TIME SERIES FORECAST PROJECT

BUSINESS REPORT

PGP DSBA – APR 2023

Submitted by:
Sruthi C

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.

Problem 1: Rose wine dataset

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

- Data:

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0
...	...	...
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

187 rows × 2 columns

Figure 1 : Rose wine data without date-time format

- The data set contains 2 columns and 187 rows of data.
- The time stamp is from Jan 1980 to July 1995.

- Time Series data:
  Time stamps were added to the data frame to make it a time series data and was set as index. The YearMonth column has been removed.

Time_stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Figure 2 : Timeseries data - Rose

- Datatype:

```
Rose    float64
dtype: object
```

Figure 3 : Datatype

- Null values: There are 2 null values in the data.

Time_stamp	Rose
1994-07-31	NaN
1994-08-31	NaN

Figure 4 : Null values

- Plotting the timeseries data:

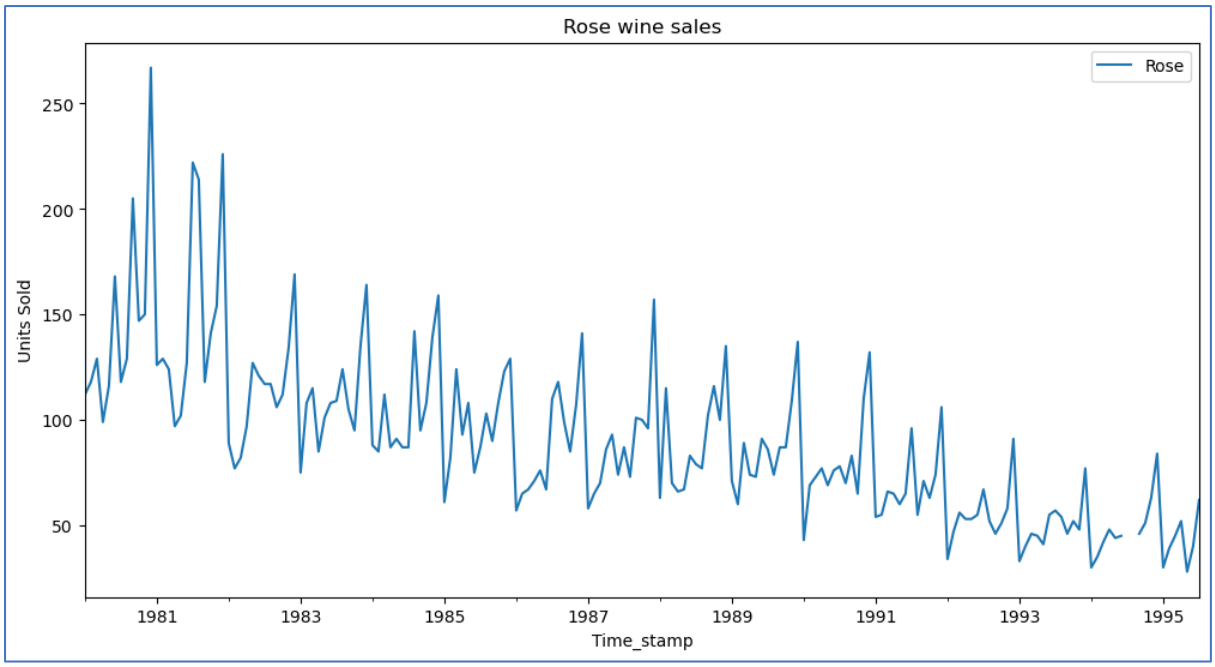


Figure 5 : Rose wine sales time series data

- A decreasing trend with the presence of a multiplicative seasonality is present in the data.
- The decreasing trend can been observed as a decline in the demand of rose wine over these years.
- The break in the plot has been observed as the presence of null values.

- Imputing null values: The missing values are imputed using polynomial interpolation of order 2. The new values are,

```
YearMonth
1994-07-01    45.364189
1994-08-01    44.279246
Name: Rose, dtype: float64
```

Figure 6 : Imputed values

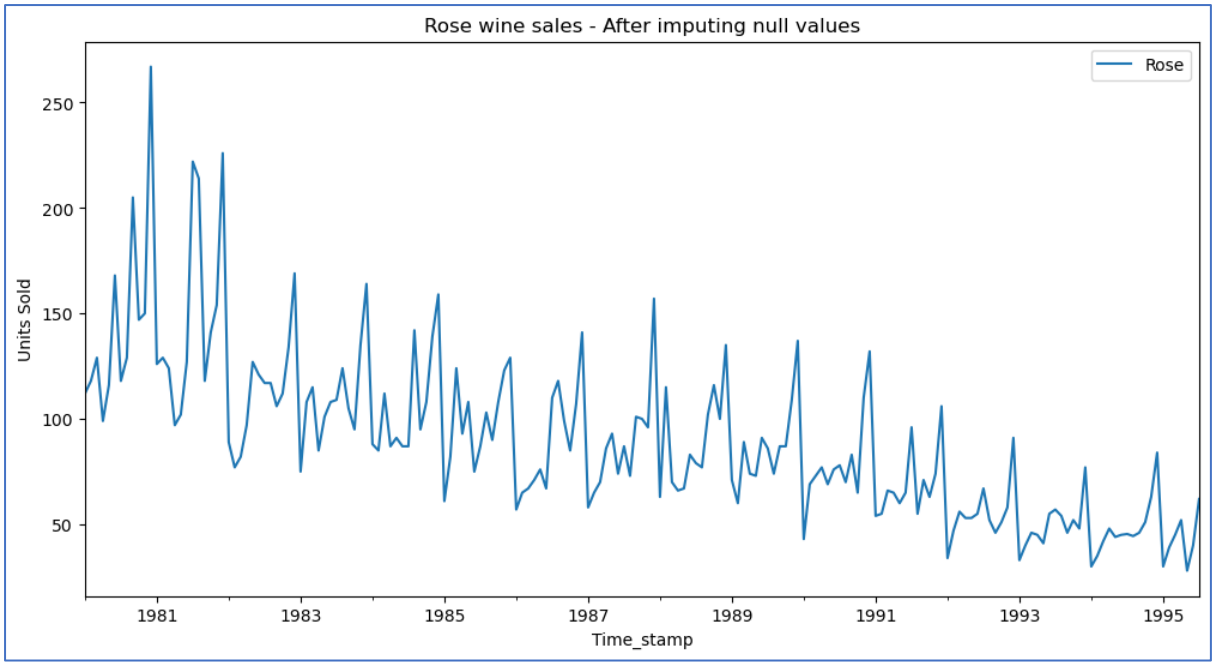


Figure 7 : Rose wine sales after imputing null values.

The plot now has no missing values.

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

- Describe:

	Rose
count	187.000000
mean	89.908161
std	39.245545
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Figure 8 : Rose wine - description

- The mean and median values are nearly the same. This may signify a decreasing trend and multiplicative seasonality.
- It shows that on an average 85 units of rose wine were sold every month in the given time period.
- Maximum of 267 units were sold.

- Yearly boxplot:

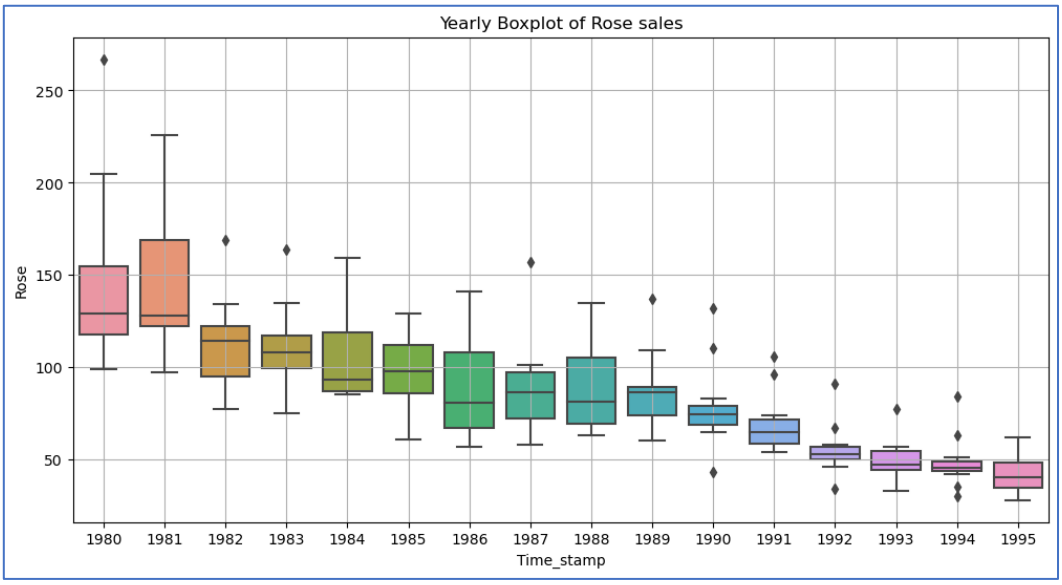


Figure 9 : Yearly boxplot of rose sales

- The yearly boxplot shows a decreasing trend.
- The outliers on the upper bound may represent sales during seasonal month.

- Monthly boxplot:

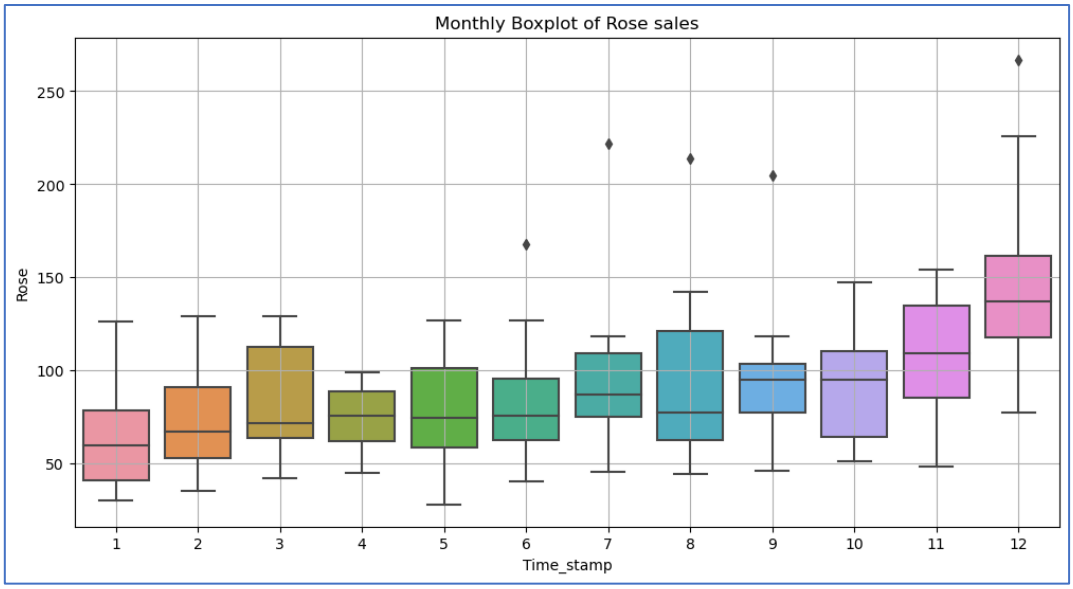


Figure 10 : Monthly boxplot of rose sales

- The monthly boxplot shows seasonality during seasonal months of November and December.
- The sale dips in January and later picks up.

- Monthly plot:

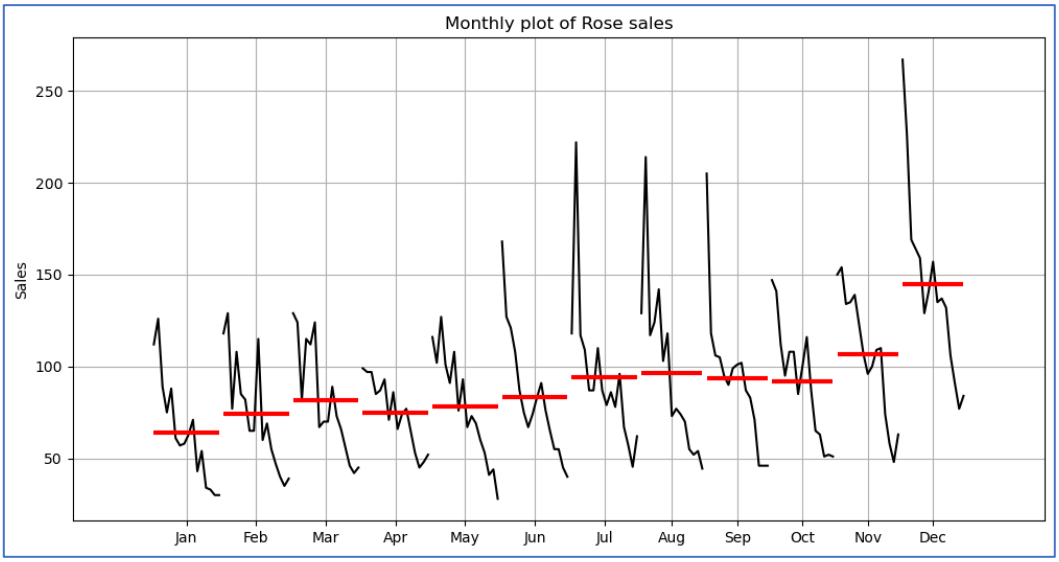


Figure 11 : Monthly of rose sales

- The monthly plot for Rose shows mean and variation of units sold each month over the years.
- Sale in months of July, August, September, and December show a higher variation than the rest.

- Monthly plot: Plot of monthly sales across years.

Time_stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_stamp												
1980	112.000000	118.000000	129.000000	99.000000	116.000000	168.000000	118.000000	129.000000	205.000000	147.000000	150.000000	267.000000
1981	126.000000	129.000000	124.000000	97.000000	102.000000	127.000000	222.000000	214.000000	118.000000	141.000000	154.000000	226.000000
1982	89.000000	77.000000	82.000000	97.000000	127.000000	121.000000	117.000000	117.000000	106.000000	112.000000	134.000000	169.000000
1983	75.000000	108.000000	115.000000	85.000000	101.000000	108.000000	109.000000	124.000000	105.000000	95.000000	135.000000	164.000000
1984	88.000000	85.000000	112.000000	87.000000	91.000000	87.000000	87.000000	142.000000	95.000000	108.000000	139.000000	159.000000
1985	61.000000	82.000000	124.000000	93.000000	108.000000	75.000000	87.000000	103.000000	90.000000	108.000000	123.000000	129.000000
1986	57.000000	65.000000	67.000000	71.000000	76.000000	67.000000	110.000000	118.000000	99.000000	85.000000	107.000000	141.000000
1987	58.000000	65.000000	70.000000	86.000000	93.000000	74.000000	87.000000	73.000000	101.000000	100.000000	96.000000	157.000000
1988	63.000000	115.000000	70.000000	66.000000	67.000000	83.000000	79.000000	77.000000	102.000000	116.000000	100.000000	135.000000
1989	71.000000	60.000000	89.000000	74.000000	73.000000	91.000000	86.000000	74.000000	87.000000	87.000000	109.000000	137.000000
1990	43.000000	69.000000	73.000000	77.000000	69.000000	76.000000	78.000000	70.000000	83.000000	65.000000	110.000000	132.000000
1991	54.000000	55.000000	66.000000	65.000000	60.000000	65.000000	96.000000	55.000000	71.000000	63.000000	74.000000	106.000000
1992	34.000000	47.000000	56.000000	53.000000	53.000000	55.000000	67.000000	52.000000	46.000000	51.000000	58.000000	91.000000
1993	33.000000	40.000000	46.000000	45.000000	41.000000	55.000000	57.000000	54.000000	46.000000	52.000000	48.000000	77.000000
1994	30.000000	35.000000	42.000000	48.000000	44.000000	45.000000	45.406283	44.419745	46.000000	51.000000	63.000000	84.000000
1995	30.000000	39.000000	45.000000	52.000000	28.000000	40.000000	62.000000	nan	nan	nan	nan	nan

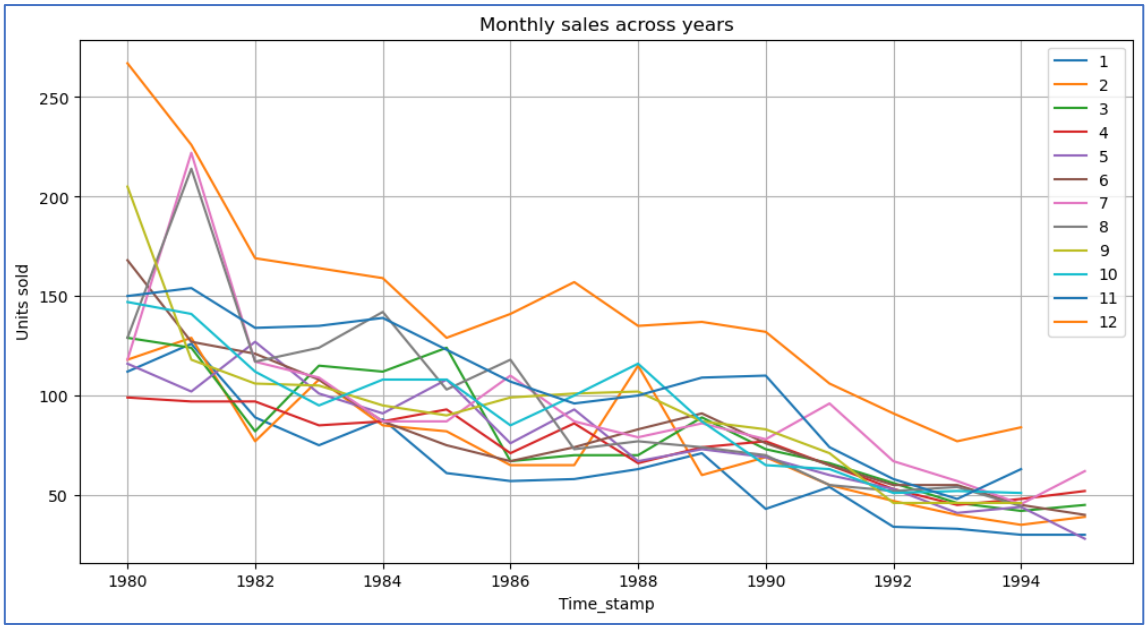


Figure 12 : Monthly sales across years

- A decreasing Trend could be observed for the different months along the Years.

- Plot of average wine sales per month and % change of sales over time

.

