Project report - Time Series Forecasting

By Prakash v Mahadole

Problem Statement:

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.

For Sparkling Dataset

- Importing all the necessary libraries
- Loading data in Data Frame ### 1. Read the data as an appropriate Time Series data and plot the data. ### Checking head of the data

```
Out[3]: YearMonth
        1980-01-01    1686
        1980-02-01    1591
        1980-03-01    2304
        1980-04-01    1712
        1980-05-01    1471
        Name: Sparkling, dtype: int64
```

Checking tail of the data

```
Out[4]: YearMonth
        1995-03-01    1897
        1995-04-01    1862
        1995-05-01    1670
        1995-06-01    1688
        1995-07-01    2031
        Name: Sparkling, dtype: int64
```


For method 2: Creating the Time Stamps and adding to the data frame to make it a Time Series Data

```
ut[6]: DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
                      '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
                      '1980-09-30', '1980-10-31',
                      ...
                      '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
                      '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
                      '1995-06-30', '1995-07-31'],
                     dtype='datetime64[ns]', length=187, freq='M')
```

Checking head of the data

Out[9]:

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Plot the Time Series to understand the behaviour of the data

- Monthly sale of sparkling wine is given for a period from January, 1980 to July, 1995
- The given data files are read as is and a date-range has been applied on the data as index
- The dataset shows significant seasonality, doesn't shows any consistent trend but has upward and downward slopes during the time period

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

Check the basic measures of descriptive statistics

	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

- The descriptive summary of the data shows that on an average 2402 units of Sparkling wines were sold each month on the given period of time. 50% of months sales varied from 1605 units to 2549 units. Maximum sale reported in a month is 7242 units.

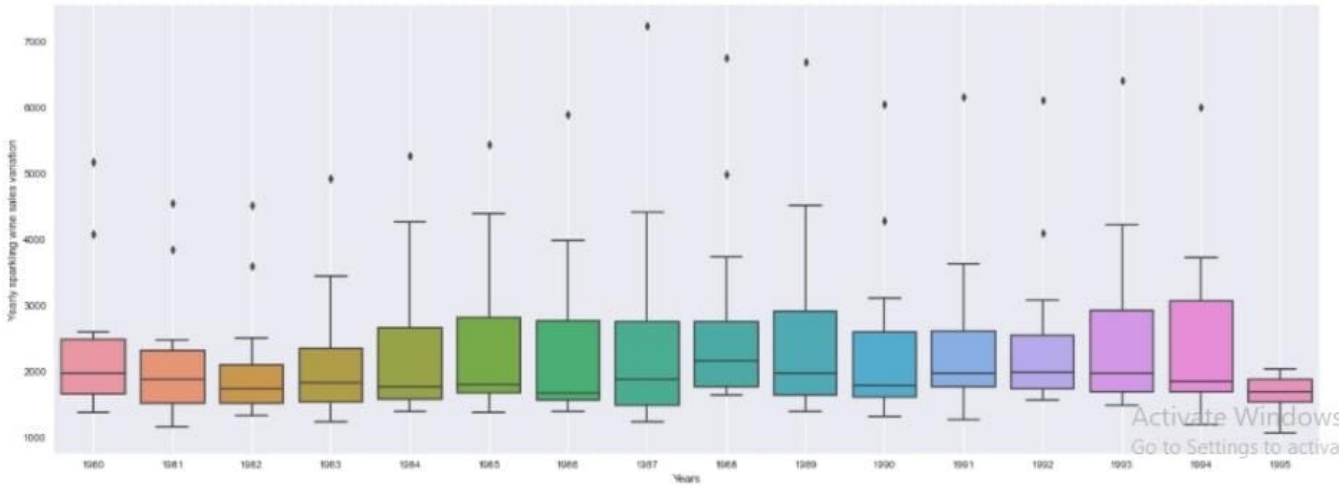
Checking for missing values

Sparkling 0

dtype: int64

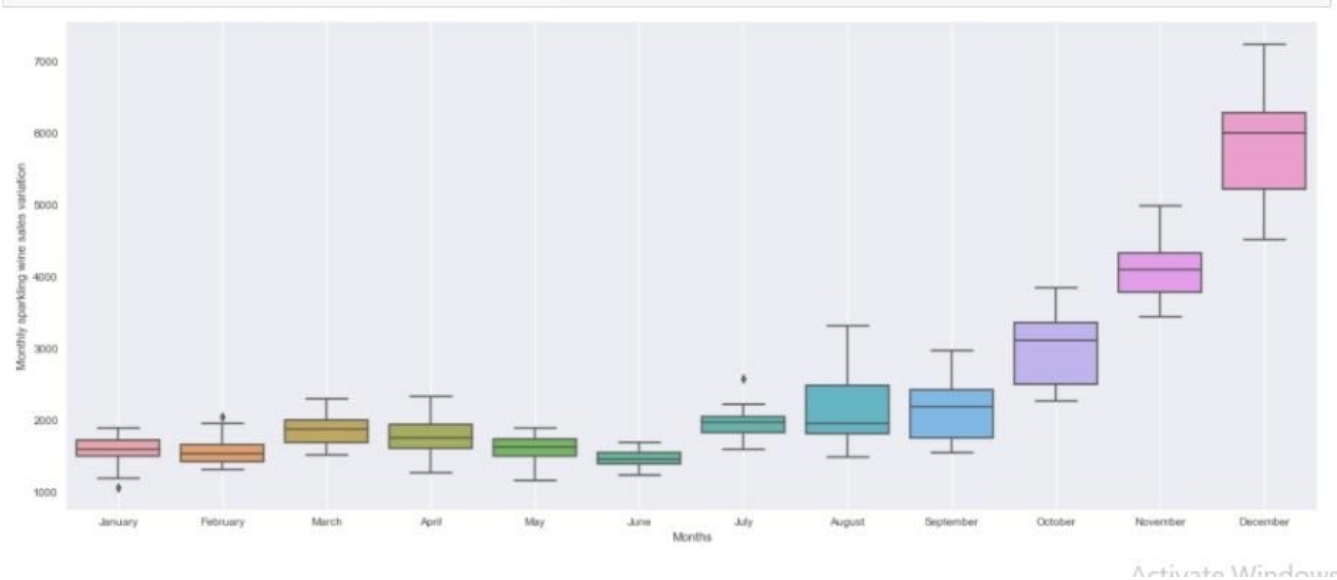
Plot a boxplot to understand the spread of sales across different years and within different months across years.

Yearly Boxplot



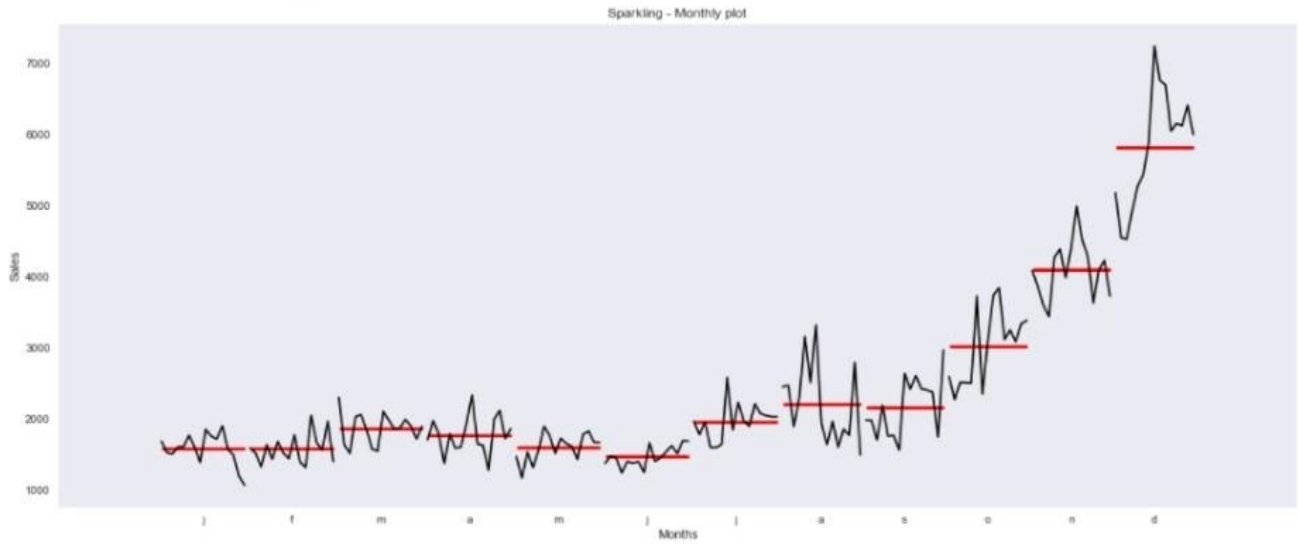
- The yearly-boxplot, shows that the average sale of Sparkling has been more or less consistent across the period, at or a little below 2000 units.

Monthly plot



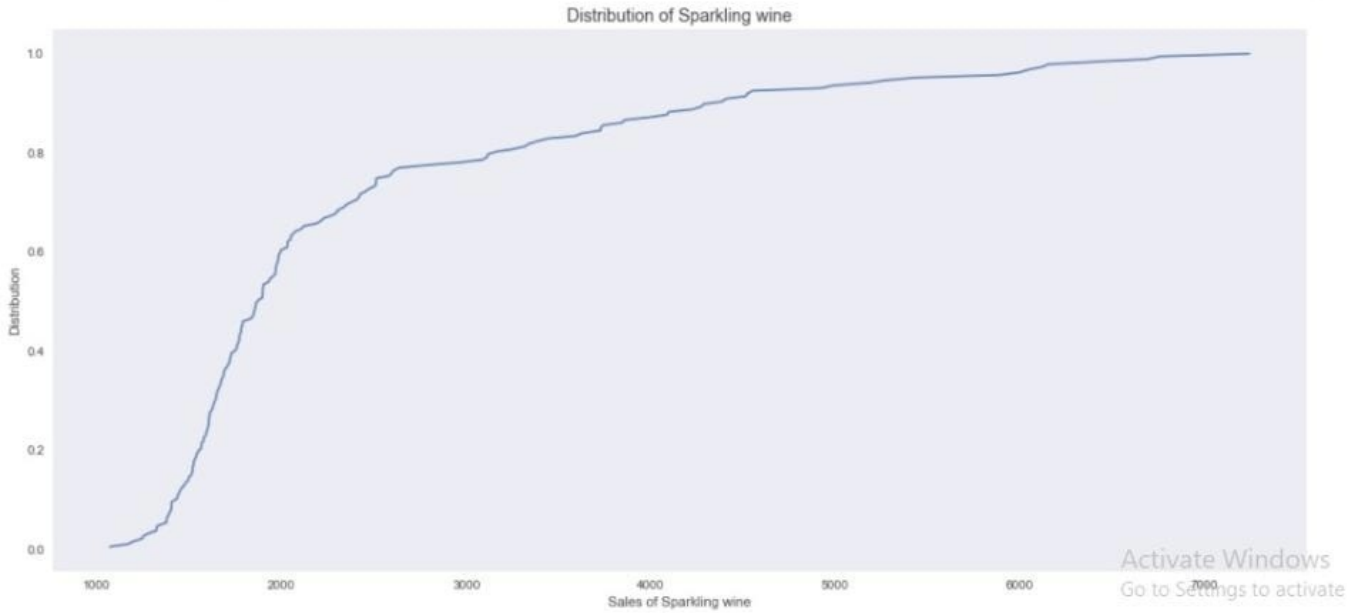
- The monthly-box-plot shows a clear seasonality during the festive seasonal months of October, November and December, which peaks in December. The sale tanks in the month of June.
- Sale in December with a mean few points below 6000, varies from 7400 to 4500 units over the years. Whereas sale in November varies from 3500 units to 5000 units and sale in October varies from 2500 to 4000 units
- The lean months from January till September shows more or less a consistent sale around 2000 units

Plot a time series monthplot to understand the spread of sales across different years and within different months across years.



- seasonality component of the time-series, with October November and December selling exponentially higher volumes.
- Sales for the months from January to July is seen to be consistent across the years, compared to the rest of the months

Plot the Empirical Cumulative Distribution.

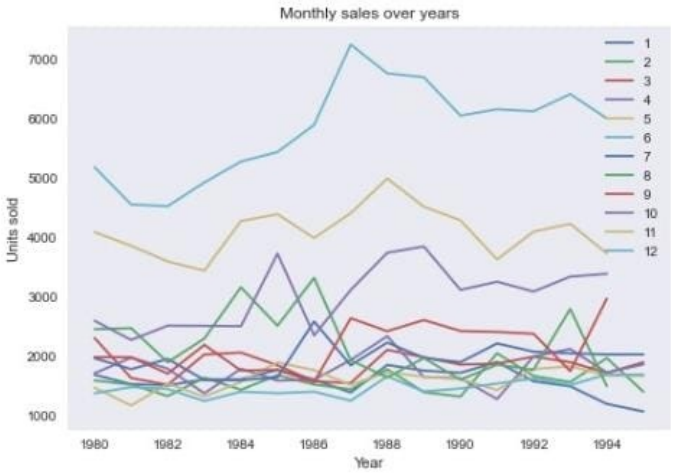


- The Empirical CDF plot shows that, in 80% of months, at least 3000 units of Sparkling wine were sold

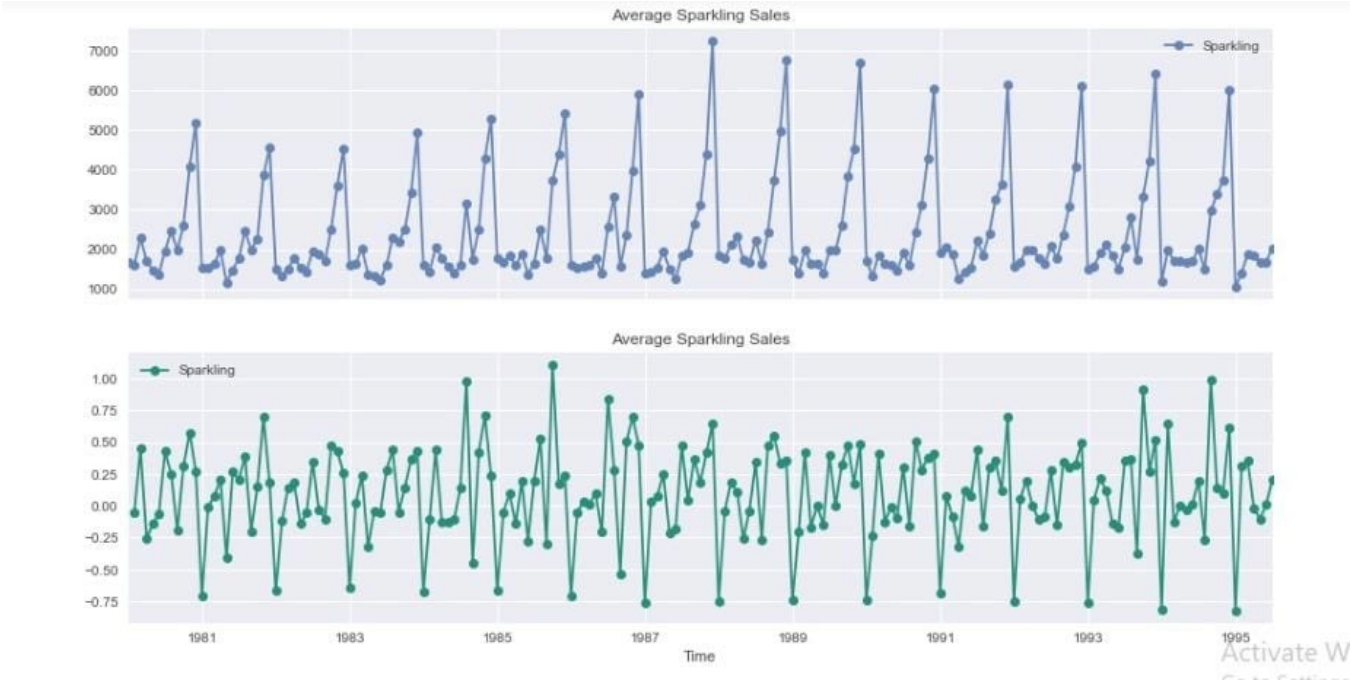
Plot a graph of monthly sales across years.

[18]:

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

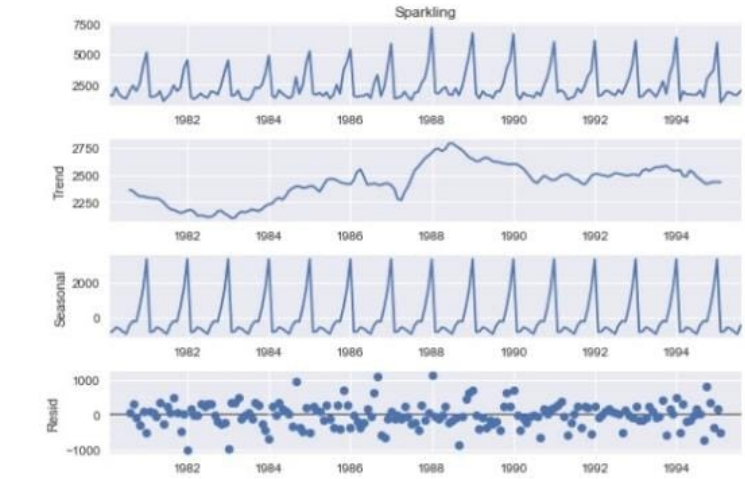


Plot the average sales per month and the month on month percentage change of sales.



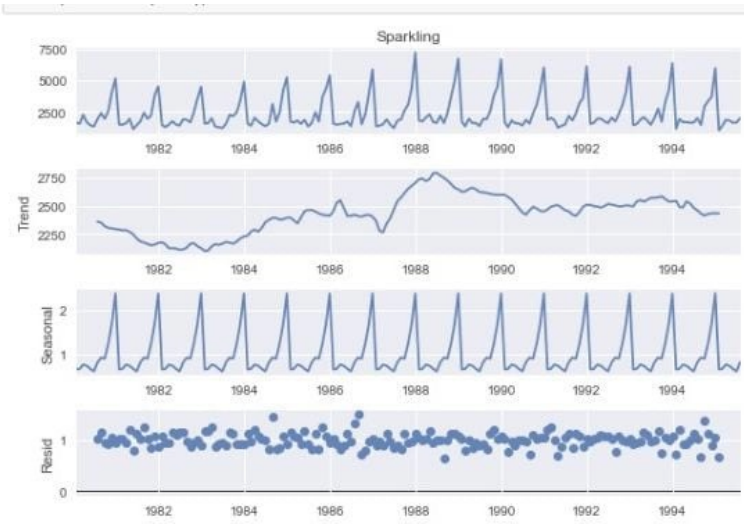
Decompose the Time Series and plot the different components.

