Sep 2021 | PG-DSBA-Online

PROJECT REPORT
TIME SERIES

SUBMITTED BY
DEV TRIPATHI

Contents

List of Figures

Problem 1

Sparkling Sales Dataset Analysis

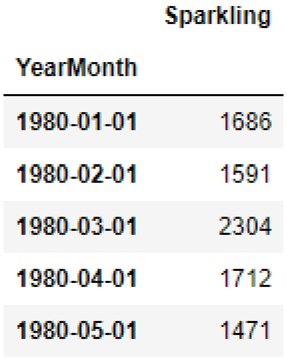
Information about dataset

The dataset contains the data of different types of wine sales in 20th century. The data for both that is Sparkling and Rose wine sales is provided by the same company. As an analyst at ABC Estate Wines, we are tasked to analyze and forecast wine sales in the 20th century.

Task 1. Read the data as an appropriate Time Series data and plot the data.

EDA

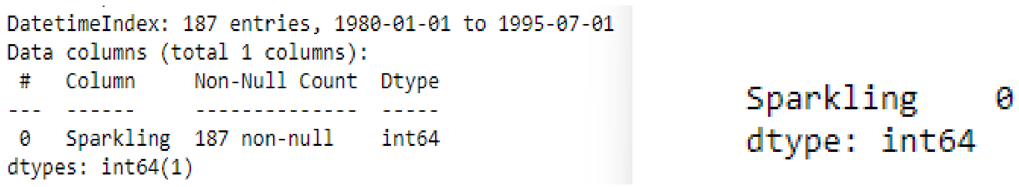
Sample of Dataset-



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

Figure 1: Sample of Sparkling Dataset

Variable Information-



```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
```

```
Sparkling    0
dtype: int64
```

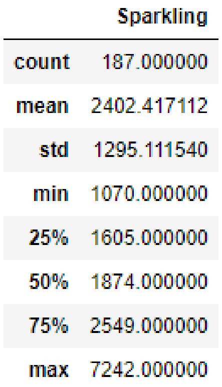
Figure 2: Feature Info and null values count of Sparkling Dataset

- The dataset contains a total of 187 entries and 2 features including the datetime feature.
- As it is shown in the list, none of the variables contains null values.
- The dataset contains values from Jan 1980 to Jul 1995 with monthly frequency.
- The only variable (named as ‘Sparkling’) represents the sales of Sparking in that month. It contains only integer type values.

Task 2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

As we have already performed null check, data types, and shape earlier we can proceed further with visualization of data.

Statistical Description of Dataset



	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

- Minimum value for the time series is 1070 and Maximum value is 7242.
- Mean value for the ‘Sparkling’ variable is 2402 and median is 1874.

Visualization of Time Series Data

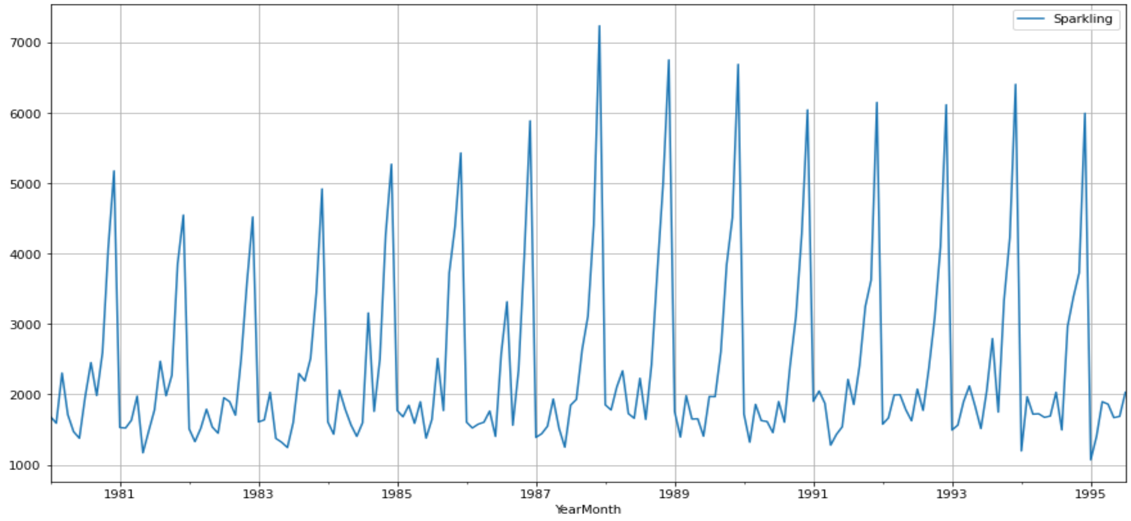


Figure 3: Plot showing the variation of Sales of Sparkling wine over the years

- As we can observe that there is seasonal component present in time series.

- If we look closely at the variations over a single year, we can say that the sales go up towards the end of the year.
- Just by looking at it we can say that there is weak or no trend at all present with in the time series. But still for verifying this statement we need to look at the decomposition of the time series data.

Further, we can look at the variation of Sales over the years. Here box plots are created to visualize the same:

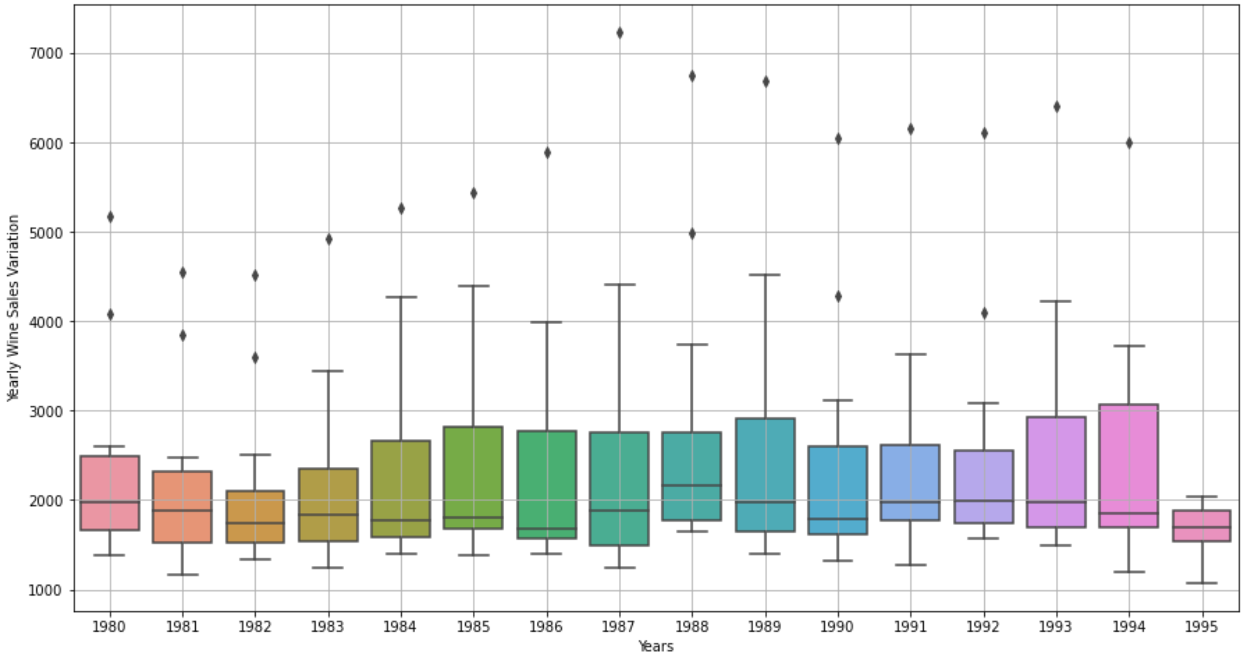


Figure 4: Box plots for visualization of variation of Sales for each year

- We can observe that over the years, for one or two months the sales are very high which are visible as outlier for each year.
- In this time series the maximum value occurred a month in the year 1987 which is 7242
- The median value of Sales is somewhere around 2000 for each year
- For years ’84 to ’87,’89 and ’94 the 75th percentile values are much higher than the 50th percentile values (median for each year).

Now, we can plot the yearly variation of sales in each month. This will give an idea of trend of time series over the years in that month:

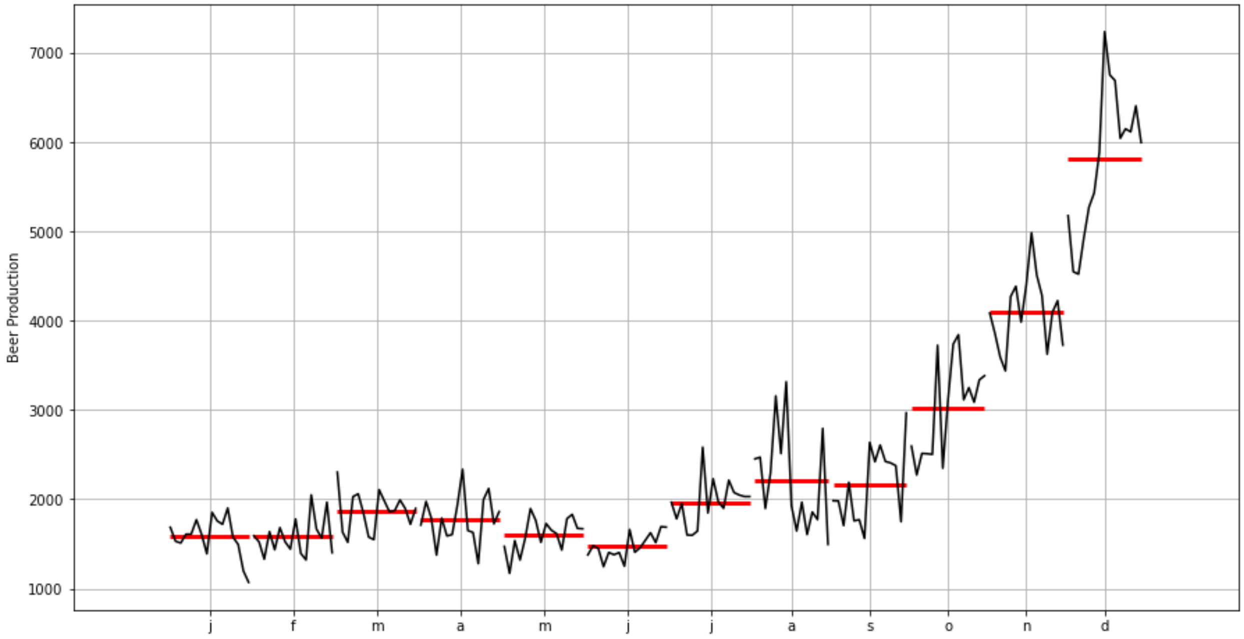
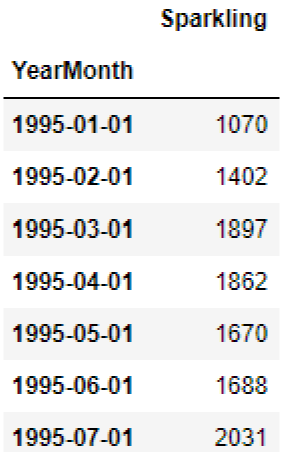


Figure 5: Plot of yearly variation of Sales over every month

- As we can observe that for the first half of each year the values are very less and after the 'July' month the Sales start go up and by the end of the year the Sales are at the peak (in a particular year)
- Also, the variation of Sales in the is first half of the year are very less compared to variation in the second half of the year.
- There is a noticeable dip present in almost each month towards the end of its yearly variation depicting that the "Sales" are going down towards the end of this time series data.



	Sparkling
YearMonth	
1995-01-01	1070
1995-02-01	1402
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

One thing to keep in mind is that the values for year 1995 are incomplete as the time series data ends on Jul 1995.

Decomposition of the time-series-

As the name suggests, decomposition simply means decomposition of time series in deterministic (trend and seasonality) and stochastic (residual) components. Decomposition for the provided time series is given below:

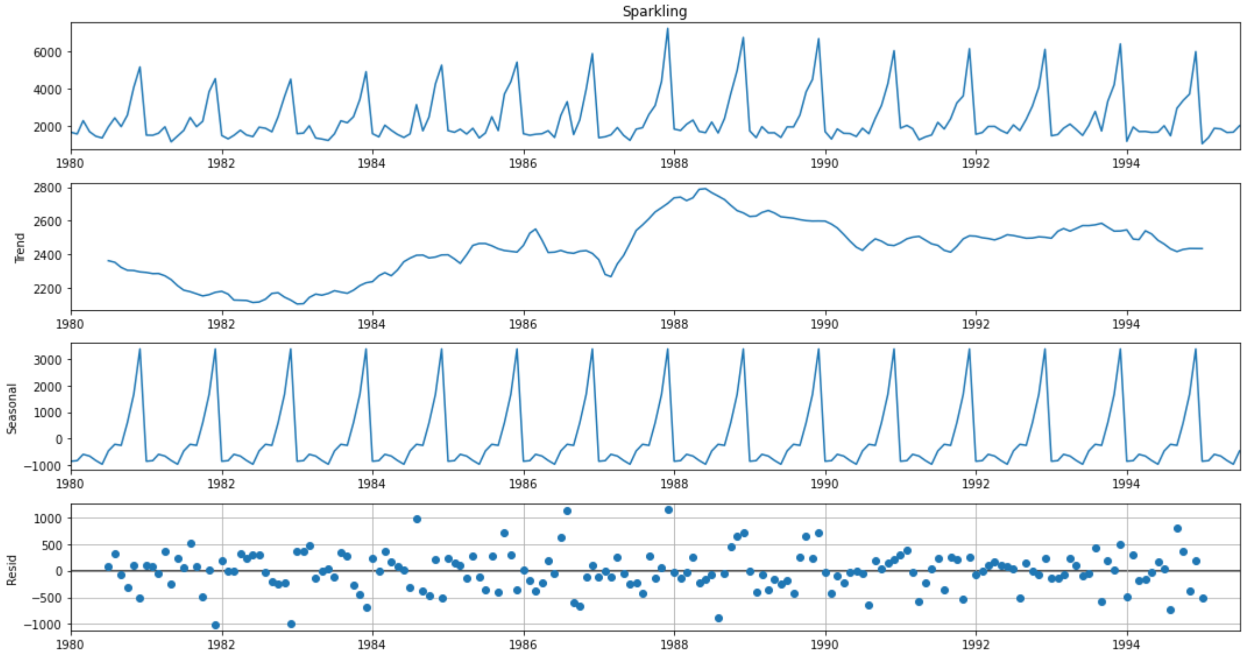


Figure 6: Decomposition of given time-series using Additive model

As we can observe that a strong seasonality component is present in the time-series. There is a steady rise from year which can be noticed There is a steady rise from the year 1980 to 1986 which can be noticed in trend after that a sudden jump in sales in the year 1987 to 1988 and after that a steady decline can be observed till year 1995. The stochastic component of time-series (residuals varying between ±1000) is high when compared with the seasonal (max. value 3000) and trend (max. value 2800) component.

For additive model, decomposition can be represented with following equation:

$$\boldsymbol{Actual\ value\ of\ time - series\ =\ Trend\ +\ Seasonality\ + Stochastic\ component}$$

Further, we can look at multiplicative model for decomposition of time series though the residuals don't seem to follow any pattern but still we can do that for cross checking:

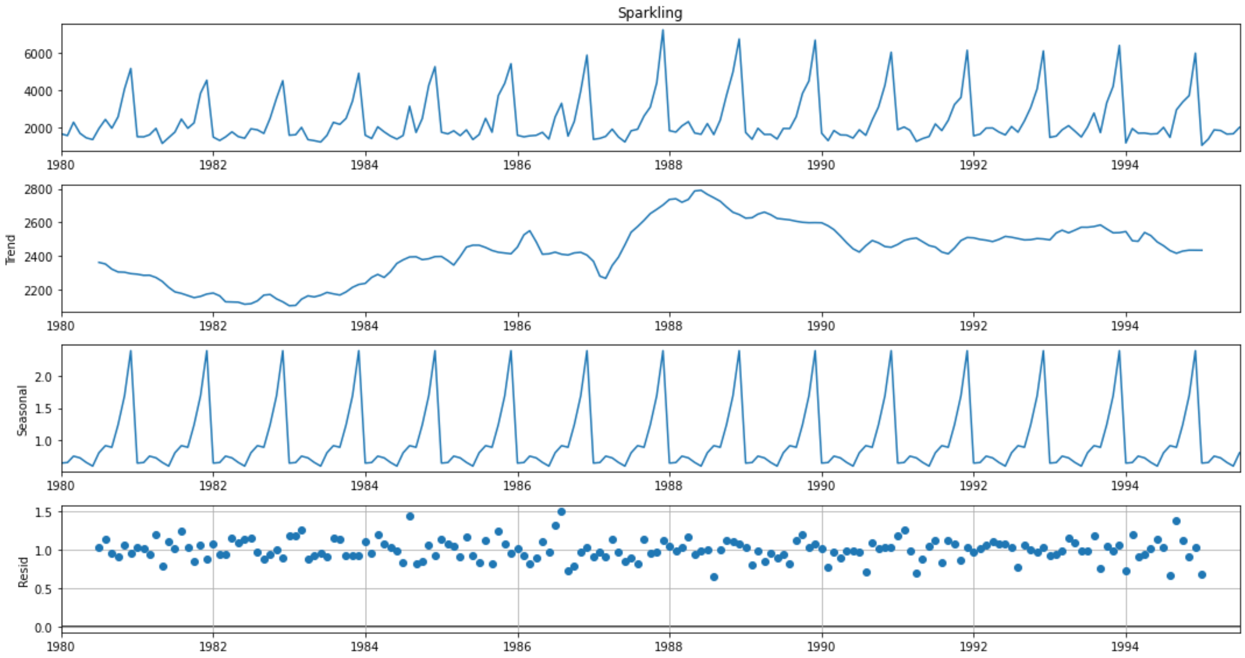


Figure 7: Decomposition of given time-series using Multiplicative model

In multiplicative models, the actual value of time-series can be obtained as:

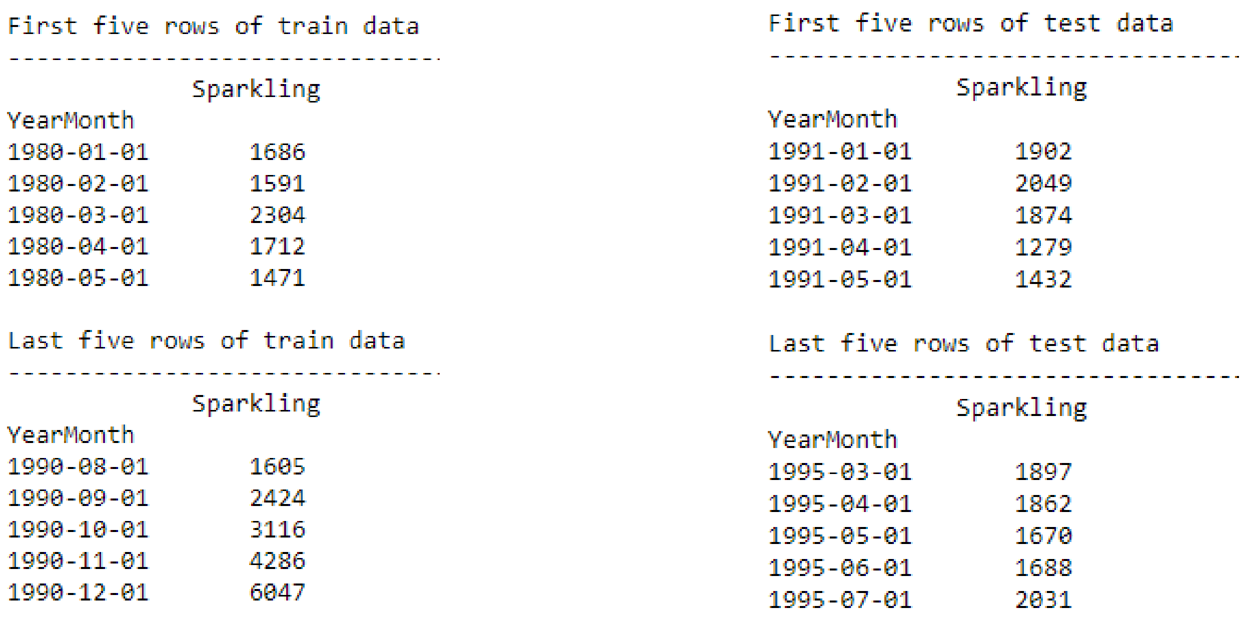
$$\boldsymbol{Actual\ value\ of\ time - series\ =\ Trend\ *\ Seasonality\ *\ Residuals}$$

As we expected, the decomposition using multiplicative model doesn't seem to perform any better as the magnitude of residual is varying from 0.6 to 1.5. But since the variation in residuals seems to be somewhat lesser than the additive model in, we can say that this model is better. Trend and seasonality component are almost same as that of additive model.

