TIME SERIES FORECASTING PROJECT REPORT

Problem 1: Sparkling Wine

By Karthik Sreeram R

MAY 22, 2021

UNIVERSITY OF TEXAS AT AUSTIN AND GREAT LAKES

Purpose
This document is the business report for my final project in the subject “Time Series Forecasting” This document gives us a detailed explanation of various approaches used, their insight and inferences. Tools used analysis: Python and Jupiter notebook. Packages used: NumPy, pandas, seaborn, os, matplotlib, stats model, sklearn and pylab

Contents

Problem 1: Sparkling Wine

Business scenario

For this assignment, the data of different types of wine sales in the 20th century is to be analyzed. Both 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.

Introduction:

The purpose of this assignment is to understand the time series data, do exploratory analysis, perform decomposition to understand the trend and seasonality of the data , train the data with different models of forecasting to predict the future sales of the sparkling wine .It will help the ABC estate to pre stock the sparkling wine for future sales based on demand (predicted sales).

Data has two fields.

YearMonth – monthly data

Sparkling – sales count of sparkling wine.

1.1) Reading the data as a Time Series data and plotting the data

a.) Dataset Head

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

Inference:

The data is month wise data starting from January 1980. The data format is year and month (YYYY-mm)

b.) Dataset Tail

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

Inference:

The data ends in July 1995. The row count and total number of months between January 1980 and July 1995 matches (i.e., 187 months)

c.) Preprocessing of time series data

1. Creating dummy month wise date data in time stamp format from the January 1980 to July 1995.

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

2. Add dummy data column to the original data set.

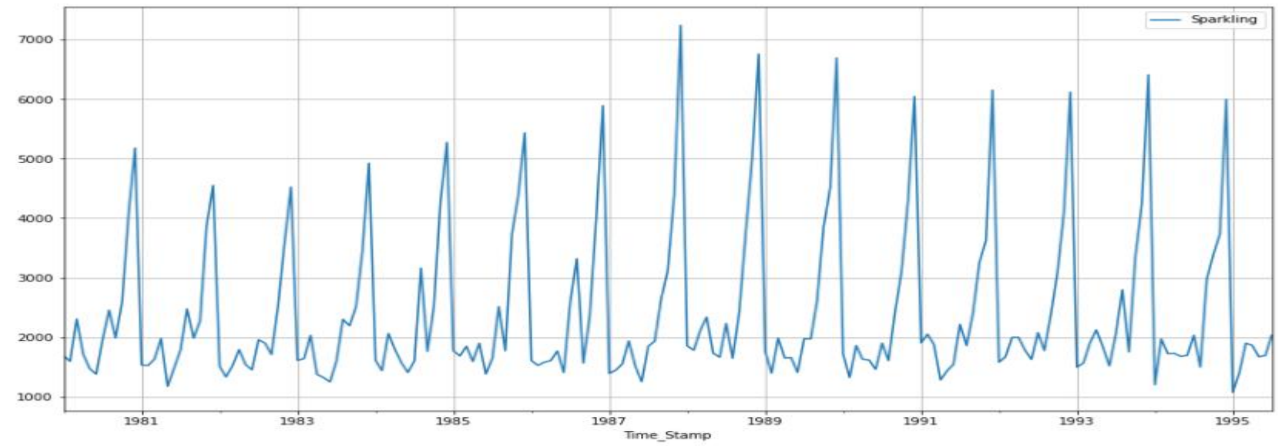
4]:

	YearMonth	Sparkling	Time_Stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

Inference:

For time series model data must be in the format YYYY-mm-dd i.e., time stamp format. Our data has date data as YYYY-mm format. So, a dummy time stamp is created to replace the YearMonth field and changed into index.

c.) Plotting the time series data.



Inference:

There is a seasonality in the data. Looks like there is no upward or downward trend in the data. There is a seasonality in the data.

1.2) Performing Exploratory Data Analysis to understand the data and to perform decomposition.

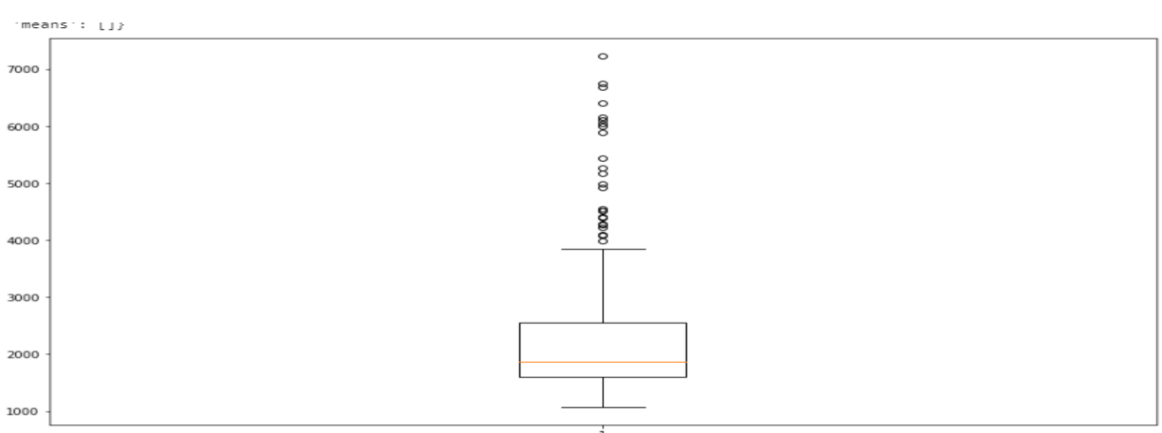
a) Description of Data

Basic Descriptive Stats of Time series

9]:

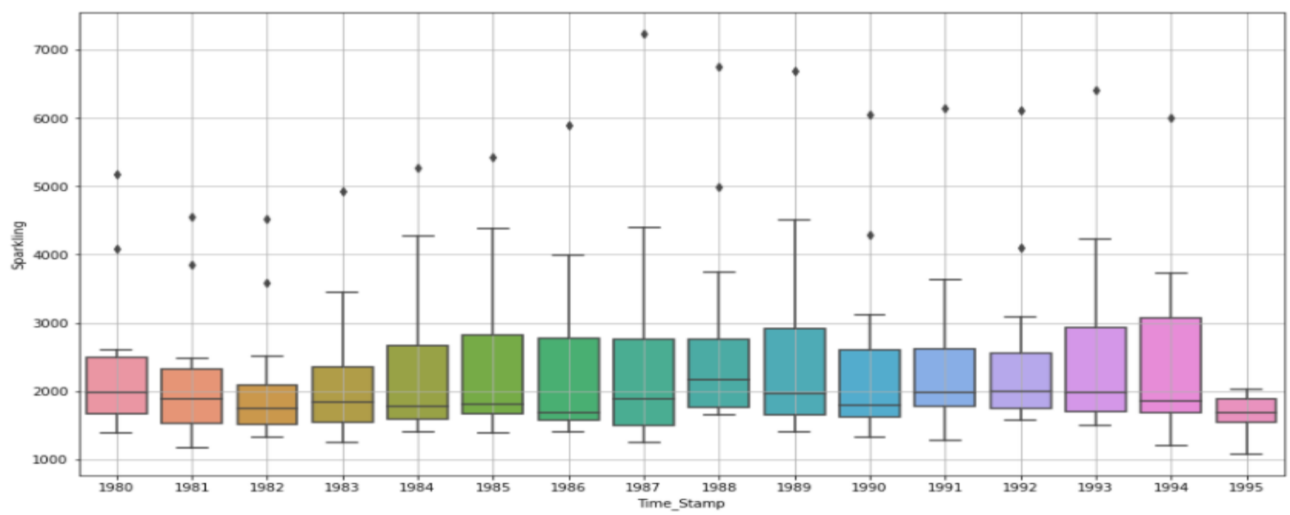
	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

b) BOX PLOT :



