Business Report on
Time Series Forecasting for Rose Wine Sales data

Yedupati Venkata Yamini

06 Aug 2023

Table of Contents:

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.
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.
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.
7. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
8. 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.
9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

List of figures:

List of tables:

Problem 1:

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 Rose Wine Sales in the 20th century.

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

Ans:

Viewing the first and last rows of original data:

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

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Table1: First and last 5 rows of original dataset

The data has been read as appropriate time series data with dates parsed using the Year Month column.
Viewing the first 5 rows of parsed time series data:

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

	YearMonth	Rose
182	1995-03-01	45.0
183	1995-04-01	52.0
184	1995-05-01	28.0
185	1995-06-01	40.0
186	1995-07-01	62.0

Table2: First and last 5 rows of parsed training and test data

Viewing the basic information of the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    datetime64[ns]
 1   Rose       185 non-null    float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 3.0 KB
```

Table3: Basic information of Rose wines dataset

- The given data set has 187 rows and 2 columns including MonthYear column and the sales of Rose Wine during the months. From 1980 to 1995.
- 'Rose' wine sales have 2 missing values for the months July and August in 1994.
- Imputed the null values with the sales data of June month in 1994.
- Basic information of the processed dataset:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    187 non-null    float64
 1   Year    187 non-null    int64
 2   Month   187 non-null    object
dtypes: float64(1), int64(1), object(1)
memory usage: 9.9+ KB
```

Table4: Basic information of null value treated Rose wines dataset.

- Basic summary of the processed dataset.

```
: count    187.000000
  mean      89.909091
  std       39.244440
  min       28.000000
  25%       62.500000
  50%       85.000000
  75%      111.000000
  max      267.000000
  Name: Rose, dtype: float64
```

Table5: Summary of Rose wine sales

Insights:
1. The columns present in the dataset in the initial glance contain monthly data of Rose Wine sales.
2. Average sales of Rose wine over the years is around 89$ monthly.

Plotting the time series data:

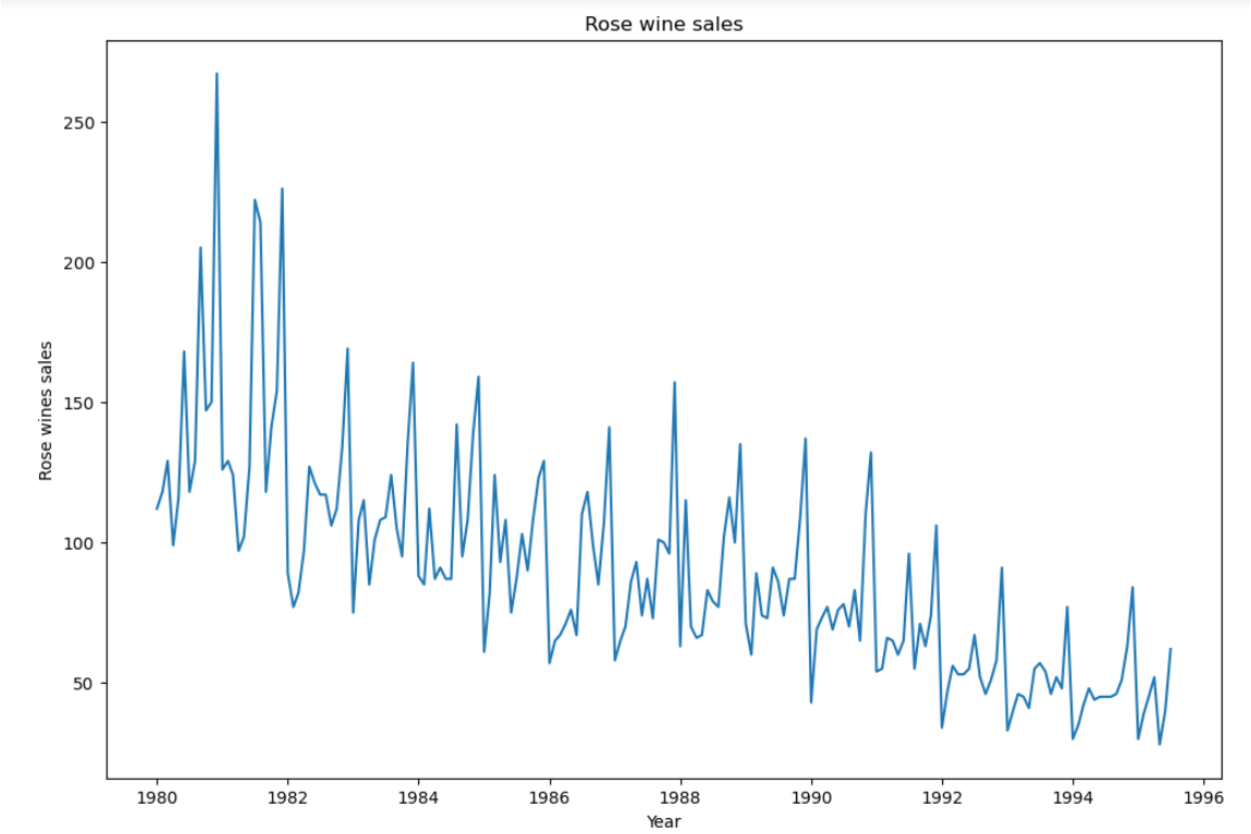


Figure1: Time series plot of Rose wine sales

Insights:

1. The wine sales show a steady decreasing trend in terms of Rose Wine sales over the years.
2. However, there is a seasonality observed in the plot. The sales of wine seem to reach peaks during particular times of the year.

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

Univariate analysis:

1. Basic summary of the processed dataset.

```
count    187.000000
mean      89.909091
std       39.244440
min       28.000000
25%       62.500000
50%       85.000000
75%      111.000000
max      267.000000
Name: Rose, dtype: float64
```

Table6: Summary of Rose wine sales dataset

Insights:

- The columns present in the dataset in the initial glance contain monthly data of Rose Wine sales.
- Average sales of Rose wine over the years is around 89$ monthly.

Bivariate analysis:

2. Yearly sales of Rose wine:

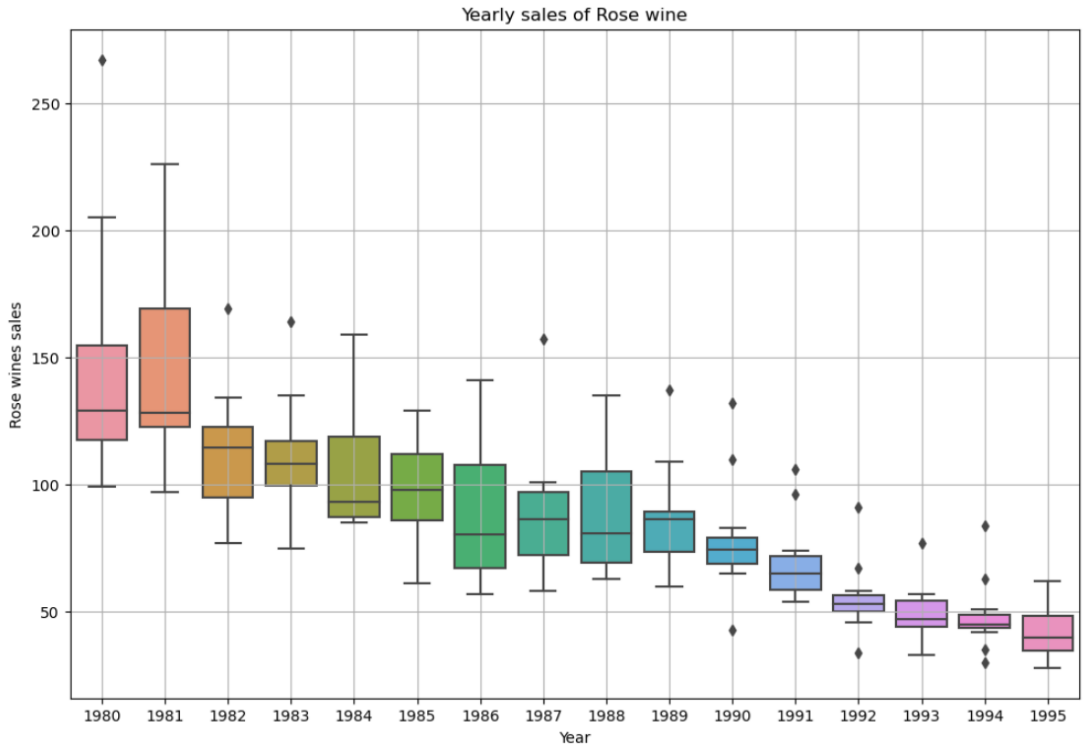


Figure2: Year wise Rose wine sales

Insights:

- Sales for the wine have been declining year by year.
- It has seemed to increase a bit in the year 1988 but there is no increase before/after that.
- The same can be visualized in a line plot as shown below

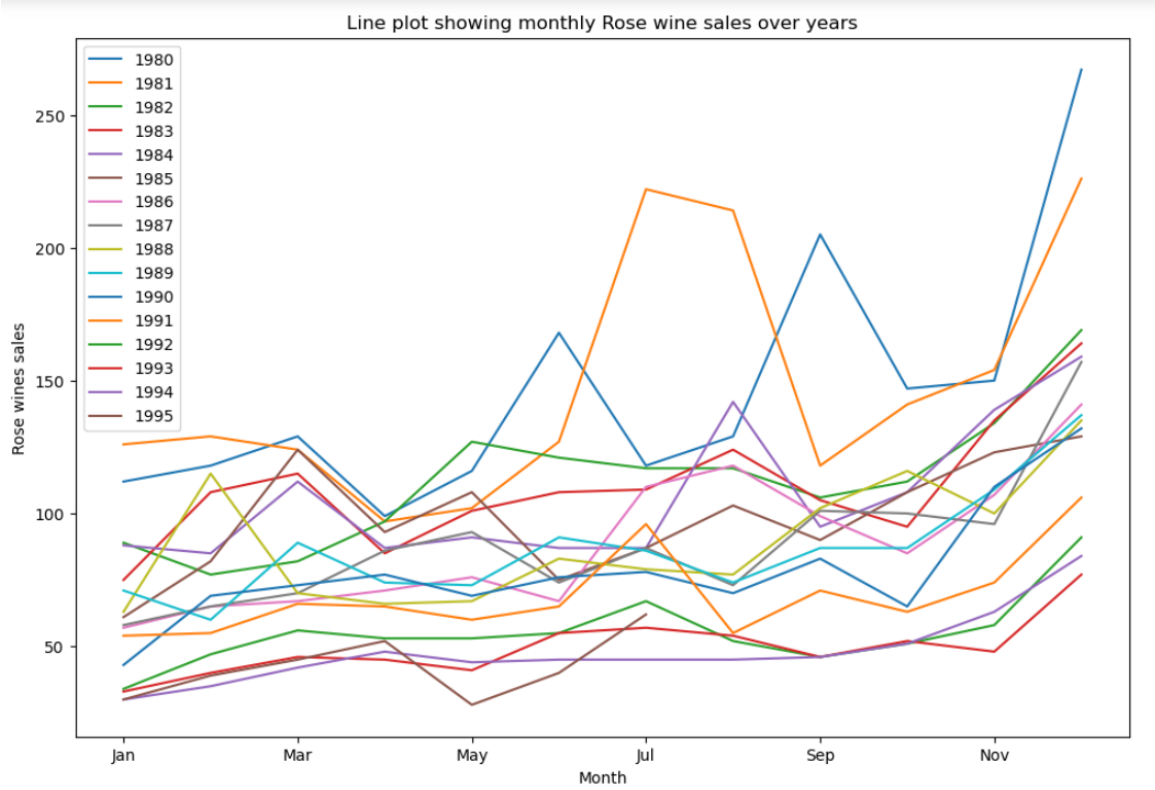


Figure3: Line plot of year wise Rose wine sales

- Every year, the company is experiencing a mild spike in Wine sales during the mid and end of the year.
- 1981 is the year that recorded the highest sales for Rose wines.