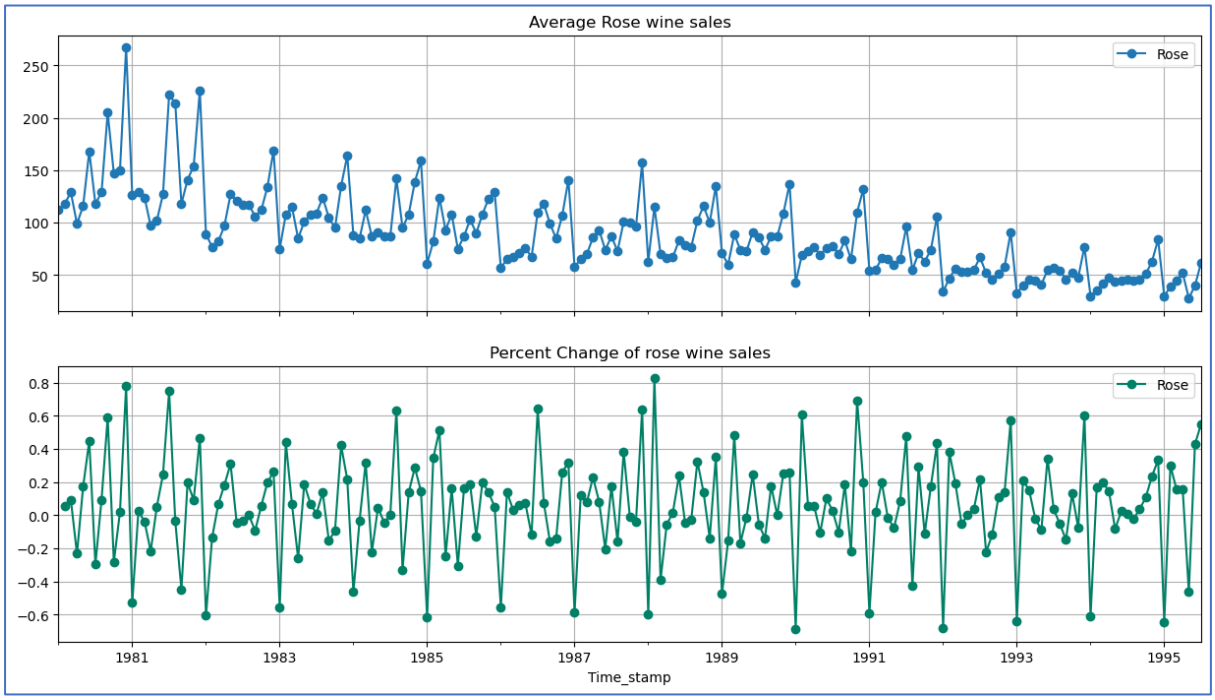


Figure 13 : Plot of average wine sales and % change of sales per month

- Suggests the presence of seasonality.

- Decompose the Time Series and plot the different components.

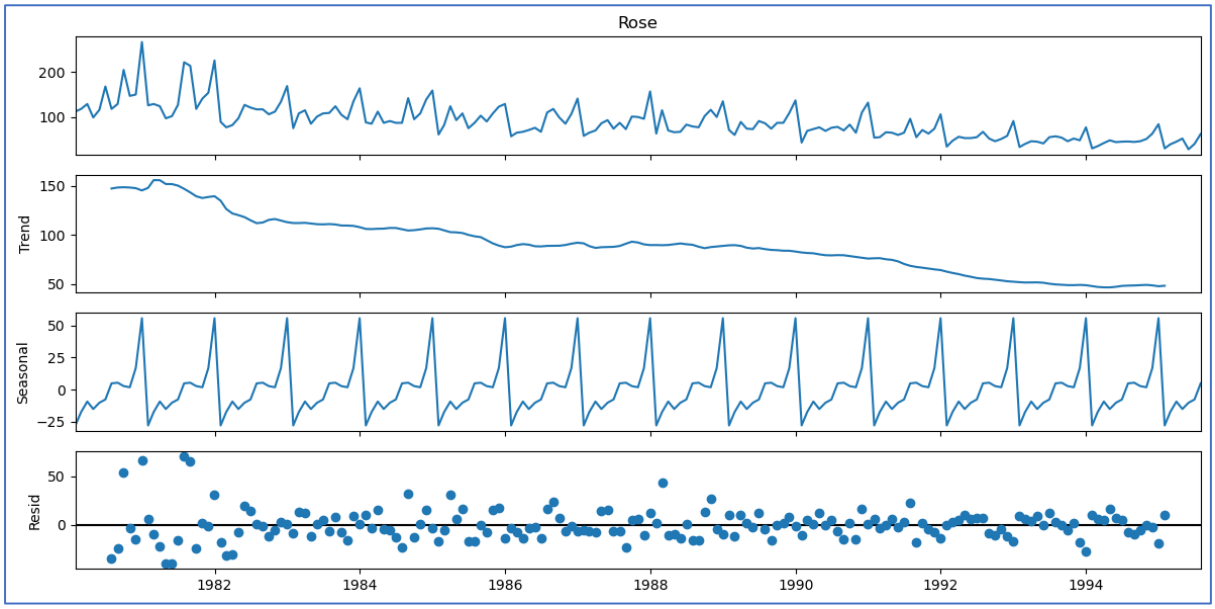


Figure 14 : Decomposition – Additive model

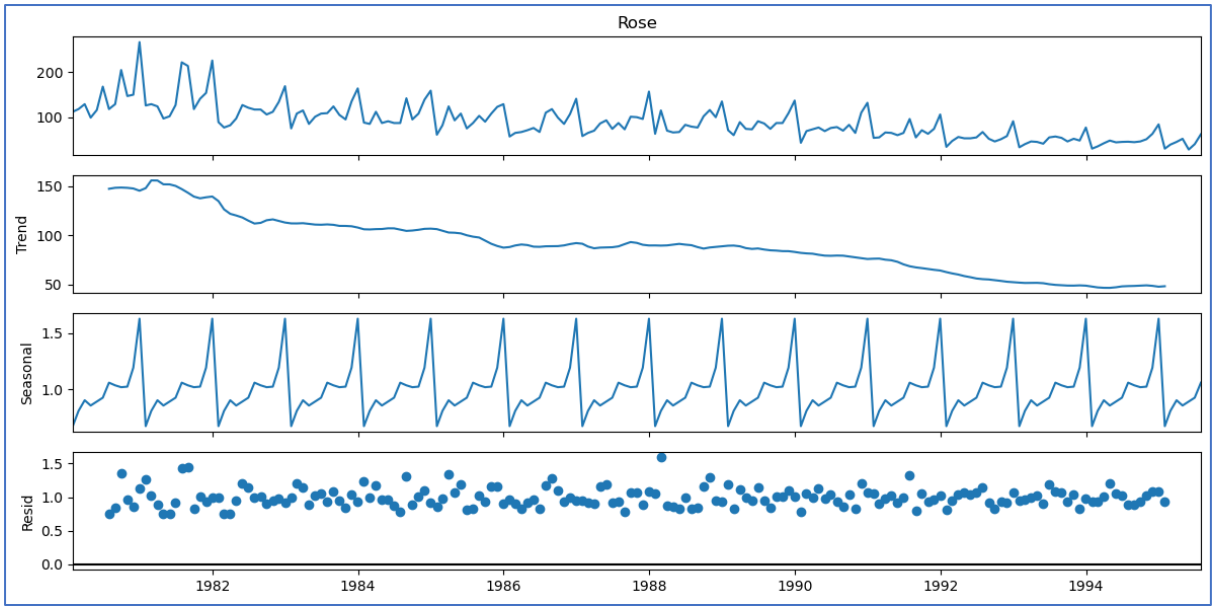


Figure 15 : Decomposition – Multiplicative model

- Shows a decreasing trend and annual seasonality.
- Early period shows higher variation than the later periods.
- The residuals show high variability across time series which is almost consistent.
- The series can be treated as a multiplicative model.

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

After splitting into train and test data, where test data starts from 1991.

```
(132, 1)
(55, 1)
```

Figure 16 : Train-test split

Train has 132 data and test has 55 data.

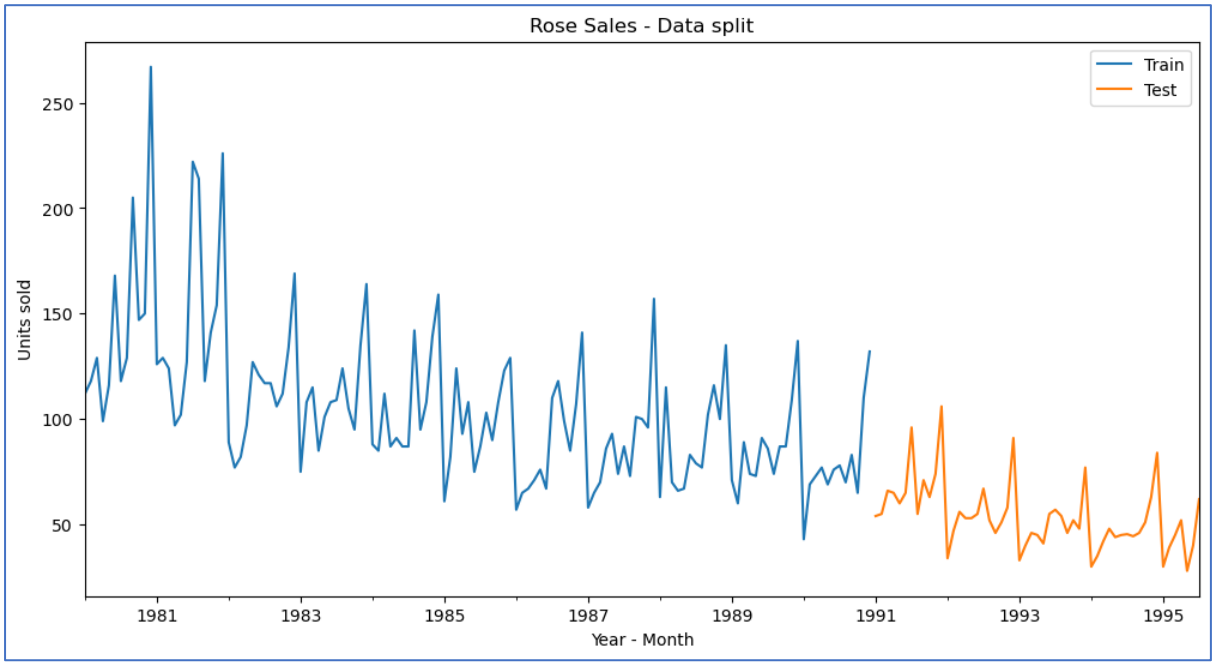


Figure 17 : Rose sales – Data split

1.4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

MODEL 1: Linear Regression

To regress rose wine sales, time instance order for test data were generated and was added to its dataset.

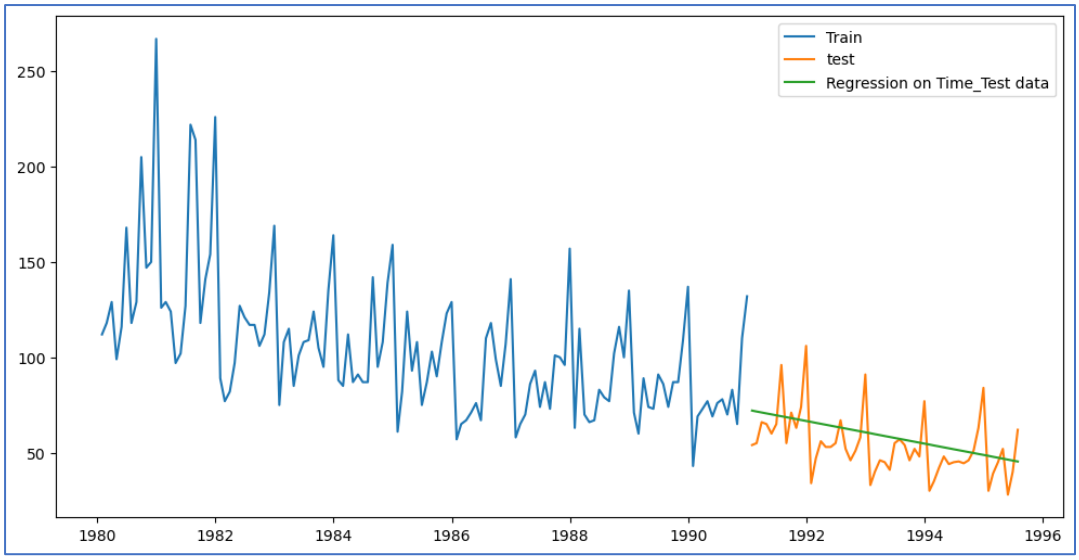


Figure 18 : Linear regression model

- The linear regression shows a downward trend.
- For regression of time series on test data, RMSE = **15.28**
- The model has captured trend but not the seasonality.

MODEL 2: Naïve Forecast

The model has taken last value from test data and fitted it on the rest of the train data and used the same value to forecast the test data.

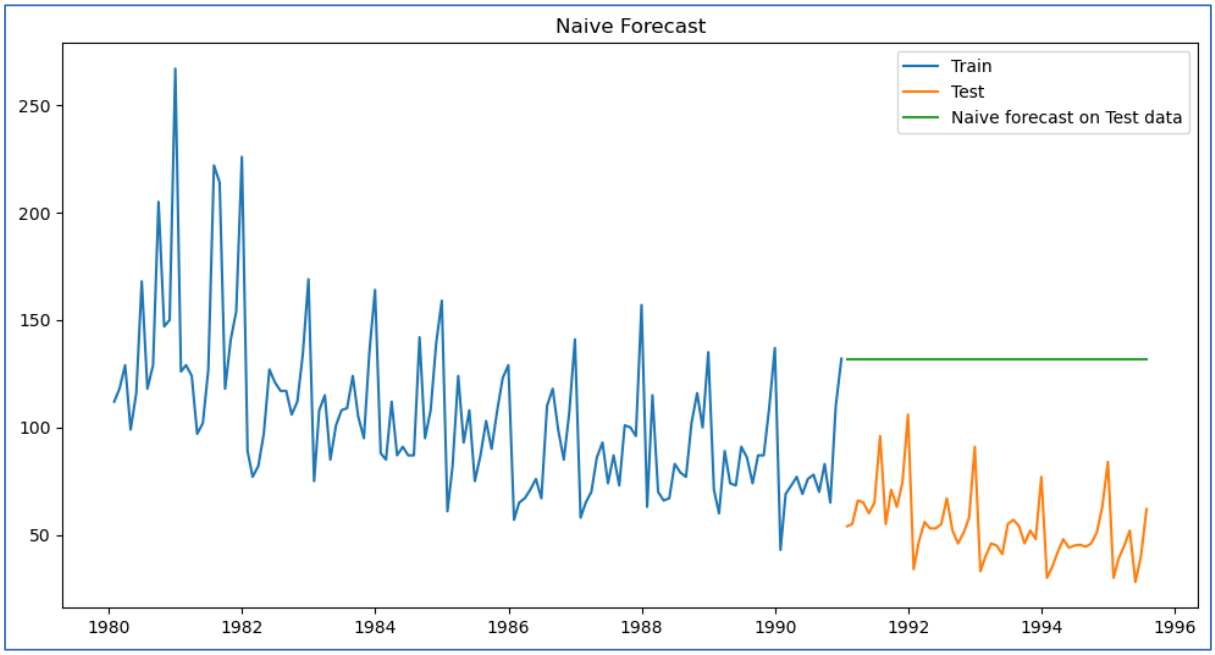


Figure 19 : Naïve forecast model

- For Naïve forecast on test data, RMSE = **79.74**
- The model neither captures trend nor seasonality.

MODEL 3: Simple average model

Here the forecast is done using mean of the time-series variable from train data.

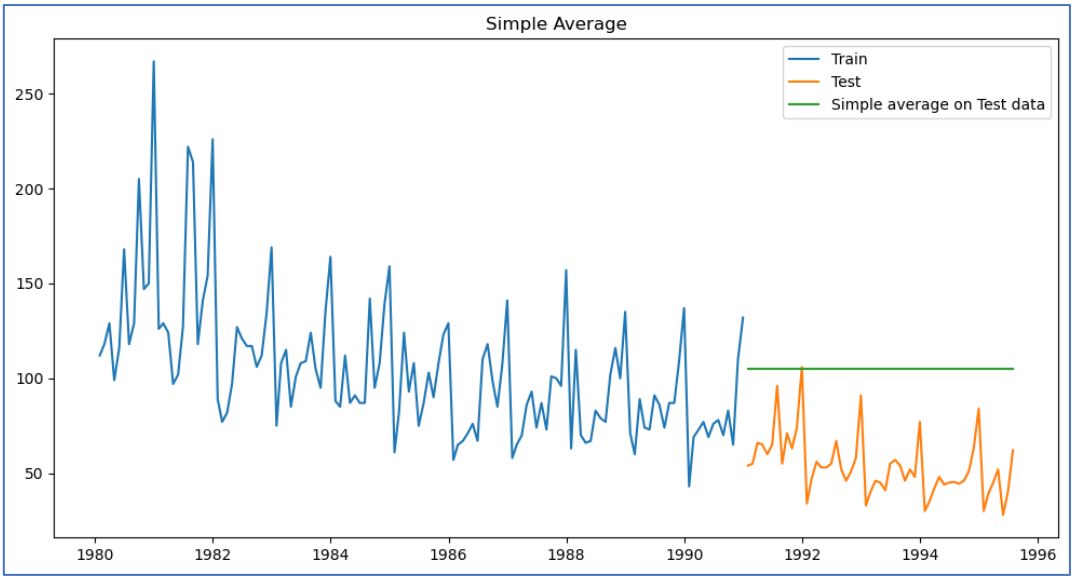


Figure 20 : Simple average model

- The model is unable to forecast.
- It's also unable to capture trend and seasonality.
- For simple average on test data, RMSE = **53.48**

<u>MODEL 4: Moving average model.</u>

In this model we calculate rolling mean for different intervals. The one with maximum accuracy is the best interval.

Here the moving average is built on 2, 4, 6, 9 points

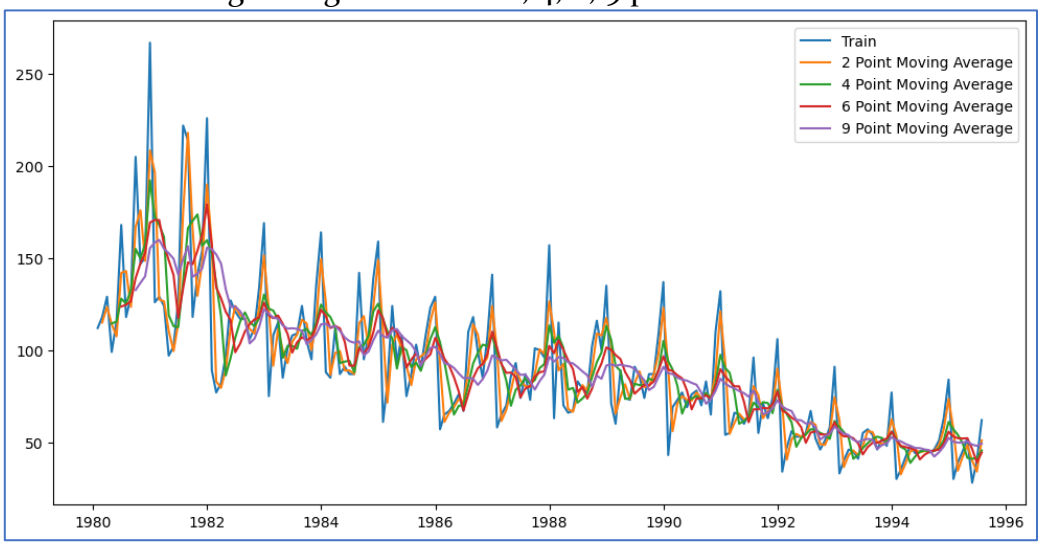


Figure 21 : Moving average model

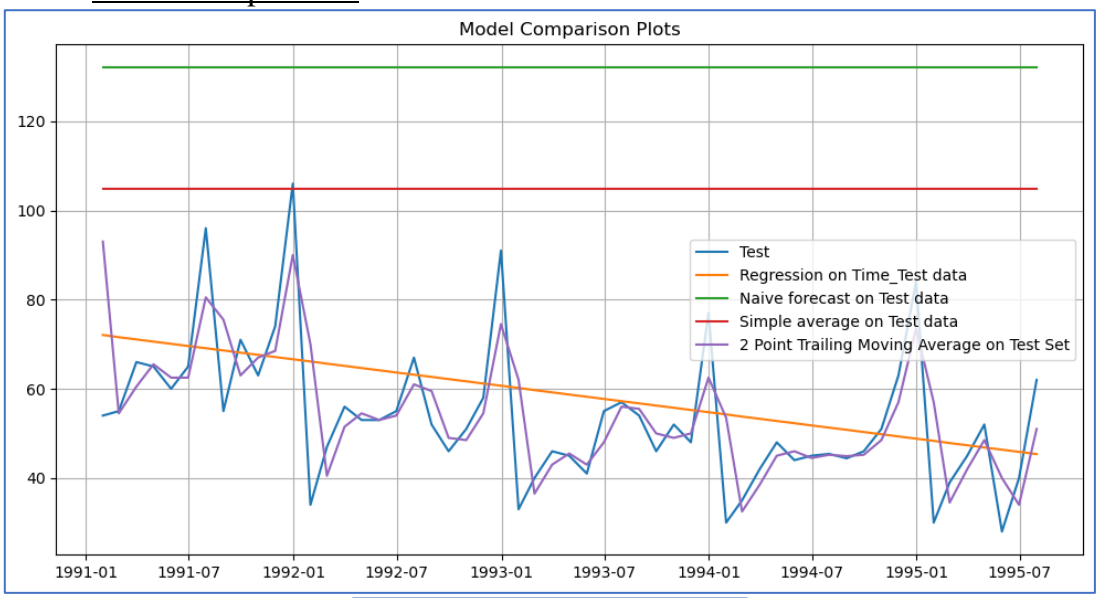
- Here accuracy is higher for lower rolling point averages.

