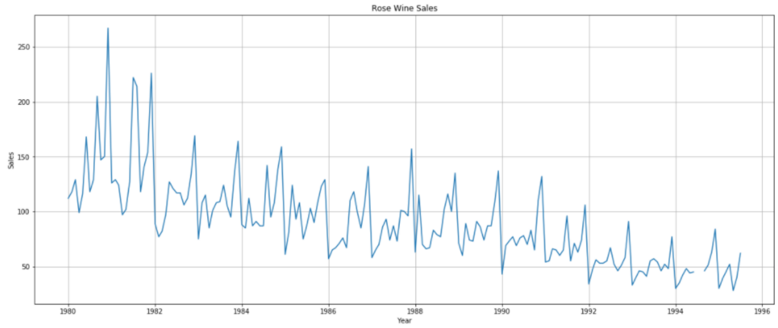
Time Series Forecasting Project

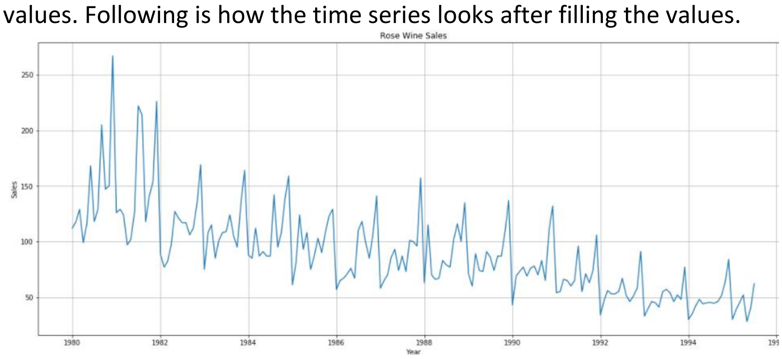
For this particular dataset, the data of different types of wine sales in the 20th century is to be analysed. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

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

The dataset is the monthly sales of rose wine of a company from 1980 to 1995. It is a time series data with frequency of one month.



The rose dataset has monthly sales data from January 1980 to July 1995. Two data points of July 1994 and August 1994 is missing. We will be using the spline interpolation method to fill in the missing values. Following is how the time series looks after filling the values.



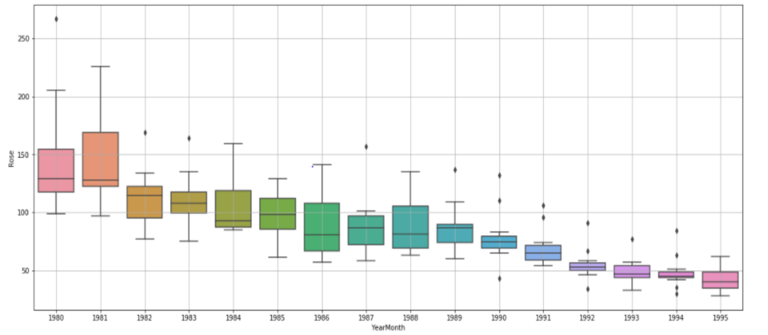
The highest sale was recorded in December 1980, after which there seems to be a downward trend over the years. Seasonality seems to be present, we l verify its using decomposition.

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

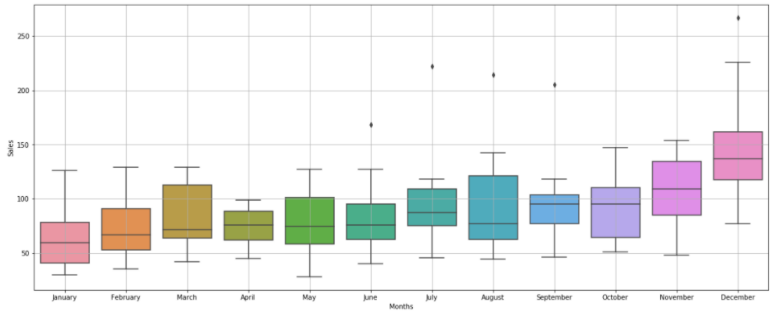
The basic measures of descriptive statistics tell us how the Sales have varied across years. For this measure of descriptive statistics we have averaged over the whole data without taking the time component into account.

	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

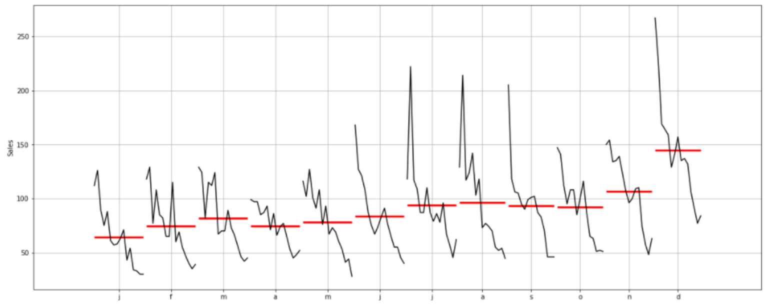
Below plot gives the average sales of the rose wine sales across the years. We can see that the average yearly sales were the highest in 1980, almost equal in 1981and then declined steadily over the years.