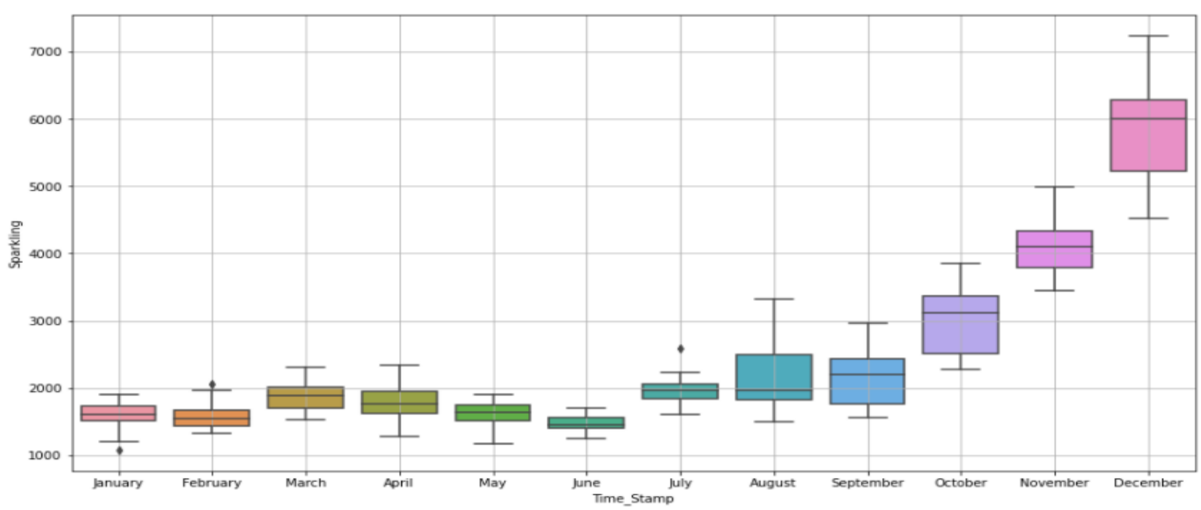
Inference:
More than 75 percent of sales quantity fall below 2549. Average sale count is 2402. The count of Sparkling field is 187 which is equal to number of months. so There is no missing data for sparkling field (sales).

c.) Box plot yearly



Inference:
Year over year comparison shows that the data has no uniform (increase or decrease) patterns. I.e., no clear trend It keeps changing.

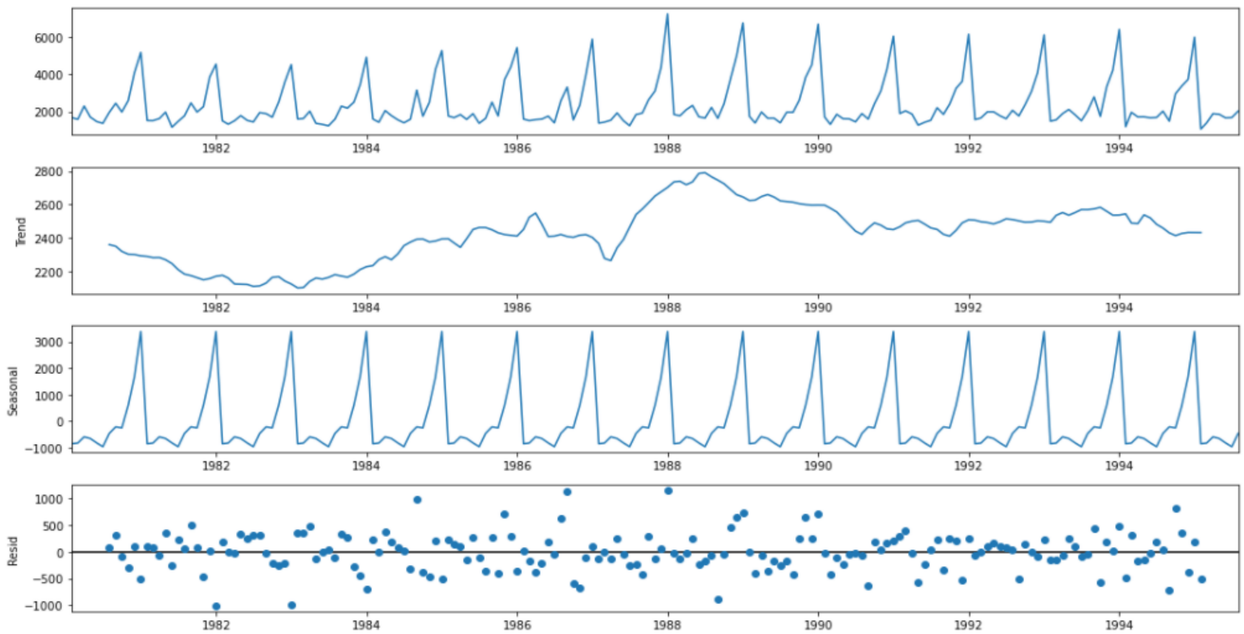
d) Box plot Monthly



Inference:
Month over month comparison shows that the data seasonality. It has increasing sales trend from July to December and decreasing sales trend from January to June . The maximum sales are during December and minimum sales is in June.

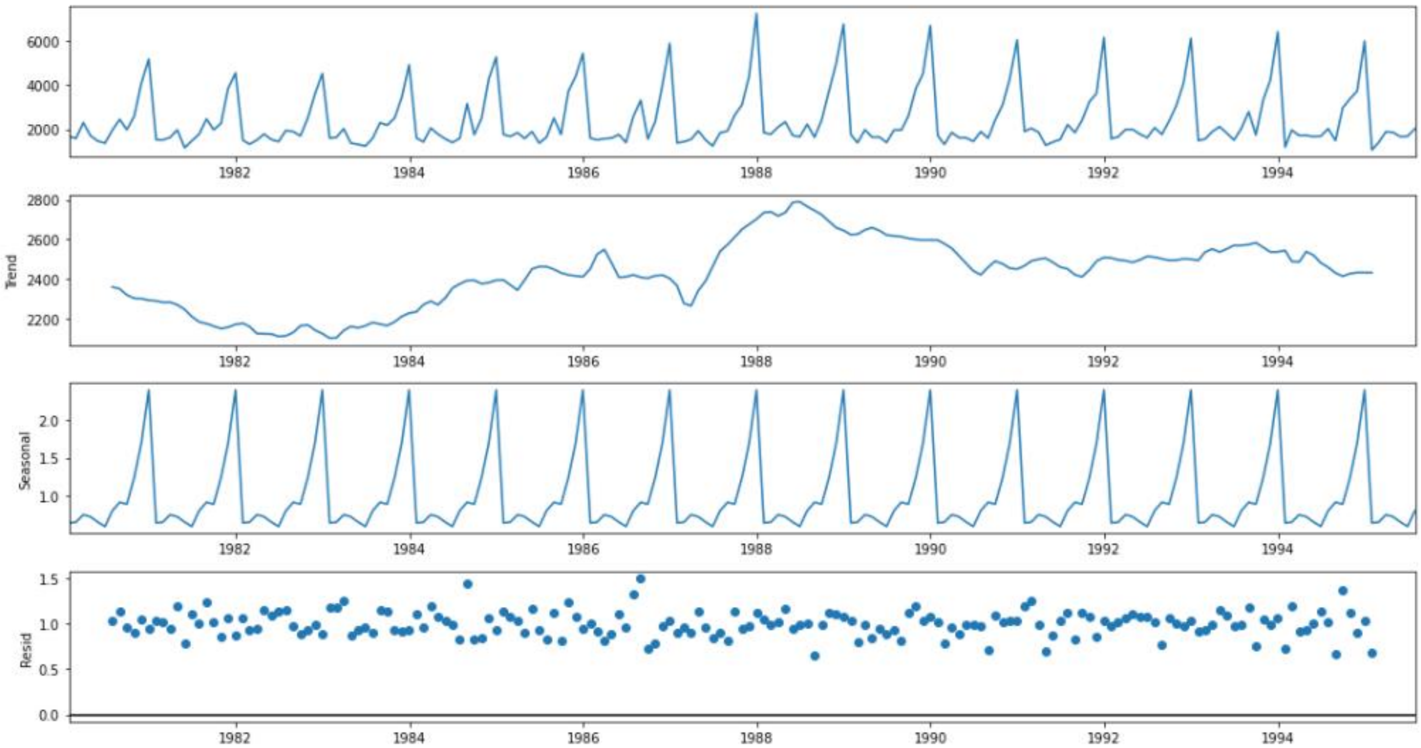
e) Decomposition of Sparkling variable time series

1. Addictive decomposition




```
Trend
 Time_Stamp
1980-01-31            NaN
1980-02-29            NaN
1980-03-31            NaN
1980-04-30            NaN
1980-05-31            NaN
1980-06-30            NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: trend, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    -854.260599
1980-02-29    -830.350678
1980-03-31    -592.356630
1980-04-30    -658.490559
1980-05-31    -824.416154
1980-06-30    -967.434011
1980-07-31    -465.502265
1980-08-31    -214.332821
1980-09-30    -254.677265
1980-10-31     599.769957
1980-11-30    1675.067179
1980-12-31    3386.983846
Name: seasonal, dtype: float64

Residual
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31     70.835599
1980-08-31    315.999487
1980-09-30    -81.864401
1980-10-31   -307.353290
1980-11-30    109.891154
1980-12-31   -501.775513
Name: resid, dtype: float64
```

2. Multiplicative Decomposition




```
Trend
 Time_Stamp
1980-01-31            NaN
1980-02-29            NaN
1980-03-31            NaN
1980-04-30            NaN
1980-05-31            NaN
1980-06-30            NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: trend, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
1980-06-30    0.603468
1980-07-31    0.809164
1980-08-31    0.918822
1980-09-30    0.894367
1980-10-31    1.241789
1980-11-30    1.690158
1980-12-31    2.384776
Name: seasonal, dtype: float64

Residual
 Time_Stamp
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    1.029230
1980-08-31    1.135407
1980-09-30    0.955954
1980-10-31    0.907513
1980-11-30    1.050423
1980-12-31    0.946770
Name: resid, dtype: float64
```

Inference:

There is no pattern in the residual. seasonality and residual components are independent of the trend. So, it is addictive.

f) Checking stationarity of whole data

```
checking seasonlity on whole data
DF test statistic is -1.798
DF test p-value is 0.7055958459932397
Number of lags used 12
fail to reject null hypothesis . its not stationary (p value is > 0.05 (alpha))
```

Inference:
Whole Data is not stationary at α = 0.05

1.3) Splitting the data into training and test. The test data starts in 1991.

a) Head and tail of train data

Head of train data

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Tail of train data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

b) Head and tail of test data

Head of test data

Time_Stamp	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Tail of test data

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

Inference
Data is split into train and split. Train data is from January 1980 December 1990. Test data is from Jan 1991.

