# TIME SERIES FORECAST

|  |  |
| --- | --- |
| S.No | Content |
|  | [**Read the data as an appropriate Time Series data and plot the data.**](#read_data_q1) |
|  | [**Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition**](#q2) |
|  | [**Split the data into training and test. The test data should start in 1991.**](#q3) |
|  | [**Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE**](#q4)**.** |
|  | [**Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.**](#q5) |
|  | [**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.**](#q6) |
|  | [**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.**](#q7) |
|  | [**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.**](#q8) |
|  | [**Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**](#q9) |

**List of figures:**

|  |  |
| --- | --- |
| s.no | List of Figures |
|  | [Figure 1:Rose wines plot](#figure1) |
|  | [Figure 2: Sparkling Wines plot](#figure2) |
|  | [Figure 3: Rose wines plot after interpolation](#figure3) |
|  | [Figure 4:*ROSE WINES EDA*](#f4)*:* |
|  | [Figure 5: *SPARKLING WINES EDA*](#f5) |
|  | [Figure 6: Decompose of Sparkling wines Additive and Multiplicative](#f6) |
|  | [Figure 7: Decompose of Rose wines Additive and Multiplicative](#f7) |
|  | [Figure 8: Autocorrelation plot for Sparkling wines](#f8) |
|  | [Figure 9: Autocorrelation plot for Rose wines](#f9) |
|  | [Figure 10](#f10): [Alpha =0.995 Simple Exponential Smoothing predictions Rose wines](#f10) |
|  | [Figure 11](#f11): [Alpha =0.995 Simple Exponential Smoothing predictions Sparkling wines](#f11) |
|  | [Figure 12:](#f12) [Double Exponential Smoothing Model forecast on the Rose wines](#f12) |
|  | [Figure 13](#f13): [Double Exponential Smoothing Model forecast on the Sparkling wines](#f13) |
|  | [Figure 14](#f14): [Triple Exponential Smoothing Model forecast on the Rose wines](#f14) |
|  | [Figure 15](#f23): [Triple Exponential Smoothing Model forecast on the Sparkling wines](#f14) |
|  | [Figure 16: Regression On Time Rose Data](#f16) |
|  | [Figure 17: Regression On Time Sparkling Data](#f17) |
|  | [Figure 18: Naive Forecast for Rose wines](#f18) |
|  | [Figure 19: Naive Forecast for Sparkling wines](#f19) |
|  | [Figure 20: Simple Average for Rose wines](#f20) |
|  | [Figure 21: Simple Average for Sparkling wines](#f21) |
|  | [Figure 22: Stationary check for Rose wines](#f22) |
|  | [Figure 23: Stationary check for Sparkling wines](#f23) |
|  | [Figure 24: Stationary check for Rose wines after 1 differencing](#f24) |
|  | [Figure 25: Stationary check for Sparkling wines after 1 differencing](#f25) |
|  | [Figure 26: Residual plot for Rose SARIMA\_6\_seasonal](#f26) |
|  | [Figure 27: Residual plot for Sparkling SARIMA\_6\_seasonal](#f27) |
|  | Figure 28: Residual for SARIMA\_12 rose wines |
|  | Figure 29: Residual for SARIMA\_12 sparkling wines |
|  | Figure 30: Residual for SARIMA rose wines best model |

**Problem:**

The data of different types of wine sales of Rose and Sparkling Wines in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century.

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

We first use the read\_csv function from pandas to read the CSV file Rose.csv, Sparkling.csv and set the first column as the index using the index\_col parameter. We also set parse\_dates=True to automatically parse the dates in the first column.

Next, we plot the time series data using matplotlib. We use plt.plot to create a line plot of the data, with the index as the x-axis and the column containing the data values as the y-axis. We also add labels to the x-axis and y-axis using plt.xlabel and plt.ylabel, and a title using plt.title. Finally, we use plt.show to display the plot.

Figure 1:Rose wines plot

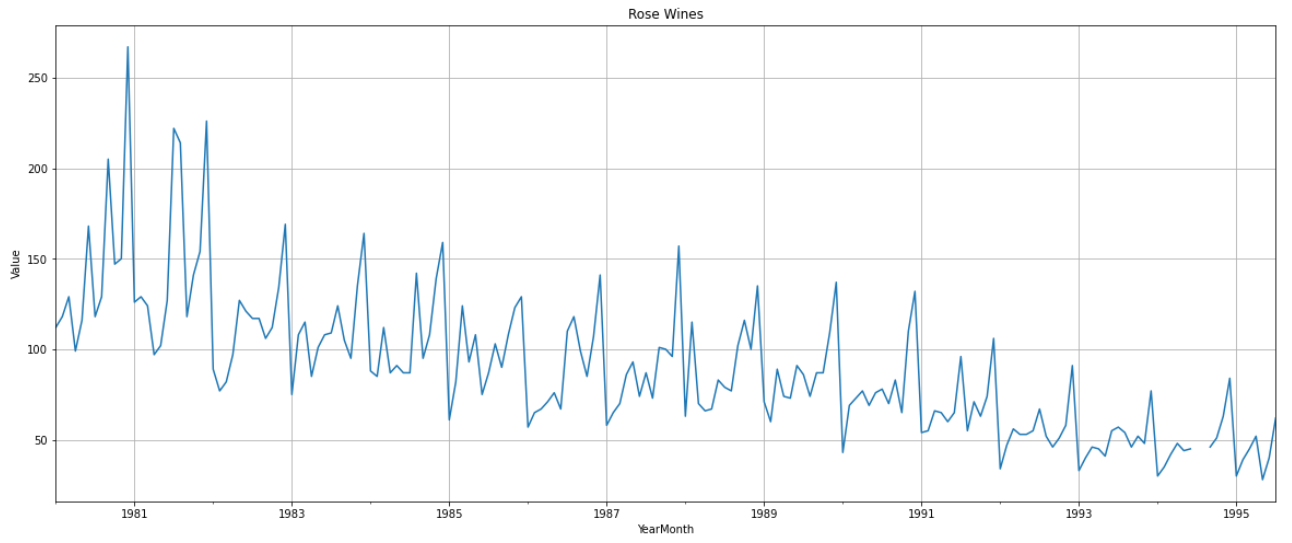
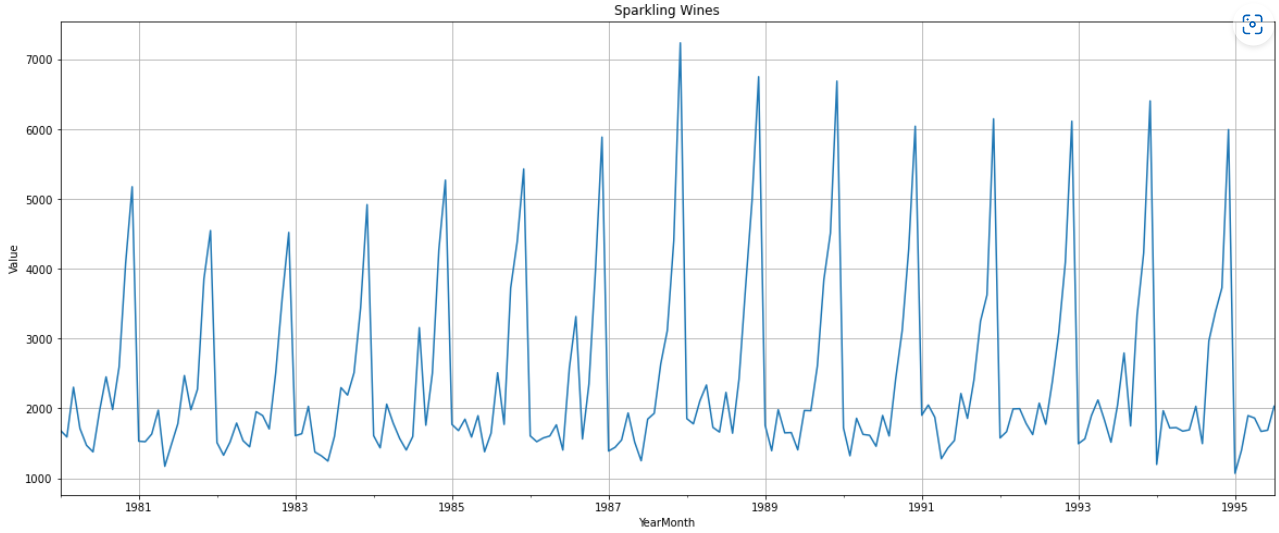


Figure 2: Sparkling Wines plot



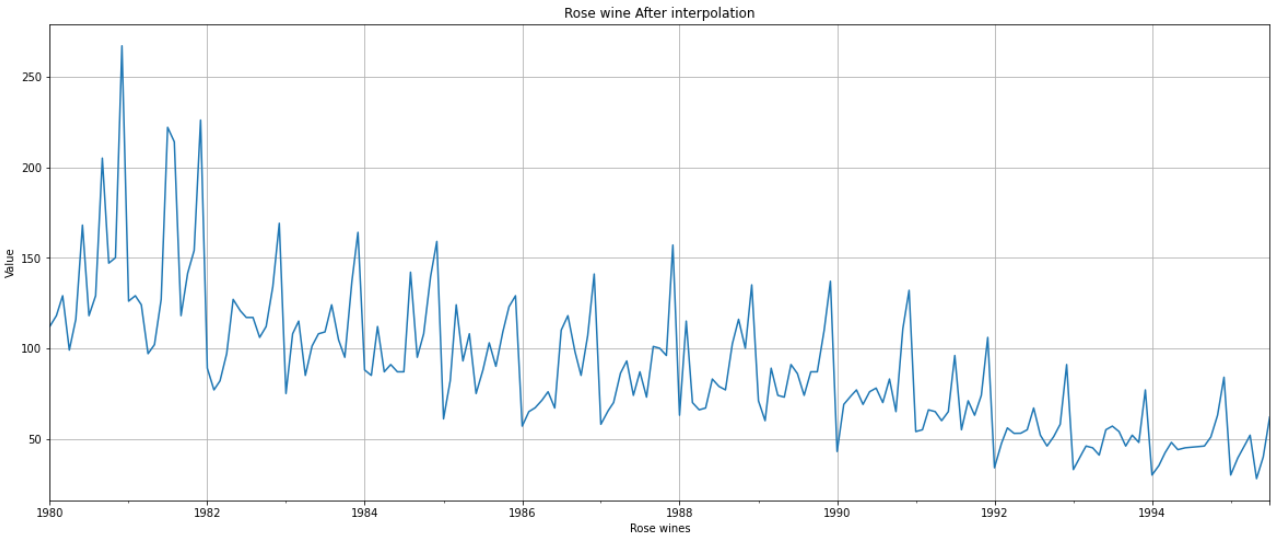
Handling Null Values:

Handling null values in time series data is important because they can affect the accuracy of forecasting models and statistical analysis.

When checked for null values and found there are two null values in the Rose wines data. Interpolation approach is used to handle this time series data

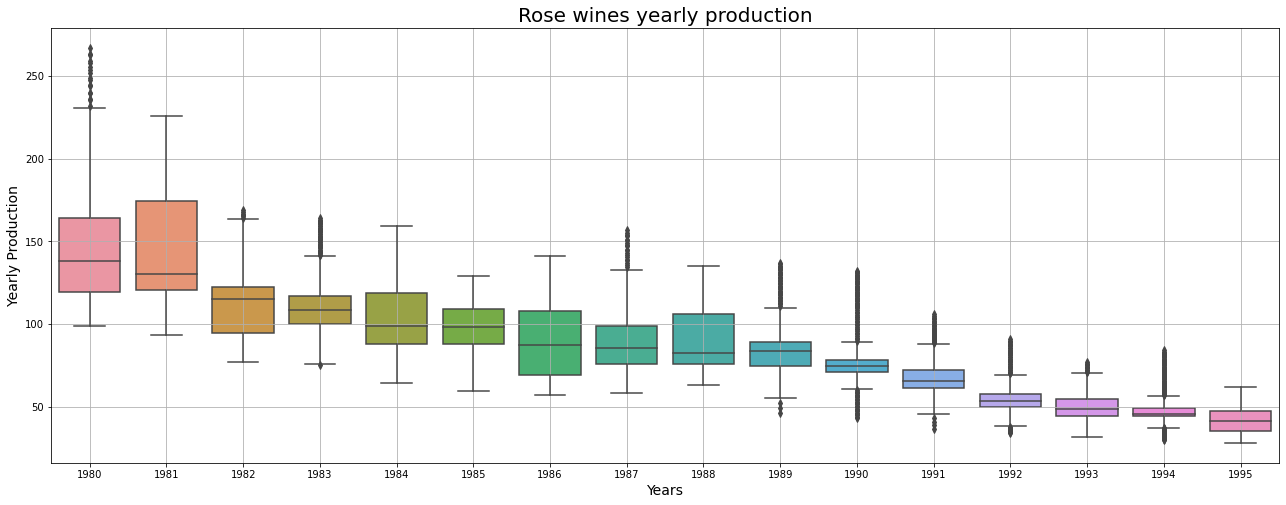
Interpolation is an estimation method based on the available data. There are various interpolation methods, such as linear interpolation, cubic spline interpolation, and nearest-neighbor interpolation. However, the choice of interpolation method depends on the nature of the data and the purpose of the analysis. Here we used a linear interpolation approach.

