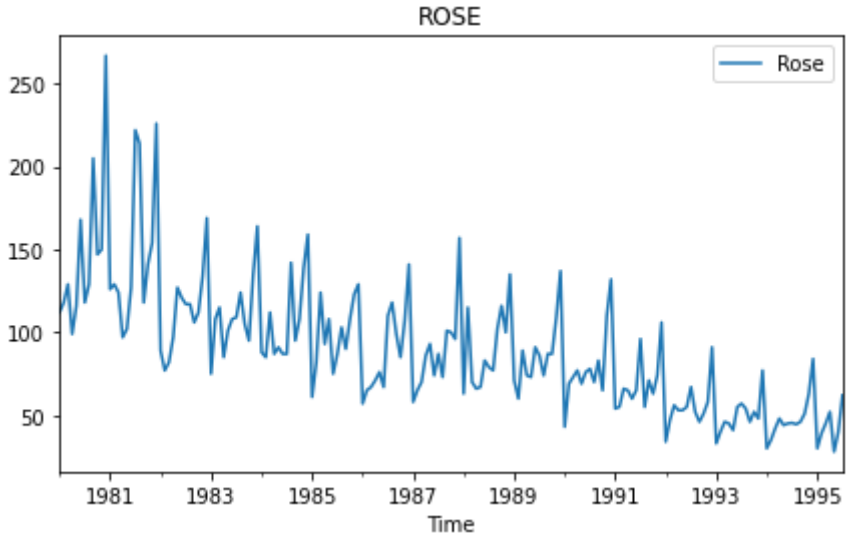
Q1. Read the data as an appropriate Time Series data and plot the data.

The csv file “Rose” is imported and the index is rest to a time stamp explicitly to get a sense of a time-series. Here is the head of the data.

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

After plotting, the data looks like following: -



The data has some sense of seasonality and a decreasing trend.

Q2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Following is the description of the data: -

```
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    187 non-null    float64
```

It is quite clear from the info. the data contains only one column of int type.

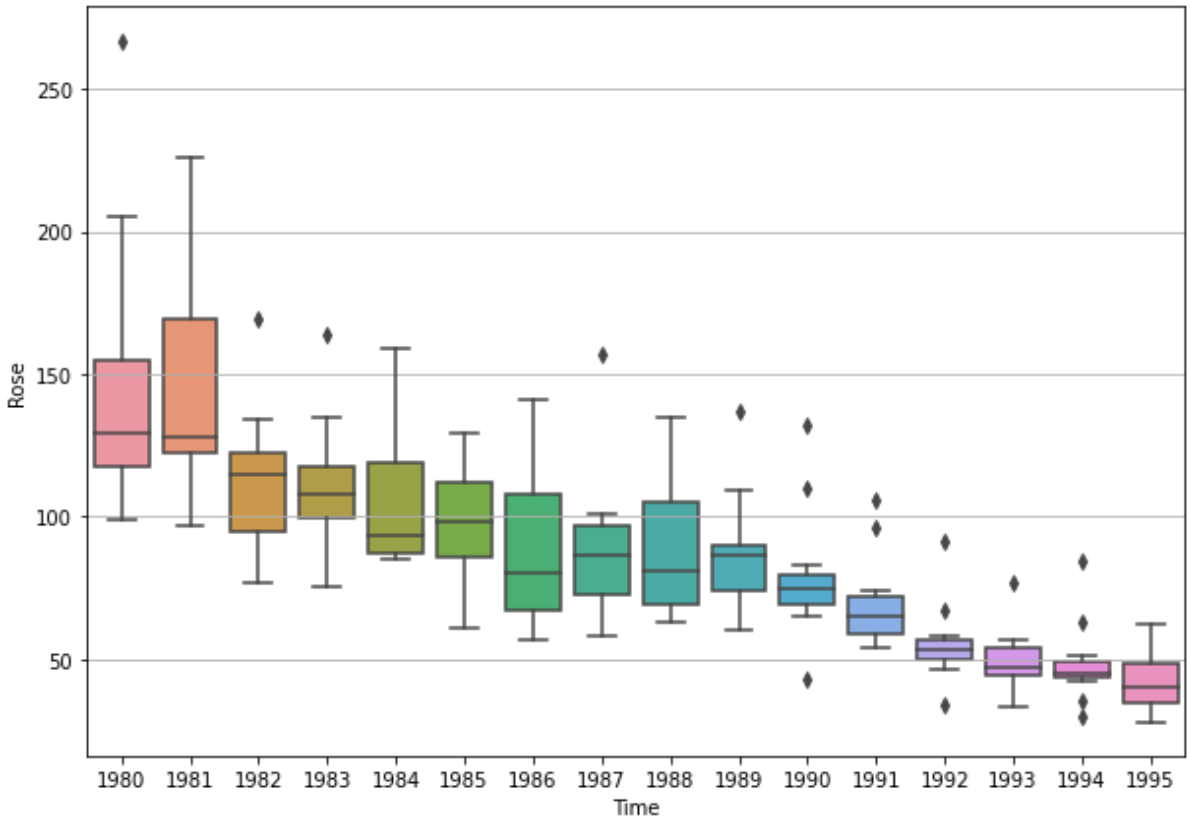
	Rose
count	187.000000
mean	89.908354
std	39.245313
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

```
df.isnull().sum()
```

```
Rose    0
dtype: int64
```

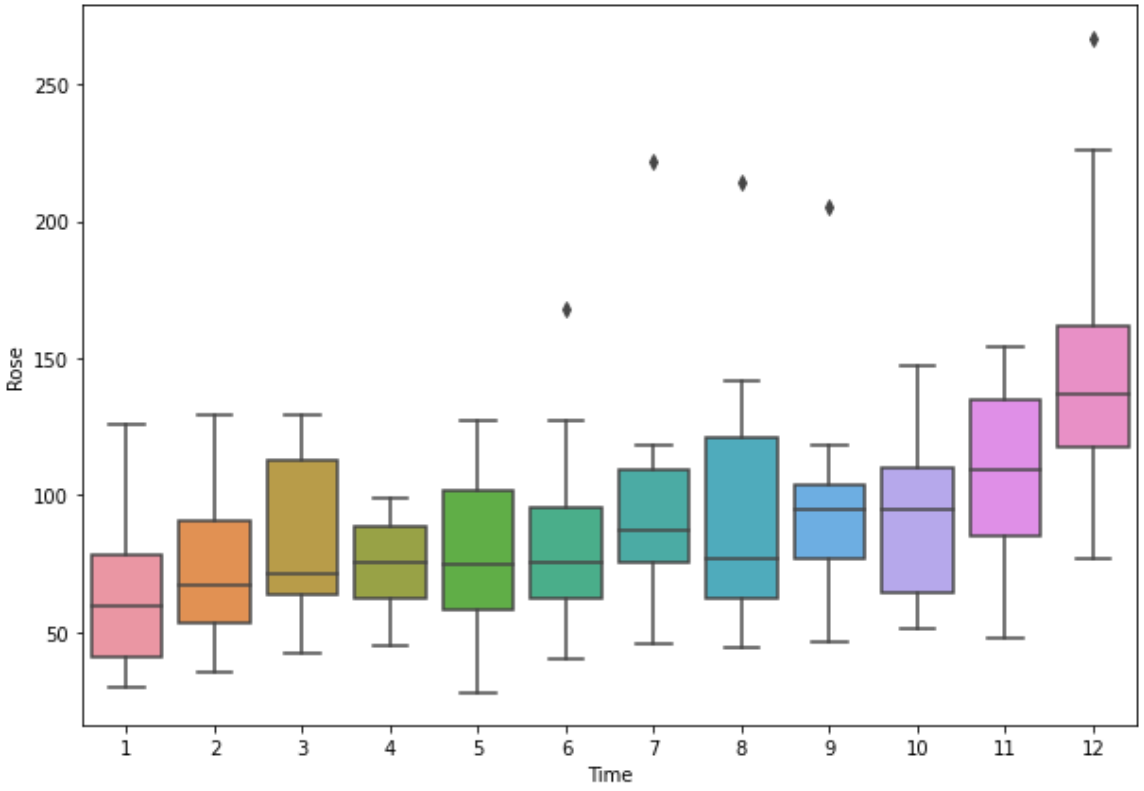
The data has no missing values. (The values were explicitly interpolated and directly changed in the csv file, as there were only two values.)

