Time Series Forecasting Business Report

- **Name**: Hemant Patidar
- **Batch**: PGPDSBA Online Sep_A 2021
- **Date**: 24/03/2022

Table of Contents

List of Figures

List of Tables

Wine Data

Executive Summary

The data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines.

As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Introduction

Purpose of our exercise would be to forecast sales in 20th century for wine data.

Data Description

- **YearMonth** – Month & year of wine sales
- **Sparkling/Rose** – Sparkling or Rose sales data respectively in the file

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

Head & Tail of the Data:

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

	YearMonth	Sparkling
182	1995-03	1897
183	1995-04	1862
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

Table 1 - Head & Tail of Sparkling Wine Data

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

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Table 2 - Head & Tail of Rose Wine Data

Data Information:

- We have 187 rows and 2 columns in each data frame
- We have 2 null values in Rose wine data
- The sales data was given from January 1980 to July 1991

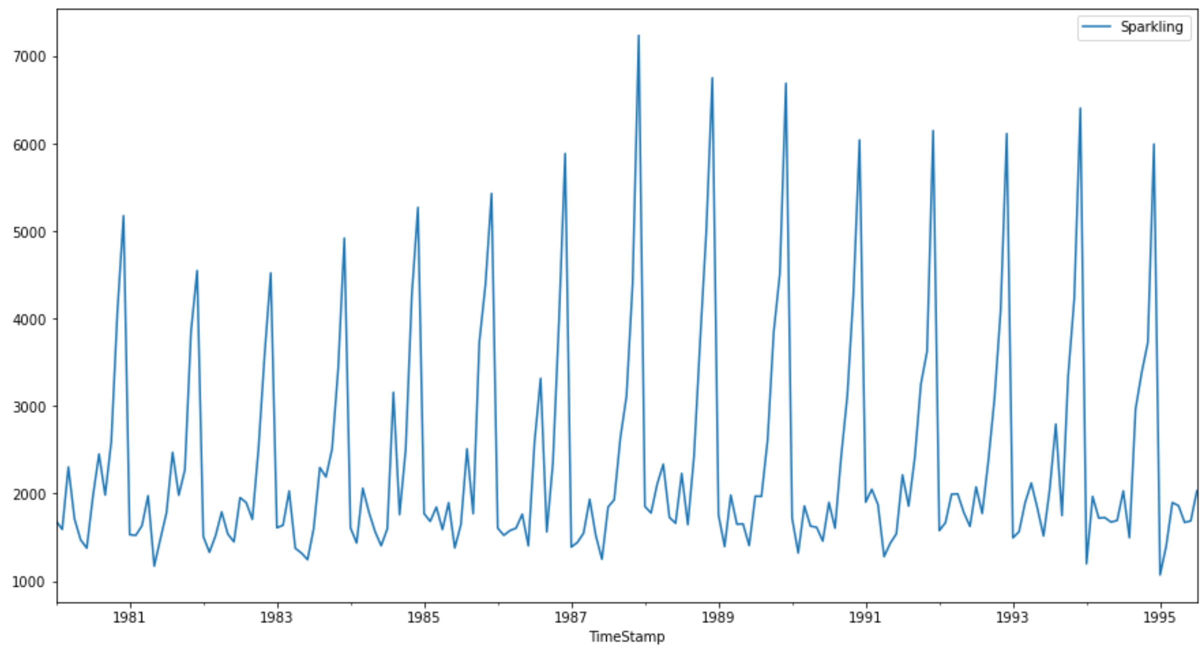


Figure I - Sparkling Wine Data Plot (Time Series)

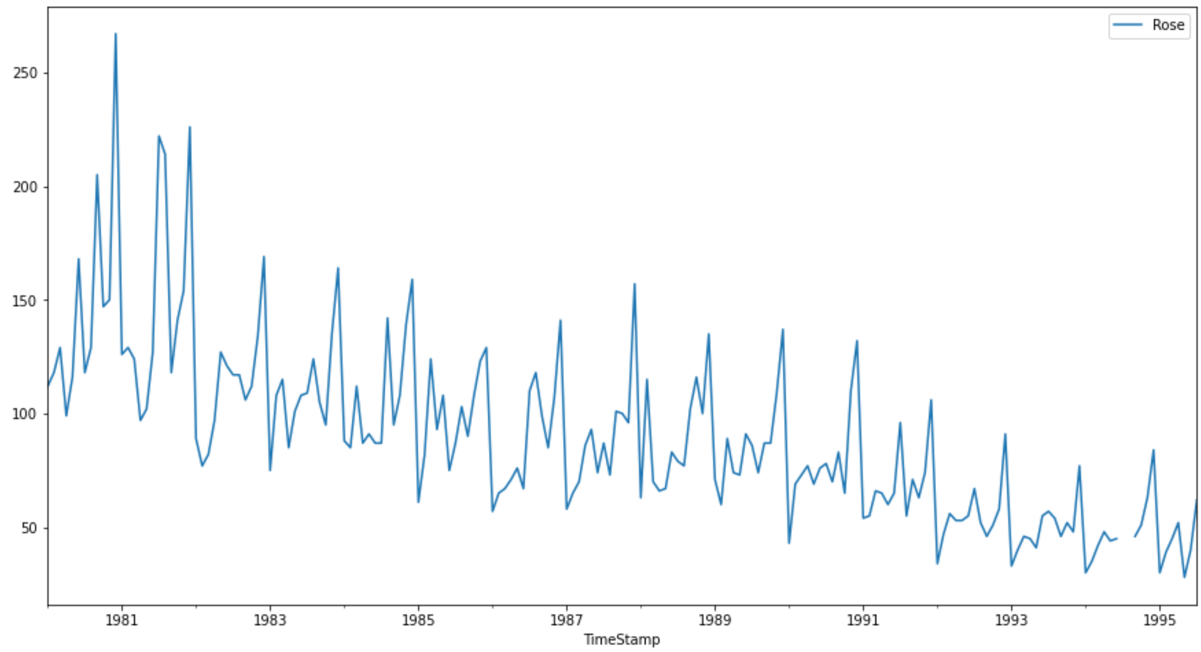


Figure II - Rose Wine Data Plot (Time Series)

As we can see from above, Rose data seems disconnected near end of 1994 (July & August), we are missing 2 values here and imputed them using linear interpolation method.

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

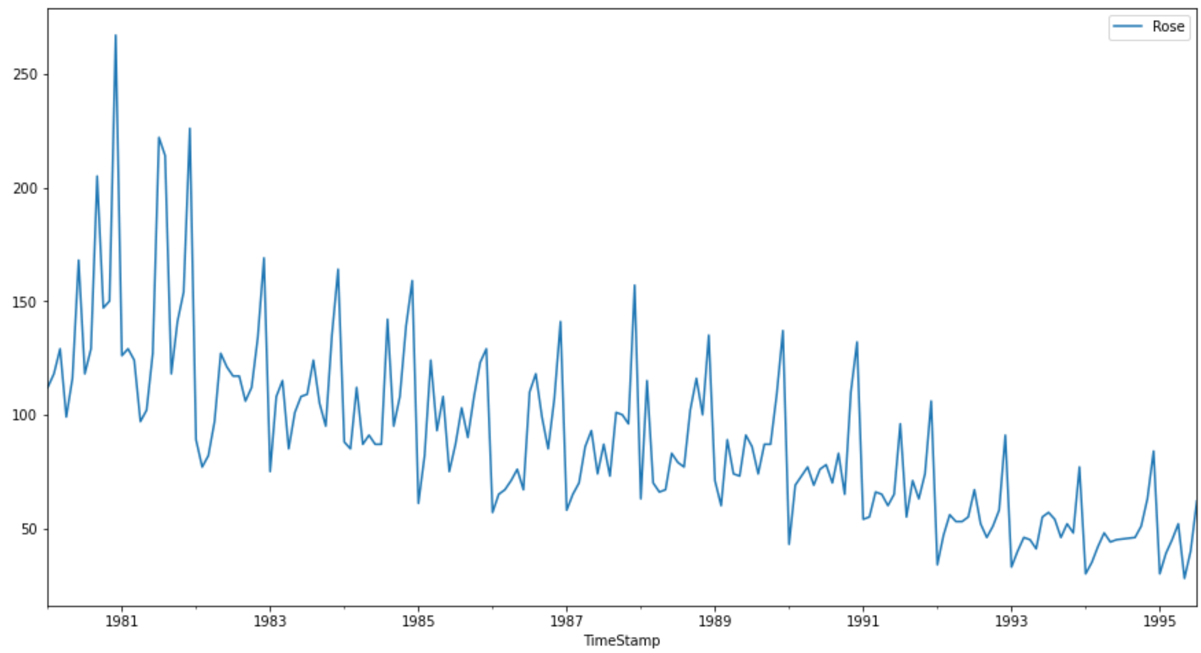


Figure III - Rose Wine Time Series (After Missing values Imputation)

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

	Rose
count	187.000000
mean	89.914439
std	39.238325
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Table 3 - Describing Sales Data (Sparkling & Rose Wine)

- The average Sparkling wine sales over the years is 2402.41, where Rose wine sales is 89.91
- Minimum Sparkling wine sale was 1070 and maximum sale was 7242
- Minimum Rose wine sale was 28 and maximum sale was 267

Time Series Decomposition

A time series can have 3 components – Trend, Seasonality & Residuals (Error), decomposition of them helps identifying impact/presence of each composition.

Additive Decomposition

If Seasonality has constant impact over time series, additive decomposition may help visualizing that.

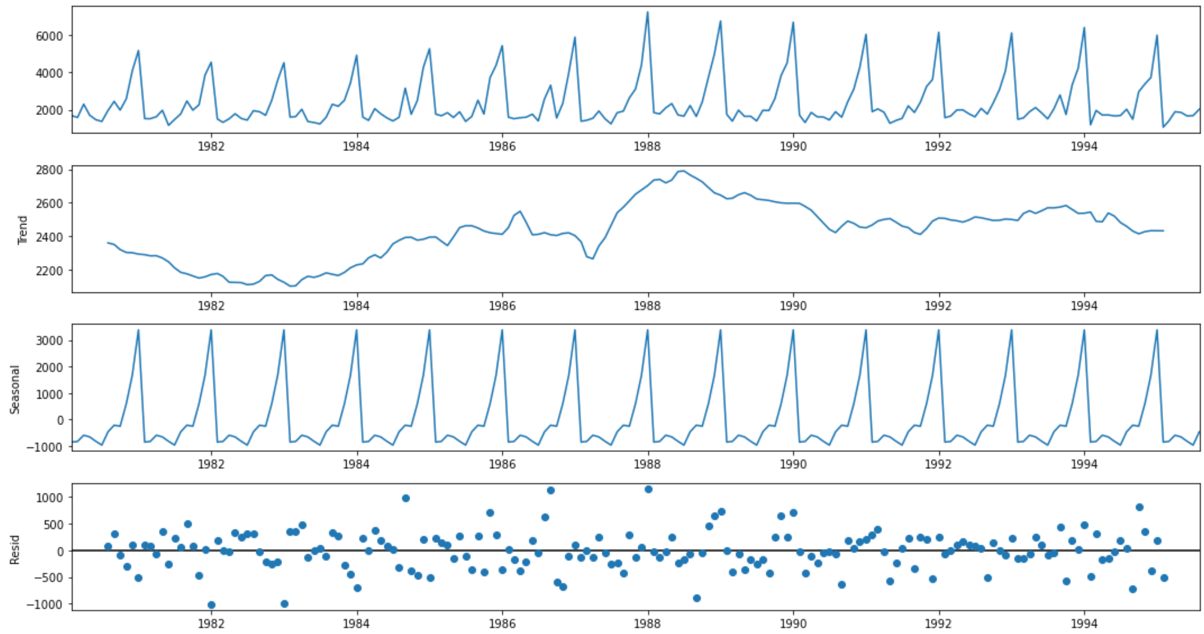


Figure IV - TS Decomposition - Sparkling (Additive)

The time series does not show a trend but a dynamic seasonality can be seen, as peaks every year are not looking same, also the residuals seem scattered. Hence, we will decompose this time series from multiplicative decomposition.

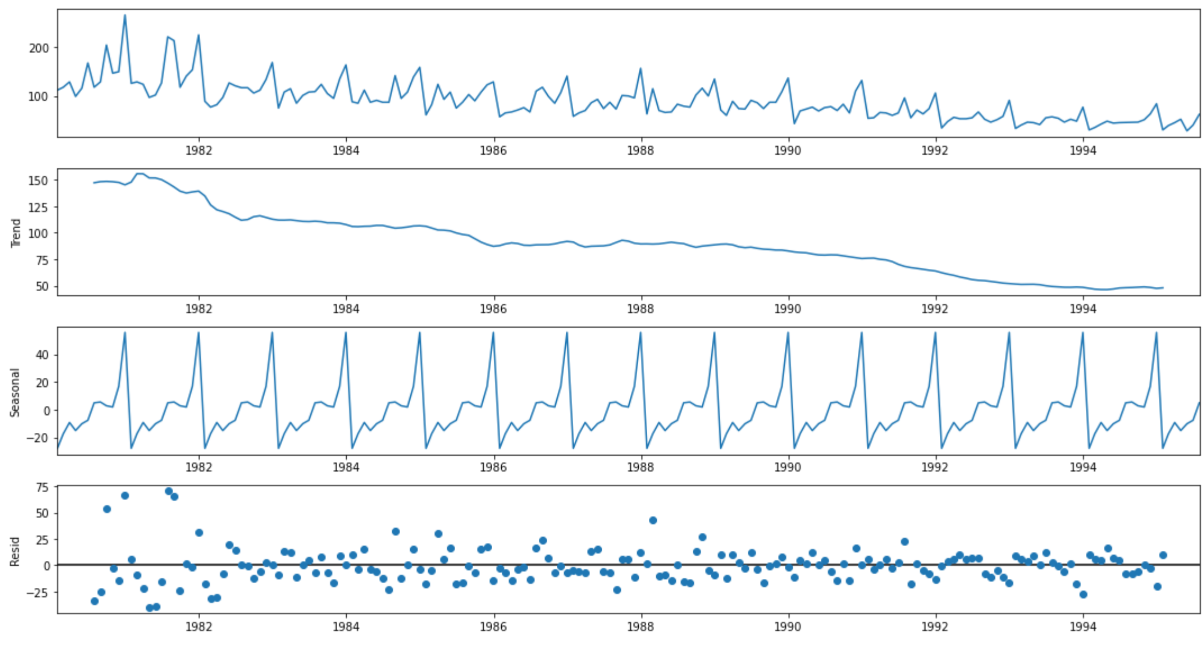


Figure V - TS Decomposition - Rose (Additive)

The time series shows a down trend over the years and seasonality doesn't seem to have constant effect, we can perform multiplicative decomposition.

Multiplicative Decomposition

If Seasonality has increased/decreased impact over time series, multiplicative decomposition may help visualizing that.

