TSF Project Report -Sparkling Wine

Parthasarathi Behura
PGP-DSBA

INDEX

s.no	Title	Page no
1	Read the data as an appropriate Time Series data and plot the data.	4-5
2	Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.	6-11
3	Split the data into training and test. The test data should start in 1991.	12-13
4	4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data.	13-20
5	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.	21-22
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.	23-26
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.	27-31
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.	32
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.	33-24
10	Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.	35

List of Tables:

1. Data dictionary
2. Rows of dataset
3. Rows of new dataset
4. Statistical summary
5. Test and train dataset

List of Plots:

1. Line plot of dataset
2. Boxplot of dataset
3. Lineplot of sales
4. Boxplot of yearly data
5. Boxplot of monthly data
6. Boxplot of weekday vise
7. Graph of monthly sales over the year
8. Correlation
9. ECDF plot
10. Decomposition addictive
11. Decomposition multiplicative
12. Train and test dataset
13. Linear regression
14. Naive approach
15. Simple average
16. Moving average
17. Simple exponential smoothing
18. Double exponential smoothing
19. Triple exponential smoothing
20. Dickey fuller test
21. Dickey fuller test after diff
22. SARIMA plots
23. PACF and ACF plot
24. PACF and ACF plot train dataset
25. Manual ARIMA plot
26. Manual SARIMA plot
27. Prediction plot

Problem Statement:

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century. By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

Objective

The primary objective of this project is to analyze and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

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

Data Dictionary:

Column name	Details
YearMonth	Dates of sales
Sparkling	Sales of sparkling wine

Table 1: data dictionary

Rows of data set;

<table>
<tr><th>Top Few Rows:</th><th>Last Few Rows:</th></tr>
<tr><td>

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

</td><td>

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

</td></tr>
</table>

Table 2: rows of data

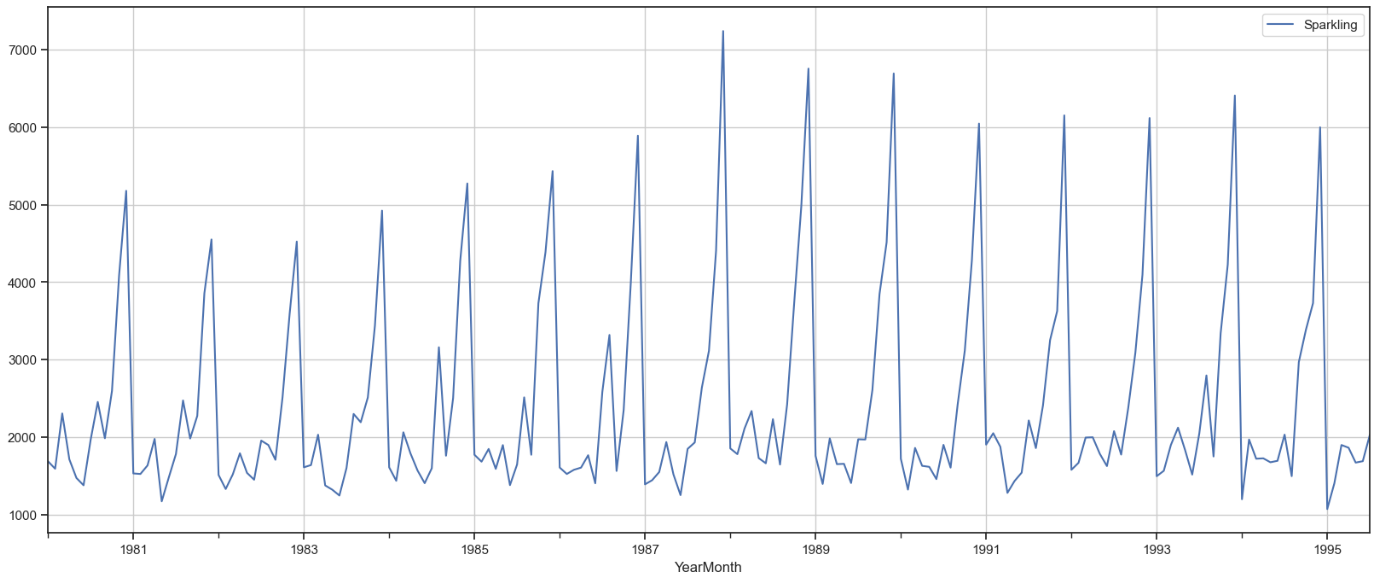
We can see data is from 1980 to 1995.

The number of Rows and Columns of the Dataset:

The dataset has 187 rows and 1 column.

Plot of the dataset:

Plot 1: dataset



Investigation of 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.

Rows of the new data set;

Table 3 : rows of new dataset

Top Few Rows:

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

Last Few Rows:

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

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

Data Type;

Index: DateTime

Sales: integer

Month: integer

Year: integer

Statistical summary:

Table 4: statistical summary of data

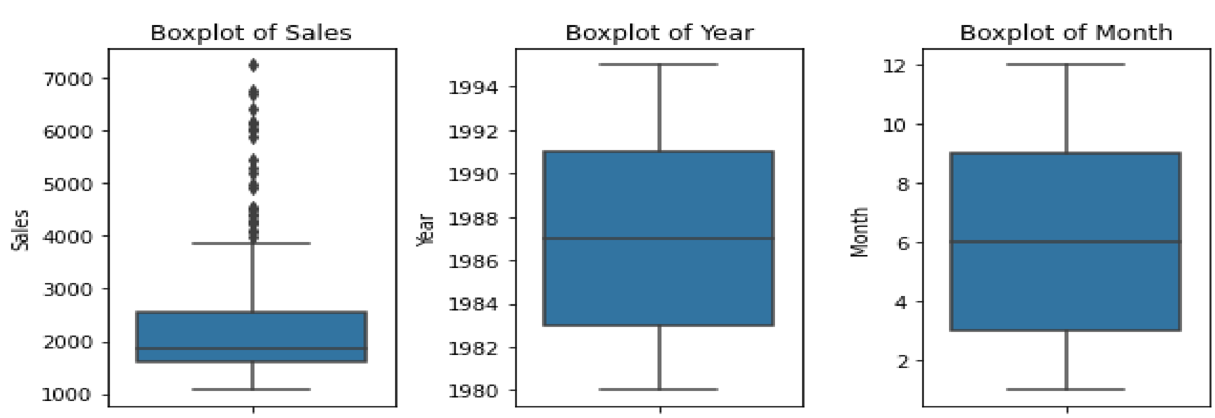
	count	mean	std	min	25%	50%	75%	max
Sales	187.0	2402.0	1295.0	1070.0	1605.0	1874.0	2549.0	7242.0
Year	187.0	1987.0	5.0	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.0	3.0	1.0	3.0	6.0	9.0	12.0

Null Value:

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

Boxplot of dataset:

Plot 2: Box plot of Data

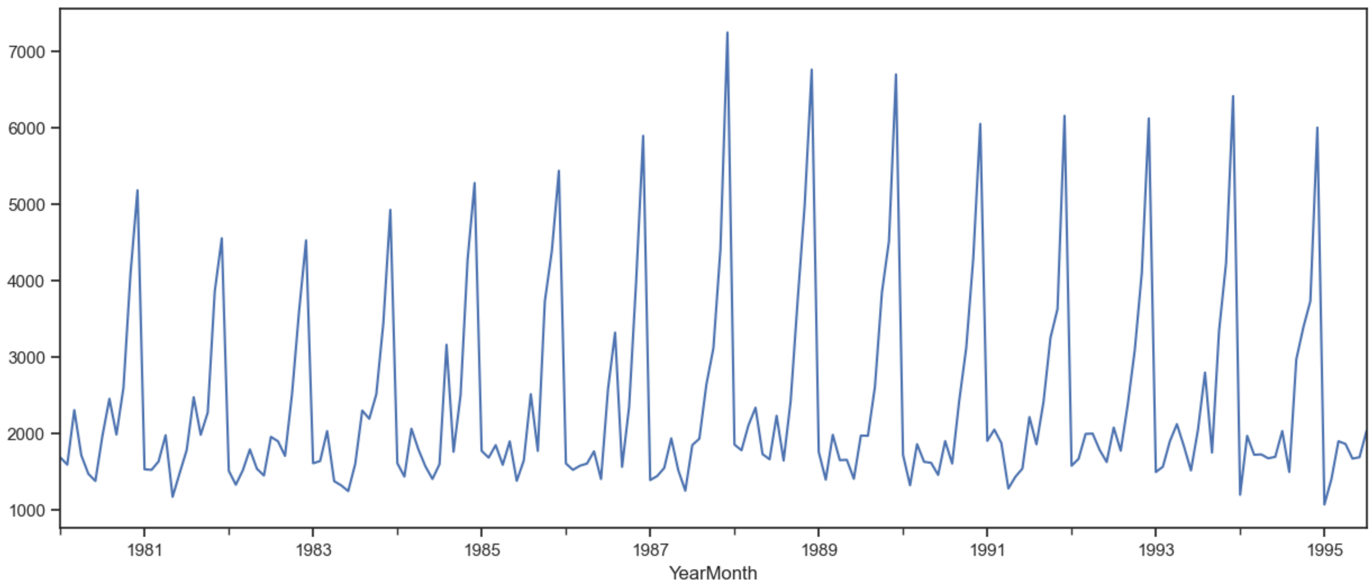


The box plot shows:

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

Line plot of sales:

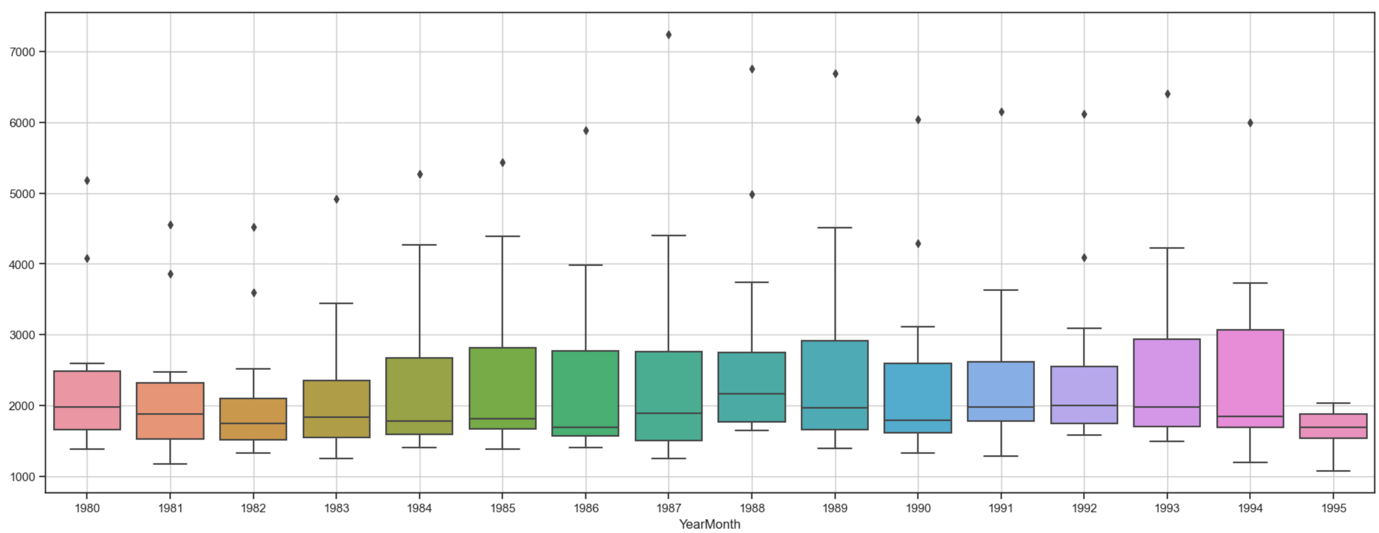
Plot 3: line plot for sales



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

Boxplot Yearly:

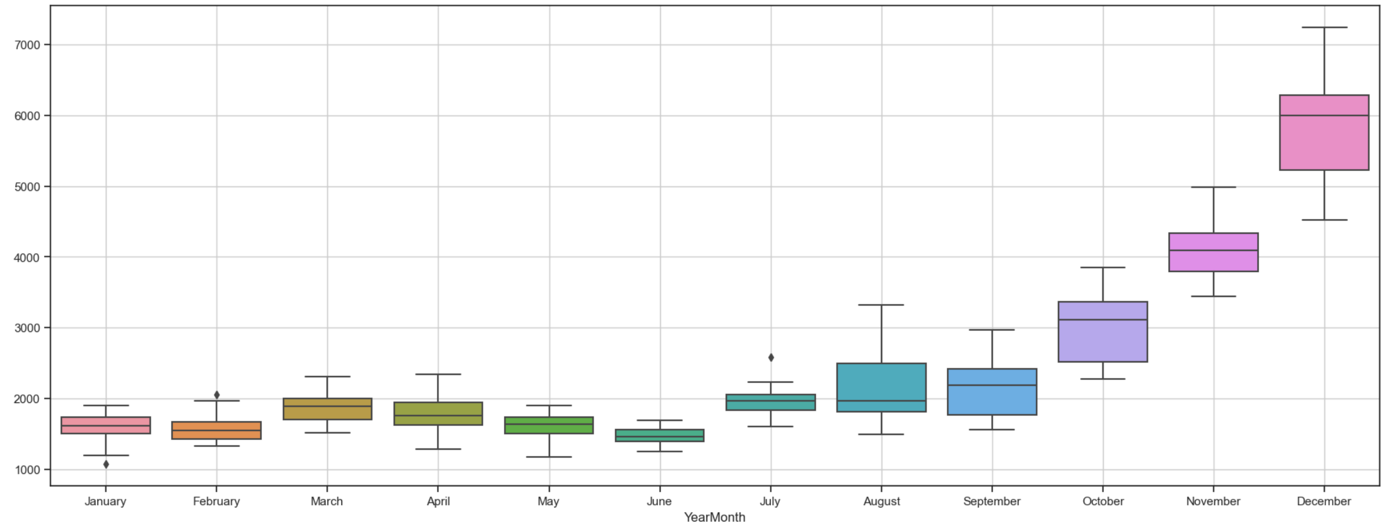
Plot 4: boxplot yearly



This yearly box plot shows there is consistency over the years and there was a peak in 1988-1989. Outliers are present in all years.

Boxplot Monthly:

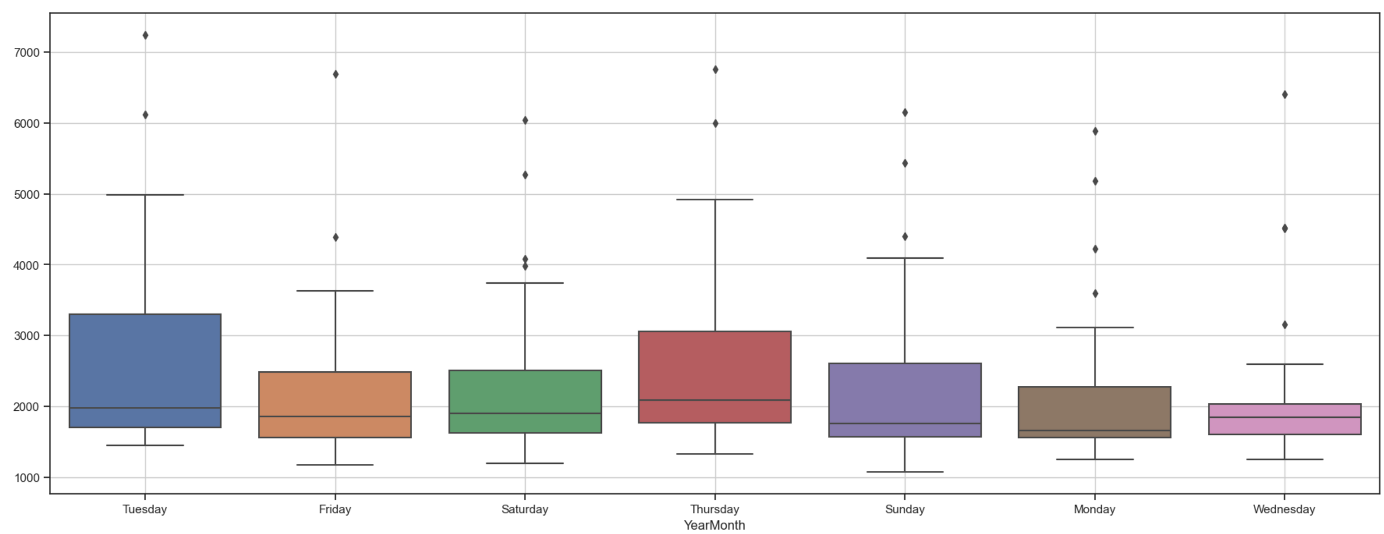
Plot 5: boxplot monthly



The plot shows that sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to July then from august the sales start to increase. Outliers are present in January, February and July.

Boxplot Weekday wise:

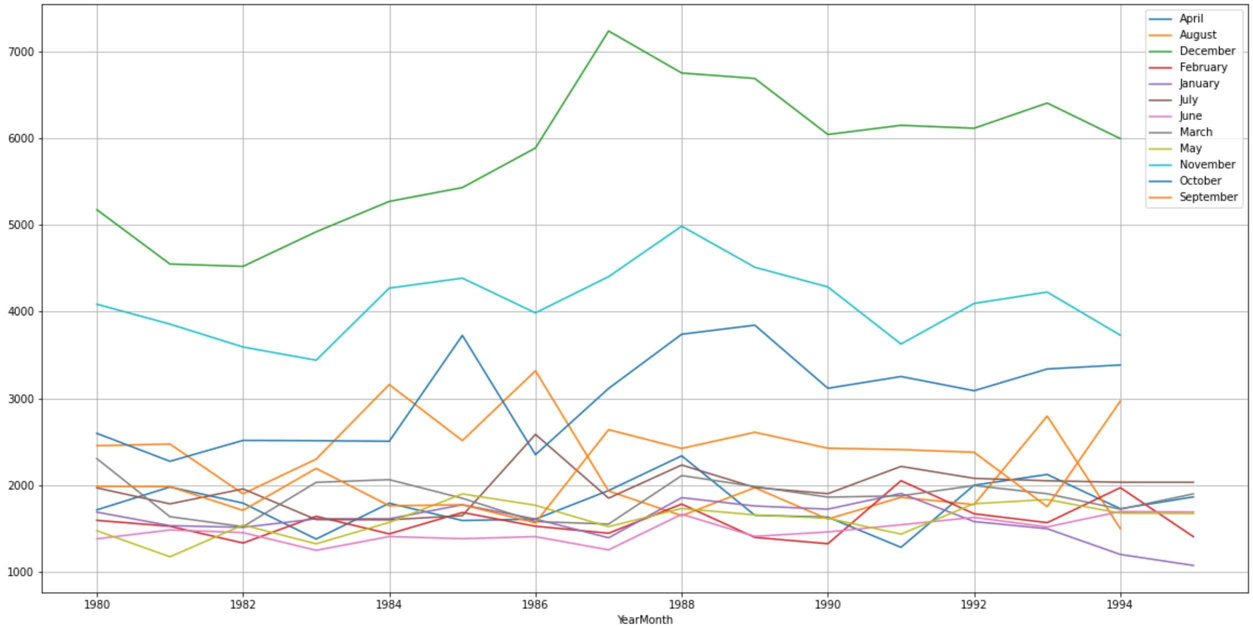
Plot 6: Boxplot weekday wise



Tuesday has more sales than other days and Wednesday has the lowest sales of the week. Outliers are present on all days which is understandable.

Graph of Monthly Sales over the years:

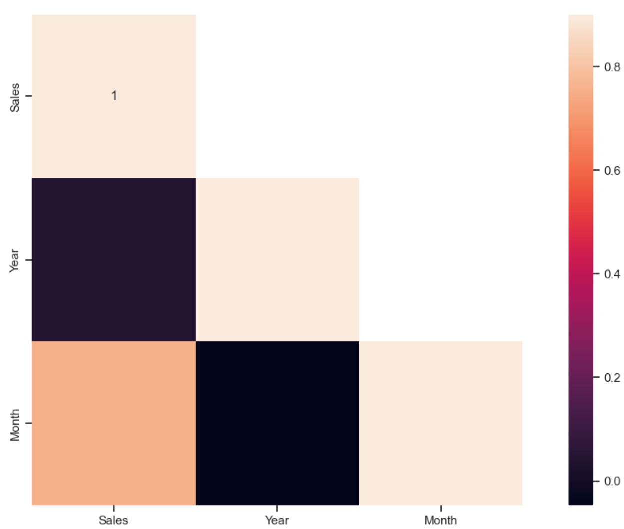
Plot 7: graph of monthly sales over the year



This plot shows that December has the highest sales over the years and the year 1988 was the year with the highest number of sales.

Correlation plot

Plot 8: Correlation plot

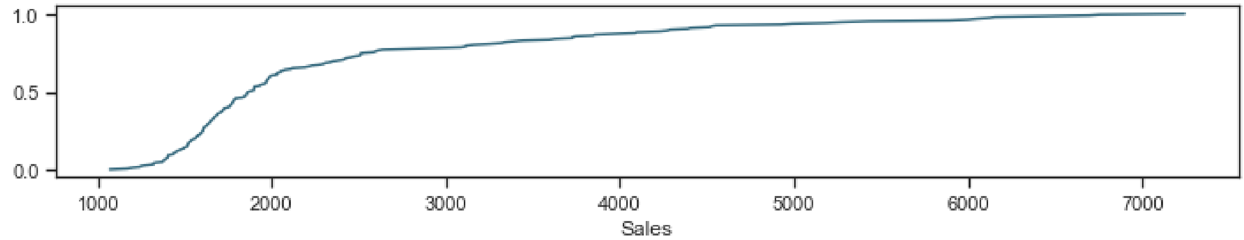


This heat map shows that there is a low correlation between sales and year. There is a more correlation between month and sales. It indicated seasonal patterns in sales

Plot ECDF: Empirical Cumulative Distribution Function

This graph shows the distribution of data.

Plot 9: ECDF plot

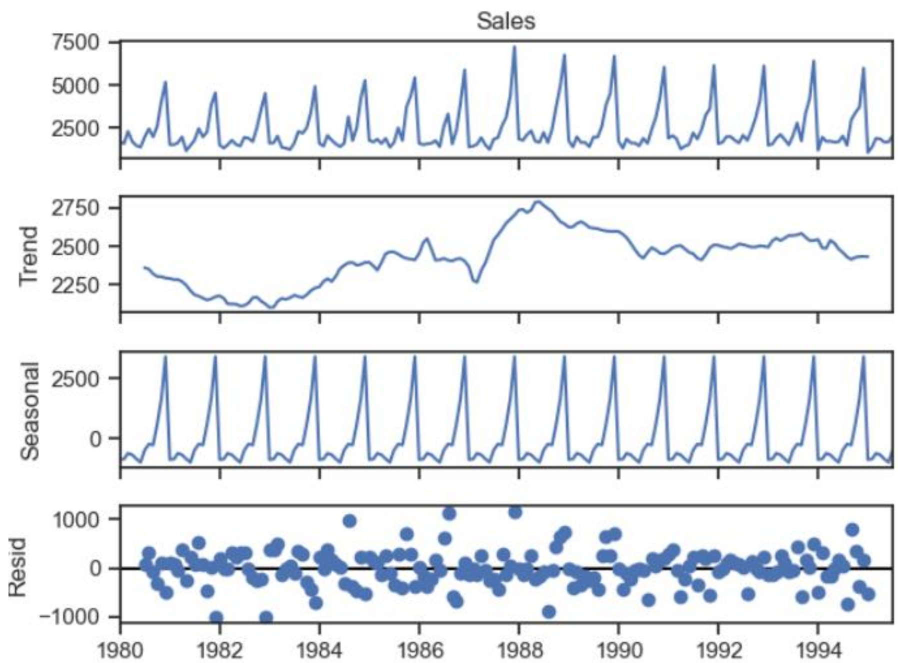


This plot shows:

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

Decomposition -Additive

Plot 10: decomposition plot addictive



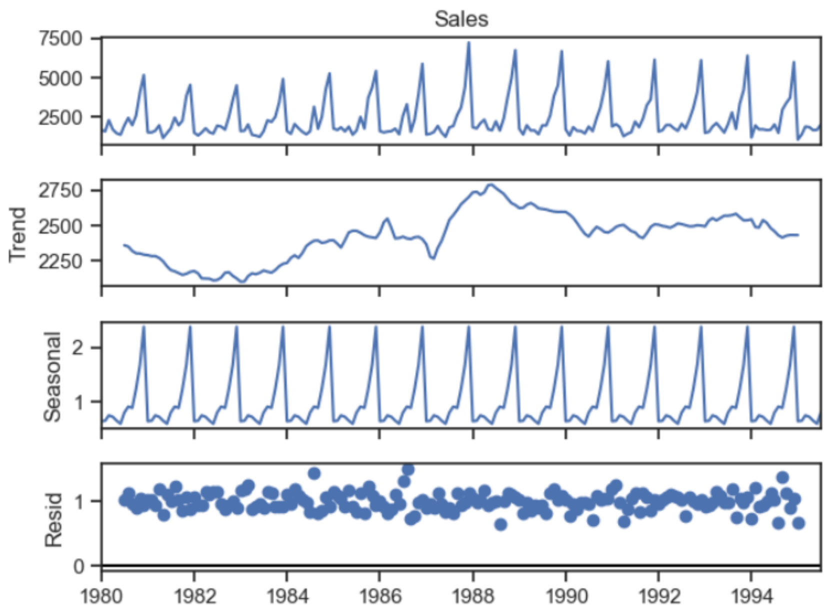
The plots show:

- 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

Plot 11: decomposition plot - mulatiplicative

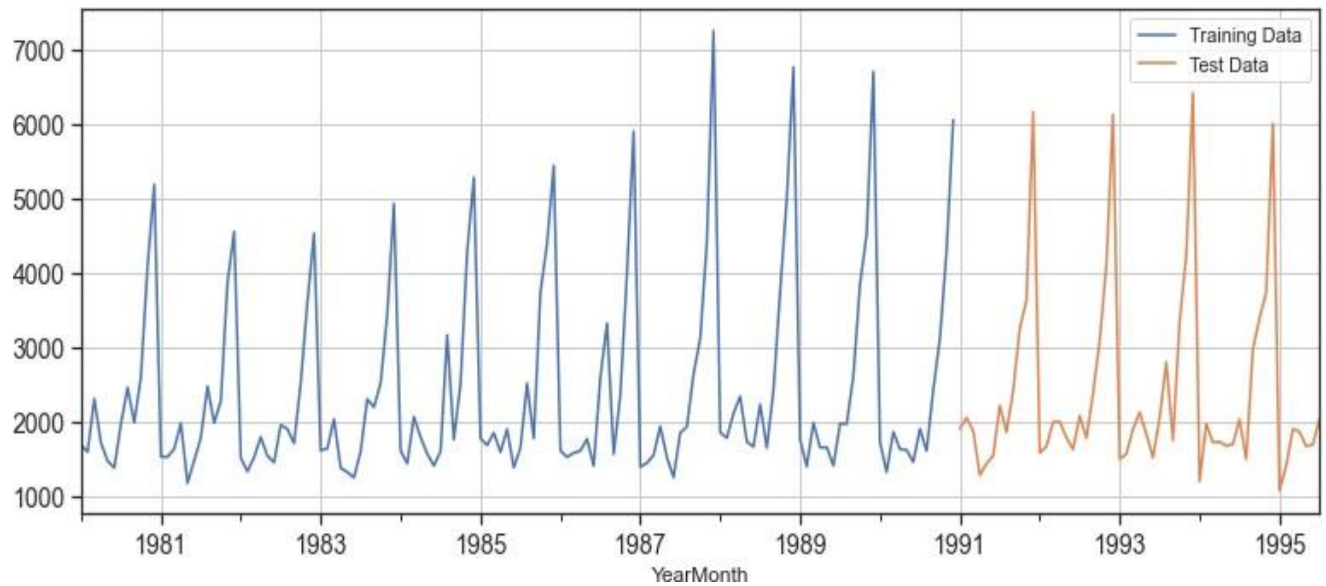


The plots show

- 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.

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

plot 12: line plot train and test dataset



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:

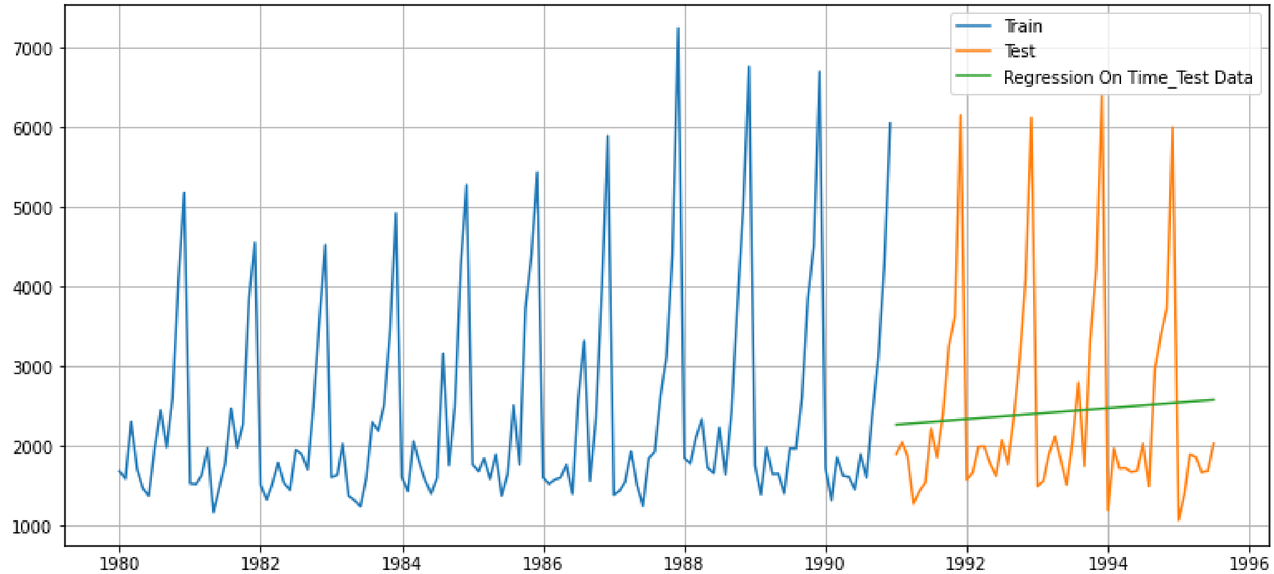
train dataset has 132 rows and 3 columns.
test dataset has 55 and 3 columns.

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression 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
- Model 2: Simple Average
- Model 3: Moving Average(MA)
- Model 4: Simple Exponential Smoothing
- Model 5: Double Exponential Smoothing (Holt's Model)
- Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

Model 1: Linear Regression

Plot 13: linear regression



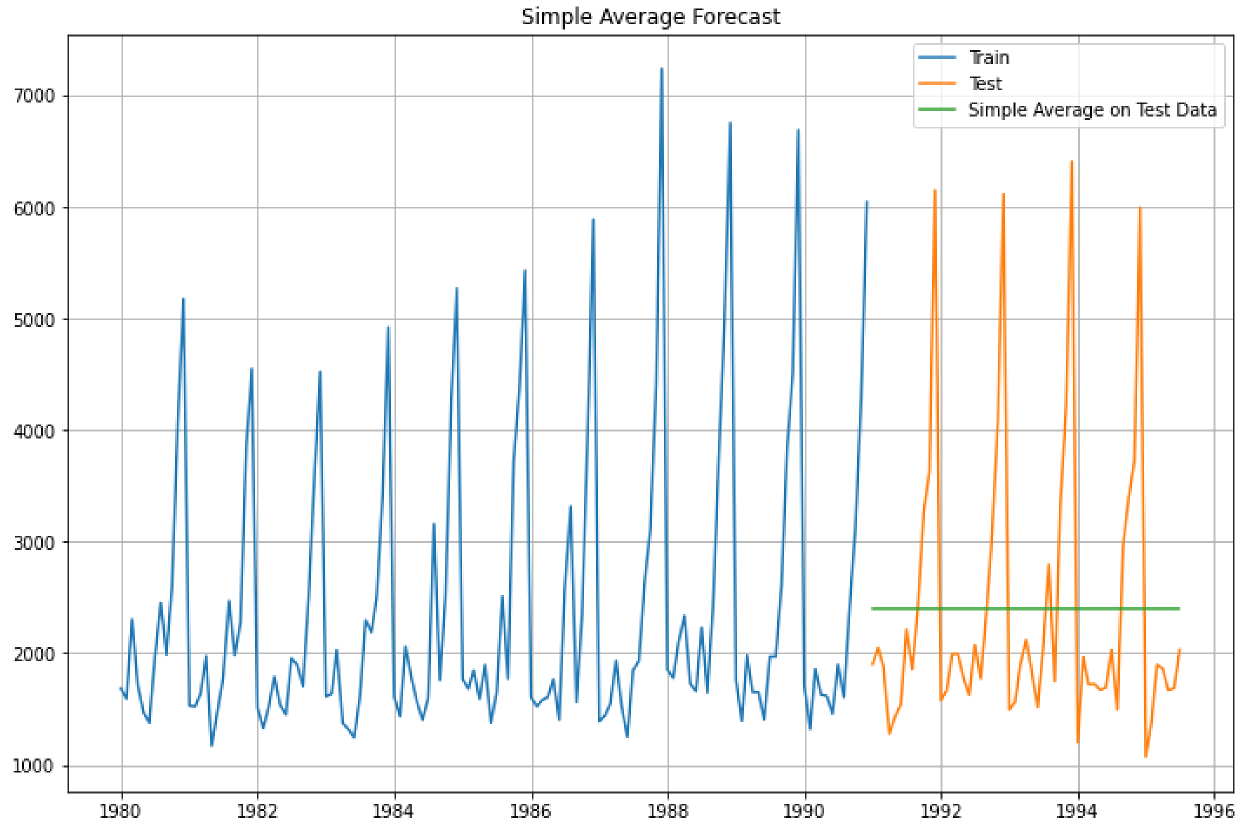
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values.

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

Linear Regression	1275.867052

Method 2: Simple Average

Plot 14: simple average



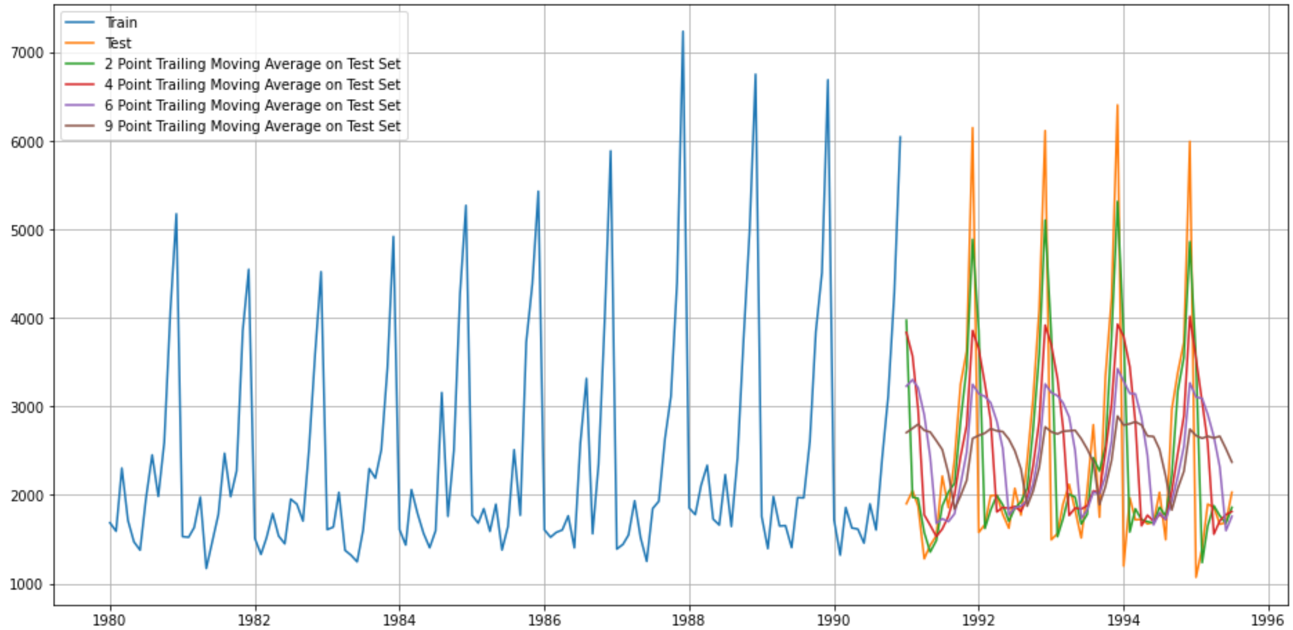
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

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

Simple Average Model	1275.081804

Method 3: Moving Average(MA)

Plot 15: moving average



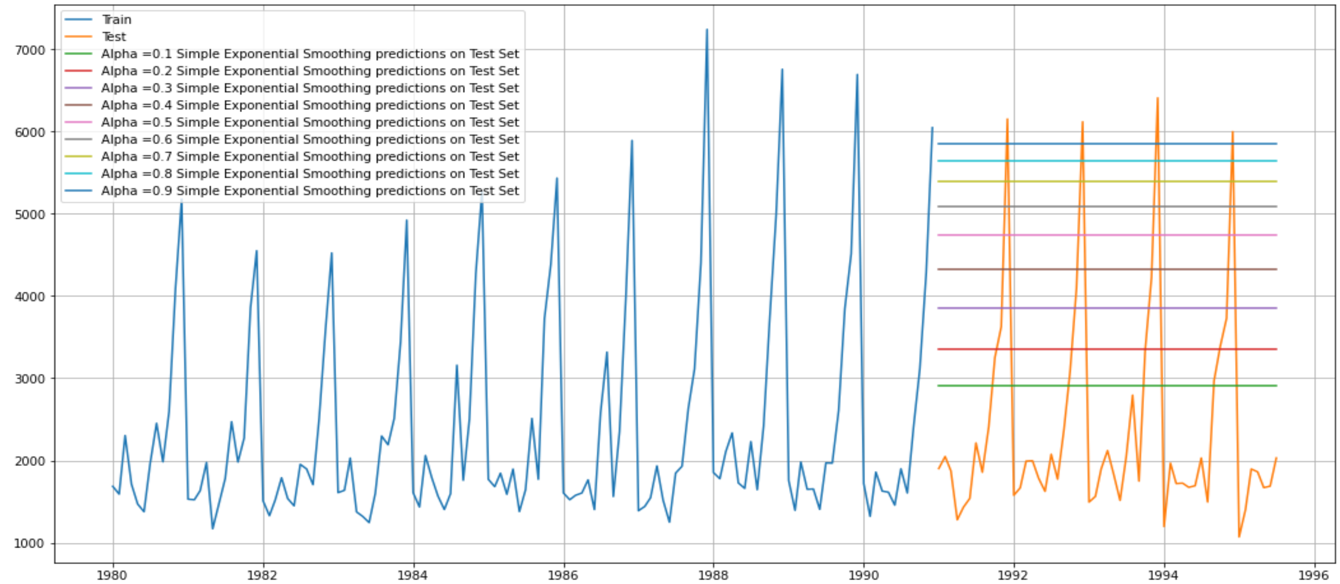
Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

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.

Method 4: Simple Exponential Smoothing

Plot 16: simple exponential smothing

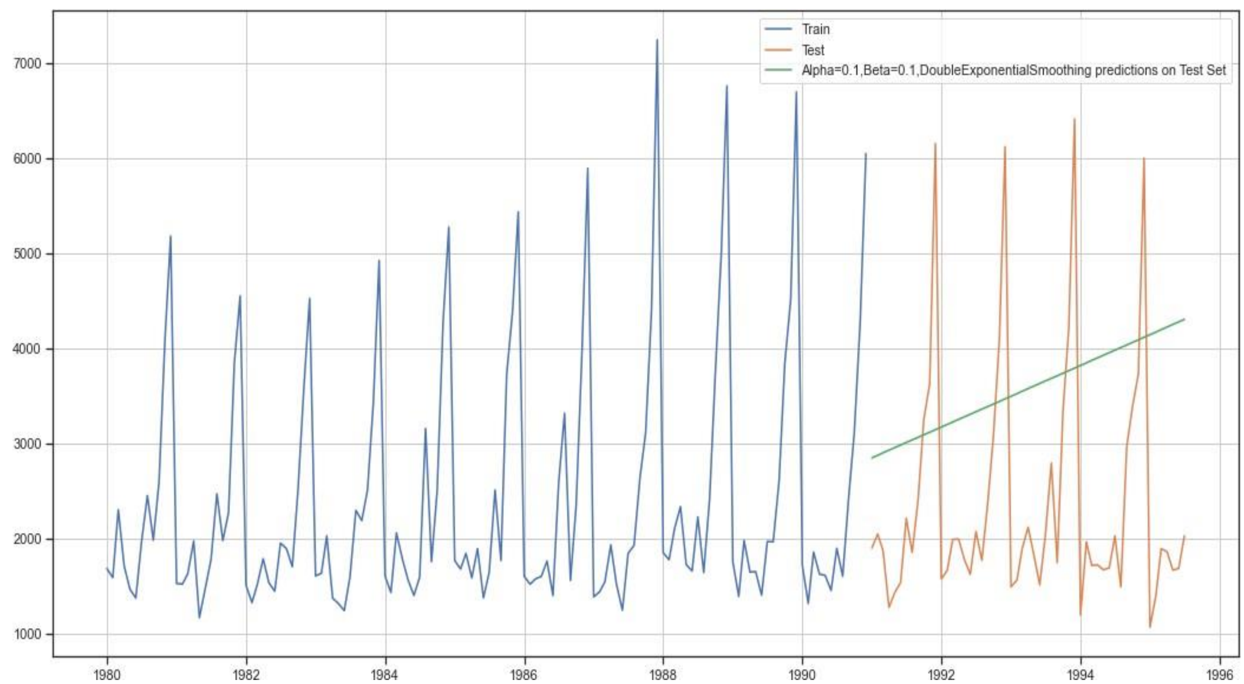


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

	Alpha Values	Train RMSE	Test RMSE
0	0.1	1333.873836	1375.393398
1	0.2	1356.042987	1595.206839
2	0.3	1359.511747	1935.507132
3	0.4	1352.588879	2311.919615
4	0.5	1344.004369	2666.351413
5	0.6	1338.805381	2979.204388
6	0.7	1338.844308	3249.944092
7	0.8	1344.462091	3483.801006
8	0.9	1355.723518	3686.794285

Method 5: Double Exponential Smoothing (Holt's Model)

Plot 18: double exponential smoothing



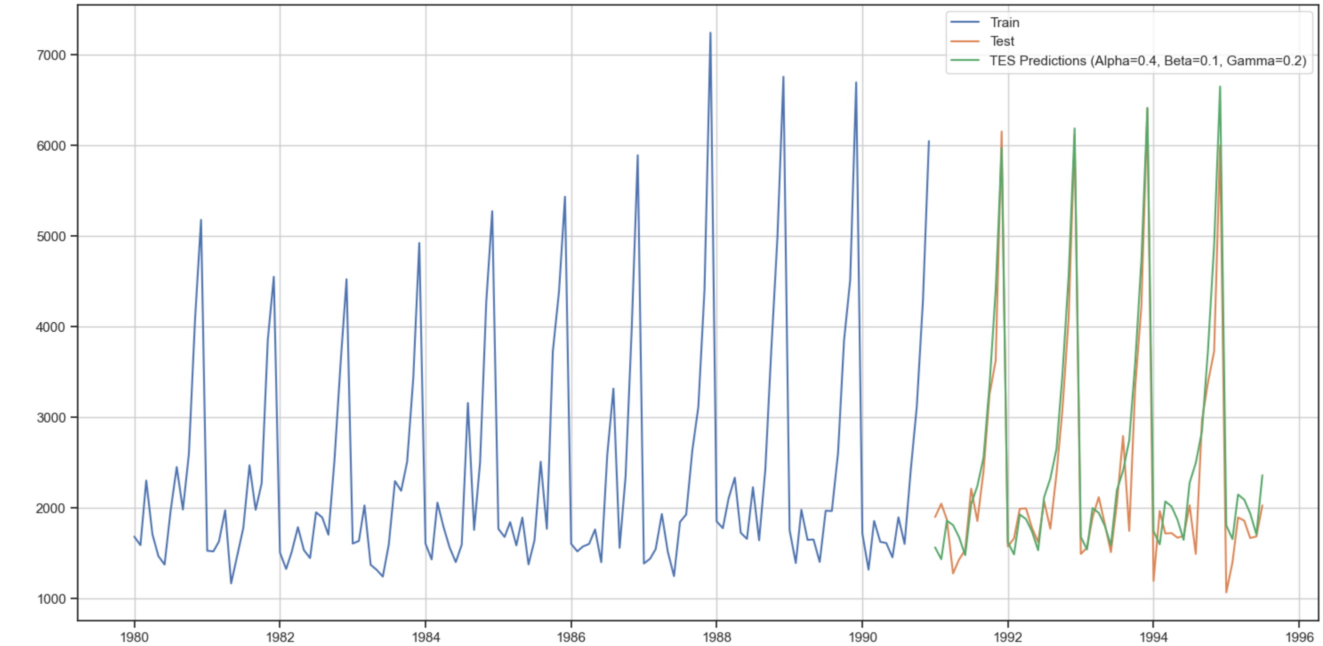
The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values.

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

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

Method 6: Triple Exponential Smoothing (Holt - Winter's Model)

Plot 19: triple exponential smoothing



Output for a best alpha, beta, and gamma values are shown by the green color line in the above plot. The best model had both a multiplicative trends, as well as a seasonality Model, which was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.4,Beta=0.1,Gamma=0.3,TripleExponentialSmoothing	341.653525

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.

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.

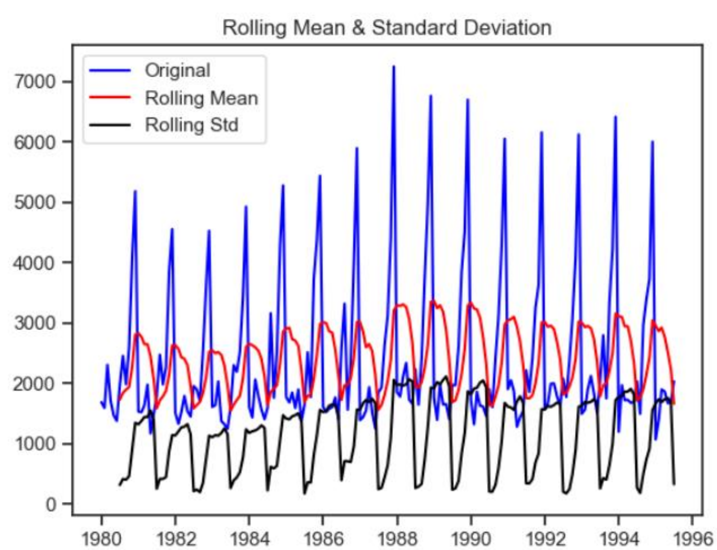
The hypothesis in a simple form for the ADF test is:

- 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.

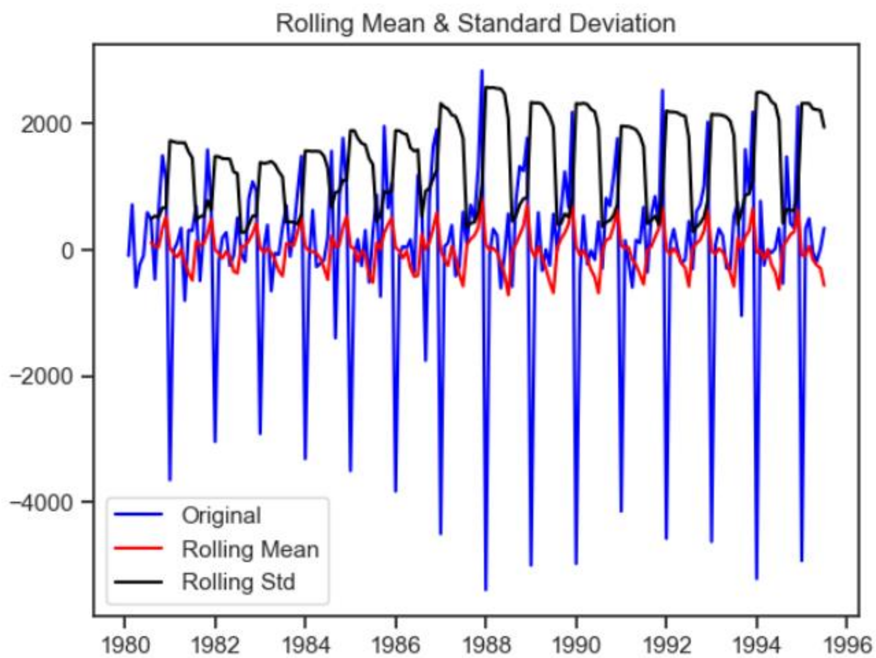
Plot 20: plot for dickey fuller test



Results of Dickey-Fuller Test:
p-value 0.601061

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

Plot 21: plot for dickey fuller test after differencing approach



Results of Dickey-Fuller Test:

p-value 0.000000

Dickey - Fuller test was 0.000, which is obviously 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. 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 the directions, indicating stationarity.

We could now proceed ahead with ARIMA/ SARIMA models, since we had made the series stationary.

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

AUTO - ARIMA 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 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)	2235.607812
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=2).

```
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 08 Dec 2024   AIC                           2213.509
Time:                        16:23:36   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:

Auto_ARIMA	1299.979692

AUTO- SARIMA Model

A similar for loop like AUTO_ARIMA with 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)

Akaike information criterion (AIC) value was evaluated for each of these models and the model with least AIC value was selected. Here only the top 5 models are shown.

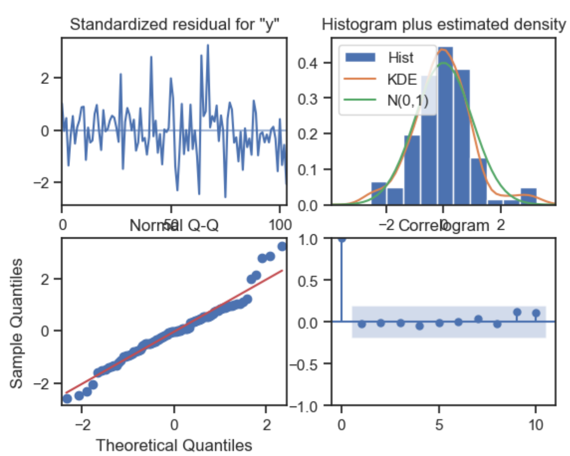
	param	seasonal	AIC
83	(1, 1, 1)	(0, 0, 3, 12)	12.000000
183	(2, 1, 3)	(1, 0, 3, 12)	20.000000
215	(3, 1, 1)	(1, 0, 3, 12)	208.047774
91	(1, 1, 1)	(2, 0, 3, 12)	259.695513
251	(3, 1, 3)	(2, 0, 3, 12)	349.867995

The summary report for the best SARIMA model with values (2,1,2)(2,0,2,12)

```
==============================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 08 Dec 2024   AIC                           1555.584
Time:                                    16:33:28   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
ar.L1         -0.6281      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      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      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
=======================================================================================
```

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

Plot 22: SARIMA plot



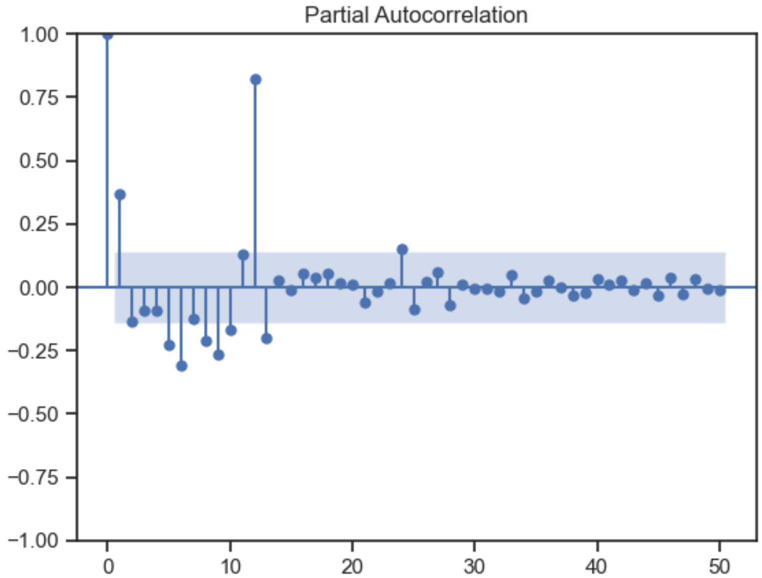
RSME of Model:

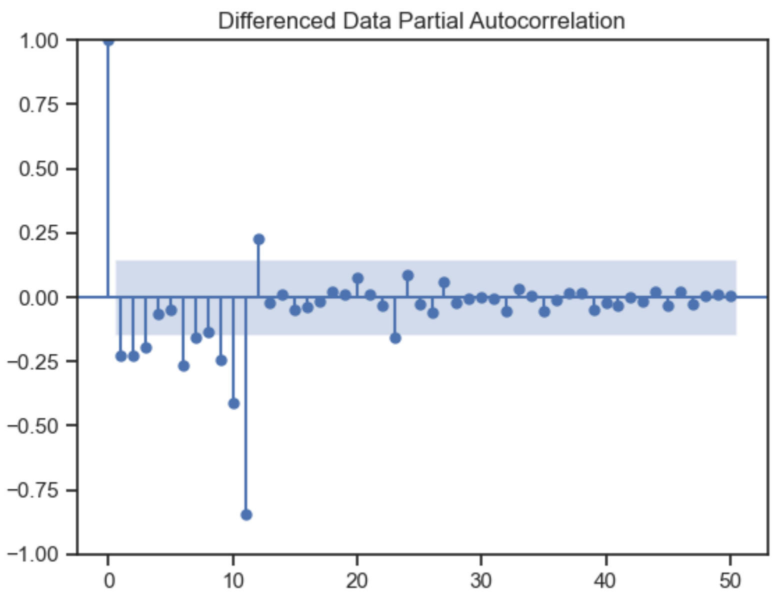
528.590608

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.

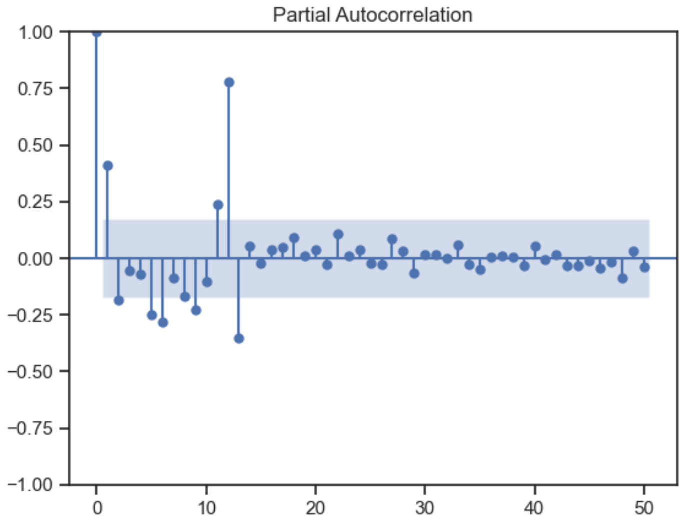
Manual- ARIMA Model

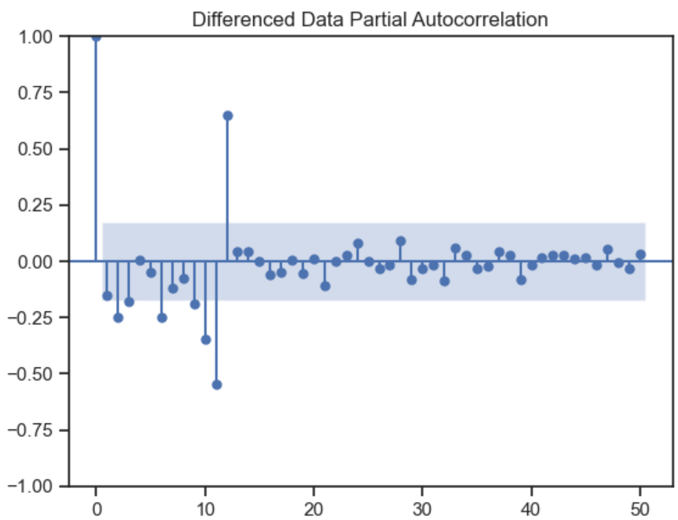
PFB the ACF plot on data





training data with diff(1):





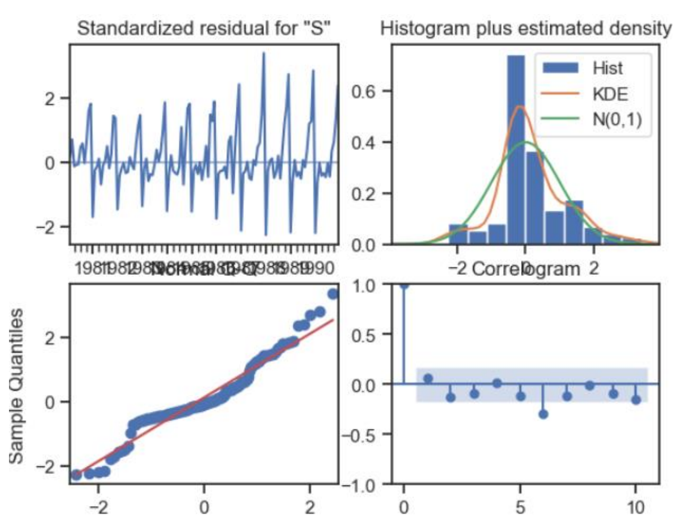
Looking at ACF plot we can see a shard decay after lag 1 for original as well as differenced data.hence we select the q value to be 1. i.e. q=1.

Looking at PACF plot we can again see significant bars till lag 1 for differenced series which is stationary in nature, post 1 the decay is large enough. Hence we choose p value to be 1. i.e. p=1. d values will be 1, since we had seen earlier that the series is stationary with lag1.

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

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, 08 Dec 2024   AIC                           2235.755
Time:                        16:33:33   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
===================================================================================
```



Model Evaluation: RSME

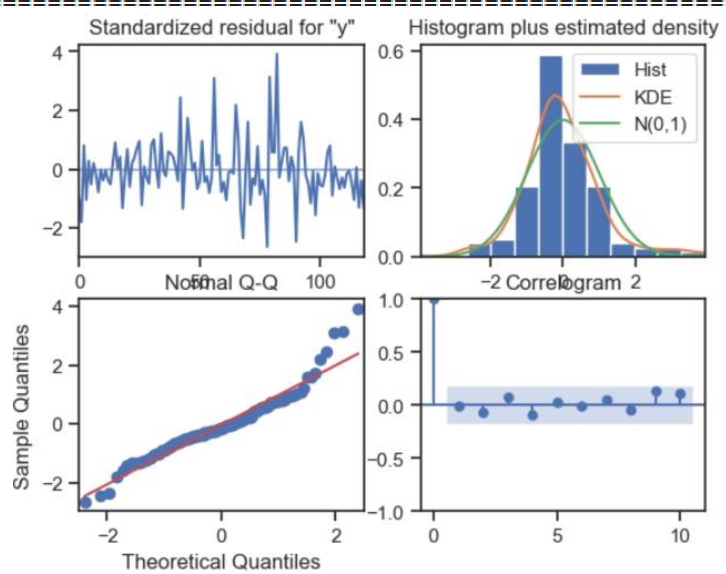
1319.936732

Manual SARIMA Model

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

Below is the summary of the manual SARIMA model

```
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                -882.088
Date:                            Sun, 08 Dec 2024   AIC                           1774.175
Time:                                    16:33:36   BIC                           1788.071
Sample:                                         0   HQIC                          1779.818
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1957      0.104      1.878      0.060      -0.009       0.400
ma.L1         -0.9404      0.053    -17.897      0.000      -1.043      -0.837
ar.S.L12       0.0711      0.242      0.294      0.769      -0.404       0.546
ma.S.L12      -0.5035      0.221     -2.277      0.023      -0.937      -0.070
sigma2       1.51e+05   1.33e+04     11.371      0.000    1.25e+05    1.77e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                45.66
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               2.61   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.56
===================================================================================
```



Model Evaluation: RSME

359.612449

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
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	341.653525
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612449
(1,1,1),(2,0,3,12),Auto_SARIMA	528.590608
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	1304.927405
ARIMA(3,1,3)	1319.936732
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670

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.

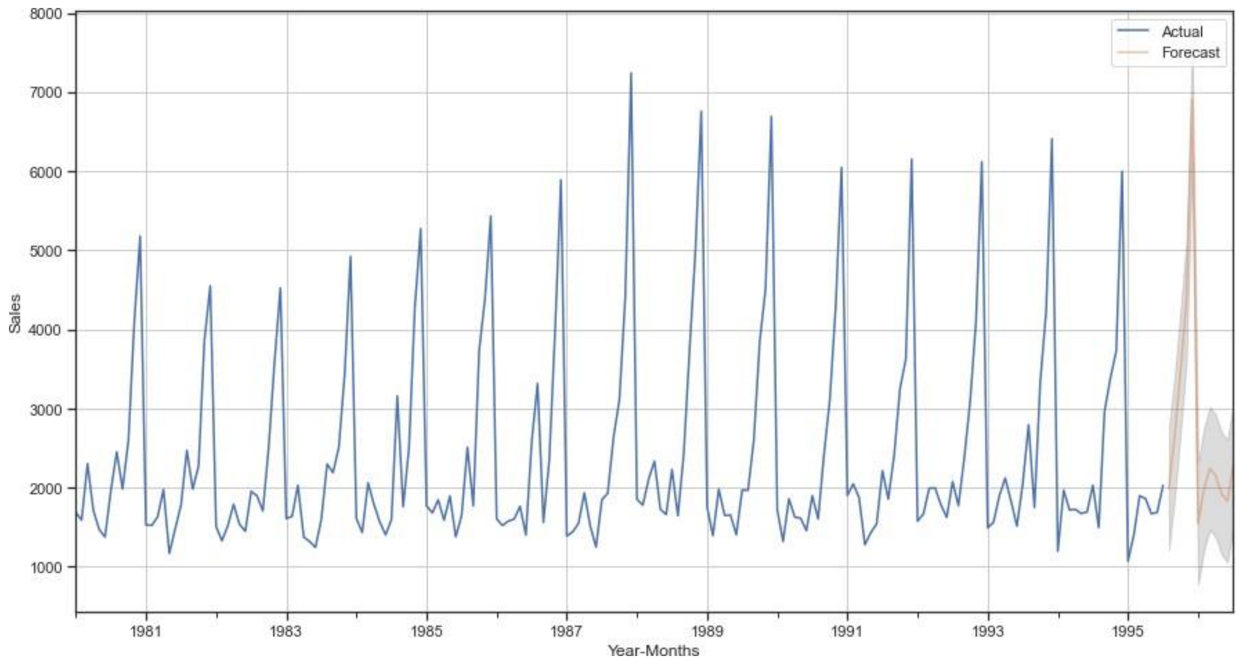
Based on the above comparison of all the various models that we had built, we can conclude that the triple exponential smoothing or the Holts-Winter model is giving us the lowest RMSE, hence it would be the most optimum model

sales predictions made by this best optimum model.

	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.

Plot 27: prediction plot



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

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

- The sales for Sparkling wine for the company are predicted to be at least the same as last year, if not more, with peak sales for next year potentially higher than this year.
- Sparkling wine has been a consistently popular wine among customers with only a very marginal decline in sales, despite reaching its peak popularity in the late 1980s.
- Seasonality has a significant impact on the sales of Sparkling wine, with sales being slow in the first half of the year and picking up from August to December.
- It is recommended for the company to run campaigns in the first half of the year when sales are slow, particularly in the months of March to July.
- Combining promotions where Sparkling wine is paired with a less popular wine such as "Rose wine" under a special offer may encourage customers to try the underperforming wine, which could potentially boost its sales and benefit the company.

-END-