Figure 3: Rose wines plot after interpolation

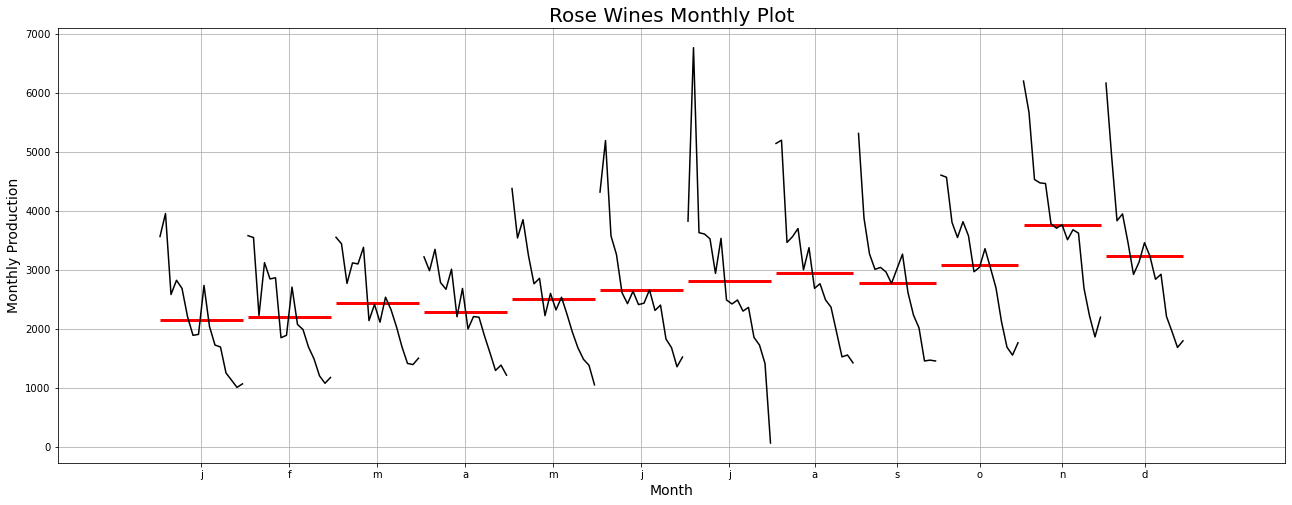


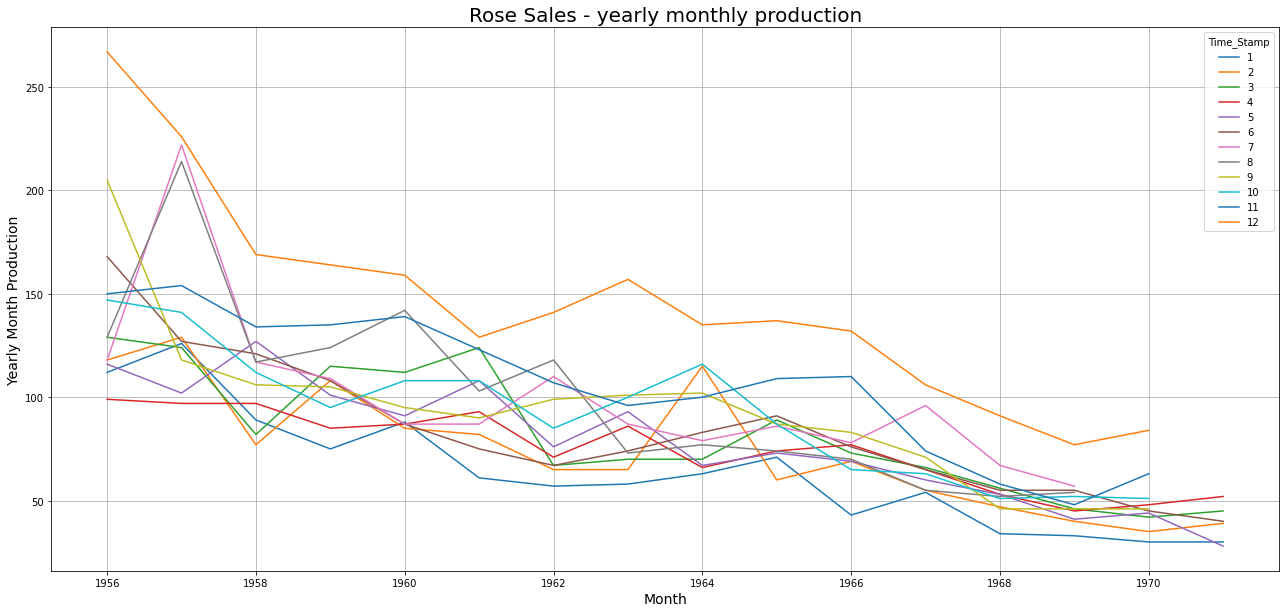
1. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition**

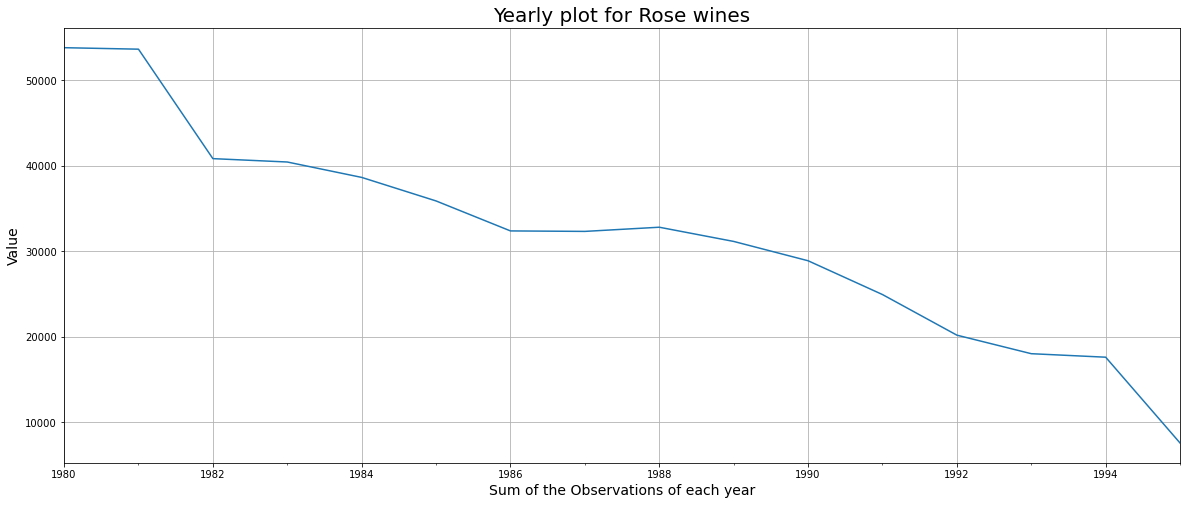
Figure 4:*ROSE WINES EDA:*







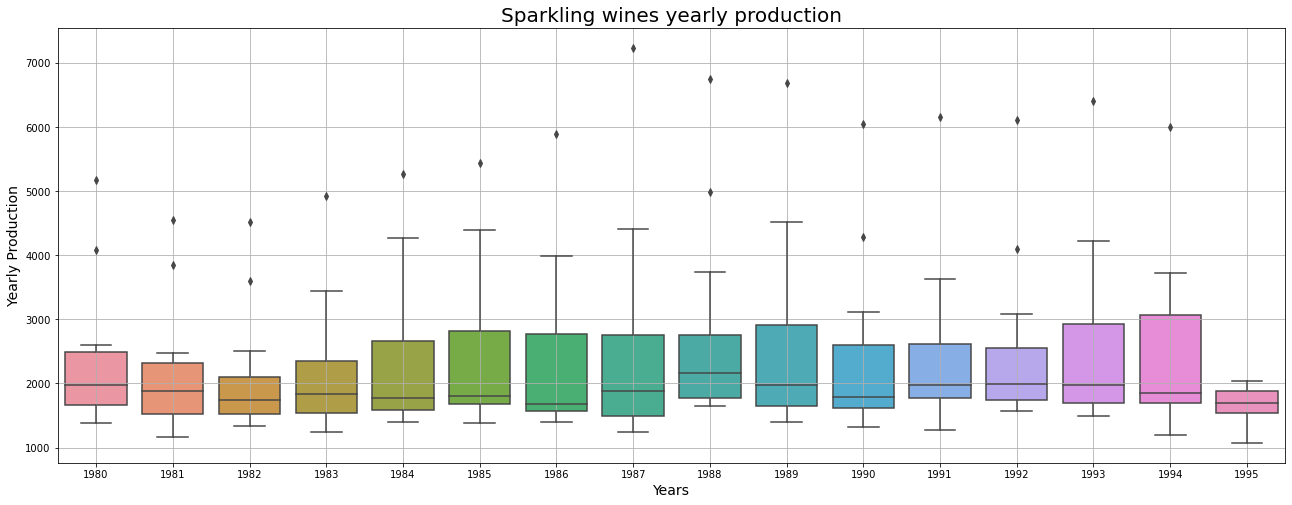


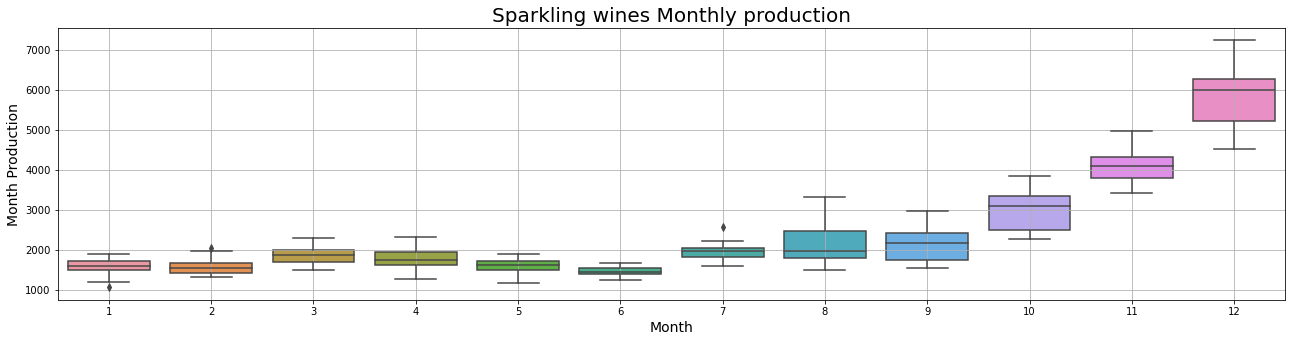


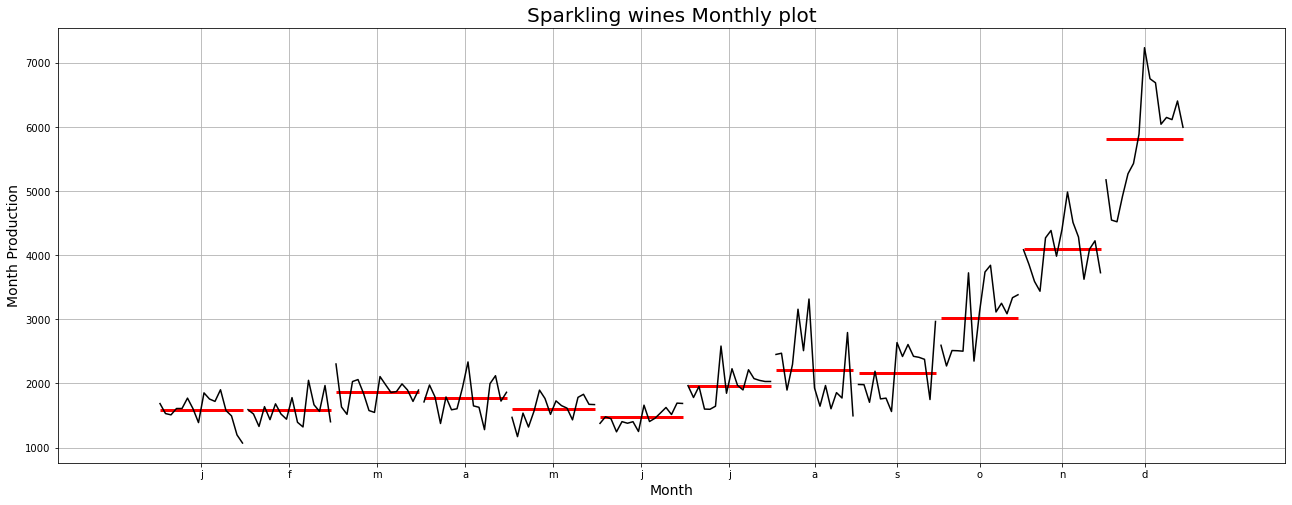
INSIGHTS:

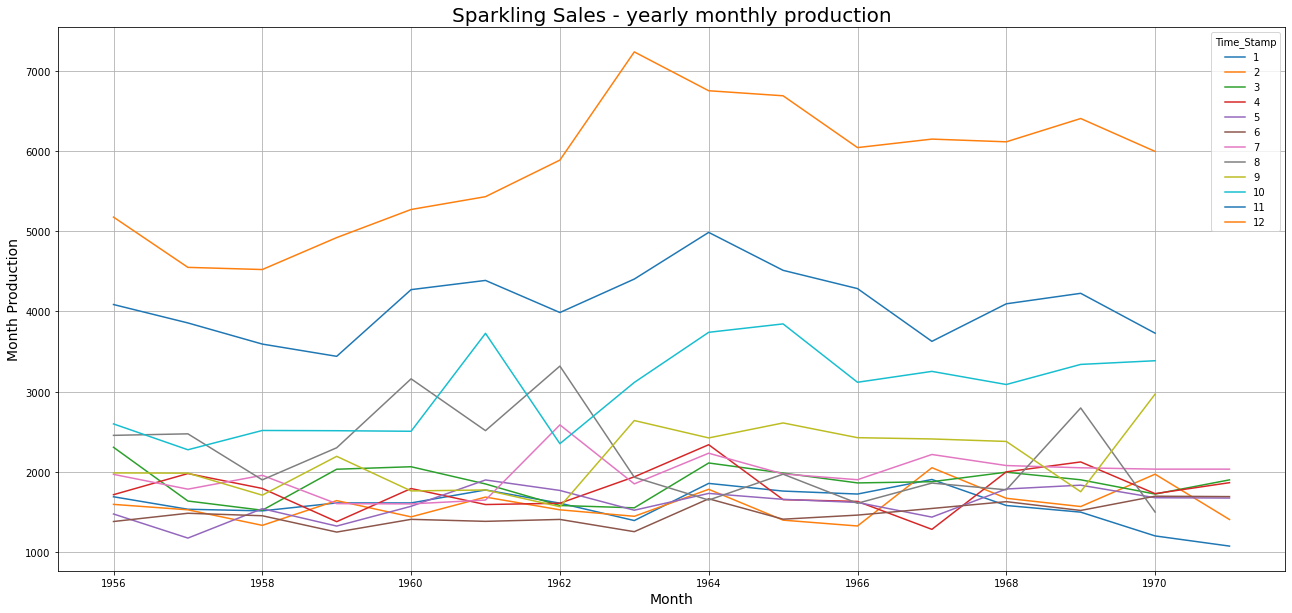
* 1. As we got to know from the Time Series plot, the boxplots over here also indicate a measure of the trend is present. Also, we see that the production of Rose wine has some outliers for certain years.
  2. The yearly boxplots also shows that the Sales have decreased towards the last few years.
  3. There is a clear distinction of 'Rose wine production' within different months spread across various years. The highest such numbers are being recorded in the month of June-August across various years.
  4. The month plot shows us the behavior of the Time Series 'Rose wine Sales' across various months. The red line is the median value.
  5. Through yearly monthly production December month have higher sales than other months across the years.

