TIME SERIES FORECASTING ASSIGNMENT

BY JASPER SHELDON M

Contents

SPARKLING WINE

ROSE WINE

SPARKLIN WINE

ROSE WINE

ACRONYMS:

RMSE: Root Mean Square Error

AIC: Alkaike Information Criteria

SES: Simple Exponential Smoothing

DES: Double Exponential Smoothing

TES: Triple Exponential Smoothing

Problem Statement:

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

SPARKLING WINE:

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

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

Table 1 Head of the Data

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

Table 2 Tail of Data

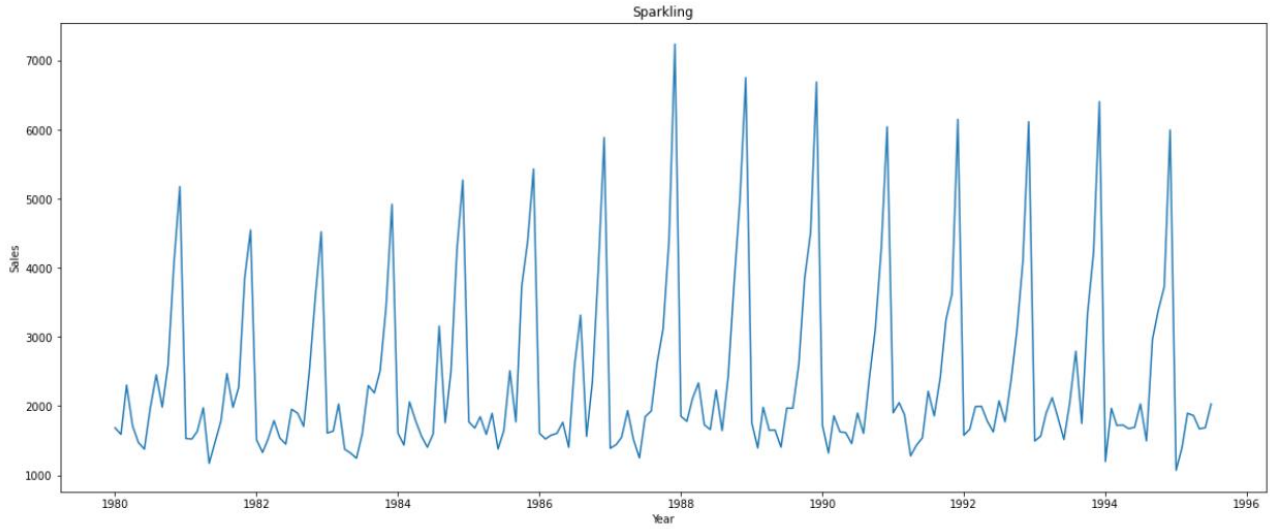

Figure 1 Overall plot of the data

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

	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

Table 3 Summary of the data

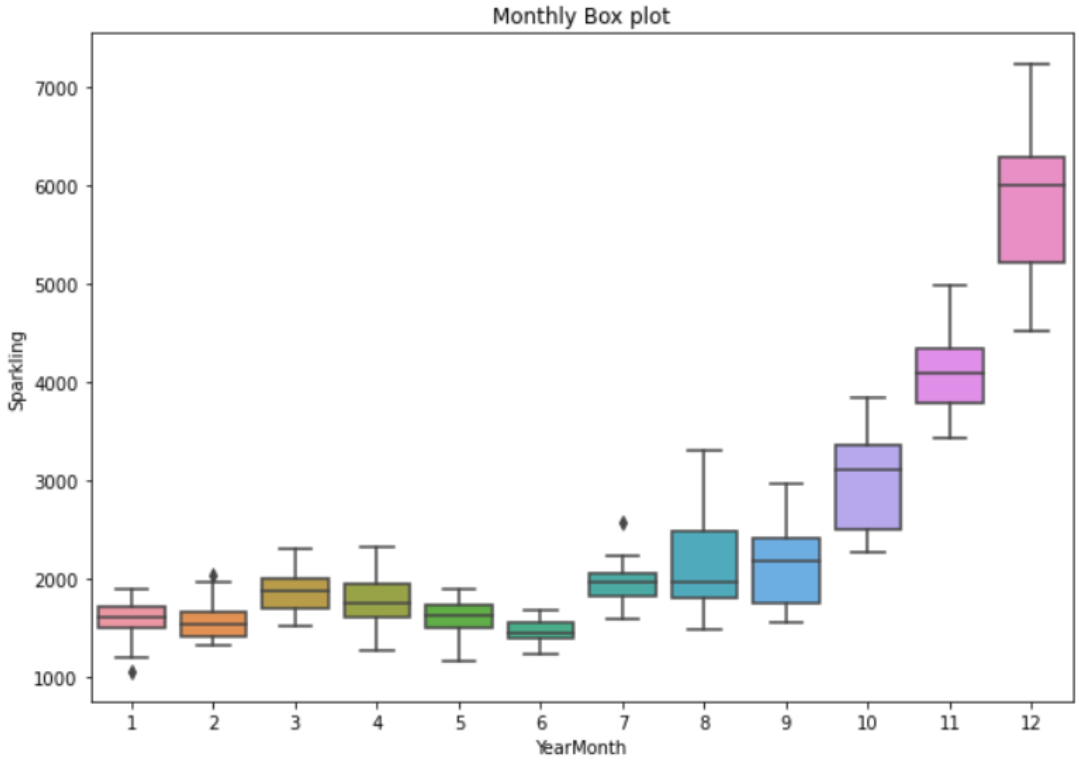

Figure 2 Boxplot throughout all months

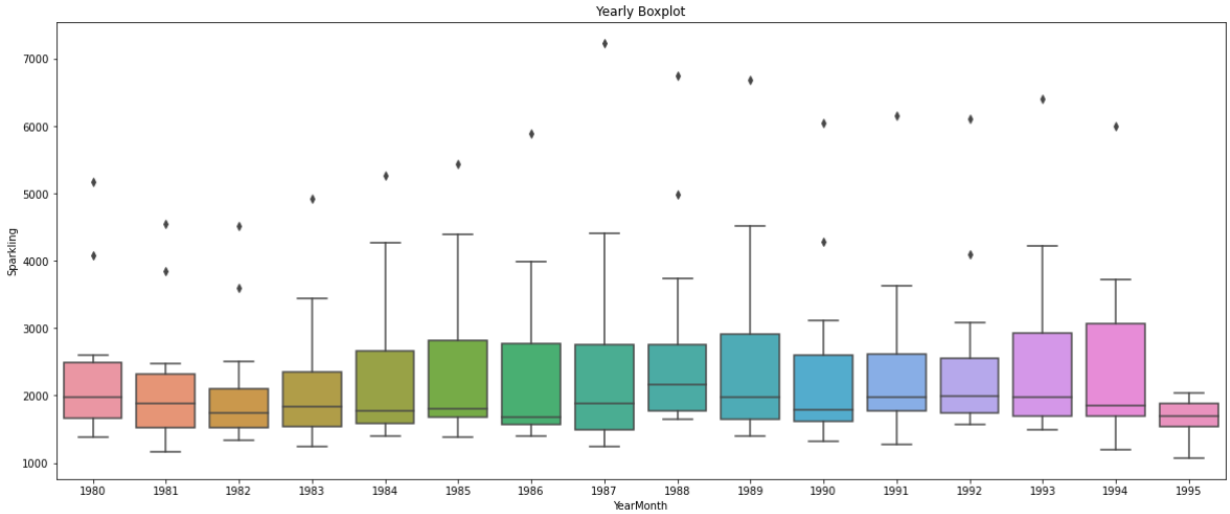

Figure 3 Boxplot throughout all years

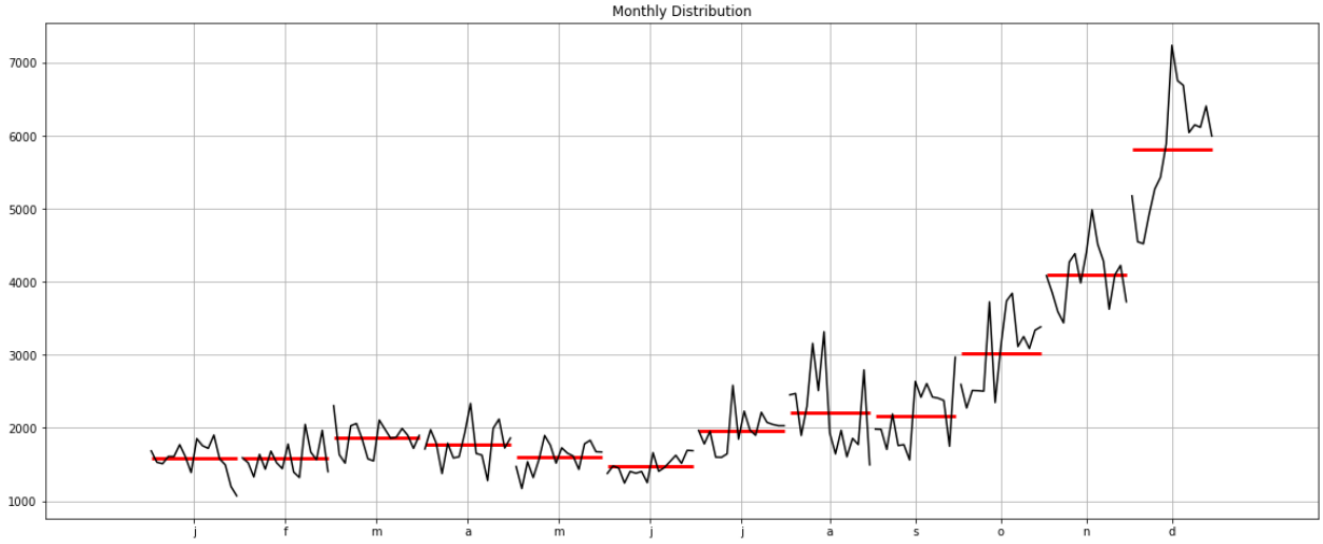

Figure 4 Monthly sales for all the years

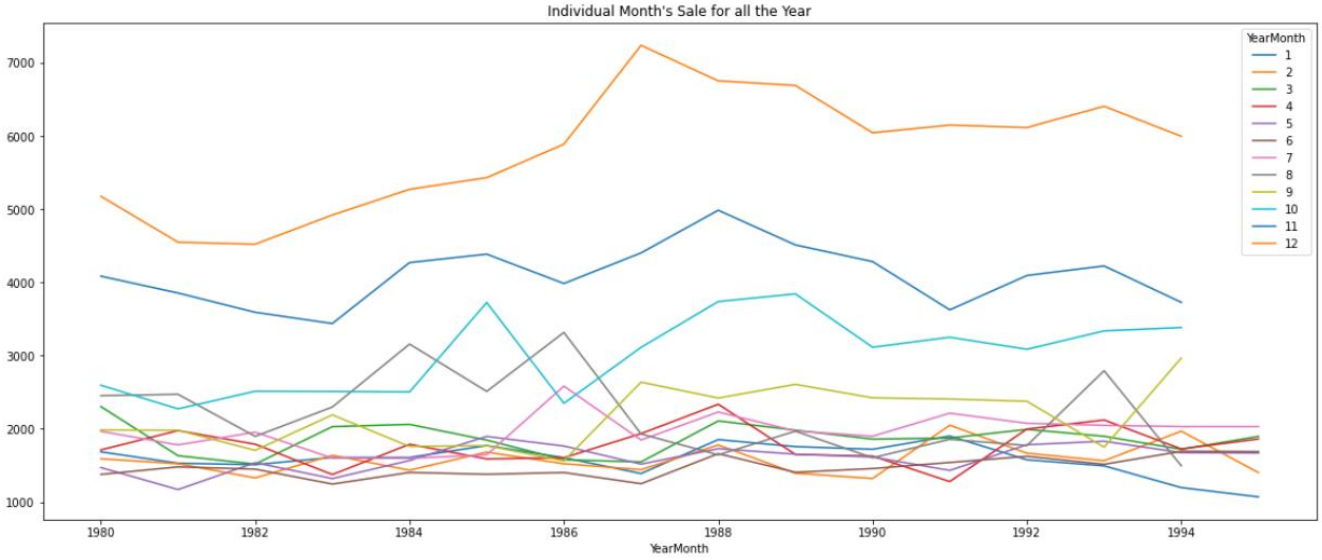

Figure 5 Monthly sales

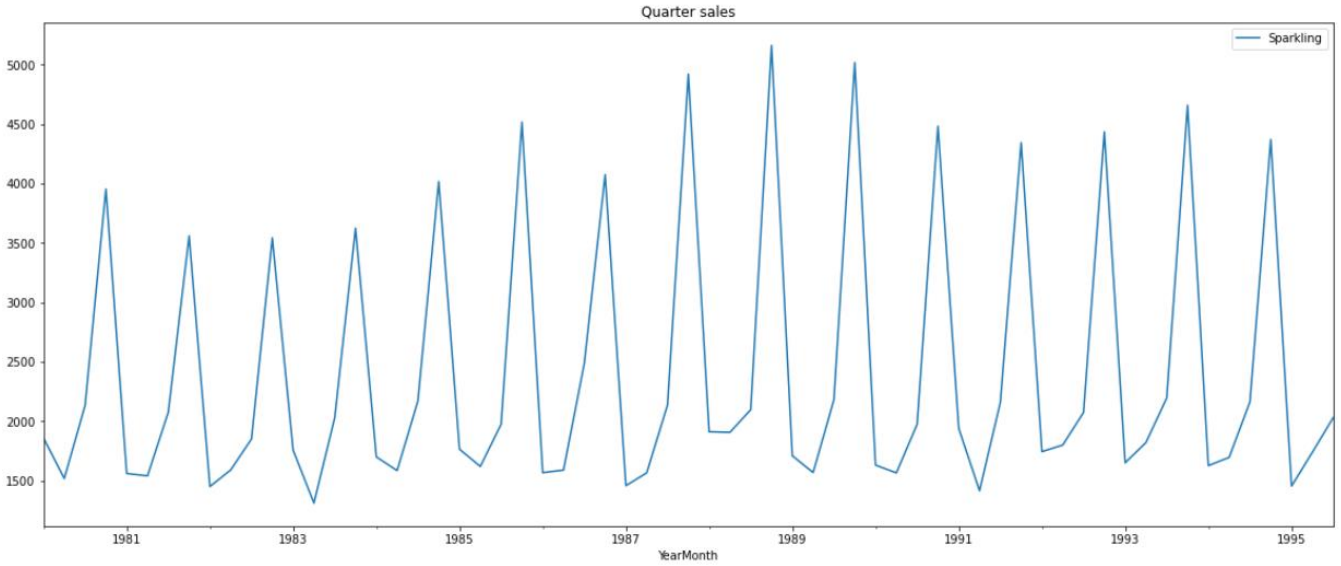

Figure 6 Quarterly Sales

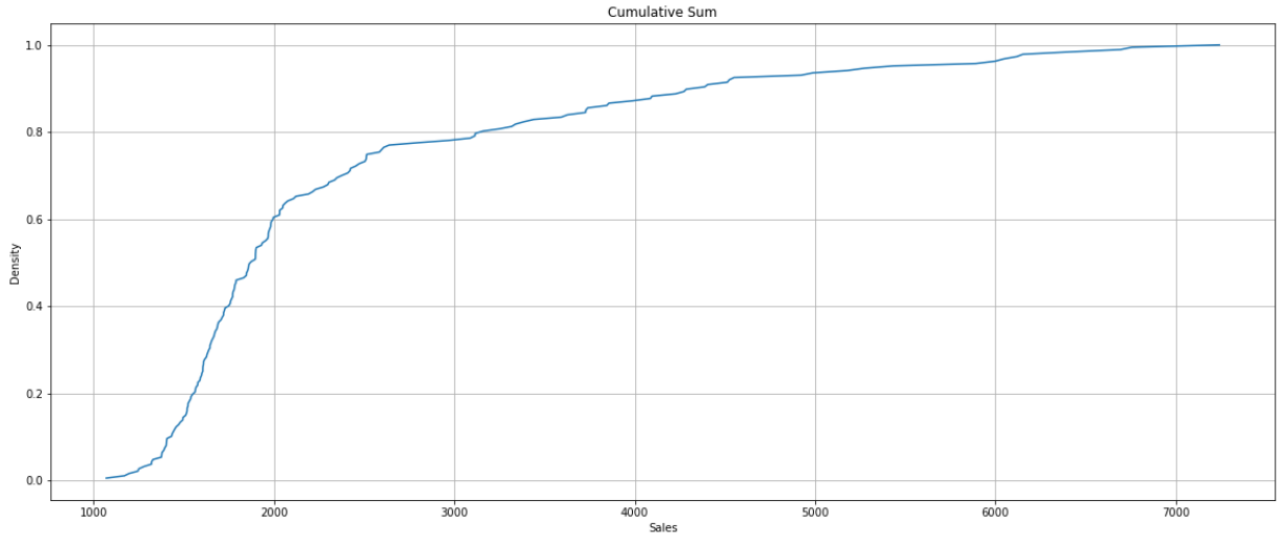

Figure 7 Cumulative Sum plot

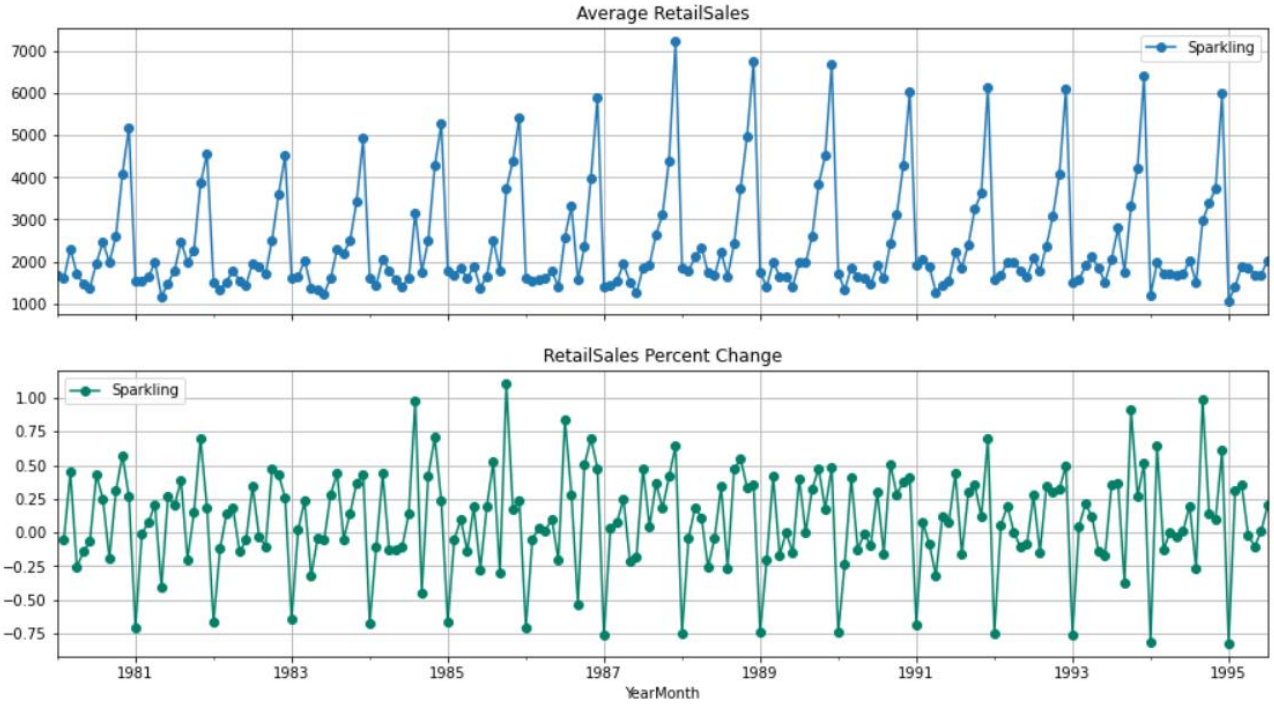

Figure 8 Plot of Avg. Sales and Percentage Change

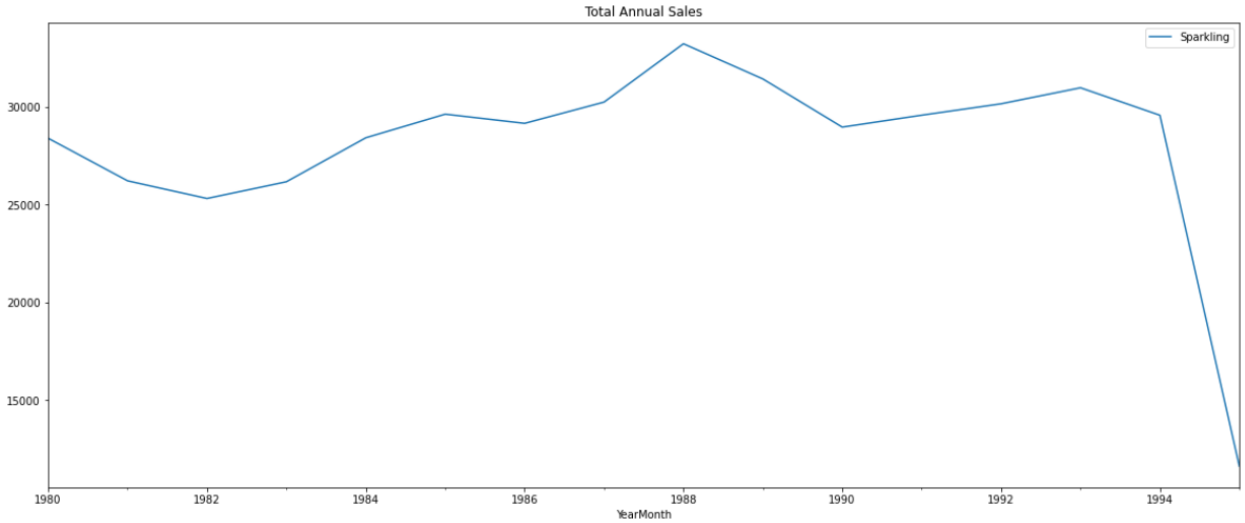

Figure 9 Total sum of sales per year

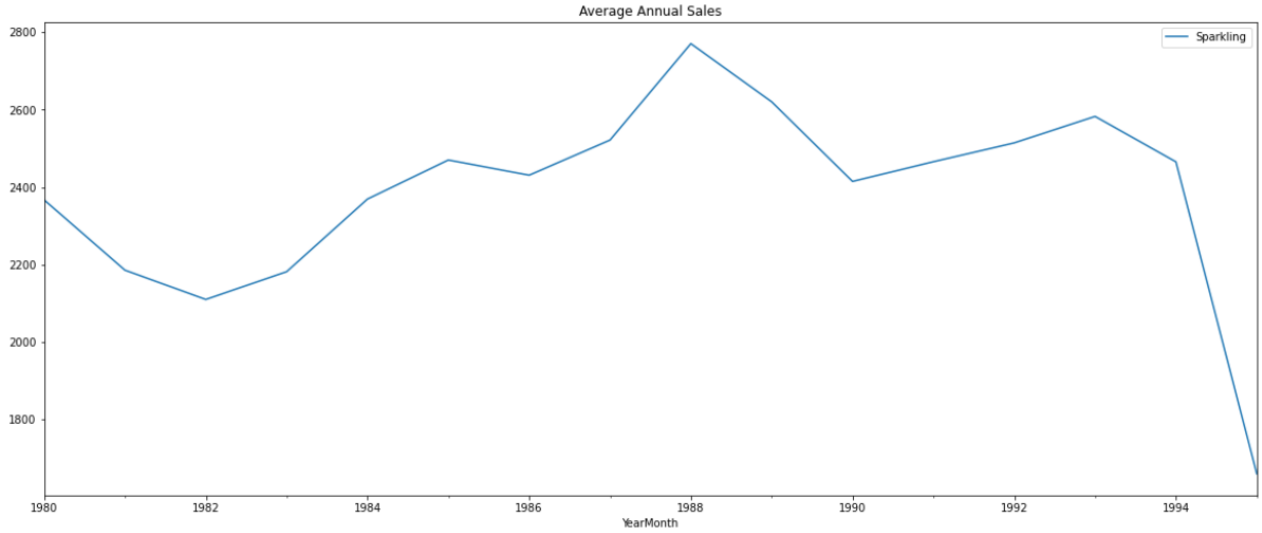

Figure 10 Average sales per year

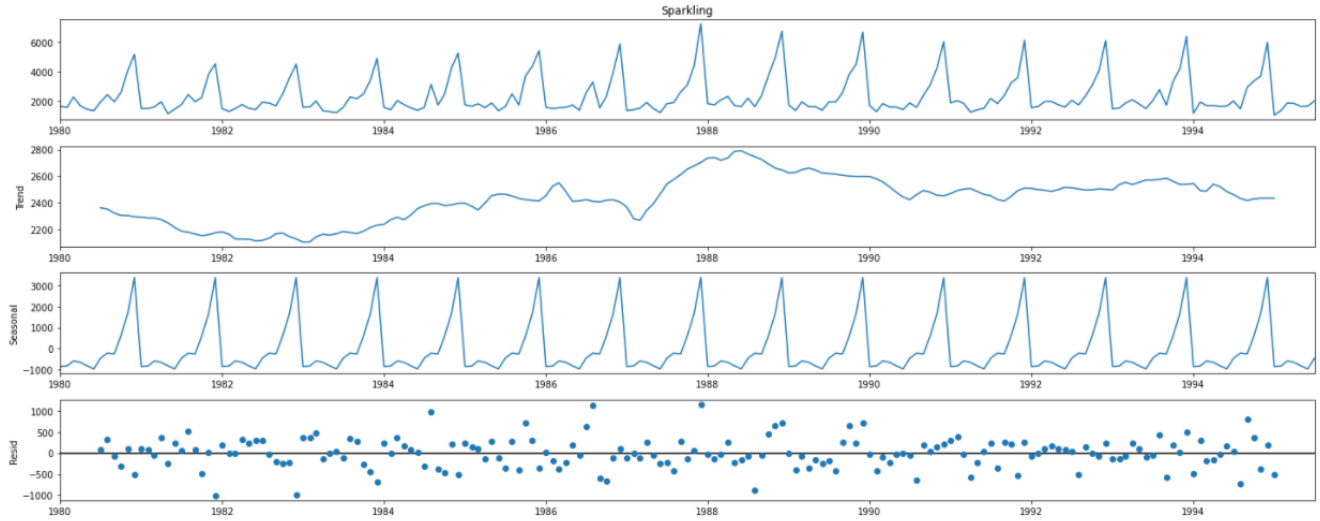

Figure 11 Seasonal Decomposition

Observations:

- ✓ There are a total of 187 data points and no missing values present in the dataset.
- ✓ The overall trend peaks at the year 1988 and it decreases a bit in the later years.
- ✓ Seasonality is present in the data, it has a small peak in the months of March and April and deeps in June and after August it starts to increase continuously and reach the highest peak in December.
- ✓ There are few outliers in every year which might be due to increased sales in December.
- ✓ From the cumulative density plot we can see 40% of the data is above 2000 and 20% of the data is above 3000.

- ✓ From the decomposition plot we can clearly see that the series contains trend and seasonality in them.
- ✓ By the end of 1994 the sales drops to the minimum value.
- ✓ The series is additive in nature.

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

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

Table 4 Train dataset

The train data consists of data from January of 1980 to December of 1990, it consists of 132 rows and 2 columns.

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Table 5 Test dataset