3. Distribution of Rose wine sales in a month every year:

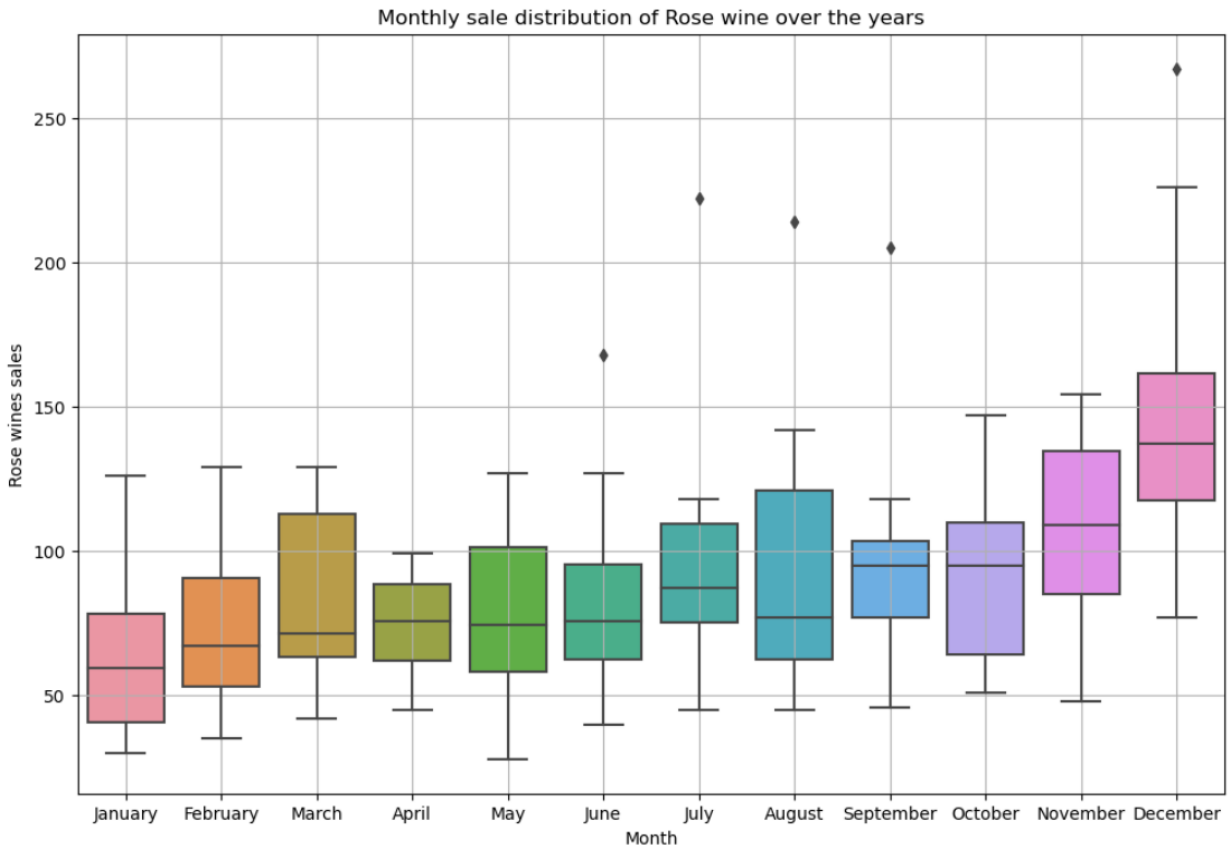


Figure4: Distribution of Rose wine sales in a month every year

Insights:

- This plot indicates that Rose wine sales are in a slightly higher margin during the months August, March, November, December and lowest during April.
- This graph can depict the seasonality of the sales in all the years.
- The sales follow an oscillating seasonal trend over the months.

4. Distribution of the wine sales indexed by months and years:

:

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

Table7: Distribution of the wine sales indexed by months and years

5. Average quarterly Rose wines sales:

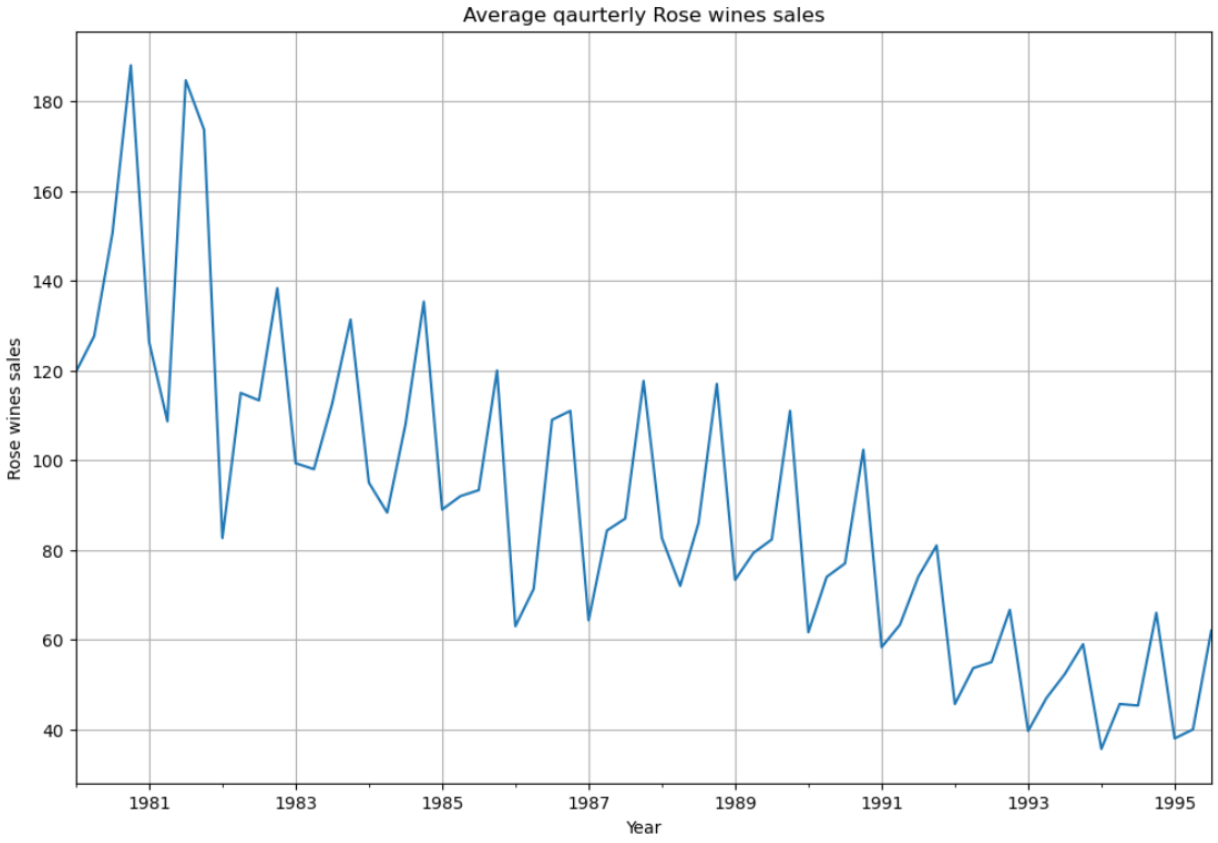


Figure5: Average quarterly Rose wines sales

Insights:

- The sales have been highest during initial years and have been decreasing steadily.

6. Line plot showing trend of sales for every month in every year:

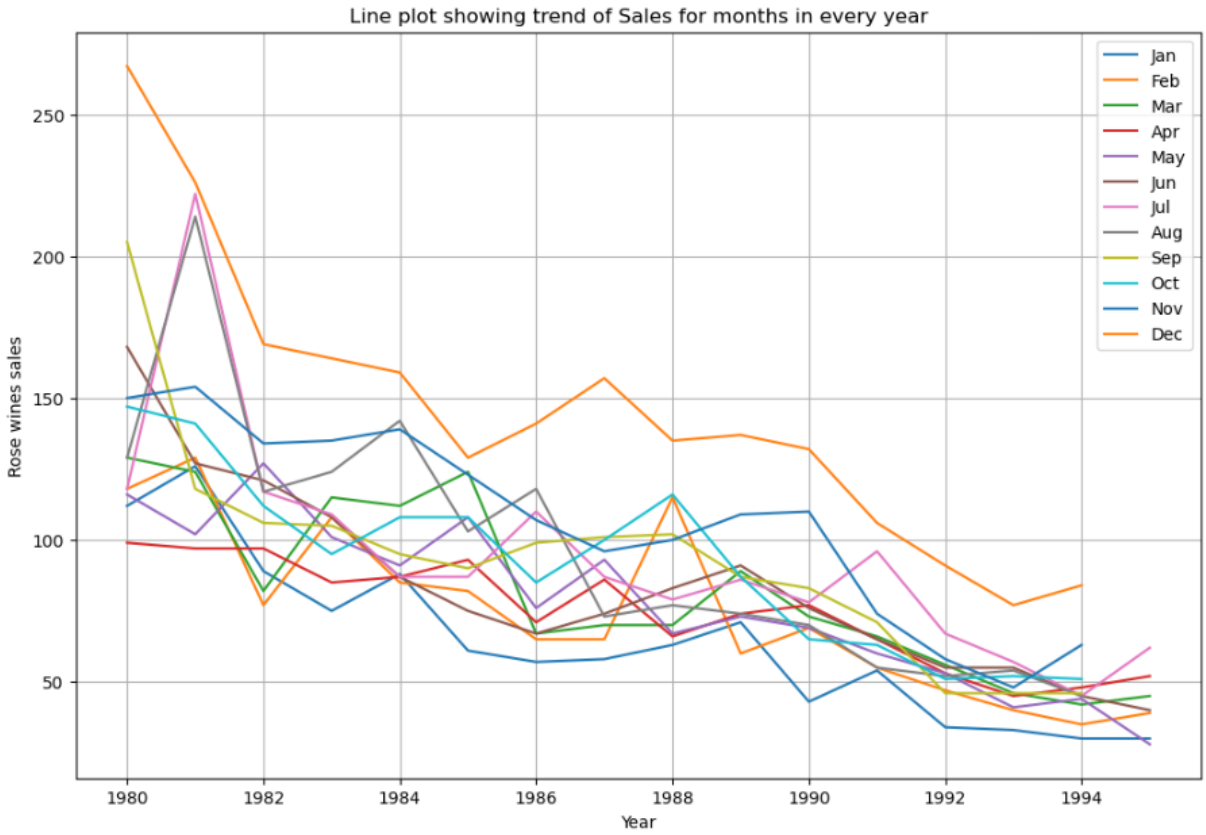


Figure6: Line plot showing trend of sales for every month in every year

Insights:

- It depicts a decreasing trend with higher sales in December.

Additive seasonal decomposition:

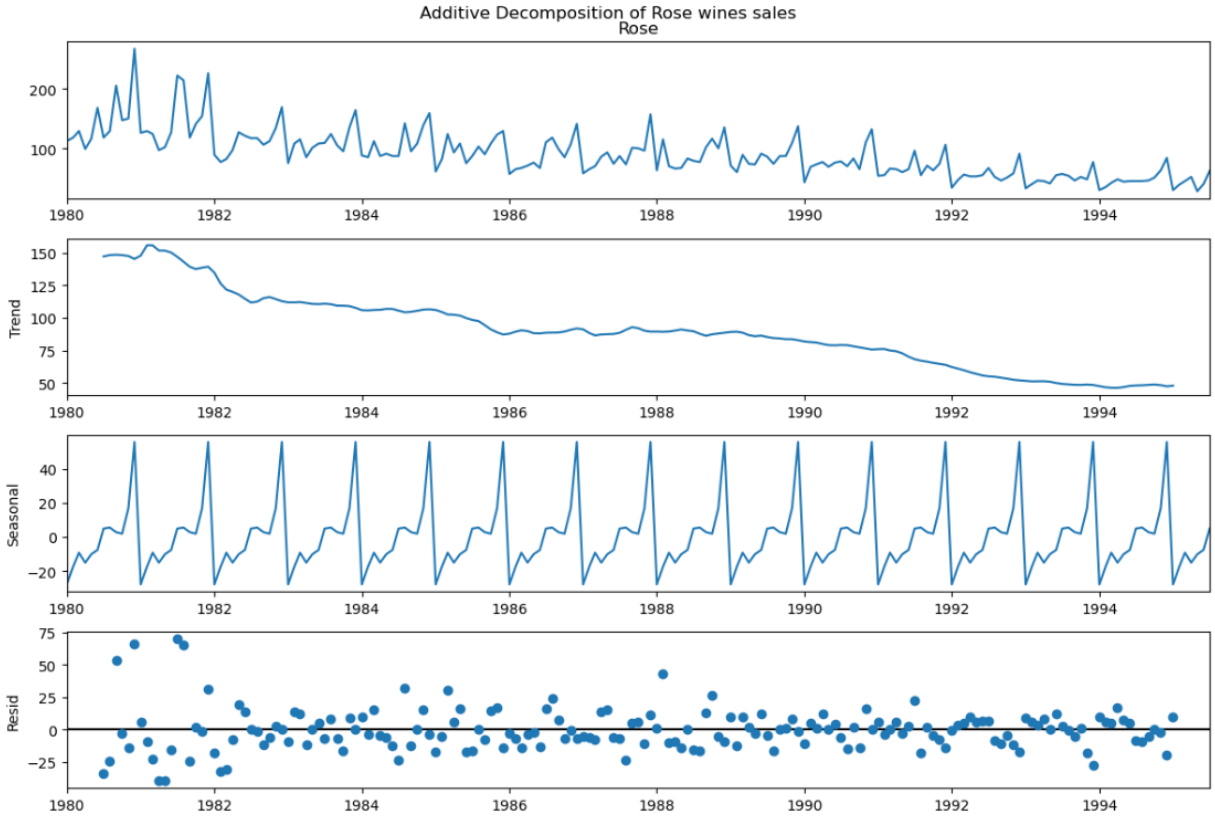


Figure7: Additive seasonal decomposition of Rose wines sales

- The initial glance on the time series plot shows us a significant decreasing trend.
- The data shows clear signs of seasonality.
- Wine sales seem to be increasing at the end of every year and low at the start of every year.
- The graph with the 'Trend' label shows the linear trend of the sales over the years.
- The seasonal component has also been captured in the third sub graph.
- The residual ideally should not show any pattern. But the residual graph for additive seasonal decomposition shows some pattern, which indicates that the data is not following additive seasonality. Seasonality must have been multiplicatively increasing over the years.

