TSF PROJECT BUSSINESS REPORT

Kratik Mehta

PGP-DSBA Feb 2022

Table of Contents

List of Figures

List of Tables

Problem Statement

Executive Summary

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

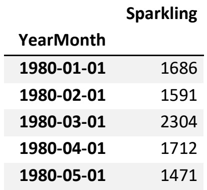
Data sets for the Problem: *Sparkling.csv* and *Rose.csv*

Sparkling Wine Sales Data

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

Sample of the Dataset

The below table shows the first 5 rows of the Sparkling dataset. To convert the data into a time series, the YearMonth column was converted into a timestamp index of the dataframe. The Sparkling column contains the sales of the mentioned wine.



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

Table 1: Sparkling Wine Data

Checking the types of variables in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

From the above output we can see that:

- There are **187 observations** in the data.
- The time series data has a **Datetime index** from **1980-01-01 to 1995-07-01**.
- The **Sparkling** column has the monthly sales values. The values are of **integer** data type.
- The dataset **has no missing values**.

Time Plot of the Time Series

The below plot shows the time plot of the Sparkling wine sales. There is **no obvious trend** present in the data. **Seasonality can be observed** clearly in the data. The **seasonal fluctuations are not constant** across the months, therefore **multiplicative models** might give better results. The trend and seasonality will be further explored during decomposition.

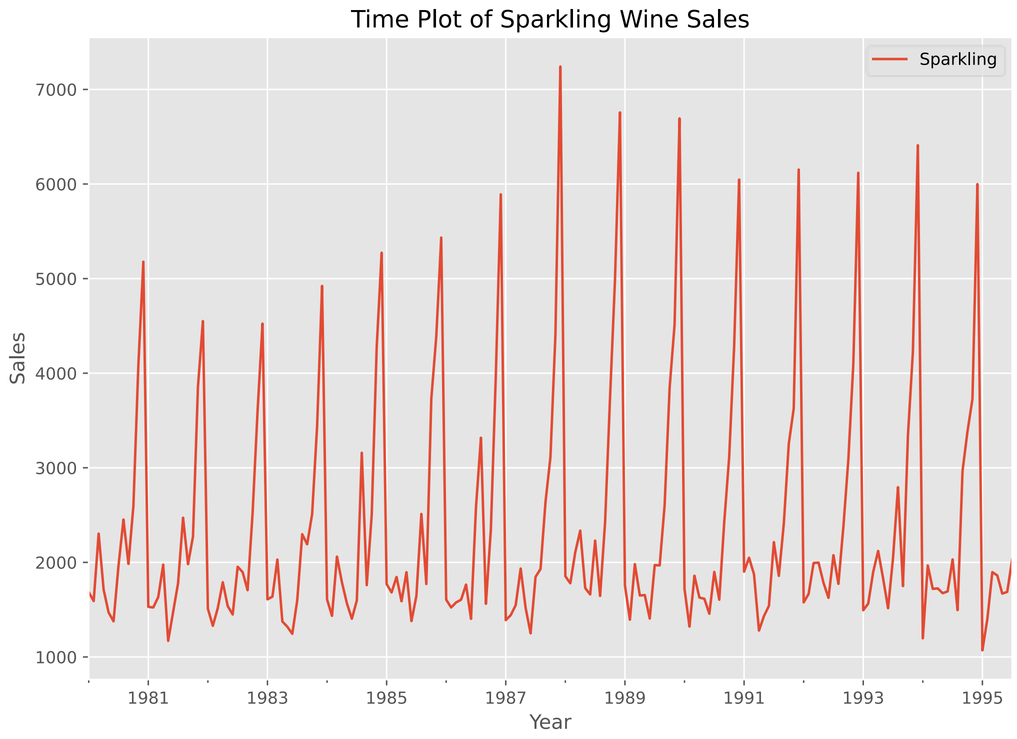
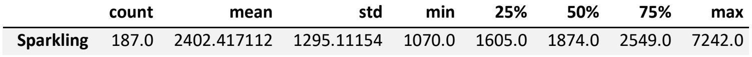


Figure 1: Time Plot of Sparkling Wine sales

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

Distribution of Sales data

The below table shows the 5-point summary of the Sparkling column:



	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

Table 2: Description of the Sparkling column

The overall mean sales of the Sparkling wine are 2402 and the standard deviation is 1295 which is quite high. The sales data is spread over a wide range from 1070 to 7242. Also, the mean and median are very different from each other indicating that the data is highly skewed. This can be confirmed from the below histogram:

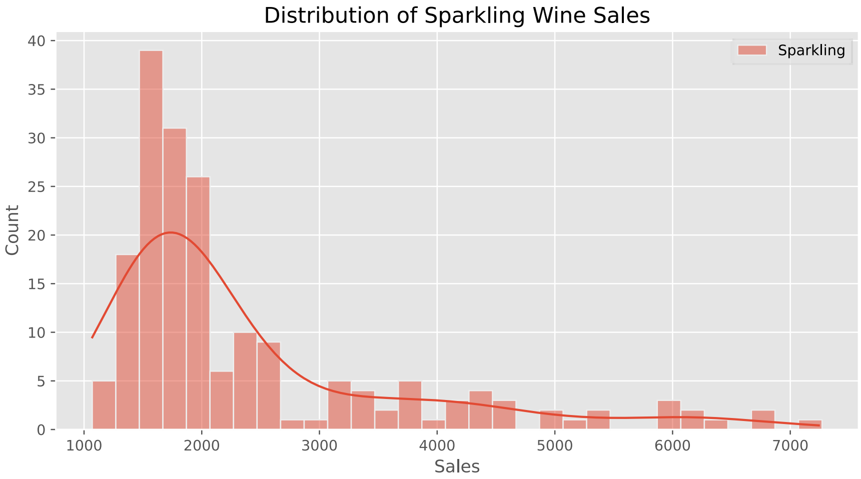


Figure 2: Distribution of the Sparkling wine sales

Yearly Pointplot for Sales

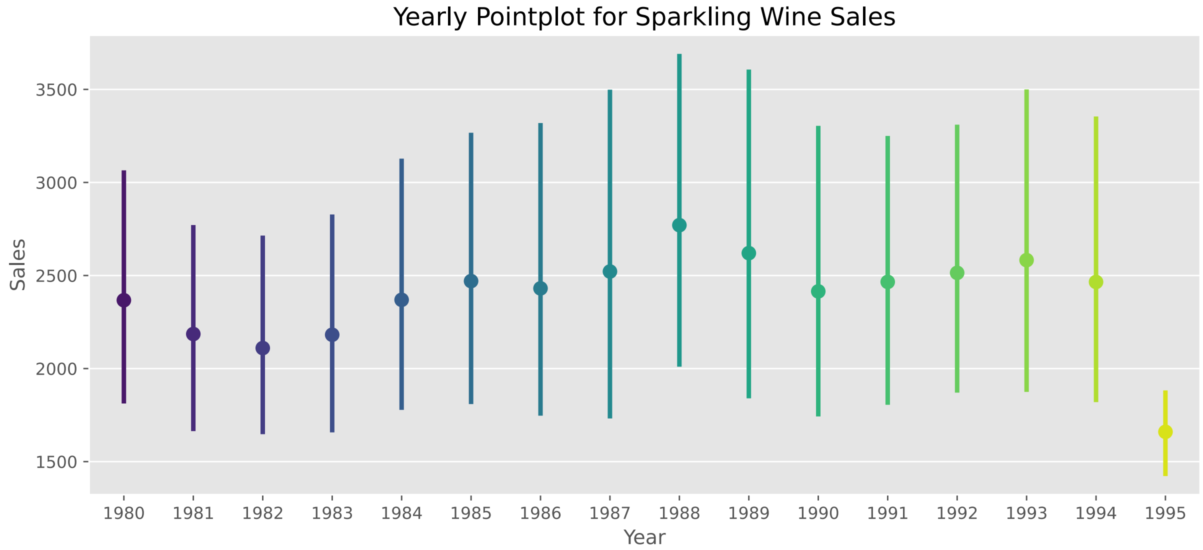


Figure 3: Yearly Point plot for Sparkling wine sales

The above point plot shows the mean sales of Sparkling wine over the years. The wine sales decreased from 1980 to 1982, then increased from 1982 to 1988 and then again decreased from 1988 to 1995. The sales were the lowest in the year 1995.

Monthly Sales Across Years

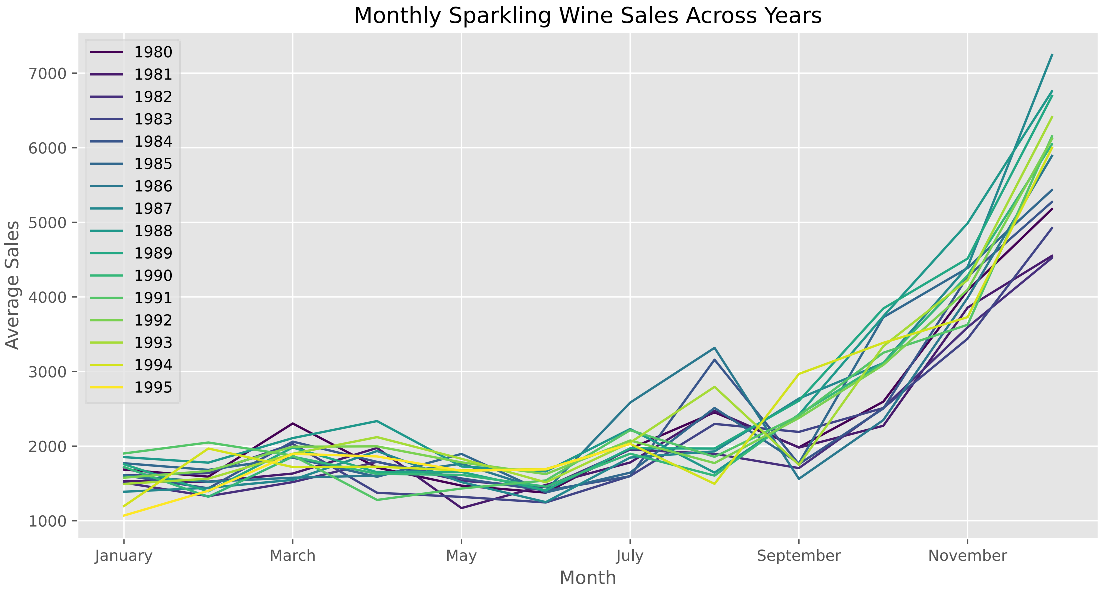


Figure 4: Monthly Sparkling Wine Sales Across Years

The above plot shows the average monthly wine sales across the years. There is a clear increasing trend in the sales over the months. For all the years, the wine sales are low in the first half of the year and then increases towards the end of the years. Also, the sales in the month of August for years 1982 and 1986 are higher than normal.

Decomposition

The original time series data of **the Sparkling wine sales was decomposed into its trend, seasonal and residual components using additive and multiplicative decomposition methods**.

Additive Decomposition

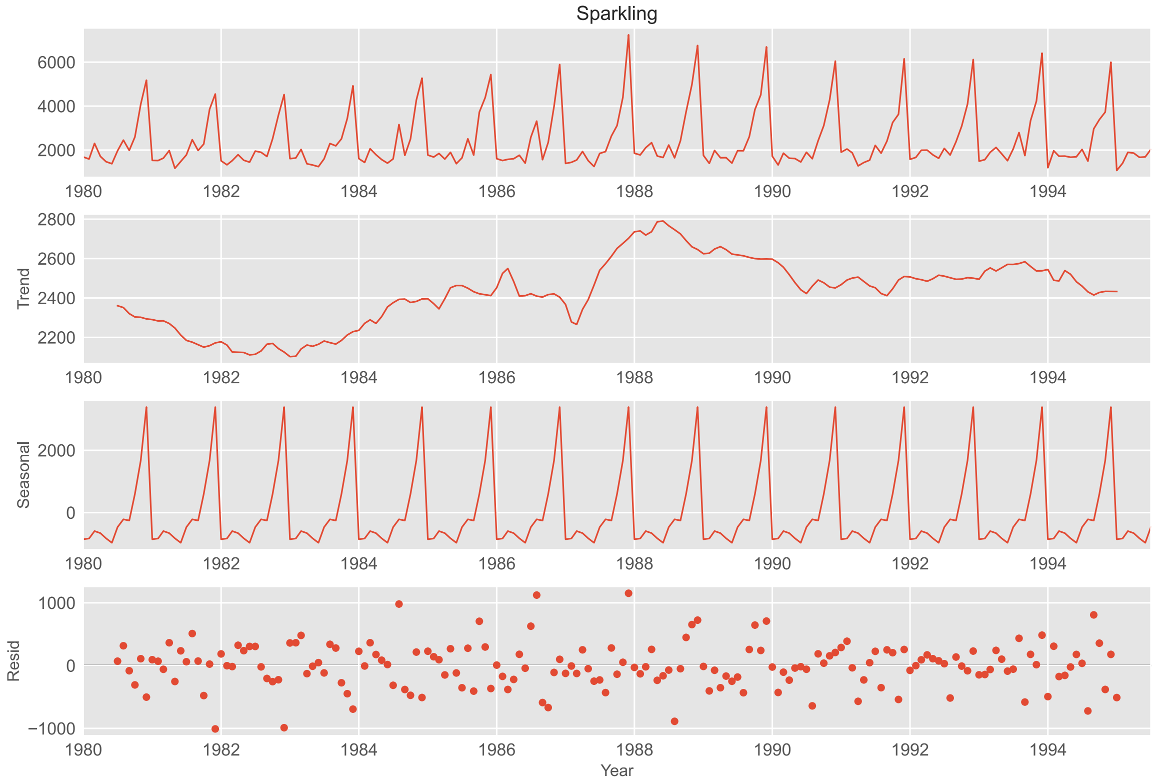


Figure 5: Additive Decomposition of Sparkling data

Multiplicative Decomposition

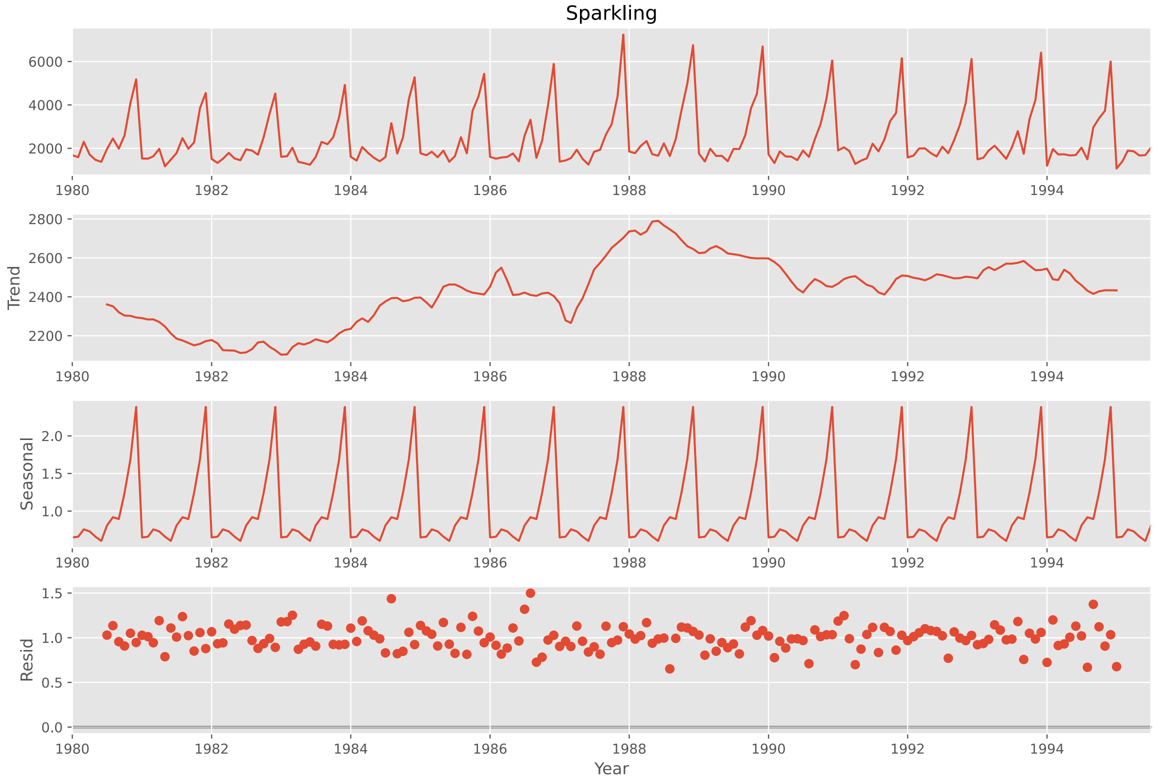


Figure 6: Multiplicative Decomposition of Sparkling data

From the above plots it can be seen that, the **data has an irregular trend**. We can also confirm that the **sales peak at the end of each year**. The **sales in the month of December are almost three times the average sales** while the **sales in the month of May is just about half of the average sales**. Also, a certain **pattern can be observed in the residual plot** indicating that **not all seasonal fluctuations are captured by the model**.

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

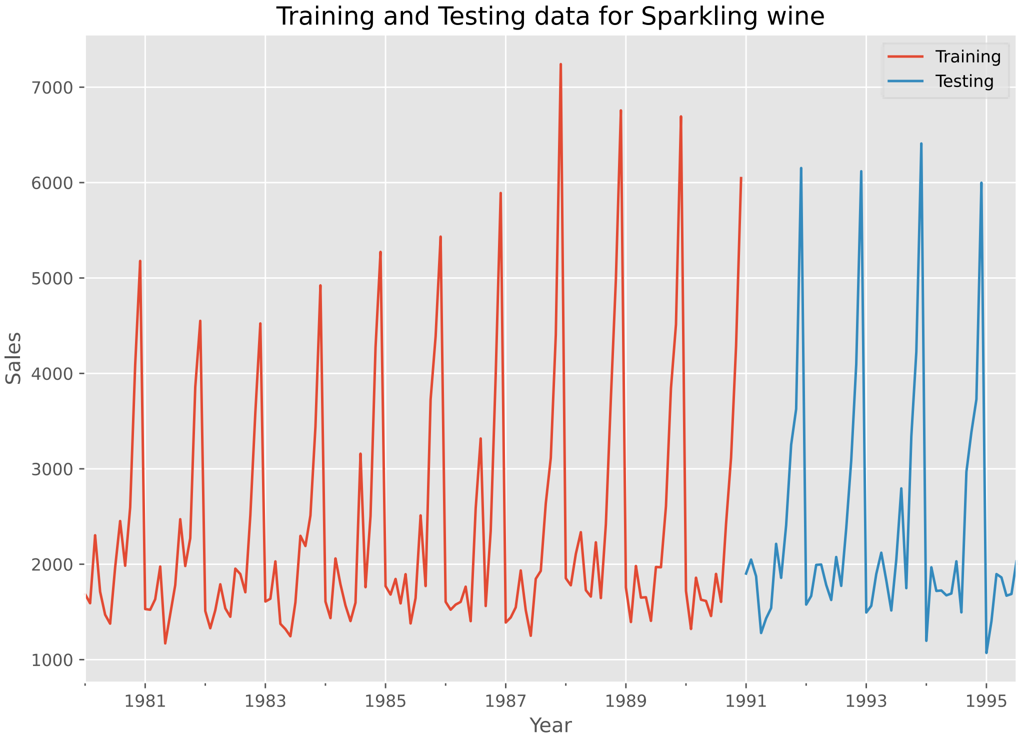


Figure 7: Training and Testing data for Sparkling wine

While splitting a time series data, the test set should contain data from the recent years. For the given problem, the **sales data from the year 1991 was put aside as a test set** and the **remaining data before 1991 was considered as a training set**. The above plot confirms that the split is done properly. The shape of the train and test sets after splitting are:

```
Shape of the training data: (132, 1)
Shape of the testing data: (55, 1)
```

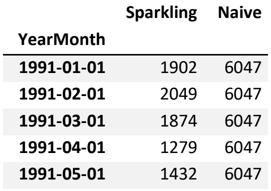
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on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

We will build various forecasting models on the Sparkling wine data and evaluate these models on the test set using RMSE metrics.

Naïve Forecast

A naïve forecast is a technique in **which the forecast for a given period is simply equal to the value observed in the previous period**. This leads to **all the forecast values in the test set to be equal to the last observation in the training set**.

The below table shows the naïve forecast for the Sparkling wine sales on the test set. We can see that the model predicted the same values for all the periods in the test set:



	Sparkling	Naive
YearMonth		
1991-01-01	1902	6047
1991-02-01	2049	6047
1991-03-01	1874	6047
1991-04-01	1279	6047
1991-05-01	1432	6047

Table 3: Naïve Forecast on the test set

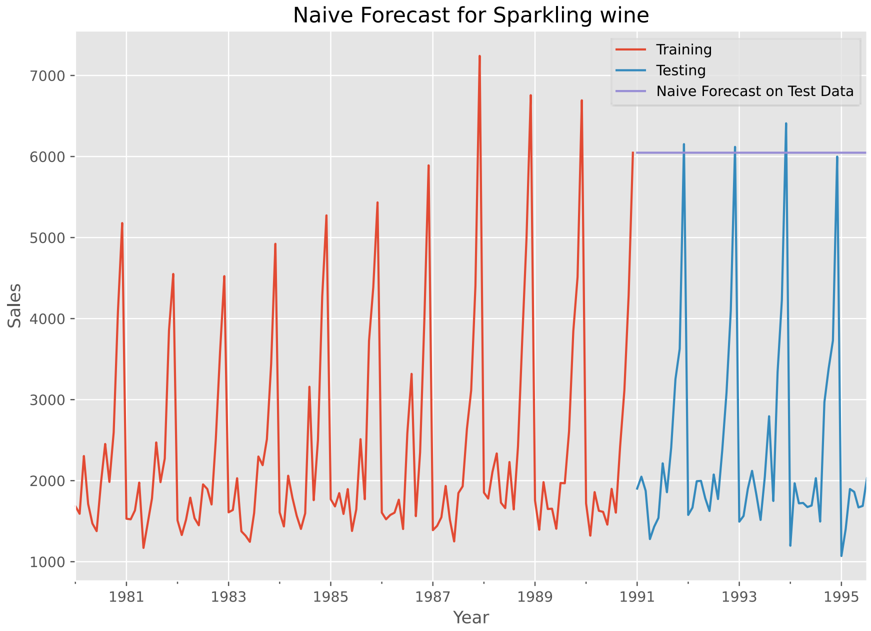


Figure 8: Naive Forecast for Sparkling wine

The above figure shows that a naïve forecast produces a straight line and does not capture either the trend or seasonality in the data as it is a very basic model. The model can be evaluated using RMSE metric and compared with other models. The below output shows the RMSE value for Naïve forecast model:

```
RMSE for Naive forecast model on Sparkling wine data: 3864.279
```

The RMSE value is a lot higher than the standard deviation of the data. Therefore, the model does not perform well.

Linear Regression

For a linear regression model, the data must contain at least one predictor variable. The Sparkling data only contains a target variable which is the sales of the Sparkling wine. Therefore, we will have to create a new predictor variable to build the model. Hence, we will regress the Sparkling column with the order of occurrence of the values.