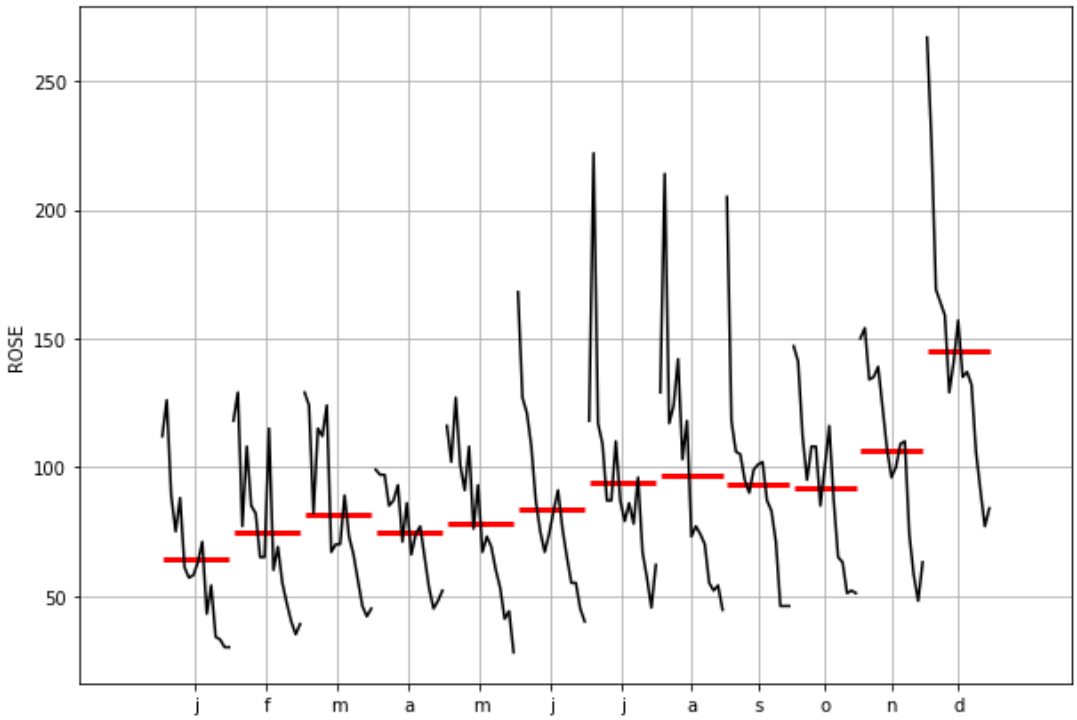
Yearly Plot: -



The above plot depicts the yearly sales of the wine. The sales is fairly high in 1981, and gradually increasing therafter.

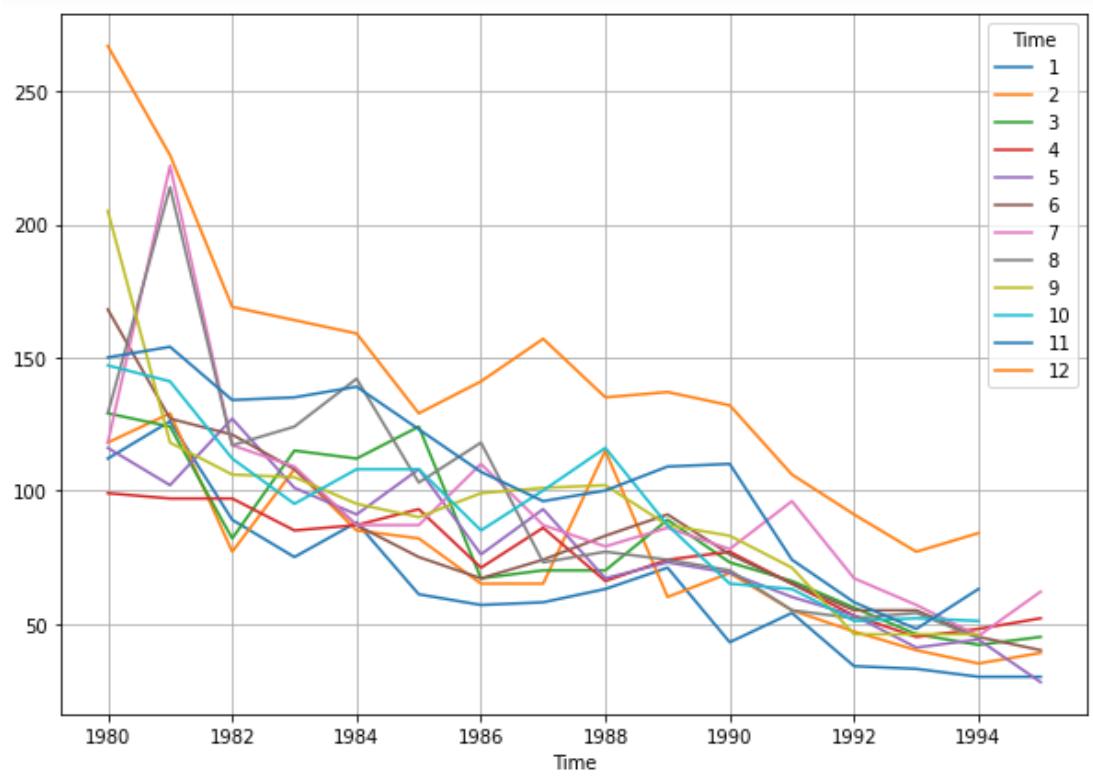
Monthly Plot





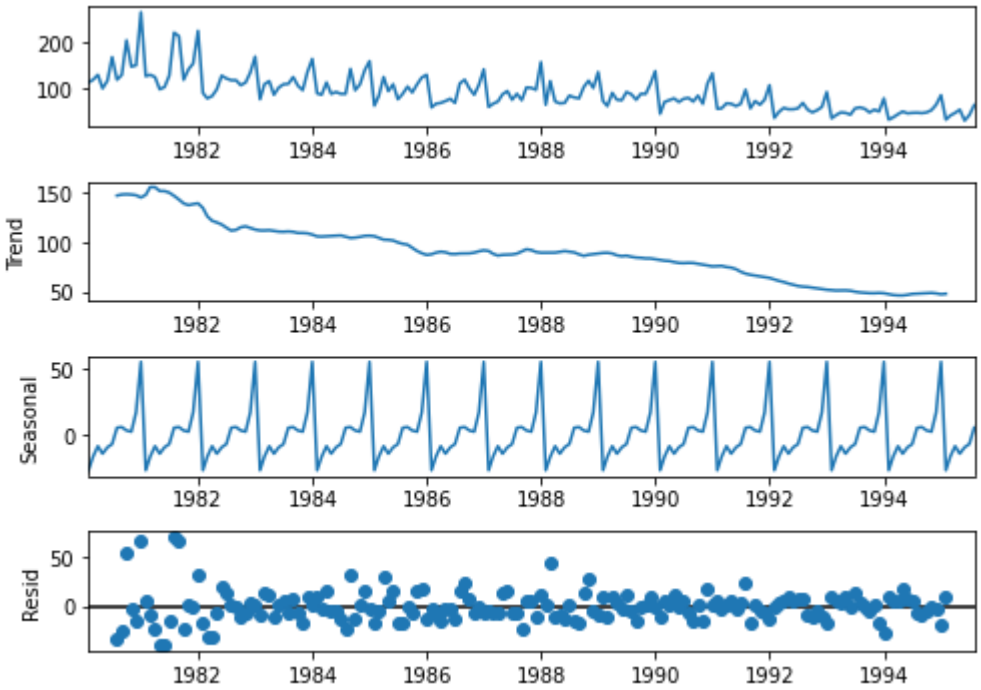
From the monthly plot we can infer that the sales are particularly skyrocketing in the month of December. It should not come as a surprise as the month of December is a festive month (Christmas + New Year). However, the month of July the sales peak, which is but obvious a positive sign for the company, but if the reason is unveiled, one could utilize that more efficiently.