Additive Decomposition



-

Multiplicative Decomposition



- As the altitude of the seasonal peaks in the observed plot is changing according to the change in trend, the time-series is assumed to be ‘multiplicative’.
- The plot of the trend component does not show a consistent trend, but an intermediary period shows an upward slope which gets consistent on the late half of time-series
- The additive model shows the seasonality with a variance of 3000 units and the multiplicative model shows a variance of 30%
- The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 10%
- If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series

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

- The train and test datasets are created with year 1991 asstarting year for test data, using index.year property of time series index

First few rows of Training Data

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Last few rows of Training Data

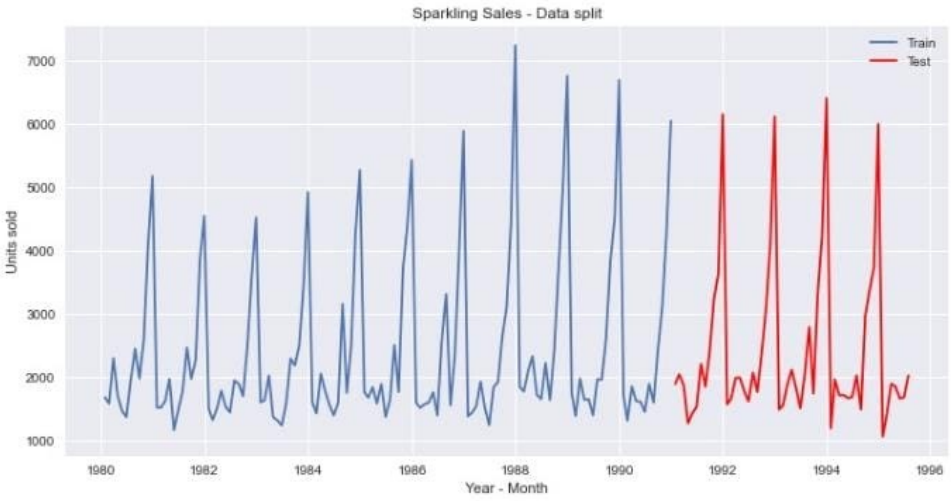
Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

First few rows of Test Data

Time_Stamp	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Last few rows of Test Data

Time_Stamp	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

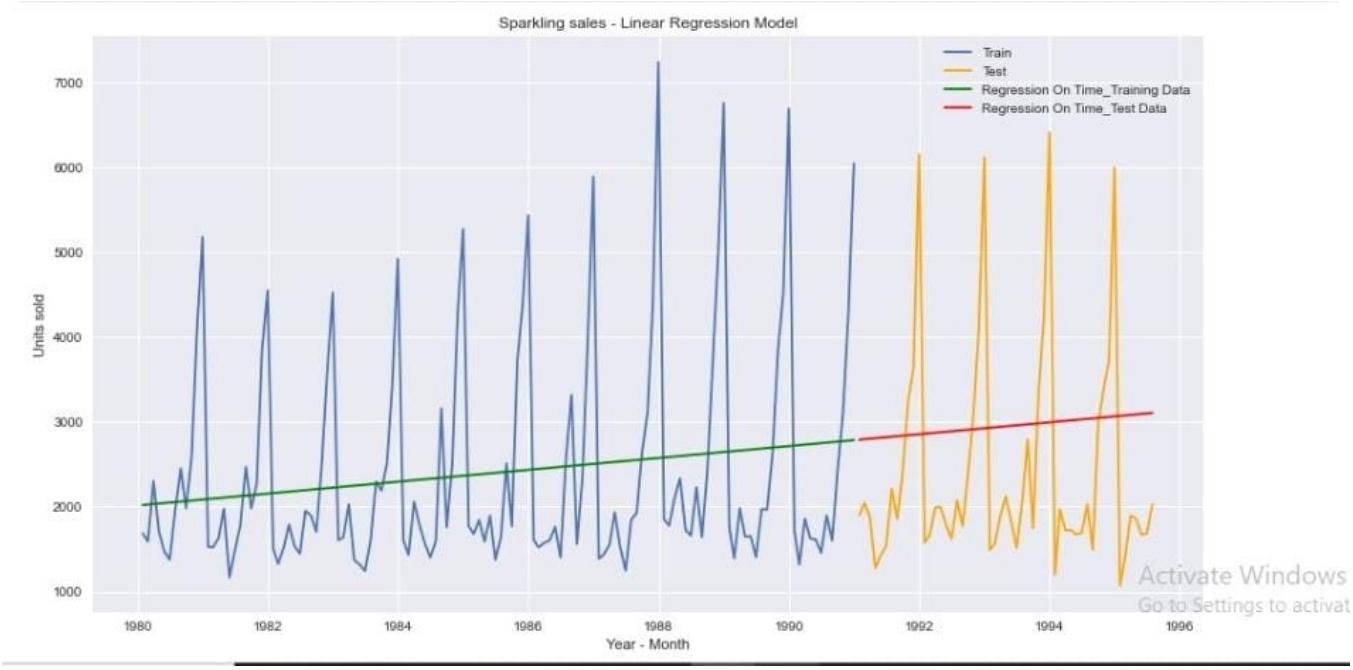


4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression

Time instances for train and test

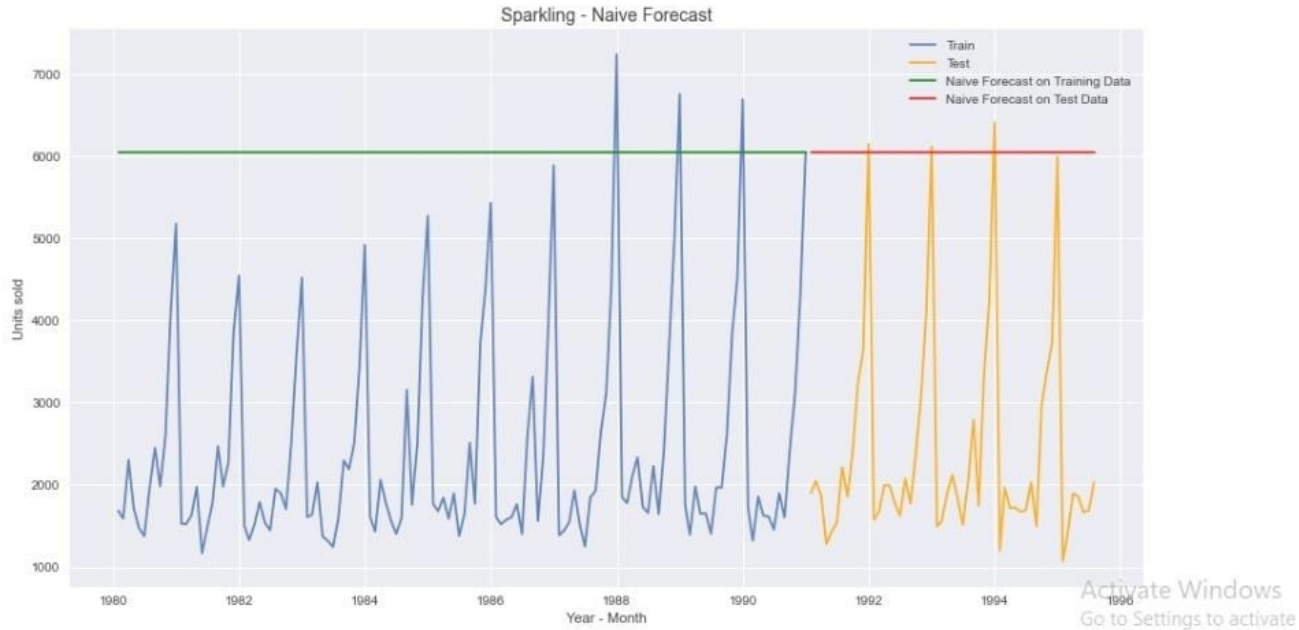
```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```



For RegressionOnTime forecast on the Sparkling Training Data: RMSE is 1279.322 and MAPE is 40.05

For RegressionOnTime forecast on the Sparkling Testing Data: RMSE is 1389.135 and MAPE is 50.15

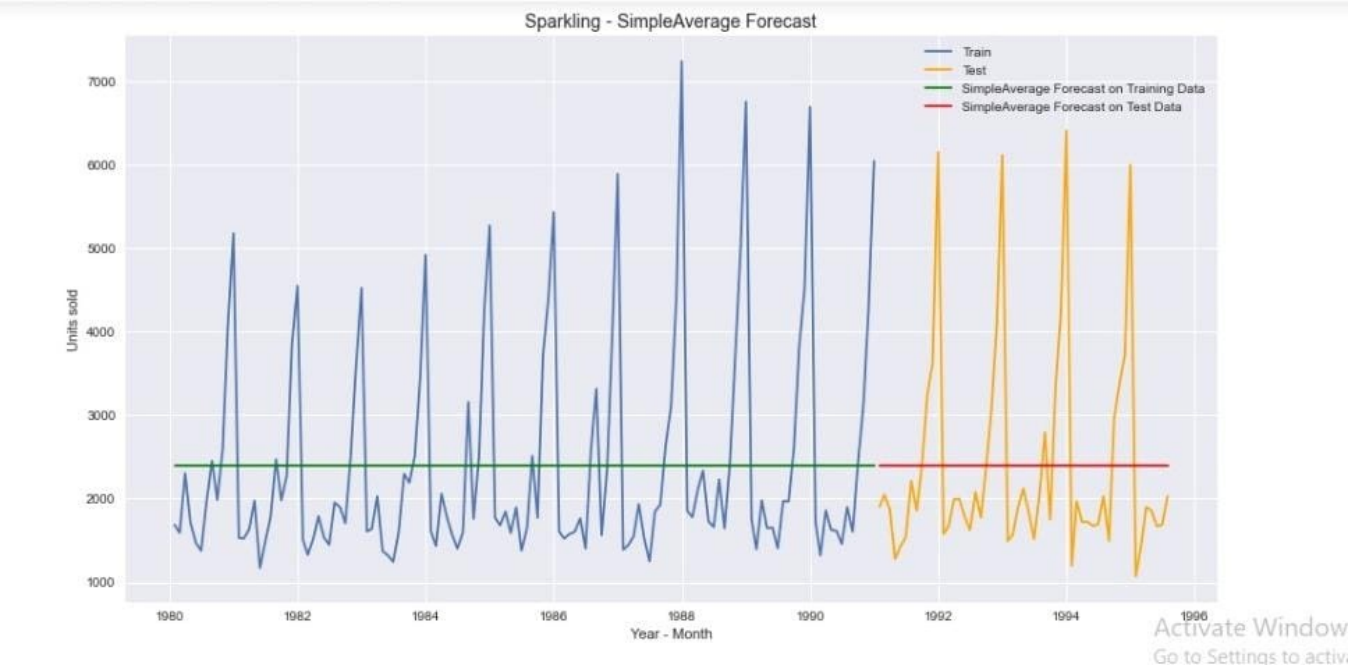
Model 2: Naive Approach



For Naive forecast on the Sparkling Training Data: RMSE is 3867.701 and MAPE is 153.17

For Naive forecast on the Sparkling Testing Data: RMSE is 3864.279 and MAPE is 152.87

Model 3: Simple Average



For Simple Average forecast on the Sparkling Training Data: RMSE is 1298.484 and MAPE is 40.36

For Simple Average forecast on the Sparkling Testing Data: RMSE is 1275.082 and MAPE is 38.90

Model 4: Moving Average

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

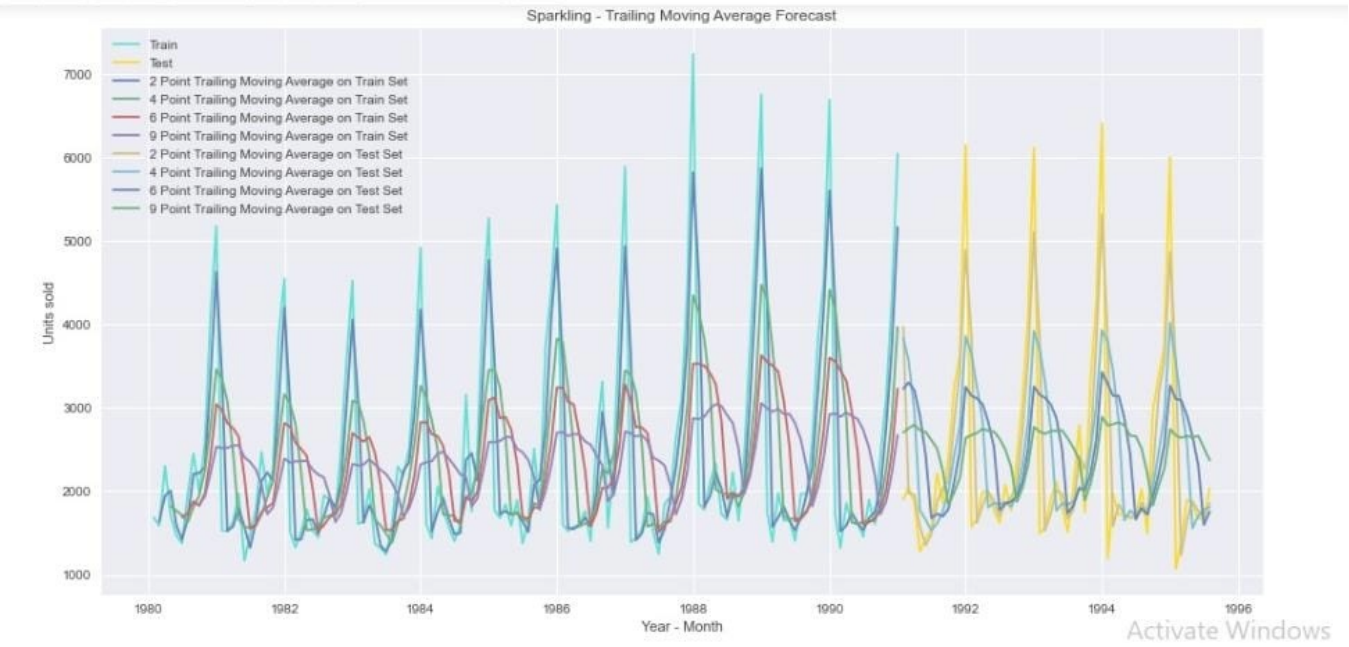
For Moving Average, we are going to average over the entire data.

- The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points ##### Trailing moving averages

ut[58]:

Time_Stamp	Sparkling	Spark_Trailing_2	Spark_Trailing_4	Spark_Trailing_6	Spark_Trailing_9
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN

Trailing Moving Average Forecast



For 2 point Moving Average Model forecast on the Training Data, rmse_spark is 813.401 mape_spark is 19.70

For 4 point Moving Average Model forecast on the Training Data, rmse_spark is 1156.590 mape_spark is 35.96

For 6 point Moving Average Model forecast on the Training Data, rmse_spark is 1283.927 mape_spark is 43.86

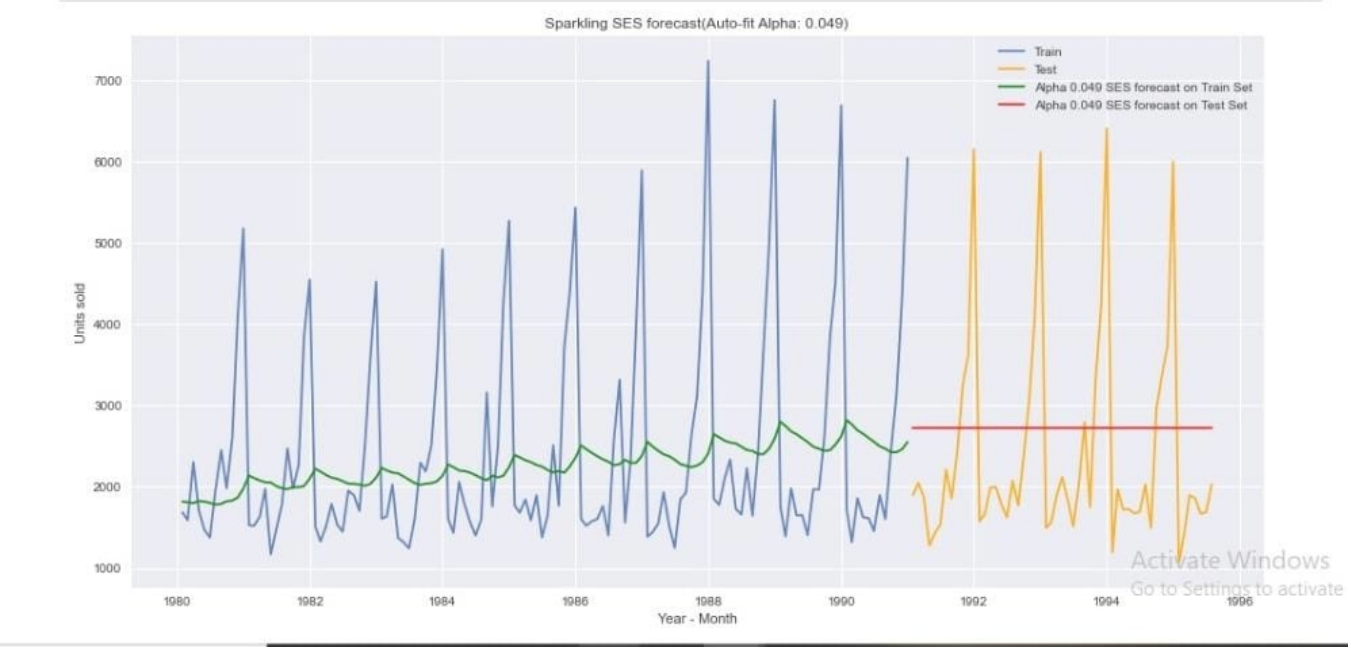
For 9 point Moving Average Model forecast on the Training Data, rmse_spark is 1346.278 mape_spark is 46.86

Model 5: Simple Exponential Smoothing

The Autofit parameters are found to be:

```
Out[68]: {'smoothing_level': 0.049607360581862936,
          'smoothing_trend': nan,
          'smoothing_seasonal': nan,
          'damping_trend': nan,
          'initial_level': 1818.535750008871,
          'initial_trend': nan,
          'initial_seasons': array([], dtype=float64),
          'use_boxcox': False,
          'lamda': None,
          'remove_bias': False}
```

SES forecast(Auto-fit Alpha: 0.049)'



For SES forecast on the Sparkling Training Data: RMSE is 1315.232 and MAPE is 39.92

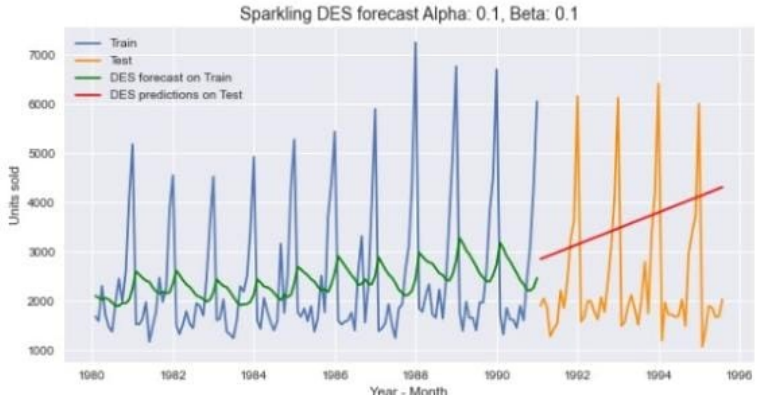
For SES forecast on the Sparkling Testing Data: RMSE is 1316.035 and MAPE is 45.47

Model 6: Double Exponential Smoothing (Holt's Model)

Out[78]:

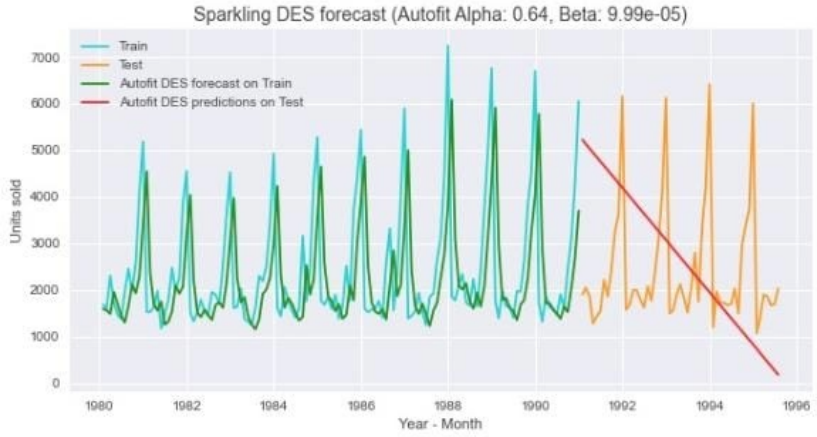
	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	1363.47	44.26	1779.42	67.23
1	0.1	0.2	1398.19	45.61	2601.54	95.50
10	0.2	0.1	1412.03	46.62	3611.77	135.41

DES forecast Alpha: 0.1, Beta: 0.1



- smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1 and beta 0.1

DES forecast (Autofit Alpha: 0.64, Beta: 9.99e-05



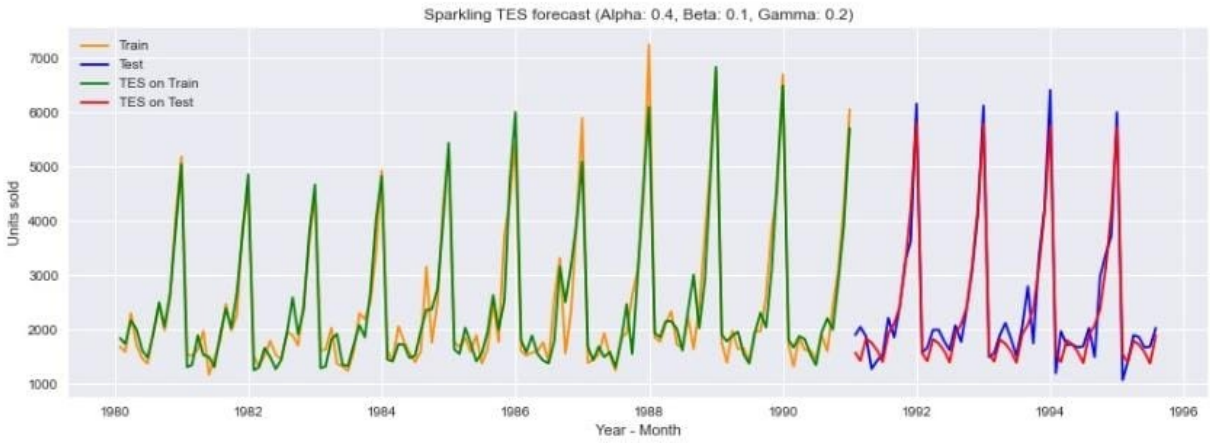
	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.1000	1363.47000	44.26	1779.420000	67.23
100	0.688571	0.0001	1349.65046	39.23	2007.238526	68.23
1	0.100000	0.2000	1398.19000	45.61	2601.540000	95.50
10	0.200000	0.1000	1412.03000	46.62	3611.770000	135.41
2	0.100000	0.3000	1431.37000	46.90	4288.430000	155.25

- The autofit model retuned higher accuracy in train dataset, but faired poorly in test, compared with the values in manual iteratio
- The model evaluation parameters of top three models from manual iteration and the autofit models are as given above
- The best model chosen as final one is with alpha 0.1 and beta 0.1

Model 7: Triple Exponential Smoothing (Holt - Winter's Model)

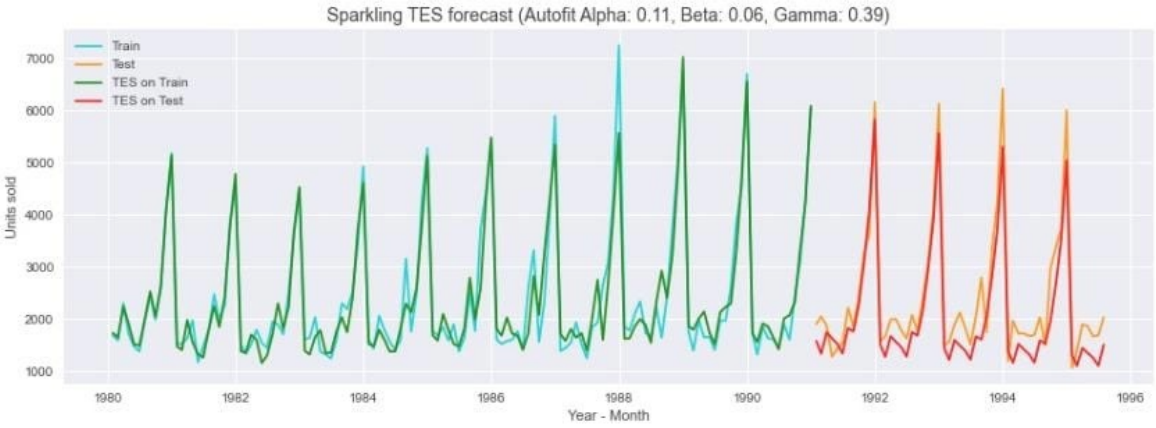
Three parameters α , β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

TES forecast (Alpha: 0.4, Beta: 0.1, Gamma: 0.2)



- The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Sparkling data contain slight trend and significant seasonality
- In first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.4, beta 0.1 and gamma 0.2

TES forecast (Autofit Alpha: 0.11, Beta: 0.06, Gamma: 0.39)



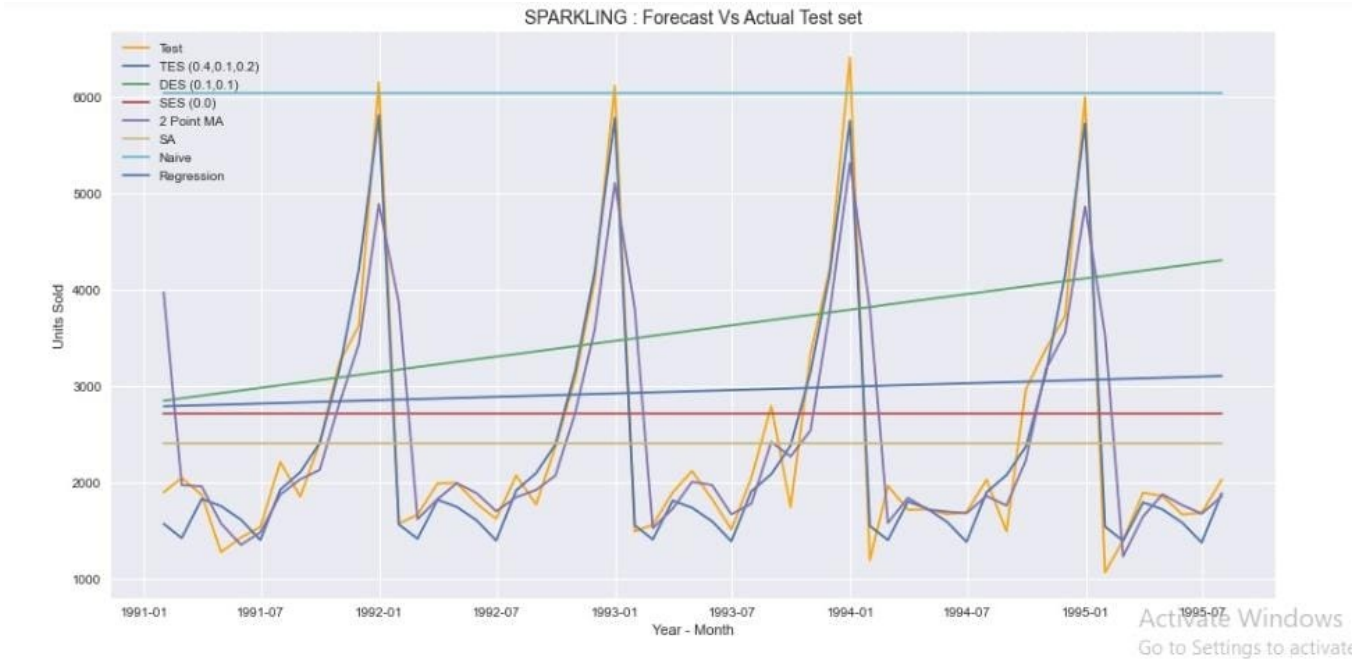
Out[102]:

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
211	0.3	0.2	0.2	378.189776	11.28	314.882349	10.10
301	0.4	0.1	0.2	373.815525	11.13	315.533374	10.41
300	0.4	0.1	0.1	371.341930	11.14	318.045761	10.24
402	0.5	0.1	0.3	390.175608	11.54	325.545203	9.99
30	0.1	0.4	0.1	403.937167	11.72	330.772119	10.56

- The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test
- The model evaluation parameters of the best models are given as above, including one from the autofit iteration
- The best model chosen as final one is the one with alpha 0.4, beta 0.1 and gamma 0.2

MODEL COMPARISON

Ploting all the above models



We can see all the models RMSE and MAPE values:

ut[104]:

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.533374	10.41
TES Alpha 0.11, Beta 0.06, Gamma 0.39	469.767970	16.40
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.049	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 9.99e-05	2007.238526	68.23
NaiveModel	3864.279352	152.87

- The accuracy of the time-series forecast models build in the previous sections of this report is as below, sorted by RMSE in test data
- The plot of the forecasts fitted on to the test data is given as well
- From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data
- 2 point trailing moving average model is also found to have fit well with a slight lag in test dataset

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.

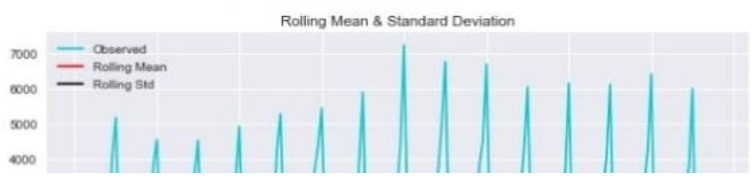
Note: Stationarity should be checked at alpha = 0.05.

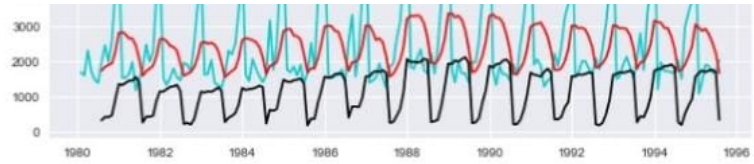
The Augmented Dicky-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

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

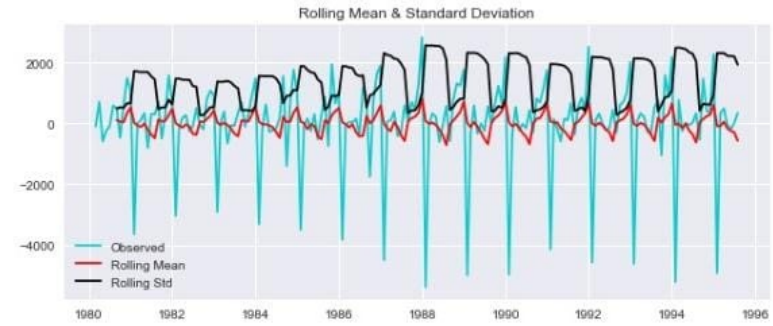
- H_0 : The Time Series has a unit root and is thus non-stationary.
- H_1 : The Time Series does not have a unit root and is thus stationary.

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.





```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```


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


We see that at 5% significant level the Time Series is non-stationary. But the seasonality is multiplicative as the Std deviation and mean varies according to the change in trend

- The ADF test is also done in this the train data and differencing of seasonal order (12), to understand if removing the multiplicity of the seasonal component will have an impact on the accuracy of model

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.

AUTO SARIMA on original data

As the data contains seasonality component we will be building SARIMA model, rather than ARIMA.

AIC values :

Out[121]:

	param	seasonal	AIC
47	(0, 1, 2)	(3, 1, 3, 12)	18.000000
187	(2, 1, 3)	(2, 1, 3, 12)	22.000000
227	(3, 1, 2)	(0, 1, 3, 12)	324.088688
163	(2, 1, 2)	(0, 1, 3, 12)	610.220194
147	(2, 1, 1)	(0, 1, 3, 12)	708.967435
252	(3, 1, 3)	(3, 1, 0, 12)	1213.282556
253	(3, 1, 3)	(3, 1, 1, 12)	1215.213499
220	(3, 1, 1)	(3, 1, 0, 12)	1215.898777
254	(3, 1, 3)	(3, 1, 2, 12)	1216.479984
236	(3, 1, 2)	(3, 1, 0, 12)	1216.859180

Model Summary – SARIMA (3, 1, 3)x(3, 1, 0, 12)

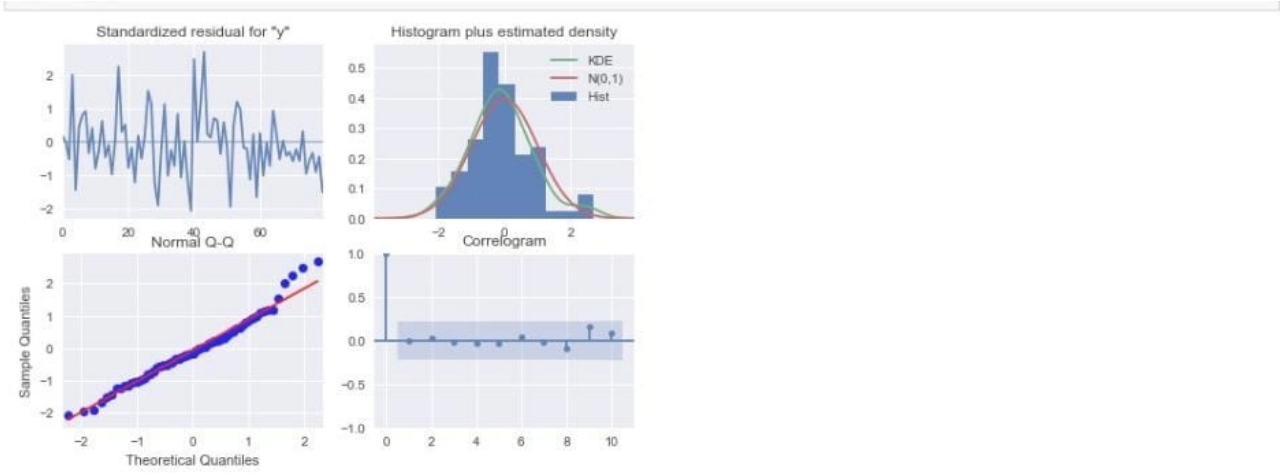
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 3)x(3, 1, [], 12)	Log Likelihood	-596.641
Date:	Fri, 08 Oct 2021	AIC	1213.283
Time:	19:14:30	BIC	1237.103
Sample:	0	HQIC	1222.833
	- 132		
Covariance Type:	opg		

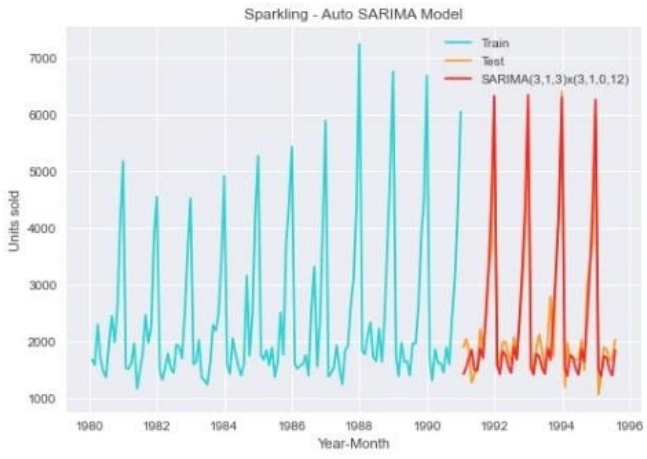
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6142	0.176	-9.177	0.000	-1.959	-1.269
ar.L2	-0.6124	0.299	-2.048	0.041	-1.199	-0.026
ar.L3	0.0861	0.161	0.536	0.592	-0.229	0.401
ma.L1	0.9855	0.472	2.089	0.037	0.061	1.910
ma.L2	-0.8738	0.166	-5.269	0.000	-1.199	-0.549
ma.L3	-0.9466	0.489	-1.936	0.053	-1.905	0.012
ar.S.L12	-0.4519	0.142	-3.191	0.001	-0.729	-0.174
ar.S.L24	-0.2341	0.144	-1.622	0.105	-0.517	0.049
ar.S.L36	-0.1008	0.122	-0.830	0.407	-0.339	0.137
sigma2	1.839e+05	8.97e+04	2.051	0.040	8136.302	3.6e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	4.06
Prob(Q):	0.93	Prob(JB):	0.13
Heteroskedasticity (H):	0.73	Skew:	0.48
Prob(H) (two-sided):	0.42	Kurtosis:	3.54

Diagnostics plot – SARIMA (3, 1, 3)x(3, 1, 0, 12)



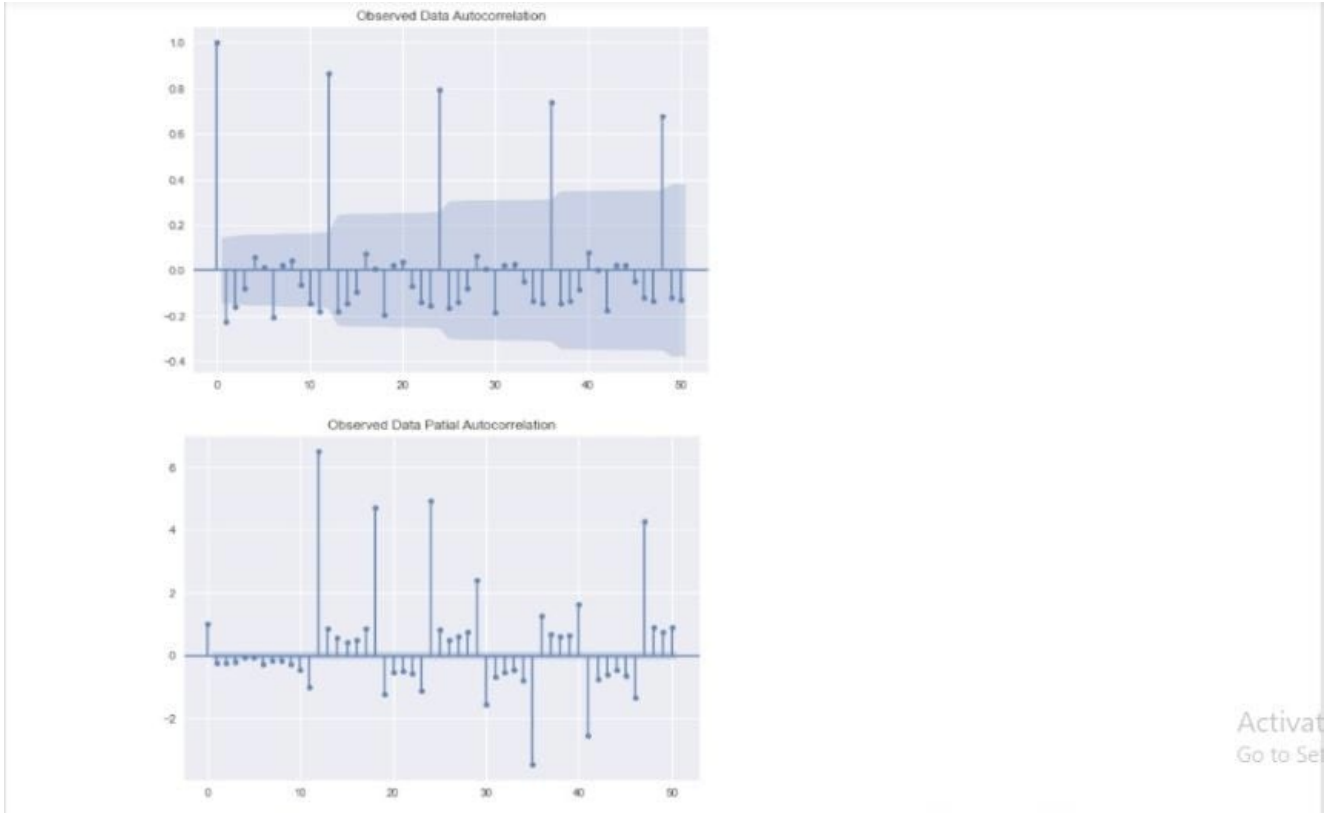
Auto SARIMA Model



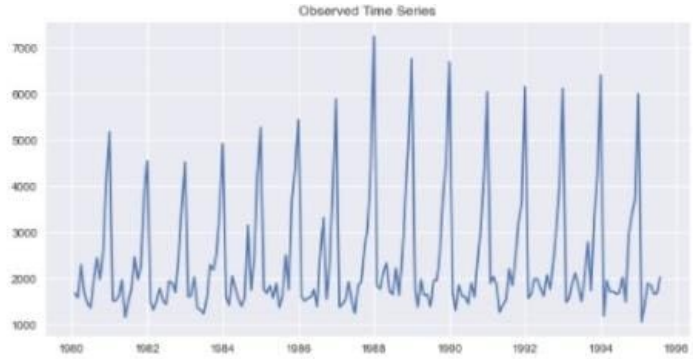
- The model built with original data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model
- As per the AIC criteria, the optimum values for final SARIMA model selected is (3, 1, 3)x(3, 1, 0, 12)
- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution
- For SARIMA forecast on the Sparkling Testing Data: RMSE is 331.642 and MAPE is 10.34
- From the p-values it can be inferred that terms AR(1), AR(2), MA(1), MA(2), MA(3) and seasonal AR(1) are significant terms, as their values are below 0.05

