Time Series Forecasting-Sparkling

25/02/2024

BALAJI S

PGP-DSBA

Module 8 - Time series

Problem:

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

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

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

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

Fig1.Heads & Tails of the Rose Dataset

- *There are 187 rows and 1 column*

Plot:

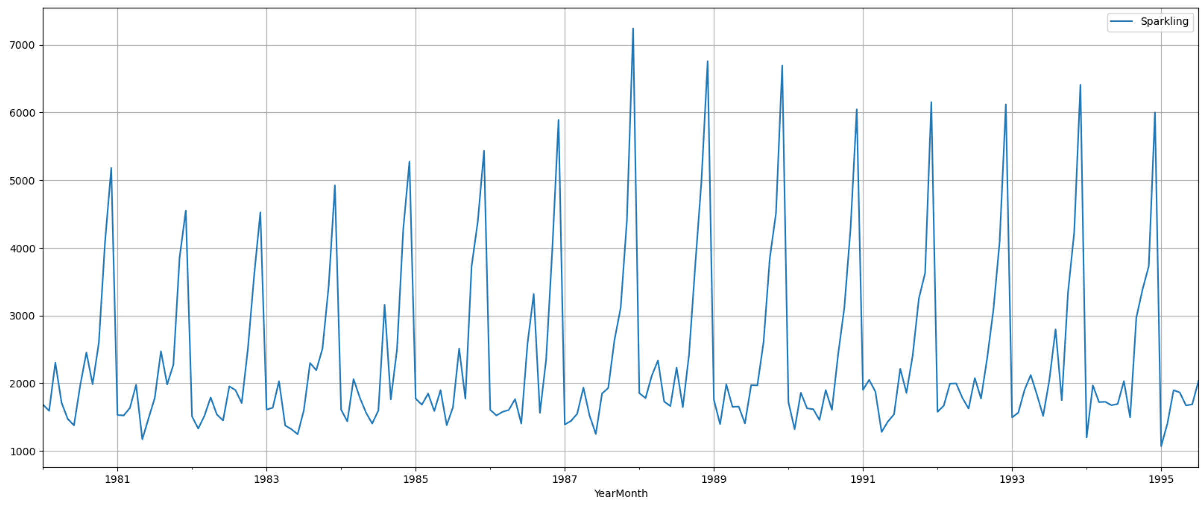


Fig 2: plot of the dataset

We have divided the dataset further by extraction month and year columns from the YearMonth column and renamed the sparkling column name to Sales for better analysis of the dataset. The new dataset has 187 rows and 3 columns.

YearMonth	Sparkling	Year	Month
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5

- Additionally, we renamed the 'Sparkling' column to 'Sales', a decision made to foster a clearer understanding and streamlined analysis of the dataset.

YearMonth	Sales	Year	Month
1980-01-01	1686	1980	1
1980-02-01	1591	1980	2
1980-03-01	2304	1980	3
1980-04-01	1712	1980	4
1980-05-01	1471	1980	5

YearMonth	Sales	Year	Month
1995-03-01	1897	1995	3
1995-04-01	1862	1995	4
1995-05-01	1670	1995	5
1995-06-01	1688	1995	6
1995-07-01	2031	1995	7

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

Data Type

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sales   187 non-null    int64
 1   Year    187 non-null    int64
 2   Month   187 non-null    int64
dtypes: int64(3)
memory usage: 5.8 KB
```

Statistical Summary:

	count	mean	std	min	25%	50%	75%	max
Sales	187.0	2402.0	1295.0	1070.0	1605.0	1874.0	2549.0	7242.0
Year	187.0	1987.0	5.0	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.0	3.0	1.0	3.0	6.0	9.0	12.0

Null values:

There are no null values present in the dataset. So we can do further analysis smoothly.