1.4) Building various time series models on the training data and evaluating the model performance using RMSE on the test data.

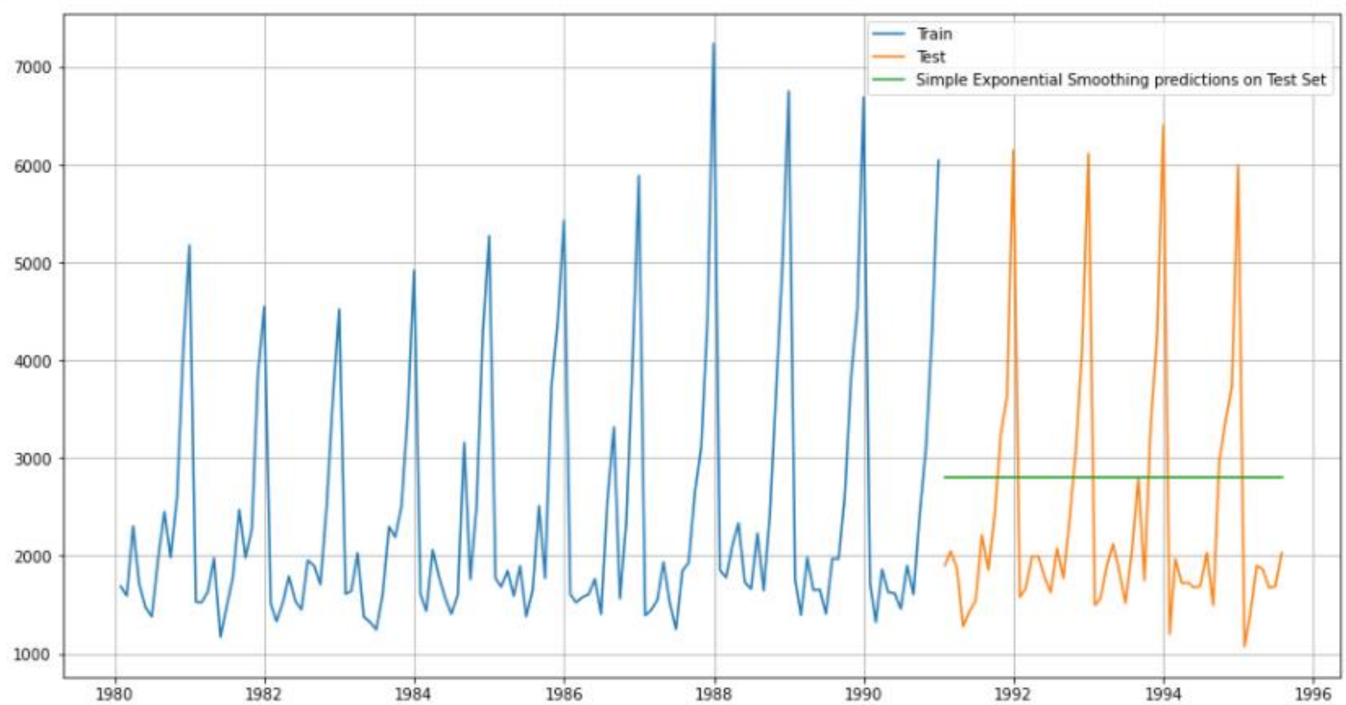
Exponential Smoothing

a.) Simple Exponential Smoothing with additive errors.

1. Autofit Params

```
: {'smoothing_level': 0.07028781460389563,
   'smoothing_trend': nan,
   'smoothing_seasonal': nan,
   'damping_trend': nan,
   'initial_level': 1763.9269926897732,
   'initial_trend': nan,
   'initial_seasons': array([], dtype=float64),
   'use_boxcox': False,
   'lamda': None,
   'remove_bias': False}
```