Multiplicative seasonal decomposition:

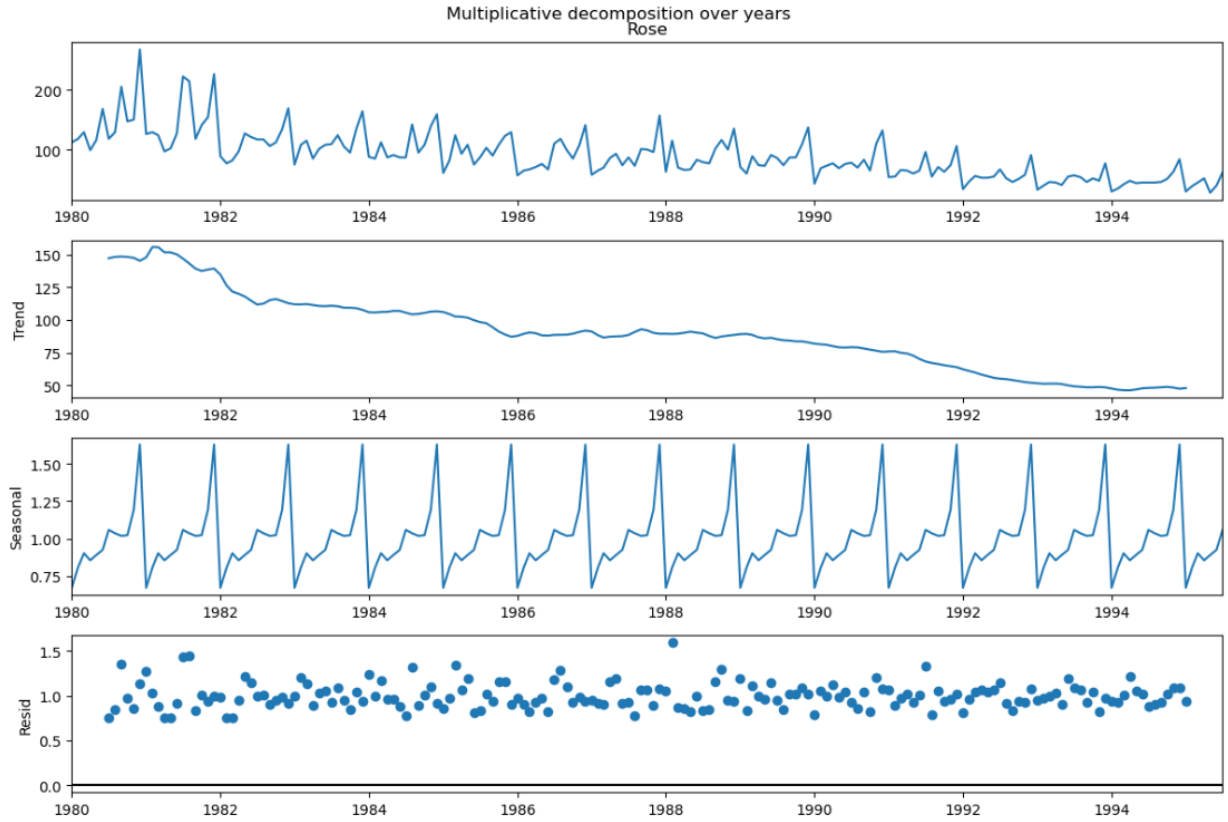


Figure8: Multiplicative seasonal decomposition of Rose wine sales

- Wine sales seem to be increasing at the end of every year and low at the start and mid of every year.
- The graph with the 'Trend' label shows the linear trend of the sales over the years.
- As described, it is decreasing.
- The seasonal component has also been captured in the third sub graph.
- The error or residual component does not show any pattern. It indicates that the data follows multiplicative seasonality.

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

First and last rows of training and test data set:

The data has been split into training and test data.
Date range for training time series: Jan-1980 to Dec-1990 (132 months)
Data range for test time series: Jan-1991 to July-1995 (55 months)

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table8: First and last 5 rows of training and test data set

4. **Build all the 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.**

4.1 Linear regression model:

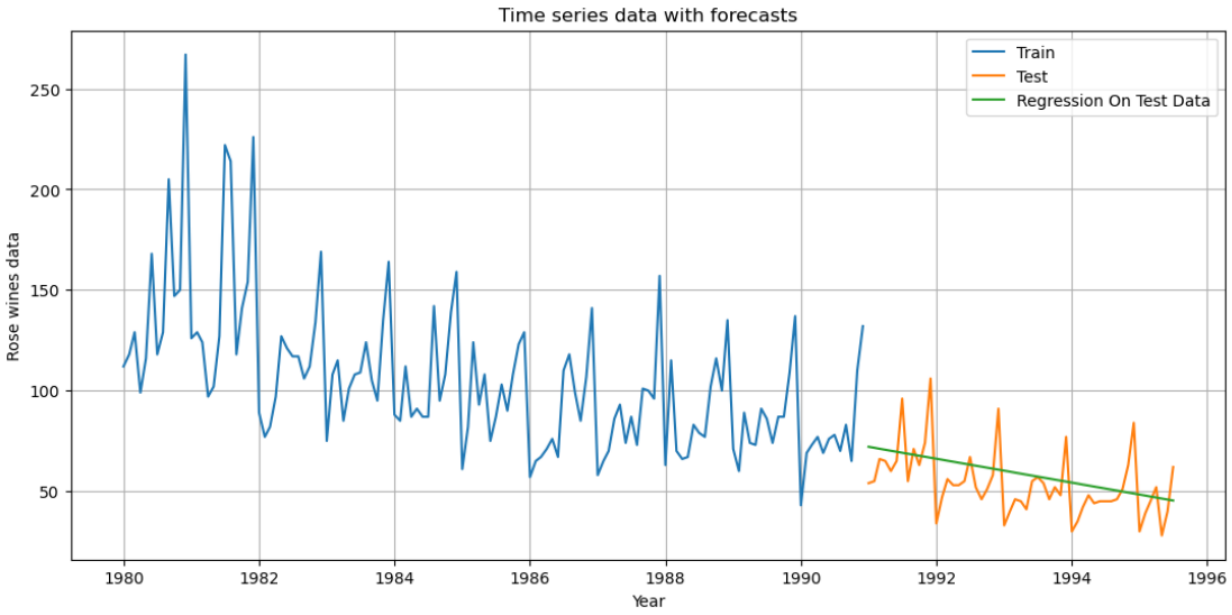
- A plain linear regression model has been applied to the training data set.
- The monthly index values of the training and test data have been converted into steps to be able to apply linear regression model.
- All default parameters are used.
    - fit_intercept=True,
    - normalize='deprecated',
    - copy_X=True,
    - n_jobs=None,
    - positive=False

- Predictions done by the linear regression model:

YearMonth	Rose	time	RegOnTime
1991-01-01	54.0	133	72.063266
1991-02-01	55.0	134	71.568888
1991-03-01	66.0	135	71.074511
1991-04-01	65.0	136	70.580133
1991-05-01	60.0	137	70.085755
1991-06-01	65.0	138	69.591377
1991-07-01	96.0	139	69.096999
1991-08-01	55.0	140	68.602621
1991-09-01	71.0	141	68.108243
1991-10-01	63.0	142	67.613866
1991-11-01	74.0	143	67.119488
1991-12-01	106.0	144	66.625110
1992-01-01	34.0	145	66.130732
1992-02-01	47.0	146	65.636354
1992-03-01	56.0	147	65.141976
1992-04-01	53.0	148	64.647598

Table9: Linear regression model predictions

- Root mean squared error for the predictions obtained on this model(on test data): 15.28
- Visualizing the line built by linear regression:



-

Figure9: Linear regression prediction plot on test data

- The equation built by linear regression model is 137.8 + (-0.49 * time_step)

4.2 Moving Average(MA) model:

- 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.
- The data constructed with different moving averages for the original is shown below

YearMonth	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1995-03-01	45.0	42.0	49.50	52.000000	49.777778
1995-04-01	52.0	48.5	41.50	52.166667	50.555556
1995-05-01	28.0	40.0	41.00	46.333333	48.666667
1995-06-01	40.0	34.0	41.25	39.000000	48.000000
1995-07-01	62.0	51.0	45.50	44.333333	49.222222

Table10: Moving average predictions on test data

The plots for different moving averages on the test set is shown below:

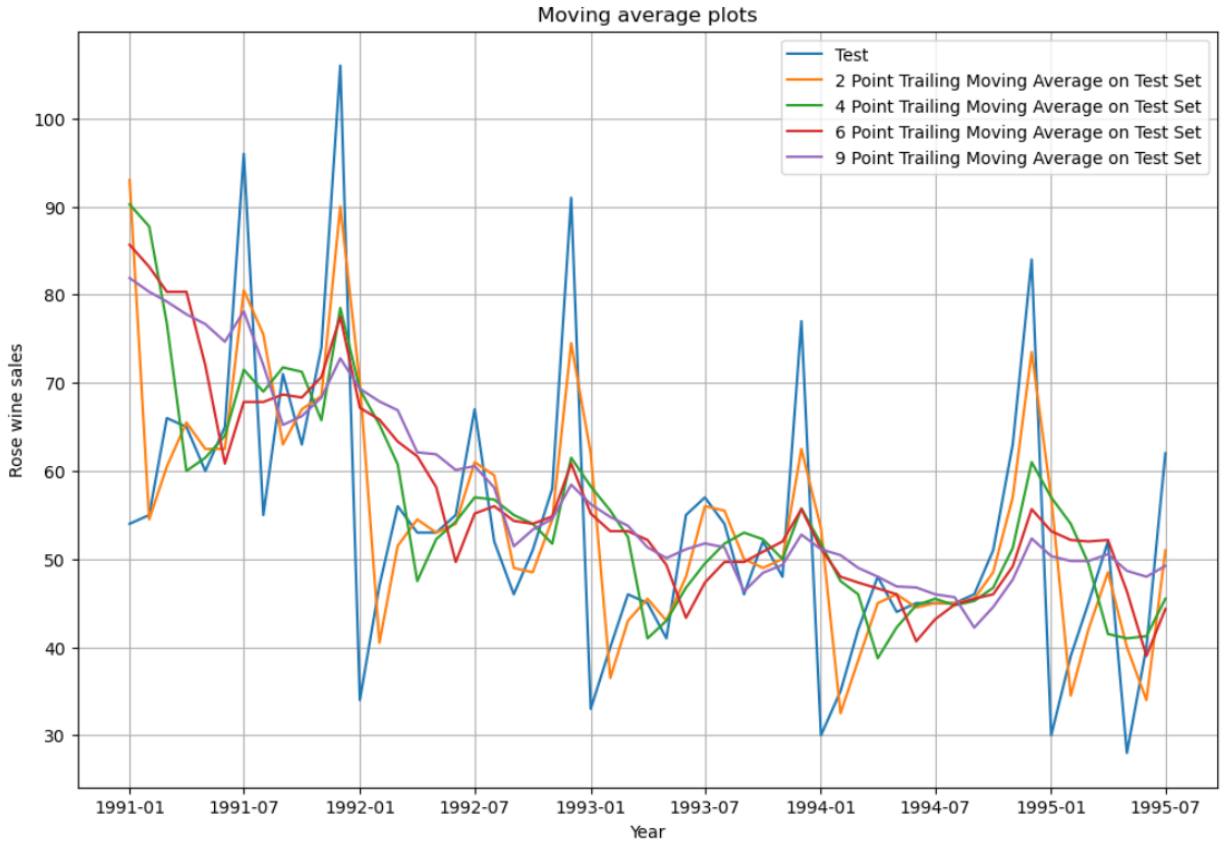


Figure10: Moving average prediction plots on test dataset

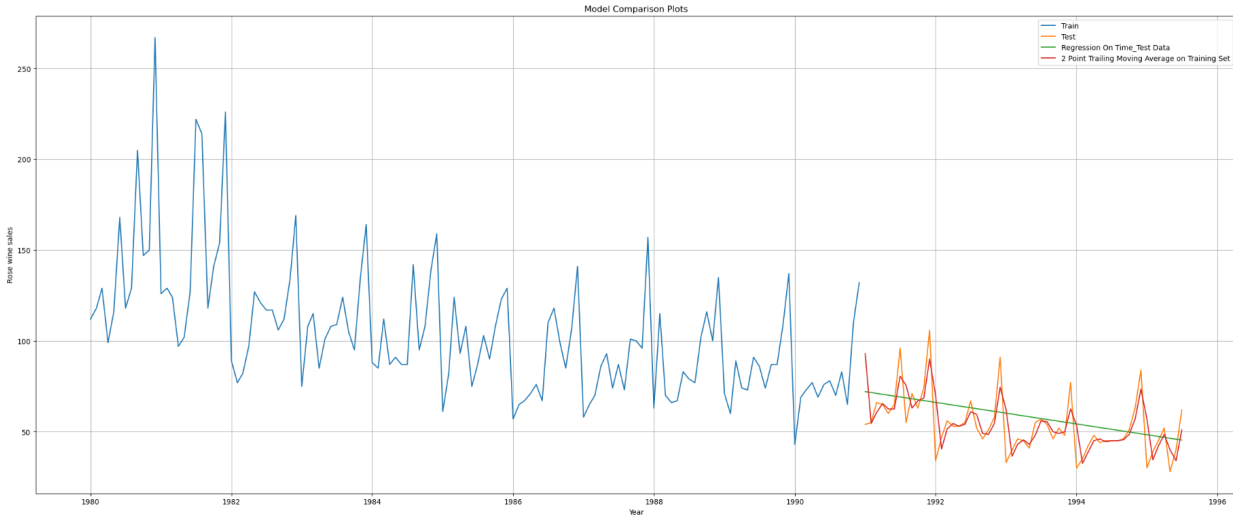
RMSE values for different averages:

	Test RMSE
RegressionOnTime	15.275732
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209

Table11: Moving average RMSE values on test data

- As we can see the 2 point trailing rolling average has the least error term compared to other rolling averages.
- This indicates that the data is mostly predictable using 2 most recent instances.

- Plot showing linear regression forecast and 2 point moving average forecast on test set:



-

Figure11: Linear regression and best moving average plot on test data

4.3 Single exponential smoothing:

Simple Exponential Smoothing is a time series forecasting method used to predict future values in a time series by giving more weight to recent observations while assigning exponentially decreasing weights to past observations. It is particularly useful for time series data that exhibit a trend or seasonality.

