TIME SERIES FORECASTING PROJECT

JAYAPREETHI R M

INDEX

INDEX FOR TABLES

INDEX FOR PICTURES

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.

Please perform the following questions on each of these two data sets separately.

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

Following is the head and tail data of the Rose csv data. The data was loaded into the codebook as a time series data. The given dataset is a time series data that starts from the beginning 1980s and ends with July of 1995.

ROSE - TIME SERIES DATA - HEAD		
	YearMonth	**Rose**
0	1980-01-01	112
1	1980-02-01	118
2	1980-03-01	129
3	1980-04-01	99
4	1980-05-01	116

ROSE - TIME SERIES DATA - TAIL		
	YearMonth	**Rose**
182	1995-03-01	45
183	1995-04-01	52
184	1995-05-01	28
185	1995-06-01	40
186	1995-07-01	62

The following is the head and tail of the Sparkling csv table. This dataset was loaded as a time series dataset. The given dataset is a time series data that starts from the beginning 1980s and ends with July of 1995.

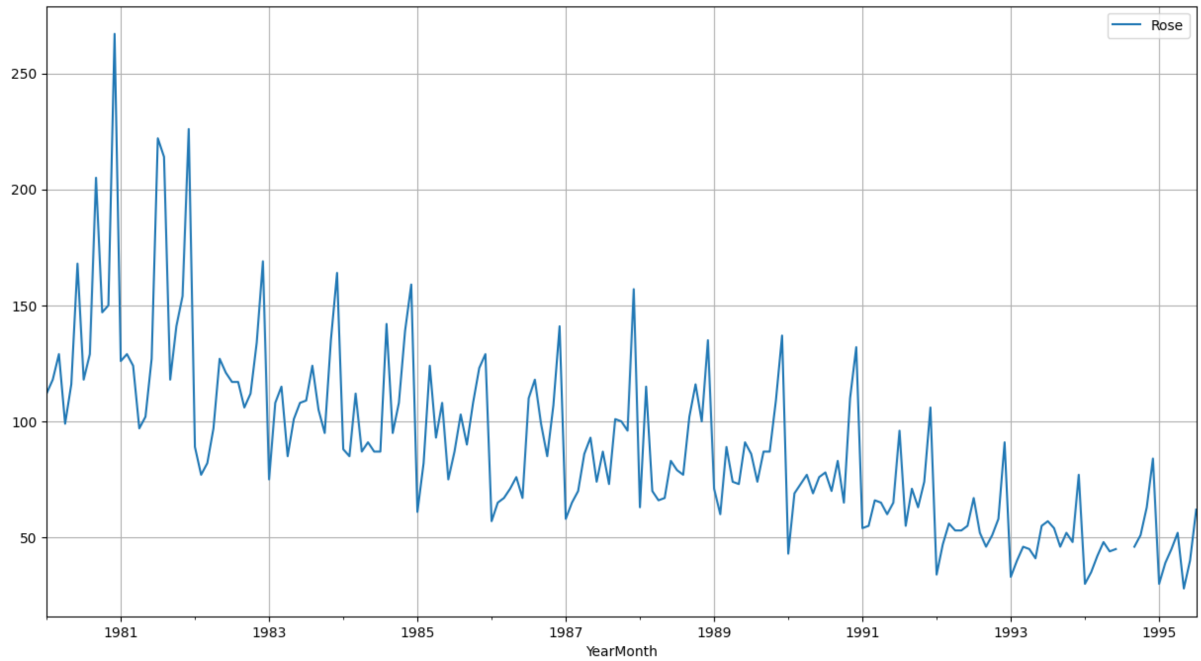
SPARKLING - TIME SERIES DATA - HEAD

	YearMonth	Sparkling
0	1980-01-01	1686
1	1980-02-01	1591
2	1980-03-01	2304
3	1980-04-01	1712
4	1980-05-01	1471

SPARKLING - TIME SERIES DATA - TAIL

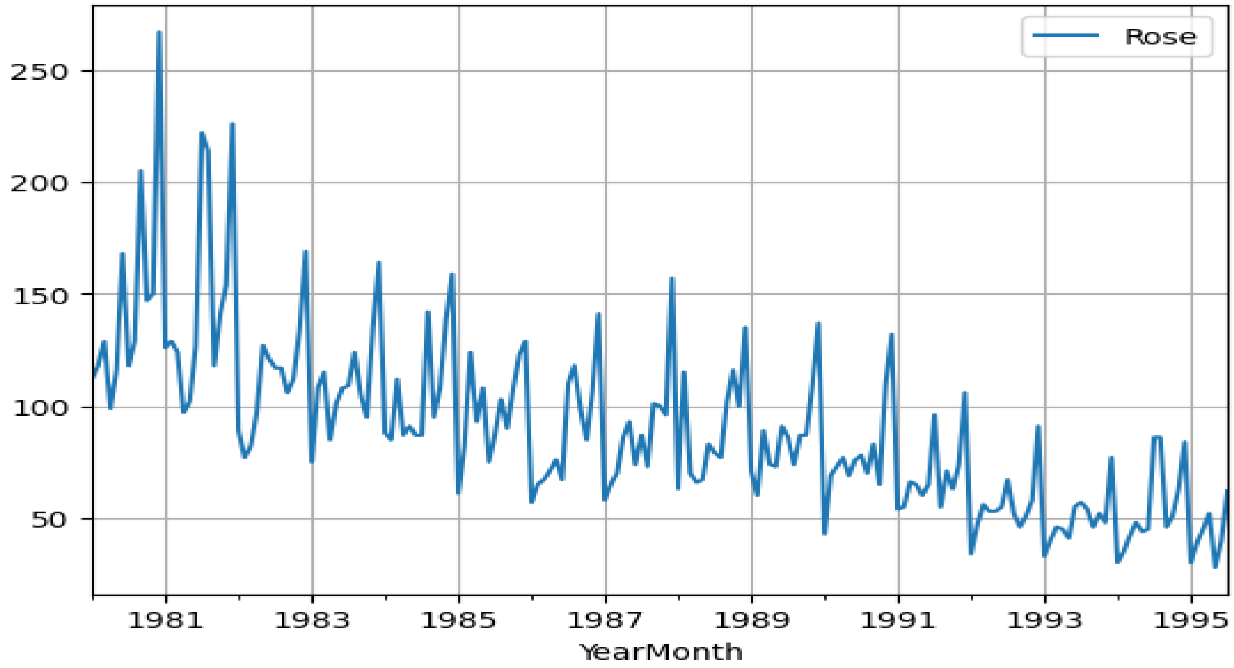
	YearMonth	Sparkling
182	1995-03-01	1897
183	1995-04-01	1862
184	1995-05-01	1670
185	1995-06-01	1688
186	1995-07-01	2031

ROSE - TIME SERIES DATA - GRAPH



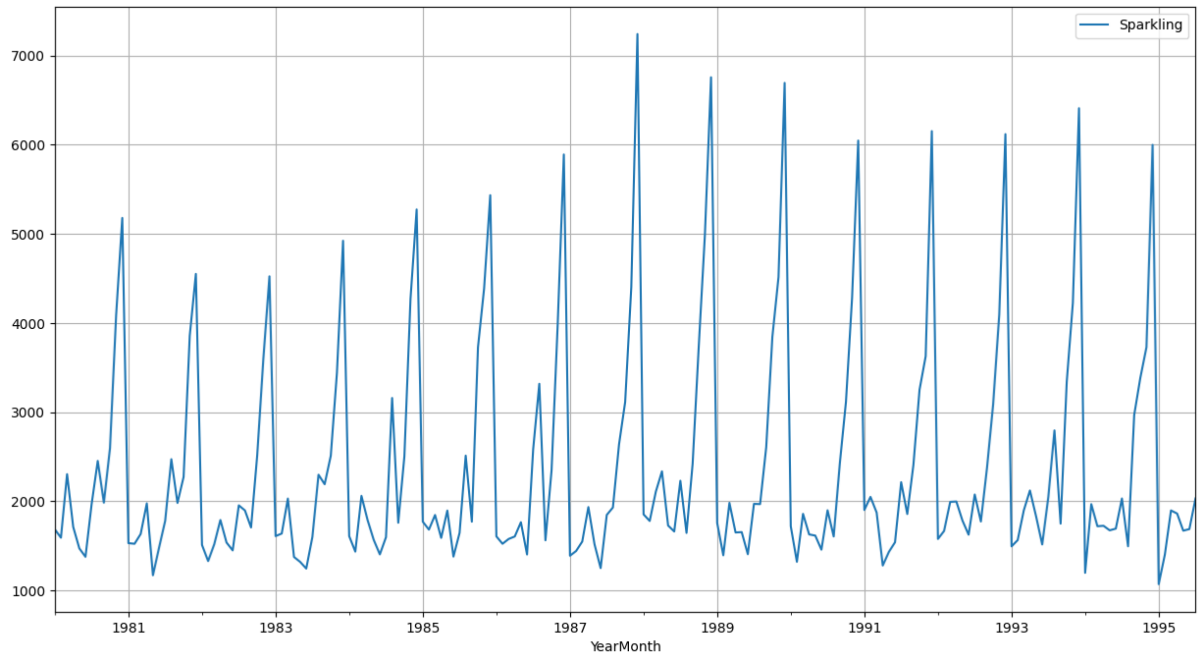
The line graph displays the Rose variable through time. There is a declining trend as time passes by. We can also notice that there are some null variables which must be treated. Given below is the graphical representation of corrected data.

ROSE - TIME SERIES DATA - CORRECTED GRAPH



On the other hand, we have the graphical representation of the Sparkling variable over the years. We can notice that the variables have oscillated between the same range with similar seasonality for years.

SPARKLING - TIME SERIES DATA - GRAPH



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

ROSE-DTYPE

YearMonth	datetime64[ns]
Rose	float64

Since, The rose dataset is a time series dataset, the YearMonth variable is the datetime variable and the Rose variable is of float data type.

SPARKLING - DTYPE

YearMonth	datetime64[ns]
Sparkling	int64

Since, The sparkling dataset is a time series dataset, the YearMonth variable is the datetime variable and the Rose variable is of integer datatype.

ROSE - SUMMARY

	Rose
count	187
mean	90.348
std	38.967
min	28
25%	63
50%	86
75%	111
max	267

After correcting the null variables there are 187 entries in total, the mean value of the entire dataset is 90.35. The variable oscillated between minimum of 28 to maximum of 267, while having a median value of 86.

