Time Series Forecasting

Business Report

Sparkling and Rose

Wine sales

Filed by:

Aditya Rishi

Batch: DSBA-Feb-21-B **For Great Learning** **Filed on: October 10,** 2021

Executive Summary

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

A glass of red wine is displayed at the Chateau La Louviere during the start of the Primeurs, a week of wine tasting, at the chateaux in Leognan, France, April 3, 2017. REUTERS/Regis Duvignau

PROBLEM SET

1. Read the data as an appropriate Time Series data and plot the data.	10,74
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.	12,80
3. Split the data into training and test. The test data should start in 1991.	22,89
4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.	23,90
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.	38,106
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.	48,112
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.	57,119
8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.	65,131
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.	67,133
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.	73,137

Sparkling Tables

1. First few rows of the dataset
2. Last few rows of the dataset
3. Creating timestamp
4. Adding timestamp to the data frame
5. Setting timestamp as index
6. Data frame after dropping YearMonth column
7. Basic measures of descriptive statistics
8. Check for missing values
9. Pivot table of monthly sales across years
10. Trend, seasonality, residuals for additive decomposition
11. Trend, seasonality, residuals for multiplicative decomposition
12. Test and training data, first and last few rows
13. Shapes of train and test sets
14. Training time and test time instances for linear regression model
15. Test and training data, first and last few rows, for linear regression model
16. Naive model train and test data with timestamp
17. Simple-average model train and test data with timestamp
18. Moving average data with timestamp
19. Simple exponential smoothing training data
20. Simple exponential smoothing forecast for fitted values and test
21. Simple exponential smoothing forecast
22. Double exponential smoothing sorted train and test RMSE and MAPE heads to get best scores (Alpha 0.1, Beta 0.1)
23. Double exponential smoothing sorted test RMSE and MAPE heads to get best scores (autofit Alpha 0.64, Beta 0.0)
24. Triple exponential smoothing sorted train and test RMSE and MAPE heads to get best scores (alpha 0.4, beta 0.1, gamma 0.2)
25. Triple exponential smoothing sorted test RMSE and MAPE heads to get best scores (alpha 0.15, beta 7.4, gamma 0.37), along with tail
26. Various models
27. Auto Sarima on original data, examples of some parameter combinations for model
28. Auto Sarima models with best AIC scores
29. Auto Sarima (3,1,3)x(3,1,0,12) prediction summary
30. Predicted and true values of time series for Auto Sarima (3,1,3)x(3,1,0,12)
31. Auto Sarima on log10, examples of some parameter combinations for model
32. Auto Sarima on log10: Models with best AIC scores
33. Auto Sarima (0,1,1)x(1,0,1,12)on log10: prediction summary
34. Predicted and true values of time series for Auto Sarima (0,1,1)x(1,0,1,12) log10 transformation model
35. Manual Sarima (3,1,1)x(1,1,2,12), prediction summary
36. Manual Sarima (3,1,1)x(1,1,2,12): predicted and true values
37. Most optimum models by RMSE
38. Most optimum model, by MAPE
39. Sarima (3,1,3)x(1,1,2,12) for forecast, prediction summary
40. Sarima (3,1,3)x(1,1,2,12) forecast on full data, tail
41. Sparkling: 12-month forecast (August 1995 to July 1996)
42. Sparkling: Forecast description

Sparkling Figures

1. Sparkling wine sales (original data)
2. Distribution of Sparkling wine (estimator of the Cumulative Distribution Function or eCDF plot)
3. Sparkling wine data distribution (distplot)
4. Sparkling wine sales by year (barplot)
5. Sparkling wine sales by month (barplot)
6. Sparkling wine average sales each day (barplot)
7. Sparkling wine overall sales each day (barplot)
8. Yearly boxplot - Sparkling
9. Monthly boxplot - Sparkling
10. Sparkling - Times series month plot
11. Sparkling-Monthly sales over years
12. Average Sparking sales
13. Sparkling Sales: Percentage change
15. Sparkling - Additive decomposition
14. Detrending, of additive: Time series, original and without seasonality
15. Sparkling - Multiplicative decomposition
16. Detrending, of multiplicative: Time series, original and without seasonality
17. Sparkling Sales - Data Split
18. Sparkling sales - Linear Regression Model
19. Sparkling - Naive Forecast
20. Sparkling - Simple Average Forecast
21. Sparkling - Trailing Moving Average & TMA Forecast
22. Simple exponential smoothing with alpha 0.1, RMSE 1375.39 (training, test, and predicted time series plot)
23. Simple exponential smoothing with alpha 0.2, RMSE 1595.21 (training, test, and predicted time series plot)
24. Simple exponential smoothing with alpha 0.3, RMSE 1935.51 (training, test, and predicted time series plot)
25. Simple exponential smoothing with alpha 0.5, RMSE 2666.35 (training, test, and predicted time series plot)
26. Simple exponential smoothing with alpha 0.99, RMSE 3847.55 (training, test, and predicted time series plot)
27. Sparkling - SES forecast (Autofit Alpha: 0.0) on train and test
28. Sparkling DES forecast (Alpha 0.1, Beta 0.1)
29. Sparkling DES forecast (Autofit Alpha 0.64, Beta 0.00)
30. Sparkling TES forecast (Alpha 0.4, Beta 0.1, Gamma 0.2)
31. Sparkling TES forecast (Autofit Alpha 0.15, Beta 7.4, Gamma 0.37)
32. Sparkling: Forecast vs Actual Test Data (all-model time series plot)
33. Dickey-Fuller Test (rolling mean and standard deviation), on original series, at 5% significant level
34. Dickey-Fuller Test (rolling mean and standard deviation) with difference of order 1
35. Dickey-Fuller Test (rolling mean and standard deviation) with differencing of seasonal order (12)
36. ADF (rolling mean and standard deviation) with differencing: Series after seasonal differencing + 1-order differencing
37. ADF (rolling mean and standard deviation) with logarithmic transformation, using log10
38. ADF (rolling mean and standard deviation) Difference of Log 10 Series after seasonal differencing
39. ADF (rolling mean and standard deviation) Difference of Log 10 Series after seasonal differencing + 1-order differencing
40. Sparkling - Autocorrelation
41. Sparkling - Differenced data autocorrelation

42. Sparkling - Partial Autocorrelation

43. Differenced Data Partial Autocorrelation

44. Sparking partial autocorrelation with ywmle method

45. Differenced data partial autocorrelation with ywmle method

46. Auto SARIMA model (3,1,3)x(3,1,0,12): Results

47. Auto SARIMA model (3,1,3)x(3,1,0,12): Diagnostics

48. Sparkling Auto Sarima(3,1,3)x(3,1,0,12)

49. Log Data Autocorrelation

50. Log data difference autocorrelation

51. Log data partial autocorrelation

52. Log data difference partial autocorrelation

53. Auto SARIMA (0,1,1)x(1,0,1,12) on log10 model: Results

54. Auto SARIMA (0,1,1)x(1,0,1,12) on log10 model: Diagnostics

55. Sparkling - Auto SARIMA (0,1,1)x(1,0,1,12) with Log Transformation

56. For Manual Sarima: Observed data autocorrelation

57. Observed data partial autocorrelation

58. Observed data partial autocorrelation with ywmle method

59. Observed time series before applying Manual SARIMA

60. Series after seasonal differencing, before applying Manual SARIMA

61. Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

62. ADF Test of stationarity on Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

63. Differenced Data Autocorrelation

64. Differenced Data Partial Autocorrelation

54. Manual SARIMA (3,1,1)x(1,1,2,12): Results

65. Manual SARIMA (3,1,1)x(1,1,2,12): Diagnostics

66. Sparkling - Manual SARIMA (3,1,1)x(1,1,2,12) forecast

67. Sparkling: Forecast vs Test Data

68. Sparkling - 12 months forecast using TES model

72. Sparkling Wine Sales - 12-month forecast

73: Sarima (3,1,3)x(1,1,2,12) model: Results

74: Sarima (3,1,3)x(1,1,2,12) model: Diagnostics

74: Sarima (3,1,3)x(1,1,2,12) model: Prediction summary

75: Sparkling: 12-month Forecast using SARIMA model

76: Sparkling: 12-month Forecast with confidence interval

Rose Tables

1. First few rows of the dataset
2. Last few rows of the dataset
3. Creating timestamp
4. Adding timestamp to data frame
5. Setting timestamp as index
6. Data frame after dropping YearMonth column
7. Basic measures of descriptive statistics
8. Check for missing values
9. 1994 sales figures, to check gap
10. 1994 figures after data interpolation
11. Descriptive statistics of converted data frame.
12. Pivot table of monthly sales across years
13. Trend, seasonality, residuals for additive decomposition
14. Trend, seasonality, residuals for multiplicative decomposition
15. Test and training data, first and last few rows
16. Shapes of train and test sets
17. Training time and test time instances for linear regression model
18. Test and training data, first and last few rows, for linear regression model
19. Naive model train and test data with timestamp
20. Simple-average model train and test data with timestamp
21. Moving average data with timestamp
22. Simple exponential smoothing forecast for fitted values and test
23. Double exponential smoothing sorted test RMSE and MAPE heads to get best scores (Alpha 0.1, Beta 0.1)
24. Double exponential smoothing sorted test RMSE and MAPE heads to get best scores (autofit Alpha 0.16, Beta 0.16)
25. Triple exponential smoothing sorted test RMSE and MAPE heads to get best scores (alpha 0.1, beta 0.2, gamma 0.2)
26. Triple exponential smoothing sorted test RMSE and MAPE heads to get best scores (alpha 0.1, beta 0.04, gamma 0.0), along with tail
27. Auto Sarima on original data, examples of some parameter combinations for model
28. Auto Sarima models with best AIC scores
29. Auto Sarima (3,1,1)x(3,1,1,12) prediction summary
30. Predicted and true values of time series for Auto Sarima (3,1,1)x(3,1,1,12)
31. Auto Sarima on log10, examples of some parameter combinations for model
32. Auto Sarima on log10: Models with best AIC scores
33. Auto Sarima (1,0,0)x(1,0,1,12)on log10: prediction summary
34. Predicted and true values of time series for Auto Sarima (1,0,0)x(1,0,1,12) log10 transformation model
35. Manual Sarima, prediction summary
36. Manual Sarima (4,1,2)x(0,1,1,12): predicted and true values
37. Manual Sarima on Log10, prediction summary
38. Manual Sarima (4,1,1)x(0,1,1,12) on Log10, predicted and true values
39. Most optimum model, by RMSE
40. Most optimum model, by MAPE
41. Sarimax model (4,1,1)x(0,1,1,12) on full data, prediction summary
42. Sarimax model on full data, fitted values

43. Forecast for August 1995 to July 1996

44. Forecast summary

Rose Figures

1. Rose wine sales (original data)
2. Rose wine sales - After interpolation
3. Distribution of rose wine (estimator of the Cumulative Distribution Function or eCDF plot)
4. Rose wine data distribution (distplot)
5. Rose wine sales by year (barplot)
6. Rose wine sales by month (barplot)
7. Rose wine average sales each day (barplot)
8. Rose wine overall sales each day (barplot)
9. Yearly boxplot
10. Monthly boxplot
11. Times series month plot
12. Monthly sales over years
13. Average rose sales
14. Rose sales: Percentage change
15. Additive decomposition
16. Detrending, of additive: Time series, original and without seasonality
17. Multiplicative decomposition
18. Detrending, of multiplicative: Time series, original and without seasonality
19. Rose Sales - Data Split
20. Rose sales - Linear Regression Model
21. Rose - Naive Forecast
22. Rose - Simple Average Forecast
23. Rose - Trailing Moving Average & TMA Forecast
24. Simple exponential smoothing with alpha 0.1, RMSE 36.83 (training, test, and predicted time series plot)
25. Simple exponential smoothing with alpha 0.2, RMSE 41.36 (training, test, and predicted time series plot)
26. Simple exponential smoothing with alpha 0.3, RMSE 47.5 (training, test, and predicted time series plot)
27. Simple exponential smoothing with alpha 0.5, RMSE 59.64 (training, test, and predicted time series plot)
28. Simple exponential smoothing with alpha 0.99, RMSE 79.5 (training, test, and predicted time series plot)
29. Rose - SES forecast (Autofit Alpha: 0.0987) on train and test
30. Rose DES forecast (Alpha 0.1, Beta 0.1)
31. Rose DES forecast (Autofit Alpha 0.16, Beta 0.16)
32. Rose TES forecast (Alpha 0.1, Beta 0.2, Gamma 0.2)
33. Rose TES forecast (Autofit Alpha 0.1, Beta 0.04, Gamma 0.0)
34. Rose: Forecast vs Actual Test Data (all-model time series plot)
35. Dickey-Fuller Test (rolling mean and standard deviation), on original series, at 5% significant level
36. Dickey-Fuller Test (rolling mean and standard deviation) with difference of order 1
37. Dickey-Fuller Test (rolling mean and standard deviation) with differencing of seasonal order (12)

38. ADF (rolling mean and standard deviation) with logarithmic transformation of the train data, using log10

39. ADF (rolling mean and standard deviation) with differenced logarithmic transformation of the train data, using log10

40. Rose - Autocorrelation

41. Rose - Differenced data autocorrelation

42. Rose - Partial Autocorrelation

43. Differenced Data Partial Autocorrelation

44. Auto SARIMA model: Results

45. Auto SARIMA model: Diagnostics

46. Rose Auto Sarima (3,1,1)x(3,1,1,12)

47 Auto SARIMA on log10 model: Results

47. Auto SARIMA on log10 model: Diagnostics

48. Rose- Auto SARIMA (1,0,0)x(1,0,1,12) with Log Transformation

49. Observed time series before applying Manual SARIMA

50. Series after seasonal differencing, before applying Manual SARIMA

51. Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

52. ADF Test of stationarity on Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

53. Differenced Data Autocorrelation

54. Differenced Data Partial Autocorrelation

54. Manual SARIMA (4,1,2)x(0,1,2,12): Results

55. Manual SARIMA (4,1,2)x(0,1,2,12): Diagnostics

56. Rose - Manual SARIMA (4,1,2)x(0,1,2,12) forecast

57. Manual SARIMA on log10: Observed logged time series, before applying model

57. Logged series after seasonal differencing, before applying Manual SARIMA log10 model

58. Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA log10 model

59. ADF Test of stationarity on logged series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

60. Log transformed data autocorrelation

61. Log transformed partial data autocorrelation

62. Differenced log transformed data autocorrelation

63. Differenced log transformed data partial autocorrelation

64. Manual SARIMA (4,1,1)x(0,1,1,12) on log10 model: Results

64. Manual SARIMA (4,1,1)x(0,1,1,12) on log10 model: Diagnostics

65. Rose - Manual SARIMA (4,1,1)x(0,1,1,12) with log transformation

66. Rose - Forecast vs Test data, 3 best models

67.Rose - 12-month forecast using TES model

68. Rose - 12-month forecast

69. Rose full data SARIMAX (4,1,1)x(0,1,1,12) model: Results

70. Rose full data SARIMAX (4,1,1)x(0,1,1,12) model: Diagnostics

71. Rose - 12 months forecast using SARIMA model

72. Rose - 12-month forecast with confidence interval

Sparkling

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

Read the data from the '.csv' file as a monthly Time Series.

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

1. First few rows of the dataset

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

2. Last few rows of the dataset

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

3. Creating timestamp with month as frequency, since we deal with sales figures here

	YearMonth	Sparkling	Time_Stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

4.

Adding timestamp to the data frame

Time_Stamp	YearMonth	Sparkling
1980-01-31	1980-01	1686
1980-02-29	1980-02	1591
1980-03-31	1980-03	2304
1980-04-30	1980-04	1712
1980-05-31	1980-05	1471
...	...	...
1995-03-31	1995-03	1897
1995-04-30	1995-04	1862
1995-05-31	1995-05	1670
1995-06-30	1995-06	1688
1995-07-31	1995-07	2031

5. Setting timestamp as index

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471
...	...
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

187 rows × 1 columns

6. Data frame after dropping YearMonth column

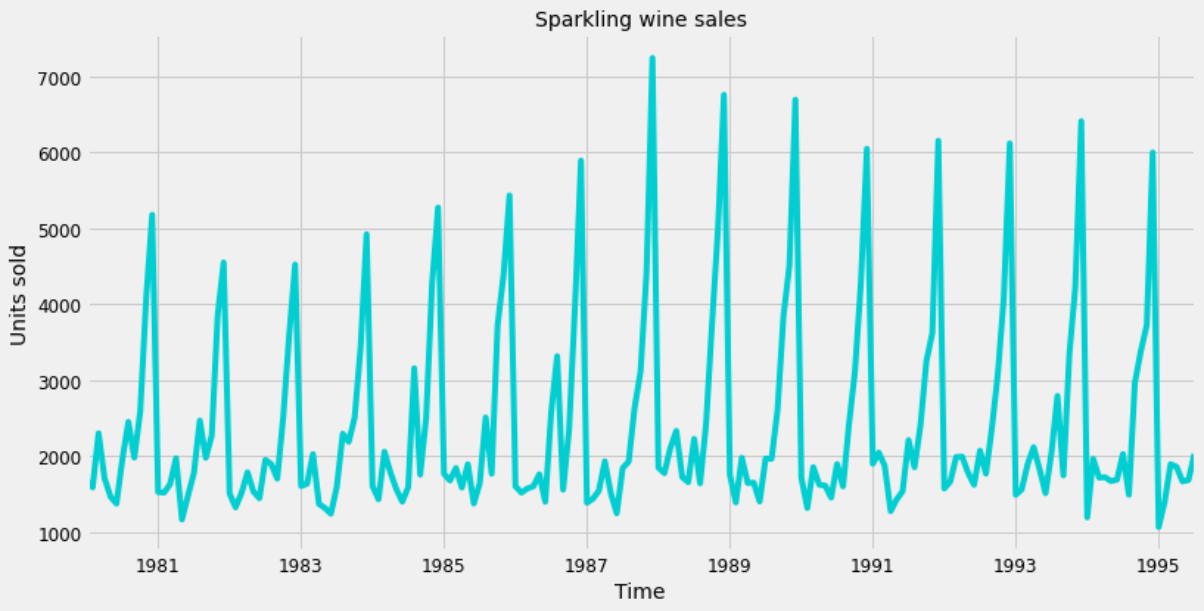
	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

7. Basic measures of descriptive statistics

```
Sparkling    0
dtype: int64
```

8. Check for missing values

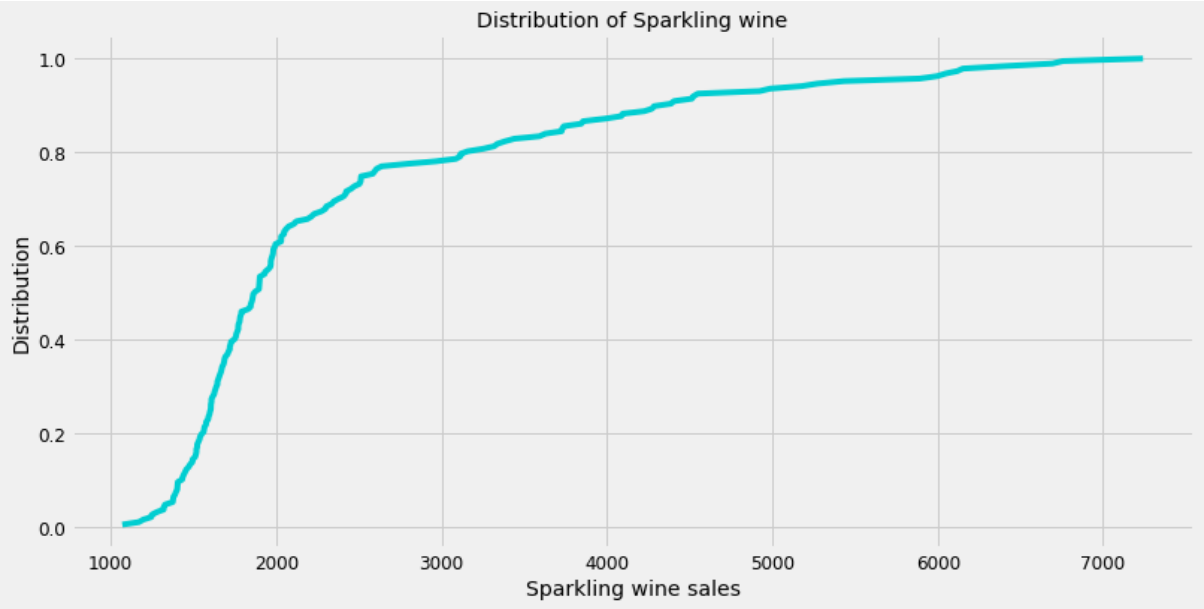
Plot the Time Series to understand the behaviour of the data.



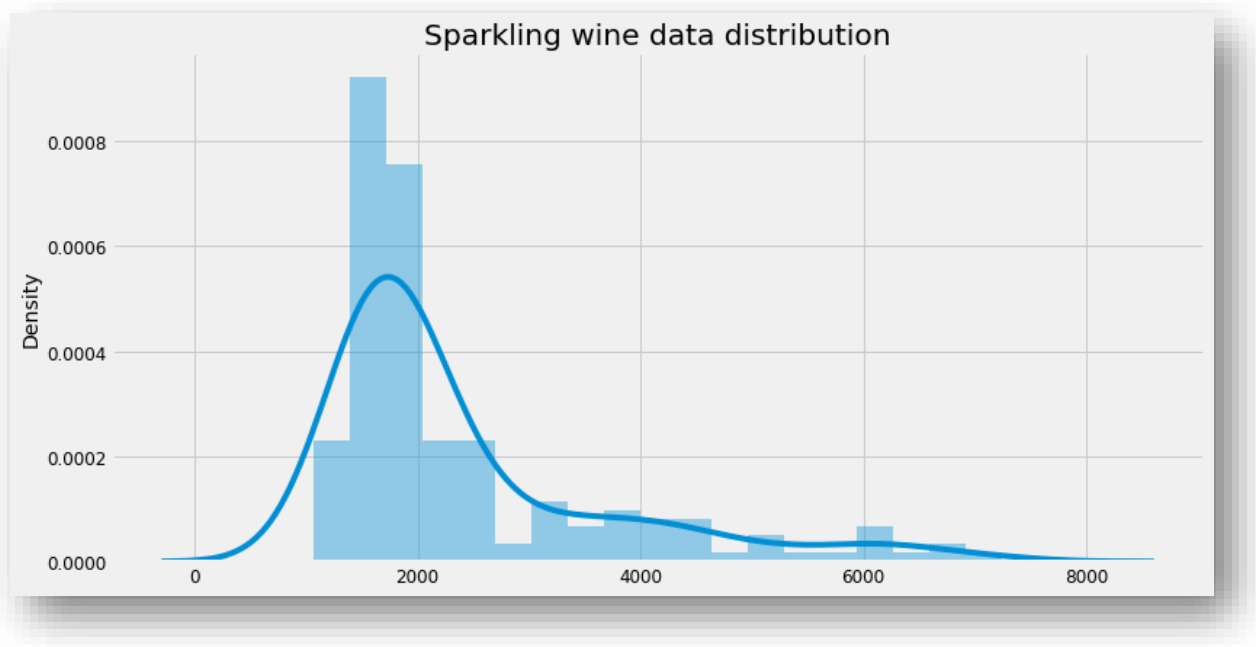
1. Sparkling wine sales (original data)

• Monthly sales of Sparkling wine are given for a period from January 1980 to July 1995
• The data loaded fine and a date-range has been applied on the set as index
• The time-series has no missing value but significant seasonality. The sales of Sparkling wine don't show any consistent trend but upward and downward slopes during the time period
• The customers have given a good response to Sparkling wine over the years

2. Perform appropriate EDA to understand the data and also perform decomposition.



2. Distribution of Sparkling wine (estimator of the Cumulative Distribution Function or eCDF plot)

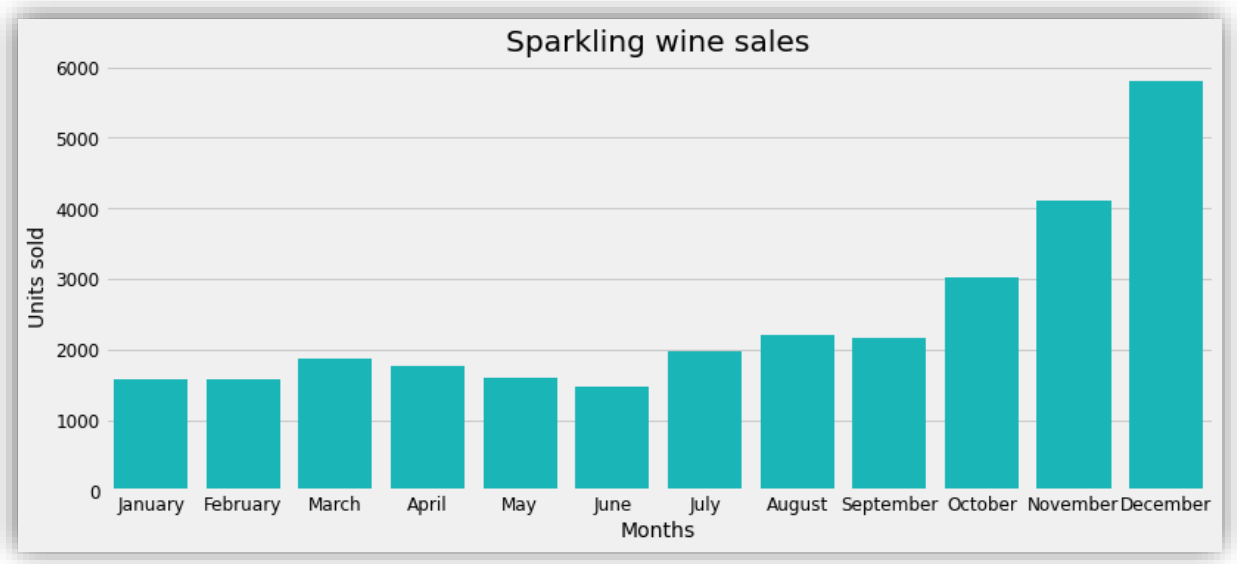


3. Sparkling wine data distribution (distplot)

Data is right skewed. Right skewed distributions occur when the long tail is on the right side of the distribution. Analysts also refer to them as positively skewed. This condition occurs because probabilities taper off more slowly for higher values. Consequently, you'll find extreme values far from the peak on the high end more frequently than on the low.

4. Sparkling wine sales by year (barplot)

• Data seems to have more or less same sales across the year. 1988 has recorded maximum sales

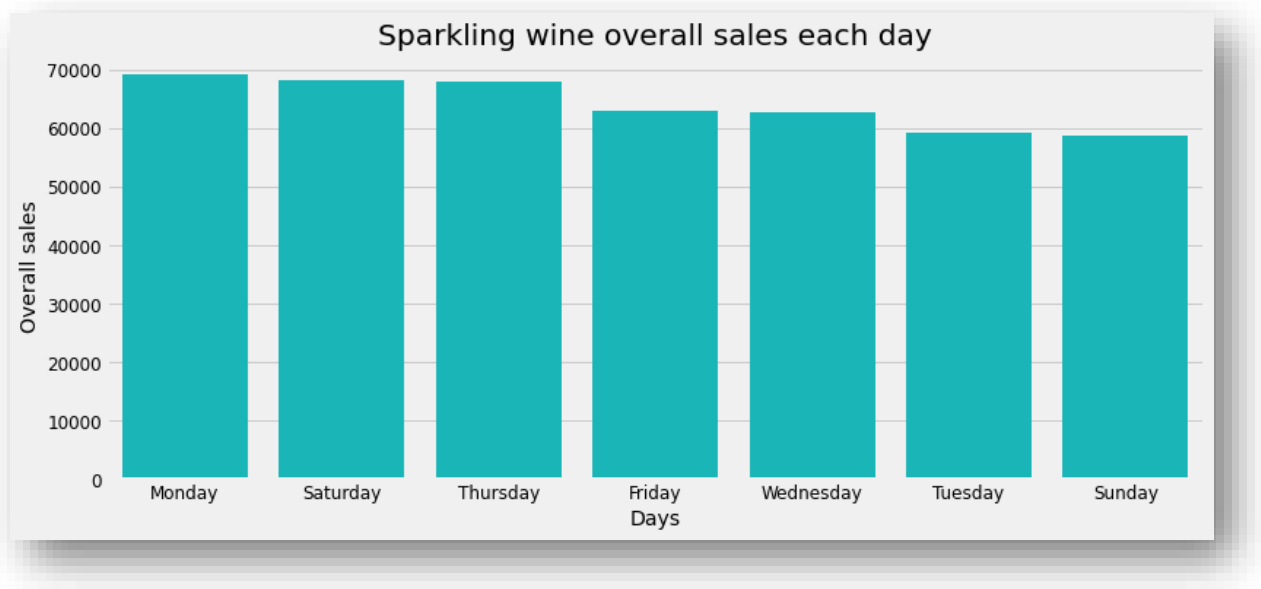


5. Sparkling wine sales by month (barplot)

• December has best sales across all months, followed by November and October, may be due to the year-end parties.

6. Sparkling wine average sales each day (barplot)

• Saturdays register highest average sales of the week.



7. Sparkling wine overall sales each day (barplot)

• **Mondays have the highest overall sales**

• The descriptive summary of the data shows that, on an average, 2,402 units of Sparkling wines were sold each month in the given period of time.

• In 50% of the months, sales varied from 1,605 units to 2,549 units. The maximum sales reported in a month were 7,242 units (checking out the yearly and monthly boxplots together, we find out that it was in December 1987).

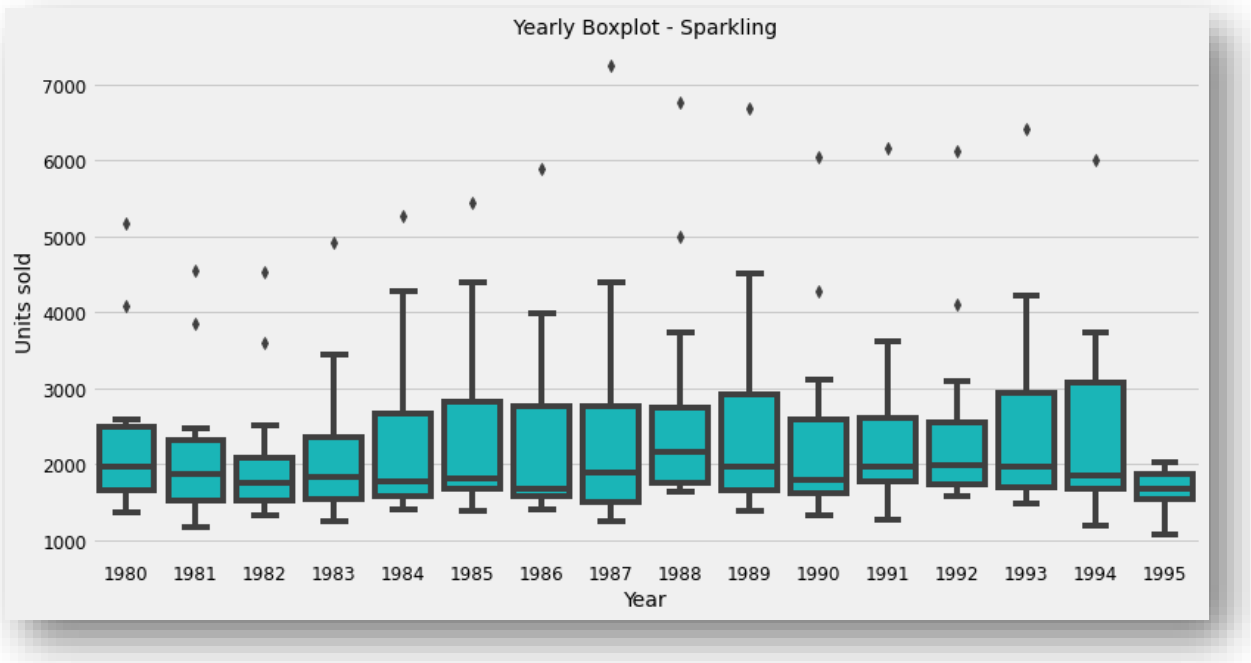
• The empirical cumulative distribution function (eCDF) plot shows that, in 80% of the months, at least 3,000 units of Sparkling wine were sold

• The yearly-boxplot, shows that the average sale of Sparkling has been more or less consistent across the period, at or a little below 2,000 units

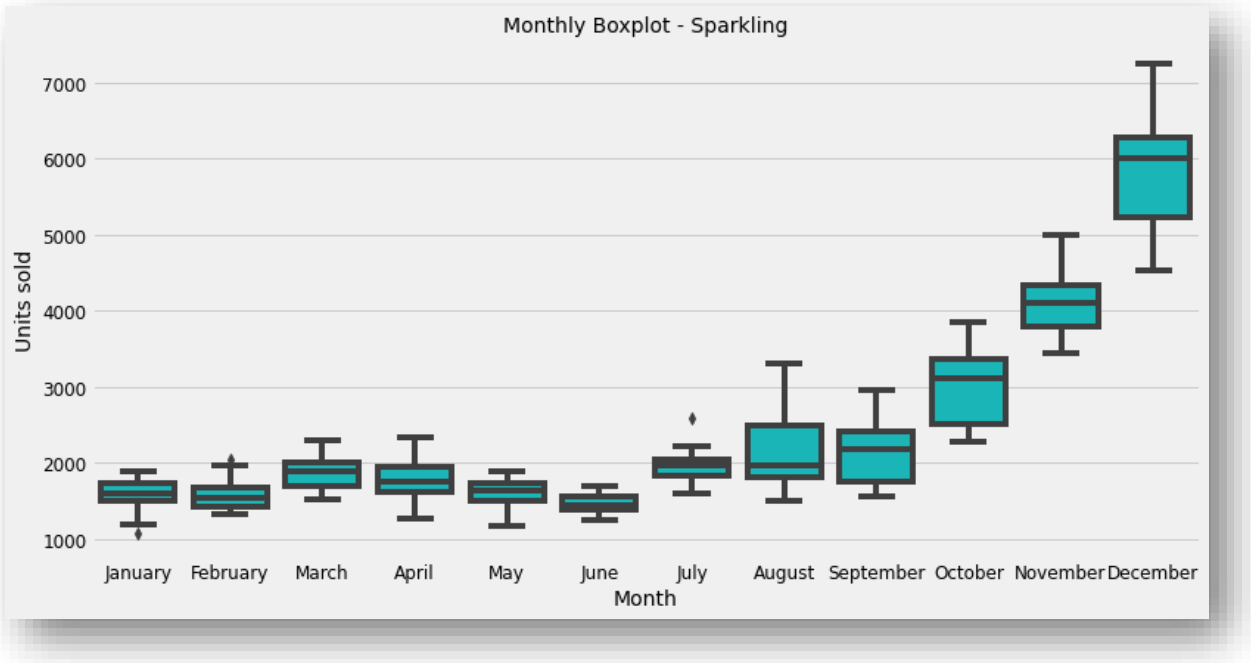
• The outliers in the yearly-boxplot represent the seasonal sales during those months, most probably

• The monthly-box-plot shows a clear seasonality during the festive-season of October, November, and December, which peaks in December. The sales tank in June.

Make a boxplot to understand the spread of wine sales across different years and within different months across years.

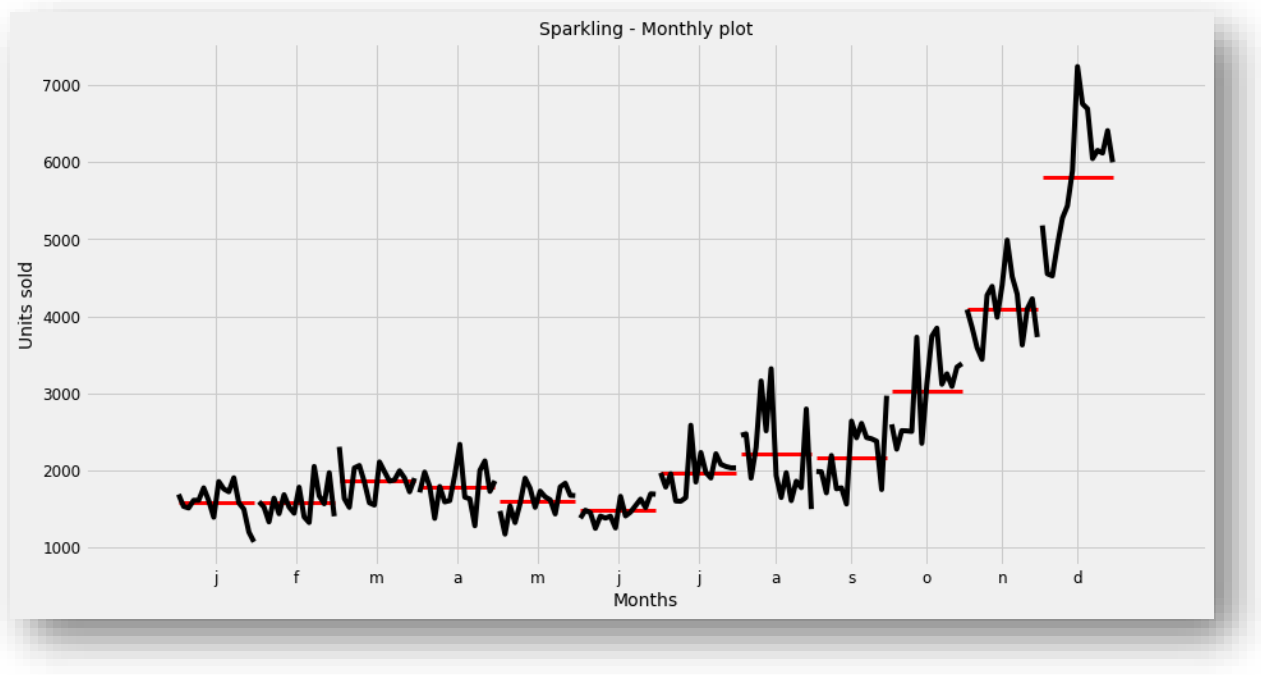


8. Yearly boxplot – Sparkling



9. Monthly boxplot – Sparkling

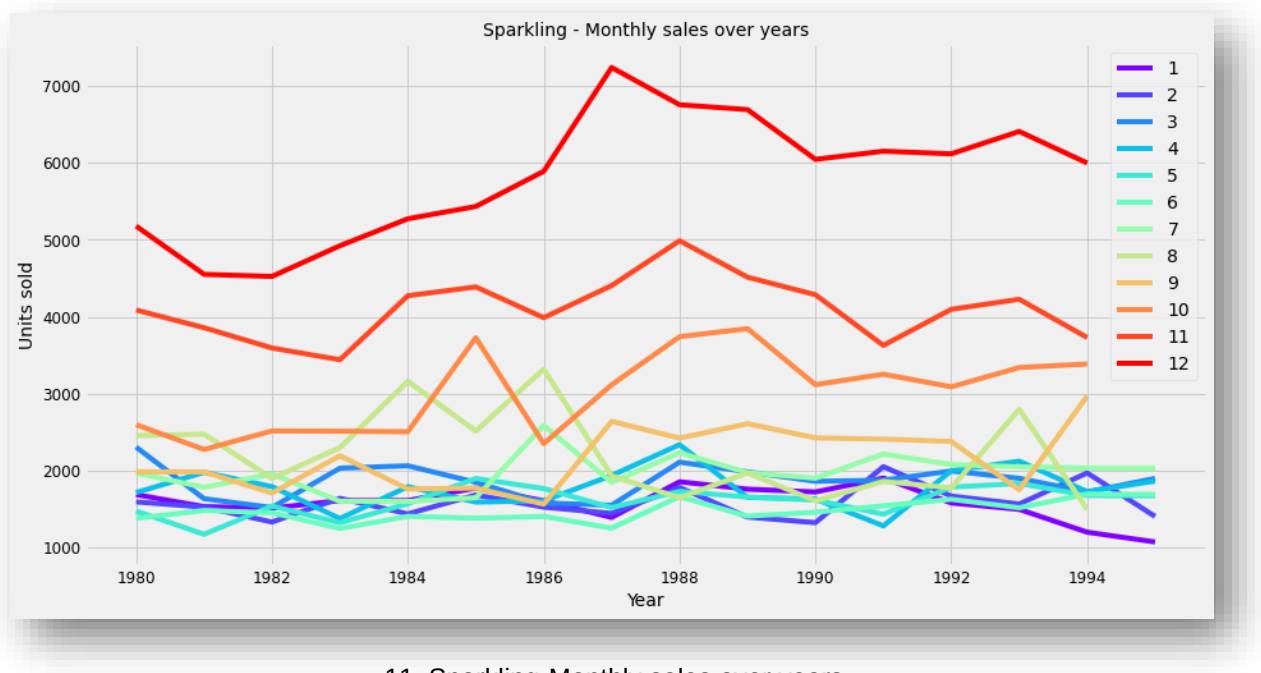
Monthplot to see the spread of sales across year, within months across years.



10. Sparkling - Month plot, tells time series' behaviour across months. The red line is the median.

Months / Years	1	2	3	4	5	6	7	8	9	10	11	12
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

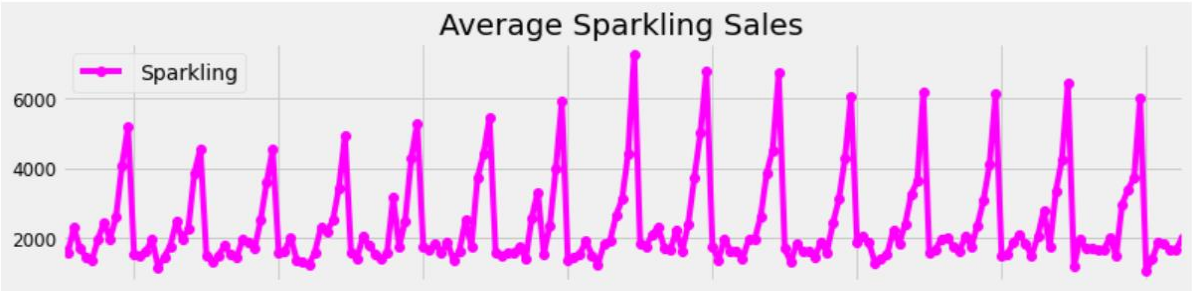
9. Plot a pivot table of monthly sales across years.



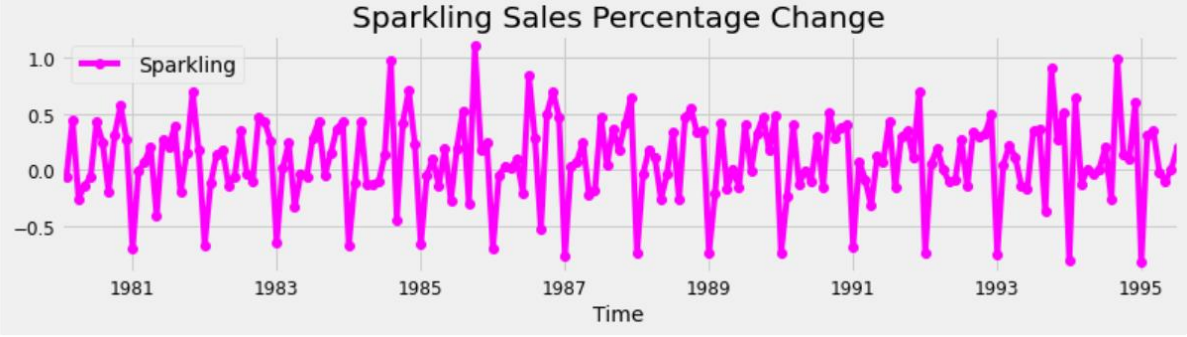
11. Sparkling-Monthly sales over years

• The monthly plot for Sparkling shows the mean and the variation of the units sold each month over the years. Sales in the seasonal months show a higher variation than the sales in the lean months.

• December sales, with mean a little below 6,000, vary from 7,400 to 4,500 units across years, while the November sales vary from 3,500 to ,5000 units, and October ones from 2,500 to 4,000 units. The lean patch of January to September shows more or less consistent sales of around 2,000 units.



12. Average Sparking sales



13. Sparkling Sales: Percentage change

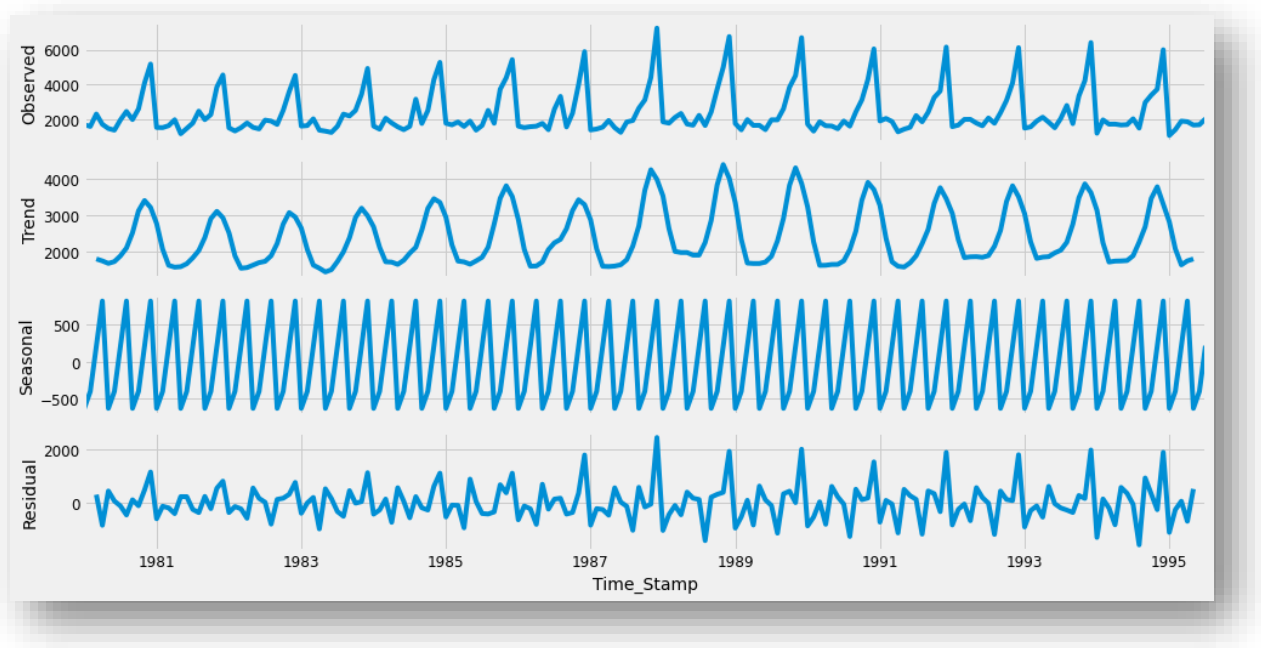
• The plot for the monthly sales over the years also shows the seasonality component of the time-series, with exponentially higher volumes of selling in October, November, and December

• If the highest volume of Sparkling wines was sold in December 1987, the least collection of this month was in 1981. Post December 1987, the sales are around an average of 6,500 units, which was around 5,000 in the early 1980's.

• The seasonal sale since 1990 has been more or less consistent around 6.000 units in December, 4,000 units in November, and 3,000 units in October

• Sales for the months from January to July are seen to be consistent across the years, when compared with the rest of the months

Decompose the Time Series and plot the different components.
If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series.



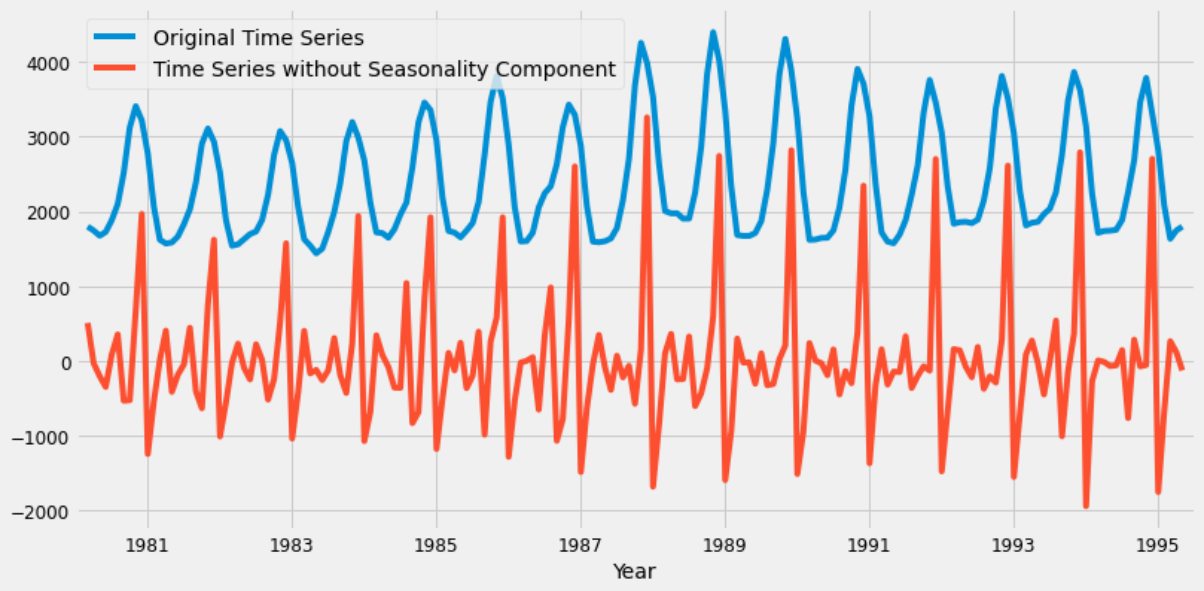
15. Sparkling - Additive decomposition

```
Trend
 Time_Stamp
1980-01-31          NaN
1980-02-29          NaN
1980-03-31     1796.375
1980-04-30     1742.750
1980-05-31     1673.750
1980-06-30     1724.125
1980-07-31     1880.875
1980-08-31     2097.375
1980-09-30     2514.875
1980-10-31     3120.750
1980-11-30     3404.750
1980-12-31     3213.875
Name: Sparkling, dtype: float64
```

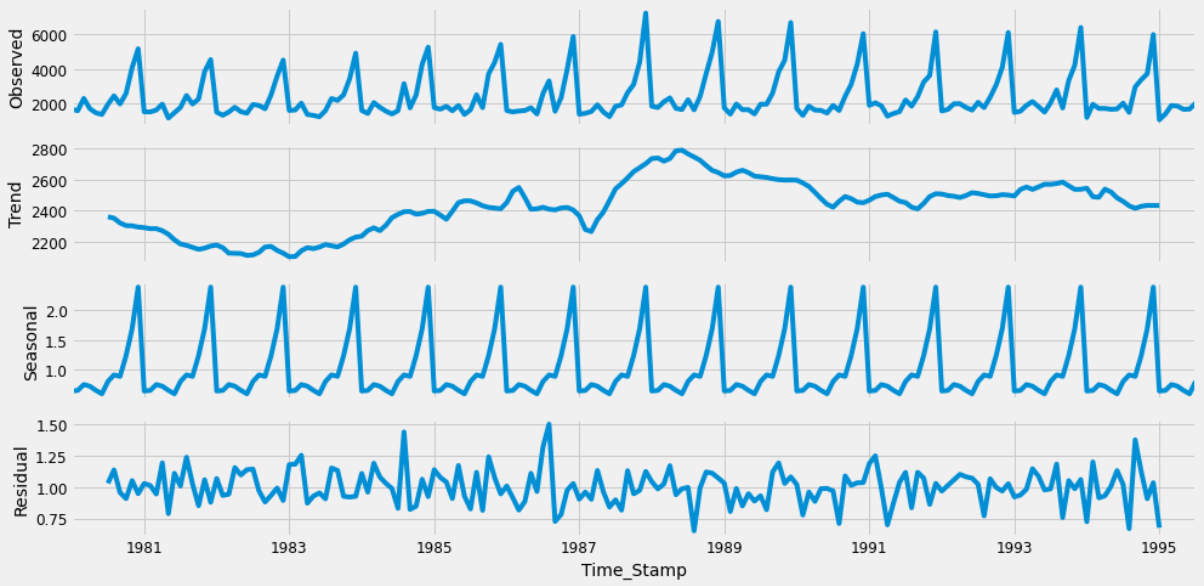
```
Seasonality
 Time_Stamp
1980-01-31   -633.469867
1980-02-29   -402.919203
1980-03-31    214.445894
1980-04-30    821.943176
1980-05-31   -633.469867
1980-06-30   -402.919203
1980-07-31    214.445894
1980-08-31    821.943176
1980-09-30   -633.469867
1980-10-31   -402.919203
1980-11-30    214.445894
1980-12-31    821.943176
Name: Sparkling, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31    293.179106
1980-04-30   -852.693176
1980-05-31    430.719867
1980-06-30     55.794203
1980-07-31   -129.320894
1980-08-31   -466.318176
1980-09-30    102.594867
1980-10-31   -121.830797
1980-11-30    467.804106
1980-12-31   1143.181824
Name: Sparkling, dtype: float64
```

10. Trend, seasonality, residuals for additive decomposition



14. Detrending, of additive: Sparkling time series, original and without seasonality



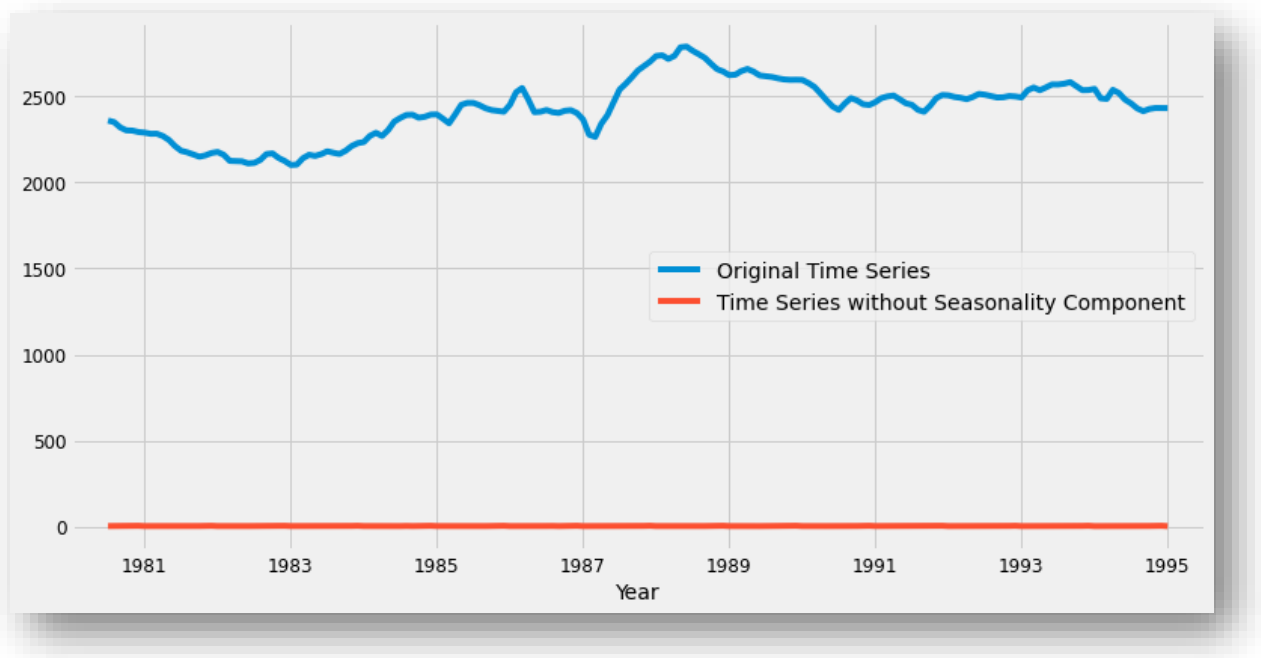
15. Sparkling - Multiplicative decomposition

```
Trend
 Time_Stamp
1980-01-31            NaN
1980-02-29            NaN
1980-03-31            NaN
1980-04-30            NaN
1980-05-31            NaN
1980-06-30            NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: Sparkling, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
1980-06-30    0.603468
1980-07-31    0.809164
1980-08-31    0.918822
1980-09-30    0.894367
1980-10-31    1.241789
1980-11-30    1.690158
1980-12-31    2.384776
Name: Sparkling, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    1.029230
1980-08-31    1.135407
1980-09-30    0.955954
1980-10-31    0.907513
1980-11-30    1.050423
1980-12-31    0.946770
Name: Sparkling, dtype: float64
```

11. Trend, seasonality, residuals for multiplicative decomposition



16. Detrending, of multiplicative: Time series, original and without seasonality

Decomposition analysis

• As the altitude of the seasonal peaks in the observed plot of Sparkling wine sales is changing according to the trend, the time-series is assumed to be 'multiplicative'

• The trend component isn not consistent. An intermediary period shows an upward slope which gets consistent in the later half of time-series

• The additive model shows the seasonality with a variance of 3,000 units, while the multiplicative model shows a variance of 30%

• The residual shows a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions

• The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 10%

• If the seasonality and residual components are independent of the trend, then you have an additive series.