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.53
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.46
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.57
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.73
```

Figure 22 : RMSE of rolling point averages

- The best interval of moving average from model is 2 point.

Model comparison:



	Test RMSE
RegressionOnTime	15.276993
Naive Model	79.742058
Simple Average model	53.484487
2 point TMA	11.529929
4 point TMA	14.457533
6 point TMA	14.571867
9 point TMA	14.732128

Figure 23 : Model Comparison

MODEL 5: Simple exponential smoothing

Simple exponential smoothing is applied when the time-series has neither trend nor seasonality.

The alpha value was closer to 1, forecasts follow actual observation closely and closer to 0, forecasts are farther from actual line and gets smoothened.

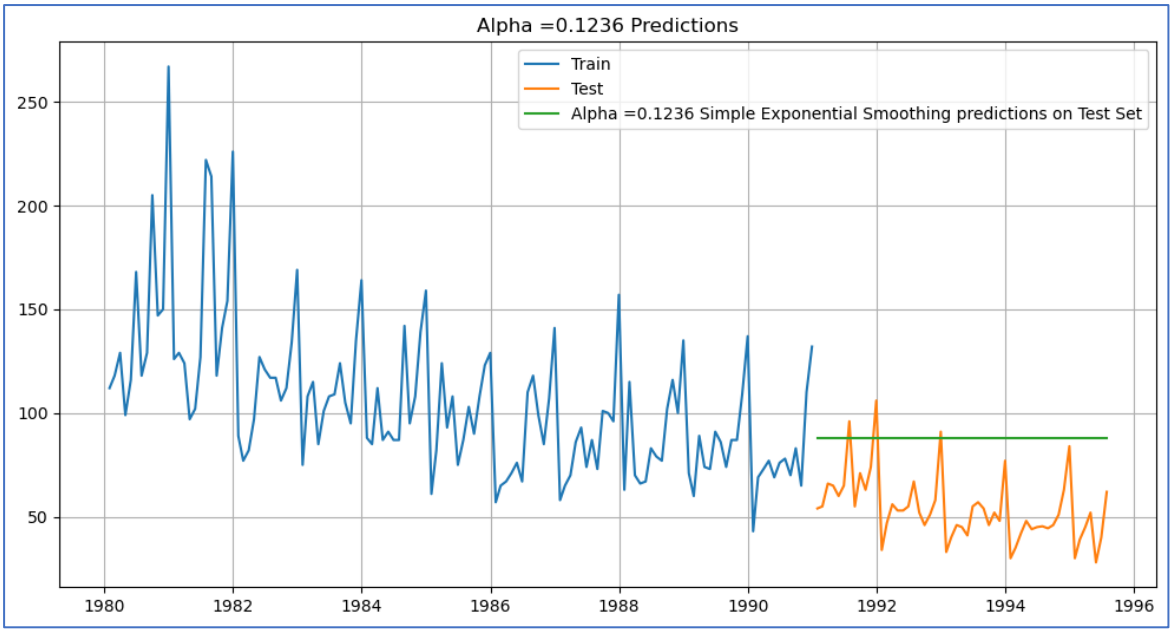


Figure 24 : SES – manual method

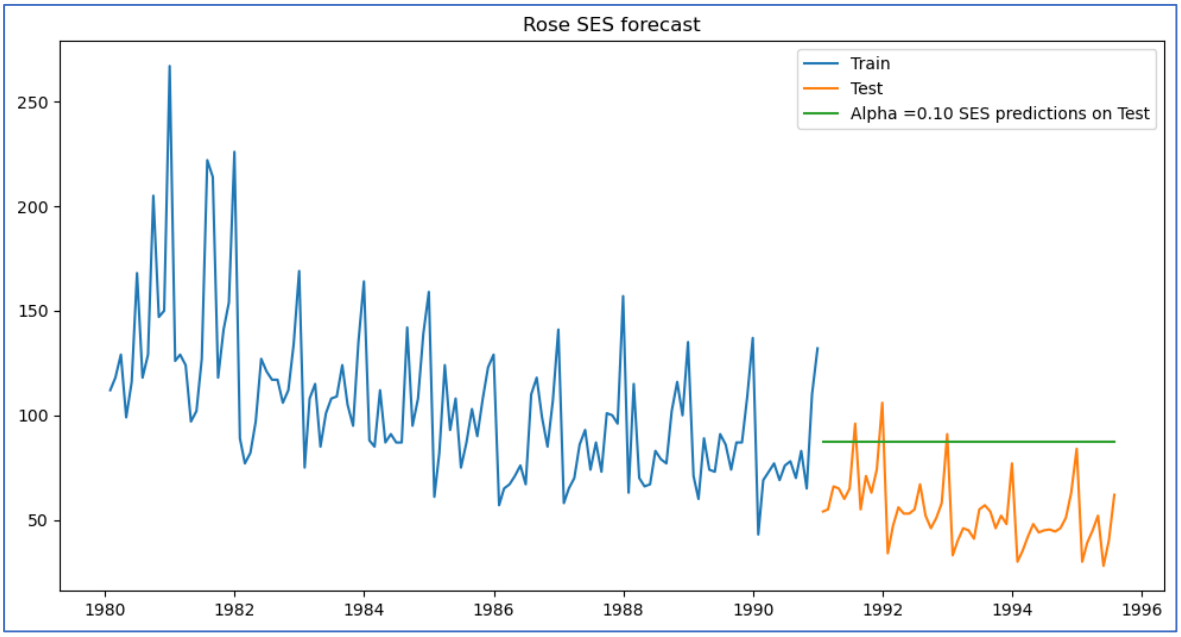


Figure 25 : SES – iterative method

- For both manual and iterative method, RMSE were almost similar.

Alpha=0.1236 ,SimpleExponentialSmoothing	37.616598
Alpha=0.10,SES_Iterative	36.852435

Figure 26 : SES – RMSE

<u>MODEL 6: Double exponential smoothing (DES)</u>

DES is applied when the data has trend but no seasonality.

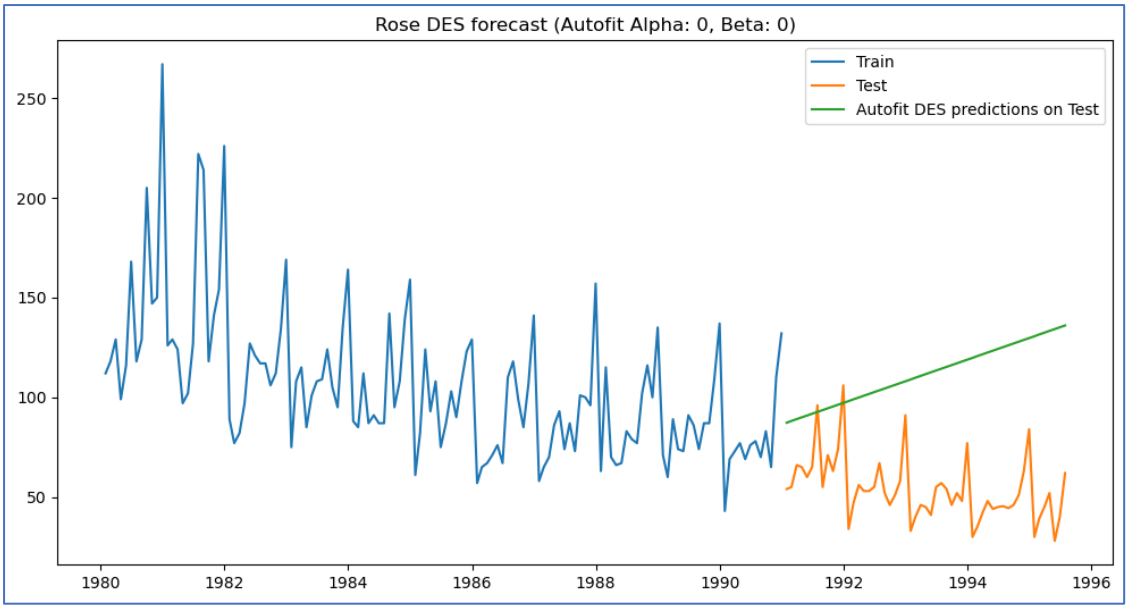


Figure 27 : DES – Optimized method

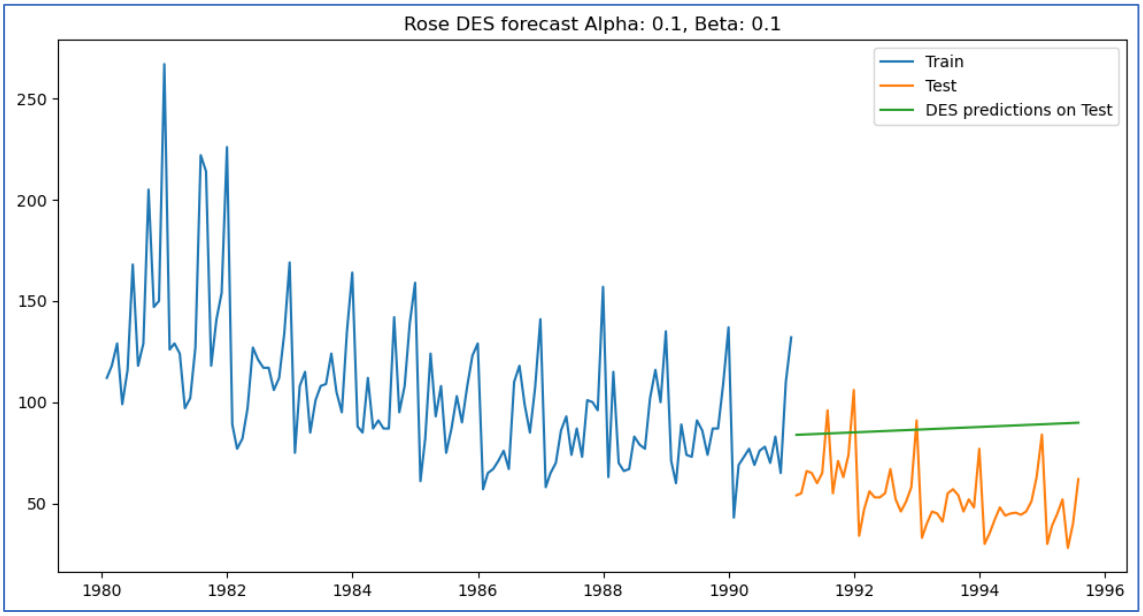


Figure 28 : DES –Iterative method

- The iterative method has lower RMSE then optimised method.

Alpha=0.0,Beta=0.0, DES Optimized	63.070429
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000

Figure 29 : DES –RMSE

MODEL 7: Triple exponential smoothing (TES)

TES is applicable when the data has both trend and seasonality.

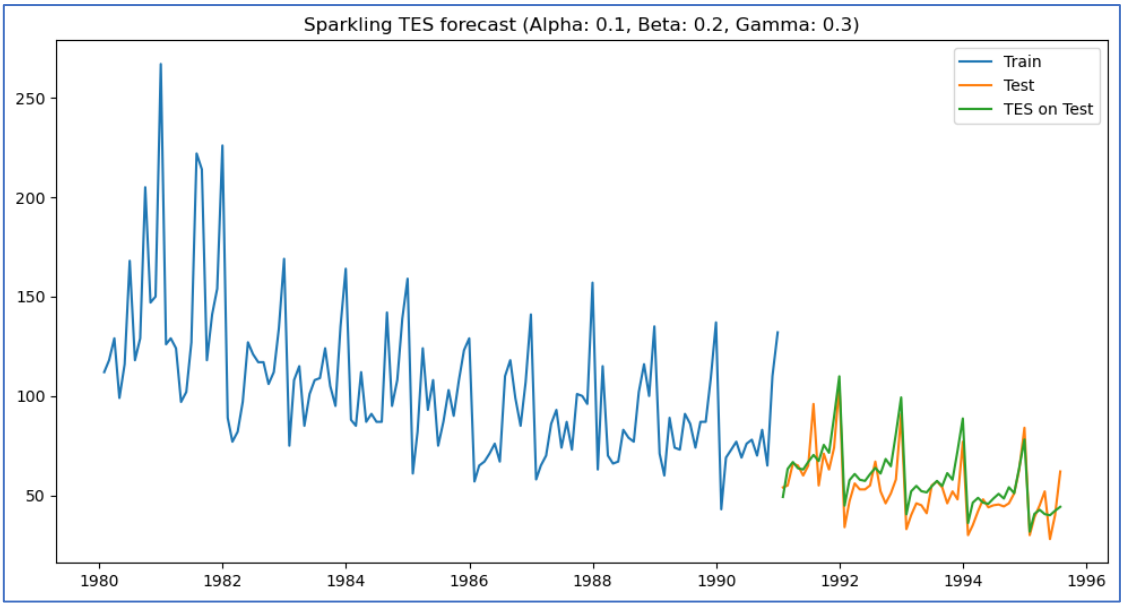


Figure 30 : TES –Iterative method

- For iterative TES on test data, RMSE = **9.88**

MODEL COMPARISONS:

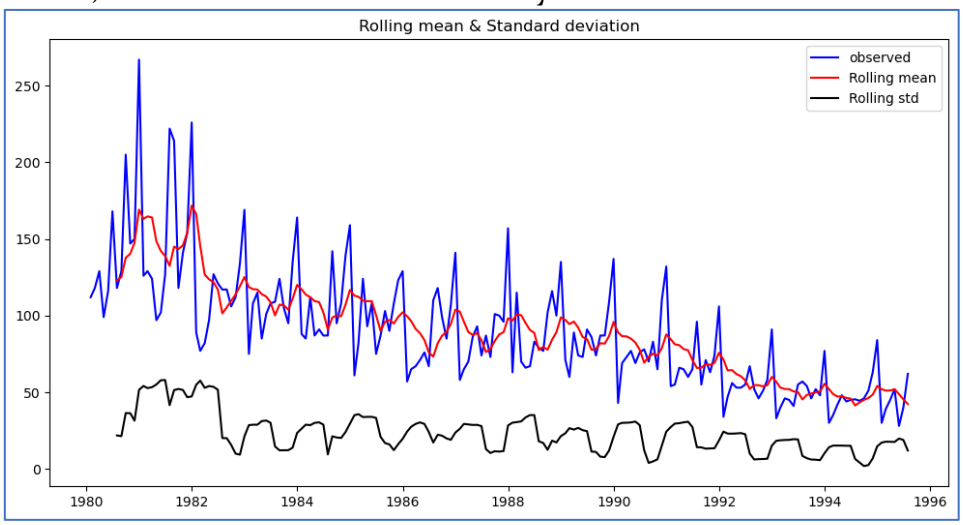
	Test RMSE
RegressionOnTime	15.276993
Naive Model	79.742058
Simple Average model	53.484487
2 point TMA	11.529929
4 point TMA	14.457533
6 point TMA	14.571867
9 point TMA	14.732128
Alpha=0.1236 ,SimpleExponentialSmoothing	37.616598
Alpha=0.10,SES_Iterative	36.852435
Alpha=0.0,Beta=0.0, DES Optimized	63.070429
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
Alpha=0.1,Beta=0.2,gamma=0.3,TES_Iterative	9.881358

Figure 31 : Model comparison

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

Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series.
Null Hypothesis: The series has a unit root, that is series is non-stationary.
Alternate Hypothesis: The series has no unit root, that is series is stationary.

If we fail to reject the null hypothesis, the series is non-stationary and if we accept the null hypothesis, the series is said to be stationary.

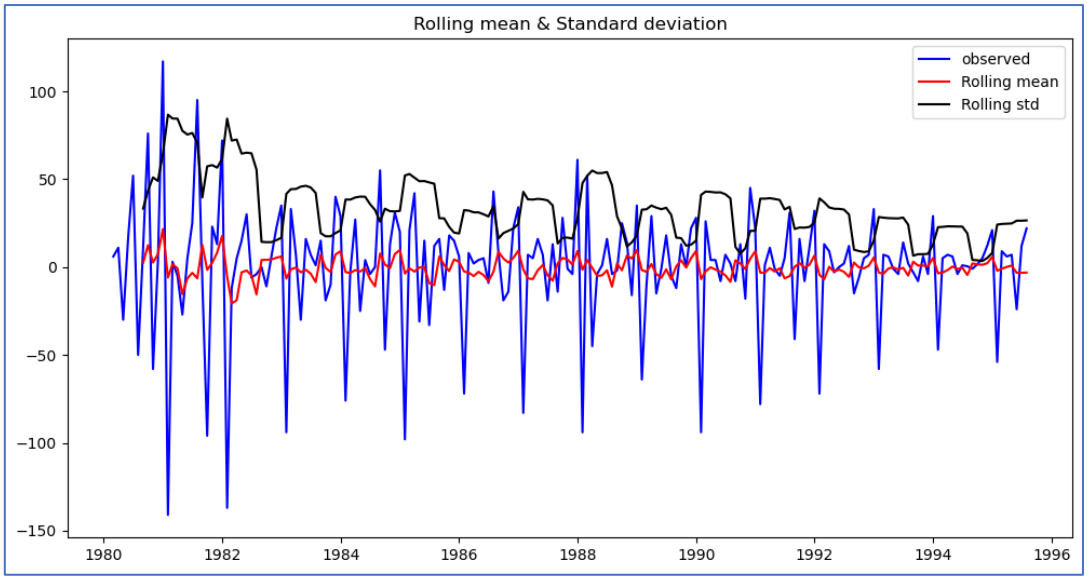



```
Result of Dickey-Fuller test:
Test Statistic                  -1.873015
p-value                          0.344860
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

Figure 32 : Dickey-Fuller test

Here the p value is greater than 0.05, hence we fail to reject null hypothesis and the series is non-stationary.

Differencing of order 1 is applied and again tested for stationarity.
For this rolling mean and standard deviation is also plotted to understand seasonality.