Year-Month Plot



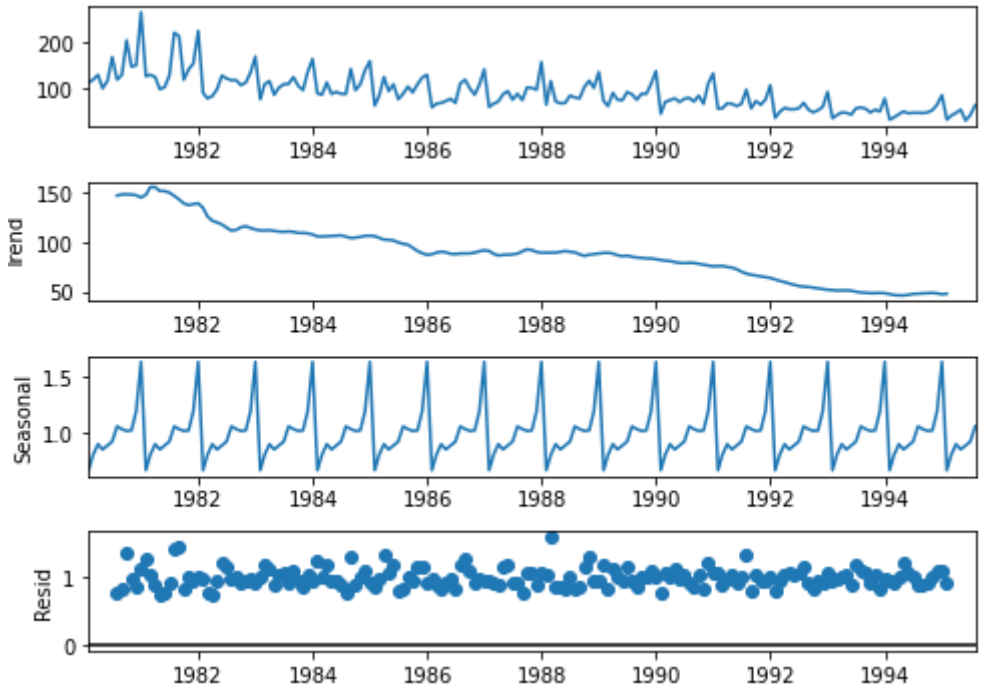
It is evident from the above pot that the sales in the month of December across all years have remained highest. However, the decreasing sales over the incoming years do not show good signs for the company.

Data decomposition: Additive



The residuals, do not show any particular pattern, they are completely random, thus the additive decomposition holds good here. However, we can't move ahead with the multiplicative decomposition.

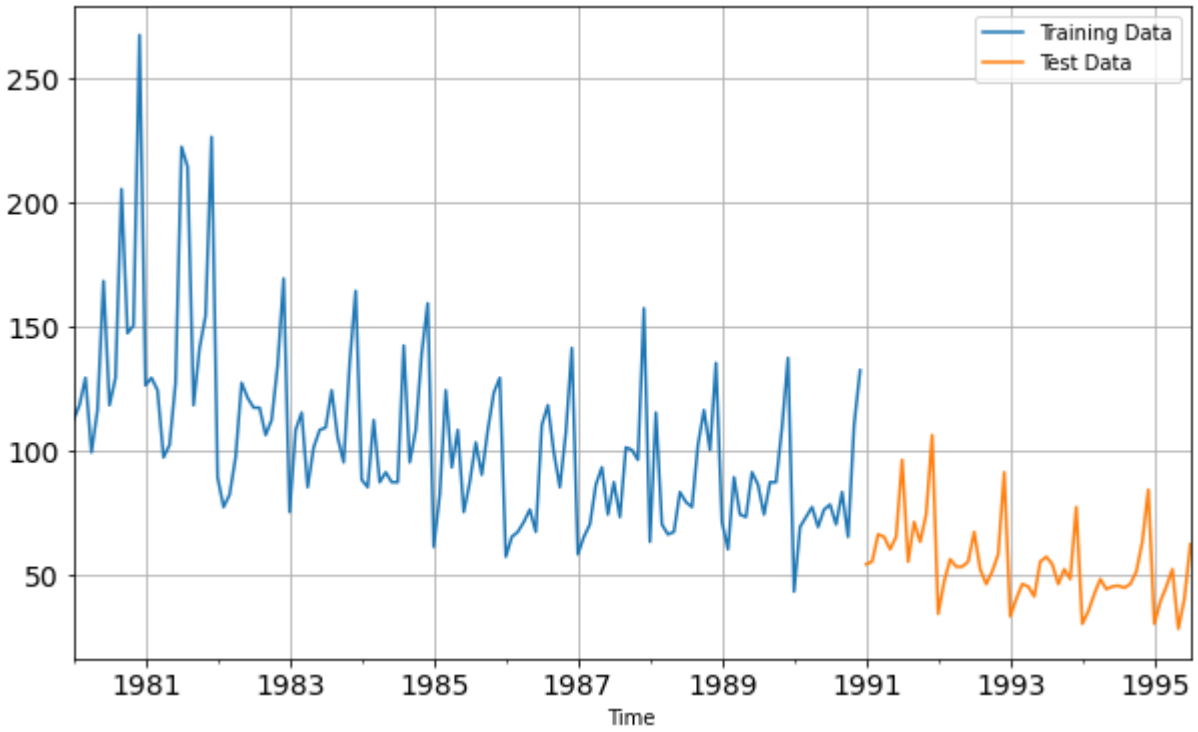
Multiplicative Decomposition:



Mostly the residuals are between 1 and 1.5. Multiplicative model too holds good here.

Q3 Split the data into training and test. The test data should start in 1991.

The data is split accordingly: -

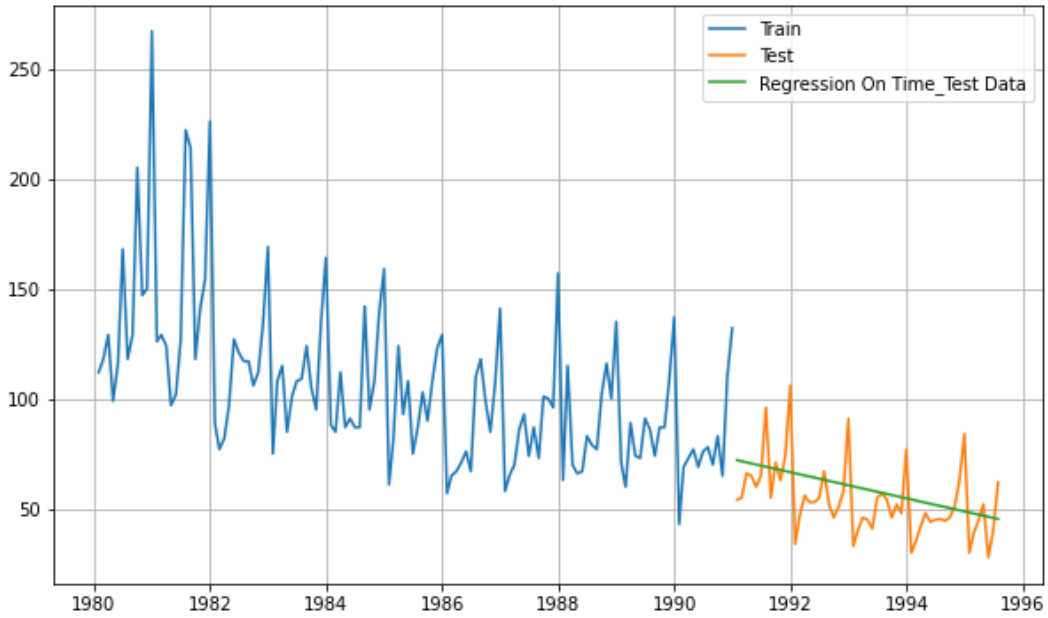


Q4 Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.

