SMDM Project: TIME SERIES FORECASTING

Student's Name: THILAK RAJ | Batch: 18 FEB 2022

Contents

Figures:

Tables:

Executive Summary (**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 Wine Sales in the 20th century.

Data set for the Problem: Sparkling.csv and Rose.csv

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

Head of the Rose 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
5	1980-06	168.0
6	1980-07	118.0
7	1980-08	129.0
8	1980-09	205.0
9	1980-10	147.0

Table 1:Header of the Rose data

- The shape of the rose data base:

```
(187, 2)
```

- Info of the rose data:

```
 0   YearMonth  187 non-null    object
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

We considered YearMonth data as index.

- Head of the rose data after indexing.

	Rose
YearMonth	
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

- We found two null values in the rose dataframe. We imputed the missing values by forward filling method.

Head of the Sparkling data:

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471
5	1980-06	1377
6	1980-07	1966
7	1980-08	2453
8	1980-09	1984
9	1980-10	2596

Table 2:Header of Sparkling data

- The shape of the sparkling data base:

  (187, 2)

- Info of the rose data:

```
<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    object
 1   Sparkling  187 non-null    int64 
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

We considered YearMonth data as index.

- Head of the rose data after indexing.

	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

- Found no null values in the sparkling dataset.

Plotting the Rose wine sales data:

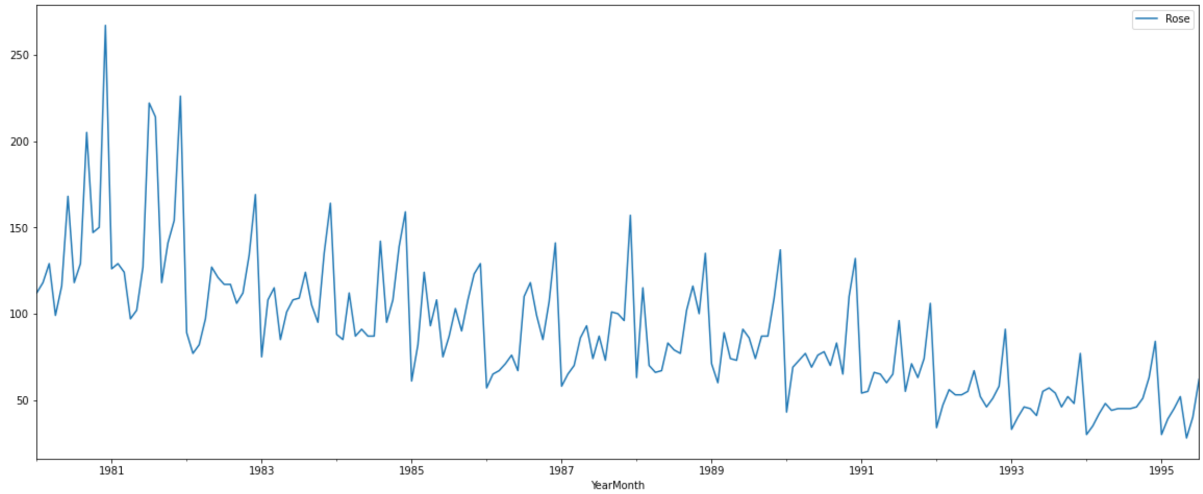


Figure 1:Plot of Rose wine data

Plotting the spark wine sales data:

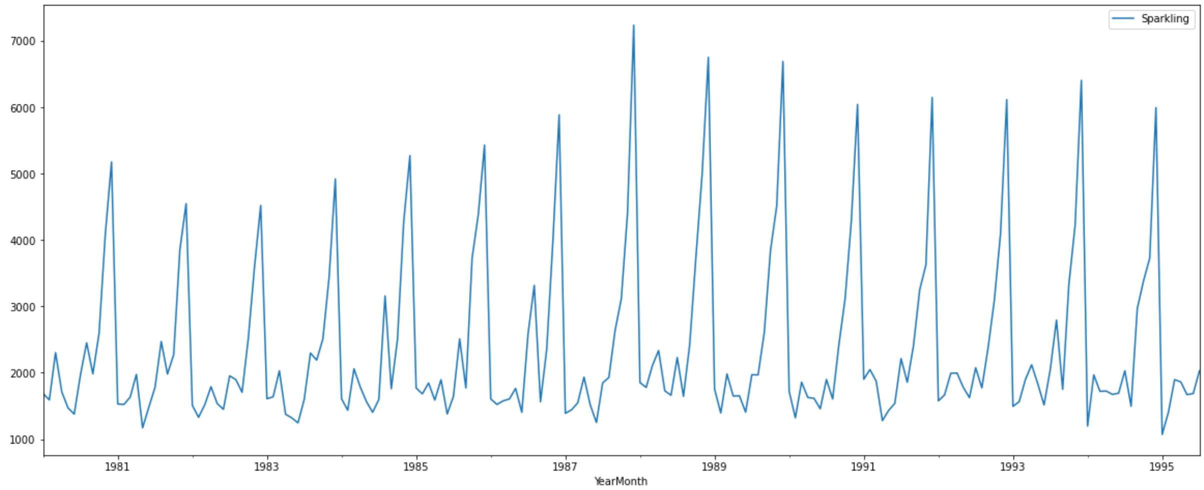


Figure 2:Plot of sparkling wine data

Observation:

Rose dataset:

- We can see that there is a slight downward trend with a seasonal pattern associated as well.
- We observe that the sales of Rose is decreasing with time

Sparking dataset

- We can see that there is almost a flat trend with a seasonal pattern associated well it.

- We observe that the pattern of the sales of the Sparkling wine is almost the same with respect to time, that is the data seems to be seasonal in nature.

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

Descriptive statistics of rose:

	Rose
count	187.000000
mean	89.909091
std	39.244440
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Table 3:Descriptive statistics of rose

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

Yearly and monthly Boxplot:

Distribution of sale of wine-Rose in each year:

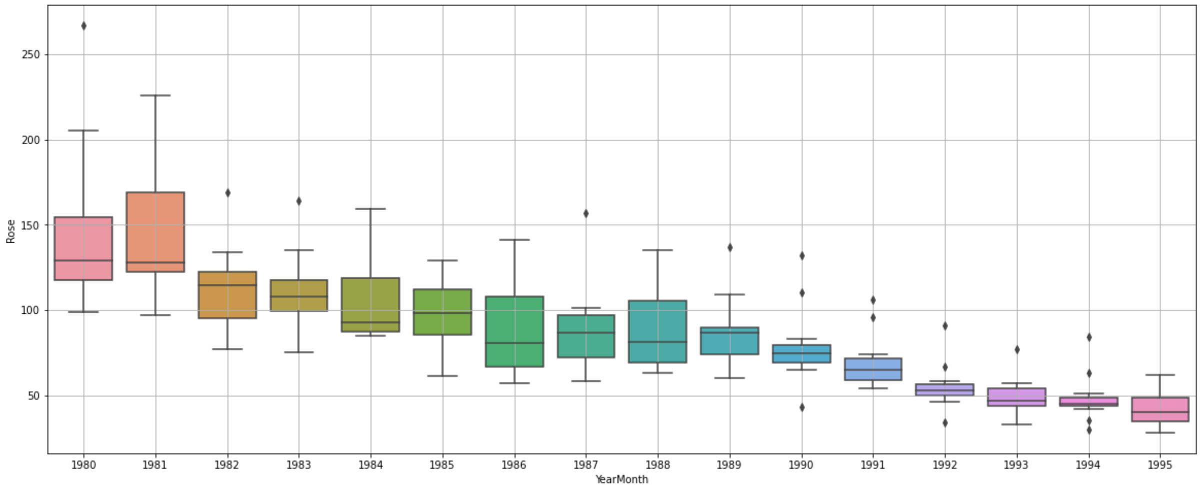


Figure 3:Distribution of sale of wine-Rose in each year

Distribution of sale of wine-Rose in each Month Trend for each month over the years in sale of wine-Rose:

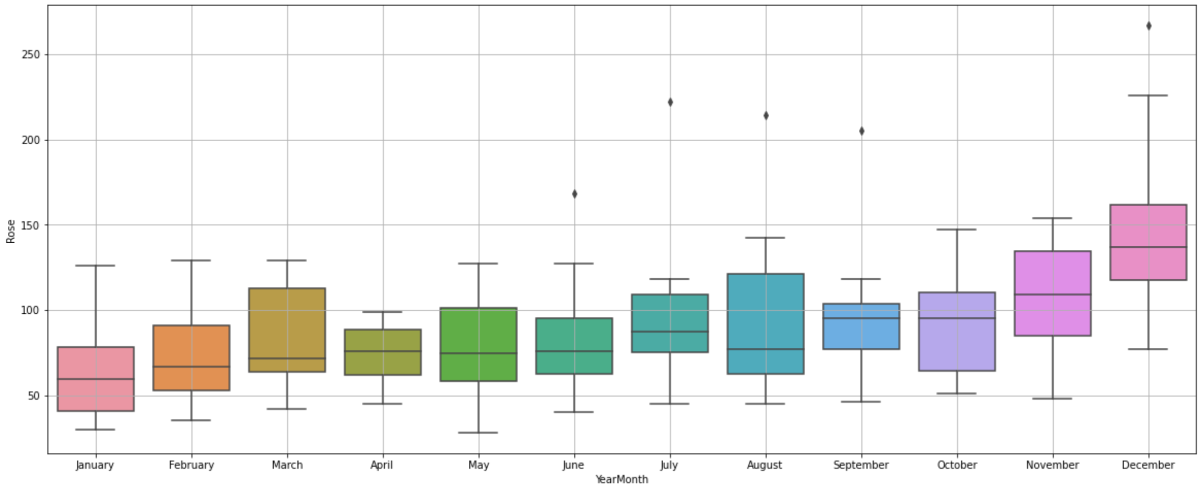


Figure 4:Distribution of sale of wine-Rose in each Month Trend for each month over the years in sale of wine-Rose.

Observation:

- The yearly boxplots also shows that the Sales have decreased towards the last few years.
- There is a clear distinction of sale within different months spread across various years. The highest such numbers are being recorded in the month of December and November across various years and least in January.
- Found outliers in both yearly and monthly data.

Plotting a month plot of the give Time Series. understand the spread of accidents across different years and within different months across years.

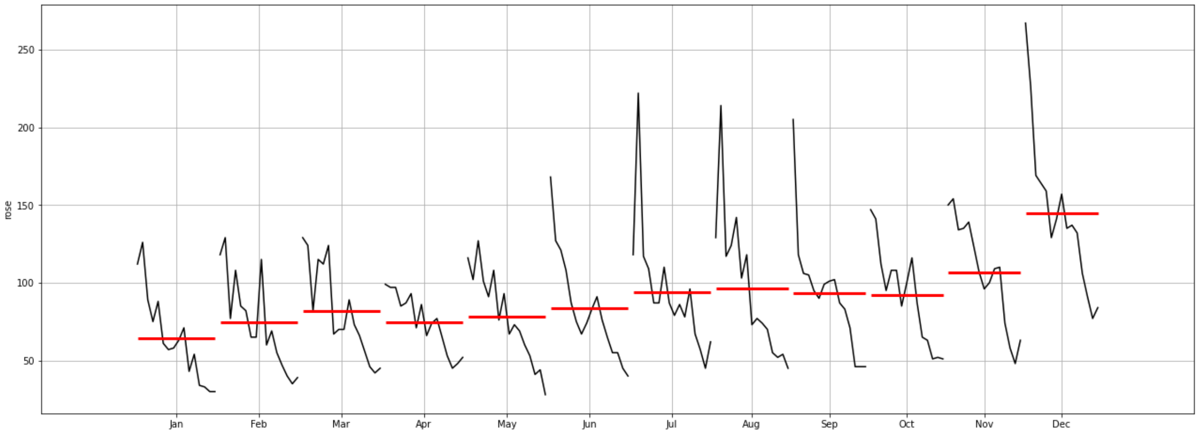


Figure 5:Plotting a month plot of the give Time Series

- This plot shows us the behaviour of the Time Series ('Rose' wine sale in this case) across various months. The red line is the median value.

Plot a graph of monthly Retail Sales across years.

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.0	45.0	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Table 4:Plot a graph of monthly Retail Sales across years.

Rose monthly sales across years:

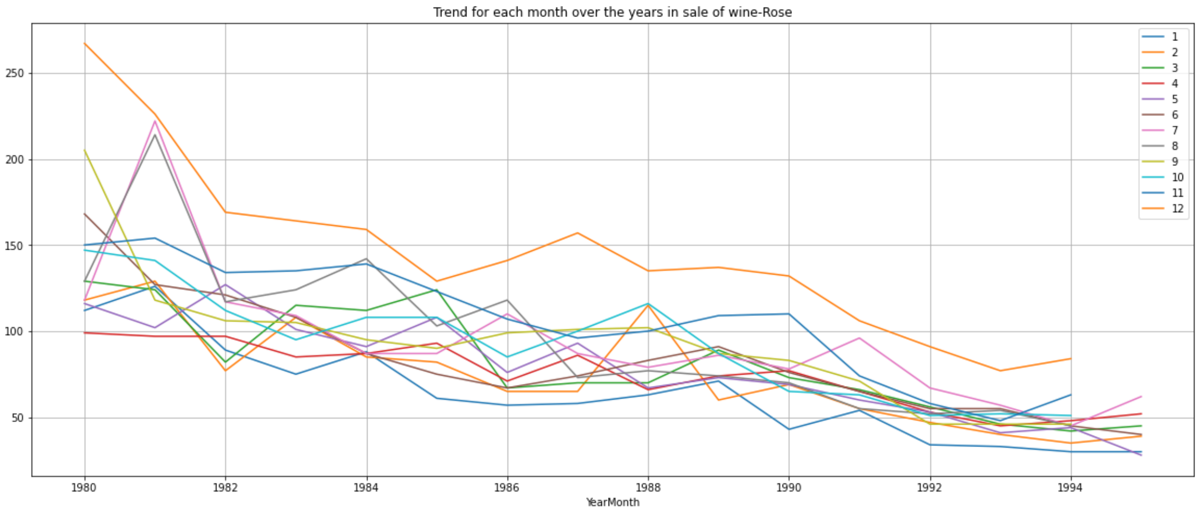


Table 5: Rose monthly salary across years

Observation:

- Overall monthly rose wine sale is also decreasing towards years.

Sparkling wine descriptive statistics of sparkling wine:

	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

Table 6:Sparkling wine descriptive statistics of sparkling wine:

Observation:

- As there is huge difference between 50%, mean and 75% with max value, there must be existence of outliers.

Yearly and monthly sale Boxplot

Distribution of sale of wine-sparkling in each year:

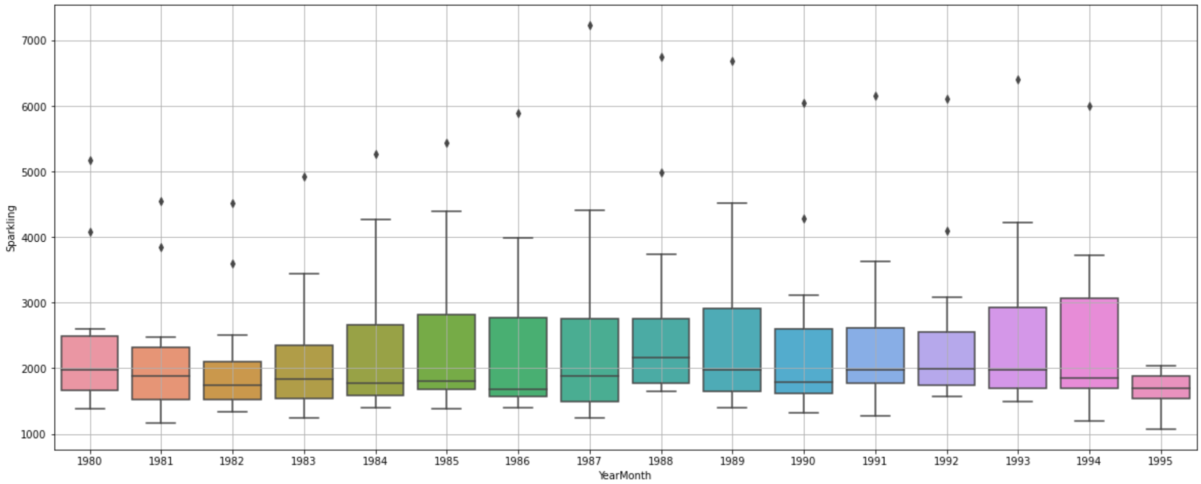


Figure 6:Distribution of sale of wine-sparkling in each year

Distribution of sale of wine-sparkling in each Month Trend for each month over the years in sale of wine-sparkling.

