Murali b

timeseries forecasting

Contents Page number

1.1 Head of the data 3  
1.2 Tail of the data 3  
1.3 Creating timestamp in the dataset 4  
1.4 Setting the timestamp at the index 4  
1.5 Shape of the dataset 5  
1.6 Plotting the graph of the dataset 5  
1.7 Description of the dataset 6  
1.8 Null values 6  
1.9 Treating Null values 6  
1.10 Yearly boxplot 7  
1.11 Monthly boxplot 7  
1.12 Graph showcasing sale of Rose 9  
1.13 Empirical cumulative distribution 9  
1.14 Avg Rose sales 10  
1.15 Additive decomposition 10  
1.16 Multiplicative decomposition 12  
1.17 first and last five rows of train and test 14  
1.18 Optimal parameters for SES 17  
1.19 Graph of forecasted values for SES 18  
1.20 RMSE values 18  
1.21 Optimal parameters for DES 18  
1.22 Graph of forecasted values for DES 19  
1.23 Test RMSE values of SES and DES 19  
1.24 Optimal parameters for TES 20  
1.25 Graph of forecasted values for SES,DES,TES 21  
1.26 RMSE values for DES,DES,TES 21  
1.27 Graph of Linear Regression model forecasting 23  
1.28 RMSE values of Linear Regression 23  
1.29 Graph of Naïve approach forecasting 24  
1.30 RMSE values of Naïve approach 24  
1.31 Graph of simple average model 25  
1.32 RMSE values of simple model 26  
1.33 Graph of moving average 27  
1.34 RMSE values for 2,4,6,9 point 27  
1.35 RMSE value for moving average 28  
1.36 ADF test using statsmodel 28  
1.37 ADF test with difference at one level 29  
1.38 Dataset plot to check stationarity 29  
1.39 Training dataset stationarity check 30  
1.40 Various ARIMA combination 30  
1.41 Sorting ARIMA values based on AIC 31  
1.42 Automated ARIMA result 32  
1.43 RMSE values for Automated ARIMA 32  
1.44 ACF plot showing seasonality 33  
1.45 Automated SARIMA AIC values 34  
1.46 SARIMAX at 6 result 35  
1.47 Plot diagnostic for SARIMA 35  
1.48 RMSE values for Automated SARIMA 36  
1.49 SARIMA values based on AIC 37  
1.50 Plot diagnostic for SARIMA at 12 38  
1.51 RMSE values for Automated SARIMA at 12 39  
1.52 ACF and PACF plot 40  
1.53 ARIMA model result 41

TIME SERIES

Problem:

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

DATASET: **ROSE**

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

Head of the data:   
Graphical user interface, application

Description automatically generated  
1.1 Head of the data  
  
Tail of the data:   
Table

Description automatically generated  
1.2 Tail of the data  
  
Creating time stamp in the dataset   
Text

Description automatically generated

Table

Description automatically generated  
1.3 Creating timestamp in the dataset   
  
Setting timestamp at the index in the dataset after dropping YearMonth column

Table

Description automatically generated  
  
Table

Description automatically generated with medium confidence  
1.4 Setting timestamp at the Index in the dataset, after dropping YearMonth column  
  
Table

Description automatically generated  
  
Shape of the dataset:   
  
  
1.5 Shape of the dataset  
  
Plotting the graph of the dataset:  
Chart

Description automatically generated  
1.6 Plotting graph of the dataset

2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.   
  
Describe the data:   
Graphical user interface, application

Description automatically generated  
1.7 Description of the dataset   
  
Checking for Null values:   
  
1.8 Null values  
  
We can see that there are 2 Null values present in the dataset.   
  
  
Treating the Null values:   
  
1.9 Treating Null values  
  
Now, we can see that there are no Null values present in the dataset.

Yearly Boxplot:   
Chart, box and whisker chart

Description automatically generated  
1.10 Yearly Boxplot showcasing the sale of Rose Wine   
  
The yearly boxplots show us how the sale has decreased over the past few years. The highest number of sales being recorded in the year 1981.   
  