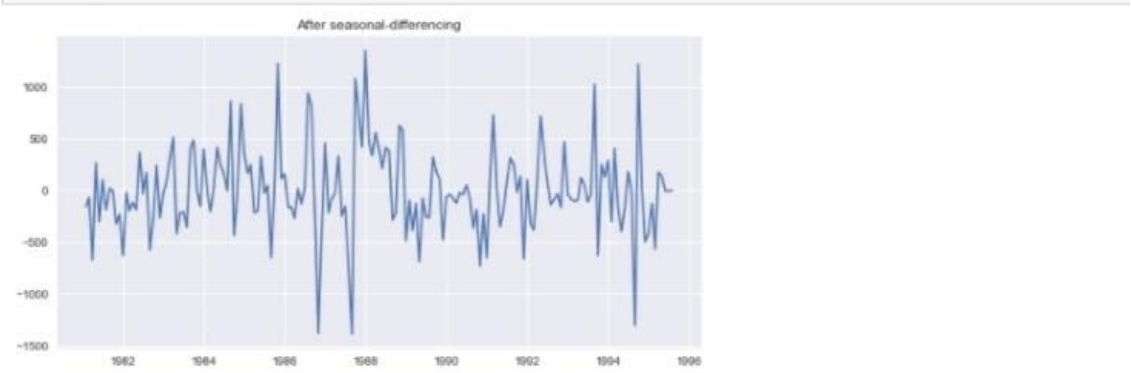
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

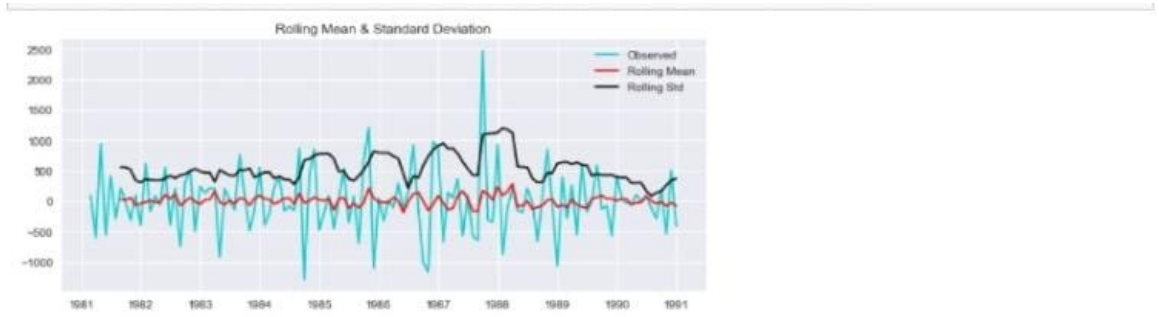
Let us look at the ACF and the PACF plots .



We see that our ACF plot at the seasonal interval (12) does not taper off quickly. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.

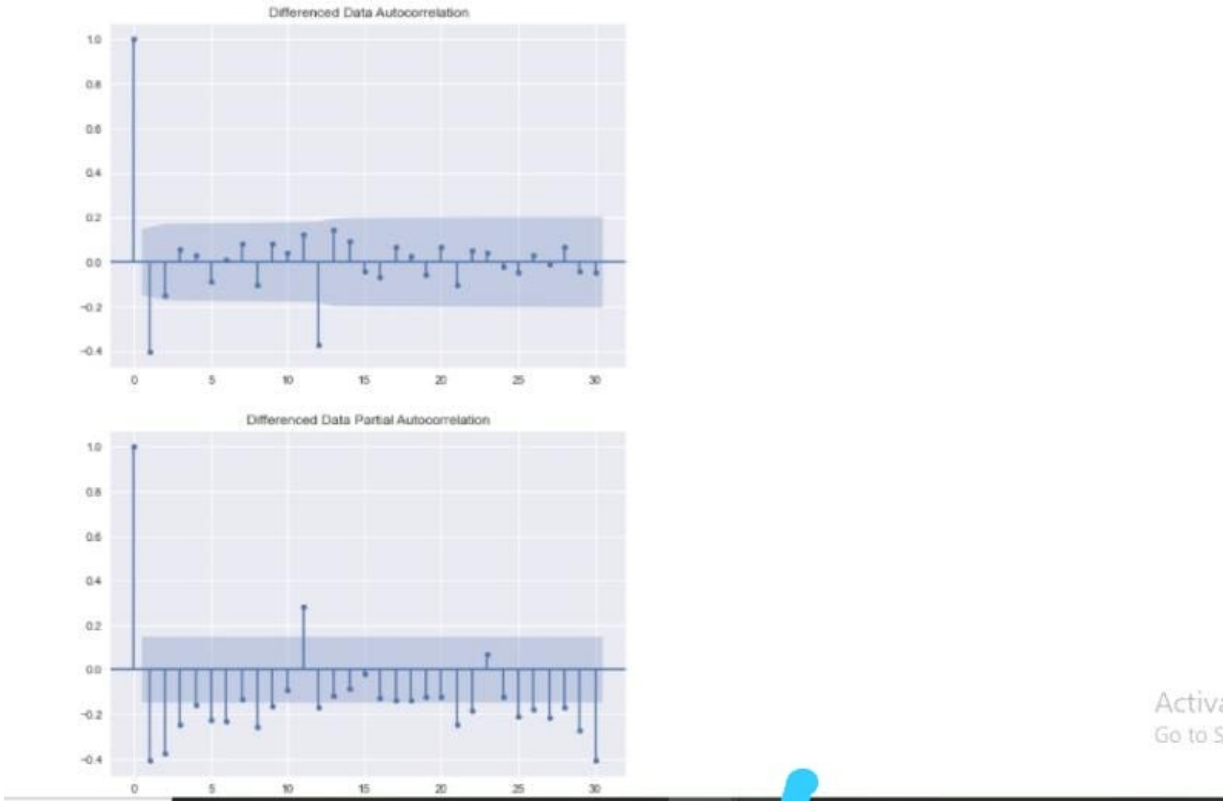







```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
#Lags Used                      10.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

- An ADF test need to be done to check the stationarity after the above differencing. With ap-value below alpha 0.05 and test statistic below critical values, it can be confirmed that thedata is stationary

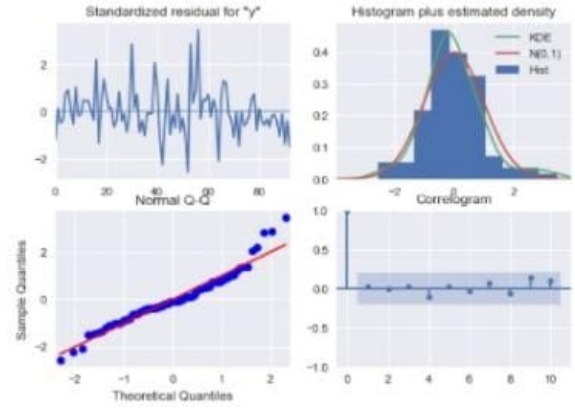


- ACF and PACF plots of the seasonal-differenced + one order differenced data is created to find the values for (p,d,q)x(P,D,Q),
- Here we have taken alpha = 0.05 and seasonal period as 12
- From the PACF plot it can be seen that till 3rd lag its significant before cut-off, so AR term ‘p = 3’ is chosen. At seasonal lag of 12, it almost cuts off, so seasonal AR ‘P = 1’
- From ACF plot it can be seen that lag 1 is significant before it cuts off, so MA term ‘q = 1’ is selected and at seasonal lag of 12, a significant lag is apparent, so kept seasonal MA term ‘Q = 1’ initially

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(1, 1, [1, 2], 12)   Log Likelihood             -693.697
Date:                                 Fri, 08 Oct 2021   AIC                        1403.394
Time:                                         19:14:40   BIC                        1423.654
Sample:                                              0   HQIC                       1411.574
                                                   - 132
Covariance Type:                                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2229      0.130      1.713      0.087      -0.032       0.478
```

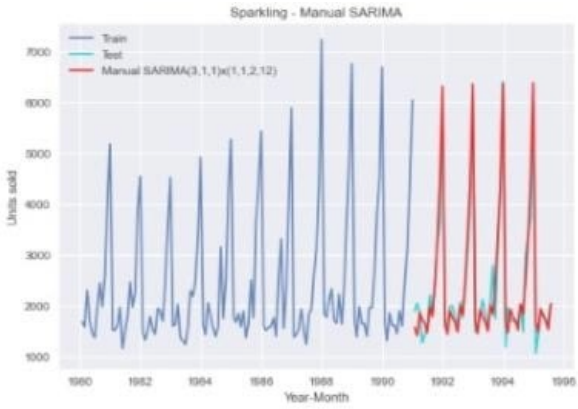
ar.L2	-0.0798	0.131	-0.607	0.544	-0.337	0.178
ar.L3	0.0921	0.122	0.756	0.450	-0.147	0.331
ma.L1	-1.0241	0.094	-10.925	0.000	-1.208	-0.840
ar.S.L12	-0.1992	0.866	-0.230	0.818	-1.897	1.499
ma.S.L12	-0.2109	0.881	-0.239	0.811	-1.938	1.516
ma.S.L24	-0.1299	0.381	-0.341	0.733	-0.877	0.617
sigma2	1.654e+05	2.62e+04	6.302	0.000	1.14e+05	2.17e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	19.66
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.81	Skew:	0.69
Prob(H) (two-sided):	0.56	Kurtosis:	4.78



- The final selected terms for SARIMA model is (3, 1, 1)x(1, 1, 2, 12)
- The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line
- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index

Forecast plotted against the test data.



For SARIMA forecast on the Sparkling Testing Data: RMSE is 324.107 and MAPE is 9.48

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.

We have 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 as below:

Sorting the results from all the models as per the RMSE values:

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.533374	10.41
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106510	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.642387	10.34
TES Alpha 0.11, Beta 0.06, Gamma 0.39	469.767970	16.40
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.049	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 9.99e-05	2007.238526	68.23
NaiveModel	3864.279352	152.87

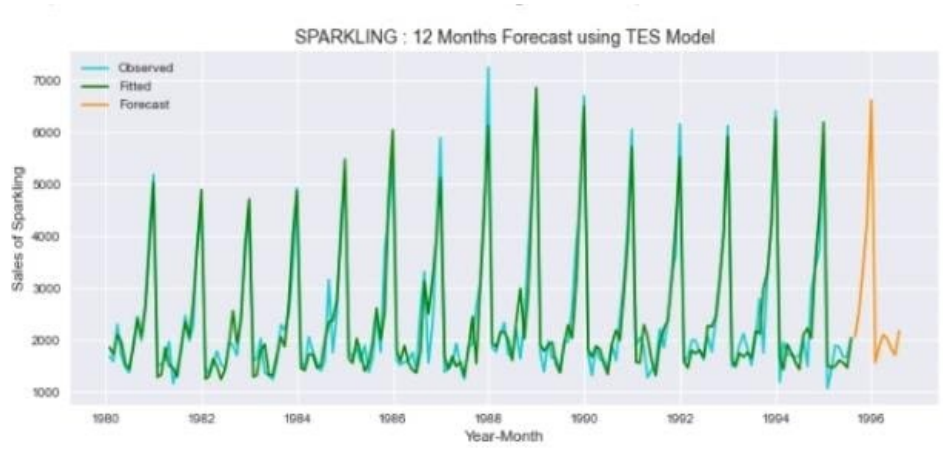
Sorting the results from all the models as per the MAPE values:

	Test RMSE	Test MAPE
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106510	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.642387	10.34
TES Alpha 0.4, Beta 0.1, Gamma 0.2	315.533374	10.41
TES Alpha 0.11, Beta 0.06, Gamma 0.39	469.767970	16.40
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
6 point TMA	1283.927428	43.86
SES Alpha 0.049	1316.035487	45.47
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.420000	67.23
DES Alpha 0.6,Beta 9.99e-05	2007.238526	68.23
NaiveModel	3864.279352	152.87

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.

TES forecast on the Sparkling Full Data: RMSE is 377.290 and MAPE is 11.36

- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data
- The SARIMA and Triple Exponential Smoothing are found to be comparable in terms of performance and fitment with the test data



- TES model alpha: 0.4, beta: 0.1 and gamma: 0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data
- The model predicts an upward trend and continuation of the seasonal surgein sales in the upcoming 12 months. According to the model the seasonal sale will be more than that of the previous year
- The 12 month prediction of the TES model is given below:

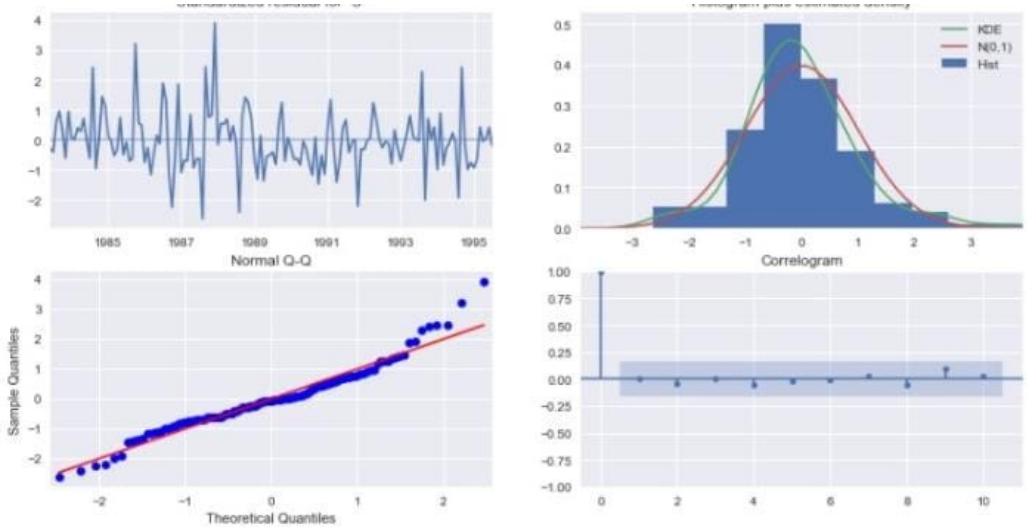
Attempt SARIMA(3,1,3)x(1,1,2,12) for forecast

SARIMAX Results

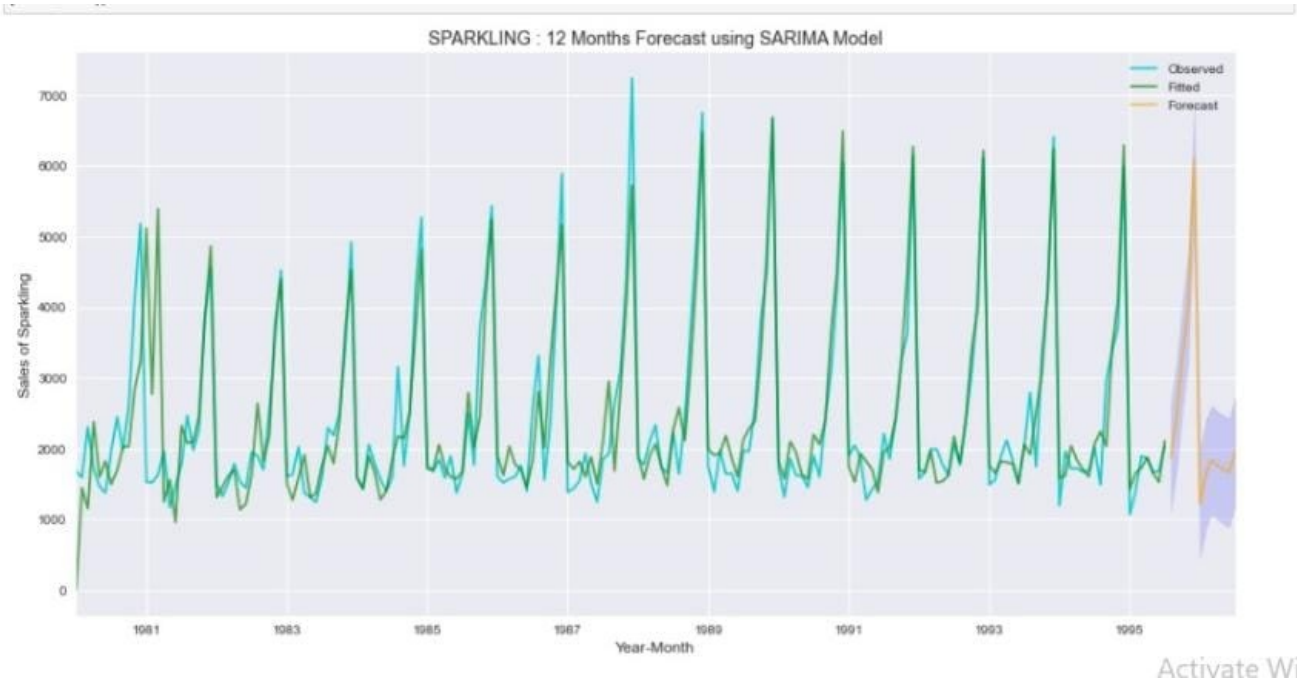
Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(3, 1, 3)x(1, 1, [1, 2], 12)	Log Likelihood	-1078.437
Date:	Fri, 08 Oct 2021	AIC	2176.875
Time:	19:15:06	BIC	2206.711
Sample:	01-31-1980	HQIC	2188.998
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4229	0.086	-4.917	0.000	-0.591	-0.254
ar.L2	-0.9093	0.053	-17.302	0.000	-1.012	-0.806
ar.L3	0.1425	0.087	1.640	0.101	-0.028	0.313
ma.L1	-0.4114	0.078	-5.283	0.000	-0.564	-0.259
ma.L2	0.4622	0.083	5.583	0.000	0.300	0.624
ma.L3	-0.9674	0.104	-9.329	0.000	-1.171	-0.764
ar.S.L12	-0.0692	0.708	-0.098	0.922	-1.457	1.318
ma.S.L12	-0.4559	0.720	-0.633	0.527	-1.867	0.955
ma.S.L24	-0.0804	0.396	-0.203	0.839	-0.856	0.696
sigma2	1.46e+05	1.05e-06	1.39e+11	0.000	1.46e+05	1.46e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	35.59
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.72	Skew:	0.66
Prob(H) (two-sided):	0.26	Kurtosis:	5.03

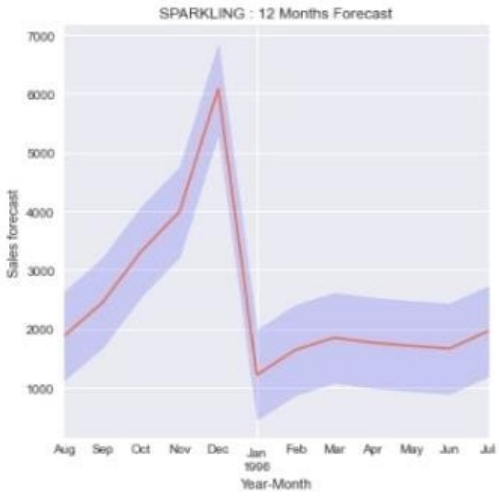


- The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot
- The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that AR(2), MA(3) terms has the highest absolute weightage. The p-values indicates that the terms AR(1), AR(2), MA(1), MA(2) and MA(3) are the most significant terms
- The rest of the p-values got values higher than alpha 0.05, which fails to reject the hull hypothesis that these terms are not significant