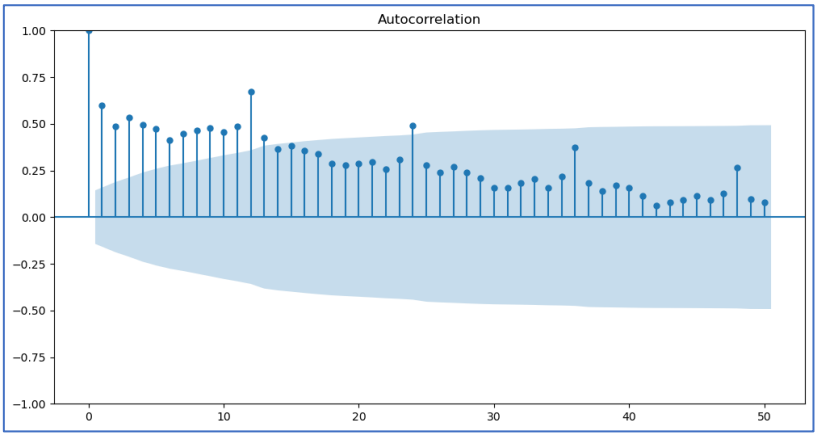

```
Result of Dickey-Fuller test:
Test Statistic                  -8.044130e+00
p-value                          1.813672e-12
#Lags Used                       1.200000e+01
Number of Observations Used      1.730000e+02
Critical Value (1%)             -3.468726e+00
Critical Value (5%)             -2.878396e+00
Critical Value (10%)            -2.575756e+00
dtype: float64
```

Figure 33 : Dickey-Fuller test with d=1

Here the p-value is less than 0.05, hence the series has now become stationary.

Auto-correlation function plots

ACF-:



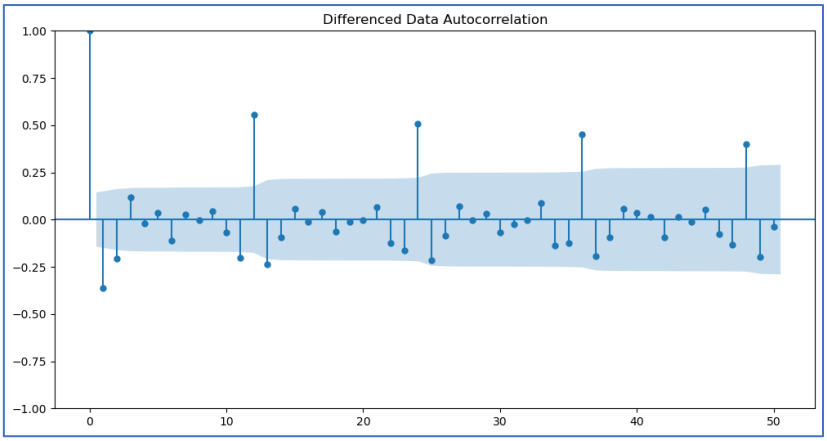
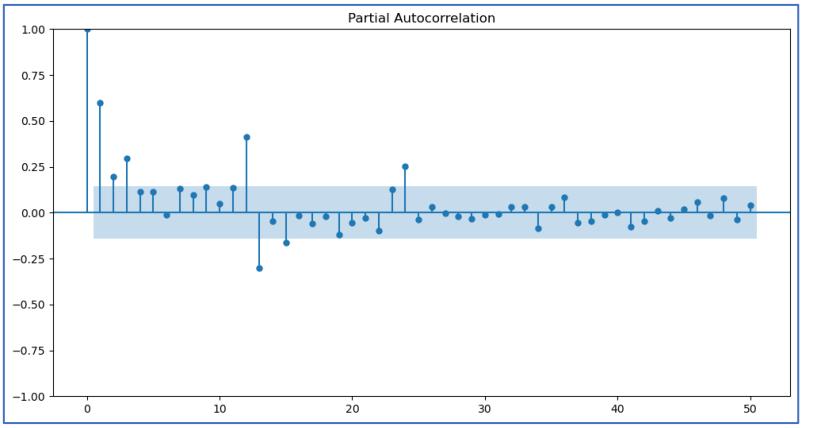


Figure 34 : Auto correlation function

PACF-:



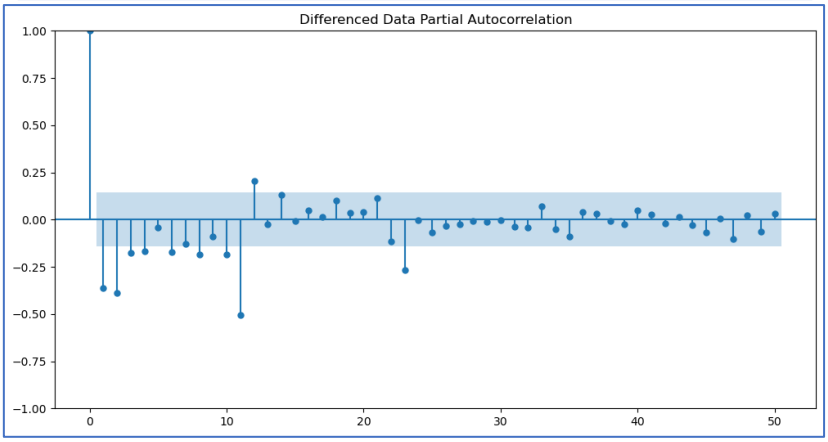


Figure 35 : Partial auto correlation function

The above plots indicate the presence of seasonality.

1.6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Auto-ARIME:

An ARIMA model was built with optimised model and found the least AIC value = 1276.9 at (0,1,2)

	param	AIC
2	(0, 1, 2)	1279.671529
5	(1, 1, 2)	1279.870723
4	(1, 1, 1)	1280.574230
7	(2, 1, 1)	1281.507862
8	(2, 1, 2)	1281.870722
1	(0, 1, 1)	1282.309832
6	(2, 1, 0)	1298.611034
3	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(0, 1, 2)	Log Likelihood	-636.836
Date:	Thu, 07 Dec 2023	AIC	1279.672
Time:	16:19:43	BIC	1288.297
Sample:	01-31-1980	HQIC	1283.176
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.6970	0.072	-9.689	0.000	-0.838	-0.556
ma.L2	-0.2042	0.073	-2.794	0.005	-0.347	-0.061
sigma2	965.8407	88.305	10.938	0.000	792.766	1138.915

Ljung-Box (L1) (Q):	0.14	Jarque-Bera (JB):	39.24
Prob(Q):	0.71	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	0.82
Prob(H) (two-sided):	0.00	Kurtosis:	5.13

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

Figure 36 : ARIMA

- The RMSE was found to be 37.33

<u>Auto-SARIMA:</u>

The model was built on train data with seasonality 12 and with different optimal parameters (p, d, q)x(P, D, Q) parameters.
The lowest AIC is 774.96 was obtained at (0, 1, 2)x(2, 1, 2, 12).

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-380.485
Date:	Thu, 07 Dec 2023	AIC	774.969
Time:	16:38:44	BIC	792.622
Sample:	0 - 132	HQIC	782.094
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9524	0.184	-5.166	0.000	-1.314	-0.591
ma.L2	-0.0764	0.126	-0.605	0.545	-0.324	0.171
ar.S.L12	0.0480	0.177	0.271	0.786	-0.299	0.395
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	187.8664	45.275	4.149	0.000	99.130	276.603

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

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

Figure 37 : SARIMA

- The RMSE was found to be 16.52

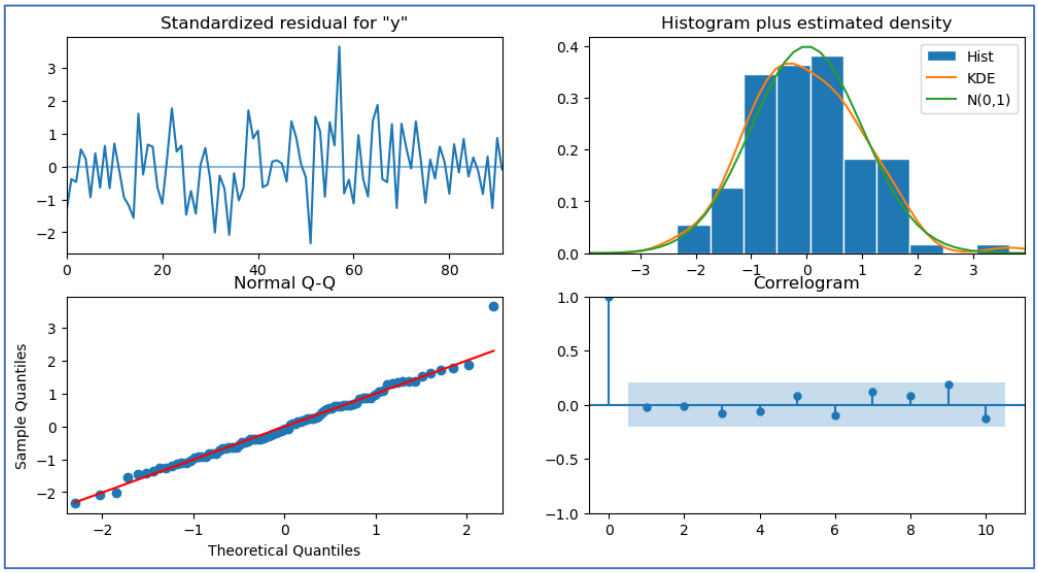


Figure 38 : Diagnostics plot

- The residuals are found to follow a mean of zero
- The histogram shows that residuals follow normal distribution.
- The normal Q-Q plot shows that the the quantiles come from a normal distribution as they almost form a straight line.
- The correlogram shows autocorrelation of residuals and there is no lag above the confidence limit.

PREDICTING ON TEST DATA:

Time_stamp	Rose	rose_forecasted
1991-01-31	54.0	44.213906
1991-02-28	55.0	62.326915
1991-03-31	66.0	67.313577
1991-04-30	65.0	63.161196
1991-05-31	60.0	66.474432

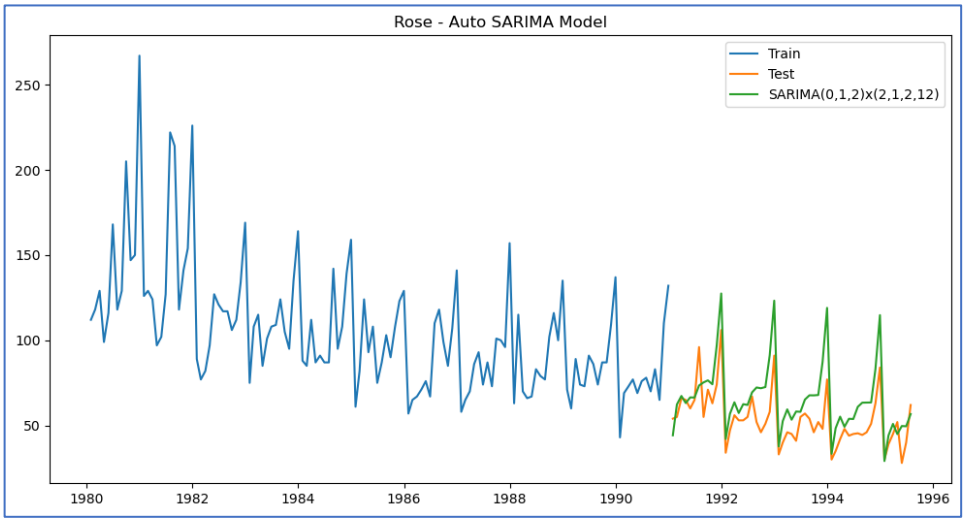


Figure 39 : Prediction on test data

- The model built with log-series data has a higher RMSE than the original train data.

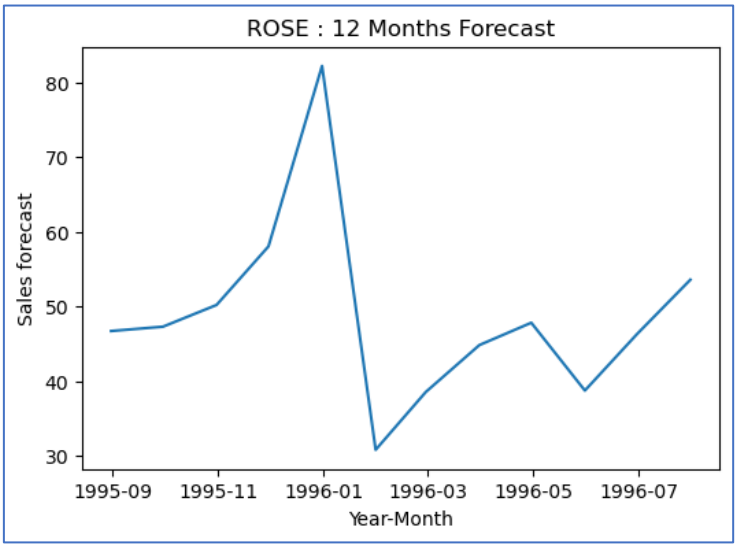
1.7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

	Test RMSE
RegressionOnTime	15.276993
Naive Model	79.742058
Simple Average model	53.484487
2 point TMA	11.529929
4 point TMA	14.457533
6 point TMA	14.571867
9 point TMA	14.732128
Alpha=0.1236 ,SimpleExponentialSmoothing	37.616598
Alpha=0.10,SES_Iterative	36.852435
Alpha=0.0,Beta=0.0, DES Optimized	63.070429
Alpha=0.1,Beta=0.1,DES_Iterative	36.950000
Alpha=0.1,Beta=0.2,gamma=0.3,TES_Iterative	9.881358
Auto_ARIMA(0, 1, 2)	37.330772
Auto_SARIMA(0, 1, 2)*(2, 1, 2, 12)	16.524205

Figure 40 : RMSE of all models on the test data

1.8. 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.

- Based on the overall model evaluation, triple exponential smoothing (Holt Winter's method) is selected for prediction as it has the lowest RMSE.
- TES model alpha: 0.1, beta: 0.2 and gamma: 0.3 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model.



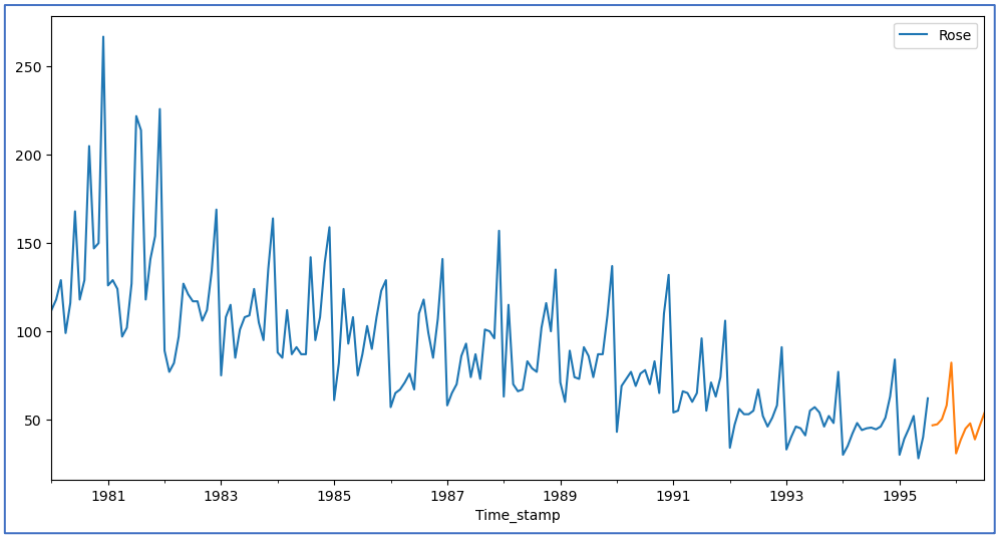


Figure 41 : 12 months forecast.

- On calculating upper and lower confidence bands at 95% confidence level, we are taking the multiplier as 1.96 as we want the plot with 95% confidence limit.

	lower_CI	prediction	upper_ci
1995-08-31	11.579038	46.697791	81.816544
1995-09-30	12.159706	47.278459	82.397212
1995-10-31	15.072439	50.191192	85.309945
1995-11-30	22.910861	58.029614	93.148367
1995-12-31	47.087920	82.206673	117.325426
1996-01-31	-4.327787	30.790966	65.909718
1996-02-29	3.413754	38.532507	73.651260
1996-03-31	9.698853	44.817606	79.936359
1996-04-30	12.689783	47.808536	82.927289
1996-05-31	3.605612	38.724365	73.843118
1996-06-30	11.130475	46.249227	81.367980
1996-07-31	18.450105	53.568858	88.687611

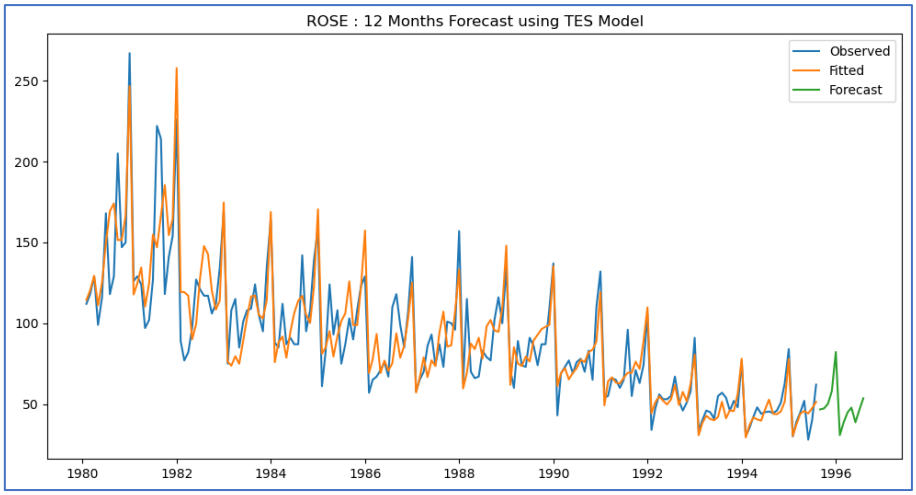


Figure 42 : 12 months forecast using TES model

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

	lower_CI	prediction	upper_ci
count	12.000000	12.000000	12.000000
mean	13.622563	48.741316	83.860069
std	12.746394	12.746394	12.746394
min	-4.327787	30.790966	65.909718
25%	8.175543	43.294296	78.413049
50%	11.869372	46.988125	82.106878
75%	15.916856	51.035609	86.154362
max	47.087920	82.206673	117.325426

	lower_CI	prediction	upper_ci
1995-08-31	11.579038	46.697791	81.816544
1995-09-30	12.159706	47.278459	82.397212
1995-10-31	15.072439	50.191192	85.309945
1995-11-30	22.910861	58.029614	93.148367
1995-12-31	47.087920	82.206673	117.325426
1996-01-31	-4.327787	30.790966	65.909718
1996-02-29	3.413754	38.532507	73.651260
1996-03-31	9.698853	44.817606	79.936359
1996-04-30	12.689783	47.808536	82.927289
1996-05-31	3.605612	38.724365	73.843118
1996-06-30	11.130475	46.249227	81.367980
1996-07-31	18.450105	53.568858	88.687611

Figure 43 : 12 months forecast data and description

- The model forecasts 584 units of rose wine in next 12 months.

- Average of 48 units per month.
- Maximum of 82 units will be sold in December 1995.
- Least units will be sold in January 1996.
- The ABC estate must look into it and take necessary actions in promoting the product.
- They must invest in marketing their wine and promoting them.

