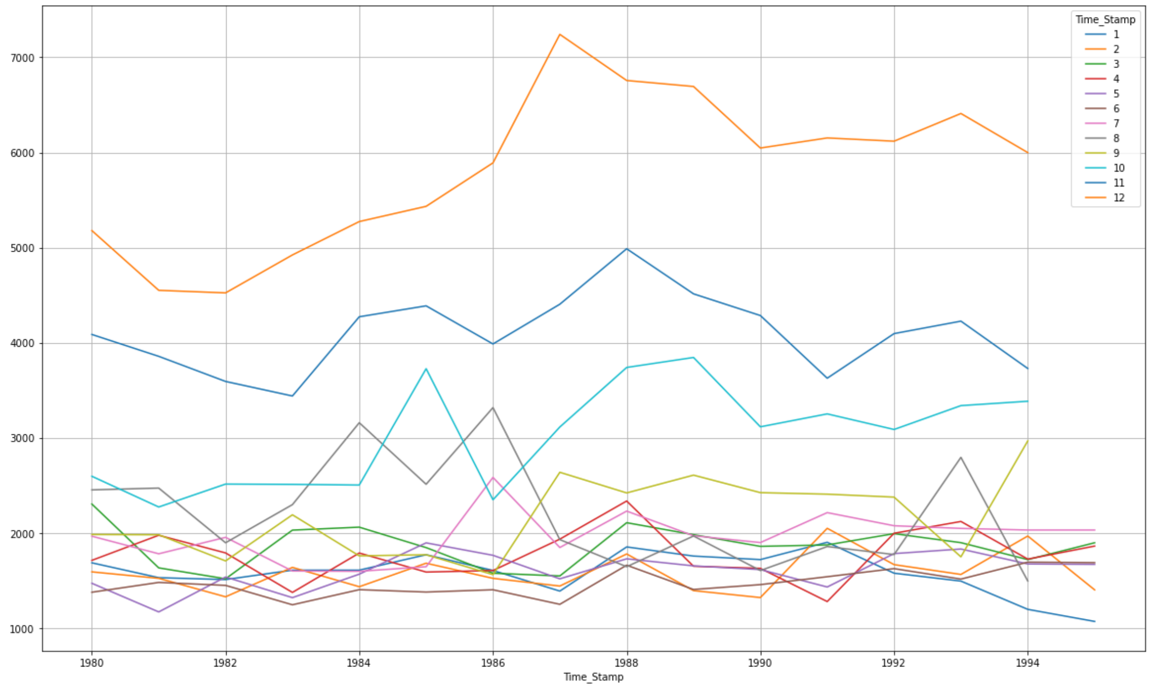


Figure 6. Month plot of ‘Sparkling’ and ‘Rose wine sales time series

Yearly sales across months

The yearly sale across month plot re-iterates the fact that the wine sale in the last month of the year is the maximum for the Sparkling type and the Rose type. The sales of the sparkling type do not show a predominant trend, whereas 'Rose' type shows a decreasing trend in sales.



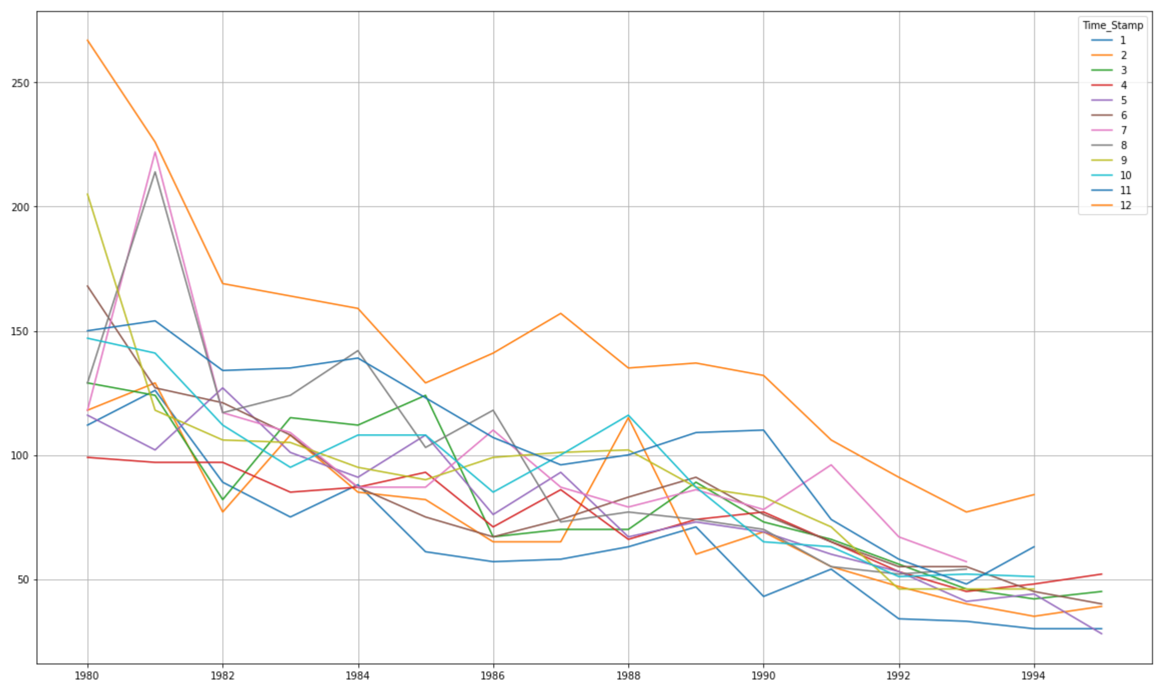


Figure 7. Yearly sales across months of the 'Sparkling' and 'Rose wine sales

RESAMPLING:

Resampling strategies are pre-processing approaches that change the original data distribution to get an overall understanding of the data.

Wine sales of ‘sparkling’ type does not exhibit a trend, it shows a steep low sale in 1994 while in 1985 the sales were maximum, while Wine sales of ‘rose’ type is high only in the beginning and consistently the sale dropped.

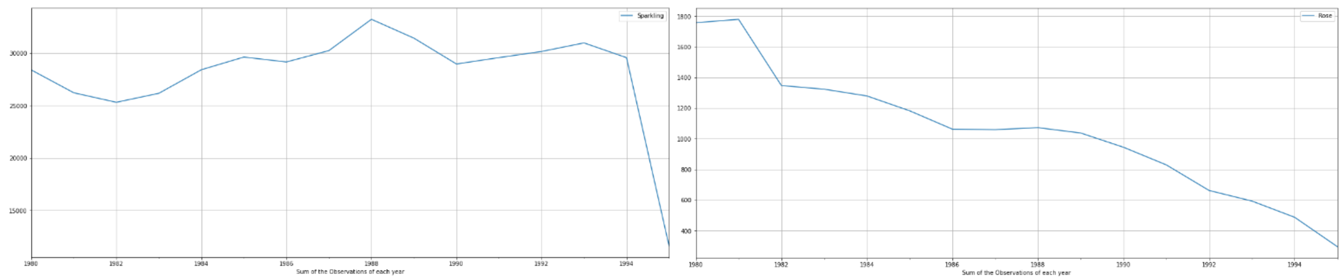


Figure 8. Resampling monthly data into yearly data – Sum of the observations

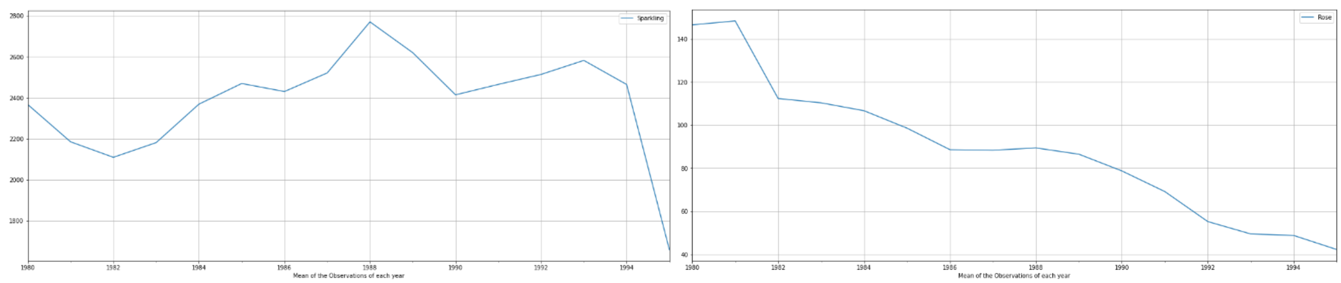


Figure 9. Resampling monthly data into yearly data – Mean of the observations

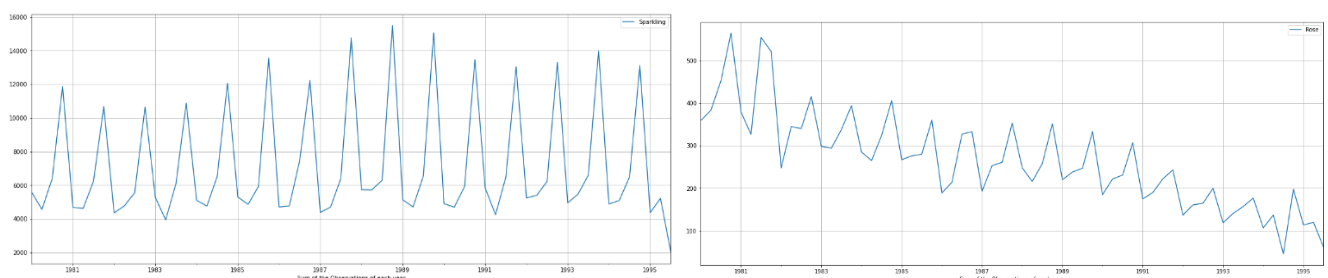


Figure 10. Resampling monthly data into Quarterly data – Sum of the observations

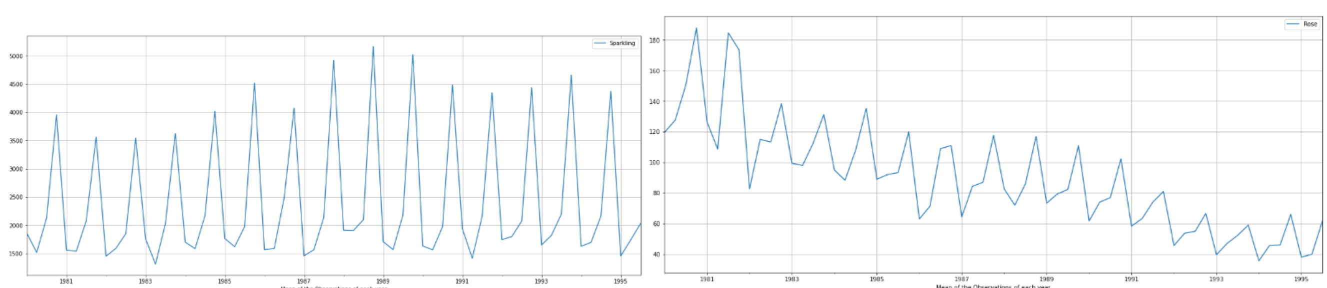
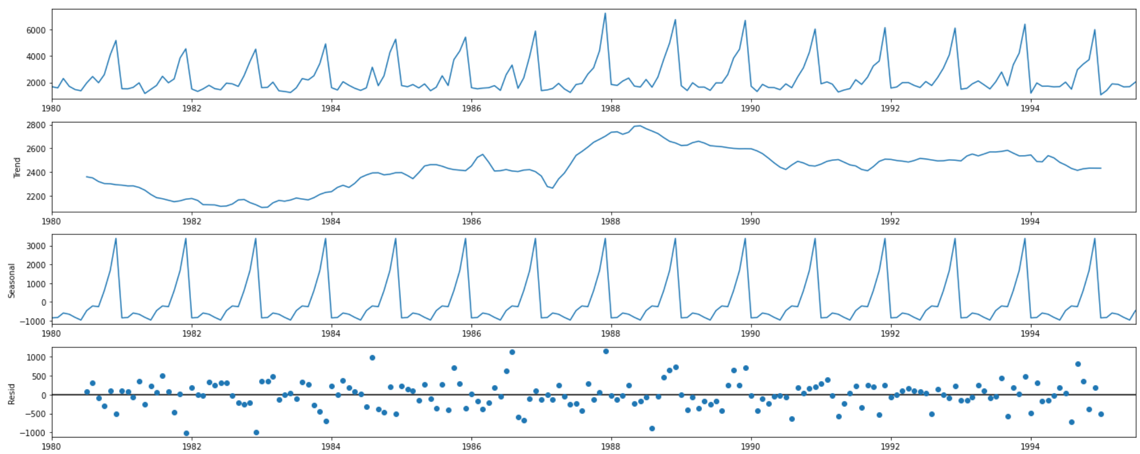


Figure 11. Resampling monthly data into Quarterly data Mean of the observations

DECOMPOSE THE TIME SERIES – SPARKLING WINE

Both additive model and multiplicative model does not show presence of trend in the dataset. The seasonality of the dataset is exhibited in the same way by both models. Multiplicative model can handle the residues well compared to the additive model.

Additive Model



Multiplicative Model

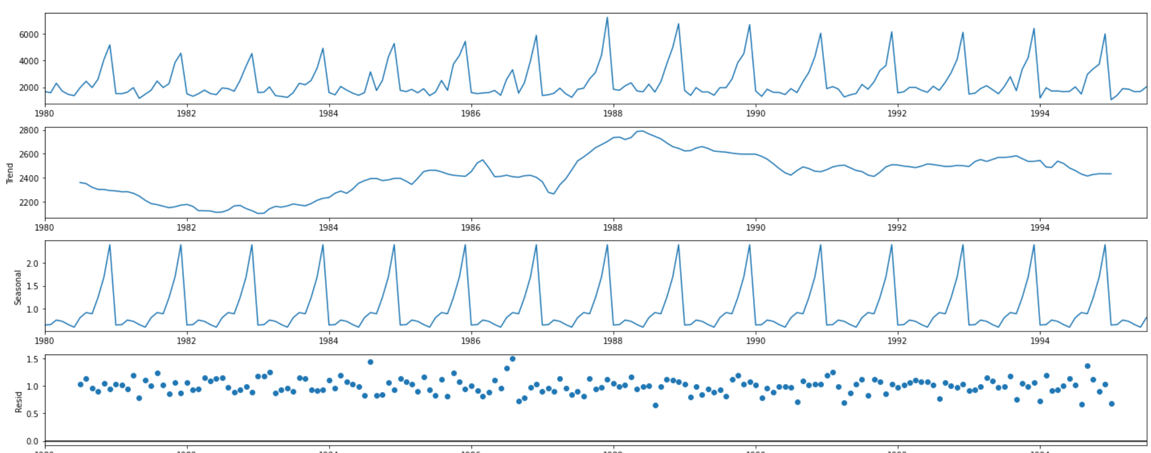


Figure 12. Decomposition of Sparkling sales

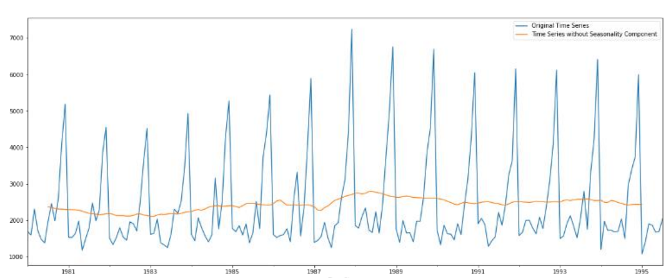


Fig.13a. Time series without seasonal factor model

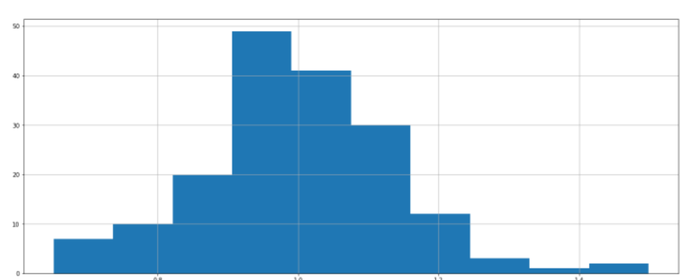


Fig. 13b Residual plot of the multiplicative

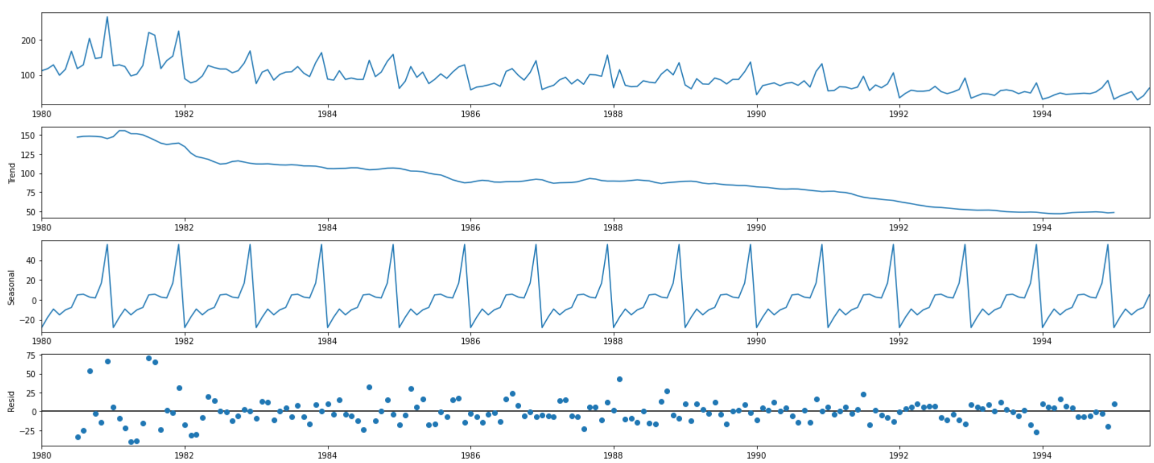
The comparison chart between original series and time series without seasonal component helps to show the predominant presence of seasonal factor in the data. Residues follow a normal distribution for multiplicative decomposition model and holds good.