- The SARIMA model is built with parameters (3, 1, 3)x(1, 1, 2, 12), is found to be the most optimal SARIMA model
- SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE and MAPE values
- For SARIMA forecast on the Sparkling Full Data: RMSE is 591.263 and MAPE is 14.86

12 months forcast:



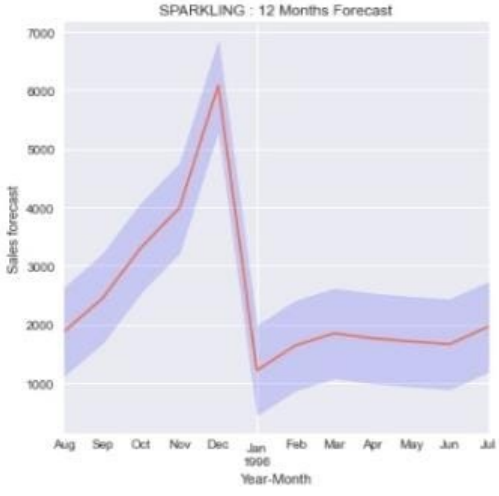
Forecasted Values for next 12 months:

	Sparkling
1995-08-31	1873.54
1995-09-30	2444.93
1995-10-31	3312.89
1995-11-30	3994.80
1995-12-31	6084.23
1996-01-31	1216.32
1996-02-29	1640.83
1996-03-31	1847.30
1996-04-30	1762.21
1996-05-31	1708.57
1996-06-30	1664.03
1996-07-31	1961.43

	Sparkling
count	12.000000
mean	2459.256667
std	1384.626967
min	1216.320000
25%	1697.435000
50%	1860.420000
75%	2661.920000
max	6084.230000

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

12 months forcast:



Forecasted Values for next 12 months:

	Sparkling
1995-08-31	1873.54
1995-09-30	2444.93
1995-10-31	3312.89
1995-11-30	3994.80
1995-12-31	6084.23
1996-01-31	1216.32
1996-02-29	1640.83
1996-03-31	1847.30
1996-04-30	1762.21
1996-05-31	1708.57
1996-06-30	1664.03
1996-07-31	1961.43

comments and the Recomondations:

- The model forecasts sale of 29510 units of Sparkling wine in 12 months intofuture. Which is an average sale of 2459 units per month
- The seasonal sale in December 1995 will hit a maximum of 6084 units, before it drops to the lowest sale in January 1996; at 1216 units.
- The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995 (October, November and December), which is a total of 13,392 units of sparkling wine is expected to be sold.
- The forecast also indicates that the year-on-year sale of sparkling wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve the sale compared to previous years
- Adding more exogenous variable into the timeseries data can improve forecasts.

For Rose Dataset:

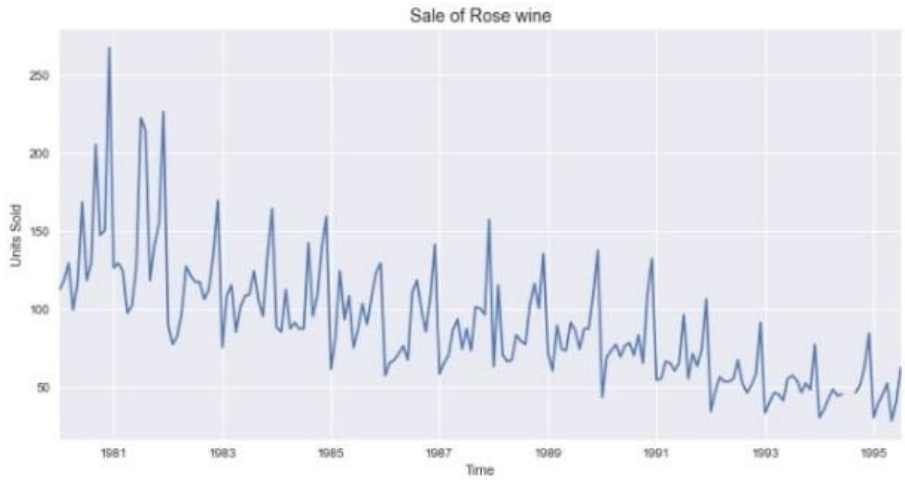
- Importing all the necessary libraries
- Loading data in Data Frame ### 1. Read the data as an appropriate Time Series data and plot the data. ### Checking head of the data

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Checking tail of the data

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Plot the Time Series to understand the behaviour of the data



- We can see Negative trend and seasonality in the above Time series data.
- The demand for Rose had been fell out-of-favour over the years

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

Checking for missing values

Rose 2

dtype: int64

There are two missing values in the dataset

The time-series got values missing for two months in 1994,

```
Out[14]: Time_Stamp
         1994-01-31    30.000000
         1994-02-28    35.000000
         1994-03-31    42.000000
         1994-04-30    48.000000
         1994-05-31    44.000000
         1994-06-30    45.000000
         1994-07-31    45.336957
         1994-08-31    45.673913
         1994-09-30    46.000000
         1994-10-31    51.000000
         1994-11-30    63.000000
         1994-12-31    84.000000
         Name: Rose, dtype: float64
```

The Rose time-series got values missing for two months in 1994, which are imputed using interpolation (linear method)

Check the basic measures of descriptive statisticsCheck the basic measures of descriptive statistics

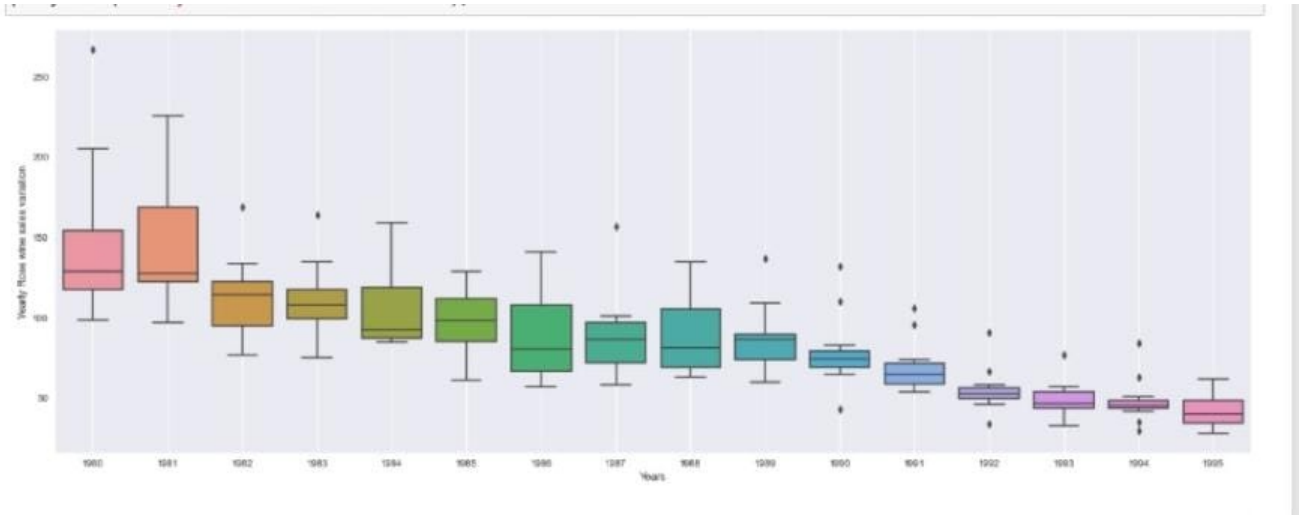
ut[16]:

	Rose
count	187.000000
mean	89.914497
std	39.238259
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

- The descriptive summary of the data shows that on an average 90 units of Rose wines were sold each month on the given period of time. 50% of months sales varied from 63 units to 112 units. Maximum sale reported in a month is 267 units and minimum of 28 units

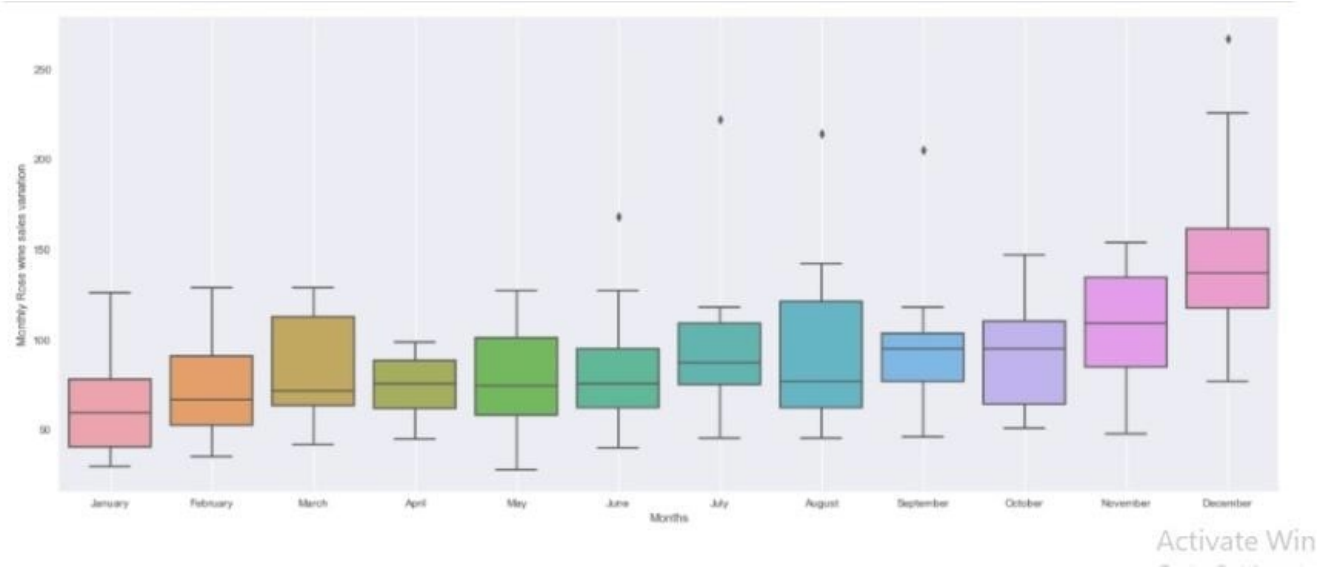
Plot a boxplot to understand the spread of sales across different years and within different months across years.

Yearly Boxplot



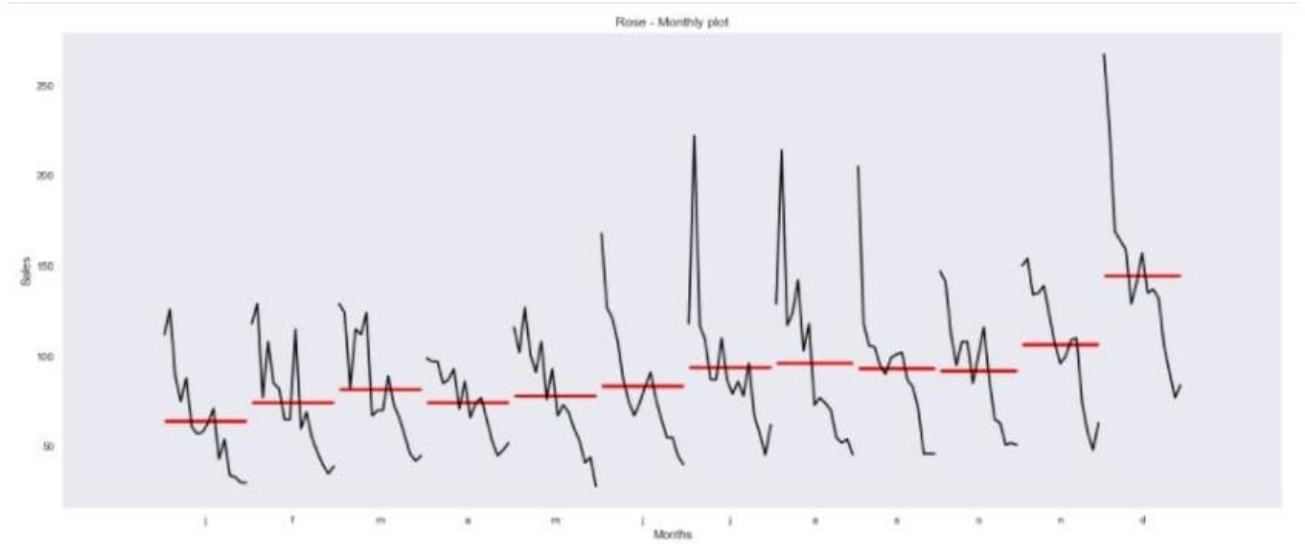
- The yearly-boxplot, shows that the average sale of Rose wine moving according to the downward trend in sales over the years. The outliers over upperbound in the yearly-boxplot most probably represent the seasonal sale during the seasonal months

Monthly plot

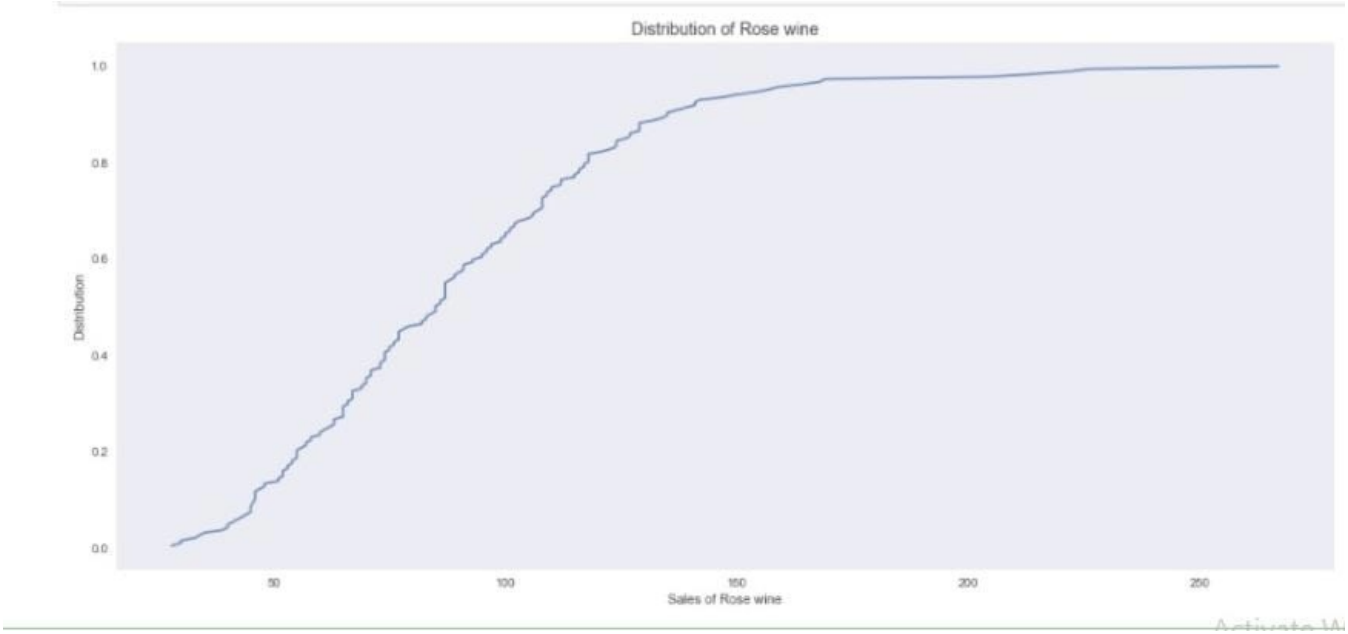


- The monthly-box-plot shows a clear seasonality during the seasonal months of November and December.
- Average sale in December is around 140 units, November is around 110 units and October is around 90 units.
- The monthly plot for Rose shows mean and variation of units sold each month over the years. Sale in months such as July, August, September and December shows a higher variation than the rest

Plot a time series monthplot to understand the spread of sales across different years and within different months across years

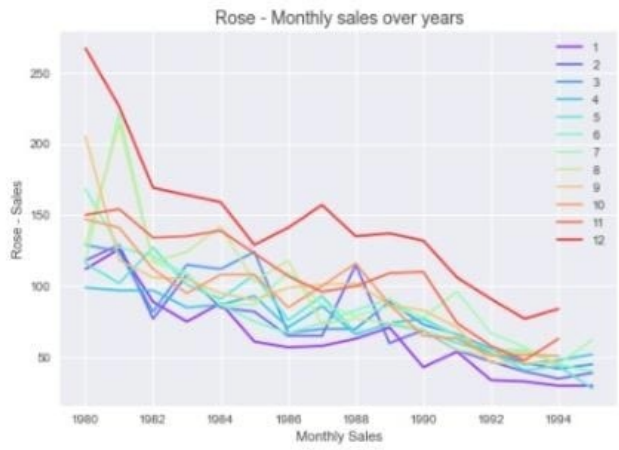


Plot the Empirical Cumulative Distribution.



- The Empirical CDF plot shows that, in 80% of months, at least 120 units of Rose wine were sold

Plot a graph of monthly sales across years

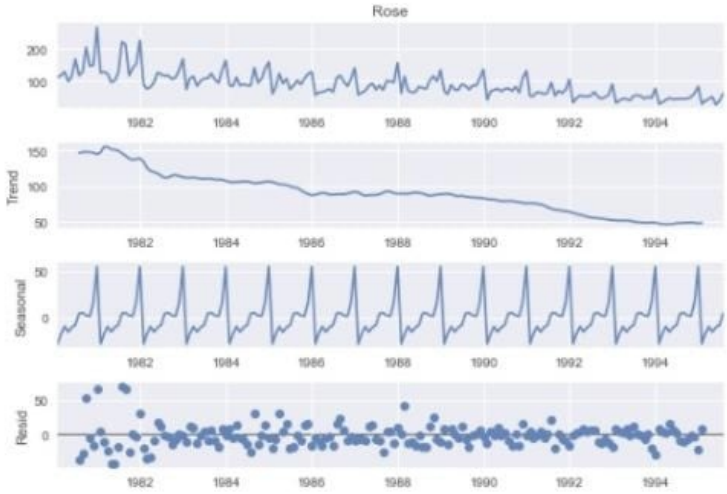


- The plot of monthly sale over the years also shows the seasonality component of the time-series, with November and December selling exponentially higher volumes than other months.
- The highest volume of Rose wines were sold in December,1980 and the least of December sale was in 1993. Though December sale picked after 1983, it consistently dipped after 1987

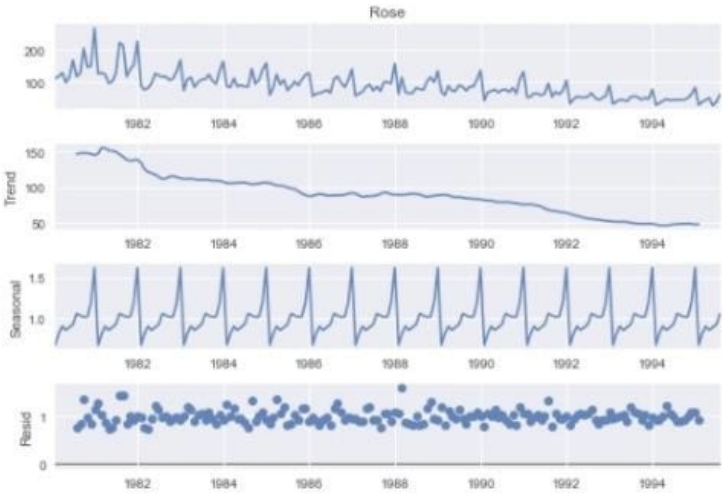
Plot the average sales per month and the month on month percentage change of sales.