The test data consists of data from January of 1991 to July of 1995, it consists of 55 rows and 2 columns.

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.

1. Linear Regression:

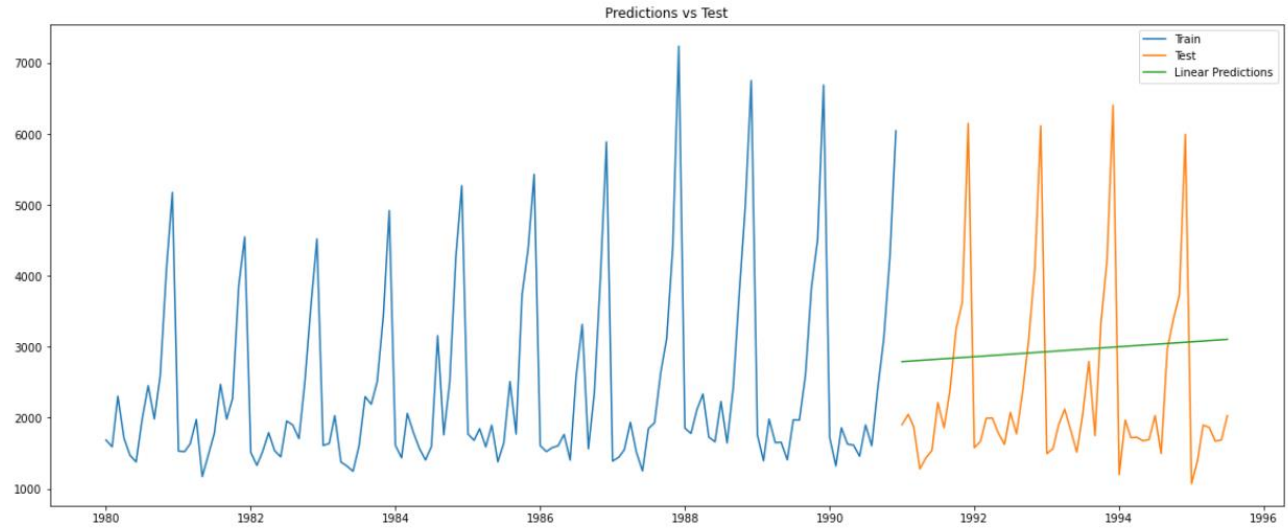

Figure 12 Performance of Linear Regression

We found the RMSE value between the test and the prediction to be RMSE: 1389.135 .The prediction shows that the model is not good enough it has more bias in it and hence high RMSE value.

2. Naïve Model:

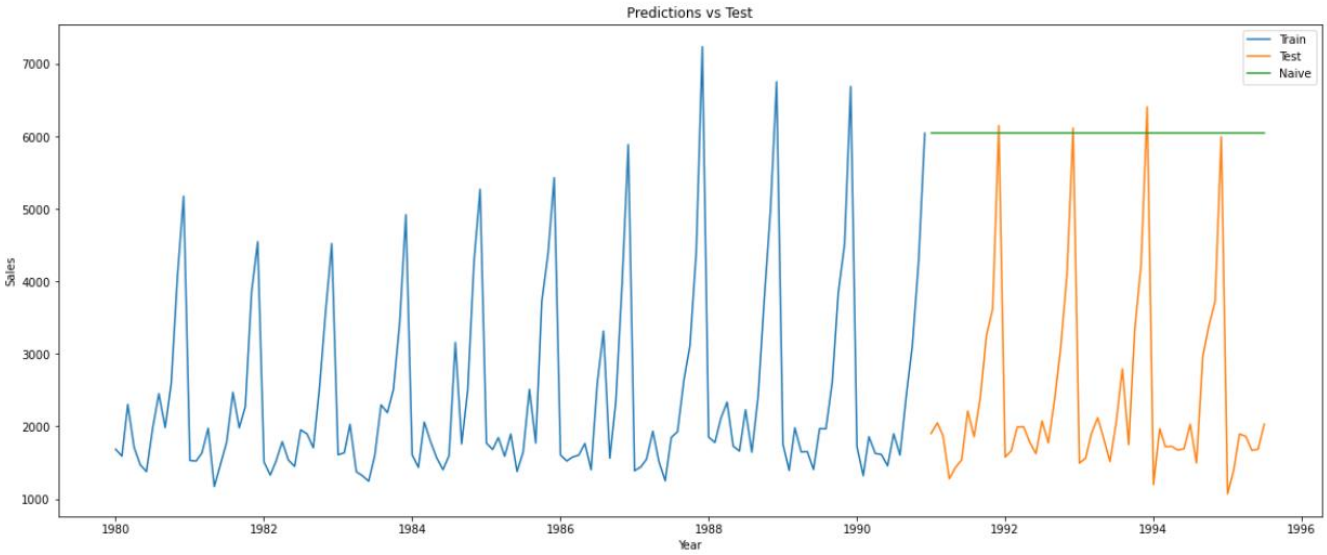

Figure 13Performance of Naive model

RMSE value between the test and prediction was found to be RMSE: 3864.279

Naïve model will assume the last value of the train data for forecasting the future values, from the plot we can see it has predicted the peak value for all the test data points. Due to this reason the RMSE is high and the model is inefficient.

3. Simple Average:

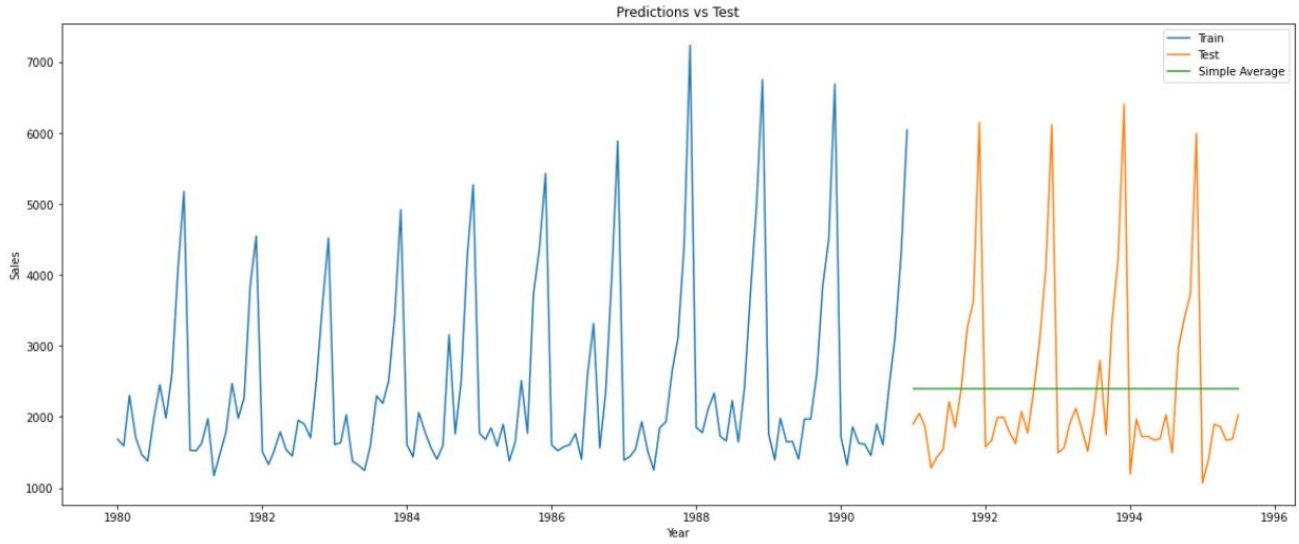

Figure 14 Performance of Simple average model

RMSE value between the test and prediction was found to be RMSE: 1275.082

Simple average model will calculate the average of the train data and will use it to forecast. This might be better than naïve model but still the model is inefficient.

4. Moving Average:

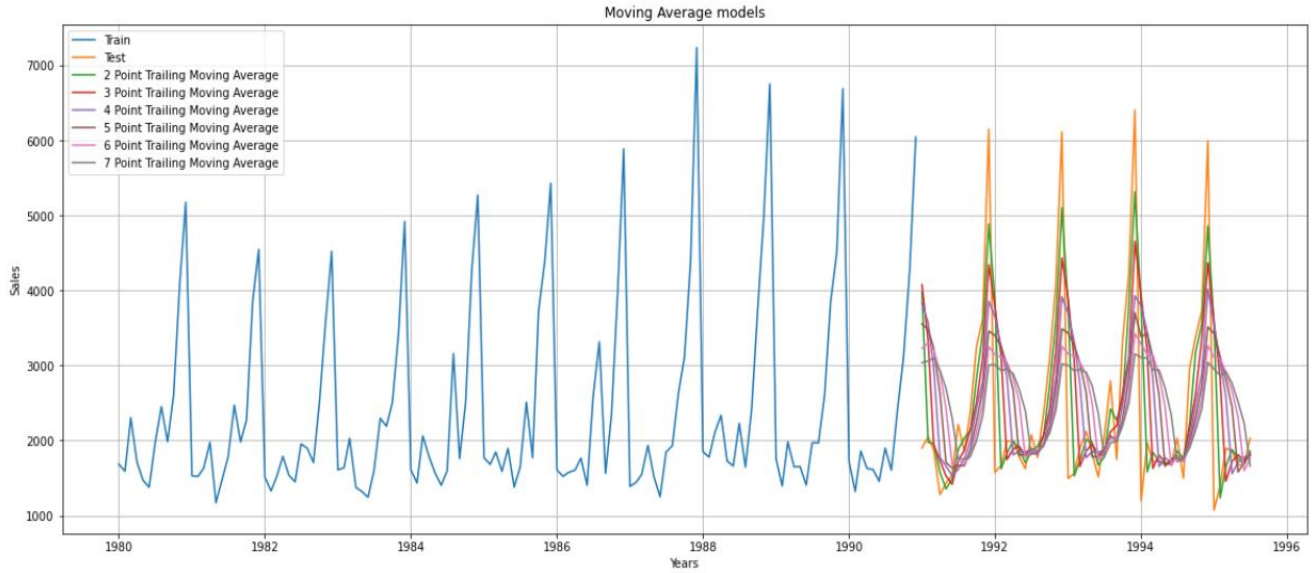

Figure 15 Different window sized moving average models

Moving average model tries to take the average within a particular window for each data point. If the window size is 2 it will take the mean of t-1 and t-2 data points. By trying out different window sizes we got the following results.

For 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401

For 3 point Moving Average Model forecast on the Training Data, RMSE is 1028.606

For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.580

For 5 point Moving Average Model forecast on the Training Data, RMSE is 1234.045

For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

For 7 point Moving Average Model forecast on the Training Data, RMSE is 1331.163

From the above we can clearly see window size of 2 is the most efficient of all.

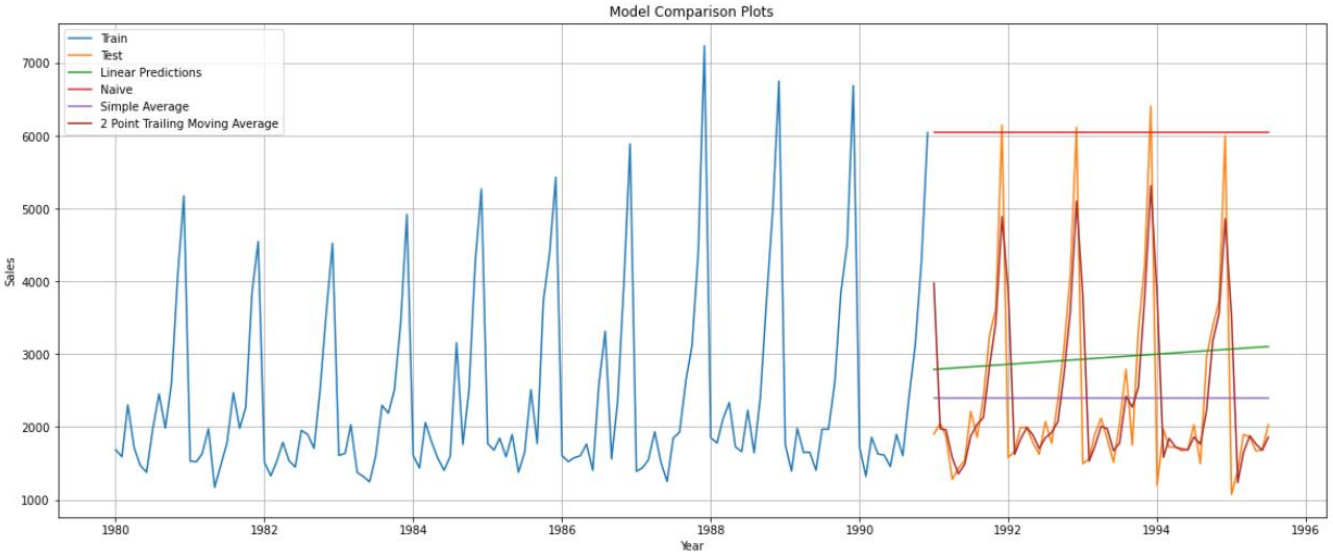

Figure 16 Model Comparison

The above plot contains the predictions of Linear, Naïve, Simple average and Moving average of window size 2. Within these models we can see moving average is the closest to the test data with the least RMSE value.

5. Simple Exponential Smoothing:

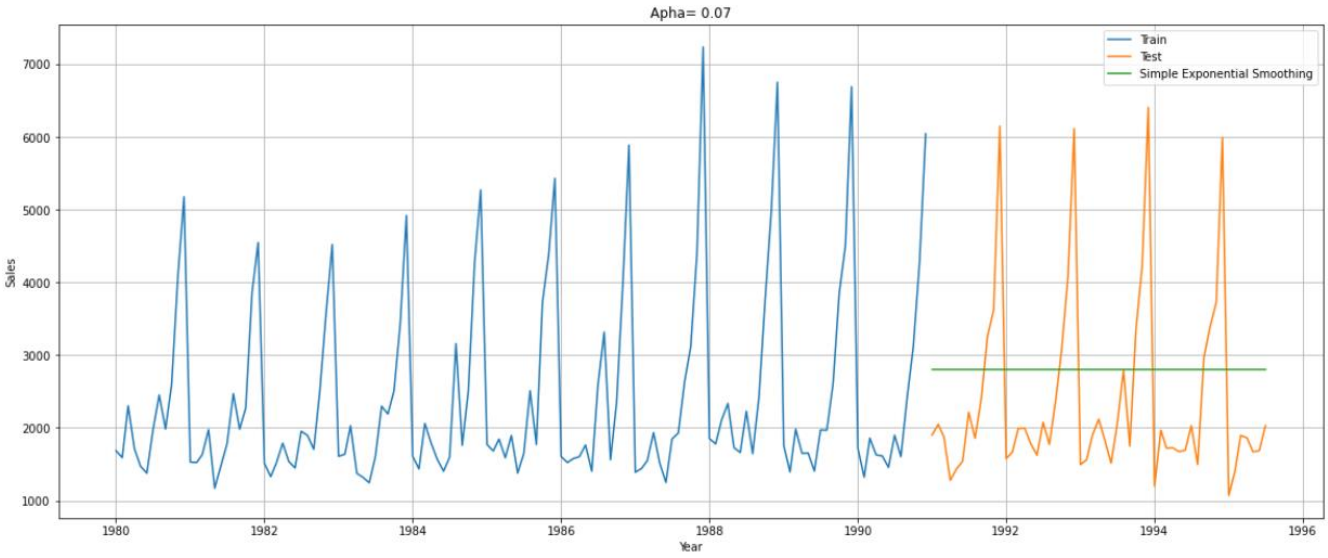

Figure 17 Simple exponential smoothing (alpha=0.07)

Smoothing technique means considering the past values and creating a relationship between the past values and the present. Exponential smoothing means giving more weightage to the recent data and least importance to old data. The importance factor diminishes exponentially as we move back in time.

