PROJECT
TIME SERIES FORECAST

TABLE OF CONTENT

List of Figure and Table

Problem 1: Sparkling wine

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze and forecast Wine Sales in the 20th century.

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

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

Figure 1: Dataset

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

Figure 2: Description of data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    datetime64[ns]
 1   Sparkling  187 non-null    int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 3.0 KB
```

Figure 4: Info of data set

	Data Type
YearMonth	datetime64[ns]
Sparkling	int64

Figure 3 : Data type

Figure 5 :Time series plot

Inference

- The data type of first column YearMonth is datetime64 and sparkling is int64, first column is converted to index column
- There are two columns with 187 rows
- Sparkling column represents sales of sparkling wine
- No null value in the data

- From the description of the data, min value is 1070, max value is 7242.
- From plot of time series observed that data has seasonality and trend

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

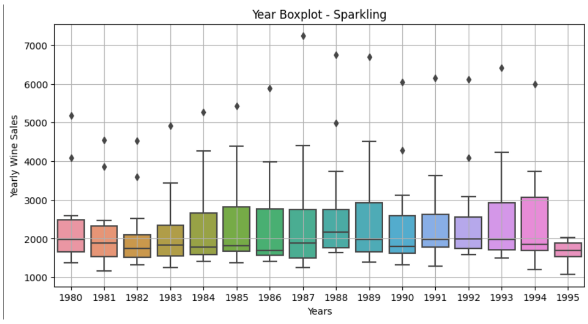


Figure 6: Boxplot of year sales

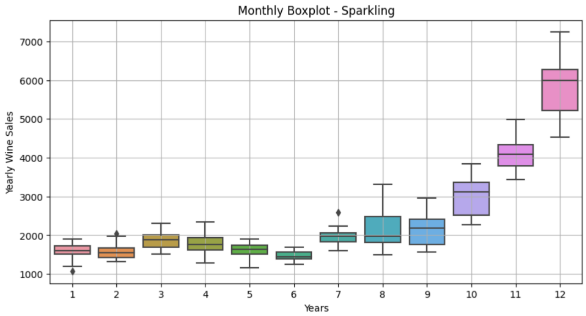


Figure 7: Boxplot of monthly sales

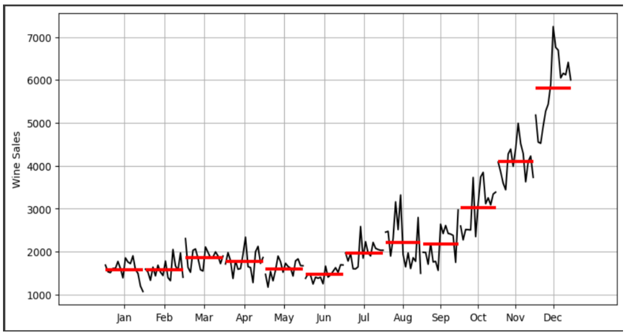


Figure7: Monthly time series

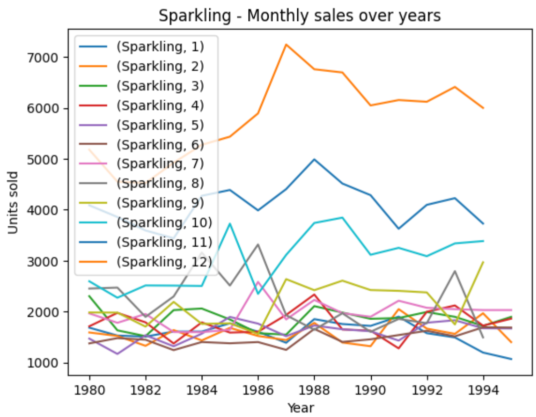


Figure8: Monthly sales over year

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

Figure9: Monthly sales over year table

Inference

- From fig 6 , Boxplot of year sales indicates the presence of outlier and presence of trend
- From Figure 7 Boxplot of monthly sales indicates few months has outlier and sales of wine rise over the months
- Figure7: Monthly time series, the sales of wine has no pattern upto September and then increased exponentially.
- Figure8 & 9: Monthly sales over year sales of wine increased from September onwards and reached highest by december.

