Business Report

Time Series

October 2021

GREAT LEARNING

Post Graduate Program in Data Science and Business Analytics [Online]
Authored by: Athisya Nadar

Table of Contents

List of Figures

List of Tables

Problem Statement:

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

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

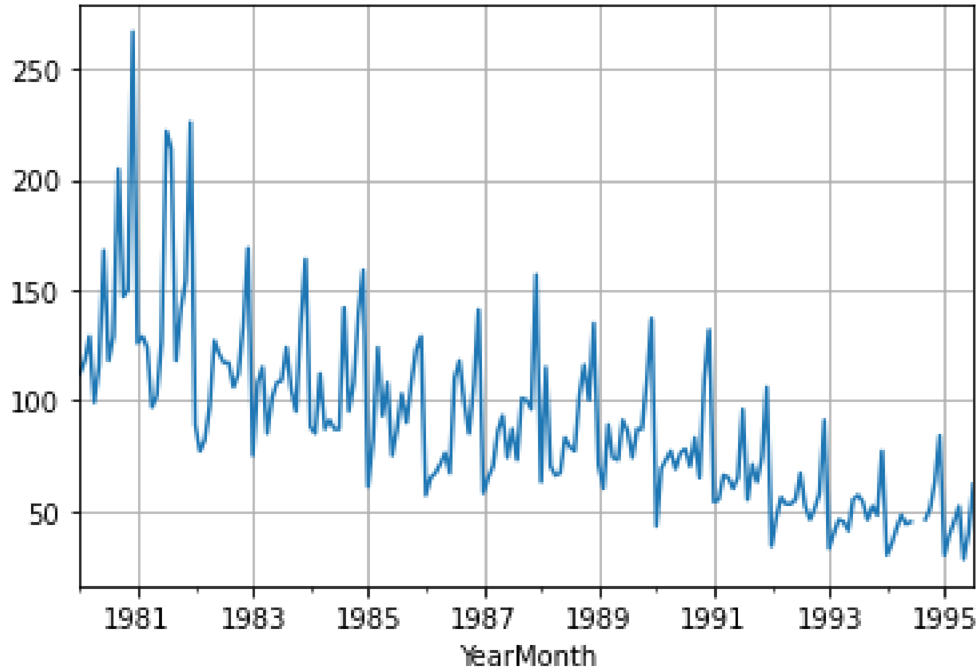
We have got the data of Rose Wines and Sparkling Wines in the 20^{th} century. Both of these data are from same company for ABC Estate Wines. We have to analyze and forecast Wine Sales in the 20^{th} century. Forecast is a statistical method to predict an attribute using historical patterns in the data.

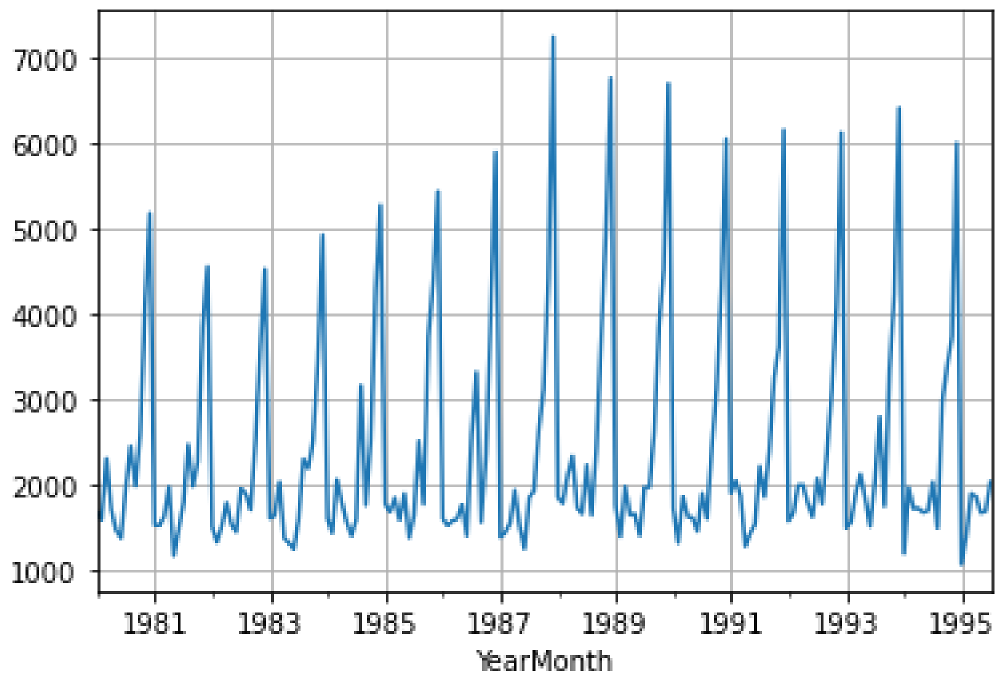
A collection of observations that has been observed at regular time intervals for a certain variable over a given duration is called a time series. Time series data has several characteristics that make it unique. These characteristics can be stated
below as: -

- All observations are dependent.
- Missing data must be imputed
- Two different types of intervals cannot be mixed

Approaches used for Time Series Forecasting: The following are two major approaches to time series forecasting.
I. Decomposition: This method is based on extraction of individual components of time series.
II. Regression: This method is based on regression on past observations.

The three important components are:
I. Trend (Long term movement)
II. Seasonal component: Intra-year stable fluctuations repeatable over the entire length of the series
III. Irregular component (Random movements)

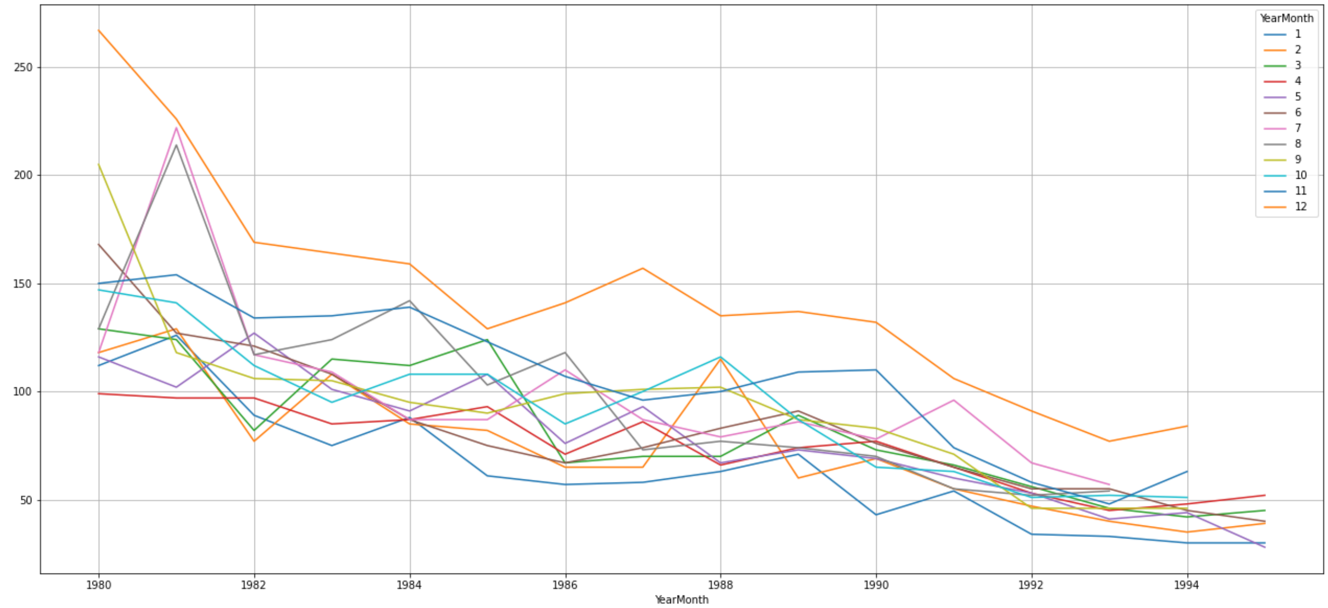


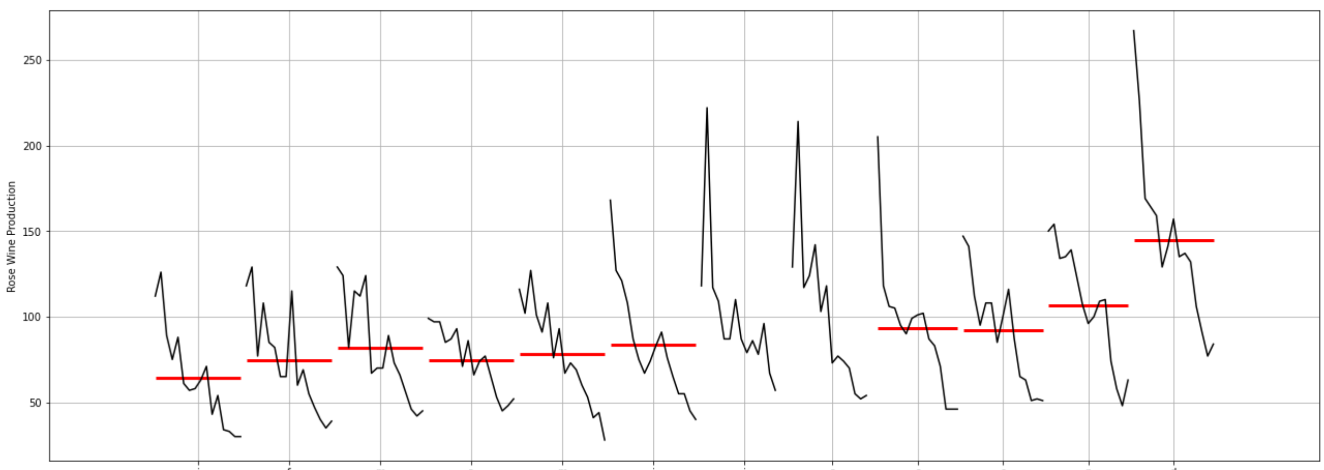


For rose wine sales, the numbers are decreasing in sales, implying presence of trend component. Intra-year stable fluctuations are indicative of seasonal component. While sparkling wine has increasing trend and multiplicative seasonality.

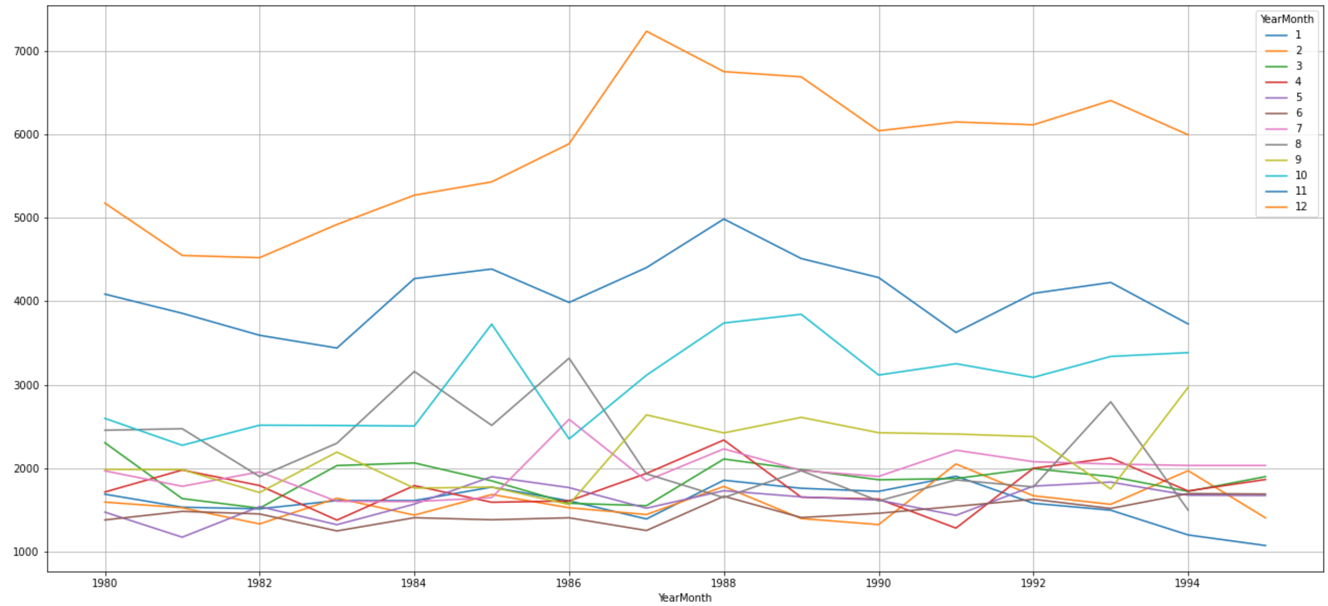
2. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

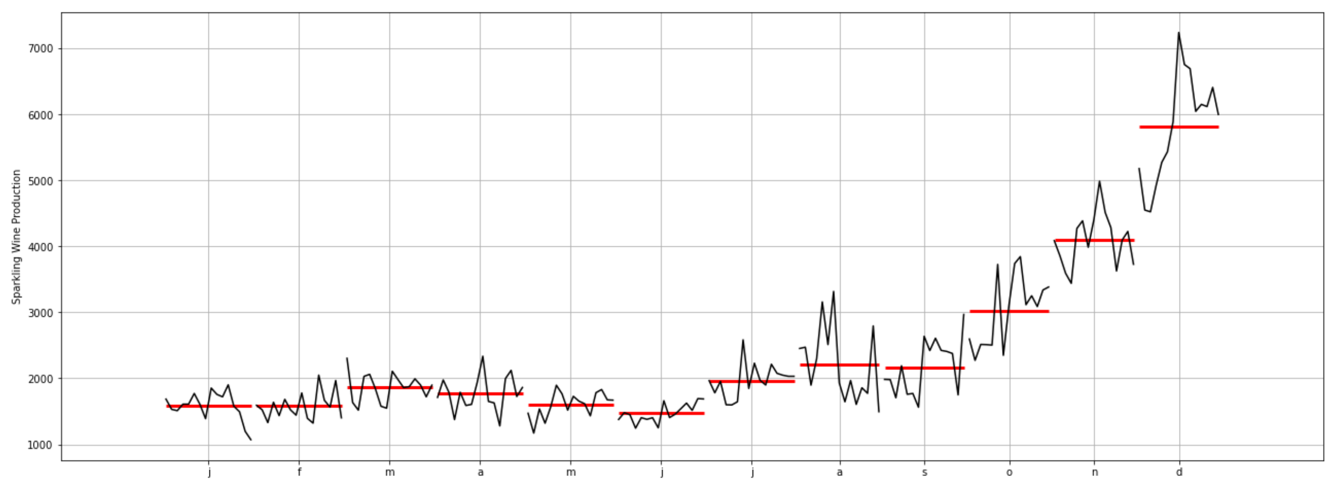
Following two plots will help us in identifying the seasonal fluctuations better:





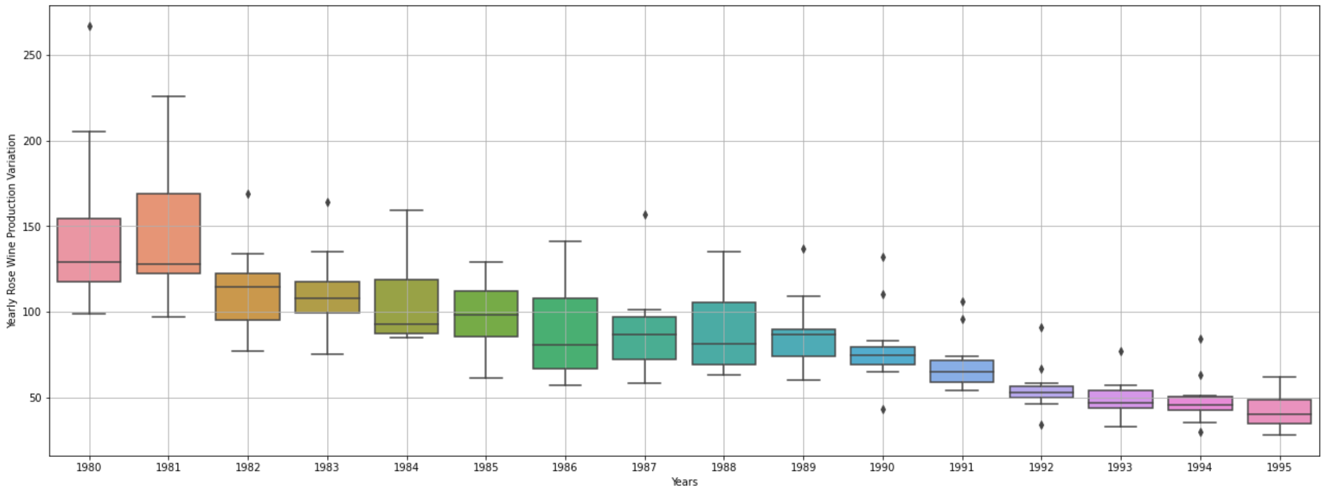
The sales of Rose wine are decreasing every year in number. The vertical lines represent monthly sales and the horizontal lines represent average sales of the given month. In all these above plots the decreasing lines that represent sales have seasonal fluctuations along with a trend.

