TSF PROJECT

RoseDataset

DSBA

BUSINESS REPORT

Sayyed Abdul Khaliq

Email : abdulkhaliq01112001@gmail.com

CONTENTS

List of Figures

List of Tables

Problem 1

Problem statement:

As an analyst at ABC Estate Wines, we have access to historical data on the sales of various types of wines throughout the 20th century. These datasets, though originating from the same company, reflect sales figures for different wine varieties. Our goal is to explore this data to identify trends, patterns, and factors that have influenced wine sales over the century. By applying data analytics and forecasting techniques, we aim to extract actionable insights that will guide strategic decision-making and help optimize sales strategies for the future.

Context

Given the historical sales data of various wine types from ABC Estate Wines spanning the 20th century, we need to conduct a comprehensive analysis to uncover trends and patterns. Additionally, we aim to build robust forecasting models to predict future wine sales. This involves:

1. Data Exploration and Preprocessing: Cleaning and organizing the historical sales data to ensure it is suitable for analysis and modelling.
2. Trend Analysis: Identifying long-term trends, seasonal patterns, and other significant factors that have influenced wine sales over the century.
3. Forecasting: Developing predictive models to forecast future sales of different wine varieties, enabling ABC Estate Wines to make informed strategic decisions.
4. Actionable Insights: Providing recommendations based on the analysis and forecasts to help optimize sales strategies and improve market positioning.

By achieving these objectives, we will support ABC Estate Wines in understanding past sales dynamics and preparing effectively for future market conditions.

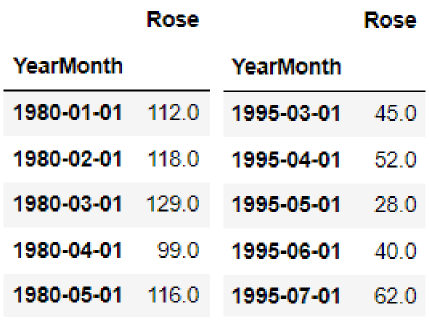
Data Description

Column name	Details
YearMonth	Dates of sales
Rose	Sales of Rose wine

Data Overview

Read the data as an appropriate time series data

Data is loaded into dataframe using pandas library and first 5 and last 5 rows were printed.



	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

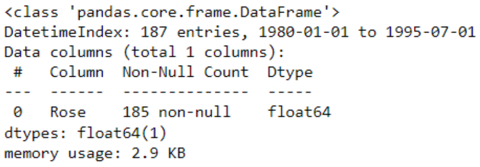
	Rose
YearMonth	
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table 1 –Dataset – Rows of Data

Check the structure of the data

- Data has 187 rows and 1 column

Check the Datatypes:

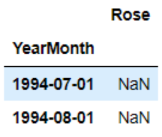


```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Table 2 –Dataset– Info

Check for and treat (if needed) missing values –

There are 2 null values in the dataset. We found the values for the months of July & August were missing for the year 1994.



	Rose
YearMonth	
1994-07-01	NaN
1994-08-01	NaN

Table 3 –Null values

We tried below approach to impute the data, these were as below.

Mean - Before & After

Treating null values is very important to do further analysis.

In this approach, instead of taking means for the 7th months across all the years, we just took mean of the 7th months values from a year before and a year after the missing value.

