TIME SERIES FORECASTING

TFA

Abstract

The time-series forecasting is a vital area that motivates continuous investigate areas of intrigued for different applications. A critical step for the time-series forecasting is the right determination of the number of past observations (lags).

Sudheendra K

Sudhi0404@gmail.com

Contents

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.

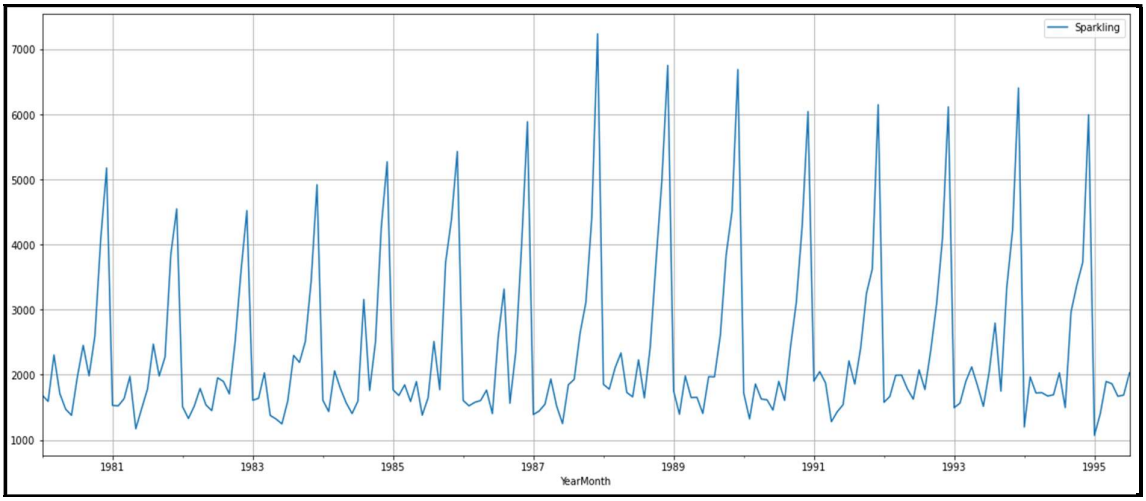
Time Series Forecasting on Sparkling Wine Dataset

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

YearMonth		YearMonth	
1980-01-01	1686	**1995-03-01**	1897
1980-02-01	1591	**1995-04-01**	1862
1980-03-01	2304	**1995-05-01**	1670
1980-04-01	1712	**1995-06-01**	1688
1980-05-01	1471	**1995-07-01**	2031

Sparkling Snippet 1 Head and Tail of the 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.*



Sparkling Fig 1 Plot of the dataset

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

Sparkling Snippet 2 Post Ingestion 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.*

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

```
<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
```

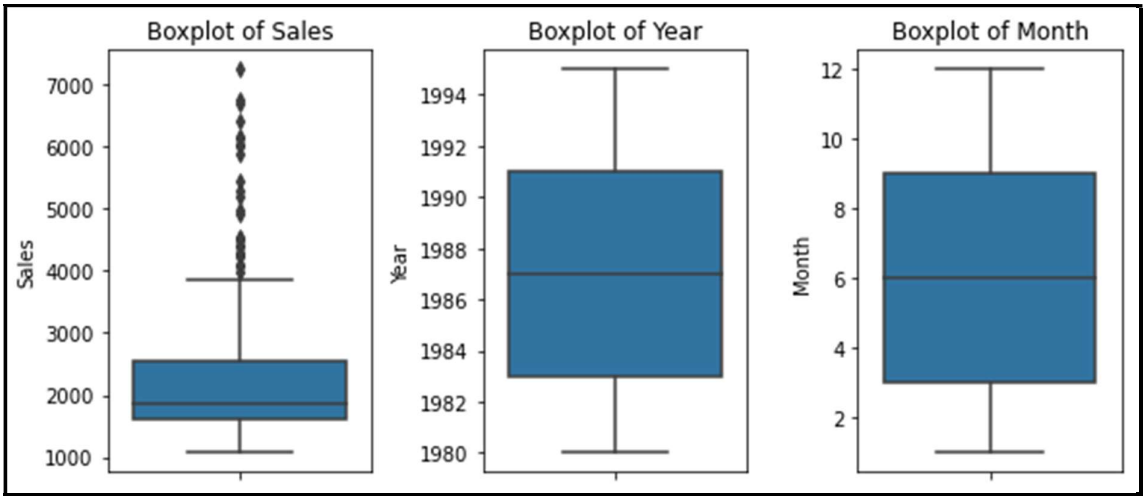
Sparkling Snippet 3 Dataset Info

- *Data Type;*
- *Index: DateTime*
- *Sales: integer*
- *Month: integer*
- *Year: integer*

	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

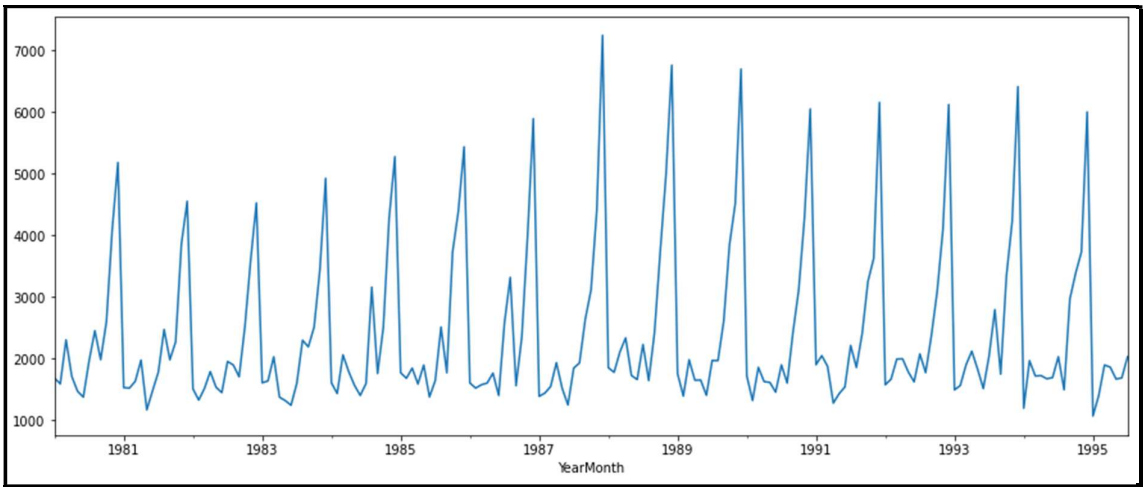
Sparkling Snippet 4 Statistical summary

- *Null Value: There are no null values present in the dataset. So we can do further analysis smoothly.*



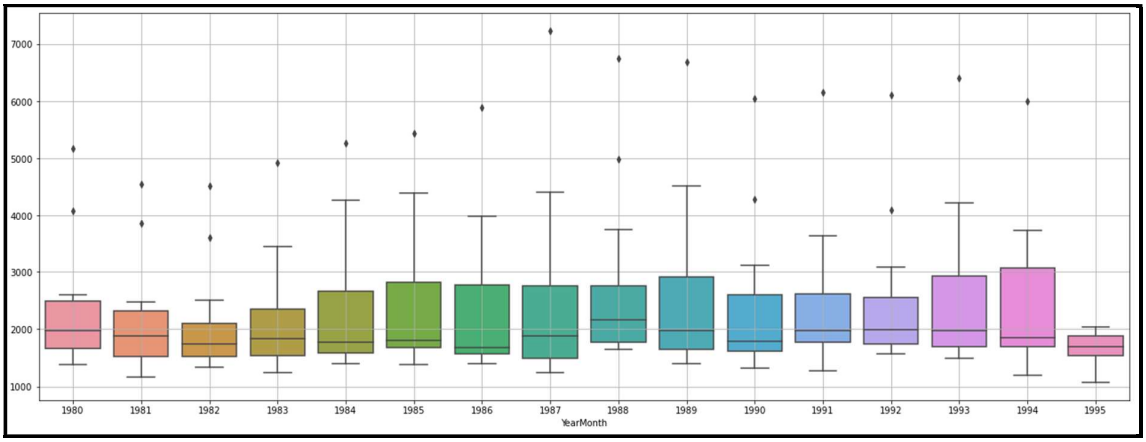
Sparkling Fig 2 BoxPlot of the dataset

- *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.*



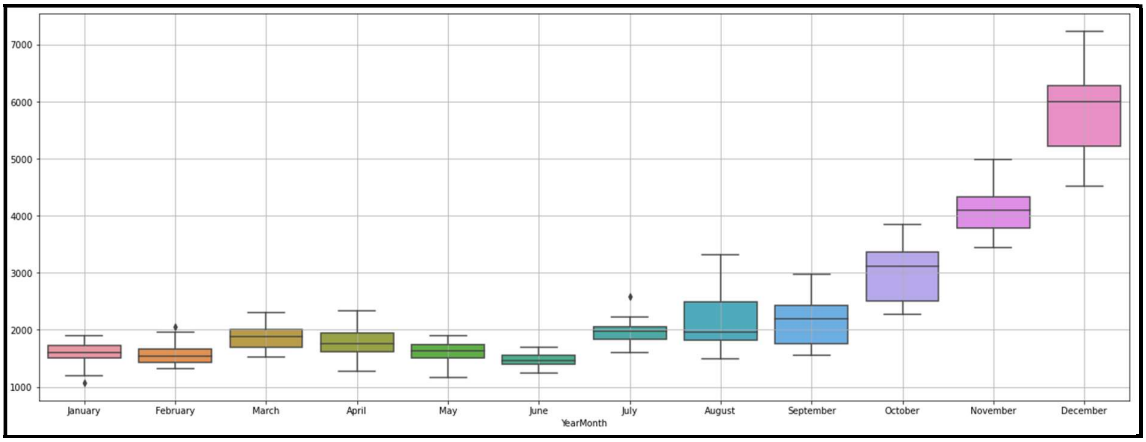
Sparkling Fig 3 Line plot of the dataset

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



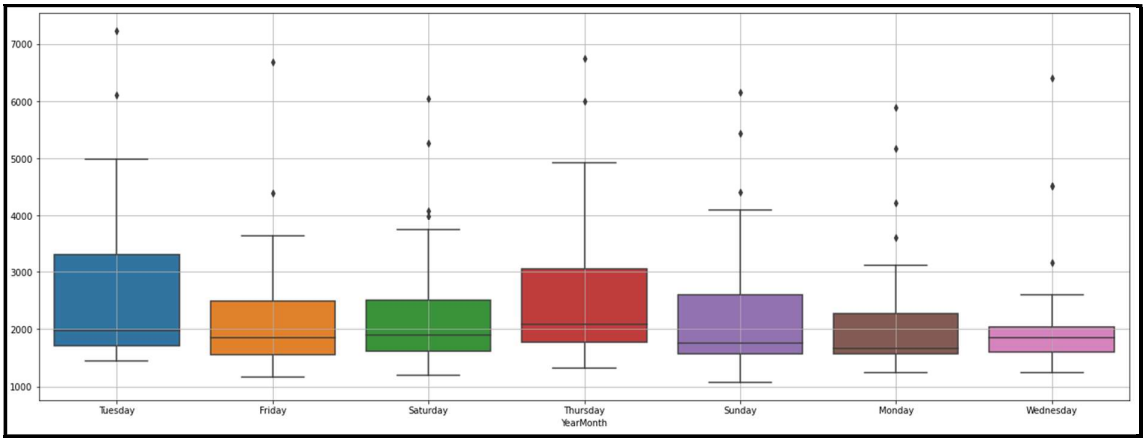
Sparkling Fig 4 Yearly Boxplot

- *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.*



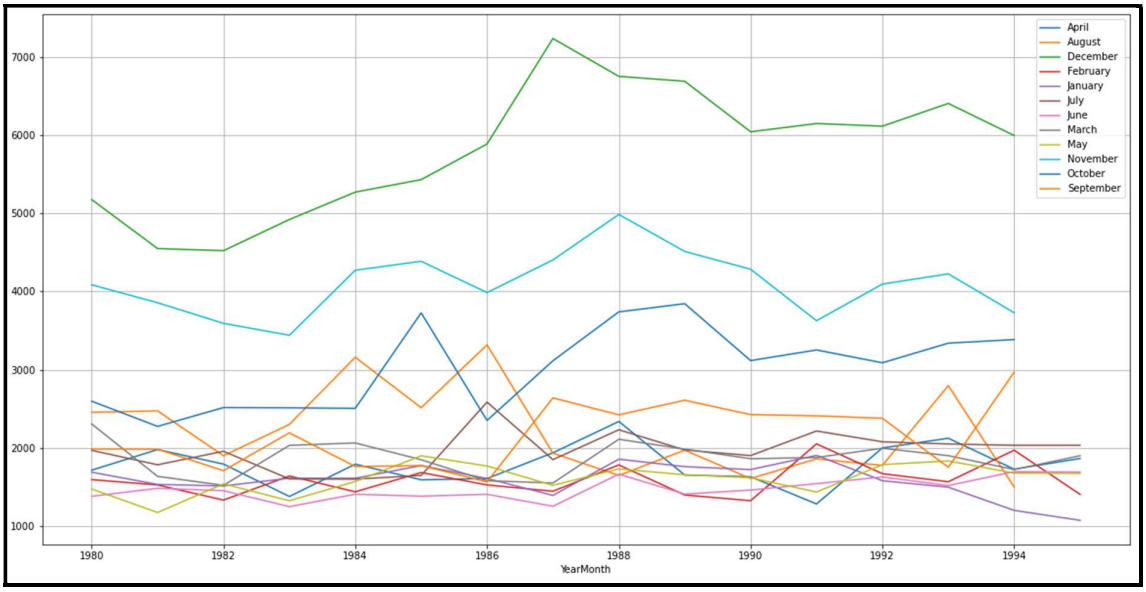
Sparkling Fig 5 Monthly Box plot

- *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.*



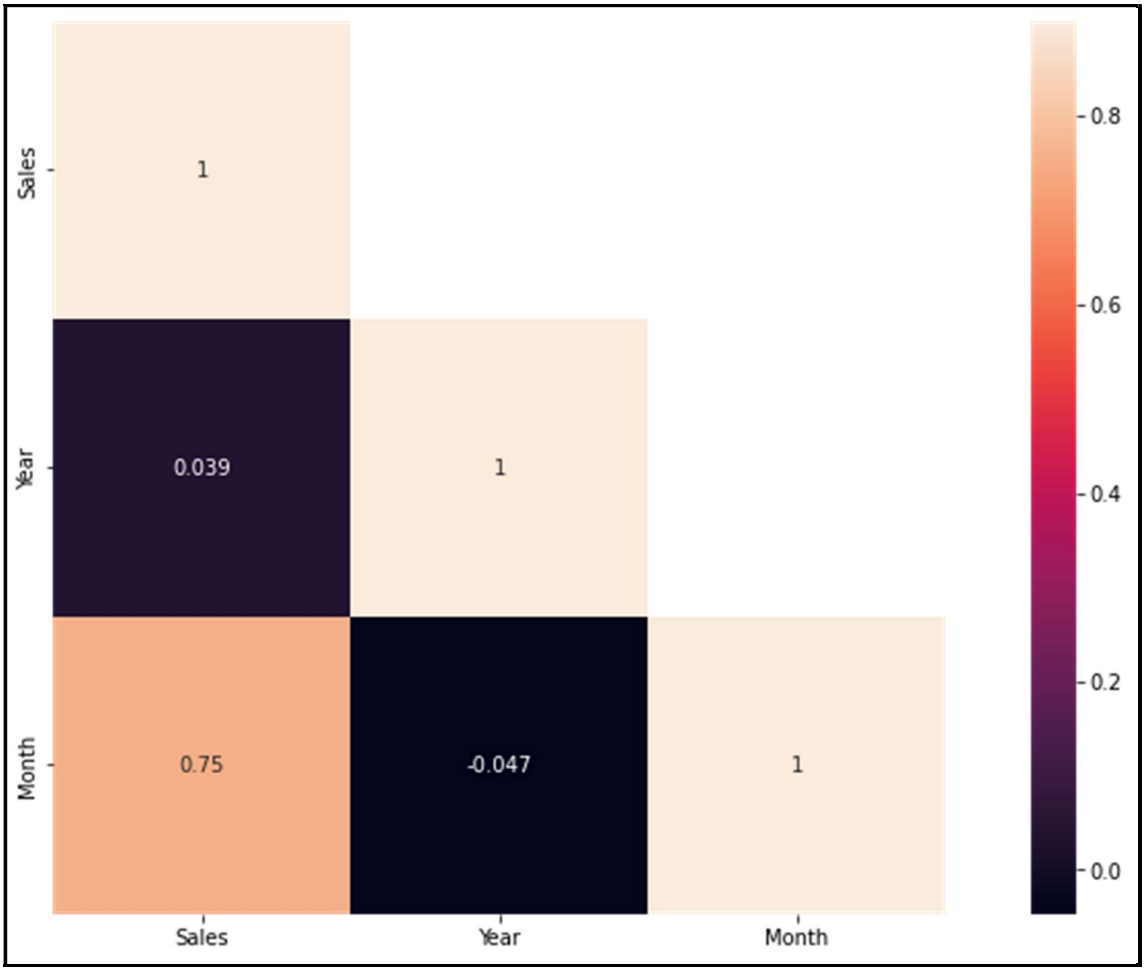
Sparkling Fig 6 Weekly Boxplot

- *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.*



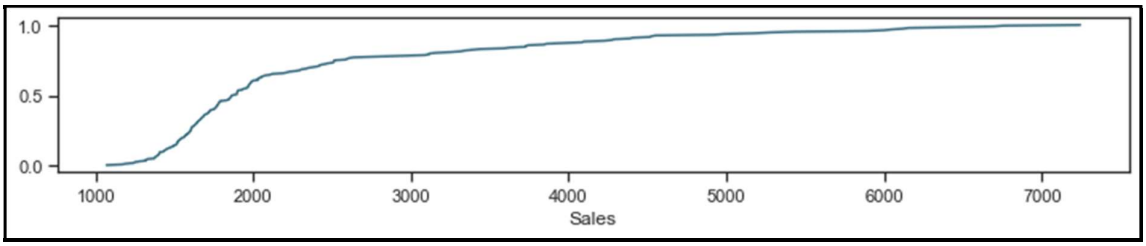
Sparkling Fig 7 Monthly Sales over the years

