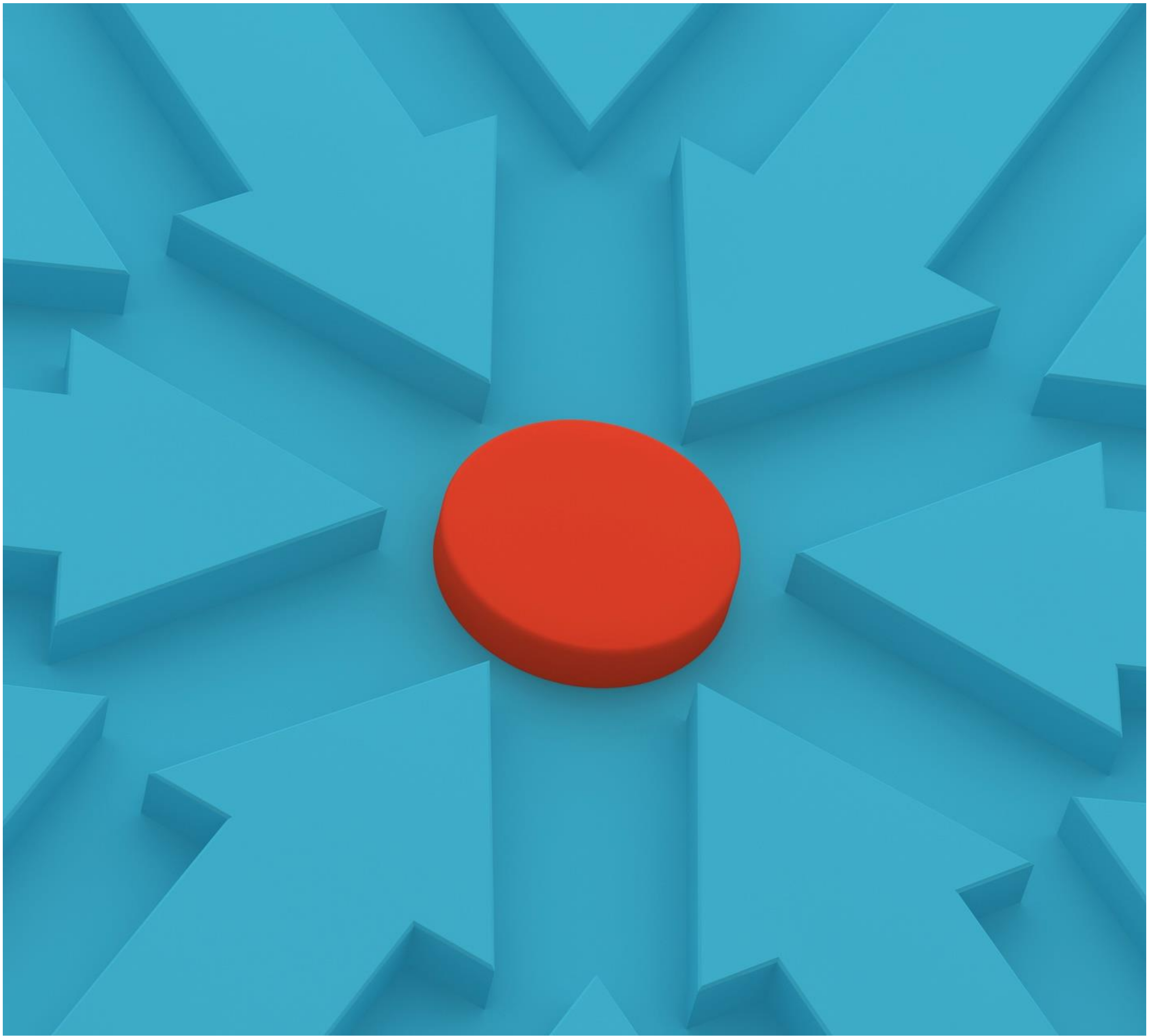


TIME SERIES FORECASTING

ROSE PROJECT REPORT

SUNDAR RAM S
PGPDSBA
ONLINE DEC_C 2021
30-JUL-2022

Contents

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF EQUATIONS

PROBLEM – TIME SERIES FORECASTING

Problem Statement

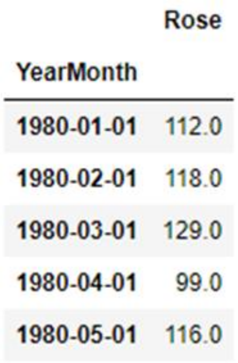
For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century.

Dataset 2 – Rose

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

Sample of the Dataset

The dataset can be read as an appropriate time series data by using the pandas read_csv method utilizing the appropriate arguments (parse_dates = True & setting the index column to be the Year Month in this case).



	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 Sample of the Dataset

The dataset contains sale data of Rose wine for 187 months starting from Jan-1980 to Jul-1995.

Dataset Description

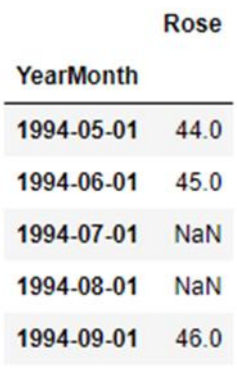
The Rose refers to the sale of Rose wine in a particular month of the year (set as index)

Variable types

The Rose contains the sale data of rose wine and is of type float64.

Check for null values

There are 2 null values in the dataset. Null values are not allowed in time series, as it is ordered data. In this particular dataset, null values are found in the Jul & August of the year 1994.



	Rose
YearMonth	
1994-05-01	44.0
1994-06-01	45.0
1994-07-01	NaN
1994-08-01	NaN
1994-09-01	46.0

Table 2 Null values - Rose

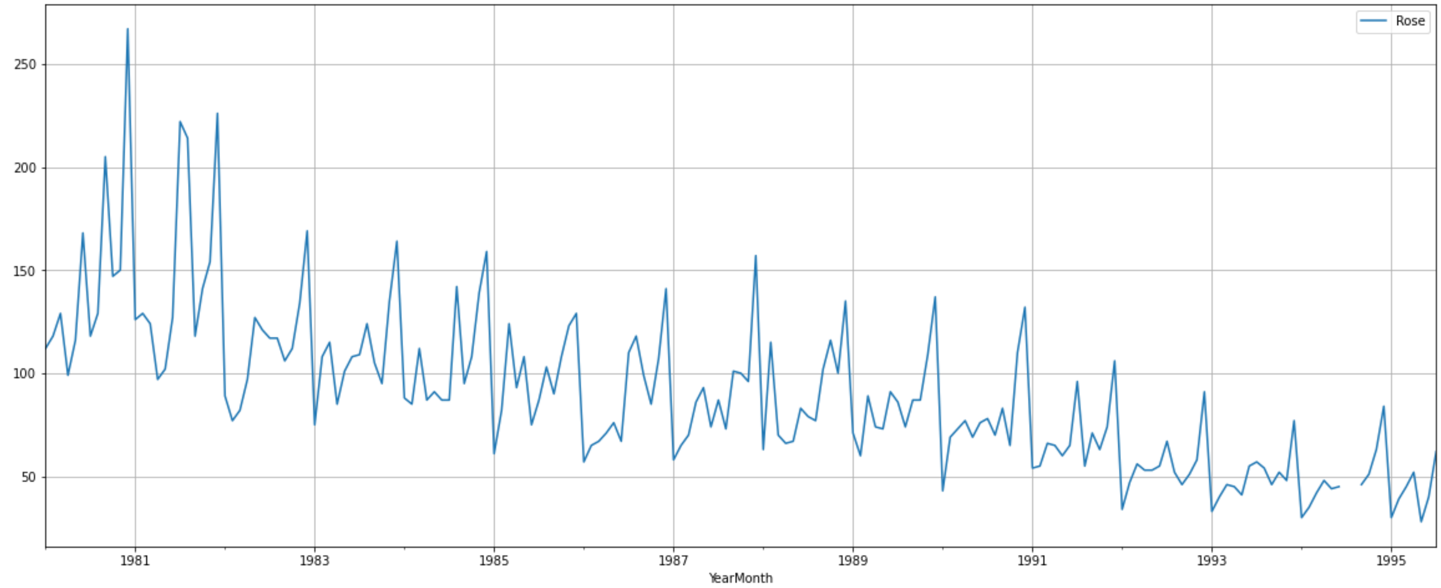


Figure 1 Time Series Plot – Rose Sale – Null values

It can be seen that there is Seasonality in the data. The sales are increasing towards the end of each year starting from the mid of the year. Hence, the data need to be imputed considering seasonality. Since, the trend has been decreasing over the years, it is the best to consider the data of previous 2 to 3 years for imputation.

Considering the average July sale for the years 1992, 1993 & 1995, the 1994 July data is imputed with 62 & considering the average August sale for the years 1991, 1992 & 1993, the 1994 August data is imputed with 53.

The plot after imputing the null values is in the below figure 2. We can see that the trend of Rose wine sale is steadily decreasing across the years. Also, there is Seasonality in the data. The sales are increasing towards the mid & peaks the end of each year and fall in the Year Starting.

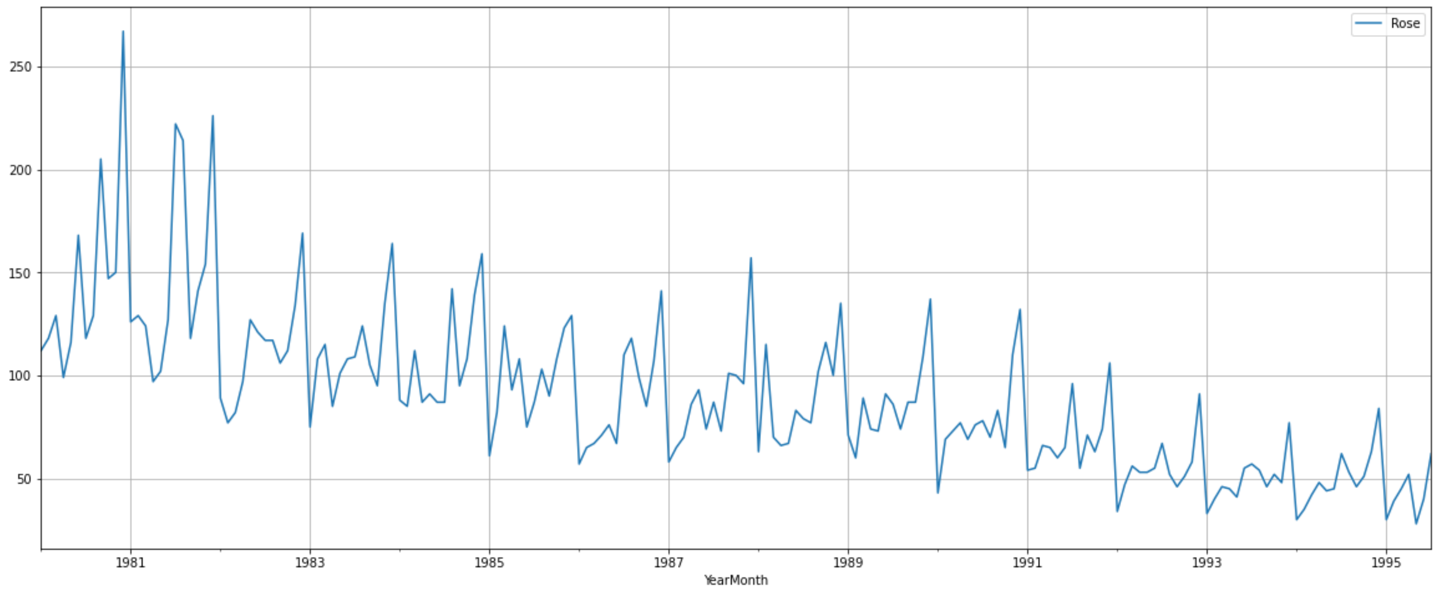
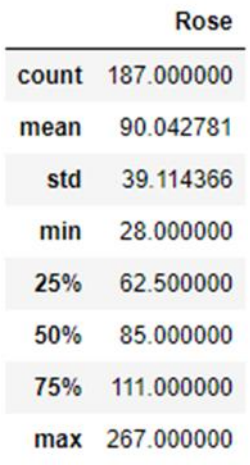


Figure 2 Time Series Data - Rose wine - After imputing Null values

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

Summary of the dataset

Summarizing briefly, there are 187 records in the sparkling dataset. The mean sales are 90 with a standard deviation of 39. The minimum sales recorded is 28 & the maximum sales recorded is 267. Also, we can see that the data is not normally distributed as the median and mean sales have difference.



	Rose
count	187.000000
mean	90.042781
std	39.114366
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Table 3 Summary Dataset

Yearly Sales Analysis – Box Plot

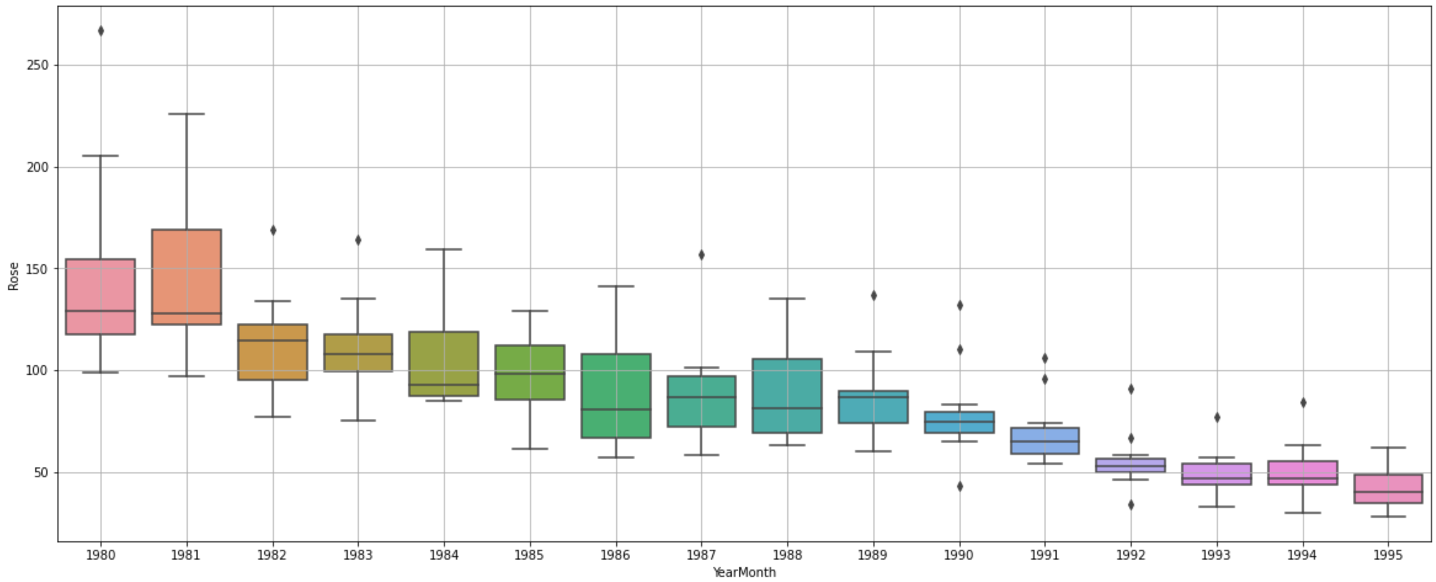


Figure 3 Yearly Sales - Box Plot

From the above plot it can be seen that the sales are decreasing across years. There are outliers present in most of the years. This would be due to the year-end sales, that was visible as seasonality in the time series plot.