Mathematical Formula:

The formula for calculating the forecasted value using Simple Exponential Smoothing is as follows:

$$F_{t+1} = \alpha \cdot Y_t + (1 - \alpha) \cdot F_t$$

Where:

- F_{t+1} is the forecasted value for the next time period $t + 1$.
- Y_t is the actual value at time period t.
- F_t is the forecasted value for the current time period t.
- α is the smoothing parameter, a value between 0 and 1 that determines the weight given to the most recent observation. A smaller α gives more weight to past observations, while a larger α gives more weight to the most recent observation.

- Simple exponential smoothing has been applied with optimized approach and a brute force approach
- The value for alpha obtained with optimized approach: 0.098
- Other parameters obtained with optimized approach:
    - {'smoothing_level': 0.0987493111726833, (alpha value)
    - 'smoothing_trend': nan, (not a number since single exponential)
    - '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}
- The value of alpha obtained with brute force approach: 0.1
- RMSE values for the simple exponential smoothing applied on the test dataset with the above alpha values:

	Test RMSE
RegressionOnTime	15.275732
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
Alpha=0.098,SimpleExponentialSmoothing	36.816889
Alpha=0.1,SimpleExponentialSmoothing	36.848694

Table12: Test RMSE values for various models

Plotting the forecast obtained by above simple exponential smoothing equations:

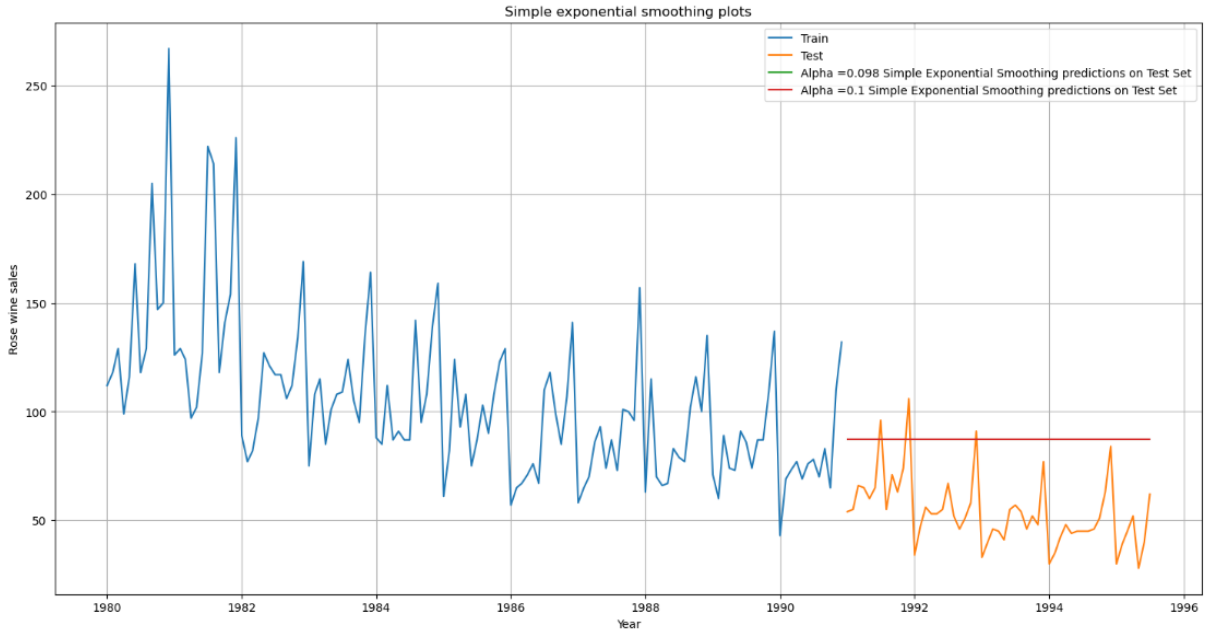


Figure12: Simple exponential smoothing model prediction plot on test data

- It is observed that the model performs the same with alpha values 0.098 and 0.1.
- The lines are the same and they have been overlapped in the above diagram.

4.4 Double exponential smoothing

Double Exponential Smoothing, also known as Holt's Linear Exponential Smoothing, is an extension of Simple Exponential Smoothing that takes into account both the level and trend of a time series. This method is particularly useful when the time series exhibits a linear trend.

Mathematical Formula:
The formula for calculating the forecasted value using Double Exponential Smoothing is as follows:

1. **Level Smoothing:**

   $$L_t = \alpha \cdot Y_t + (1 - \alpha) \cdot (L_{t-1} + T_{t-1})$$

   Where:

   - L_t is the smoothed level (or the estimate of the level) at time t.
   - Y_t is the actual value at time t.
   - L_{t-1} is the smoothed level at time $t - 1$.
   - T_{t-1} is the trend at time $t - 1$.
   - α is the smoothing parameter for the level, similar to Simple Exponential Smoothing.

2. **Trend Smoothing:**

   $$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}$$

   Where:

   - T_t is the smoothed trend (or the estimate of the trend) at time t.
   - β is the smoothing parameter for the trend.

3. **Forecast Calculation:**

   $$F_{t+h} = L_t + h \cdot T_t$$

   Where:

   - F_{t+h} is the forecasted value for h periods beyond time t.

- Double exponential smoothing has been applied with optimized approach and a brute force approach
- The value for alpha obtained with optimized approach: 0.0175
- The value for beta obtained with optimized approach: 0.00003
- Parameters obtained with optimal approach:
    - {'smoothing_level': 0.017549790270679714,
    - 'smoothing_trend': 3.236153800377395e-05,
    - 'smoothing_seasonal': nan,
    - 'damping_trend': nan,
    - 'initial_level': 138.82081494774005,
    - 'initial_trend': -0.492580228245491,
    - 'initial_seasons': array([], dtype=float64),
    - 'use_boxcox': False,
    - 'lamda': None,
    - 'remove_bias': False}
- The value of alpha obtained with brute force approach: 0.1
- The value of beta obtained with brute force approach: 0.1

- RMSE values for the simple exponential smoothing applied on the dataset with the above alpha values:

	Test RMSE
RegressionOnTime	15.275732
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
Alpha=0.098,SimpleExponentialSmoothing	36.816889
Alpha=0.1,SimpleExponentialSmoothing	36.848694
Alpha =0.0175, Beta = 0.00003,DoubleExponentialSmoothing	36.816889
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.944741

Table13: Test RMSE values for various models

Plotting the forecast obtained by above simple exponential smoothing equations:

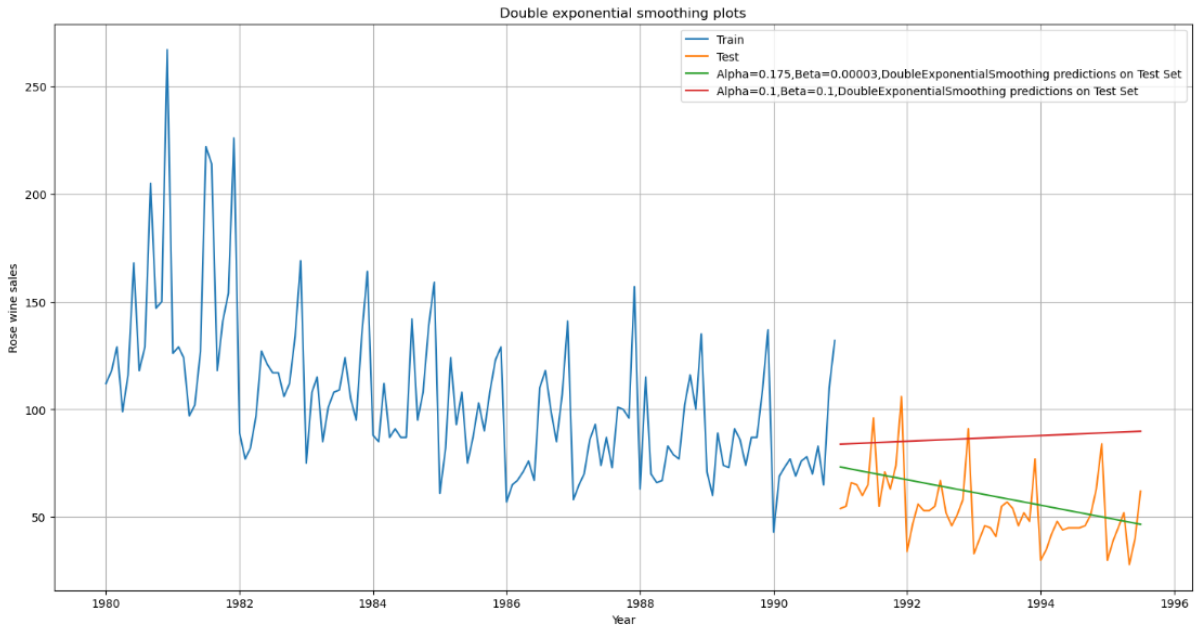


Figure13: Double exponential smoothing model prediction plots on test data

- It is observed that the model performs the same with above 2 specified alpha and beta combinations.

- RMSE value is around 36 for double exponential smoothing model.
- However, till the current status, 2 point moving average seems to give the best performance.

4.5 Triple exponential smoothing:

Triple Exponential Smoothing, also known as Holt-Winters Exponential Smoothing, is an extension of Double Exponential Smoothing that includes a seasonality component. This method is particularly useful when the time series exhibits trend and seasonality.

Mathematical Formula:
The formula for calculating the forecasted value using Triple Exponential Smoothing is as follows:

1. **Level Smoothing:**
   $$L_t = \alpha \cdot (Y_t - S_{t-m}) + (1 - \alpha) \cdot (L_{t-1} + T_{t-1})$$
2. **Trend Smoothing:**
   $$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}$$
3. **Seasonal Smoothing:**
   $$S_t = \gamma \cdot (Y_t - L_t) + (1 - \gamma) \cdot S_{t-m}$$
4. **Forecast Calculation:**
   $$F_{t+h} = L_t + h \cdot T_t + S_{t-m+h_m}$$

Where:

- L_t is the smoothed level (or the estimate of the level) at time t.
- T_t is the smoothed trend (or the estimate of the trend) at time t.
- S_t is the smoothed seasonal component (or the estimate of the seasonality) at time t.
- Y_t is the actual value at time t.
- S_{t-m} represents the seasonal component at time $t - m$, where m is the number of seasons (e.g., for monthly data, $m = 12$ for yearly seasonality).
- h is the number of periods ahead for forecasting.
- h_m is the corresponding seasonal index for h.

- Triple exponential smoothing has been applied with optimized approach and a brute force approach
- The value for alpha obtained with optimized approach: 0.715
- The value for beta obtained with optimized approach: 0.045
- The value for beta obtained with optimized approach: 0.00007
- Hyper parameters for the above optimized approach are:
    - {'smoothing_level': 0.0715106306609405,
    - 'smoothing_trend': 0.04529179757535142,
    - 'smoothing_seasonal': 7.244325029450242e-05,
    - 'damping_trend': nan,
    - 'initial_level': 130.40839142502193,
    - 'initial_trend': -0.77985743179386,
    - 'initial_seasons': array([0.86218996, 0.977675 , 1.0687727 , 0.93403881, 1.050625 ,
    - 1.14410977, 1.25836944, 1.33937772, 1.26778766, 1.24131254,
    - 1.44724625, 1.99553681]),
    - 'use_boxcox': False,
    - 'lamda': None,
    - 'remove_bias': False}

- The value of alpha obtained with brute force approach: 0.1
- The value of beta obtained with brute force approach: 0.2
- The value of beta obtained with brute force approach: 0.1
- RMSE values for the simple exponential smoothing applied on the dataset with the above alpha values:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.236464
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
Alpha=0.0715,Beta=0.045,Gamma=0.00007,TripleExponentialSmoothing	20.182721
Alpha=0.098,SimpleExponentialSmoothing	36.816889
Alpha =0.0175, Beta = 0.00003,DoubleExponentialSmoothing	36.816889
Alpha=0.1,SimpleExponentialSmoothing	36.848694
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.944741

-

Table14: Test RMSE values for various models

Plotting the forecast obtained by above simple exponential smoothing equations:

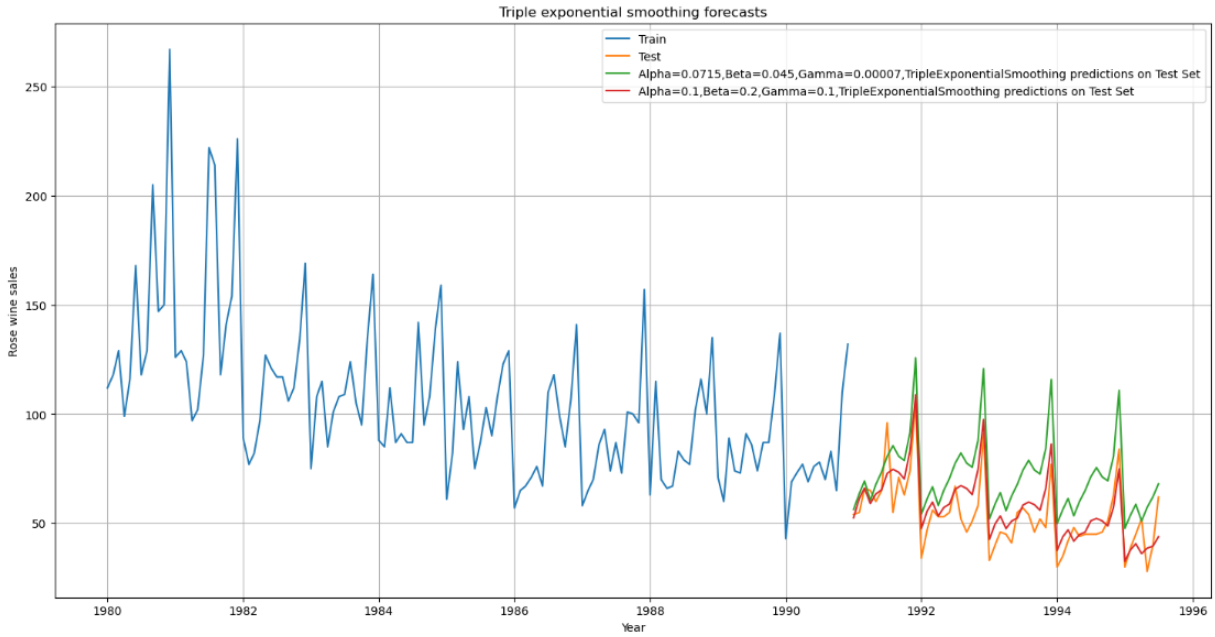


Figure14: Triple exponential smoothing model prediction plots on test data

- It is observed that the model performs better when alpha - 0.1, beta - 0.2 and gamma - 0.1 with RMSE value of 9.236