- *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.*



Sparkling Fig 8 Correlation Plot

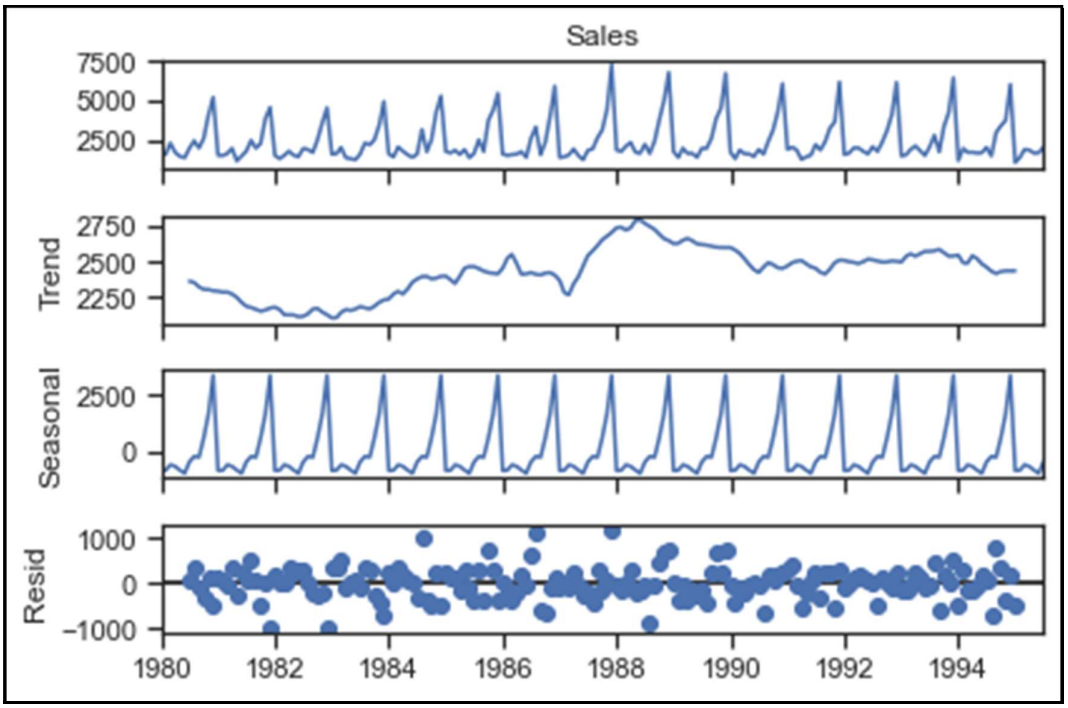
- *This heat map shows that there is a low correlation between sales and year.*
- *There is a more correlation between month and sales.*
- *It indicated seasonal patterns in sales.*



Sparkling Fig 9 ECDF plot

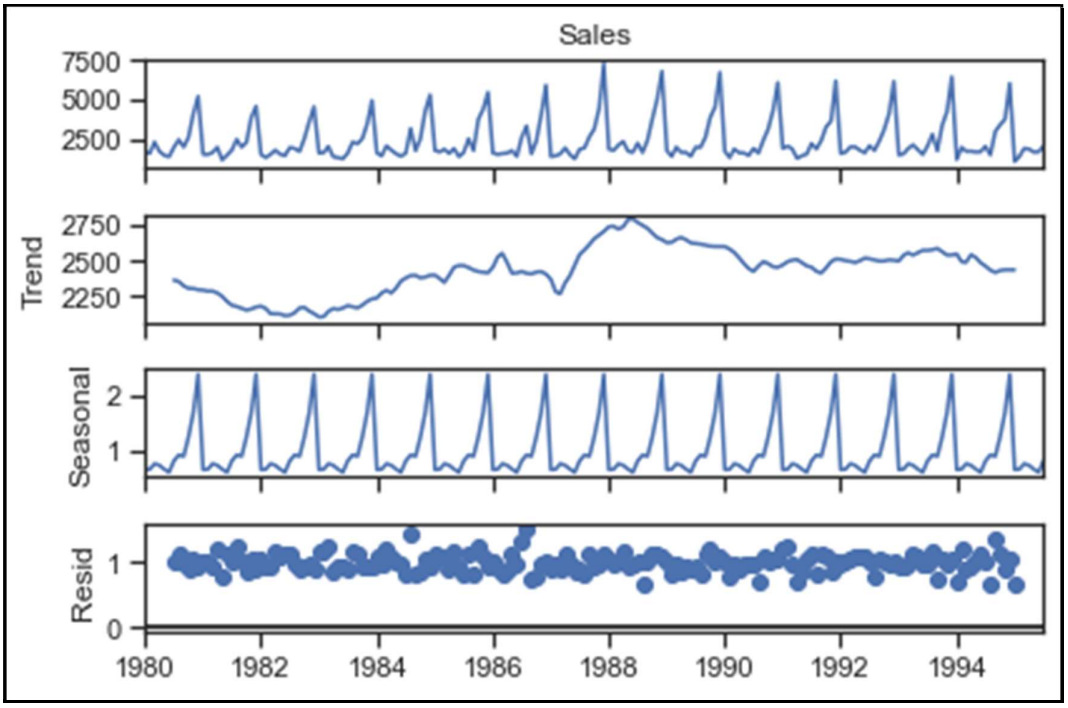
This plot shows:

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



Sparkling Fig 10 Additive Decomposition Plot

- *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.*

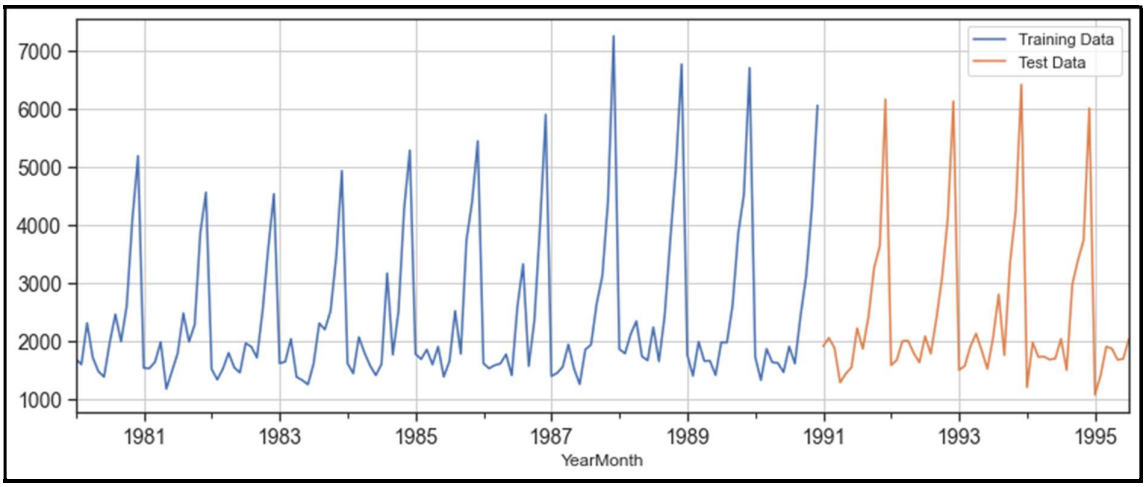


Sparkling Fig 11 Multiplicative Decomposition Plot

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.



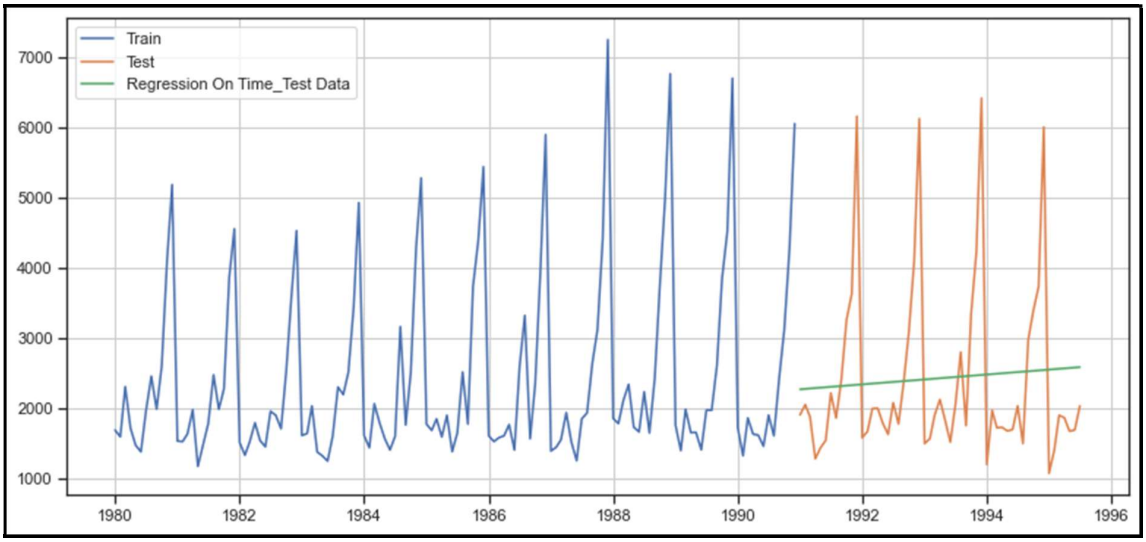
Sparkling Fig 12 Line Plot train and test dataset

- *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 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)*

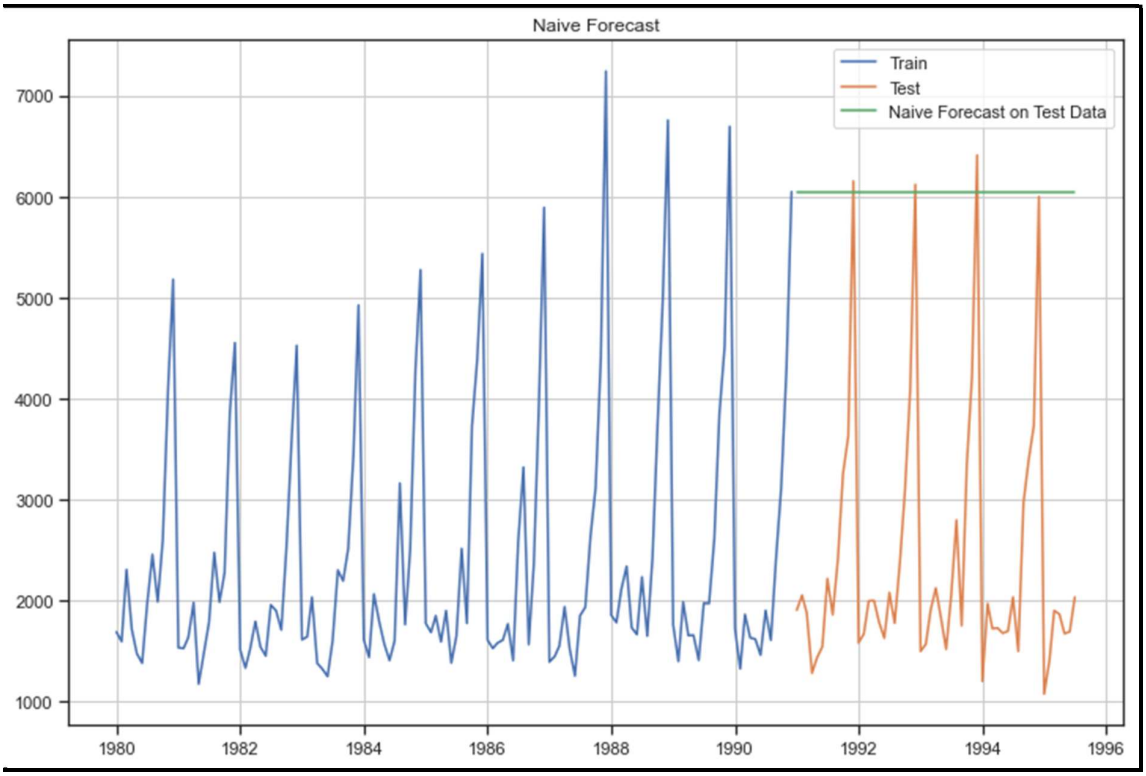
Model 1: Linear Regression



Sparkling Fig 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.*
- ***RMSE: Linear Regression is 1275.867052***

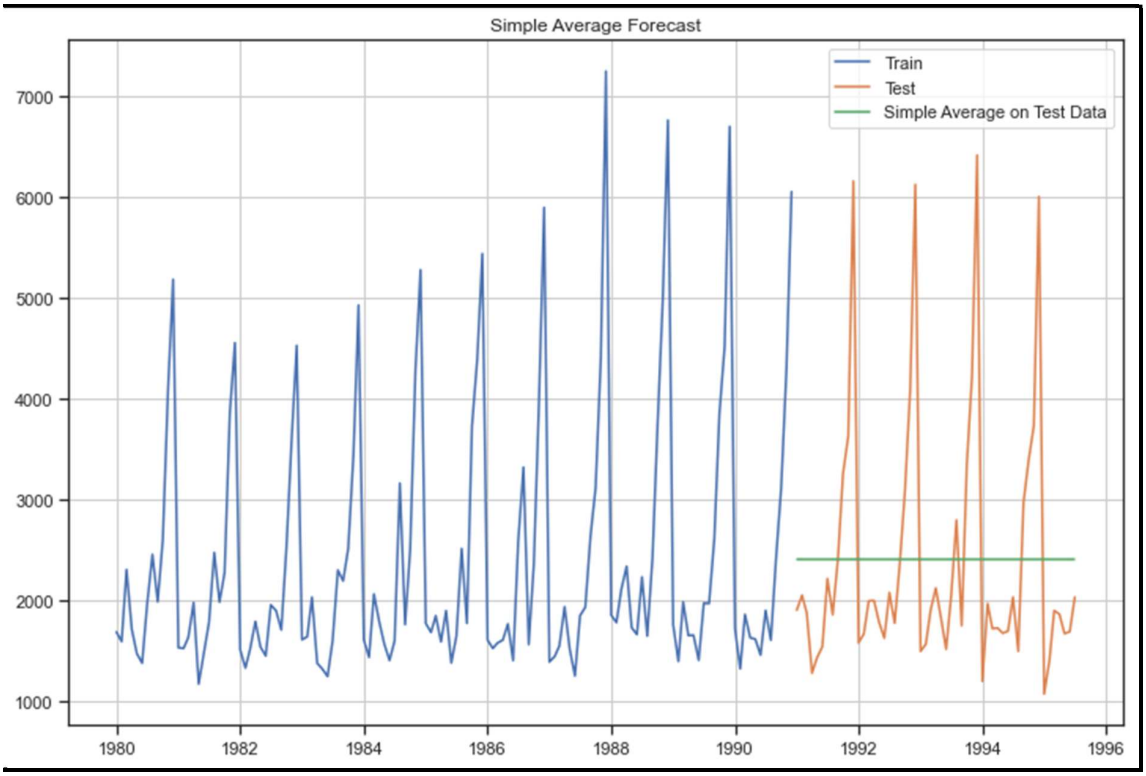
Model 2: Naïve Approach



Sparkling Fig 14 Naïve Forecast

- *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.*
- ***RMSE: Naïve Model is 3864.279352***

