Time Series Modelling

ASSIGMENT

Submitted By

Vinit Sharma

Table of Contents

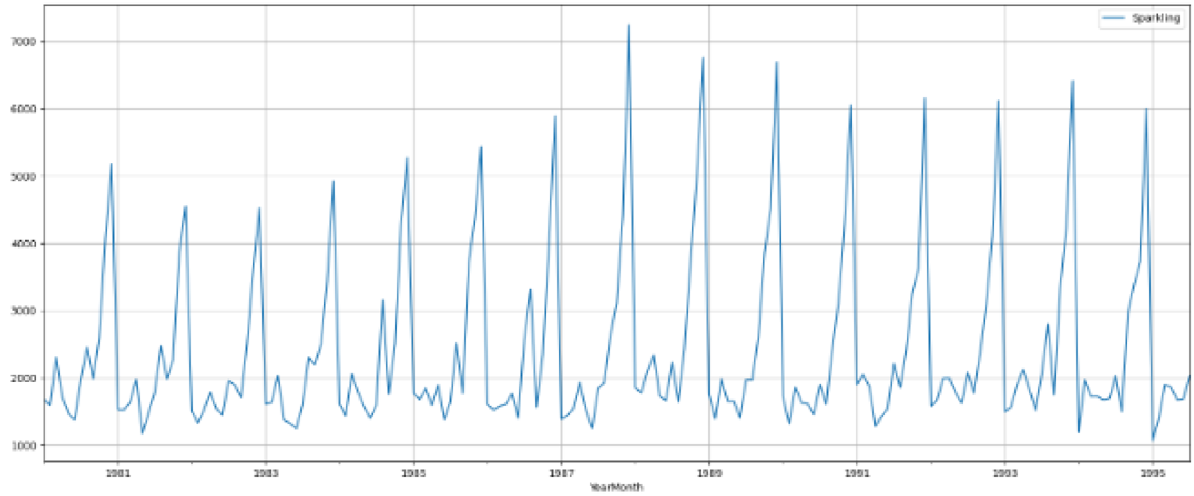
Problem 1: Sparkling Dataset

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

1. **Read the data as an appropriate Time Series data and plot the data.**
   Sparkling dataset has 187 data points with one variable as int datatype and other one is datetime datatype.

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

Below plot shows the Time Series plot which include trends and seasonality in the plot. Plot is created in a range of 1981 to 1995.



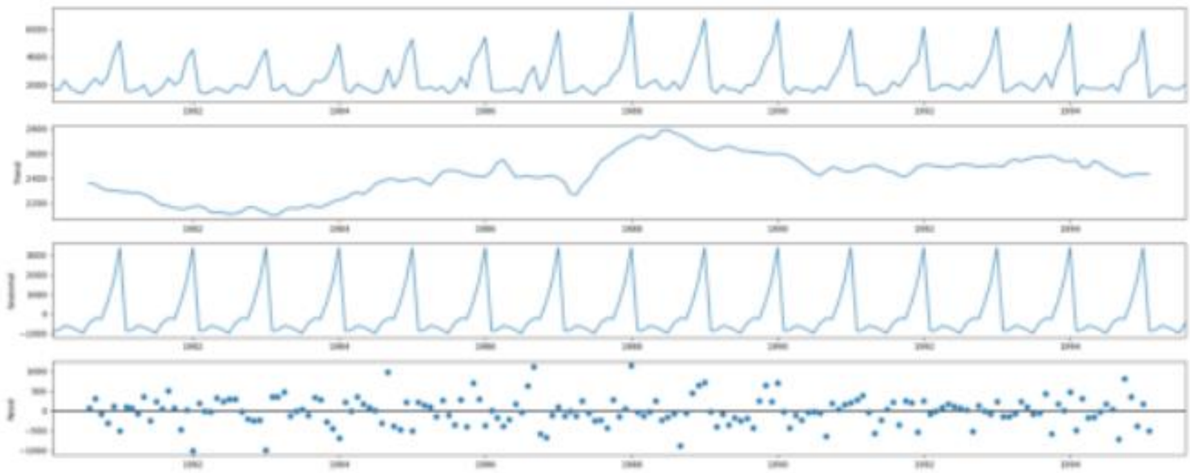
2. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Dataset description is shown in below table. Table includes mean, std, percentile information.

	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

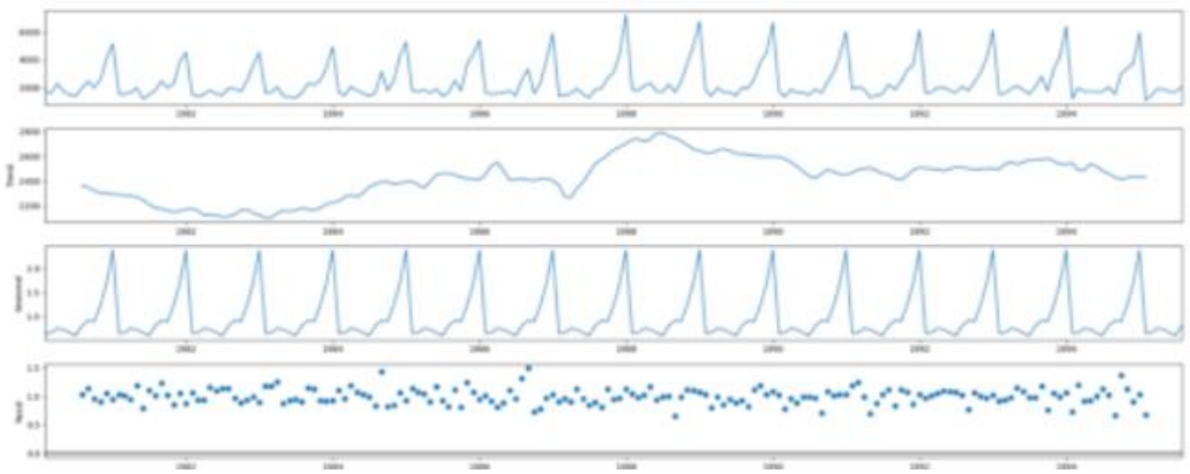
Additive Decomposition:

We have tried to decompose the dataset with respect to additive and multiplicative method. In a first approach, we can look for the error plot segment which signifies that error spread is more. So this approach is not suitable for the further process.



Multiplicative Decomposition:

In this approach, we can look for the error plot segment which signifies that error spread is less. So this approach is suitable for the further process. Error plot is almost flat and seasonality plot shows some sign of repeatability.



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

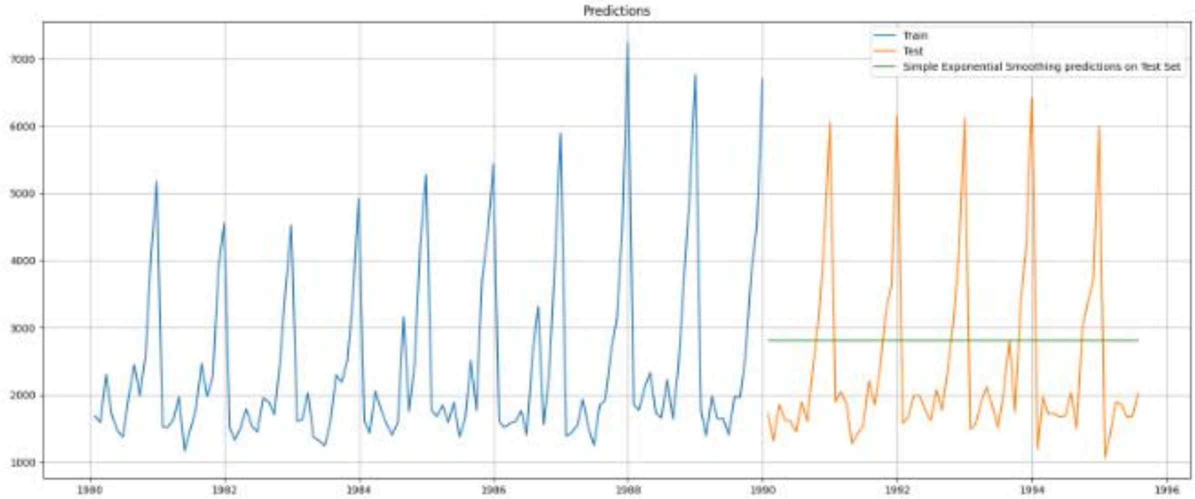
```
train = df[df.index<='1990']
test = df[df.index>'1990']
```

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.**

   All the exponential smoothing models have been built on the training dataset and their necessary performance have been evaluated on the test dataset.