Task 3. Split the data into training and test. The test data should start from the year 1991.

Splitting the time-series in train and test set

The train and test split in a time-series model is based upon the timestamp associated with the variable. As instructed, here we have split the time-series data in train and test where test set is starting from year 1991:



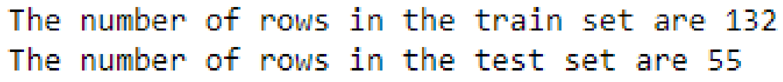
```
First five rows of train data
------------------------------
             Sparkling
YearMonth
1980-01-01        1686
1980-02-01        1591
1980-03-01        2304
1980-04-01        1712
1980-05-01        1471

Last five rows of train data
------------------------------
             Sparkling
YearMonth
1990-08-01        1605
1990-09-01        2424
1990-10-01        3116
1990-11-01        4286
1990-12-01        6047
```

```
First five rows of test data
--------------------------------
             Sparkling
YearMonth
1991-01-01        1902
1991-02-01        2049
1991-03-01        1874
1991-04-01        1279
1991-05-01        1432

Last five rows of test data
--------------------------------
             Sparkling
YearMonth
1995-03-01        1897
1995-04-01        1862
1995-05-01        1670
1995-06-01        1688
1995-07-01        2031
```

Figure 8: First and last five rows of Train and Test set



```
The number of rows in the train set are 132
The number of rows in the test set are 55
```

Figure 9: Shape of Train and Test set of Time-series

We can also plot the train and test set on the same axis to check whether the split is correct or not:

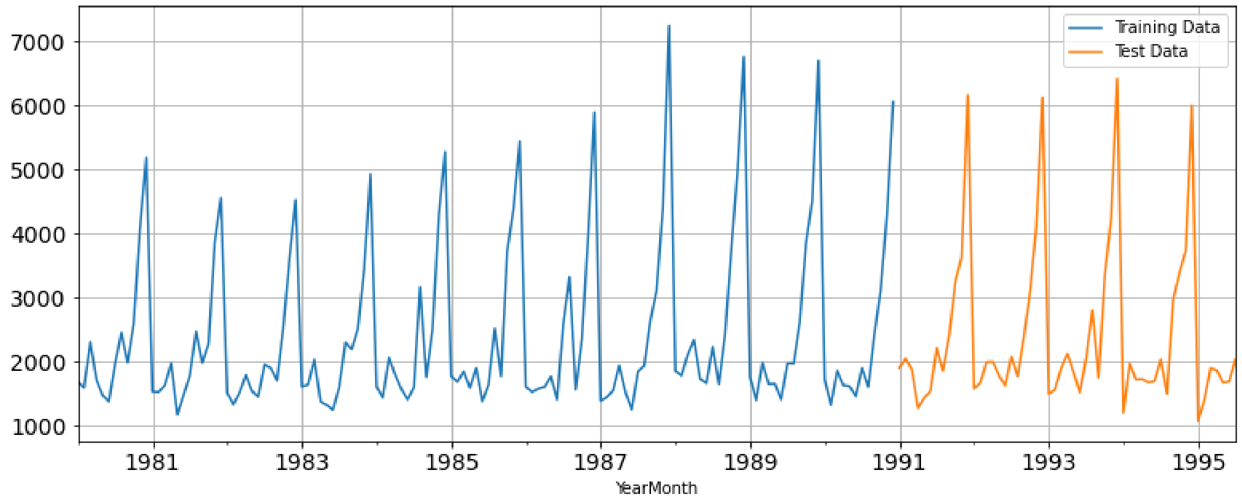


Figure 10: Plot of Train and Test data on same axis

As we can see that train and test set split is correct, hence we may proceed further to modelling part.

Task 4. Build various 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.

Modelling

As per the rubric let's start with the modelling part:

Exponential Smoothing Models

Simple Exponential Smoothing Models (SES)

SES or one-parameter exponential smoothing is applicable to time series which do not contain either of trend or seasonality.

$$\hat{Y}_{(t+1)} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots, 0 < \alpha < 1$$

where, α is the smoothing parameter for the level. In reality such a series is hard to find. This is a one- step-ahead forecast where all the forecast values are identical.

Here we have built two models for SES, one is with automated optimization and other one is optimization using grid-search method for which minimum the RMSE value on test set (unseen data) was looked for among all the models trained. The parameter obtained from automated optimization is $\alpha = 0.0496$ and from the manual grid-search method we found out that minimum RMSE value on test set was obtained for $\alpha = 0.02$.

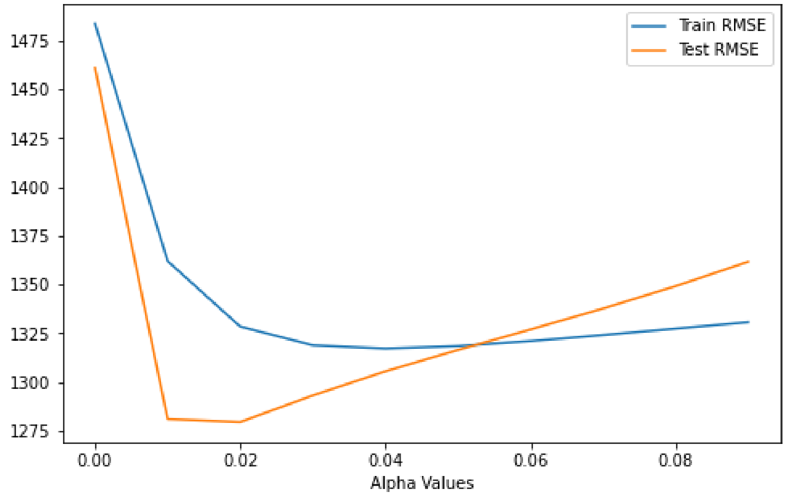


Figure 11: Plot showing variation of RMSE value with Level (α) for train and test set

For visualization of the forecast, we can plot the train and test set along with the forecast obtained from SES models:

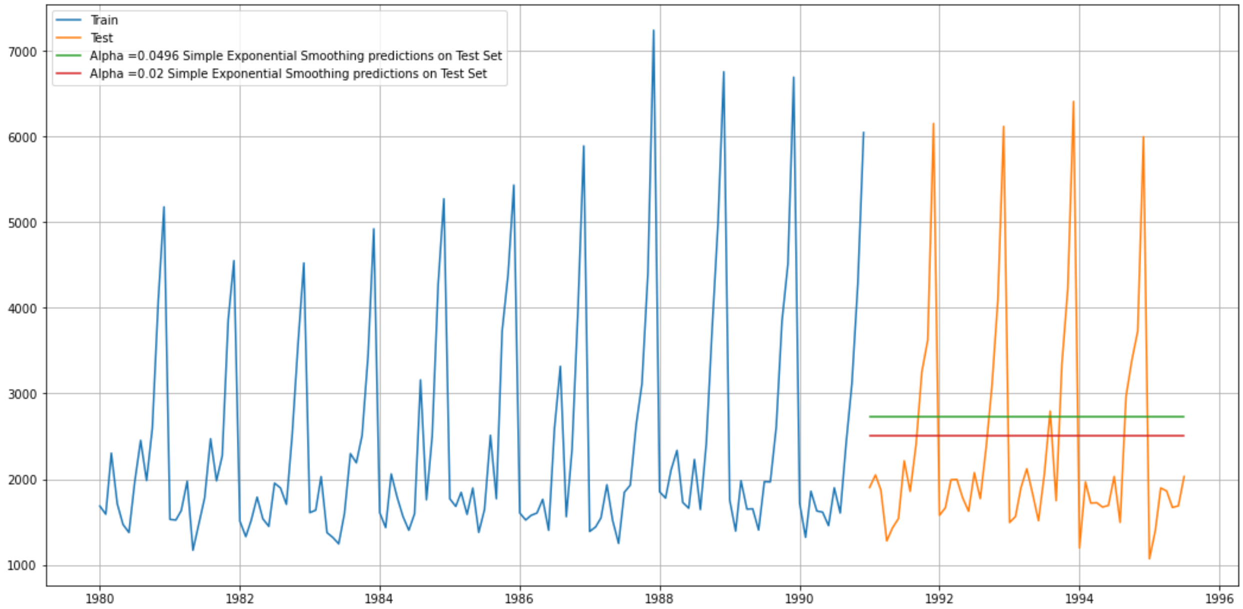


Figure 12: Plot for visualizing the forecasts obtained from SES models

Double Exponential Smoothing Models (DES)

This method is an extension of SES method, proposed by **Holt** in **1957**. Hence, this method is also known as Holt's model. This method is applicable where trend is present in the data but no seasonality. The forecast values are given as:

Forecast equation: $\hat{Y}_{t+1} = l_t + b_t$

Level Equation: $l_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1},\ 0 < \alpha\ < \ 1$

Trend Equation: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},\ 0\ <\ \beta\ <\ 1$

where, l_t is the estimate of level and b_t is the trend estimate. α is the smoothing parameter for the level and β is the smoothing parameter for trend. Now, similar to SES models we have built two models here also, one is automated optimization and the other one manual grid search model. The parameters obtained are following:

- For automated optimization model: $\alpha = 0.688$ and $\beta = 9.9 * 10^{-05}$
- For manually optimized model: $\alpha = 0.1$ and $\beta = 0.01$

For visualization of the forecast, we can plot the train and test set along with the forecast obtained from DES models:

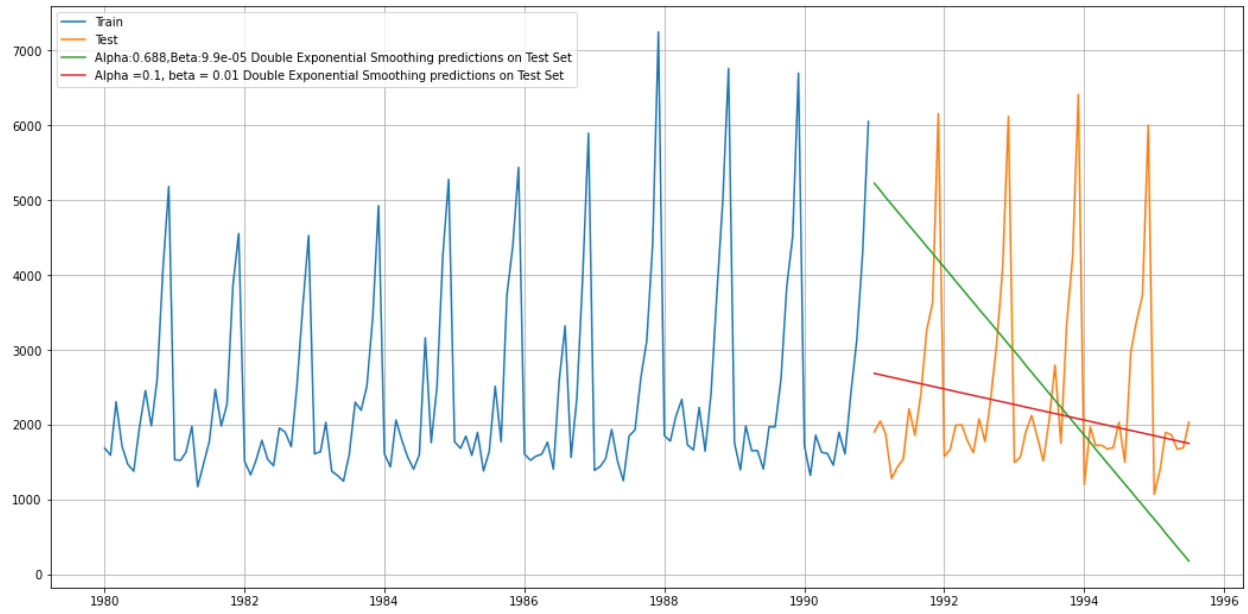


Figure 13: Plot for visualizing the forecasts obtained from DES models

Triple Exponential Smoothing Models (TES)

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

Forecast equation: $Y_{t+1} = l_t + b_t + s_{t-m(k+1)}$

Level Equation:$l_t = \alpha(Y_t - s_{t-m}) + \alpha(1-\alpha)Y_{t-1}, 0 \ < \ \alpha \ < \ 1$

Trend Equation: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, 0 \ < \ \beta \ < \ 1$

Seasonal Equation: $s_t = \gamma(Y_t - l_{t-1} - b_{t-1}) \ + (1-\gamma)s_{t-m}, 0 \ < \ \gamma \ < \ 1$

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, β and γ. Here m is frequency of the seasonality component. This is a general method and a true multi-step ahead forecast.

Similar to aforementioned models, we have built two models here also, one is automated optimization and the other one manual grid search model. The parameters obtained are following:

- For automated optimization model: $\alpha = 0.11, \beta = 0.06 \ and \ \gamma = 0.4$
- For manually optimized model: $\alpha = 0.5, \beta = 0.027 \ and \ \gamma = 0.3$

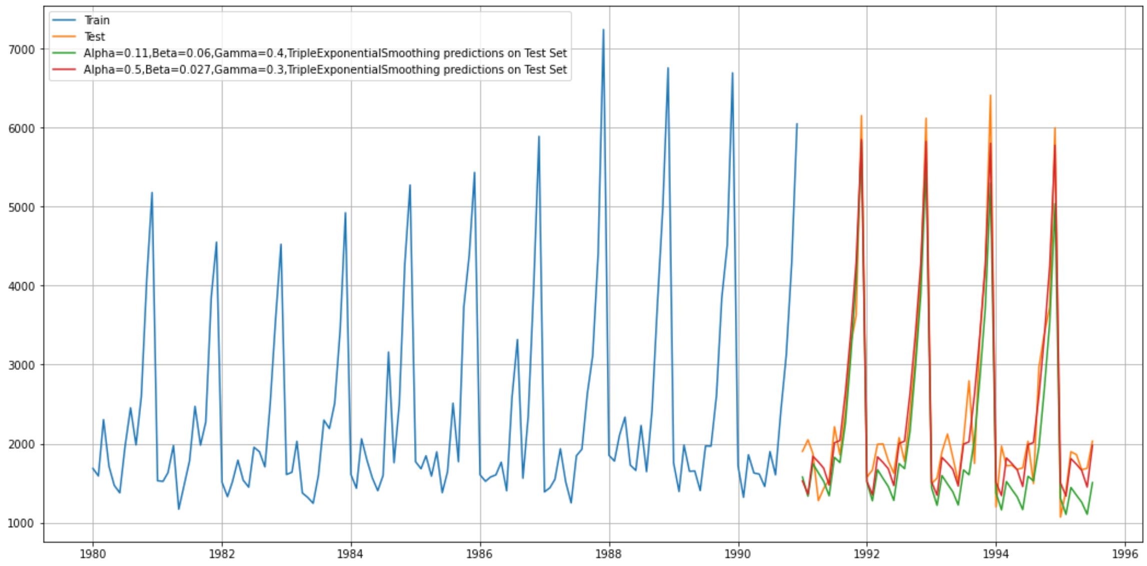


Figure 14: Plot for visualizing the forecasts obtained from TES models

As expected, we can observe that the models are doing very good job in finding correct the seasonal component and forecasting based on that for required steps (i.e., equal to length of test set) but still if we look closely, we can observe that the model with 'red' line is closer to actual values.

Linear Regression model

We cannot directly apply linear regression on a time-series data as the data has an order of occurrence, hence we must retain the order while fitting the data to regression model. Which means another variable has to be created which retains this order while fitting in the model. Once this is done, we can good to proceed further with fitting the values. After building the model we can forecast using this model for the steps size of 55 (i.e., length of test set). Following plot can be obtained by plotting the forecast against the actual series:

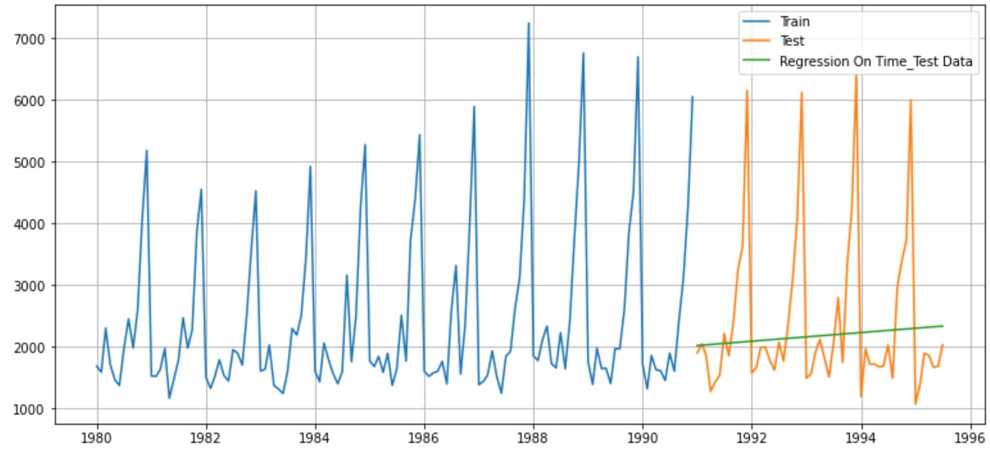


Figure 15: Plotting the values obtained after forecasting for length of test set

Naïve Forecast Model

As the name suggest, this is probably the simplest model to build. It simply states that tomorrow's value will exactly be equal to today's value. Hence, the last value of train set will become the prediction for any duration. This can be easily understood from the following plot:

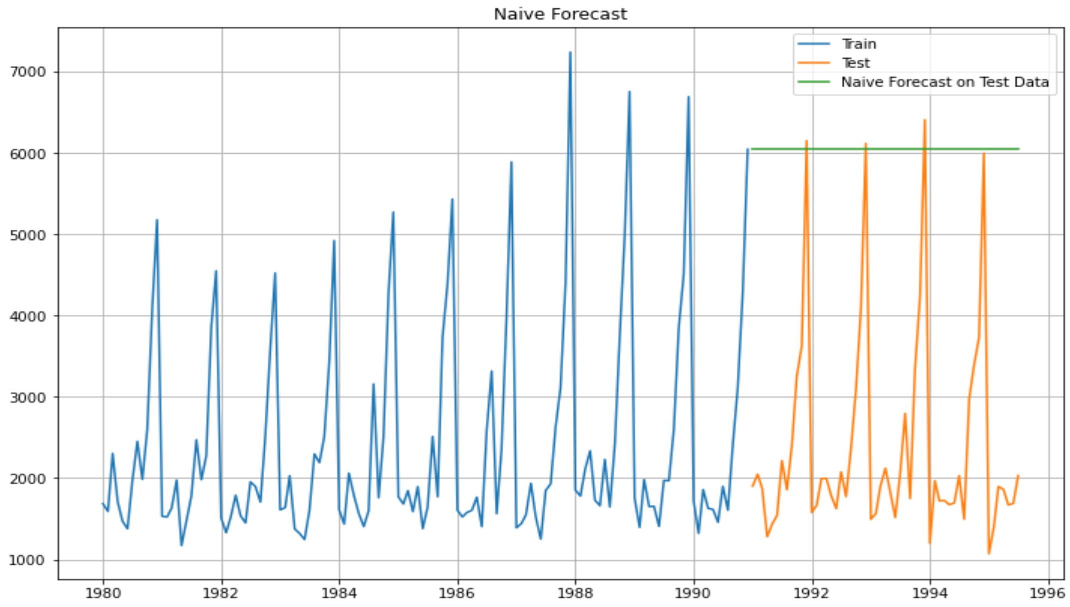


Figure 16: Plot of Naive forecast for test set time duration (steps)

Simple Average Model

In this model we simply average the values provided for a duration and forecast that same value over the required time duration.