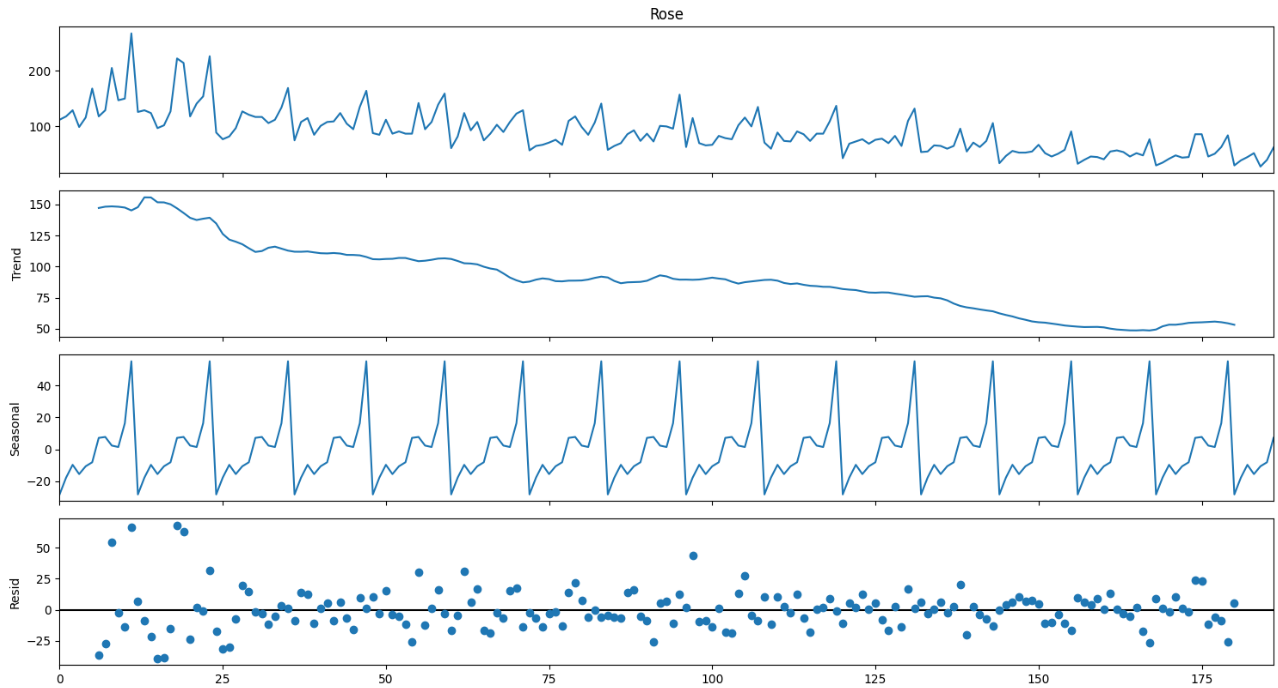
SPARKLING - SUMMARY

Sparkling	
count	187
mean	2402.417
std	1295.112
min	1070

25%	1605
50%	1874
75%	2549
max	7242

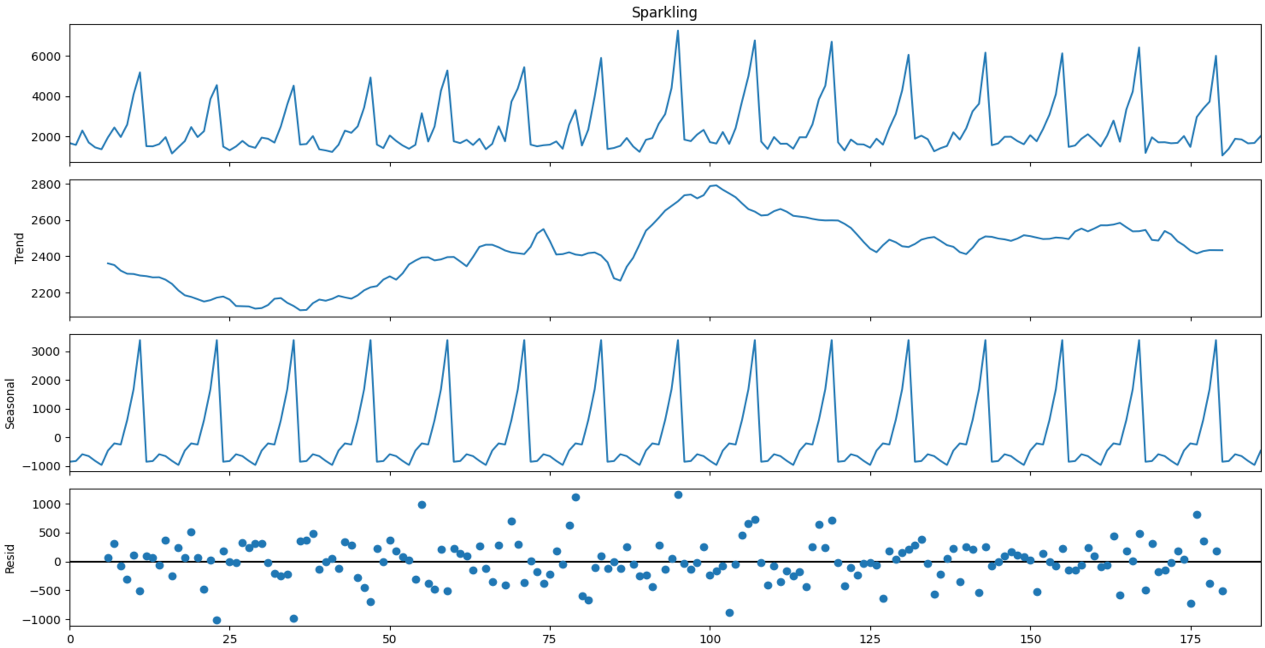
After correcting the null variables there are 187 entries in total, the mean value of the entire dataset is 2402.42. The variable oscillated between minimum of 1070 to maximum of 7242, while having a median value of 1874.

ROSE- ADDITIVE DECOMPOSITION



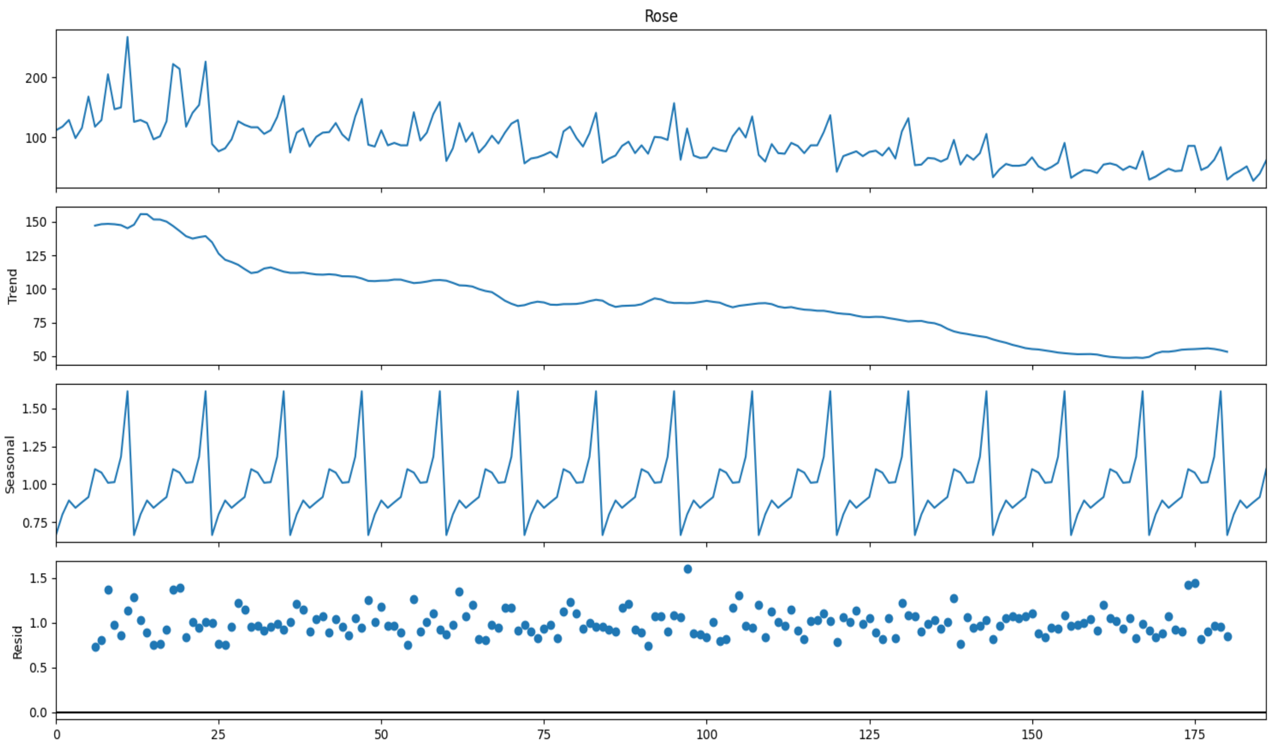
Additive decomposition posits that time series data can be expressed as the aggregate of its constituent elements. In this context, Y represents the time series data, T signifies the trend-cycle component, S denotes the seasonal component, and R represents the residual component. Here, we can see that the additive decomposition for the Rose variable has a declining trend. The seasonal component has similar oscillations and the residual is scattered along the 0 values.

SPARKLING - ADDITIVE DECOMPOSITION



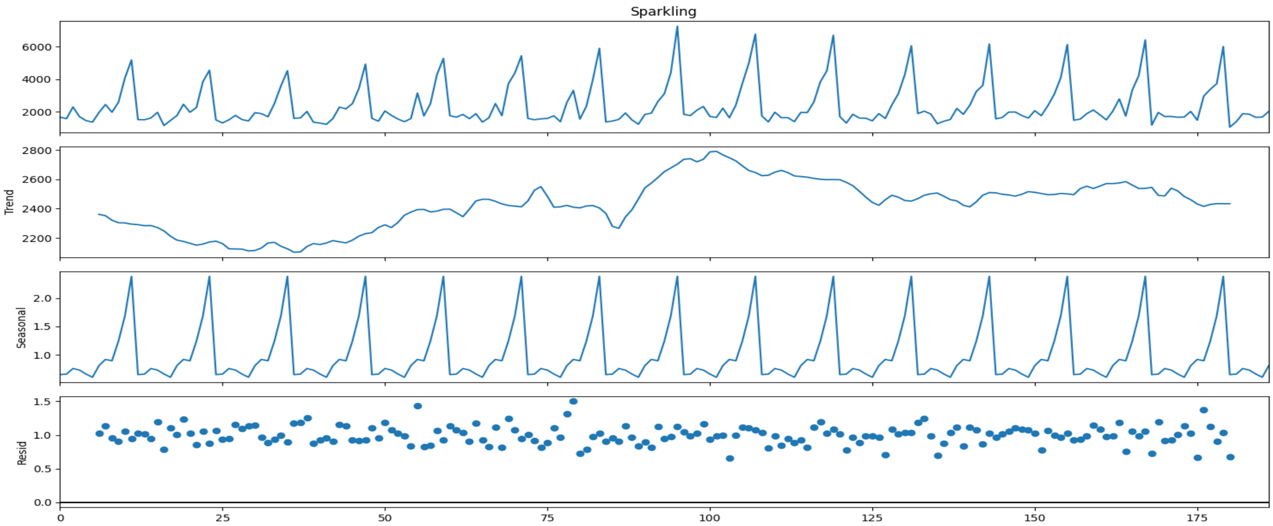
Sparkling variable has a changing trend where there was a downward trend at start. Later it went through a huge upwards trend and remained the same at somewhat similar level. The seasonality remains the same and the residuals are scattered around the midpoint with some outliers.

ROSE - MULTIPLICATIVE DECOMPOSITION



In the multiplicative model, the original time series is portrayed as the result of multiplying its trend, seasonal, and irregular components. In this model, the trend component retains the same units as the original series, while the seasonal and irregular components are dimensionless factors distributed around a value of 1. Rose variable has a declining trend. The seasonality of the variable remains the same and the residuals are much closely packed than the additive model.

SPARKLING - MULTIPLICATIVE DECOMPOSITION



The sparkling variable has similar trend and seasonality to its additive model. However, the residual is much closer to each other than the additive model.
Since, the residuals are much readable in the multiplicative model, The multiplicative model is preferred with both the variables.

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

ROSE

Training Data			Test Data		
	YearMonth	**Rose**		**YearMonth**	**Rose**
0	1980-01-01	112	**133**	1991-02-01	55
1	1980-02-01	118	**134**	1991-03-01	66
2	1980-03-01	129	**135**	1991-04-01	65
3	1980-04-01	99	**136**	1991-05-01	60
4	1980-05-01	116	**137**	1991-06-01	65

Above given is the training and test data for the Rose variable. The test variable starts from 1991 whereas the training variable ends in 1980.
SPARKLING

Training Data			Test Data		
	YearMonth	**Sparkling**		**YearMonth**	**Sparkling**
0	1980-01-01	1686	**133**	1991-02-01	2049
1	1980-02-01	1591	**134**	1991-03-01	1874

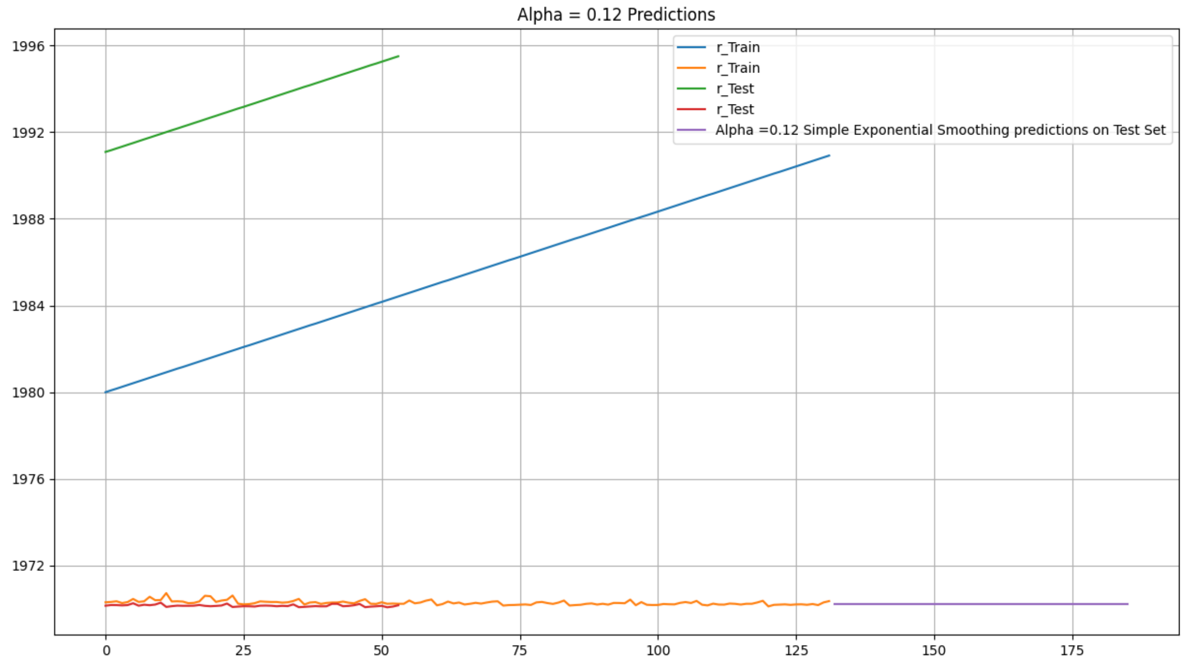
2	1980-03-01	2304	**135**	1991-04-01	1279
3	1980-04-01	1712	**136**	1991-05-01	1432
4	1980-05-01	1471	**137**	1991-06-01	1540

Above given is the training and test data for the Sparkling variable. The test variable starts from 1991 whereas the training variable ends in 1980.

