Time Series Forecast
Project Report

Contents

List of Figures

List of Tables

Problem:

ABC Estate Wines has been a leader in the wine industry for many years, offering high-quality wines to consumers all around the world. As the company continues to expand its reach and grow its customer base, it is essential to analyse market trends and forecast future sales to ensure continued success.

In this report, we will focus on analysing the sales data for sparkling wine in the 20th century. As an analyst for ABC Estate Wines, I have been tasked with reviewing this data to identify patterns, trends, and opportunities for growth in the sparkling wine market.

This knowledge will help us to make informed decisions about how to position our products in the market, optimize our sales strategies, and forecast future sales trends.

Overall, this report aims to provide valuable insights into the sparkling wine market and how ABC Estate Wines can continue to succeed in this highly competitive industry.

Time Series Forecast Using Sparkling Data

1.1 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: Sparkling Data Dictionary

Data Set

Given dataset doesn't have time index.

Top Few Rows

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

Last Few Rows

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

Table 2: Rows of Sparkling Data Set

So, I have parsed the YearMonth column to create a timestamp.

Top Few Rows

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

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

Table 3: Rows of Sparkling Data after adding index

Shape:
We have 187 rows and 1 column in our Data set.

Plot of the Dataset

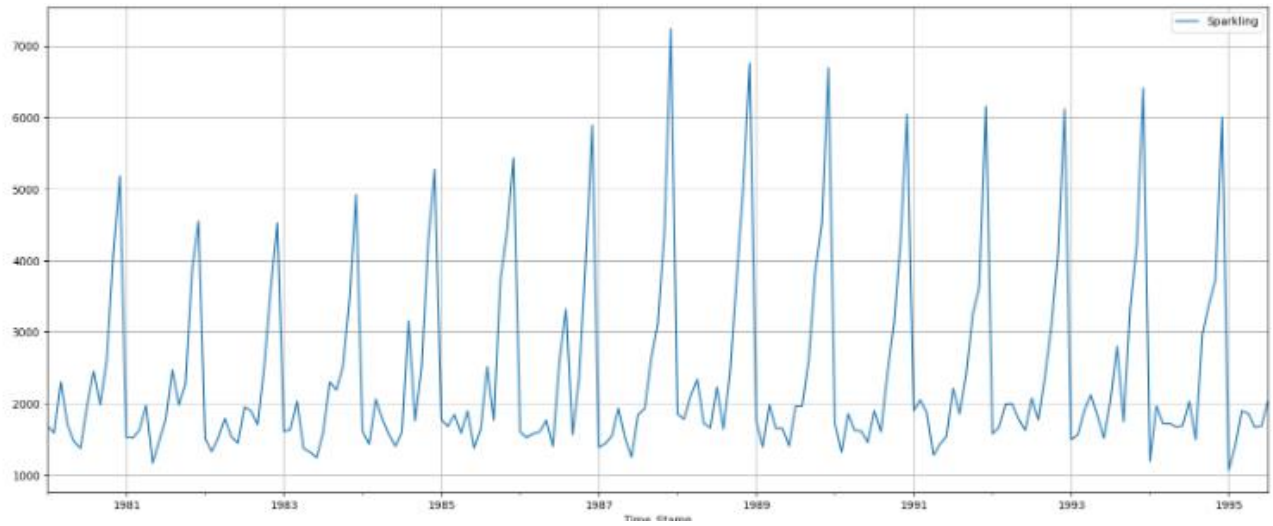

Figure 1: Plot of Sparkling Dataset

Observation:

- We can notice that the increasing trend in the initial years.
- Also, Dataset seems to contain Seasonality.

Post Ingestion 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

Top Few Rows

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

Last Few Rows

Time_Stamp	Sales	Year	Month
1995-03-31	1897	1995	3
1995-04-30	1862	1995	4
1995-05-31	1670	1995	5
1995-06-30	1688	1995	6
1995-07-31	2031	1995	7

Table 4: Rows of Sparkling Dataset after adding Sales and Year

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

Null Value Check

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

There are no null values present in the given data set.

Data Types:

```
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sales   187 non-null    int64
 1   Year    187 non-null    int32
 2   Month   187 non-null    int32
dtypes: int32(2), int64(1)
```

Index: DateTime
Sales: Integer
Month: Integer
Year: Integer

Summary of Dataset:

	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

Figure 2: Summary of Sparkling Dataset

Boxplot of Sales:

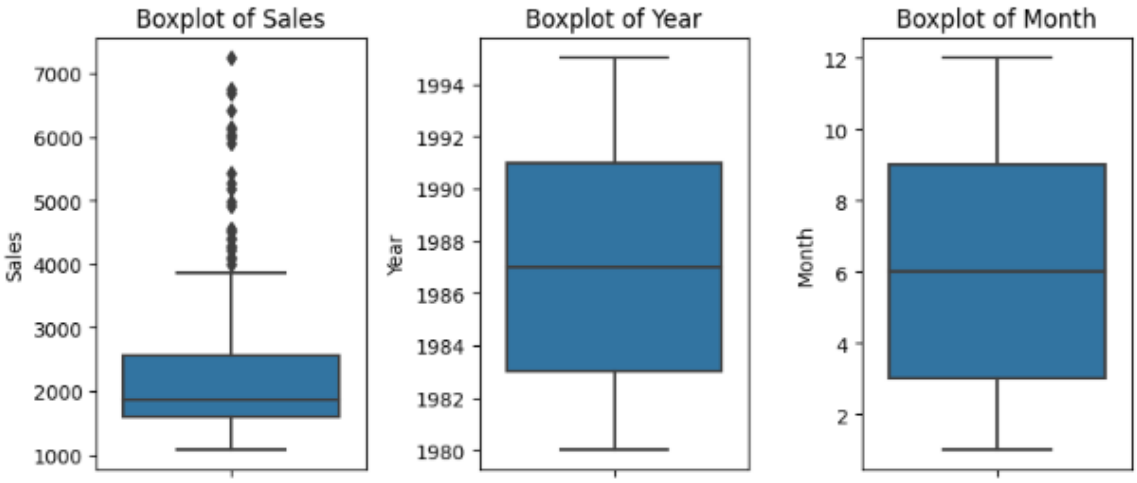

Figure 3: Box Plot of Sparkling Sales

- There are outliers in the sales boxplot that we can handle, but we have decided against doing so because they don't significantly affect the time series model.

Line Plot of Sales

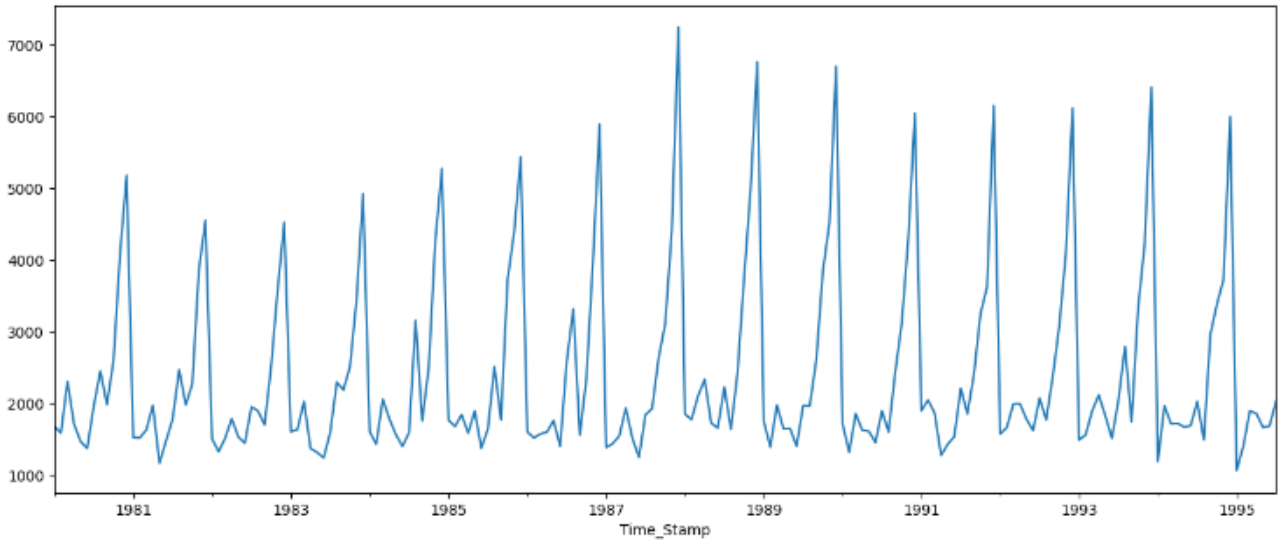

Figure 4: Line Plot of Sparkling Sales

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

Box Plot of Sales yearly

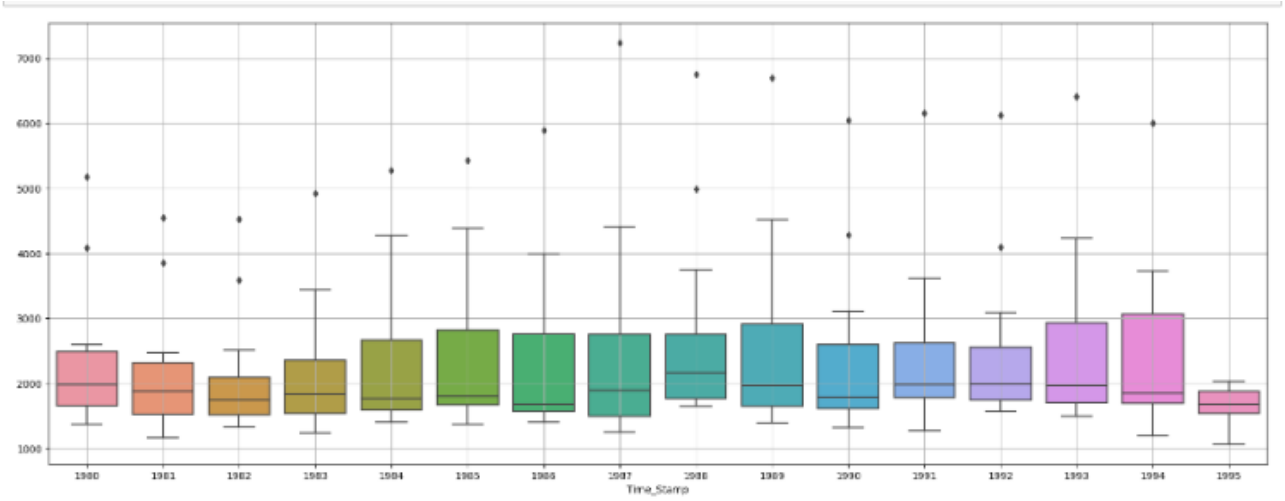

Figure 5: Box Plot of Sparkling Sales Yearly

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

Box Plot Monthly

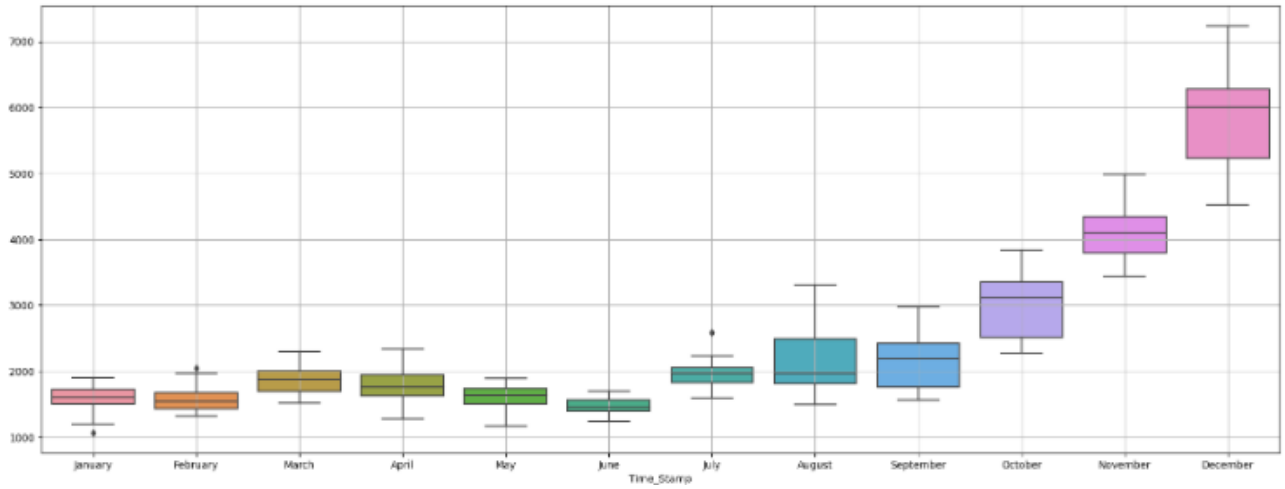

Figure 6: Box Plot of Sparkling Sales Monthly

- 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 Week Day vise:

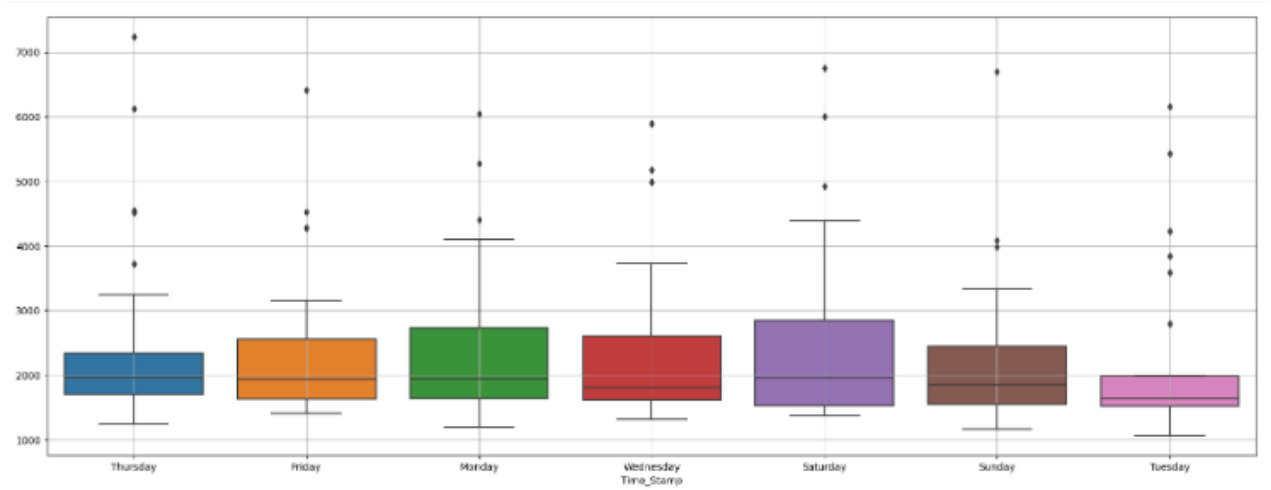

Figure 7: Box Plot of Sparkling Sales Weekly

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

Monthly Sales over the years

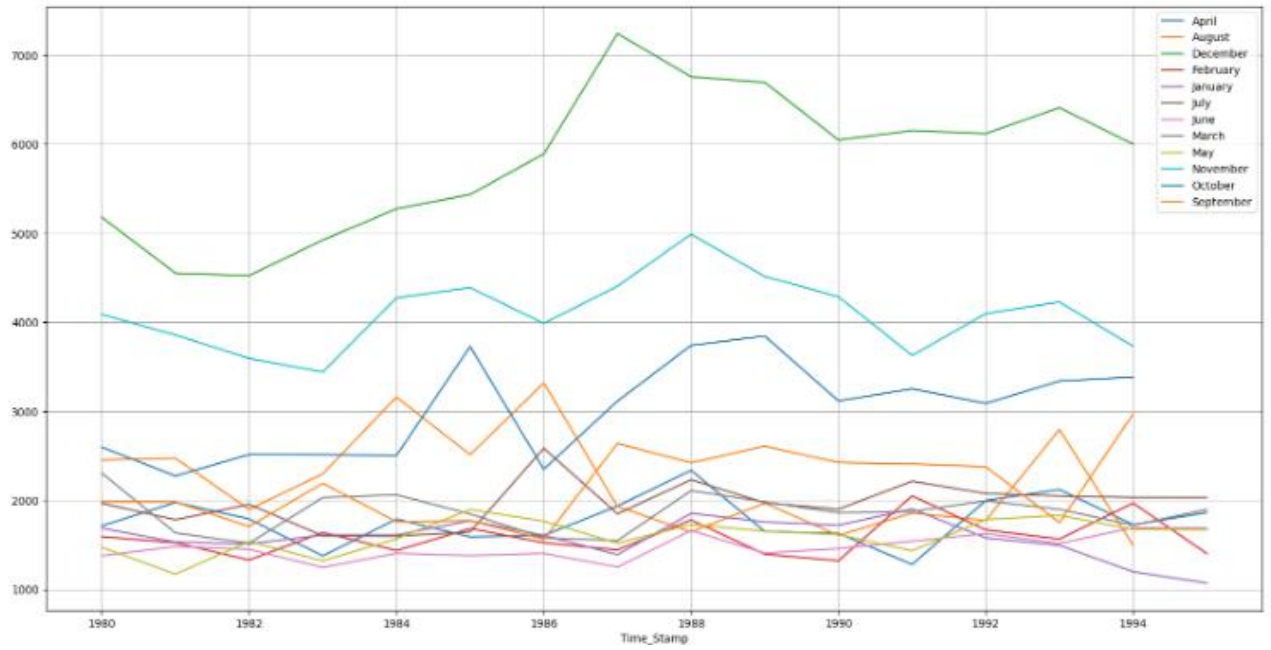

Figure 8: Graph of monthly Sparkling sales over the year

- 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:

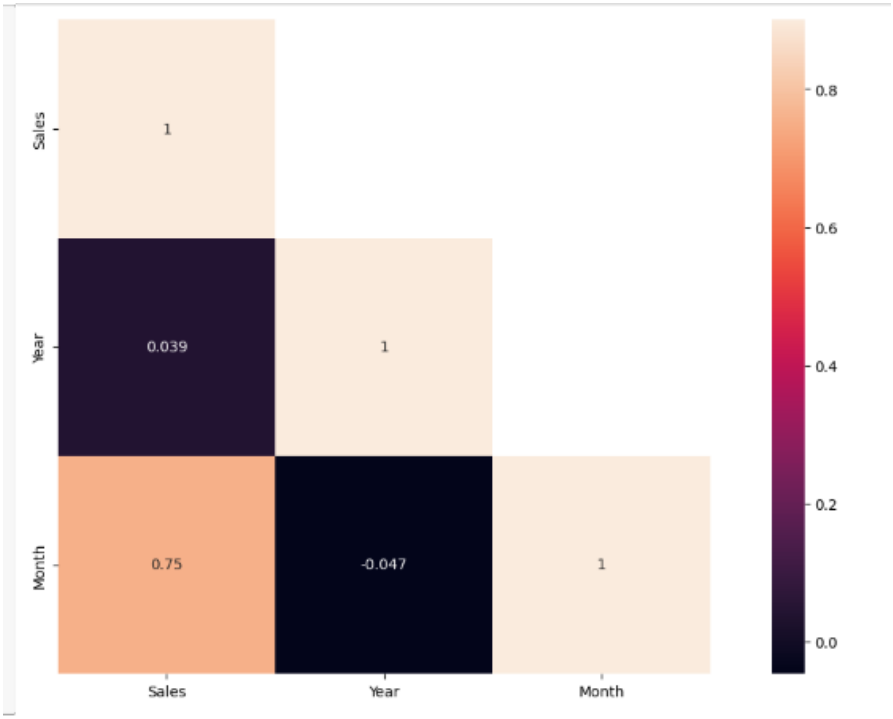

Figure 9: Sparkling Correlation Plot

- 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 sale

Decomposition-Additive

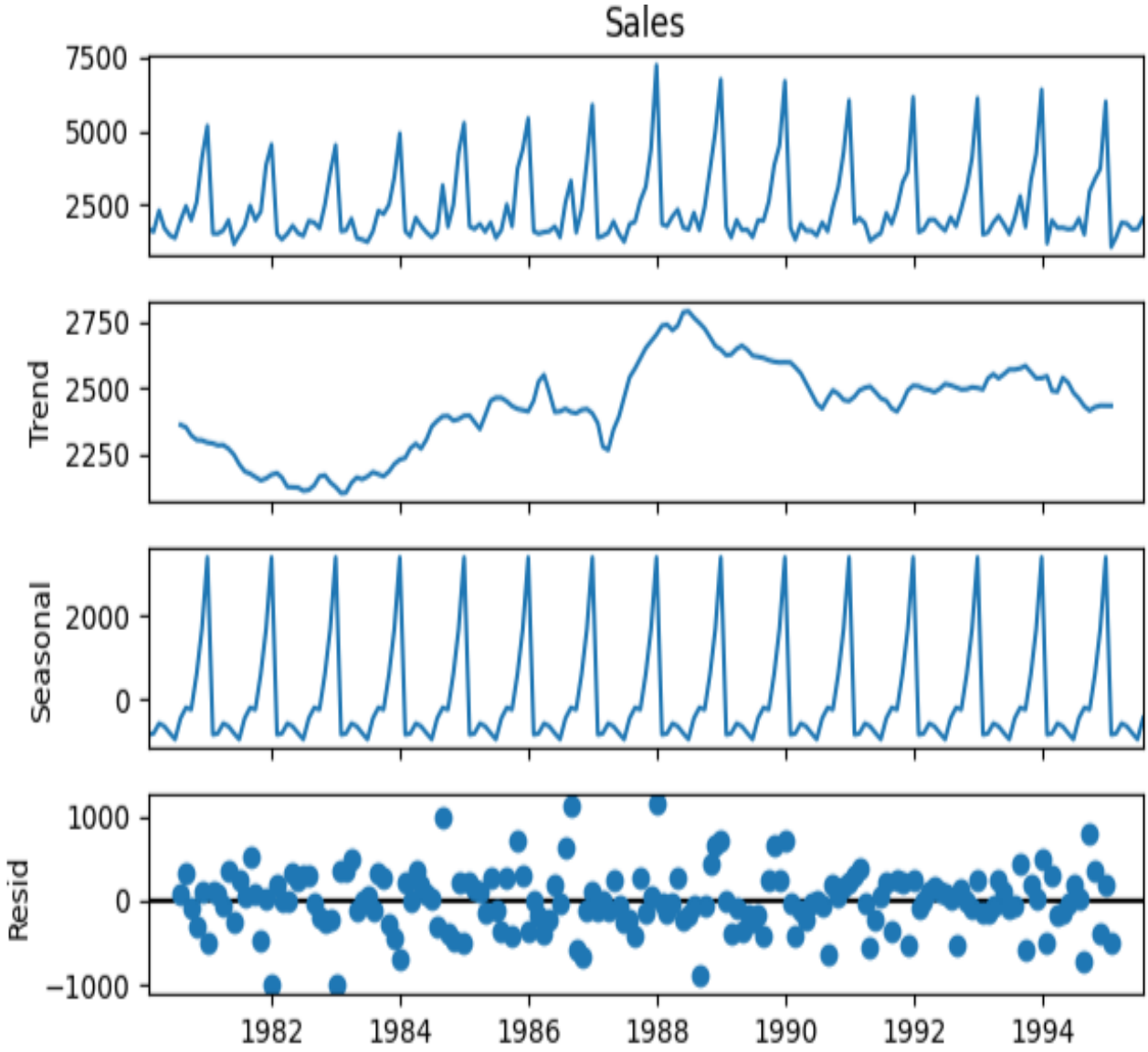

Figure 10: Sparkling Decomposition Plot Addictive

- 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

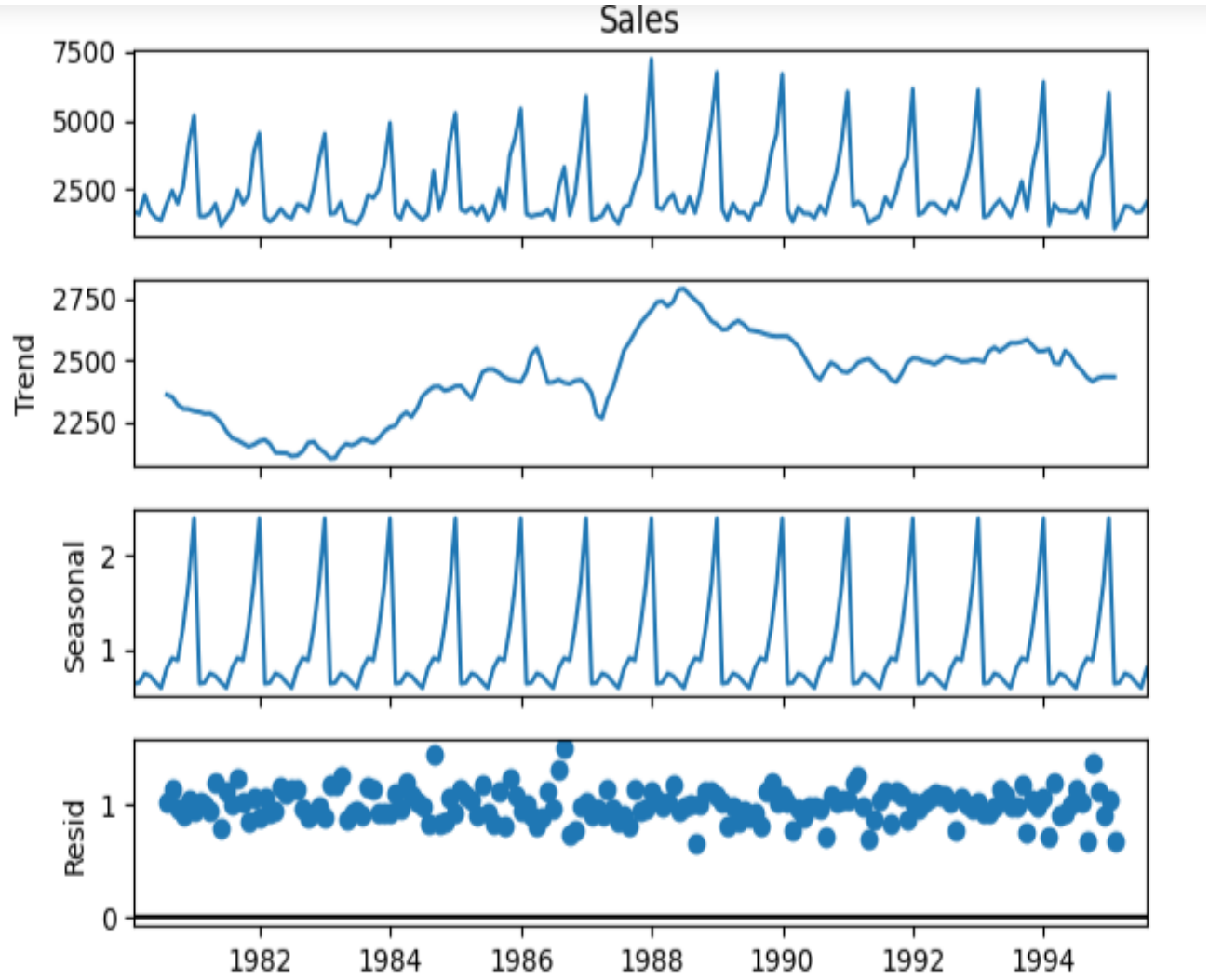

Figure 11: Sparkling Decomposition Plot Multiplicative

- Peak year 1988-1989
- It also shows that the trend has declined over the year after 1988-1989.
- Residue is spread and is in approximate 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

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

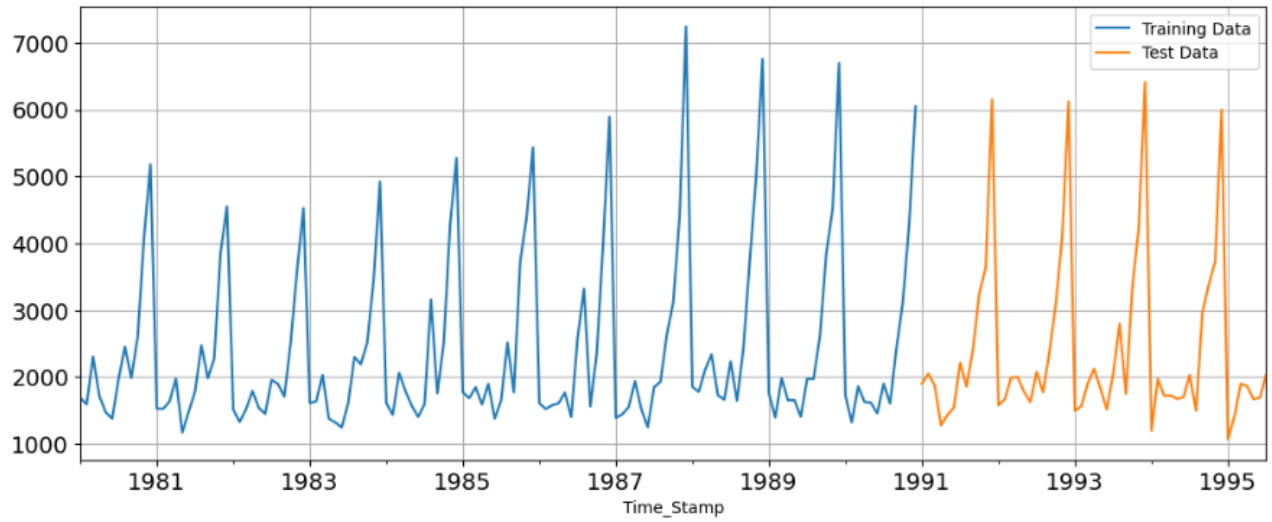

Figure 12: Line Plot of Sparkling Train and Test Dataset

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

Few Rows of Datasets

Train Dataset

First few rows of Training Data

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

Last few rows of Training Data

Time_Stamp	Sales	Year	Month
1990-08-31	1605	1990	8
1990-09-30	2424	1990	9
1990-10-31	3116	1990	10
1990-11-30	4286	1990	11
1990-12-31	6047	1990	12

Test Dataset

First few rows of Test Data

Time_Stamp	Sales	Year	Month
1991-01-31	1902	1991	1
1991-02-28	2049	1991	2
1991-03-31	1874	1991	3
1991-04-30	1279	1991	4
1991-05-31	1432	1991	5

Last few rows of Test Data

Time_Stamp	Sales	Year	Month
1995-03-31	1897	1995	3
1995-04-30	1862	1995	4
1995-05-31	1670	1995	5
1995-06-30	1688	1995	6
1995-07-31	2031	1995	7

Table 5: Rows of Sparkling Train and Test Dataset

1.4 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)

Model 1: Linear Regression

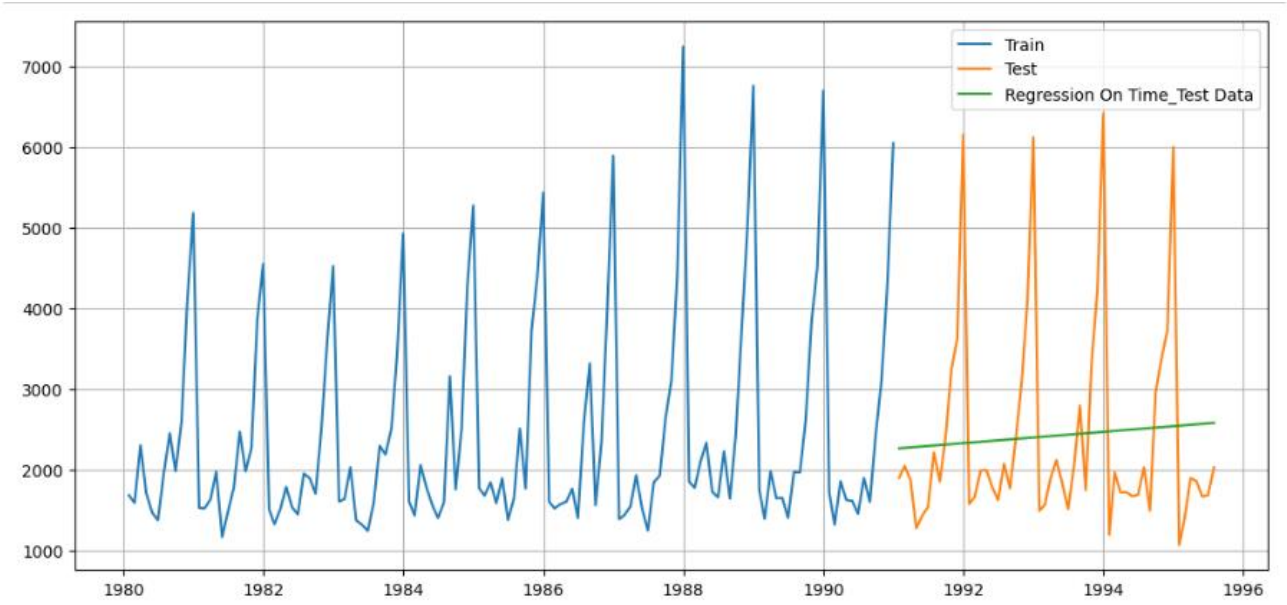

Figure 13: Linear Regression Plot of Sparkling

- 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 Model 3864.279

Model 2: Naive Approach:

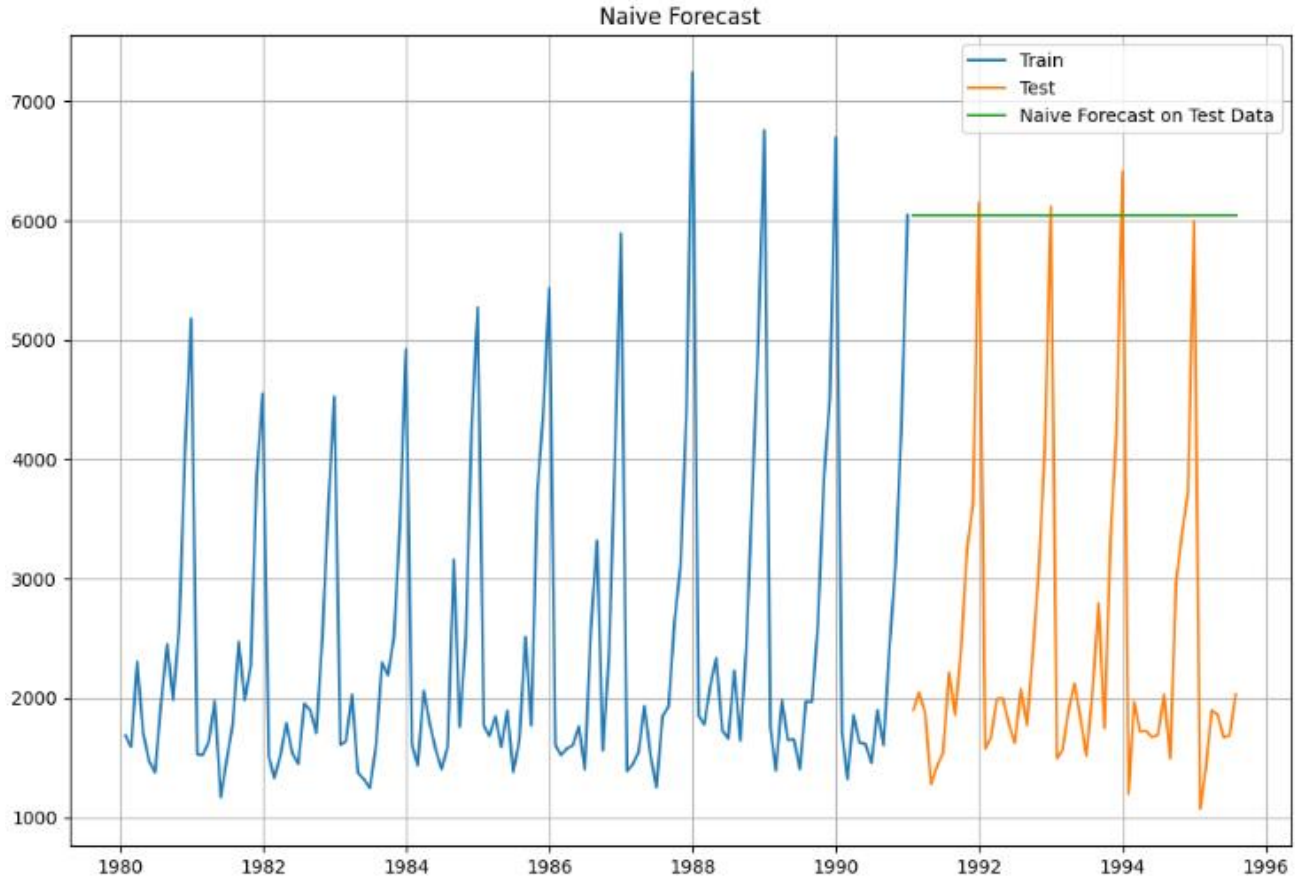

Figure 14: Naïve Approach Plot of Sparkling

- 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.

Naïve Model 3864.279

Method 3: Simple Average

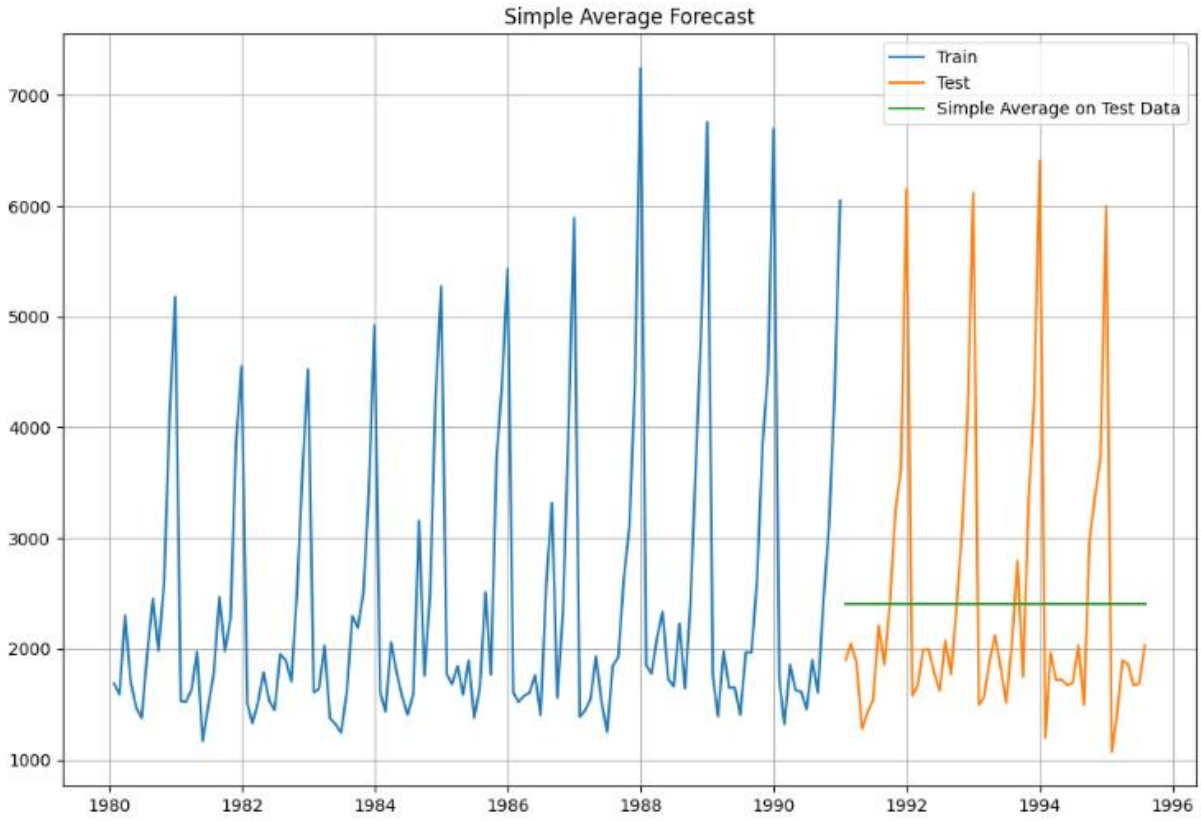

Figure 15: Sparkling Plot of Simple Average

- 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 1275.0818

Method 4: Moving Average (MA)

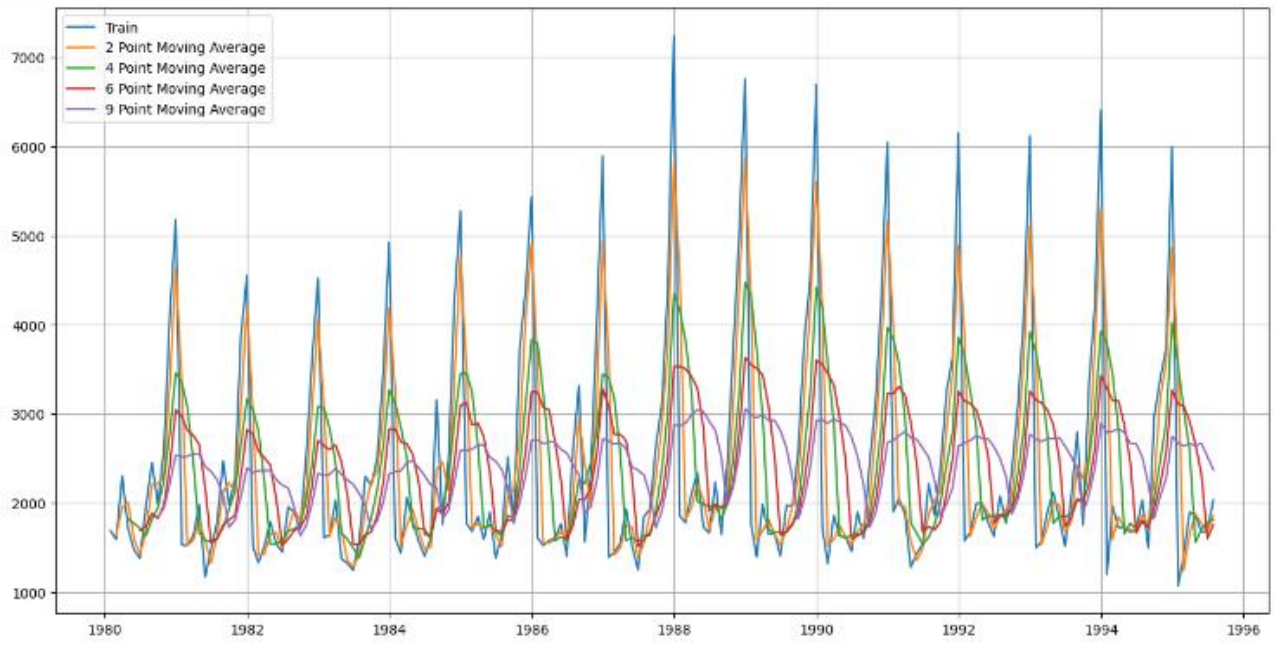

Figure 16: Sparkling Plot of Moving Average

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

2 Point Trailing Moving Average ➔ 813.400684
4 Point Trailing Moving Average ➔ 1156.589694
6 Point Trailing Moving Average ➔ 1283.927428
9 Point Trailing Moving Average ➔ 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 5: Simple Exponential Smoothing

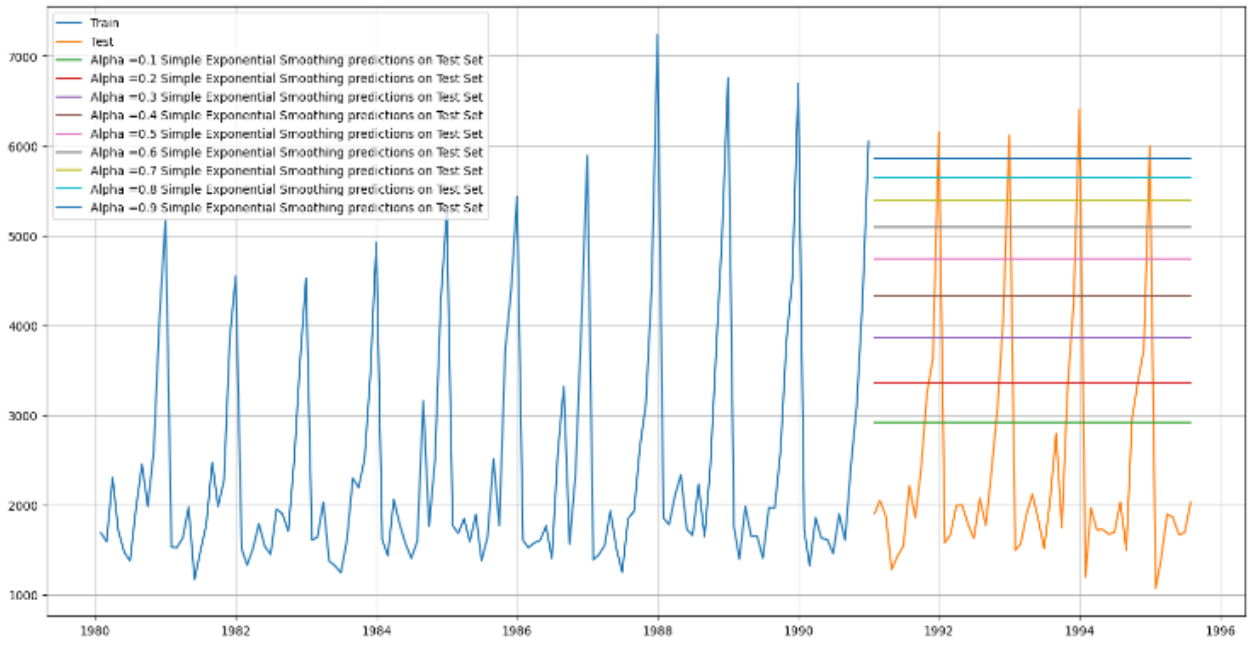

Figure 17: Sparkling Plot of Simple Exponential Smoothing

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

Table 6: Various alpha value and RMSE value of Sparkling Dataset

Alpha Value	Test RMSE
0.1	1375.393398
0.2	1595.206839
0.3	1935.507132
0.4	2311.919615
0.5	2666.351413
0.6	2979.204388
0.7	3249.944092
0.8	3483.801006

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

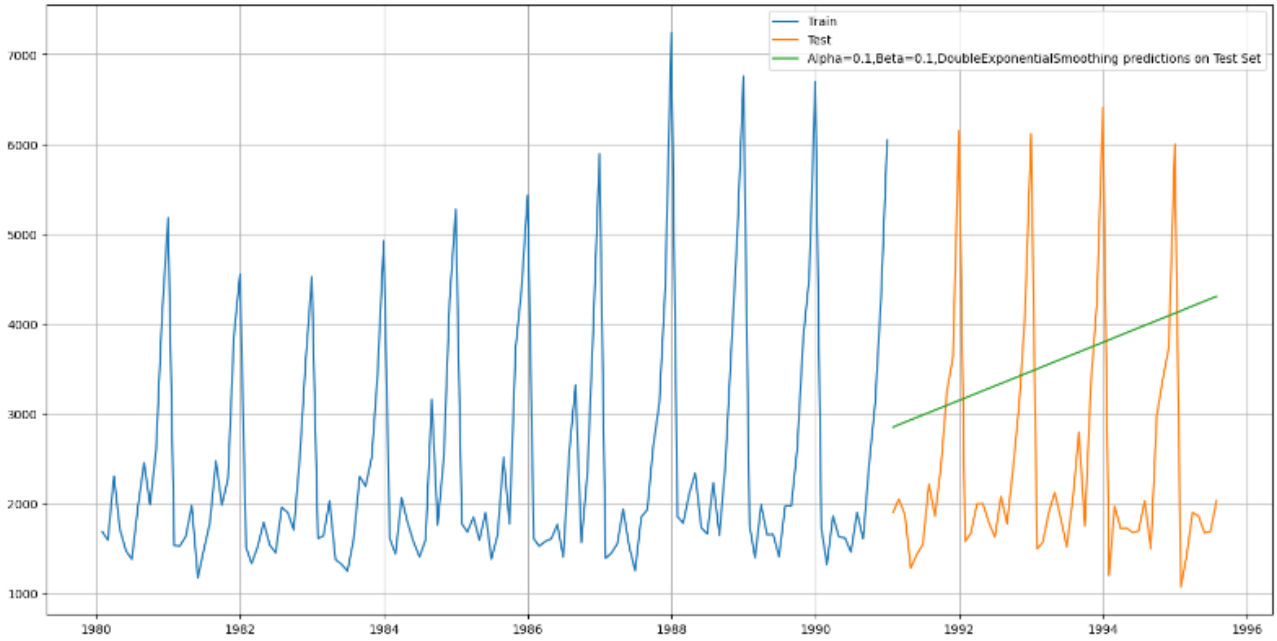

Figure 18: Sparkling Plot of Double Exponential Smoothing

- 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, Double Exponential Smoothing = 1778.564670

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

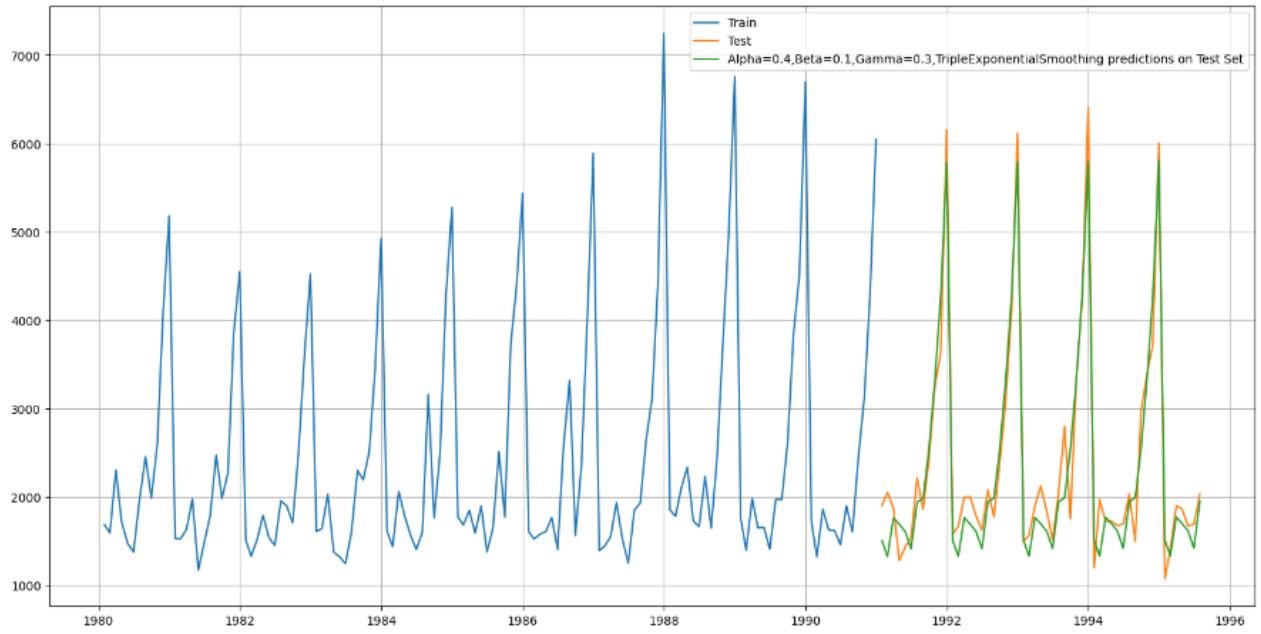

Figure 19: Sparkling Plot of Triple Exponential Smoothing

- Output for a best alpha, beta, and gamma values are shown by the green colour 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, Triple Exponential Smoothing 317.434302

1.5 Apply 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 a 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

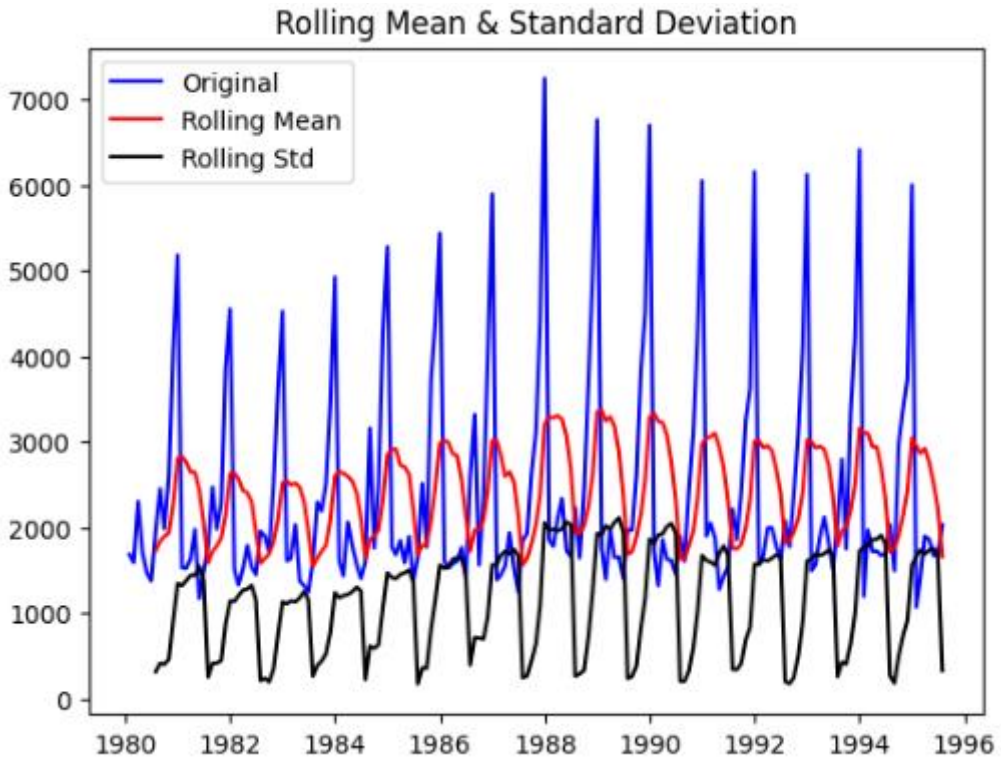

Figure 20: Sparkling Plot for Dicky Fuller Test

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

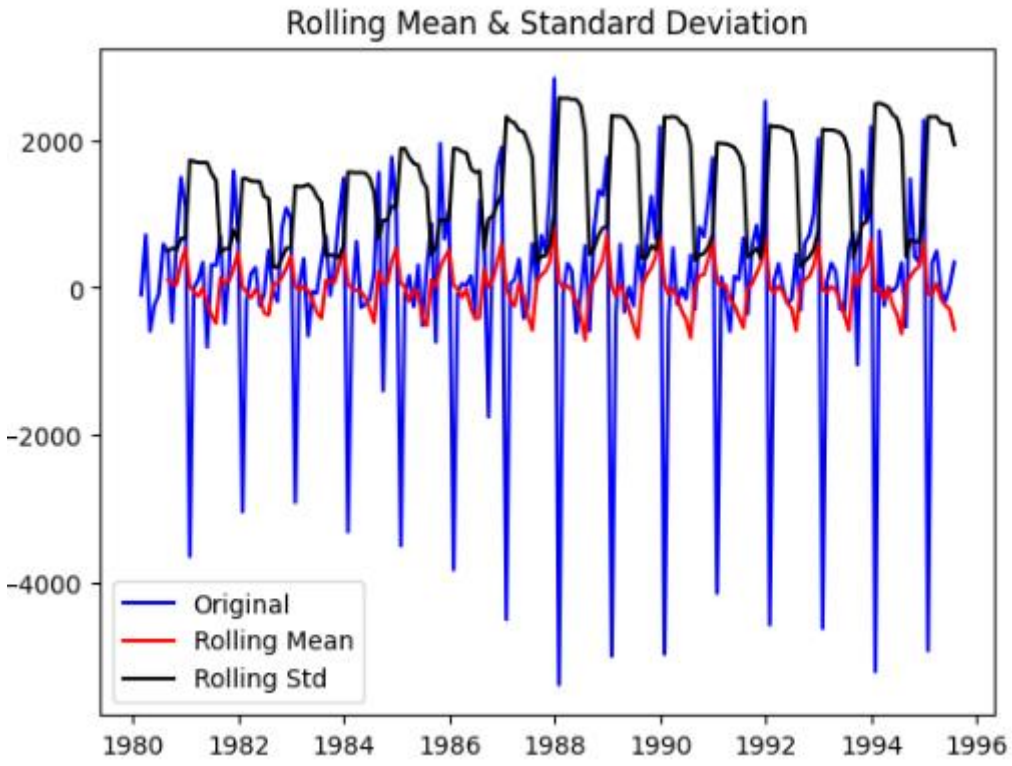

Figure 21: Sparkling Plot for Dicky Fuller Test after differencing approach

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

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

AUTO - 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)	2213.509213
15	(3, 1, 3)	2221.460263
14	(3, 1, 2)	2230.803671
11	(2, 1, 3)	2232.880717
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)	2235.498529
7	(1, 1, 3)	2235.607809
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

Figure 22: Ascending values of AIC for different Param values for Sparkling

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:	Thu, 24 Aug 2023	AIC	2213.509
Time:	08:44:15	BIC	2227.885
Sample:	01-31-1980 - 12-31-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.781	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.740	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.215	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.108	0.000	0.785	1.215
sigma2	1.099e+06	2e-07	5.51e+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

Figure 23: Auto ARIMA Summary Report of Sparkling

RMSE values are as below:

Auto ARIMA 1299.978401

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
87	(1, 1, 1)	(1, 0, 3, 12)	434.151130
247	(3, 1, 3)	(1, 0, 3, 12)	946.108510
220	(3, 1, 1)	(3, 0, 0, 12)	1387.788331
237	(3, 1, 2)	(3, 0, 1, 12)	1388.602612
221	(3, 1, 1)	(3, 0, 1, 12)	1388.681485

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

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Thu, 24 Aug 2023	AIC	1555.584
Time:	09:09:26	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	-4.735	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.840	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

Figure 24: Auto SARIMA Summary Report

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.

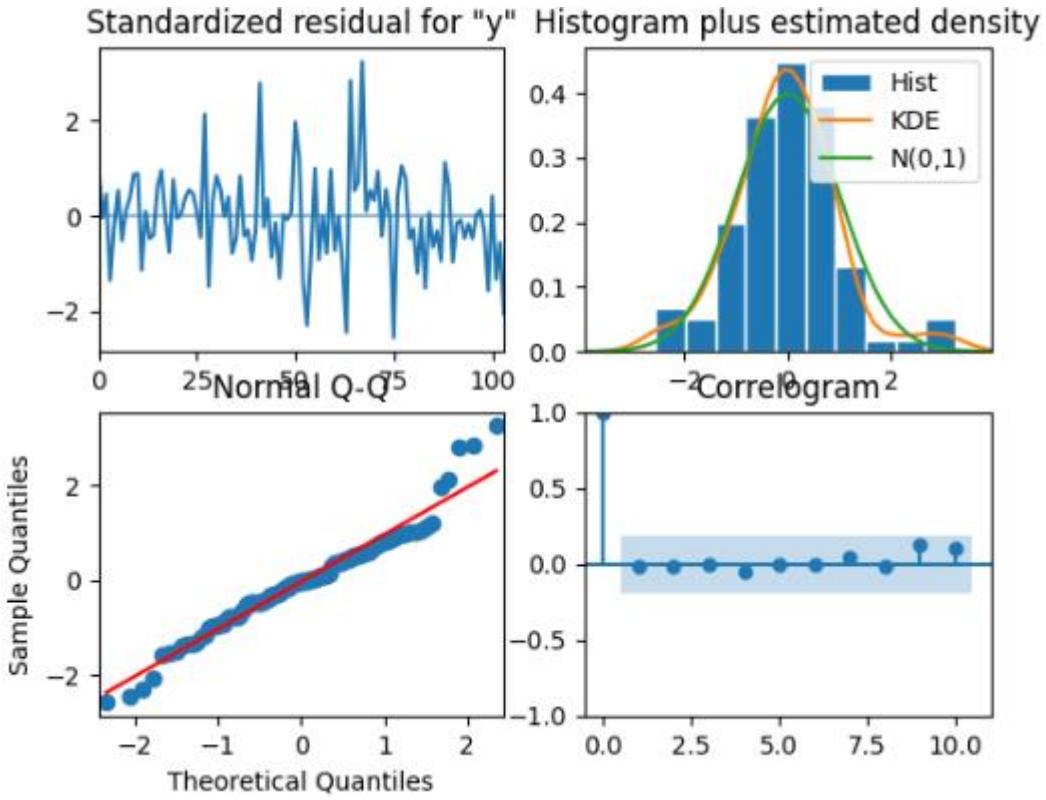

Figure 25: SARIMA Plot

RSME of Model: 528.6069474180102

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

Model	RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	51.080941
2 Point Trailing Moving Average	11.589082
4 Point Trailing Moving Average	14.506190
6 Point Trailing Moving Average	14.558008
9 Point Trailing Moving Average	14.558008
Alpha=0.1, Simple Exponential Smoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, Double Exponential Smoothing	36.510010
Alpha=0.08621, Beta=1.3722, Gamma=0.4763, Tripple Exponential Smoothing Auto Fit	37.192623
Alpha=0.2, Beta=0.7, Gamma=0.2, Triple Exponential Smoothing	8.992350
Auto ARIMA	36.420803

Table 7: Various Model Report for RMSE values

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

Based on the 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

Figure 26: Sparkling Sales Predictions

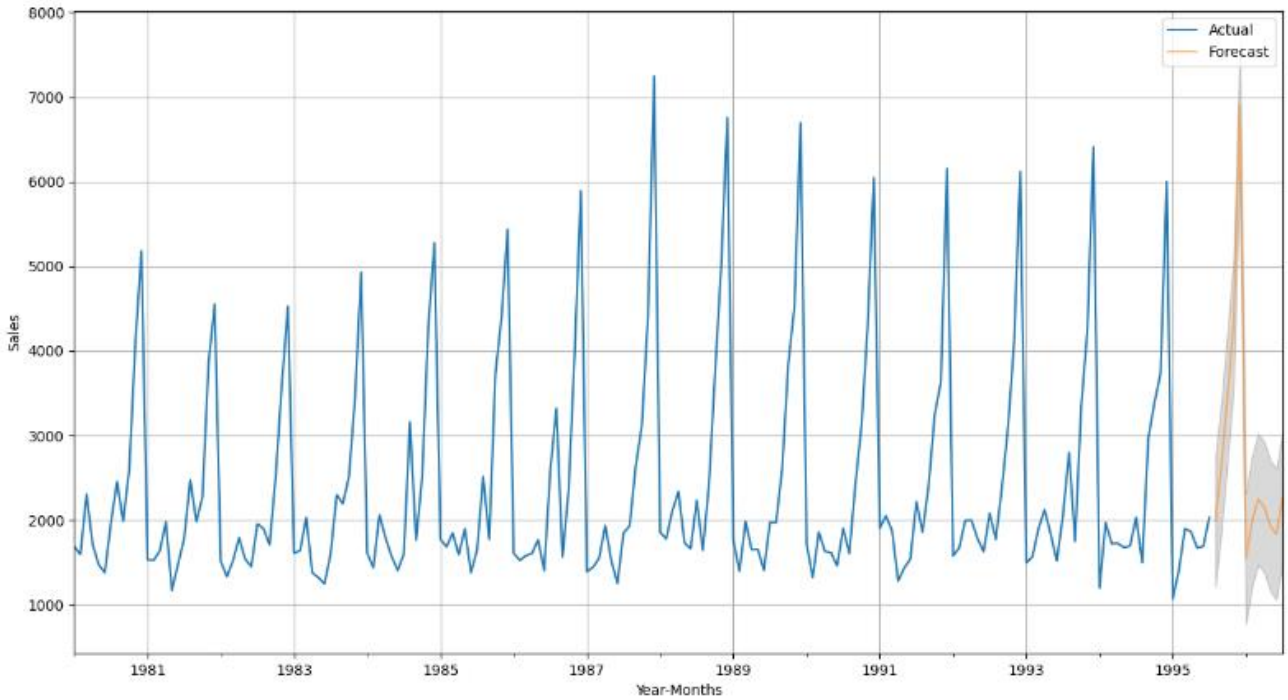

Figure 27: Plot of Sales Predictions

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

Insights:

- 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.

Problem Statement:

ABC Estate Wines has been a leader in the rose wine industry for many years, offering high-quality wines to consumers all around the world. As the company continues to expand its reach and grow its customer base, it is essential to analyse market trends and forecast future sales to ensure continued success.

In this report, we will focus on analysing the sales data for rose wine in the 20th century. As an analyst for ABC Estate Wines, I have been tasked with reviewing this data to identify patterns, trends, and opportunities for growth in the wine market. This knowledge will help us to make informed decisions about how to position our products in the market, optimize our sales strategies, and forecast future sales trends.

Overall, this report aims to provide valuable insights into the wine market and how ABC Estate Wines can continue to succeed in this highly competitive industry.

Time Series Forecast Using Rose Data

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

Data dictionary:

Column	Details
YearMonth	Dates of Sales
Sparkling	Sales of rose wine.

Table 8: Rose Data Dictionary

Rows of Data Set

Top few Rows:

	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

Last few Rows

	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

Table 9: Rose Dataset

Number of Rows and Columns of Dataset:
The dataset has 187 rows and 1 column.

Plot of the Dataset:

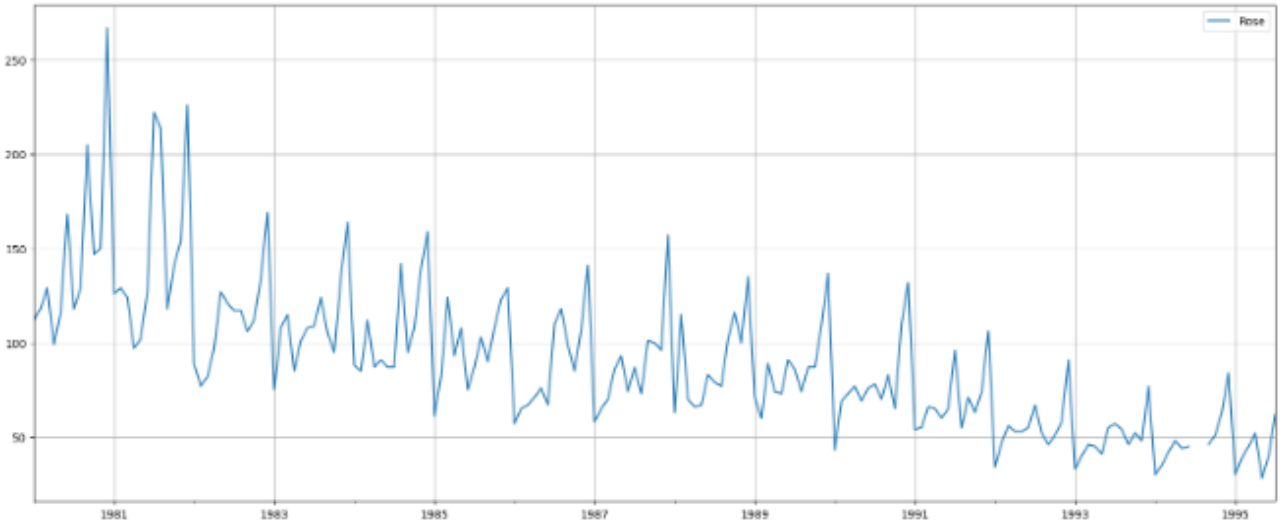

Figure 28: Plot of Rose Sales Dataset

Insights:
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.

Rows of new data set:

Top few Rows of Data set	Last few Rows of Data set

	Sales	Year	Month
1980-01-01	112.0	1980	1
1980-02-01	118.0	1980	2
1980-03-01	129.0	1980	3
1980-04-01	99.0	1980	4
1980-05-01	116.0	1980	5

	Sales	Year	Month
1995-03-01	45.0	1995	3
1995-04-01	52.0	1995	4
1995-05-01	28.0	1995	5
1995-06-01	40.0	1995	6
1995-07-01	62.0	1995	7

Table 10: Rose Data Set after adding Year and Month

Number of Rows and Columns of Dataset: The dataset has 187 rows and 3 columns.

2.2 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:

	count	mean	std	min	25%	50%	75%	max
Sales	185.0	90.0	39.0	28.0	63.0	86.0	112.0	267.0
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

Figure 29: Summary of Rose Sales Dataset

Null Value:
There are 2 null values present in sales the dataset.
We found the values for the months of July & August were missing for the year 1994

	Sales	Year	Month
1994-07-01	NaN	1994	7
1994-08-01	NaN	1994	8

Figure 30: Null values of Rose Sales Dataset

We tried following approaches to impute the data, these were as below.
Mean - Before & After

- Treating null values is very important to do further analysis.
- In this approach, instead of taking means for the 7th months across all the years, we just took mean of the 7th months values from a year before and a year after the missing value.
- Similar steps were taken for 8th month.

Boxplot of dataset:

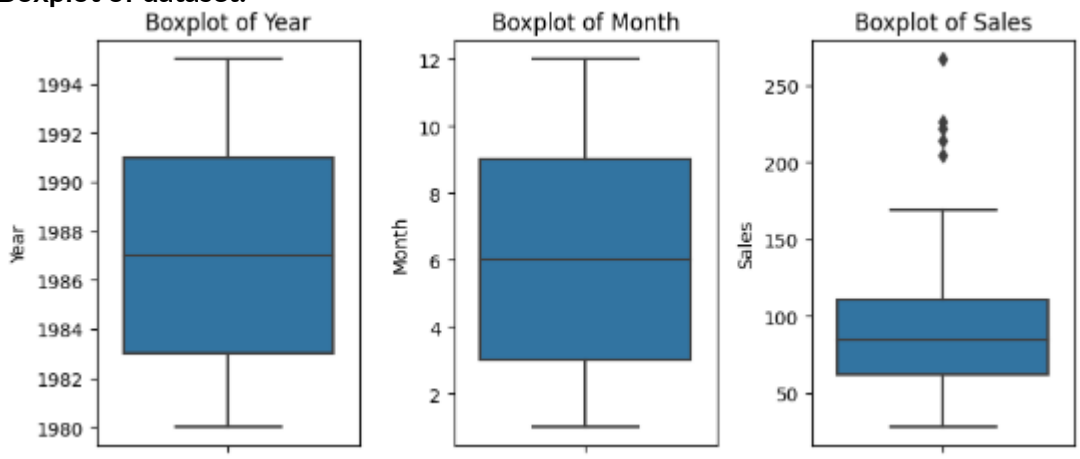

Figure 31: Box Plot of Rose Sales Dataset

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:

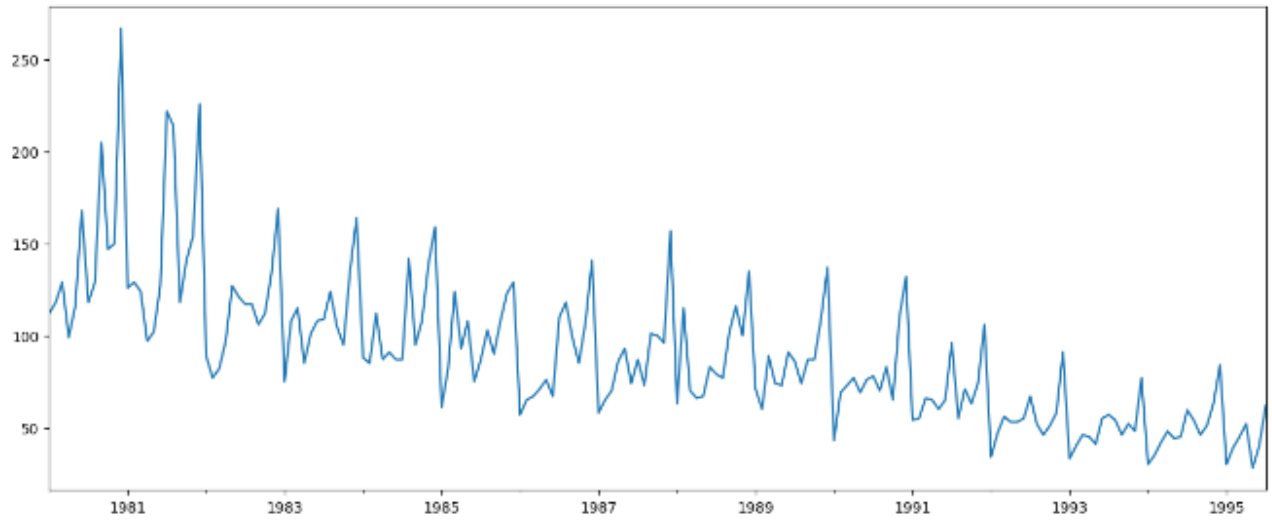

Figure 32: Line Plot of Rose Sales Dataset

- The line plot shows the patterns of trend and seasonality and also shows that there was a peak in the year 1981.

Boxplot Yearly:

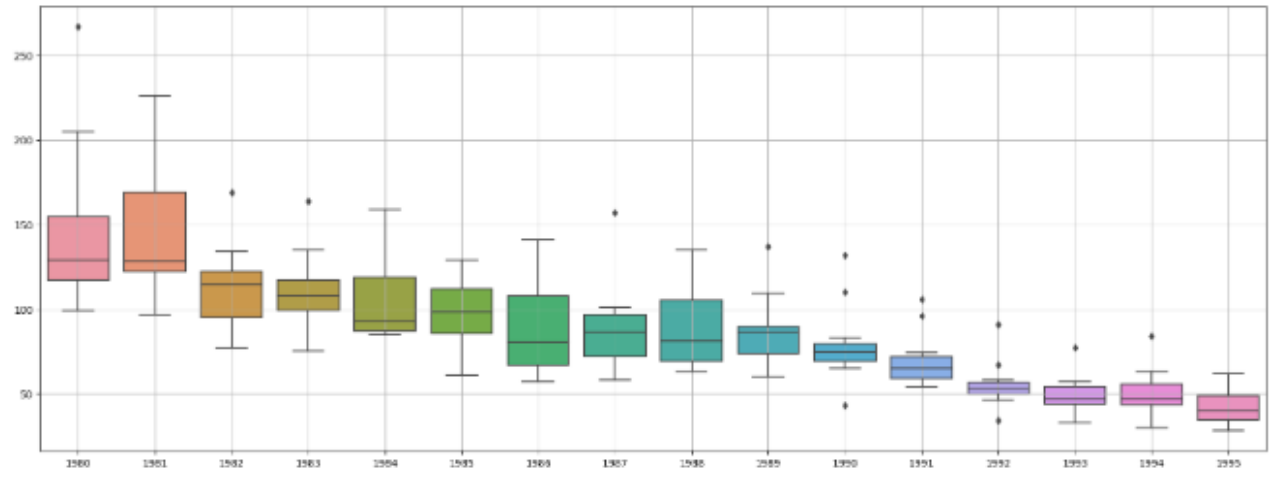

Figure 33: Box Plot of Yearly Rose Sales Dataset

- This yearly box plot shows there is consistency over the years and there was a peak in 1980-1981. Outliers are present in almost all years.

Boxplot Monthly:

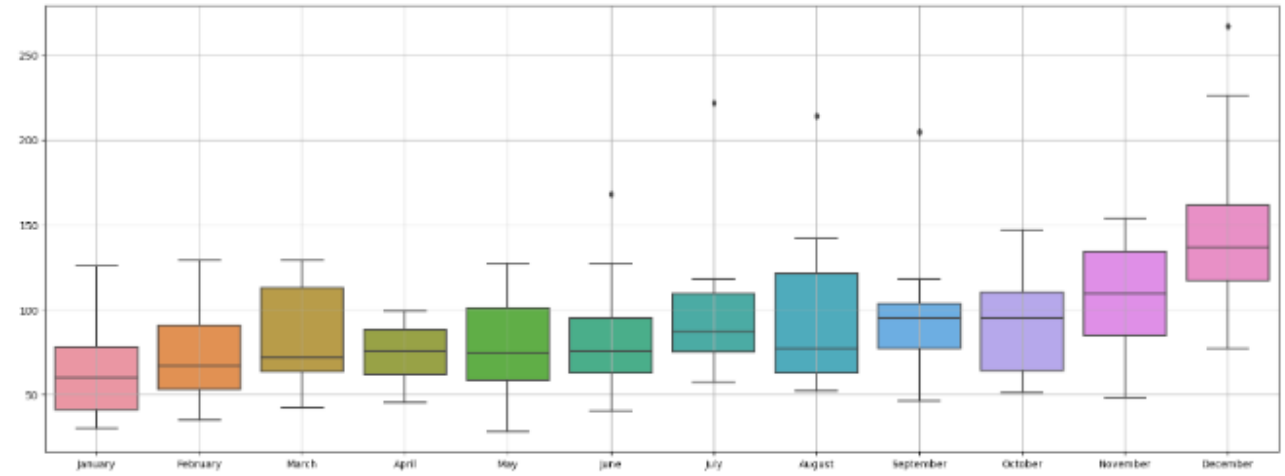

Figure 34: Box Plot of Monthly Rose Sales Dataset

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 June, July, august, September and December.

Boxplot Weekday vise:

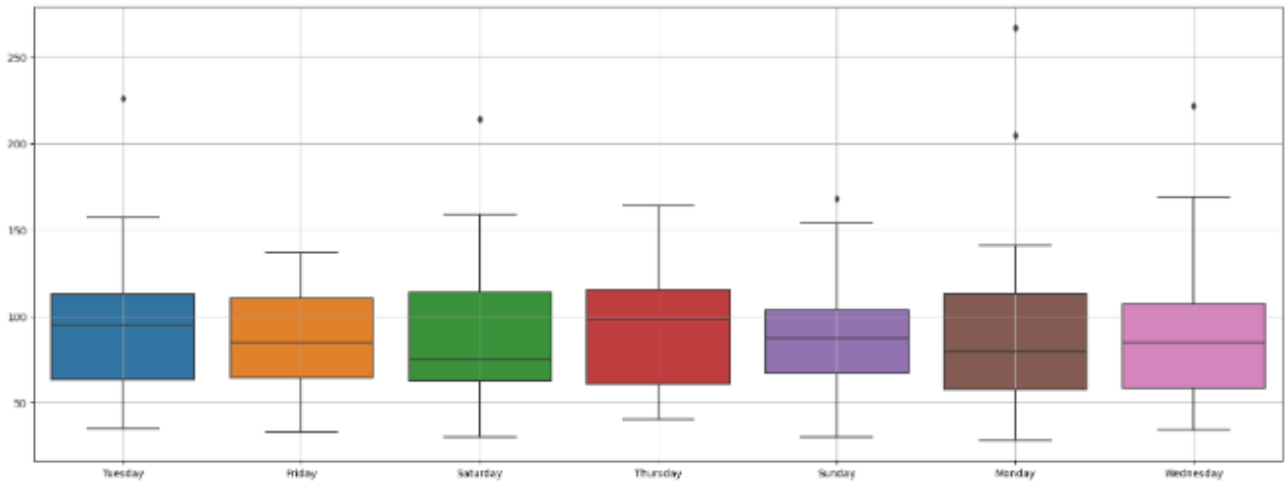

Figure 35: Box Plot of Weekdayvise Rose Sales Dataset

Tuesday has more sales than other days and Wednesday has the lowest sales of the week. Outliers are present on all days except Friday and Thursday

Graph of Monthly Sales over the years:

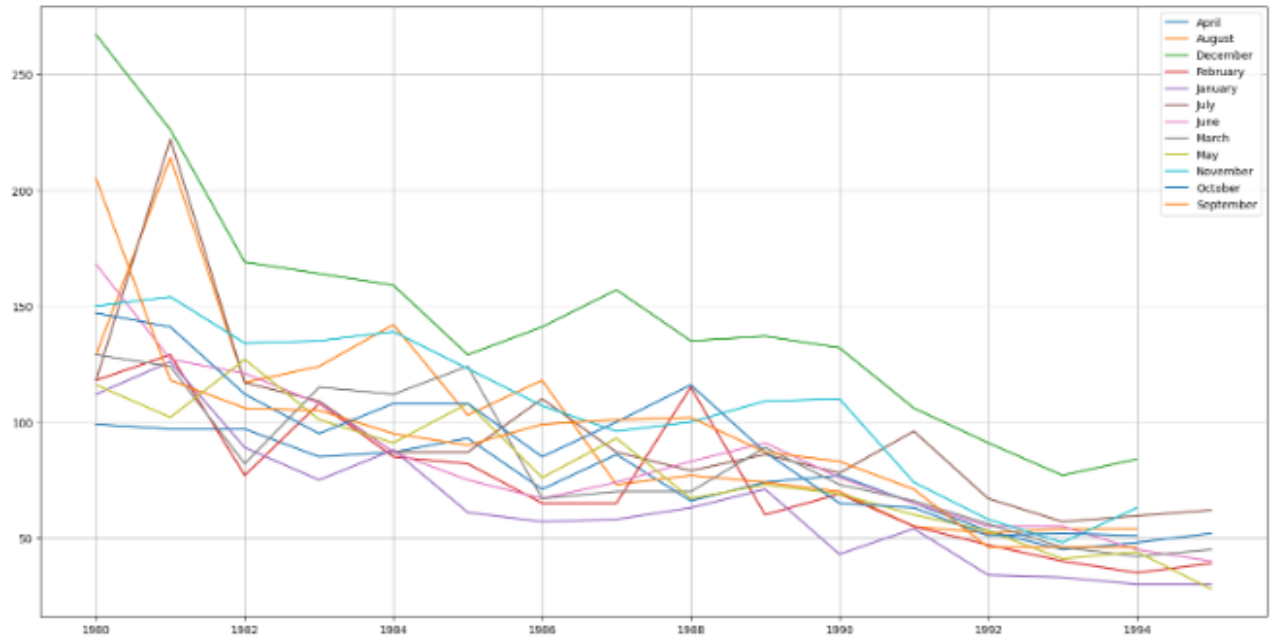

Figure 36: Graph of Monthly Rose Sales Over the years

This plot shows that December has the highest sales over the years and the year 1981 was the year with the highest number of sales.

Correlation plot

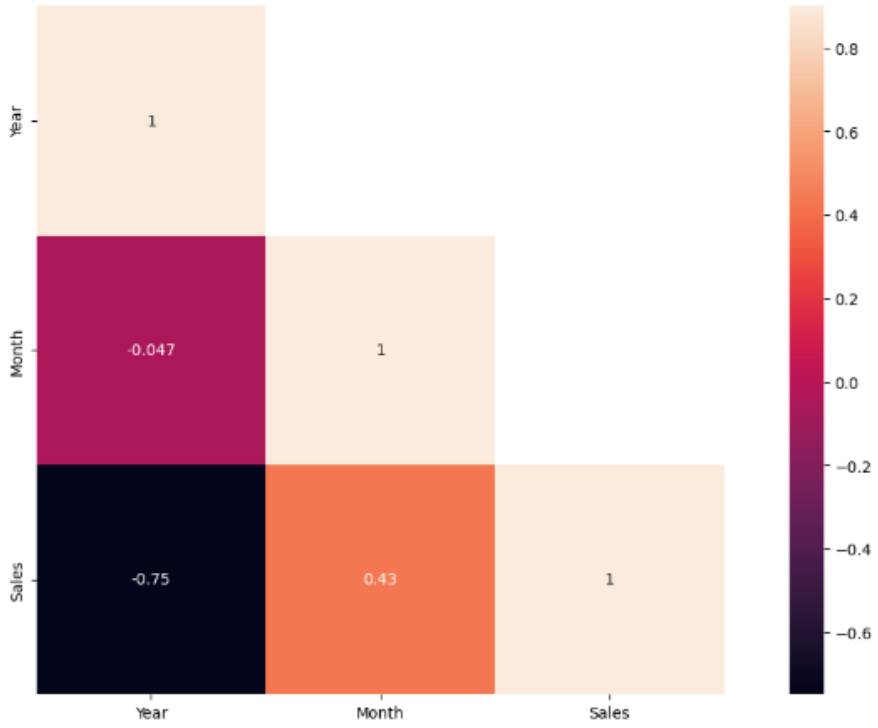

Figure 37: Correlation Plot of Rose Sales

- This heat map shows that there was little correlation between Sales and the Years data, there significantly more correlation between the month and Sales columns.
- Clearly indicating a seasonal pattern in our Sales data. Certain months have higher sales, while certain months have lesser

Decomposition -Additive

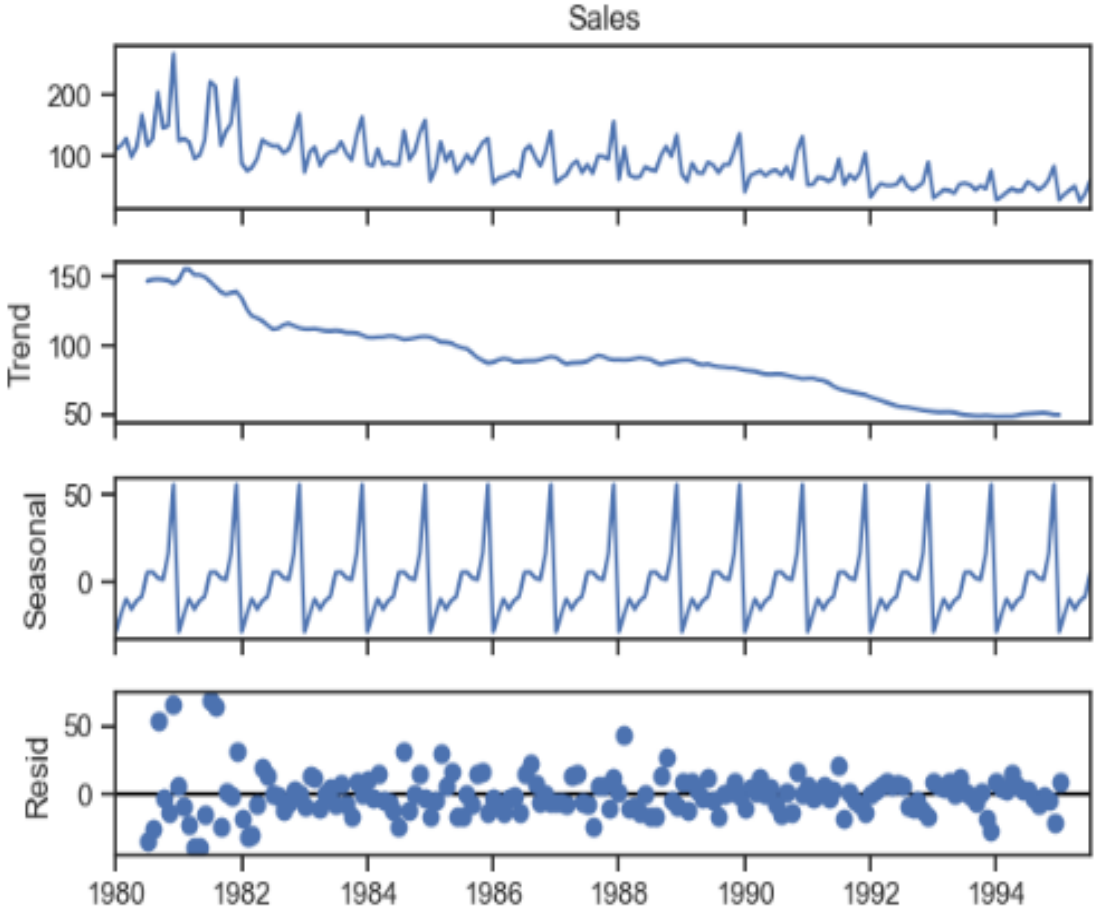

Figure 38: Decomposition Additive Rose Sales

The plots show:

- Peak year 1981.
- It also shows that the trend has declined over the year after 1981.
- Residue is spread and is not in a straight line.
- Both trend and seasonality are present.

Decomposition-Multiplicative

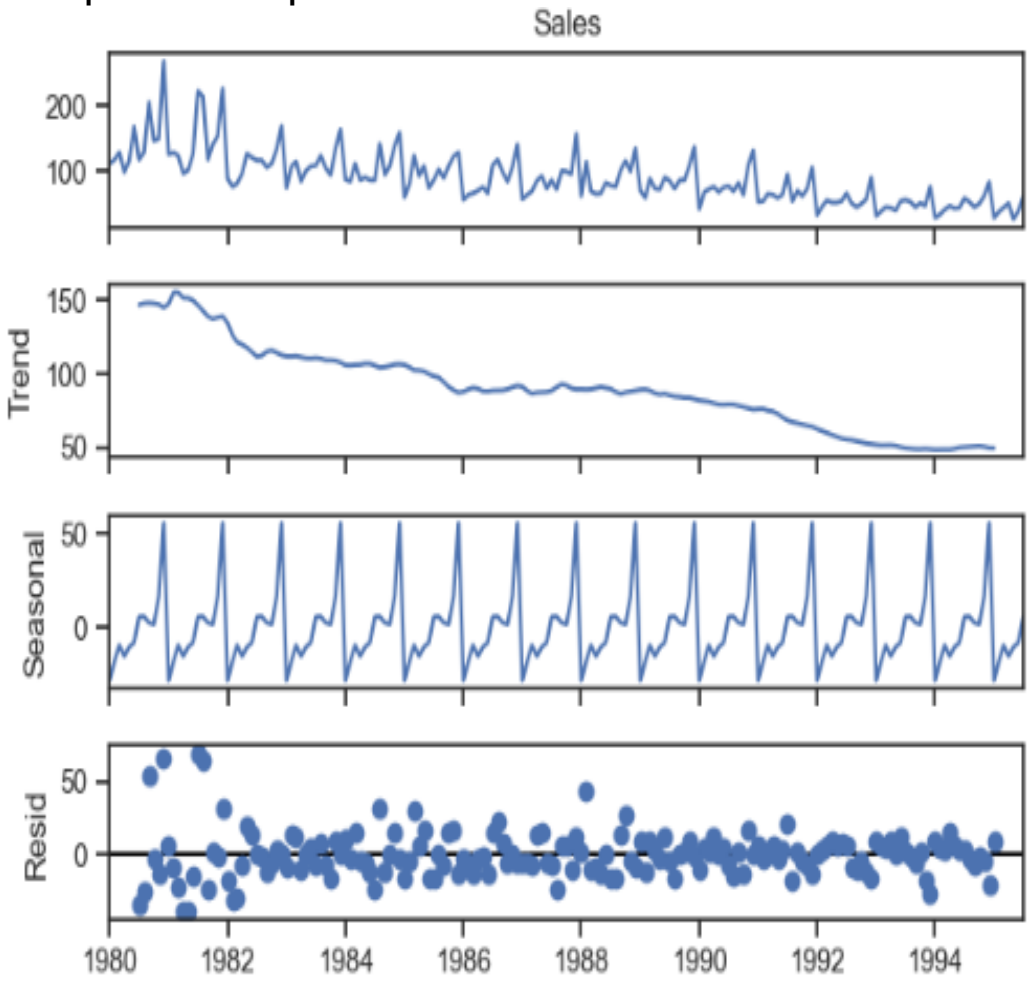

Figure 39: Decomposition Additive Rose Sales

The plots show:

- Peak year 1981.
- It also shows that the trend has declined over the year after 1981.
- Residue is spread and is in approx a straight line.
- Both trend and seasonality are present.
- Reside is 0 to 1, while for additive is 0 to 50.
- So multiplicative model is selected owing to a more stable residual plot and lower range of residuals

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

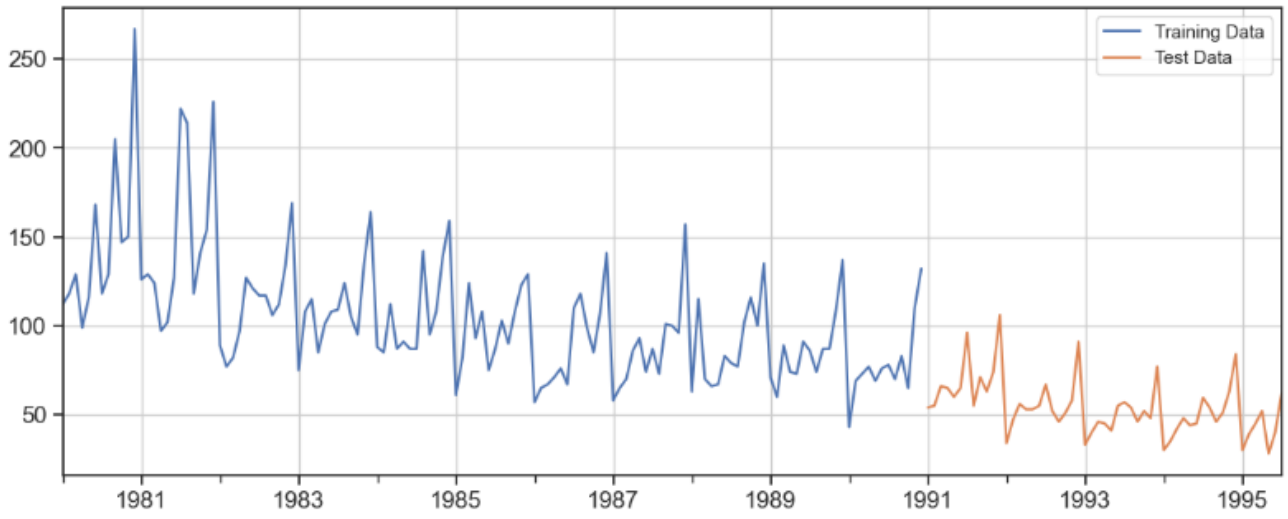

Figure 40: Training and Test data set of Rose Sales

Data split from 1980-1990 is training data, then 1991 to 1995 is training data.

Rows and Columns:
Train dataset has 132 rows and 3 columns.
Test dataset has 55 and 3 columns.

Few Rows of datasets: Table 5: train and test dataset rows

Train Dataset

First few rows of Training Data

	Year	Month	Sales
1980-01-01	1980	1	112.0
1980-02-01	1980	2	118.0
1980-03-01	1980	3	129.0
1980-04-01	1980	4	99.0
1980-05-01	1980	5	116.0

Last few rows of Training Data

	Year	Month	Sales
1990-08-01	1990	8	70.0
1990-09-01	1990	9	83.0
1990-10-01	1990	10	65.0
1990-11-01	1990	11	110.0
1990-12-01	1990	12	132.0

Test Dataset

First few rows of Test Data

	Year	Month	Sales
1991-01-01	1991	1	54.0
1991-02-01	1991	2	55.0
1991-03-01	1991	3	66.0
1991-04-01	1991	4	65.0
1991-05-01	1991	5	60.0

Last few rows of Test Data

	Year	Month	Sales
1995-03-01	1995	3	45.0
1995-04-01	1995	4	52.0
1995-05-01	1995	5	28.0
1995-06-01	1995	6	40.0
1995-07-01	1995	7	62.0

Table 11: Rose Data Set after splitting Train and Test

2.4 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)

Model 1: Linear Regression

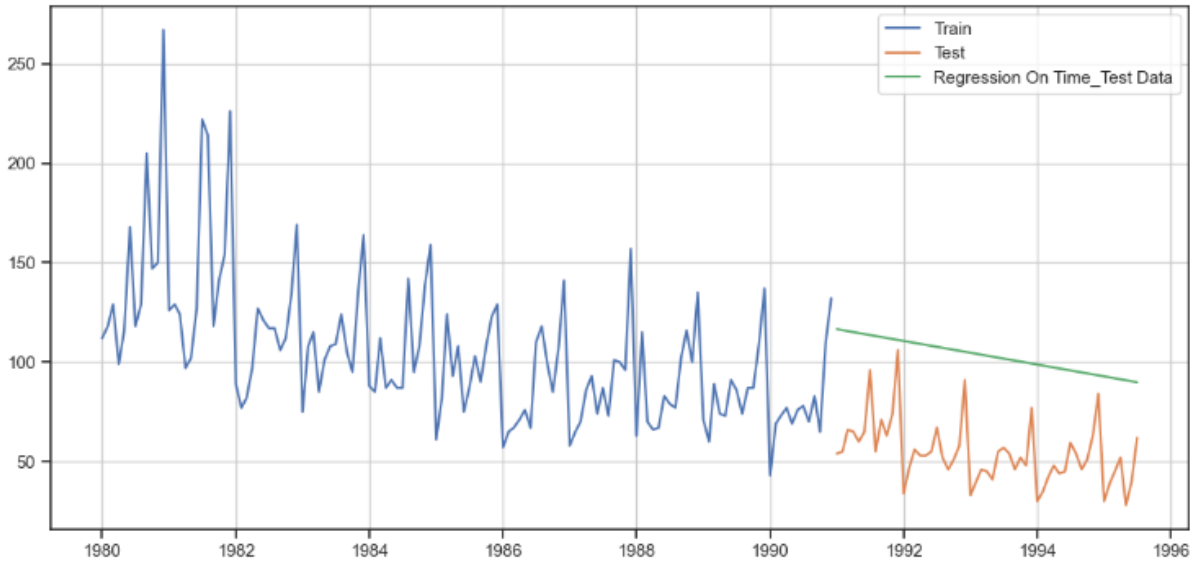

Figure 41: Linear Regression Plot of Rose Sales

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 51.800941

Model 2: Naive Approach

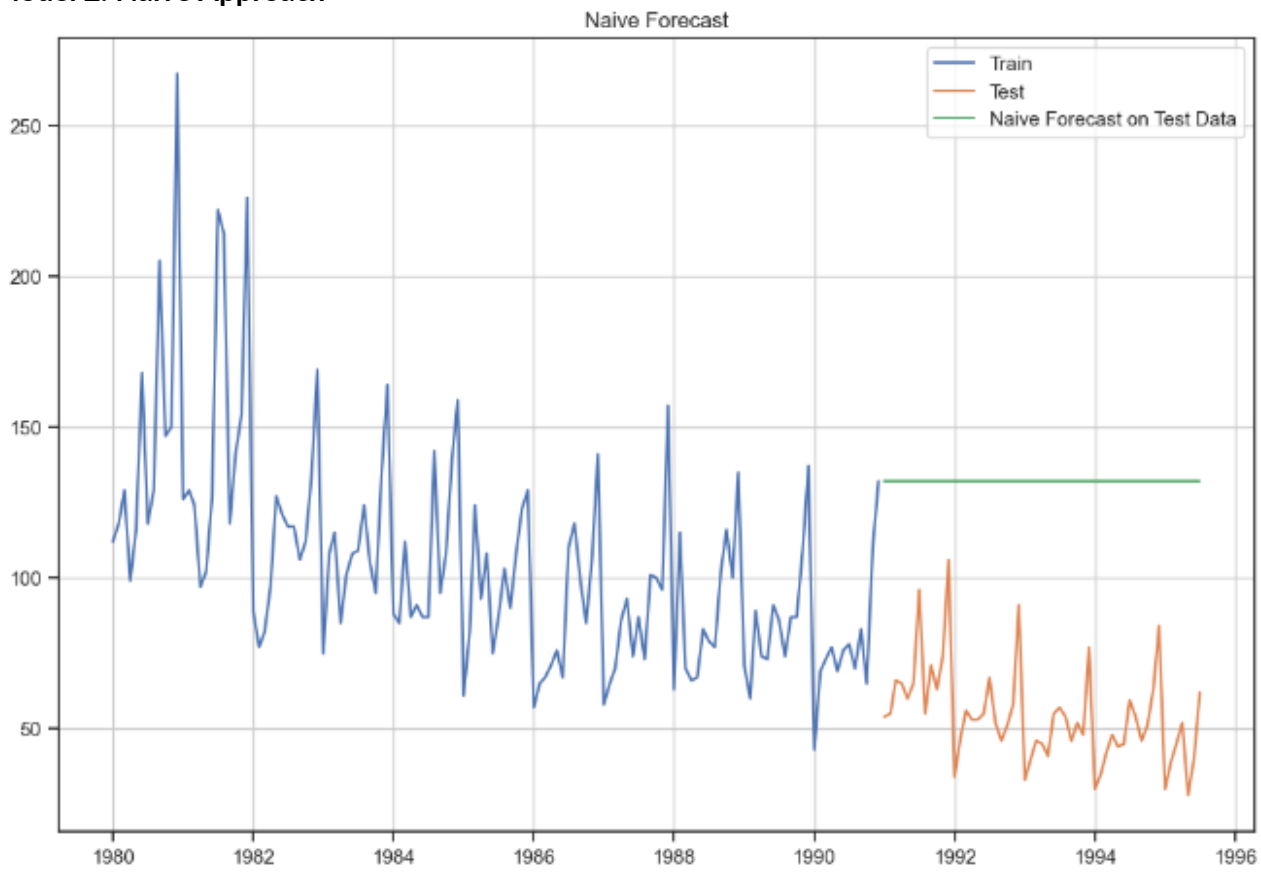

Figure 42: Naive Approach Plot of Rose Sales

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.

Naïve Model 79.304391

Model 3: Simple Average

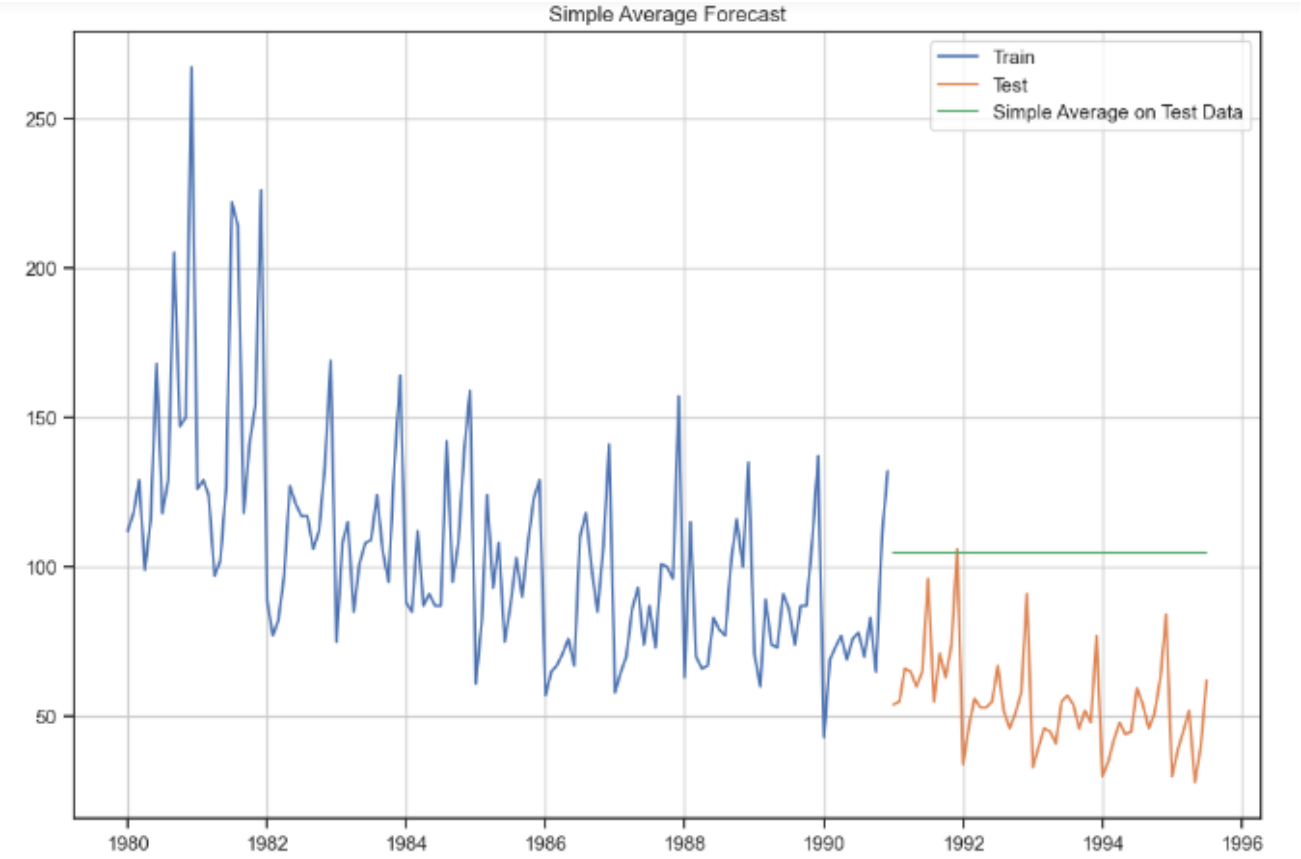

Figure 43: Simple Average Plot of Rose Sales

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 53.049755

Method 4: Moving Average (MA)

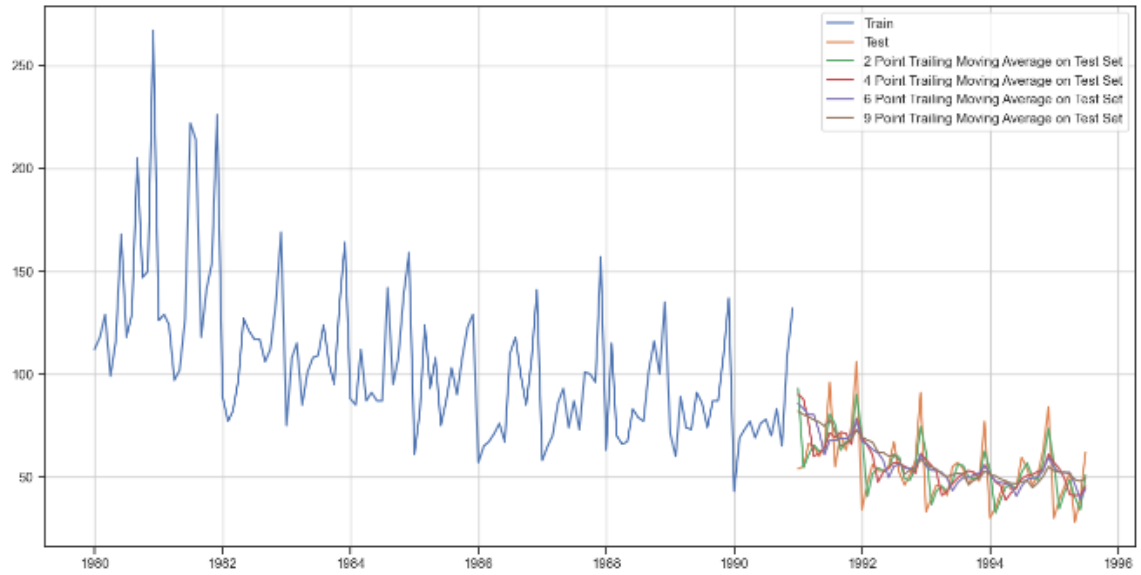

Figure 44: Moving Average Plot of Rose Sales

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

2 Point Trailing Moving Average 11.589082
4 Point Trailing Moving Average 14.506190
6 Point Trailing Moving Average 14.558008
9 Point Trailing Moving Average 14.797139

- We created 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. Higher the rolling window, smoother will be its curve, since more values are being taken into account.

Model 5: Simple Exponential Smoothing

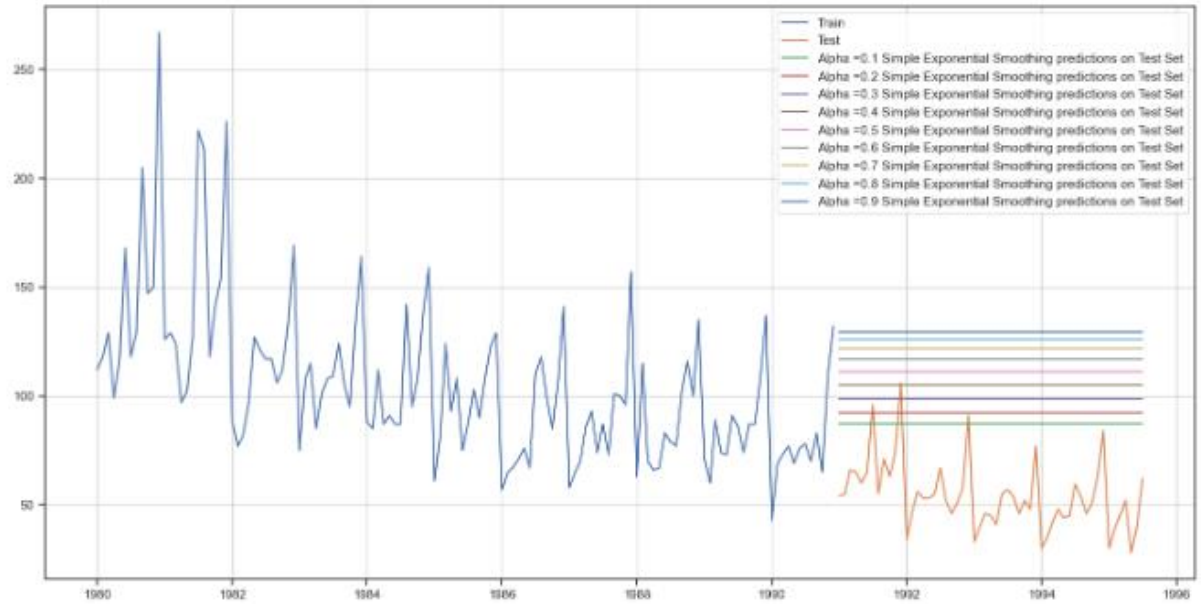

Figure 45: Simple Exponential Smoothing Plot of Rose Sales

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

Alpha=0.1, Simple Exponential Smoothing Model 36.429535

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

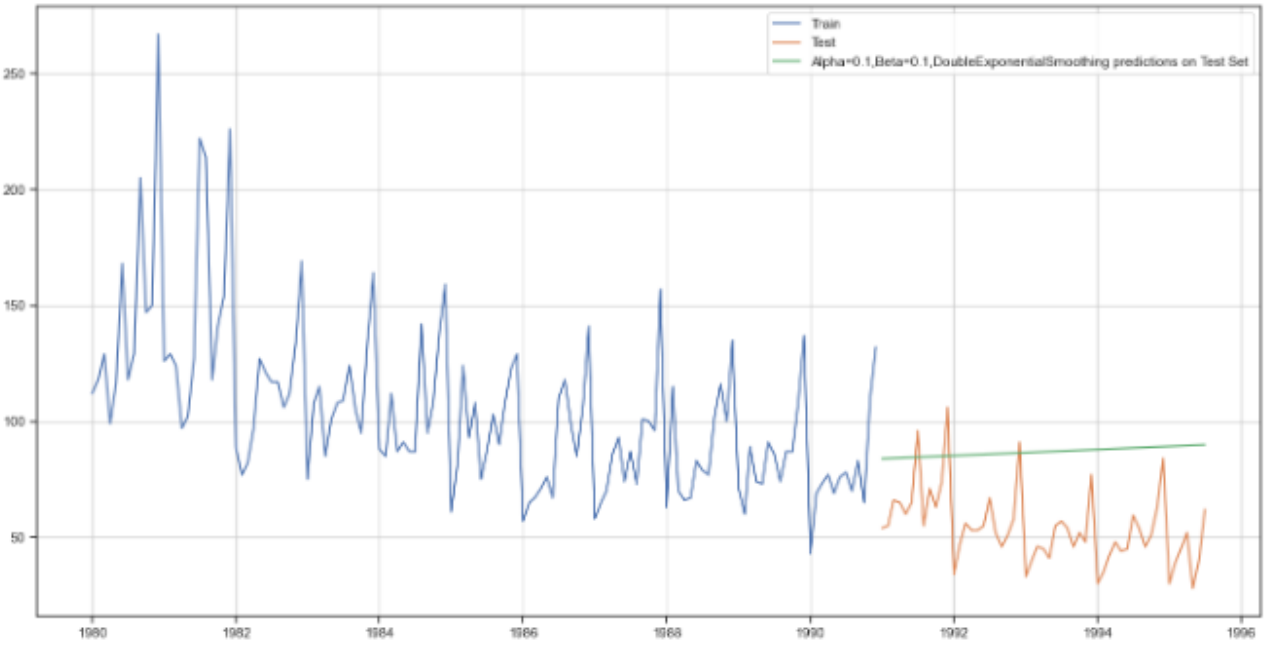

Figure 46: Double Exponential Smoothing Plot of Rose Sales

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

Alpha=0.1, Beta=0.1, Double Exponential Smoothing Model 36.510010

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

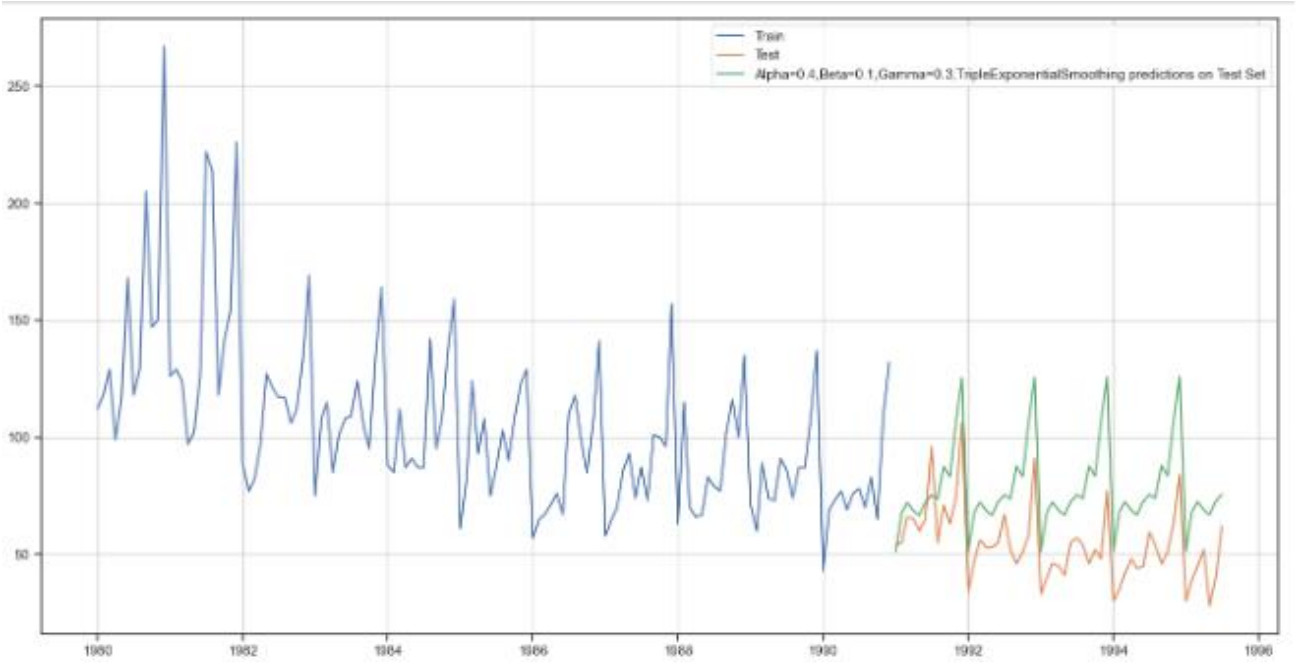

Figure 47: Triple Exponential Smoothing Plot of Rose Sales

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

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

Alpha=0.4, Beta=0.1, Gamma=0.3, Triple Exponential Smoothing Model 36.510010

2.5 Apply 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 a 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.

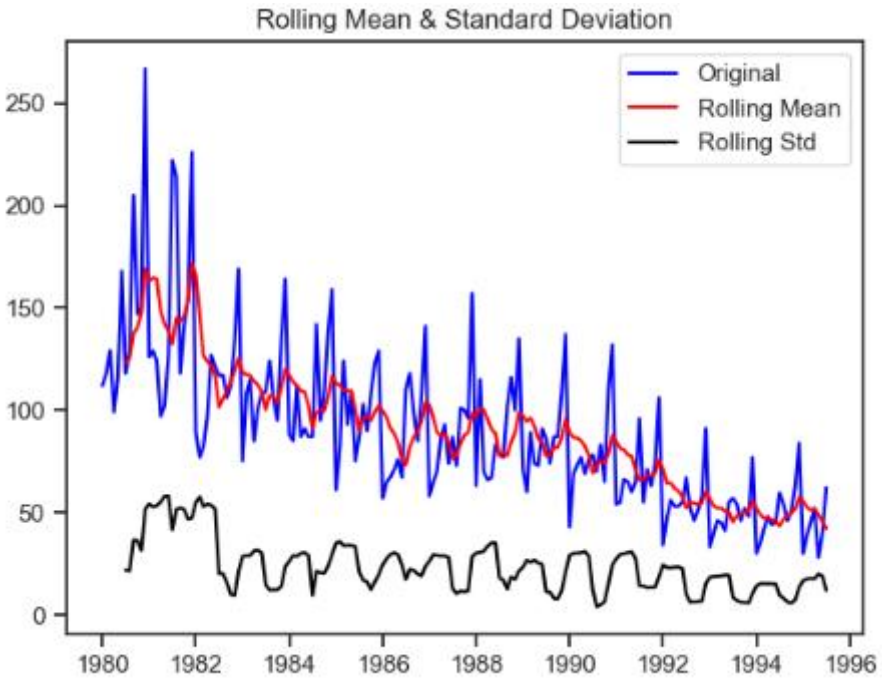

Figure 48: Rolling Mean & Standard Deviation Plot of Rose

Results of Dickey-Fuller Test:
Test Statistic -1.892338
p-value 0.335674 23
we failed to reject the null hypothesis, which implies the Series is not stationary in nature. 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.

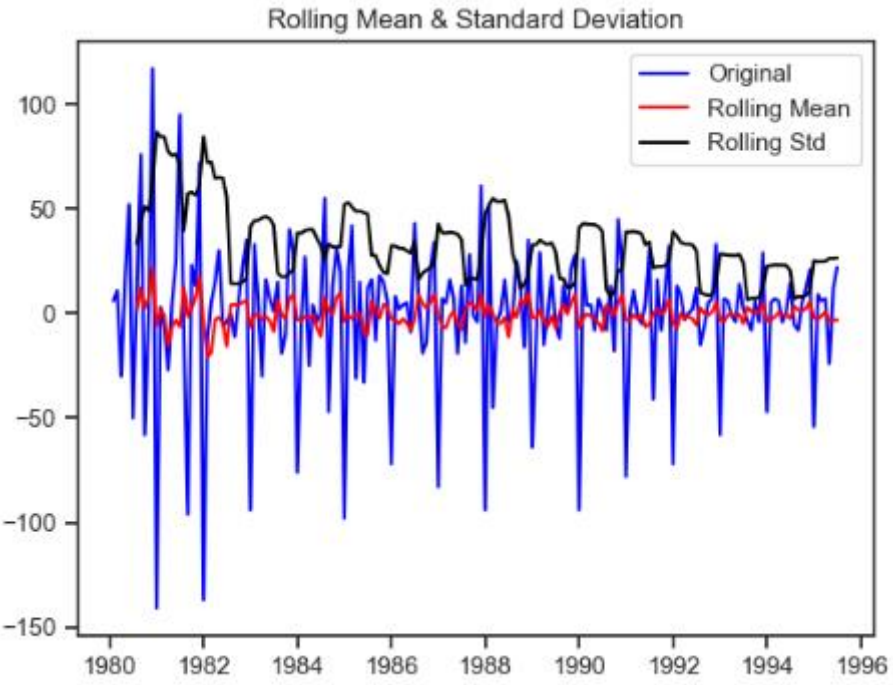

Figure 49: Dicky Fuller Test after diff

Results of Dickey-Fuller Test:
Test Statistic -8.032729e+00
p-value 1.938803e-12

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.

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

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

AUTO - ARIMA 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
11	(2, 1, 3)	1274.694964
15	(3, 1, 3)	1278.668417
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.574230
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

Figure 50: AIC value of different Params

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, 3)	Log Likelihood	-631.347
Date:	Fri, 25 Aug 2023	AIC	1274.695
Time:	21:27:09	BIC	1291.946
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.705
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6781	0.084	-20.037	0.000	-1.842	-1.514
ar.L2	-0.7289	0.084	-8.704	0.000	-0.893	-0.565
ma.L1	1.0449	0.674	1.550	0.121	-0.277	2.367
ma.L2	-0.7716	0.136	-5.669	0.000	-1.038	-0.505
ma.L3	-0.9046	0.612	-1.478	0.139	-2.104	0.295
sigma2	858.3737	567.984	1.511	0.131	-254.854	1971.602

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.44
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

Figure 51: ARIMA Summary

RMSE values are as below: 36.42079120523518

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
222	(3, 1, 1)	(3, 0, 2, 12)	774.400285
238	(3, 1, 2)	(3, 0, 2, 12)	774.880938
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.495330
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018

The summary report for the best SARIMA model with values (3,1,1) (3,0,2,12)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)   Log Likelihood                -377.200
Date:                                 Fri, 25 Aug 2023   AIC                            774.400
Time:                                         22:02:53   BIC                            799.618
Sample:                                              0   HQIC                           784.578
                                                   - 132
Covariance Type:                                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0464      0.126      0.367      0.714      -0.202       0.294
ar.L2         -0.0060      0.120     -0.050      0.960      -0.241       0.229
ar.L3         -0.1808      0.098     -1.837      0.066      -0.374       0.012
ma.L1         -0.9370      0.067    -13.904      0.000      -1.069      -0.805
ar.S.L12       0.7639      0.165      4.639      0.000       0.441       1.087
ar.S.L24       0.0840      0.159      0.527      0.598      -0.229       0.397
ar.S.L36       0.0727      0.095      0.764      0.445      -0.114       0.259
ma.S.L12      -0.4968      0.250     -1.988      0.047      -0.987      -0.007
ma.S.L24      -0.2191      0.210     -1.044      0.296      -0.630       0.192
sigma2       192.1606     39.630      4.849      0.000     114.487     269.834
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                 1.64
Prob(Q):                              0.58   Prob(JB):                         0.44
Heteroskedasticity (H):               1.11   Skew:                             0.33
Prob(H) (two-sided):                  0.77   Kurtosis:                         3.03
===================================================================================

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

Figure 52: SARIMA Summary

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.

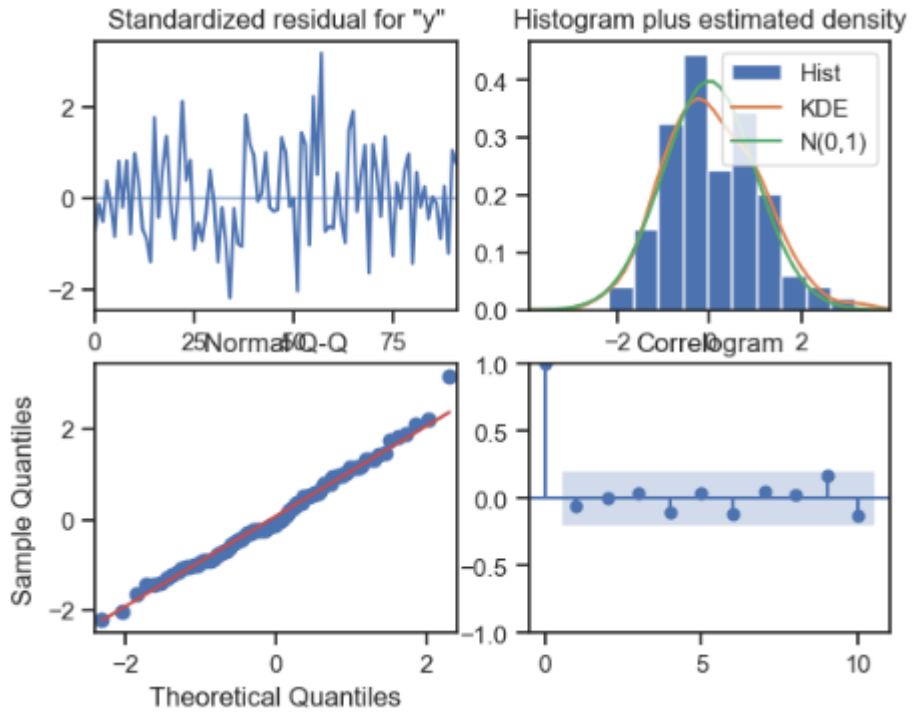

Figure 53: SARIMA Plot

RSME of Model: 18.53502803217281

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

Model	RMSE
Linear Regression	1275.867052
Naive Model	3864.279352
Simple Average Model	1275.081804
2 Point Trailing Moving Average	813.400684
4 Point Trailing Moving Average	1156.589694
6 Point Trailing Moving Average	1283.927428
9 Point Trailing Moving Average	1346.278315
Alpha=0.1, Simple Exponential Smoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, Double Exponential Smoothing	1778.564670
Alpha=0.08621, Beta=1.3722, Gamma=0.4763, Tripple Exponential Smoothing Auto Fit	1304.927405
Alpha=0.4, Beta=0.1, Gamma=0.2, Triple Exponential Smoothing	317.434302
Auto ARIMA	1299.978684
(1,1,1), (1,0,3,12), Auto SARIMA	528.602375
ARIMA (3,1,3)	1319.936734
(1,1,1) (1,1,1,12), Manual SARIMA	359.612449

Table 12: Various alpha value and RMSE value of Rose Dataset

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

Based on the 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	36.096841
1995-09-01	34.999961
1995-10-01	36.289937
1995-11-01	43.126839
1995-12-01	61.593978
1996-01-01	24.293852
1996-02-01	31.406019
1996-03-01	37.545514
1996-04-01	39.735393
1996-05-01	33.753457
1996-06-01	38.868148
1996-07-01	43.093112

Figure 54: Future Predictions of Rose Sales

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

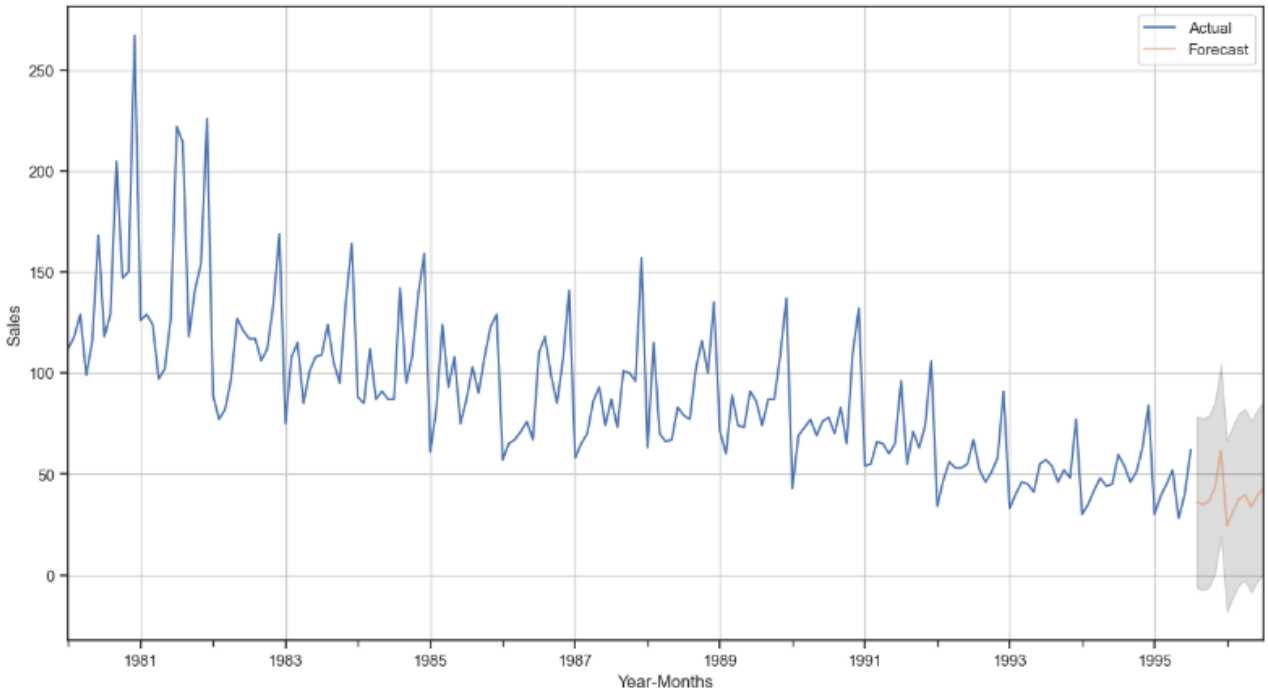

Figure 55: Plot of Future Predictions of Rose Sales

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

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

- The rose wine variety for the company has been seeing a noticeable decline in sales for more than ten years, according to a review of wine sales statistics.
- According to the forecasts of the best model, this tendency is anticipated to persist into the future as well.
- Seasonal fluctuations have a significant impact on wine sales, with sales rising during festival season and falling during the busiest winter month of January.
- Due to the year's slow sales during this time, the company should think about launching efforts to increase wine consumption.
- Since sales are low from April to June, campaigns during this time period may produce the best results for the business. Boosting sales during this time would also improve the wine's performance on the market year-round.
- Running campaigns during peak periods (such as during festivals) might not generate significant impact on sales, as they are already high during this time of the year.
- Campaigns during peak winter time (January) are not recommended as people are less likely to purchase wine due to climatic reasons, and running campaigns during this period may not change people's opinion.
- The company should also consider exploring reasons behind the decline in popularity of the Rose wine variety, and if needed, revamp its production and marketing strategies to regain the market share.