Simple Exponential smoothing smoothens only the value of the time series, from the plot above we can see the prediction is higher than simple average. Although since the data has seasonality this is inefficient we found the RMSE value to be 1338.008. The smoothing parameter for the model is denoted by alpha and found out to be 0.07.

6. Double Exponential Smoothing:

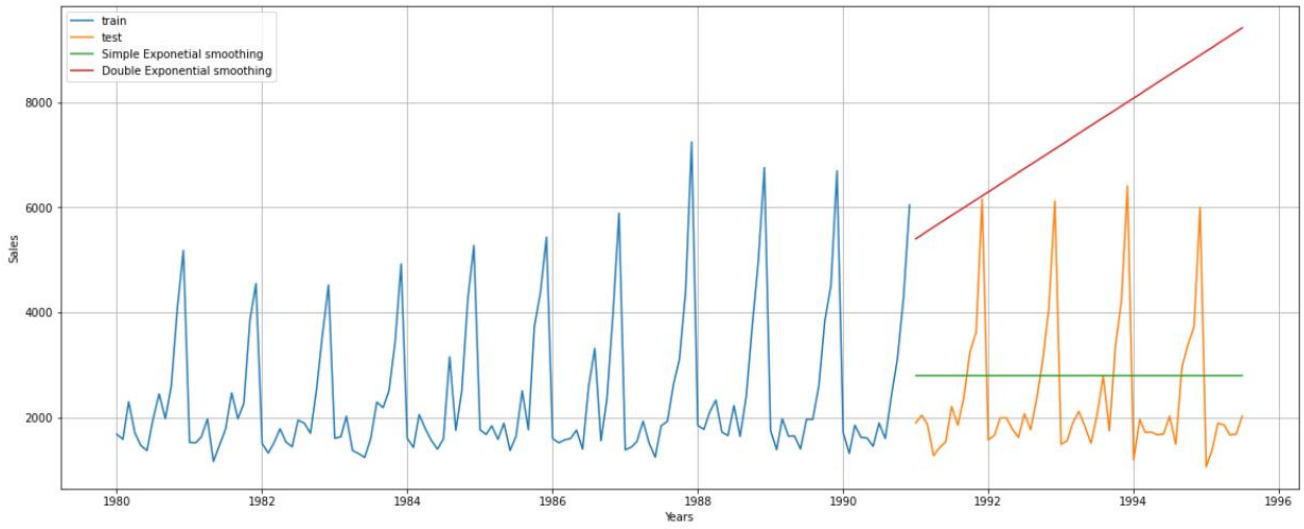

Figure 18 double exponential smoothing

In Double exponential smoothing we'll consider both the value and the trend for smoothing and both will be used to forecast the future values. Similar to SES this model also find it difficult to predict due to the presence of seasonality. As this model doesn't involve seasonality in prediction.

The RMSE value was found to be 5291.880 and the parameter for smoothing level is denoted by alpha which is found out to be 0.665. Parameter for smoothing trend is denoted by Beta and found out to be 0.0001.

7. Holt winter's model:

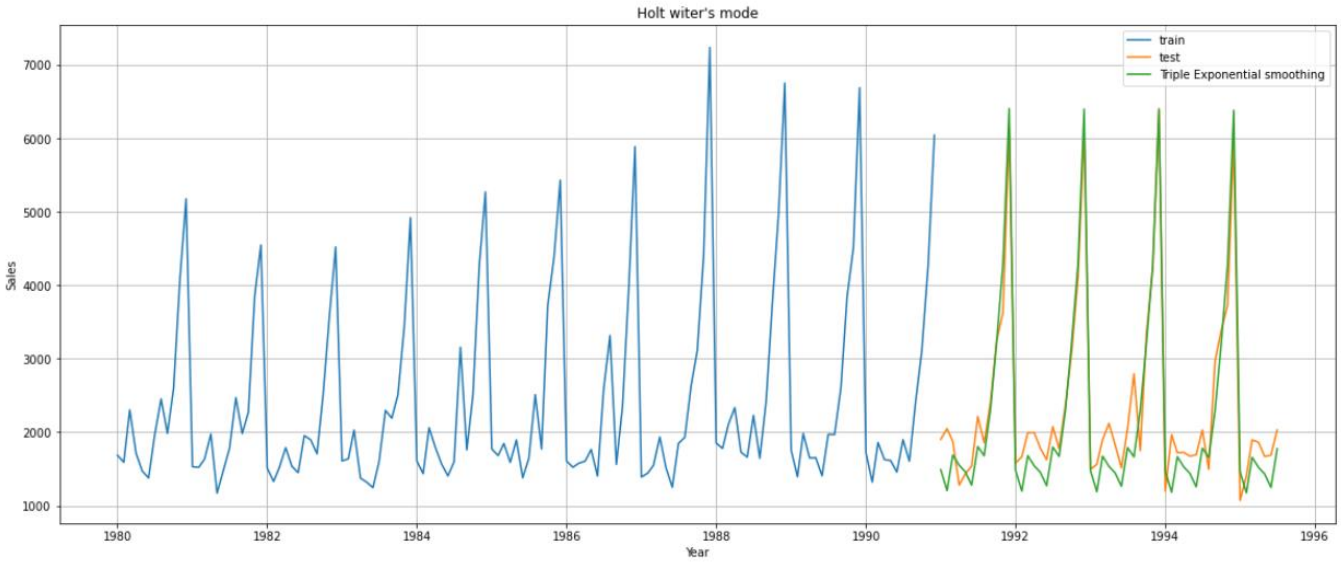

Figure 19 Holt Winter's Performance

Unlike the above two models Holt winter's model will smoothen all the 3 components that is value, trend and seasonality. So because of this reason this model will perform better than the other models. By evaluating the prediction with the test data we found the RMSE value to be 378.951 which is the least from all the above models. Similar to the previous model it denotes value and trend as alpha and beta and seasonality as gamma. The value for all 3 are as mentioned, alpha=0.111, beta=0.012 and gamma=0.460.

Observation:

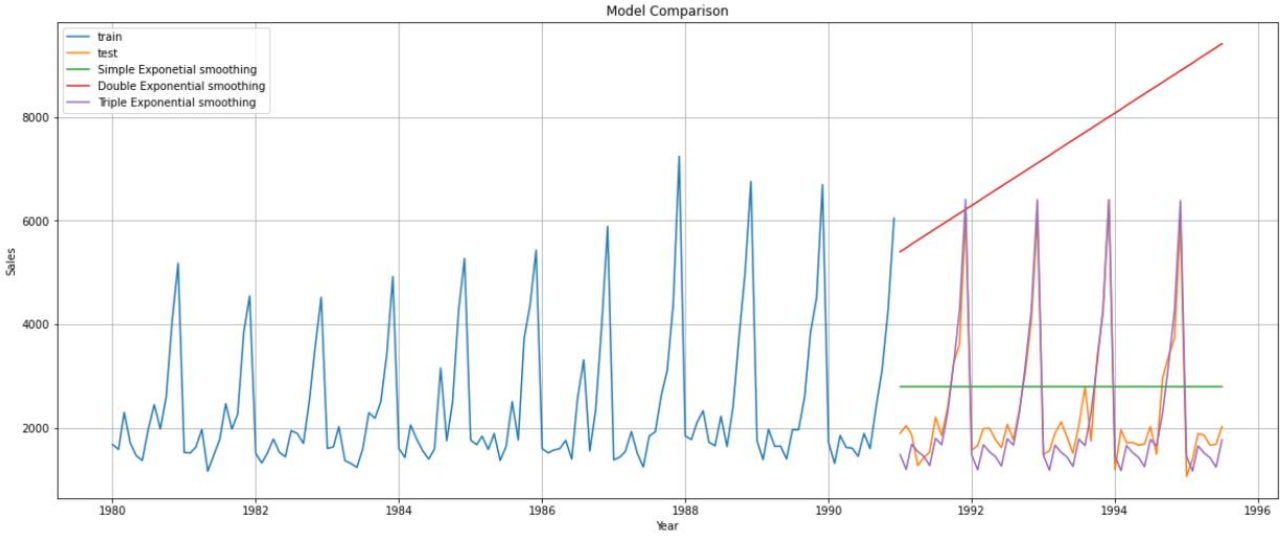

Figure 20 Exponential model comparisons

Within the exponential methods we can clearly see the Holt winter's method (Triple exponential smoothing) is the closest to the test data and with the least RMSE value. The prediction traces the test data very closely.

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.

A series is said to be stationary when it has all constant statistical properties such as variance and auto correlation over time. We can use a statistical test called Dickey Fuller test to check for stationarity. The hypothesis for the test are as follows:

H_0: The series is non – stationary

H_a: The series is stationary

After performing the test on the series we found the test statistic to be -1.798 and p value to be 0.705. As we can see the p value is greater than the alpha value 0.05 we can conclude we don't have enough evidence to reject the null hypothesis, therefore the series is not stationary.

In order to make the series stationary we can take the first difference on the series and check for stationarity. If it still doesn't become stationary we can keep performing the differencing or try any transformation on the series. In the following series we have performed the first differences and the series is shown below.

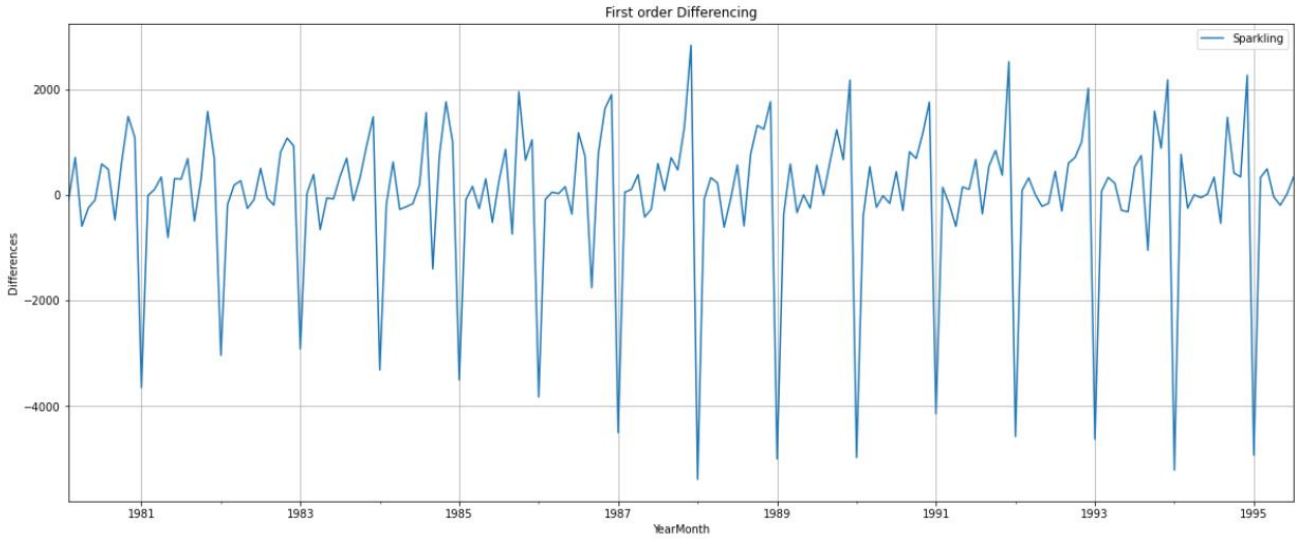

Figure 21 Differentiated Time Series

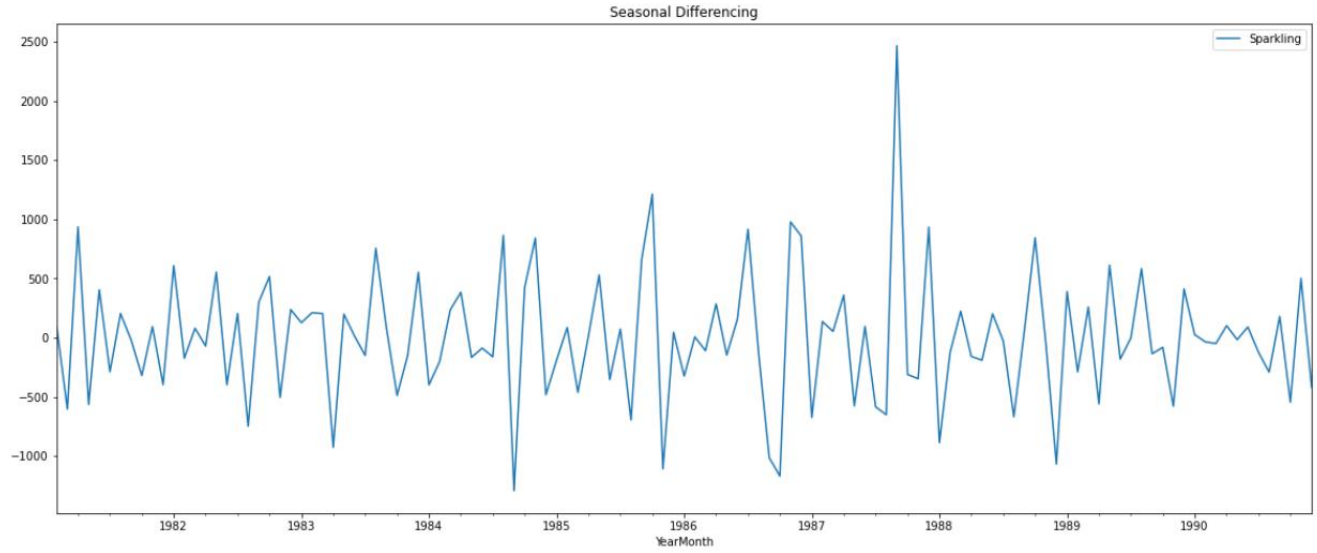

Figure 22 Series after 12 Differencing

After differentiating and performing the test we found the test statistic to be -7.96 and p value to be $8.47e^{-11}$. As the p value is very insignificant we can conclude we have enough evidence to reject the null hypothesis and hence we can conclude the series is stationary. Although the series has become stationary we can see the data has some trend in it so we must perform the seasonal differencing to remove the trend. The above results were for the entire dataset, the same test were done on the train dataset and found the p value to be same as $8.47e^{-11}$.

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.

A brute force approach is used to find the best parameters for non-seasonal component and the seasonal component of the sarima model and the models with the best AIC values are shown below.

	param	seasonal	AIC
203	(1, 1, 2)	(0, 1, 2, 12)	1382.347780
95	(0, 1, 2)	(0, 1, 2, 12)	1382.484254
209	(1, 1, 2)	(1, 1, 2, 12)	1384.137874
311	(2, 1, 2)	(0, 1, 2, 12)	1384.317618
101	(0, 1, 2)	(1, 1, 2, 12)	1384.398867

Table 6 Sarima models with AIC values

From the table above we can see the model (1, 1, 2) (0, 1, 2, 12) has the least AIC value so we proceed with trying this model on train data and find its performance on the test data.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-685.174
Date:	Fri, 17 Dec 2021	AIC	1382.348
Time:	17:36:17	BIC	1397.479
Sample:	01-01-1980 - 12-01-1990	HQIC	1388.455
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

Table 7 Summary of the auto Sarima model

From the summary table above we can see it has significant values for coefficients but the p value for ma.L1 (moving average for lag 1) and ma.S.L24 (moving average for seasonality lag 24) are significant so we can say these parameters might not be significant enough to perform well. The model might perform moderately.

As discussed above after building the model when tried on the test data we found the RMSE (Root mean squared error) to be 382.577. Which is greater than Holt winter's model.

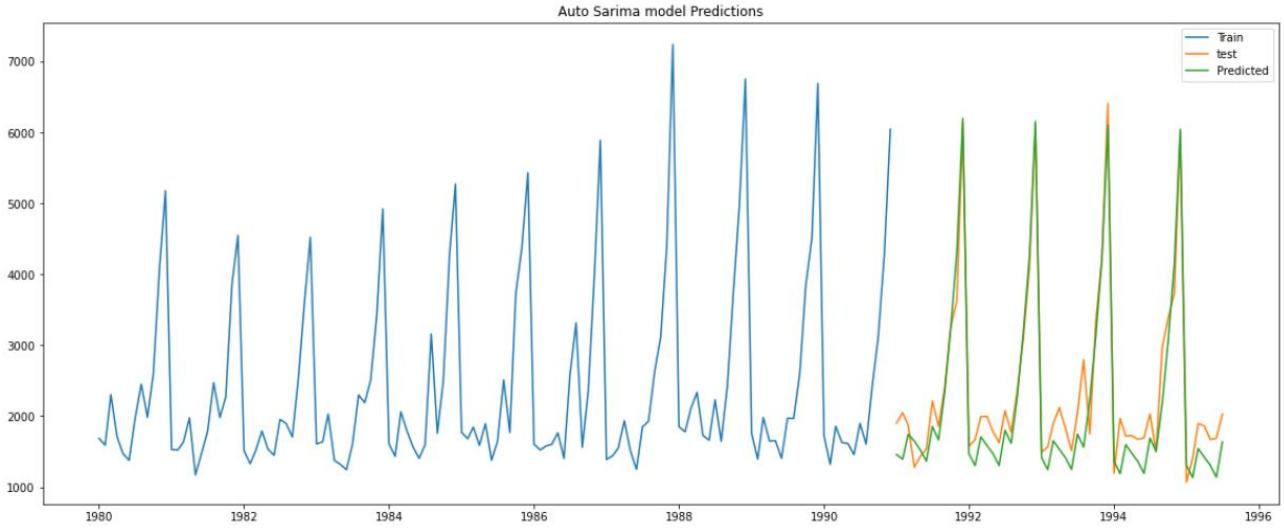

Figure 23 Performance of the auto sarima model

Different models with varying the seasonality were tried. 6 and 24 were tried for seasonal factor and the following is the result.

	RMSE	AIC
SARIMAX(1, 1, 2)X(0, 1, 2, 12)	382.576734	1382.347780
SARIMAX(2, 1, 3)x(1, 2, 3, 6)	558.318086	1640.930651
SARIMAX(1, 1, 2)x(0, 1, 2, 24)	320.549565	870.869858

Table 8 Sarima models

From the table above we can see the model with seasonality 6 is not performing well whereas model with seasonality as 24 has the least RMSE and AIC but the model could not be explained as we cannot have 24 seasons within a year so we are ignoring the model and choosing the model with seasonality 12 as the best model.

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.