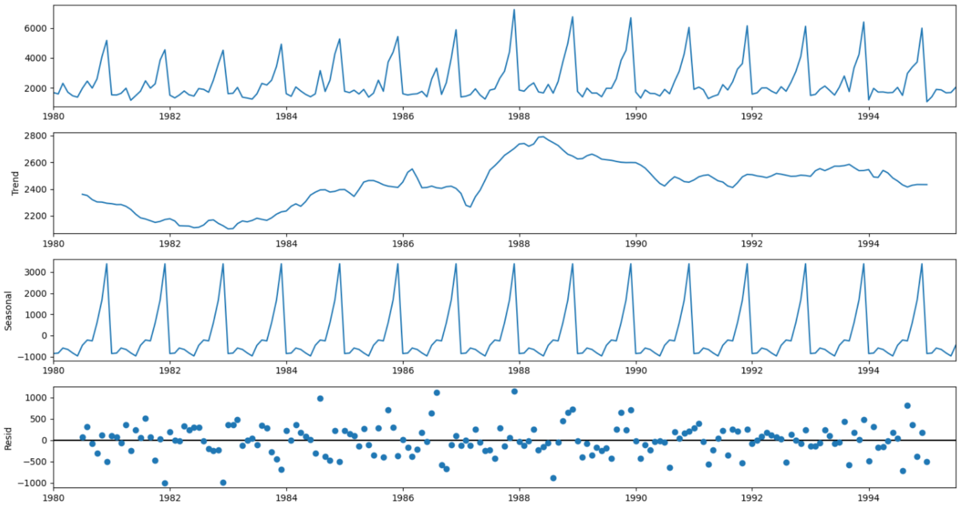


Figure10: Additive decomposition

- Upto 1988 trend is increasing then decreasing
- Seasonality is present is data.
- Residual has no pattern

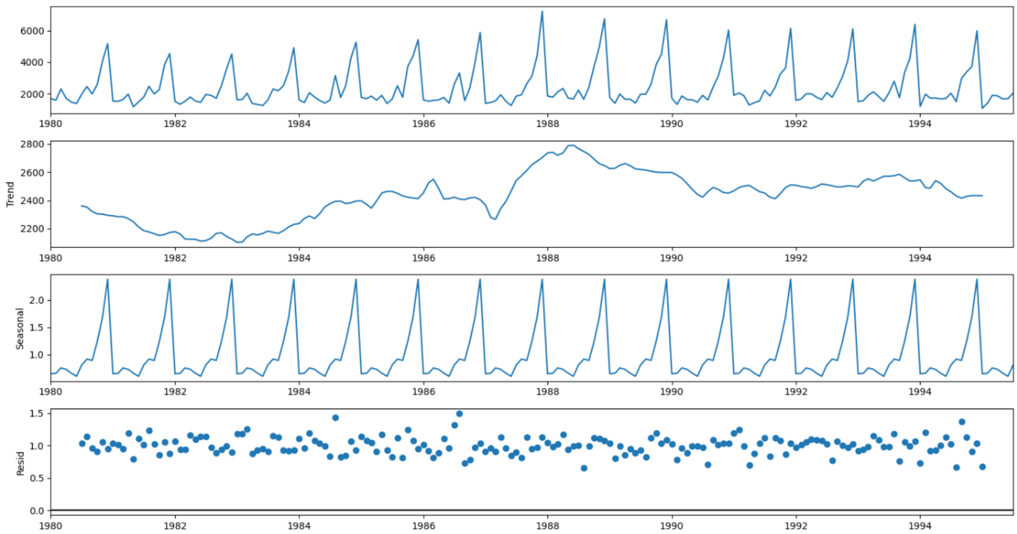


Figure11: Multiplication decomposition

- Upto 1988 trend is increasing then decreasing
- Seasonality is present is data.
- Residual has no pattern

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

The data is split to training and test data, and test data split starts at 1991.

First few rows of Training Data

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

Last few rows of Training Data

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

First few rows of Test Data

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Last few rows of Test Data

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

Fig 12: Head of train and test data Fig 13: Tail of train and test data

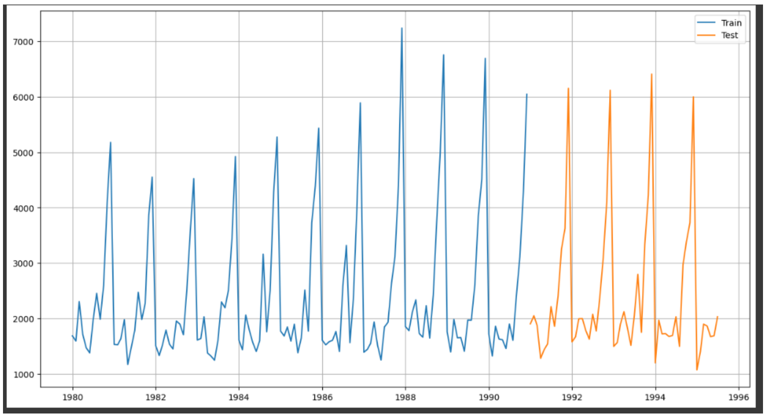


Fig 14: Plot of train and test data

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

Model 1: Linear Regression

First Few Rows of Train Data

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

Last Few Rows of Train Data

YearMonth	Sparkling	time
1990-08-01	1605	128
1990-09-01	2424	129
1990-10-01	3116	130
1990-11-01	4286	131
1990-12-01	6047	132

YearMonth	Sparkling	forecast_lr
1991-01-01	1902	2791.652
1991-02-01	2049	2797.485