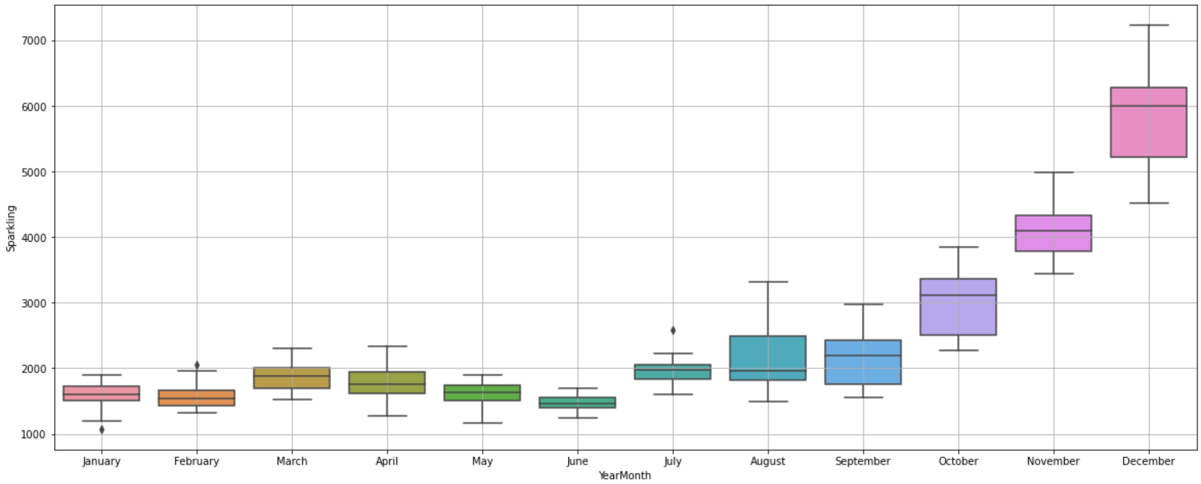


Figure 7:Distribution of sale of wine-sparkling in each Month Trend for each month over the years in sale of wine-sparkling.

Observation:

- The yearly boxplots also shows that the Sales is almost same last few years.
- There is a clear distinction of sale within different months spread across various years.
- Sales are gradually increasing after every September till December. The highest average sales are being recorded in December and November.
- The least sales are being recorded in the first seven months every year least in the month of June
- Found outliers in both yearly and monthly sale data.

Plotting a month plot of the given Time Series. understand the spread of accidents across different years and within different months across years.

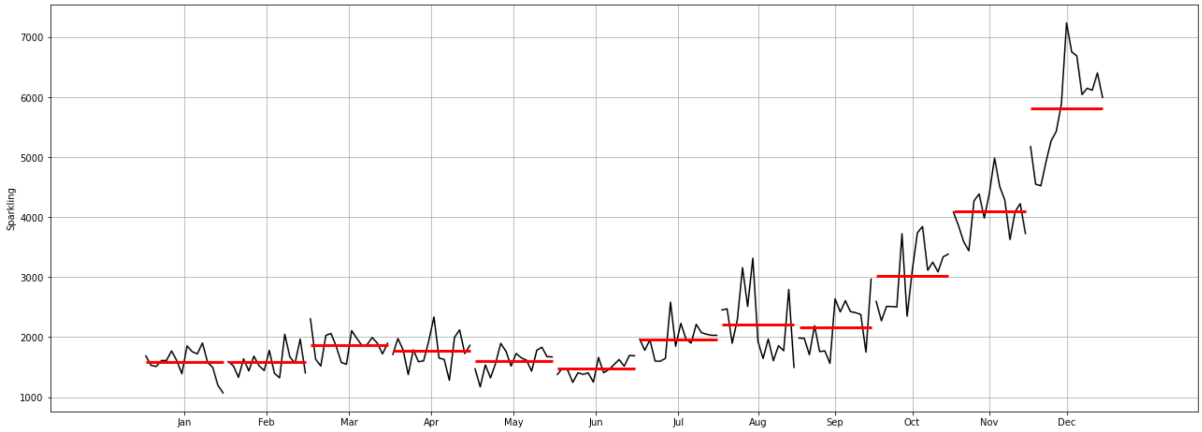


Figure 8: plot of the given Time Series.

Plot a graph of monthly Sales across years:

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
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

Table 7:Plot a graph of monthly Sales across years

Rose monthly sales across years:

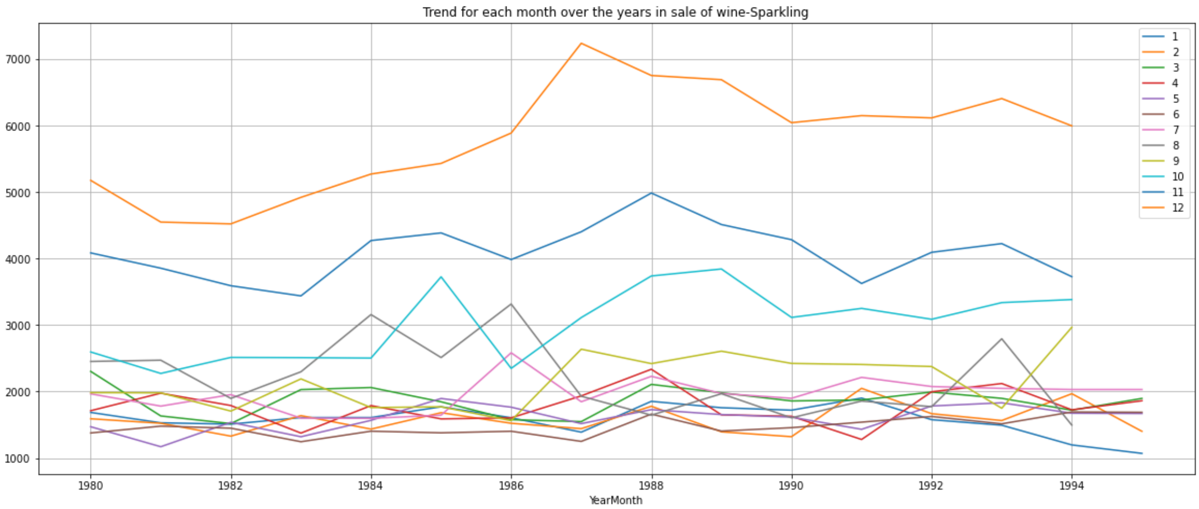


Figure 9:Rose monthly sales across years.

Observation:

- Overall monthly rose wine sale is almost same towards years.

Plotting actual data and rolling mean together (average of 12 months) for rose data:

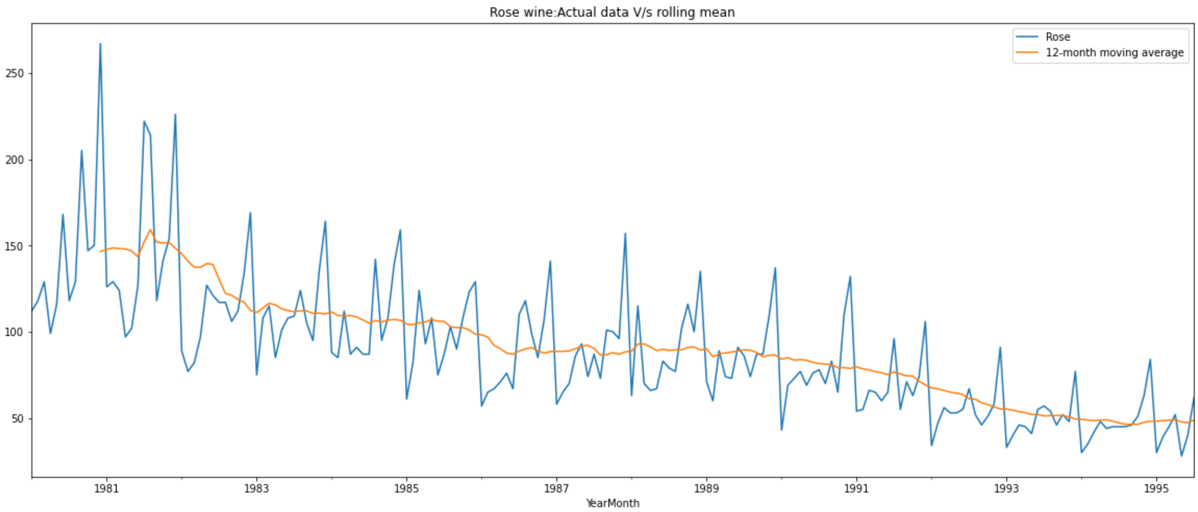


Figure 10:Rose wine: Actual data V/s rolling mean

Plotting actual data and rolling mean together (average of 12 months) for sparkling data:

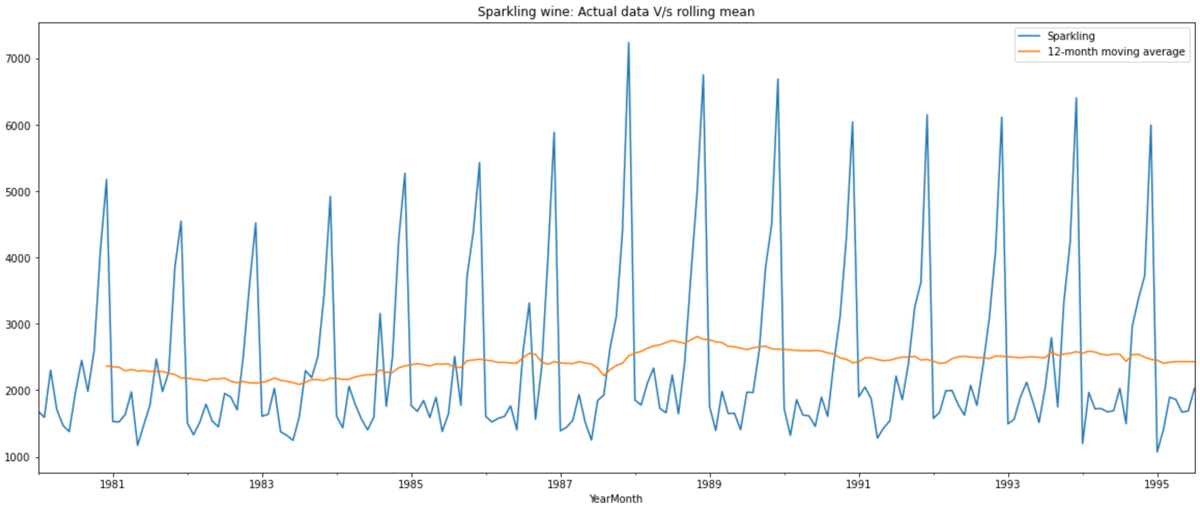


Figure 11:Sparkling wine: Actual data V/s rolling mean

Observation:

- Hence from the plot we can observe that the mean value of Rose dataset is decreasing over time whereas the for the sparkling dataset, it is almost constant.
- Now we will filter and remove the trend and cyclic component from each data using Hedrick filter method and then we will decompose the dataset using seasonal decomposition method

Applying Hodrick-Prescott filter before decomposition:

- Now we will filter and remove the trend and cyclic component from each data using Hodrick-Prescott smoothing parameter. filter method and then we will decompose the dataset using seasonal decomposition method
- The Hodrick–Prescott filter or Hodrick–Prescott decomposition is a mathematical tool that is used in time series analysis and modelling. This filter is mainly useful in removing the cyclic component from time-series data.
- Applying the Hodrick–Prescott filter in time series allows us to obtain a smooth time series from time series that has time series components like trend cycle and noise in large quantities.

Hpfilter: Rose data plot:

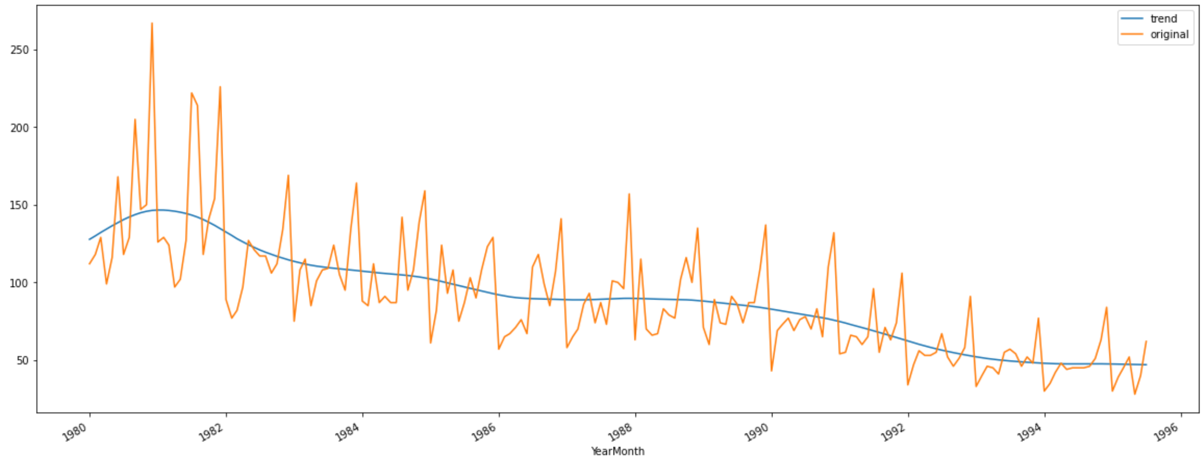


Figure 12: Hpfilter: Rose data plot

Hpfilter: Sparkling data plot:

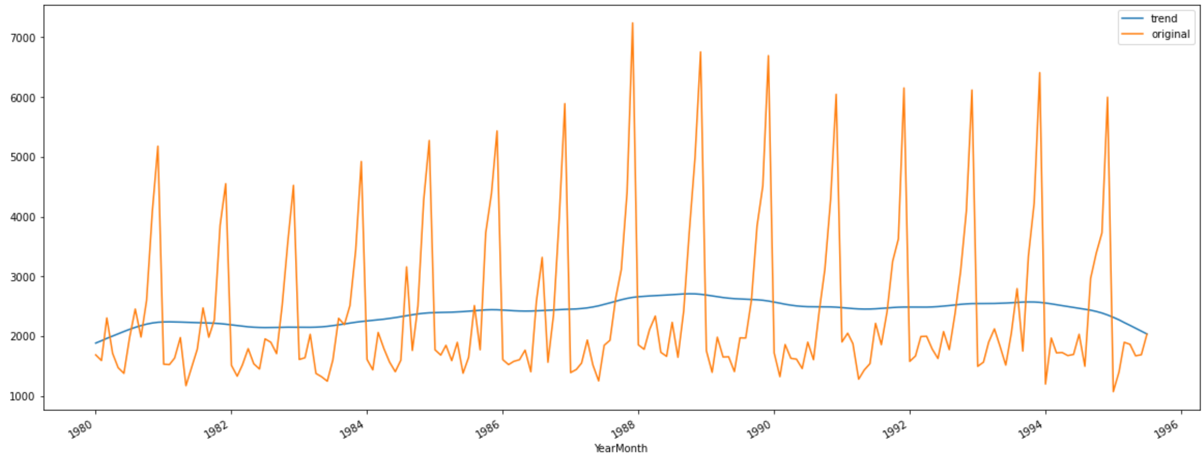


Figure 13: Hpfilter: Sparkling data plot

Seasonal decompose of the time series data:

Decomposition of Rose-wine sales into Trend, Seasonal and Residual Distribution each year.

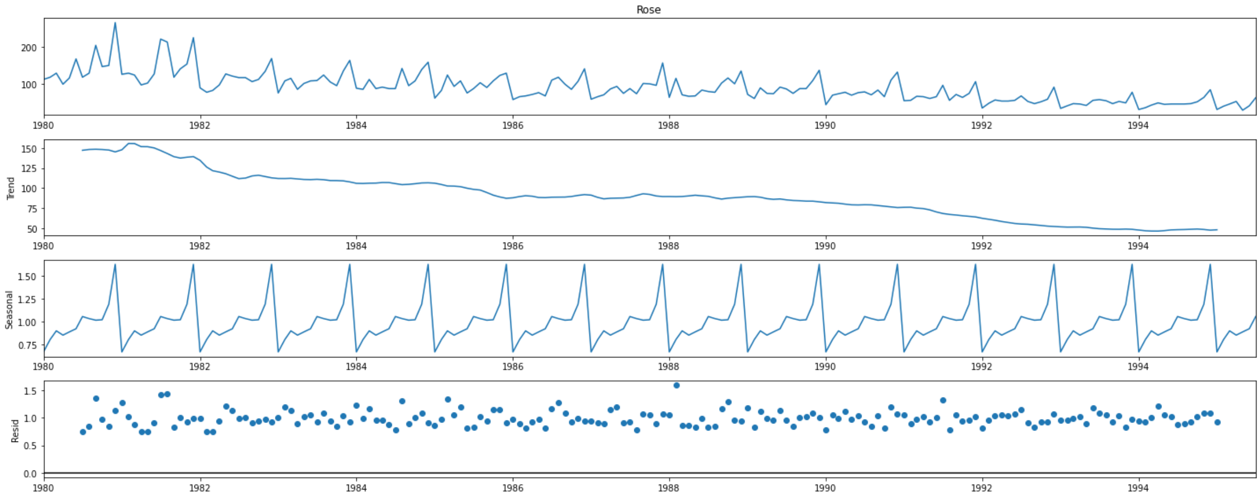


Figure 14:Decomposition of Rose wine sales data into Trend, Seasonal and ResidualDistribution

Decomposition of Sparkling-wine sales into Trend, Seasonal and Residual Distribution each year.