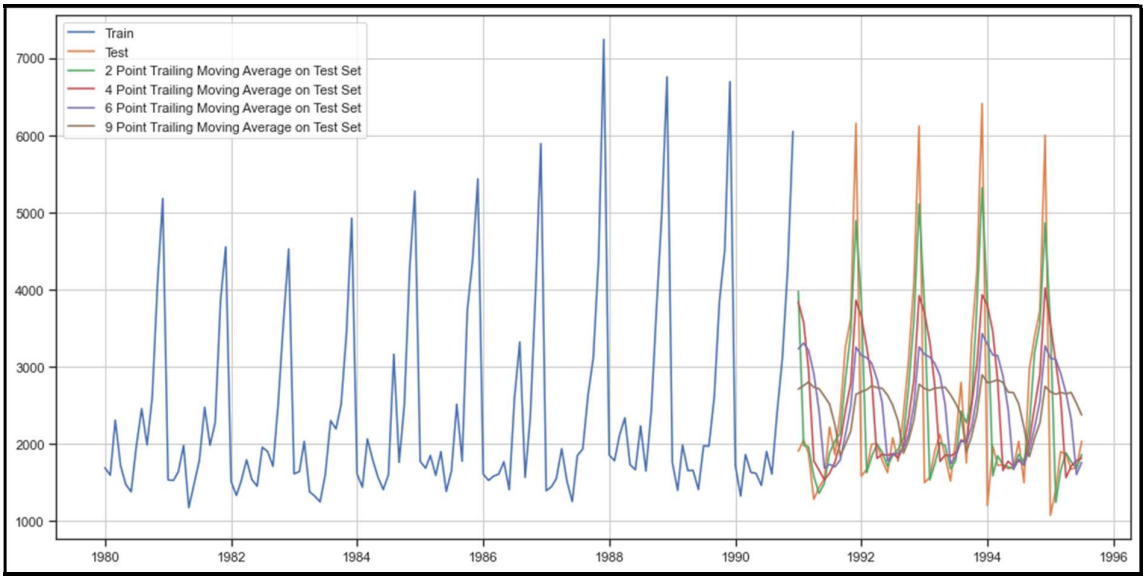
Model 3: Simple Average



Sparkling Fig 15 Simple Average Forecast

- *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.*
- ***RMSE: Simple Average is 1275.081804***

Model 4: Moving Average (MA)



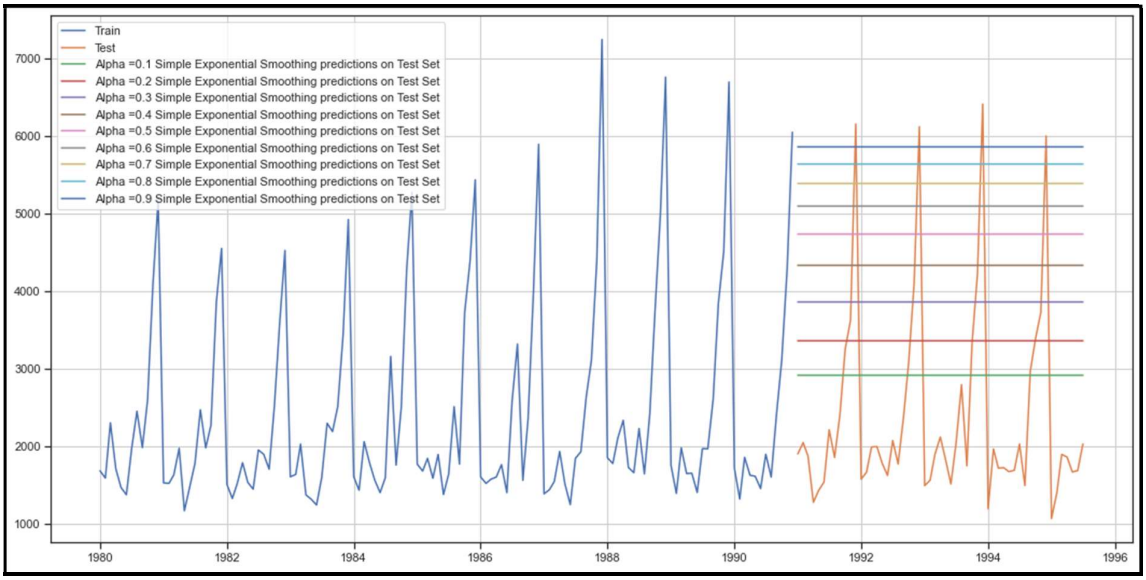
Sparkling Fig 16 Moving Average Forecast

- *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.*

Model 5: Simple Exponential Smoothing

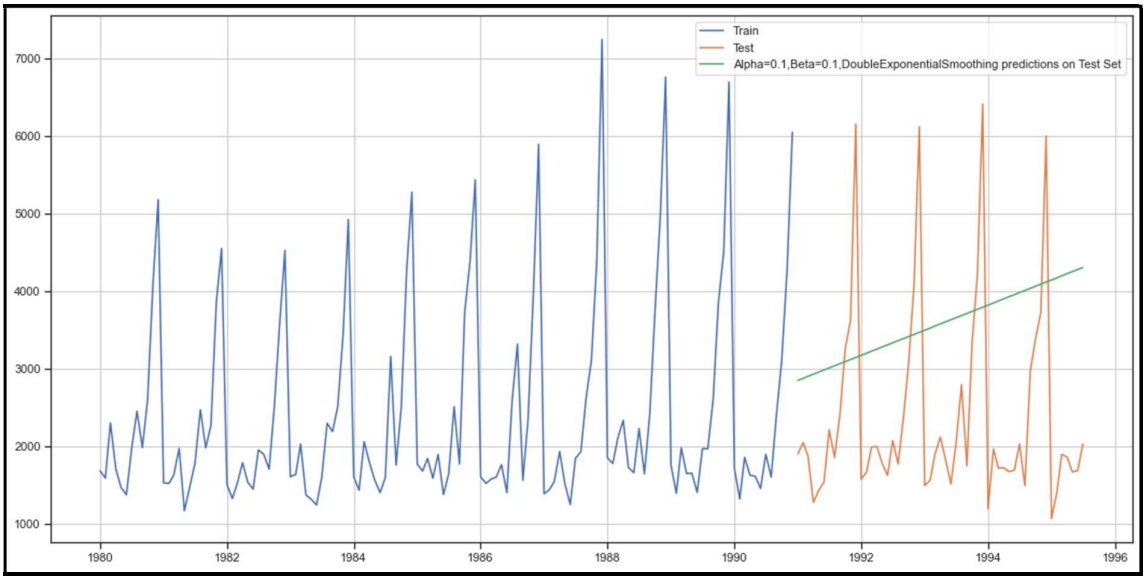


Sparkling Fig 17 Simple Exponential smoothing Forecast

- *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

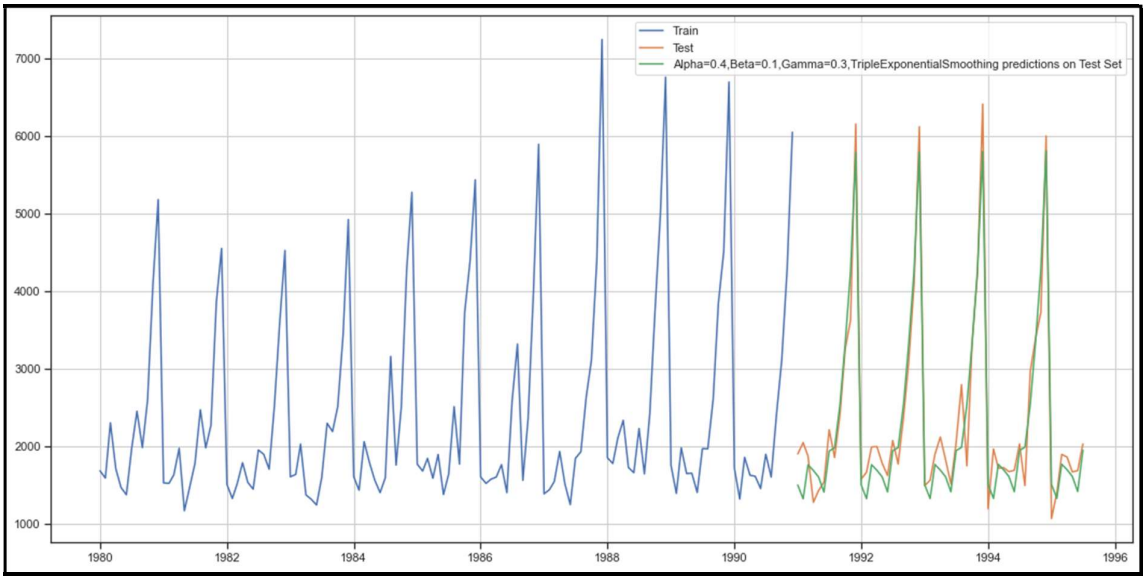
Model 6: Double Exponential Smoothing (Holt's Model)



Sparkling Fig 18 Double Exponential Smoothing Forecasting

- *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*

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



Sparkling Fig 19 Triple Exponential Smoothing Forecast

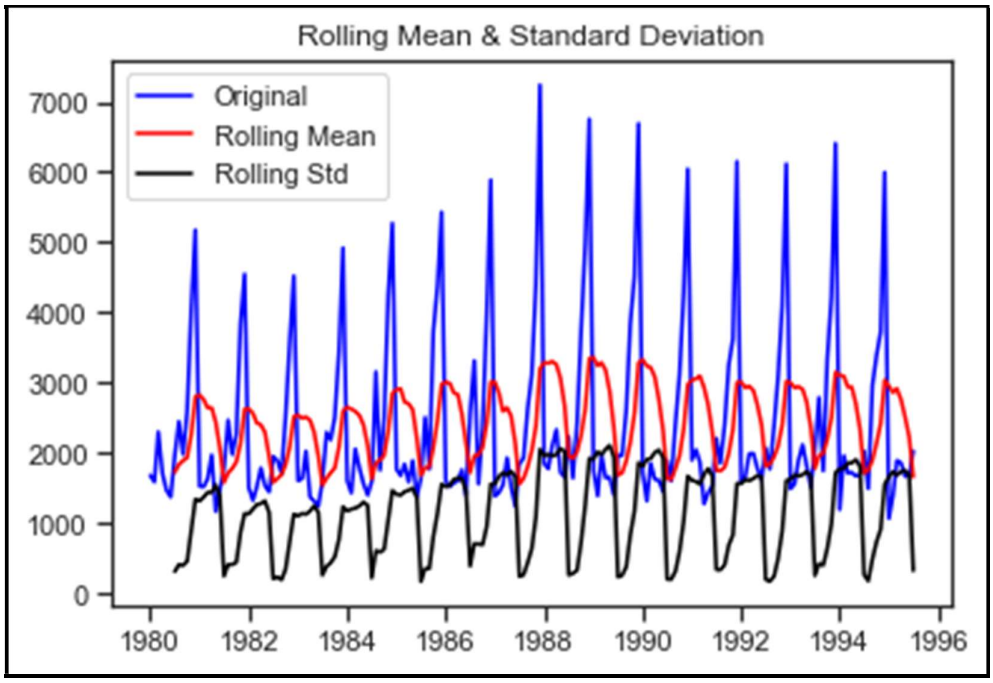
- *Output for a best alpha, beta, and gamma values are shown by the green color line in the above plot.*
- *The best model had both a multiplicative trends, as well as a seasonality Model, which was evaluated using the RMSE metric. Below is the RMSE calculated for this model.*
- *Alpha=0.4, Beta=0.1, Gamma=0.3, TripleExponentialSmoothing 317.434302*

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.*

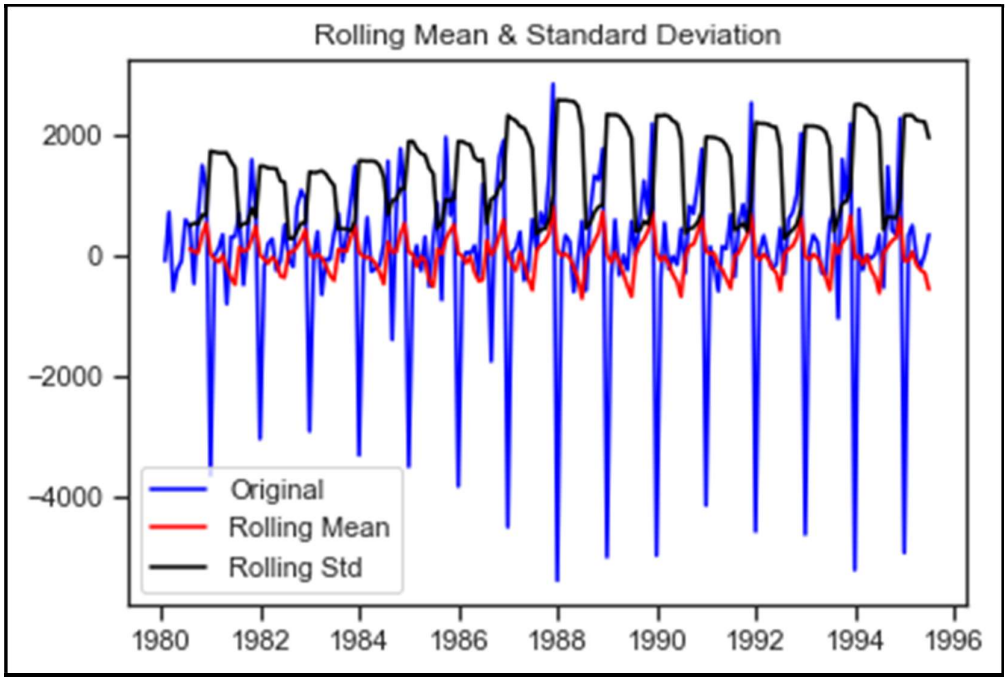
- *The hypothesis in a simple form for the ADF test is:*
- *H0: The Time Series has a unit root and is thus non-stationary.*
- *H1: The Time Series does not have a unit root and is thus stationary.*
- *We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.*
- *We see that at 5% significant level the Time Series is non-stationary.*



Sparkling Fig 20 Dicky fuller test plot

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.*



Sparkling Fig 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)	2213.509213
15	(3, 1, 3)	2221.459263
14	(3, 1, 2)	2230.808088
11	(2, 1, 3)	2232.836355
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)	2235.498605
7	(1, 1, 3)	2235.607804
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).*

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Thu, 06 Jul 2023	AIC	2213.509
Time:	16:53:33	BIC	2227.885
Sample:	01-01-1980 - 12-01-1990	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	-18.217	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-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

Sparkling Snippet 5 Summary report of ARIMA model

- ***RMSE values is: 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
28	(0, 1, 1)	(3, 0, 0, 12)	1428.460768
44	(0, 1, 2)	(3, 0, 0, 12)	1428.599341
29	(0, 1, 1)	(3, 0, 1, 12)	1428.872799
30	(0, 1, 1)	(3, 0, 2, 12)	1429.589188
45	(0, 1, 2)	(3, 0, 1, 12)	1429.744837