```
Sales     0
Year      0
Month     0
dtype: int64
```

Boxplot of the dataset:

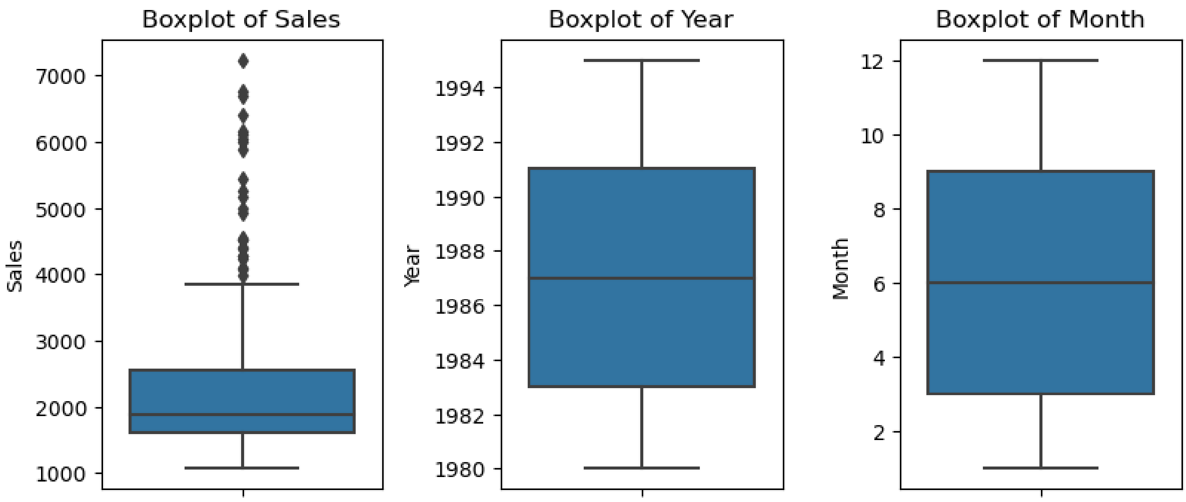
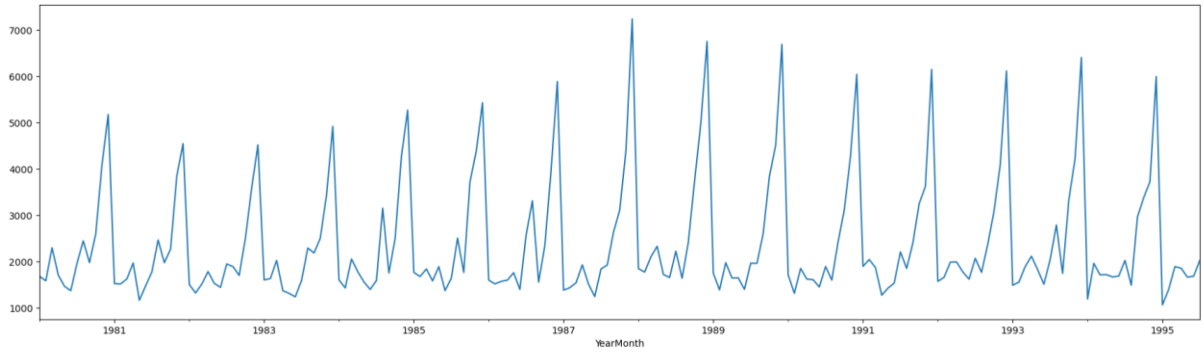


Fig 3: Boxplot

The box plot shows:

●Sales boxplot has outliers we can treat them but we are choosing not to treat them as they do not have much effect on the time series model.

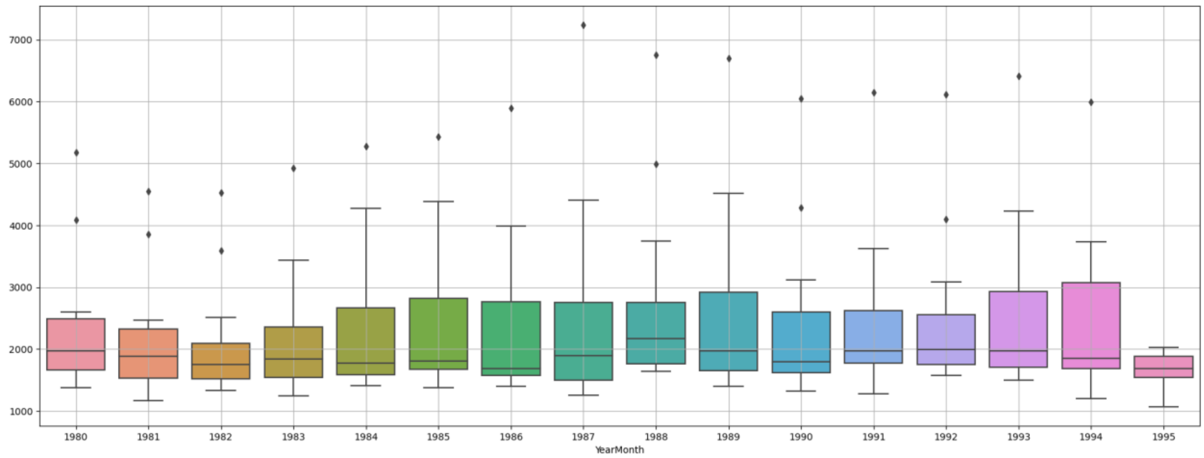
Line plot of sales:



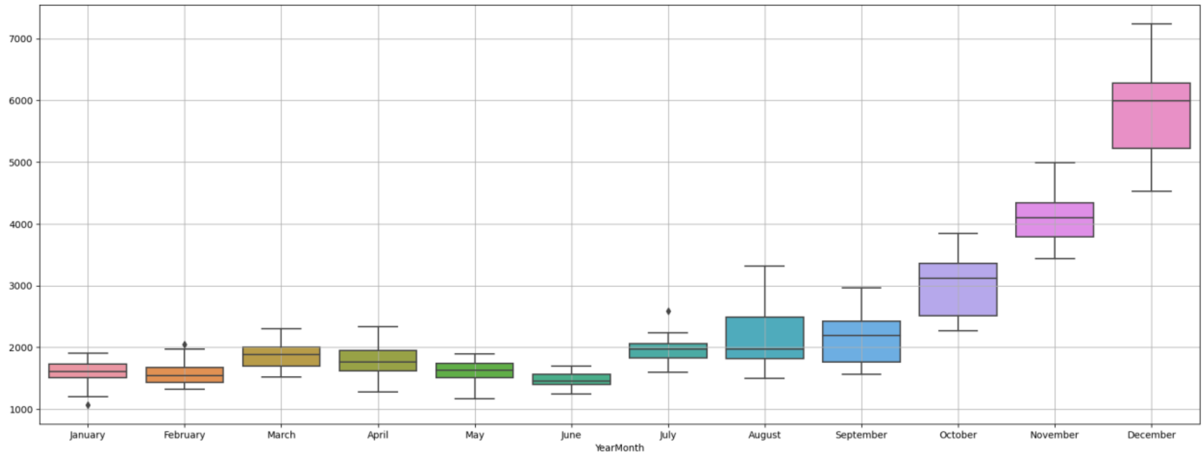
- *The line plot shows the patterns of trend and seasonality and also shows that there was a peak in the year 1988.*

Boxplot

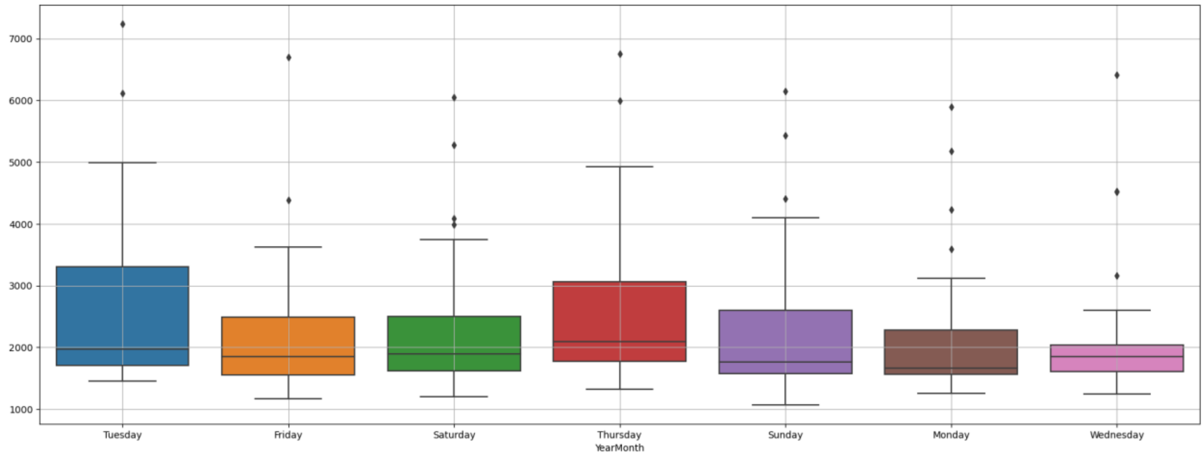
yearly



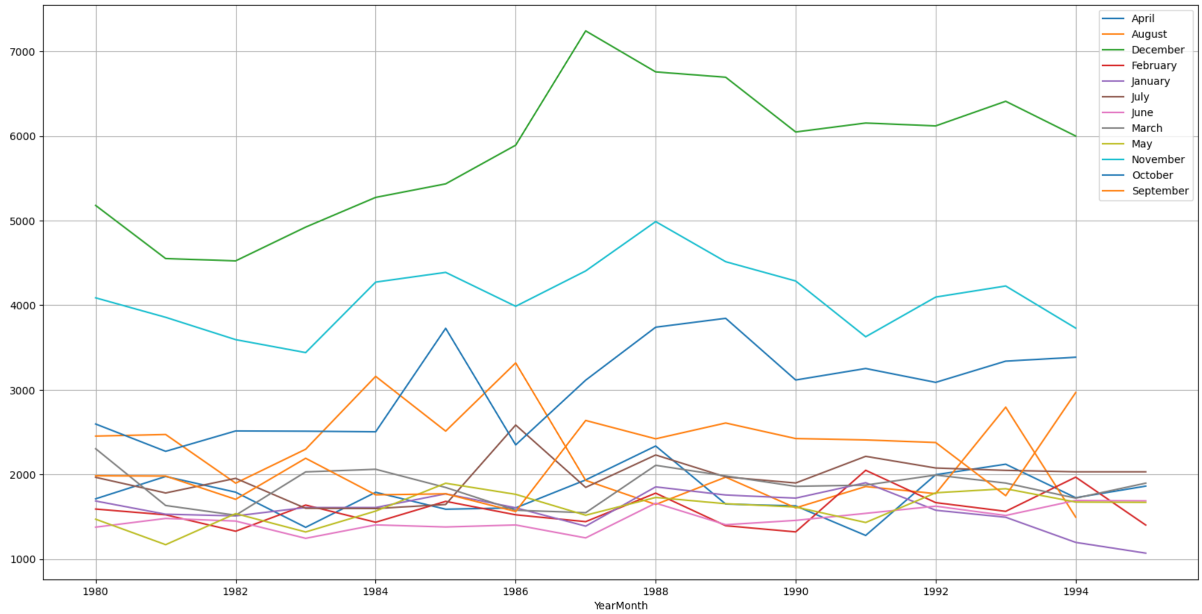
Monthly



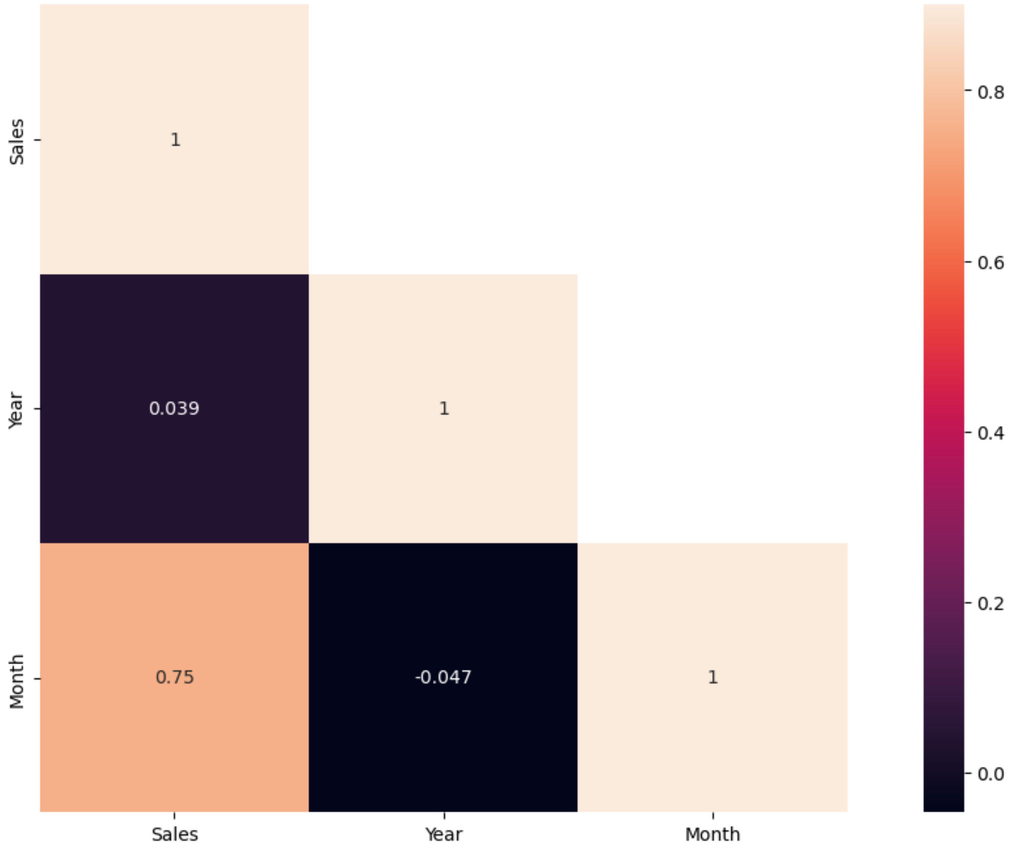
Weekly



Graph of Monthly sales across years

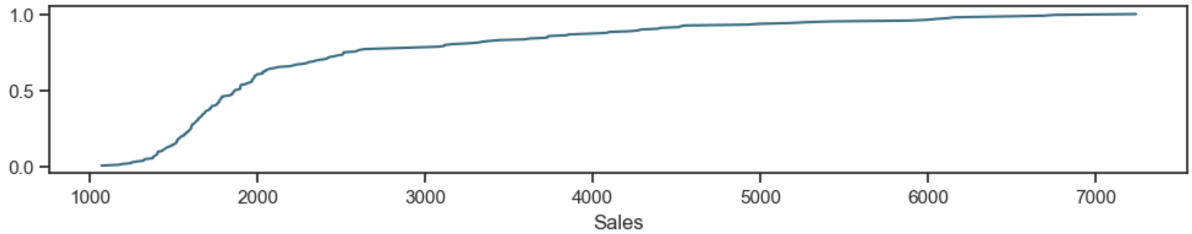


CORRELATION:



This heat map shows that there is a low correlation between sales and year. there is a more correlation between month and sales. It indicated seasonal patterns in sales

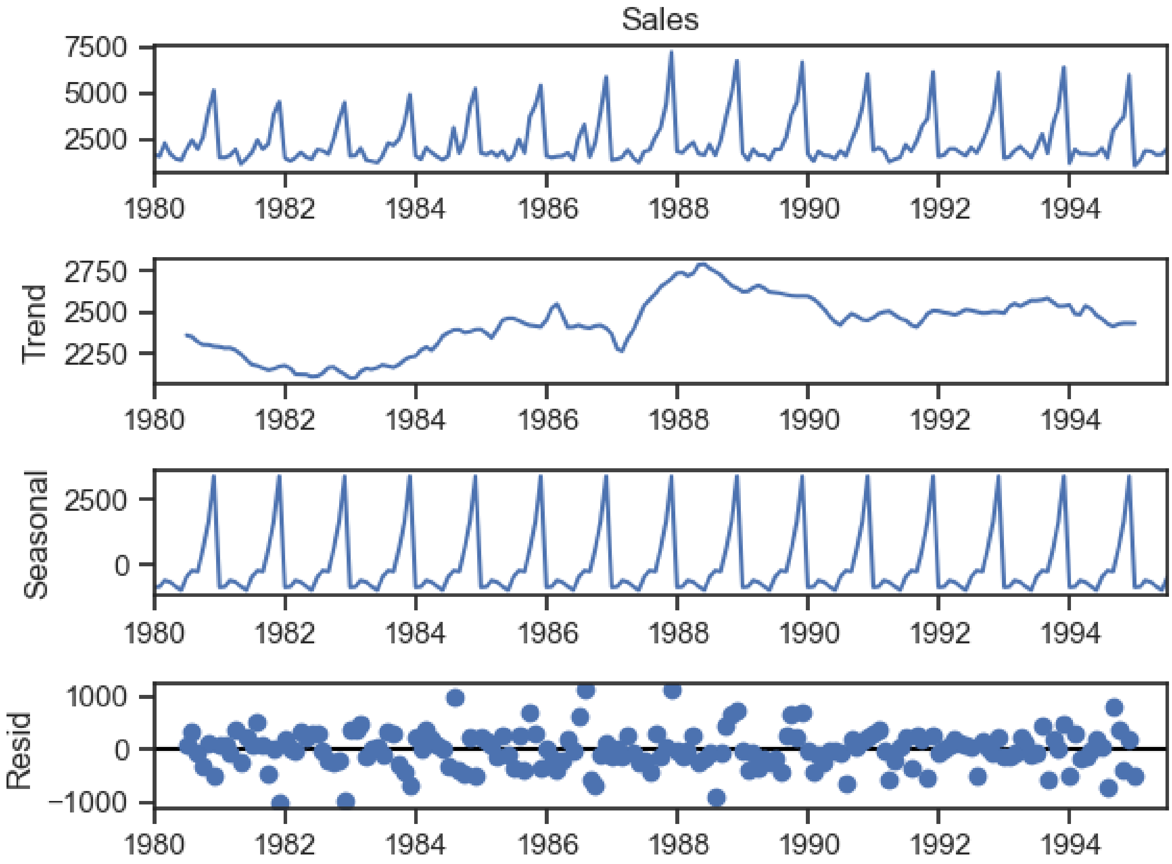
Plot ECDF: Empirical Cumulative Distribution Function.



This plot shows:

- More than 50% of sales have been less than 2000
- Highest values is 7000
- Approx 80% of sales have been less than 3000

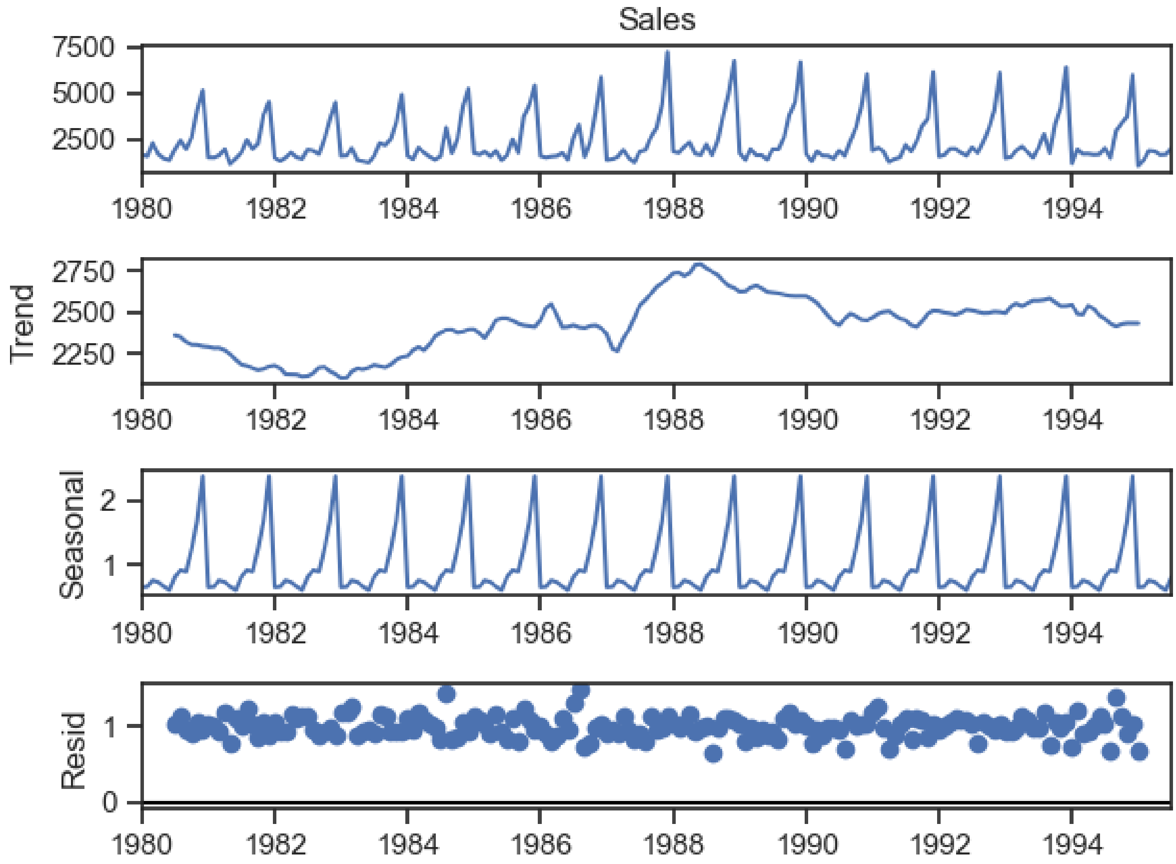
Decomposition -addictive



The plots show:

- Peak year 1988-1989
- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is not in a straight line.
- Both trend and seasonality are present.

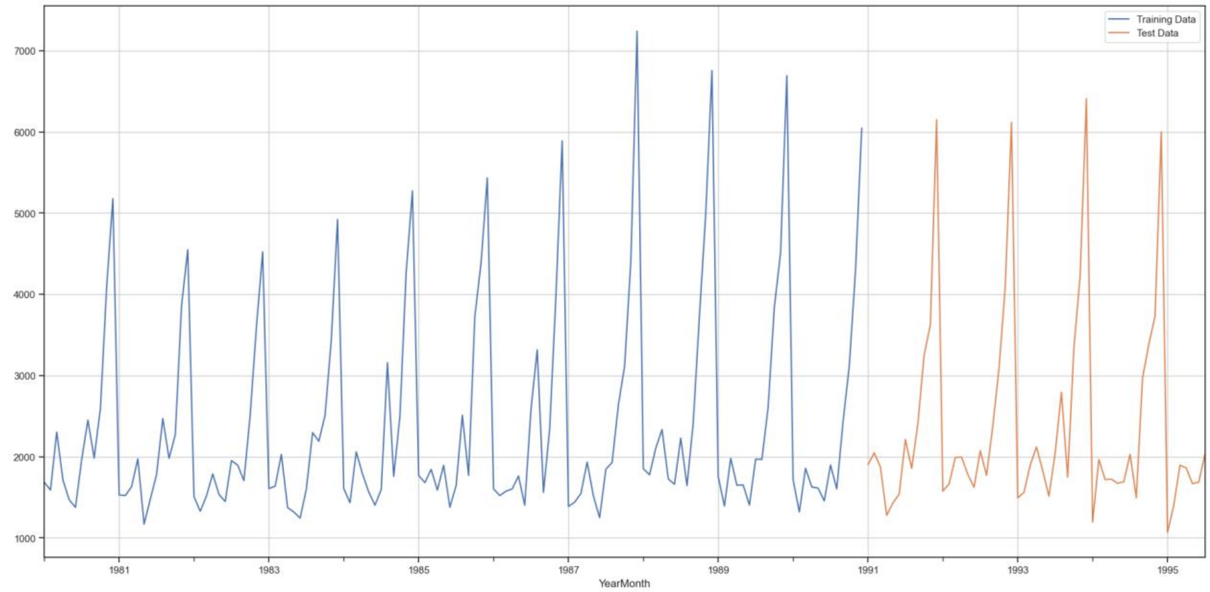
Decomposition -multiplicative



The plots show

- Peak year 1988-1989
- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is in approx. a straight line.
- Both trend and seasonality are present.
- Reside is 0 to 1, while additive is 0 to 1000.
- So multiplicative model is selected owing to a more stable residual plot and lower range of residuals.

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



As per the instructions given in the project we have split the data, around 1991. With training data from 1980 to 1990 December. Test data starts from the first month of January 1991 till the end.

Rows and Columns:

- The train dataset has 132 rows and 3 columns.
- The test dataset has 55 and 3 columns.