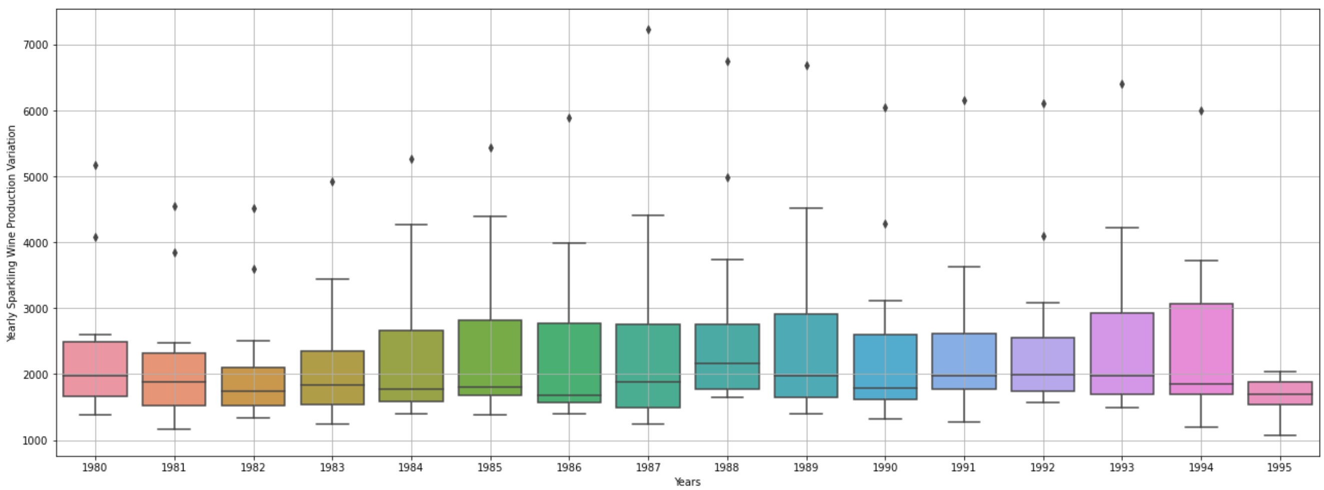




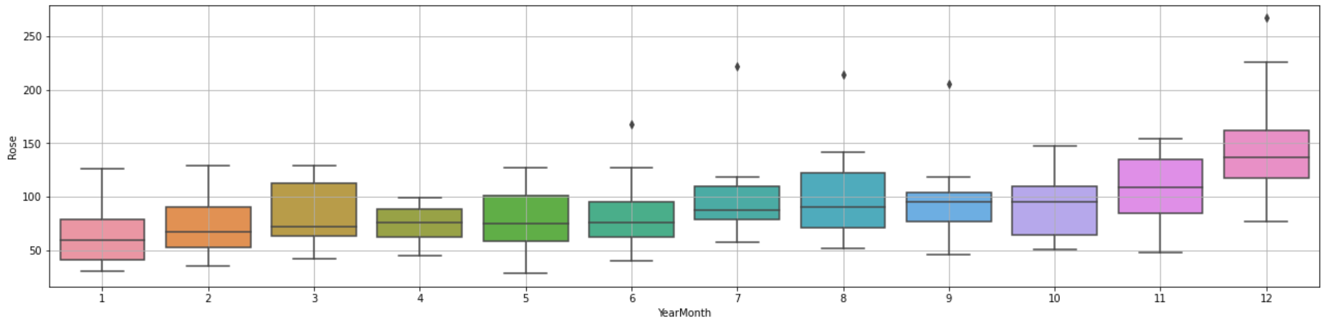
The sales of Sparkling wine are increasing every year in number. The vertical lines represent monthly sales and the horizontal lines represent average sales of the given month. In all these above plots the increasing lines that represent sales have seasonal fluctuations along with a trend.

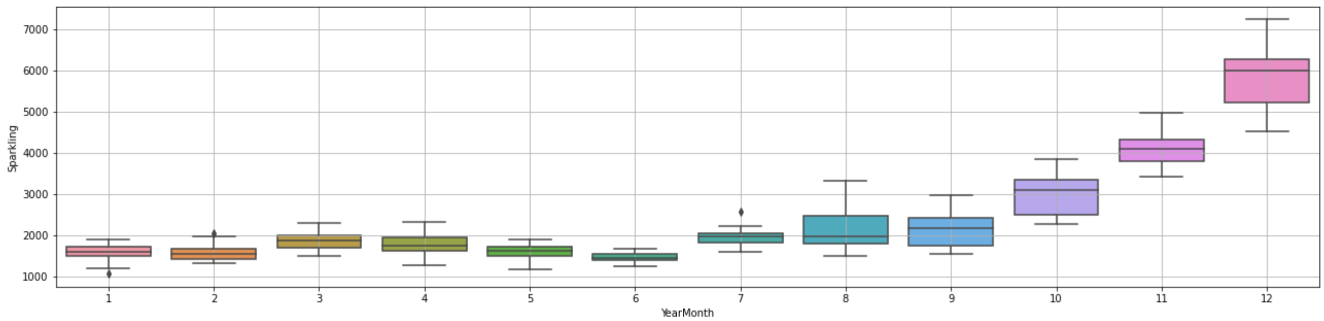
Let us plot a year-on-year boxplot for the Wine production:





Let us plot a monthly boxplot for the Wine production taking all the years into account.



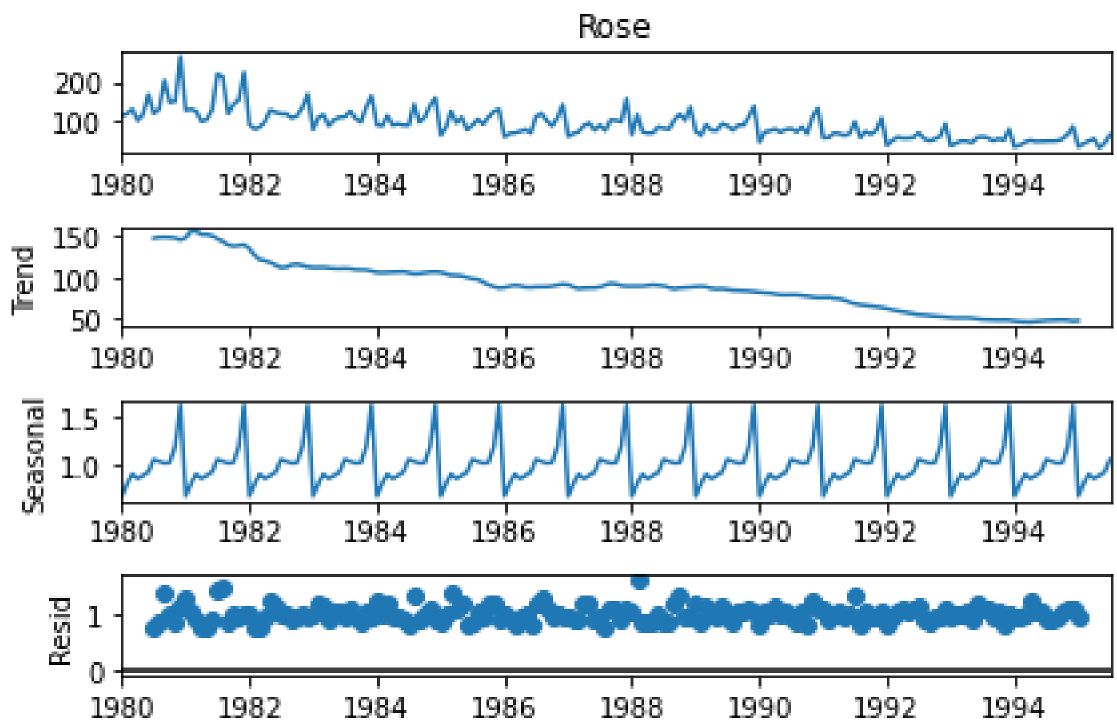


Let us look at the table of sales for rose and sparkling wine:

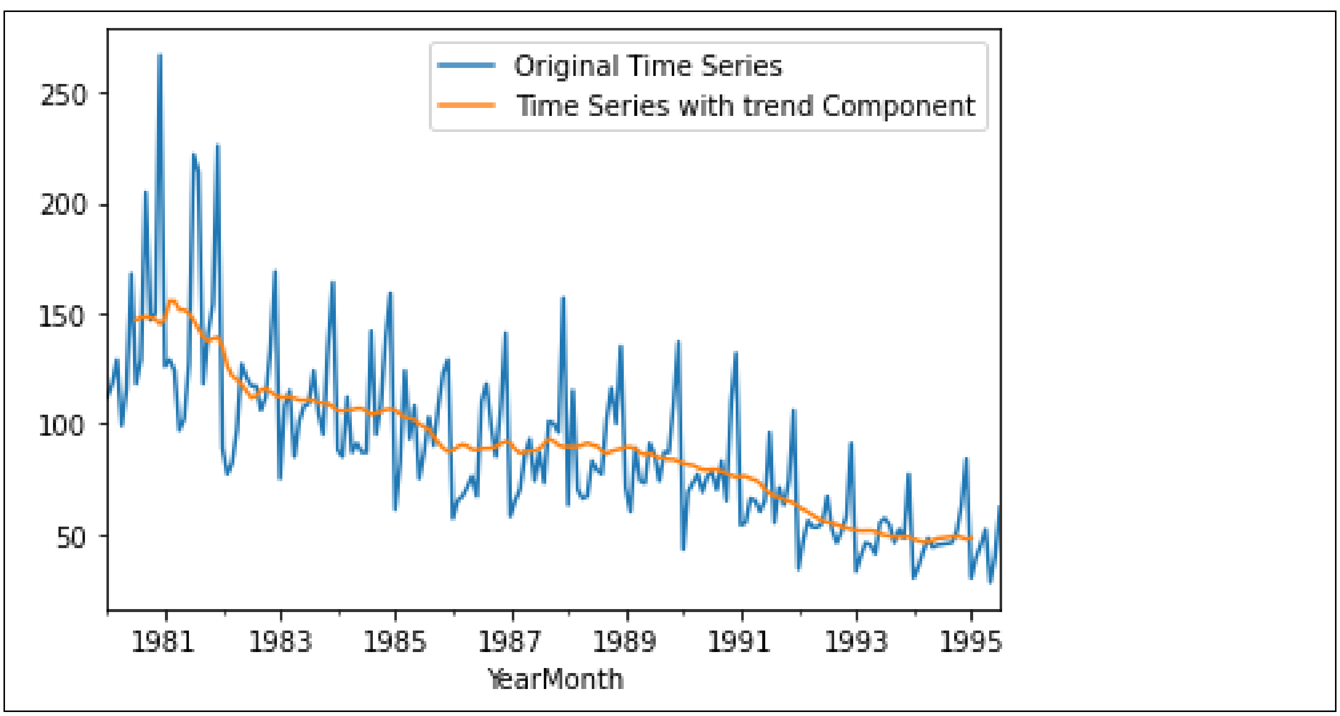
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	NaN	NaN	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

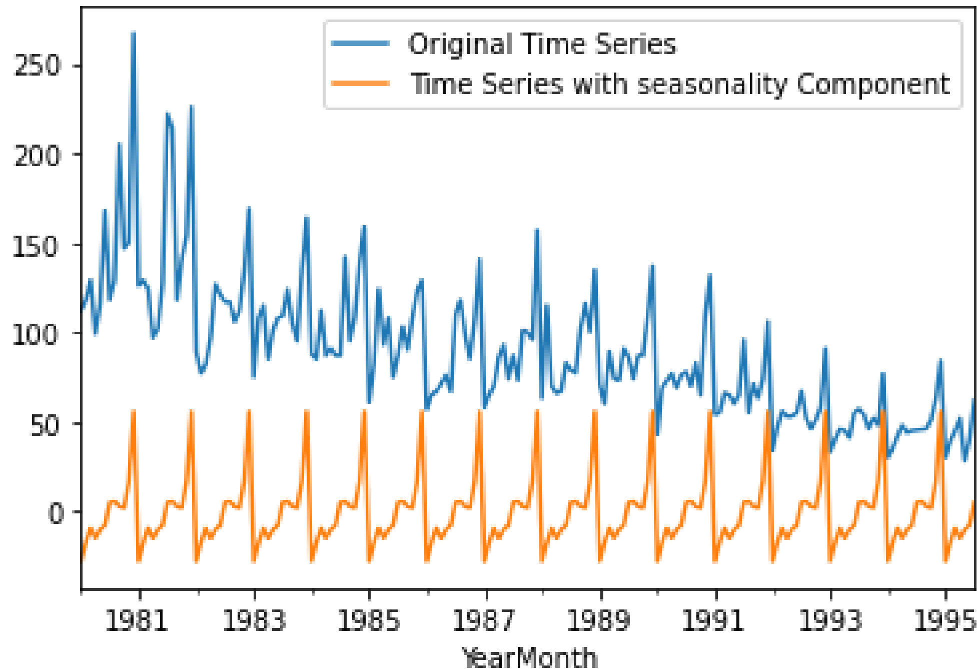
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

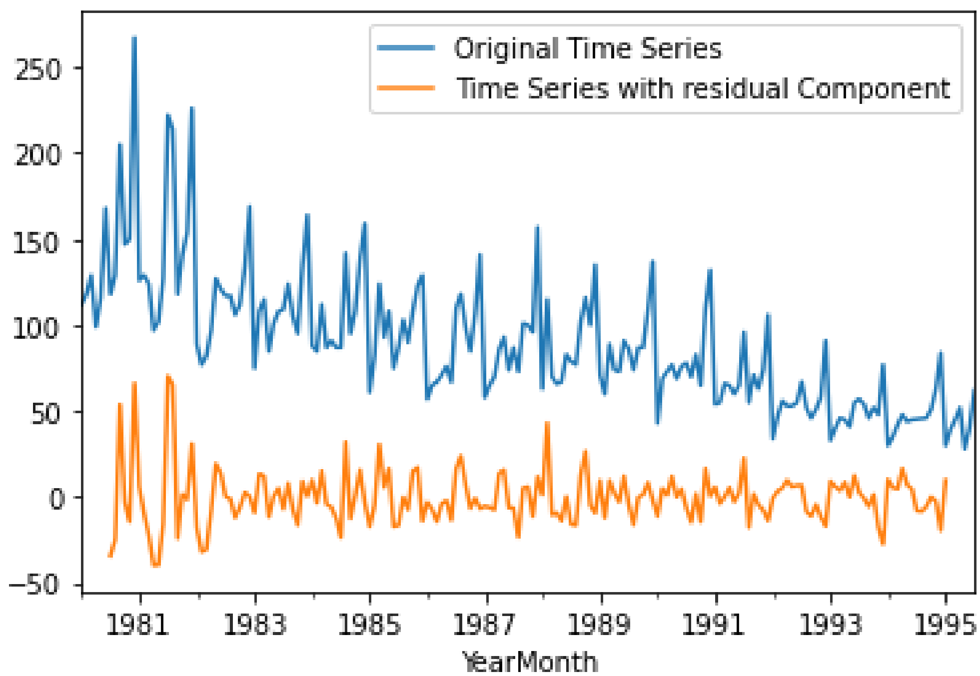
Now let us decompose the Rose Wine sales data and look at the components:



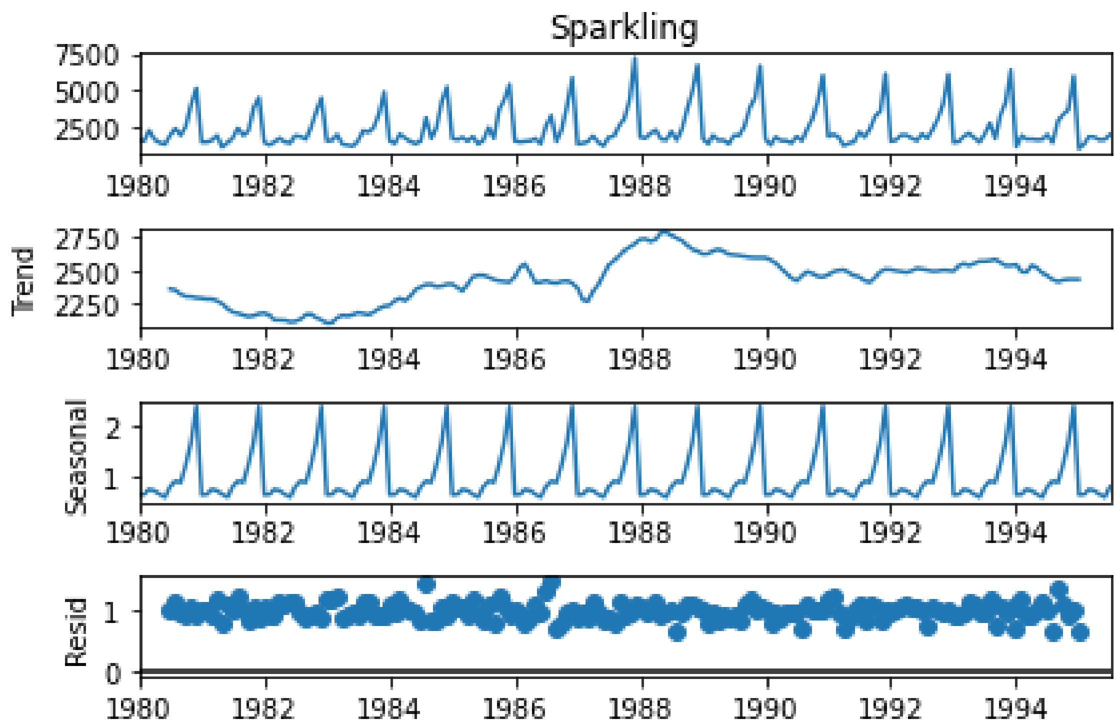
When we plot the components against the original series the trend can be viewed:



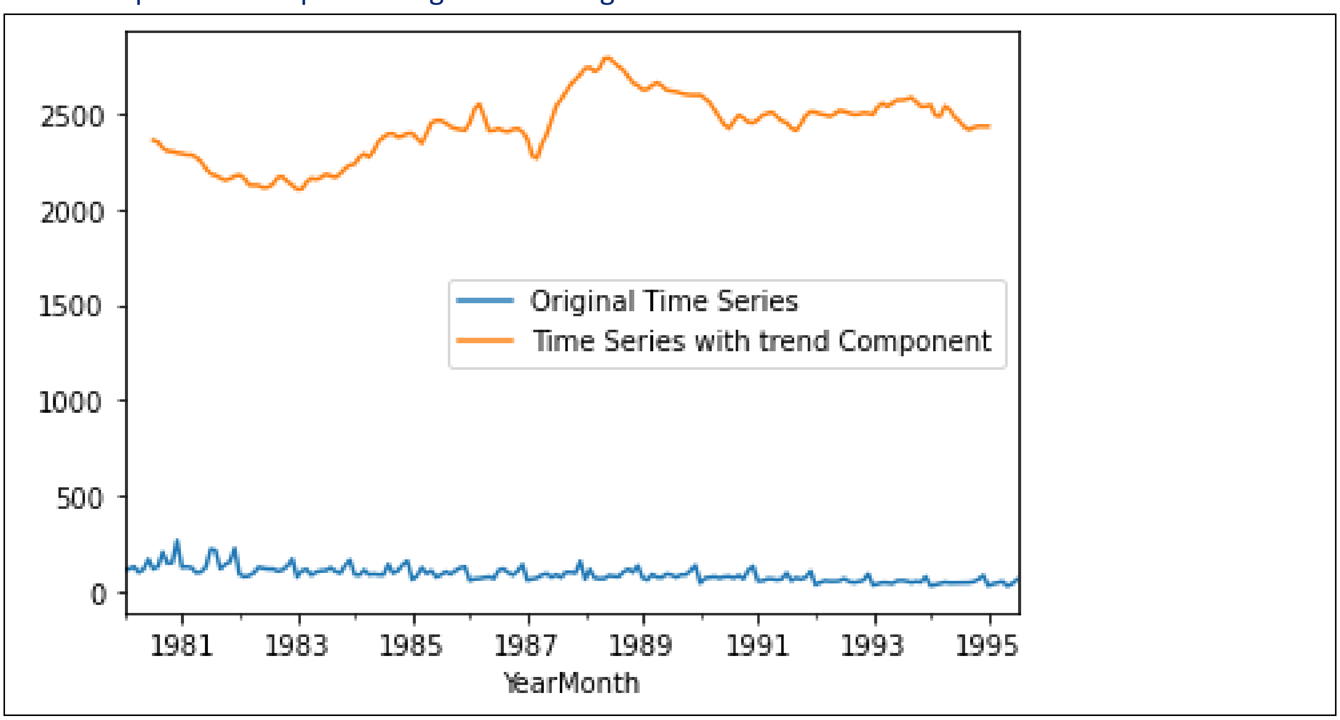


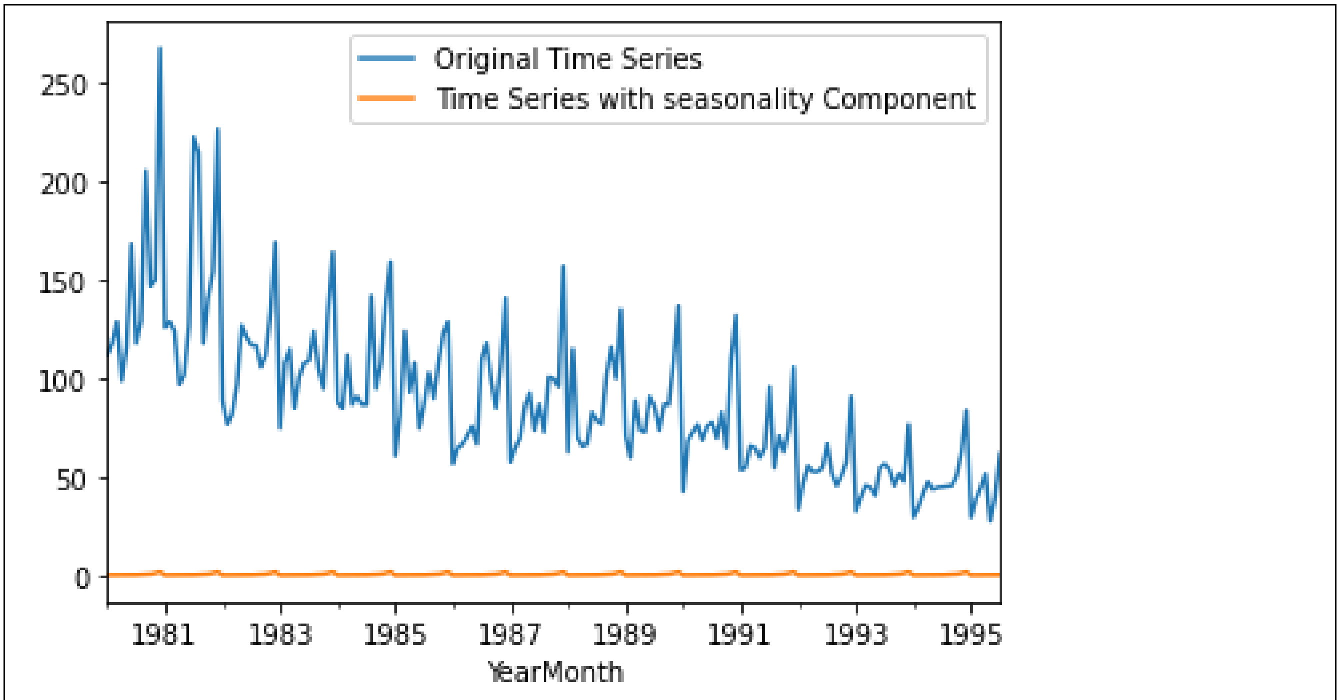


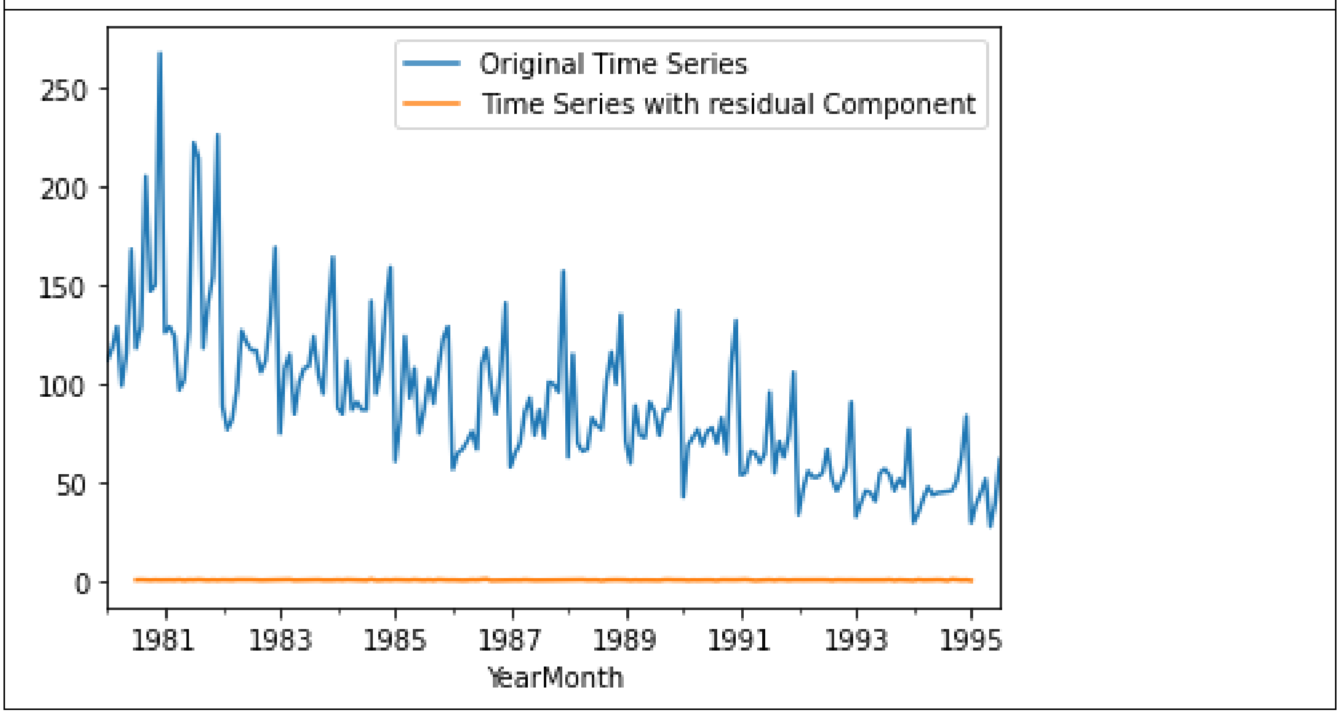
Now let us decompose the Sparkling Wine sales data and look at the components:



When we plot the components against the original series the trend can be viewed:



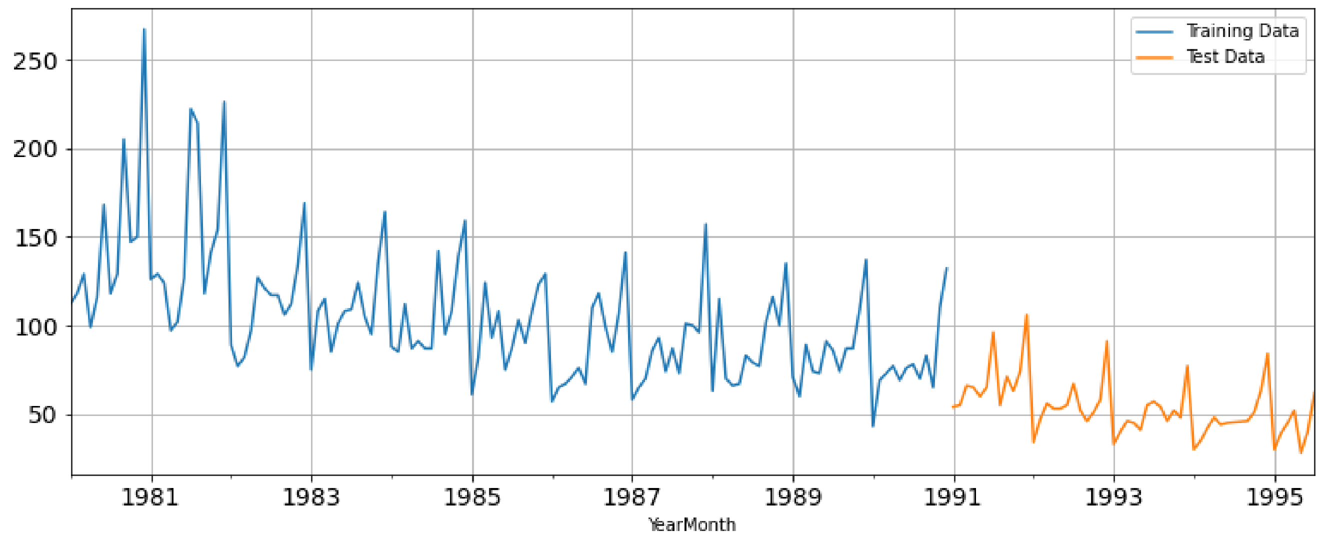


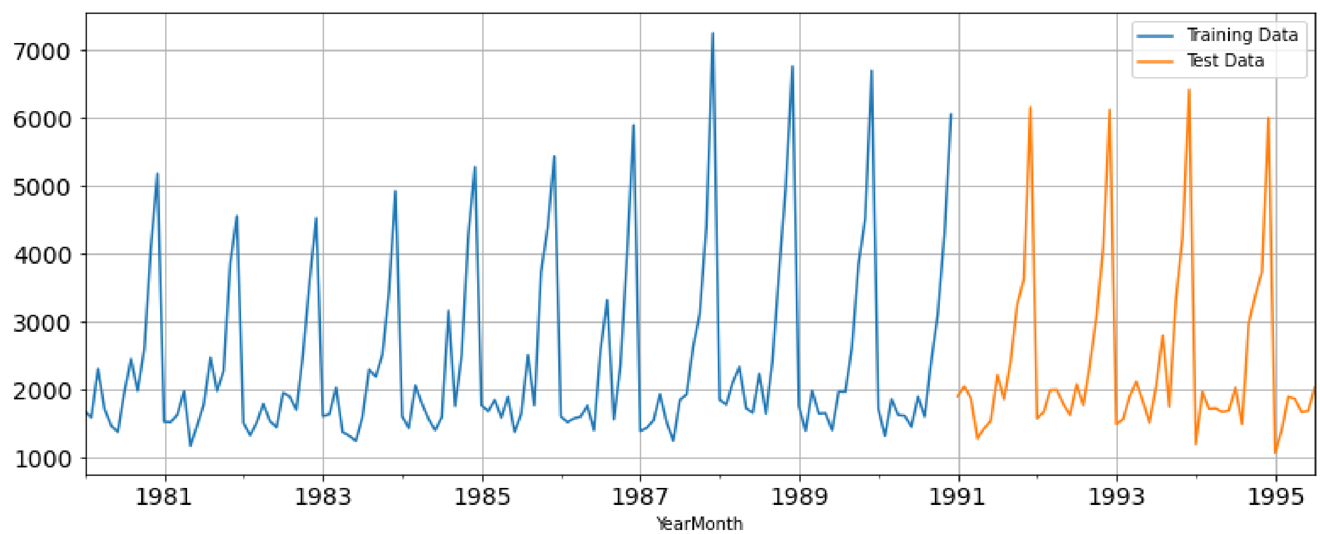


Now decomposition method is applied to identify and separate out the three components (i.e trend, seasonality and irregular components) from the given series to observe their independent properties.

