BHANU PRATAP REDDY

TIME SERIES FORECASTING

GRADED PROJECT

FEBRUARY 19 **2022**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Data from sale of different types of wines over the course of a particular period in the 20^{th} century is to be analyzed and based on the data forecast.

INTRODUCTION

We have the data for the sales of Sparkling Wine and Rose wine from the start of the year of 1980 to mid-1995. We shall utilize this data to first help build and evaluate appropriate time forecasting models and then forecast on test data. All of this shall be done for Sparking Wine sales and Rose wine Sales separately and independently and at the end of our analysis we shall state findings for our models and additional measure the company ABC estate Wines must take for future sales.

DATA DICTIONARY

The data dictionary gives us information regarding the variables present in the given dataset. we are analyzing and a brief explanation and what that variable contains or means. Therefore, the System measures used are given as:

- **YearMonth** – Indicates the Month and Year which the sales happened.
- **Sparkling** – Represents the Value of sales that occurred in a particular month for Sparkling wine.
- **Rose** - Represents the Value of sales that occurred in a particular month for Rose wine.

DATA IMPORT AND PRE-PROCESSING

After properly Initializing the date range and setting it as the index, we import data from the given file and here is a sample of it to verify whether or not data has been imported properly:

YearMonth	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Figure 2: *Data Sample of Sparkling wine Time series*

YearMonth	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Figure 1: *Data Sample of Rose wine Time series*

Checking for the Number of entries:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
```

Figure 3: *Sparkling Wine data info.*

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
```

Figure 4: *Rose Wine data info.*

Both Time series Contain 187 entries in their data set going from 31-01-1980 to 31-07-1995 but it appears there are 2 Null values in the Rose time series dataset.

For the missing values in Rose wine time series dataset we shall impute it utilizing forward fill since the nearby data to the missing values over a window have a very low deviation. Hence the final five point summary of the data series after imputation is given as:

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

Figure 6: *Sparkling Wine 5-point Data summary.*

	Rose
count	187.000000
mean	89.909091
std	39.244440
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Figure 5: *Rose Wine 5-point Data summary.*

It now appears all the data is in order and in the proper format for further processing. Now we shall plot the Time series data and perform and Exploratory Data analysis before we do any further processing.

TIME SERIES PLOT & EDA

With the data in the proper format the time series plot for each of them can be given as:

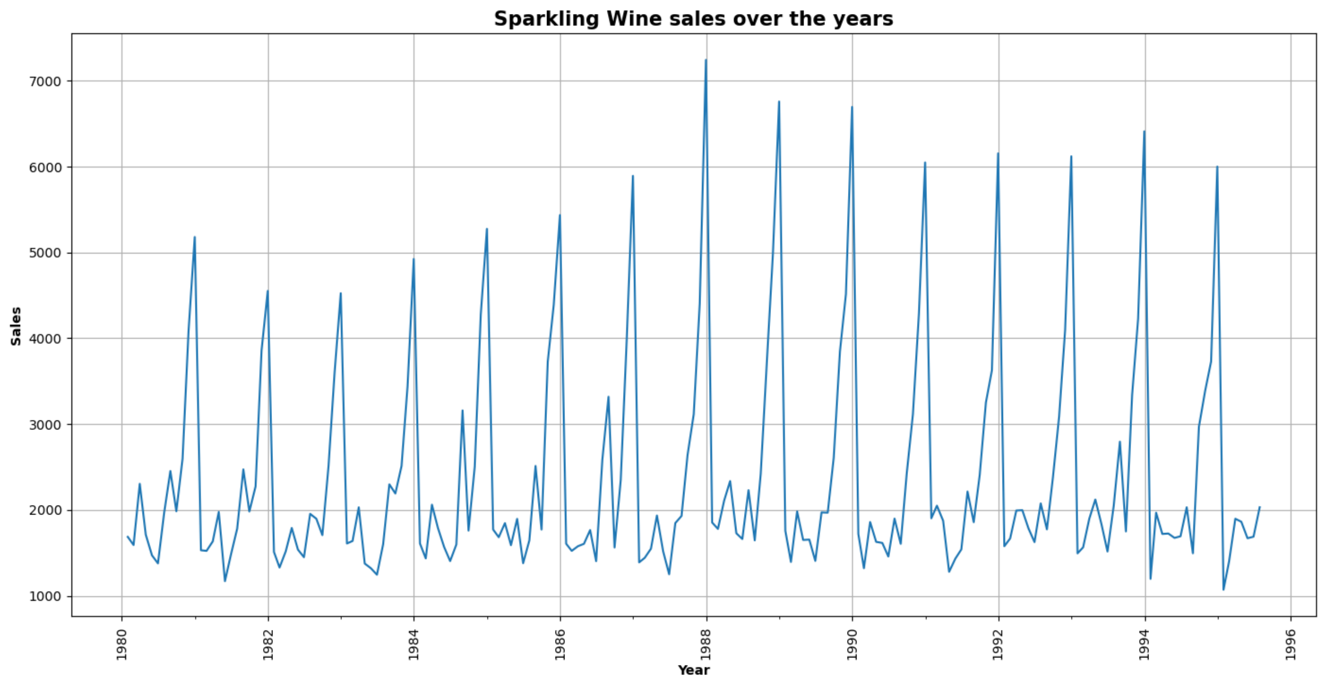

Figure 7: *Sparkling Wine Time Series Plot*

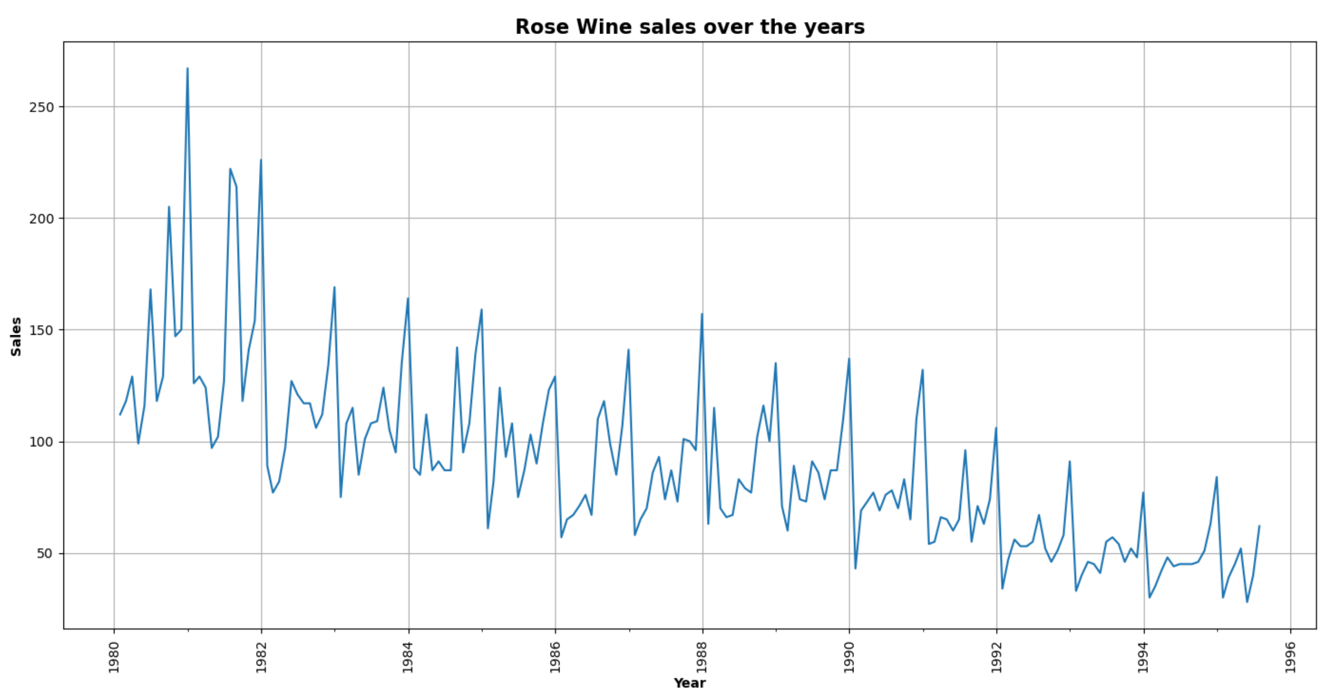

Figure 8: *Rose Wine Time Series Plot*

For both time series we can clearly see that they have a certain trend and seasonality let us have more closer look at the sales with respect to either years or by months over the years i.e.,:

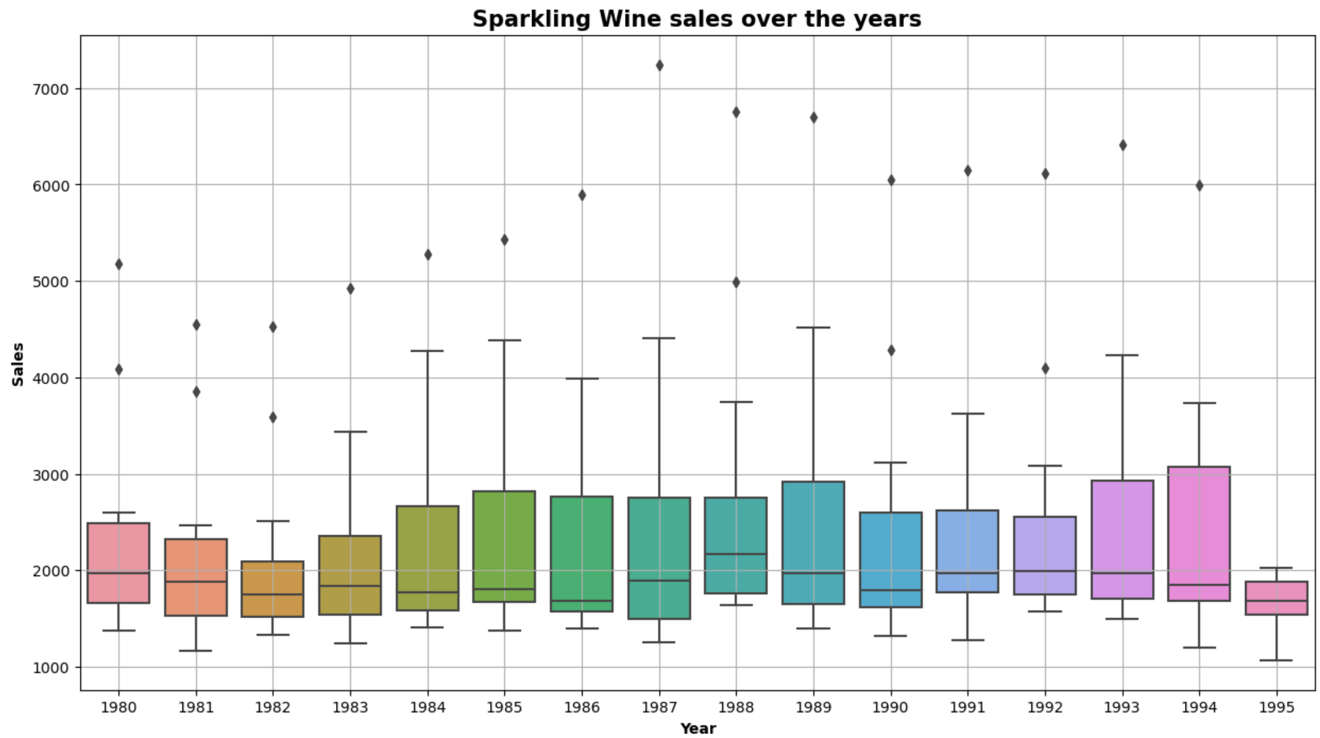

Figure 9: *Sparkling Wine yearly sales plot.*

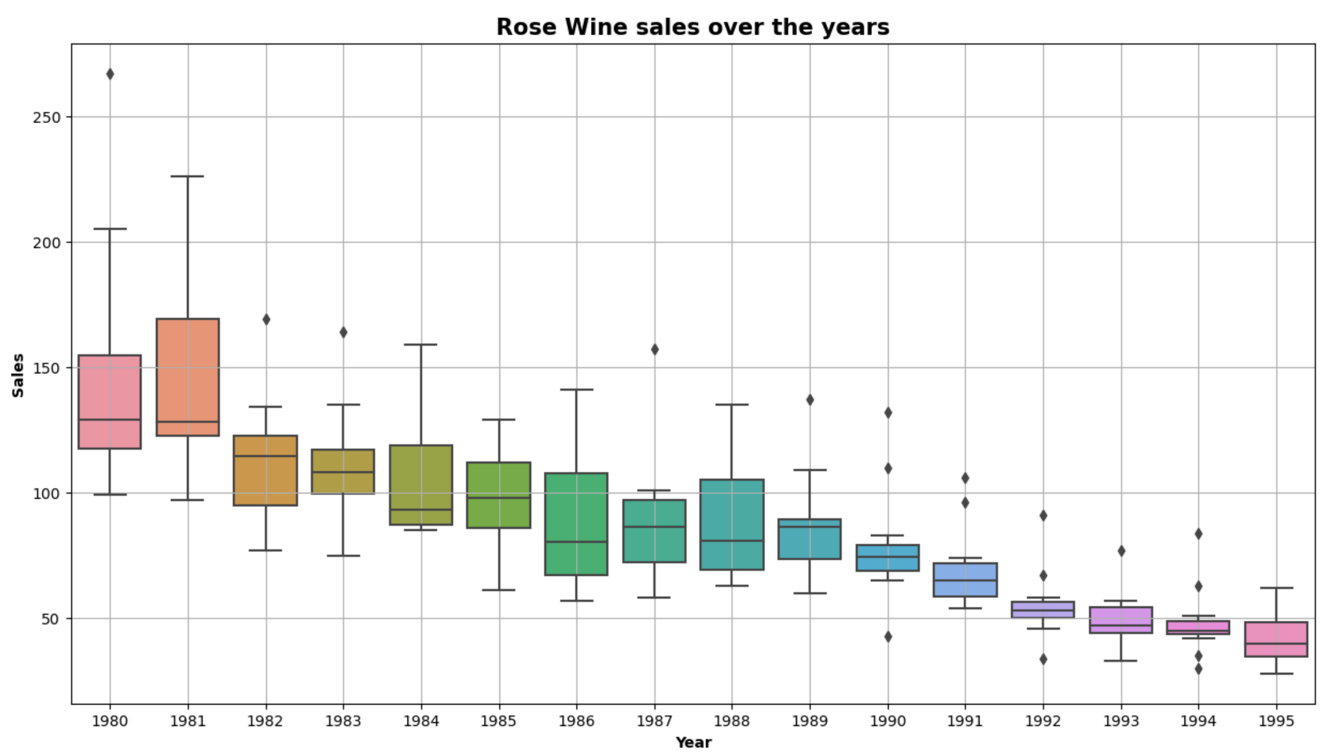

Figure 10: *Rose Wine yearly sales plot.*

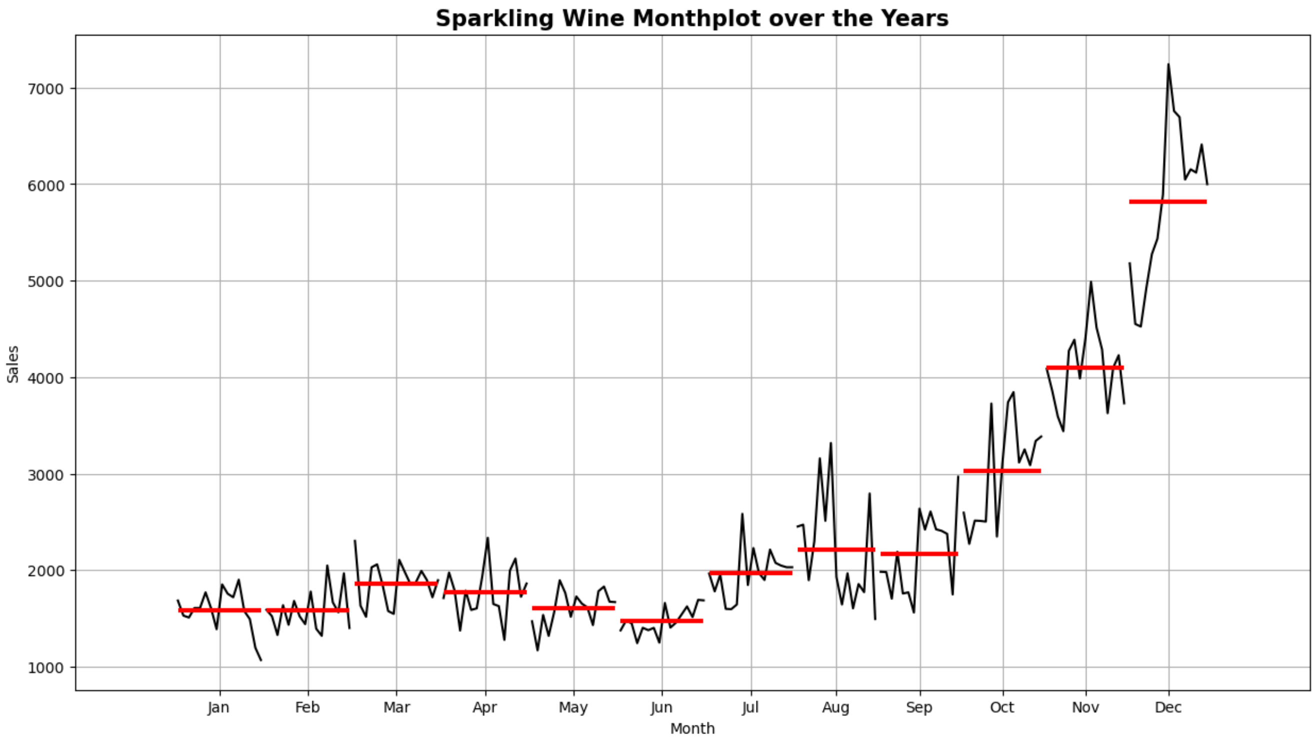

Figure 11: *Sparkling Wine Monthly Sales over the years.*

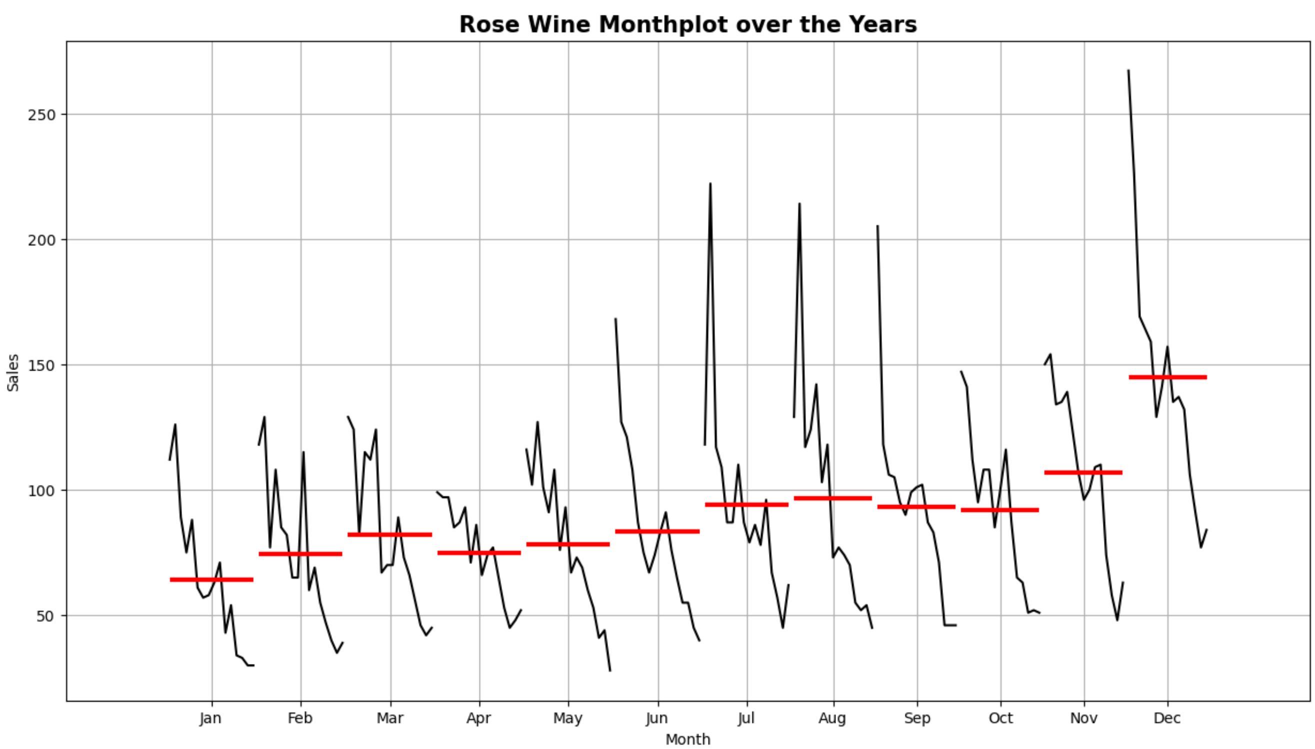

Figure 12: *Rose Wine Monthly Sales over the years.*

Sales Graph over the years as shown in Figure 9 and Figure 10 show the trend for both Sparkling and Rose wines and the very clear thing to note about the trend here is that it is somewhat constant for Sparkling wine with no clear decrease and trend is somewhat decreasing for Rose Wine over the years.
We shall have a more detailed look at this when we decompose this time series in the next section.

The Month plot for the time series over the years helps us visualize the seasonality present in both the time series.
As observer in Figure 11 and Figure 12 Sparkling wine has an increasing seasonality over the year i.e., start slow in January and exponentially Move to highest amount at the end of the year in December. Whereas Rose wine Starts slow In January and gradually increases up to July and then takes a dip in sales up to November and then shoots up to its highest peak in December.

TIME SERIES DECOMPOSITION

As we discussed earlier a little bit about trend and seasonality of the data we know it is important to decompose a time series into is elemental components i.e.,:

- T_t, The trend component at time t, which reflects the long-term progression of the series (secular variation). A trend exists when there is a persistent increasing or decreasing direction in the data. The trend component does not have to be linear.

- S_t, The seasonal component at time t, reflecting seasonality (seasonal variation). A seasonal pattern exists when a time series is influenced by seasonal factors. Seasonality occurs over a fixed and known period (e.g., the quarter of the year, the month, or day of the week).

- I_t, The irregular component (or "noise") at time t, which describes random, irregular influences. It represents the residuals or remainder of the time series after the other components have been removed.

- Y_t, The resultant time series which is given as a function of its components based on the Model Utilized i.e.

    - $Y_t = T_t + S_t + I_t$ for an additive model where the variations around the trend do not vary with the level of the time series.

    - $Y_t = T_t \ x \ S_t \ x \ I_t$ for a multiplicative model where the trend is proportional to the level of the time series.

Based on what we have already gathered from the EDA it looks like:

- Sparkling Wine time series data would be best described by a Multiplicative model.

- Rose Wine time series data also would be best described by a Multiplicative model.

Hence, we have:

- Sparkling wine Time Series Decomposition represented as:

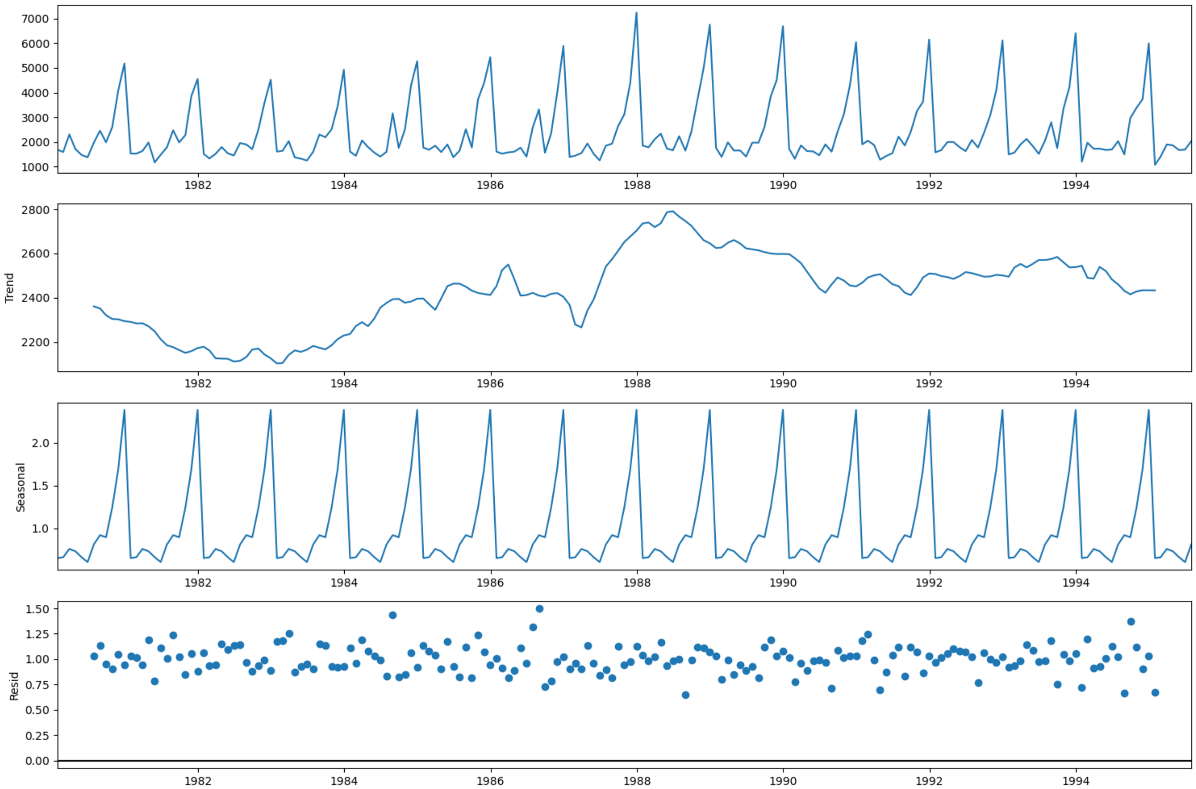

Figure 13: *Time series Decomposition in multiplicative method for Sparkling Wine Time Series.*

- Rose wine Time Series Decomposition represented as

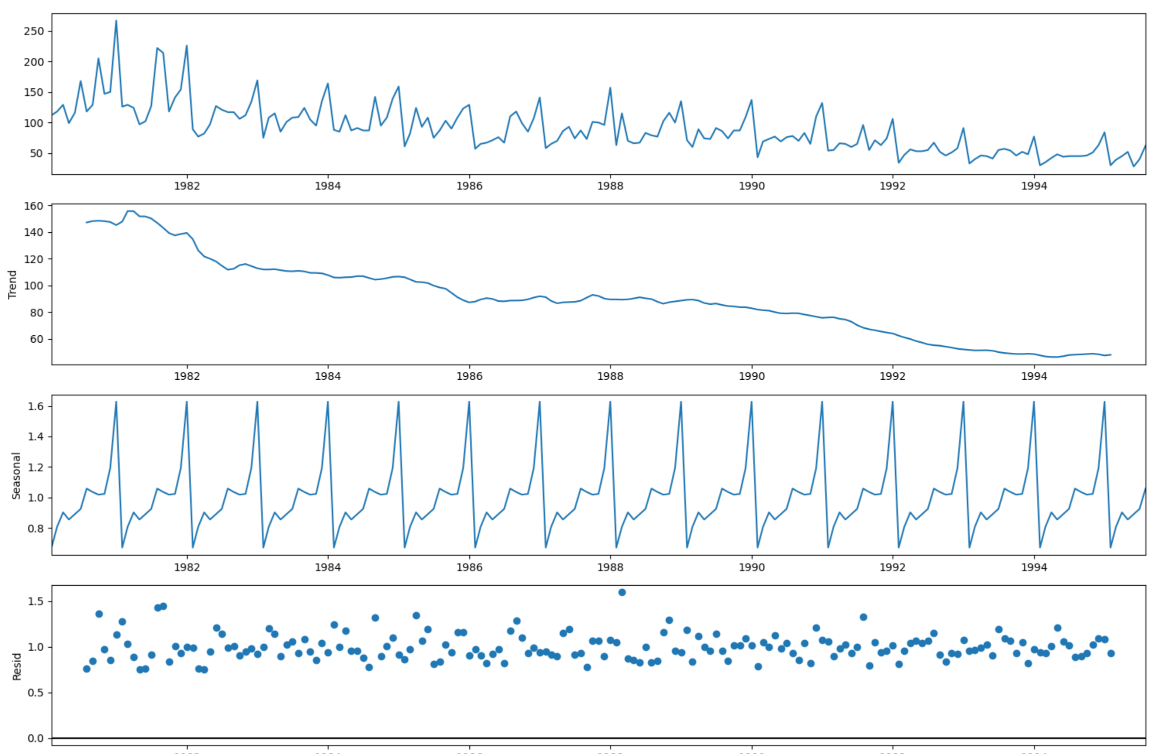

Figure 14: *Time series Decomposition in multiplicative method for Rose Wine Time series.*

TRAIN TEST SPLIT

We have to split our data into train and test in order to help train and evaluate models for the time series forecasting. As per the recommendation we will split of all data after 1991 as test series hence this can be given as:

- Train Test Split for Sparkling Wine Data:

First few rows of Training Data

YearMonth	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Figure 16: *Sparkling Wine Train set first five Rows.*

Last few rows of Training Data

YearMonth	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Figure 15: *Sparkling Wine Train set last five Rows.*

First few rows of Test Data

YearMonth	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Figure 18: *Sparkling Wine Test set first five Rows.*

Last few rows of Test Data

YearMonth	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

Figure 17: *Sparkling Wine Test set last five Rows.*

The above shows the splits of train and test data and at which point they occurred.

But this can be better visualized as follows:

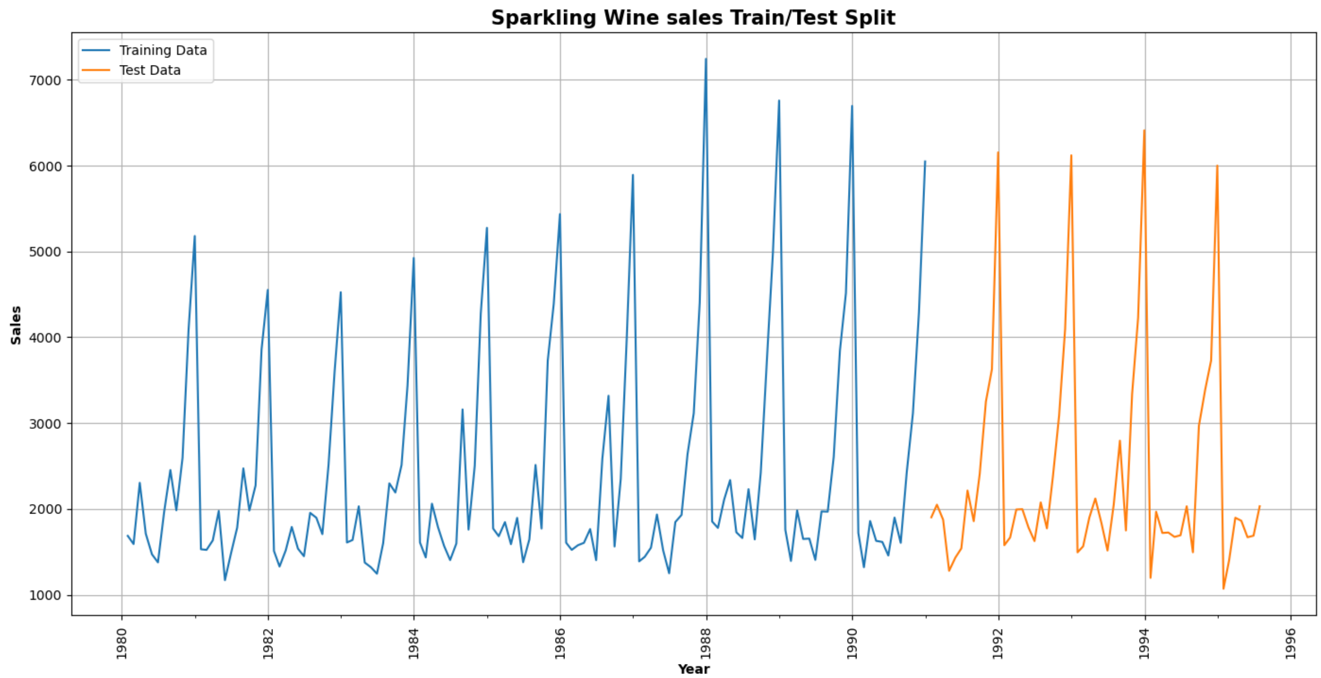

Figure 19: *Sparkling wine Train test split plot*

- Train Test Split for Rose Wine Data:

First few rows of Training Data

YearMonth	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Figure 21: *Rose Wine Train set first five Rows.*

Last few rows of Training Data

YearMonth	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

Figure 20: *Rose Wine Train set last five Rows.*

First few rows of Test Data

YearMonth	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Figure 23: *Rose Wine Test set first five Rows*

Last few rows of Test Data

YearMonth	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Figure 22: *Rose Wine Test set last five Rows.*

The above shows the splits of train and test data and at which point they occurred. But this can be better visualized as follows:

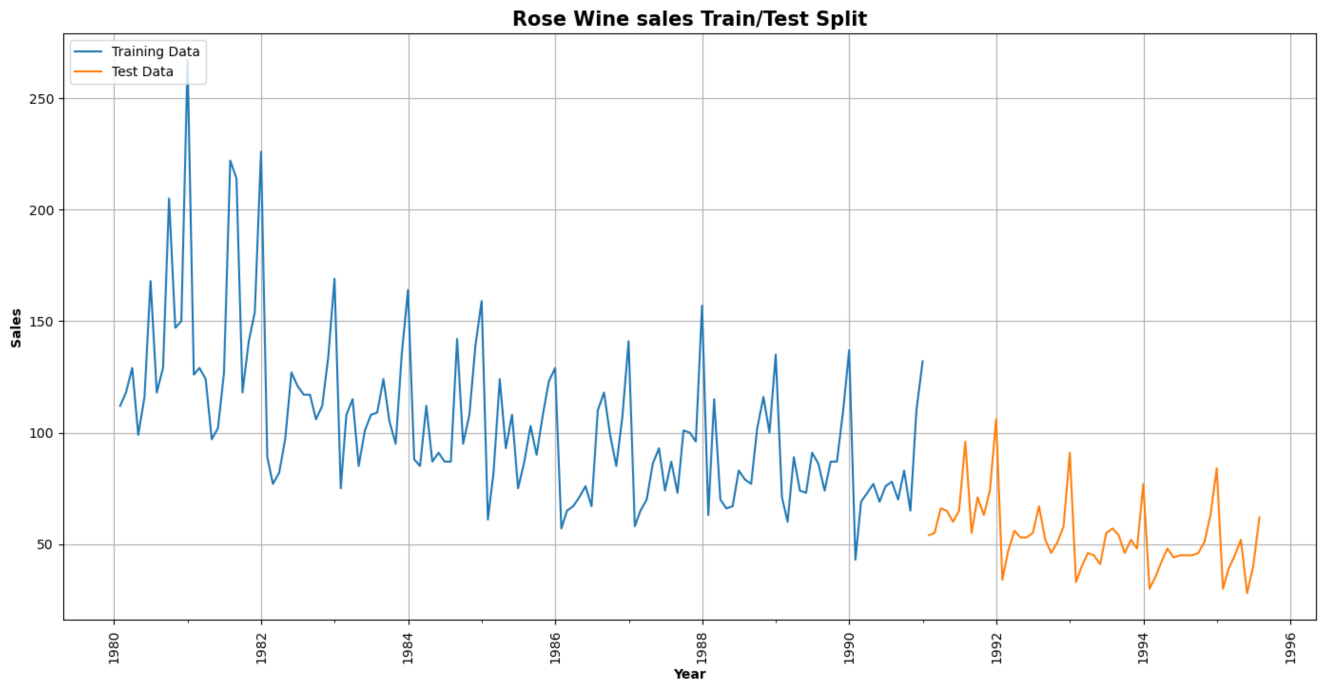

Figure 24: *Rose wine Train test split plot.*

FORECAST MODELS EVALUATION

For the data from Sparkling wine and Rose wine time series the following models will be built and evaluated upon RMSE (Root Mean Squared Error) based on the previous train test split we performed:

- Linear Regression
- Naïve Forecast
- Simple Average
- Moving Window Average (for 2 pt, 4pt, 6pt and 9pt)
- Single Exponential Smoothing (SES)
- Double Exponential Smoothing (DES, Holt's Model)
- Triple Exponential Smoothing (TES, Holt – Winter's Model)

We will also look at ARIMA and SARIMA models but will dedicate another complete section for that. In this section we ill discuss with respect to the above-mentioned models only.

SPARKLING WINE MODELS EVALUATION

In order to avoid clutter in the representation we will see the visualization of each model's performance grouped appropriately before we discuss their performance i.e., RMSE.

- Visualization of performance of Linear Regression, Naïve and Simple Average:

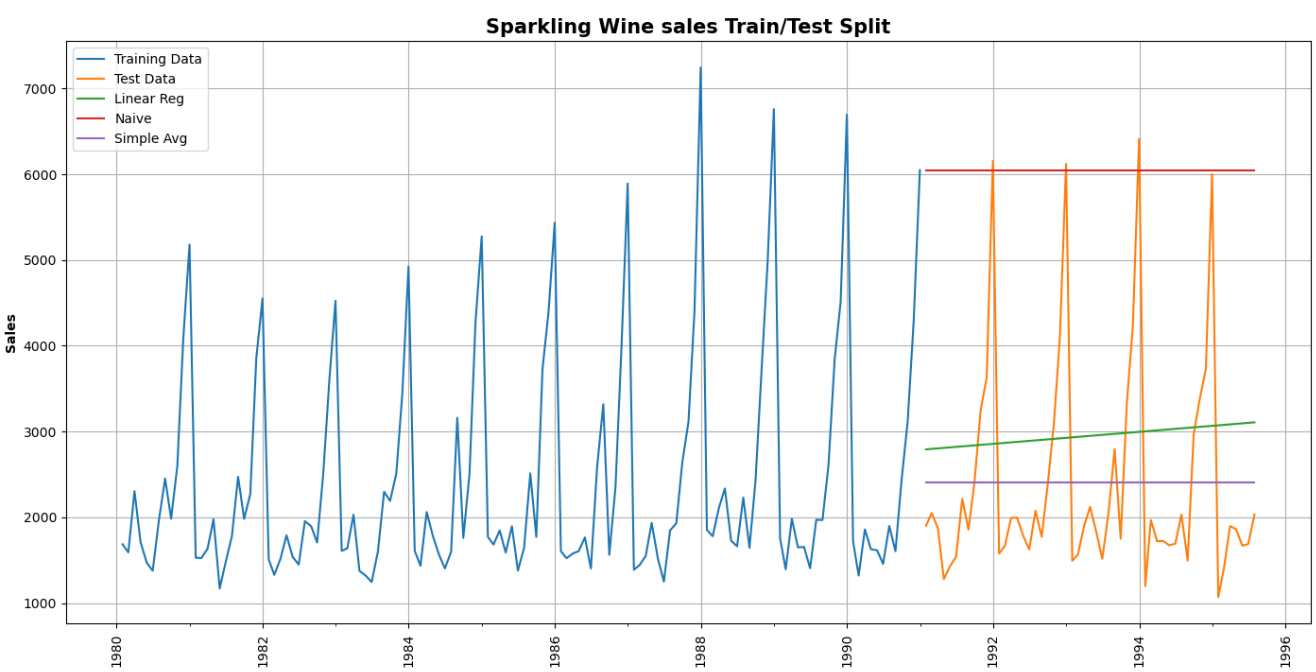

Figure 25: *Plot of Performance of Linear Reg, Naive and Simple Avg vs Test data in Sprakling Wine Time Series.*

- Visualization of performance of Moving Window average for 2, 4, 6 and 9 points:

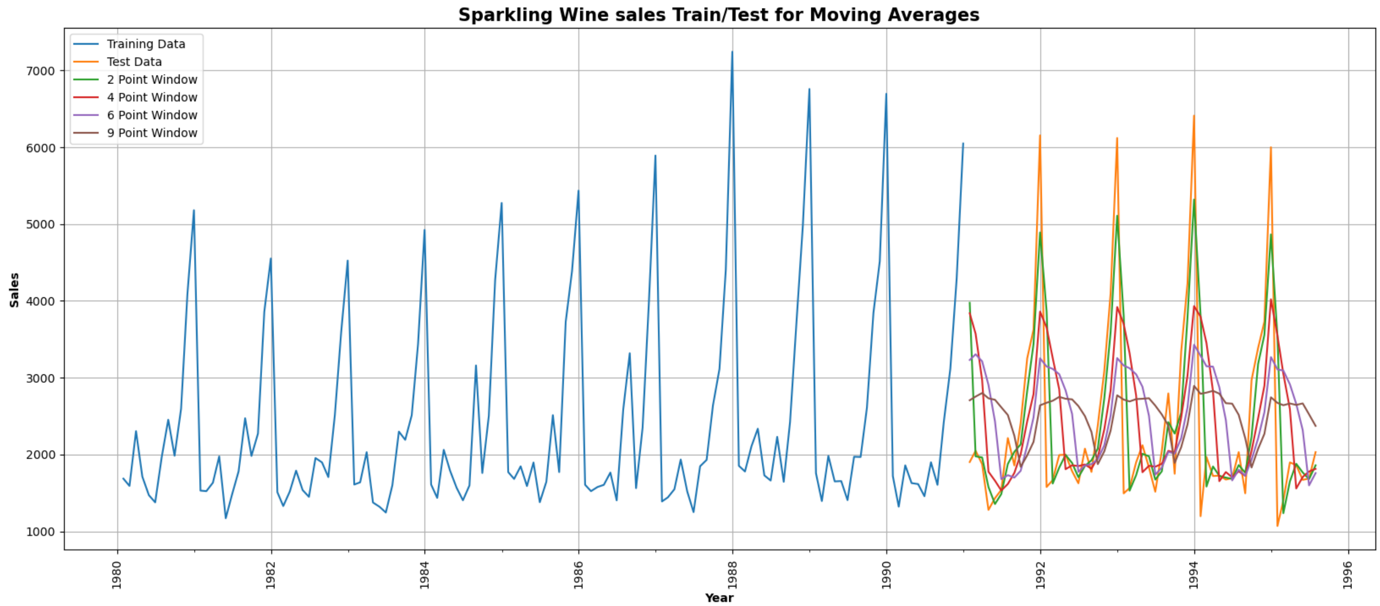

Figure 26: *Plot of Performance of Moving average Window vs Test data in Sparkling Wine Time Series.*

- Visualization of performance of SES, DES and TES:

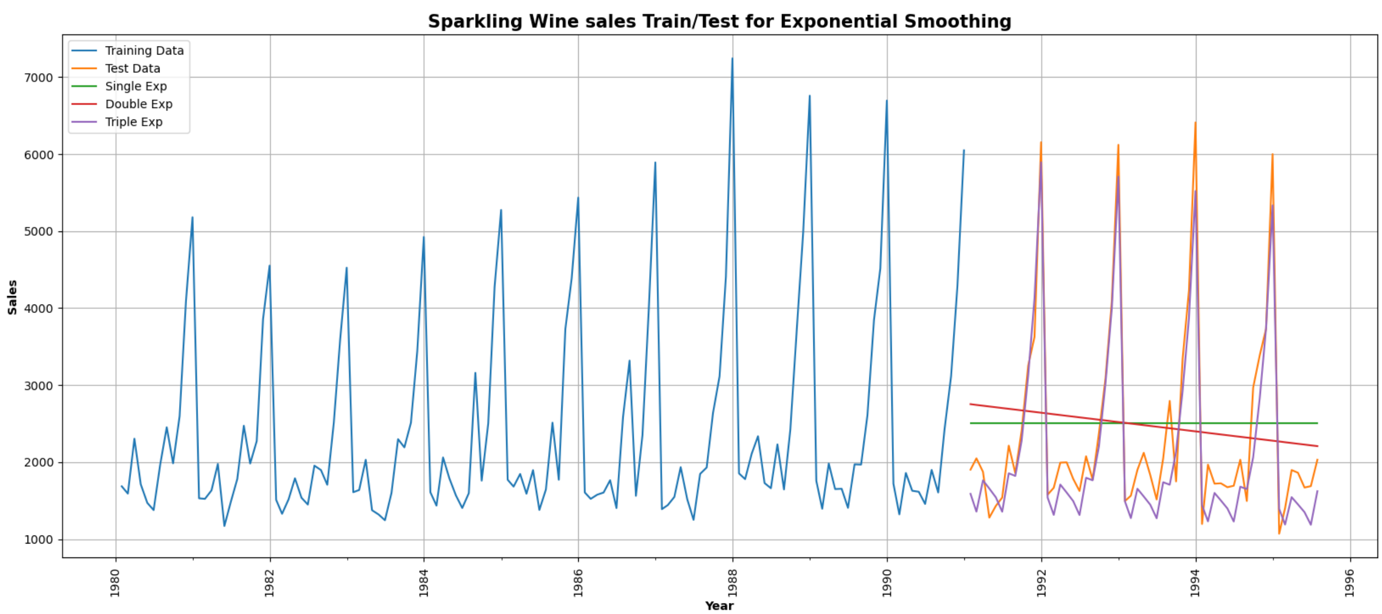

Figure 27: *Plot of Performance of Exponential Smoothing vs Test data in Sparkling Wine Time Series.*

The RMSE Values of each model's performance with respect to original test data after being trained arranged in ascending order (lower RMSE indicates better performance):

Table 1: *Table Showing RMSE values of All Smoothing, Moving Average, and regression models for Sparkling Wine Time series.*

MODEL	**RMSE**
Sparkling - TES, α = 0.111, β = 0.049, γ = 0.362	402.9362
Sparkling - 2 pt Moving Avg	813.4007
Sparkling - 4 pt Moving Avg	1156.59
Sparkling - Simple Average	1275.082
Sparkling - SES, α = 0.02	1279.495
Sparkling - 6 pt Moving Avg	1283.927
Sparkling - DES, α = 0.1, β = 0.01	1292.437
Sparkling - 9 pt Moving Avg	1346.278
Sparkling - Linear Regression	1389.135
Sparkling - Max Error Possible	2403.780
Sparkling - Naive Forecast	3864.279

The RMSE Values along with the visualizations confirm that the best performing model is the Triple exponential Smoothing i.e., The Hold-Winters model with values of Level, trend and seasonality were **α = 0.111, β = 0.049** and **γ = 0.362.**

Now let us Look at the same models but for the Rose Wine time series.

ROSE WINE MODELS EVALUATION

In order to avoid clutter in the representation we will see the visualization of each model's performance grouped appropriately before we discuss their performance i.e., RMSE.