1991-03-01	1874	2803.317
1991-04-01	1279	2809.150
1991-05-01	1432	2814.983

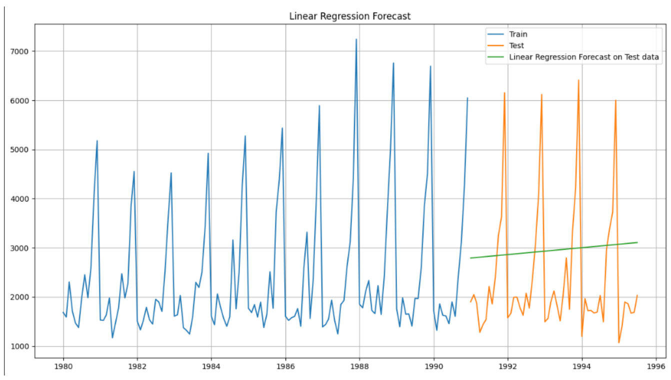


Fig 15: train and test data Fig 16: Sample forecast data Fig 17: Plot of forecasted sales in LD Model

- The data is split accordingly and numerical time instance is generated
- Head and tail is shown in fig 15
- Fig 16 is the forecasted value
- Figure 17 us the Plot of forecasted sales in LD Model , green line indicated LD which is not matching with actual value
- RMSE calculated is 1389.13

Model 2: Naive Approach

YearMonth	Sparkling	forecast_naive
1991-01-01	1902	6047
1991-02-01	2049	6047
1991-03-01	1874	6047
1991-04-01	1279	6047
1991-05-01	1432	6047

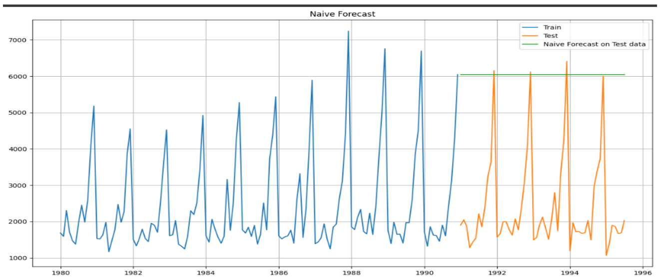


Fig 18: Forecasted value table Figure 19: Plot of forecasted sales in Naive model

- The forecast value is same throughout
- Forecasted value is a straight line and not matching with actual value
- RMSE is 3864.27

Model 3: Simple average approach

YearMonth	Sparkling	forecast_sa
1991-01-01	1902	2403.78
1991-02-01	2049	2403.78
1991-03-01	1874	2403.78
1991-04-01	1279	2403.78
1991-05-01	1432	2403.78

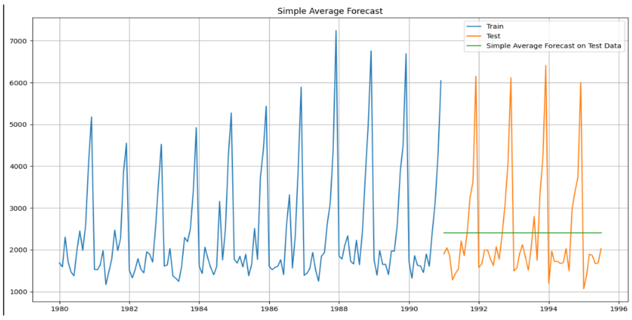


Fig 20: Forecasted value table Figure 21: Plot of forecasted sales in Simple average model

- The forecast value is same throughout
- Forecasted value is a straight line and not matching with actual value
- RMSE is 1275.08

Model 4: Simple Exponential Smoothing