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

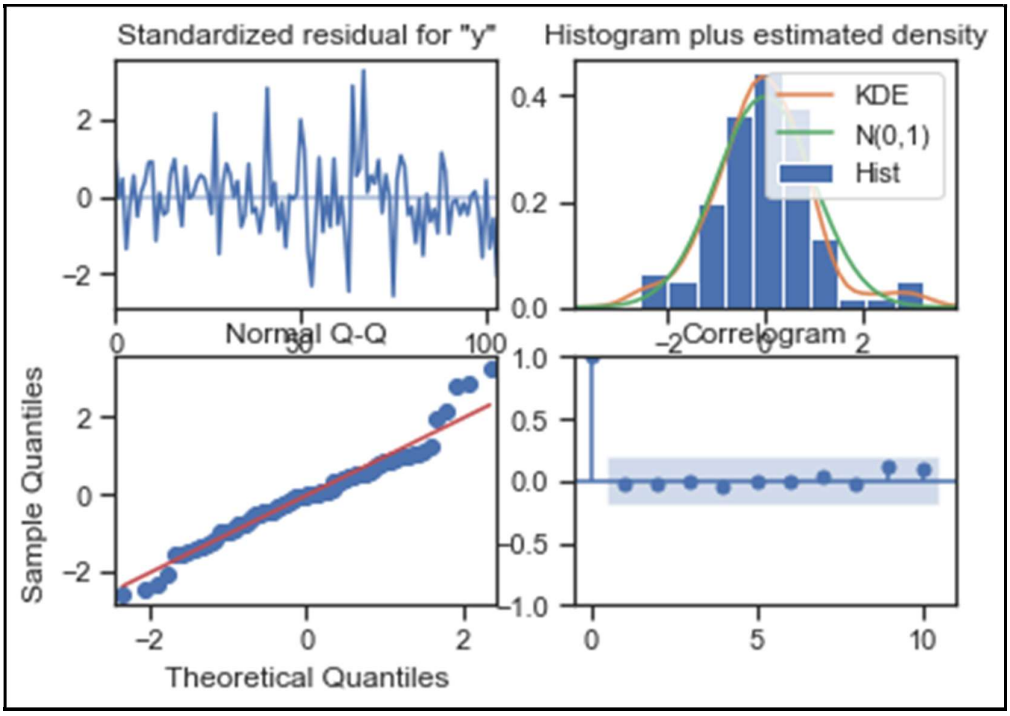
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Thu, 06 Jul 2023	AIC	1555.584
Time:	16:56:30	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	-2.464	0.014	-1.128	-0.128
ma.L1	-0.1040	0.225	-0.463	0.644	-0.545	0.337
ma.L2	-0.7276	0.154	-4.736	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.838	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.*

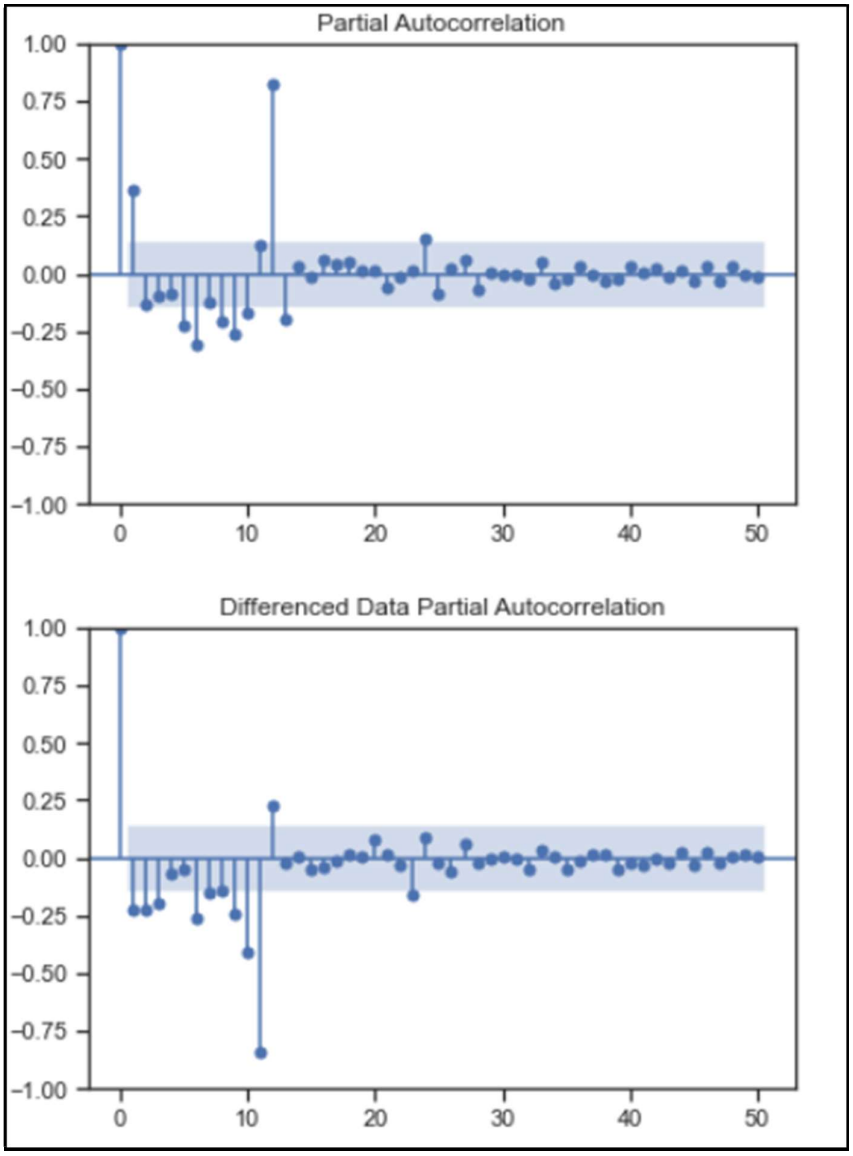


Sparkling Fig 22 SARIMA Plot

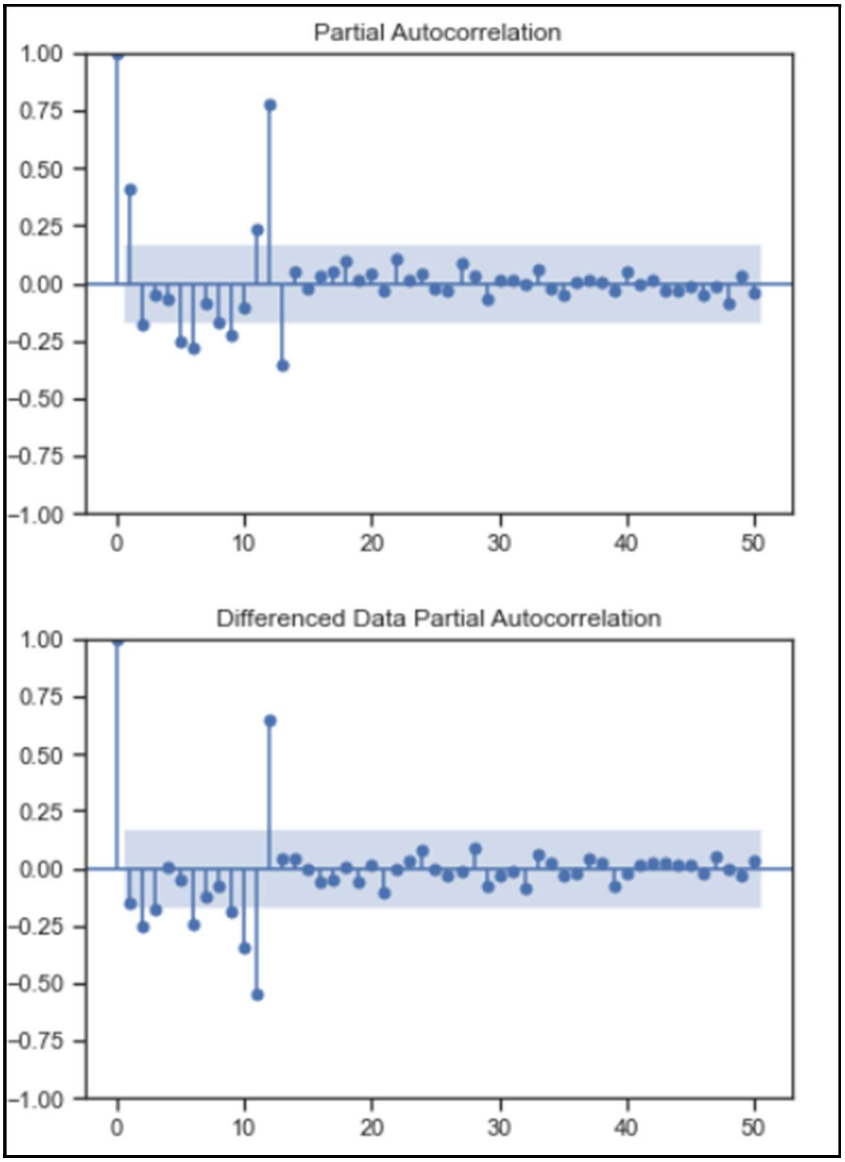
- ***RSME of Model: 528.6069474180102***

7.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



Sparkling Fig 23 ACF plot



Sparkling Fig 24 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*

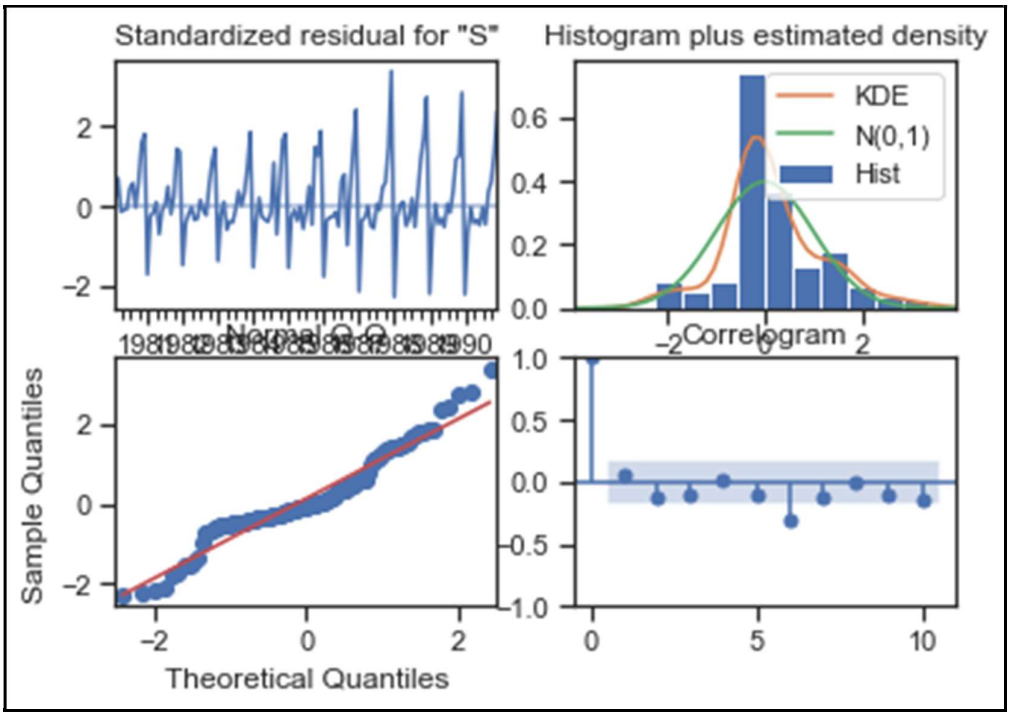
SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1114.878
Date:	Thu, 06 Jul 2023	AIC	2235.755
Time:	16:56:33	BIC	2244.381
Sample:	01-01-1980 - 12-01-1990	HQIC	2239.260
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

Summary from this manual ARIMA model.



Sparkling Fig 25 Manual ARIMA plot

- ***Model Evaluation: RSME is 1319.9367298218867***

Manual SARIMA Model

- *SARIMAX (1, 1, 1) x (1, 1, 1, 12)*

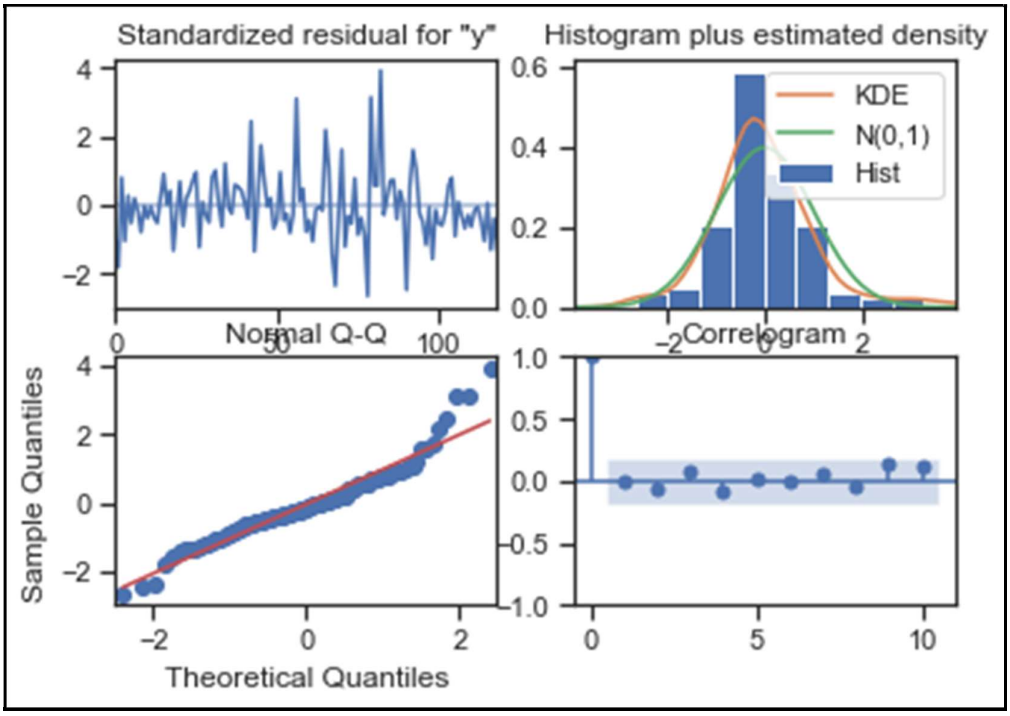
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-882.088
Date:	Thu, 06 Jul 2023	AIC	1774.175
Time:	16:56:34	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

Sparkling Snippet 6 Summary of the manual SARIMA model



Sparkling Fig 26 Manual SARIMA Plot

- ***Model Evaluation: RSME is 359.612454***

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.

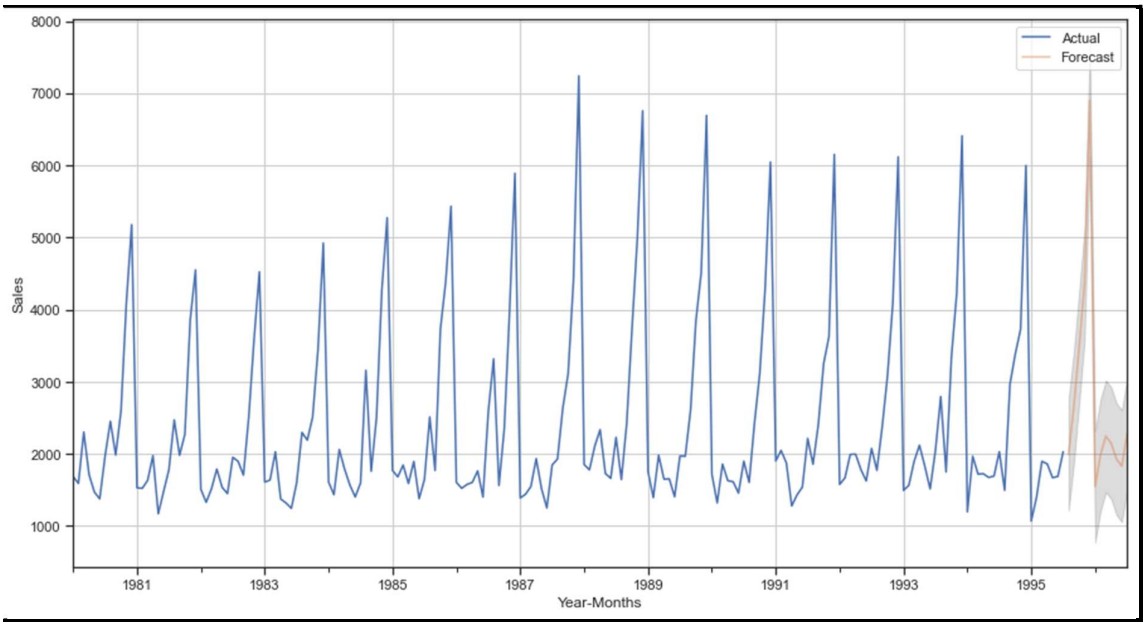
	Test RMSE
Linear Regression	1275.867052
Naive Model	3864.279352
Simple Average Model	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1304.927405
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Auto_ARIMA	1299.980241
(1,1,1),(2,0,3,12),Auto_SARIMA	528.661982
ARIMA(3,1,3)	1319.936735
(1,1,1)(1,1,1,12),Manual_SARIMA	359.612456

- *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
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

Sparkling Snippet 7 Sales Prediction

- *sales predictions made by this best optimum model.*



Sparkling Fig 27 Prediction Plot

- *the sales prediction on the graph along with the confidence intervals. Predictions, 1 year into the future are shown in orange color, while the confidence interval has been shown in grey color.*