• If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series

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

First few rows of Training Data

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

Last few rows of Training Data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

First few rows of Test Data

Time_Stamp	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Last few rows of Test Data

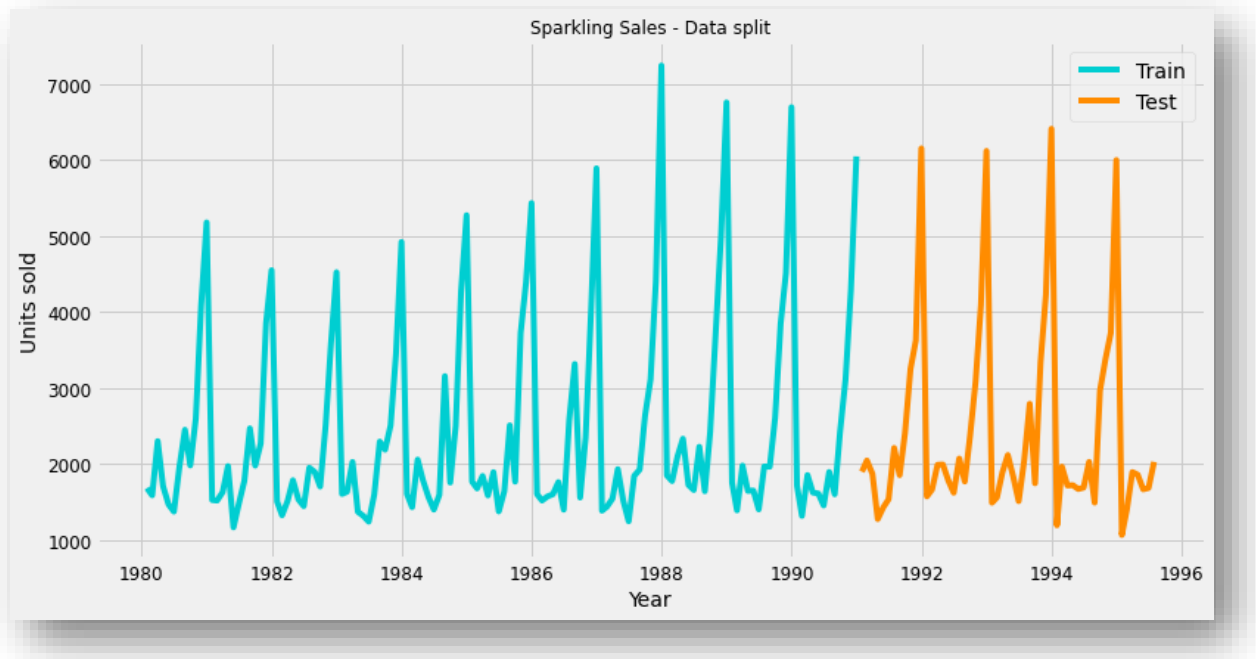
Time_Stamp	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

Train	(132, 1)
Test	(55, 1)

13. Shapes of train and test sets

12. Test and training data, first and last few rows

Plotting data split



17. Sparkling Sales - Data Split. The train and test datasets are created with year 1991 as starting year for test data, using index.year property of time series index

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

Model 1: Linear Regression

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

14. Training time and test time instances for linear regression model

First few rows of Training Data

Time_Stamp	Sparkling	time
1980-01-31	1686	1
1980-02-29	1591	2
1980-03-31	2304	3
1980-04-30	1712	4
1980-05-31	1471	5

Last few rows of Training Data

Time_Stamp	Sparkling	time
1990-08-31	1605	128
1990-09-30	2424	129
1990-10-31	3116	130
1990-11-30	4286	131
1990-12-31	6047	132

First few rows of Test Data

Time_Stamp	Sparkling	time
1991-01-31	1902	133
1991-02-28	2049	134
1991-03-31	1874	135
1991-04-30	1279	136
1991-05-31	1432	137

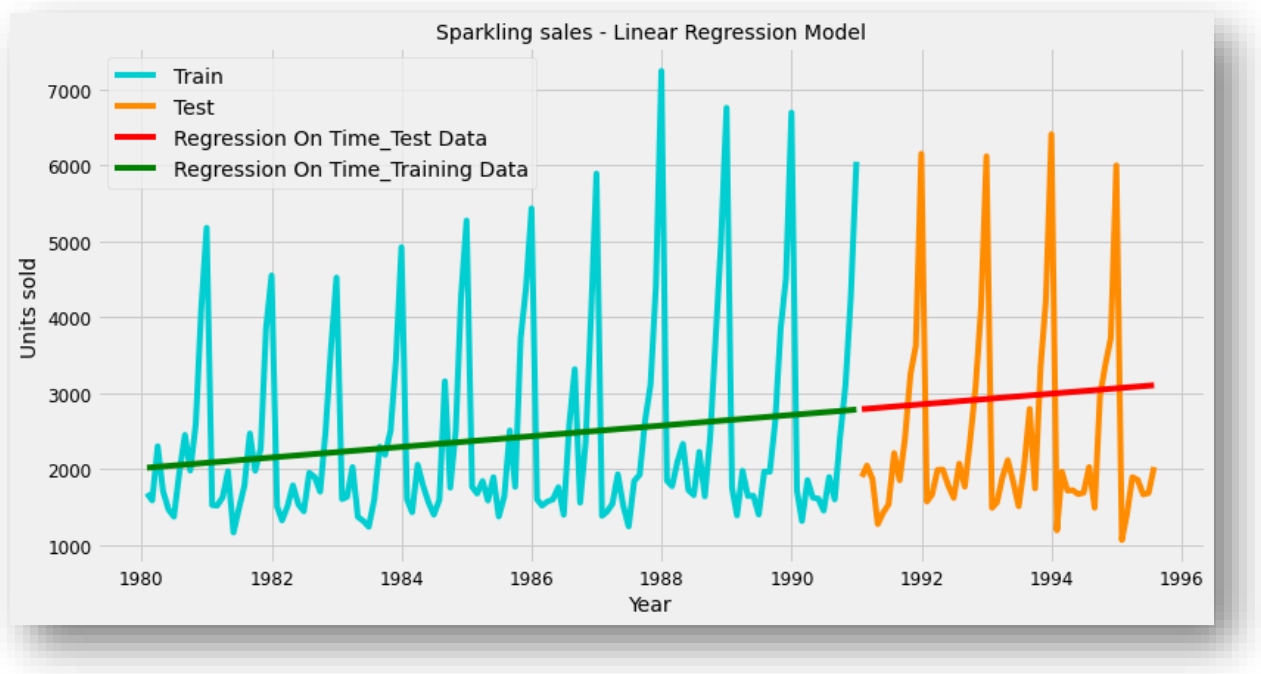
Last few rows of Test Data

Time_Stamp	Sparkling	time
1995-03-31	1897	183
1995-04-30	1862	184
1995-05-31	1670	185
1995-06-30	1688	186
1995-07-31	2031	187

15. Training and test data, first and last few rows, for linear regression model

- Numerical time instance order for both training and test set were generated and the values added to the respective datasets
- The linear regression plot shows a gradual upward trend in forecast of Sparkling wine, consistent with the observed trend which was not apparent visually
- The RMSE **(root mean square error)** and MAPE **(mean absolute percentage error)** values for Train and Test are given in the opposite box.
- 50% of the forecast is erroneous

Model evaluation		
	RMSE	MAPE
Train	3867.322	40.05
Test	1389.135	50.15



18. Sparkling sales - Linear Regression Model

Model 2: Naive forecast

• In naive model, the prediction for tomorrow is the same as today and the prediction Rose for day after tomorrow is tomorrow.

• Since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

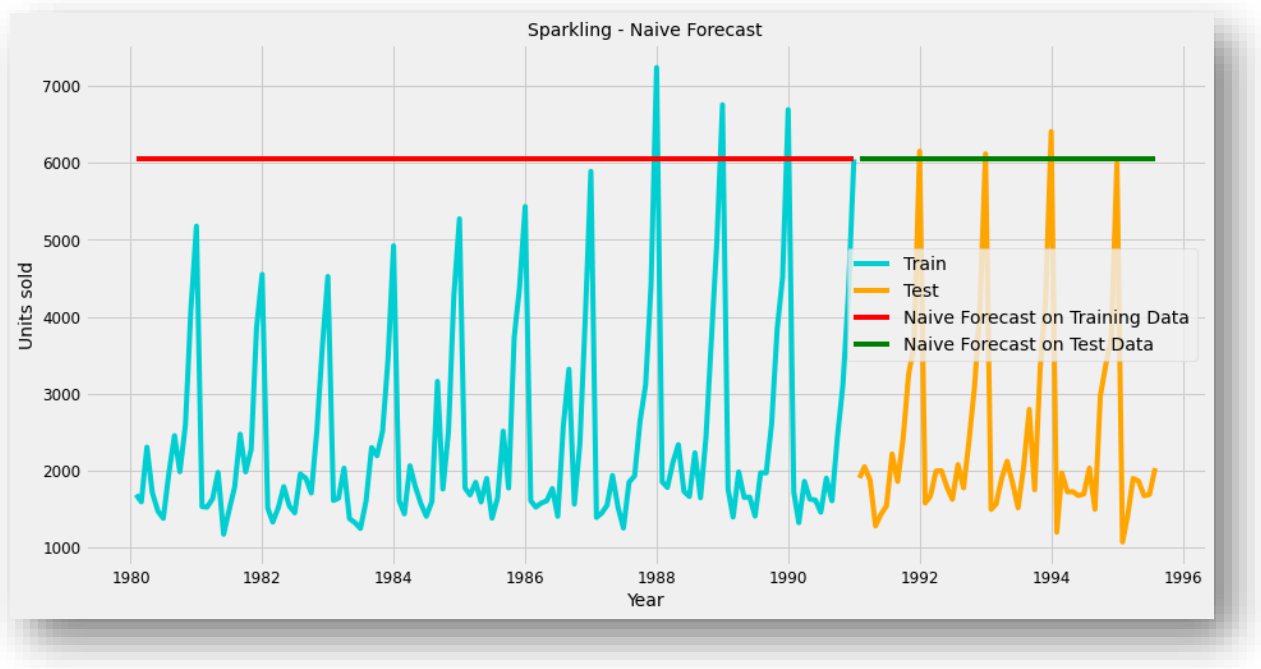
• The model has taken the last value from the test set and fitted it on the rest of the train time period and used the same value to forecast the test set.

• The performance metrices above shows a very poor fitment and high percentage of error.

```
Time_Stamp
1980-01-31    6047
1980-02-29    6047
1980-03-31    6047
1980-04-30    6047
1980-05-31    6047
Name: spark_naive, dtype: int64
```

```
Time_Stamp
1991-01-31    6047
1991-02-28    6047
1991-03-31    6047
1991-04-30    6047
1991-05-31    6047
Name: spark_naive, dtype: int64
```

16. Naive model train and test data with timestamp



19. Sparkling - Naive Forecast

Model evaluation		
	RMSE	MAPE
Train	3867.701	153.17
Test	3864.279	152.87

Model 3: Simple Average

• In the Simple Average model, the forecast is done using the mean of the time-series variable from the training set.

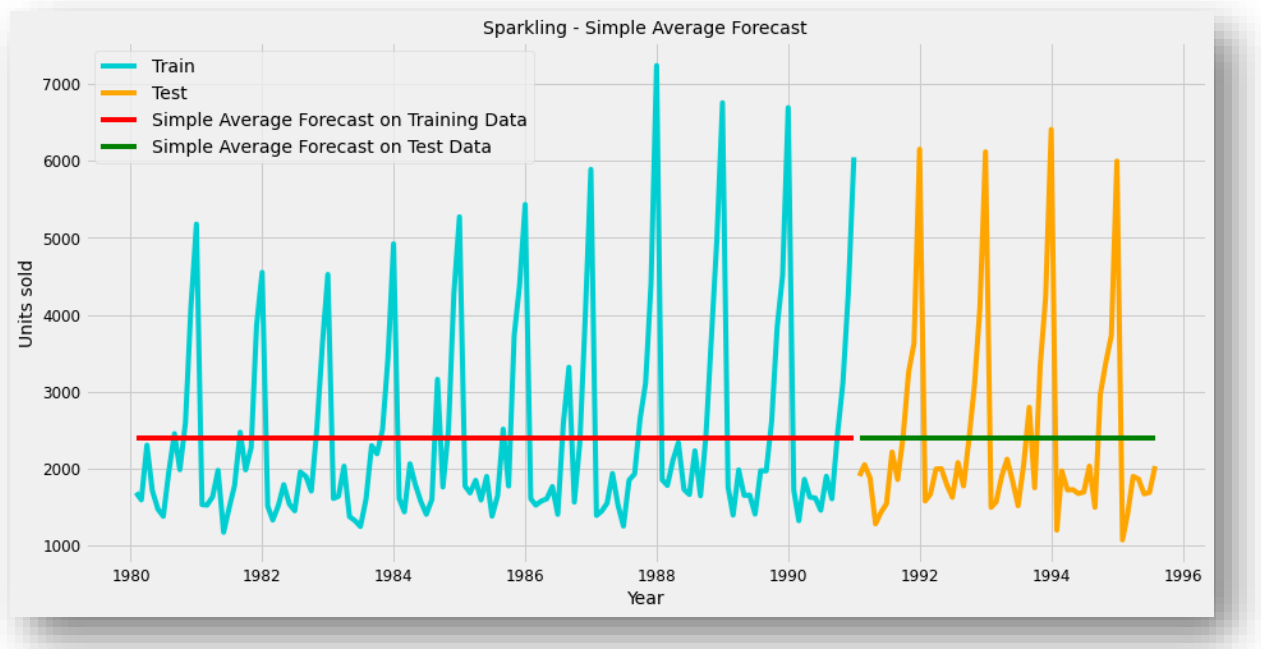
• The model is capable of neither forecasting nor capturing the trend and seasonality present in the dataset.

• For Sparkling the RMSE and MAPE is consistent in both test and train datasets.

```
Time_Stamp
1980-01-31    2403.780303
1980-02-29    2403.780303
1980-03-31    2403.780303
1980-04-30    2403.780303
1980-05-31    2403.780303
Name: spark_mean_forecast, dtype: floa
```

```
Time_Stamp
1991-01-31    2403.780303
1991-02-28    2403.780303
1991-03-31    2403.780303
1991-04-30    2403.780303
1991-05-31    2403.780303
Name: spark_mean_forecast, dtype: float64
```

17. Simple-average model train and test data with timestamp



20. Sparkling - Simple Average Forecast

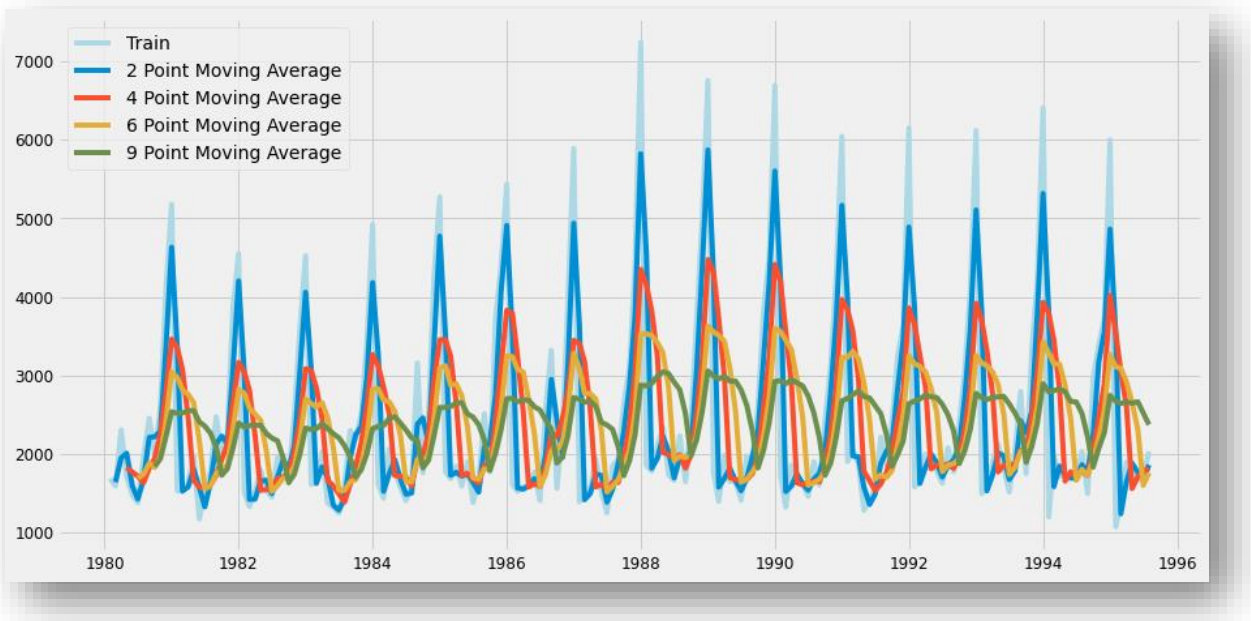
Model evaluation		
	RMSE	MAPE
Train	1298.484	40.36
Test	1275.082	38.9

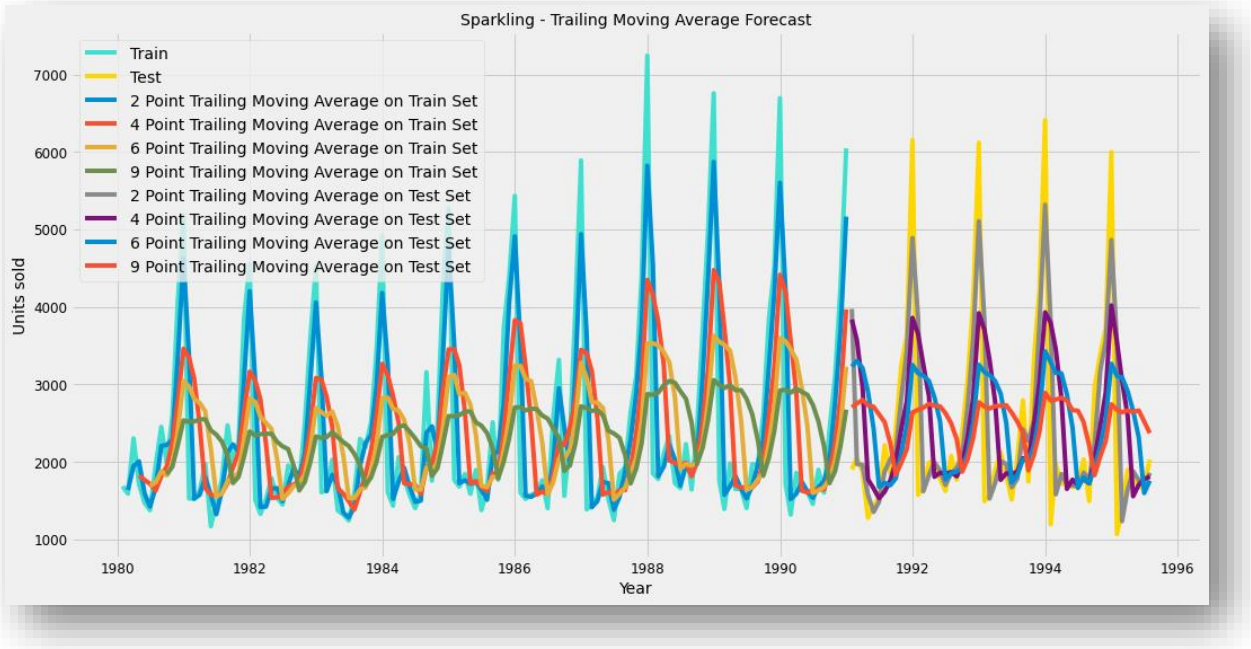
Model 4: Moving Average

Time_Stamp	Sparkling	Spark_Trailing_2	Spark_Trailing_4	Spark_Trailing_6	Spark_Trailing_9
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN

18. Moving average data with timestamp

- For the moving average model, we calculate rolling means (or trailing moving averages) for different intervals. The best interval is gauged by the maximum accuracy (or the minimum error)
- The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points
- For Sparkling dataset the accuracy is found to be higher with the lower rolling point averages
- In moving average forecasts the values can be fitted with a delay of n number of points
- The Root Mean Squared Error and Mean Absolute Percentage Error of the test set are given below
- The best interval of moving average from the model is 2-point





21. Sparkling - Trailing Moving Average & TMA Forecast

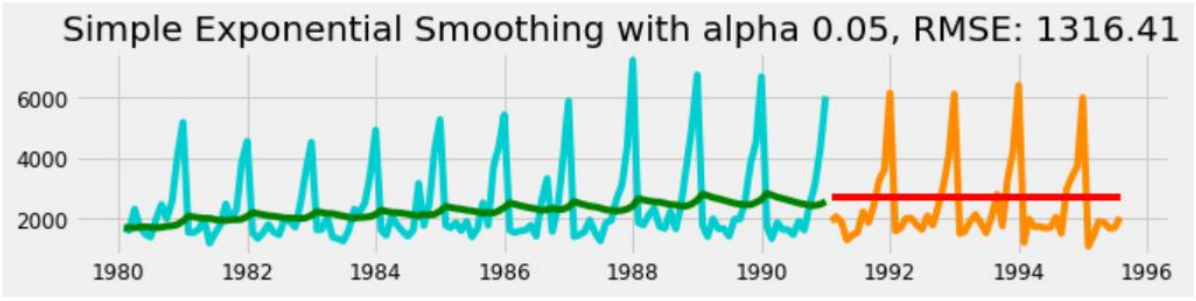
Model evaluation		
Model	RMSE	MAPE
2-point MA	813.401	19.7
4-point MA	1156.59	35.96
6-point MA	1283.927	43.86
9-point MA	1346.278	46.86

Model 5: Simple Exponential Smoothing

```
Time_Stamp
1980-01-31    1686
1980-02-29    1591
1980-03-31    2304
1980-04-30    1712
1980-05-31    1471
Name: Sparkling, dtype: int64
```

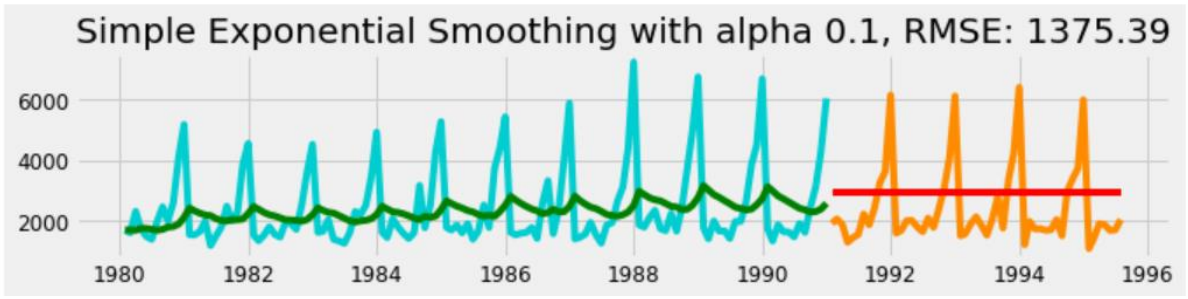
19. Simple exponential smoothing training data

```
Test: For alpha = 0.05,  RMSE is 1316.4117 MAPE is 45.50
For smoothing level = 0.05,  Initial level 1686.00
```



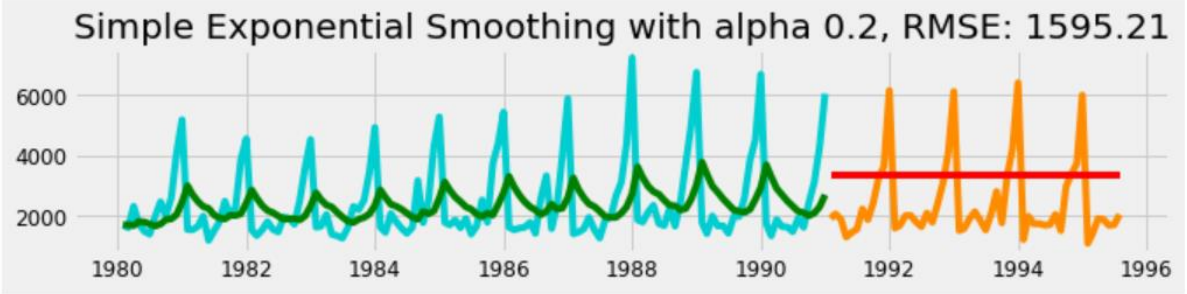
22.1 Simple exponential smoothing with alpha 0.05, RMSE 1316.41
(training, test, and predicted time series plot)

```
Test: For alpha = 0.10,  RMSE is 1375.3934 MAPE is 49.53
For smoothing level = 0.10,  Initial level 1686.00
```



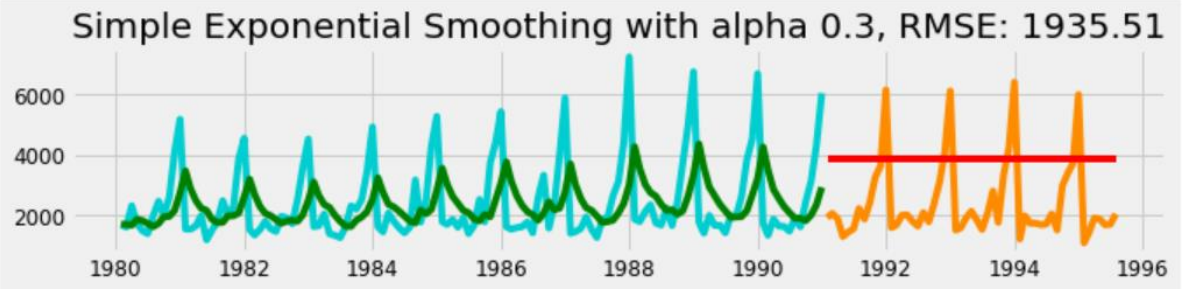
22.2 Simple exponential smoothing with alpha 0.1, RMSE 1375.39
(training, test, and predicted time series plot)

```
Test: For alpha = 0.20,  RMSE is 1595.2068 MAPE is 60.46
For smoothing level = 0.20,  Initial level 1686.00
```



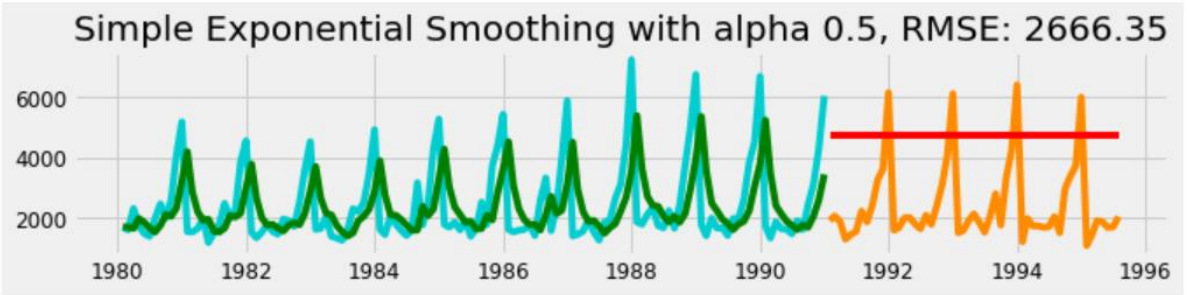
23. Simple exponential smoothing with alpha 0.2, RMSE 1595.21
(training, test, and predicted time series plot)

```
Test: For alpha = 0.30,  RMSE is 1935.5071 MAPE is 75.66
For smoothing level = 0.30,  Initial level 1686.00
```

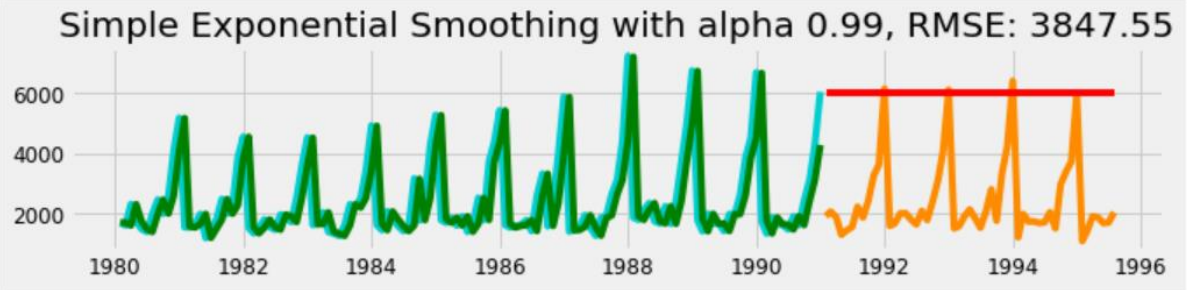


24. Simple exponential smoothing with alpha 0.3, RMSE 1935.51
(training, test, and predicted time series plot)

```
Test: For alpha = 0.50,  RMSE is 2666.3514 MAPE is 106.27
For smoothing level = 0.50,  Initial level 1686.00
```



25. Simple exponential smoothing with alpha 0.5, RMSE 2666.35
(training, test, and predicted time series plot)



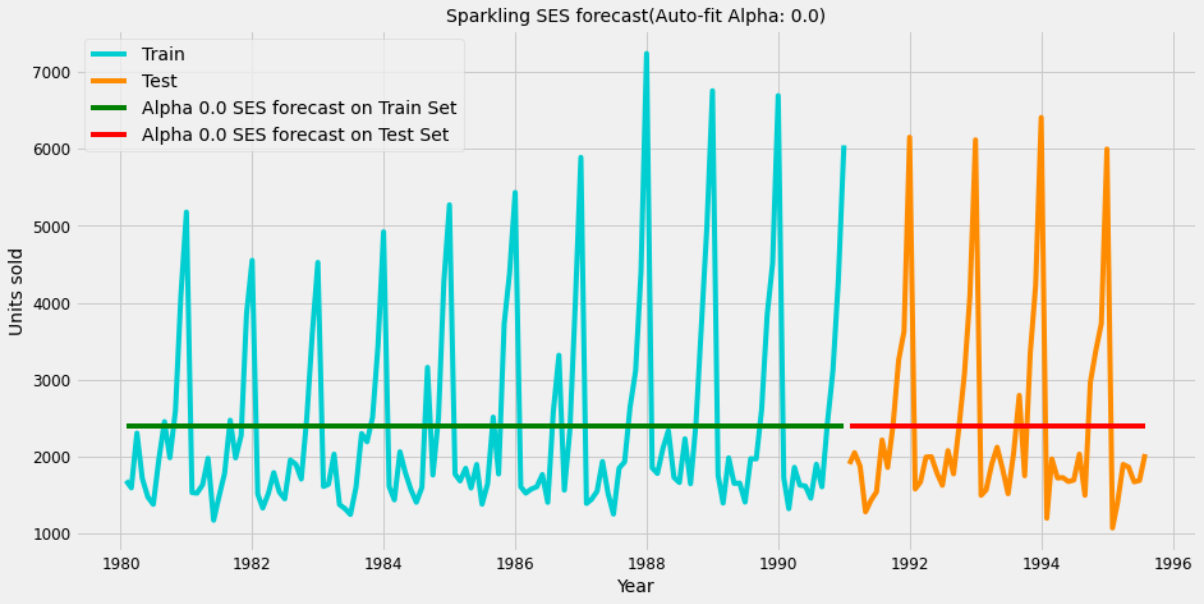
26. Simple exponential smoothing with alpha 0.99, RMSE 3847.55
(training, test, and predicted time series plot)

Time_Stamp	Sparkling	predict_spark
1980-01-31	1686	2403.78287
1980-02-29	1591	2403.78287
1980-03-31	2304	2403.78287
1980-04-30	1712	2403.78287
1980-05-31	1471	2403.78287

20. Simple exponential smoothing forecast for fitted values train set

Time_Stamp	Sparkling	predict_spark
1991-01-31	1902	2403.78287
1991-02-28	2049	2403.78287
1991-03-31	1874	2403.78287
1991-04-30	1279	2403.78287
1991-05-31	1432	2403.78287

21. Simple exponential smoothing forecast for fitted values test set



27. Sparkling - SES forecast (Autofit Alpha: 0.0) on train and test

Model evaluation		
	RMSE	MAPE
Train	1298.484	40.36
Test	1275.082	38.9

Simple Exponential Smoothing is applied if the time-series has neither a trend nor seasonality, which is not the case with the given data

• The forecasting using smoothing levels of alpha between 0 and 1 are as below, where the smoothing levels are passed manually

• For alpha value closer to 1, forecast follows the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened

• For Sparkling, test RMSE is found to be higher for values closer to zero, which is same as in Simple average forecast

On the second iteration, the model was ran without passing a value for alpha and used parameters 'optimized=True, use_brute=True'

• The autofit model picked 0.0 as the smoothing parameter and retuned consistent RMSE values in train and test datasets, which is higher in accuracy than in first iteration

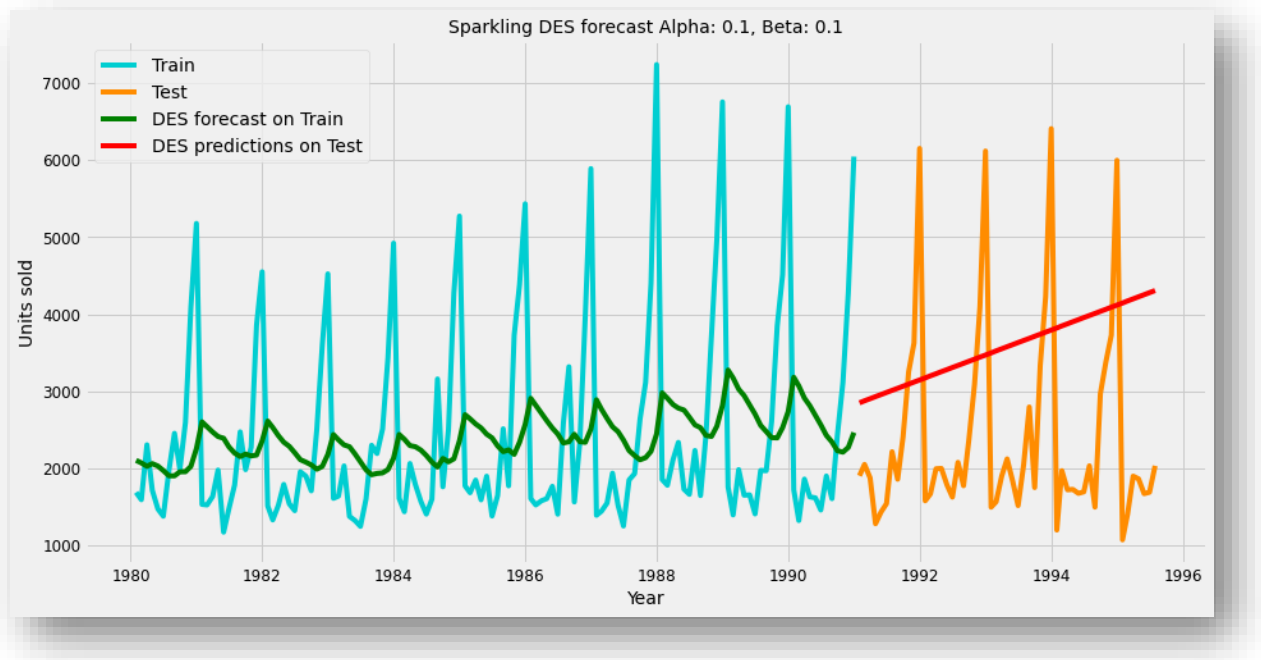
• As the smoothing level is 0.0, we got a completely smoothened out forecast with an initial value 2403.79 applied across the series

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

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	1363.47	44.26	1779.43	67.23
1	0.1	0.2	1401.76	45.65	2599.79	95.44
10	0.2	0.1	1412.03	46.62	3611.77	135.41
2	0.1	0.3	1435.33	46.85	4290.13	155.32
20	0.3	0.1	1428.27	46.92	5908.19	223.50

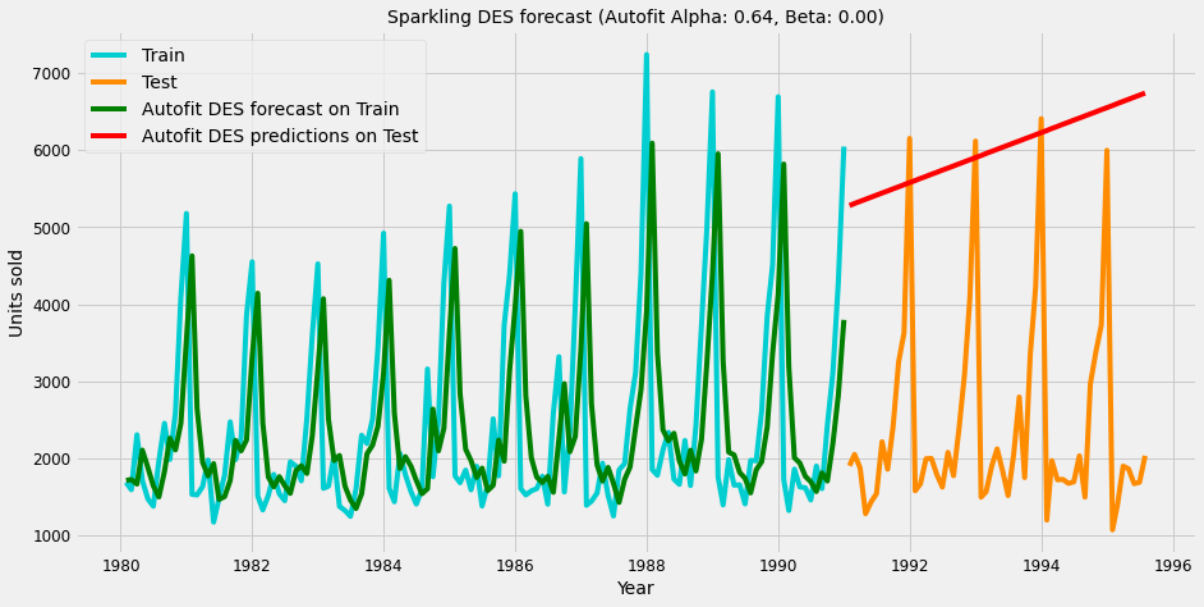
	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	1363.47	44.26	1779.43	67.23
1	0.1	0.2	1401.76	45.65	2599.79	95.44
10	0.2	0.1	1412.03	46.62	3611.77	135.41
2	0.1	0.3	1435.33	46.85	4290.13	155.32
3	0.1	0.4	1471.35	48.26	6041.56	219.06

22. Double exponential smoothing sorted train and test RMSE and MAPE heads to get best scores (Alpha 0.1, Beta 0.1)



28. Sparkling DES forecast (Alpha 0.1, Beta 0.1)

Letting the DES model auto fit



29. Sparkling DES forecast (Autofit Alpha 0.64, Beta 0.00)

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.1	1363.47000	44.26	1779.430000	67.23
1	0.100000	0.2	1401.76000	45.65	2599.790000	95.44
10	0.200000	0.1	1412.03000	46.62	3611.770000	135.41
100	0.647814	0.0	1337.48427	39.11	3850.779835	152.05
2	0.100000	0.3	1435.33000	46.85	4290.130000	155.32

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.1	1363.47000	44.26	1779.430000	67.23
1	0.100000	0.2	1401.76000	45.65	2599.790000	95.44
10	0.200000	0.1	1412.03000	46.62	3611.770000	135.41
100	0.647814	0.0	1337.48427	39.11	3850.779835	152.05
2	0.100000	0.3	1435.33000	46.85	4290.130000	155.32

23. Double exponential smoothing sorted test RMSE and MAPE heads
to get best scores (autofit Alpha 0.64, Beta 0.0)

- Auto-fitted model shows better RMSE and MAPE in train, but not the best fit in test set

	Test RMSE	Test MAPE
DES Alpha 0.1,Beta 0.1	1779.430000	67.23
DES Alpha 0.6,Beta 0.0	3850.779835	152.05

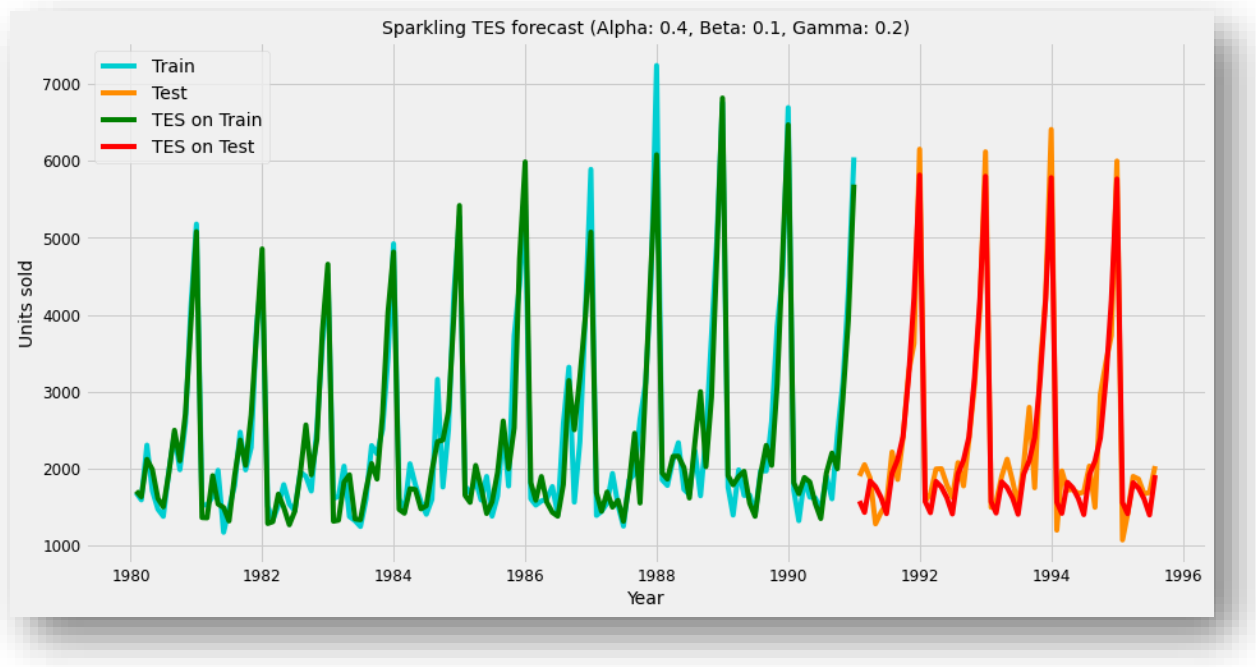
• The Double Exponential Smoothing models is applicable when data has trend, but no seasonality.
• Sparkling data contain slight trend component and very significant seasonality
• In first iteration, smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values given later with alpha 0.1 and beta 0.1
• On the second iteration the model was allowed to chose the optimized values using parameters 'optimized=True, use_brute=True'
• The autofit model retuned higher accuracy in train dataset, but faired poorly in test, compared with the values in manual iteration
• The model evaluation parameters of top three models from manual iteration and the autofit models are as given above
• The best model chosen as final one is with alpha 0.1 and beta 0.1

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

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
301	0.4	0.1	0.2	373.281410	11.05	312.211065	10.20
211	0.3	0.2	0.2	377.346884	11.23	315.195008	10.07
300	0.4	0.1	0.1	370.807398	11.06	318.281165	10.00
402	0.5	0.1	0.3	390.181794	11.54	325.690520	9.99
403	0.5	0.1	0.4	401.059753	11.55	343.321915	11.07

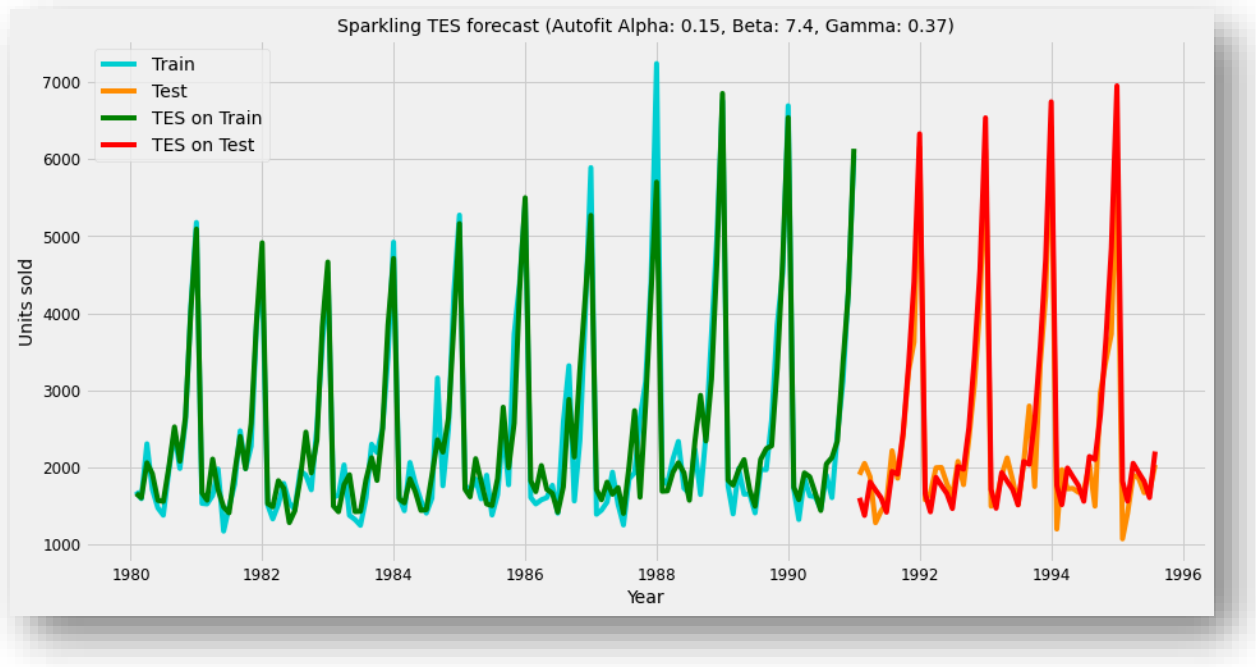
	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
402	0.5	0.1	0.3	390.181794	11.54	325.690520	9.99
300	0.4	0.1	0.1	370.807398	11.06	318.281165	10.00
211	0.3	0.2	0.2	377.346884	11.23	315.195008	10.07
301	0.4	0.1	0.2	373.281410	11.05	312.211065	10.20
401	0.5	0.1	0.2	384.608362	11.44	344.182644	10.67

24. Triple exponential smoothing sorted train and test RMSE and MAPE heads to get best scores (alpha 0.4, beta 0.1, gamma 0.2)



30. Sparkling TES forecast (Alpha 0.4, Beta 0.1, Gamma 0.2)

Attempting to autofit the TES model



31. Sparkling TES forecast (Autofit Alpha 0.15, Beta 7.4, Gamma 0.37)

•The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Sparkling data contain slight trend and significant seasonality

• In first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.4, beta 0.1 and gamma 0.2

• On the second iteration the model was allowed to chose the optimized values using parameters 'optimized=True, use_brute=True'

• The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test

• The model evaluation parameters of the best models are given as above, including one from the autofit iteration. The best model is the one with alpha 0.4, beta 0.1 and gamma 0.2

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
301	0.4	0.1	0.2	373.281410	11.05	312.211065	10.20
211	0.3	0.2	0.2	377.346884	11.23	315.195008	10.07
300	0.4	0.1	0.1	370.807398	11.06	318.281165	10.00
402	0.5	0.1	0.3	390.181794	11.54	325.690520	9.99
403	0.5	0.1	0.4	401.059753	11.55	343.321915	11.07

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
402	0.5	0.1	0.3	390.181794	11.54	325.690520	9.99
300	0.4	0.1	0.1	370.807398	11.06	318.281165	10.00
211	0.3	0.2	0.2	377.346884	11.23	315.195008	10.07
301	0.4	0.1	0.2	373.281410	11.05	312.211065	10.20
401	0.5	0.1	0.2	384.608362	11.44	344.182644	10.67

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
996	1.00	1.0	0.70	93702.603162	1031.58	1.383388e+06	25559.44
997	1.00	1.0	0.80	104677.046328	1039.17	3.959727e+06	55484.79
998	1.00	1.0	0.90	161779.762665	1427.45	5.643065e+05	8672.79
999	1.00	1.0	1.00	239920.545915	1304.98	1.211460e+05	3173.80
1000	0.15	0.0	0.37	353.379117	10.18	3.841977e+02	11.94

25. Triple exponential smoothing sorted test RMSE and MAPE heads to get best scores (alpha 0.15, beta 7.4, gamma 0.37), along with tail

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	312.211065	10.20
TES Alpha 0.15, Beta 0.00, Gamma 0.37	384.197750	11.94

Model evaluation

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	312.211065	10.20
TES Alpha 0.15, Beta 0.00, Gamma 0.37	384.197750	11.94
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
SES Alpha 0.00	1275.081813	38.90
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.430000	67.23
DES Alpha 0.6,Beta 0.0	3850.779835	152.05
NaiveModel	3864.279352	152.87

26. Various exponential and other models

Model comparison

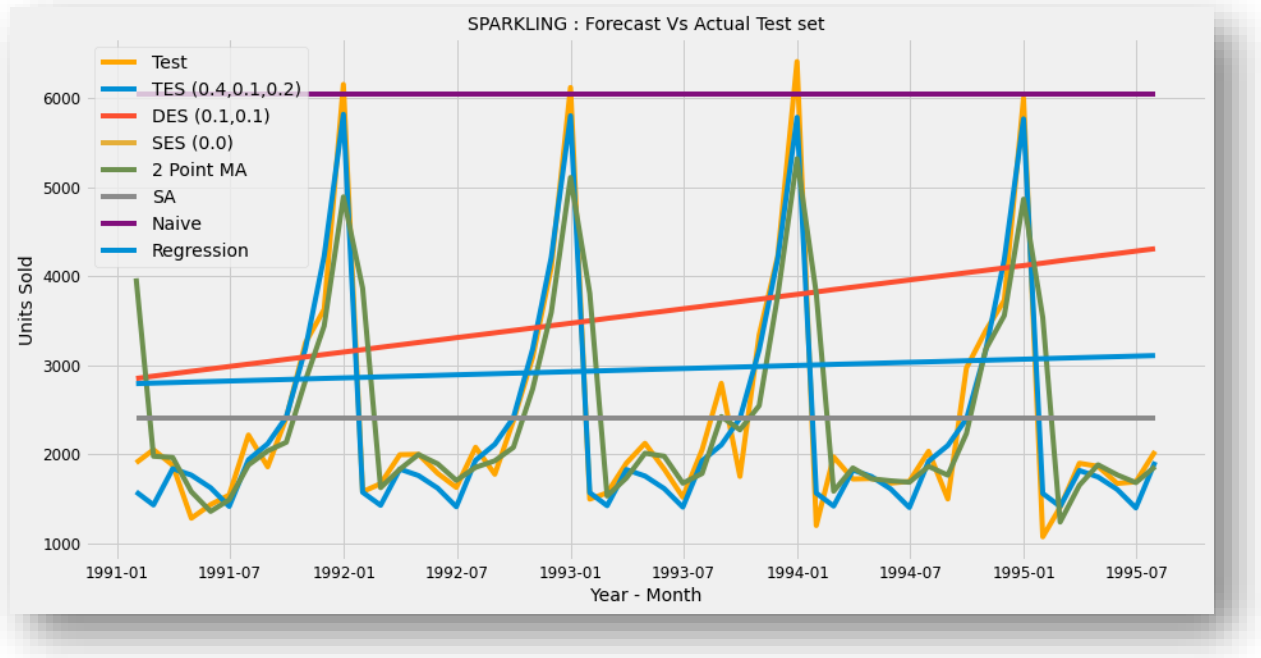
• The accuracy of the time-series forecast models built in the previous sections of this report is on the left, sorted by RMSE in test data

• The plot of the forecasts fitted on to the test data is also given

• From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data

• 2-point trailing moving average model also fits well with a slight lag in test dataset

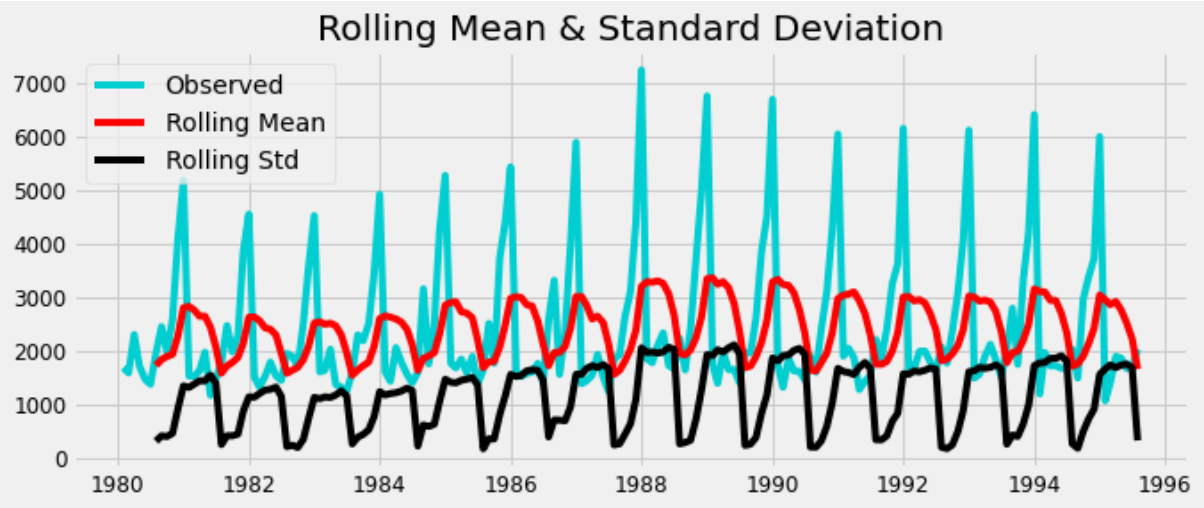
Plot all above models



32. Sparkling: Forecast vs Actual Test Data (various-model time series plot)

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.