Following models are built: -

1. Linear Regression
2. Naïve Model
3. Simple average
4. Exponential Smoothing

Linear Regression: -

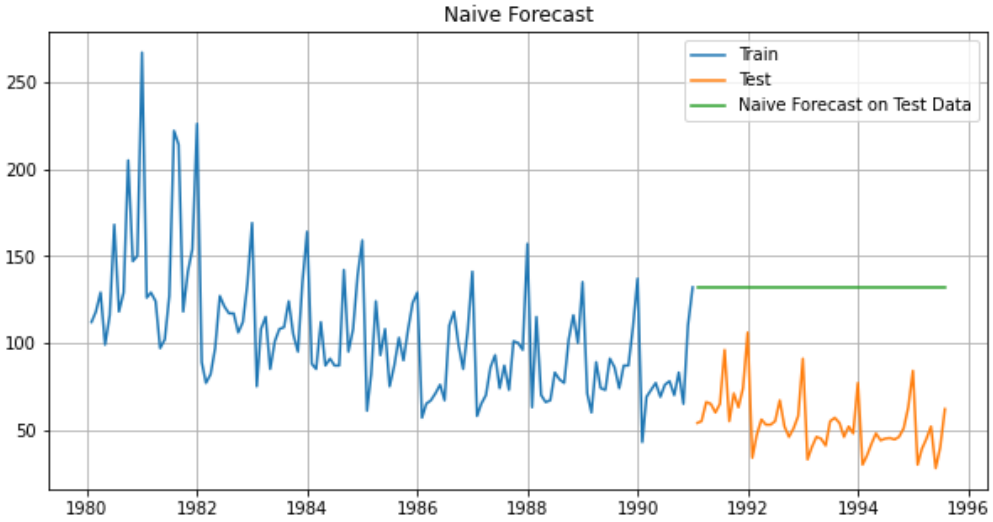


The regression model considers the sales as the target variable and the time stamp as the independent variable. It just takes into account the possible trend and predicts accordingly. Not very accurate.

	Test RMSE
RegressionOnTime	15.276693

The RMSE score is very good, and definitely could be used for model building, but again we need to keep in mind this does not capture trend and seasonality.

Naïve Approach



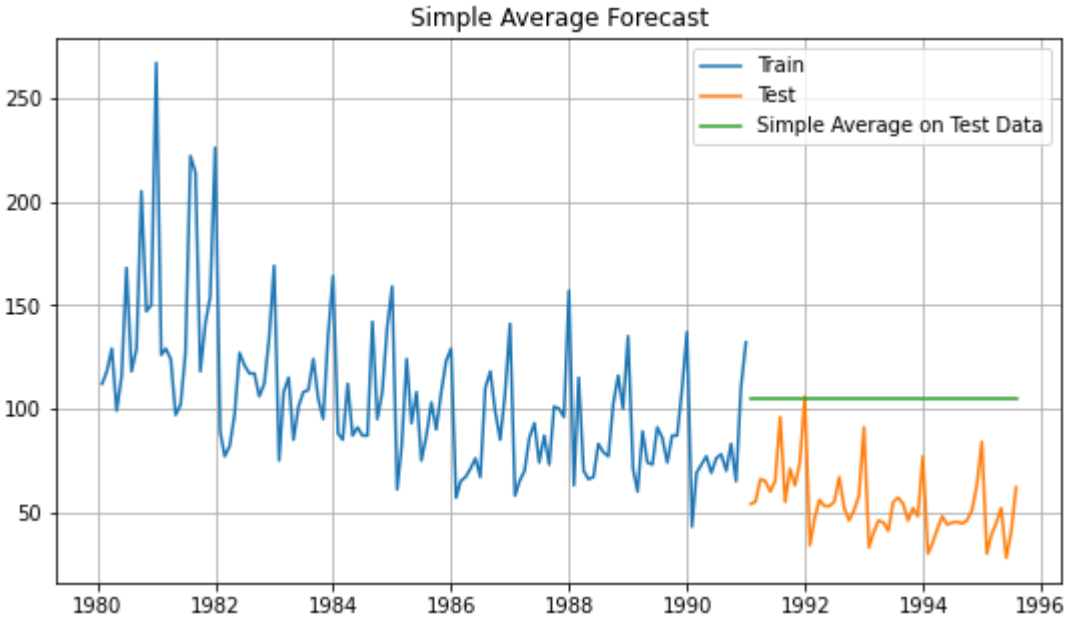
The Naïve approach just takes the latest value and presents it as upcoming forecast.

	Test RMSE
RegressionOnTime	15.276693
NaiveModel	79.741326

Evident that the model seems almost of no use, with an even higher RMSE.

Simple average

It considers the average on the whole data and presents it as the forecast.

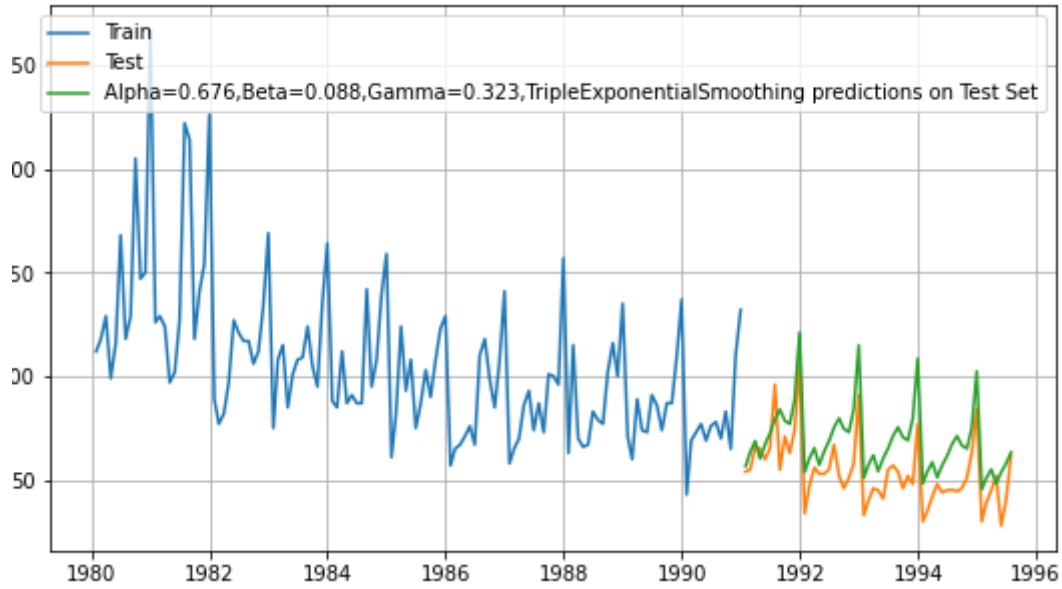


	Test RMSE
RegressionOnTime	15.276693
NaiveModel	79.741326
SimpleAverageModel	53.483727

The simple average has better RMSE than the NAÏVE model, but still less than the Linear regression model.

Exponential Smoothing

The data has some amount of trend and an evident seasonality. A triple exponential smoothing is the one that could fit it the best.



As can be seen, the testing data and the predictions are so close.

	Test RMSE
RegressionOnTime	15.276693
NaiveModel	79.741326
SimpleAverageModel	53.483727
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	17.400641

The RMSE has significantly decreased.

Q5. 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.

H0: The time series has a unit root and is not stationary.