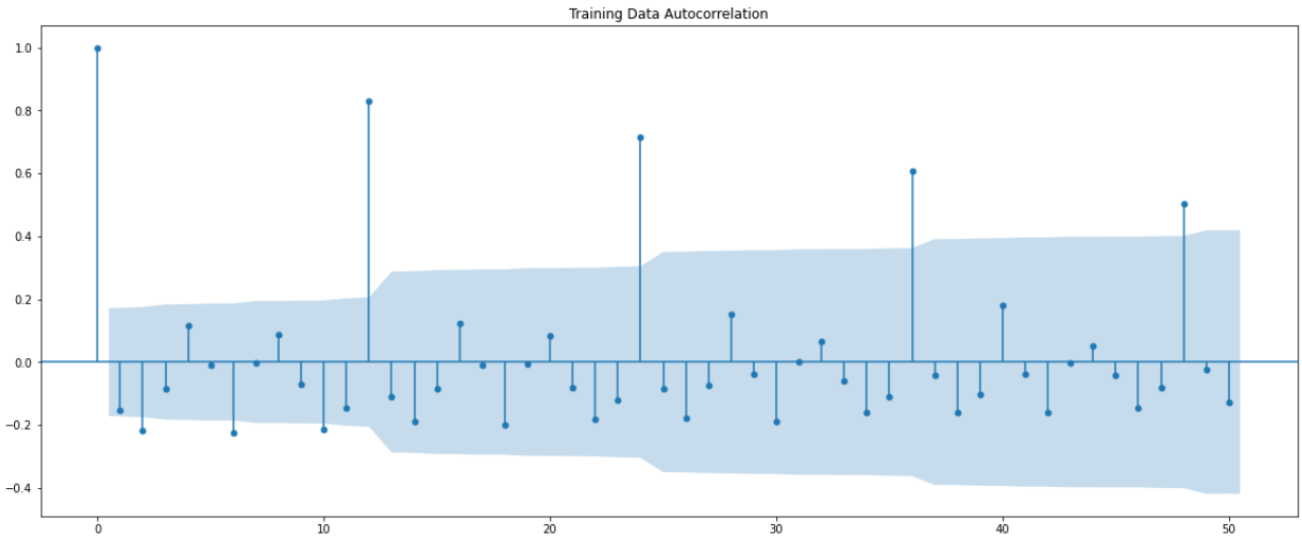

Figure 24 ACF plot on 1st Differentiated series

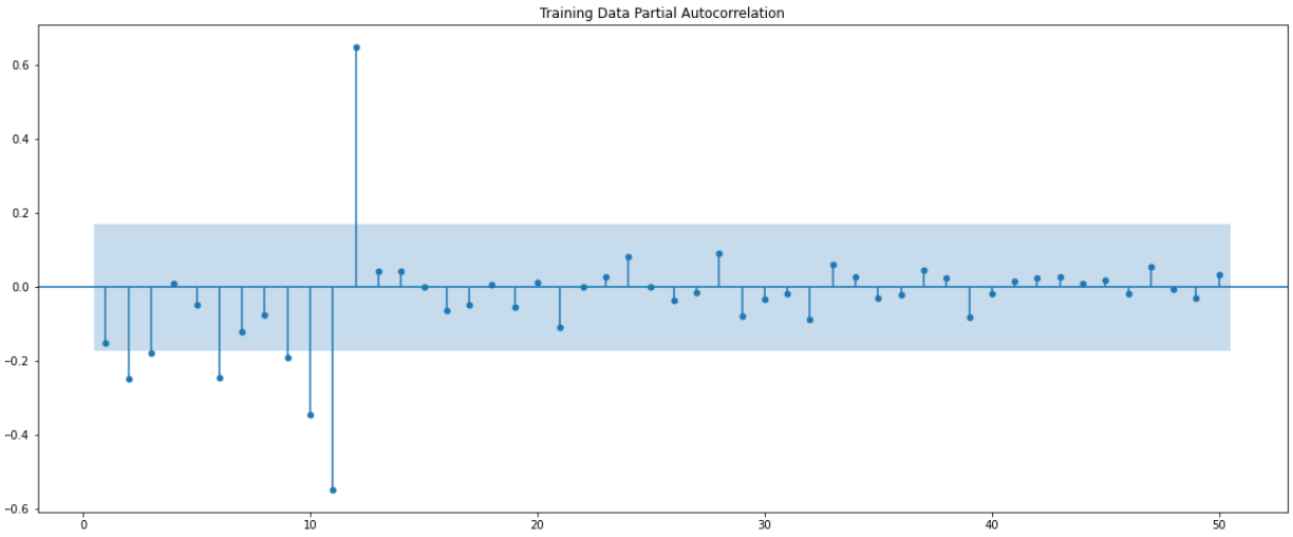

Figure 25 PACF plot on 1st Differentiated series

From acf plot we can find the value for moving average (q) and from pacf plot we can find the value for auto correlation (p) by finding the first lag where it cuts off to 0. From the acf plot we can see the first lag itself cuts off to 0 so the q value will be 0 and in the pacf plot the first lag itself cuts off to 0 so the p value is also 0.

We use to first order differencing to make the series stationary so d=1.

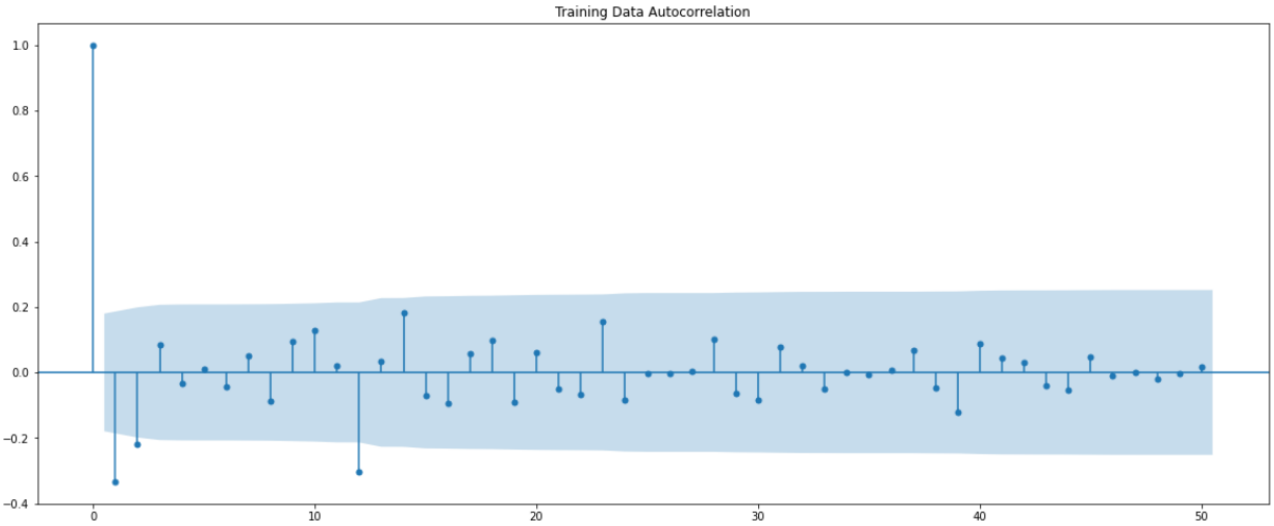

Figure 26 Acf plot seasonal differentiated data

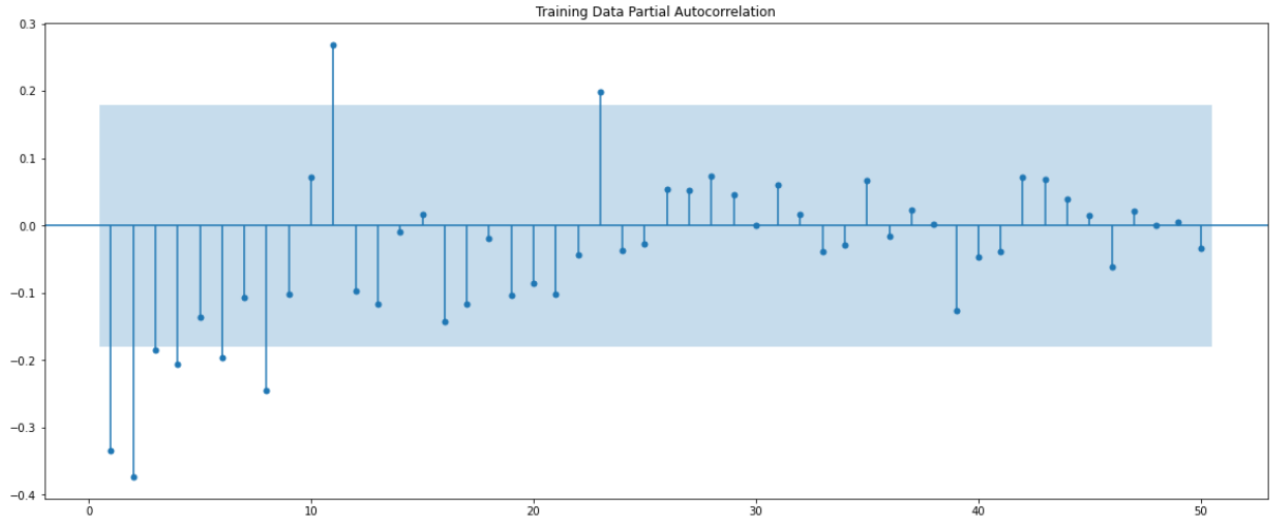

Figure 27 Pacf plot seasonal differentiated data

From the fig24 we can find the value of moving average of seasonal component (Q) by using the same method as before. (Lag at which first cuts off to 0) from the pacf plot we can find the auto correlation parameter (P) for seasonal component. From the plots we can see Q to be 2 and P to be 4. As we have done seasonal differencing (12) D is 1 and Seasonal parameter(S) is 12.

```
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(4, 1, [1, 2], 12)   Log Likelihood                -544.303
Date:                                 Sat, 18 Dec 2021   AIC                           1102.605
Time:                                         22:18:37   BIC                           1118.444
Sample:                                     01-01-1980   HQIC                          1108.904
                                          - 12-01-1990
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12       0.3381      0.215      1.573      0.116      -0.083       0.759
ar.S.L24      -0.4912      0.137     -3.580      0.000      -0.760      -0.222
ar.S.L36      -0.2017      0.104     -1.940      0.052      -0.405       0.002
ar.S.L48      -0.2922      0.164     -1.786      0.074      -0.613       0.029
ma.S.L12      -0.9563      0.230     -4.155      0.000      -1.407      -0.505
ma.S.L24       0.9977      0.141      7.070      0.000       0.721       1.274
sigma2      1.91e+05   1.01e-06    1.9e+11      0.000    1.91e+05    1.91e+05
==========================================================================================
Ljung-Box (L1) (Q):                   5.13   Jarque-Bera (JB):                 24.18
Prob(Q):                              0.02   Prob(JB):                          0.00
Heteroskedasticity (H):               0.35   Skew:                              0.99
Prob(H) (two-sided):                  0.01   Kurtosis:                          5.06
==========================================================================================
```

From this summary we see there are no non seasonal component parameters only seasonal component parameters are present and from the p value only two are insignificant in predicting.

This could be a moderate predictor.

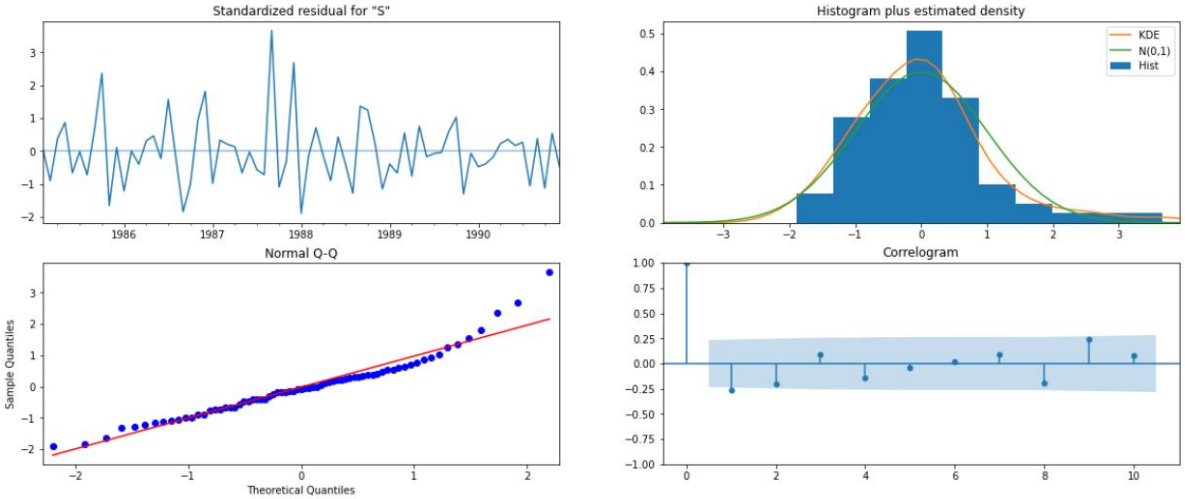

Figure 28 Manual sarima model diagnostics

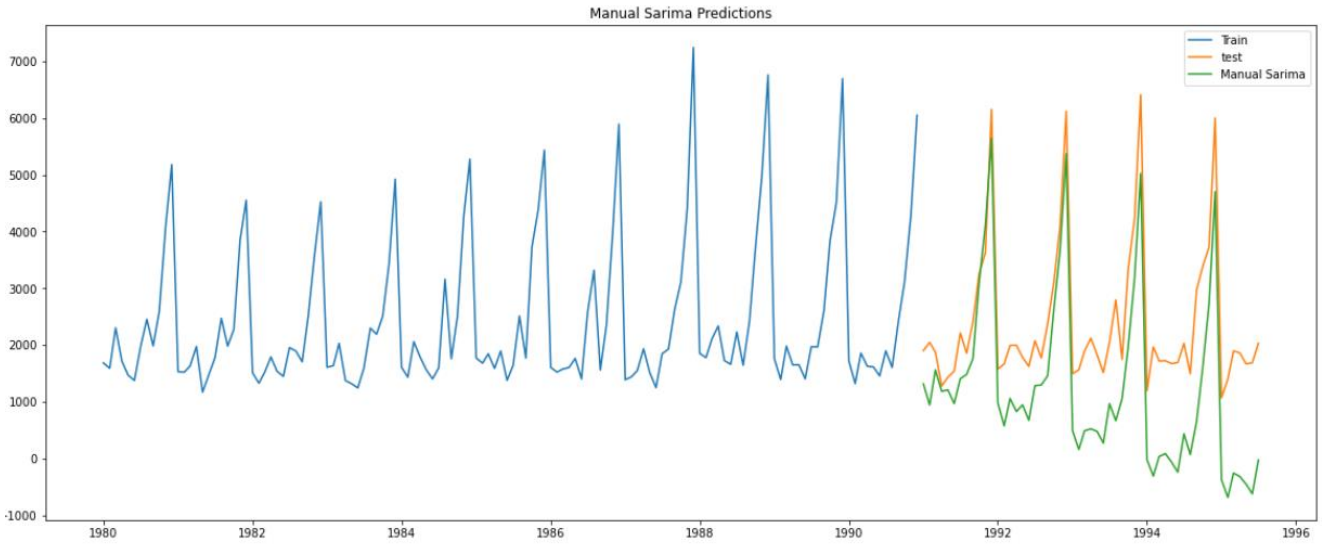

Figure 29 Manual Sarima performance on test data

From fig 26 we can see the standard error doesn't follow normal distribution and AIC is found to be 1102.605 and RMSE to be 1136.559.

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.

	RMSE
Linear_Regression	1389.135175
Naive	3864.279352
Simple_Avg	1275.081804
Trailing_2	813.400684
Trailing_3	1028.605756
Trailing_4	1156.589694
Trailing_5	1234.045344
Trailing_6	1283.927428
Trailing_7	1331.163342
Simple_Exponential_Smoothing	1338.008384
Double_Exponential_Smoothing	5291.879833
Holt_winter	378.951023
Auto_SARIMA(1,1,2)(0,1,2,12)	382.576734
Manual_SARIMA(0, 1, 0)(4, 1, 2, 12)	1336.558510

Table 9 Models Summary

From the table 8 we can see Holt winter model has the least RMSE value 378.951 we can choose this model to train on the entire dataset and forecast into the future.

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.

Holt winter model is trained on the entire dataset and found the following parameters.

Smoothing level: 0.0760,

Smoothing trend: 0.0326,

Smoothing seasonal: 0.377

RMSE on the fitted values of the model is 365.307 which is better compared to the other models.

	Forecast	ci_lower	ci_upper
1995-08-01	1877.431801	1159.508682	2595.354921
1995-09-01	2405.285747	1687.362628	3123.208866
1995-10-01	3242.105698	2524.182579	3960.028817
1995-11-01	3922.189524	3204.266404	4640.112643
1995-12-01	6118.502404	5400.579284	6836.425523
1996-01-01	1262.618990	544.695870	1980.542109
1996-02-01	1592.137914	874.214795	2310.061033
1996-03-01	1831.652945	1113.729826	2549.576064
1996-04-01	1806.470072	1088.546953	2524.393191
1996-05-01	1651.723185	933.800065	2369.646304
1996-06-01	1586.507708	868.584588	2304.430827
1996-07-01	1977.014975	1259.091856	2694.938095

Table 10 Forecast of Holt winter model

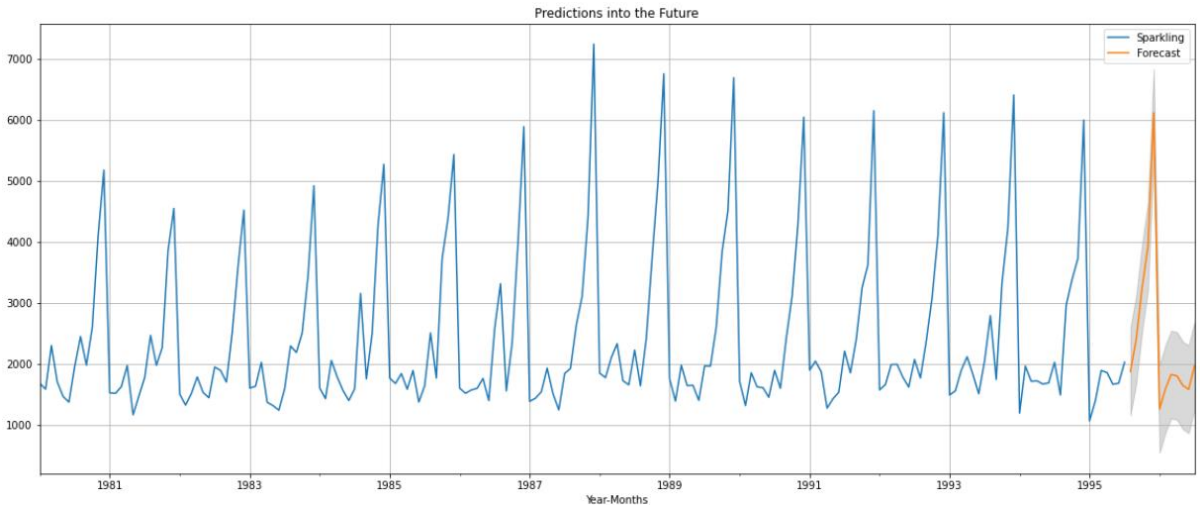

Figure 30 Forecast of Holt winter model

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

SUMMARY:

The data was read by parsing the date as index column, after that once we tried to understand the data we found that there were no missing values in the data and the entire series has a average of 2402.47 and standard deviation of 1295.11. The standard deviation is high indicating there might be high variations in the series. After that the data was explored and found the following insights from it.

- ✓ There are a total of 187 data points and no missing values present in the dataset.
- ✓ The overall trend peaks at the year 1988 and it decreases a bit in the later years.
- ✓ Seasonality is present in the data, it has a small peak in the months of March and April and deeps in June and after August it starts to increase continuously and reach the highest peak in December.
- ✓ There are few outliers in every year which might be due to increased sales in December.
- ✓ From the cumulative density plot we can see 40% of the data is above 2000 and 20% of the data is above 3000.
- ✓ From the decomposition plot we can clearly see that the series contains trend and seasonality in them.
- ✓ The series is additive in nature.

After finding the insights the data was divided into train and test. Data below 1991 was considered as Training data and above 1991 was considered as test data.

Various models like Linear regression, Naïve, Simple average ,moving average ,Exponential smoothing methods and SARIMA models were trained using the train data and tested with the test data. Each models' performance was measured using RMSE as the metric. In the above methods Regression needs an input variable to train on so the number of months were calculated for all the

data points and were passed as input variable for the model. For moving average different window size were tried and found out 2 is the optimum size.

While building the SARIMA model we must make sure the series is stationary. Stationary means all the statistical property like variance and correlation must not be dependent on time. We found out by default the series was not stationary so the first differencing was done and checked for stationarity. To check for stationary we use Dickey fuller test as the statistical test with null hypothesis – The series is not stationary. Even after taking the first order differencing the series showed trend in it so seasonal differencing was also performed before building the model.

PACF and ACF plots were plot and the values for the non-seasonal component were found out using the first order differentiated data. The values p and q were found out by finding the first cut from both the plots. For seasonal component same method were used to find P and Q from the seasonal differentiated series and the seasonal value S is found out by looking at the ACF plot by finding the regular intervals where we get a significant spikes.

A version of automated SARIMA was also built by using AIC(Alkaike Information Criteria). A brute force approach was done trying out different values for the model and found the model with the least AIC value. The least AIC means the better the model is. After building these to SARIMA models the RMSE value was calculated on the test data.

A table with all the models and its RMSE on the test data was created. From the table Holt winter model showed the least RMSE on the test data and it was trained on the entire data. Once trained on the entire data we found the smoothing parameters to be Smoothing level: 0.0760, Smoothing trend: 0.0326, Smoothing seasonal: 0.377 and the RMSE on the fitted values to be 365.07.

The final model has low RMSE on the train data indicating it has trained properly, it also has very good values for smoothing parameter indicating this will be a good predictor. This model is then used to forecast 12 months into the future. Holt winter method in python doesn't not calculate confidence interval for the predictions so it was manually done using the below formula:

$CI = Forecast \mp 1.96(std.error)$ @95% confidence interval

After finding the lower and upper confidence interval the forecasted values were plotted along with the confidence interval values next to the original data.

BUSINESS INSIGHTS:

- ✓ The sale of the Sparkling wine is more or less stable over the years.
- ✓ The sales increase after August month and spikes at December it might be due to Christmas and Holidays.
- ✓ The sales was more at 1988 than present it can be due to other competitors in the market.
- ✓ At present the overall sale has started to decrease the can investigate to find the reason.

RECOMMENDATION:

- ✓ To improve the sales company can establish more shops and gain more dealers by scaling up the distribution to many places.
- ✓ As the sales are more during Christmas and Holidays offering discount can increase customers
- ✓ Company can take effort to sell the wine in airports as people travelling can get it easily while travelling.
- ✓ Reaching many restaurant and request them to suggest sparkling wine to their customers may help.
- ✓ A separate team can be formed to investigate the drop in sales in order to avoid further drop.