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.

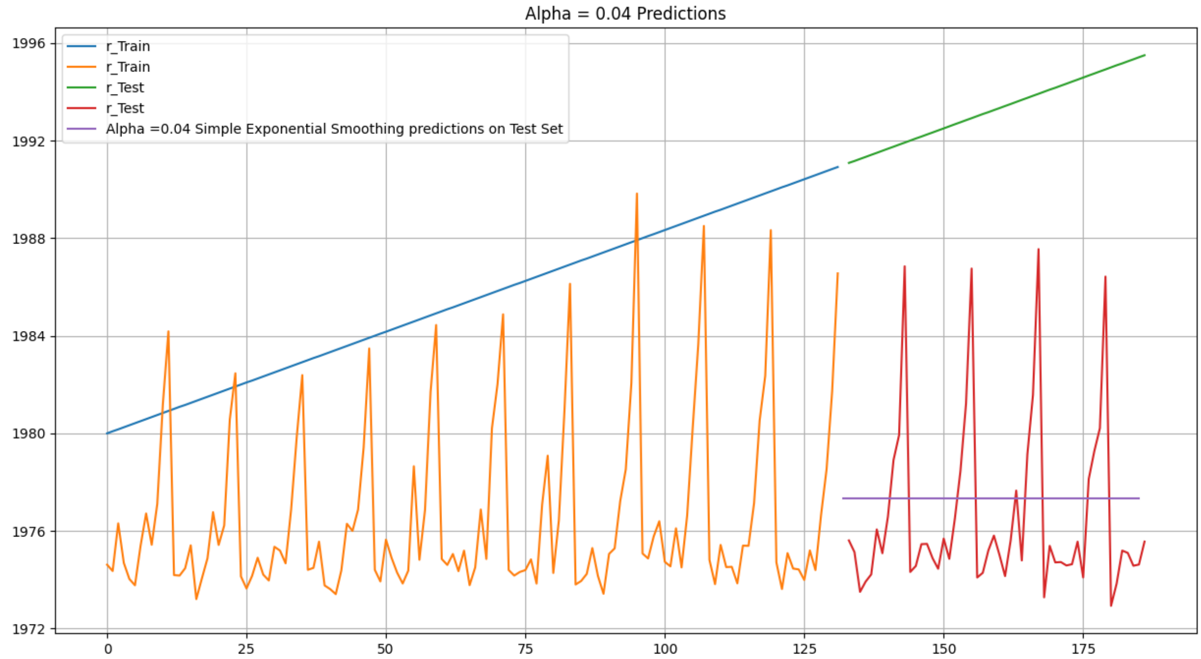
Simple Exponential Smoothing with additive errors

ROSE



Single (or simple) exponential smoothing is used for time-series data with no seasonality or trend. It requires a single smoothing parameter that controls the rate of influence from historical observations (indicated with a coefficient value between 0 and 1). The Rose variable, at 0.12 Alpha (suggested by the best parameters) has performed moderately, since it is just a flat line. It has an RMSE of 36.759.

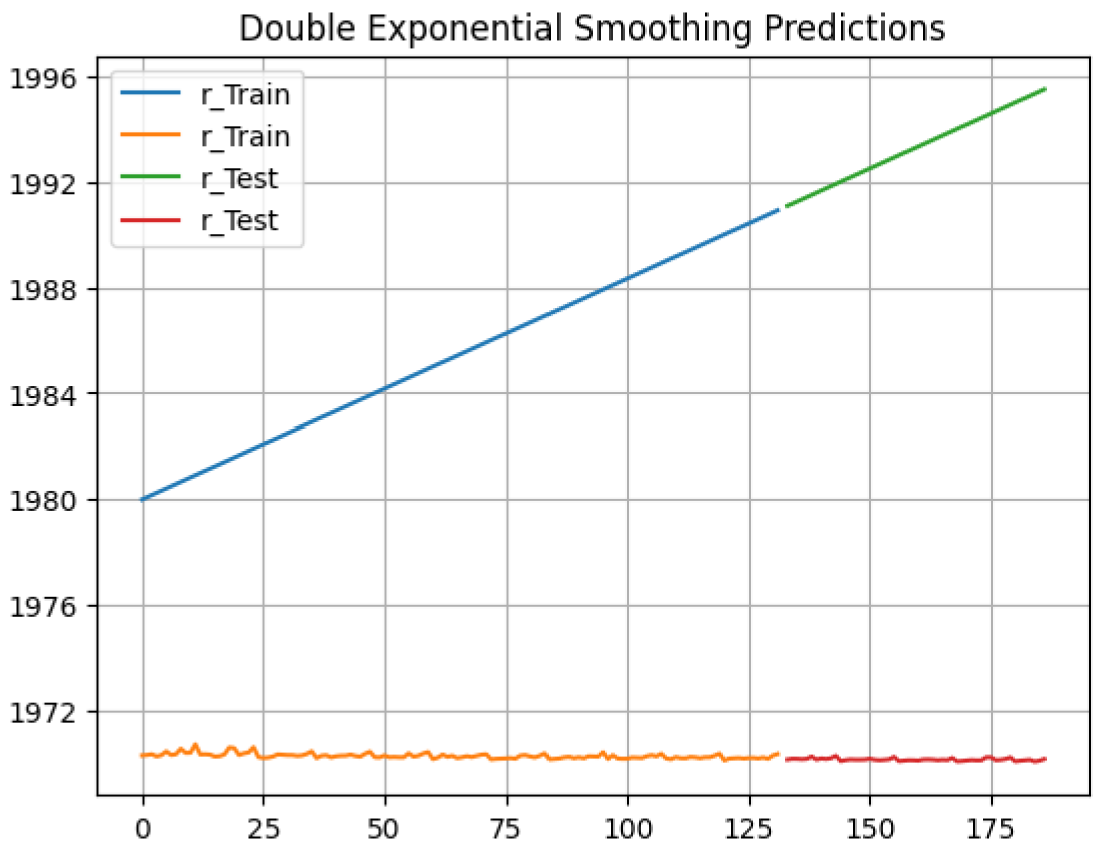
SPARKLING



The Rose variable, at 0.04 Alpha (suggested by the best parameters) has performed poorly, since it is just a flat line that does not predict the future. It has an RMSE of 1312.7 .

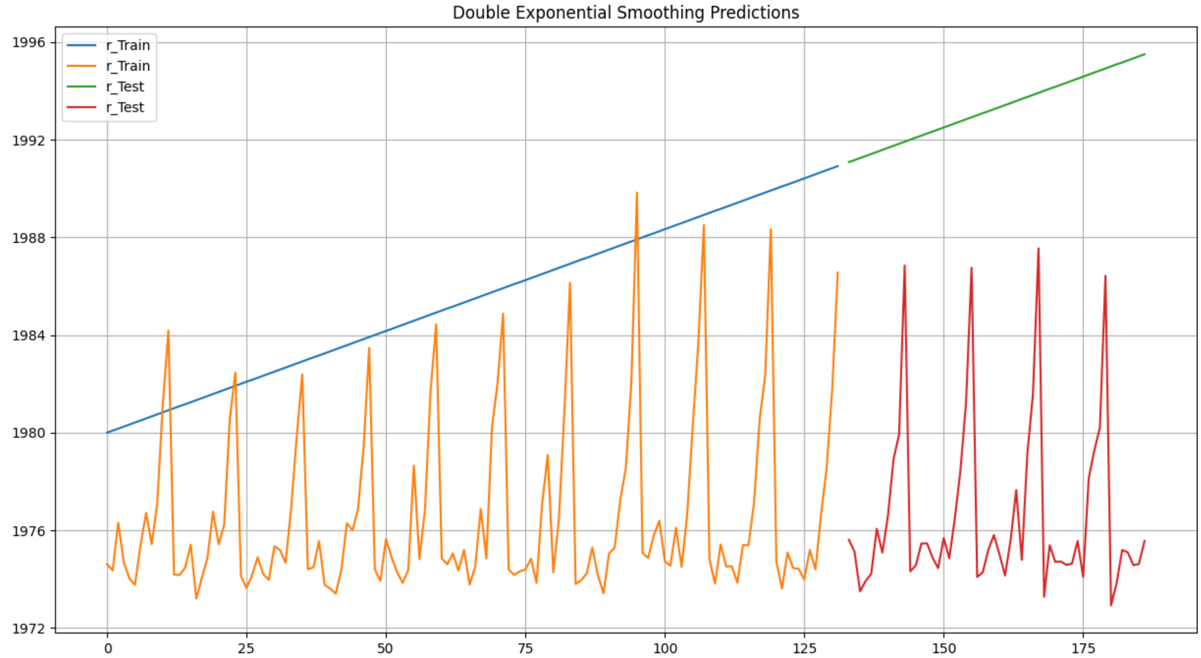
Double Exponential Smoothing model

ROSE



Double exponential smoothing employs a level component and a trend component at each period. Double exponential smoothing uses two weights, (also called smoothing parameters), to update the components at each period. The model has done well at predicting the Rose variable since it shows some seasonality. It has an RMSE of 16.71.

SPARKLING



The model has also performed well at best parameters on the Sparkling variable. It has an RMSE of 5249.12.

Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

ROSE

The additive Holt-Winters model is identical to the multiplicative model, except that seasonality is considered to be additive. This means that the forecasted value for each data element is the sum of the baseline, trend, and seasonality components. Here, the model for the rose variable has an RMSE of 74.004.

SPARKLING

THe model built on Sparkling variable has an RMSE of 574.14.

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

ROSE

The seasonal multiplicative method multiplies the trended forecast by the seasonality, producing the Holt-Winters' multiplicative forecast. The model built of rose has an RMSE of 74.004.

SPARKLING

The model built on the rose variable has an RMSE of 574.14.

LINEAR REGRESSION

ROSE

A time series regression forecasts a time series as a linear relationship with the independent variables. y t = X t β + ϵ t. The linear regression model assumes there is a linear relationship between the forecast variable and the predictor variables. The Linear regression model built on rose variable has an RMSE of 734.0

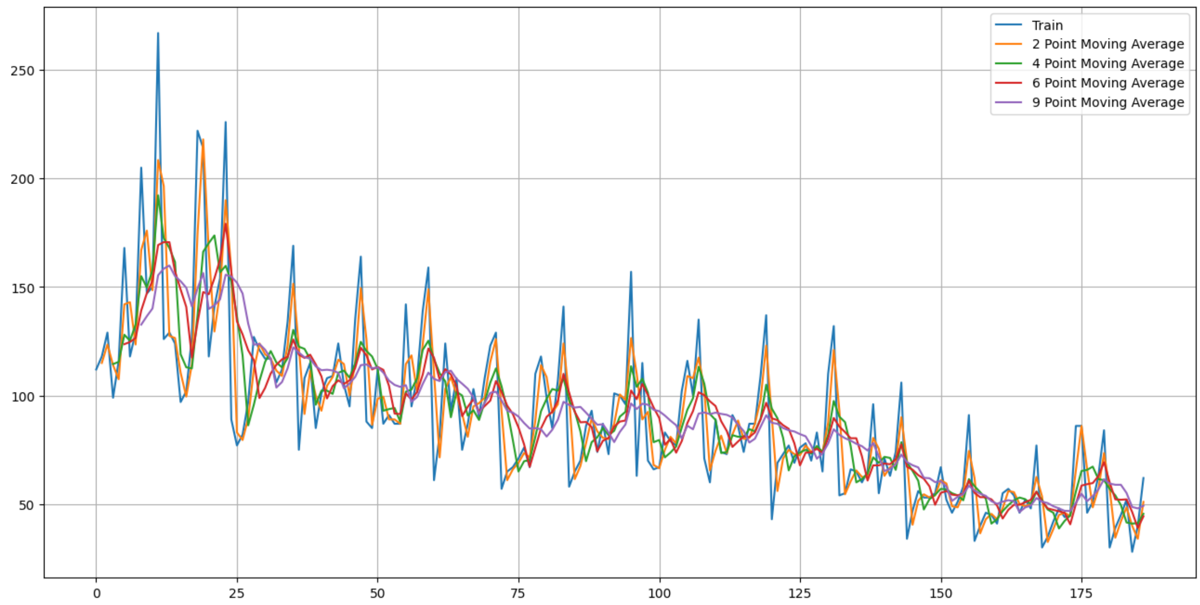
SPARKLING

Whereas, the linear regression model built on sparkling variables has an RMSE of 736.1.