Original series

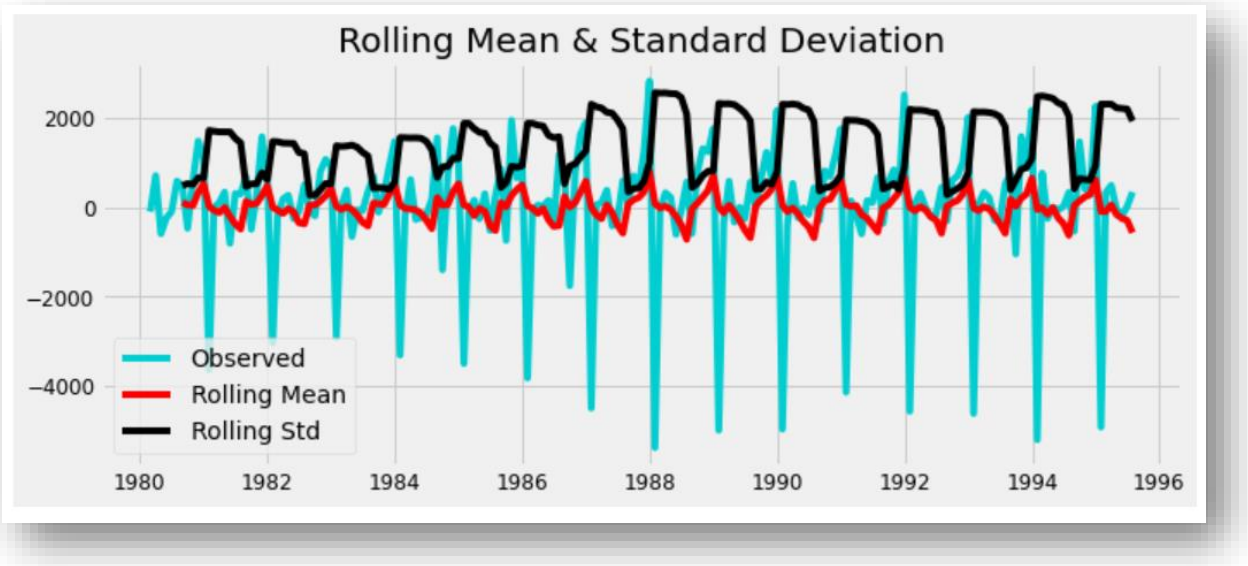



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

33. Dickey-Fuller Test (rolling mean and standard deviation), on original series, at 5% significant level

- Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determines the presence of unit root in the series to understand if the series is stationary or not
- **Null Hypothesis:** The series has a unit root, that is series is non-stationary
- **Alternative Hypothesis:** The series has no unit root, that is series is stationary
- If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary
- The ADF test on the original Sparkling series retuned the values where p-value is greater than alpha .05 so we fail to reject the null hypothesis.

Differenced series



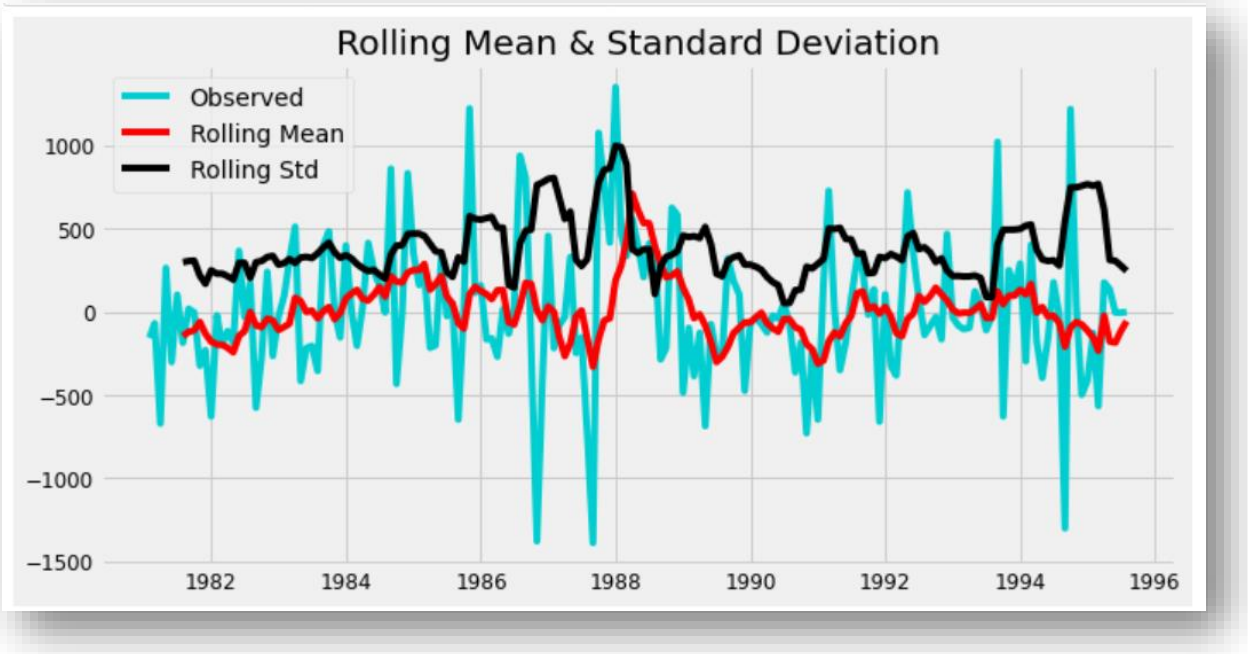

```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical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

34. Dickey-Fuller Test (rolling mean and standard deviation) with difference of order 1

We see that at 5% significant level the time series is non-stationary. But the seasonality is multiplicative as the standard deviation and mean vary according to the change in trend. Let us take a difference of order 1 and check whether the time series is stationary or not.

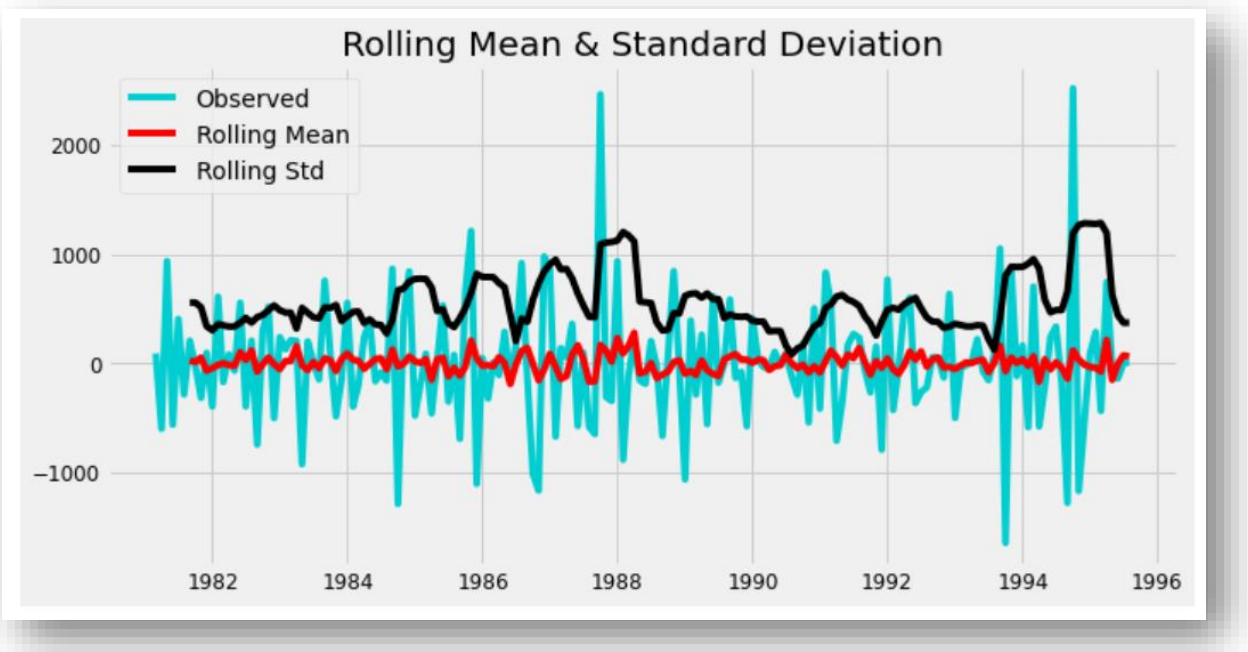
- Differencing of order 1 is applied on the Sparkling series and tested for stationarity. At an order of differencing 1, the series is found to be stationary
- The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if its multiplicative or additive in character
- The altitude of rolling mean and std dev is seen changing according to change in slope, which indicates multiplicity
- The ADF test is also done in this exercise with logarithmic transformation of the train data and differencing of seasonal order (12), to understand if removing the multiplicity of the seasonal component will have an impact on the accuracy of model

Difference of original series




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.460165
p-value                          0.000232
#Lags Used                      11.000000
Number of Observations Used    163.000000
Critical Value (1%)             -3.471119
Critical Value (5%)             -2.879441
Critical Value (10%)            -2.576314
dtype: float64
```

35. Dickey-Fuller Test (rolling mean and standard deviation)
original series with differencing of seasonal order (12)

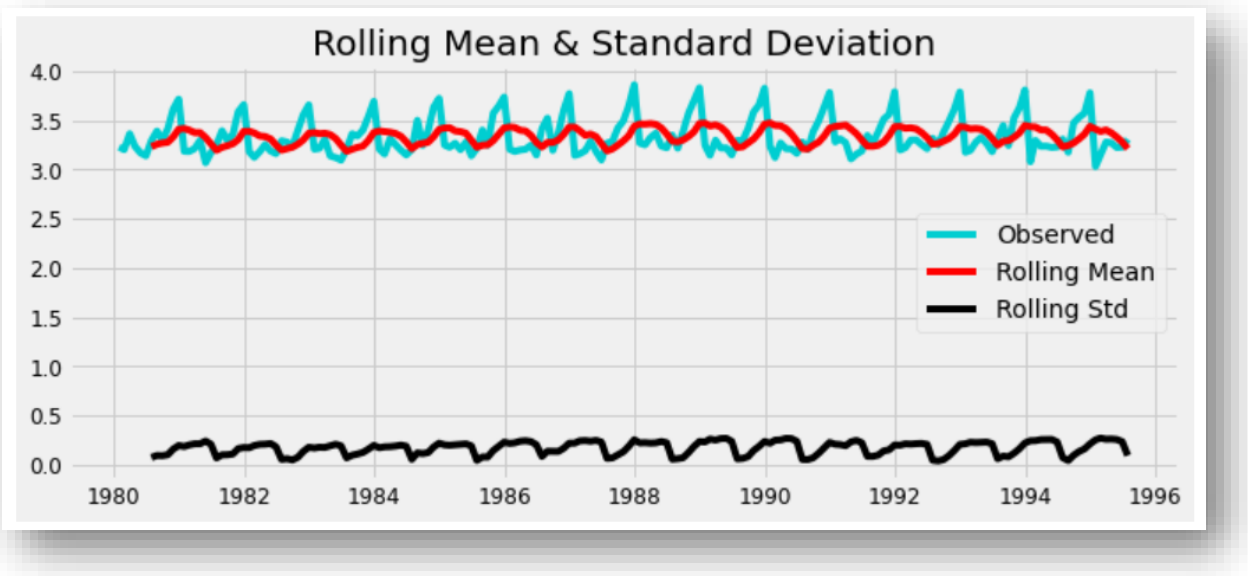



```
Results of Dickey-Fuller Test:
Test Statistic                  -5.113533
p-value                          0.000013
#Lags Used                      11.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value (10%)            -2.576373
dtype: float64
```

36. ADF (rolling mean and standard deviation) with differencing:
Series after seasonal differencing + 1-order differencing

We see that at αα = 0.05 the Time Series is indeed stationary. But seasonality is multiplicative

Log of series

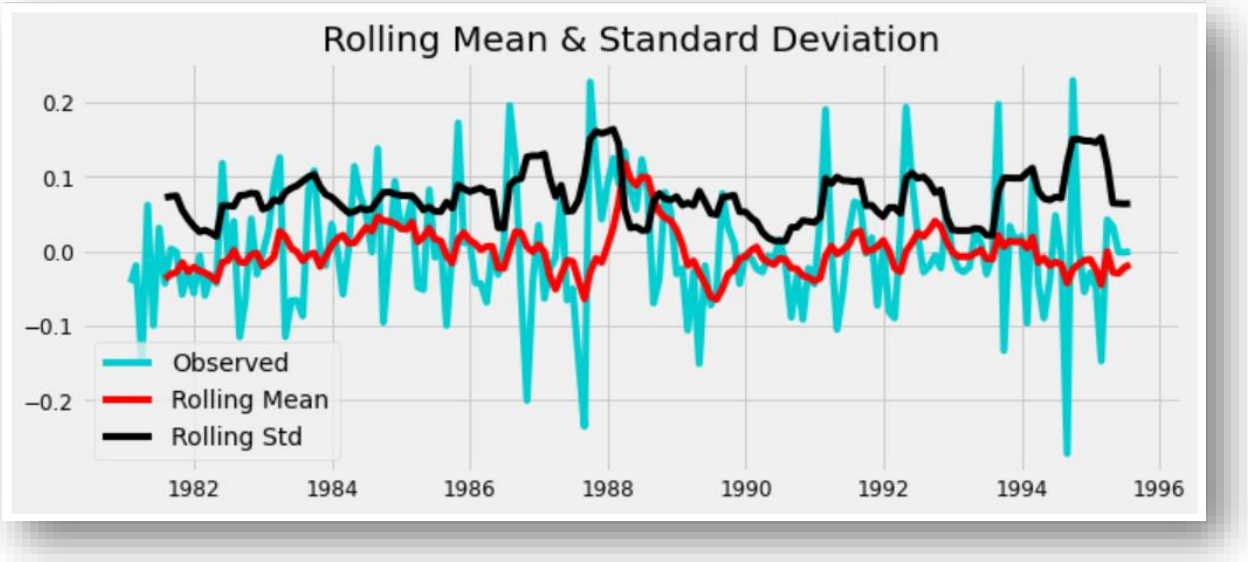



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.749630
p-value                          0.405740
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

37. ADF (rolling mean and standard deviation) with logarithmic transformation, using log10

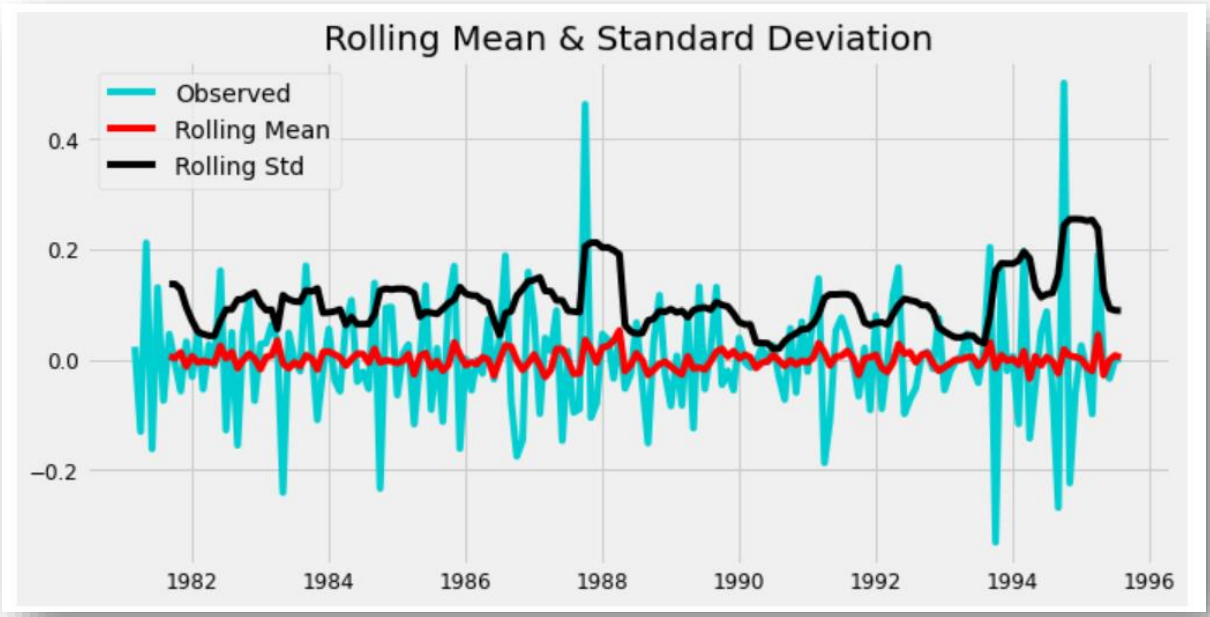
Seasonality is now additive but non-stationary

Difference of log of series




```
Results of Dickey-Fuller Test:
Test Statistic                  -5.183811
p-value                          0.000009
#Lags Used                      11.000000
Number of Observations Used    163.000000
Critical Value (1%)             -3.471119
Critical Value (5%)             -2.879441
Critical Value (10%)            -2.576314
dtype: float64
```

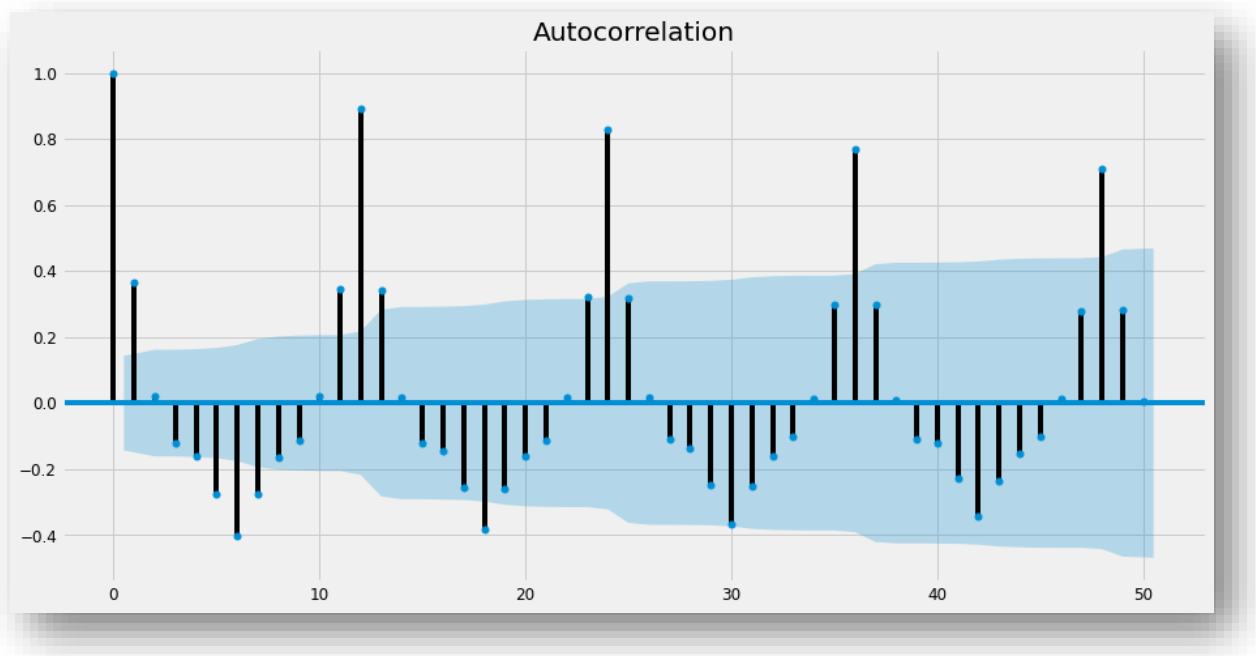
38. ADF (rolling mean and standard deviation) Difference of Log 10 Series after seasonal differencing



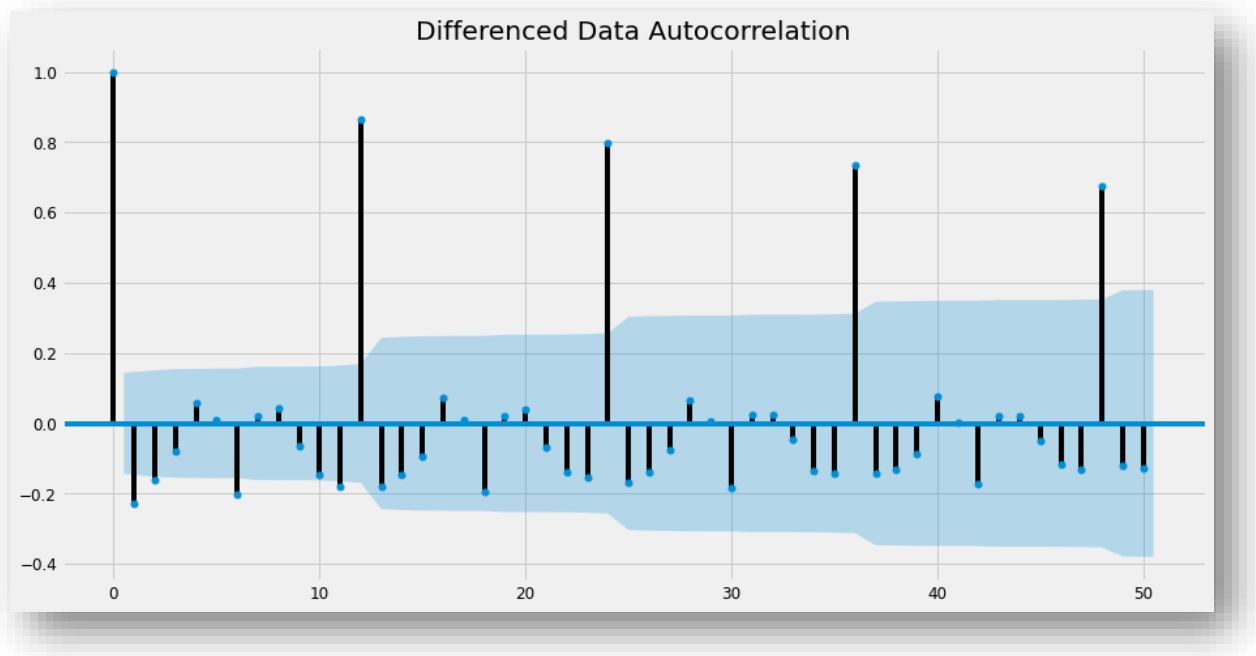

```
Results of Dickey-Fuller Test:
Test Statistic                  -5.254601
p-value                          0.000007
#Lags Used                      12.000000
Number of Observations Used    161.000000
Critical Value (1%)             -3.471633
Critical Value (5%)             -2.879665
Critical Value (10%)            -2.576434
dtype: float64
```

39. ADF (rolling mean and standard deviation) Difference of Log 10 Series after seasonal differencing + 1-order differencing

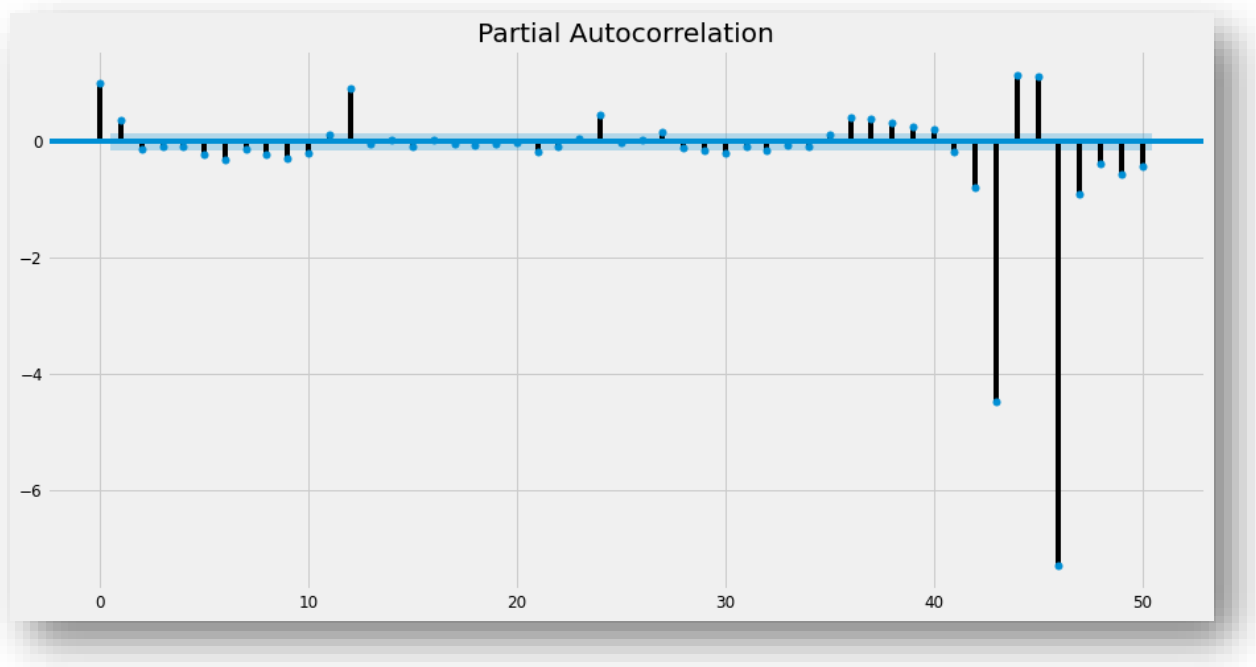
Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.



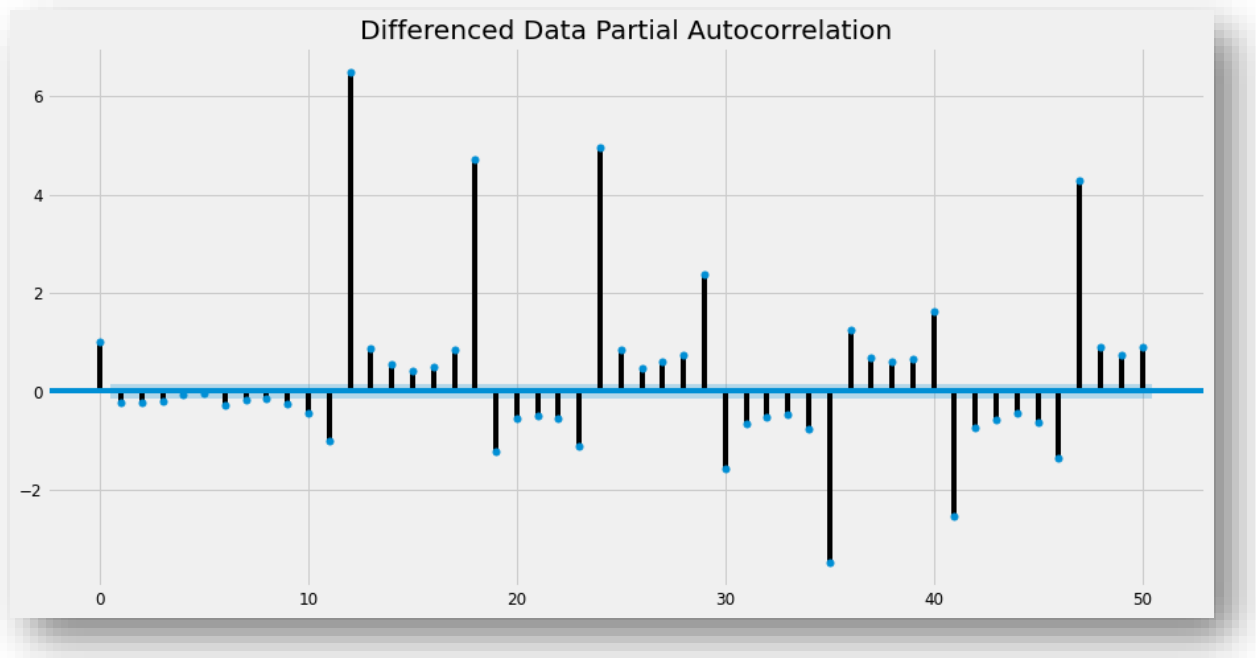
40. Sparkling – Autocorrelation



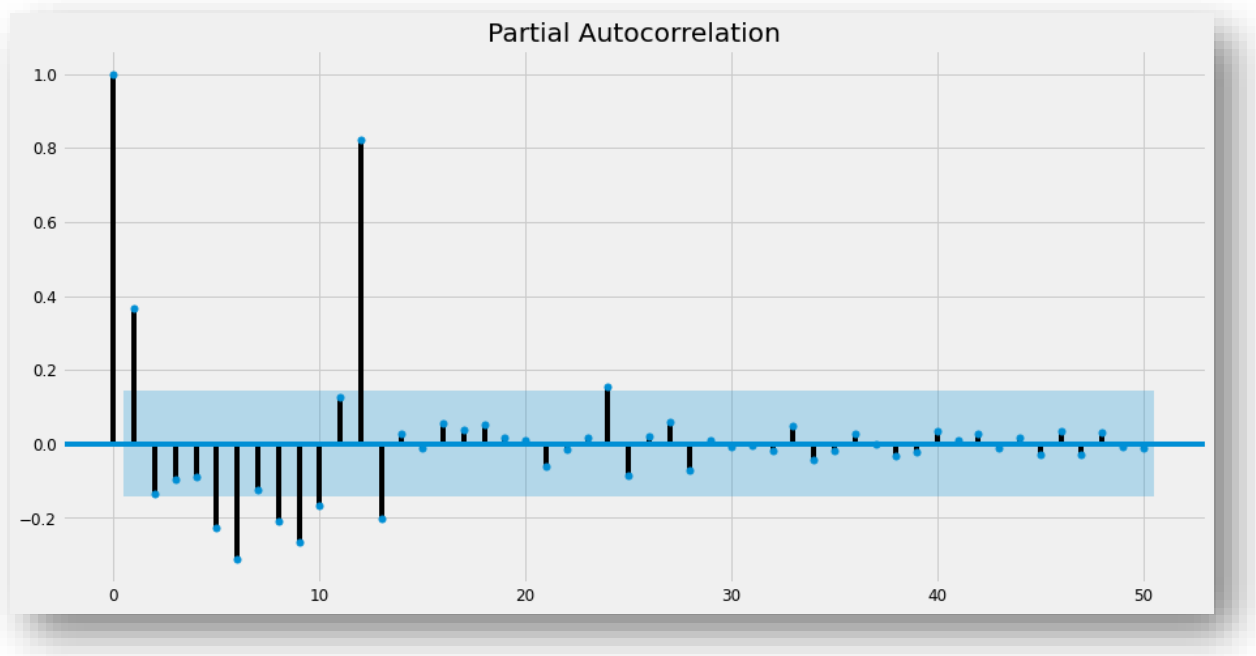
41. Sparkling - Differenced data autocorrelation



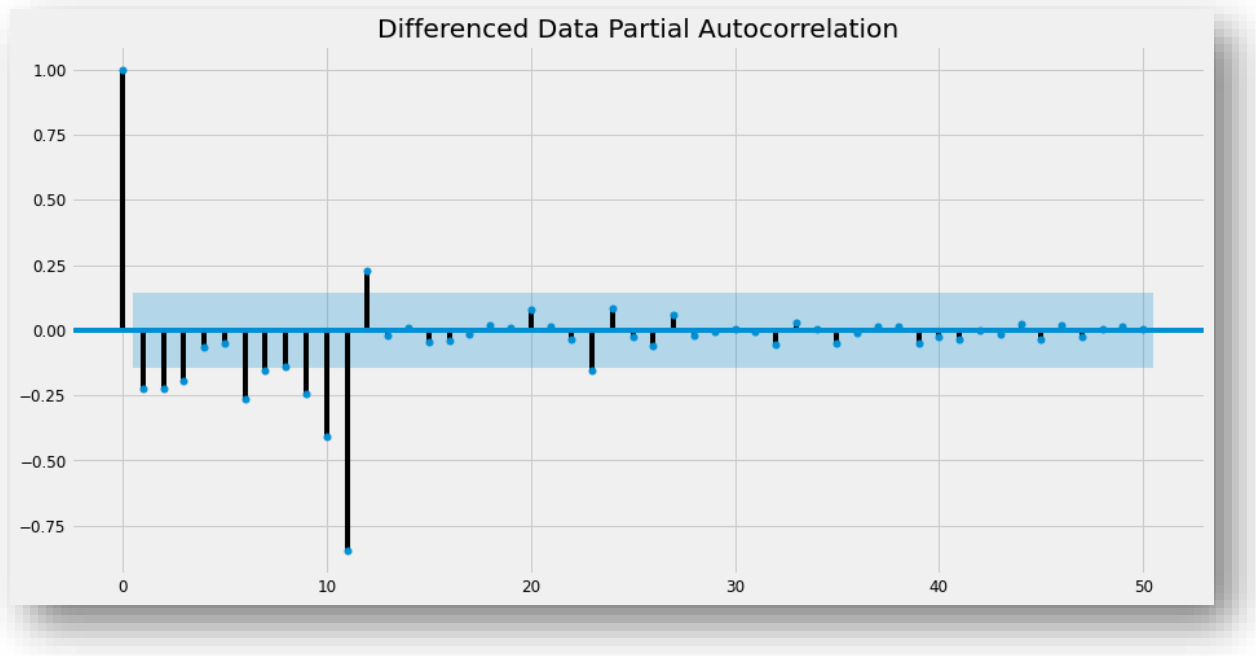
42. Sparkling - Partial Autocorrelation



43. Differenced Data Partial Autocorrelation



44. Sparking partial autocorrelation with ywmle method



45. Differenced data partial autocorrelation with ywmle method

From the above plots, we can say there is monthly seasonality in the data

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.

Model 8: SARIMA

AUTO SARIMA on original data

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 1, 1, 12)
Model: (0, 1, 2)(0, 1, 2, 12)
Model: (0, 1, 3)(0, 1, 3, 12)
Model: (1, 1, 0)(1, 1, 0, 12)
Model: (1, 1, 1)(1, 1, 1, 12)
Model: (1, 1, 2)(1, 1, 2, 12)
Model: (1, 1, 3)(1, 1, 3, 12)
Model: (2, 1, 0)(2, 1, 0, 12)
Model: (2, 1, 1)(2, 1, 1, 12)
Model: (2, 1, 2)(2, 1, 2, 12)
Model: (2, 1, 3)(2, 1, 3, 12)
Model: (3, 1, 0)(3, 1, 0, 12)
Model: (3, 1, 1)(3, 1, 1, 12)
Model: (3, 1, 2)(3, 1, 2, 12)
Model: (3, 1, 3)(3, 1, 3, 12)
```

Why no ARIMA

As the Sparkling series of data contains seasonality component, we choose to build a SARIMA model instead of an ARIMA model

ARIMA can be limited in forecasting extreme values. While the model is adept at modelling seasonality and trends, outliers are difficult to forecast for ARIMA for the very reason that they lie outside of the general trend as captured by the model.

27. Auto Sarima on original data, examples of some parameter combinations for model

	param	seasonal	AIC
252	(3, 1, 3)	(3, 1, 0, 12)	1213.282563
253	(3, 1, 3)	(3, 1, 1, 12)	1215.213343
220	(3, 1, 1)	(3, 1, 0, 12)	1215.898777
254	(3, 1, 3)	(3, 1, 2, 12)	1216.480085
236	(3, 1, 2)	(3, 1, 0, 12)	1216.859180

28. Auto Sarima models with best AIC scores

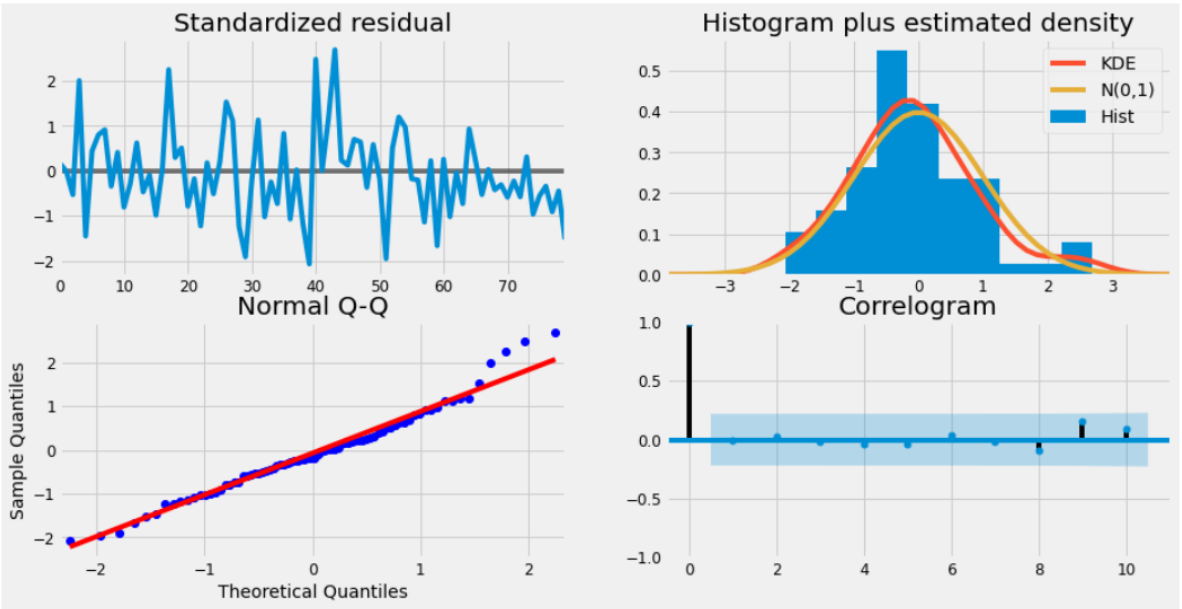
Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 3)x(3, 1, 0, 12)	Log Likelihood	-596.641
Date:	Fri, 08 Oct 2021	AIC	1213.283
Time:	08:02:03	BIC	1237.103
Sample:	0	HQIC	1222.833
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6138	0.176	-9.174	0.000	-1.959	-1.269
ar.L2	-0.6117	0.299	-2.046	0.041	-1.198	-0.026
ar.L3	0.0864	0.161	0.538	0.590	-0.228	0.401
ma.L1	0.9855	0.471	2.090	0.037	0.061	1.910
ma.L2	-0.8737	0.166	-5.266	0.000	-1.199	-0.549
ma.L3	-0.9465	0.489	-1.937	0.053	-1.904	0.011
ar.S.L12	-0.4519	0.142	-3.191	0.001	-0.729	-0.174
ar.S.L24	-0.2343	0.144	-1.624	0.104	-0.517	0.049
ar.S.L36	-0.1008	0.122	-0.829	0.407	-0.339	0.138
sigma2	1.839e+05	8.96e+04	2.052	0.040	8274.924	3.59e+05

Ljung-Box (Q):	23.20	Jarque-Bera (JB):	4.06
Prob(Q):	0.98	Prob(JB):	0.13
Heteroskedasticity (H):	0.73	Skew:	0.48
Prob(H) (two-sided):	0.42	Kurtosis:	3.54

46. Auto SARIMA model (3,1,3)x(3,1,0,12): Results



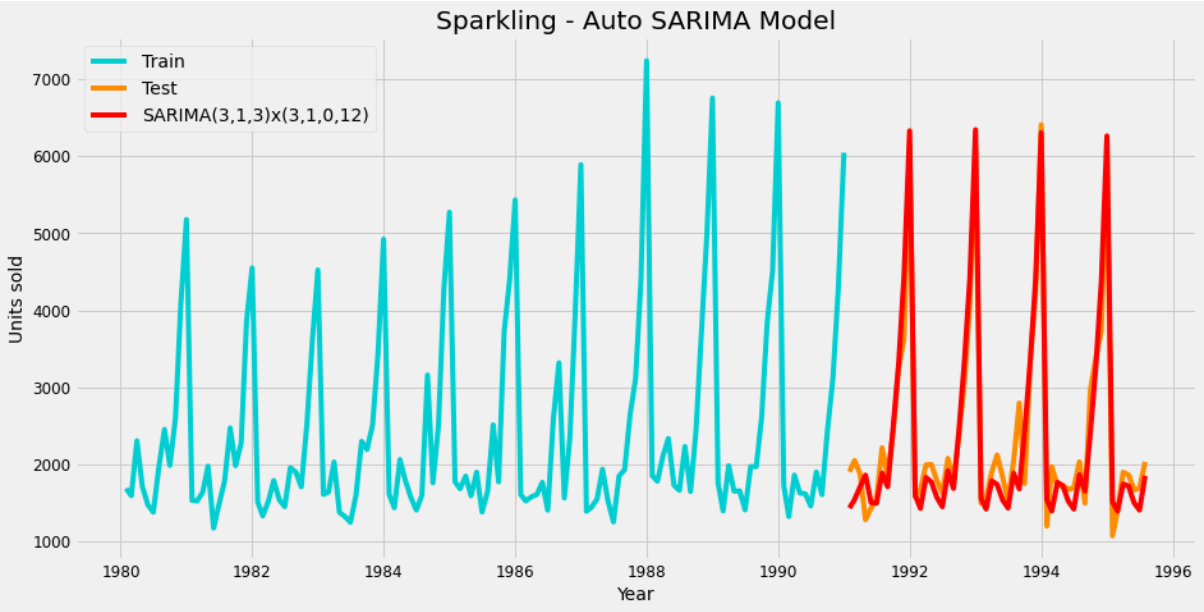
47. Auto SARIMA model (3,1,3)x(3,1,0,12): Diagnostics

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1430.435992	431.198057	585.303331	2275.568654
1	1540.310097	458.412112	641.838869	2438.781326
2	1707.333372	460.179283	805.398551	2609.268193
3	1858.864004	466.747778	944.055170	2773.672839
4	1501.503828	467.048273	586.106034	2416.901622

29. Auto Sarima (3,1,3)x(3,1,0,12) prediction summary

Time_Stamp	Sparkling	spark_forecasted
1991-01-31	1902	1430.435992
1991-02-28	2049	1540.310097
1991-03-31	1874	1707.333372
1991-04-30	1279	1858.864004
1991-05-31	1432	1501.503828

30. Predicted and true values of time series for Auto Sarima (3,1,3)x(3,1,0,12)



48. Sparkling Auto Sarima(3,1,3)x(3,1,0,12)

Model analysis

• Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as an element of multiplicity in seasonality is suspected

• The model built with original data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model

• The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values

• The top three models with lowest AIC values are as given. As per the AIC criteria, the optimum values for final SARIMA model selected is (3, 1, 3)x(3, 1, 0, 12)

• The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution

• The Normal Q -Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line

• The correlogram shows the autocorrelation of the residuals and there are no significant lags above the confidence index

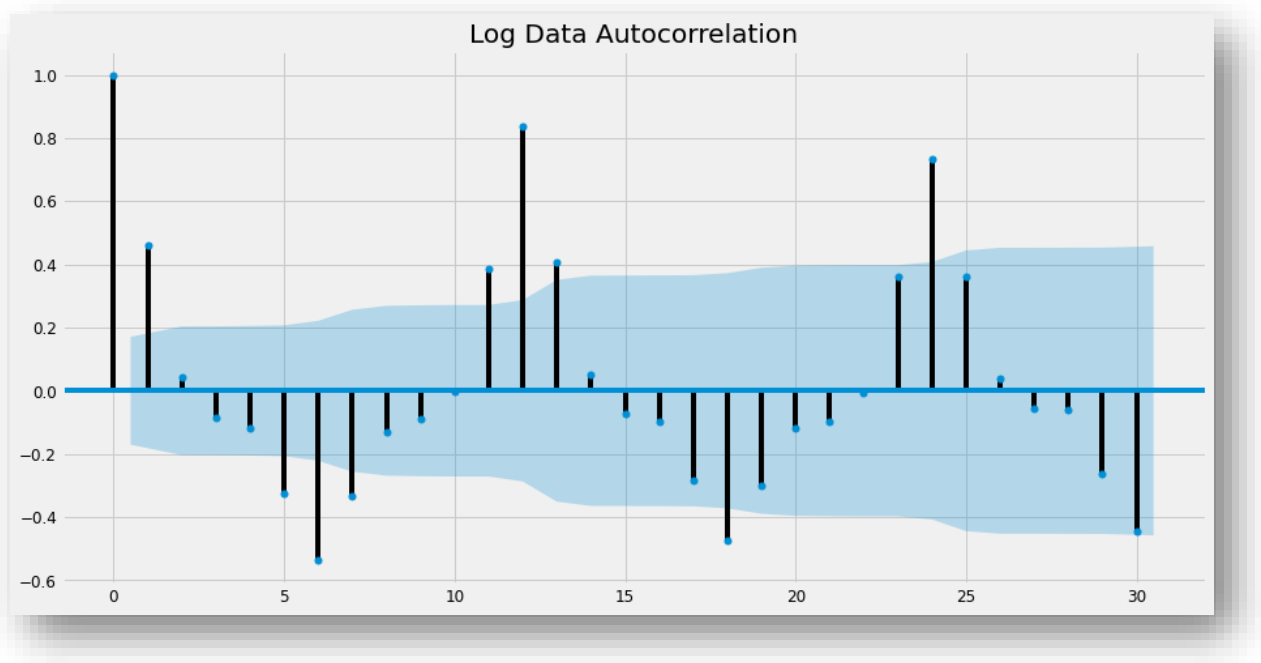
• The RMSE and MAPE values of the automated SARIMA models built are given here

• The diagnostics plot of the selected model is given in the next slide

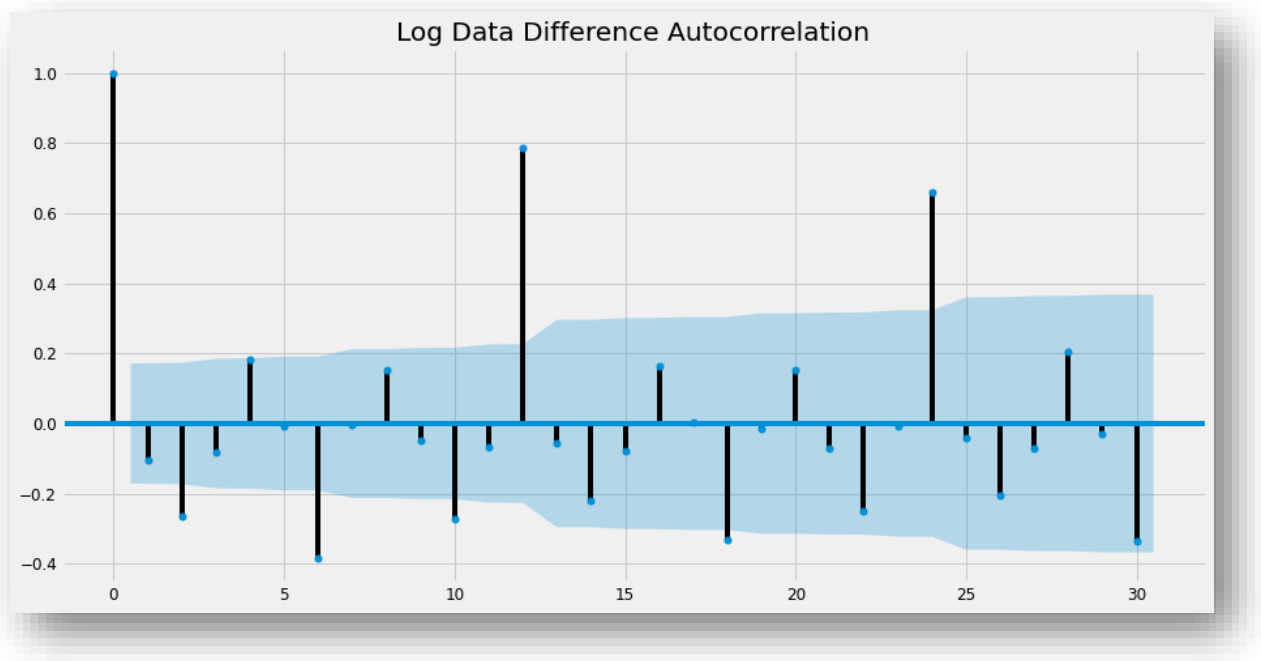
• Looking at the model summary helps us infer that AR(1), MA(1), MA(3), MA(2) terms have the highest absolute weightage.

• From the p-values it can be inferred that terms AR(1), AR(2), MA(1), MA(2), MA(3) and seasonal AR(1) are significant terms, as their values are below 0.05

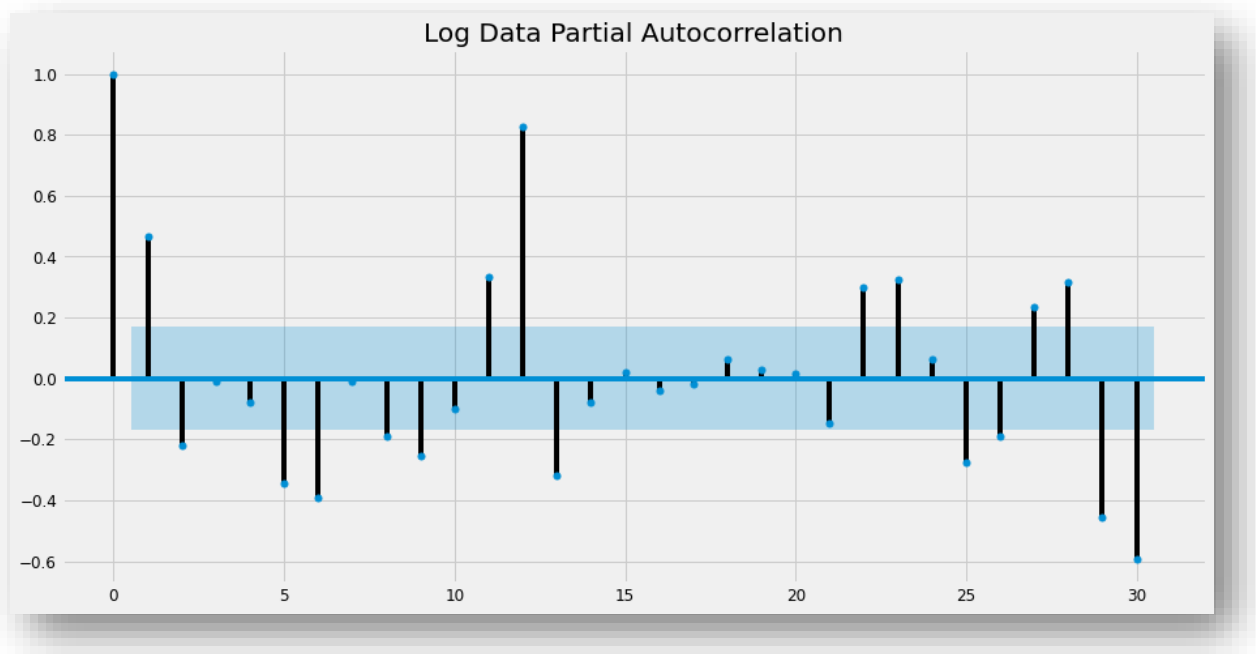
AUTO SARIMA on Log



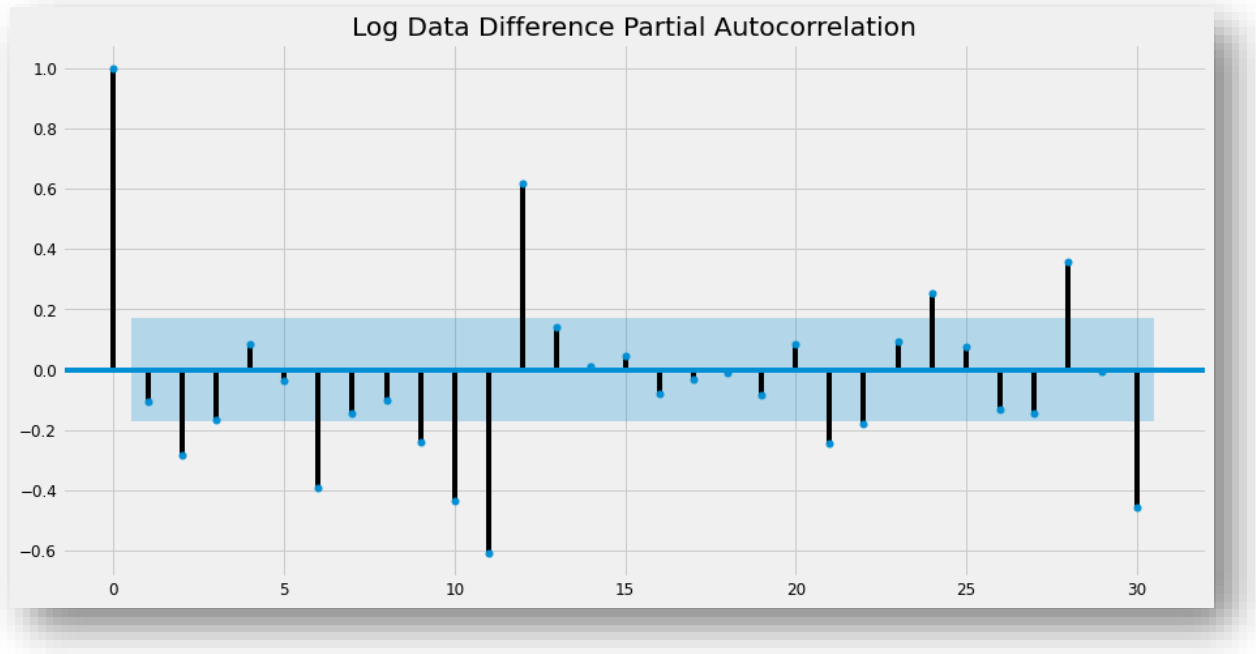
49. Log Data Autocorrelation



50. Log data difference autocorrelation



51. Log data partial autocorrelation



52. Log data difference partial autocorrelation

We see that there can be a seasonality of 12. We will run our auto SARIMA models by setting seasonality as 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)(0, 1, 0, 12)
Model: (1, 1, 1)(0, 1, 1, 12)
Model: (1, 1, 2)(0, 1, 2, 12)
Model: (2, 1, 0)(1, 0, 0, 12)
Model: (2, 1, 1)(1, 0, 1, 12)
Model: (2, 1, 2)(1, 0, 2, 12)
```