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 sales for Sparkling wine for the company are predicted to be atleast 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.*

Time Series Forecasting on Rose Wine Dataset

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

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

Rose Snippet 1 Head and Tail of the dataset

- *Number of Rows and Columns of Dataset has 187 rows and 1 column.*
- *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 the dataset.*

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

Rose Snippet 2 Post Ingestion of the dataset

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

```
<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   185 non-null    float64
 1   Year    187 non-null    int64
 2   Month   187 non-null    int64
dtypes: float64(1), int64(2)
memory usage: 5.8 KB
```

Rose Snippet 3 Dataset Info

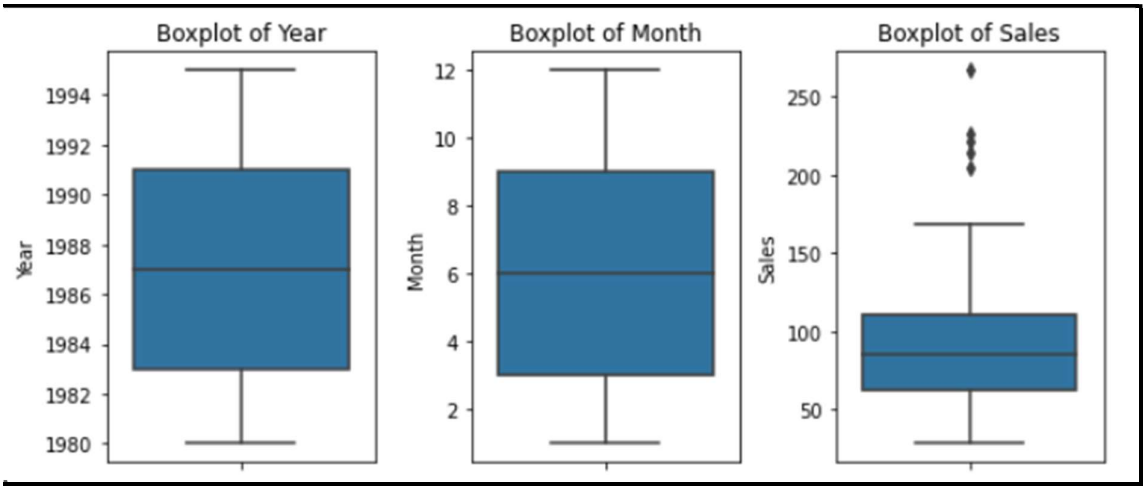
- *Data Type;*
- *Index: DateTime*
- *Sales: integer*
- *Month: integer*
- *Year: integer*

	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

Rose Snippet 4 Statistical Summary

- *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.*
- *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.*

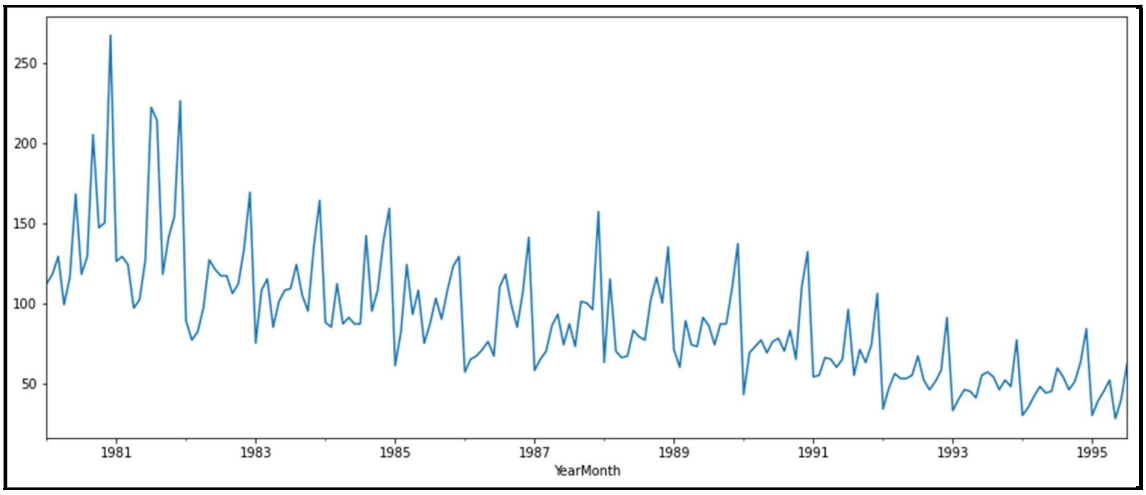
- *In this approach, instead of taking means for the 7th months across all the years, we just took mean of the 7th months values from a year before and a year after the missing value.*
- *Similar steps were taken for 8th month.*



Rose Fig 1 Box Plot

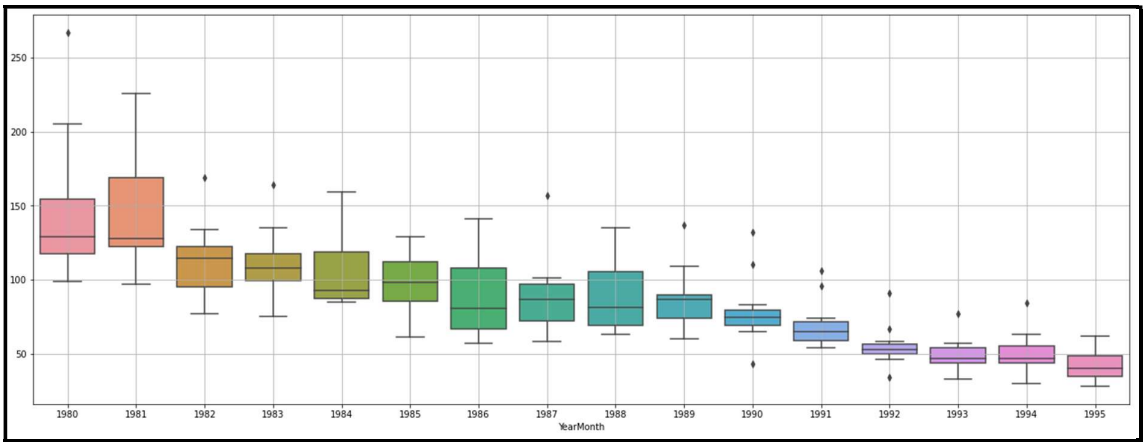
The box plot shows:

- *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.*



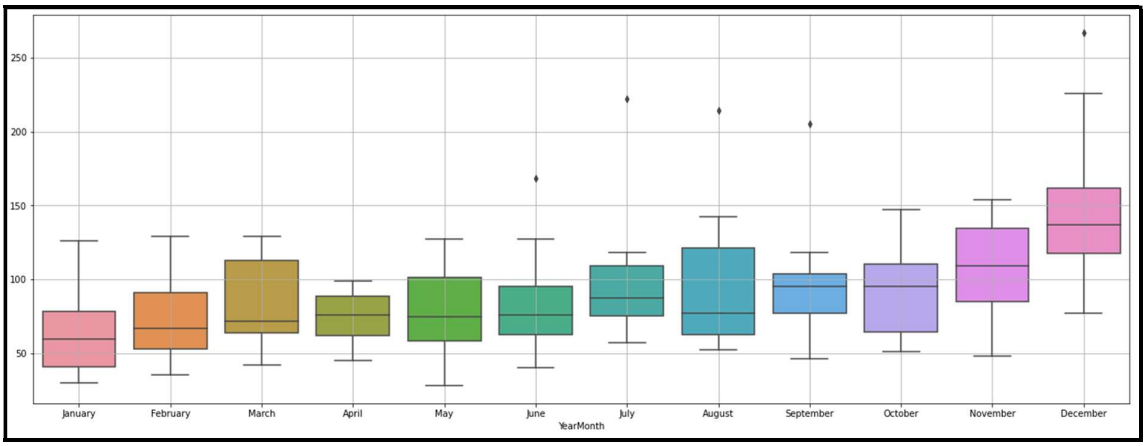
Rose Fig 2 Line Plot

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



Rose Fig 3 Yearly Box plot

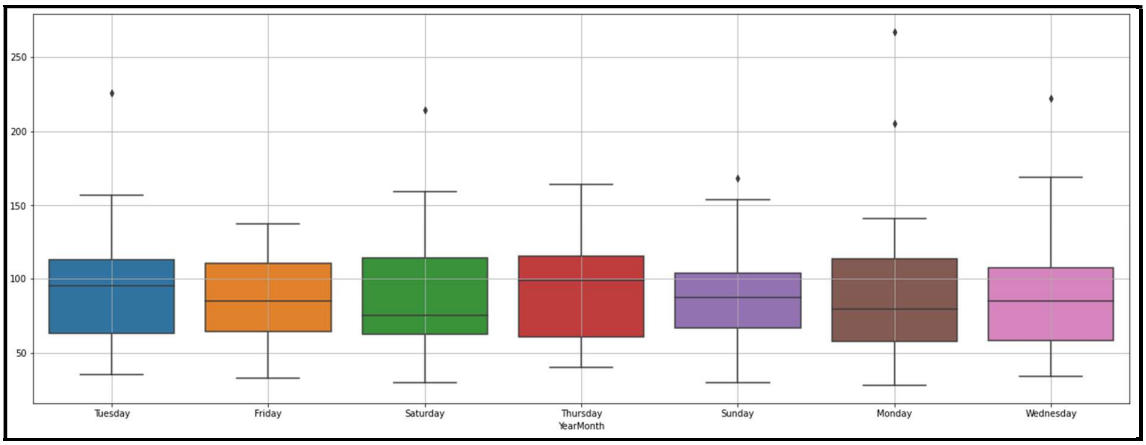
- *This yearly box plot shows there is consistency over the years and there was a peak in 1980-1981. Outliers are present in almost all years.*



Rose Fig 4 Monthly Box plot

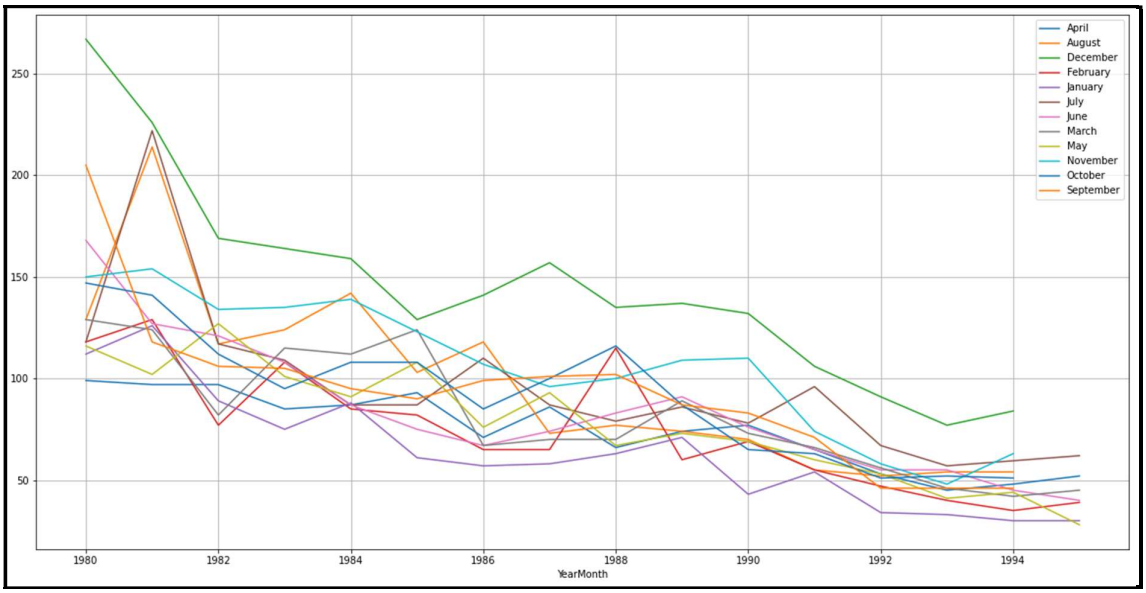
- *The plot shows that sales are highest in the month of December and lowest in the month of January. Sales are consistent from January to July then from*

august the sales start to increase. Outliers are present in June, July, august, September and December.