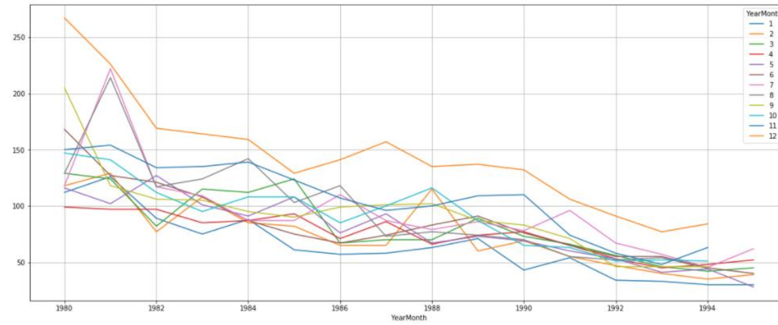
Below plot gives the average monthly sales of rose wine. We can conclude that highest wine sales were in the month of December, followed by November, and the lowest in January. Hence we can say there is seasonality in our rose data set.



Below plot is another way of plotting the monthly median, highest and lowest values. December has the highest median also the highest peak.



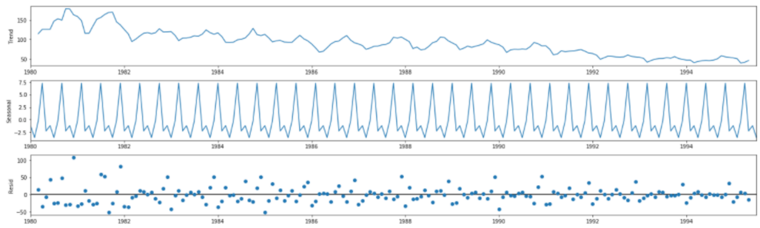
The below plot gives the monthly sales across the years. December has the highest sale across all years from 1980 to 1994



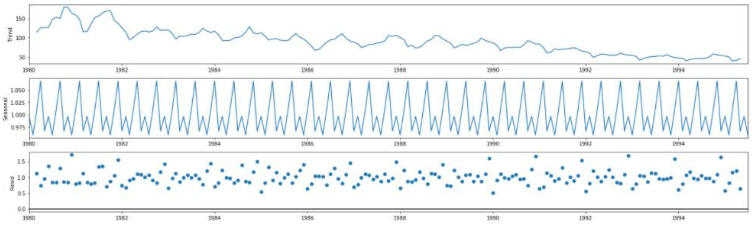
Decomposition

We will be doing decomposition by dividing the time series into trend, seasonality and residual/ noise component. Below is the plot for additive model and multiplicative model.

Additive Model



Multiplicative Model



After decomposition in both additive and multiplicative model we can say that there is a downward trend in the sales of rose wine over the years. There is also a definitive seasonality present in our dataset. There is also a significant amount of unexplained component/error in both the models. The spread of residual is high in our model (Slightly greater spread in the additive model).

3. Split the data into training and test.

Next we have divided the rose and sparking datasets into train and test data. Test data starts in 1991, following is the shape of the test and train data.

	Train	Test
Rose	(132,2)	(55,2)

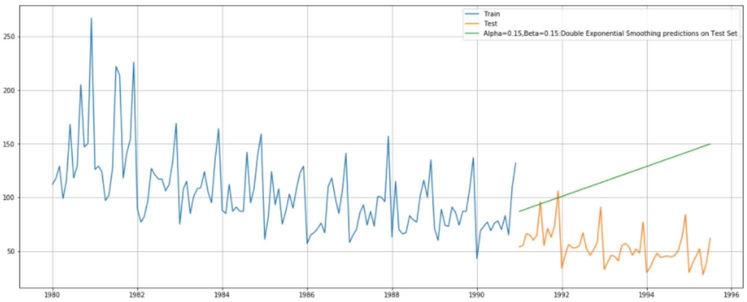
4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Exponential Smoothing Models

Here we will be building multiple exponential smoothing models. First we have Single Exponential Smoothing (SES) model, SES has only level, i.e. the local mean and the level is determined by the value of alpha. This model is suitable for no clear trend or seasonality.
Next we have double exponential smoothing, also known as Holt's model. This model has 2 parameters, level and trend captured by alpha and beta. This model is not suitable if the dataset has seasonality.
Then we have triple exponential smoothing (TES), also known as Holt's Winter model. This model has 3 parameters, level, trend and seasonality captured by alpha, beta and gamma. The value of alpha, beta and gamma lie between 0 and 1.
Looking at our data, we will go ahead with both Double Exponential Smoothing and Triple Exponential Smoothing, and select the one with the lowest RMSE.