Monthly Sales Analysis

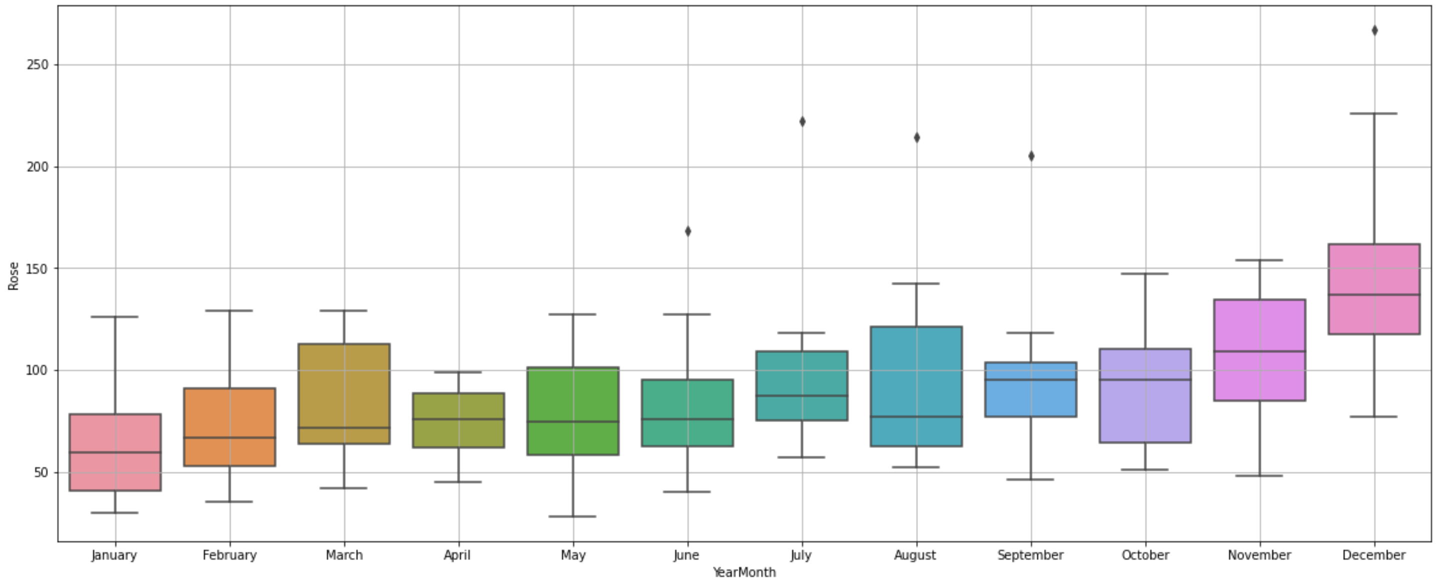


Figure 4 Monthly sales across years - Box Plot

From the above plot it can be seen that the sales increase towards the end of the year. It can also be seen that the median sale is minimum in January. The sale is highest in December. The median sale is found to be in an increasing trend across months with an exception in August.

Monthly Sales Across Years

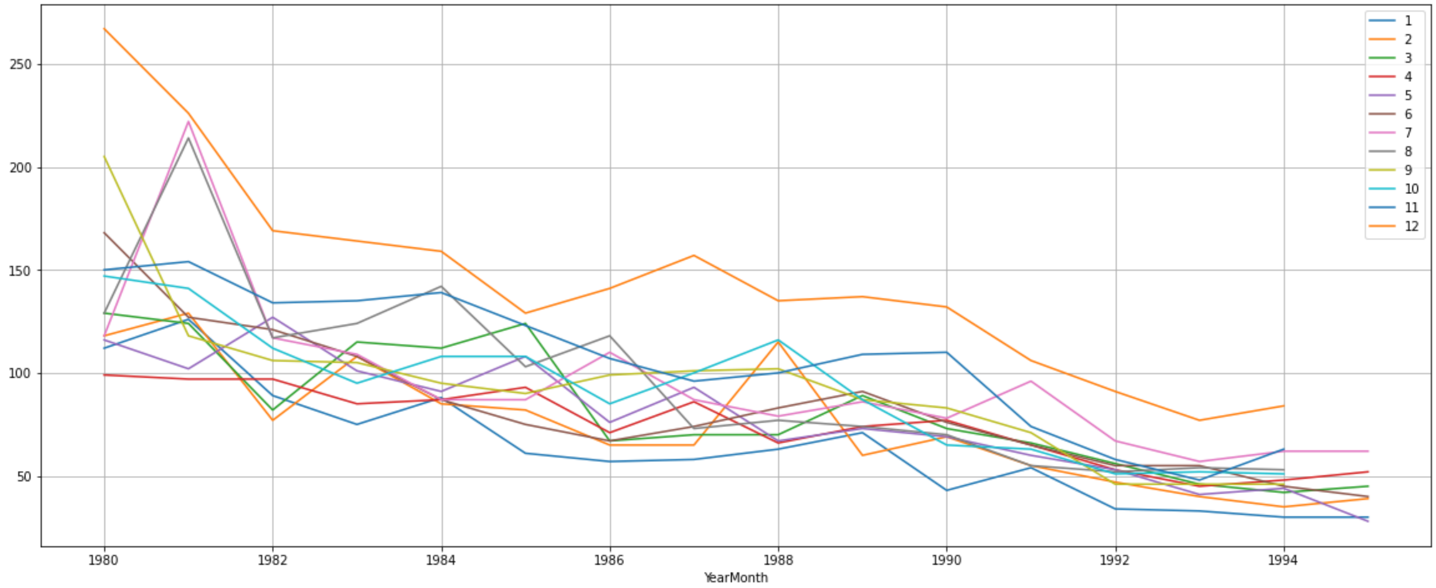


Figure 5 Monthly Sales across years

From the above plot it can be seen that sales across years in December is the highest. The highest December sale happened in the year 1980. After that the sales have been on a decreasing trend across years. Similarly, the sales in other months are also having a decreasing trend across years.

Decomposition

Decomposition is a process of decomposing the Time Series in to its components of Trend, Seasonality & Error. There are two models in decomposition – Additive and Multiplicative.

Additive Model

Sales = Trend + Seasonality + Error

Equation 1 Additive Model – Decomposition

Additive model is depicted by the above equation. It is the sum of Trend, seasonality & error.

Multiplicative Model

Sales = Trend * Seasonality * Error

Equation 2 Multiplicative Model – Decomposition

Multiplicative model is depicted by the above equation. It is the product of Trend, seasonality & error.

The model that bests suits depends on the time series itself. If there is an additive increase or decrease in Sales/ seasonal sales across time, it can be visualized in the additive time series. In case of a multiplicative increase or decrease in sales/ seasonal sales across time, it can be visualized in the multiplicative time series.

The below plot depicts the additive decomposition of the Time series. The trend can be seen to be decreasing over the years. The seasonality can be seen similar to the original series. There is a seasonal increase in sales towards the end of the year and drop in the year start. We can also see that the error is widespread, and there is some seasonal pattern of error in the middle years of the dataset. However, we can understand that there is still data that is yet to be captured. The error is centered across 0 with a wide range of values. Hence, we need to explore the multiplicative model also.

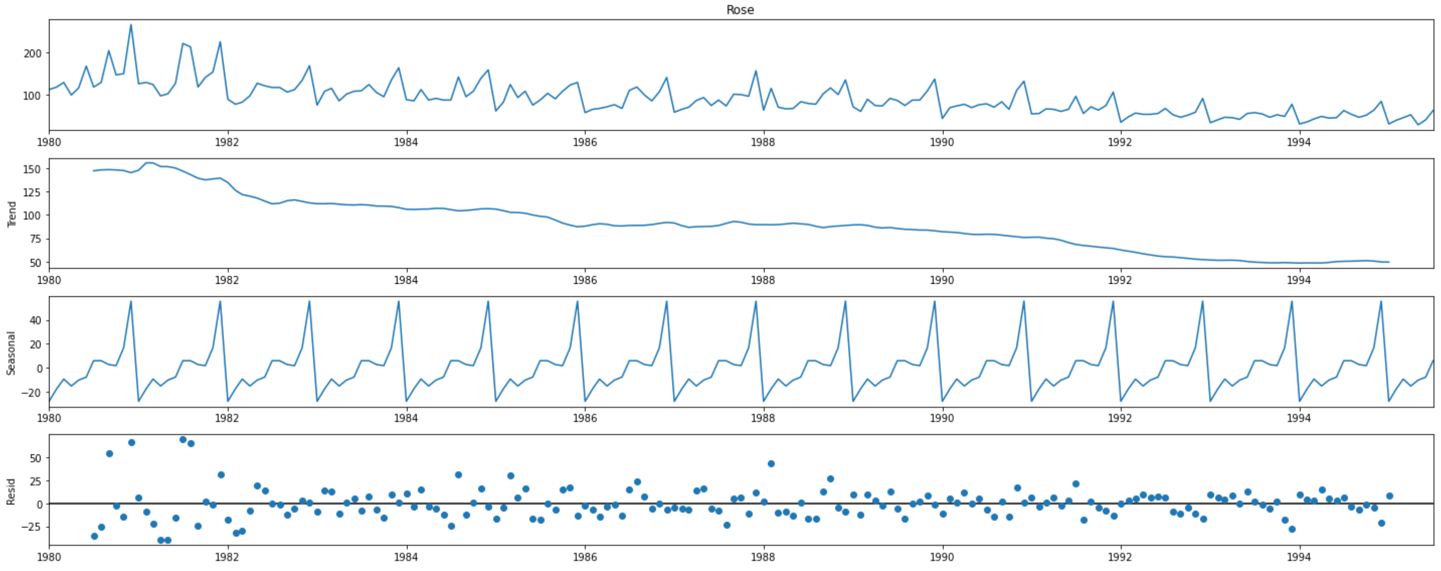


Figure 6 Additive Decomposition - Rose TS

The below plot depicts the multiplicative decomposition of the Time series. The trend & seasonality can be seen to be same as that in the additive model. Here, the error component can be seen to flattened. Error is seen to be less spread and is centered across 1 with minimal range. Hence this is evident that, this Time series is more or less a multiplicative model.

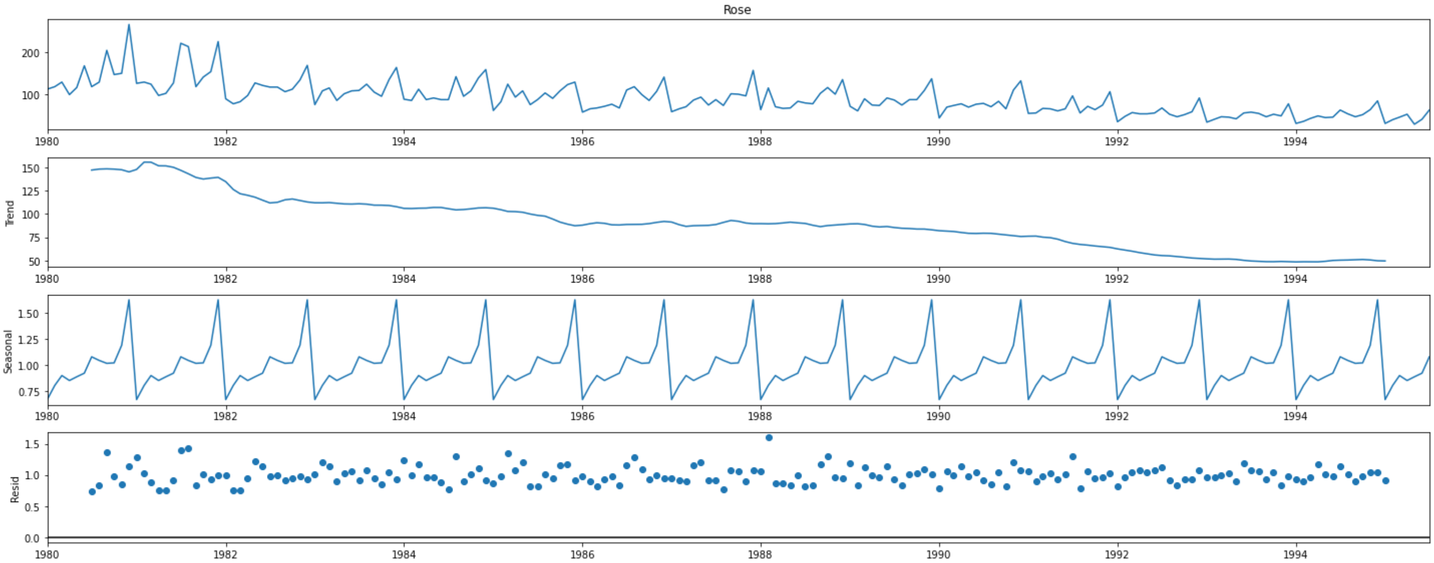
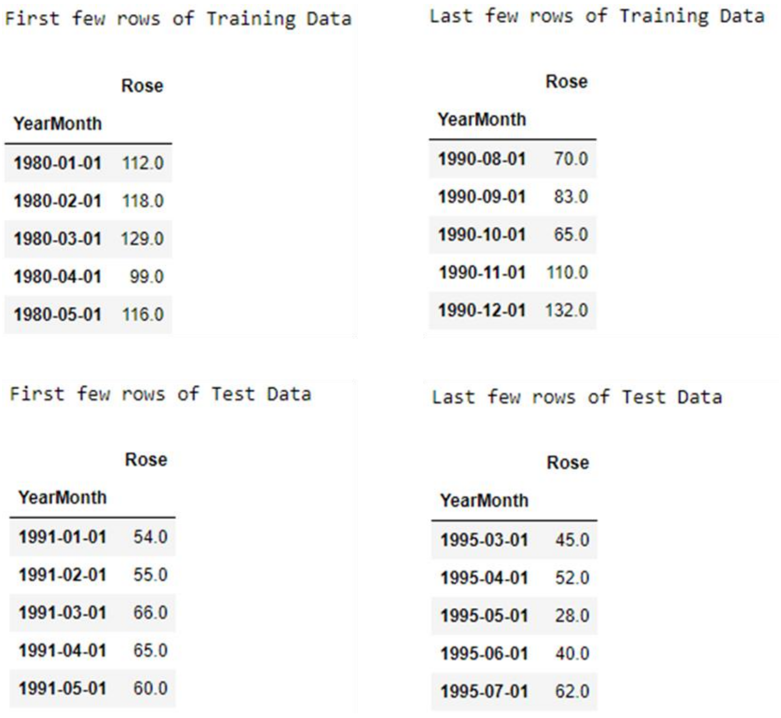