The below tables show the training and test set after adding the new predictor variables:

	Sparkling	Time
YearMonth		
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

Table 4: Training set for Linear Regression

	Sparkling	Time
YearMonth		
1991-01-01	1902	133
1991-02-01	2049	134
1991-03-01	1874	135
1991-04-01	1279	136
1991-05-01	1432	137

Table 5: Testing set for Linear Regression

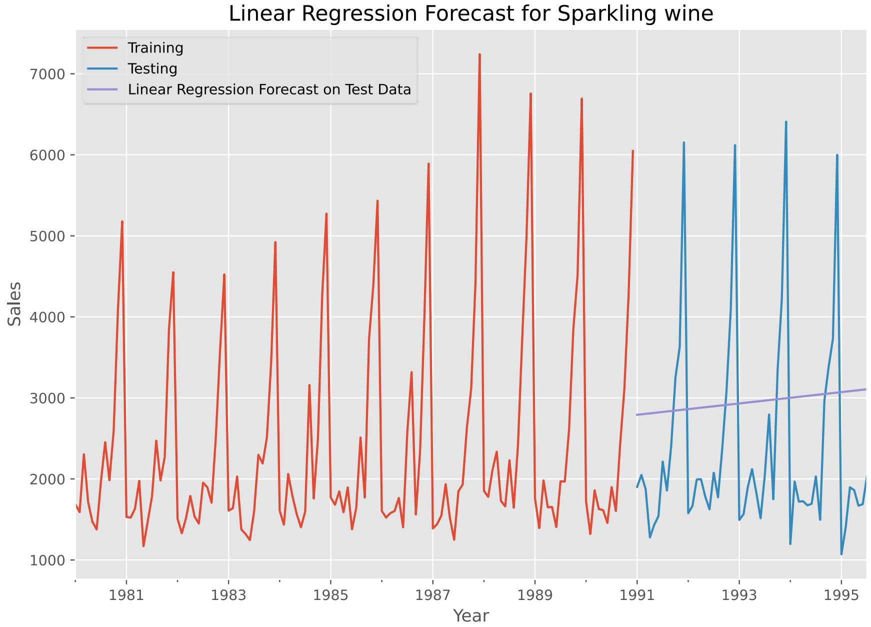


Figure 9: Linear Regression Forecast for Sparkling wine

The linear regression model fits an inclined line through the data such that the RMSE is minimized. From the above figure we can see that the model only captures the trend of the time series data. The below output shows the RMSE value for Linear Regression model:

```
RMSE for Linear Regression forecast model on Sparkling wine data: 1389.135
```

The RMSE value for the Linear Regression model is a lot better than that of the Naïve forecast model but still comparable to the standard deviation of the data. Therefore, this model performs a lot better that Naïve forecast model.

Simple Average Model

In a Simple Average forecasting technique, the mean of the training data is used as the forecast for all the periods in the test data. The below table shows the forecast for the Simple Average model:

	Sparkling	Average Forecast
YearMonth		
1991-01-01	1902	2403.780303
1991-02-01	2049	2403.780303
1991-03-01	1874	2403.780303
1991-04-01	1279	2403.780303
1991-05-01	1432	2403.780303

Table 6: Simple Average Forecast on the test set

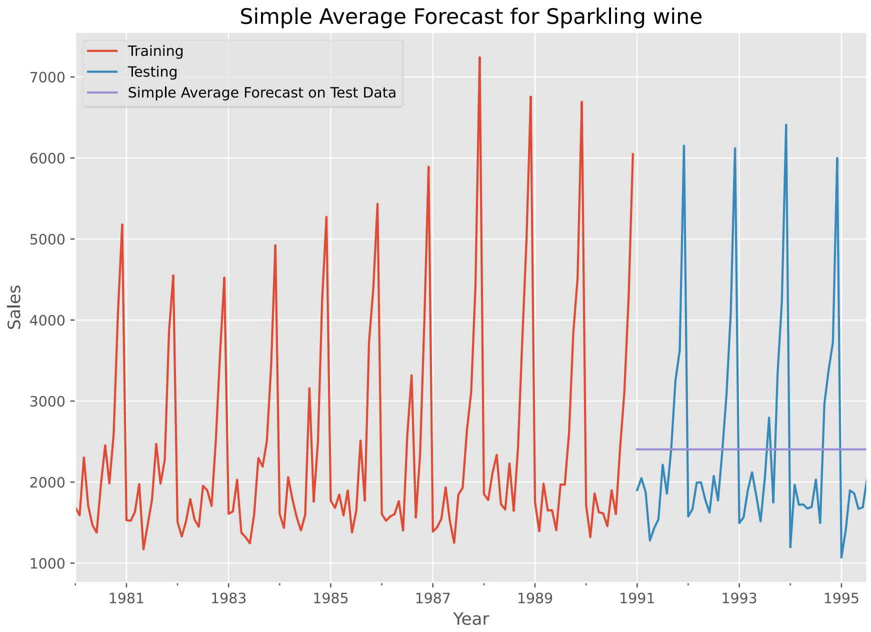


Figure 10: Simple Average Forecast for Sparkling wine

From the above figure we can see that, the Simple Average forecast is a straight line which passes through the mean of the training data. The model does not capture either the trend or seasonality in the data. The RMSE value for the model is shown below:

```
RMSE for Simple Average forecast model on Sparkling wine data: 1275.082
```

The RMSE value for the Simple Average model is slightly lower than that of the Linear Regression model.

Moving Average (MA) Models

In a Moving Average forecasting technique, the forecast for a given period is the average of predetermined number of previous periods k. The moving averages smoothens the seasonal fluctuations in the time series data. The forecasts for first k-1 periods are null values as k number of data points are required to find the averages. Therefore, we will find various moving averages for the complete data and then split the data into training and test sets so that the test set does not contain any null values. We will calculate 2-, 4-, 6-, and 8-point moving averages for the entire data.

The below two tables show the various moving average forecast for the entire data and the test data. For the entire data we can confirm that first few rows contain null values.

	Sparkling	2-MA	4-MA	6-MA	8-MA
YearMonth					
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Table 7: Moving Average Forecast for entire data

	Sparkling	2-MA	4-MA	6-MA	8-MA
YearMonth					
1991-01-01	1902	3974.5	3837.75	3230.000000	2842.000
1991-02-01	2049	1975.5	3571.00	3304.000000	2916.000
1991-03-01	1874	1961.5	2968.00	3212.333333	2912.875
1991-04-01	1279	1576.5	1776.00	2906.166667	2872.125
1991-05-01	1432	1355.5	1658.50	2430.500000	2748.125

Table 8: Moving Average Forecast for test data

The smoothening effect produced by the moving averages depend on the window of rolling mean. Larger the window more is the smoothening effect. This can be confirmed from the below figure.

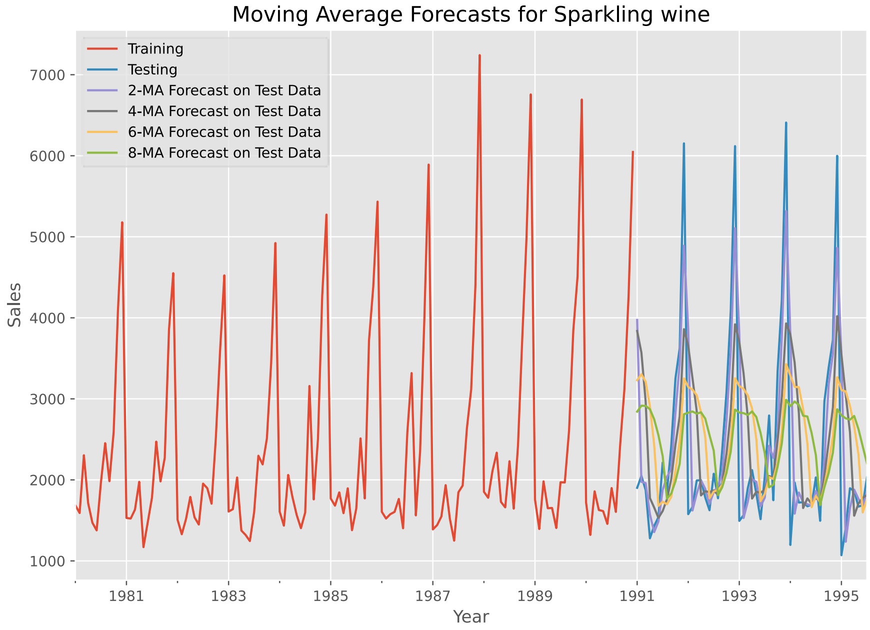


Figure 11: Moving Average Forecasts for Sparkling wine

The 2-point Moving Average is the closest to the test data and hence, it is expected to have the lowest RMSE value. The RMSE values for all the moving average models are:

```
RMSE for 2-MA forecast model on Sparkling wine data: 813.401
RMSE for 4-MA forecast model on Sparkling wine data: 1156.590
RMSE for 6-MA forecast model on Sparkling wine data: 1283.927
RMSE for 8-MA forecast model on Sparkling wine data: 1342.568
```

The 2-point Moving Average model has the lowest RMSE among the all the MA models. It also has the lowest RMSE among all the previous built models like Simple Average, etc. Therefore, we will consider 2-MA model for further model comparisons.

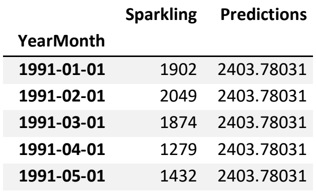
Simple Exponential Smoothing (SES) Model

Simple Exponential Smoothing (SES), is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha, also called the smoothing factor or smoothing coefficient. Parameter alpha ranges between 0 and 1. Large values of alpha means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

The model was built using least squares method of optimization to ensure convergence and the following parameters were found to be optimum:

```
{'smoothing_level': 0.000,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 2403.780,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Here, the **smoothing_trend** and **smoothing_seasonal** parameters are NaN because we are using Simple Exponential Smoothing model. The **smoothing_level** parameter is the alpha value. The model found the optimal alpha value to be zero i.e., the model used all the previous period data to make predictions into the future. The below table shows the predictions done by the model:



	Sparkling	Predictions
YearMonth		
1991-01-01	1902	2403.78031
1991-02-01	2049	2403.78031
1991-03-01	1874	2403.78031
1991-04-01	1279	2403.78031
1991-05-01	1432	2403.78031

Table 9: SES Forecast for test data

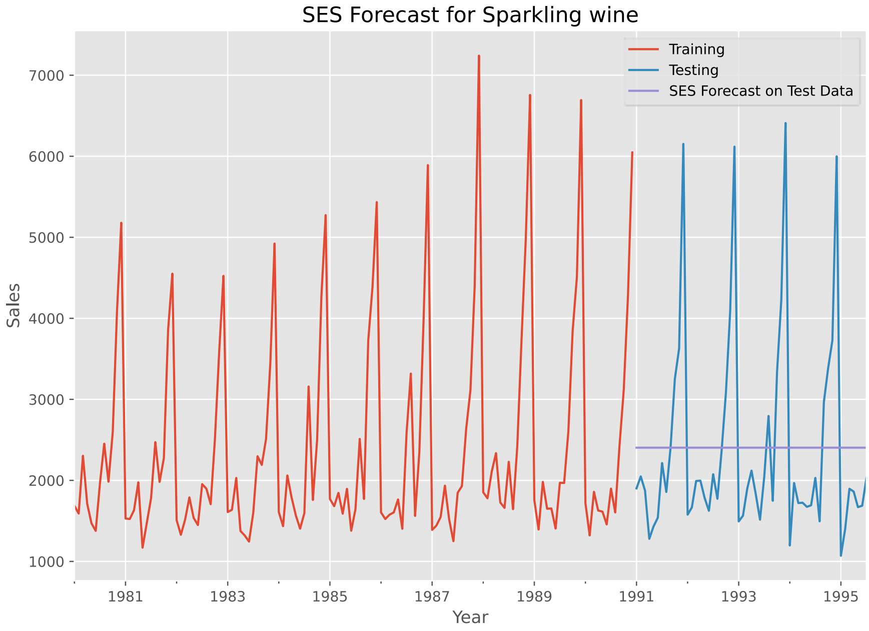


Figure 12: SES Forecast for Sparkling wine

From the above figure we see that the forecast done by the Simple Exponential Smoothing model is a straight line. This is due the fact that SES model is only used for data which does not have any trend or seasonality. From the prediction value we can infer that the model works exactly similar to the Simple Average model. The RMSE value for the Simple Exponential Smoothing model is:

```
RMSE for SES forecast model on Sparkling wine data: 1275.082
```

As expected, the RMSE value of the SES model is same as that of Simple Average model.

Holt (Double Exponential Smoothing) Model

Double exponential smoothing (Holt’s model) employs a level component and a trend component at each period. Holt’s model is used for data that has a trend but no seasonality component. Double exponential smoothing uses two smoothing parameters, alpha and beta, to update the components at each period. Alpha is the level smoothing parameter while beta is the trend smoothing parameter. Both parameter range between 0 and 1. Large values means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

The model was built using least squares method of optimization to ensure convergence and the following parameters were found to be optimum:

```
{'smoothing_level': 0.648,
 'smoothing_trend': 0.000,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1670.400,
 'initial_trend': 27.155,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Here, the **smoothing_seasonal** parameter is NaN because we are using Simple Exponential Smoothing model. The **smoothing_level** parameter is the alpha value while the **smoothing_trend** parameter is the beta value. The model found the optimal alpha value to be 0.648 and beta value as zero. The below table shows the predictions done by the model:

	Sparkling	Predictions
YearMonth		
1991-01-01	1902	5281.993752
1991-02-01	2049	5309.148538
1991-03-01	1874	5336.303323
1991-04-01	1279	5363.458108
1991-05-01	1432	5390.612894

Table 10: DES Forecast for test data

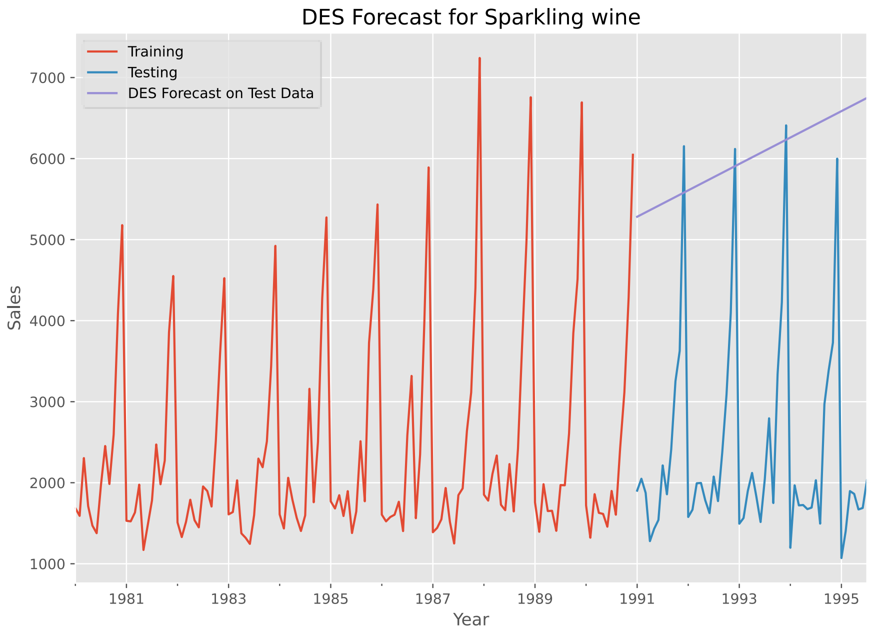


Figure 13: DES Forecast for Sparkling wine

From the above figure we see that the forecast done by the Double Exponential Smoothing model is an inclined line. This is due the fact that DES model is only used for data which has trend but no seasonality. The RMSE value for the Double Exponential Smoothing model is:

```
RMSE for DES forecast model on Sparkling wine data: 3854.073
```

The RMSE for the Holt's model is almost the same as that of the Naïve forecast model.

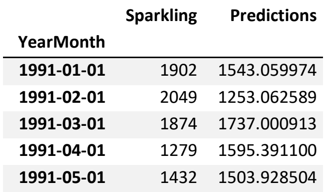
Holt-Winters (Triple Exponential Smoothing) Model

Triple exponential smoothing (Holt-Winters model) employs level, trend and seasonal components at each period. Holt-Winters model is used for data that has both trend and seasonality components. Triple exponential smoothing uses three smoothing parameters, alpha, beta and gamma, to update the components at each period. Alpha is the level smoothing parameter, beta is the trend smoothing parameter and gamma is the seasonal smoothing parameter. All of these parameters range between 0 and 1. Large values means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

Following parameters were found to be optimum:

```
{'smoothing_level': 0.082,
 'smoothing_trend': 0.000,
 'smoothing_seasonal': 0.474,
 'damping_trend': nan,
 'initial_level': 3325.132,
 'initial_trend': 0.288,
 'initial_seasons': array([-1663.164, -1738.205, -1271.782, -1485.487, -1845.598, -1850.294,
        -1368.608,  -843.923, -1293.058,  -727.119,   694.499,  1710.512]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

The **smoothing_level** parameter is the alpha value, the **smoothing_trend** parameter is the beta value and the **smoothing_seasonal** parameter is the gamma value. The model found the optimal alpha value to be 0.082, beta value as zero and gamma value to be 0.474. The below table shows the predictions done by the model:



	Sparkling	Predictions
YearMonth		
1991-01-01	1902	1543.059974
1991-02-01	2049	1253.062589
1991-03-01	1874	1737.000913
1991-04-01	1279	1595.391100
1991-05-01	1432	1503.928504

Table 11: TES Forecast for test data

From the below figure we see that the forecast done by the Triple Exponential Smoothing model closely follows the actual test data. This is due the fact that TES model considers both trend and seasonality.

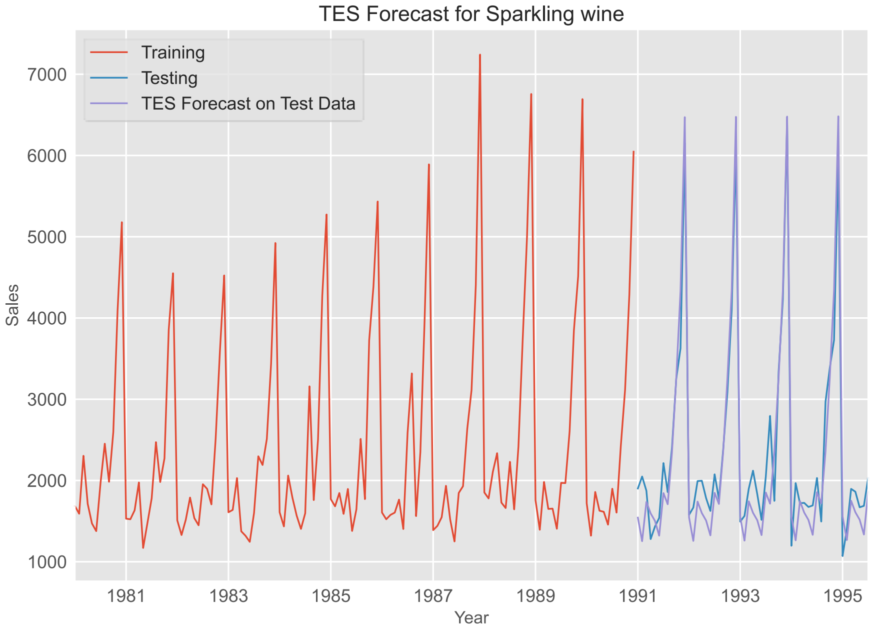


Figure 14: TES Forecast for Sparkling wine

The RMSE value for the Triple Exponential Smoothing model is:

```
RMSE for TES forecast model on Sparkling wine data: 357.725
```

The RMSE for the Holt-Winters model is the lowest so far.

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.

Stationarity means that the statistical properties of a time series do not change over time. It does not mean that the series does not change over time, just that the **way** it changes does not itself change over time. Forecasting models like ARIMA or SARIMA requires that the time series data is stationary.

Checking for Stationarity on Whole data

The **Augmented Dickey-Fuller test is a unit root test** which determines whether there is a unit root and subsequently **whether the series is non-stationary**. It is a hypothesis test with the null and alternate hypothesis as follows:

H_0 *: The Time Series has a unit root and is thus non-stationary.*
H_1 *: The Time Series does not have a unit root and is thus stationary.*

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value. The below output shows the result of the ADF test performed on the Sparkling data:

```
Results of Dickey-Fuller Test:
Test Statistic          -1.798262
p-value                  0.705596
#Lags Used              12.000000
Critical Value (5%)     -3.436029
```

The **p-value from the ADF test is greater than 0.05** and therefore **we fail to reject the Null hypothesis**. Hence, the **Sparkling data is non-stationary**.

Differencing and Checking for Stationarity of Whole data

Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality thereby making the series stationary. As our data is a monthly data, one seasonal period is of 12 months. Therefore, to make the data stationary, we will take a seasonal difference of 12 (D=1) and an additional level difference of 1 (d=1). The below output shows the result of the ADF test performed on differenced Sparkling data:

```
Results of Dickey-Fuller Test:
Test Statistic          -5.121008
p-value                  0.000123
#Lags Used              11.000000
Critical Value (5%)     -3.437946
```

The **p-value from the ADF test is less than 0.05** and therefore **we reject the Null hypothesis**. Therefore, taking the **difference of the data has made the data stationary**.

Checking for Stationarity on Training data

The below output shows the result of the ADF test performed on the training data:

```
Results of Dickey-Fuller Test:
Test Statistic          -2.061798
p-value                  0.567411
#Lags Used              12.000000
Critical Value (5%)     -3.448049
```

The **p-value from the ADF test is greater than 0.05** and therefore **we fail to reject the Null hypothesis**. Hence, the **training data is non-stationary**.

Differencing and Checking for Stationarity of Training data

The same differencing as above (d=1 and D=1) is applied to the training data. Below figure shows the differenced training data:

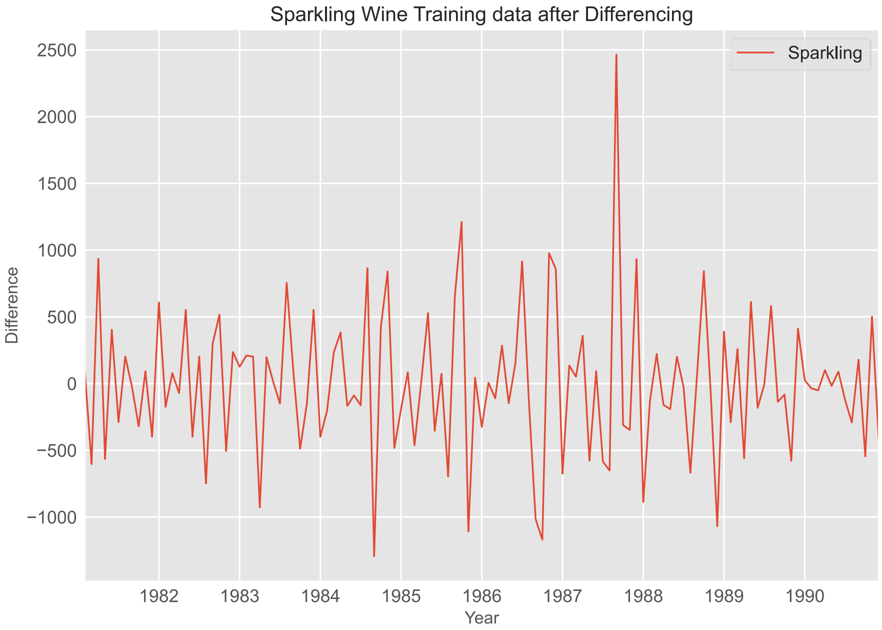


Figure 15: Sparkling Wine Training data after Differencing

From the above figure we see that, the trend and seasonality components in the training data have been reduced. This can be confirmed by performing the ADF test on the differenced training data:

```
Results of Dickey-Fuller Test:
Test Statistic          -3.467900
p-value                  0.042955
#Lags Used              10.000000
Critical Value (5%)     -3.451953
```

The **p-value from the ADF test is less than 0.05** and therefore **we reject the Null hypothesis**. Therefore, taking the **difference of the training data has made the data stationary**.

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.

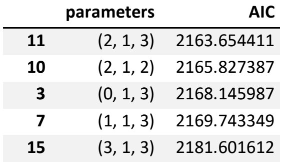
Automated ARIMA(p, d, q) Model

ARIMA stands for Auto Regressive Integrated Moving Average model. There are **total three parameters related to three different components** of the model. The **p parameter indicates the lag used in the Auto Regressive component**. The **d parameter is the order of level differencing applied to make the data stationary**. The **q parameter is the lag used in the Moving Average component**. Here, **we are using a value of d = 1**. Various combinations of the p, d and q parameter are used to build various ARIMA models and the **combination that give the lowest AIC value is used to evaluate the model on the test data**.