DES

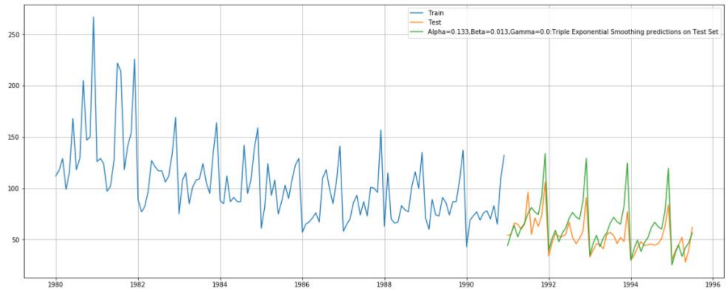
After running the double exponential smoothing, we found out the value of alpha = 0.15 and beta = 0.15. DES doesn't consider the seasonality component; it only has level and trend. Its RMSE value is 70.5. Following is the plot of prediction vs. expected.



Based on the plot and RMSE, DES doesn't seem to make sense, let's try TES.

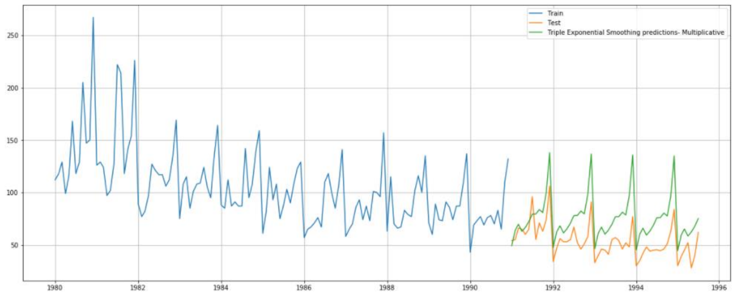
TES

Running TES with additive seasonality on rose data set, we found out the following values. Alpha=0.133, Beta=0.013, Gamma=0.0. It gives a better RSME of 16.5. Following is the plot of the predicted values vs. actual values post 1991.



Trend seems to be captured in the model, however there seems to be error at the peaks, as our forecasted model shows higher peaks Running TES with multiplicative seasonality on rose data set, we found out the following values. Alpha= 7.54e-11, Beta= 4.89e-14 , Gamma= 0.21. It gives a lower RSME of 25.2.

Following is the plot of the predicted values vs. actual values post 1991.



Larger discrepancies in the actual vs. predicted plot as compared to additive TES

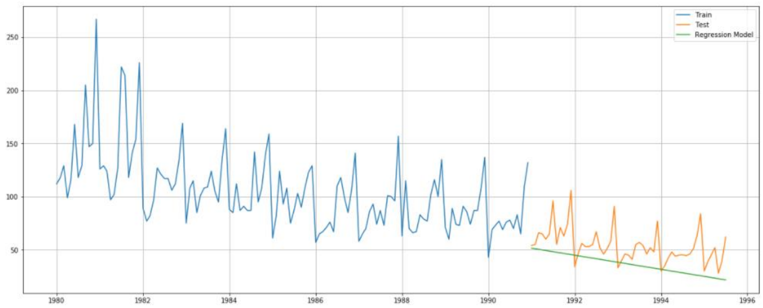
RMSE comparison

	Test RMSE
TES Additive	16.475042
TES Multiplicative	25.210414
DES	70.599377

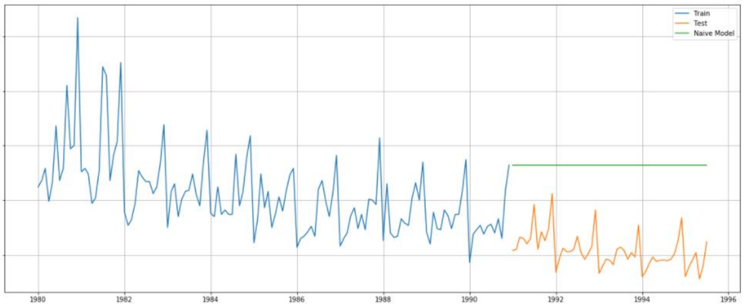
TES additive so far gives the lowest RMSE for Rose dataset

Regression, Naïve and Average models

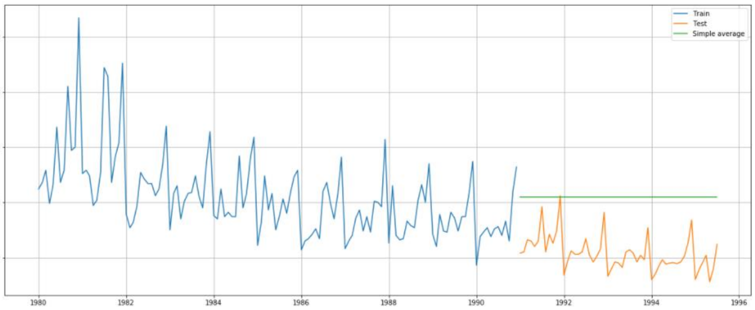
Building a regression model to predict the values from 1991. The RMSE is 22.57 and below is the plot.