2. Simple Exponential Smoothing prediction plot on test data



3. RMSE Score of Simple Exponential Smoothing:

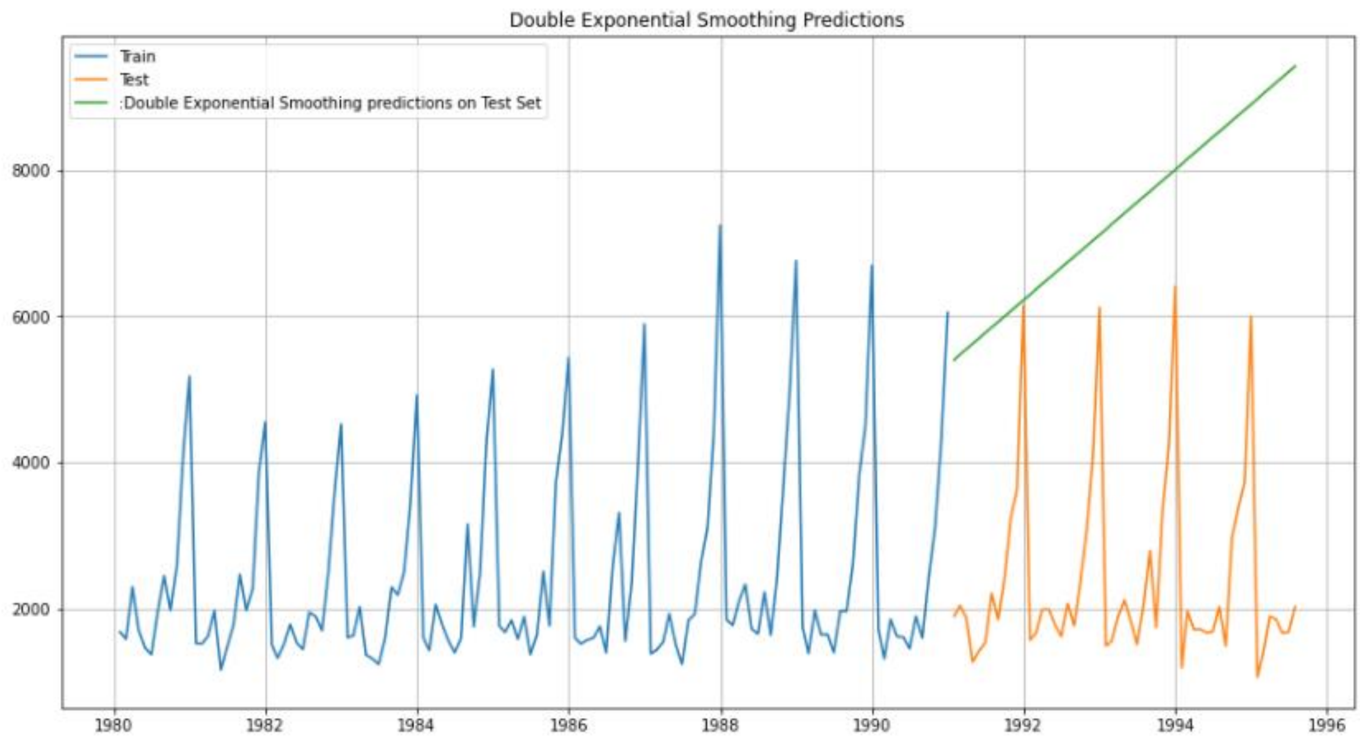
```
SES RMSE: 1338.0046232563645
```

b.) Double Exponential Smoothing - Holt's linear method with additive errors

1. Autofit Params

```
{'smoothing_level': 0.6649999999999999,
 'smoothing_trend': 0.0001,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1502.1999999999998,
 'initial_trend': 74.87272727272733,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Double Exponential Smoothing prediction plot on test data



3. RMSE Score of double Exponential Smoothing:

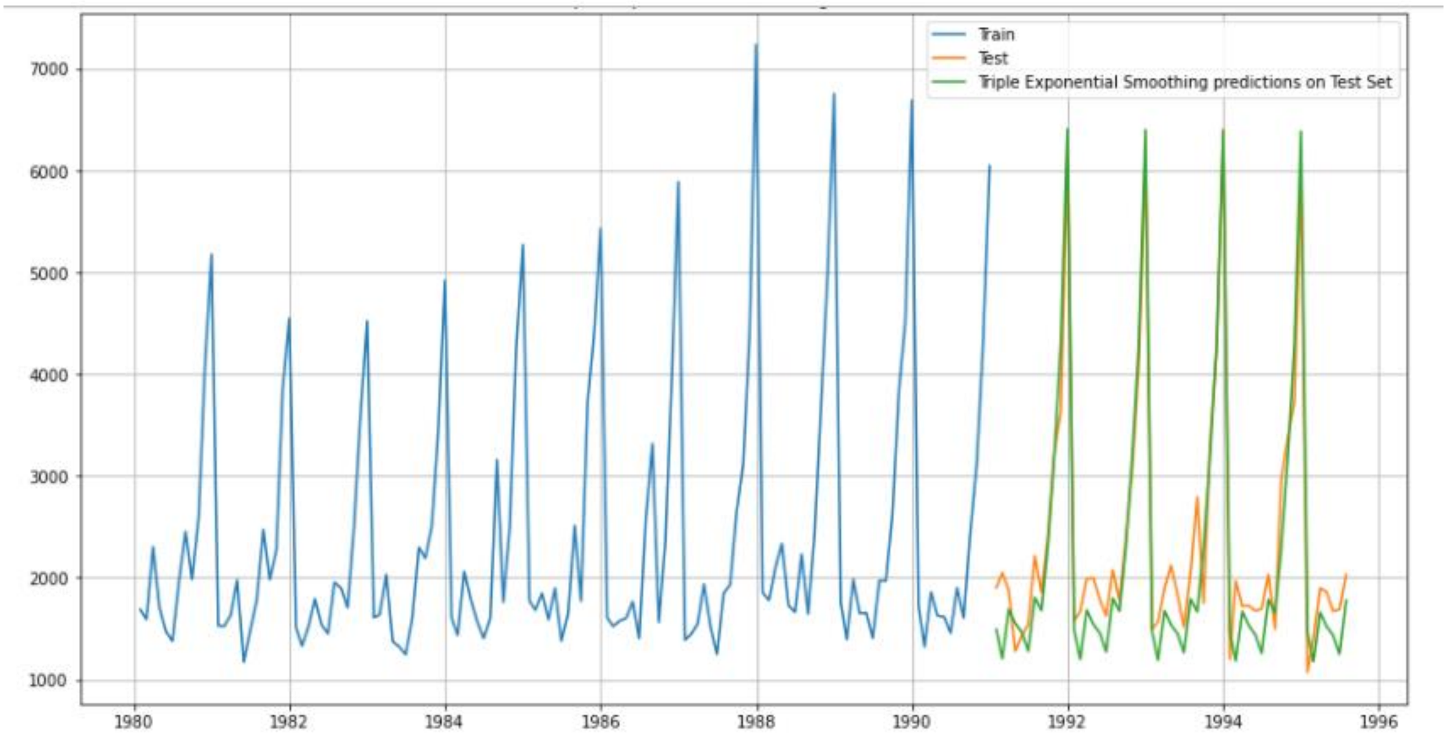
```
DES RMSE: 5291.879833226911
```

c.) Triple Exponential Smoothing (addictive) - Holt Winter's linear method with additive errors

1. Autofit Params

```
{'smoothing_level': 0.11127217859992398,
 'smoothing_trend': 0.01236078328796452,
 'smoothing_seasonal': 0.4607177642170641,
 'damping_trend': nan,
 'initial_level': 2356.578308185137,
 'initial_trend': -0.01853556812789492,
 'initial_seasons': array([-636.23360535, -722.98363367, -398.6436108 , -473.43084469,
        -808.42525514, -815.35024951, -384.23066987,   72.99508063,
        -237.44272911,  272.32584554, 1541.3782103 , 2590.0775386 ]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Triple Exponential Smoothing prediction plot on test data