ROSE WINE:

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

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

Table 11 Head of the Data

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

Table 12 Tail of Data

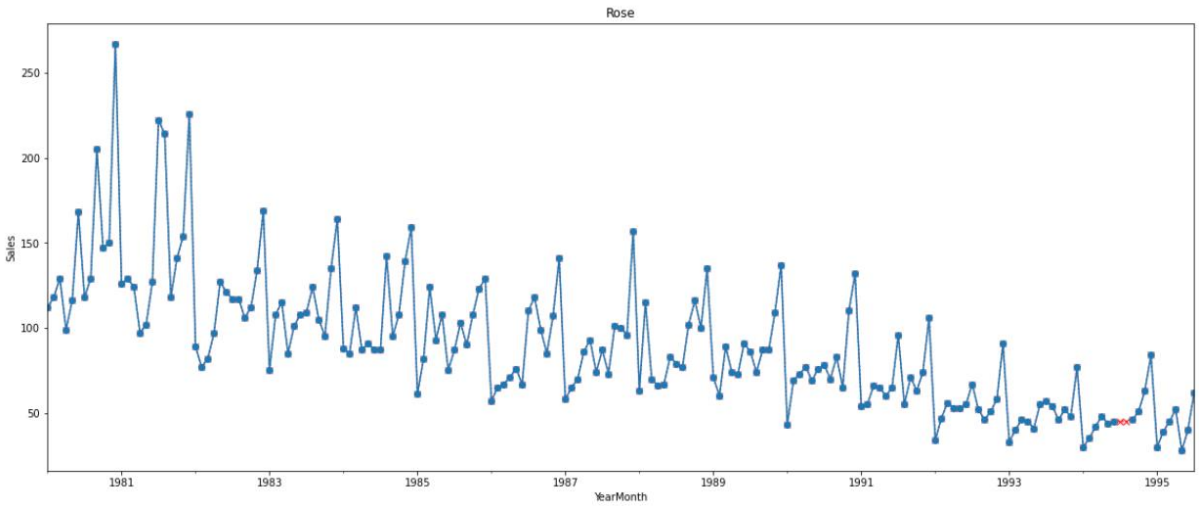

Figure 31 Overall plot of the data

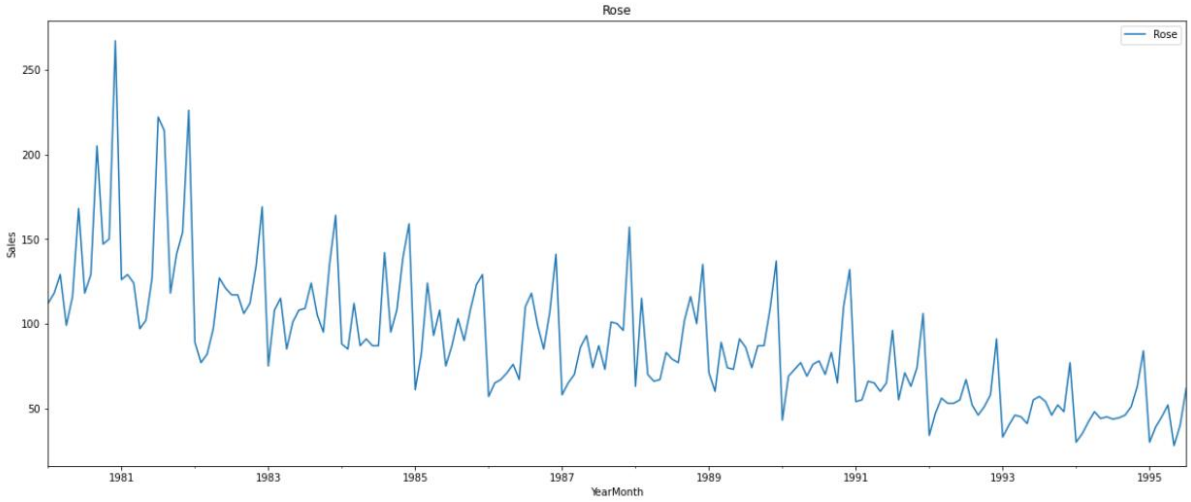

Figure 32 Series after imputation

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

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

Table 13 Summary of the series

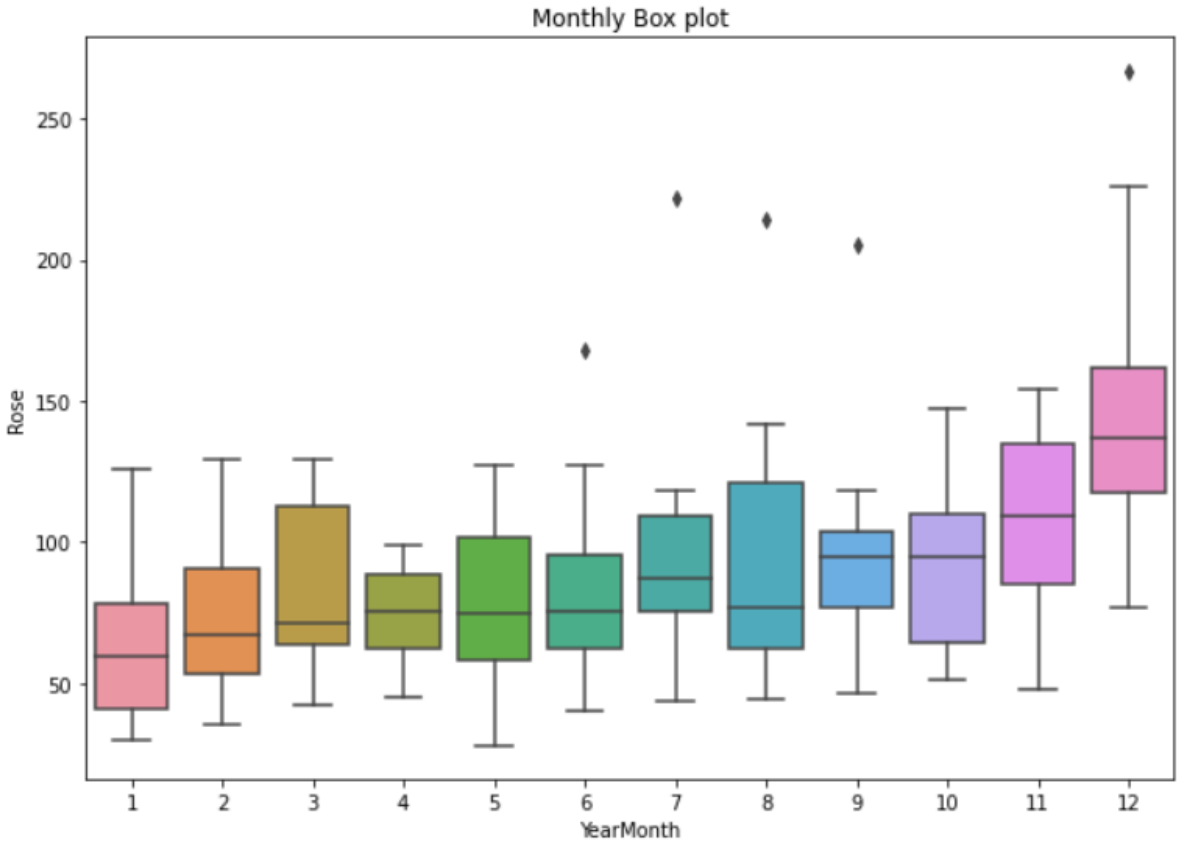

Figure 33 Boxplot throughout all months

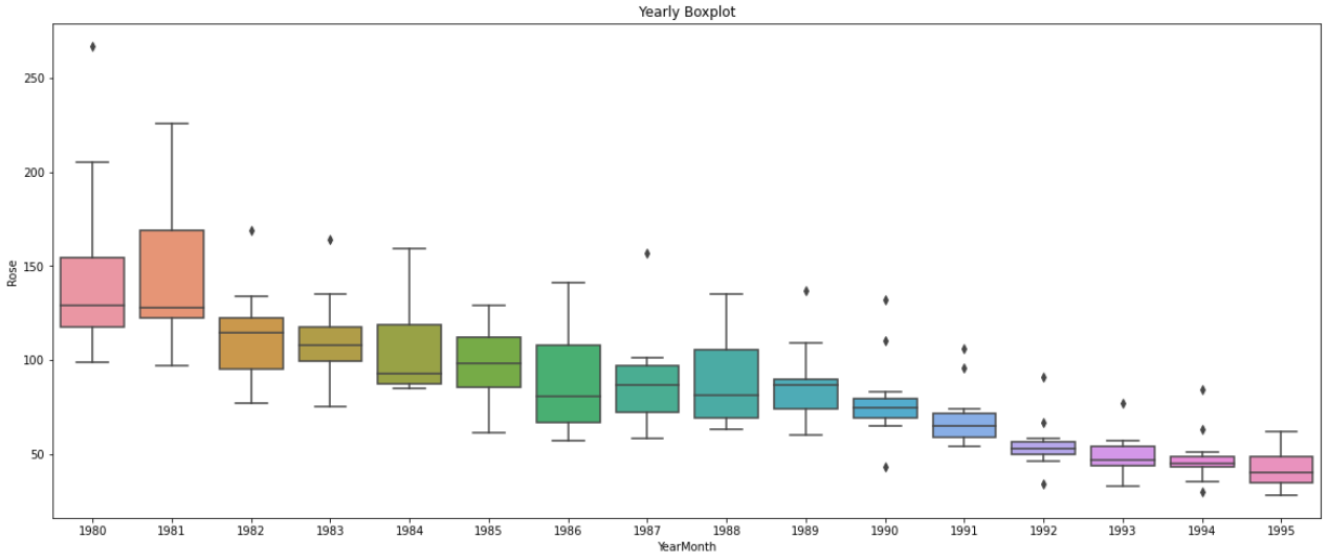

Figure 34 Boxplot throughout all years

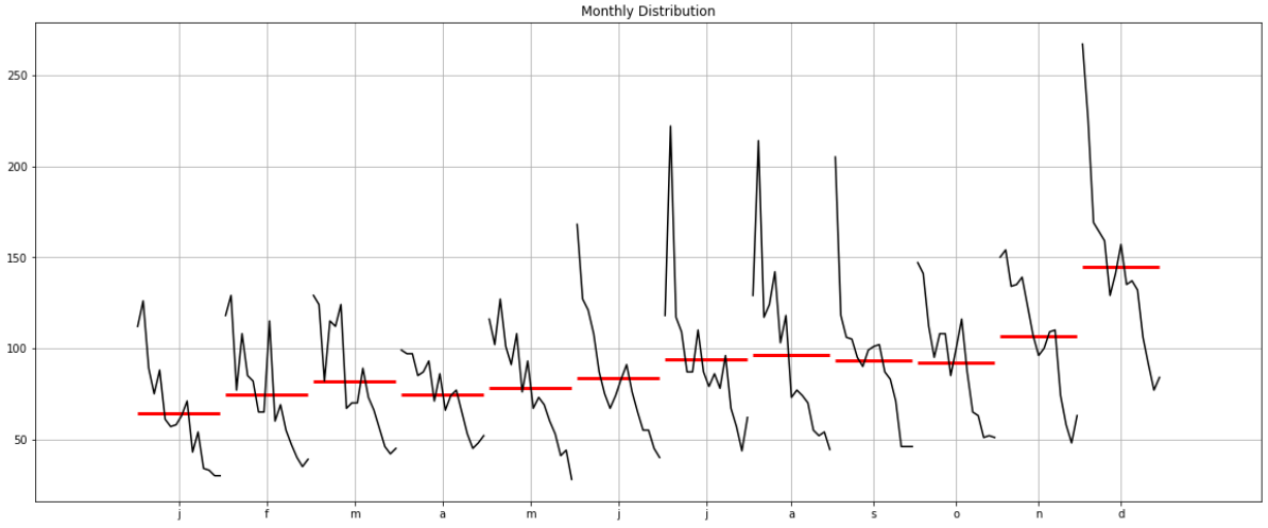

Figure 35 Monthly sales for all the years

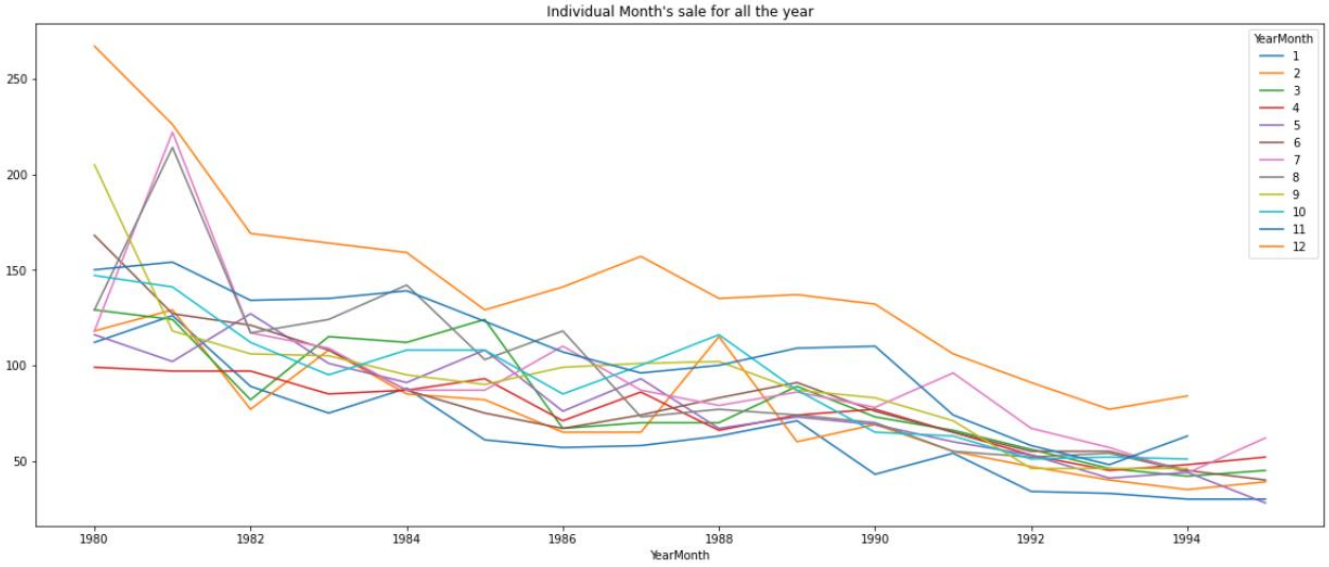

Figure 36 Monthly sales

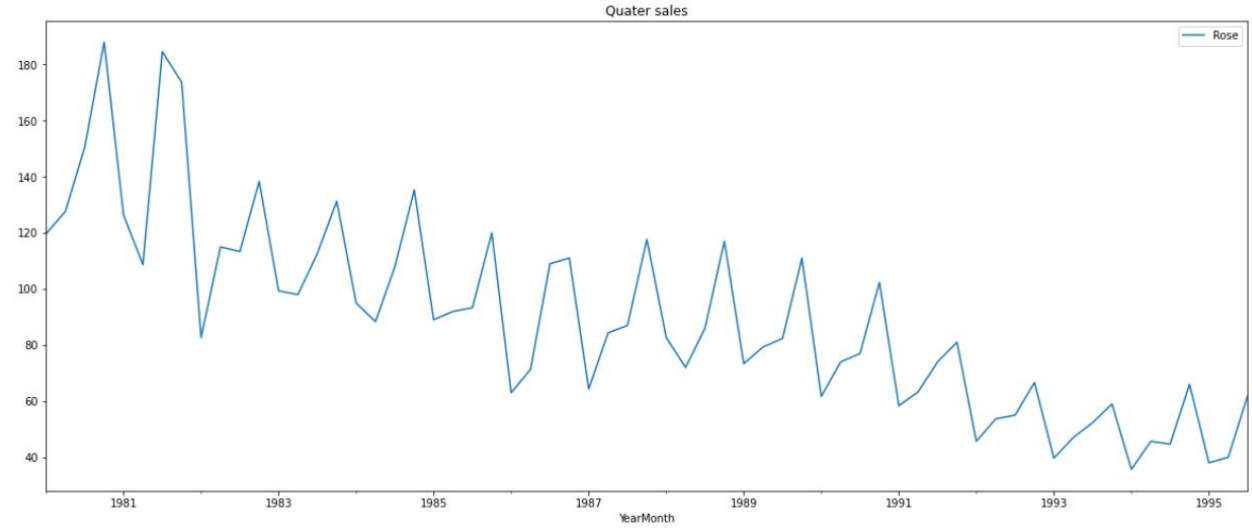

Figure 37 Quarterly Sales

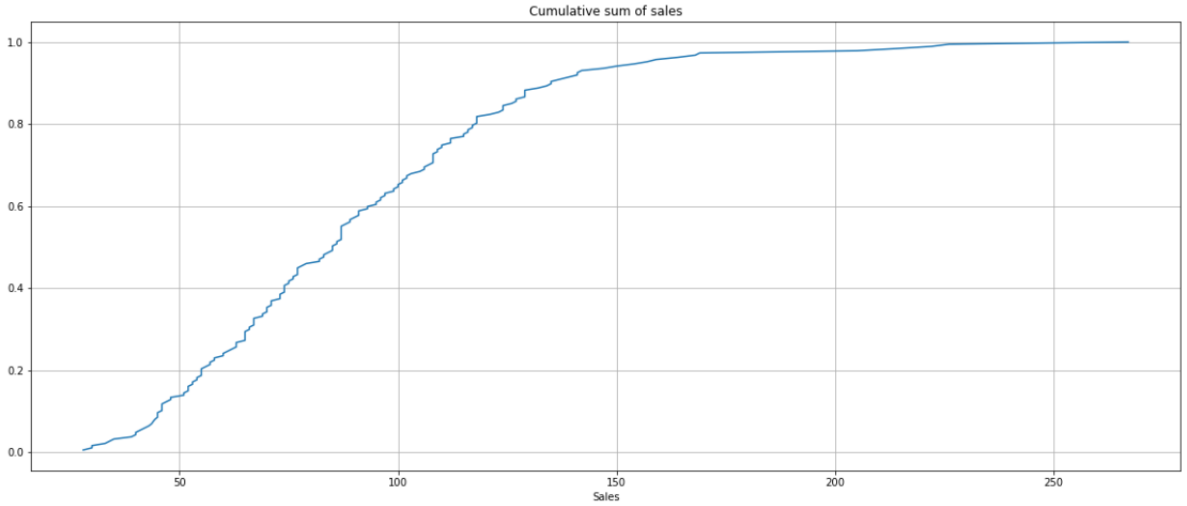

Figure 38 Cumulative Sum plot

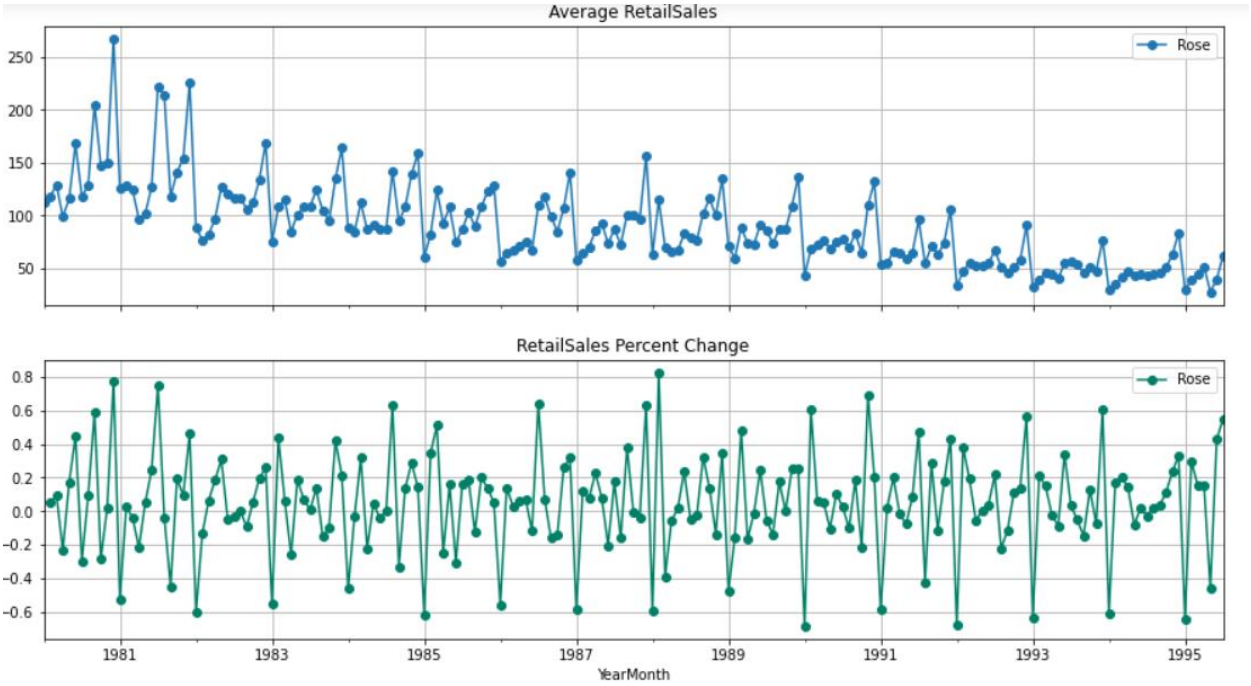

Figure 39 Plot of Avg. Sales and Percentage Change

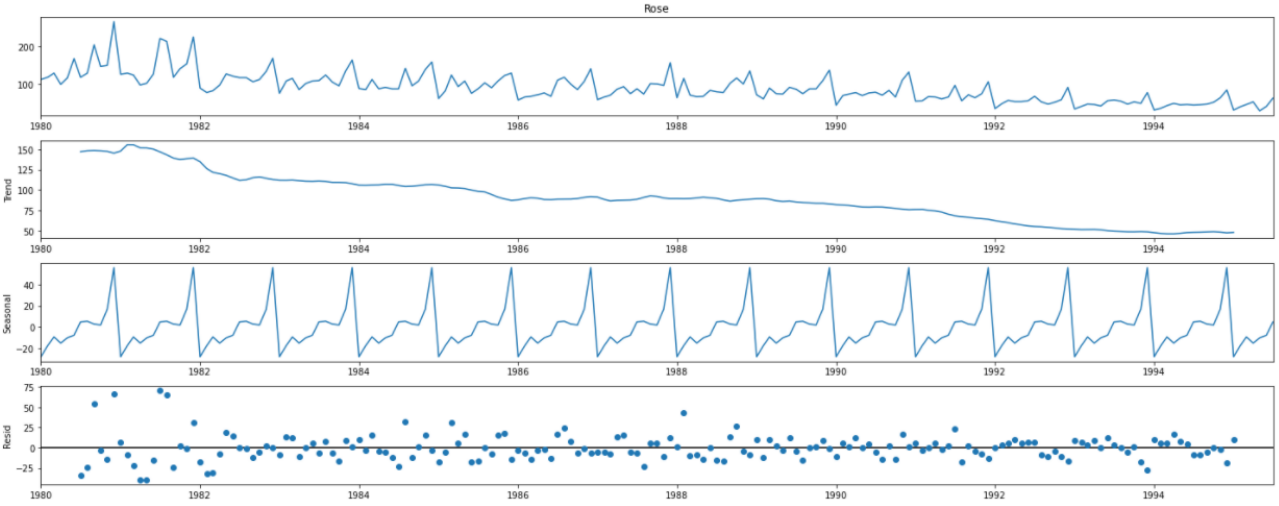

Figure 40 Seasonal Decomposition

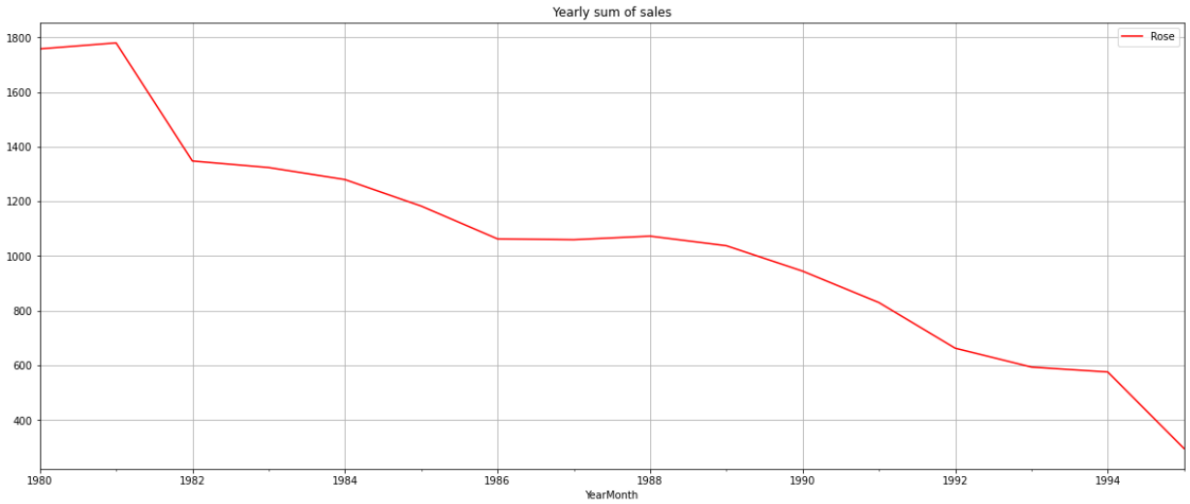

Figure 41 Annual sum of sales

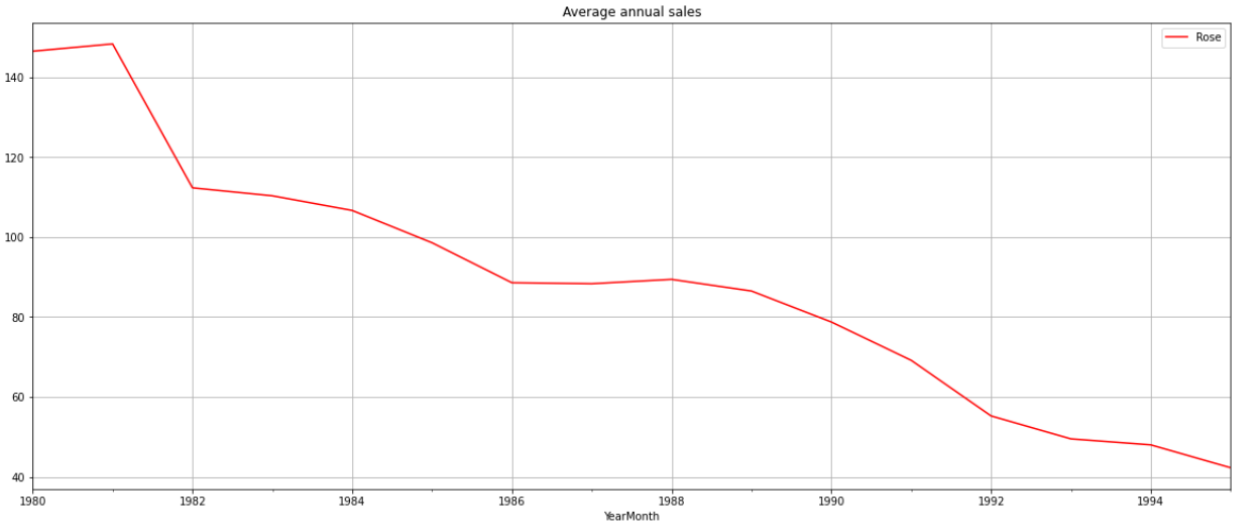

Figure 42 Annual average sales

Observations:

- ✓ There are 185 data points and 2 missing values in the dataset.
- ✓ The overall trend is decreasing, the sales keeps decreasing every year.
- ✓ The maximum sale is in 1981 and the least is in 1995
- ✓ Very few outliers present indicating there are sudden increase in the sales.
- ✓ Seasonality is found in the series small peak in sales is found in March and sales increases after August.
- ✓ While comparing the month plot we can see most of sales have happened in June, July, September and December.
- ✓ From the cumulative density plot we can see 38% of the data is above 100 and only 20% of the data is above 140.
- ✓ From the decomposition plot we can clearly see that the series contains trend and seasonality in them.
- ✓ The series is additive in nature.

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

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

Table 14 Train dataset

The train data consists of data from January of 1980 to December of 1990, it consists of 132 rows and 2 columns.

YearMonth	Rose
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

Table 15 Test dataset

The test data consists of data from January of 1991 to July of 1995, it consists of 55 rows and 2 columns.

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.

8. Linear Regression:

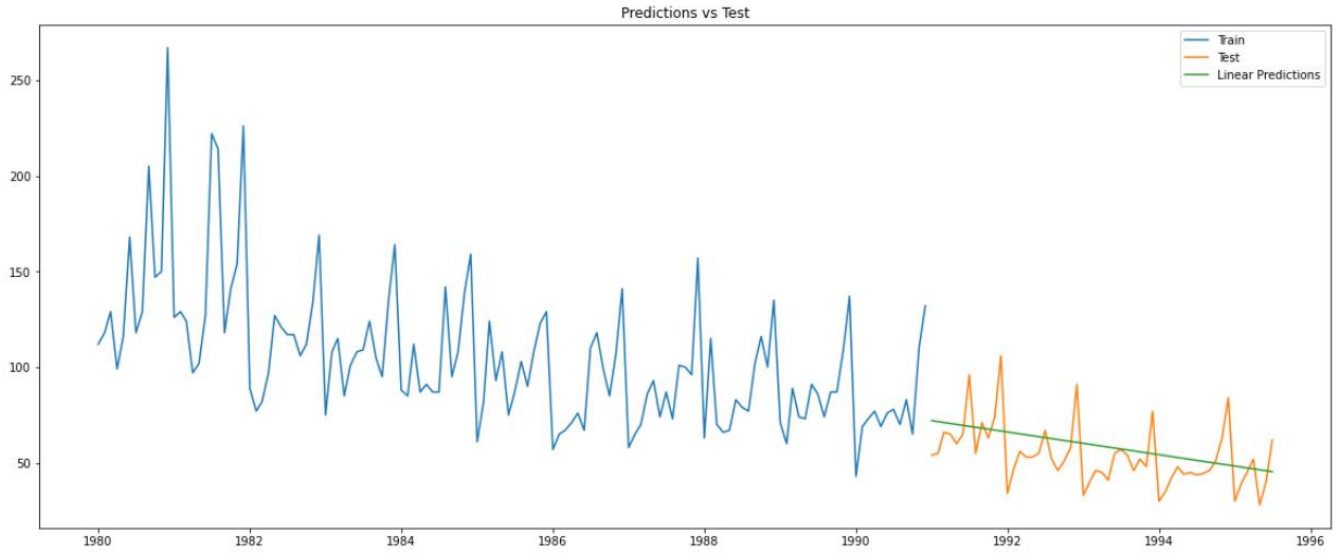

Figure 43 Performance of Linear Regression

We found the RMSE value between the test and the prediction to be RMSE: 15.135.The prediction shows that the model is not good enough it has more bias in it and hence high RMSE value.

9. Naïve Model:

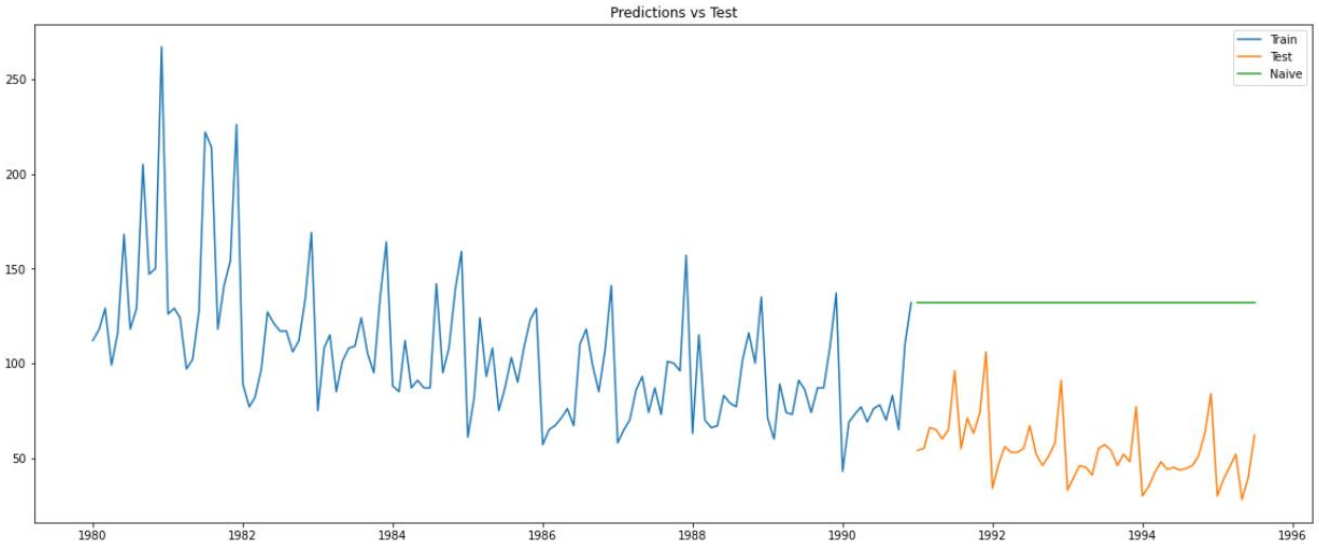

Figure 44Performance of Naive model

RMSE value between the test and prediction was found to be RMSE: 79.777

Naïve model will assume the last value of the train data for forecasting the future values, from the plot we can see it has predicted the peak value for all the test data points. Due to this reason the RMSE is high and the model is inefficient.

10. Simple Average:

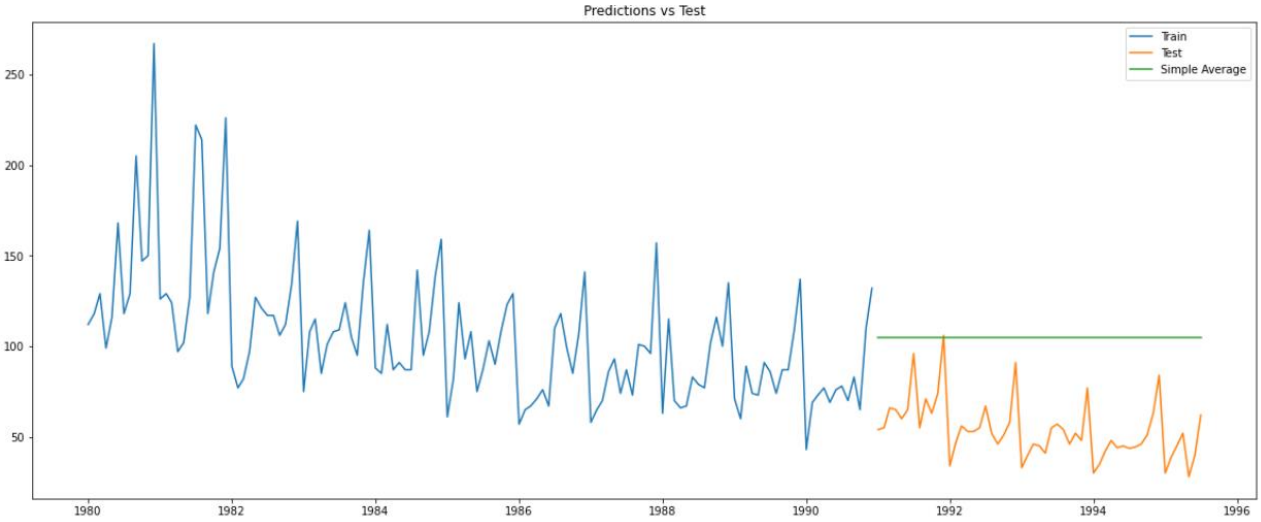

Figure 45 Performance of Simple average model

RMSE value between the test and prediction was found to be RMSE: 53.520

Simple average model will calculate the average of the train data and will use it to forecast. This might be better than naïve model but still the model is inefficient.

11. Moving Average:

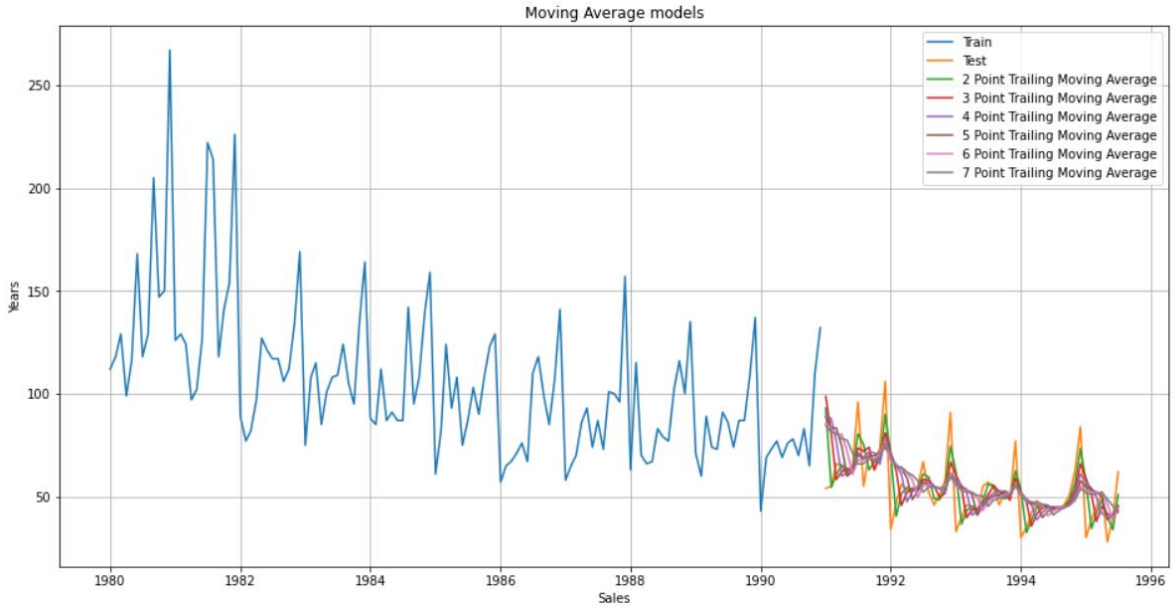

Figure 46 Different window sized moving average models

Moving average model tries to take the average within a particular window for each data point. If the window size is 2 it will take the mean of t-1 and t-2 data points. By trying out different window sizes we got the following results.

For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.530179857353627

For 3 point Moving Average Model forecast on the Training Data, RMSE is 14.129476303199834

For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.462022626370864

For 5 point Moving Average Model forecast on the Training Data, RMSE is 14.490389532601862

For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.58683041892348

For 7 point Moving Average Model forecast on the Training Data, RMSE is 15.077137800349332

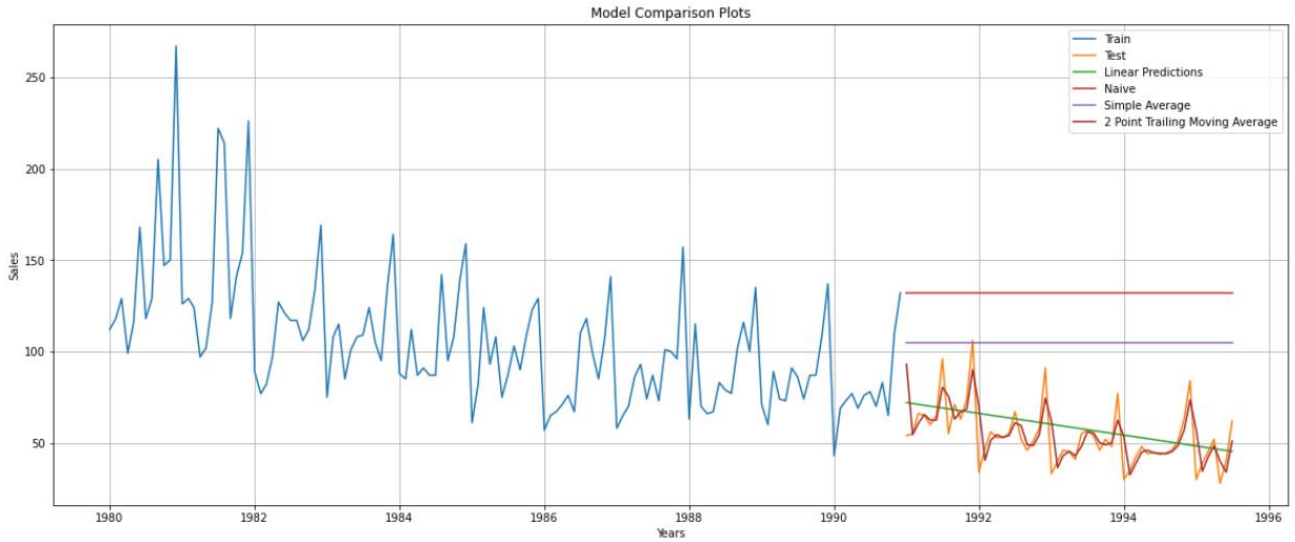

Figure 47 Model Comparison

The above plot contains the predictions of Linear, Naïve, Simple average and Moving average of window size 2. Within these models we can see moving average is the closest to the test data with the least RMSE value.

12. Simple Exponential Smoothing:

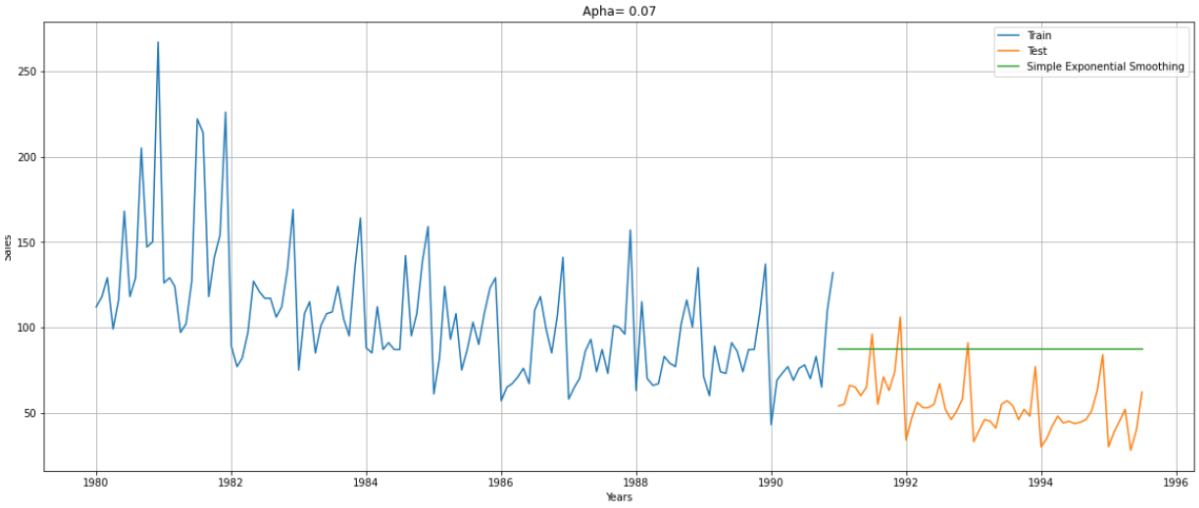

Figure 48 Simple exponential smoothing (alpha=0.07)

Smoothing technique means considering the past values and creating a relationship between the past values and the present. Exponential smoothing means giving more weightage to the recent data and least importance to old data. The importance factor diminishes exponentially as we move back in time.