Monthly Boxplot:   
  
Chart, box and whisker chart

Description automatically generated  
1.11 Monthly boxplot showcasing the sale of Rose wine   
  
We can see that highest sale was recorded in the month of December then was followed by November, October, and August.   
  
  
Monthly sales across the years:   
Chart, line chart

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Chart, line chart

Description automatically generated  
1.12 Graph showcasing the sale of Rose wine across the year for every month

From the above plot, it can be concluded that December has the highest sale of Sparkling wine.   
  
  
Plotting the Empirical Cumulative Distribution:   
Chart, line chart

Description automatically generated

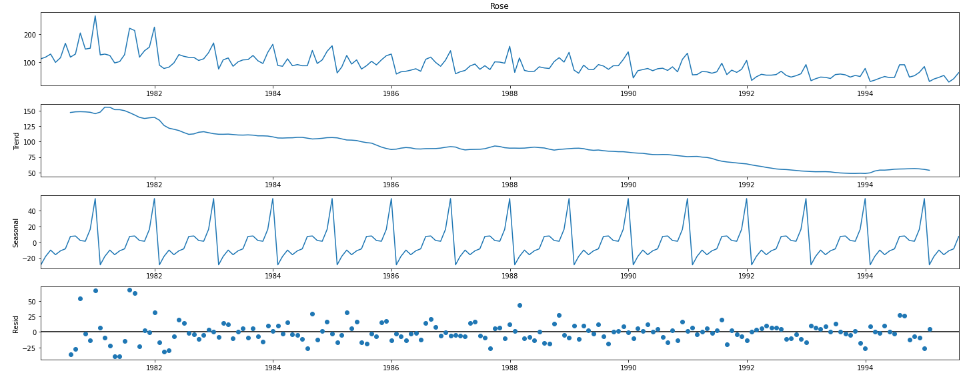
1.13 Empirical cumulative distribution graph

The above graph tells us what percentage of data points refer to what number of Sales.

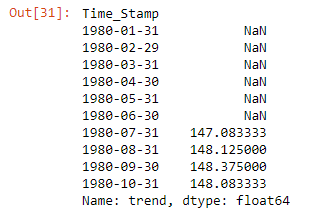
Plotting the average Sparkling Sales per month:   
Chart, scatter chart

Description automatically generated  
Chart, box and whisker chart

Description automatically generated  
1.14 Graph showcasing average rose sales and its percentage change over the years   
  
  
The above two plots show us the Average Sparkling Sales and the Percentage change of Sparkling Sales with respect to the time. So, from the EDA we can observe that there are 187 rows and 2 columns. Two null values were present initially so we treated the null values, and the data starts from year 1980 and will end in 1995. Later, we then created a time stamp and then replaced it at index while dropping the YearMonth column. The plot shows that the sale for Rose has been declining over the past few years. The lowest sale was in the year 1995 and highest was in the year 1980 in December month followed by November and October.

Decomposition:   
  
Additive Decomposition:   


1.15 Additive Decomposition

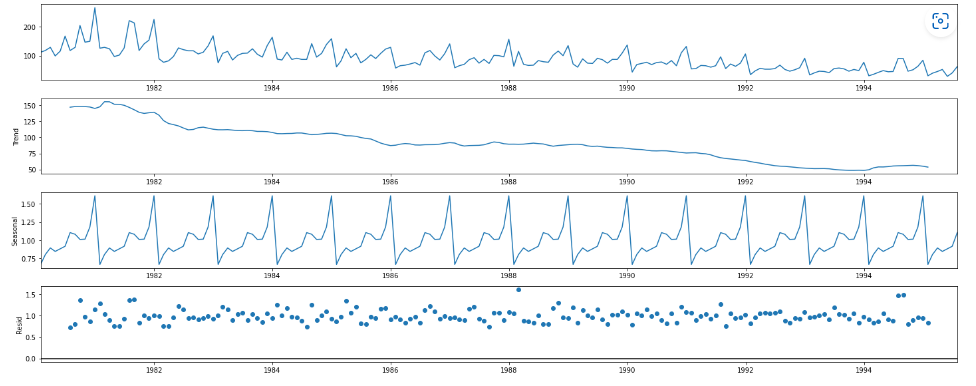
  
Text, table

Description automatically generated

Text, table

Description automatically generated

We can observe that the residuals are located around 0 from the plot of the residuals in the decomposition.