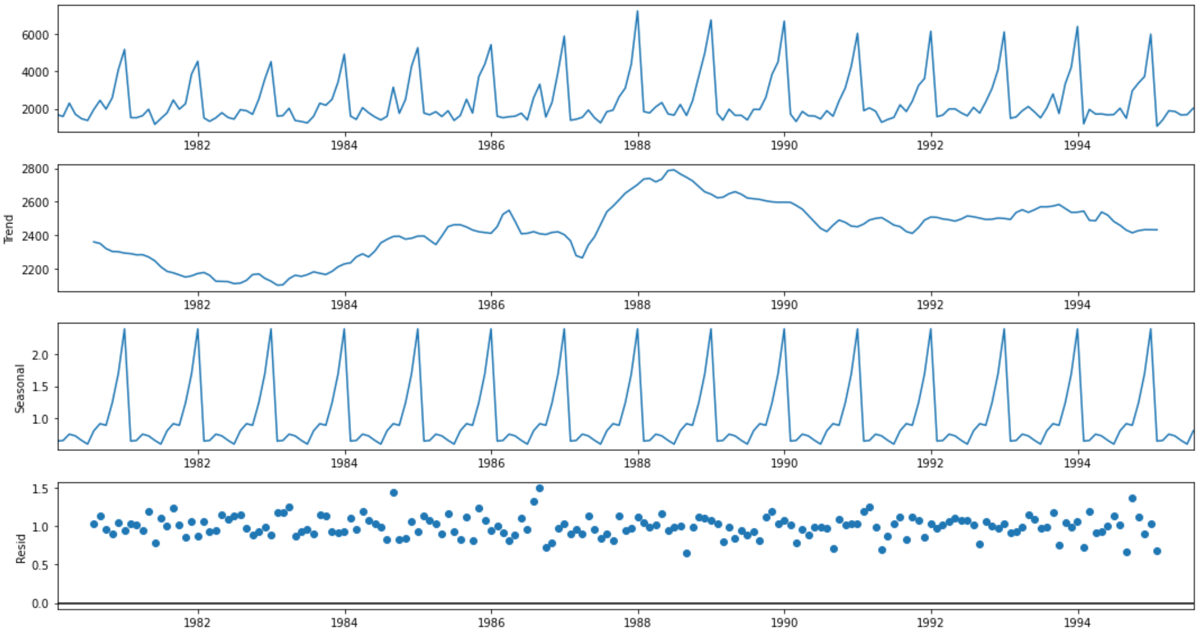


Figure VI - TS Decomposition - Sparkling (Multiplicative)

Residuals are now looking into a band of 0.5 to 1.5, and it can be concluded that Sparkling wine sales do not follow a specific trend in these years.

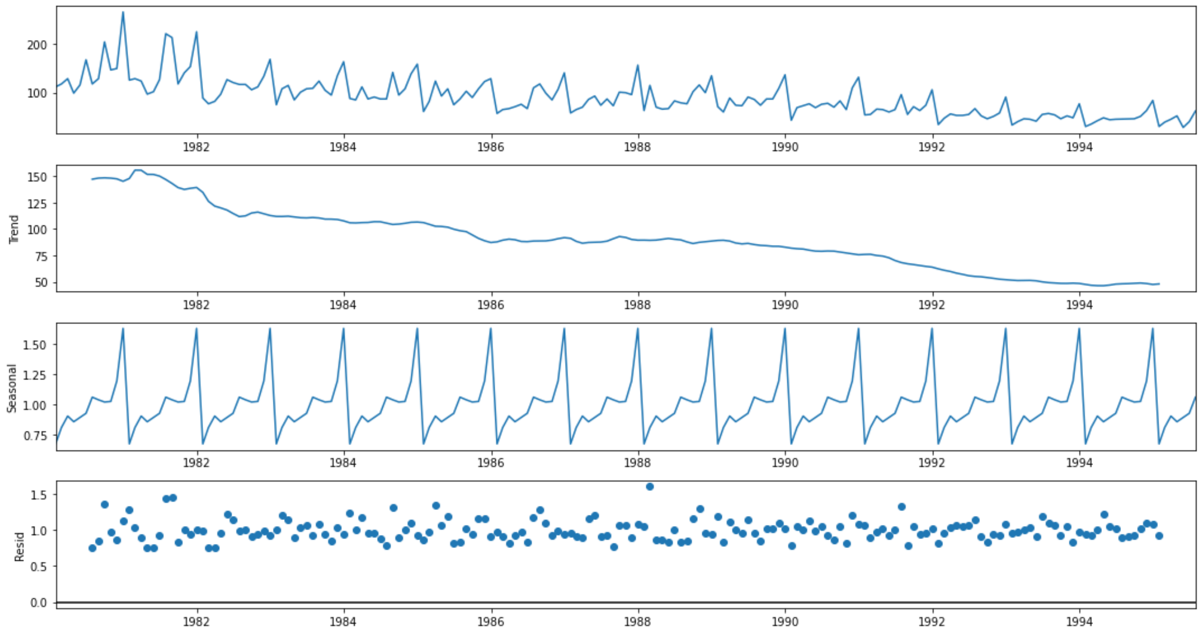


Figure VII - TS Decomposition - Rose (Multiplicative)

Residuals are now looking into a band of 0.5 to 1.5, and it can be concluded that Rose wine sales follow a down trend in these years.

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

Time series data cannot be sampled randomly for training and testing as models should be able to interpret/identify trend by any training data.

Both data frames (Sparking & Rose) were divided into Training and Testing set –

Training Set – Data before 1991 – 132 Rows

Testing Set – Data After 1991 – 55 Rows

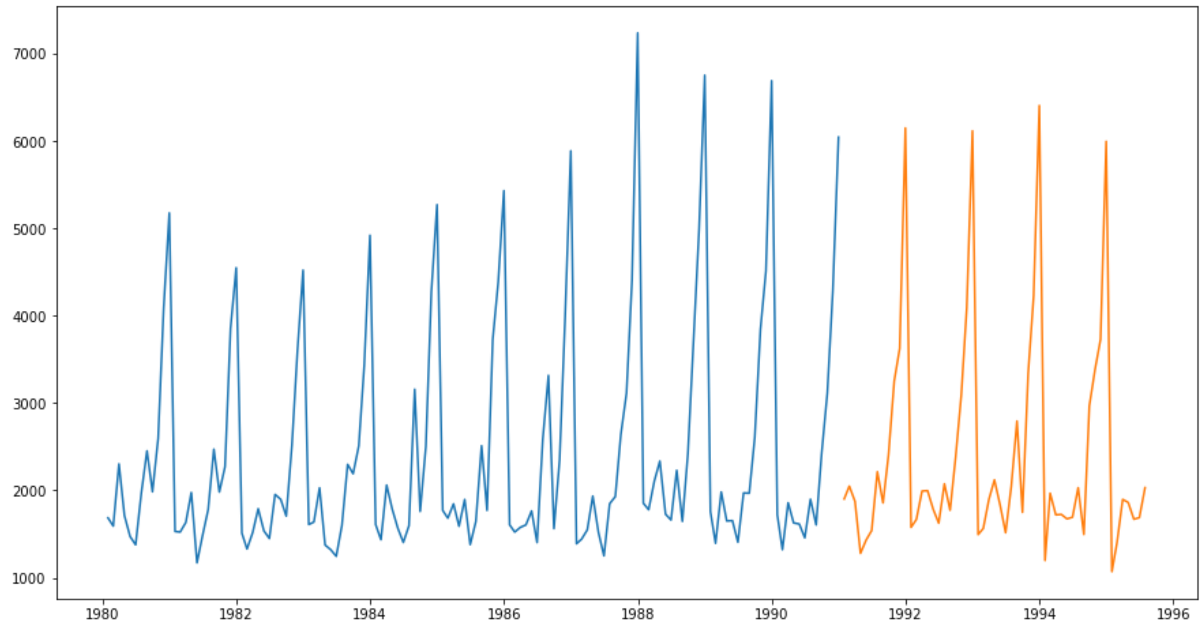


Figure VIII - Sparkling Wine Data Split

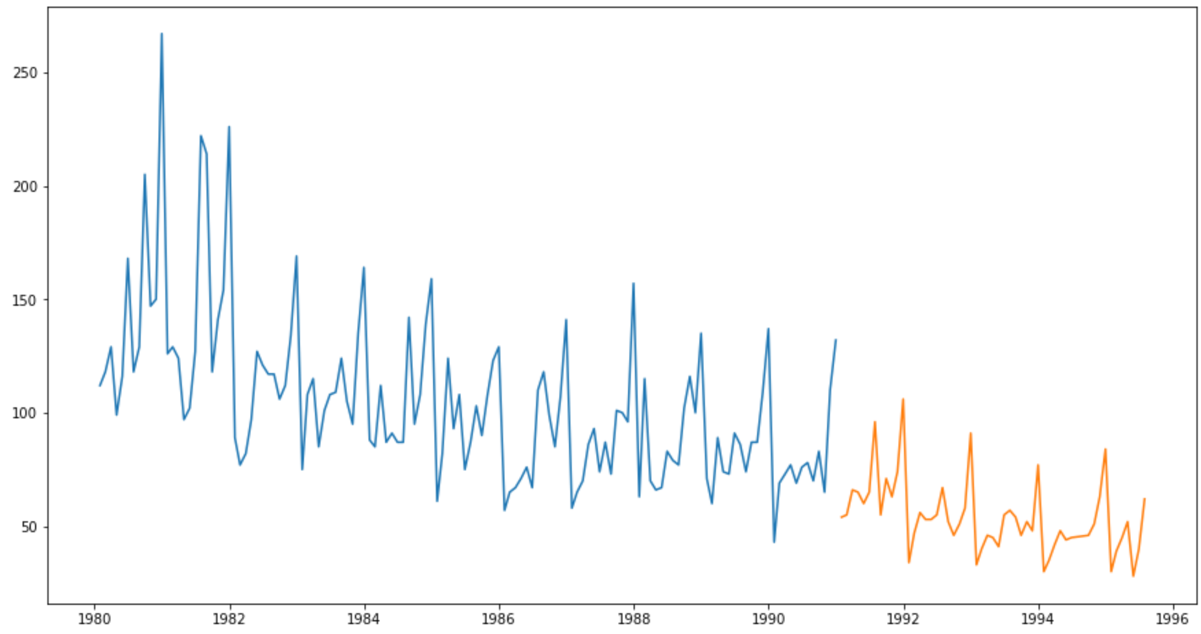


Figure IX - Rose Wine Data Split

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

Linear Regression Model

A time index was created for linear prediction from both data frames, the Training time was from 0 to 132 whereas testing time from 133 onwards (till 187)

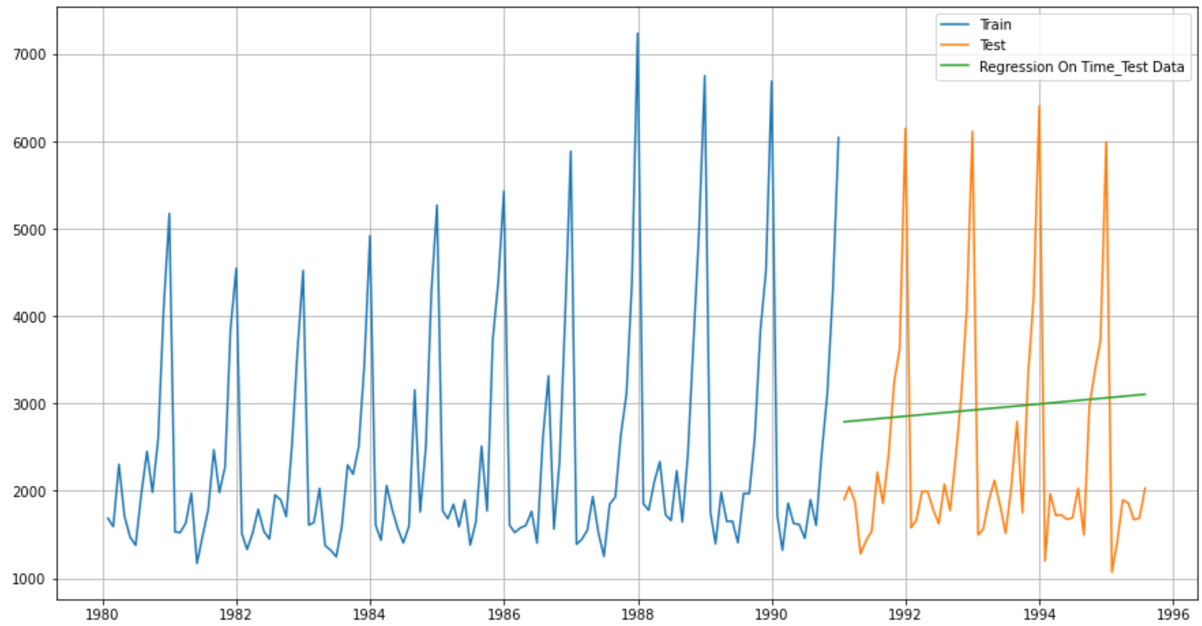


Figure X - Linear Regression Prediction (Sparkling Wine)

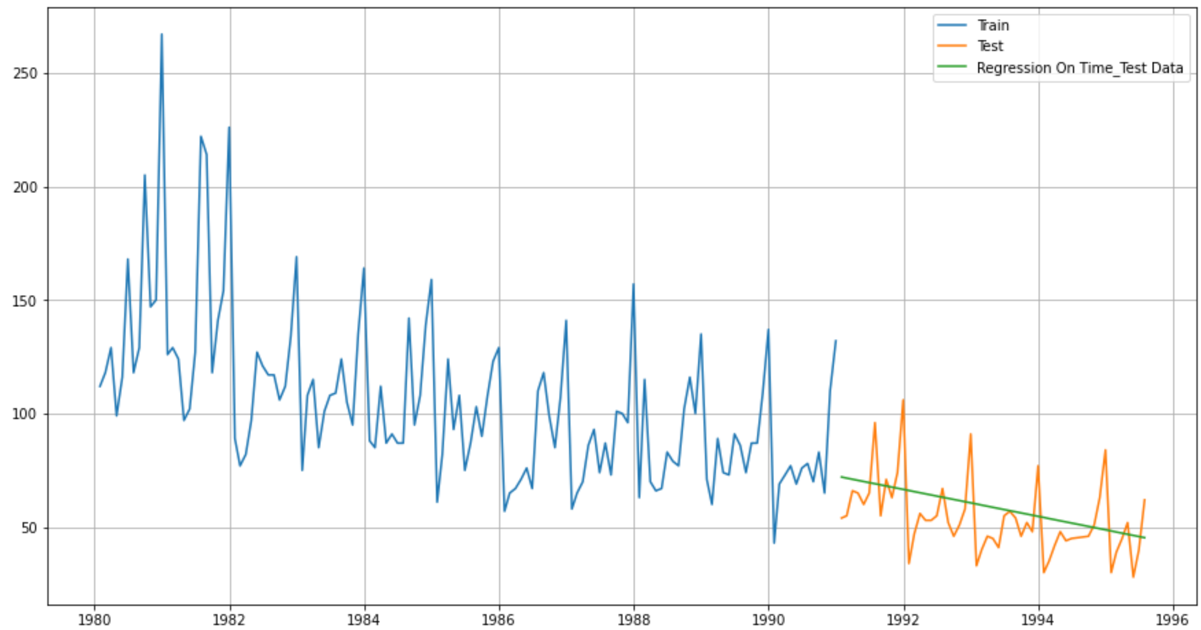


Figure XI - Linear Regression Prediction (Rose Wine)

RMSE (Test Data)

Sparkling Wine – 1389.13

Rose Wine – 15.26

Naïve Forecast Model

In Naïve Forecasting, Model pulls last known value (from training set) and use that as-is for future predictions.