Figure 7 Multiplicative Decomposition – Rose TS

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

The train test split of time series is not similar to the train test split of the other data. In time series, the data order is important & the data split cannot be randomized. Hence, as per requirement the data before the year 1991 is split as training data set and the data from 1991 is split as test data.



```
First few rows of 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

```
Last few rows of Training Data
```

	Rose
YearMonth	
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

```
First few rows of Test Data
```

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

```
Last few rows of Test Data
```

	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

Figure 8 Sample of the train & test data

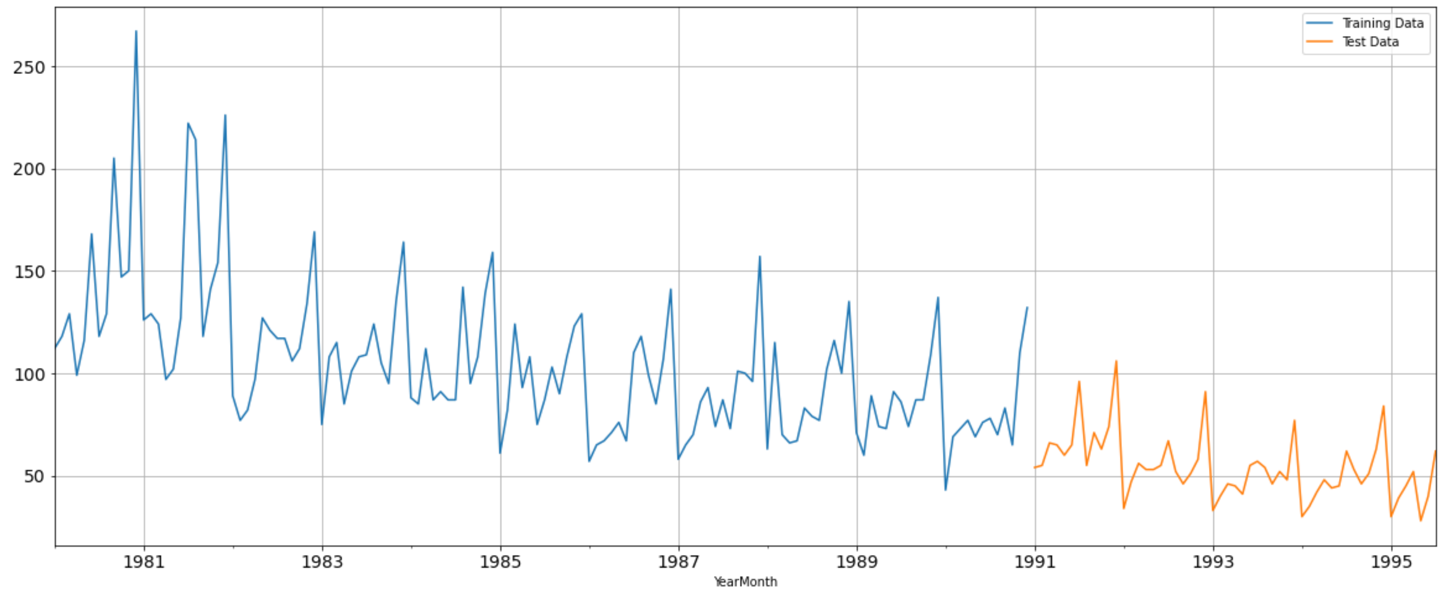


Figure 9 Train & Test Data plot

It is difficult to predict the future observations if such an instance has not happened in the past. From our train-test split we are predicting likewise behavior as compared to the past years.

1.4 <u>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.</u>

LINEAR REGRESSION:

Before proceeding with the Linear Regression model, we need to make some changes to the dataset. Linear regression involves predicting the dependent variable y based on the independent variable x. But in the current data set there is only one dependent data. Hence the independent data here is the time stamp. Thus, we need to generate numerical time instance order for the time stamp which will act as our independent variable.

Now, having generated the numerical time instances for training and test data, we can proceed to modelling. There are 132 training time instances and 55 testing time instances. The Linear Regression model is available in the Sci Kit Learn library of python. The model is fit with the training Time instances as independent variable and training sparkling sales as the dependent variable. Post fitting and training the model with the train data, the model is tested with predictions on the test data.

As the name suggests, the predicted values form a straight line when plotted. The Linear Regression predicted plot is as below when plotted against the original dataset. The RMSE values are found to be 15.303.

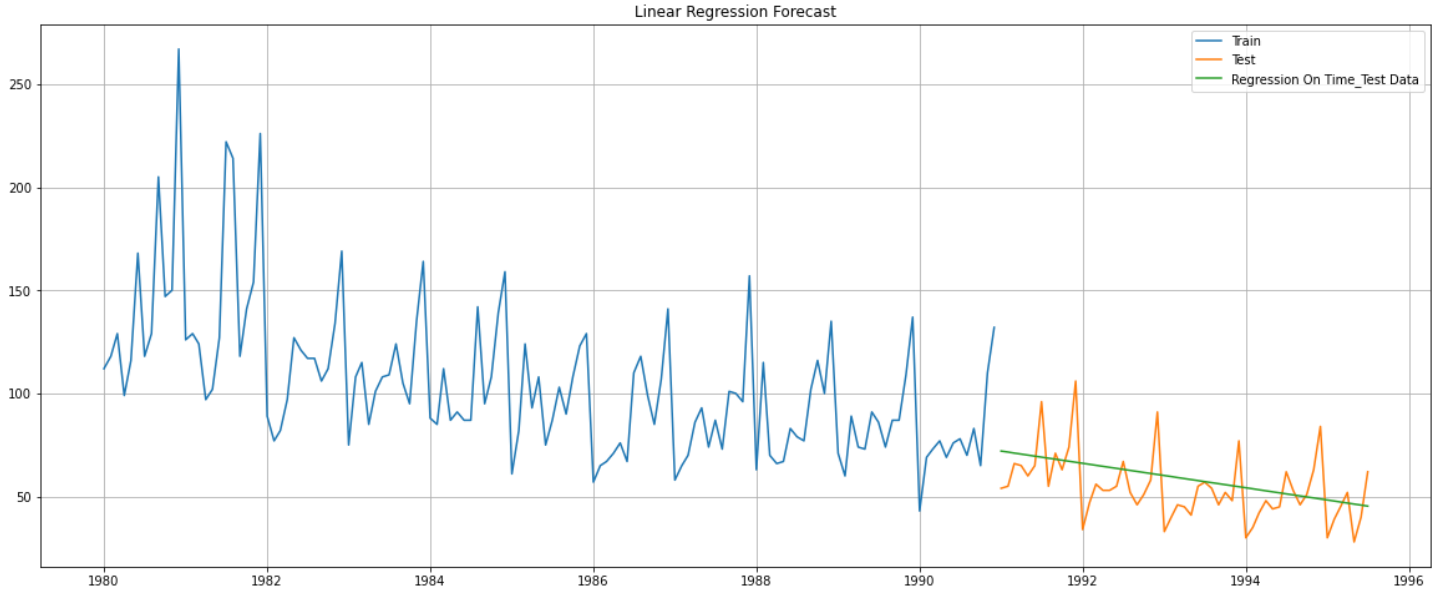


Figure 10 Linear Regression model

NAIVE MODEL:

Naïve model, as the name suggests is very naïve in approach. It simply considers the last seen value as the value for upcoming predictions. The naïve value in this case is 132. Hence there will be a horizontal line parallel to the time axis. Plot as below. RMSE for the same is 79.282

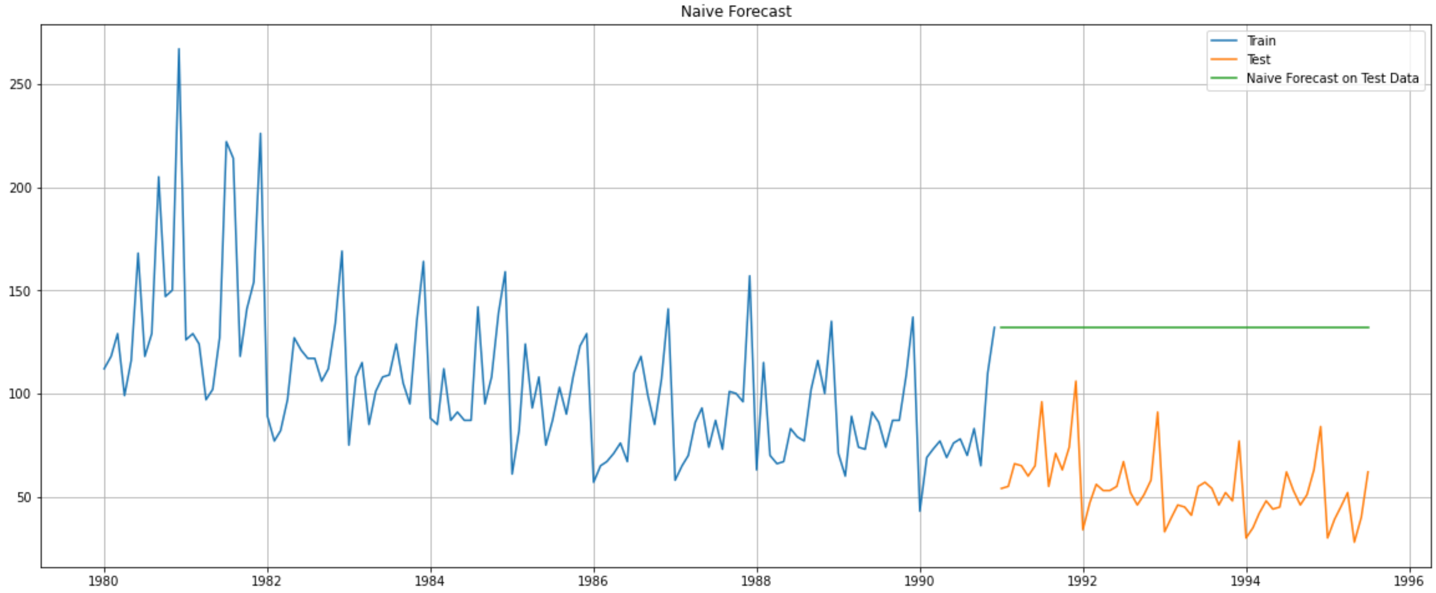


Figure 11 Naive Model

SIMPLE AVERAGE MODEL:

A Simple average model is also naïve in approach, except that it predicts all the future values to be the mean value of the available data, which is 104.94 in this case. This is also a straight line parallel to the time axis. Plot as below. RMSE value is 53.03.

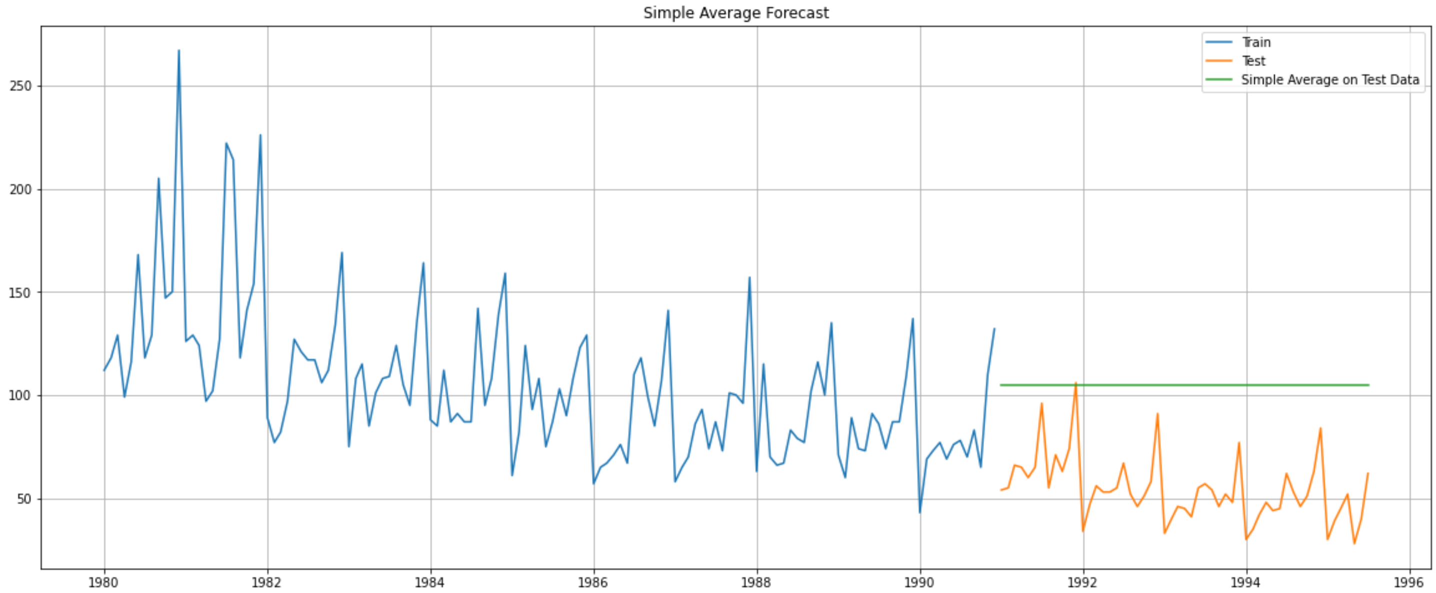


Figure 12 Simple Average Model

SIMPLE EXPONENTIAL SMOOTHING MODEL:

Exponential smoothing methods consist of flattening time series data. 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). One or more parameters control how fast the weights decay. These parameters have values between 0 and 1.

Simple Exponential Smoothing is applicable to a Time series that neither has a trend nor seasonality. It is the simplest of the exponentially smoothing methods and is naturally called simple exponential smoothing (SES). In Single ES, the forecast at time (t + 1) is given by Winters,1960

$$Ft+1=\alpha Yt+(1-\alpha)Ft$$

Equation 2 Simple Exponential Smoothing

Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

The stats model library in python has separate module for Time series analysis, in which there is a model for Simple Exponential Smoothing. This model is imported, built and fit with the TS training data under optimized parameters, so that the best hyperparameters are found. The hyper parameter here is the alpha value, that comes out to be 0.0987. The hyper parameters for Trend and seasonality will be Not a Number values, as Simple exponential smoothing considers only level term.

The fit model is then used for predicting the values to the length of the test set and are plotted. The RMSE value is found to be 36.381.

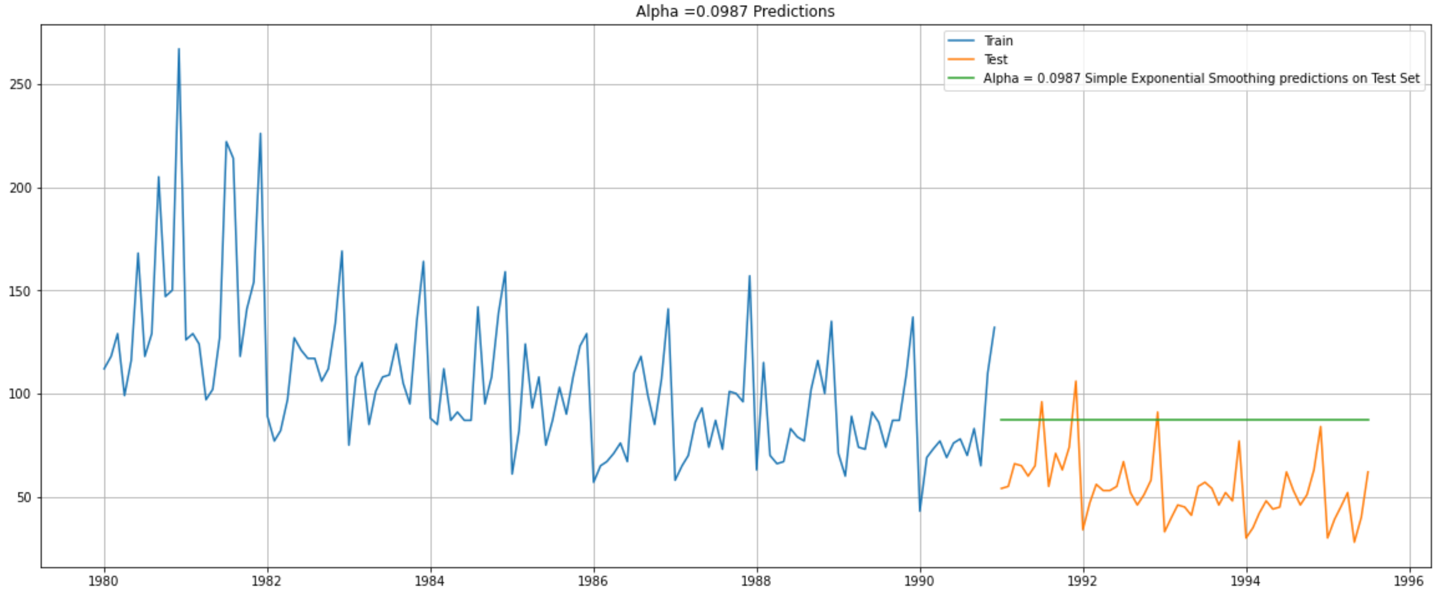


Figure 13 Simple Exponential Smoothing

However, on iterating for different values of alpha, we can see that, 0.06 has the lowest RMSE on test data. The model plot is as below.

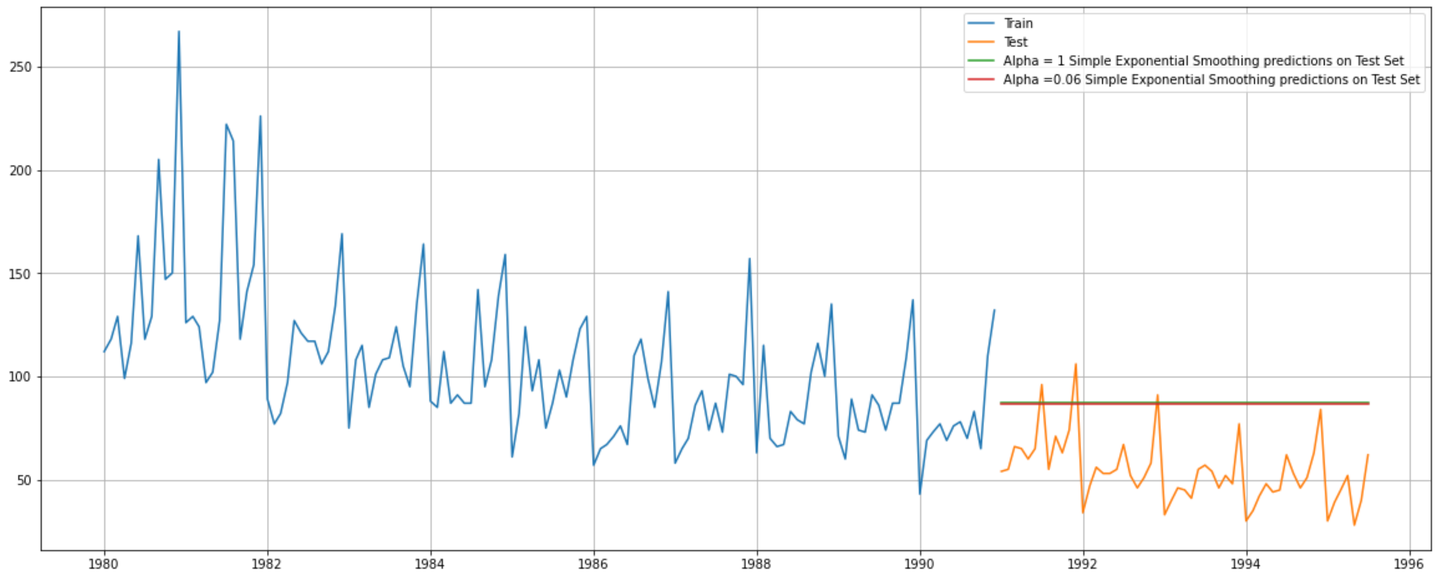


Figure 14 SES @ alpha = 0.06

DOUBLE EXPONENTIAL SMOOTHING MODEL:

One of the drawbacks of the simple exponential smoothing is that the model does not do well in the presence of the trend. This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters and is applicable when data has Trend but no seasonality. It is also called the Holt's Linear method. However, this model is built for understanding purposes. Two separate components are considered: Level (error) and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series. A second smoothing parameter β corresponds to the trend series. Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short-term average value or level and the other for capturing the trend.

$$Lt = \alpha Yt+(1-\alpha)Ft\mathrm{Lt{=}\alpha Yt{+}(1{-}\alpha)Ft}$$

Equation 3 Level Equation

$$Tt=\beta(Lt-Lt-1)+(1-\beta)Tt-1\mathrm{Tt{=}\beta(Lt{-}Lt{-}1){+}(1{-}\beta)Tt{-}1}$$

Equation 4 Trend Equation

Here, α and β are the smoothing constants for level and trend, respectively,

- 0 <α < 1 and 0 < β < 1.

The forecast at time t + 1 & t + n is given by

$$Ft+1=Lt+Tt\mathrm{Ft+1=Lt+Tt}$$

$$Ft+n=Lt+nTt$$

Equation 5 Forecast Equation - Holt's Linear Method

The stats model library in python has separate module for Time series analysis, in which there is a model for Double Exponential Smoothing (Holt). This model is imported, built and fit with the TS training data. The hyper parameters here are the alpha & trend coefficient values. The seasonality term will not be considered here. On iterating and finding the Test RMSEs for various values of the alpha & Beta, it can be seen that the Test RMSE is the lowest when alpha = 0.03 & Beta = 0.24.

The fit model is then used for predicting the values to the length of the test set and are plotted. The RMSE value is found to be 16.049.

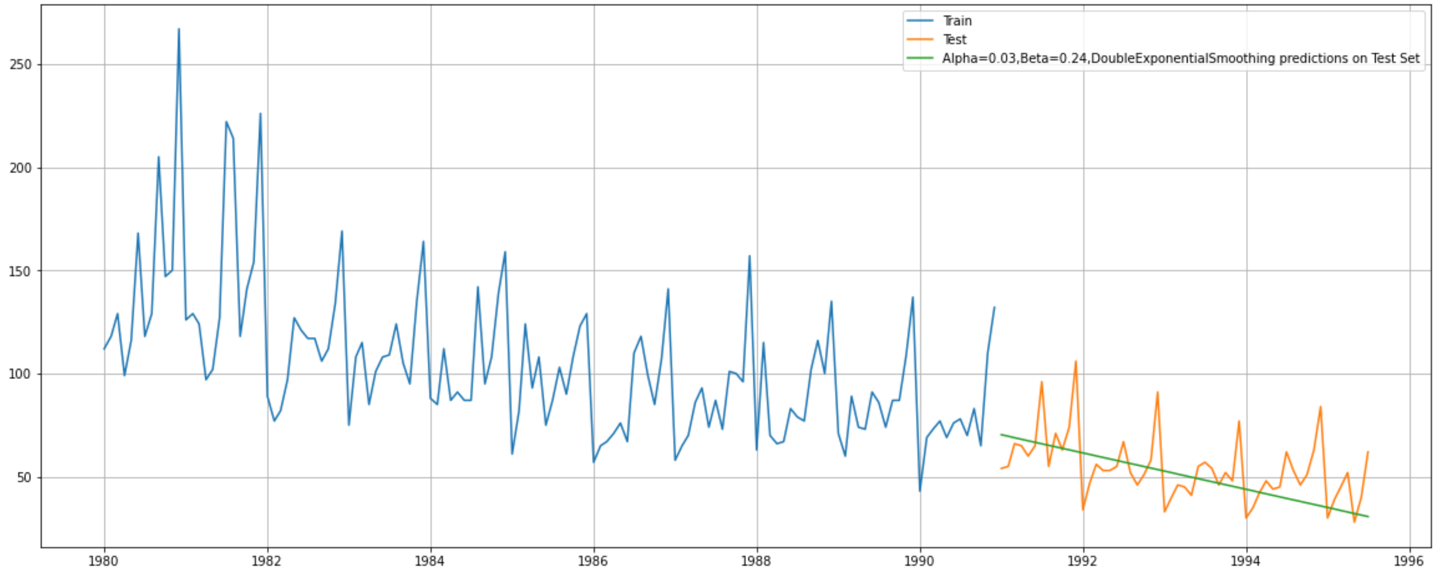


Figure 15 Double Exponential Smoothing plot

TRIPLE EXPONENTIAL SMOOTHING MODEL:

One of the drawbacks of the double exponential smoothing is that the model does not do well in the presence of the seasonality. This model is an extension of SES & DES known as Triple Exponential smoothing model or Holt’s Winter model, which estimates three smoothing parameters and is applicable when data has both trend and seasonality. However, the seasonality model here needs to be multiplicative as we have seen during decomposition. On fitting the model, we find the optimized parameter values of alpha, beta & gamma values to be 0.064, 0.053 & 0.00 respectively. The Test RMSE for this model was found to be 20.707. Model plot as in Figure 16. However, on iterating for different values of alpha, beta & gamma, we can see that, the lowest RMSE on test data is found when alpha = 0.03, beta = 0.66, gamma = 0.24. The model plot is as in the figure 17. The Test RMSE for this model is found to be 8.72 and is much better than SES, DES & even the first TES.

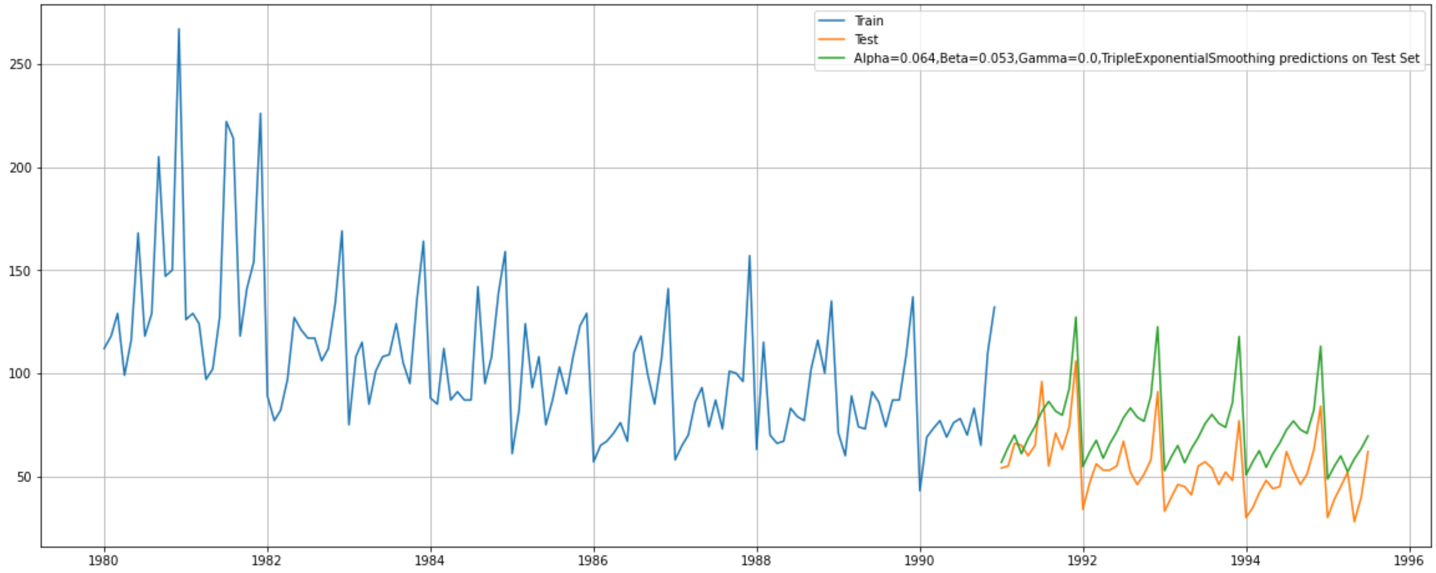


Figure 16 Triple Exponential Smoothing Model

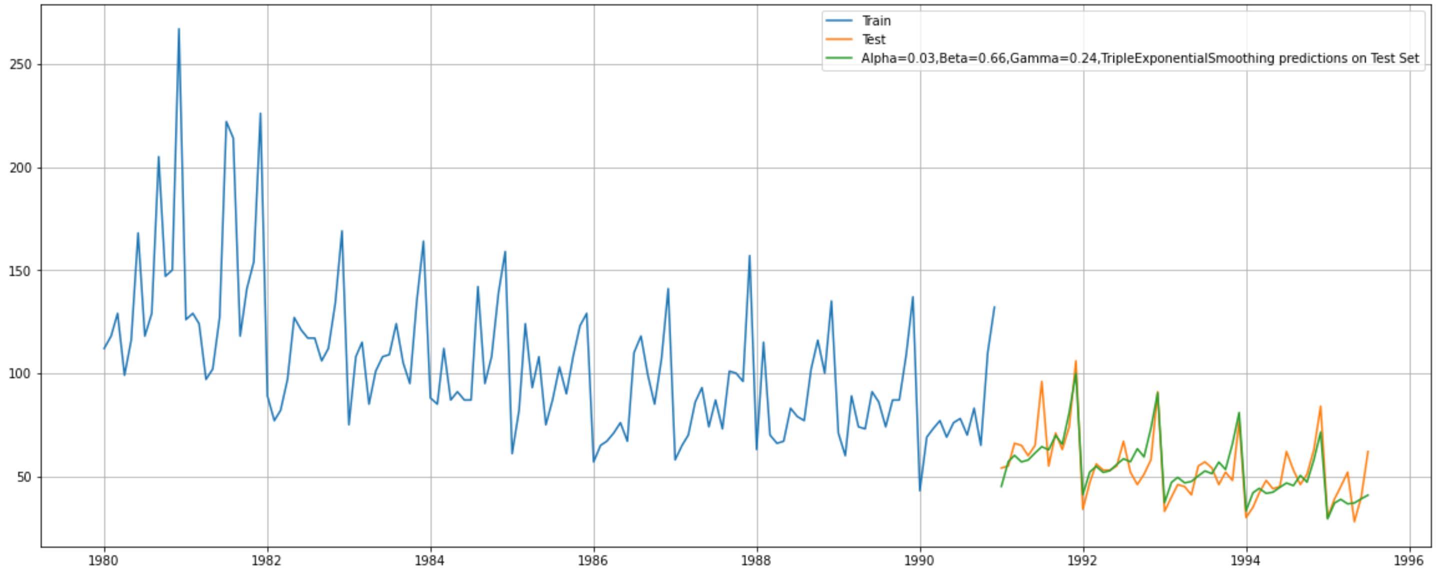


Figure 17 TES - alpha = 0.03, beta = 0.66, gamma = 0.24

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

Auto-regression means regression of a variable on itself. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR models

A Time Series is considered to be stationary whose statistical properties such as the variance and (auto) correlation are all constant over time. The properties of a stationary time series do not depend on time. The (auto) correlation observations only depend on how far apart these observations are and not where they are.

Stationary Time Series allows us to essentially have "copies" of things which enables us to do build appropriate statistical models for forecasting. One more intuitive understanding of the importance of stationarity is that the coefficients of the AR model should not be biased because the Time Series has a pronounced trend or seasonality.

To check whether the series is stationary, we use the Augmented Dickey Fuller (ADF)test whose null and alternate hypothesis are as below.

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

At our desired level of significance (chosen alpha value), in this case 0.05, we can test for stationarity using the ADF test. The ADF test works on the principle of finding the probability that a unit-root is present in the AR model. If a unit root is present in a Time Series, the Time Series shows a systematic pattern which is unpredictable thereby violating the idea of a stationary Time Series. At a very basic level, a process can be written as a series of monomials and each of these monomials corresponds to a root. If one of these roots is 1, then that can be said as a unit root. This is a very intuitive definition of a unit root.