```
Rows of dataset:
First few rows of Training Data
            Sales  Year  Month
YearMonth
1980-01-01   1686  1980      1
1980-02-01   1591  1980      2
1980-03-01   2304  1980      3
1980-04-01   1712  1980      4
1980-05-01   1471  1980      5

Last few rows of Training Data
            Sales  Year  Month
YearMonth
1990-08-01   1605  1990      8
1990-09-01   2424  1990      9
1990-10-01   3116  1990     10
1990-11-01   4286  1990     11
1990-12-01   6047  1990     12
```

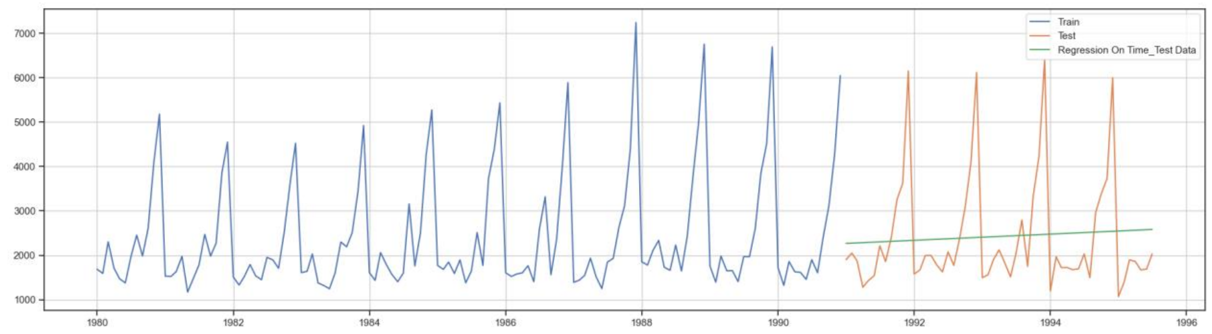
```
First few rows of Test Data
            Sales  Year  Month
YearMonth
1991-01-01   1902  1991      1
1991-02-01   2049  1991      2
1991-03-01   1874  1991      3
1991-04-01   1279  1991      4
1991-05-01   1432  1991      5