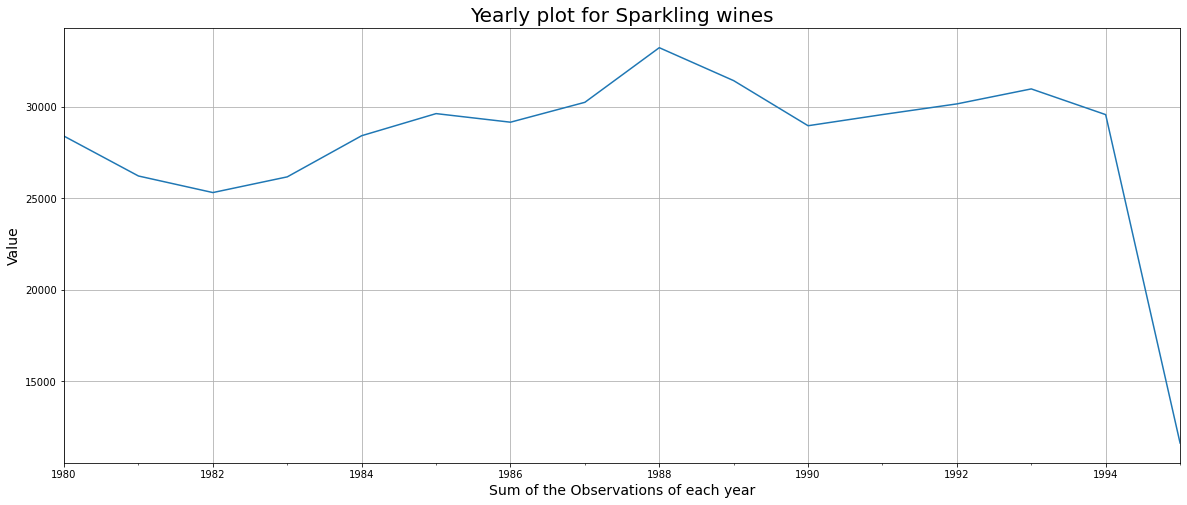
Figure 5: *SPARKLING WINES EDA*











INSIGHTS:

* 1. As we got to know from the Time Series plot, the boxplots over here also indicate a measure of the trend is present. Also, we see that the production of Rose wine has some outliers for certain years.
  2. The yearly boxplots also show that the Sales are not very much varying through years.
  3. There is a clear distinction between 'Sparkling wine production' within different months spread across various years. The highest such numbers are being recorded in the month of December across various years.
  4. The monthly plot shows there is an increase in the sales at end of the year.
  5. The month plot shows us the behavior of the Time Series 'Sparkling wine Sales' across various months. The red line is the median value.
  6. Through yearly monthly production December month have higher sales than other months across the years.

**DECOMPOSITION OF TIME SERIES**

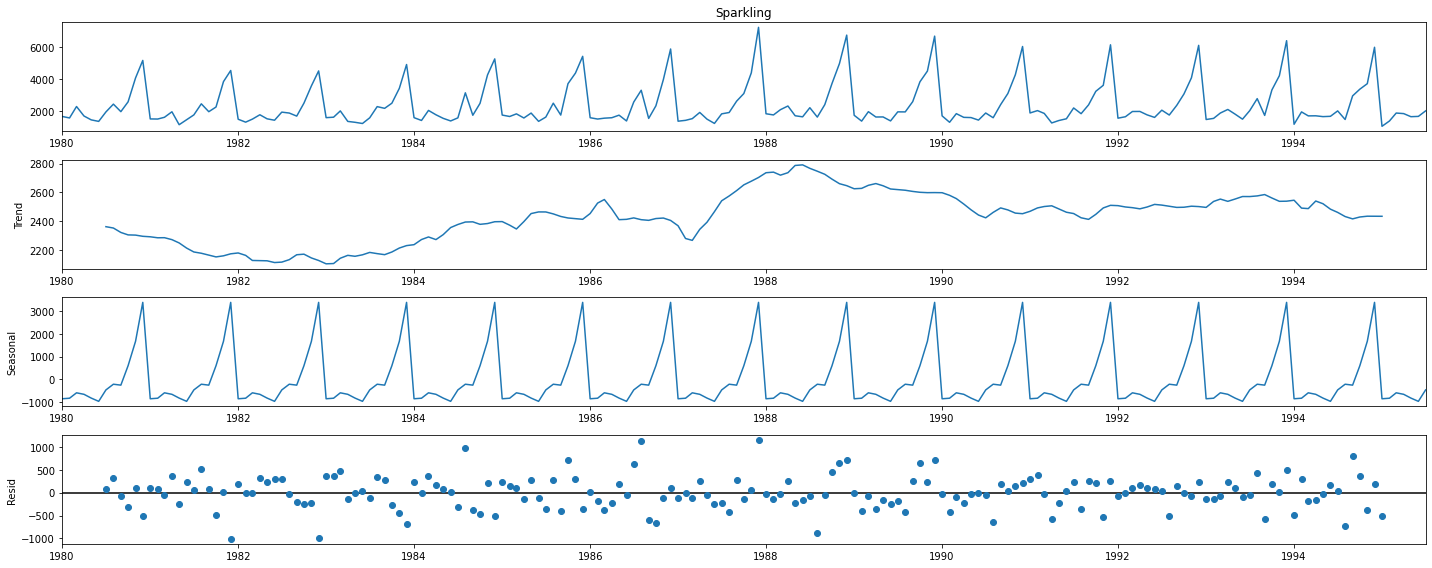
Decomposition is a statistical method used to decompose a time series into its underlying components, which typically include a trend, a seasonal component, and a residual component.

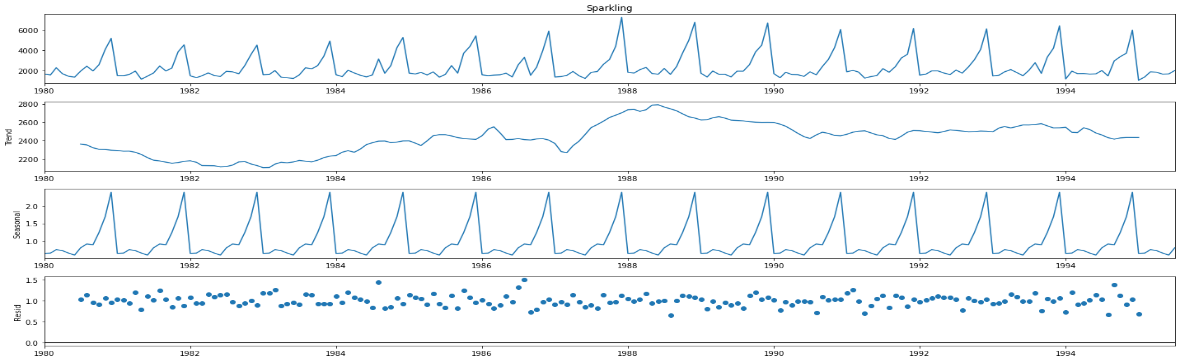
The *trend component* represents the long-term behavior of the time series

The *seasonal component* represents the seasonal patterns in the data

The *residual component* represents the random noise in the data that cannot be explained by the trend or seasonal patterns.

The purpose of decomposing a time series is to separate out these underlying components in order to better understand the patterns and trends in the data. This can be useful for analyzing and forecasting time series data, as well as identifying anomalies and outliers.

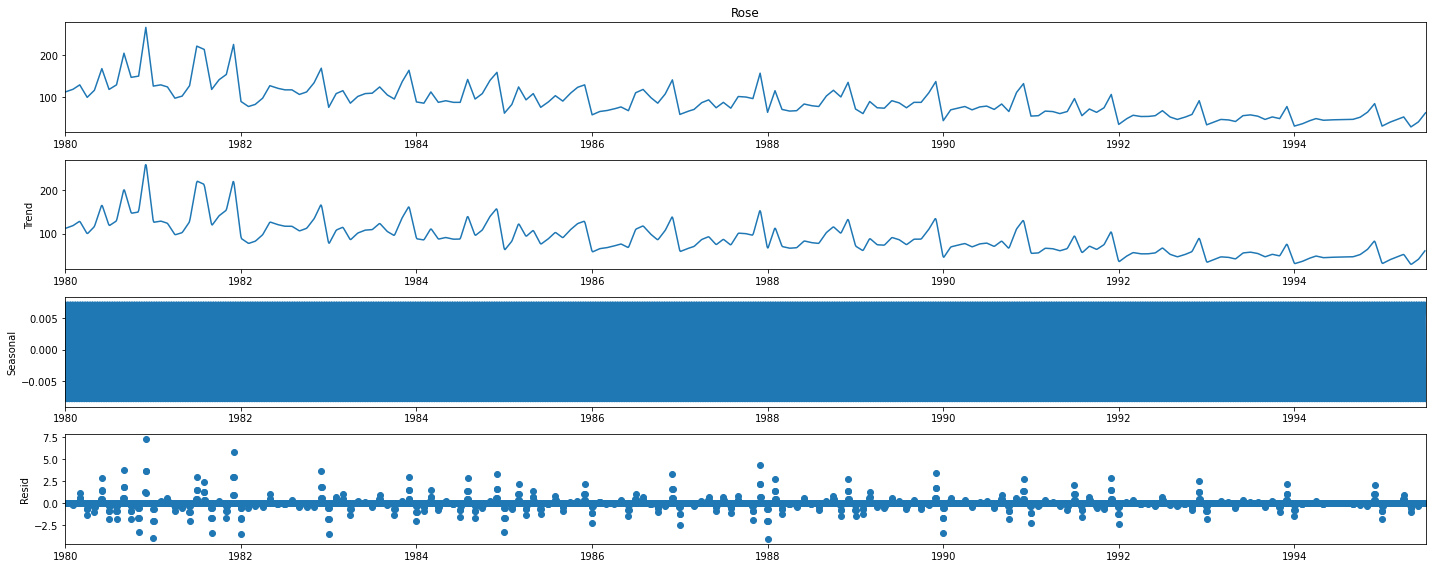
Figure 6: Decompose of Sparkling wines Additive and Multiplicative

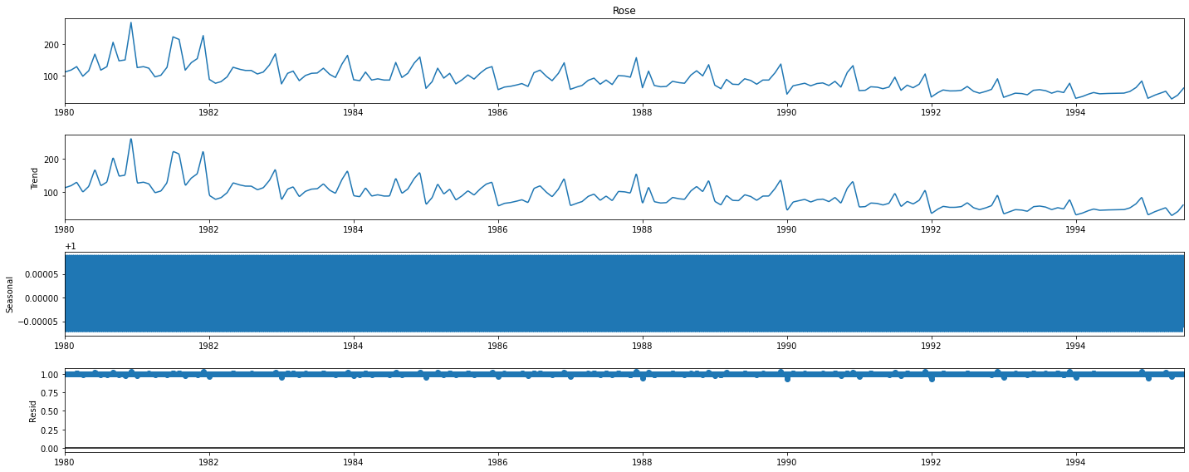


Insights:

1. We see that the residuals are located around 0 from the plot of the residuals in the decomposition for Addition
2. We see that the residuals are located around 1 from the plot of the residuals in the decomposition for Multiplicative
3. There is seasonality and some trend.
4. For the multiplicative series, we see that a lot of residuals are located around 1. Thus Multiplicative Decomposition is the right way to decompose the time series
5. Also, it is evident that there is a 6-month seasonality in the data from the above plots

Figure 7: Decompose of Rose wines Additive and Multiplicative



****

Insights:

1. If the seasonal component of a time series appears as a relatively constant or flat line, it indicates that the data does not exhibit any significant seasonal patterns
2. We see that the residuals are located around 0 from the plot of the residuals in the decomposition for Addition
3. We see that the residuals are located around 1 from the plot of the residuals in the decomposition for Multiplicative

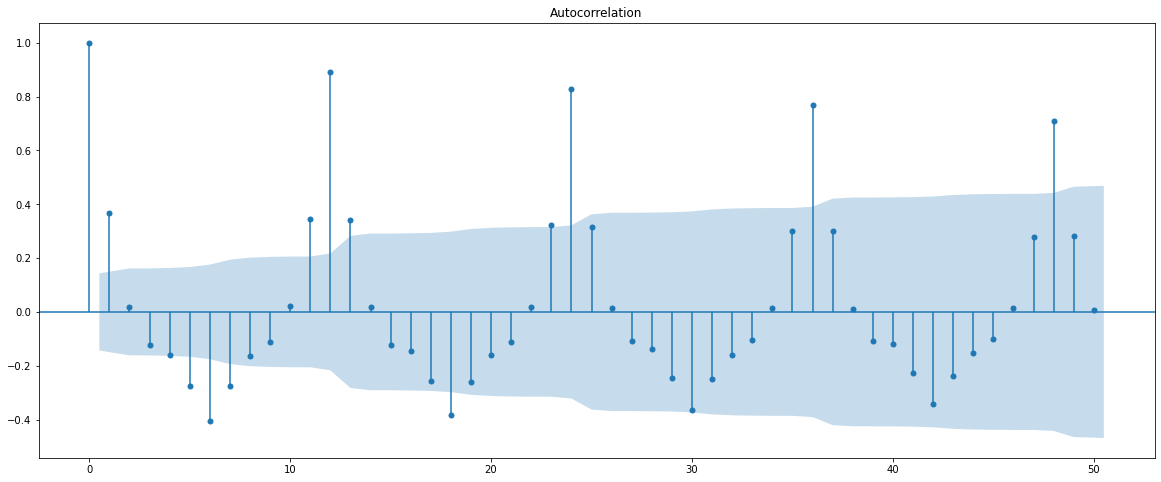
**AUTOCORRELATION:**

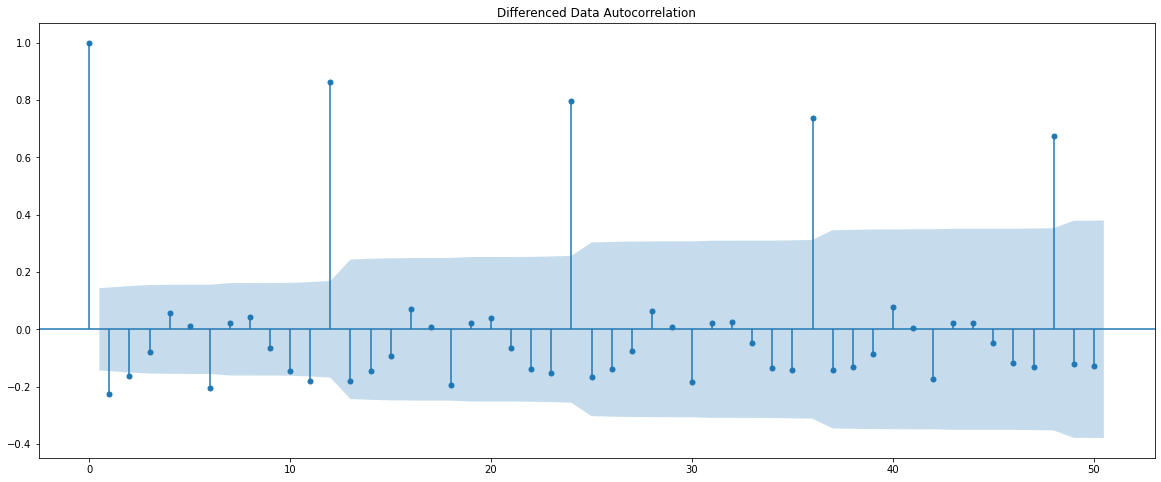
The ACF is a plot of the correlation between the time series and its lagged values. It is a useful tool for identifying patterns of autocorrelation in the data, which can help inform further analysis and modeling.