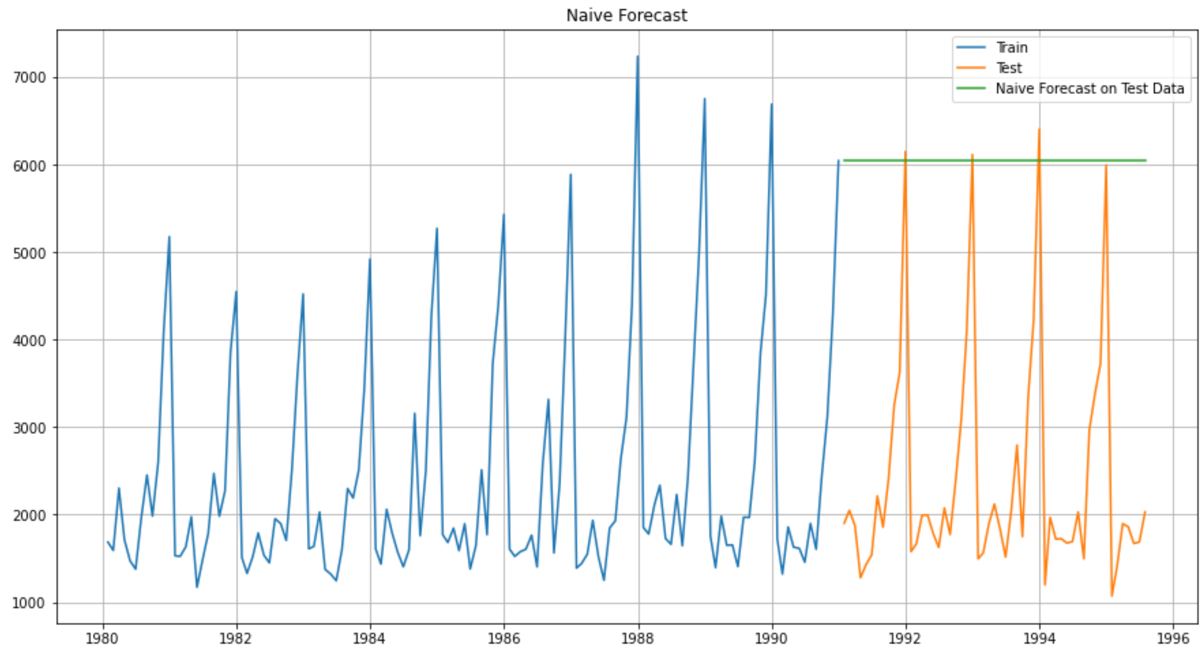


Figure XII - Naive Forecast (Sparkling Wine)

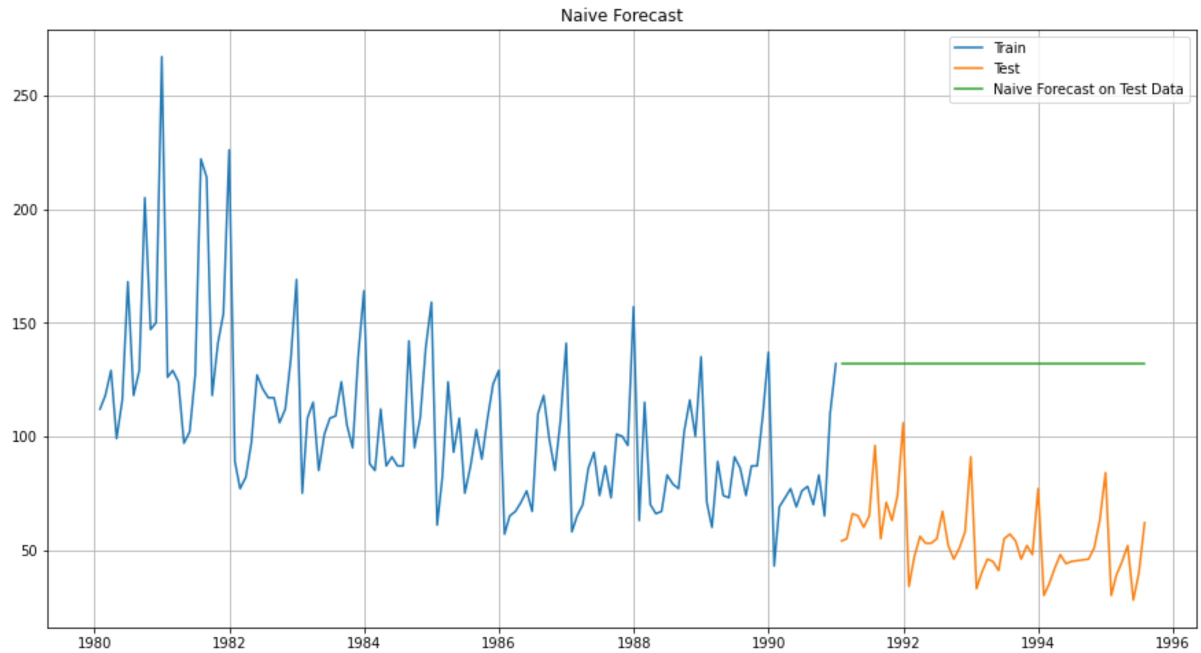


Figure XIII - Naive Forecast (Rose Wine)

RMSE (Test Data)

Sparkling Wine – 3864.27

Rose Wine – 5993.16

If RMSE compared against linear regression model, both Sparkling and Rose wine data has worsened prediction by Naïve forecasting model.

Simple Average Model

This model takes average of previously identified sales and shows that as prediction for future. It doesn't account for any trend or seasonality.

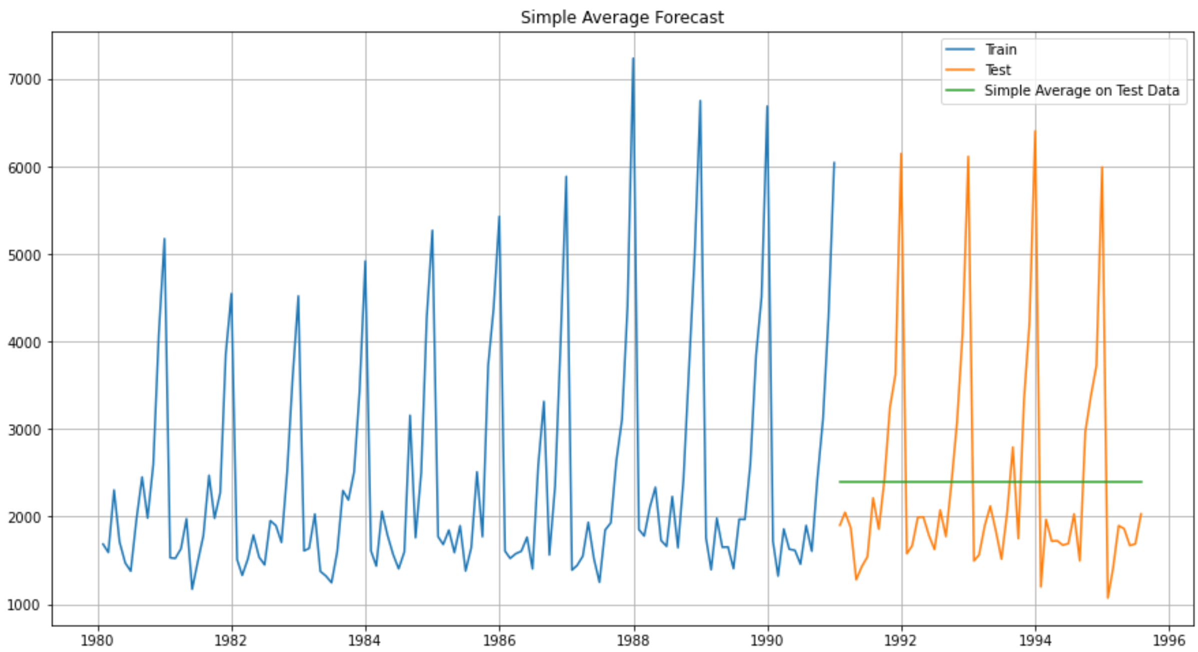


Figure XIV - Simple Average (Sparkling Wine)

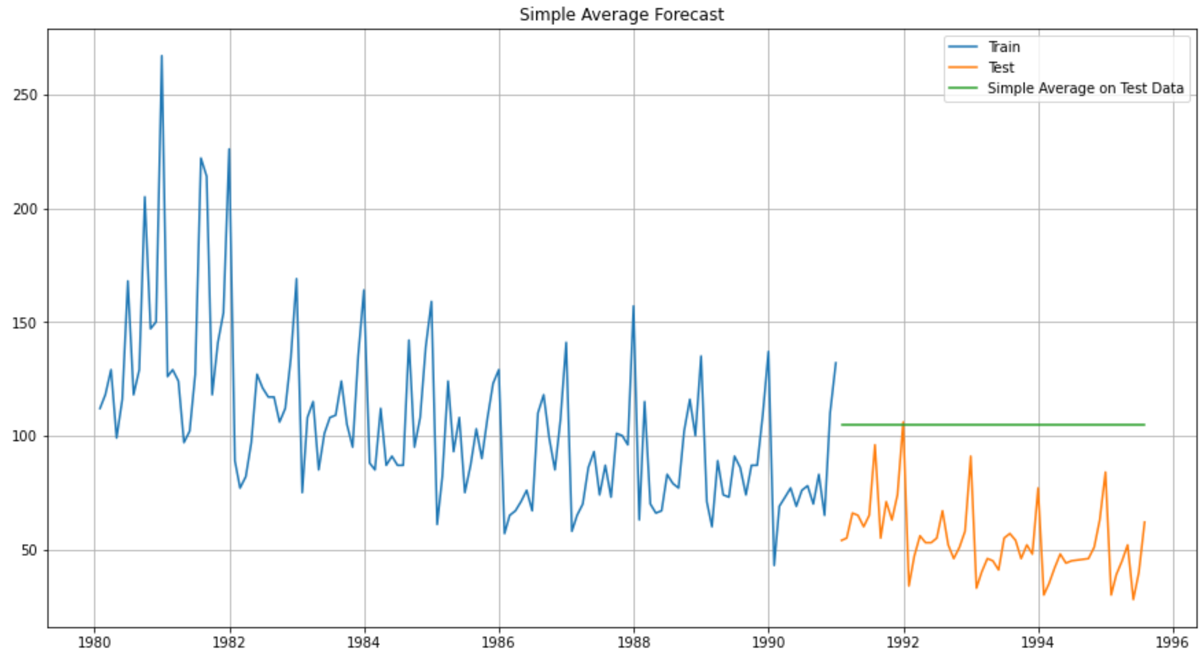


Figure XV - Simple Average (Rose Wine)

RMSE (Test Data)

Sparkling Wine – 1275.08

Rose Wine – 53.46

Moving average model has better RMSE on Sparkling wine data compared to linear regression and Naïve forecasting models.

Moving Average Model

Moving Average models taken last n values and predict next outcome, and larger the n-value, model tend to smooth the curve and approach towards simple average.

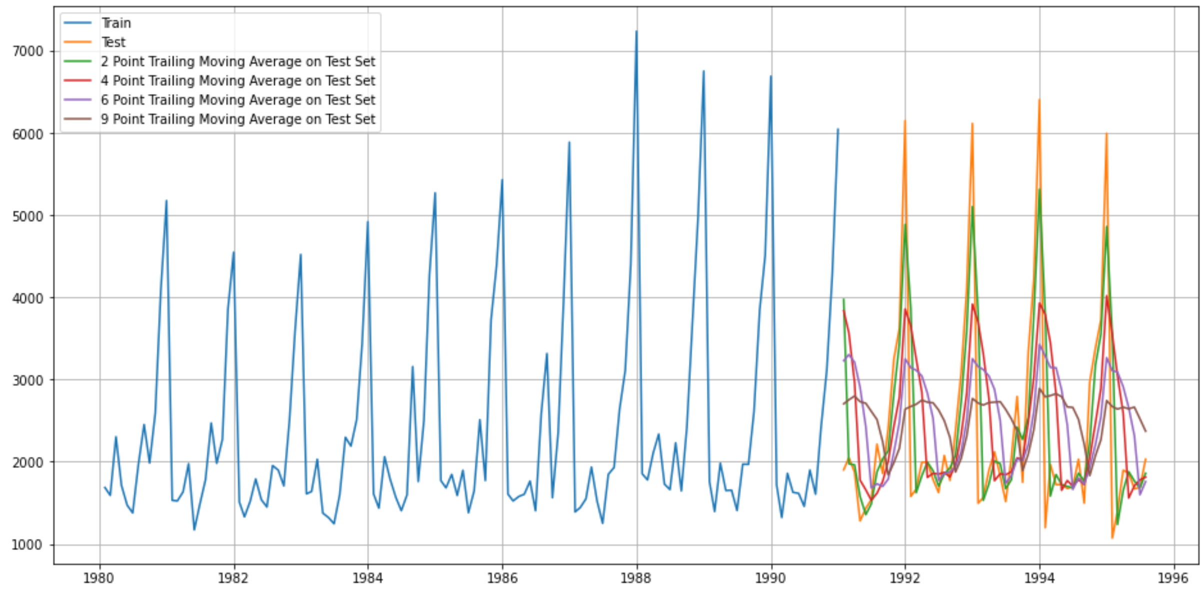


Figure XVI - n-point Moving Average (Sparkling Wine)