   Simple Exponential Smoothing:
   Based on the process, we have used the smoothing level parameter. This test doesnot include the trend and seasonality parameter in it. Respective plot for the time series also plotted with SES line.

```
{'smoothing_level': 0.04847975339291667,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 2152.0542614313003,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



Double Exponential Smoothing:

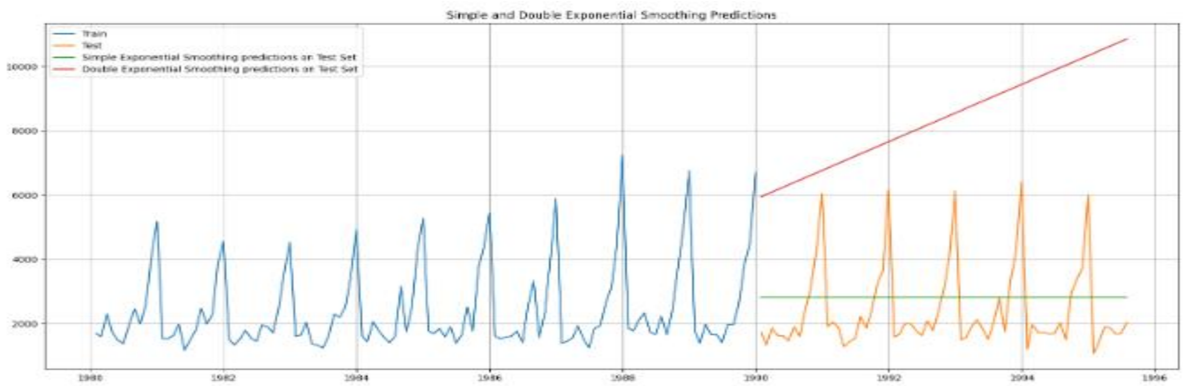
Based on the process, we have used the smoothing level and smoothing trend parameter. This test does not include the seasonality parameter in it. Respective plot for the time series also plotted with DES line.

```
# Initializing the Double Exponential Smoothing Model
model_DES = Holt(train,initialization_method='estimated')
# Fitting the model
model_DES = model_DES.fit()

print('')
print('==Holt model Exponential Smoothing Estimated Parameters ==')
print('')
print(model_DES.params)
```

```
==Holt model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.6649999999999999, 'smoothing_trend': 0.0001, 'smoothing_seasonal': nan, 'damping_trend': nan, 'initial_le
vel': 1502.1999999999998, 'initial_trend': 74.87272727272733, 'initial_seasons': array([], dtype=float64), 'use_boxcox': False,
'lamda': None, 'remove_bias': False}
```



Triple Exponential Smoothing: (A, A, A)

Based on the process, we have used the smoothing level, smoothing trend and seasonality parameter. This test used additive decomposition approach for the dataset. Respective plot for the time series also plotted with TES line.

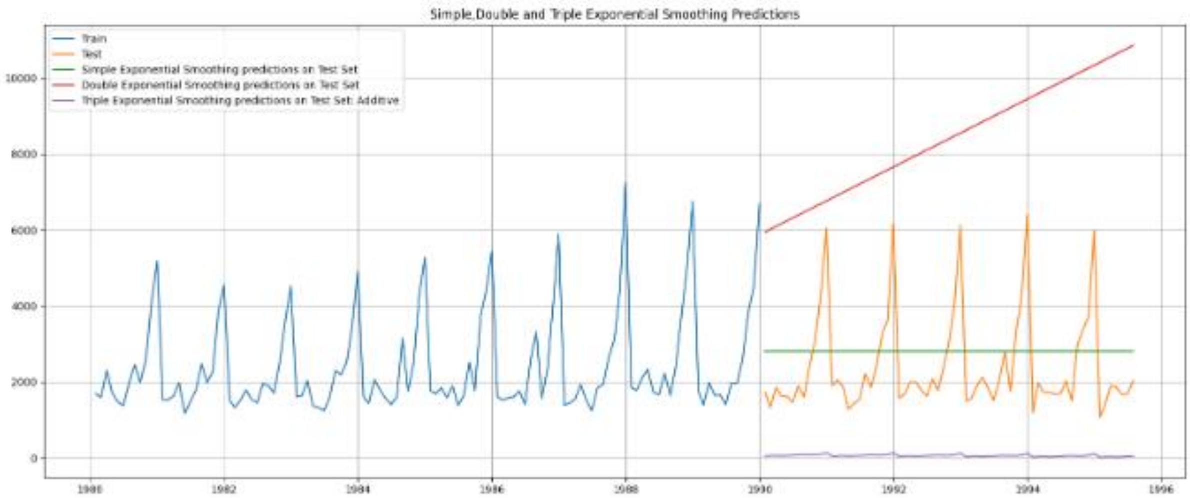
```
# Initializing the Double Exponential Smoothing Model
model_TES = ExponentialSmoothing(train,trend='additive',seasonal='additive',initialization_method='estimated')
# Fitting the model
model_TES = model_TES.fit()

print('')
print('==Holt Winters model Exponential Smoothing Estimated Parameters ==')
print('')
print(model_TES.params)
```

```
C:\Python\Anaconda\lib\site-packages\statsmodels\tsa\base\tsa_model.py:539: ValueWarning: No frequency information was provide
d, so inferred frequency M will be used.
  % freq, ValueWarning)
```

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.0759640279371264, 'smoothing_trend': 0.04336101054036127, 'smoothing_seasonal': 0.47864368464705426, 'dam
ping_trend': nan, 'initial_level': 2356.512698405284, 'initial_trend': -2.15237363936188, 'initial_seasons': array([-636.371339
4 , -723.0906225 , -398.39409242, -473.5382789 ,
       -808.59758782, -815.51336094, -384.2643042 ,   73.12842207,
       -237.65218686,  272.25793688, 1541.69781336, 2590.31364122]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```



Triple Exponential Smoothing: (A, A, M)

Based on the process, we have used the smoothing level, smoothing trend and seasonality parameter. This test used multiplicative decomposition approach for the dataset. Respective plot for the time series also plotted with TES line.

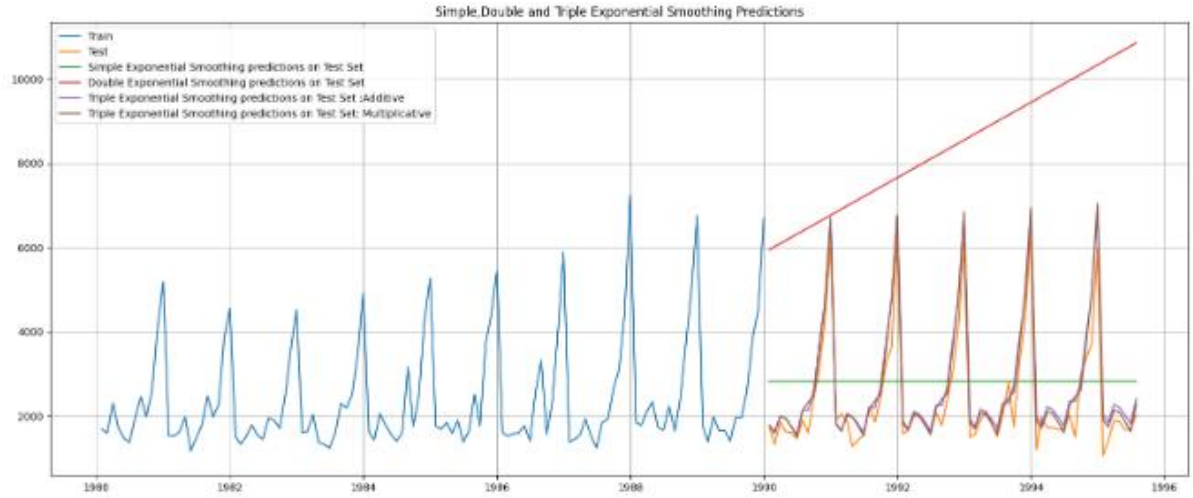
```
# Initializing the Double Exponential Smoothing Model
model_TES_am = ExponentialSmoothing(train,trend='add',seasonal='multiplicative',initialization_method='estimated')
# Fitting the model
model_TES_am = model_TES_am.fit()

print('')
print('==Holt Winters model Exponential Smoothing Estimated Parameters ==')
print('')
print(model_TES_am.params)
```

```
C:\Python\Anaconda\lib\site-packages\statsmodels\tsa\base\tsa_model.py:539: ValueWarning: No frequency information was provide
d, so inferred frequency M will be used.
  % freq, ValueWarning)
```

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.07569902291934616, 'smoothing_trend': 0.06488708837787201, 'smoothing_seasonal': 0.30803660340080713, 'da
mping_trend': nan, 'initial_level': 2356.499891168232, 'initial_trend': -13.236336129547544, 'initial_seasons': array([0.716293
28, 0.68549666, 0.8975794 , 0.80530055, 0.65656175,
       0.64725986, 0.87806682, 1.13921336, 0.91790558, 1.22969087,
       1.90962736, 2.44918653]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```



Below table shows the performance of all the smoothing model. We have seen that TES with multiplicative decomposition approach has minimum error at place.

	Test RMSE
SES	1355.557634
DES	6290.069982
TES	473.871025
TES:Multiplicative	455.360502

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.

Check for stationarity of the whole Time Series data.

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

The hypothesis in a simple form for the ADF test is:

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

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

Check for stationarity of the Training Data Time Series.

Let us plot the training data once.