DECOMPOSE THE TIME SERIES – ROSE WINE

As ‘Rose’ wine type has missing values, decomposition can be performed only after the missing values are interpolated. Various interpolation techniques can be used. Here, ‘Linear’ interpolation is used.

1994-03-01	42.000000
1994-04-01	48.000000
1994-05-01	44.000000
1994-06-01	45.000000
1994-07-01	45.333333
1994-08-01	45.666667
1994-09-01	46.000000
1994-10-01	51.000000

Additive Model



Multiplicative Model

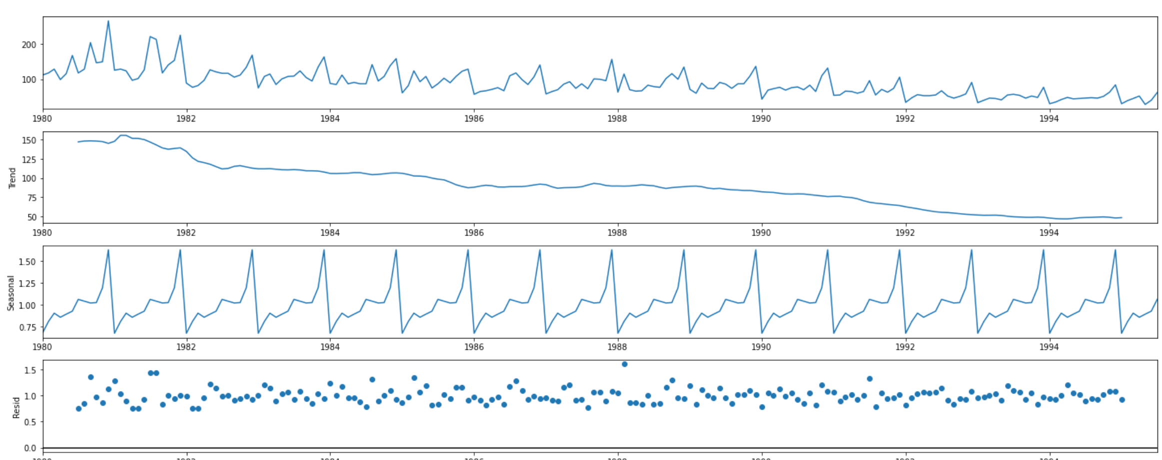
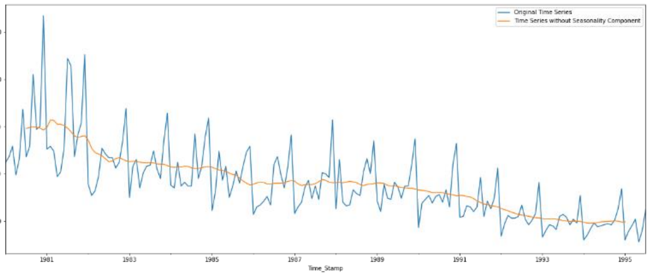


Figure 14. Decomposition of Rose wine sales



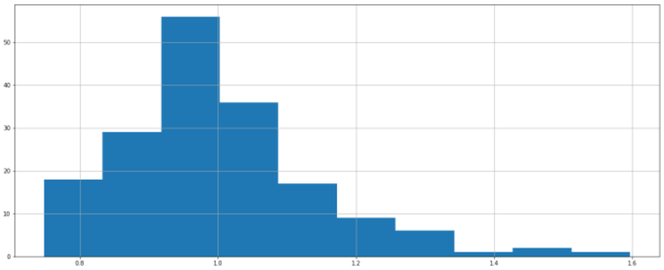


Fig.15a. Time series without seasonal factor   Fig. 15b Residual plot of the multiplicative model

The comparison chart between original series and time series without seasonal component helps to show the predominant presence of seasonal factor as well as the decreasing trend in the data. Residues follow a normal distribution for multiplicative decomposition model and holds good.

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

First few rows of Training Data

Time_Stamp	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

First few rows of Test Data

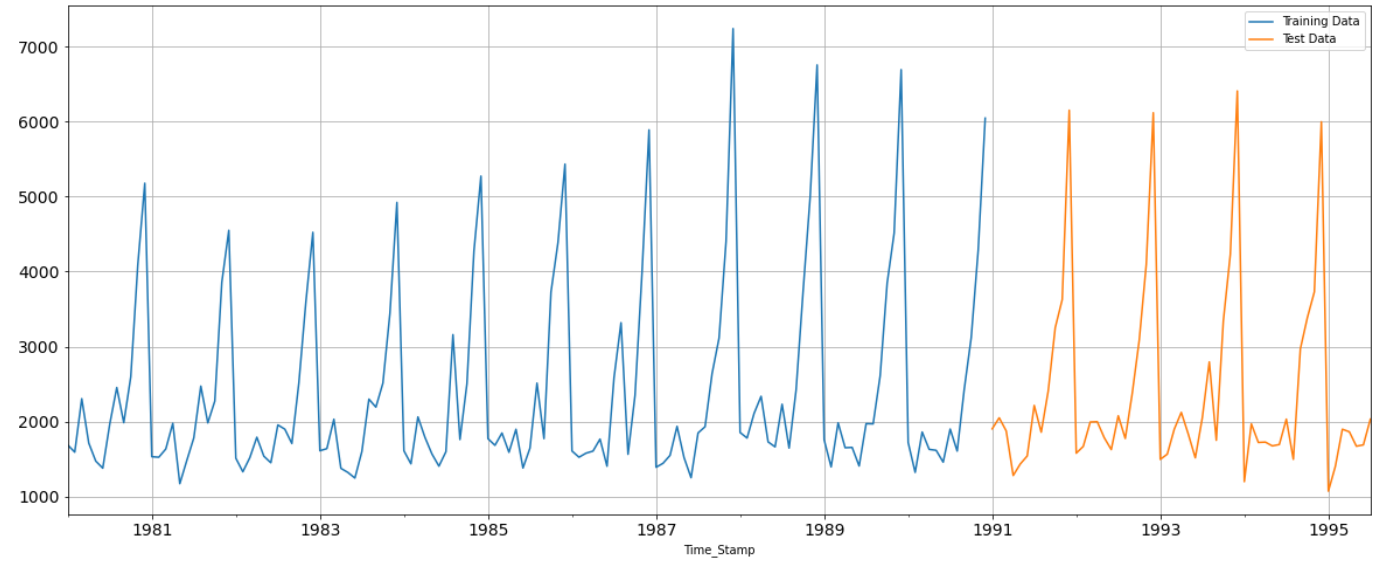
Time_Stamp	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

First few rows of Training Data

Time_Stamp	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

First few rows of Test Data

Time_Stamp	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0



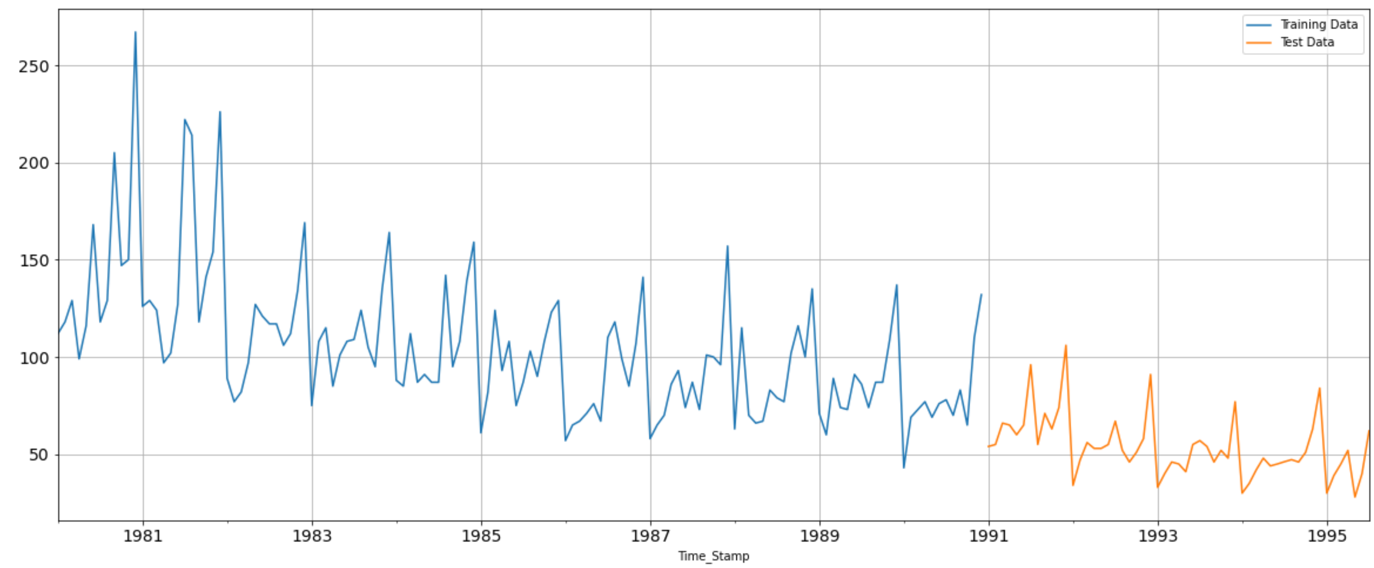


Fig. 16 Train and Test Data plot of the 'Sparkling' and 'Rose' wine sales

MODELING

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, and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Simple 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.

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. 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.

SimpleExpSmoothing class must be instantiated and passed the training data. The fit() function is then called providing the fit configuration, the alpha value, smoothing_level Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be none(N), additive (N), additive damped (Ad), Multiplicative (M) or multiplicative damped (Md). One or more parameters control how fast the weights decay. These parameters have values between 0 and 1

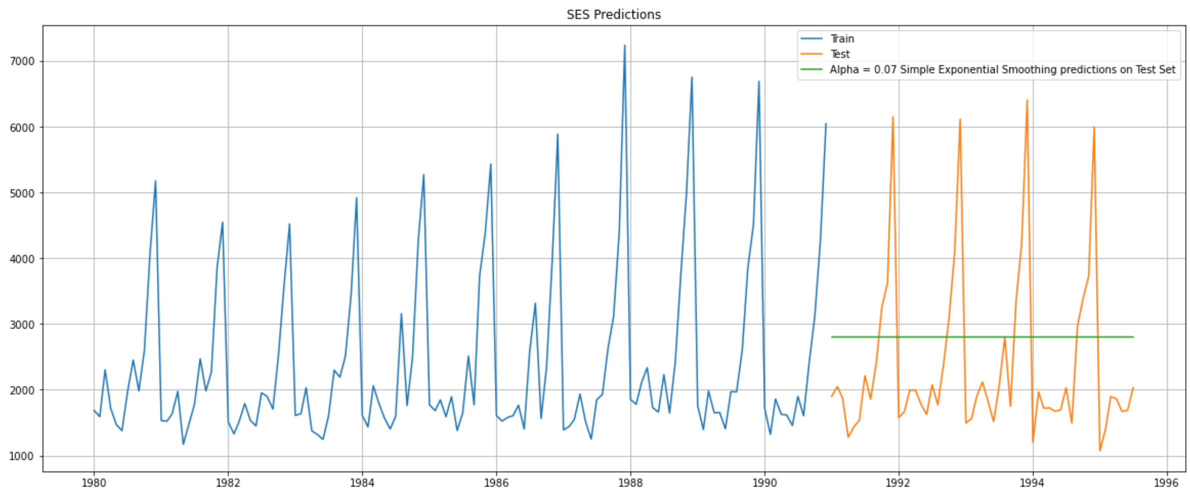


Fig. 17a SES plot of the 'Sparkling' wine sales – Alpha 0.07

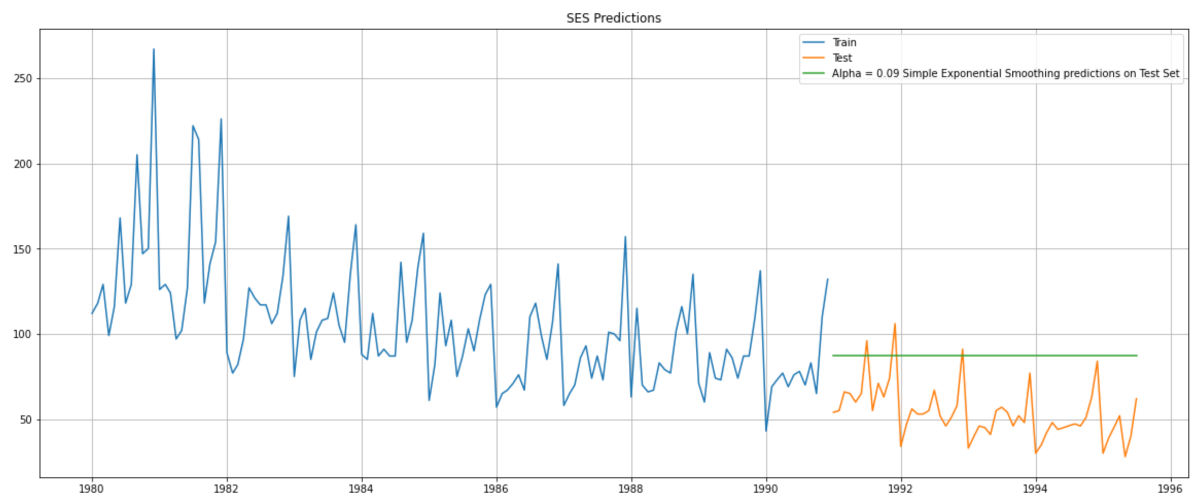


Fig. 17b SES plot of the 'Rose' wine sales– Alpha 0.09

SES MODEL EVALUATION

Sparkling Wine

	Test RMSE
Alpha = 0.07 SES	1338.008384

Rose Wine

	Test RMSE
Alpha = 0.09 SES	36.748402

Double Exponential Smoothing - ETS(A, A, N) - Holt's linear method with additive error

Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecating the short term avarage value or level and the other for capturing the trend.

- Intercept or Level equation, LtLt is given by: $Lt=\alpha Yt+(1-\alpha)Ft$Lt=αYt+(1−α)Ft
- Trend equation is given by $Tt=\beta(Lt-Lt-1)+(1-\beta)Tt-1$Tt=β(Lt−Lt−1)+(1−β)Tt−1

The forecast at time t + 1 is given by

- $Ft+1=Lt+Tt$Ft+1=Lt+Tt
- $Ft+n=Lt+nTt$

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.

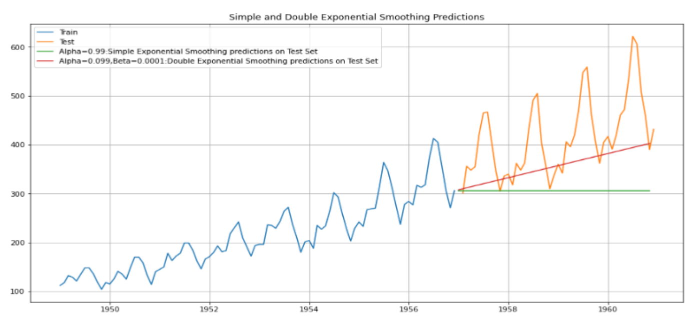


Fig. 18a DES plot of the ‘Sparkling’ wine sales

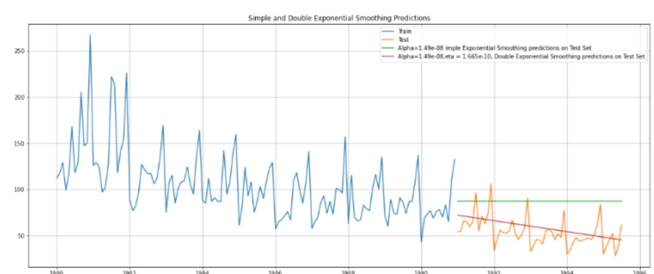


Fig. 18b DES plot of the ‘Rose’ wine sales

DES MODEL EVALUATION

Sparkling Wine

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833

Rose Wine

	Test RMSE
Alpha = 0.09 SES	36.748402
Alpha=1.49e-08,eta = 1.665e-10	15.255480

INFERENCES:

Double Exponential Smoothing has done well (lesser RMSE) in ‘Rose’ wine type when compared to the Simple Exponential Smoothing as the Double Exponential Smoothing model has picked up the trend component as well in the Rose data set and trend is not present in the ‘Sparkling’ wine type.

Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

Because Seasonality can be additive or multiplicative, HW model can be additive or multiplicative Simultaneously smooths the level, trend, and seasonality. Three separate smoothing parameters, α: Smooths level; 0 < α < 1 • β: Smooths trend; 0 < β < 1 • γ : Smooths seasonality; 0 < γ < 1

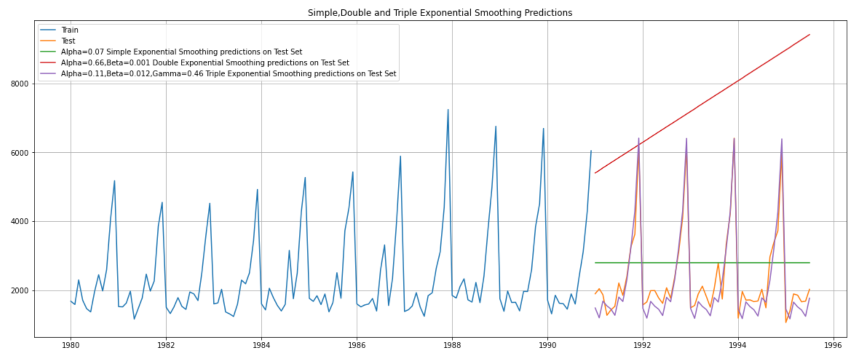


Fig. 19a TES plot of the ‘Sparkling’ wine sales

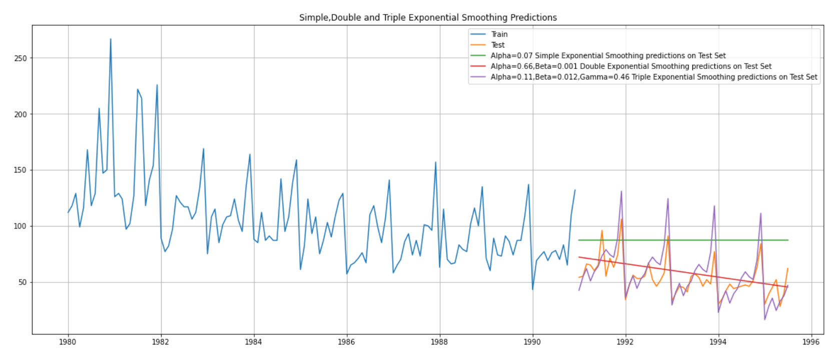


Fig. 19b TES plot of the ‘Rose’ wine sales

TES MODEL EVALUATION – ADDITIVE ERROR

Sparkling Wine

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008

Rose Wine

	Test RMSE
Alpha = 0.07 SES	36.748402
Alpha=0.66,Beta=0.0001 DES	15.255480
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	14.222850

INFERENCES:

Triple Exponential Smoothing has performed the best on the test as expected since the data had both trend and seasonality for both the data sets.

Holt-Winters - ETS(A, A, M) - Holt Winter's linear method with 'multiplicative' errors

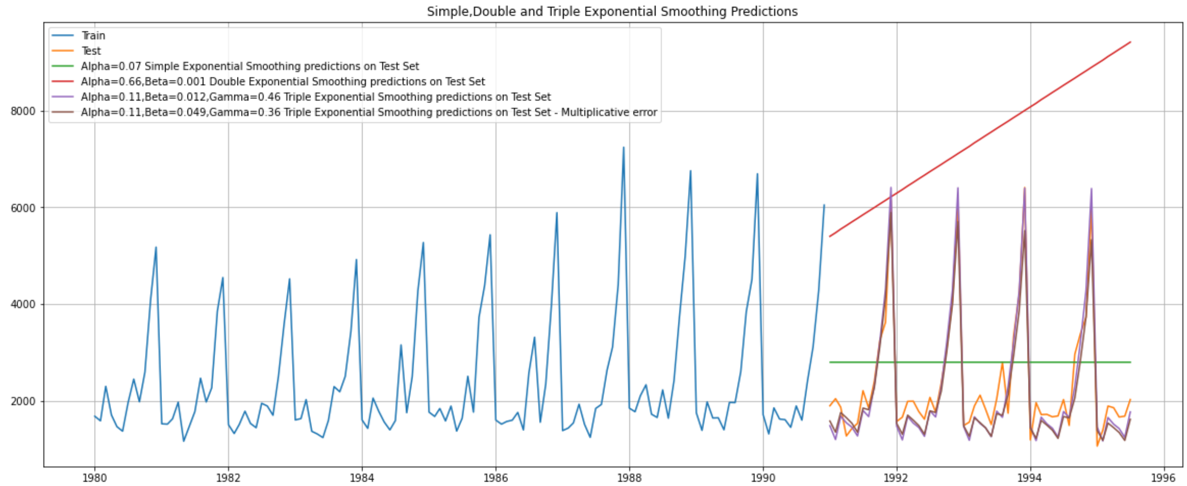


Fig. 20a TES plot of the 'Sparkling' wine sales – Multiplicative error

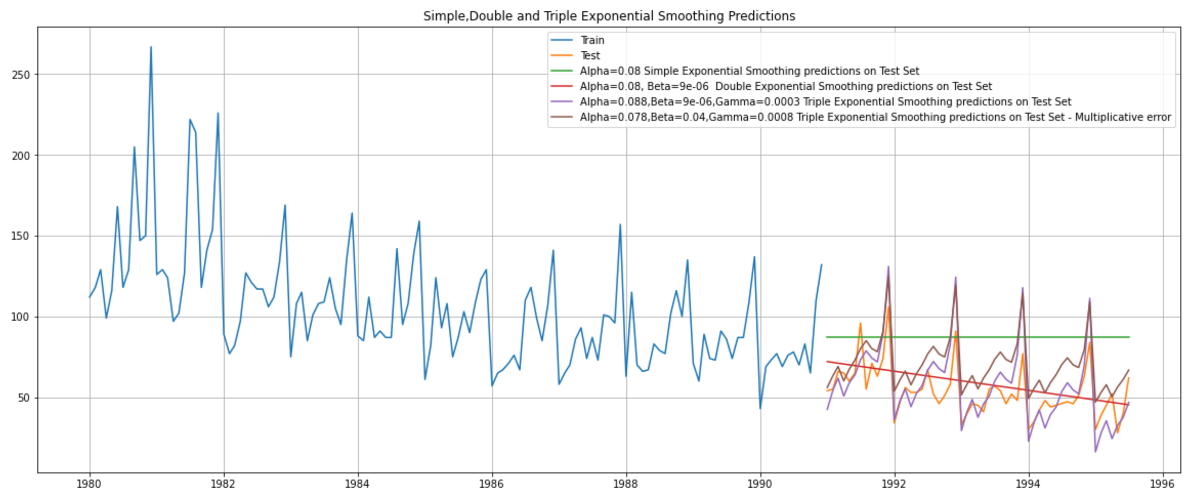


Fig. 20b TES plot of the 'Rose' wine sales – Multiplicative error

TES MODEL EVALUATION – MULTIPLICATIVE ERROR

'Sparkling' wine

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	402.938530

'Rose' wine

	Test RMSE
Alpha = 0.07 SES	36.748402
Alpha=0.66,Beta=0.0001 DES	15.255480
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	14.222850
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	19.337756

We see that the multiplicative error model has not done that well when compared to the additive error in the Triple Exponential Smoothing model.

HYPERPARAMETER TUNING OF SES, DES, TES MODEL

It is done by building smoothing model by taking the best alpha, beta, and gamma [all in the range of 0.1 to 1 taking an interval of 0.1] in terms of the least RMSE. Then, evaluation is done for the same model on the test data with 'optimized = False' within the '.fit()' command.

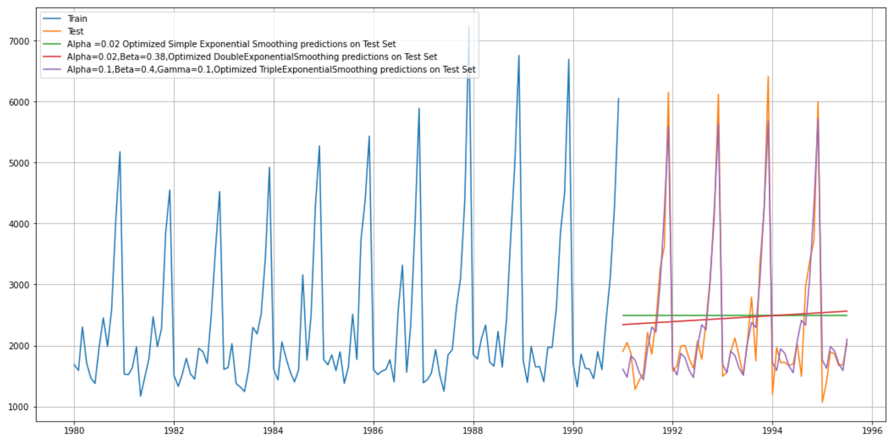


Fig. 21a Plot of the 'Sparkling' wine sales – Hyperparameter tuning

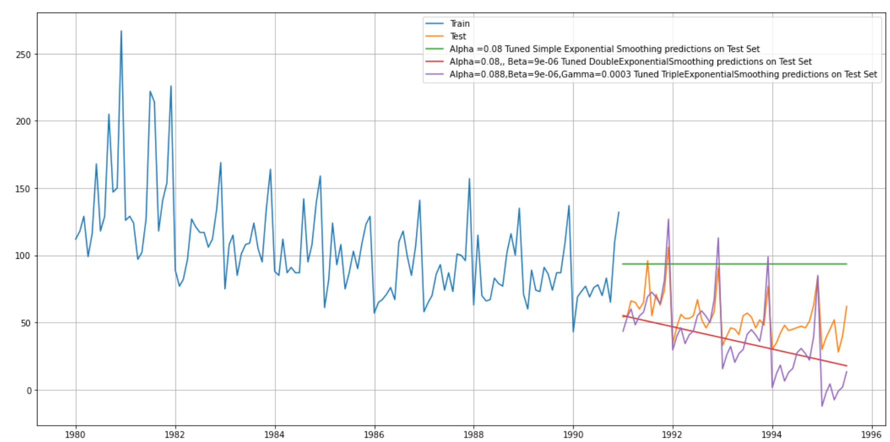


Fig. 21b Plot of the 'Rose' wine sales – Hyperparameter tuning

TES MODEL EVALUATION – Hyperparameter tuning

'Sparkling' wine

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	402.938530
Alpha=0.02 Optimized SES	1278.497798
Alpha=0.02,Beta=0.38 Optimized DES	1275.874751
Alpha=0.1,Beta=0.4,Gamma=0.1:Optimized TES	342.934716

'Rose' wine

	Test RMSE
Alpha = 0.09 SES	36.748402
Alpha=1.49e-08,Beta = 1.665e-10 DES	15.255480
Alpha=0.088,Beta=9e-06,,Gamma=0.0003 TES	14.222850
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	19.337756
Alpha=0.07 Tuned SES	36.387162
Alpha=0.04,Beta=0.47 Tuned DES	14.455710
Alpha=0.1,Beta=0.4,Gamma=0.3 Tuned TES	11.986168

Build a Linear Regression model

To perform linear regression on the target variable against the order of the occurrence, we need to modify training data before fitting it into a linear regression. Numerical time instance order for both the training and test set must be generated.

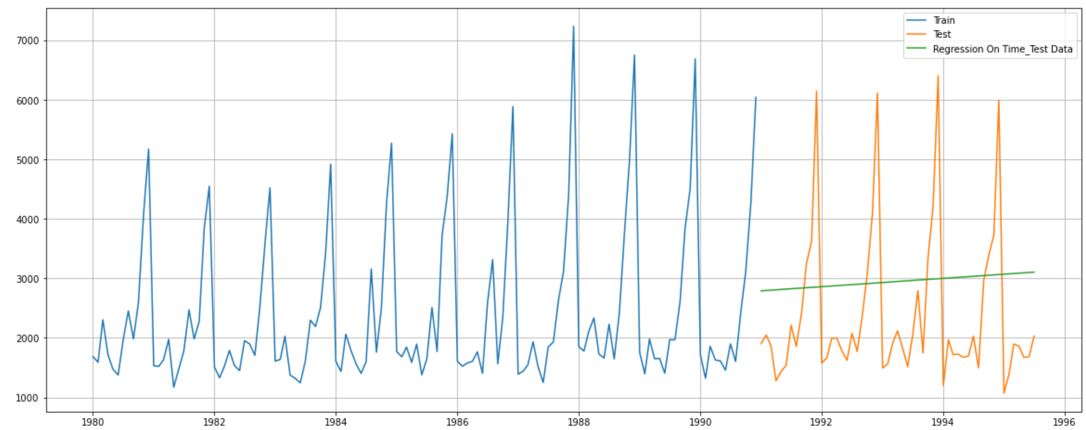


Fig. 22a Linear Regression Plot of the 'Sparkling' wine sales

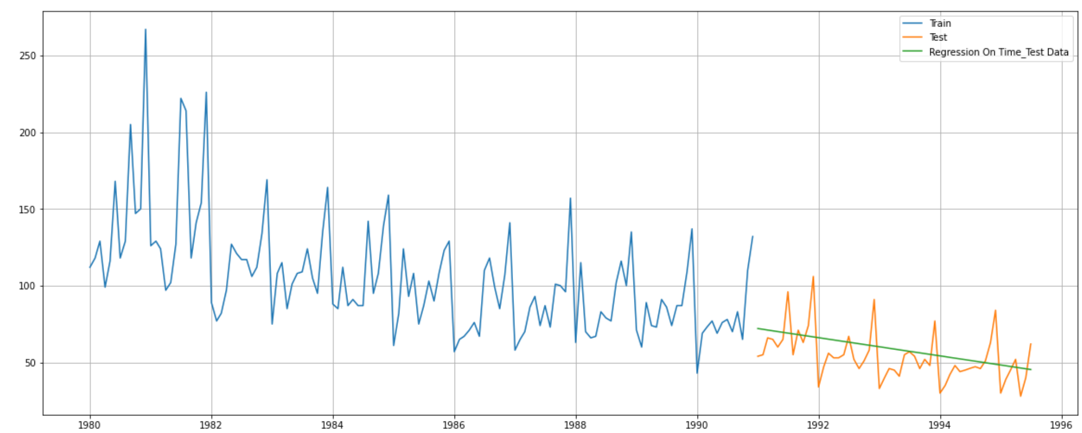


Fig. 22b Linear Regression Plot of the 'Rose' wine sales

TES MODEL EVALUATION – Hyperparameter tuning

'Sparkling' wine

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	402.938530
Alpha=0.02 Optimized SES	1278.497798
Alpha=0.02,Beta=0.38 Optimized DES	1275.874751
Alpha=0.1,Beta=0.4,Gamma=0.1:Optimized TES	342.934716
Linear RegressionOnTime	1389.135175

'Rose' wine

	Test RMSE
Alpha = 0.09 SES	36.748402
Alpha=1.49e-08,Beta = 1.665e-10 DES	15.255480
Alpha=0.088,Beta=9e-06,,Gamma=0.0003 TES	14.222850
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	19.337756
Alpha=0.07 Tuned SES	36.387162
Alpha=0.04,Beta=0.47 Tuned DES	14.455710
Alpha=0.1,Beta=0.4,Gamma=0.3 Tuned TES	11.986168
Linear RegressionOnTime	15.255492

Build a Naive Model

For 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.