Similarly, instead of taking means for the 8th months across all the years, we just took mean of the 8th months values from a year before and a year after the missing value.

Plot the data:

After treating null values, final data looks like below:

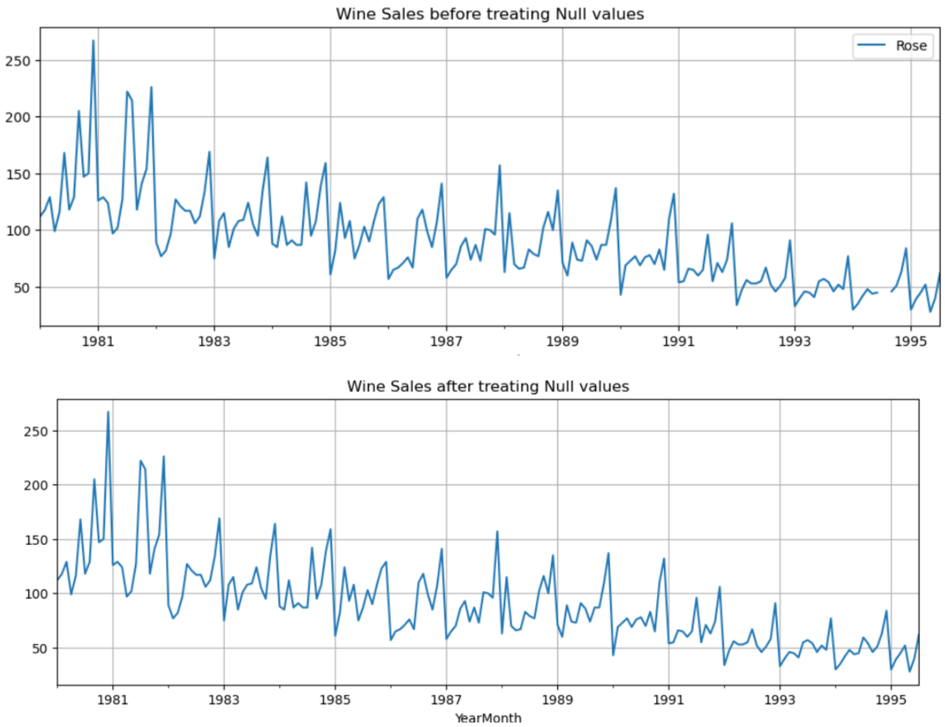


Fig 1– Time series before and after treating null values

Perform EDA :

We have divided the dataset further by extraction month and year columns from the YearMonth column for better analysis of the dataset. The new dataset has 187 rows and 3 columns.

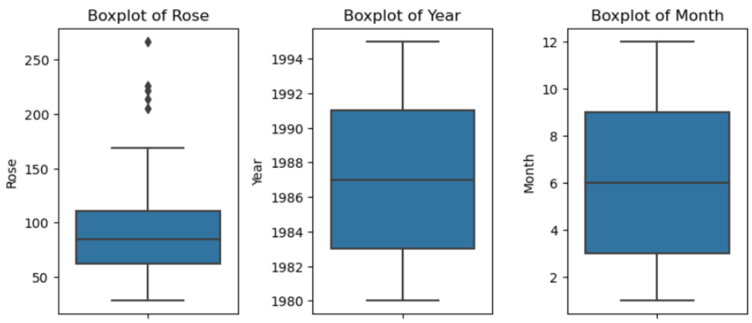


Fig 2 :Box plot of data

The box plot shows:

Boxplot has outliers we can treat them but we are choosing not to treat them as they do not give much effect on the time series model.

Boxplot Yearly:

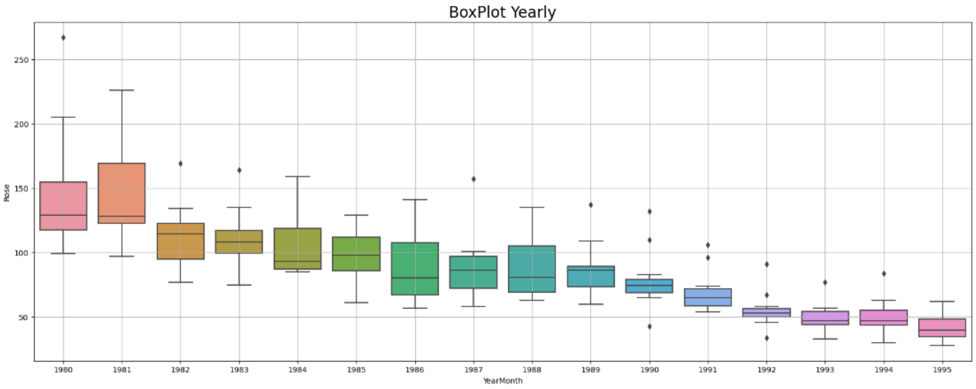


Fig 3 :Boxplot Yearly

This yearly box plot shows there is gradual decrease in sales over the years. Outliers are present in few years.

Boxplot Monthly:

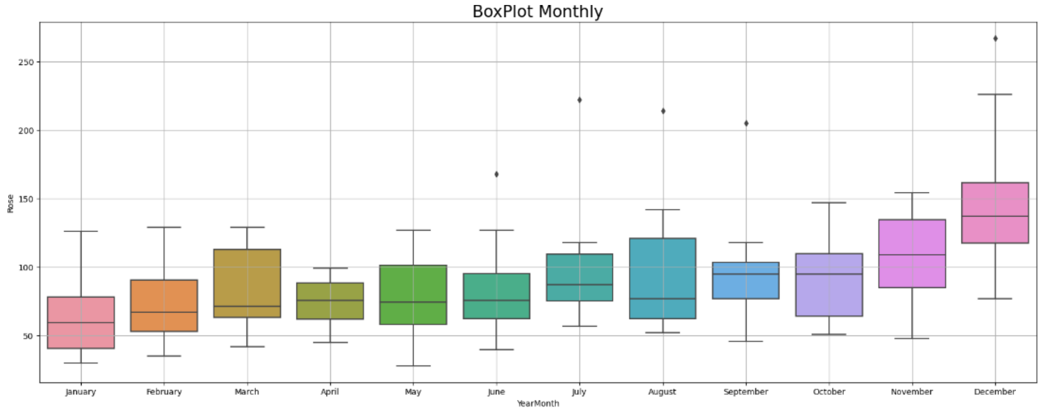


Fig 4 :Boxplot Monthly

The plot shows that sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to Sept then from October the sales start to increase due to holiday season. Outliers are present in June to September and Decemeber.

Monthly Sales over the years:

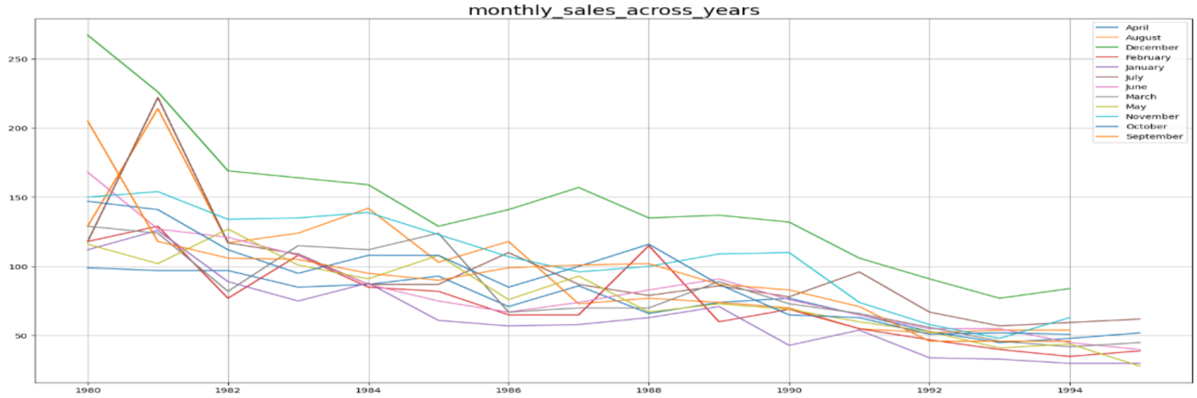


Fig 5: Monthly Sales over the years

This plot shows that December has the highest sales over the years and the year 1980 was the year with the highest number of sales and then sales started to decrease.

Decomposition -Additive

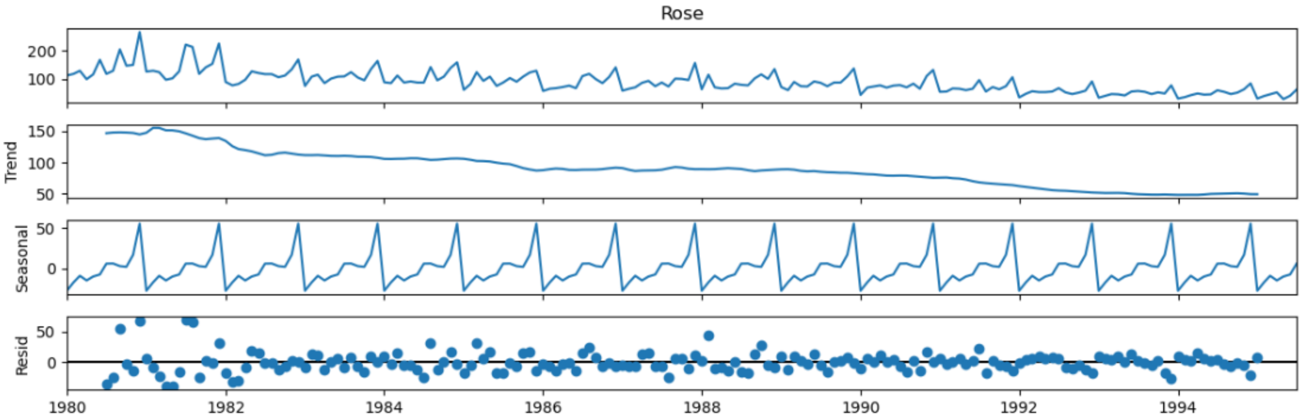


Fig 6: Decomposition -Additive

The plots show:

- Peak year 1981
- It also shows that the trend has declined over the year after 1981
- Residue is spread and is not in a straight line.
- Both trend and seasonality are present.

Decomposition-Multiplicative

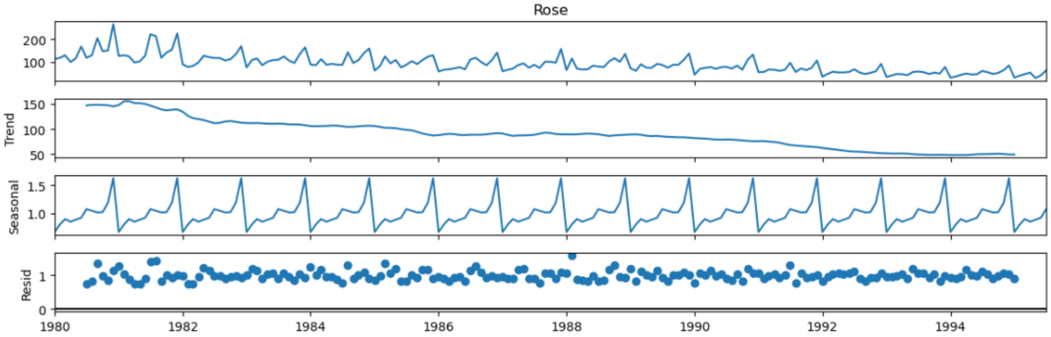


Fig 7: Decomposition – Multiplicative

The plots show

- Peak year 1981
- It also shows that the trend has declined over the year after 1981.
- Residue is spread and is in approx a straight line.
- Both trend and seasonality are present.
- So multiplicative model is selected owing to a more stable residual plot and lower range of residuals.

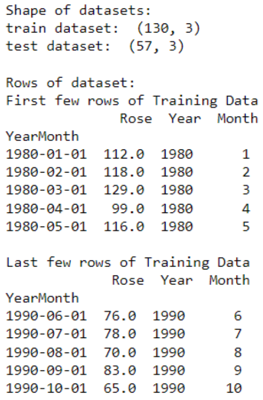
Train-test split:

In time series forecasting, splitting the data into training and testing sets is crucial for evaluating model performance. Unlike random splits in typical machine learning tasks, time series data require careful handling due to the temporal dependencies.

1. **Chronologically Split Data**: Ensure the training set precedes the test set.
2. **Typical Ratios**: Use 70-80% of the data for training, and 20-30% for testing.
3. **Prevent Data Leakage**: Always ensure the model only sees past data during training.

This approach helps in accurately evaluating the model's forecasting ability and ensures realistic performance metrics.

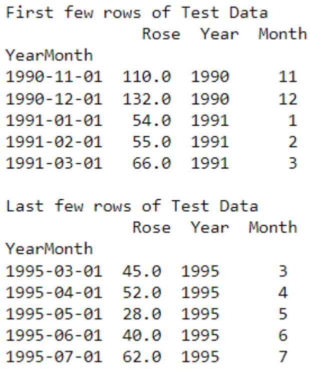
Details of train and test data are as follows:



```
Shape of datasets:
train dataset:  (130, 3)
test dataset:  (57, 3)

Rows of dataset:
First few rows of Training Data
              Rose  Year  Month
YearMonth
1980-01-01  112.0  1980      1
1980-02-01  118.0  1980      2
1980-03-01  129.0  1980      3
1980-04-01   99.0  1980      4
1980-05-01  116.0  1980      5

Last few rows of Training Data
             Rose  Year  Month
YearMonth
1990-06-01  76.0  1990      6
1990-07-01  78.0  1990      7
1990-08-01  70.0  1990      8
1990-09-01  83.0  1990      9
1990-10-01  65.0  1990     10
```



```
First few rows of Test Data
              Rose  Year  Month
YearMonth
1990-11-01  110.0  1990     11
1990-12-01  132.0  1990     12
1991-01-01   54.0  1991      1
1991-02-01   55.0  1991      2
1991-03-01   66.0  1991      3

Last few rows of Test Data
             Rose  Year  Month
YearMonth
1995-03-01  45.0  1995      3
1995-04-01  52.0  1995      4
1995-05-01  28.0  1995      5
1995-06-01  40.0  1995      6
1995-07-01  62.0  1995      7
```

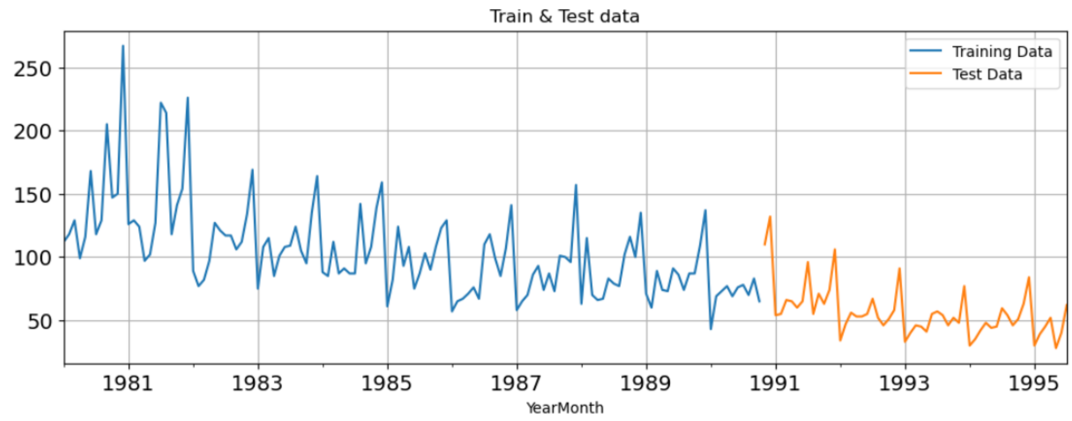


Fig 8 –Train & Test Plot

Model building:

Build forecasting models:

Evaluation Metrics:

When evaluating time series forecasting models, it's important to use metrics that capture both the accuracy of the predictions and the goodness of fit of the model. Two commonly used metrics are Root Mean Squared Error (RMSE) and Akaike Information Criterion (AIC).

1**. Root Mean Squared Error (RMSE)**

RMSE measures the average magnitude of the errors between the predicted and actual values. It is the square root of the average of squared differences between prediction and actual observation.

2. Akaike Information Criterion (AIC)

AIC is a measure of the relative quality of a statistical model for a given set of data. It balances the model's goodness of fit with its complexity, penalizing models with more parameters to avoid overfitting.

Linear regression

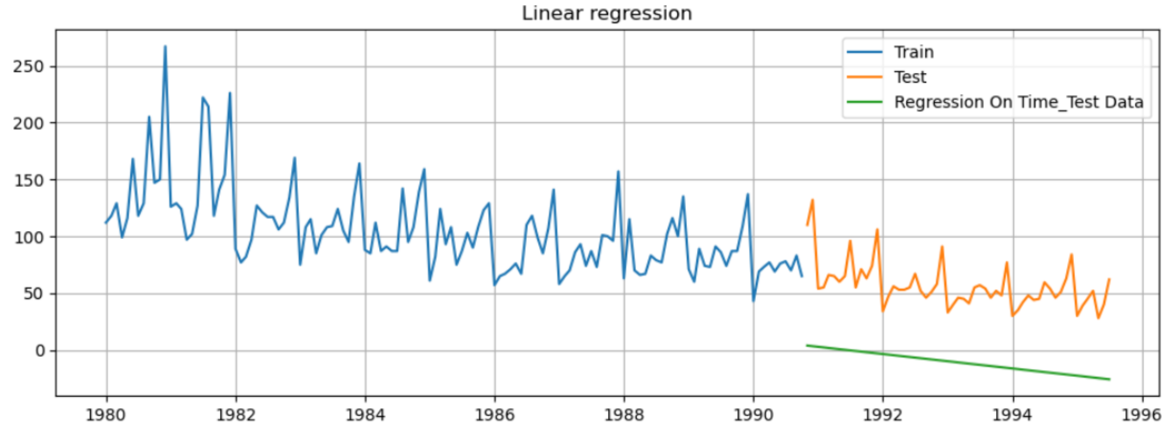


Fig 9 – Linear regression

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 69.749.

Simple Average Forecast

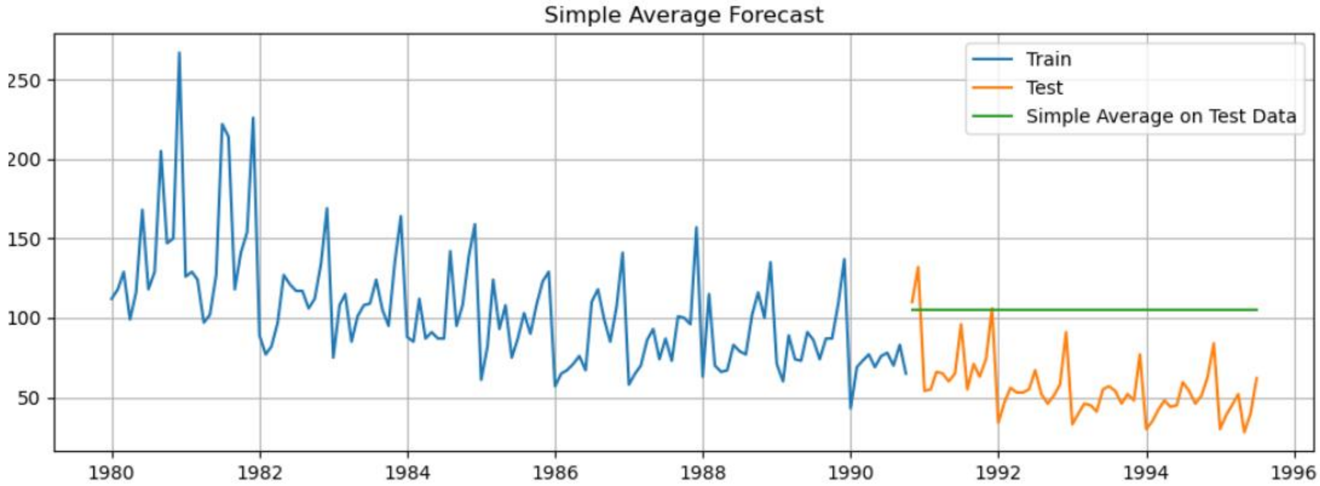


Fig 10– Simple Average Forecast

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 52.010

Moving Average Forecast

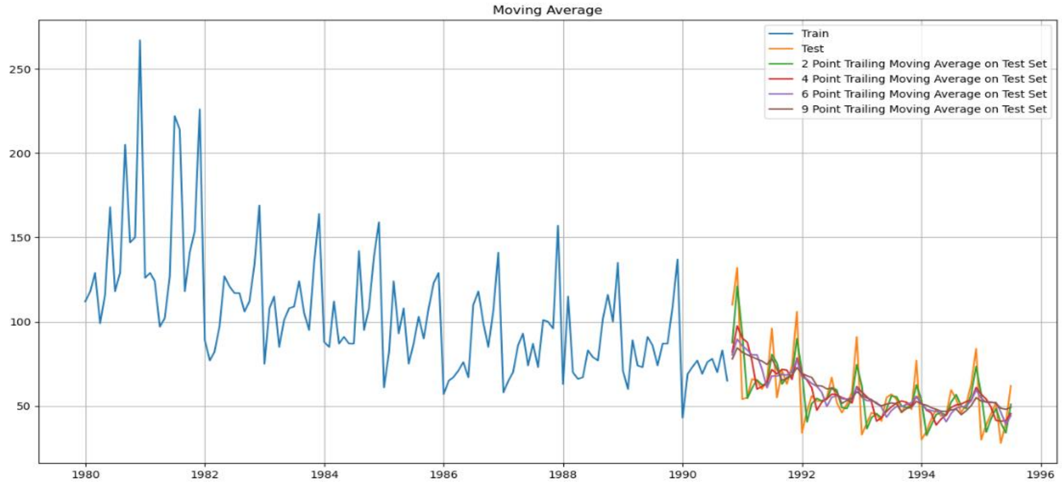


Fig 11–Moving Average Forecast

We have made multiple moving average models with rolling windows varying from 2 to 9. Rolling average is a better method than simple average as it takes into account only the previous n values to make the prediction, where n is the rolling window defined. This takes into account the recent trends and is in general more accurate. The higher the rolling window, the smoother will be its curve, since more values are being taken into account

Based on the below RMSE scores, 2 point trailing Moving Average is selected.
For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.857
For 4 point Moving Average Model forecast on the Training Data, RMSE is 15.417
For 6 point Moving Average Model forecast on the Training Data, RMSE is 15.855
For 9 point Moving Average Model forecast on the Training Data, RMSE is 16.402

RMSE calculated for 2 point trailing Moving Average model is 11.857

Exponential Models (Single, Double, Triple):

In time series forecasting using Exponential Smoothing models in Python's statsmodels library, setting the parameter optimized=True allows the model to automatically find the best values for the smoothing parameters (alpha, beta, gamma) that minimize the error measure (e.g., Mean Squared Error).

Simple Exponential Smoothing:

Parameters:

```
{'smoothing_level': 0.1027210637690678,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.26261326012983,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
```

'lamda': None,
'remove_bias': False}

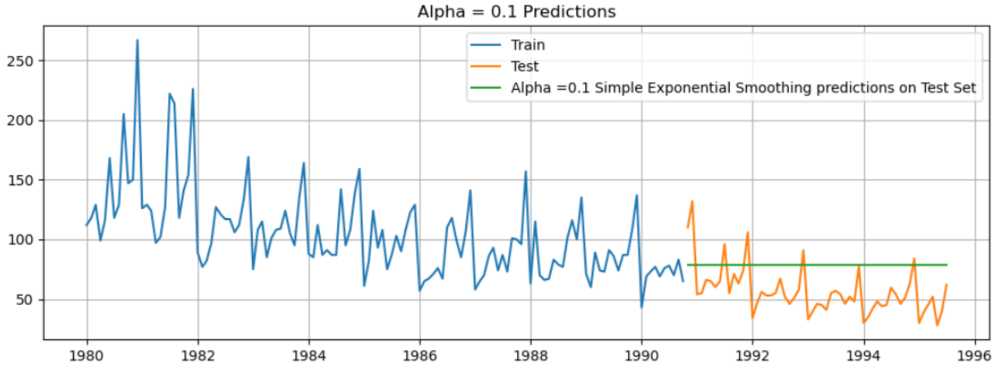


Fig 12– Simple Exponential Smoothing

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 29.827370

Double Exponential Smoothing:

Parameters:

```
{'smoothing_level': 1.4901161193847656e-08,
 'smoothing_trend': 1.1142151039754693e-10,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 139.3541894587564,
 'initial_trend': -0.529189091674724,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

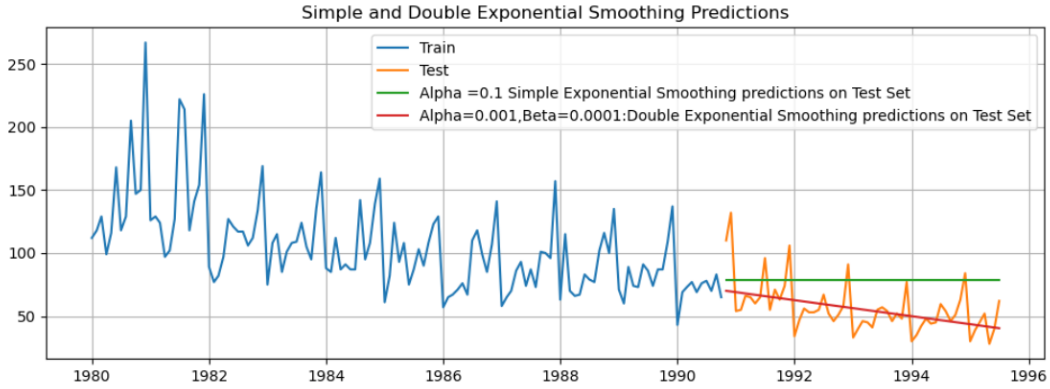


Fig 13– Double Exponential Smoothing

It is clear the predicted values are very far off from the actual values

RMSE calculated for this model is 17.467105511059792

Triple exponential Smoothing (Additive):

Parameters:

```
{'smoothing_level': 0.08821436937988711,
 'smoothing_trend': 1.800307905077045e-05,
 'smoothing_seasonal': 0.0003555282543494388,
 'damping_trend': nan,
 'initial_level': 146.89163094636658,
 'initial_trend': -0.5560607891056817,
 'initial_seasons': array([-31.04964235, -18.73445697, -10.76315366, -21.43688134, -12.5757113 ,  -6.95772026,
2.76047953,   8.92527792, 4.8839166 ,   2.94720422,  19.88703135,  63.66923412]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

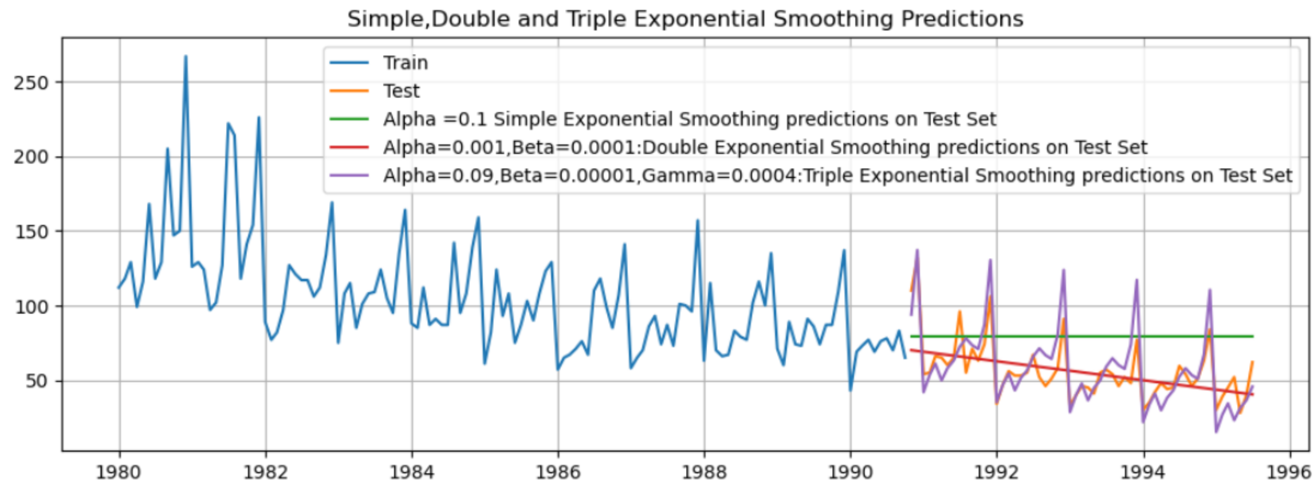


Fig 14– Triple Exponential Smoothing (Additive)

RMSE calculated for this model is 13.848022

Triple exponential Smoothing (Multiplicative):

Parameters:

```
{'smoothing_level': 0.10169520380955438,
 'smoothing_trend': 0.0007058642829212204,
 'smoothing_seasonal': 1.2230189753146669e-05,
 'damping_trend': nan,
 'initial_level': 127.00110122061433,
 'initial_trend': -0.5130818887786344,
 'initial_seasons': array([0.86329436, 0.9758005 , 1.06696094, 0.93565741, 1.05039692, 1.13162774, 1.24587724, 1.33247998,
1.24729237, 1.22701271, 1.41107894, 1.94589043]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

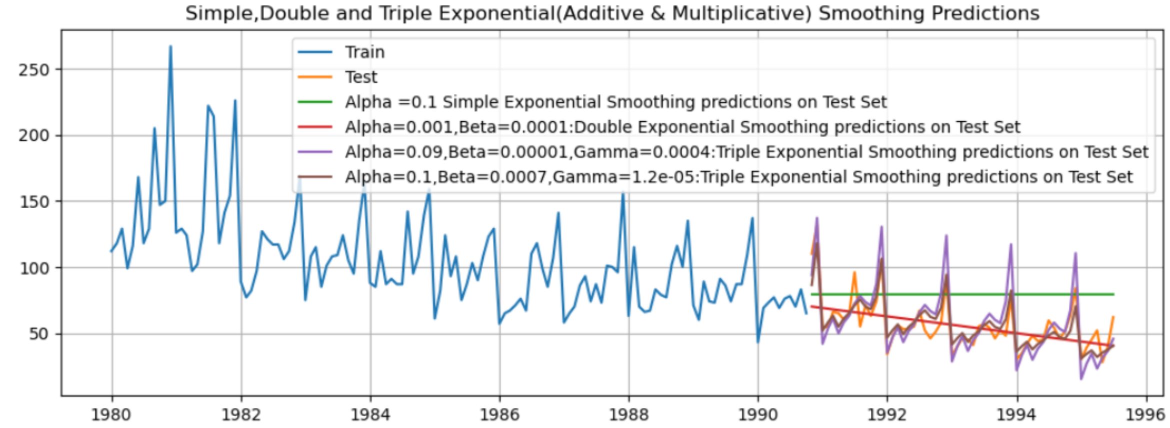
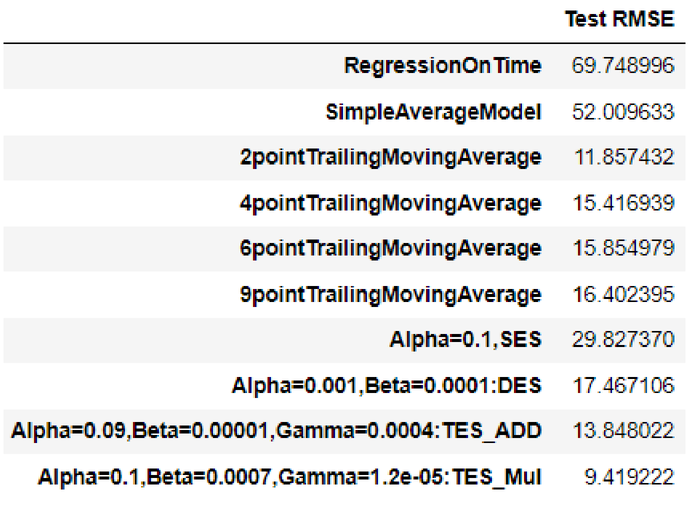


Fig 15– Triple exponential Smoothing (Multiplicative)

TES with Multiplicative seasonality fit better on test data than additive seasonality based on the above plot.

RMSE calculated for this model is 9.419222

Check the performance of the models built:



	Test RMSE
RegressionOnTime	69.748996
SimpleAverageModel	52.009633
2pointTrailingMovingAverage	11.857432
4pointTrailingMovingAverage	15.416939
6pointTrailingMovingAverage	15.854979
9pointTrailingMovingAverage	16.402395
Alpha=0.1,SES	29.827370
Alpha=0.001,Beta=0.0001:DES	17.467106
Alpha=0.09,Beta=0.00001,Gamma=0.0004:TES_ADD	13.848022
Alpha=0.1,Beta=0.0007,Gamma=1.2e-05:TES_Mul	9.419222

Table 4 –Model comparison

Till now, The best model had both a multiplicative seasonality with additive trends which was evaluated using the RMSE metric (9.42).

Check for Stationarity:

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

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

- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

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

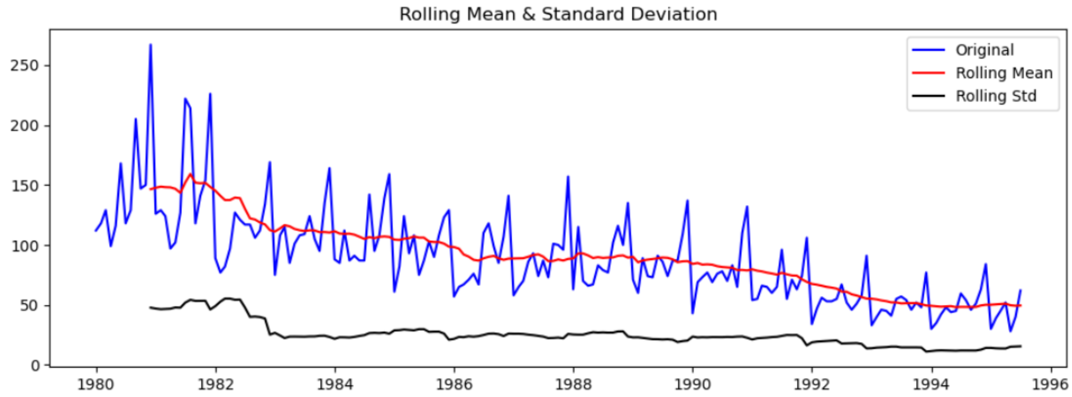


Fig 16– ADF test- Original data

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.892338
p-value                        0.335674
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)            -3.468726
Critical Value (5%)            -2.878396
Critical Value (10%)            -2.575756
```

The p-value 0.33 is very large, and not smaller than 0.05. We see that at 5% significant level the Time Series is non-stationary.

First-Order differencing:

In order to try and make the series stationary we used the differencing approach. We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped

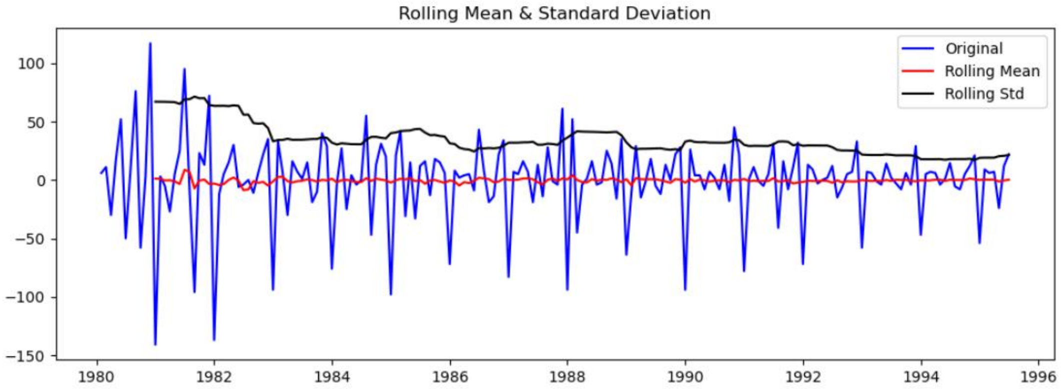


Fig 17– ADF test- Differenced data

```
Results of Dickey-Fuller Test:
Test Statistic                 -8.032729e+00
p-value                         1.938803e-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
```

Dickey - Fuller test was 0.000, which is obviously less than 0.05. Hence the null hypothesis that the series is not stationary at difference = 1 was rejected, which implied that the series has indeed become stationary after we performed the differencing. Null hypothesis was rejected since the p-value was less than alpha i.e. 0.05.

Also the rolling mean plot was a straight line this time around. Also the series looked more or less the same from both the directions, indicating stationarity.

Generate ACF & PACF Plot and find the AR, MA values

Original data: (Non-Stationary)

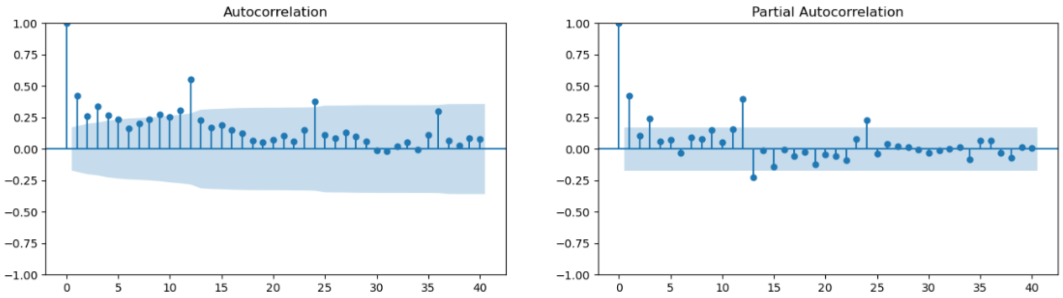


Fig 18–ACF and PACF Plots- Original data

Differenced data: (Stationary):

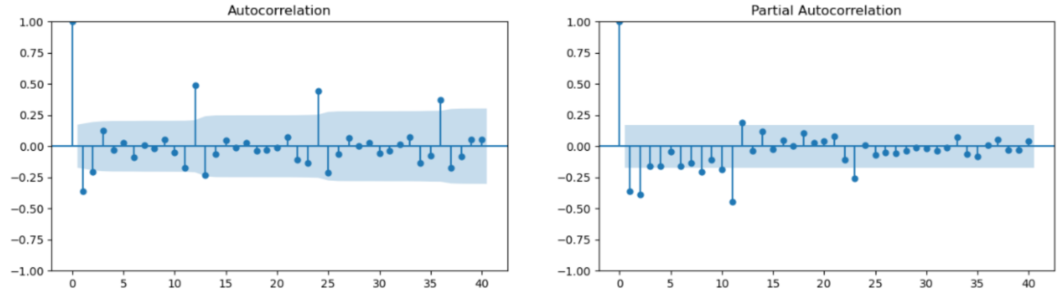


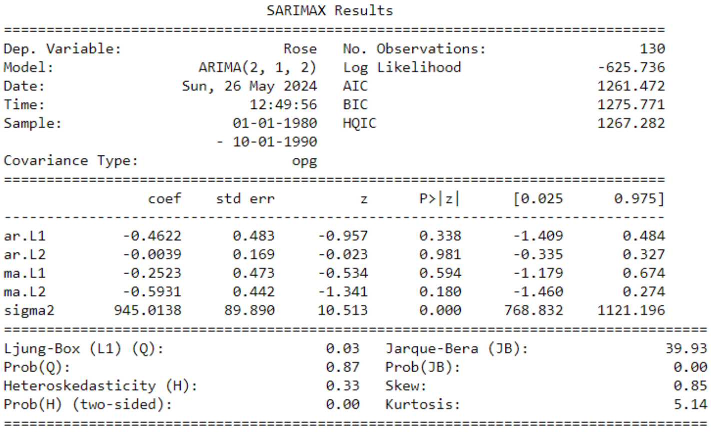
Fig 19–ACF and PACF Plots- Differenced data

- Looking at ACF plot we can see a decay after lag 2 for differenced data. hence we select the q value to be 2. i.e. q=2.
- Looking at PACF plot we can again see significant bars till lag 2 for differenced series which is stationary in nature. Hence we choose p value to be 2. i.e. p=2.
- d values will be 1 as data is stationary after first order differencing.

Manual- ARIMA Model

Based on ACF and PACF plots, the values selected for manual ARIMA:- p=2, d=1, q=2

Summary from this manual ARIMA model:



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  130
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -625.736
Date:                Sun, 26 May 2024   AIC                           1261.472
Time:                        12:49:56   BIC                           1275.771
Sample:                    01-01-1980   HQIC                          1267.282
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4622      0.483     -0.957      0.338      -1.409       0.484
ar.L2         -0.0039      0.169     -0.023      0.981      -0.335       0.327
ma.L1         -0.2523      0.473     -0.534      0.594      -1.179       0.674
ma.L2         -0.5931      0.442     -1.341      0.180      -1.460       0.274
sigma2       945.0138     89.890     10.513      0.000     768.832    1121.196
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                39.93
Prob(Q):                              0.87   Prob(JB):                         0.00
Heteroskedasticity (H):               0.33   Skew:                             0.85
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.14
===================================================================================
```

Diagnostic Plots:

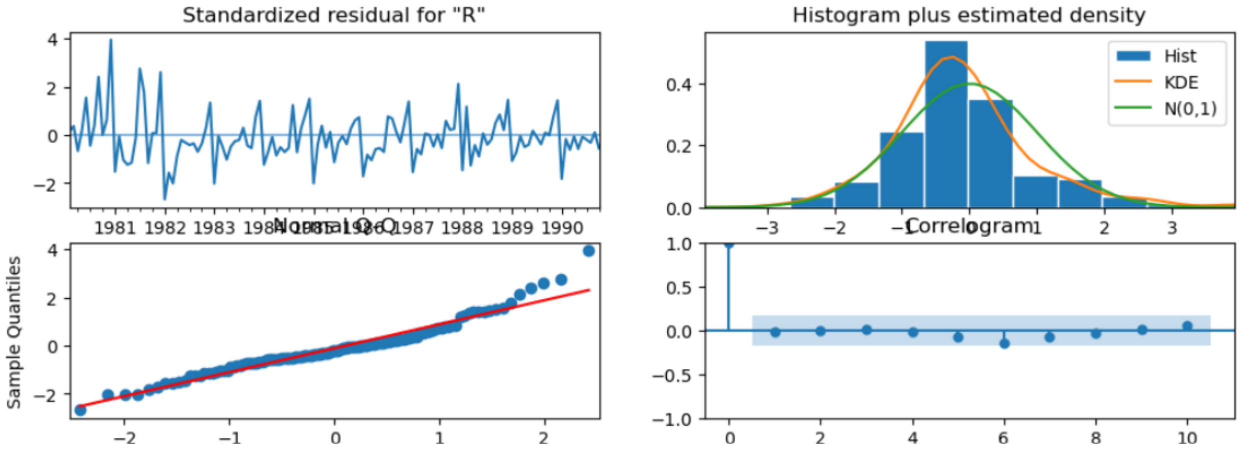


Fig 20–Manual Arima (Diagnostic Plots)

Residuals shows normal distribution and no correlations.. So it is average fit

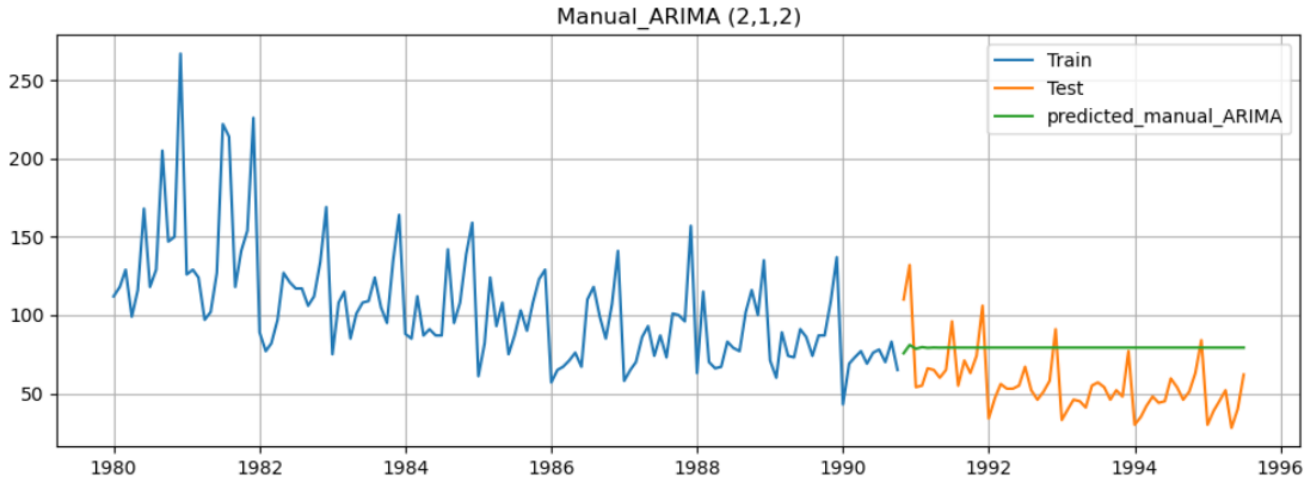


Fig 21–Manual Arima

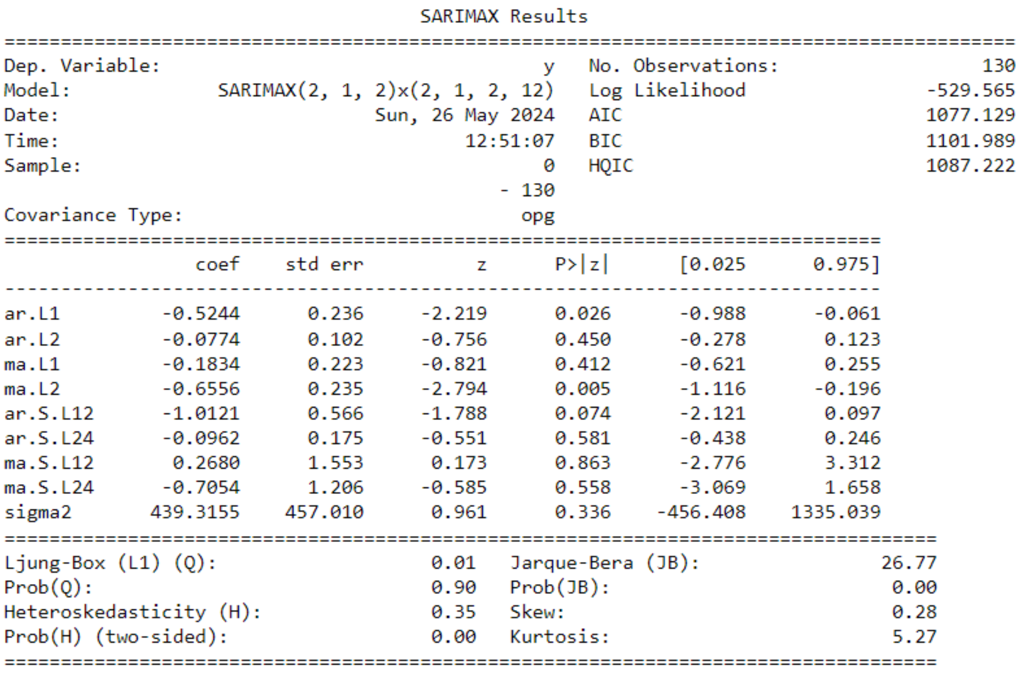
RMSE calculated for this model is 30.086215

Manual- SARIMA Model:

Identified Seasonal Parameters: P=2and Q=2were determined from the ACF and PACF plots.

Final parameters : SARIMAX(2, 1, 2)x(2, 1, 2, 12)

Summary from this manual SARIMA model:



```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  130
Model:             SARIMAX(2, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -529.565
Date:                            Sun, 26 May 2024   AIC                           1077.129
Time:                                    12:51:07   BIC                           1101.989
Sample:                                         0   HQIC                          1087.222
                                            - 130                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5244      0.236     -2.219      0.026      -0.988      -0.061
ar.L2         -0.0774      0.102     -0.756      0.450      -0.278       0.123
ma.L1         -0.1834      0.223     -0.821      0.412      -0.621       0.255
ma.L2         -0.6556      0.235     -2.794      0.005      -1.116      -0.196
ar.S.L12      -1.0121      0.566     -1.788      0.074      -2.121       0.097
ar.S.L24      -0.0962      0.175     -0.551      0.581      -0.438       0.246
ma.S.L12       0.2680      1.553      0.173      0.863      -2.776       3.312
ma.S.L24      -0.7054      1.206     -0.585      0.558      -3.069       1.658
sigma2       439.3155    457.010      0.961      0.336    -456.408    1335.039
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                26.77
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               0.35   Skew:                             0.28
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.27
===================================================================================
```

Diagnostic Plots:

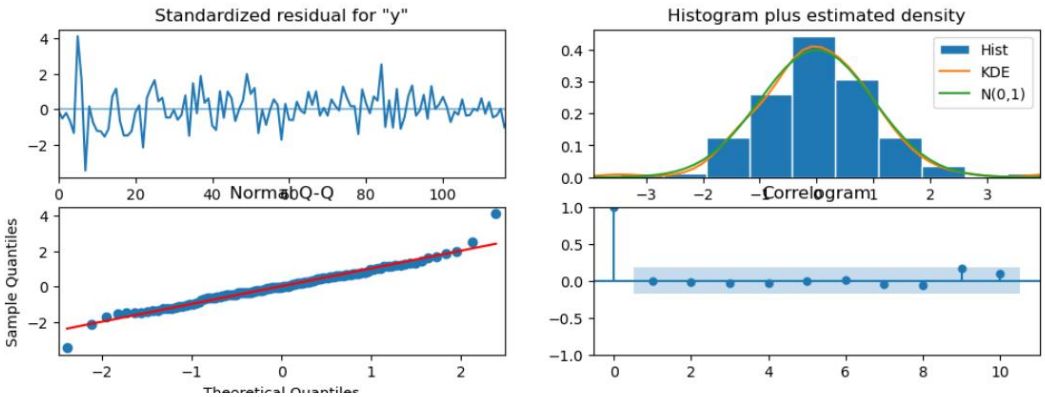


Fig 22–Manual Sarima (Diagnostic Plots)

Residuals shows normal distribution and no correlations.. So it is good fit

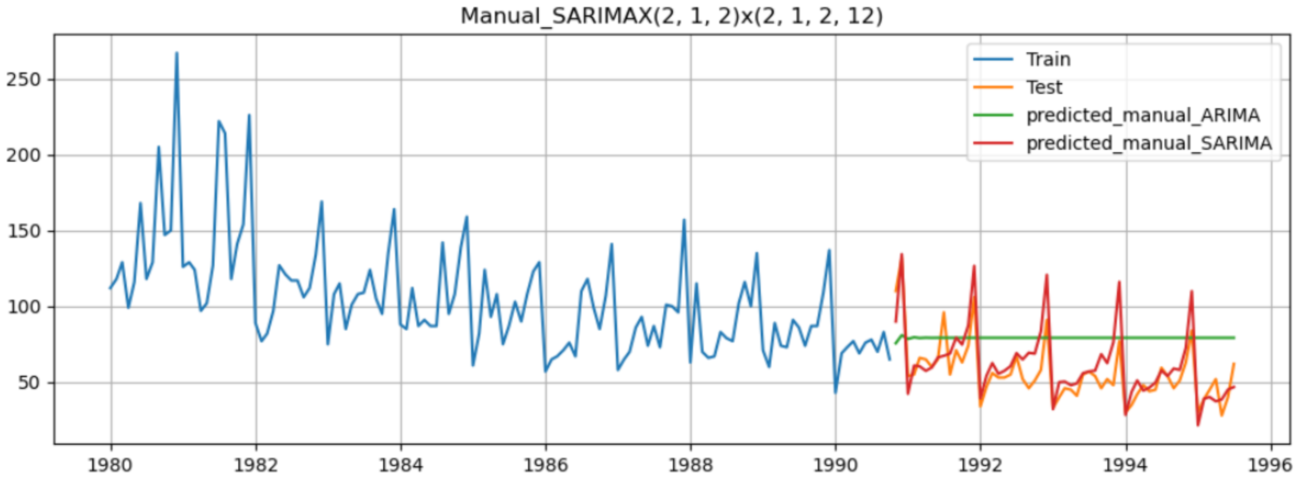


Fig 23–Manual Sarima

RMSE calculated for this model is 13.412014

Auto ARIMA:

We employed a for loop for determining the optimum values of p,d,q,
where p is the order of the AR (Auto-Regressive) part of the model,
while q is the order of the MA (Moving Average) part of the model.
d is the differencing that is required to make the series stationary.

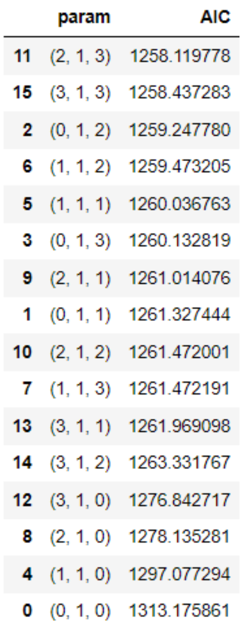
p,q values in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for d, since we had already determined d to be 1, while checking for stationarity using the ADF test.

Some parameter combinations for the Model...

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)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected.

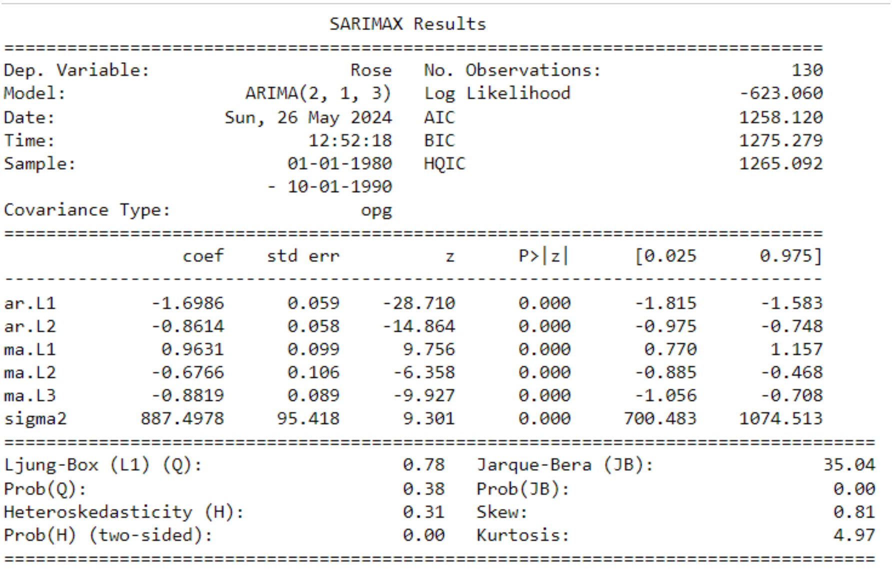
Best Model:



	param	AIC
11	(2, 1, 3)	1258.119778
15	(3, 1, 3)	1258.437283
2	(0, 1, 2)	1259.247780
6	(1, 1, 2)	1259.473205
5	(1, 1, 1)	1260.036763
3	(0, 1, 3)	1260.132819
9	(2, 1, 1)	1261.014076
1	(0, 1, 1)	1261.327444
10	(2, 1, 2)	1261.472001
7	(1, 1, 3)	1261.472191
13	(3, 1, 1)	1261.969098
14	(3, 1, 2)	1263.331767
12	(3, 1, 0)	1276.842717
8	(2, 1, 0)	1278.135281
4	(1, 1, 0)	1297.077294
0	(0, 1, 0)	1313.175861

Table 5 – Auto arima models

Based on the above results , Parameters of the best model are (2,1,3)
The summary report for the ARIMA model with values (p=2,d=1,q=3).



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  130
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -623.060
Date:                Sun, 26 May 2024   AIC                           1258.120
Time:                        12:52:18   BIC                           1275.279
Sample:                    01-01-1980   HQIC                          1265.092
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6986      0.059    -28.710      0.000      -1.815      -1.583
ar.L2         -0.8614      0.058    -14.864      0.000      -0.975      -0.748
ma.L1          0.9631      0.099      9.756      0.000       0.770       1.157
ma.L2         -0.6766      0.106     -6.358      0.000      -0.885      -0.468
ma.L3         -0.8819      0.089     -9.927      0.000      -1.056      -0.708
sigma2       887.4978     95.418      9.301      0.000     700.483    1074.513
===================================================================================
Ljung-Box (L1) (Q):                   0.78   Jarque-Bera (JB):                35.04
Prob(Q):                              0.38   Prob(JB):                         0.00
Heteroskedasticity (H):               0.31   Skew:                             0.81
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.97
===================================================================================
```

Diagnostic plots:

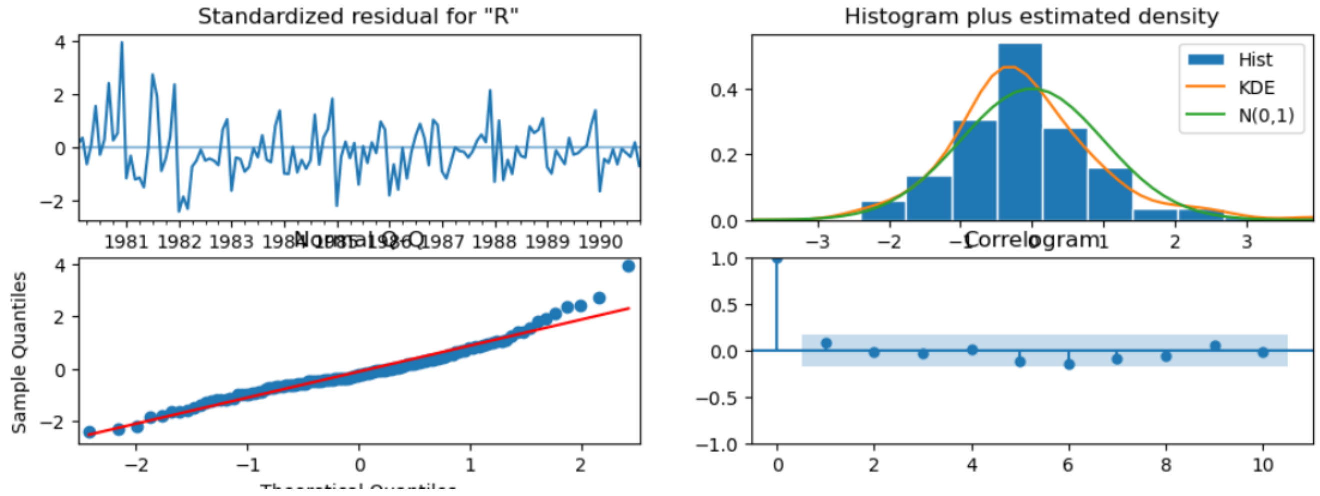


Fig 24 –Auto arima- Diagnostic plots

Residuals shows normal distribution and no correlations.. So it is good fit

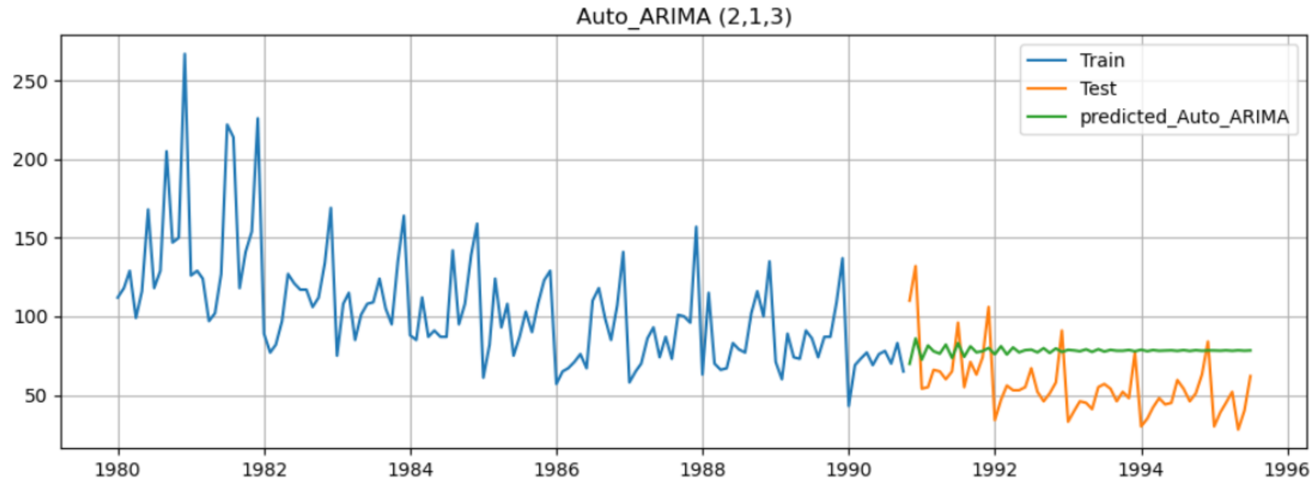


Fig 25–Auto arima

RMSE calculated for this model is 29.184835

Auto SARIMA:

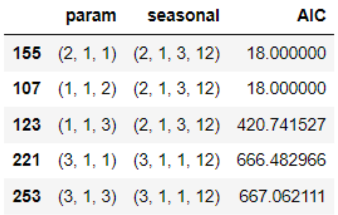
A similar for loop like AUTO_ARIMA with below values was employed.

p = q = range(0, 4)
d = range(0, 2)
P = Q = range(0, 4)
D = range(1 2)
s = 12

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 1, 1, 12)
Model: (0, 1, 2)(0, 1, 2, 12)
Model: (0, 1, 3)(0, 1, 3, 12)
Model: (1, 1, 0)(1, 1, 0, 12)
Model: (1, 1, 1)(1, 1, 1, 12)
Model: (1, 1, 2)(1, 1, 2, 12)
Model: (1, 1, 3)(1, 1, 3, 12)
Model: (2, 1, 0)(2, 1, 0, 12)
Model: (2, 1, 1)(2, 1, 1, 12)
Model: (2, 1, 2)(2, 1, 2, 12)
Model: (2, 1, 3)(2, 1, 3, 12)
Model: (3, 1, 0)(3, 1, 0, 12)
Model: (3, 1, 1)(3, 1, 1, 12)
Model: (3, 1, 2)(3, 1, 2, 12)
Model: (3, 1, 3)(3, 1, 3, 12)
```

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected.

Best Model:

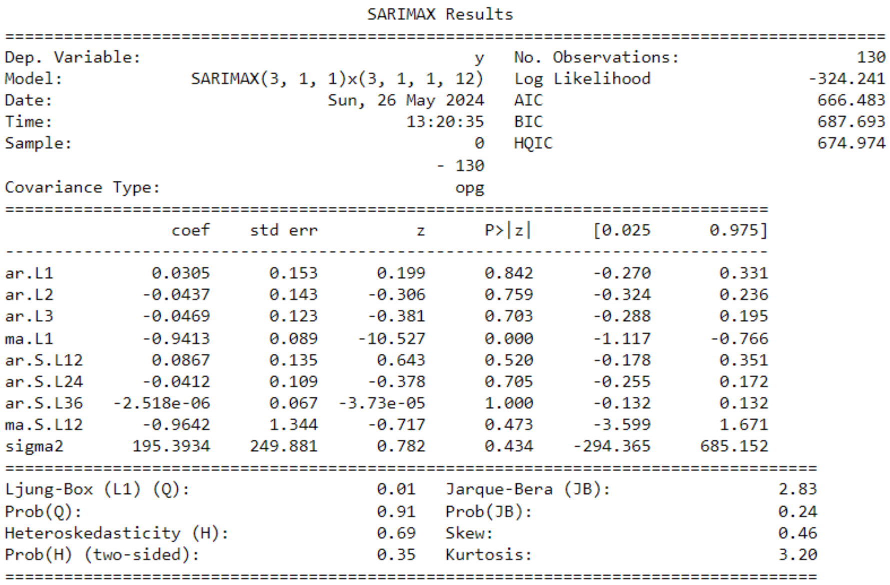


	param	seasonal	AIC
155	(2, 1, 1)	(2, 1, 3, 12)	18.000000
107	(1, 1, 2)	(2, 1, 3, 12)	18.000000
123	(1, 1, 3)	(2, 1, 3, 12)	420.741527
221	(3, 1, 1)	(3, 1, 1, 12)	666.482966
253	(3, 1, 3)	(3, 1, 1, 12)	667.062111

Table 5 – Auto Sarima models

However first 3 models showed low AIC but RMSE was very much higher. The 4th Best model gave low RMSE among others. So Parameters of the best model are (3, 1, 1)x(3, 1, [1], 12)

The summary report for the SARIMA model



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  130
Model:             SARIMAX(3, 1, 1)x(3, 1, 1, 12)   Log Likelihood                -324.241
Date:                            Sun, 26 May 2024   AIC                            666.483
Time:                                    13:20:35   BIC                            687.693
Sample:                                         0   HQIC                           674.974
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0305      0.153      0.199      0.842      -0.270       0.331
ar.L2         -0.0437      0.143     -0.306      0.759      -0.324       0.236
ar.L3         -0.0469      0.123     -0.381      0.703      -0.288       0.195
ma.L1         -0.9413      0.089    -10.527      0.000      -1.117      -0.766
ar.S.L12       0.0867      0.135      0.643      0.520      -0.178       0.351
ar.S.L24      -0.0412      0.109     -0.378      0.705      -0.255       0.172
ar.S.L36   -2.518e-06      0.067  -3.73e-05      1.000      -0.132       0.132
ma.S.L12      -0.9642      1.344     -0.717      0.473      -3.599       1.671
sigma2       195.3934    249.881      0.782      0.434    -294.365     685.152
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 2.83
Prob(Q):                              0.91   Prob(JB):                         0.24
Heteroskedasticity (H):               0.69   Skew:                             0.46
Prob(H) (two-sided):                  0.35   Kurtosis:                         3.20
===================================================================================
```

Diagnostic plots:

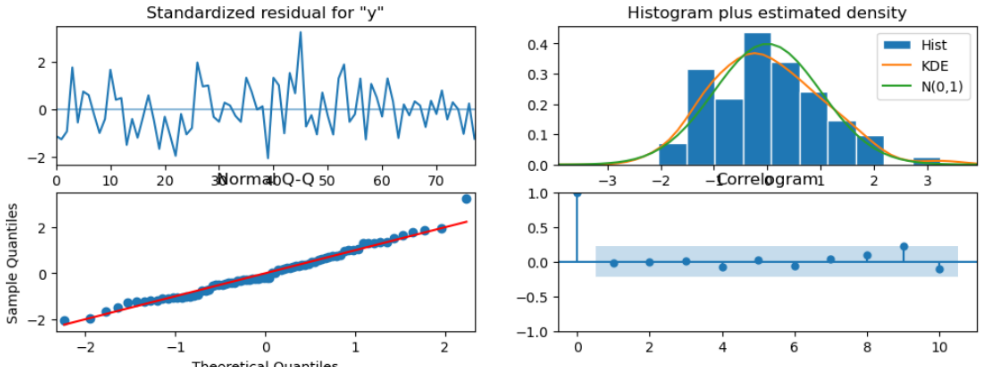


Fig 26 –Auto Sarima- Diagnostic plots

Residuals shows normal distribution and no correlations.. So it is good fit

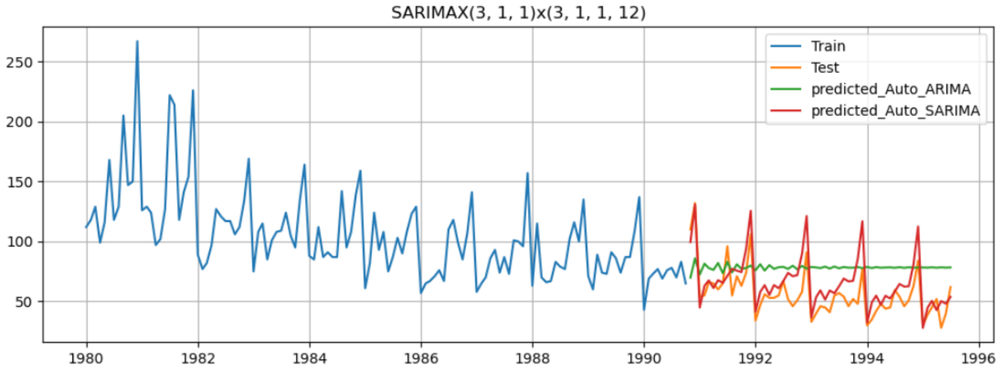
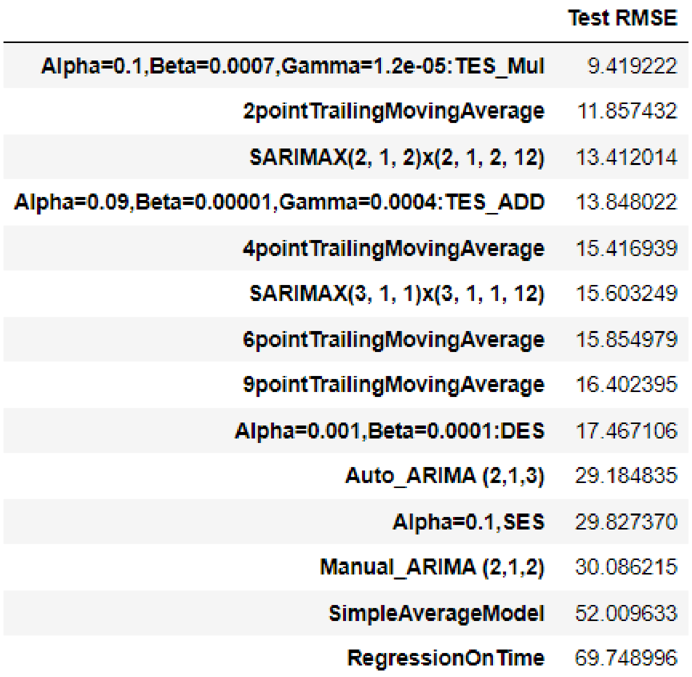


Fig 27–Auto Sarima

RMSE calculated for this model is 15.603249

Compare the performance of the models:



	Test RMSE
Alpha=0.1,Beta=0.0007,Gamma=1.2e-05:TES_Mul	9.419222
2pointTrailingMovingAverage	11.857432
SARIMAX(2, 1, 2)x(2, 1, 2, 12)	13.412014
Alpha=0.09,Beta=0.00001,Gamma=0.0004:TES_ADD	13.848022
4pointTrailingMovingAverage	15.416939
SARIMAX(3, 1, 1)x(3, 1, 1, 12)	15.603249
6pointTrailingMovingAverage	15.854979
9pointTrailingMovingAverage	16.402395
Alpha=0.001,Beta=0.0001:DES	17.467106
Auto_ARIMA (2,1,3)	29.184835
Alpha=0.1,SES	29.827370
Manual_ARIMA (2,1,2)	30.086215
SimpleAverageModel	52.009633
RegressionOnTime	69.748996

Table 6– Model Evaluation

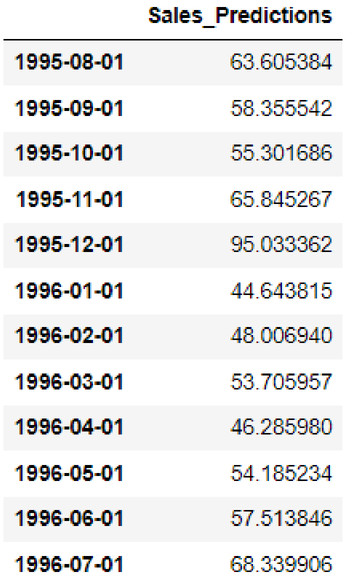
Analysis of Results

- Triple Exponential Smoothing (TES) - Multiplicative (Alpha=0.1, Beta=0.0007, Gamma=1.2e-05) has the lowest RMSE of 9.419222, making it the best-performing model among the given options.
- 2-point Trailing Moving Average is the second best with an RMSE of 11.857432, but it is significantly higher than the TES_Mul model. But this model is good for forecast and computationally expensive as train data needs to be appended new data and refit.
- SARIMAX models (SARIMAX(2, 1, 2)x(2, 1, 2, 12) and SARIMAX(3, 1, 1)x(3, 1, 1, 12)) also perform well but do not outperform TES_Mul.

Conclusion:
The Triple Exponential Smoothing (TES) with Multiplicative seasonality model (Alpha=0.1, Beta=0.0007, Gamma=1.2e-05) is the best model among the provided options based on its lowest Test RMSE of 9.419222. This model is recommended for forecasting wine sales for ABC Estate Wines.

Rebuild the best model using the entire data:

Entire data has been fit into TES model and forecasted for next 12 months.



	Sales_Predictions
1995-08-01	63.605384
1995-09-01	58.355542
1995-10-01	55.301686
1995-11-01	65.845267
1995-12-01	95.033362
1996-01-01	44.643815
1996-02-01	48.006940
1996-03-01	53.705957
1996-04-01	46.285980
1996-05-01	54.185234
1996-06-01	57.513846
1996-07-01	68.339906

Table 8 – Final predictions

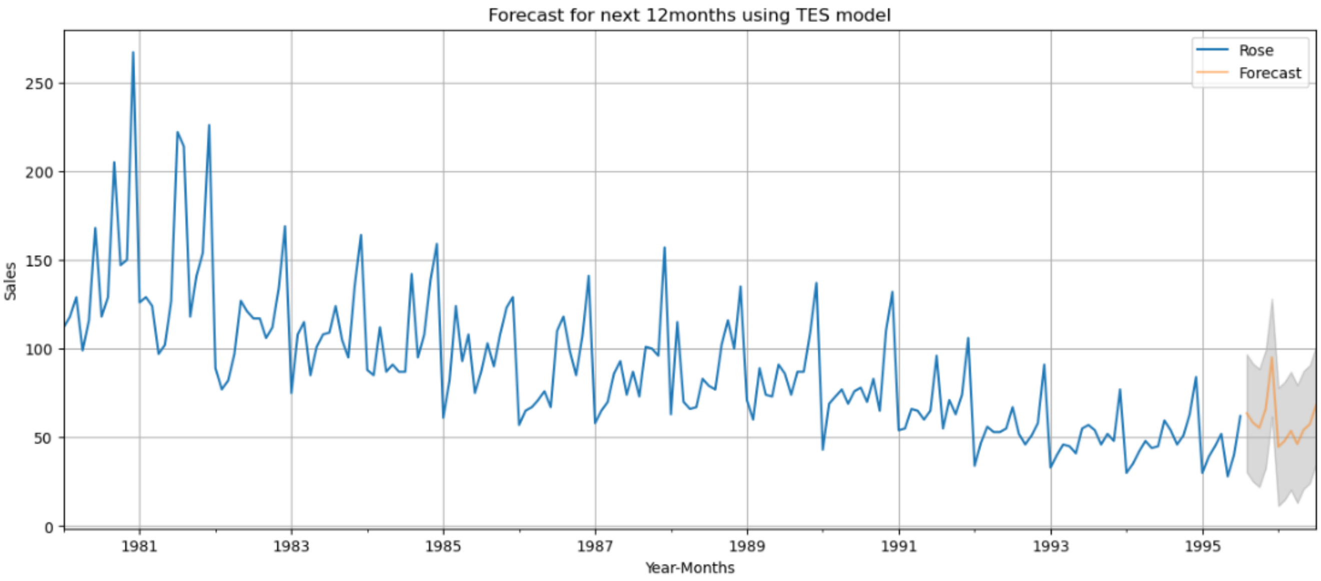


Fig 28–Final forecast

Actionable Insights & Recommendations

1. **Rose Wine Strategy:** Investigate and address the decline in Rose wine sales, potentially revamping production and marketing strategies to regain market share.
2. **Trend Continuation:** The downward trend in Rose wine sales is projected to continue, necessitating strategic interventions to reverse this pattern.
3. **Seasonal Sales Influence:** Wine sales increase during the festival season and drop during the peak winter month of January.
4. **Year-Round Campaigns**: Implement campaigns to boost wine consumption throughout the year, particularly during periods of subdued sales.
5. **Target Lean Periods:** Focus on promotional campaigns from April to June to maximize impact, as sales are typically low during these months.
6. **Peak Period Strategy:** Avoid heavy campaigns during festivals when sales are already high, as they are less likely to have a significant impact.
7. **Winter Campaigns:** Avoid campaigns in January due to low consumer interest, driven by climatic factors.