Last few rows of Test Data
            Sales  Year  Month
YearMonth
1995-03-01   1897  1995      3
1995-04-01   1862  1995      4
1995-05-01   1670  1995      5
1995-06-01   1688  1995      6
1995-07-01   2031  1995      7
```

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

- Model 1:Linear Regression
- Model 2: Naive Approach
- Model 3: Simple Average
- Model 4: Moving Average(MA)
- Model 5: Simple Exponential Smoothing
- Model 6: Double Exponential Smoothing (Holt's Model)
- Model 7: Triple Exponential Smoothing (Holt - Winter's Model)

LINEAR REGRESSION

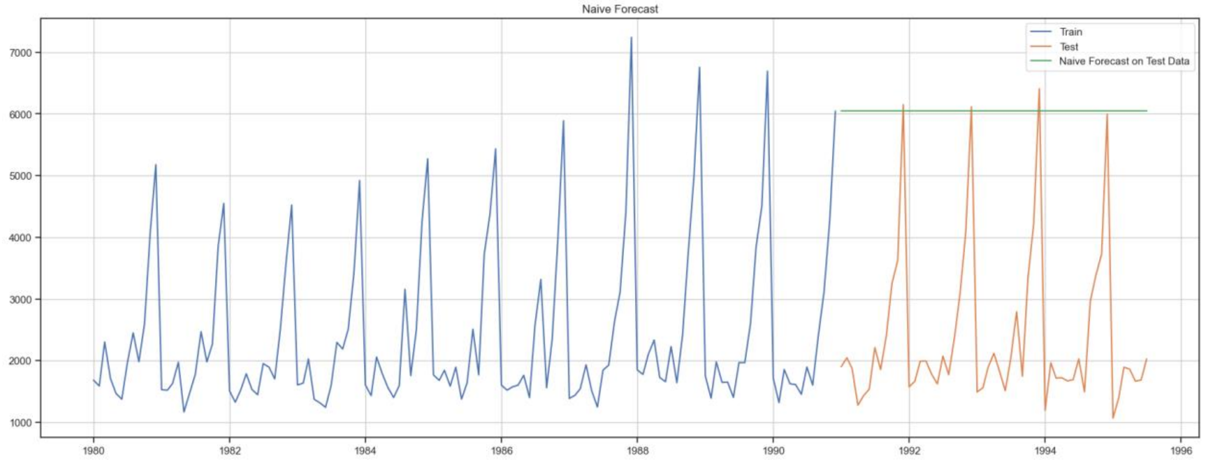


The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

	Test RMSE
Linear Regression	1275.867052

NAÏVE APPROACH

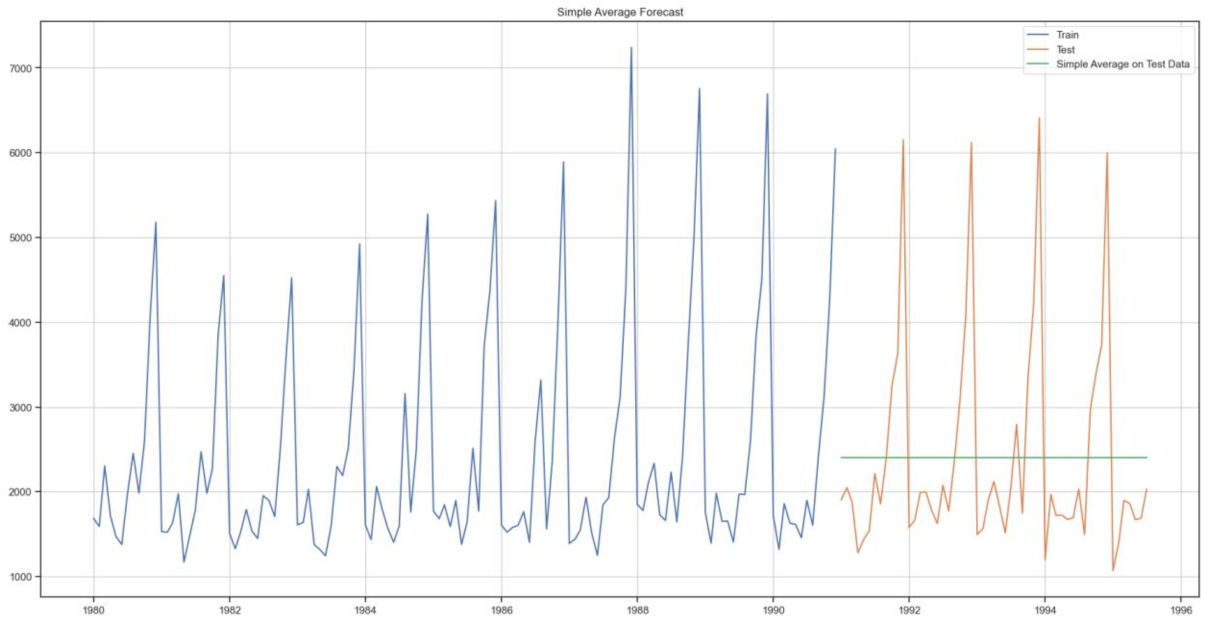


The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

Naive Model	3864.279352

SIMPLE AVERAGE:

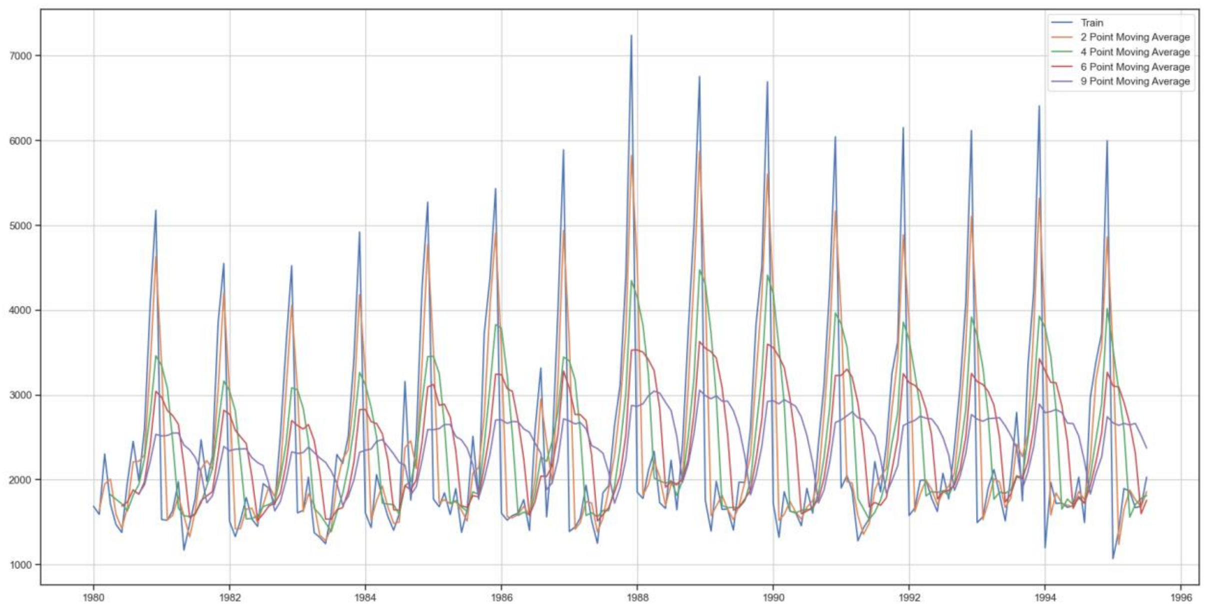


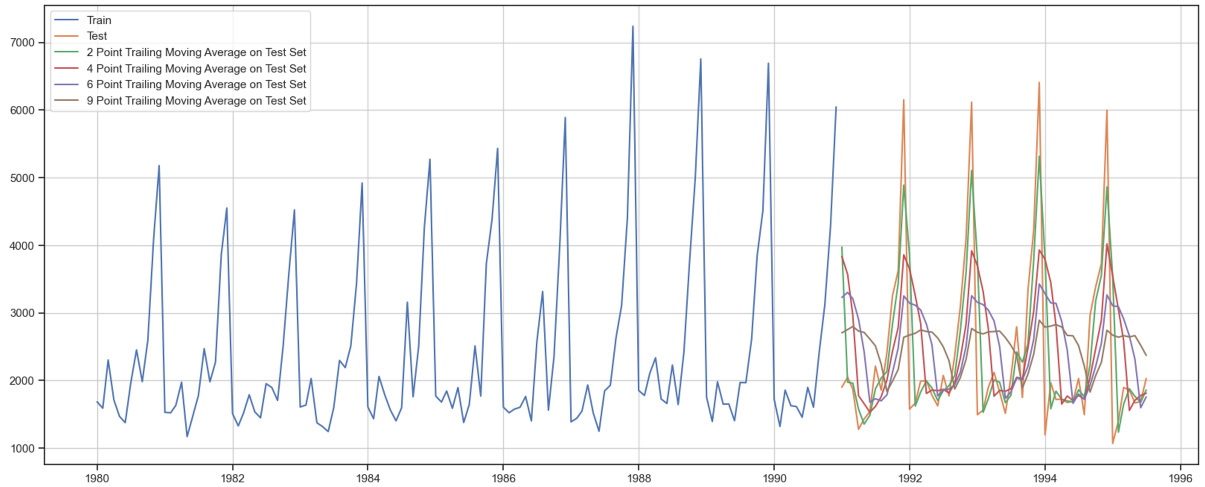
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

Simple Average Model	1275.081804

MOVING AVERAGE:



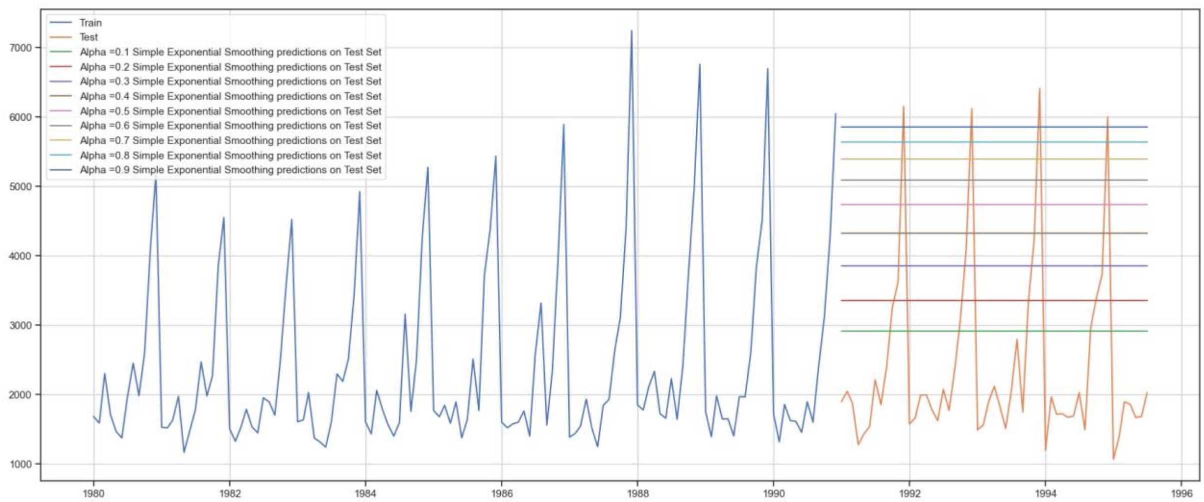


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

- We have made multiple moving average models with rolling windows varying from 2 to 9.
- Rolling average is a better method than simple average as it takes into account only the previous n values to make the prediction, where n is the rolling window defined.
- This takes into account the recent trends and is in general more accurate.
- The higher the rolling window, the smoother will be its curve, since more values are being taken into account.

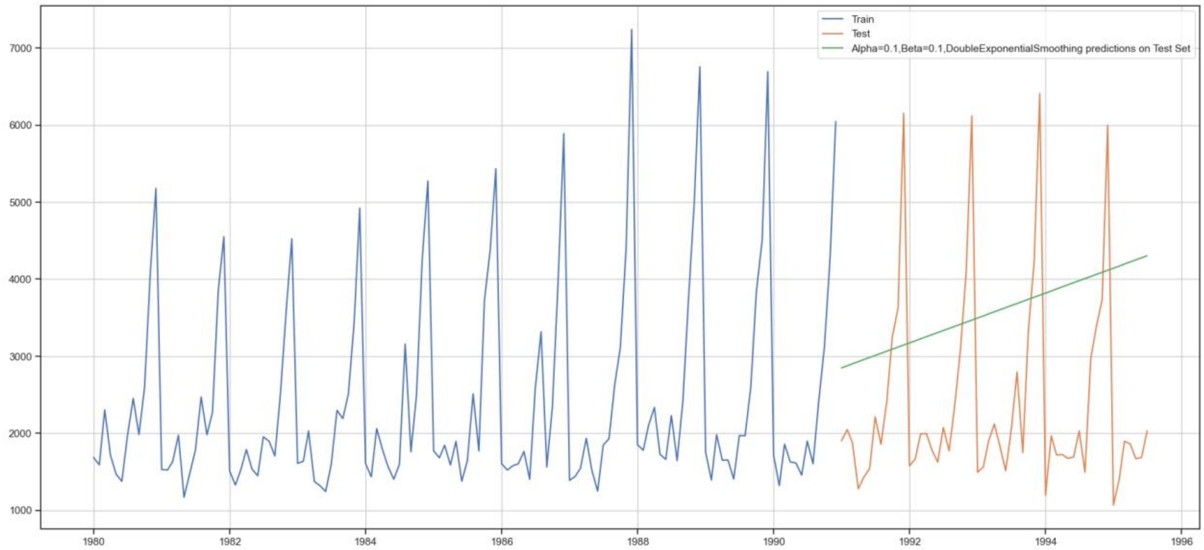
SIMPLE EXPONENTIAL SMOOTHING:



The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.1,SimpleExponentialSmoothing	1375.393398

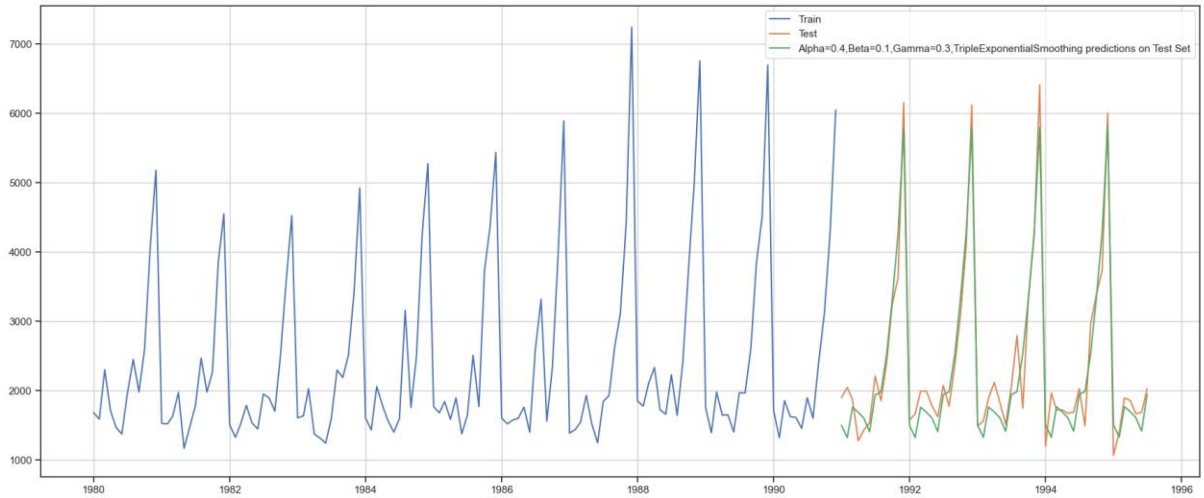
Double Exponential smoothing(Holt’s model)



The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing 1778.564670

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



Output for best alpha, beta, and gamma values is shown by the green color line in the
above plot. The best model had both multiplicative trend as well as seasonality. So far this is the best model

The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

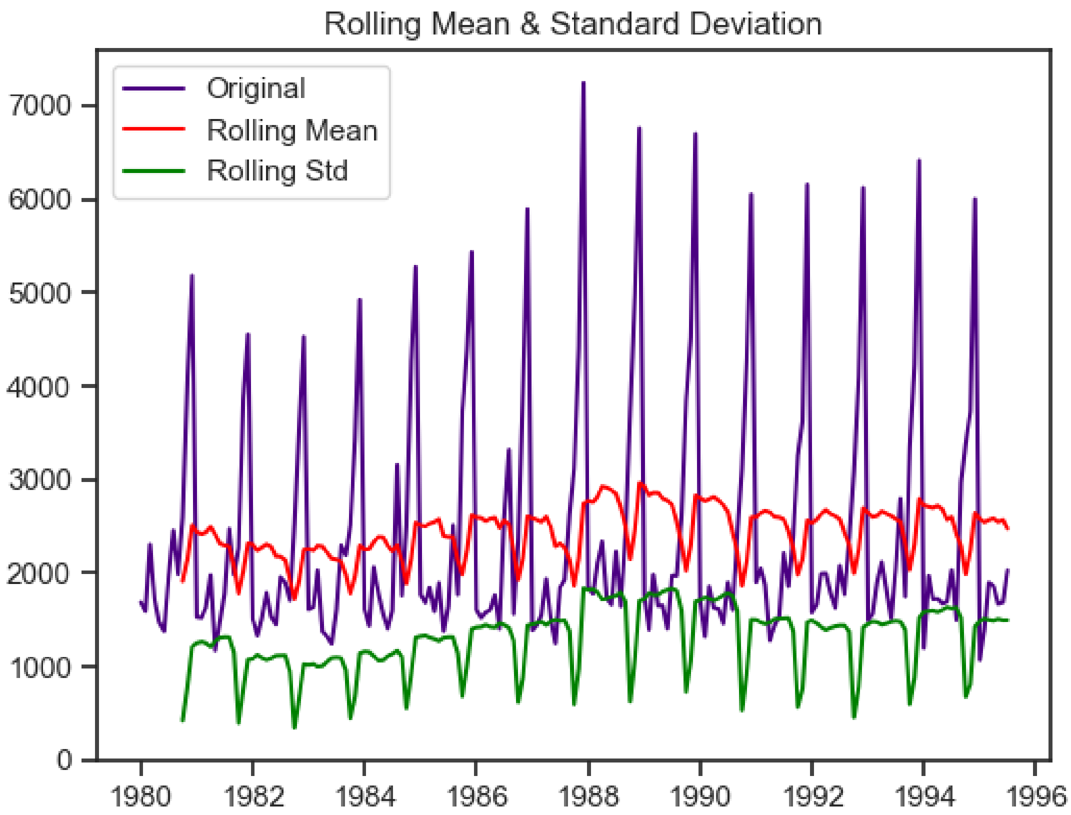
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302

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.

Check for stationarity of the whole Time Series data.
The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root
and subsequently whether the series is non-stationary.
The hypothesis in a simple form for the ADF test is:

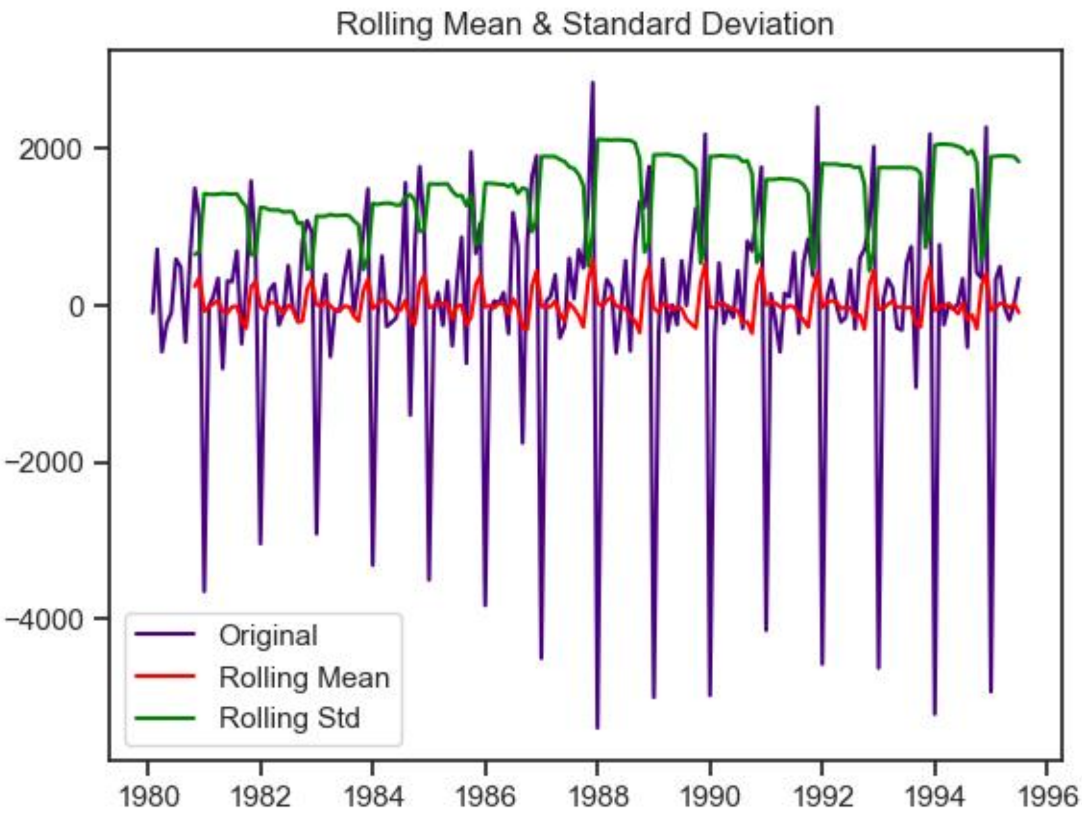
- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the
p-value of this test to be less than the α value.
We see that at 5% significant level the Time Series is non-stationary.