```
Some examples of the parameter combinations for the Model
Model: (0, 1, 0)
```

```
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
```

The below table shows the AIC values of different parameter combinations in ascending order:



	parameters	AIC
11	(2, 1, 3)	2163.654411
10	(2, 1, 2)	2165.827387
3	(0, 1, 3)	2168.145987
7	(1, 1, 3)	2169.743349
15	(3, 1, 3)	2181.601612

Table 12: AIC values for ARIMA model

The parameter combination of p=2, d=1, and q=3 gives the lowest AIC value. We will use this model for making predictions on the test data. The below output shows the model summary of ARIMA(2, 1, 3) model:

```
                               SARIMAX Results                                
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood               -1075.827
Date:                Thu, 22 Sep 2022   AIC                           2163.654
Time:                        12:06:03   BIC                           2180.720
Sample:                    01-01-1980   HQIC                          2170.588
                         - 12-01-1990                                         
Covariance Type:                  opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8475      0.120     -7.083      0.000      -1.082      -0.613
ar.L2         -0.4809      0.110     -4.372      0.000      -0.696      -0.265
ma.L1          0.3505      0.062      5.640      0.000       0.229       0.472
ma.L2         -0.3754      0.071     -5.253      0.000      -0.515      -0.235
ma.L3         -0.9169      0.047    -19.510      0.000      -1.009      -0.825
sigma2      1.253e+06   1.59e+05      7.897      0.000    9.42e+05    1.56e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.64   Jarque-Bera (JB):                 9.77
Prob(Q):                              0.42   Prob(JB):                         0.01
Heteroskedasticity (H):               2.54   Skew:                             0.67
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.25
===================================================================================

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

We see that all the coefficients for the AR and MA model have p-values less than 0.05. Therefore, all the coefficients are significant. The below figure shows the diagnostic plots for standardized residuals of the Sparkling variable:

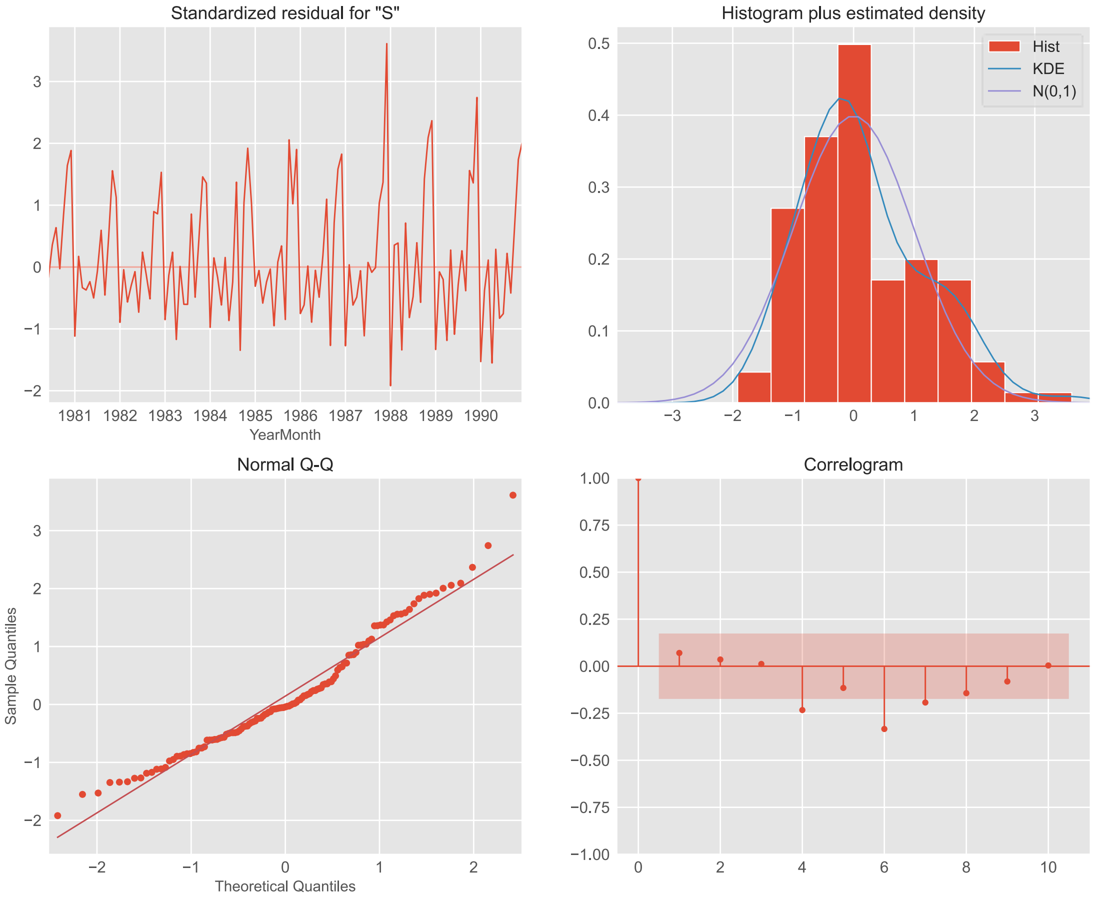


Figure 16: Diagnostic plots for Auto ARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Auto ARIMA forecast model on Sparkling wine data: 1294.096
```

The RMSE of the Auto ARIMA model is almost similar to the Simple Average and SES models built above. This is expected as the ARIMA model also do not take into consideration the seasonal fluctuations in the data. From the below figure we see that the predictions fluctuate in the begin and then flattens out at the end.

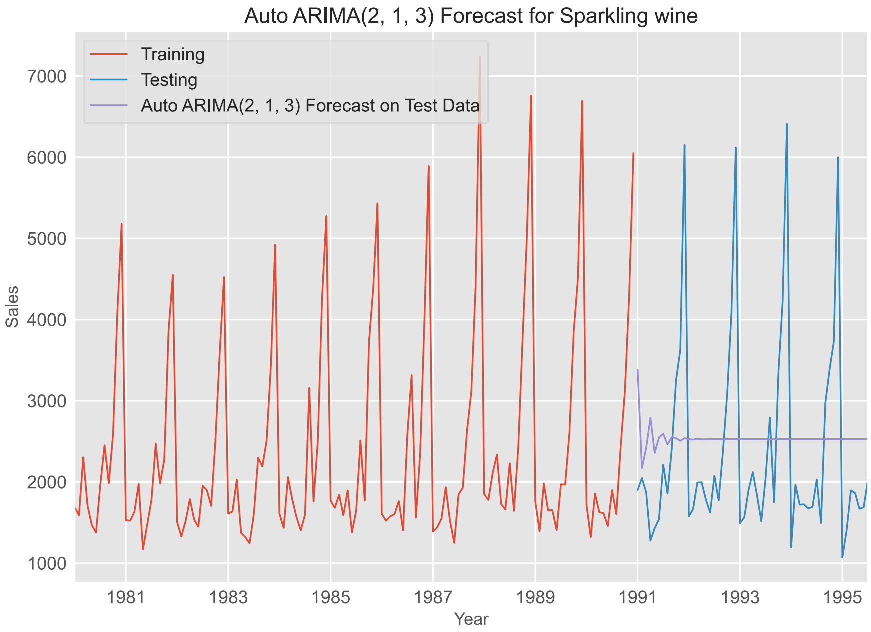


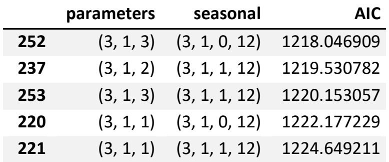
Figure 17: Auto ARIMA(2, 1, 3) Forecast for Sparkling wine

Automated SARIMA(p, d, q)(P, D, Q, F) Model

SARIMA stands for Seasonal Auto Regressive Integrated Moving Average model. It is an extension of the ARIMA model. There are total 7 parameters used to define the model. The parameters (p, d, q) are the same as the ARIMA model. The parameters (P, D, Q) are the seasonal counterparts of (p, d, q). The parameter F is the seasonality of the data which is 12 in our case. Here, **we are using a value of d = 1 and D = 1**. Various combinations of the p, d, q, P, D, and Q parameter are used to build various SARIMA models and the **combination that give the lowest AIC value is used to evaluate the model on the test data**.

```
Some examples of the parameter combinations for the Model
Model: (0, 1, 0)(0, 1, 0, 12)
Model: (0, 1, 1)(0, 1, 1, 12)
Model: (0, 1, 2)(0, 1, 2, 12)
Model: (0, 1, 3)(0, 1, 3, 12)
Model: (1, 1, 0)(1, 1, 0, 12)
```

The below table shows the AIC values of different parameter combinations in ascending order:



	parameters	seasonal	AIC
252	(3, 1, 3)	(3, 1, 0, 12)	1218.046909
237	(3, 1, 2)	(3, 1, 1, 12)	1219.530782
253	(3, 1, 3)	(3, 1, 1, 12)	1220.153057
220	(3, 1, 1)	(3, 1, 0, 12)	1222.177229
221	(3, 1, 1)	(3, 1, 1, 12)	1224.649211

Table 13: AIC values for SARIMA model

The parameter combination of p=3, d=1, q=3, P=3, D=1, Q=0, and F=12 gives the lowest AIC value. We will use this model for making predictions on the test data. The below output shows the model summary of SARIMA(3, 1, 3)(3, 1, 0, 12) model:

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                 Sparkling   No. Observations:          132
Model:             SARIMAX(3, 1, 3)x(3, 1, [], 12)   Log Likelihood            -596.483
Date:                           Thu, 22 Sep 2022   AIC                       1212.966
Time:                                   12:06:29   BIC                       1236.786
Sample:                               01-01-1980   HQIC                      1222.516
                                    - 12-01-1990                                      
Covariance Type:                             opg                                      
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6134      0.140    -11.524      0.000      -1.888      -1.339
ar.L2         -0.6101      0.269     -2.265      0.023      -1.138      -0.082
ar.L3          0.0877      0.147      0.597      0.550      -0.200       0.376
ma.L1          0.9952      0.230      4.329      0.000       0.545       1.446
ma.L2         -0.8770      0.158     -5.554      0.000      -1.186      -0.568
ma.L3         -0.9600      0.182     -5.276      0.000      -1.317      -0.603
ar.S.L12      -0.4521      0.130     -3.489      0.000      -0.706      -0.198
ar.S.L24      -0.2356      0.132     -1.780      0.075      -0.495       0.024
ar.S.L36      -0.1018      0.111     -0.918      0.358      -0.319       0.115
sigma2      1.669e+05   2.28e-06   7.33e+10      0.000    1.67e+05    1.67e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 3.98
Prob(Q):                              0.93   Prob(JB):                         0.14
Heteroskedasticity (H):               0.73   Skew:                             0.48
Prob(H) (two-sided):                  0.43   Kurtosis:                         3.53
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 4.39e+26. Standard
errors may be unstable.
```

We see that the coefficients ar.L3, ar.S.L24 and ar.S.L36 have p-values greater than 0.05. Therefore, these coefficients are not significant. The below figure shows the diagnostic plots for standardized residuals of the Sparkling variable:

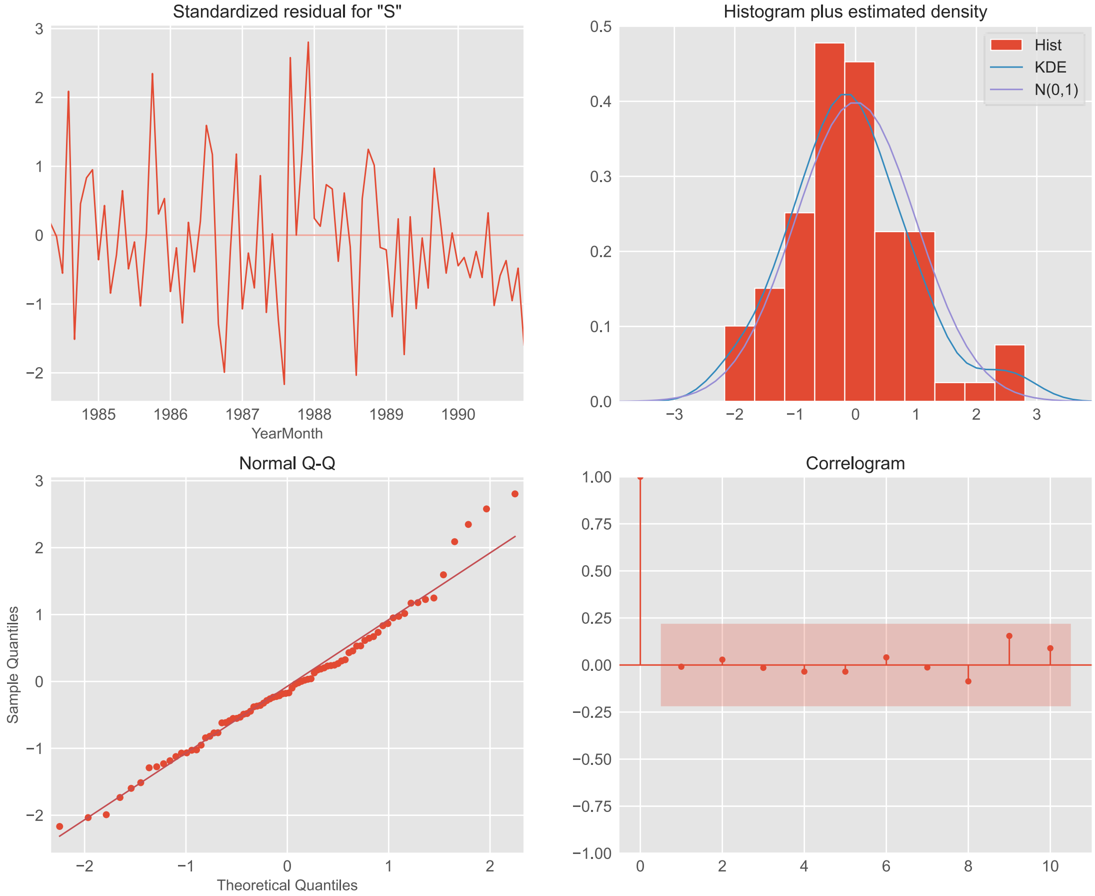


Figure 18: Diagnostic plots for Auto SARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Auto SARIMA forecast model on Sparkling wine data: 329.782
```

The RMSE of the Auto SARIMA model is the lowest out of all the models built above. From the below figure we see that the predictions follow the test data very closely.

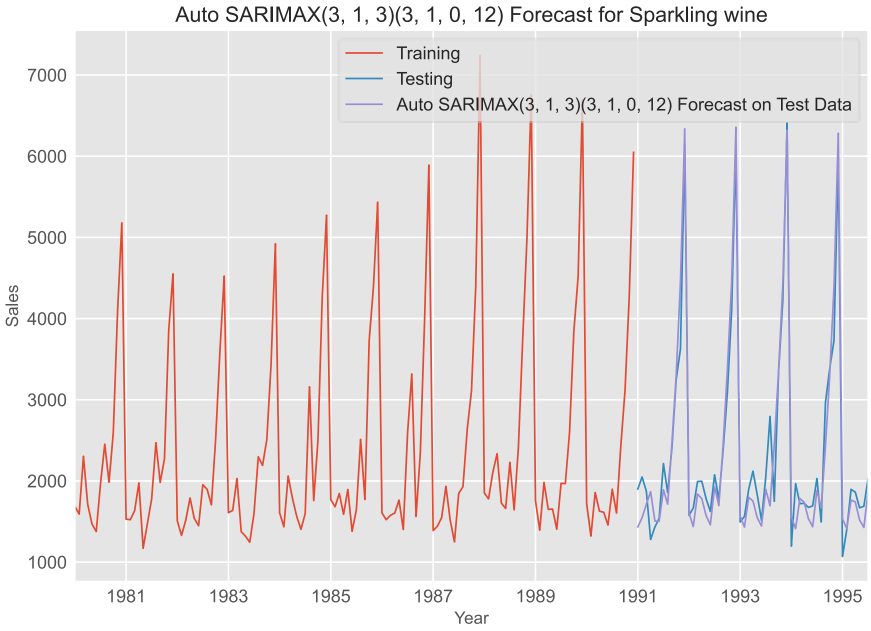


Figure 19: Auto SARIMAX(3, 1, 3)(3, 1, 0, 12) Forecast for Sparkling wine

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.

PACF Plot

PACF or **Partial Autocorrelation Function is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed**. The Auto-Regressive parameter in an ARIMA/SARIMA model is **'p' which comes from the significant lag before which the PACF plot cuts-off to 0**.

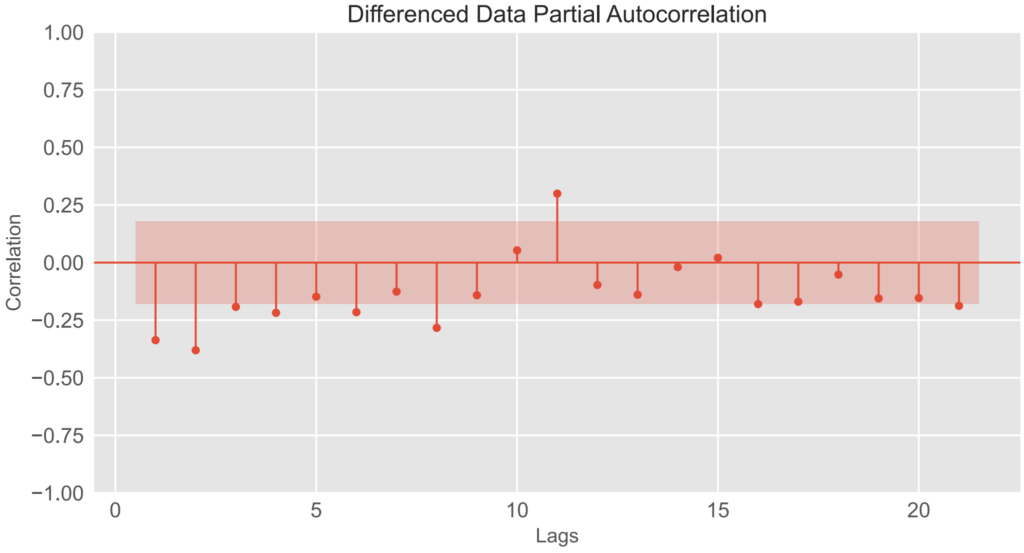


Figure 20: Differenced Data Partial Autocorrelation

From the above PACF plot we see that, the PACF cuts-off after 2nd lag. Therefore, **we will use p = 2** for the models. Also, the PACF for 12th lag is inside the 0.05 interval. Therefore, **we will use P = 0** for the models.

ACF Plot

ACF or **Autocorrelation Function is a way to measure the linear relationship between an observation at time *t* and the observations at previous times**. The Moving-Average parameter in an ARIMA model is **'q' which comes from the significant lag before the ACF plot cuts-off to 0**.

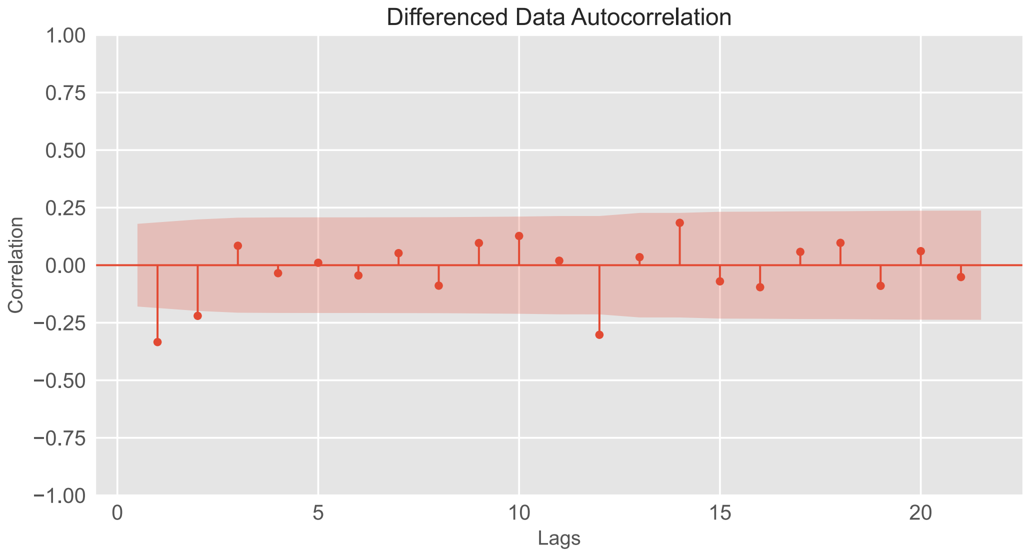


Figure 21: Differenced Data Autocorrelation

From the above ACF plot we see that, the ACF cuts-off after 2nd lag. Therefore, **we will use q = 2** for the models. Also, after the 12th lag, the ACF cuts-off. Therefore, **we will use Q = 1** for the models.

Manual ARIMA(p, d, q) Model

The best ARIMA model suggested by the above ACF and PACF plots is ARIMA(2, 1, 2). The below output shows the model summary of ARIMA(2, 1, 2) model:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:               Sparkling   No. Observations:                  132
Model:                  ARIMA(2, 1, 2)   Log Likelihood               -1077.914
Date:                 Thu, 22 Sep 2022   AIC                           2165.827
Time:                         15:42:22   BIC                           2180.088
Sample:                     01-01-1980   HQIC                          2171.621
                          - 12-01-1990
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3098      0.045     28.859      0.000       1.221       1.399
ar.L2         -0.5559      0.072     -7.674      0.000      -0.698      -0.414
ma.L1         -1.9910      0.110    -18.060      0.000      -2.207      -1.775
ma.L2          0.9993      0.111      9.032      0.000       0.782       1.216
sigma2      1.093e+06   2.05e-07   5.32e+12      0.000    1.09e+06    1.09e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                14.64
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               2.54   Skew:                             0.63
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================

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

```
[2] Covariance matrix is singular or near-singular, with condition number 1.83e+28. Standard
errors may be unstable.
```

We see that all the coefficients for the AR and MA model have p-values less than 0.05. Therefore, all the coefficients are significant. The below figure shows the diagnostic plots for standardized residuals of the Sparkling variable:

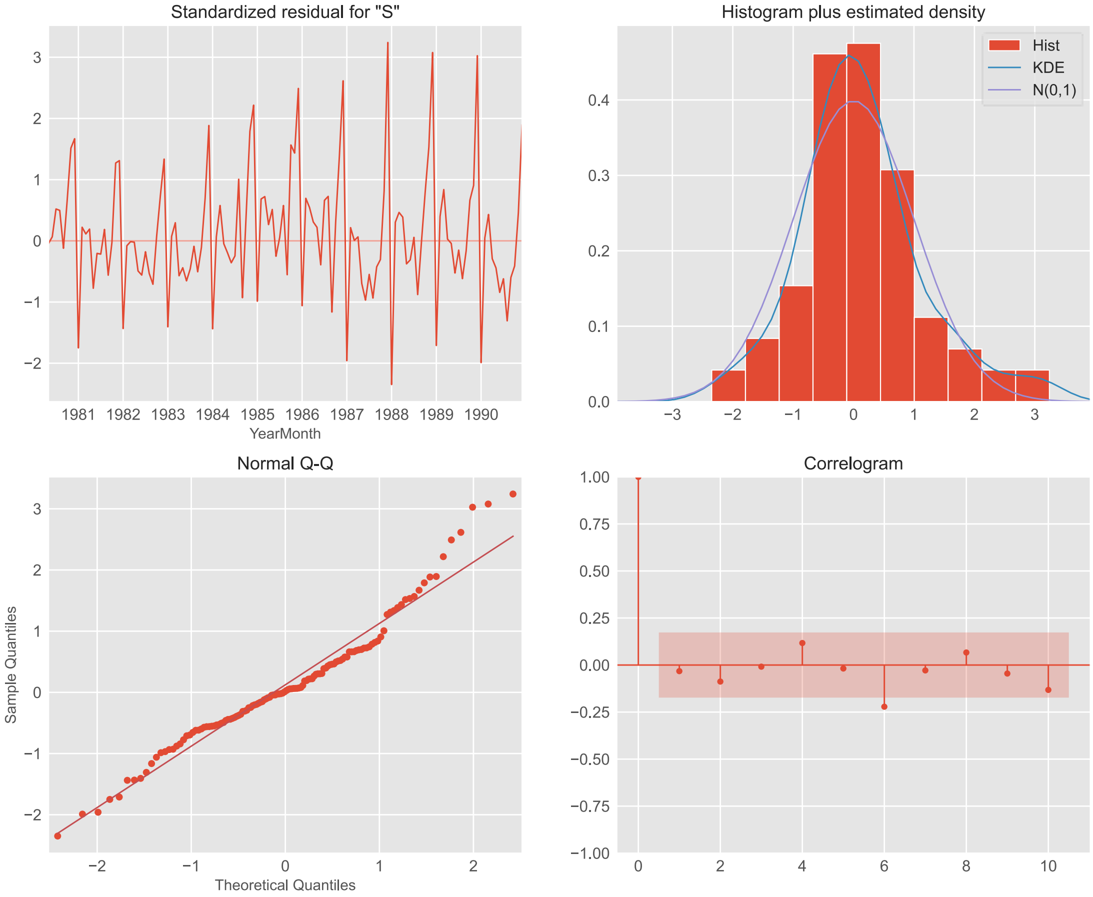


Figure 22: Diagnostic plots for Manual ARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Manual ARIMA forecast model on Sparkling wine data: 1300.135
```

The RMSE of the Manual ARIMA model is slightly greater than that of the Auto ARIMA model. From the below figure we see that the predictions fluctuate in the begin and then flattens out at the end.