The ARIMA model can be modelled as an autoregressive polynomial (of order ‘p’) which has ‘d’ roots on the unit circle. Correspondingly, we calculate a (modified) version of the t-statistic with appropriate degrees of freedom and compare it with the empirical values to conclude whether a unit root is present (and subsequently the Time Series can be said to be non-stationary).

In python, the augmented dickey fuller test is available in the stats model library under time series analysis as adfuller. We already know that the series has seasonality and the same has been verified using the dickey fuller test where the t statistic value turned out to be -1.88 & the corresponding p-value is 0.33. Here p-value > 0.05 and hence, we fail to reject the null hypothesis. Therefore, the time series is not stationary.

We can take appropriate levels of differencing to make a Time Series stationary. We can try various mathematical transformations to make the series stationary.

- Apply transformation and/or differencing.
- Check for stationarity.
- If the time series is not stationary repeat the process of differencing.

In this case, on performing a first order differencing and checking for stationarity it is found that the t statistic value is -8.02 & the corresponding p value is 1.98 e-12 which is less than 0.05. Hence, we can reject the null hypothesis & conclude that the first order differenced time series is stationary.

1.6 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.

ARIMA:

AR – Auto Regression, I – Integration & MA – Moving Average

We look at the Partial Auto-Correlations of a stationary Time Series to understand the order of an Auto-Regressive models. For an AR model, the PACF (Partial Auto Correlation Function) values cuts-off after a certain lag. The PACF values closes to 0 (at appropriate confidence intervals for the PACF plots) beyond that order (or lag). Partial auto-correlation of lag k, is the correlation between $\boldsymbol{Yt}$ and $\boldsymbol{Yt-k}$ when the influence of all intermediate values ($\boldsymbol{Yt-1}$, $\boldsymbol{Yt-2}$, ⋯, $\boldsymbol{Yt-k+1}$) is removed from both $\boldsymbol{Yt\ and\ Yt-k}$

Similarly for the Partial Autocorrelation of order p is the corresponding covariance values taking into account the intermediate values divided by the subsequent variance values.

For a MA model, the error component is modelled. A MA(q) model (Moving Average model of order q) can be written as:

Yt(hat) = et + α*1 et−1* + α*2 et−2* + …. + α*q et−q*

Equation 6 MA Model

We look at the Auto-Correlations of a stationary Time Series to understand the order of a Moving-Average models. For a MA model, the ACF (Auto Correlation Function) values cut-off at a certain lag. The ACF values closes to 0 (at appropriate confidence intervals for the ACF plots) beyond that order (or lag). Auto-correlation of lag k, $\boldsymbol{\rho}_{\mathbf{qk}}$, is the correlation between $\boldsymbol{Yt}$and $\boldsymbol{Yt-k}$. This particular function does not depend on ‘t’ since the Time Series is stationary. For building MA models, we look at the ACF plots and determine the order of the MA model.

The ACF(0) = 1 since this is the correlation of series with itself (without any lags). The PACF(1) = ACF (1) as for the PACF of order 1 we do not need to factor out the effect of another lags in between.

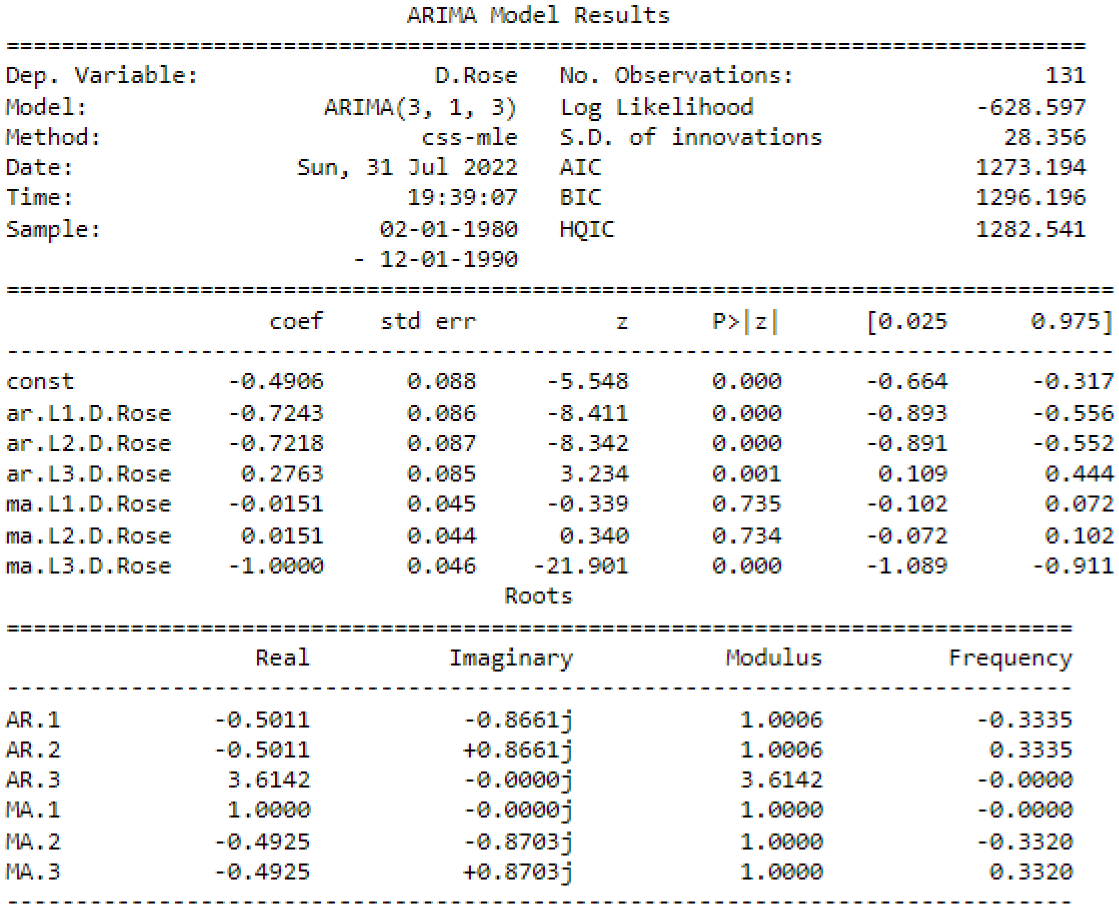
ARIMA models can be built keeping the Akaike Information Criterion (AIC) in mind as well. In this case, we choose the ‘p’ and ‘q’ values to determine the AR and MA orders respectively which gives us the lowest AIC value. Lower the AIC better is the model. In python we try different orders of ‘p’ and ‘q’ to arrive to this conclusion. Even for such a way of choosing the ‘p’ and ‘q’ values, we must make sure that the series is stationary. The formula for calculating the AIC is

2k –2ln(L),

Equation 7 AIC formula

where k is the number of parameters to be estimated and L is the likelihood estimation.

On the first hand, we need to find the best value of p & q that has the lowest AIC value. To do this, we first need to create the possible combinations of p, d, q, wherein d remains a constant and is equal to 1, as it is the order of differencing to stationarize the series. On creating an ARIMA model and running the model for different combinations of p, d, q, we get the AIC scores of all the models. The model with the least AIC score is selected and rerun separately. In this case, the least AIC score (1273.19) was found for p = 3, d = 1 & q = 3. The summary for the ARIMA model is in the below figure 18. We can see that there are 3 autoregression coefficients with both being significant & 3 moving average coefficients and both significant. The ARIMA model and also the AIC values are displayed in the ARIMA summary. On predicting the test values for ARIMA model and evaluating it, we see that the test RMSE is 16.21.