```
train.plot(grid=True);
```

```
dftest = adfuller(train,regression='ct')
print('DF test statistic is %3.3f' %dftest[0])
print('DF test p-value is' ,dftest[1])
print('Number of lags used' ,dftest[2])
```

```
DF test statistic is -2.710
DF test p-value is 0.23197916198156393
Number of lags used 12
```

The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationarize the Time Series.

```
dftest = adfuller(train.diff().dropna(),regression='ct')
print('DF test statistic is %3.3f' %dftest[0])
print('DF test p-value is' ,dftest[1])
print('Number of lags used' ,dftest[2])
```

```
DF test statistic is -7.794
DF test p-value is 2.1438260690703056e-10
Number of lags used 11
```

Based on p-value, we can now reject the null hypothesis.

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 MODEL:

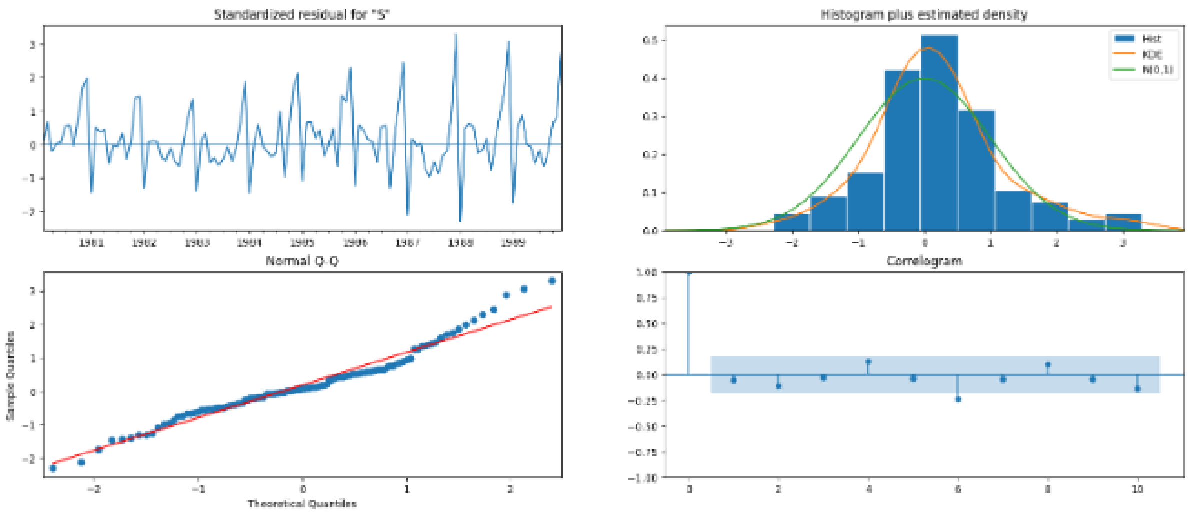
Based on analysis, we have figured out parameter (2,1,2) shows lowest AIC value based on which further analysis needs to be done.

	param	AIC
10	(2, 1, 2)	2012.666187
15	(3, 1, 3)	2016.372299
11	(2, 1, 3)	2025.152612
14	(3, 1, 2)	2025.729053
9	(2, 1, 1)	2027.872565

Here, we have test statistics on model which shows all the AR and MR are significant for the process.

```
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  120
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1001.333
Date:                Wed, 25 May 2022   AIC                           2012.666
Time:                        22:00:30   BIC                           2026.562
Sample:                    01-31-1980   HQIC                          2018.309
                         - 12-31-1989
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3149      0.050     26.110      0.000       1.216       1.414
ar.L2         -0.5712      0.089     -6.405      0.000      -0.746      -0.396
ma.L1         -1.9391      0.054    -35.639      0.000      -2.046      -1.832
ma.L2          0.9487      0.054     17.448      0.000       0.842       1.055
sigma2       1.16e+06   3.91e-08   2.97e+13      0.000    1.16e+06    1.16e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                13.31
Prob(Q):                              0.58   Prob(JB):                         0.00
Heteroskedasticity (H):               2.63   Skew:                             0.58
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.16
===================================================================================
```

Significant diagnostics have been performed for the model performance.



SARIMA MODEL:

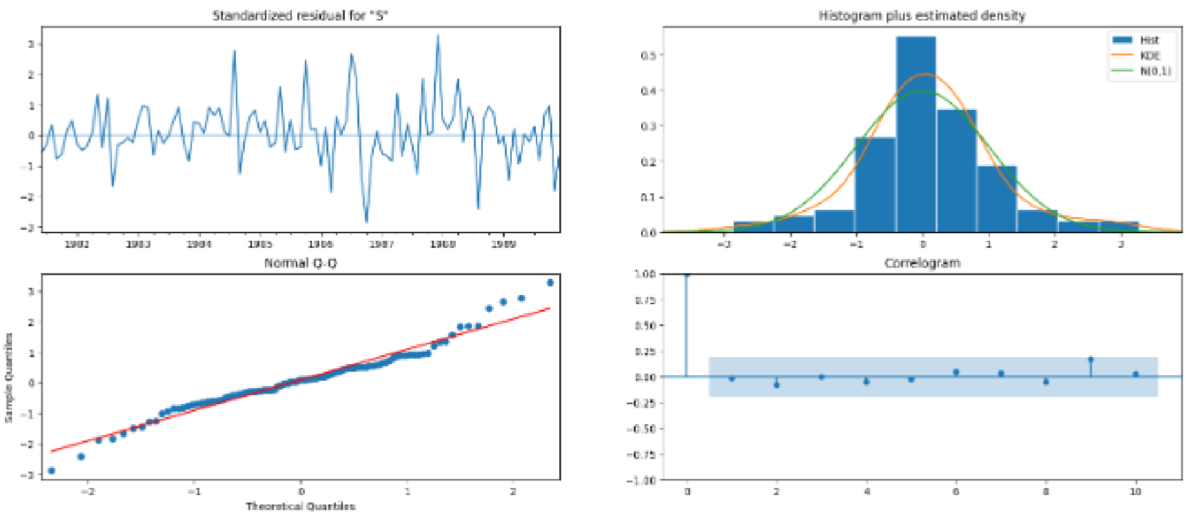
Based on analysis, we have figured out parameter (2,1,2)(3,0,3,4) shows lowest AIC value based on which further analysis needs to be done. This model includes the seasonality factor in the analysis.

	param	seasonal	AIC
255	(3, 1, 3)	(3, 0, 3, 4)	1532.703232
63	(0, 1, 3)	(3, 0, 3, 4)	1534.726268
127	(1, 1, 3)	(3, 0, 3, 4)	1535.121149
191	(2, 1, 3)	(3, 0, 3, 4)	1536.910520
251	(3, 1, 3)	(2, 0, 3, 4)	1541.039257

Here, we have test statistics on model which shows all the AR and MR are significant for the process. Some AR and MR values are more than 0.05 which could show less significance in the analysis.

