TIME SERIES FORECASTING PROJECT REPORT

Problem 2: Rose Wine

By Karthik Sreeram R

MAY 22, 2021

UNIVERSITY OF TEXAS AT AUSTIN AND GREAT LAKES

Purpose
This document is the business report for my final project in the subject “Time. Series Forecasting” This document gives us a detailed explanation of various approaches used, their insight and inferences. Tools used analysis: Python and Jupiter notebook. Packages used: NumPy, pandas, seaborn, os, matplotlib, stats model, sklearn and pylab

Contents

Problem 2: Rose Wine

Business scenario

For this assignment, the data of different types of wine sales in the 20th century is to be analyzed. Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century.

Introduction:

The purpose of this assignment is to understand the time series data, do exploratory analysis, perform decomposition to understand the trend and seasonality of the data , train the data with different models of forecasting to predict the future sales of the sparkling wine .It will help the ABC estate to pre stock the Rose wine for future sales based on demand (predicted sales).

Data has two fields.

YearMonth – monthly data

Rose– sales count of sparkling wine.

2.1) Reading the data as a Time Series data and plotting the data

a.) Dataset Head

]:

	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

Inference:

The data is month wise data starting from January 1980. The data format is year and month (YYYY-mm)

b.) Dataset Tail

:

	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

Inference:

The data ends in July 1995. The row count and total number of months between January 1980 and July 1995 matches (i.e., 187 months)

c.) Preprocessing of time series data

1. Creating dummy month wise date data in time stamp format from the January 1980 to July 1995.

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

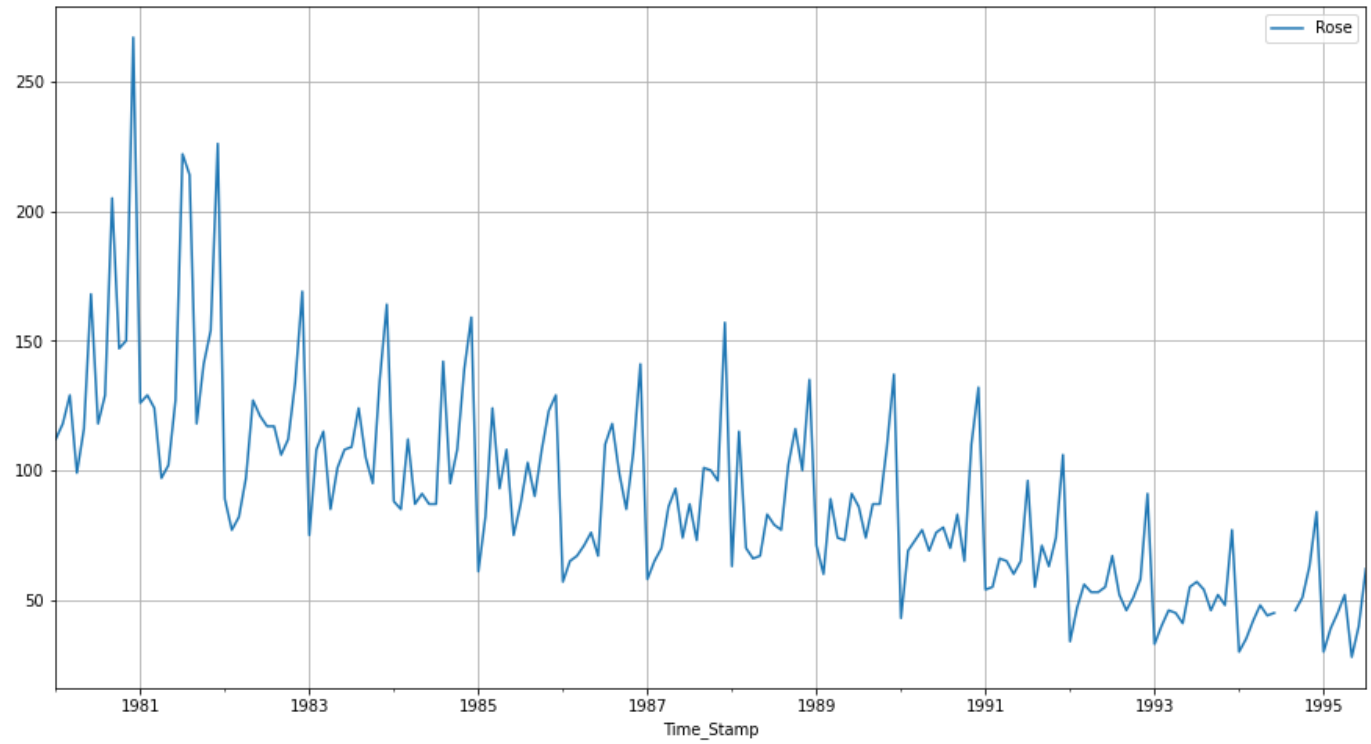
2. Add dummy date column to the original data set.

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

Inference:

For time series model data must be in the format YYYY-mm-dd i.e., time stamp format. Our data has date data as YYYY-mm format. So, a dummy time stamp is created to replace the YearMonth field and changed into index.

c.) Plotting the time series data.



Inference:

Looks like there is a downward trend in the data. Data has few missing points in between 1993 and 1995.

2.2) Performing Exploratory Data Analysis to understand the data and to perform decomposition.

a) Description of Data

Basic Descriptive Stats of Time series

	Rose
count	185.000
mean	90.395
std	39.175
min	28.000
25%	63.000
50%	86.000
75%	112.000
max	267.000

Count is 185 but date count is 187. So, there are two missing data

b) Handling Missing data:

1. Count of missing data

```
Rose    2
dtype: int64
```

2. row of missing data

0]:

Time_Stamp	Rose
1994-07-31	NaN
1994-08-31	NaN

3. count of rose sales before and after the missing data

Time_Stamp	Rose
1994-05-31	44.0
1994-06-30	45.0
1994-07-31	NaN
1994-08-31	NaN
1994-09-30	46.0
1994-10-31	51.0

4. After Handling Missing data

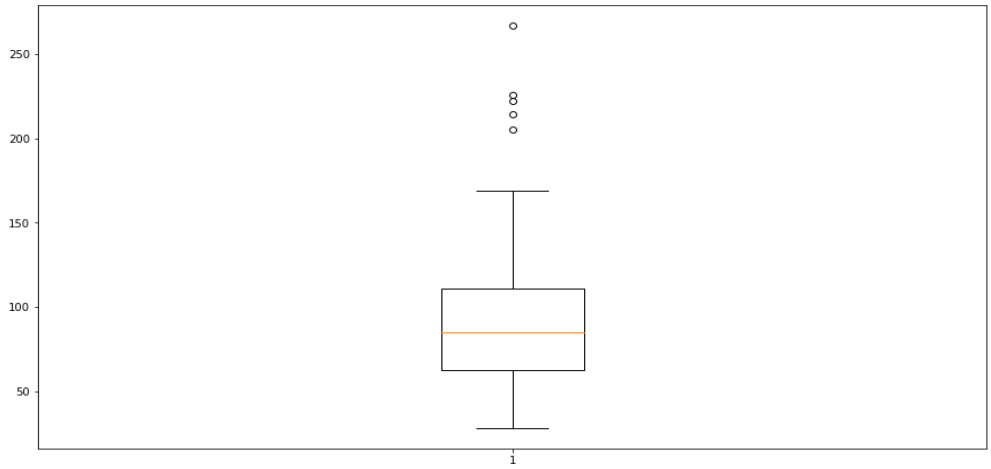
Time_Stamp	Rose
1994-05-31	44.000000
1994-06-30	45.000000
1994-07-31	45.333333
1994-08-31	45.666667
1994-09-30	46.000000
1994-10-31	51.000000

```
count of missing null values
Rose    0
dtype: int64
```

Inference:

There were two missing data. It was handled using Linear Interpolation technique.

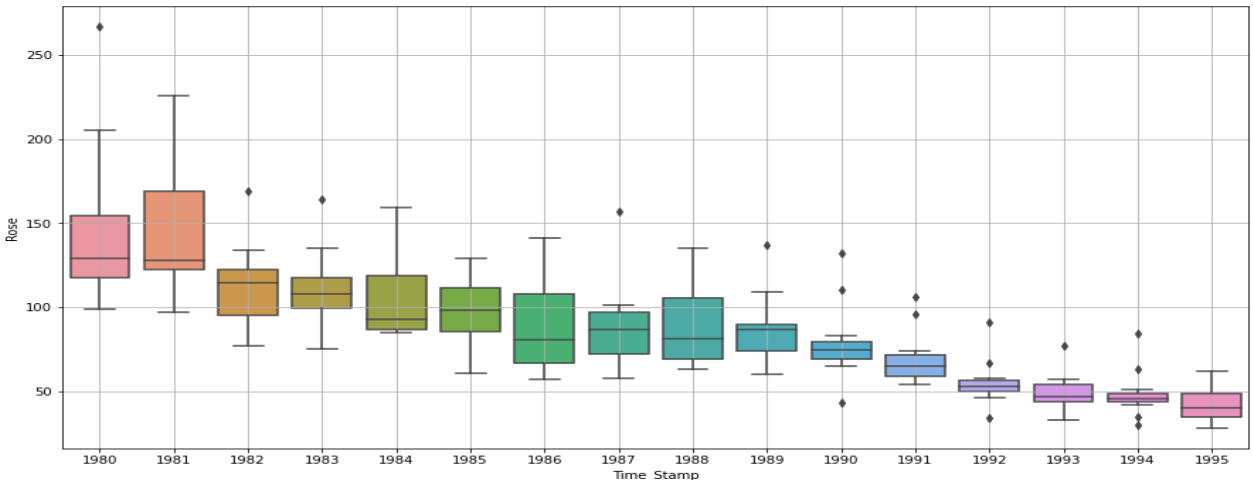
c) BOX PLOT:



Inference:

More than 75 percent of sales quantity fall below 112. Average sale count is 90. There was missing data and its handled.

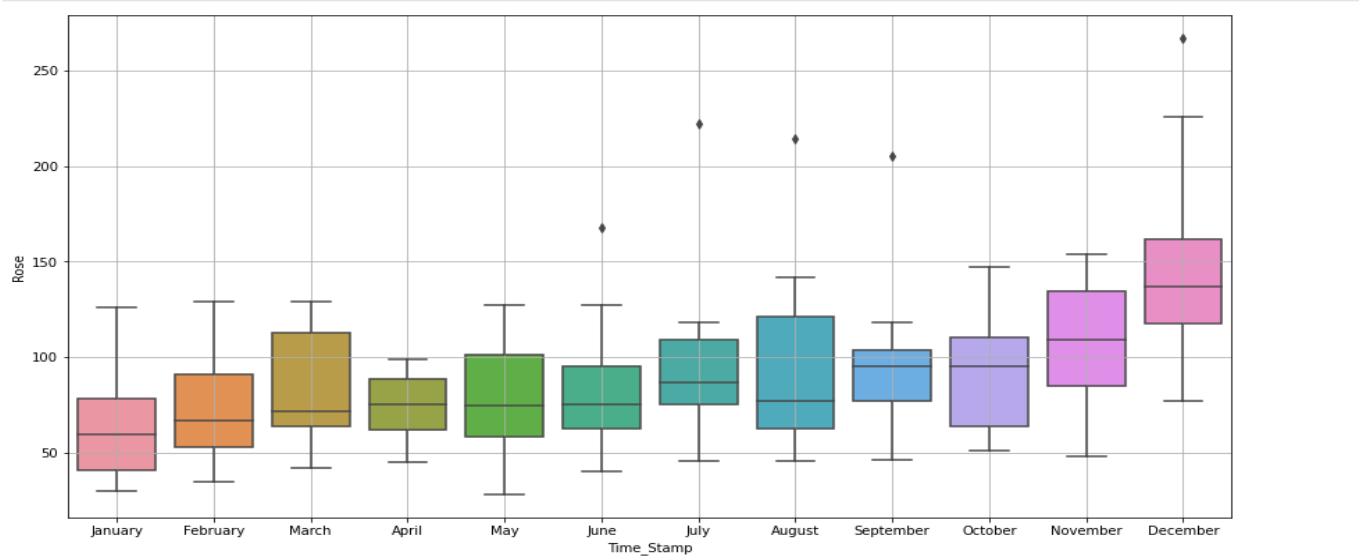
d.) Box plot yearly



Inference:

Year over year comparison shows that the data has a decreasing sales patterns .

e) Box plot Monthly

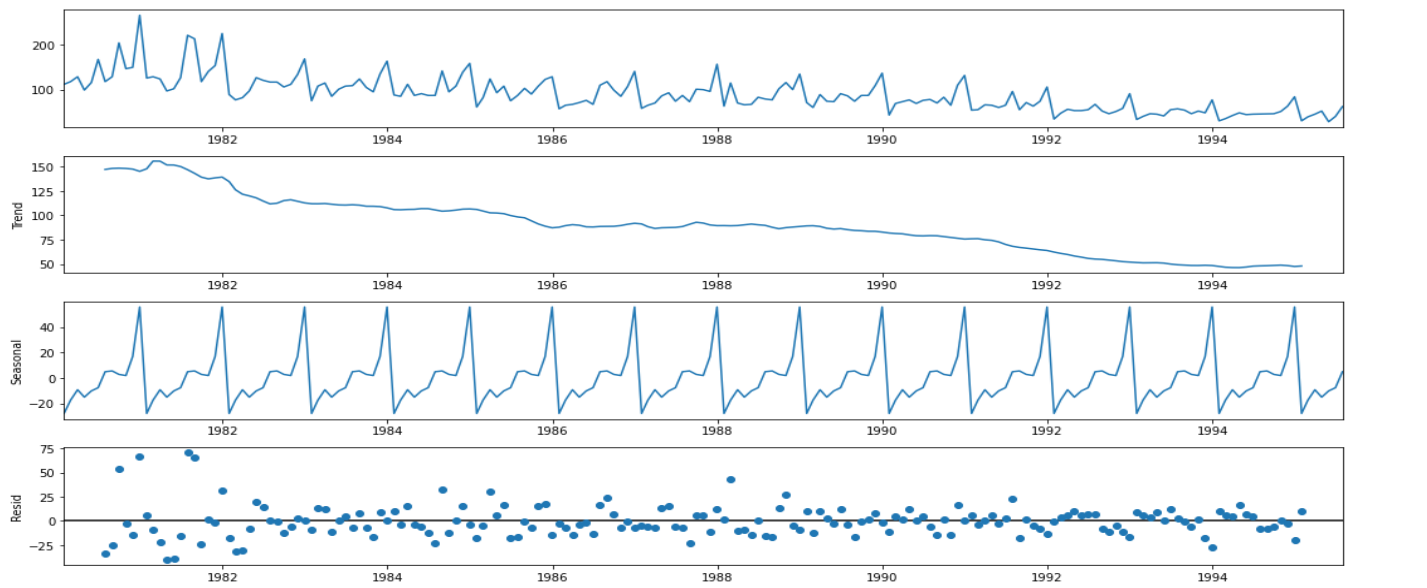


Inference:

Month over month comparison shows that the data has very less seasonality pattern. The maximum sales are during December and minimum sales is in April.

f) Decomposition of Rose variable time series

1. Addictive decomposition

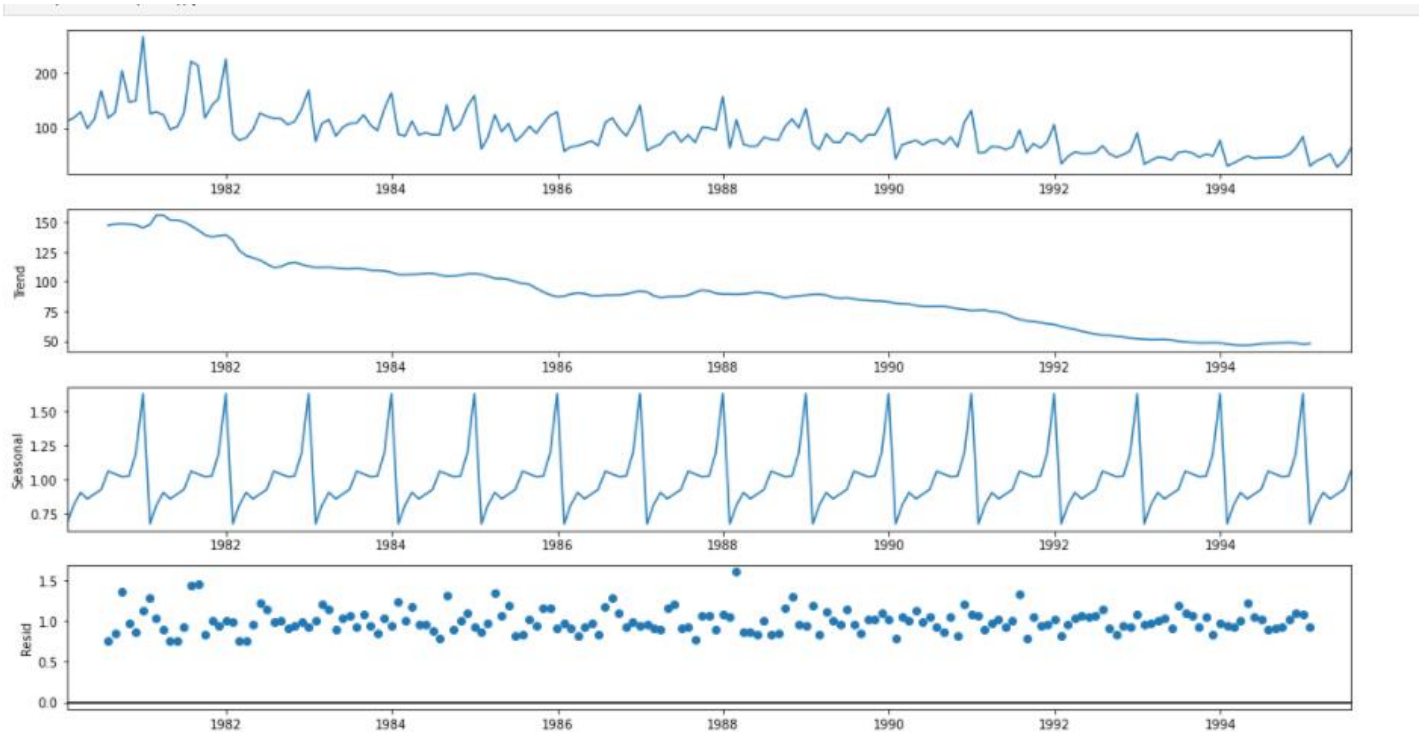


```
Trend
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
Name: trend, dtype: float64

Seasonality
 Time_Stamp
1980-01-31   -27.908647
1980-02-29   -17.435632
1980-03-31    -9.285830
1980-04-30   -15.098330
1980-05-31   -10.196544
1980-06-30    -7.678687
1980-07-31     4.896908
1980-08-31     5.499686
1980-09-30     2.774686
1980-10-31     1.871908
1980-11-30    16.846908
1980-12-31    55.713575
Name: seasonal, dtype: float64

Residual
 Time_Stamp
1980-01-31          NaN
1980-02-29          NaN
1980-03-31          NaN
1980-04-30          NaN
1980-05-31          NaN
1980-06-30          NaN
1980-07-31   -33.980241
1980-08-31   -24.624686
1980-09-30    53.850314
1980-10-31    -2.955241
1980-11-30   -14.263575
1980-12-31    66.161425
Name: resid, dtype: float64
```

2. Multiplicative Decomposition



```
Trend
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
Name: trend, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    0.670111
1980-02-29    0.806163
1980-03-31    0.901164
1980-04-30    0.854024
1980-05-31    0.889415
1980-06-30    0.923985
1980-07-31    1.058038
1980-08-31    1.035881
1980-09-30    1.017648
1980-10-31    1.022573
1980-11-30    1.192349
1980-12-31    1.628646
Name: seasonal, dtype: float64

Residual
 Time_Stamp
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    0.758258
1980-08-31    0.840720
1980-09-30    1.357674
1980-10-31    0.970771
1980-11-30    0.853378
1980-12-31    1.129646
Name: resid, dtype: float64
```

Inference:

seasonality and residual components are independent of the trend. So, it is addictive.

g) Checking stationarity of whole data

```
checking seasonlity on whole data
DF test statistic is -2.240
DF test p-value is 0.46713716277931483
Number of lags used 13
fail to reject null hypothesis . its not stationary (p value is > 0.05 (alpha))
```

Inference:

Whole Data is not stationary at α = 0.05

2.3) Splitting the data into training and test. The test data starts in 1991.

a) Head and tail of train data

Head of train data

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

Tail of train data

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

b) Head and tail of test data

Head of test data

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

Tail of test data

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

Inference

Data is split into train and split. Train data is from January 1980 December 1990. Test data is from Jan 1991.

2.4) Building various time series models on the training data and evaluating the model performance using RMSE on the test data.

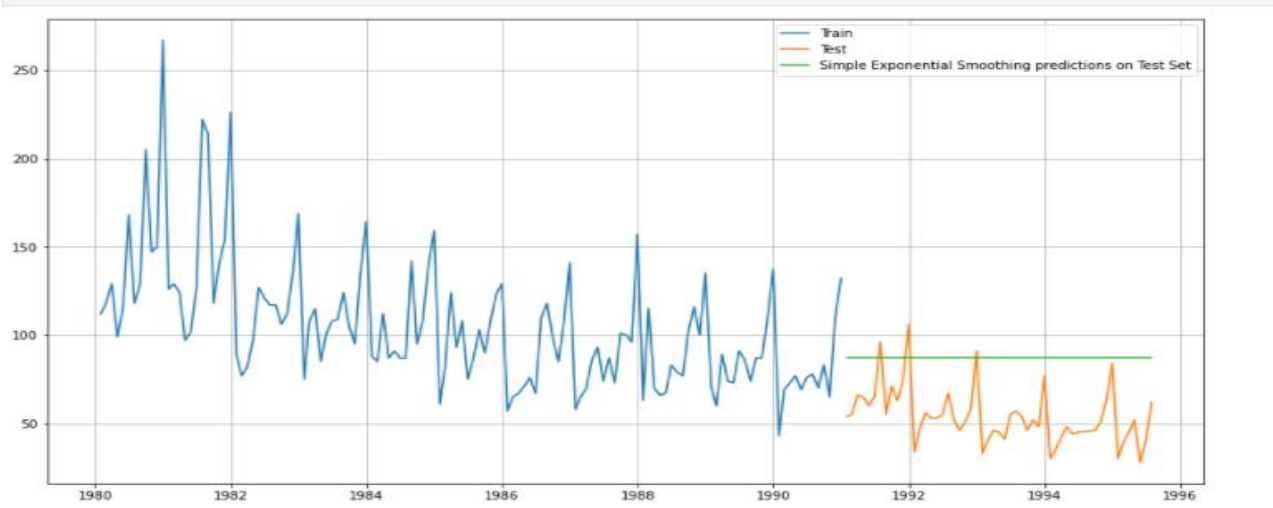
Exponential Smoothing

a.) Simple Exponential Smoothing with additive errors.

1. Autofit Params

```
{'smoothing_level': 0.09874920899865502,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.3871074301239,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Simple Exponential Smoothing prediction plot on test data



3. RMSE Score of Simple Exponential Smoothing:

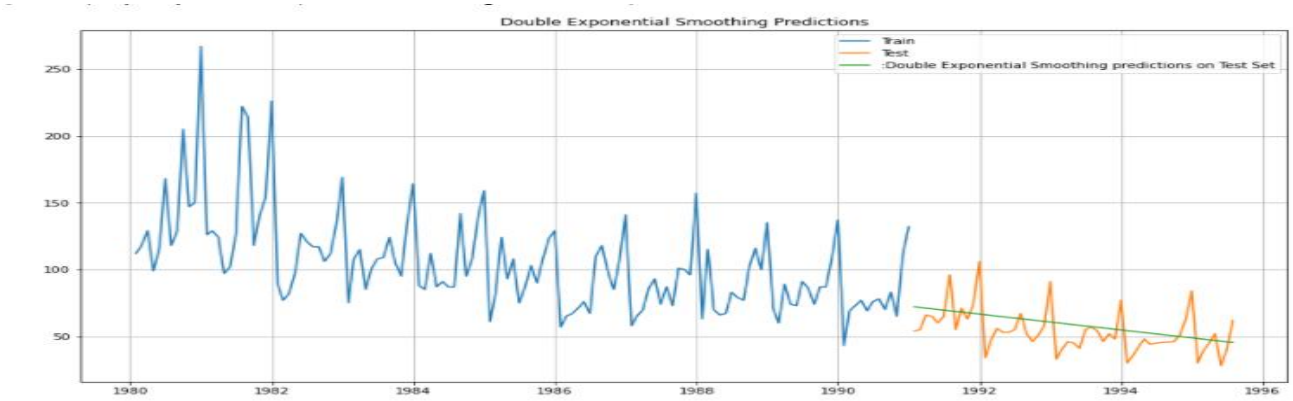
```
SES RMSE: 36.79622482462249
```

b.) Double Exponential Smoothing - Holt's linear method with additive errors

1. Autofit Params

```
{'smoothing_level': 1.4901161193847656e-08,
 'smoothing_trend': 5.448169774560283e-09,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.81762949544608,
 'initial_trend': -0.494350728399512 3,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Double Exponential Smoothing prediction plot on test data



3. RMSE Score of double Exponential Smoothing:

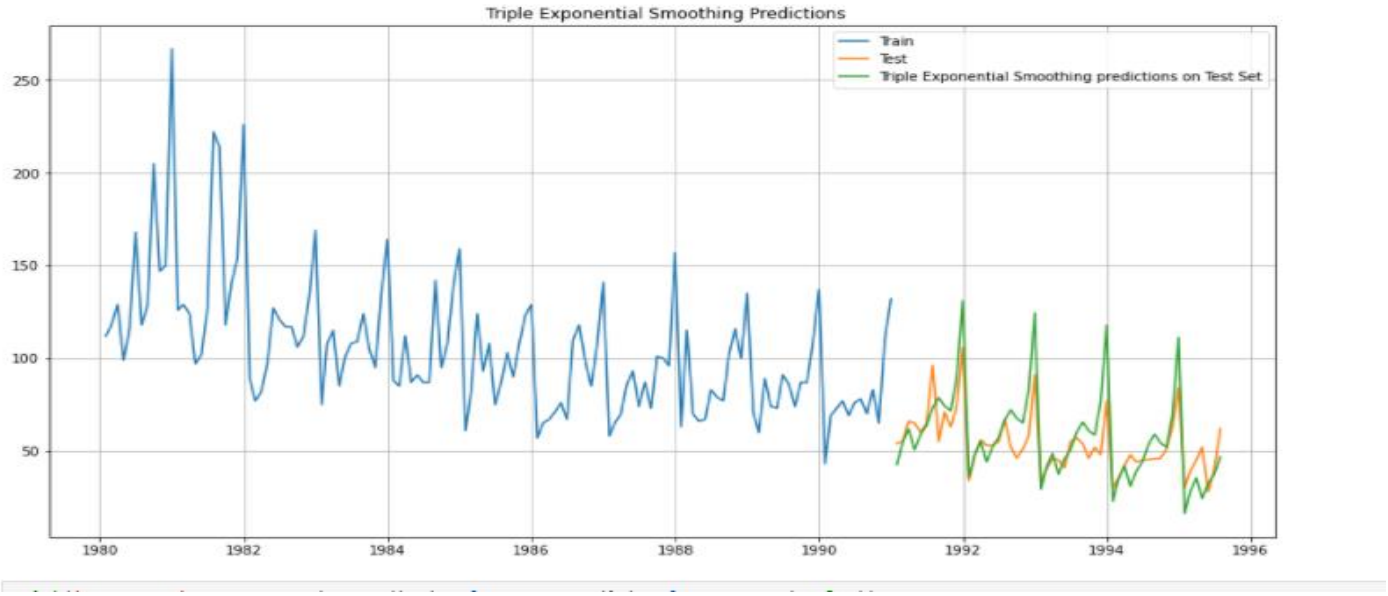
```
DES RMSE: 15.27096843395306
```

c.) Triple Exponential Smoothing (addictive) - Holt Winter's linear method with additive errors

1. Autofit Params

```
{'smoothing_level': 0.08491574907842013,
 'smoothing_trend': 5.5205494088745035e-06,
 'smoothing_seasonal': 0.0005477182208247348,
 'damping_trend': nan,
 'initial_level': 147.05898703809248,
 'initial_trend': -0.5496981430927392,
 'initial_seasons': array([-31.16021285, -18.81317648, -10.81406896, -21.41413199,
       -12.6036696 ,  -7.23553106,   2.76744902,   8.85548059,
         4.83969803,   2.95125217,  21.07934859,  63.31472515]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Triple Exponential Smoothing prediction plot on test data



3. RMSE Score of Triple Exponential Smoothing (Addictive errors):

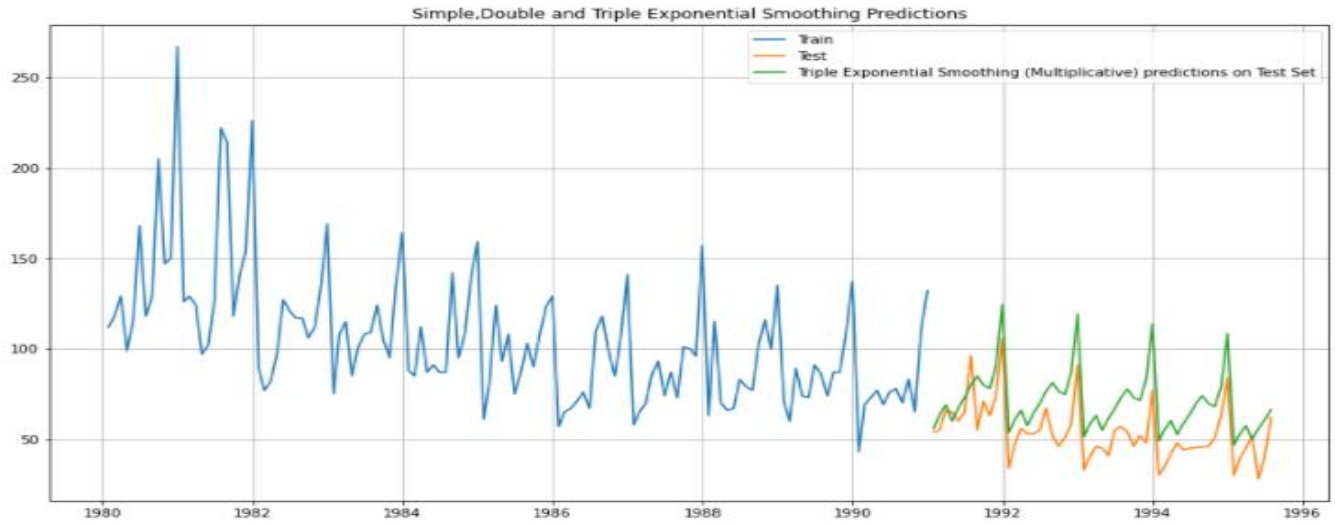
```
TES RMSE: 14.24323950074202
```

d.) Triple Exponential Smoothing (Multiplicative) - Holt Winter's linear method

1. Autofit Params

```
{'smoothing_level': 0.07736040004765096,
 'smoothing_trend': 0.03936496779735522,
 'smoothing_seasonal': 0.0008375039104357999,
 'damping_trend': nan,
 'initial_level': 156.90674503596637,
 'initial_trend': -0.9061396720042346,
 'initial_seasons': array([0.7142168 , 0.80982439, 0.88543128, 0.77363782, 0.87046319,
        0.94699283, 1.04196135, 1.11012703, 1.04835489, 1.0276963 ,
        1.19783562, 1.6514144 ]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Triple Exponential Smoothing prediction plot on test data



3. RMSE Score of Triple Exponential Smoothing (Multiplicative):

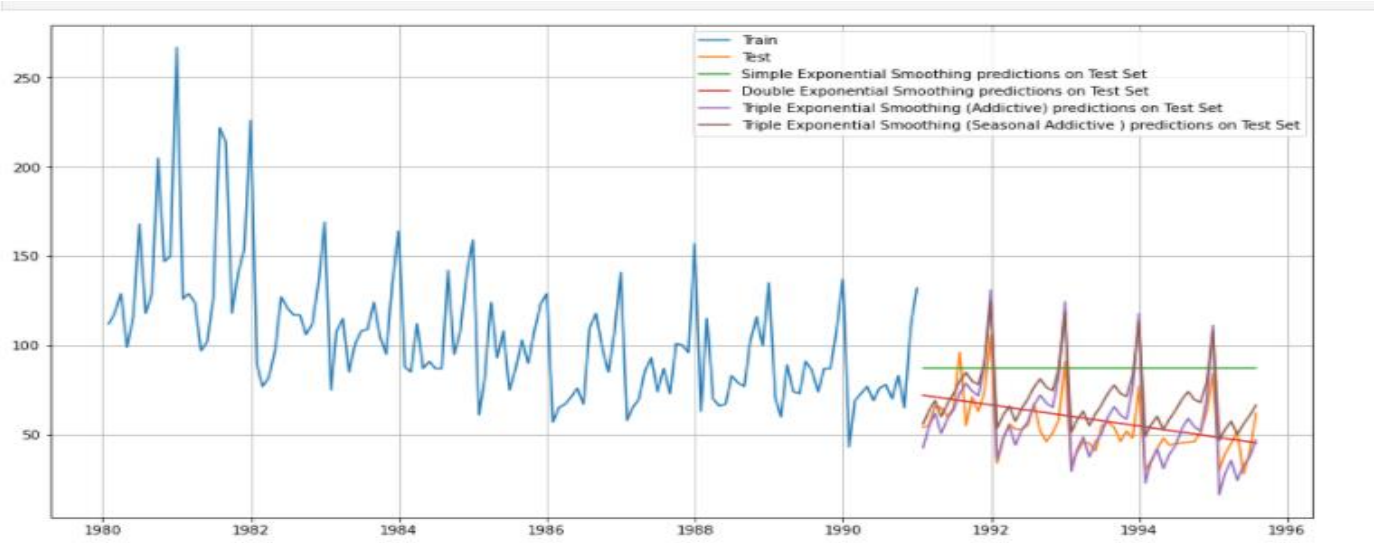
```
TES_am RMSE: 19.113110215160134
```

Inference on Exponential Smoothing:

1. Comparison Table of RMSE OF Different model

	Test RMSE
SES	36.796225
DES	15.270968
TES A	14.243240
TES SM	19.113110

2. Comparison of prediction on Multiple model



3. Insights

In exponential smoothing, Holt Winter's linear method with additive errors (Triple Exponential Smoothing (addictive)) is the best model based on least RMSE score.

Regression, Naïve forecast, and simple average models.

a.) Linear Regression.

1. Creating linear instance (according to date)

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 5
8, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110,
111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176,
177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]
```

2. Head of data set after adding linear Instance.

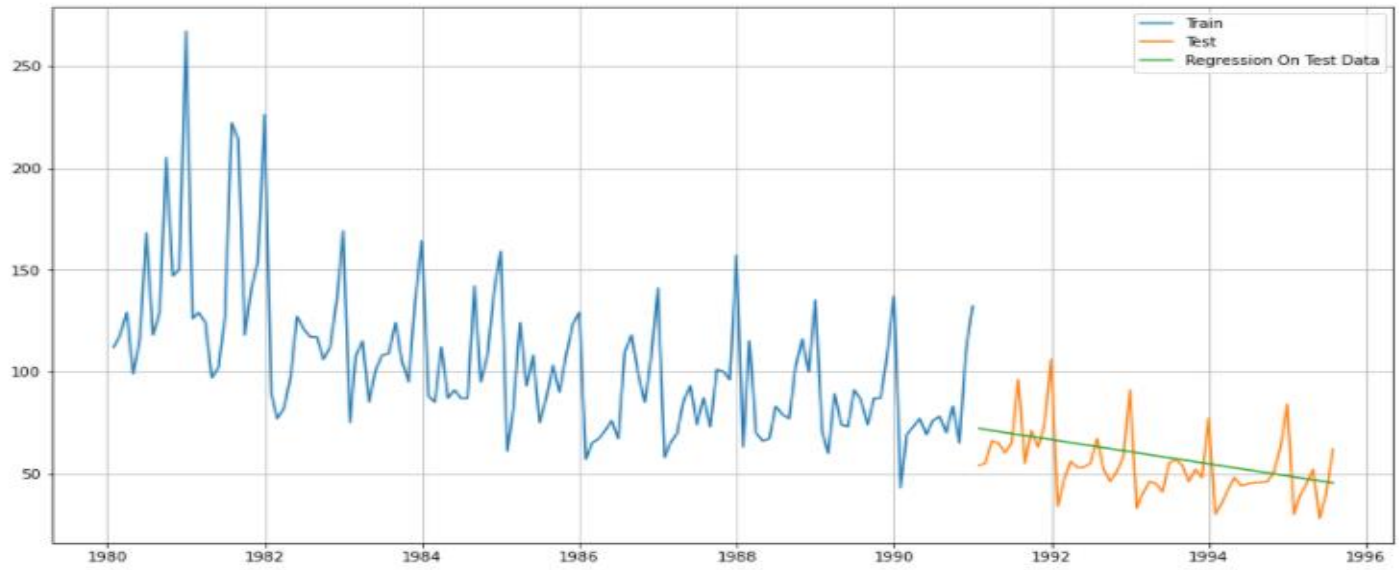
]:

Time_Stamp	Rose	time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

3. Building Linear Regression

```
LinearRegression()
```

4. Linear Regression prediction plot on test data



5. RMSE Score of Linear Regression:

```
For Regression  forecast on the Test Data,  RMSE is 15.269
```

b.) Naïve Approach

1. Tail of train data

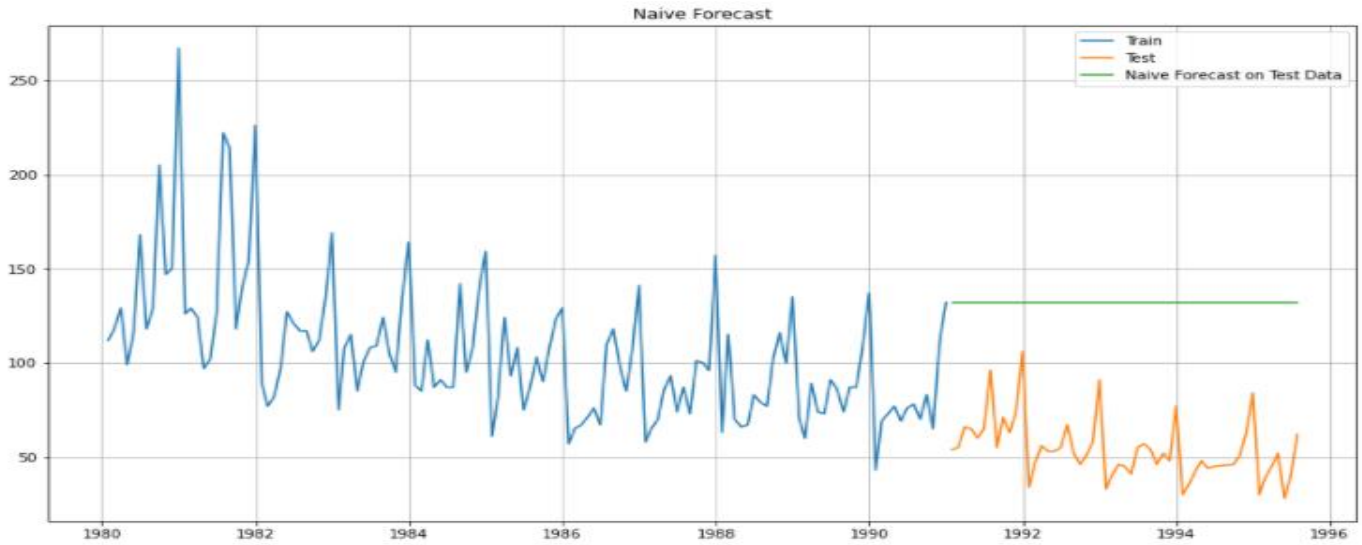
]:

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

2. Head of Test data after applying Naïve Approach.

```
]: Time_Stamp
   1991-01-31    132.0
   1991-02-28    132.0
   1991-03-31    132.0
   1991-04-30    132.0
   1991-05-31    132.0
   Name: naive, dtype: float64
```

3. Naïve Approach prediction plot on test data



4. RMSE Score of Naïve Approach:

```
For Naive Forecast ,  RMSE is 79.719
```

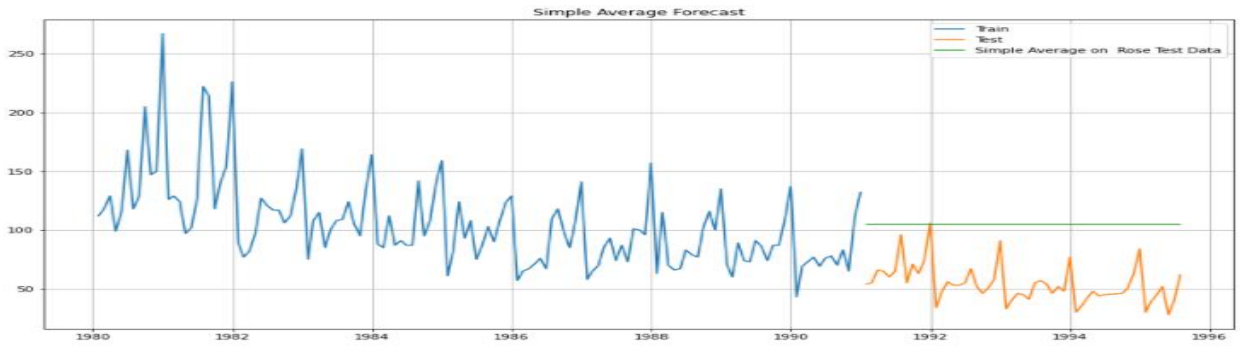
c.) Simple Average

1. Head of data set after creating mean forecast.

]:

Time_Stamp	Rose	mean_forecast
1991-01-31	54.0	104.939394
1991-02-28	55.0	104.939394
1991-03-31	66.0	104.939394
1991-04-30	65.0	104.939394
1991-05-31	60.0	104.939394

2. Simple average prediction plot on test data



3. RMSE Score of simple Average:

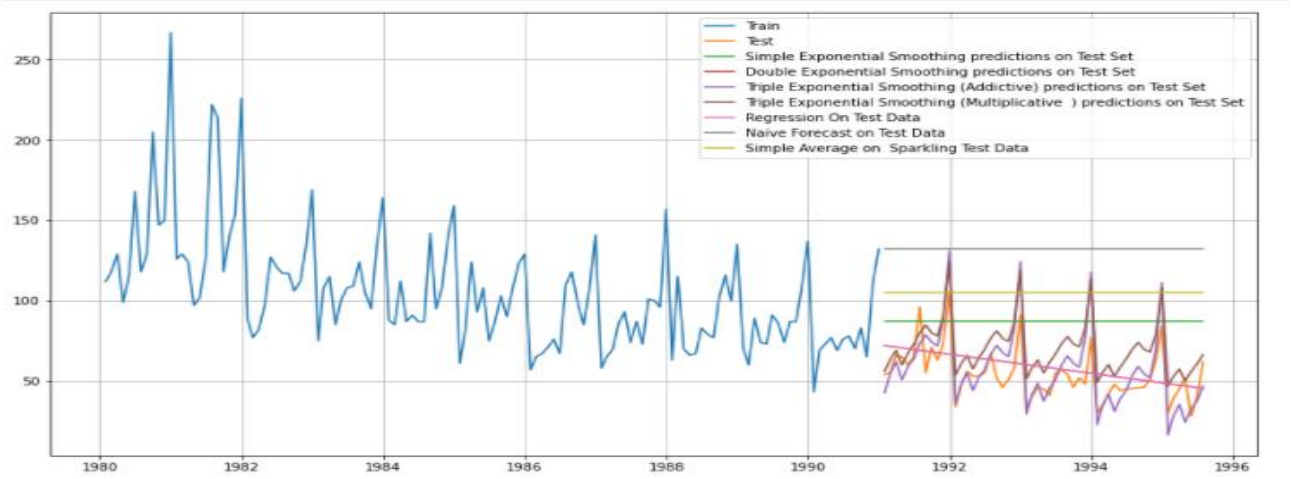
```
For Simple Average forecast on the Test Data,  RMSE is 53.461
```

Inference on above given models:

1. Comparison Table of RMSE OF Different model

	Test RMSE
SES	36.796225
DES	15.270968
TES A	14.243240
TES SM	19.113110
Regression	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570

2. Comparison of prediction on Multiple model



3. Insights

In above analyzed models, Holt Winter's linear method with additive errors (Triple Exponential Smoothing (addictive)) is the best model based on least RMSE score. I.e. it will be able to forecast the sales with least errors compared to other analyzed models.

2.5) Checking and changing the training data into stationary data using appropriate statistical tests and methods .Stationarity is checked at alpha = 0.05.

a.) Augmented Dickey–Fuller test Hypothesis for stationary data

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

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

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the alpha value.

b.) ADF test on train data

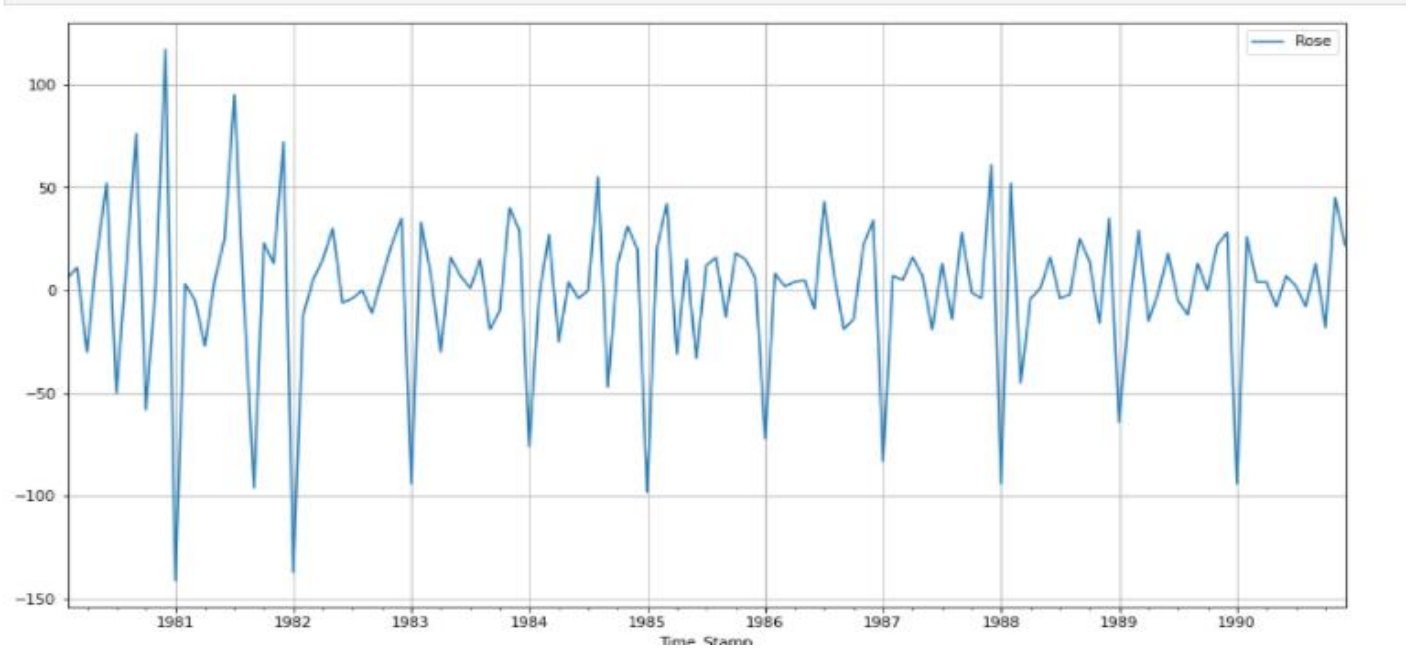
```
DF test statistic is -1.686
DF test p-value is 0.7569093051047049
Number of lags used 13
fail to reject null hypoathes . its not stationary
```

The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationarize the Time Series.

c.) ADF test on train data after one differencing

```
DF test statistic is -6.804
DF test p-value is 3.894831356783106e-08
Number of lags used 12
reject Null hypothesis  ie its statioary
```

d.) Plotting of Data set after one differencing.



Comment / Inference:

Actual Train data is not stationary. After one differencing data has become stationary

Testing for stationarity is very important because the whole results of the regression might be fabricated thus predcation may not be proper if data is nonstationary.

2.6) Building 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 evaluating the performance of this model on the test data using RMSE.

a.) ARIMA model

1.Creating different parameters for the model.

```
Examples of the parameter combinations for the Model
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary.

2.Building Arima model with different parameters

```
ARIMA(0, 1, 0) - AIC:1333.154672912435
ARIMA(0, 1, 1) - AIC:1282.3098319748347
ARIMA(0, 1, 2) - AIC:1279.671528853577
ARIMA(0, 1, 3) - AIC:1280.5453761734668
ARIMA(1, 1, 0) - AIC:1317.350310538156
ARIMA(1, 1, 1) - AIC:1280.5742295380078
ARIMA(1, 1, 2) - AIC:1279.870723423192
ARIMA(1, 1, 3) - AIC:1281.8707223309964
ARIMA(2, 1, 0) - AIC:1298.6110341604892
ARIMA(2, 1, 1) - AIC:1281.5078621868524
ARIMA(2, 1, 2) - AIC:1281.870722226448
D:\anocondal\lib\site-packages\statsmodels
  warnings.warn("Maximum Likelihood optim
ARIMA(2, 1, 3) - AIC:1274.6951715918117
ARIMA(3, 1, 0) - AIC:1297.4810917271716
ARIMA(3, 1, 1) - AIC:1282.4192776271911
D:\anocondal\lib\site-packages\statsmodels
  warn('Non-stationary starting autoregres
D:\anocondal\lib\site-packages\statsmodels
  warn('Non-invertible starting MA parame
ARIMA(3, 1, 2) - AIC:1283.7207405977156
ARIMA(3, 1, 3) - AIC:1278.6552365209318
```

3.Head of the Arima models with AIC score in ascending order

	param	AIC
11	(2, 1, 3)	1274.695172
15	(3, 1, 3)	1278.655237
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

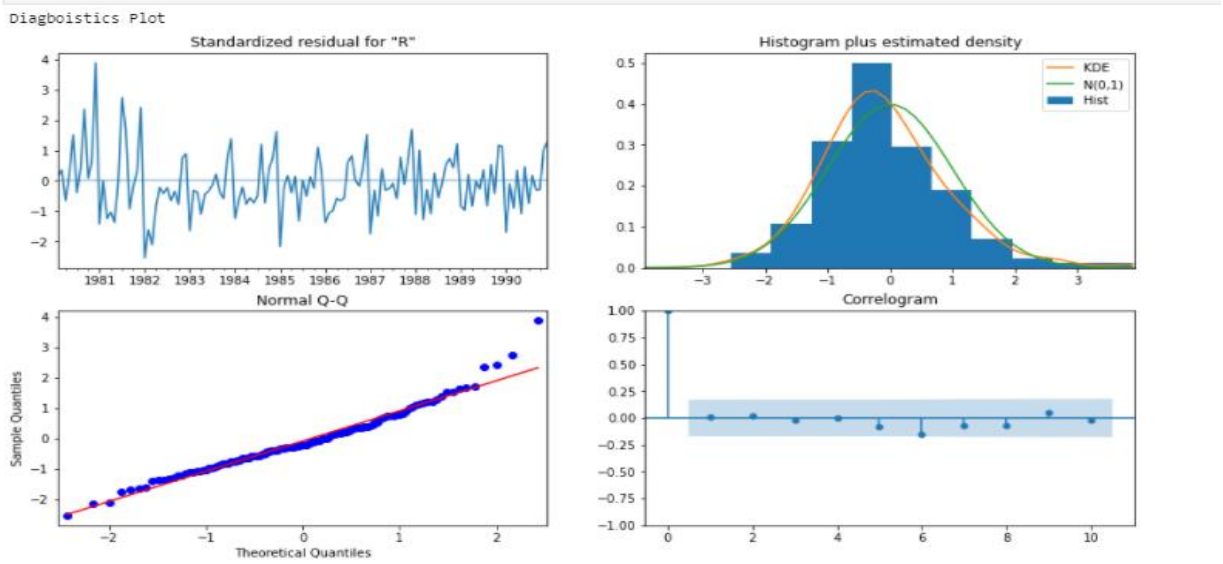
Model with parameter (2,1,3) has the least AIC score. So, it is the best parameter.

4.Summary of the model after fitting it with the best parameters

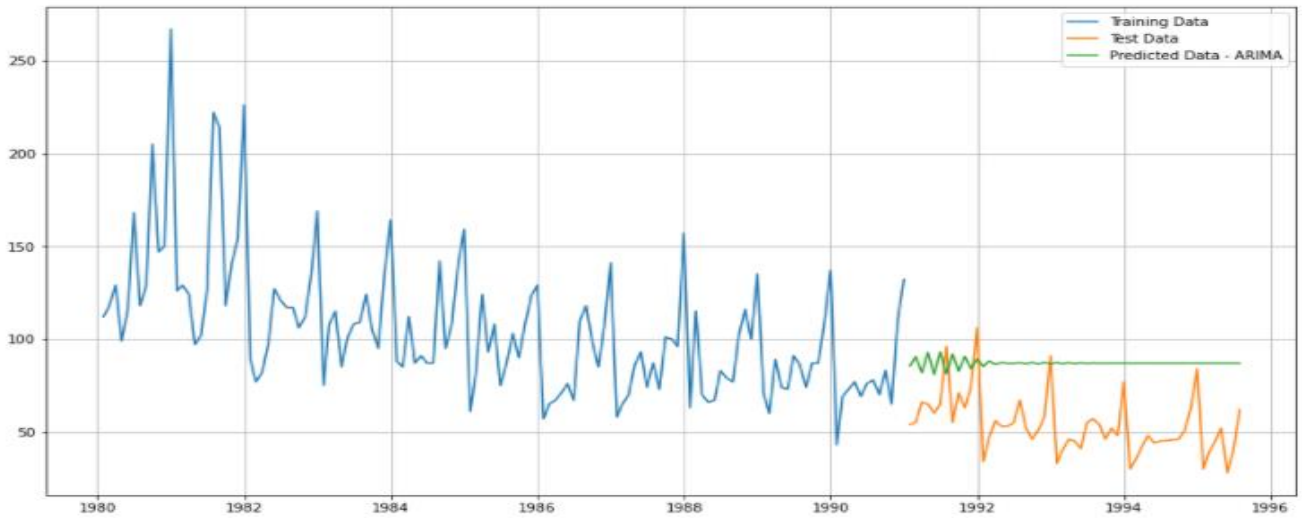
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 23 May 2021   AIC                           1274.695
Time:                        16:56:22   BIC                           1291.946
Sample:                    01-31-1980   HQIC                          1281.705
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6780      0.084    -20.002      0.000      -1.842      -1.514
ar.L2         -0.7288      0.084     -8.687      0.000      -0.893      -0.564
ma.L1          1.0447      0.638      1.638      0.102      -0.206       2.295
ma.L2         -0.7719      0.133     -5.790      0.000      -1.033      -0.511
ma.L3         -0.9046      0.579     -1.564      0.118      -2.039       0.229
sigma2       860.3442    537.587      1.600      0.110    -193.307    1913.995
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.46
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).
```

5.Diagnostics Plot



6.ARIMA predication Plot on test data



7.RMSE score on test data for Arima using lowest Akaike Information Criteria

```
RMSE: 36.81375470121735
```

b.) SARIMA model

1.Creating different parameters for the model.

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary.

As we are not applying any additional differentiate for seasonal data D value stays as zero

2.Building Sarima model with different parameters

In the below images we have shown only the few parameter combinations

```
SARIMA(0, 1, 0)x(0, 0, 0, 6) - AIC:1323.9657875279158
SARIMA(0, 1, 0)x(0, 0, 1, 6) - AIC:1264.4996261113863
SARIMA(0, 1, 0)x(0, 0, 2, 6) - AIC:1144.7077471827379
SARIMA(0, 1, 0)x(0, 0, 3, 6) - AIC:1081.2713830625291
SARIMA(0, 1, 0)x(1, 0, 0, 6) - AIC:1274.7897737087985
SARIMA(0, 1, 0)x(1, 0, 1, 6) - AIC:1241.7870945149107
SARIMA(0, 1, 0)x(1, 0, 2, 6) - AIC:1146.3093266721787
SARIMA(0, 1, 0)x(1, 0, 3, 6) - AIC:1058.98617431244
SARIMA(0, 1, 0)x(2, 0, 0, 6) - AIC:1137.9167236212038
SARIMA(0, 1, 0)x(2, 0, 1, 6) - AIC:1137.4533629515267
SARIMA(0, 1, 0)x(2, 0, 2, 6) - AIC:1117.0224426127638
SARIMA(0, 1, 0)x(2, 0, 3, 6) - AIC:1058.804820642388
SARIMA(0, 1, 0)x(3, 0, 0, 6) - AIC:1072.5465834695267
SARIMA(0, 1, 0)x(3, 0, 1, 6) - AIC:1061.3687765139675
SARIMA(0, 1, 0)x(3, 0, 2, 6) - AIC:1058.0425053414538
SARIMA(0, 1, 0)x(3, 0, 3, 6) - AIC:1058.8803339857695
SARIMA(0, 1, 1)x(0, 0, 0, 6) - AIC:1263.5369097383968
SARIMA(0, 1, 1)x(0, 0, 1, 6) - AIC:1201.383254802955
SARIMA(0, 1, 1)x(0, 0, 2, 6) - AIC:1097.190821775279
```

3.Head of the Sarima models with AIC score in ascending order

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	951.744297
59	(0, 1, 3)	(2, 0, 3, 6)	952.073632
251	(3, 1, 3)	(2, 0, 3, 6)	952.582110
191	(2, 1, 3)	(3, 0, 3, 6)	953.205627
123	(1, 1, 3)	(2, 0, 3, 6)	953.684953

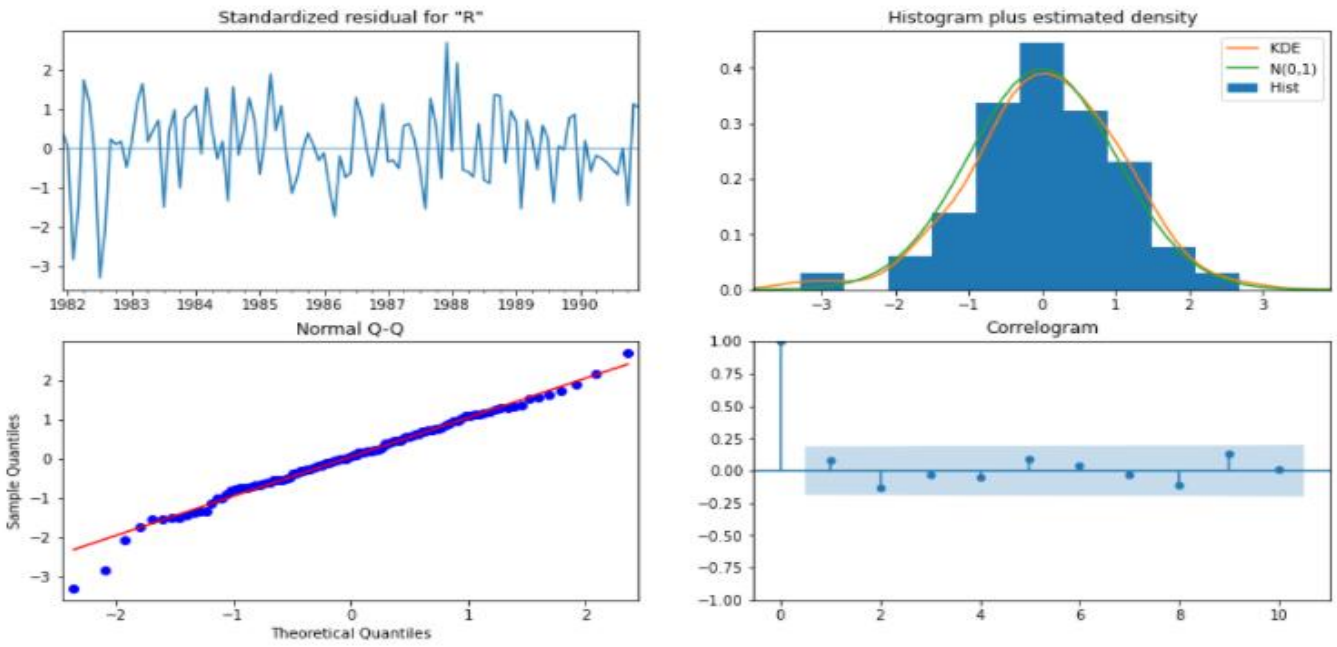
Model with parameter (2,1,3) (2,0,3,6) has the least AIC score. So, it is the best parameter.

4.Summary of the model after fitting it with the best parameters

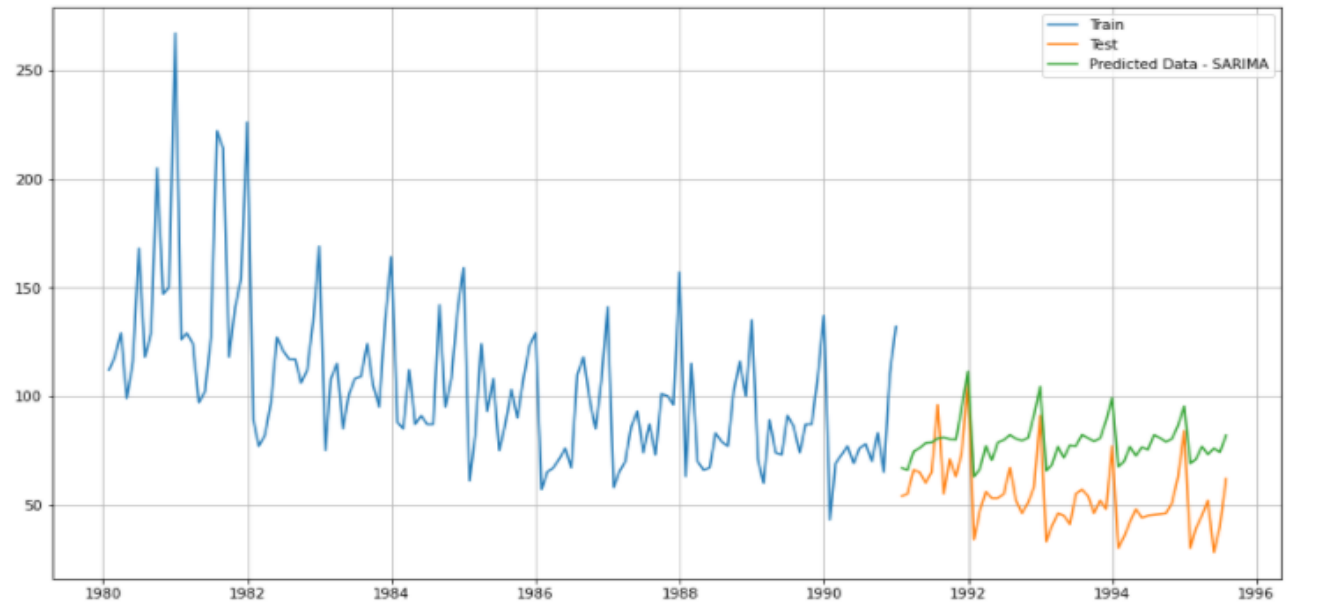
```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                              Rose   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -464.872
Date:                           Sun, 23 May 2021   AIC                             951.744
Time:                                   17:32:20   BIC                             981.349
Sample:                                 01-31-1980   HQIC                          963.750
                                      - 12-31-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5027      0.083     -6.081      0.000      -0.665      -0.341
ar.L2         -0.6628      0.084     -7.918      0.000      -0.827      -0.499
ma.L1         -0.3714    192.592     -0.002      0.998    -377.846     377.103
ma.L2          0.2033    121.054      0.002      0.999    -237.058     237.464
ma.L3         -0.8320    160.186     -0.005      0.996    -314.791     313.127
ar.S.L6       -0.0838      0.049     -1.720      0.085      -0.179       0.012
ar.S.L12       0.8099      0.052     15.464      0.000       0.707       0.913
ma.S.L6        0.1702      0.248      0.687      0.492      -0.316       0.656
ma.S.L12      -0.5646      0.199     -2.835      0.005      -0.955      -0.174
ma.S.L18       0.1710      0.143      1.198      0.231      -0.109       0.451
sigma2       260.8021   5.02e+04      0.005      0.996   -9.82e+04    9.87e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):                 4.77
Prob(Q):                              0.40   Prob(JB):                         0.09
Heteroskedasticity (H):               0.54   Skew:                            -0.36
Prob(H) (two-sided):                  0.06   Kurtosis:                         3.73
===================================================================================

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

5.Diagnostics Plot



6.SARIMA predication Plot on test data

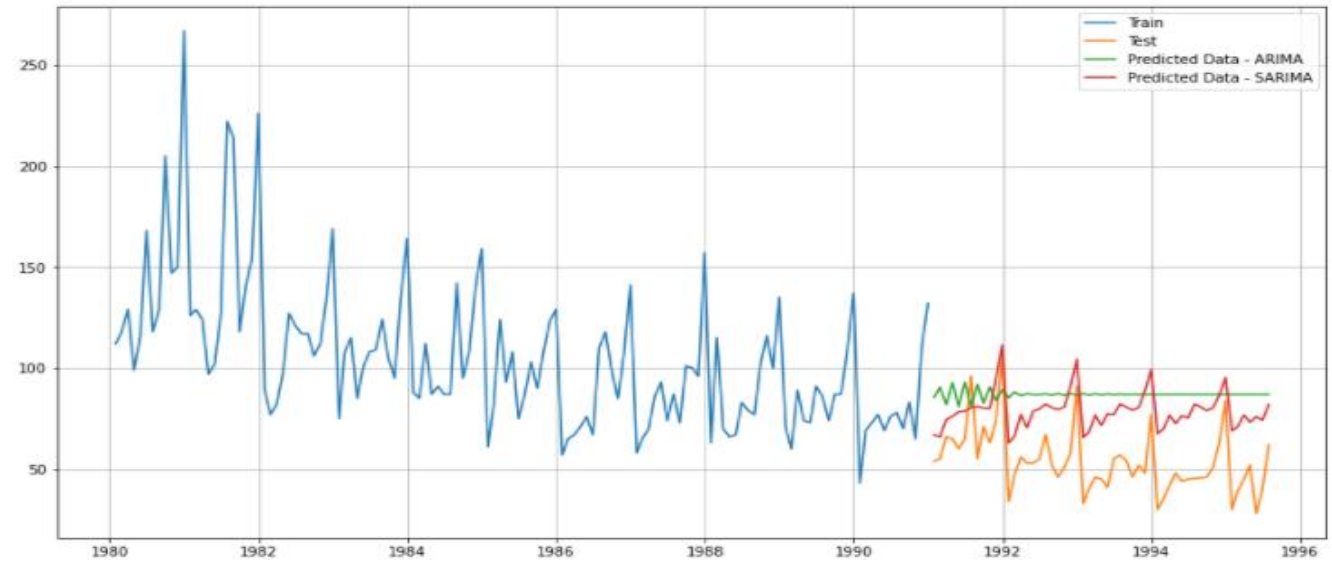


7.RMSE score of Sarima using lowest Akaike Information Criteria

```
RMSE: 27.124404049535265
```

Inference:

	RMSE
ARIMA(2,1,3)	36.813755
SARIMA(2,1,3)(2,0,3,6)	27.124404

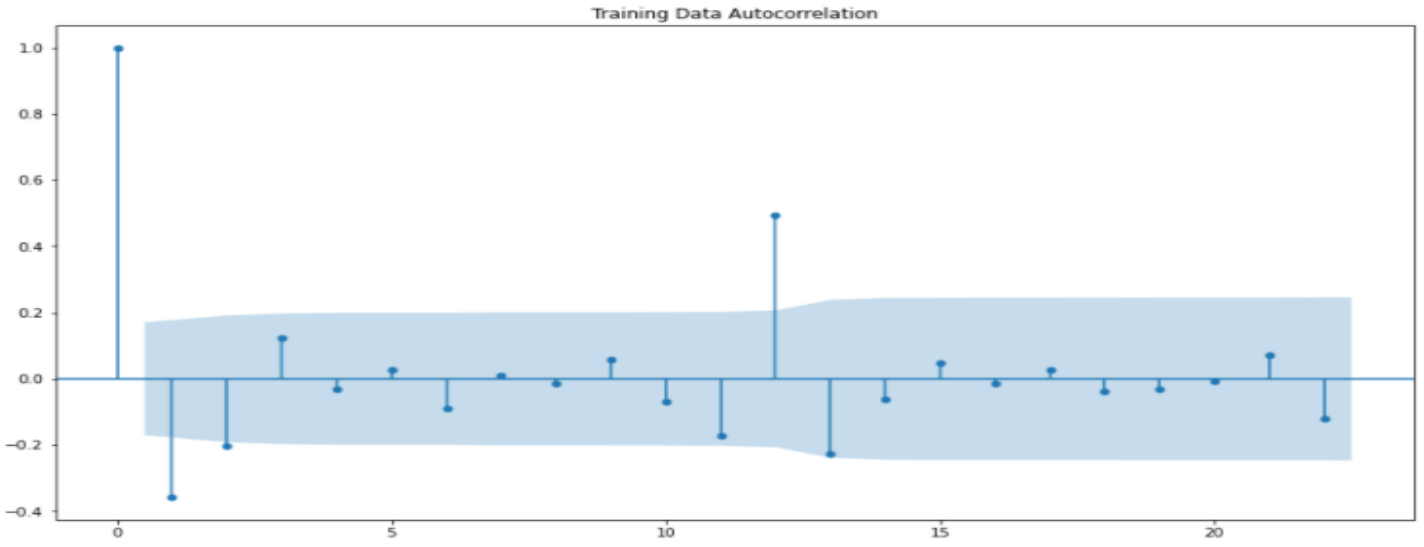


RMSE has reduced in comparison to ARIMA when seasonality was introduced.

2.7) Building ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluating this model on the test data using RMSE.

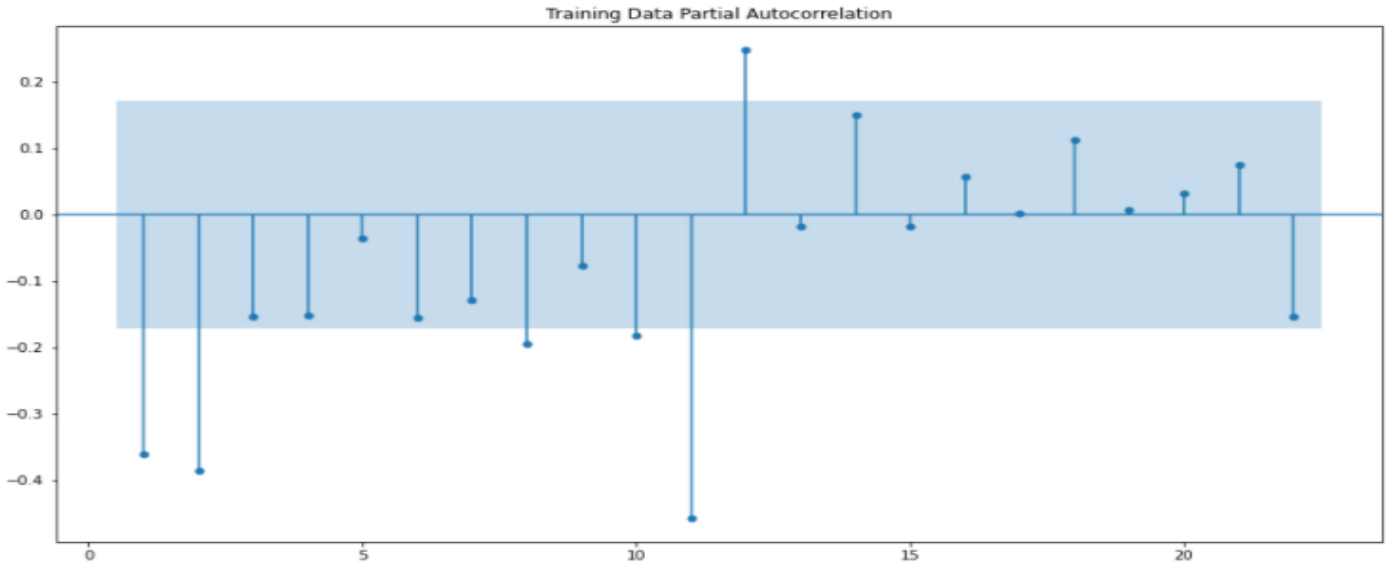
a.) ARIMA model method using cut-off points of ACF and PACF.

1.ACF plot of train data.



Based on acf plot q value is 2 . (ie 2nd lag is out of confidence level and next lag drops below confidence level)

2.PACF plot of train data



Based on PACF plot p value is 2. (ie 2nd lag is out of confidence level and next lag drops below confidence level)

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary.

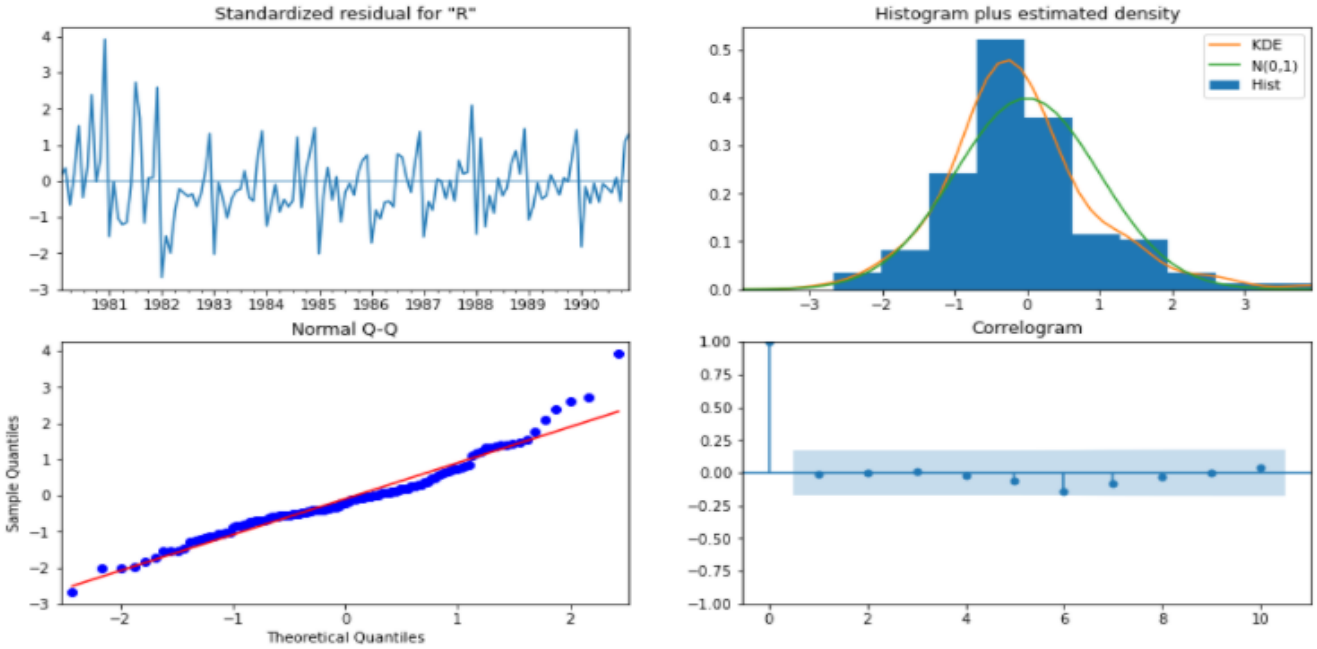
So, the parameter is (2,1,2)

3.Summary of the model after fitting it with the best parameters (2,1,2):

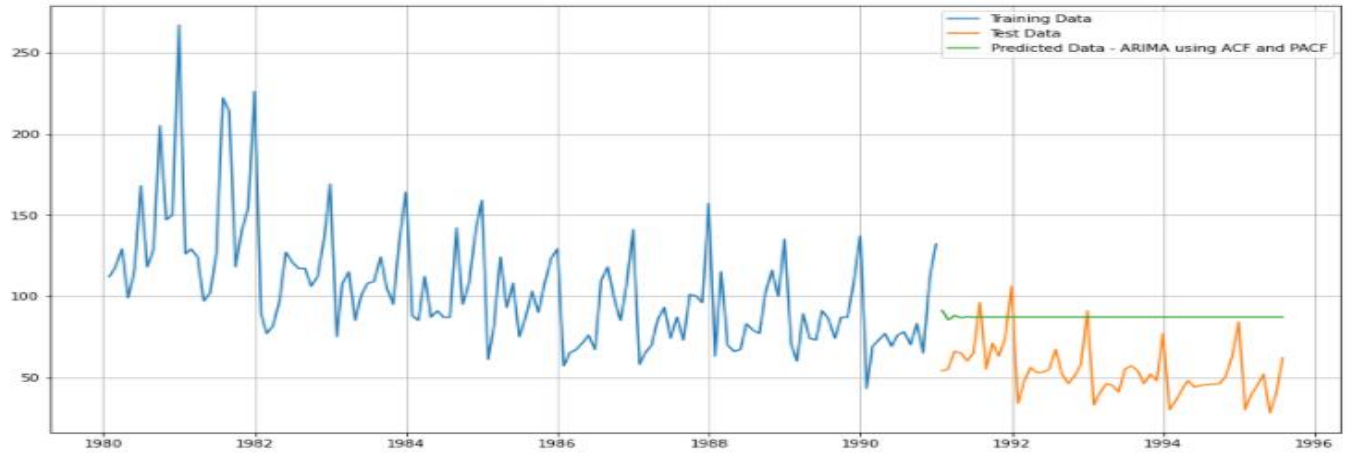
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Sun, 23 May 2021   AIC                           1281.871
Time:                        17:51:48   BIC                           1296.247
Sample:                    01-31-1980   HQIC                          1287.712
                         - 12-31-1990
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4540      0.469     -0.969      0.333      -1.372       0.464
ar.L2          0.0001      0.170      0.001      0.999      -0.334       0.334
ma.L1         -0.2541      0.459     -0.554      0.580      -1.154       0.646
ma.L2         -0.5984      0.430     -1.390      0.164      -1.442       0.245
sigma2       952.1601     91.424     10.415      0.000     772.973    1131.347
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                34.16
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.94
===================================================================================

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

4.Diagnostics Plot



5.ARIMA predication Plot on test data

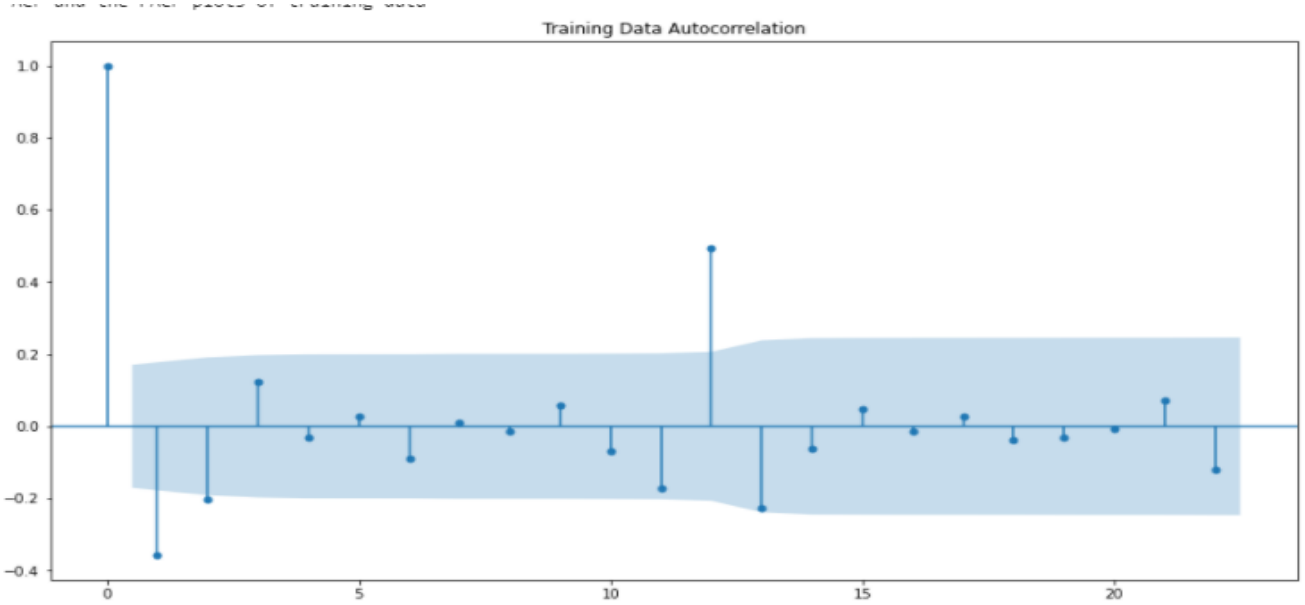


6.RMSE score on test data for Arima using ACF AND PACF (Manual)

```
RMSE: 36.87119661682952
```

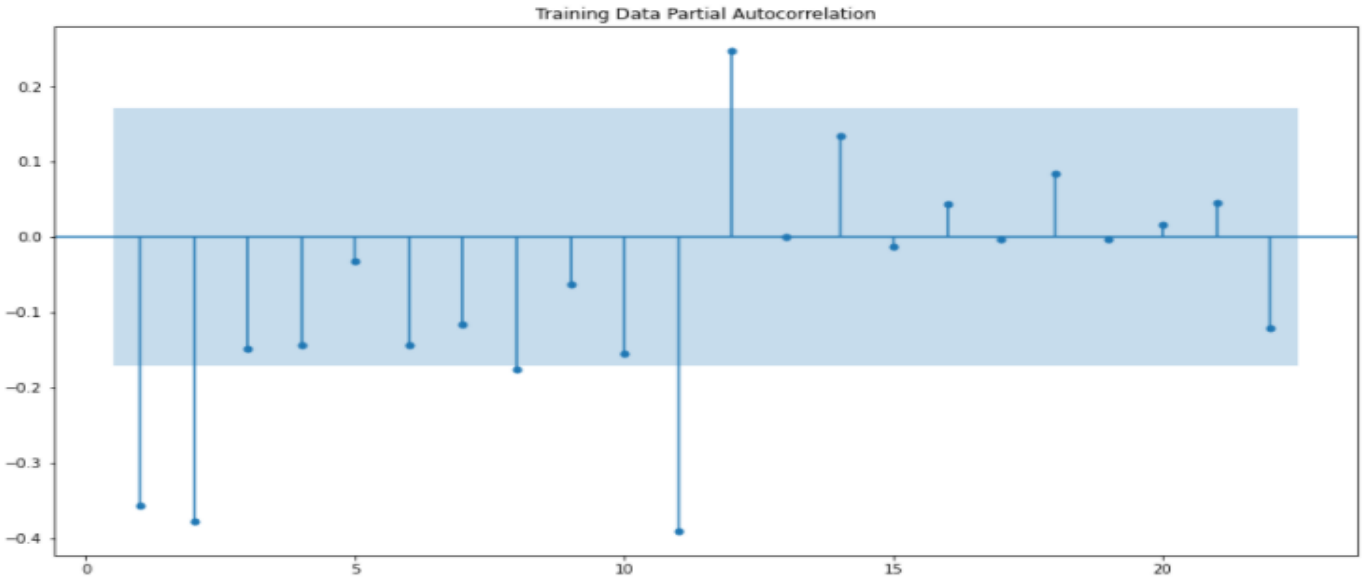
b.) SARIMA model cut-off points of ACF and PACF

1.ACF plot of train data.



Based on ACF plot q value is 2. (ie 2nd lag is out of confidence level and next lag drops below confidence level) Q value is 3 After 3 lag there is huge drop .

2.PACF plot of train data



Based on PACF plot p value is 2. (i.e. 2nd lag is out of confidence level and next lag drops below confidence level) and P value is 0 cos there is a no significant seasonality trend

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary.

We do not differentiate again D is 0.

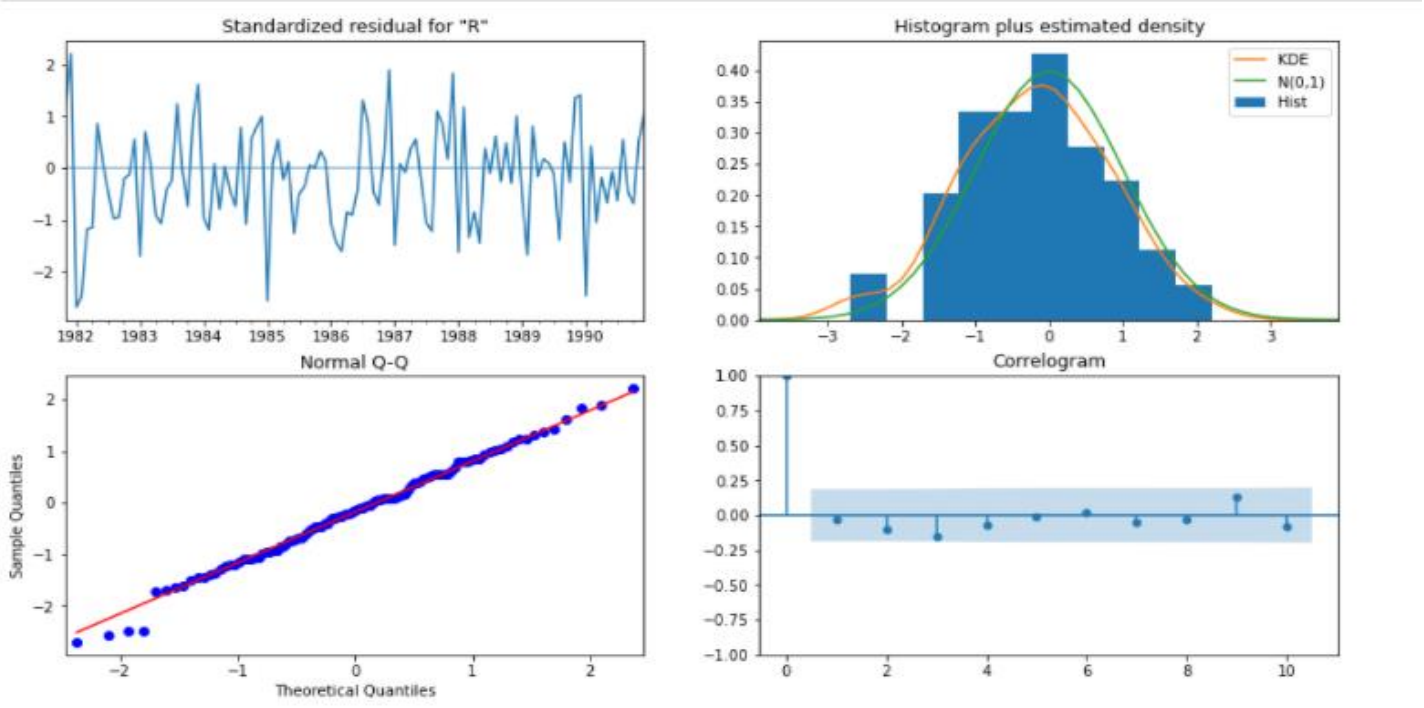
So, the parameter is (2,1,2) (0, 0, 3, 6)

3.Summary of the model after fitting it with the best parameters (2,1,2) (0, 0, 3, 6)

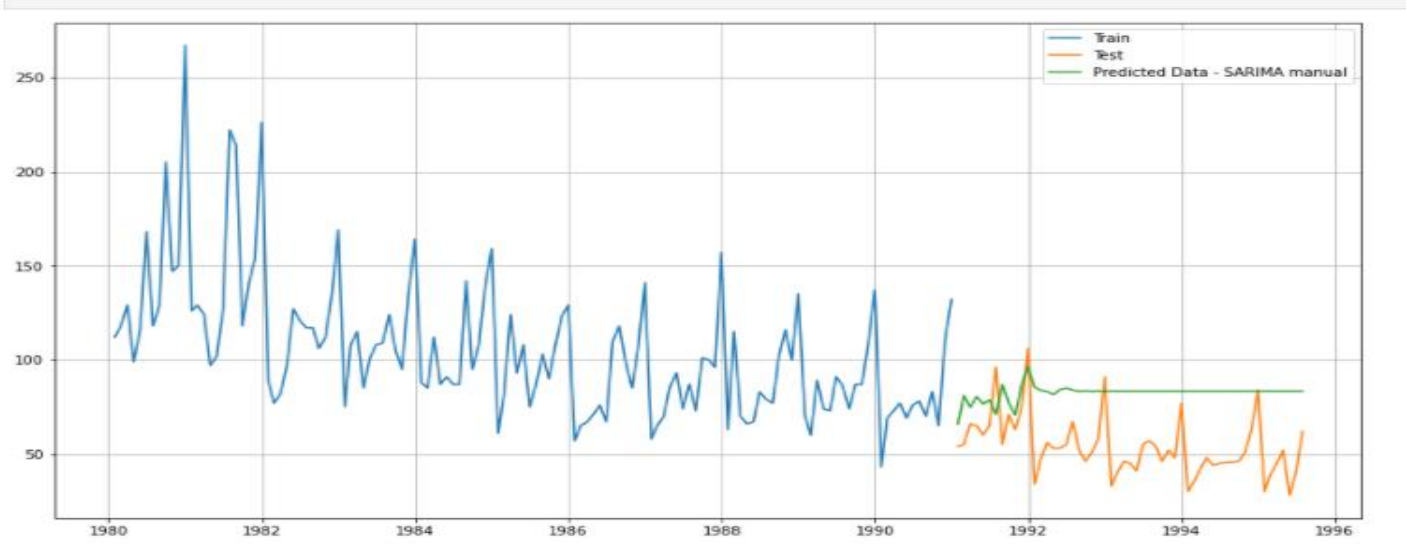
```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(0, 0, [1, 2, 3], 6)   Log Likelihood                -494.745
Date:                                   Sun, 23 May 2021   AIC                           1005.489
Time:                                           18:01:33   BIC                           1027.093
Sample:                                     01-31-1980   HQIC                          1014.252
                                          - 12-31-1990
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5896      0.081     -7.267      0.000      -0.749      -0.431
ar.L2          0.1404      0.088      1.597      0.110      -0.032       0.313
ma.L1         -0.1155      0.101     -1.138      0.255      -0.314       0.083
ma.L2         -0.7726      0.102     -7.600      0.000      -0.972      -0.573
ma.S.L6       -0.1745      0.131     -1.328      0.184      -0.432       0.083
ma.S.L12       0.6238      0.121      5.140      0.000       0.386       0.862
ma.S.L18       0.0814      0.157      0.520      0.603      -0.226       0.388
sigma2       441.1057     66.974      6.586      0.000     309.839     572.372
===================================================================================
Ljung-Box (L1) (Q):                   0.13   Jarque-Bera (JB):                 0.38
Prob(Q):                              0.72   Prob(JB):                         0.83
Heteroskedasticity (H):               0.80   Skew:                            -0.13
Prob(H) (two-sided):                  0.49   Kurtosis:                         2.89
===================================================================================

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

4.Diagnostics Plot



5.SARIMA predication Plot on test data (manual)



6.RMSE score of Sarima using ACF AND PACF cutoff (Manual)

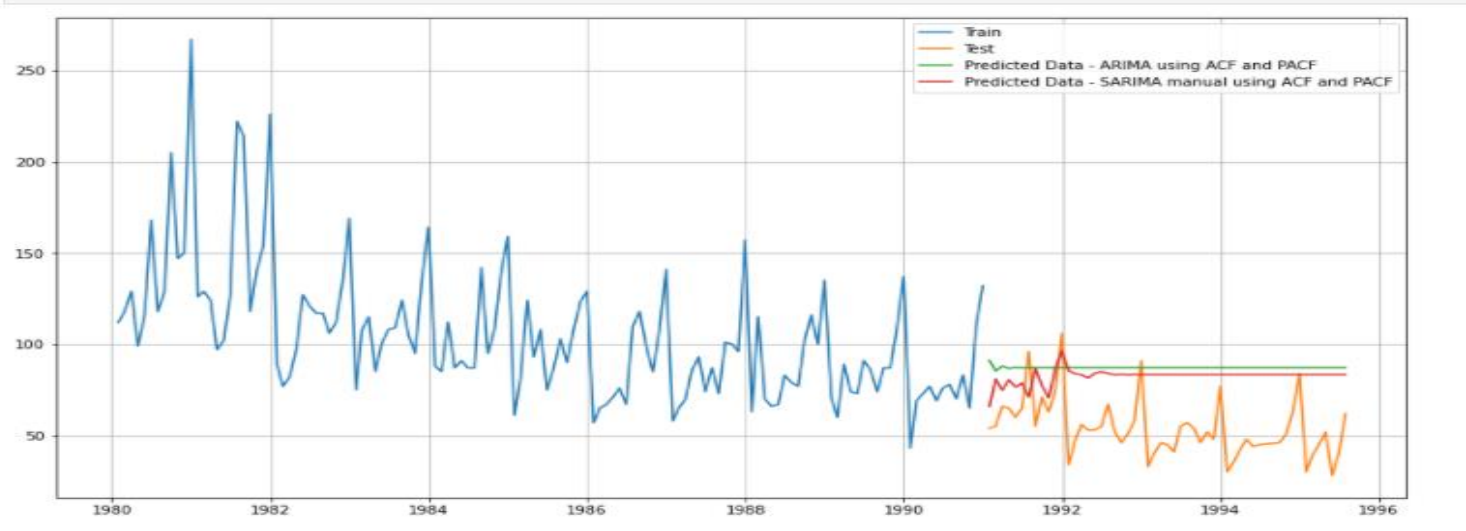
```
RMSE: 33.05392044028609
```

Inference:

1. Comparison Table of RMSE OF Different model

	RMSE
ARIMA M(2,1,2)	36.871197
SARIMA M(2,1,2)(0,0,3,6)	33.053920

2. Comparison graph of Different model



3. Insights

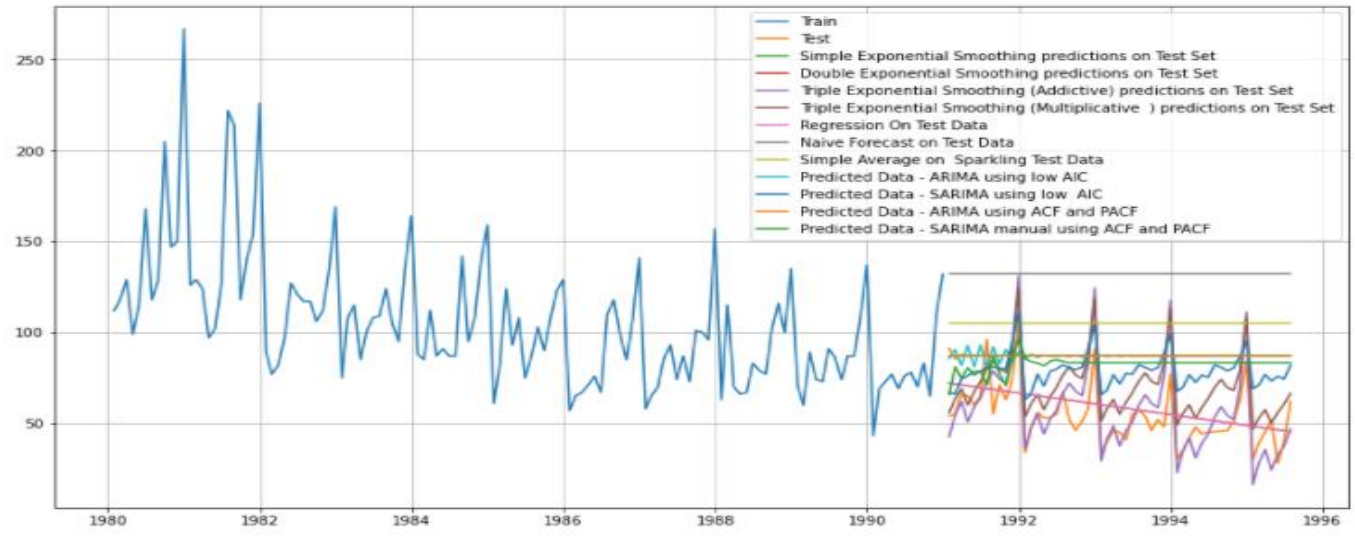
RMSE has reduced in comparison to ARIMA when seasonality was introduced.

2.8) Building a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Model Performance comparison table

	Test RMSE	MAPE
SES (Alpha=0.0987)	36.796225	75.909183
DES (Alpha=1.4901e-08,Beta=5.4481e-09)	15.270968	30.073896
TES (Alpha=0.0849,Beta=5.52e-06,Gamma=0.0005) : Addictive	14.243240	42.850275
TES(Alpha=0.0773,Beta=0.0393,Gamma=0.0008): Multiplicative	19.113110	47.756101
Regression	15.268955	30.067644
NaiveModel	79.718773	164.846275
SimpleAverage	53.460570	110.587957
Arima (2,1,3) : Low AIC	36.813755	75.840581
Sarima (2, 1, 3)(2, 0, 3,6), :Low AIC	27.124404	55.239519
Arima (2,1,2) : cut-off points of ACF and PACF	36.871197	76.056213
Sarima (2, 1, 2)(0, 0, 3,6), :cut-off points of ACF and PACF	33.053920	67.354994

Forecasting comparison plot.



Inference:

Based on the above table, we can see that Triple exponential Smoothing (Alpha=0.0849,Beta=5.52e-06,Gamma=0.0005) : Addictive has the least RMSE . So, we will choose that as optimized model.

2.9) Based on the model-building exercise, building the most optimum model on the complete data and predicting 12 months into the future with appropriate confidence intervals/bands.

Best model is TES (Alpha=0.0849,Beta=5.52e-06,Gamma=0.0005) : Addictive (Based on RMSE)

Summary of optimized model on complete data

ExponentialSmoothing Model Results

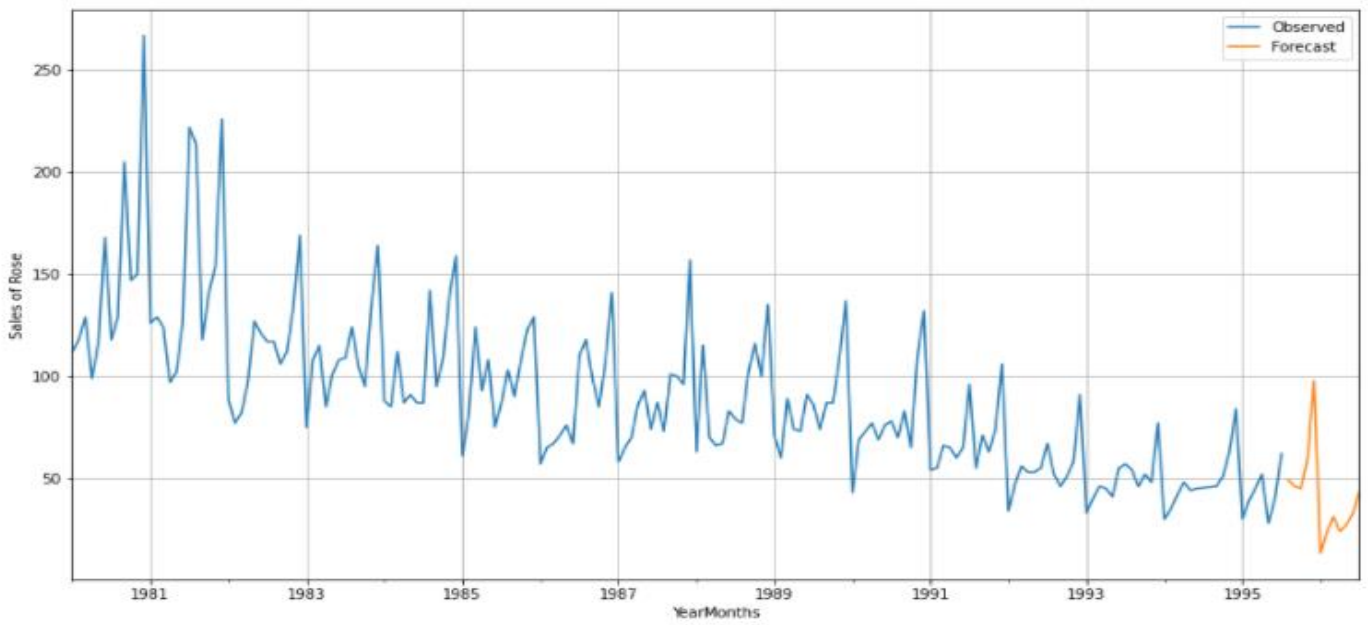
Dep. Variable:	Rose	**No. Observations:**	187
Model:	ExponentialSmoothing	**SSE**	58378.195
Optimized:	True	**AIC**	1106.051
Trend:	Additive	**BIC**	1157.749
Seasonal:	Additive	**AICC**	1110.123
Seasonal Periods:	12	**Date:**	Sun, 23 May 2021
Box-Cox:	False	**Time:**	18:21:01
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.0849157	alpha	False
smoothing_trend	5.5205e-06	beta	False
smoothing_seasonal	0.0005477	gamma	False
initial_level	145.99879	l.0	True
initial_trend	-0.5425658	b.0	True
initial_seasons.0	-28.039827	s.0	True
initial_seasons.1	-17.186813	s.1	True
initial_seasons.2	-9.0838598	s.2	True
initial_seasons.3	-15.732811	s.3	True
initial_seasons.4	-11.818749	s.4	True
initial_seasons.5	-5.7767801	s.5	True
initial_seasons.6	5.4113147	s.6	True
initial_seasons.7	5.3084715	s.7	True
initial_seasons.8	2.6761590	s.8	True
initial_seasons.9	1.9527664	s.9	True
initial_seasons.10	17.095964	s.10	True
initial_seasons.11	55.909796	s.11	True

Forecasted value

```
1995-08-31    49.304696
1995-09-30    46.129961
1995-10-31    44.864079
1995-11-30    59.464640
1995-12-31    97.735616
1996-01-31    13.243890
1996-02-29    23.554355
1996-03-31    31.114713
1996-04-30    23.923410
1996-05-31    27.294667
1996-06-30    32.794011
1996-07-31    43.432328
Freq: M, dtype: float64
```

Plot the forecast (mean value) of the whole data.



Forecasted value description.

```
count    12.000000
mean     41.071364
std      22.146617
min      13.243890
25%      26.451853
50%      38.113170
75%      46.923645
max      97.735616
dtype: float64
```

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

Analytical Insights:

- I have formatted the data into time series data by creating time stamp instead of YearMonth date field. It is a monthly data from January 1980 to July 1995.
- Did exploratory analysis and found out that there is two missing data.It was handled using Linear Interpolation technique .
- More than 75 percent of data has sales count of 112 and below with average sales count being 90.
- performed decomposition to understand that there is a decreasing uniform trend.
- There is a seasonality in the data. Peak sales are in December and least sales are in April . this suggest that the sales of Sparkling wine follow a festive seasonality.

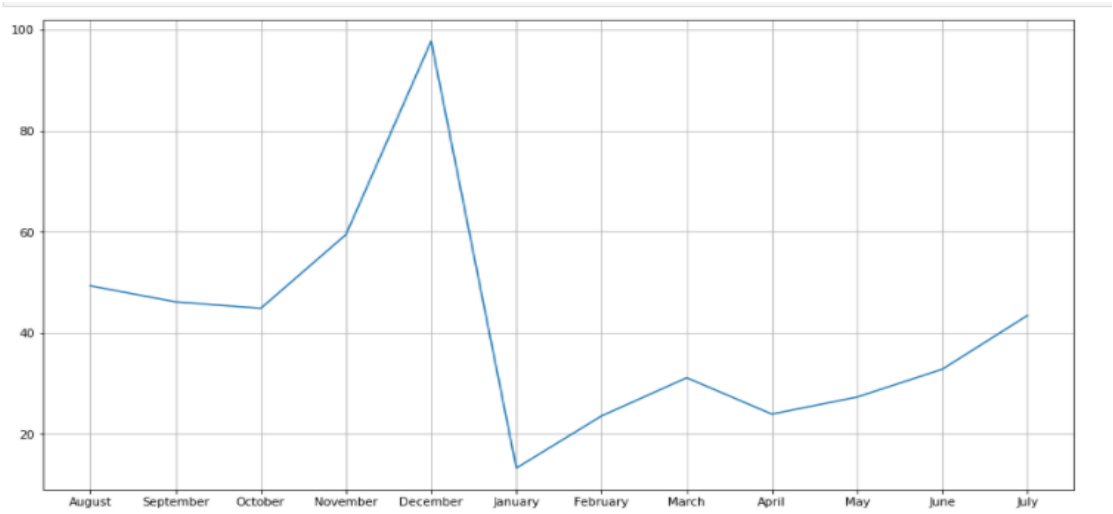
- Have split the data into train and test. Train data is from January 1980 December 1990. Test data is from Jan 1991 till July 1995 (end of the data set).

- Using Stationary test, we have found that the dataset is not stationary. After one differencing It becomes stationary.

- Have analyzed the data using various time series models and found out Triple exponential Smoothing (Alpha=0.0849, Beta=5.52e-06, Gamma=0.0005) : Addictive is the most optimized model with least RMSE score.

- Have predicted the data for next 12 months using the Triple exponential Smoothing (Alpha=0.0849, Beta=5.52e-06, Gamma=0.0005) : Addictive model (The most optimized model). The forecasted average sales will be in December 1995 with the count of 97. The average sales count will be 41 for each month.

- Forecasting for 12 months (Month wise sales count:



Business Recommendations:

- I suggest the "ABC Estate Wines "to have average stock up a minimum of 45 rose wine every month.

- I suggest them to increase the stock by 10 percent every month for the time July to December and to maintain minimum stock count of 50 (45 (average wine sales) plus 5 (for prediction error)) from January till June

- I recommend them to have a minimum stock of 150 rose wine during December.

- As the rose wine sales has festive seasonality (the sales are maximum in December so it may be due to Christmas and new year) I suggest them to spend more on marketing during the month of October to December. It may help them to increase sales.

- As the sales are least during January till July, I suggest them to give discount or voucher during these months. It may increase the sales.

Thank you.