BUSINESS REPORT

ON

TIME SERIES FORECASTING

By Kshitij Nishant

Table of Contents

Contents

List of Figures

Figures

This Particular Report is for Rose.csv

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.

Data set for the Problem: Sparkling.csv and Rose.csv

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

[I have made report on one dataset at a time]

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

Answer:

After Interpolation:

Head:

	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

Tail:

	YearMonth	Rose
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

I converted them into date format:

	YearMonth	Rose	Date
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

We have converted the data into date format and given the column name as Date.
I also dropped the column YearMonth as we got the month year and date format in one column named Time_Stamp:

Date	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

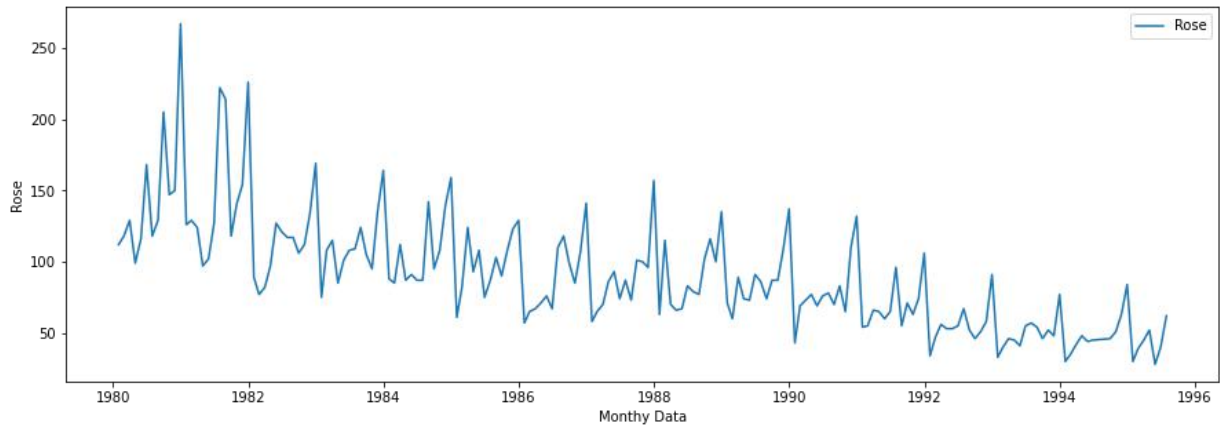

Fig1

• Rose wine sales shows a decreasing trend in the initial years which stabilizes after few years and again shows a decreasing trend
• Rose wines sales shows seasonality in the data trend and pattern seem to repeat on yearly basis

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

Answer:

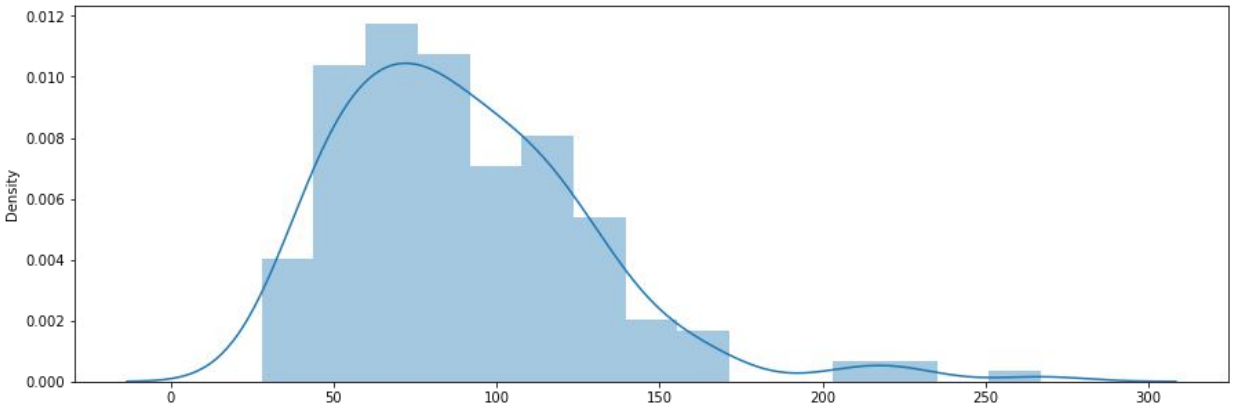

Fig 2

Data is skewed towards left.

Description of data:

	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

• There are 187 observations which represent the monthly sales of respective wines form the year 1980 to July 1995.

• The data has two variables the year/month of sales and the sales for the respective month of the year.

• Mean, min, max values for sparkling wine sales are greater than rose wine sales.

Shape of data: (187, 2)

Null value check: We found 2 null values.There are no null values in data set Rose after interpolation.

Distribution of sale of wine-Rose in each year via BoxPlot:

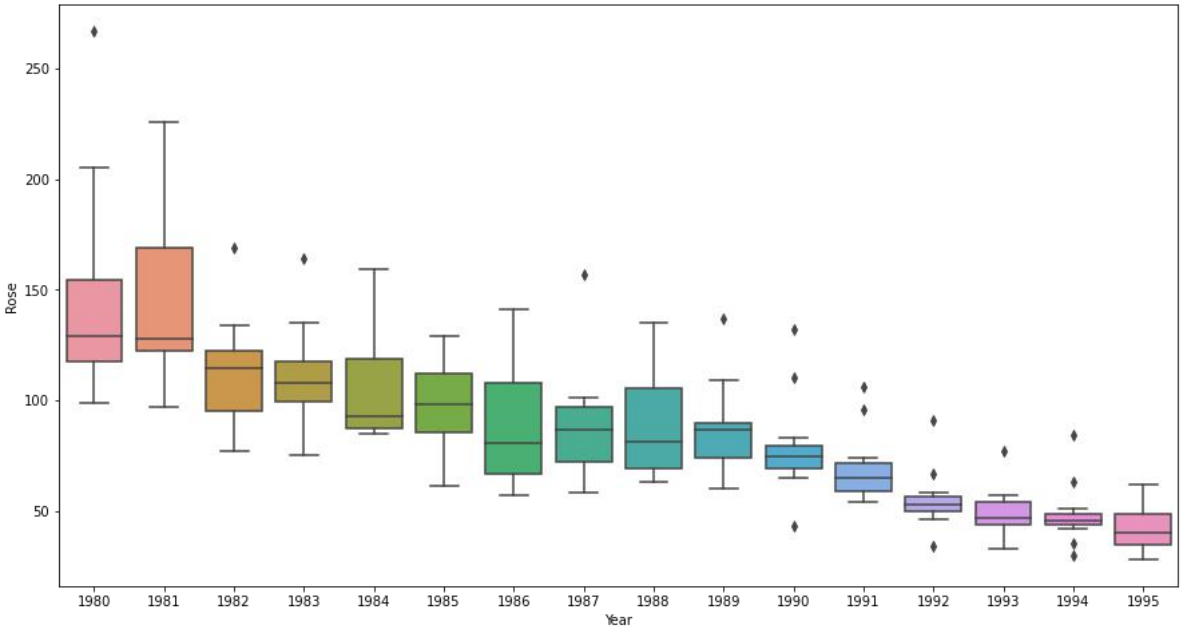

Fig3

- Outliers are present when looking at corresponding year wise data.

Distribution of sale of wine-Rose in each year via BarPlot:

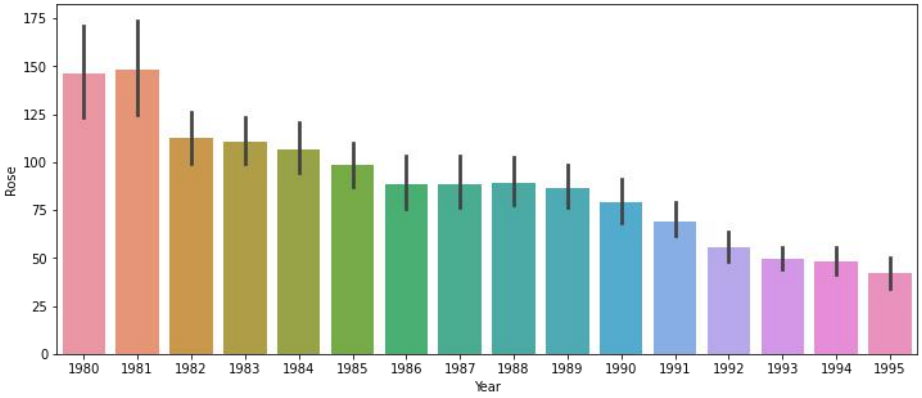

Fig4

- Data seems to have gradual decrease in sales across the year. 1981 has recorded maximum sales.

Distribution of sale of wine-Rose in each Month via BarPlot:

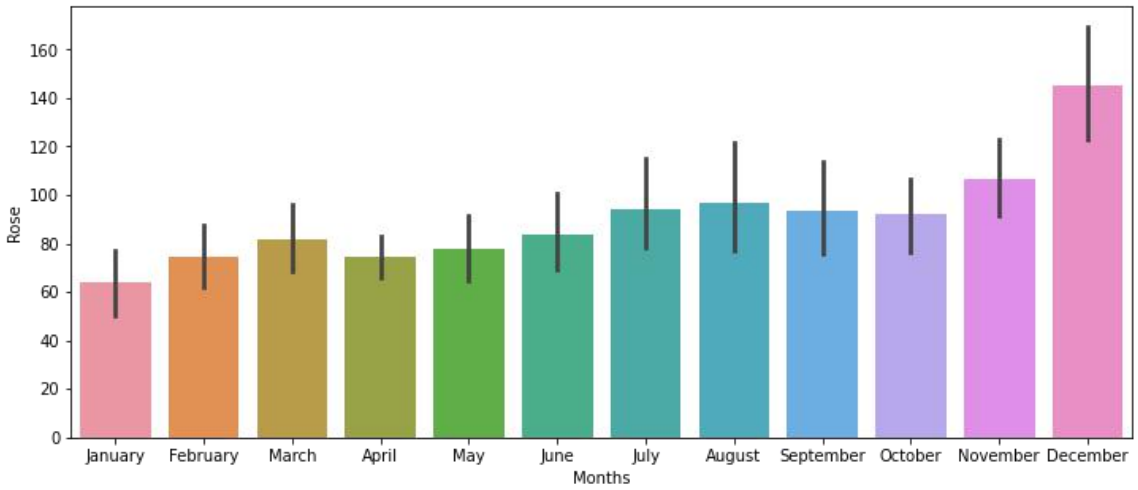

Fig5

- December have greatest amount of sales across all the months followed be November and August.

- Greater in sales may be due to the celebration in year end.

Distribution of sale of wine-Rose in each Month via BoxPlot:

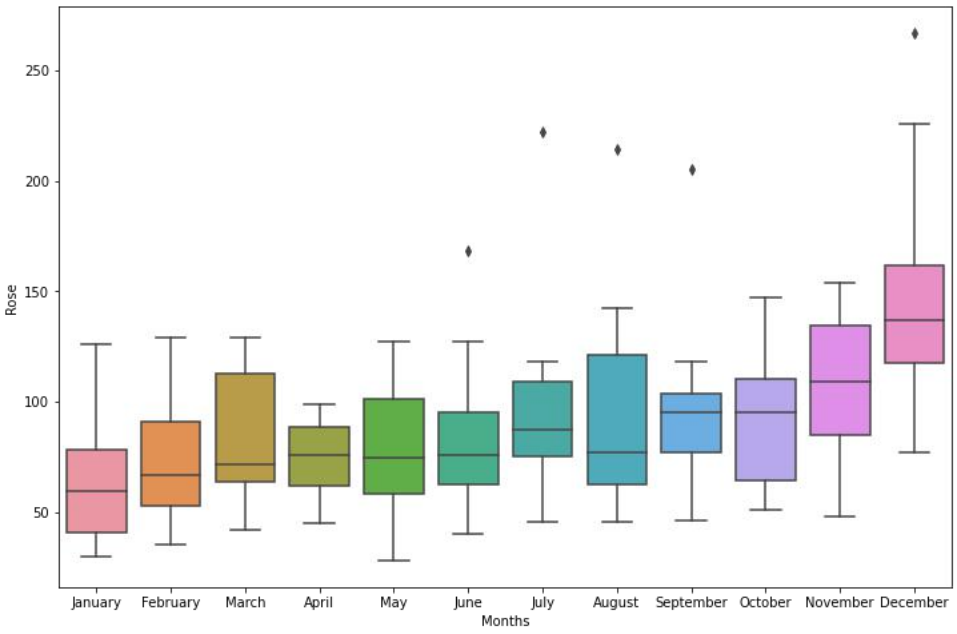

Fig6

- Box plot is also shows us that December has recorded most number of sales.

Distribution of average sale of wine-Rose of each Day via BarPlot:

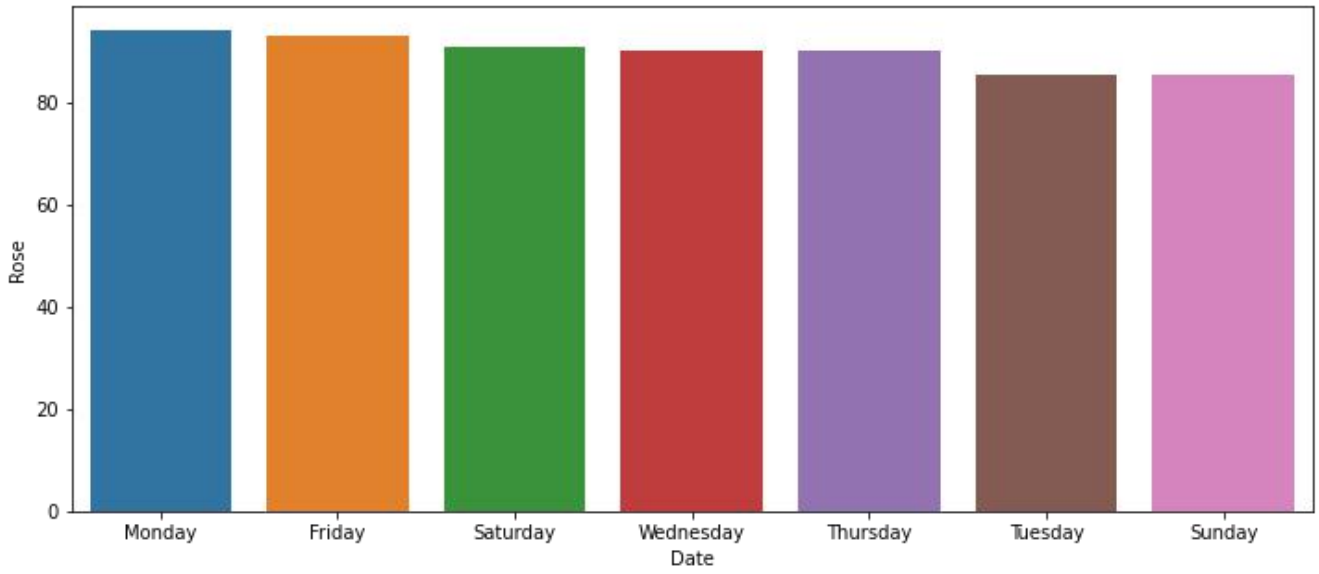

Fig7

- Monday registers highest average sales of beer throughout the whole week.