- Visualization of performance of Linear Regression, Naïve and Simple Average:

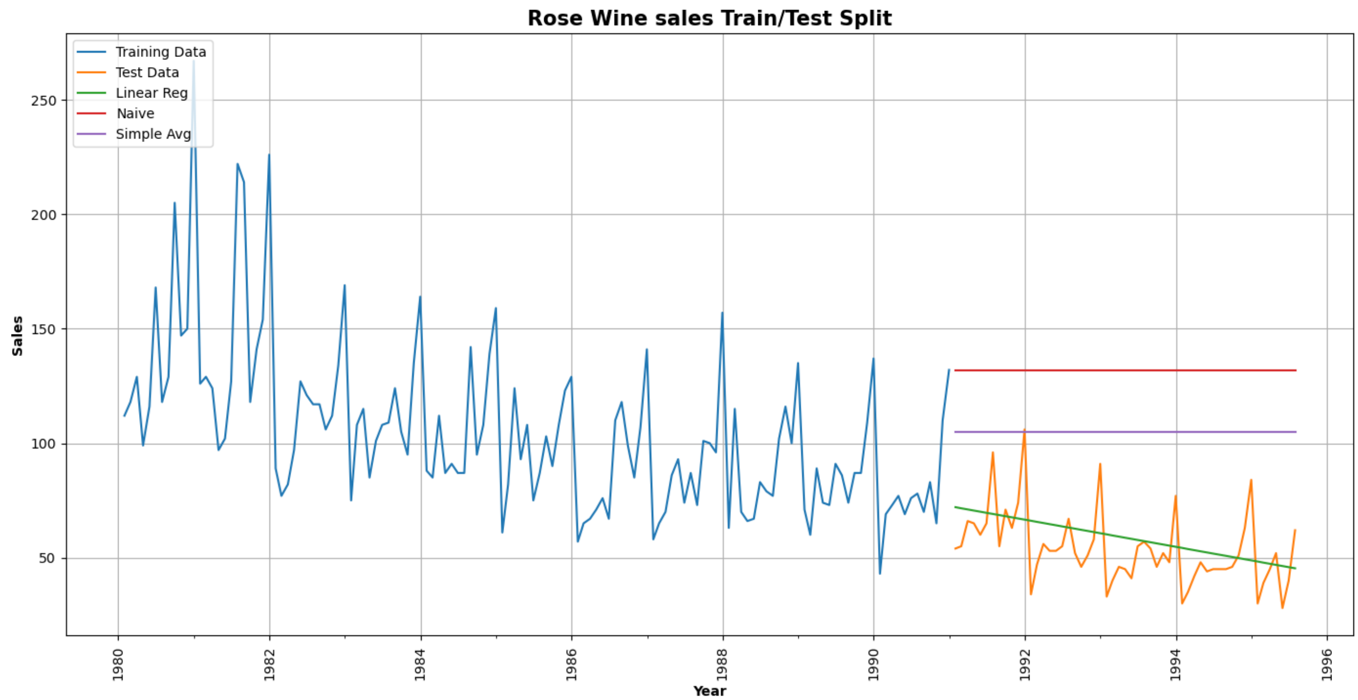

Figure 28: *Plot of Performance of Linear Reg, Naive and Simple Avg vs Test data in Rose Wine Time Series.*

- Visualization of performance of Moving Window average for 2, 4, 6 and 9 points:

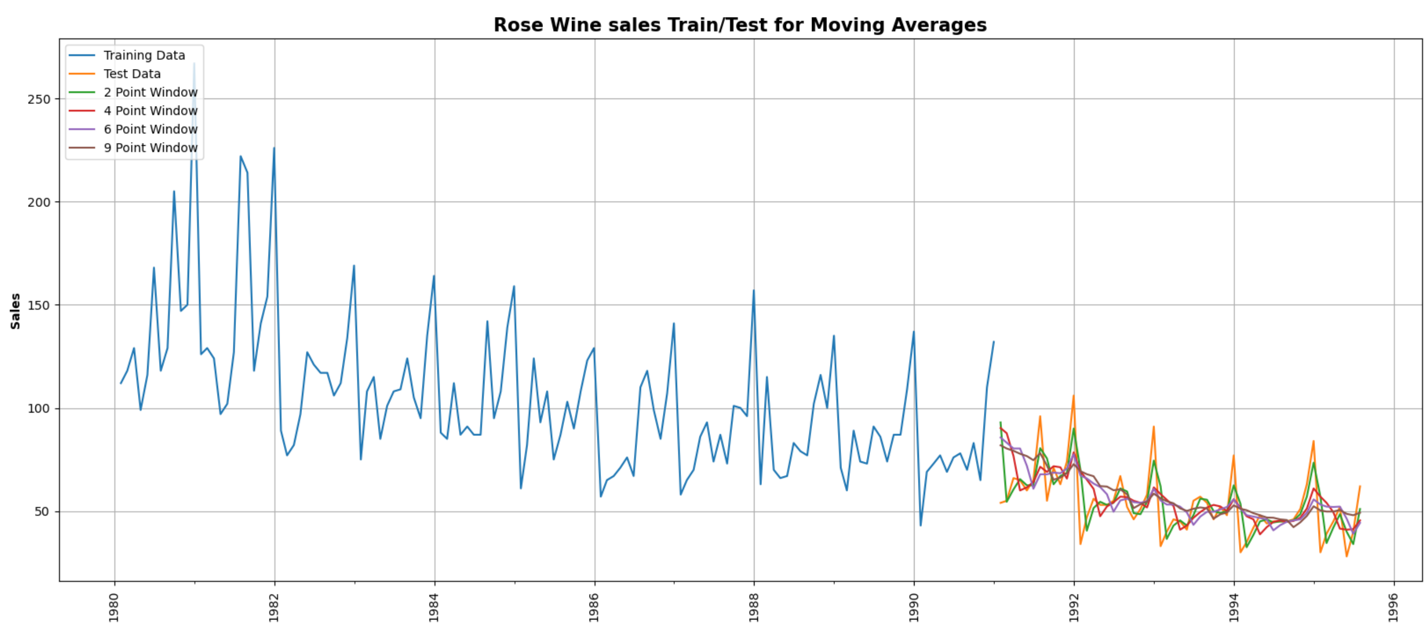

Figure 29: *Plot of Performance of Moving average Window vs Test data in Rose Wine Time Series.*

- Visualization of performance of SES, DES and TES:

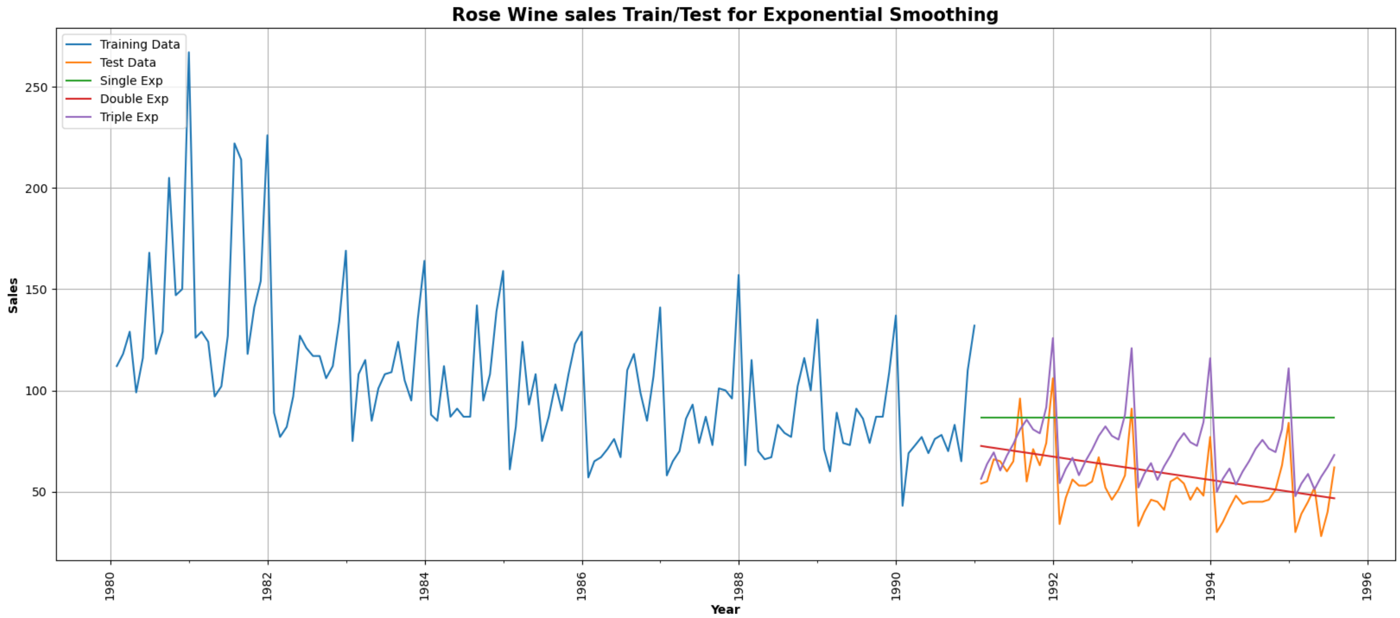

Figure 30: *Plot of Performance of Exponential Smoothing vs Test data in Rose Wine Time Series.*

The RMSE Values of each model's performance with respect to original test data after being trained arranged in ascending order (lower RMSE indicates better performance):

Table 2: *Table Showing RMSE values of All Smoothing, Moving Average, and regression models for Rose Wine Time series.*

MODELS	RMSE
Rose - 2 pt Moving Avg	11.52940904
Rose - 4 pt Moving Avg	14.45522082
Rose - 6 pt Moving Avg	14.57200859
Rose - 9 pt Moving Avg	14.73120938
Rose - Linear Regression	15.2757316
Rose - DES, α = 0.0001, β = 1.05e-16	15.57701993
Rose - TES, α = 0.071, β = 0.045, γ = 8.38e-5	20.21571908
Rose - SES, α = 0.07	36.45643859
Rose - Simple Average	53.48085658
Rose - Naive Forecast	79.73855005
Rose - Max Error Possible	104.9393939

The RMSE Values along with the visualizations confirm that the best performing model is the 2-point moving average model.

ARIMA & SARIMA MODELS

ARIMA and SARIMA (Seasonal ARIMA) can be compared to an alloy a similarity as ARIMA is the combination of Auto Regression Techniques and Moving Average Techniques to form a model which can better predict/forecast a time series.

One of the most important things for ARIMA/SARIMA is for the time series to be stationary hence we will check for stationarity before we proceed to build the ARIMA and SARIMA models.

Note: Although there is no need to build ARIMA models in our case since we have clearly established that our Time series data has seasonality, we will however build ARIMA models for the sake of comparison with a SARIMA model.

STATIONARITY

The stationarity for a given time series can be checked with the Dickey Fuller Test. Whose Hypothesis can be stated as:

$$\boldsymbol{Null\ Hypothesis}: \quad \boldsymbol{H_0}: The\ Given\ Time\ series\ is\ not\ Stationary$$
$$\boldsymbol{Alternate\ Hypothesis}: \quad \boldsymbol{H_A}: The\ Given\ Time\ series\ is\ Stationary$$

- The Dickey Fuller Test on Sparkling Wine data (Difference - 0):

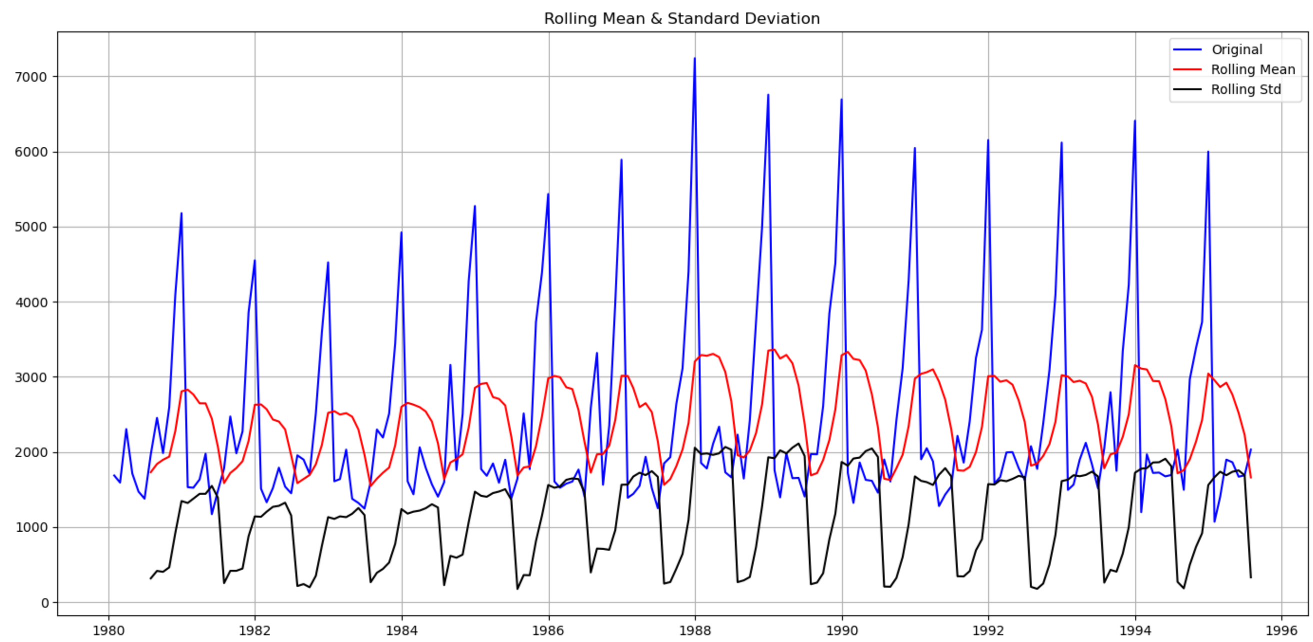

Figure 31: *Rolling mean and std, Deviation - d = 0 - Sparkling Wine Time Series.*

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.36
p-value                          0.60
#Lags Used                      11.00
Number of Observations Used    175.00
Critical Value (1%)             -3.47
Critical Value (5%)             -2.88
Critical Value (10%)            -2.58
```

Figure 32: *Dickey Fuller test Statistic - d = 0 - Sparkling Wine Time Series.*

$$p > 0.05, Hence\ Series\ is\ not\ Stationary.$$

- The Dickey Fuller Test on Sparkling Wine data (Difference - 1):

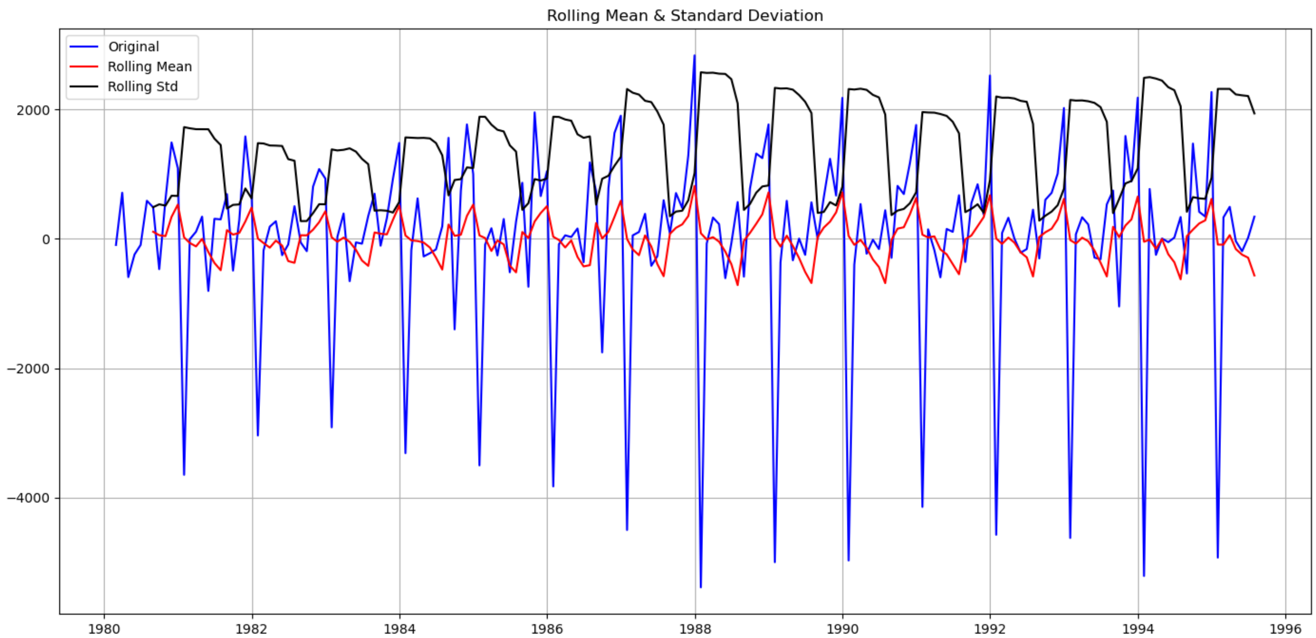

Figure 33: *Rolling mean and std, Deviation - d =1 - Sparkling Wine Time Series.*