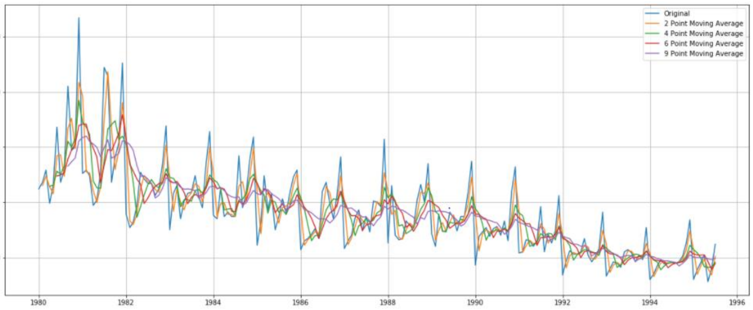
The regression model captures trend beautifully but not the seasonality being a linear regression model.
The Naïve forecast takes the previous value and keeps the forecast constant at the same value, hence the name naïve. This is how the forecasting plot of naïve looks like.



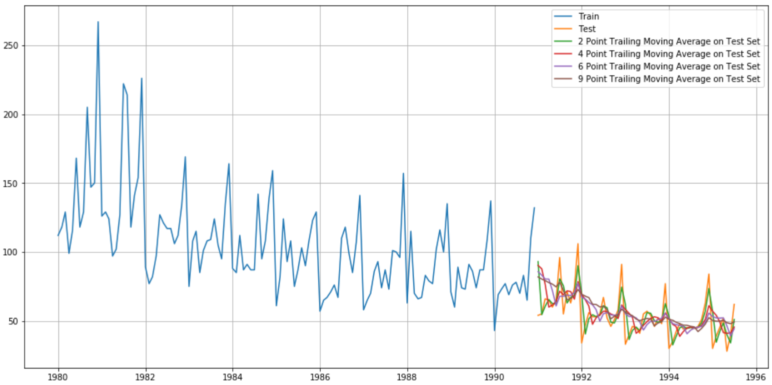
The RMSE of naïve model is as expected really high at 79.7
Simple average model as the name suggests takes the average of the train data and plots the same mean as the test forecast.



104.9 is the mean of the historical/training data which is the constant forecast for the test data. The RSME of Simple Average model is 53.4.
Next we will be calculating the moving or rolling averages for different intervals for rose dataset. Following is the plot for moving averages for interval 2, 4, 6 and 9.



We will split this moving average data into train and test at 1991 and plot the predicted data after 1991 and calculate the RMSE.



From the plot we can see that the green plot i.e. the 2 point moving average appears closest to the orange, i.e. the test plot. We can verify the same after calculating the RMSE for all moving averages.

```
Moving Average Model forecast on the Training Data - RMSE 11.530
Moving Average Model forecast on the Training Data - RMSE 14.457
Moving Average Model forecast on the Training Data - RMSE 14.572
Moving Average Model forecast on the Training Data - RMSE 14.732
```

Model Performance on test data

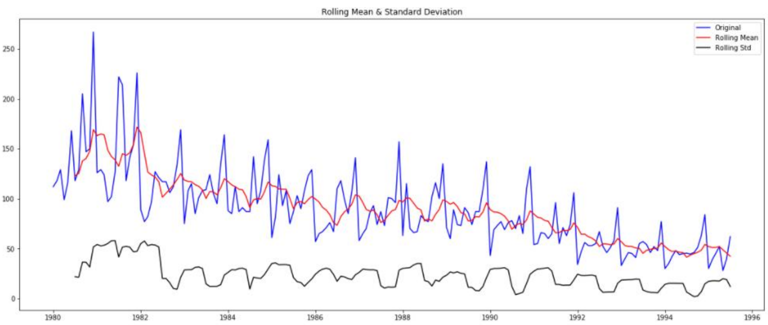
	Test RMSE
2 point Moving Average	11.529811
4 point Moving Average	14.457115
6 point Moving Average	14.571789
9 point Moving Average	14.731914
TES Additive	16.475042
Regression	22.573067
TES Multiplicative	25.210414
Simple Average Model	53.483727
DES	70.599377
Naive Model	79.741326

2 point moving average gives the lowest RMSE for Rose Data

5. Check for the stationarity of the data. If the data is found to be non-stationary, take appropriate steps to make it stationary.

A time series is 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 lag between the two periods.
To check if the data is stationary we will do the Augmented Dickey Fuller test. For that we need to define the null and alternate hypothesis.

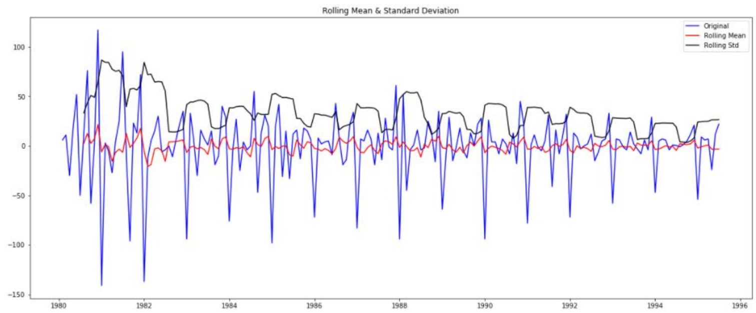
Null Hypothesis (H0) – The time series data is not stationary
Alternate Hypothesis (H1) – The time series data is stationary



Based on the above plot while standard deviation looks more or less constant, the rolling mean does not seem to be constant, there is to be a downward trend present. We can verify the same using the below t-statistics and p value.