```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(3, 1, 3)   Log Likelihood                -628.597
Method:                       css-mle   S.D. of innovations             28.356
Date:                Sun, 31 Jul 2022   AIC                           1273.194
Time:                        19:39:07   BIC                           1296.196
Sample:                    02-01-1980   HQIC                          1282.541
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4906      0.088     -5.548      0.000      -0.664      -0.317
ar.L1.D.Rose    -0.7243      0.086     -8.411      0.000      -0.893      -0.556
ar.L2.D.Rose    -0.7218      0.087     -8.342      0.000      -0.891      -0.552
ar.L3.D.Rose     0.2763      0.085      3.234      0.001       0.109       0.444
ma.L1.D.Rose    -0.0151      0.045     -0.339      0.735      -0.102       0.072
ma.L2.D.Rose     0.0151      0.044      0.340      0.734      -0.072       0.102
ma.L3.D.Rose    -1.0000      0.046    -21.901      0.000      -1.089      -0.911
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.5011           -0.8661j            1.0006           -0.3335
AR.2           -0.5011           +0.8661j            1.0006            0.3335
AR.3            3.6142           -0.0000j            3.6142           -0.0000
MA.1            1.0000           -0.0000j            1.0000           -0.0000
MA.2           -0.4925           -0.8703j            1.0000           -0.3320
MA.3           -0.4925           +0.8703j            1.0000            0.3320
-----------------------------------------------------------------------------
```

Figure 18 ARIMA Model - p=3, d=1, q=3

SARIMA:

In SARIMA, The ARIMA is same as the ARIMA model, while ‘S’ stands for Seasonal.

For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the 'F' parameter indicates the seasonality/seasonal effects over a particular period. For deciding the 'P' and 'Q' values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of 'F' and see where these cut-offs (for appropriate confidence interval bands). We can also estimate 'p', 'q', 'P' and 'Q' by looking at the lowest AIC values. The seasonal parameter 'F' can be determined by looking at the ACF plots. The ACF plot is expected to show a spike at multiples of 'F' thereby indicating a presence of seasonality.

For SARIMA also the stationarity is important. We saw that the stationarity is established after the first differencing. Hence d = 1. However, the seasonal Differencing 'D' will be zero, as the Time series became stationary on the first normal differencing. From the ACF plot on the first order differenced training data, it can be seen that the seasonal parameter is 6 from the Differenced ACF plot in figure 19. Similar to the ARIMA model, the best p, q, P, Q are determined through iterating the SARIMA model for different combinations of the 4 values, keeping the d=1, D=0 & F=6.

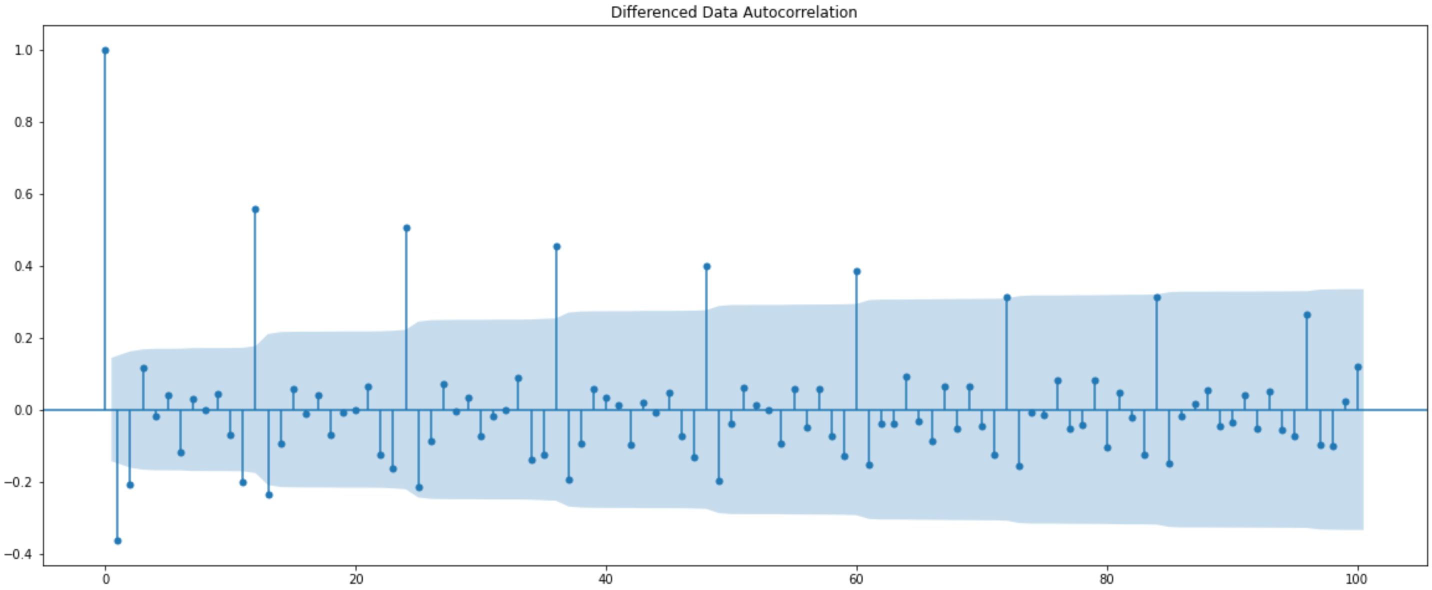
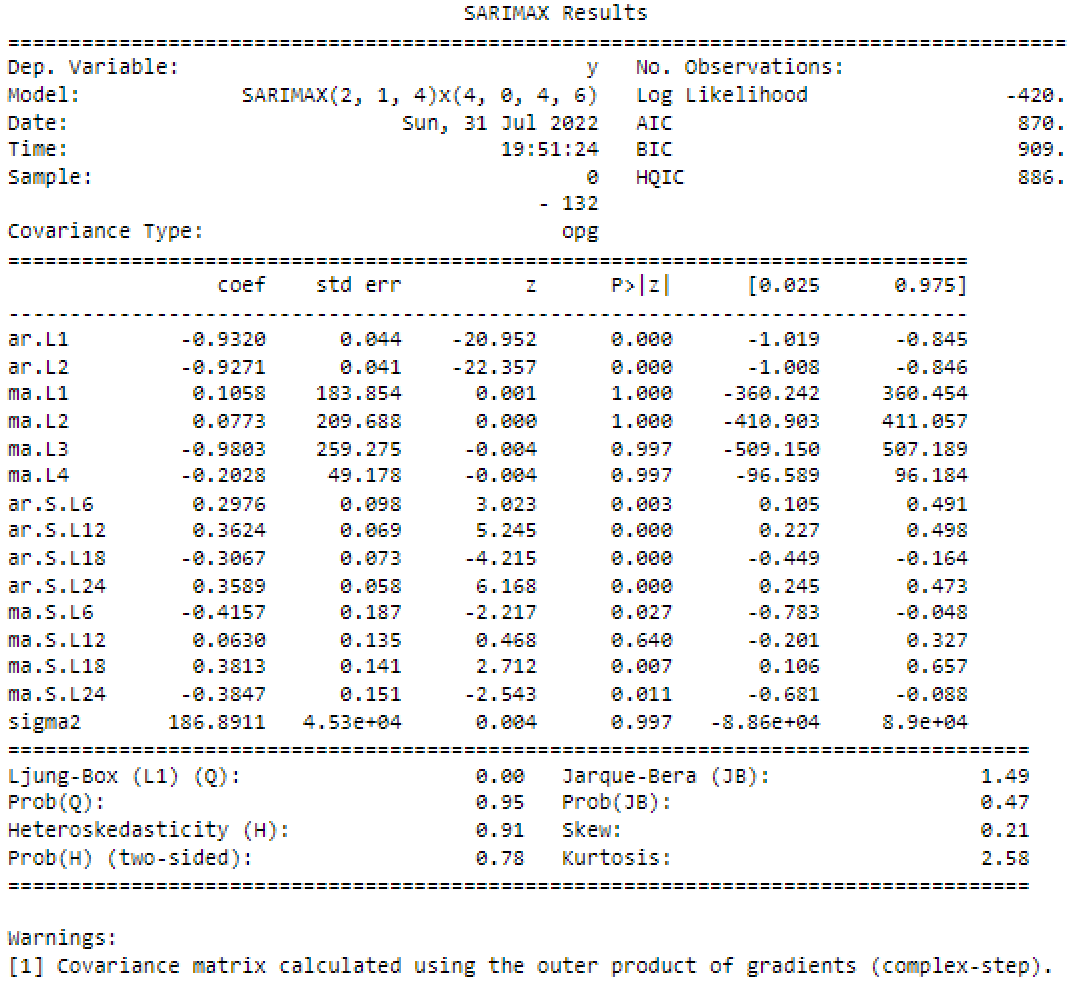


Figure 19 ACF plot on differenced Data

The hyper parameters in SARIMA model are the order of p,d,q, seasonal order P,D,Q,F. the enforce stationarity remains false, as the data is already stationary at d = 1 & the invertibility is also set as False as the invertibility is not needed. It can be seen that the lowest AIC value (870.449) occurred for the values p = 2, q = 4, P = 4, & Q = 4.

On re-running the SARIMA model for p = 2, q = 4, P = 4, & Q = 4, d = 1, D = 0 & F = 6, we get the summary as below. On prediction for the length of test values & evaluating the RMSE for this model it can be seen that the value is 28.032. The diagnostic plot is also shown in the figure 21. It can be seen that the SARIMA model is almost similar to the Normal distribution. The differences in the both are leading to the RMSE values. However, the same is low, when compared to the other models. It can also be seen that, most of the points at the centre are concentrated near the Line, indicating that most of the information are captured.