```
Results of Dickey-Fuller Test:
Test Statistic                 -45.05
p-value                          0.00
#Lags Used                      10.00
Number of Observations Used    175.00
Critical Value (1%)             -3.47
Critical Value (5%)             -2.88
Critical Value (10%)            -2.58
```

Figure 34: *Dickey Fuller test Statistic - d = 1 - Sparkling Wine Time Series.*

$$p < 0.05, Hence\ Series\ is\ Stationary.$$

Similarly for Rose wine time series data.

- The Dickey Fuller Test on Rose Wine data (Difference - 0):

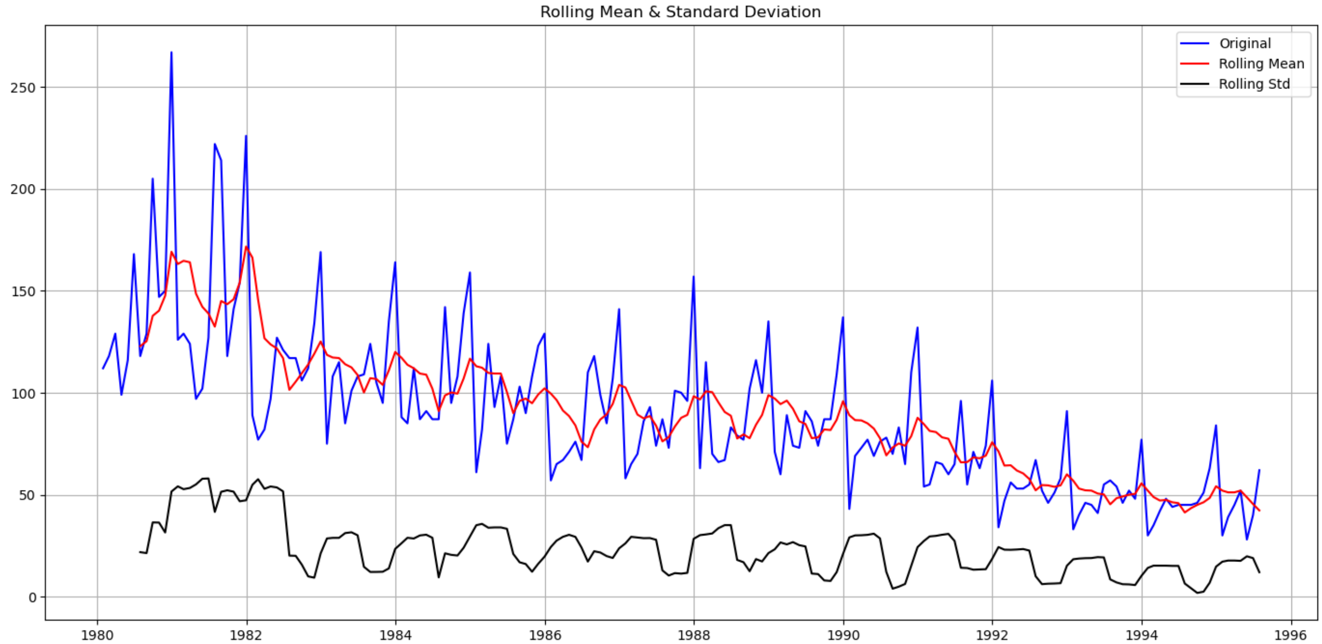

Figure 35: *Rolling mean and std, Deviation - d =0 - Rose Wine Time Series.*

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.87
p-value                          0.34
#Lags Used                      13.00
Number of Observations Used    173.00
Critical Value (1%)             -3.47
Critical Value (5%)             -2.88
Critical Value (10%)            -2.58
```

Figure 36: *Dickey Fuller test Statistic - d = 0 - Rose Wine Time Series.*

$$p > 0.05, Hence\ Series\ is\ not\ Stationary.$$

- The Dickey Fuller Test on Rose Wine data (Difference - 1):

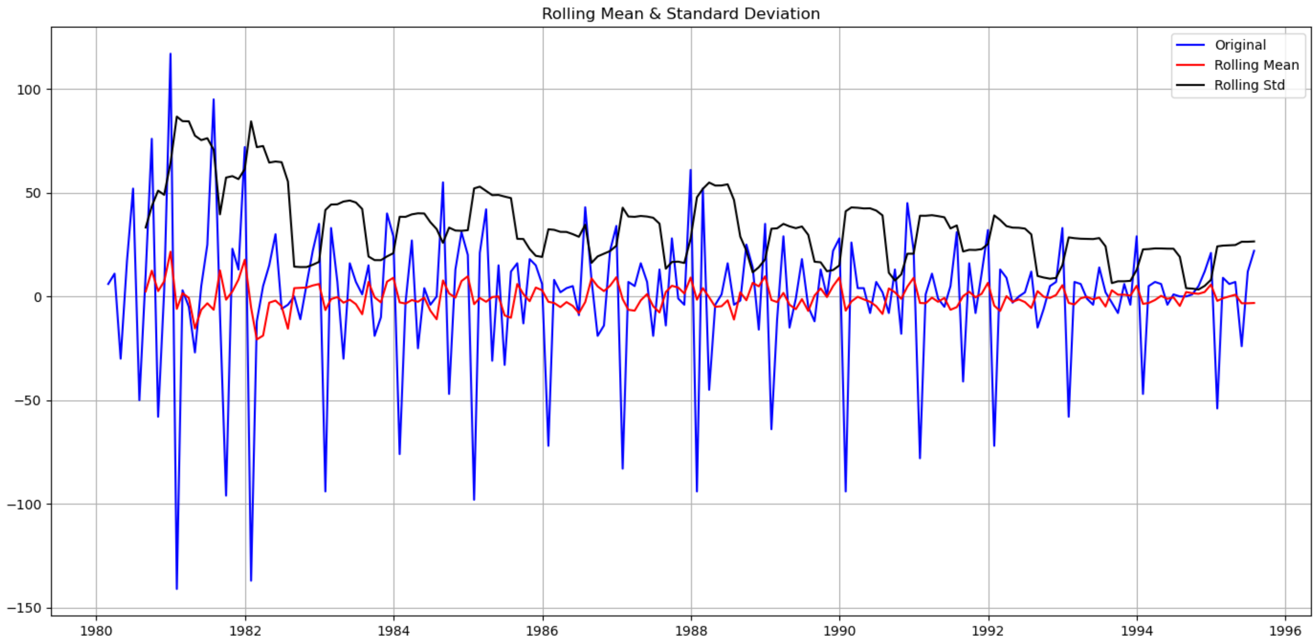

Figure 37: *Rolling mean and std, Deviation - d =1 - Rose Wine Time Series.*

Results of Dickey-Fuller Test:

Test Statistic	-8.04
p-value	0.00
#Lags Used	12.00
Number of Observations Used	173.00
Critical Value (1%)	-3.47
Critical Value (5%)	-2.88
Critical Value (10%)	-2.58

Figure 38: *Dickey Fuller test Statistic - d = 1 - Rose Wine Time Series.*

$$p < 0.05, Hence\ Series\ is\ Stationary.$$

AUTO REGRESSION & MOVING AVERAGE ORDERS (p & q)

Before we proceed with modelling ARIMA/SARIMA we must first determine the orders p and q for the model. This can be done by plotting the Autocorrelation and Partial Autocorrelation of the data.

- Autocorrelation Plot and Differenced Autocorrelation Plot for Sparkling Wine Time series.

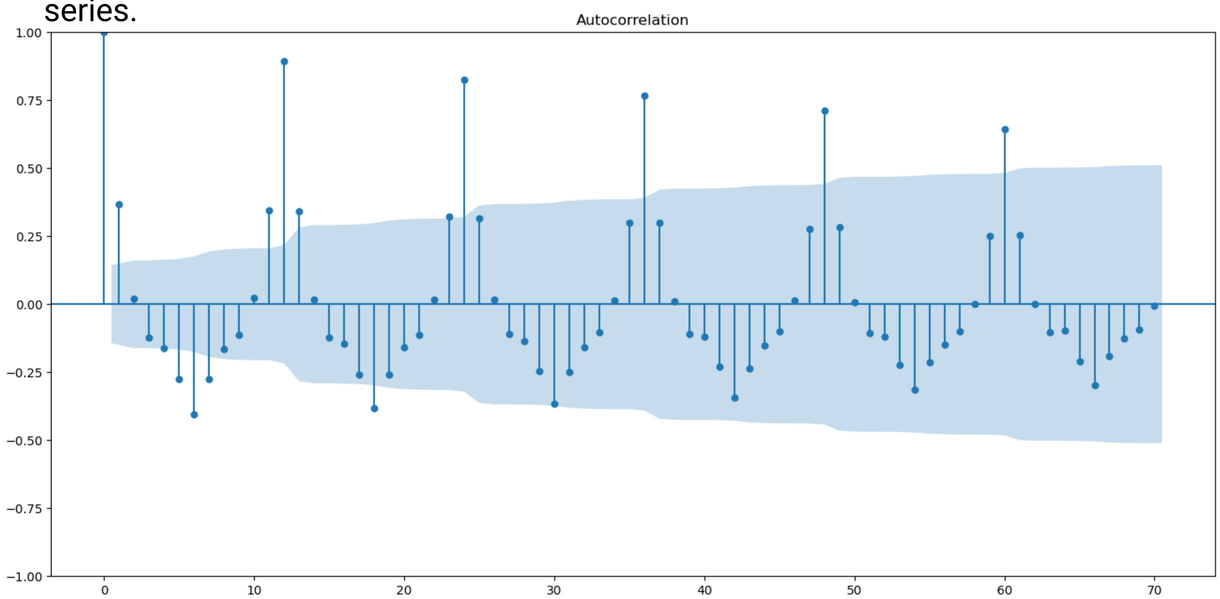

Figure 39: *Auto Correlation plot - Difference of 0 - Sparkling Wine Time Series.*

$$q = 1.$$

- Partial Autocorrelation Plot and Differenced Partial Autocorrelation Plot for Sparkling Wine Time series:

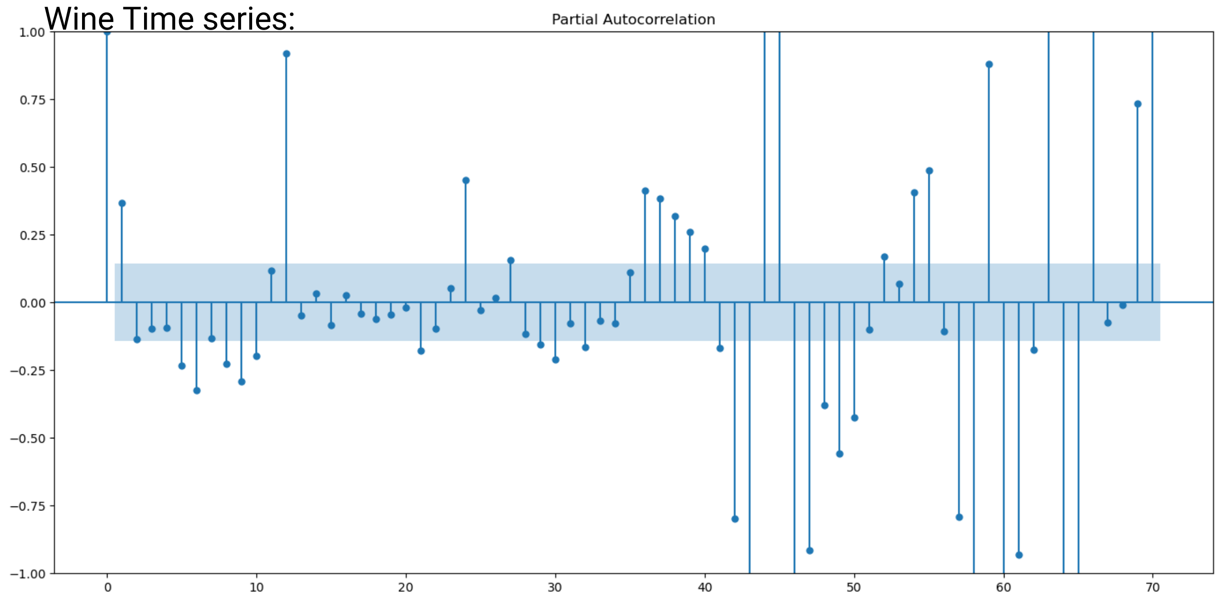

Figure 40: *Partial Auto Correlation plot - Difference of 0 - Sparkling Wine Time Series.*

$$p = 1.$$

- Autocorrelation Plot and Differenced Autocorrelation Plot for Rose Wine Time series.

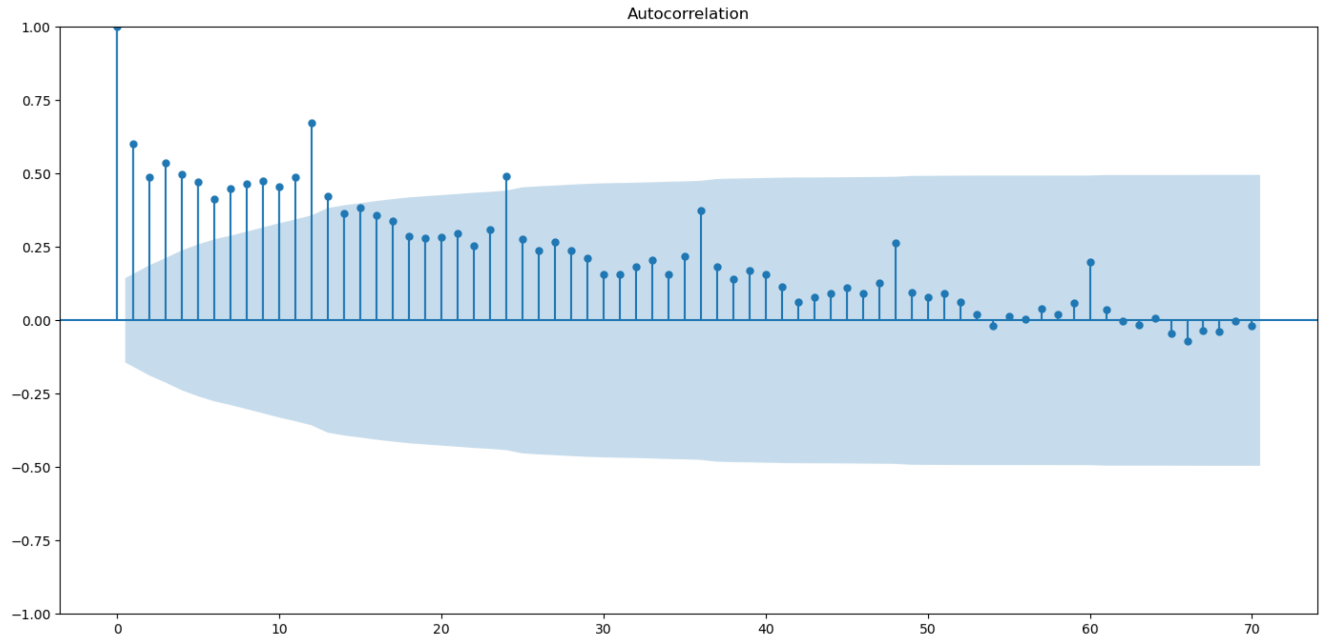

Figure 41: *Auto Correlation plot - Difference of 0 - Rose Wine Time Series.*

$$q = 13.$$

- Partial Autocorrelation Plot and Differenced Partial Autocorrelation Plot for Rose Wine Time series:

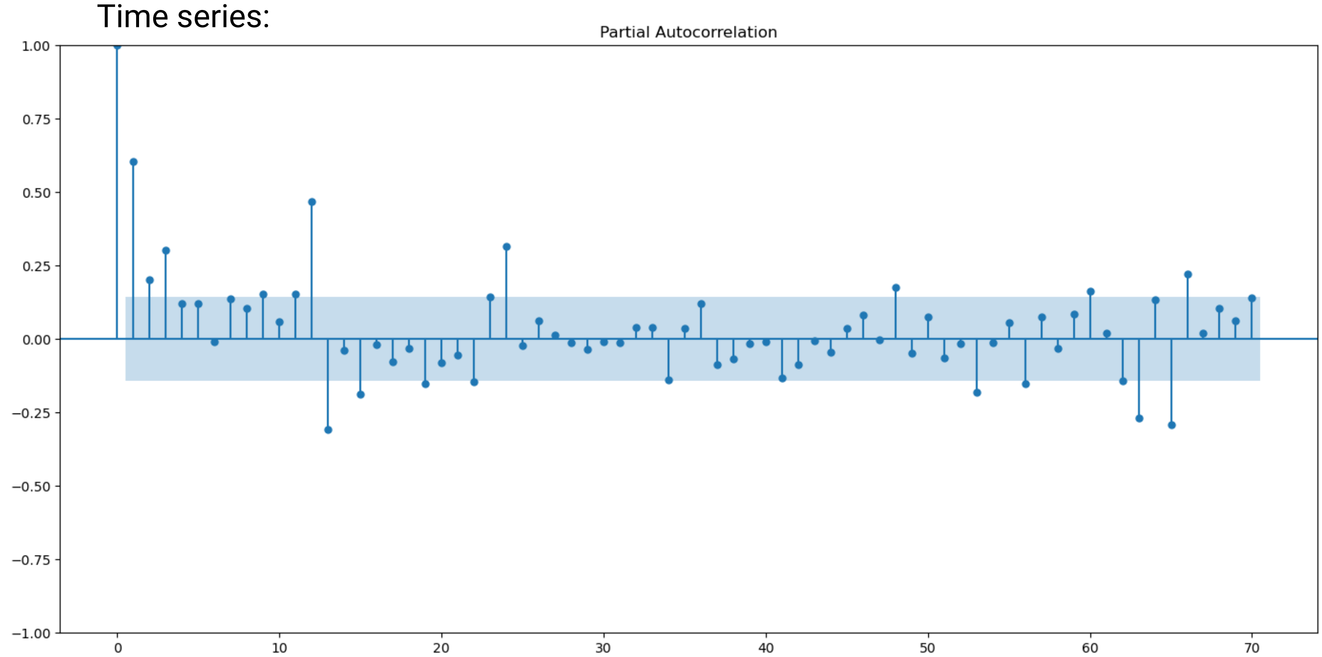

Figure 42: *Partial Auto Correlation plot - Difference of 0 - Rose Wine Time Series.*

$$p = 3.$$

- Autocorrelation Plot and Differenced Autocorrelation Plot for Sparkling Wine Time series of Seasonality 12 (Differenced by 12):