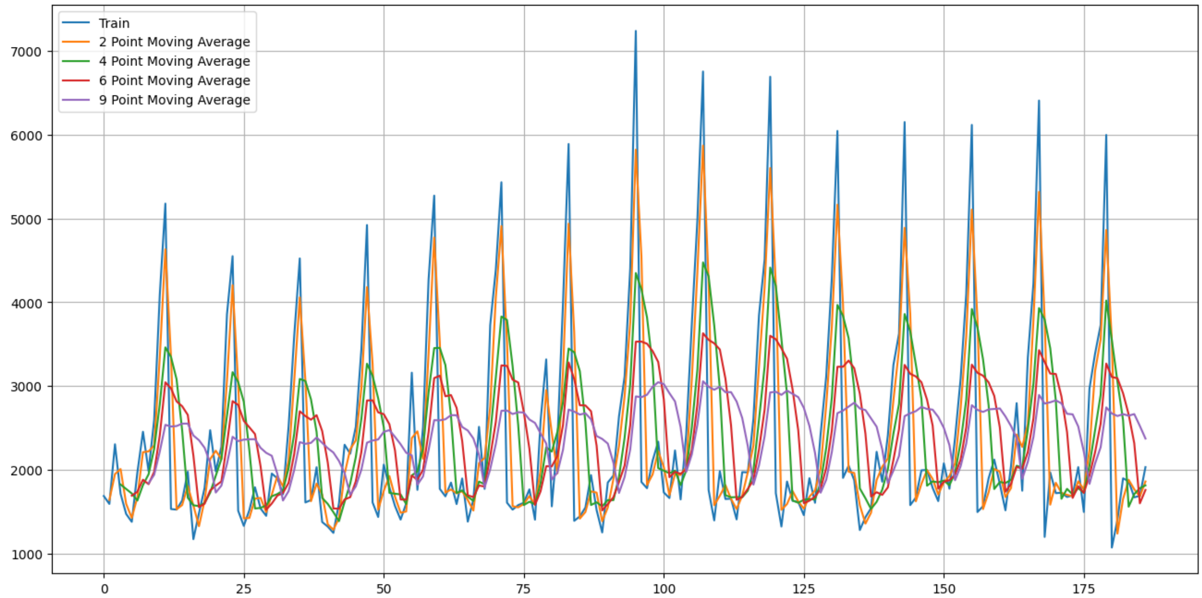
MOVING AVERAGE

ROSE

A moving average is a series of averages, calculated from historic data. Moving averages can be calculated for any number of time periods, for example a three-month moving average, a seven-day moving average, or a four-quarter moving average. The basic calculations are the same.A moving average process, or the moving average model, states that the current value is linearly dependent on the current and past error terms. We can see that the moving average model built on the rose variable has the following graphical representation and is able to replicate the original model to an extent. It is quite hard to discern the lines. But, one major takeaway is that as the moving average increases by points, the seasonality seems to decrease. The moving avg. model has following RMSE: For the 2 point Moving Average Model forecast on the Training Data, RMSE is 11.064. For the 4 point Moving Average Model forecast on the Training Data, RMSE is 14.922. For the 6 point Moving Average Model forecast on the Training Data, RMSE is 15.234. For the 9 point Moving Average Model forecast on the Training Data, RMSE is 15.862.



SPARKLING

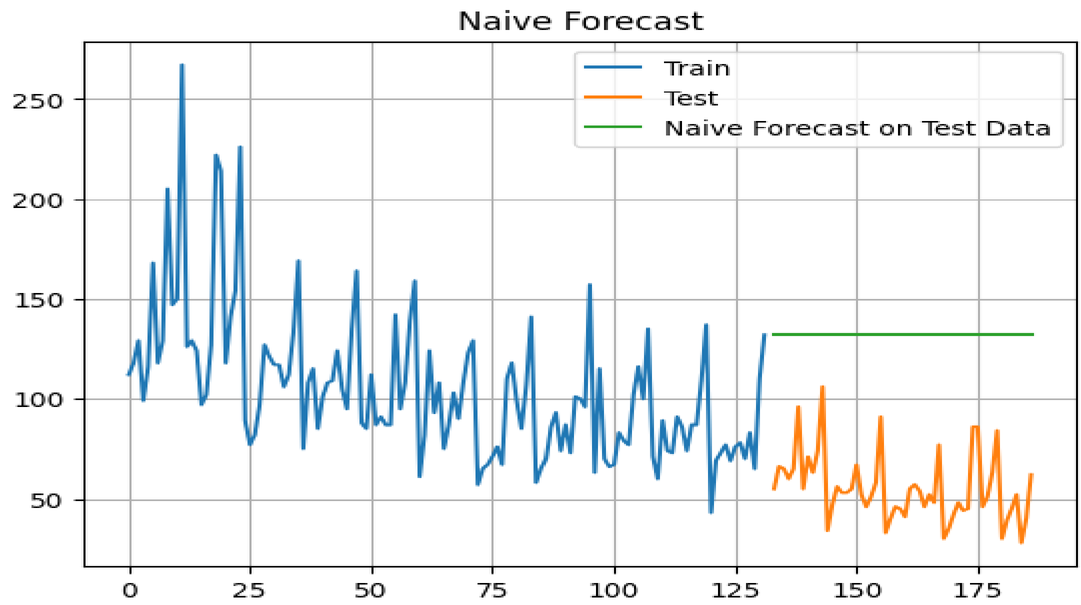


We can see that the moving average model built on the sparkling variable has the following graphical representation and is able to replicate the original model to an extent. It is quite hard to discern the lines. But, one major takeaway is that as the moving average increases by points, the seasonality seems to decrease. For the 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401. For the 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590. For the 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927. For the 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278.

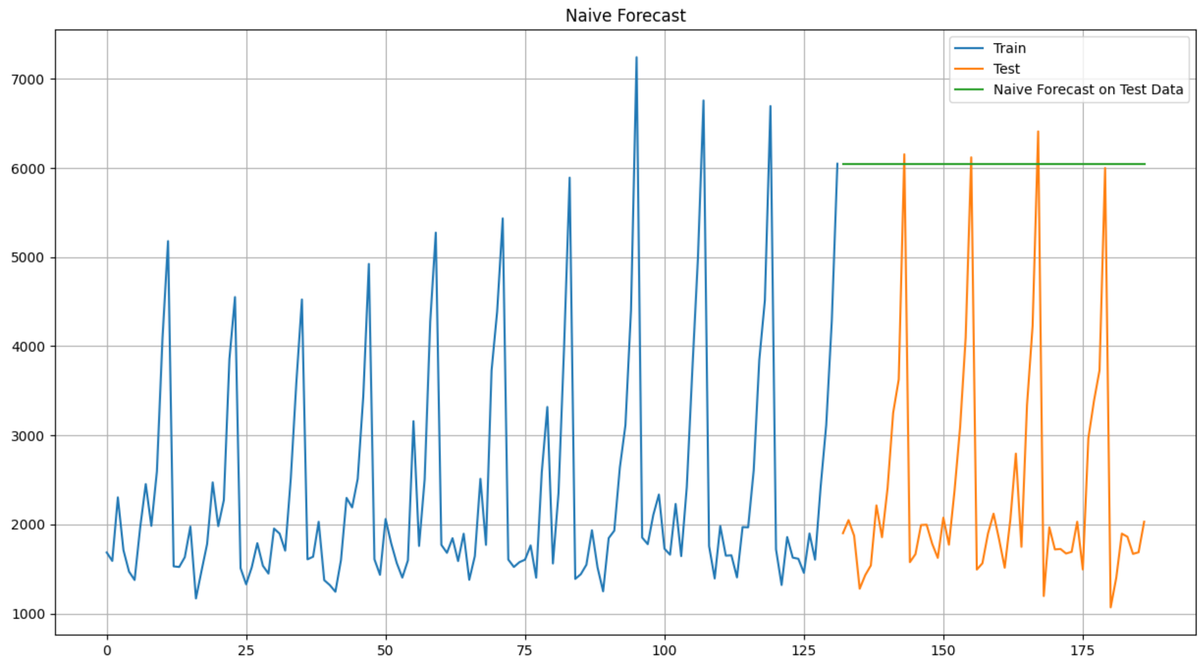
NAIVE MODEL

ROSE

The simple naive model predicts the next values as the last observed value. From the naive model built on the rose variable, we can see that the model has not been able to predict the test variable very well. It has just projected a straight line which does not predict the intricacies of test data. The model has an RMSE of 78.494.



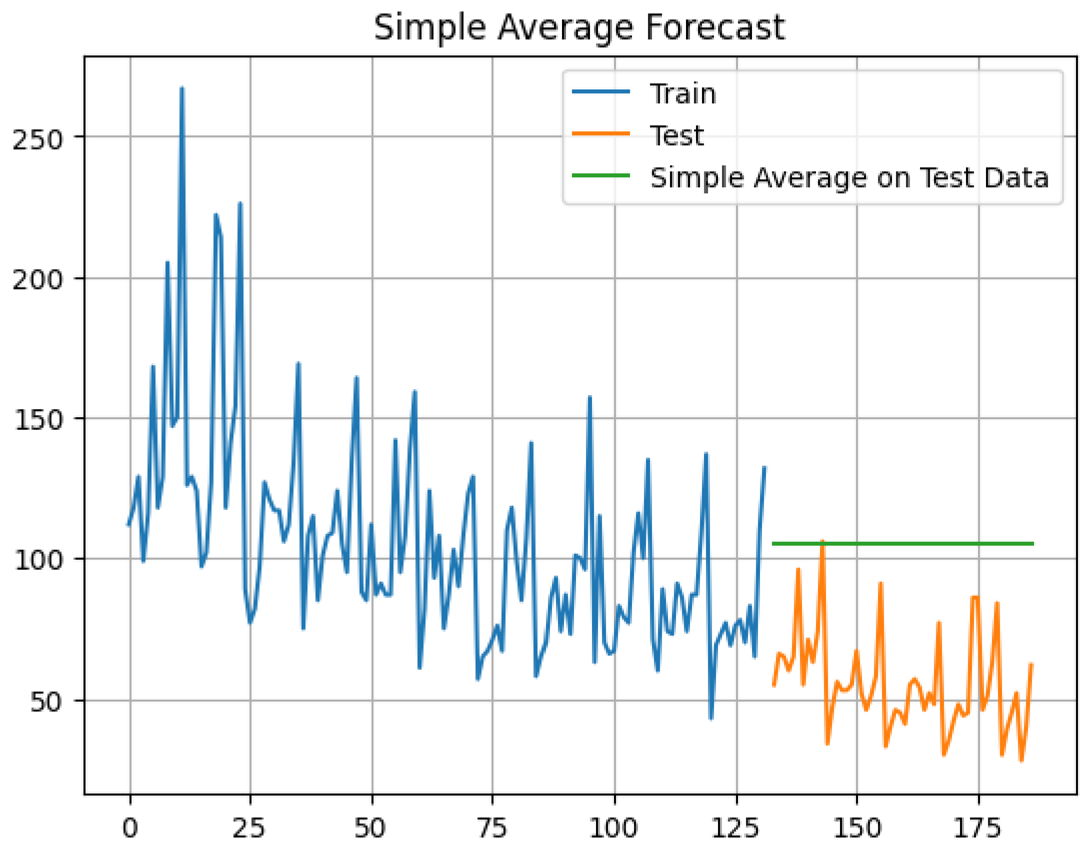
SPARKLING



From the Naive model built on sparkling variables we can see that the model has not performed well in predicting the test data. It has just projected a straight line which does not predict the intricacies of test data. It has an RMSE of 3864.279.