The differenced time series is obtained by taking the first difference of the original time series, which can be useful for removing trend and seasonality from the data.

The plot can be helpful for identifying patterns of autocorrelation in the residual component of the time series.

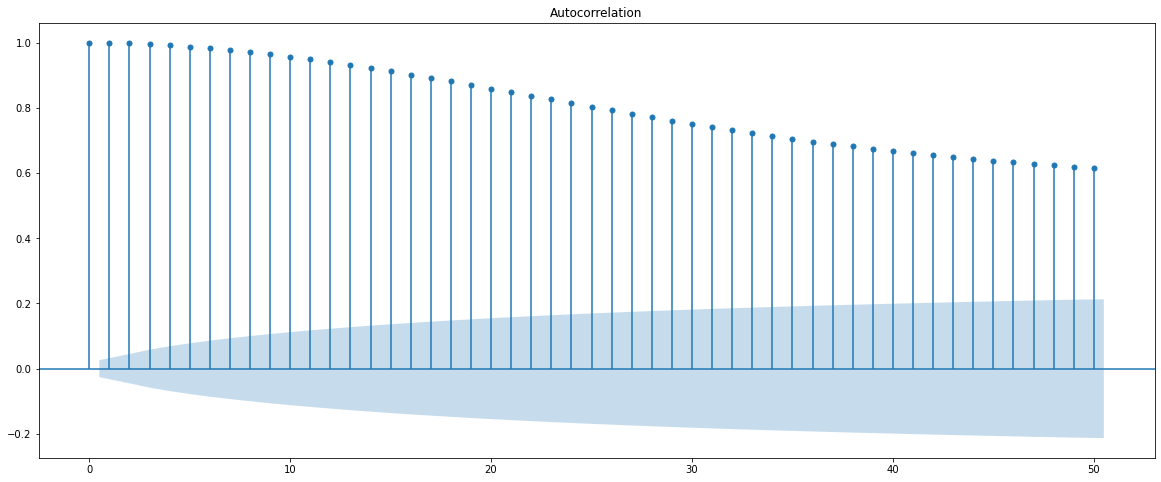
Figure 8: Autocorrelation plot for Sparkling wines

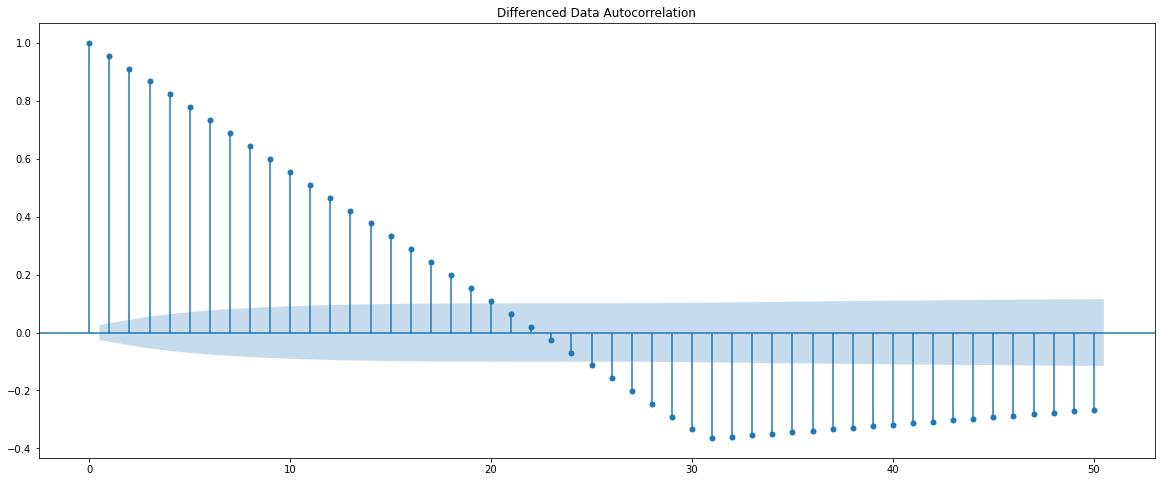




From this plot, we see that values for the ACF are within 95% confidence interval (represented by the solid gray line) for lags = 15, which verifies that our data have some autocorrelation.

Figure 9: Autocorrelation plot for Rose wines





Insights:

1. In the first graph, there are high positive correlations that only slowly decline a little with increasing lags. This indicates a lot of autocorrelation and you will need to take that into account in your modeling.
2. In second plot, older data has less impact than new ones. However as you can see here it is not always the case. Moreover the ACF function drops below zero.
3. A negative autocorrelation implies that if a past value is above average the newer value is more likely to be below average.

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

Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

By splitting the time series into a training set and a test set, you can evaluate the performance of a forecasting model by comparing its predictions for the test set to the actual values in the test set. The training set is used to estimate the parameters of the model, while the test set is used to evaluate its predictive accuracy.

It is important to choose an appropriate split point for the time series based on the specific context of the analysis.

Train set: The year of the index is less than 1991, and assigns them to the variable train. This creates a training set consisting of the earlier part of the time series.

Test set: The year of the index is greater than or equal to 1990, and assigns to the variable test. This creates a test set consisting of the later part of the time series.

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

## **Building different models and comparing the accuracy metrics**

**Exponential Smoothing Models:**

Simple Exponential Smoothing (SES) is a time series forecasting method that uses a weighted average of past observations to make future predictions.

The basic idea behind SES is to forecast the next value in the time series as a weighted average of the previous observations, where the weights assigned to each observation are determined by a smoothing parameter, typically denoted by the symbol alpha (α).

F(t+1) = α \* Y(t) + (1-α) \* F(t)

where:

F(t+1) is the forecast for the next period (t+1)

Y(t) is the actual value of the time series at period t

F(t) is the forecast for the current period (t)

α is the smoothing parameter, with values between 0 and 1.

First, an instance of SimpleExpSmoothing is instantiated and passed the training data. Next, the fit() function is called, giving the fit configuration, especially the alpha value. The fit() function returns an instance of the HoltWintersResults class containing the learned coefficients. The forecast() or the predict() function is then called on the result object to make a forecast

The Prediction for Rose wines is 56.5 and the Prediction for Rose wines is 2724.9

**For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Rose wines Data,**

**RMSE is 13.719**

**For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Sparkling wines Data,**

**RMSE is 1316.035**

Figure 10: Alpha =0.995 Simple Exponential Smoothing predictions Rose wines

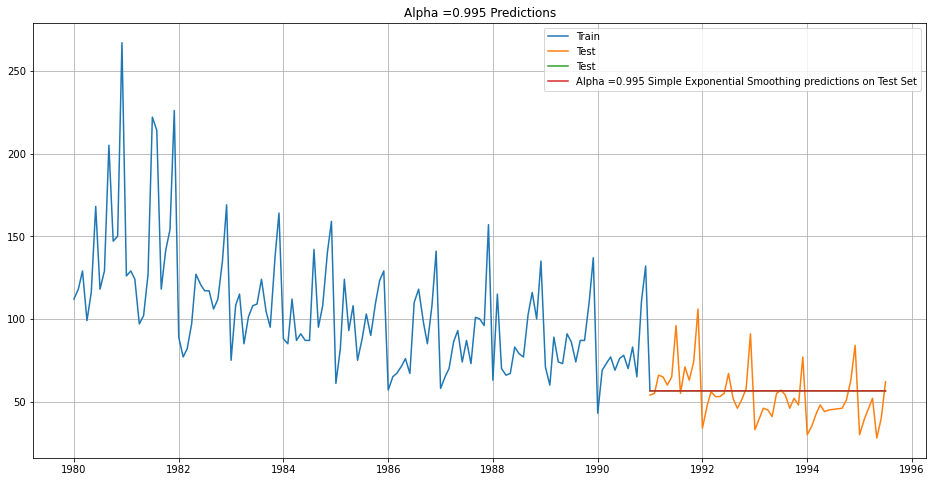
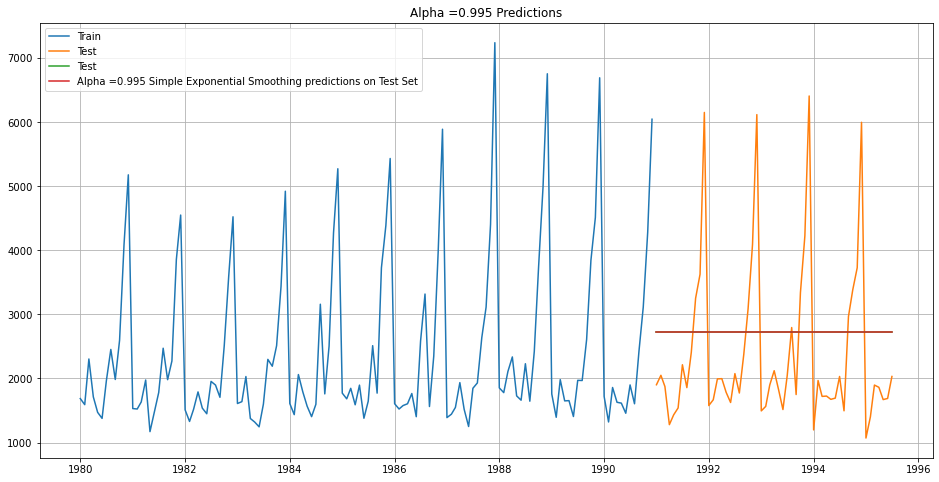


Figure 11: Alpha =0.995 Simple Exponential Smoothing predictions Sparkling wines



**Double Exponential Smoothing:**

Double Exponential Smoothing, also known as Holt's method, is a time series forecasting method that is an extension of Simple Exponential Smoothing (SES).

The basic idea behind Double Exponential Smoothing is to forecast the next value in the time series as a weighted average of the previous observations and their trends, where the weights assigned to each observation and trend are determined by two smoothing parameters, typically denoted by the symbols alpha (α) and beta (β).

F(t+1) = α \* Y(t) + (1-α) \* (L(t) + T(t))

T(t+1) = β \* (L(t+1) - L(t)) + (1-β) \* T(t)

L(t+1) = α \* Y(t+1) + (1-α) \* L(t) + β \* (L(t+1) - L(t))

where:

F(t+1) is the forecast for the next period (t+1)

Y(t) is the actual value of the time series at period t

L(t) is the estimated level of the time series at period t

T(t) is the estimated trend of the time series at period t

α is the level smoothing parameter, with values between 0 and 1.

β is the trend smoothing parameter, with values between 0 and 1.

**For Alpha =0.1, Beta =0.1 Double Exponential Smoothing Model forecast on the Rose wines Test Data, RMSE is 35.91**

**For Alpha =0.1, Beta =0.1 Double Exponential Smoothing Model forecast on the Sparkling wines Test Data, RMSE is 1916.34**

Figure 12: Double Exponential Smoothing Model forecast on the Rose wines

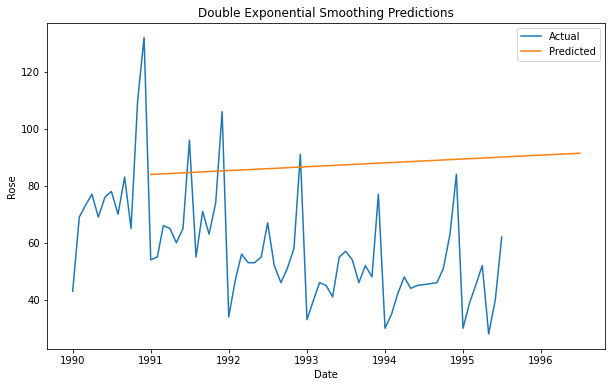
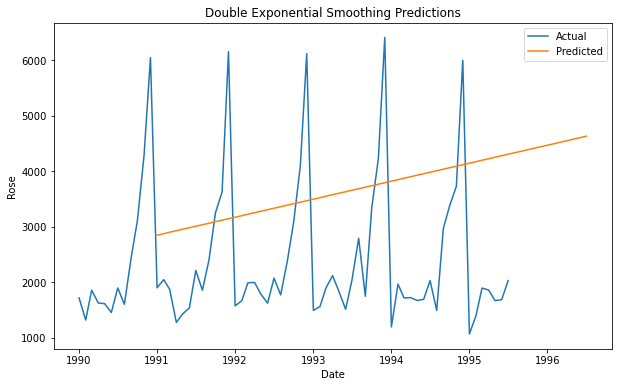


Figure 13: Double Exponential Smoothing Model forecast on the Sparkling wines



**Triple Exponential Smoothing:**

Triple Exponential Smoothing, also known as Holt-Winters method, is an extension of Double Exponential Smoothing that incorporates seasonality into the forecasting model. Like Double Exponential Smoothing, Triple Exponential Smoothing estimates the level and trend of the time series, but it also adds a seasonal component to capture periodic variations in the data.