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(2, 1, 4)x(4, 0, 4, 6)   Log Likelihood                -420.225
Date:                           Sun, 31 Jul 2022   AIC                            870.450
Time:                                   19:51:24   BIC                            909.824
Sample:                                        0   HQIC                           886.394
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9320      0.044    -20.952      0.000      -1.019      -0.845
ar.L2         -0.9271      0.041    -22.357      0.000      -1.008      -0.846
ma.L1          0.1058    183.854      0.001      1.000    -360.242     360.454
ma.L2          0.0773    209.688      0.000      1.000    -410.903     411.057
ma.L3         -0.9803    259.275     -0.004      0.997    -509.150     507.189
ma.L4         -0.2028     49.178     -0.004      0.997     -96.589      96.184
ar.S.L6        0.2976      0.098      3.023      0.003       0.105       0.491
ar.S.L12       0.3624      0.069      5.245      0.000       0.227       0.498
ar.S.L18      -0.3067      0.073     -4.215      0.000      -0.449      -0.164
ar.S.L24       0.3589      0.058      6.168      0.000       0.245       0.473
ma.S.L6       -0.4157      0.187     -2.217      0.027      -0.783      -0.048
ma.S.L12       0.0630      0.135      0.468      0.640      -0.201       0.327
ma.S.L18       0.3813      0.141      2.712      0.007       0.106       0.657
ma.S.L24      -0.3847      0.151     -2.543      0.011      -0.681      -0.088
sigma2       186.8911   4.53e+04      0.004      0.997   -8.86e+04     8.9e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 1.49
Prob(Q):                              0.95   Prob(JB):                         0.47
Heteroskedasticity (H):               0.91   Skew:                             0.21
Prob(H) (two-sided):                  0.78   Kurtosis:                         2.58
===================================================================================

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

Figure 20 SARIMA Summary for (2,1,4) (4,0,4,6)

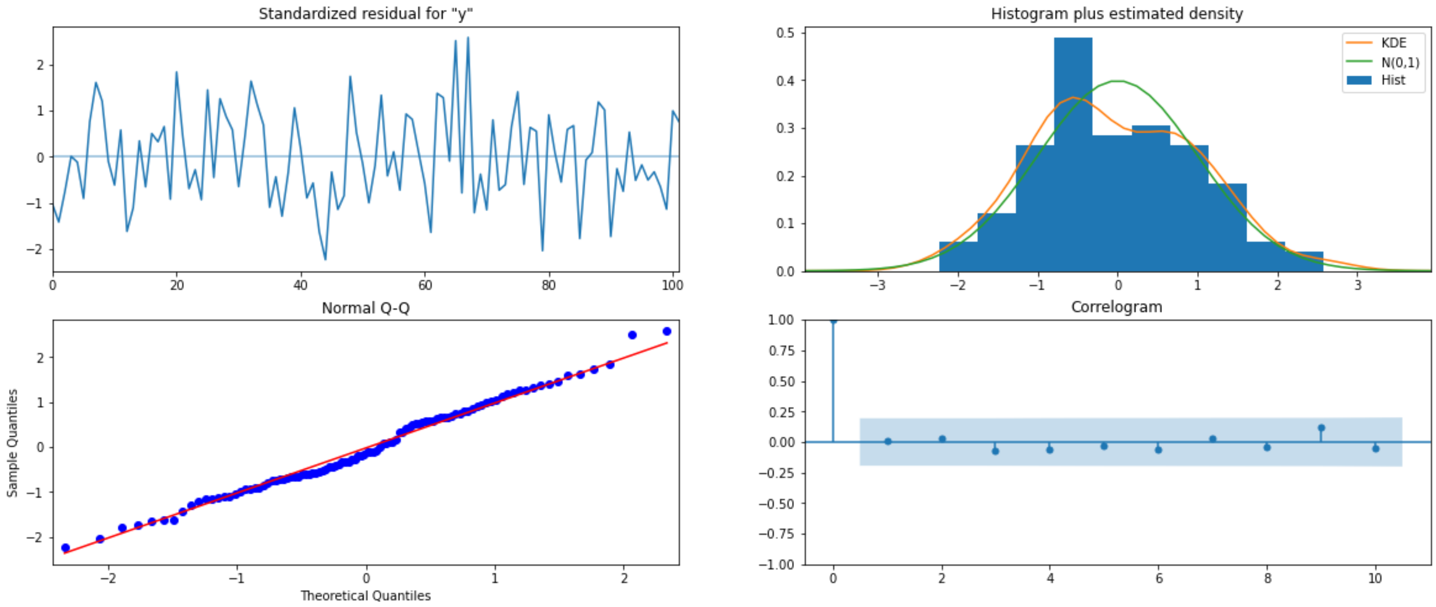


Figure 21 Diagnostic Plot – SARIMA

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

MANUAL ARIMA:

An ARIMA model consists of the Auto-Regressive (AR) part and the Moving Average (MA) part after we have made the Time Series stationary by taking the correct degree/order of differencing. The AR order is selected by looking at where the PACF plot cuts-off (for appropriate confidence interval bands) and the MA order is selected by looking at where the ACF plots cuts-off (for appropriate confidence interval bands). The correct degree or order of difference gives us the value of 'd' while the 'p' value is for the order of the AR model and the 'q' value is for the order of the MA model. This is the Box-Jenkins methodology for building the ARIMA models.

The ACF & PACF plots for the differenced data are as below.

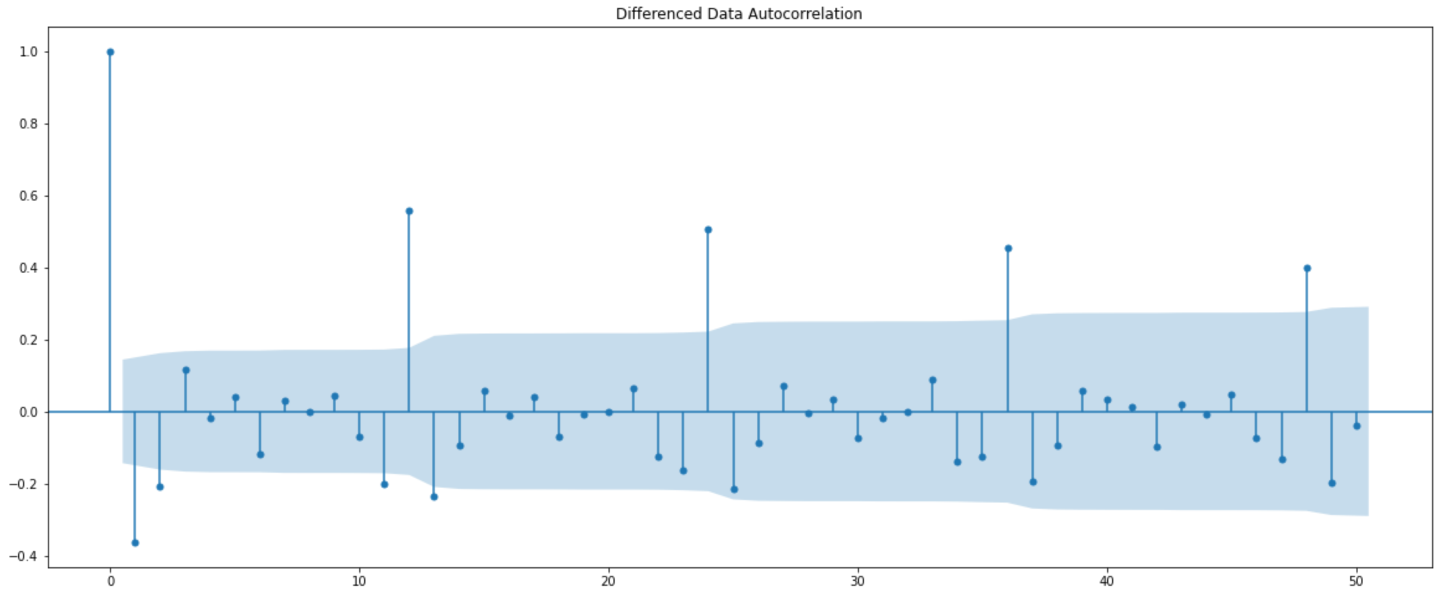


Figure 22 ACF Plot

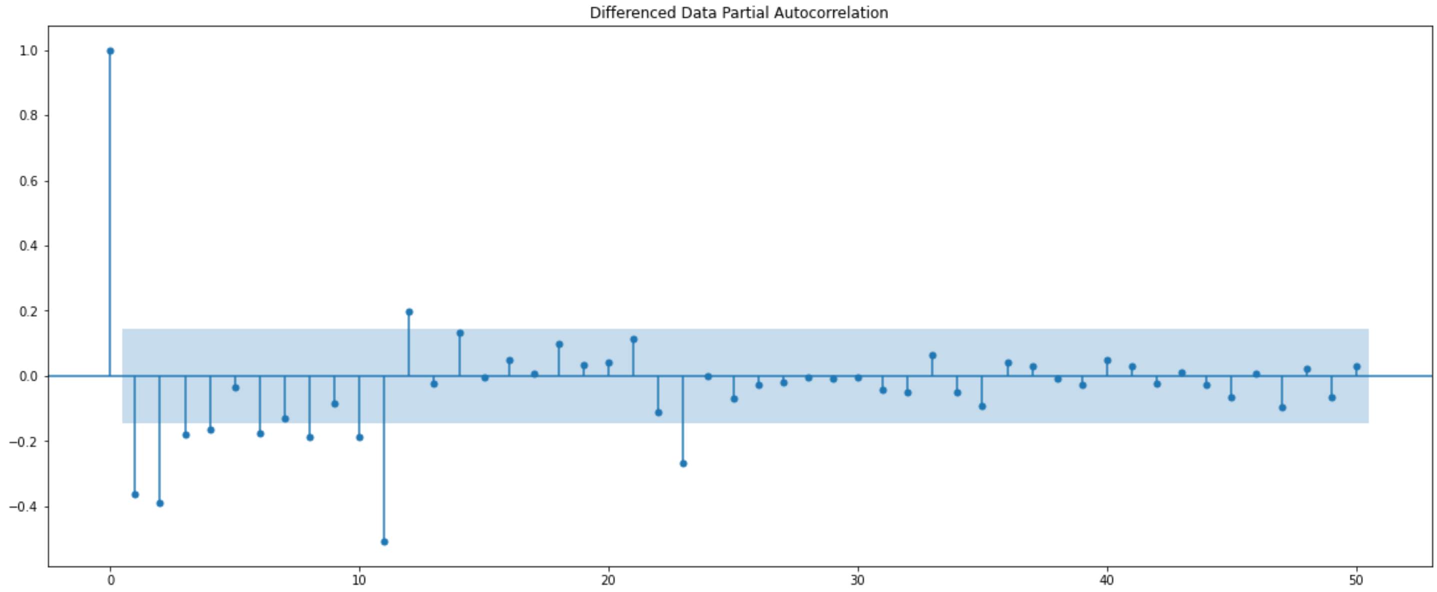


Figure 23 PACF Plot

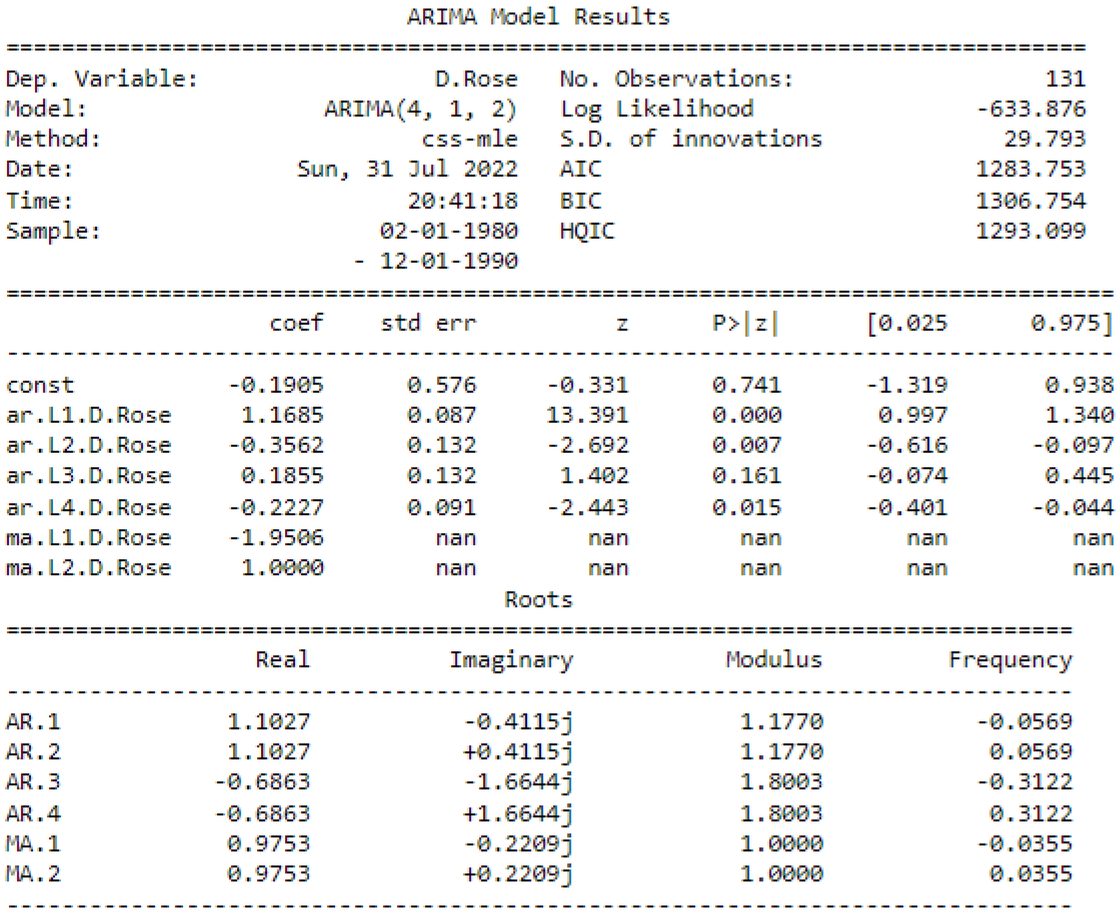
Here, we have taken alpha=0.05.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.

By looking at the above plots, we can say that the PACF plot cuts-off at lag 4 and ACF plot cuts-off at lag 2.

Hence, p = 4 & q = 2 with d = 1.

On building the ARIMA model for the above parameters, the summary is displayed as in the below figure.