Multiplicative Decomposition:   
  
1.16 Multiplicative Decomposition

Table

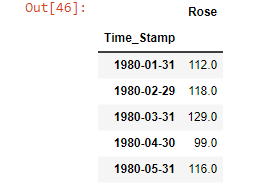
Description automatically generated

Graphical user interface, text, table

Description automatically generated  
Table

Description automatically generated

For the multiplicative series, we see that a lot of residuals are located around 1.

3. Split the data into training and test. The test data should start in 1991.   
  
A picture containing text

Description automatically generated

Table

Description automatically generatedA picture containing table

Description automatically generated  
1.17 First and last five rows of Training and test data

Splitting the data into train and test and plotting the training and test data. Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

Plotting the Autocorrelation and the Partial Autocorrelation function plots:  
Chart, timeline

Description automatically generated

Timeline

Description automatically generated

From the above graphs, we can say that there seems to be a seasonality in the data.

Checking for stationarity of the Training Data Time Series:  
A picture containing graphical user interface

Description automatically generated  
We see that the series is not stationary at 𝛼 = 0.05.

Graphical user interface

Description automatically generated  
We see that after taking a difference of order 1 the series have become stationary at 𝛼 = 0.05.

Train set Info:   
Text

Description automatically generated  
  
  
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.   
  
Simple Exponential Smoothing model:   
  
Optimal parameters for SES:  
Text

Description automatically generated  
1.18 Optimal parameters for Single Exponential smoothing   
  
Table

Description automatically generated  
  
SimpleExpSmoothing class must be instantiated and passed the training data. The fit() function is then called providing the fit configuration, the alpha value, smoothing\_level. If this is omitted or set to None, the model will automatically optimize the value.   
  
Plotting the Training data, Test data and the forecasted values:   
  
Graphical user interface, chart, application

Description automatically generated  
1.19 Graph of forecasted values for Simple Exponential Smoothing   
  
It is being found that the optimal smoothing level is at 0.0987. The above-mentioned graph showcases the forecasted values for both the train and test dataset using SES.

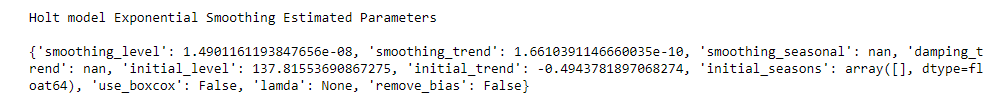
RMSE:



A picture containing shape

Description automatically generated  
1.20 RMSE values

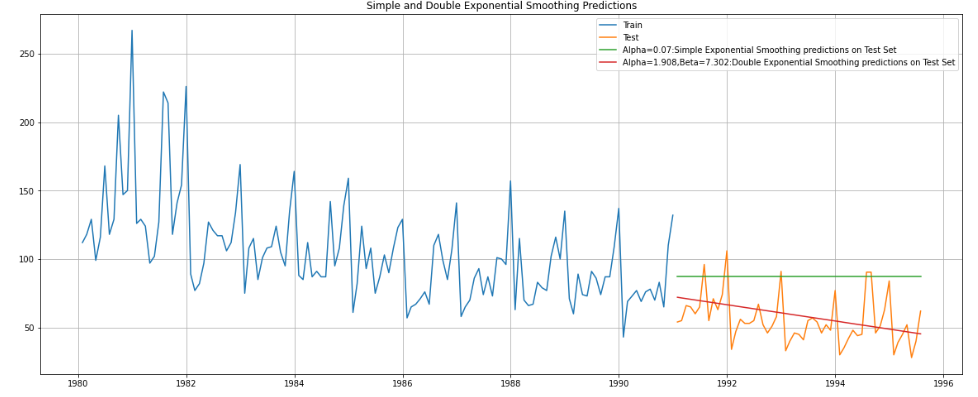
Double Exponential Smoothing:

Optimal parameters for DES:   
  
  
1.21 Optimal parameters for Double Exponential Smoothing

Graphical user interface, text

Description automatically generated

We can see that the optimal smoothing level for alpha is 1.490 and beta is 1.661

Plotting the Training data, Test data and the forecasted values:   
  
1.22 graph of forecasted values for Double Exponential Smoothing

We can see here that the double exponential smoothing is picking up the trend component along with the level component.

RMSE values:

  
Graphical user interface, text

Description automatically generated with medium confidence  
1.23 Test RMSE value of SES and DES

Inferences which can be made here is that we see that the DES has done well when compared to the SES. This is because the DES model was able to pick up the trend component available in the data.

Triple Exponential Smoothing:

Optimal parameters:  
Text

Description automatically generated with medium confidence  
1.24 Optimal parameters for Triple Exponential Smoothing

Table

Description automatically generated  
It is found that the optimized smoothing level is at alpha= 0.089 , beta= 0.0002 , gamma = 0.0034 based on the test set.

Plotting the Training data, Test data and the forecasted values:  
Graphical user interface, chart, application

Description automatically generated   
1.25 Graph of forecasted values for Single, Double and Triple Exponential Smoothing  
  
We see that the Triple Exponential Smoothing is picking up the seasonal component as well as the trend.

RMSE value:   
  
Text

Description automatically generated  
1.26 RMSE values for Single, Double and Triple Exponential Smoothing  
  
We see that the multiplicative seasonality model has not done that well when compared to the additive seasonality Triple Exponential Smoothing model.

Linear Regression:  
  
Training and Test time instances:   
Scatter chart

Description automatically generated

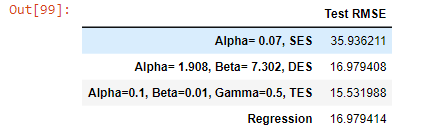
Table

Description automatically generated

For regression model to be built, we take both the train and test data and fit it using Linear Regression.

Plotting the regression model forecasting:  
Chart, line chart

Description automatically generated  
1.27 Graph of Linear regression model forecasting

RMSE value:   
  
  
  
1.28 RMSE values of Linear Regression along with other models.   
  
We can observe that RMSE value for Regression is high when compared to TES but very low when compared to SES and DES. So overall regression is not best fit model used for forecasting.

Naive Approach:   
  
Text

Description automatically generated

Chart

Description automatically generated  
1.29 Graph of Naïve Approach forecasting

For Naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

RMSE value:   
  
  
Graphical user interface, text, application

Description automatically generated  
1.30 RMSE value: Naïve Approach along with previously built models   
  
From the above output, we can see that the RMSE for Naïve Bayes is highest when compared to other models built. However, this cannot be used for best fit for overall predictions.

Simple Average:   
  
Table

Description automatically generated  
  
Chart

Description automatically generated  
1.31 Graph of Simple Average model

For this simple average method, we will forecast by using the average of the training values.

RMSE value:   
  
Graphical user interface, text, application

Description automatically generated  
1.32 RMSE values of Simple Model

The RMSE value for Simple Average is low as compared to Naïve mode but is higher than TES.

Moving Average:   
Table

Description automatically generated with medium confidence

Plotting on both the Training and Test data:  
Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated  
1.33 Graph of moving average model forecasting

From the above plots, we can see that the 2-point trailing moving average is really performing well on test data when compared to the other trailing points.