3. RMSE Score of Triple Exponential Smoothing (Addictive errors):

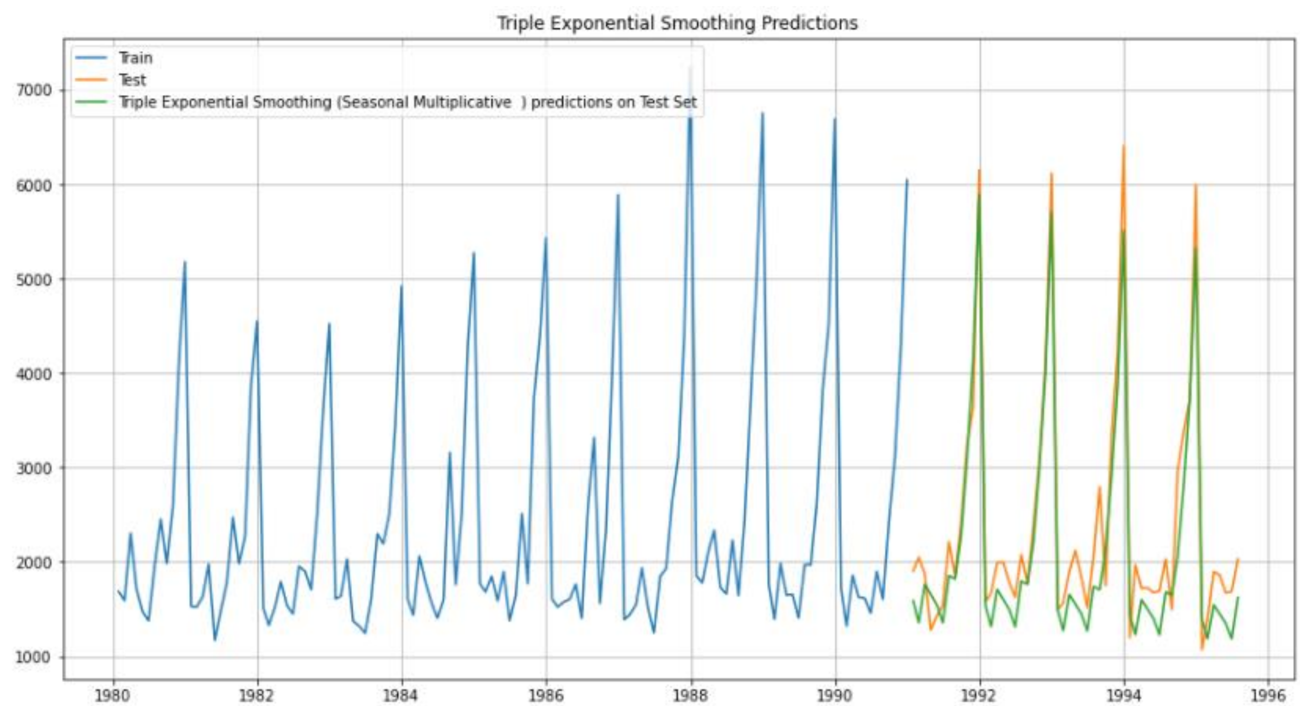
```
TES RMSE: 378.6262408893861
```

d.) Triple Exponential Smoothing (Multiplicative) - Holt Winter's linear method

1. Autofit Params

```
{'smoothing_level': 0.11119949831569428,
 'smoothing_trend': 0.049430920023313805,
 'smoothing_seasonal': 0.3620525701498937,
 'damping_trend': nan,
 'initial_level': 2356.5264391986907,
 'initial_trend': -9.443690175376352,
 'initial_seasons': array([0.71325627, 0.68332509, 0.90537798, 0.80561841, 0.65639659,
        0.65451508, 0.88690241, 1.13423953, 0.91927727, 1.21396745,
        1.86941738, 2.3734461 ]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

2. Triple Exponential Smoothing prediction plot on test data



3. RMSE Score of Triple Exponential Smoothing (Multiplicative):

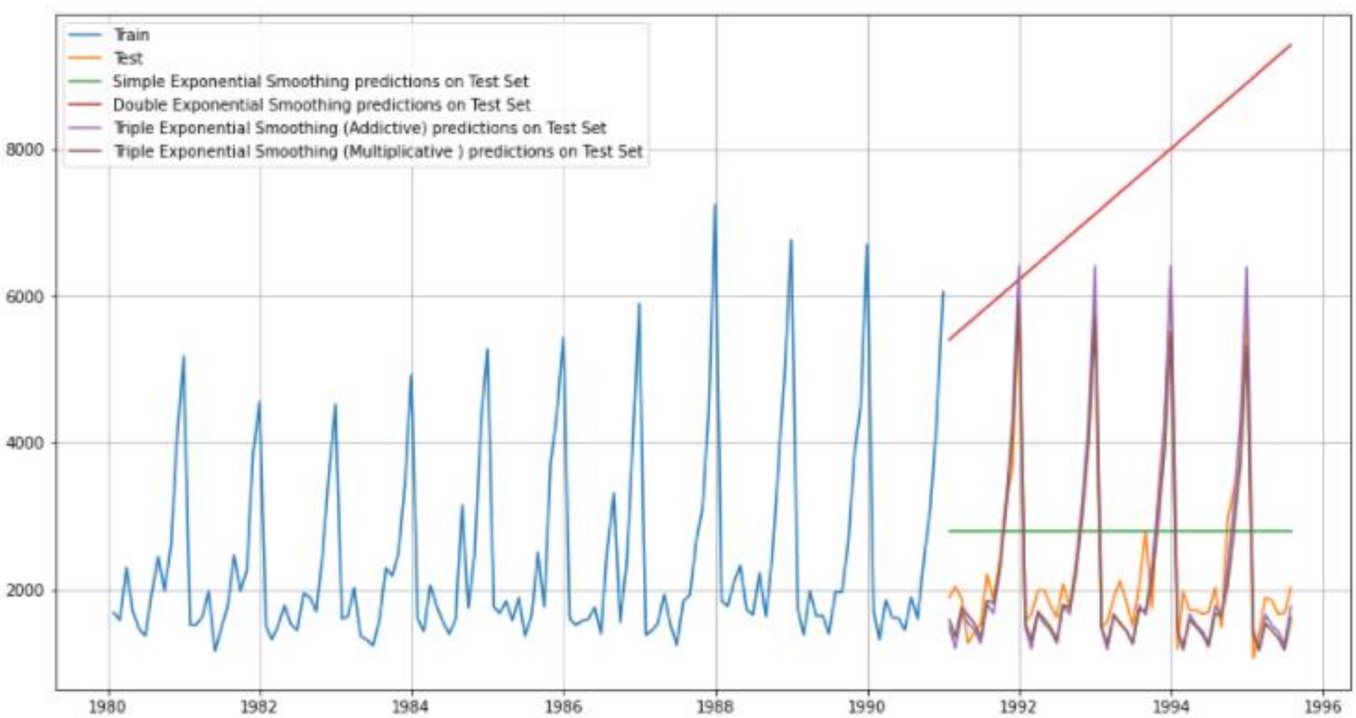
```
TES_am RMSE: 403.7062277856435
```

Inference on Exponential Smoothing:

1. Comparison Table of RMSE OF Different model

	Test RMSE
SES	1338.004623
DES	5291.879833
TES A	378.626241
TES SM	403.706228

2. Comparison of prediction on Multiple model



3. Insights

In exponential smoothing, Holt Winter's linear method with additive errors (Triple Exponential Smoothing (addictive)) is the best model based on least RMSE score.

Regression, Naïve forecast, and simple average models.

a.) Linear Regression.

1. Creating linear instance (according to date)

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 5
8, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110,
111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176,
177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]
```

2. Head of data set after adding linear Instance.

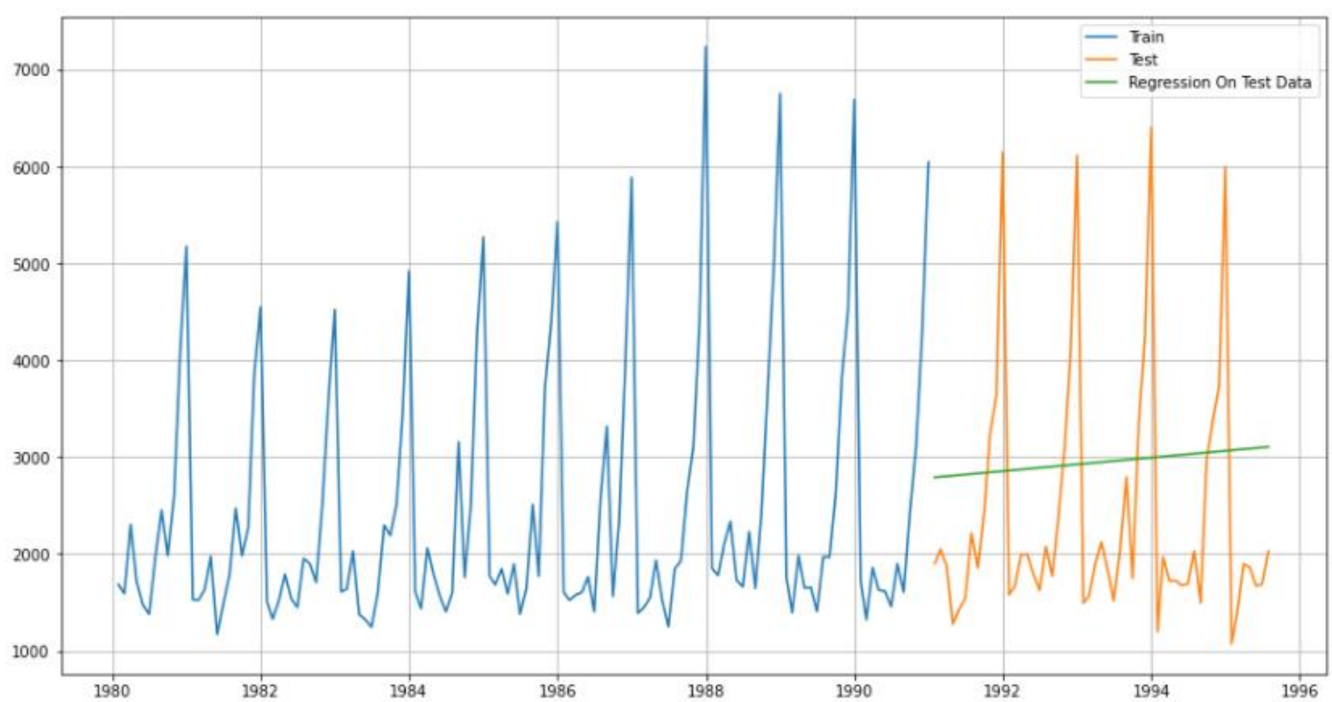
1]:

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

3. Building Linear Regression

```
LinearRegression()
```

4. Linear Regression prediction plot on test data



5. RMSE Score of Linear Regression:

```
For Regression  forecast on the Test Data,  RMSE is 1389.135
```

b.) Naïve Approach

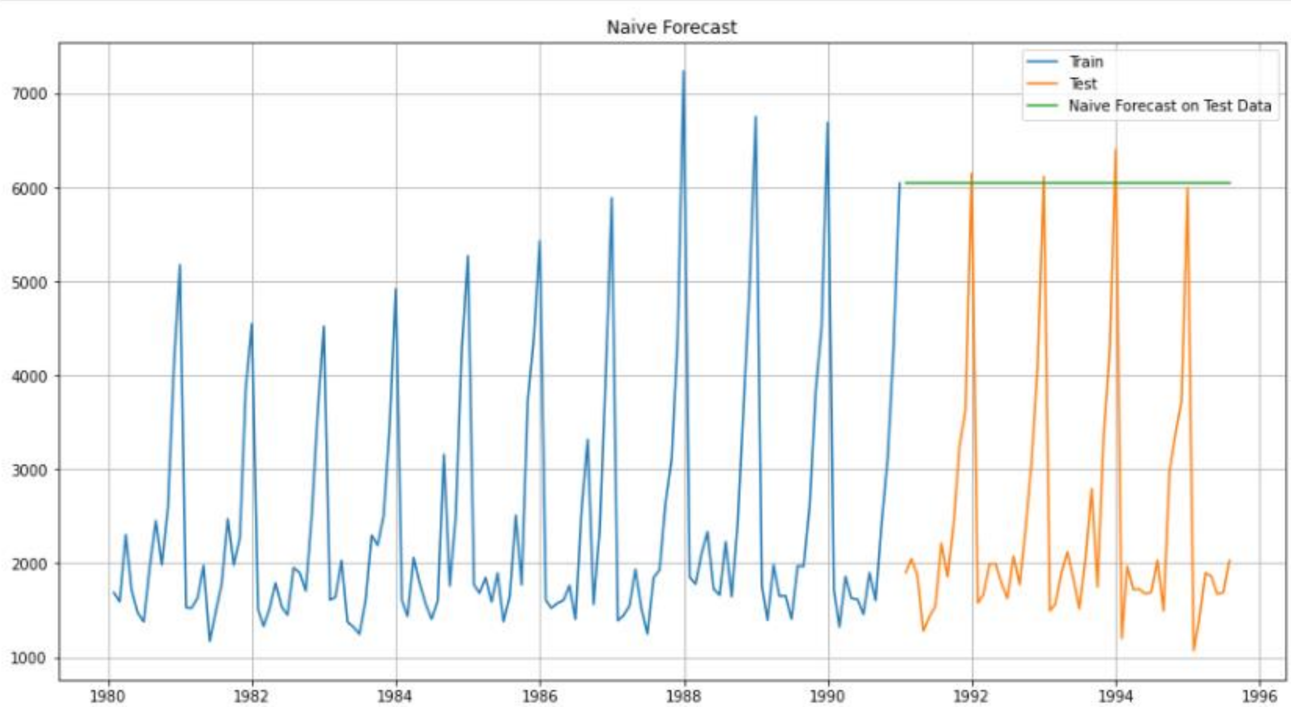
1. Tail of train data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