YearMonth	Sparkling	forecast_ses_optimized
1991-01-01	1902	2804.663
1991-02-01	2049	2804.663
1991-03-01	1874	2804.663
1991-04-01	1279	2804.663
1991-05-01	1432	2804.663

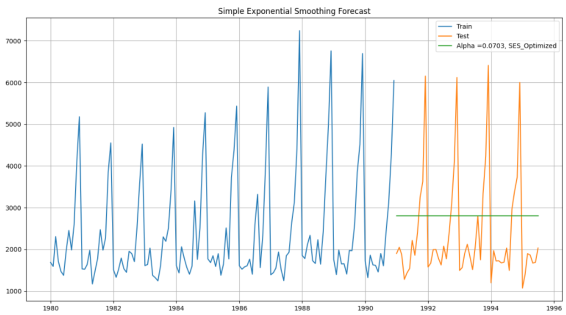


Fig 22: Forecasted value table Figure 23: Plot of forecasted sales in SES

- The model doesnt consider trend and seasonality
- The forecasted value is same throughout
- Forecasted value is a straight line and not matching with actual value
- RMSE is 1338

Optimizing Alpha based on Test RMSE

- Different values are used to find the lowest RMSE. From it best alpha is 0.1
- Forecasted value is a straight line and not matching with actual value
- RMSE is 1375.39

	Alpha_Values	RMSE_Train	RMSE_Test
0	0.1	1336.428	1375.393
1	0.2	1356.950	1595.207
2	0.3	1359.953	1935.507
3	0.4	1352.862	2311.920
4	0.5	1344.203	2666.351

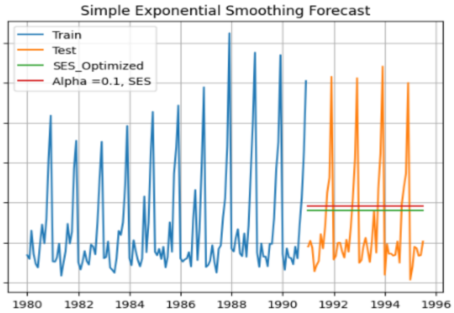


Fig 24: Forecasted value table   Figure 25: Plot of forecasted sales in SES optimized alpha

Model5: Double Exponential Smoothing

	Sparkling	forecast_des_optimized
YearMonth		
1991-01-01	1902	5401.733
1991-02-01	2049	5476.005
1991-03-01	1874	5550.277
1991-04-01	1279	5624.550
1991-05-01	1432	5698.822

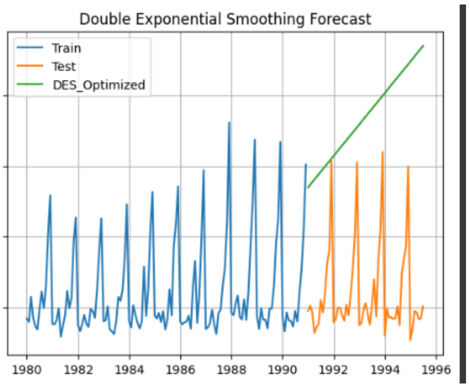


Fig 26: Forecasted value table   Figure 27: Plot of forecasted sales in DES

	Sparkling	forecast_des_optimized	forecast_des
YearMonth			
1991-01-01	1902	5401.733	2847.662