RMSE value: Text

Description automatically generated  
1.34 RMSE values for 2,4, 6 and 9 point moving average  
  
So compared to 4,6 and 9 Moving Average RMSE, we can see that the RMSE for 2 point Moving Average has a very low value.

Table

Description automatically generated  
1.35 RMSE values for moving average model and all the previously built models

When compared to previously built models, we can see that the Moving Average has the lowest RMSE values but still it’s an unstable method for forecasting.

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.

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:

H0 : The Time Series has a unit root and is thus non-stationary.

H1 : The Time Series does not have a unit root and is thus stationary. We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

ADF Test using statsmodel:  
Text

Description automatically generated  
1.36 ADF test using statsmodel

We see that at 5% significant level the Time Series is non-stationary.

ADF Test with difference at one level:  
Text, letter

Description automatically generated  
1.37 ADF Test with difference at one level

So, after using one level differencing, we can see that the data is now stationary.

Plot to check stationarity:   
Graphical user interface, chart, line chart

Description automatically generated  
1.38 Dataset plot to check stationarity

Checking the stationarity of the training dataset:  
Chart

Description automatically generated

We see that at 5% significant level the Time Series is non-stationary.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.  
Graphical user interface

Description automatically generated  
1.39 Training Dataset stationarity check

We see that at 𝛼 = 0.05 the Time Series is indeed stationary.

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.

Automated ARIMA:  
  
Text

Description automatically generated  
1.40 Various ARIMA combination   
So we use different parameter combinations to find the value of p,d,q to build the model. In automated ARIMA usually Python does the permutation and combination to arrive at the best values for p,d,q.

Various ARIMA combination:



Text

Description automatically generated

Sorting the ARIMA values based on the lowest AIC:

Table

Description automatically generated

1.41 Sorting the ARIMA values based on the lowest AIC

We can see that the automated ARIMA model can be built using 0,1,2 as it has the lowest AIC value.

Automated ARIMA result:   
Table

Description automatically generated with medium confidence  
1.42 Automated ARIMA result

RMSE value:   
  
  
Table

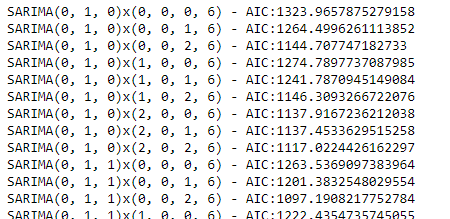
Description automatically generated with medium confidence  
1.43 RMSE values for Automated ARIMA and all the other models built

Automated SARIMA:   
Chart, box and whisker chart

Description automatically generated  
1.44 ACF plot showing seasonality   
  
We see that there can be a seasonality of 6 as well as 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

SARIMA at 6:  
Text

Description automatically generated  
  


Automated SARIMA AIC values in ascending order:   


Text, letter

Description automatically generated  
1.45 Automated SARIMA AIC values in ascending order

SARIMA at 6 results:  
Table

Description automatically generated  
1.46 SARIMAX at 6 result

SARIMA Plot:   
Chart, line chart

Description automatically generated  
1.47 Plot diagnostic for SARIMA

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

RMSE value:   
Table

Description automatically generated

Table

Description automatically generated  
1.48 RMSE values for Automated SARIMA and all the other models built

Setting the seasonality as 12 for the second iteration of the auto SARIMA model:   
Text