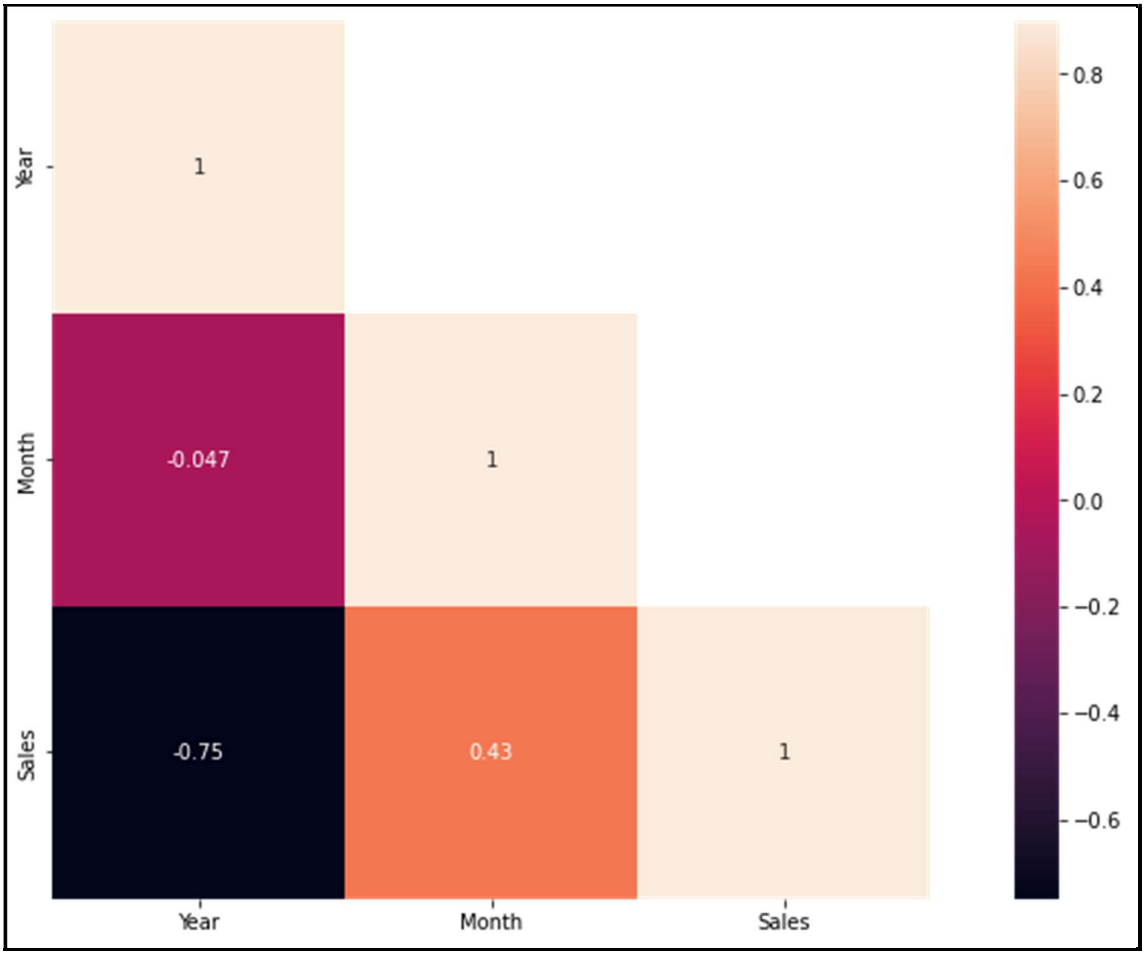
Rose Fig 5 Weekly Box Plot

- *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.*



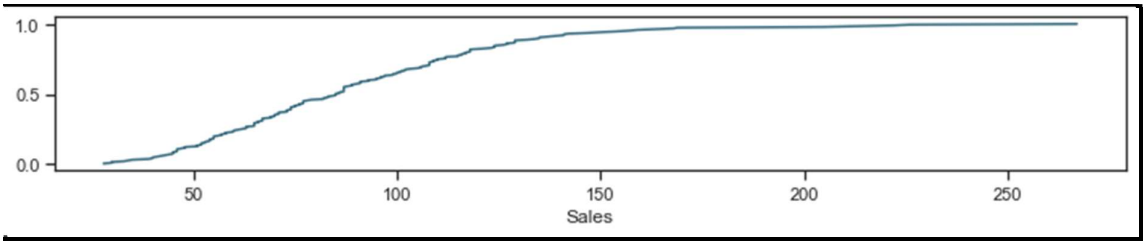
Rose Fig 6 Monthly Sales over years

- *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.*



Rose Fig 7 Correlation Plot

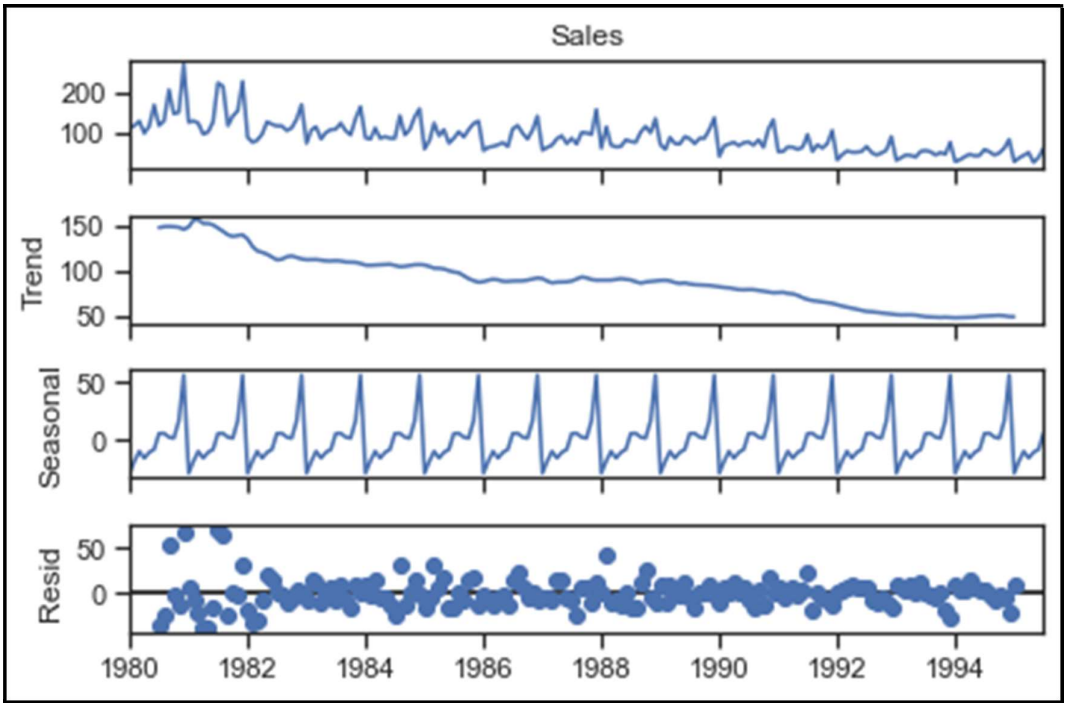
- *This heat map shows that there was little correlation between Sales and the Years data, there significantly more correlation between the month and Sales columns.*
- *Clearly indicating a seasonal pattern in our Sales data. Certain months have higher sales, while certain months have lesser.*



Rose Fig 8 ECDF

This plot shows:

- *50% sales has been less 100*
- *Highest vales is 250*
- *Aprox 90% sales has been less than 150*

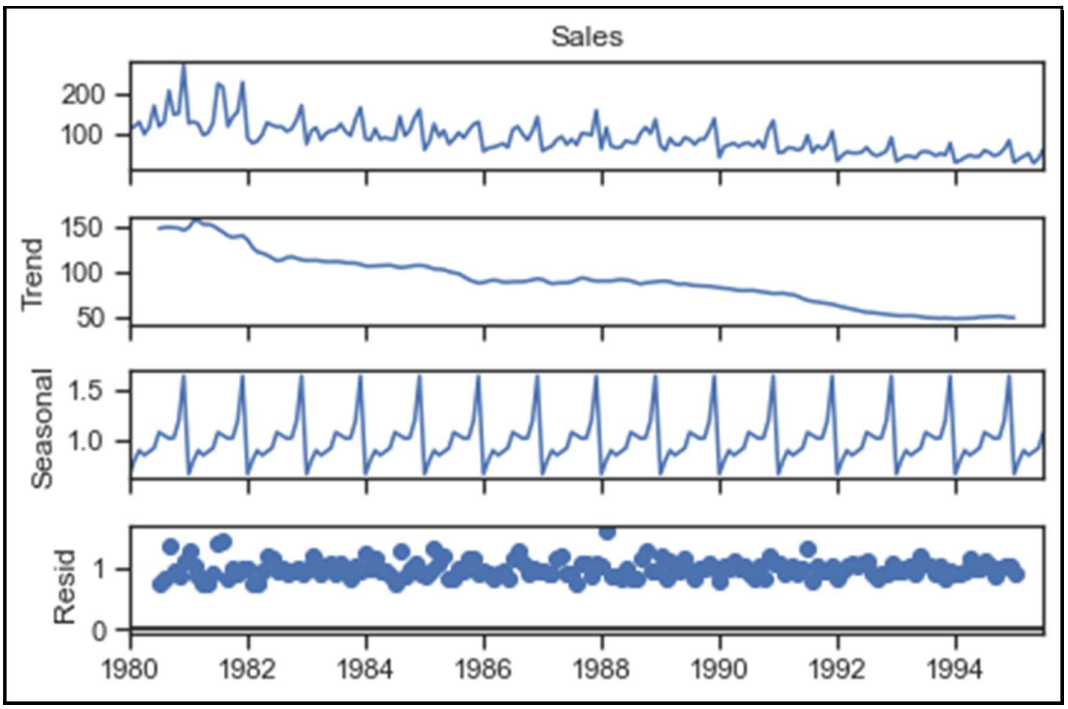


Rose Fig 9 Addictive Decomposition

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.*

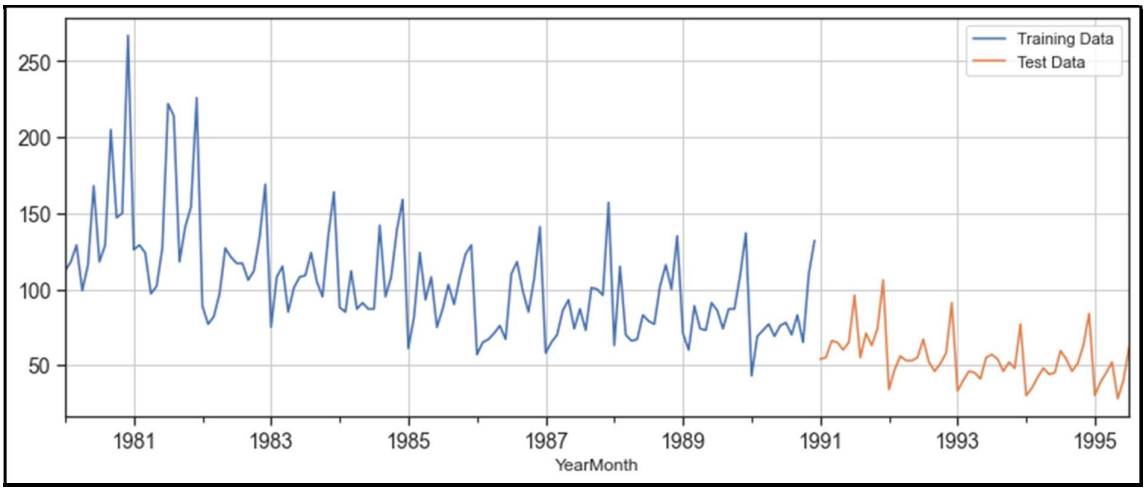


Rose Fig 10 Multiplicative Decomposition

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.*

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



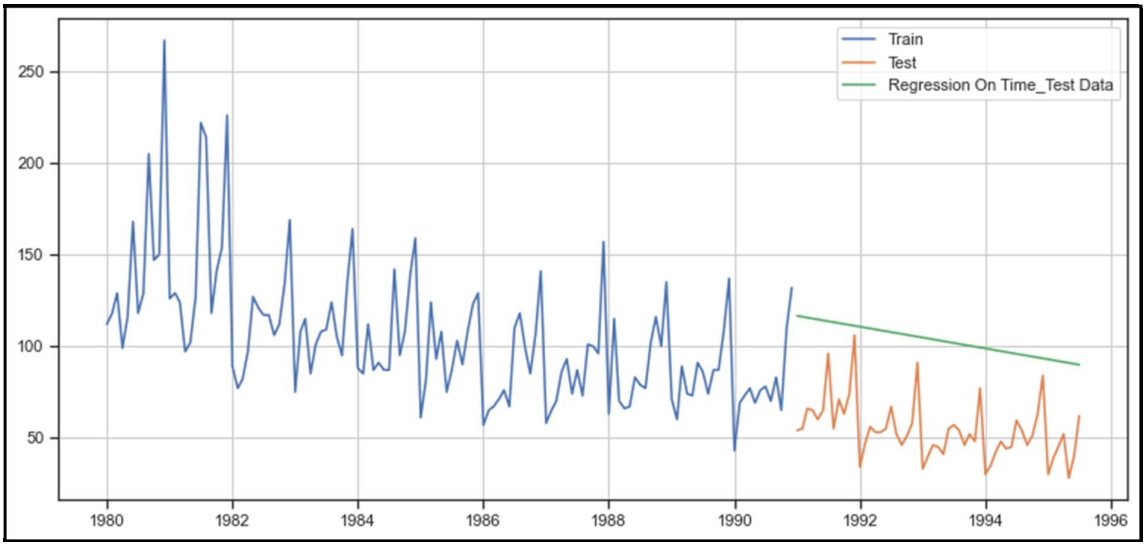
Rose Fig 11 Training and Test

- *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.*

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)*

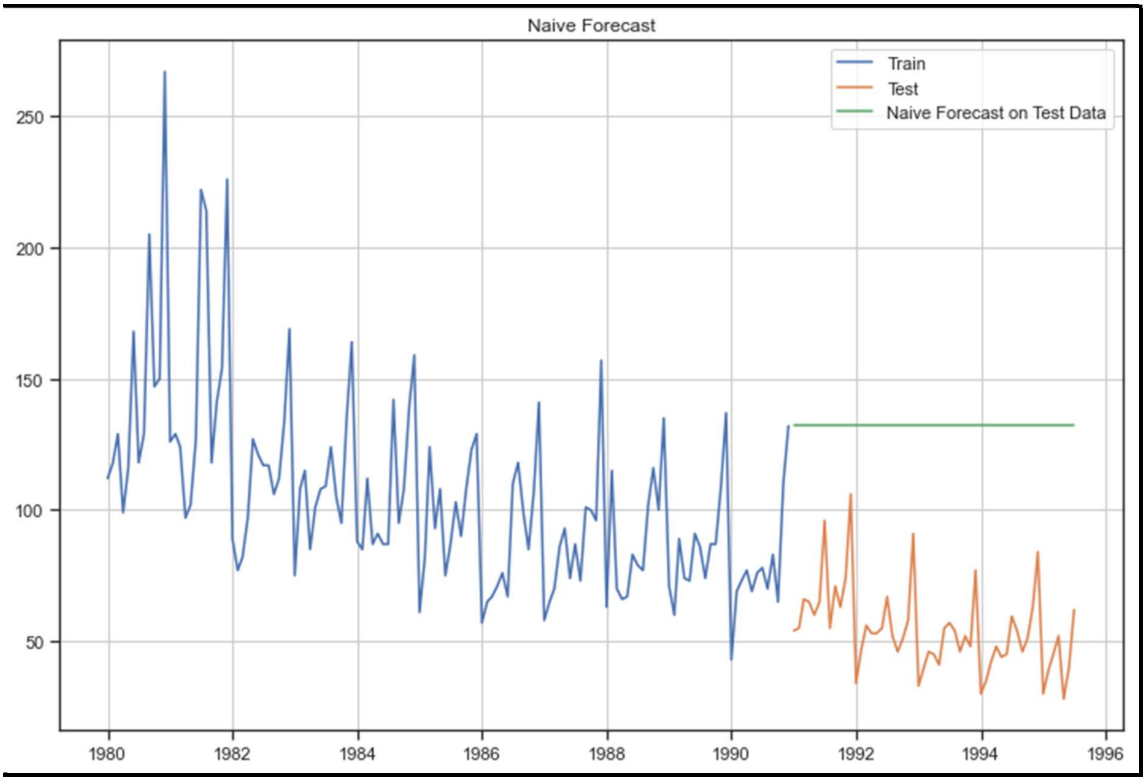
Model 1: Linear Regression



Rose Fig 12 Linear Regression Forecast

- *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 51.080941***

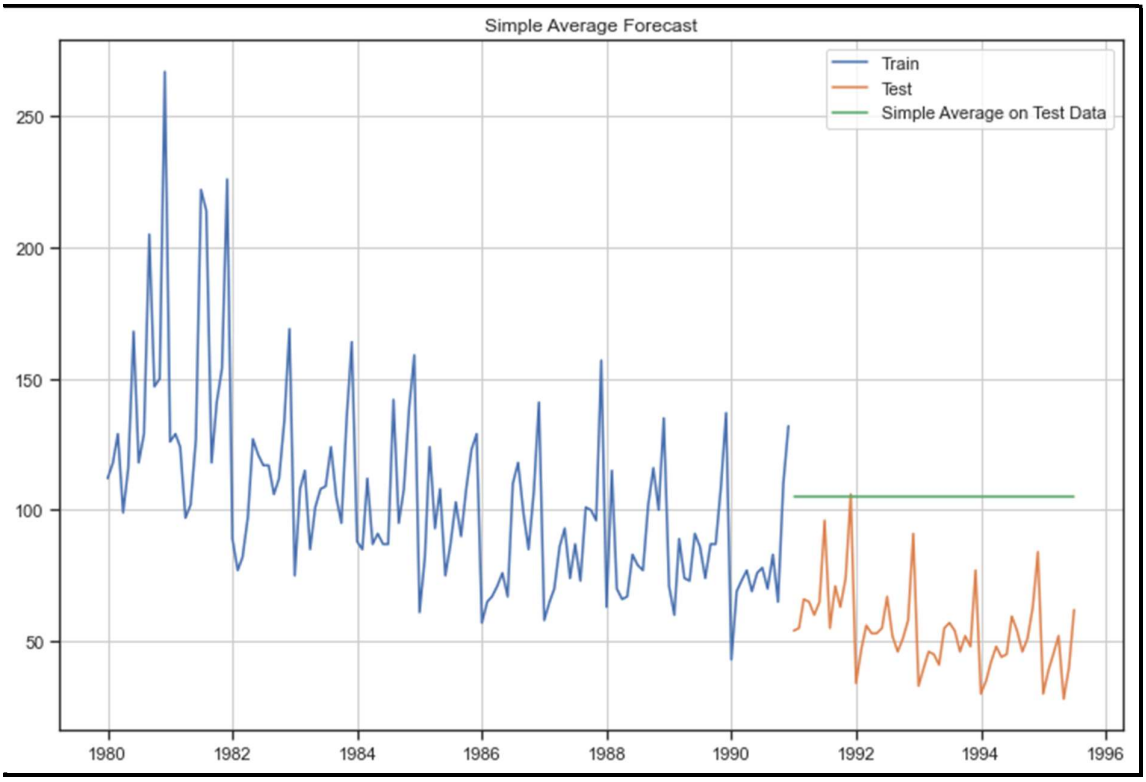
Model 2: Naive Approach



Rose Fig 13 Naïve Forecast

- *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.*
- ***Naive Model 79.304391***

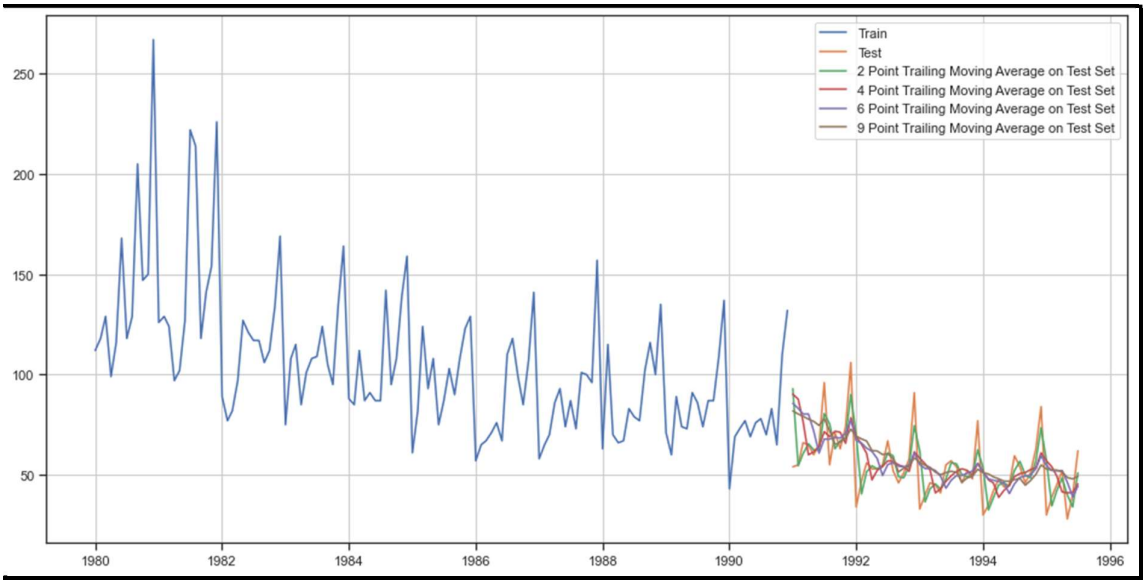
Model 3: Simple Average



Rose Fig 14 Simple Average Forecast

- *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***

Model 4: Moving Average (MA)