31. Auto Sarima on log10, examples of some parameter combinations for model

	param	seasonal	AIC
25	(0, 1, 1)	(1, 0, 1, 12)	-284.472032
79	(1, 1, 1)	(1, 0, 1, 12)	-282.517333
43	(0, 1, 2)	(1, 0, 1, 12)	-281.567994
97	(1, 1, 2)	(1, 0, 1, 12)	-279.611721
133	(2, 1, 1)	(1, 0, 1, 12)	-278.288233

32. Auto Sarima on log10: Models with best AIC scores

Statespace Model Results

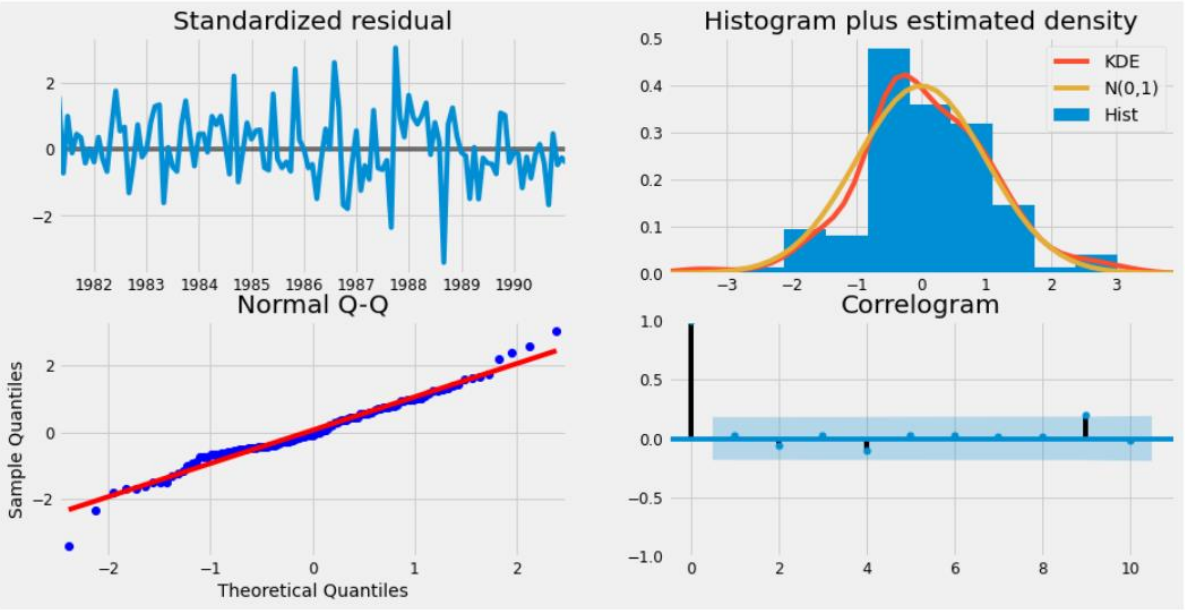
Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Log Likelihood	146.236
Date:	Fri, 08 Oct 2021	AIC	-284.472
Time:	08:04:52	BIC	-273.423
Sample:	01-31-1980	HQIC	-279.986
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8966	0.045	-19.862	0.000	-0.985	-0.808
ar.S.L12	1.0112	0.020	49.871	0.000	0.971	1.051
ma.S.L12	-0.6489	0.075	-8.629	0.000	-0.796	-0.502
sigma2	0.0045	0.001	7.842	0.000	0.003	0.006

Ljung-Box (Q):	40.45	Jarque-Bera (JB):	5.26
Prob(Q):	0.45	Prob(JB):	0.07
Heteroskedasticity (H):	1.43	Skew:	-0.00
Prob(H) (two-sided):	0.27	Kurtosis:	4.04

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

53. Auto SARIMA (0,1,1)x(1,0,1,12) on log10 model: Results



54. Auto SARIMA (0,1,1)x(1,0,1,12) on log10 model: Diagnostics

Predict on the Test Set using this model

Time_Stamp	Sparkling	spark_forecasted
1991-01-31	1902	1430.435992
1991-02-28	2049	1540.310097
1991-03-31	1874	1707.333372
1991-04-30	1279	1858.864004
1991-05-31	1432	1501.503828

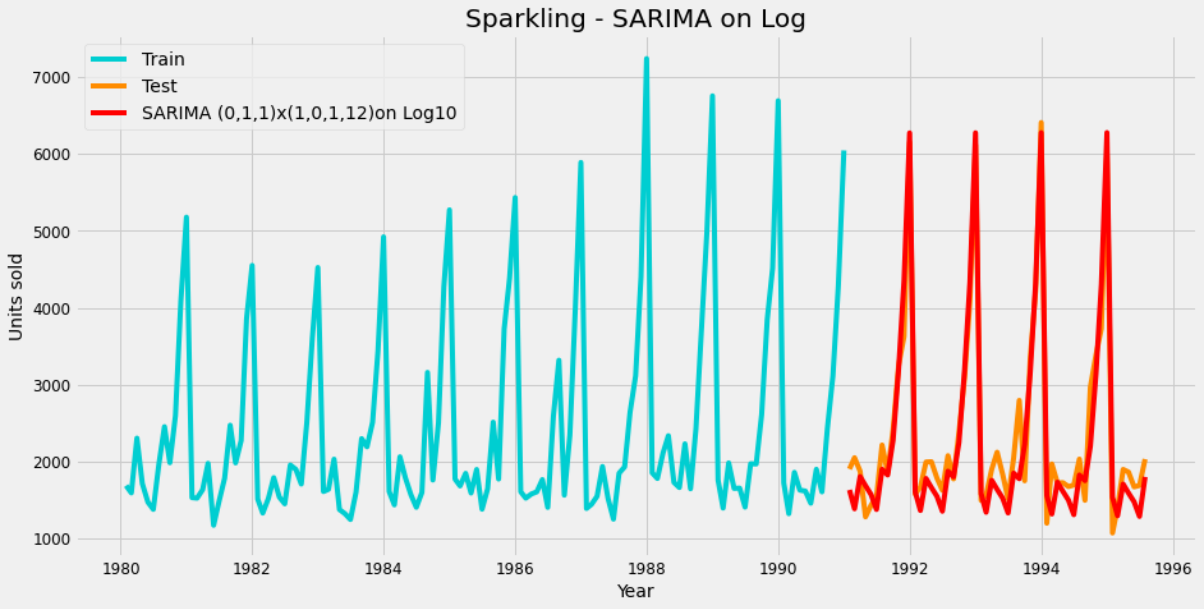
Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-31	3.212031	0.067108	3.080502	3.343560
1991-02-28	3.141307	0.067465	3.009078	3.273535
1991-03-31	3.256285	0.067821	3.123359	3.389211
1991-04-30	3.226732	0.068175	3.093112	3.360351
1991-05-31	3.195787	0.068527	3.061478	3.330097

33. Auto Sarima (0,1,1)x(1,0,1,12)on log10: test head and prediction summary

Extract the predicted and true values of our time series. We need to change the scale of the logarithmic scale to the original scale by raising the predicted values to the power of 10

Time_Stamp	Sparkling	spark_forecasted	spark_log_forecasted
1991-01-31	1902	1430.435992	1629.412404
1991-02-28	2049	1540.310097	1384.544633
1991-03-31	1874	1707.333372	1804.202350
1991-04-30	1279	1858.864004	1685.511077
1991-05-31	1432	1501.503828	1569.594099

34. Predicted and true values of time series for
Auto Sarima (0,1,1)x(1,0,1,12) log10 transformation model

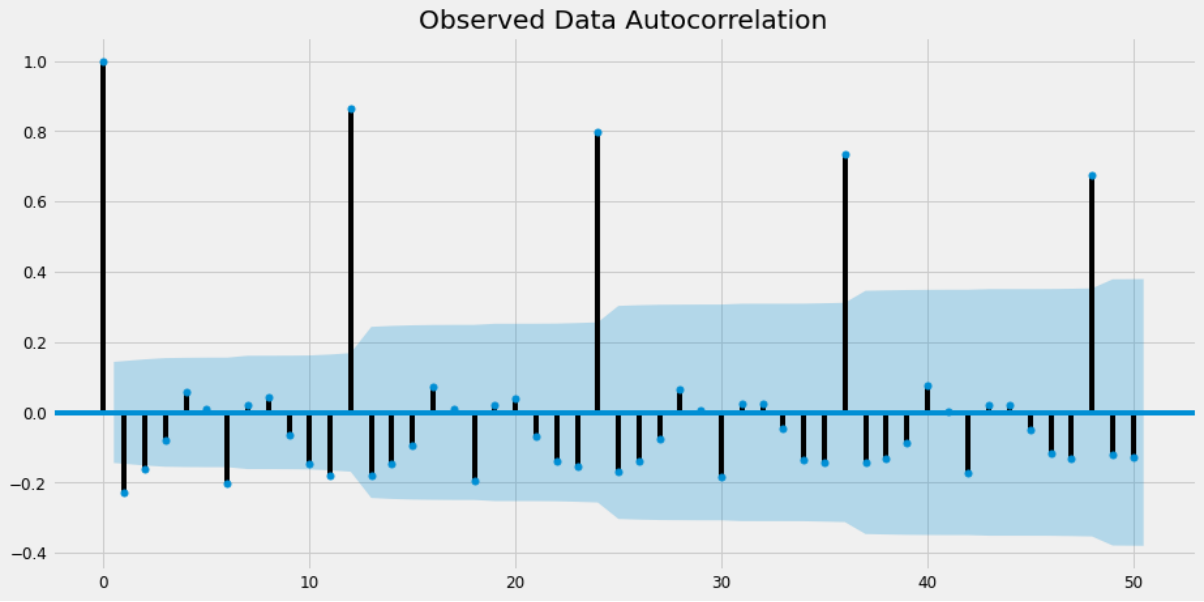


55. Sparkling - Auto SARIMA (0,1,1)x(1,0,1,12) with Log Transformation

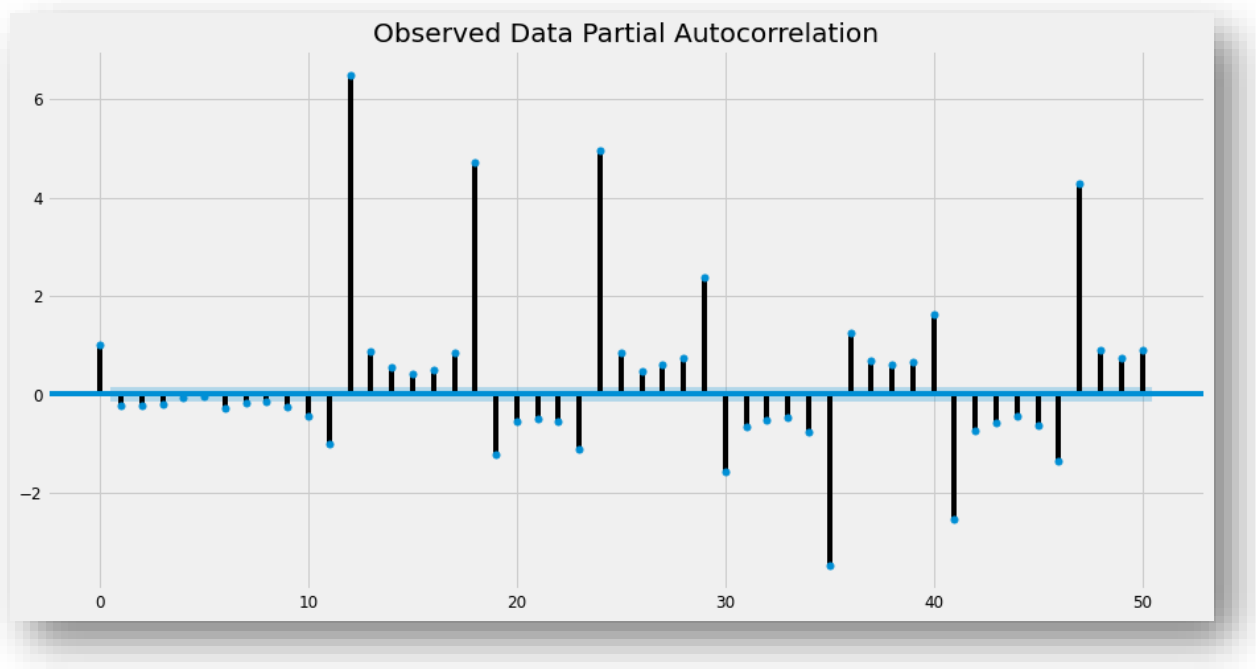
	Test RMSE	Test MAPE
Auto SARIMA(3,1,3)x(3,1,0,12)	331.695786	10.34
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.801449	11.19

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.

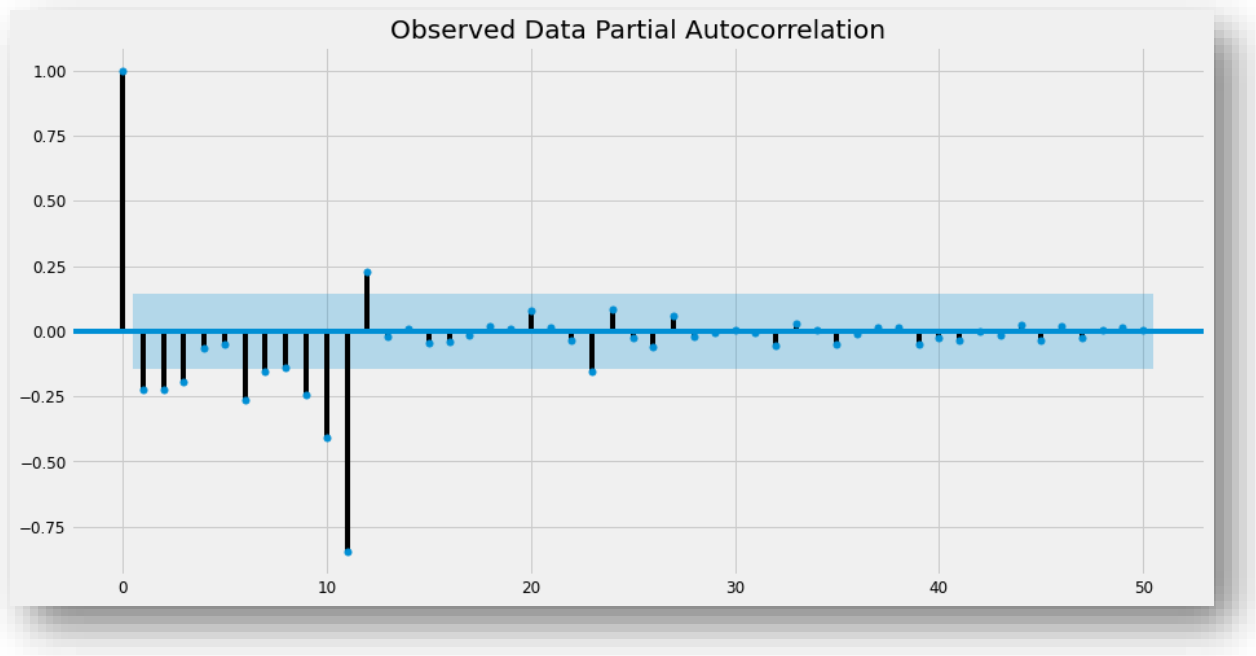
Manual SARIMA



56. For Manual Sarima: Observed data autocorrelation

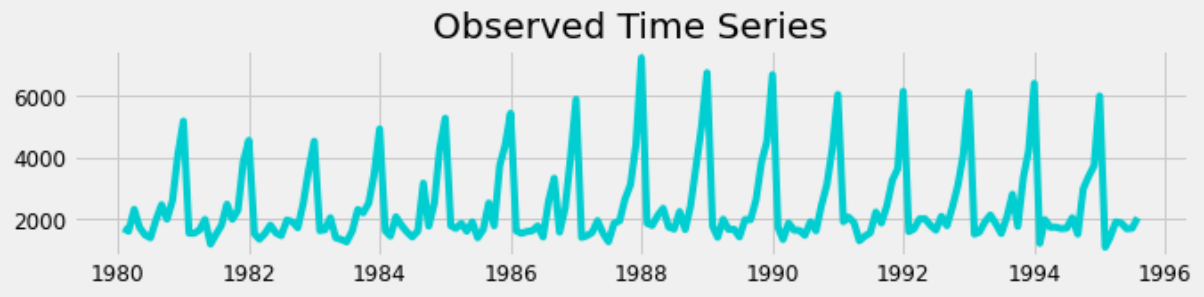


57. Observed data partial autocorrelation



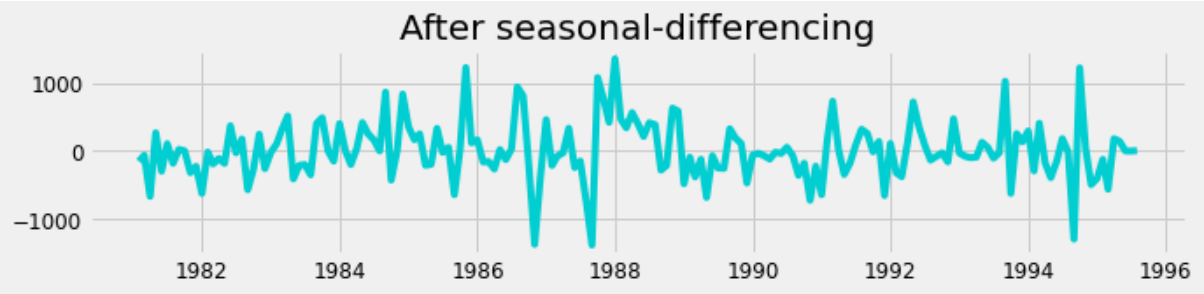
58. Observed data partial autocorrelation with ywmle method

We see that our ACF plot at the seasonal interval (12) does not taper off quickly. So, we go ahead and take a seasonal differencing of the original series. But before that, let us look at the original series.



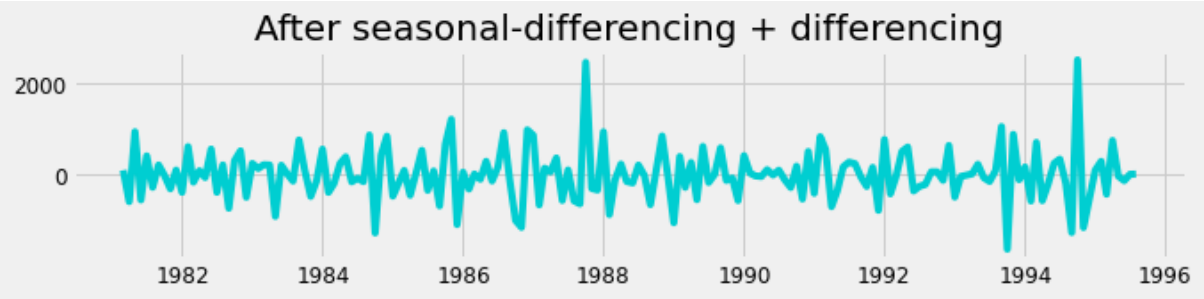
59. Observed time series before applying Manual SARIMA

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



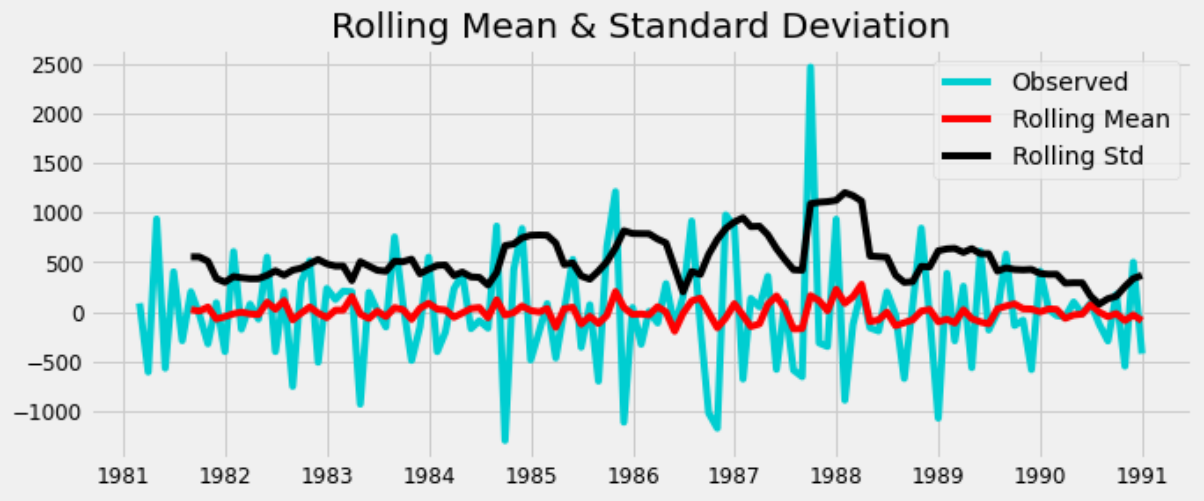
60. Series after seasonal differencing, before applying Manual SARIMA

The marginal trend in the data is still seen



61. Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

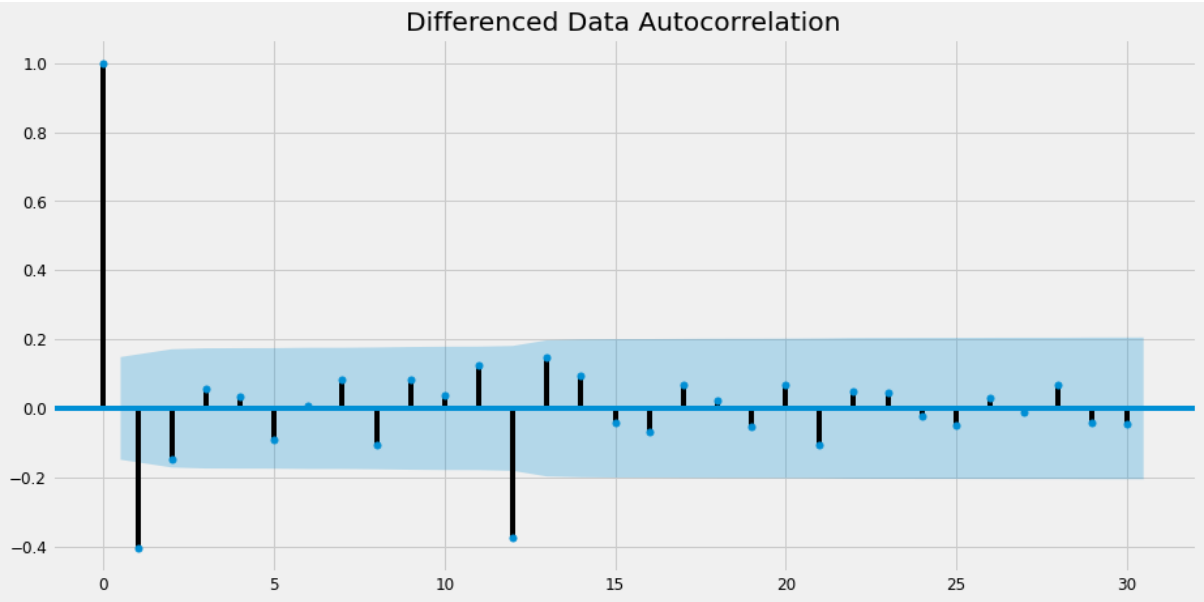
Now we see that there is almost no trend present in the data. Seasonality alone is present. Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.



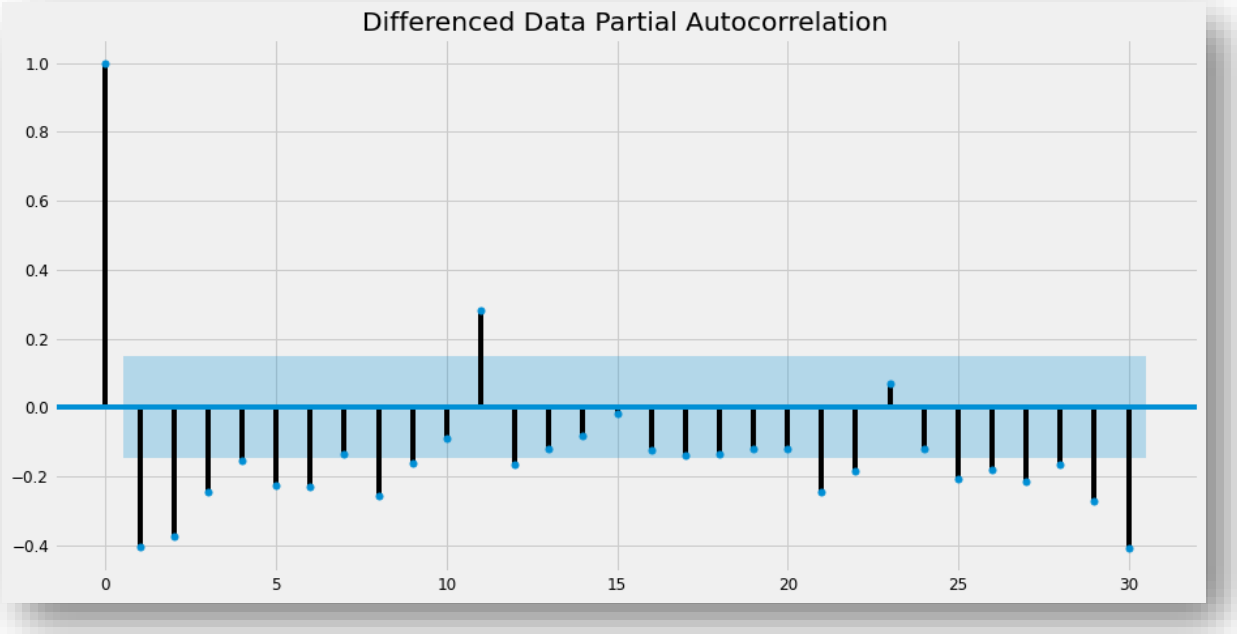

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
#Lags Used                      10.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

62. ADF Test of stationarity on Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA

Checking the ACF and the PACF plots for the new modified Time Series.



63. Differenced Data Autocorrelation



64. Differenced Data Partial Autocorrelation

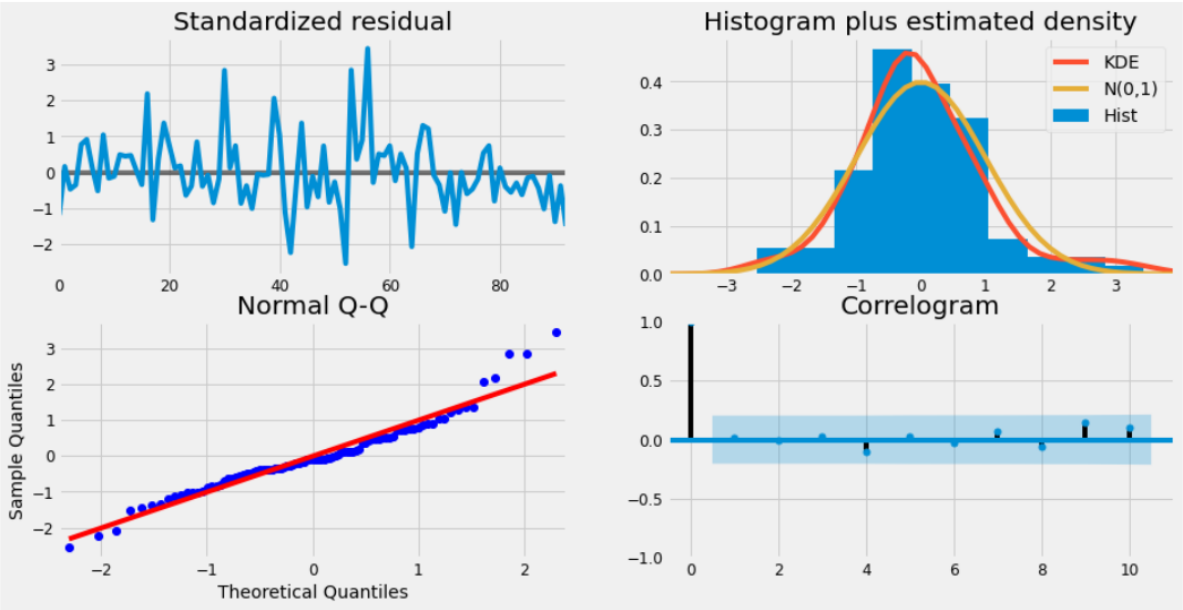
Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(1, 1, 2, 12)	Log Likelihood	-693.697
Date:	Fri, 08 Oct 2021	AIC	1403.394
Time:	08:04:59	BIC	1423.654
Sample:	0	HQIC	1411.574
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2229	0.130	1.713	0.087	-0.032	0.478
ar.L2	-0.0798	0.131	-0.607	0.544	-0.337	0.178
ar.L3	0.0921	0.122	0.756	0.450	-0.147	0.331
ma.L1	-1.0241	0.094	-10.925	0.000	-1.208	-0.840
ar.S.L12	-0.1992	0.866	-0.230	0.818	-1.897	1.499
ma.S.L12	-0.2109	0.881	-0.239	0.811	-1.938	1.516
ma.S.L24	-0.1299	0.381	-0.341	0.733	-0.877	0.617
sigma2	1.654e+05	2.62e+04	6.302	0.000	1.14e+05	2.17e+05

Ljung-Box (Q):	24.16	Jarque-Bera (JB):	19.66
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	0.81	Skew:	0.69
Prob(H) (two-sided):	0.56	Kurtosis:	4.78

54. Manual SARIMA (3,1,1)x(1,1,2,12): Results



65. Manual SARIMA (3,1,1)x(1,1,2,12): Diagnostics

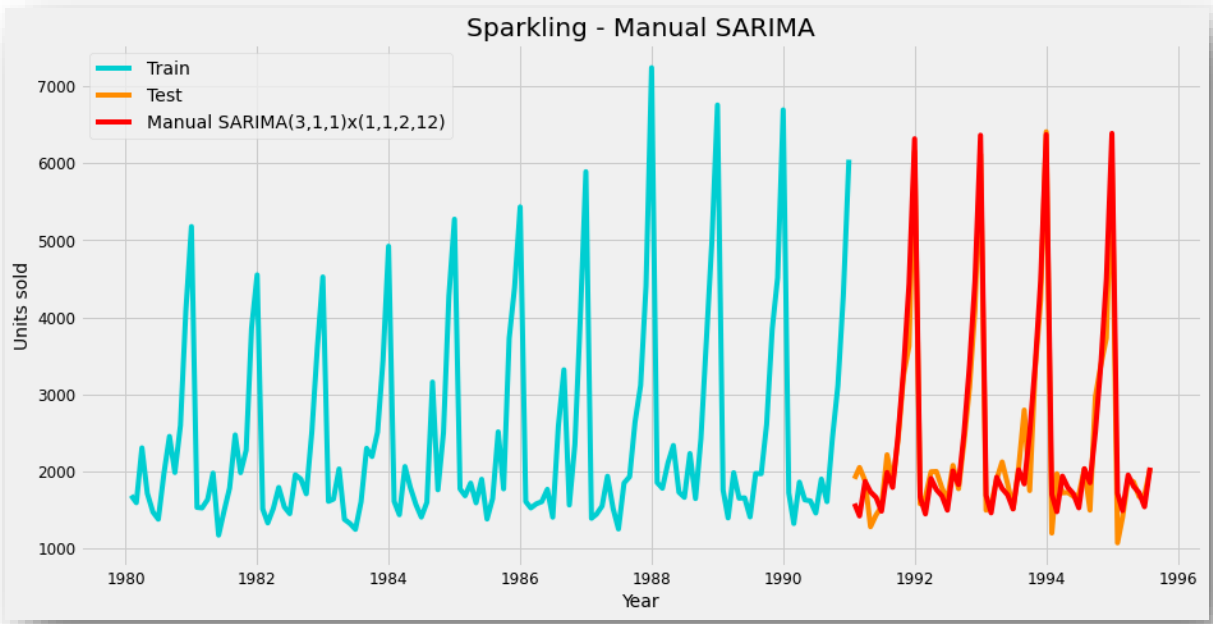
The model diagnostics plot looks okay.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1579.910092	416.594409	763.400053	2396.420130
1	1419.154450	429.113793	578.106870	2260.202030
2	1868.144134	429.104434	1027.114896	2709.173371
3	1731.472282	430.973097	886.780534	2576.164031
4	1659.822690	431.906077	813.302335	2506.343045

35. Manual Sarima (3,1,1)x(1,1,2,12), prediction summary

Time_Stamp	Sparkling	spark_forecasted	spark_log_forecasted	spark_manual_forecasted
1991-01-31	1902	1430.435992	1629.412404	1579.910092
1991-02-28	2049	1540.310097	1384.544633	1419.154450
1991-03-31	1874	1707.333372	1804.202350	1868.144134
1991-04-30	1279	1858.864004	1685.511077	1731.472282
1991-05-31	1432	1501.503828	1569.594099	1659.822690

36. Manual Sarima (3,1,1)x(1,1,2,12): predicted and true values



66. Sparkling - Manual SARIMA (3,1,1)x(1,1,2,12) forecast

From the ACF plot of the observed/ train data, it can be inferred that at seasonal interval of 12, the plot is not quickly tapering off. So a seasonal differencing of 12 has to be taken

• From the plots below an apparent slight trend is still existing after differencing of seasonal order of 12. With a further differencing of order one, no trend is present

• An ADF test need to be done to check the stationarity after the above differencing. With a p-value below alpha 0.05 and test statistic below critical values, it can be confirmed that the data is stationary

• ACF and PACF plots of the seasonal-differenced + one order differenced data is created to find the values for (p,d,q)x(P,D,Q), continued on next slide…

Here we have taken alpha = 0.05 and seasonal period as 12

• From the PACF plot it can be seen that till 3rd lag its significant before cut-off, so AR term 'p = 3' is chosen. At seasonal lag of 12, it almost cuts off, so seasonal AR 'P = 1'

• From ACF plot it can be seen that lag 1 is significant before it cuts off, so MA term 'q = 1' is selected and at seasonal lag of 12, a significant lag is apparent, so kept seasonal MA term 'Q = 1' initially

The seasonal MA term 'Q' was later optimized to 2, by validating model performance, as the data might be under-differenced

• The final selected terms for SARIMA model is (3, 1, 1)x(1, 1, 2, 12)

• The diagnostic plot for the model is given, which clearly shows a normal distribution of residuals, where more values are around zero

• The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line

• The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index

The model summary indicates that that only MA(1) term used in the model is significant in terms of p-values

• From the multiple iterations of SARIMA models, we also get a comparison of the models in terms of its accuracy attributes of RMSE and MAPE

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

	Test RMSE	Test MAPE
TES Alpha 0.4, Beta 0.1, Gamma 0.2	312.211065	10.20
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106792	9.48
Auto SARIMA(3,1,3)x(3,1,0,12)	331.695786	10.34
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.801449	11.19
TES Alpha 0.15, Beta 0.00, Gamma 0.37	384.197750	11.94
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
SES Alpha 0.00	1275.081813	38.90
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.430000	67.23
DES Alpha 0.6,Beta 0.0	3850.779835	152.05
NaiveModel	3864.279352	152.87

37. Most optimum models by RMSE

	Test RMSE	Test MAPE
Manual SARIMA(3,1,1)x(1,1,2,12)	324.106792	9.48
TES Alpha 0.4, Beta 0.1, Gamma 0.2	312.211065	10.20
Auto SARIMA(3,1,3)x(3,1,0,12)	331.695786	10.34
Auto SARIMA(0,1,1)x(1,0,1,12)-Log10	336.801449	11.19
TES Alpha 0.15, Beta 0.00, Gamma 0.37	384.197750	11.94
2 point TMA	813.400684	19.70
4 point TMA	1156.589694	35.96
SimpleAverage	1275.081804	38.90
SES Alpha 0.00	1275.081813	38.90
6 point TMA	1283.927428	43.86
9 point TMA	1346.278315	46.86
RegressionOnTime	1389.135175	50.15
DES Alpha 0.1,Beta 0.1	1779.430000	67.23
DES Alpha 0.6,Beta 0.0	3850.779835	152.05
NaiveModel	3864.279352	152.87

38. Most optimum model, by MAPE

Model comparison

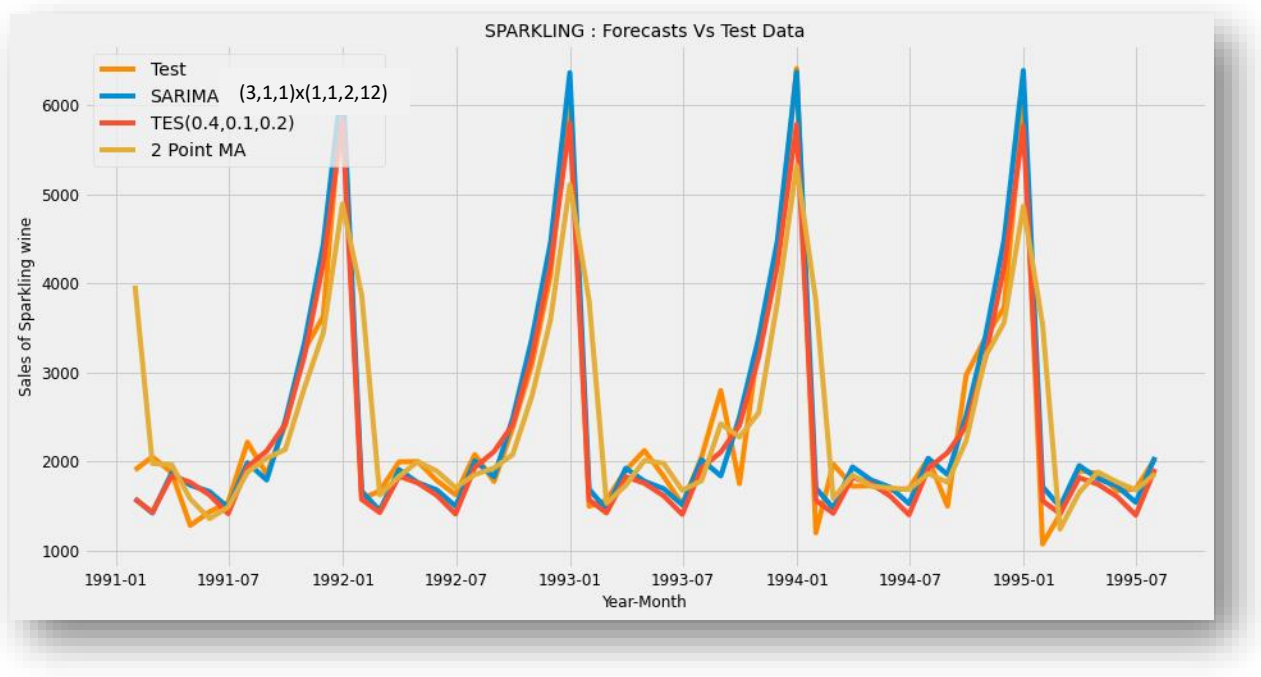
• The overall comparison of all the time-series forecast models is listed in increasing order of RMSE against test data or in the order of decreasing accuracy

• Triple Exponential Smoothing is found to be the best model, followed by SARIMA

• The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted below against the test data

• The SARIMA and Triple Exponential Smoothing are found to be comparable in terms of performance and fitment with the test data

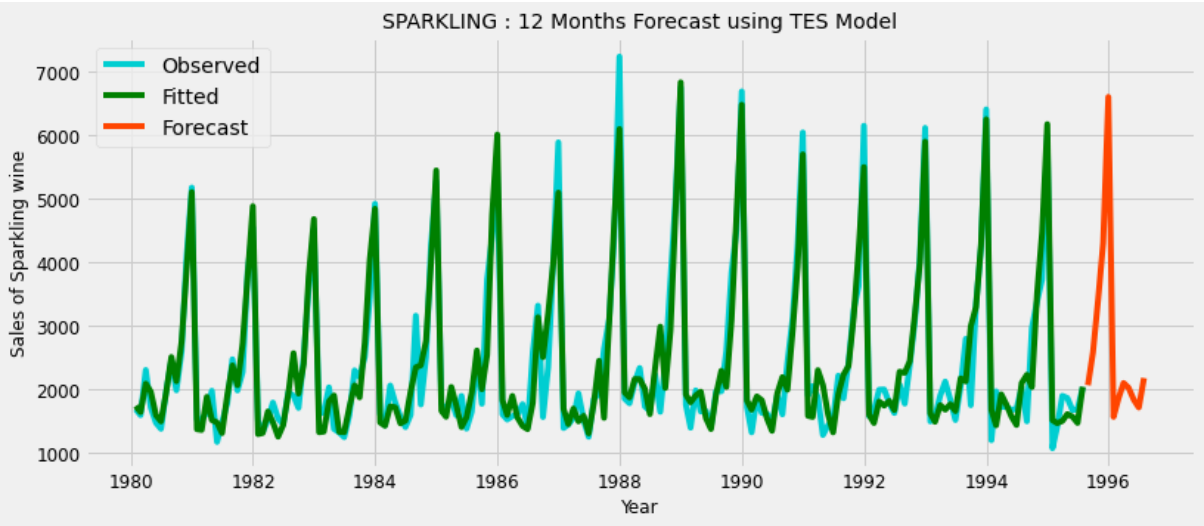
Plot all the forecast



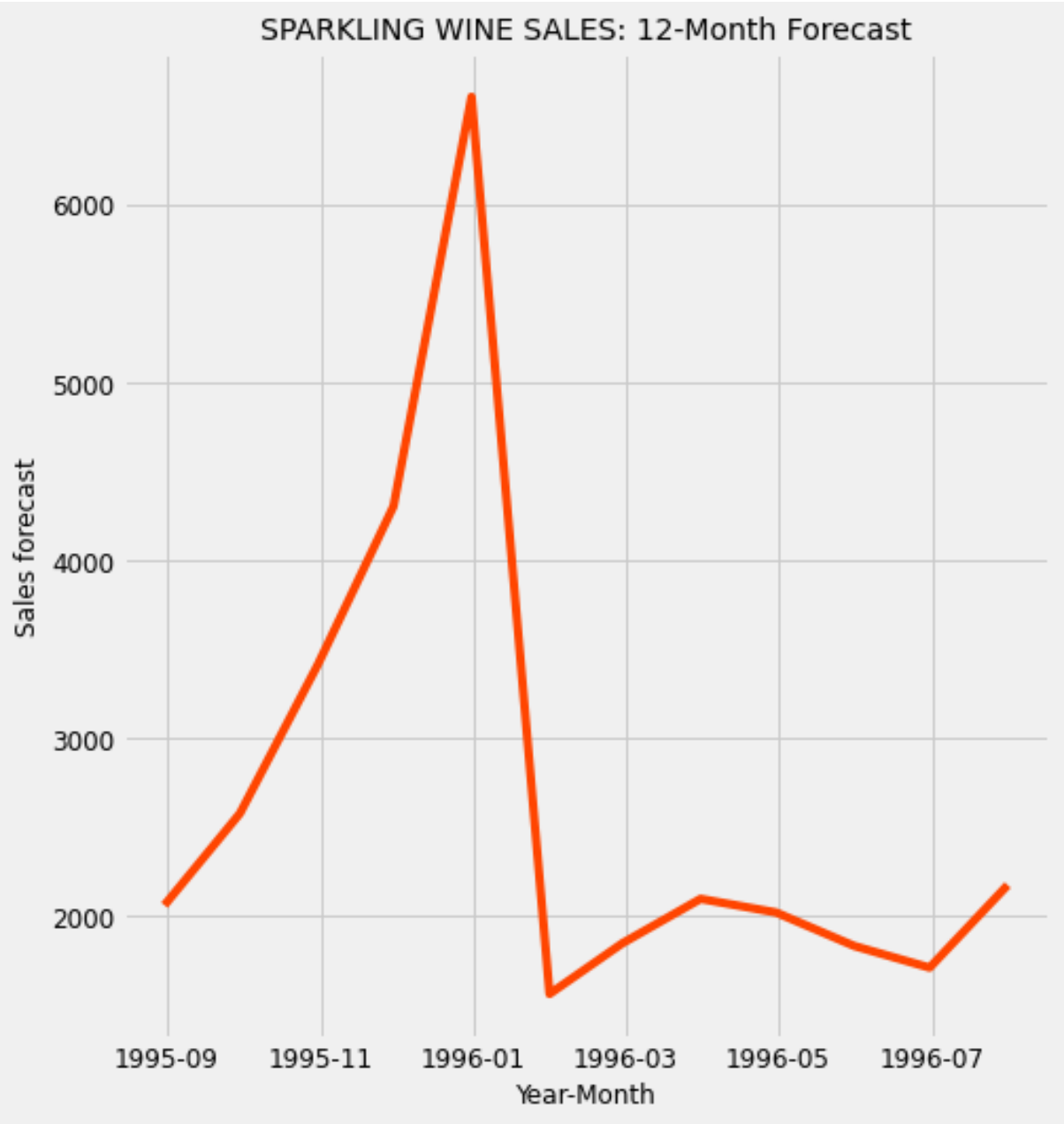
67. Sparkling: Forecast vs Test Data

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

Build model on all of the dataset using best model, which is TES



68. Sparkling - 12 months forecast using TES model



72. Sparkling Wine Sales - 12-month forecast, using TES model

Trying out SARIMA(3,1,3)x(1,1,2,12) for the forecast

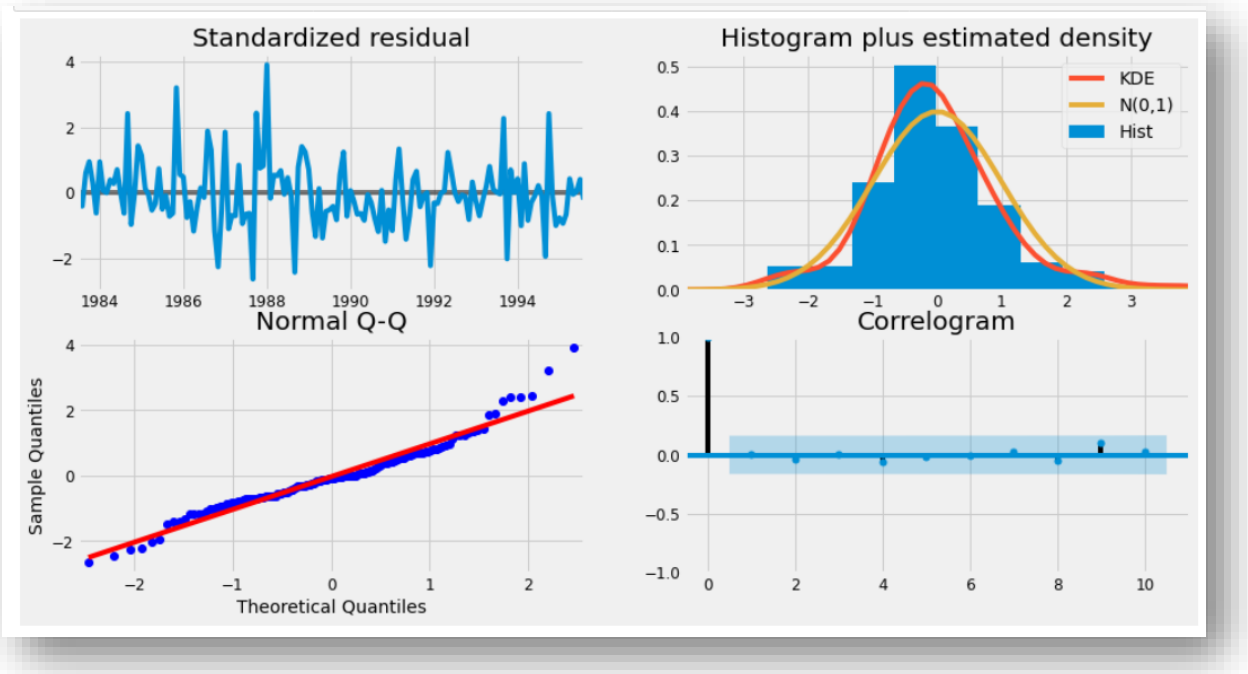
Statespace Model Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(3, 1, 3)x(1, 1, 2, 12)	Log Likelihood	-1078.437
Date:	Fri, 08 Oct 2021	AIC	2176.875
Time:	08:05:18	BIC	2206.711
Sample:	01-31-1980	HQIC	2188.998
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4229	0.086	-4.916	0.000	-0.591	-0.254
ar.L2	-0.9094	0.053	-17.295	0.000	-1.012	-0.806
ar.L3	0.1425	0.087	1.639	0.101	-0.028	0.313
ma.L1	-0.4114	0.078	-5.281	0.000	-0.564	-0.259
ma.L2	0.4622	0.083	5.582	0.000	0.300	0.625
ma.L3	-0.9674	0.104	-9.325	0.000	-1.171	-0.764
ar.S.L12	-0.0701	0.709	-0.099	0.921	-1.460	1.319
ma.S.L12	-0.4549	0.721	-0.631	0.528	-1.868	0.958
ma.S.L24	-0.0811	0.396	-0.205	0.838	-0.858	0.696
sigma2	1.461e+05	1.05e-06	1.39e+11	0.000	1.46e+05	1.46e+05

Ljung-Box (Q):	17.11	Jarque-Bera (JB):	35.58
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.72	Skew:	0.66
Prob(H) (two-sided):	0.26	Kurtosis:	5.03

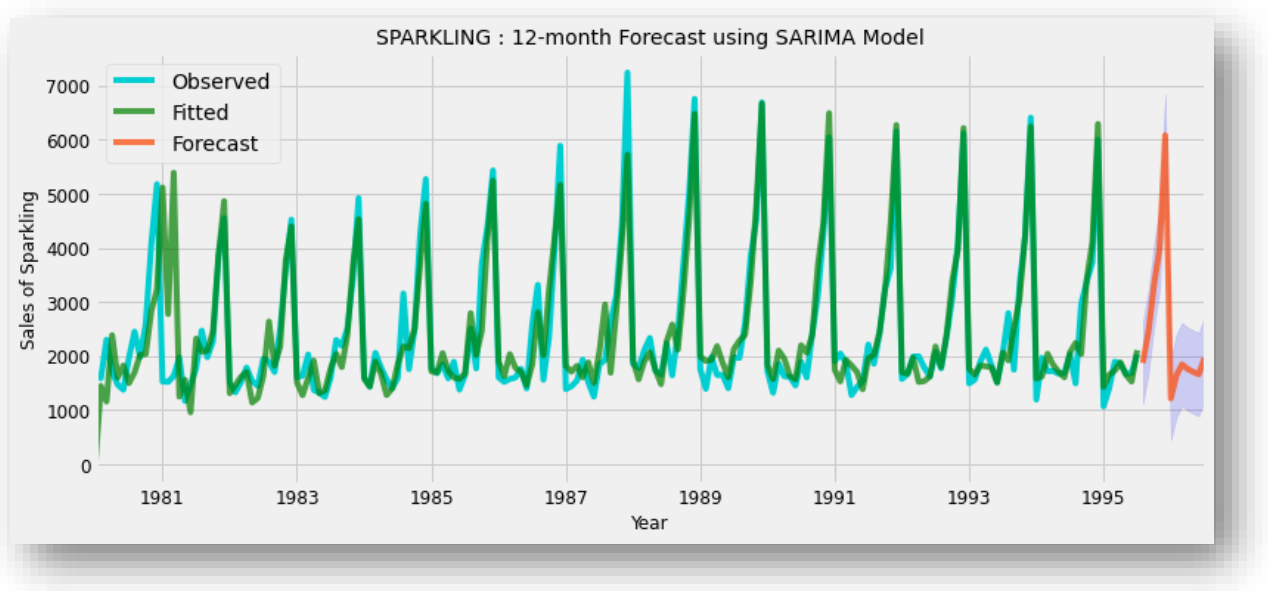
73: Sarima (3,1,3)x(1,1,2,12) model: Results



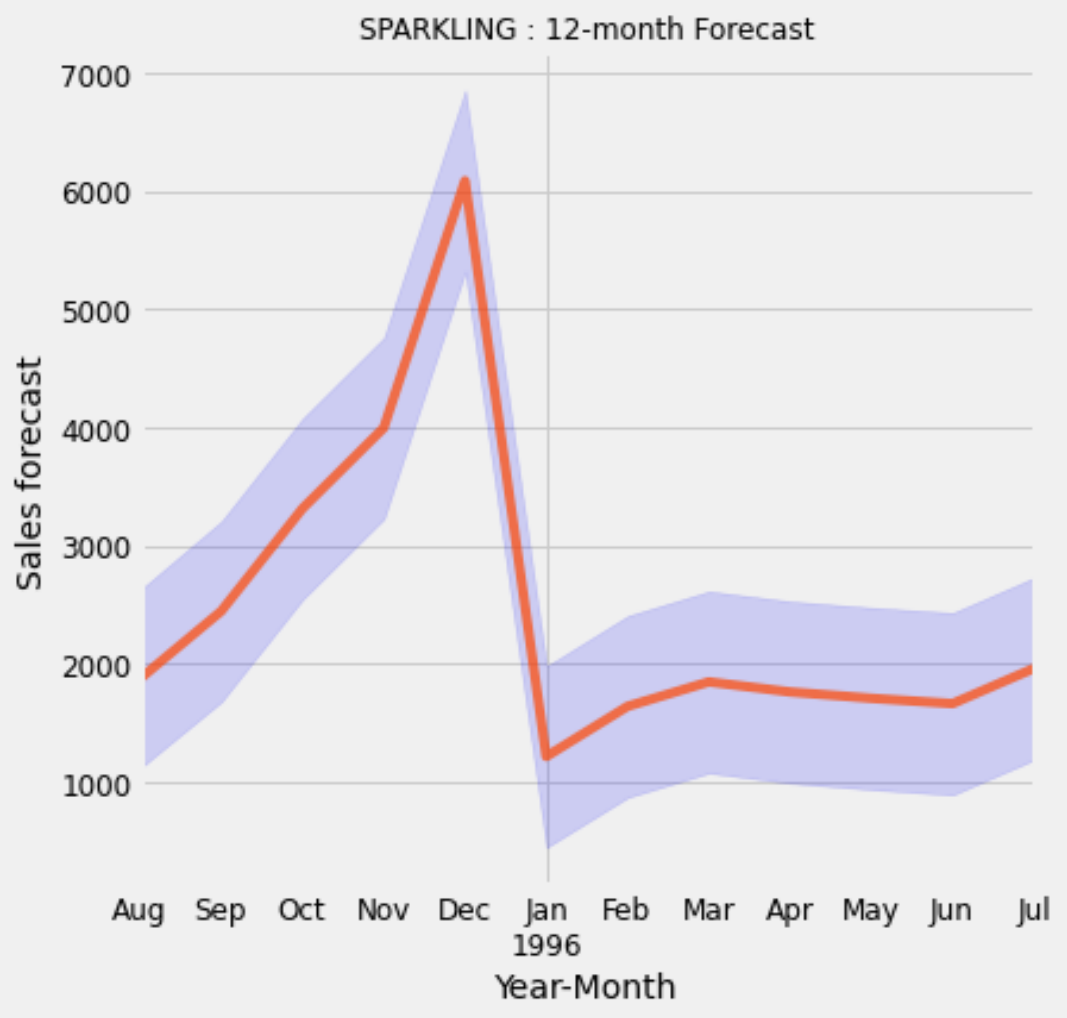
74: Sarima (3,1,3)x(1,1,2,12) model: Diagnostics

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1873.317897	384.477893	1119.755074	2626.880720
1995-09-30	2445.126073	389.491583	1681.736599	3208.515547
1995-10-31	3312.737952	390.186433	2547.986596	4077.489308
1995-11-30	3994.660130	390.695547	3228.910930	4760.409331
1995-12-31	6084.204109	390.825246	5318.200703	6850.207515

39. Sarima (3,1,3)x(1,1,2,12) for forecast, prediction summary



75: Sparkling: 12-month Forecast using manual SARIMA model (3,1,3)x(1,1,2,12)



76: Sparkling: 12-month Forecast with confidence interval,
using manual SARIMA model (3,1,3)x(1,1,2,12)

The forecast

• Based on the overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) and SARIMA are selected for final prediction into 12 months in future

• TES model alpha: 0.4, beta: 0.1 and gamma: 0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data

• The model predicts an upward trend and continuation of the seasonal surge in sales in the upcoming 12 months. According to the model the seasonal sale will be more than that of the previous year

• The 12-month prediction of the TES model is plotted

• The SARIMA model is built with parameters (3, 1, 3)x(1, 1, 2, 12), is found to be the most optimal SARIMA model

•SARIMA model has reflected the trend and seasonality of the series continuing

into the future year as well. The seasonal altitude predicted us more conservative than TES model

• SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE and MAPE values

Model evaluation		
	RMSE	MAPE
TES Forecast	376.821	11.3
SARIMA Forecast	591.245	14.86

	Sparkling
1995-08-31	1873.32
1995-09-30	2445.13
1995-10-31	3312.74
1995-11-30	3994.66
1995-12-31	6084.20
1996-01-31	1216.28
1996-02-29	1640.58
1996-03-31	1847.34
1996-04-30	1762.21
1996-05-31	1708.40
1996-06-30	1663.96
1996-07-31	1961.46

40. Sarima (3,1,3)x(1,1,2,12) 12-month forecast (August 1995 to July 1996) on full data

	Sparkling
count	12.000000
mean	2459.190000
std	1384.631728
min	1216.280000
25%	1697.290000
50%	1860.330000
75%	2662.032500
max	6084.200000

41. Sparkling: Forecast description

```
Sparkling    29510.28
dtype: float64
```

Forecast of total sales

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

Final model

The SARIMA model (3,1,3)x(1,1,2,12) built on the complete Sparkling timeseries is chosen, as prediction provide confidence interval which give better explainability and confidence to the forecasts

• The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot

• The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that AR(2), MA(3) terms has the highest absolute weightage.