```
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  120
Model:             SARIMAX(3, 1, 3)x(3, 0, 3, 4)   Log Likelihood                -753.352
Date:                          Wed, 25 May 2022   AIC                           1532.703
Time:                                  22:07:30   BIC                           1566.955
Sample:                              01-31-1980   HQIC                          1546.576
                                   - 12-31-1989
Covariance Type:                            opg
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
ar.L1         -1.7540      0.119    -14.771      0.000      -1.987      -1.521
ar.L2         -0.9075      0.200     -4.543      0.000      -1.299      -0.516
ar.L3         -0.0180      0.109     -0.165      0.869      -0.232       0.196
ma.L1          0.8115      0.707      1.147      0.251      -0.575       2.198
ma.L2         -1.0510      0.127     -8.264      0.000      -1.300      -0.802
ma.L3         -1.0823      0.769     -1.407      0.159      -2.590       0.425
ar.S.L4       -0.0057      0.015     -0.392      0.695      -0.034       0.023
ar.S.L8       -0.0276      0.015     -1.875      0.061      -0.056       0.001
ar.S.L12       1.0644      0.015     72.954      0.000       1.036       1.093
ma.S.L4        0.0405      0.689      0.059      0.953      -1.309       1.390
ma.S.L8        0.0692      0.715      0.097      0.923      -1.332       1.471
ma.S.L12      -1.0863      0.736     -1.475      0.140      -2.530       0.357
sigma2      7.447e+04   1.99e-05   3.75e+09      0.000    7.45e+04    7.45e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 9.94
Prob(Q):                              0.88   Prob(JB):                         0.01
Heteroskedasticity (H):               2.88   Skew:                             0.27
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.42
===================================================================================
```

Significant diagnostics have been performed for the model performance.



Based on analysis of Both ARIMA and SARIMA, we have evaluated the performance which shows SARIMA is performing better than ARIMA model on lower AIC value.

	RMSE	MAPE
ARIMA(2,1,2)	1339.707781	53.172870
SARIMA(3,1,3)(3,0,3,4)	852.835727	21.466912

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.**

ARIMA MODEL:

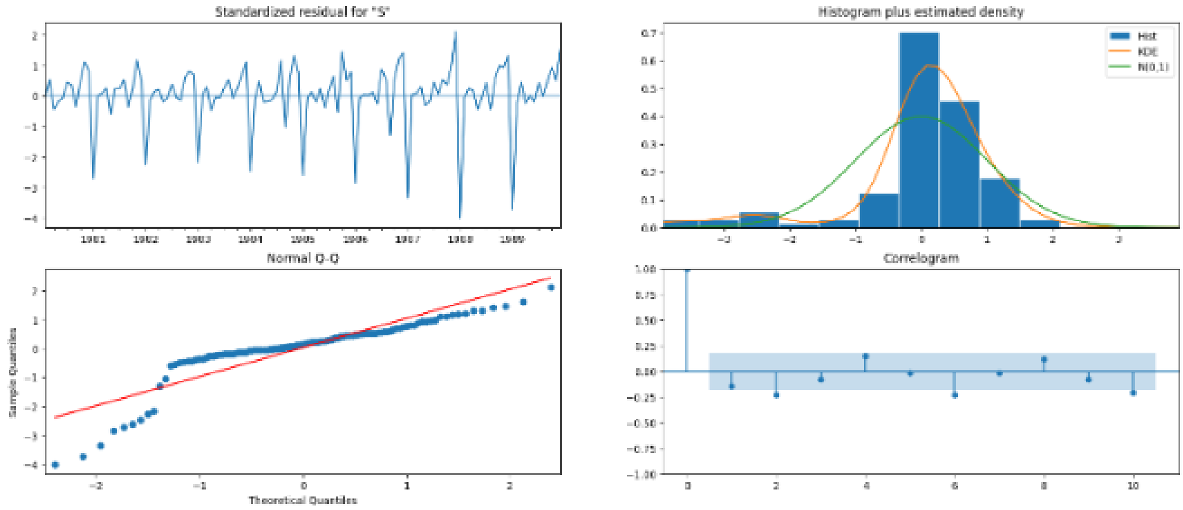
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.

By looking at the above plots, we will take the value of p and q to be 0 and 0 respectively.

Here, we have test statistics on model which shows all the AR and MR are significant for the process.

```
Dep. Variable:              Sparkling   No. Observations:                  120
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1026.816
Date:                Wed, 25 May 2022   AIC                           2055.631
Time:                        22:10:34   BIC                           2058.411
Sample:                    01-31-1980   HQIC                          2056.760
                         - 12-31-1989
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.814e+06   1.32e+05     13.795      0.000    1.56e+06    2.07e+06
===================================================================================
Ljung-Box (L1) (Q):                   2.56   Jarque-Bera (JB):               170.25
Prob(Q):                              0.11   Prob(JB):                         0.00
Heteroskedasticity (H):               2.48   Skew:                            -1.06
Prob(H) (two-sided):                  0.01   Kurtosis:                         7.53
===================================================================================
```

Significant diagnostics have been performed for the model performance.



Below table shows the performance of the SARIMA model on certain parameters value.

	RMSE	MAPE
ARIMA(0,1,0)	4482.058965	234.266414

SARIMA MODEL:

Here, we have taken alpha=0.05.

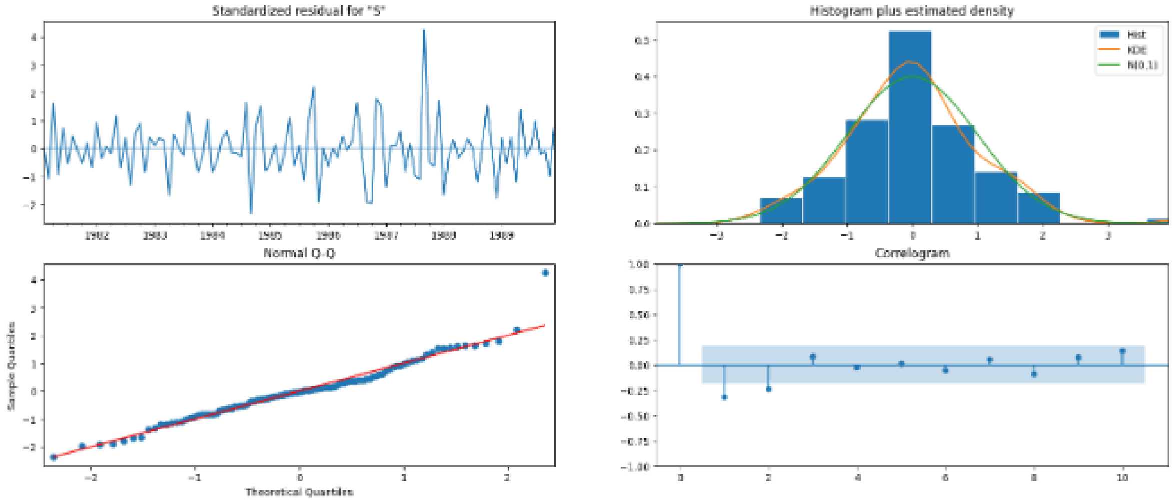
We are going to take the seasonal period as 4 or its multiple e.g. 8. We are taking the p value to be 0 and the q value also to be 0 as the parameters same as the ARIMA model.

- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2.
- The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 1.

Here, we have test statistics on model which shows all the AR and MR are significant for the process. Some AR and MR values are more than 0.05 which could show less significance in the analysis.

```
==========================================================================================
Dep. Variable:                         Sparkling   No. Observations:                  120
Model:             SARIMAX(0, 1, 0)x(2, 1, [1], 4)   Log Likelihood                -829.058
Date:                           Wed, 25 May 2022   AIC                           1666.117
Time:                                   22:18:07   BIC                           1676.808
Sample:                               01-31-1980   HQIC                          1670.451
                                    - 12-31-1989
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L4       -0.9655      0.036    -26.958      0.000      -1.036      -0.895
ar.S.L8       -0.9722      0.041    -23.917      0.000      -1.052      -0.893
ma.S.L4       -0.0473      0.103     -0.461      0.645      -0.248       0.154
sigma2      3.135e+05   3.54e+04      8.862      0.000    2.44e+05    3.83e+05
===================================================================================
Ljung-Box (L1) (Q):                  11.11   Jarque-Bera (JB):                24.96
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               2.09   Skew:                             0.62
Prob(H) (two-sided):                  0.03   Kurtosis:                         5.02
===================================================================================
```

Significant diagnostics have been performed for the model performance.



Below table shows the performance of the SARIMA model on certain parameters value.

	RMSE	MAPE
SARIMA(0,1,0)(2,1,1,4)	556.543882	23.245745

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

Performance of all the built models with their corresponding parameters and RMSE value has been captured in the table for the test data. Based on this table, we can suggest SARIMA model based on Lowest AIC perform much better than any other built model.

	RMSE	MAPE
ARIMA(2,1,2)	1339.707781	53.172870
SARIMA(3,1,3)(3,0,3,4)	852.835727	21.466912
ARIMA(0,1,0)	4482.058965	234.266414
SARIMA(0,1,0)(2,1,1,4)	556.543882	23.245745

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.**

SARIMA model with lowest AIC parameter is the most optimum model for complete dataset and it is used to predict 12 months into the future with appropriate CI.

```
full_data_model = sm.tsa.statespace.SARIMAX(df['Sparkling'],
                                order=(3,1,3),
                                seasonal_order=(3, 0, 3, 4),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results_full_data_model = full_data_model.fit(maxiter=1000)
print(results_full_data_model.summary())
```

Evaluate the model on the whole data and predict 12 months into the future (till the end of next year).