$$\hat{y}_{t+1} = y_t$$

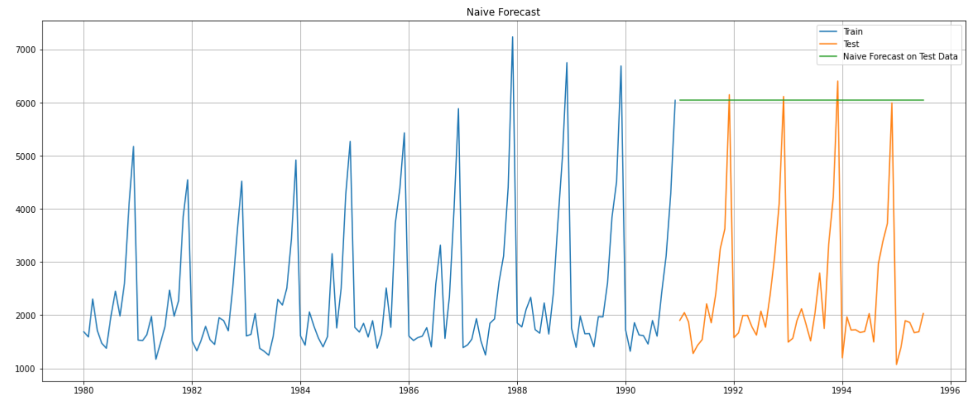


Fig. 23a Naive Plot of the 'Sparkling' wine sales

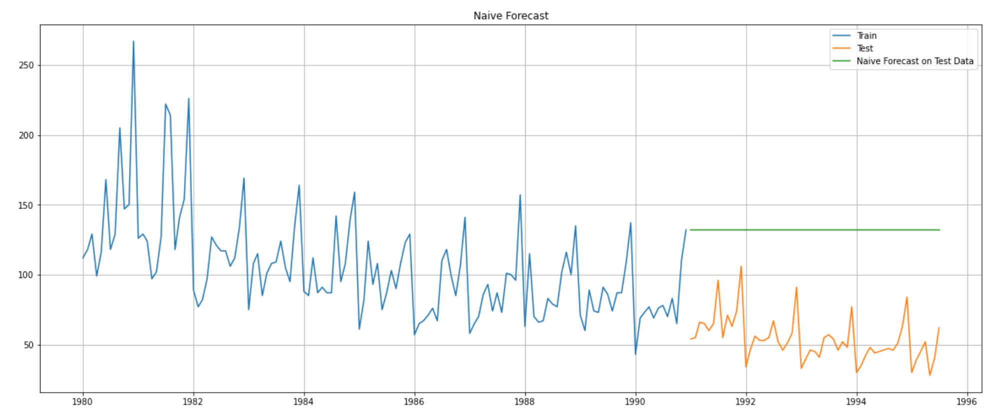


Fig. 23b Naive Plot of the 'Rose' wine sales

Naive MODEL EVALUATION – Hyperparameter tuning

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	402.938530
Alpha=0.02 Optimized SES	1278.497798
Alpha=0.02,Beta=0.38 Optimized DES	1275.874751
Alpha=0.1,Beta=0.4,Gamma=0.1:Optimized TES	342.934716
Linear RegressionOnTime	1389.135175
NaiveModel	3864.279352

	Test RMSE
Alpha = 0.09 SES	36.748402
Alpha=1.49e-08,Beta = 1.665e-10 DES	15.255480
Alpha=0.088,Beta=9e-06,,Gamma=0.0003 TES	14.222850
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	19.337756
Alpha=0.07 Tuned SES	36.387162
Alpha=0.04,Beta=0.47 Tuned DES	14.455710
Alpha=0.1,Beta=0.4,Gamma=0.3 Tuned TES	11.986168
Linear RegressionOnTime	15.255492
NaiveModel	79.672475

Build a Simple Average Model

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

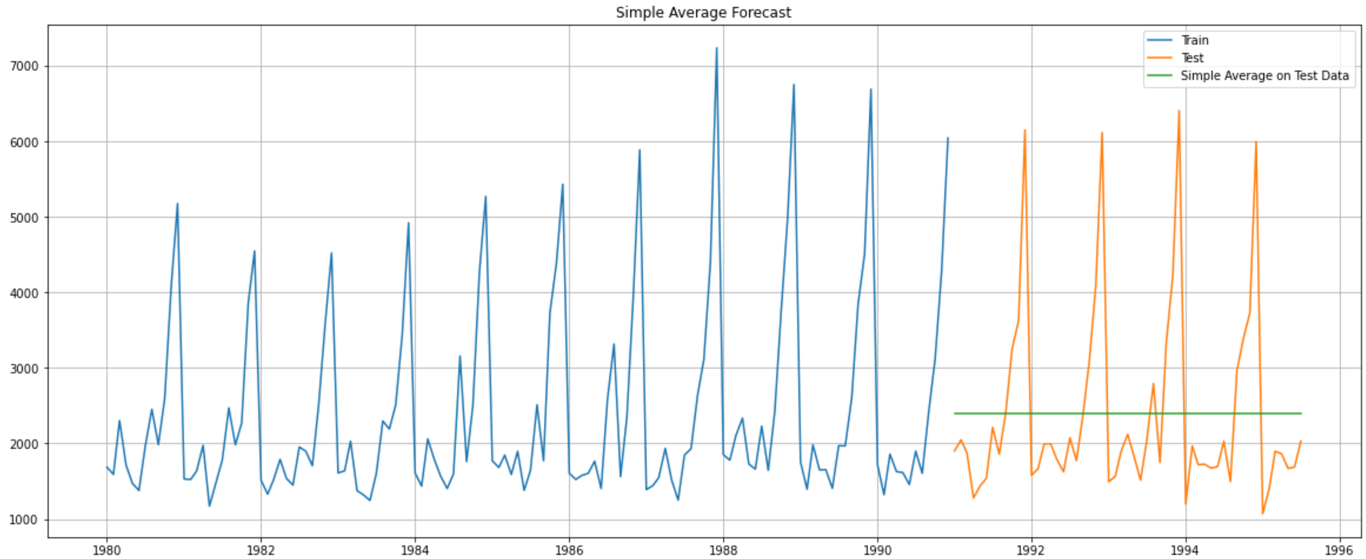


Fig. 24a Simple Average Plot of the 'Sparkling' wine sales

MODEL EVALUATION:

For Simple Average forecast on the Test Data of the 'Sparkling' wine sales, RMSE is 1275.082

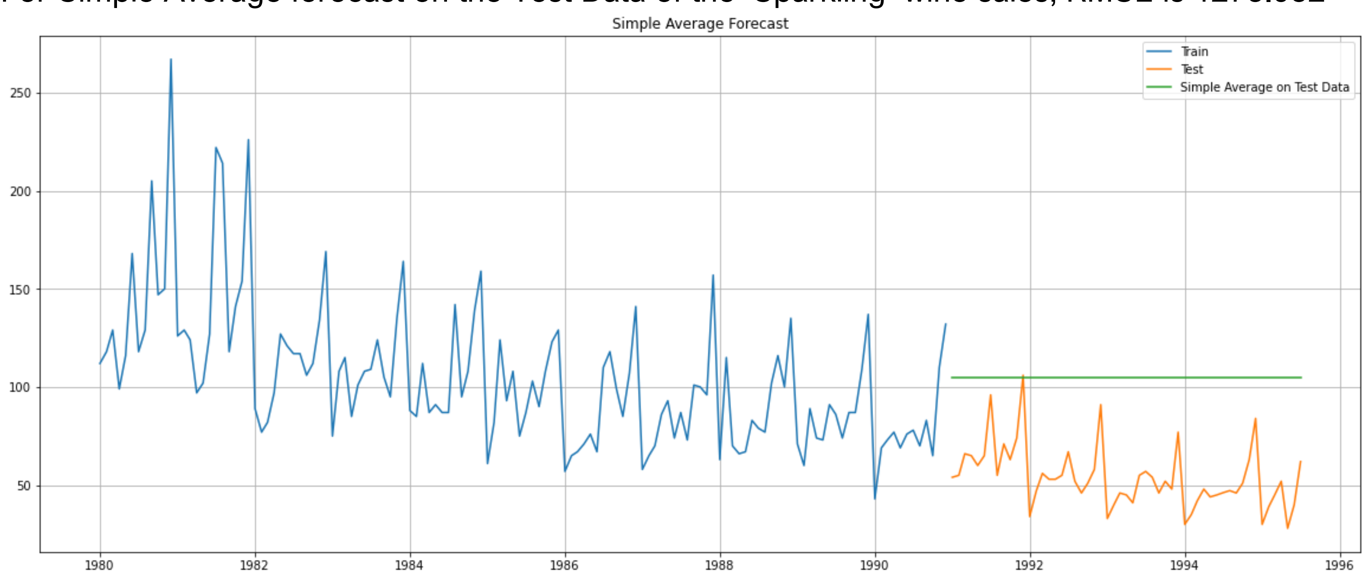


Fig. 24b Simple Average Plot of the 'Rose' wine sales

MODEL EVALUATION - For Simple Average forecast on the Test Data, RMSE is 53.413

Build a Moving Average Model

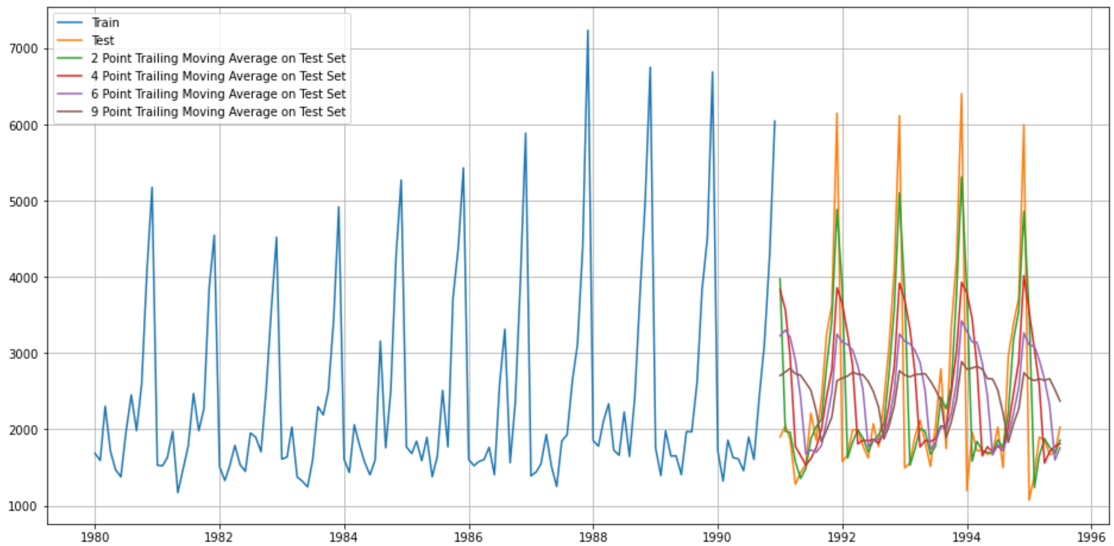


Fig. 25a Moving Average Plot of the 'Sparkling' wine sales

MODEL EVALUATION:

For 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278

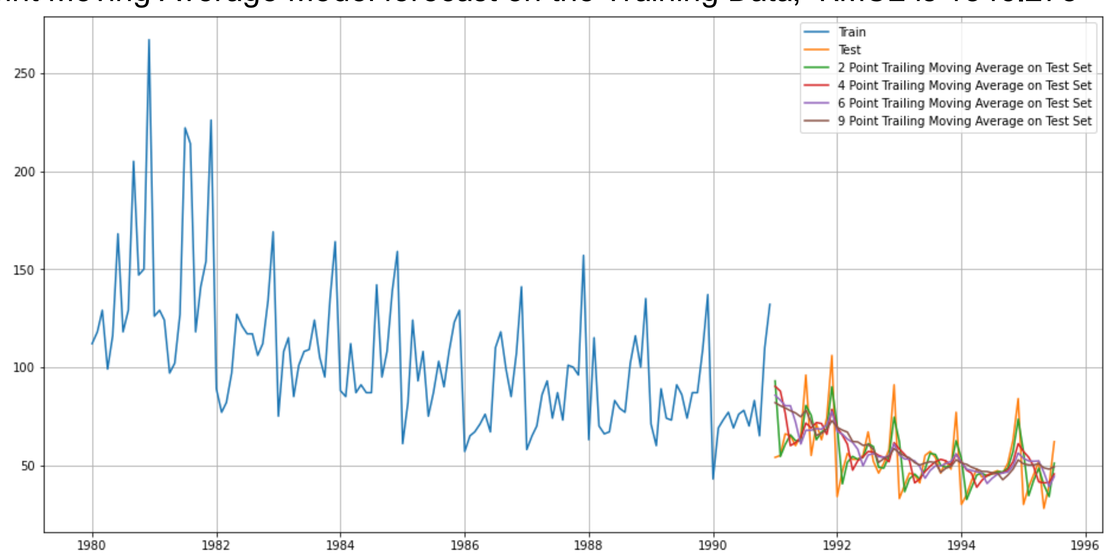


Fig. 25b Moving Average Plot of the 'Rose' wine sales

MODEL EVALUATION:
For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.530
For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.444
For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.555
For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.722

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and 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 Time Series is stationary whose statistical properties such as the variance and (auto) correlation are all constant over time. The properties of a stationary time series do not depend on time.

To check whether the series is stationary, we use the Augmented Dickey Fuller (ADF)test whose null and alternate hypothesis can be simplified to

Null Hypothesis H0: Time Series is non-stationary

Alternate Hypothesis Ha: Time Series is stationary

SPARKLING WINE SALES:

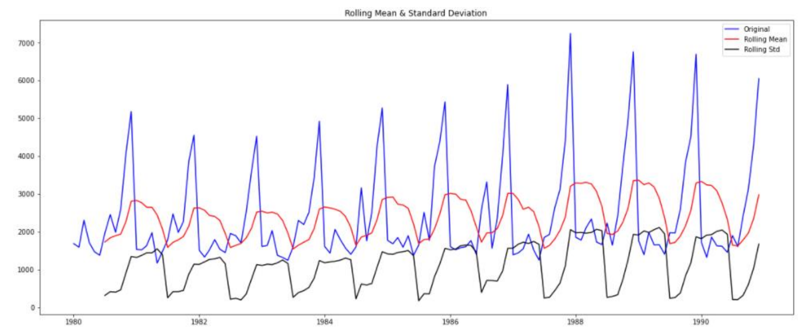


Fig. 26a Rolling mean and Standard Deviation Plot of the 'Sparkling' wine

```
Results of Dickey-Fuller Test:
Test Statistic                 -1.208926
p-value                         0.669744
#Lags Used                     12.000000
Number of Observations Used   119.000000
Critical Value (1%)            -3.486535
Critical Value (5%)            -2.886151
Critical Value (10%)           -2.579896
dtype: float64
```

We see that at 5% significant level the Time Series is non-stationary as 'p' value is 0.6 which is g reater than α = 0.05. We need to use differencing technique to make the data stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