SIMPLE AVERAGE

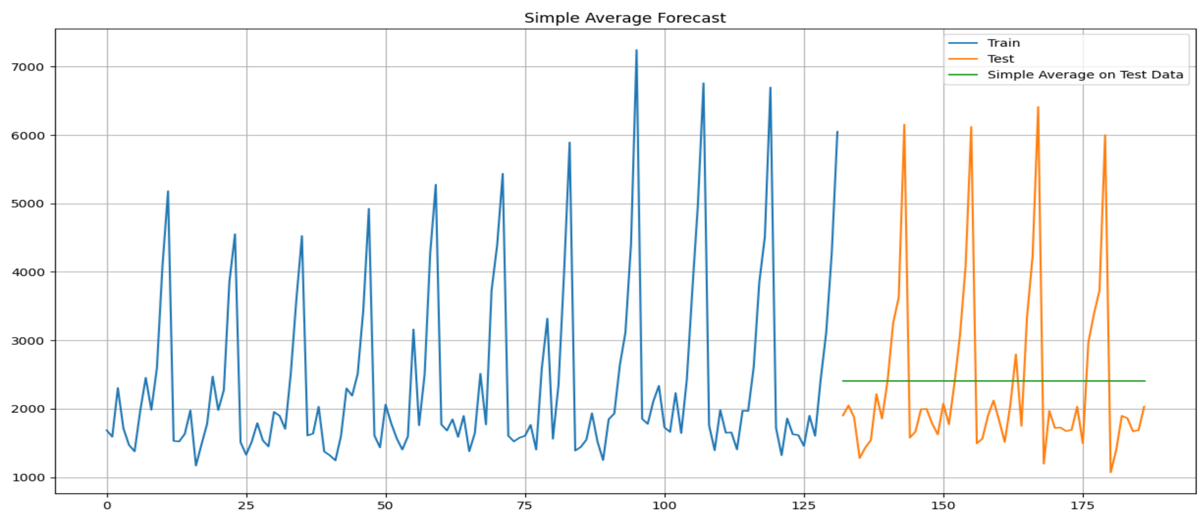
ROSE



SMA is one of the simplest forecasting methods that forecasts the future value of a time series data using the average of the past N observations. Here, N is the hyperparameter. The basic assumption of averaging models is that the series has a slow varying mean. Here, the simple average model built on the rose variable has not performed well, since it does not predict the test dataset very well. The model has an RMSE of 52.396.

SPARKLING

Here, the simple average model built on the rose variable has not performed well, since it does not predict the test dataset very well. The model has an RMSE of 1275.082.



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.

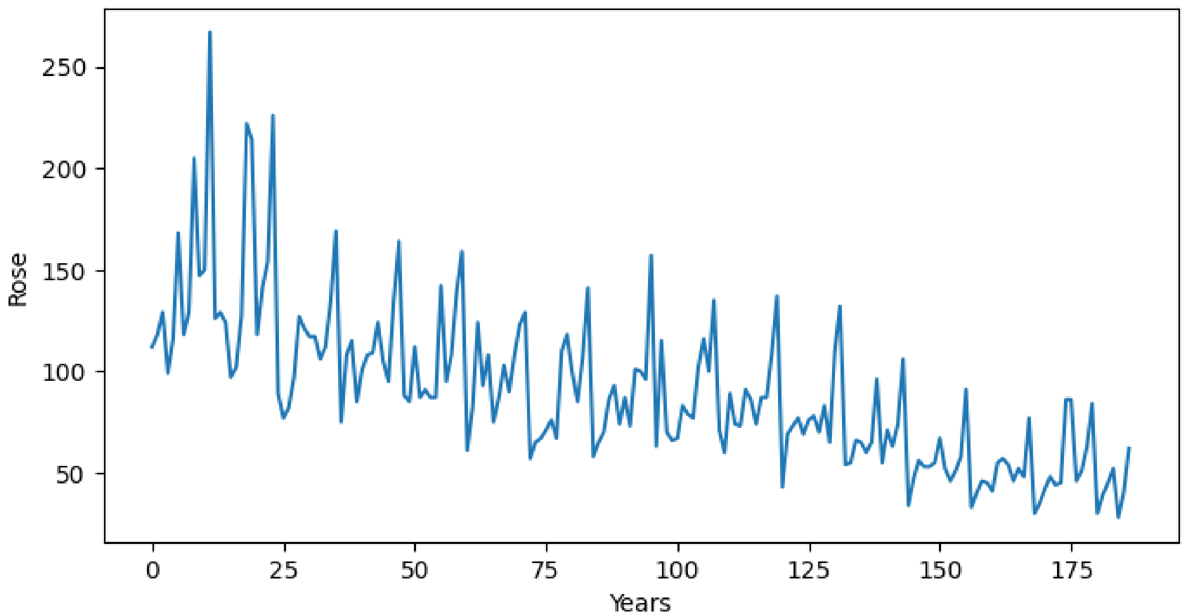
ROSE

The Dickey–Fuller test tests the null hypothesis that a unit root is present in an autoregressive (AR) time series model. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

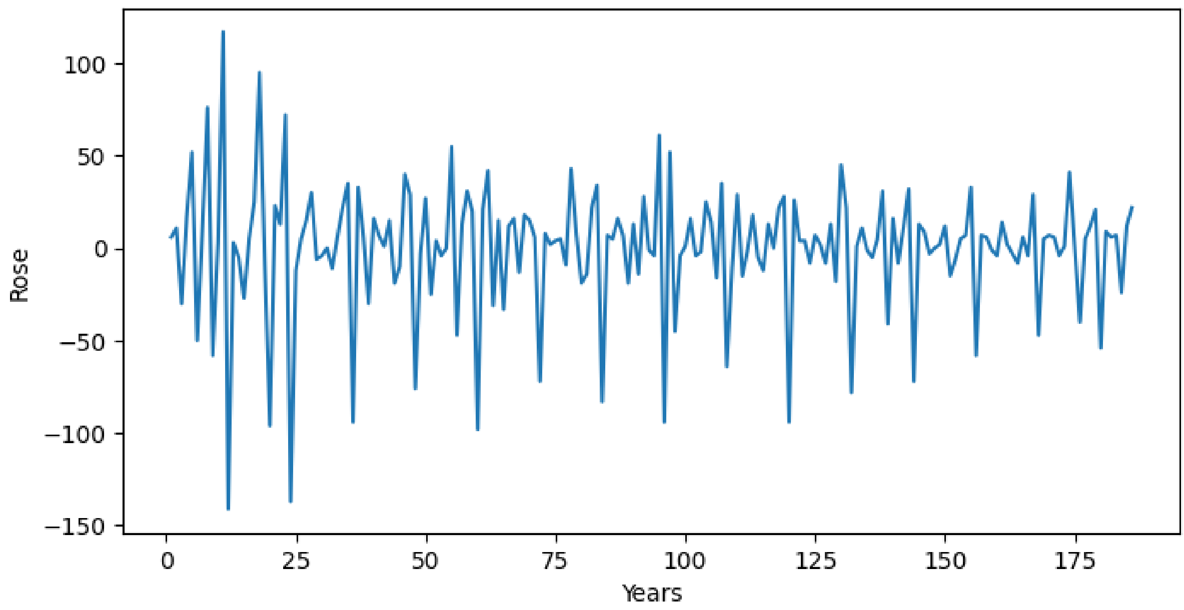
Let us take the Dicky Fuller test for the Rose variable. Results are as follows:
Ho:The given dataset is stationary.
Ha: The given dataset is not stationary.

While doing the Dicky Fuller test with an alpha of 0.05, the results are:
DF test statistic is -1.933
DF test p-value is 0.3167.

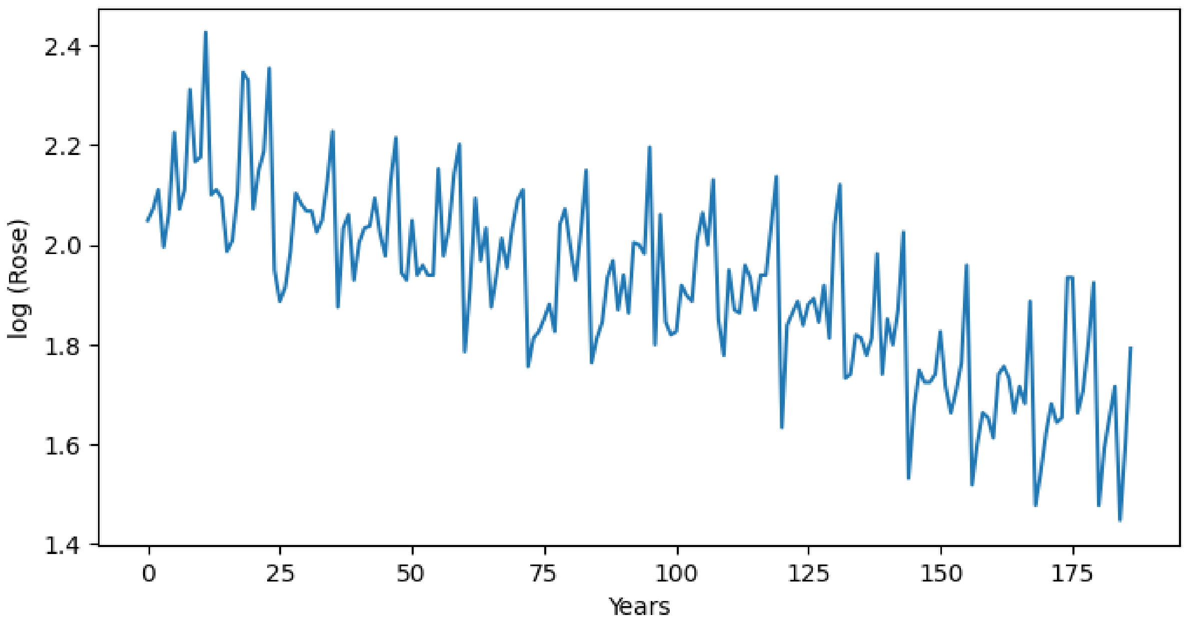
Therefore Since the test p-value is higher than 0.05, we fail to reject alternative hypotheses. The given dataset is non-stationary in nature.



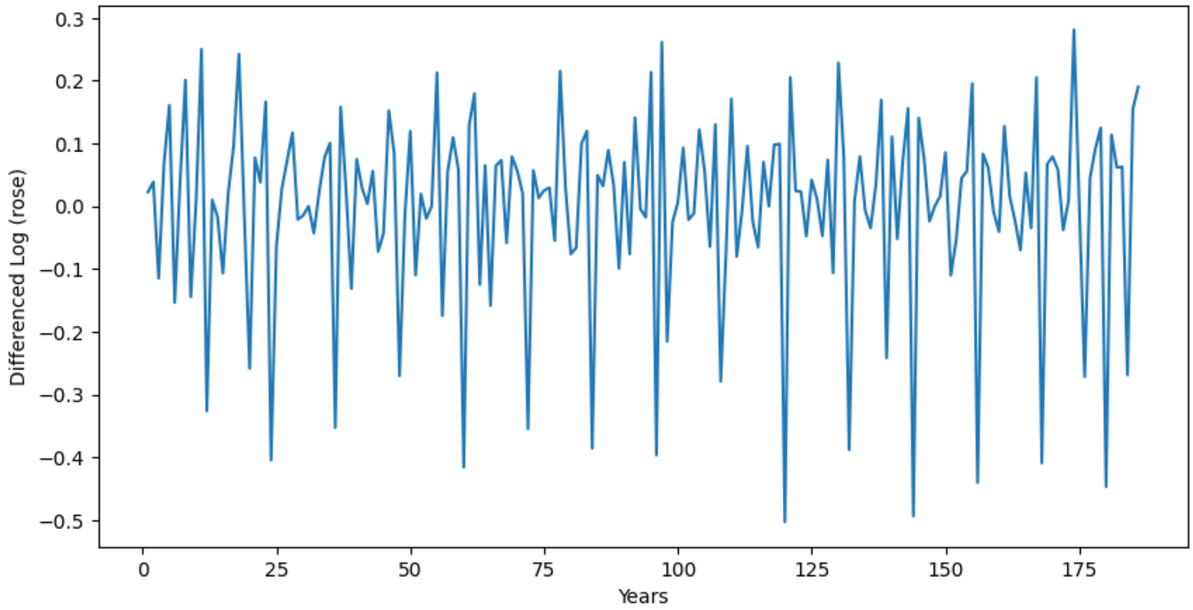
The above given diagram shows the data in its entirety. We can see that there is a declining trend in the data with seasonalities.



The above given diagram shows us the first difference model, where it shows us the difference between each month,that is month on month increase in sales. We can see that There seems to be a higher difference at the beginning that slowly decreases along the year.



The logarithmic model is where the log10 function is applied on the dataset. Here, the negative values of the differenced model are removed. Thus, giving us a stable model that can be valued in between parallel lines. This diagram shows us the seasonality and the trend but also by regulating it.