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

For Wine Sales series, the data till 1991 is used for training purpose and the data after 1991 is used for testing purpose. After splitting train and test data the data for Sparkling and Rose wine sales can be plotted as follows:





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

1) Holt-Winter's method (Triple Exponential Smoothing)

This table represents a summary table of decomposition methods stating its conditions, advantages and disadvantages:

Table 1 represents a summary table of decomposition methods stating its conditions, advantages and disadvantages.

Table 1: Summary Table for Exponential Smoothing Models

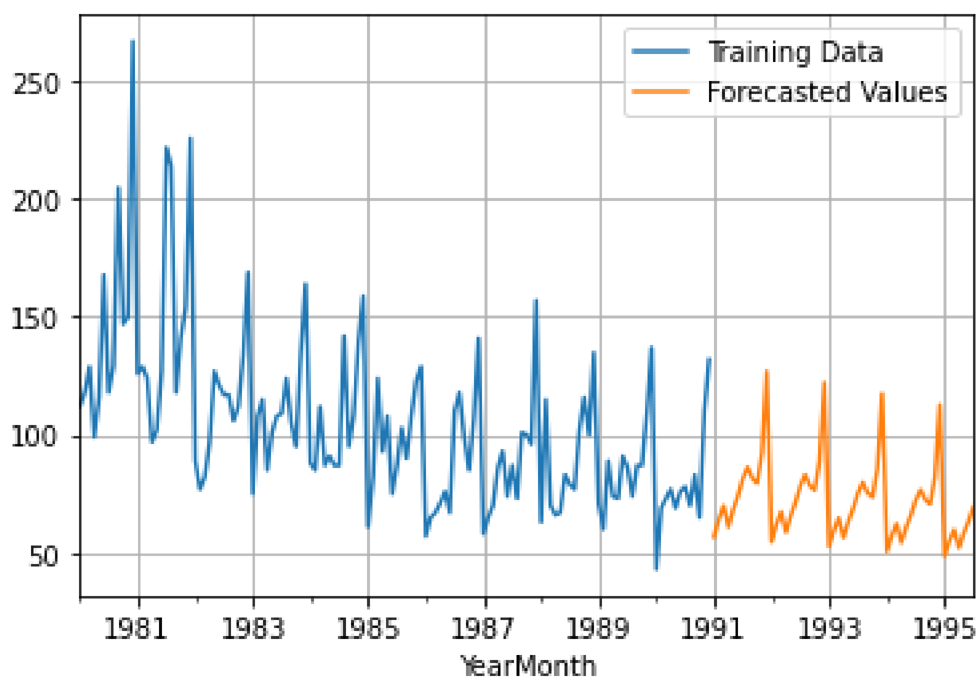
Method	**When to consider a particular method?**		**Advantages**	**Disadvantages**
	There is trend in data	**There is seasonality in data**		
Decomposition Method	May/May not be	May/may not be	Helps in understanding pattern of time series components individually.	Does not generate values for a few initial and last data points
Simple Exponential method	No	No	Suitable when data has no clear presence of trend and seasonality	Takes time in calibrating value of parameter α
Holt's Method	Yes	No	Adjusts level and trend both	Takes time in calibrating value of parameter α, β.
Holt Winter's Method	Yes	Yes	Adjust level, trend and seasonality simultaneously	Takes time in calibrating value of parameter α, β, γ.

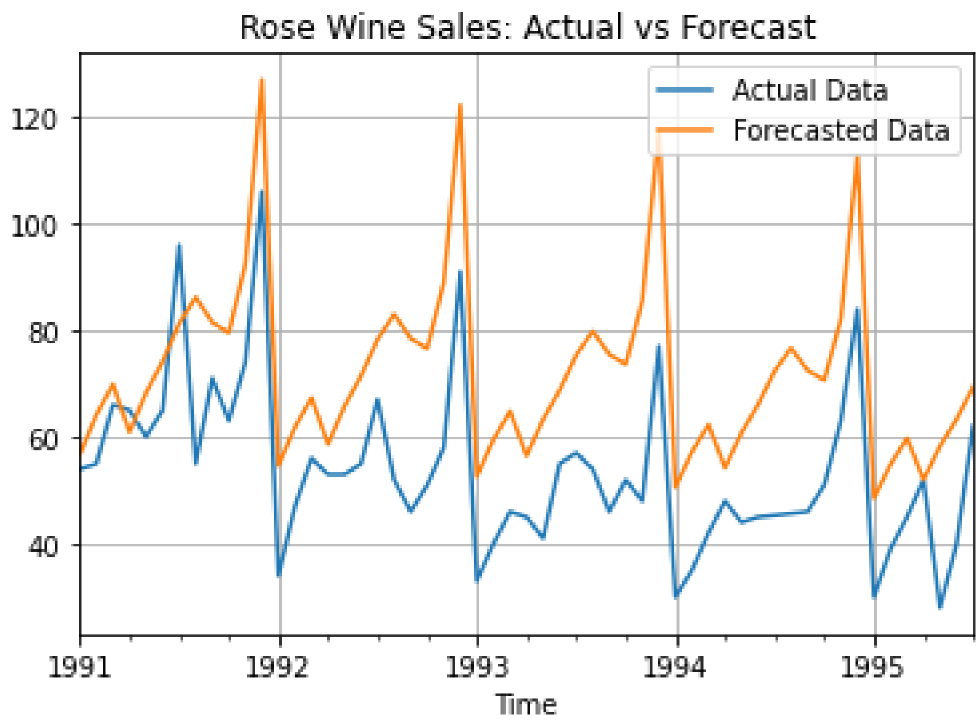
Since there is trend and seasonality we will consider Holt-Winter's method (Triple Exponential Smoothing) :

This is an extension of Holt's method when seasonality is found in the data.
Forecast equation:$Yt+1=lt+bt+st-m(k+1)$
Level Equation:$lt=\alpha(Yt-st-m)+\alpha(1-\alpha)Yt-1$, $0 < \alpha < 1$
Trend Equation:$bt=\beta(lt-lt-1)+(1-\beta)bt-1$, $0 < \beta < 1$
Seasonal Equation: $\gamma(Yt-lt-1-bt-1)+(1-\gamma)st-m$, $0 < \gamma < 1$

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, β and γ. This is a general method and a true multi-step ahead forecast.

	name	param	optimized
smoothing_level	alpha	0.065694	True
smoothing_trend	beta	0.051929	True
smoothing_seasonal	gamma	0.000004	True
initial_level	l.0	54.109855	True
initial_trend	b.0	-0.334720	True
initial_seasons.0	s.0	2.082823	True
initial_seasons.1	s.1	2.363267	True
initial_seasons.2	s.2	2.582102	True
initial_seasons.3	s.3	2.257027	True
initial_seasons.4	s.4	2.537575	True
initial_seasons.5	s.5	2.766400	True
initial_seasons.6	s.6	3.041018	True
initial_seasons.7	s.7	3.234346	True
initial_seasons.8	s.8	3.067473	True
initial_seasons.9	s.9	3.001641	True
initial_seasons.10	s.10	3.498938	True
initial_seasons.11	s.11	4.825525	True





	Test RMSE	Test MAPE
TripleExponentialSmoothing	21.01962	38.7431

Based on a rule of thumb, it can be said that RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately.

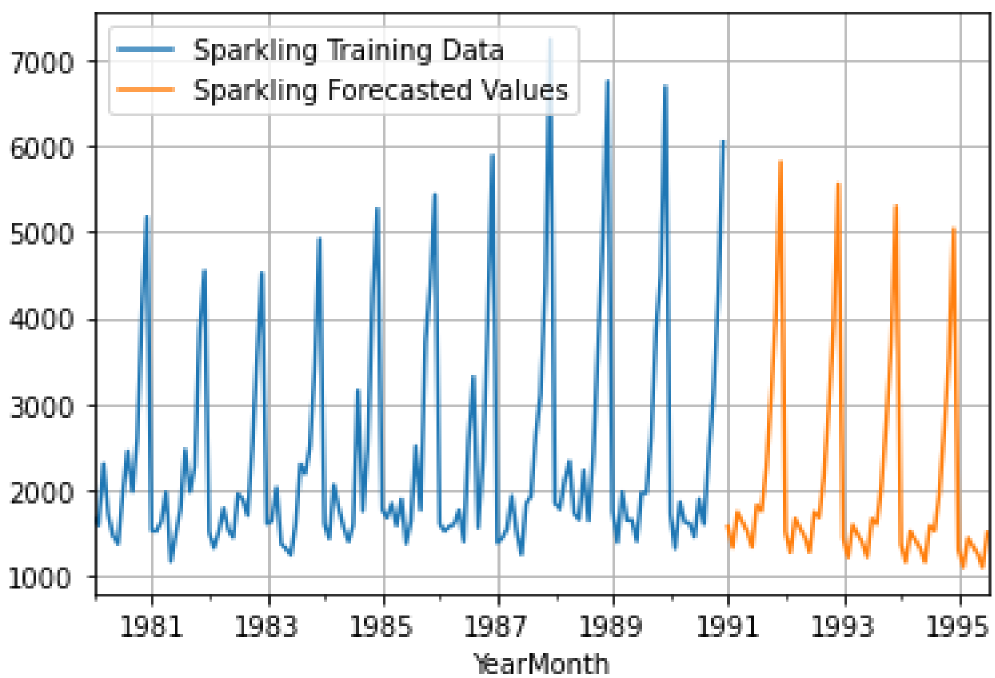
MAPE value of 0 to 10% is good, 10 to 20 is average and anything above 20 is worst.
So clearly this model is giving us only reasonable forecasting.

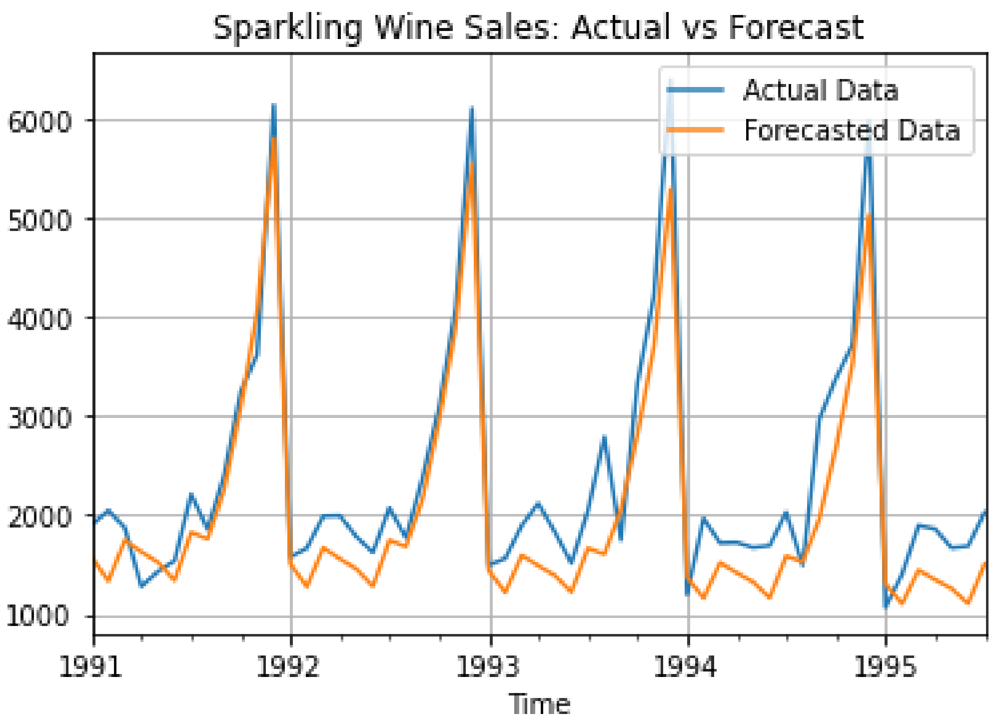
MAPE	Interpretation
<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

Source: Lewis (1982, p. 40)

Now let us look at the sparkling Wine data:

	name	param	optimized
smoothing_level	alpha	0.111108	True
smoothing_trend	beta	0.061729	True
smoothing_seasonal	gamma	0.395048	True
initial_level	l.0	1639.934066	True
initial_trend	b.0	-12.224946	True
initial_seasons.0	s.0	1.064020	True
initial_seasons.1	s.1	1.023521	True
initial_seasons.2	s.2	1.406719	True
initial_seasons.3	s.3	1.201655	True
initial_seasons.4	s.4	0.975930	True
initial_seasons.5	s.5	0.971002	True
initial_seasons.6	s.6	1.318974	True
initial_seasons.7	s.7	1.695889	True
initial_seasons.8	s.8	1.389529	True
initial_seasons.9	s.9	1.814764	True
initial_seasons.10	s.10	2.851500	True
initial_seasons.11	s.11	3.624705	True

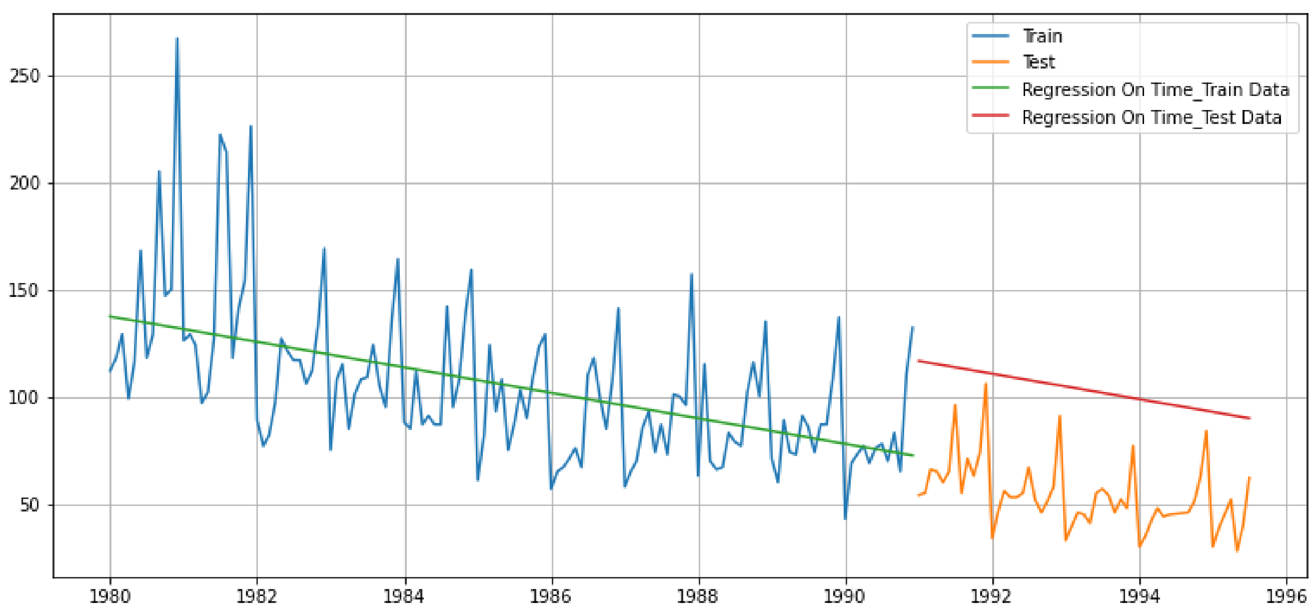




	Test RMSE	Test MAPE
TripleExponentialSmoothing	469.76797	17.580255

Since the Test MAPE is between 10 and 20 we have achieved good forecasting for sparkling Wine using Triple Exponential Smoothening model.