The p-values indicates that the terms AR(1), AR(2), MA(1), MA(2) and MA(3) are the most significant terms

• The rest of the p-values got values higher than alpha 0.05, which fails to reject the hull hypothesis that these terms are not significant

Recommendations to the wine company

• The model forecasts sale of 29510 units of Sparkling wine in the coming 12 months, at an average of 2,459 units a month

• The seasonal sales in December 1995 will hit a maximum of 6,084 units, before it drops to the lowest in January 1996; at 1216 units.

• The wine company is recommended to ramp up their procurement and production line in accordance with the above forecasts for the third quarter of 1995 (October, November and December), which is a total of 13,392 units of sparkling wine expected to be sold.

• The forecast also indicates that the year-on-year sale of Sparkling wine is not showing an upward trend. The winery must come up with a new marketing plan to improve the sales

• Adding more exogenous variables into the timeseries data can improve forecast

Rose

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

Read the data from the '.csv' file as a monthly Time Series.

	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

1. First few rows of the dataset

	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

2. Last few rows of the dataset

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

3. Creating timestamp with month as frequency, since we deal with sales figures here

	YearMonth	Rose	Time_Stamp
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

4. Adding timestamp to data frame

Time_Stamp	YearMonth	Rose
1980-01-31	1980-01	112.0
1980-02-29	1980-02	118.0
1980-03-31	1980-03	129.0
1980-04-30	1980-04	99.0
1980-05-31	1980-05	116.0
...	...	...
1995-03-31	1995-03	45.0
1995-04-30	1995-04	52.0
1995-05-31	1995-05	28.0
1995-06-30	1995-06	40.0
1995-07-31	1995-07	62.0

Monthly sales of Rose wine are given for a period from January, 1980 to July, 1995.

• The given dataset loaded correctly and a date-range was applied on it as index

5. Setting timestamp as index

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
...	...
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

187 rows × 1 columns

6. Data frame after dropping YearMonth column

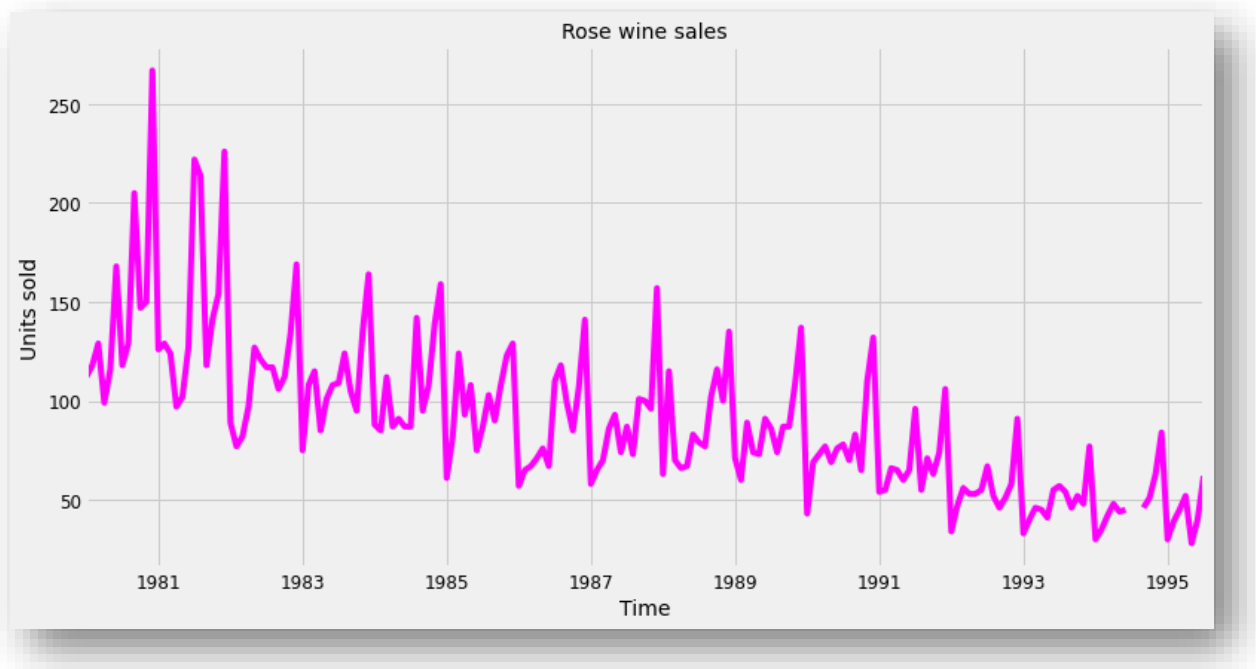
	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

7. Basic measures of descriptive statistics

```
Rose     2
dtype: int64
```

8. Check for missing values

- The Rose time-series has values missing for two months. Let's check which values



1. Rose wine sales (original data)

Time series data must be contiguous. Bure there's a gap in the 1994 figures. Let us check which

Time_Stamp	Rose
1994-01-31	30.0
1994-02-28	35.0
1994-03-31	42.0
1994-04-30	48.0
1994-05-31	44.0
1994-06-30	45.0
1994-07-31	NaN
1994-08-31	NaN
1994-09-30	46.0
1994-10-31	51.0
1994-11-30	63.0
1994-12-31	84.0

9. 1994 sales figures, to check gap

Since the data has monthly frequency, we can resample at a shorter frequency such as day or daily to get a better prediction. Some of the alias for time series frequency to be used are:-

B: Business-day frequency

D: Calendar-day frequency

M: Month-end frequency

MS: Month-start frequency

Q: Quarter-end frequency

QS: Quarter-start frequency

H: Hourly frequency

A: Year-end frequency

```
Time_Stamp
1994-01-31    30.000000
1994-02-28    35.000000
1994-03-31    42.000000
1994-04-30    48.000000
1994-05-31    44.000000
1994-06-30    45.000000
1994-07-31    45.336957
1994-08-31    45.673913
1994-09-30    46.000000
1994-10-31    51.000000
1994-11-30    63.000000
1994-12-31    84.000000
Name: Rose, dtype: float64
```

10. 1994 figures after data interpolation in July and August, (7th and 8th month)

The missing values were imputed using linear interpolation because the gap to bridge is small, meaning there's very small difference between the previous and next figures, of just 1 (45 and 46), so any imputed value must lie between these numbers). Besides the graph is also jagged and not smooth for us to consider any other method such as spline. Linear imputation ignores the index and treats the values as equally spaced.

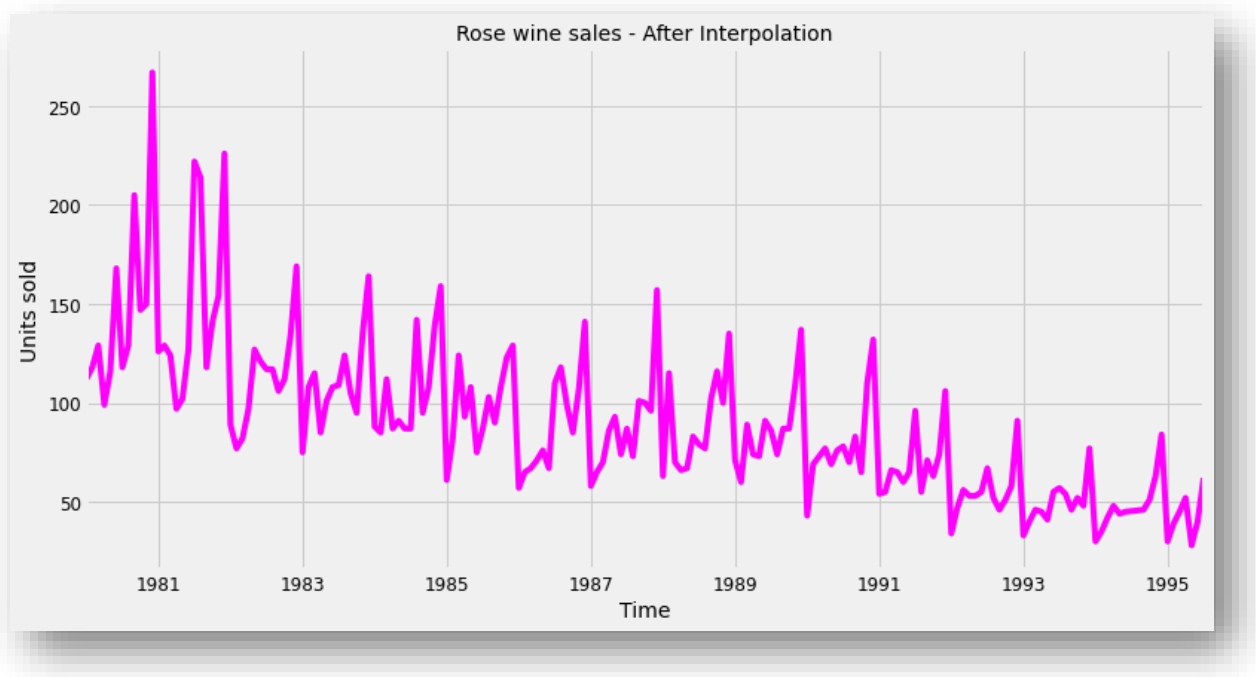
• Rose data after interpolation for year 1994 is also plotted.

	Rose
count	187.000000
mean	89.914497
std	39.238259
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

11. Descriptive statistics of converted data frame.

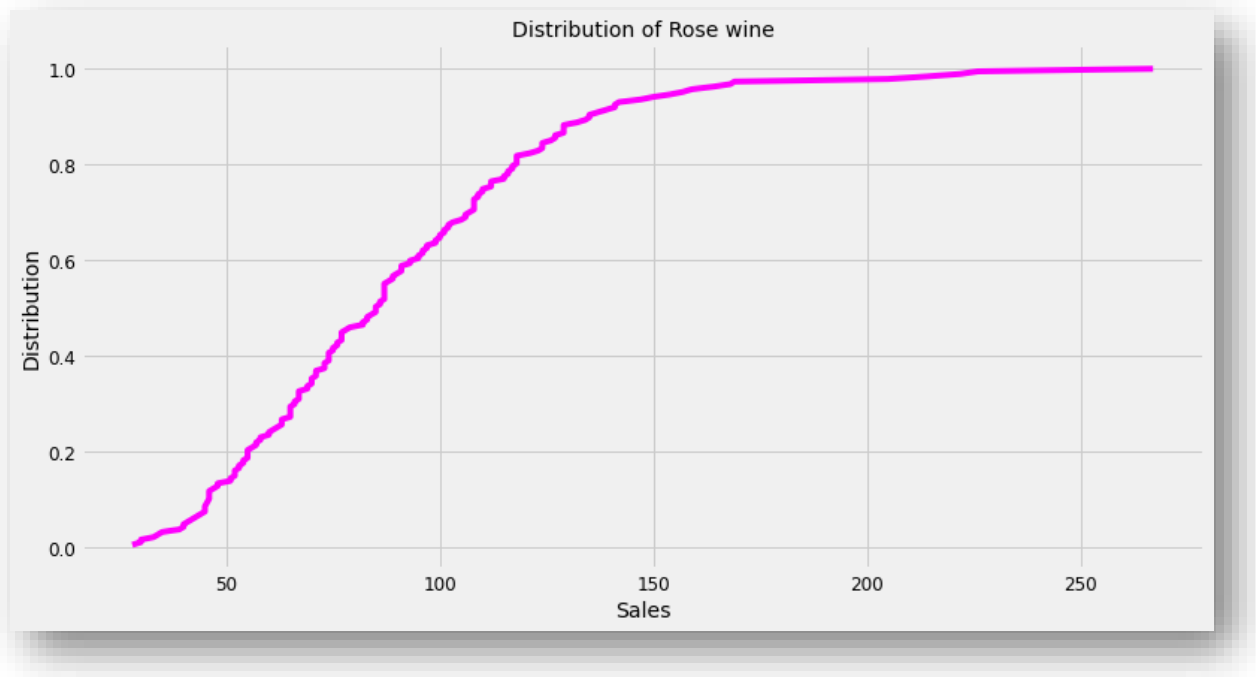
The descriptive summary of the data shows that, on an average, 90 units of Rose wine were sold each month in the given period of time, and 50% of the months, the sales varied from 63 to 112 units. The maximum-ever sale reported in a month is 267 units and the minimum is 28 units

Plot the Time Series to understand the behaviour of the data.



2. Rose wine sales - After interpolation

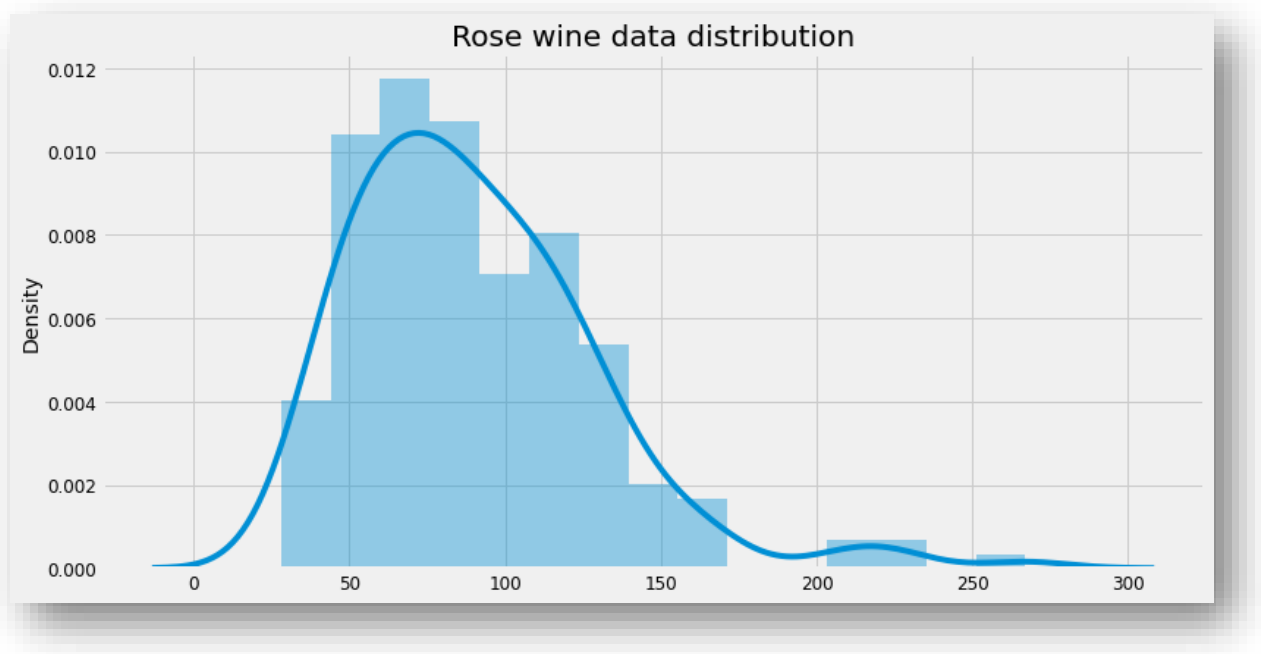
- The dataset shows significant seasonality. The sales of Rose show an evident downward trend
- The demand for Rose wine has been falling out of favour with the customers over the years



3. Distribution of rose wine (estimator of the Cumulative Distribution Function or eCDF plot)

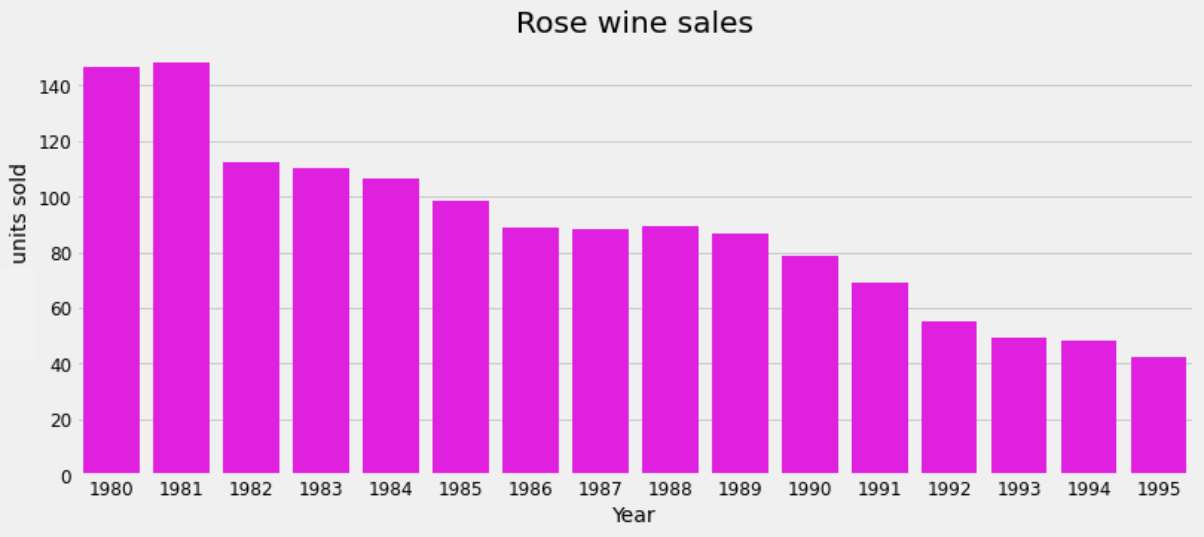
The Empirical CDF plot shows that, in 80% of months, at least 120 units of Rose wine were sold

2. Perform appropriate EDA to understand the data and also perform decomposition.



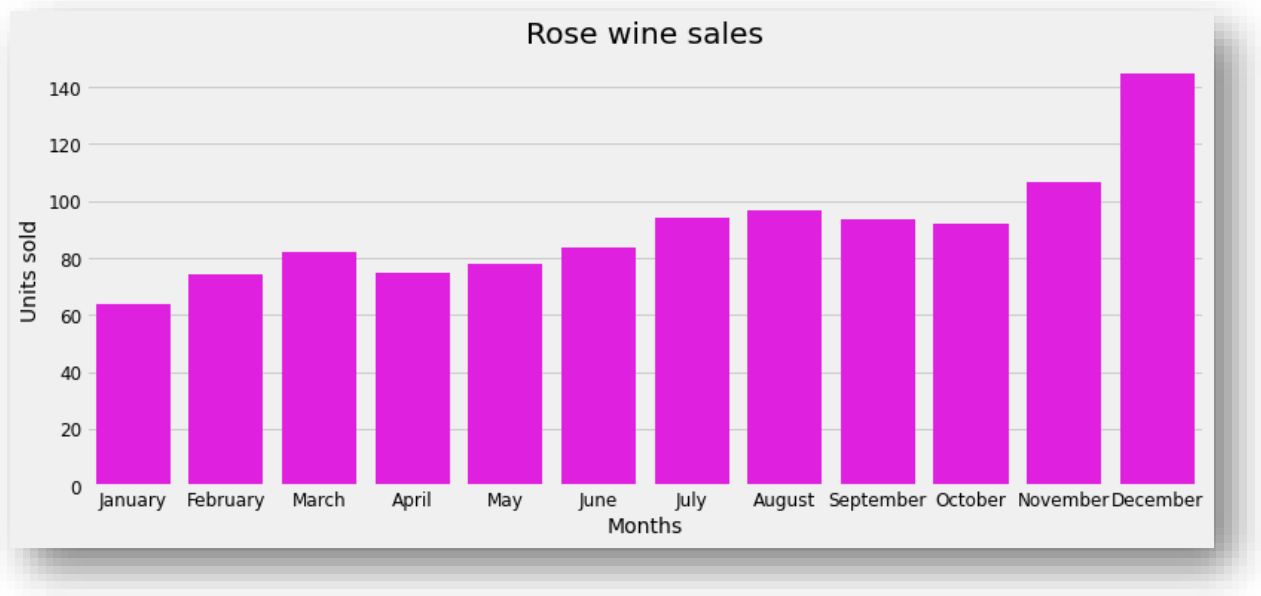
4. Rose wine data distribution (distplot)

Data found to be right skewed. The highest frequencies are for 50 to 90 units of sales. Right skewed distributions occur when the long tail is on the right side of the distribution. Analysts also refer to them as positively skewed. This condition occurs because probabilities taper off more slowly for higher values. Consequently, you'll find extreme values far from the peak on the high end more frequently than on the low.



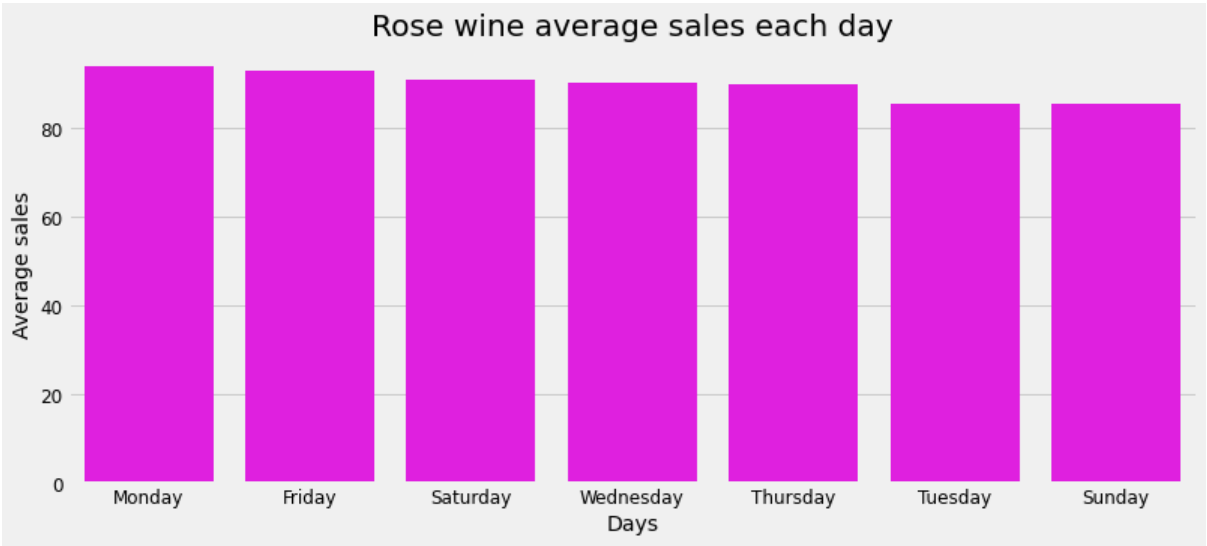
5. Rose wine sales by year (barplot)

Best sales were in 1981 with 1980 the next best. Towards the end, there's a period of decline



6. Rose wine sales by month (barplot)

December is the best sales month, the happiest season being November and December



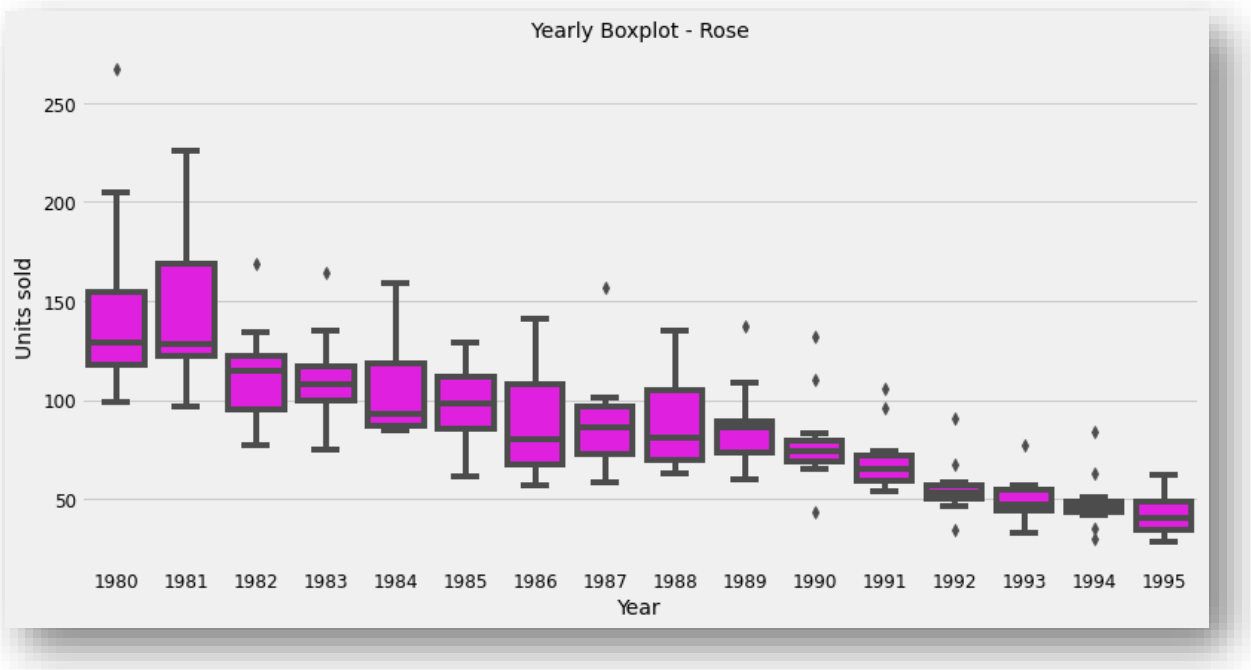
7. Rose wine average sales each day (barplot)

Mondays and Fridays are best part, although the entire week is not much different

8. Rose wine overall sales each day (barplot)

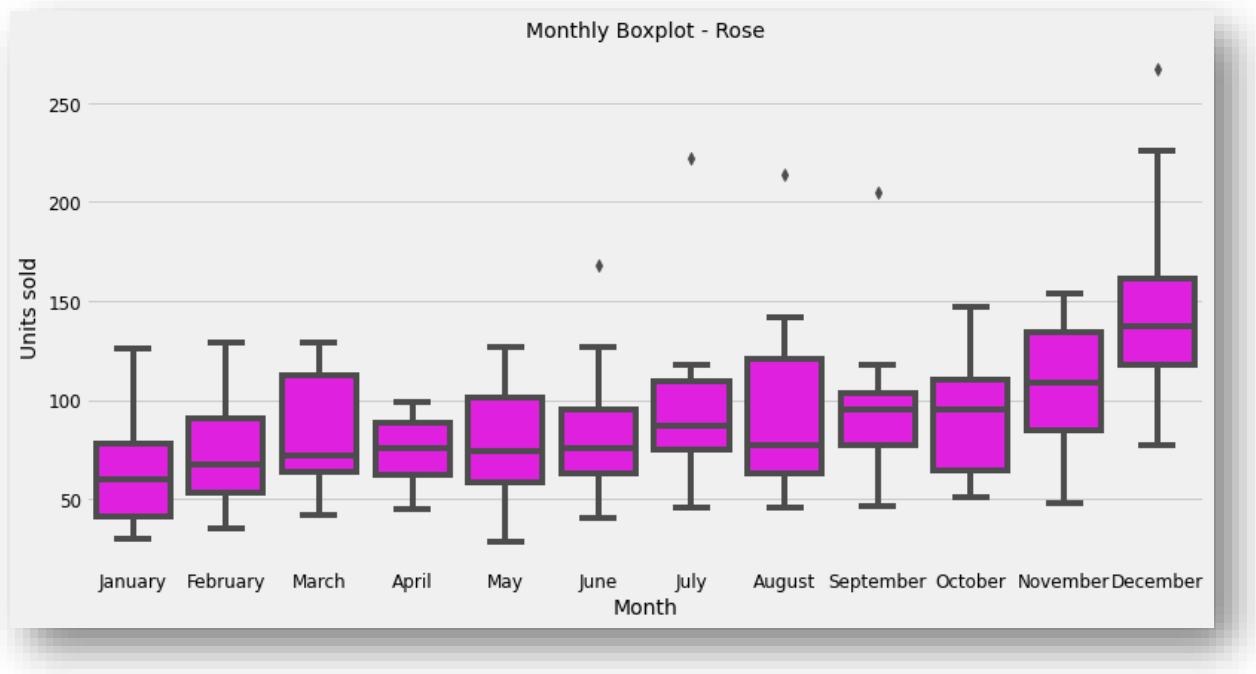
Mondays and Saturdays sold the most Rose wines overall

Make a boxplot to understand the spread of wine sales across different years and within different months across years.



9. Yearly boxplot

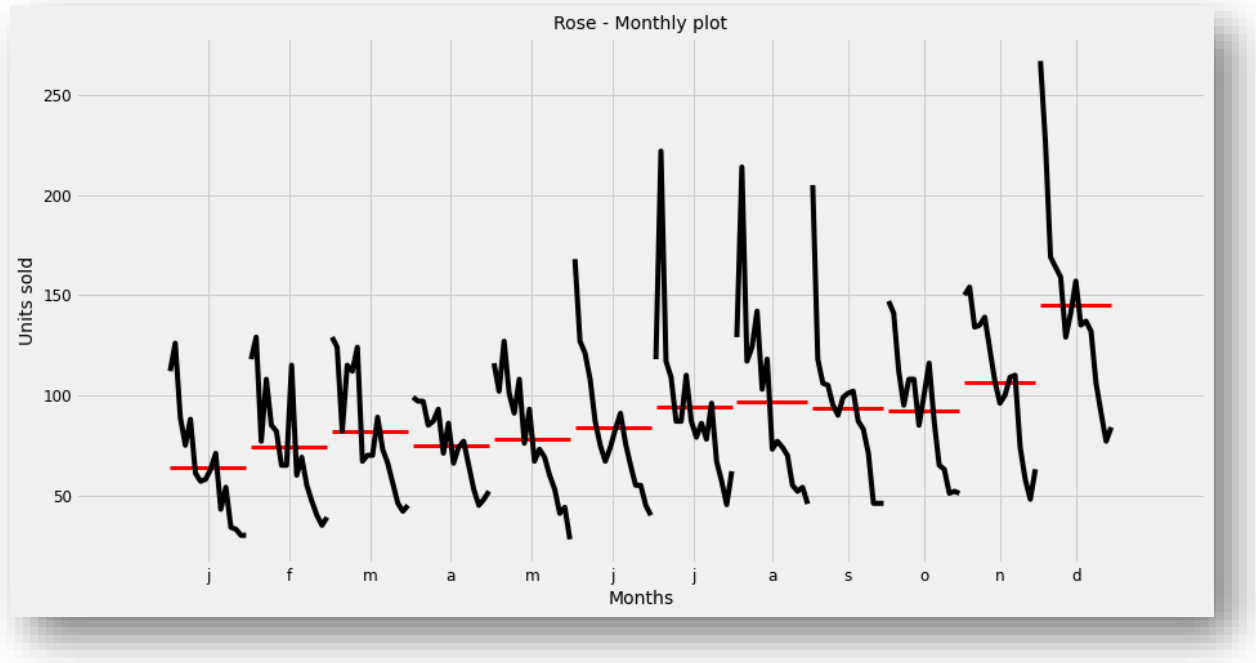
The yearly-boxplot, shows that the average sale of Rose wine moving according to the downward trend in sales over the years. The outliers over upper-bound in the yearly boxplot most probably represent the seasonal sale during the seasonal months



10. Monthly boxplot

The monthly-box-plot shows a clear seasonality during the months of November and December. The sales tank in January but picks up in due course of the year. Average sales are around 140 units in December, around 110 units in November, and 90 in October

Plot a time series monthplot to understand the spread of sales across different years and within different months across years.



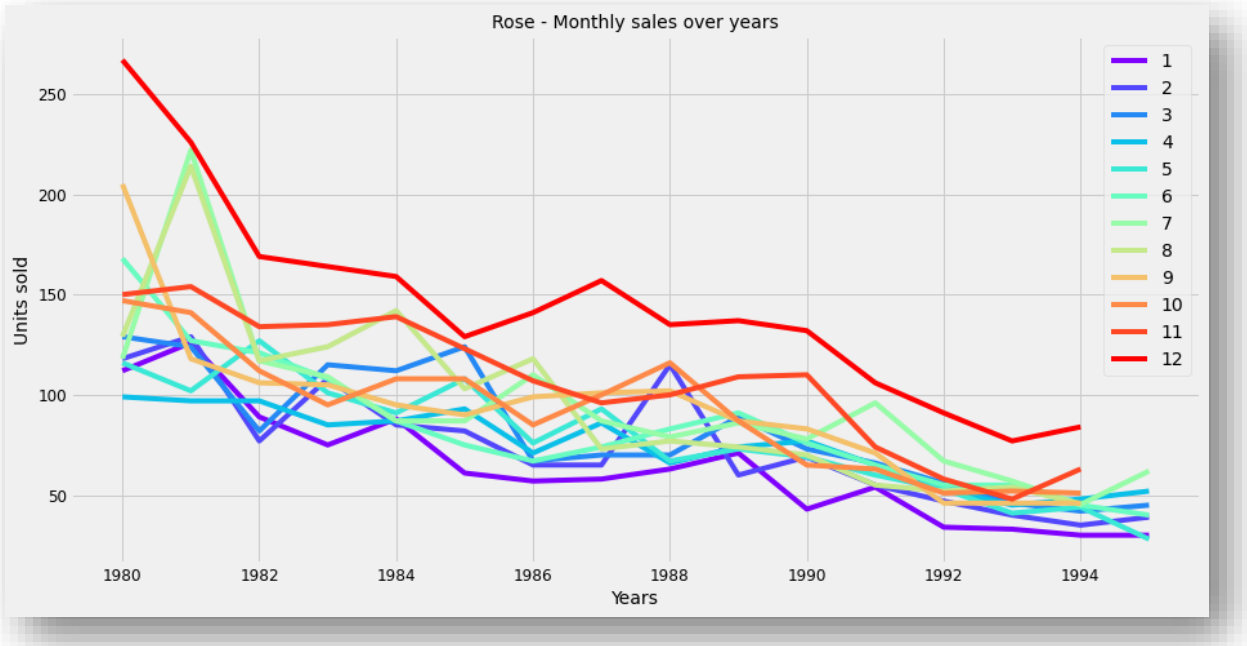
11. Times series month plot

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

• Sales in December with a mean a few points below 100, vary from 75 to 270 units over the years, while the average sales are below or closer to 100 units (above 50) for the rest of the year

Plot a graph of monthly sales across years. Pivot table first.

Months / Years	1	2	3	4	5	6	7	8	9	10	11	12
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.336957	45.673913	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN

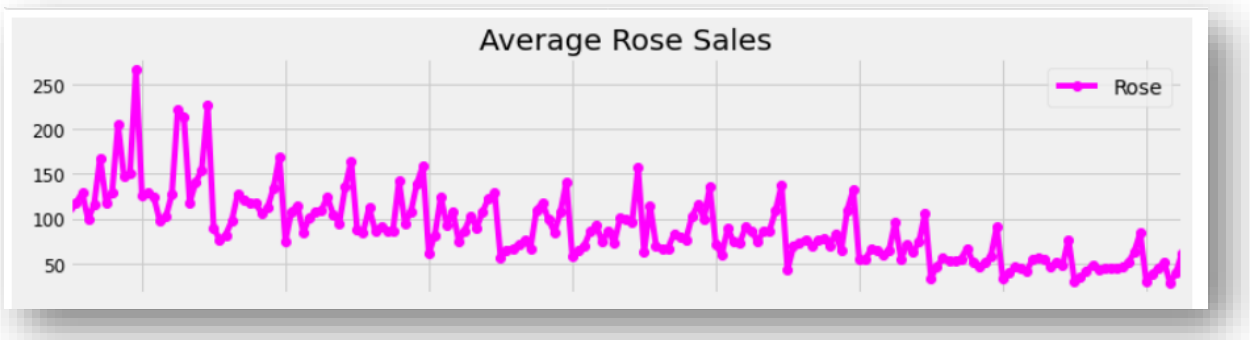


12. Monthly sales over years

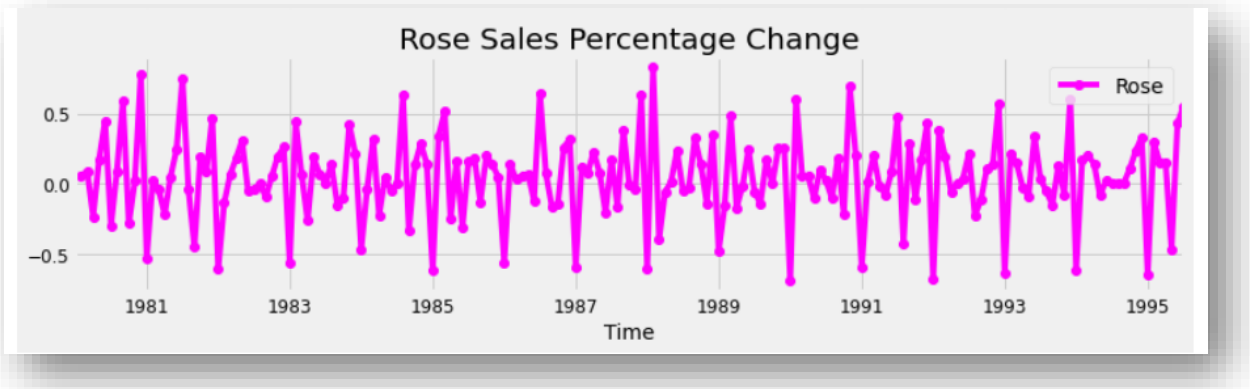
• The plot of monthly sale over the years also shows the seasonality component of the time-series, with November and December selling exponentially higher volumes than the other months.

• The highest volume of Rose wine was sold in December 1980 and the least of December sales was in 1993. Even though the December sales picked after 1983, these dipped consistently after 1987

Plot the average sales per month and the month on month percentage change of sales.



13. Average rose sales



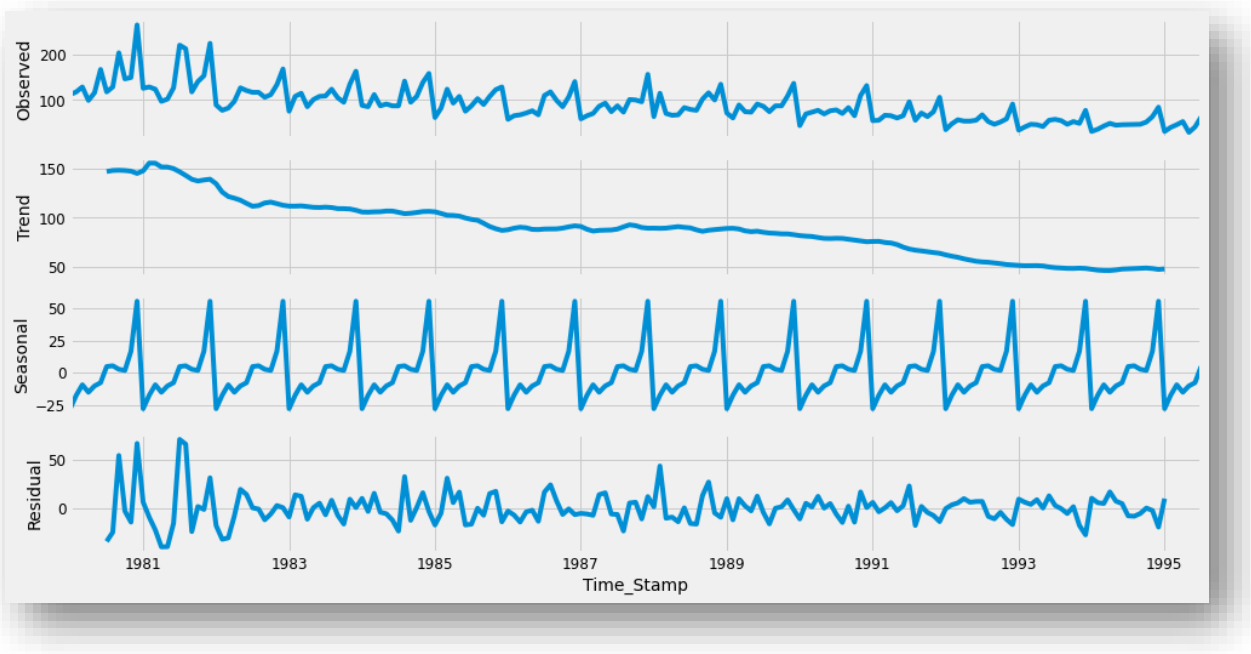
14. Rose sales: Percentage change

Average sales are on the decline. The percentage changes every annual season

Decompose the Time Series and plot the different components.

If the seasonality and residual components are independent of the trend, then you have an additive series. If the seasonality and residual components are in fact dependent, meaning they fluctuate on trend, then you have a multiplicative series.