Here, the differenced logarithmic model combines both the previous functions to our dataset, that is to perform differencing on the log model. This makes the model stationary in nature, where there are 0 mean differences with equal trend.

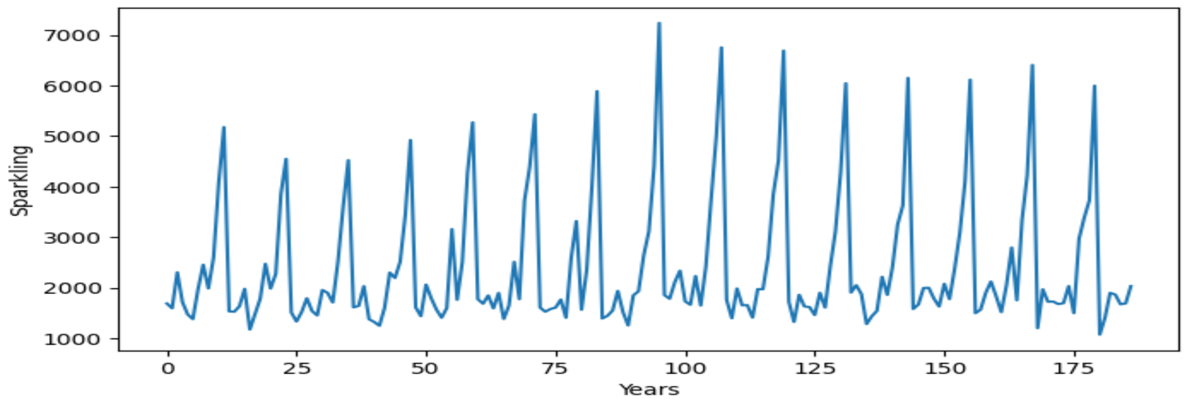
SPARKLING

Let us take the Dicky Fuller test for the Rose variable. Results are as follows:

Ho:The given dataset is stationary.
Ha: The given dataset is not stationary.

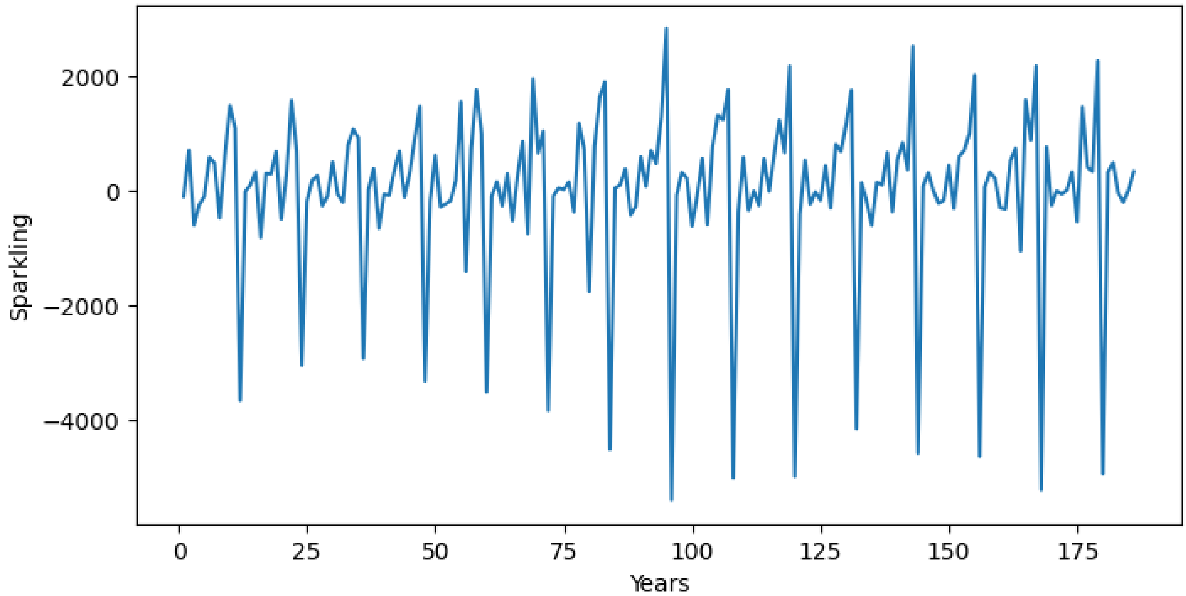
While doing the Dicky Fuller test with an alpha of 0.05, the results are:
DF test statistic is -1.360
DF test p-value is 0.6011

Therefore Since the test p-value is higher than 0.05, we fail to reject alternative hypotheses. The given dataset is non-stationary in nature.

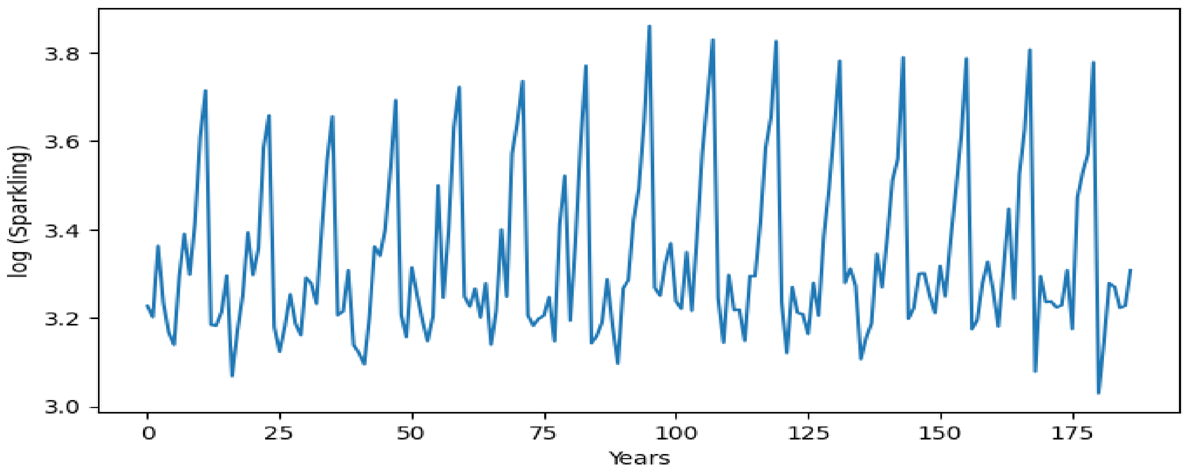


The above given diagram shows the data in its entirety. We can see that there is an increasing trend at the beginning that slows down in the data along with seasonalities.

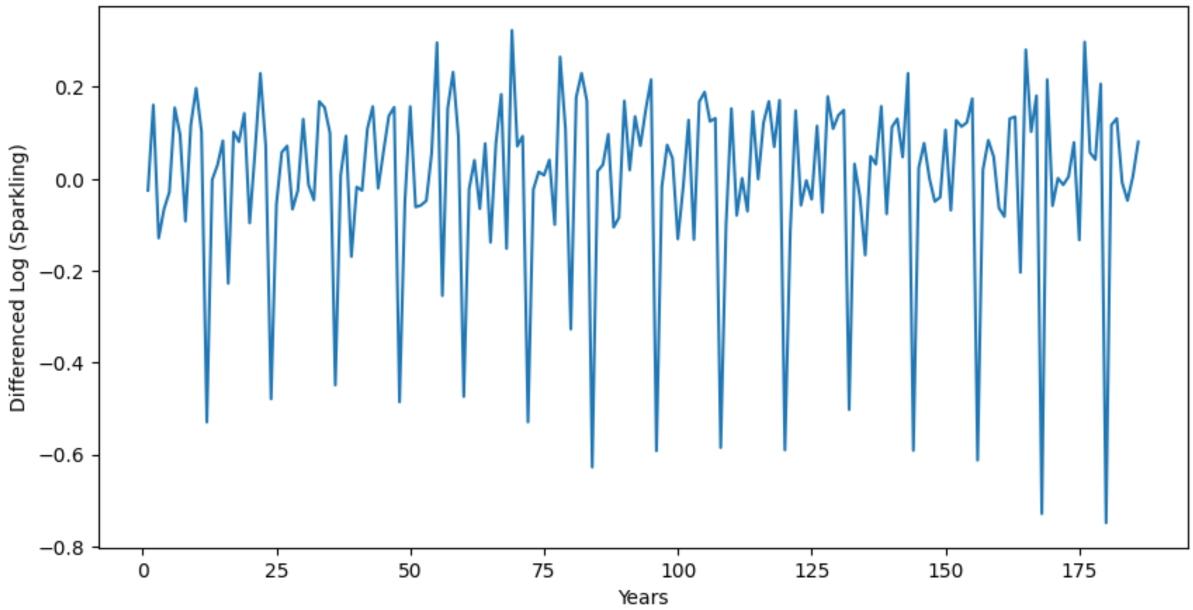
The below given diagram shows us the first difference model, where it shows us the difference between each month,that is month on month increase in sales. We can see that There seems to be a lesser difference which further increases and decreases as it goes along the year.



The logarithmic model is where the log10 function is applied on the dataset. Here, the negative values of the differenced model are removed. Thus, giving us a stable model that can be valued in between parallel lines, where the differences are reduced. This diagram shows us the seasonality and the trend but also by regulating it.



Here, given below, the differenced logarithmic model combines both the previous functions to our dataset, that is to perform differencing on the log model. This attempts to makes the model stationary in nature, where there are lesser mean differences with equal trend.



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.

ROSE

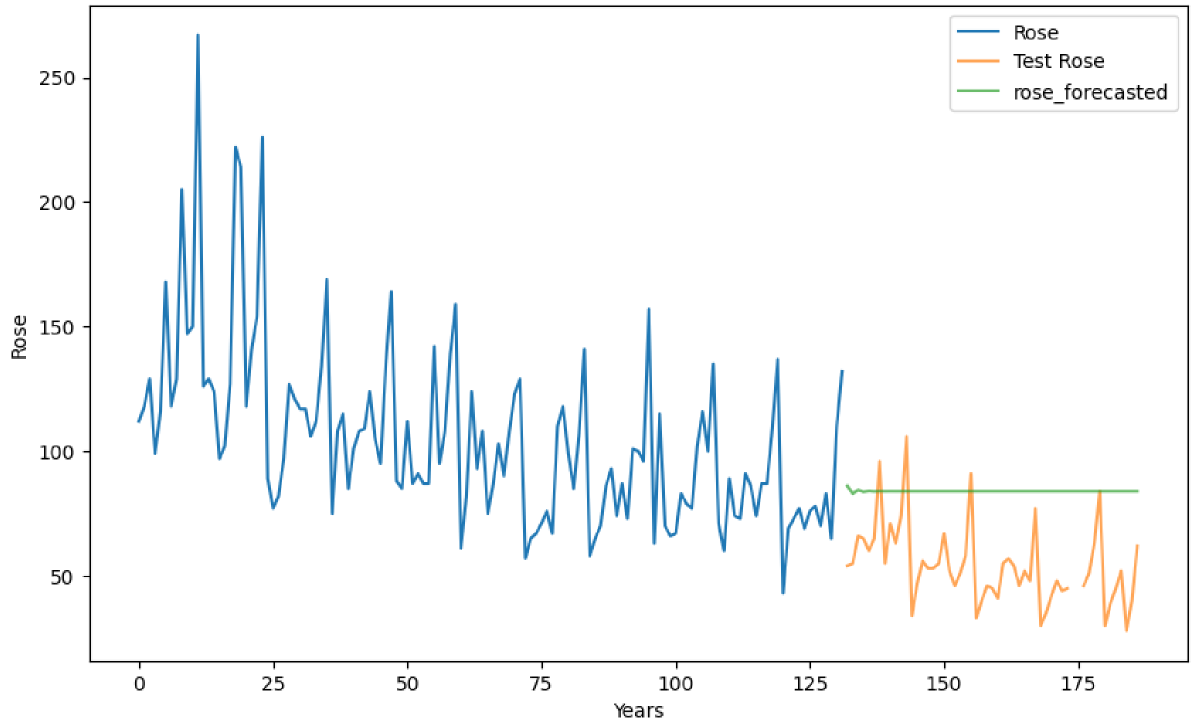
First, different pdq (p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and. q is the number of lagged forecast errors in the prediction equation) values were generated with s (seasonality) value set at 12 since the year is separated into 12 months. Out of all these combinations, different sets of parameters and AIC were calculated and imputed into a dataframe.

	param	AIC
4	(1, 1, 2)	-186.195065
3	(1, 1, 1)	-186.101119
1	(1, 0, 2)	-185.365917
9	(2, 1, 1)	-184.829175
7	(2, 0, 2)	-184.730802