F(t+1) = L(t) + T(t) + S(t+m)

L(t) = α \* (Y(t) - S(t-m)) + (1 - α) \* (L(t-1) + T(t-1))

T(t) = β \* (L(t) - L(t-1)) + (1 - β) \* T(t-1)

S(t) = γ \* (Y(t) - L(t)) + (1 - γ) \* S(t-m)

F(t+1) is the forecast for the next period (t+1)

Y(t) is the actual value of the time series at period t

L(t) is the estimated level of the time series at period t

T(t) is the estimated trend of the time series at period t

S(t) is the estimated seasonal component of the time series at period t

α, β, and γ are the smoothing parameters for level, trend, and seasonality, respectively, with values between 0 and 1.

m is the number of periods in each season.

The parameters for Rose wines is as follows:

|  |  |
| --- | --- |
| 'smoothing\_level' | 0.06571007449183297, |
| 'smoothing\_trend' | 0.051867105713176015, |
| 'smoothing\_seasonal' | 0.0015637515713898, |
| 'damping\_trend' | nan, |
| 'initial\_level' | 47.81887301367471, |
| 'initial\_trend' | -0.2961562797665537, |
| 'initial\_seasons' | array([2.35763018, 2.67367218, 2.92146068, 2.55308191, 2.87099548, 3.13124987, 3.44178442, 3.66118656, 3.47154364, 3.39670325, 3.95879831, 5.46173463]), |
| 'use\_boxcox' | False, |
| 'lamda' | None, |
| 'remove\_bias' | False |

The parameters for Sparkling wines is as follows:

|  |  |
| --- | --- |
| 'smoothing\_level' | 0.11057044018305404, |
| 'smoothing\_trend' | 0.06076609768412894, |
| 'smoothing\_seasonal' | 0.39187601902826213, |
| 'damping\_trend' | nan, |
| 'initial\_level' | 1621.806699459997, |
| 'initial\_trend' | -8.175193580026457, |
| 'initial\_seasons' | array([1.07304448, 1.02730389, 1.39469706, 1.20333193, 0.98008967, 0.97664181, 1.39261648, 1.70888984, 1.37519684, 1.81953245, 2.82880203, 3.59142504]), |
| 'use\_boxcox' | False, |
| 'lamda' | None, |
| 'remove\_bias' | False |

Figure 14: Triple Exponential Smoothing Model forecast on the Rose wines

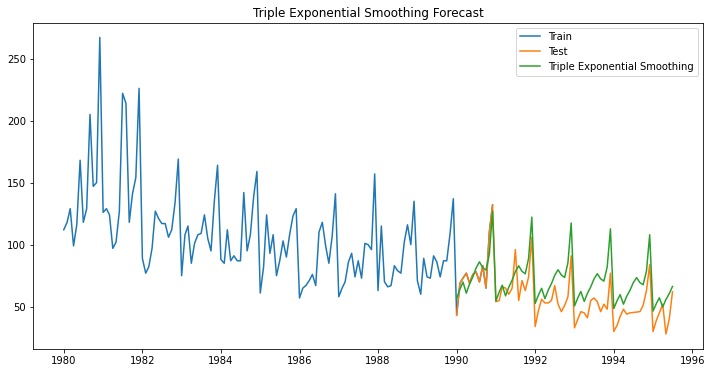
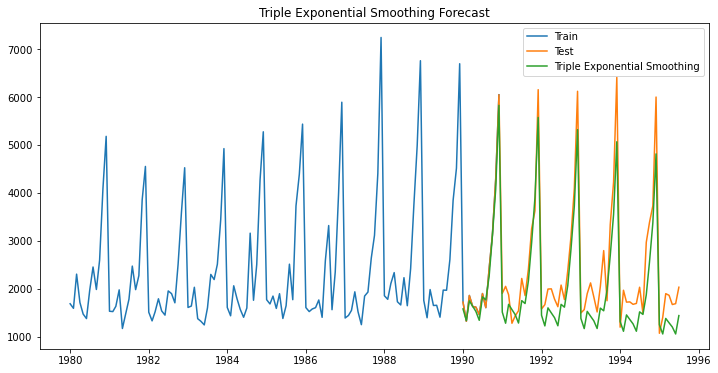
****

Figure 15: Triple Exponential Smoothing Model forecast on the Sparkling wines



**LINEAR REGRESSION:**

* For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.
* We should create time indices for the training and test sets, and adds them as a new column to the respective DataFrames.
* As there are 132 training data, the test set starts at index 133 in the original DataFrame
* This creates a new column called 'time' and populates it with the corresponding time index values.
* Next fit the linear regression model using the fit() method from sci-kit-learn’s LinearRegression class. The independent variable for the model is the 'time' column of the LinearRegression\_train DataFrame, and the dependent variable is the 'Rose' and 'Sparkling' columns.
* **For the Regression on Time forecast on the Rose wines data, RMSE is 16.817**
* **For the Regression on Time forecast on the Sparkling wines data, RMSE is 1416.381**

Figure 16: Regression On Time Rose Data

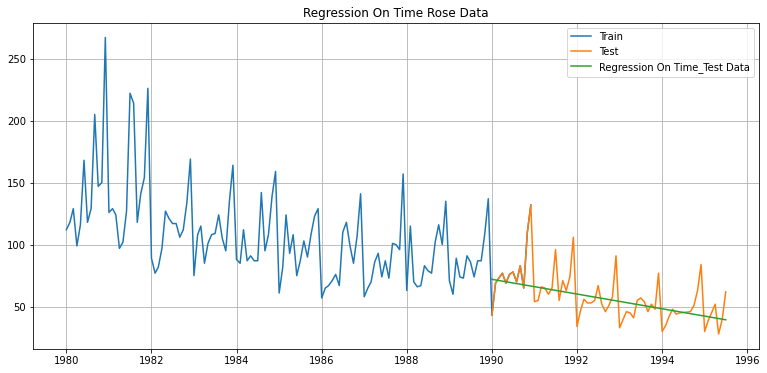
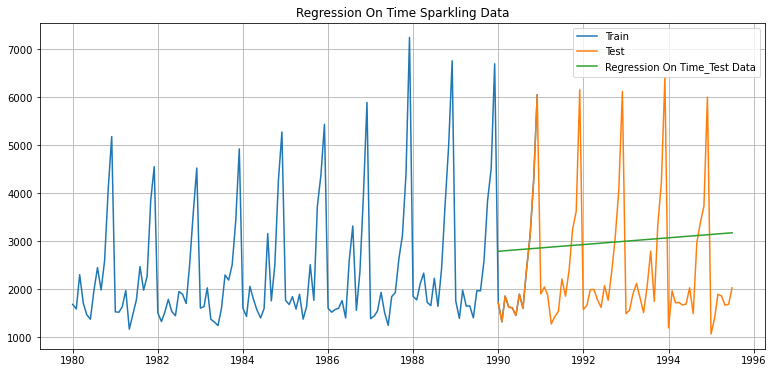


Figure 17: Regression On Time Sparkling Data



**NAÏVE Approach:**

* This is a simple baseline model that assumes the time series will not change from the last observed value in the training set. While it can be useful for comparison with more complex models, it may not be suitable for all time series and may not provide accurate forecasts in many cases.
* In this approach, a new column called 'naive' to the NaiveModel\_test DataFrame, which contains the naive forecast for the 'Rose' time series. The naive forecast is simply the last observed value from the training set, which is used as a prediction for all future time periods in the test set.
* Next, convert the 'Rose' column of the train DataFrame to a numpy array, and then extracts the last value from this array. This value is then set as the value of the 'naive' column for all rows.
* Here the naïve values is set to 132 for Rose wines and 6047 for Sparkling wines.
* **For the Naive approach forecast on the Test Data, RMSE is 5988.718**
* **For the Naive approach forecast on the Test Data, RMSE is 3866.980**

Figure 18: Naive Forecast for Rose wines

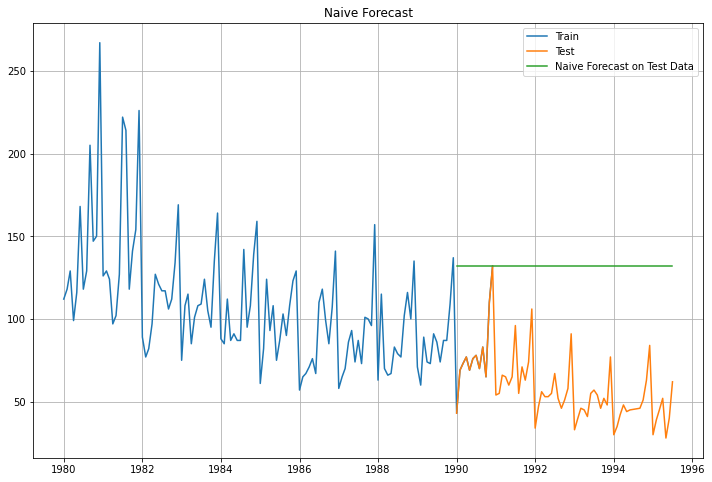
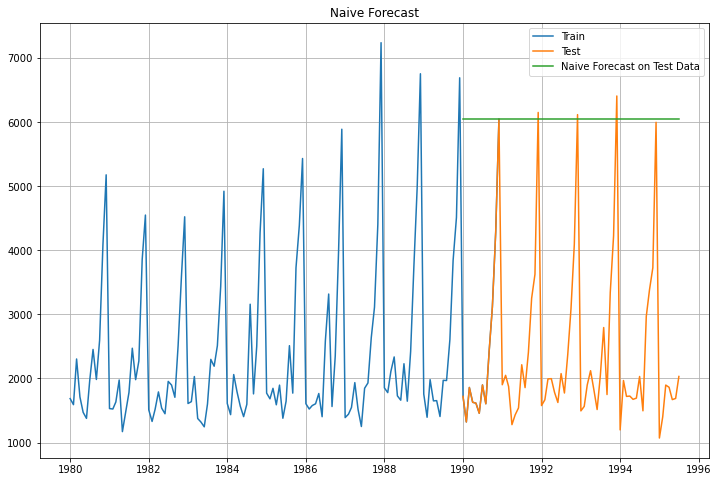


Figure 19: Naive Forecast for Sparkling wines

****

**Simple Average:**

* Simple average forecasting is a basic technique used to forecast time series data. It involves calculating the average value of a given time series over a specific period and using this average value to forecast future values of the time series.
* To use the simple average method for forecasting, the following steps can be taken:
* Split the time series data into a training set and a test set. The training set should contain historical data up until a 1991 in time, while the test set should contain data for the remaining time period that is after 1991.
* Calculate the average of the time series in the training set.Use this average value as the forecast for all future time periods in the test set.
* Calculate the forecast errors by comparing the forecasted values to the actual values in the test set.
* The forecast value for Rose wines is 104.9 and for Sparkling wines is 2403.78
* **For Simple Average forecast on the Rose wines Data, RMSE is 2345.548**
* **For Simple Average forecast on the Sparkling wines Data, RMSE is 2345.548**

Figure 20: Simple Average for Rose wines

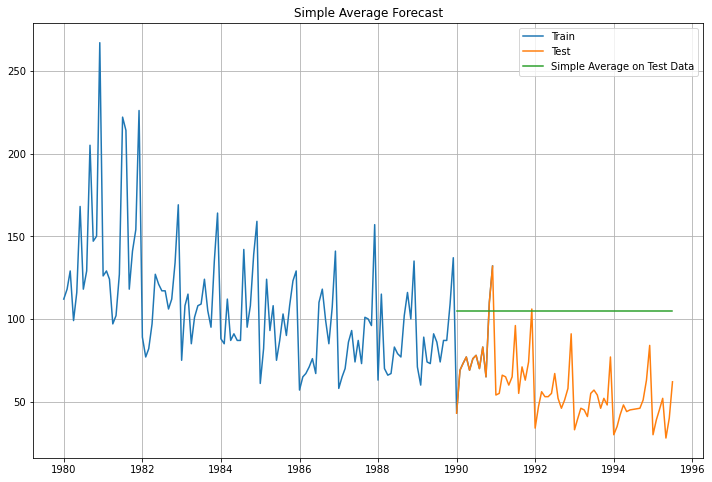
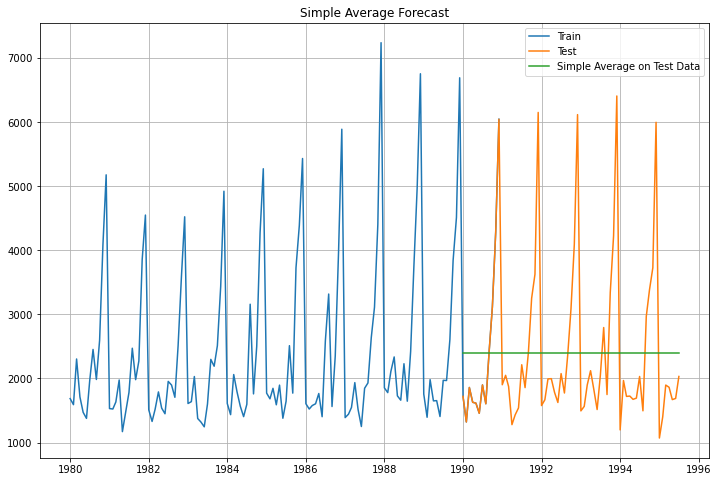


Figure 21: Simple Average for Sparkling wines



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

Stationarity is an important assumption for many time series modeling techniques. A stationary time series is one whose statistical properties, such as the mean and variance, remain constant over time. In contrast, a non-stationary time series may exhibit trends, seasonality, or other patterns that change over time.

To check for the stationarity of a time series, we can perform statistical tests such as the Augmented Dickey-Fuller (ADF) test or the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. These tests can help determine whether a time series is stationary or not and can provide insights into how to make the data stationary.

**The Null hypothesis (H0) of the ADF test is that the time series is non-stationary**

**The Alternative hypothesis (Ha) is that the time series is stationary**

The test statistic is compared to a critical value at a given significance level (alpha) to determine whether the null hypothesis should be rejected or not. If the p-value of the test statistic is less than alpha, we reject the null hypothesis and conclude that the time series is stationary. Otherwise, we fail to reject the null hypothesis and conclude that the time series is non-stationary.