1991-02-01	2049	5476.005	2874.629
1991-03-01	1874	5550.277	2901.596
1991-04-01	1279	5624.550	2928.563
1991-05-01	1432	5698.822	2955.530

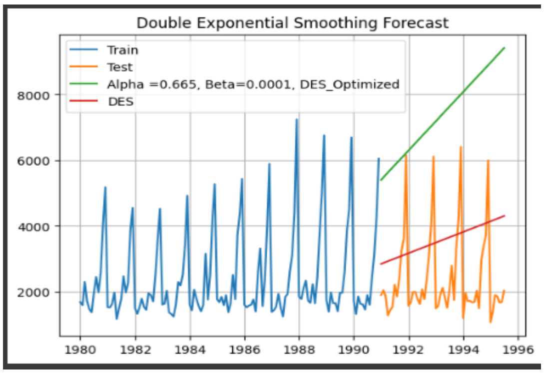


Fig 28: Forecasted value table   Figure 29: Plot of forecasted sales in DES optimized alpha, beta

- This model has trend but no seasonality the rmse without optimisation is 5291.87
- After optimizing alpha and beta between 0 to 1. The best model is optimised, RMSE : 1777.73

Fig 30: Triple Exponential Smoothing with Additive trend & Additive seasonality
Fig 31: Triple Exponential Smoothing with Additive trend & Multiplication seasonality

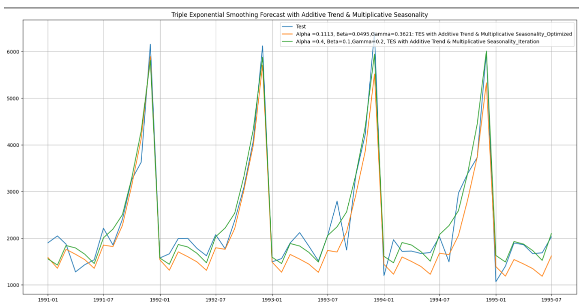
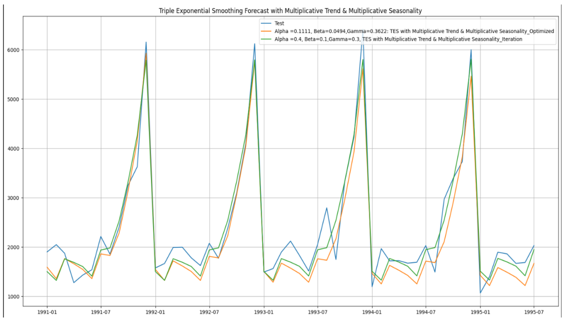
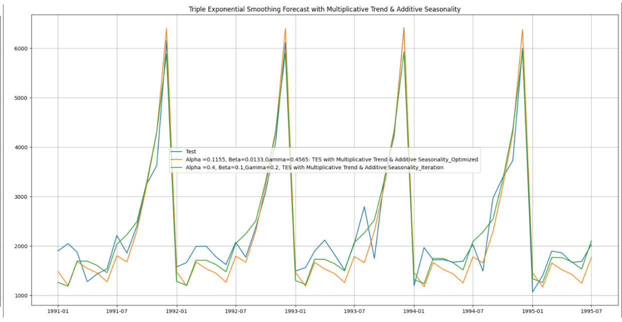


Fig 32: Triple Exponential Smoothing with Multiplication trend & Additive seasonality

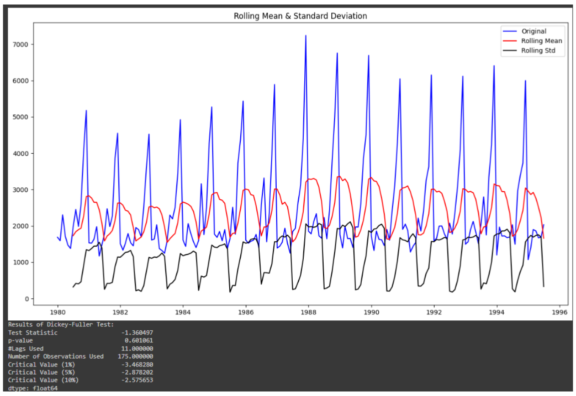
Fig 33: Triple Exponential Smoothing with Multiplication trend & Multiplication seasonality





- RMSE of Triple Exponential Smoothing with Additive trend & Additive seasonality–342.93
- RMSE of Triple Exponential Smoothing with Additive trend & Multiplication seasonality–317.4
- RMSE of Triple Exponential Smoothing with Multiplication trend & Additive seasonality-326.57
- RMSE of Triple Exponential Smoothing with Multiplication trend & Multiplication seasonality-341.65

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.