Simple Exponential smoothing smoothens only the value of the time series, from the plot above we can see the prediction is higher than simple average. Although since the data has seasonality this is inefficient we found the RMSE value to be 36.858. The smoothing parameter for the model is denoted by alpha and found out to be 0.098.

13. Double Exponential Smoothing:

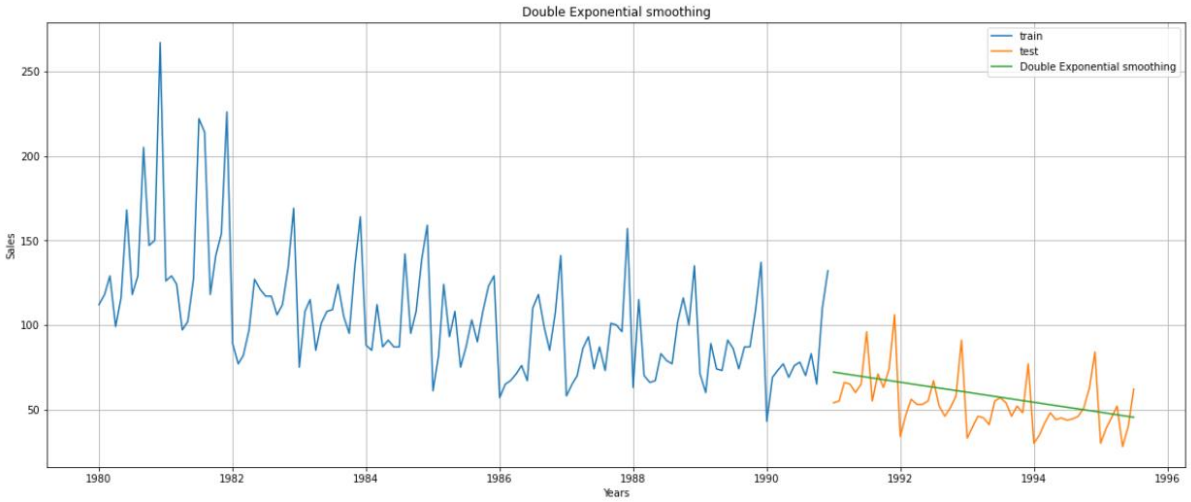

Figure 49 Double exponential smoothing

In Double exponential smoothing we'll consider both the value and the trend for smoothing and both will be used to forecast the future values. Similar to SES this model also find it difficult to predict due to the presence of seasonality. As this model doesn't involve seasonality in prediction.

The RMSE value was found to be 15.291 and the parameter for smoothing level is denoted by alpha which is found out to be 1.490e-08. Parameter for smoothing trend is denoted by Beta and found out to be 1.661e-10.

14. Holt winter's model:

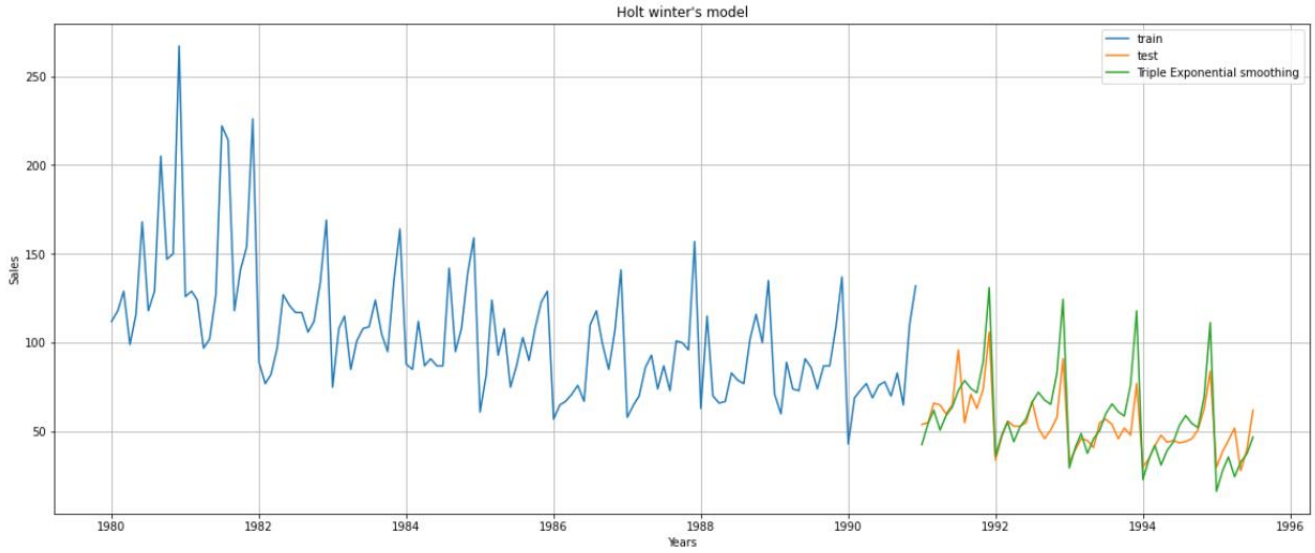

Figure 50 Holt Winter's Performance

Unlike the above two models Holt winter's model will smoothen all the 3 components that is value, trend and seasonality. So because of this reason this model will perform better than the other models. By evaluating the prediction with the test data we found the RMSE value to be 14.291 which is the least from all the above models. Similar to the previous model it denotes value and trend as alpha and beta and seasonality as gamma. The value for all 3 are as mentioned, alpha=0.089, beta=0.0002 and gamma=0.003.

Observation:

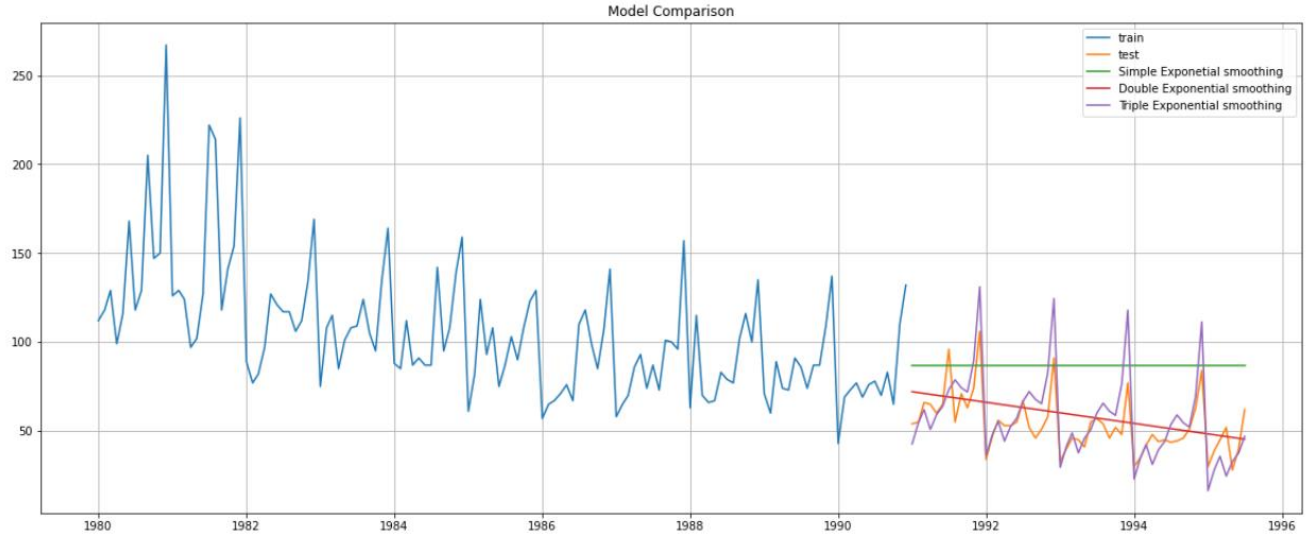

Figure 51 Exponential model comparisons

Within the exponential methods we can clearly see the Holt winter's method (Triple exponential smoothing) is the closest to the test data and with the least RMSE value. The prediction traces the test data very closely.

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.

A series is said to be stationary when it has all constant statistical properties such as variance and auto correlation over time. We can use a statistical test called Dickey Fuller test to check for stationarity. The hypothesis for the test are as follows:

H_0: The series is non – stationary

H_a: The series is stationary

After performing the test on the series we found the test statistic to be -2.240 and p value to be 0.467. As we can see the p value is greater than the alpha value 0.05 we can conclude we don't have enough evidence to reject the null hypothesis, therefore the series is not stationary.

In order to make the series stationary we can take the first difference on the series and check for stationarity. If it still doesn't become stationary we can keep performing the differencing or try any transformation on the series. In the following series we have performed the first order differencing and the series is shown below.

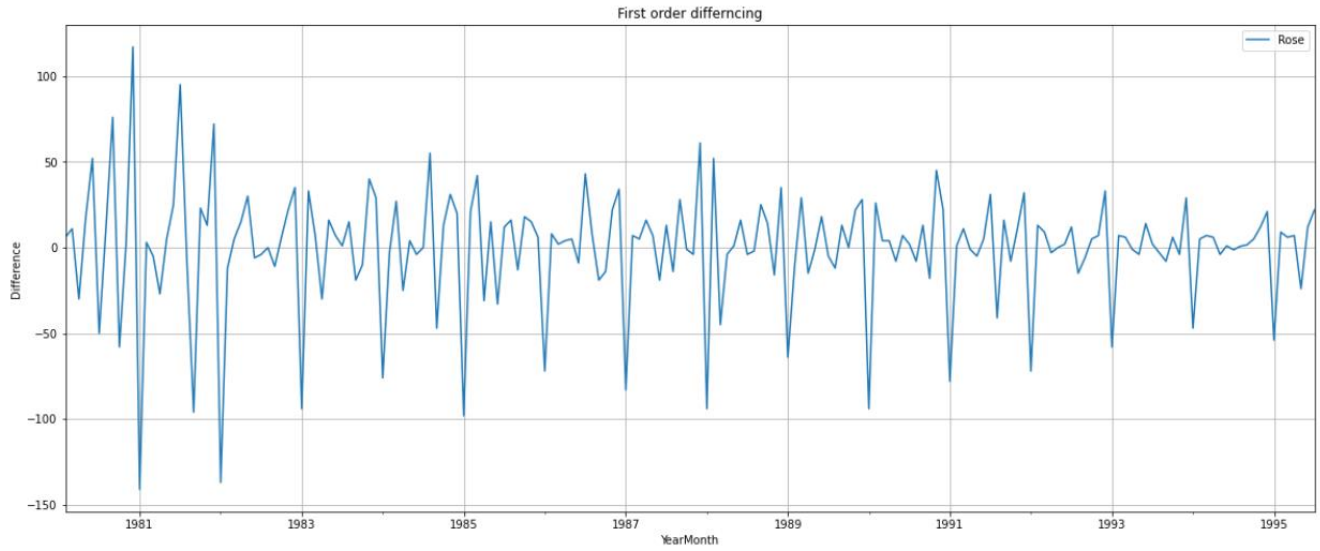

Figure 52 Differentiated Time Series

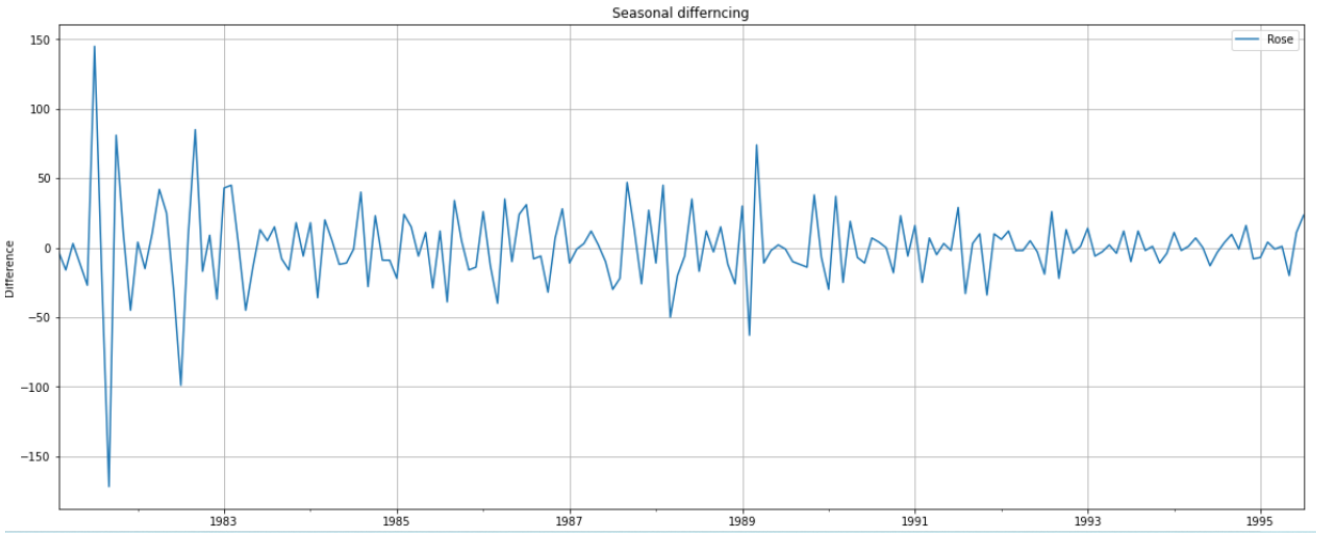

Figure 53 Series after 12 Differencing

After differentiating and performing the test we found the test statistic to be -8.160 and p value to be $3.048e^{-11}$. As the p value is very insignificant we can conclude we have enough evidence to reject the null hypothesis and hence we can conclude the series is stationary. Although the series has become stationary we can see the data has some trend in it so we must perform the seasonal differencing to remove the trend. The above results were for the entire dataset, the same test were done on the train dataset and found the p value to be $3.894e^{-8}$.Hence the training data has become 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.

From the acf plot of the series we can find the seasonal(S) value by looking at regular lags where we have a significant correlation.

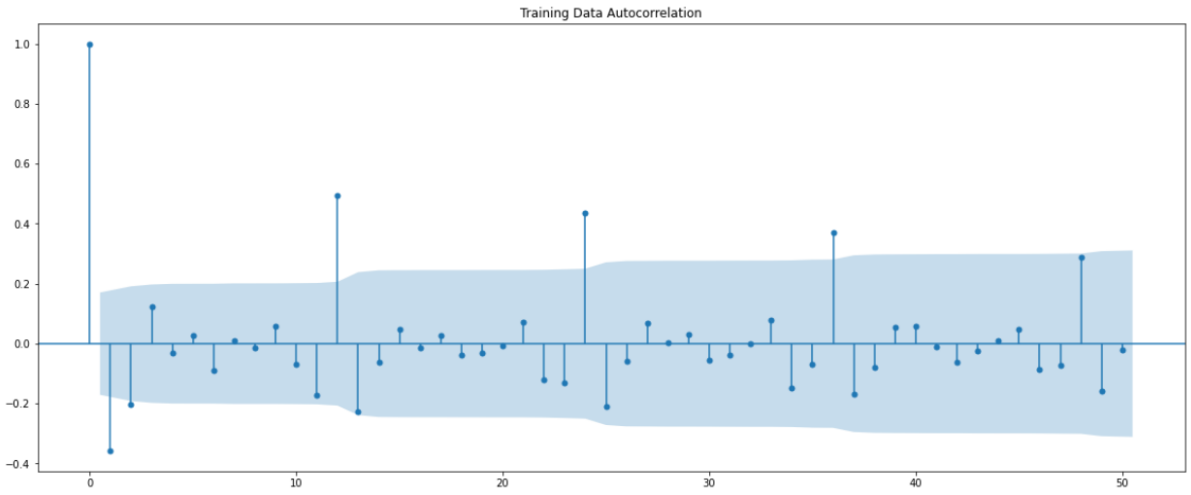

Figure 54 Acf plot to find the seasonal value

We can see there are significant spikes at lag 12 and its multiples so we can say the series has seasonality at every 12^{th} lag so S= 12.A brute force approach is used to find the best parameters for non-seasonal component and the seasonal component of the sarima model and the models with the best AIC values are shown below.

	param	seasonal	AIC
107	(0, 1, 2)	(2, 1, 2, 12)	774.969119
215	(1, 1, 2)	(2, 1, 2, 12)	776.940164
323	(2, 1, 2)	(2, 1, 2, 12)	776.996100
269	(2, 0, 2)	(2, 1, 2, 12)	780.716944
161	(1, 0, 2)	(2, 1, 2, 12)	780.992967

Table 16 Sarima models with AIC values

From the table above we can see the model (0,1,2)(2,1,2,12) has the least AIC value so we proceed with trying this model on train data and find its performance on the test data.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-380.485
Date:	Wed, 15 Dec 2021	AIC	774.969
Time:	14:13:47	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	-0.0763	0.126	-0.605	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	187.8596	45.271	4.150	0.000	99.131	276.588

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

Table 17 Summary of the auto Sarima model

From the summary table above looking at the p value we can see most of the value is greater than 0.05 which denotes most of the parameters cannot be very good predictors except for ma.L1(moving average lag 1) and ma.S.L12(moving average Seasonal component lag 12).Hence this model may not perform well in predicting the test data.

As discussed above after building the model when tried on the test data we found the RMSE (Root mean squared error) to be 16.556. which is greater than holt winter's model.

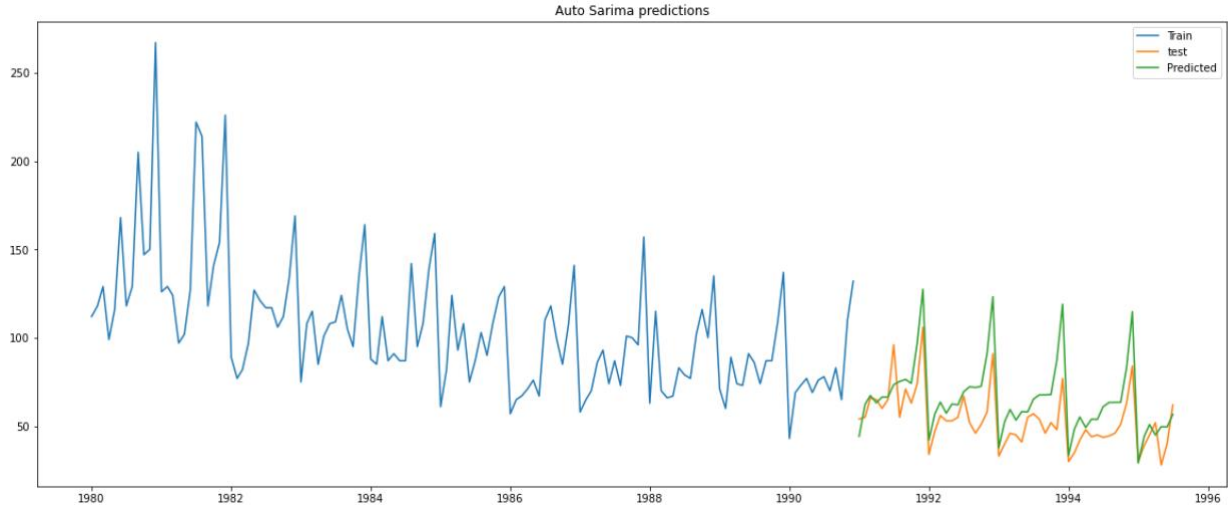

Figure 55 Performance of the auto sarima model

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.