Additive Decomposition



15. Rose- Additive decomposition

Time_Stamp	Trend
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	147.083333
1980-08-31	148.125000
1980-09-30	148.375000
1980-10-31	148.083333
1980-11-30	147.416667
1980-12-31	145.125000

Name: Rose, dtype: float64

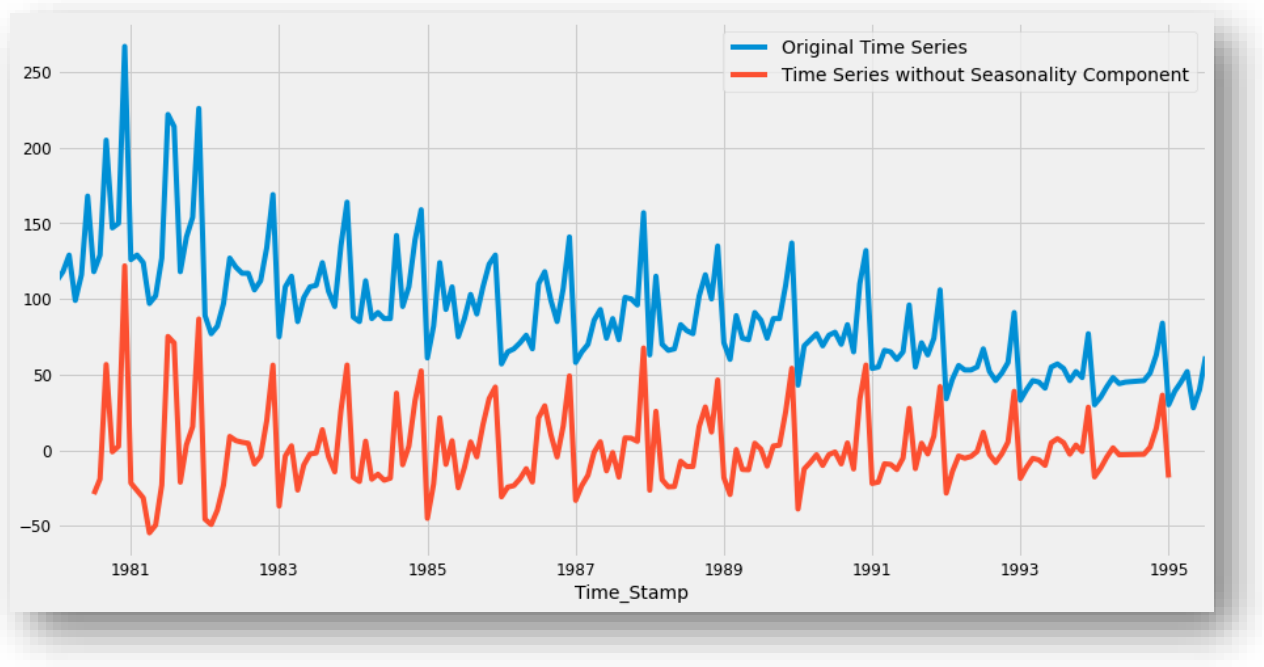
Time_Stamp	Seasonality
1980-01-31	-27.908708
1980-02-29	-17.435675
1980-03-31	-9.285895
1980-04-30	-15.098395
1980-05-31	-10.196609
1980-06-30	-7.678752
1980-07-31	4.897089
1980-08-31	5.500109
1980-09-30	2.774625
1980-10-31	1.871848
1980-11-30	16.846848
1980-12-31	55.713514

Name: Rose, dtype: float64

Time_Stamp	Residual
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	-33.980423
1980-08-31	-24.625109
1980-09-30	53.850375
1980-10-31	-2.955181
1980-11-30	-14.263514
1980-12-31	66.161486

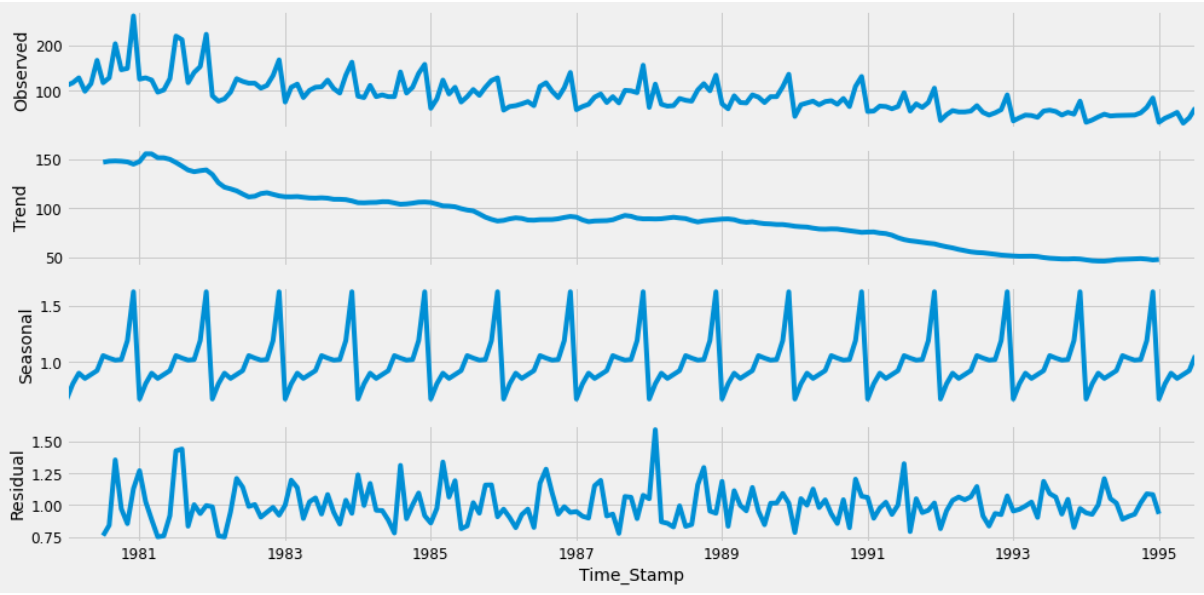
Name: Rose, dtype: float64

13. Trend, seasonality, residuals for additive decomposition



16. Detrending, of additive: Time series, original and without seasonality

Multiplicative Decomposition



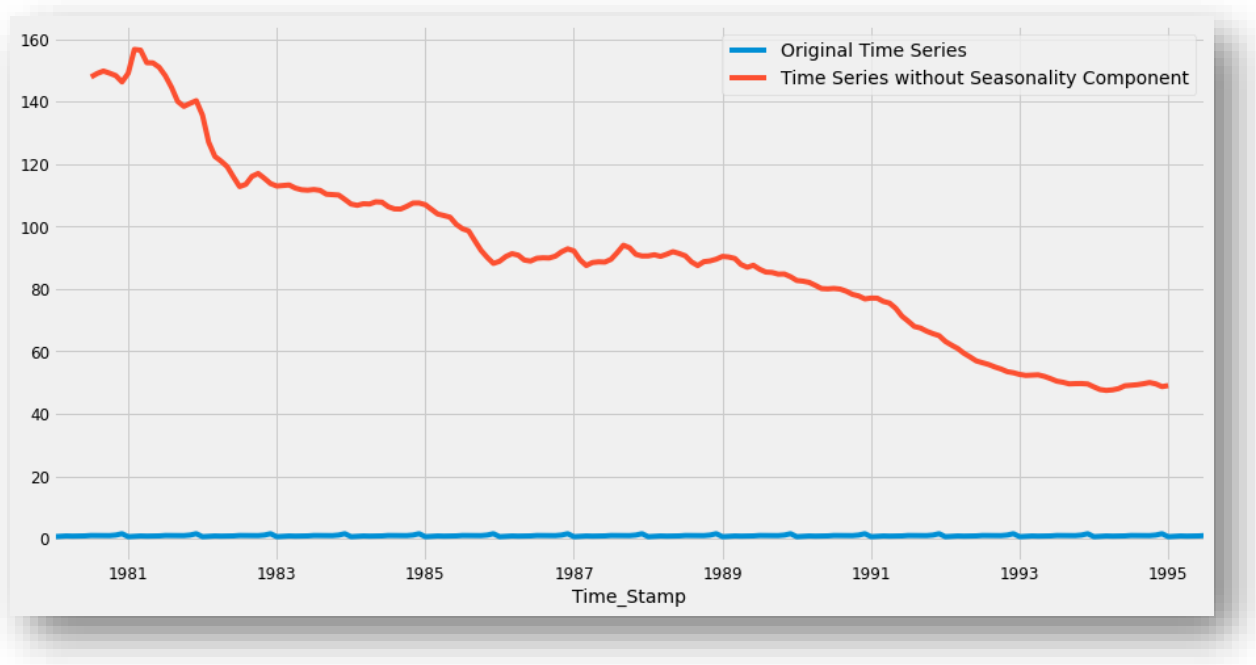
17. Multiplicative decomposition for Rose

```
Trend
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
Name: Rose, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31    0.670111
1980-02-29    0.806163
1980-03-31    0.901163
1980-04-30    0.854023
1980-05-31    0.889414
1980-06-30    0.923984
1980-07-31    1.058042
1980-08-31    1.035890
1980-09-30    1.017647
1980-10-31    1.022572
1980-11-30    1.192347
1980-12-31    1.628644
Name: Rose, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    0.758256
1980-08-31    0.840713
1980-09-30    1.357675
1980-10-31    0.970772
1980-11-30    0.853379
1980-12-31    1.129647
Name: Rose, dtype: float64
```

14. Trend, seasonality, residuals for multiplicative decomposition



18. Detrending, of multiplicative: Time series, original and without seasonality

Decomposition analysis

• The observed plot of the decomposition diagrams shows visible annual seasonality and a downward trend. The early period of the plot shows higher variation than in the later periods

• The trend diagram shows a downward trend overall. Exponential dips can be seen between 1981 and 1983 and later from 1991 to 1993

• Seasonal components are quite visible and consistent in both the observed and seasonal charts of the diagrams. The additive chart shows variance in seasonality from -20 to 50 units and the multiplicative model shows variance of 16%

• The residuals show a pattern of high variability across the period of time-series, which is more or less consistent in both additive and multiplicative decompositions

• The variance in residuals shows higher variance in the early period of the series, which explains the higher variance in observed plot at same time period

• The additive model shows a mean variance around 0 and the multiplicative model shows a variance around 15%

• As the seasonality peaks are consistently reducing its altitude in consistent with trend, the series can be treated as multiplicative in model-building

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

First few rows of Training Data

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

Last few rows of Training Data

Time_Stamp	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

First few rows of Test Data

Time_Stamp	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

Last few rows of Test Data

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

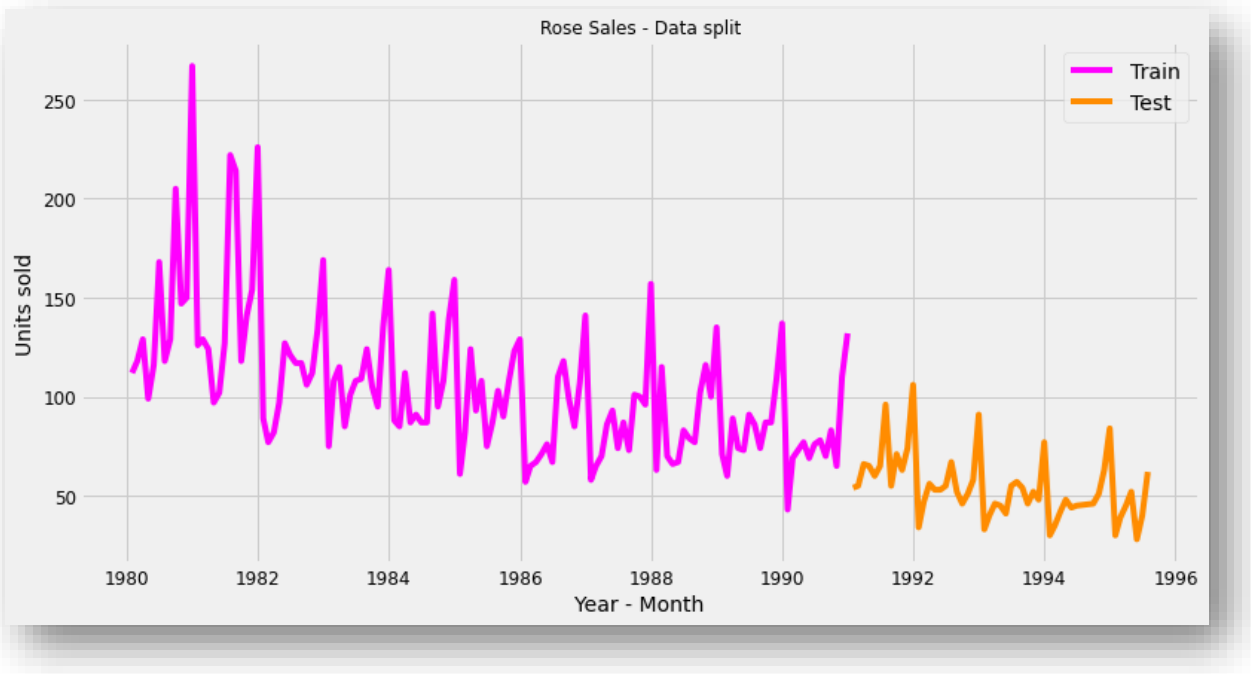
15. Test and training data, first and last few rows

• The train and test datasets are created with 1991 as starting year for the test data, using index.year property of time series index

• The plots for train-and-test split for Sparkling and Rose time-series are given

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

16. Shapes of train and test sets



19. Rose Sales - Data Split

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

Model 1: Linear Regression

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

17. Training time and test time instances for linear regression model

First few rows of Training Data

Time_Stamp	Rose	time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

Last few rows of Training Data

Time_Stamp	Rose	time
1990-08-31	70.0	128
1990-09-30	83.0	129
1990-10-31	65.0	130
1990-11-30	110.0	131
1990-12-31	132.0	132

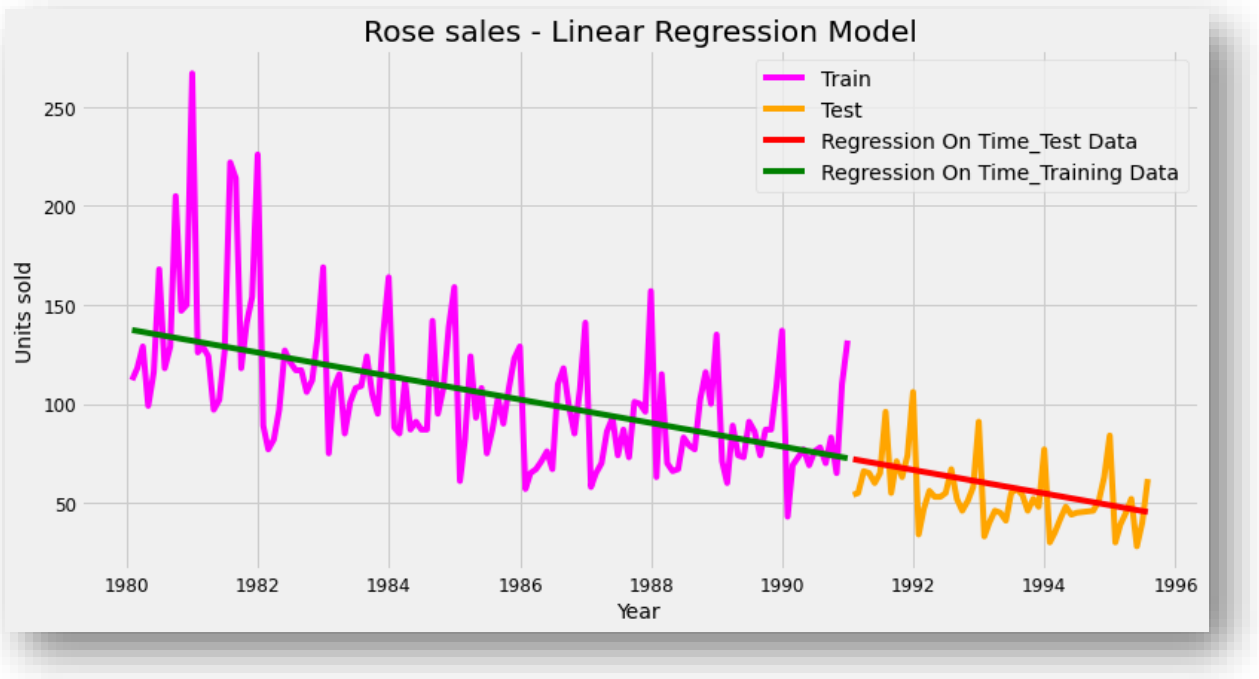
First few rows of Test Data

Time_Stamp	Rose	time
1991-01-31	54.0	133
1991-02-28	55.0	134
1991-03-31	66.0	135
1991-04-30	65.0	136
1991-05-31	60.0	137

Last few rows of Test Data

Time_Stamp	Rose	time
1995-03-31	45.0	183
1995-04-30	52.0	184
1995-05-31	28.0	185
1995-06-30	40.0	186
1995-07-31	62.0	187

18. Test and training data, first and last few rows, for linear regression model



20. Rose sales - Linear Regression Model

The linear regression on the Rose dataset shows an apparent downward trend as consistent with the observed time-series

• The RMSE and MAPE of the forecast is given above. The model leaves a 23% error in forecast against test set

• The model has successfully captured the trend of both the series, but does not reflects the seasonality

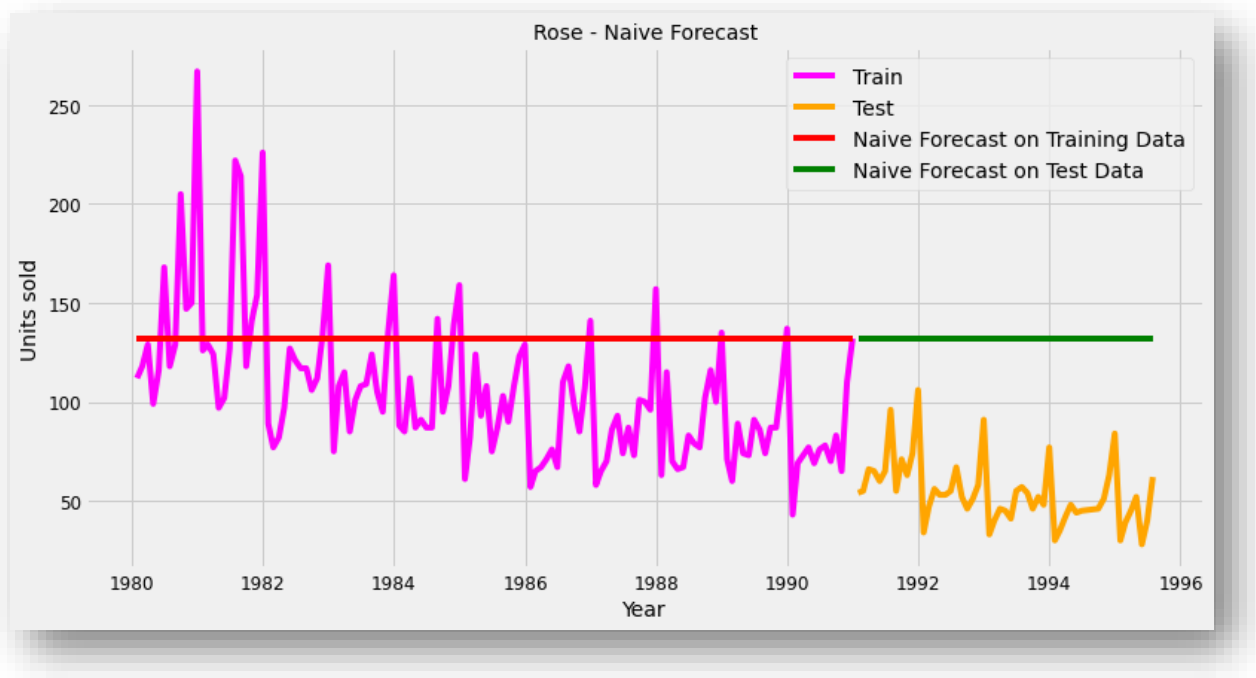
Model evaluation		
	RMSE	MAPE
Train	30.718	21.22
Test	15.269	22.82

Model 2: Naive forecast

```
Time_Stamp
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
Name: rose_naive, dtype: float64
```

```
Time_Stamp
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
Name: rose_naive, dtype: float64
```

19. Naive model train and test data with timestamp



21. Rose - Naive Forecast

• As Rose dataset has a downward trend, the percentage of error is less in train and very high in test

• The model captures neither the trend nor the seasonality of the given data

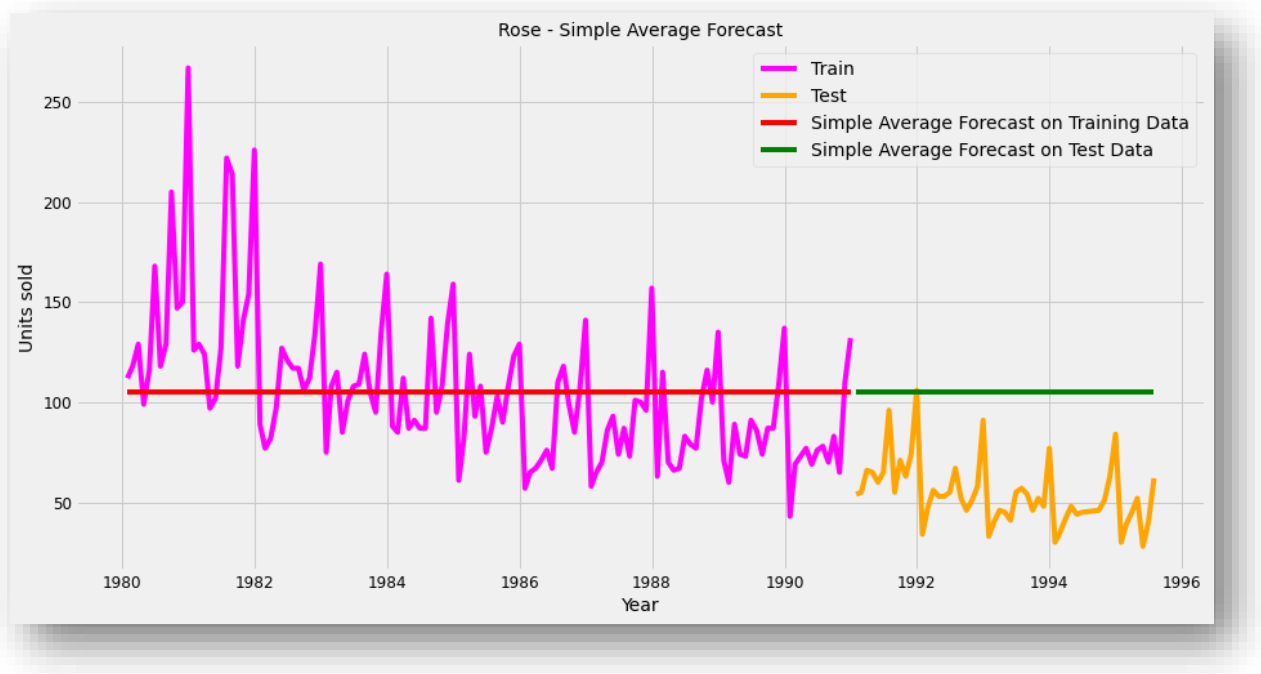
Model evaluation		
	RMSE	MAPE
Train	45.064	36.38
Test	79.719	145.1

Model 3: Simple Average

```
Time_Stamp
1980-01-31    104.939394
1980-02-29    104.939394
1980-03-31    104.939394
1980-04-30    104.939394
1980-05-31    104.939394
Name: rose_mean_forecast, dtype: float64
```

```
Time_Stamp
1991-01-31    104.939394
1991-02-28    104.939394
1991-03-31    104.939394
1991-04-30    104.939394
1991-05-31    104.939394
Name: rose_mean_forecast, dtype: float64
```

20. Simple-average model train and test data with timestamp



22. Rose - Simple Average Forecast

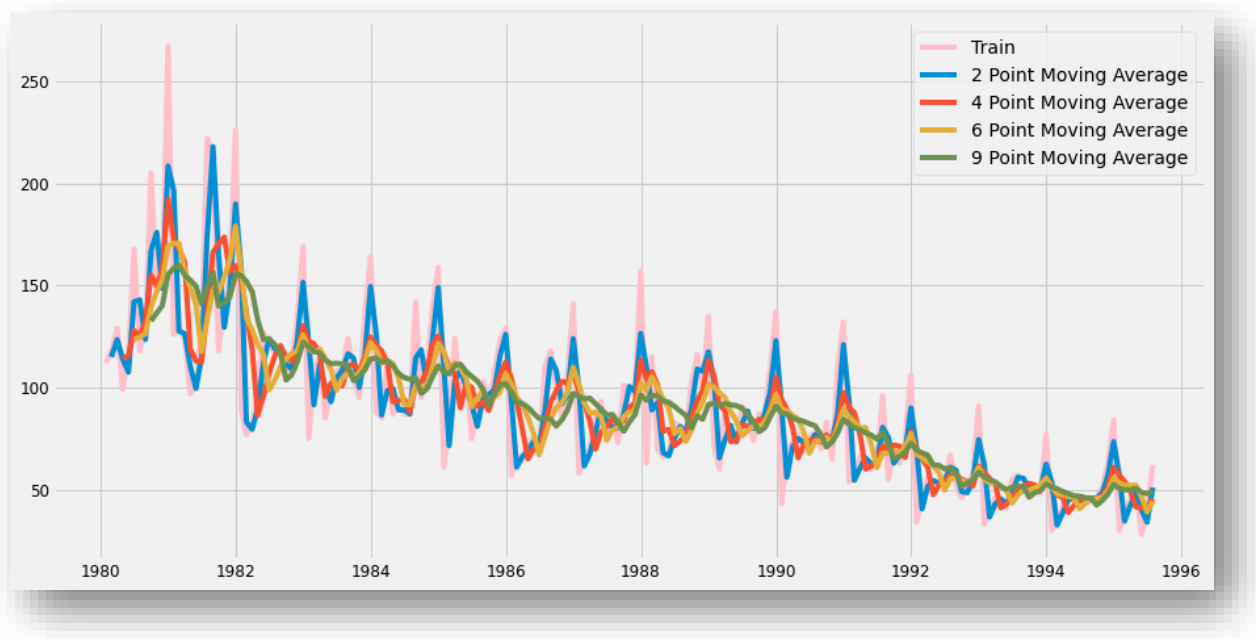
- For Rose dataset, the model forecast is almost 100% error in test data and 25% in train
- Due to the downward trend the performance in train data set is better than the test dataset

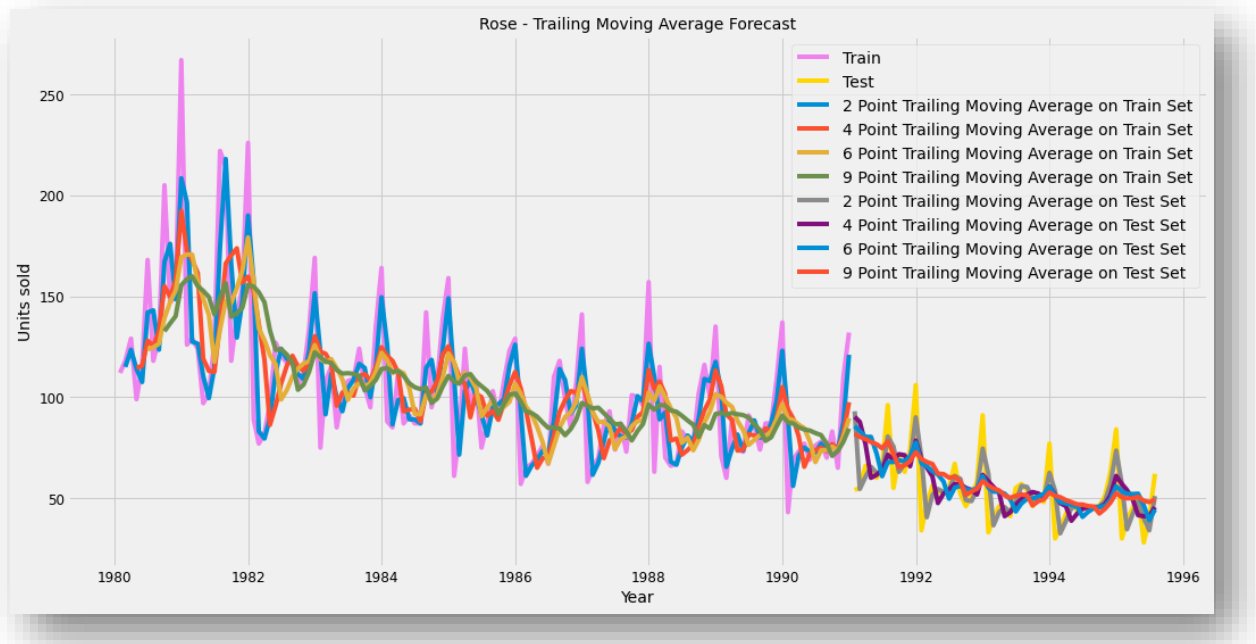
Model evaluation		
	RMSE	MAPE
Train	36.034	25.39
Test	53.46	94.93

Model 4: Moving Average

Time_Stamp	Rose	Rose_Trailing_2	Rose_Trailing_4	Rose_Trailing_6	Rose_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

21. Moving average data with timestamp





23. Rose - Trailing Moving Average & TMA Forecast

For the moving average model, we are going to calculate rolling means (or trailing moving averages) for different intervals.

• The best interval can be determined by the maximum accuracy (or the minimum error)

• The moving average models are built for trailing 2 points, 4 points, 6 points and 9 points

• For Rose dataset the accuracy is found to be higher with the lower rolling point averages

• In moving average forecasts the values can be fitted with a delay of n number of points

• The Root Mean Squared Error and Mean Absolute Percentage Error of the test set are given below
• The best interval of moving average from the model is 2 point

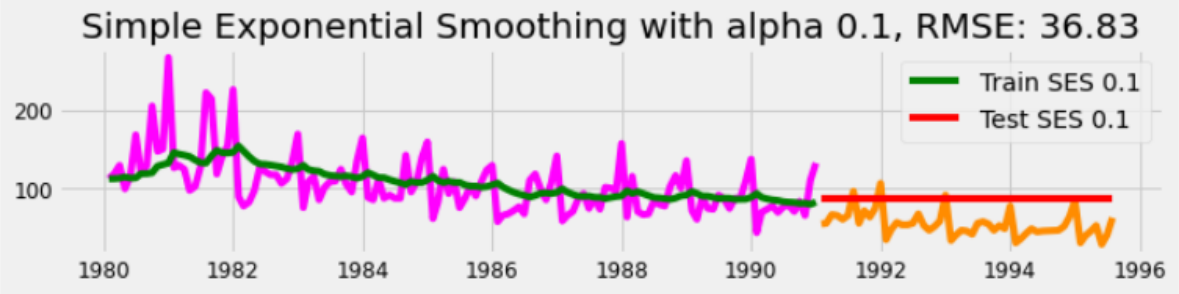
Model evaluation		
Model	RMSE	MAPE
2-point MA	11.529	13.54
4-point MA	14.451	19.49
6-point MA	14.566	20.82
9-point MA	14.728	21.01

Model 5: Simple Exponential Smoothing

```
Time_Stamp
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
Name: Rose, dtype: float64
```

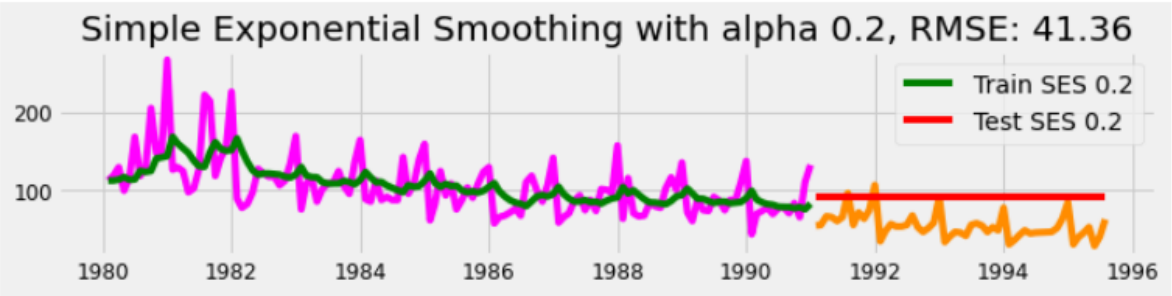
22. Simple exponential smoothing train data

```
Test: For alpha = 0.10,  RMSE is 36.8278 MAPE is 63.94
For smoothing level = 0.10,  Initial level 112.00
```



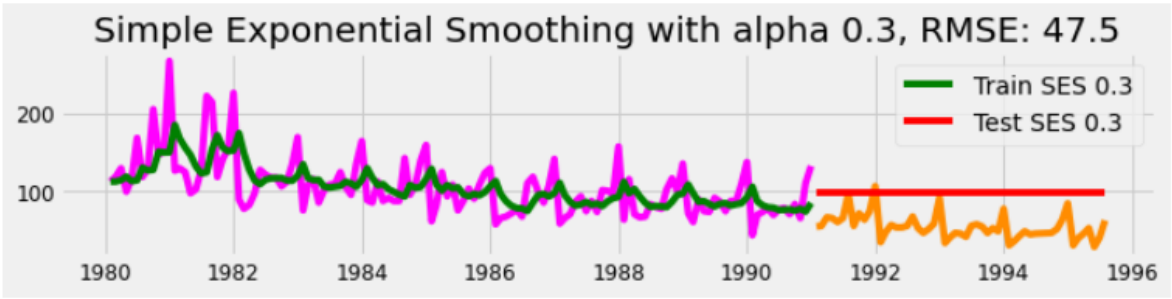
24. Simple exponential smoothing with alpha 0.1, RMSE 36.83
(training, test, and predicted time series plot)

```
Test: For alpha = 0.20,  RMSE is 41.3617 MAPE is 72.21
For smoothing level = 0.20,  Initial level 112.00
```



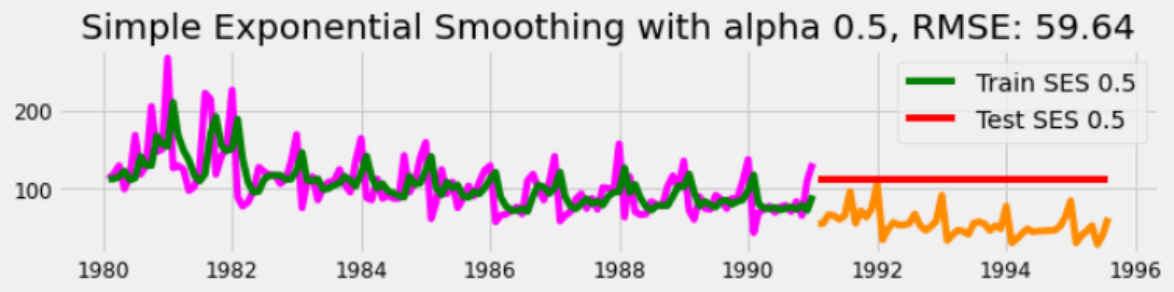
25. Simple exponential smoothing with alpha 0.2, RMSE 41.36
(training, test, and predicted time series plot)

```
Test: For alpha = 0.30,  RMSE is 47.5046 MAPE is 83.71
For smoothing level = 0.30,  Initial level 112.00
```



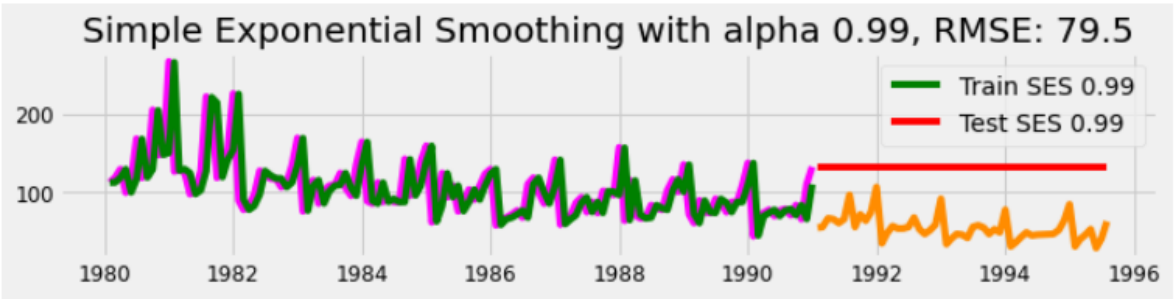
26. Simple exponential smoothing with alpha 0.3, RMSE 47.5
(training, test, and predicted time series plot)

```
Test: For alpha = 0.50,  RMSE is 59.6416 MAPE is 106.81
For smoothing level = 0.50,  Initial level 112.00
```



27. Simple exponential smoothing with alpha 0.5, RMSE 59.64
(training, test, and predicted time series plot)

```
Test: For alpha = 0.99,  RMSE is 79.4985 MAPE is 144.69
For smoothing level = 0.99,  Initial level 112.00
```

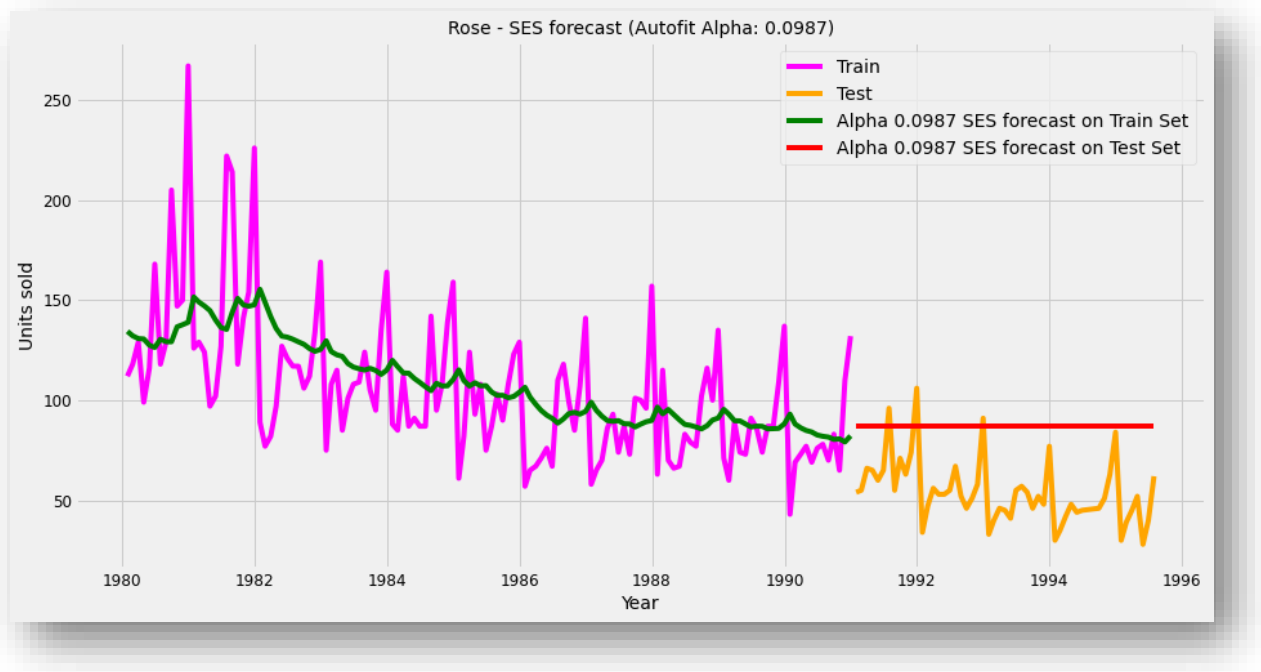


28. Simple exponential smoothing with alpha 0.99, RMSE 79.5
(training, test, and predicted time series plot)

Time_Stamp	Rose	predict_rose
1980-01-31	112.0	134.387120
1980-02-29	118.0	132.176391
1980-03-31	129.0	130.776472
1980-04-30	99.0	130.601045
1980-05-31	116.0	127.480441

Time_Stamp	Rose	predict_rose
1991-01-31	54.0	87.105003
1991-02-28	55.0	87.105003
1991-03-31	66.0	87.105003
1991-04-30	65.0	87.105003
1991-05-31	60.0	87.105003

22. Simple exponential smoothing forecast for fitted values and test



29. Rose - SES forecast (Autofit Alpha: 0.0987) on train and test

Simple Exponential Smoothing is usually applied if the time-series has neither a trend nor seasonality, which is not the case with the given data

• The forecasting using smoothing levels or alpha between 0 and 1 are as below, where the values were passed manually

• For alpha value closer to 1, forecast follows the actual observation closely and closer to 0, forecasts are farther from actual and line gets smoothened

• The test RMSE is found to be higher for values closer to zero

On the second iteration, the model was ran without passing a value for alpha and used parameters 'optimized=True, use_brute=True'

• The autofit model picked 0.098 as the smoothing parameter and retuned consistent RMSE values in train and test datasets, which is consistent with alpha 0.1 in first iteration

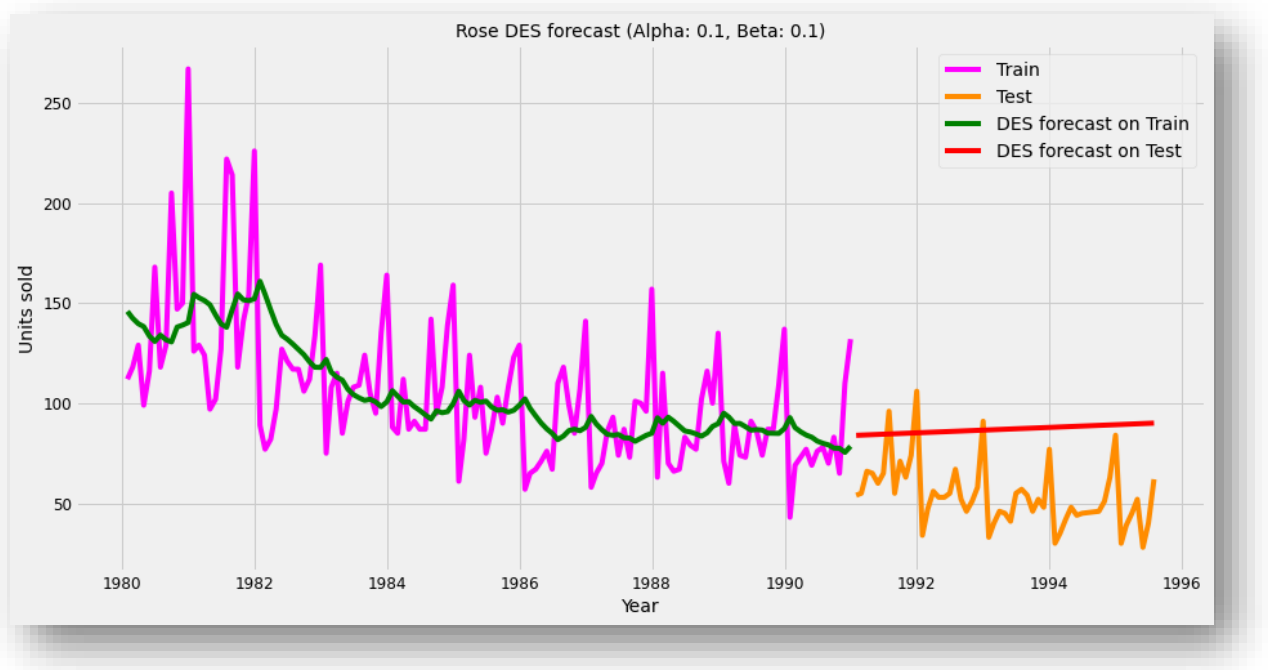
Model evaluation		
	RMSE	MAPE
Train	31.501	22.73
Test	36.796	63.88

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

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	32.026565	22.78	37.056911	64.02
1	0.1	0.2	33.450729	24.45	48.688399	83.09
10	0.2	0.1	32.796403	23.06	65.731352	113.20
2	0.1	0.3	33.145789	24.46	78.156381	131.24
20	0.3	0.1	33.528397	23.47	98.653063	170.12

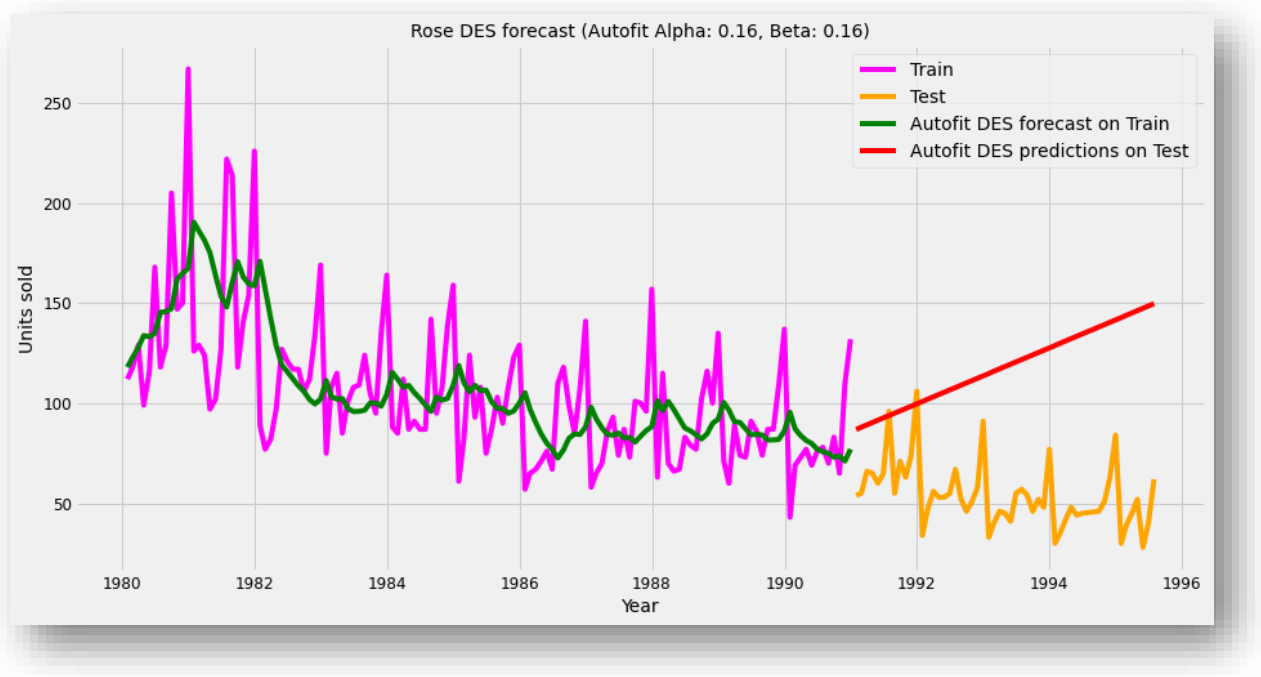
	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.1	0.1	32.026565	22.78	37.056911	64.02
1	0.1	0.2	33.450729	24.45	48.688399	83.09
10	0.2	0.1	32.796403	23.06	65.731352	113.20
2	0.1	0.3	33.145789	24.46	78.156381	131.24
3	0.1	0.4	33.262191	24.68	99.583210	165.53

23. Double exponential smoothing sorted test RMSE and MAPE heads to get best scores (Alpha 0.1, Beta 0.1)



30. Rose DES forecast (Alpha 0.1, Beta 0.1)

Attempting autofit



31. Rose DES forecast (Autofit Alpha 0.16, Beta 0.16)

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.100000	32.026565	22.78	37.056911	64.02
1	0.100000	0.200000	33.450729	24.45	48.688399	83.09
10	0.200000	0.100000	32.796403	23.06	65.731352	113.20
100	0.157895	0.157895	33.074575	23.99	70.572197	120.25
2	0.100000	0.300000	33.145789	24.46	78.156381	131.24

	Alpha	Beta	Train RMSE	Train MAPE	Test RMSE	Test MAPE
0	0.100000	0.100000	32.026565	22.78	37.056911	64.02
1	0.100000	0.200000	33.450729	24.45	48.688399	83.09
10	0.200000	0.100000	32.796403	23.06	65.731352	113.20
100	0.157895	0.157895	33.074575	23.99	70.572197	120.25
2	0.100000	0.300000	33.145789	24.46	78.156381	131.24

24. Double exponential smoothing sorted test RMSE and MAPE heads to get best scores (autofit Alpha 0.16, Beta 0.16)

The Double Exponential Smoothing models is applicable when data has trend, but no seasonality. Rose data contain significant trend component and seasonality

• In first iteration, smoothing level (alpha) and trend (beta) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1 and beta 0.1

• On the second iteration the model was allowed to chose the optimized values using parameters 'optimized=True, use_brute=True'

The autofit model retuned higher accuracy in train dataset, on par with the best models from iteration 1, but faired behind in the test accuracy scores

• The model evaluation parameters of the best models are given as above

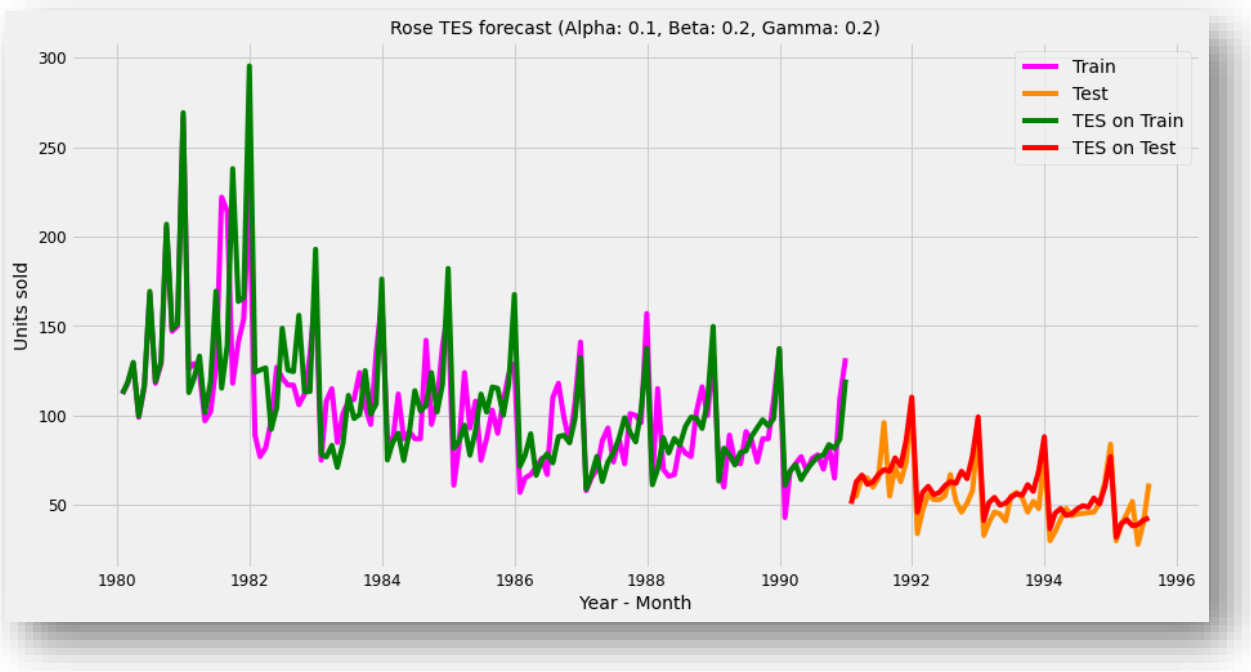
• The best model chosen as final one is the one with alpha 0.1 and beta 0.1

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

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
11	0.1	0.2	0.2	24.365597	15.36	9.640616	13.96
12	0.1	0.2	0.3	23.969166	15.13	9.935672	14.21
10	0.1	0.2	0.1	25.529854	16.06	9.943512	14.39
142	0.2	0.5	0.3	27.631767	17.87	10.026322	14.34
151	0.2	0.6	0.2	28.289836	18.09	10.031754	13.62

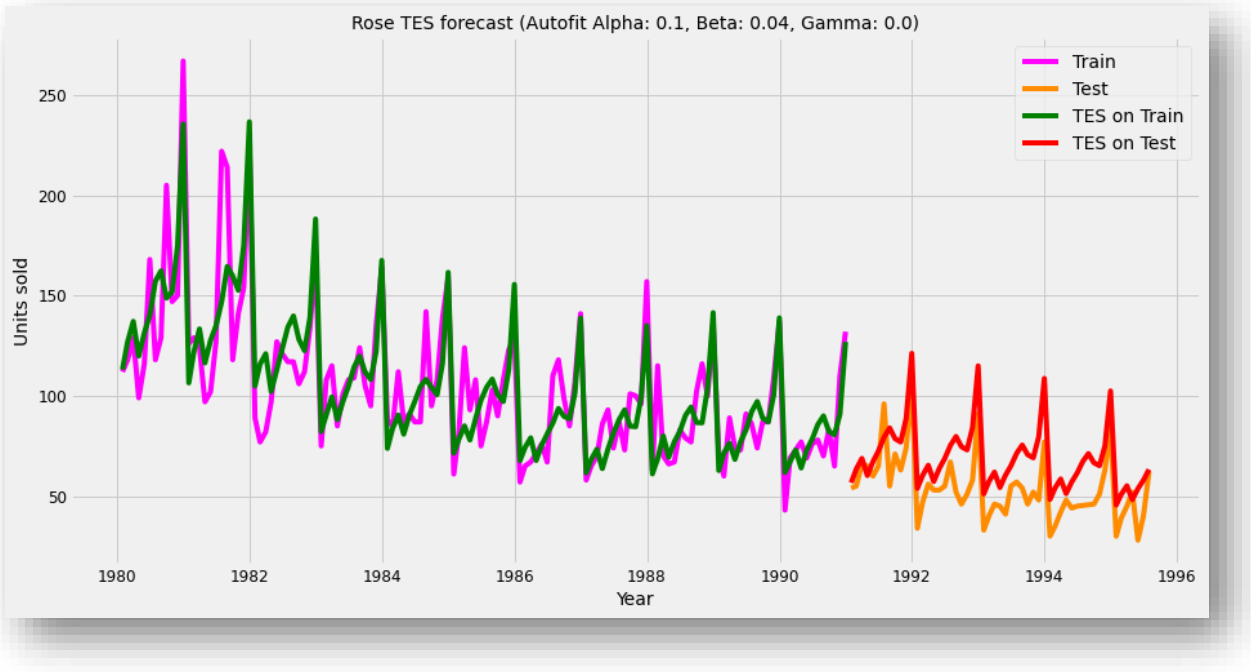
	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
151	0.2	0.6	0.2	28.289836	18.09	10.031754	13.62
223	0.3	0.3	0.4	24.209084	16.78	10.169742	13.67
11	0.1	0.2	0.2	24.365597	15.36	9.640616	13.96
214	0.3	0.2	0.5	24.580627	16.87	10.413322	14.00
12	0.1	0.2	0.3	23.969166	15.13	9.935672	14.21

25. Triple exponential smoothing sorted test RMSE and MAPE heads to get best scores (alpha 0.1, beta 0.2, gamma 0.2)



32. Rose TES forecast (Alpha 0.1, Beta 0.2, Gamma 0.2)

Attempt autofit



33. Rose TES forecast (Autofit Alpha 0.1, Beta 0.04, Gamma 0.0)

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
11	0.1	0.2	0.2	24.365597	15.36	9.640616	13.96
12	0.1	0.2	0.3	23.969166	15.13	9.935672	14.21
10	0.1	0.2	0.1	25.529854	16.06	9.943512	14.39
142	0.2	0.5	0.3	27.631767	17.87	10.026322	14.34
151	0.2	0.6	0.2	28.289836	18.09	10.031754	13.62

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
151	0.2	0.6	0.2	28.289836	18.09	10.031754	13.62
223	0.3	0.3	0.4	24.209084	16.78	10.169742	13.67
11	0.1	0.2	0.2	24.365597	15.36	9.640616	13.96
214	0.3	0.2	0.5	24.580627	16.87	10.413322	14.00
12	0.1	0.2	0.3	23.969166	15.13	9.935672	14.21

	Alpha	Beta	Gamma	Train RMSE	Train MAPE	Test RMSE	Test MAPE
996	1.000000	1.000000	0.7	30724.126331	4617.55	23029.955358	11836.01
997	1.000000	1.000000	0.8	1218.755446	493.75	9626.710890	8580.97
998	1.000000	1.000000	0.9	14150.253251	2303.41	9691.905408	7916.16
999	1.000000	1.000000	1.0	1768.254189	614.79	8138.618610	6811.03
1000	0.106096	0.048438	0.0	18.578860	13.21	17.369211	28.88

26. Triple exponential smoothing sorted test RMSE and MAPE heads to get best scores (alpha 0.1, beta 0.04, gamma 0.0), along with tail

•The Triple Exponential Smoothing models (Holt-Winter's Model) is applicable when data has both trend and seasonality. Rose data contain both trend and seasonality significantly

• In first iteration, smoothing level (alpha), trend (beta) and seasonality (gamma) are fitted to the model iteratively from values 0.1 to 1 and the best combination was chosen based on the RMSE and MAPE values, which is as below with alpha 0.1, beta 0.2 and gamma 0.2

• On the second iteration the model was allowed to chose the optimized values using parameters 'optimized=True, use_brute=True'

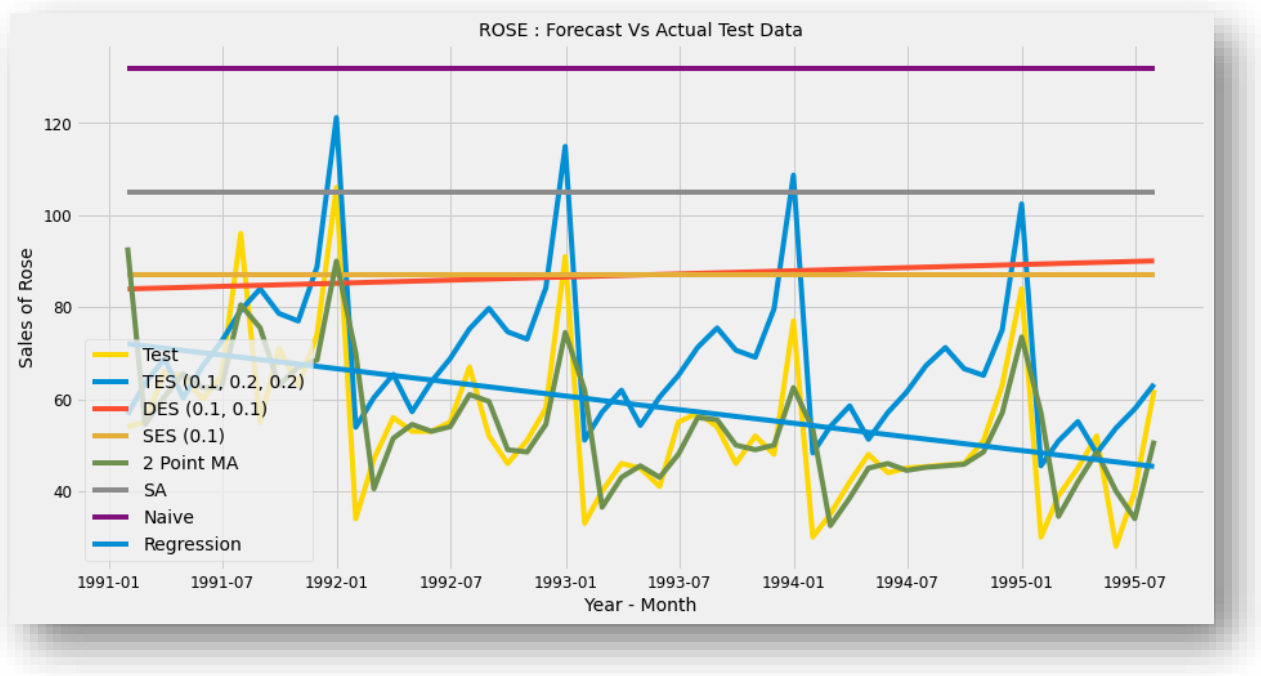
•The autofit model retuned higher accuracy in train dataset, much higher than the values from iteration 1, but faired poorly in accuracy in test