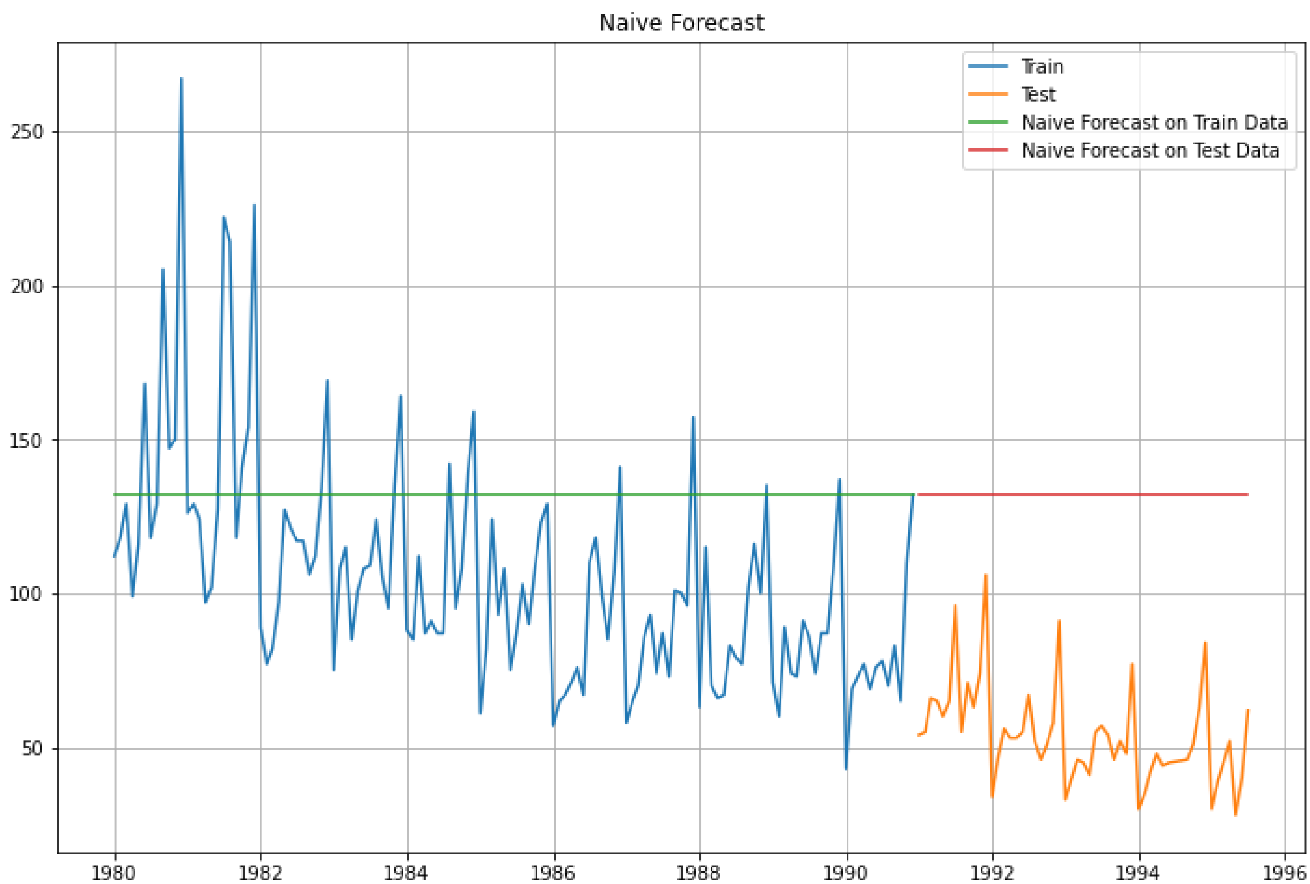
2) Linear Regression:



	Test RMSE	Test MAPE
RegressionOnTime	51.433312	91.64

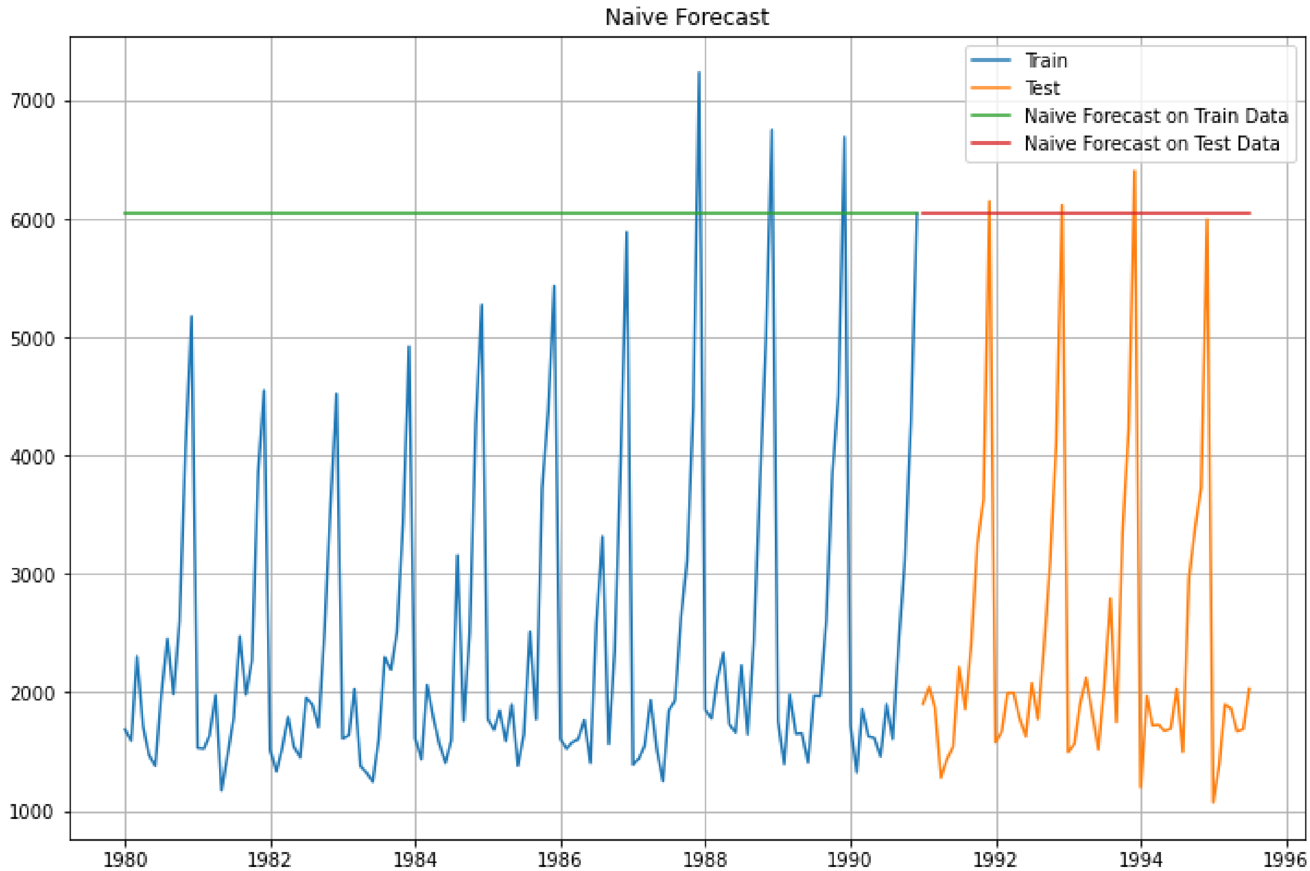
For linear Regression on Rose Wine Sales Data, we have got inaccurate model

Model 2: Naive Approach:



	Test RMSE	Test MAPE
RegressionOnTime	51.433312	91.64
NaiveModel	79.718773	145.10

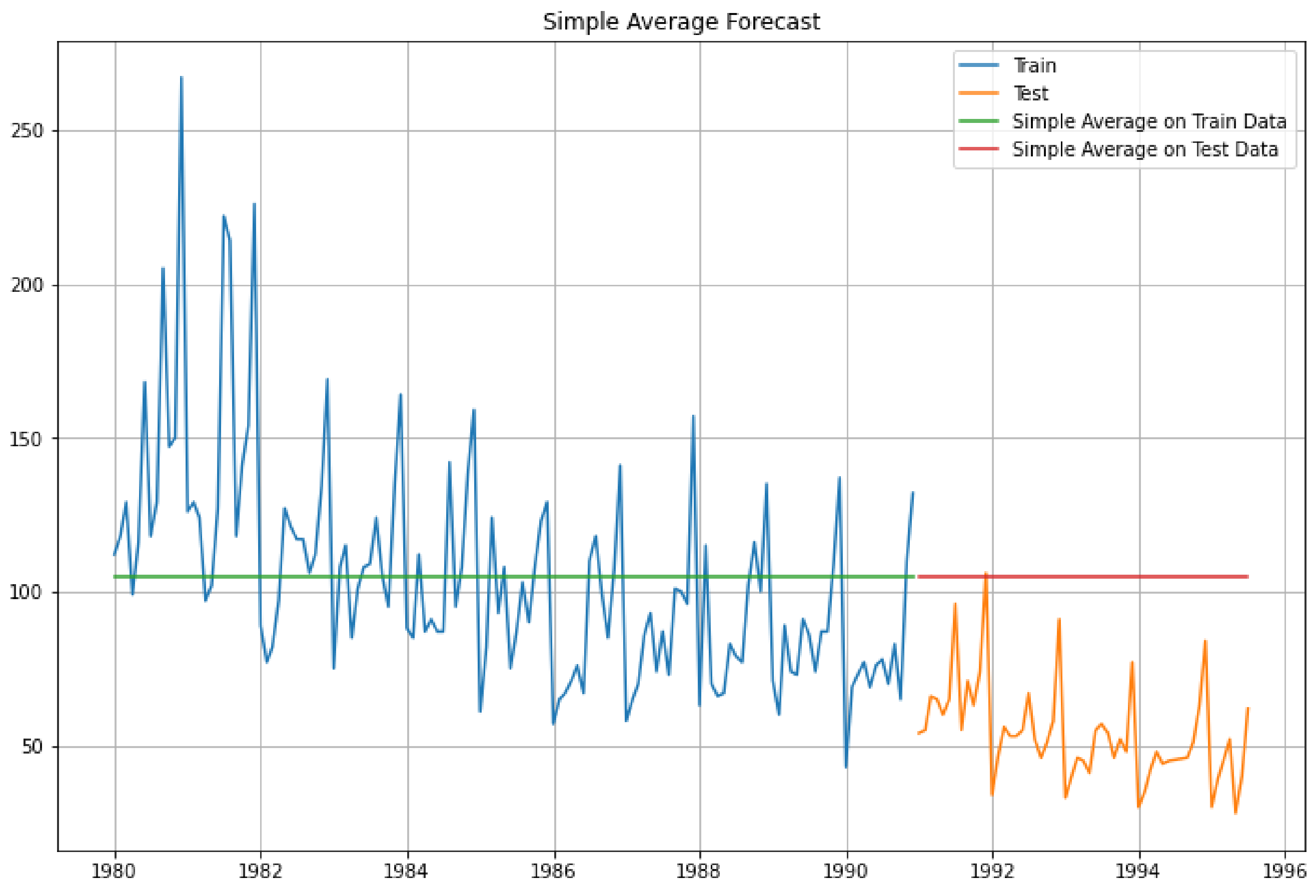
Again, for naïve approach for Rose Wine we have not got good RMSE or MAPE, it is a overfit model.



	Test RMSE	Test MAPE
RegressionOnTime	1275.867052	39.16
NaiveModel	3864.279352	152.87

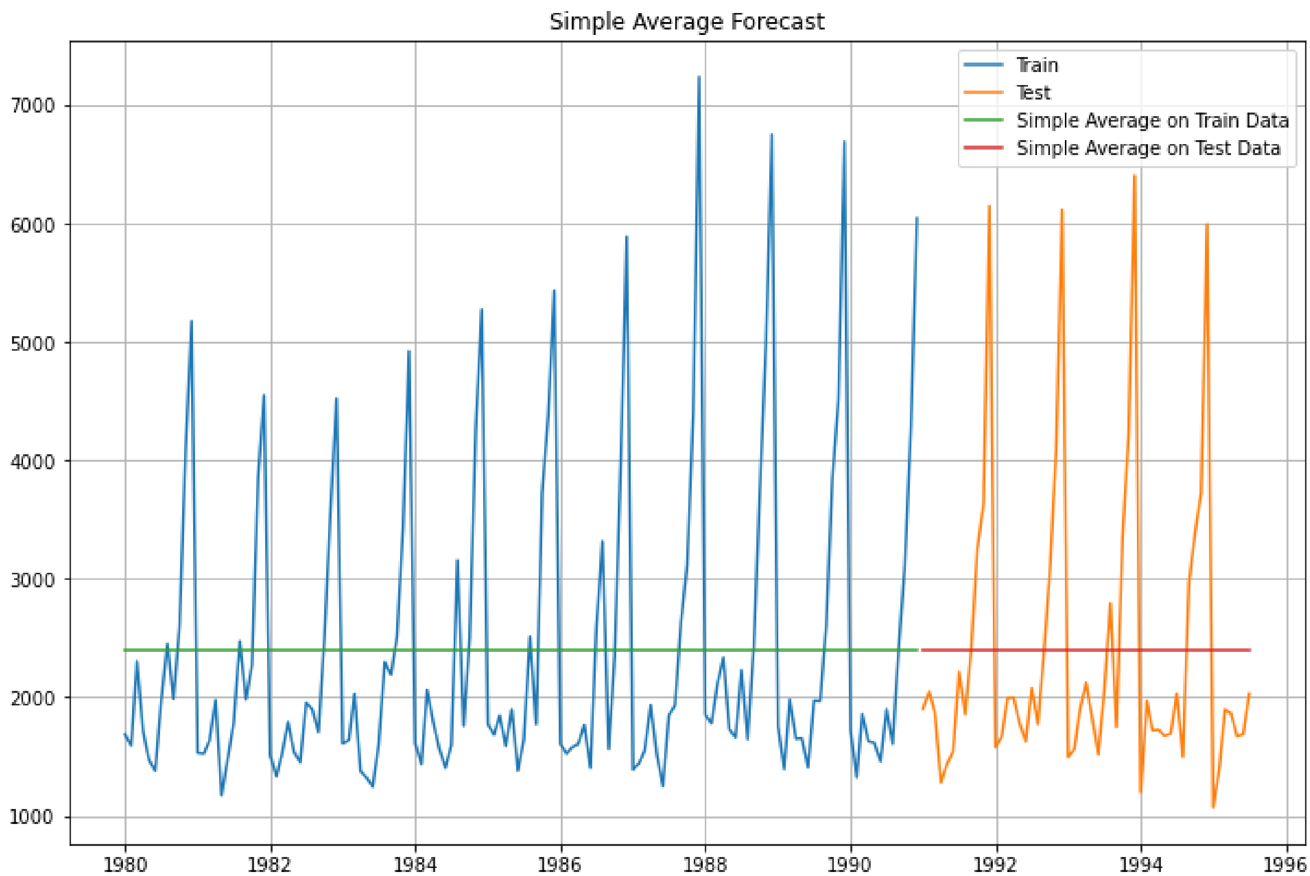
For sparkling wine data compared to Regression naïve approach has yielded inaccurate results

Method 3: Simple Average



	Test RMSE	Test MAPE
RegressionOnTime	51.433312	91.64
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93

For Rose Wine data, the forecast by Simple Average method is giving worst model in terms of RMSE and MAPE.



	Test RMSE	Test MAPE
RegressionOnTime	1275.867052	39.16
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90

For Sparkling Wine data, the forecast by Simple Average method is giving reasonable model in terms of MAPE.