Problem 2: Sparkling wine dataset

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

- Data:

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471
...	...	...
182	1995-03	1897
183	1995-04	1862
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

187 rows × 2 columns

Figure 44 : Sparkling wine data without date-time format

- The data set contains 2 columns and 187 rows of data.
- The time stamp is from Jan 1980 to July 1995.

- Time Series data:
  Time stamps were added to the data frame to make it a time series data and was set as index. The YearMonth column has been removed.

Time_stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Figure 45 : Sparkling wine - Timeseries data

- Datatype:

```
Sparkling    int64
dtype: object
```

Figure 46 : Sparkling wine data type

- Null values: There are no null values in the data.
- Plotting the timeseries data:

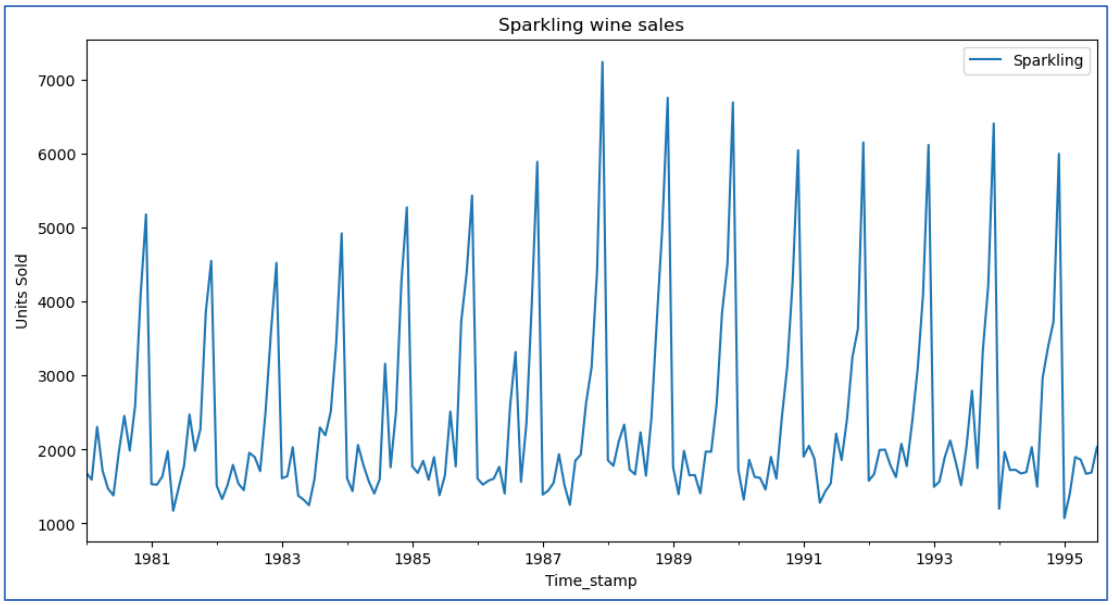


Figure 47 : Sparkling wine sales

- The plot shows no significant trend but shows high seasonality.

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

- Describe:

	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

Figure 48 : Sparkling wine description

- It shows that on an average 2402 units of sparkling wine were sold every month in the given time period.
- Maximum of 7242 units were sold.
- Minimum of 1070 units were sold.

- Yearly boxplot:

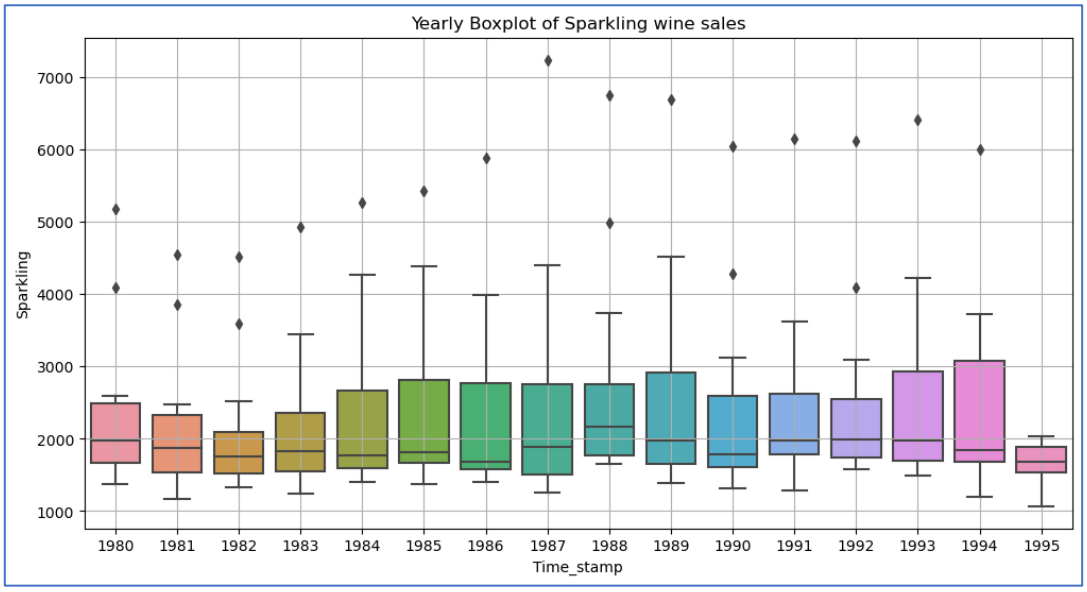


Figure 49 : Yearly boxplot of sparkling wine

- The yearly boxplot shows that sparkling wine has been consistent across the given time period.
- The outliers on the upper bound may represent sales during seasonal months.

- Monthly boxplot:

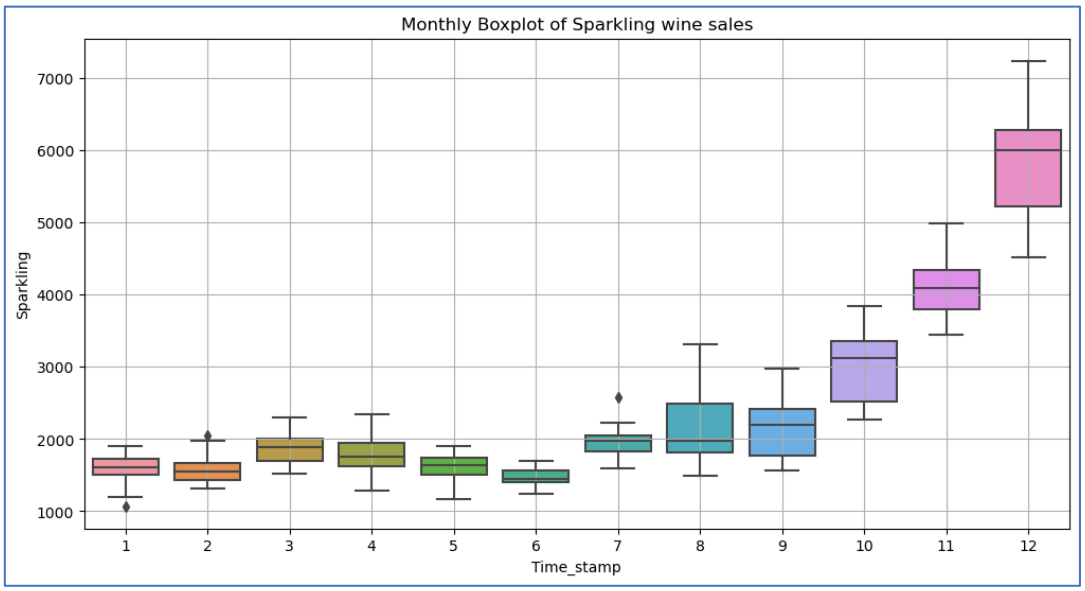


Figure 50 : Monthly boxplot of sparkling wine

- The monthly boxplot shows seasonality during seasonal months of October, November, and December.
- The sale starts to increase from the month of July.

- Monthly plot:

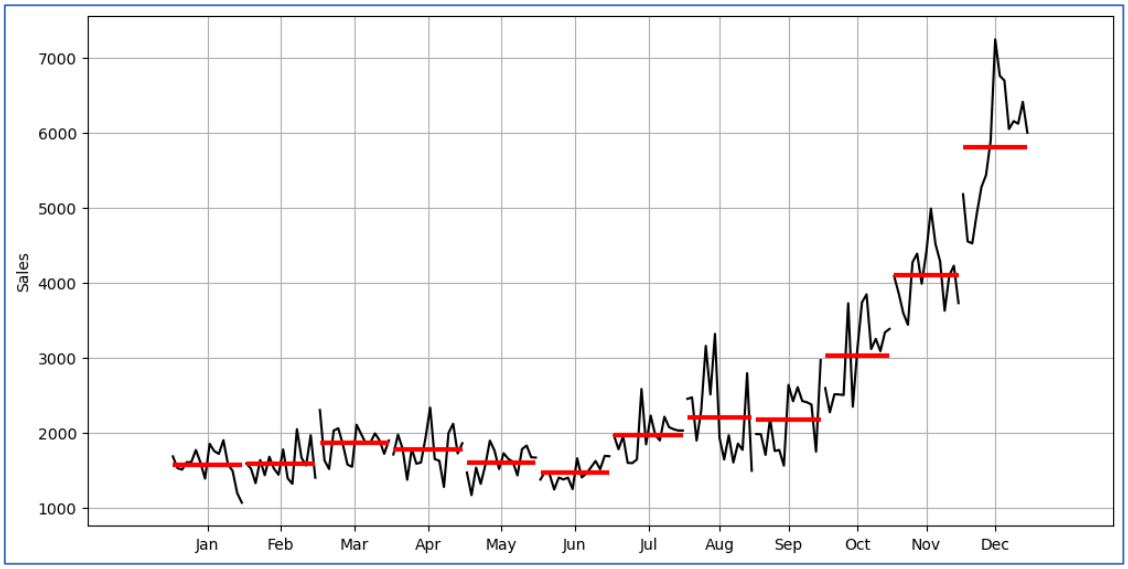


Figure 51 : Monthly plot of sparkling wine

- The monthly plot for sparkling wine shows mean and variation of units sold each month over the years.
- Sales in seasonal months show higher variation than the lean months.
- January to September shows a consistent sale more or less.

- Monthly plot: Plot of monthly sales across years.

mp	1	2	3	4	5	6	7	8	9	10	11	1
mp												
980	1686.000000	1591.000000	2304.000000	1712.000000	1471.000000	1377.000000	1966.000000	2453.000000	1984.000000	2596.000000	4087.000000	5179.00000
981	1530.000000	1523.000000	1633.000000	1976.000000	1170.000000	1480.000000	1781.000000	2472.000000	1981.000000	2273.000000	3857.000000	4551.00000
982	1510.000000	1329.000000	1518.000000	1790.000000	1537.000000	1449.000000	1954.000000	1897.000000	1706.000000	2514.000000	3593.000000	4524.00000
983	1609.000000	1638.000000	2030.000000	1375.000000	1320.000000	1245.000000	1600.000000	2298.000000	2191.000000	2511.000000	3440.000000	4923.00000
984	1609.000000	1435.000000	2061.000000	1789.000000	1567.000000	1404.000000	1597.000000	3159.000000	1759.000000	2504.000000	4273.000000	5274.00000
985	1771.000000	1682.000000	1846.000000	1589.000000	1896.000000	1379.000000	1645.000000	2512.000000	1771.000000	3727.000000	4388.000000	5434.00000
986	1606.000000	1523.000000	1577.000000	1605.000000	1765.000000	1403.000000	2584.000000	3318.000000	1562.000000	2349.000000	3987.000000	5891.00000
987	1389.000000	1442.000000	1548.000000	1935.000000	1518.000000	1250.000000	1847.000000	1930.000000	2638.000000	3114.000000	4405.000000	7242.00000
988	1853.000000	1779.000000	2108.000000	2336.000000	1728.000000	1661.000000	2230.000000	1645.000000	2421.000000	3740.000000	4988.000000	6757.00000
989	1757.000000	1394.000000	1982.000000	1650.000000	1654.000000	1406.000000	1971.000000	1968.000000	2608.000000	3845.000000	4514.000000	6694.00000
990	1720.000000	1321.000000	1859.000000	1628.000000	1615.000000	1457.000000	1899.000000	1605.000000	2424.000000	3116.000000	4286.000000	6047.00000
991	1902.000000	2049.000000	1874.000000	1279.000000	1432.000000	1540.000000	2214.000000	1857.000000	2408.000000	3252.000000	3627.000000	6153.00000
992	1577.000000	1667.000000	1993.000000	1997.000000	1783.000000	1625.000000	2076.000000	1773.000000	2377.000000	3088.000000	4096.000000	6119.00000
993	1494.000000	1564.000000	1898.000000	2121.000000	1831.000000	1515.000000	2048.000000	2795.000000	1749.000000	3339.000000	4227.000000	6410.00000
994	1197.000000	1968.000000	1720.000000	1725.000000	1674.000000	1693.000000	2031.000000	1495.000000	2968.000000	3385.000000	3729.000000	5999.00000
995	1070.000000	1402.000000	1897.000000	1862.000000	1670.000000	1688.000000	2031.000000	nan	nan	nan	nan	na

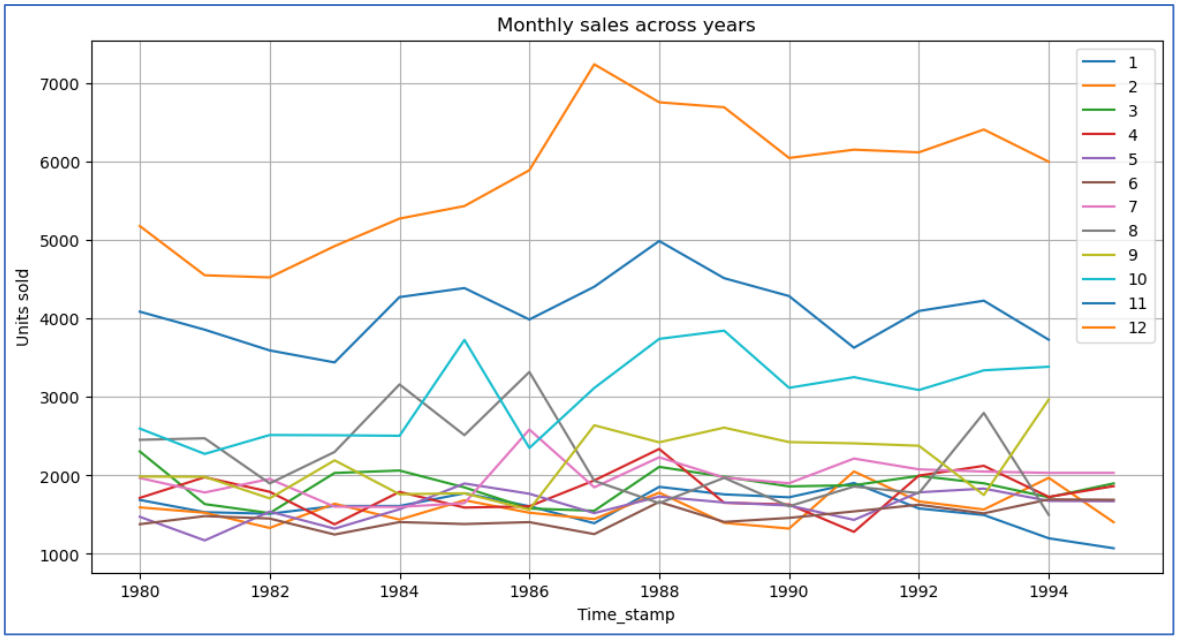


Figure 52 : Monthly sales across years

- Plot of monthly sales shows seasonality.
- October, November, December sells exponentially higher volume.
- Highest volume was sold in December 1987 and least in December 1981.
- Sales from January to July is seen to be consistent across years.

- Plot of average wine sales per month and % change of sales over time

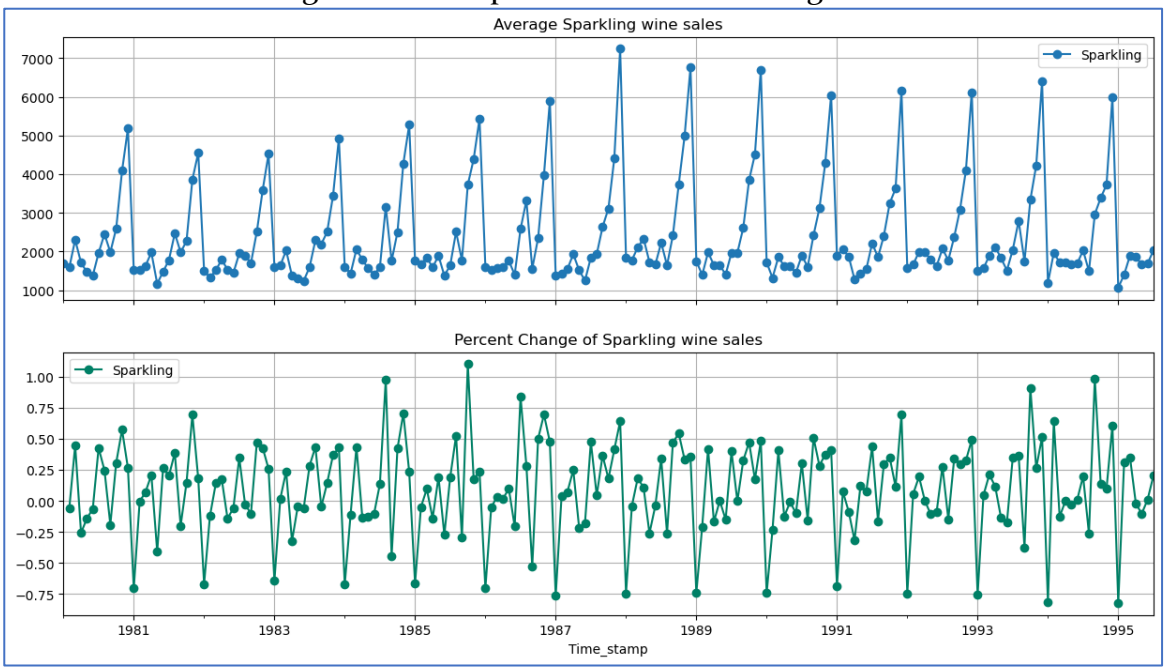


Figure 53 : Average and change of sales per month

- Suggests the presence of seasonality.

- Decompose the Time Series and plot the different components.