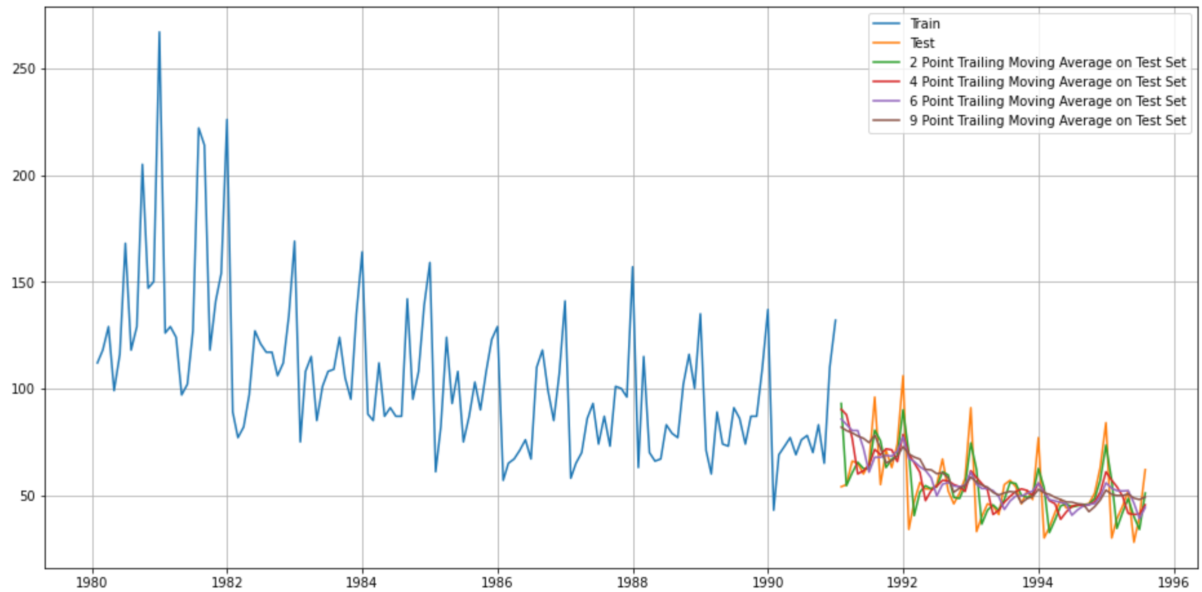


Figure XVII - n-point Moving Average (Rose Wine)

From above plots, it's evident that lower n-point moving average follows the series whereas n-point approaches towards larger values, the model tends to smooth the curve (by removing short duration noise)

RMSE (Test Data)

Models	Test RMSE (Sparkling)	Test RMSE (Rose)
2pointTrailingMovingAverage	813.40	11.53
4pointTrailingMovingAverage	1,156.59	14.45
6pointTrailingMovingAverage	1,283.93	14.57
9pointTrailingMovingAverage	1,346.28	14.73

Table 4 - Moving Average RMSE values

We can see that, 2-point Moving average model has better (lower) RMSE values on both Sparkling and Rose wine data

Simple Exponential Smoothing with additive errors

Exponential smoothing methods consist of flattening time series data.

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

The methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be none(N), additive (N), additive damped (Ad), Multiplicative (M) or multiplicative damped (Md).

SES model is applicable when data has no Trend and no seasonality

Parameters (Sparkling Wine Data)

With optimized SES model, initial level for Sparkling data comes as – 1764.013, and smoothing level as – 0.070

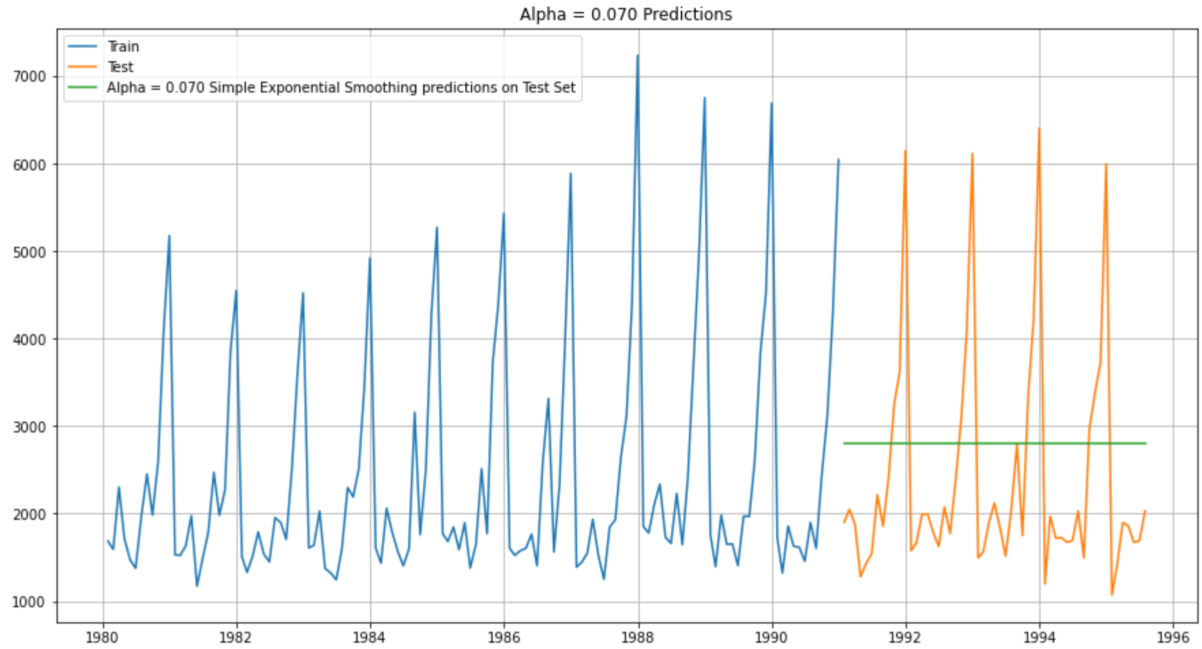


Figure XVIII - SES Model (Sparkling Wine)

Parameters (Rose Wine Data)

With optimized SES model, initial level for Rose wine data comes as – 134.38, and smoothing level as – 0.098

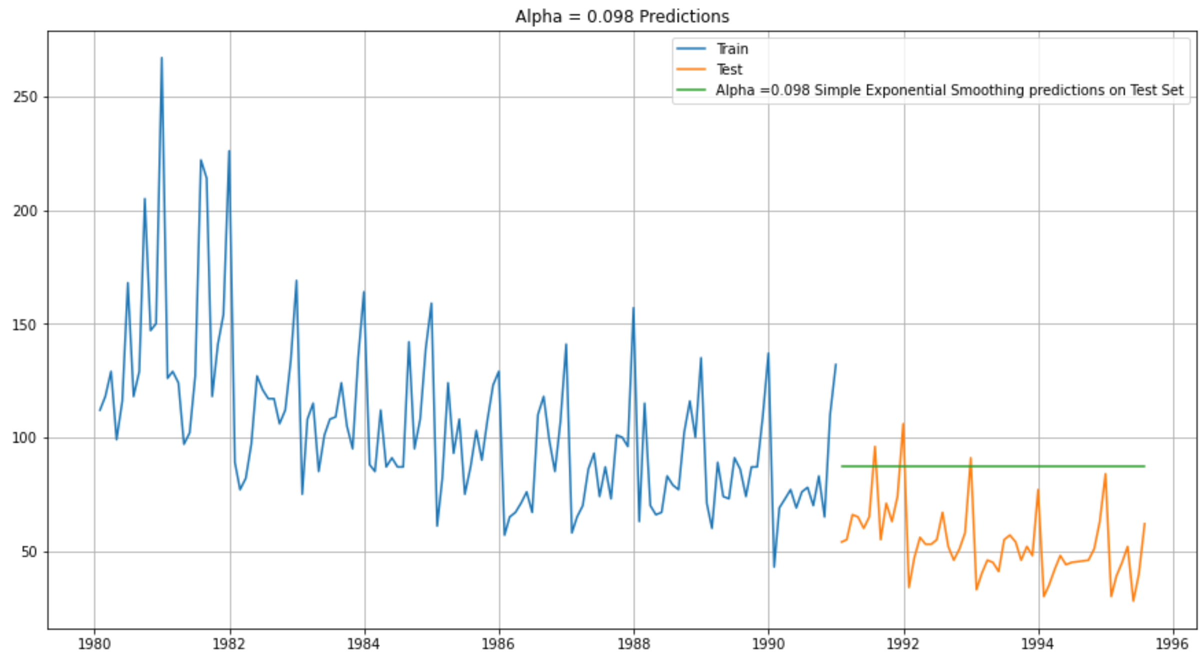


Figure XIX - SES Model (Rose Wine)

RMSE (Test Data)

Sparkling Wine – 1338.00

Rose Wine – 36.79

Double Exponential Smoothing (Halt's Linear)

Applicable when data has Trend but no seasonality

Parameters (Sparkling Wine Data)

With optimized DES model –

Initial level – 1502.19

Initial Trend – 74.87

Smoothing level – 0.66

Smoothing Trend – 0.0001

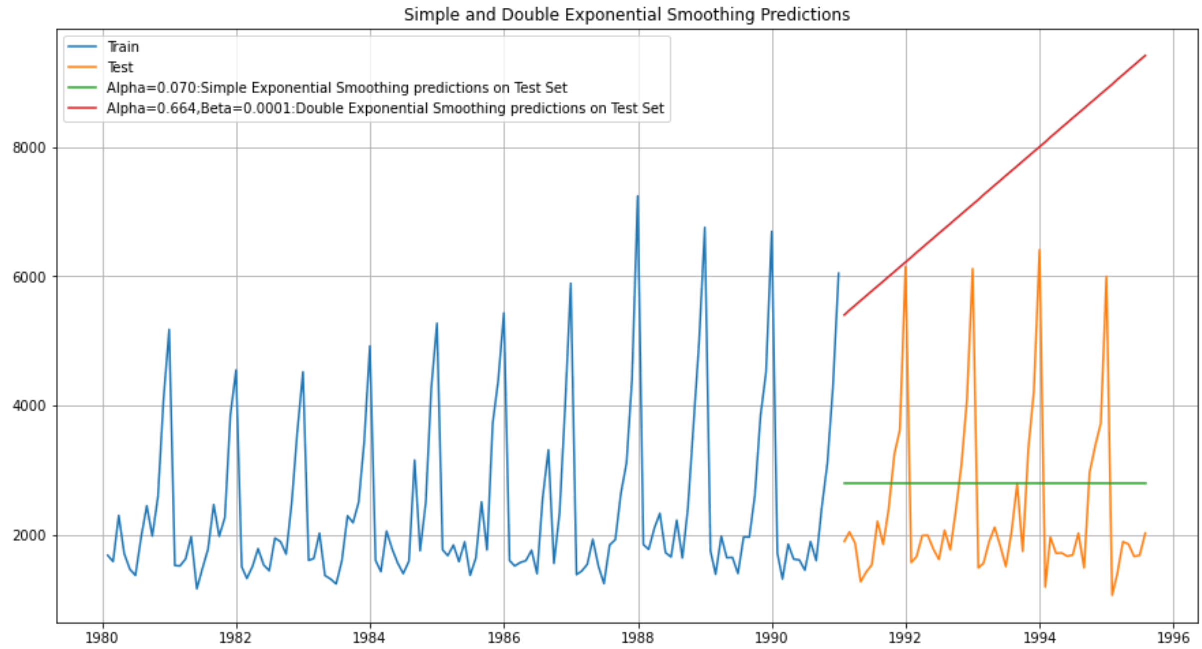


Figure XX - DES Model (Sparkling Wine)

Parameters (Rose Wine Data)

With optimized DES model –

Initial level – 137.81

Initial Trend – 0.49 (-) damping

Smoothing level – 1.49 x 10^{-8}

Smoothing Trend – 1.66 x 10^{-8}

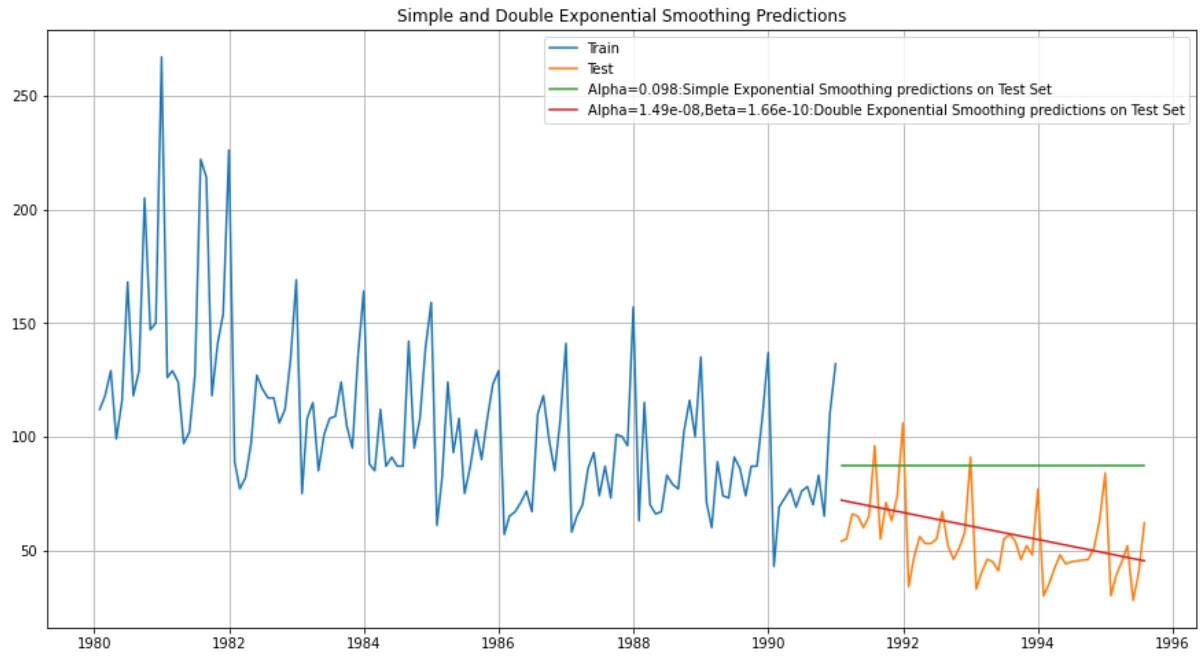


Figure XXI - DES Model (Rose Wine)

RMSE (Test Data)

Sparkling Wine – 5291.87

Rose Wine – 15.26

As we stated above, the DES works better in trending time series, since Sparkling wine data doesn't have a trend, we have bad RMSE value, whereas with damping trend on Rose wine data, we have better (lower) RMSE value.

Triple Exp (Holt Winter's linear) method with additive errors

If time series data has both trend and seasonality, Triple exponential smoothing model works better.