Decompose the Time Series and plot the different components

Additive Decomposition



Multiplicative Decomposition



- The observed plot of the decomposition diagrams shows visible annual seasonality and a downward trend. The early period of the plot shows higher variation than in the later periods

- Seasonal components are quite visible and consistent in both the observed and seasonal charts of the diagrams. The additive chart shows variance in seasonality from -20 to 50 units and the multiplicative model shows variance of 16%
- The residuals shows a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions
- As the seasonality peaks are consistently reducing its altitude in consistent with trend, the series can be treated as multiplicative in model building

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

Spliting the data into training and testing, The train and test datasets are created with year 1991 as starting year for test data,

The data head and the tail for the training and testing data:

First few rows of Training Data

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Last few rows of Training Data

Time_Stamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

First few rows of Test Data

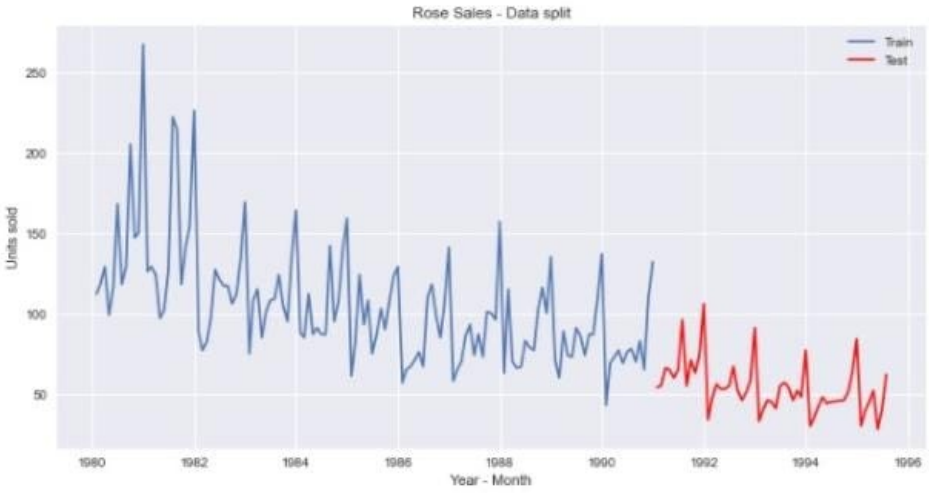
First few rows of Test Data

Time_Stamp	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Last few rows of Test Data

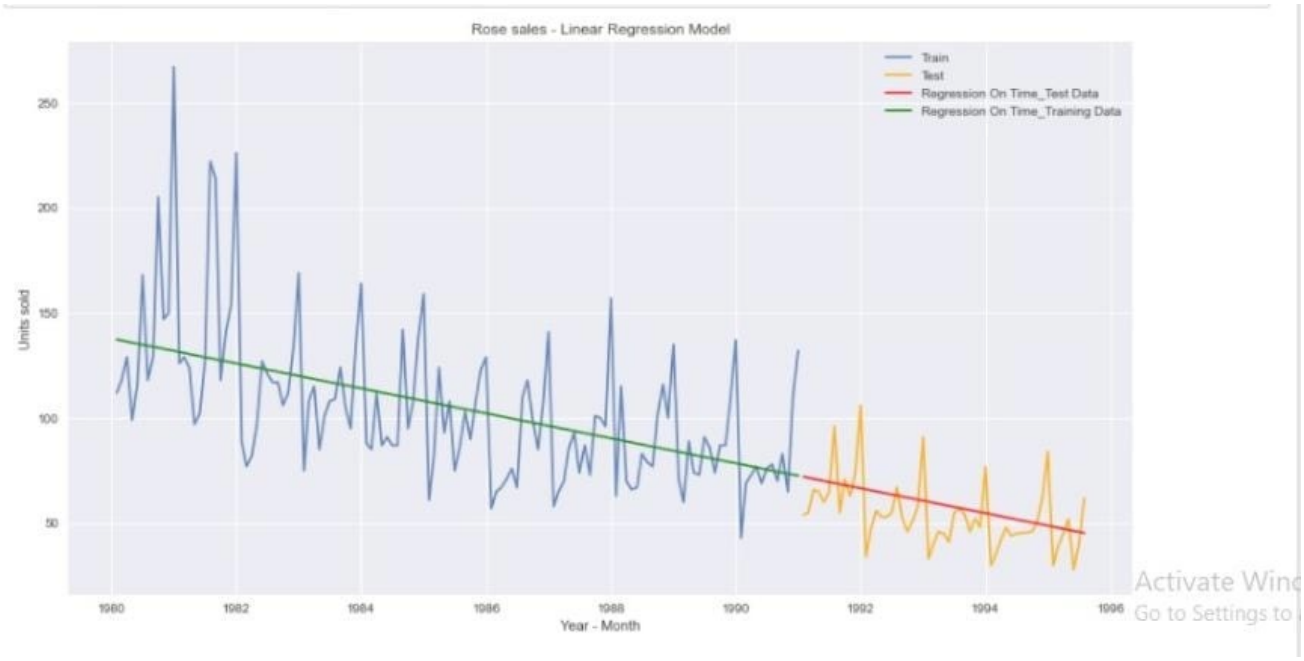
Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Rose Sales - Data split



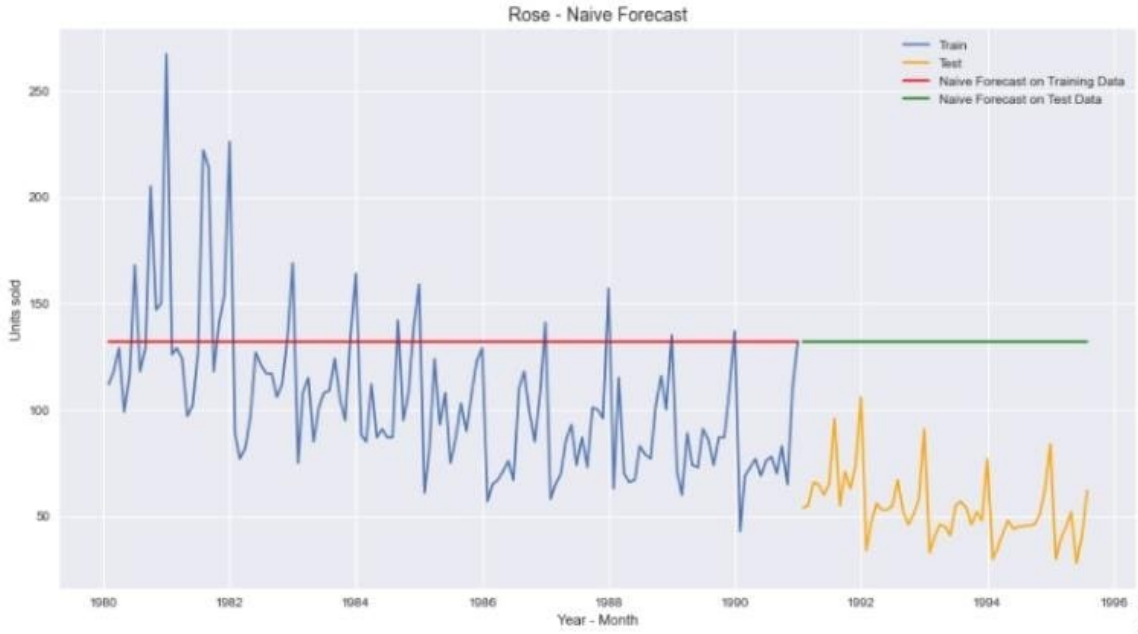
4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression



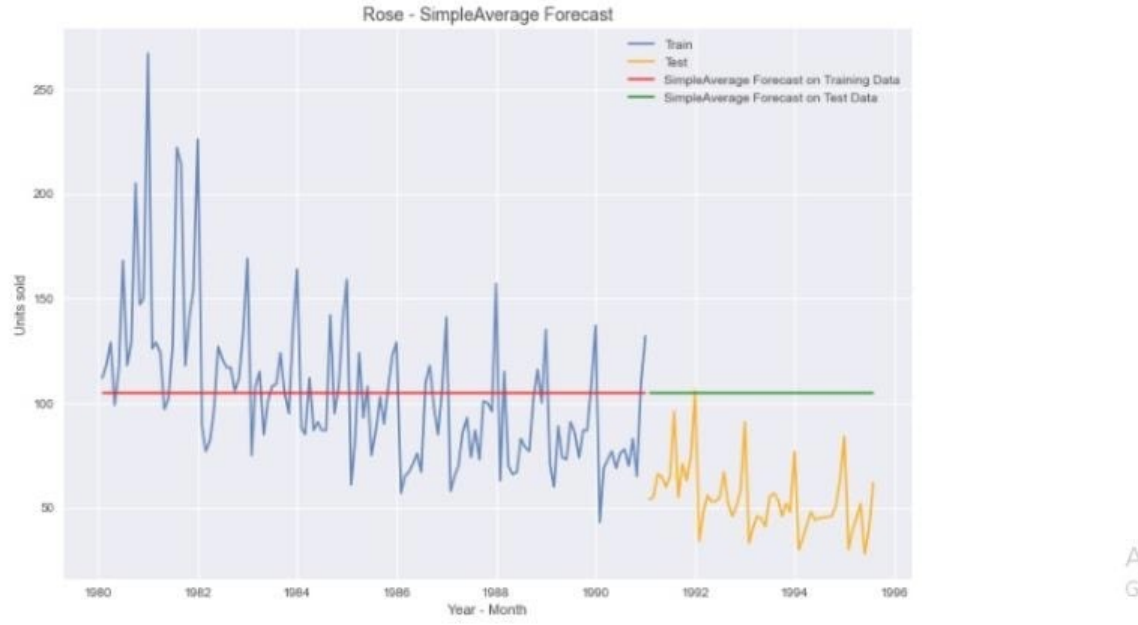
- shows an apparent downward trend as consistent with the observed time-series
- The model has successfully captured the trend of both the series, but does not reflects the seasonality
- For RegressionOnTime forecast on the Rose Training Data: RMSE is 30.718 and MAPE is 21.22
- For RegressionOnTime forecast on the Rose testing Data: RMSE is 15.269 and MAPE is 22.82

Model 2: Naive forecast



- As data set has a downward trend the percentage of error in train is lesser and is very high in test
- The model does not capture the trend nor seasonality of the given datasets
- For Naive forecast on the Rose Training Data: RMSE is 45.064 and MAPE is 36.38
- For Naive forecast on the Rose Testing Data: RMSE is 79.719 and MAPE is 145.10

Model 3: Simple Average



- The model forecast is almost 100% error in test data and 25% in train
- For Simple Average forecast on the Rose Training Data: RMSE is 36.034 and MAPE is 25.39
- For Simple Average forecast on the Rose Testing Data: RMSE is 53.460 and MAPE is 94.93

Model 4: Moving Average

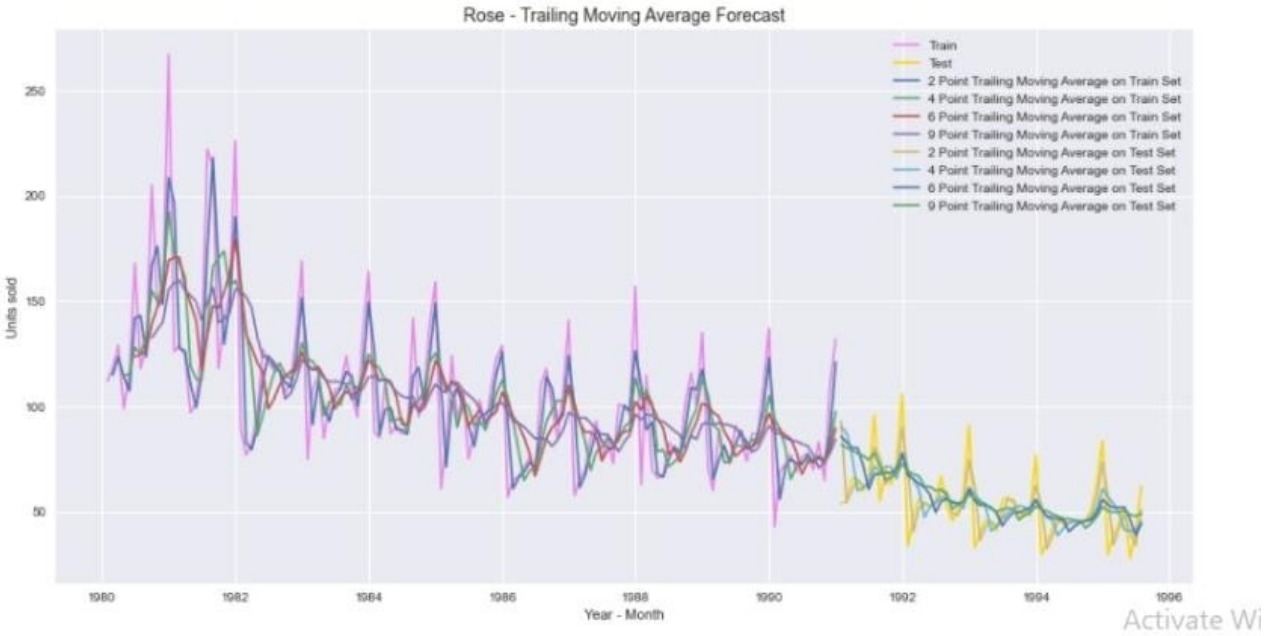
The best interval can be determined by the maximum accuracy

we are going to calculate rolling means for different intervals:

Out[61]:

Time_Stamp	Rose	Rose_Trailing_2	Rose_Trailing_4	Rose_Trailing_6	Rose_Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN

- The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points



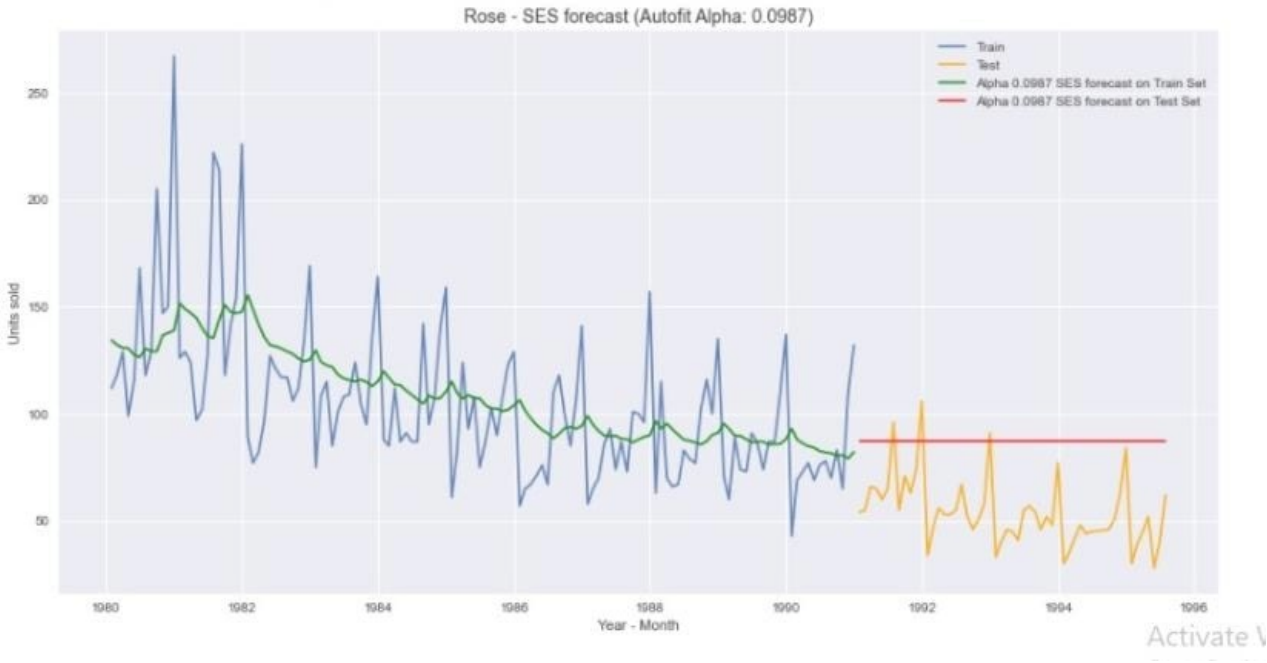
- For 2 point Moving Average Model forecast on the Training Data, rmse_rose is 11.529 mape_rose is 13.54
- For 4 point Moving Average Model forecast on the Training Data, rmse_rose is 14.451 mape_rose is 19.49
- For 6 point Moving Average Model forecast on the Training Data, rmse_rose is 14.566 mape_rose is 20.82
- For 9 point Moving Average Model forecast on the Training Data, rmse_rose is 14.728 mape_rose is 21.01

Model 5: Simple Exponential Smoothing

The parameters are found to be:

Out[70]:
```
{'smoothing_level': 0.0987493111726833,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38720226208358,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Simple Exponential Smoothing is usually applied if the time-series has neither a trend nor seasonality, which is not the case with the given

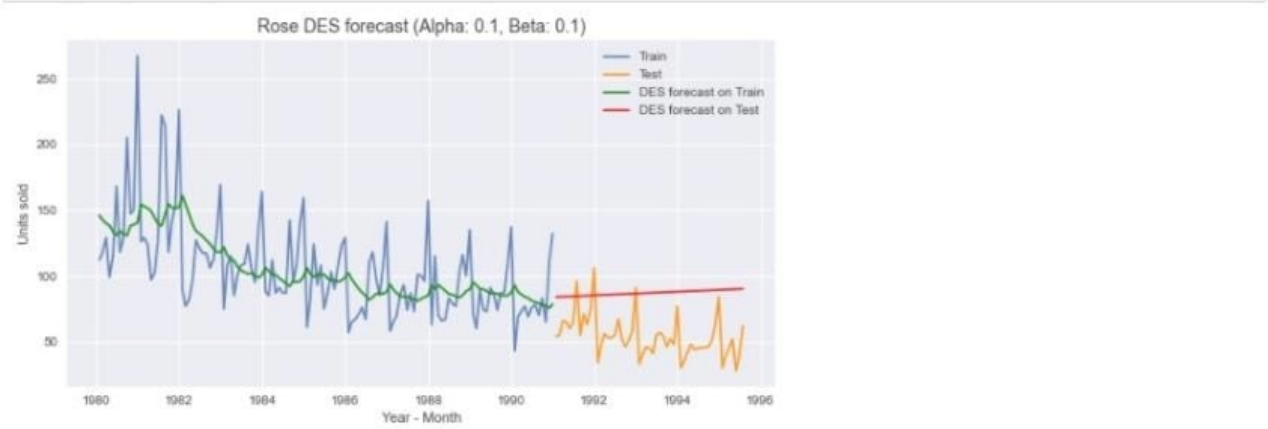


data

- For SES forecast on the Rose Training Data: RMSE is 31.501 and MAPE is 22.73
- For SES forecast on the Rose Testing Data: RMSE is 36.796 and MAPE is 63.88

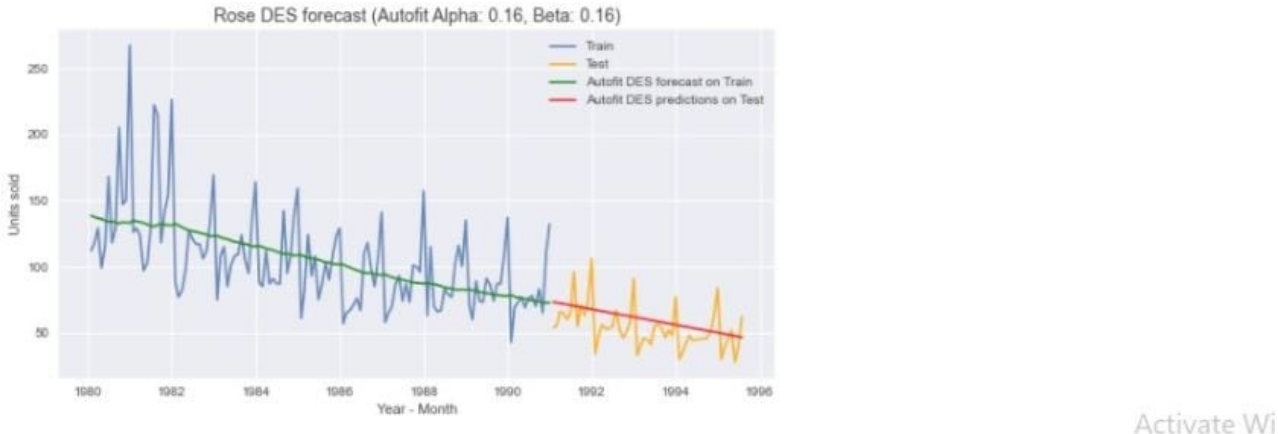
Model 6: Double Exponential Smoothing (Holt's Model)

- The Double Exponential Smoothing models is applicable when data has trend, but no seasonality. Rose data contain significant trend component and seasonality



- On the second iteration the model was allowed to chose the optimized values using parameters ‘optimized=True, use_brute=True

Rose DES forecast (Autofit Alpha: 0.16, Beta: 0.16)



- The autofit model retuned higher accuracy in train dataset, on par with the best models from iteration 1, but faired behind in the test accuracy scores

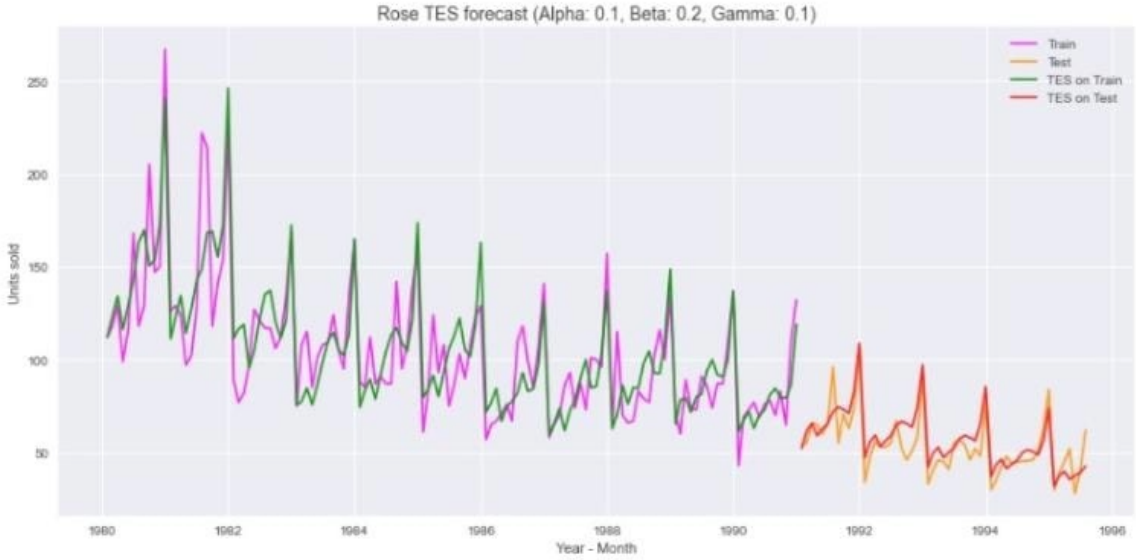
Out[89]:

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
100	0.01755	0.000032	30.890794	21.61	15.706968	24.12
0	0.10000	0.100000	32.026565	22.78	37.056911	64.02
1	0.10000	0.200000	32.685228	23.63	48.806921	83.29
10	0.20000	0.100000	32.796403	23.06	65.731352	113.20
2	0.10000	0.300000	32.925494	24.23	78.209401	131.33

Model 7: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters α , β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

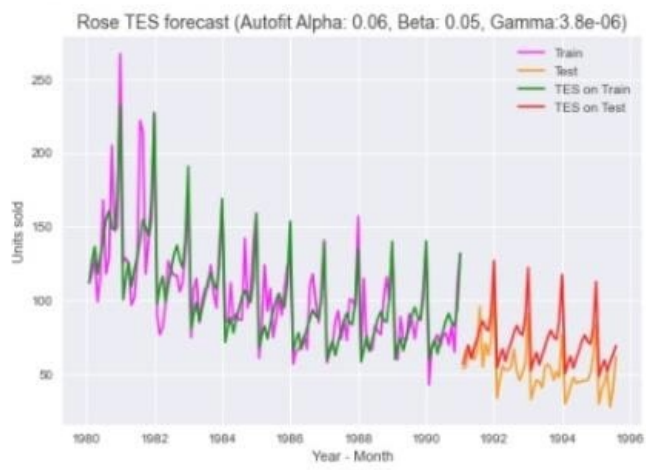
- The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Rose data contain both trend and seasonality significantly



- In first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1, beta 0.2 and gamma 0.1

On the second iteration the model was allowed to chose the optimized values using parameters 'optimized=True, use_brute=True'

Autofit model of TES



- The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test

Out[106]:

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
10	0.1	0.2	0.1	19.651464	14.31	9.171621	13.19
11	0.1	0.2	0.2	20.140683	14.66	9.493832	13.68
151	0.2	0.6	0.2	22.793871	17.02	9.682585	13.71
142	0.2	0.5	0.3	23.300524	17.35	9.885717	14.21
12	0.1	0.2	0.3	20.725703	14.88	9.896169	14.16

- The model evaluation parameters of the best models are given as above, including one from the autofit iteration
- The best model chosen as final one is the one with alpha 0.1, beta 0.2 and gamma 0.1

MODEL COMPARISON

Ploting all the above models

[109]:

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.1	9.493832	13.68
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
DES Alpha 0.01, Beta 3.2e-05	15.706968	24.12
TES Alpha 0.06, Beta 0.05, Gamma 3.8e-06	21.019341	35.16
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

- The accuracy of the time-series forecast models build in the previous sections of this report is as below, sorted by RMSE in test data
- be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data
- 2 point trailing moving average model is also found to have fit well with a slight lag in test dataset

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.

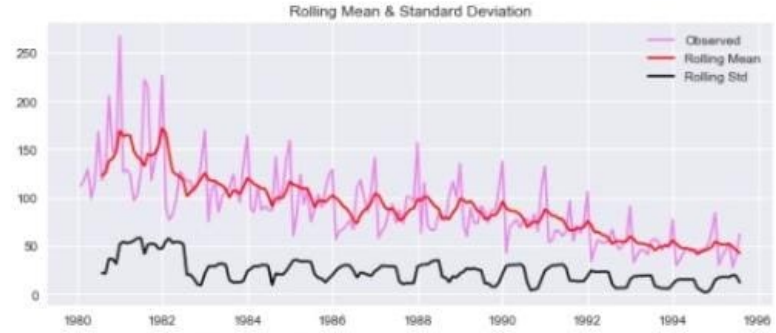
Note: Stationarity should be checked at alpha = 0.05.

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

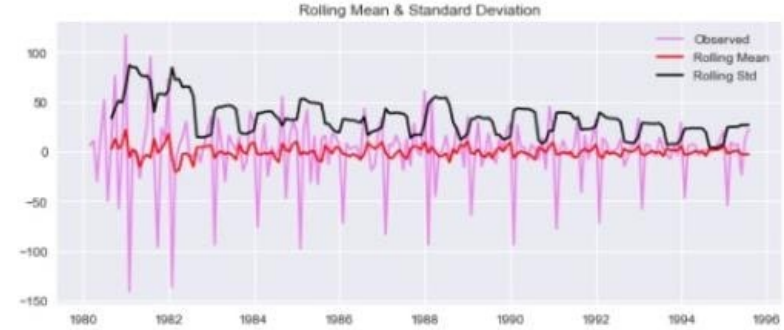
The hypothesis in a simple form for the ADF test is:

- 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.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876719
p-value                          0.343091
#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
dtype: float64
```


```
Results of Dickey-Fuller Test:
Test Statistic                -8.044395e+00
p-value                        1.810868e-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
dtype: float64
```


- The ADF test on the original Rose series retuned the below values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis.
- Differencing of order one is applied on the Sparkling series as above and tested for stationarity
- At an order of differencing 1, the series is found to be stationary as above
- The plot of rolling mean and standard deviation indicates that the seasonality is multiplicative as the altitude of plot varies with respect to trend

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.

AUTO SARIMA on original data

As the data contains seasonality component we will be building SARIMA model, rather than ARIMA.

:[125]:

	param	seasonal	AIC
163	(2, 1, 2)	(0, 1, 3, 12)	104.422530
251	(3, 1, 3)	(2, 1, 3, 12)	511.234313
99	(1, 1, 2)	(0, 1, 3, 12)	572.552977
247	(3, 1, 3)	(1, 1, 3, 12)	587.250447
243	(3, 1, 3)	(0, 1, 3, 12)	592.526851
255	(3, 1, 3)	(3, 1, 3, 12)	653.840917
221	(3, 1, 1)	(3, 1, 1, 12)	681.362808
253	(3, 1, 3)	(3, 1, 1, 12)	681.610083
254	(3, 1, 3)	(3, 1, 2, 12)	681.964120
222	(3, 1, 1)	(3, 1, 2, 12)	682.320697

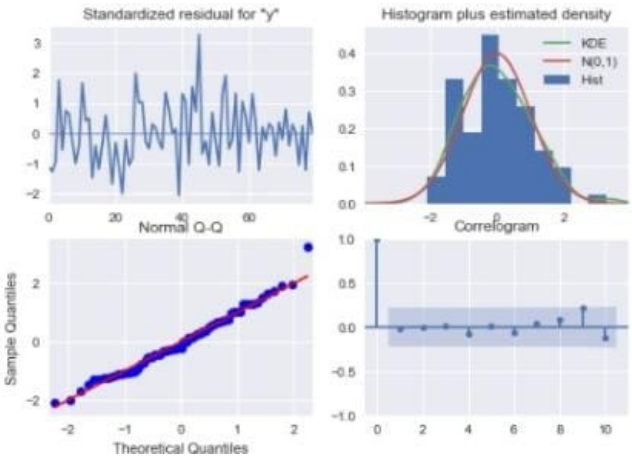
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 1, 1, 12)	Log Likelihood	-331.681
Date:	Fri, 08 Oct 2021	AIC	681.363
Time:	19:15:45	BIC	702.801
Sample:	0	HQIC	689.958
	- 132		
Covariance Type:	opg		

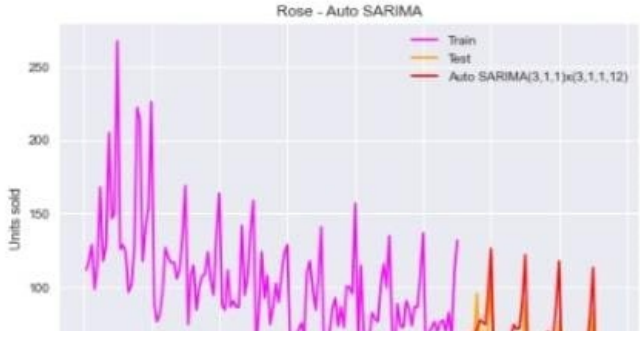
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0173	0.151	0.114	0.909	-0.279	0.314
ar.L2	-0.0426	0.141	-0.302	0.762	-0.319	0.234
ar.L3	-0.0575	0.119	-0.485	0.628	-0.290	0.175
ma.L1	-0.9388	0.085	-11.107	0.000	-1.104	-0.773
ar.S.L12	0.0907	0.126	0.720	0.471	-0.156	0.337
ar.S.L24	-0.0436	0.108	-0.406	0.685	-0.254	0.167
ar.S.L36	-3.594e-05	0.053	-0.001	0.999	-0.103	0.103
ma.S.L12	-0.9997	183.450	-0.005	0.996	-360.555	358.555
sigma2	185.4302	3.4e+04	0.005	0.996	-6.65e+04	6.68e+04

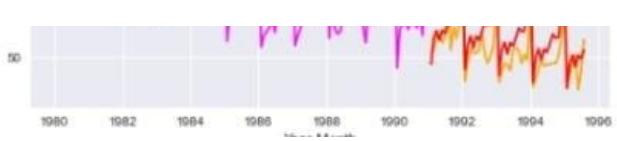
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	2.56
Prob(Q):	0.91	Prob(JB):	0.28
Heteroskedasticity (H):	0.56	Skew:	0.42
Prob(H) (two-sided):	0.13	Kurtosis:	3.22

- The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values
- The best AIC values selected for the best model is (3, 1, 1)x(3, 1, 1, 12)



- The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line

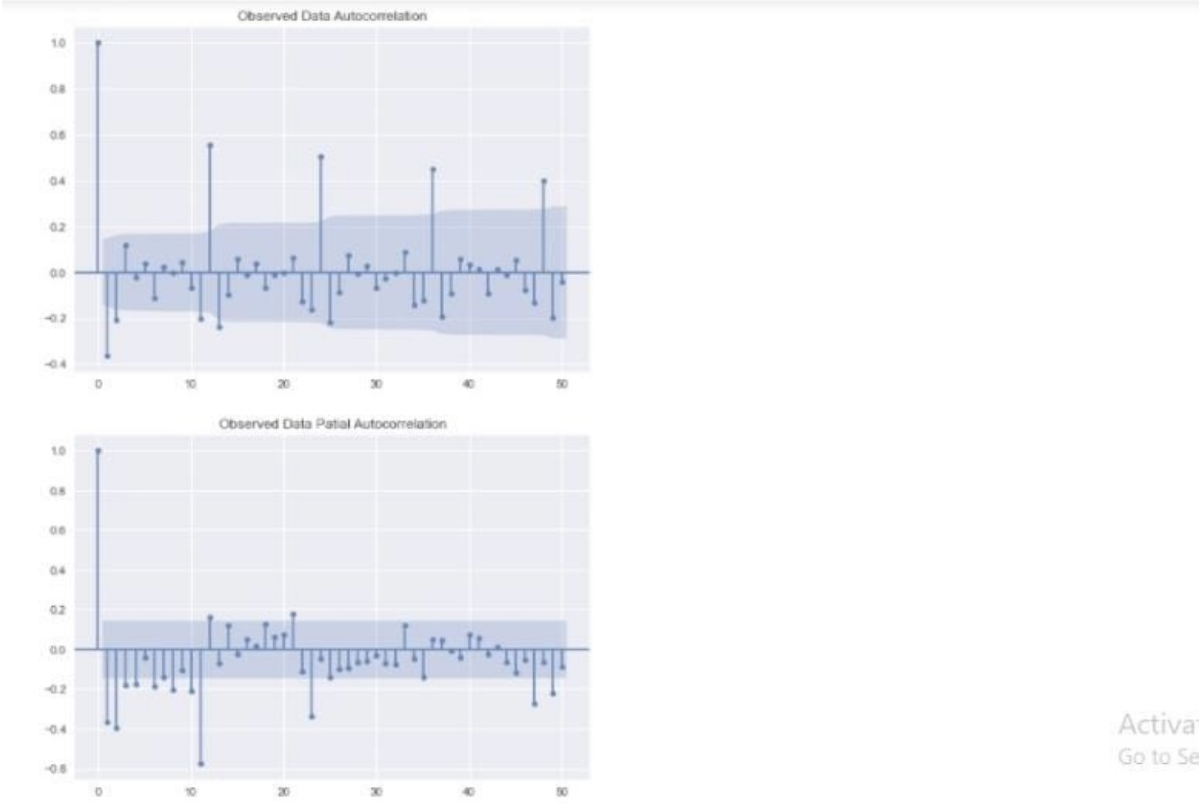




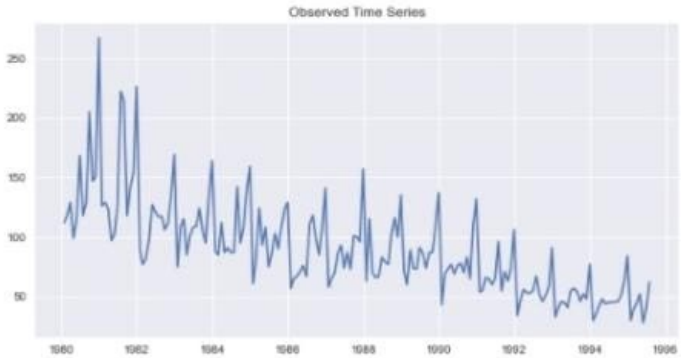
- For SARIMA forecast on the SRose Testing Data: RMSE is 16.824 and MAPE is 25.48

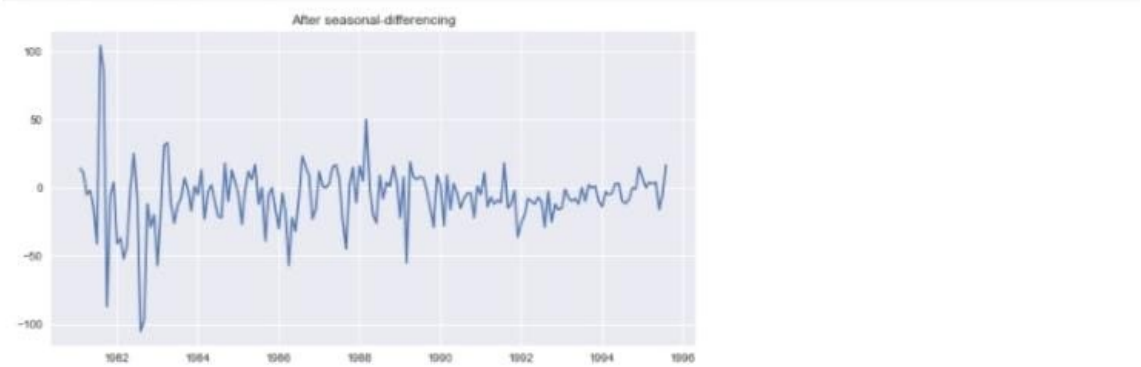
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

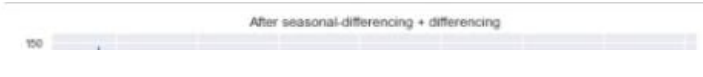
Let us look at the ACF and the PACF plots.

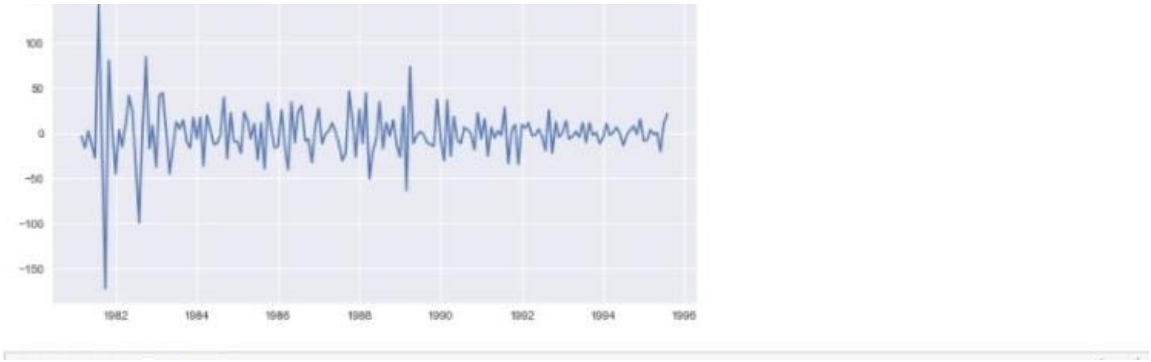


- We see that our ACF plot at the seasonal interval (12) does not taper off quickly. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.