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

ARIMA: Auto Regressive Integrated Moving Average (ARIMA) models are applied on time series data when the current value is assumed to be correlated to past values and past prediction errors. Therefore, these models are used in defining current value as a linear combination of past values and past prediction errors.

A series is said to be stationary if its mean and variance are constant over a period of time and, the correlation between the two time periods depends only on the distance or lag between the two periods. Mathematically, let Yt be a time series with these properties:

Mean: $E(Yt)=\mu$

Variance: $Var(Yt)=E(Yt-\mu)$ 2=σ 2

Correlation: $\rho k=E[(Yt-\mu)(Yt+k-\mu)/(\sigma t\sigma t+k)]$

Where ρk is the correlation (or auto-correlation) at lag k between the values of Yt and $Yt+k$ So, if mean, variance and correlation (or auto-correlation) of time series data is constant (at different lags) no matter at what point of time it is measured; i.e. if they are time invariant, the series is called a stationary time series. A series not possessing these properties is termed as a non-stationary time series.

Since ARIMA model requires a stationary series, a formal stationarity test needs to be applied to the time series under consideration.

Augmented Dickey-Fuller Test: A formal test to check whether time series data follows stationary series.

H0: Time series is non-stationary

H1: Time series is stationary

Results of Dickey-Fuller Test: (Rose Wine)

Test Statistic	-1.876699
p-value	0.343101
#Lags Used	13.000000
Number of Observations Used	173.000000
Critical Value (1%)	-3.468726
Critical Value (5%)	-2.878396
Critical Value (10%)	-2.575756

Results of Dickey-Fuller Test: (Sparkling Wine)

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

We see that at 5% significant level the Time Series is non-stationary in case of both Rose wine and Sparkling Wine.

If the series is non-stationary, stationarize the Time Series by taking a difference of the Time Series. Then we can use this particular differenced series to train the ARIMA models. We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there. You can look at other kinds of transformations as part of making the time series stationary like taking logarithms.

Hence, a stationarization is necessary. Often differencing a non-stationary time series leads to a stationary series. After differencing,

Results of Dickey-Fuller Test: (Rose Wine)
Test Statistic -7.966534e+00
p-value 2.855044e-12
#Lags Used 1.200000e+01
Number of Observations Used 1.700000e+02
Critical Value (1%) -3.469413e+00
Critical Value (5%) -2.878696e+00
Critical Value (10%) -2.575917e+00

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

We see that at alpha = 0.05 the Time Series is indeed stationary.

6. **Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

$ARIMA(p,d,q)$ Model: ARIMA is defined by 3 parameters
p : No of autoregressive terms

d : No of differencing to stationarize the series
q: No of moving average terms

ACF and PACF used together to identify the order of the ARMA. Seasonal ACF and PACF examines correlations for seasonal data.

When the current value of variable can be expressed as a linear function of its past values then, it is known as an auto-regression process. PACF is used for identifying the value of p.

When the current value of the series is a function of past forecast errors this model is known as a moving average model. ACF is used for identifying the value of q.

ARMA model is a combination of two basic processes i.e. AR and MA.

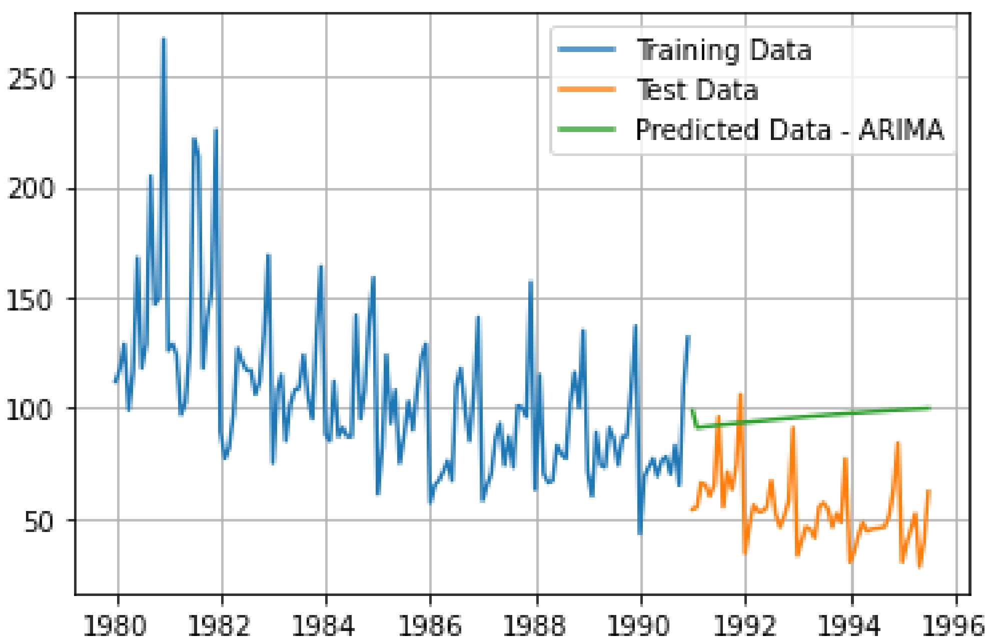
$\boldsymbol{ARIMA}$ $(\boldsymbol{p,d,q})$: ARIMA model is an advance version of ARMA model where $d > 0$ indicates that the original series is non-stationary and d differencing is required to make is stationary. We will also define the models with centered Yt, centering is done by subtracting the mean of the stationary series.

SARIMA(p,d,q)(P,D,Q)[m]:
Seasonal ARIMA model with seasonal frequency m:
Seasonal ARIMA models are more complex models with seasonal adjustments. These models are used when time series data has significant seasonality. The most general form of seasonal ARIMA is $ARIMA(p,d,q)*ARIMA(P,D,Q)$[m], where P, D, Q are defined as seasonal AR component, seasonal difference and seasonal MA component respectively. And, 'm'represents the frequency (time interval) at which the data is observed. For example, a monthly series will have m = 12.

Seasonal ACF and PACF may be used to understand seasonality.

	param	AIC
5	(1, 0, 2)	1292.053210
8	(2, 0, 2)	1292.248055
7	(2, 0, 1)	1292.937195
4	(1, 0, 1)	1294.510585
3	(1, 0, 0)	1301.546304
6	(2, 0, 0)	1302.346074
1	(0, 0, 1)	1305.468406
2	(0, 0, 2)	1306.586679
0	(0, 0, 0)	1324.899703

```
                              ARMA Model Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                     ARMA(1, 2)   Log Likelihood                -641.027
Method:                       css-mle   S.D. of innovations             30.999
Date:                Sun, 07 Nov 2021   AIC                           1292.053
Time:                        00:11:34   BIC                           1306.467
Sample:                    01-01-1980   HQIC                          1297.910
                         - 12-01-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        107.8488     15.807      6.823      0.000      76.867     138.830
ar.L1.Rose     0.9861      0.018     53.660      0.000       0.950       1.022
ma.L1.Rose    -0.6873      0.098     -6.989      0.000      -0.880      -0.495
ma.L2.Rose    -0.2007      0.094     -2.129      0.033      -0.386      -0.016
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0141           +0.0000j            1.0141            0.0000
MA.1            1.1009           +0.0000j            1.1009            0.0000
MA.2           -4.5247           +0.0000j            4.5247            0.5000
-----------------------------------------------------------------------------
```

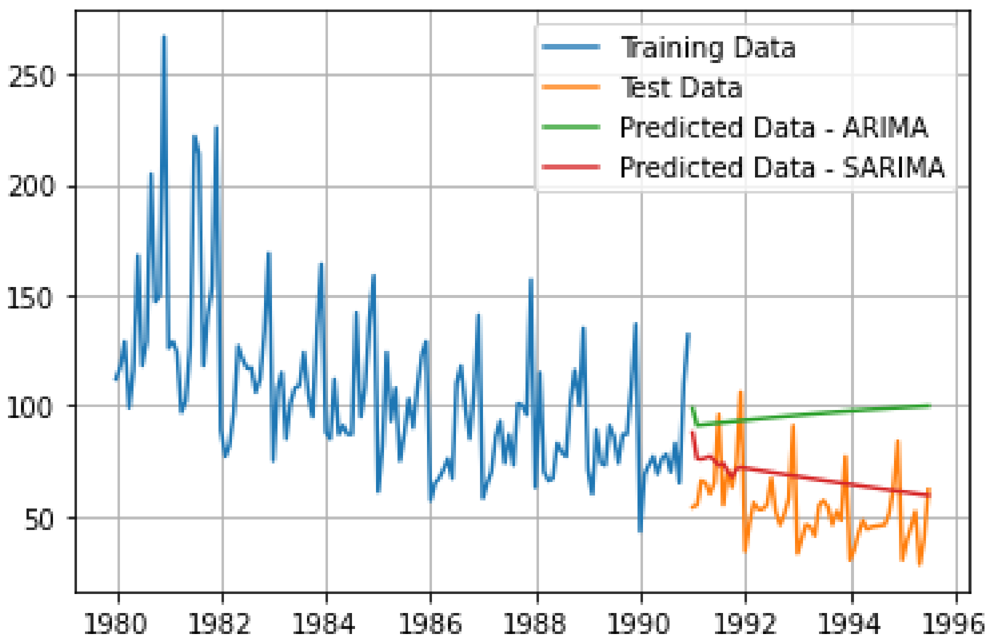


	param	seasonal	AIC
47	(1, 0, 2)	(0, 0, 2, 5)	1141.298095
50	(1, 0, 2)	(1, 0, 2, 5)	1149.743093
74	(2, 0, 2)	(0, 0, 2, 5)	1150.498512
53	(1, 0, 2)	(2, 0, 2, 5)	1150.915100
77	(2, 0, 2)	(1, 0, 2, 5)	1151.455691
...	...	...	...
10	(0, 0, 1)	(0, 0, 1, 5)	1380.987202
18	(0, 0, 2)	(0, 0, 0, 5)	1426.844550
1	(0, 0, 0)	(0, 0, 1, 5)	1453.701942
9	(0, 0, 1)	(0, 0, 0, 5)	1481.819865
0	(0, 0, 0)	(0, 0, 0, 5)	1607.530754

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                132
Model:             SARIMAX(1, 0, 2)x(0, 0, 2, 5)   Log Likelihood                -564.649
Date:                              Sun, 07 Nov 2021   AIC                           1141.298
Time:                                      00:11:58   BIC                           1157.973
Sample:                                  01-01-1980   HQIC                          1148.069
                                       - 12-01-1990                                         
Covariance Type:                                opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9955      0.001    908.989      0.000       0.993       0.998
ma.L1         -0.7507      0.116     -6.460      0.000      -0.978      -0.523
ma.L2         -0.2493      0.080     -3.118      0.002      -0.406      -0.093
ma.S.L5        0.0524      0.092      0.569      0.569      -0.128       0.233
ma.S.L10      -0.0918      0.117     -0.784      0.433      -0.321       0.138
sigma2       745.1061      0.000   4.83e+06      0.000     745.106     745.106
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                11.67
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.57   Skew:                             0.45
Prob(H) (two-sided):                  0.08   Kurtosis:                         4.24
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.26e+21. Standard errors may be unstable.
```

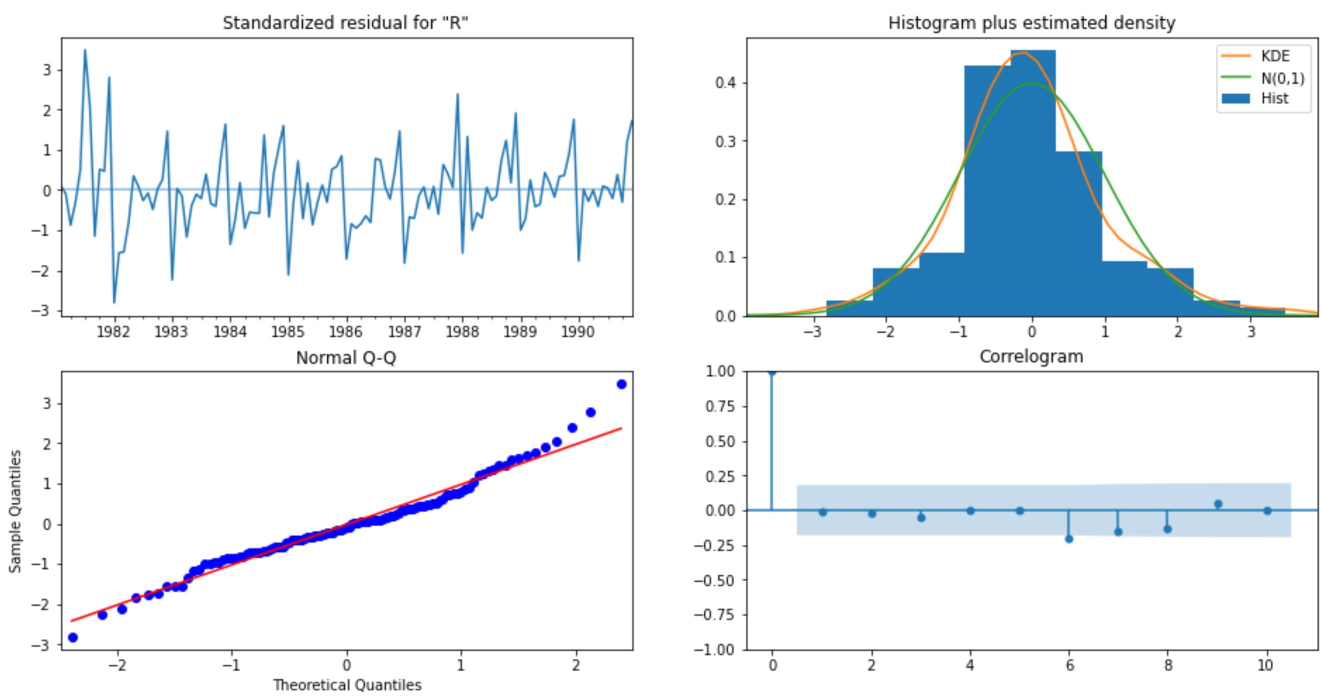
	Test RMSE
ARIMA(1, 0, 2)	45.450517
SARIMA(1, 0, 2)(0, 0, 2)5	20.055333



For Rose Wine Comparing RMSE for ARIMA and SARIMA, we can say that SARIMA is a better model because 20% is better than 45%.

In terms of AIC, lower the AIC better the model.

1141.298 < 1292.053 , again SARIMA is a better model.



Now let us check the Sparkling Wine Data:

	param	AIC
8	(2, 0, 2)	2200.904841
7	(2, 0, 1)	2236.590818
6	(2, 0, 0)	2244.799915
1	(0, 0, 1)	2245.268851
2	(0, 0, 2)	2245.343218
4	(1, 0, 1)	2245.949094
5	(1, 0, 2)	2246.012193
3	(1, 0, 0)	2247.348276
0	(0, 0, 0)	2271.203212