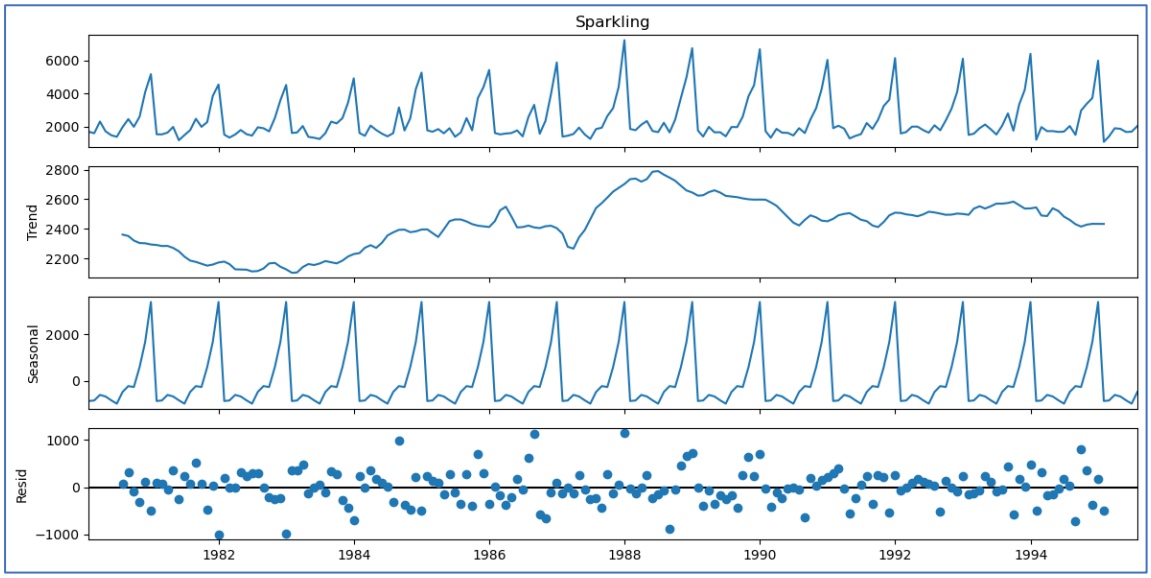


Figure 54 : Decomposition by additive method

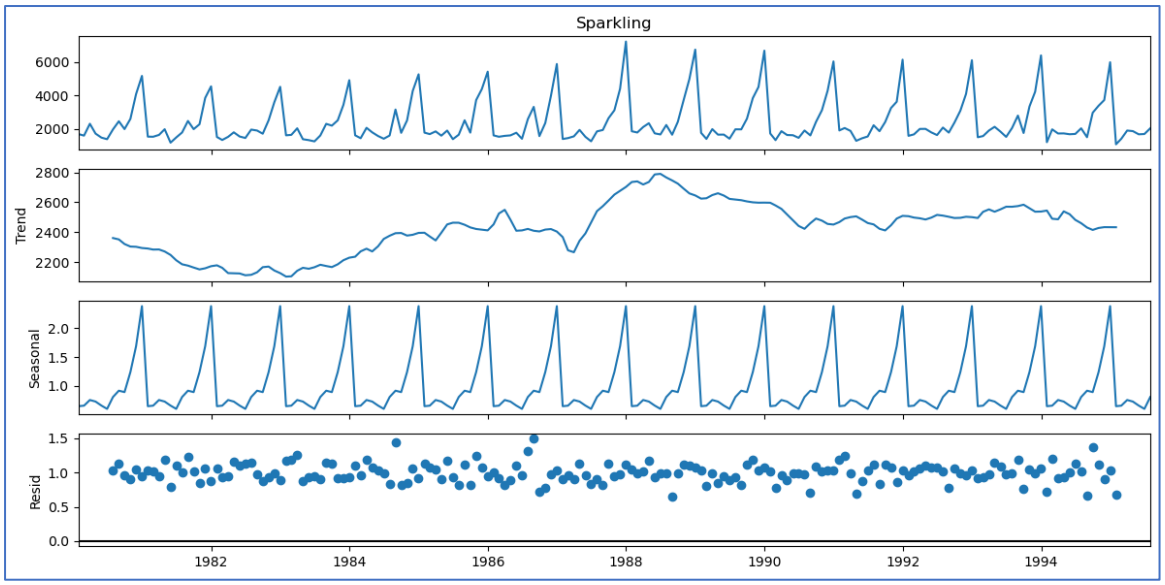


Figure 55 : Decomposition by multiplicative method

- The series can be treated as a multiplicative model.
- Plot shows a consistent trend, but an intermediary period shows an upward trend which gets consistent on the latter half of the time-series
- The residuals show pattern of high variability across time series which is almost consistent in both additive and multiplicative methods.

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

After splitting into train and test data, where test data starts from 1991.

```
(132, 1)
(55, 1)
```

Train has 132 data and test has 55 data.

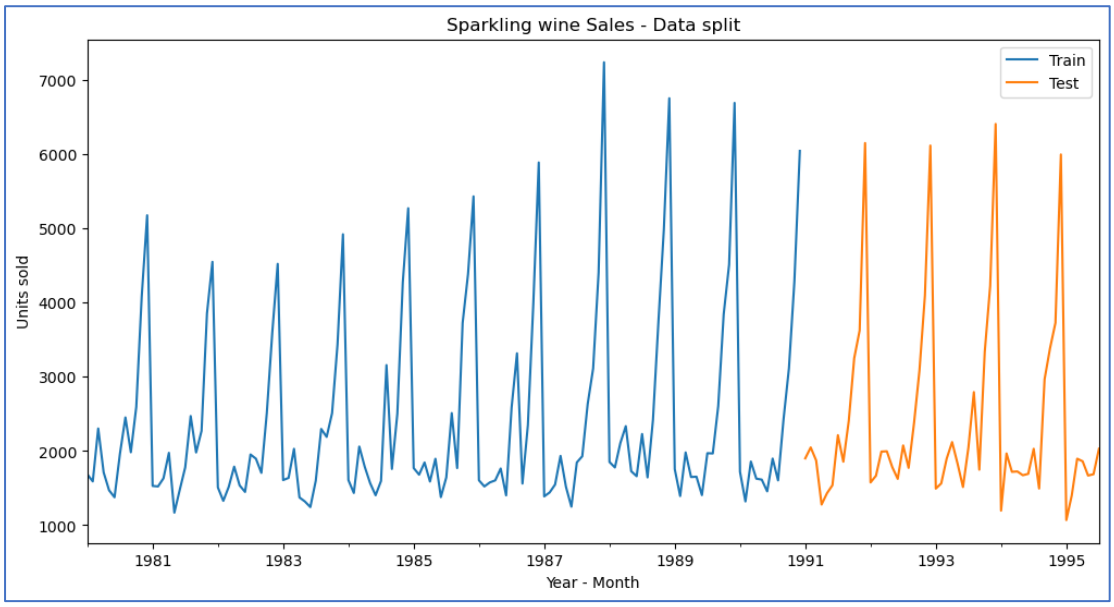


Figure 56 : Sparkling wine sales data split

2.4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

MODEL 1: Linear Regression

To regress sparkling wine sales, time instance order for test data were generated and was added to its dataset.

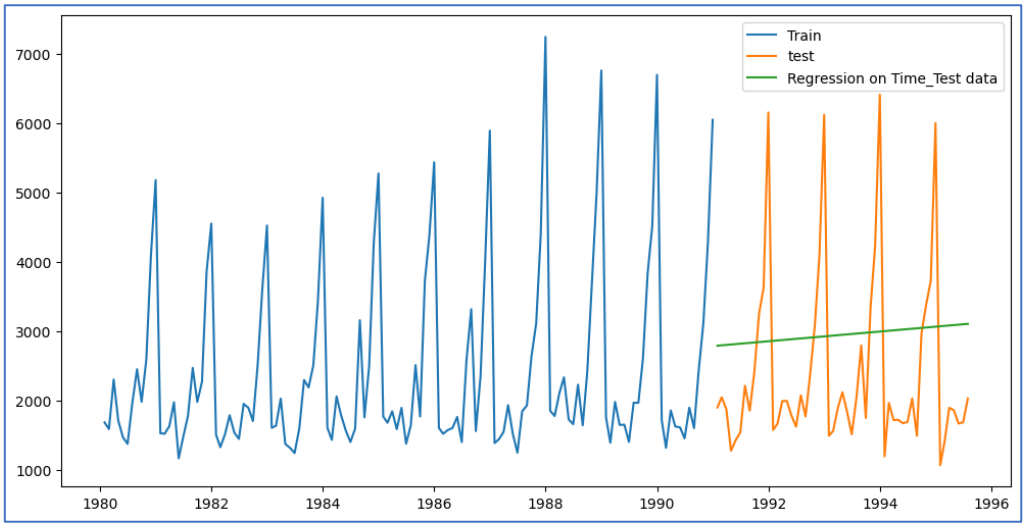


Figure 57 : Linear regression model

- The linear regression shows a gradual upward trend which later becomes constant.
- For regression of time series on test data, RMSE = **1389.1**

MODEL 2: Naïve Forecast

The model has taken last value from test data and fitted it on the rest of the train data and used the same value to forecast the test data.

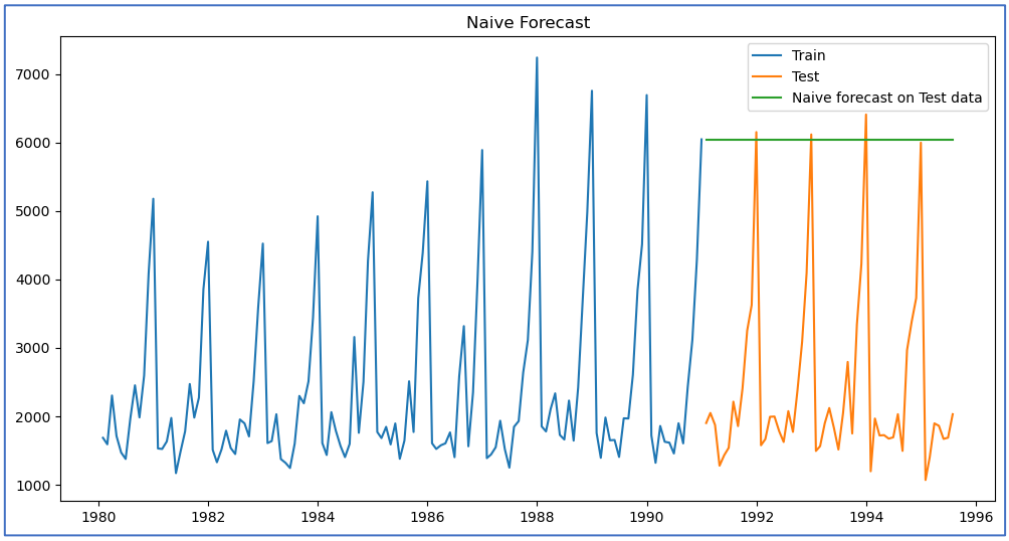


Figure 58 : Naïve forecast model

- For Naïve forecast on test data, RMSE = **3864.28**
- The model neither captures trend nor seasonality.

MODEL 3: Simple average model

Here the forecast is done using mean of the time-series variable from train data.

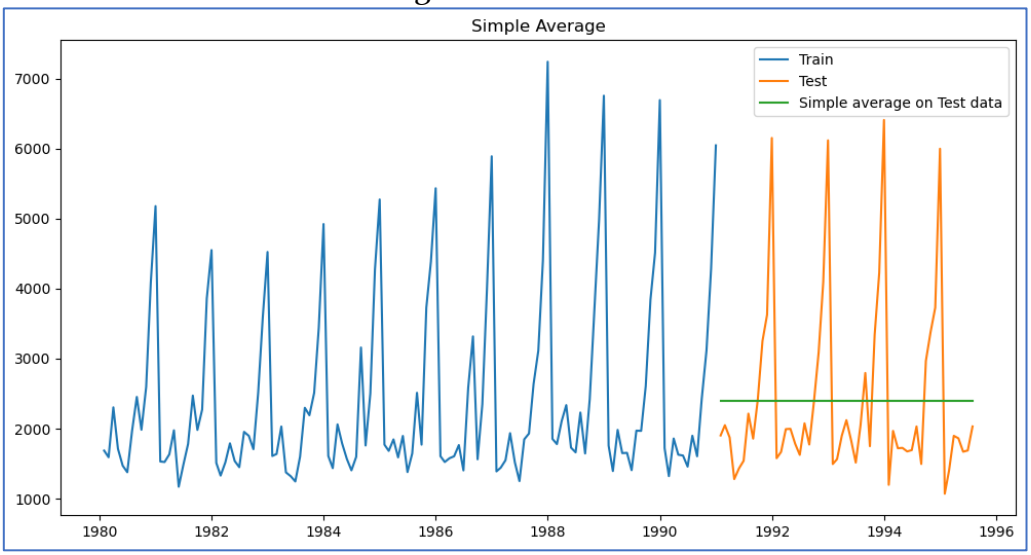


Figure 59 : Simple average model

- The model is unable to forecast.
- It's also unable to capture trend and seasonality.
- For simple average on test data, RMSE = **1275.08**

MODEL 4: Moving average model.

In this model we calculate rolling mean for different intervals. The one with maximum accuracy is the best interval.

Here the moving average is built on 2, 4, 6, 9 points

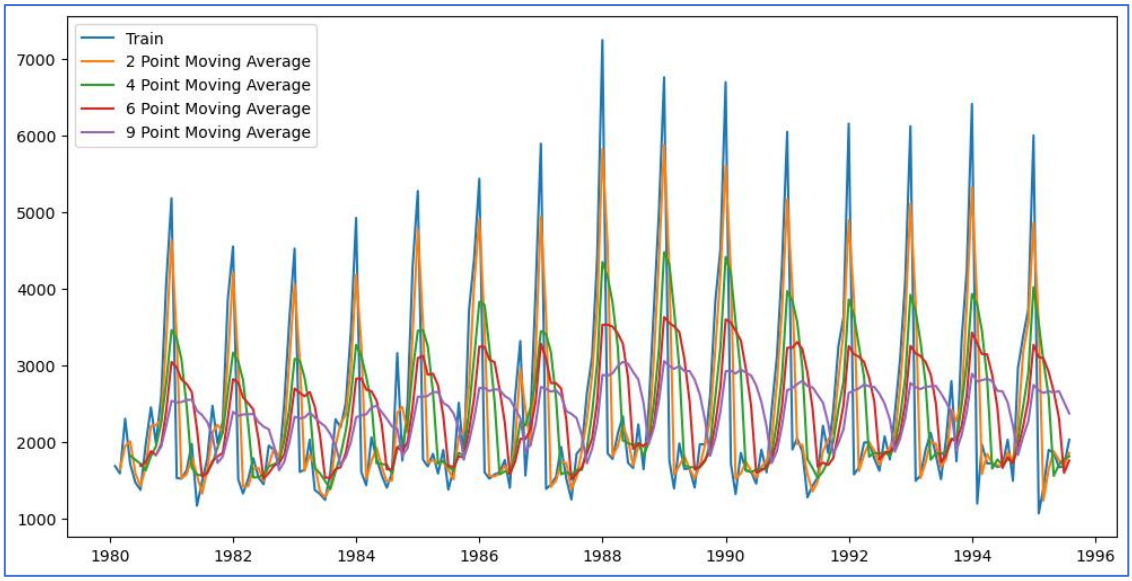


Figure 60 : Moving average model

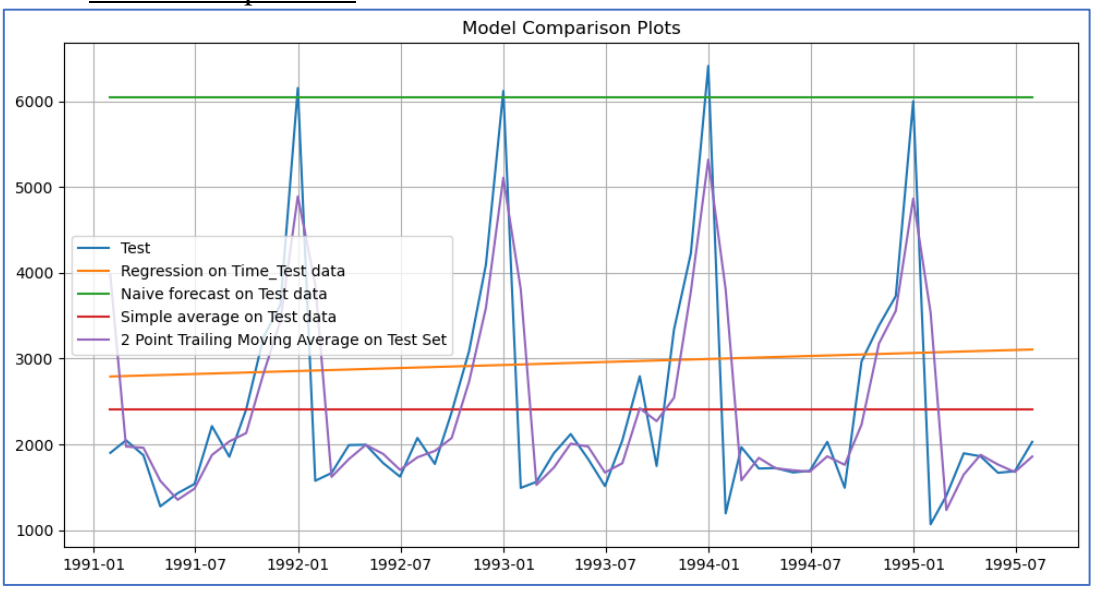
- Here accuracy is higher for lower rolling point averages.