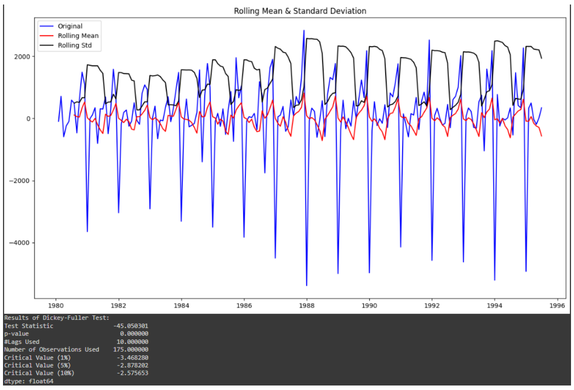


Fig 34 Dickey Fuller test on whole time series Fig 35 Dickey Fuller test on differenced whole time series

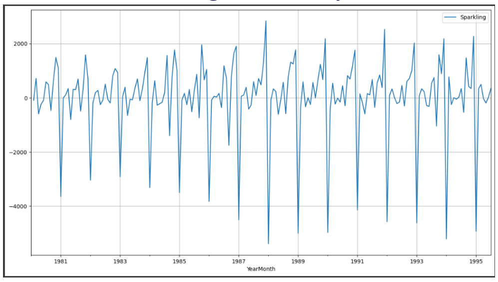
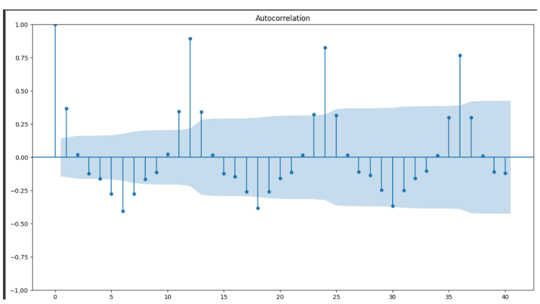


Fig 36:Plot of differenced whole time series



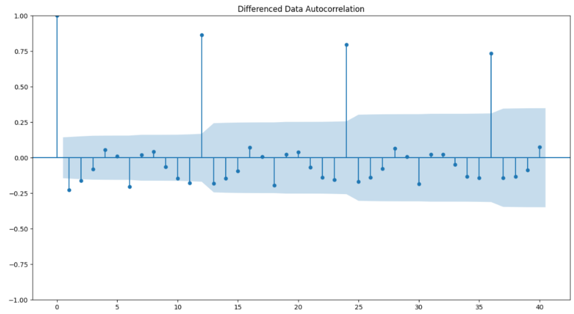
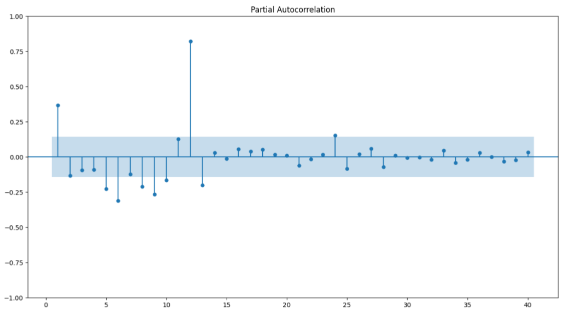


Fig 37:Autocorrelation plot of whole time series Fig 38:Autocorrelation plot of differenced whole time series



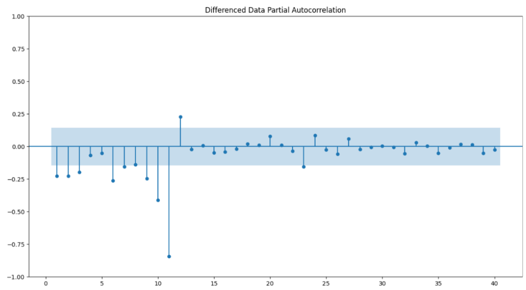
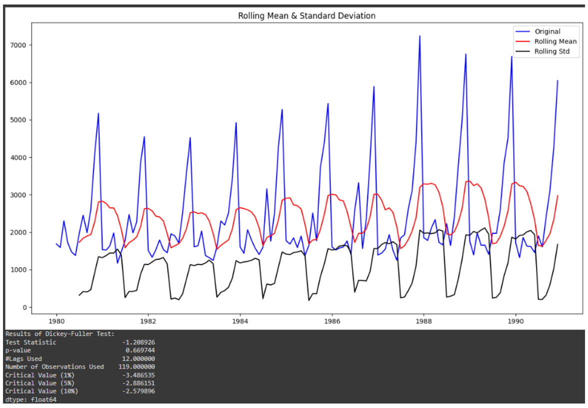


Fig 39:Partial Autocorrelation plot of whole time series Fig 40: Partial Autocorrelation plot of differenced whole time series



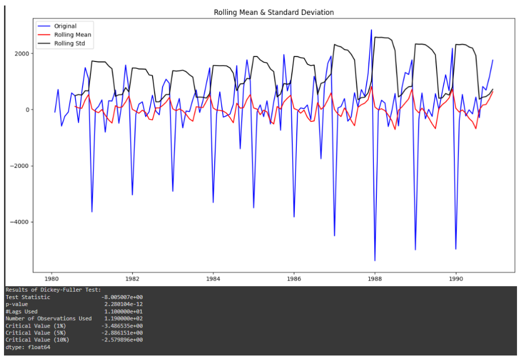


Fig 41: Stationarity of Train Dataset Fig 42: Stationarity of differencing Train Dataset

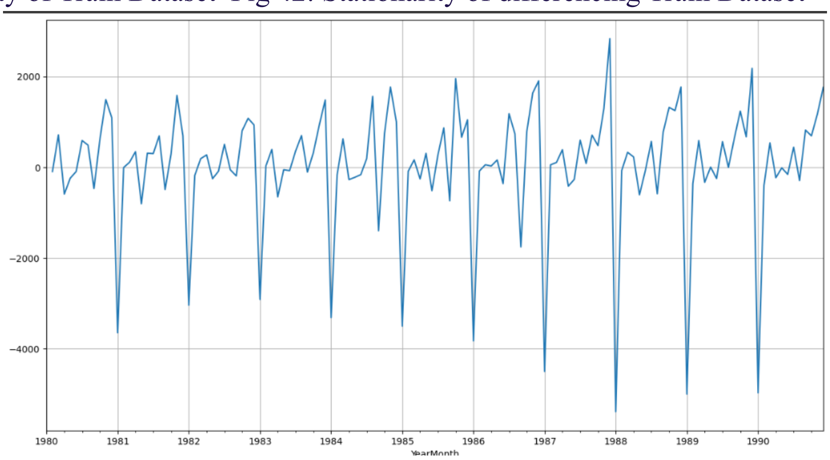


Fig 43: Differenced Time series

Inteference

- ADF is a test to find if the model is stationary or not
- Ho states model is non stationary, H1 states model is stationary

- If alpha is <=0.5 then null value can be disproved
- Fig 34 ,p is 0.6 means model whole time series is non stationary
- Fig 35 ,p is 0.66 means model Partial Autocorrelation plot of whole time series is non stationary
- Fig 34 ,p is 2.2 means model Partial Autocorrelation plot of differenced whole time series
- is non 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.