H1: The time series does not have a unit root and is stationary.

From the ADF test, we can arrive at a conclusion that p>.05 and we can't reject NULL, thus the series is not stationary. We need to follow certain steps of differencing to make it stationary.

```
dftest = adfuller(df,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])
```

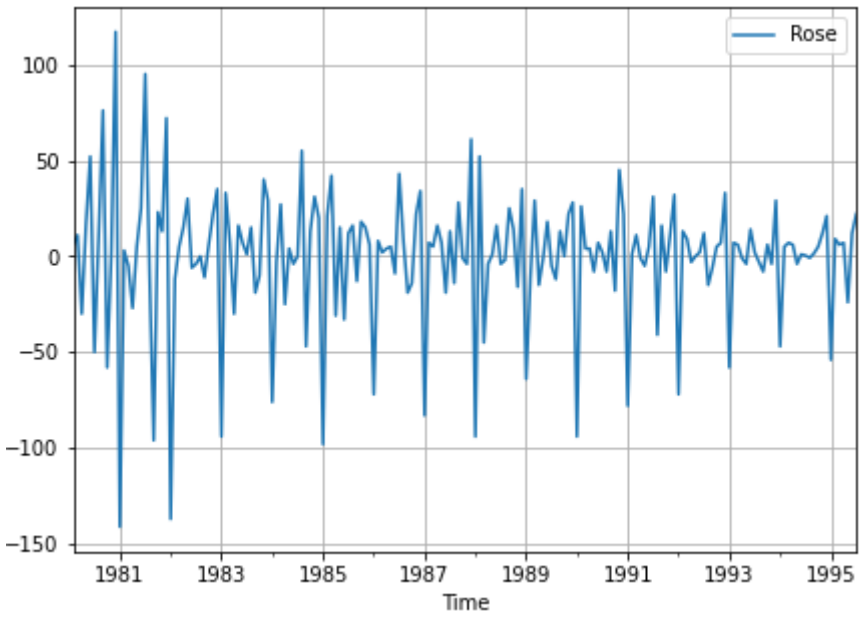
```
DF test statistic is -2.242
DF test p-value is 0.466437102037184
Number of lags used 13
```

```
dftest = adfuller(df.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 -8.161
DF test p-value is 3.034192419073321e-11
Number of lags used 12
```

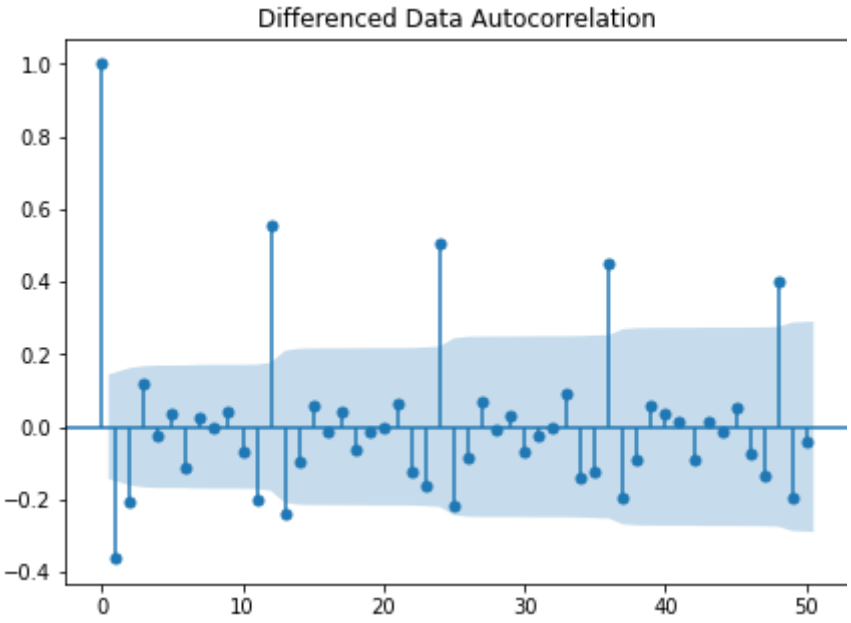
After one level of differencing, we can see the p value is<0.05 and we can thus reject the NULL hypothesis. Now the series is stationary.

As can be seen, after one level of differencing, the series is stationary.

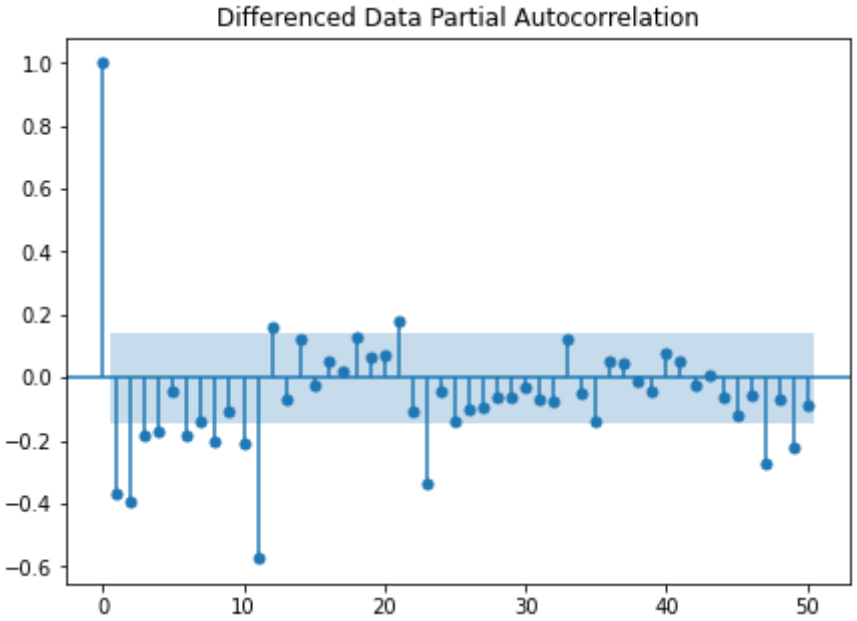


Q6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

To begin with one needs to check the ACF and PACF plot to get a rough idea about the p and q values. (this has been plotted on the entire data)



From the ACF plot, we can predict the q value to be 2.



From the PACF plots, the p value comes around 3.

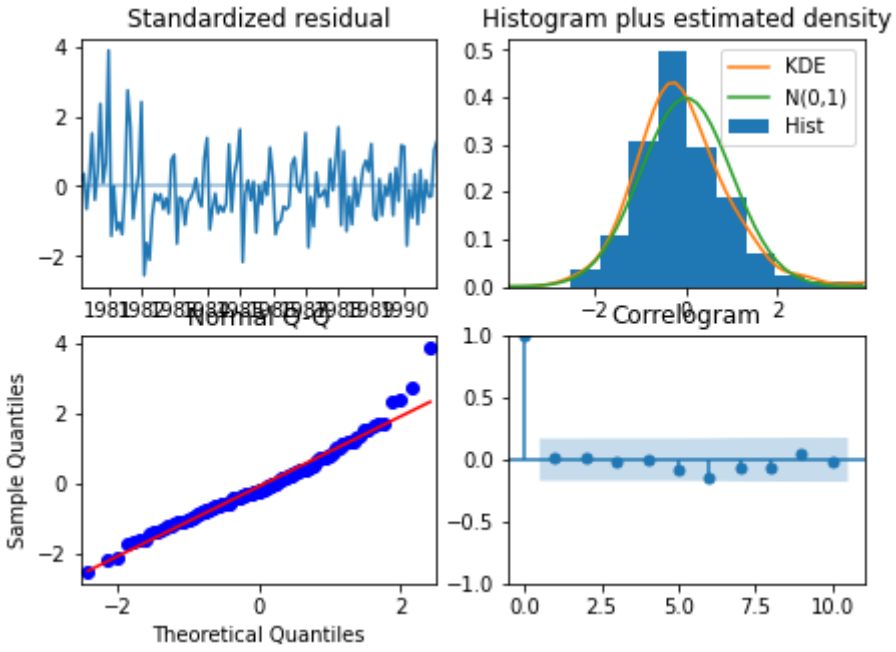
Building the Automated ARIMA model: -

	param	AIC
11	(2, 1, 3)	1274.695127
15	(3, 1, 3)	1278.663118
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

The least AIC comes for parameters (2,1,3). We already know, data is not stationary, and we need to apply differencing to make it stationary, thus d = 1.