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.873307
p-value                          0.344721
#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
```

The p value is 0.34 which is greater than 0.05. So at 95% confidence interval we fail to reject the null hypothesis, hence we can conclude that the data is not stationary.
Next we will take the difference of order 1 and check again if the data is stationary. Below is the t- statistic and p value for the difference of 1.



This is the plot of mean and std of difference of time series. Both mean and standard deviation looks constant, we can verify the same using t-stat and p-val.

```
Results of Dickey-Fuller Test:
Test Statistic                 -8.044136e+00
p-value                         1.813615e-12
#Lags Used                      1.200000e+01
Number of Observations Used     1.730000e+02
Critical Value (1%)            -3.468726e+00
Critical Value (5%)            -2.878396e+00
Critical Value (10%)           -2.575756e+00
```

The p-value is very low, lower than 0.05. So at this level we can with surety reject the null hypothesis and can conclude that at difference =1, the time series is now stationary.

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information

Auto Regressive Integrated Moving Average(ARIMA) explains a given time series based on its own past values and correlations between the values. It has 3 components- p, d and q, where p is the Auto regressive component, q is the moving average component and d is the order of difference taken between values to make the series stationary.

We have seen above that order of difference, i.e. d = 1. So to find p and q, we will run an iterative loop for values of p and q ranging between 0 and 3 and value of d as 1 and select the values which gives us the lowest Akaike Information Criteria (AIC). AIC is a mathematical method of evaluating how well the model fits the data, the lower the AIC and the better fit the model.

	param	AIC
5	(0, 1, 2)	1276.835372
11	(1, 1, 2)	1277.359233
10	(1, 1, 1)	1277.775749
16	(2, 1, 1)	1279.045689
17	(2, 1, 2)	1279.298694

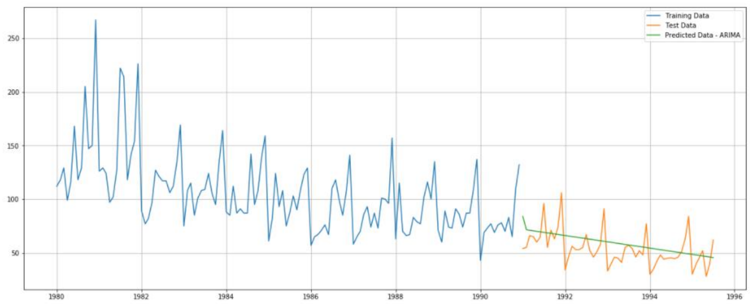
For the value of p,d,q as (0,1,2) we get the lowest AIC value, hence we will build the ARIMA model with these parameters.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -634.418
Method:                       css-mle   S.D. of innovations             30.167
Date:                Wed, 27 Jan 2021   AIC                           1276.835
Time:                        01:18:35   BIC                           1288.336
Sample:                    02-01-1980   HQIC                          1281.509
                         - 12-01-1990
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -0.4885      0.085     -5.742      0.000      -0.655      -0.322
ma.L1.D.Rose   -0.7601      0.101     -7.499      0.000      -0.959      -0.561
ma.L2.D.Rose   -0.2398      0.095     -2.518      0.012      -0.427      -0.053
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -4.1695           +0.0000j            4.1695            0.5000
-----------------------------------------------------------------------------
```

The above gives the summary of our ARIMA model, as p is 0 there is no auto regressive component. q= 2 hence we have 2 moving average components and from the plot we can ensure that p value is less than 0.05, hence both the ma components are significant further confirming that our model looks good.

Next we will calculate the RMSE of this model.

	Test RMSE
ARIMA(0,1,2)	15.625788



The green line denotes the ARIMA model predicted value. The line captures the trend beautifully, but it doesn't capture the seasonality, for that we have the SARIMA model.
Seasonal Auto Regressive Integrated Moving Average as the name suggests is ARIMA with the additional seasonality component. It has the ARIMA p,d and q, it also has P,Q , D and S components. S denotes the seasonality. To find the adequate paramaters, we will again be running the iterative model for all parameters value.

	param	seasonal	AIC
26	(0, 1, 2)	(2, 1, 2, 12)	774.969120
107	(0, 1, 2)	(2, 1, 2, 12)	774.969120
269	(0, 1, 2)	(2, 1, 2, 12)	774.969120
53	(1, 1, 2)	(2, 1, 2, 12)	776.940110
377	(1, 1, 2)	(2, 1, 2, 12)	776.940110
...	...	...	...
270	(1, 0, 0)	(0, 0, 0, 12)	1331.248484
163	(0, 0, 0)	(0, 0, 1, 12)	1342.887980
198	(0, 0, 2)	(0, 0, 0, 12)	1426.844550
180	(0, 0, 1)	(0, 0, 0, 12)	1481.819865
162	(0, 0, 0)	(0, 0, 0, 12)	1607.530754