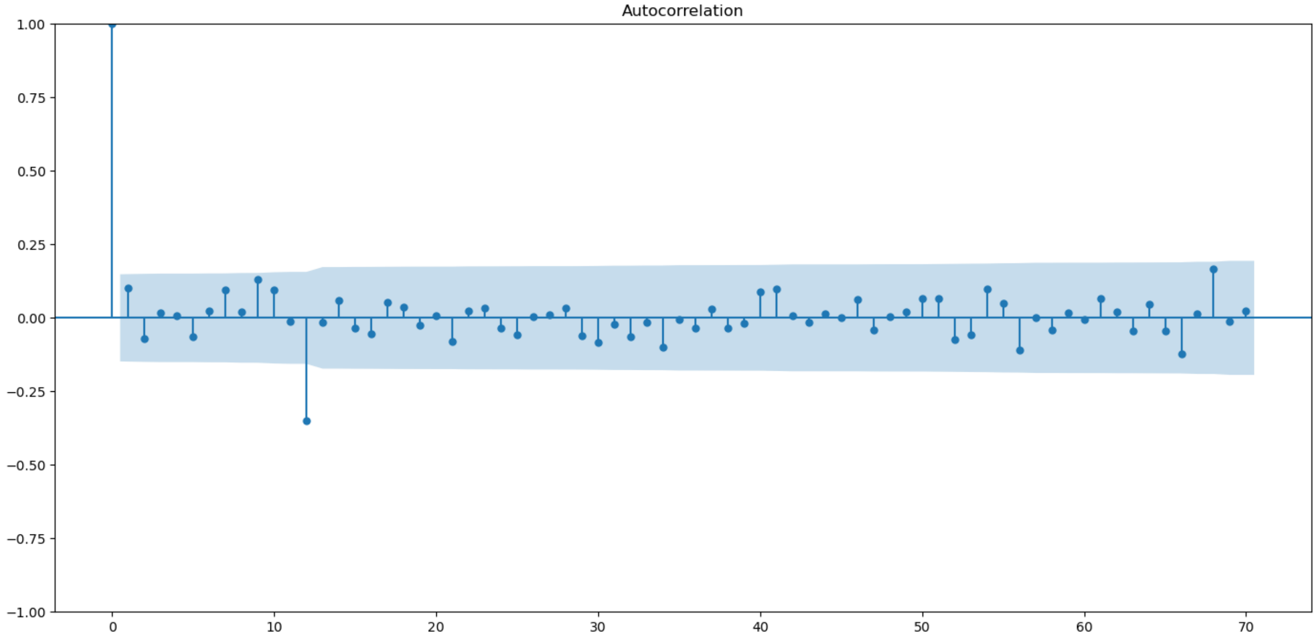

Figure 43: *Auto Correlation plot - Difference of 12 - Sparkling Wine Time Series.*

$$Q = 0.$$

- Partial Autocorrelation Plot and Differenced Partial Autocorrelation Plot for Sparkling Wine Time series (Differenced by 12):

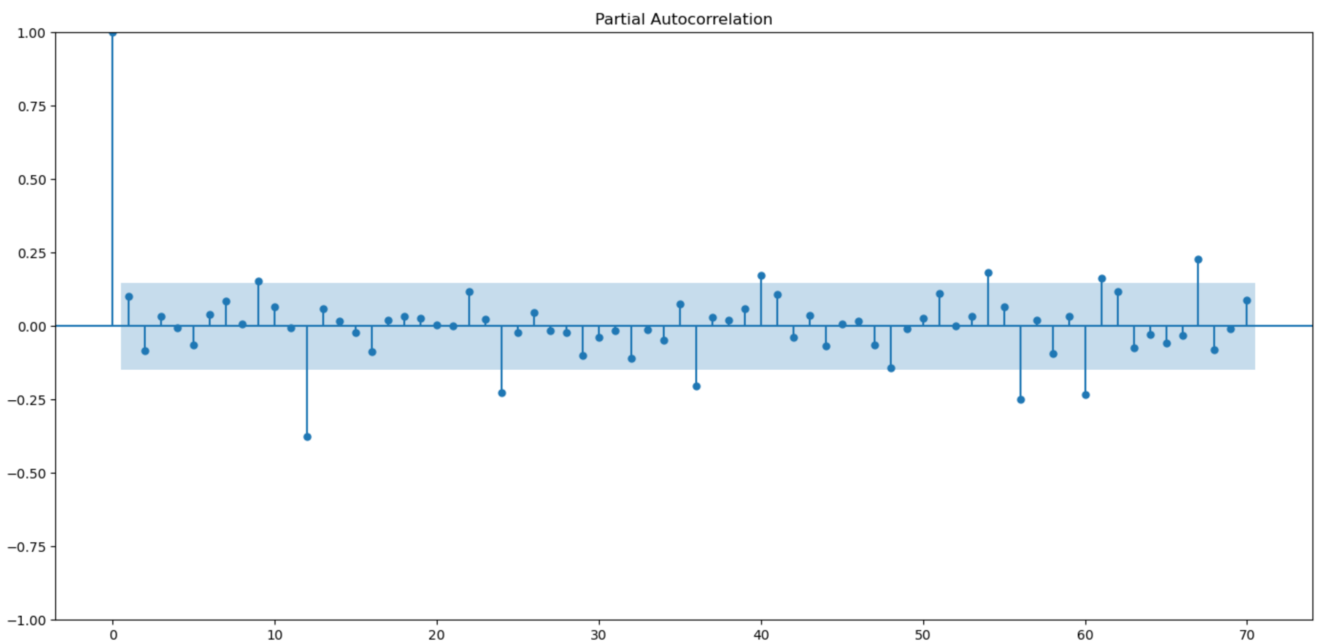

Figure 44: *Partial Auto Correlation plot - Difference of 12 - Sparkling Wine Time Series.*

$$P = 0.$$

- Autocorrelation Plot and Differenced Autocorrelation Plot for Rose Wine Time series of Seasonality 12 (Differenced by 12):

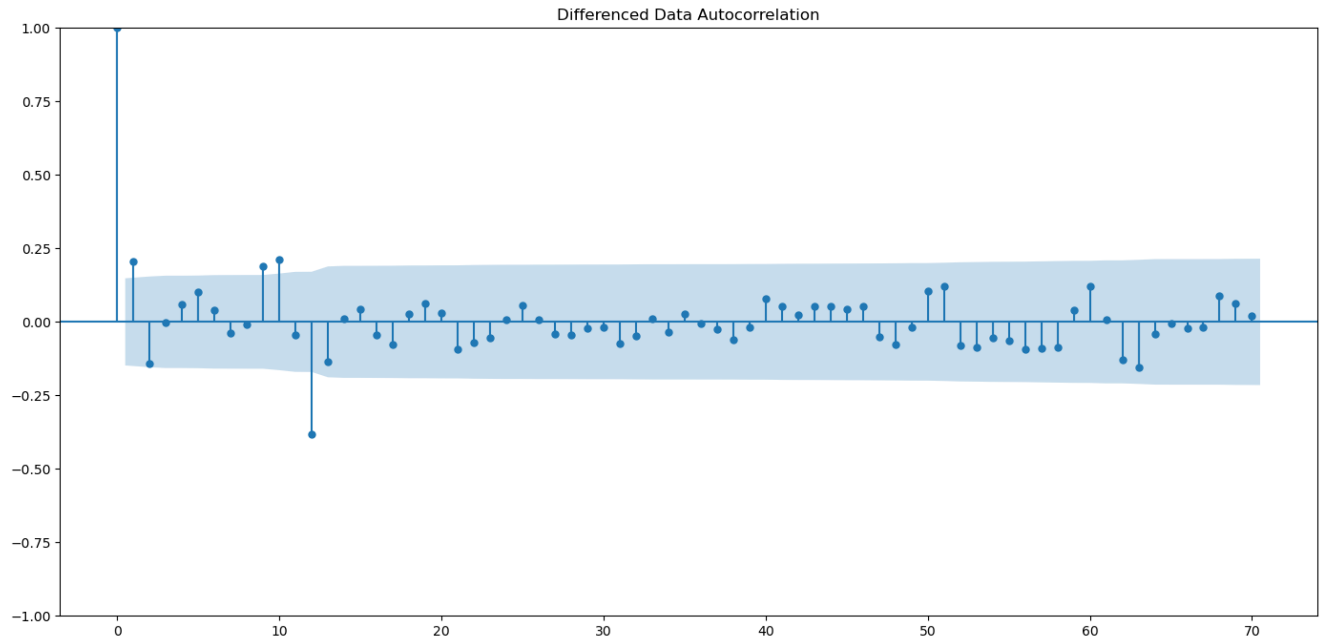

Figure 45: *Auto Correlation plot - Difference of 12 - Rose Wine Time Series.*

$$Q = 1.$$

- Partial Autocorrelation Plot and Differenced Partial Autocorrelation Plot for Rose Wine Time series (Differenced by 12):

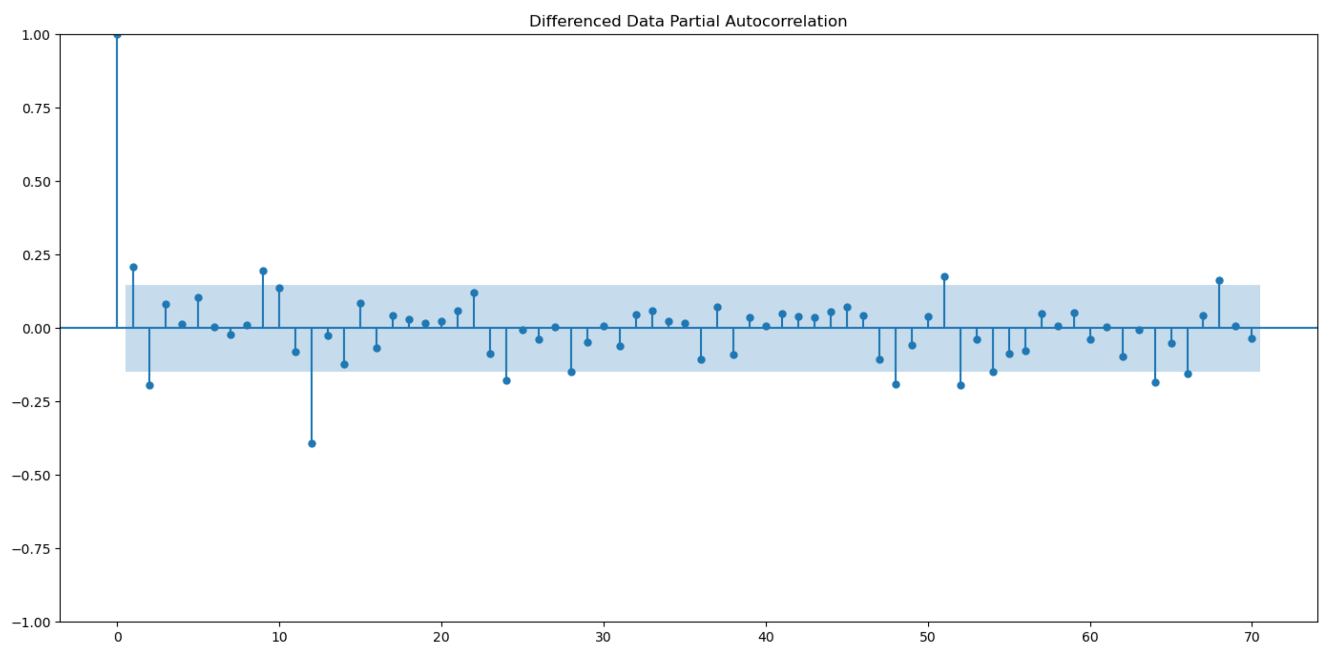

Figure 46: *Partial Auto Correlation plot - Difference of 12 - Sparkling Wine Time Series.*

$$P = 2.$$

ARIMA MODEL

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Although for our time series it requires the use of SARIMA Model, we will however also Model with ARIMA comparison purposes.

We shall try to build an ARIMA model in 2 different methods:

- **Auto AIC**: Where we gather AIC values for iterated Values of p, q and d and generate the model based on the lowest AIC score for a particular combination value of (p, q, d).

- **Manual**: In this method we manually generate the value of (p, q, d) from the plots of Auto Correlation and Partial Auto Correlation.

Therefore proceeding, **for Sparkling wine time series data**:

- The Lowest AIC values and their Combinations:
-

	param	AIC
10	(2, 1, 2)	2213.51
15	(3, 1, 3)	2221.46
14	(3, 1, 2)	2230.78
11	(2, 1, 3)	2232.90
9	(2, 1, 1)	2233.78

Figure 47: *Lowest AIC Values - ARIMA - Sparkling Wine Time Series.*

As seen above it seems for p = 2, d = 1 and q = 2 we have the lowest AIC. Hence, we generate the Model for this value and the resultant summary Report.

Note about the summary Report:

- Top box - Gives the overall summary.
- Middle box - Gives the coefficients and tells if the variables are significant.
- Lower box (Roots) - If the model is stable and good for forecasting. Between Imaginary & Real you draw a circle and check if the values/roots are inside the circle.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sat, 18 Feb 2023   AIC                           2213.509
Time:                        13:34:49   BIC                           2227.885
Sample:                    01-31-1980   HQIC                          2219.351
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.781      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.218      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================
```

Figure 48: *Summary report for (2,1,2) - ARIMA - Sparkling Wine Time Series.*

And the resultant **RMSE** calculated for this is 1299.98.

- For the manual method looking at the Autocorrelation and partial autocorrelation in Figure 39 and Figure 40 the value of (p, d, q) = (1, 1, 1). The resultant summary Report:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -1114.878
Date:                Sat, 18 Feb 2023   AIC                           2235.755
Time:                        13:34:49   BIC                           2244.381
Sample:                    01-31-1980   HQIC                          2239.260
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4494      0.043     10.366      0.000       0.364       0.534
ma.L1         -0.9996      0.102     -9.811      0.000      -1.199      -0.800
sigma2      1.401e+06   7.57e-08   1.85e+13      0.000     1.4e+06     1.4e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.50   Jarque-Bera (JB):                10.42
Prob(Q):                              0.48   Prob(JB):                         0.01
Heteroskedasticity (H):               2.64   Skew:                             0.46
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.03
===================================================================================
```

Figure 49: *Summary report for (1,1,1) - ARIMA - Sparkling Wine Time Series.*

And the resultant **RMSE** calculated for this is 1319.94.