Out of all these combinations, we can see that the first combination has the lowest AIC on training data. Therefore an ARIMA/SARIMA model was created using the given parameters.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(1, 1, 2)   Log Likelihood                  97.098
Date:                Sun, 03 Sep 2023   AIC                           -186.195
Time:                        13:58:12   BIC                           -174.694
Sample:                             0   HQIC                          -181.522
                                - 132
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5038      0.288     -1.750      0.080      -1.068       0.061
ma.L1         -0.2168      0.266     -0.814      0.416      -0.739       0.305
ma.L2         -0.6340      0.222     -2.862      0.004      -1.068      -0.200
sigma2         0.0131      0.002      8.119      0.000       0.010       0.016
==============================================================================
```

We can discern from the results that the moving average.L2 is significant compared to other parameters.



From the diagrammatic representation, we can see that the SARIMA model did not perform well in predicting the rose test data. The Root Mean Squared Error of our forecasts is 33.838, which is not terrible in comparison with other predictive models.

SPARKLING

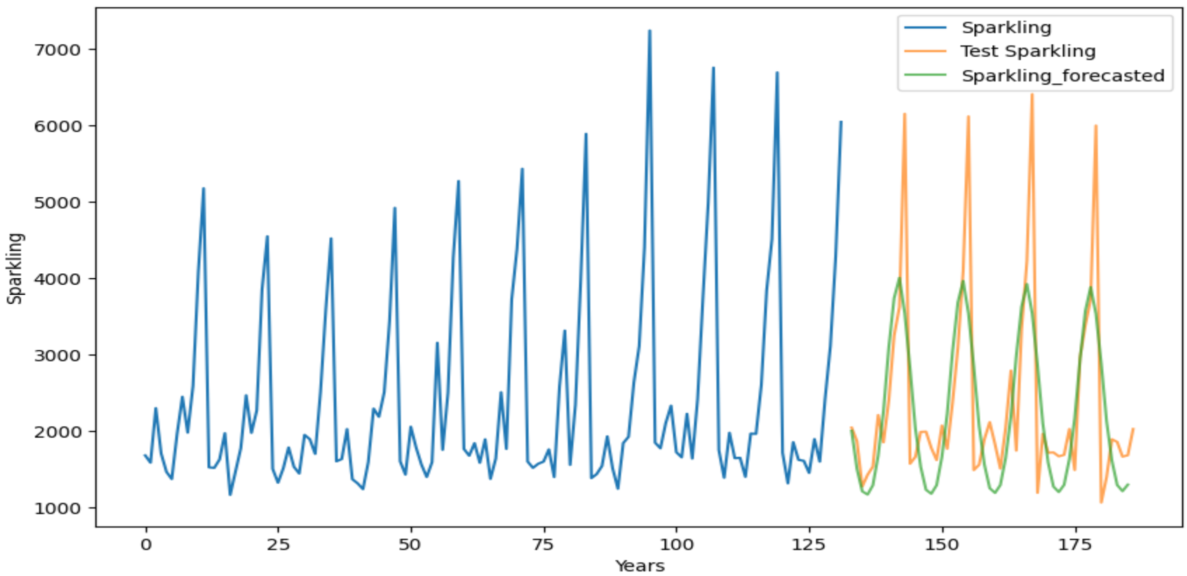
First, different pdq (p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and. q is the number of lagged forecast errors in the prediction equation) values were generated with s (seasonality) value set at 12 since the year is separated into 12 months. Out of all these combinations, different sets of parameters and AIC were calculated and imputed into a dataframe.

	param	AIC
14	(3, 0, 3)	-136.429366
8	(2, 0, 3)	-130.749109
17	(3, 1, 3)	-116.361803
6	(2, 0, 1)	-106.955371
13	(3, 0, 2)	-103.194396

Out of all these combinations, we can see that the first combination has the lowest AIC on training data. Therefore an ARIMA/SARIMA model was created using the given parameters.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 0, 3)   Log Likelihood                  76.215
Date:                Sun, 03 Sep 2023   AIC                           -136.429
Time:                        14:58:52   BIC                           -113.367
Sample:                             0   HQIC                          -127.058
                                - 132
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.3341      0.010    331.860      0.000       3.314       3.354
ar.L1          0.7373      0.383      1.925      0.054      -0.013       1.488
ar.L2          0.7244      0.660      1.097      0.273      -0.570       2.019
ar.L3         -0.9913      0.381     -2.600      0.009      -1.739      -0.244
ma.L1         -0.7929      0.373     -2.126      0.034      -1.524      -0.062
ma.L2         -0.7942      0.670     -1.185      0.236      -2.108       0.519
ma.L3          0.9838      0.397      2.478      0.013       0.206       1.762
sigma2         0.0173      0.003      5.662      0.000       0.011       0.023
==============================================================================
```

From the SARIMA model, we can see that the ar.L2 and ma.L2 are statistically significant compared to other parameters.



From the diagrammatic representation, we can see that the SARIMA model performs decently well in predicting the sparkling test data. However it did not predict the intricate seasonalities in the dataset. The Root Mean Squared Error of our forecasts is 948.84, which is not terrible in comparison with other predictive models.

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

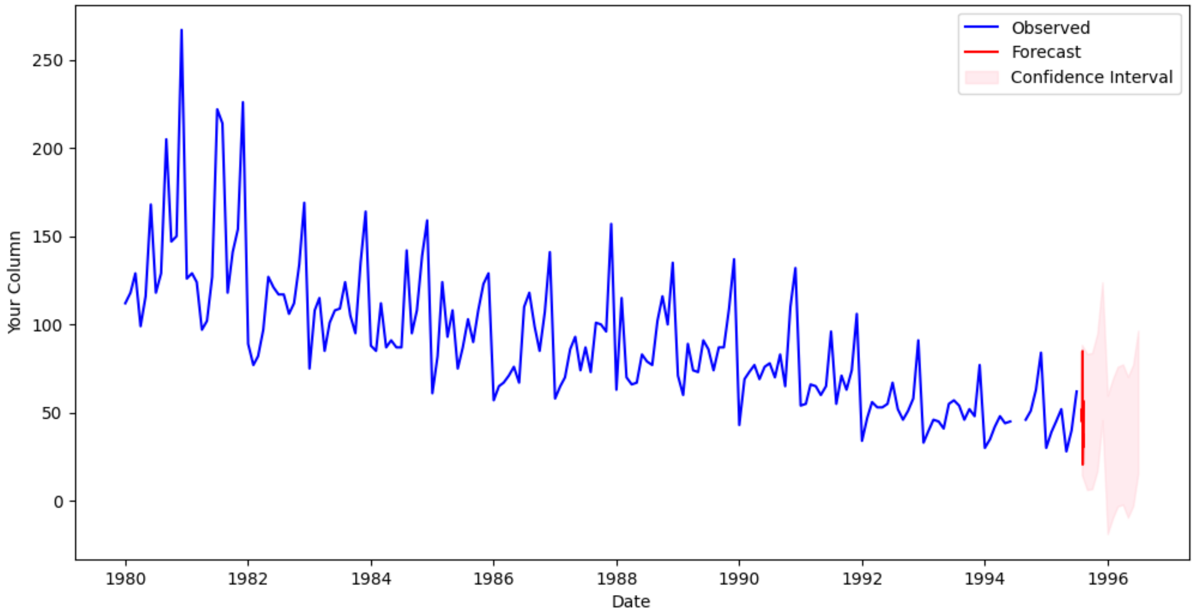
From the model building exercise: Following are the RMSE of all the models predicting the Rose test data.

MODEL	RMSE
Simple Exponential Smoothing with additive errors	36.759
Double Exponential Smoothing model	16.71
Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors	74.004
Holt-Winters - ETS(A, A, M) - Holt Winter's linear method with multiplicative error	74.004
LINEAR REGRESSION	734
MOVING AVERAGE (2 point)	11.064
MOVING AVERAGE (4 point)	14.922
MOVING AVERAGE (6 point)	15.234
MOVING AVERAGE (9 point)	15.862
NAIVE MODEL	78.494

SIMPLE AVERAGE	52.396
SARIMA	33.838

From the table, we can see that the double exponential model has the lowest root mean square error. Thus, mathematically the Double exponential model is the best model to predict the test data most effectively. However, I have decided to use the SARIMA model to predict 12 months into the future. This is because I believe that the SARIMA model has better performance in capturing and regulating all the components and parameters of the test data. Furthermore, the RMSE of the SARIMA model is not particularly worse in comparison with the other models.

Following are the results of the same:



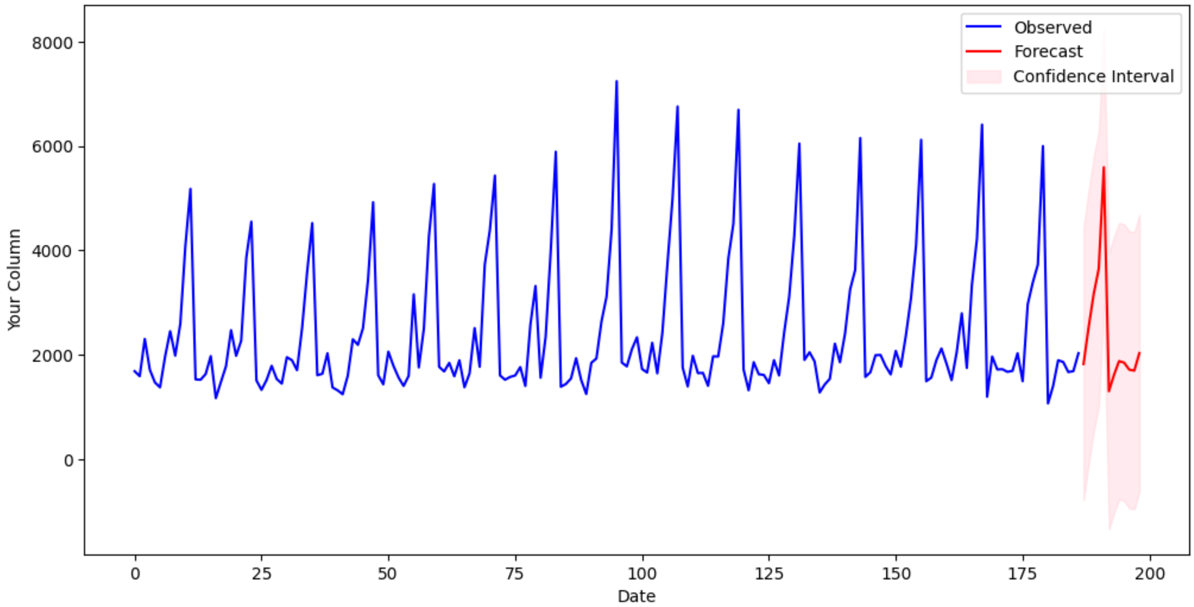
From the model building exercise: Following are the RMSE of all the models predicting the Sparkling test data.

MODEL	RMSE
Simple Exponential Smoothing with additive errors	1312.7
Double Exponential Smoothing model	5249.12
Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors	574.14
Holt-Winters - ETS(A, A, M) - Holt Winter's linear method with multiplicative error	574.14
LINEAR REGRESSION	736.1
MOVING AVERAGE (2 point)	813.401

MOVING AVERAGE (4 point)	1156.59
MOVING AVERAGE (6 point)	1283.927
MOVING AVERAGE (9 point)	1346.278
NAIVE MODEL	3864.279
SIMPLE AVERAGE	1275.082
SARIMA	948.84

From the table, we can see that both the Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive and multiplicative error has the lowest root mean square error. Thus, mathematically both the Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive and multiplicative error is the best model to predict the test data most effectively. However, I have decided to use the SARIMA model to predict 12 months into the future. This is because I believe that the SARIMA model has better performance in capturing and regulating all the components and parameters of the test data. Furthermore, the RMSE of the SARIMA model is not particularly worse in comparison with the other models.