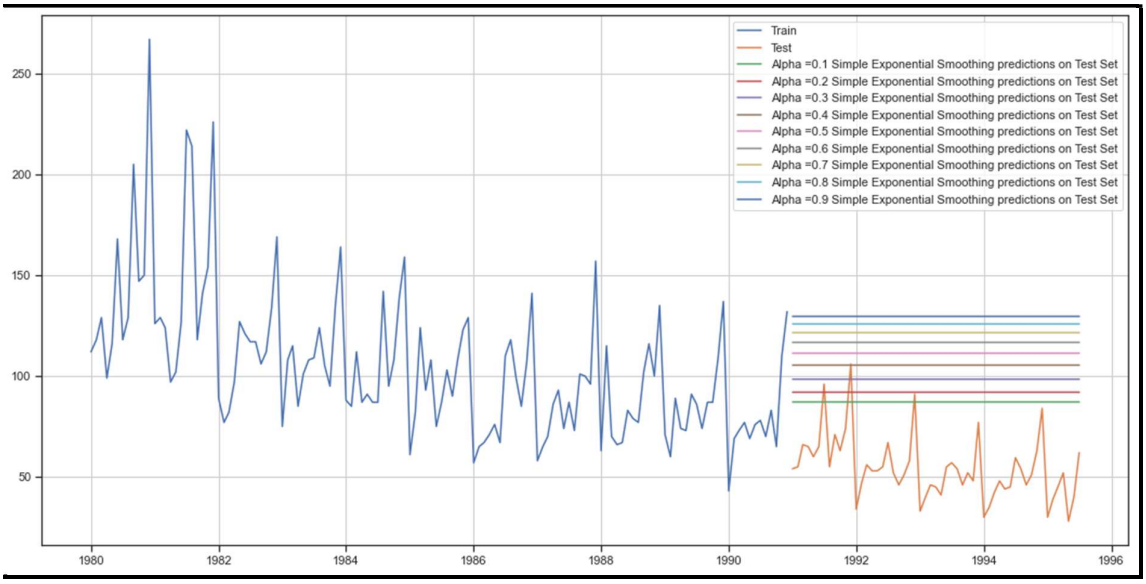
Rose Fig 15 Moving Average Forecast

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

2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	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.*

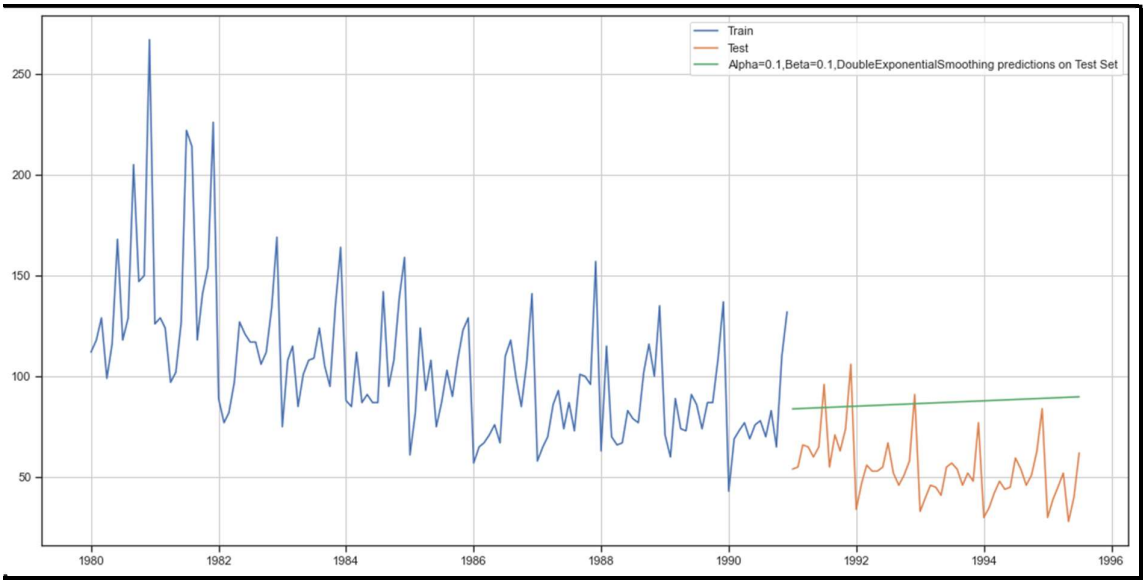
Model 5: Simple Exponential Smoothing



Rose Fig 16 Simple Exponential Smoothing

- *Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.*
- ***Alpha=0.1, SimpleExponentialSmoothing 36.429535***

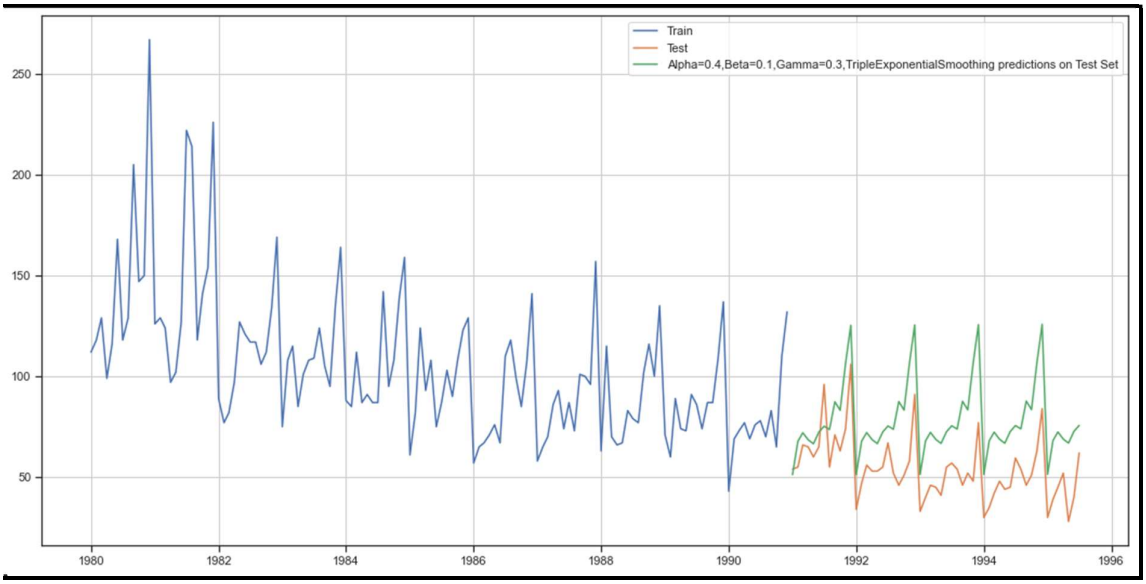
Model 6: Double Exponential Smoothing (Holt's Model)



Rose Fig 17 Double Exponential Smoothing Forecasting

- *Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.*
- ***Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing36.510010***

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



Rose Fig 18 Triple Exponential Smoothing Forecasting

- *Output for best alpha, beta and gamma values is shown by the green color line in the above plot. 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=0.3,TripleExponentialSmoothing 8.992350***

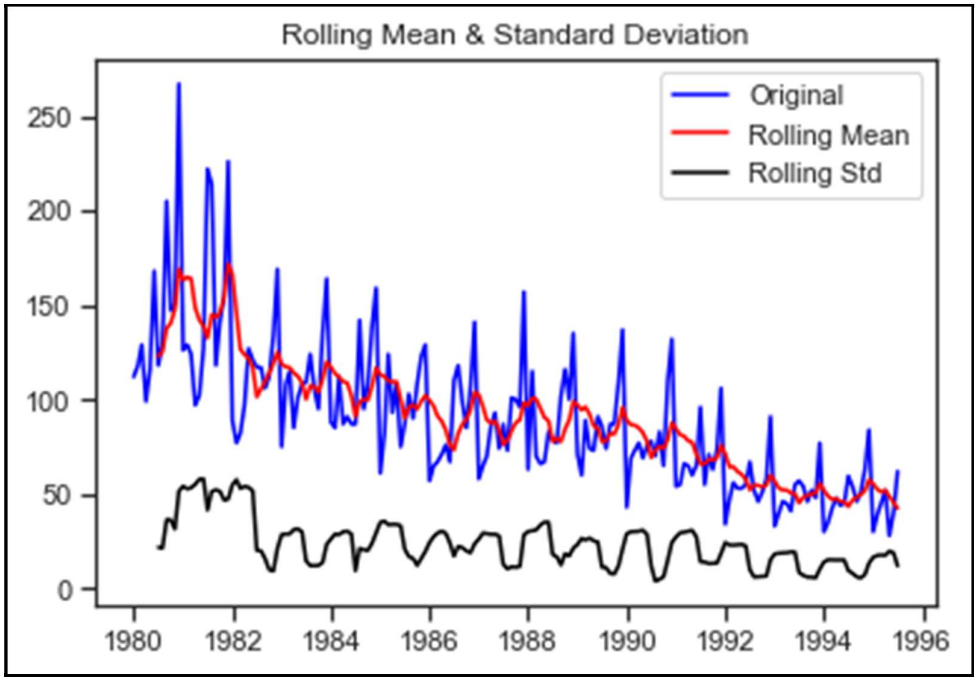
5.Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and 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.*

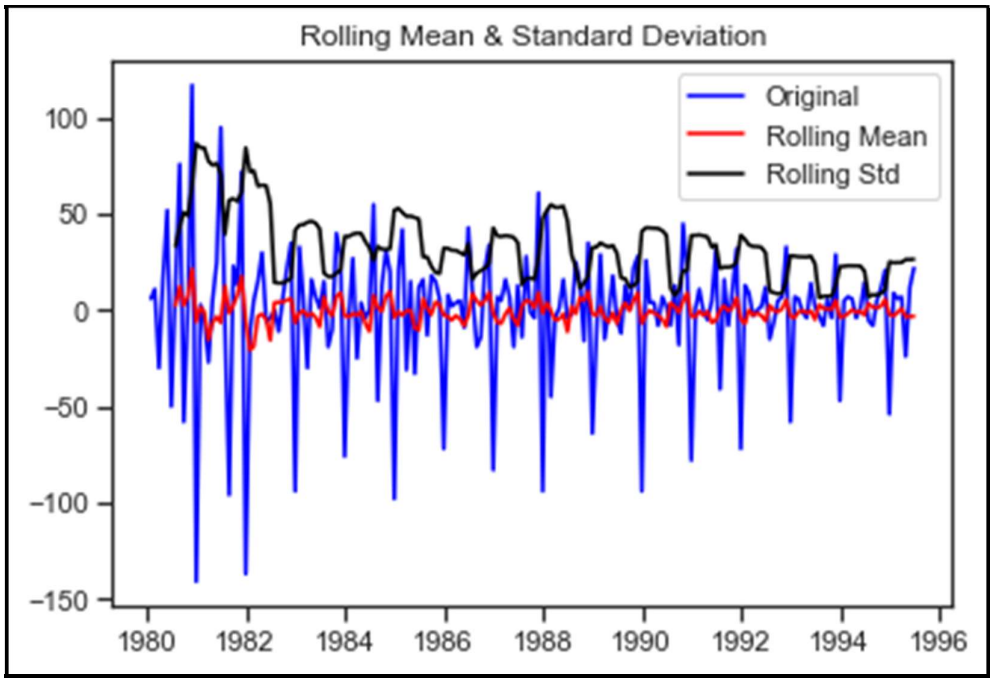


Rose Fig 19 Dickey-Fuller Test

Results of Dickey-Fuller Test:

- *Test Statistic -1.892338*
- *p-value 0.335674*
- *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.*



Rose Fig 20 Dickey 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.*

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)	1274.695149
15	(3, 1, 3)	1278.669962
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

- *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	-631.348
Date:	Thu, 06 Jul 2023	AIC	1274.695
Time:	16:43:58	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	-1.6777	0.084	-20.037	0.000	-1.842	-1.514
ar.L2	-0.7285	0.084	-8.701	0.000	-0.893	-0.564
ma.L1	1.0445	0.650	1.606	0.108	-0.230	2.319
ma.L2	-0.7724	0.134	-5.772	0.000	-1.035	-0.510
ma.L3	-0.9049	0.590	-1.533	0.125	-2.061	0.252
sigma2	858.9672	547.433	1.569	0.117	-213.981	1931.916

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.48
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

- ***RMSE values are: 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)	774.880935
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)	775.561019

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

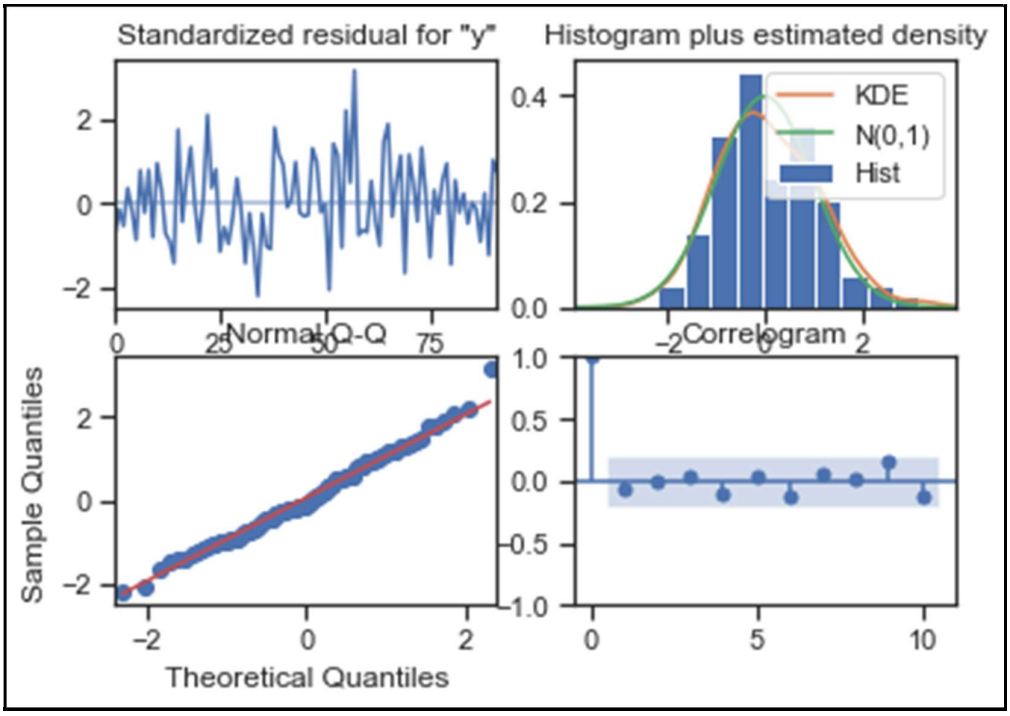
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-377.200
Date:	Thu, 06 Jul 2023	AIC	774.400
Time:	16:55:57	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	-0.201	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	-13.905	0.000	-1.069	-0.805
ar.S.L12	0.7639	0.165	4.640	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	192.1578	39.629	4.849	0.000	114.486	269.830