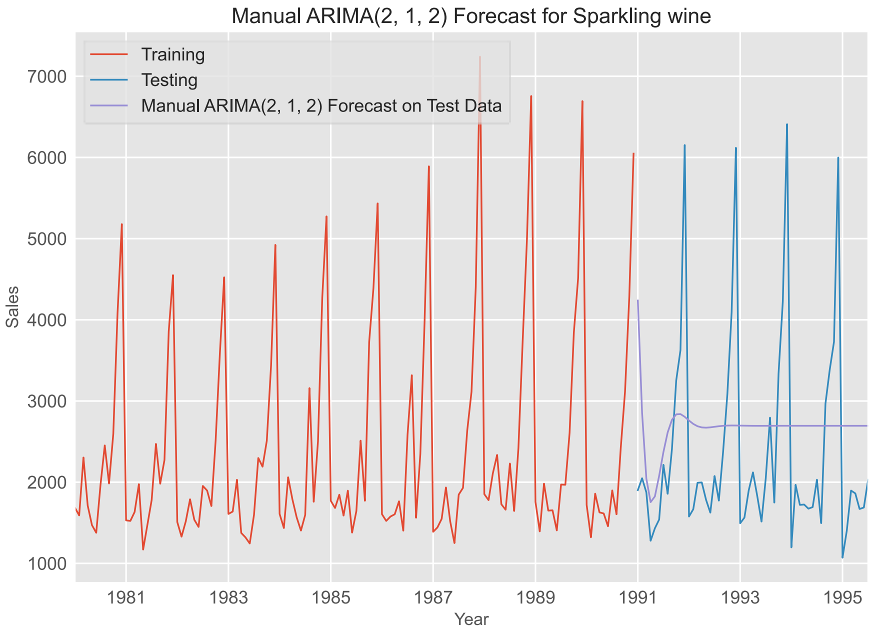


Figure 23: Manual ARIMA(2, 1, 2) Forecast for Sparkling wine

Manual SARIMA(p, d, q)(P, D, Q, F) Model

The best SARIMA model suggested by the above ACF and PACF plots is SARIMA(2, 1, 2)(0, 1, 1, 12). The below output shows the model summary of SARIMA(2, 1, 2)(0, 1, 1, 12) model:

```
                                      SARIMAX Results                                       
============================================================================================
Dep. Variable:                            Sparkling   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(0, 1, [1], 12)   Log Likelihood                -772.473
Date:                              Thu, 22 Sep 2022   AIC                           1556.947
Time:                                      15:19:27   BIC                           1572.813
Sample:                                  01-01-1980   HQIC                          1563.375
                                       - 12-01-1990                                         
Covariance Type:                                opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5504      0.295     -1.868      0.062      -1.128       0.027
ar.L2         -0.0136      0.153     -0.089      0.929      -0.314       0.287
ma.L1         -0.1657      0.279     -0.594      0.553      -0.713       0.381
ma.L2         -0.6739      0.259     -2.606      0.009      -1.181      -0.167
ma.S.L12      -0.4581      0.082     -5.615      0.000      -0.618      -0.298
sigma2       1.65e+05    2.2e+04      7.504      0.000    1.22e+05    2.08e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                15.54
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               1.03   Skew:                             0.56
Prob(H) (two-sided):                  0.94   Kurtosis:                         4.52
===================================================================================

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

We see that the coefficients ar.L1, ar.L2 and ma.L1 have p-values greater than 0.05. Therefore, these coefficients are not significant. The below figure shows the diagnostic plots for standardized residuals of the Sparkling variable:

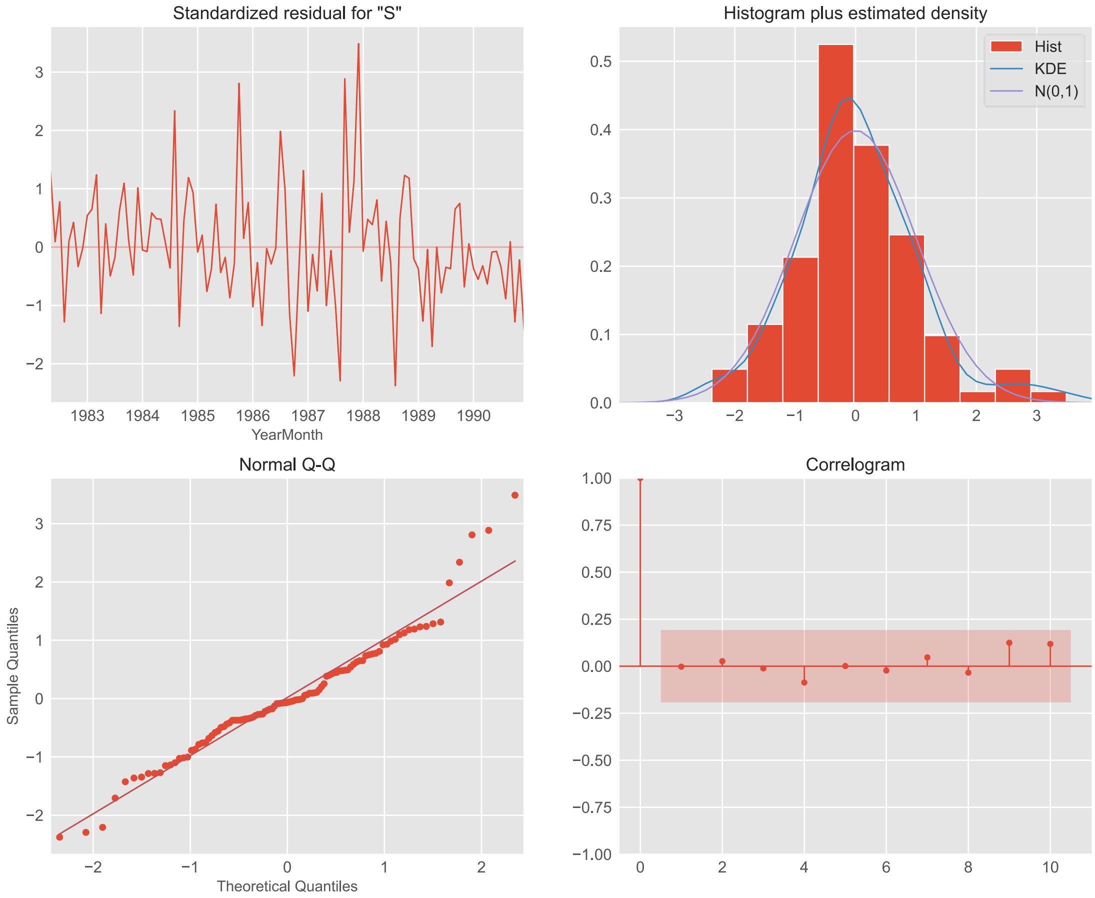


Figure 24: Diagnostic plots for Manual SARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Manual SARIMA forecast model on Sparkling wine data: 452.981
```

The RMSE of the Manual SARIMA model is greater than that of the Auto SARIMA model built above. From the below figure we see that the predictions follow the test data closely.

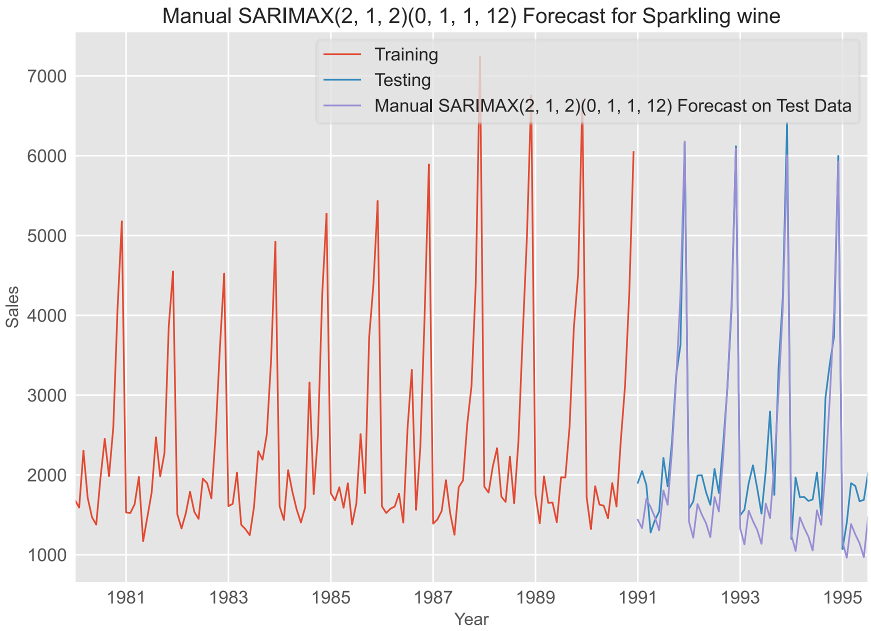
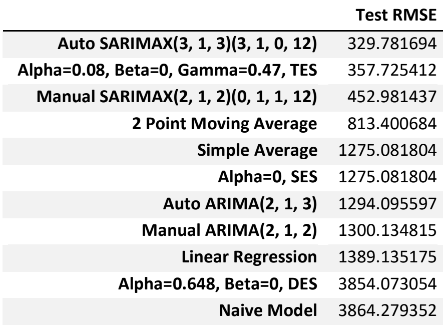


Figure 25: Manual SARIMAX(2, 1, 2)(0, 1, 1, 12) Forecast for Sparkling wine

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.

The below table shows the summary of RMSE values of the all the models built above in ascending order:



	Test RMSE
Auto SARIMAX(3, 1, 3)(3, 1, 0, 12)	329.781694
Alpha=0.08, Beta=0, Gamma=0.47, TES	357.725412
Manual SARIMAX(2, 1, 2)(0, 1, 1, 12)	452.981437
2 Point Moving Average	813.400684
Simple Average	1275.081804
Alpha=0, SES	1275.081804
Auto ARIMA(2, 1, 3)	1294.095597
Manual ARIMA(2, 1, 2)	1300.134815
Linear Regression	1389.135175
Alpha=0.648, Beta=0, DES	3854.073054
Naive Model	3864.279352

Table 14: RMSE values of all the models

Therefore, the most optimum model for the Sparkling wine dataset is the SARIMA(3, 1, 3)(3, 1, 0, 12) model as it has the lowest RMSE value. Hence, we will use this model to train on the complete data and predict 12 months into the future.

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.

We will use the most optimum model i.e., SARIMA(3, 1, 3)(3, 1, 0, 12) to train on the complete dataset. The below output shows the summary of the final model:

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  187
Model:             SARIMAX(3, 1, 3)x(3, 1, [], 12)   Log Likelihood                -998.042
Date:                             Thu, 22 Sep 2022   AIC                           2016.083
Time:                                     19:21:06   BIC                           2045.136
```

```
Sample:                               01-01-1980   HQIC                           2027.890
                                    - 07-01-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.0051      0.104     -9.705      0.000      -1.208      -0.802
ar.L2         -0.8249      0.102     -8.057      0.000      -1.026      -0.624
ar.L3          0.1033      0.088      1.169      0.243      -0.070       0.276
ma.L1          0.2016      0.105      1.924      0.054      -0.004       0.407
ma.L2         -0.1296      0.090     -1.445      0.149      -0.306       0.046
ma.L3         -0.9668      0.090    -10.786      0.000      -1.143      -0.791
ar.S.L12      -0.5570      0.074     -7.522      0.000      -0.702      -0.412
ar.S.L24      -0.2808      0.118     -2.371      0.018      -0.513      -0.049
ar.S.L36      -0.1628      0.088     -1.859      0.063      -0.335       0.009
sigma2      1.488e+05   9.65e-07   1.54e+11      0.000    1.49e+05    1.49e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                42.47
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               0.55   Skew:                             0.73
Prob(H) (two-sided):                  0.05   Kurtosis:                         5.33
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.06e+26. Standard
errors may be unstable.
```

We see that the coefficients ar.L3, ma.L1, ma.L2 and ar.S.L36 have p-values greater than 0.05. Therefore, these coefficients are not significant. The below figure shows the diagnostic plots for standardized residuals of the Sparkling variable:

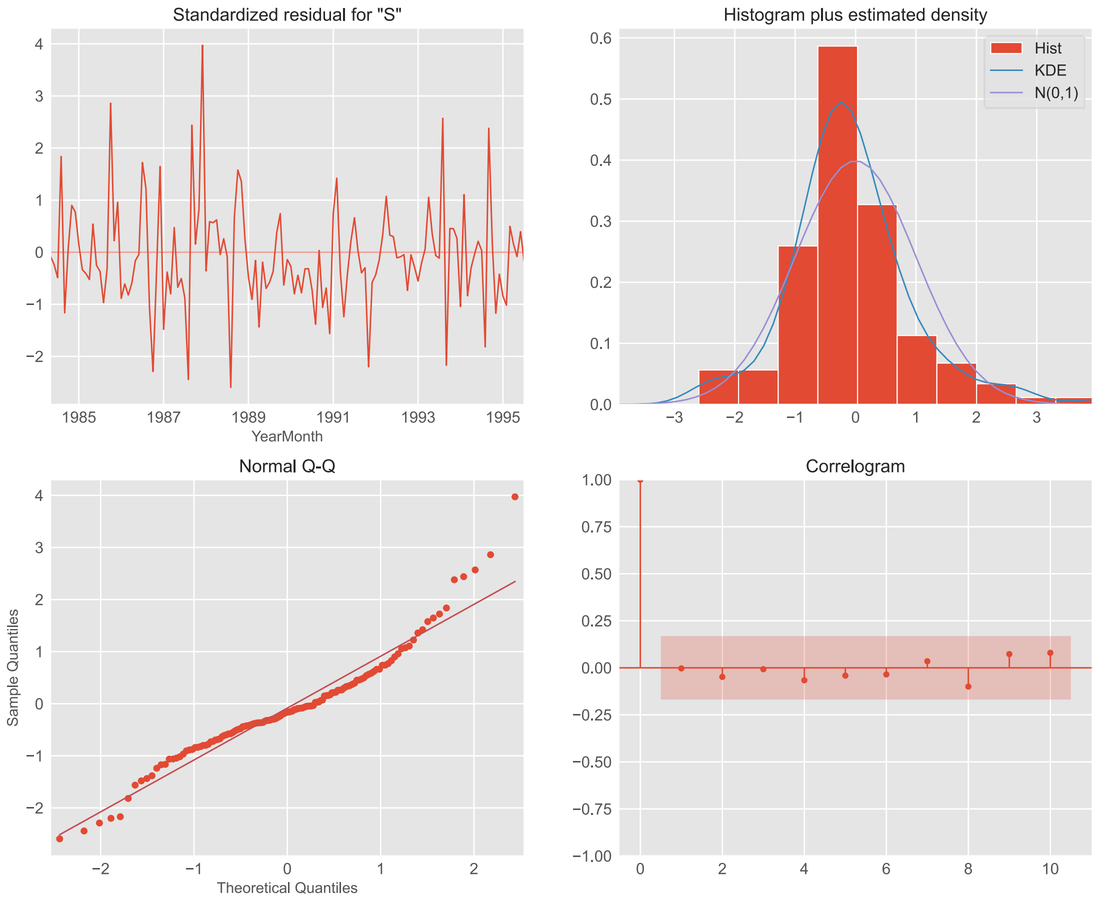
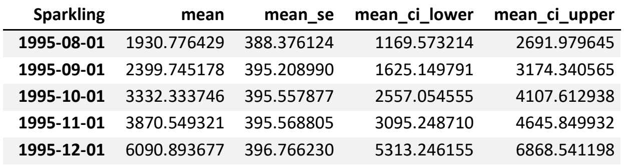


Figure 26: Diagnostic plots for Final SARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern. The above model was then used to predict 12 months into the future. The below output shows the mean, standard error and 95% confidence intervals of the predictions:



Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1930.776429	388.376124	1169.573214	2691.979645
1995-09-01	2399.745178	395.208990	1625.149791	3174.340565
1995-10-01	3332.333746	395.557877	2557.054555	4107.612938
1995-11-01	3870.549321	395.568805	3095.248710	4645.849932
1995-12-01	6090.893677	396.766230	5313.246155	6868.541198

Table 15: Future predictions of 12 months

The below output shows the RMSE value calculated using the actual data and the fitted values:

```
RMSE for the final model on complete data: 612.763
```

The below figure shows the future forecast for 12 months for the Sparkling wine sales with 95% confidence interval band. We see that the model has captured the trend and seasonality of the data properly. The confidence interval band is narrow for the first half of the prediction while it is wider for the second half.

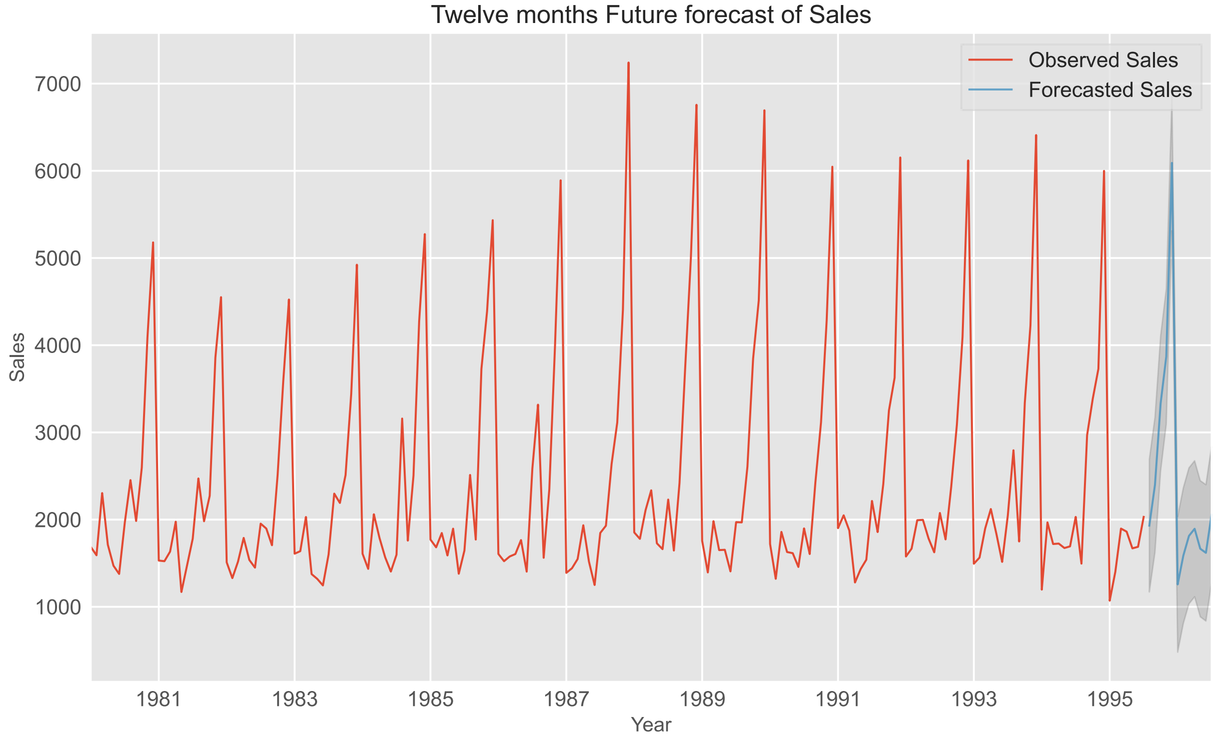


Figure 27: Twelve months Future forecast of Sparkling wine sales

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

We were tasked to analyse and forecast Sparkling Wine Sales in the 20th century for ABC Estate Wines and predict the sales for next 12 months. The time series data was read into a dataframe and proper Datetime Index were provided for the dataframe. The data contained monthly sales of Sparkling wine from 1980 to 1995. While doing the EDA and decomposing the data, we found that wine sales peaked during fourth quarter of every year. The Sparkling wine sales had no clear trend over the years. The sales data from 1991 was used as test data for evaluating the models. Various basic models like Naïve forecast, Linear Regression, Simple and Moving Average, Exponential Smoothing, etc. were built and evaluated using RMSE. The data was found to be non-stationary and proper differencing was applied to make the data stationary. Automated and Manual versions of both ARIMA and SARIMA models were built and evaluated. The SARIMA(3, 1, 3)(3, 1, 0, 12) model was found to be the best model and hence was used to forecast 12 months into the future.

Inferences

1. Sales peak in the last quarter of every year. Especially in December, the sales are almost 3 times the yearly average sales. This could be due to the holiday season.
2. Sales are the lowest during the first quarter of every year. This might be due the after effects of the holidays in the last quarter of previous year.
3. The average sales in the year 1995 was the lowest among all the years. This is due to the fact that only first seven months of data was present in the data.
4. The Sparkling wine sales are expected to reach above 6000 in December 1995.

Recommendations

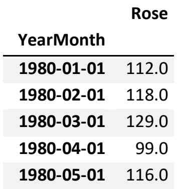
1. The company should prepare for the high demand during the last quarter of the year by increasing production and stocking their inventory well before that period, to allow the wine to age.
2. The company can provide offers during the first half of the year to increase the sale in that period.
3. The general trend in the sales over the years is not increasing. The reason behind this observation should be investigated further. The company can try to improve the quality of the Sparkling wine.

Rose Wine Sales Data

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

Sample of the Dataset

The below table shows the first 5 rows of the Rose dataset. To convert the data into a time series, the YearMonth column was converted into a timestamp index of the dataframe. The Rose column contains the sales of the mentioned wine.



	Rose
YearMonth	
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

Table 16: Rose Wine Data

Checking the types of variables in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

From the above output we can see that:

- There are **187 observations** in the data.
- The time series data has a **Datetime index** from **1980-01-01 to 1995-07-01**.
- The **Rose** column has the monthly sales values. The values are of **float** data type.
- The dataset **has two missing values**.

Imputing Missing Values

As seen from the above output, the **dataset has two missing values**. In a timeseries data, we cannot drop the missing values as it creates a gap in the order of the timestamp. Here the **missing values were imputed by the linear interpolation method**. After interpolation, the missing values in the dataset are zero as shown below:

```
Rose    0
dtype: int64
```

Time Plot of the Time Series

The below plot shows the time plot of the Rose wine sales. There is a clear **decreasing trend** present in the data. **Seasonality can be observed** clearly in the data. The **seasonal fluctuations are decreasing with the sales** across the months, therefore **multiplicative models** might give better results. The trend and seasonality will be further explored during decomposition.

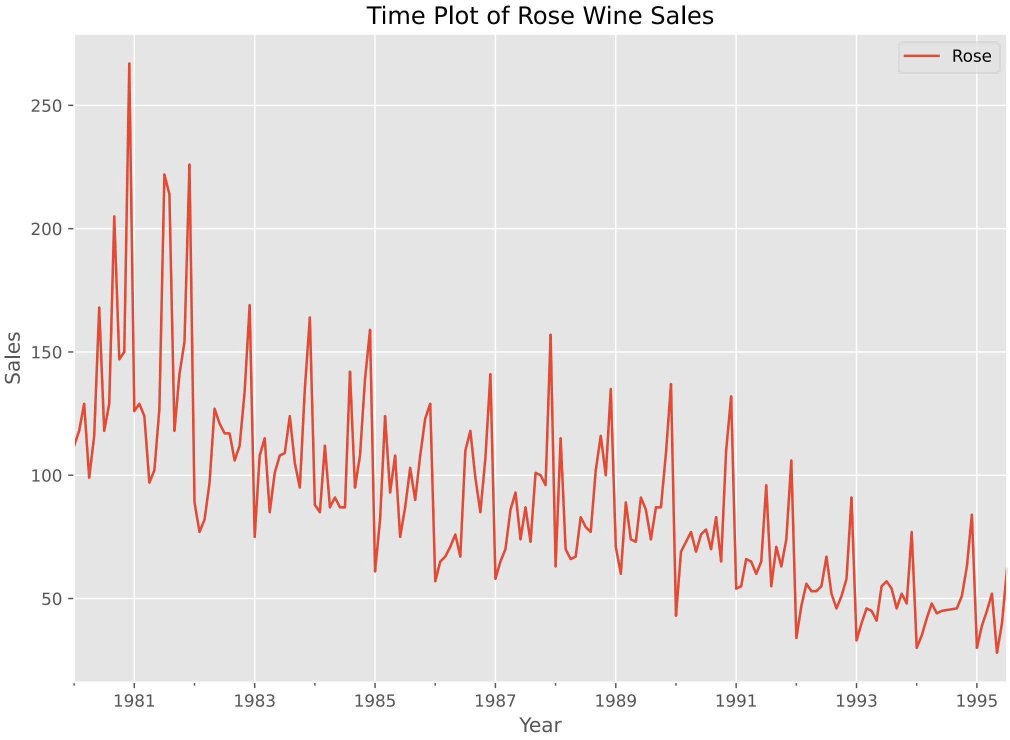


Figure 28: Time Plot of Rose Wine Sales

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

Distribution of Sales data

The below table shows the 5-point summary of the Rose column:

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.914439	39.238325	28.0	62.5	85.0	111.0	267.0

Table 17: Description of the Rose column

The overall mean sales of the Rose wine are 89.9 and the standard deviation is 39 which is quite high. The sales data is spread over a wide range from 28 to 267. Also, the mean and median are different from each other indicating that the data is skewed. This can be confirmed from the below histogram:

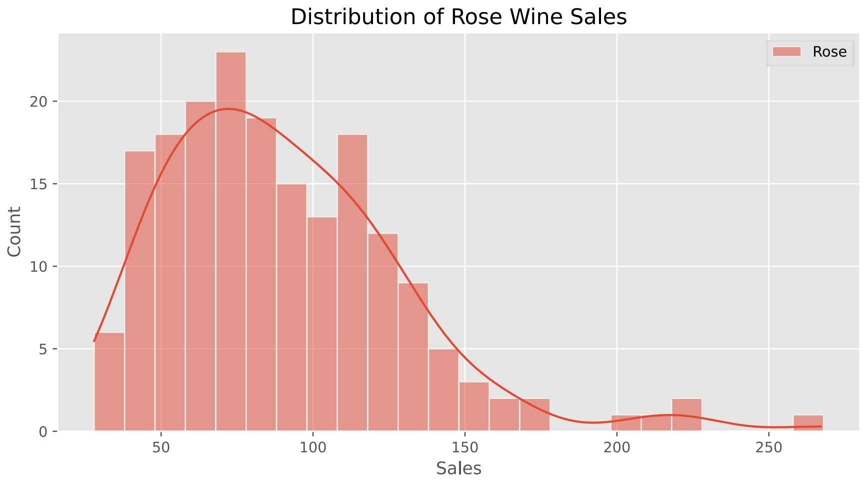


Figure 29: Distribution of Rose Wine Sales

Yearly Pointplot for Sales

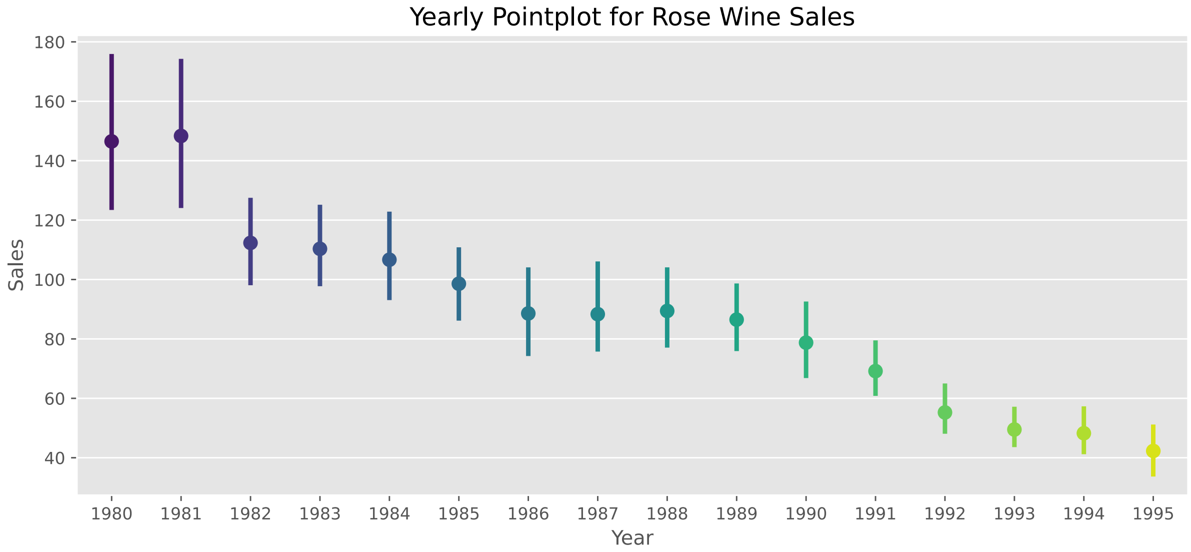


Figure 30: Yearly Pointplot for Rose Wine Sales

The above point plot shows the mean sales of Rose wine over the years. The wine **sales decreased continuously from 1980 to 1995**. Year 1981 had the highest average sales followed by year 1980. Also, the **spread of the data decreases as the mean sales decreases** indicating that a **multiplicative model would work better**.

Monthly Sales Across Years

The below plot shows the average monthly wine sales across the years. There is a slight increasing trend in the sales over the months. For all the years, the wine sales are low in the first half of the year and then increases towards the end of the years. Also, the sales in the month of July and August for years 1983 are higher than normal.

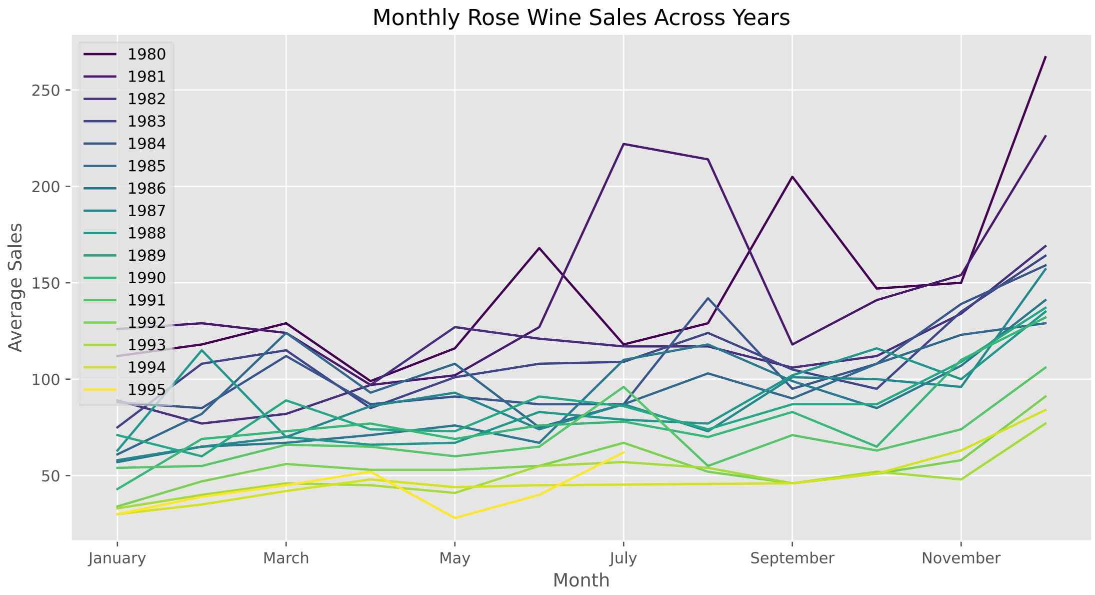


Figure 31: Monthly Rose Wine Sales Across Years

Decomposition

The original time series data of **the Rose wine sales was decomposed into its trend, seasonal and residual components using a multiplicative decomposition method**.

Additive Decomposition

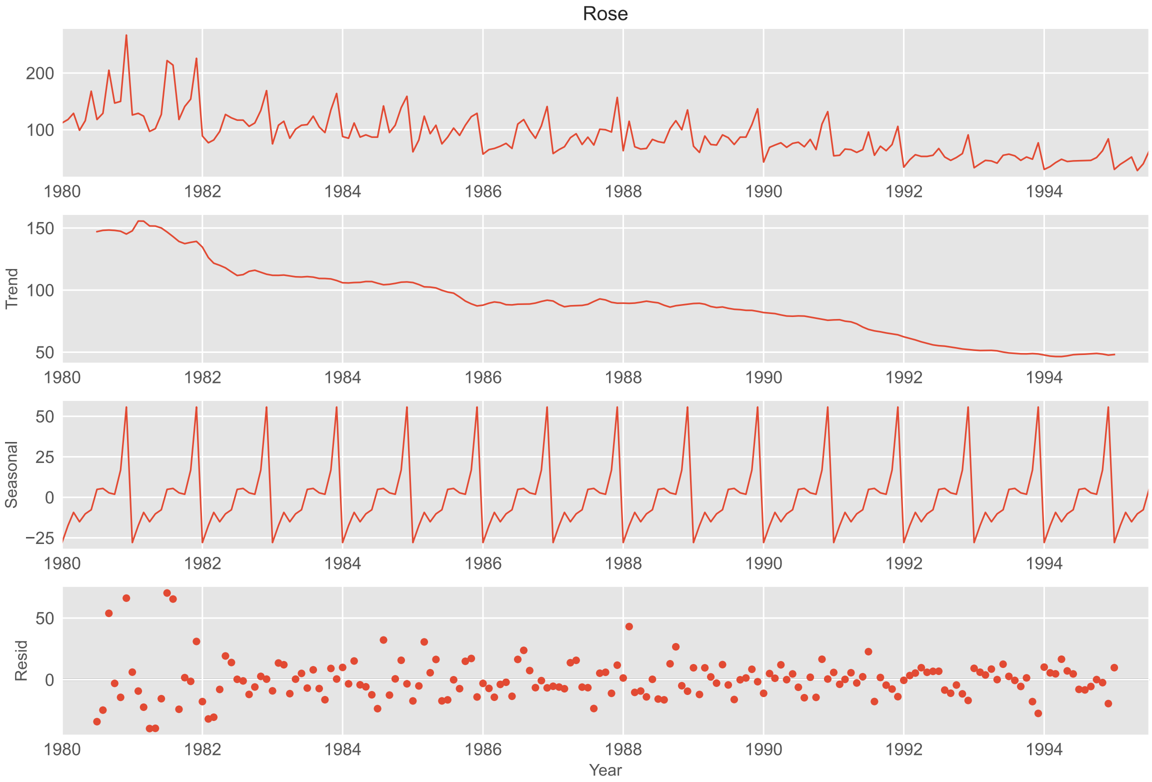


Figure 32: Additive Decomposition of Rose data

Multiplicative Decomposition

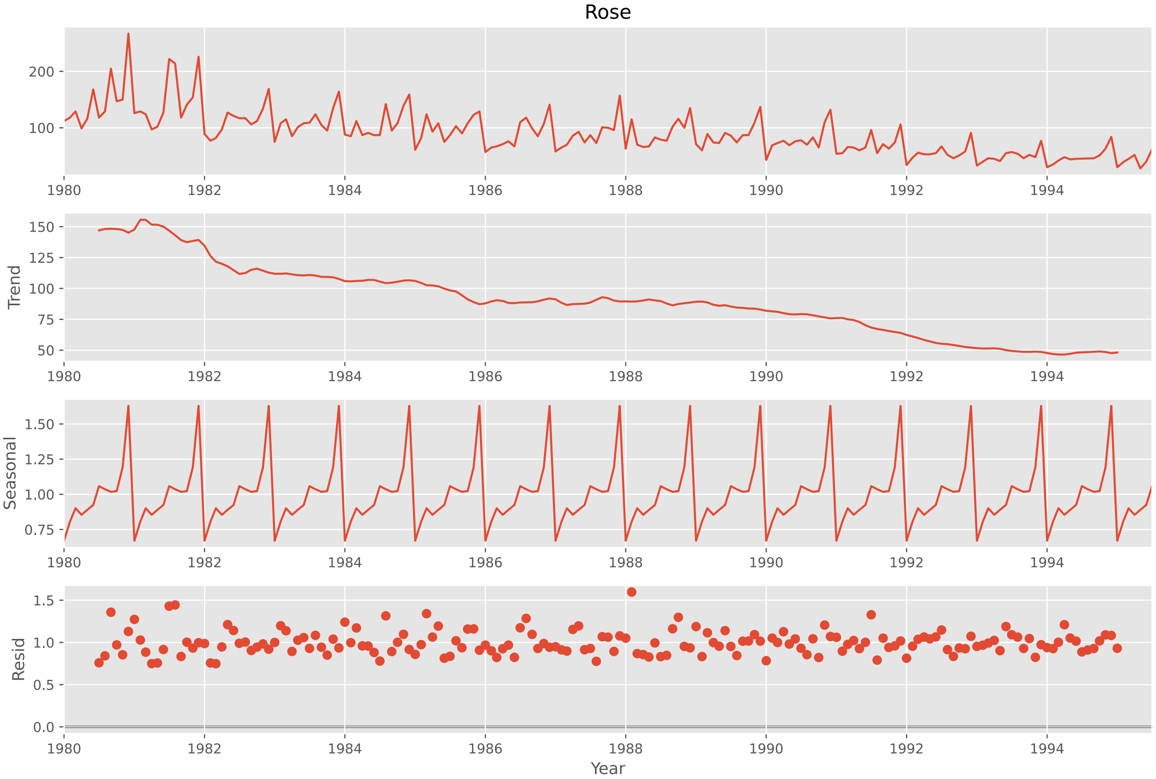


Figure 33: Multiplicative Decomposition of Rose data

From the above plots it can be seen that, the **data has a clear decreasing trend**. We can also confirm that the **sales peak at the end of each year**. The **sales in the month of December are almost 1.6 times the average sales** while the **sales in the month of January is just about 0.6 times of the average sales**. Also, a certain **pattern can be observed in the residual plot** indicating that **not all seasonal fluctuations are captured by the model**.

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

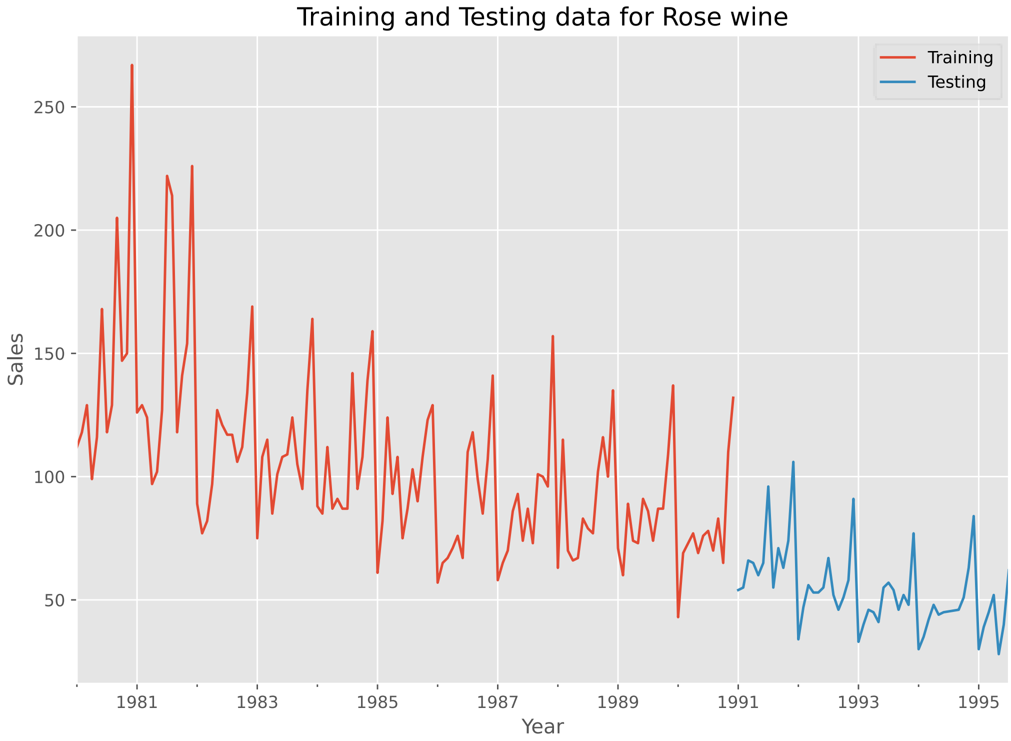


Figure 34: Training and Testing data for Rose wine

While splitting a time series data, the test set should contain data from the recent years. For the given problem, the **sales data from the year 1991 was put aside as a test set** and the **remaining data before 1991 was considered as a training set**. The above plot confirms that the split is done properly. The shape of the train and test sets after splitting are:

```
Shape of the training data: (132, 1)
Shape of the testing data: (55, 1)
```

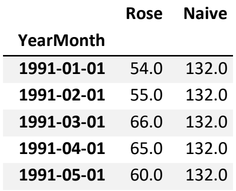
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on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

We will build various forecasting models on the Rose wine data and evaluate these models on the test set using RMSE metrics.

Naïve Forecast

A naïve forecast is a technique in **which the forecast for a given period is simply equal to the value observed in the previous period**. This leads to **all the forecast values in the test set to be equal to the last observation in the training set**.

The below table shows the naïve forecast for the Rose wine sales on the test set. We can see that the model predicted the same values for all the periods in the test set:



YearMonth	Rose	Naive
1991-01-01	54.0	132.0
1991-02-01	55.0	132.0
1991-03-01	66.0	132.0
1991-04-01	65.0	132.0
1991-05-01	60.0	132.0

Table 18: Naïve Forecast on the test set

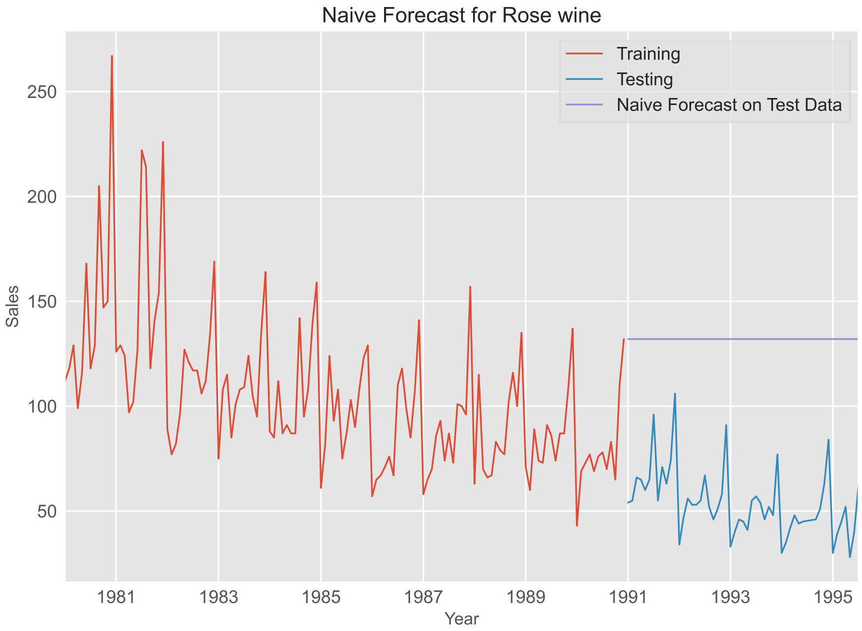


Figure 35: Naive Forecast for Rose wine

The above figure shows that a naïve forecast produces a straight line and does not capture either the trend or seasonality in the data as it is a very basic model. The model can be evaluated using RMSE metric and compared with other models. The below output shows the RMSE value for Naïve forecast model:

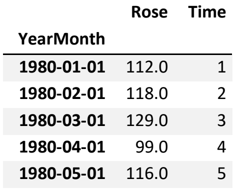
```
RMSE for Naive forecast model on Rose wine data: 79.719
```

The RMSE value is a lot higher than the standard deviation of the data. Therefore, the model does not perform well.

Linear Regression

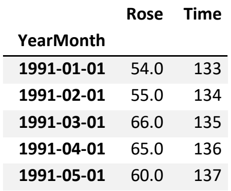
For a linear regression model, the data must contain at least one predictor variable. The Rose data only contains a target variable which is the sales of the Rose wine. Therefore, we will have to create a new predictor variable to build the model. Hence, we will regress the Rose column with the order of occurrence of the values.

The below tables show the training and test set after adding the new predictor variables:



YearMonth	Rose	Time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Table 19: Training set for Linear Regression



	Rose	Time
YearMonth		
1991-01-01	54.0	133
1991-02-01	55.0	134
1991-03-01	66.0	135
1991-04-01	65.0	136
1991-05-01	60.0	137

Table 20: Testing set for Linear Regression

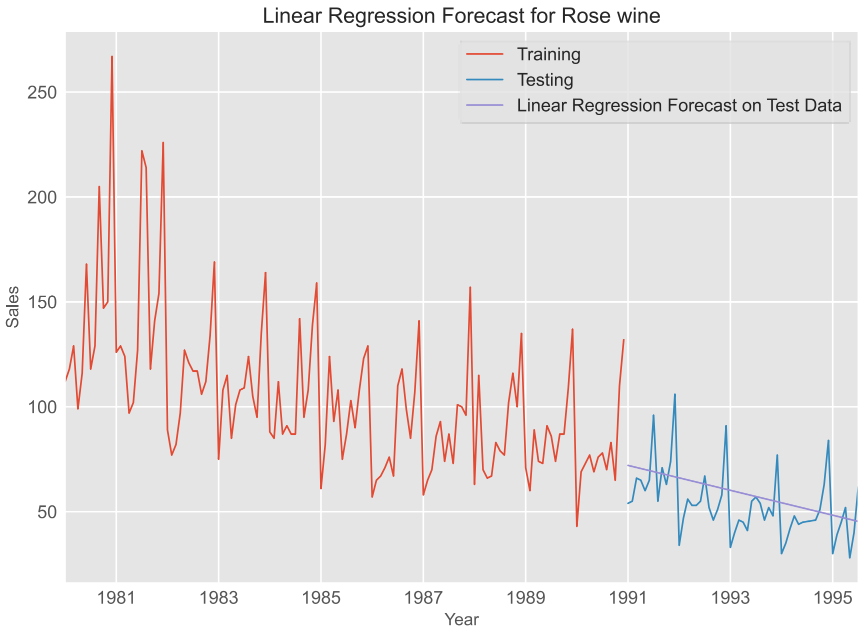


Figure 36: Linear Regression Forecast for Rose wine

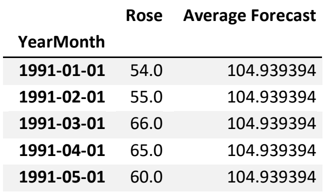
The linear regression model fits an inclined line through the data such that the RMSE is minimized. From the above figure we can see that the model only captures the trend of the time series data. The below output shows the RMSE value for Linear Regression model:

```
RMSE for Linear Regression forecast model on Rose wine data: 15.269
```

The RMSE value for the Linear Regression model is a lot better than that of the Naïve forecast model. Therefore, this model performs a lot better that Naïve forecast model.

Simple Average Model

In a Simple Average forecasting technique, the mean of the training data is used as the forecast for all the periods in the test data. The below table shows the forecast for the Simple Average model:



	Rose	Average Forecast
YearMonth		
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394

Table 21: Simple Average Forecast on the test set

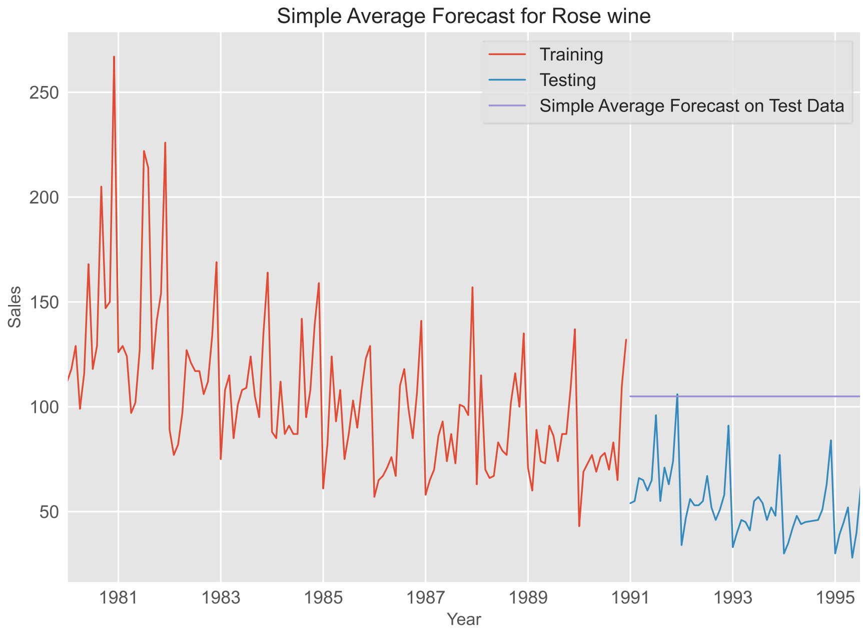


Figure 37: Simple Average Forecast for Rose wine

From the above figure we can see that, the Simple Average forecast is a straight line which passes through the mean of the training data. The model does not capture either the trend or seasonality in the data. The RMSE value for the model is shown below:

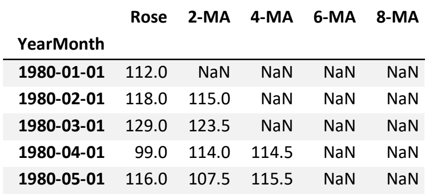
```
RMSE for Simple Average forecast model on Rose wine data: 53.461
```

The RMSE value for the Simple Average model is lower than that of the Naïve Forecast model but a lot greater than the Linear Regression model.

Moving Average (MA) Models

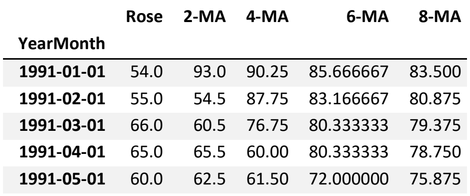
In a Moving Average forecasting technique, the forecast for a given period is the average of predetermined number of previous periods k. The moving averages smoothens the seasonal fluctuations in the time series data. The forecasts for first k-1 periods are null values as k number of data points are required to find the averages. Therefore, we will find various moving averages for the complete data and then split the data into training and test sets so that the test set does not contain any null values. We will calculate 2-, 4-, 6-, and 8-point moving averages for the entire data.

The below two tables show the various moving average forecast for the entire data and the test data. For the entire data we can confirm that first few rows contain null values.



YearMonth	Rose	2-MA	4-MA	6-MA	8-MA
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

Table 22: Moving Average Forecast for entire data



YearMonth	Rose	2-MA	4-MA	6-MA	8-MA
1991-01-01	54.0	93.0	90.25	85.666667	83.500
1991-02-01	55.0	54.5	87.75	83.166667	80.875
1991-03-01	66.0	60.5	76.75	80.333333	79.375
1991-04-01	65.0	65.5	60.00	80.333333	78.750
1991-05-01	60.0	62.5	61.50	72.000000	75.875

Table 23: Moving Average Forecast for test data

The smoothening effect produces by the moving averages depend on the window of rolling mean. Larger the window more is the smoothening effect. This can be confirmed from the below figure.

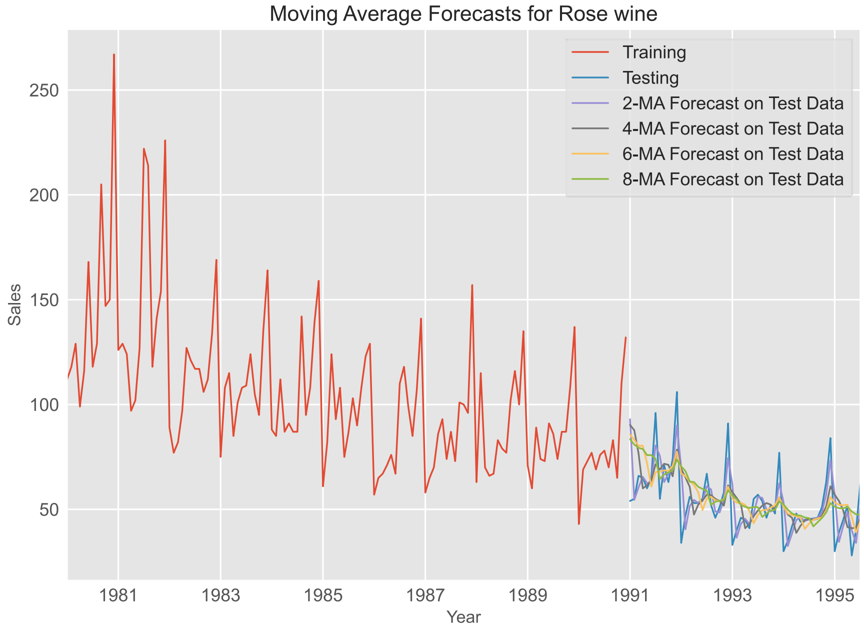


Figure 38: Moving Average Forecasts for Rose wine

The 2-point Moving Average is the closest to the test data and hence, it is expected to have the lowest RMSE value. The RMSE values for all the moving average models are:

```
RMSE for 2-MA forecast model on Rose wine data: 11.529
RMSE for 4-MA forecast model on Rose wine data: 14.451
RMSE for 6-MA forecast model on Rose wine data: 14.566
RMSE for 8-MA forecast model on Rose wine data: 14.805
```

The 2-point Moving Average model has the lowest RMSE among the all the MA models. It also has the lowest RMSE among all the previous built models like Simple Average, etc. **Therefore, we will consider 2-MA model for further model comparisons**.

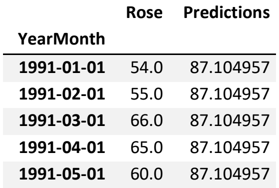
Simple Exponential Smoothing (SES) Model

Simple Exponential Smoothing (SES), is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha, also called the smoothing factor or smoothing coefficient. Parameter alpha ranges between 0 and 1. Large values of alpha means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

The model was built using least squares method of optimization to ensure convergence and the following parameters were found to be optimum:

```
{'smoothing_level': 0.099,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.387,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Here, the **smoothing_trend** and **smoothing_seasonal** parameters are NaN because we are using Simple Exponential Smoothing model. The **smoothing_level** parameter is the alpha value. The model found the optimal alpha value to be 0.099 i.e., the model used most of the previous period data to make predictions into the future. The below table shows the predictions done by the model:



	Rose	Predictions
YearMonth		
1991-01-01	54.0	87.104957
1991-02-01	55.0	87.104957
1991-03-01	66.0	87.104957
1991-04-01	65.0	87.104957
1991-05-01	60.0	87.104957

Table 24: SES Forecast for test data

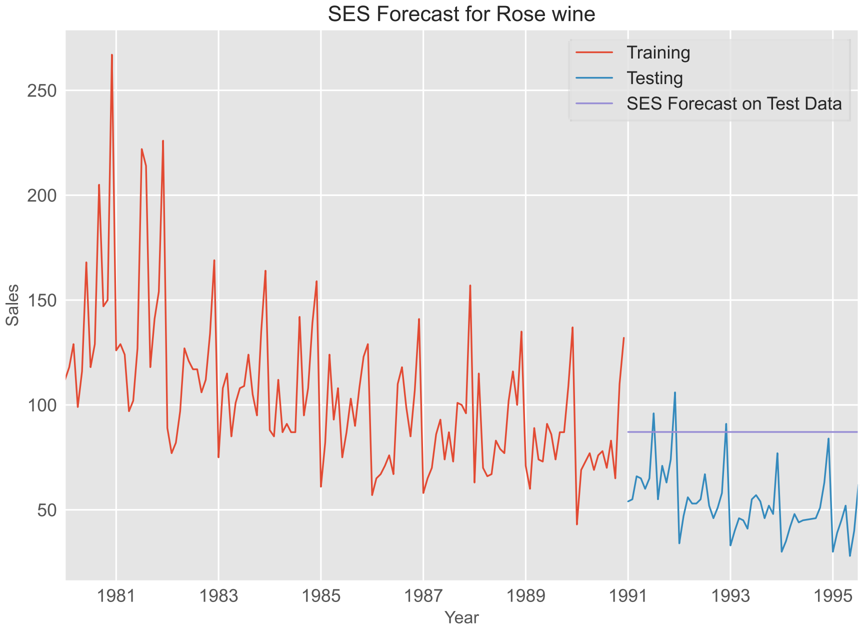


Figure 39: SES Forecast for Rose wine

From the above figure we see that the forecast done by the Simple Exponential Smoothing model is a straight line. This is due the fact that SES model is only used for data which does not have any trend or seasonality. The RMSE value for the Simple Exponential Smoothing model is:

```
RMSE for SES forecast model on Rose wine data: 36.796
```

The RMSE value of the SES model is lower than that of Simple Average model.

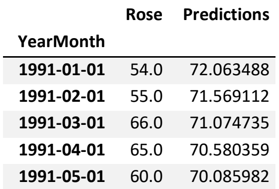
Holt (Double Exponential Smoothing) Model

Double exponential smoothing (Holt’s model) employs a level component and a trend component at each period. Holt’s model is used for data that has a trend but no seasonality component. Double exponential smoothing uses two smoothing parameters, alpha and beta, to update the components at each period. Alpha is the level smoothing parameter while beta is the trend smoothing parameter. Both parameter range between 0 and 1. Large values means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

The model was built using least squares method of optimization to ensure convergence and the following parameters were found to be optimum:

```
{'smoothing_level': 0.000,
 'smoothing_trend': 0.000,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.816,
 'initial_trend': -0.494,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Here, the **smoothing_seasonal** parameter is NaN because we are using Simple Exponential Smoothing model. The **smoothing_level** parameter is the alpha value while the **smoothing_trend** parameter is the beta value. The model found the optimal alpha value to be zero and beta value as zero. The below table shows the predictions done by the model:



	Rose	Predictions
YearMonth		
1991-01-01	54.0	72.063488
1991-02-01	55.0	71.569112
1991-03-01	66.0	71.074735
1991-04-01	65.0	70.580359
1991-05-01	60.0	70.085982

Table 25: DES Forecast for test data

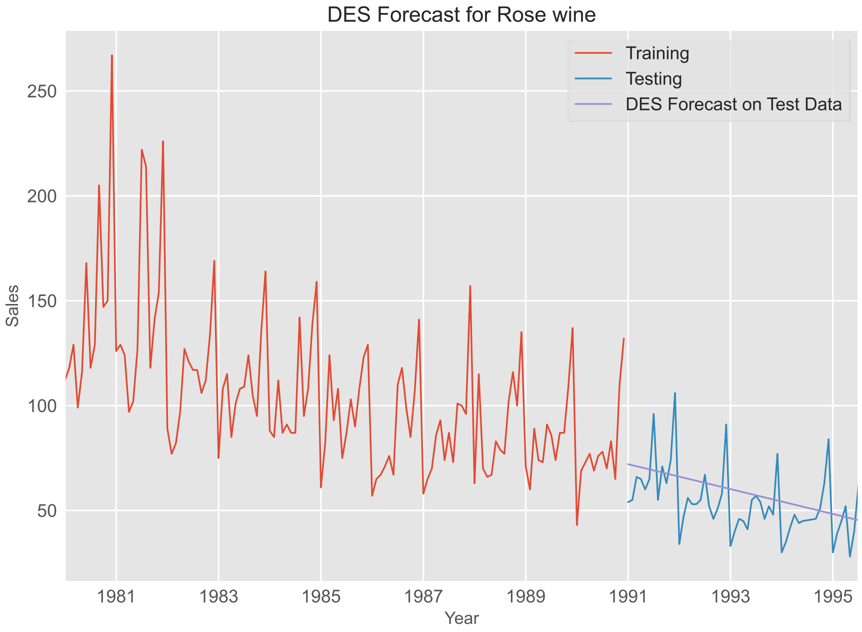


Figure 40: DES Forecast for Rose wine

From the above figure we see that the forecast done by the Double Exponential Smoothing model is an inclined line. This is due the fact that DES model is only used for data which has trend but no seasonality. The RMSE value for the Double Exponential Smoothing model is:

```
RMSE for DES forecast model on Rose wine data: 15.269
```

The RMSE for the Holt's model is almost the same as that of the Linear Regression model.

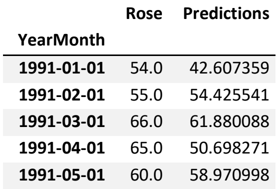
Holt-Winters (Triple Exponential Smoothing) Model

Triple exponential smoothing (Holt-Winters model) employs level, trend and seasonal components at each period. Holt-Winters model is used for data that has both trend and seasonality components. Triple exponential smoothing uses three smoothing parameters, alpha, beta and gamma, to update the components at each period. Alpha is the level smoothing parameter, beta is the trend smoothing parameter and gamma is the seasonal smoothing parameter. All of these parameters range between 0 and 1. Large values means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

Following parameters were found to be optimum:

```
{'smoothing_level': 0.089,
 'smoothing_trend': 0.000,
 'smoothing_seasonal': 0.000,
 'damping_trend': nan,
 'initial_level': -4371.952,
 'initial_trend': -0.549,
 'initial_seasons': array([4487.565, 4499.933, 4507.936, 4497.303, 4506.125, 4511.583,
        4521.495, 4527.59 , 4523.593, 4521.687, 4539.782, 4582.058]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

The **smoothing_level** parameter is the alpha value, the **smoothing_trend** parameter is the beta value and the **smoothing_seasonal** parameter is the gamma value. The model found the optimal alpha value to be 0.089, beta value as zero and gamma value as zero. The below table shows the predictions done by the model:



	Rose	Predictions
YearMonth		
1991-01-01	54.0	42.607359
1991-02-01	55.0	54.425541
1991-03-01	66.0	61.880088
1991-04-01	65.0	50.698271
1991-05-01	60.0	58.970998

Table 26: TES Forecast for test data

From the below figure we see that the forecast done by the Triple Exponential Smoothing model closely follows the actual test data. This is due the fact that TES model considers both trend and seasonality.

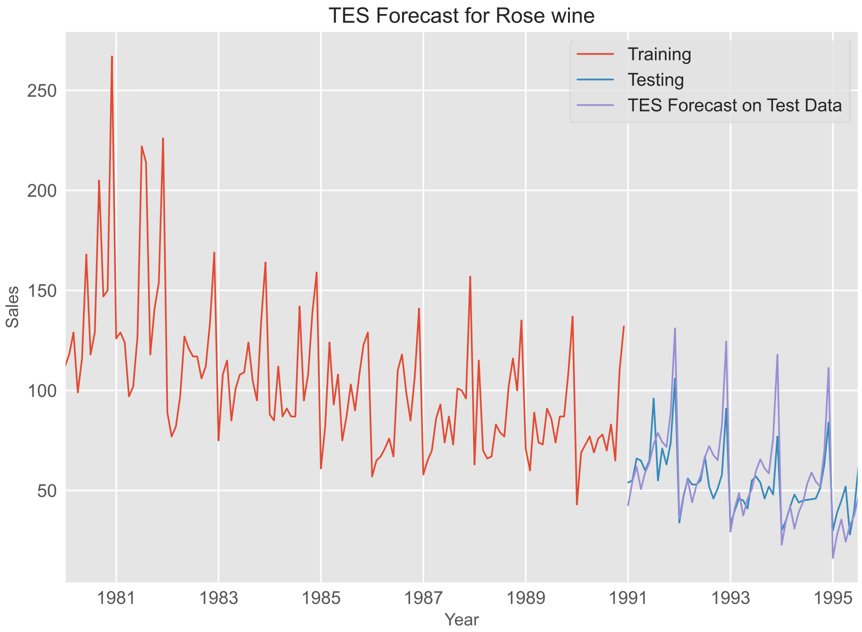


Figure 41: TES Forecast for Rose wine

The RMSE value for the Triple Exponential Smoothing model is:

```
RMSE for TES forecast model on Rose wine data: 14.257
```

The RMSE for the Holt-Winters model is the second lowest so far.

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.

Stationarity means that the statistical properties of a time series do not change over time. It does not mean that the series does not change over time, just that the **way** it changes does not itself change over time. Forecasting models like ARIMA or SARIMA requires that the time series data is stationary.

Checking for Stationarity on Whole data

The **Augmented Dickey-Fuller test is a unit root test** which determines whether there is a unit root and subsequently **whether the series is non-stationary**. It is a hypothesis test with the null and alternate hypothesis as follows:

H_0 : *The Time Series has a unit root and is thus non-stationary.*
H_1 : *The Time Series does not have a unit root and is thus stationary.*

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value. The below output shows the result of the ADF test performed on the Rose data:

```
Results of Dickey-Fuller Test:
Test Statistic          -2.240431
p-value                  0.467137
#Lags Used              13.000000
Critical Value (5%)     -3.436179
```

The **p-value from the ADF test is greater than 0.05** and therefore **we fail to reject the Null hypothesis**. Hence, the **Rose data is non-stationary**.

Differencing and Checking for Stationarity of Whole data

Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality thereby making the series stationary. As our data is a monthly data, one seasonal period is of 12 months. Therefore, to make the data stationary, we will take a seasonal difference of 12 (D=1) and an additional level difference of 1 (d=1). The below output shows the result of the ADF test performed on differenced Rose data:

```
Results of Dickey-Fuller Test:
Test Statistic          -4.550678
p-value                  0.001246
#Lags Used              11.000000
Critical Value (5%)     -3.437946
```

The **p-value from the ADF test is less than 0.05** and therefore **we reject the Null hypothesis**. Therefore, taking the **difference of the data has made the data stationary**.

Checking for Stationarity on Training data

The below output shows the result of the ADF test performed on the training data:

```
Results of Dickey-Fuller Test:
Test Statistic          -1.686149
p-value                  0.756909
#Lags Used              13.000000
Critical Value (5%)     -3.448373
```

The **p-value from the ADF test is greater than 0.05** and therefore **we fail to reject the Null hypothesis**. Hence, the **training data is non-stationary**.

Differencing and Checking for Stationarity of Training data

The same differencing as above (d=1 and D=1) is applied to the training data. Below figure shows the differenced training data:

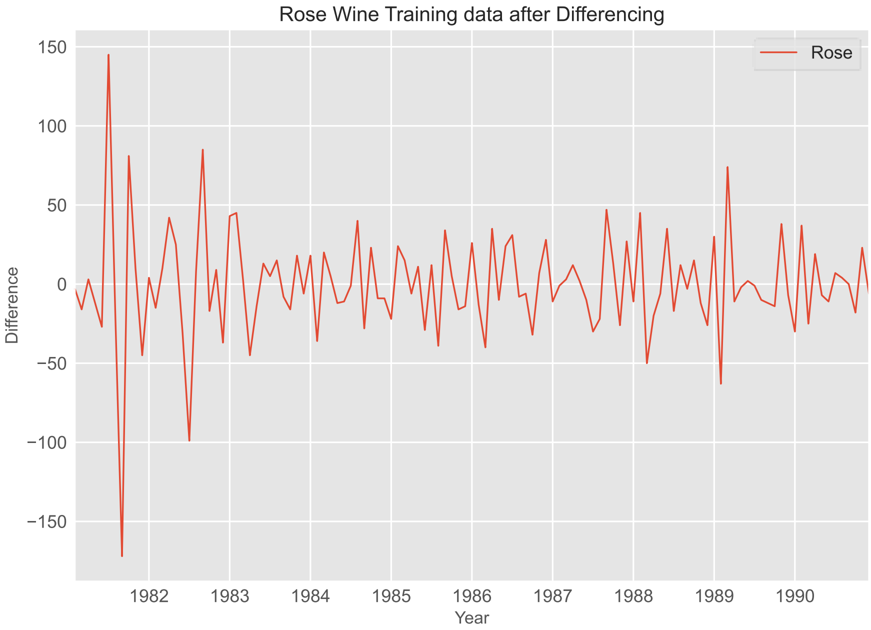


Figure 42: Rose Wine Training data after Differencing

From the above figure we see that, the trend and seasonality components in the training data have been reduced. This can be confirmed by performing the ADF test on the differenced training data:

```
Results of Dickey-Fuller Test:
Test Statistic           -3.631572
p-value                   0.027278
#Lags Used               11.000000
Critical Value (5%)      -3.452348
```

The **p-value from the ADF test is less than 0.05** and therefore **we reject the Null hypothesis**. Therefore, taking the **difference of the training data has made the data stationary**.