```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -633.876
Method:                       css-mle   S.D. of innovations             29.793
Date:                Sun, 31 Jul 2022   AIC                           1283.753
Time:                        20:41:18   BIC                           1306.754
Sample:                    02-01-1980   HQIC                          1293.099
                         - 12-01-1990
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -0.1905      0.576     -0.331      0.741      -1.319       0.938
ar.L1.D.Rose   1.1685      0.087     13.391      0.000       0.997       1.340
ar.L2.D.Rose  -0.3562      0.132     -2.692      0.007      -0.616      -0.097
ar.L3.D.Rose   0.1855      0.132      1.402      0.161      -0.074       0.445
ar.L4.D.Rose  -0.2227      0.091     -2.443      0.015      -0.401      -0.044
ma.L1.D.Rose  -1.9506        nan        nan        nan         nan         nan
ma.L2.D.Rose   1.0000        nan        nan        nan         nan         nan
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1027           -0.4115j            1.1770           -0.0569
AR.2            1.1027           +0.4115j            1.1770            0.0569
AR.3           -0.6863           -1.6644j            1.8003           -0.3122
AR.4           -0.6863           +1.6644j            1.8003            0.3122
MA.1            0.9753           -0.2209j            1.0000           -0.0355
MA.2            0.9753           +0.2209j            1.0000            0.0355
-----------------------------------------------------------------------------
```

Figure 24 Manual ARIMA Summary

On prediction to the length of test data and evaluating the Test RMSE, it can be seen that the value of RMSE is 33.56.

MANUAL SARIMA:

Similar to the ARIMA model, we can follow the Box-Jenkins method over here as well to decide the ‘p’, ‘q’, ‘P’ and ‘Q’ values. For deciding the ‘P’ and ‘Q’ values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of ‘F’ and see where these cut-offs (for appropriate confidence interval bands). The ACF & PACF plots for the differenced data are as below.

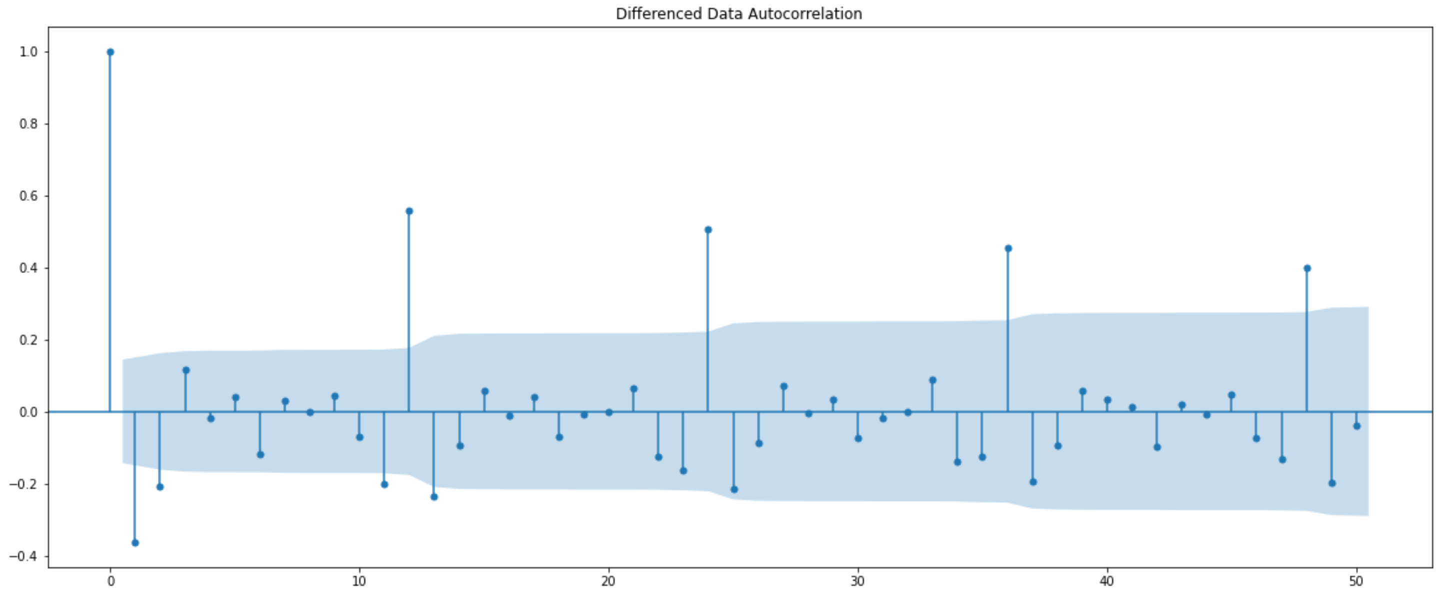


Figure 25 ACF Plot

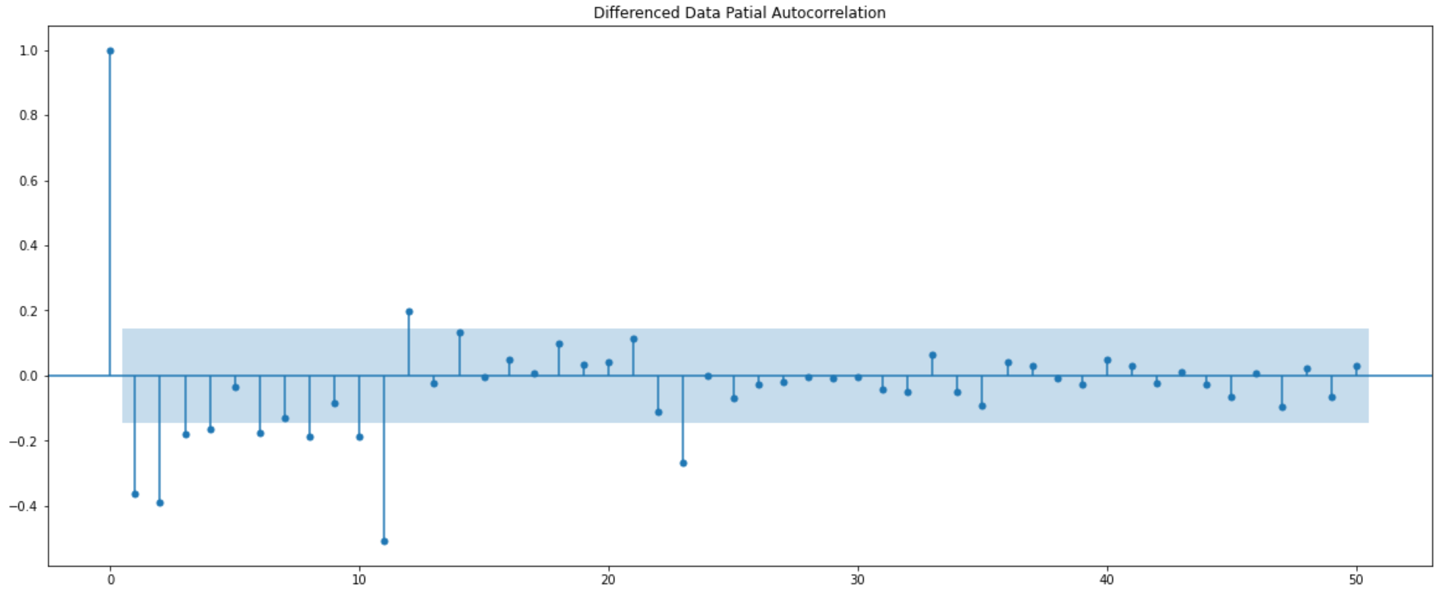
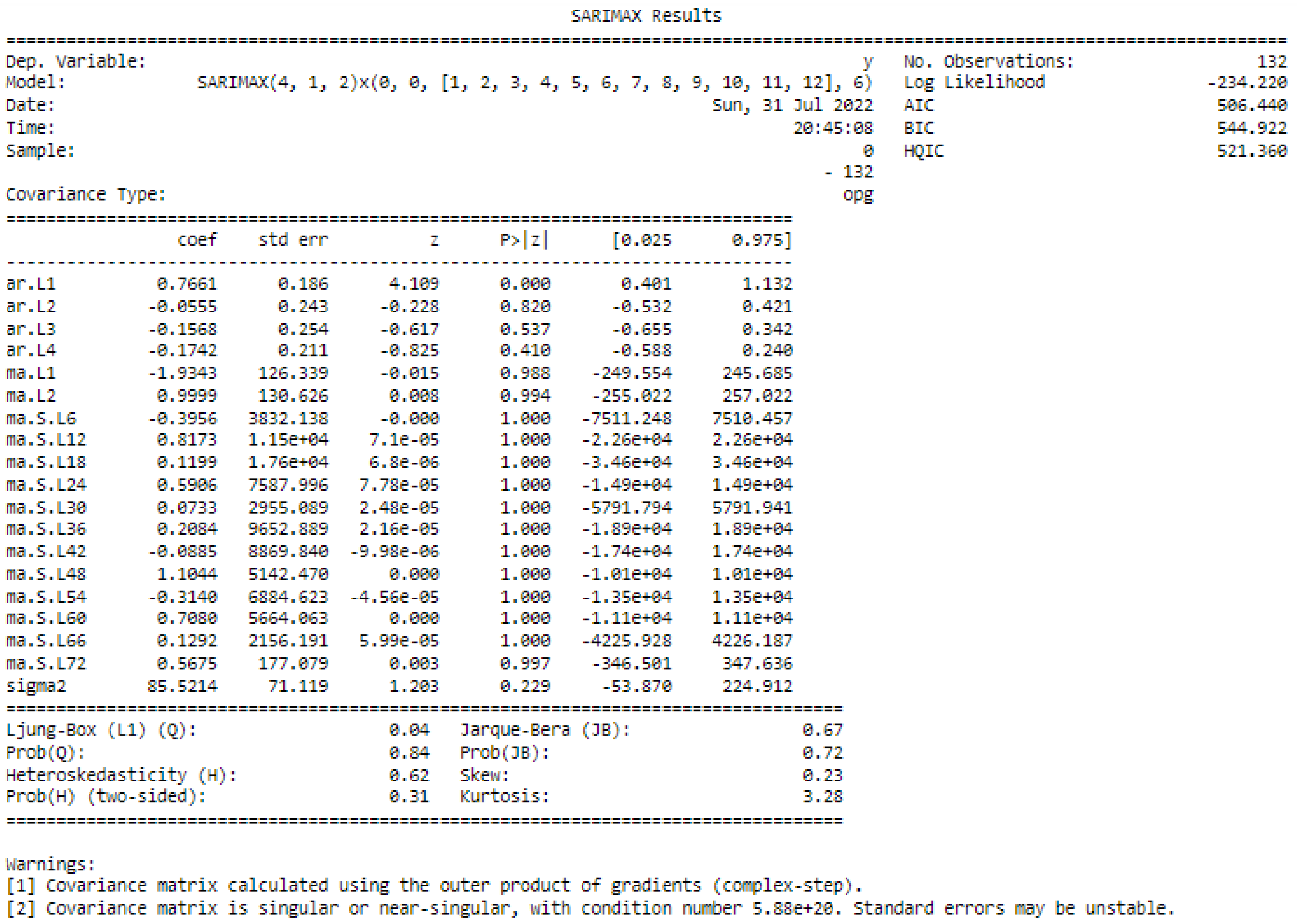


Figure 26 PACF Plot

Here from ACF plot we can see that there is a seasonal pattern at every 12th lag. The plot is cut off at every 12th lag. Hence, Q = 12. From PACF plot we can see that there is no specific seasonal pattern at every 12th lag or any other lag. Hence, P = 0. D=0, as the stationarity has been established after differencing in the ARIMA model. Also, F can be tested at both 6.

The model summary at a seasonality of 6 is as shown in the figure below.



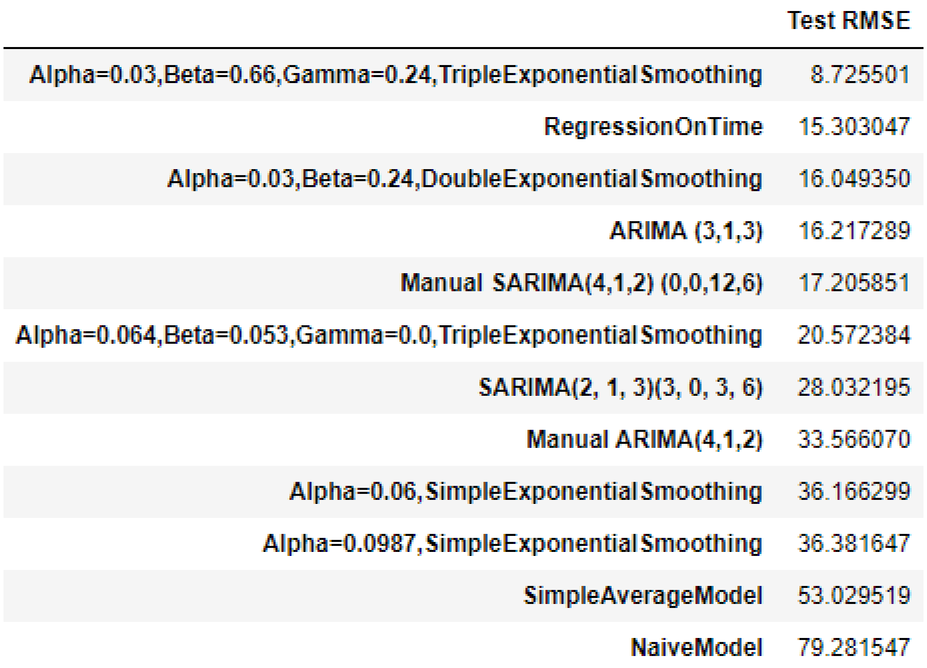
```
                                               SARIMAX Results
==================================================================================================================
Dep. Variable:                                                     y   No. Observations:                      132
Model:             SARIMAX(4, 1, 2)x(0, 0, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], 6)   Log Likelihood      -234.220
Date:                                               Sun, 31 Jul 2022   AIC                                506.440
Time:                                                       20:45:08   BIC                                544.922
Sample:                                                            0   HQIC                               521.360
                                                               - 132
Covariance Type:                                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7661      0.186      4.109      0.000       0.401       1.132
ar.L2         -0.0555      0.243     -0.228      0.820      -0.532       0.421
ar.L3         -0.1568      0.254     -0.617      0.537      -0.655       0.342
ar.L4         -0.1742      0.211     -0.825      0.410      -0.588       0.240
ma.L1         -1.9343    126.339     -0.015      0.988    -249.554     245.685
ma.L2          0.9999    130.626      0.008      0.994    -255.022     257.022
ma.S.L6       -0.3956   3832.138     -0.000      1.000   -7511.248    7510.457
ma.S.L12       0.8173   1.15e+04    7.1e-05      1.000   -2.26e+04    2.26e+04
ma.S.L18       0.1199   1.76e+04    6.8e-06      1.000   -3.46e+04    3.46e+04
ma.S.L24       0.5906   7587.996   7.78e-05      1.000   -1.49e+04    1.49e+04
ma.S.L30       0.0733   2955.089   2.48e-05      1.000   -5791.794    5791.941
ma.S.L36       0.2084   9652.889   2.16e-05      1.000   -1.89e+04    1.89e+04
ma.S.L42      -0.0885   8869.840  -9.98e-06      1.000   -1.74e+04    1.74e+04
ma.S.L48       1.1044   5142.470      0.000      1.000   -1.01e+04    1.01e+04
ma.S.L54      -0.3140   6884.623  -4.56e-05      1.000   -1.35e+04    1.35e+04
ma.S.L60       0.7080   5664.063      0.000      1.000   -1.11e+04    1.11e+04
ma.S.L66       0.1292   2156.191   5.99e-05      1.000   -4225.928    4226.187
ma.S.L72       0.5675    177.079      0.003      0.997    -346.501     347.636
sigma2        85.5214     71.119      1.203      0.229     -53.870     224.912
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 0.67
Prob(Q):                              0.84   Prob(JB):                         0.72
Heteroskedasticity (H):               0.62   Skew:                             0.23
Prob(H) (two-sided):                  0.31   Kurtosis:                         3.28
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 5.88e+20. Standard errors may be unstable.
```

Figure 27 Manual SARIMA @ Seasonality – 6

The Test RMSEs for both e models – Seasonality @ 6 is 17.20.

1.8 <u>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.</u>

The Test RMSEs of all the models are combined in a single data frame and sorted by the RMSEs value in the ascending order. The Data Frame is as below.



	Test RMSE
Alpha=0.03,Beta=0.66,Gamma=0.24,TripleExponentialSmoothing	8.725501
RegressionOnTime	15.303047
Alpha=0.03,Beta=0.24,DoubleExponentialSmoothing	16.049350
ARIMA (3,1,3)	16.217289
Manual SARIMA(4,1,2) (0,0,12,6)	17.205851
Alpha=0.064,Beta=0.053,Gamma=0.0,TripleExponentialSmoothing	20.572384
SARIMA(2, 1, 3)(3, 0, 3, 6)	28.032195
Manual ARIMA(4,1,2)	33.566070
Alpha=0.06,SimpleExponentialSmoothing	36.166299
Alpha=0.0987,SimpleExponentialSmoothing	36.381647
SimpleAverageModel	53.029519
NaiveModel	79.281547

Table 4 Test RMSEs of various models

It can be seen that the Triple Exponential Smoothing model with alpha – 0.03, Beta = 0.66 & Gamma = 0.24 has the least RMSE and hence considered the best model for forecasting in this case.

1.9 Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

From the Table, we can see that the Holt's Winter model performs the best. Extending this model to the full data set, we can see that the RMSE is 18.68. On further predicting for the next 12 months, it can be seen that the same pattern is repeated. These predicted may/ may not happen in actual. Hence, it is always suggested to build this forecast on a confidence interval of 95%. Here we are taking the multiplier to be 1.96 as we want to plot with respect to 95% confidence intervals. The plot with confidence band is as below.

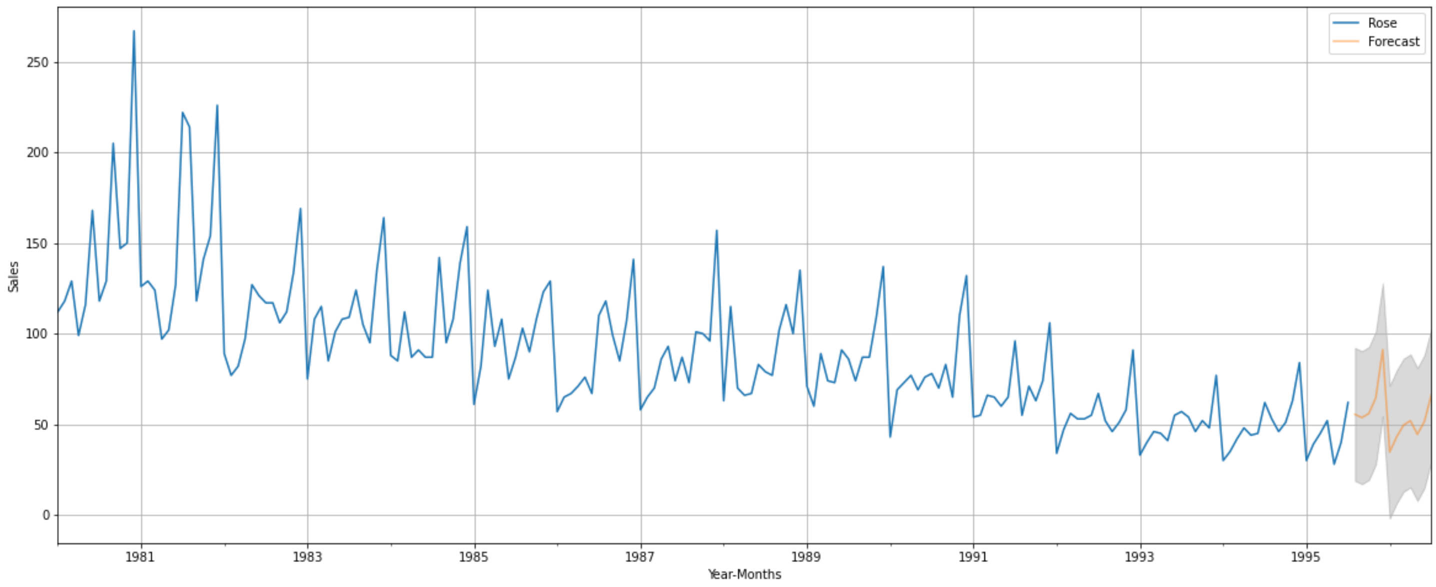


Figure 28 Prediction TS with CI

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

The model thus built on the full data has predictions of 12 months that are similar to the data in the past.

There is seasonality in December as usual in the past.

Following are the insights and suggestions for the company to take up for the upcoming sales:

- The average sale of Rose wine is seen to be between 50 to 80 for each month in the first half of the year. Hence, the company need to stock a minimum of 70~100 numbers every month in the first half of the year. However, this trend is seen to be decreasing with time. Hence, the optimum amount would be stocking 60~90 nos.
- The sale is seen to pick up the pace and peaks in the month of December. However, due to the decreasing trend, we can go with stocking the average of the last 3 years sales in the December (90 units).
- Though the average sales are decreasing in Rose wine, the company should ensure stocks a little more than average in the second half of the year, so that it doesn't miss out on customers.
- The company could also consider looking for some additional ingredient or an additional method through which it can reattract customers to boom the Rose wine sales.