Description automatically generated



Text

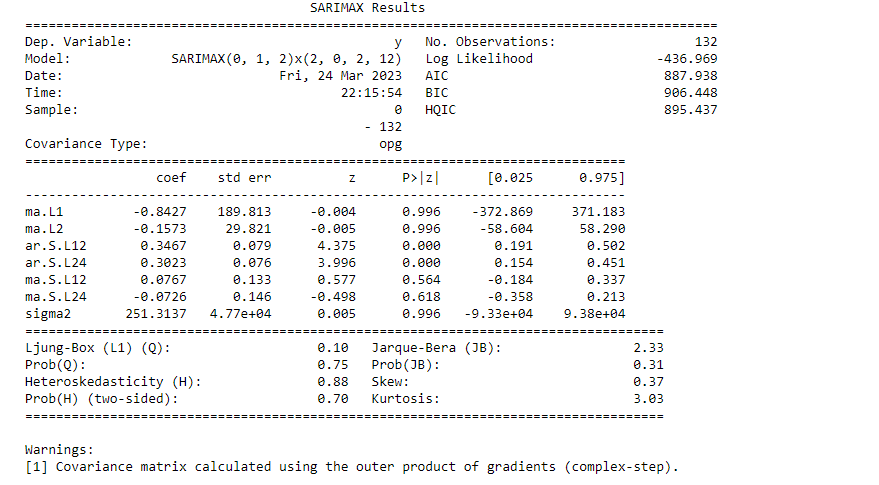
Description automatically generated

SARIMA values based on the lowest AIC value:

Text, letter

Description automatically generated  
1.49 SARIMA values based on the lowest AIC value:

SARIMA at 12 results:



Plot of SARIMA 12:   
  
Chart

Description automatically generated  
1.50 Plot diagnostic for SARIMA at 12

Like the last iteration of the model where the seasonality parameter was taken as 6, here also we see that the model diagnostics plot does not indicate any remaining information that we can get.

RMSE value:

Table

Description automatically generated



Table

Description automatically generated  
1.51 RMSE values for Automated SARIMA at 12 and all the other models built

We see that the RMSE value has only increased when the seasonality parameter was changed to 12, and there is not much difference in the model diagnostics of the two models.

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.   
  
ARIMA manual:   
Chart, timeline

Description automatically generated

1.52 ACF and 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 will take the value of p and q to be 0 and 0 respectively.

ARIMA model results:   
Table

Description automatically generated

1.53 ARIMA model Result

RMSE value:



Table

Description automatically generated  
1.54 RMSE values for Manual ARIMA and all the other models built

SARIMA Manual:   
  
Timeline

Description automatically generated with medium confidence  
1.55 Differenced Data : ACF and PACF plot

We see that our ACF plot at the seasonal interval (6) does not taper off. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.

Plot graph for Series:   
Graphical user interface, chart

Description automatically generated  
1.56 Series Graph

We see that there is a trend and a seasonality. So, now we take a seasonal differencing and check the series.

Timestamp:   
Graphical user interface

Description automatically generated  
1.57 Series Graph with difference at 6

We see that there might be a slight trend which can be noticed in the data. So we take a differencing of first order on the seasonally differenced series.

Graphical user interface, chart, line chart

Description automatically generated  
1.58 Series graph with order 1 difference for seasonality at 6 difference series

Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.

Graphical user interface

Description automatically generated with medium confidence  
1.59 Stationarity of dataset for building SARIMA model

Checking the ACF and the PACF plots:   
Chart, timeline

Description automatically generated

1.60 ACF and PACF plot for SARIMA at 6 newly modified Time Series

Here, we have taken alpha=0.05. We are going to take the seasonal period as 6. We will keep the p(2) and q(2) parameters same as the ARIMA model. The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in a SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0. By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0.

Manual SARIMA result:   
Table

Description automatically generated  
1.61 Manual SARIMA result

Plot of SARIMA at 6:  
Chart, line chart

Description automatically generated  
1.62 Plot diagnostic for SARIMA at 6

RMSE value:   


Table

Description automatically generated  
1.63 RMSE values for manual SARIMA at 6 and all the other models built

Sarima at 12:   
Graphical user interface, chart, line chart

Description automatically generated  
1.64 Graph for SARIMA at 12

We see that there might be a slight trend which can be notices in the data. So, we take a differencing of first order on the seasonality differenced series.