Parameters (Sparkling Wine Data)

With TES model on multiplicative seasonality and additive trend –

Initial level – 2356.49

Initial Trend – 10.18 (-)

Smoothing level – 0.111

Smoothing Trend – 0.049

Smoothing Seasonal – 0.362 (With a seasonal array)

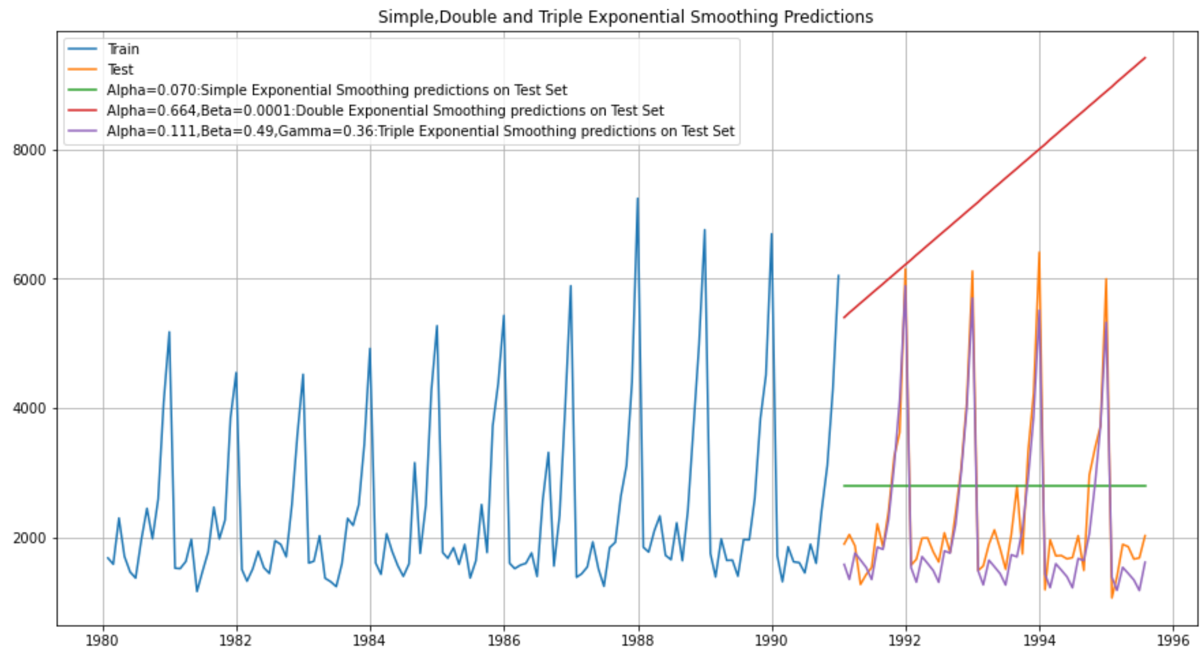


Figure XXII - TES Model (Sparkling Wine)

Parameters (Rose Wine Data)

With TES model on multiplicative seasonality and additive trend –

Initial level – 130.40

Initial Trend – 0.779 (-)

Smoothing level – 0.0715

Smoothing Trend – 0.045

Smoothing Seasonal – 7.244 x 10^{-5} (With a seasonal array)

RMSE (Test Data)

Sparkling Wine – 404.28

Rose Wine – 20.15

As we have seasonality in Sparkling wine data (which was not being accounted in SES and DES), TES has better (lower) RMSE score here.

Model	Parameters (Sparkling)	Test RMSE (Sparkling)
RegressionOnTime		1,389.14
NaiveModel		3,864.28
SimpleAvgModel		1,275.08
2pointTrailingMovingAverage		813.40
4pointTrailingMovingAverage		1,156.59
6pointTrailingMovingAverage		1,283.93
9pointTrailingMovingAverage		1,346.28
SES	Alpha = 0.070	1,338.01
DES	Alpha = 0.66, Beta = 0.0001	5,291.88
TES	Alpha = 0.111, Beta = 0.049, Gamma = 0.362	404.29

Table 5 - RMSE Comparison (Sparkling Wine)

TES (Triple Exponential Smoothing) model has best (lowest) RMSE value, whereas 2-point MA model also performs better.

Model	Parameters (Rose)	Test RMSE (Rose)
RegressionOnTime		15.27
NaiveModel		5,993.17
SimpleAvgModel		53.46
2pointTrailingMovingAverage		11.53
4pointTrailingMovingAverage		14.45
6pointTrailingMovingAverage		14.57
9pointTrailingMovingAverage		14.73
SES	Alpha = 0.098	36.80
DES	Alpha = 1.49e-8, Beta = 1.66e-8	15.27
TES	Alpha = 0.0715, Beta = 0.045, Gamma = 7.24e-5	20.16

Table 6 - RMSE Comparison (Rose Wine)

Models built on Rose wine data mostly have better (lower) RMSE values on n-point MA, Linear regression, DES & TES model, but 2-point MA has lowest RMSE among them.

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.

The Augmented Dickey-Fuller test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis for the ADF test is:

$H0$: The Time Series has a unit root and is thus non-stationary.

$H1$: The Time Series does not have a unit root and is thus stationary.

With time series data as-is, we get below values in metrics –

Metrics	Sparkling Wine	Rose Wine
Tstats	-1.798	-2.24
P-value	0.705	0.467
# of Lags	12	13

Table 7 - P-Value of Stationarity (Without Difference)

P-Value of Sparkling wine is 70.5% and Rose wine is 46.7%, so we failed to reject null hypothesis at 5% significant level. Which means both time series are not stationary.

In order to make them stationary, we can take difference within them. With 1-level difference –

Metrics	Sparkling Wine	Rose Wine
Tstats	-44.912	-8.162
P-value	0	3.01E-11
# of Lags	10	12

Table 8 - P-Value of Stationarity (After difference)

P-Value is now lesser than significant level 5%, hence we can consider time series with 1-level difference as stationary time series.

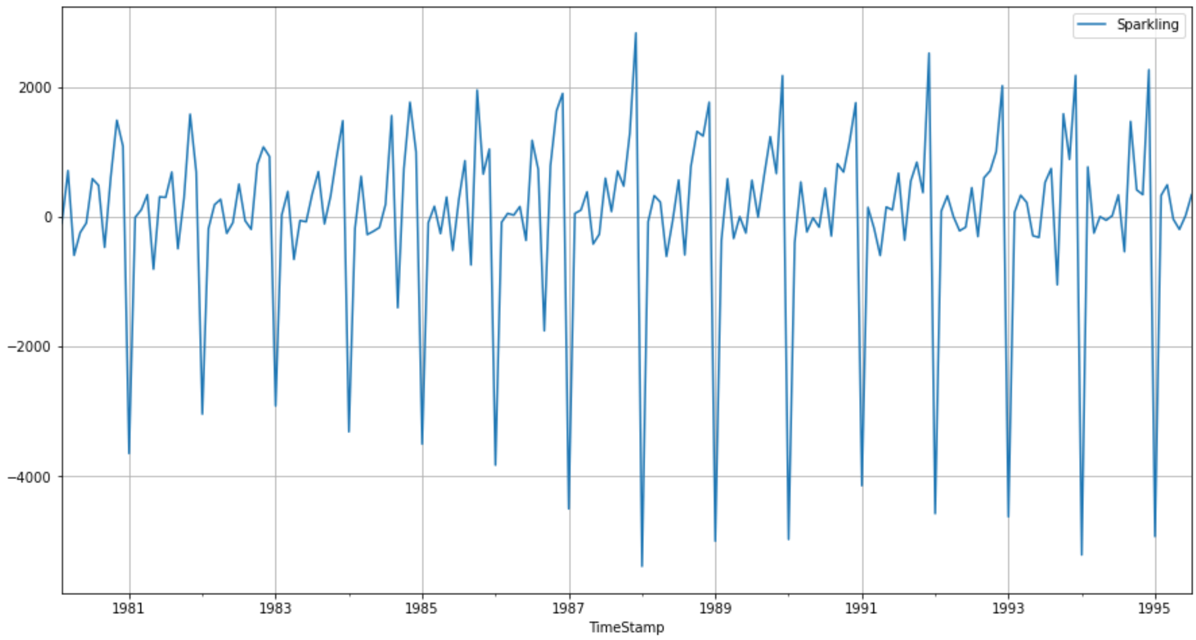


Figure XXIII - Stationary TS (Sparkling Wine)

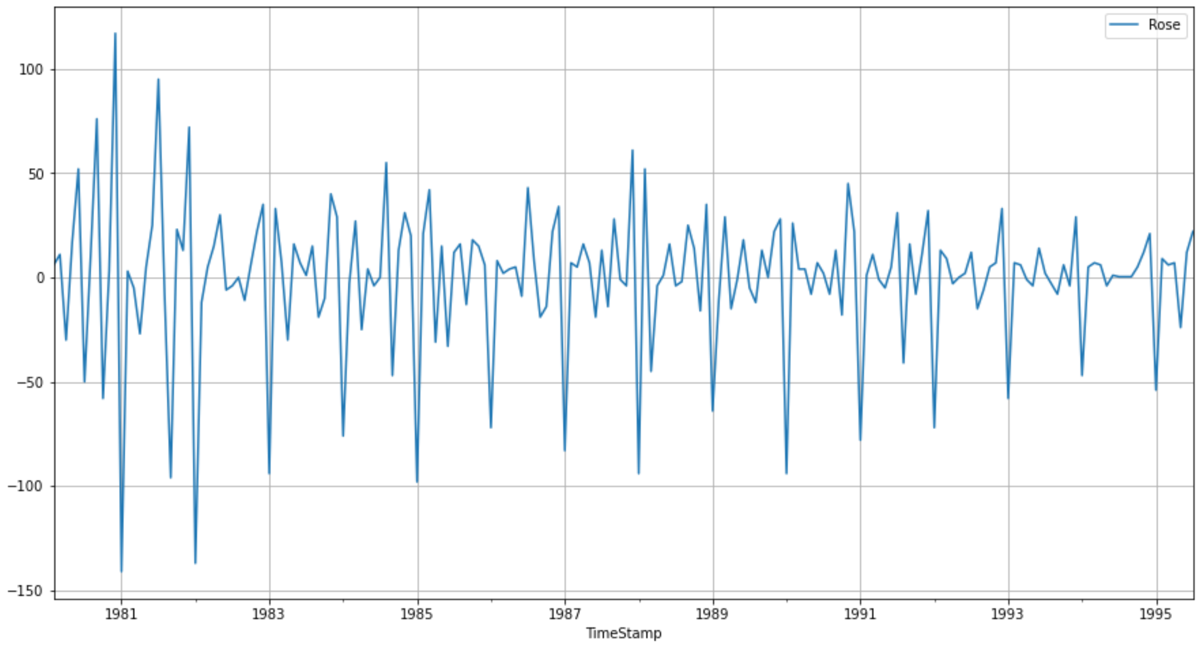


Figure XXIV - Stationary TS (Rose Wine)

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.

The Akaike information criterion is an estimator of prediction error and thereby relative quality of statistical models for a given set of data.

Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. A lower AIC score is better.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model

where:

p is the number of autoregressive terms,

d is the number of nonseasonal differences needed for stationarity

q is the number of lagged forecast errors in the prediction equation.

We will calculate AIC on list of p, d, q parameters… Since 1-level difference gives us stationary time series, hence d will be always 1.

And p, q can be taking in range 0 – 3

Sparkling Wine AIC (ARIMA)

Param (p,d,q)	AIC (Sparkling)
(2, 1, 2)	2,213.51
(3, 1, 3)	2,221.46
(3, 1, 2)	2,230.76
(2, 1, 3)	2,232.92
(2, 1, 1)	2,233.78

Table 9 - AIC Table (ARIMA) for Sparkling Wine TS (Low - High)

ARIMA Model (Sparkling)