Distribution of daily sale of wine-Rose of each day via BarPlot:

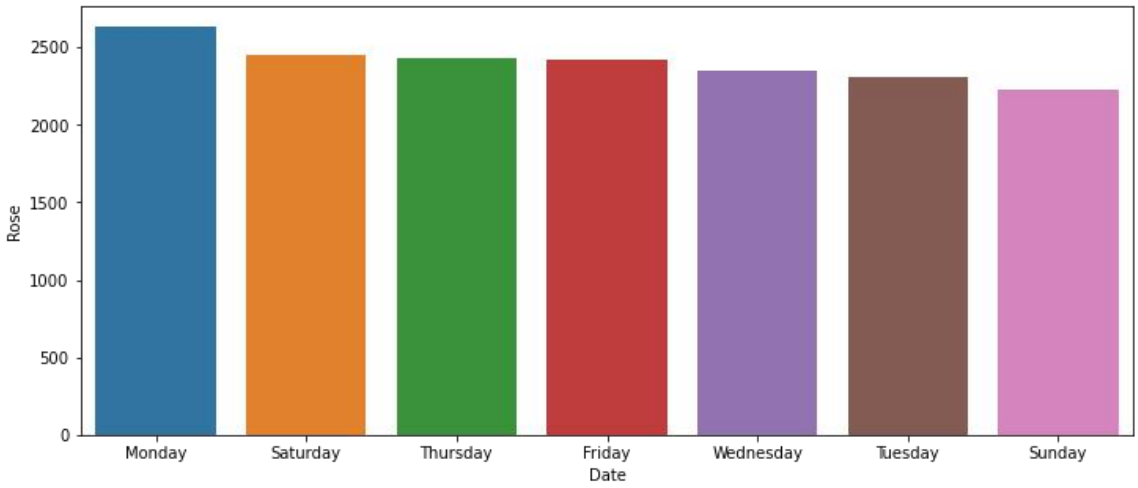

Fig8

- Monday has the highest sales overall.

Time series monthplot to understand the spread of Rose Sales across different years and within different months across years:

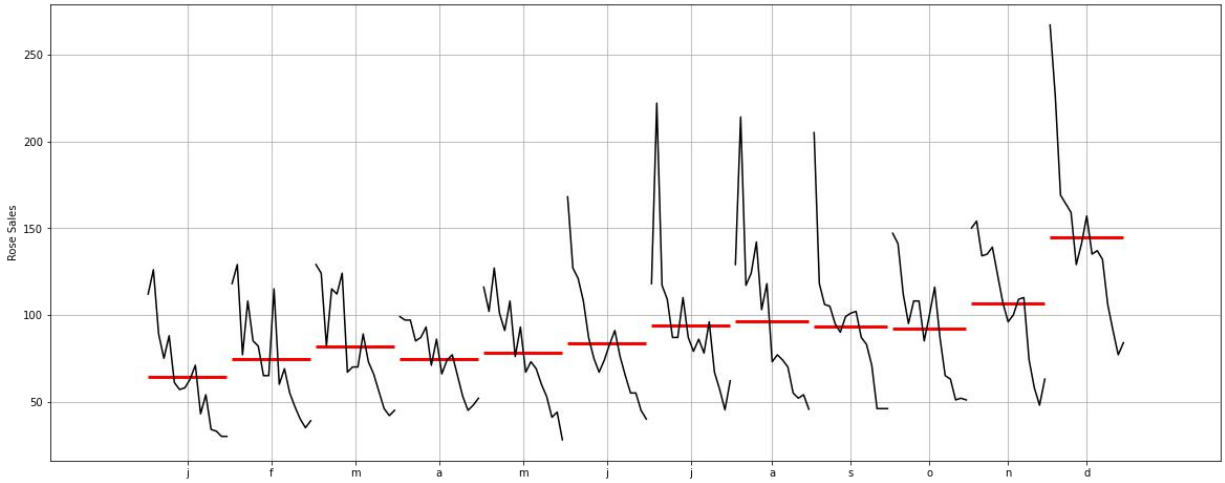

Fig9

Graph of monthly Rose's Sales across years:

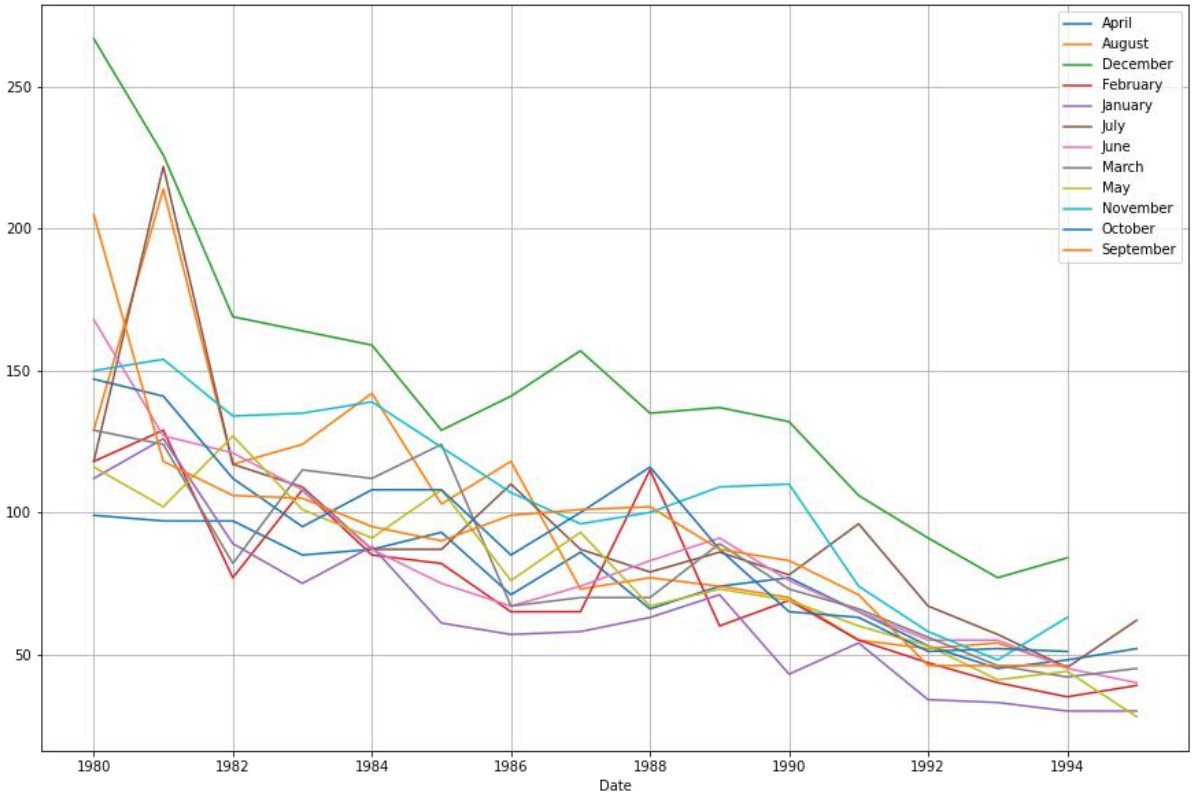

Fig 10

- Dec registers the highest amount of sales.

Empirical Cumulative Distribution:

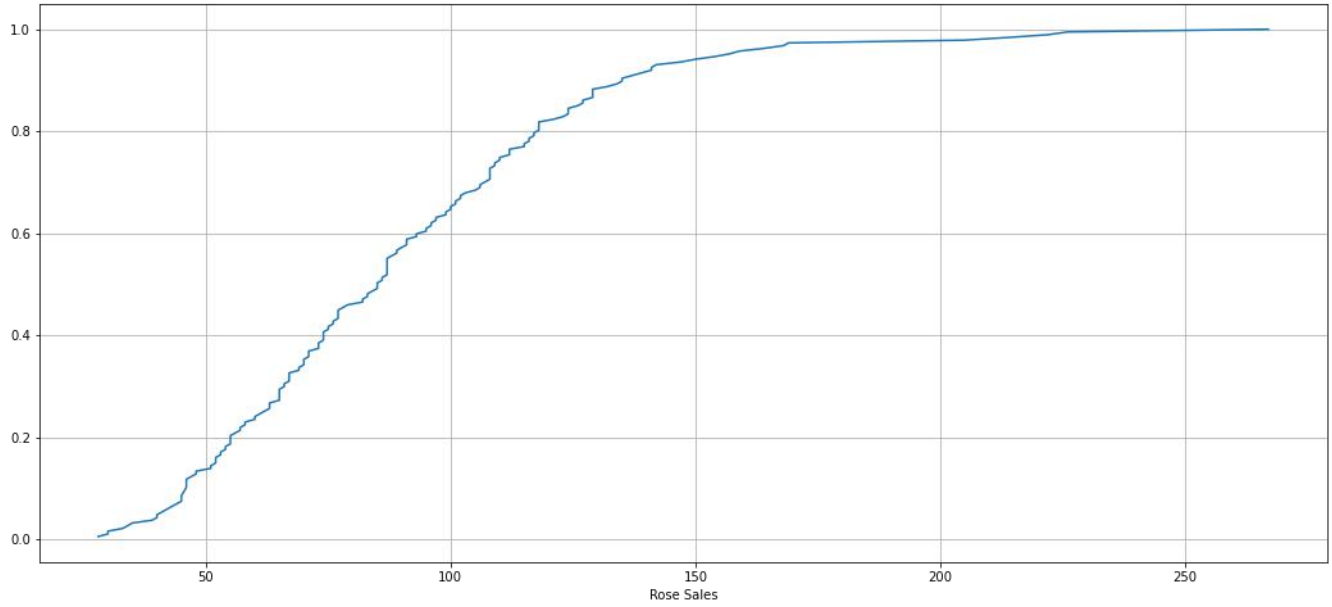

Fig 11

Average Sparkling Sales per month and the month on month percentage change of Sparkling Sales:

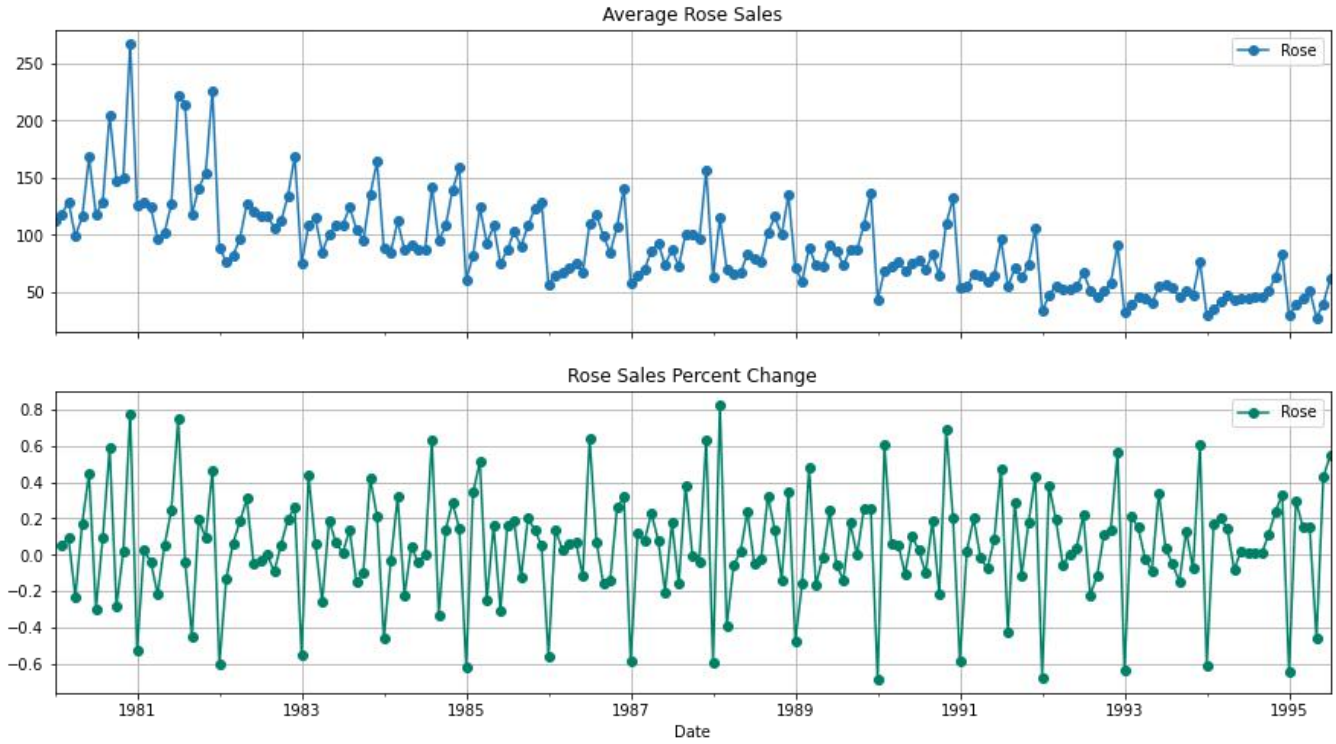

Fig 12

The median values keep increasing from January to December months. The Average Sales value also shows a decreasing trend

Additive Decomposition of dataset:

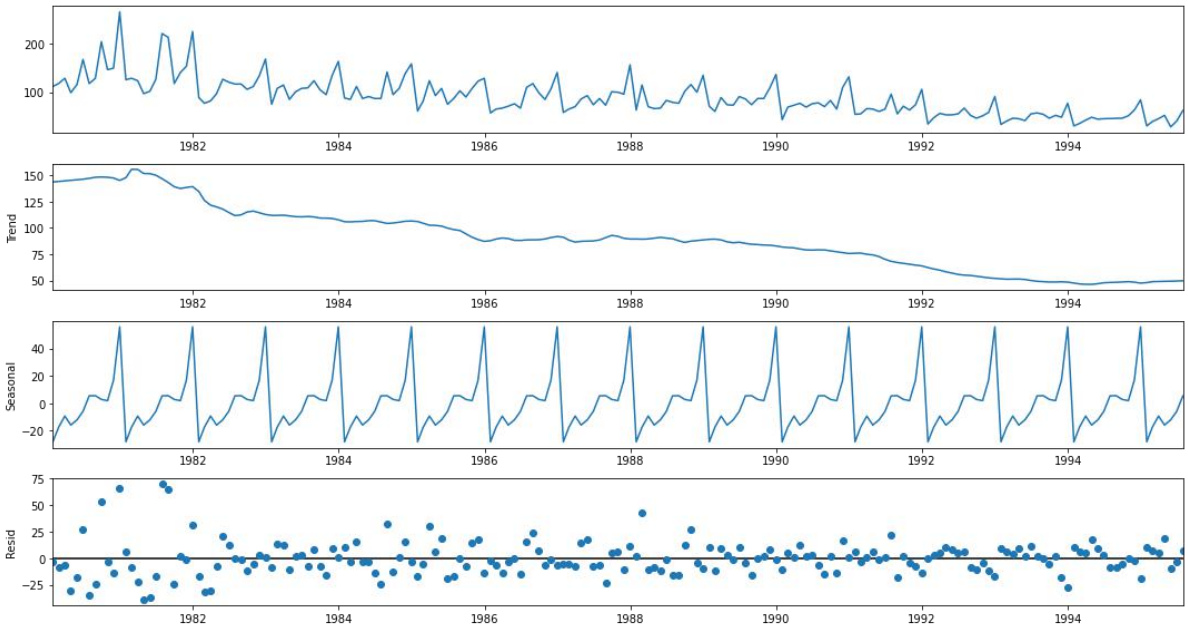

Fig 13

Multiplicative Decomposition of dataset:

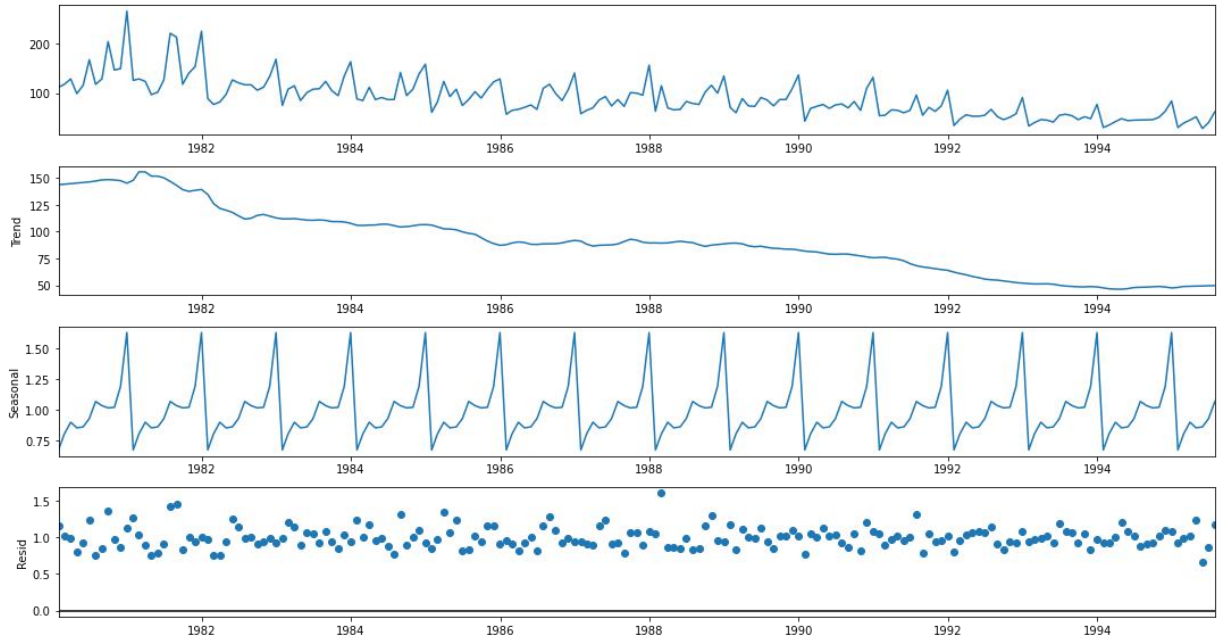

Fig 14

- For additive we see the residual values don't make any pattern and there is no increasing treand or seasonality but for Multiplicative model we see the residual make some form of pattern.
- So I decided to choose additive model is better for forecasting Rose.csv.

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

Answer:

Tail of Train Set:

Date	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

Head of Test Set:

Date	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Train-Test Plot

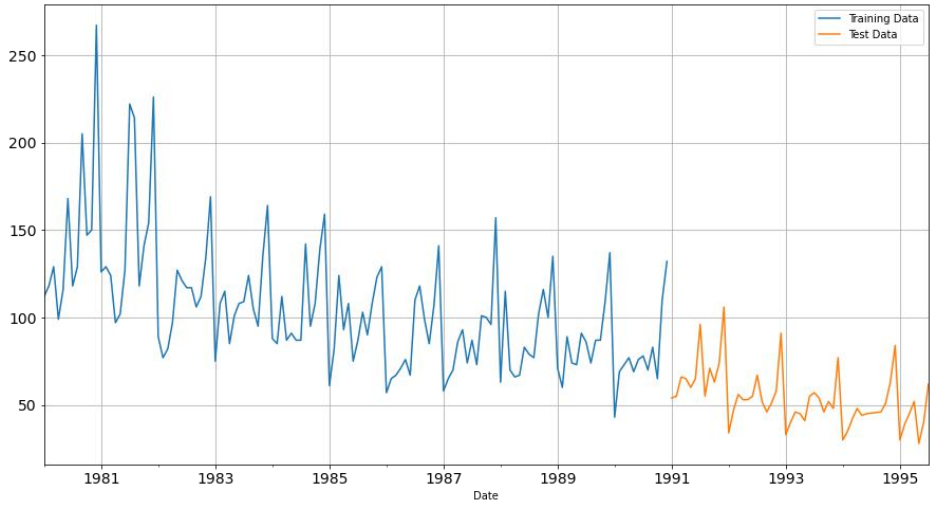

Fig15

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.

Answer:

Model 1: Linear Regression

Train set after Predictions:

Date	Rose	train_time	RegOnTime
1980-01-31	112.0	1	137.321144
1980-02-29	118.0	2	136.826766
1980-03-31	129.0	3	136.332388
1980-04-30	99.0	4	135.838010
1980-05-31	116.0	5	135.343632

Test Set after Predictions:

Date	Rose	test_time	RegOnTime
1991-01-31	54.0	133	72.063266
1991-02-28	55.0	134	71.568888
1991-03-31	66.0	135	71.074511
1991-04-30	65.0	136	70.580133
1991-05-31	60.0	137	70.085755

Plotting Training,Testing and values obtained from regression model:

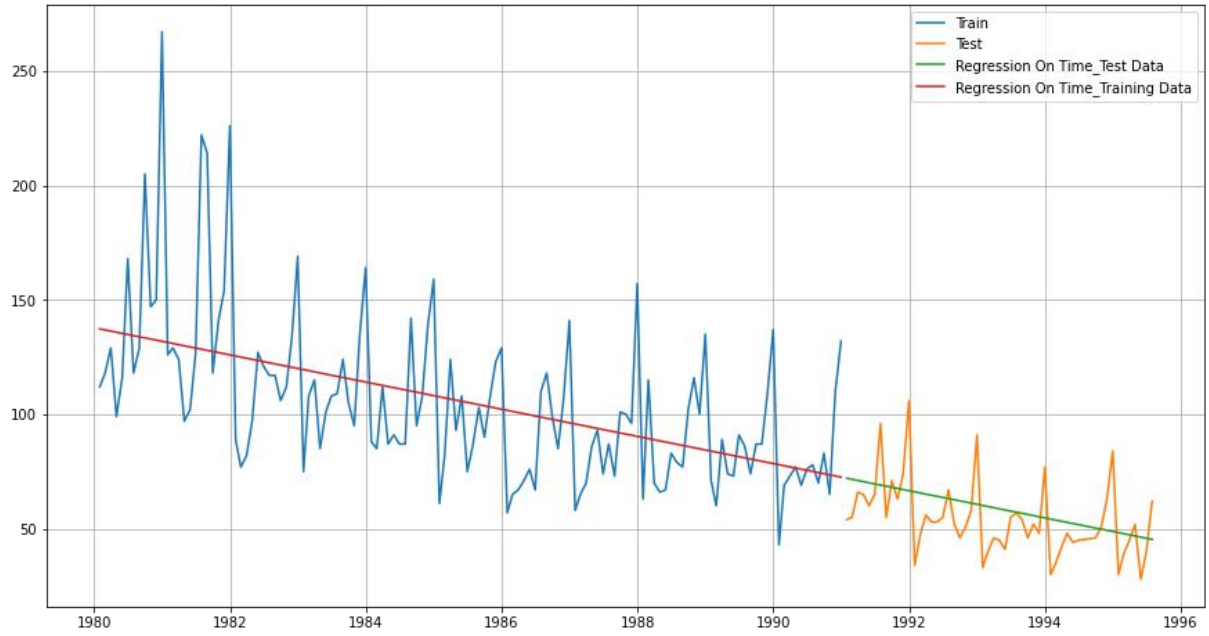

Fig16

Model Evaluation:

For RegressionOnTime forecast on the Training Data, RMSE is 30.718 and MAPE is 21.22

For RegressionOnTime forecast on the Test Data, RMSE is 15.269 and MAPE is 22.82

Model 2: Naive Approach

Train set Head after Naive Predictions:

Date	
1980-01-31	132.0
1980-02-29	132.0
1980-03-31	132.0
1980-04-30	132.0
1980-05-31	132.0

Test Set Head after Naive Predictions:

Date	
1991-01-31	132.0
1991-02-28	132.0
1991-03-31	132.0
1991-04-30	132.0
1991-05-31	132.0

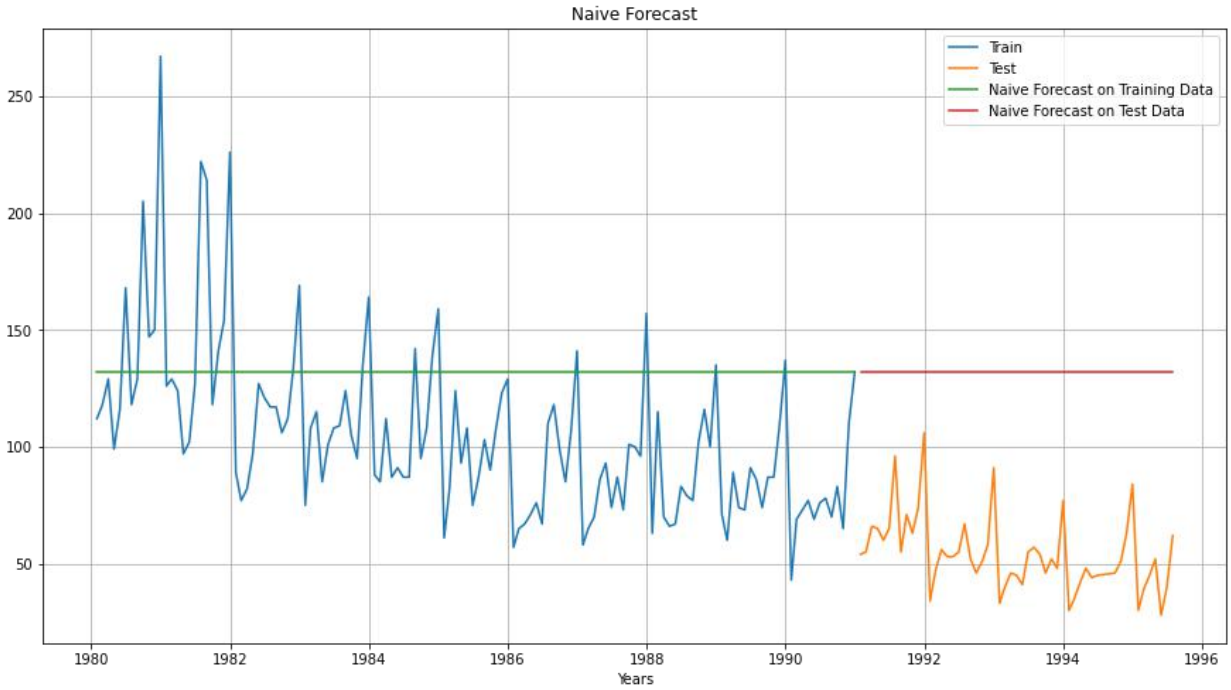

Fig 17

Model Evaluation:

For Naive Model forecast on the Training Data, RMSE is 45.064 and MAPE is 36.38

For Naive forecast on the Test Data, RMSE is 79.719 and MAPE is 145.10

Method 3: Simple Average

Simple Avg Train set:

Date	Rose	mean_forecast
1980-01-31	112.0	104.939394
1980-02-29	118.0	104.939394
1980-03-31	129.0	104.939394
1980-04-30	99.0	104.939394
1980-05-31	116.0	104.939394

Simple Avg Test Set:

Date	Rose	mean_forecast
1991-01-31	54.0	104.939394
1991-02-28	55.0	104.939394
1991-03-31	66.0	104.939394
1991-04-30	65.0	104.939394
1991-05-31	60.0	104.939394