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

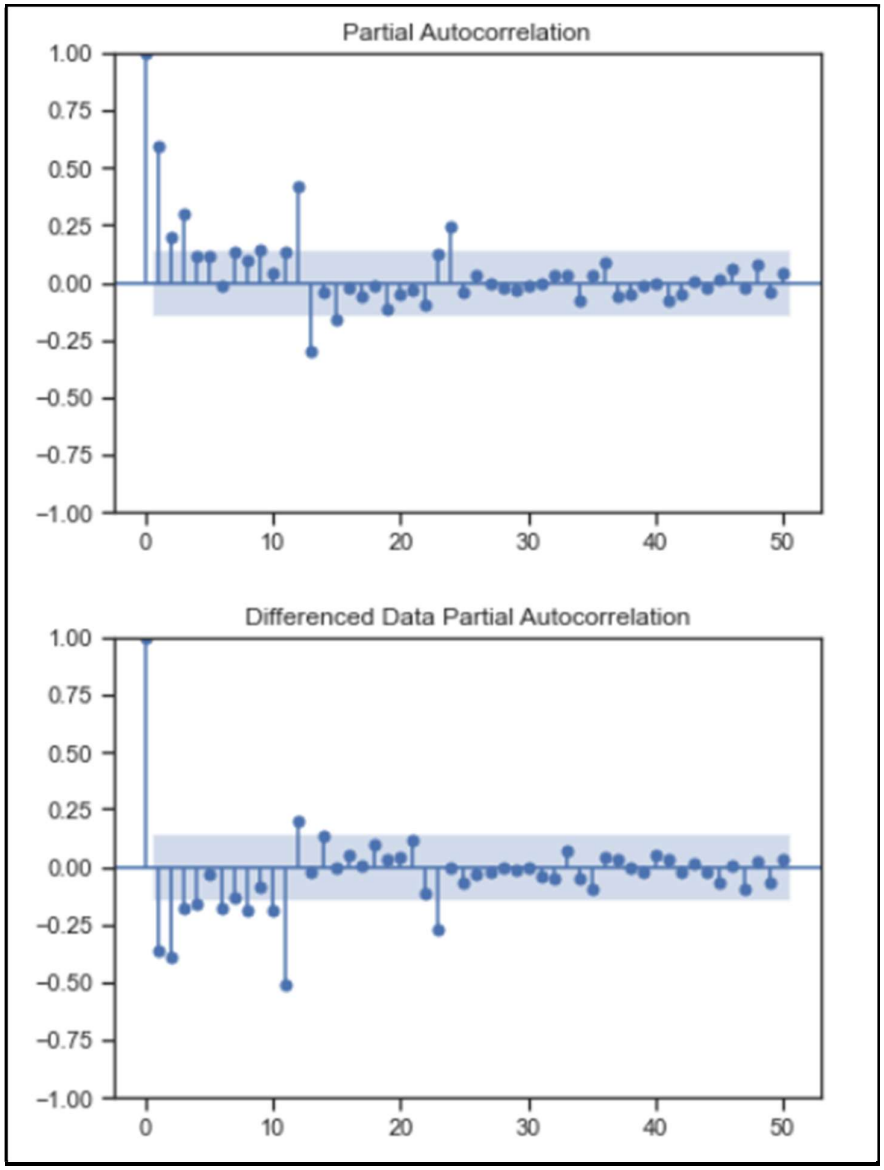
- *We also plotted the graphs for the residual to determine if any further information can be extracted or all the usable information has already been extracted.*
- *Below were the plots for the best auto SARIMA model.*



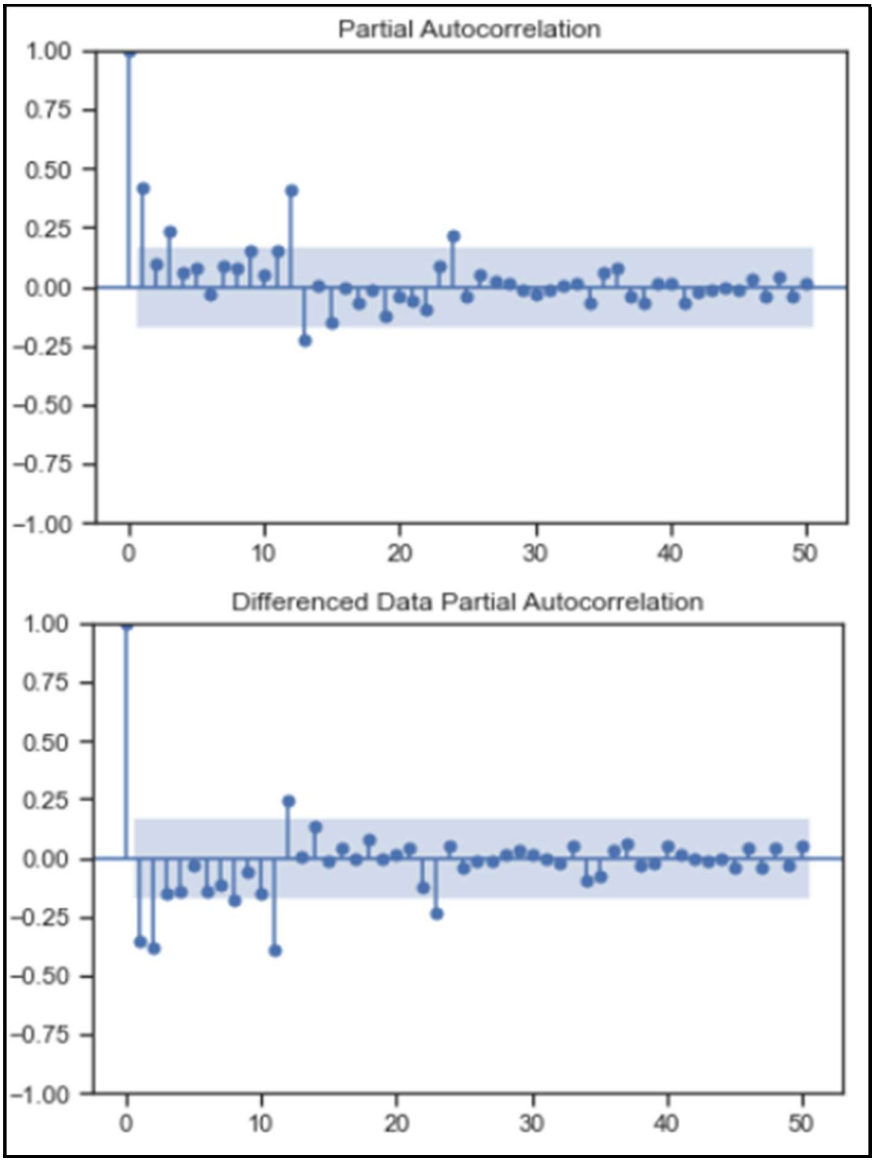
Rose Fig 21 SARIMA Plot

- ***RSME of Model: 18.53502803217281***

Manual- ARIMA Model



*Rose Fig 22 **PACF and ACF plots***



Rose Fig 23 PACF and ACF plot of train date

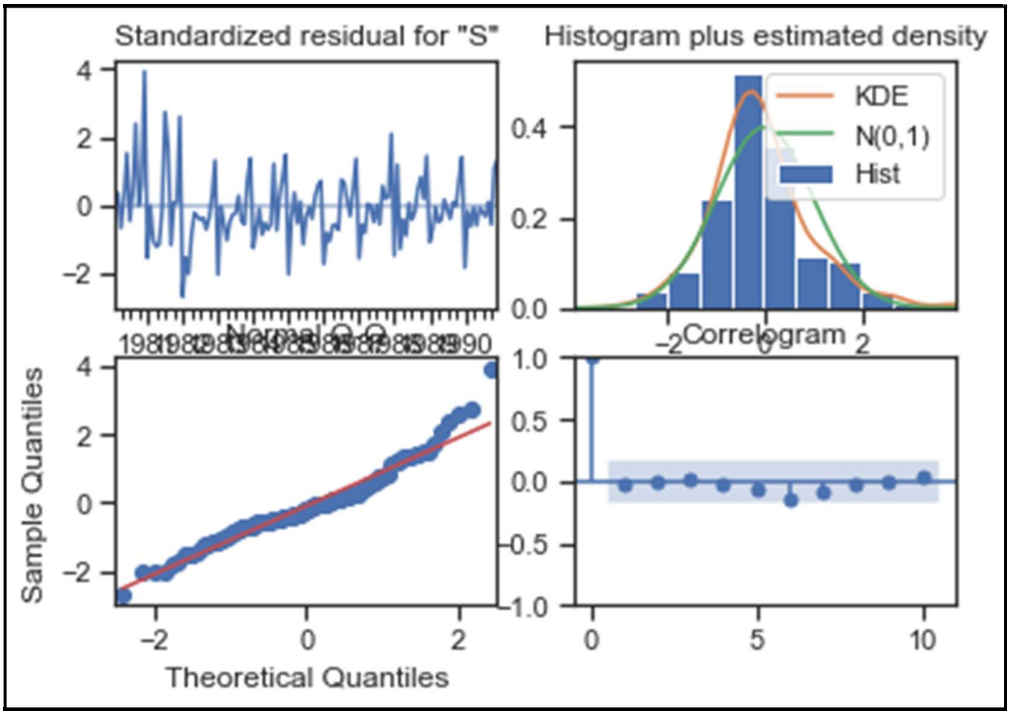
- *Hence the values selected for manual ARIMA:- p=2, d=1, q=2*
- *summary from this manual ARIMA model.*

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Thu, 06 Jul 2023	AIC	1281.871
Time:	16:56:01	BIC	1296.247
Sample:	01-01-1980	HQIC	1287.712
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4540	0.469	-0.969	0.333	-1.372	0.464
ar.L2	0.0001	0.170	0.001	0.999	-0.334	0.334
ma.L1	-0.2541	0.459	-0.554	0.580	-1.154	0.646
ma.L2	-0.5984	0.430	-1.390	0.164	-1.442	0.245
sigma2	952.1601	91.424	10.415	0.000	772.973	1131.347

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	34.16
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.79
Prob(H) (two-sided):	0.00	Kurtosis:	4.94



Rose Fig 24 Manual ARIMA model plots

- ***Model Evaluation: RSME 36.47322487814613***

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.*

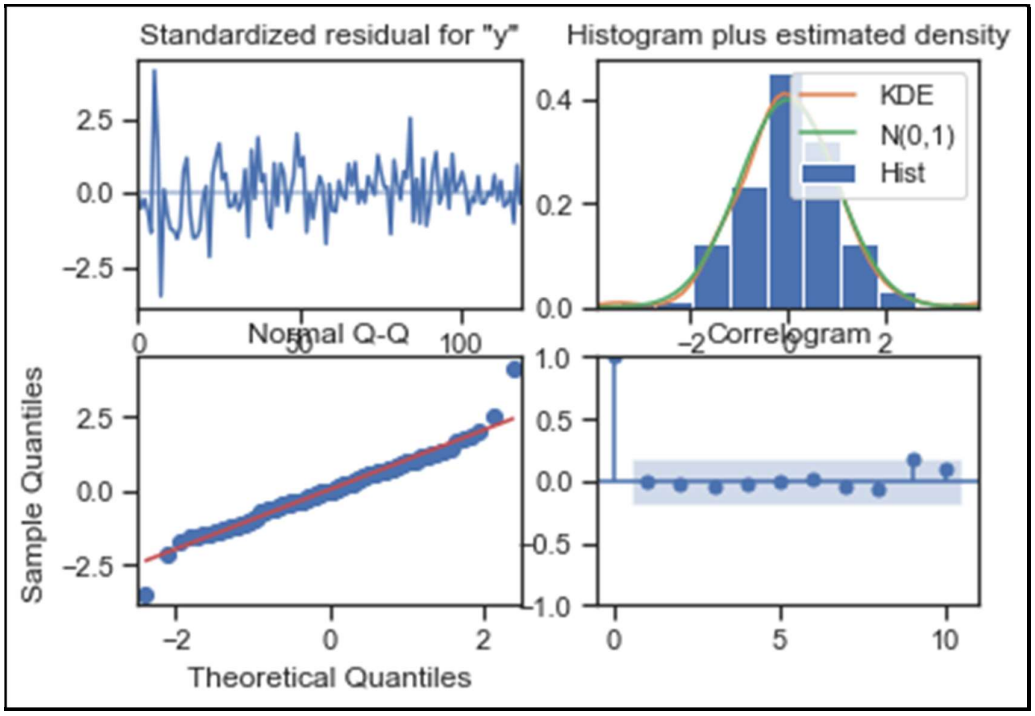
  SARIMAX(2, 1, 2)x(2, 1, 2, 12)
- *Below is the summary of the manual SARIMA model*

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-538.016
Date:	Thu, 06 Jul 2023	AIC	1094.031
Time:	16:56:08	BIC	1119.044
Sample:	0	HQIC	1104.188
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5490	0.228	-2.408	0.016	-0.996	-0.102
ar.L2	-0.0744	0.099	-0.752	0.452	-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	-1.0134	0.524	-1.935	0.053	-2.040	0.013
ar.S.L24	-0.1001	0.175	-0.572	0.568	-0.444	0.243
ma.S.L12	0.2911	47.514	0.006	0.995	-92.834	93.416
ma.S.L24	-0.7081	33.752	-0.021	0.983	-66.861	65.445
sigma2	430.2704	2.02e+04	0.021	0.983	-3.92e+04	4.01e+04

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	27.16
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	0.26
Prob(H) (two-sided):	0.00	Kurtosis:	5.28



Rose Fig 25 Manual SARIMA plots

- ***Model Evaluation: RSME 14.975041301618377***

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

	Test RMSE
Linear Regression	51.080941
Naive Model	79.304391
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	37.192623
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.992350
Auto_ARIMA	36.412720
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535503
ARIMA(3,1,3)	36.473225
(2,1,2)(2,1,2,12),Manual_SARIMA	14.973695

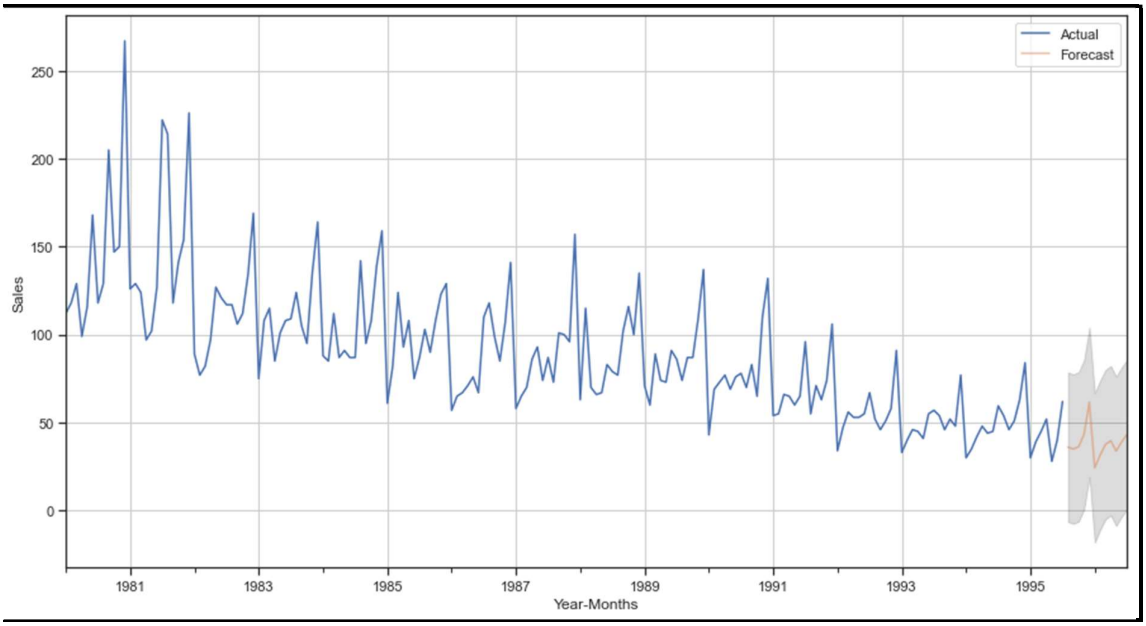
- *We can clearly see that triple exponential smoothing model with alpha 0.1, beta 0.7 and gamma 0.2 is the best as it he the lowest RSME score.*

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

- *The sales prediction on the graph along with the confidence intervals.*



Rose Fig 26 Prediction plot

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

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 analysis of the wine sales data indicates a clear downward trend for the Rose wine variety for the company, which has been declining in popularity for more than a decade.*
- *This trend is expected to continue in the future as well, based on the predictions of the most optimal model.*
- *Wine sales are highly influenced by seasonal changes, with sales increasing during festival season and dropping during peak winter time i.e. January.*
- *The company should consider running campaigns to boost the consumption of the wine during the rest of the year, as sales are subdued during this period.*
- *Campaigns during the lean period (April to June) might yield maximum results for the company, as sales are low during this period, and boosting them would increase the overall performance of the wine in the market across the year.*
- *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.*