```
predicted_manual_SARIMA_full_data = results_full_data_model.get_forecast(steps=12)
```

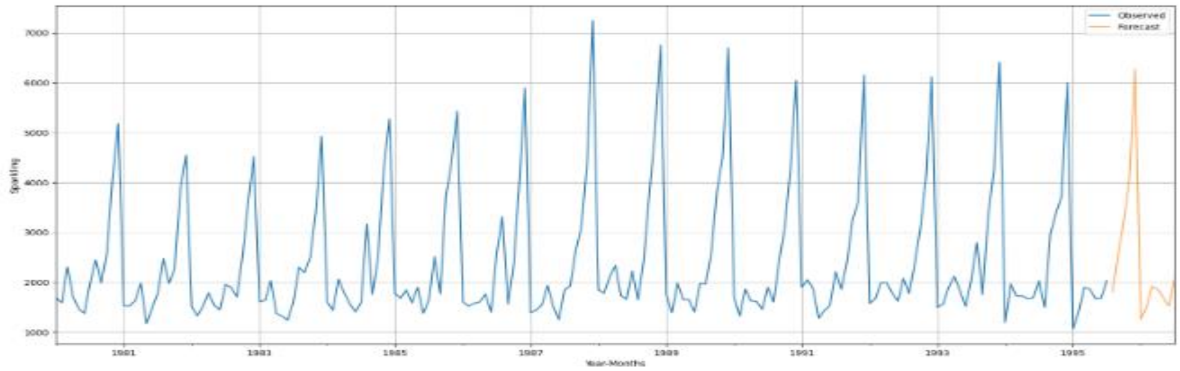
```
pred_full_manual_SARIMA_date = predicted_manual_SARIMA_full_data.summary_frame(alpha=0.05)
pred_full_manual_SARIMA_date.head()
```

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1807.580567	361.539969	1098.975249	2516.185885
1995-09-30	2547.640258	366.780928	1828.762849	3266.517667
1995-10-31	3227.953700	366.876542	2508.888891	3947.018509
1995-11-30	4064.965835	368.206420	3343.294513	4786.637158
1995-12-31	6274.879647	368.425000	5552.779916	6996.979378

```
rmse = mean_squared_error(df['Sparkling'],results_full_data_model.fittedvalues,squared=False)
print('RMSE of the Full Model',rmse)
```

```
RMSE of the Full Model 525.810355093653
```

This plot signifies about the behaviour of the future predict months.



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

- In all the built smoothing model, TES with multiplicative decomposition approach has minimum error at place
- We can suggest that SARIMA model based on Lowest AIC perform much better than any other time series-built model.
- Based on the result, companies wine sales in Sparkling segment takes a heavy spike in a certain quarter of the year. If companies supply chain should be strong to support that demand and by providing some offers (wedding special and buy one get one kind), company can boost its sales revenue.

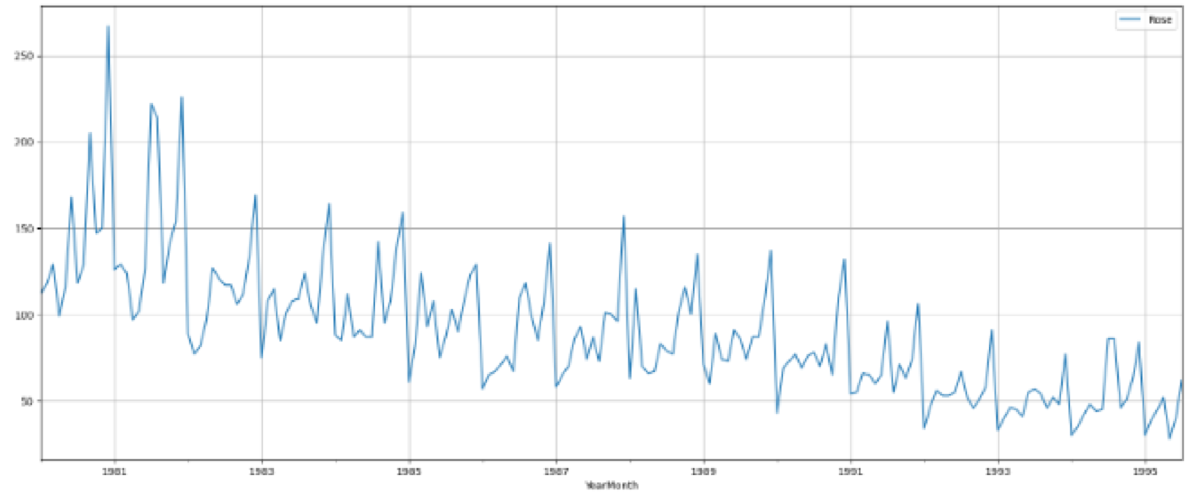
Problem 2: Rose Dataset

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

1. **Read the data as an appropriate Time Series data and plot the data.**
   Rose dataset has 187 data points with one variable as float datatype and other one is datetime datatype.

	YearMonth	Rose
0	1980-01-31	112.0
1	1980-02-29	118.0
2	1980-03-31	129.0
3	1980-04-30	99.0
4	1980-05-31	116.0

Below plot shows the Time Series plot which include trends and seasonality in the plot. Plot is created in a range of 1981 to 1995.

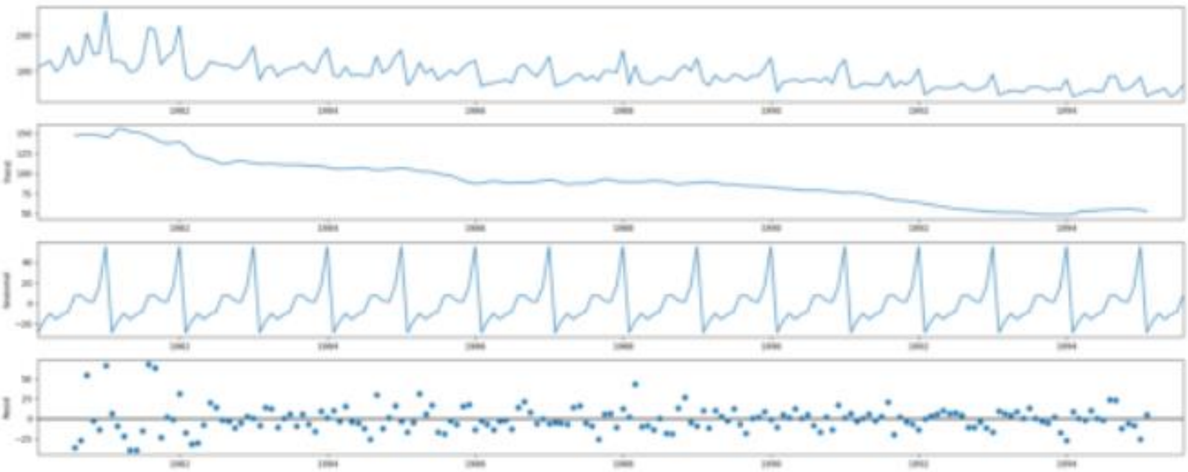


2. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

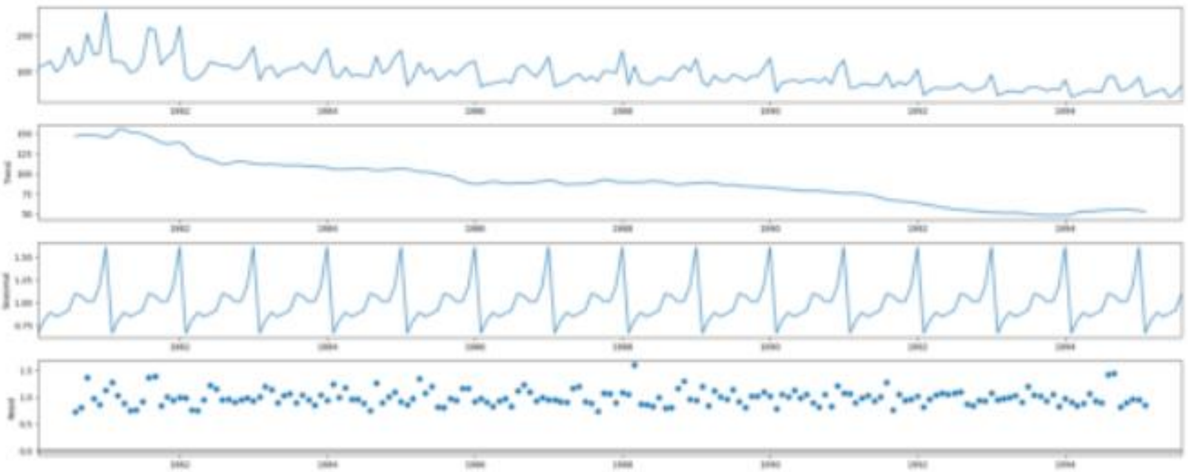
Dataset description is shown in below table. Table includes mean, std, percentile information.

	Rose
count	187.000000
mean	90.347594
std	38.966791
min	28.000000
25%	63.000000
50%	86.000000
75%	111.000000
max	267.000000

Additive Decomposition:
We have tried to decompose the dataset with respect to additive and multiplicative method. In a first approach, we can look for the error plot segment which signifies that error spread is more. So this approach is not suitable for the further process. In the plot, we can clearly sees that we have trend in the dataset.



Multiplicative Decomposition:
In this approach, we can look for the error plot segment which signifies that error spread is less. So this approach is suitable for the further process. Error plot is almost flat and seasonality plot shows some sign of repeatability. We also have a clear trend in the plot.



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