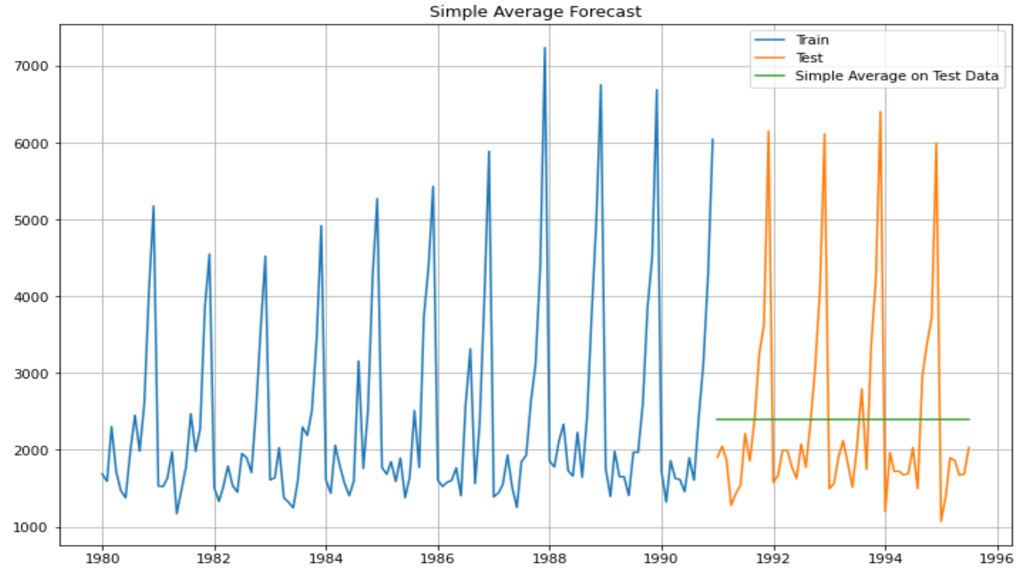


Figure 17: Plot of Simple average model over the test set time-duration

Task 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.

Check for Stationarity of Time Series

Stationary process: A process is said to be stationary if its mean and variance are constant over a period of time and, the correlation between the two time periods depends only on the distance or lag between the two periods. Mathematically, let Y_t be a time series with these properties:

Mean: $E(Y_t) = \mu$

Variance: $Var(Y_t) = E(Y_t - \mu)^2 = \sigma^2$

Correlation: $\rho_k = E\left[\frac{(Y_t-\mu)(Y_{t+k}-\mu)}{(\sigma_t \sigma_{t+k})}\right]$

Where ρ_k is the correlation (or auto-correlation) at lag k between the values of $Y_t\ and\ Y_{t+k}$.

Augmented Dickey-Fuller test: The Augmented Dickey-Fuller test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis in a simple form for the ADF test is:

- **Null hypothesis ($\boldsymbol{H_0}$):** The Time Series has a unit root and is thus non-stationary.
- **Alternate Hypothesis ($\boldsymbol{H_1}$):** The Time Series does not have a unit root and is thus stationary.

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

Now, after performing the ADF test for the whole dataset we obtained following results:

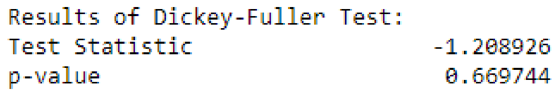
```
ADF test statistic is -1.798
ADF test p-value is 0.7055958459932417
```

Since, the p-value is coming out to be 0.71 which is greater than our significance level α which is 0.05 the null hypothesis can not be rejected. Thus, at the significance level of 0.05 we can say time series is non-stationary. Further we have taken the difference of time series with 1 lag for achieving the stationarity goal and again performed ADF test on this series, results of which are shown below:

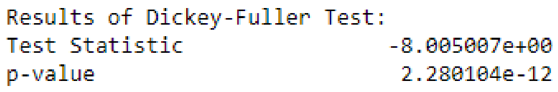
```
ADF test statistic is -44.912
ADF test p-value is 0.0
```

As we can see that the p-value (0.0) is less than the significance level (i.e., 0.05); we can reject the null hypothesis at this significance level. Hence, we can say that differenced time series with 1 lag is stationary.

Since we will be building the model based on train data only, we should also check for the stationarity of train set. The ADF test results on train set and differenced train set with 1 lag are given below:



```
Results of Dickey-Fuller Test:
Test Statistic                 -1.208926
p-value                         0.669744
```



```
Results of Dickey-Fuller Test:
Test Statistic               -8.005007e+00
p-value                       2.280104e-12
```

Figure 18: Results obtained from ADF test on Train set (left) and differenced series with 1 lag of train set (right)

As it is obvious from the results obtained that the original train set of time series is non-stationary since the p-value is 0.66 which is greater than the significance level 0.05. On the other hand, the differenced time series with one lag of train set is stationary. Hence, for building the SARIMA models we will be using the 'd' parameter as 1 since we did the differentiation only one time to make the series stationary.

Task 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.

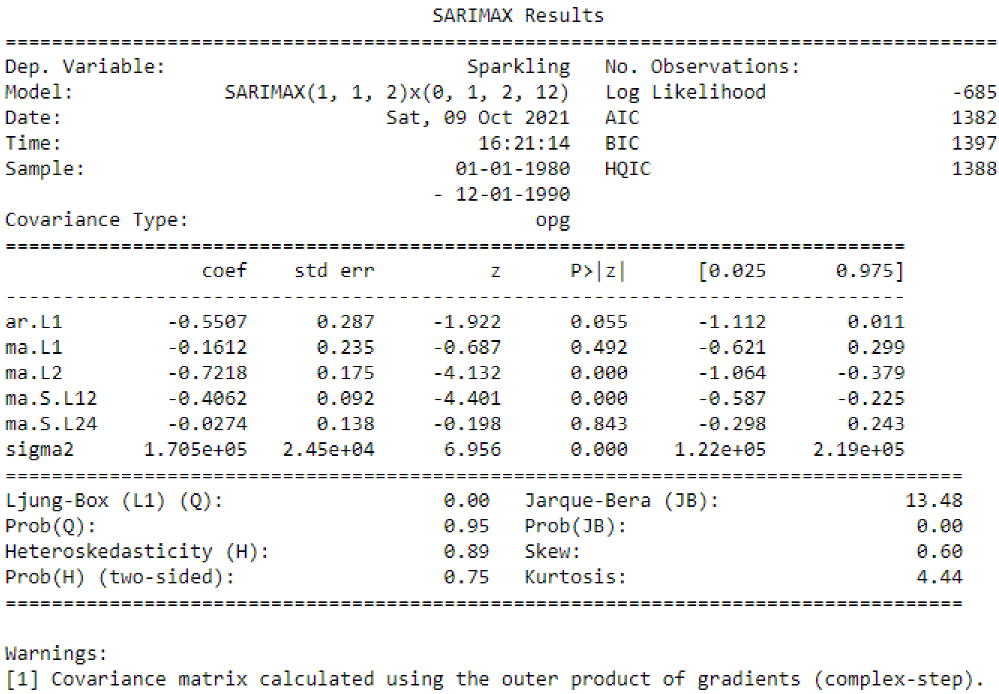
SARIMA Models

Seasonal ARIMA models are more complex models with seasonal adjustments. Since, these models are used when time series data has significant seasonality, we can use these models as our time series data contains significant seasonality.

The most general form of seasonal ARIMA is $ARIMA(p, d, q) * ARIMA(P, D, Q)[m]$, where P, D, Q are defined as seasonal AR component, seasonal difference and seasonal MA component respectively. And, 'm' represents the frequency (time interval) at which the data is observed.

As we saw from the visualizations that our time series data has yearly frequency, we will be putting 'm' as 12.

Now, for building the automated version of the SARIMA model, we had to pick the parameter's range of values first and after that we had to pick the best model based on the Akaike Information Criteria (AIC) built using those parameters. After that we used this model to forecast for 55 steps (i.e., length of test set) and calculated RMSE value for the model which came out to be 382.57.



```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                              Sparkling   No. Observations:              132
Model:             SARIMAX(1, 1, 2)x(0, 1, 2, 12)   Log Likelihood               -685.174
Date:                            Sat, 09 Oct 2021   AIC                          1382.348
Time:                                    16:21:14   BIC                          1397.479
Sample:                                01-01-1980   HQIC                         1388.455
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5507      0.287     -1.922      0.055      -1.112       0.011
ma.L1         -0.1612      0.235     -0.687      0.492      -0.621       0.299
ma.L2         -0.7218      0.175     -4.132      0.000      -1.064      -0.379
ma.S.L12      -0.4062      0.092     -4.401      0.000      -0.587      -0.225
ma.S.L24      -0.0274      0.138     -0.198      0.843      -0.298       0.243
sigma2      1.705e+05   2.45e+04      6.956      0.000    1.22e+05    2.19e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                13.48
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.89   Skew:                             0.60
Prob(H) (two-sided):                  0.75   Kurtosis:                         4.44
===================================================================================

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

Figure 19: Results summary obtained for (1,1,2) (0,1,2)12 SARIMA

Task 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.

Now, for selecting the parameters manually we must first find out the values of parameters (p,d,q and seasonal parameters P,D,Q) by looking at Auto Correlation and Partial Auto Correlation plots:

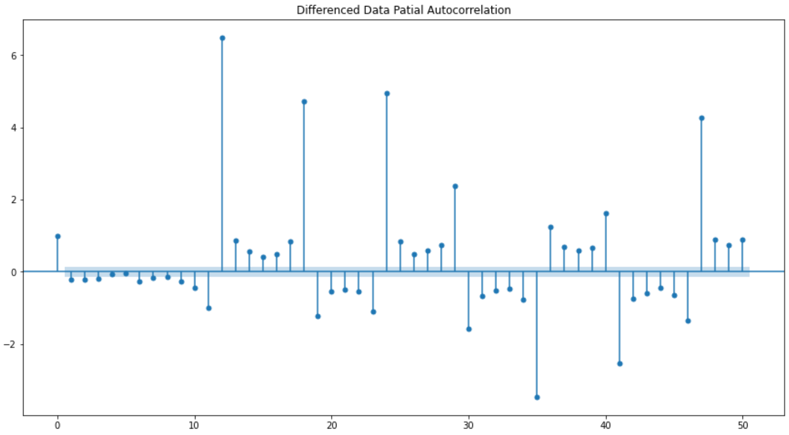


Figure 20:Auto Correlation plot for difference series with 1 lag

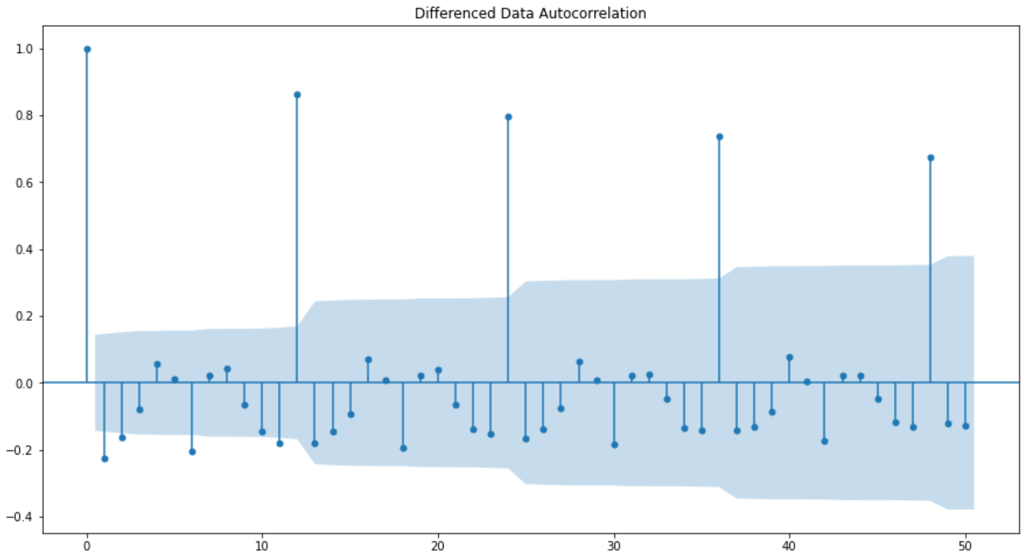


Figure 21: Partial Auto Correlation plot for difference series with 1 lag

Now, we can go ahead and take first seasonal differencing of the original series.

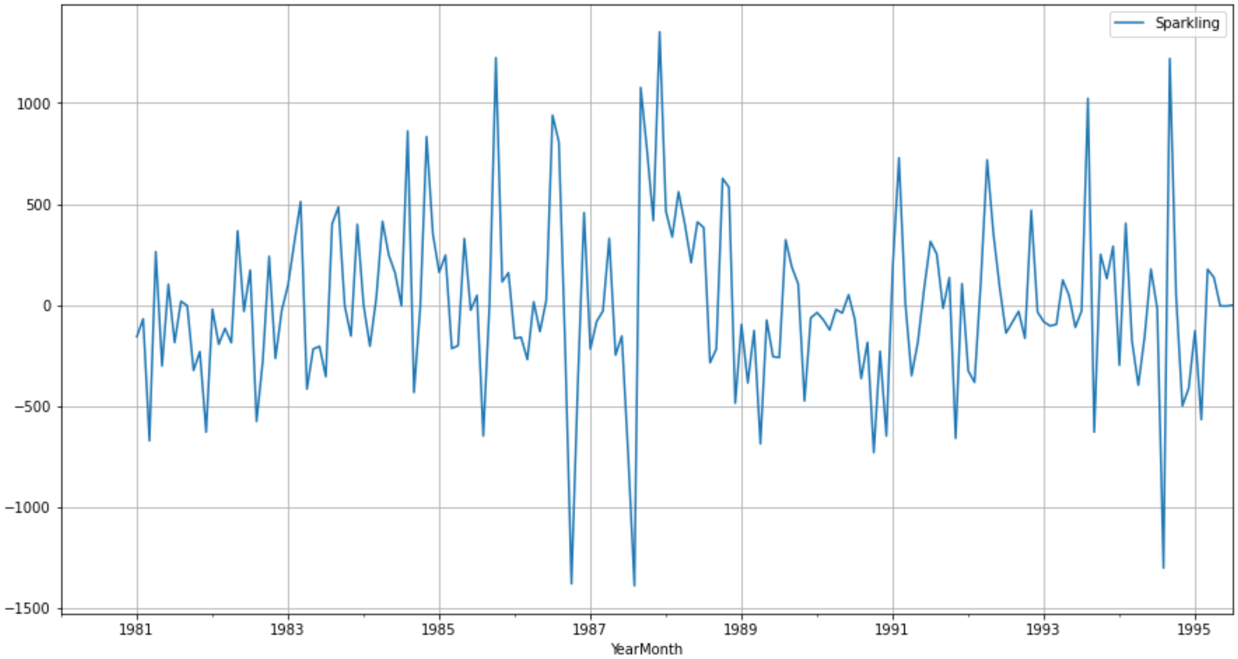


Figure 22: Plot of first seasonal difference series

As there is still some trend is visible in the first seasonal difference series, we can look at the first difference series of this series:

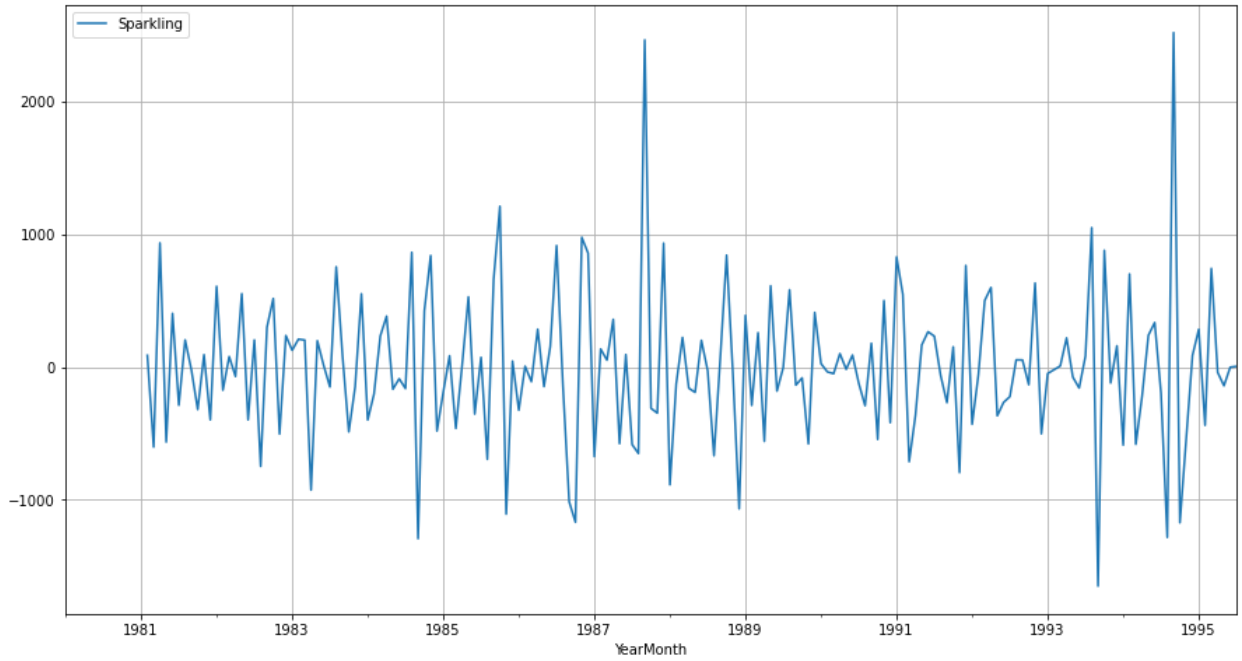
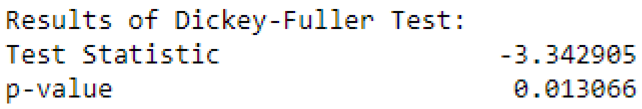


Figure 23: Plot of first seasonal difference first difference series

Now we see that there is almost no trend present in the data. Seasonality component only is present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.