The diagnostic for each method is given as:

- Auto:

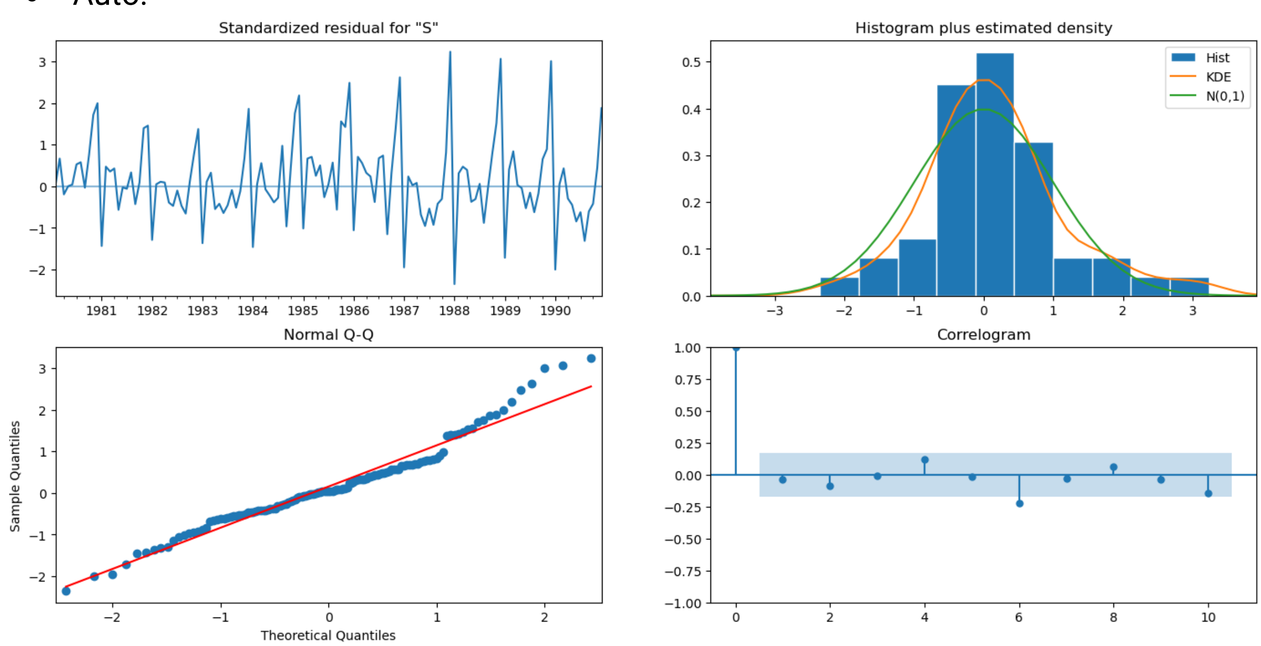

Figure 50: *Visualization Summary for (2,1,2) - ARIMA - Sparkling Wine Time Series.*

- Manual:

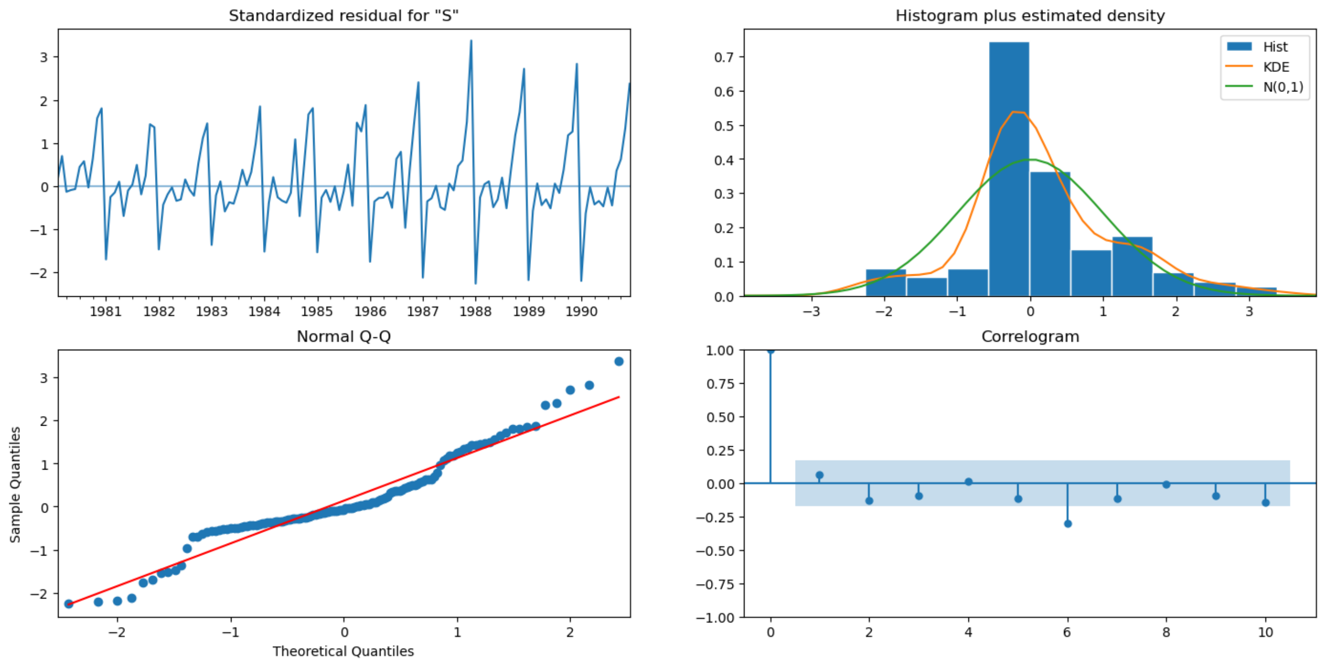

Figure 51: *Visualization Summary for (1,1,1) - ARIMA - Sparkling Wine Time Series.*

Therefore proceeding, **for Rose wine time series data**:

- The Lowest AIC values and their Combinations

	param	AIC
11	(2, 1, 3)	1274.70
15	(3, 1, 3)	1278.66
2	(0, 1, 2)	1279.67
6	(1, 1, 2)	1279.87
3	(0, 1, 3)	1280.55
5	(1, 1, 1)	1280.57

Figure 52: *Lowest AIC Values - ARIMA - Rose Wine Time Series.*

As seen above it seems for p = 2, d = 1 and q = 3 we have the lowest AIC. Hence, we generate the Model for this value and the resultant summary Report.

- Summary report:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 19 Feb 2023   AIC                           1274.696
Time:                        14:48:00   BIC                           1291.947
Sample:                    01-31-1980   HQIC                          1281.706
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6774      0.084    -20.019      0.000      -1.842      -1.513
ar.L2         -0.7283      0.084     -8.691      0.000      -0.893      -0.564
ma.L1          1.0446      0.593      1.762      0.078      -0.117       2.206
ma.L2         -0.7716      0.130     -5.929      0.000      -1.027      -0.517
ma.L3         -0.9043      0.537     -1.685      0.092      -1.956       0.147
sigma2       859.7707    497.996      1.726      0.084    -116.284    1835.825
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.37
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

Figure 53: *Summary report for (2,1,3) - ARIMA - Rose Wine Time Series.*

And the resultant **RMSE** calculated for this is 36.84.

- For the manual method looking at the Autocorrelation and partial autocorrelation in Figure 41 and Figure 42 the value of (p, d, q) = (3, 1, 13). The resultant summary Report:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                ARIMA(3, 1, 13)   Log Likelihood                -616.154
Date:                Sat, 18 Feb 2023   AIC                           1266.309
Time:                        13:42:26   BIC                           1315.187
Sample:                    01-31-1980   HQIC                          1286.170
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0376      0.181     -0.207      0.836      -0.393       0.318
ar.L2         -0.2489      0.188     -1.326      0.185      -0.617       0.119
ar.L3          0.0163      0.156      0.104      0.917      -0.290       0.322
ma.L1         -0.6417      0.291     -2.205      0.027      -1.212      -0.071
ma.L2         -0.1309      0.501     -0.261      0.794      -1.113       0.851
ma.L3          0.0017      0.578      0.003      0.998      -1.131       1.134
ma.L4         -0.1826      0.306     -0.597      0.551      -0.782       0.417
ma.L5          0.1800      0.395      0.456      0.649      -0.594       0.954
ma.L6         -0.1358      0.177     -0.767      0.443      -0.483       0.211
ma.L7         -0.0249      0.278     -0.090      0.929      -0.569       0.520
ma.L8         -0.0049      0.269     -0.018      0.986      -0.533       0.523
ma.L9          0.2629      0.337      0.781      0.435      -0.397       0.923
ma.L10        -0.0144      0.485     -0.030      0.976      -0.965       0.936
ma.L11        -0.1812      0.370     -0.490      0.624      -0.906       0.544
ma.L12         0.6710      0.380      1.768      0.077      -0.073       1.415
ma.L13        -0.6336      0.244     -2.598      0.009      -1.112      -0.156
sigma2       609.8095    205.619      2.966      0.003     206.804    1012.815
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                31.78
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.35   Skew:                             0.78
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.84
===================================================================================
```

Figure 54: *Summary report for (3,1,13) - ARIMA - Rose Wine Time Series.*

And the resultant **RMSE** calculated for this is 35.

D

The diagnostic for each method is given as:

- Auto:

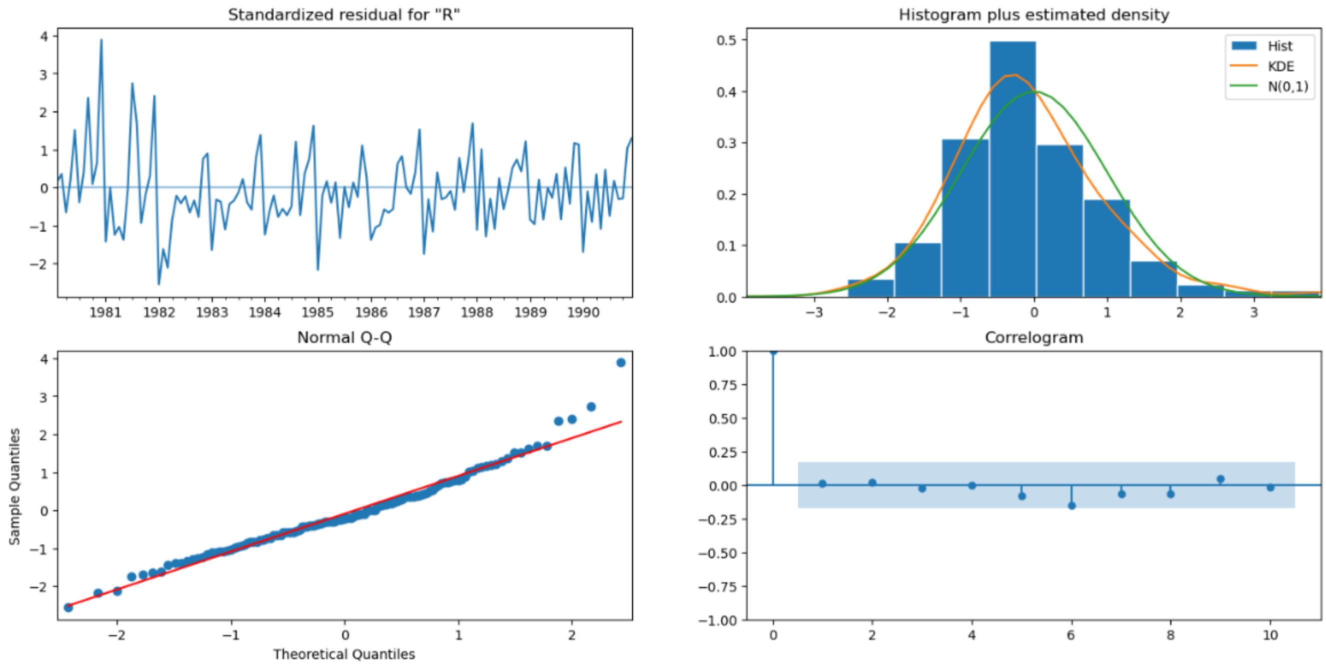

Figure 55: *Visualization Summary for (2,1,3) - ARIMA - Rose Wine Time Series*

- Manual:

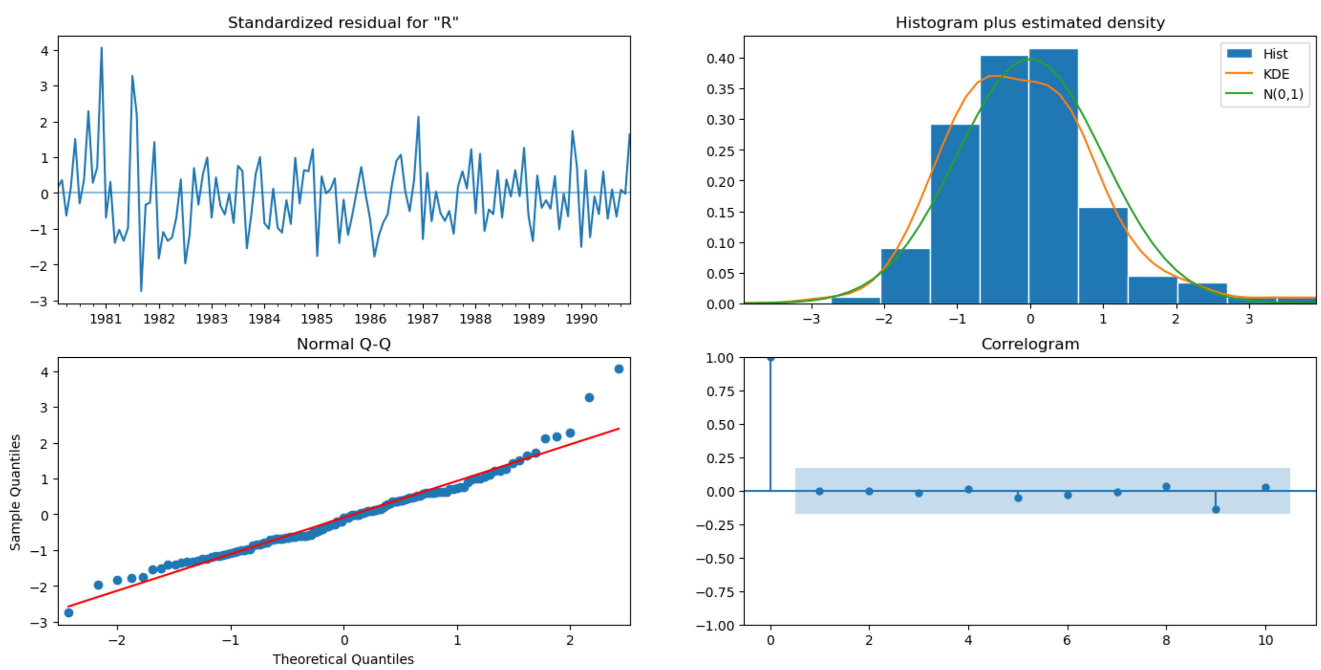

Figure 56: *Visualization Summary for (3,1,13) - ARIMA - Rose Wine Time Series*

SARIMA MODEL

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

SARIMA contains both Trend and Seasonal Elements which need to be configured i.e.,

TREND ELEMENTS

There are three trend elements that require configuration.
They are the same as the ARIMA model; specifically:

- p: Trend autoregression order.
- d: Trend difference order.
- q: Trend moving average order.

SEASONAL ELEMENTS

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

- P: Seasonal autoregressive order.
- D: Seasonal difference order.
- Q: Seasonal moving average order.
- m: The number of time steps for a single seasonal period.

A SARIMA Model notation can be given as:

$$SARIMA(p, d, q)(P, D, Q, m)$$

Note: *We will build ARIMA and SARIMA models for seasonality of 6 as well for comparison purposes, we will not discuss much about the process (refer to notebook for the work on seasonality of 6)*

Similar to ARIMA we will build SARIMA with both the auto and manual method.

Therefore proceeding, **for Sparkling wine time series data**:

- With a Seasonality of 12 the Lowest AIC values and their Combinations:

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.58
53	(1, 1, 2)	(2, 0, 2, 12)	1555.93
26	(0, 1, 2)	(2, 0, 2, 12)	1557.12
23	(0, 1, 2)	(1, 0, 2, 12)	1557.16
77	(2, 1, 2)	(1, 0, 2, 12)	1557.34

Figure 57: *Lowest AIC Values - SARIMA - Sparkling Wine Time Series.*

As seen above it seems for p = 1, d = 1, q = 2 and P = 1, D = 0, Q = 2, m = 12 we have the lowest AIC. Hence, we generate the Model for this value and the resultant summary Report.

- Summary report:

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sat, 18 Feb 2023   AIC                           1555.584
Time:                                    13:36:03   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6281      0.255     -2.463      0.014      -1.128      -0.128
ma.L1         -0.1041      0.225     -0.463      0.643      -0.545       0.337
ma.L2         -0.7276      0.154     -4.734      0.000      -1.029      -0.426
ar.S.L12       1.0439      0.014     72.840      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1355      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.400      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
===================================================================================
```

Figure 58: *Summary report for (1,1,2) (1,0,2,12) - SARIMA - Sparkling Wine Time Series.*

And the resultant **RMSE** calculated for this is 528.59.