• The model evaluation parameters of the best models are given as above, including one from the autofit iteration

• The best model chosen as final one is the one with alpha 0.1, beta 0.2 and gamma 0.2

All exponential and other models, so far

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.640616	13.96
2 point TMA	11.529278	13.54
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
TES Alpha 0.11, Beta 0.05, Gamma 0.00	17.369211	28.88
SES Alpha 0.01	36.796022	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
DES Alpha 0.16, Beta 0.16	70.572197	120.25
NaiveModel	79.718559	145.10



34. Rose: Forecast vs Actual Test Data (all-model time series plot)

The accuracy of the time-series forecast models build in the previous sections of this report is as below, sorted by RMSE in test data

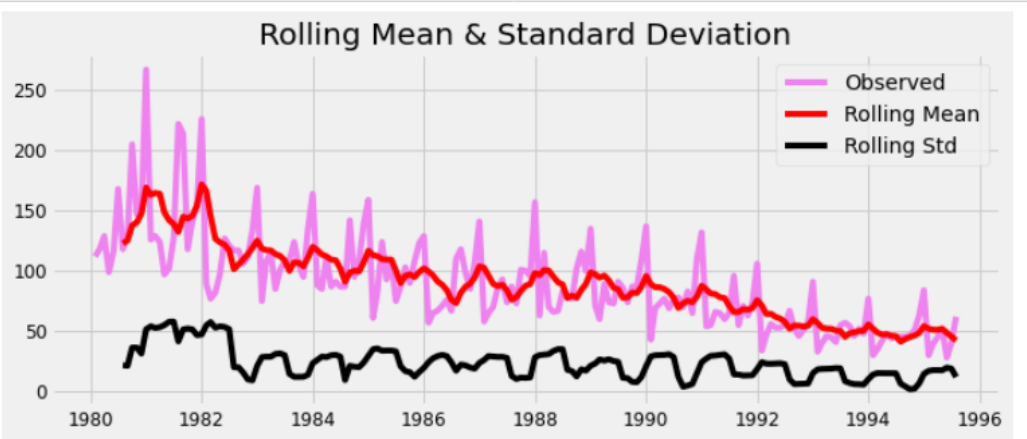
• The plot of the forecasts fitted on to the test data is given as well

• From the comparison of accuracy values and the plot it can be inferred that Triple Exponential Smoothing is the best model, which has trend as well as seasonality components fitting well with the test data

• 2 point trailing moving average model is also found to have fit well with a slight lag in test dataset

ARIMA Models

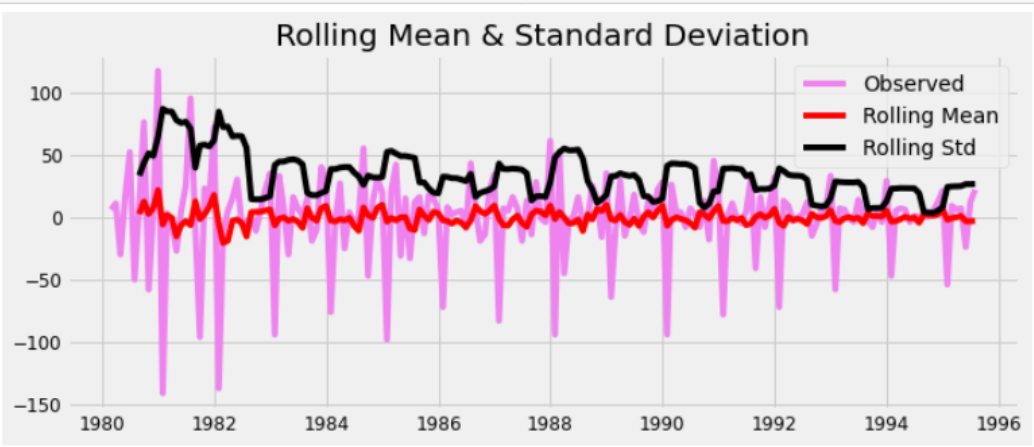
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and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




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876719
p-value                          0.343091
#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
```

35. Dickey-Fuller Test (rolling mean and standard deviation),
on original series, at 5% significant level

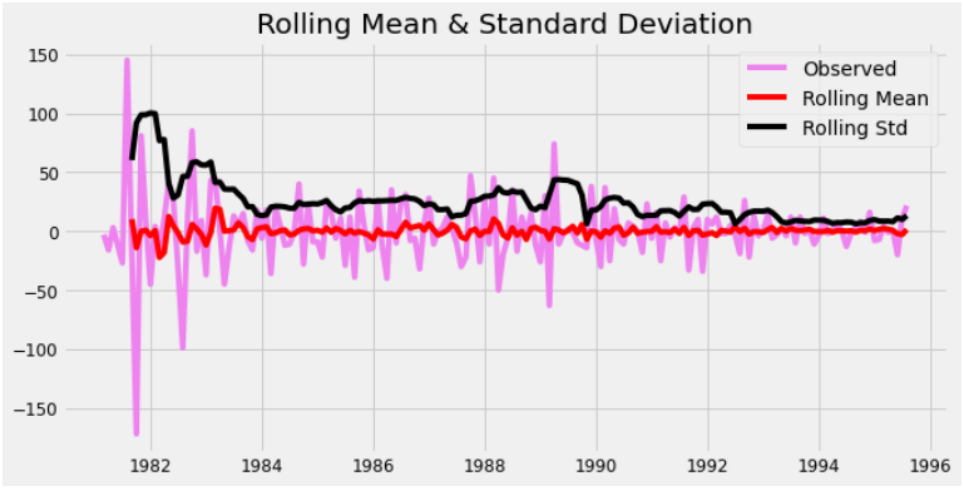
We see that at 5% significant level, the time series is non-stationary, as the p-value is high




```
Results of Dickey-Fuller Test:
Test Statistic                -8.044395e+00
p-value                        1.810868e-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
```

36. Dickey-Fuller Test (rolling mean and standard deviation) with difference of order 1

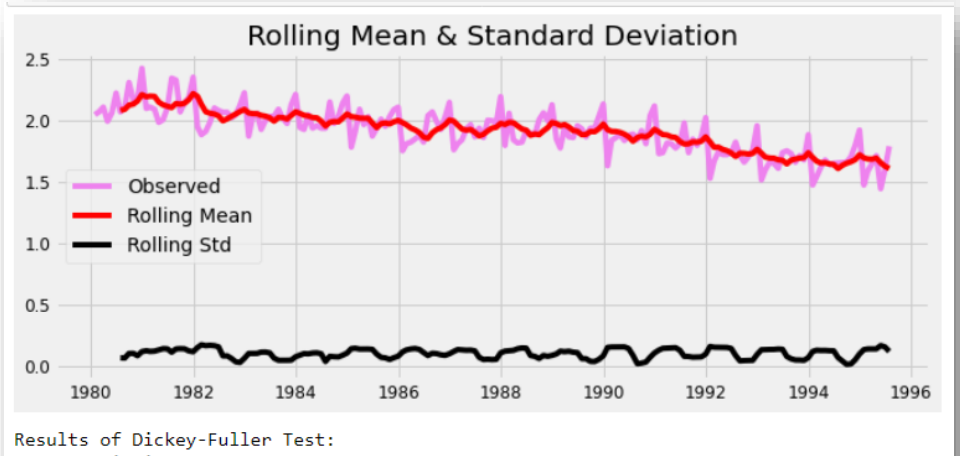
At difference of order 1, Rose time series is stationary with no trend.
We have visible seasoality but it it additive or mutiplicative?




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.605732
p-value                          0.000126
#Lags Used                      11.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value (10%)            -2.576373
dtype: float64
```

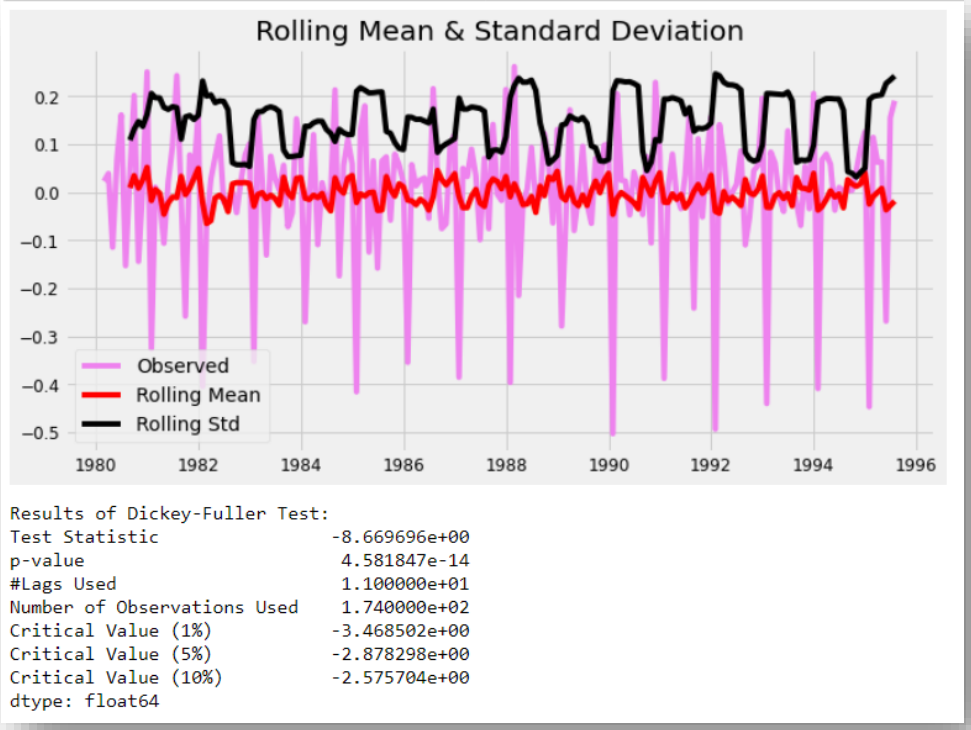
37. Dickey-Fuller Test (rolling mean and standard deviation)
with differencing of seasonal order (12)

Series now stationary




```
Results of Dickey-Fuller Test:
Test Statistic                  -0.422555
p-value                          0.906264
#Lags Used                      12.000000
Number of Observations Used    174.000000
Critical Value (1%)             -3.468502
Critical Value (5%)             -2.878298
Critical Value (10%)            -2.575704
dtype: float64
```

38. ADF (rolling mean and standard deviation)
with logarithmic transformation of the train data, using log10.
Logged series not stationary



39. ADF (rolling mean and standard deviation) with differenced logarithmic transformation of the train data, using log10. P is low, series is now stationary.

Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determines the presence of unit root in the series to understand if the series is stationary or not

• 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

•P-Value > alpha .05
• Test statistic > Critical values

Fail to reject the null hypothesis.
The series is non-stationary

IF

• P-Value < alpha .05
• Test statistic < Critical values

Reject the null hypothesis.
The series is stationary

• If we fail to reject the null hypothesis, it can say that the series is non-stationary and if we accept the null hypothesis, it can say that the series is stationary

• The ADF test on the original Rose series retuned the below values, where p-value is greater than alpha .05 so we fail to reject the null hypothesis

Differencing of order one is applied on the Rose series as above and tested for stationarity

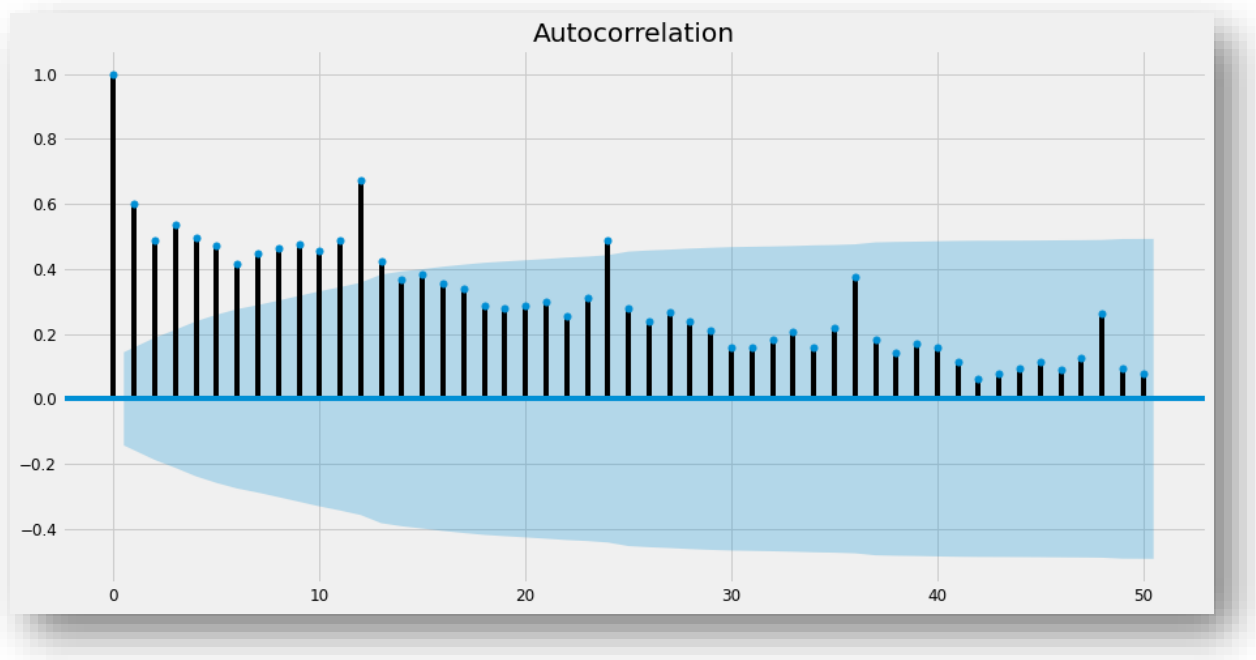
• At an order of differencing 1, the series is found to be stationary as above

• The rolling mean and standard deviation is also plotted to understand the component of seasonality and to ascertain if its multiplicative or additive in character

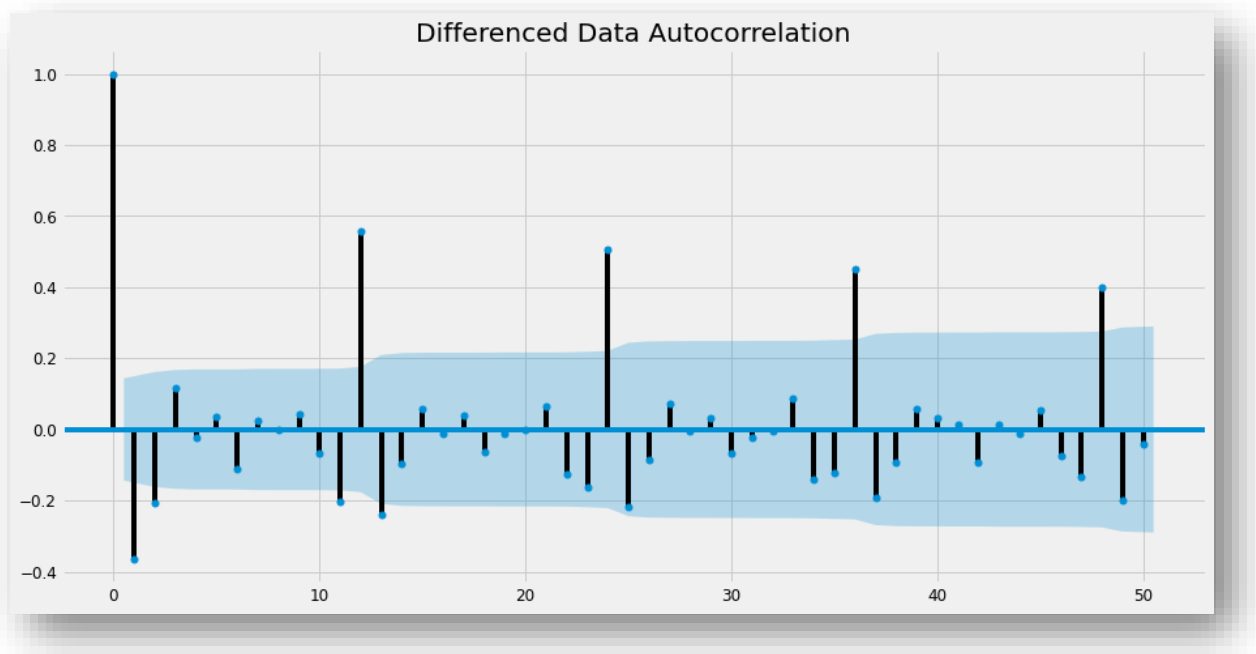
• The plot of rolling mean and standard deviation indicates that the seasonality is multiplicative as the altitude of plot varies with respect to trend

• The ADF test is also done in this exercise with logarithmic transformation of the train data and differencing of seasonal order (12), to understand if removing the multiplicity of the seasonal component will have an impact on the accuracy of model

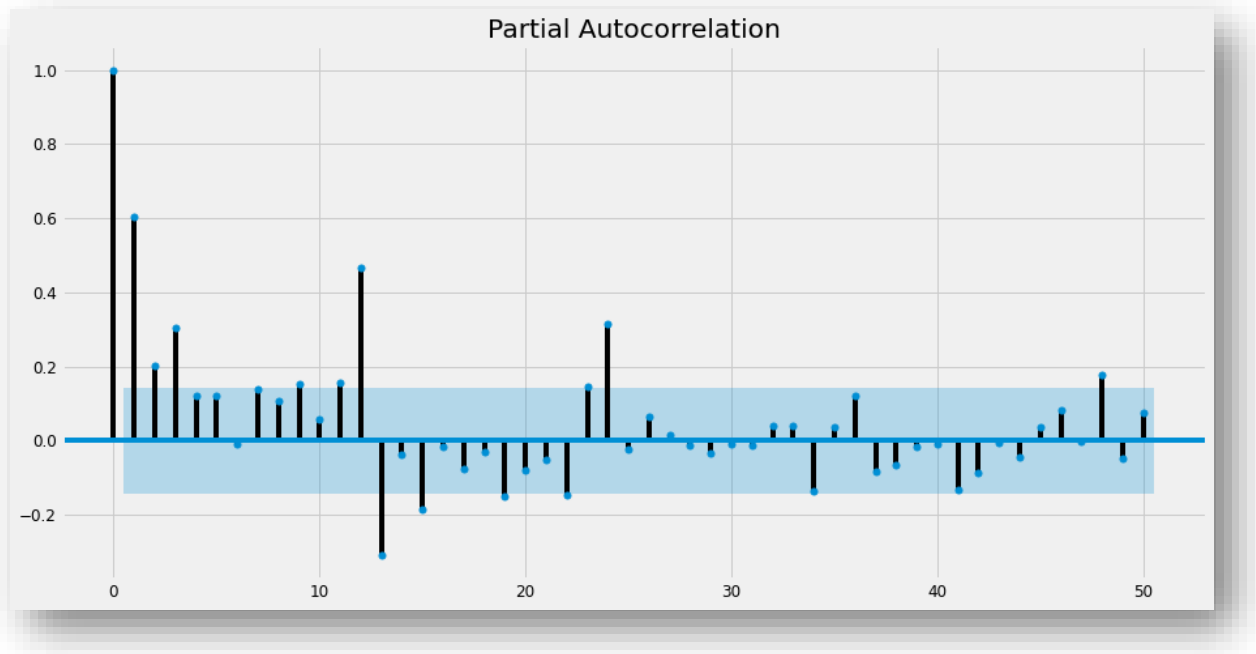
Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.



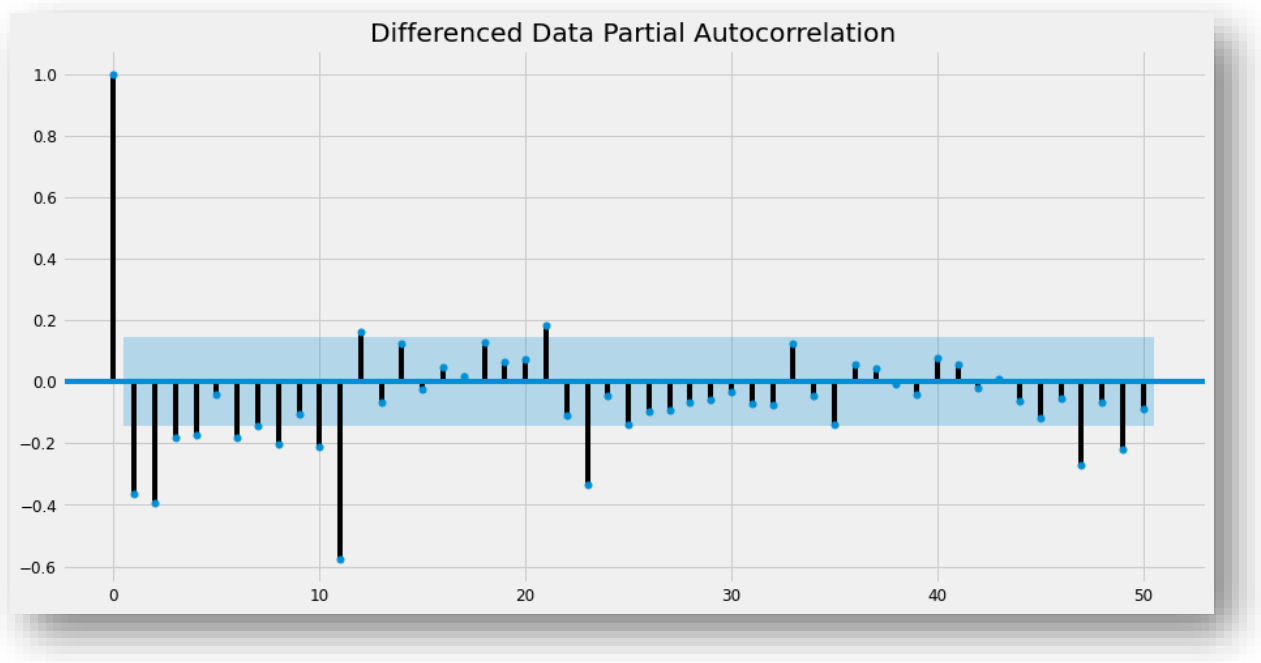
40. Rose – Autocorrelation



41. Rose - Differenced data autocorrelation



42. Rose - Partial Autocorrelation



43. Differenced Data Partial Autocorrelation

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.

Model 8: SARIMA

Why no ARIMA

As the Rose series of data contains seasonality component, we choose to build a SARIMA model instead of an ARIMA model

ARIMA can be limited in forecasting extreme values. While the model is adept at modelling seasonality and trends, outliers are difficult to forecast for ARIMA for the very reason that they lie outside of the general trend as captured by the model.

AUTO SARIMA on original data

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 1, 1, 12)
Model: (0, 1, 2)(0, 1, 2, 12)
Model: (0, 1, 3)(0, 1, 3, 12)
Model: (1, 1, 0)(1, 1, 0, 12)
Model: (1, 1, 1)(1, 1, 1, 12)
Model: (1, 1, 2)(1, 1, 2, 12)
Model: (1, 1, 3)(1, 1, 3, 12)
Model: (2, 1, 0)(2, 1, 0, 12)
Model: (2, 1, 1)(2, 1, 1, 12)
Model: (2, 1, 2)(2, 1, 2, 12)
Model: (2, 1, 3)(2, 1, 3, 12)
Model: (3, 1, 0)(3, 1, 0, 12)
Model: (3, 1, 1)(3, 1, 1, 12)
Model: (3, 1, 2)(3, 1, 2, 12)
Model: (3, 1, 3)(3, 1, 3, 12)
```

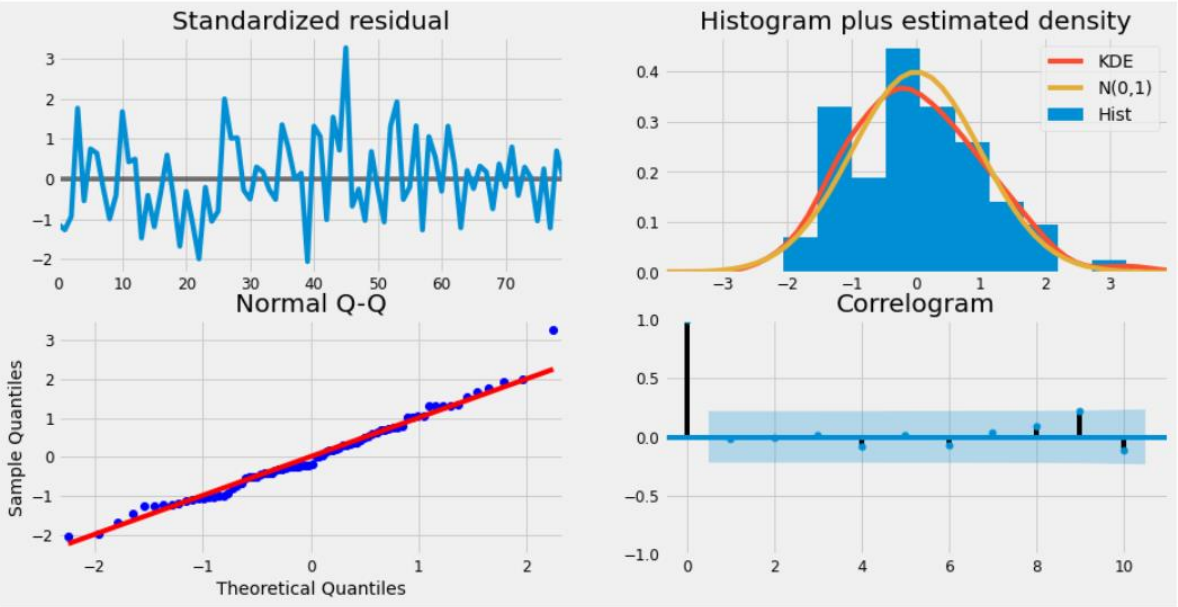
Auto Sarima on original data, examples of some parameter combinations for model

	param	seasonal	AIC
221	(3, 1, 1)	(3, 1, 1, 12)	681.362817
253	(3, 1, 3)	(3, 1, 1, 12)	681.607795
254	(3, 1, 3)	(3, 1, 2, 12)	681.972961
222	(3, 1, 1)	(3, 1, 2, 12)	682.320699
237	(3, 1, 2)	(3, 1, 1, 12)	683.211700

28. Auto Sarima models with best AIC scores

```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 1, 1, 12)   Log Likelihood                -331.681
Date:                            Fri, 08 Oct 2021   AIC                            681.363
Time:                                    07:05:55   BIC                            702.801
Sample:                                         0   HQIC                           689.958
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0171      0.151      0.113      0.910      -0.279       0.313
ar.L2         -0.0425      0.141     -0.302      0.763      -0.319       0.234
ar.L3         -0.0574      0.119     -0.484      0.629      -0.290       0.175
ma.L1         -0.9388      0.084    -11.113      0.000      -1.104      -0.773
ar.S.L12       0.0906      0.126      0.720      0.471      -0.156       0.337
ar.S.L24      -0.0438      0.108     -0.407      0.684      -0.255       0.167
ar.S.L36    -3.87e-05      0.052     -0.001      0.999      -0.103       0.103
ma.S.L12      -0.9998    253.299     -0.004      0.997    -497.456     495.456
sigma2       185.4082   4.69e+04      0.004      0.997   -9.18e+04    9.22e+04
==========================================================================================
Ljung-Box (Q):                       42.97   Jarque-Bera (JB):                 2.56
Prob(Q):                              0.35   Prob(JB):                         0.28
Heteroskedasticity (H):               0.56   Skew:                             0.42
Prob(H) (two-sided):                  0.13   Kurtosis:                         3.22
==========================================================================================
```

44. Auto SARIMA (3,1,1)x(3,1,1,2) model: Results



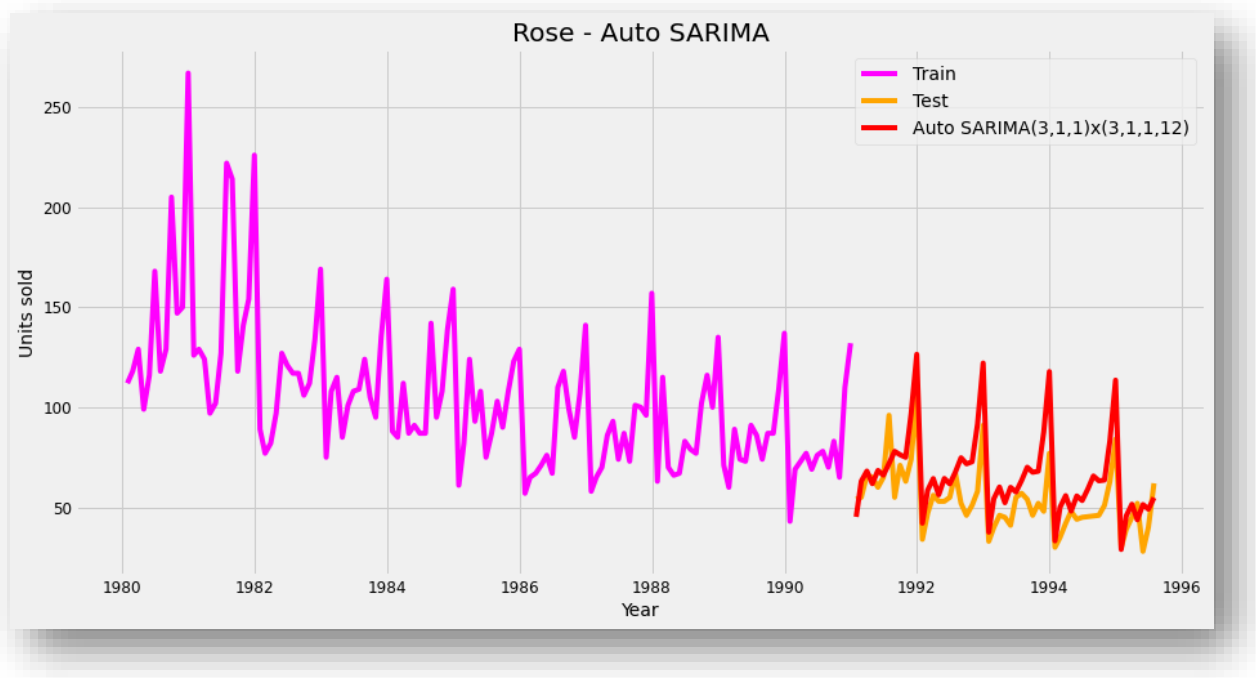
45. Auto SARIMA model: Diagnostics

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	45.225140	14.457938	16.888102	73.562179
1	63.054674	14.502678	34.629947	91.479401
2	68.114801	14.452769	39.787894	96.441709
3	61.822527	14.449015	33.502979	90.142076
4	68.435877	14.469562	40.076058	96.795697

29. Auto Sarima (3,1,1)x(3,1,1,12) prediction summary

Time_Stamp	Rose	rose_auto_forecasted
1991-01-31	54.0	45.225140
1991-02-28	55.0	63.054674
1991-03-31	66.0	68.114801
1991-04-30	65.0	61.822527
1991-05-31	60.0	68.435877

30. Predicted and true values of time series for Auto Sarima (3,1,1)x(3,1,1,12)



46. Rose Auto Sarima (3,1,1)x(3,1,1,12)

Auto SARIMA on log

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

Auto Sarima on log10, examples of some parameter combinations for model

	param	seasonal	AIC
115	(1, 0, 0)	(1, 0, 1, 12)	-257.620745
7	(0, 0, 0)	(1, 0, 1, 12)	-256.170282
133	(1, 0, 1)	(1, 0, 1, 12)	-255.482061
25	(0, 0, 1)	(1, 0, 1, 12)	-254.978844
223	(2, 0, 0)	(1, 0, 1, 12)	-253.620649

32. Auto Sarima on log10: Models with best AIC scores

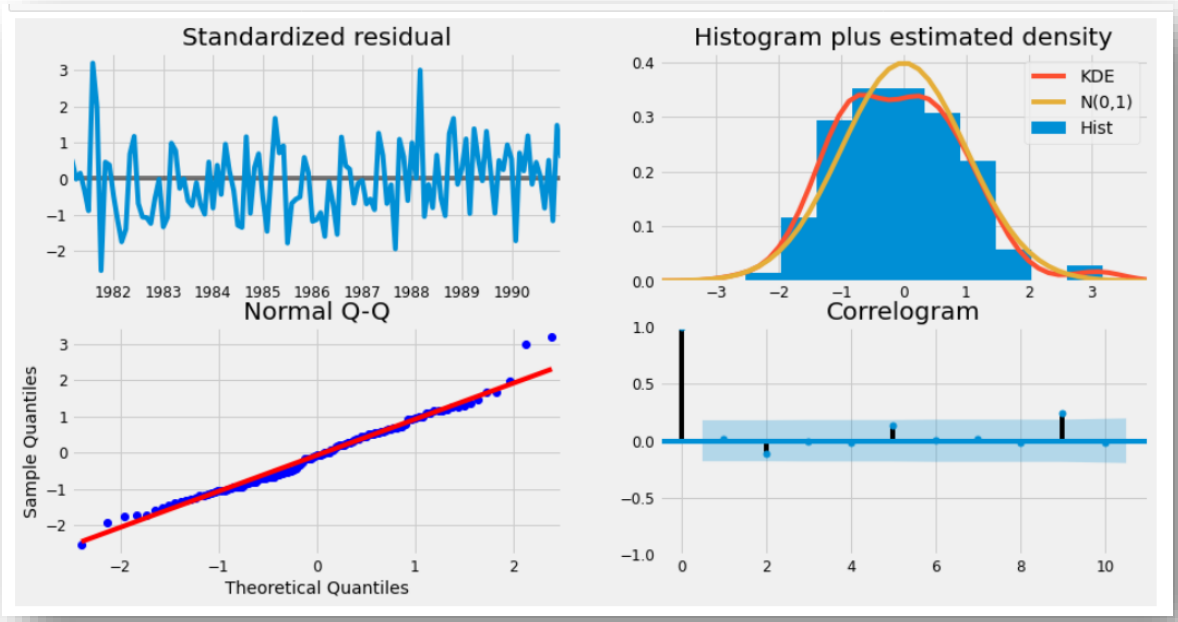
Statespace Model Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(1, 0, 0)x(1, 0, 1, 12)	Log Likelihood	132.810
Date:	Fri, 08 Oct 2021	AIC	-257.621
Time:	07:13:35	BIC	-246.504
Sample:	01-31-1980	HQIC	-253.107
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1690	0.078	2.180	0.029	0.017	0.321
ar.S.L12	0.9872	0.001	751.773	0.000	0.985	0.990
ma.S.L12	-0.9410	0.350	-2.688	0.007	-1.627	-0.255
sigma2	0.0052	0.002	2.890	0.004	0.002	0.009

Ljung-Box (Q):	24.28	Jarque-Bera (JB):	4.00
Prob(Q):	0.98	Prob(JB):	0.14
Heteroskedasticity (H):	0.86	Skew:	0.40
Prob(H) (two-sided):	0.64	Kurtosis:	3.40

Auto SARIMA on log10 model: Results



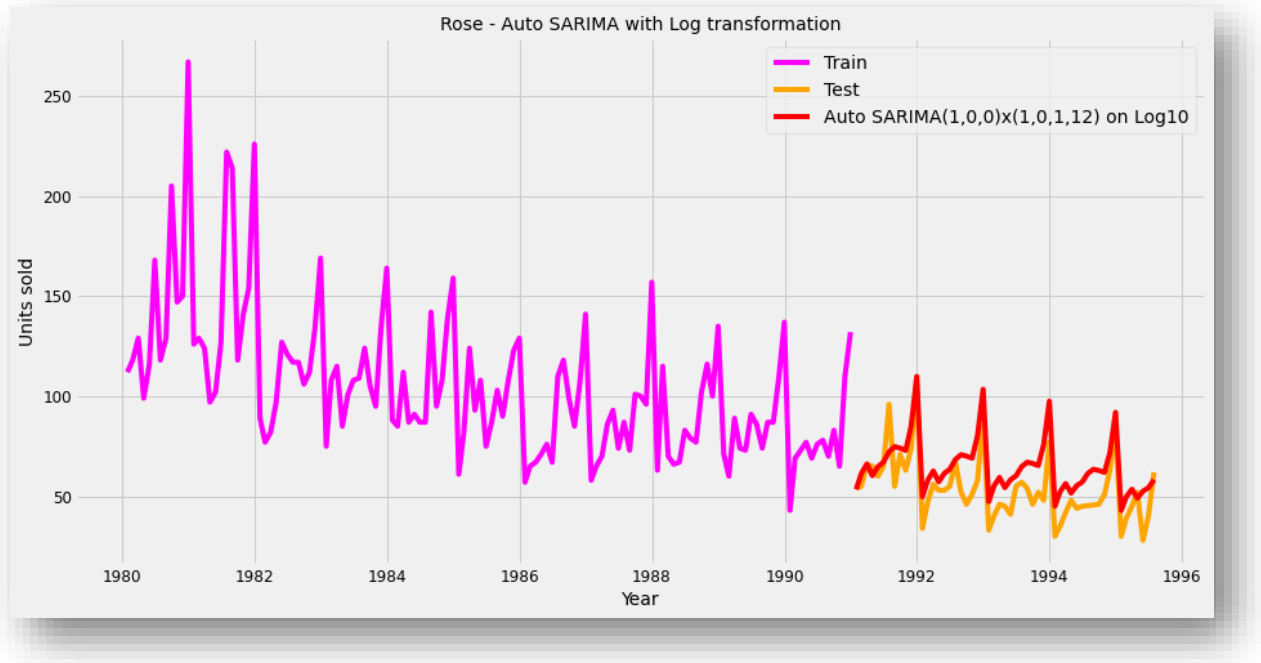
47. Auto SARIMA on log10 model: Diagnostics

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-31	1.727954	0.073887	1.583139	1.872769
1991-02-28	1.787585	0.074674	1.641227	1.933943
1991-03-31	1.820432	0.074696	1.674030	1.966834
1991-04-30	1.780330	0.074697	1.633927	1.926733
1991-05-31	1.811704	0.074697	1.665300	1.958107

Auto Sarima (1,0,0)x(1,0,1,12)on log10: prediction summary

Time_Stamp	Rose	rose_auto_forecasted	rose_log_auto_forecasted
1991-01-31	54.0	45.225140	53.450768
1991-02-28	55.0	63.054674	61.317622
1991-03-31	66.0	68.114801	66.135121
1991-04-30	65.0	61.822527	60.301793
1991-05-31	60.0	68.435877	64.819197

34. Predicted and true values of time series for
Auto Sarima (1,0,0)x(1,0,1,12) log10 transformation model



48. Rose- Auto SARIMA (1,0,0)x(1,0,1,12) with Log Transformation

Auto Arima models analysis

Please, remember, that as the Rose series of data contain seasonality component, we will build SARIMA model rather than ARIMA

• Two iterations of automated SARIMA models were attempted in this exercise, one with original data and another with log transformation of the data, as the seasonality got apparent multiplicity

• The model built with log transformed data is found to be higher in accuracy scores of RMSE and MAPE, which is selected as the final model

• To handle multiplicity of seasonality, the data was log transformed to make it additive

• The optimal parameters for (p, d, q)x(P, D, Q) were selected in accordance with the lowest Akaike Information Criteria (AIC) values • The top three models with lowest AIC values are as given here and the final selected one is (1, 0, 0)x(1, 0, 1, 12)

• The diagnostics plot of the model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution

• The Normal Q -Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line

• The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index

• The RMSE and MAPE values of the automated SARIMA models built are given here

• The diagnostics plot of the selected model is given

• From the model summary, it can be inferred that seasonal AR(2) term has the highest weightage, followed by seasonal MA(2)

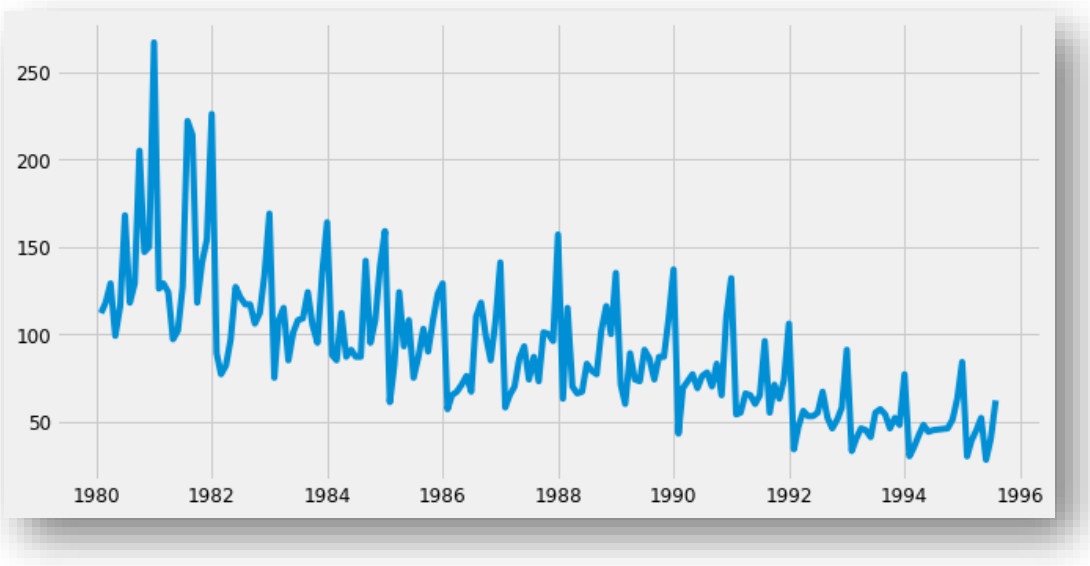
• From the p-values it can be inferred that all the AR and MA terms are significant as the values are below .05

	Test RMSE	Test MAPE
Auto SARIMA(3,1,1)x(3,1,1,12)	16.822203	25.48
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.589879	21.92

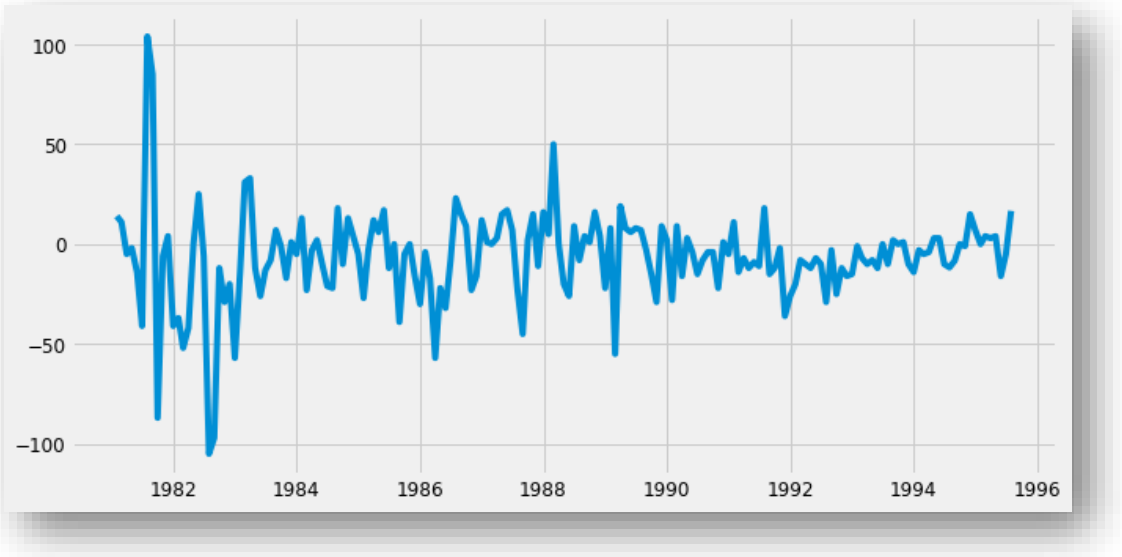
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Manual SARIMA

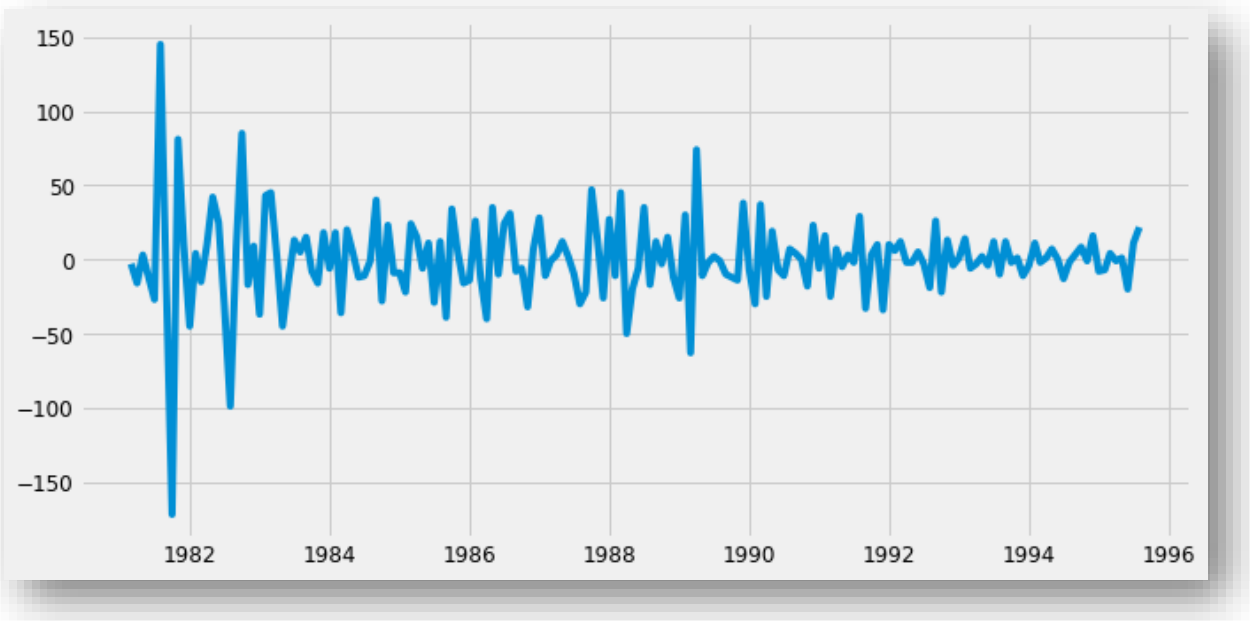
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.



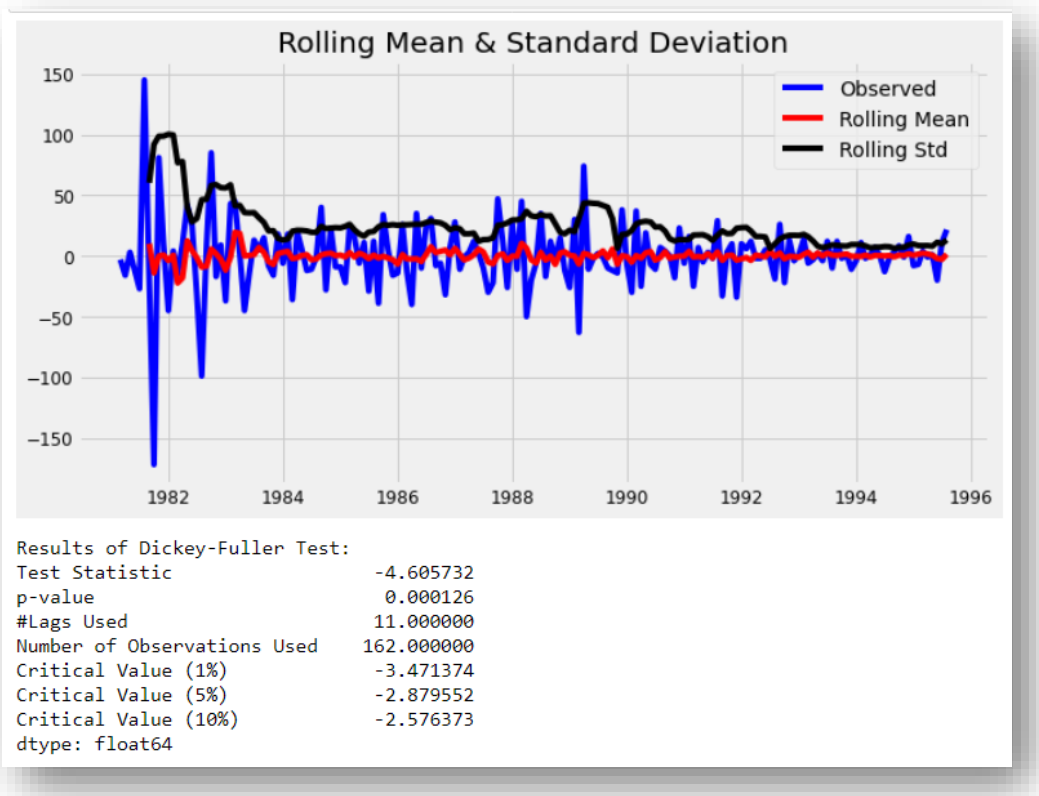
Observed time series before applying Manual SARIMA



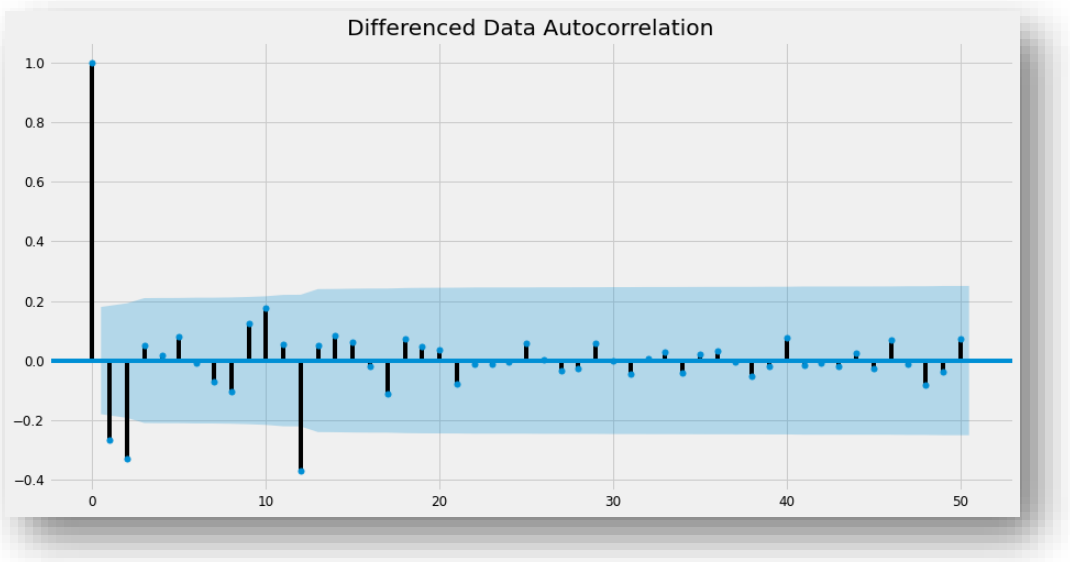
50. Series after seasonal differencing, before applying Manual SARIMA



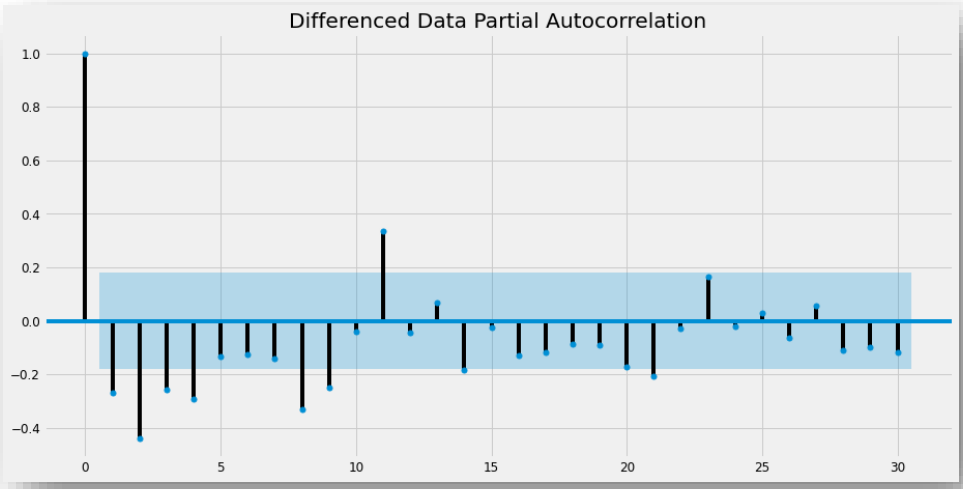
51. Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA



52. ADF Test of stationarity on Series after seasonal differencing + 1-order differencing, before applying Manual SARIMA
p is low, series is stationary



53. Differenced Data Autocorrelation



54. Differenced Data Partial Autocorrelation

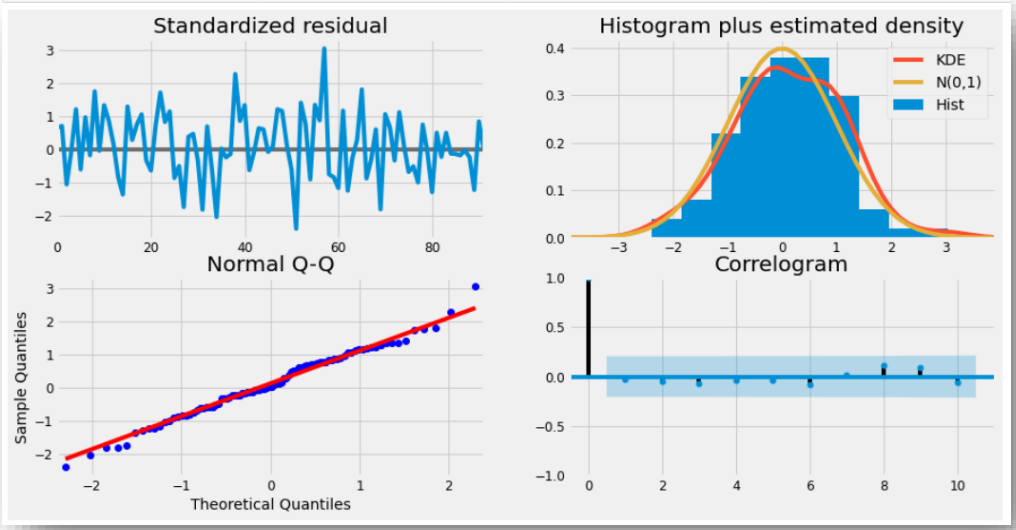
Statespace Model Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-384.369
Date:	Fri, 08 Oct 2021	AIC	786.737
Time:	07:13:41	BIC	809.433
Sample:	0	HQIC	795.898
	- 132		
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 (Q):	27.59	Jarque-Bera (JB):	0.01
Prob(Q):	0.93	Prob(JB):	0.99
Heteroskedasticity (H):	0.76	Skew:	-0.01
Prob(H) (two-sided):	0.46	Kurtosis:	3.06

54. Manual SARIMA (4,1,2)x(0,1,2,12): Results



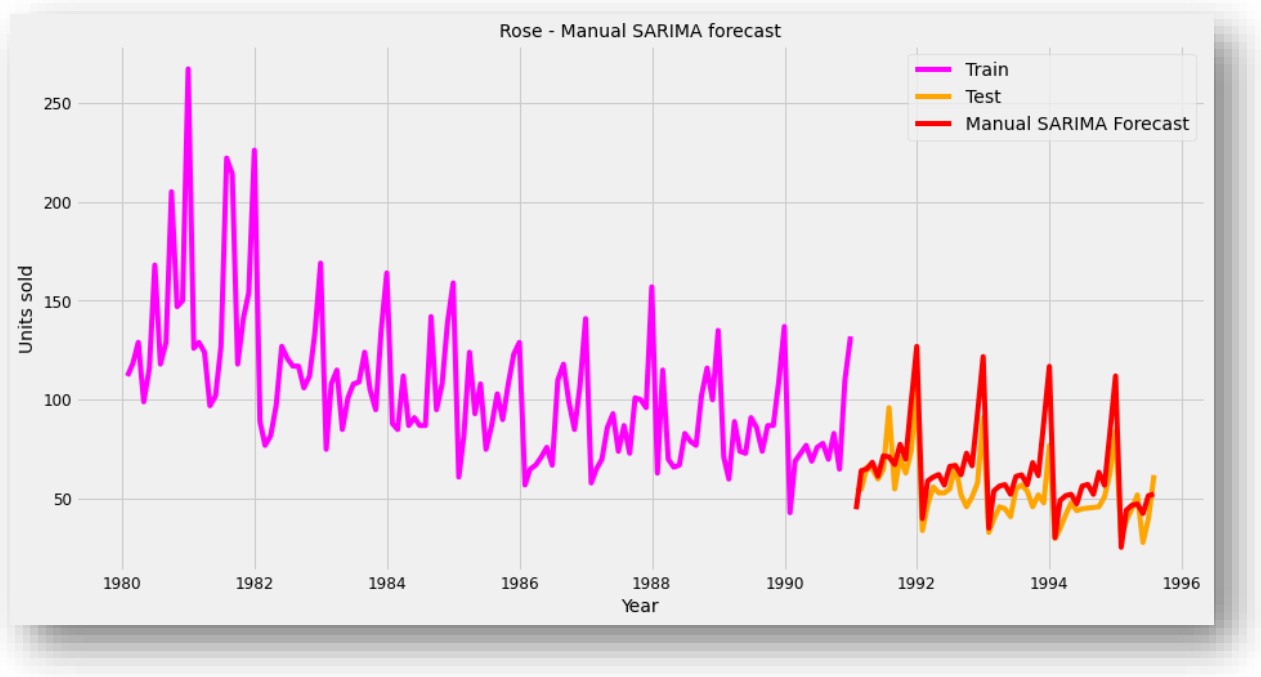
55. Manual SARIMA (4,1,2)x(0,1,2,12): Diagnostics

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	44.733041	15.552666	14.250376	75.215705
1	64.208693	16.000767	32.847767	95.569620
2	65.110689	16.074606	33.605041	96.616337
3	68.453063	16.150995	36.797694	100.108432
4	61.423433	16.154555	29.761087	93.085780

Manual Sarima, prediction summary

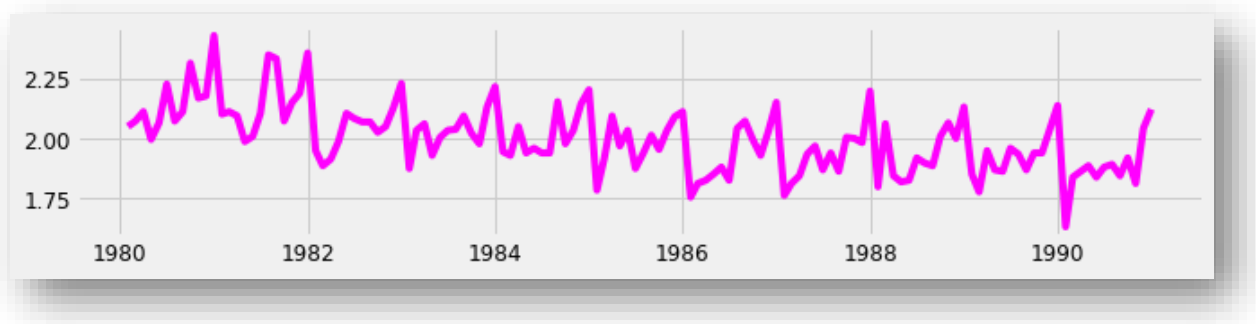
Time_Stamp	Rose	rose_auto_forecasted	rose_log_auto_forecasted	rose_manual_forecasted
1991-01-31	54.0	45.225140	53.450768	44.733041
1991-02-28	55.0	63.054674	61.317622	64.208693
1991-03-31	66.0	68.114801	66.135121	65.110689
1991-04-30	65.0	61.822527	60.301793	68.453063
1991-05-31	60.0	68.435877	64.819197	61.423433

36. Manual Sarima (4,1,2)x(0,1,1,12): predicted and true values

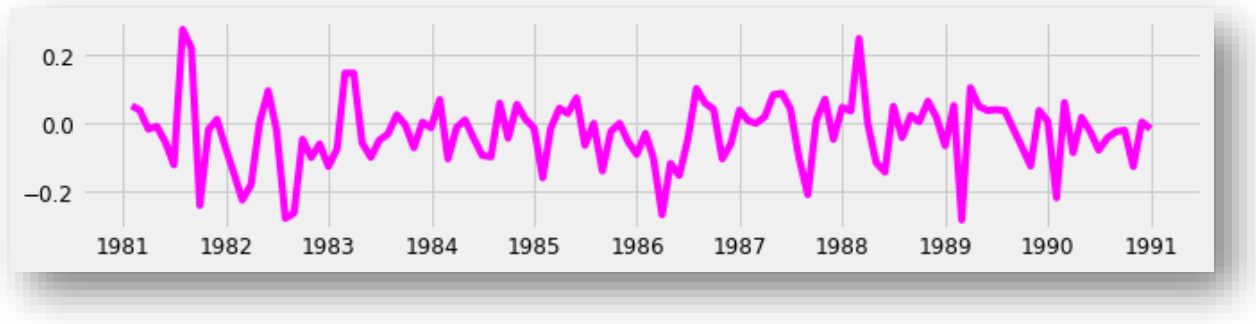


56. Rose - Manual SARIMA (4,1,2)x(0,1,2,12) forecast

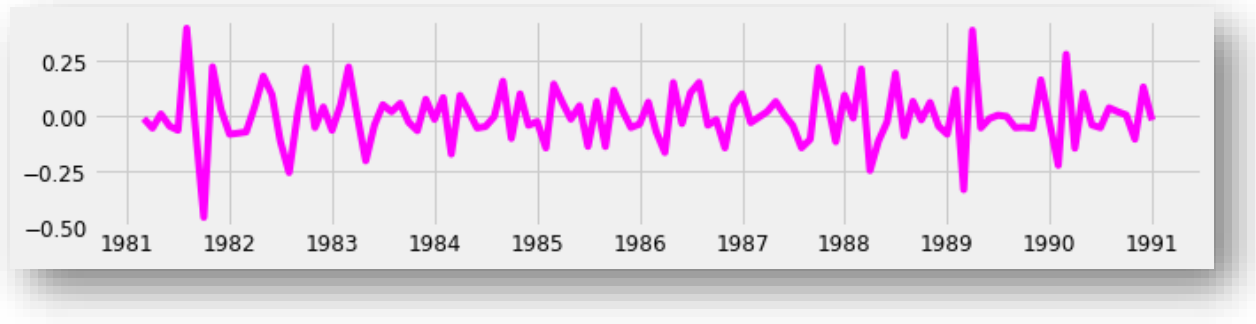
Manual SARIMA on Log



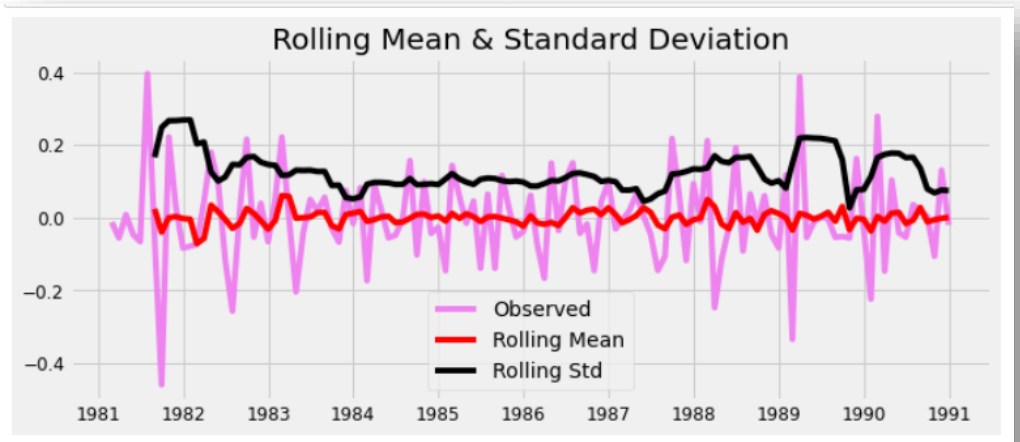
57. Manual SARIMA on log10: Observed logged time series, before applying model



57. Logged series after seasonal differencing, before applying Manual SARIMA log10 model

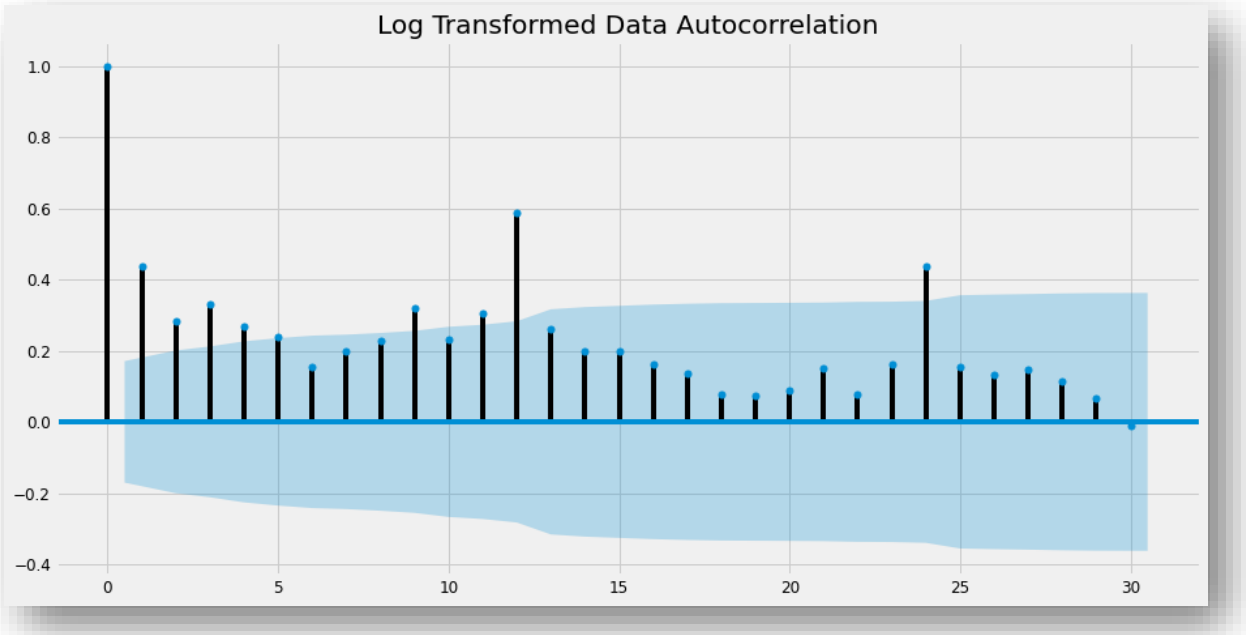


58. Series after seasonal differencing + 1-order differencing,
before applying Manual SARIMA log10 model

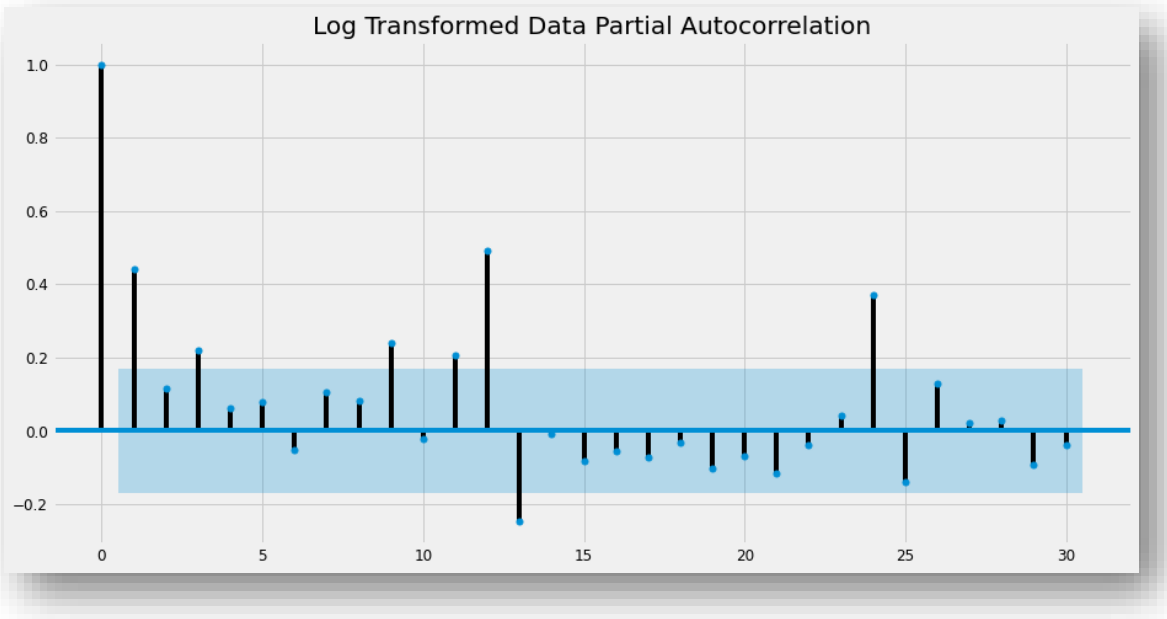