```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
```

As the p-value is < α i.e., 0.05, we can reject the null hypothesis which means this time series is stationary.

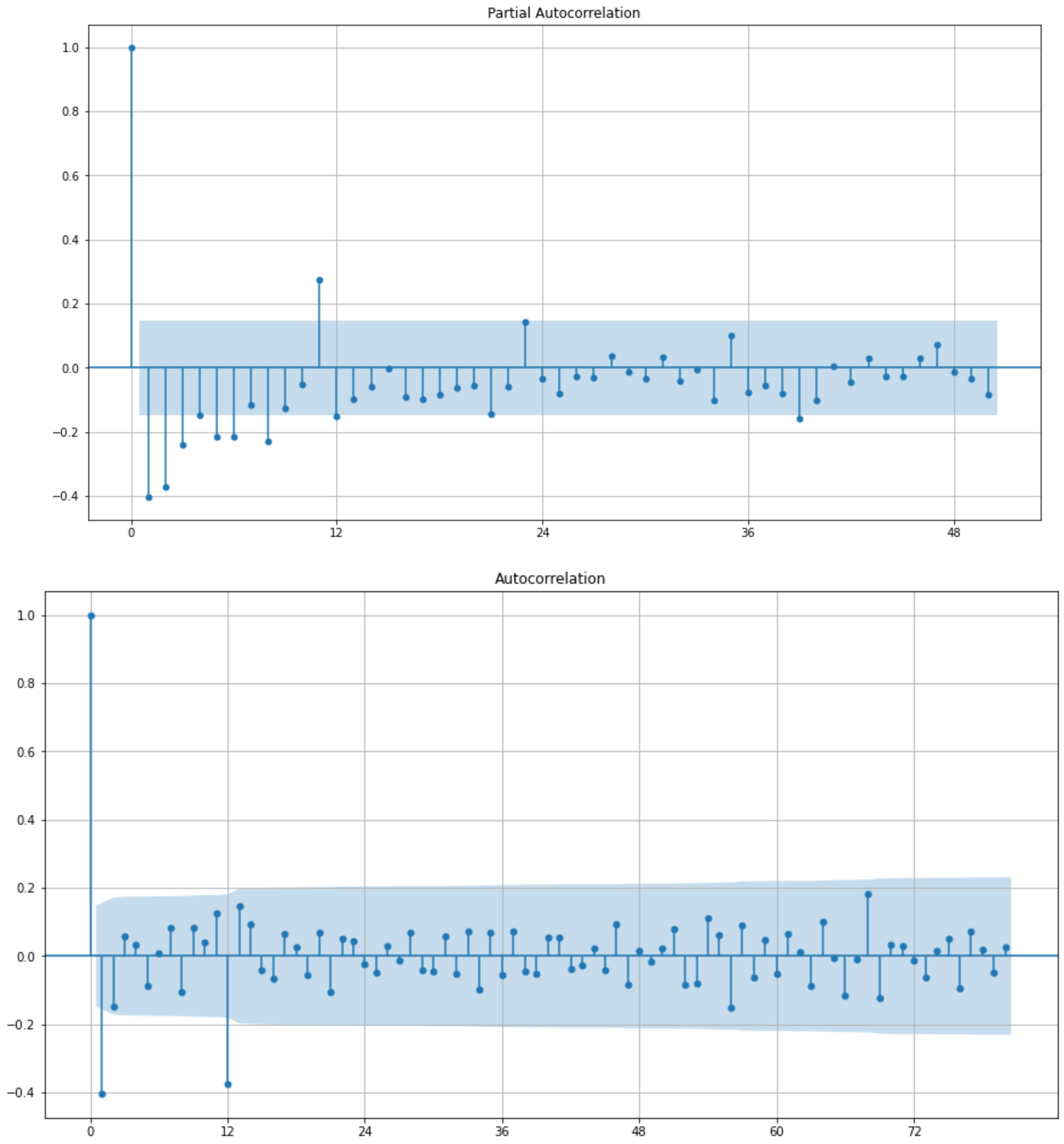
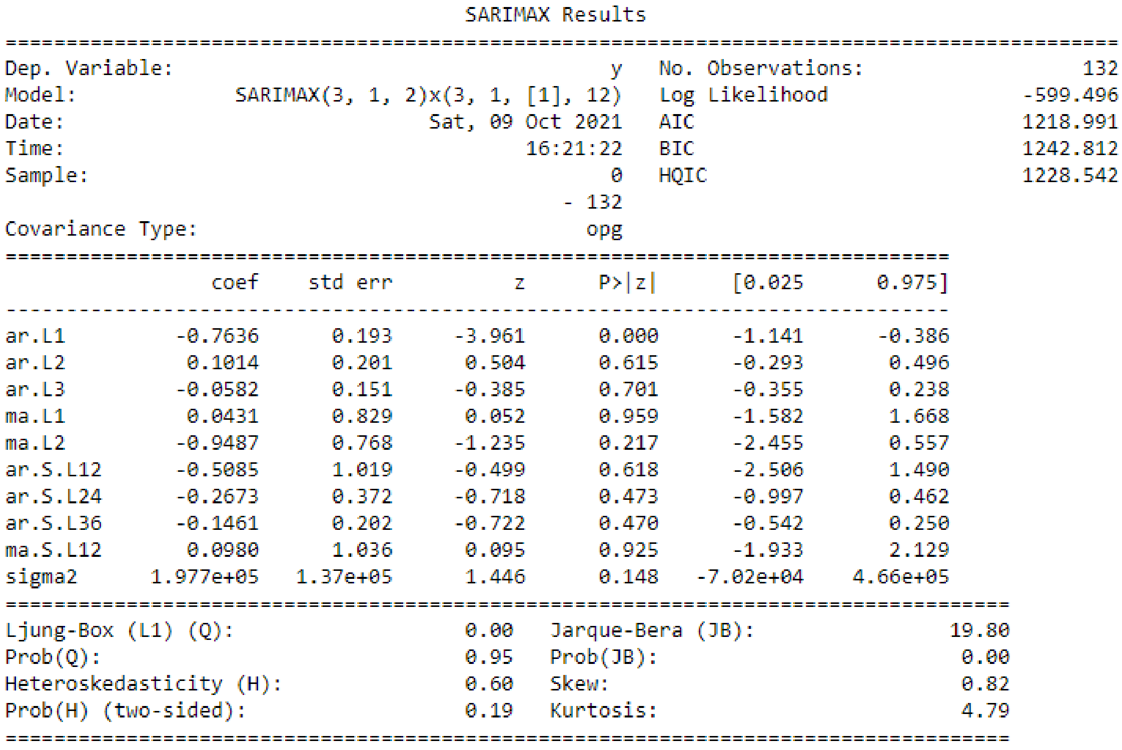


Figure 24: Plot of first difference and seasonal first difference series, PACF and ACF

Here, we have taken alpha=0.05. We are going to take the seasonal period as 12. We will keep the p=3 and q=2 (obtained from ACF and PACF plots of first difference series). The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).

Now, as we can see from the figure 24 that the PACF plot cuts off after 1 seasonal lag and ACF plot cuts off after 1 seasonal lag also. Hence, we can say that the value of P is 3 and Q is 1. Using these parameters, we have built the model results of which is given below:



```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 132
Model:             SARIMAX(3, 1, 2)x(3, 1, [1], 12)   Log Likelihood                -599.496
Date:                             Sat, 09 Oct 2021   AIC                           1218.991
Time:                                     16:21:22   BIC                           1242.812
Sample:                                          0   HQIC                          1228.542
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7636      0.193     -3.961      0.000      -1.141      -0.386
ar.L2          0.1014      0.201      0.504      0.615      -0.293       0.496
ar.L3         -0.0582      0.151     -0.385      0.701      -0.355       0.238
ma.L1          0.0431      0.829      0.052      0.959      -1.582       1.668
ma.L2         -0.9487      0.768     -1.235      0.217      -2.455       0.557
ar.S.L12      -0.5085      1.019     -0.499      0.618      -2.506       1.490
ar.S.L24      -0.2673      0.372     -0.718      0.473      -0.997       0.462
ar.S.L36      -0.1461      0.202     -0.722      0.470      -0.542       0.250
ma.S.L12       0.0980      1.036      0.095      0.925      -1.933       2.129
sigma2      1.977e+05   1.37e+05      1.446      0.148   -7.02e+04    4.66e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                19.80
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               0.60   Skew:                             0.82
Prob(H) (two-sided):                  0.19   Kurtosis:                         4.79
===================================================================================
```

Figure 25:Results summary of (3,1,2) (3,1,1,12) SARIMA model

As we can see that the AIC value for the model is coming out to be 1218.991. The distribution of residuals is close to normal as it can be observed from Jarque-Bera probability value. Ljung-Box test probability is also high which means residuals are independent. The RMSE value obtained for this model was 330.90.

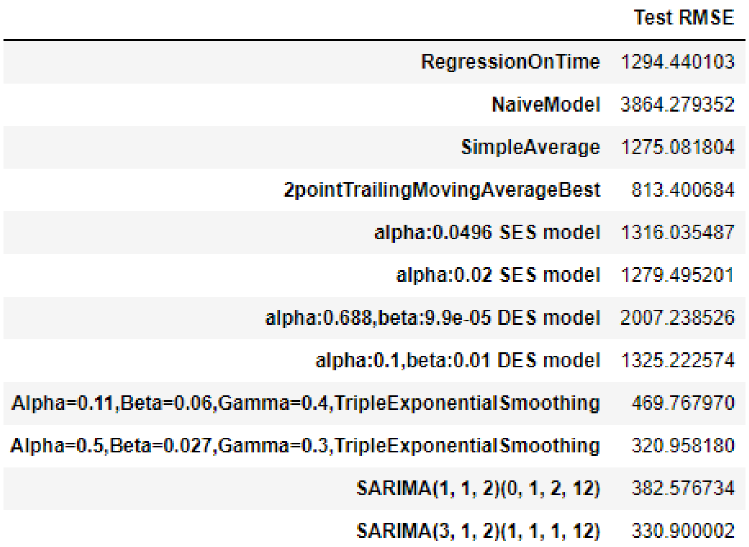
Task 8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Results

Now, as have built several models for the same, we can compare these models based on the performance on test set. The performance metric chosen for the same is Root Mean Squared Error Value which can be computed as:

$$RMSE = \sqrt{E\left(\hat{Y} - Y\right)^2} = \sqrt{\frac{\Sigma\left(\hat{Y} - Y\right)^2}{N}}$$

The RMSE values for the different models built are given in following figure:



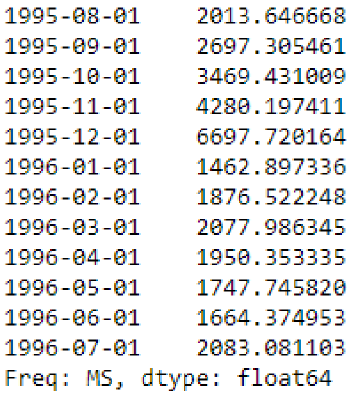
	Test RMSE
RegressionOnTime	1294.440103
NaiveModel	3864.279352
SimpleAverage	1275.081804
2pointTrailingMovingAverageBest	813.400684
alpha:0.0496 SES model	1316.035487
alpha:0.02 SES model	1279.495201
alpha:0.688,beta:9.9e-05 DES model	2007.238526
alpha:0.1,beta:0.01 DES model	1325.222574
Alpha=0.11,Beta=0.06,Gamma=0.4,TripleExponentialSmoothing	469.767970
Alpha=0.5,Beta=0.027,Gamma=0.3,TripleExponentialSmoothing	320.958180
SARIMA(1, 1, 2)(0, 1, 2, 12)	382.576734
SARIMA(3, 1, 2)(1, 1, 1, 12)	330.900002

Figure 26: Data Frame containing different models built & corresponding RMSE value obtained

Task 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.

Based on this analysis, we found out that the Triple Exponential Smoothing model with parameters $\alpha = 0.5, \beta = 0.027\ and\ \gamma = 0.3$ gave least RMSE value on the test set (i.e., 320.96). Hence, we can conclude that this model is the most optimized one among the rest. Now, for forecasting into the future, we must train another model with these parameters on the whole data else we will be losing the data in test set if we used the previous model built using only the train data.

After building this final model, we can obtain the forecast for the required number of months (12 months). Forecast values obtained are given as:



```
1995-08-01    2013.646668
1995-09-01    2697.305461
1995-10-01    3469.431009
1995-11-01    4280.197411
1995-12-01    6697.720164
1996-01-01    1462.897336
1996-02-01    1876.522248
1996-03-01    2077.986345
1996-04-01    1950.353335
1996-05-01    1747.745820
1996-06-01    1664.374953
1996-07-01    2083.081103
Freq: MS, dtype: float64
```

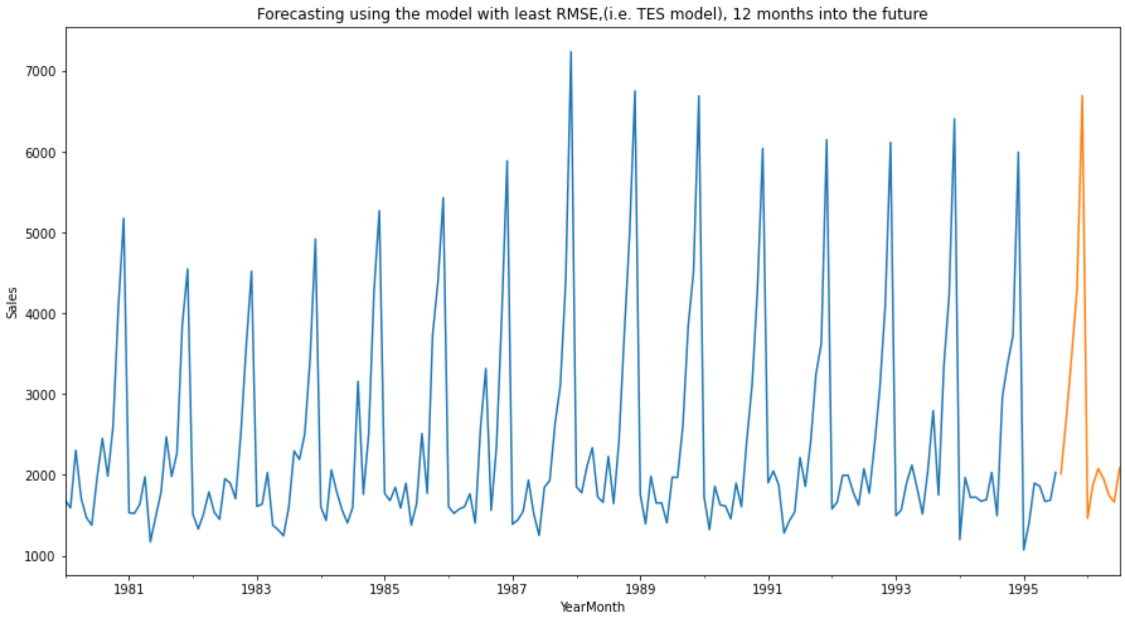


Figure 27: Plot of forecasted values using the final model for 12 months

Task 10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Inferences and Suggestions

Following inferences can be drawn from this analysis:

- As it can be observed that huge spikes are present in plot of time series towards each year's end.
- One of the reasons could be the holiday season which happens every at the end of each year.
- Also, a huge jump is observed in the trend in year 1988, after which a steady decrease is observed in the trend.
- Based on this analysis we can say that the Triple Exponential Smoothing model performed best among various models built.
- Although the Sales pattern is very similar every year, the company should try to find the possible reasons for the steady decline in the sales which is observed after year 1988.
- As it is obvious from the seasonal patterns that the sales goes up towards the end of the year, suitable business strategies should be laid out like availability of product should be insured for the increased demand and discount & offers can also be provided to attract more customers.

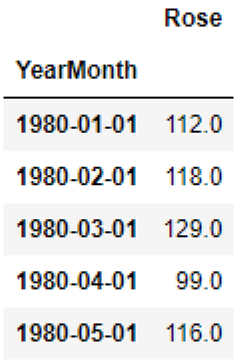
Problem 2

Rose Wine Sales Dataset Analysis

Task 11. Read the data as an appropriate Time Series data and plot the data.

EDA

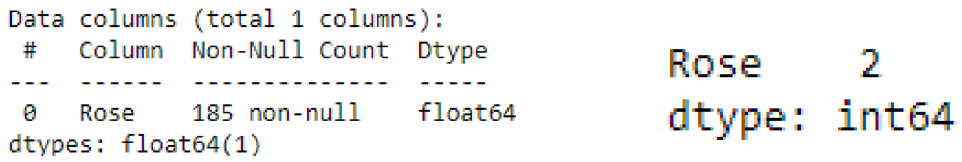
Sample of Dataset-



	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

Figure 28: Sample of Rose Wine Dataset

Variable Information-



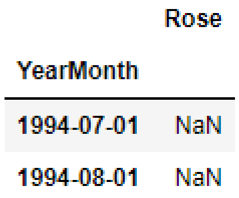
```
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    185 non-null    float64
dtypes: float64(1)
```