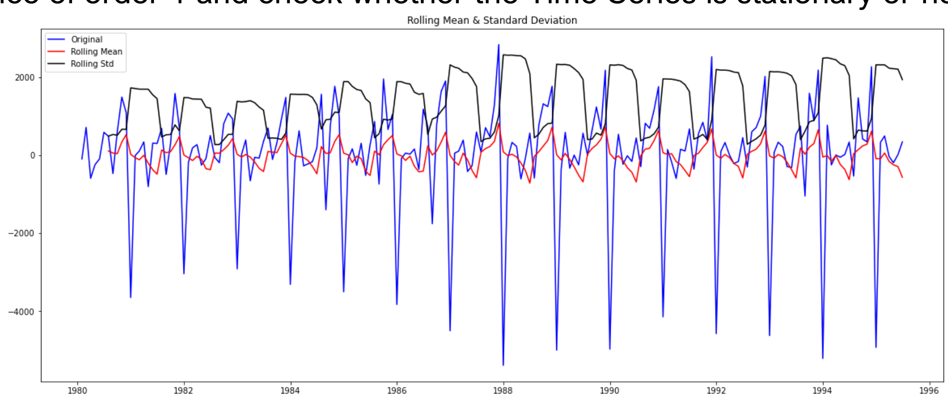


Fig. 27a Rolling mean & Standard Deviation Plot of the 'Sparkling' wine – Differenced series

```
Results of Dickey-Fuller Test:
Test Statistic                 -45.050301
p-value                          0.000000
#Lags Used                      10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

We see that after taking a difference of order 1 the series have become stationary at α = 0.05
We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there.

ROSE WINE SALES:

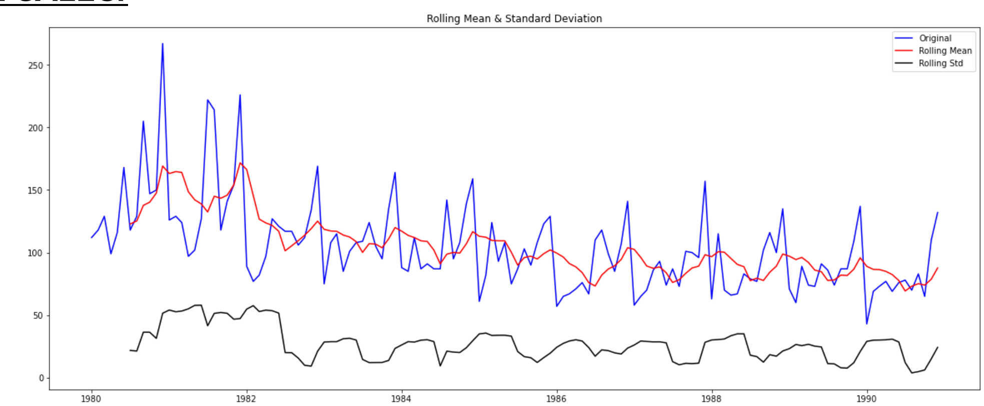


Fig. 26b Rolling mean and Standard Deviation Plot of the ‘Rose’ wine

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

We see that at 5% significant level the Time Series is non-stationary as 'p' value is 0.21 which is greater than α = 0.05. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

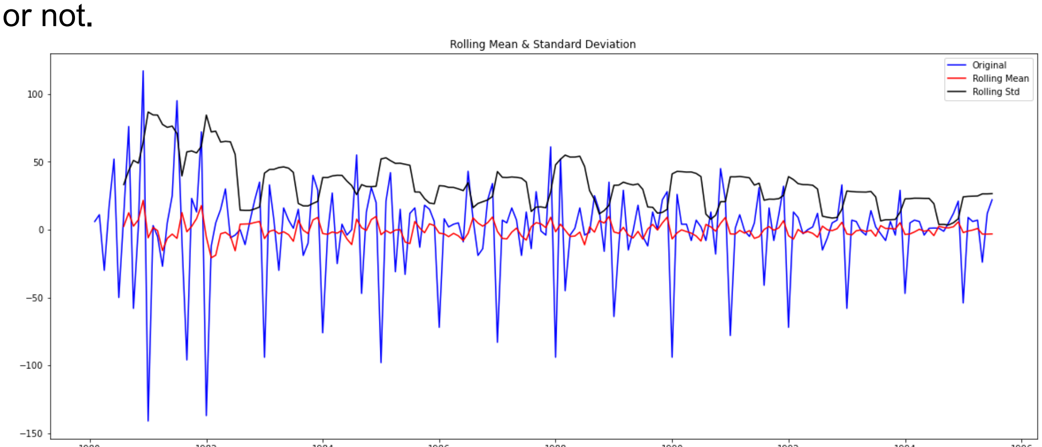


Fig. 27b Rolling mean & Standard Deviation Plot of the ‘Rose’ wine – Differenced series

```
Results of Dickey-Fuller Test:
Test Statistic                 -8.044819e+00
p-value                         1.806379e-12
#Lags Used                      1.200000e+01
Number of Observations Used     1.730000e+02
Critical Value (1%)            -3.468726e+00
Critical Value (5%)            -2.878396e+00
Critical Value (10%)           -2.575756e+00
dtype: float64
```

We see that after taking a difference of order 1 the series have become stationary at α = 0.05

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.

ARIMA – Auto Regressive Integrated Moving Average is a form of regression analysis.
The model's goal is to predict future values by examining the differences between values in the series instead of through actual values.
AR-Auto regressive refers to the tendency of the time series data to regress on itself when considering past or future lags.
I-Integration refers to the technique of differencing the data to bring in stationarity.
MA – Moving Average refers to the relation between the observation and the error from a lagged data perspective.
The commonly used terminology is p,d,q in sequence p- number of lag observations to consider, d-number of times differencing is to be done, q-size of the moving average window or order of moving average.

ARIMA MODELING

ARIMA models can be built keeping the Akaike Information Criterion (AIC) in mind as well. In this case, we choose the 'p' and 'q' values to determine the AR and MA orders respectively which gives us the lowest AIC value. Lower the AIC better is the model.

Coding languages tries different orders of 'p' and 'q' to arrive to this conclusion. Remember, even for such a way of choosing the 'p' and 'q' values, we must make sure that the series is stationary. The formula for calculating the AIC is 2k –2ln(L), where k is the number of parameters to be estimated and L is the likelihood estimation

SPARKLING DATA SET:

	param	AIC
8	(2, 1, 2)	2210.618562
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906439
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.309
Method:                       css-mle   S.D. of innovations           1012.730
Date:                Sun, 25 Jul 2021   AIC                           2210.619
Time:                        12:19:17   BIC                           2227.870
Sample:                    02-01-1980   HQIC                          2217.628
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5843      0.518     10.790      0.000       4.570       6.599
ar.L1.D.Sparkling     1.2700      0.074     17.048      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5604      0.074     -7.620      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9978      0.042    -47.093      0.000      -2.081      -1.915
ma.L2.D.Sparkling     0.9978      0.042     23.501      0.000       0.915       1.081
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1333           -0.7073j            1.3359           -0.0888
AR.2            1.1333           +0.7073j            1.3359            0.0888
MA.1            1.0004           +0.0000j            1.0004            0.0000
MA.2            1.0019           +0.0000j            1.0019            0.0000
-----------------------------------------------------------------------------
```

RMSE of ARIMA model – Sparkling wine sales - 1374.546023727508

ROSE DATA SET:

	param	AIC
2	(0, 1, 2)	1276.835373
5	(1, 1, 2)	1277.359229
4	(1, 1, 1)	1277.775753
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -633.649
Method:                       css-mle   S.D. of innovations             29.975
Date:                Sun, 25 Jul 2021   AIC                           1279.299
Time:                        17:15:23   BIC                           1296.550
Sample:                    02-01-1980   HQIC                          1286.309
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4911      0.081     -6.076      0.000      -0.649      -0.333
ar.L1.D.Rose    -0.4383      0.218     -2.015      0.044      -0.865      -0.012
ar.L2.D.Rose     0.0269      0.109      0.246      0.806      -0.188       0.241
ma.L1.D.Rose    -0.3316      0.203     -1.633      0.102      -0.729       0.066
ma.L2.D.Rose    -0.6684      0.201     -3.332      0.001      -1.062      -0.275
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -2.0290           +0.0000j            2.0290            0.5000
AR.2           18.3389           +0.0000j           18.3389            0.0000
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.4961           +0.0000j            1.4961            0.5000
-----------------------------------------------------------------------------
```

RMSE of ARIMA model – ROSE Dataset 15.342048385136696

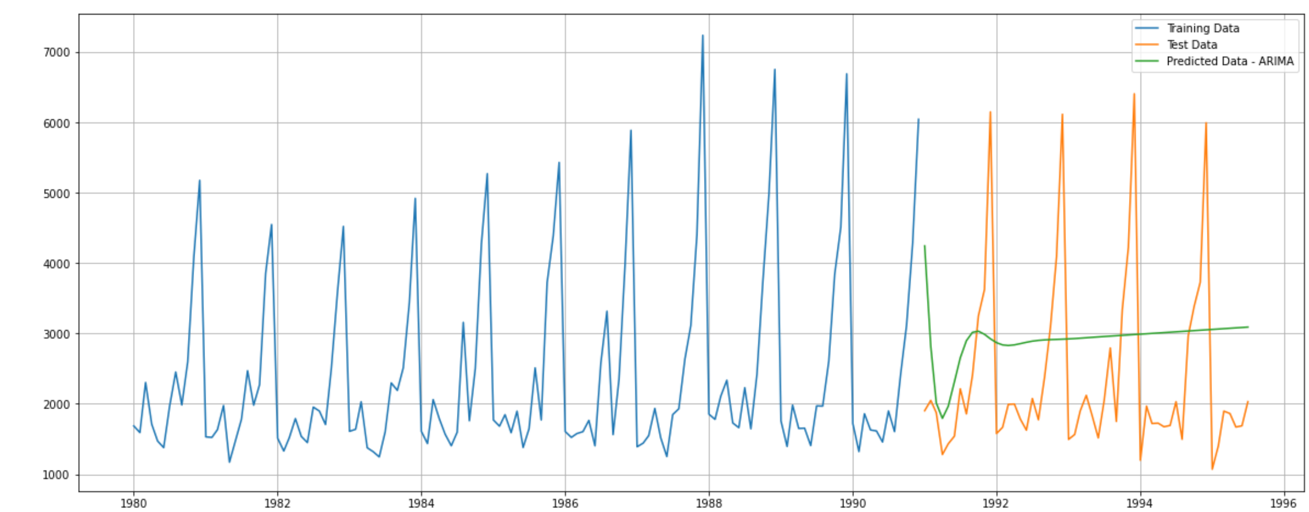


Fig. 28a ARIMA Plot of the 'Sparkling' wine sales dataset

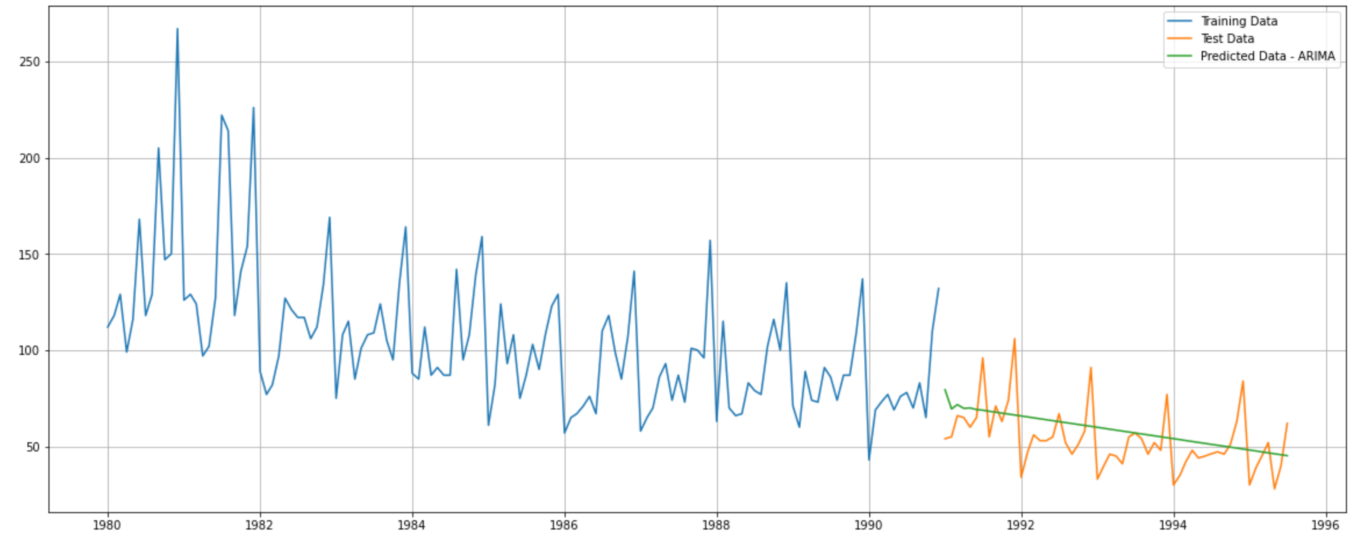


Fig. 28b ARIMA Plot of the ‘Rose’ wine sales dataset

Build an SARIMA model

SPARKLING DATASET

Seasonal ARIMA (p,d,q)(P,D,Q)F Model

For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the ‘F’ parameter indicates the seasonality/seasonal effects over a particular period. We can follow the Box-Jenkins method over here as well to decide the ‘p’, ‘q’, ‘P’ and ‘Q’ values.

For the SARIMA models, we can also estimate ‘p’, ‘q’, ‘P’ and ‘Q’ by looking at the lowest AIC values.

•The seasonal parameter ‘F’ can be determined by looking at the ACF plots. The ACF plot is expected to show a spike at multiples of ‘F’ thereby indicating a presence of seasonality.

•Also, for Seasonal models, the ACF and the PACF plots are going to behave a bit different and they will not always continue to decay as the number of lags increase.

Let us look at the ACF plot to understand the seasonal parameter for the SARIMA model

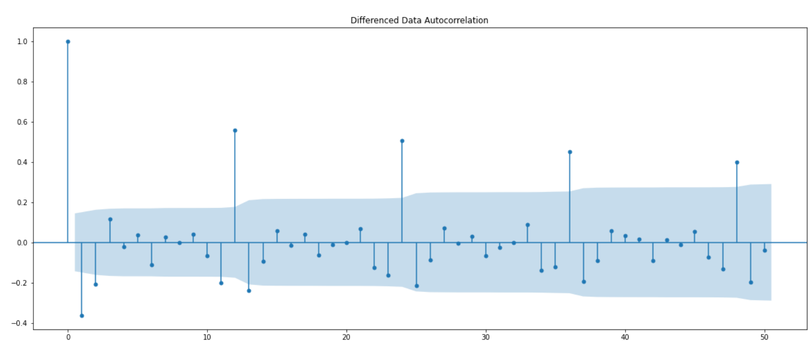


Fig. 29a ACF Plot of the ‘Sparkling’ wine sales dataset

We will run our auto SARIMA models by setting seasonality at 12 looking at the ACF plot.

For deciding the 'P' and 'Q' values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of 'F' and see where these cut-off (for appropriate confidence interval bands)

SPARKLING DATASET

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584248
53	(1, 1, 2)	(2, 0, 2, 12)	1556.076790
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121579
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 25 Jul 2021   AIC                           1555.584
Time:                                    12:21:00   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6283      0.255     -2.464      0.014      -1.128      -0.128
ma.L1         -0.1040      0.225     -0.463      0.644      -0.545       0.337
ma.L2         -0.7277      0.154     -4.736      0.000      -1.029      -0.427
ar.S.L12       1.0439      0.014     72.834      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
===================================================================================

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

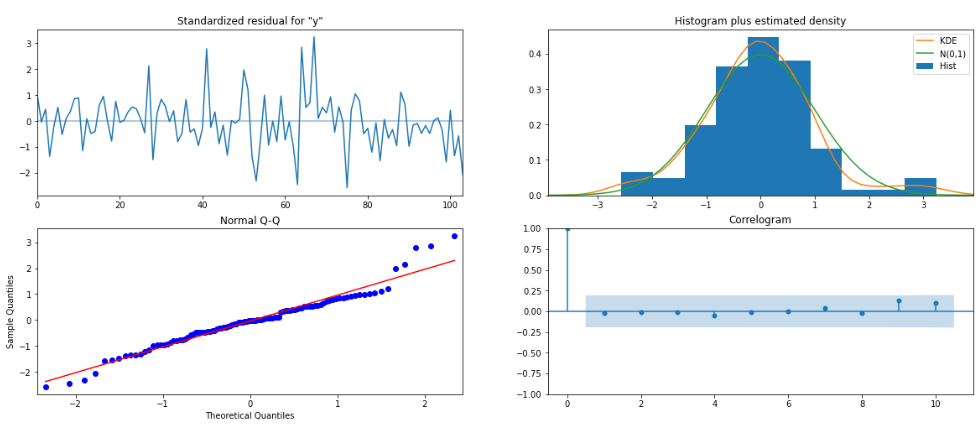


Fig.30a Diagnostics plot of the auto ARIMA (AIC) – Sparkling dataset

ROSE DATASET

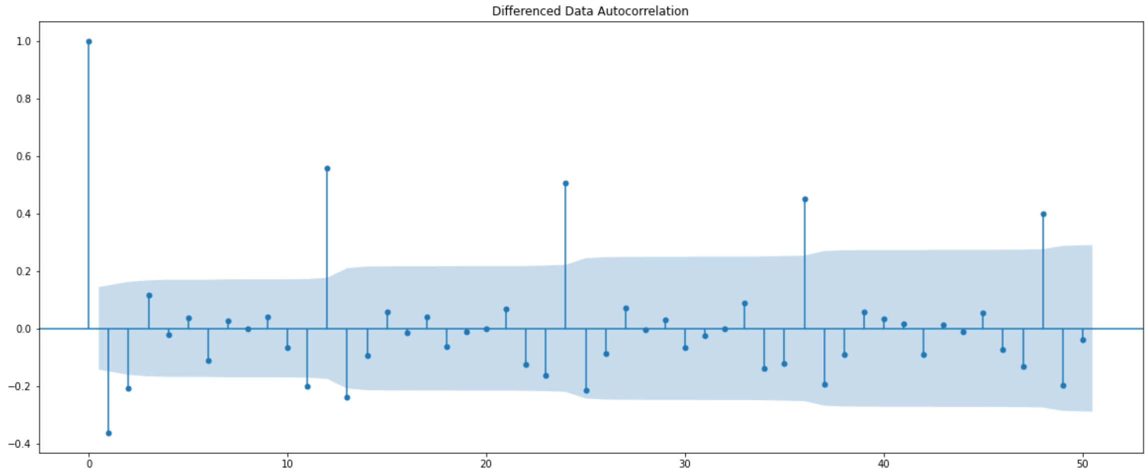


Fig. 29b ACF Plot of the 'Rose' wine sales dataset