```
train = df2[df2.index<='1990']
test = df2[df2.index>'1990']
```

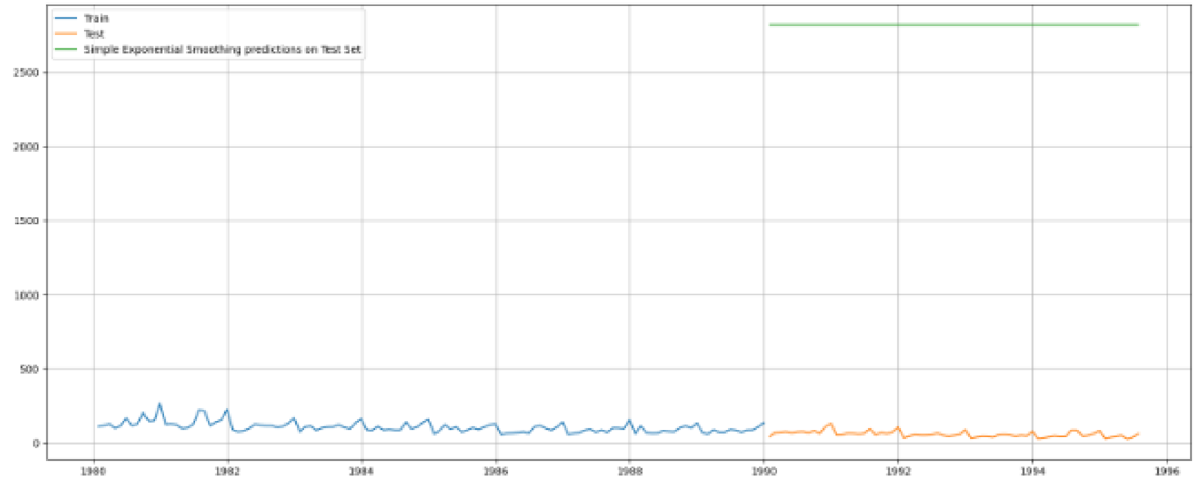
4. **Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance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

All the exponential smoothing models have been built on the training dataset and their necessary performance have been evaluated on the test dataset.

Simple Exponential Smoothing:

Based on the process, we have used the smoothing level parameter. This test doesnot include the trend and seasonality parameter in it. Respective plot for the time series also plotted with SES line.

{'smoothing_level': 0.04847975339291667,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 2152.054261431303,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}

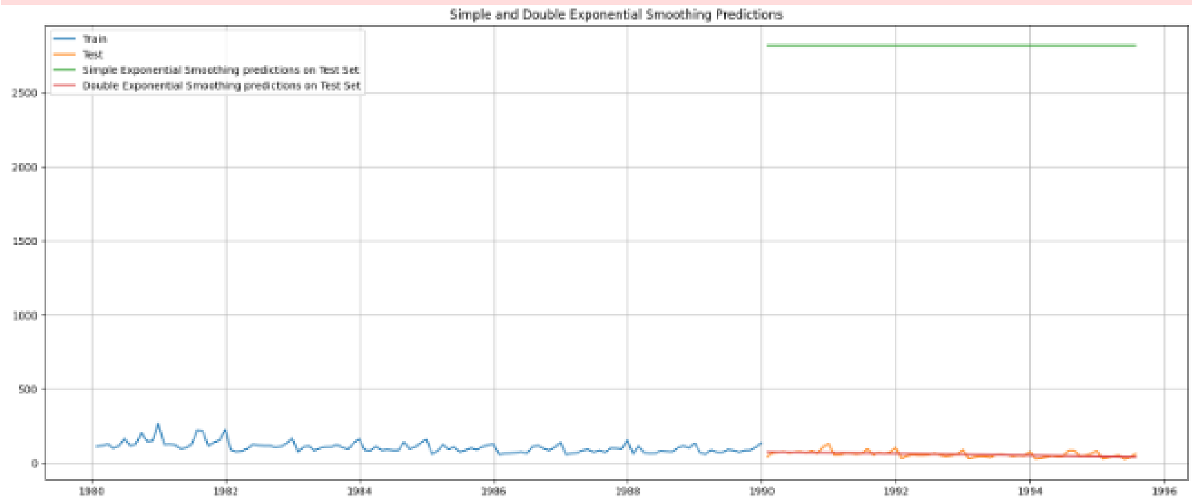


Double Exponential Smoothing:

Based on the process, we have used the smoothing level and smoothing trend parameter. This test does not include the seasonality parameter in it. Respective plot for the time series also plotted with DES line.

```
==Holt model Exponential Smoothing Estimated Parameters ==
```

{'smoothing_level': 1.4901161193847656e-08, 'smoothing_trend': 6.680818251431411e-09, 'smoothing_seasonal': nan, 'damping_tren
d': nan, 'initial_level': 138.93243921985837, 'initial_trend': -0.5185804177349018, 'initial_seasons': array([], dtype=float6
4), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

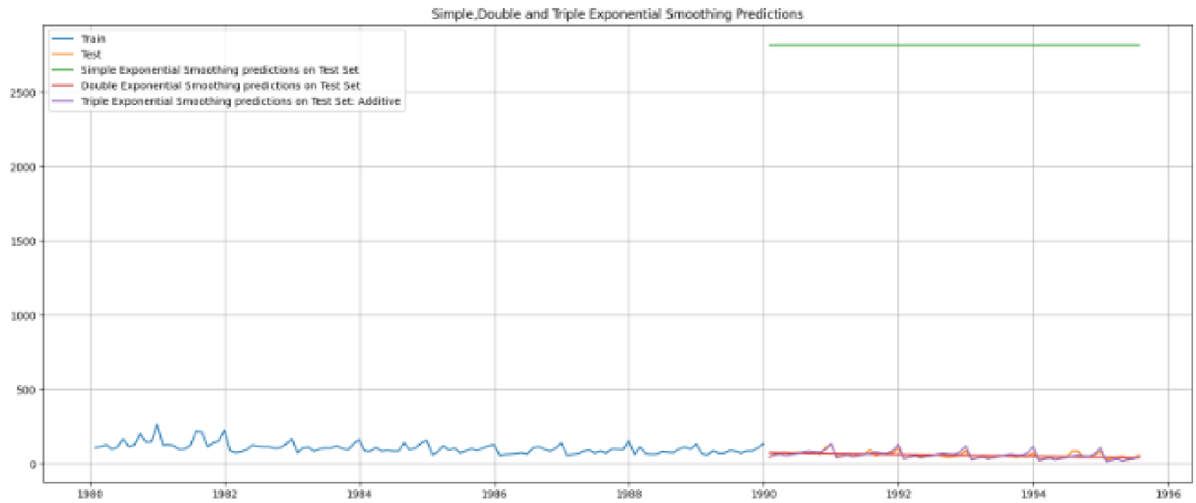


Triple Exponential Smoothing: (A, A, A)

Based on the process, we have used the smoothing level, smoothing trend and seasonality parameter. This test used additive decomposition approach for the dataset. Respective plot for the time series also plotted with TES line.

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.09497210888874562, 'smoothing_trend': 7.465095965273274e-05, 'smoothing_seasonal': 0.0003521866705400393,
'damping_trend': nan, 'initial_level': 146.58611350466705, 'initial_trend': -0.5551780400208214, 'initial_seasons': array([-30.
49580985, -19.51332275, -11.15657296, -23.27538705,
       -12.84253406,  -7.64478742,   3.10608786,  10.50245295,
         4.77490993,   4.38565829,  19.65816521,  63.91978977]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

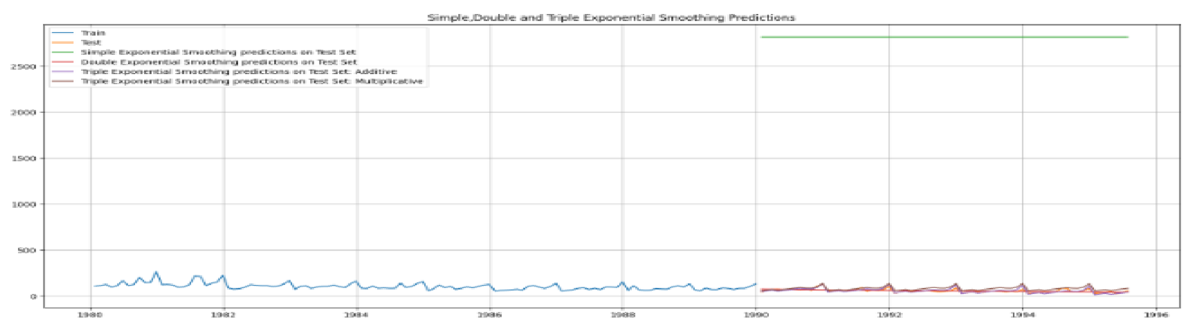


Triple Exponential Smoothing: (A, A, M)

Based on the process, we have used the smoothing level, smoothing trend and seasonality parameter. This test used multiplicative decomposition approach for the dataset. Respective plot for the time series also plotted with TES line.

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.061212896166004546, 'smoothing_trend': 0.06121289609390582, 'smoothing_seasonal': 1.9472419683677065e-07,
'damping_trend': nan, 'initial_level': 132.95858964065945, 'initial_trend': -0.826106263183913, 'initial_seasons': array([0.862
90308, 0.96224949, 1.05305424, 0.90797979, 1.03682312,
       1.12969751, 1.24700591, 1.33682102, 1.25441347, 1.24019411,
       1.41573663, 1.97074129]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```



Below table shows the performance of all the smoothing model. We have seen that TES with additive decomposition approach has minimum error at place.

	Test RMSE
SES	2755.358150
DES	17.876669
TES:Additive	14.694018
TES: Multiplicative	28.363216

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.

Check for stationarity of the whole Time Series data.

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

The hypothesis in a simple form for the ADF test is:

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

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