```
Rose     2
dtype: int64
```

Figure 29: Feature Info and null values count of Sparkling Dataset

- The dataset contains a total of 187 entries and 2 features including the datetime feature.
- As it is shown in the list, 2 null values are there in Rose column.
- The dataset contains values from Jan 1980 to Jul 1995 with monthly frequency.
- The only variable (named as ‘Rose’) represents the sales of Rose wine in that month. It contains only float type values.

Before proceeding further to decomposition, we must impute these null values. Let's have a look at those rows which contain null value.



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

There are many methods available to impute the values in a time series data. Here we have used the interpolation method using second order Spline because it was simply following the previous years patterns (this conclusion was drawn only from visual interpretation). After imputation the time series (for last few years) looks like:

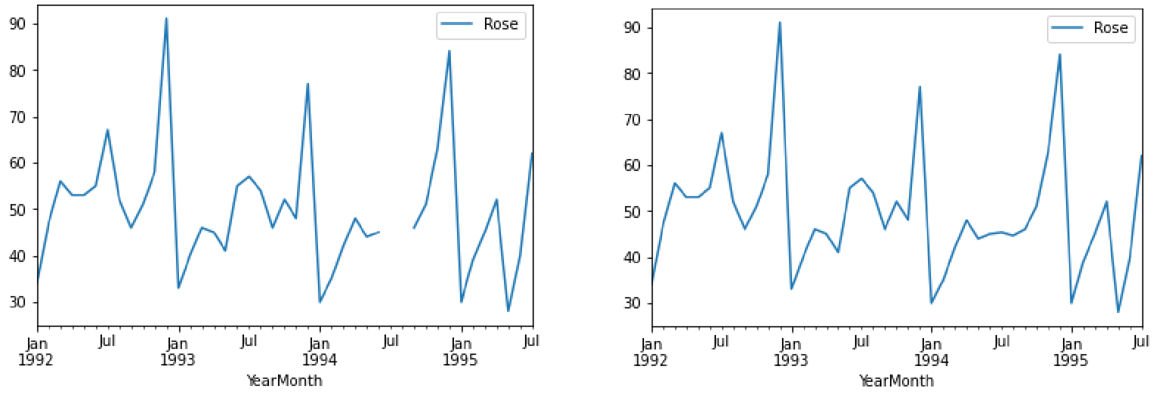


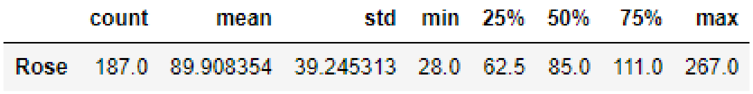
Figure 30: Time series data (for years>1992) before and after imputing values

Task 12. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

As we have already performed null check, data types, and shape earlier we can proceed further with visualization of data.

Statistical Description of Dataset

- Minimum value for the time series is 1070 and Maximum value is 7242.
- Mean value for the 'Sparkling' variable is 2402 and median is 1874.



	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.908354	39.245313	28.0	62.5	85.0	111.0	267.0

Visualization of Time Series Data

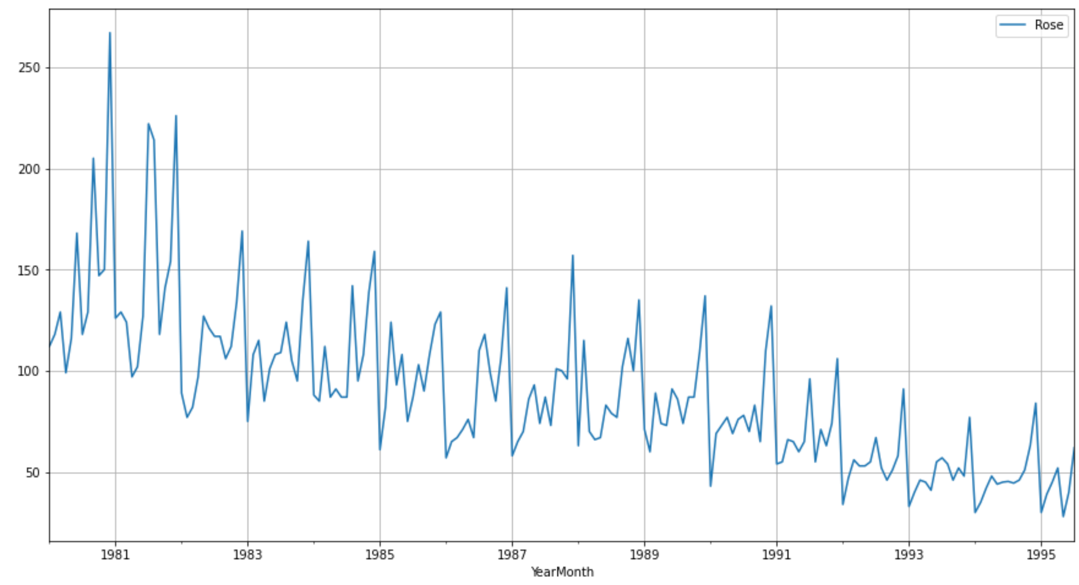


Figure 31: Plot showing the variation of Sales of Rose wine over the years

- As we can observe that there is seasonal component present in time series.
- If we look closely at the variations over a single year, we can say that the sales go up towards the end of the year.
- Just by looking at it we can say that there is weak a strong trend component present with in the time series.

Further, we can look at the variation of Sales in over the years. Here box plots are created to visualize the same:

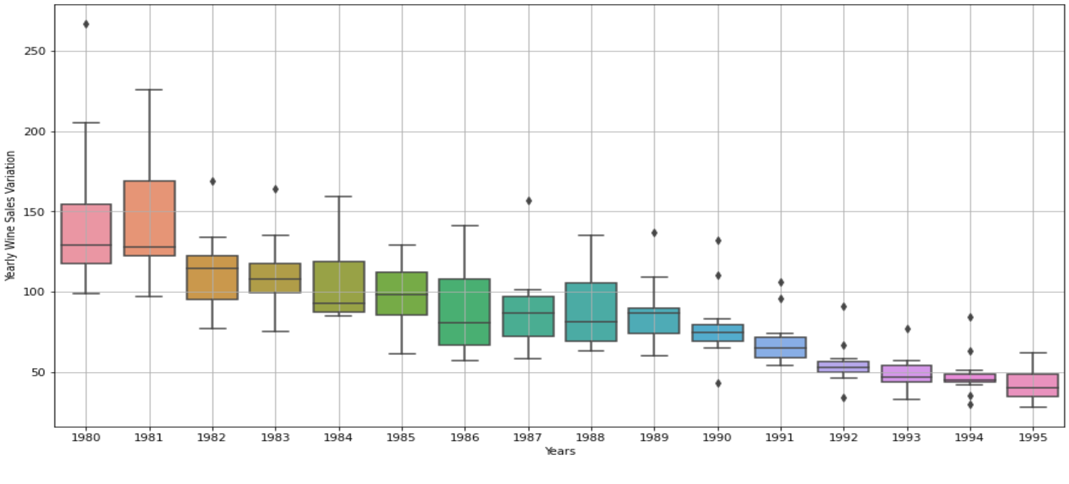


Figure 32: Box plots for visualization of variation of Sales for each year

- We can observe that over the years, for one or two months the sales are high which appear as outlier for each year.
- A decreasing trend in Sales of Rose wine can be seen easily.
- In this time series the maximum value occurred a month in the year 1980 which is 260
- Towards the end of the time series boxes are getting smaller and smaller which indicates that the Sales for these years are similar throughout the year (except for one or two months which appear as outliers)

Now, we can plot the yearly variation of sales in each month. This will give an idea of trend of time series over the years in that month:

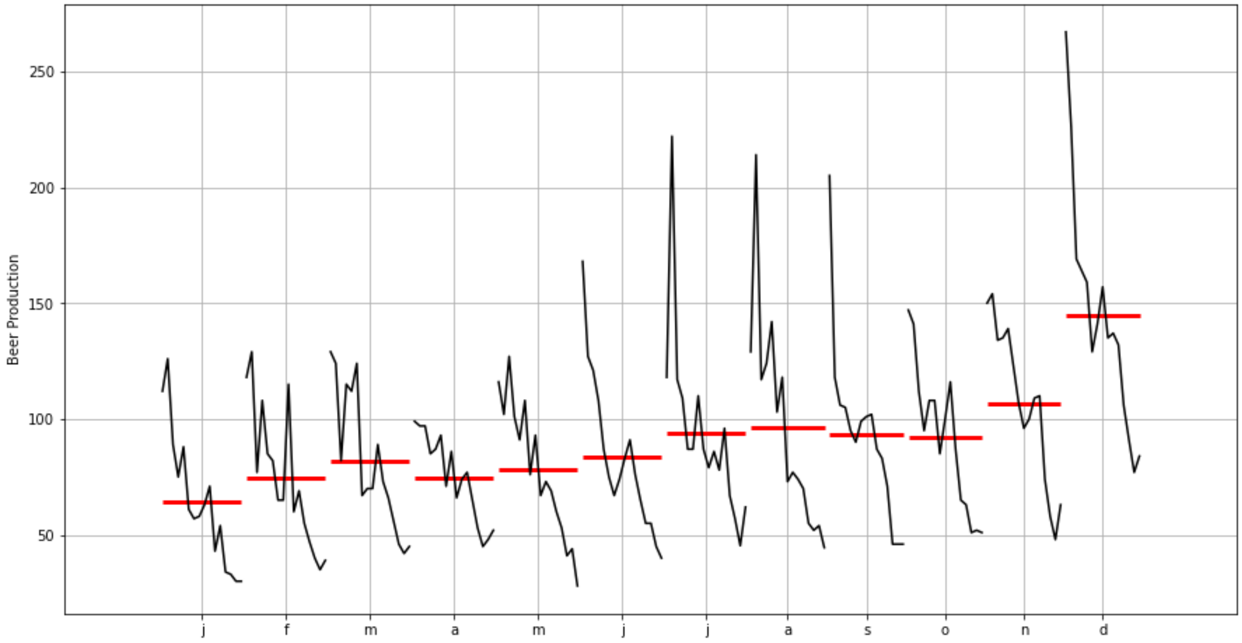
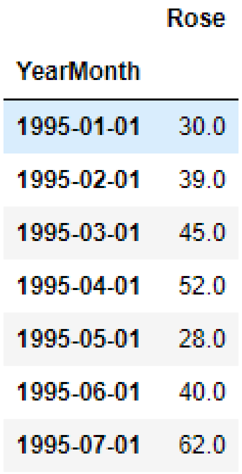


Figure 33: Plot of yearly variation of Sales over every month

- As we can observe that for the first half of each year the values are very less and after the 'July' month the Sales starts go up and by the end of the year the Sales are at the peak (in a particular year)
- Also, the variation of Sales in the is first six months are very less compared to variation in the last six months.
- There is a noticeable dip present in almost each month towards the end of its yearly variation depicting that the "Sales" are going down towards the end of this time series data.



	Rose
YearMonth	
1995-01-01	30.0
1995-02-01	39.0
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

One thing to keep in mind is that the values for year 1995 are incomplete as the time series data ends on Jul 1995.

Decomposition of the time-series-

As the name suggests, decomposition simply means decomposition of time series in deterministic (trend and seasonality) and stochastic (residual) components. Decomposition for the provided time series is given below:

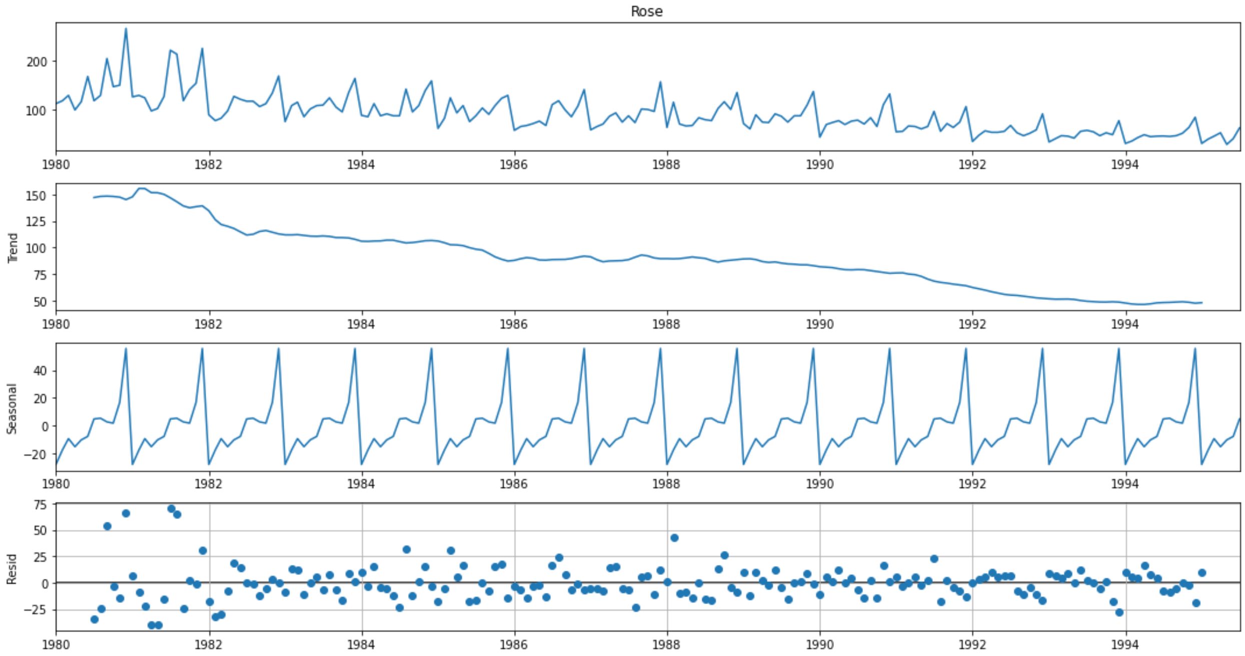


Figure 34: Decomposition of given time-series using Additive model

As we can observe that a strong seasonality component is present in the time-series. The decline of Sales is clearly visible in the trend component. The stochastic component of time-series (residuals varying between -25 to +75) is high when compared with the seasonal (max. value 40) and trend (max. value 150) component.

For additive model, decomposition can be represented with following equation:

$$\boldsymbol{Actual\ value\ of\ time - series\ =\ Trend\ +\ Seasonality\ + Stochastic\ component}$$

Further, we can look at multiplicative model for decomposition of time series though the residuals don't seem to follow any pattern but still we can do that for cross checking:

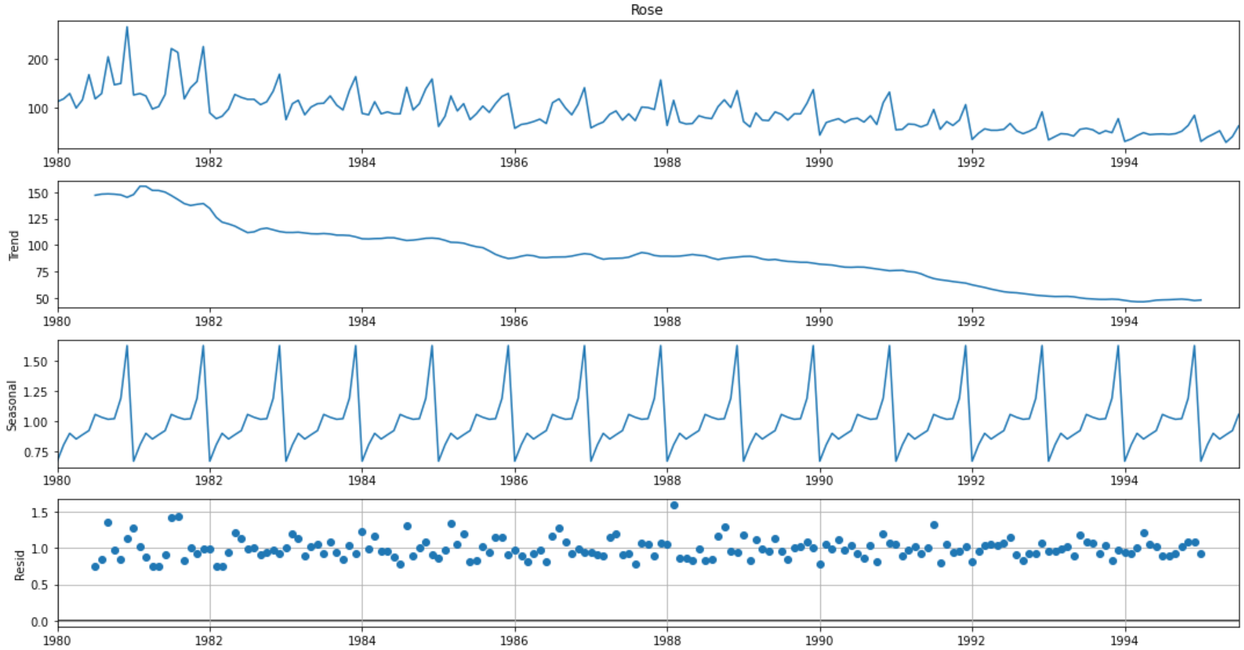


Figure 35: Decomposition of given time-series using Multiplicative model

In multiplicative models, the actual value of time-series can be obtained as:

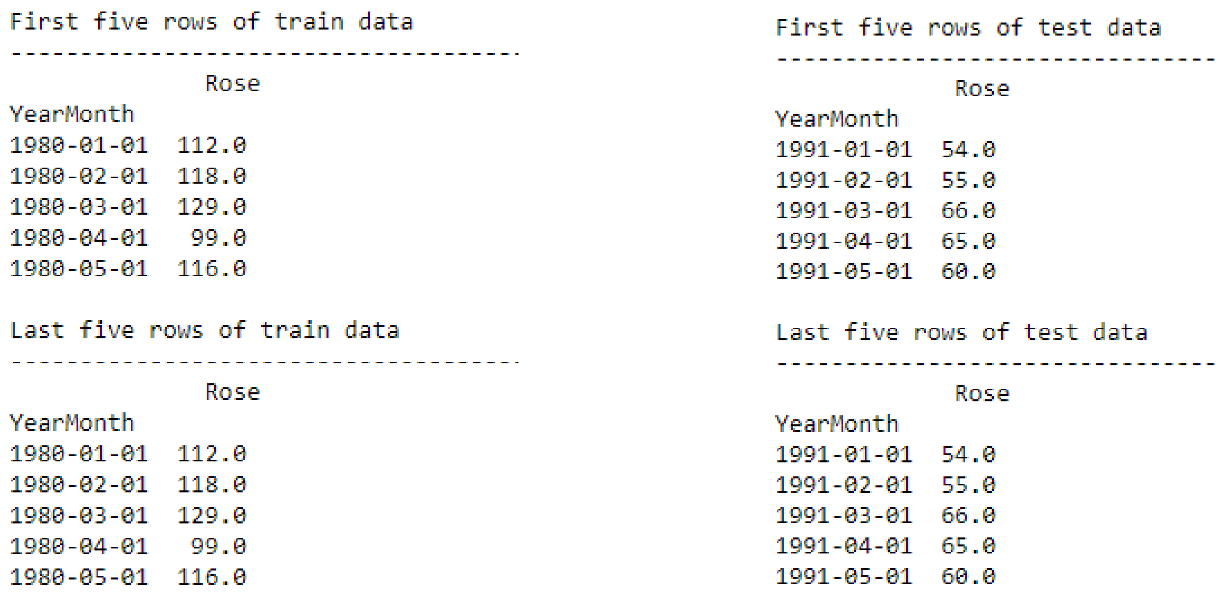
$$\boldsymbol{Actual\ value\ of\ time - series\ =\ Trend\ *\ Seasonality\ *\ Residuals}$$

As we expected, the decomposition using multiplicative model doesn't seem to perform any better as the magnitude of residual is varying from 0.7 to 1.6. But since the variation in residuals seems to be somewhat lesser than the additive model in, we can say that this model is better. Trend and seasonality component are almost same as that of additive model.

Task 13. Split the data into training and test. The test data should start from the year 1991.

Splitting the time-series in train and test set

The train and test split in a time-series model is based upon the timestamp associated with the variable. As instructed, here we have split the time-series data in train and test where test set is starting from year 1991:



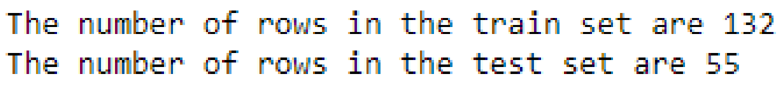
```
First five rows of train data
------------------------------------
              Rose
YearMonth
1980-01-01  112.0
1980-02-01  118.0
1980-03-01  129.0
1980-04-01   99.0
1980-05-01  116.0