2. Head of Test data after applying Naïve Approach.

```
Time_Stamp
1991-01-31    6047
1991-02-28    6047
1991-03-31    6047
1991-04-30    6047
1991-05-31    6047
```

3. Naïve Approach prediction plot on test data



4. RMSE Score of Naïve Approach:

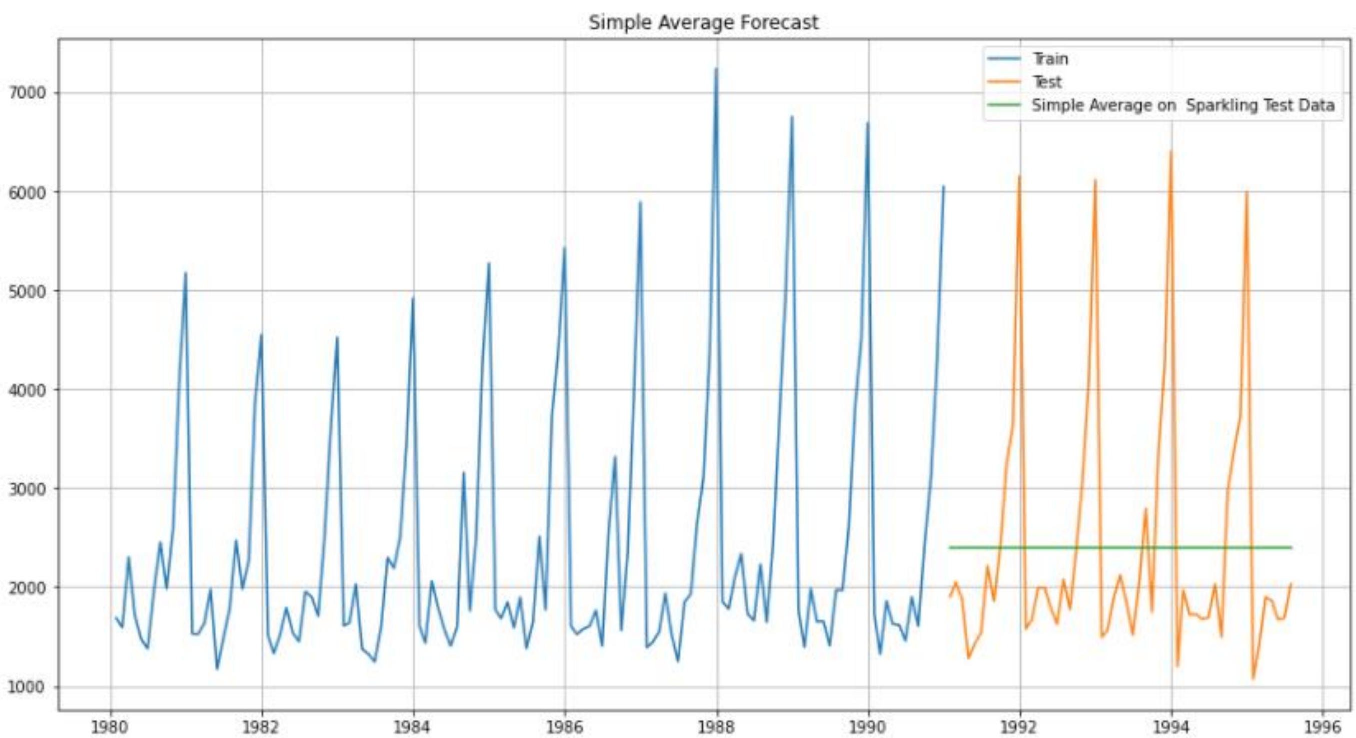
```
For Naive forecast on the sparking Test Data,  RMSE is 3864.279
```

c.) Simple Average

1. Head of data set after creating mean forecast.

Time_Stamp	Sparkling	mean_forecast
1991-01-31	1902	2403.780303
1991-02-28	2049	2403.780303
1991-03-31	1874	2403.780303
1991-04-30	1279	2403.780303
1991-05-31	1432	2403.780303

2. Simple average prediction plot on test data



3. RMSE Score of simple Average:

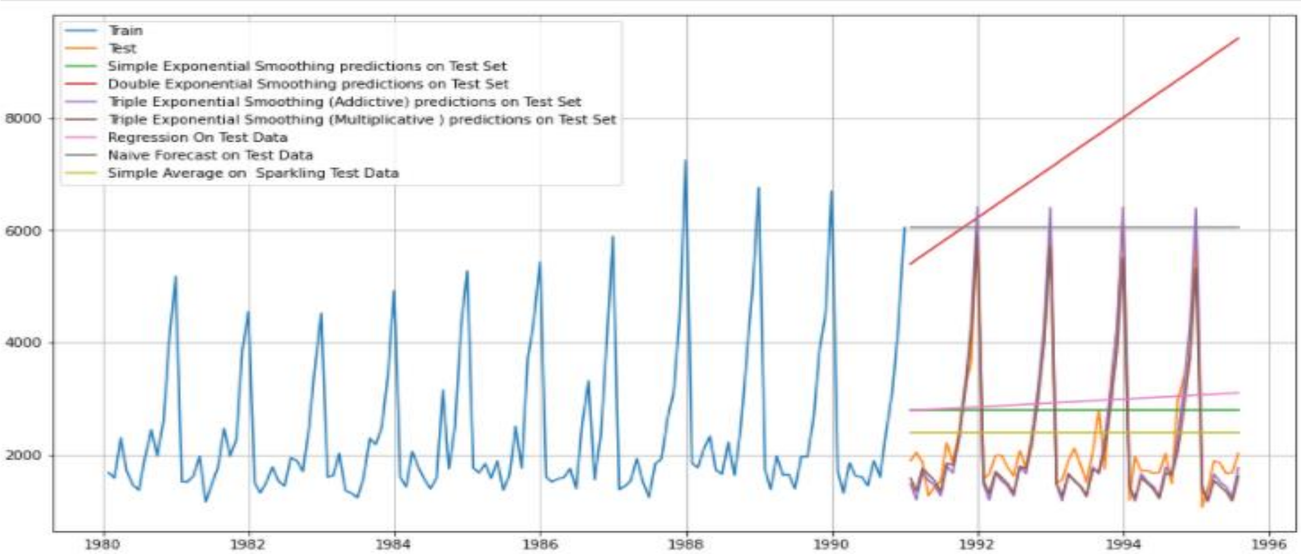
```
For Simple Average forecast on the Test Data,  RMSE is 1275.082
```

Inference on above given models:

1. Comparison Table of RMSE OF Different model

	Test RMSE
SES	1338.004623
DES	5291.879833
TES A	378.626241
TES SM	403.706228
Regression	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

2. Comparison of prediction on Multiple model



3. Insights

In above analyzed models, Holt Winter's linear method with additive errors (Triple Exponential Smoothing (addictive)) is the best model based on least RMSE score. I.e. it will be able to forecast the sales with least errors compared to other analyzed models.