```
                              ARMA Model Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                     ARMA(1, 2)   Log Likelihood               -1118.006
Method:                       css-mle   S.D. of innovations           1152.409
Date:                Sun, 07 Nov 2021   AIC                           2246.012
Time:                        02:21:38   BIC                           2260.426
Sample:                    01-01-1980   HQIC                          2251.869
                         - 12-01-1990
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
const            2401.9307     79.985     30.030      0.000    2245.163    2558.699
ar.L1.Sparkling     0.7623      0.101      7.520      0.000       0.564       0.961
ma.L1.Sparkling    -0.3897      0.100     -3.908      0.000      -0.585      -0.194
ma.L2.Sparkling    -0.4260      0.067     -6.393      0.000      -0.557      -0.295
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.3119           +0.0000j            1.3119            0.0000
MA.1            1.1415           +0.0000j            1.1415            0.0000
MA.2           -2.0564           +0.0000j            2.0564            0.5000
-----------------------------------------------------------------------------
```

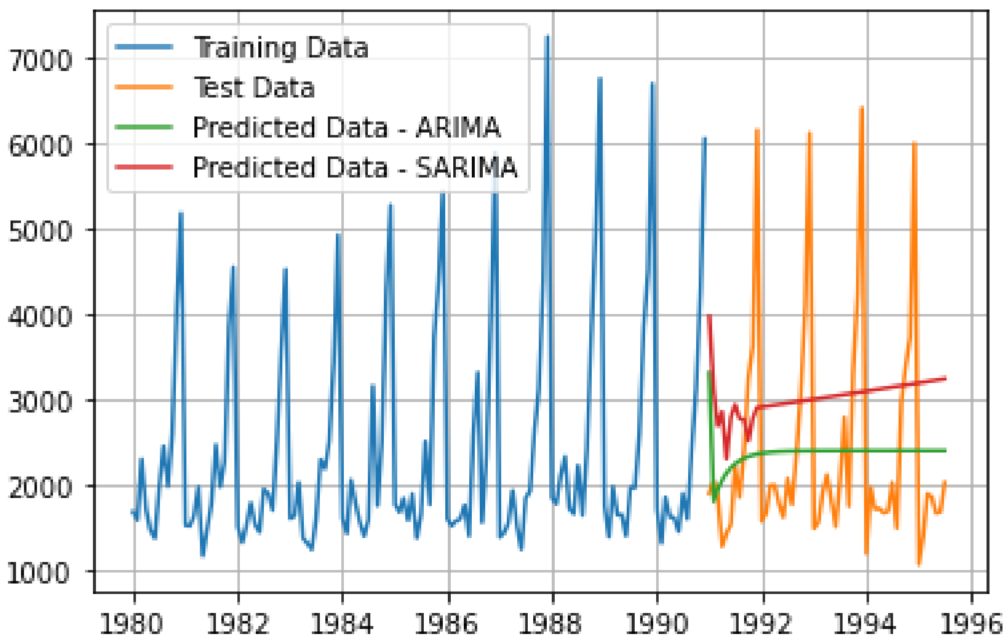
	param	seasonal	AIC
47	(1, 0, 2)	(0, 0, 2, 5)	1141.298095
50	(1, 0, 2)	(1, 0, 2, 5)	1149.743093
74	(2, 0, 2)	(0, 0, 2, 5)	1150.498512
53	(1, 0, 2)	(2, 0, 2, 5)	1150.915100
77	(2, 0, 2)	(1, 0, 2, 5)	1151.455691
...	...	...	...
10	(0, 0, 1)	(0, 0, 1, 5)	1380.987202
18	(0, 0, 2)	(0, 0, 0, 5)	1426.844550
1	(0, 0, 0)	(0, 0, 1, 5)	1453.701942
9	(0, 0, 1)	(0, 0, 0, 5)	1481.819865
0	(0, 0, 0)	(0, 0, 0, 5)	1607.530754

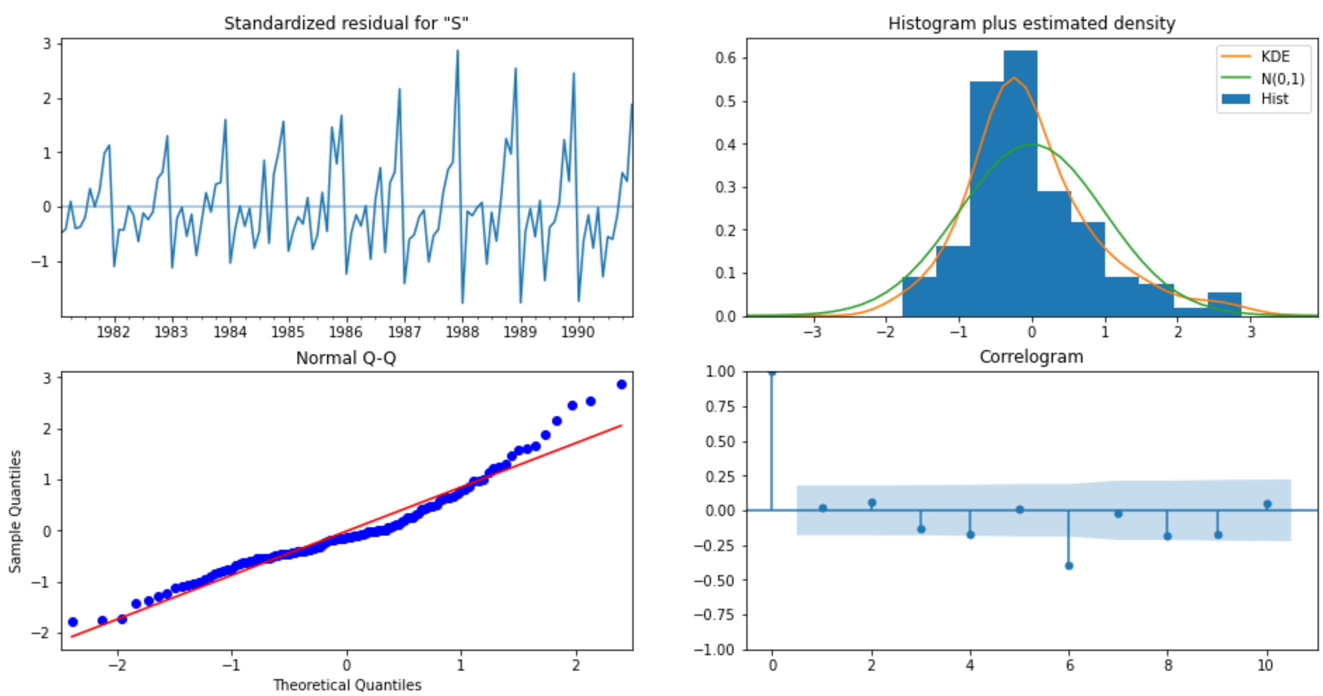
```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  132
Model:             SARIMAX(1, 0, 2)x(0, 0, 2, 5)   Log Likelihood               -1011.756
Date:                            Sun, 07 Nov 2021   AIC                           2035.513
Time:                                    02:24:45   BIC                           2052.187
Sample:                                01-01-1980   HQIC                          2042.284
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0026      0.002    517.869      0.000       0.999       1.006
ma.L1         -0.6705      0.141     -4.762      0.000      -0.947      -0.395
ma.L2         -0.3189      0.195     -1.633      0.103      -0.702       0.064
ma.S.L5       -0.2750      0.209     -1.315      0.189      -0.685       0.135
ma.S.L10      -0.1401      0.158     -0.884      0.377      -0.451       0.171
sigma2       1.77e+06   7.94e-08   2.23e+13      0.000    1.77e+06    1.77e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                21.42
Prob(Q):                              0.85   Prob(JB):                         0.00
Heteroskedasticity (H):               3.25   Skew:                             0.85
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.18
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.36e+28. Standard errors may be unstable.
```

	Test RMSE
ARIMA(1, 0, 2)	1277.139529
SARIMA(1, 0, 2)(0, 0, 2)5	1441.613869

If the noise is small, as estimated by RMSE, this generally means our model is good at predicting our observed data, and if RMSE is large, this generally means our model is failing to account for important features underlying our data.



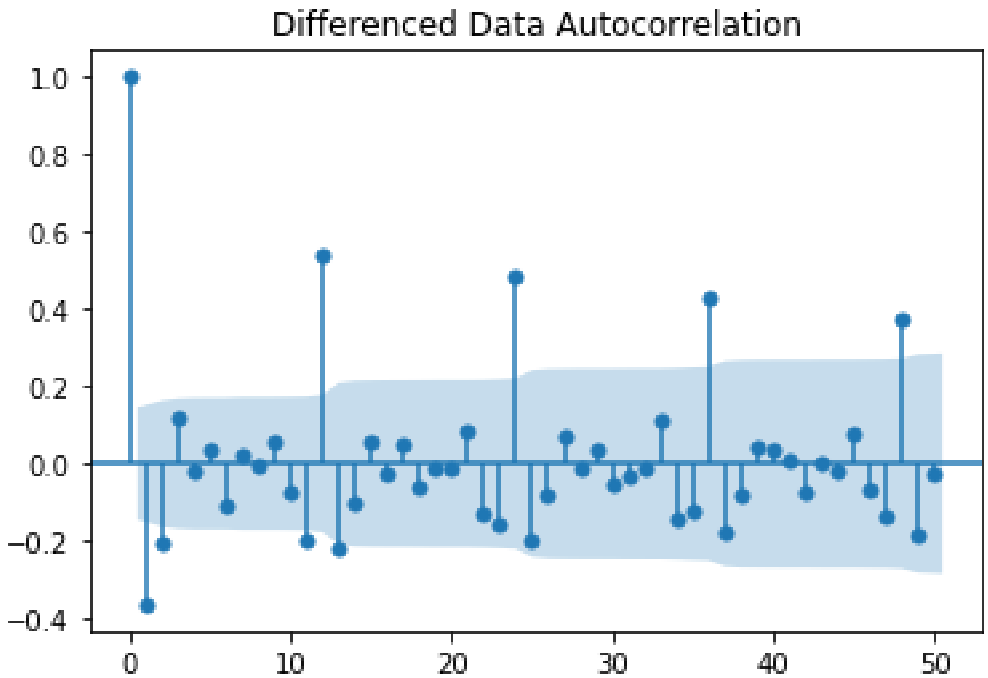


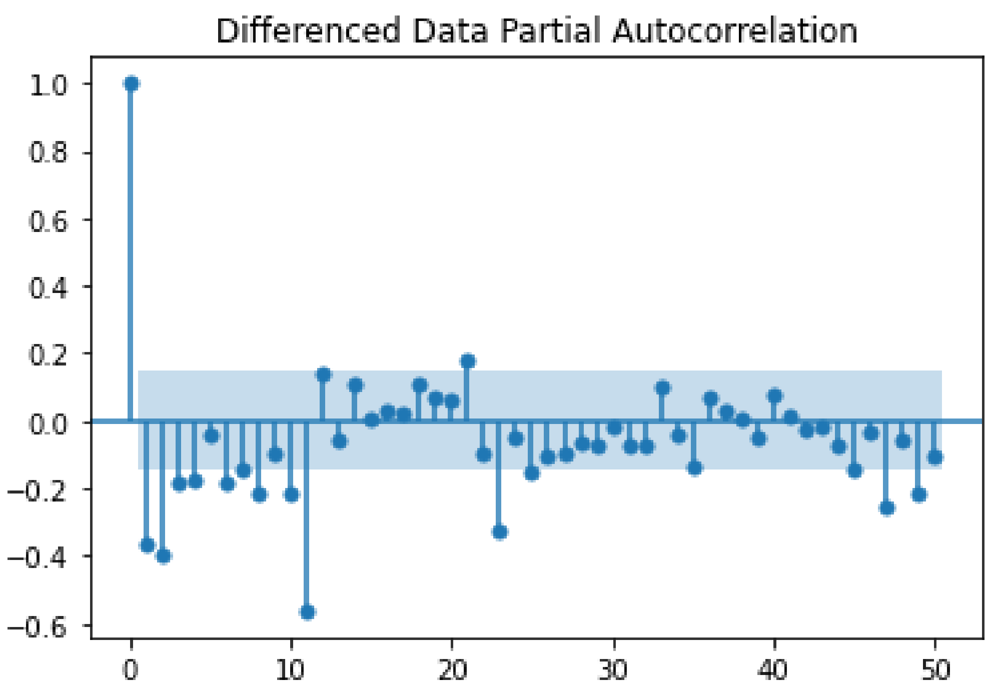
In terms of AIC, lower the AIC better the model.

2035.513< 2246.012, again SARIMA is a better model for sparkling wine data.

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.**

First Let us plot the ACF and PACF plot on the wine sales training data:





Here, we have taken alpha=0.05.

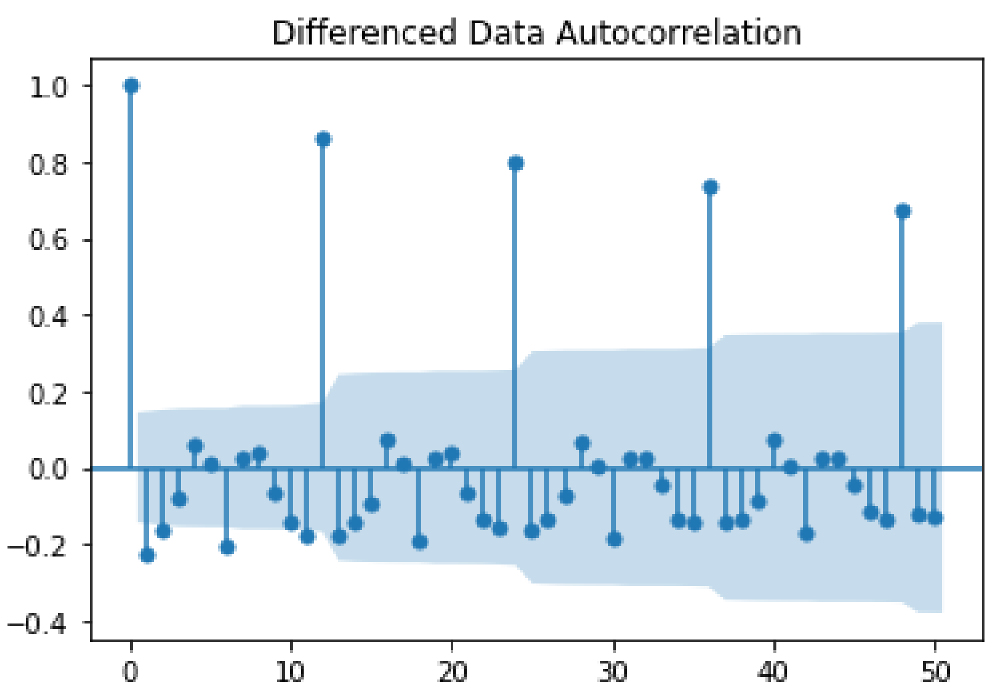
* The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.

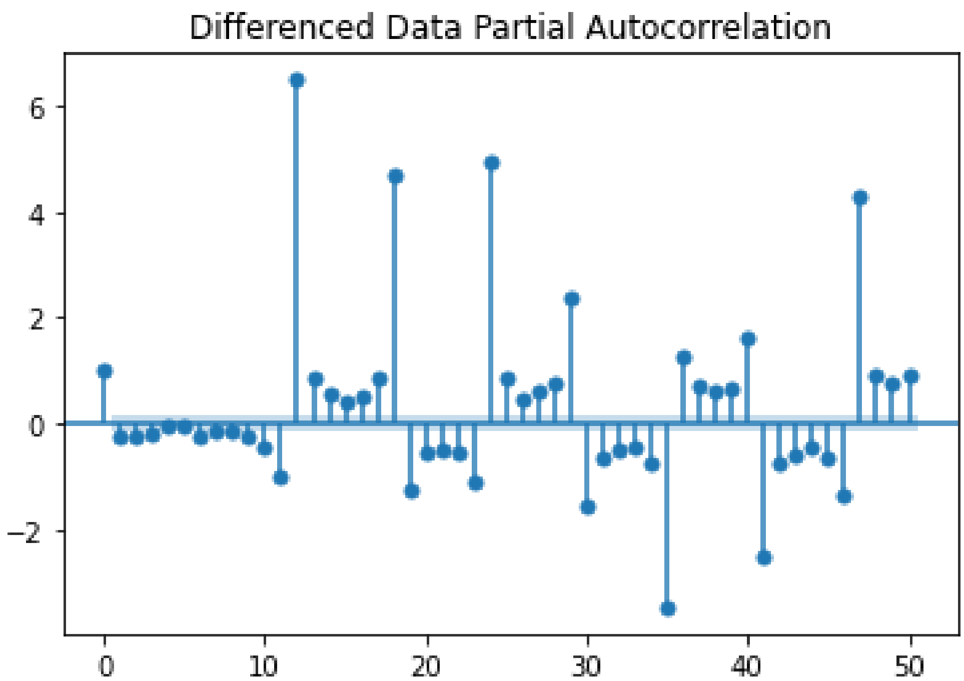
* The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.

By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -665.576
Method:                           css   S.D. of innovations             38.931
Date:                Sun, 07 Nov 2021   AIC                           1335.153
Time:                        15:56:24   BIC                           1340.903
Sample:                    02-01-1980   HQIC                          1337.489
                         - 12-01-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1527      3.401      0.045      0.964      -6.514       6.819
==============================================================================
```

With RMSE of 84.133011 and AIC of 1277.776 we have got a bad model.






```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.Sparkling   No. Observations:                  186
Model:                  ARIMA(0, 1, 0)   Log Likelihood               -1618.671
Method:                            css   S.D. of innovations           1456.212
Date:                 Sun, 07 Nov 2021   AIC                           3241.342
Time:                         18:04:02   BIC                           3247.793
Sample:                     02-01-1980   HQIC                          3243.956
                          - 07-01-1995
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.8548    106.775      0.017      0.986    -207.420     211.129
==============================================================================
```

With RMSE of 1441.613869 and AIC of 3241.342 , this is a bad model.