```
Examples of some parameter combinations for Model.
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
```

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.903048
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -446.366
Date:                            Sun, 25 Jul 2021   AIC                            906.732
Time:                                    17:15:52   BIC                            925.243
Sample:                                         0   HQIC                           914.231
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1141      0.369     -0.310      0.757      -0.837       0.608
ma.L1         -0.6700    295.772     -0.002      0.998    -580.373     579.033
ma.L2         -0.3300     97.673     -0.003      0.997    -191.765     191.105
ar.S.L12       0.6255      0.059     10.547      0.000       0.509       0.742
ma.S.L12      -0.1613      0.126     -1.285      0.199      -0.407       0.085
ma.S.L24       0.1133      0.134      0.844      0.399      -0.150       0.376
sigma2       299.9783   8.87e+04      0.003      0.997   -1.74e+05    1.74e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):                 0.59
Prob(Q):                              0.78   Prob(JB):                         0.74
Heteroskedasticity (H):               0.71   Skew:                             0.08
Prob(H) (two-sided):                  0.31   Kurtosis:                         3.33
===================================================================================

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

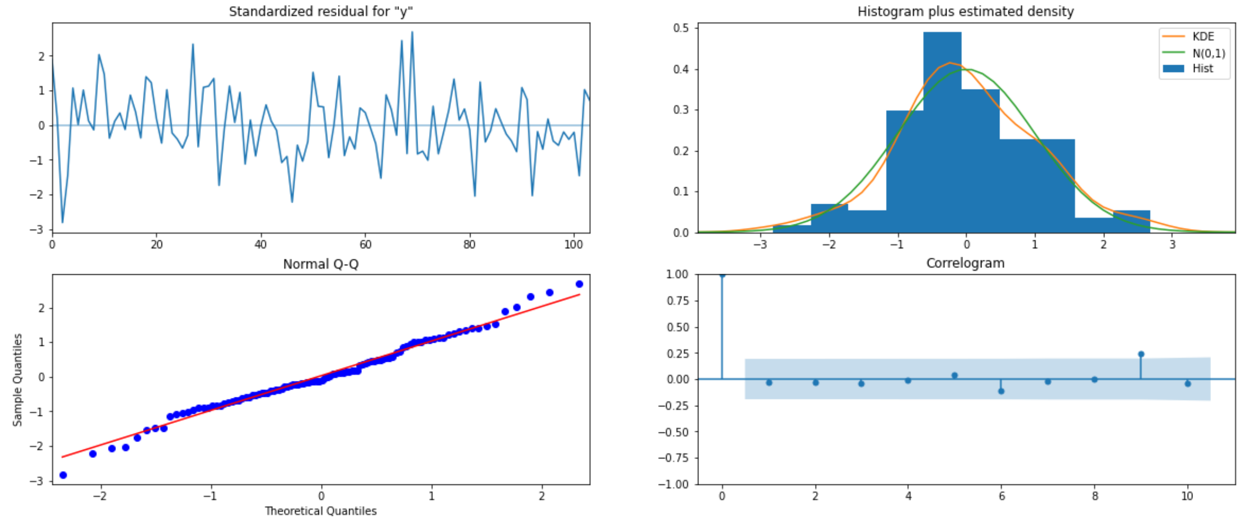


Fig.30b Diagnostics plot of the auto ARIMA (AIC) – Rose dataset

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers.
RMSE of SARIMA model 31.42979344641327

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.

ARIMA - SPARKLING DATASET

An ARIMA model consists of the Auto-Regressive (AR) part and the Moving Average (MA) part after we have made the Time Series stationary by taking the correct degree/order of differencing.

•The AR order is selected by looking at where the PACF plot cuts-off (for appropriate confidence interval bands) and the MA order is selected by looking at where the ACF plots cuts-off (for appropriate confidence interval bands)

•The correct degree or order of difference gives us the value of ‘d’ while the ‘p’ value is for the order of the AR model and the ‘q’ value is for the order of the MA model.

•This is the Box-Jenkins methodology for building the ARIMA models.

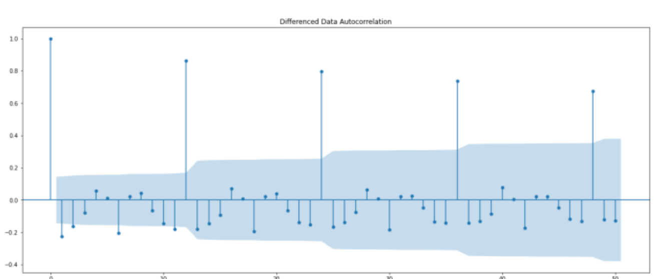


Fig.31a ACF Plot – Sparkling dataset

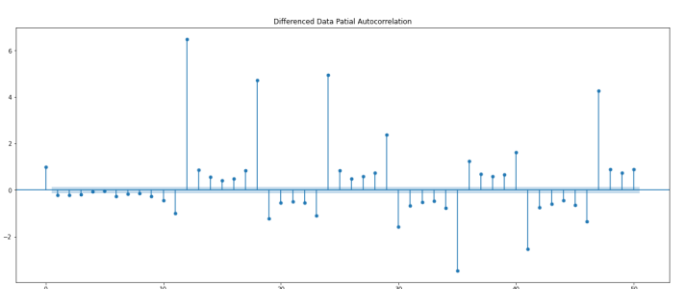


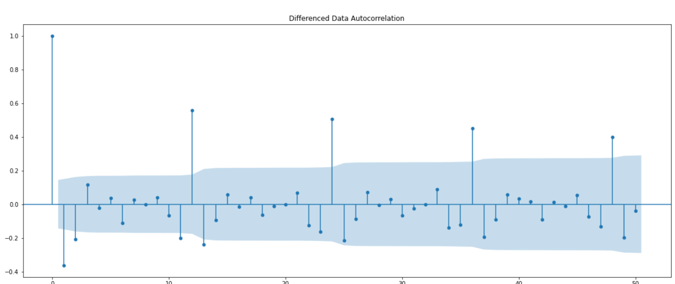
Fig.31b PACF Plot – Sparkling dataset

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.309
Method:                       css-mle   S.D. of innovations           1012.730
Date:                Sun, 25 Jul 2021   AIC                           2210.619
Time:                        12:21:04   BIC                           2227.870
Sample:                    02-01-1980   HQIC                          2217.628
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5843      0.518     10.790      0.000       4.570       6.599
ar.L1.D.Sparkling     1.2700      0.074     17.048      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5604      0.074     -7.620      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9978      0.042    -47.093      0.000      -2.081      -1.915
ma.L2.D.Sparkling     0.9978      0.042     23.501      0.000       0.915       1.081
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1333           -0.7073j            1.3359           -0.0888
AR.2            1.1333           +0.7073j            1.3359            0.0888
MA.1            1.0004           +0.0000j            1.0004            0.0000
MA.2            1.0019           +0.0000j            1.0019            0.0000
-----------------------------------------------------------------------------
```

RMSE of manual ARIMA (2,1,2) 1374.546023727508

ARIMA - ROSE DATASET

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



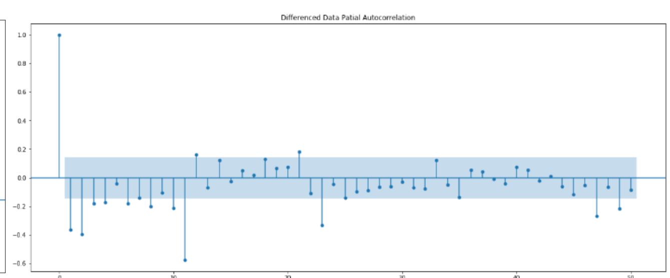


Fig.32a ACF Plot – Rose dataset

Fig.32b PACF Plot – Rose dataset

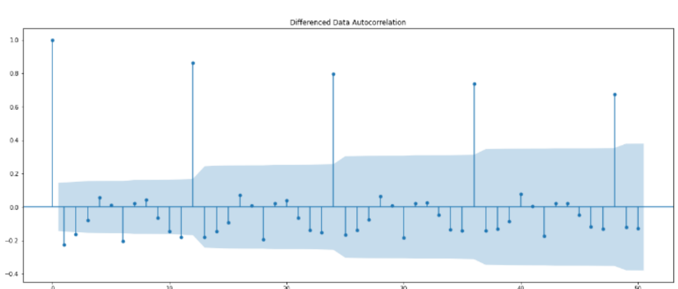
```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -633.649
Method:                       css-mle   S.D. of innovations             29.975
Date:                Sun, 25 Jul 2021   AIC                           1279.299
Time:                        17:15:53   BIC                           1296.550
Sample:                    02-01-1980   HQIC                          1286.309
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4911      0.081     -6.076      0.000      -0.649      -0.333
ar.L1.D.Rose    -0.4383      0.218     -2.015      0.044      -0.865      -0.012
ar.L2.D.Rose     0.0269      0.109      0.246      0.806      -0.188       0.241
ma.L1.D.Rose    -0.3316      0.203     -1.633      0.102      -0.729       0.066
ma.L2.D.Rose    -0.6684      0.201     -3.332      0.001      -1.062      -0.275
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -2.0290           +0.0000j            2.0290            0.5000
AR.2           18.3389           +0.0000j           18.3389            0.0000
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.4961           +0.0000j            1.4961            0.5000
-----------------------------------------------------------------------------
```

RMSE of manual ARIMA (2,1,2) 15.342048385136696

SARIMA - SPARKLING DATASET

Seasonal ARIMA (p,d,q)(P,D,Q)F Model - For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the ‘F’ parameter indicates the seasonality/seasonal effects over a particular period.

We can follow the Box-Jenkins method over here as well to decide the ‘p’, ‘q’, ‘P’ and ‘Q’ values. For deciding the ‘P’ and ‘Q’ values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of ‘F’ and see where these cut-off (for appropriate confidence interval bands)



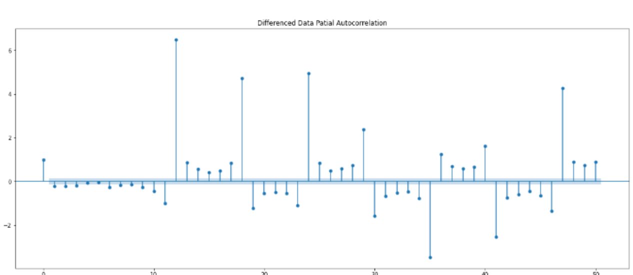


Fig.33a ACF Plot – Differenced data

Fig.33b PACF Plot – Differenced data

We take a seasonal differencing (12) and check the series

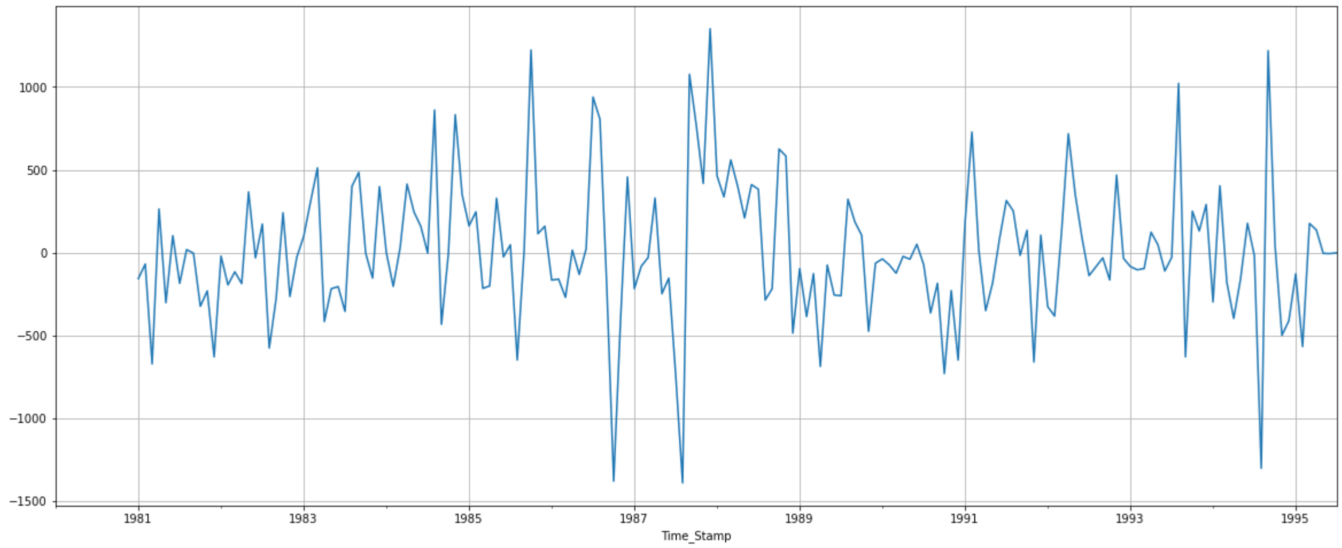


Fig.34a Seasonal Differenced data

We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

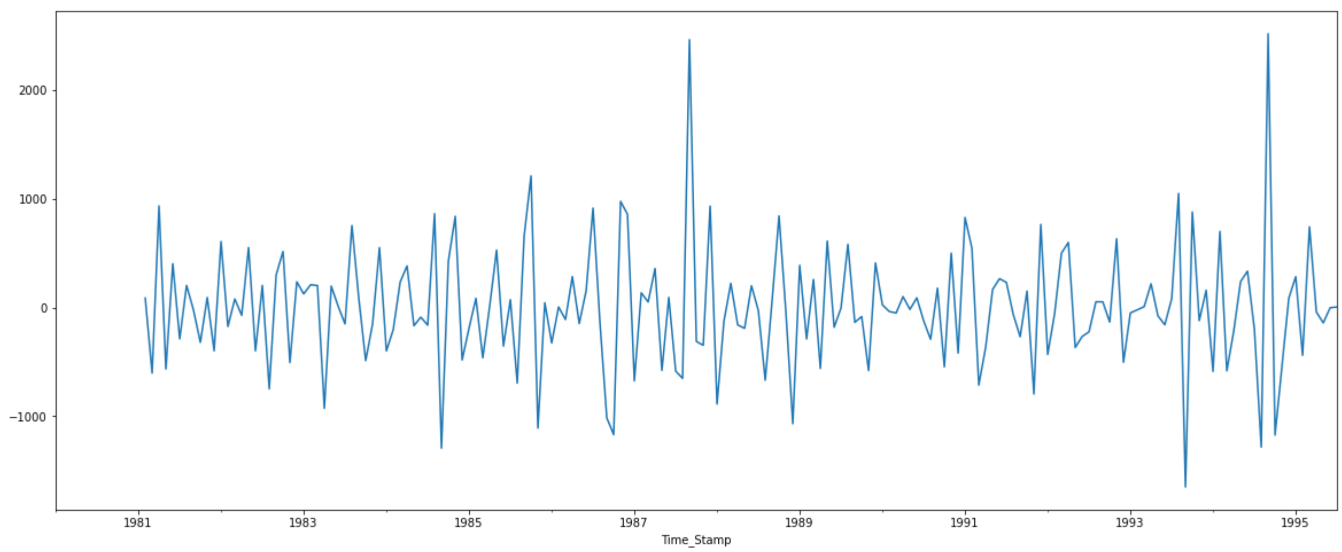


Fig.35a Differencing of first order on the seasonally differenced series

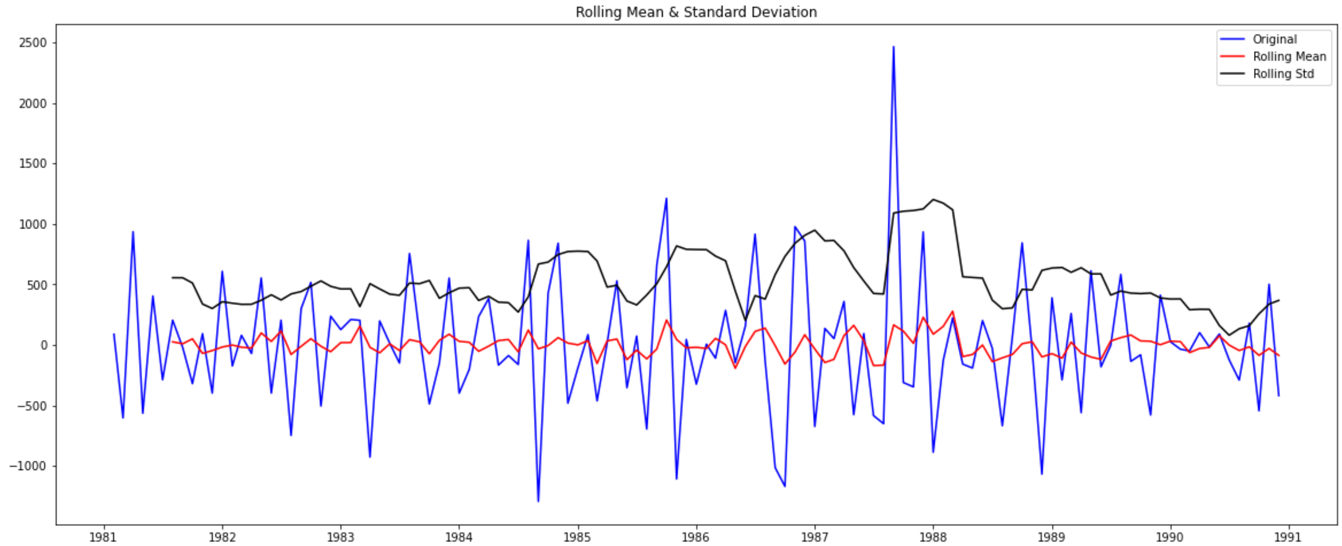


Fig. 36 Rolling mean & Standard Deviation Plot of the ‘Sparkling’ wine – SARIMA

```
Results of Dickey-Fuller Test:
Test Statistic                 -3.342905
p-value                         0.013066
#Lags Used                     10.000000
Number of Observations Used   108.000000
Critical Value (1%)            -3.492401
Critical Value (5%)            -2.888697
Critical Value (10%)           -2.581255
dtype: float64
```