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.

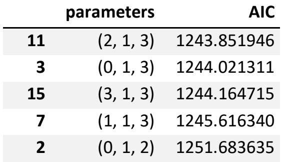
Automated ARIMA(p, d, q) Model

ARIMA stands for Auto Regressive Integrated Moving Average model. There are **total three parameters related to three different components** of the model. The **p parameter indicates the lag used in the Auto Regressive component**. The **d parameter is the order of level differencing applied to make the data stationary**. The **q parameter is the lag used in the Moving Average component**. Here, **we are using a value of d = 1**. Various combinations of the p, d and q parameter are used to build various ARIMA models and the **combination that give the lowest AIC value is used to evaluate the model on the test data**.

```
Some examples of the parameter combinations for the Model
Model: (0, 1, 0)
```

```
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
```

The below table shows the AIC values of different parameter combinations in ascending order:



	parameters	AIC
11	(2, 1, 3)	1243.851946
3	(0, 1, 3)	1244.021311
15	(3, 1, 3)	1244.164715
7	(1, 1, 3)	1245.616340
2	(0, 1, 2)	1251.683635

Table 27: AIC values for ARIMA model

The parameter combination of p=2, d=1, and q=3 gives the lowest AIC value. We will use this model for making predictions on the test data. The below output shows the model summary of ARIMA(2, 1, 3) model:

```
                               SARIMAX Results                                
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -615.926
Date:                Sat, 24 Sep 2022   AIC                           1243.852
Time:                        13:41:10   BIC                           1260.917
Sample:                    01-01-1980   HQIC                          1250.785
                         - 12-01-1990                                         
Covariance Type:                  opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.7004      0.043    -39.121      0.000      -1.786      -1.615
ar.L2         -0.8639      0.048    -17.973      0.000      -0.958      -0.770
ma.L1          0.9718      0.116      8.402      0.000       0.745       1.198
ma.L2         -0.6822      0.108     -6.314      0.000      -0.894      -0.470
ma.L3         -0.8991      0.102     -8.839      0.000      -1.098      -0.700
sigma2       885.7623      0.000   4.39e+06      0.000     885.762     885.763
===================================================================================
Ljung-Box (L1) (Q):                   1.03   Jarque-Bera (JB):                23.77
Prob(Q):                              0.31   Prob(JB):                         0.00
Heteroskedasticity (H):               0.36   Skew:                             0.76
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.47
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.97e+21. Standard errors may be unstable.
```

We see that all the coefficients for the AR and MA model have p-values less than 0.05. Therefore, all the coefficients are significant. The below figure shows the diagnostic plots for standardized residuals of the Rose variable:

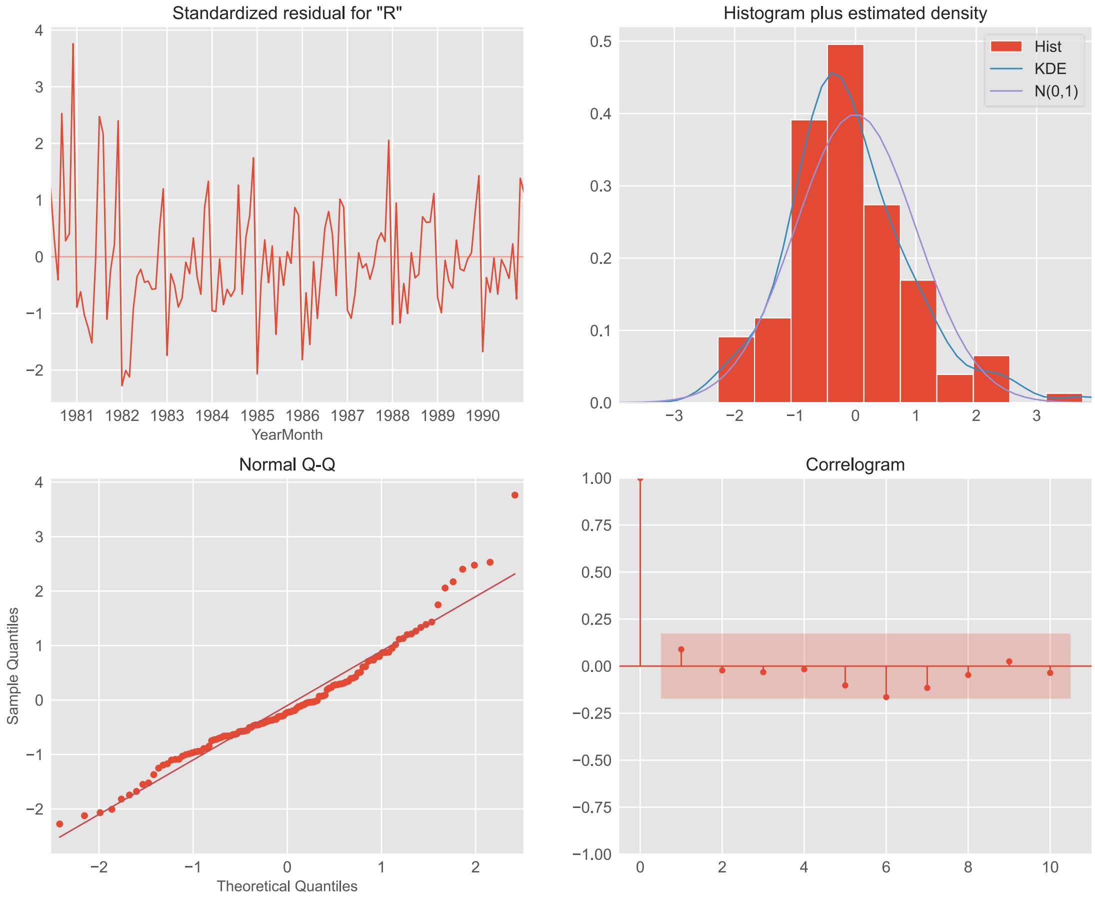


Figure 43: Diagnostic plots for Auto ARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Auto ARIMA forecast model on Rose wine data: 36.239
```

The RMSE of the Auto ARIMA model is almost similar to the SES model built above. This is expected as the ARIMA model also do not take into consideration the seasonal fluctuations in the data. From the below figure we see that the predictions fluctuate in the begin and then flattens out at the end.

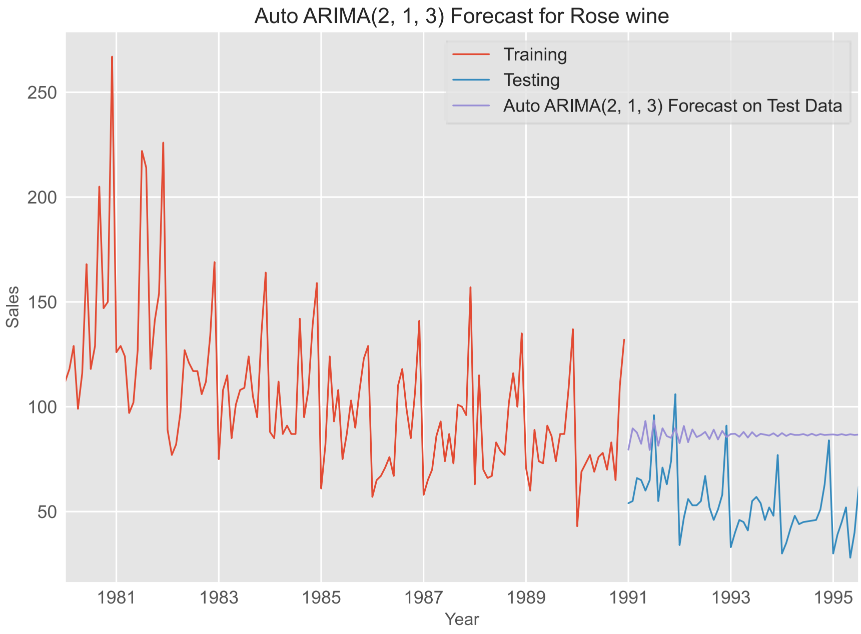


Figure 44: Auto ARIMA(2, 1, 3) Forecast for Rose wine

Automated SARIMA(p, d, q)(P, D, Q, F) Model

SARIMA stands for Seasonal Auto Regressive Integrated Moving Average model. It is an extension of the ARIMA model. There are total 7 parameters used to define the model. The parameters (p, d, q) are the same as the ARIMA model. The parameters (P, D, Q) are the seasonal counterparts of (p, d, q). The parameter F is the seasonality of the data which is 12 in our case. Here, **we are using a value of d = 1 and D = 1**. Various combinations of the p, d, q, P, D, and Q parameter are used to build various SARIMA models and the **combination that give the lowest AIC value is used to evaluate the model on the test data**.

```
Some examples of the parameter combinations for the Model
Model: (0, 1, 0)(0, 1, 0, 12)
Model: (0, 1, 1)(0, 1, 1, 12)
Model: (0, 1, 2)(0, 1, 2, 12)
Model: (0, 1, 3)(0, 1, 3, 12)
Model: (1, 1, 0)(1, 1, 0, 12)
```

The below table shows the AIC values of different parameter combinations in ascending order:

	parameters	seasonal	AIC
221	(3, 1, 1)	(3, 1, 1, 12)	681.362807
253	(3, 1, 3)	(3, 1, 1, 12)	681.608485
254	(3, 1, 3)	(3, 1, 2, 12)	681.985200
222	(3, 1, 1)	(3, 1, 2, 12)	682.320701
237	(3, 1, 2)	(3, 1, 1, 12)	683.211700

Table 28: AIC values for SARIMA model

The parameter combination of p=3, d=1, q=1, P=3, D=1, Q=1, and F=12 gives the lowest AIC value. We will use this model for making predictions on the test data. The below output shows the model summary of SARIMA(3, 1, 1)(3, 1, 1, 12) model:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 1, 1, 12)   Log Likelihood                -331.681
Date:                            Sat, 24 Sep 2022   AIC                            681.363
Time:                                    13:43:31   BIC                            702.801
Sample:                                01-01-1980   HQIC                           689.958
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1           0.0173      0.151      0.114      0.909      -0.279       0.314
ar.L2          -0.0426      0.141     -0.302      0.763      -0.319       0.234
ar.L3          -0.0575      0.119     -0.485      0.628      -0.290       0.175
ma.L1          -0.9388      0.085    -11.108      0.000      -1.104      -0.773
ar.S.L12        0.0907      0.126      0.721      0.471      -0.156       0.337
ar.S.L24       -0.0437      0.108     -0.406      0.685      -0.255       0.167
ar.S.L36    -3.662e-05      0.053     -0.001      0.999      -0.103       0.103
ma.S.L12       -0.9999    428.099     -0.002      0.998    -840.059     838.059
sigma2        185.3876   7.94e+04      0.002      0.998   -1.55e+05    1.56e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 2.56
Prob(Q):                              0.91   Prob(JB):                         0.28
Heteroskedasticity (H):               0.56   Skew:                             0.42
Prob(H) (two-sided):                  0.13   Kurtosis:                         3.22
===================================================================================

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

We see that almost all the coefficients have p-values greater than 0.05. Therefore, these coefficients are not significant. The below figure shows the diagnostic plots for standardized residuals of the Rose variable:

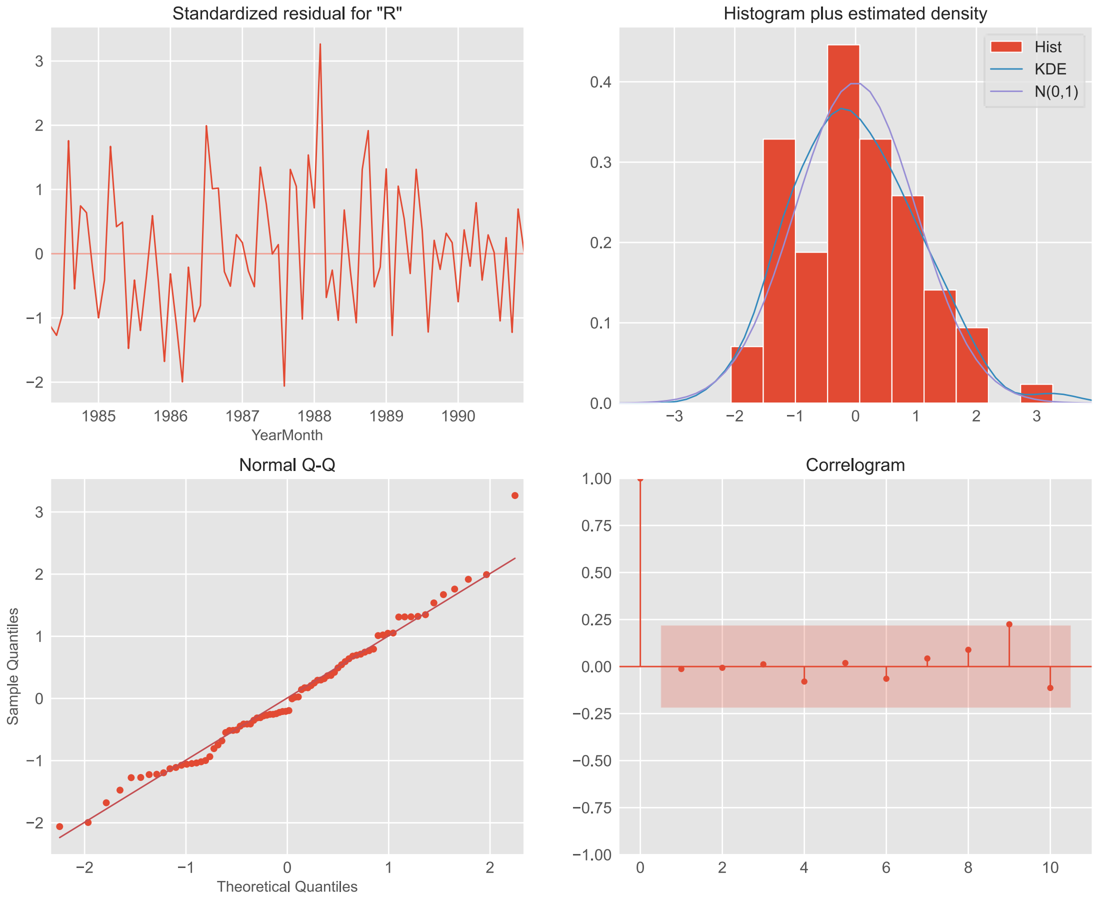


Figure 45: Diagnostic plots for Auto SARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Auto SARIMA forecast model on Rose wine data: 16.824
```

From the below figure we see that the predictions follow the test data closely.

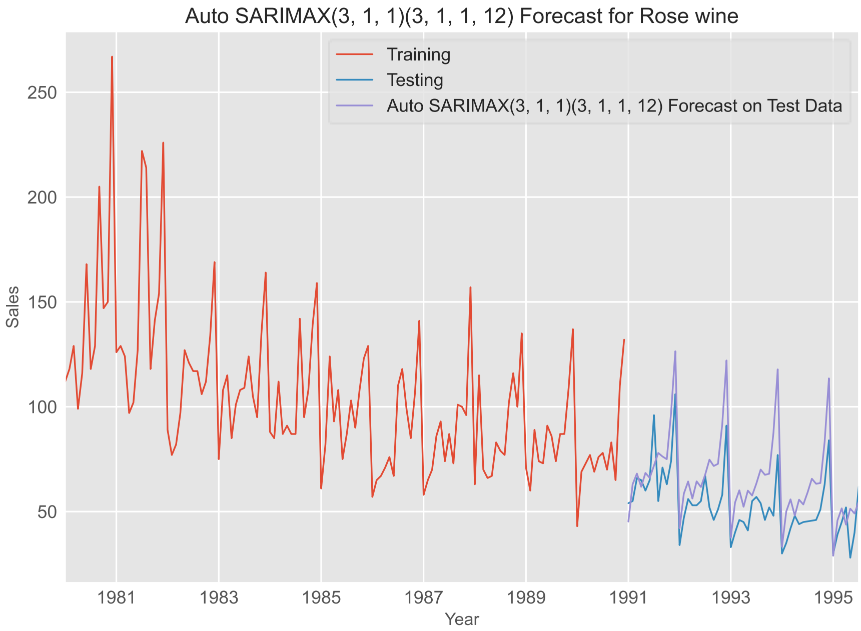


Figure 46: Auto SARIMAX(3, 1, 1)(3, 1, 1, 12) Forecast for Rose wine

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.

PACF Plot

PACF or **Partial Autocorrelation Function is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed**. The Auto-Regressive parameter in an ARIMA/SARIMA model is **'p' which comes from the significant lag before which the PACF plot cuts-off to 0**.

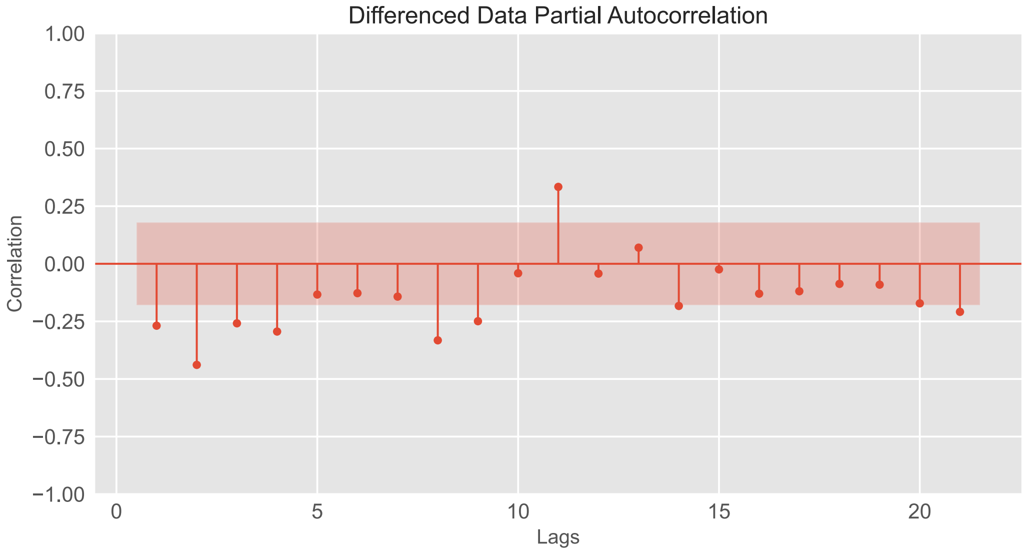


Figure 47: Differenced Data Partial Autocorrelation

From the above PACF plot we see that, the PACF cuts-off after 4th lag. Using p = 4 leads to convergence issues and requires a lot of iterations. Therefore, **we will use p = 3** for the models. Also, the PACF for 12th lag is inside the 0.05 interval. Therefore, **we will use P = 0** for the models.

ACF Plot

ACF or **Autocorrelation Function is a way to measure the linear relationship between an observation at time *t* and the observations at previous times**. The Moving-Average parameter in an ARIMA model is **'q' which comes from the significant lag before the ACF plot cuts-off to 0**.

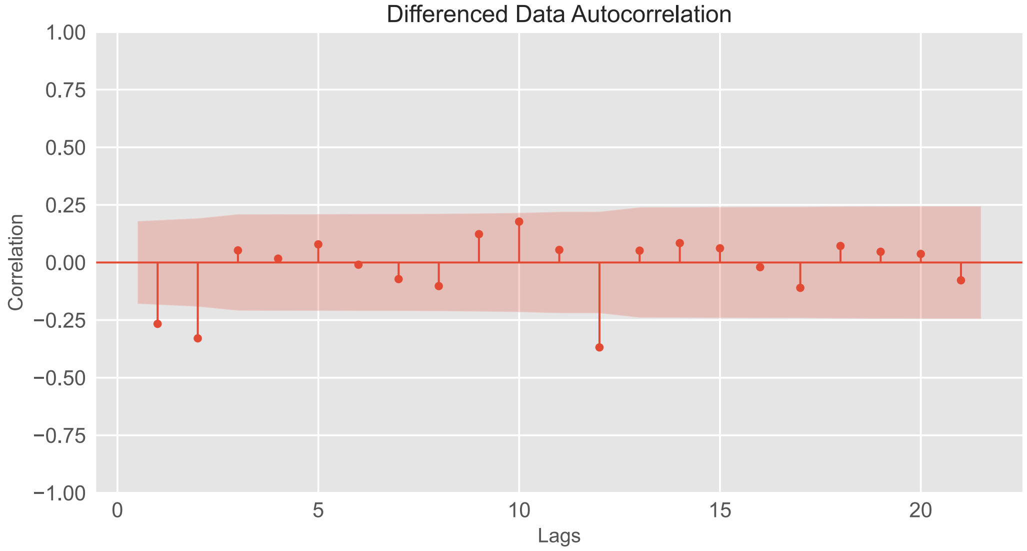


Figure 48: Differenced Data Autocorrelation

From the above ACF plot we see that, the ACF cuts-off after 2nd lag. Therefore, **we will use q = 2** for the models. Also, after the 12th lag, the ACF cuts-off. Therefore, **we will use Q = 1** for the models.