- For the manual method looking at the Autocorrelation and partial autocorrelation in Figure 43 and Figure 44 the value of (p, d, q) = (1, 1, 1) and (P, D, Q, m) = (0, 0, 0, 12). The resultant summary Report:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  132
Model:               SARIMAX(1, 1, 1)   Log Likelihood               -1099.467
Date:                Sun, 19 Feb 2023   AIC                           2204.934
Time:                        14:47:55   BIC                           2213.513
Sample:                             0   HQIC                          2208.420
                                - 132
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4324      0.106      4.074      0.000       0.224       0.640
ma.L1         -0.9865      0.080    -12.291      0.000      -1.144      -0.829
sigma2      1.756e+06   2.14e+05      8.215      0.000    1.34e+06    2.17e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.74   Jarque-Bera (JB):                11.75
Prob(Q):                              0.39   Prob(JB):                         0.00
Heteroskedasticity (H):               2.69   Skew:                             0.55
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.00
===================================================================================
```

Figure 59: *Summary report for (1,1,1) (0,0,0,12) - SARIMA - Sparkling Wine Time Series.*

And the resultant **RMSE** calculated for this is 1325.34.

Therefore proceeding, **for Rose wine time series data**:

- With a Seasonality of 12 the Lowest AIC values and their Combinations:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.94
80	(2, 1, 2)	(2, 0, 2, 12)	890.67
69	(2, 1, 1)	(2, 0, 0, 12)	896.52
53	(1, 1, 2)	(2, 0, 2, 12)	896.69
78	(2, 1, 2)	(2, 0, 0, 12)	897.35

Figure 60: *Lowest AIC Values - SARIMA - Rose Wine Time Series*

As seen above it seems for p = 0, d = 1, q = 2 and P = 2, D = 0, Q = 2, m = 12 we have the lowest AIC. Hence, we generate the Model for this value and the resultant summary Report.

- Summary report:

```
                                      SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 19 Feb 2023   AIC                            887.938
Time:                                    14:48:38   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                            - 132                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.766     -0.004      0.996    -372.778     371.092
ma.L2         -0.1573     29.813     -0.005      0.996     -58.590      58.276
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.77e+04      0.005      0.996   -9.32e+04    9.37e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================
```

Figure 61: *Summary report for (0,1,2) (2,0,2,12) - SARIMA - Rose Wine Time Series.*

And the resultant **RMSE** calculated for this is 26.95.

- For the manual method looking at the Autocorrelation and partial autocorrelation in Figure 45 and Figure 46 the value of (p, d, q) = (3, 1, 11) and (P, D, Q, m) = (2, 0, 1, 12). The resultant summary Report:

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  187
Model:             SARIMAX(3, 1, 11)x(2, 0, [1], 12)   Log Likelihood                -636.803
Date:                            Sun, 19 Feb 2023   AIC                           1309.605
Time:                                    14:48:50   BIC                           1364.846
Sample:                                         0   HQIC                          1332.038
                                            - 187
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2056      1.356      0.152      0.879      -2.451       2.863
ar.L2         -0.5893      0.758     -0.778      0.437      -2.074       0.896
ar.L3         -0.2379      1.045     -0.228      0.820      -2.287       1.811
ma.L1         -0.6339      1.142     -0.555      0.579      -2.873       1.605
ma.L2         -0.0184      0.988     -0.019      0.985      -1.954       1.918
ma.L3          0.6271      0.415      1.512      0.131      -0.186       1.440
ma.L4         -1.0986      1.159     -0.948      0.343      -3.369       1.172
ma.L5          0.5559      1.894      0.293      0.769      -3.156       4.268
ma.L6         -0.5937      1.361     -0.436      0.663      -3.261       2.074
ma.L7          0.5525      1.186      0.466      0.641      -1.772       2.877
ma.L8         -0.2429      1.075     -0.226      0.821      -2.349       1.863
ma.L9          0.2893      0.619      0.467      0.640      -0.925       1.504
ma.L10        -0.1242      0.481     -0.258      0.796      -1.068       0.819
ma.L11         0.0308      0.221      0.139      0.889      -0.402       0.464
ar.S.L12       0.4508      0.094      4.792      0.000       0.266       0.635
ar.S.L24       0.2592      0.068      3.818      0.000       0.126       0.392
ma.S.L12       0.0296      0.124      0.238      0.812      -0.214       0.273
sigma2       116.7591     39.785      2.935      0.003      38.783     194.736
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 4.84
Prob(Q):                              0.95   Prob(JB):                         0.09
Heteroskedasticity (H):               0.30   Skew:                             0.29
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.63
===================================================================================
```

Figure 62: *Summary report for (3,1,11) (2,0,1,12) - SARIMA - Rose Wine Time Series.*

And the resultant **RMSE** calculated for this is 27.18.

PREDICTION ON FULL DATA

Using the best model we created earlier based on the RMSE Values we shall use it and predict for a period of 12 months into the for both Sparking Wine Time series and Rose Wine Time Series.

Let's look into the RMSE values to help us choose the proper model for prediction (Note: The final RMSE Value is also included in this table so that it will be easily accessible from list of tables in contents section rather than having to create a separate table and increase confusion):

- Sparkling Wine Time Series:

Table 3: *Table Showing RMSE values of All ARIMA and SARIMA models for Sparkling Wine Time series.*

MODELS	RMSE
Sparkling - Auto - SARIMA (1,1,2)(1,0,2,12)	528.5883
Sparkling - Prediction - 12 Months	539.9718
Sparkling - Auto - SARIMA (1,1,2)(2,0,2,6)	626.9047
Sparkling - Auto - ARIMA (2,1,2)	1299.98
Sparkling - Manual - ARIMA (1,1,1)	1319.937
Sparkling - Manual - SARIMA (1,1,1)(0,0,0,12)	1325.336
Sparkling - Max Error Possible	2403.78

- Rose Wine Time Series:

Table 4: *Table Showing RMSE values of All ARIMA and SARIMA models for Rose Wine Time series.*

MODELS	RMSE
Rose - Auto – SARIMA (1,1,2)(2,0,2,6)	26.15945
Rose - Auto - SARIMA (0,1,2)(2,0,2,12)	26.94902
Rose - Manual - SARIMA (3,1,11)(2,0,1,12)	27.17843
Rose - Prediction - 12 Months	27.67695
Rose - Manual - SARIMA (3,1,5)(1,0,1,6)	31.8131
Rose - Manual - ARIMA (3,1,13)	35.00425
Rose - Auto - ARIMA (2,1,3)	36.83851
Rose - Max Error Possible	104.9394

From the Tables above we can clearly see that **Sparkling - Auto - SARIMA (1,1,2)(1,0,2,12)** is the best model for Sparking wine time series and **Rose - Manual - SARIMA(0,1,2)(2,0,2,12)** is the best model for Rose wine Time series.

- Prediction for Sparkling wine Time series for 12 months:

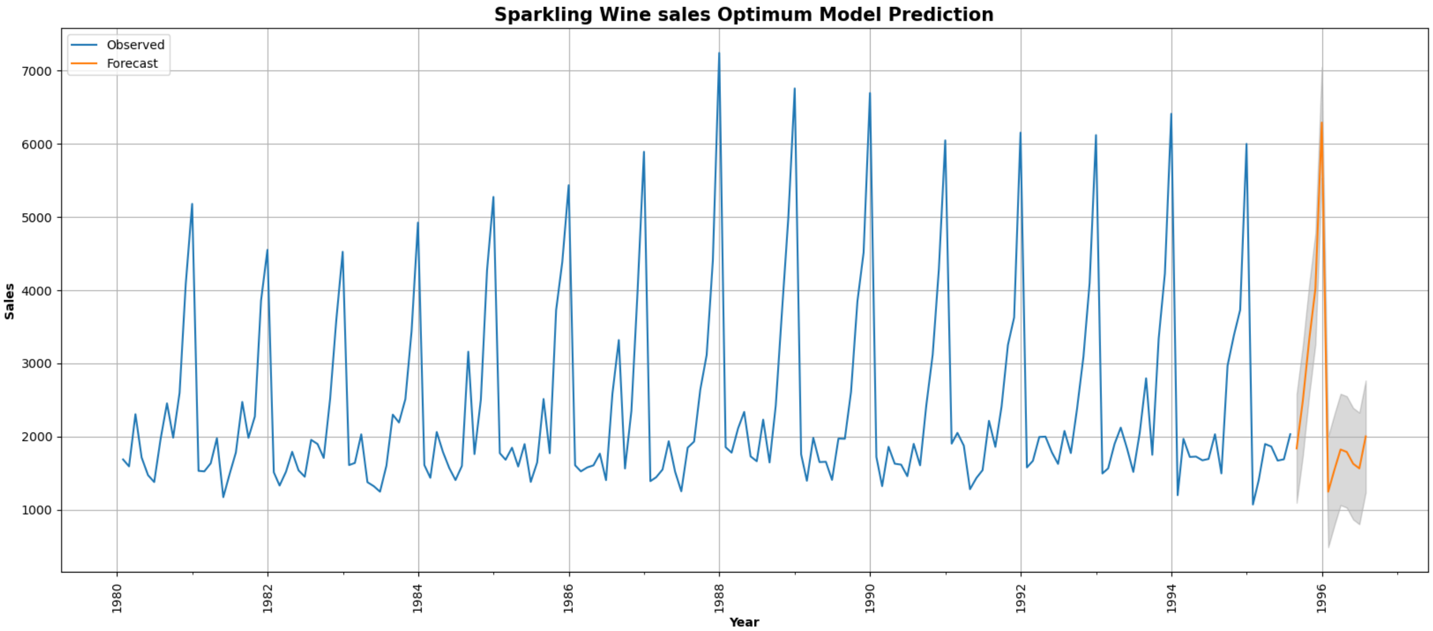

Figure 63: *Visualization of 12 Month Prediction of Sparkling Wine Time Series.*

- The summary Report for the same:

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  187
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)  Log Likelihood               -1173.413
Date:                            Sat, 18 Feb 2023  AIC                           2360.827
Time:                                    20:34:29  BIC                           2382.309
Sample:                                        0   HQIC                          2369.551
                                           - 187
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6608      0.242     -2.730      0.006      -1.135      -0.186
ma.L1         -0.2742      0.200     -1.368      0.171      -0.667       0.119
ma.L2         -0.8111      0.227     -3.572      0.000      -1.256      -0.366
ar.S.L12       1.0157      0.012     84.471      0.000       0.992       1.039
ma.S.L12      -1.3874      0.338     -4.102      0.000      -2.050      -0.724
ma.S.L24      -0.1460      0.146     -1.000      0.317      -0.432       0.140
sigma2      5.947e+04   1.84e+04      3.231      0.001    2.34e+04    9.55e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                27.47
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               1.03   Skew:                             0.52
Prob(H) (two-sided):                  0.93   Kurtosis:                         4.76
===================================================================================
```

Figure 64: *Summary Report of 12 Month Prediction of Sparkling Wine Time Series*

And the resultant **RMSE** calculated for this is 539.97.

- Prediction for Rose wine Time series for 12 months:

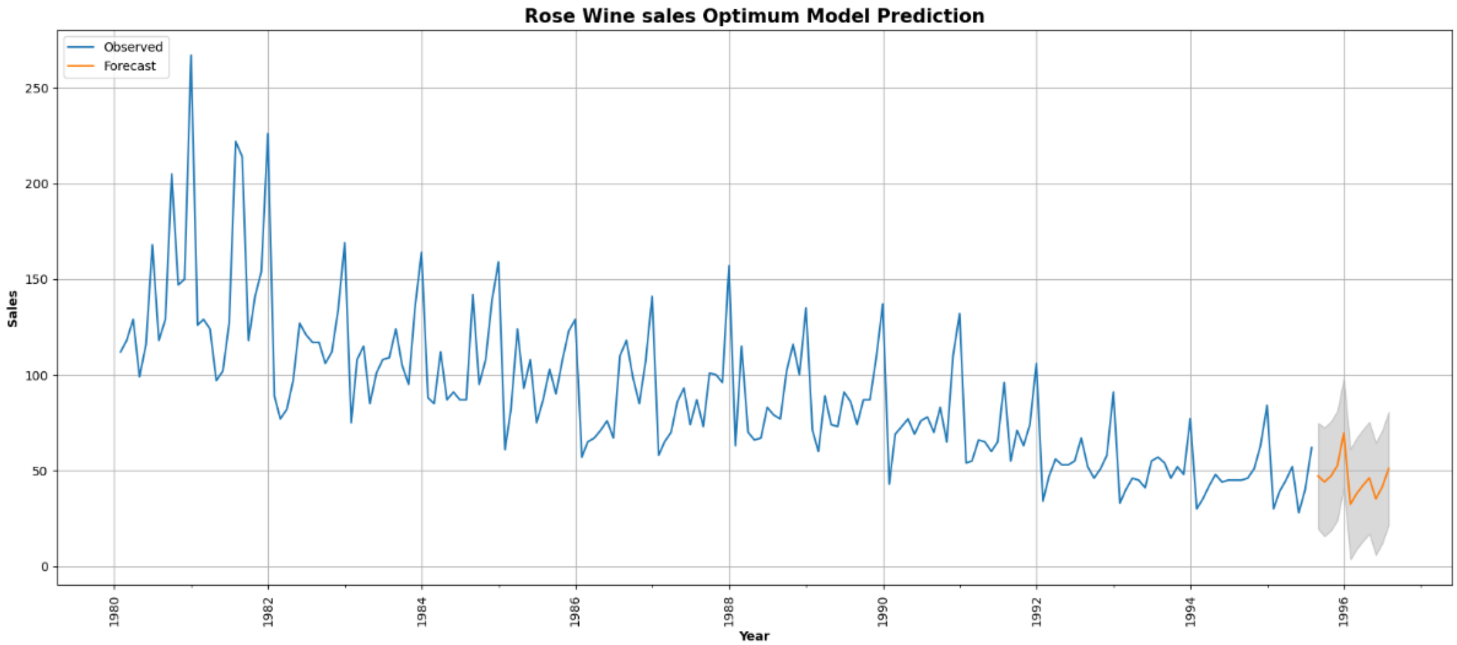

Figure 65: *Visualization of 12 Month Prediction of Rose Wine time series.*

The summary Report for the same:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                    187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                  -647.372
Date:                          Sun, 19 Feb 2023   AIC                             1308.744
Time:                                  16:38:45   BIC                             1330.226
Sample:                                       0   HQIC                            1317.468
                                          - 187
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7656      0.088     -8.739      0.000      -0.937      -0.594
ma.L2         -0.1389      0.081     -1.716      0.086      -0.298       0.020
ar.S.L12       0.3980      0.052      7.624      0.000       0.296       0.500
ar.S.L24       0.3369      0.049      6.831      0.000       0.240       0.434
ma.S.L12       0.0116      0.089      0.130      0.897      -0.163       0.187
ma.S.L24      -0.1476      0.099     -1.489      0.137      -0.342       0.047
sigma2       199.2829     21.436      9.297      0.000     157.269     241.297
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 8.74
Prob(Q):                              0.75   Prob(JB):                         0.01
Heteroskedasticity (H):               0.26   Skew:                             0.50
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.57
===================================================================================
```

Figure 66: *Summary Report of 12 Month Prediction of Rose Wine Time Series*

And the resultant **RMSE** calculated for this is 27.67.

CONCLUSION AND RECOMMENDATION

SPARKLING WINE TIME SERIES

- Considering ALL the models we built it would appear that the best performing model for sparkling wine time series is a Holt – Winter's Model (TES) **α = 0.111, β = 0.049** and **γ = 0.362.** with a RMSE of **402.9362** closely followed by SARIMA (1,1,2)(1,0,2,12) with a RMSE of **528.5883**.

- Based on the Prediction of sparkling wine time series it would appear that there is not much change in trend or seasonality to account for, Hence there is no need for taking and measures for mitigation of loss in sales as there maybe none.

ROSE WINE TIME SERIES

- Considering ALL the models we built it appears that the best model is the 2 point Moving average window with RMSE **11.52**, but this model can be hardly used for prediction and the next best performing model which can be used for prediction is the Holt – Winter's Model (TES) α = 0.071, β = 0.045, γ = 8.38e-5, with and RMSE **20.12** both of which performed extremely well compared to any of the ARIMA/SARIMA models whose best performing model was SARIMA (0,1,2)(2,0,2,12) of RMSE **26.95**.

- Based on the Prediction of Rose wine time series there appears to be a downward trend in sales therefore measures must be taken to mitigate this loss. Also looking at the data this trend has been occurring for quite a few years which means whatever measure are taken must be severe and completely different from what was done in the past, if the measures are not extreme and novel it is possible with quite a high certainty the downward trend in sales will continue.