Last five rows of train 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
```

```
First five 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 five 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
```

Figure 36: First and last five rows of Train and Test set



```
The number of rows in the train set are 132
The number of rows in the test set are 55
```

Figure 37: Shape of Train and Test set of Time-series

We can also plot the train and test set on the same axis to check whether the split is correct or not:

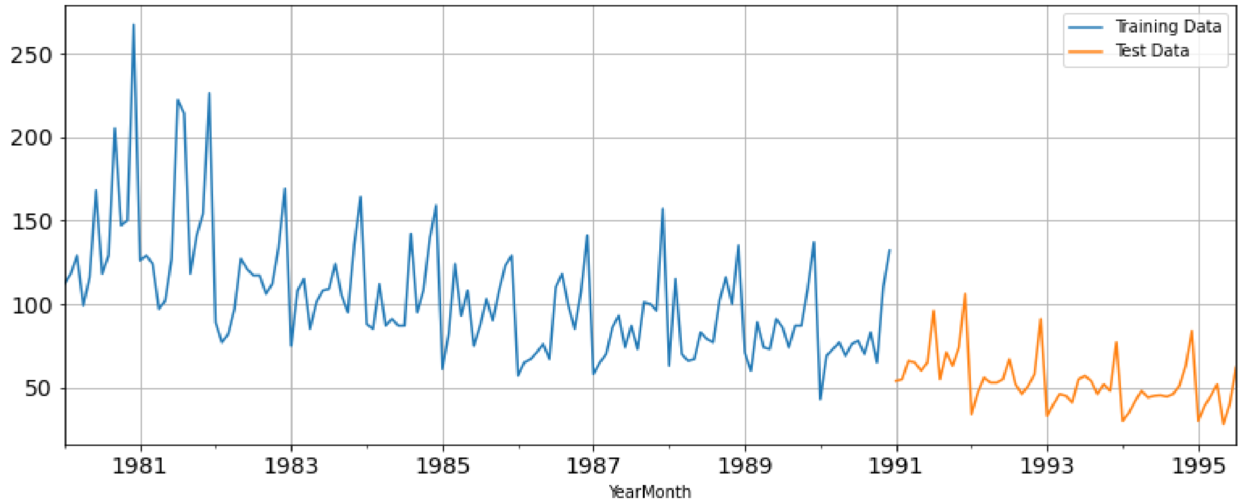


Figure 38: Plot of Train and Test data on same axis

As we can see that train and test set split is correct, hence we may proceed further to modelling part.

Task 14. Build various 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.

Modelling

As per the rubric let's start with the modelling part:

Exponential Smoothing Models

Simple Exponential Smoothing Models (SES)

SES or one-parameter exponential smoothing is applicable to time series which do not contain either of trend or seasonality.

$$\hat{Y}_{(t+1)} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots, 0 < \alpha < 1$$

where, α is the smoothing parameter for the level. In reality such a series is hard to find. This is a one step-ahead forecast where all the forecast values are identical.

Here we have built two models for SES, one is with automated optimization and other one is optimization using grid-search method for which minimum the RMSE value on test set (unseen data) was looked for among all the models trained. The parameter obtained from automated optimization is $\alpha = 0.0987$ and from the manual grid-search method we found out that minimum RMSE value on test set was obtained for $\alpha = 0.07$.

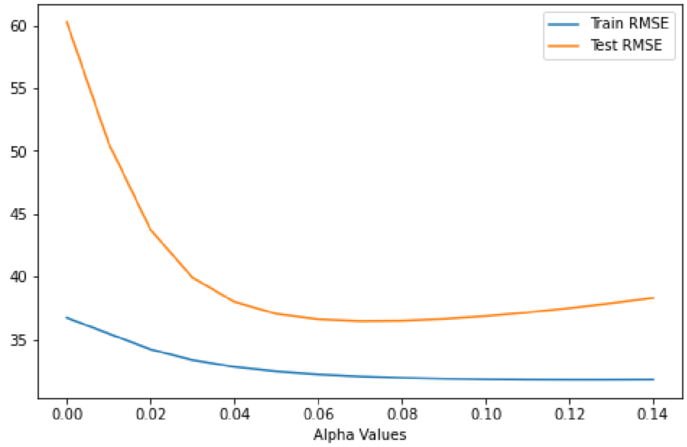


Figure 39: Plot showing variation of RMSE value with Level (α) for train and test set

For visualization of the forecast, we can plot the train and test set along with the forecast obtained from SES models:

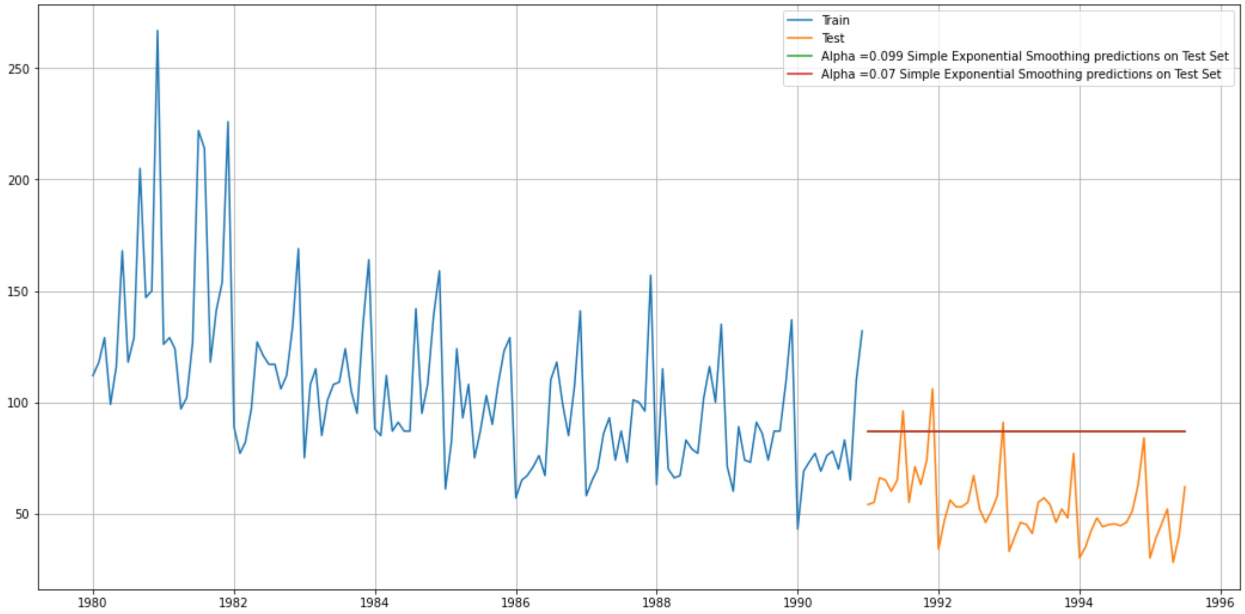


Figure 40: Plot for visualizing the forecasts obtained from SES models

As we can observe that the plots for both the α values are overlapping each other.

Double Exponential Smoothing Models (DES)

This method is an extension of SES method, proposed by **Holt** in **1957**. Hence, this method is also known as Holt’s model. This method is applicable where trend is present in the data but no seasonality. The forecast values are given as:

Forecast equation: $\hat{Y}_{t+1} = l_t + b_t$

Level Equation: $l_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1},\ 0 < \alpha\ <\ 1$

Trend Equation: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},\ 0\ <\ \beta\ <\ 1$

where, l_t is the estimate of level and b_t is the trend estimate. α is the smoothing parameter for the level and β is the smoothing parameter for trend. Now, similar to SES models we have built two models here also, one is automated optimization and the other one manual grid search model. The parameters obtained are following:

- For automated optimization model: $\alpha = 0.0176$ and $\beta = 3.24 * 10^{-05}$
- For manually optimized model: $\alpha = 0.04$ and $\beta = 0.04$

For visualization of the forecast, we can plot the train and test set along with the forecast obtained from DES models:

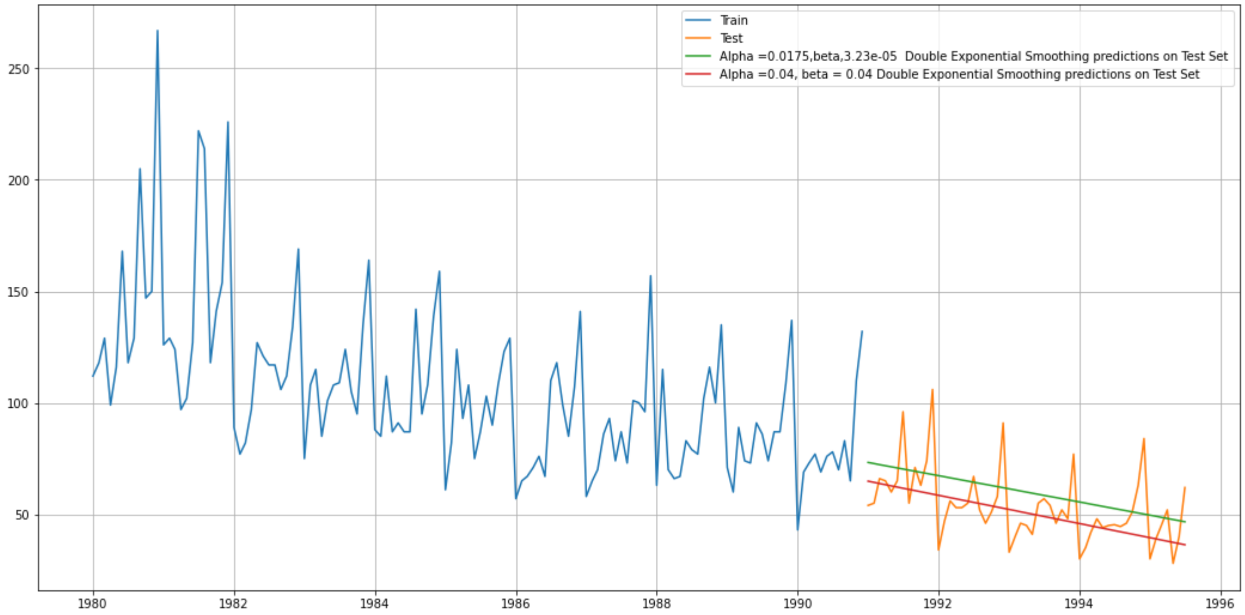


Figure 41: Plot for visualizing the forecasts obtained from DES models

As we can see that the automated model is slightly above the actual values, whereas the model obtained after grid search is closer to actual observations.

Triple Exponential Smoothing Models (TES)

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

Forecast equation: $Y_{t+1} = l_t + b_t + s_{t-m(k+1)}$

Level Equation:$l_t = \alpha(Y_t - s_{t-m}) + \alpha(1-\alpha)Y_{t-1}, 0 \ < \ \alpha \ < \ 1$

Trend Equation: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, 0 \ < \ \beta \ < \ 1$

Seasonal Equation: $s_t = \gamma(Y_t - l_{t-1} - b_{t-1}) \ + (1-\gamma)s_{t-m}, 0 \ < \ \gamma \ < \ 1$

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters α, β and γ. Here m is frequency of the seasonality component. This is a general method and a true multi-step ahead forecast.

Similar to aforementioned models, we have built two models here also, one is automated optimization and the other one manual grid search model. The parameters obtained are following:

- For automated optimization model: $\alpha = 0.066, \beta = 0.052 \ and \ \gamma = 3.88 * 10^{-6}$
- For manually optimized model: $\alpha = 0.1, \beta = 0.19 \ and \ \gamma = 0.019$

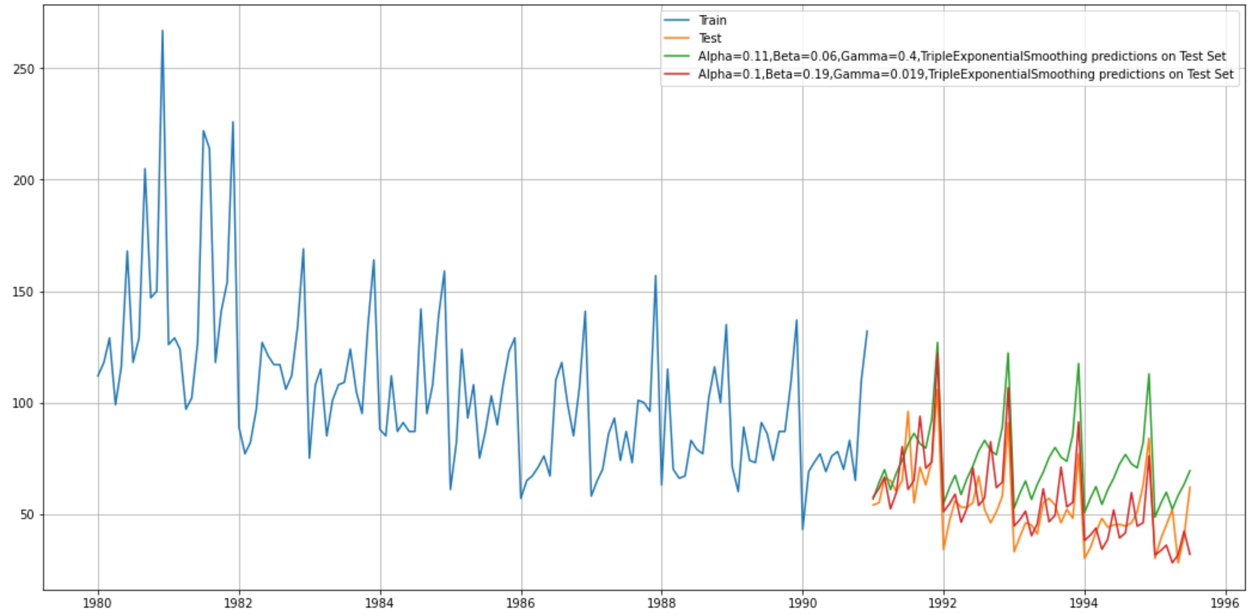


Figure 42: Plot for visualizing the forecasts obtained from TES models

As expected, we can observe that the models are doing very good job in finding correct the seasonal component and forecasting based on that for required steps (i.e., equal to length of test set) but still if we look closely, we can observe that the model with ‘red’ line is closer to actual values.

Linear Regression model

We cannot directly apply linear regression on a time-series data as the data has an order of occurrence, hence we must retain the order while fitting the data to regression model. Which means another variable has to be created which retains this order while fitting in the model. Once this is done, we can good to proceed further with fitting the values. After building the model we can forecast using this model for the steps size of 55 (i.e., length of test set). Following plot can be obtained by plotting the forecast against the actual series:

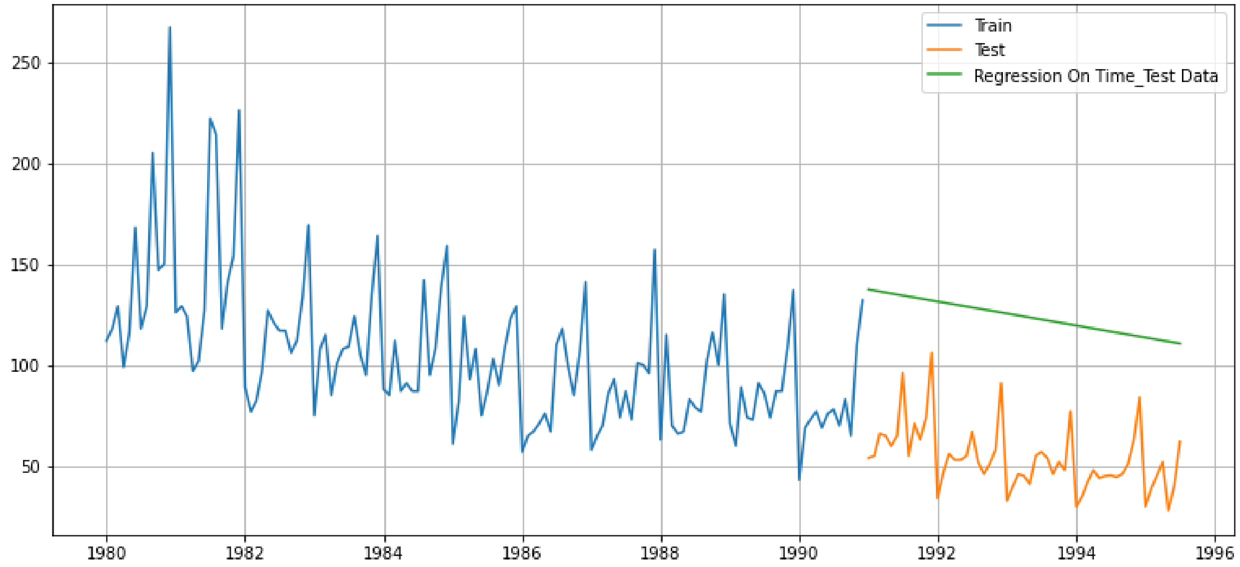


Figure 43: Plotting the values obtained after forecasting for length of test set

Naïve Forecast Model

As the name suggest, this is probably the simplest model to build. It simply states that tomorrow’s value will exactly be equal to today’s value. Hence, the last value of train set will become the prediction for any duration. This can be easily understood from the following plot:

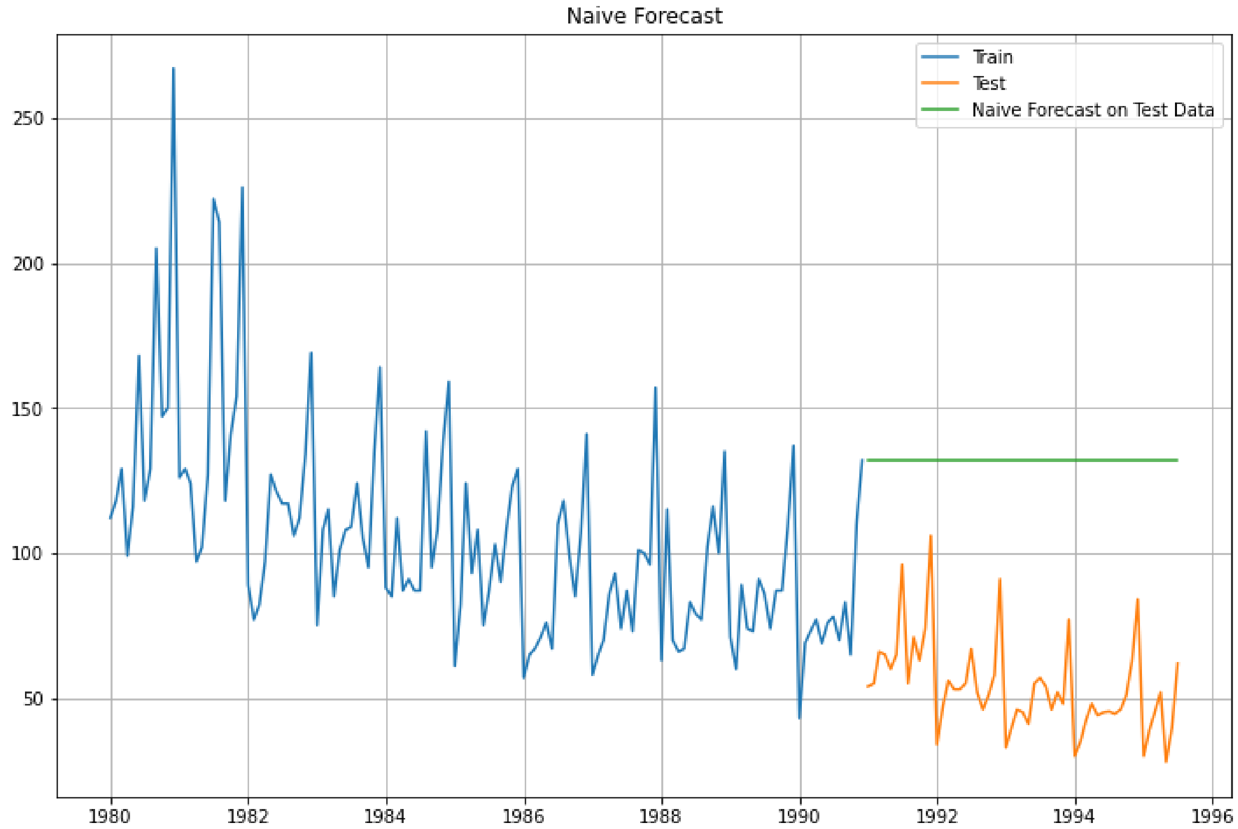


Figure 44: Plot of Naive forecast for test set time duration (steps)

Simple Average Model

In this model we simply average the values provided for a duration and forecast that same value over the required time duration.

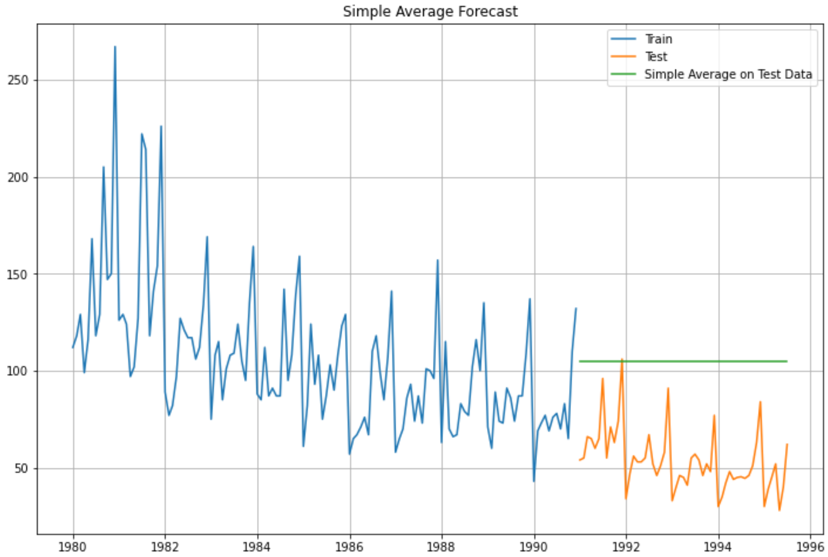


Figure 45: Plot of Simple average model over the test set time-duration

Task 15. 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.

Check for Stationarity of Time Series

Stationary process: A process is said to be stationary if its mean and variance are constant over a period of time and, the correlation between the two time periods depends only on the distance or lag between the two periods. Mathematically, let Y_t be a time series with these properties:

Mean: $E(Y_t) = \mu$

Variance: $Var(Y_t) = E(Y_t - \mu)^2 = \sigma^2$

Correlation: $\rho_k = E\left[\frac{(Y_t-\mu)(Y_{t+k}-\mu)}{(\sigma_t\sigma_{t+k})}\right]$

Where ρ_k is the correlation (or auto-correlation) at lag k between the values of $Y_t \; and \; Y_{t+k}$.

Augmented Dickey-Fuller test: The Augmented Dickey-Fuller test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis in a simple form for the ADF test is:

- **Null hypothesis ($\boldsymbol{H_0}$):** The Time Series has a unit root and is thus non-stationary.
- **Alternate Hypothesis ($\boldsymbol{H_1}$):** The Time Series does not have a unit root and is thus stationary.

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

Now, after performing the ADF test for the whole dataset we obtained following results:

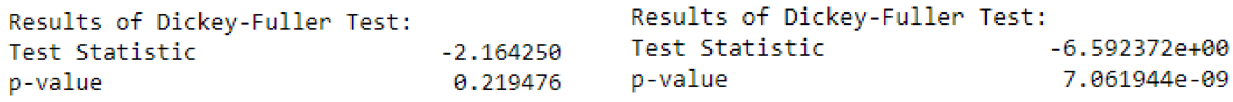
```
ADF test statistic is -2.242
ADF test p-value is 0.46643710204731487
```

Since, the p-value is coming out to be 0.47 which is greater than our significance level α which is 0.05 the null hypothesis cannot be rejected. Thus, at the significance level of 0.05 we can say time series is non-stationary. Further we have taken the difference of time series with 1 lag for achieving the stationarity goal and again performed ADF test on this series, results of which are shown below:

```
ADF test statistic is -8.161
ADF test p-value is 3.034192412609764e-11
```

As we can see that the p-value (≈ 0) is less than the significance level (i.e., 0.05); we can reject the null hypothesis at this significance level. Hence, we can say that differenced time series with 1 lag is stationary.

Since we will be building the model based on train data only, we should also check for the stationarity of train set. The ADF test results on train set and differenced train set with 1 lag are given below:



```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
```

```
Results of Dickey-Fuller Test:
Test Statistic                -6.592372e+00
p-value                        7.061944e-09
```

Figure 46: Results obtained from ADF test on Train set (left) and differenced series with 1 lag of train set (right)

As it is obvious from the results obtained that the original train set of time series is non-stationary since the p-value is 0.2195 which is greater than the significance level 0.05. On the other hand, the differenced time series with one lag of train set is stationary. Hence, for building the SARIMA models we will be using the 'd' parameter as 1 since we did the differentiation only one time to make the series stationary.

Task 16. 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.

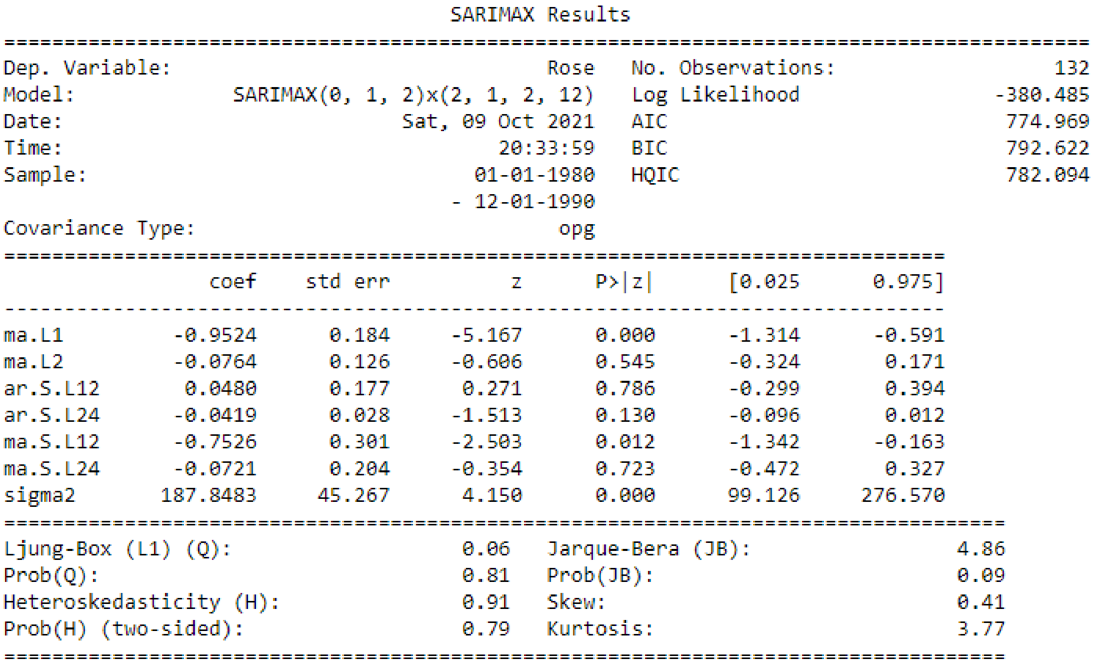
SARIMA Models

Seasonal ARIMA models are more complex models with seasonal adjustments. Since, these models are used when time series data has significant seasonality, we can use these models as our time series data contains significant seasonality.

The most general form of seasonal ARIMA is $ARIMA(p,d,q) * ARIMA(P,D,Q)[m]$, where P, D, Q are defined as seasonal AR component, seasonal difference and seasonal MA component respectively. And, 'm' represents the frequency (time interval) at which the data is observed.

As we saw from the visualizations that our time series data has yearly frequency, we will be putting 'm' as 12.

Now, for building the automated version of the SARIMA model, we had to pick the parameter's range of values first and after that we had to pick the best model based on the Akaike Information Criteria (AIC) built using those parameters. After that we used this model to forecast for 55 steps (i.e., length of test set) and calculated RMSE value for the model which came out to be 16.523.



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                132
Model:             SARIMAX(0, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -380.485
Date:                              Sat, 09 Oct 2021   AIC                            774.969
Time:                                      20:33:59   BIC                            792.622
Sample:                                  01-01-1980   HQIC                           782.094
                                       - 12-01-1990
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9524      0.184     -5.167      0.000      -1.314      -0.591
ma.L2         -0.0764      0.126     -0.606      0.545      -0.324       0.171
ar.S.L12       0.0480      0.177      0.271      0.786      -0.299       0.394
ar.S.L24      -0.0419      0.028     -1.513      0.130      -0.096       0.012
ma.S.L12      -0.7526      0.301     -2.503      0.012      -1.342      -0.163
ma.S.L24      -0.0721      0.204     -0.354      0.723      -0.472       0.327
sigma2       187.8483     45.267      4.150      0.000      99.126     276.570
===================================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                 4.86
Prob(Q):                              0.81   Prob(JB):                         0.09
Heteroskedasticity (H):               0.91   Skew:                             0.41
Prob(H) (two-sided):                  0.79   Kurtosis:                         3.77
===================================================================================
```

Figure 47: Results summary obtained for (0,1,2) (2,1,2)12 SARIMA

As we can see that the AIC value for this model is 774.969 and the model is also satisfying almost all the tests such as Ljung-box test, Heteroskedasticity and Jarque-Bera test.

Task 17. 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.

Now, for selecting the parameters manually we must first find out the values of parameters (p,d,q and seasonal parameters P,D,Q) by looking at Auto Correlation and Partial Auto Correlation plots:

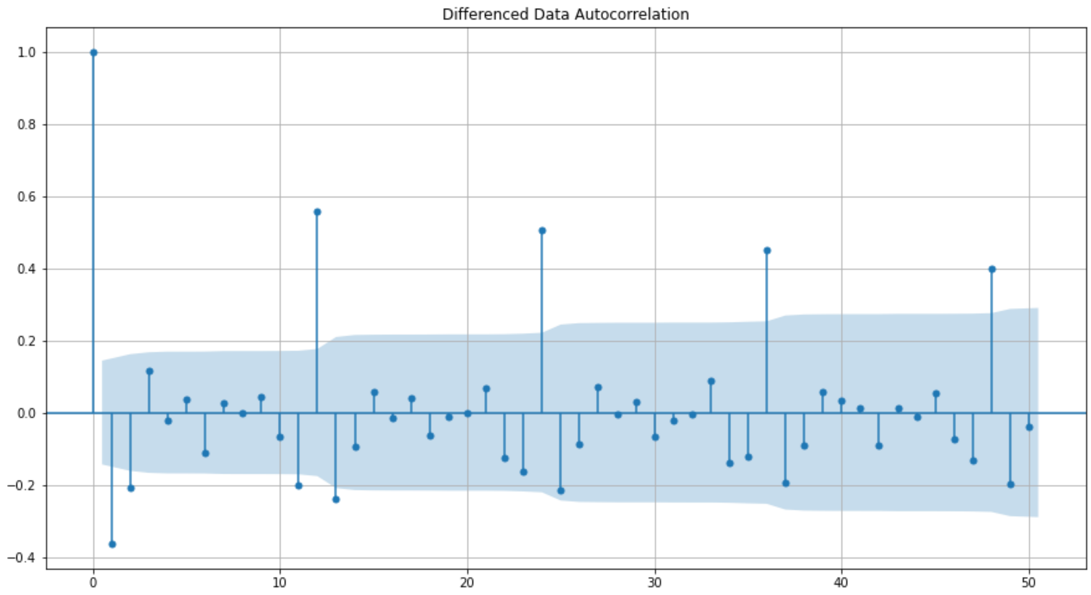


Figure 48:Auto Correlation plot for difference series with 1 lag

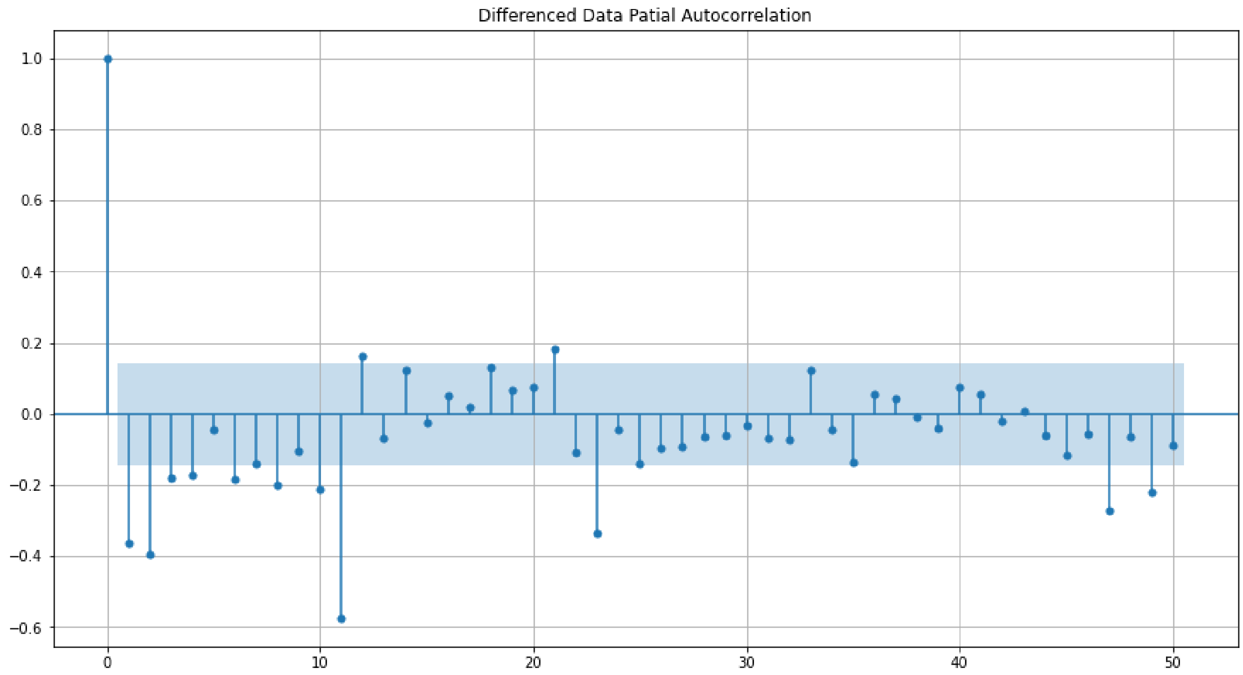


Figure 49: Partial Auto Correlation plot for difference series with 1 lag

Now, if we look at the first seasonal difference series, we can see that some amount is trend is still present in the series:

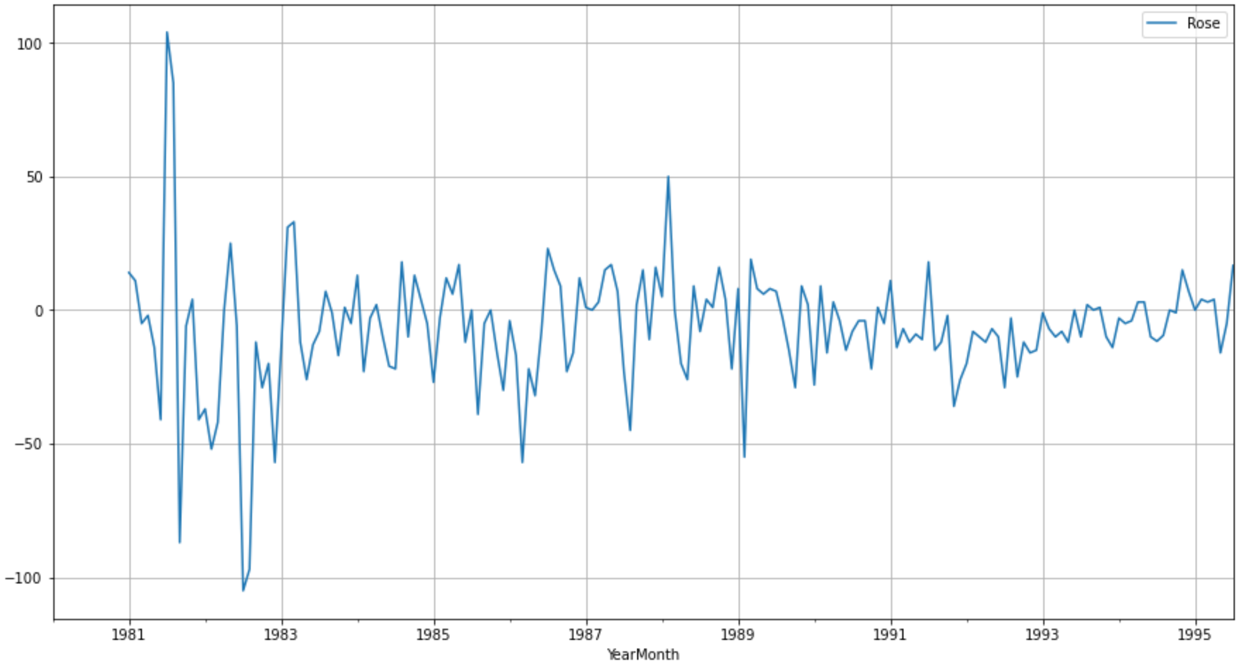


Figure 50: Plot of first seasonal difference series

Hence, we can try taking the first difference to eliminate the trend present in this series.