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(2, 1, [1, 2], 12)   Log Likelihood              -722.996
Date:                           Sun, 25 Jul 2021   AIC                           1455.991
Time:                                   12:21:11   BIC                           1468.708
Sample:                                        0   HQIC                          1461.128
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12      -0.2445      0.879     -0.278      0.781      -1.967       1.478
ar.S.L24      -0.2107      0.257     -0.820      0.412      -0.714       0.293
ma.S.L12      -0.1220      0.860     -0.142      0.887      -1.807       1.563
ma.S.L24       0.0444      0.502      0.088      0.930      -0.940       1.029
sigma2      2.806e+05    3.2e+04      8.764      0.000    2.18e+05    3.43e+05
===================================================================================
Ljung-Box (L1) (Q):                  12.20   Jarque-Bera (JB):                37.03
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.76   Skew:                             0.77
Prob(H) (two-sided):                  0.44   Kurtosis:                         5.66
===================================================================================

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

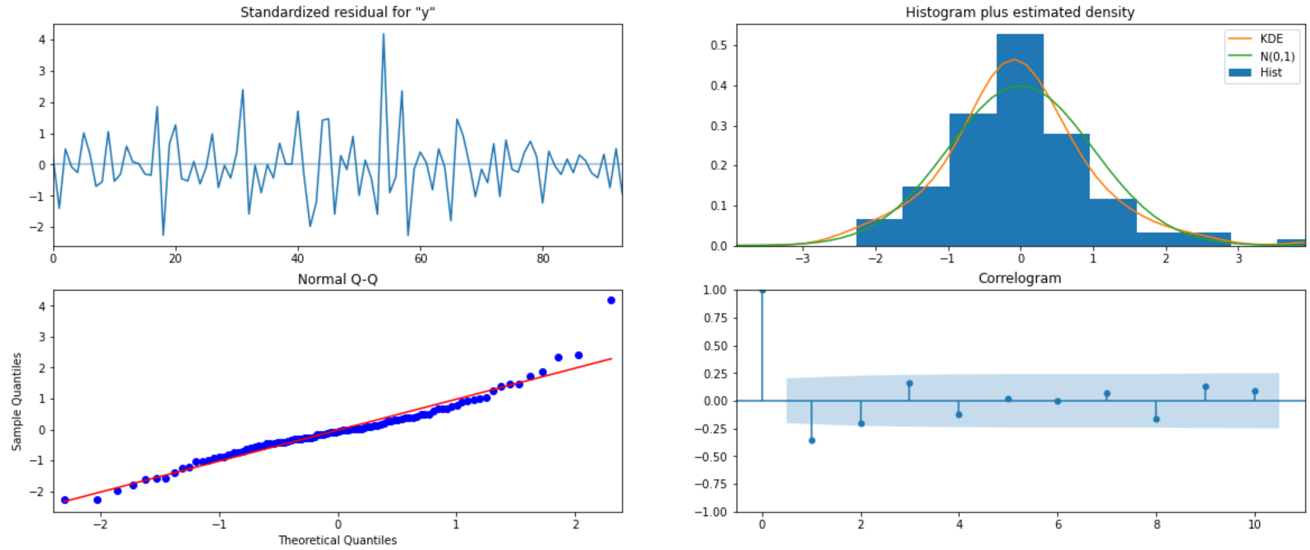
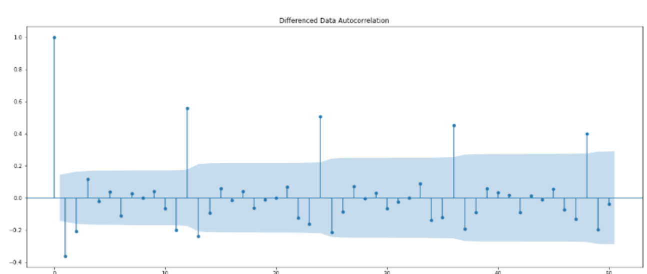


Fig. 37 Diagnostics Plot of the 'Sparkling' wine – SARIMA

RMSE Manual SARIMA 1757.7266967386927

SARIMA - ROSE DATASET



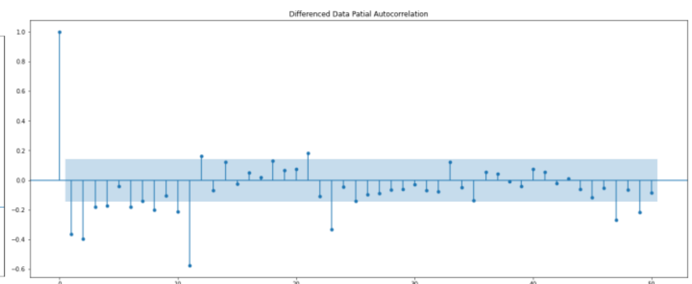


Fig.38a ACF Plot – Differenced data

Fig.38b PACF Plot – Differenced data

Now we take a seasonal differencing and check the series

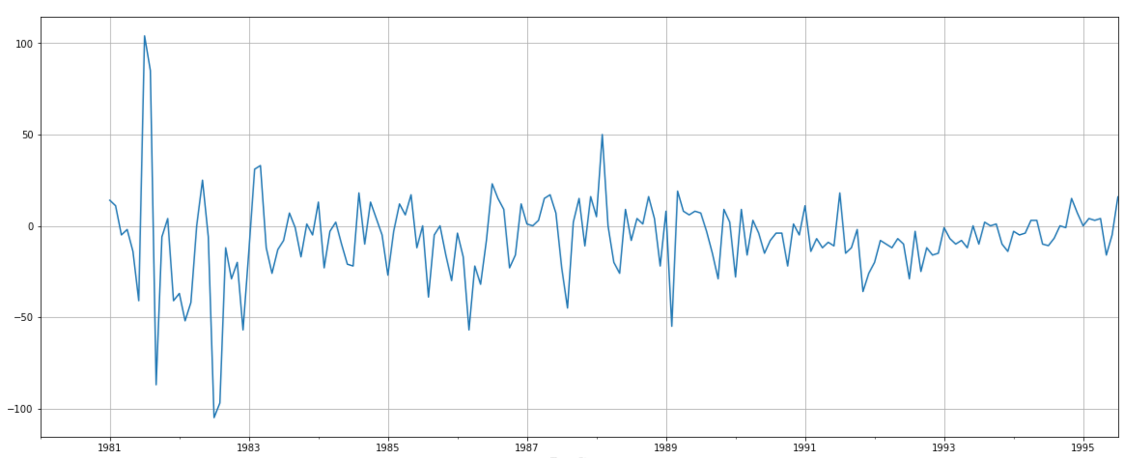


Fig.39 Seasonal Differenced data

We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

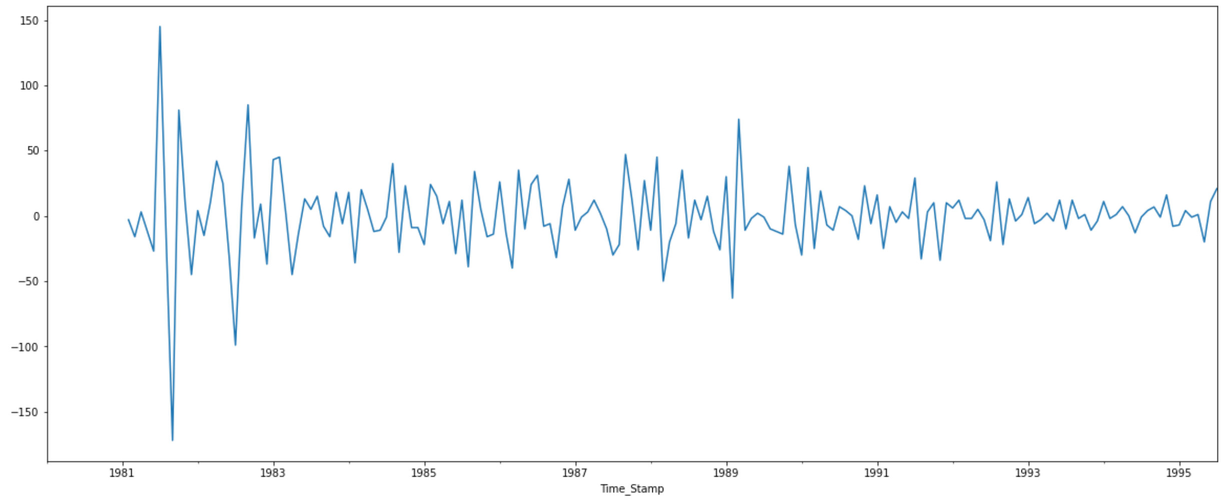


Fig.40 Differencing of first order on the seasonally differenced series

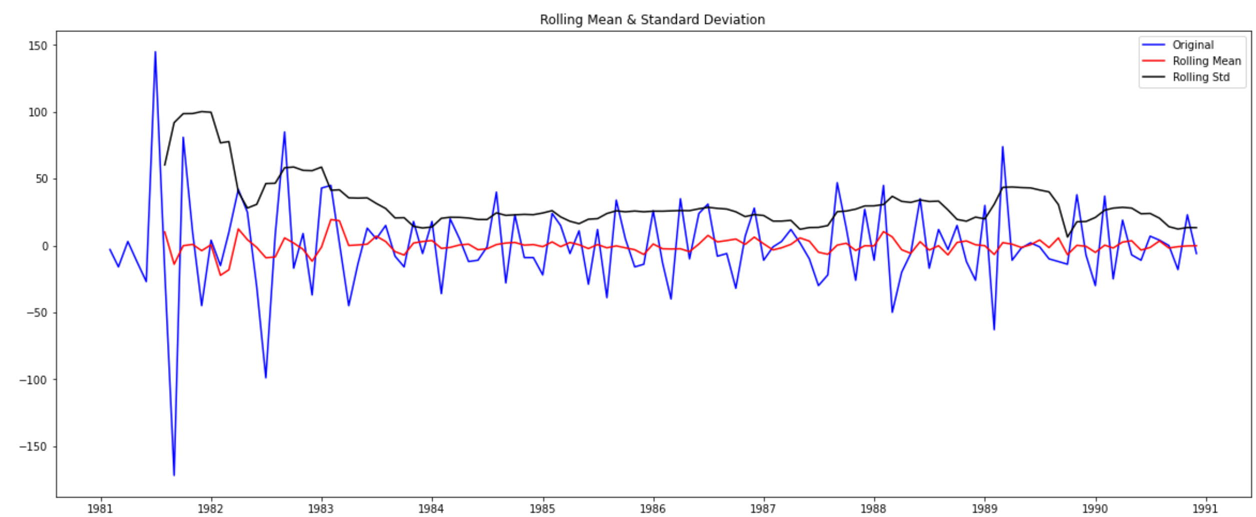
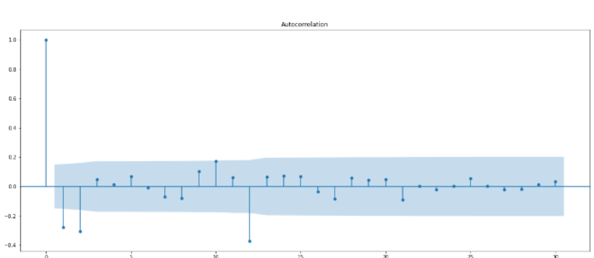
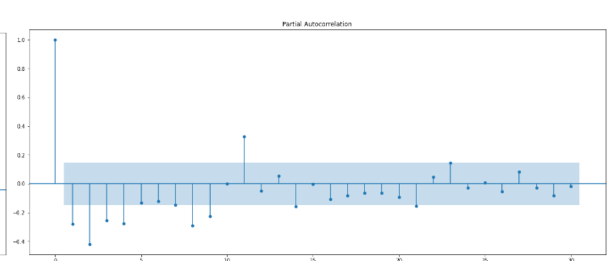


Fig. 41 Rolling mean & Standard Deviation Plot of the 'Rose' wine – SARIMA

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.692348
p-value                          0.004222
#Lags Used                      11.000000
Number of Observations Used    107.000000
Critical Value (1%)             -3.492996
Critical Value (5%)             -2.888955
Critical Value (10%)            -2.581393
dtype: float64
```





```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(2, 1, [1, 2], 12)   Log Likelihood            -416.314
Date:                            Sun, 25 Jul 2021   AIC                            842.628
Time:                                    17:15:56   BIC                            855.345
Sample:                                         0   HQIC                           847.765
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12       0.0063      0.433      0.015      0.988      -0.842       0.855
ar.S.L24      -0.0467      0.063     -0.745      0.456      -0.170       0.076
ma.S.L12      -0.6530      0.480     -1.361      0.174      -1.593       0.287
ma.S.L24      -0.0983      0.298     -0.329      0.742      -0.683       0.487
sigma2       379.3885     66.867      5.674      0.000     248.333     510.444
===================================================================================
Ljung-Box (L1) (Q):                  18.51   Jarque-Bera (JB):                 0.81
Prob(Q):                              0.00   Prob(JB):                         0.67
Heteroskedasticity (H):               1.13   Skew:                             0.22
Prob(H) (two-sided):                  0.74   Kurtosis:                         3.12
===================================================================================

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

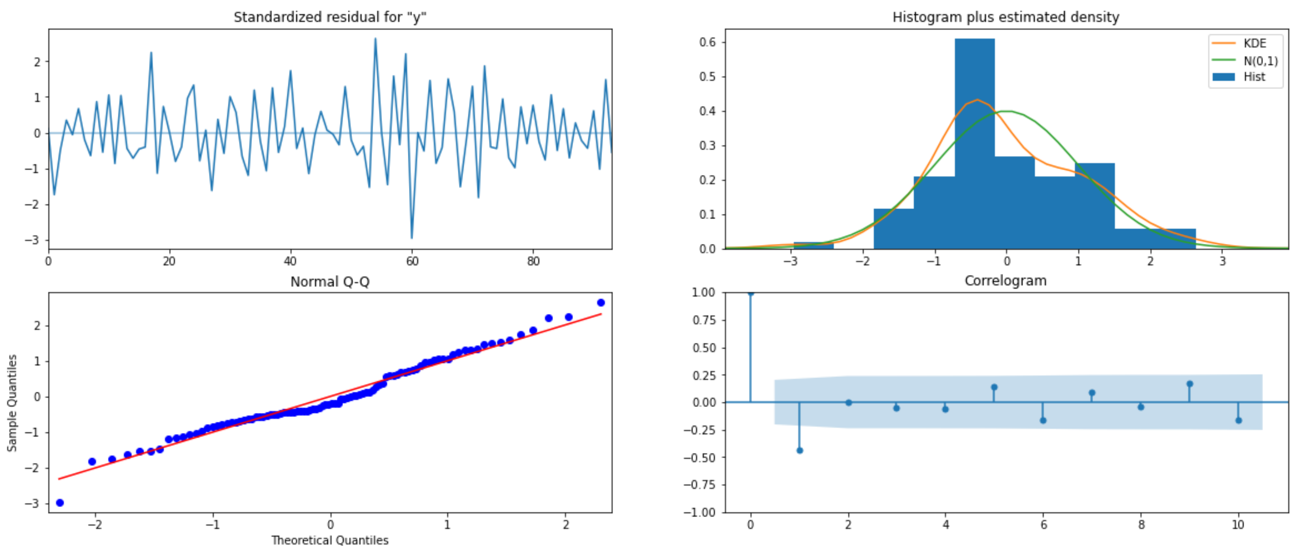


Fig. 42 Diagnostics Plot of the ‘Rose’ wine – SARIMA

RMSE Manual SARIMA 15.004343981604407

INFERENCE

8. 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.