Building, the model yields the following summary: -

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sat, 24 Apr 2021   AIC                           1274.695
Time:                        00:13:26   BIC                           1291.946
Sample:                    01-31-1980   HQIC                          1281.705
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6779      0.084    -20.050      0.000      -1.842      -1.514
ar.L2         -0.7289      0.084     -8.710      0.000      -0.893      -0.565
ma.L1          1.0448      0.662      1.577      0.115      -0.254       2.343
ma.L2         -0.7718      0.135     -5.717      0.000      -1.036      -0.507
ma.L3         -0.9047      0.601     -1.505      0.132      -2.083       0.274
sigma2       858.1766    557.564      1.539      0.124    -234.628    1950.981
===================================================================================
Ljung-Box (Q):                      101.06   Jarque-Bera (JB):                24.45
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```



The residuals show a little deviation from the original.

Prediction on the test data: -

	RMSE	MAPE
ARIMA(2,1,2)	36.839348	75.933747

Yields a bit higher RMSE.

Building automated SARIMA model: -

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	951.744322
59	(0, 1, 3)	(2, 0, 3, 6)	952.073632
251	(3, 1, 3)	(2, 0, 3, 6)	952.582717
191	(2, 1, 3)	(3, 0, 3, 6)	953.205610
123	(1, 1, 3)	(2, 0, 3, 6)	953.684951

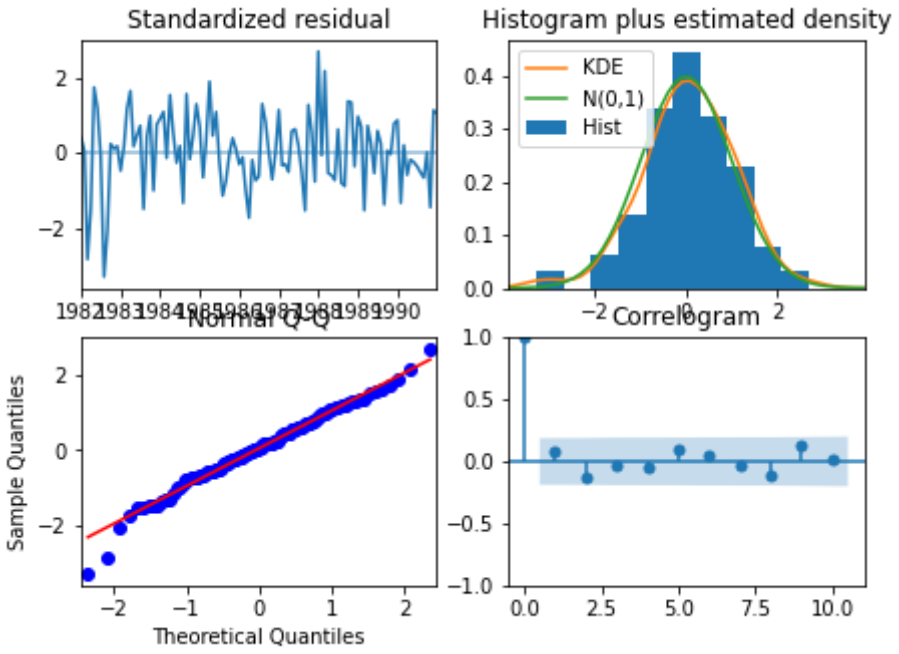
Building model based on the least AIC values: -

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 0, 3, 6)	Log Likelihood	-464.872
Date:	Sat, 24 Apr 2021	AIC	951.744
Time:	00:36:34	BIC	981.349
Sample:	01-31-1980 - 12-31-1990	HQIC	963.750
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5027	0.083	-6.076	0.000	-0.665	-0.341
ar.L2	-0.6627	0.084	-7.913	0.000	-0.827	-0.499
ma.L1	-0.3714	67.008	-0.006	0.996	-131.704	130.961
ma.L2	0.2033	42.138	0.005	0.996	-82.386	82.792
ma.L3	-0.8319	55.791	-0.015	0.988	-110.179	108.516
ar.S.L6	-0.0837	0.049	-1.718	0.086	-0.179	0.012
ar.S.L12	0.8099	0.052	15.455	0.000	0.707	0.913
ma.S.L6	0.1697	0.246	0.690	0.490	-0.313	0.652
ma.S.L12	-0.5642	0.198	-2.853	0.004	-0.952	-0.177
ma.S.L18	0.1707	0.142	1.198	0.231	-0.109	0.450
sigma2	261.0331	1.75e+04	0.015	0.988	-3.4e+04	3.45e+04

Ljung-Box (Q):	26.07	Jarque-Bera (JB):	4.77
Prob(Q):	0.96	Prob(JB):	0.09
Heteroskedasticity (H):	0.54	Skew:	-0.36
Prob(H) (two-sided):	0.06	Kurtosis:	3.73



Quite a bit deviation can be seen in the sample vs theoretical quantities.