486 rows × 3 columns

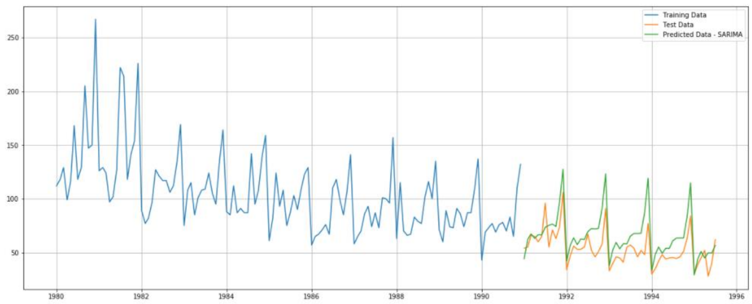
Here we ran iterations for all possible values and it gave us value of AIC for 486 different combinations, and for the (0,1,2) and (2,1,2,12) parameters and seasonal parameters we get the lowest AIC value of 774. Hence we will build our SARIMA model with these parameters.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -380.485
Date:                            Wed, 27 Jan 2021   AIC                            774.969
Time:                                    14:17:17   BIC                            792.622
Sample:                                01-01-1980   HQIC                           782.094
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9524      0.184     -5.168      0.000      -1.314      -0.591
ma.L2         -0.0764      0.126     -0.605      0.545      -0.324       0.171
ar.S.L12       0.0480      0.177      0.271      0.786      -0.299       0.394
ar.S.L24      -0.0419      0.028     -1.513      0.130      -0.096       0.012
ma.S.L12      -0.7526      0.301     -2.503      0.012      -1.342      -0.163
ma.S.L24      -0.0721      0.204     -0.354      0.723      -0.472       0.327
sigma2       187.8577     45.268      4.150      0.000      99.135     276.581
===================================================================================
Ljung-Box (Q):                       31.31   Jarque-Bera (JB):                 4.86
Prob(Q):                              0.84   Prob(JB):                         0.09
Heteroskedasticity (H):               0.91   Skew:                             0.41
Prob(H) (two-sided):                  0.79   Kurtosis:                         3.77
===================================================================================
```

The above gives the summary of our SARIMA model. Based on our p value we can infer that the 2 moving average , first and second seasonal auto regression and second seasonal moving average are not significant, hence we will drop them and run the model again for (0,1,1) X(0,1,1,12) to find the new AIC.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 12)   Log Likelihood                -454.537
Date:                            Wed, 27 Jan 2021   AIC                            915.073
Time:                                    14:53:05   BIC                            923.035
Sample:                                01-01-1980   HQIC                           918.299
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8930      0.047    -18.881      0.000      -0.986      -0.800
ma.S.L12      -0.6052      0.076     -8.010      0.000      -0.753      -0.457
sigma2       330.4814     46.689      7.078      0.000     238.973     421.990
===================================================================================
Ljung-Box (Q):                       34.34   Jarque-Bera (JB):                 0.17
Prob(Q):                              0.72   Prob(JB):                         0.92
Heteroskedasticity (H):               0.49   Skew:                             0.01
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.20
===================================================================================
```

After running the model again, we see that even though all the components are significant, the AIC has increased. Hence we will keep our original SARIMA model of parameters (0,1,2) X (2,1,2,12) and calculate the RMSE.

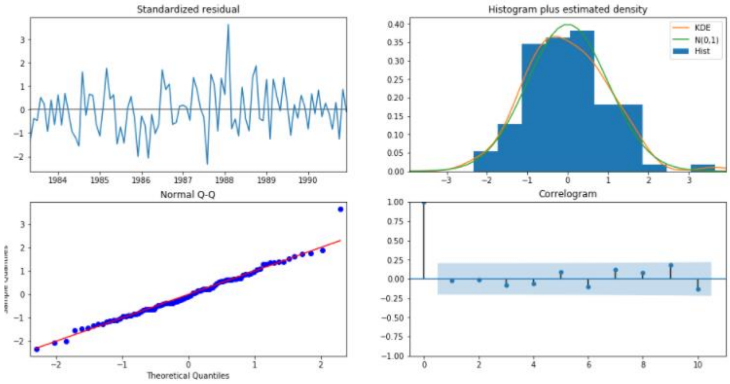


In the above plot the green line shows the SARIMA model predicted data, while the orange line shows the original test data. As we can see the trend was captured correctly using ARIMA, now SARIMA does capture the yearly seasonal component also, but the peaks are much higher than original series. Then we calculate the RMSE value.