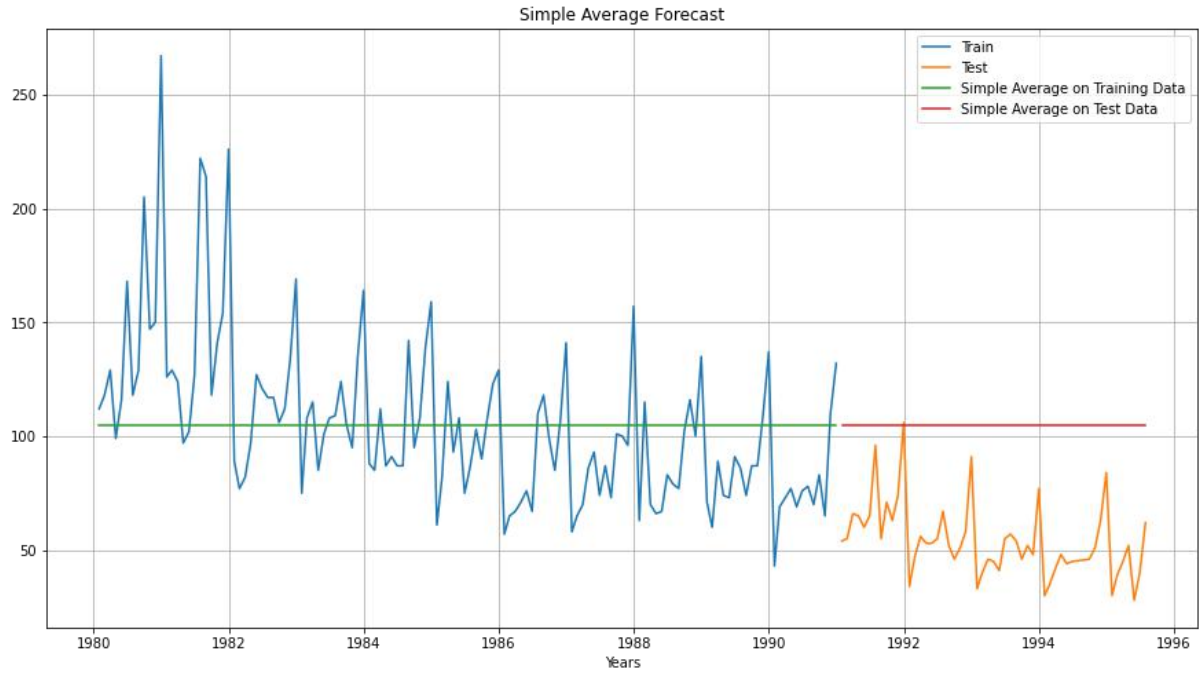

Fig 18

Model Evaluation of Simple Average:

For Simple Average Model forecast on the Training Data, RMSE is 36.034 and MAPE is 25.39

For Simple Average forecast on the Test Data, RMSE is 53.461 and MAPE is 94.93

Method 4: Moving Average(MA)

Moving Avg Train Set:

Date	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

Trailing data set:

Date	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN

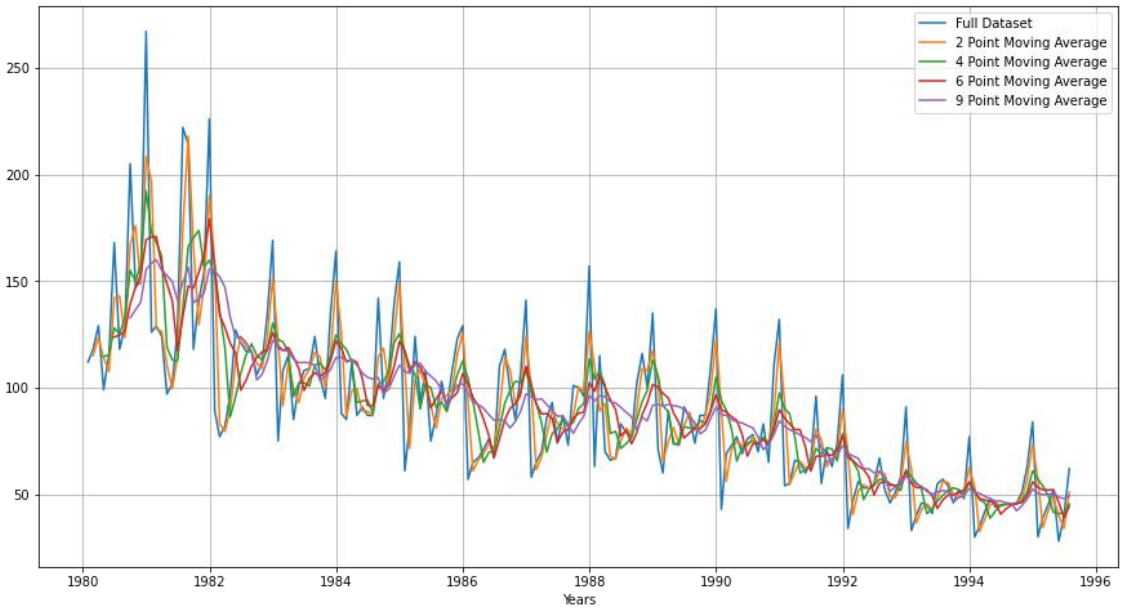

Fig 19

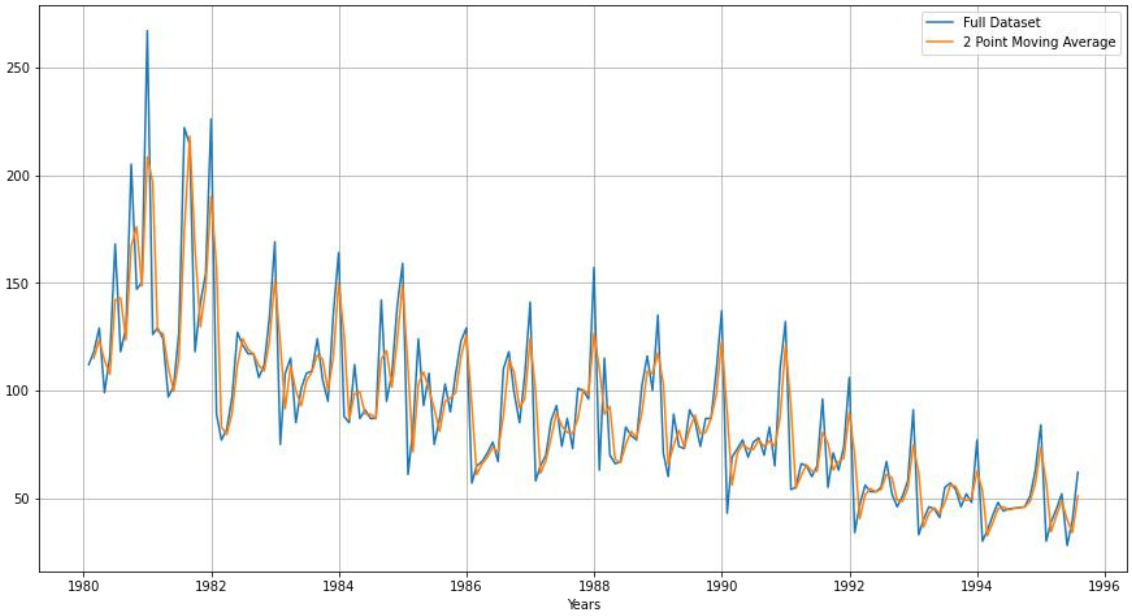

Fig 20

We see 2 point moving average gives best results here:

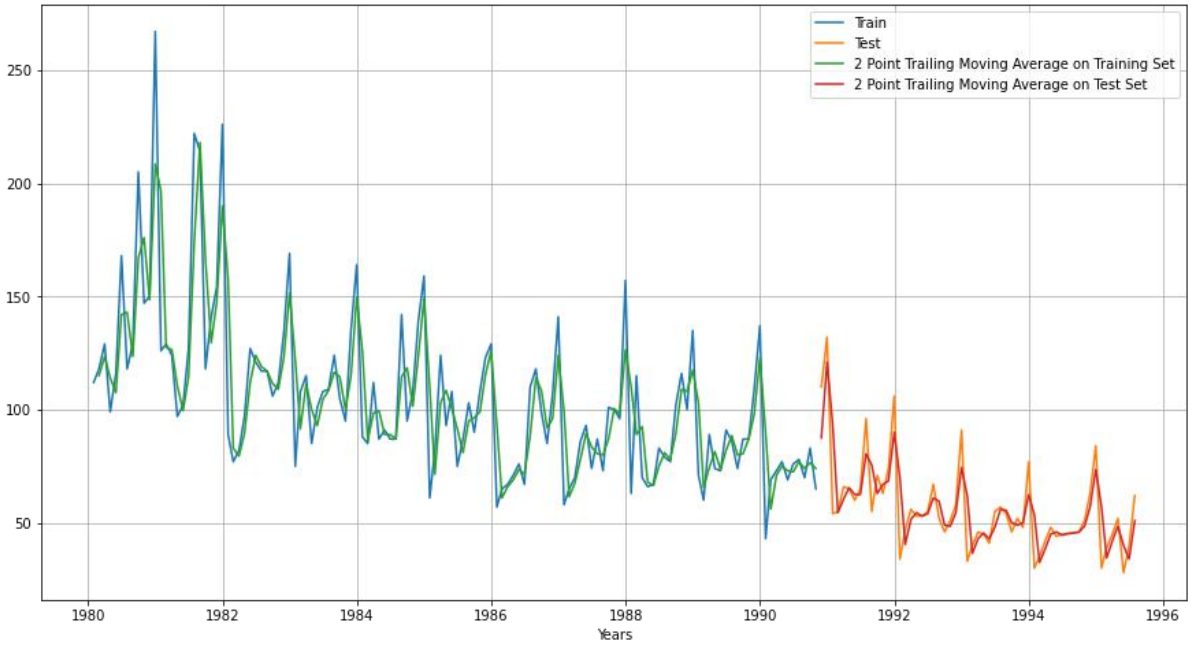

Fig 21

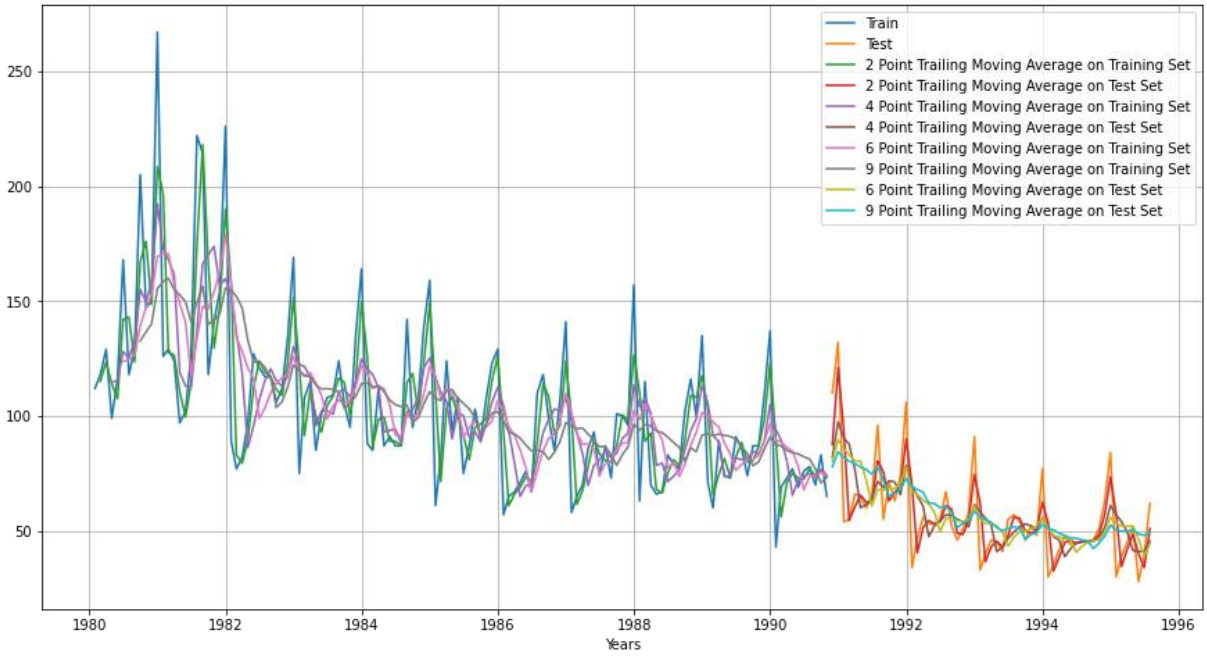

Fig 22

For 2 point Moving Average Model forecast on the Testing Data, RMSE is 11.529 and MAPE is 13.54
For 4 point Moving Average Model forecast on the Testing Data, RMSE is 14.451 and MAPE is 19.49
For 6 point Moving Average Model forecast on the Testing Data, RMSE is 14.566 and MAPE is 20.82

For 9 point Moving Average Model forecast on the Testing Data, RMSE is 14.728 and MAPE is 21.01

Method 5: Simple Exponential Smoothing

SES Test Set after forecast:

Date	Rose	predict
1991-01-31	54.0	87.104983
1991-02-28	55.0	87.104983
1991-03-31	66.0	87.104983
1991-04-30	65.0	87.104983
1991-05-31	60.0	87.104983

Smoothing Level: 0.09

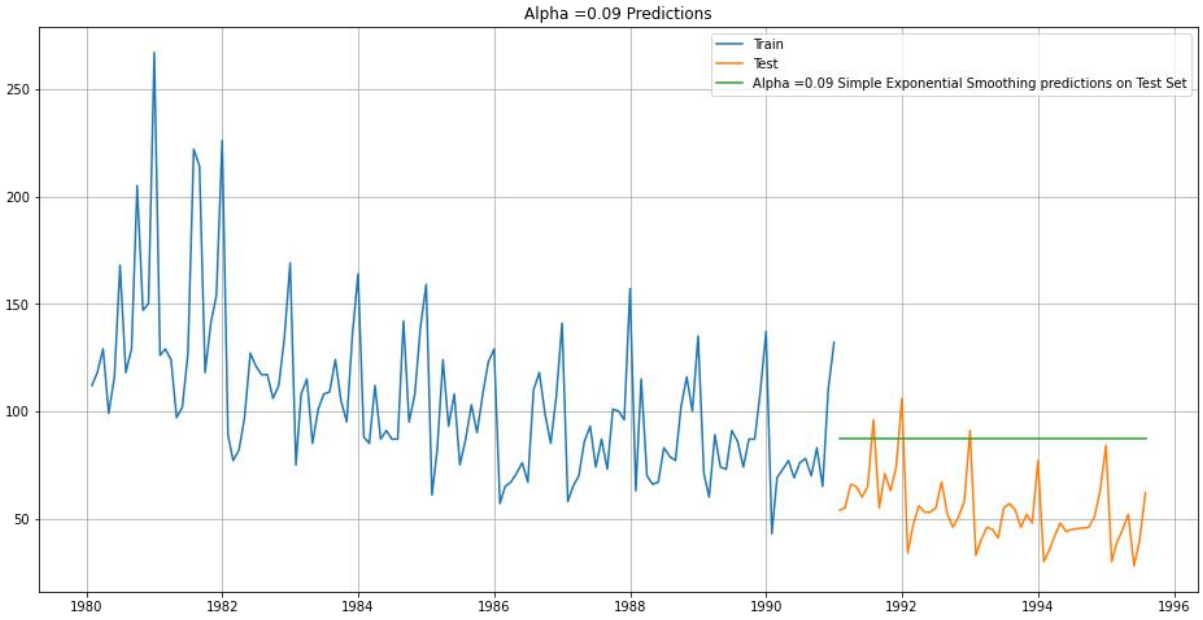

Fig 23

Model Evaluation for alpha = 0.04:

For Alpha =0.09 Simple Exponential Smoothing Model forecast on the Training Data, RMSE is 36.796 and MAPE is 63.88

After SES Tuning:

	Alpha Values	Train RMSE	Test RMSE	Test MAPE
0	0.3	32.470164	47.504821	83.71
1	0.4	33.035130	53.767406	95.50
2	0.5	33.682839	59.641786	106.81
3	0.6	34.441171	64.971288	117.04
4	0.7	35.323261	69.698162	126.07
5	0.8	36.334596	73.773992	133.83
6	0.9	37.482782	77.139276	140.22

The RMSE for Alpha = 0.3 was not better than Alpha= 0.09

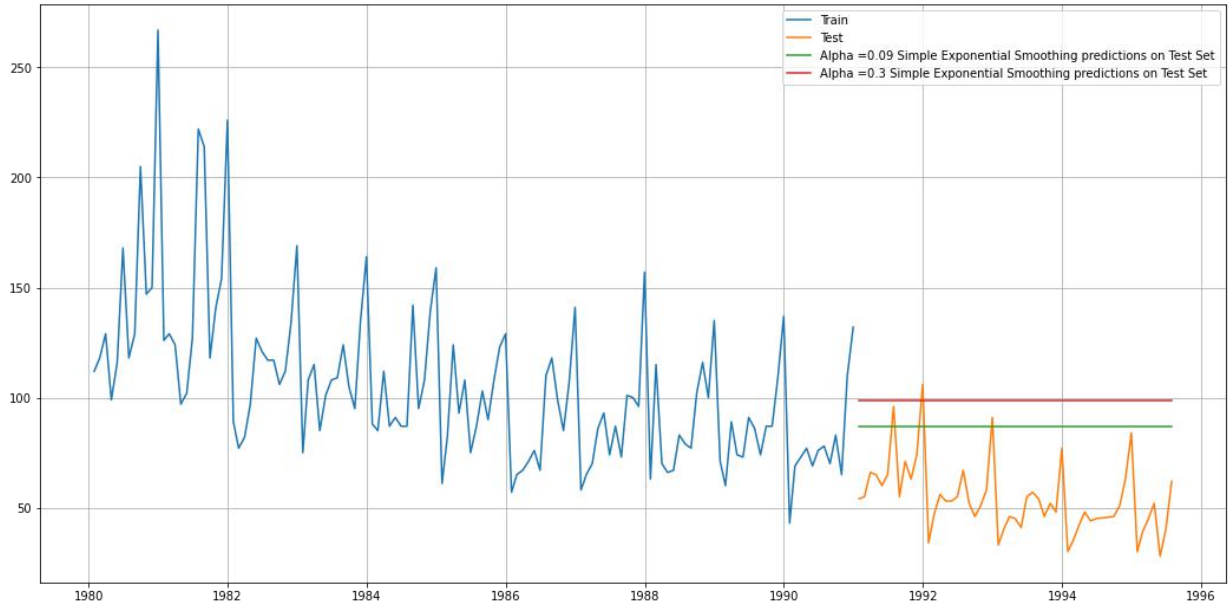

Fig 24

Method 6: Double Exponential Smoothing (Holt's Model)

Best Values after Predictions:

	Alpha Values	Beta Values	Train RMSE	Test RMSE	Test MAPE
0	0.3	0.3	35.944983	265.567594	442.50
8	0.4	0.3	36.749123	339.306534	565.42
1	0.3	0.4	37.393239	358.750942	593.91
16	0.5	0.3	37.433314	394.272629	657.17
24	0.6	0.3	38.348984	439.296033	732.29

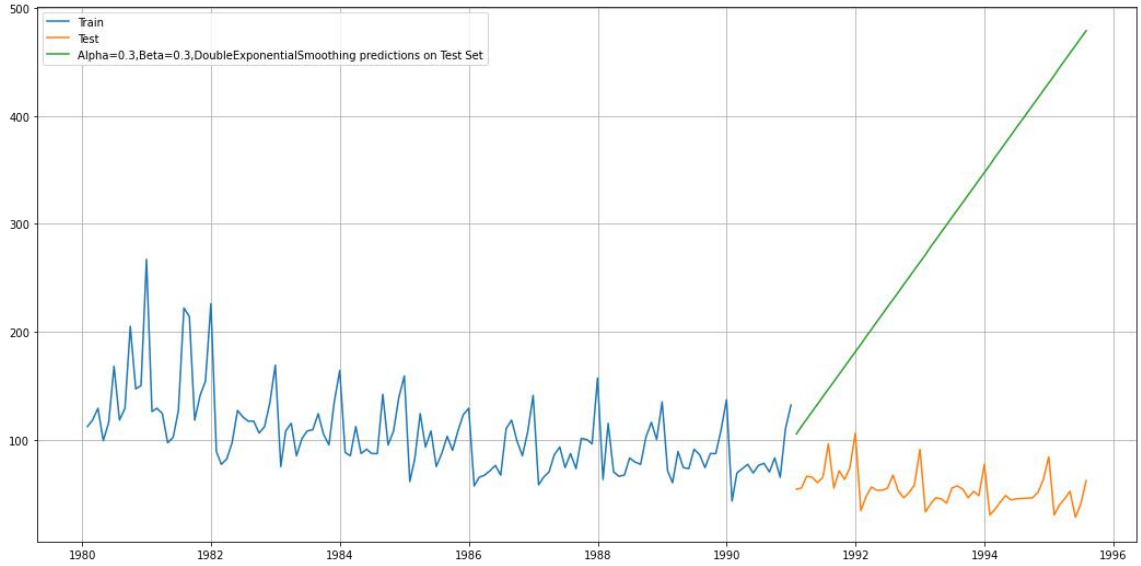

Fig 25

Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

''smoothing_level': 0.08485622209289158,
'smoothing_trend': 0.0005280630369796539,
'smoothing_seasonal': 0.00676452679451911

Train set after fitting values:

Date	Rose	auto_predict
1980-01-31	112.0	115.401460
1980-02-29	118.0	126.870859
1980-03-31	129.0	133.689243
1980-04-30	99.0	121.941182
1980-05-31	116.0	128.235113

Test Set after prediction:

Date	Rose	auto_predict
1991-01-31	54.0	42.605820
1991-02-28	55.0	54.376558
1991-03-31	66.0	61.934184
1991-04-30	65.0	50.608812
1991-05-31	60.0	58.838643

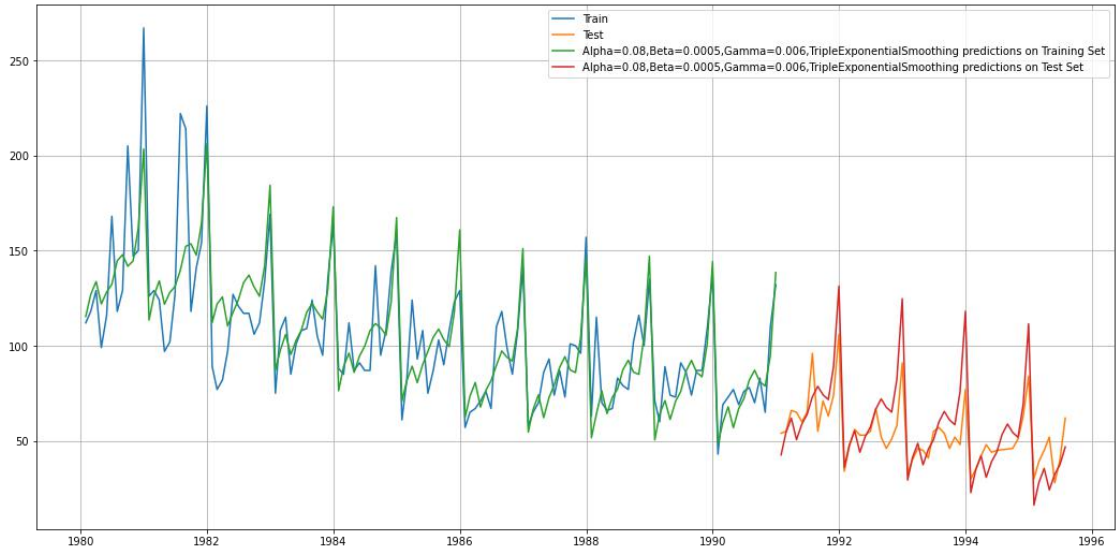

Fig 26

Model Evaluation:

For Alpha: 0.08,Beta: 0.0005 and Gamma:0.006, Triple Exponential Smoothing Model forecast on the Training Data, RMSE is 19.542 MAPE is 13.37

For Alpha: 0.08,Beta: 0.0005 and Gamma:0.006,Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 14.257 MAPE is 19.23

Model Evaluation after tuning:

Best Params after Tuning:

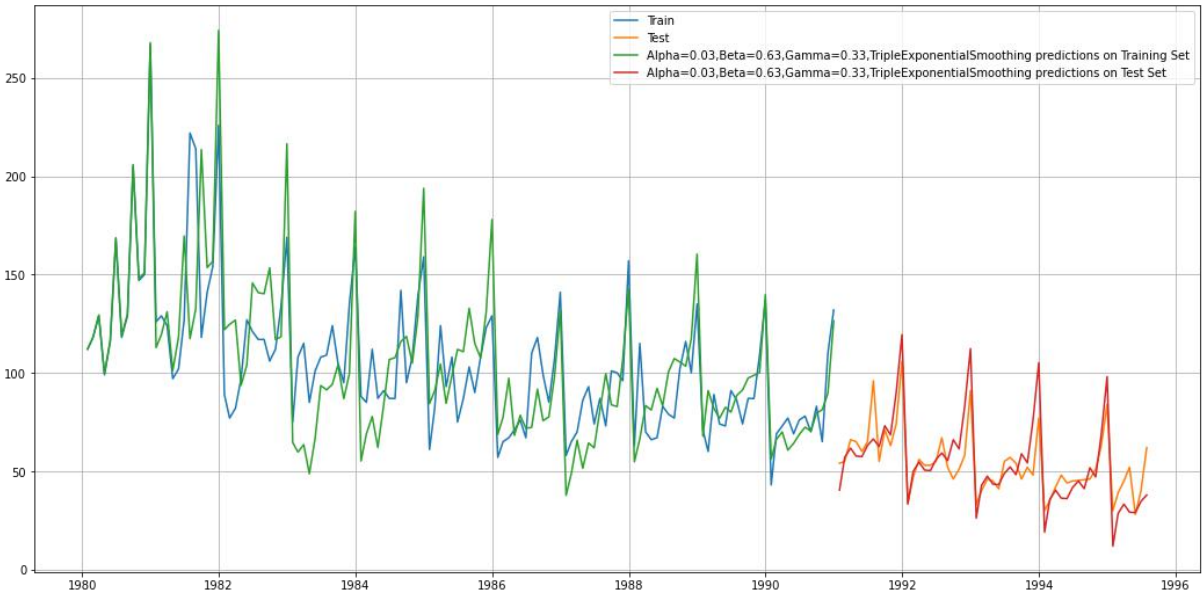
Train
Test
Alpha=0.03,Beta=0.63,Gamma=0.33,TripleExponentialSmoothing predictions on Training Set
Alpha=0.03,Beta=0.63,Gamma=0.33,TripleExponentialSmoothing predictions on Test Set
250
200
150
100
50
0
1980
1982
1984
1986
1988
1990
1992
1994
1996

Fig 27

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.