1.5) Checking and changing the training data into stationary data using appropriate statistical tests and methods .Stationarity is checked at alpha = 0.05.

a.) Augmented Dickey–Fuller test Hypothesis for stationary data

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

- H_0 : The Time Series has a unit root and is thus non-stationary.
- H_1 : 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 alpha value.

b.) ADF test on train data

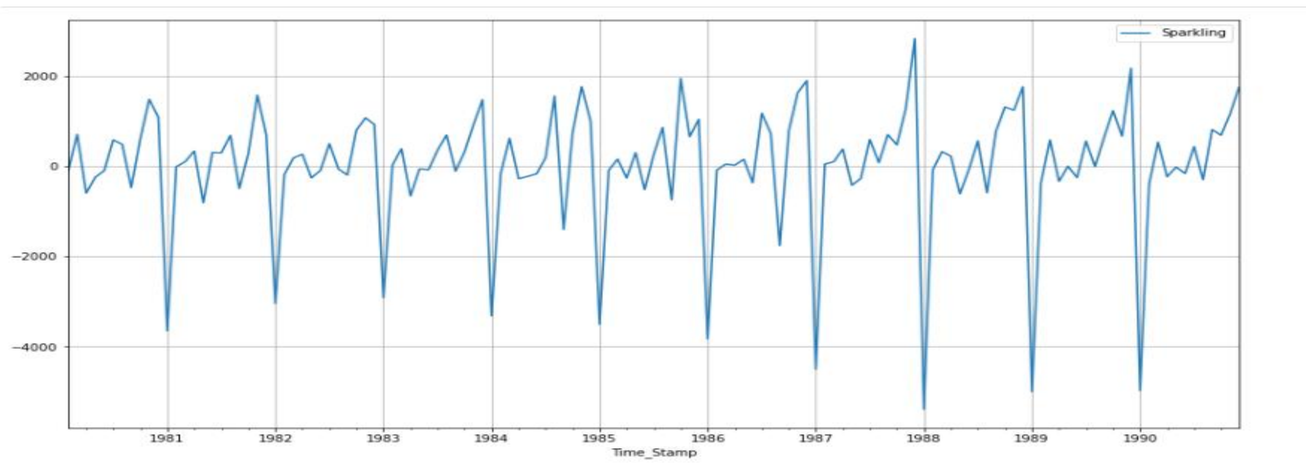
```
DF test statistic is -2.062
DF test p-value is 0.5674110388593719
Number of lags used 12
fail to reject null hypoathes . its not stationary
```

The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationarize the Time Series.

c.) ADF test on train data after one differencing

```
DF test statistic is -7.968
DF test p-value is 8.479210655514579e-11
Number of lags used 11
reject Null hypothesis  ie its statioary
```

d.) Plotting of Data set after one differencing.



Comment / Inference:

Actual Train data is not stationary. After one differencing data has become stationary

Testing for stationarity is very important because the whole results of the regression might be fabricated thus predcation may not be proper if data is nonstationary.

1.6) Building 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 evaluating the performance of this model on the test data using RMSE.

The data has some seasonality so ideally, we should build a SARIMA model. But for confirmation I am building an ARIMA model by looking at the minimum AIC criterion.

a.) ARIMA model

1.Creating different parameters for the model.

```
Examples of the parameter combinations for the Model
Model: (0, 1, 0)
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)
```

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary.

2.Building Arima model with different parameters

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.0600155922552
ARIMA(0, 1, 2) - AIC:2234.4083231242757
ARIMA(0, 1, 3) - AIC:2233.994857747629
ARIMA(1, 1, 0) - AIC:2266.608539319009
ARIMA(1, 1, 1) - AIC:2235.755094669173
D:\anocondal\lib\site-packages\statsmodel
  warn('Non-invertible starting MA parame
ARIMA(1, 1, 2) - AIC:2234.5272004516546
ARIMA(1, 1, 3) - AIC:2235.607808732252
ARIMA(2, 1, 0) - AIC:2260.3657439680865
ARIMA(2, 1, 1) - AIC:2233.777626209401
ARIMA(2, 1, 2) - AIC:2213.509212785536
D:\anocondal\lib\site-packages\statsmodel
  warnings.warn("Maximum Likelihood optim
ARIMA(2, 1, 3) - AIC:2232.8110262733007
ARIMA(3, 1, 0) - AIC:2257.723378997941
ARIMA(3, 1, 1) - AIC:2235.498940717457
D:\anocondal\lib\site-packages\statsmodel
  warnings.warn("Maximum Likelihood optim
ARIMA(3, 1, 2) - AIC:2230.754792087503
ARIMA(3, 1, 3) - AIC:2221.4554497355275
```

2.Head of the Arima models with AIC score in ascending order

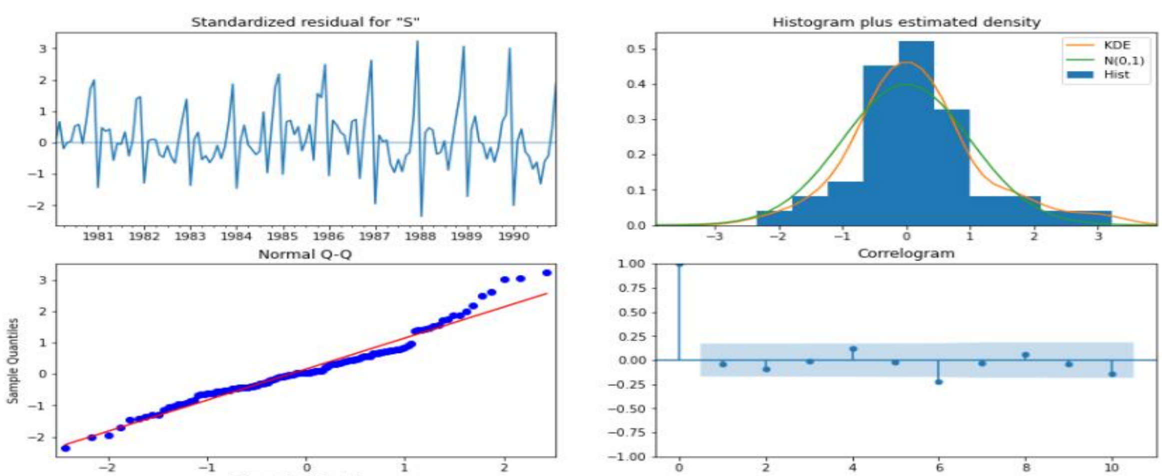
	param	AIC
10	(2, 1, 2)	2213.509213
15	(3, 1, 3)	2221.455450
14	(3, 1, 2)	2230.754792
11	(2, 1, 3)	2232.811026
9	(2, 1, 1)	2233.777626

Model with parameter (2,1,2) has the least AIC score. So, it is the best parameter

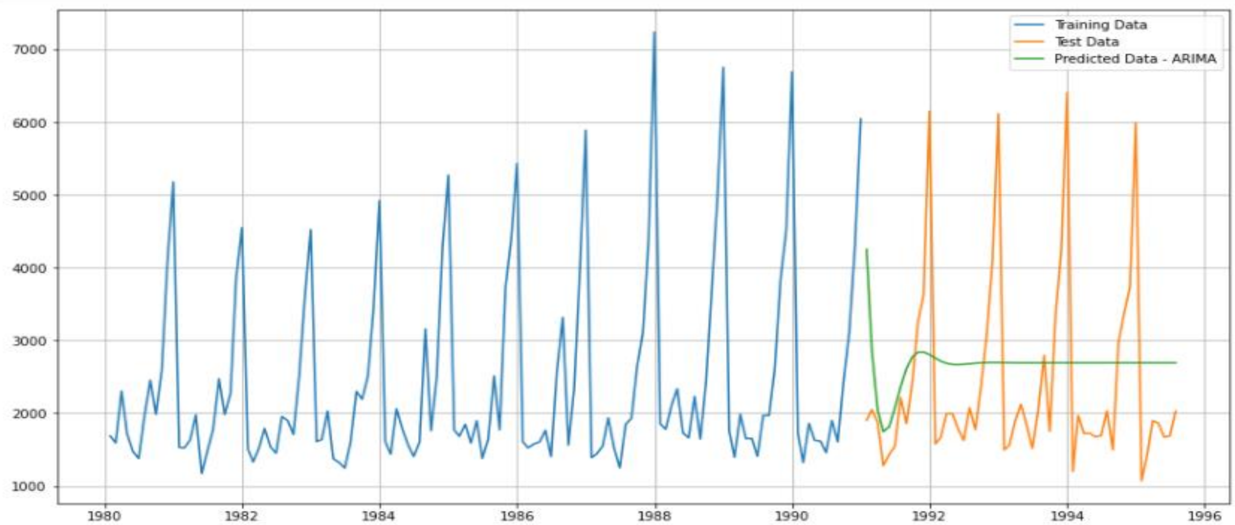
3.Summary of the model after fitting it with the best parameters

```
  warnings.warn('No frequency information was'
                                   SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sat, 22 May 2021   AIC                           2213.509
Time:                        22:35:21   BIC                           2227.885
Sample:                    01-31-1980   HQIC                          2219.351
                         - 12-31-1990
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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.41e+28. Standard errors may be unstable.
```

4.Diagnostics Plot



5.ARIMA predication Plot on test data



6.RMSE score on test data for Arima using lowest Akaike Information Criteria

```
RMSE: 1299.9808693008124
```

b.) SARIMA model

1.Creating different parameters for the model.

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary.

As we are not applying any additional differentiate for seasonal data D value stays as zero

2.Building Sarima model with different parameters

In the below images we have shown only the few parameter combinations