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 813.40
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1156.59
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1283.93
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 1346.28
```

Figure 61 : RMSE of Moving average model

- The best interval of moving average from model is 2 point.

<u>Model comparison:</u>



	Test RMSE
RegressionOnTime	1389.135175
Naive Model	3864.279352
Simple Average model	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315

Figure 62 : Model comparison

MODEL 5: Simple exponential smoothing

Simple exponential smoothing is applied when the time-series has neither trend nor seasonality.

The alpha value was closer to 1, forecasts follow actual observation closely and closer to 0, forecasts are farther from actual line and gets smoothened.

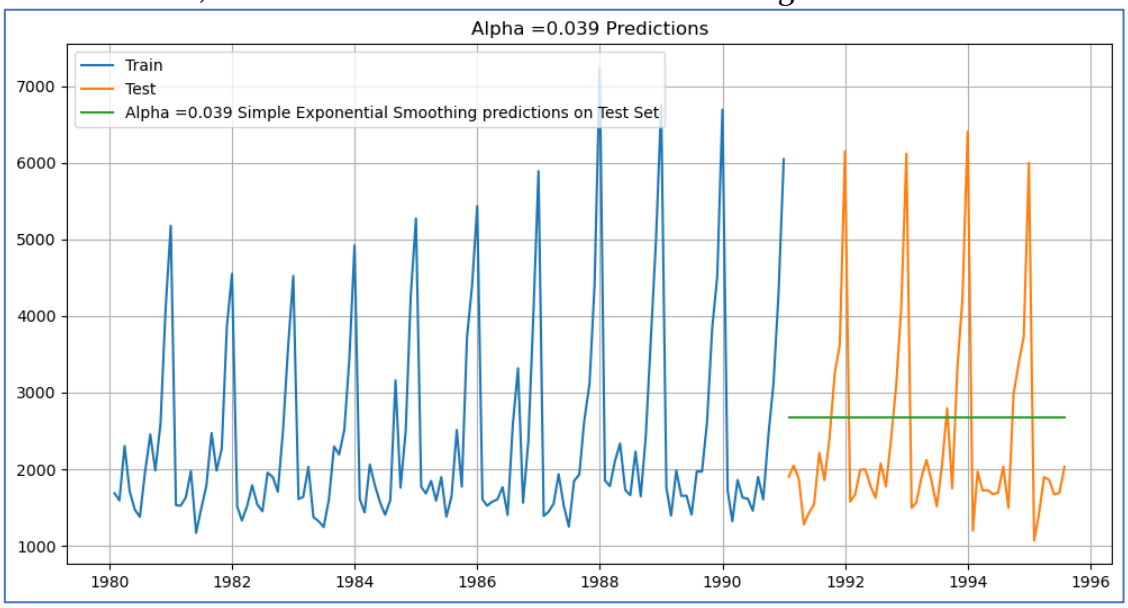


Figure 63 : Simple exponential smoothing- Manual method

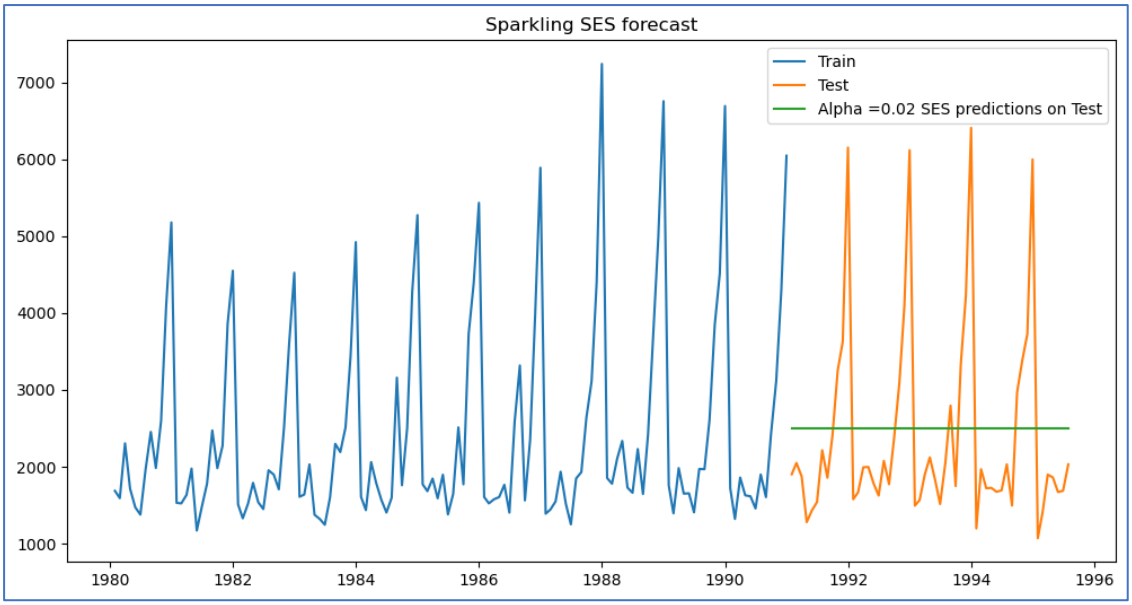


Figure 64 : Simple exponential smoothing-Iterative method

- RMSE:

Alpha=0.039 ,SimpleExponentialSmoothing	1304.927405
Alpha=0.02,SES_Iterative	1279.495201

Figure 65 : RMSE - SES

MODEL 6: Double exponential smoothing (DES)

DES is applied when the data has trend but no seasonality.

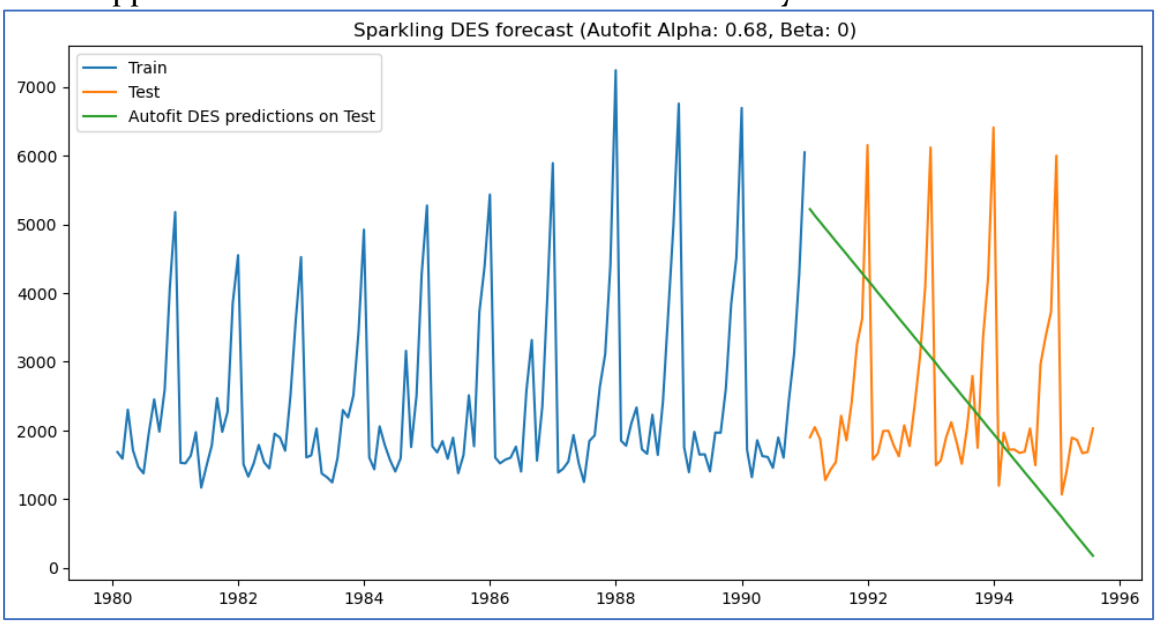


Figure 66 : Double exponential smoothing – Optimized model

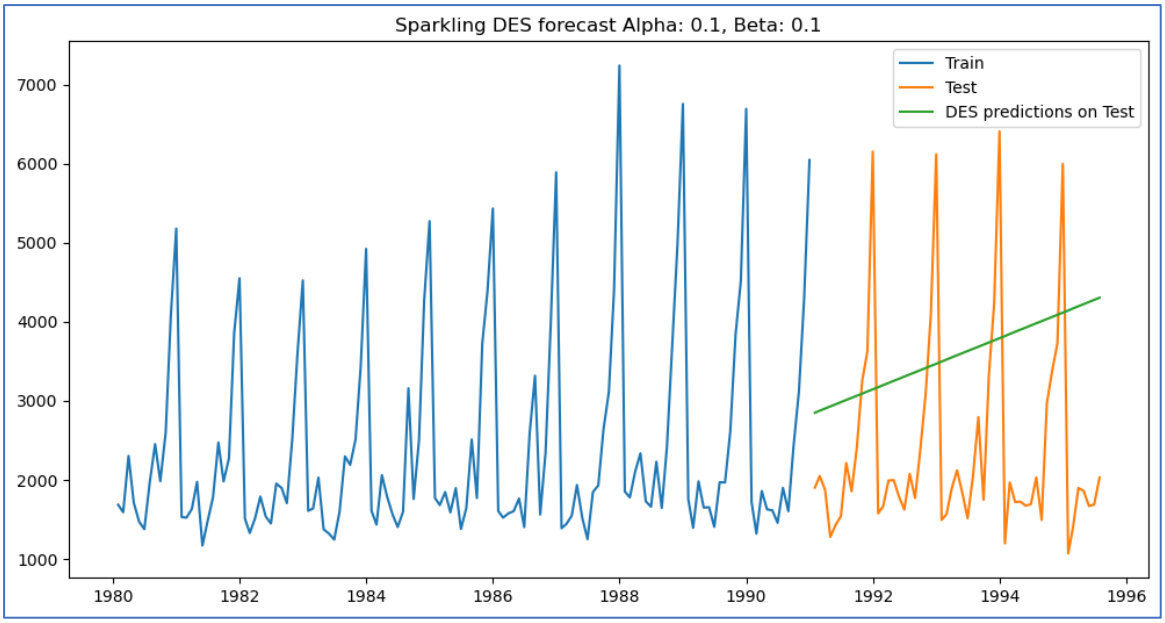


Figure 67 : Double exponential smoothing –Iterative model

- The iterative method has lower RMSE then optimised method.

Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.1,Beta=0.1,DES_Iterative	1778.560000

Figure 68 : RMSE - DES

MODEL 7: Triple exponential smoothing (TES)

TES is applicable when the data has both trend and seasonality.

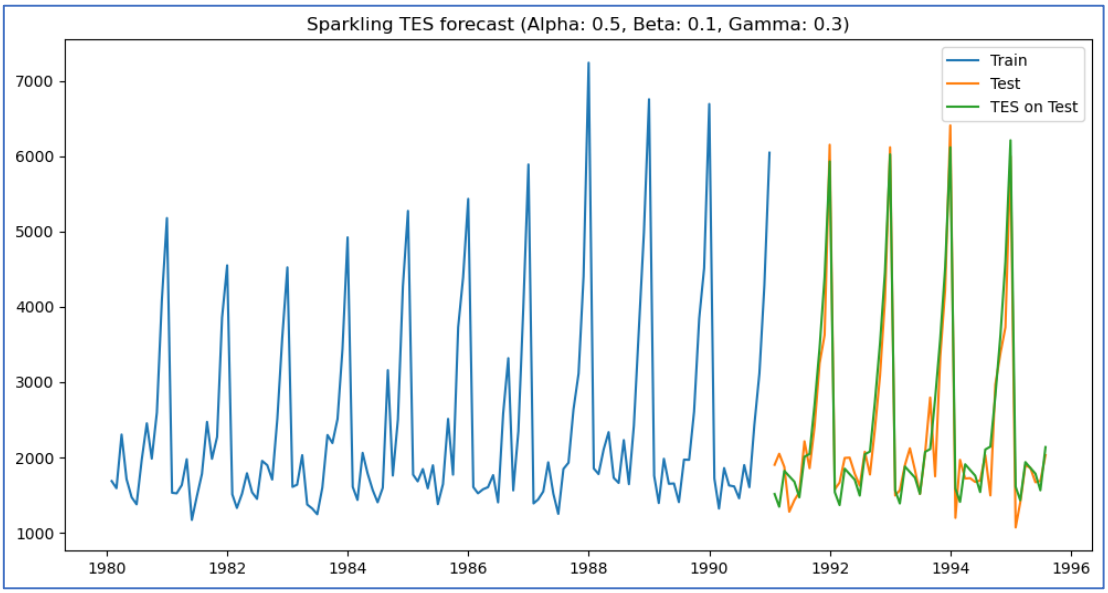


Figure 69 : Triple exponential smoothing – Iterative method

- For iterative TES on test data, RMSE = **345.9**

MODEL COMPARISONS:

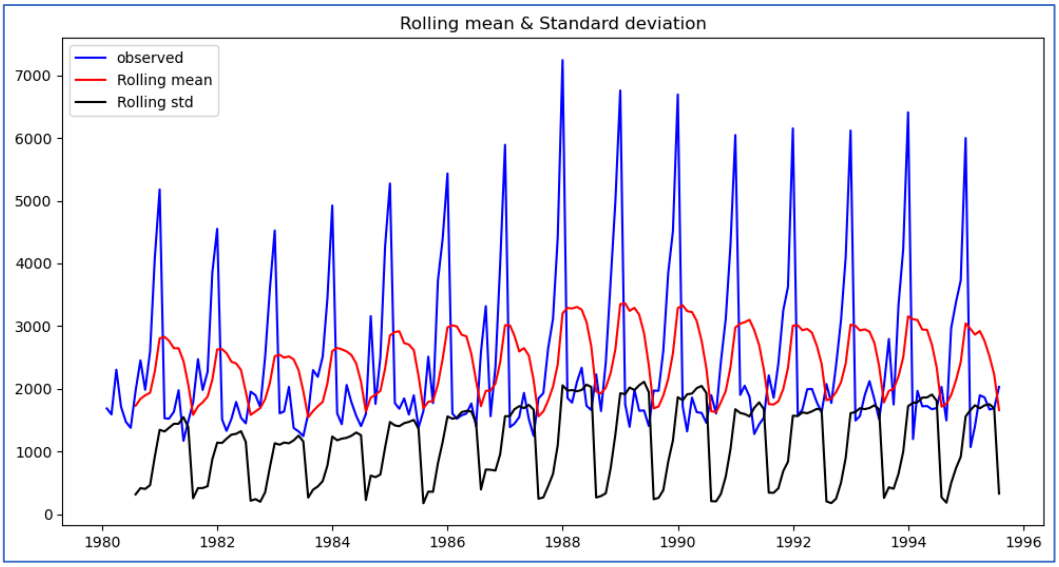
	Test RMSE
RegressionOnTime	1389.135175
Naive Model	3864.279352
Simple Average model	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha=0.039 , SimpleExponentialSmoothing	1304.927405
Alpha=0.02, SES_Iterative	1279.495201
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.1,Beta=0.1,DES_Iterative	1778.560000
Alpha=0.5,Beta=0.1,gamma=0.3,TES_Iterative	345.913415

Figure 70 : Model comparison

2.5. 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.

Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series.
Null Hypothesis: The series has a unit root, that is series is non-stationary.
Alternate Hypothesis: The series has no unit root, that is series is stationary.

If we fail to reject the null hypothesis, the series is non-stationary and if we accept the null hypothesis, the series is said to be stationary.

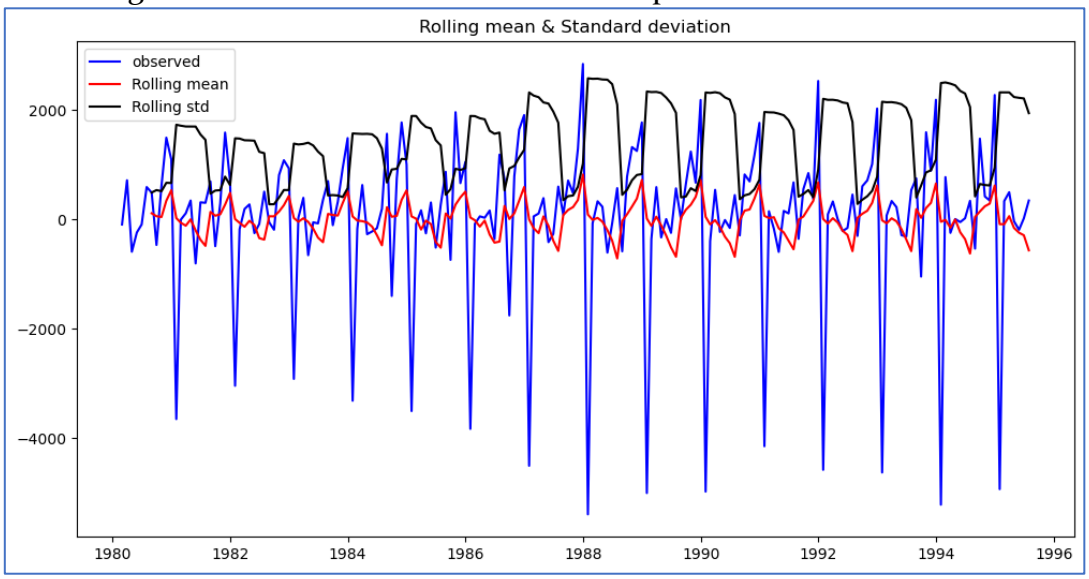



```
Result of Dickey-Fuller test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

Figure 71 : Dickey Fuller test

Here the p value is greater than 0.05, hence we fail to reject null hypothesis and the series is non-stationary.

Differencing of order 1 is applied and again tested for stationarity.
For this rolling mean and standard deviation is also plotted to understand seasonality.



```
Result of Dickey-Fuller test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

Figure 72 : Dickey Fuller test with d=1

Here the p-value is less than 0.05, hence the series has now become stationary.

Auto-correlation function plots

ACF-:

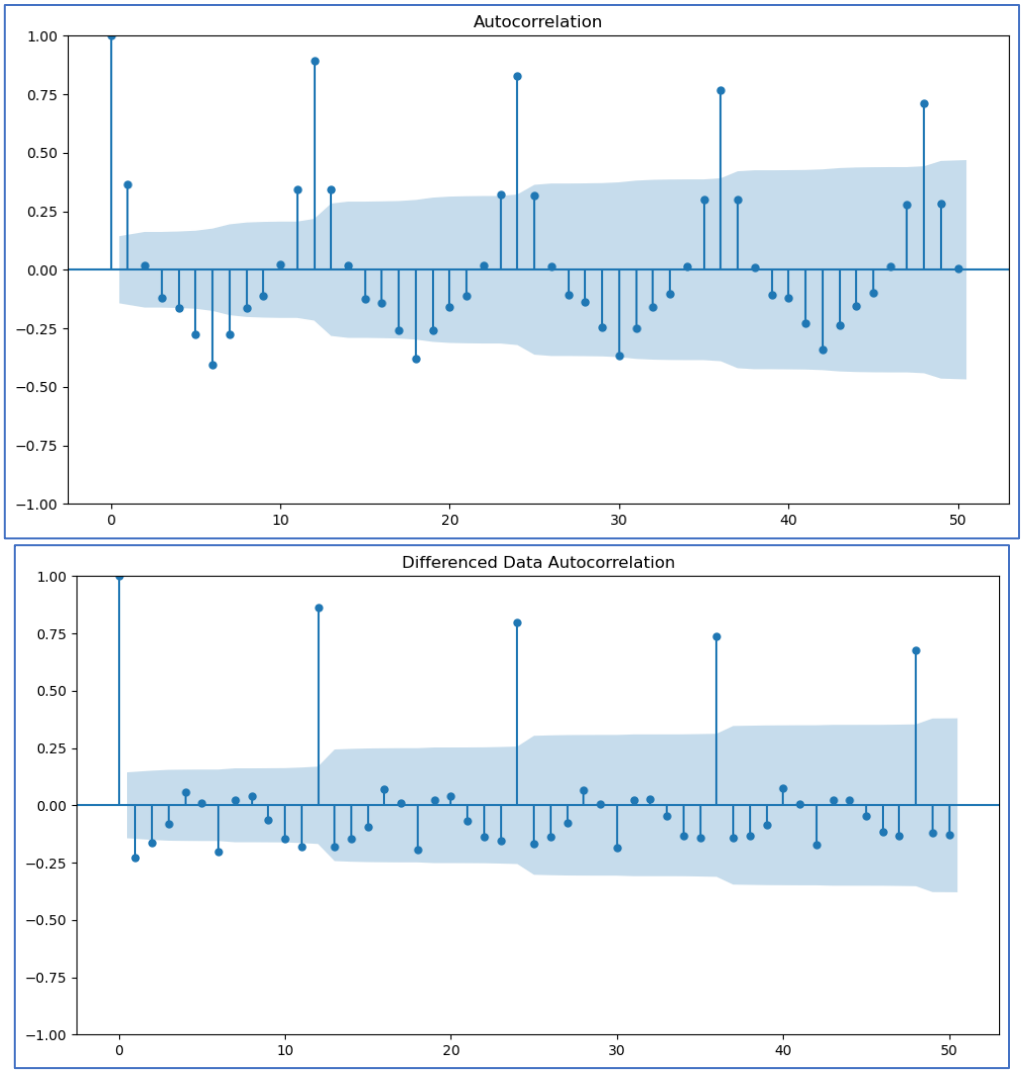


Figure 73 : ACF

PACF-:

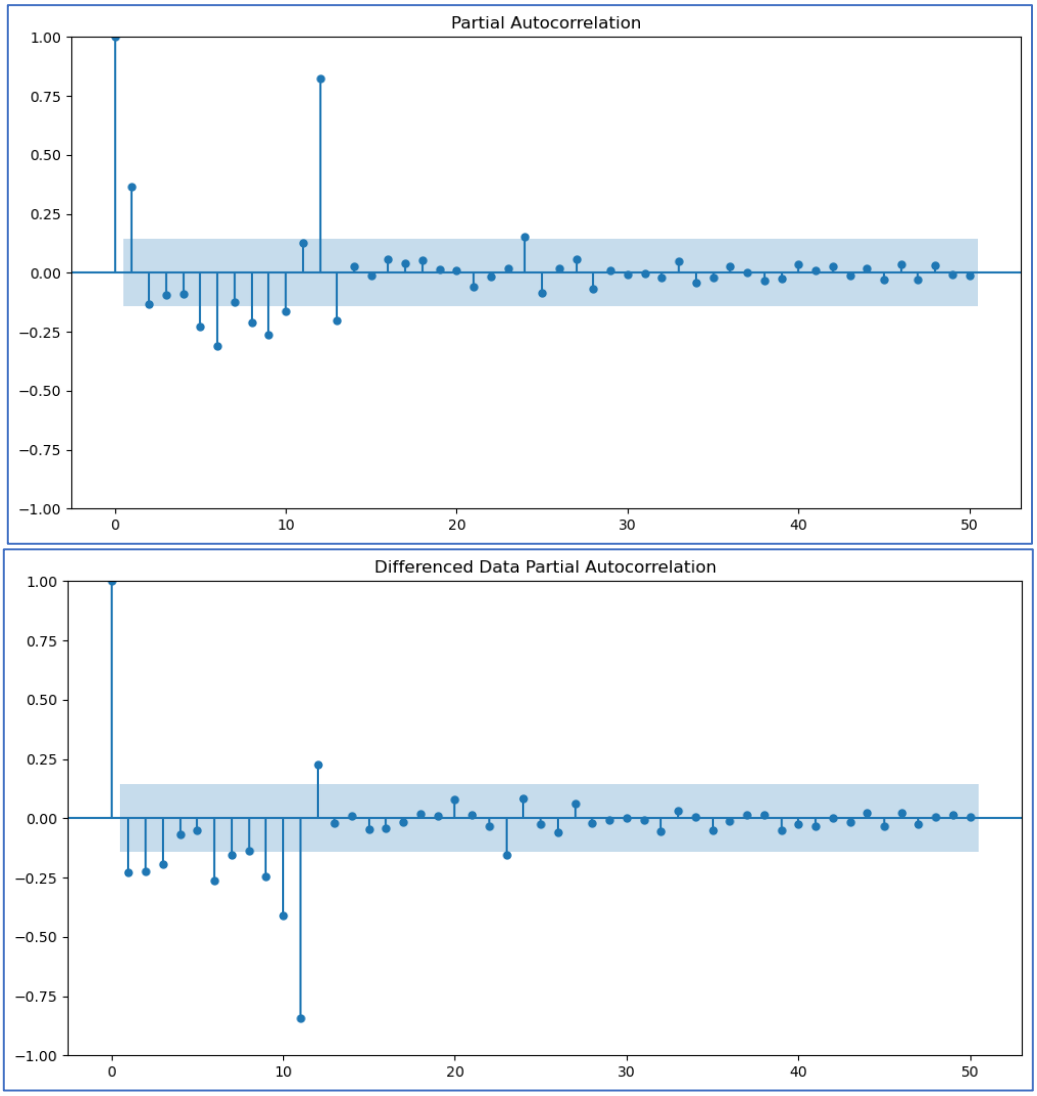


Figure 74 : PACF

The above plots indicate the presence of seasonality.

2.6. 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.

Auto-ARIMA:

An ARIMA model was built with optimised model and found the least AIC value = 2213.5 at (2,1,2)

	param	AIC
8	(2, 1, 2)	2213.509212
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095
6	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
3	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Fri, 08 Dec 2023   AIC                           2213.509
Time:                        11:21:12   BIC                           2227.885
Sample:                    01-31-1980   HQIC                          2219.351
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.217      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
====
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                 1
4.46
Prob(Q):                              0.67   Prob(JB):
0.00
Heteroskedasticity (H):               2.43   Skew:
0.61
Prob(H) (two-sided):                  0.00   Kurtosis:
4.08
===================================================================================
====
```