```
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
```

- In order to try and make the series stationary we used the differencing approach.
- We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped




```
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
```

- Dickey-Fuller test was 0.000, which is less than 0.05. Hence the null hypothesis that the series is not stationary at difference = 1 was rejected, which implied that the series has indeed become stationary after we performed the differencing.
- The null hypothesis was rejected since the p-value was less than alpha i.e. 0.05.
- Also, the rolling mean plot was a straight line this time around. Also, the series looked more or less the same from both directions, indicating stationarity

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

AUTO - ARIMA model

We employed a for loop for determining the optimum values of p,d,q, where p is the order of the AR (Auto-Regressive) part of the model, while q is the order of the MA (Moving Average) part of the model. d is the differencing that is required to make the series stationary. p,q values in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for d, since we had already determined d to be 1, while checking for stationarity using the ADF test.

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

Akaike information criterion (AIC) value was evaluated for each of these models and the model with the least AIC value was selected

	param	AIC
10	(2, 1, 2)	2213.509217
15	(3, 1, 3)	2221.456643
14	(3, 1, 2)	2230.792548
11	(2, 1, 3)	2232.982762
9	(2, 1, 1)	2233.777626
3	(0, 1, 3)	2233.994858
2	(0, 1, 2)	2234.408323
6	(1, 1, 2)	2234.527200
13	(3, 1, 1)	2235.498987
7	(1, 1, 3)	2235.607810
5	(1, 1, 1)	2235.755095
12	(3, 1, 0)	2257.723379
8	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
4	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

The summary report for the ARIMA model with values (p=2,d=1,q=3).

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 25 Feb 2024   AIC                           2213.509
Time:                        15:28:19   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.786      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.731      0.000      -0.701      -0.417
ma.L1         -1.9916      0.110    -18.184      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.093      0.000       0.784       1.215
sigma2      1.099e+06      2e-07   5.49e+12      0.000      1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================
```

RMSE values are as below: `1299.9796919669707`

AUTO- SARIMA Model

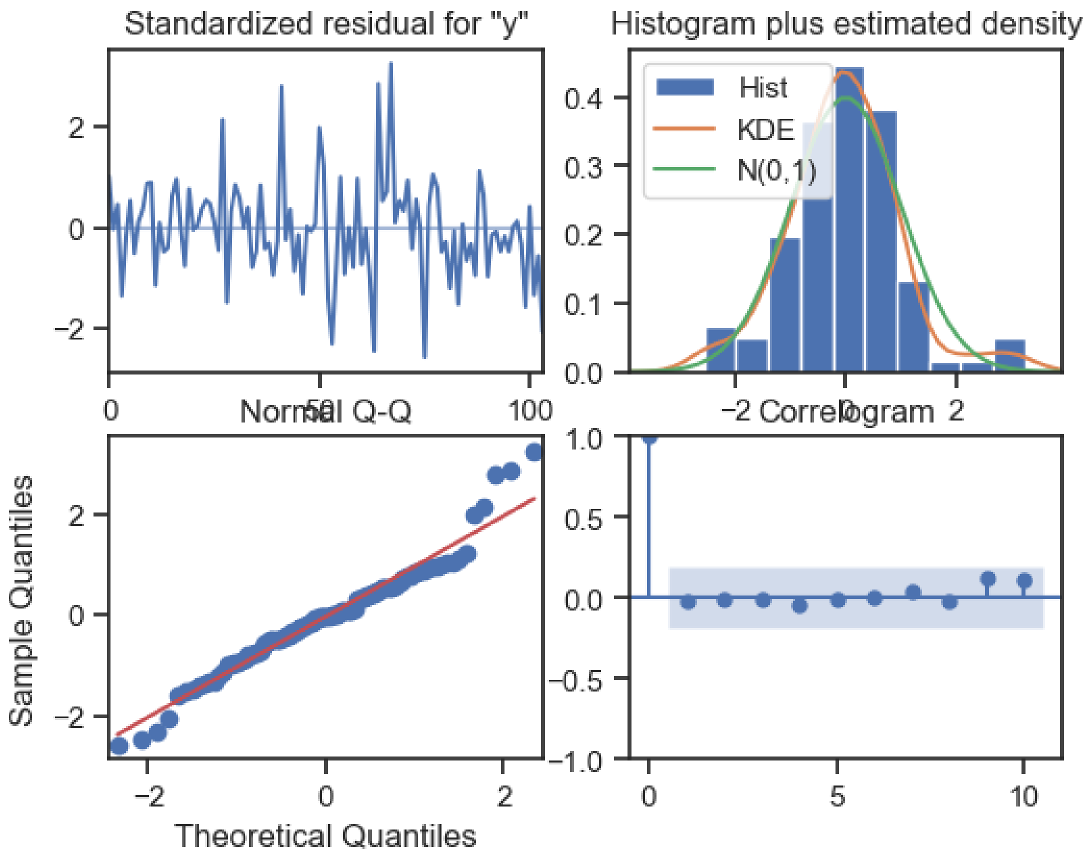
A similar for loop like AUTO_ARIMA with the below values was employed, resulting in the models shown below.

p = q = range(0, 4) d= range(0,2) D = range(0,2) pdq = list(itertools.product(p, d, q))
model_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, D, q))]

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

We also plotted the graphs for the residual to determine if any further information can be extracted or if all the usable information has already been extracted. Below are the plots
for the best auto SARIMA model.