```
Results of Dickey-Fuller Test:
Test Statistic                  -3.910109
p-value                          0.001962
#Lags Used                      11.000000
Number of Observations Used    107.000000
Critical Value (1%)             -3.492996
Critical Value (5%)             -2.888955
Critical Value (10%)            -2.581393
dtype: float64
```

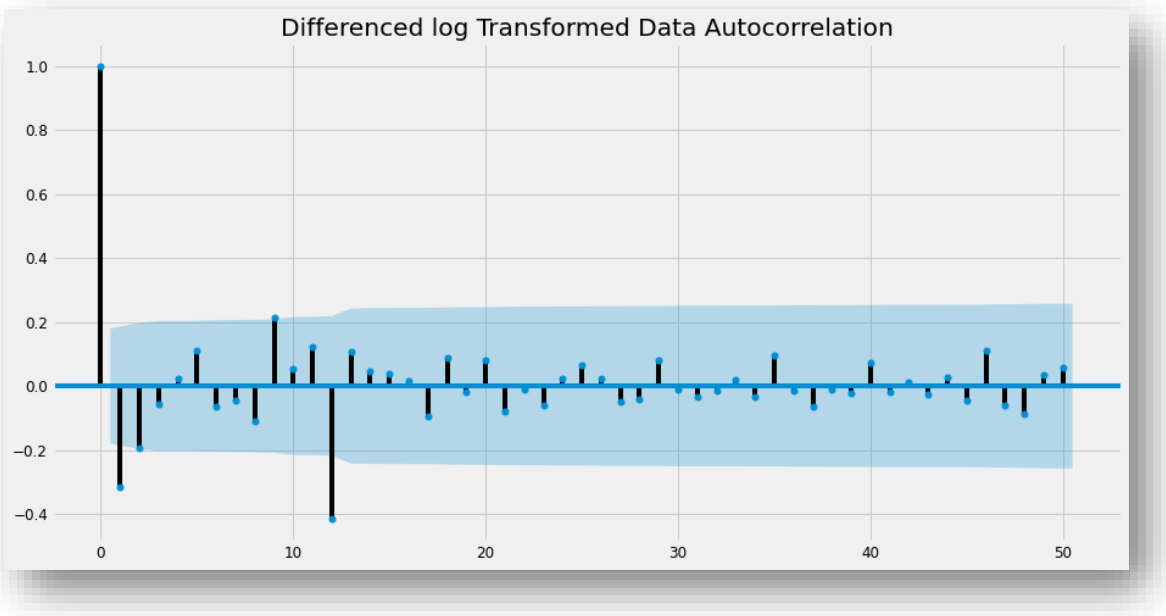
59. ADF Test of stationarity on logged series after seasonal differencing + 1-order differencing, before applying Manual SARIMA. p is low, which means that the series is stationary



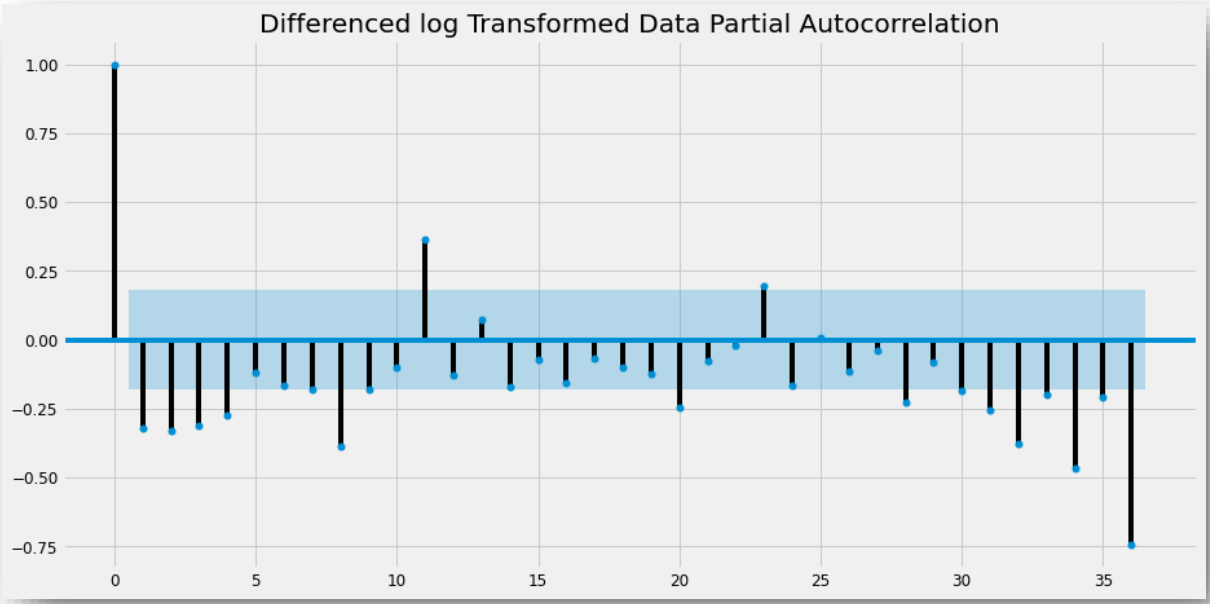
60. Log transformed data autocorrelation



61. Log transformed partial data autocorrelation



62. Differenced log transformed data autocorrelation



63. Differenced log transformed data partial autocorrelation

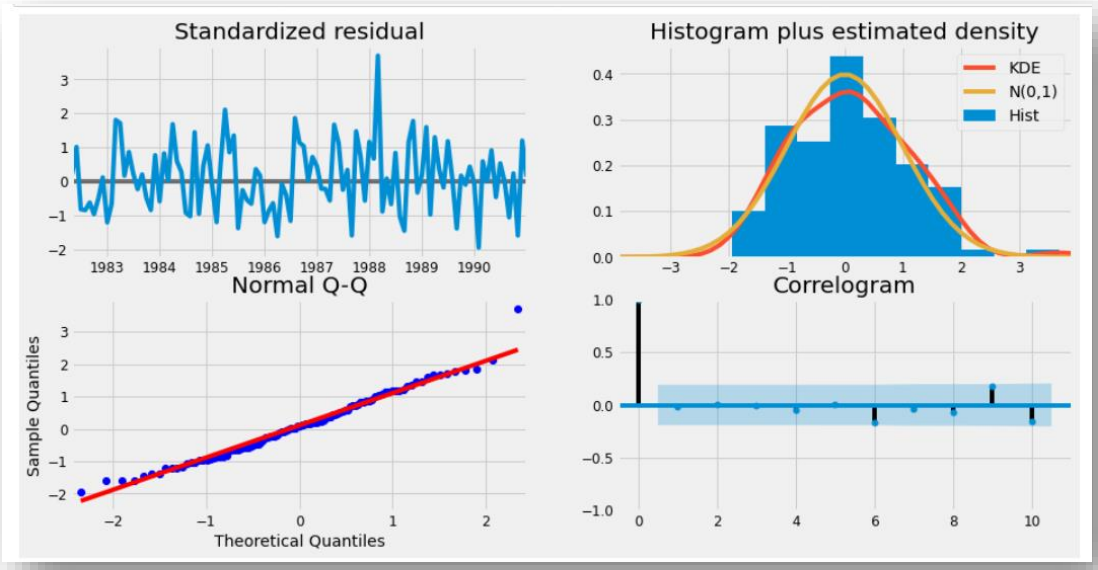
Statespace Model Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)	Log Likelihood	128.764
Date:	Fri, 08 Oct 2021	AIC	-243.528
Time:	07:13:46	BIC	-224.950
Sample:	01-31-1980 - 12-31-1990	HQIC	-236.000
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0013	0.118	-0.011	0.991	-0.232	0.229
ar.L2	-0.1549	0.126	-1.226	0.220	-0.403	0.093
ar.L3	-0.1594	0.113	-1.416	0.157	-0.380	0.061
ar.L4	-0.1503	0.121	-1.241	0.215	-0.388	0.087
ma.L1	-0.8435	0.074	-11.395	0.000	-0.989	-0.698
ma.S.L12	-0.9990	23.823	-0.042	0.967	-47.692	45.694
sigma2	0.0041	0.098	0.042	0.966	-0.187	0.195

Ljung-Box (Q):	45.58	Jarque-Bera (JB):	3.83
Prob(Q):	0.25	Prob(JB):	0.15
Heteroskedasticity (H):	1.60	Skew:	0.44
Prob(H) (two-sided):	0.17	Kurtosis:	3.34

64. Manual SARIMA (4,1,1)x(0,1,1,12) on log10 model: Results



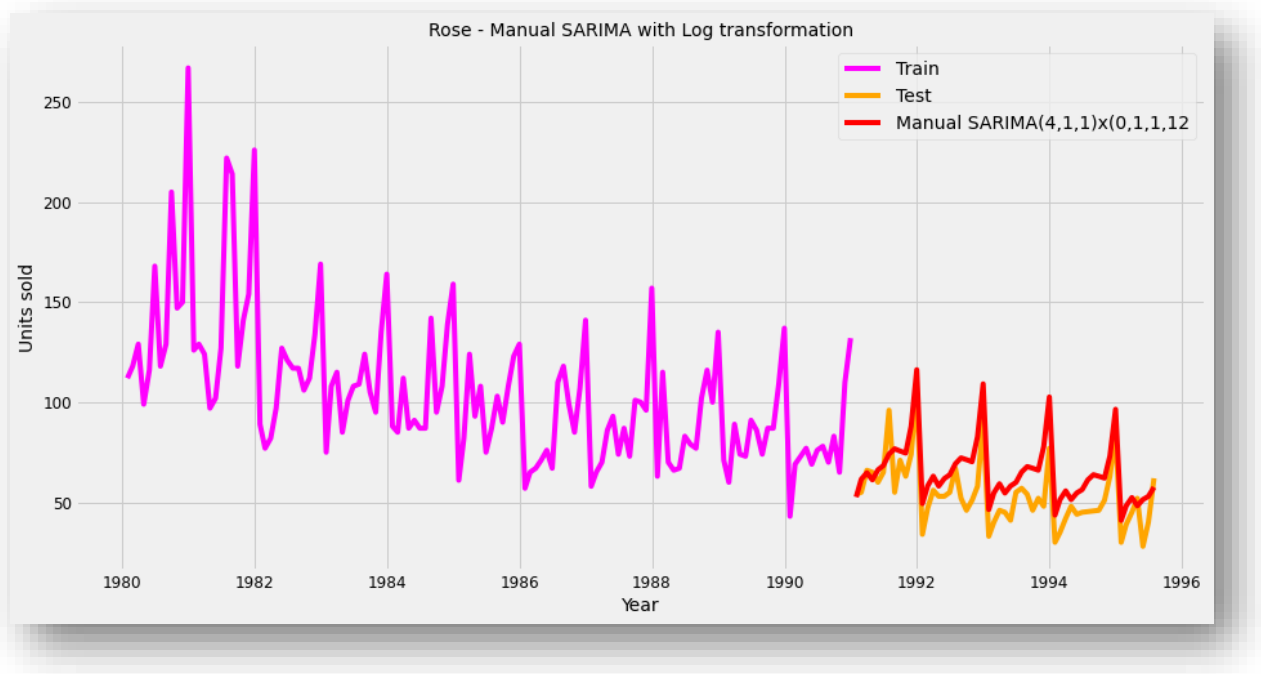
64. Manual SARIMA (4,1,1)x(0,1,1,12) on log10 model: Diagnostics

	Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-31	1.721104	0.067180	1.589433	1.852775
1991-02-28	1.789043	0.067984	1.655795	1.922290
1991-03-31	1.811286	0.067982	1.678043	1.944528
1991-04-30	1.786831	0.068004	1.653546	1.920116
1991-05-31	1.820831	0.067848	1.687852	1.953811

37. Manual Sarima on Log10, prediction summary

Time_Stamp	Rose	rose_auto_forecasted	rose_log_auto_forecasted	rose_manual_forecasted	rose_log_manual_forecasted
1991-01-31	54.0	45.225140	53.450768	44.733041	52.614317
1991-02-28	55.0	63.054674	61.317622	64.208693	61.523709
1991-03-31	66.0	68.114801	66.135121	65.110689	64.756832
1991-04-30	65.0	61.822527	60.301793	68.453063	61.211239
1991-05-31	60.0	68.435877	64.819197	61.423433	66.195958

38. Manual Sarima (4,1,1)x(0,1,1,12) on Log10, predicted and true values



65. Rose - Manual SARIMA (4,1,1)x(0,1,1,12) with log transformation

Manual SARIMA analysis

Log transformation of the data is done to handle multiplicity of seasonality

• From the ACF plot of the log transformed data, it can be seen that at seasonal interval of 12, the plot is not quickly tapering off. So we need to take a seasonal differencing of 12

• From the plots below it can be seen that a slight trend is still existing after differencing of seasonal order of 12. With a further differencing of order one, no trend is present

• We have done an ADF test to check the stationarity after the above differencing. With a p-value below alpha 0.05 and test statistic below critical values, it can be confirmed that the data is stationary

• ACF and PACF plots of the seasonal-differenced + one order differenced data were created to find our the values for (p,d,q)x(P,D,Q)

Here we have taken alpha = 0.05 and seasonal period as 12

• From the PACF plot it can be seen that till lag 4 is significant before cut-off, so AR term ‘p = 4’ is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’

• From ACF plot, lag 1 and 2 are significant before it cuts off, so lets keep MA term ‘q = 1’ and at seasonal lag of 12, a significant lag is apparent, so lets keep ‘Q = 1’

• The final selected terms for SARIMA model is (4, 1, 1)x(0, 1, 1, 12), as inferred from the ACF and PACF plots

• The diagnostic plot for the model is as below, which clearly shows a normal distribution of residuals, where more values are around zero

• The Normal Q-Q plot also shows that the quantiles come from a normal distribution as the points forms roughly a straight line

• The correlogram shows the autocorrelation of the residuals and there are no points significant above the confidence index

• Forecast was fitted with the test data

•The model summary indicates that that none of the terms used in the model are significant in terms of p-values

• From the multiple iterations of SARIMA models, below is the comparison of the models in terms of its accuracy attributes of RMSE and MAPE

	Test RMSE	Test MAPE
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.589879	21.92
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.176466	23.10
Auto SARIMA(3,1,1)x(3,1,1,12)	16.822203	25.48

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

	Test RMSE	Test MAPE
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.640616	13.96
2 point TMA	11.529278	13.54
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.589879	21.92
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.176466	23.10
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
RegressionOnTime	15.268885	22.82
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
Auto SARIMA(3,1,1)x(3,1,1,12)	16.822203	25.48
TES Alpha 0.11, Beta 0.05, Gamma 0.00	17.369211	28.88
SES Alpha 0.01	36.796022	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
DES Alpha 0.16, Beta 0.16	70.572197	120.25
NaiveModel	79.718559	145.10

39. Most optimum model, by RMSE

	Test RMSE	Test MAPE
2 point TMA	11.529278	13.54
TES Alpha 0.1, Beta 0.2, Gamma 0.2	9.640616	13.96
4 point TMA	14.451364	19.49
6 point TMA	14.566269	20.82
9 point TMA	14.727594	21.01
Auto SARIMA(1,0,0)x(1,0,1,12)-Log10	13.589879	21.92
Manual SARIMA(4,1,2)x(0,1,1,12)	15.377144	22.16
RegressionOnTime	15.268885	22.82
Manual SARIMA(4,1,1)x(0,1,1,12)-Log10	14.176466	23.10
Auto SARIMA(3,1,1)x(3,1,1,12)	16.822203	25.48
TES Alpha 0.11, Beta 0.05, Gamma 0.00	17.369211	28.88
SES Alpha 0.01	36.796022	63.88
DES Alpha 0.10, Beta 0.10	37.056911	64.02
SimpleAverage	53.460350	94.93
DES Alpha 0.16, Beta 0.16	70.572197	120.25
NaiveModel	79.718559	145.10

40. Most optimum model, by MAPE

Model comparison

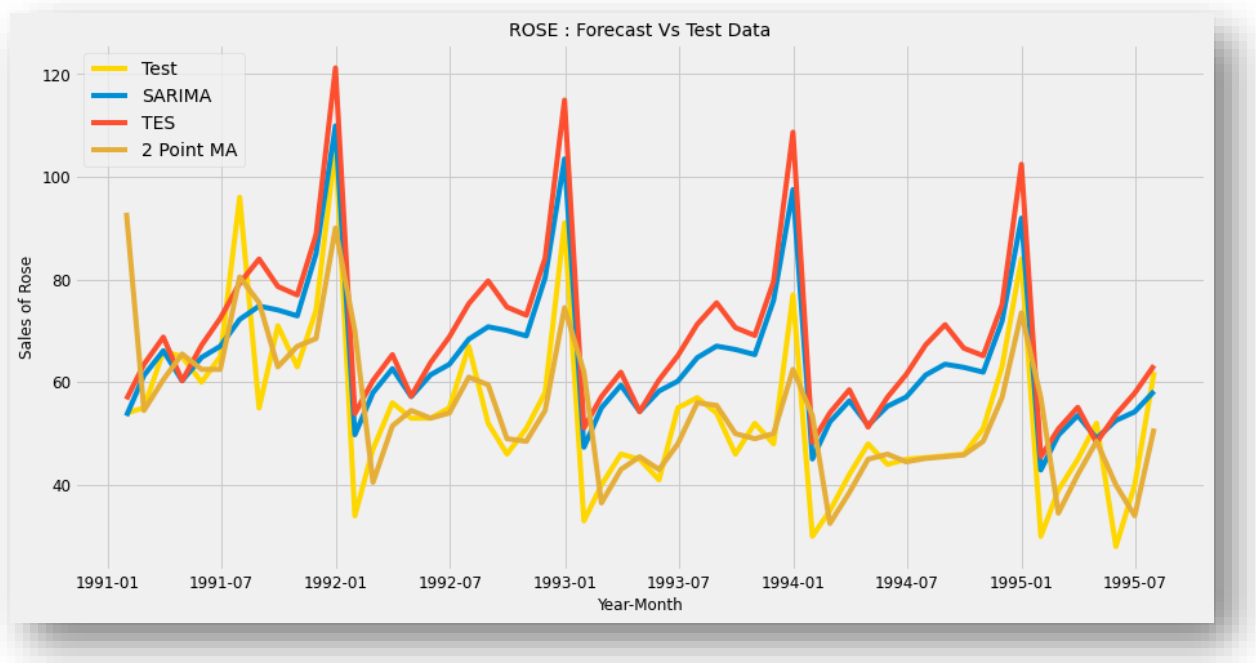
The overall comparison of all the time-series forecast models are listed below in accordance with increasing RMSE against test data or in the order of decreasing accuracy

• Triple Exponential Smoothing is found to be the best model, followed by 2 point Moving Average

• The best of SARIMA, Triple Exponential Smoothing and Moving Average models are plotted above against the test data

• 2 point trailing moving average is found to be having the best fitment against the test data, through with a lag of 2 and falling short at times

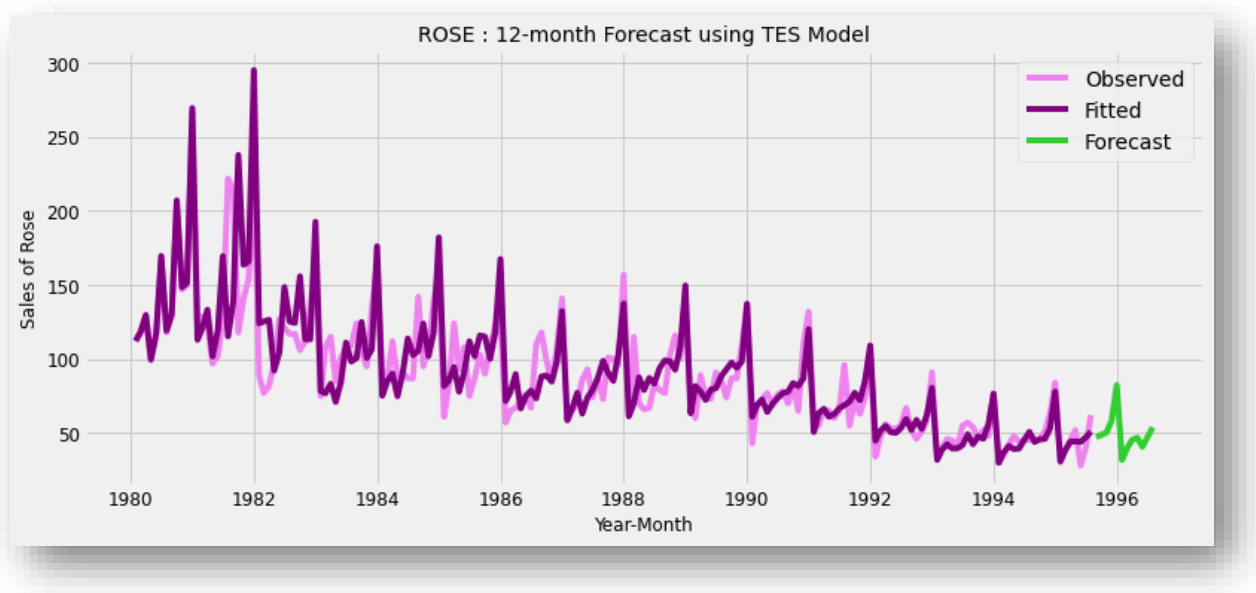
• Both SARIMA and TES forecasts are a bit higher than the actuals at any given point in time



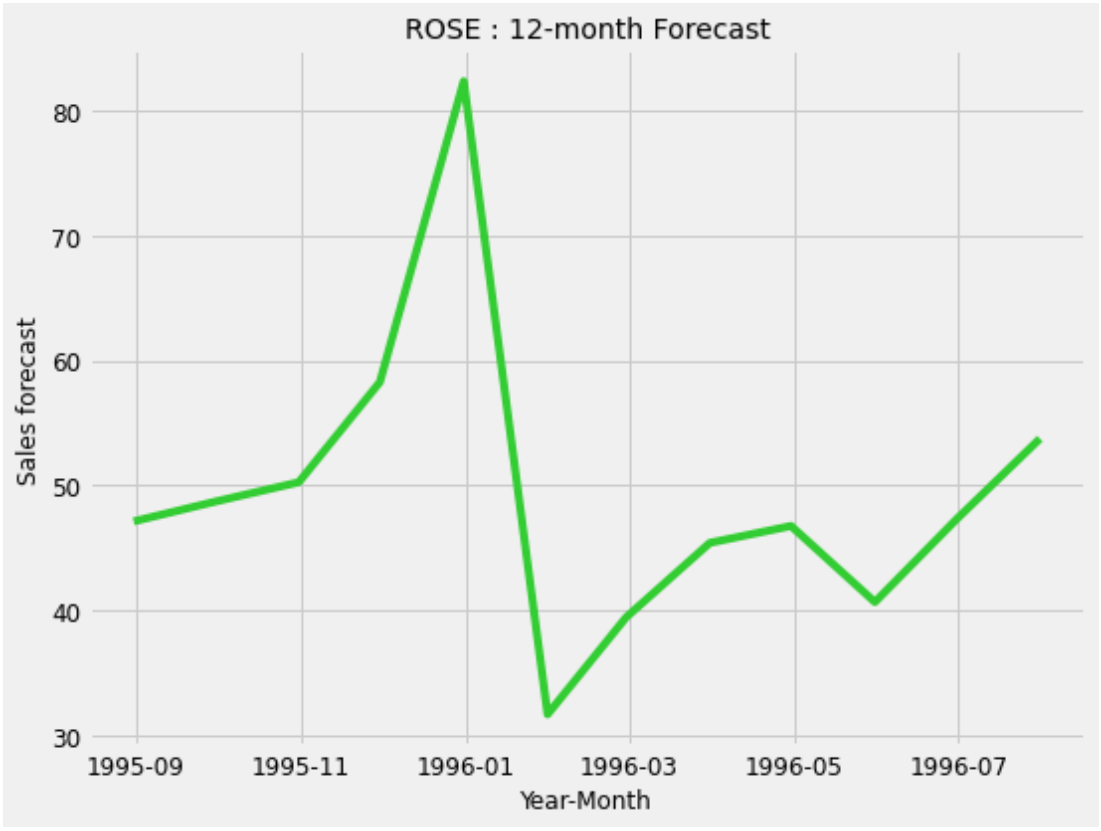
66. Rose - Forecast vs Test data, 3 best models

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

Build TES Model on Rose full data



67.Rose - 12-month forecast using TES model



68. Rose - 12-month forecast

Trying another SARIMA

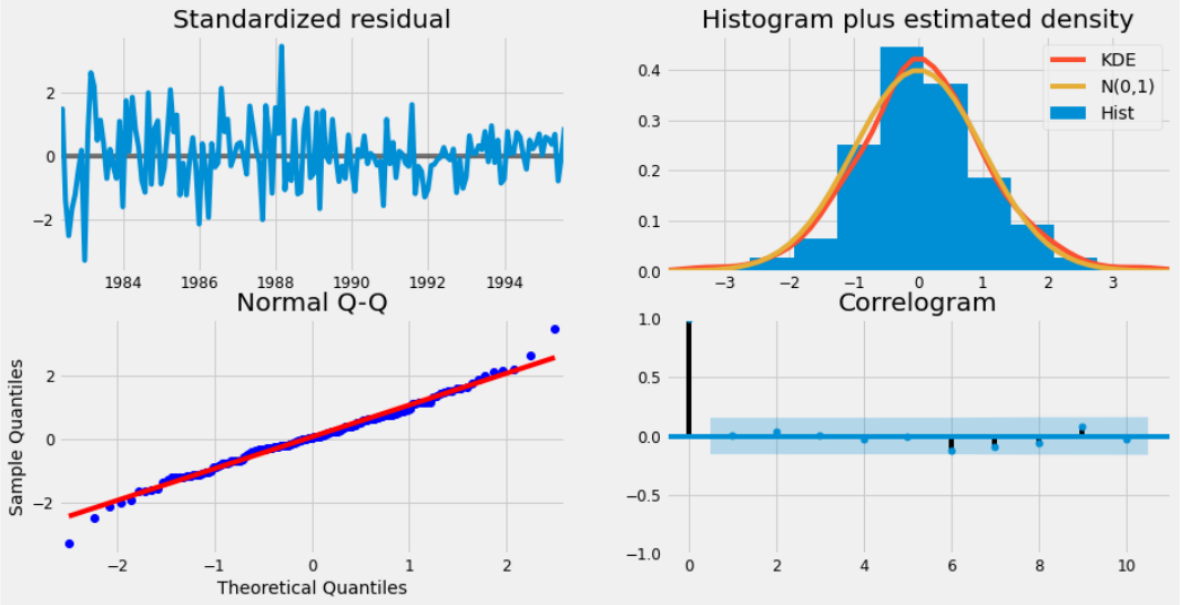
Statespace Model Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(4, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-664.135
Date:	Fri, 08 Oct 2021	AIC	1342.270
Time:	07:13:49	BIC	1363.796
Sample:	01-31-1980	HQIC	1351.011
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0914	0.084	1.093	0.274	-0.072	0.255
ar.L2	-0.1077	0.077	-1.393	0.164	-0.259	0.044
ar.L3	-0.1314	0.076	-1.729	0.084	-0.280	0.018
ar.L4	-0.1071	0.078	-1.375	0.169	-0.260	0.046
ma.L1	-0.8270	0.055	-14.901	0.000	-0.936	-0.718
ma.S.L12	-0.5963	0.059	-10.122	0.000	-0.712	-0.481
sigma2	232.4248	24.359	9.542	0.000	184.682	280.168

Ljung-Box (Q):	35.39	Jarque-Bera (JB):	5.30
Prob(Q):	0.68	Prob(JB):	0.07
Heteroskedasticity (H):	0.22	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.89

69. Rose full data SARIMA (4,1,1)x(0,1,1,12) model: Results



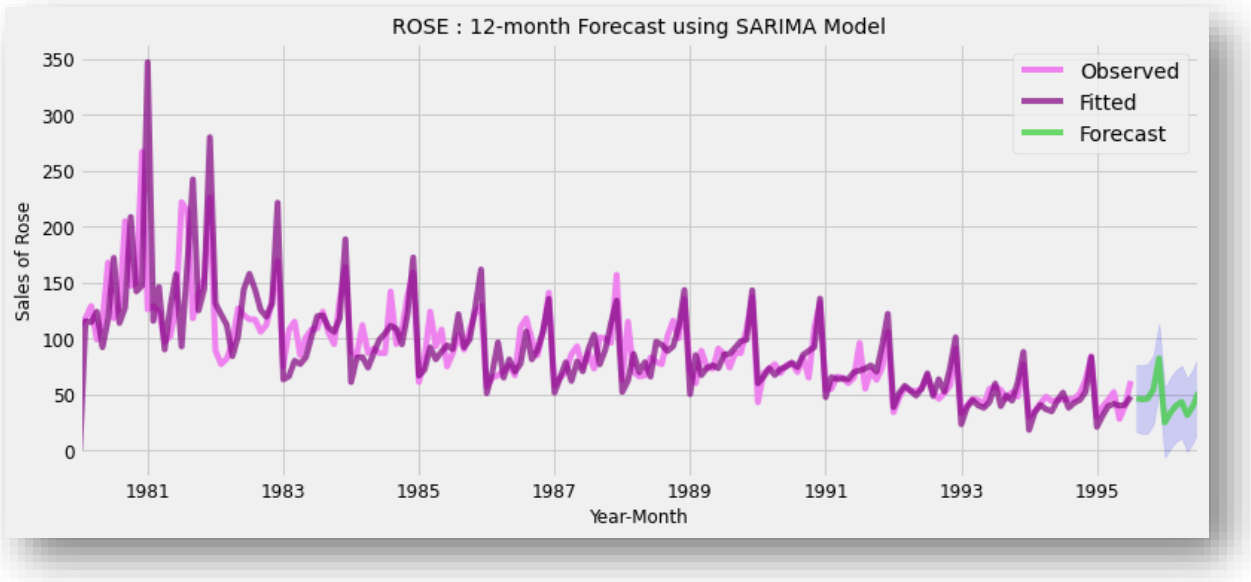
70. Rose full data SARIMA (4,1,1)x(0,1,1,12) model: Diagnostics

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	46.540808	15.245485	16.660206	76.421410
1995-09-30	45.514601	15.769107	14.607720	76.421483
1995-10-31	46.226234	15.827871	15.204177	77.248291
1995-11-30	54.319829	15.831175	23.291297	85.348362
1995-12-31	82.214863	15.835193	51.178456	113.251271

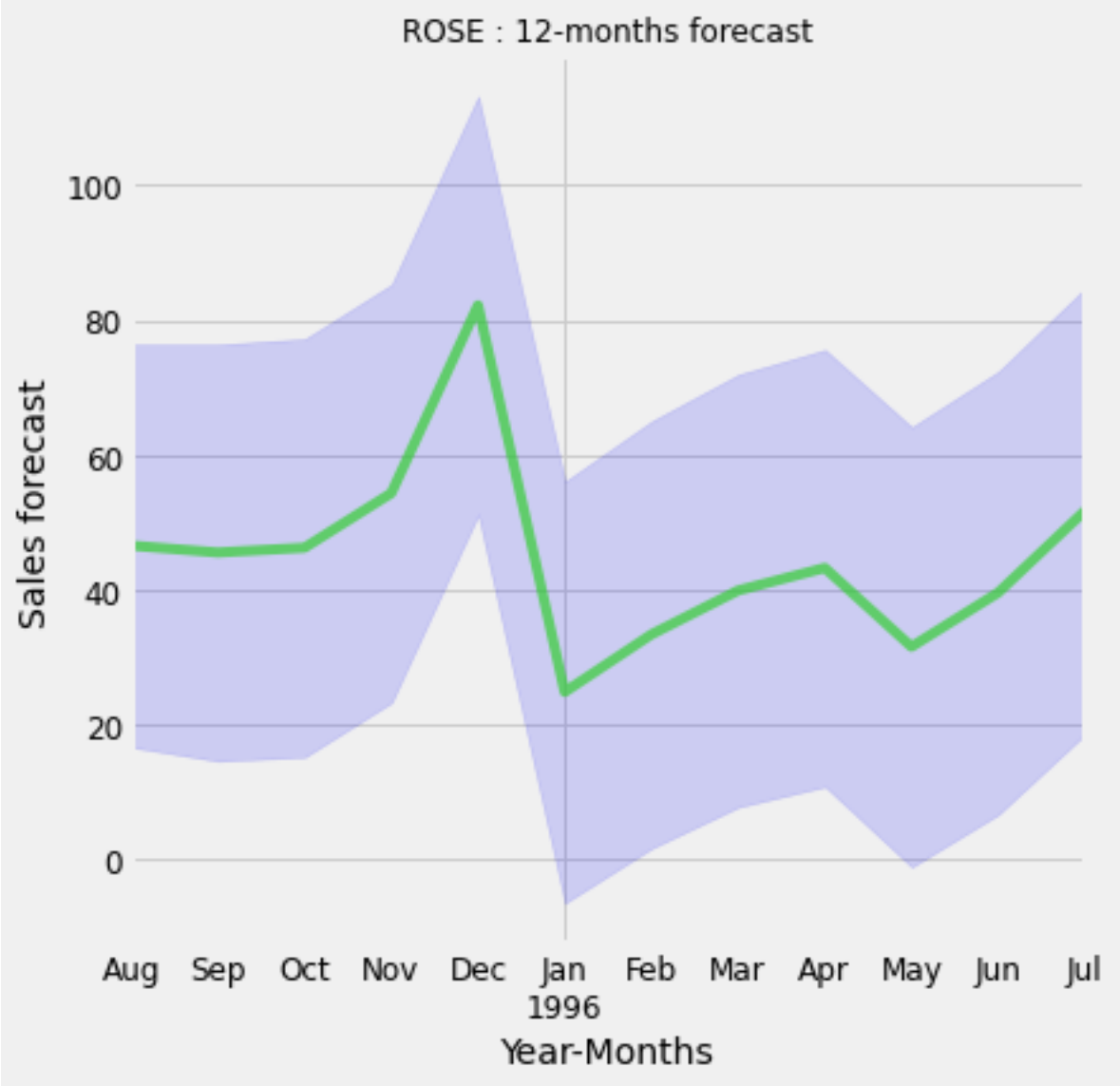
41. Sarimax model (4,1,1)x(0,1,1,12) on full data, prediction summary

```
Time_Stamp
1980-01-31      0.000000
1980-02-29    115.410430
1980-03-31    114.398803
1980-04-30    123.947737
1980-05-31     92.219741
dtype: float64
```

42. Sarimax model on full data, fitted values



71. Rose - 12 months forecast using SARIMA model



72. Rose - 12-month forecast with confidence interval

Predict future

Based on the overall model evaluation and comparison, Triple Exponential Smoothing (Holt Winter's) and SARIMA are selected for final prediction into 12 months in future

• TES model alpha: 0.1, beta: 0.2 and gamma: 0.2 & trend: 'additive', seasonal: 'multiplicative' is found to be the best model in terms of accuracy scored against the full data

• The model predicts continuation of the trend in sales and seasonality in year end sales. The prediction shows a stabilization of downward trend, as the sales will be almost same as previous observed year • The 12 month prediction of the TES model is as below

• The SARIMA model is built with parameters (4, 1, 1)x(0, 1, 1, 12), is found to be the most optimal SARIMA model for the complete time-series

• SARIMA model has also reflected the trend and seasonality of the series continuing into the future year as well.

• SARIMA model is seen to have better fitment with the most recent observed data and shows high variations in the farthest periods of observations, which explains the high RMSE and MAPE values

• The RMSE and MAPE values of the two models are as below

Model evaluation		
	RMSE	MAPE
TES Forecast	20.881	14.48
SARIMA Forecast	30.676	19.4

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

Final model

The SARIMA model is chosen as the final model for prediction on Rose dataset, as it provide confidence interval and better explainability of the model

• The diagnostics plot of the model shows that the residuals follow a normal distribution with most values around mean zero. The residuals also follow a straight line in normal QQ plot

• The model summary also provides valuable insights in the model. From the snapshot of summary below it can be understood that MA(1) and seasonal MA(1) term has the highest weightage. The p-values indicates that the terms MA(1) and Seasonal MA(1) are the most significant terms

• The rest of the p-values got values higher than alpha 0.05, which fails to reject the hull hypothesis that these terms are not significant

• Prediction on the Rose time-series is on a wider confidence band than Sparkling

Recommendations for the company

The model forecasts sale of 539 units of Rose wine in 12 months into future. Which is an average sale of 45 units per month

• The seasonal sale in December 1995 will reach a maximum of 82 units, before it drops to the lowest sale in January 1996; at 25 units.

• Unlike Sparkling wine, Rose wine sells very low number of units and the standard deviation is only 14.5. Which means that higher demand does not impact procurement and production

• Apart from higher sale in November and December months, Rose sales will be above average in the summer months of July and August

• The winery should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions

	ROSE
1995-08-31	46.54
1995-09-30	45.51
1995-10-31	46.23
1995-11-30	54.32
1995-12-31	82.21
1996-01-31	24.81
1996-02-29	33.35
1996-03-31	39.87
1996-04-30	43.23
1996-05-31	31.53
1996-06-30	39.56
1996-07-31	51.70

43. Forecast for August 1995 to July 1996

	ROSE
count	12.000000
mean	44.905000
std	14.473222
min	24.810000
25%	38.007500
50%	44.370000
75%	47.830000
max	82.210000

44. Forecast summary

```
ROSE    538.86
dtype: float64
```

Total sales forecast

BIBLIOGRAPHY

https://towardsdatascience.com/3-facts-about-time-series-forecasting-that-surprise-experienced-machine-learning-practitioners-69c18ee89387

https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q

https://people.duke.edu/~rnau/seasarim.htm

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.interpolate.html

Great Learning courseware and notes of mentor Milind Desai

Thank You