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

	Test RMSE	Test MAPE	RMSE
RegressionOnTime	51.433312	91.6400	NaN
NaiveModel	79.718773	145.1000	NaN
SimpleAverageModel	53.460570	94.9300	NaN
2pointTrailingMovingAverage	11.531555	13.6100	NaN
4pointTrailingMovingAverage	14.428913	19.6800	NaN
6pointTrailingMovingAverage	14.783389	21.1000	NaN
9pointTrailingMovingAverage	14.820724	20.9500	NaN
Alpha=0.995,SimpleExponentialSmoothing	36.796227	63.8800	NaN
Alpha=0.3,SimpleExponentialSmoothing	47.504821	83.7100	NaN
Alpha=0.4,SimpleExponentialSmoothing	53.767406	95.5000	NaN
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	155.814991	278.1600	NaN
TripleExponentialSmoothing	21.019620	38.7431	NaN
ARIMA(1, 0, 2)	45.450517	NaN	NaN
SARIMA(1, 0, 2)(0, 0, 2)5	20.055333	NaN	NaN
SARIMAX(1, 0, 2)(0, 0, 2)5	0.000000	NaN	NaN
ARIMA(0,1,0)	NaN	NaN	84.133011

After evaluating the RMSE , MAPE and AIC values for each model we can conclude that SARIMA model is giving us good forecast for rose wine.

Similarly for Sparkling Wine,

	RMSE	MAPE
RegressionOnTime	1275.867052	39.16
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
TripleExponentialSmoothing	469.76797	17.580255
ARIMA(1, 0, 2)	1277.139529	na
SARIMA(1, 0, 2)(0, 0, 2)5	1441.613868706947	na

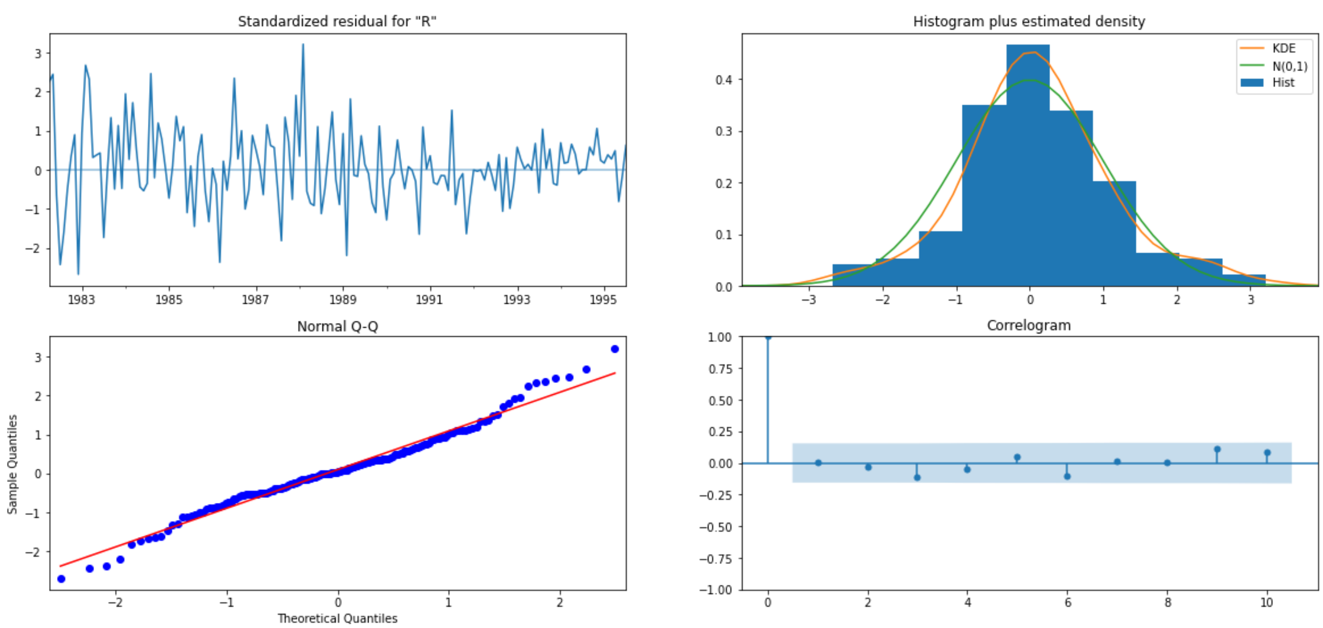
After evaluating the RMSE , MAPE and AIC values for each model we can conclude that Triple Exponential Smoothening model is giving us good forecast for Sparkling wine.

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.**

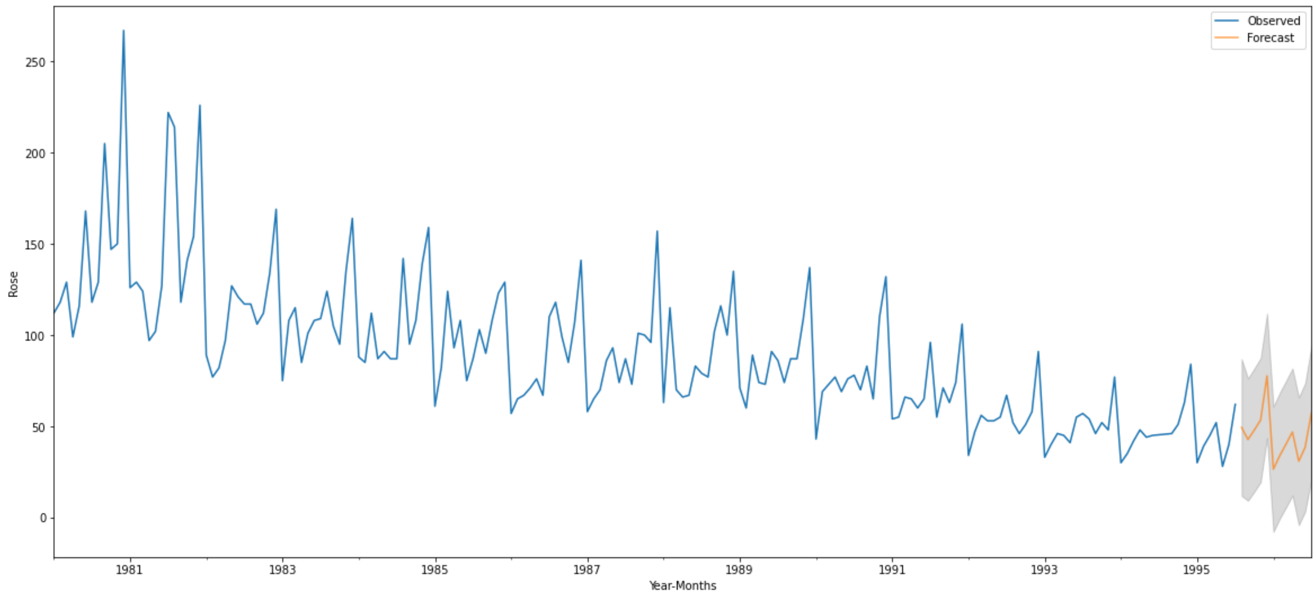
Let us now build the optimum model:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                187
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                -667.172
Date:                              Sun, 07 Nov 2021   AIC                           1344.344
Time:                                      22:11:43   BIC                           1359.720
Sample:                                  01-01-1980   HQIC                          1350.588
                                       - 07-01-1995
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1947      0.086      2.264      0.024       0.026       0.363
ma.L1         -0.9137      0.045    -20.497      0.000      -1.001      -0.826
ar.S.L12      -0.4062      0.048     -8.421      0.000      -0.501      -0.312
ma.S.L12       0.0068      0.095      0.071      0.943      -0.179       0.193
sigma2       270.3561     26.275     10.290      0.000     218.859     321.853
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 5.87
Prob(Q):                              0.91   Prob(JB):                         0.05
Heteroskedasticity (H):               0.20   Skew:                             0.19
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.86
===================================================================================

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



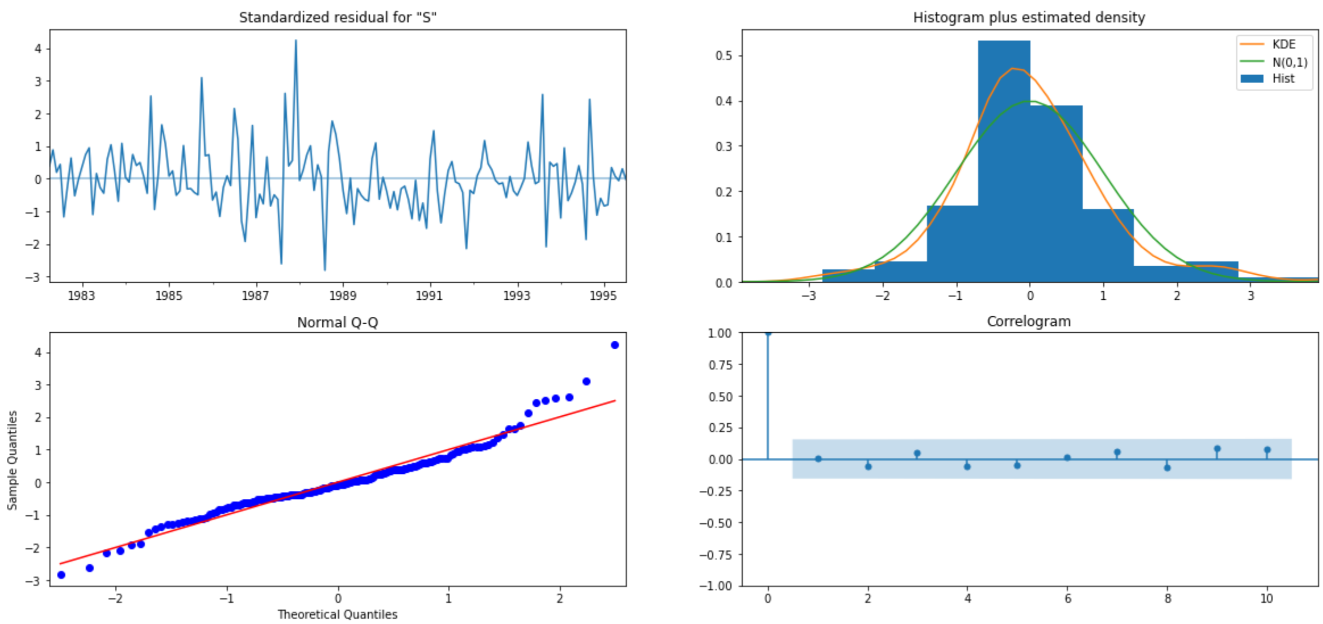
Residuals should never form a pattern; they should be random. With RMSE of 20.055333 and AIC of 1728.719 we have got a good forecast for rose wine sales.

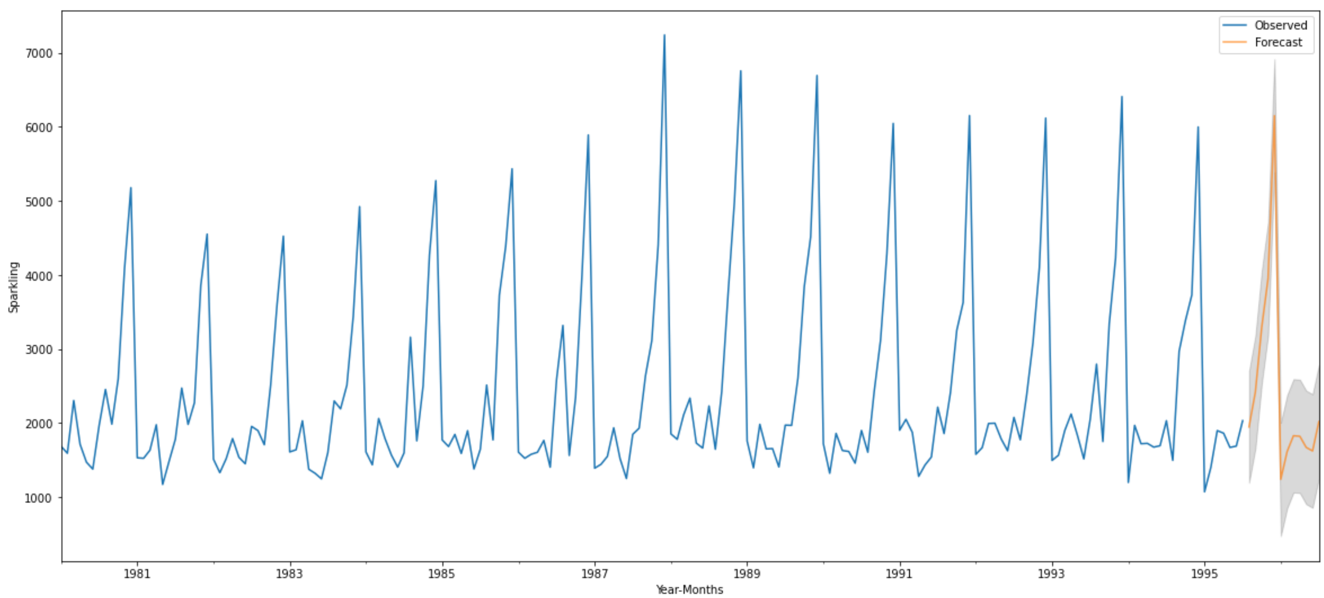


Now let us forecast for sparkling wine data:

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  187
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood               -1180.382
Date:                            Sun, 07 Nov 2021   AIC                           2370.765
Time:                                    22:12:41   BIC                           2386.141
Sample:                                01-01-1980   HQIC                          2377.009
                                     - 07-01-1995                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1151      0.082      1.398      0.162      -0.046       0.276
ma.L1         -0.9655      0.033    -29.501      0.000      -1.030      -0.901
ar.S.L12      -0.0903      0.122     -0.743      0.457      -0.329       0.148
ma.S.L12      -0.4995      0.113     -4.430      0.000      -0.720      -0.278
sigma2      1.477e+05   1.17e+04     12.575      0.000    1.25e+05    1.71e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                52.61
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               0.92   Skew:                             0.69
Prob(H) (two-sided):                  0.75   Kurtosis:                         5.44
===================================================================================

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





We have forecasted 12 months into the future using this model for sparkling Wine with 82% accuracy.

10. **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.**

- The wine market is a highly competitive market due to a large number of global and domestic companies operating in various countries. After analyzing the ABC Estate Rose and sparkling Wine sales, we can see that Rose wine sales have been dropping significantly every month, which means customers prefer other variety of wine. Highest Sales for Rose wines happen every December and lowest sales happens in January every year. Similarly, Sparkling wine sales is the highest in December and lowest in June every year. The marketing and promotion calendar needs to target these months to increase sales.

- When we look at the monthly average plot, again there is increasing trend in October, November and December every year for both rose and sparkling wine. To capitalize on this ABC estate should run targeted campaigns with special offers during this time of the year.

- We have got 80% accuracy for rose wine and 82% accuracy for sparkling wine forecast, which can be used to leverage the demand and supply and keep cost down.

- For future sales, when we look at the rose wine market Rosé remains an important wine sales driver, particularly during warm-weather months, which means ABC estate can market this during summer

- Luxury brands have started to develop accurate social media strategies to engage tech-savvy young consumers that seek greater value for money, more personalization, and integrated digital access. As the living standards are increasing globally, product dependency of these products at marriages, parties, and social gatherings is anticipated to drive the market growth in the coming years.

Thank you !