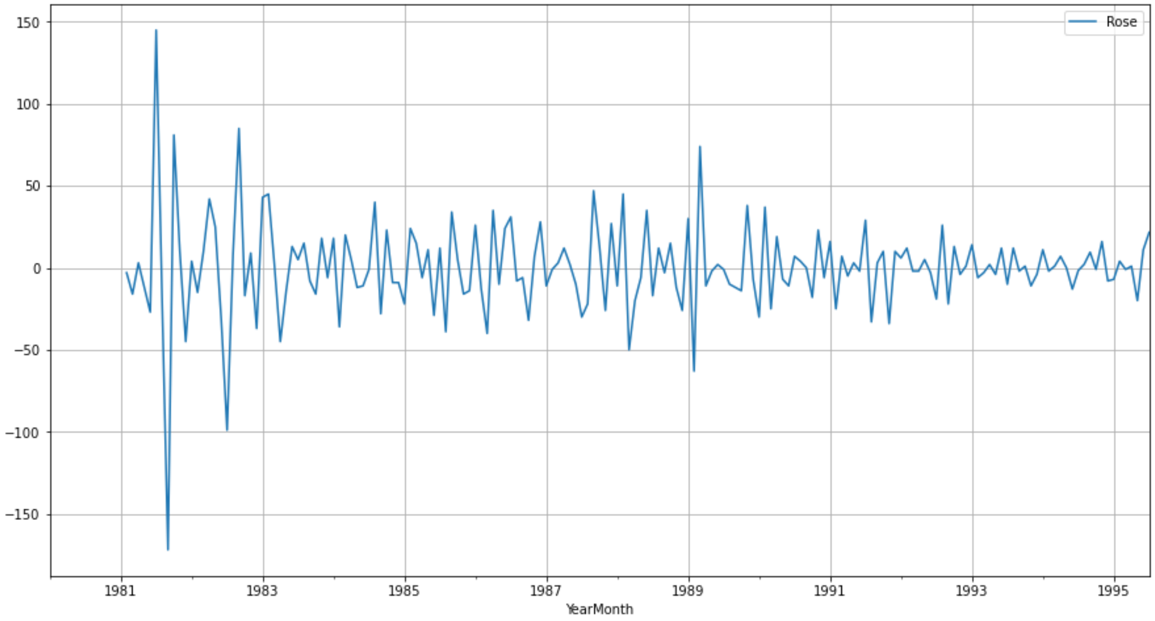
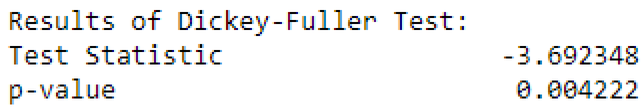


Figure 51:Plot of first difference first seasonal difference series

Now we see that there is almost no trend present in the data. Seasonality component only is present in the data. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.



```
Results of Dickey-Fuller Test:
Test Statistic                  -3.692348
p-value                          0.004222
```

As the p-value is < α i.e., 0.05, we can reject the null hypothesis which means this time series is stationary.

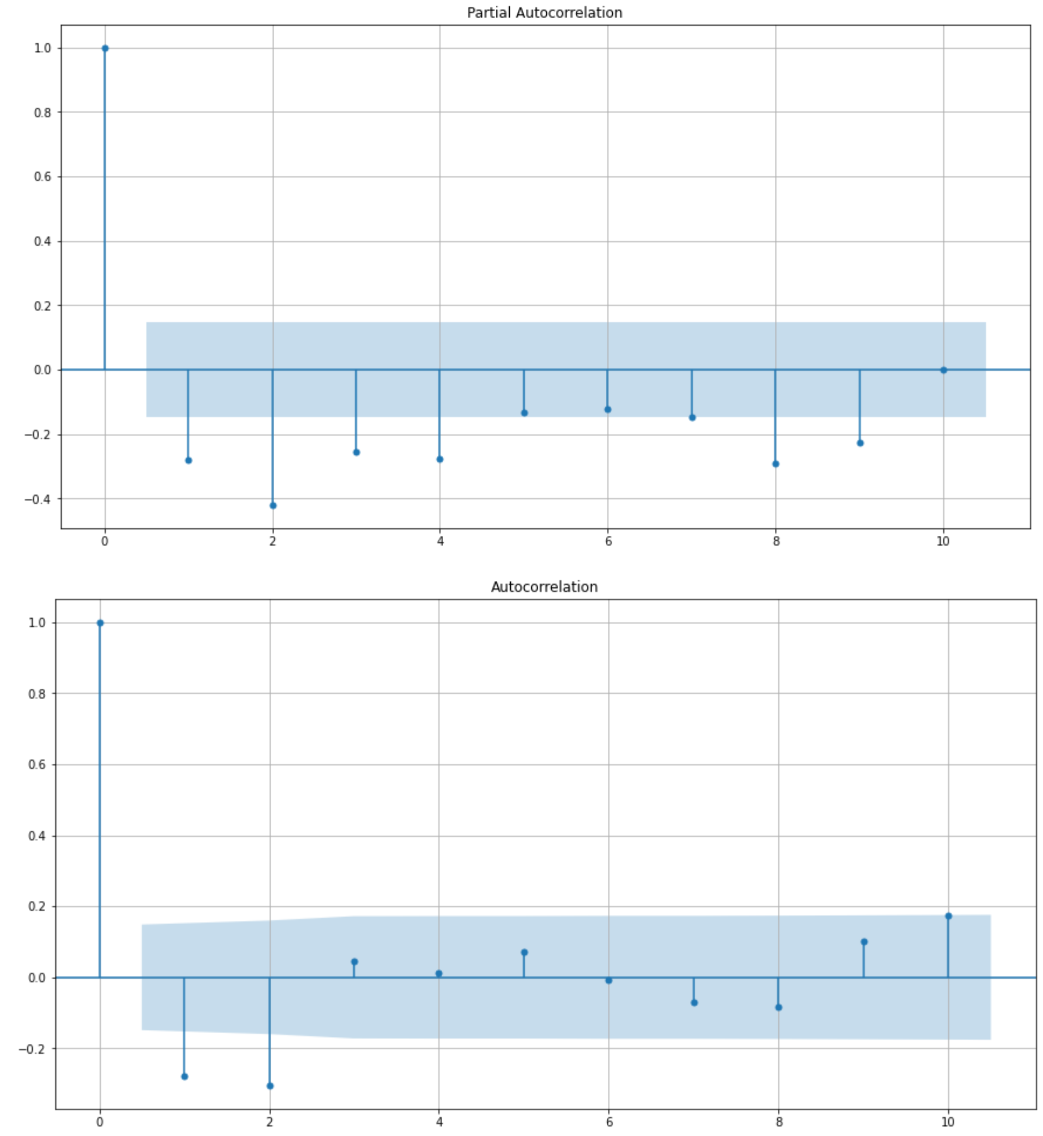
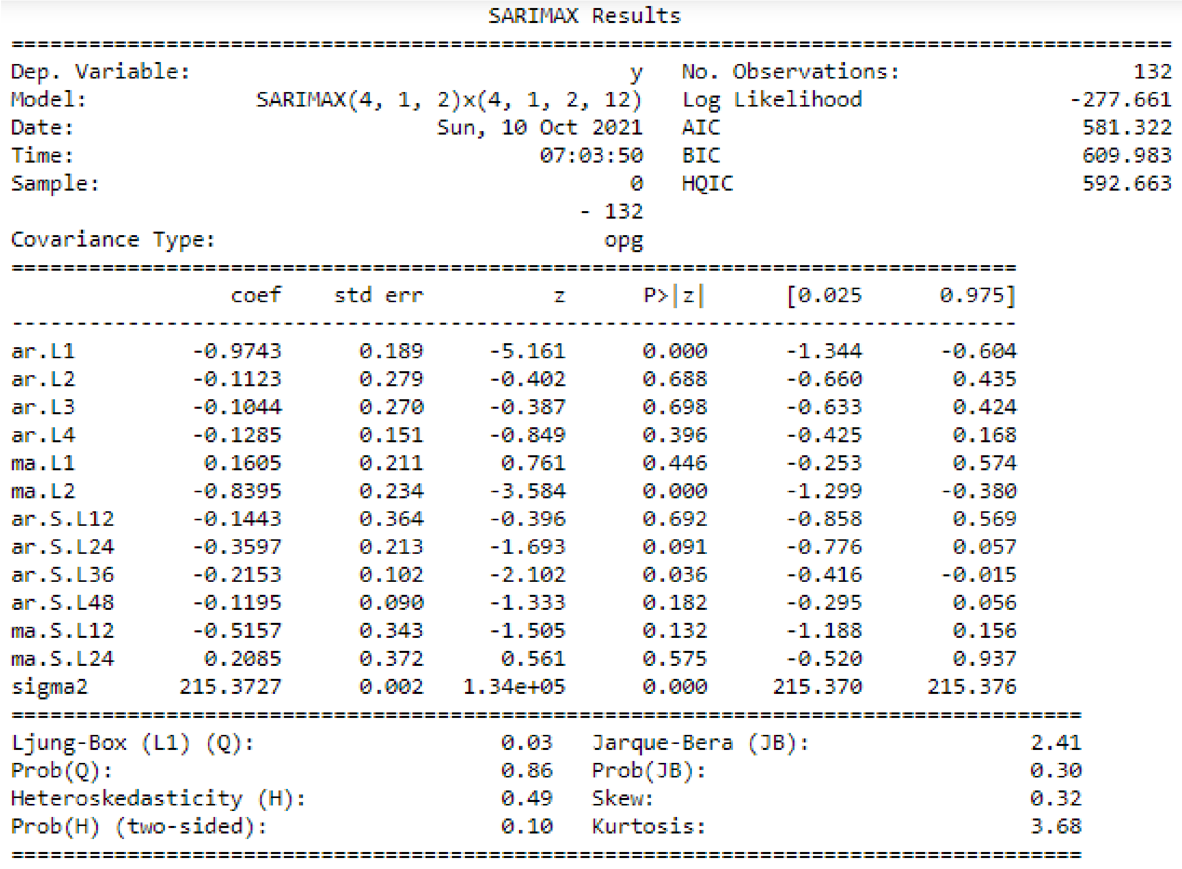


Figure 52: Plot of first difference and seasonal first difference series, PACF and ACF

Here, we have taken alpha=0.05. We are going to take the seasonal period as 12. We will keep the p=4 and q=2 (obtained from ACF and PACF plots of first difference series). The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).

Now, as we can see from the figure 51 that the PACF plot cuts off after 4 lags and ACF plot cuts off after 2 lags, hence, the value of ‘P’ is 4 and ‘Q’ is 2. Using these parameters, we have built the model results of which is given below:



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(4, 1, 2)x(4, 1, 2, 12)   Log Likelihood                -277.661
Date:                            Sun, 10 Oct 2021   AIC                            581.322
Time:                                    07:03:50   BIC                            609.983
Sample:                                         0   HQIC                           592.663
                                          - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9743      0.189     -5.161      0.000      -1.344      -0.604
ar.L2         -0.1123      0.279     -0.402      0.688      -0.660       0.435
ar.L3         -0.1044      0.270     -0.387      0.698      -0.633       0.424
ar.L4         -0.1285      0.151     -0.849      0.396      -0.425       0.168
ma.L1          0.1605      0.211      0.761      0.446      -0.253       0.574
ma.L2         -0.8395      0.234     -3.584      0.000      -1.299      -0.380
ar.S.L12      -0.1443      0.364     -0.396      0.692      -0.858       0.569
ar.S.L24      -0.3597      0.213     -1.693      0.091      -0.776       0.057
ar.S.L36      -0.2153      0.102     -2.102      0.036      -0.416      -0.015
ar.S.L48      -0.1195      0.090     -1.333      0.182      -0.295       0.056
ma.S.L12      -0.5157      0.343     -1.505      0.132      -1.188       0.156
ma.S.L24       0.2085      0.372      0.561      0.575      -0.520       0.937
sigma2       215.3727      0.002   1.34e+05      0.000     215.370     215.376
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                 2.41
Prob(Q):                              0.86   Prob(JB):                         0.30
Heteroskedasticity (H):               0.49   Skew:                             0.32
Prob(H) (two-sided):                  0.10   Kurtosis:                         3.68
===================================================================================
```

Figure 53:Results summary of (4,1,2) (4,1,2,12) SARIMA model

As we can see that the AIC value for the model is coming out to be 581.332. The distribution of residuals is close to normal as it can be observed from Jarque-Bera probability value. Ljung-Box test probability is also high which means residuals are independent. The RMSE value obtained for this model was 17.54.

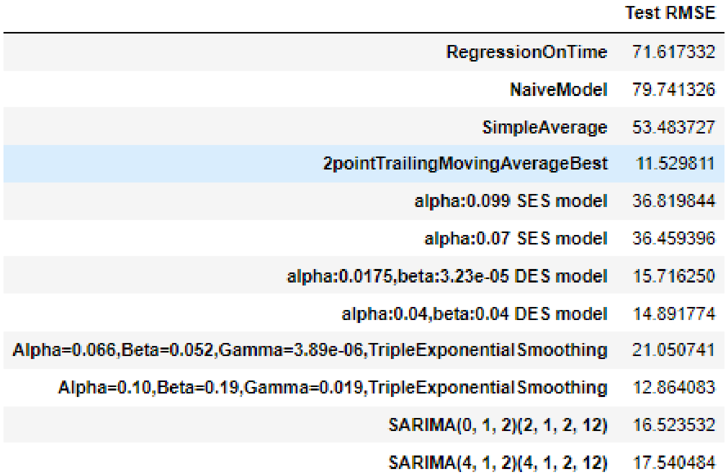
Task 18. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Results

Now, as have built several models for the same, we can compare these models based on the performance on test set. The performance metric chosen for the same is Root Mean Squared Error Value which can be computed as:

$$RMSE = \sqrt{E\left(\hat{Y} - Y\right)^2} = \sqrt{\frac{\Sigma\left(\hat{Y} - Y\right)^2}{N}}$$

The RMSE values for the different models built are given in following table:



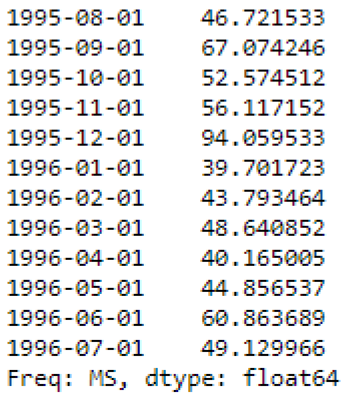
	Test RMSE
RegressionOnTime	71.617332
NaiveModel	79.741326
SimpleAverage	53.483727
2pointTrailingMovingAverageBest	11.529811
alpha:0.099 SES model	36.819844
alpha:0.07 SES model	36.459396
alpha:0.0175,beta:3.23e-05 DES model	15.716250
alpha:0.04,beta:0.04 DES model	14.891774
Alpha=0.066,Beta=0.052,Gamma=3.89e-06,TripleExponentialSmoothing	21.050741
Alpha=0.10,Beta=0.19,Gamma=0.019,TripleExponentialSmoothing	12.864083
SARIMA(0, 1, 2)(2, 1, 2, 12)	16.523532
SARIMA(4, 1, 2)(4, 1, 2, 12)	17.540484

Figure 54: Data Frame containing different models built & corresponding RMSE value obtained

Task 19. 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.

Based on this analysis, we found out that the Triple Exponential Smoothing model with parameters $\alpha = 0.1, \beta = 0.19\ and\ \gamma = 0.019$ gave least RMSE value on the test set (i.e., 12.864). Hence, we can conclude that this model is the most optimized one among the rest. Now, for forecasting into the future, we must train another model with these parameters on the whole data else we will be losing the data in test set if we used the previous model built using only the train data.

After building this final model, we can obtain the forecast for the required number of months (12 months). Forecast values obtained are given as:



```
1995-08-01    46.721533
1995-09-01    67.074246
1995-10-01    52.574512
1995-11-01    56.117152
1995-12-01    94.059533
1996-01-01    39.701723
1996-02-01    43.793464
1996-03-01    48.640852
1996-04-01    40.165005
1996-05-01    44.856537
1996-06-01    60.863689
1996-07-01    49.129966
Freq: MS, dtype: float64
```

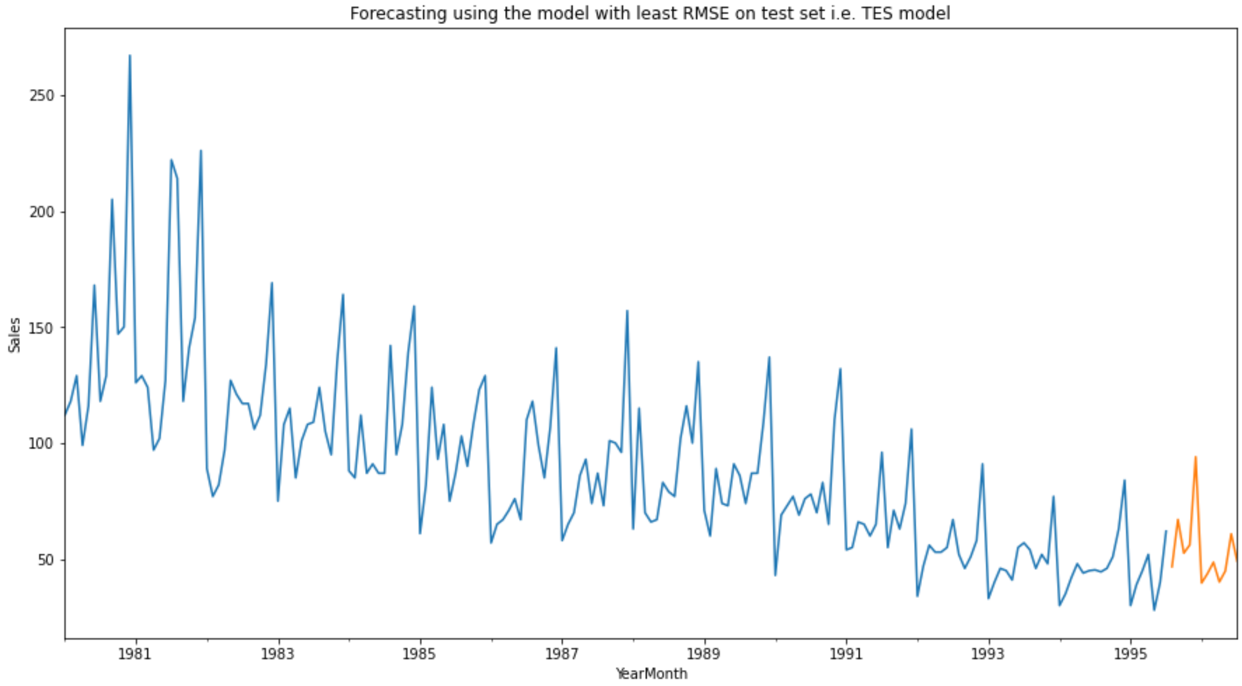


Figure 55: Plot of forecasted values using the final model for 12 months

Task 20. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Inferences and Suggestions

Based on this analysis following inferences can be drawn:

- As it was pointed out earlier in the analysis that the Sales is following a declining trend.
- In the 80's the seasonality component doesn't have much impact on the actual data. Residuals have high magnitude for 80's.
- After year 1988 the seasonal patterns are clearly visible in the plot of Rose wine Sales time series.
- The seasonality pattern observed indicates that the wine sales go up in the towards the end of the year. One possible reason could be that it is holiday season.
- Since, the seasonality is present in the time-series, in this analysis we chose to build a SARIMA model rather than building an ARIMA model.
- Based on this analysis we can conclude that the Triple Exponential Smoothing model performed better than even more complex models like SARIMA.
- As it is obvious from the seasonal patterns that the sales go up towards the end of the year, the availability of product should be insured for the increased demand and discount & offers can also be provided to attract more customers.
- As a declining trend is present in the time-series with a steep slope, the company should try to find out what is the possible reasons like is it the quality issue or other wines are getting more popular this product or steep price etc. and based on those, future strategies of business should be laid out.