Plotting the best variants of all exponential smoothing models built till now:

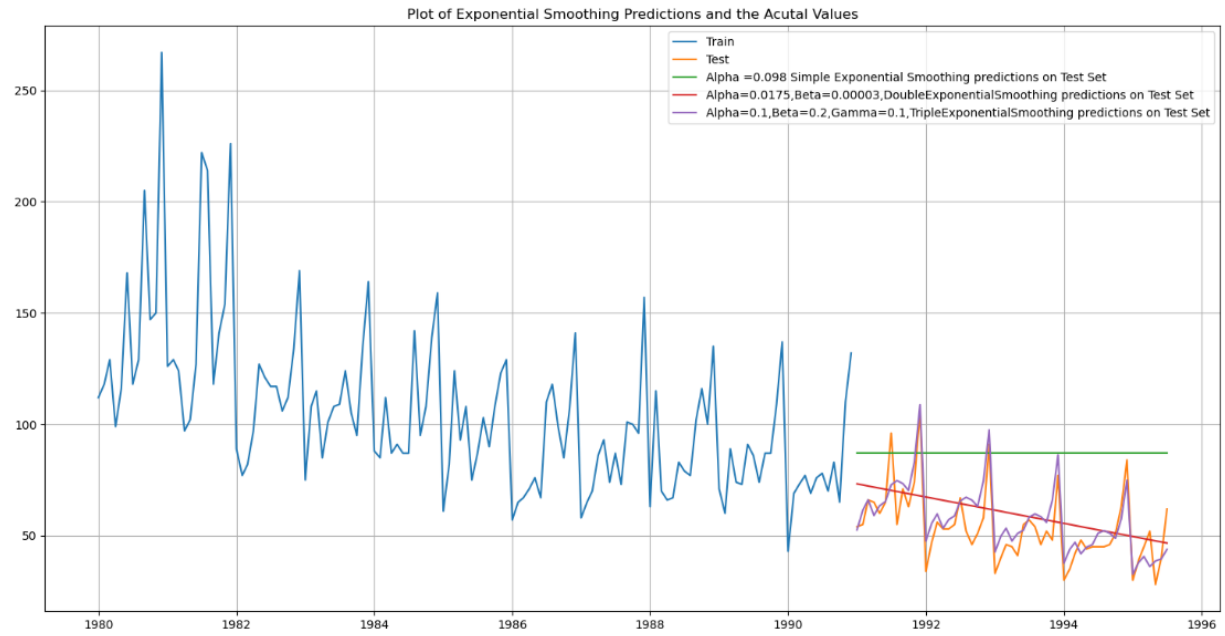


Figure15: All smoothing model prediction plots on test data

- Till the current status, triple exponential smoothing with alpha - 0.1, beta - 0.2 and gamma - 0.1 seems to give the best performance with RMSE 9.236

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**

- Models like ARIMA and SARIMA work on data that is stationary. Data is said to be stationary when it shows no trend and significant variance in the distribution of data over time.
- If data is found to be non-stationary we use differencing to make the data stationary, if the data shows significant variance, we use transformations like log transformation in order to stabilize the magnitude and variance of the data.
- Dicky Fuller Test on the timeseries is run to check for stationarity of data.

- **Null Hypothesis** *H*0: Time Series is non-stationary.
- **Alternate Hypothesis** ***Ha***: Time Series is stationary.

- *So Ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected .*

- Applying a dickey-fuller test on the data gives the following results. (considering 0.05 to be the level of confidence)
  - Test Statistic -1.874856
  - p-value 0.343981
  - #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

- P-value is > 0.05 which means we fail to reject null hypothesis. The data is non stationary.

- Applying first order differencing on the data and verifying for stationarity (considering 0.05 to be the level of confidence)
  - Test Statistic -8.044139e+00
  - p-value 1.813580e-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

- P-value obtained now is ~0 which means we reject null hypothesis.
- Applied first order differencing on the data and the data is now stationary.

<u>Plot of non stationary time series data</u>:

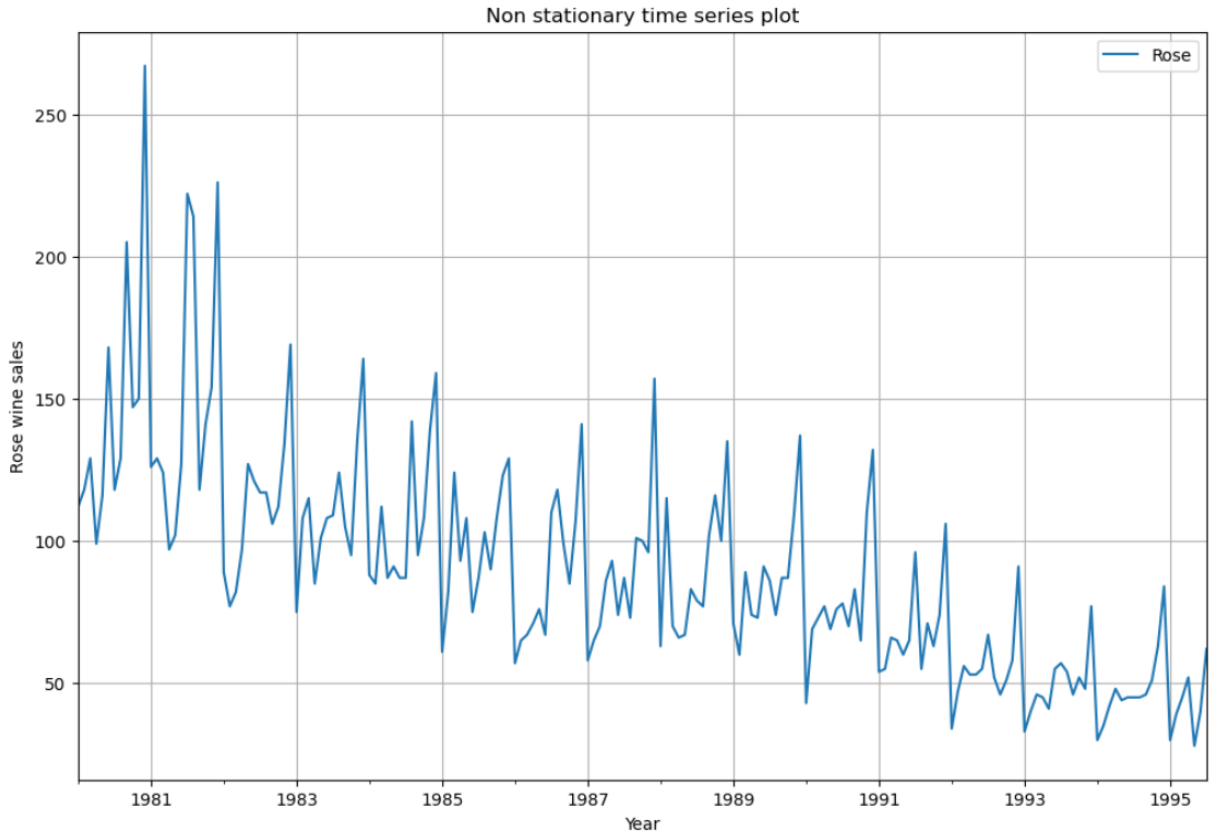


Figure16: Non stationary time series plot

Plot of first order differenced time series data:

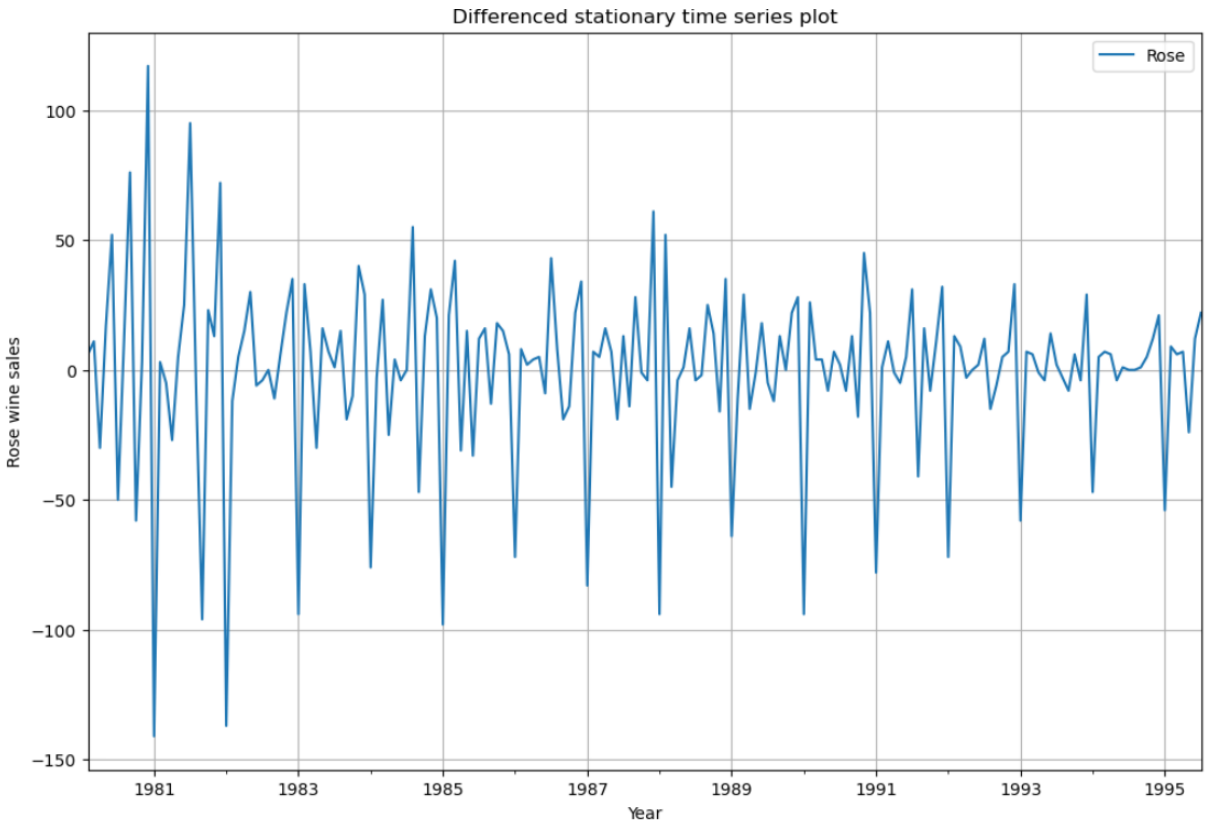


Figure17: First order differenced time series plot

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.**

Prerequisite: The training data has to be determined to be stationary or not.

- Given training time series data is non stationary.
- Plotting the distribution of non stationary time series data:

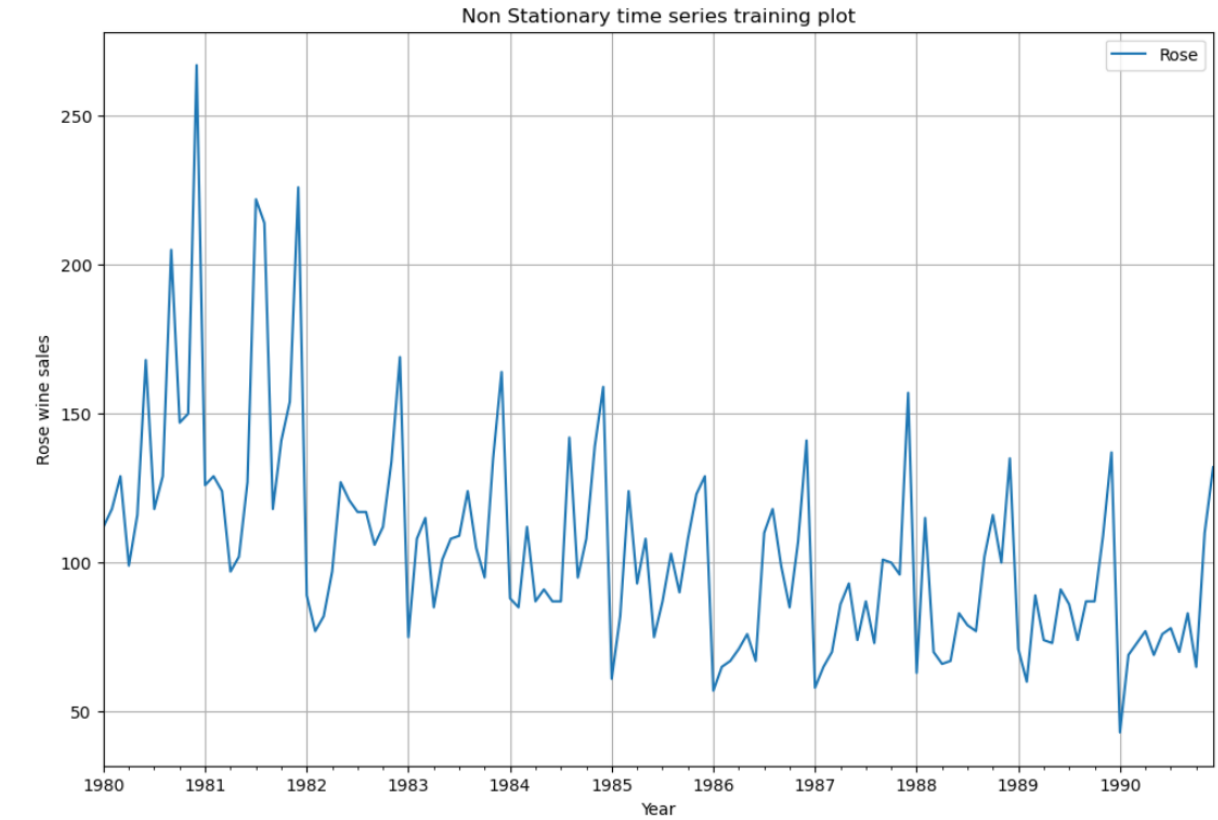


Figure18: Non stationary time series training plot

- First and last rows of training and test data:

First few rows of Training Data

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

First few rows of Test Data

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Training Data

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

Last few rows of Test Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table15: Training and test data

- Applying dickey-fuller test on training data:

- The p-value obtained is > 0.05 which means the training time series is non stationary. Need to make the data stationary in order for ARIMA and SARIMA models to use it.
- Applied first order differencing on the training data.
- Applied dickey - fuller test on training data after differencing.
    - Test Statistic -6.592372e+00
    - p-value 7.061944e-09
    - #Lags Used 1.200000e+01
    - Number of Observations Used 1.180000e+02
    - Critical Value (1%) -3.487022e+00
    - Critical Value (5%) -2.886363e+00
    - Critical Value (10%) -2.580009e+00
    - dtype: float64

- Since p-value is < 0.05, we reject null hypothesis and time series is considered to be stationary.
- Differenced stationary time series training plot:

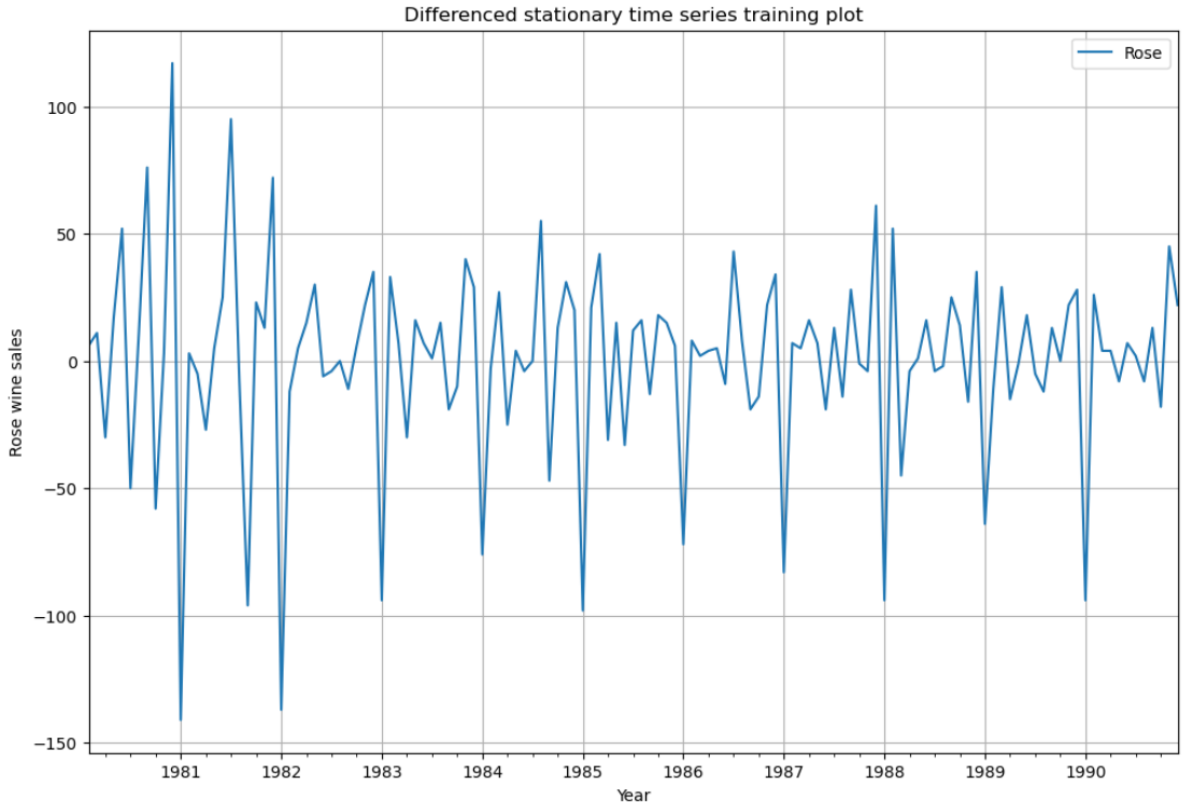


Figure19: Differenced time series stationary training plot

- We use the training dataset directly because we are using ARIMA and SARIMA models which have an inbuilt option to difference the time series.

ARIMA model:

- ARIMA:- **Auto Regressive Integrated Moving Average** is a way of modeling time series data for forecasting or predicting future data points.
- Improving AR Models by making Time Series stationary through Moving Average Forecasts
- ARIMA models consist of 3 components:-
  - **AR model:** The data is modeled based on past observations.
  - **Integrated component:** Whether the data needs to be differenced/transformed.
  - **MA model:** Previous forecast errors are incorporated into the model.

Figure20: ARIMA model

- For the given data set, we consider p,d,q to represent level component, differencing component and trend component respectively.
    - p: The number of lag observations included in the model (order of autoregressive terms).
    - d: The degree of differencing, which is the number of times the data is differenced (order of integration).
    - q: The size of the moving average window (order of moving average terms).
- Ranges considered for p,d,q are (0,5).
- An automated ARIMA model has been built with the following combinations/parameters.
    - AR component(p): Range(0,5)
    - MA component(q): Range(0,5)
    - Differencing component(d): 1
    - Enforce_stationarity: False (not to enforce stationarity on the AR components of the model)
    - Enforce_intvertibility: False(not to enforce invertibility on the MA components on the model)
- It notes the AIC scores for the data with each set of p,q,d values.
- The top 5 combinations with least AIC scores obtained are

	param	AIC
58	(2, 1, 3)	1274.696103
37	(1, 2, 2)	1278.058579
14	(0, 2, 4)	1278.431001
108	(4, 1, 3)	1278.451407
39	(1, 2, 4)	1278.606955

Table16: ARIMA model best AIC values

- Therefore, best hyper parameters for p,d,q for the given data set are taken as 2,1,3
- AIC metric for this combination is 1274.696

- Applying ARIMA on the model with the above obtained parameters.
- Summary obtained:

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.348
Date:	Sat, 05 Aug 2023	AIC	1274.696
Time:	17:31:39	BIC	1291.947
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.706
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6773	0.084	-19.962	0.000	-1.842	-1.513
ar.L2	-0.7281	0.084	-8.661	0.000	-0.893	-0.563
ma.L1	1.0460	0.684	1.530	0.126	-0.294	2.386
ma.L2	-0.7698	0.138	-5.565	0.000	-1.041	-0.499
ma.L3	-0.9037	0.620	-1.458	0.145	-2.119	0.311
sigma2	859.7747	576.982	1.490	0.136	-271.090	1990.640

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.16
Prob(Q):	0.89	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.70
Prob(H) (two-sided):	0.00	Kurtosis:	4.56

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

-

Figure21: ARIMA model summary on Rose Wine sales

- The equation obtained by ARIMA model can be written as follows:

ARIMA(2,1,3)

$= (-1.6773 * Yt-1 - 0.7281 * Yt-2$

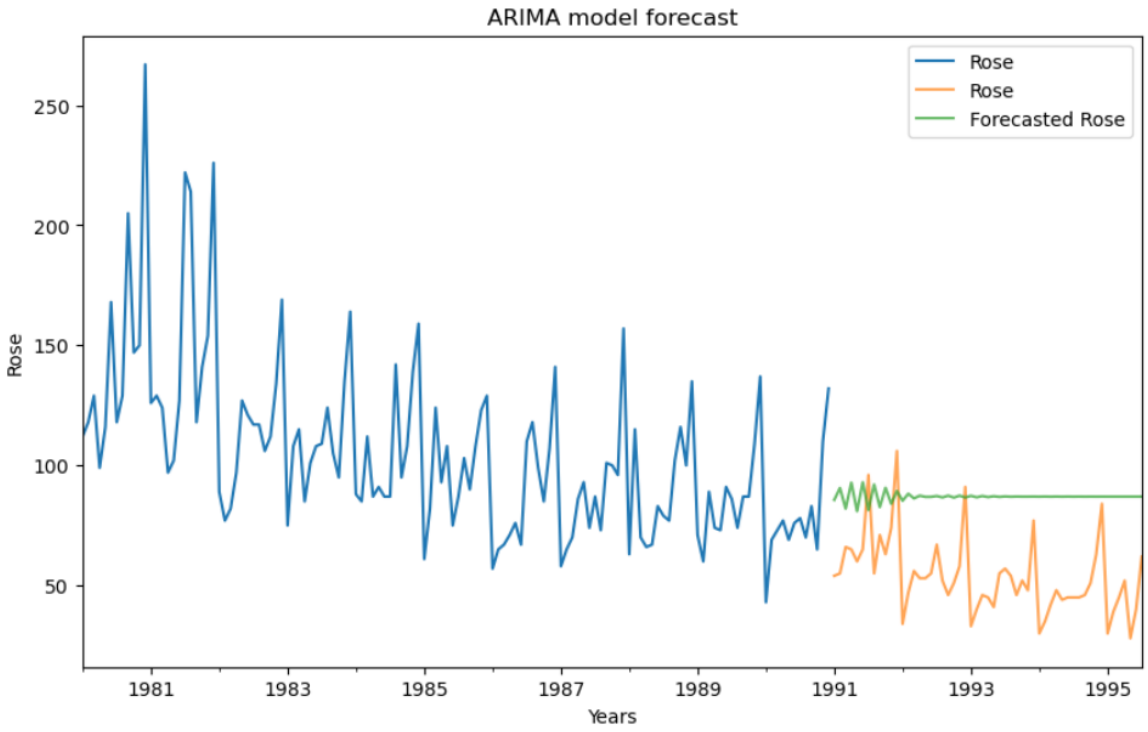
- $(+1.0460 * et-1 - 0.7698 * et-2 - 0.9037 * et-3$

- A sample of the forecasted values on the actual test data is shown below:

YearMonth	Rose	Rose_forecasted
1991-01-01	54.0	85.624392
1991-02-01	55.0	90.570708
1991-03-01	66.0	81.982766
1991-04-01	65.0	92.786210
1991-05-01	60.0	80.917993

Table17: Sample of forecasted values by ARIMA model

- Root mean squared error for the above built ARIMA model is 36.858
- Forecast plot by ARIMA model on test data:



-

Figure22: ARIMA model prediction forecast

- Table showing all models built till now and their best RMSE values:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.236464
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
Alpha=0.0715,Beta=0.045,Gamma=0.00007,TripleExponentialSmoothing	20.182721
Alpha=0.098,SimpleExponentialSmoothing	36.816889
Alpha =0.0175, Beta = 0.00003,DoubleExponentialSmoothing	36.816889
Alpha=0.1,SimpleExponentialSmoothing	36.848694
Best ARIMA Model : ARIMA(2,1,3)	36.858265
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.944741

Table18: Sorted RMSE values for all forecasting models

- It is observed that the ARIMA model has captured the test data but has failed to accurately predict the test values when compared to 2-point moving average or exponential smoothing models.
- This is because the dataset has shown string seasonality but ARIMA models do not consider seasonality in training and forecasting.

SARIMA model:

- The ARIMA models can be extended/improved to handle seasonal components of a data series.
- The seasonal autoregressive moving average model is given by: **SARIMA (p, d, q)(P, D, Q)F**
- The above model consists of:
    - Autoregressive and moving average components (p, q)
    - Seasonal autoregressive and moving average components (P, Q)
    - The ordinary and seasonal difference components of order ‘d’ and ‘D’
    - Seasonal frequency ‘F’
- - The value for the parameters (p,d,q) and (P, D, Q) can be decided by comparing different values for each and taking **the lowest AIC value** for the model build.

- The value for F can be consolidated by ACF plot.