```
dftest = adfuller(train_1,regression='ct')
print('DF test statistic is %3.3f' %dftest[0])
print('DF test p-value is' ,dftest[1])
print('Number of lags used' ,dftest[2])
```

```
DF test statistic is -1.321
DF test p-value is 0.8827240404113933
Number of lags used 13
```

```
dftest = adfuller(train_1.diff().dropna(),regression='ct')
print('DF test statistic is %3.3f' %dftest[0])
print('DF test p-value is' ,dftest[1])
print('Number of lags used' ,dftest[2])
```

```
DF test statistic is -6.392
DF test p-value is 3.1814938121222416e-07
Number of lags used 12
```

Based on p-value, we can now reject the null hypothesis.

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 MODEL:

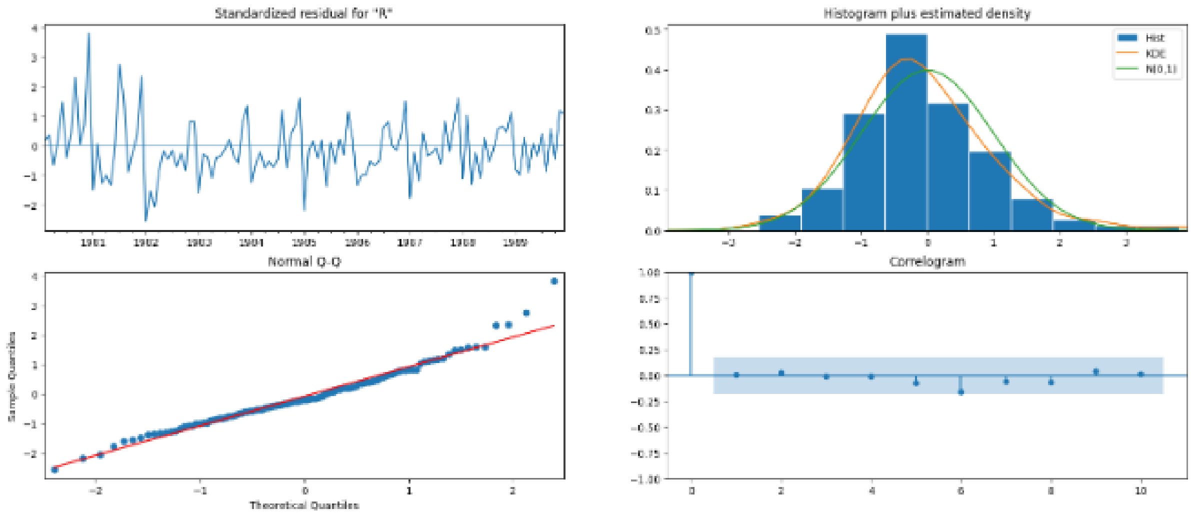
Based on analysis, we have figured out parameter (2,1,3) shows lowest AIC value based on which further analysis needs to be done.

	param	AIC
11	(2, 1, 3)	1162.387543
15	(3, 1, 3)	1166.943160
2	(0, 1, 2)	1167.498636
6	(1, 1, 2)	1167.897377
5	(1, 1, 1)	1168.359028

Here, we have test statistics on model which shows all the AR and MR are significant for the process.

```
Dep. Variable:                   Rose   No. Observations:                  120
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -575.194
Date:                Wed, 25 May 2022   AIC                           1162.388
Time:                        22:54:23   BIC                           1179.062
Sample:                    01-31-1980   HQIC                          1169.159
                         - 12-31-1989
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6499      0.089    -18.557      0.000      -1.824      -1.476
ar.L2         -0.6968      0.091     -7.636      0.000      -0.876      -0.518
ma.L1          1.0466      0.726      1.442      0.149      -0.376       2.469
ma.L2         -0.7700      0.145     -5.299      0.000      -1.055      -0.485
ma.L3         -0.9037      0.662     -1.366      0.172      -2.201       0.393
sigma2       877.3869    628.179      1.397      0.162    -353.820    2108.594
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                21.14
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               0.39   Skew:                             0.69
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.53
===================================================================================
```

Significant diagnostics have been performed for the model performance.



Below table shows the performance of the ARIMA model on certain parameters value.

	RMSE	MAPE
ARIMA(2,1,2)	38.844627	74.407596

SARIMA MODEL:

	param	seasonal	AIC
227	(3, 1, 2)	(0, 0, 3, 12)	677.210818
220	(3, 1, 1)	(3, 0, 0, 12)	680.640009
222	(3, 1, 1)	(3, 0, 2, 12)	681.625669
221	(3, 1, 1)	(3, 0, 1, 12)	681.905114
252	(3, 1, 3)	(3, 0, 0, 12)	682.245331

Here, we have test statistics on model which shows all the AR and MR are significant for the process.

```
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                  120
Model:             SARIMAX(3, 1, 2)x(0, 0, [1, 2, 3], 12)   Log Likelihood              -329.605
Date:                              Wed, 25 May 2022   AIC                            677.211
Time:                                      23:16:05   BIC                            698.649
Sample:                                  01-31-1980   HQIC                           685.806
                                       - 12-31-1989
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9516      3.361     -0.283      0.777      -7.540       5.637
ar.L2          0.6325    524.798      0.001      0.999   -1027.953    1029.218
ar.L3         -0.2495    314.044     -0.001      0.999    -615.764     615.265
ma.L1          1.0620      3.403      0.312      0.755      -5.607       7.731
ma.L2         -1.0122      0.008   -128.531      0.000      -1.028      -0.997
ma.S.L12    -9.81e+13   2.17e-08  -4.52e+21      0.000   -9.81e+13   -9.81e+13
ma.S.L24   -1.948e+13   1.26e-13  -1.55e+26      0.000   -1.95e+13   -1.95e+13
ma.S.L36    -3.18e+14   9.48e-17  -3.35e+30      0.000   -3.18e+14   -3.18e+14
sigma2      1072.0748    990.871      1.082      0.279    -869.996    3014.146
===================================================================================
Ljung-Box (L1) (Q):                   8.82   Jarque-Bera (JB):              2541.51
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.00   Skew:                            -4.50
Prob(H) (two-sided):                  0.00   Kurtosis:                        29.11
===================================================================================
```

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.**

ARIMA MODEL:

Here, we have taken alpha=0.05.

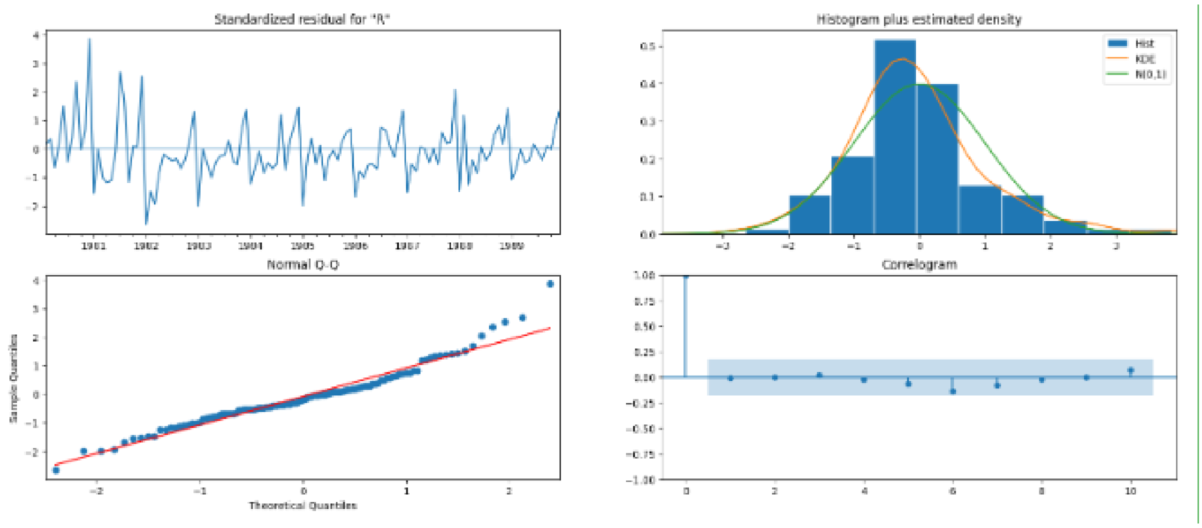
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 2.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.

By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.

Here, we have test statistics on model which shows all the AR and MR are significant for the process.