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 25 Feb 2024   AIC                           1555.584
Time:                                    15:39:21   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6282      0.255     -2.463      0.014      -1.128      -0.128
ma.L1         -0.1041      0.225     -0.463      0.643      -0.545       0.337
ma.L2         -0.7276      0.154     -4.734      0.000      -1.029      -0.426
ar.S.L12       1.0439      0.014     72.841      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1355      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.400      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
-----------------------------------------------------------------------------------
```



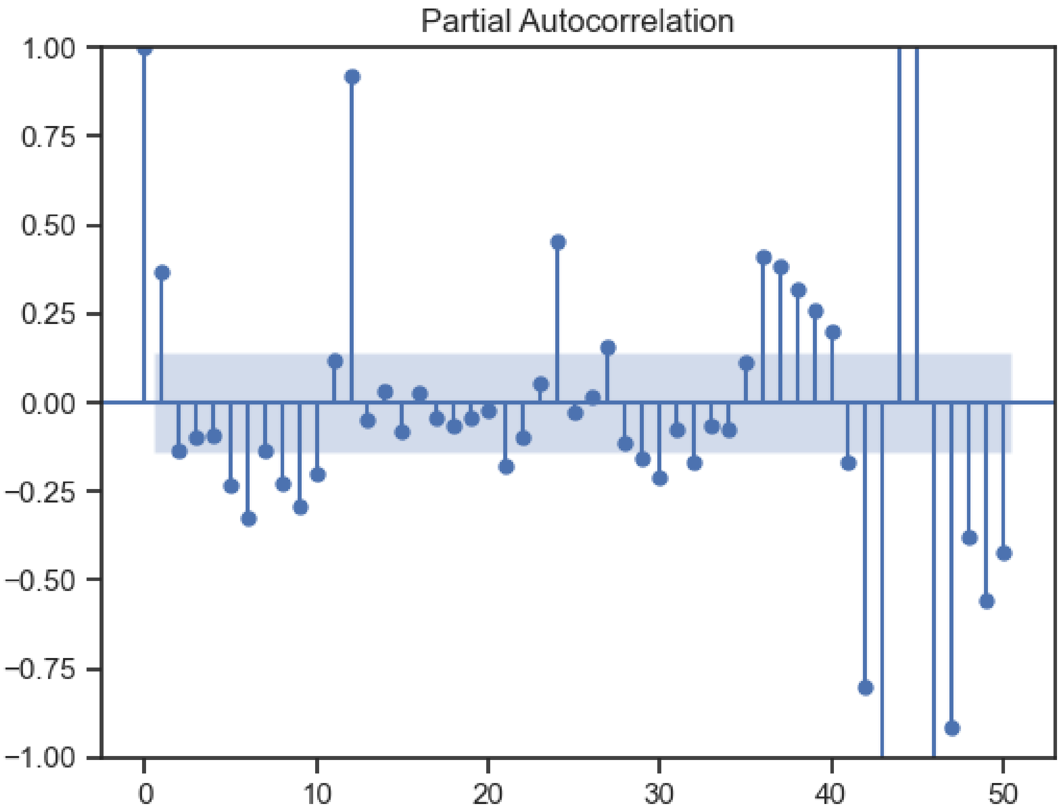
528.6029223625152

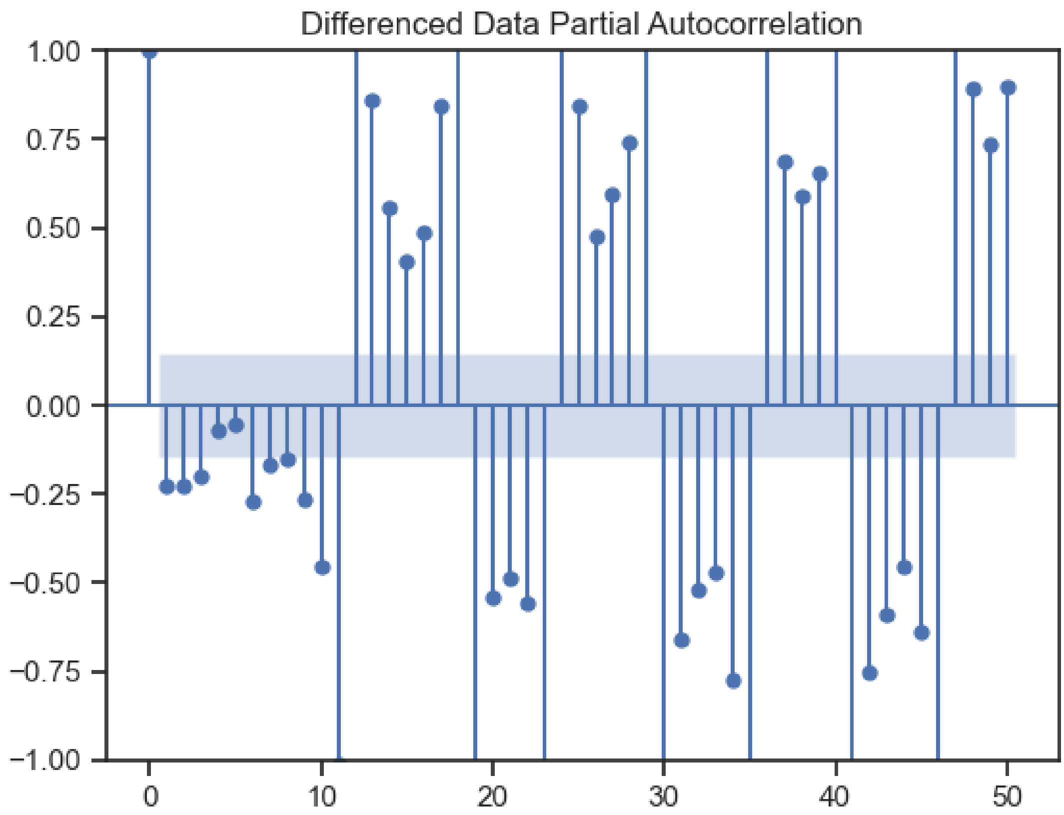
RSME of Model: ______________________

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

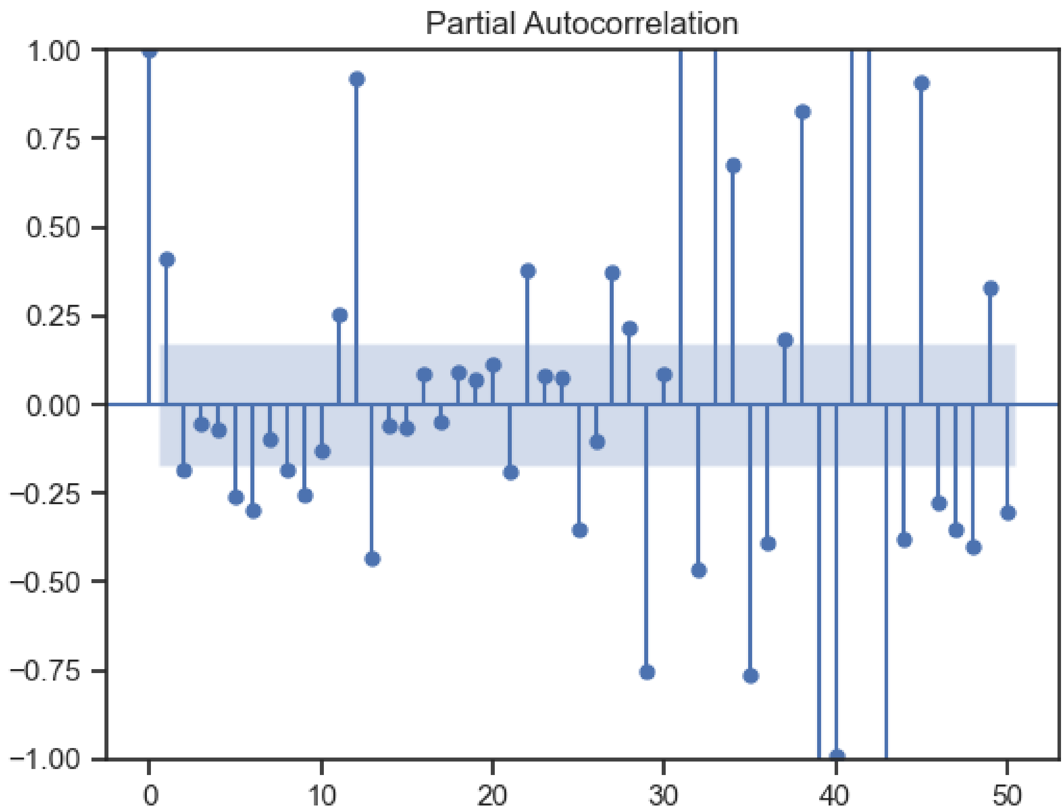
Manual- ARIMA Model

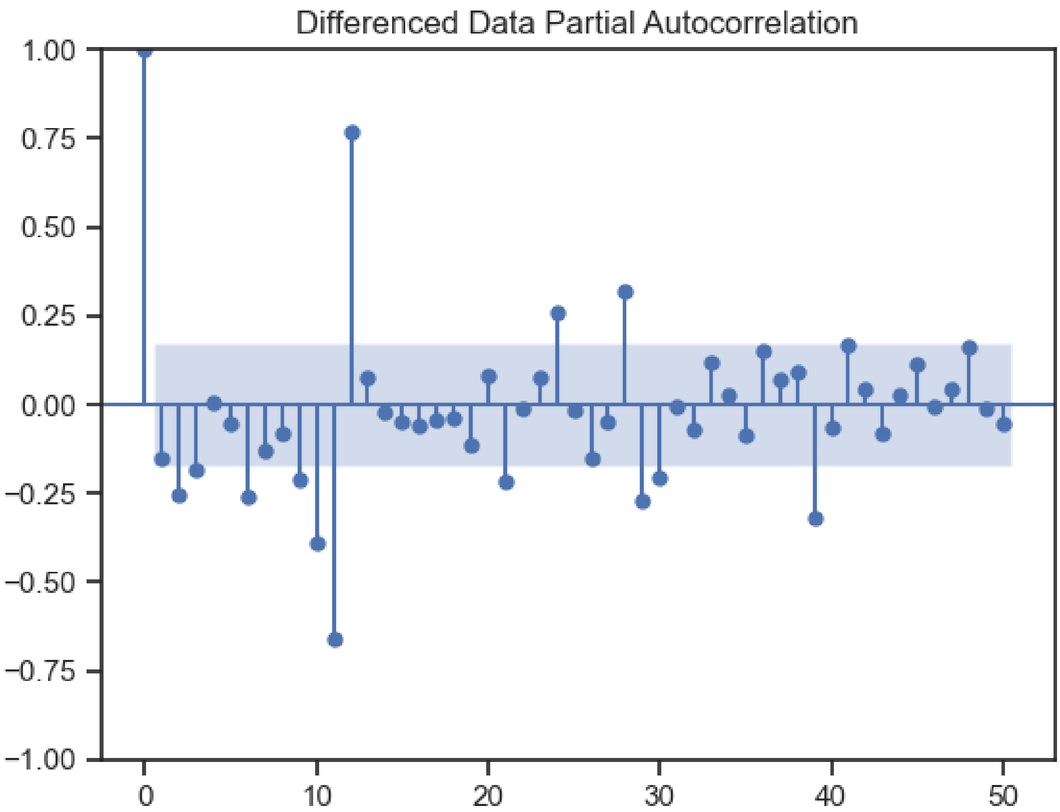
PACF the ACF plot on data :





PACF and ACF plot of train date:



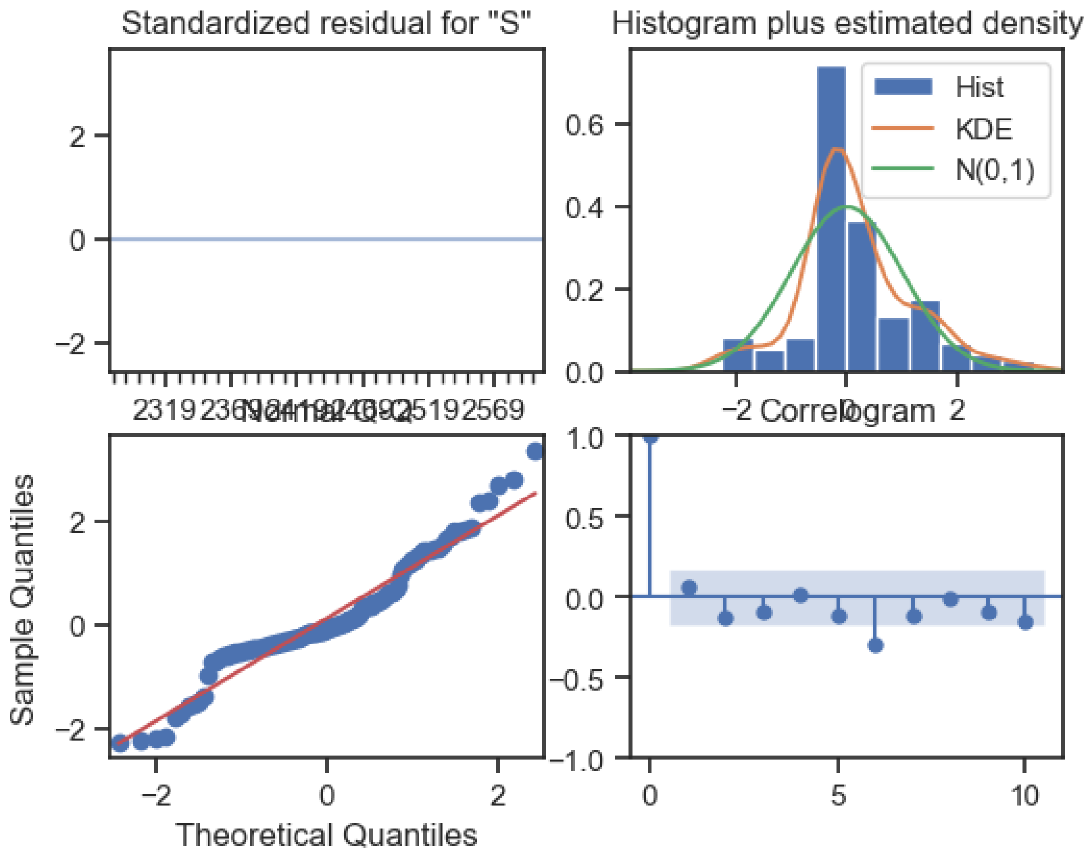


Hence the values selected for manual ARIMA:- p=2, d=1, q=2

summary from this manual ARIMA model:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -1114.878
Date:                Sun, 25 Feb 2024   AIC                           2235.755
Time:                        15:40:26   BIC                           2244.381
Sample:                    01-01-1980   HQIC                          2239.260
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4494      0.043     10.366      0.000       0.364       0.534
ma.L1         -0.9996      0.102     -9.811      0.000      -1.199      -0.800
sigma2      1.401e+06   7.57e-08   1.85e+13      0.000     1.4e+06     1.4e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.50   Jarque-Bera (JB):                10.42
Prob(Q):                              0.48   Prob(JB):                         0.01
Heteroskedasticity (H):               2.64   Skew:                             0.46
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.03
```

manual Arima model plots:



Model Evaluation: RSME: `1319.936733819979`

Manual SARIMA Model

Looking at the ACF and PACF plots for training data, we can clearly see significant spikes at lags 12,24,36,48 etc, indicating a seasonality of 12. The parameters used for manual

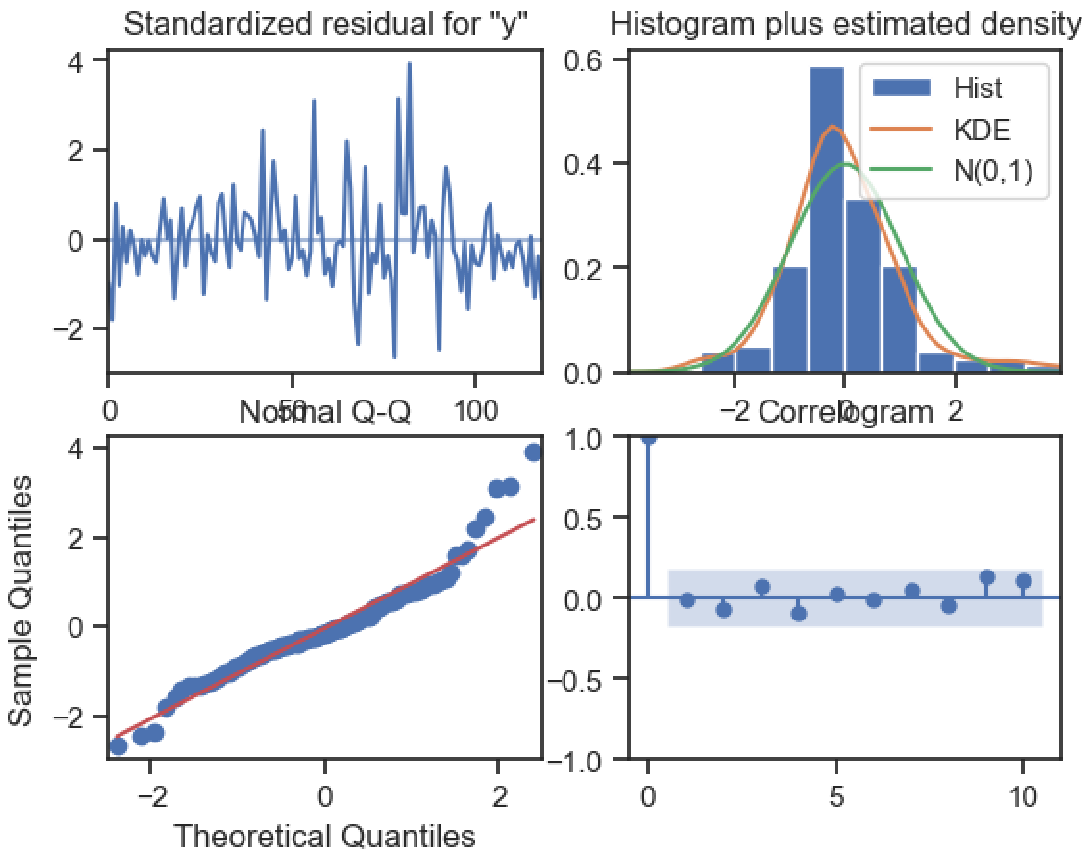
SARIMA model are as below.

SARIMAX(2, 1, 2)x(2, 1, 2, 12)

Below is the summary of the manual SARIMA model

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -538.016
Date:                            Sun, 25 Feb 2024   AIC                           1094.031
Time:                                    14:59:36   BIC                           1119.044
Sample:                                         0   HQIC                          1104.188
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5491      0.228     -2.408      0.016      -0.996      -0.102
ar.L2         -0.0744      0.099     -0.753      0.451      -0.268       0.119
ma.L1         -0.1703      0.216     -0.787      0.431      -0.594       0.254
ma.L2         -0.6694      0.228     -2.937      0.003      -1.116      -0.223
ar.S.L12      -1.0135      0.524     -1.935      0.053      -2.040       0.013
ar.S.L24      -0.1003      0.175     -0.572      0.567      -0.444       0.243
ma.S.L12       0.2906     20.998      0.014      0.989     -40.864      41.445
ma.S.L24      -0.7076     14.965     -0.047      0.962     -30.038      28.623
sigma2       430.5088   8838.340      0.049      0.961   -1.69e+04    1.78e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                27.15
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               0.33   Skew:                             0.26
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.28
```

manula sarima plots:



Model Evaluation: RSME `359.61244606979693`

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.

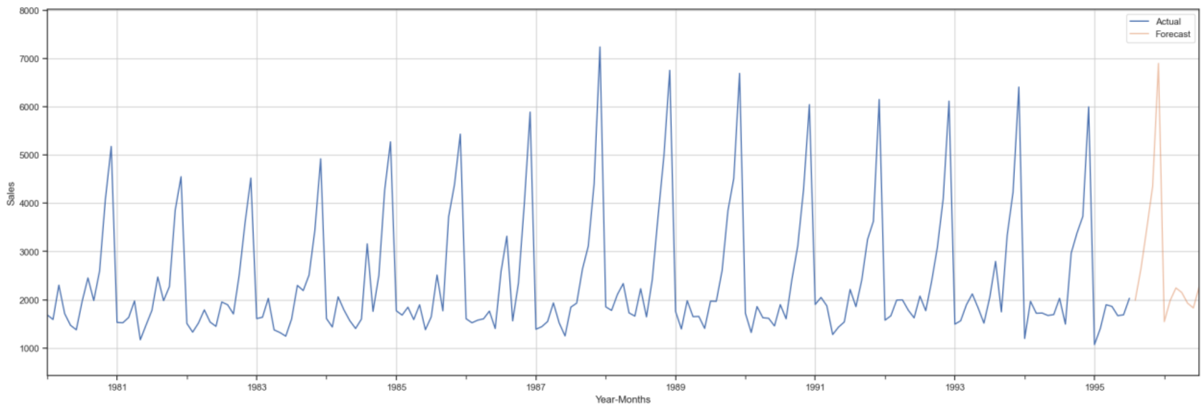
	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612446
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612446
(1,1,1),(2,0,3,12),Auto_SARIMA	528.602922
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple Average Model	1275.081804
Linear Regression	1275.867052
6pointTrailingMovingAverage	1283.927428
Auto_ARIMA	1299.979692
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1316.034674
ARIMA(3,1,3)	1319.936734
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Naive Model	3864.279352

We can see that the triple exponential smoothing model with ***alpha 0.1, beta 0.7, and gamma 0.2*** *is the best as it he the lowest RSME score.*

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

	Sales_Predictions
1995-08-01	1988.782193
1995-09-01	2652.762887
1995-10-01	3483.872246
1995-11-01	4354.989747
1995-12-01	6900.103171
1996-01-01	1546.800546
1996-02-01	1981.361768
1996-03-01	2245.459724
1996-04-01	2151.066942
1996-05-01	1929.355815
1996-06-01	1830.619260
1996-07-01	2272.156151

The sales prediction on the graph along with the confidence intervals. PFB the graph.



Predictions, 1 year into the future are shown in orange color, while the confidence interval has been shown in grey color.

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

- Sparkling wine sales are expected to be at least the same as last year or even higher next year. It's always been a popular choice, with sales dropping slightly but not by much.
- Sales pick up in the second half of the year, especially from August to December. So it might be a good idea to focus on marketing during the first half of the year.
- To encourage people to buy less popular wine like Rose wine, you can have promotions where people can get a good deal if they buy both Sparkling and Rose wines together. This might help boost sales of Rose wine.