Figure 22: Stationary check for Rose wines

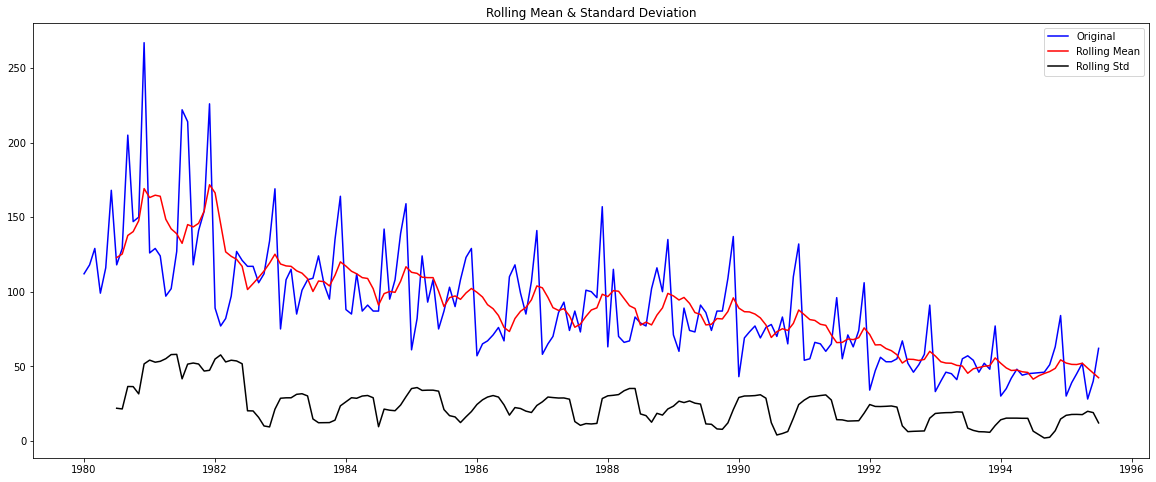


Table 1: Results of ADF for Rose wines

|  |
| --- |
| Results of Dickey-Fuller Test: |
| Test Statistic -1.876691 |
| p-value 0.343105 |
| #Lags Used 13.000000 |
| Number of Observations Used 173.000000 |
| Critical Value (1%) -3.468726 |
| Critical Value (5%) -2.878396 |
| Critical Value (10%) -2.575756 |
| dtype: float64 |

Figure 23: Stationary check for Sparkling wines

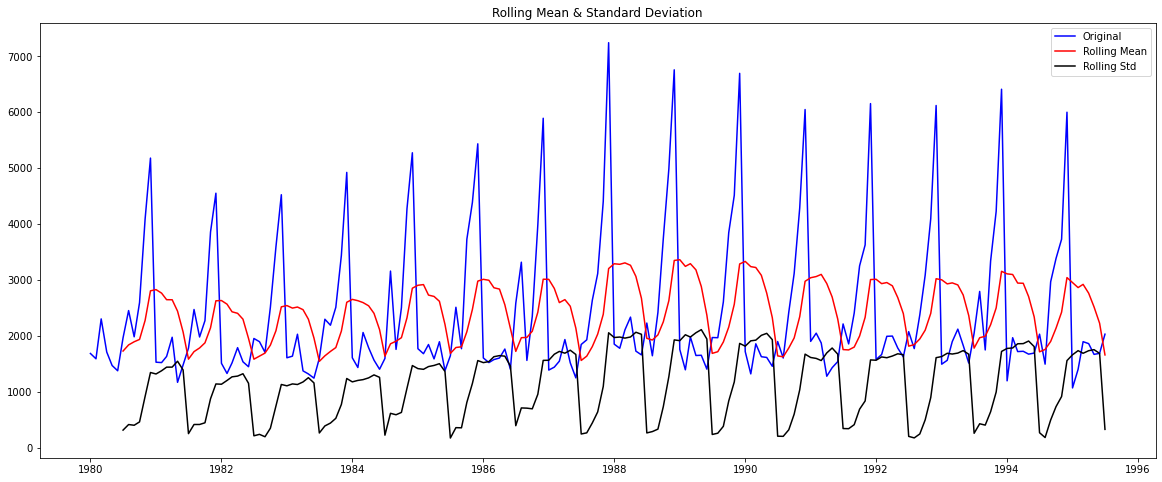


Table 2: Results of ADF for Sparkling wines

**Results of Dickey-Fuller Test for Sparkling wines:**

|  |  |
| --- | --- |
| Test Statistic | -1.360497 |
| p-value | 0.601061 |
| #Lags Used | 11.000000 |
| Number of Observations Used | 175.000000 |
| Critical Value (1%) | -3.468280 |
| Critical Value (5%) | -2.878202 |
| Critical Value (10%) | -2.575653 |
| dtype: float64 |  |

From the above check for stationary, we could see p-value of the statistical test for both the wines is greater than the chosen significance level of 0.05.

we fail to reject the null hypothesis and conclude that the time series is non-stationary. This means that the statistical properties of the time series, such as the mean and variance, are not constant over time, and there may be trends or patterns present that change over time.

Here we are using differencing to remove trends or seasonality from the data. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

Figure 24: Stationary check for Rose wines after 1 differencing

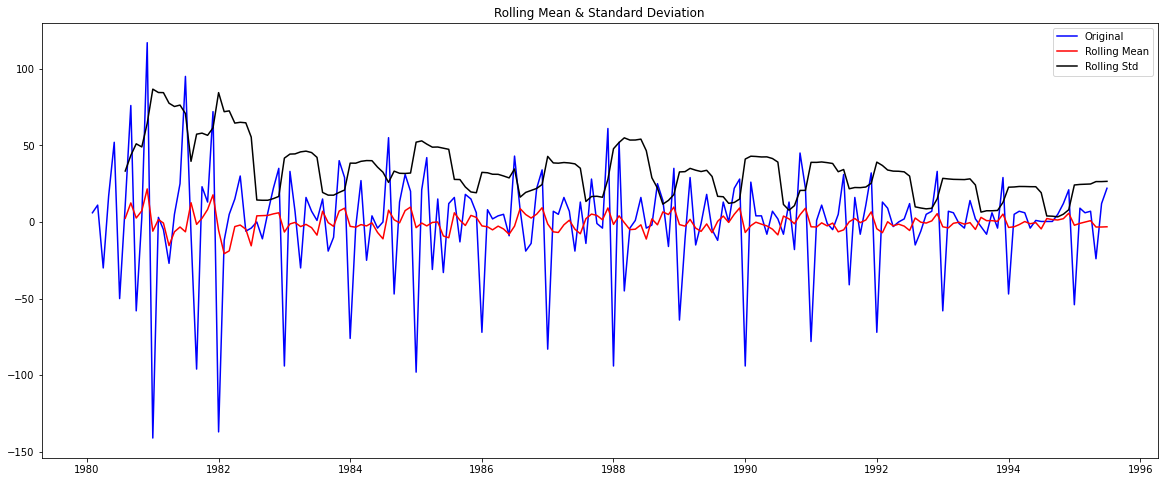


Table 3: Results of Dickey-Fuller Test after difference for Rose wines:

|  |  |
| --- | --- |
| Test Statistic -8.044389e+00 |  |
| p-value 1.810924e-12 |  |
| #Lags Used 1.200000e+01 |  |
| Number of Observations Used 1.730000e+02 |  |
| Critical Value (1%) -3.468726e+00 |  |
| Critical Value (5%) -2.878396e+00 |  |
| Critical Value (10%) -2.575756e+00 |  |
| dtype: float64 |  |

Figure 25: Stationary check for Sparkling wines after 1 differencing

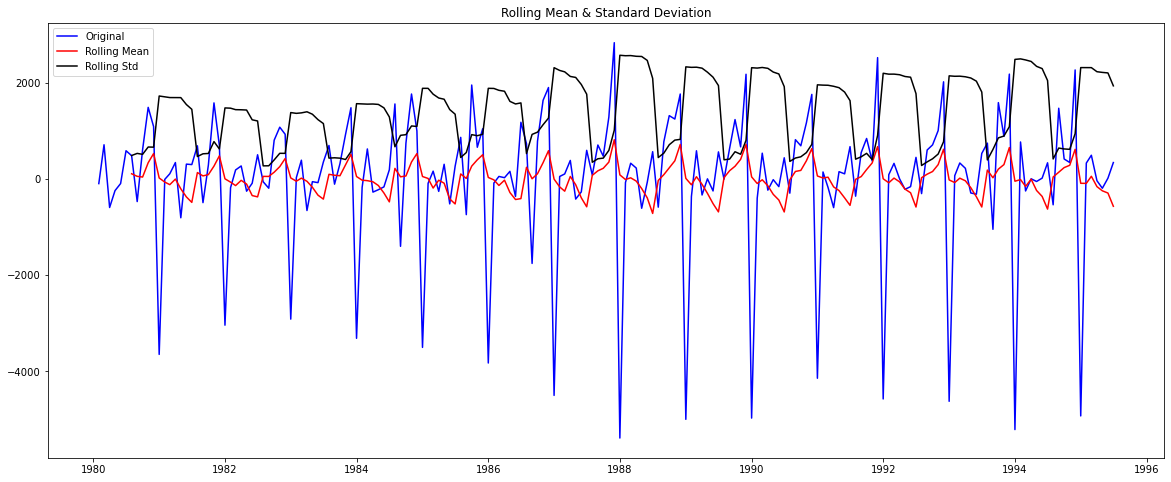


Table 4: Results of Dickey-Fuller Test after difference for Sparkling wines

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

Here, after 1 differencing the P-values are smaller, it means that the null hypothesis can be rejected with high confidence. This suggests that the time series is likely stationary and the statistical properties, such as the mean and variance, are constant over time. In this case, we can proceed to build time series models on the data without the need for further preprocessing steps to make it 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.

**ARIMA (Autoregressive Integrated Moving Average) model:**

* ARIMA is a popular and powerful time series forecasting model that combines the concepts of autoregression, differencing, and moving average to model the linear relationship between past values and future values of a time series. It involves three components:
* Autoregression (AR) component: This component uses the past values of the time series to predict future values. It assumes that the future value of the time series is a linear combination of its past values.
* Integrated (I) component: This component involves differencing the time series to make it stationary. Differencing means computing the difference between consecutive values of the time series.
* Moving Average (MA) component: This component uses the errors of the predictions made by the autoregressive component to make future predictions. It assumes that the errors are normally distributed and have zero mean and constant variance.
* The ARIMA model is usually denoted as ARIMA(p,d,q), where p is the order of the autoregressive component, d is the degree of differencing, and q is the order of the moving average component.

Table 5: Sorted ARIMA AIC values for Rose wines

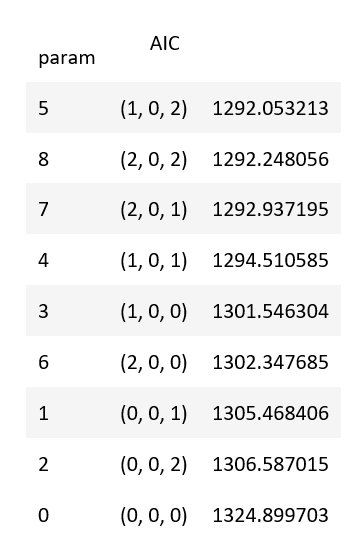


Table 6: Summay of ARIMA for Rose wines

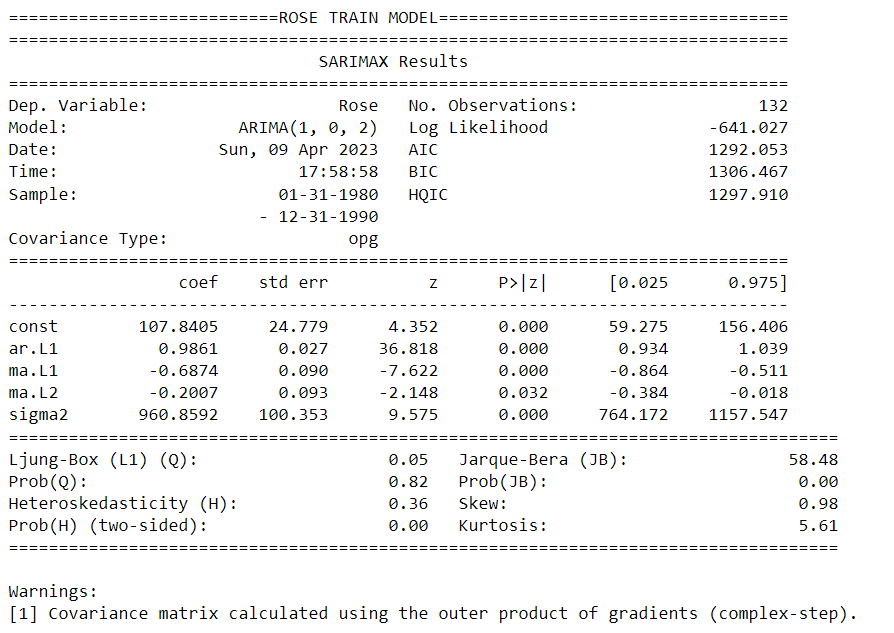


Table 7: Sorted ARIMA AIC values for Rose wines

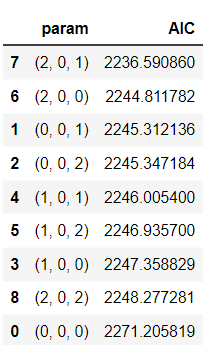
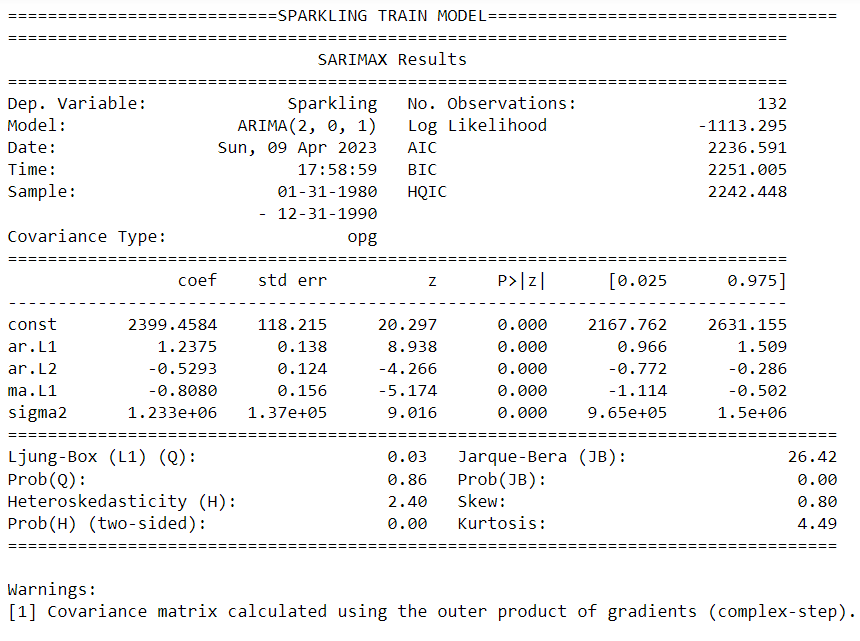


Table 8: Summary of ARIMA for Rose wines



The p-value represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming that the null hypothesis is true.