The arima model is created with p =0 to 3, q = 0 to 3, d =1

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sat, 22 Apr 2023	AIC	2213.509
Time:	13:36:04	BIC	2227.885
Sample:	01-01-1980	HQIC	2219.351
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.782	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

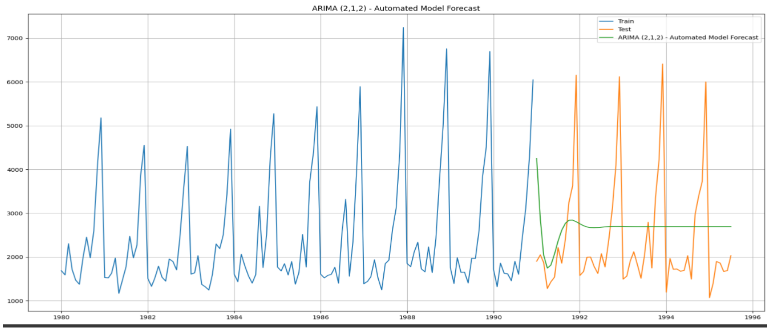
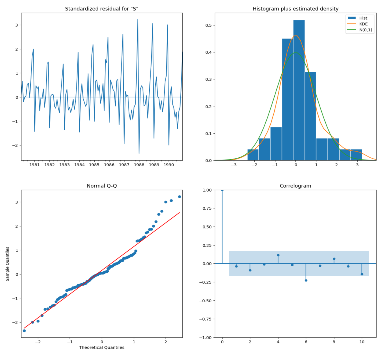


Fig 44: summary automated arima model Fig 45 Plot of forecasted sales automated arima model



	param	AIC
10	(2, 1, 2)	2213.5
15	(3, 1, 3)	2221.5
14	(3, 1, 2)	2230.8
11	(2, 1, 3)	2232.8
9	(2, 1, 1)	2233.8

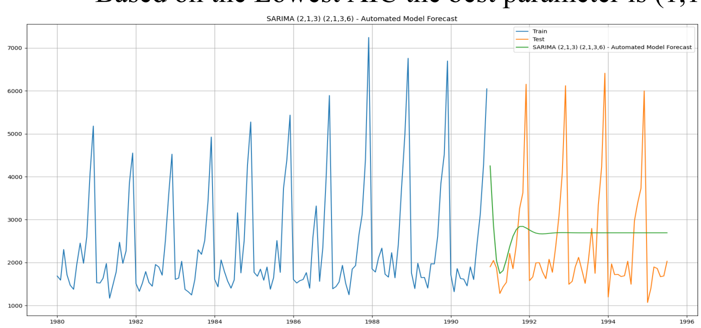
Fig 46:diagnostic automated arima model Fig 47:Order based on AIC

- From this observed actual not matching with forecast, it is thus not a good model for forecasting.
- RMSE is 1299.9
- Based on the Lowest AIC the best parameter is (2,1,2)

SARIMA Model

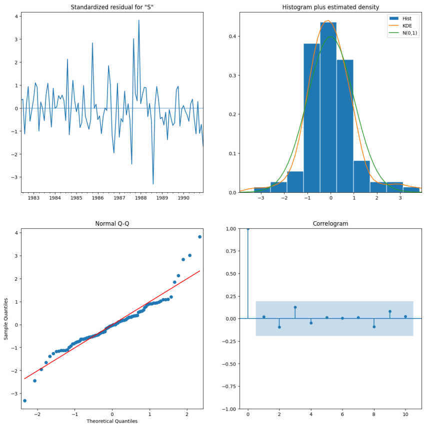
Following value p between 0 to 3, q between 0 to 3, d between 0 to 1

- From this observed actual not matching with forecast, it is thus not a good model for forecasting.
- RMSE is 784.1
- Based on the Lowest AIC the best parameter is (1,1,3,6) AIC = 1541.6



	param	param_seasonal	AIC
495	(3, 1, 3)	(1, 1, 3, 6)	1541.6
111	(0, 1, 3)	(1, 1, 3, 6)	1544.0
367	(2, 1, 3)	(1, 1, 3, 6)	1545.7
503	(3, 1, 3)	(2, 1, 3, 6)	1545.8
239	(1, 1, 3)	(1, 1, 3, 6)	1545.9

Fig 48: Plot of forecasted sales automated sarima model Fig 51::Order based on AIC



SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sat, 22 Apr 2023	AIC	2213.509
Time:	13:36:04	BIC	2227.885
Sample:	01-01-1980	HQIC	2219.351
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.782	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

Fig 49:diagnostic automated sarima model Fig 50: summary automated sarima model

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

	Test RMSE
Triple Exponential Smoothing Forecast with Additive Trend & Multiplicative Seasonality Alpha =0.4, Beta=0.1,Gamma=0.2	317.434
Triple Exponential Smoothing Forecast with Multiplicative Trend & Multiplicative Seasonality Alpha =0.4, Beta=0.1,Gamma=0.3	326.580
Triple Exponential Smoothing Forecast with Multiplicative Trend & Additive Seasonality Alpha =0.4, Beta=0.1,Gamma=0.2	341.654
Triple Exponential Smoothing Forecast with additive trend & additive seasonality Alpha =0.1, Beta=0.4,Gamma=0.1,	342.935
Triple Exponential Smoothing Forecast with Additive Trend & Additive Seasonality	378.626
Triple Exponential Smoothing Forecast with Multiplicative Trend & Additive Seasonality	379.646
Triple Exponential Smoothing Forecast with Multiplicative Trend & Multiplicative Seasonality	380.393
Triple Exponential Smoothing Forecast with Additive Trend & Multiplicative Seasonality	403.706
forecast_SARIMA_auto	867.384
Simple Average	1275.082
forecast_ARIMA_auto	1299.980
Simple Exponential Smoothing Forecast	1338.005
Simple Exponential Smoothing Forecast with alpha =0.1	1375.393
Linear Regression	1389.135
Double Exponential Smoothing Forecast Alpha =0.665, Beta=0.0001	1777.735
Naive Forecast	3864.279
Double Exponential Smoothing Forecast	5291.880

Fig 51: summary of all models

From the above the table, the best model is the triple exponential smoothing with additive trend and multiplicative seasonality with the parameters with alpha = 0.4, beta =0.1 and gamma = 0.2

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.