Let's plot an ACF plot on the given data:

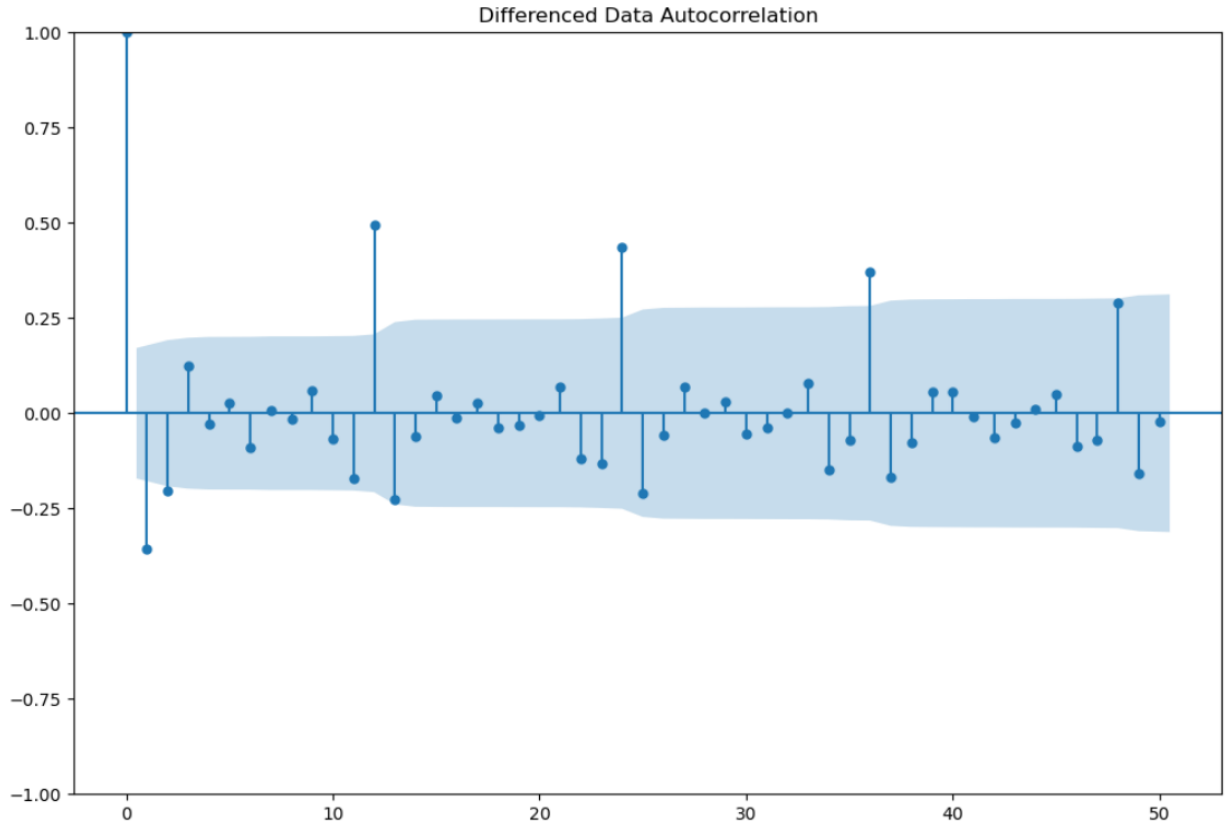


Figure23: ACF plot for Rose wine sales data

- ACF plots help us understand the correction of time series values with their own past values.
- From the ACF plot, we observed there is a trend in the data which is not significantly increasing or decreasing.
- It is also observed that there is strong correlation among the time series values at regular lag intervals. For example at lag = 6, there is strong correlation.
- This indicates that there is strong seasonality in the data that needs to be considered when choosing our final model.

- SARIMA model also takes into account the seasonality along with the AR, MA and differencing terms.
- An automated SARIMA model has been built with the following combinations/parameters.
    - AR component(p/P): Range(1,4)
    - MA component(q/Q): Range(2,5)

- Differencing component(d): (0,2)
- Seasonal differencing component: (0,2)
- Seasonal component(F): (6)
- Enforce_stationarity: False (not to enforce stationarity on the AR components of the model)
- Enforce_intvertibility: False(not to enforce invertibility on the MA components on the model)

Results with SARIMA model having seasonality = 6:

- It notes the AIC scores for the data with each set of (p,q,d)(P,Q,D,F) values.
- The top 5 combinations with least AIC scores obtained are

	param	seasonal	AIC
39	(1, 1, 5)	(1, 1, 5, 6)	750.810359
43	(1, 1, 5)	(2, 1, 5, 6)	751.940253
47	(1, 1, 5)	(3, 1, 5, 6)	752.175766
87	(2, 1, 5)	(1, 1, 5, 6)	752.689194
91	(2, 1, 5)	(2, 1, 5, 6)	753.690108

Table19: SARIMA model best AIC values

- Therefore, best hyper parameters for p,d,q for the given training data set are taken as (1,1,5) (1,1,5,6)
- AIC metric for this combination is 750.81

- Applying SARIMA on the model with the above obtained parameters.
- Summary obtained:

SARIMAX Results

Dep. Variable:	y		No. Observations:		132
Model:	SARIMAX(1, 1, 5)x(1, 1, 5, 6)		Log Likelihood		-362.405
Date:	Sun, 06 Aug 2023		AIC		750.810
Time:	11:34:23		BIC		783.163
Sample:	0		HQIC		763.851
	- 132				
Covariance Type:	opg				

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7447	0.189	-3.933	0.000	-1.116	-0.374
ma.L1	-0.9507	0.762	-1.248	0.212	-2.444	0.543
ma.L2	-1.2555	0.616	-2.037	0.042	-2.463	-0.048
ma.L3	0.2020	0.597	0.338	0.735	-0.968	1.372
ma.L4	0.1888	0.380	0.497	0.619	-0.555	0.933
ma.L5	0.6126	0.295	2.073	0.038	0.034	1.192
ar.S.L6	-0.9678	0.017	-56.240	0.000	-1.002	-0.934
ma.S.L6	0.1206	54.961	0.002	0.998	-107.601	107.842
ma.S.L12	-0.8339	61.436	-0.014	0.989	-121.246	119.578
ma.S.L18	0.1862	15.704	0.012	0.991	-30.593	30.966
ma.S.L24	-0.2257	25.921	-0.009	0.993	-51.030	50.579
ma.S.L30	-0.2479	13.547	-0.018	0.985	-26.800	26.304
sigma2	62.2028	3417.630	0.018	0.985	-6636.230	6760.635

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	2.90
Prob(Q):	0.87	Prob(JB):	0.23
Heteroskedasticity (H):	0.89	Skew:	0.38
Prob(H) (two-sided):	0.75	Kurtosis:	3.44

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

Figure24: SARIMA model summary on Rose wine sales

- The equation obtained by SARIMA model can be written as follows:
- The terms ar.L1, ar.L2, ma.L1, ma.L2, ma.L3, ma.L4 indicate the orders of AR and MA components of the actual data. (p and q values)
- The terms ar.S.L6, ar.S.L12, ma.S.L6, ma.S.L12. ma.S.L18, ma.S.L24 indicate the AR and MA components of the previous seasonal data. (P and Q values)

- Root mean squared error for the above built ARIMA model is **15.31**
- Table showing all models built till now and their best RMSE values:

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.236464
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
SARIMA(1,1,1)(1,1,5,6)	15.308698
Alpha=0.0715,Beta=0.045,Gamma=0.00007,TripleExponentialSmoothing	20.182721
Alpha=0.098,SimpleExponentialSmoothing	36.816889
Alpha =0.0175, Beta = 0.00003,DoubleExponentialSmoothing	36.816889
Alpha=0.1,SimpleExponentialSmoothing	36.848694
Best ARIMA Model : ARIMA(2,1,3)	36.858265
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.944741

Table19: Test RMSE values of all forecasting models

- Sample of the predicted values using SARIMA forecast:

:

YearMonth	Rose	SARIMA forecasted
1991-01-01	54.0	50.824747
1991-02-01	55.0	58.995885
1991-03-01	66.0	64.457852
1991-04-01	65.0	64.225173
1991-05-01	60.0	73.880947

Table20: Sample of forecasted values by SARIMA model

- SARIMA model prediction plot on test data:

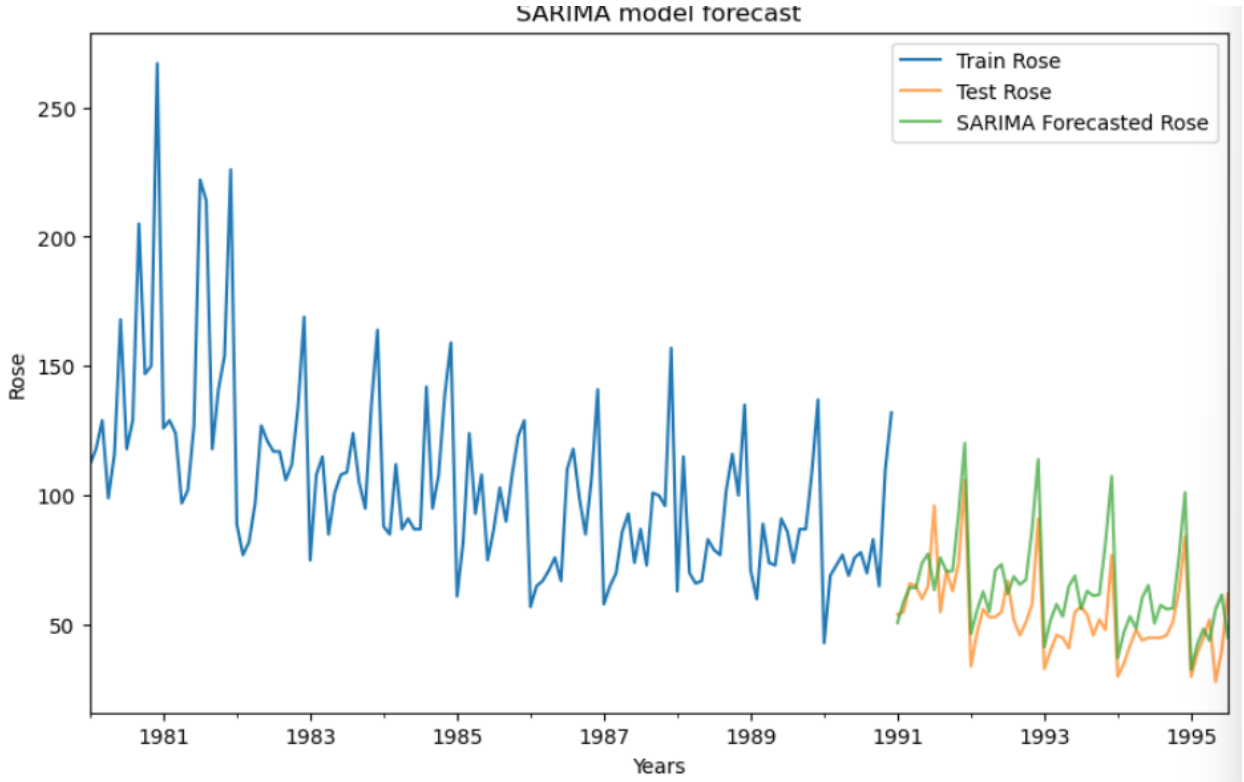


Figure24: SARIMA model prediction plot on test data

- The diagnostics for the above built SARIMA model with seasonality = 6 is:

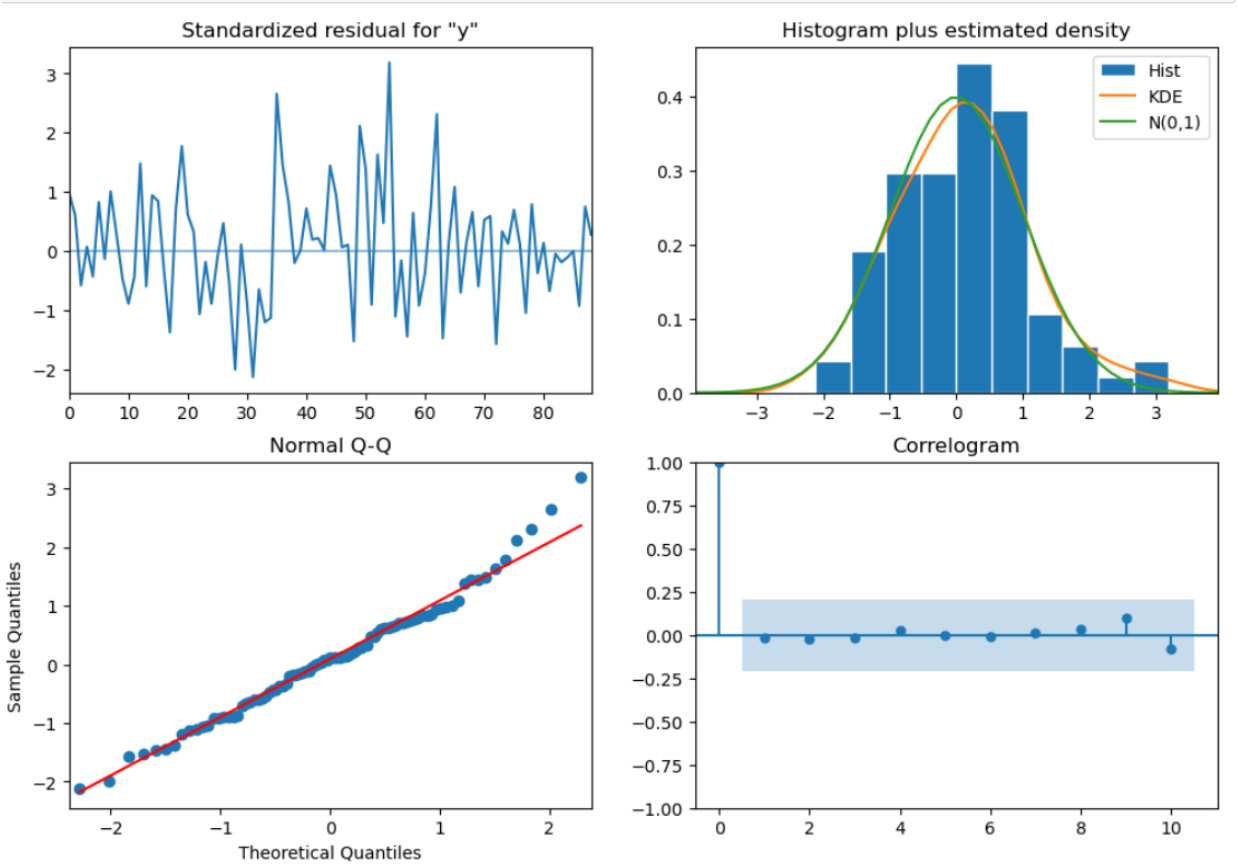


Figure25: SARIMA model diagnostics plot

- It is observed that the SARIMA model has captured the test data more accurately than the ARIMA model as we have included seasonal components to forecast the values. This made the forecasts more accurately predictable.