In this case, the null hypothesis is that the coefficients of the model (including the intercept and the autoregressive and moving average terms) are equal to zero.

The p-value is less than the significance level (usually set to 0.05), So we can reject the null hypothesis and conclude that the corresponding coefficient is statistically significant, meaning that it has a significant effect on the dependent variable. In the summary output you provided, all the coefficients have p-values less than 0.05, which indicates that they are statistically significant.

**The RMSE values for ARIMA model on Rose wines is 43.96**

**The RMSE values for ARIMA model on Sparkling wines is 1291.6**

**SARIMA (Seasonal Autoregressive Integrated Moving Average):**

SARIMA is an extension of the ARIMA model that incorporates seasonal factors in addition to the autoregressive, integrated, and moving average terms. It is used for time series forecasting when the data exhibits seasonality.

A SARIMA model is defined by three parameters: (p,d,q)x(P,D,Q)s, where:

p: the order of the autoregressive (AR) term

d: the degree of differencing required to make the time series stationary

q: the order of the moving average (MA) term

P: the order of the seasonal autoregressive (SAR) term

D: the degree of seasonal differencing required to make the time series stationary

Q: the order of the seasonal moving average (SMA) term

s: the number of time steps in each season

The model parameters can be determined using statistical methods such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Once the parameters are estimated, the model can be used to make forecasts.

Table 9: Sorted SARIMA AIC values for Rose wines

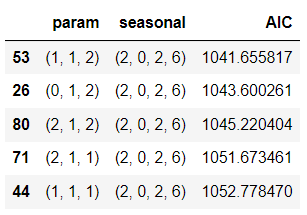


Table 10: Summary of SARIMA 6 seasonal for Rose wines

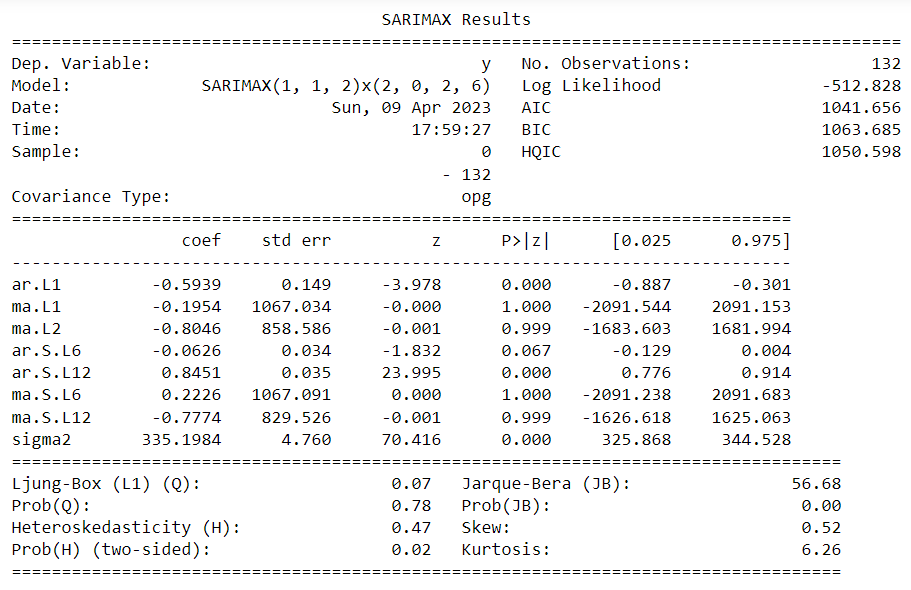
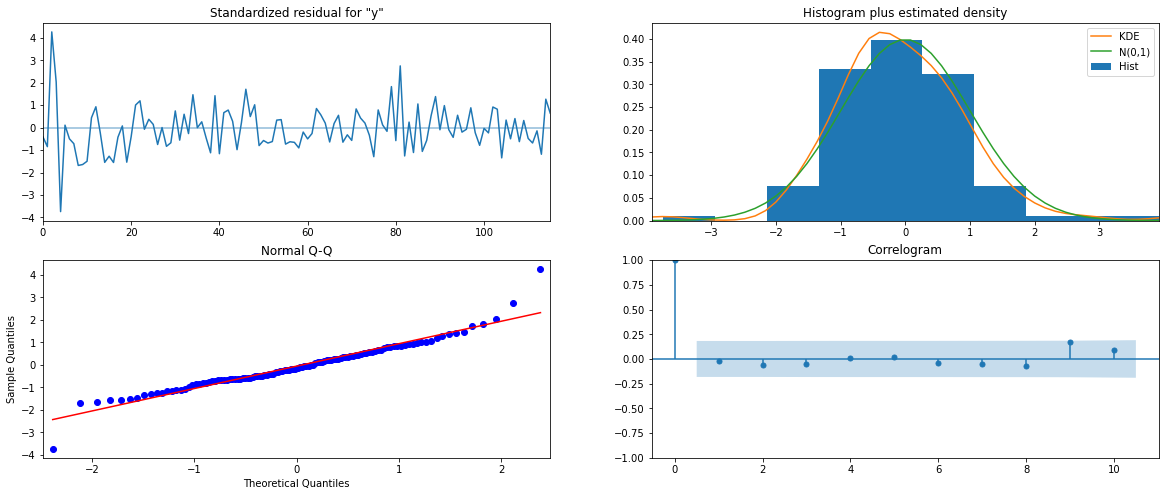


Figure 26: Residual plot for Rose SARIMA\_6\_seasonal



From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

Table 11: Sorted SARIMA AIC values for Sparkling wines

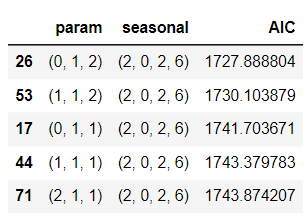


Table 12: Summary of SARIMA 6 seasonal for Sparkling wines

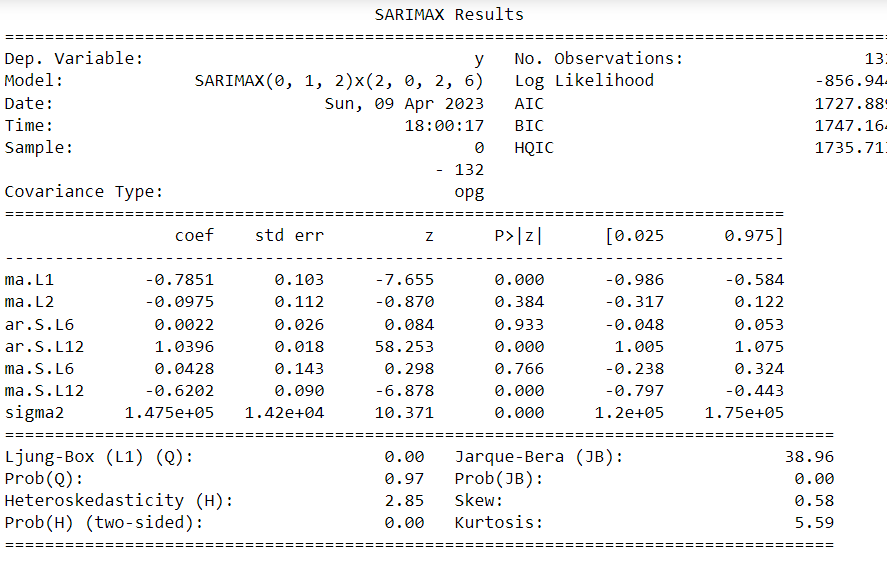
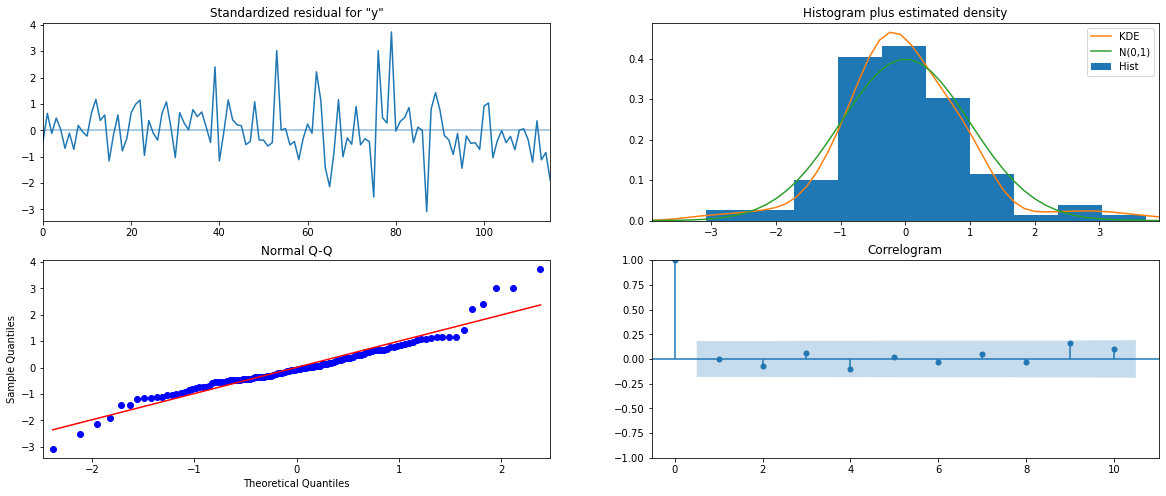


Figure 27: Residual plot for Sparkling SARIMA\_6\_seasonal



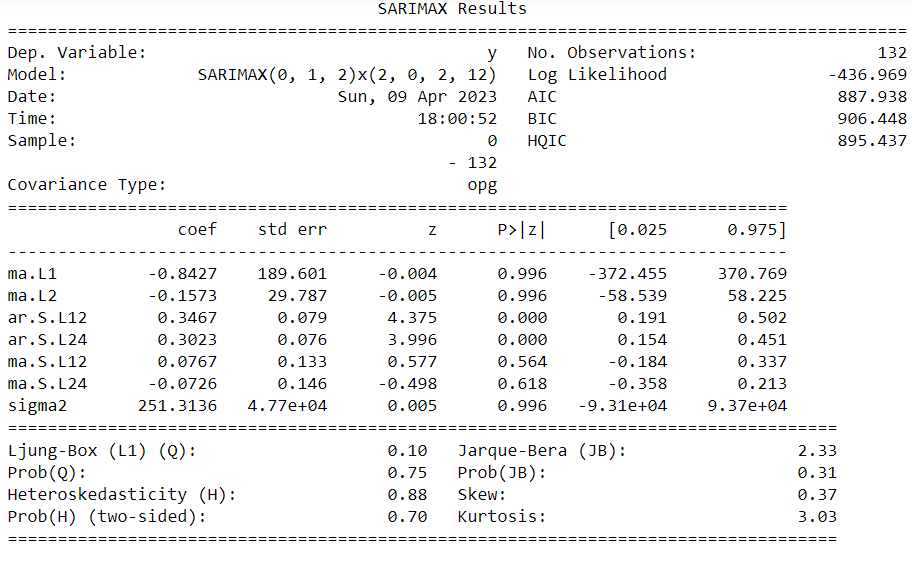
* The p-values for the coefficients in the SARIMA model can be used to determine if the coefficients are statistically significant. A p-value less than 0.05 indicates that the coefficient is statistically significant at a 95% confidence level. In the provided summary, the p-values for ar.L1, ar.S.L12, and the constant are less than 0.05, indicating that they are statistically significant.
* For SARIMA rose wines,ma.L1 is greater than 0.05.
* Ljung-Box test for autocorrelation at lag 1 (Q statistic): 0.07 with a p-value of 0.78, indicating no significant autocorrelation at this lag.
* Jarque-Bera test for normality: JB statistic of 56.68 with a p-value of 0.00, indicating that the residuals are not normally distributed.
* Heteroskedasticity test: The probability of the null hypothesis of homoskedasticity (equal variance) is rejected at the 0.02 significance level, indicating the presence of heteroskedasticity in the residuals.
* For SARIMA sparkling wines, the p-values for ma.L1 and ma.L2 are greater than 0.05, indicating that they are not statistically significant
* Ljung-Box test for autocorrelation at lag 1 (Q statistic): 0.07 with a p-value of 0.78, indicating no significant autocorrelation at this lag.
* Jarque-Bera test for normality: JB statistic of 56.68 with a p-value of 0.00, indicating that the residuals are not normally distributed.
* Heteroskedasticity test: The probability of the null hypothesis of homoskedasticity (equal variance) is rejected at the 0.02 significance level, indicating the presence of heteroskedasticity in the residuals.

The Prediction is done for 6 months, now will do for 12 months seasonality.

Table 13: Sorted values of SARIMA AIC 12 seasonality Rose wines data



Table 14: Summary SARIMA Rose wines data 12 seasonality



The second model is a SARIMAX(0,1,2)x(2,0,2,12) model with 132 observations. The model includes two moving average terms (MA) with lags 1 and 2, and two autoregressive terms (AR) with lags 12 and 24. The seasonal period is 12. The log likelihood of the model is -436.969, the Akaike Information Criterion (AIC) is 887.938, the Bayesian Information Criterion (BIC) is 906.448, and the Ljung-Box test has a p-value of 0.75.