Model will be built on params against lowest AIC, which is (2,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 24 Apr 2022   AIC                           2213.509
Time:                        01:25:26   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    -18.217      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
===================================================================================
```

Table 10 - Lowest AIC ARIMA (Sparkling)

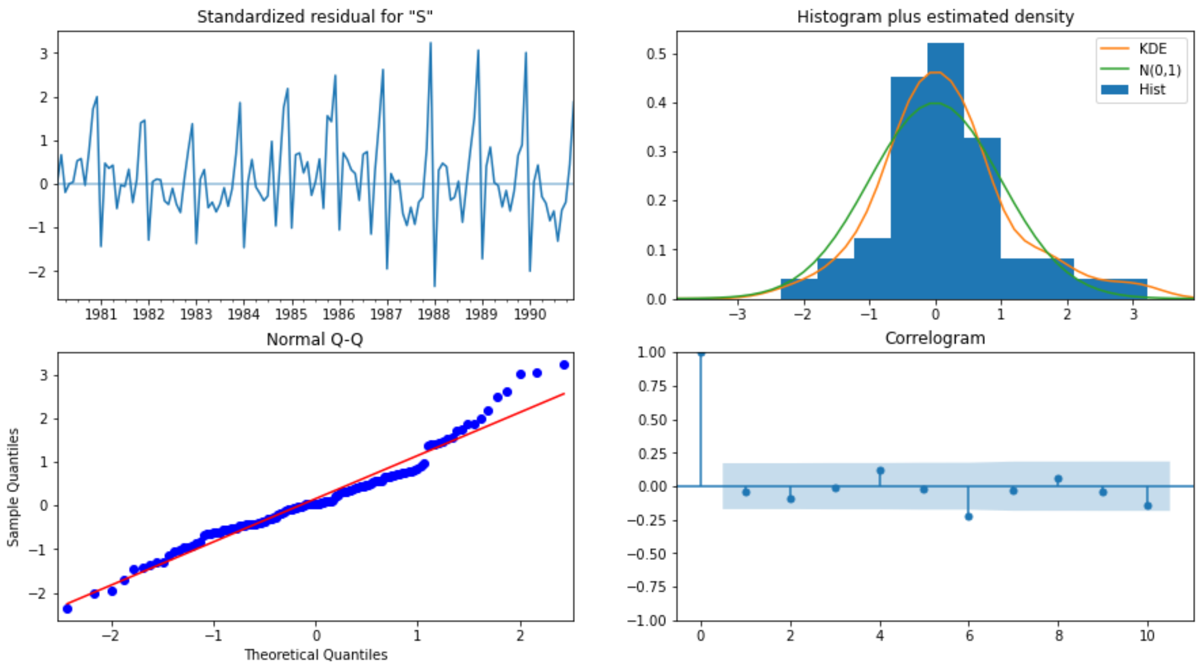


Figure XXV - Lowest AIC ARIMA Diagnosis (Sparkling)

RMSE & MAPE Value (Test Data)

RMSE: 1299.979569

MAPE: 47.099932

Rose Wine AIC (ARIMA)

Param (p,d,q)	AIC (Rose)
(2, 1, 3)	1,274.69
(3, 1, 3)	1,278.66
(0, 1, 2)	1,279.67
(1, 1, 2)	1,279.87
(0, 1, 3)	1,280.55

Table 11 - AIC Table (ARIMA) for Rose TS (Low - High)

ARIMA Model (Rose)

Model will be built on params against lowest AIC, which is (2,1,3)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.347
Date:                Sun, 24 Apr 2022   AIC                           1274.695
Time:                        01:25:27   BIC                           1291.946
Sample:                    01-31-1980   HQIC                          1281.705
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6781      0.084    -20.035      0.000      -1.842      -1.514
ar.L2         -0.7289      0.084     -8.703      0.000      -0.893      -0.565
ma.L1          1.0450      0.685      1.527      0.127      -0.297       2.387
ma.L2         -0.7716      0.137     -5.636      0.000      -1.040      -0.503
ma.L3         -0.9046      0.622     -1.455      0.146      -2.123       0.314
sigma2       858.3595    576.845      1.488      0.137    -272.237    1988.956
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.45
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

Table 12 - Lower AIC ARIMA (Rose)

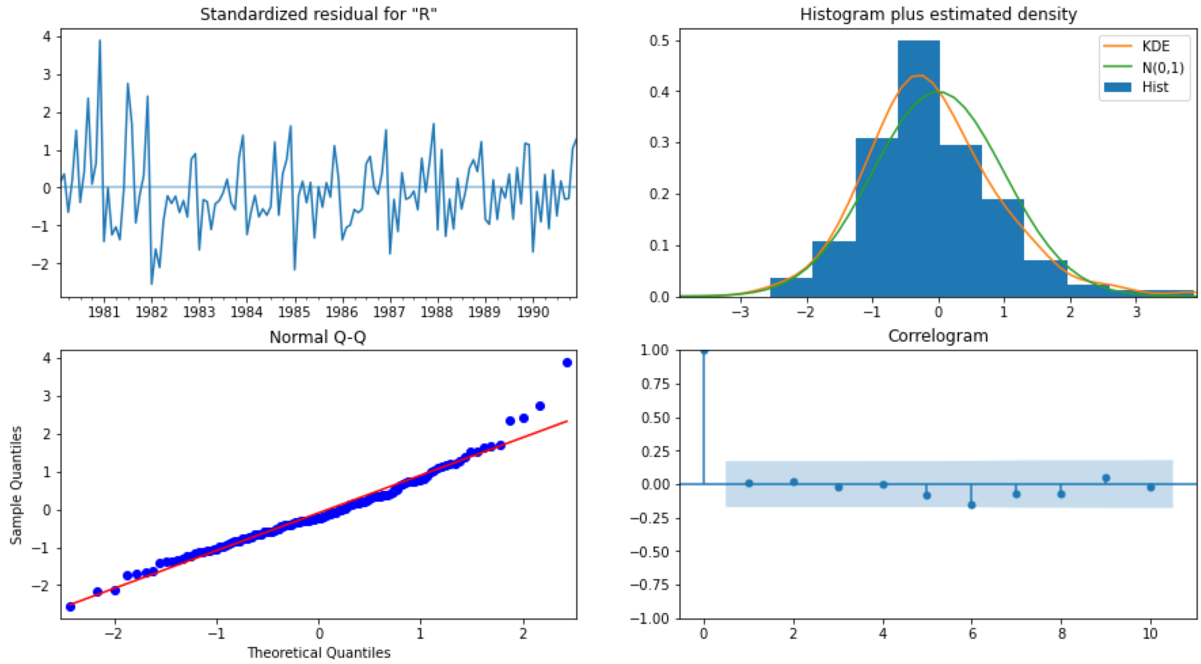


Figure XXVI - Lowest AIC ARIMA Diagnosis (Rose)

RMSE & MAPE Value (Test Data)

RMSE: 36.81

MAPE: 75.84

Seasonal ARIMA (SARIMA)

Trend Elements

p: Trend autoregression order. (0-3)

d: Trend difference order. (0-3)

q: Trend moving average order. (1)

Seasonal Elements

P: Seasonal autoregressive order. (0-3)

D: Seasonal difference order. (0)

Q: Seasonal moving average order. (0-3)

m: The number of time steps for a single seasonal period. (6)

We will calculate AIC on list of p, d, q parameters… Since 1-level difference gives us stationary time series, hence d will be always 1.

And p, q can be taking in range 0 – 3 and seasonality being repeated at every 6 months so, m would be 6.

Sparkling Wine AIC (SARIMA)

Trend Param	Seasonal Param	AIC (Sparkling)
(2, 1, 3)	(2, 0, 3, 6)	1,629.15
(3, 1, 3)	(2, 0, 3, 6)	1,631.01
(0, 1, 3)	(2, 0, 3, 6)	1,633.33
(1, 1, 3)	(2, 0, 3, 6)	1,633.97
(0, 1, 3)	(3, 0, 3, 6)	1,635.05

Table 13 - AIC Table (SARIMA) for Sparkling TS (Low - High)

SARIMA Model (Sparkling)

Model will be built on params against lowest AIC, which is $(2,1,3) \times (2,0,3)_6$

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -803.575
Date:                            Sun, 24 Apr 2022   AIC                           1629.150
Time:                                    01:40:29   BIC                           1658.755
Sample:                                01-31-1980   HQIC                          1641.156
                                     - 12-31-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.7450      0.065    -27.006      0.000      -1.872      -1.618
ar.L2         -0.7871      0.069    -11.384      0.000      -0.923      -0.652
ma.L1          1.0833      0.165      6.580      0.000       0.761       1.406
ma.L2         -0.7526      0.123     -6.139      0.000      -0.993      -0.512
ma.L3         -0.8884      0.112     -7.962      0.000      -1.107      -0.670
ar.S.L6       -0.0107      0.029     -0.364      0.716      -0.068       0.047
ar.S.L12       1.0381      0.022     47.708      0.000       0.995       1.081
ma.S.L6        0.1216      0.179      0.678      0.498      -0.230       0.473
ma.S.L12      -0.5765      0.099     -5.848      0.000      -0.770      -0.383
ma.S.L18       0.0883      0.139      0.633      0.527      -0.185       0.362
sigma2      1.323e+05   1.91e-06   6.93e+10      0.000    1.32e+05    1.32e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                15.25
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               1.50   Skew:                             0.38
Prob(H) (two-sided):                  0.23   Kurtosis:                         4.66
===================================================================================
```

Table 14 - Lowest AIC SRIMA (Sparkling)

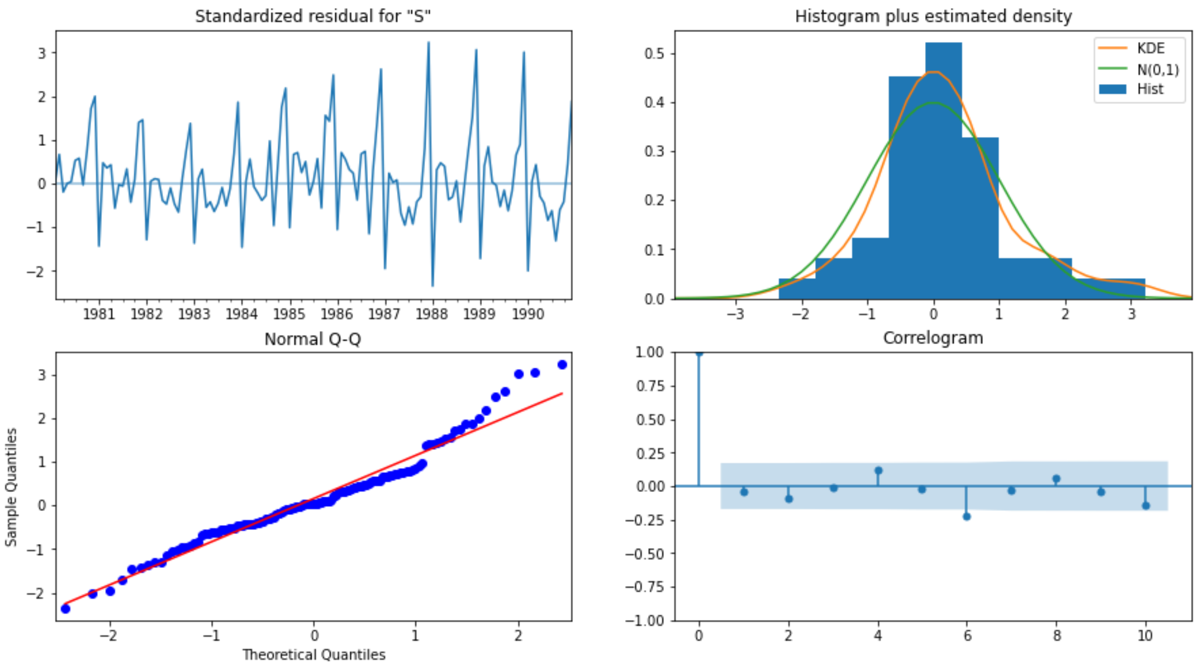


Figure XXVII - Lowest AIC SARIMA Diagnosis (Sparkling)

RMSE & MAPE Value (Test Data)

RMSE: 812.747

MAPE: 35.757

Rose Wine AIC (SARIMA)

Trend	Seasonal	AIC (Rose)
(2, 1, 3)	(2, 0, 3, 6)	951.744298
(0, 1, 3)	(2, 0, 3, 6)	952.073632
(3, 1, 3)	(2, 0, 3, 6)	952.582104
(1, 1, 3)	(2, 0, 3, 6)	953.684951
(0, 1, 3)	(3, 0, 3, 6)	954.049162

Table 15 - AIC Table (SARIMA) for Rose TS (Low - High)

SARIMA Model (Rose)

Model will be built on params against lowest AIC, which is (2,1,3) x $(2,0,3)_6$

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  Rose   No. Observations:               132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -464.872
Date:                            Sun, 24 Apr 2022   AIC                            951.744
Time:                                    01:40:33   BIC                            981.349
Sample:                                01-31-1980   HQIC                           963.750
                                     - 12-31-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5026      0.083     -6.079      0.000      -0.665      -0.341
ar.L2         -0.6627      0.084     -7.916      0.000      -0.827      -0.499
ma.L1         -0.3714    215.453     -0.002      0.999    -422.651     421.908
ma.L2          0.2033    135.412      0.002      0.999    -265.199     265.606
ma.L3         -0.8319    179.184     -0.005      0.996    -352.026     350.362
ar.S.L6       -0.0838      0.049     -1.720      0.085      -0.179       0.012
ar.S.L12       0.8099      0.052     15.463      0.000       0.707       0.913
ma.S.L6        0.1702      0.248      0.687      0.492      -0.316       0.656
ma.S.L12      -0.5645      0.199     -2.835      0.005      -0.955      -0.174
ma.S.L18       0.1710      0.143      1.198      0.231      -0.109       0.451
sigma2       260.8103   5.62e+04      0.005      0.996    -1.1e+05     1.1e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):                 4.77
Prob(Q):                              0.40   Prob(JB):                         0.09
Heteroskedasticity (H):               0.54   Skew:                            -0.36
Prob(H) (two-sided):                  0.06   Kurtosis:                         3.73
===================================================================================
```

Table 16 - Lowest AIC SARIMA (Rose)

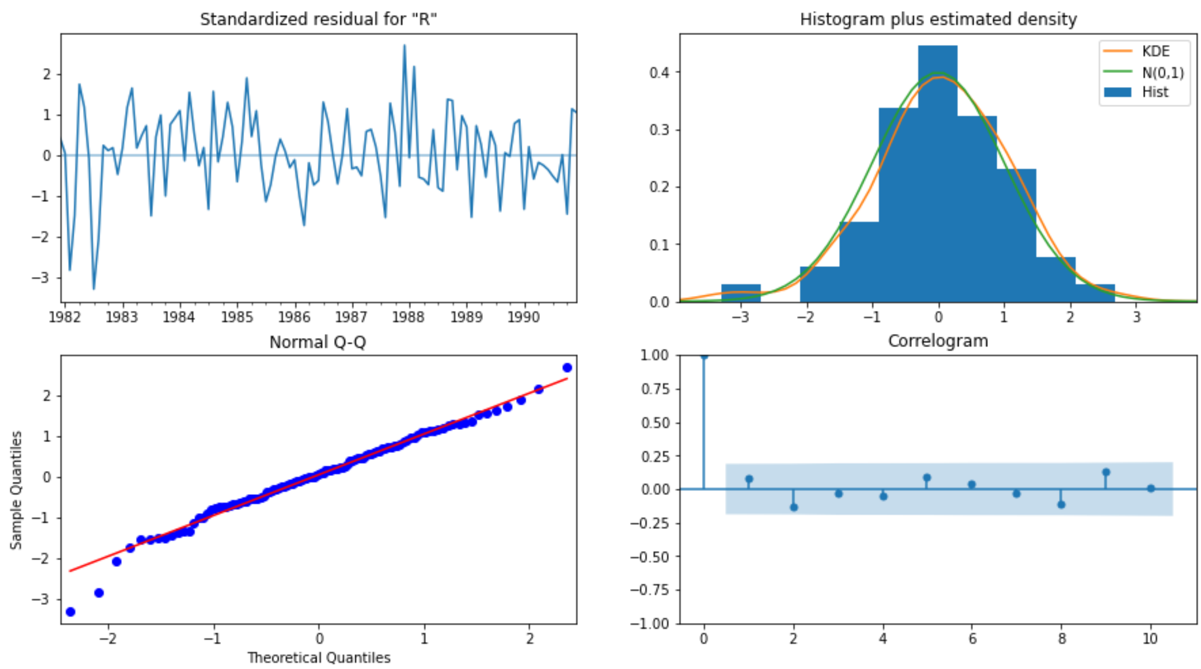


Figure XXVIII - Lowest AIC SARIMA Diagnosis (Rose)

RMSE & MAPE Value (Test Data)

RMSE: 27.124

MAPE: 55.24

Model Comparison

Models	Parameter (Sparkling)	RMSE (Sparkling)	MAPE (Sparkling)
ARIMA - Lowest AIC	(2,1,2)	1299.979569	47.099932
SARIMA - Lowest AIC	(2, 1, 3) x (2,0,3,6)	812.74728	35.757186

Table 17 - ARIMA vs SARIMA Comparison (Sparkling)

Models	Parameter (Rose)	RMSE (Rose)	MAPE (Rose)
ARIMA - Lowest AIC	(2,1,3)	36.817423	75.848378
SARIMA - Lowest AIC	(2, 1, 3) x (2,0,3,6)	27.124997	55.240791

Table 18 - ARIMA vs SARIMA Comparison (Rose)

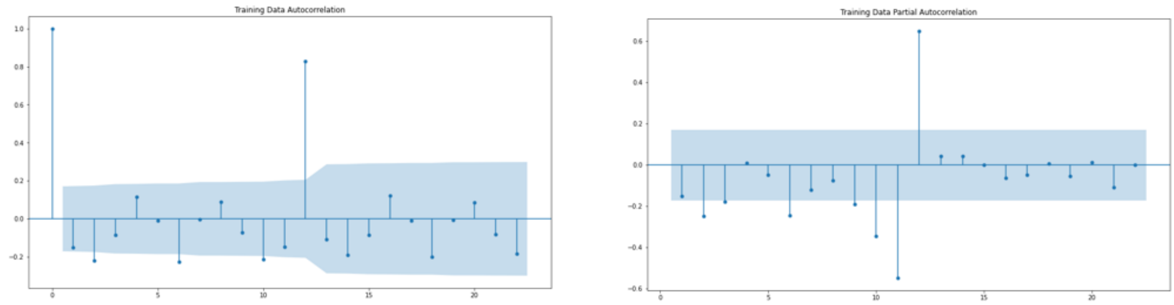
SARIMA works better than ARIMA for Sparkling & Rose Dataset (Since both have seasoning SARIMA could capture the fluctuations)

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

ACF is an (complete) auto-correlation function which gives us values of autocorrelation of any series with its lagged values.

PACF is a partial auto-correlation function. Basically, instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)).

Sparkling Wine Data - ARIMA



ACF Plot

PACF Plot

Figure XXIX - ACF & PACF Plot (Sparkling Wine)

With above plot, we can see the 1st cross in ACF is happening at 2 (ignored 0 valued) and PACF at 2.

We can build model on top of them with param (p,d,q) – (2,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.832
Date:                Sat, 23 Apr 2022   AIC                           2267.663
Time:                        18:06:19   BIC                           2270.538
Sample:                    01-31-1980   HQIC                          2268.831
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.885e+06   1.29e+05     14.658      0.000    1.63e+06    2.14e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.07   Jarque-Bera (JB):               198.83
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.65
===================================================================================
```