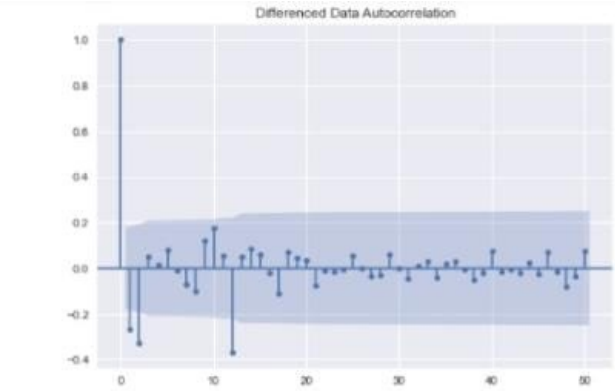


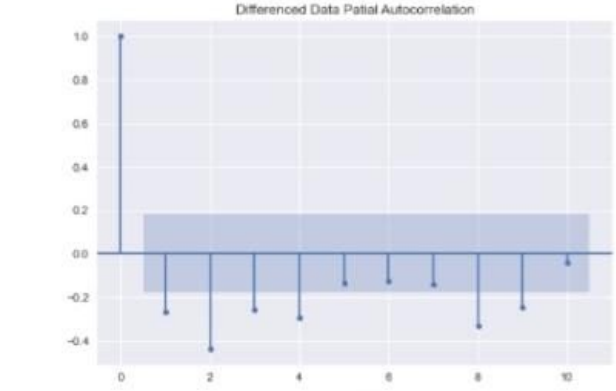


- From the above plots it can be seen that a slight trend is still existing after differencing of seasonal order of 12. With a further differencing of order one, no trend is present


```
Results of Dickey-Fuller Test:
Test Statistic                -4.605732
p-value                        0.000126
#Lags Used                    11.000000
Number of Observations Used  162.000000
Critical Value (1%)           -3.471374
Critical Value (5%)           -2.879552
Critical Value (10%)          -2.576373
dtype: float64
```


- Have done an ADF test to check the stationarity after the above differencing. With a pvalue below alpha 0.05 and test statistic below critical values, it can be confirmed that thedata is stationary





- Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till lag 4 is significant before cut-off, so AR term ‘p = 4’ is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’.

- From ACF plot, lag 1 and 2 are significant before it cuts off, so lets keep MA term ‘q = 1’ and at seasonal lag of 12, a significant lag is apparent, so lets keep ‘Q = 1'.
- The final selected terms for SARIMA model is (4, 1, 1)x(0, 1, 1, 12), as inferred from the ACF and PACF plots.

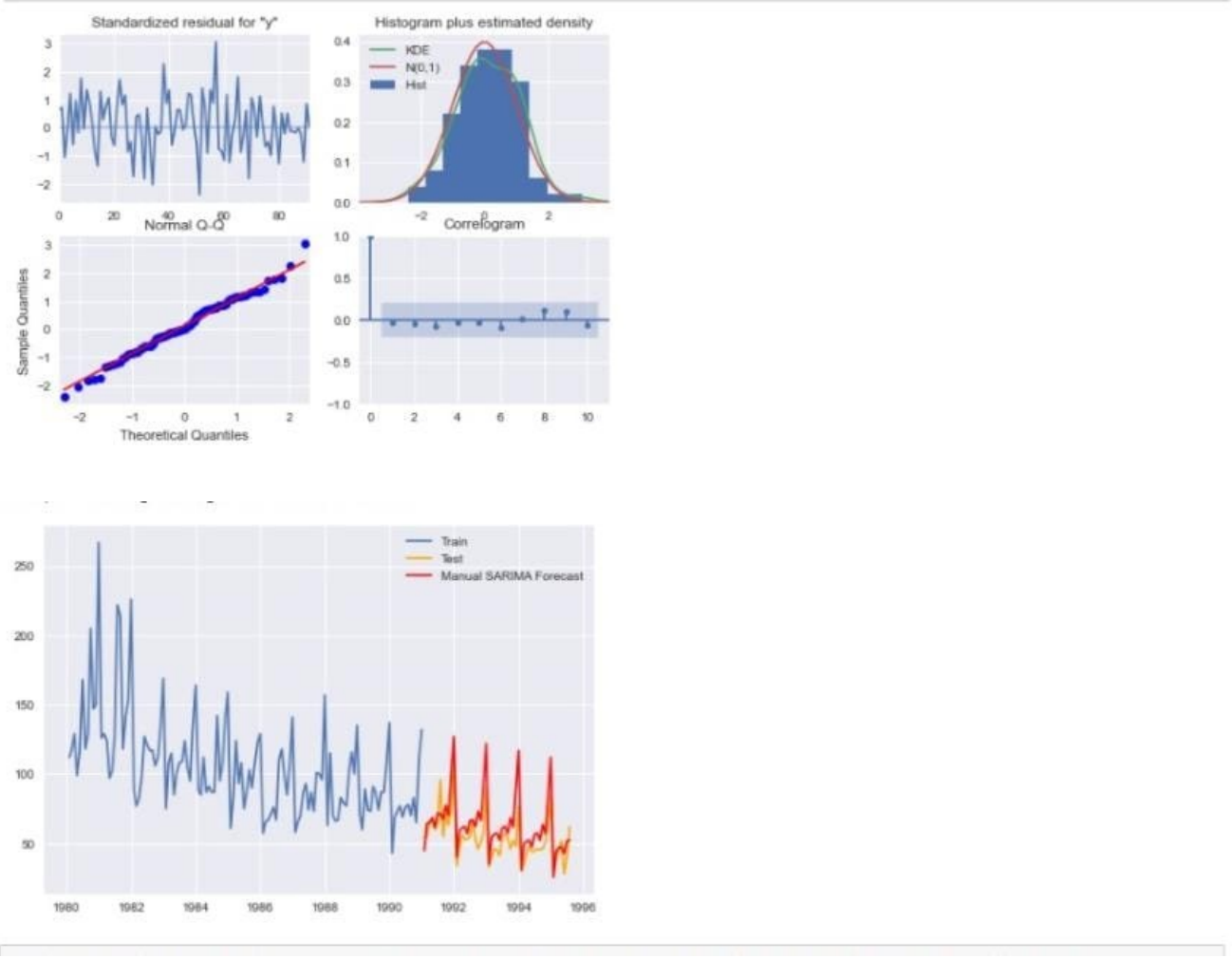
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-384.369
Date:	Fri, 08 Oct 2021	AIC	786.737
Time:	19:16:28	BIC	809.433
Sample:	0	HQIC	795.898
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8967	0.132	-6.814	0.000	-1.155	-0.639
ar.L2	0.0165	0.171	0.097	0.923	-0.319	0.352
ar.L3	-0.1132	0.174	-0.650	0.515	-0.454	0.228
ar.L4	-0.1598	0.116	-1.380	0.168	-0.387	0.067
ma.L1	0.1508	0.174	0.866	0.387	-0.191	0.492
ma.L2	-0.8492	0.164	-5.166	0.000	-1.171	-0.527
ma.S.L12	-0.3907	0.102	-3.848	0.000	-0.590	-0.192
ma.S.L24	-0.0887	0.091	-0.977	0.329	-0.267	0.089
sigma2	238.9649	0.001	2.02e+05	0.000	238.963	238.967

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	0.01
Prob(Q):	0.80	Prob(JB):	0.99
Heteroskedasticity (H):	0.76	Skew:	-0.01
Prob(H) (two-sided):	0.46	Kurtosis:	3.06

- The final selected terms for SARIMA model is (4, 1, 2)x(0, 1, 2, 12), as inferred from the ACF and PACF plots
- The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero
- The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line
- The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index



- For SARIMA forecast on the Rose Testing Data: RMSE is 15.377 and MAPE is 22.16

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.

We have 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 as below:

Sorting the results from all the models as per the RMSE values:

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.1	9.493832	13.68
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
DES Alpha 0.01, Beta 3.2e-05	15.706968	24.12
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823819	25.48
TES Alpha 0.06, Beta 0.05, Gamma 3.8e-06	21.019341	35.16
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

Sorting the results from all the models as per the RMSE values:

	Test RMSE	Test MAPE
2 point TMA	11.529278	13.54
TES Alpha 0.1, Beta 0.2, Gamma 0.1	9.493832	13.68
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
RegressionOnTime	15.268885	22.82
DES Alpha 0.01, Beta 3.2e-05	15.706968	24.12
Auto SARIMA(3,1,1)x(3,1,1,12)	16.823819	25.48
TES Alpha 0.06, Beta 0.05, Gamma 3.8e-06	21.019341	35.16
SES Alpha 0.01	36.796004	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
NaiveModel	79.718559	145.10

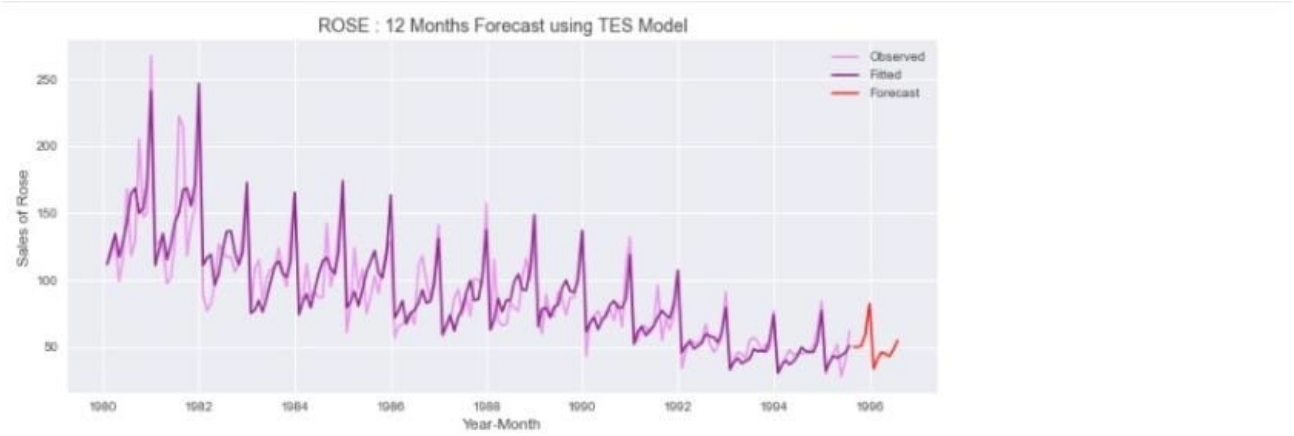
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands

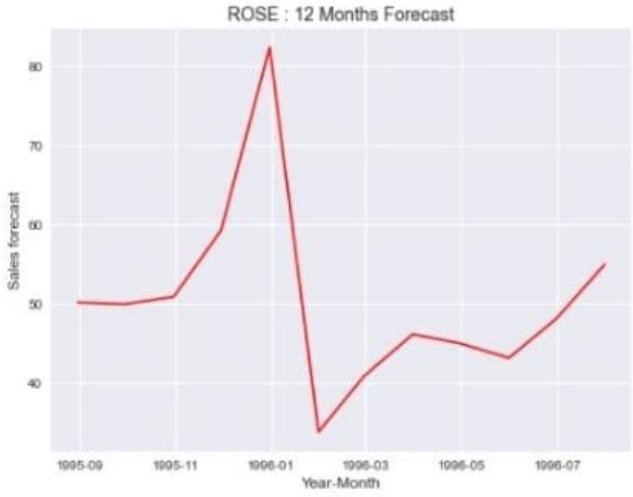
The overall comparison of all the time-series forecast models are listed below in accordance with increasing RMSE against test data or in the order of decreasing accurac

- Triple Exponential Smoothing is found to be the best model, followed by 2 point Moving Average
- The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data

We have selected Triple Exponential Smoothing (Holt Winter's) and SARIMA for final prediction into 12 months in future

The 12 month prediction of the TES model is as below:





- TES model alpha: 0.1, beta: 0.2 and gamma: 0.1 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data

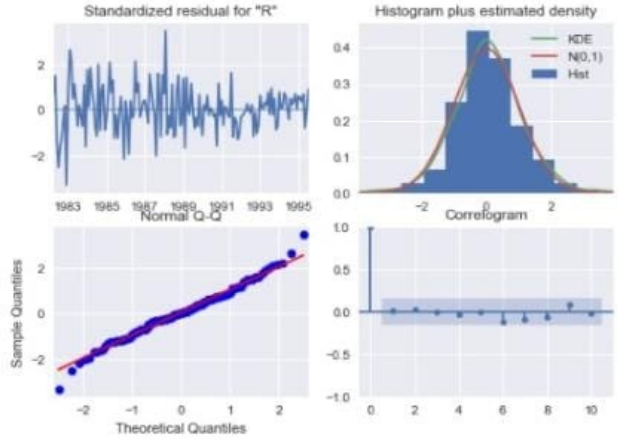
SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-664.135
Date:	Fri, 08 Oct 2021	AIC	1342.270
Time:	19:16:39	BIC	1363.796
Sample:	01-31-1980 - 07-31-1995	HQIC	1351.011
Covariance Type:	opg		

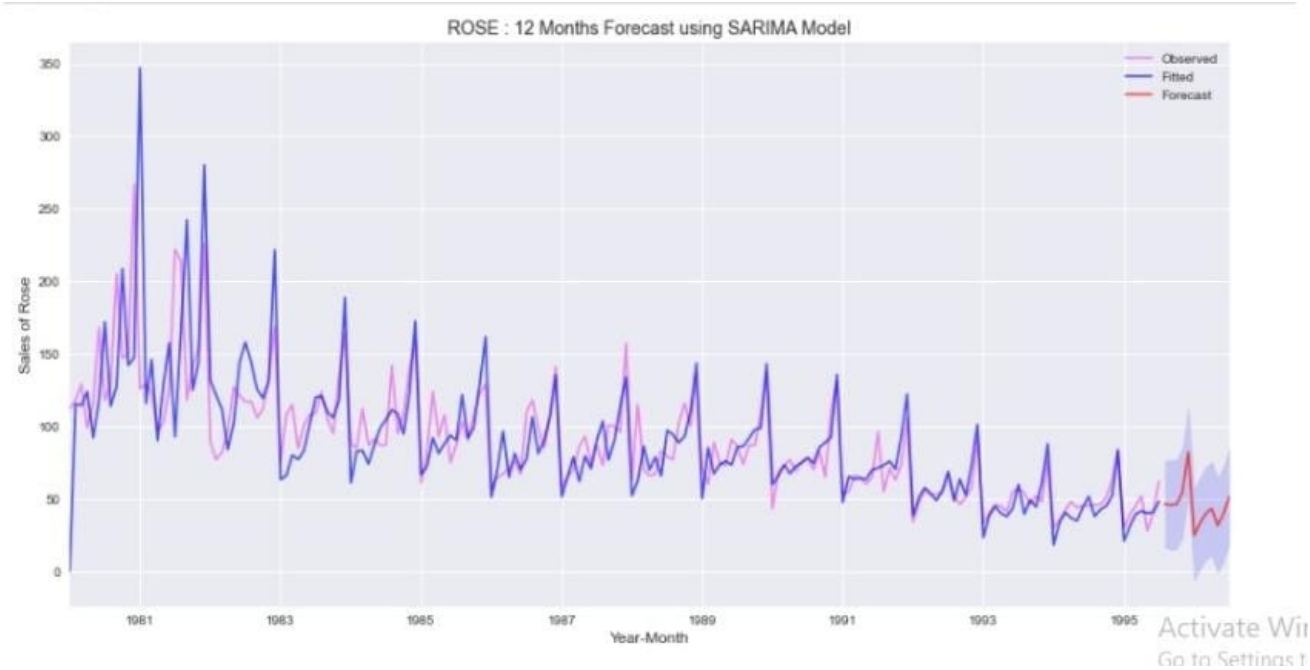
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0914	0.084	1.093	0.274	-0.072	0.255
ar.L2	-0.1077	0.077	-1.393	0.164	-0.259	0.044
ar.L3	-0.1314	0.076	-1.729	0.084	-0.280	0.018
ar.L4	-0.1071	0.078	-1.375	0.169	-0.260	0.046
ma.L1	-0.8270	0.055	-14.901	0.000	-0.936	-0.718
ma.S.L12	-0.5963	0.059	-10.122	0.000	-0.712	-0.481
sigma2	232.4248	24.359	9.542	0.000	184.682	280.168

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	5.30
Prob(Q):	0.93	Prob(JB):	0.07
Heteroskedasticity (H):	0.22	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.89

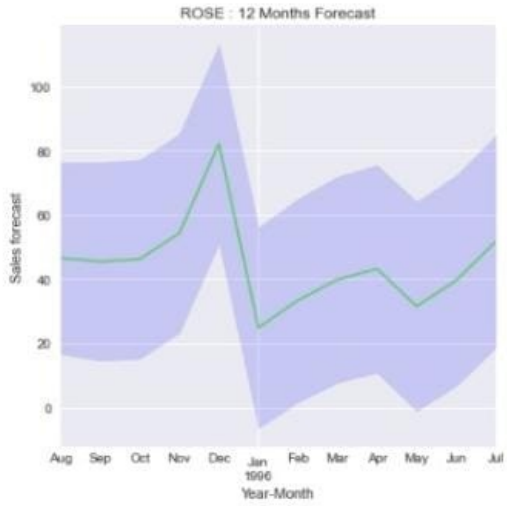
- The SARIMA model is built with parameters (4, 1, 1)x(0, 1, 1, 12), is found to be the most optimal SARIMA model for the complete time-series



- The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot.
- The rest of the p-values got values higher than alpha 0.05, which fails to reject the null hypothesis that these terms are not significant.



12 months forcast:



- SARIMA model has also reflected the trend and seasonality of the series continuing into the future year as well.
- The SARIMA model is chosen as the final model for prediction on Rose dataset, as it provide confidence interval and better explainability of the model

Forecasted Values for next 12 months:

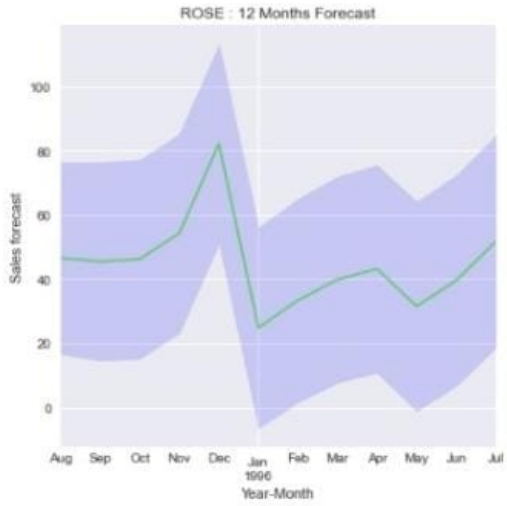
177]:

	ROSE
1995-08-31	46.54
1995-09-30	45.51
1995-10-31	46.23
1995-11-30	54.32
1995-12-31	82.21
1996-01-31	24.81
1996-02-29	33.35
1996-03-31	39.87
1996-04-30	43.23
1996-05-31	31.53
1996-06-30	39.56
1996-07-31	51.70

	ROSE
count	12.000000
mean	44.905000
std	14.473222
min	24.810000
25%	38.007500
50%	44.370000
75%	47.830000
max	82.210000

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

12 months forcast:



Forecasted Values for next 12 months:

177]:

	ROSE
1995-08-31	46.54
1995-09-30	45.51
1995-10-31	46.23
1995-11-30	54.32
1995-12-31	82.21
1996-01-31	24.81
1996-02-29	33.35
1996-03-31	39.87
1996-04-30	43.23
1996-05-31	31.53
1996-06-30	39.56
1996-07-31	51.70

- The model forecasts sale of 539 units of Rose wine in 12 months into future. Which is an average sale of 45 units per month
- The seasonal sale in December 1995 will reach a maximum of 82 units, before it drops to the lowest sale in January 1996; at 25 units.
- Unlike Sparkling wine, Rose wine sells very low number of units and the standard deviation is only 14.5. Which means that higher demand does not impact procurement and production
- Apart from higher sale in November and December months, Rose sales will be above average in the summer months of July and August
- The winery should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions.