Time Series Analysis Business Report

DSBA

'Rose'

Vikash Kumar

Aug'22 Batch

Contents

Table Content

Figures Content

Problem Statement:

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

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

Time Series Analysis & Forecasting of Rose Wine Sales

QN1 Read the data as an appropriate Time Series data and plot the data.

- To enable us to address the business issue, we have imported necessary libraries:
  - Basic - from Numpy and Pandas
  - Data visualization - Matplotlib, Seaborn,
  - General – Python standard warnings
  - Statistics - scipy and scikit learn for time series modelling and plotting time series specific plots
- The dataset in csv format 'Sparkling' was loaded in python as time series data and we have verified that the same is properly loaded. The Header and Tail of dataset is verified for completeness.

	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

TABLE 1 - HEADER OF DATASET

```
<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: SHAPE & INFORMATION OF THE DATASET

Insights:

Out of 187 entries only 185 are present in row hence its assumed that Two TS value are missing. Data TS range is 1980-01-01 to 1995-07-01

- The shape of the dataset and the information of number of columns, number of records as well as data type, was verified

```
Shape of Rose a TS Data : (187, 1)
```

TABLE 3: SHAPE & INFORMATION OF THE DATASET

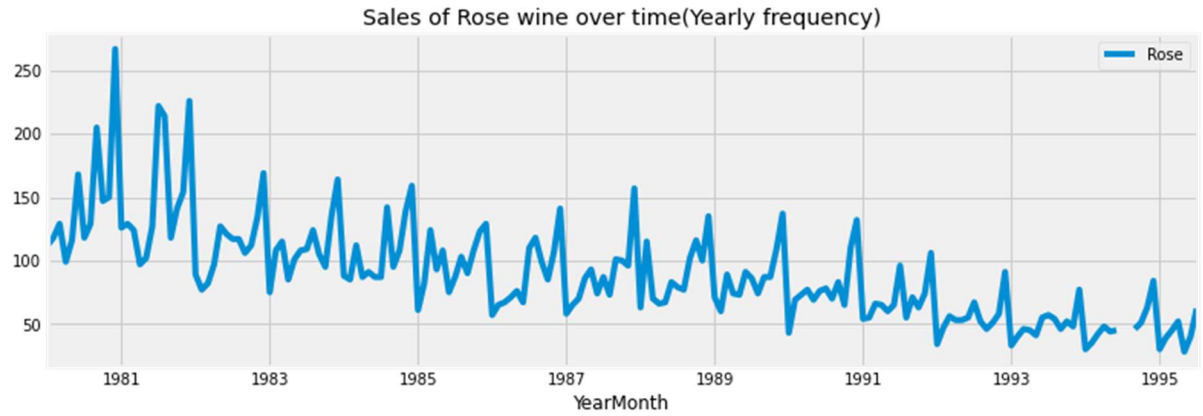


FIGURE 1 YEARLY SALES PATTERN

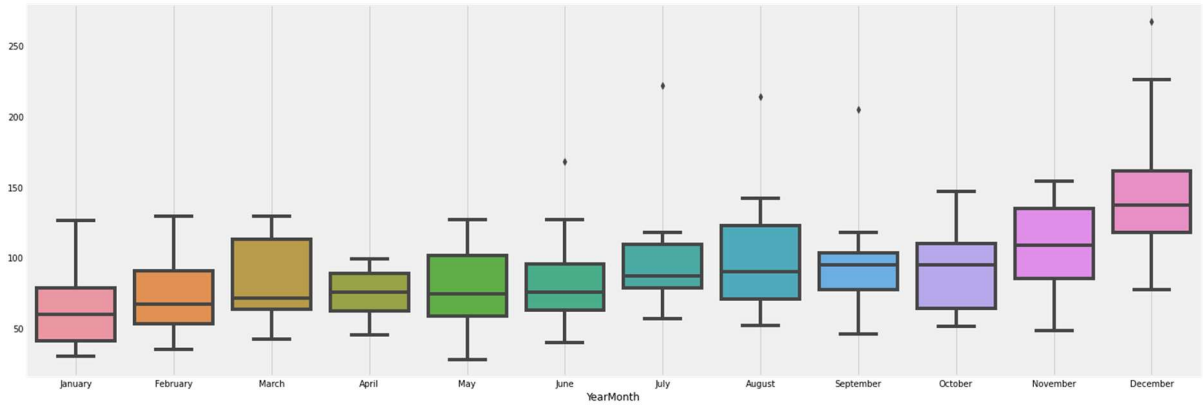


FIGURE 2 MONTHLY SALES PATTERN

Plot Insights:

1.The Time series data showing the sales of Rose wine TS data showing some Missing values in year 1994.

2.There is year after year sales degrowing till last available data record.

3.The maximum sales was in year 1981 while the minimum sales was in 1995.

4.The sales are in range of above 250 to lowest less than 50.

5.Seasonality year wise can be well display by plot.

QN2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

- The descriptive statistics showing 5-point summary is verified

	Rose
count	185
mean	90.394595
std	39.175344
min	28
25%	63
50%	86
75%	112
max	267

TABLE 4 DATA DESCRIPTION

Insights:

1.There is low difference in mean and median and data seems NO skewed.

2.There is low gap between mean sales and median sales, which can be infer as data are in normal distribution

Checking The Null Values in TS Dataset

```
Rose    2
dtype: int64
```

Treatment of Missing Values by Interpolate function of pandas, Interpolate Missing Values But Only Up Two Values

```
Rose    0
dtype: int64
```

TABLE 5 MISSING VALUES

INSIGHTS:

1.There is two missing value has been detected in TS data in Year 1994.
2. This has been filled by interpolation method to go further in modelling.

Decomposition of Rose

These are the components of a time series

Trend - Consistent upwards or downwards slope of a time series
Seasonality - Clear periodic pattern of a time series (like sine function)
Noise/Error - Outliers or missing values White noise has...

Constant mean Constant variance Zero auto-correlation at all lags

Multiplicative Decomposition

1.A Multiplicative model suggests that the components are multiplied together.
2.A Multiplicative model is non-linear such as quadratic or exponential.
3.Changes increase or decrease over time.
4.A non-linear seasonality has an increasing or decreasing frequency (width of the cycles) and / or amplitude (height of the cycles) over time.

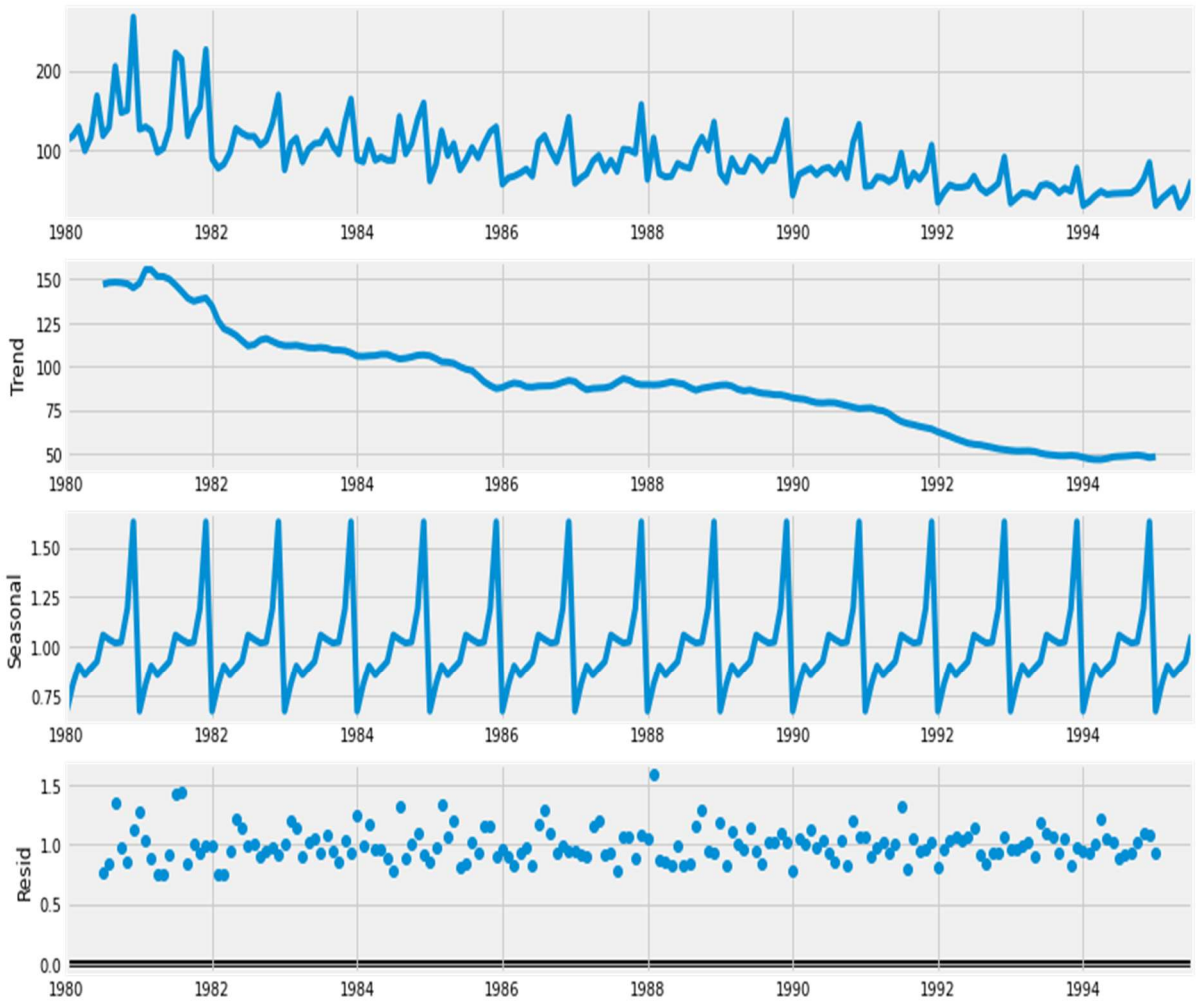


FIGURE 3 SEASONAL DECOMPOSITION MULTIPLICATIVE

Additive Decomposition

An additive model suggests that the components are added together.
An additive model is linear where changes over time are consistently made by the same amount.
A linear seasonality has the same frequency (width of the cycles) and amplitude (height of the cycles).

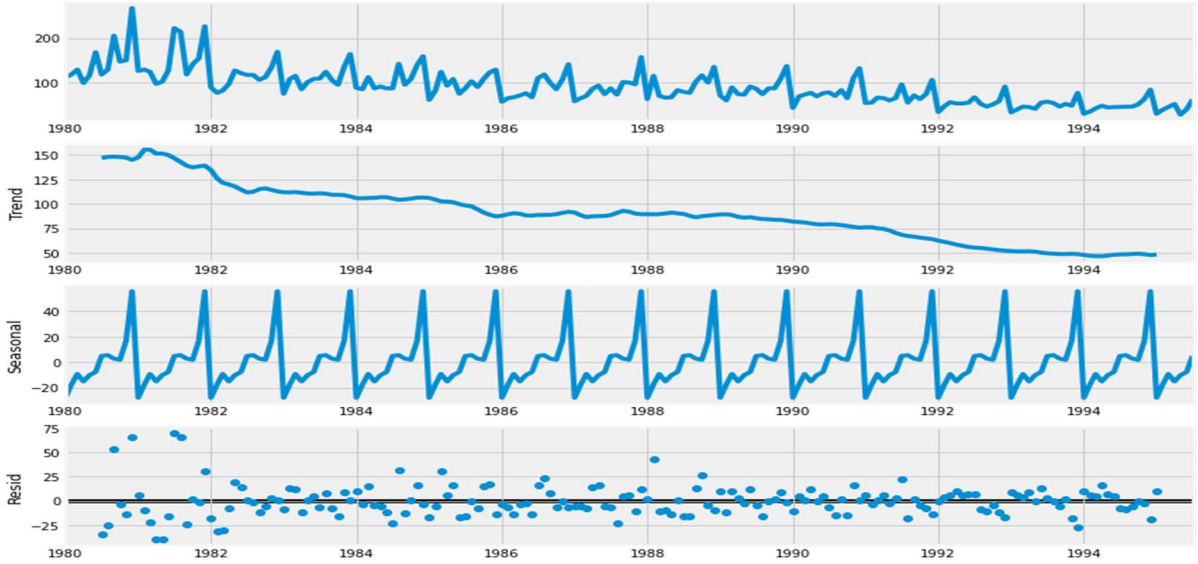


FIGURE 4 SEASONAL DECOMPOSITION ADDITIVE

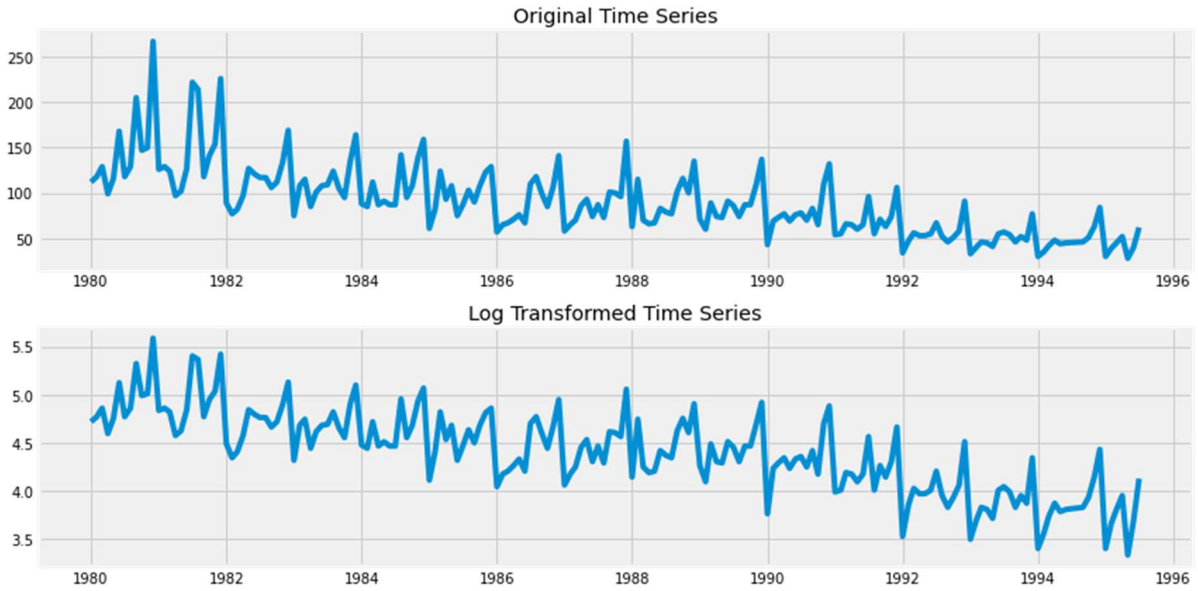


TABLE 6 ORIGINAL VS LOG TRANSFORMED PLOT

Insights:

Some of our key observations from Decomposition plots:
1.Multiplicative model is more effective over additive model to elaborate the time series dataset.
2.Running the above code performs the decomposition, and plots the 4 resulting series.

3.We observe that the trend and seasonality are clearly separated.
4.Trend: The year wise data in plot follow the sharp Decreasing sales trend, it can be considered as Degrowth of Rose sales.
5.Seasonality: Seasonal plot displays a fairly consistent month-on-month pattern.
6.The Multiplicative model are performing better in decomposition because the residual are more align to centre line.
7.The sale of Rose touches the bottom in year of 1995 as continuous downward trend.

QN3.Split the data into training and test. The test data should start in 1991.

Train Test Split

Before a forecast method is proposed, the method needs to be validated. For that purpose, data has to be split into two sets i.e. training and testing.
Training data helps in identifying and fitting right model(s) and test data is used to validate the same.
In case of time series data, the test data is the most recent part of the series so that the ordering in the data is preserved.

Training Data

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

... ...

1990-08-01 70.0

1990-09-01 83.0

1990-10-01 65.0

1990-11-01 110.0

1990-12-01 132.0

132 rows × 1 columns

Test Data

Rose

YearMonth

1991-01-01 54.000000

1991-02-01 55.000000

1991-03-01 66.000000

1991-04-01 65.000000

1991-05-01 60.000000

1991-06-01 65.000000

1991-07-01 96.000000

1991-08-01 55.000000

1991-09-01 71.000000

1991-10-01 63.000000

1991-11-01 74.000000

1991-12-01 106.000000

1992-01-01 34.000000

1992-02-01 47.000000

1992-03-01 56.000000

1992-04-01 53.000000

1992-05-01 53.000000

1992-06-01 55.000000

1992-07-01 67.000000

1992-08-01 52.000000

1992-09-01 46.000000

1992-10-01 51.000000

1992-11-01 58.000000

1992-12-01 91.000000

1993-01-01 33.000000

1993-02-01 40.000000

1993-03-01 46.000000

1993-04-01 45.000000

1993-05-01 41.000000

1993-06-01 55.000000

1993-07-01 57.000000

1993-08-01 54.000000

1993-09-01 46.000000

1993-10-01 52.000000

1993-11-01 48.000000

1993-12-01 77.000000

1994-01-01 30.000000

1994-02-01 35.000000

1994-03-01 42.000000

1994-04-01 48.000000

1994-05-01 44.000000

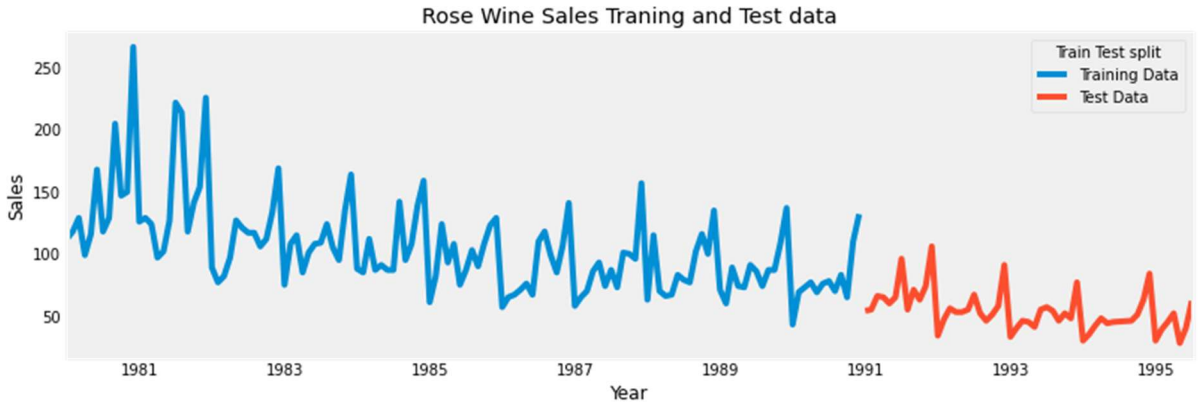
1994-06-01 45.000000

1994-07-01 45.333333

1994-08-01	45.666667
1994-09-01	46.000000
1994-10-01	51.000000
1994-11-01	63.000000
1994-12-01	84.000000
1995-01-01	30.000000
1995-02-01	39.000000
1995-03-01	45.000000
1995-04-01	52.000000
1995-05-01	28.000000
1995-06-01	40.000000
1995-07-01	62.000000

```
Shape of Rose of Train Data: (132, 1)
Shape of Spark of Test Data : (55, 1)
```

TABLE 7 TRAIN TEST SPLIT DATA



The Disjunction at 1991 showing the split of train and test data in Time series.

FIGURE 5 TRAIN TEST SPLIT PLOT

QN4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Exponential Smoothing Method

Smoothing is a technique applied to time series to remove the fine-grained variation between time steps.

Exponential smoothing is a time series forecasting method for univariate data. Exponential smoothing forecasting methods consist of flattening time series data and are similar in that a prediction is a weighted sum of past observations, Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations.

Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be none(N), additive (N), additive damped (Ad), Multiplicative (M) or multiplicative damped (Md).

There are three main types of exponential smoothing time series forecasting methods.

SES - ETS(A, N, N) - Simple Exponential Smoothing with additive errors

A simple method that assumes no systematic structure, an extension that explicitly handles trends, and the most advanced approach that add support for seasonality.

Single Exponential Smoothing, SES for short, also called Simple Exponential Smoothing, is a time series forecasting method for univariate data without a trend or seasonality.

$$\hat{Y}t+1 = \alpha Yt+\alpha(1-\alpha)Yt-1+\alpha(1-\alpha)2Yt-1+\cdots,\ 0<\alpha<1$$

It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

This parameter controls the rate at which the influence of the observations at prior time steps decay exponentially.
Alpha is often set to a value between 0 and 1.
Large values mean that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

 # created class

 # Fitting the Simple Exponential Smoothing model and asking python to choose the optimal parameters

 # Check the parameters

```
{'smoothing_level': 0.09874963957110783,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38708961485827,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

```
Plotting the Training data, Test data and the forecasted values
```

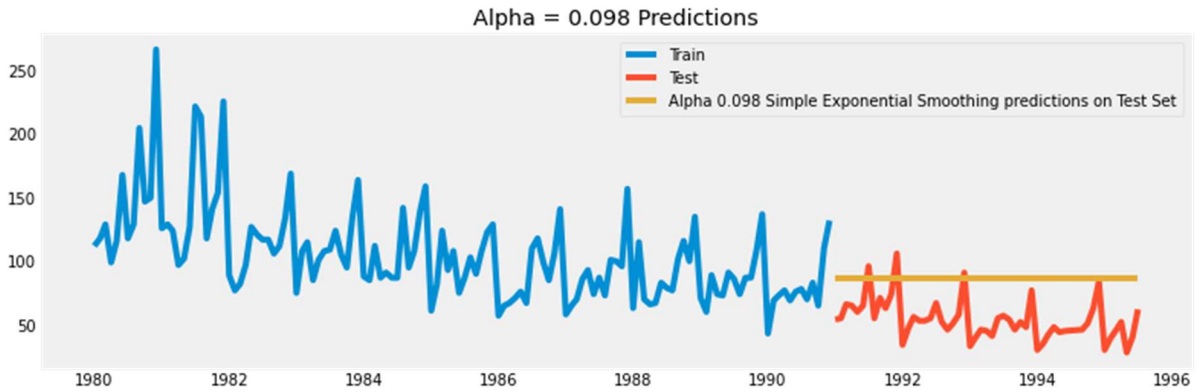


FIGURE 6 SIMPLE EXP SMOOTHING PLOT

Insights:

SES RMSE: 36.796235605069505
The simple Exponential model predicting the straight line, which does not support the level trend and seasonality complete missing.

SES Model Accuracy Check by generating RMSE values

	Test RMSE
Alpha=0.098,SES	**36.796236**

Double Exponential Smoothing or Holt - ETS(A, A, N) - Holt's linear method with additive errors

Double Exponential Smoothing is an extension to Exponential Smoothing that explicitly adds support for trends in the univariate time series.

Forecast Equation: $\hat{Y}t+1=lt+bt$

Level Equation: $lt=\alpha Yt+\alpha(1-\alpha)Yt-1$, $0 < \alpha < 1$

Trend Equation: $bt=\beta(lt-lt-1)+(1-\beta)bt-1$, $0 < \beta < 1$

where, lt is the estimate of level and bt is the trend estimate.

$\alpha$is the smoothing parameter for the level and β is the smoothing parameter for trend.

In addition to the alpha parameter for controlling smoothing factor for the level, an additional smoothing factor is added to control the decay of the influence of the change in trend called beta (b).

The method supports trends that change in different ways: an additive and a multiplicative, depending on whether the trend is linear or exponential respectively.

Double Exponential Smoothing with an additive trend is classically referred to as Holt's linear trend model, named for the developer of the method Charles Holt.

Additive Trend: Double Exponential Smoothing with a linear trend. Multiplicative Trend: Double Exponential Smoothing with an exponential trend. For longer range (multi-step) forecasts, the trend may continue on unrealistically. As such, it can be useful to dampen the trend over time.

Hyperparameters:

Alpha: Smoothing factor for the level.

Beta: Smoothing factor for the trend.

Trend Type: Additive or multiplicative.

Dampen Type: Additive or multiplicative.

Phi: Damping coefficient.

Initializing the Double Exponential Smoothing Model

-----Holt model Exponential Smoothing Estimated Parameters ---

{'smoothing_level': 1.4901247095597348e-08, 'smoothing_trend': 7.3896641488640725e-09, 'smoothing_seasonal': nan, 'damping_trend': nan, 'initial_level': 137.81551313502814, 'initial_trend': -0.4943777717865305, 'initial_seasons': array([], dtype=float64), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

TABLE 8 HOLT EXPO SMOOTHING PARAMETERS

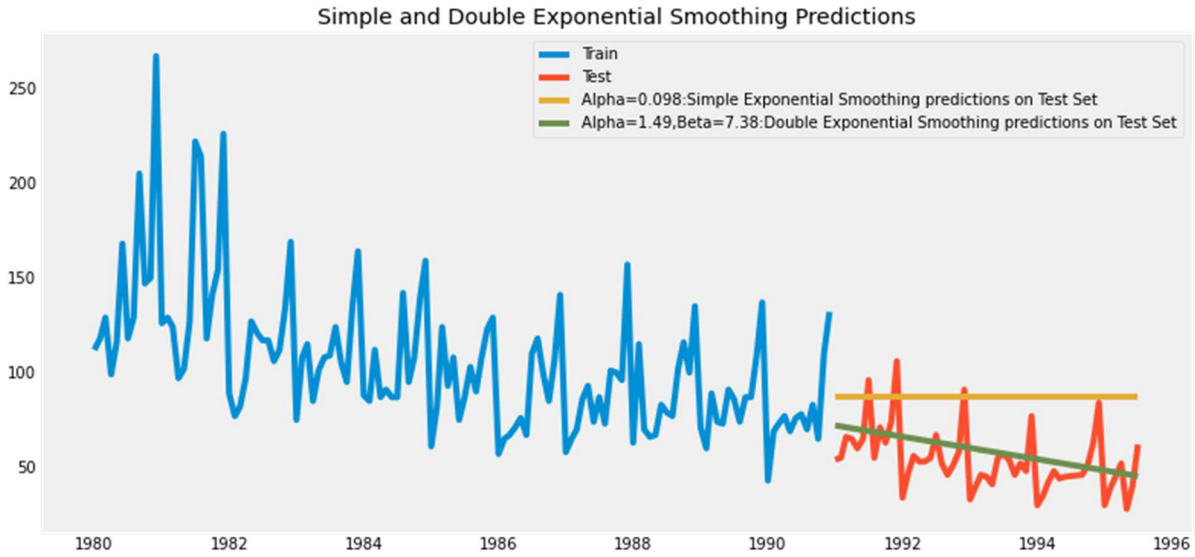


FIGURE 7 DOUBLE EXPO SMOOTHING PLOT

DES Model Accuracy

```
SES RMSE (calculated using statsmodels): 24.081702453785567
```

RMSE Model Summary

	Test RMSE
Alpha=0.098, SES	36.796236
Alpha=1.49, Beta=7.38:DES	15.268957

Triple Exponential Smoothing or Holt-Winters - ETS(A, A, A) - **Holt Winter's** linear method with additive errors

Triple Exponential Smoothing is an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series.

This is an extension of Holt's method when seasonality is found in the data.

$$\text{Forecast Equation}: Yt+1 = lt + bt + st - m(k+1)$$

$$\text{Level Equation}: lt = \alpha(Yt - st - m) + \alpha(1-\alpha)Yt-1, \quad 0 < \alpha < 1$$

$$\text{Trend Equation}: bt = \beta(lt - lt-1) + (1-\beta)bt-1, \quad 0 < \beta < 1$$

$$\text{Seasonal Equation}: \gamma(Yt - lt-1 - bt-1) + (1-\gamma)st-m, \quad 0 < \gamma < 1$$

This is also known as three parameters exponential or triple exponential because of the three

smoothing parameters α, β and γ. This is a general method and a true multi-step ahead forecast.

In addition to the alpha and beta smoothing factors, a new parameter is added called gamma (g) that controls the influence on the seasonal component.

As with the trend, the seasonality may be modelled as either an additive or multiplicative process for a linear or exponential change in the seasonality.

Additive Seasonality: Triple Exponential Smoothing with a linear seasonality. Multiplicative Seasonality: Triple Exponential Smoothing with an exponential seasonality. Triple exponential smoothing is the most advanced variation of exponential smoothing and through configuration, it can also develop double and single exponential smoothing models.

This method is sometimes called Holt-Winters Exponential Smoothing, named for two contributors to the method: Charles Holt and Peter Winters.

Hyperparameters:

```
Alpha: Smoothing factor for the level.

Beta: Smoothing factor for the trend.

Gamma: Smoothing factor for the seasonality.

Trend Type: Additive or multiplicative.

Dampen Type: Additive or multiplicative.

Phi: Damping coefficient.

Seasonality Type: Additive or multiplicative.

Period: Time steps in seasonal period.
```

Initializing the Triple Exponential Smoothing Model

```
---Holt Winters model Exponential Smoothing Estimated Paramete
rs----

{'smoothing_level': 0.09467987567540882, 'smoothing_trend': 2.
31999683285252e-05, 'smoothing_seasonal': 0.000417528569192231
4, 'damping_trend': nan, 'initial_level': 146.40142527639352,
'initial_trend': -0.5464913833622084, 'initial_seasons': array
([-31.19268548, -18.83344765, -10.84745053, -21.48718886,
       -12.67654312,  -7.19154248,   2.65454402,   8.80233514,
         4.79913097,   2.91389547,  21.00157004,  63.18716583]
), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

TABLE 9 HOLT WINTER SMOOTHING MODEL PARAMETERS

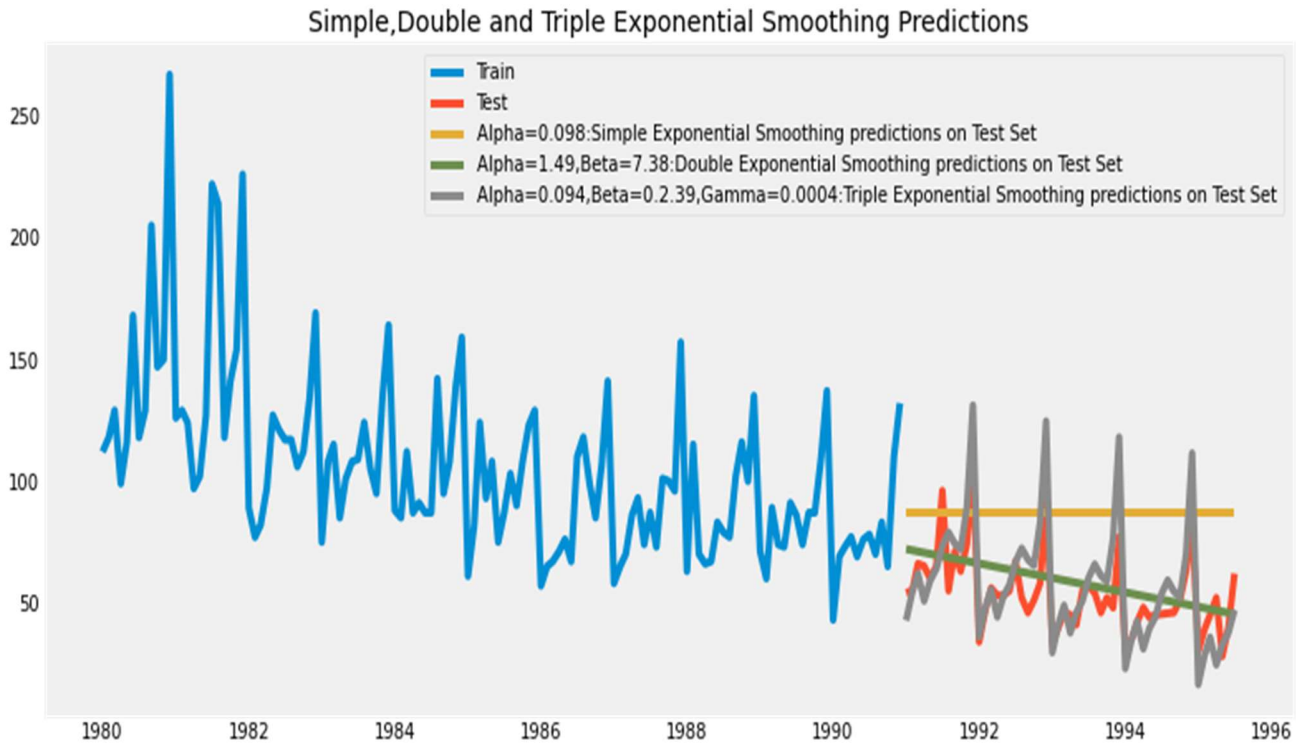


FIGURE 8 TRIPLE EXP SMOOTHING PLOT

```
TES RMSE: 14.278440376982086
```

Inferences:

1.Triple Exponential Smoothing with Lowest AIC(Alpha=0.094,Beta=2.39,Gamma=0.0004) has given the Low RMSE value as 14.27.

2.Which is comparatively better over last two models RSME score.

3.Lower the Error model better.

4.In plot the clear trend and seasonality cane be defined by this model.

RMSE Models Summary	
	Test RMSE
Alpha=0.098,SES	36.796236
Alpha=1.49,Beta=7.38:DES	15.268957
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	14.27844

TABLE 10 RSME MODEL SUMMARY

Holt-Winters - ETS(A, A, M) - Holt Winter's linear method

ETS(A, A, M) model

Initializing the Holt Winter Exponential Smoothing Model parameter Search

```
----Holt Winters model Exponential Smoothing Estimated Paramet
ers---

{'smoothing_level': 0.07130285749243212, 'smoothing_trend': 0.
04550837652110988, 'smoothing_seasonal': 8.385716703273524e-05
, 'damping_trend': nan, 'initial_level': 163.60092654560762, '
initial_trend': -0.9804841883026134, 'initial_seasons': array(
[0.68714163, 0.77936108, 0.85184662, 0.74446365, 0.8372947 ,
       0.91182237, 1.00282327, 1.06745268, 1.01025249, 0.98957
378,
       1.1535151 , 1.59037115]), 'use_boxcox': False, 'lamda':
 None, 'remove_bias': False}
```

TABLE 11 HOLT WINTER LINEAR METHOD PARAMETERS

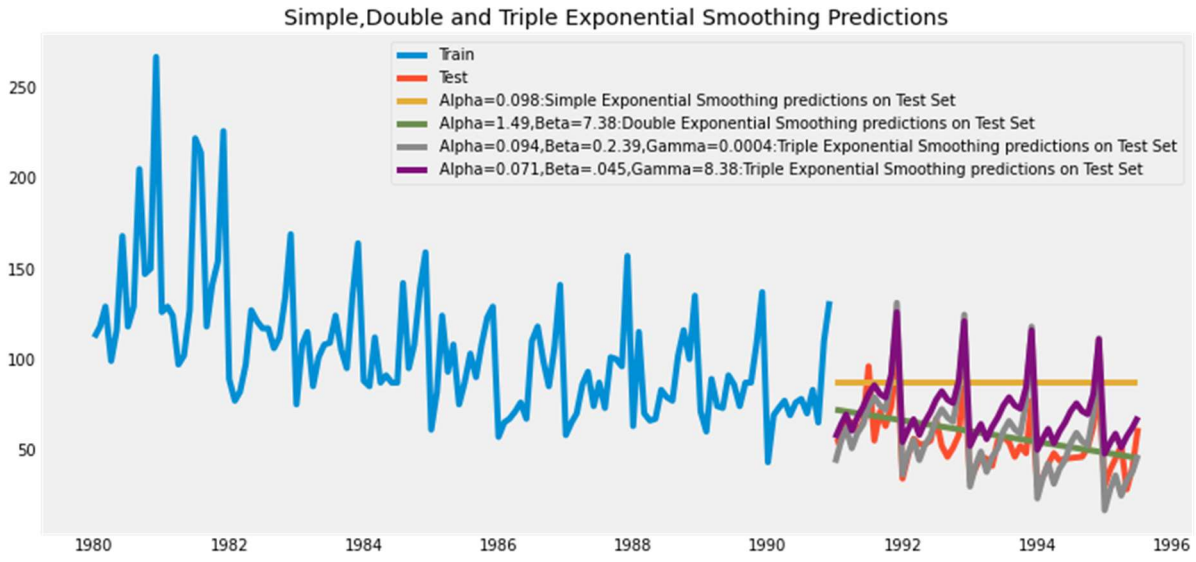


FIGURE 9 HOLT WINTER PLOT

TES_am RMSE: 20.189764216068276

Inferences:

1.Simple Exponential model with alpha value 0.098 the RMSE value is 36.79

2.Double Exponential model with Alpha=1.49,Beta=7.38 giving value of 15.26 which does not defining correctly to trend and seasonality.

3.Triple Exponential Model with Alpha=0.094,Beta=0.2.39,Gamma=0.0004 values giving RMSE of 14.27.
4.In Holt Winters model Exponential Smoothing model with Alpha=0.071,Beta=.045,Gamma=8.38 giving RMSE of 20.18\

Among all the Triple Exponential Smoothing has performed the best to 14.27.It also clear defining the value, Trend and Seasonality of TS dataset.

RMSE Models Summary	
	Test RMSE
Alpha=0.098,SES	**36.796236**
Alpha=1.49,Beta=7.38:DES	**15.268957**
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	**14.27844**
Alpha=0.071,Beta=.045,Gamma=8.38:TES	**20.189764**

TABLE 12 RMSE SUMMARY TABLE

Model 1**: Linear Regression**

A time series regression forecasts a time series as a linear relationship with the independent variables. The linear regression model assumes there is a linear relationship between the forecast variable and the predictor variables

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

```
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    43
1991-02-01  55.0    44
1991-03-01  66.0    45
```

```
1991-04-01   65.0     46
1991-05-01   60.0     47

Last few rows of Test Data
             Rose  time
YearMonth
1995-03-01   45.0     93
1995-04-01   52.0     94
1995-05-01   28.0     95
1995-06-01   40.0     96
1995-07-01   62.0     97
```

TABLE 13 LR TRAIN TEST DATA

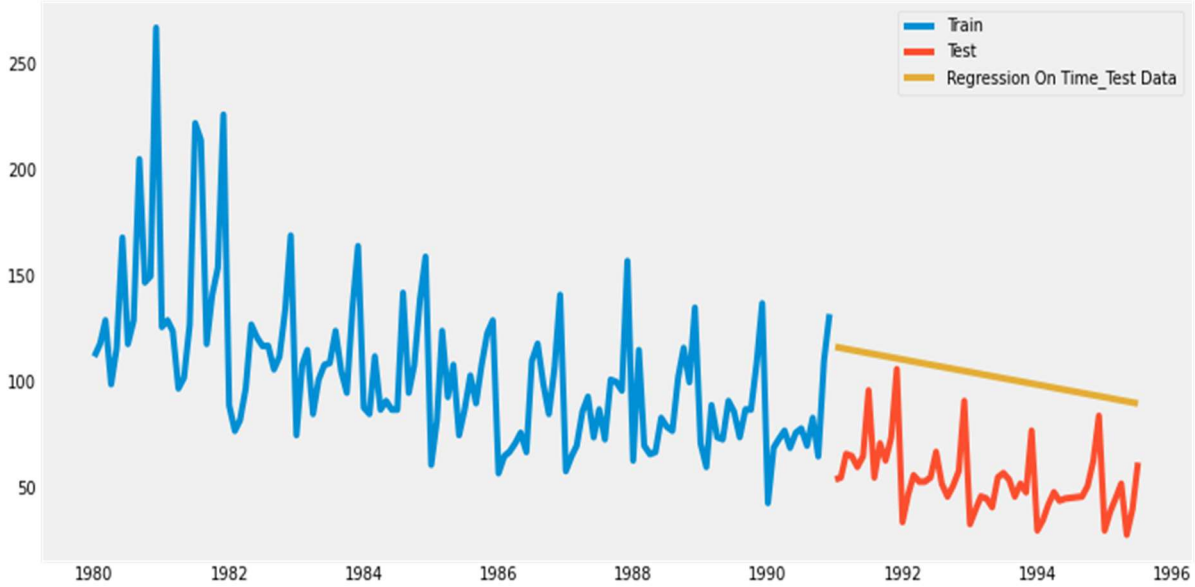


FIGURE 10 LINEAR REGRESSION PLOT

For RegressionOnTime forecast on the Test Data, RMSE is 51.433

Inference: The Regression Model having the RMSE value 51.43 which are higher among all till now.

RMSE Models Summary	
	Test RMSE
Alpha=0.098,SES	**36.796236**
Alpha=1.49,Beta=7.38:DES	**15.268957**
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	**14.27844**
Alpha=0.071,Beta=.045,Gamma=8.38:TES	**20.189764**
Regression On Time	**51.433312**

TABLE 14 RSME MODELS SUMMARY

Model 2: Naive Approach: y^t+1=yt

The Naive Bayes method is a classification algorithm that uses Bayes' theorem to predict the probability of a class given a set of features. However, there is a method called Naive Method which uses the most recent value as the forecasted value for the next time step. The assumption followed by this method is that its value tomorrow is equal to its value today

Naïve Model Test Data head
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

Name: naive, dtype: float64

TABLE 15 TEST DATA NAIVE MODEL

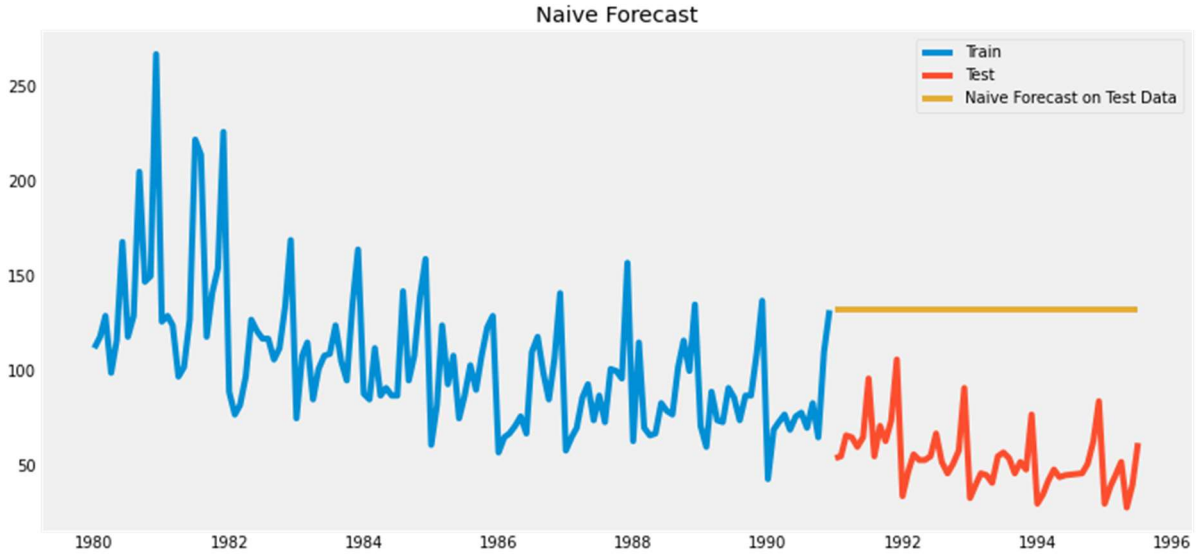


FIGURE 11 NAIVE BAYES METHOD PLOT

For Naïve Model forecast on the Test Data, RMSE is 79.719

Inferences:
The Naive model uses the last value for forecasting hence the RSME value are too hight to 7 9.719

RMSE Summary of Models

RMSE Models Summary	
	Test RMSE
Alpha=0.098,SES	**36.796236**
Alpha=1.49,Beta=7.38:DES	**15.268957**
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	**14.27844**
Alpha=0.071,Beta=.045,Gamma=8.38:TES	**20.189764**
Regression On Time	**51.433312**
Naive Model	**79.718773**

TABLE 16 RMSE MODEL SUMMARY

Simple Average Model

The method is very simple: average the data by months or quarters or years and them calculate the average for the period. Then find out, what percentage it is to the grand average.
Top 5 Test Dataset of SAM

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 17 SIMPLE AVERAGE

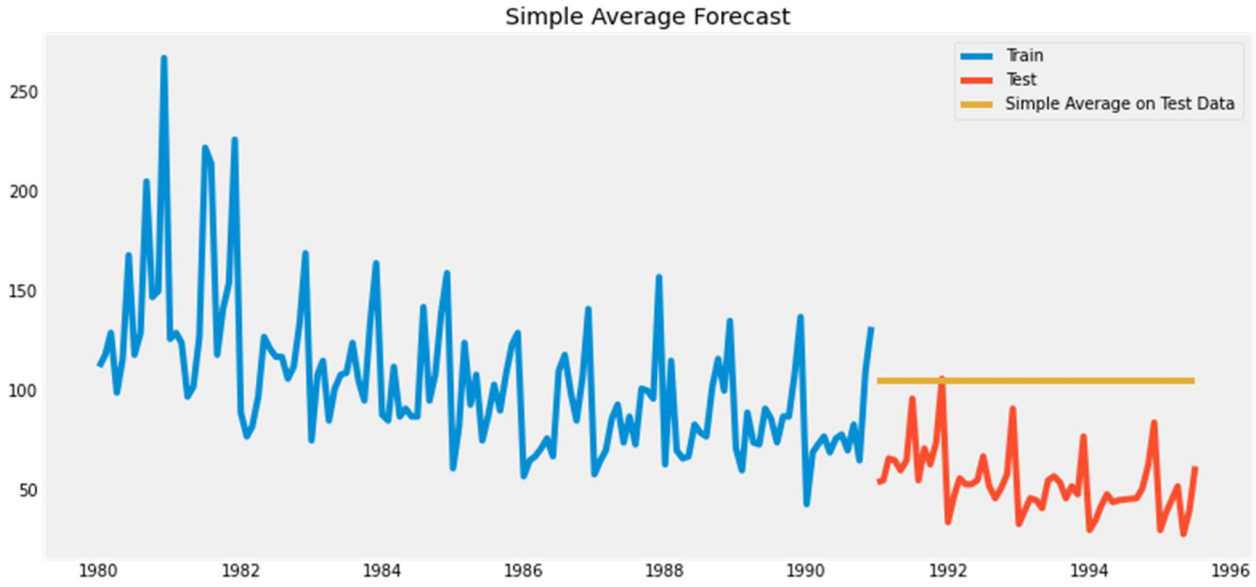


FIGURE 12 SIMPLE AVERAGE PLOT

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

RMSE Summary of Models

RMSE Models Summary	
	Test RMSE
Alpha=0.098,SES	**36.796236**
Alpha=1.49,Beta=7.38:DES	**15.268957**
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	**14.27844**
Alpha=0.071,Beta=.045,Gamma=8.38:TES	**20.189764**
Regression On Time	**51.433312**
Naive Model	**79.718773**
Simple Average	**53.46057**

TABLE 18 RMSE MODELS SUMMARY

Moving Average Forecast

The moving average is a statistical method used for forecasting long-term trends. The technique represents taking an average of a set of numbers in a given range while moving the range. In simple terms, a moving average plot takes the average of several different points in the data set and then plots it over time.

Two main types of moving averages:

1) Centred moving average - calculated as the average of raw observations at, before and after time, t.
2) Trailing moving average - uses historical observations and is used on time series forecasting.

The rolling () function on the Series Pandas object will automatically group observations into a window.

The main advantage of the moving average method is that it takes into account all previous values when predicting future values. This helps to reduce the effect of outliers when making predictions and also makes it easier to identify seasonal patterns in a time-series data set. The moving average method is an effective tool for short-term forecasting due to its flexibility and ease of use. Its ability to take into account all past values when making predictions ensures accuracy while its ability to identify seasonal patterns means that it can be used effectively for long-term forecasting too

This algorithm helps us to forecast new observations based on a time series. This algorithm uses smoothing methods. The moving average algorithm is used only on time series that DOESN'T have a trend. This method is by far the easiest and it consists of making the arithmetic mean of the last ,,n'' observations contained by the time series to forecast the next observation. We use the following

$$formula: MA\_(t+1)=(\sum\_(i=t-n)^t x\_i)/n$$

We need to find the optimal number,, n" of observation to be used in the forecast. We can find it by checking the square error mean of multiple ,,n'' observations. We should start at 3 observations and we can go up to half of the data set size + 1.

Moving Average train Data

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

TABLE 19 MOVING AVERAGE TEST DATA SET

YearMonth	Rose	moving_avg_forecast
1991-01-01	54.000000	79.666667
1991-02-01	55.000000	78.500000
1991-03-01	66.000000	77.916667
1991-04-01	65.000000	76.916667
1991-05-01	60.000000	76.166667
1991-06-01	65.000000	75.250000
1991-07-01	96.000000	76.750000
1991-08-01	55.000000	75.500000
1991-09-01	71.000000	74.500000
1991-10-01	63.000000	74.333333
1991-11-01	74.000000	71.333333
1991-12-01	106.000000	69.166667
1992-01-01	34.000000	67.500000
1992-02-01	47.000000	66.833333
1992-03-01	56.000000	66.000000
1992-04-01	53.000000	65.000000
1992-05-01	53.000000	64.416667
1992-06-01	55.000000	63.583333
1992-07-01	67.000000	61.166667
1992-08-01	52.000000	60.916667
1992-09-01	46.000000	58.833333
1992-10-01	51.000000	57.833333

1992-11-01	58.000000	56.500000
1992-12-01	91.000000	55.250000
1993-01-01	33.000000	55.166667
1993-02-01	40.000000	54.583333
1993-03-01	46.000000	53.750000
1993-04-01	45.000000	53.083333
1993-05-01	41.000000	52.083333
1993-06-01	55.000000	52.083333
1993-07-01	57.000000	51.250000
1993-08-01	54.000000	51.416667
1993-09-01	46.000000	51.416667
1993-10-01	52.000000	51.500000
1993-11-01	48.000000	50.666667
1993-12-01	77.000000	49.500000
1994-01-01	30.000000	49.250000
1994-02-01	35.000000	48.833333
1994-03-01	42.000000	48.500000
1994-04-01	48.000000	48.750000
1994-05-01	44.000000	49.000000
1994-06-01	45.000000	48.166667
1994-07-01	45.333333	47.194444
1994-08-01	45.666667	46.500000
1994-09-01	46.000000	46.500000
1994-10-01	51.000000	46.416667
1994-11-01	63.000000	47.666667
1994-12-01	84.000000	48.250000
1995-01-01	30.000000	48.250000
1995-02-01	39.000000	48.583333
1995-03-01	45.000000	48.833333
1995-04-01	52.000000	49.166667
1995-05-01	28.000000	47.833333
1995-06-01	40.000000	47.416667
1995-07-01	62.000000	48.805556

TABLE 20 MOVING AVERAGE FORECAST

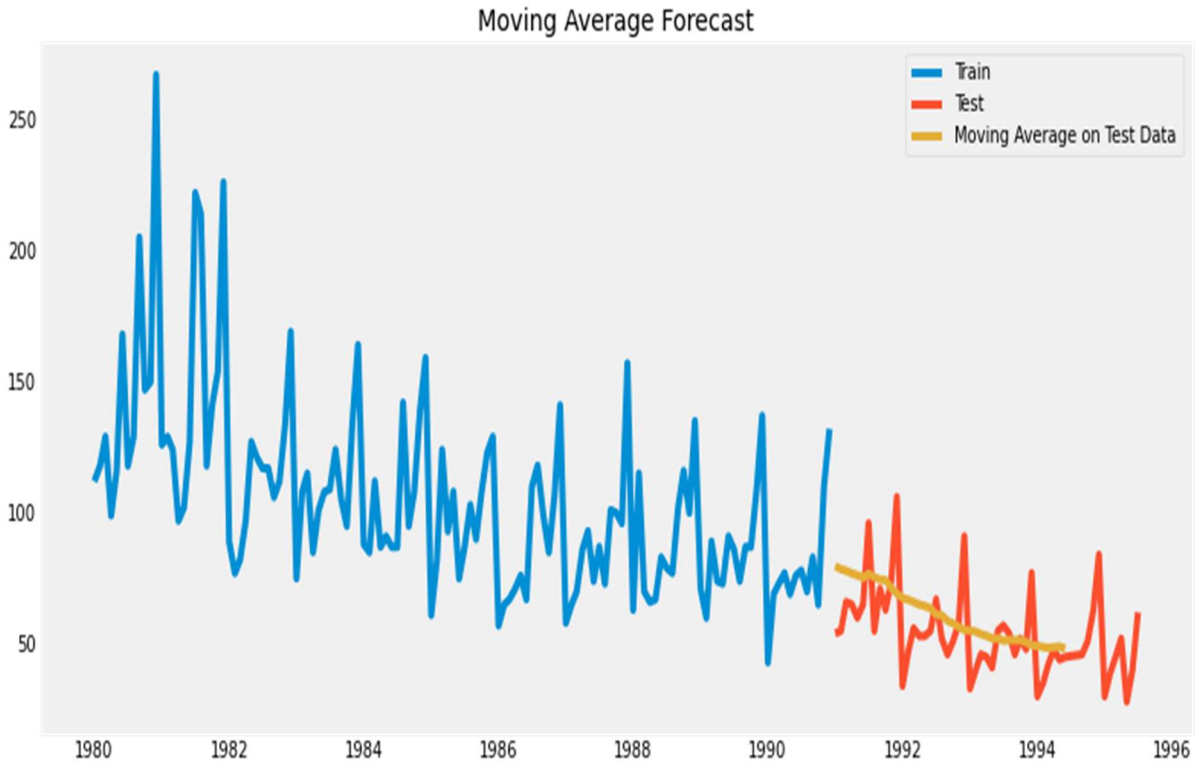


FIGURE 13 MOVING AVERAGE PLOT

For Moving Average forecast on the Test Data, RMSE is 53.461

RMSE Summary of Models

RMSE Models Summary	
	Test RMSE
Alpha=0.098,SES	36.796236
Alpha=1.49,Beta=7.38:DES	15.268957
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	14.27844
Alpha=0.071,Beta=.045,Gamma=8.38:TES	20.189764
Regression On Time	51.433312
Naive Model	79.718773
Simple Average	53.46057
Moving Average	15.236052

TABLE 21 RMSE SUMMARY TABLE

Trailing Moving Average

	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
YearMonth					
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	114.50	NaN	NaN
1980-05-01	116.0	107.5	115.50	NaN	NaN
...	...	...	...	...	...
1995-03-01	45.0	42.0	49.50	52.000000	NaN
1995-04-01	52.0	48.5	41.50	52.166667	NaN
1995-05-01	28.0	40.0	41.00	46.333333	48.666667
1995-06-01	40.0	34.0	41.25	39.000000	48.000000
1995-07-01	62.0	51.0	45.50	44.333333	49.222222

TABLE 22 TRAILING MOVING AVERAGE

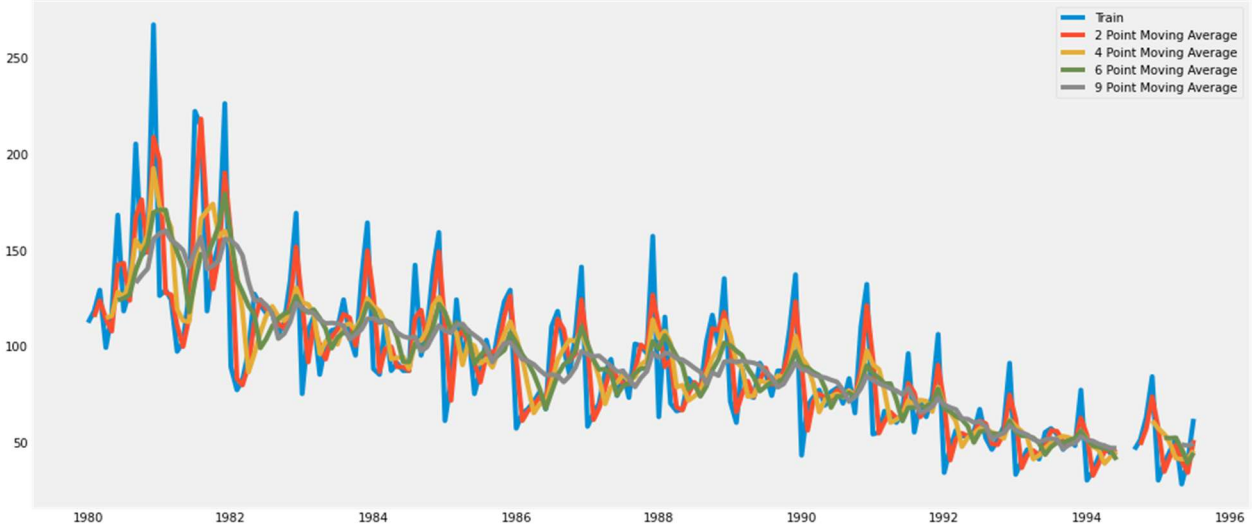


FIGURE 14 TRAILING MOVING AVERAGE FORECAST PLOT

```
Moving Average of train 5166.5
              and test  1859.5
```

Last five Data of Trailing Moving average

```
YearMonth
1990-08-01     74.0
1990-09-01     76.5
1990-10-01     74.0
1990-11-01     87.5
1990-12-01    121.0
Name: Trailing_2, dtype: float64
```

TABLE 23 TRAILING MOVING AVERAGE LAST FIVE DATA

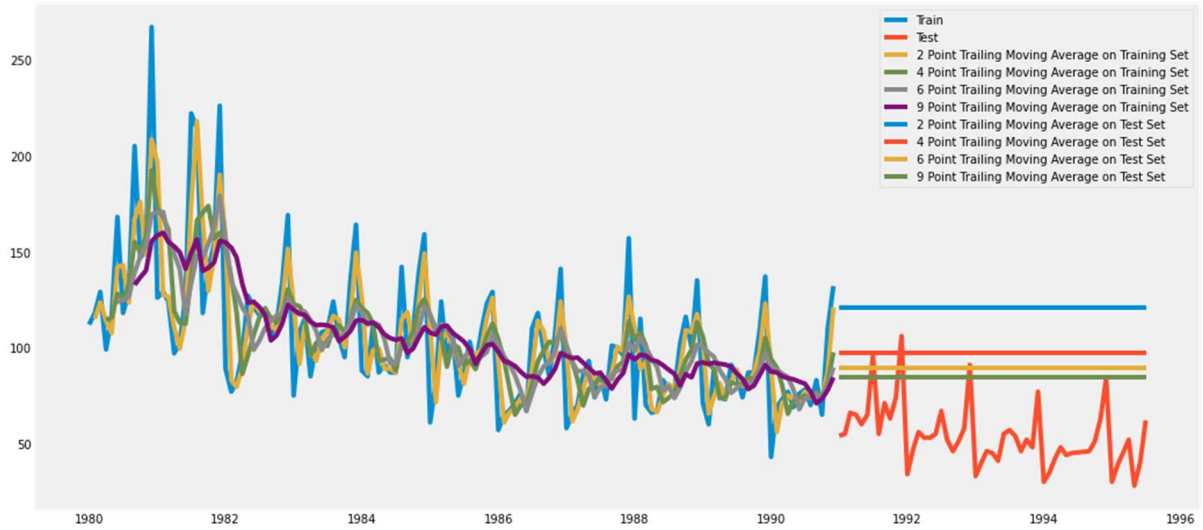


FIGURE 15 TRAILING MA ON POINTS PLOT

The Tabulating the RMSE Values of Trailing Average point wise forecast models

For 2 point Moving Average Model forecast on the Test Data, RMSE is 68.970
For 4 point Moving Average Model forecast on the Test Data, RMSE is 46.404
For 6 point Moving Average Model forecast on the Test Data, RMSE is 39.126
For 9 point Moving Average Model forecast on the Test Data, RMSE is 34.411

TABLE 24 RMSE TRAILING AVERAGE POINT WISE PLOT

Inference:

1.The RMSE values are above 34 in all label of data but among all trailing average the value of RMSE is Lowest for 9th point 34.41.
2. Which still higher than the other TES models.

RMSE Summary of Models

	Test RMSE
ALPHA=0.70,SES	1338.000861
ALPHA=.665, BETA=0.0001:DES	5291.879833
ALPHA=0.11, BETA=0.012,GAMMA=0.460:TES	378.625883
ALPHA=0.111, BETA=.0493, GAMMA=.362: TES	402.936179
REGRESSIONONTIME	1275.867052
Naive Model	3864.279352
Simple Average	1275.081804
Moving Average	1267.92533
2pointTrailingMovingAverage	3046.976092
4pointTrailingMovingAverage	2021.85588
6pointTrailingMovingAverage	1521.61125
9pointTrailingMovingAverage	1304.618912

TABLE 25 RMSE SUMMARY

QN5.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.

Checking for stationarity of the whole Time Series data.

Stationarity

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time.

Stationarity means that the autocorrelation of lag 'k' depends on k, but not on time t.

Let Xt denote the time series at time t.

Autocorrelation of lag k is the correlation between Xt and X(t−k)

Strong stationarity: is a stochastic process whose unconditional joint probability distribution does not change when shifted in time. Consequently, parameters such as mean and variance also do not change over time.
Weak stationarity: is a process where mean, variance, autocorrelation are constant throughout the time Stationarity is important as non-stationary series that depend on time have too many parameters to account for when modelling the time series. diff() method can easily convert a non-stationary series to a stationary series.

We will try to decompose seasonal component of the above decomposed time series.

Check for Stationarity

There are multiple tests that can be used to check stationarity.

ADF(Augmented Dicky Fuller Test)

KPSS

PP (Phillips-Perron test)

Let's just perform the ADF which is the most commonly used one

Dickey-Fuller Test - Dicky Fuller Test on the timeseries is run to check for stationarity of data.

Hypothesis

Null Hypothesis H0: Time Series is non-stationary.
Alternate Hypothesis Ha: Time Series is stationary.

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

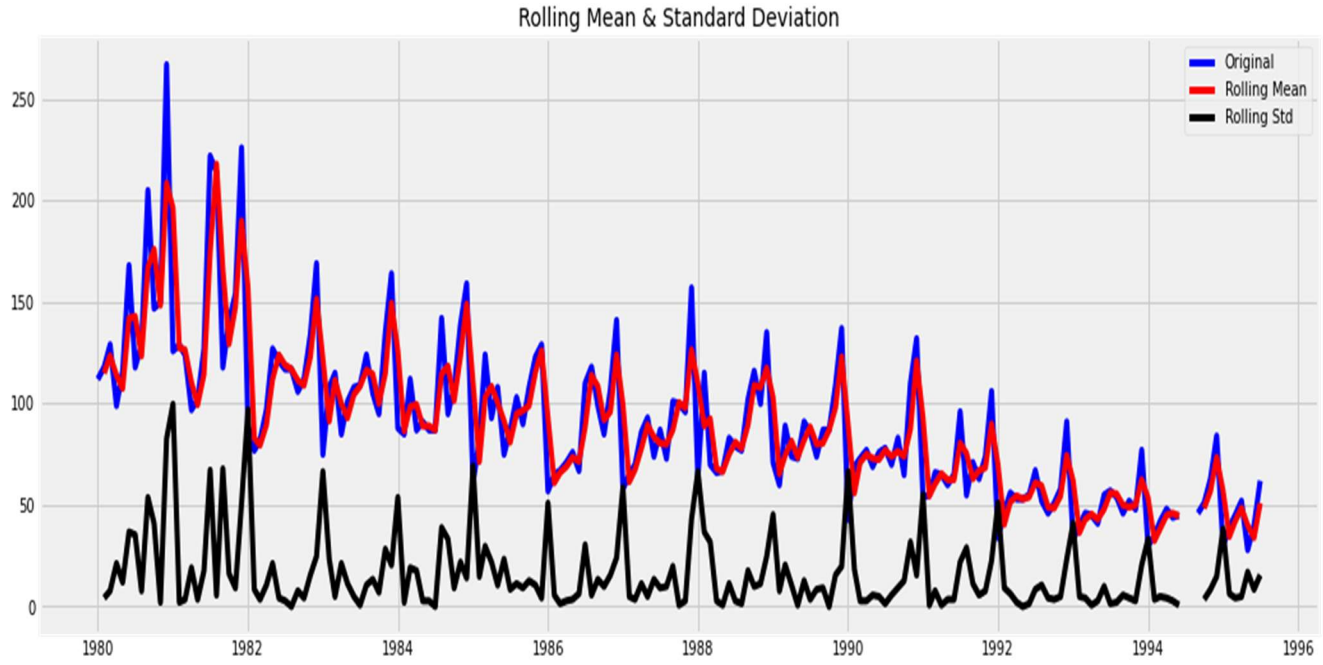


FIGURE 16 DATA STATIONARY PLOT

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#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
dtype: float64
```

TABLE 26 DICKEY FULLER TEST

Insight:

In this case, since the p-value is 0.343101 which is greater than 0.05, we cannot reject the null hypothesis that the time series has a unit root and conclude that it is non-stationary.

Making a Time Series Stationary by - Differencing 'd'

Differencing 'd' is done on a non-stationary time series data one or more times to convert it into stationary. (d=1) 1st order differencing is done where the difference between the current and previous (1 lag before) series is taken and then checked for stationarity using the ADF(Augmented Dicky Fueller) test. If differenced time series is stationary, we proceed with AR modelling. Else we do (d=2) 2nd order differencing, and this process repeats till we get a stationary time series

1st order differencing equation is : $y_t = y_t - y_{t-1}$
2nd order differencing equation is : $y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$ and so on…
The variance of a time series may also not be the same over time. To remove this kind of non-stationarity, we can transform the data. If the variance is increasing over time, then a log transformation can stabilize the variance.
Let us take a difference of order 1 and check whether the Time Series is stationary or not.

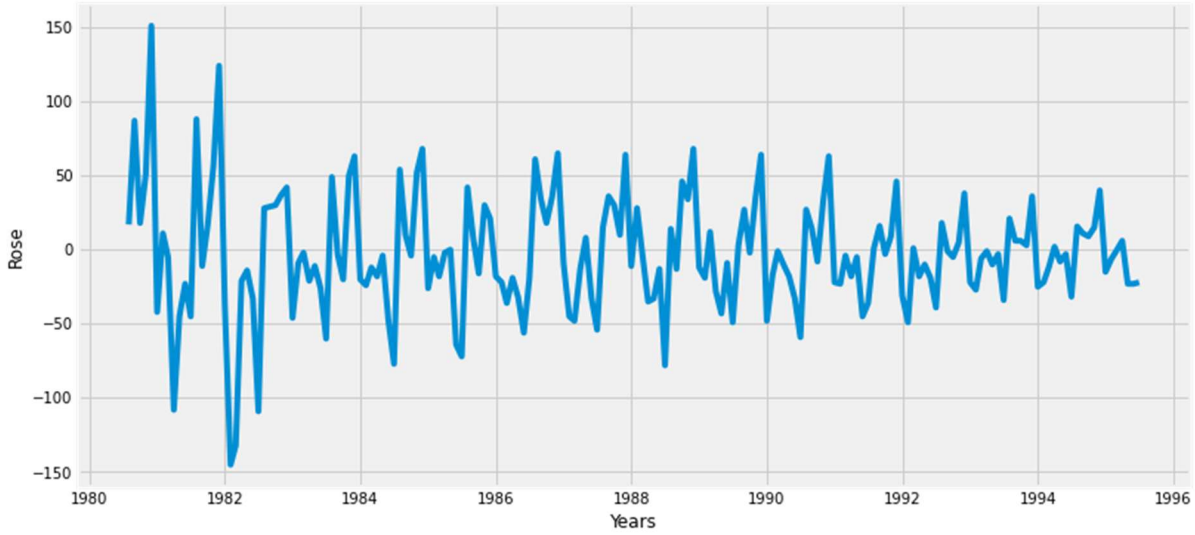


FIGURE 17 DICKEY FULLER TEST PLOT

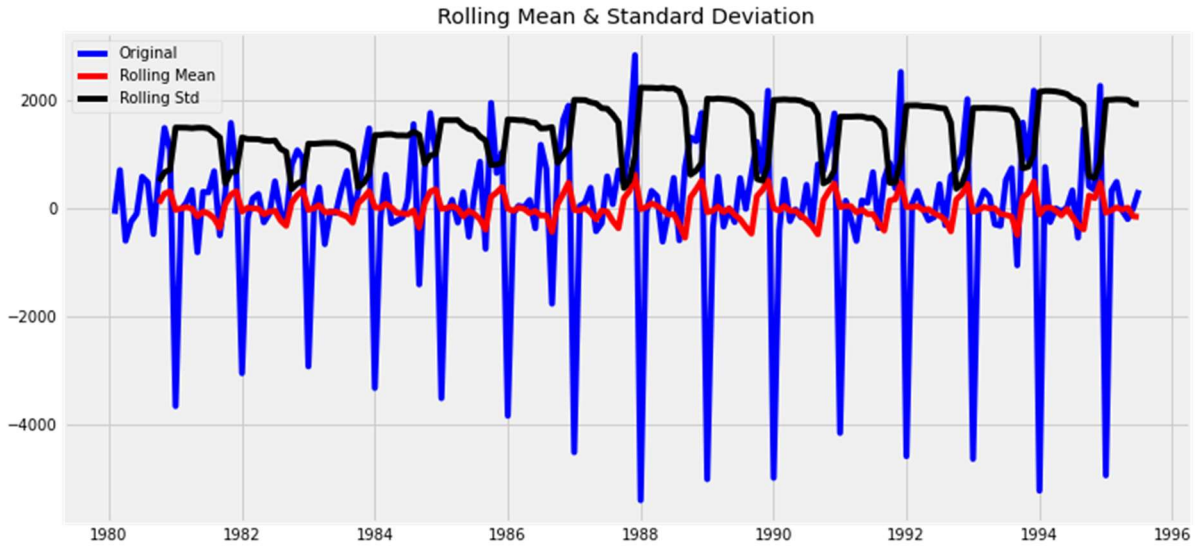


FIGURE 18 AFTER DICKEY FULLER CORRECTION PLOT

```
Results of Dickey-Fuller Test:
Test Statistic                  -8.044392e+00
p-value                          1.810895e-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 27 AFTER DICKEY FULLER CORRECTION

Insight : Now, We can see that at α= 0.05 the Time Series is indeed stationary.

Plot the Autocorrelation function plots on the whole data.

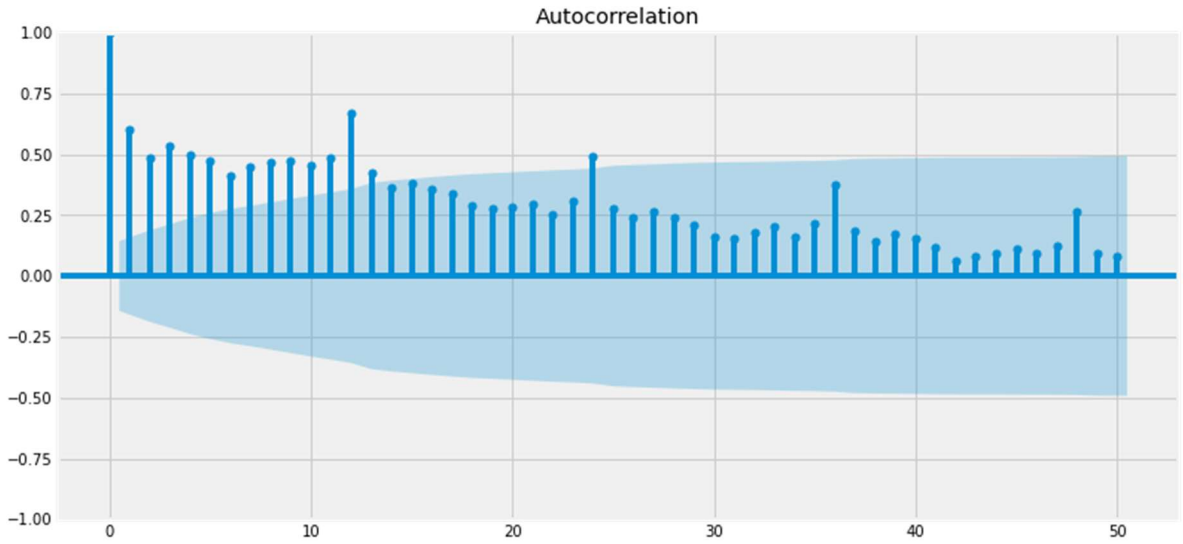


FIGURE 19 AUTOCORRELATION

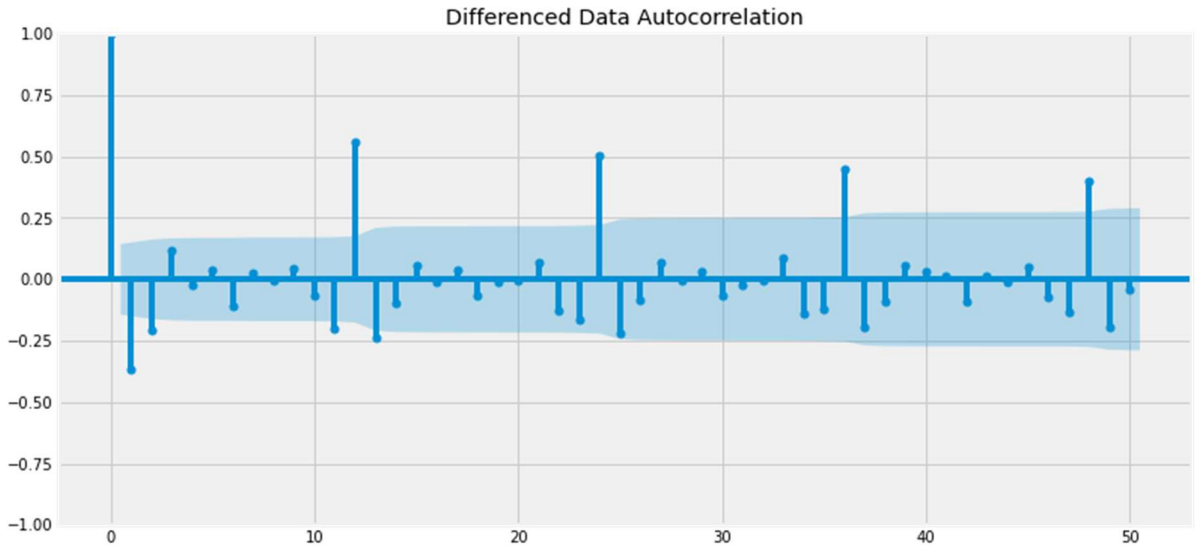


FIGURE 20 DIFFERENCE IN AUTOCORRELATION

QN 6 Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Auto Regressive (AR) Models

Autoregression means regression of a variable on itself which means Autoregressive models use previous time period values to predict the current time period values.

An autoregressive (AR) model is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc

The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term); thus the model is in the form of a stochastic difference equation.

One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. An AR(p) model (Auto-Regressive model of order p) can be written as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots\ldots + \phi_p y_{t-p} + \varepsilon_t$$

ε_t is an error term which is an independent and identically distributed random variable (or in other words, a white noise) with the parameters mean = 0 and standard deviation = σ

The ϕ are regression coefficients multiplied by lagged time series variable, which captures the effect of the input variable on the output, provided intermediate values do not change.

Choose the order 'p' of AR model

We look at the Partial Autocorrelations of a stationary Time Series to understand the order of Auto-Regressive models. For an AR model, 2 ways to identify order of 'p':

1) PACF Approach: the PACF method where the (Partial Auto Correlation Function) values cut off and become zero after a certain lag. PACF vanishes if there is no regression coefficient that far back. The cut-off value where this happens can be taken as the order of AR as ‘p’. This can be seen from a PACF plot.

If the 2nd PACF vanishes (cut off in PACF) then the 2nd coefficient is not considered and thus ‘p’ is 1.

If the 3rd PACF vanishes (cuts off in PACF) then the 3rd coefficient is not considered and thus ‘p’ is 2 and so on…

Partial Autocorrelation of order 2 = Partial autocorrelation of lag 2 = Correlation between Xt and Xt−2 holding Xt−1 fixed.

2) Lowest AIC Approach: the lowest Akaike Information Criteria (AIC) value compared among different orders of ‘p’ is considered.

Moving Average (MA) Models

Moving average model considers past residual values to predict the current time period values. These past residuals are past prediction errors. For a MA model, the residual or error component is modelled. The moving average model MA(q) of order q can be represented as:

$$y_t=\varepsilon_t+\theta_1\varepsilon_{t-1}+\ldots\ldots+\theta_q\varepsilon_{t-q}$$

Where y_t time series variable, θ are numeric coefficients multiplied to lagged residuals and ε is the residual term considered as a purely random process with mean 0, variance σ^2 and $Cov(\varepsilon_{t-1},\varepsilon_{t-q})= 0$.

Choose the order 'q' of MA model

We look at the Autocorrelations of a stationary Time Series to understand the order of Moving Average models. For a MA model,

1) ACF Approach: ACF (Autocorrelation Function) values cut off at a certain lag. ACF vanishes, and there are no coefficients that far back; thus, the cut-off value where this happens is taken as the order of MA as ‘q’. This can be seen from the ACF plot.

2) Lowest AIC Approach: the lowest Akaike Information Criteria (AIC) value compared among different orders of ‘q’ is considered

ARIMA model

Auto Regressive Integrated Moving Average is a way of modelling time series data for forecasting or predicting future data point An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity. ARIMA model is of the form: ARIMA(p,d,q): p is AR parameter, d is differential parameter, q is MA parameter

ARIMA(1,0,0)

$$yt = a1yt-1 + \epsilon t$$

ARIMA(1,0,1)

$$yt = a1yt-1 + \epsilon t + b1\epsilon t-1$$

ARIMA(1,1,1)

$$\Delta yt = a1\Delta yt-1 + \epsilon t + b1\epsilon t-1 \text{ where } \Delta yt = yt - yt-1$$

Improving AR Models by making Time Series stationary through Moving AVerage Forecasts
ARIMA models consist of 3 components: -
AR model: The data is modelled based on past observations.
Integrated component: Whether the data needs to be differenced/transformed.
MA model: Previous forecast errors are incorporated into the model.

Building an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC

Getting a combination of different parameters of p and q in the range of 0 and 2, We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
```

TABLE 28 ARIMA BEST COMBINATION

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748333
ARIMA(0, 1, 2) - AIC:1279.671528853574
ARIMA(1, 1, 0) - AIC:1317.3503105381546
ARIMA(1, 1, 1) - AIC:1280.5742295380066
ARIMA(1, 1, 2) - AIC:1279.870723423191
ARIMA(2, 1, 0) - AIC:1298.6110341604885
ARIMA(2, 1, 1) - AIC:1281.5078621868474
ARIMA(2, 1, 2) - AIC:1281.8707222264484
```

TABLE 29 AIC VALUE WITH COMBINATIONS

AIC in Descending Order for get the AIC value minimum.

param	AIC
2 (0, 1, 2)	1279.671529
5 (1, 1, 2)	1279.870723
4 (1, 1, 1)	1280.574230
7 (2, 1, 1)	1281.507862
8 (2, 1, 2)	1281.870722
1 (0, 1, 1)	1282.309832
6 (2, 1, 0)	1298.611034
3 (1, 1, 0)	1317.350311
0 (0, 1, 0)	1333.154673

TABLE 30 BEST PARAMETERS

```
                SARIMAX Results
==============================================================================
=======
Dep. Variable:          Rose   No. Observations:                  132
Model:        ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date         Sun, 09 Apr 2023   AIC                           1279.672
Time:                16:29:29   BIC                           1288.297
Sample:            01-01-1980   HQIC                          1283.176
                 - 12-01-1990
Covariance Type:           opg
==============================================================================
=======
                 coef    std err          z      P>|z|      [0.025
 0.975]
------------------------------------------------------------------------------
-------
ma.L1         -0.6970      0.072     -9.689      0.000      -0.838
 -0.556
ma.L2         -0.2042      0.073     -2.794      0.005      -0.347
 -0.061
sigma2       965.8407     88.305     10.938      0.000     792.766    1
138.915
==============================================================================
============
Ljung-Box (L1) (Q):    4                 Jarque-Bera (JB):     39.24
Prob(Q):               0.71                Prob(JB):           0.00
Heteroskedasticity (H):0.36                Skew:               0.82
Prob(H) (two-sided):   .00                 Kurtosis:           5.13
==============================================================================
============

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

TABLE 31 ARIMA MODEL RESULTS

```
RMSE for the ARIMA  Model is   37.30647972368469
```

Inferences:

Criteria to choose the best fit model is the lowest/minimum AIC value For ARIMA (p, d, q) we got (2, 1, 2),model with the least AIC of 1279.671529

Here,

p = non-seasonal AR order = 0,
d = non-seasonal differencing = 1,
q = non-seasonal MA order = 2,
S = time span of repeating seasonal pattern = 12

The ARIMA model with order 0,1,2 has AIC of 1279.671529

which is pretty higher than few models.

RMSE Summary of Models

Summary of RMSE of All Models	
	Test RMSE
Alpha=0.098,SES	36.796236
Alpha=1.49,Beta=7.38:DES	15.268957
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	14.27844
Alpha=0.071,Beta=.045,Gamma=8.38:TES	20.189764
Regression On Time	51.433312
Naive Model	79.718773
Simple Average	53.46057
Moving Average	15.236052
2pointTrailingMovingAverage	68.970159
4pointTrailingMovingAverage	46.403626
6pointTrailingMovingAverage	39.126446
9pointTrailingMovingAverage	34.410938
ARIMA(2,1,2)	37.30648

TABLE 32 RMSE SUMMARY OF MODELS

Building an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC)

SAMRIMA Model

SARIMA stands for Seasonal Autoregressive Integrated Moving Average. It is a statistical analysis model that uses time-series data to either better understand the data set or to predict future trends. SARIMA is an extension of ARIMA (Autoregressive Integrated Moving Average) model and is used when the time series data has seasonal frequency yet also by non-seasonal differencing in Univariate data. It adds three new hyperparameters to specify

the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

A typical SARIMA model equation looks like the following –

SARIMA(p,d,q)x(P,D,Q)lag

The parameters for these types of models are as follows:

p and seasonal P: indicate the number of AR terms (lags of the stationary series)

d and seasonal D: indicate differencing that must be done to stationary series

q and seasonal Q: indicate the number of MA terms (lags of the forecast errors)

lag: indicates the seasonal length in the data

Seasonality S(P, D, Q, s), where s is simply the season's length. This component requires the parameters P and Q which are the same as p and q, but for the seasonal component. Finally, D is the order of seasonal integration representing the number of differences required to remove seasonality from the series.

Combining all, we get the SARIMA (p, d, q)(P, D, Q, s) model.
Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model.

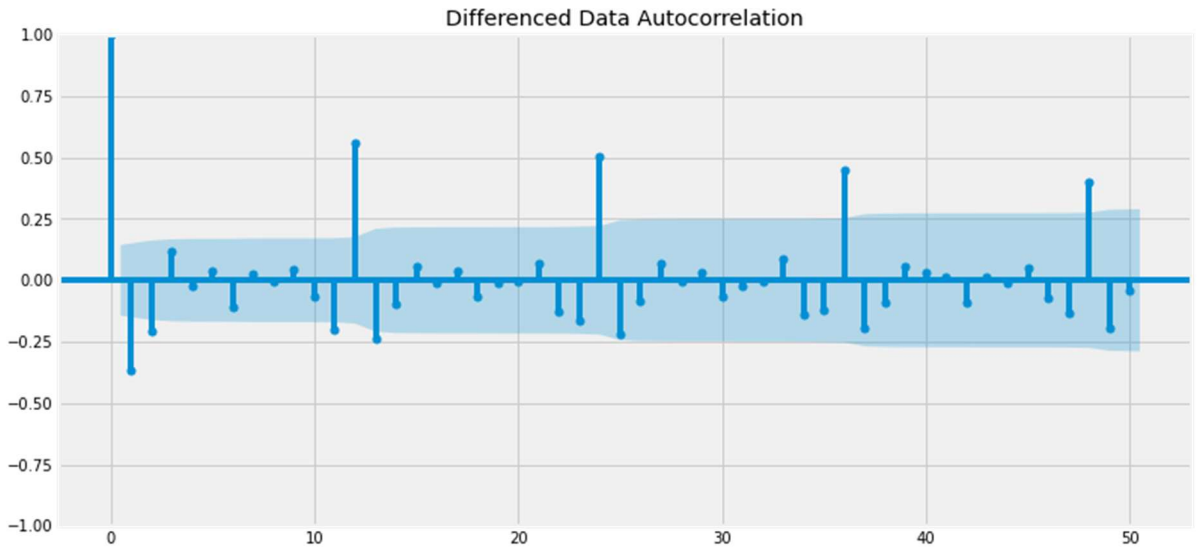


FIGURE 21 SEASONAL PARAMETER OF SARIMA

INSIGHT:

We see that there can be a seasonality at 12. But from the decomposition at the start we ascertained that visually it looks like the seasonality =12 and thus using the same

Setting the seasonality as 12 to estimate parameters using auto SARIMA model

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

TABLE 33 PARAMETER COMBINATIONS

Getting the lowest AIC by The best combination using param by a loop method.

The best 5 are as follows:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.871767
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

TABLE 34 BEST AIC FOR SARIMA

```
SARIMAX Results
==========================================================================
==================
Dep. Variable:  y.                                Observations:       132
Model:SARIMAX(0, 1, 2)x(2, 0, 2, 12)        Log Likelihood    -436.969
Date:     Sun, 09 Apr 2023                        AIC          887.938
Time:16:14:46                                     BIC          906.448
Sample:  0                                        HQIC     895.437- 132

Covariance Type: opg
==========================================================================
=======
                 coef    std err          z      P>|z|      [0.025
 0.975]
--------------------------------------------------------------------------
-------
ma.L1         -0.8427    190.063     -0.004      0.996    -373.359
371.673
ma.L2         -0.1573     29.860     -0.005      0.996     -58.682
 58.367
ar.S.L12       0.3467      0.079      4.375      0.000       0.191
  0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154
  0.451
```

```
ma.S.L12       0.0767      0.133       0.577       0.564       -0.184
  0.337
ma.S.L24      -0.0726      0.146      -0.498       0.618       -0.358
  0.213
sigma2       251.3137   4.78e+04       0.005       0.996   -9.34e+04     9
.39e+04
==============================================================================
============
Ljung-Box (L1) (Q): 0.10                   Jarque-Bera (JB): 2.33
Prob(Q): 0.75                                   Prob(JB): 0.31
Heteroskedasticity (H) 0.88                        Skew: 0.37
Prob(H) (two-sided): 0.70                       Kurtosis: 3.03
==============================================================================
============

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

TABLE 35 SARIMA MODEL SUMMARY

Plotting the assumption of Models

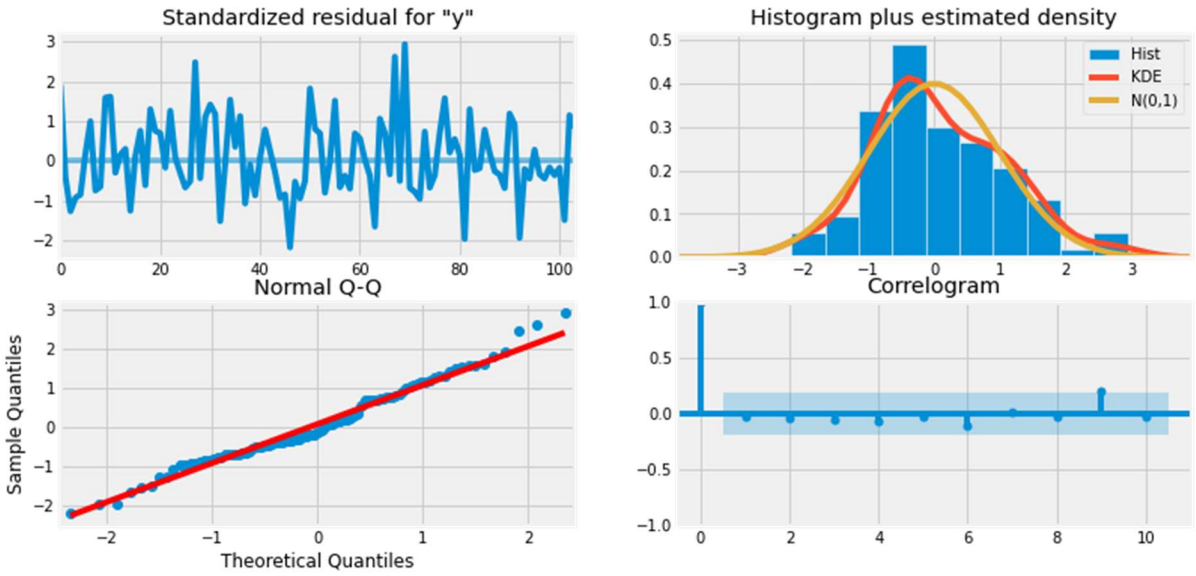


FIGURE 22 ASSUMPTION OF MODELS

Y	MEAN	MEAN_SEMEAN	_CI_LOWER	MEAN_CI_UPPER
0	62.867265	15.928501	31.647976	94.086554
1	70.541190	16.147659	38.892360	102.190021
2	77.356411	16.147657	45.707585	109.005237
3	76.208814	16.147657	44.559988	107.857640
4	72.747398	16.147657	41.098572	104.396224

TABLE 36 CI OF SARIMA MODEL

The RMSE of SARIMA model with lowest AIC is 26.928362237820917

Inferences:

Criteria to choose the best fit model is the lowest/minimum AIC value For ARIMA(p, d, q) × (P, D, Q)S, we got SARIMAX(0, 1, 2)x(2, 0, 2, 12)model with the least AIC of 887.937509

Here,

```
p = non-seasonal AR order = 0,

d = non-seasonal differencing = 1,

q = non-seasonal MA order = 2,

P = seasonal AR order = 2,

D = seasonal differencing = 0,

Q = seasonal MA order = 2,

S = time span of repeating seasonal pattern = 12
```

Insights:

In this case, our model diagnostics suggests that the model residuals are normally distributed based on the following:

1.The KDE plot of the residuals on the top right is almost similar with the normal distribution.
2.The Q-Q-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0, 1). Again, this is a strong indication that the residuals are normally distributed.
3.The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.
4.Those observations coupled with the fact that there are no spikes outside the insignificant zone for both ACF and PACF plots lead us to conclude that that residuals are random with no information or juice in them and our model produces a satisfactory fit that could help us understand our time series data and forecast future values.

It seems that our SARIMA model is working fine.

From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.

Summary of RMSE of All Models	
	Test RMSE
Alpha=0.098,SES	36.796236
Alpha=1.49,Beta=7.38:DES	15.268957
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	14.27844
Alpha=0.071,Beta=.045,Gamma=8.38:TES	20.189764
Regression On Time	51.433312
Naive Model	79.718773
Simple Average	53.46057
Moving Average	15.236052
2pointTrailingMovingAverage	68.970159
4pointTrailingMovingAverage	46.403626
6pointTrailingMovingAverage	39.126446
9pointTrailingMovingAverage	34.410938
ARIMA(2,1,2)	37.30648
SARIMA(0,1,2)(2,0,2,12)	26.928362

TABLE 37 RMSE MODEL SUMMARY

Evaluate the model on the whole and predict 12 months into the future

Time Series Assumptions Table of Contents

Some of the most common assumptions made for time series are based on the common sense. But always Keep in mind one thing

1.Forecast is done by keeping in mind that the market and the other conditions are not going to change in the future.
2.There will be not any change in the market.
3.But the change is gradual and not a drastic change.
4.Situations like recession in 2008 US market will send the forecasts into a tizzy.
5.Events like demonetization would throw the forecasts into disarray Based on the data available ,
we should not try to forecast for more than a few periods ahead.

Building the most optimum model on the Full Data.

We have used the number of methods to get the model having minimum RMSE.In This on the basis of RSME table we find **Triple Exponential Smoothening Modelling** is perfect for this solution with the Triple Exponential Smoothing with multiplicative seasonality with the following parameters:

$\alpha = 0.094,$

$\beta = 2.32$ and

$\gamma = 0.0004$

RMSE of the TES Full Model 22.411790148276328

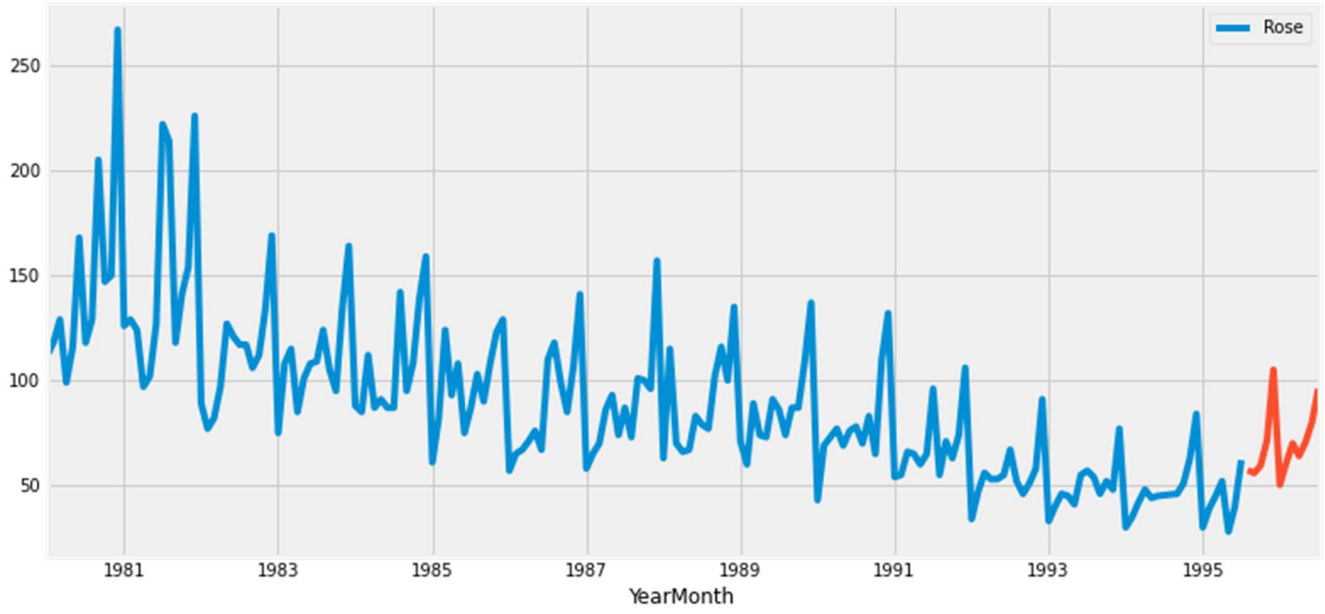


FIGURE 23 TES PROJECTION ON FULL DATA

One assumption that we have made over here while calculating the confidence bands is that the standard deviation of the forecast distribution is almost equal to the residual standard deviation.

Calculation The CI for the forecast

	lower_CI	prediction	upper_ci
1995-08-01	13.227126	57.272146	101.317167
1995-09-01	11.834505	55.879525	99.924545
1995-10-01	15.388069	59.433089	103.478109
1995-11-01	27.441866	71.486886	115.531906
1995-12-01	60.988252	105.033272	149.078292

TABLE 38 CI ON TES PROJECTION ON FULL MODELS

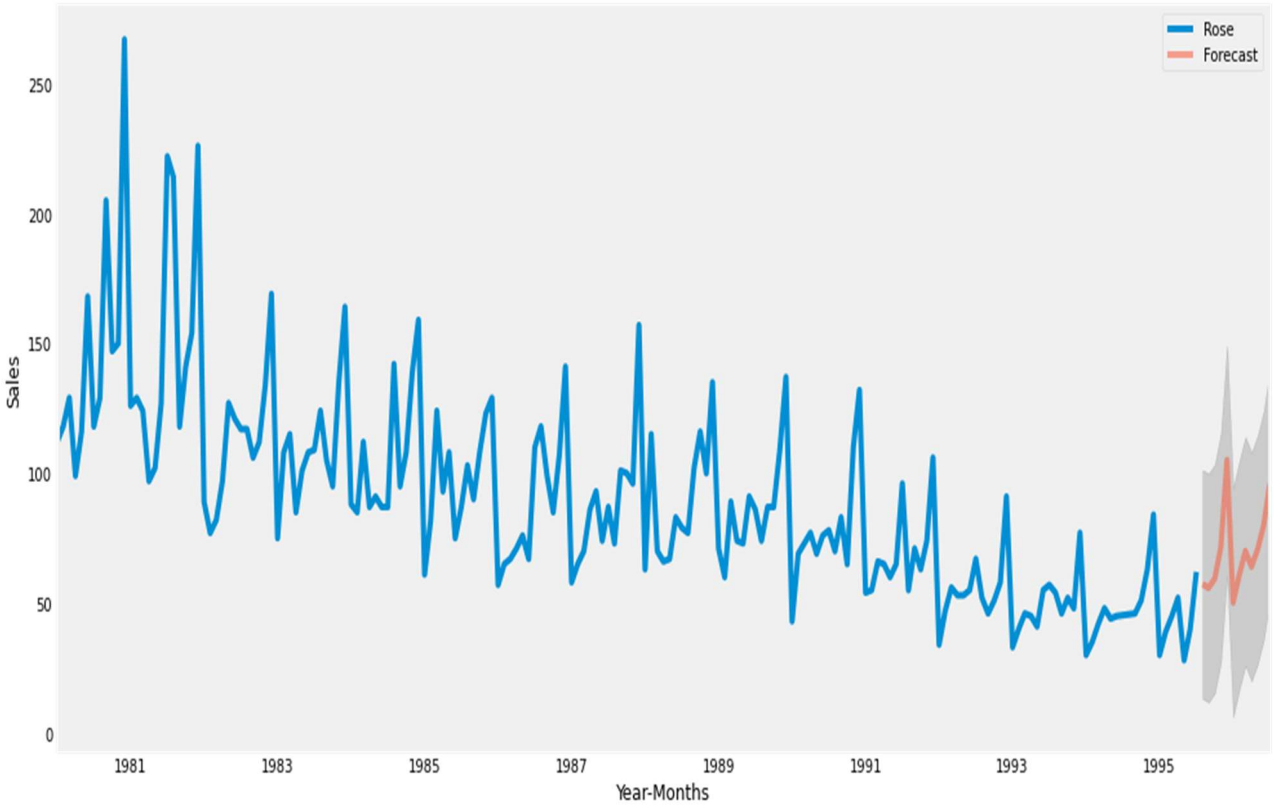


FIGURE 24 TES FORECAST ON FULL DATA WITH CI

Inference:

TES stands as best model to determine a good forecast on the full Rose TS dataset.22.411 is the RMSE in on th full dataset. Which is lower than the most of training and test models performed here. CI Intervals: The lower_CI and upper_CI columns represent the lower and upper bounds of the confidence interval for the prediction column respectively. A confidence interval is a range of values that we can be confident contains the true population parameter with a certain degree of certainty. Our case we have considered the confidence level is 95%. Which is plotted as well.

So, for example, if we take the first row of your data (1995-08-01), we can say that we are 95% confident that the true value of prediction lies between 13.227126 and 101.317167

RMSE Tabular Summary of all performed Models

	Test RMSE
Alpha=0.098,SES	36.796236
Alpha=1.49,Beta=7.38:DES	15.268957
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	14.278440
Alpha=0.071,Beta=.045,Gamma=8.38:TES	20.189764
Regression On Time	51.433312
Naive Model	79.718773
Simple Average	53.460570
Moving Average	15.236052
2pointTrailingMovingAverage	68.970159
4pointTrailingMovingAverage	46.403626
6pointTrailingMovingAverage	39.126446
9pointTrailingMovingAverage	34.410938
ARIMA(2,1,2)	37.306480
SARIMA(0,1,2)(2,0,2,12)	26.928362
RMSE of the TES Full Model	22.411790

Table 39 RMSE SUMMARY OF ALL MODELS

QN-8 Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Sorting

	Test RMSE
Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES	14.278440
Moving Average	15.236052
Alpha=1.49,Beta=7.38:DES	15.268957
Alpha=0.071,Beta=.045,Gamma=8.38:TES	20.189764
RMSE of the TES Full Model	22.411790
SARIMA(0,1,2)(2,0,2,12)	26.928362
9pointTrailingMovingAverage	34.410938
Alpha=0.098,SES	36.796236
ARIMA(2,1,2)	37.306480
6pointTrailingMovingAverage	39.126446
4pointTrailingMovingAverage	46.403626
Regression On Time	51.433312
Simple Average	53.460570
2pointTrailingMovingAverage	68.970159
Naive Model	79.718773

TABLE 40 RMSE VALUE IN DESCENDING ORDER

Insights:

1.The TES model on full dataset is having the lowest RMSE(14.27) Value with Parameter **Alpha=0.094,Beta=0.2.39,Gamma=0.0004:TES**

2.While the highest RMSE(79.71) value is outcome of DES Model with Naïve Model.

QN 9 Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.(for Both Sparkling and Rose TS dataset)

1.The three fundamental steps to model a time series are building a model for time series, validating the model and using the model to forecast future values/impute missing values.

2.The first step in time series modelling is to read the data ,account for existing seasons (a recurring pattern over a given period of time) and trends (upward or downward movement in the data). Accounting for these embedded patterns is what we call making the data stationary.

Steps:

a. Collecting the data and cleaning it

b. Preparing Visualization with respect to time vs key feature

c. Normal check of data shape, data Information, Description and plot (Yearly and Monthly),Missing values

d. Performed and plot Decomposition by both additive and multiplicative methods and log transformed.

e. Compared the normal plot with log plot

f. identified the patterns in time series data

3.Check for stationary of data and make it stationary by prescribe common method. Once the data is stationary, the next step is modelling to establish a base level forecast. This can be done using various techniques such as exponential smoothing methods as Simple, Double and Triple exponential (moving average),Linear regression, Naive Model, Simple Average, Moving average.

4.For the modelling of ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data. Considered the best param combination to get the lowest RMSE.

a. Data break up into Train and test as per desired year

b. All types of Exponential methods performed

c. Tuned the models on hypermeter by Param search

d. hyperparameters:

Alpha: Smoothing factor for the level.

Beta: Smoothing factor for the trend.

Gamma: Smoothing factor for the seasonality.

Trend Type: Additive or multiplicative.

Dampen Type: Additive or multiplicative.

Phi: Damping coefficient.

Seasonality Type: Additive or multiplicative.

Period: Time steps in seasonal period.

e. Performed the Models and Evaluate them individually and plotting

ARIMA and SARIMA model performed on low AIC value

Uses the combination parameter to get the best combination

Improving AR Models by making Time Series stationary through Moving Average Forecasts

ARIMA models consist of 3 components: -

AR model: The data is modelled based on past observations.

Integrated component: Whether the data needs to be differenced/transformed.

MA model: Previous forecast errors are incorporated into the model.

for SARIMA, we first need to define a few parameters and a range of values for other parameters to generate

a list of all possible combinations of p, q, d, P, Q, D, s.

we get the SARIMA (p, d, q)(P, D, Q, s) model.

5.The final step in modelling is evaluating model accuracy. This has been done by using various statistical measures root mean squared error (RMSE).Generated a table of RMSE for the whole modelling outcome and sorted out in descending order to get the lowest RMSE value at top.

6.Than Run the best model on test data set than on the whole dataset to predict 12 months into the future (till the end of next year) as desired.

Business Interpretations and Actionable Insights

Business Interpretations of Sparkling

1.In Yearly Box plot there is not fix growth and degrowth trend whereas trend can be observed while in Monthly sales analysis by plotting the data and the Seasonality can be seen clearly.
2. There are sharp increase in sale from month of August every year which touches the peak in December. While the sales peak is in month of December, which might be due to Holidays, some strong festival and new year celebration.
3.We can also observe a bit jump in sales in month of March and April the actual reason not mentioned in data record.
4.Month June can be considered as poor month in year for Sparkling sales view point.
5.As per sales perspective the minimum sales was in year 1983 was less than 2000 while the highest sales are recorded in year 1988 Dec.
6.After 1983 there was increasing trend where as sales has started shooting from 1987 which cannot be accidental.

Business Interpretations of Rose

1.In Yearly Box plot there is degrowth trend whereas trend can be observed while in Monthly sales analysis by plotting the data and the Seasonality can be seen clearly.
2. There are normal increase in sale in month of August every year and hiting the peak in month of December. 3.We can also observe lowest sales in Month January and a little jump in sales in month of March and dip in April the actual reason of fluctuation in sales has not mentioned in data record.
4.The Rose wine is showing continuous downward trend in sales throughout the time period. After year 1988 the sales are in sharp decline trend which is cause of worry and need address immediately. 5.The Rose sales are going down year by year this means the brand has not having a minimum loyal customer base.

Over all Actionable Insights

A. For this we have to go through the factors apart from data prevailing in market, like Competition, Demand Trend, Pricing, Lucrative Schemes.
B.As per data modelling and projection the Sparkling sales will not going to record the degrowth in sales but not showing the healthy growth too.
C. Company have to find the actual factor(s) behind this sales jump from 1987 and sales dip of June 1983.
D. Company should analyse the physical market conditions of 1987 to 1988, because that was neither their entrance period nor their stage of wrapping up.
E. Market survey and competitor pricing or other relevant lever has to find to use to incorporate in planning. These are not insight of analysis but there was a trend for two year which should be analysed or added for further forecast.
F .In year 1988 both brands(Sparkling and Rose) has generated good number of sales, but

after 1988 the Rose decline and in case of Sparkling after 1990 sales has showing the good increasing trend.
G. After analysis of Both the brand it can be infer as company is supporting more to brand Sparkling more than Rose. Which may be one of reason for continuous downward trend.
H. Need to analyse the not picking of seasonal sales from Aug to Dec in every year of Rose as per Sparkling pattern.
I. The forecast has been shared for the both brand on the basis of past data modelling. Best Model suggesting that next 12 month sales are on growth pattern . Which would be helpful analyse the other physical factors also for further growth of both the Brands.

Note: ***The Jupyter notebook has been attached here which contains the all type of calculations, process and libraries formulas used, Assumptions considered, Verification, Model evaluation, RMSE generation and Model comparison, AIC value extraction. Model performance. On these vary basis the Inferences has been made to translate the data into actions***.

----------THE END----------