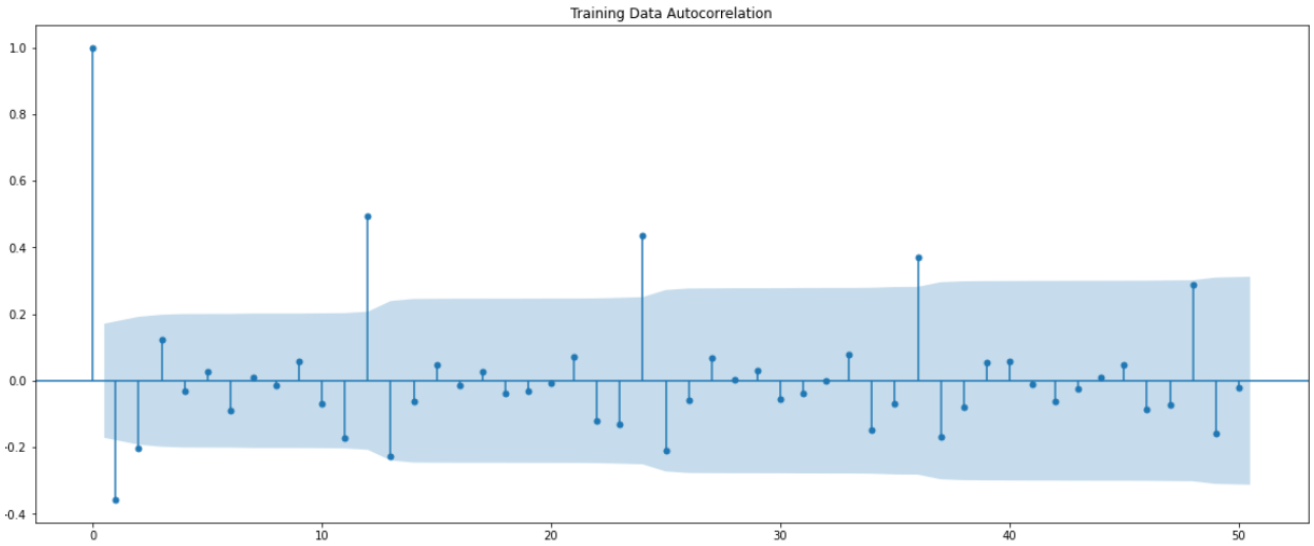

Figure 56 ACF plot on 1st Differentiated series

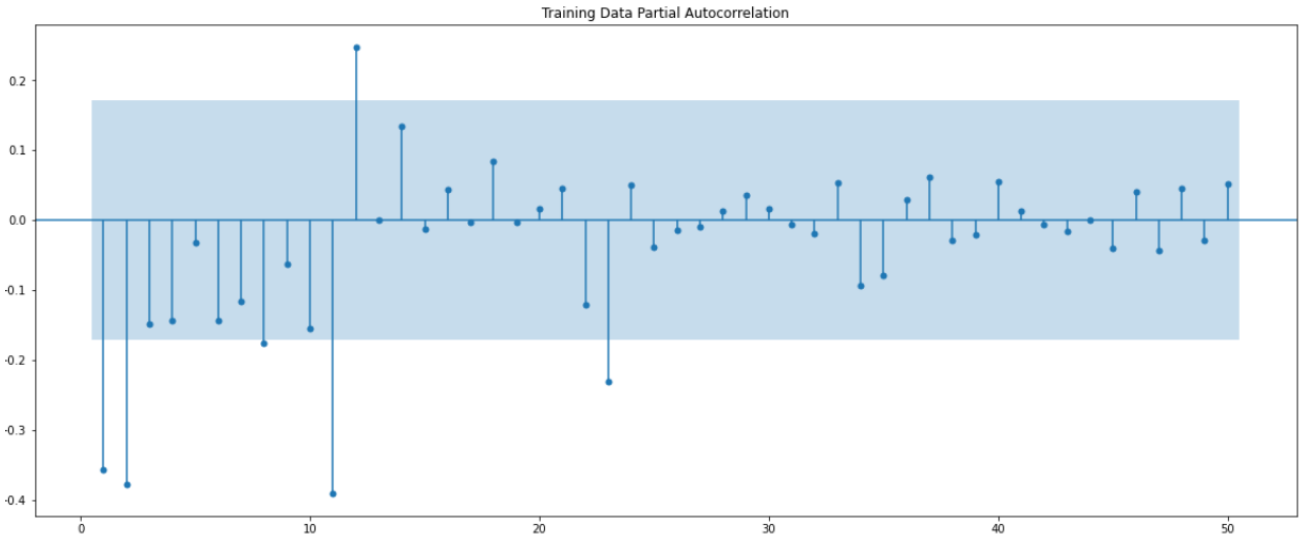

Figure 57 PACF plot on 1st Differentiated series

From acf plot we can find the value for moving average (q) and from pacf plot we can find the value for auto correlation(p) by finding the first lag where it cuts off to 0. From the acf plot we can see from 3rd lag it cuts off to 0 so the q value will be 2 and in the pacf plot the 3rd lag cuts off to 0 so the p value is also 2.

We use to first order differencing to make the series stationary so d=1.

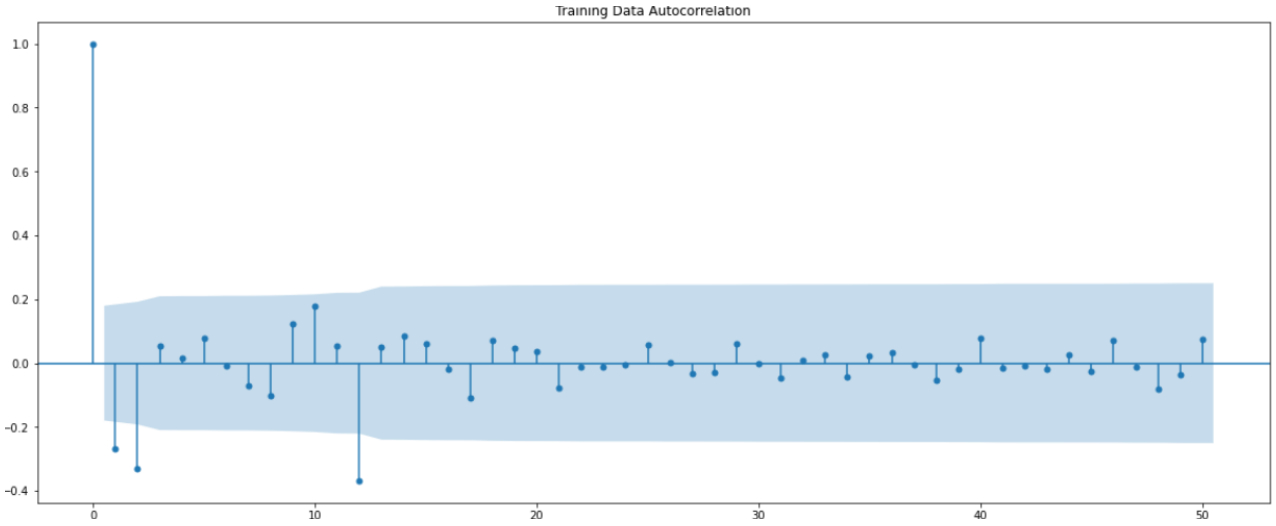

Figure 58 Acf plot seasonal differentiated data

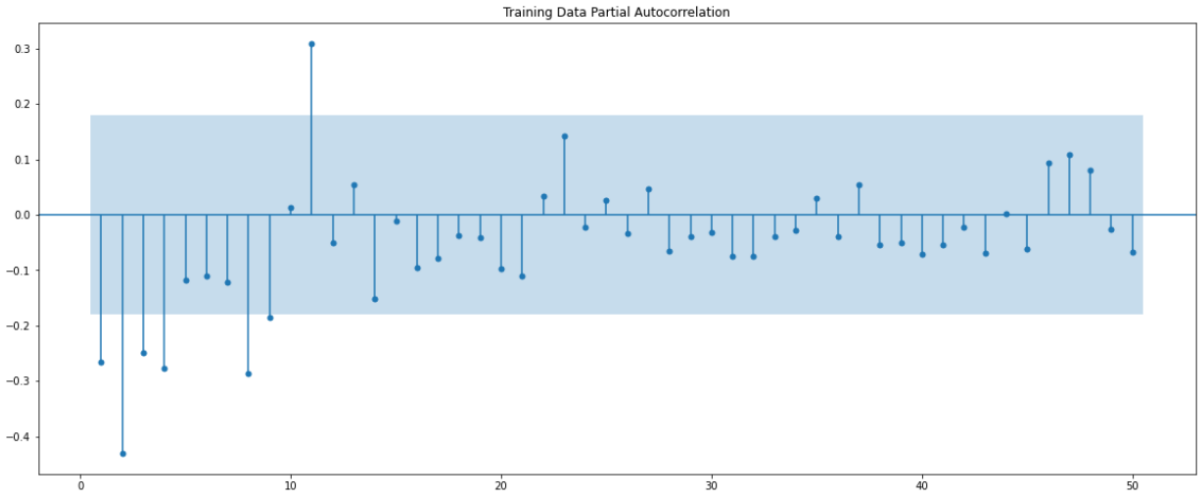

Figure 59 Pacf plot seasonal differentiated data

From the fig29 we can find the value of moving average of seasonal component (Q) by using the same method as before. (Lag at which first cuts off to 0) from the pacf plot we can find the auto correlation parameter (P) for seasonal component. From the plots we can see Q to be 2 and P to be 4. As we have done seasonal differencing (12) D is 1 and Seasonal parameter(S) is 12.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(4, 1, 2, 12)	Log Likelihood	-284.473
Date:	Sun, 19 Dec 2021	AIC	590.945
Time:	14:30:15	BIC	615.520
Sample:	01-01-1980 - 12-01-1990	HQIC	600.695
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9797	0.225	-4.348	0.000	-1.421	-0.538
ar.L2	-0.1283	0.143	-0.897	0.370	-0.409	0.152
ma.L1	0.0197	0.247	0.080	0.936	-0.464	0.503
ma.L2	-0.8815	0.194	-4.553	0.000	-1.261	-0.502
ar.S.L12	-0.7331	0.198	-3.704	0.000	-1.121	-0.345
ar.S.L24	-0.0714	0.173	-0.414	0.679	-0.410	0.267
ar.S.L36	0.0761	0.088	0.865	0.387	-0.096	0.249
ar.S.L48	-0.0064	0.021	-0.306	0.759	-0.047	0.035
ma.S.L12	-0.3449	0.714	-0.483	0.629	-1.744	1.054
ma.S.L24	-0.8949	0.567	-1.579	0.114	-2.006	0.216
sigma2	146.4883	113.972	1.285	0.199	-76.893	369.869

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	6.02
Prob(Q):	0.91	Prob(JB):	0.05
Heteroskedasticity (H):	0.61	Skew:	0.53
Prob(H) (two-sided):	0.25	Kurtosis:	3.98

Table 18 Manual Sarima summary

From this summary we see there are only 3 significant predictors and others are insignificant so this model might not be a good predictor.

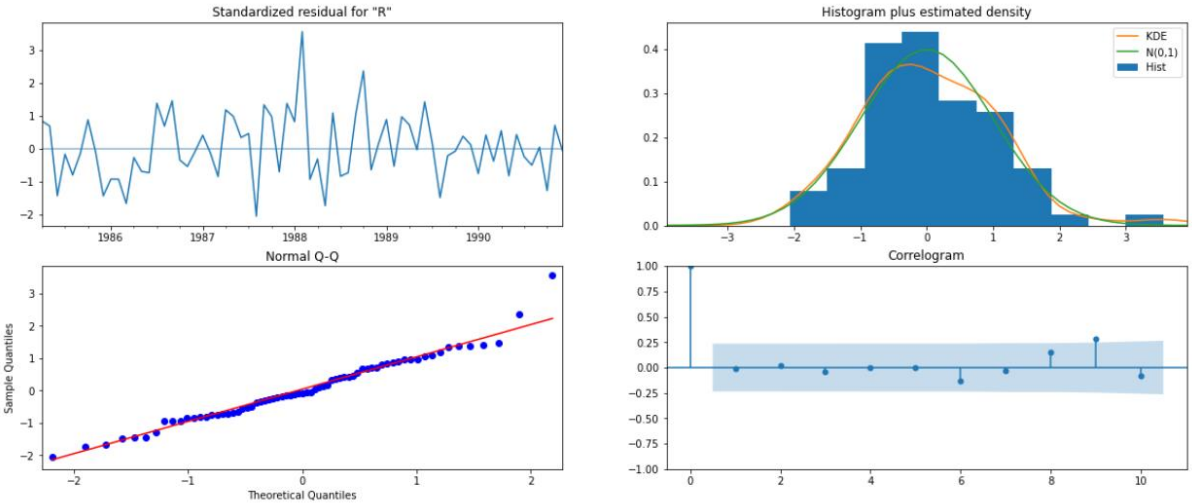

Figure 60 Manual sarima model diagnostics

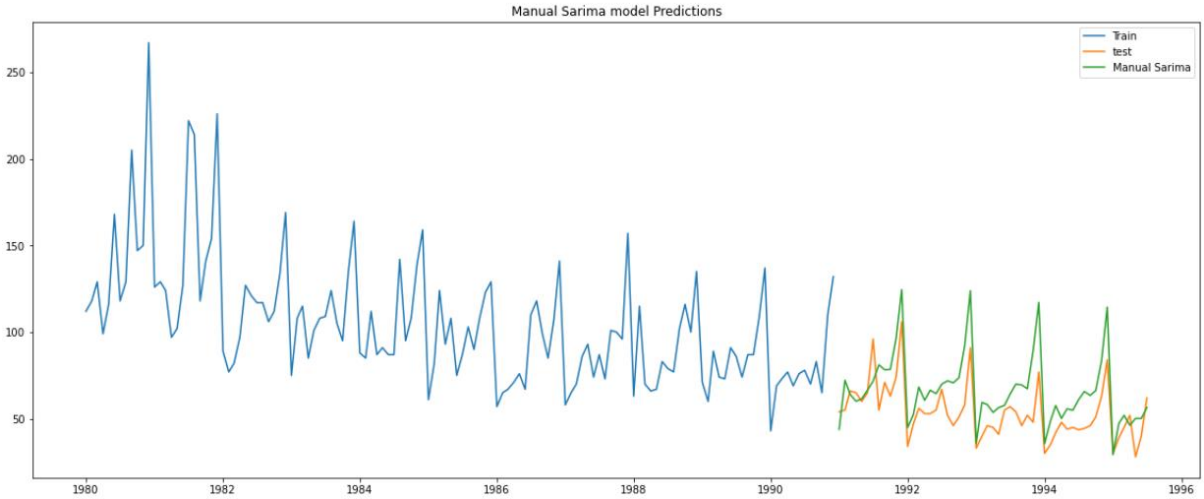

Figure 61 Manual Sarima performance on test data

From fig 26 we can see the standard error doesn't follow normal distribution and AIC is found to be 590.945 and RMSE to be 17.398.

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.

	RMSE
Linear_Regression	15.291
Naive	79.777
Simple_Avg	53.521
Trailing_2	11.530
Trailing_3	14.129
Trailing_4	14.462
Trailing_5	14.490
Trailing_6	14.587
Trailing_7	15.077
Simple_Exponential_Smoothing	36.858
Double_Exponential_Smoothing	15.291
Holt_winter	14.291
Auto_SARIMA(0, 1, 2)x(2, 1, 2, 12)	16.556
SARIMAX(2, 1, 2)x(4, 1, 2, 12)	17.399

Table 19 Models Summary

From the table 8 we can see Moving average model with window size of 2 has the least RMSE value 11.530 we cannot choose this model because we cannot forecast into the future so we are going ahead with second best model that is Holt winter's model with RMSE value of 14.291.

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.

Holt winter model is trained on the entire dataset and found the following parameters.

Smoothing level: 0.1167,

Smoothing trend: 0.0077,

Smoothing seasonal: 0.0023

RMSE on the fitted values of the model is 17.755 which is better compared to the other models.

	Forecast	ci_lower	ci_upper
1995-08-01	50.790631	15.902663	85.678599
1995-09-01	47.721093	12.833125	82.609061
1995-10-01	46.477600	11.589632	81.365568
1995-11-01	61.126872	26.238904	96.014839
1995-12-01	99.395342	64.507374	134.283310
1996-01-01	15.107371	-19.780597	49.995339
1996-02-01	25.350487	-9.537481	60.238455
1996-03-01	32.887418	-2.000550	67.775386
1996-04-01	25.690153	-9.197815	60.578121
1996-05-01	29.089136	-5.798832	63.977104
1996-06-01	34.563125	-0.324843	69.451093
1996-07-01	45.170766	10.282798	80.058734

Table 20 Forecast of Holt winter model

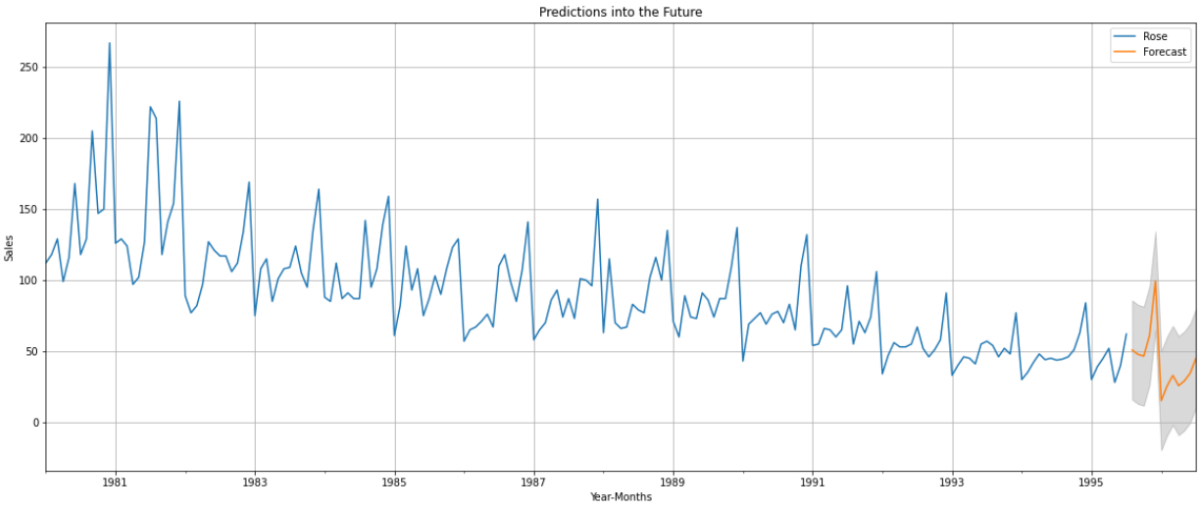

Figure 62 Forecast of Holt winter model

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

SUMMARY:

The data was read by parsing the date as index column, after that once we tried to understand the data we found that there were 2 missing values in the data and the entire series has an average of 90.394 and standard deviation of 39.175. The standard deviation is high indicating there might be high variations in the series. After that the data was explored and found the following insights from it.

- ✓ There are 185 data points and 2 missing values in the dataset.
- ✓ The overall trend is decreasing, the sales keeps decreasing every year.
- ✓ The maximum sale is in 1981 and the least is in 1995
- ✓ Very few outliers present indicating there are sudden increase in the sales.
- ✓ Seasonality is found in the series small peak in sales is found in March and sales increases after August.
- ✓ While comparing the month plot we can see most of sales have happened in June, July, September and December.
- ✓ From the cumulative density plot we can see 38% of the data is above 100 and only 20% of the data is above 140.
- ✓ From the decomposition plot we can clearly see that the series contains trend and seasonality in them.
- ✓ The series is additive in nature.

After finding the insights the missing values were filled by using the method interpolate then the data was divided into train and test. Data below 1991 was considered as Training data and above 1991 was considered as test data.

Various models like Linear regression, Naïve, Simple average ,moving average ,Exponential smoothing methods and SARIMA models were trained using the train data and tested with the test data. Each models' performance was measured using RMSE as the metric. In the above methods

Regression needs an input variable to train on so the number of months were calculated for all the data points and were passed as input variable for the model. For moving average different window size were tried and found out 2 is the optimum size.

While building the SARIMA model we must make sure the series is stationary. Stationary means all the statistical property like variance and correlation must not be dependent on time. We found out by default the series was not stationary so the first differencing was done and checked for stationarity. To check for stationary we use Dickey fuller test as the statistical test with null hypothesis – The series is not stationary. Even after taking the first order differencing the series showed trend in it so seasonal differencing was also performed before building the model.

PACF and ACF plots were plot and the values for the non-seasonal component were found out using the first order differentiated data. The values p and q were found out by finding the first cut from both the plots. For seasonal component same method were used to find P and Q from the seasonal differentiated series and the seasonal value S is found out by looking at the ACF plot by finding the regular intervals where we get a significant spikes.

A version of automated SARIMA was also built by using AIC(Alkaike Information Criteria). A brute force approach was done trying out different values for the model and found the model with the least AIC value. The least AIC means the better the model is. After building these to SARIMA models the RMSE value was calculated on the test data.

A table with all the models and its RMSE on the test data was created. From the table Holt winter model showed the least RMSE on the test data and it was trained on the entire data. Once trained on the entire data we found the smoothing parameters to be Smoothing level: 0.1167, Smoothing trend: 0.007, Smoothing seasonal: 0.0023 and the RMSE on the fitted values to be 17.755.

The final model has low RMSE on the train data indicating it has trained properly, it also has very good values for smoothing parameter indicating this will be a good predictor. This model is then used to forecast 12 months into the future. Holt winter method in python doesn't not calculate confidence interval for the predictions so it was manually done using the below formula:

$CI = Forecast \mp 1.96(std.error)$ @95% confidence interval

After finding the lower and upper confidence interval the forecasted values were plotted along with the confidence interval values next to the original data.

BUSINESS INSIGHTS:

- ✓ The sale of the Rose wine keeps decreasing over the years.
- ✓ The sales spike at July, August and December it might be due to Christmas and Holidays.
- ✓ The sales was more at 1980 than present it can be due to other competitors in the market.
- ✓ At present the overall sale has reached the least value.

RECOMMENDATION:

- ✓ Considering the company's benefit discontinuing the Wine will be advisable as the sale reached its lowest value.
- ✓ But in case of getting out from the situation analysing the market and competitors will give us the reason why the product has failed and we can make those changes in the product or release the same product with those changes under a new name.
- ✓ People are preferring during festival season so we must make sure some of the properties must be preserved and we must try to add on latest flavours and smell to it.