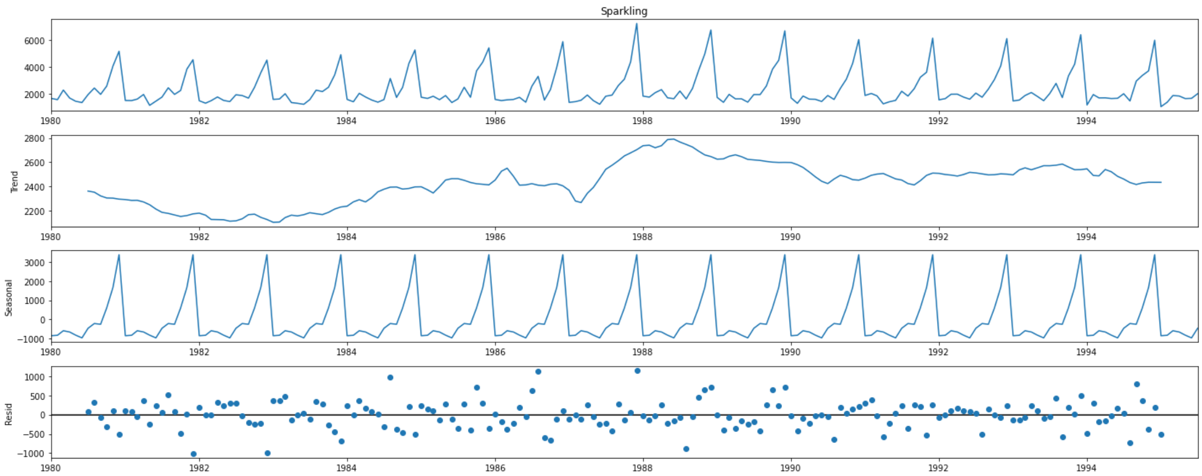


Figure 15:Decomposition of Sparkling wine sales data into Trend, Seasonal and Residual Distribution

Seasonal decomposition plots observations:

- The trend factor in the Rose dataset has a decreasing trend whereas the trend for the sparking dataset, initially it decreases a bit and then increases again becomes constant after 1990s.So we consider it as the almost constant trend.
- We also observed that both the dataset contains seasonal element in them.
- We found scattered residual component in both the dataset.

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

We split both the data into training and testing data sets

```
Shape of the train data:
(132, 1)
Shape of the test data:
(55, 1)
```

After splitting train and test data set , train data set has 132 records and test has 55 records. We splitted data in almost 70:30 ratio.

Rose data-plotting train and test split data together:

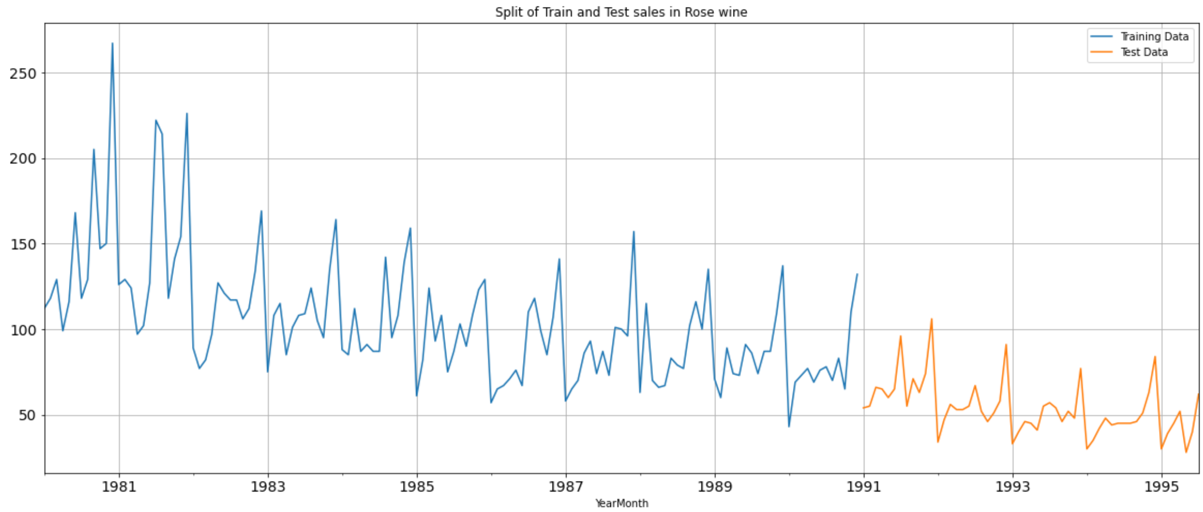


Figure 16:Split of Train and Test sales in Rose wine

Sparkling data-plotting train and test split data together:

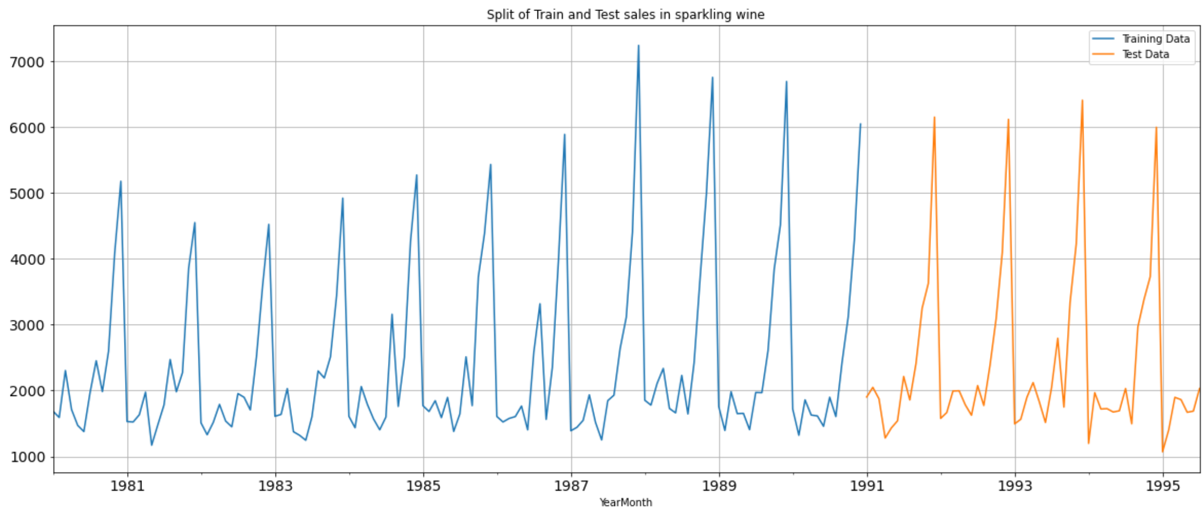


Figure 17:Split of Train and Test sales in sparkling win

Q4: 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.

Simple Exponential smoothing- Rose wine data:

- Autofit parameter to build best simple exponential smoothing model.

```
{'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}
```

- Below is the prediction header data using SES model.

YearMonth	Rose	predict
1991-01-01	54.0	87.104983
1991-02-01	55.0	87.104983
1991-03-01	66.0	87.104983
1991-04-01	65.0	87.104983
1991-05-01	60.0	87.104983

Table 8: Rose- SES model autofit prediction

Plotting Alpha =0.0987 Simple Exponential Smoothing predictions on Test Set.

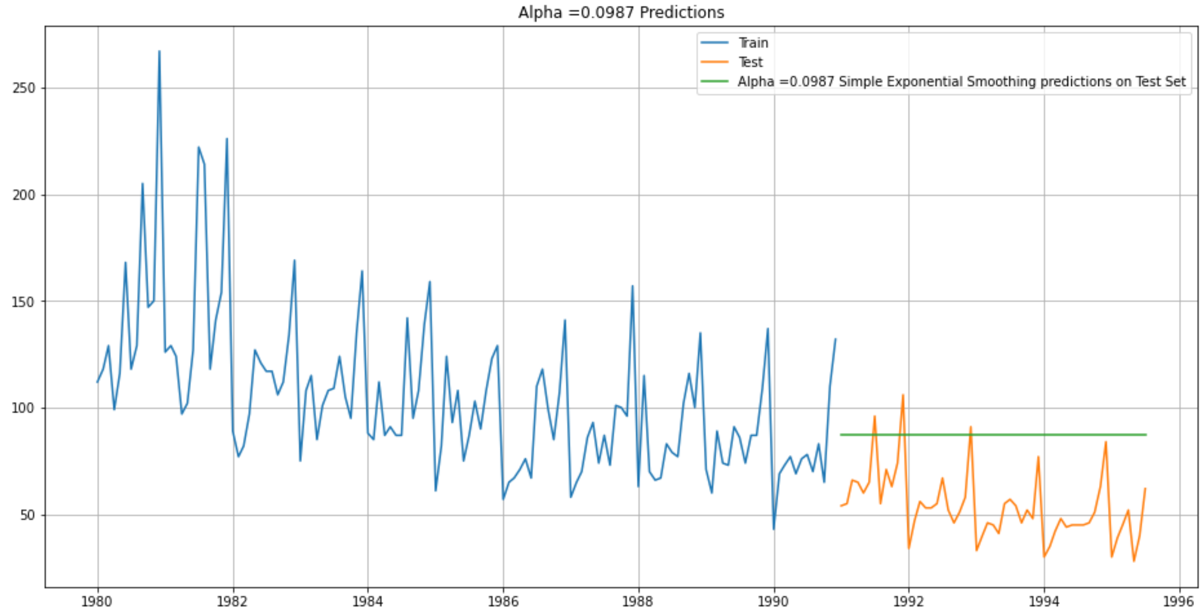


Table 9:SES Alpha =0.0987 Predictions-Rose

Model Evaluation for α = 0.0987: Simple Exponential Smoothing.

RMSE:

For Alpha =0.0987 Simple Exponential Smoothing Model forecast on the Rose Test Data, RMSE is **36.817**

Setting different alpha values.

- A smaller value (closer to 0) creates a smoother (slowly changing) line similar to a moving average with a large number of periods.
- A high value for alpha tracks the data more closely by giving more weight to recent data.
- We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

Below are the Test RMSE values for different values of α.

	Alpha Values	Train RMSE	Test RMSE
0	0.3	32.470164	47.525251
1	0.4	33.035130	53.787686
2	0.5	33.682839	59.661932
3	0.6	34.441171	64.991324
4	0.7	35.323261	69.718108
5	0.8	36.334596	73.793865
6	0.9	37.482782	77.159094

Table 10: SES Test RMSE for different α values.

- We found lowest Test RMSE value for alpha=0.0987 and alpha=0.3. So we are ignoring other alpha values.

Plotting the SES prediction for Rose:

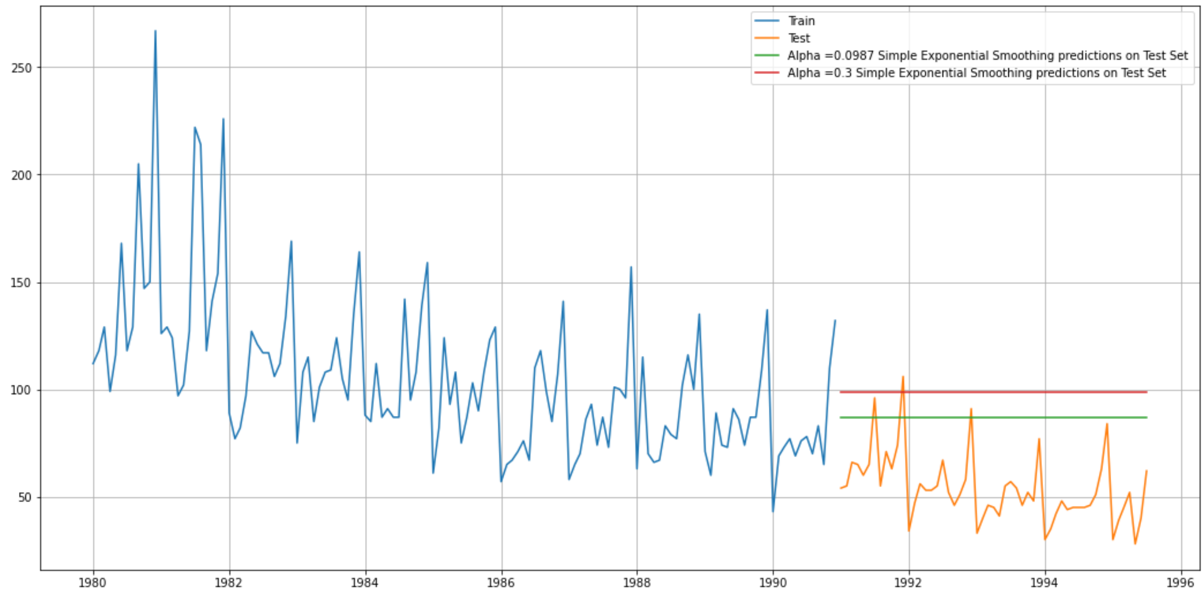


Figure 18:Simple Exponential Smoothing predictions for different alpha values on Test Set-Rose

Above shows the SES model prediction on different alpha values on rose data. We can see that the RMSE value is better on alpha=0.0987.

	Rose Test RMSE
Alpha=0.0987,SimpleExponentialSmoothing	36.816889
Alpha=0.3,SimpleExponentialSmoothing	47.525251

Simple Exponential smoothing- Sparkling wine data:

- Autofit parameter to build best simple exponential smoothing model.

```
{'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}
```

- Below is the prediction header data using SES model.

	Sparkling	predict
YearMonth		
1991-01-01	1902	2724.932624
1991-02-01	2049	2724.932624
1991-03-01	1874	2724.932624
1991-04-01	1279	2724.932624
1991-05-01	1432	2724.932624

Table 11:Sparkling- SES model autofit prediction

Plotting Alpha =0.0496 Simple Exponential Smoothing predictions on Test Set.

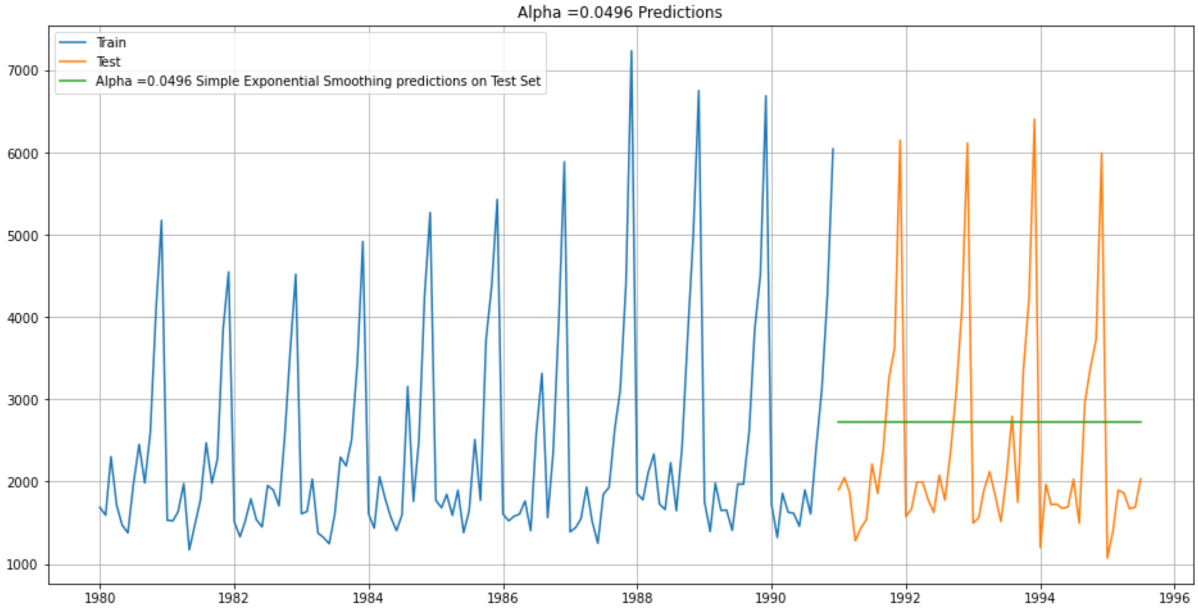


Figure 19:SES Alpha =0.0987 Predictions-Sparkle

Model Evaluation for α = 0.0496: Simple Exponential Smoothing.

RMSE:

For Alpha =0.0496 Simple Exponential Smoothing Model forecast on the Sparkle Test Data, RMSE is 1316.035

Setting different alpha values.

Below are the Test RMSE values for different values of α.

	Alpha Values	Train RMSE	Test RMSE
0	0.3	1359.511747	1935.507132
1	0.4	1352.588879	2311.919615
2	0.5	1344.004369	2666.351413
3	0.6	1338.805381	2979.204388
4	0.7	1338.844308	3249.944092
5	0.8	1344.462091	3483.801006
6	0.9	1355.723518	3686.794285

Table 12:SES Test RMSE for different α values-Sparkle

- We found lowest Test RMSE value for alpha=0.0496 and alpha=0.3. So, we are ignoring other alpha values.

Plotting the SES prediction for Rose:

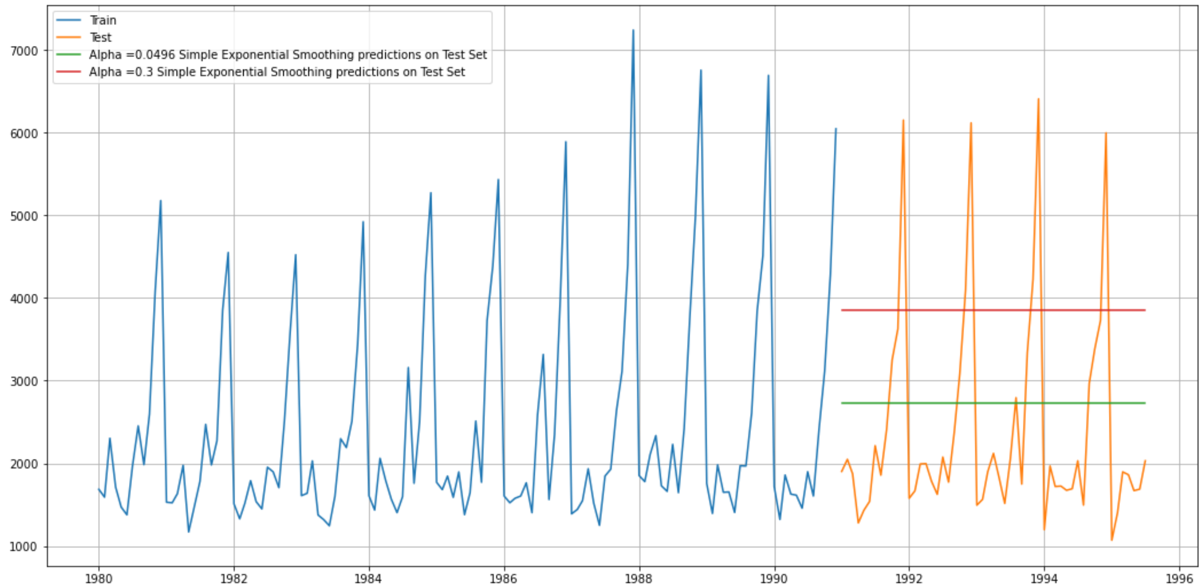


Figure 20:Simple Exponential Smoothing predictions for different alpha values on Test Set-Sparkle

Above shows the SES model prediction on different alpha values on rose data. We can see that the RMSE value is better on alpha=0.0496.

	Sparkling Test RMSE
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha=0.3,SimpleExponentialSmoothing	1935.507132

Double Exponential Smoothing (Holt's Model)-Rose Data:

Two parameters **α** and **β** are estimated in this model. Level and Trend are accounted for in this model.

Finding best parameters using autofit:

```
{'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}
```

Below is the head of the prediction:

	Rose	predict
YearMonth		
1991-01-01	54.0	73.259732
1991-02-01	55.0	72.767150
1991-03-01	66.0	72.274569
1991-04-01	65.0	71.781987
1991-05-01	60.0	71.289405

Table 13:DES prediction-Rose

Plotting the DES prediction-Rose:

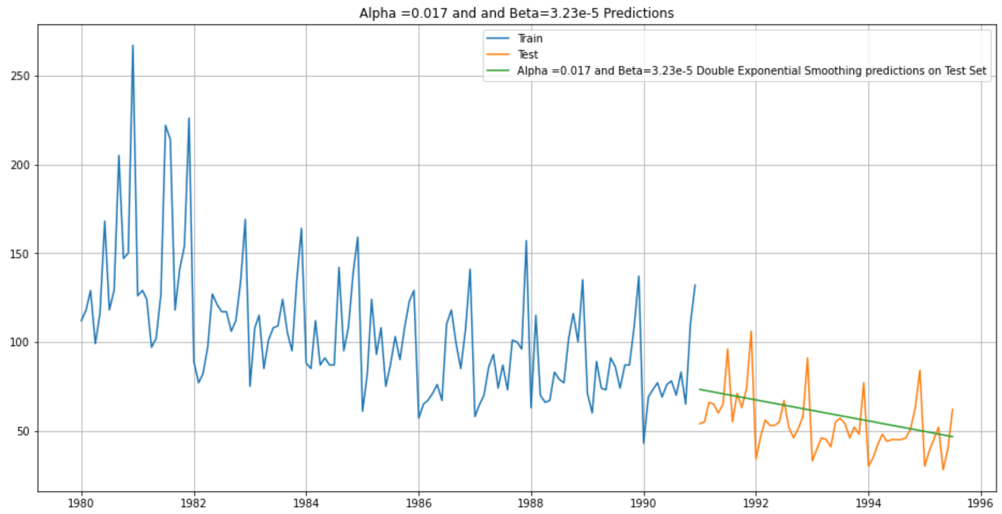


Figure 21:Alpha =0.017 and Beta=3.23e-5 DES Predictions-Rose

Model Evaluation for Alpha =0.017 and Beta=3.23e-5: Double Exponential Smoothing.

RMSE:

For Alpha =0.017 and Beta=3.23e-5 Double Exponential Smoothing model forecast on the Rose Test Data, RMSE is 15.715

Double Exponential Smoothing (Holt's Model)-Sparkling Data:

Finding best parameters using autofit:

```
{'smoothing_level': 0.6885714285714285,
 'smoothing_trend': 9.999999999999999e-05,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': -95.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Below is the head of the prediction

YearMonth	Sparkling	predict
1991-01-01	1902	5221.278699
1991-02-01	2049	5127.886554
1991-03-01	1874	5034.494409
1991-04-01	1279	4941.102264
1991-05-01	1432	4847.710119

Table 14: DES model prediction-Rose

Plotting the DES prediction-Rose:

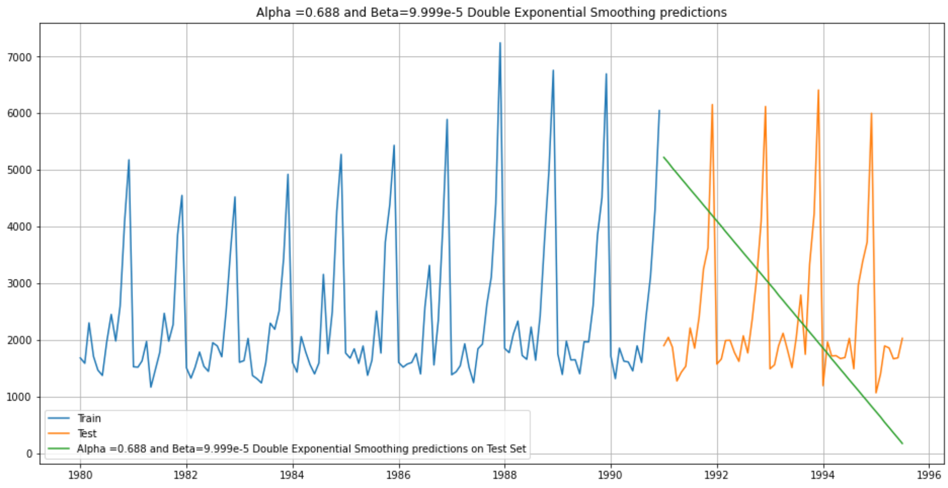


Figure 22:Alpha =0.688 and Beta=9.999e-5 DES predictions

Model Evaluation for Alpha =0.688 and Beta=9.999e-5 Double Exponential Smoothing. *RMSE*:

For Alpha =0.688 and Beta=9.999e-5 Double Exponential Smoothing model forecast on the Rose Test Data, RMSE is 2007.239

We also tried to build model with different alpha and beta combinations. Below is the best five Test RMSE result of it.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1592.292788	18259.110704
8	0.4	0.3	1569.338606	23878.496940
1	0.3	0.4	1682.573828	26069.841401
16	0.5	0.3	1530.575845	27095.532414
24	0.6	0.3	1506.449870	29070.722592

Observation: In the prediction data we saw that, the prediction of best parameter grid forecasting data is in descending order remaining all the alpha and beta combination predictions are increasing. Also, we found RMSE for the different alpha and beta values and sorted in Test RMSE in descending order. But we couldn't find the better result in this task. So we are considering only Alpha =0.688 and Beta=9.999e-5 Double Exponential Smoothing model with RMSE of 2007.239

Plotting Alpha=0.3, Beta=0.3, Double Exponential Smoothing predictions on Test Set

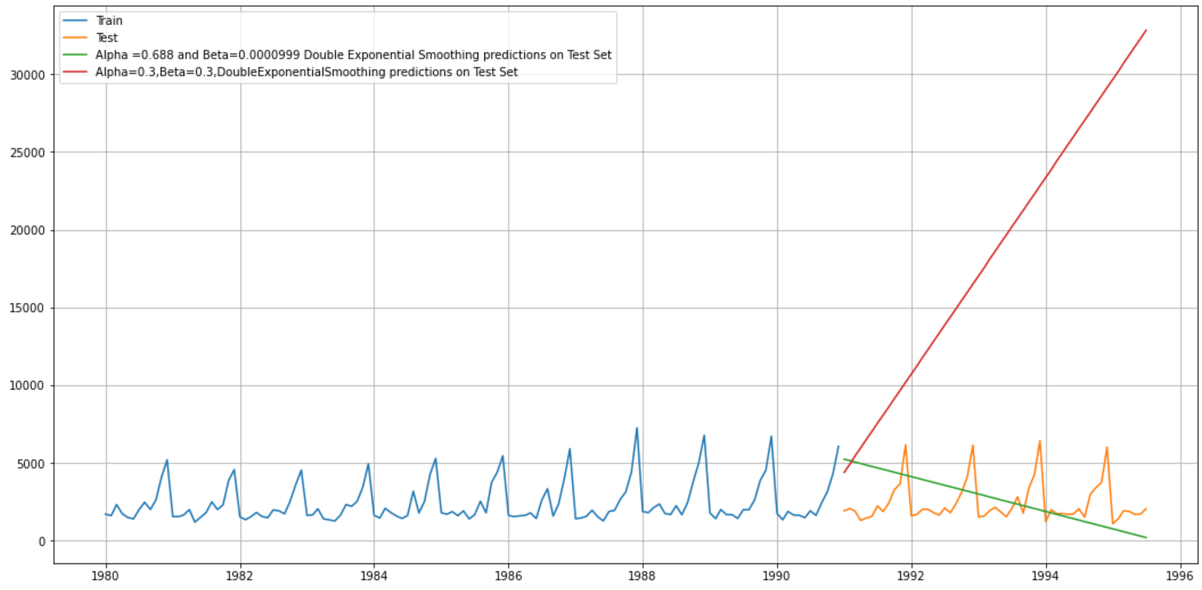


Figure 23:Alpha=0.3, Beta=0.3, DES predictions on Test Set-Sparkling

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

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

Finding best parameters using autofit:

```
{'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}
```

Below is the head of the prediction

	Rose	predict
YearMonth		
1991-01-01	54.0	56.321655
1991-02-01	55.0	63.664690
1991-03-01	66.0	69.374024
1991-04-01	65.0	60.435528
1991-05-01	60.0	67.758341

Table15:TES Prediction-Rose data

Plotting the TES prediction-Rose:

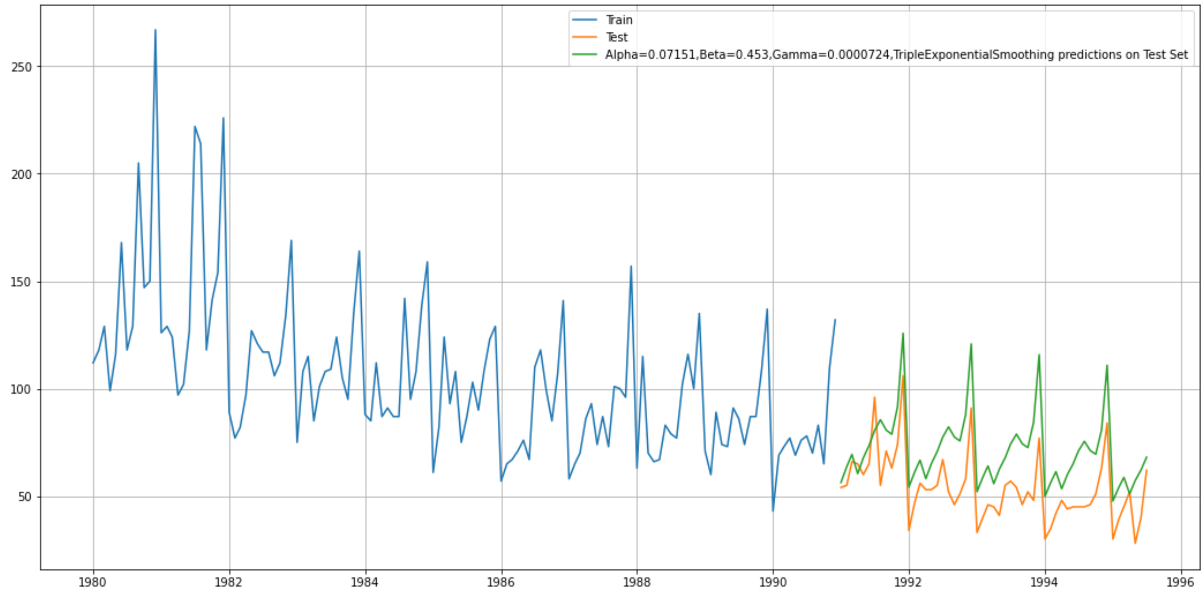


Figure 24:Alpha=0.07151, Beta=0.453, Gamma=0.0000724 TES predictions-Rose

Model Evaluation for Alpha=0.07151, Beta=0.453, Gamma=0.0000724 TES model.

RMSE:

For Alpha=0.07151, Beta=0.453, Gamma=0.0000724, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 20.183

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

Finding best parameters using autofit:

```
{'smoothing_level': 0.11133818361298699,
 'smoothing_trend': 0.049505131019509915,
 'smoothing_seasonal': 0.3620795793580111,
 'damping_trend': nan,
 'initial_level': 2356.4967888704355,
 'initial_trend': -10.187944726007238,
 'initial_seasons': array([0.71296382, 0.68242226, 0.90755008, 0.80515228, 0.65597218,
        0.65414505, 0.88617935, 1.13345121, 0.92046306, 1.21337874,
        1.87340336, 2.37811768]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Below is the head of the prediction:

YearMonth	Sparkling	predict
1991-01-01	1902	1587.497468
1991-02-01	2049	1356.394925
1991-03-01	1874	1762.929755
1991-04-01	1279	1656.165933
1991-05-01	1432	1542.002730

Table 16: TES prediction-Sparkling.

Plotting the DES prediction-Sparkling:

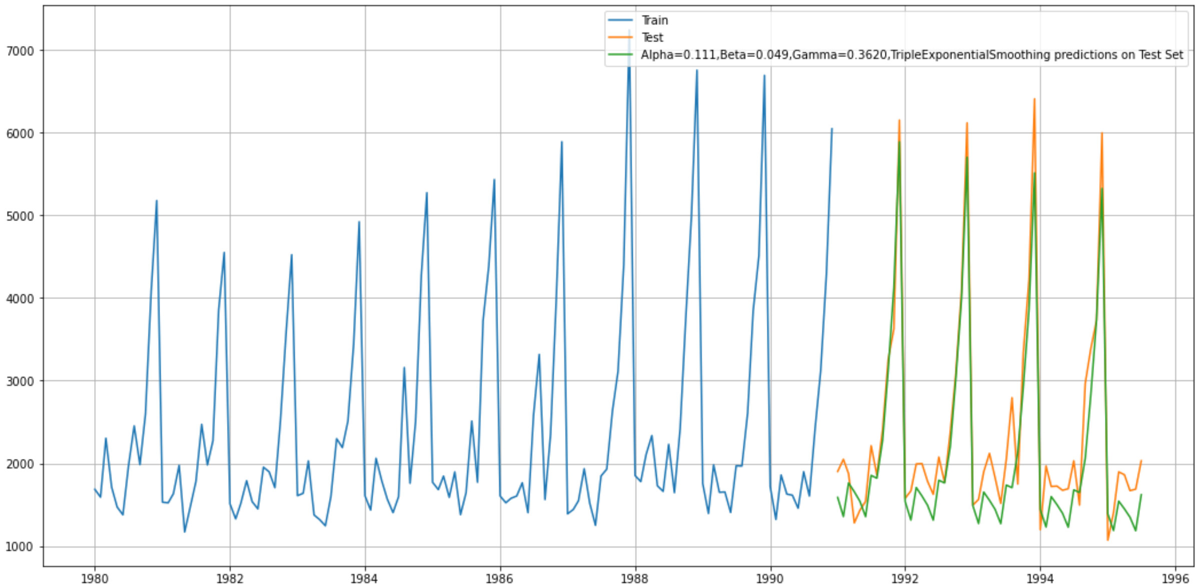


Figure 25:Alpha=0.111,Beta=0.049,Gamma=0.3620,TES model-Sparkling

Model Evaluation for Alpha=0.111,Beta=0.049,Gamma=0.3620 TES- Sparkling.

RMSE:

For Alpha=0.111, Beta=0.049, Gamma=0.3620, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 404.287

Linear Regression Models-Rose:

Note: For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

We have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

```
First few rows of Training Data
              Rose  time
YearMonth
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     5

Last few rows of Training Data
              Rose  time
YearMonth
1990-08-01   70.0   128
1990-09-01   83.0   129
1990-10-01   65.0   130
1990-11-01  110.0   131
1990-12-01  132.0   132

First few rows of Test Data
             Rose  time
YearMonth
1991-01-01  54.0   133
1991-02-01  55.0   134
1991-03-01  66.0   135
1991-04-01  65.0   136
1991-05-01  60.0   137

Last few rows of Test Data
             Rose  time
YearMonth
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   187
```

Table 17: Sample data after imputing the time column to rose train and test data

Now that our training and test data has been modified, let us go ahead use $LinearRegression$ to build the model on the training data and test the model on the test data.

We built the Linear Regression model on Rose data below is the plot of the prediction on test data.

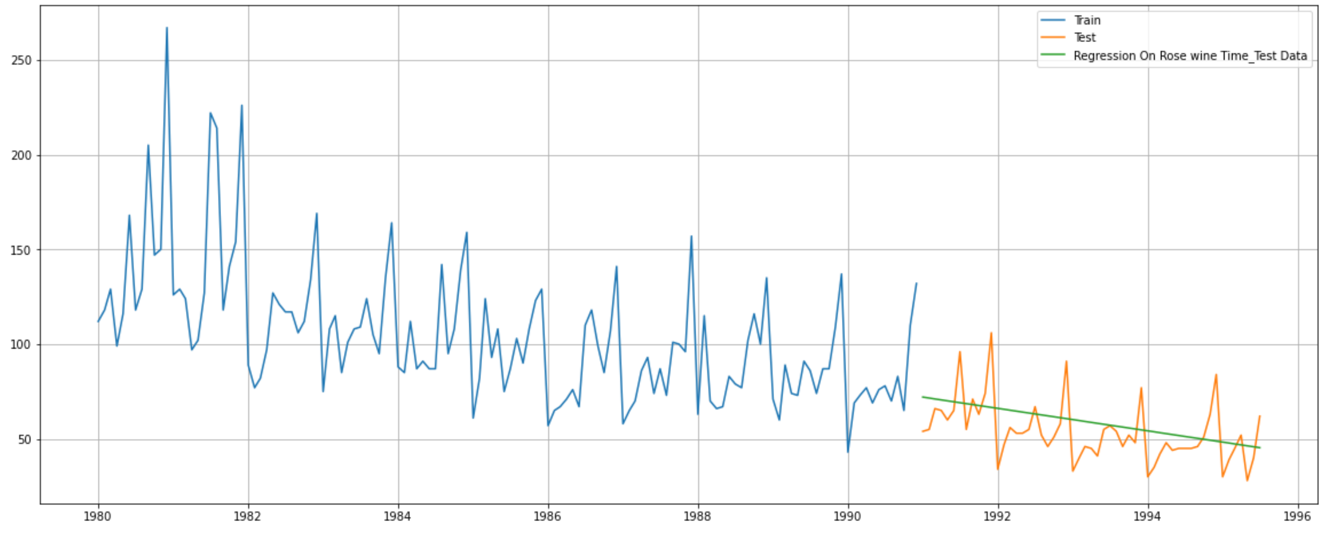


Figure 26:Regression on Rose wine Time-Test Data

Model Evaluation for Linear Regression model-Rose.

RMSE:

For Regression on Time forecast on the Test Data, RMSE is 15.276

Linear Regression Models-Sparkling:

We have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

```
First few rows of Training Data
            Sparkling  time
YearMonth
1980-01-01       1686     1
1980-02-01       1591     2
1980-03-01       2304     3
1980-04-01       1712     4
1980-05-01       1471     5

Last few rows of Training Data
            Sparkling  time
YearMonth
1990-08-01       1605   128
1990-09-01       2424   129
1990-10-01       3116   130
1990-11-01       4286   131
1990-12-01       6047   132

First few rows of Test Data
            Sparkling  time
YearMonth
1991-01-01       1902   133
1991-02-01       2049   134
1991-03-01       1874   135
1991-04-01       1279   136
1991-05-01       1432   137

Last few rows of Test Data
            Sparkling  time
YearMonth
1995-03-01       1897   183
1995-04-01       1862   184
1995-05-01       1670   185
1995-06-01       1688   186
1995-07-01       2031   187
```

Table 18:Sample data after imputing the time column to sparkling train and test data

Now that our training and test data has been modified, let us go ahead use $LinearRegression$ to build the model on the training data and test the model on the test data.

We built the Linear Regression model on Sparkling data below is the plot of the prediction on test data.

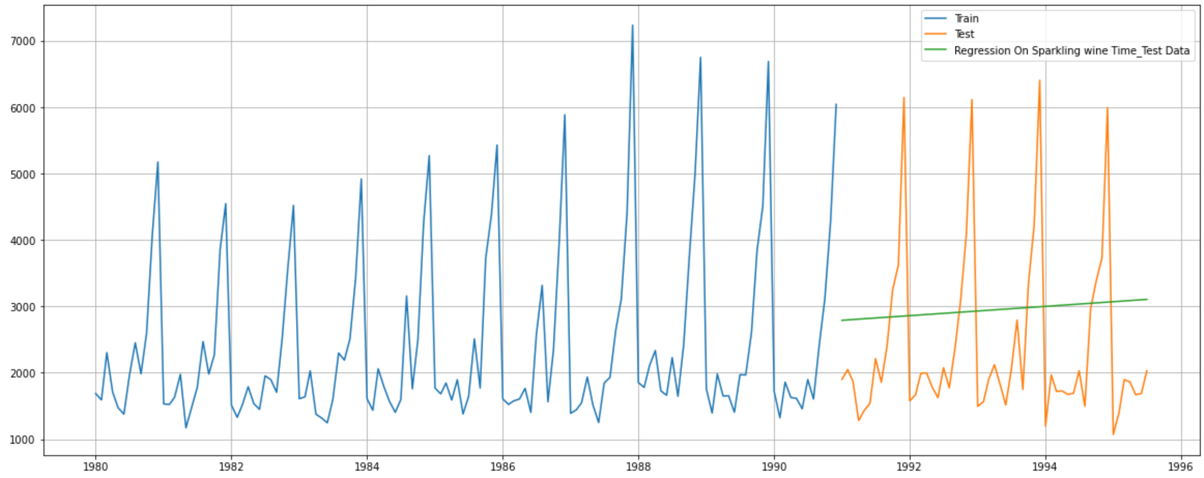


Figure 27:Regression On Sparkling wine Time-Test Data.

Model Evaluation for Linear Regression model-Sparkling.

RMSE:

For Regression on Time forecast on the Sparkling Test Data, RMSE is 1389.135

naïve forecast models-Rose data

Note: For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

Below is the Naïve model prediction.

```
YearMonth
1991-01-01    132.0
1991-02-01    132.0
1991-03-01    132.0
1991-04-01    132.0
1991-05-01    132.0
Freq: MS, Name: naive, dtype: float64
```

Table 19: Naive prediction-Rose

Plotting the Naïve model prediction.

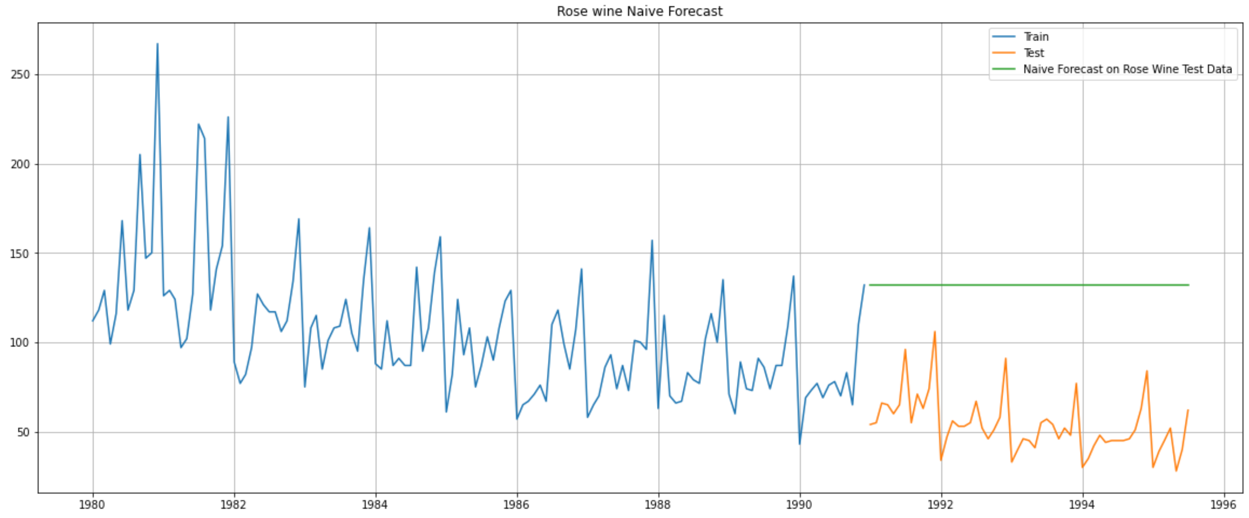


Figure 28:Naive Forecast on Rose Wine Test Data

Model Evaluation for Naïve model prediction.

RMSE:

For naïve forecast on the Test Data, RMSE is 79.739

Naïve forecast models-Sparkling data

Below is the Naïve model prediction.

```
YearMonth
1991-01-01    6047
1991-02-01    6047
1991-03-01    6047
1991-04-01    6047
1991-05-01    6047
Freq: MS, Name: naive, dtype: int64
```

Table 20: Naive prediction-Sparkling

Plotting the Naïve model prediction.

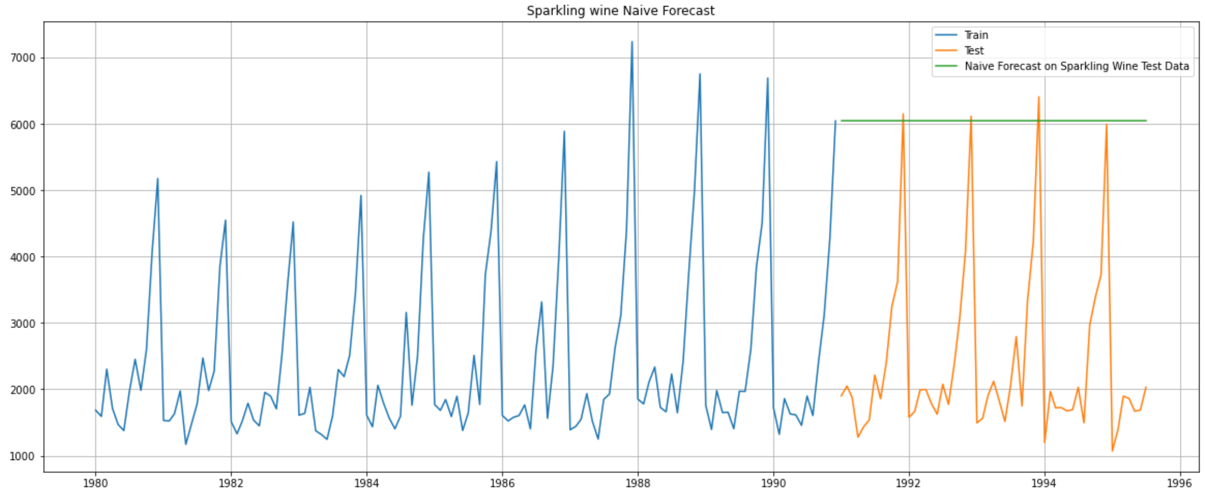


Figure 29:Naive Forecast on Sparkling Wine Test Data

Model Evaluation for Naïve model prediction.

RMSE:

For naïve forecast on the Sparkling wine Test Data, RMSE is 3864.279

Simple Average Model-Rose data:

Note: For this particular simple average method, we will forecast by using the average of the training values.

Below is the Simple average model prediction.

YearMonth	Rose	mean_forecast
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394

Table 21: Simple average model prediction-Rose

Plotting the Simple Average model forecast.

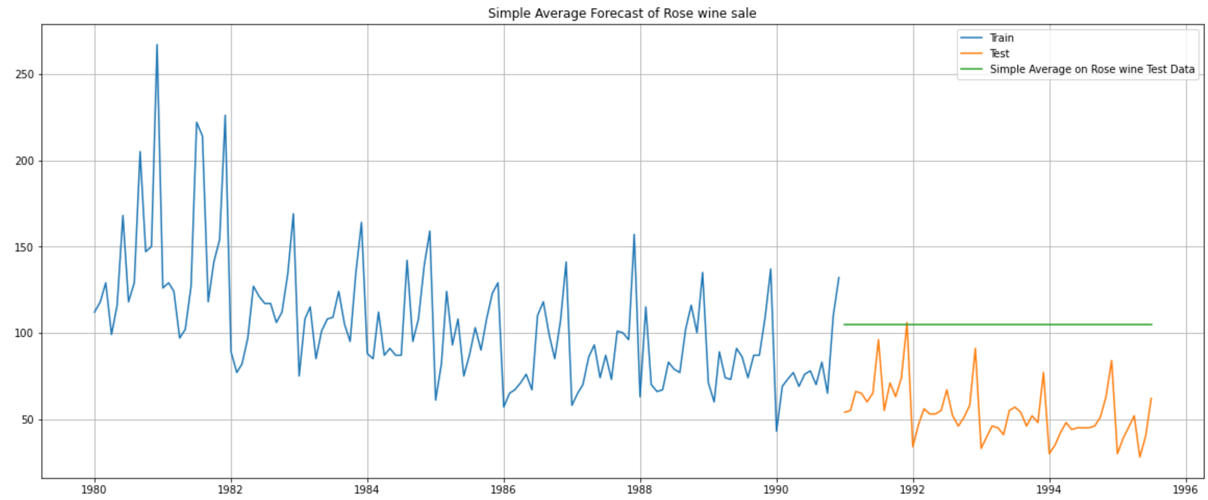


Figure 30:Simple Average Forecast of Rose wine sale

Model Evaluation for Simple Average Model-Rose

RMSE:

For simple average forecast on the Rose wine Test Data, RMSE is 53.481

Simple Average Model-Sparkling data:

Below is the Simple average model prediction.

YearMonth	Sparkling	mean_forecast
1991-01-01	1902	2403.780303
1991-02-01	2049	2403.780303
1991-03-01	1874	2403.780303
1991-04-01	1279	2403.780303
1991-05-01	1432	2403.780303

Table 22:Simple average model prediction-Sparkling

Plotting the Simple Average model forecast-Sparkling

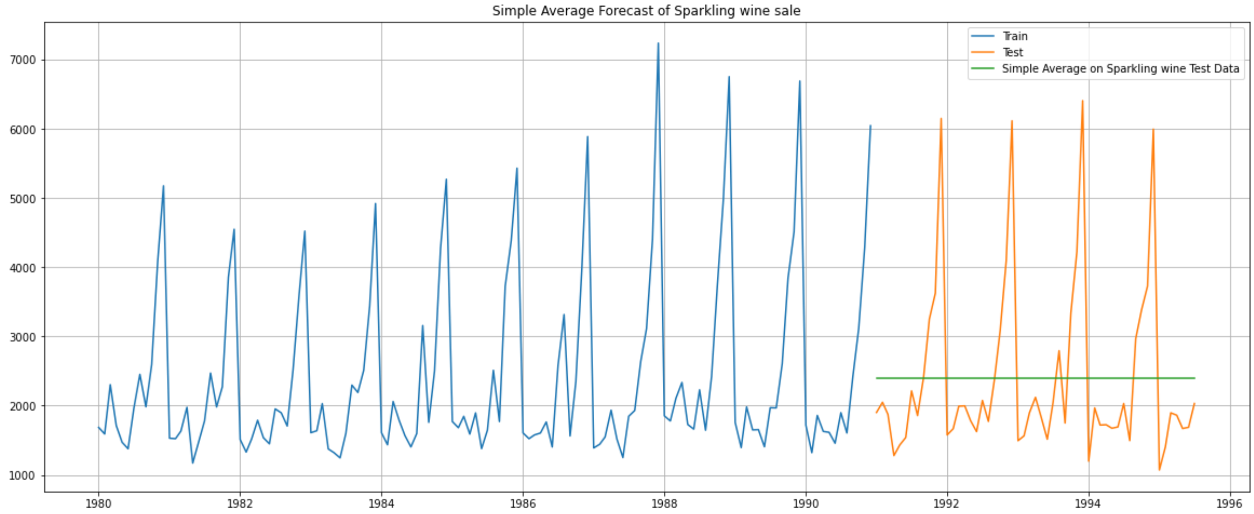


Figure 31:Simple Average Forecast of Sparkling wine

Model Evaluation of the Simple Average model forecast-Sparkling

RMSE:

For simple average forecast on the Sparkling wine Test Data, RMSE is 1275.082

Moving Average (MA)-Rose Data:

Finding out moving average for different point trailing.

YearMonth	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

Table 23: moving average for different point trailing-Rose

Plotting Moving Average Forecast of Rose data:

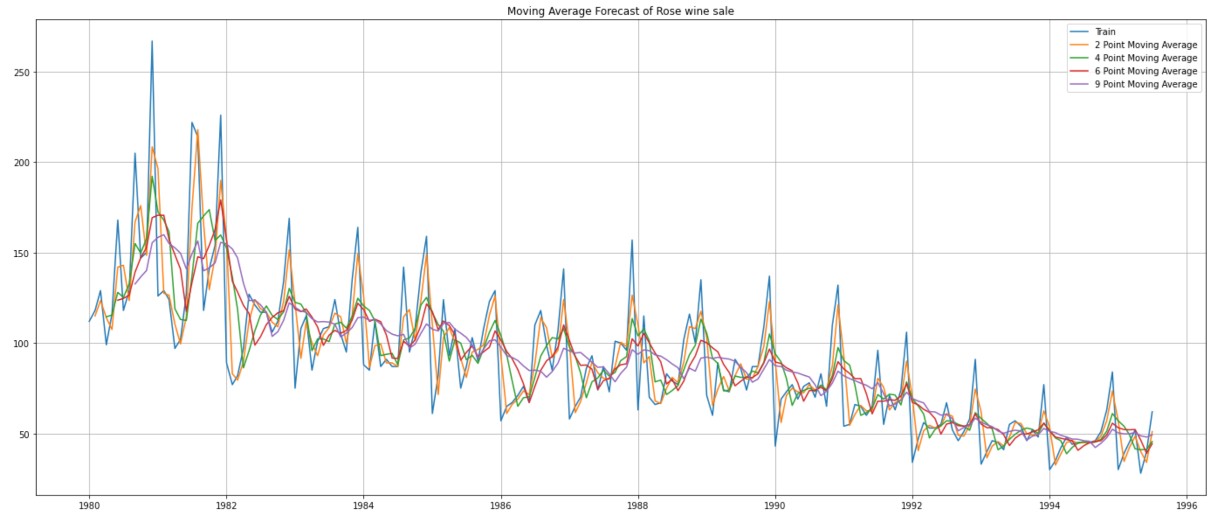


Figure 32: moving Average Forecast of Rose wine sale

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

Plotting Moving Average Forecast of Rose wine sale according to test and train data

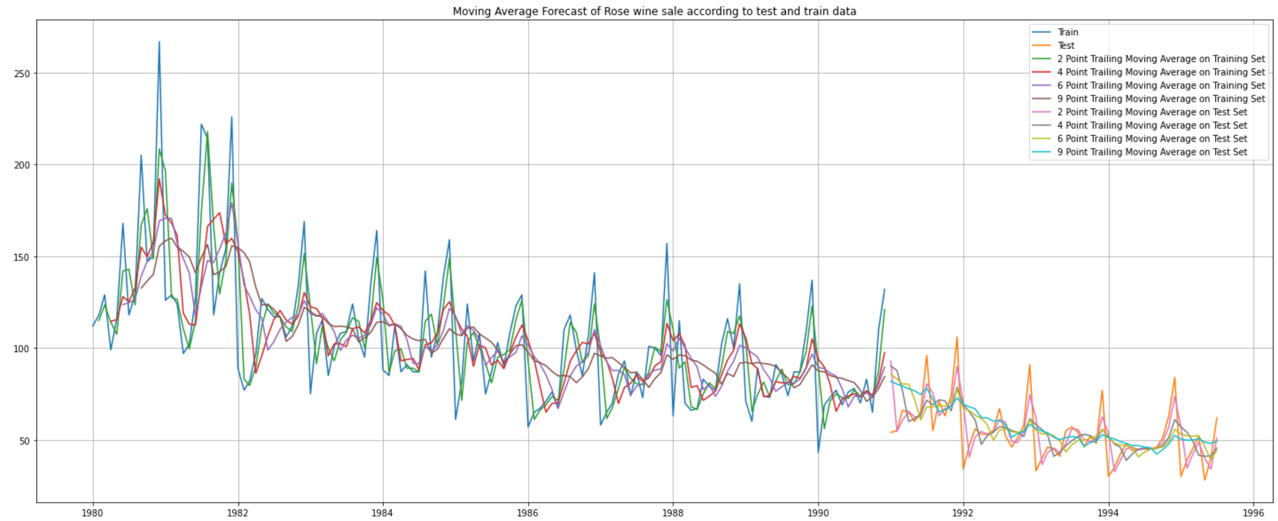


Figure 33:Moving Average Forecast of Rose wine sale according to test and train data

Model Evaluation Moving Average Forecast of Rose wine sale

RMSE:

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.455
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.572
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.731
```

Moving Average (MA)-Sparkling Data

Finding out moving average for different point trailing.

YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Table 24:moving average for different point trailing-Sparkling

Plotting Moving Average Forecast of Sparkling data:

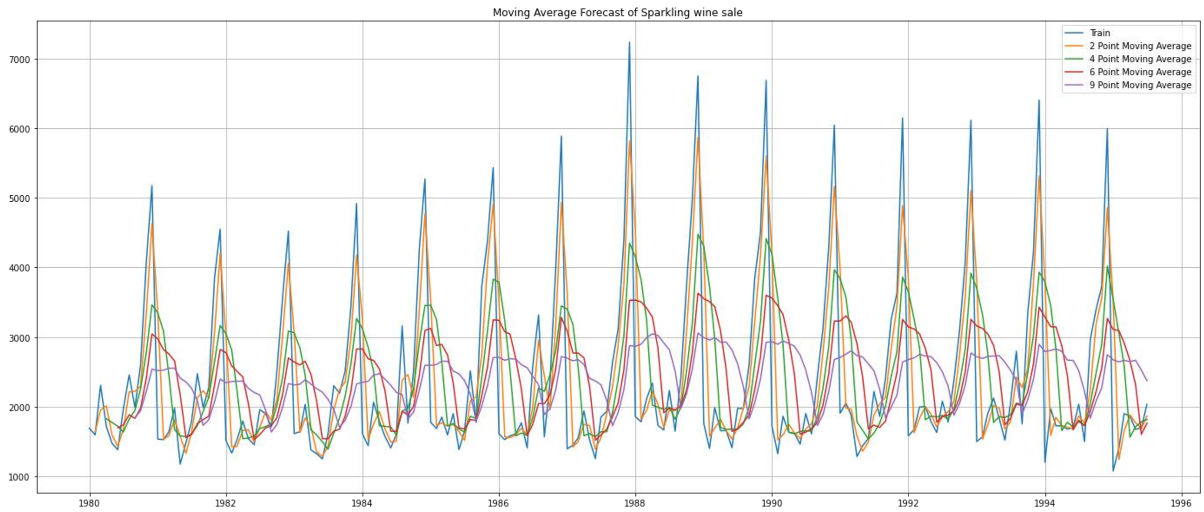


Figure 34: moving Average Forecast of Sparkling wine sale

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

Plotting Moving Average Forecast of Sparkling wine sale according to test and train data

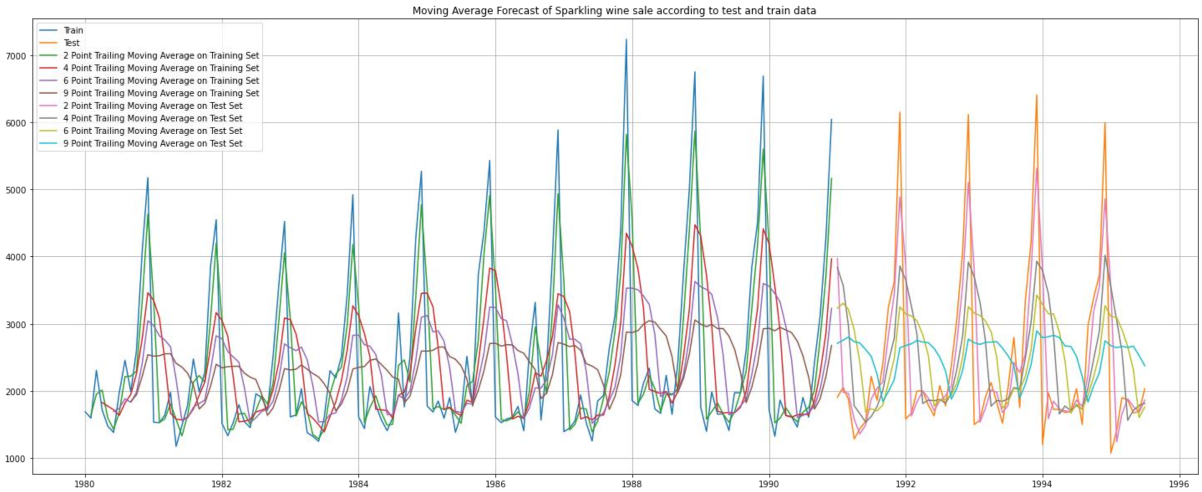


Figure 35::Moving Average Forecast of Sparkling wine sale according to test and train data

Model Evaluation Moving Average Forecast of Sparkling wine sale

RMSE:

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 1346.278
```

Plot all the models and compare the Time Series plots with test RMSE-Rose.

```
Sorted by RMSE values on the Rose wine Test Data:
```

	Rose Test RMSE
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
Alpha =0.017 and Beta=3.23e-5 Double Exponential Smoothing model	15.715112
Alpha=0.07151,Beta=0.453,Gamma=0.0000724, Triple Exponential Smoothing	20.182721
Alpha=0.0987,SimpleExponentialSmoothing	36.816889
Alpha=0.3,SimpleExponentialSmoothing	47.525251
Simple average forecast	53.480857
NaiveModel	79.738550

Table 25:RMSE values on the Rose wine Test Data

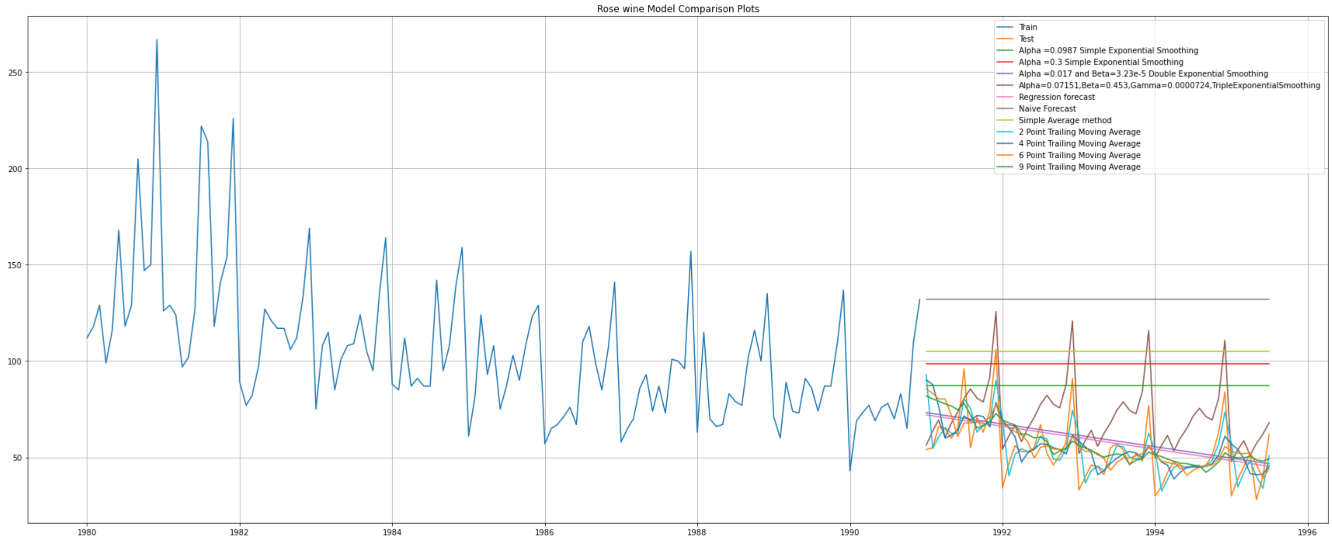


Figure 36: Rose wine Model Comparison Plots

Observation:

- Triple Exponential smoothing work better compare to simple and double exponential smoothing as it is including level, trend and seasonality.
- Naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today. As it is not including trend and seasonality, the RMSE is highest among all the methods followed by simple average forecast.
- Rolling moving average work better among all the other models and regression model RMSE is almost matches with rolling moving average in RMSE.
- From the above data, we can consider 2-point trailing moving average model is the best model with lower RMSE 11.53.

Plot all the models and compare the Time Series plots with test RMSE-Sparkling

```
Sorted by RMSE values on the Sparkling wine Test Data:
```

	Sparkling Test RMSE
Alpha=0.111,Beta=0.049,Gamma=0.3620 Triple Exponential Smoothing	404.286809
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple average forecast	1275.081804
6pointTrailingMovingAverage	1283.927428
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
Alpha=0.3,SimpleExponentialSmoothing	1935.507132
Alpha=0.688,Beta=0.0000999,DoubleExponentialSmoothing	2007.238526
NaiveModel	3864.279352

Table 26:RMSE values on the Sparkling wine Test Data

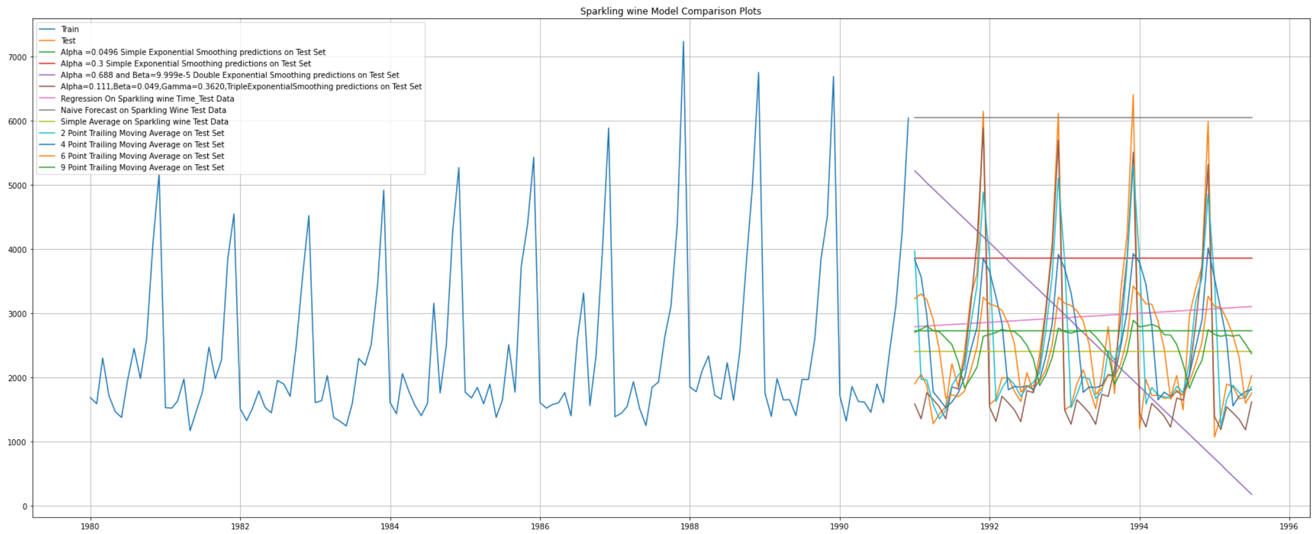


Figure 37: Sparkling wine Model Comparison Plots

Observation:

- Triple Exponential smoothing work better com pare to simple and double exponential smoothing as it is including level, trend and seasonality.
- Naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today. As it is not including trend and seasonality, the RMSE is highest among all the methods followed by Alpha=0.688, Beta=0.0000999, Double Exponential Smoothing which can be considered as poor models.
- From the above data, we can consider Alpha=0.111, Beta=0.049, Gamma=0.3620 Triple Exponential Smoothing is the best model with lower RMSE 404.29.

Q5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

To check whether the series is stationary, we use the Augmented Dickey Fuller (ADF)test whose null and alternate hypothesis can be simplified to

`Null Hypothesis H0`: Time series is not stationery

`Alternative Hypothesis H1`: Time series is stationery

Test for stationarity of the series - Dicky Fuller test-Rose data:

Plot rolling statistics:

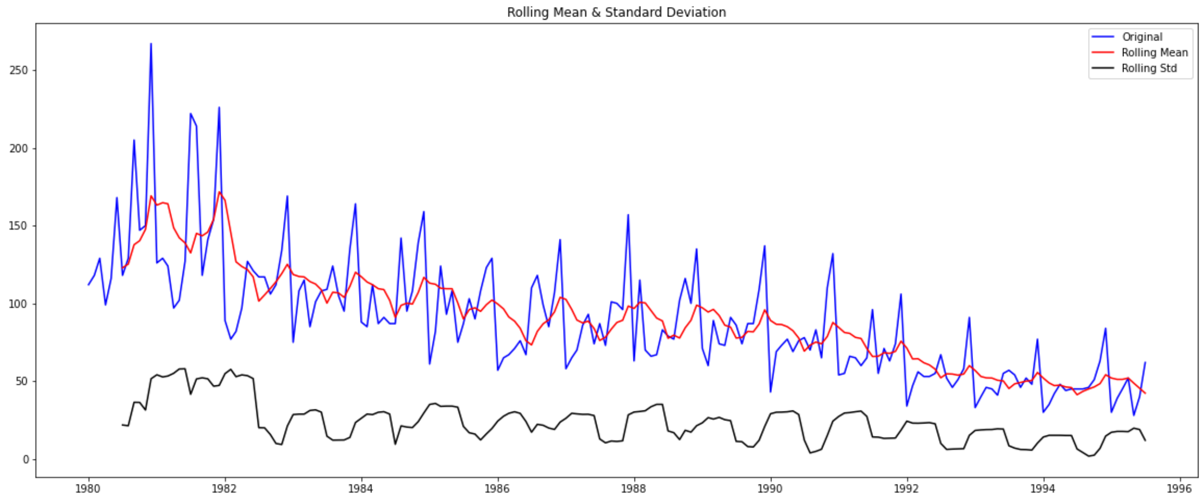


Figure 38:Rolling Mean & Standard Deviation-Rose

Results of Dickey-Fuller Test:

```
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
dtype: float64
```

Table 27: Results of Dickey-Fuller Test-Rose

since p-value> 0.05, at alpha 0.05, time series is not stationary. We can take next levels of differencing (a difference of order 1) to make a Time Series stationary.

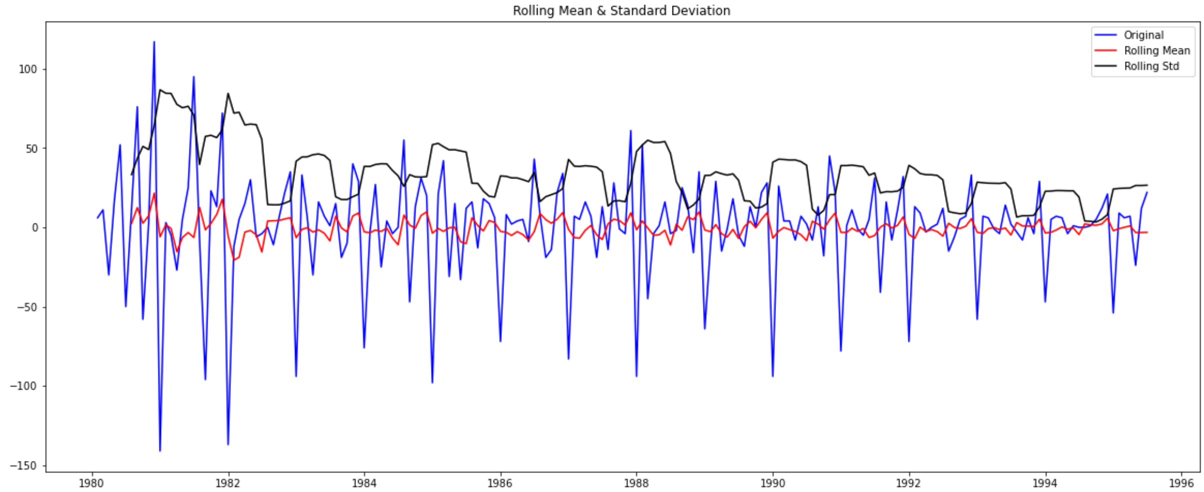


Figure 39: Rolling Mean & Standard Deviation (Next level)-Rose

Results of Dickey-Fuller Test:

```
Results of Dickey-Fuller Test:
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
dtype: float64
```

Table 28: Results of Dickey-Fuller Test (Next level)-Rose

Observation: After next level of levels of differencing p-value <0.05. We see that at α = 0.05 the Time Series is indeed stationary therefore series is stationary.

Test for stationarity of the series - Dicky Fuller Test-Sparkling data:

Plot rolling statistics:

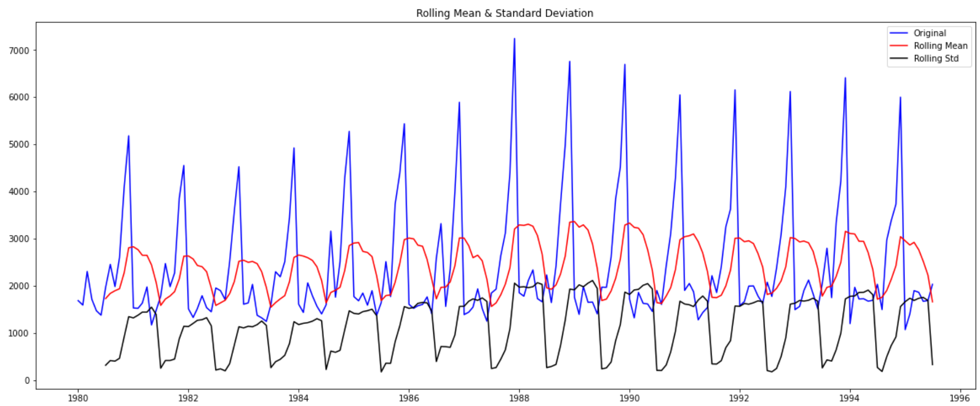


Figure 40: Rolling Mean & Standard Deviation-Sparkling

Results of Dickey-Fuller Test-Sparkling

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	

Table 29:Results of Dickey-Fuller Test-Sparkling

since p-value> 0.05, at alpha 0.05, time series is not stationary. We can take next levels of differencing (a difference of order 1) to make a Time Series stationary.

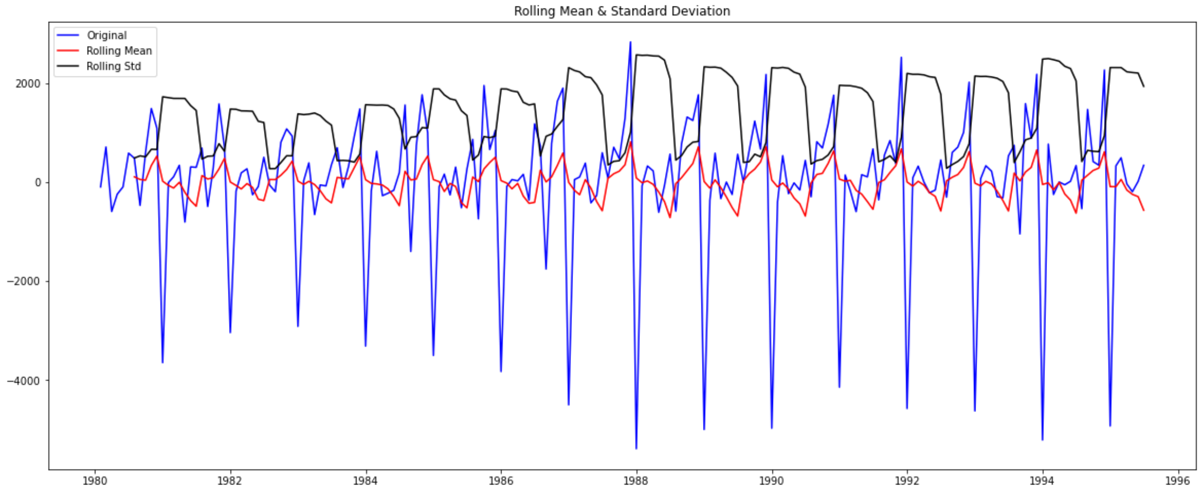


Figure 41: Rolling Mean & Standard Deviation (Next level)-Sparkling

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	

Table 30: Results of Dickey-Fuller Test (Next level)-Sparkling

Observation: After next level of levels of differencing p-value <0.05 therefore series is stationary.

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

Note: The data has some seasonality so ideally, we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model both by looking at the minimum AIC criterion and by looking at the ACF and the PACF plots.

Rose Data:

Since the data is not stationery and seasonal in nature therefore, we choose SARIMAX model.

SARIMAX Result for Rose Data:

Dep. Variable:	y	**No. Observations:**	132
Model:	SARIMAX(1, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-594.281
Date:	Sun, 04 Sep 2022	**AIC**	1200.562
Time:	15:21:31	**BIC**	1217.813
Sample:	01-01-1980	**HQIC**	1207.572
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5364	0.166	-3.241	0.001	-0.861	-0.212
ma.L1	-0.1584	0.146	-1.085	0.278	-0.444	0.128
ma.L2	-0.7064	0.133	-5.327	0.000	-0.966	-0.447
ar.S.L12	0.9865	0.020	49.830	0.000	0.948	1.025
ma.S.L12	-0.7978	0.144	-5.558	0.000	-1.079	-0.516
sigma2	438.1831	56.485	7.758	0.000	327.475	548.891

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	58.92
Prob(Q):	0.99	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.30	**Skew:**	0.73
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.94

Sparkling data:

Since the data is not stationery and seasonal in nature therefore, we choose SARIMAX model.

SARIMAX Result for Sparkling Data:

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 0, 1)x(0, 1, 1, 12)	Log Likelihood	-885.179
Date:	Sun, 04 Sep 2022	AIC	1778.358
Time:	16:37:23	BIC	1789.508
Sample:	01-01-1980	HQIC	1782.886
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	37.3695	26.640	1.403	0.161	-14.843	89.582
ma.L1	0.1882	0.090	2.090	0.037	0.012	0.365
ma.S.L12	-0.4763	0.071	-6.675	0.000	-0.616	-0.336
sigma2	1.478e+05	1.38e+04	10.681	0.000	1.21e+05	1.75e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	48.14
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	3.15	Skew:	0.84
Prob(H) (two-sided):	0.00	Kurtosis:	5.60

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

Q7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Rose Data:

From the above auto_arima method we got best Sarima model (1,1,2)(1,0,1)[12] for Rose data.

SARIMAX Result:

Dep. Variable:	Rose	**No. Observations:**	132
Model:	SARIMAX(1, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-594.281
Date:	Sun, 04 Sep 2022	**AIC**	1200.562
Time:	17:19:40	**BIC**	1217.813
Sample:	01-01-1980	**HQIC**	1207.572
	- 12-01-1990		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	-0.5364	0.166	-3.241	0.001	-0.861	-0.212
ma.L1	-0.1584	0.146	-1.085	0.278	-0.444	0.128
ma.L2	-0.7064	0.133	-5.327	0.000	-0.966	-0.447
ar.S.L12	0.9865	0.020	49.830	0.000	0.948	1.025
ma.S.L12	-0.7978	0.144	-5.558	0.000	-1.079	-0.516
sigma2	438.1831	56.485	7.758	0.000	327.475	548.891

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	58.92
Prob(Q):	0.99	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.30	**Skew:**	0.73
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.94

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

Plotting SARIMAX prediction against actual values.

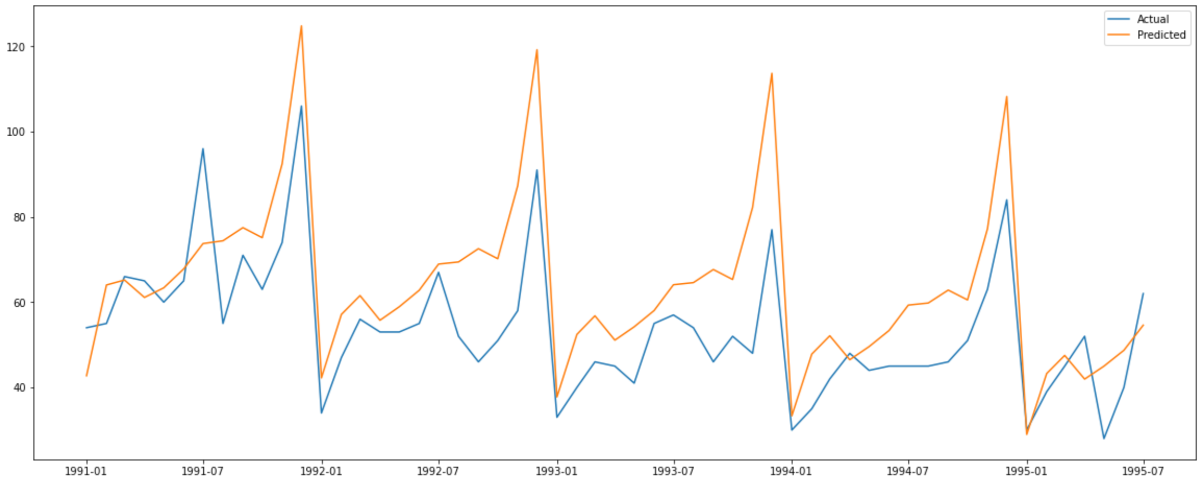


Figure 42:SARIMAX prediction against actual values-Rose

Model Evaluation of SARIMAX:

RMSE:

The RMSE of SARIMAX model is 14.578281918727992

Sparkling Data:

From the above auto_arima method we got best SARIMA model (0,0,1)(0,1,1)[12] for Sparkling data

Dep. Variable:	Sparkling	**No. Observations:**	132
Model:	SARIMAX(0, 0, 1)x(0, 1, 1, 12)	**Log Likelihood**	-886.229
Date:	Sun, 04 Sep 2022	**AIC**	1778.458
Time:	17:57:05	**BIC**	1786.820
Sample:	01-01-1980	**HQIC**	1781.854
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.2040	0.087	2.354	0.019	0.034	0.374
ma.S.L12	-0.4254	0.069	-6.180	0.000	-0.560	-0.290
sigma2	1.504e+05	1.29e+04	11.618	0.000	1.25e+05	1.76e+05

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	43.91
Prob(Q):	0.94	**Prob(JB):**	0.00
Heteroskedasticity (H):	3.12	**Skew:**	0.81
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.48

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

Plotting SARIMAX prediction against actual values.

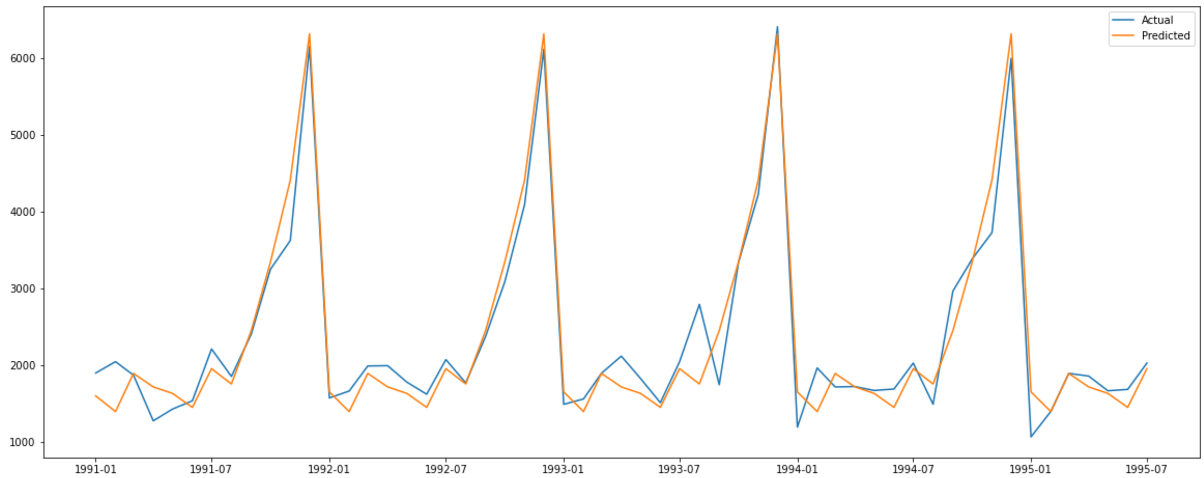


Figure 43:SARIMAX prediction against actual values-Sparkling

Model Evaluation of SARIMAX:

RMSE:

The RMSE of SARIMAX model is 323.7008774953012

ARIMA MODEL- Rose data:

AIC Values for different value combinations.

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748312
ARIMA(0, 1, 2) - AIC:1279.6715288535818
ARIMA(0, 1, 3) - AIC:1280.5453761734655
ARIMA(1, 1, 0) - AIC:1317.3503105381492
ARIMA(1, 1, 1) - AIC:1280.5742295380032
ARIMA(1, 1, 2) - AIC:1279.870723423191
ARIMA(1, 1, 3) - AIC:1281.8707223310003
ARIMA(2, 1, 0) - AIC:1298.6110341605004
ARIMA(2, 1, 1) - AIC:1281.5078621868474
ARIMA(2, 1, 2) - AIC:1281.8707222264284
ARIMA(2, 1, 3) - AIC:1274.6949119626274
ARIMA(3, 1, 0) - AIC:1297.48109172717
ARIMA(3, 1, 1) - AIC:1282.4192776271977
ARIMA(3, 1, 2) - AIC:1283.720740597716
ARIMA(3, 1, 3) - AIC:1278.6588655941036
```

Sorting the above AIC values in the ascending order to get the parameters for the minimum AIC values.

	param	AIC
27	(2, 1, 3)	1274.694912
11	(2, 1, 3)	1274.694912
15	(3, 1, 3)	1278.658866
31	(3, 1, 3)	1278.658866
2	(0, 1, 2)	1279.671529
18	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
22	(1, 1, 2)	1279.870723
19	(0, 1, 3)	1280.545376
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.57423
21	(1, 1, 1)	1280.57423
25	(2, 1, 1)	1281.507862
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
26	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
23	(1, 1, 3)	1281.870722
17	(0, 1, 1)	1282.309832
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
29	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
30	(3, 1, 2)	1283.720741

Result for best parameters (order= (2,1,3))

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.347
Date:                Sun, 04 Sep 2022   AIC                           1274.695
Time:                        19:17:51   BIC                           1291.946
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6781      0.084    -20.035      0.000      -1.842      -1.514
ar.L2         -0.7289      0.084     -8.703      0.000      -0.893      -0.565
ma.L1          1.0450      0.685      1.527      0.127      -0.297       2.387
ma.L2         -0.7716      0.137     -5.636      0.000      -1.040      -0.503
ma.L3         -0.9046      0.622     -1.455      0.146      -2.123       0.314
sigma2       858.3595    576.845      1.488      0.137    -272.237    1988.956
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.45
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================

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

Model Evaluation of ARIMA model-Rose data:

RMSE:

The RMSE for ARIMA model is 36.83800830441206

ARIMA MODEL- Sparkling data:

AIC Values for different value combinations.

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.060015591336
ARIMA(0, 1, 2) - AIC:2234.4083231283275
ARIMA(0, 1, 3) - AIC:2233.994857753515
ARIMA(1, 1, 0) - AIC:2266.6085393190087
ARIMA(1, 1, 1) - AIC:2235.755094673383
ARIMA(1, 1, 2) - AIC:2234.527200452466
ARIMA(1, 1, 3) - AIC:2235.607816390617
ARIMA(2, 1, 0) - AIC:2260.365743968086
ARIMA(2, 1, 1) - AIC:2233.777626239922
ARIMA(2, 1, 2) - AIC:2213.509212306332
ARIMA(2, 1, 3) - AIC:2232.921136688177
ARIMA(3, 1, 0) - AIC:2257.72337899794
ARIMA(3, 1, 1) - AIC:2235.498878057432
ARIMA(3, 1, 2) - AIC:2230.759636959836
ARIMA(3, 1, 3) - AIC:2221.4566102276085
```

Sorting the above AIC values in the ascending order to get the parameters for the minimum AIC values.

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.45661
14	(3, 1, 2)	2230.759637
11	(2, 1, 3)	2232.921137
9	(2, 1, 1)	2233.777626
3	(0, 1, 3)	2233.994858
2	(0, 1, 2)	2234.408323
6	(1, 1, 2)	2234.5272
13	(3, 1, 1)	2235.498878
7	(1, 1, 3)	2235.607816
5	(1, 1, 1)	2235.755095
12	(3, 1, 0)	2257.723379
8	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
4	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

Result for best parameters (order= (2,1,2))

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 04 Sep 2022   AIC                           2213.509
Time:                        19:59:10   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.781      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.217      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.24e+27. Standard errors may be unstable.
```

Model Evaluation of ARIMA model-Sparkling data:

RMSE:

The RMSE for ARIMA model is 1299.9795689481477

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

Rose wine sale data:

Sorted by RMSE values on the Rose wine Test Data:

	Rose Test RMSE
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
SARIMAX	14.578282
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
Alpha =0.017 and Beta=3.23e-5 Double Exponential Smoothing model	15.715112
Alpha=0.07151,Beta=0.453,Gamma=0.0000724, Triple Exponential Smoothing	20.182721
Alpha=0.0987,SimpleExponentialSmoothing	36.816889
ARIMA(2,1,1)	36.838008
Alpha=0.3,SimpleExponentialSmoothing	47.525251
Simple average forecast	53.480857
NaiveModel	79.738550

Table 31:Sorted RMSE values on the Rose wine Test Data:

Sparkling wine sale data:

Sorted by RMSE values on the Sparkling wine Test Data:

	Sparkling Test RMSE
SARIMAX	323.700877
Alpha=0.111,Beta=0.049,Gamma=0.3620 Triple Exponential Smoothing	404.286809
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple average forecast	1275.081804
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,1)	1299.979569
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
Alpha=0.3,SimpleExponentialSmoothing	1935.507132
Alpha=0.688,Beta=0.0000999,DoubleExponentialSmoothing	2007.238526
NaiveModel	3864.279352

Table 32:Sorted RMSE values on the Sparkling wine Test Data

Observation:

- Rose wine data: 2pointTrailingMovingAverage is the best among all other models with lowest RMSE of 11.529409
- Since Rose data set has clear component of seasonality SARIM. Therefore, SARIMA model with parameters (1,1,2) x (0,1,0,12) is selected for forecasting time line series.
- Sparkling wine data: SARIMAX is the best among all other models with lowest RMSE of 323.700877
- Rose wine data: Naïve Model has the lowest RMSE of 79.738550
- Sparkling wine data: Naïve Model has the lowest RMSE of 3864.279352

Q9: Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future.

Rose date:

Since Rose data set has clear component of seasonality SARIM. Therefore, SARIMA model with parameters (1,1,2) x (0,1,0,12) is selected for forecasting time line series.

SARIMAX Result

SARIMAX Results

Dep. Variable:	Rose	**No. Observations:**	187
Model:	SARIMAX(1, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-814.860
Date:	Sun, 04 Sep 2022	**AIC**	1641.720
Time:	20:45:36	**BIC**	1661.074
Sample:	01-01-1980	**HQIC**	1649.563
	- 07-01-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5159	0.155	-3.333	0.001	-0.819	-0.213
ma.L1	-0.1849	0.144	-1.288	0.198	-0.466	0.097
ma.L2	-0.6722	0.126	-5.339	0.000	-0.919	-0.425
ar.S.L12	0.9827	0.014	72.553	0.000	0.956	1.009
ma.S.L12	-0.7589	0.091	-8.361	0.000	-0.937	-0.581
sigma2	336.9175	29.039	11.602	0.000	280.002	393.833

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	171.56
Prob(Q):	0.96	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.14	**Skew:**	0.79
Prob(H) (two-sided):	0.00	**Kurtosis:**	7.43

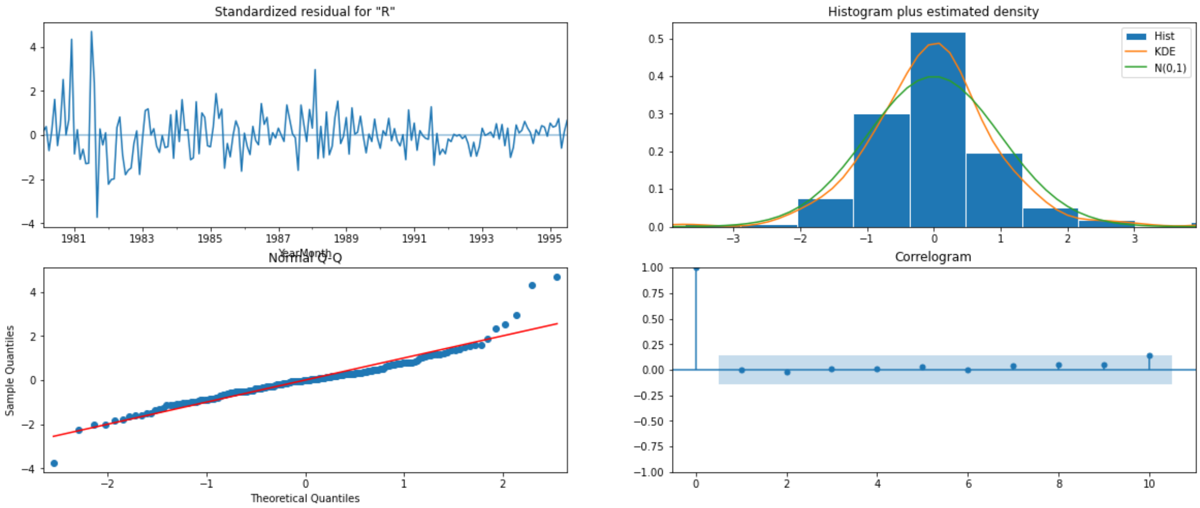
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table 33: Rose SARIMAX Best result

Prediction for 12 months- Rose Data:

```
1995-08-01    43.309050
1995-09-01    42.655788
1995-10-01    44.791222
1995-11-01    54.184940
1995-12-01    81.724082
1996-01-01    20.823473
1996-02-01    29.578875
1996-03-01    35.776714
1996-04-01    36.547765
1996-05-01    30.183710
1996-06-01    36.419928
1996-07-01    46.471402
Freq: MS, Name: Rose SARIMA 12 months Predictions, dtype: float64
```

Diagnostic model plot:



Observation: The above diagnostics plots look fine.

Plotting the 12 months data prediction:

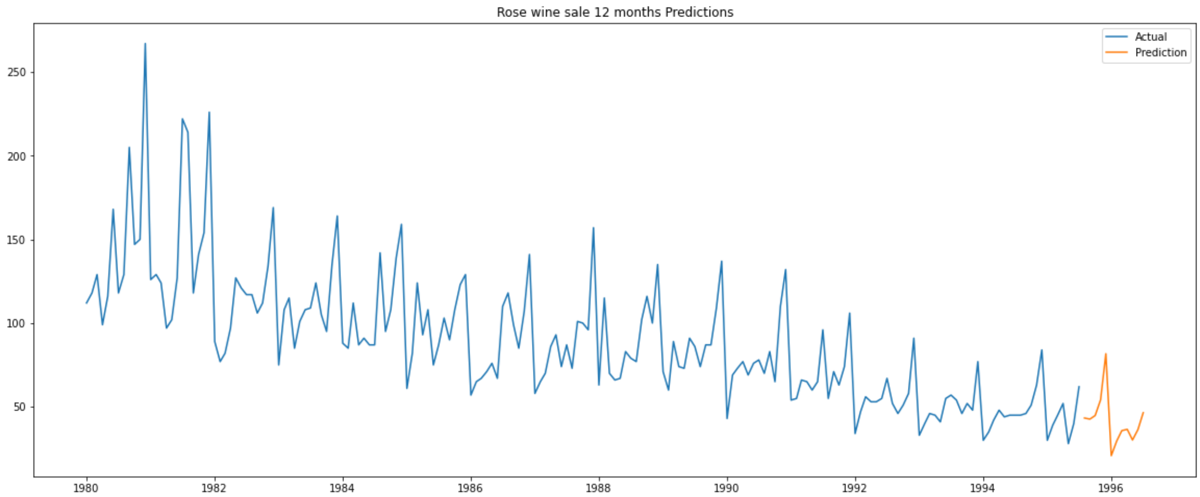


Figure 44:the 12 months data prediction-Rose

Sparkling date:

Since Sparkling data set has clear component of seasonality SARIM. Therefore, SARIMA model with parameters (0,0,1) (0,1,1)[12] is selected for forecasting time line series.

SARIMAX RESULT:

Dep. Variable:	Sparkling	**No. Observations:**	187
Model:	SARIMAX(1, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-1532.670
Date:	Sun, 04 Sep 2022	**AIC**	3077.341
Time:	23:19:47	**BIC**	3096.695
Sample:	01-01-1980	**HQIC**	3085.184
	- 07-01-1995		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.9580	1.333	0.719	0.472	-1.655	3.571
ma.L1	-1.9341	1.457	-1.328	0.184	-4.789	0.921
ma.L2	0.9346	1.434	0.652	0.515	-1.876	3.745
ar.S.L12	0.8962	0.168	5.346	0.000	0.568	1.225
ma.S.L12	-0.1936	0.589	-0.329	0.742	-1.348	0.961
sigma2	1.826e+06	5.58e-06	3.28e+11	0.000	1.83e+06	1.83e+06

Ljung-Box (L1) (Q):	5.88	**Jarque-Bera (JB):**	2583.15
Prob(Q):	0.02	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.35	**Skew:**	2.74
Prob(H) (two-sided):	0.00	**Kurtosis:**	20.41

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.42e+27. Standard errors may be unstable.

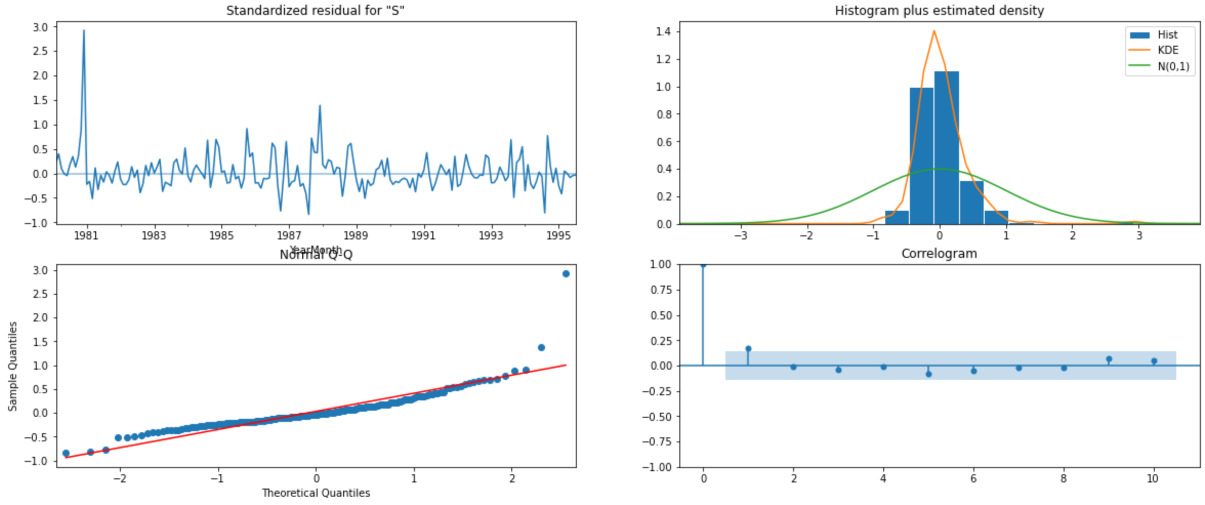
Table 34: Sparkling SARIMAX Best result

Prediction for 12 months- Rose Data:

```
1995-08-01    1788.362510
1995-09-01    2701.198690
1995-10-01    3234.930098
1995-11-01    3627.652810
1995-12-01    5588.417041
1996-01-01    1265.909399
1996-02-01    1606.791928
1996-03-01    1931.627812
1996-04-01    1913.302536
1996-05-01    1763.193454
1996-06-01    1769.671889
1996-07-01    2074.065962
Freq: MS, Name: Sparkling SARIMA 12 months Predictions, dtype: float64
```

Table 35: 12 months data prediction- Sparkling data

Diagnostic model plot:



Observation: The above diagnostics plots look fine.

lotting the 12 months data prediction:

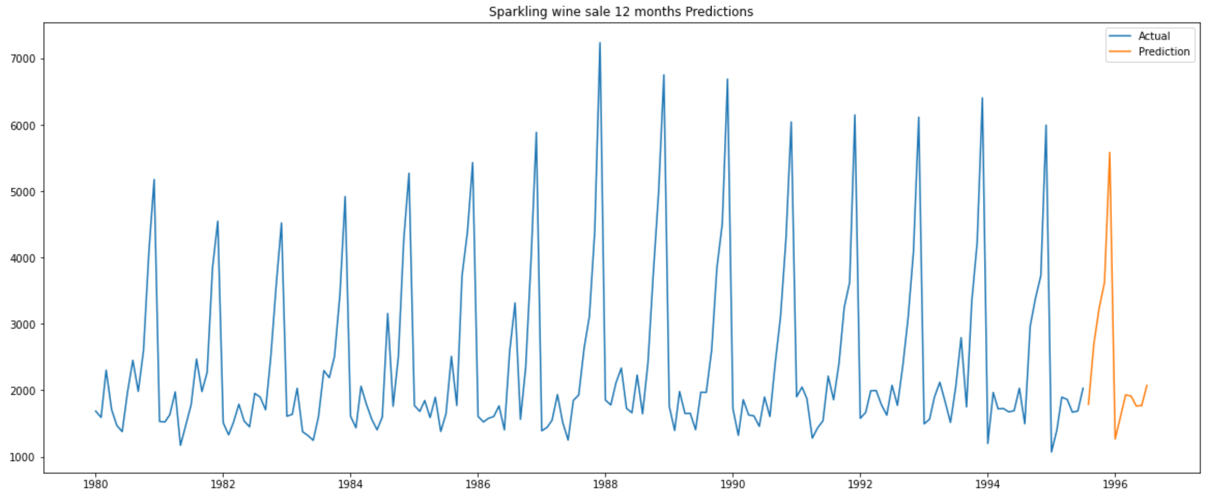


Figure 45: Sparkling wine sale 12 months Predictions

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

Insights:

- On observing RMSE values of different types of model (Rose) that we developed such as regression (ARM), simple average models (SAM), and exponential smoothing model (ESM) on the training data, we found that SARIMAX model outperformed all the other model because it had the lowest RMSE score (considering seasonality).

- Two major observations that we see in the dataset are as follows:
  1. The Sales of Rose wine is decreasing with time. May be the quality of the wine is decreasing or the competitors have manufactured same quality wine with low price. Hence Rose wine's company needs to develop the quality of the wine and also work on the cost of it so that customer will be inclined to buy this wine.
  2. In case of Sparkling wine, the sales are following seasonal pattern but the sales value is not decreasing. It is also found that the sales of the Sparkling wine always reach its peak at the end of the year. Hence company should work on developing marketing schemes towards the end of the year so that more customers are impelled to buy the same.

- Trend in sales of Rose is continuously decreasing over the period. Detailed study may be required to see whether decreasing trend is due to change in customer preference or due to substitution. Seasonality of sales is observed, and higher sales is maintained in the end of the year. Some promotion schemes and improvement / quality enhancers in the product can be examined so as to attract new young generation customers.
- Sales in Sparking does not have uniform trend but increased in some years and decreased later. Business study may be done to find why sales are not increasing and what the contributing factors. Study can also include to see which wine product has substituted/ had higher sales in the years of low sales of Sparkling. With promotion and focussed effort with micro detailing it may be feasible to increase the sales. Sales of Sparkling wine higher in the later part of the year. This may be due to climatic condition of the geography under study.

---E N D--