Figure XXX - Manual ARIMA (Sparkling)

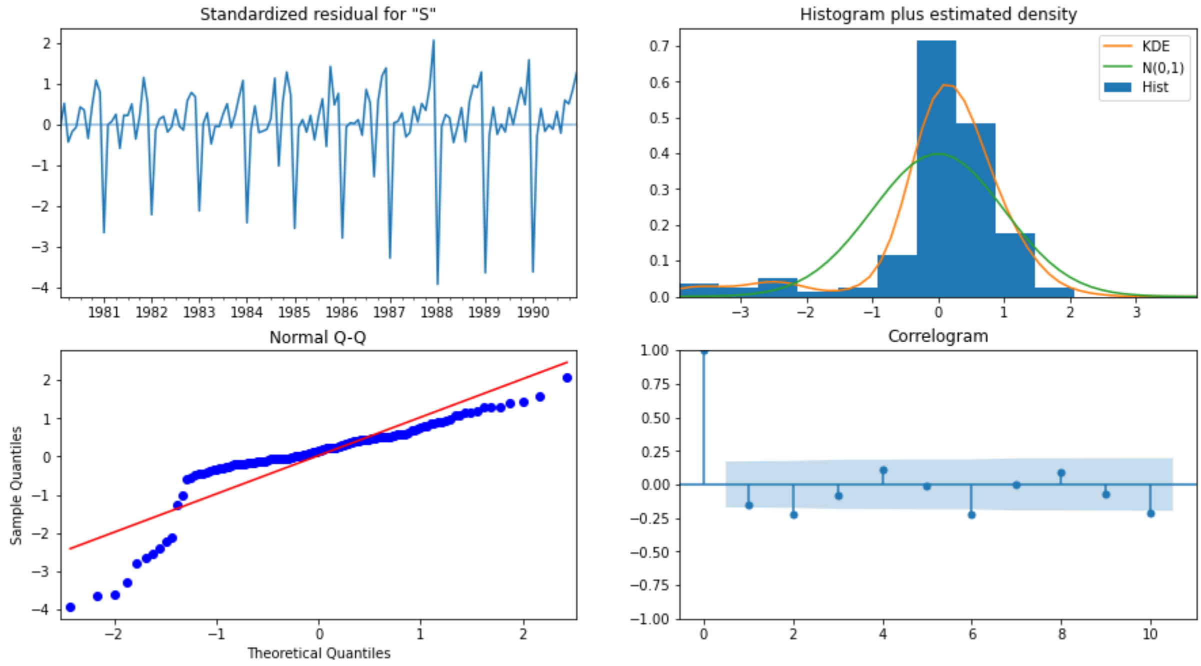


Figure XXXI - Manual ARIMA Diagnostic (Sparkling)

RMSE & MAPE Values

RMSE: 3864.279

MAPE: 201.327

Rose Wine Data - ARIMA

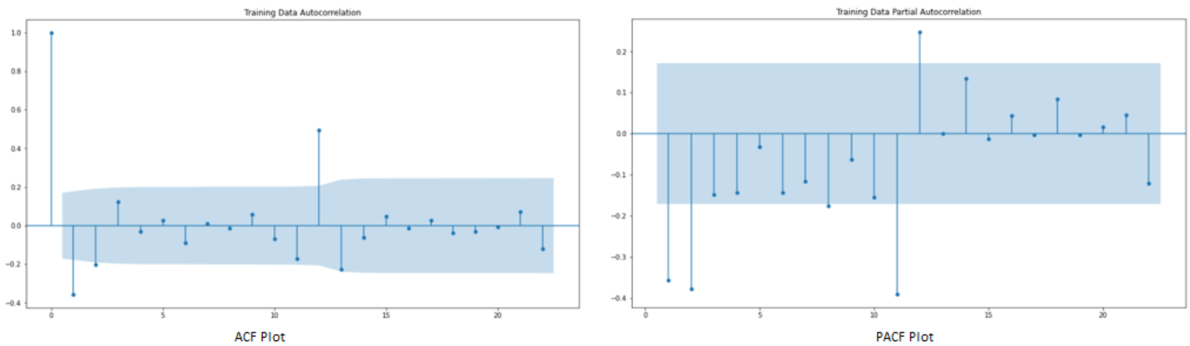


Figure XXXII - ACF & PACF Plot (Rose Wine)

With above plot, we can see the 1st cross in ACF is happening at 2 (ignored 0 valued) and PACF at 2.

We can build model on top of them with param (p,d,q) – (2,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Sat, 23 Apr 2022   AIC                           1281.871
Time:                        18:07:06   BIC                           1296.247
Sample:                    01-31-1980   HQIC                          1287.712
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4540      0.469     -0.969      0.333      -1.372       0.464
ar.L2          0.0001      0.170      0.001      0.999      -0.334       0.334
ma.L1         -0.2541      0.459     -0.554      0.580      -1.154       0.646
ma.L2         -0.5984      0.430     -1.390      0.164      -1.442       0.245
sigma2       952.1601     91.424     10.415      0.000     772.973    1131.347
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                34.16
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.94
===================================================================================
```

Figure XXXIII - Manual ARIMA (Rose)

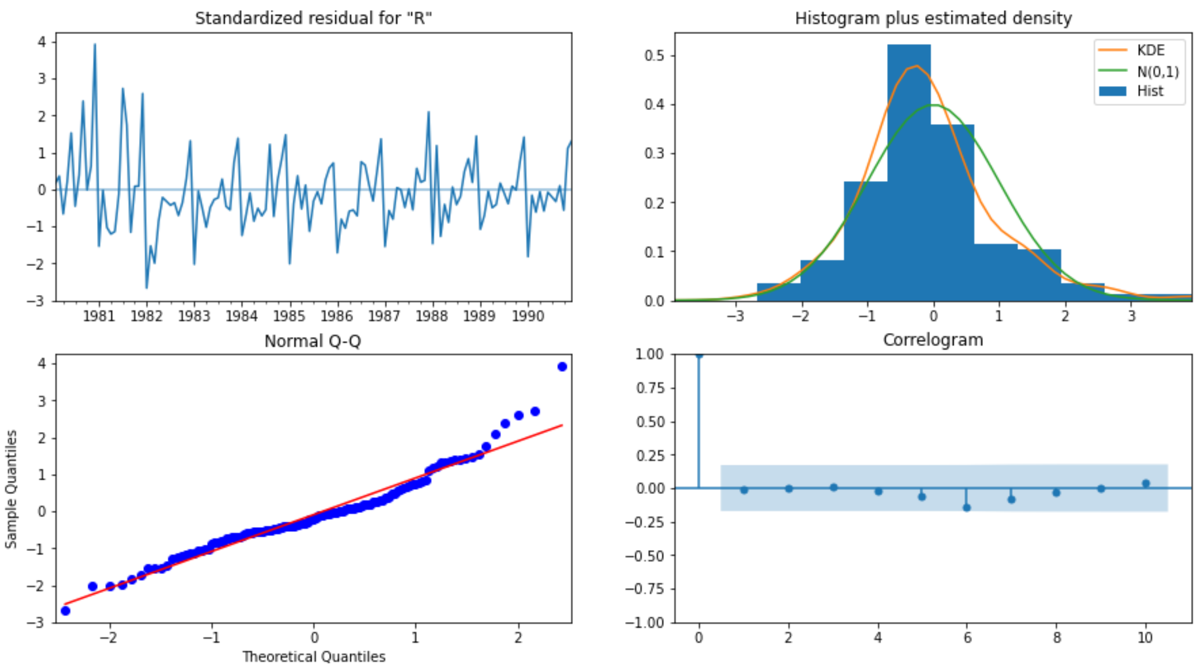


Figure XXXIV - Manual ARIMA Diagnostic (Rose)

RMSE & MAPE Values

RMSE: 36.871

MAPE: 76.056

Sparkling Wine Data - SARIMA

We have already identified trend parameter (p,d,q) – (2,1,2) from ACF & PACF plots, seasonal parameters will also be same and seasonality can be seen at every 6 months.

Parameters – (2,1,2) x (2,0,2,6)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:               SARIMAX(0, 1, 0)   Log Likelihood               -1124.680
Date:                Sat, 23 Apr 2022   AIC                           2251.360
Time:                        18:38:37   BIC                           2254.227
Sample:                    01-31-1980   HQIC                          2252.525
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.899e+06   1.31e+05     14.543      0.000    1.64e+06    2.16e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.04   Jarque-Bera (JB):               194.29
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.60
===================================================================================
```

Figure XXXV - Manual SARIMA (Sparkling)

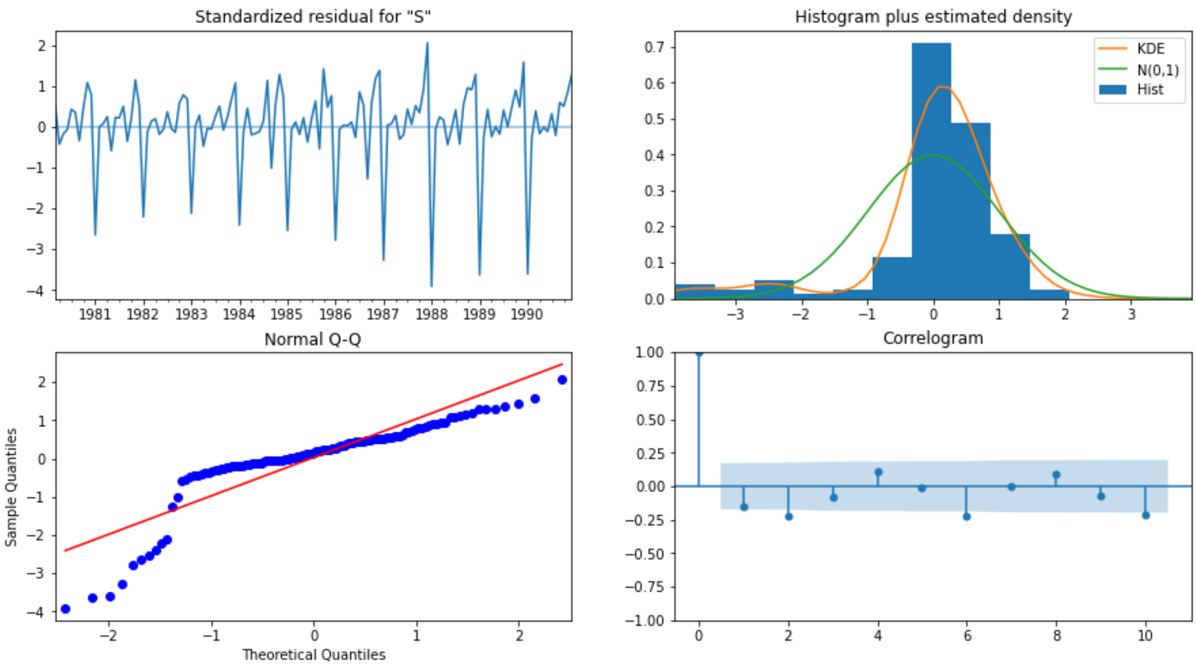


Figure XXXVI - Manual SARIMA (Sparkling)

RMSE & MAPE Values

RMSE: 2640.80

MAPE: 96.012

Rose Wine Data - SARIMA

We have already identified trend parameter (p,d,q) – (2,1,2) from ACF & PACF plots, seasonal parameters will also be same and seasonality can be seen at every 6 months.