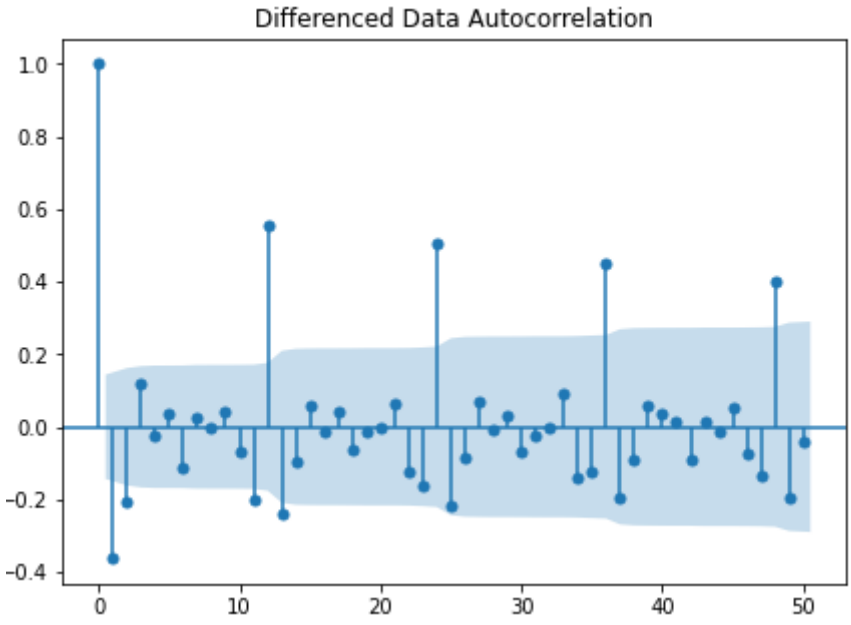
Predicting on the test data: -

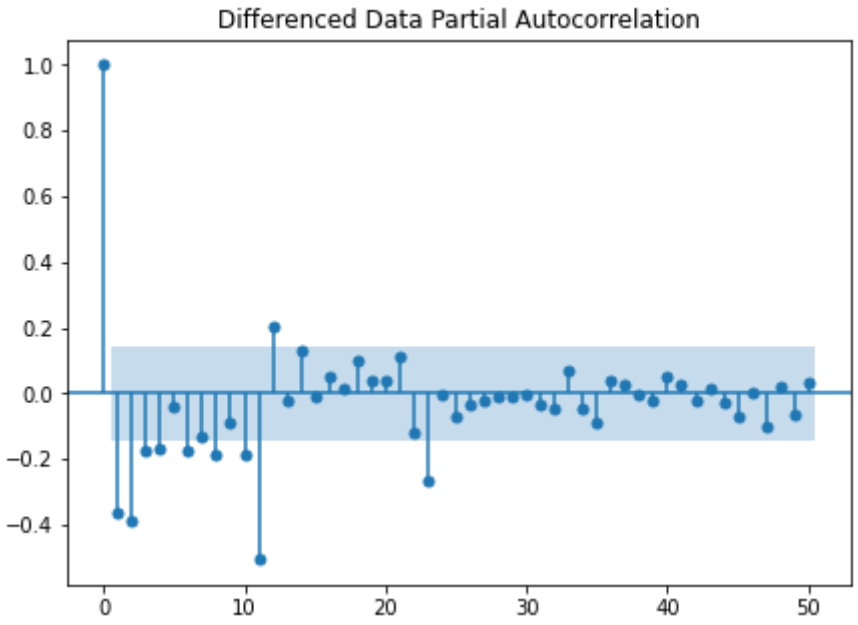
	RMSE	MAPE
ARIMA(2,1,2)	36.839348	75.933747
SARIMA(2,1,3)(2,0,3,6)	27.149923	55.319863

The RMSE have significantly dropped with the SARIMA Model.

Q7. 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.

Looking at the ACF and PACF plot to determine p and q: -





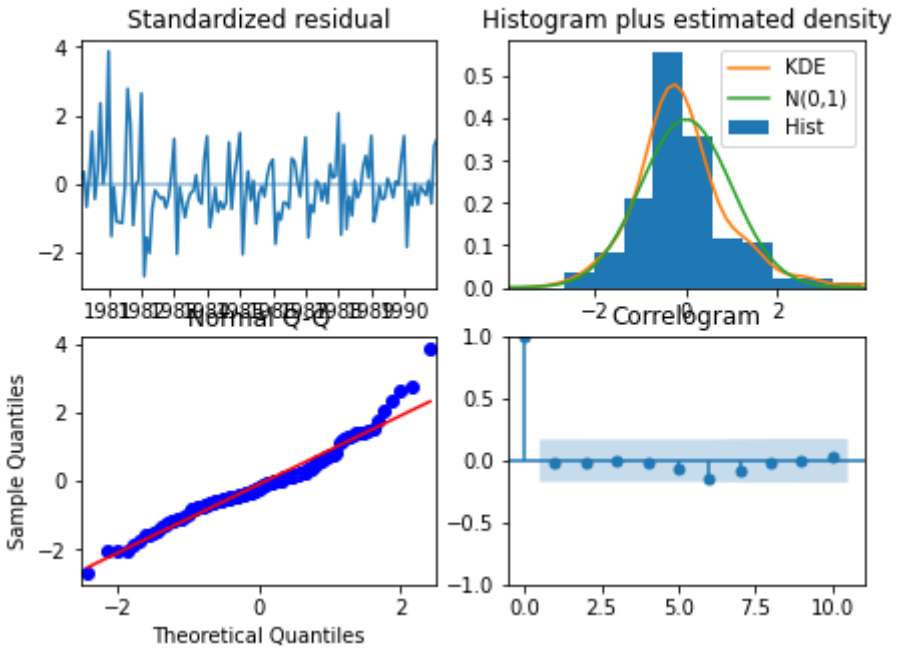
From the above two plots q = 2, and p = 4.

The manual model yields the following results: -

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(4, 1, 2)	Log Likelihood	-635.859
Date:	Sat, 24 Apr 2021	AIC	1285.718
Time:	00:40:18	BIC	1305.845
Sample:	01-31-1980	HQIC	1293.896
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3838	0.923	-0.416	0.677	-2.192	1.425
ar.L2	0.0046	0.258	0.018	0.986	-0.502	0.511
ar.L3	0.0414	0.113	0.366	0.714	-0.180	0.263
ar.L4	-0.0054	0.177	-0.031	0.976	-0.353	0.342
ma.L1	-0.3239	0.933	-0.347	0.729	-2.153	1.505
ma.L2	-0.5407	0.874	-0.619	0.536	-2.254	1.172
sigma2	951.1524	93.870	10.133	0.000	767.170	1135.135



	RMSE	MAPE
ARIMA(2,1,2)	36.839348	75.933747
SARIMA(2,1,3)(2,0,3,6)	27.149923	55.319863
ARIMA(3,1,2)	37.061202	76.498431

Prediction on the test data, shows higher value than automated ARIMA, but still not the best.

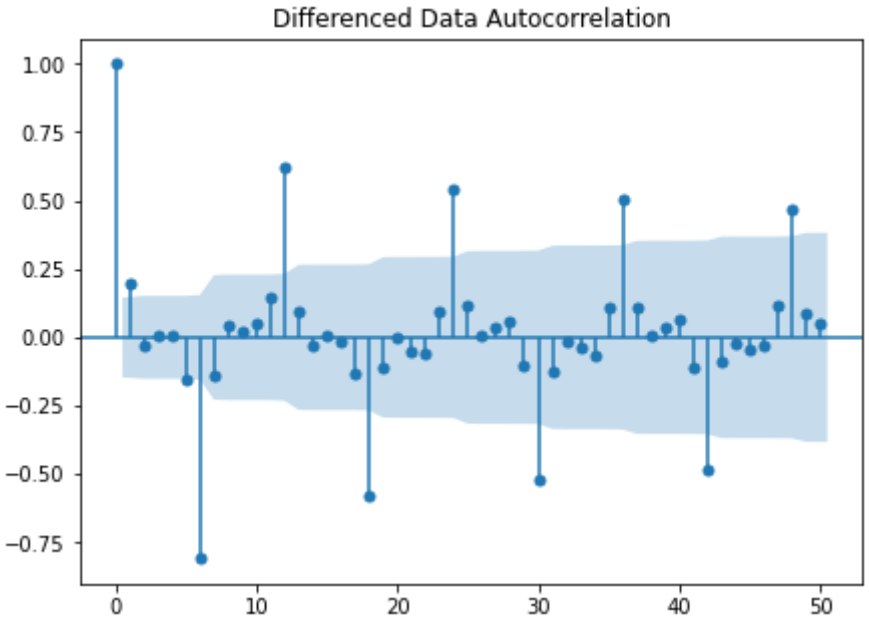
Moving ahead with the manual SARIMA.

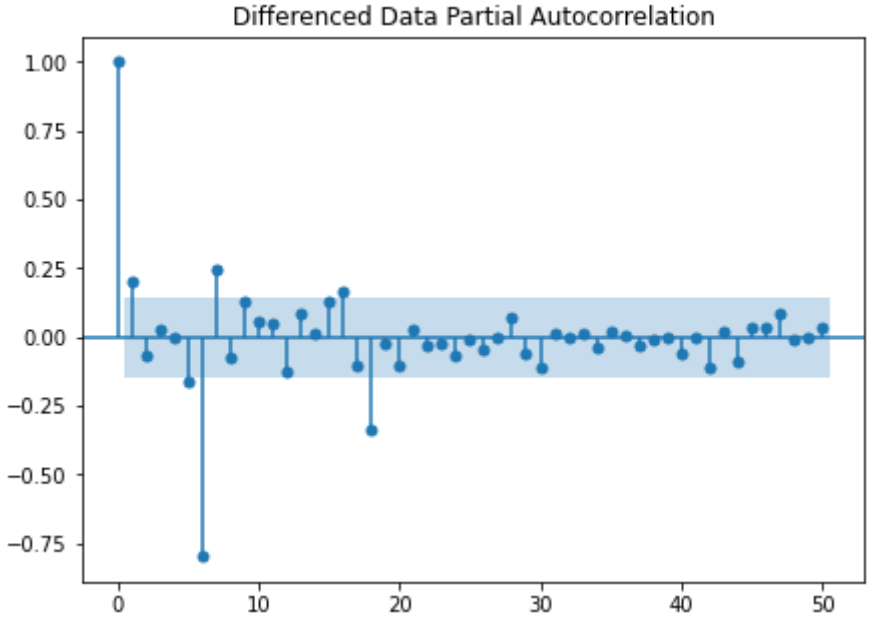
Looking at the stationarity: -

```
DF test statistic is -9.255
DF test p-value is 9.724172783305914e-14
Number of lags used 6
```

The data is stationary and thus D = 0.