```
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  120
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -579.948
Date:                Wed, 25 May 2022   AIC                           1169.896
Time:                        22:57:16   BIC                           1183.792
Sample:                    01-31-1980   HQIC                          1175.539
                         - 12-31-1989
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4423      0.494     -0.896      0.370      -1.410       0.526
ar.L2          0.0043      0.184      0.024      0.981      -0.356       0.365
ma.L1         -0.2552      0.483     -0.528      0.597      -1.202       0.692
ma.L2         -0.5948      0.453     -1.313      0.189      -1.482       0.293
sigma2       988.1544     99.393      9.942      0.000     793.348    1182.960
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                29.97
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               0.35   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.88
===================================================================================
```

Significant diagnostics have been performed for the model performance.



Below table shows the performance of the ARIMA model on certain parameters value.

	RMSE	MAPE
ARIMA(2,1,2)	39.159403	75.173677

SARIMA MODEL:

Here, we have taken alpha=0.05.

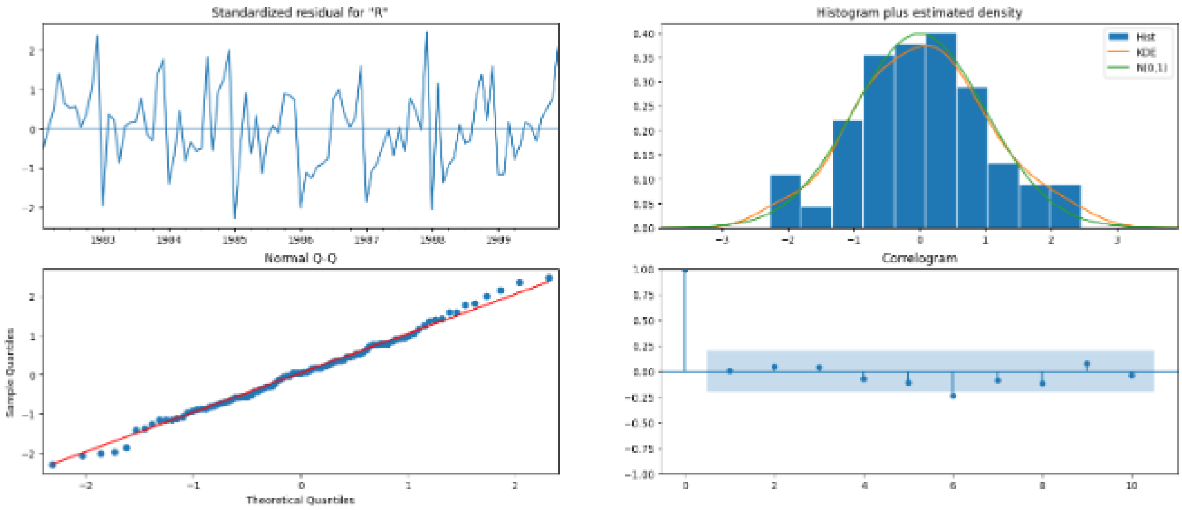
We are going to take the seasonal period as 11 or its multiple e.g. 22. We are taking the p value to be 2 and the q value also to be 2 as the parameters same as the ARIMA model.

- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2.
- The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 2.

Here, we have test statistics on model which shows all the AR and MR are significant for the process. Some AR and MR values are more than 0.05 which could show less significance in the analysis.

```
Dep. Variable:                           Rose   No. Observations:                  120
Model:             SARIMAX(2, 0, 2)x(2, 0, 2, 11)   Log Likelihood                -436.535
Date:                        Wed, 25 May 2022   AIC                            891.070
Time:                                23:24:57   BIC                            914.055
Sample:                            01-31-1980   HQIC                           900.357
                                 - 12-31-1989
Covariance Type:                          opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9567      0.275      3.482      0.000       0.418       1.495
ar.L2          0.0404      0.274      0.147      0.883      -0.497       0.578
ma.L1         -0.7668      7.610     -0.101      0.920     -15.683      14.149
ma.L2         -0.2328      1.798     -0.129      0.897      -3.757       3.291
ar.S.L11       0.0291      0.137      0.213      0.832      -0.239       0.297
ar.S.L22      -0.0054      0.158     -0.034      0.973      -0.315       0.304
ma.S.L11      -0.1231      0.186     -0.662      0.508      -0.487       0.241
ma.S.L22      -0.1789      0.214     -0.835      0.404      -0.599       0.241
sigma2       540.6313   4108.819      0.132      0.895   -7512.507    8593.769
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 0.25
Prob(Q):                              0.98   Prob(JB):                         0.88
Heteroskedasticity (H):               1.08   Skew:                             0.09
Prob(H) (two-sided):                  0.82   Kurtosis:                         2.84
===================================================================================
```

Significant diagnostics have been performed for the model performance.



Below table shows the performance of the SARIMA model on certain parameters value.

	RMSE	MAPE
SARIMA(2,0,2)(2,0,2,11)	2.600800e+01	4.803361e+01

8. **Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**
Performance of all the built models with their corresponding parameters and RMSE value has been captured in the table for the test data. Based on this table, we can suggest SARIMA model based on ACF and PACF plot perform much better than any other built model.

	RMSE	MAPE
ARIMA(2,1,3)	3.884463e+01	7.440760e+01
ARIMA(2,1,2)	3.915940e+01	7.517368e+01
SARIMA(1,1,3)(3,0,3,6)	2.060948e+33	1.285628e+33
SARIMA(2,0,2)(2,0,2,11)	2.600800e+01	4.803361e+01

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.**

SARIMA model with ACF and PACF cut off point parameter is the most optimum model for complete dataset and it is used to predict 12 months into the future with appropriate CI.

```
full_data_model = sm.tsa.statespace.SARIMAX(df2['Rose'],
                                order=(2,0,2),
                                seasonal_order=(2, 0, 2, 11),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results_full_data_model = full_data_model.fit(maxiter=1000)
print(results_full_data_model.summary())
```

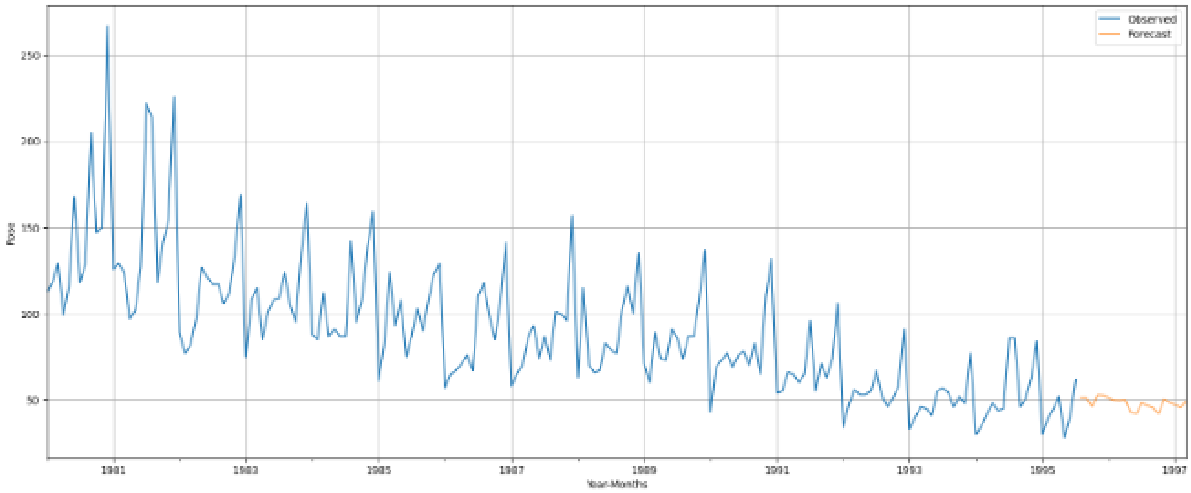
```
pred_full_manual_SARIMA_date = predicted_manual_SARIMA_full_data.summary_frame(alpha=0.05)
pred_full_manual_SARIMA_date.head()
```

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	51.133971	21.889031	8.232258	94.035684
1995-09-30	51.103433	22.212479	7.567774	94.639092
1995-10-31	46.241902	22.218317	2.694800	89.789004
1995-11-30	52.865439	22.238139	9.279488	96.451391
1995-12-31	52.741578	22.255414	9.121768	96.361388

```
rmse = mean_squared_error(df2['Rose'],results_full_data_model.fittedvalues,squared=False)
print('RMSE of the Full Model',rmse)
```

```
RMSE of the Full Model 30.858872322018925
```

This plot signifies about the behaviour of the future predict months.



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

- In all the built smoothing model, TES with additive decomposition approach has minimum error at place

- We can suggest that SARIMA model based on ACF and PACF cut off plot perform much better than any other time series-built model.

- Based on the result, companies wine sales in Rose segment takes a downtrend in a year by year. We have seen some sudden spikes in a year may be due to some special occasion. If companies works on its quality and marketing campaign, company can stablies the sales downfall. As per predicted plot, Rose wine sales will be stable for the coming year.