Answer:

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:
• H0 : The Time Series has a unit root and is thus non-stationary.
• H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.(0.05)

Check for stationarity of the Whole Data Time Series:

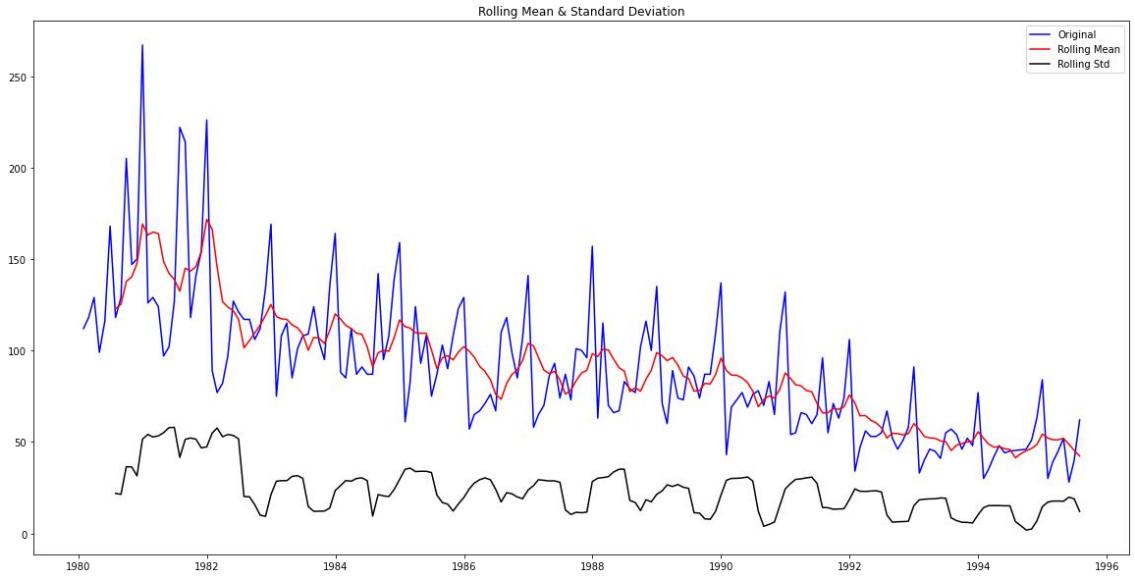

Fig 28

Results 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

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

Let us take a difference of order 1 and check whether the Time Series is stationary or not:

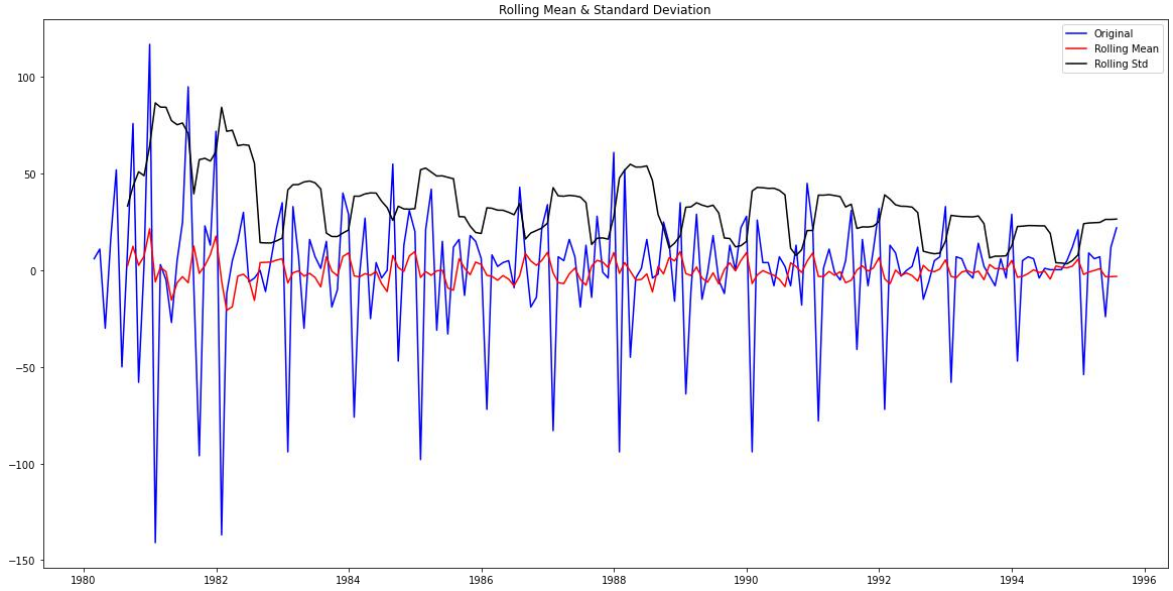

Fig 29

Results 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

We see that at alpha = 0.05 when taking difference of order 1 the Time Series is indeed stationary.

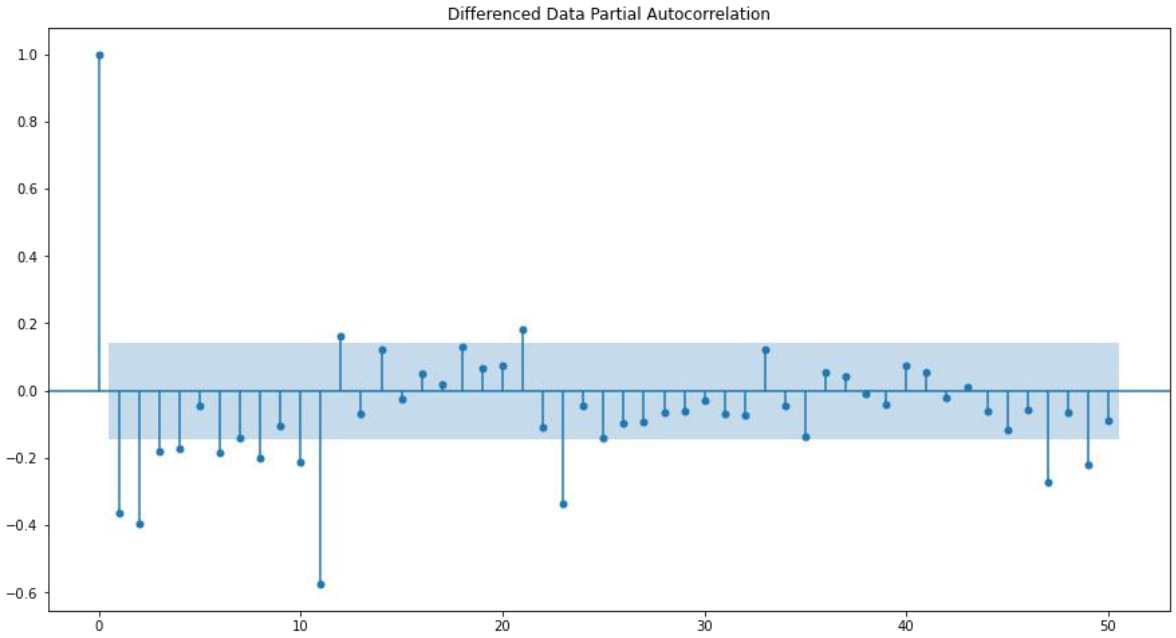
Differenced Data Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
10
20
30
40
50

Fig 30

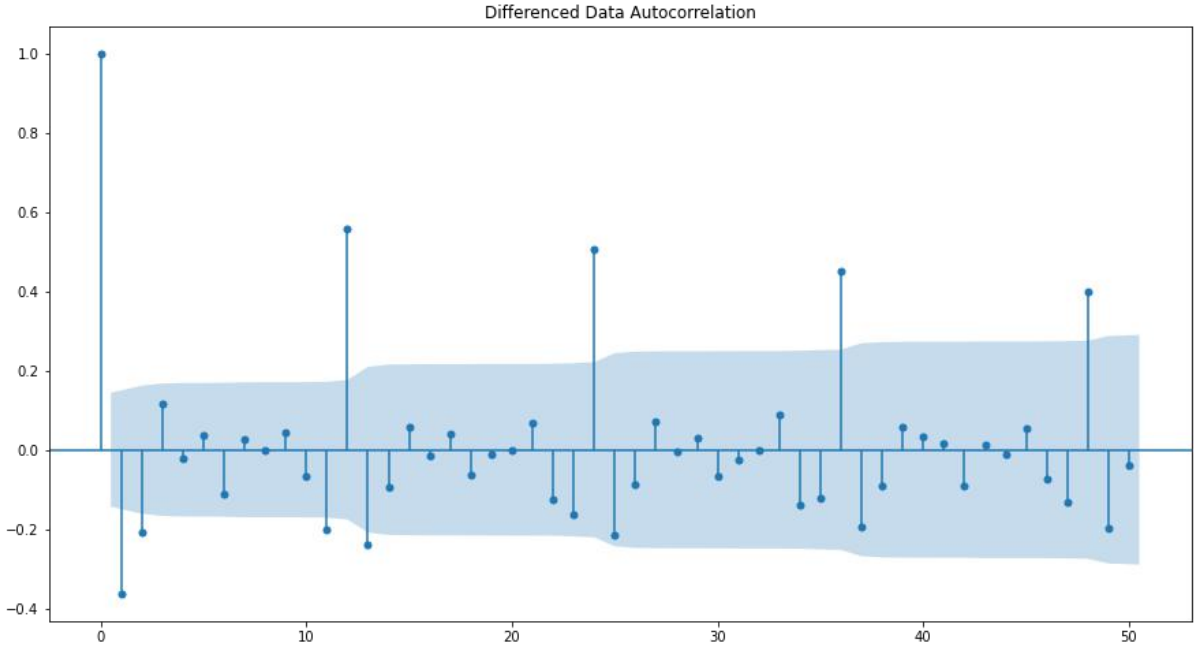
Differenced Data Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
40
50

Fig 31

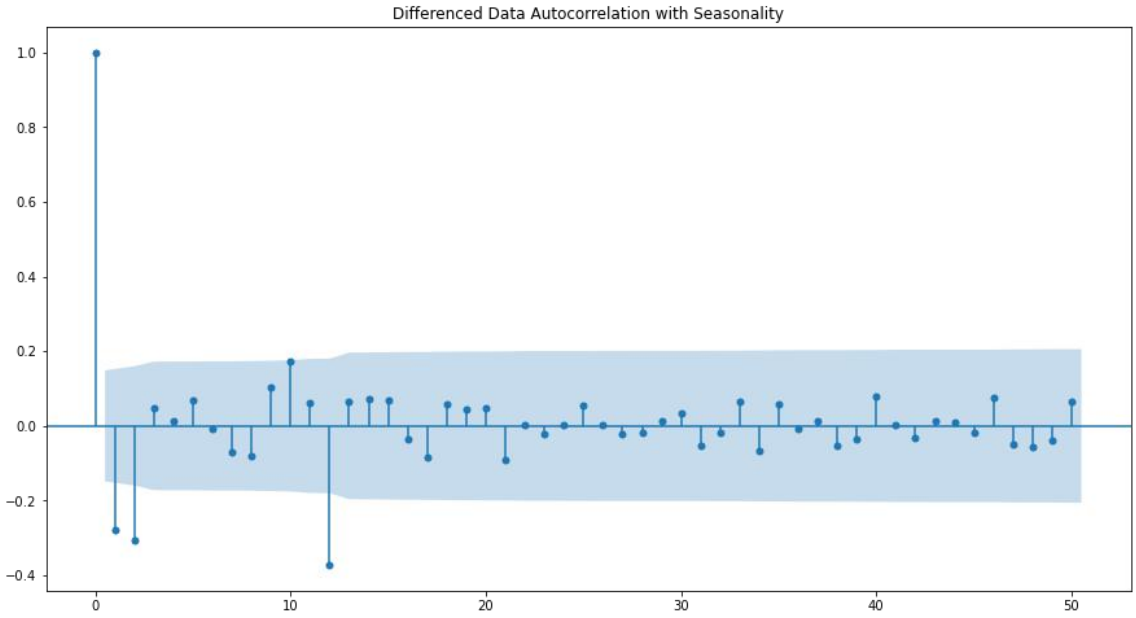

Fig 32

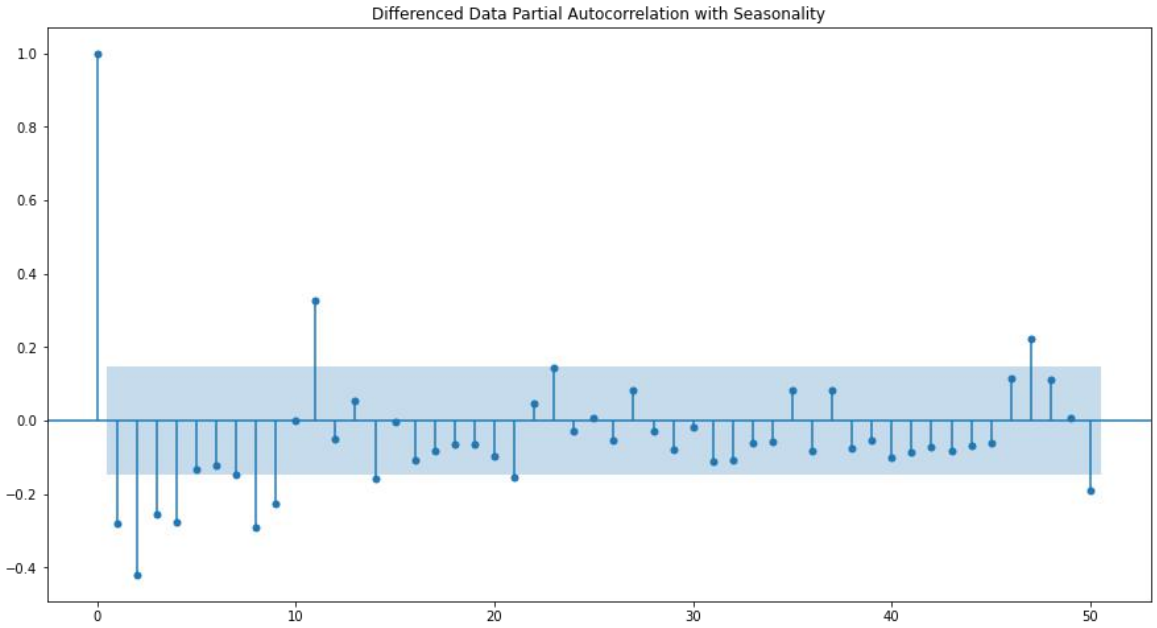

Fig 33

We observe the ACF plot for Sparkling Sales and observe seasonality at intervals 12, hence we run the Automated SARIMA models at seasonality 12.

When we build manual ARIMA model for Rose Sales based on the ACF and PACF plots. Hence we chose the AR parameter p = 4 and P = 0, Moving average parameter q = 2 and Q = 0 and d =1 and D= 1 based on the plots.

Check for stationarity of the Training Data Time Series:

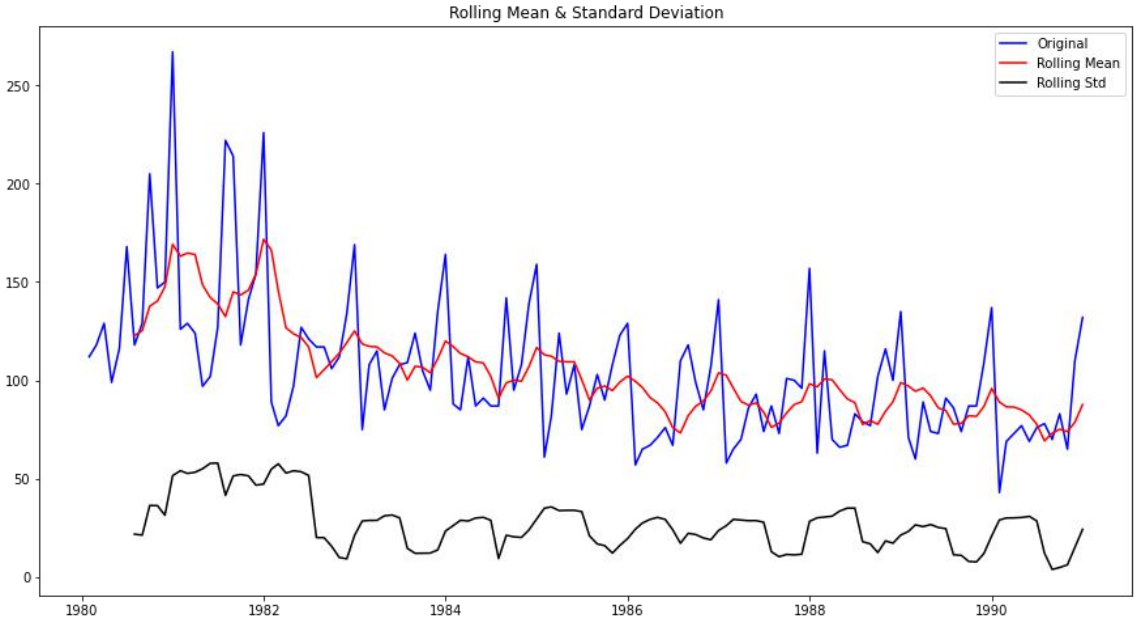

Fig 34

Results of Dickey-Fuller Test:

Test Statistic	-1.208926
p-value	0.669744
#Lags Used	12.000000
Number of Observations Used	119.000000
Critical Value (1%)	-3.486535
Critical Value (5%)	-2.886151
Critical Value (10%)	-2.579896

We see that the Train series is not stationary at alpha = 0.05.