SARIMA at 12 1st order difference on seasonally differenced series

Graphical user interface, chart, application, scatter chart

Description automatically generated  
1.65 Graph for dataset SARIMA at 12 1st order difference on seasonally differenced series

Graphical user interface

Description automatically generated

ACF and PACF plot : SARIMA at 12

Timeline

Description automatically generated with medium confidence  
  
1.66 ACF and PACF plots for newly modified time series SARIMA at 12

SARIMA results:   
Table

Description automatically generated  
1.67 SARIMA at 12 result

SARIMA at 12 Plot:   
Chart, line chart

Description automatically generated

1.68 Plot diagnostics SARIMA at 12

RMSE value:   
Table

Description automatically generated



Graphical user interface, table

Description automatically generated  
1.69 RMSE values for Manual SARIMA at 12 and all the other models built

8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.   
  
Table

Description automatically generated  
1.70 Dataframe models built along with their RMSE

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.

Building the most optimum model on the Full Data.

List of prediction for next 12 months

Graphical user interface, text, table

Description automatically generated  
1.71 List of prediction for next 12 months

Chart, line chart

Description automatically generated  
1.72 Graph of prediction on full dataset

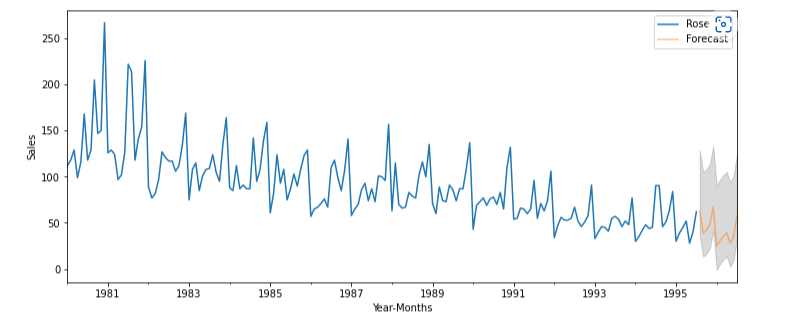
Building the most optimum model on the Full Data. We see that the best model is the 2 Point Trailing Moving average. We see that the best model is TES with multiplicative seasonally with the parameters alpha = 0.1 , beta = 0.01 and gamma= 0.05. The same thing is put into a dataframe for building the most optimum model on the complete data. The RMSE score for the full model is 18.970

Graph of prediction on full dataset:   
Chart

Description automatically generated  
1.73 Graph of prediction on full dataset

Class Intervals:   
 Text

Description automatically generated  
1.74 Class intervals

Graph of class intervals and next 12 months prediction:   


Graphical user interface, chart

Description automatically generated

1.75 Graph of class intervals and next 12 months prediction

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

It can be noted that from the given data, given that Rose wine has the highest sale in the year 1980 in the month of December followed by November and October which had a greater number of Rose wine sales when compared to other months. The decomposition of the data was first done using. Additive decomposition but it resulted in residuals to have pattern around 0 so when multiplicative method was applied the data did not show any pattern for the residuals and the components were located around 1. The month of January, February, June, and May have the lowest number of Rose wine sales across the years. Several models were built such as SES, DES, TES, SA, MA, Linear Regression. The lowest RMSE value is optimal for creating the best model for overall predictions. These models were built on training data and performance was checked on test data using RMSE. Based on the models built such as TES with parameters alpha=0.1 , beta=0.01, gamma=0.50 had the lowest AIC value at 18.970.

Based on the predictions from the models, with the use of TES as the best fit model so future sales predictions for next 12 months can be done. So, the company now has an insight on the profit, loss, production quantity, materials needed to produce Rose wine. Based on the insights the company can forecast which month will have higher or lower production with respect to the demand and supply for the month. Various discounts and offers can be given to both the sale team as well as people who purchase the product to boost the sales. To make the business more profitable, the company can use various strategies such as advertisements in Radio, Television, Local newspapers, pamphlets etc to grow the consumer base, and to boost sale and production for the time where the sale is low. Special offers at certain events can be discussed to uplift the production at low seasons. In future, the company can have low production for months with lowest sale to avoid wastage which in turn helps with the profit.