Figure 28: Residual for SARIMA\_12 rose wines

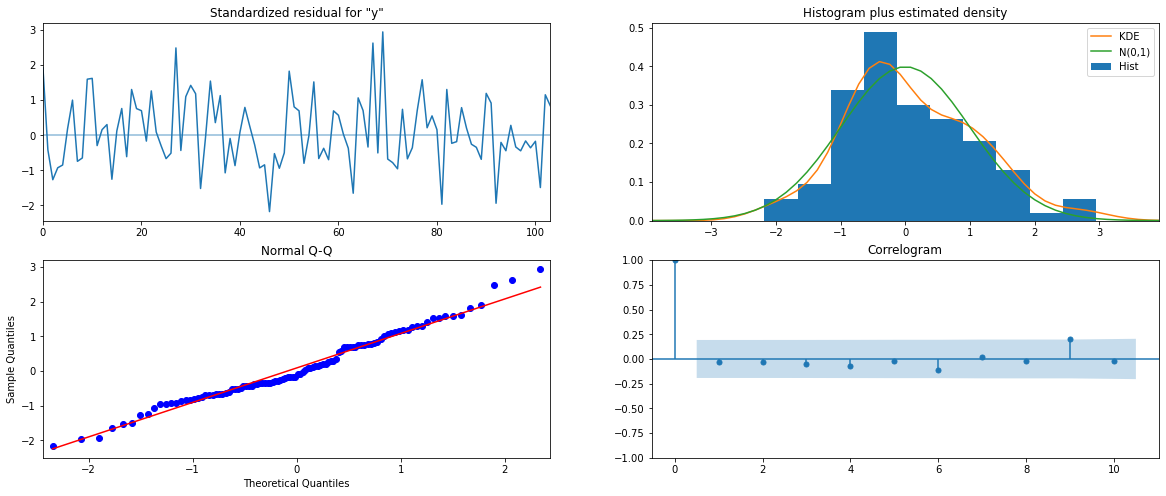


Table 15: Sorted values of SARIMA AIC 12 seasonality Sparkling wines data

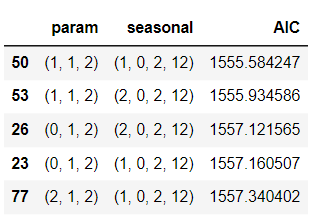
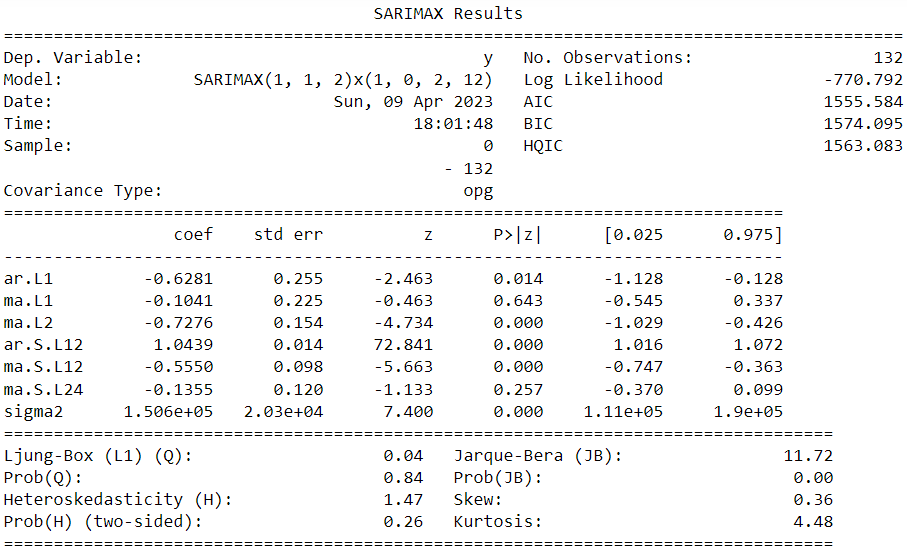
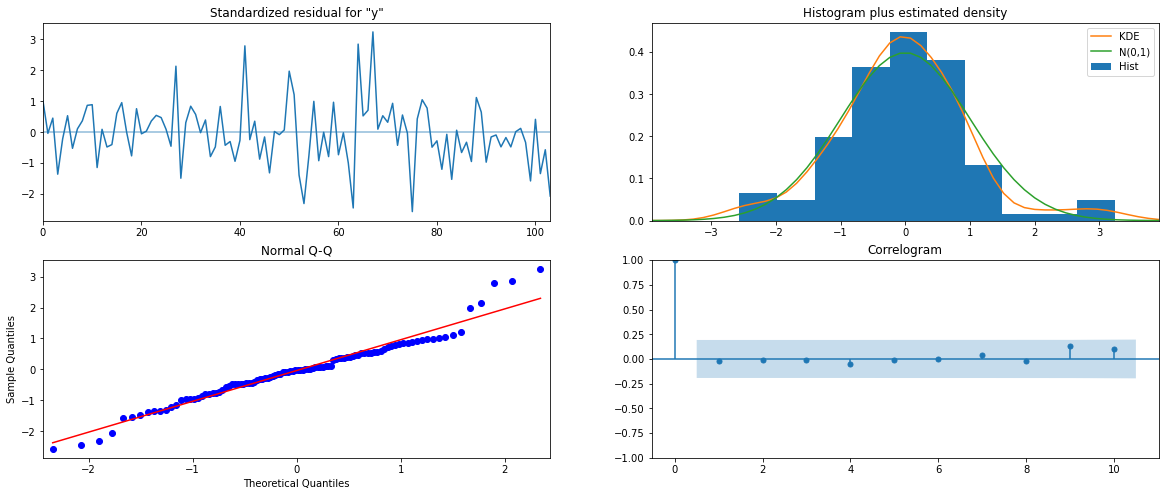


Table 16: Summary SARIMA Sparkling wines data 12 seasonality



The model's log-likelihood is -770.792, and the AIC and BIC are 1555.584 and 1574.095, respectively. These values indicate that the model fits the data relatively well. The residual variance (sigma2) is estimated at 1.506e+05, and the model's diagnostic tests show no evidence of autocorrelation or heteroscedasticity in the residuals. However, the Jarque-Bera test suggests some mild deviation from normality in the residuals.

Figure 29: Residual for SARIMA\_12 sparkling wines



**7. 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.**

Table 17: Models and their RSME values for Rose wines

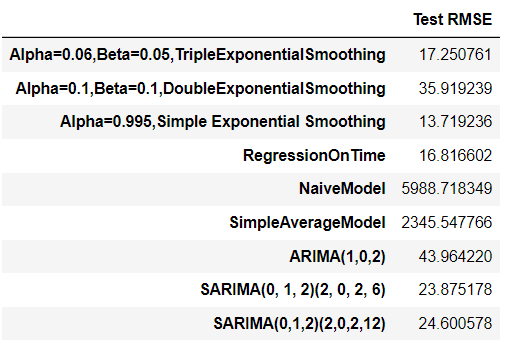
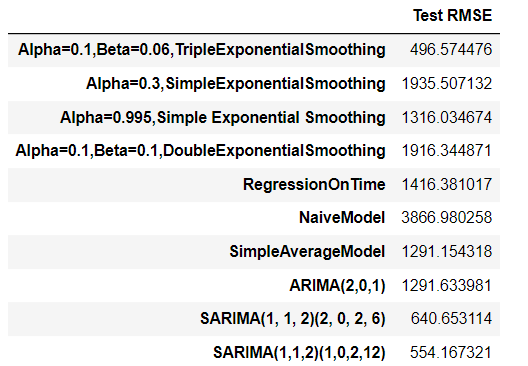
****

Table 18: Models and their RSME values for Sparkling wines



* For Rose wines data, comparing ARIMA/SARIMA models, SARIMA\_6 seasonality is best model
* For Sparkling wines data, comparing ARIMA/SARIMA models, SARIMA\_12 seasonality is best model

**8.** **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.**

Table 19: Summary for Rose wines best model

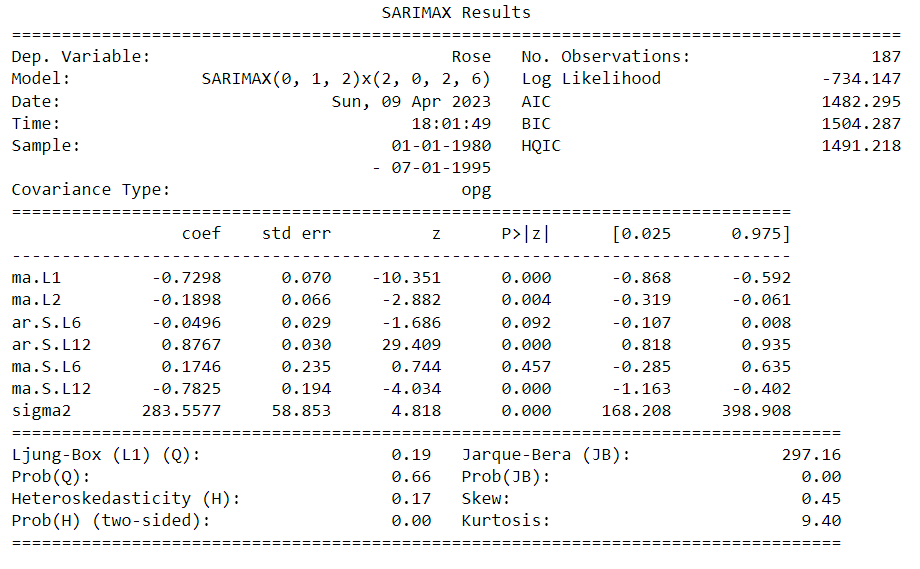
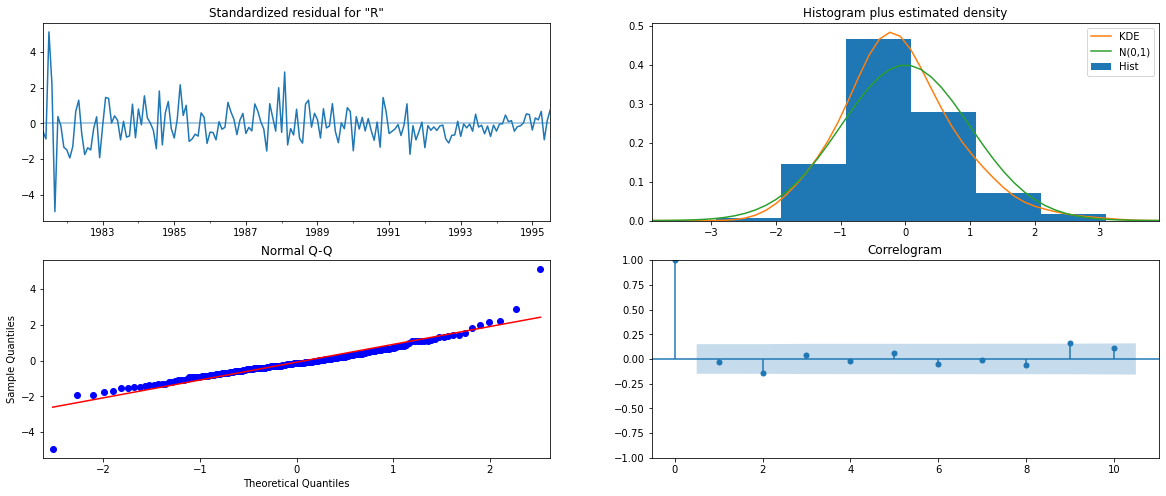


Figure 30: Residual for SARIMA rose wines best model



* The residuals are in shaded regions, which indicates that the residuals are uncorrelated.
* From Normal Q-Q plot we can say that the data is normally distributed with some skewness. The points are are on the line.

Table 20: Summary for Sparkling wines best model

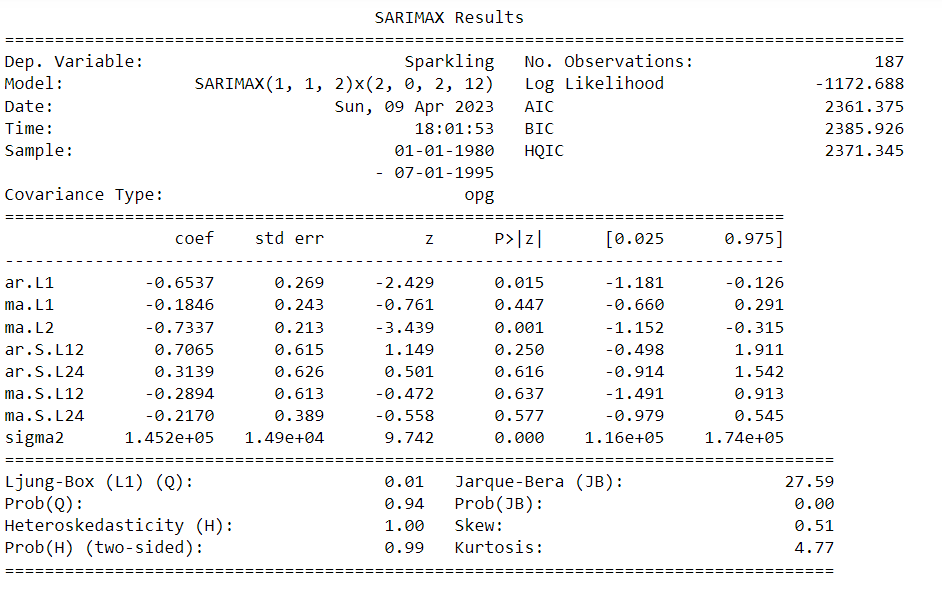
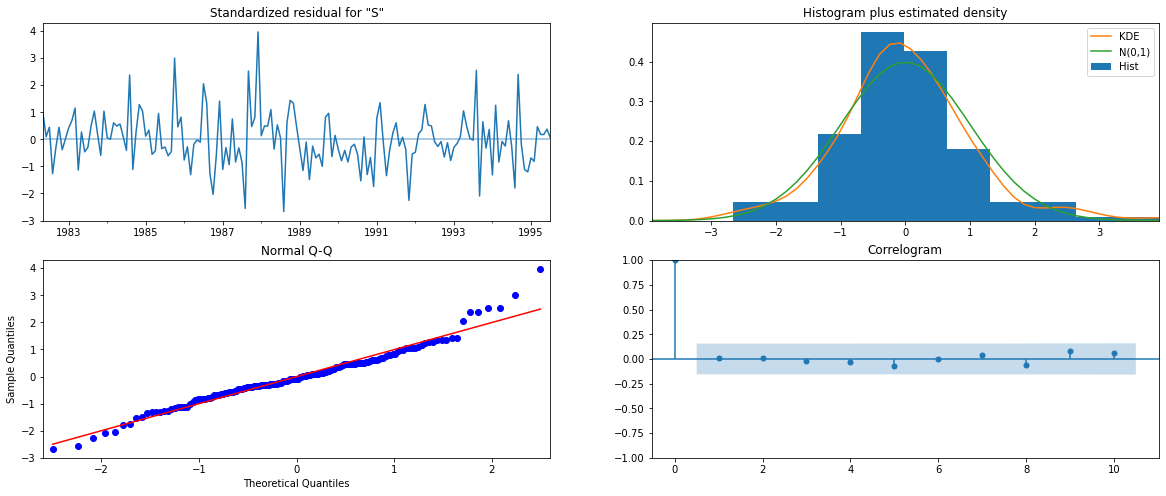


Figure 31: Residual for SARIMA Sparkling wines best model



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

1. Rose Wines no seasonality.
2. Sparkling wines have seasonality.
3. There is a trend in Rose wines
4. There is no trend in Sparkling wines
5. The sales are increased from August Sparkling wines.
6. There is decline trend for Rose wines.
7. In recent years, the sales for Rose wines is decreased. By advertising and new add-ons can increase the sales.
8. For Sparkling wines, there is high demand at the end of the years. The advertising and offers might in year starting might help in increasing sales.
9. From the predictions made, the sales in future for Rose wines is better and some creative ideas can help for increasing sales.