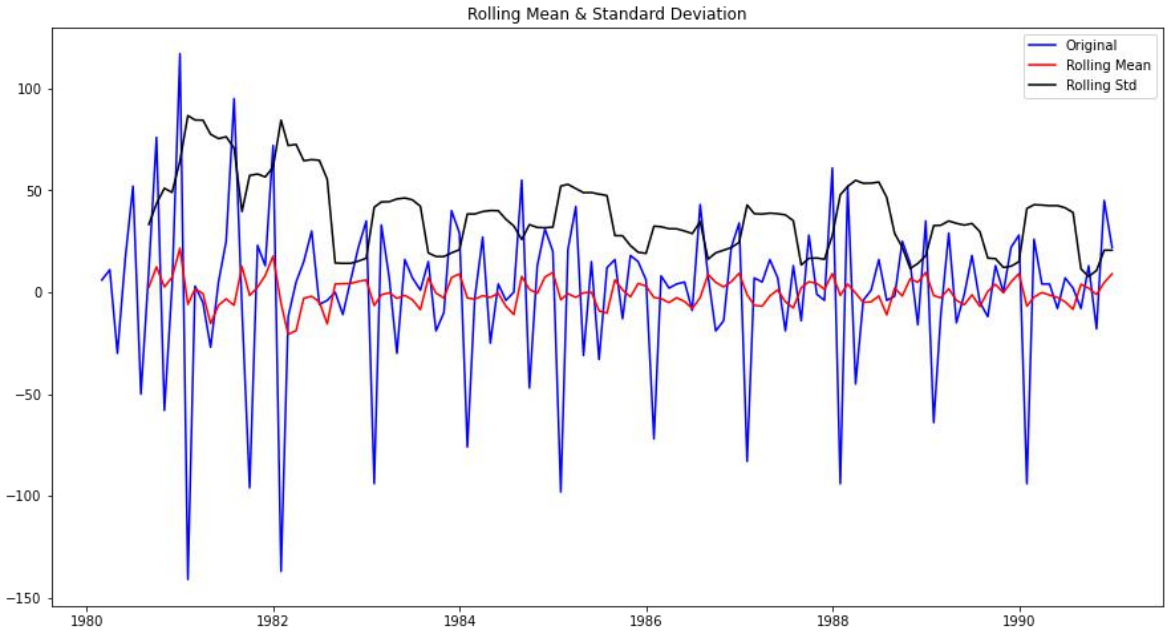

Fig 35

Results of Dickey-Fuller Test:

Test Statistic	-8.005007e+00
p-value	2.280104e-12
#Lags Used	1.100000e+01
Number of Observations Used	1.190000e+02
Critical Value (1%)	-3.486535e+00
Critical Value (5%)	-2.886151e+00
Critical Value (10%)	-2.579896e+00

We see that after taking a difference of order 1 the series have become stationary at alpha = 0.05

Note: If the series is non-stationary, stationarize the Time Series by taking a difference of the Time Series. Then we can use this particular differenced series to train the ARIMA models. We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there

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.

Answer:

1. Automated version of ARIMA for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (0, 1, 4)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (1, 1, 4)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (2, 1, 4)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
Model: (3, 1, 4)
Model: (4, 1, 0)
Model: (4, 1, 1)
Model: (4, 1, 2)
Model: (4, 1, 3)
Model: (4, 1, 4)
```

Best ARIMA Params by sorting lowest AIC to top:

	param	AIC
18	(3, 1, 3)	1273.194238

19	(3, 1, 4)	1274.337930
2	(0, 1, 2)	1276.835373
7	(1, 1, 2)	1277.359223
6	(1, 1, 1)	1277.775748
3	(0, 1, 3)	1278.074260

ARIMA SUMMARY:

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(3, 1, 3)	Log Likelihood	-628.597
Method:	css-mle	S.D. of innovations	28.356
Date:	Sun, 20 Mar 2022	AIC	1273.194
Time:	23:17:26	BIC	1296.196
Sample:	02-29-1980 - 12-31-1990	HQIC	1282.541

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4906	0.088	-5.546	0.000	-0.664	-0.317
ar.L1.D.Rose	-0.7241	0.086	-8.379	0.000	-0.893	-0.555
ar.L2.D.Rose	-0.7215	0.087	-8.307	0.000	-0.892	-0.551
ar.L3.D.Rose	0.2766	0.086	3.226	0.001	0.109	0.445
ma.L1.D.Rose	-0.0154	0.045	-0.345	0.730	-0.103	0.072
ma.L2.D.Rose	0.0154	0.045	0.346	0.729	-0.072	0.103
ma.L3.D.Rose	-1.0000	0.046	-21.811	0.000	-1.090	-0.910

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.5011	-0.8661j	1.0007	-0.3335
AR.2	-0.5011	+0.8661j	1.0007	0.3335
AR.3	3.6111	-0.0000j	3.6111	-0.0000
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.4923	-0.8704j	1.0000	-0.3319
MA.3	-0.4923	+0.8704j	1.0000	0.3319

Test rmse for arima is 15.989214539681338
Test mape for arima is 26.09

2. Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC):

We observe the ACF plot for Rose Sales and observe seasonality at intervals 12, hence we run the Automated SARIMA models at seasonality 12.

Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)

Best SARIMA Params by sorting lowest AIC to top:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686920
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

SARIMA SUMMARY:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-436.969
Date:	Sun, 20 Mar 2022	AIC	887.938
Time:	23:18:41	BIC	906.448
Sample:	0	HQIC	895.437
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]

ma.L1	-0.8427	189.826	-0.004	0.996	-372.895	371.209
ma.L2	-0.1573	29.823	-0.005	0.996	-58.609	58.294
ar.S.L12	0.3467	0.079	4.375	0.000	0.191	0.502
ar.S.L24	0.3023	0.076	3.996	0.000	0.154	0.451
ma.S.L12	0.0767	0.133	0.577	0.564	-0.184	0.337
ma.S.L24	-0.0726	0.146	-0.498	0.618	-0.358	0.213
sigma2	251.3137	4.77e+04	0.005	0.996	-9.33e+04	9.38e+04

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	2.33
Prob(Q):	0.75	Prob(JB):	0.31
Heteroskedasticity (H):	0.88	Skew:	0.37
Prob(H) (two-sided):	0.70	Kurtosis:	3.03

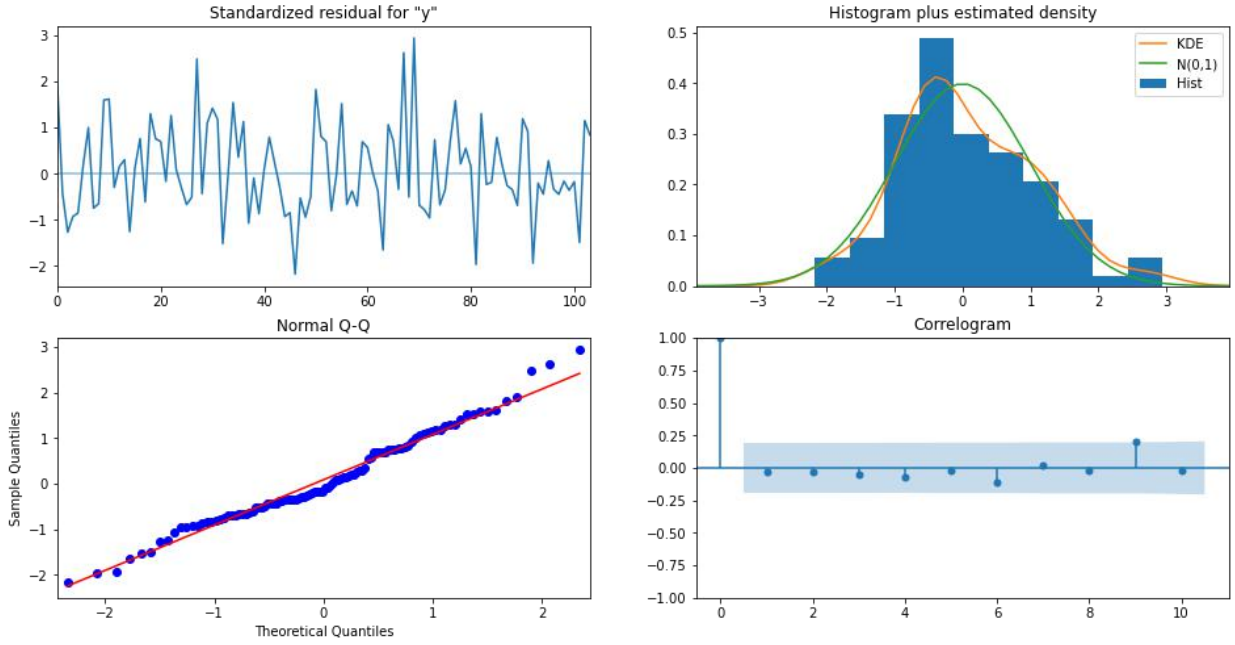

Fig 36

Inference from Model diagnostics confirms that:

- The model residuals are normally distributed
- Standardized residual – Do not display any obvious seas onality
- Histogram plus estimated density - The KDE plot of th e residuals is similar with the normal distribution, hence the model residuals are normally distributed based
- Normal Q-Q plot – There is an ordered distribution of residuals (blue dots) following the linear trend of the sampl es taken from a standard normal distribution with N(0, 1)

- Correlogram – The time series residuals have low correlation with lagged versions of itself

Test rmse for SARIMA is 26.928361739818897
Test mape for SARIMA is 46.6

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.

Answer:

1. Manual ARIMA Model

When we build manual ARIMA model for Rose Sales based on the ACF and PACF plots. Hence we chose the AR parameter p = 1, Moving average parameter q = 2 and d =1 based on the ACF/PACF plots.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(1, 1, 2)	Log Likelihood	-633.680
Method:	css-mle	S.D. of innovations	29.977
Date:	Sun, 20 Mar 2022	AIC	1277.359
Time:	23:18:58	BIC	1291.735
Sample:	02-29-1980 - 12-31-1990	HQIC	1283.201

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4921	0.079	-6.224	0.000	-0.647	-0.337
ar.L1.D.Rose	-0.4163	0.222	-1.874	0.061	-0.852	0.019
ma.L1.D.Rose	-0.3616	0.189	-1.913	0.056	-0.732	0.009
ma.L2.D.Rose	-0.6383	0.186	-3.437	0.001	-1.002	-0.274

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-2.4022	+0.0000j	2.4022	0.5000
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	-1.5665	+0.0000j	1.5665	0.5000

Test rmse for Manual arima is 15.284561918351923
Test mape for Manual arima is 22.63

2. Manual SARIMA Model

We observe the ACF plot for Rose Sales and observe seasonality at intervals 12, hence we run the SARIMA models at seasonality 12.

When we build manual SARIMA model for Rose Sales based on the ACF and PACF plots. Hence we chose the AR parameter p = 4 and P = 0, Moving average parameter q = 2 and Q = 2 and d =1 and D= 1 based on the plots.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-384.369
Date:	Sun, 20 Mar 2022	AIC	786.737
Time:	23:19:05	BIC	809.433
Sample:	0 - 132	HQIC	795.898
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8967	0.132	-6.814	0.000	-1.155	-0.639
ar.L2	0.0165	0.171	0.097	0.923	-0.319	0.352
ar.L3	-0.1132	0.174	-0.650	0.515	-0.454	0.228
ar.L4	-0.1598	0.116	-1.380	0.168	-0.387	0.067
ma.L1	0.1508	0.174	0.866	0.387	-0.191	0.492
ma.L2	-0.8492	0.164	-5.166	0.000	-1.171	-0.527
ma.S.L12	-0.3907	0.102	-3.848	0.000	-0.590	-0.192
ma.S.L24	-0.0887	0.091	-0.977	0.329	-0.267	0.089
sigma2	238.9649	0.001	2.02e+05	0.000	238.963	238.967

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	0.01
Prob(Q):	0.80	Prob(JB):	0.99
Heteroskedasticity (H):	0.76	Skew:	-0.01
Prob(H) (two-sided):	0.46	Kurtosis:	3.06

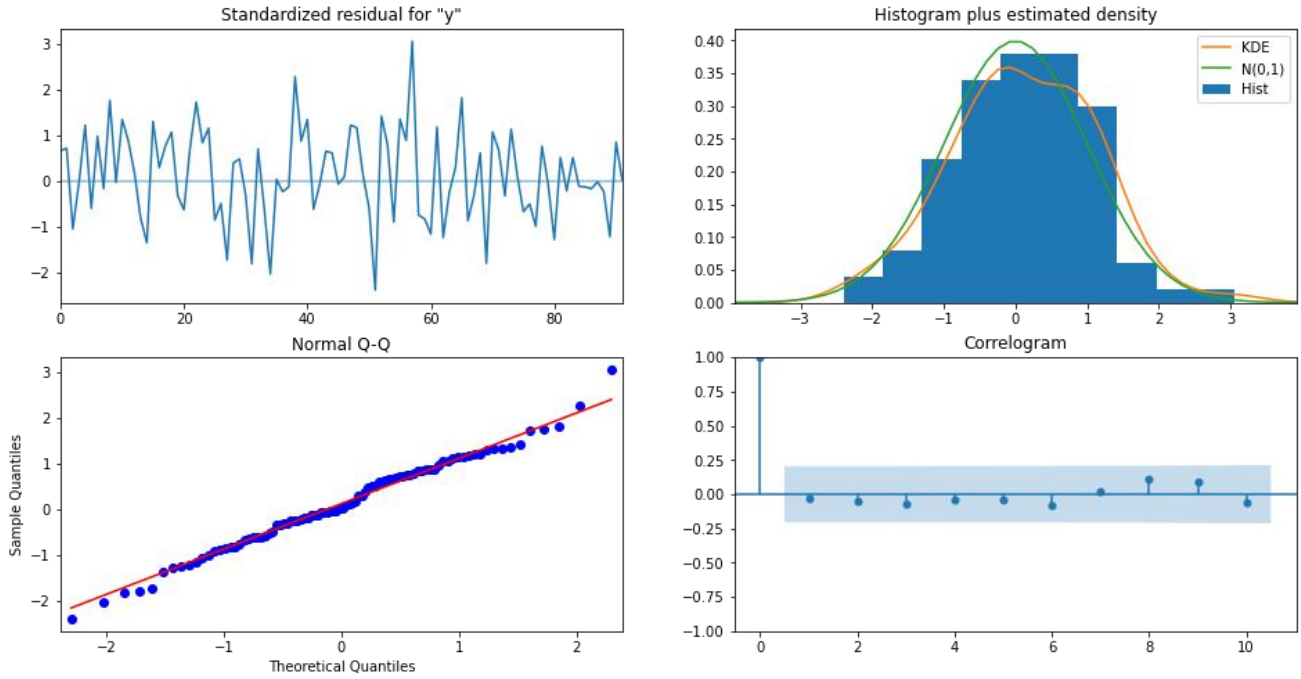

Fig 37

Model diagnostics confirms that the model residuals are normal y distributed. Standardized residual do not display any obviouss easonality,
Histogram plus estimated density - The KDE plot has normal distribution ,
Normal Q-Q plot – There is an ordered distribution of residual s (blue dots) following the linear trend ,
Correlogram – The time series residuals have low correlation with lagged versions of itself

Test rmse for Manual sarima is 15.377251633894176
Test mape for Manual sarima is 22.16

I built various models by tweaking the parameters by looking at the ACF and PACF plots of which I showed the best mode l in the report.

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

Answer:

Sorted by RMSE values on the Test Data:

	Test RMSE	Test MAPE
2pointTrailingMovingAverage	11.529278	13.54
Alpha=0.08,Beta=0.0005,Gamma=0.006,TripleE	14.257122	19.23
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
RegressionOnTime	15.268955	22.82
ARIMA(1,1,2)	15.284562	22.63
SARIMA(4, 1, 2)(0, 1, 2, 12)	15.377252	22.16
ARIMA (3,1,3)	15.989215	26.09
Alpha=0.03,Beta=0.63,Gamma=0.33,TripleExpo	16.012742	23.64
SARIMA(0, 1, 2)(2, 0, 2, 12)	26.928362	46.60
Alpha =0.09 Simple Exponential Smoothing Model	36.796227	63.88
SimpleAverageModel	53.460570	94.93
NaiveModel	79.718773	145.10
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.567594	442.50

Sorted by MAPE values on the Test Data:

	Test RMSE	Test MAPE
2pointTrailingMovingAverage	11.529278	13.54
Alpha=0.08,Beta=0.0005,Gamma=0.006,TripleExp	14.257122	19.23
4pointTrailingMovingAverage	14.451403	19.49
6pointTrailingMovingAverage	14.566327	20.82
9pointTrailingMovingAverage	14.727630	21.01
SARIMA(4, 1, 2)(0, 1, 2, 12)	15.377252	22.16
ARIMA(1,1,2)	15.284562	22.63
RegressionOnTime	15.268955	22.82
Alpha=0.03,Beta=0.63,Gamma=0.33,TripleExpo.	16.012742	23.64
ARIMA (3,1,3)	15.989215	26.09
SARIMA(0, 1, 2)(2, 0, 2, 12)	26.928362	46.60
Alpha =0.09 Simple Exponential Smoothing Model	36.796227	63.88
SimpleAverageModel	53.460570	94.93
NaiveModel	79.718773	145.10
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.567594	442.50

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

Answer:

[Although we are seeing that the best model as per RMSE is t he 2pointTrailingMovingAverage as it is giving us the least R MSE value. But the moving average models are actually quite naive for prediction model and assumes that the trend and seas onality components of the time series have already been remov ed or adjusted for. Hence we will going to choose the second best model which comes out to be Triple Exponential Model. I ts RMSE value is close to the 2pointTrailingMovingAverage va lue hence we can choose this model as well.]

We see that the best model is the Triple Exponential Smoothin g with additive seasonality with the parameters alpha = 0.08, beta = 0.0005 and gamma = 0.006.

RMSE when Model run on full dataset: 17.723677428972273
MAPE when Model run on full dataset : 13.74

Predictions on 12 months into future:

1995-08-31	49.268273
1995-09-30	46.138792
1995-10-31	44.885229
1995-11-30	59.486958
1995-12-31	97.791122
1996-01-31	13.357765
1996-02-29	23.642647
1996-03-31	31.177139
1996-04-30	23.957475
1996-05-31	27.268934
1996-06-30	32.760269
1996-07-31	43.313327

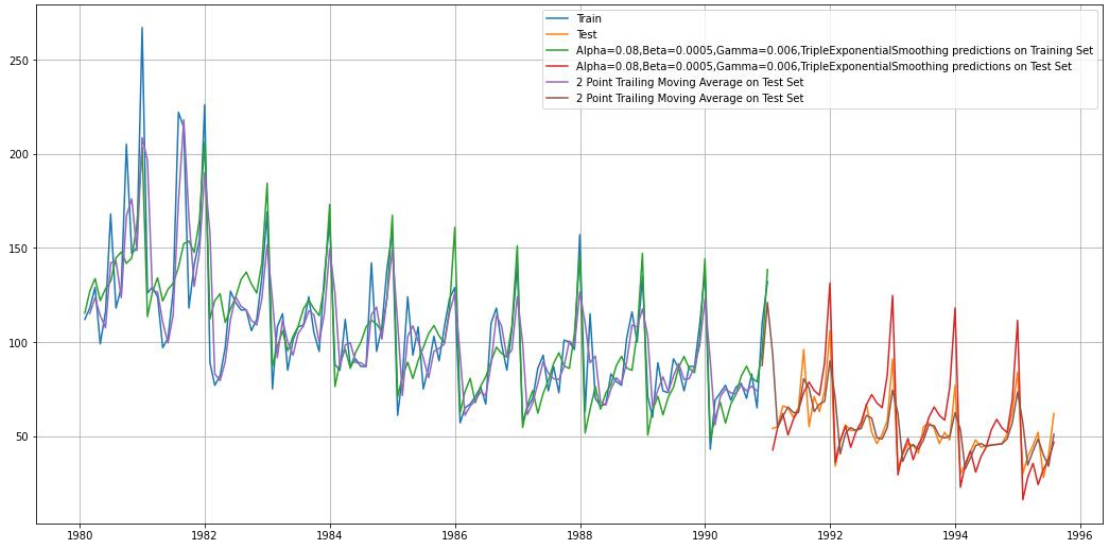

Fig 38

I have calculated the upper and lower confidence bands at 95% confidence level.

The percentile function under numpy lets us calculate these and adding and subtracting from the predictions gives us the necessary confidence bands for the predictions.

	lower_CI	prediction	upper_ci
1995-08-31	33.082166	49.268273	136.525750
1995-09-30	29.952685	46.138792	133.396269
1995-10-31	28.699122	44.885229	132.142706
1995-11-30	43.300851	59.486958	146.744434
1995-12-31	81.605014	97.791122	185.048598

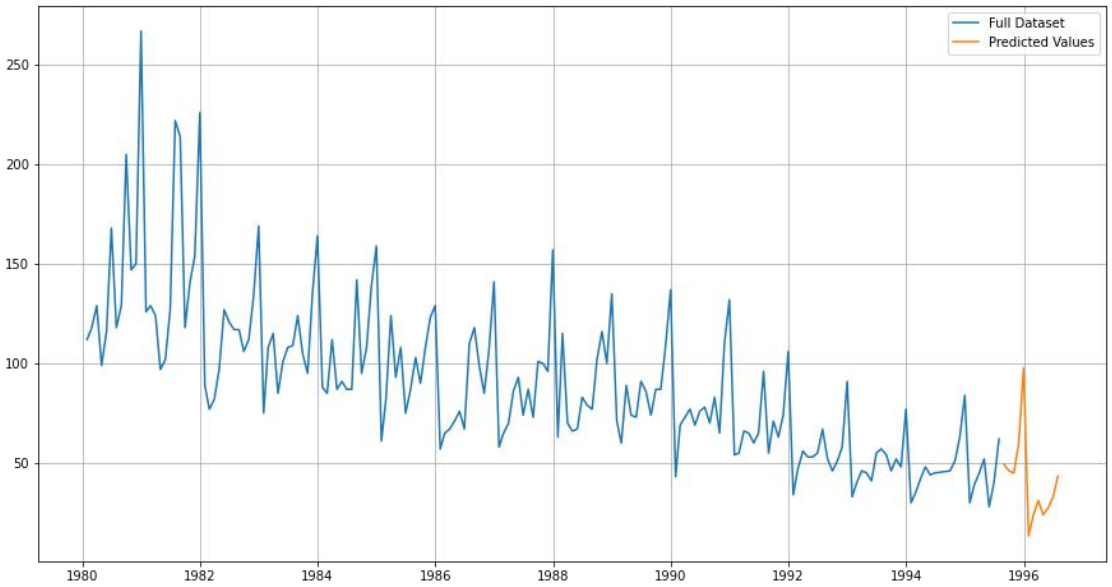
Full Dataset
Predicted Values
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996

Fig 39

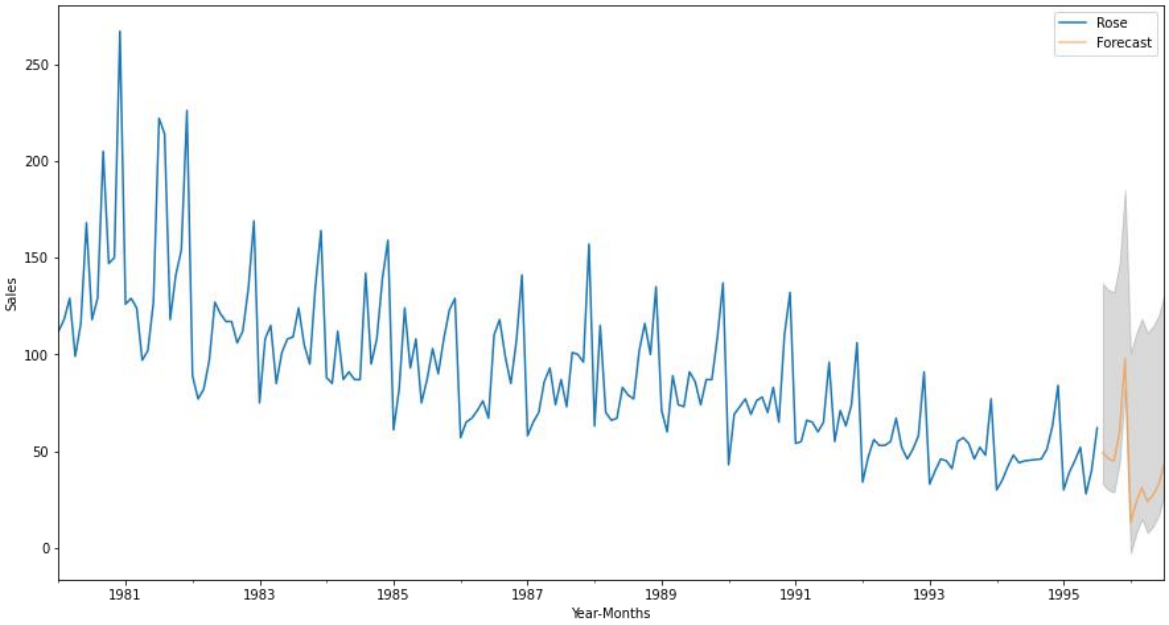
Rose
Forecast
Sales
250
200
150
100
50
0
1981
1983
1985
1987
1989
1991
1993
1995
Year-Months

Fig 40

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

Answer:

1. Triple Exponential Model is performing best in this case giving us the least error.

2. Looking at the bar plot, we can see that on December months the sales are highest. We can use this insights to increase our sales further.

3. We can introduce certain offers in November, December months to attaract more customers.

4. On Mondays mean sales of the wine is highest. We can give certain offers to attract more customers.

5. Year 1981 has the highest sales recorded till data. We can go back to find out the reasons to which pushed the sales so much.

6. We can also see in the year 1981, 1983 and 1990 the Wine sales in the month of October, November remained high after which it started declining, needs to pay attention for those months sales as well.

7.Looking at the prediction, we can say that the sales figure will be more or less same as that of previous year. Hence some important measures have to be taken to increase the trend. As the trend has been declining throughout the years.

8.Both the models are built considering the Trend and Seasonality in to account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years.

9.The company should use the prediction results and capitalize on the high demand seasons and ensure to source and supply t he high demand.

10.The company should use the prediction results to plan the l ow demand seasons to stock as per the demand.

11. The price of rose wine may be expensive than sparkling s o seasonal discounts can help improve the sales of rose wine. Products that are discounted should be highlighted so consumer s can see the savings prominently. Discounts can compel consu mers to buy.

12.As we know how the seasonality is in the prediction compa ny cannot have the same stock through the year. You should c reate a dynamic consumer experience with fresh point-of-sale materials and wellstocked displays. Displays need to look fresh and interesting and tell a compelling story about why the consumer should purchase the product.

12. Seasonal memberships and discounts can be introduced. Co nsumers get very excited about savings and appreciate discount s being passed on. Many prominent retailers also have loyalty programs or club member cards that create excitement. A club-member price brings consumers back and improve sales.

13. Events and tastings help draw consumers to your store and generate sales. Retailers with economies of scale successfully sample consumers on more profitable wines. Some even compa rison-taste customers on national brands that are more expensiv eto demonstrate they are offering a less expensive but superior product.

14.And bringing in celebrities, sommeliers or trade reps for tast ings can help create excitement and drive
traffic