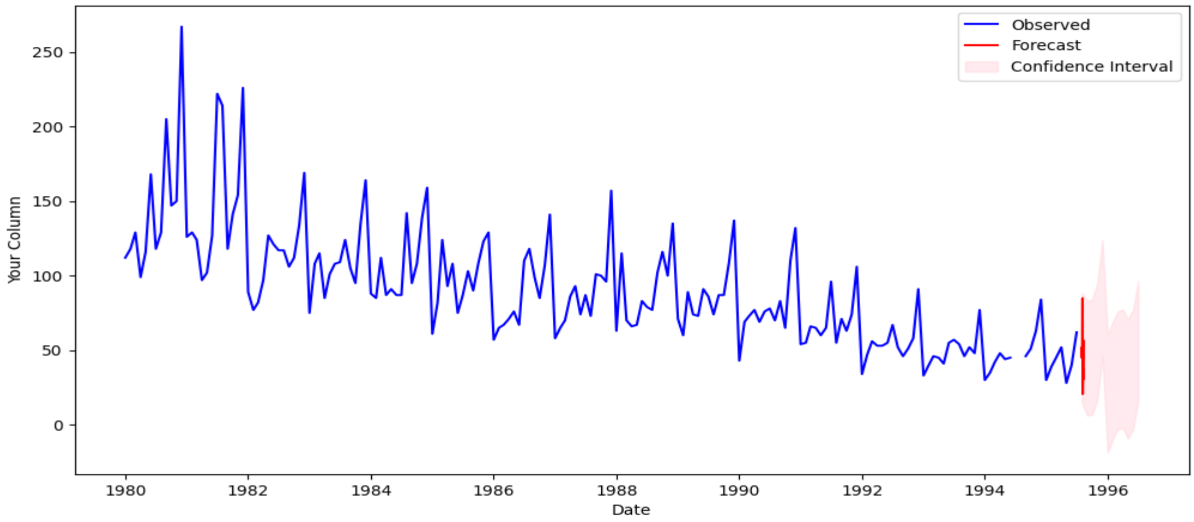
Following are the results of the same:

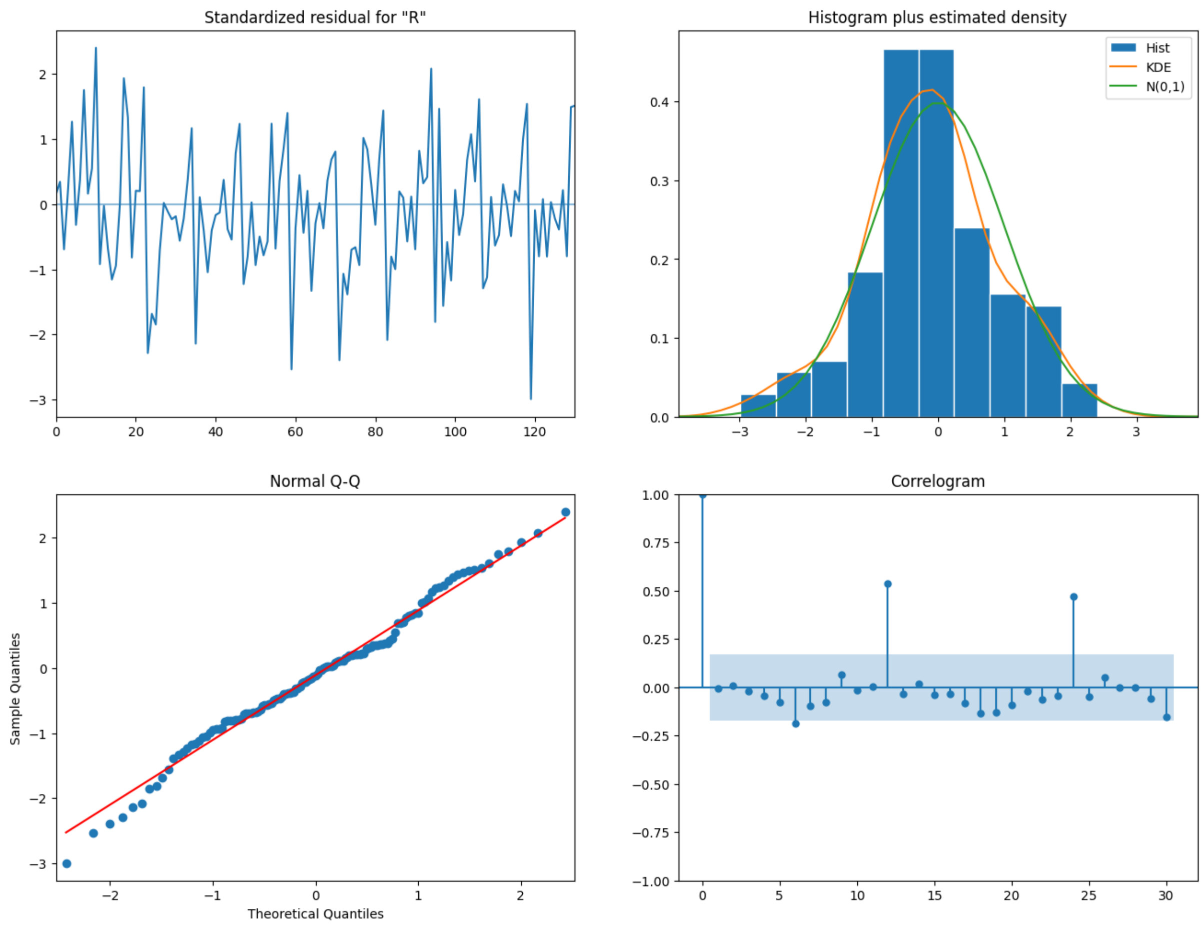


The inference and business insights are discussed in the next section.

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

ROSE





INFERENCE

1. Above given diagram is the representation of the model forecast 12 months into the future. Here, in this prediction we can see that the Rose variable will follow its downward trend and have similar seasonalities.
2. At first, it might have a preliminary increase and later decrease. Thus, following this pattern.

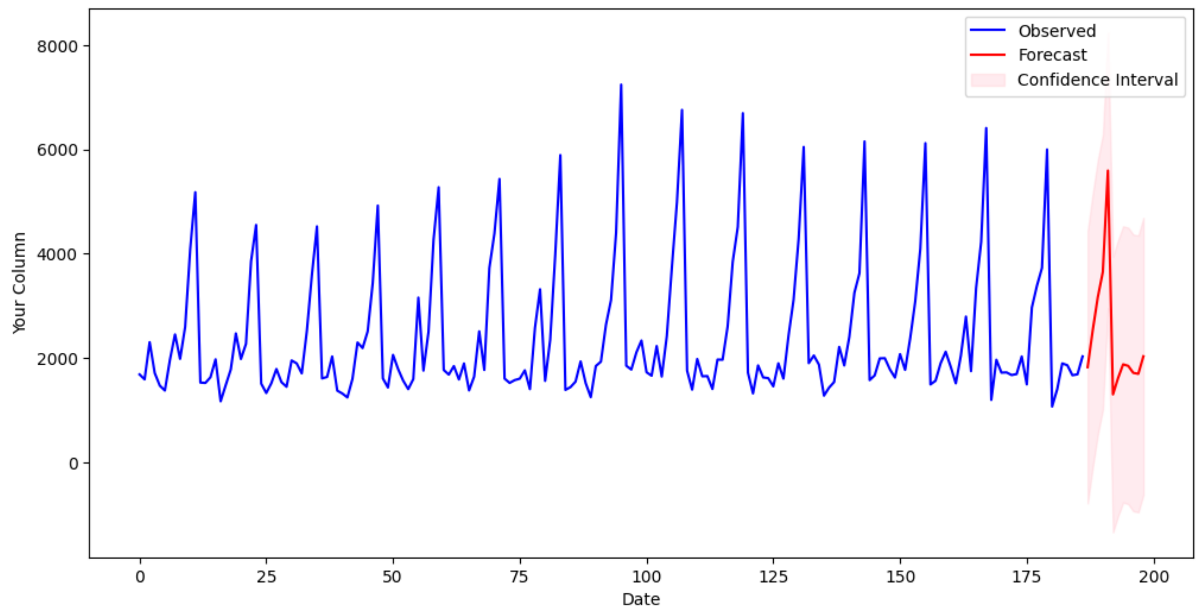
3. There seems to be a certain residual distribution that is uncounted by the model as evidenced in the Normal Q-Q plot where the residual points deviate from the normal distribution line.
4. The correlogram also has a similar phenomenon where there are outliers from the blue shaded area representing normal distribution.

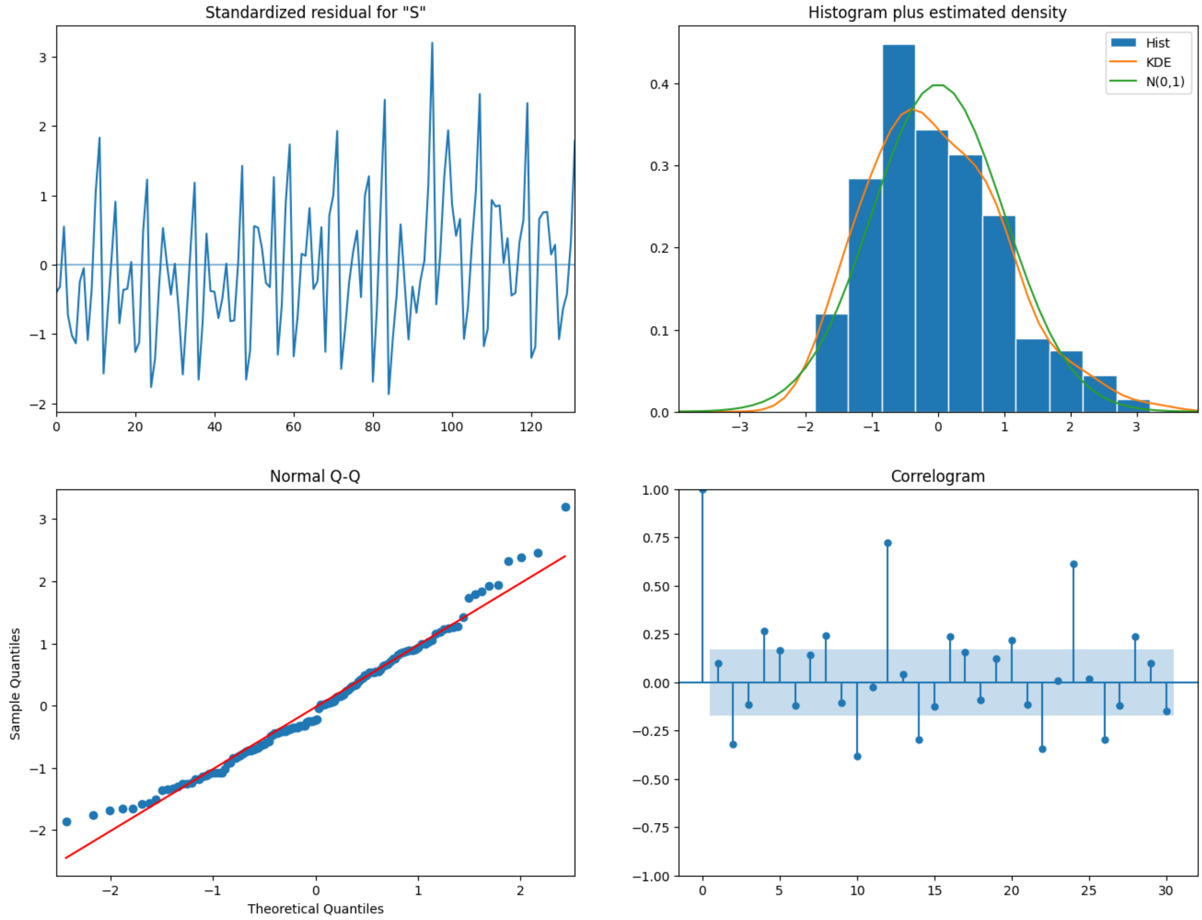
BUSINESS INSIGHTS

Following are the business insights for the above found inferences:

1. The company must try to improve their overall sales, since Rose sales have been on decline since the 1980s.
2. The company must find the root causes of this issue. It must try to take in the seasonal changes in the data that repeats throughout the year.
3. There are also certain factors of the data that are not accounted for in the model prediction. It is necessary to investigate what these deviations are and what effect do they have on sales.

SPARKLING





INFERENCE

1. Above given diagram is the representation of the model forecast 12 months into the future. Here, in this prediction we can see that the Sparkling variable will follow its minute changes trend and have similar seasonalities with some differences in the middle.
2. There seems to be some differences between the months in the dataset, That is there are some upticks in Sparkling sales.
3. There seems to be a certain residual distribution that is uncounted by the model as evidenced in the Normal Q-Q plot where the residual points deviate from the normal distribution line at the beginning and the end of the dataset.
4. The correlogram also has a similar phenomenon where there are some significant number of outliers from the blue shaded area representing normal distribution.

BUSINESS INSIGHTS

Following are the business insights for the above found inferences:

1. The company must try to improve their overall sales, since Sparkling sales have stagnated since the 1980s.
2. The company must find the root causes of this issue. It must try to take in the seasonal changes in the data that repeats throughout the year.

3. There are some minor deviations in the dataset which accounts for the increase in Sparkling sales. The root cause of these differences and their effect on sales must be investigated.