SPARKLING DATASET

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	402.938530
Alpha=0.02 Optimized SES	1278.497798
Alpha=0.02,Beta=0.38 Optimized DES	1275.874751
Alpha=0.1,Beta=0.4,Gamma=0.1:Optimized TES	342.934716
Linear RegressionOnTime	1389.135175
NaiveModel	3864.279352
Simple Average Model	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
ARIMA(2, 1, 2)	1374.546024
SARIMA(1,1,2)(1,0,2,12)	528.611364
Manual ARIMA(2,1,2)	1374.546024
SARIMA(0,1,0)(2,1,2,12)	1757.726697

ROSE DATASET

	Test RMSE
Alpha = 0.09 SES	36.748402
Alpha=1.49e-08,Beta = 1.665e-10 DES	15.255480
Alpha=0.088,Beta=9e-06,,Gamma=0.0003 TES	14.222850
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	19.337756
Alpha=0.07 Tuned SES	36.387162
Alpha=0.04,Beta=0.47 Tuned DES	14.455710
Alpha=0.1,Beta=0.4,Gamma=0.3 Tuned TES	11.986168
Linear RegressionOnTime	15.255492
NaiveModel	79.672475
Simple Average Model	53.413298
2pointTrailingMovingAverage	11.529985
4pointTrailingMovingAverage	14.444375
6pointTrailingMovingAverage	14.554986
9pointTrailingMovingAverage	14.721520
ARIMA(2, 1, 2)	15.342048
SARIMA(1,1,2)(1,0,2,12)	31.429793
Manual ARIMA(2,1,2)	15.342048
SARIMA(0,1,0)(2,1,2,12)	15.004344

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

SPARKLING DATASET

```
                                       SARIMAX Results
==========================================================================================
Dep. Variable:                             Sparkling   No. Observations:                187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood              -1173.399
Date:                           Sun, 25 Jul 2021   AIC                          2360.798
Time:                                   12:21:15   BIC                          2382.281
Sample:                               01-01-1980   HQIC                         2369.522
                                     - 07-01-1995
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8322      0.079    -10.517      0.000      -0.987      -0.677
ma.L2         -0.1231      0.082     -1.495      0.135      -0.284       0.038
ar.S.L12       0.6497      0.670      0.970      0.332      -0.664       1.963
ar.S.L24       0.3700      0.683      0.542      0.588      -0.969       1.709
ma.S.L12      -0.2320      0.669     -0.347      0.729      -1.543       1.079
ma.S.L24      -0.2436      0.425     -0.573      0.566      -1.076       0.589
sigma2      1.467e+05   1.29e+04     11.342      0.000    1.21e+05    1.72e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                42.80
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               0.94   Skew:                             0.60
Prob(H) (two-sided):                  0.83   Kurtosis:                         5.24
===================================================================================

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

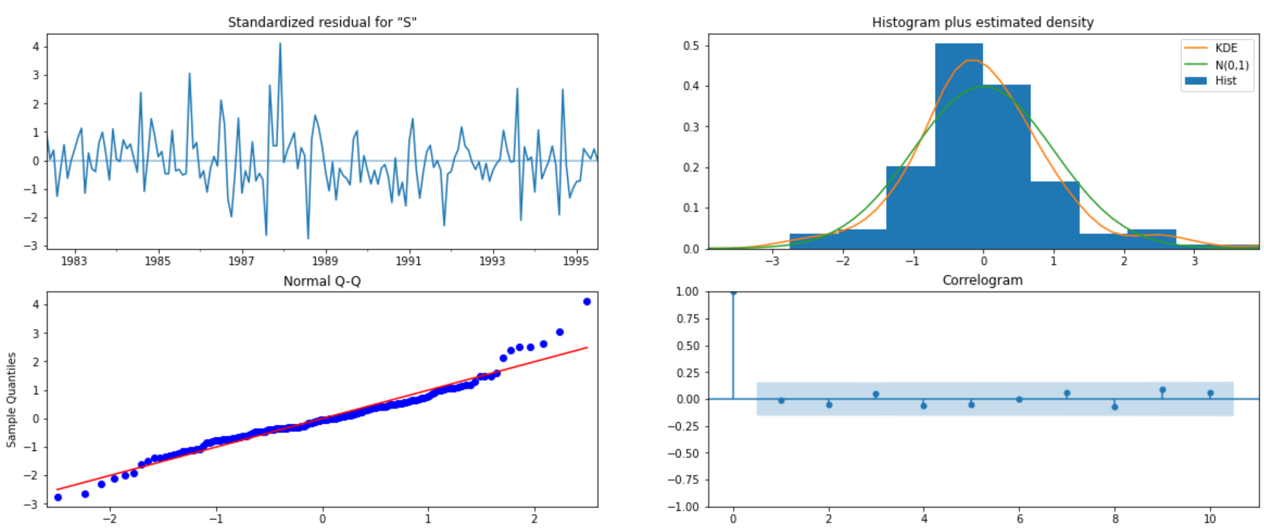


Fig. 43 Diagnostics Plot of the 'Sparkling' wine – Full Model SARIMA

RMSE of the Full Model 524.6502508543183

Predict 12 months into the future with appropriate confidence intervals/bands.

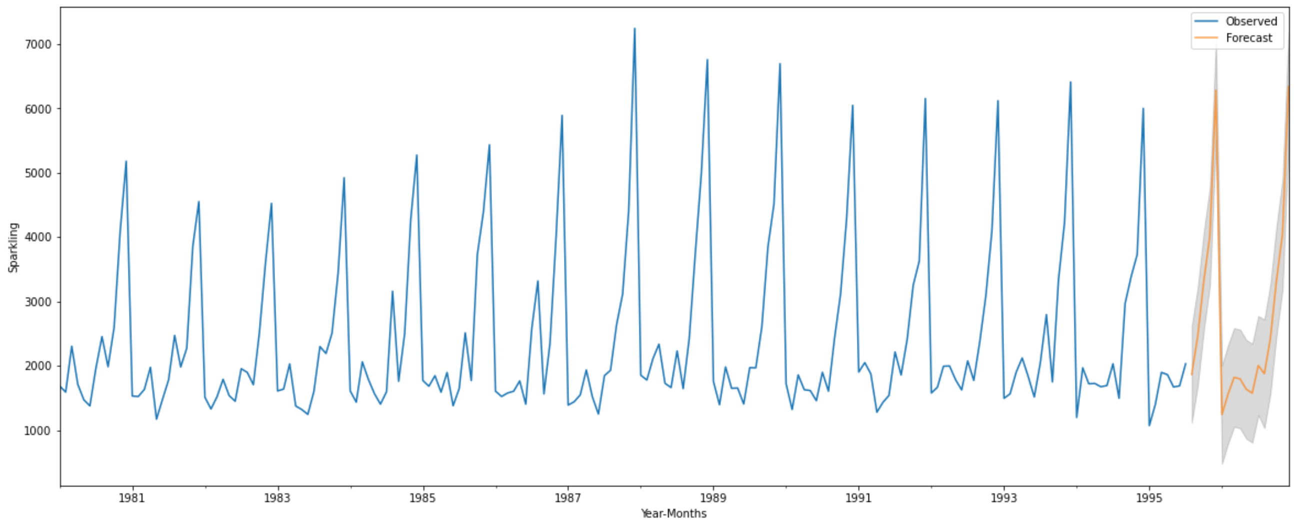


Fig. 44 Forecast plot of the 'Sparkling' wine – Full Model SARIMA

ROSE DATASET

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -647.280
Date:                              Sun, 25 Jul 2021   AIC                           1308.559
Time:                                      17:15:58   BIC                           1330.042
Sample:                                  01-01-1980   HQIC                          1317.283
                                       - 07-01-1995
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7666      0.088     -8.755      0.000      -0.938      -0.595
ma.L2         -0.1381      0.081     -1.706      0.088      -0.297       0.021
ar.S.L12       0.3972      0.052      7.615      0.000       0.295       0.499
ar.S.L24       0.3362      0.049      6.817      0.000       0.240       0.433
ma.S.L12       0.0147      0.089      0.165      0.869      -0.160       0.189
ma.S.L24      -0.1453      0.099     -1.466      0.143      -0.340       0.049
sigma2       199.0703     21.376      9.313      0.000     157.174     240.967
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 8.76
Prob(Q):                              0.75   Prob(JB):                         0.01
Heteroskedasticity (H):               0.26   Skew:                             0.50
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.58
===================================================================================

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

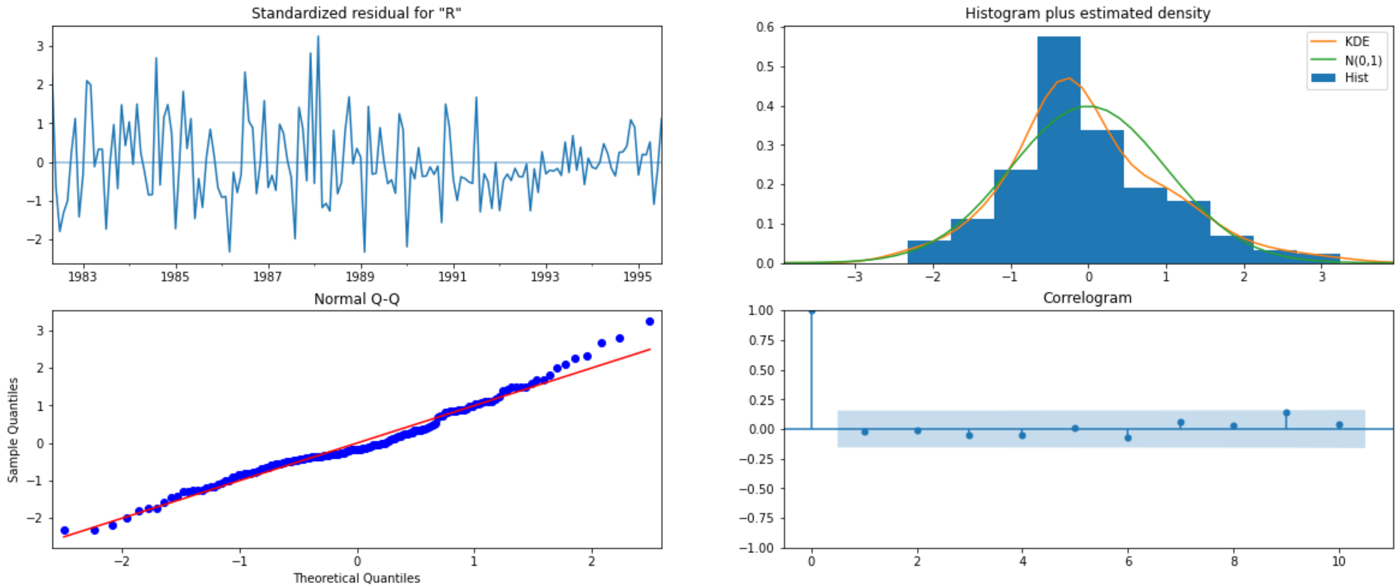


Fig. 45 Diagnostics Plot of the ‘Rose’ wine – Full Model SARIMA

RMSE of the Full Model 27.67120666670762

<u>Predict 12 months into the future with appropriate confidence intervals/bands.</u>

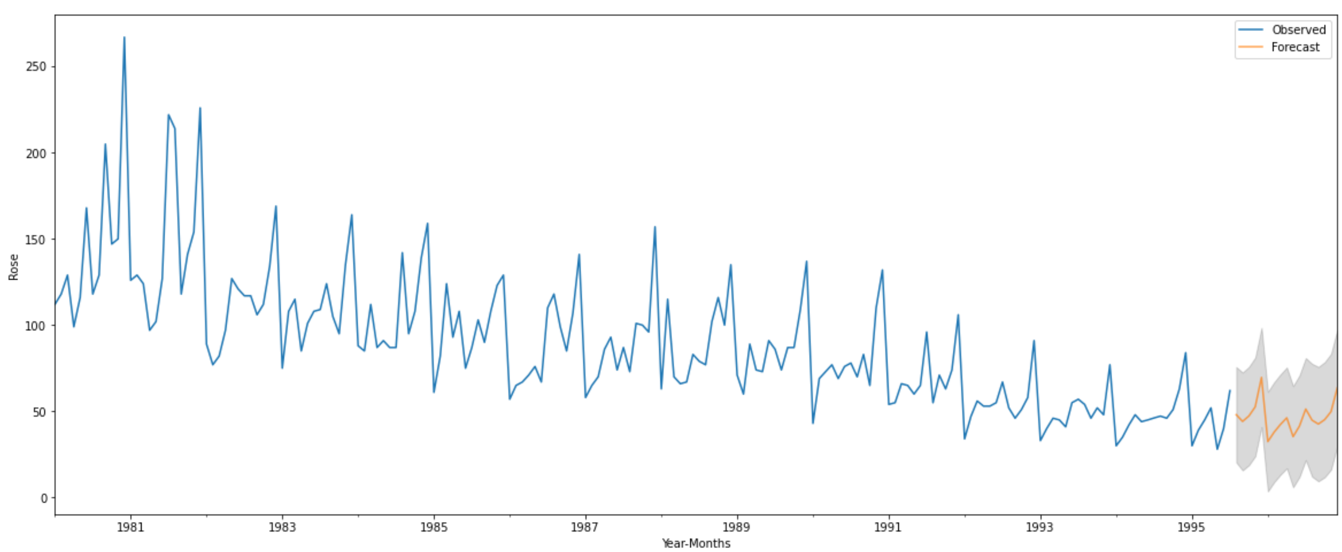


Fig. 46 Forecast plot of the ‘Sparkling’ wine – Full Model SARIMA

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

Sparkling Dataset:

RMSE (342.9) values for the Holt's Winter Exponential smoothing method when tuned to the best hyperparameters is the lowest compared with all the other forecasting models.

Rose Dataset:

RMSE (14.2) values for the Holt's Winter Exponential smoothing method (Triple Exponential Smoothing) is the lowest compared with all the other forecasting models.

Business Insights:

- The predicted 12-month future sales forecast follows the immediate last year for both 'Sparkling' and 'Rose' wine sales type.

- Holt Winters' exponential smoothing method perform well provided the hyperparameters are tuned appropriately.

- Different models give a different level of output for the predicted output. Hence domain knowledge should be used to compliment the findings from models.

- As the sales are more in last quarter, discount can be provided to customers in the other quarters and the loss thus obtained can be set right by hiking the prices when the demand is more.

- 'Rose' wine type consistently had a decreasing trend in the sales. Root cause analysis can be initiated, and steps can be taken to improve the failure modes in the product.

- We were able to create a generalized model that can forecast the future with minimum errors in the predicted outcome.

	Test RMSE
Alpha = 0.07 SES	1338.008384
Alpha=0.66,Beta=0.0001 DES	5291.879833
Alpha=0.11,Beta=0.012,Gamma=0.46 TES	378.626008
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	402.938530
Alpha=0.02 Optimized SES	1278.497798
Alpha=0.02,Beta=0.38 Optimized DES	1275.874751
Alpha=0.1,Beta=0.4,Gamma=0.1:Optimized TES	342.934716
Linear RegressionOnTime	1389.135175
NaiveModel	3864.279352
Simple Average Model	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
ARIMA(2, 1, 2)	1374.546024
SARIMA(1,1,2)(1,0,2,12)	528.611364
Manual ARIMA(2,1,2)	1374.546024
SARIMA(0,1,0)(2,1,2,12)	1757.726697
Full Model SARIMA(0,1,2)(2,0,2,12)	524.650251

	Test RMSE
Alpha = 0.09 SES	36.748402
Alpha=1.49e-08,Beta = 1.665e-10 DES	15.255480
Alpha=0.088,Beta=9e-06,,Gamma=0.0003 TES	14.222850
Alpha=0.11,Beta=0.049,Gamma=0.36 TES_AM	19.337756
Alpha=0.07 Tuned SES	36.387162
Alpha=0.04,Beta=0.47 Tuned DES	14.455710
Alpha=0.1,Beta=0.4,Gamma=0.3 Tuned TES	11.986168
Linear RegressionOnTime	15.255492
NaiveModel	79.672475
Simple Average Model	53.413298
2pointTrailingMovingAverage	11.529985
4pointTrailingMovingAverage	14.444375
6pointTrailingMovingAverage	14.554986
9pointTrailingMovingAverage	14.721520
ARIMA(2, 1, 2)	15.342048
SARIMA(1,1,2)(1,0,2,12)	31.429793
Manual ARIMA(2,1,2)	15.342048
SARIMA(0,1,0)(2,1,2,12)	15.004344
Full Model SARIMA(0,1,2)(2,0,2,12)	27.671207