Looking at the seasonal ACF and PACF with differencing 6: -





From the above two plots P = 2 and Q can't be determined.

Finally, we have: -

p = 1

q = 2

d = 1

P = 2

Q = 1

D = 0

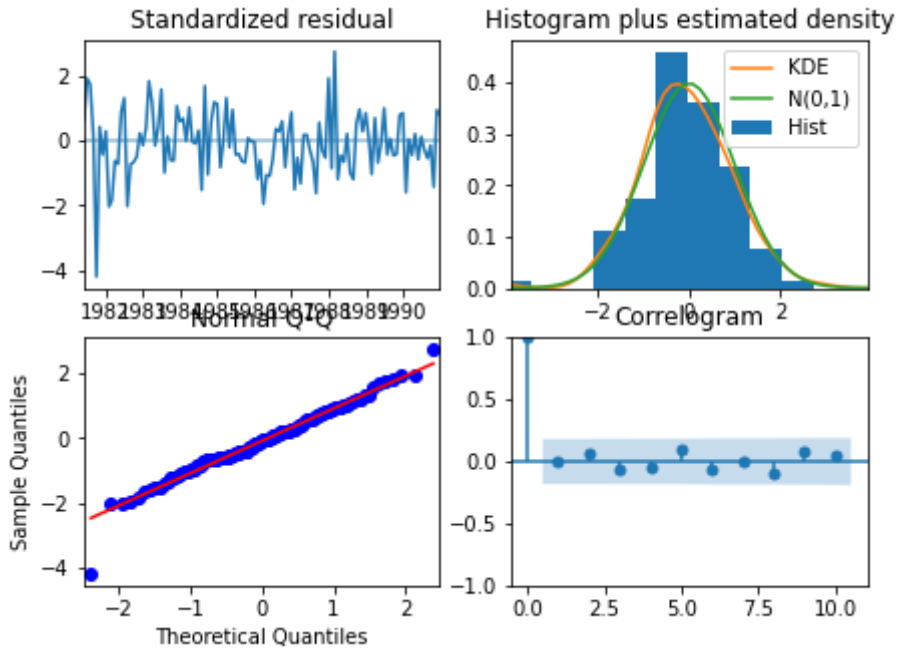
The final model yields the following results: -

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(2, 0, [1], 6)	Log Likelihood	-518.906
Date:	Sat, 24 Apr 2021	AIC	1055.811
Time:	00:44:13	BIC	1080.593
Sample:	01-31-1980 - 12-31-1990	HQIC	1065.871
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.1198	0.107	10.419	0.000	0.909	1.330
ar.L2	-0.3391	0.121	-2.795	0.005	-0.577	-0.101
ar.L3	0.1543	0.073	2.116	0.034	0.011	0.297
ma.L1	-1.9963	13.718	-0.146	0.884	-28.883	24.890
ma.L2	1.0001	13.744	0.073	0.942	-25.938	27.939
ar.S.L6	-0.4895	0.070	-6.988	0.000	-0.627	-0.352
ar.S.L12	0.3964	0.070	5.637	0.000	0.259	0.534
ma.S.L6	0.7074	0.088	8.032	0.000	0.535	0.880
sigma2	395.7069	5448.067	0.073	0.942	-1.03e+04	1.11e+04

Ljung-Box (Q):	27.79	Jarque-Bera (JB):	15.66
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	0.55	Skew:	-0.33
Prob(H) (two-sided):	0.06	Kurtosis:	4.68



Finally, plotting predicting on the test data, the least value for manual SARIMA is obtained: -

	RMSE	MAPE
ARIMA(2,1,2)	36.839348	75.933747
SARIMA(2,1,3)(2,0,3,6)	27.149923	55.319863
ARIMA(3,1,2)	37.061202	76.498431
SARIMA(2,1,2)(1,0,1,6)	27.511177	55.619355

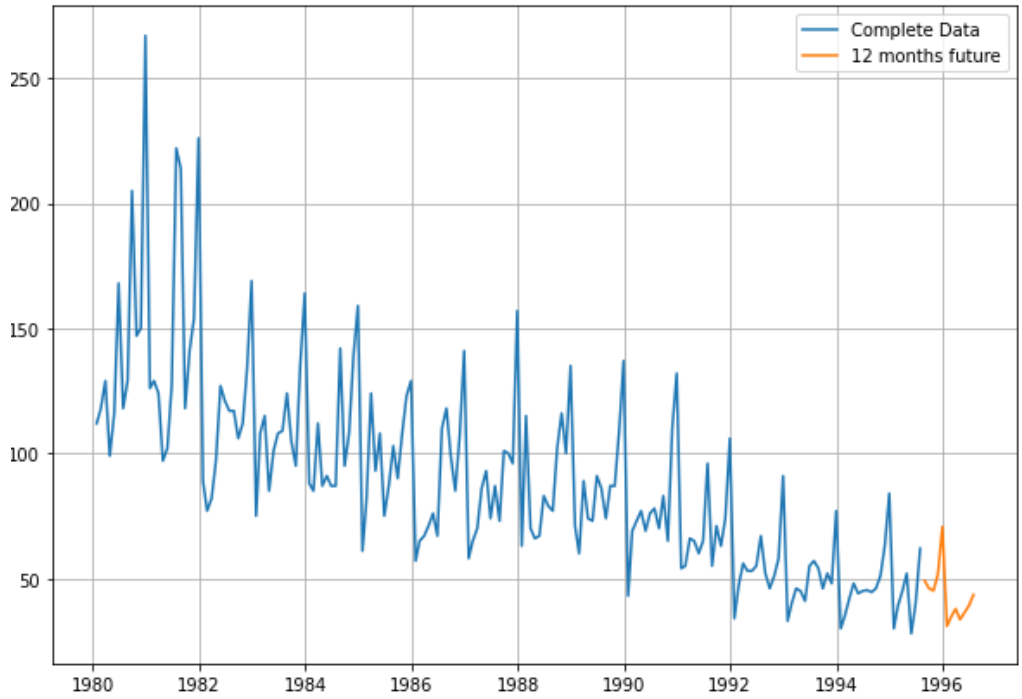
Q8 Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

MODEL	RMSE
Linear Regression	15.27
Naïve Approach	79.47
Simple Average	53.48
Exponential Smoothing	17.40
Automated ARIMA (2,1,2)	36.83
Automated SARIMA (2,1,3)(2,0,3,6)	27.14
Manual ARIMA(3,1,2)	37.06
Manual SARIMA(2,1,2)(1,0,1,6)	27.51

Although, here Linear Regression might seem as the best model, but we also need to take into account the seasonality and trend which is best captured by the Exponential smoothing model, thus making it the best choice.

Q9. 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.

Building the model on the whole data: -



Finally, the prediction into the future looks like above.

Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

The best model capturing the trend and seasonality is the Exponential smoothing model. Firstly, the company must figure out the reason for declining sales and act upon it. In the monthly plot, a high spike was observed in the months of July, which the company should try to target.

Company must bring out exciting offers and discounts for the wine, to increase the sales. The company could give certain credit points on every purchase which can be used during the next sales. Additional discounts could be provided on the peak season, i.e., month of December where sales are the highest.

As far as future is concerned, there is more dip expected in sales in the coming 12 months, which the company must look into seriously and act upon the suggestions accordingly.

The best possible was is to run a root cause analysis and considering an fishbone diagram to identify the cause of declining sales and target that directly.