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

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.236464
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
SARIMA(1,1,1)(1,1,5,6)	15.308698
Alpha=0.0715,Beta=0.045,Gamma=0.00007,TripleExponentialSmoothing	20.182721
Alpha=0.098,SimpleExponentialSmoothing	36.816889
Alpha =0.0175, Beta = 0.00003,DoubleExponentialSmoothing	36.816889
Alpha=0.1,SimpleExponentialSmoothing	36.848694
Best ARIMA Model : ARIMA(2,1,3)	36.858265
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.944741

Table21: Table showing Test RMSE values of all best forecasting models

- The above indicates all models used for forecasting Rose wines sales with parameters specified.
- The data has been sorted according to the ascending RMSE values.
- It is evident that the Triple exponential smoothing model with parameters alpha = 0.1, beta = 0.2, gamma = 0.1 renders the most optimal model with the best accuracy among all.

8**. 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.**

- Since, triple exponential smoothing is the most optimal model, constructed a triple exponential model on full data.

- Parameters used: alpha = 0.1, beta = 0.2, gamma = 0.1
- RMSE obtained for the model built: 17.025
- This model is used to forecast the values for the next 12 months from the last month of the dataset.
- The forecast ranges from ‘1995-08-01’ to ‘1996-07-01’
- Below is the forecasted data:

```
1995-08-01    49.964554
1995-09-01    49.842233
1995-10-01    50.806489
1995-11-01    59.168191
1995-12-01    82.325490
1996-01-01    33.706342
1996-02-01    40.784007
1996-03-01    46.085277
1996-04-01    44.929079
1996-05-01    43.085914
1996-06-01    48.007422
1996-07-01    54.866211
Freq: MS, dtype: float64
```

Table22: Table showing forecasted Rose wine sales for next 12 months (from Aug 1995)

- The forecast along with current data looks like below:

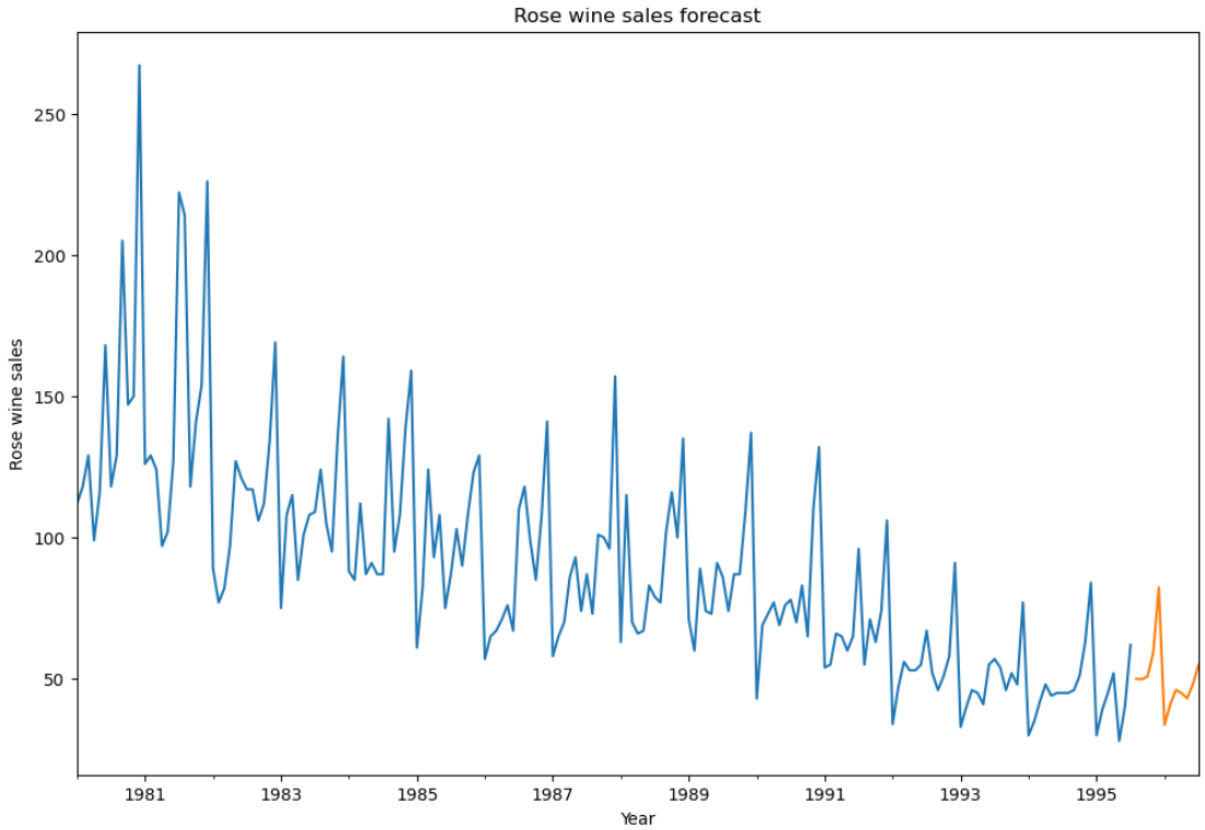
- 

Figure26: Time series forecast plot for next year data with current data

- Plotting the range of forecasted values for the next 12 months with a confidence interval of 95%

	lower_CI	prediction	upper_ci
1995-08-01	16.512944	49.964554	83.416165
1995-09-01	16.390623	49.842233	83.293844
1995-10-01	17.354879	50.806489	84.258100
1995-11-01	25.716580	59.168191	92.619801
1995-12-01	48.873880	82.325490	115.777101
1996-01-01	0.254731	33.706342	67.157952
1996-02-01	7.332396	40.784007	74.235618
1996-03-01	12.633667	46.085277	79.536888
1996-04-01	11.477468	44.929079	78.380690
1996-05-01	9.634303	43.085914	76.537525
1996-06-01	14.555811	48.007422	81.459032
1996-07-01	21.414601	54.866211	88.317822

Table23: Forecasted values with 95% confidence intervals

- The plot for the same is shown below:

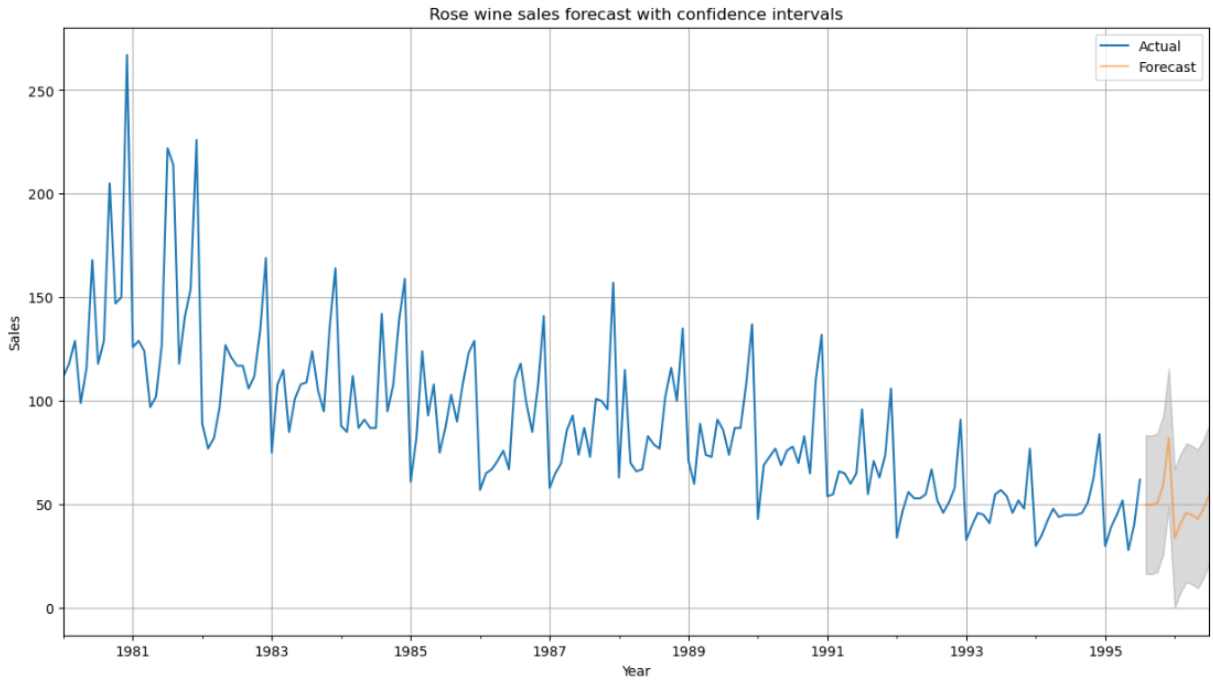


Figure27: Forecast plot for next 12 months data with confidence intervals

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

Ans:

- The given dataset depicts Rose wines sales on a monthly basis for the years 1908-1995.
- The same has been read in proper datetime format. Exploratory data analysis has been done on the project.
- It showed that the data has two missing values.
- The data has been decomposed. There is a significant downward trend observed but a significant multiplicative seasonality observed in the data.
- The sales have been decreasing constantly over the years.
- Several models like Linear Regression, Exponential smoothing methods, Moving average methods, ARIMA, SARIMA have been trained with the dataset given.
- Of all the models, the models which account for seasonality have performed better compared to other models.
- The RMSE scores obtained for all the models built are shown below.

	lower_CI	prediction	upper_ci
1995-08-01	16.512944	49.964554	83.416165
1995-09-01	16.390623	49.842233	83.293844
1995-10-01	17.354879	50.806489	84.258100
1995-11-01	25.716580	59.168191	92.619801
1995-12-01	48.873880	82.325490	115.777101
1996-01-01	0.254731	33.706342	67.157952
1996-02-01	7.332396	40.784007	74.235618
1996-03-01	12.633667	46.085277	79.536888
1996-04-01	11.477468	44.929079	78.380690
1996-05-01	9.634303	43.085914	76.537525
1996-06-01	14.555811	48.007422	81.459032
1996-07-01	21.414601	54.866211	88.317822

Table24: Forecasted values with 95% confidence intervals

- The best model for forecasting sales for Rose wines data is triple exponential smoothing.
- The forecast range for the sales of next 12 months has obtained as below:

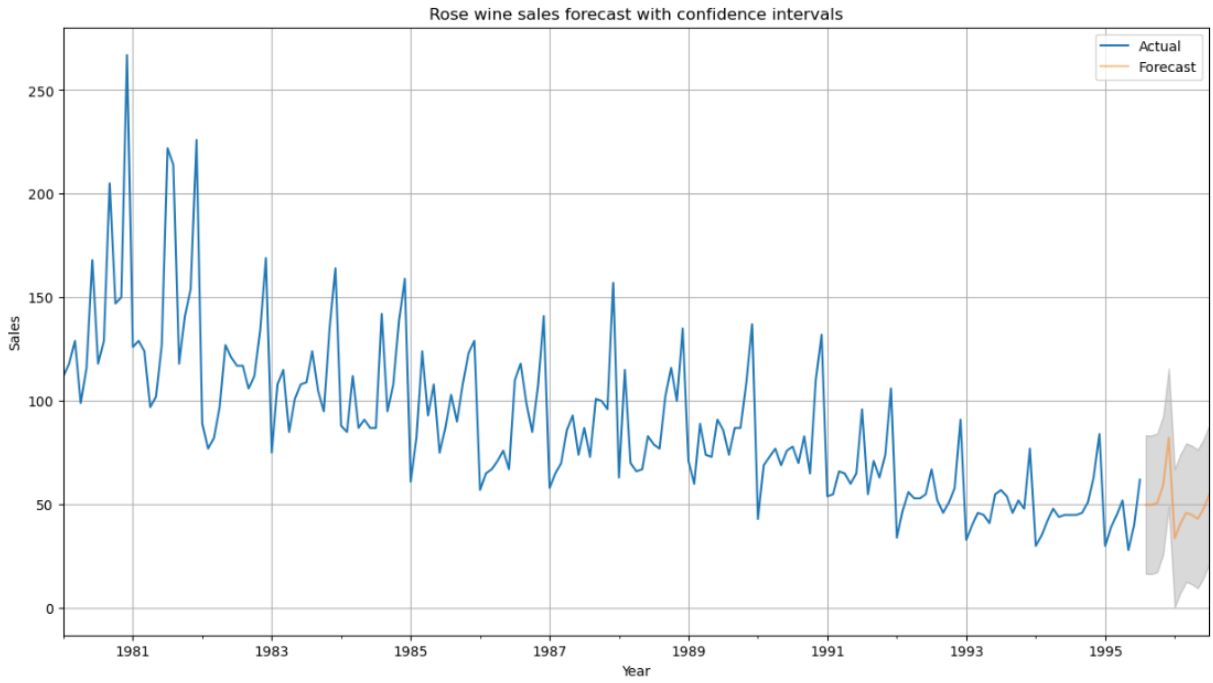


Figure28: Forecast plot for next 12 months data with confidence intervals

- The forecast range indicates that the data is going to see the similar sales observed for the year 1995.

Actionable insights:

- To increase rose wine sales, the company has to focus on the taste and pricing of the wine
- Feedback has to be taken from the customers who purchase the wine and who stopped purchasing the wine.
- Since there is a seasonality, the sales see mild perks during mid year and see high sales during year end, so attracting customers with loyalty and promotion schemes during those times can help improve the sales.
- Since there is no clue on the increasing trend anywhere over the years, the product has to be revisited and revamped and tried to help experiment and enhance business sales.