Manual ARIMA(p, d, q) Model

The best ARIMA model suggested by the above ACF and PACF plots is ARIMA(3, 1, 2). The below output shows the model summary of ARIMA(3, 1, 2) model:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(3, 1, 2)   Log Likelihood                -622.692
Date:                Sat, 24 Sep 2022   AIC                           1257.385
Time:                        13:44:47   BIC                           1274.497
Sample:                    01-01-1980   HQIC                          1264.337
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7952      0.084     -9.459      0.000      -0.960      -0.630
ar.L2          0.0810      0.131      0.620      0.535      -0.175       0.337
ar.L3         -0.1237      0.102     -1.213      0.225      -0.324       0.076
ma.L1         -0.1055      1.327     -0.079      0.937      -2.707       2.496
ma.L2         -1.1034      1.464     -0.754      0.451      -3.973       1.766
sigma2       768.1444   1026.846      0.748      0.454   -1244.438    2780.727
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                30.98
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.36   Skew:                             0.84
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.73
===================================================================================

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

We see that almost all the coefficients have p-values greater than 0.05. Therefore, these coefficients are not significant. The below figure shows the diagnostic plots for standardized residuals of the Rose variable:

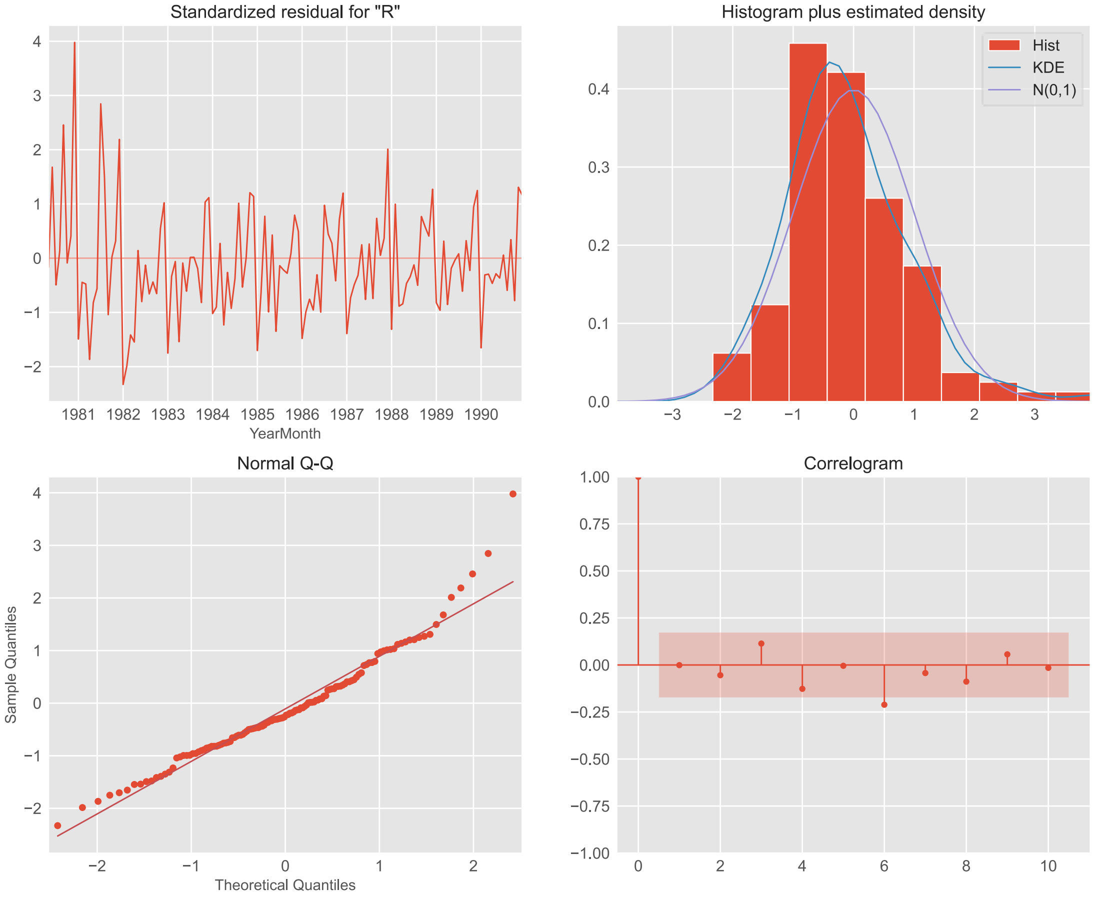


Figure 49: Diagnostic plots for Manual ARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Manual ARIMA forecast model on Rose wine data: 36.342
```

The RMSE of the Manual ARIMA model is slightly greater than that of the Auto ARIMA model. From the below figure we see that the predictions fluctuate till the end.

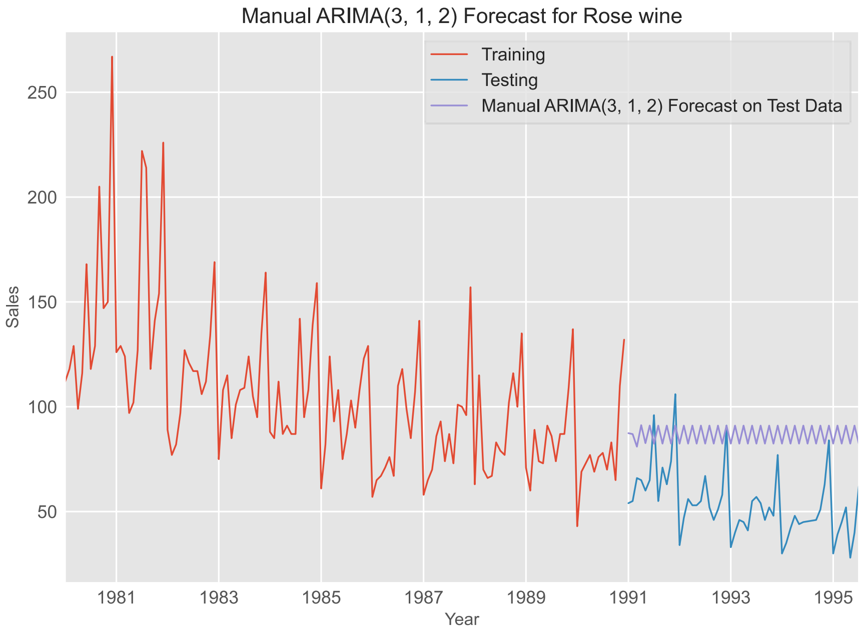
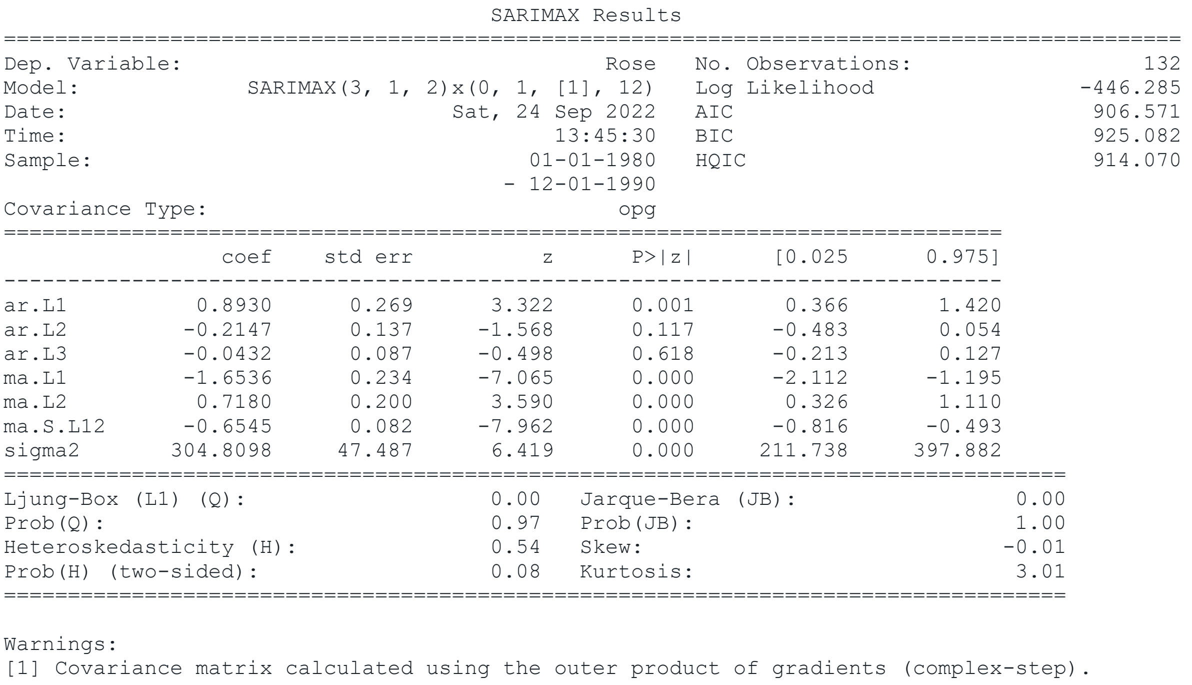


Figure 50: Manual ARIMA(3, 1, 2) Forecast for Rose wine

Manual SARIMA(p, d, q)(P, D, Q, F) Model

The best SARIMA model suggested by the above ACF and PACF plots is SARIMA(3, 1, 2)(0, 1, 1, 12). The below output shows the model summary of SARIMA(3, 1, 2)(0, 1, 1, 12) model:



```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(0, 1, [1], 12)   Log Likelihood                -446.285
Date:                              Sat, 24 Sep 2022   AIC                            906.571
Time:                                      13:45:30   BIC                            925.082
Sample:                                  01-01-1980   HQIC                           914.070
                                       - 12-01-1990                                         
Covariance Type:                                opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8930      0.269      3.322      0.001       0.366       1.420
ar.L2         -0.2147      0.137     -1.568      0.117      -0.483       0.054
ar.L3         -0.0432      0.087     -0.498      0.618      -0.213       0.127
ma.L1         -1.6536      0.234     -7.065      0.000      -2.112      -1.195
ma.L2          0.7180      0.200      3.590      0.000       0.326       1.110
ma.S.L12      -0.6545      0.082     -7.962      0.000      -0.816      -0.493
sigma2       304.8098     47.487      6.419      0.000     211.738     397.882
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 0.00
Prob(Q):                              0.97   Prob(JB):                         1.00
Heteroskedasticity (H):               0.54   Skew:                            -0.01
Prob(H) (two-sided):                  0.08   Kurtosis:                         3.01
===================================================================================

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

We see that the coefficients ar.L2 and ar.L3 have p-values greater than 0.05. Therefore, these coefficients are not significant. The below figure shows the diagnostic plots for standardized residuals of the Rose variable:

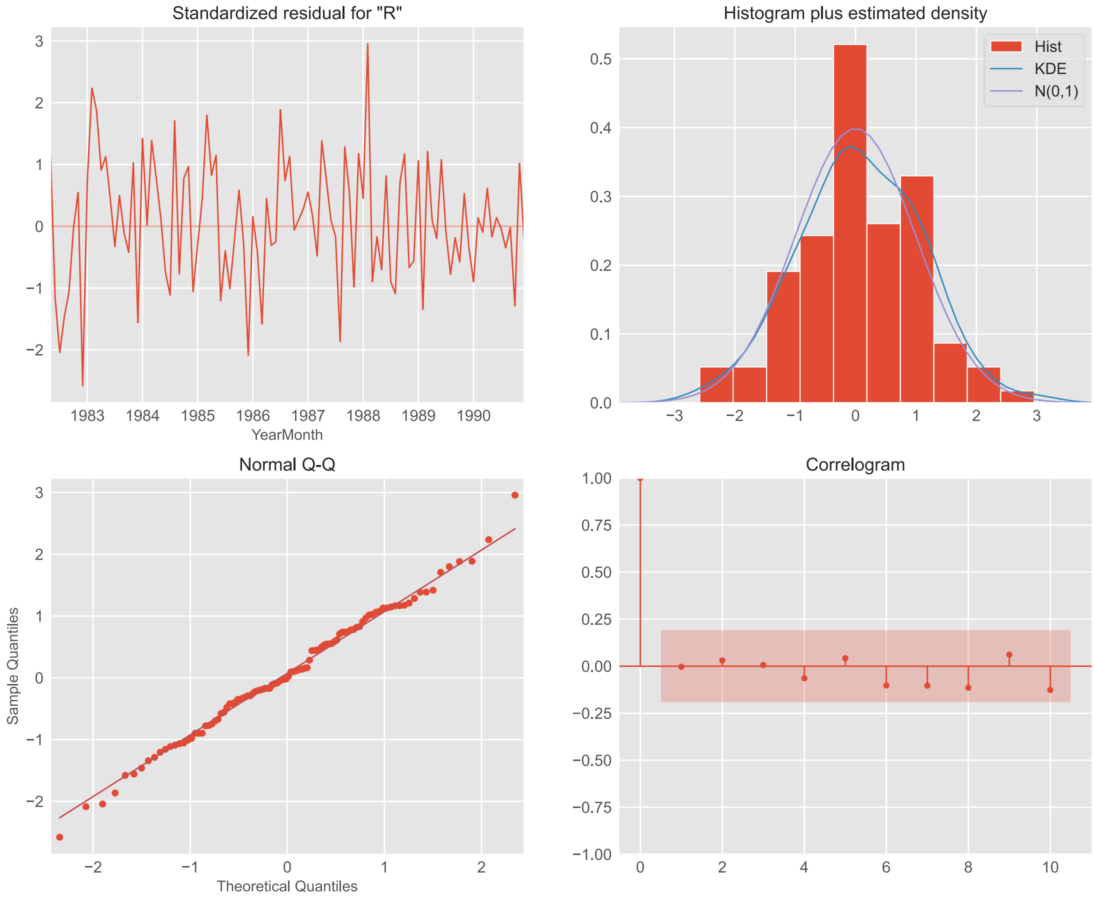


Figure 51: Diagnostic plots for Manual SARIMA model

From the above plot we confirm that the residuals are random and follow a normal distribution. Correlogram of residuals indicates that they are stationary in nature and have no pattern.

```
RMSE for Manual SARIMA forecast model on Rose wine data: 14.403
```

The RMSE of the Manual SARIMA model is lower than that of the Auto SARIMA model built above. From the below figure we see that the predictions follow the test data closely.

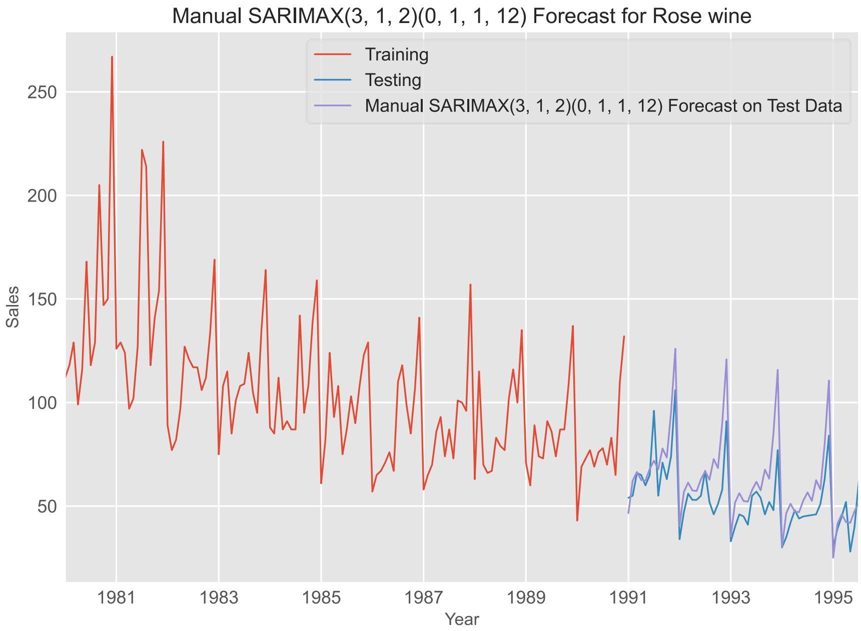
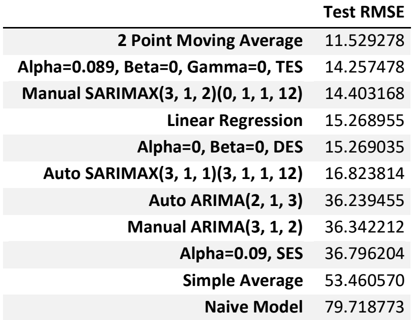


Figure 52: Manual SARIMAX(3, 1, 2)(0, 1, 1, 12) Forecast for Rose wine

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.

The below table shows the summary of RMSE values of the all the models built above in ascending order:



	Test RMSE
2 Point Moving Average	11.529278
Alpha=0.089, Beta=0, Gamma=0, TES	14.257478
Manual SARIMAX(3, 1, 2)(0, 1, 1, 12)	14.403168
Linear Regression	15.268955
Alpha=0, Beta=0, DES	15.269035
Auto SARIMAX(3, 1, 1)(3, 1, 1, 12)	16.823814
Auto ARIMA(2, 1, 3)	36.239455
Manual ARIMA(3, 1, 2)	36.342212
Alpha=0.09, SES	36.796204
Simple Average	53.460570
Naive Model	79.718773

Table 29: RMSE values of all the models

Therefore, the most optimum model for the Rose wine dataset is the 2-Point Moving Average model as it has the lowest RMSE value. The second-best optimum model is Triple Exponential Smoothing model. Hence, we will use these two models to train on the complete data and predict 12 months into the future.

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.

The most optimum model for the Rose wine dataset was found to be the 2-Point Moving Average model. Moving Average models are used to see the general trend in the data by reducing the seasonal fluctuations. But while making predictions into the future, the MA models only predict the last known moving average from the training data, for all the future periods. This can be confirmed from the below output and the figure:

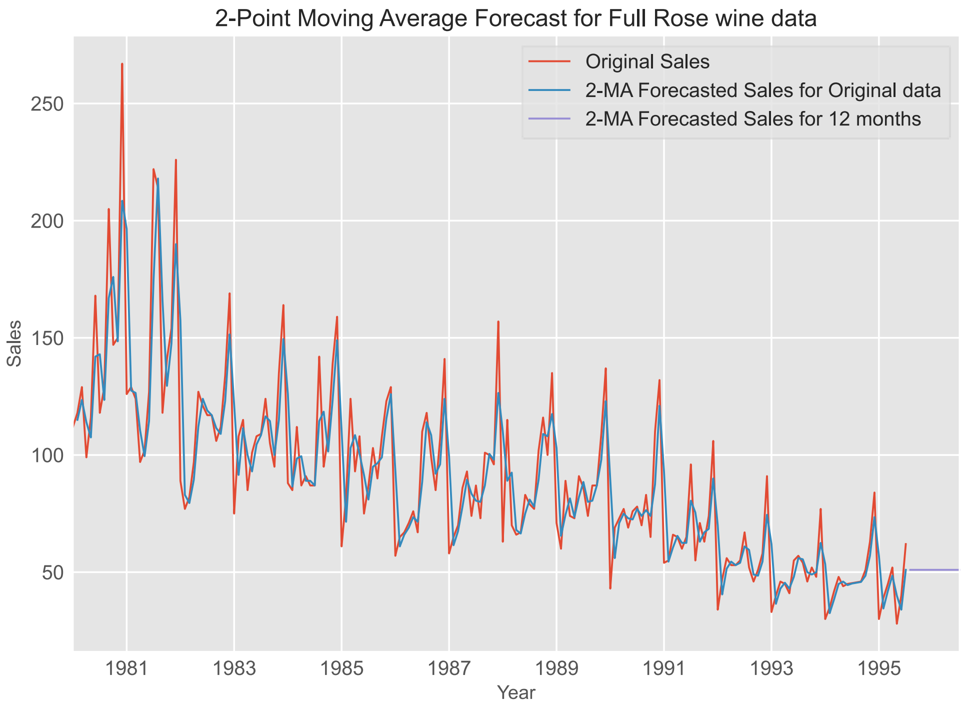


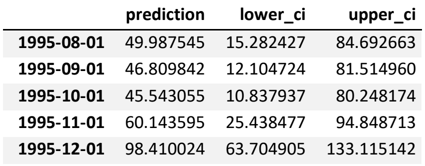
Figure 53: 2-Point Moving Average Forecast for Full Rose wine data

```
1995-08-01    51.0
1995-09-01    51.0
1995-10-01    51.0
1995-11-01    51.0
1995-12-01    51.0
Freq: MS, Name: Predictions, dtype: float64
```

Therefore, we will use the second-best model TES for future prediction for 12 months. The below output shows the most optimum parameters found by the TES model:

```
{'smoothing_level': 0.097,
 'smoothing_trend': 0.000,
 'smoothing_seasonal': 0.000,
 'damping_trend': nan,
 'initial_level': -8311426.530,
 'initial_trend': -0.538,
 'initial_seasons': array([8311544.241, 8311555.091, 8311563.191, 8311556.541, 8311560.454,
        8311566.491, 8311577.675, 8311577.608, 8311574.968, 8311574.239,
        8311589.377, 8311628.181]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

The model has found the alpha parameter to be 0.097, beta parameter to be zero and gamma parameter to be zero as the most optimum parameters. The below table and figure shows the next 12 months of predictions done by the final TES model with 95% confidence intervals:



	prediction	lower_ci	upper_ci
1995-08-01	49.987545	15.282427	84.692663
1995-09-01	46.809842	12.104724	81.514960
1995-10-01	45.543055	10.837937	80.248174
1995-11-01	60.143595	25.438477	94.848713
1995-12-01	98.410024	63.704905	133.115142

Table 30: Predictions of final TES model

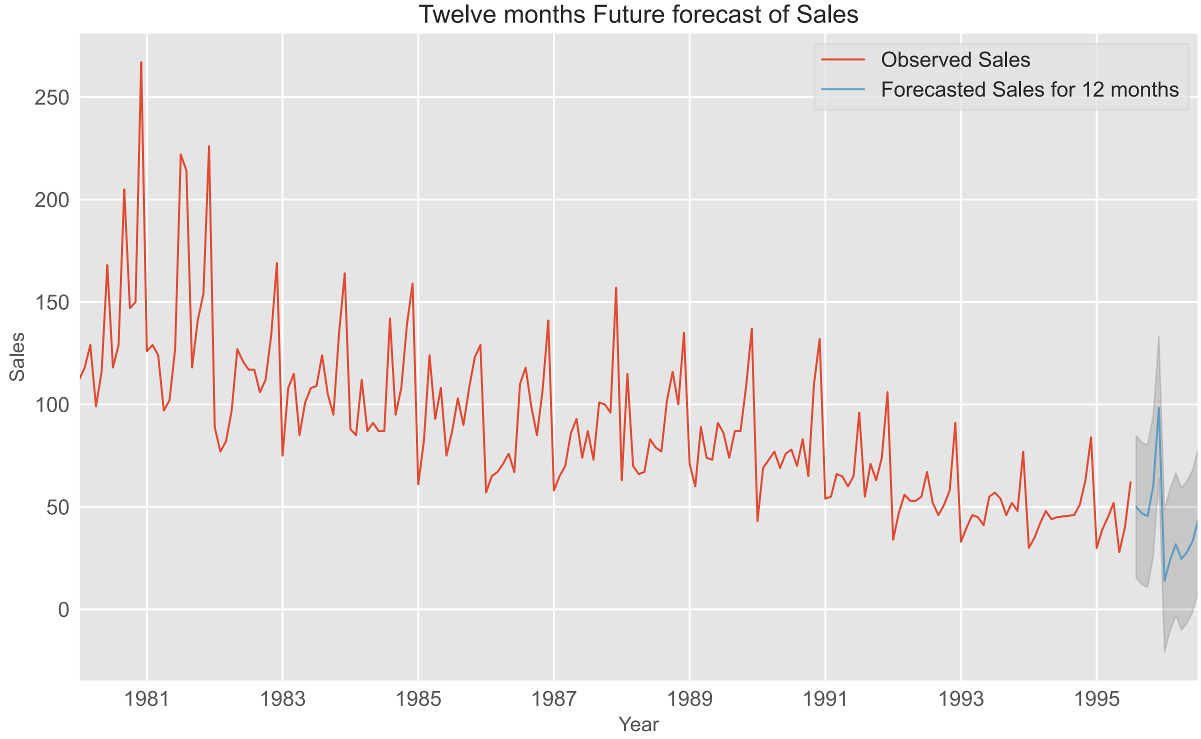


Figure 54: Twelve months Future forecast of Rose wine sales

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

We were tasked to analyse and forecast Rose Wine Sales in the 20th century for ABC Estate Wines and predict the sales for next 12 months. The time series data was read into a dataframe and proper Datetime Index were provided for the dataframe. The data contained monthly sales of Rose wine from 1980 to 1995. The data contained two missing values which were imputed using linear interpolation. While doing the EDA and decomposing the data, we found that wine sales peaked during fourth quarter of every year. The Rose wine sales show a decreasing trend over the years. The sales data from 1991 was used as test data for evaluating the models. Various basic models like Naïve forecast, Linear Regression, Simple and Moving Average, Exponential Smoothing, etc. were built and evaluated using RMSE. The data was found to be non-stationary and proper differencing was applied to make the data stationary. Automated and Manual versions of both ARIMA and SARIMA models were built and evaluated. The 2-Point Moving Average and Triple Exponential Smoothing models were found to be the best models. The 2-Point Moving Average model when used for forecasting did not give good predictions and hence TES model was used to forecast 12 months into the future.

Inferences

1. Sales peak in the last quarter of every year. Especially in December, the sales are almost 2 times the yearly average sales. This could be due to the holiday season.
2. Sales are the lowest during the first quarter of every year. This might be due the after effects of the holidays in the last quarter of previous year.
3. The Rose wine sales shows a continuous decreasing trend over the years indicating decreasing popularity of the wine.
4. The Rose wine sales are expected to reach up to 100 in December 1995.

Recommendations

1. The company should prepare for the high demand during the last quarter of the year by increasing production and stocking their inventory well before that period, to allow the wine to age.
2. The company can provide offers during the first half of the year to increase the sale in that period.
3. The general trend in the sales over the years is decreasing continuously indicating that the Rose wine is losing its popularity. The company can try to improve the quality of the wine and try different taste variants.