	Test RMSE
SARIMA (0, 1, 2)x(2, 1, 2, 12)	16.523415

The RMSE for our SARIMA model is slightly higher than ARIMA, as we can see form the graph that the seasonality component was a extrapolated a bit too much.

Next we will plot the diagnostics of SARIMA.



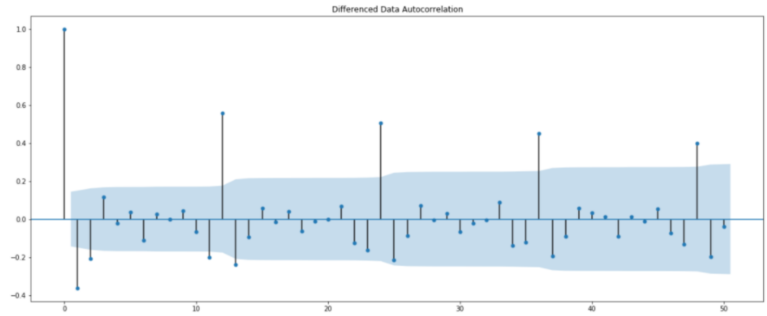
- The Standardized residual does not have any set pattern
- The Histogram is more or less normally distributed
- In the Normal Q-Q graph the blue dots lie on the red line
- In the Correlogram there is no data point outside the significance interval and values are all close to 0

Based on the above deductions we can conclude that this model is stable.

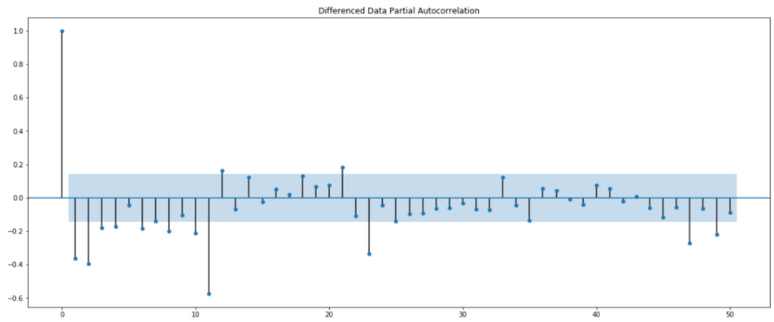
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data

Next we will create the auto correlation (ACF) and partial auto correlation (PACF) plots. The auto correlation plot gives us the correlations of the present value with the previous values at time steps or lags. The partial auto correlation removes the indirect correlation between time steps, i.e. removes the in linear correlations for the current and past observations and gives only the direct correlation. The PACF plot gives us the p value, i.e. the auto regressive value and ACF plot gives us q value, i.e. moving average.

PACF should ideally have a gradual drop over time and ACF should have a sharp and sudden drop.
We will be plotting the PACF and ACF plot for the rose time series differenced by 1.



Based on the above ACF curve, we can conclude the q value as 2 as after the first auto correlation, there are two values outside of the significance area. Looking at the plot, there could be a seasonality of 12.



Based on the above PACF plot, we can say that the p value could be 4, and difficult to comment on the seasonality.
So with the pdq value of (4,1,2) we will create ARIMA and SARIMA models.
ARIMA (4,1,2)

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -633.876
Method:                       css-mle   S.D. of innovations             29.793
Date:                Wed, 27 Jan 2021   AIC                           1283.753
Time:                        13:05:23   BIC                           1306.754
Sample:                    02-01-1980   HQIC                          1293.099
                         - 12-01-1990
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975
--------------------------------------------------------------------------------
const         -0.1905      0.576     -0.331      0.741      -1.319       0.93
ar.L1.D.Rose   1.1685      0.087     13.391      0.000       0.997       1.34
ar.L2.D.Rose  -0.3562      0.132     -2.693      0.007      -0.616      -0.09
ar.L3.D.Rose   0.1855      0.132      1.402      0.161      -0.074       0.44
ar.L4.D.Rose  -0.2227      0.091     -2.443      0.015      -0.401      -0.04
ma.L1.D.Rose  -1.9506        nan        nan        nan         nan         na
ma.L2.D.Rose   1.0000        nan        nan        nan         nan         na
                                    Roots
```

The first AR component and 4th AR component do not seem to be significant. RMSE appears very high, not a great model.

	Test RMSE
ARIMA (4, 1, 2)ACF, PACF	33.972477

SARIMA (4,1,2)X(4,1,2)X12

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(4, 1, 2)x(4, 1, 2, 12)   Log Likelihood                -277.661
Date:                            Wed, 27 Jan 2021   AIC                            581.322
Time:                                    13:28:19   BIC                            609.983
Sample:                                01-01-1980   HQIC                           592.663
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9746      0.199     -4.901      0.000      -1.364      -0.585
ar.L2         -0.1126      0.285     -0.395      0.693      -0.671       0.446
ar.L3         -0.1043      0.277     -0.377      0.706      -0.647       0.439
ar.L4         -0.1283      0.162     -0.792      0.428      -0.446       0.189
ma.L1          0.1605     82.712      0.002      0.998    -161.953     162.274
ma.L2         -0.8395     69.457     -0.012      0.990    -136.974     135.294
ar.S.L12      -0.1453      0.364     -0.399      0.690      -0.859       0.568
ar.S.L24      -0.3593      0.227     -1.585      0.113      -0.804       0.085
ar.S.L36      -0.2152      0.106     -2.039      0.041      -0.422      -0.008
ar.S.L48      -0.1196      0.093     -1.282      0.200      -0.302       0.063
ma.S.L12      -0.5148      0.343     -1.500      0.134      -1.188       0.158
ma.S.L24       0.2072      0.373      0.555      0.579      -0.525       0.939
sigma2       215.3813   1.78e+04      0.012      0.990   -3.47e+04    3.52e+04
```