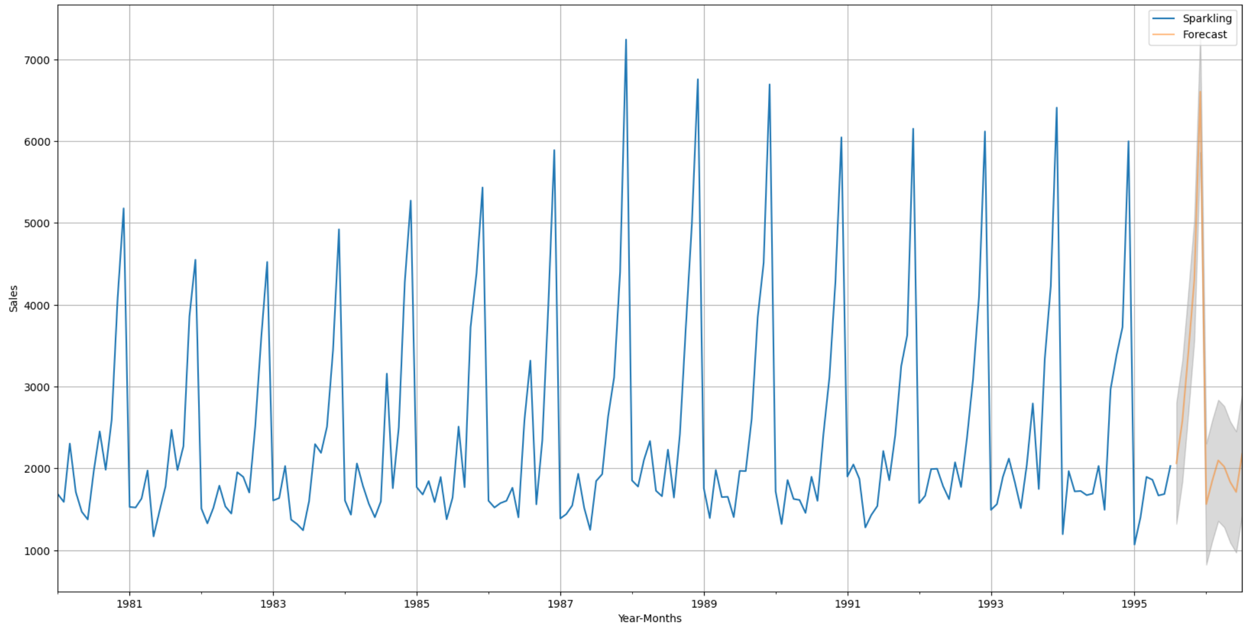


Fig 52: Forecast with 95% confidence to 12 month in future

	lower_CI	Forecast	upper_ci
1995-08-01	1323.0	2063.4	2803.9
1995-09-01	1838.9	2579.4	3319.9
1995-10-01	2676.2	3416.7	4157.1
1995-11-01	3564.0	4304.5	5044.9
1995-12-01	5864.4	6604.9	7345.3
1996-01-01	824.1	1564.5	2305.0
1996-02-01	1109.3	1849.8	2590.2
1996-03-01	1358.4	2098.9	2839.3
1996-04-01	1282.0	2022.4	2762.9
1996-05-01	1094.1	1834.5	2575.0
1996-06-01	971.9	1712.4	2452.9
1996-07-01	1436.0	2176.4	2916.9

Fig 53: Forecast with 95% confidence to 12 month in future tabl

Best model is the triple exponential smoothing with additive trend and multiplicative seasonality with the parameters with alpha = 0.4, beta =0.1 and gamma = 0.2

- RMSE is 377.3
- The value is forecasted to 12 month interval with this model

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 Best model is the triple exponential smoothing with additive trend and multiplicative seasonality with the parameters with alpha = 0.4, beta =0.1 and gamma = 0.2
- RMSE is 377.3
- Plot of 12 month forecast is following similar pattern as original data
- Sales is increasing gradually then exponentially as per the pattern observed the actual data
- We can expect that sales has no pattern till september

Measures to increase sales:

- During peak season maximum stock should be available for more sales
- Find out the reason why sales is less in other months

- Since the trend is upwards the forecasted value is good