Figure 75 : ARIMA

- The RMSE was found to be 1299.980

Auto-SARIMA:

The model was built on train data with seasonality 12 and with different optimal parameters (p, d, q)x(P, D, Q) parameters.
The lowest AIC is 774.96 was obtained at (0, 1, 2)x(2, 1, 2, 12).

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-686.012
Date:	Fri, 08 Dec 2023	AIC	1386.024
Time:	11:26:58	BIC	1403.676
Sample:	0	HQIC	1393.148
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8016	0.192	-4.186	0.000	-1.177	-0.426
ma.L2	-0.2427	0.132	-1.843	0.065	-0.501	0.015
ar.S.L12	-0.2064	1.067	-0.194	0.847	-2.297	1.884
ar.S.L24	-0.0869	0.181	-0.480	0.631	-0.442	0.268
ma.S.L12	-0.2099	1.059	-0.198	0.843	-2.285	1.865
ma.S.L24	-0.0397	0.451	-0.088	0.930	-0.924	0.845
sigma2	1.619e+05	2.14e+04	7.558	0.000	1.2e+05	2.04e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	27.55
Prob(Q):	0.85	Prob(JB):	0.00
Heteroskedasticity (H):	0.88	Skew:	0.78
Prob(H) (two-sided):	0.73	Kurtosis:	5.19

Figure 76 : SARIMA

- The RMSE was found to be **327.524**

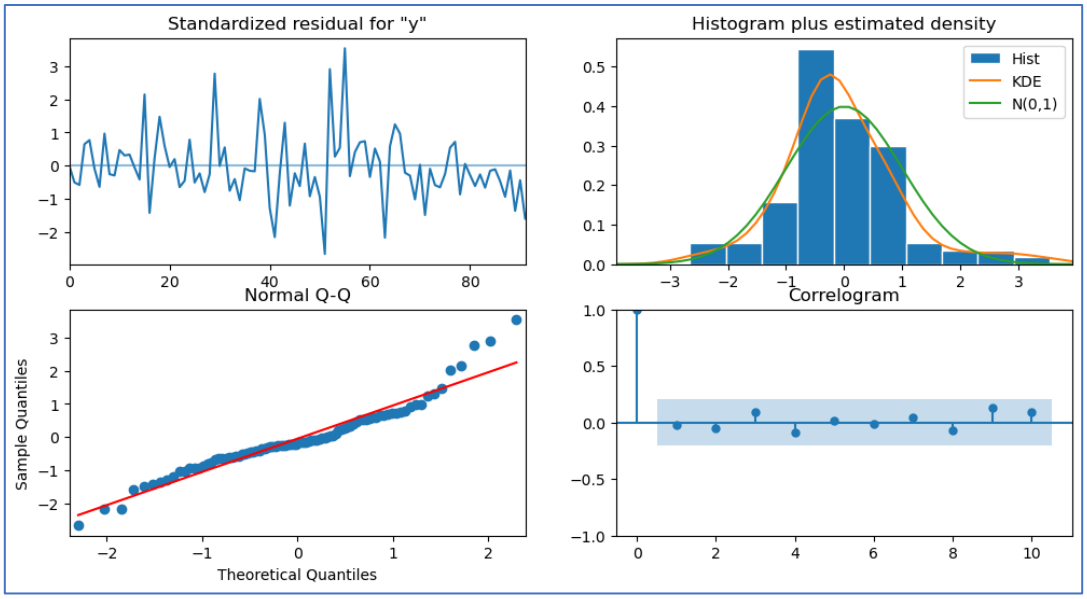


Figure 77 : Diagnostic plot

- The residuals are found to follow a mean of zero
- The histogram shows that residuals follow normal distribution.
- The normal Q-Q plot shows that the quantiles come from a normal distribution as they almost form a straight line.
- The correlogram shows autocorrelation of residuals and there is no lag above the confidence limit.

PREDICTING ON TEST DATA:

Time_stamp	Sparkling	sparkling_forecasted
1991-01-31	1902	1575.977119
1991-02-28	2049	1408.288350
1991-03-31	1874	1898.483136
1991-04-30	1279	1745.269764
1991-05-31	1432	1632.303072

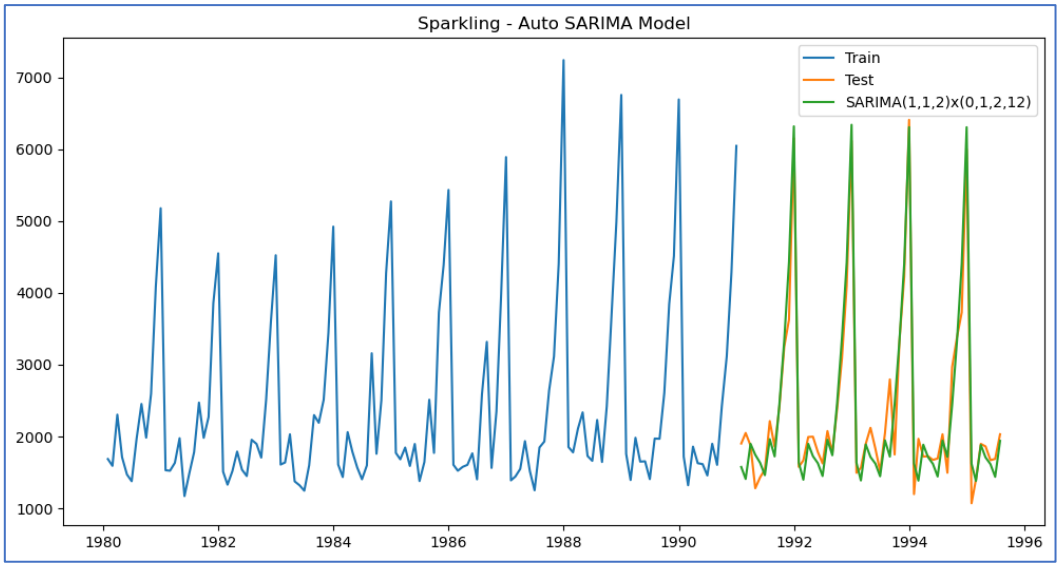


Figure 78 : SARIMA prediction on test data

The model built with log-series data has a lower RMSE than the original train data.

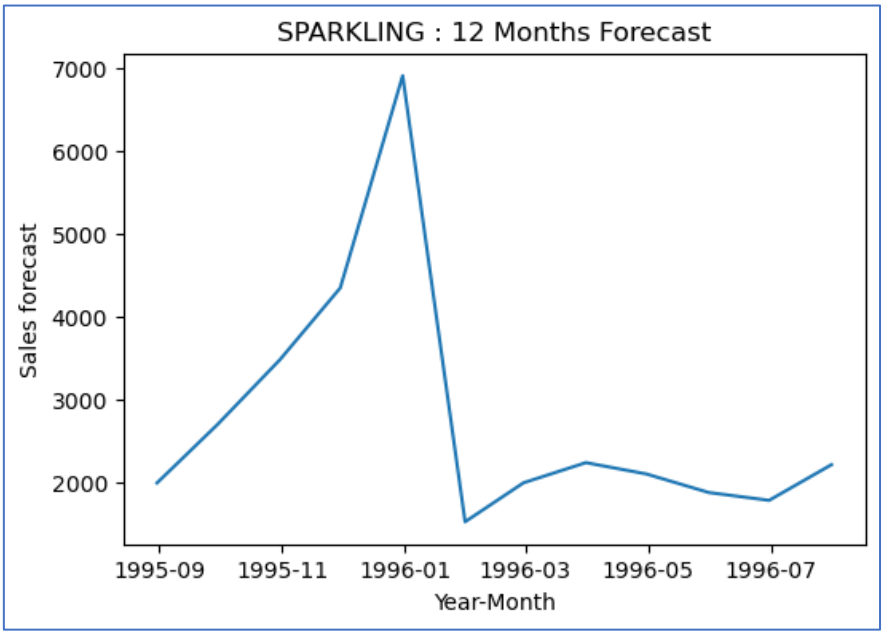
2.7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

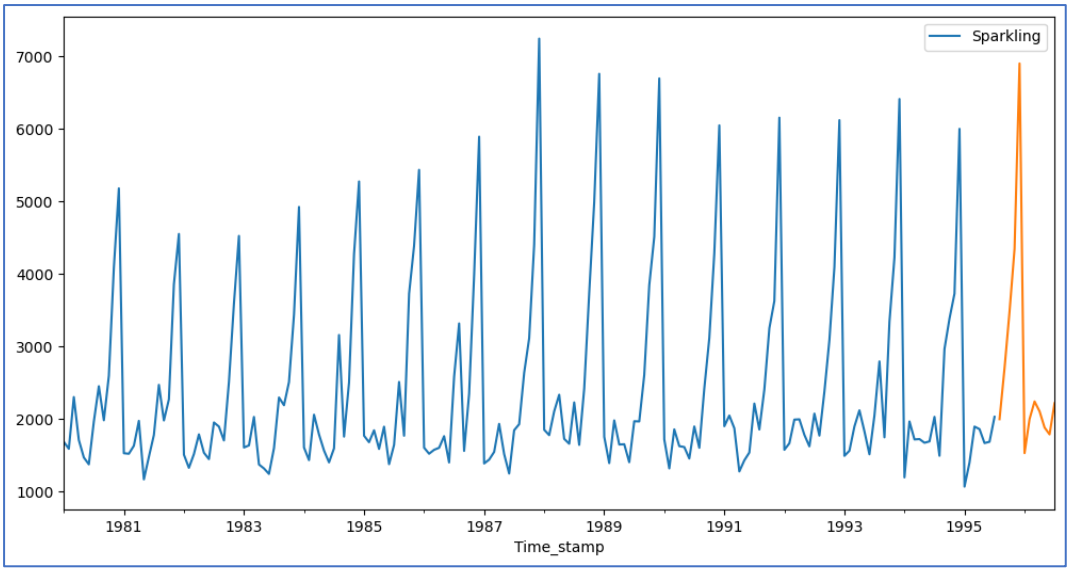
	Test RMSE
RegressionOnTime	1389.135175
Naive Model	3864.279352
Simple Average model	1275.081804
2 point TMA	813.400684
4 point TMA	1156.589694
6 point TMA	1283.927428
9 point TMA	1346.278315
Alpha=0.039 , SimpleExponentialSmoothing	1304.927405
Alpha=0.02,SES_Iterative	1279.495201
Alpha=0.68,Beta=0.0, DES Optimized	2007.238526
Alpha=0.1,Beta=0.1,DES_Iterative	1778.560000
Alpha=0.5,Beta=0.1,gamma=0.3,TES_Iterative	345.913415
Auto_ARIMA(2, 1, 2)	1299.980041
Auto_SARIMA(1, 1, 2)*(0, 1, 2, 12)	327.524217

Figure 79 : RMSE of all models

2.8. 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.

- Based on the overall model evaluation, triple exponential smoothing (Holt Winter's method) is selected for prediction as it has the lowest RMSE.
- TES model alpha: 0.1, beta: 0.2 and gamma: 0.3 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model.





- *Figure 80 : Sparkling – 12 months forecast*

- On calculating upper and lower confidence bands at 95% confidence level, we are taking the multiplier as 1.96 as we want the plot with 95% confidence limit.

	lower_CI	prediction	upper_ci
1995-08-31	1205.518357	1999.209720	2792.901082
1995-09-30	1907.715655	2701.407018	3495.098380
1995-10-31	2687.412962	3481.104325	4274.795688
1995-11-30	3552.120280	4345.811643	5139.503005
1995-12-31	6106.190745	6899.882108	7693.573471
1996-01-31	737.799229	1531.490592	2325.181955
1996-02-29	1207.022578	2000.713941	2794.405303
1996-03-31	1449.576227	2243.267590	3036.958953
1996-04-30	1313.553559	2107.244922	2900.936285
1996-05-31	1090.275557	1883.966920	2677.658283
1996-06-30	995.954921	1789.646284	2583.337647
1996-07-31	1425.677466	2219.368829	3013.060192

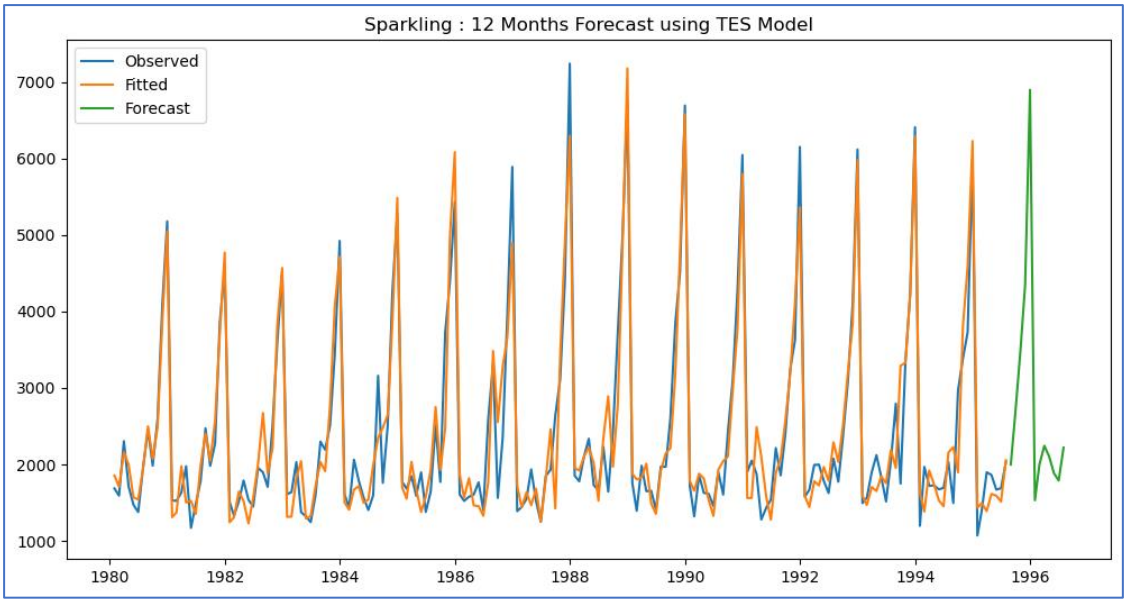


Figure 81 : 12 months forecast using TES

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

	lower_CI	prediction	upper_ci
count	12.000000	12.000000	12.000000
mean	1973.234795	2766.926157	3560.617520
std	1523.608988	1523.608988	1523.608988
min	737.799229	1531.490592	2325.181955
25%	1176.707657	1970.399020	2764.090382
50%	1369.615513	2163.306875	2956.998238
75%	2102.639982	2896.331345	3690.022707
max	6106.190745	6899.882108	7693.573471

	lower_CI	prediction	upper_ci
1995-08-31	1205.518357	1999.209720	2792.901082
1995-09-30	1907.715655	2701.407018	3495.098380
1995-10-31	2687.412962	3481.104325	4274.795688
1995-11-30	3552.120280	4345.811643	5139.503005
1995-12-31	6106.190745	6899.882108	7693.573471
1996-01-31	737.799229	1531.490592	2325.181955
1996-02-29	1207.022578	2000.713941	2794.405303
1996-03-31	1449.576227	2243.267590	3036.958953
1996-04-30	1313.553559	2107.244922	2900.936285
1996-05-31	1090.275557	1883.966920	2677.658283
1996-06-30	995.954921	1789.646284	2583.337647
1996-07-31	1425.677466	2219.368829	3013.060192

```
lower_CI      23678.817534
prediction    33203.113889
upper_ci      42727.410244
dtype: float64
```

- *Figure 82 : Forecast data and its description*

- The model forecasts 33203 units of rose wine in next 12 months.
- Average of 2766 units per month.
- Maximum of 6899 units will be sold in December 1995.
- Least units will be sold in January 1996.
- The forecast indicates that the one year sale of sparkling wine is not showing an upward trend.
- The ABC estate must look into it and take necessary actions in promoting the product.
- They must invest in marketing their wine and promoting them.