Parameters – (2,1,2) x (2,0,2,6)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -513.610
Date:                             Sat, 23 Apr 2022   AIC                           1045.220
Time:                                     18:39:58   BIC                           1070.003
Sample:                                 01-31-1980   HQIC                          1055.281
                                      - 12-31-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0479      0.120      8.749      0.000       0.813       1.283
ar.L2         -0.2225      0.134     -1.662      0.097      -0.485       0.040
ma.L1         -1.9992    664.781     -0.003      0.998   -1304.946    1300.947
ma.L2          1.0000    665.048      0.002      0.999   -1302.471    1304.471
ar.S.L6       -0.1127      0.026     -4.305      0.000      -0.164      -0.061
ar.S.L12       0.7999      0.024     33.808      0.000       0.754       0.846
ma.S.L6        0.2936    665.014      0.000      1.000   -1303.110    1303.697
ma.S.L12      -0.7063    469.766     -0.002      0.999    -921.430     920.017
sigma2       315.2293      0.143   2198.061      0.000     314.948     315.510
===================================================================================
Ljung-Box (L1) (Q):                   0.18   Jarque-Bera (JB):               120.40
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               0.44   Skew:                            -0.31
Prob(H) (two-sided):                  0.01   Kurtosis:                         7.95
===================================================================================
```

Figure XXXVII - Manual SARIMA (Rose)

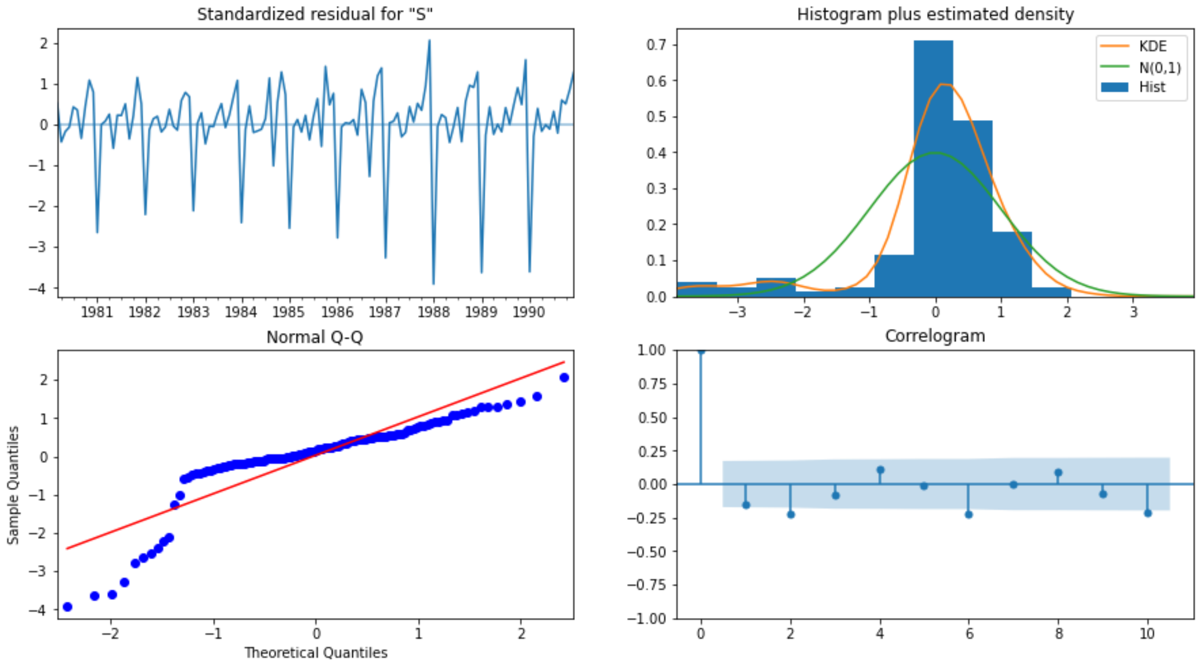


Figure XXXVIII - Manual SARIMA (Rose)

RMSE & MAPE Values

RMSE: 30.63

MAPE: 62.65

Model	Parameter	RMSE (Sparkling)	MAPE (Sparkling)	RMSE (Rose)	MAPE (Rose)
ARIMA - Manual	(2,1,2)	3,864.28	201.33	36.87	76.06
SARIMA - Manual	(2,1,2) x (2,0,2,6)	2,640.81	96.01	30.63	62.66

Table 19 - SARIMA RMSE Values

SARIMA works better in both cases of Sparkling and Rose wine.

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

RMSE Values on Sparkling Wine Data

Model	Parameters (Sparkling)	Test RMSE (Sparkling)
RegressionOnTime		1,389.14
NaiveModel		3,864.28
SimpleAvgModel		1,275.08

2pointTrailingMovingAverage		813.4
4pointTrailingMovingAverage		1,156.59
6pointTrailingMovingAverage		1,283.93
9pointTrailingMovingAverage		1,346.28
SES	Alpha = 0.070	1,338.01
DES	Alpha = 0.66, Beta = 0.0001	5,291.88
TES	Alpha = 0.111, Beta = 0.049, Gamma = 0.362	404.29
ARIMA - Lowest AIC	(2,1,2)	1299.979569
SARIMA - Lowest AIC	(2,1,3) x (2,0,3,6)	812.74728
ARIMA - Manual (PCF/APCF)	(2,1,2)	3864.279352
SARIMA - Manual (PCF/APCF)	(2,1,2) x (2,0,2,6)	2640.806467

Table 20 - RMSE Values of Models (Sparkling Wine Data)

TES (Triple exponential smoothing) model has best RMSE score compared to other models followed by SARIMA and 2-point MA model.

RMSE Values on Rose Wine Data

Model	Parameters (Rose)	Test RMSE (Rose)
RegressionOnTime		15.27
NaiveModel		5,993.17
SimpleAvgModel		53.46
2pointTrailingMovingAverage		11.53
4pointTrailingMovingAverage		14.45
6pointTrailingMovingAverage		14.57
9pointTrailingMovingAverage		14.73
SES	Alpha = 0.098	36.80
DES	Alpha = 1.49e-8, Beta = 1.66e-8	15.27
TES	Alpha = 0.0715, Beta = 0.045, Gamma = 7.24e-5	20.16
ARIMA - Lowest AIC	(2,1,3)	36.82
SARIMA - Lowest AIC	(2,1,3) x (2,0,3,6)	27.12
ARIMA - Manual (PCF/APCF)	(2,1,2)	36.87
SARIMA - Manual (PCF/APCF)	(2,1,2) x (2,0,2,6)	30.63

Table 21 - RMSE Values of Models (Rose Wine Data)

On Rose wine dataset, n-point moving average model performs better compared to others, and many models have very little RMSE difference in them.

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

Both Sparkling & Rose wine time series being better predicted by SARIMA (at lowest AIC) model, so we will build same model using all data to predict next 12 months' outcome.

Sparkling Wine SARIMA (At AIC) Model

Trend Parameter (p,d,q) – (2,1,3)

Seasoning Parameter (P,D,Q,m) – (2,0,3,6)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                             Sparkling   No. Observations:              187
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood               -1208.621
Date:                           Sat, 23 Apr 2022   AIC                           2439.243
Time:                                   21:56:25   BIC                           2473.341
Sample:                               01-31-1980   HQIC                          2453.086
                                    - 07-31-1995
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3210      0.498      0.645      0.519      -0.655       1.297
ar.L2          0.2481      0.406      0.612      0.541      -0.547       1.043
ma.L1         -1.3728      0.747     -1.837      0.066      -2.838       0.092
ma.L2         -0.1287      0.771     -0.167      0.867      -1.640       1.382
ma.L3          0.4632      0.554      0.836      0.403      -0.623       1.550
ar.S.L6        0.0091      0.019      0.471      0.638      -0.029       0.047
ar.S.L12       1.0180      0.012     87.518      0.000       0.995       1.041
ma.S.L6       -0.3198      0.187     -1.711      0.087      -0.686       0.047
ma.S.L12      -0.8539      0.113     -7.569      0.000      -1.075      -0.633
ma.S.L18      -0.0879      0.129     -0.679      0.497      -0.342       0.166
sigma2      8.707e+04   3.51e+04      2.483      0.013    1.83e+04    1.56e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                36.66
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               1.17   Skew:                             0.50
Prob(H) (two-sided):                  0.56   Kurtosis:                         5.09
===================================================================================
```

Table 22 - SARIMA Sparkling Full Model

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1823.310454	374.978205	1088.366677	2558.254231
1995-09-30	2371.642353	380.303053	1626.262066	3117.022640
1995-10-31	3256.128140	380.526002	2510.310882	4001.945399
1995-11-30	4019.177846	381.308921	3271.826094	4766.529598
1995-12-31	6273.278578	381.709487	5525.141731	7021.415425

Table 23 - Head of Predicted Values (Sparkling Wine Sales)

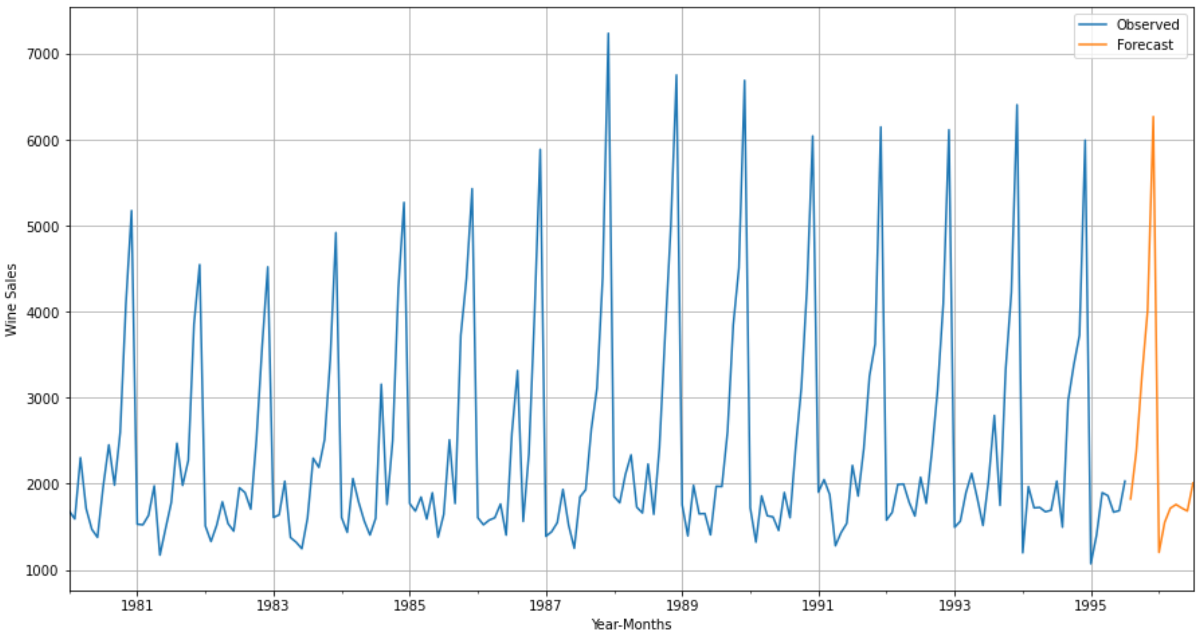


Figure XXXIX - Predicted Value Plot at 95% CI (Sparkling Wine Sales)

Rose Wine SARIMA (At AIC) Model

Trend Parameter (p,d,q) – (2,1,3)

Seasoning Parameter (P,D,Q,m) – (2,0,3,6)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  187
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -675.235
Date:                             Sat, 23 Apr 2022   AIC                           1372.470
Time:                                     22:04:33   BIC                           1406.568
Sample:                                 01-31-1980   HQIC                          1386.312
                                      - 07-31-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5266      0.060     -8.710      0.000      -0.645      -0.408
ar.L2         -0.6852      0.054    -12.778      0.000      -0.790      -0.580
ma.L1         -0.2424      0.072     -3.362      0.001      -0.384      -0.101
ma.L2          0.2346      0.074      3.189      0.001       0.090       0.379
ma.L3         -0.7580      0.075    -10.100      0.000      -0.905      -0.611
ar.S.L6       -0.0544      0.034     -1.624      0.104      -0.120       0.011
ar.S.L12       0.8636      0.034     25.280      0.000       0.797       0.931
ma.S.L6        0.1711     36.778      0.005      0.996     -71.912      72.254
ma.S.L12      -0.6540     30.480     -0.021      0.983     -60.393      59.085
ma.S.L18       0.1746      6.464      0.027      0.978     -12.494      12.844
sigma2       197.7000   7268.318      0.027      0.978    -1.4e+04    1.44e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.35   Jarque-Bera (JB):                18.29
Prob(Q):                              0.56   Prob(JB):                         0.00
Heteroskedasticity (H):               0.21   Skew:                            -0.30
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.52
===================================================================================
```

Table 24 - SARIMA Rose Wine Sales Full Model

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	52.736646	14.292107	24.724631	80.748662
1995-09-30	45.723229	14.668511	16.973476	74.472982
1995-10-31	48.576440	14.905834	19.361542	77.791338
1995-11-30	54.282017	14.908419	25.062054	83.501980
1995-12-31	72.336184	15.004548	42.927811	101.744557

Table 25 - Head of Predicted Values (Rose Wine Sales)

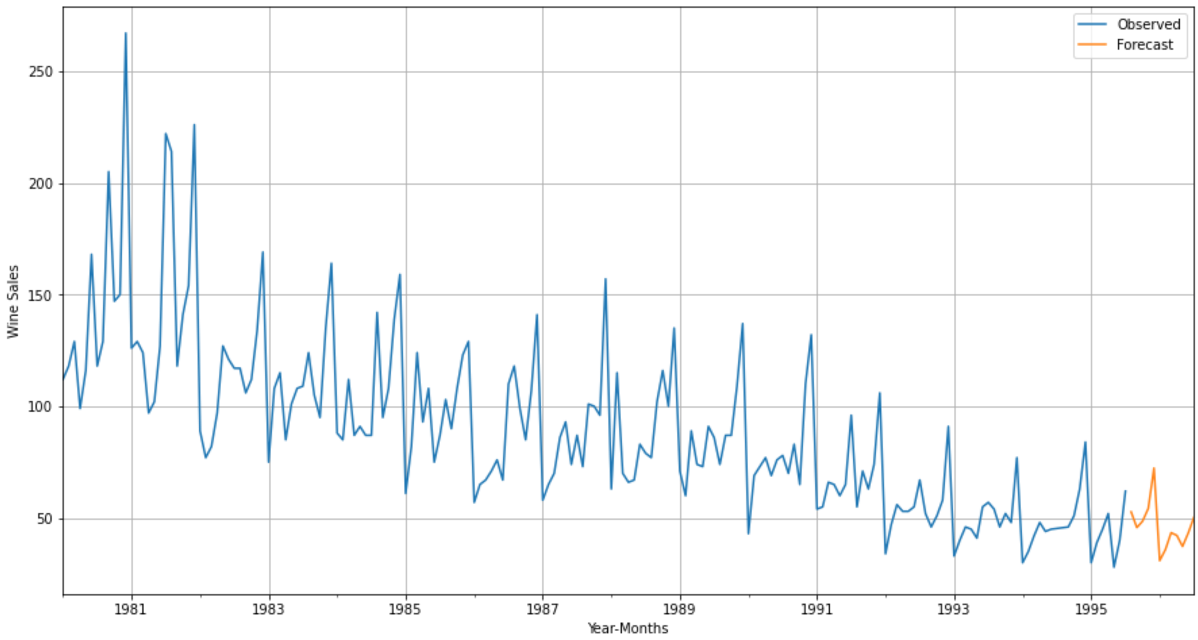


Figure XL - Predicted Value Plot at 95% CI (Rose Wine Sales)

Both 12 months' data prediction was done with 95% confidence interval.

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

Insights

- Rose wine data has 2 missing values at July'94 & Aug'94, we imputed them using linear interpolation.
- Sparkling wine doesn't show a trend, but a significant seasonality can be seen there.
- Rose wine sales has seasonality as well as a down trend, which means people are moving away (not linking) from Rose wine.
- Spikes can be seen in mid & year-end, they are seasons at which people usually drink more and sales goes higher (It can be interpreted as, in winters [year-end], people need to warm themselves and sales would go higher)

Actions (Model building)

- Linear Regression Model – Presence of seasonality on both sales data induce error in prediction from this model, as Linear regression model can work better in trending data.
- Naïve Forecast Model – The model predicts future based on last known values, hence it doesn't work on time series where trend or seasonality is present (does not work efficiently on our wine sales data)

- Simple Average Model – Takes overall average for prediction, so futuristic trend and seasonality gets ignored, since we have both in wine sales data, this model also has large RMSE
- N-point trailing Moving Average Model – by Moving average models, we are trying to follow both trend and seasonality, the model works better on lower n-points and smoothens with increased n-points
- Exponential Smoothing:
  - SES (Simple Exponential Smoothing) – This model works when we don't have trend or seasonality in time series, with both present on sales data, we have high RMSE computed from this model.
  - DES (Double Exponential Smoothing) – Works when we have trend but no seasonality, but we have seasonality present on both Sparkling and Rose wine sales, it doesn't give us better RMSE.
  - TES (Triple Exponential Smoothing) – Works better when we have both trend and seasonality available in our data, and keep RMSE on lower side.
- ARIMA/SARIMA – Both sales data have seasonality, hence SARIMA works better compared to ARIMA.

Recommendations

- ABC Estate company that produces Rose wine should adjust their flavours or try giving discount to attract customers in purchasing more.
- ABC Estate should offer an increasing discount on successive purchases of Sparkling wine, so that they'll have an up-trend in their sales.
- Company can offer more discount on non-seasonal months and prices can be adjusted when seasonality comes in play.