SARIMA with p and q value from ACF and PACF plots gives the lowest AIC, though most of the components are greater than 0.05, hence insignificant. Let us calculate RMSE to check if this model is good.

	Test RMSE
SARIMA (4, 1, 2)X(4,1,2) 12,ACF, PACF	17.534342

Test RMSE, not too bad but we got a lower RMSE in SARIMA (0,1,2)X(2,1,2,12)

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
2 point Moving Average	11.529811
4 point Moving Average	14.457115
6 point Moving Average	14.571789
9 point Moving Average	14.731914
ARIMA(0,1,2)	15.625788
TES Additive	16.475042
SARIMA (0, 1, 2)x(2, 1, 2, 12)	16.523415
SARIMA (4, 1, 2)X(4,1,2) 12,ACF, PACF	17.534342
Regression	22.573067
TES Multiplicative	25.210414
ARIMA (4, 1, 2)ACF, PACF	33.972477
Simple Average Model	53.483727
DES	70.599377
Naive Model	79.741326

Moving averages give the lowest RMSE, but due to less stability and slight over fitting nature of moving average model, we will not consider it as our final model.

The final model we will select based on the RMSE value would be ARIMA (0, 1, 2) as with SARIMA our model is not able to clearly explain the seasonality component in our rose time series data.

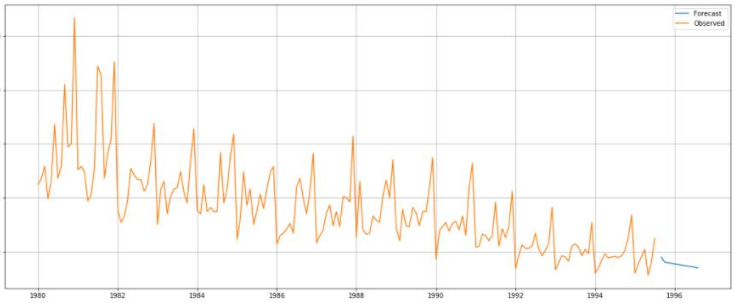
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

We will now be building the ARIMA (0, 1, 2) model on our entire rose data and predict the next 12 months. Below is the summary.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  186
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -876.963
Method:                       css-mle   S.D. of innovations             26.650
Date:                Wed, 27 Jan 2021   AIC                           1761.927
Time:                        16:27:24   BIC                           1774.830
Sample:                    02-01-1980   HQIC                          1767.156
                         - 07-01-1995
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.5231      0.043    -12.030      0.000      -0.608      -0.438
ma.L1.D.Rose    -0.7923      0.082     -9.617      0.000      -0.954      -0.631
ma.L2.D.Rose    -0.2076      0.081     -2.572      0.010      -0.366      -0.049
                                    Roots
```

The AIC of our final model = 1761. All components are significant. RMSE value = 101.48

Below is the forecasted plot for next 12 months.



The forecasted value is from 1995-08 to 1996-08

10. Insights, Findings and Recommendations

Following are the insights, findings and recommendations:

- There is a very clear downward trend from 1980 to 1995.
- There is seasonality present in our dataset, we have peaks in December followed by November, and the least in January. We can infer that people prefer to have rose wine around Christmas time, as rose wine is considered as dessert wine.
- The lowest sale of rose wine is in January year after year. This could be due to some New Year resolutions and the end of vacation. Would recommend having marketing campaigns, discounts and offers in the month of Jan to increase consumption.
- From 1980 to 1982 the peaks of rose wine consumption in Nov/Dec was very high, but the peaks dropped significantly after that. Even though there is a overall drop in consumption, there seems to a huge drop in peaks consumption. Need to study further what happened in 1982, it could be due to any following factors:
- Competitor Launch
- Increased price
- Poor supply chain management
- Drop in brand equity
- The seasonality, i.e. high in December and low in January was high in earlier years, and the seasonality 1992 onwards is reducing and the series is becoming less volatile.

- With the reduction of seasonality have built a model without seasonality for predicted of next 12 months. With reduced volatility a linear line would give us higher accuracy.
- Even though December has higher consumption in recent years (1992 onwards) as compared to other months, over the years sales in December are reducing. Would suggest conducting a market research on Christmas wine/alcohol preferences and identify competition to create a better positioning strategy.