```
SARIMA(0, 1, 0)x(0, 0, 0, 6) - AIC:2251.3597196862966
SARIMA(0, 1, 0)x(0, 0, 1, 6) - AIC:2152.378076171629
SARIMA(0, 1, 0)x(0, 0, 2, 6) - AIC:1955.6355536889994
SARIMA(0, 1, 0)x(0, 0, 3, 6) - AIC:1863.7845154972952
SARIMA(0, 1, 0)x(1, 0, 0, 6) - AIC:2164.4097581959904
SARIMA(0, 1, 0)x(1, 0, 1, 6) - AIC:2079.5599844431026
SARIMA(0, 1, 0)x(1, 0, 2, 6) - AIC:1926.9360121327015
SARIMA(0, 1, 0)x(1, 0, 3, 6) - AIC:1803.3929094952937
SARIMA(0, 1, 0)x(2, 0, 0, 6) - AIC:1839.4012986872267
SARIMA(0, 1, 0)x(2, 0, 1, 6) - AIC:1841.1993617510623
D:\anocondal\lib\site-packages\statsmodels\base\model.p
  warnings.warn("Maximum Likelihood optimization failed
SARIMA(0, 1, 0)x(2, 0, 2, 6) - AIC:1810.9177805661222
SARIMA(0, 1, 0)x(2, 0, 3, 6) - AIC:1725.5376425549302
SARIMA(0, 1, 0)x(3, 0, 0, 6) - AIC:1748.762266815527
SARIMA(0, 1, 0)x(3, 0, 1, 6) - AIC:1750.6879953816626
SARIMA(0, 1, 0)x(3, 0, 2, 6) - AIC:1739.4489858030327
SARIMA(0, 1, 0)x(3, 0, 3, 6) - AIC:1725.0138750179908
```

2.Head of the Sarima models with AIC score in ascending order

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	1629.052955
59	(0, 1, 3)	(2, 0, 3, 6)	1633.327871
191	(2, 1, 3)	(3, 0, 3, 6)	1634.400488
251	(3, 1, 3)	(2, 0, 3, 6)	1634.617364
63	(0, 1, 3)	(3, 0, 3, 6)	1635.058644

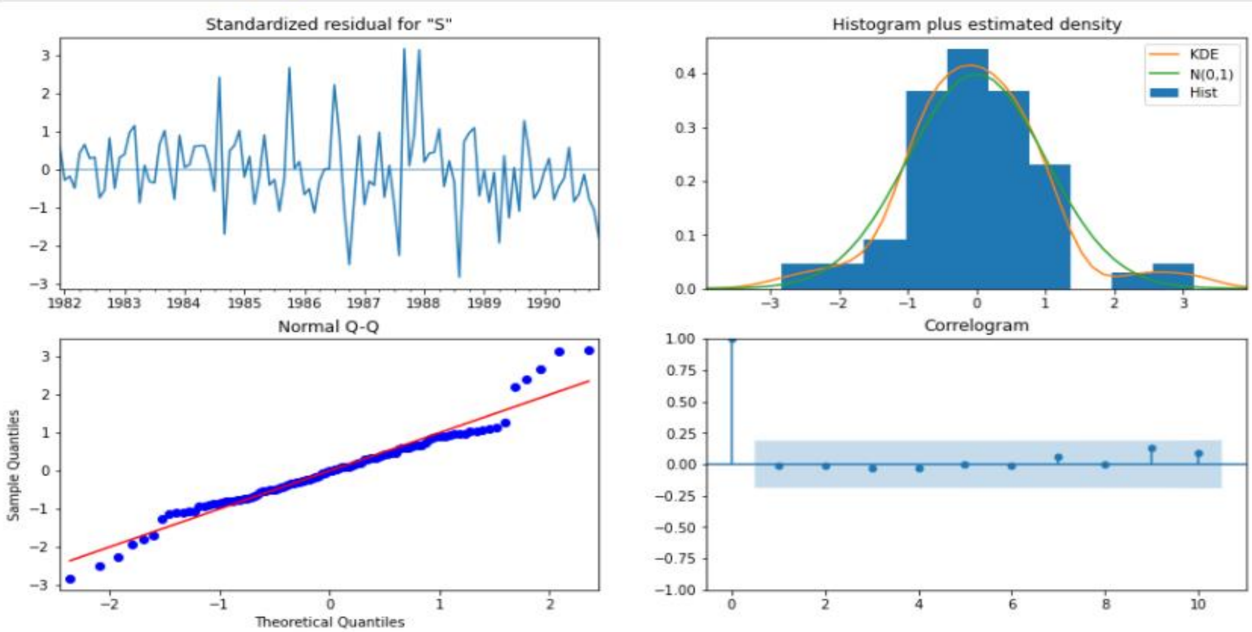
Model with parameter (2,1,3) (2,0,3,6) has the least AIC score. So, it is the best parameter

3.Summary of the model after fitting it with the best parameters

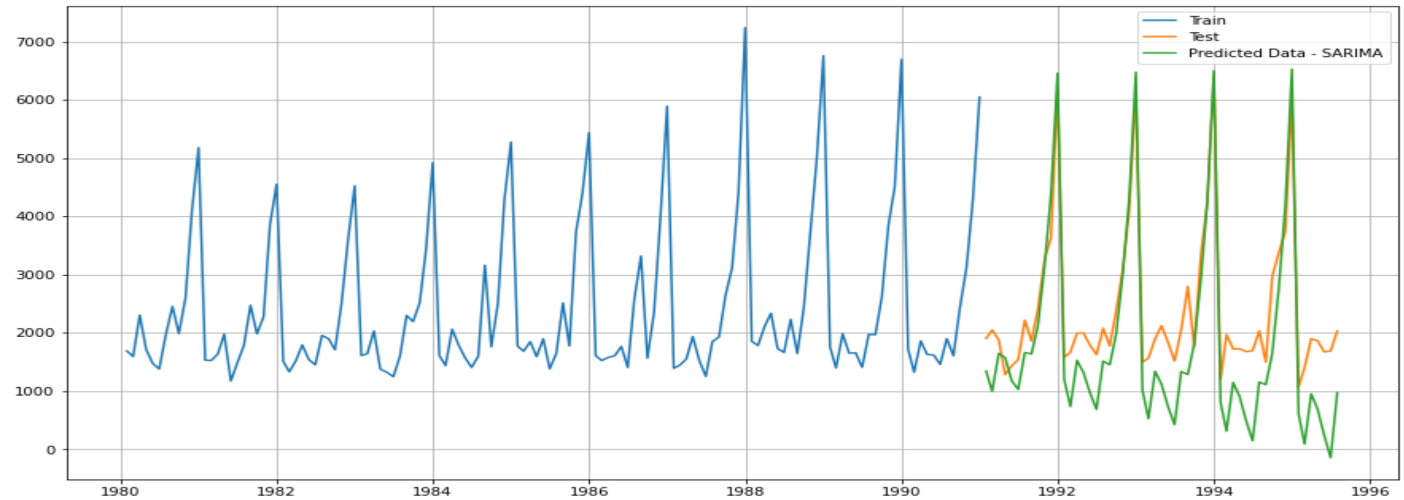
```
  warnings.warn('No frequency information was
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                             Sparkling   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -803.526
Date:                           Sat, 22 May 2021   AIC                           1629.053
Time:                                   22:39:07   BIC                           1658.658
Sample:                               01-31-1980   HQIC                          1641.059
                                    - 12-31-1990
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.7440      0.090    -19.397      0.000      -1.920      -1.568
ar.L2         -0.7864      0.087     -9.041      0.000      -0.957      -0.616
ma.L1          1.0816      1.406      0.769      0.442      -1.675       3.838
ma.L2         -0.7534      0.209     -3.608      0.000      -1.163      -0.344
ma.L3         -0.8875      1.268     -0.700      0.484      -3.374       1.599
ar.S.L6       -0.0112      0.030     -0.372      0.710      -0.070       0.048
ar.S.L12       1.0386      0.023     45.564      0.000       0.994       1.083
ma.S.L6        0.3732      0.264      1.412      0.158      -0.145       0.891
ma.S.L12      -0.7620      0.205     -3.710      0.000      -1.165      -0.359
ma.S.L18       0.1092      0.165      0.663      0.507      -0.214       0.432
sigma2      1.031e+05   1.51e+05      0.682      0.496   -1.93e+05       4e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                14.91
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               1.50   Skew:                             0.38
Prob(H) (two-sided):                  0.23   Kurtosis:                         4.65
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.12e+14. Standard errors may be unstable.
```

```
results_auto_SARIMA.plot_diagnostics();
```

4.Diagnostics Plot



5.SARIMA predication Plot on test data



6.RMSE score of Sarima using lowest Akaike Information Criteria

```
RMSE: 820.6238123232409
```

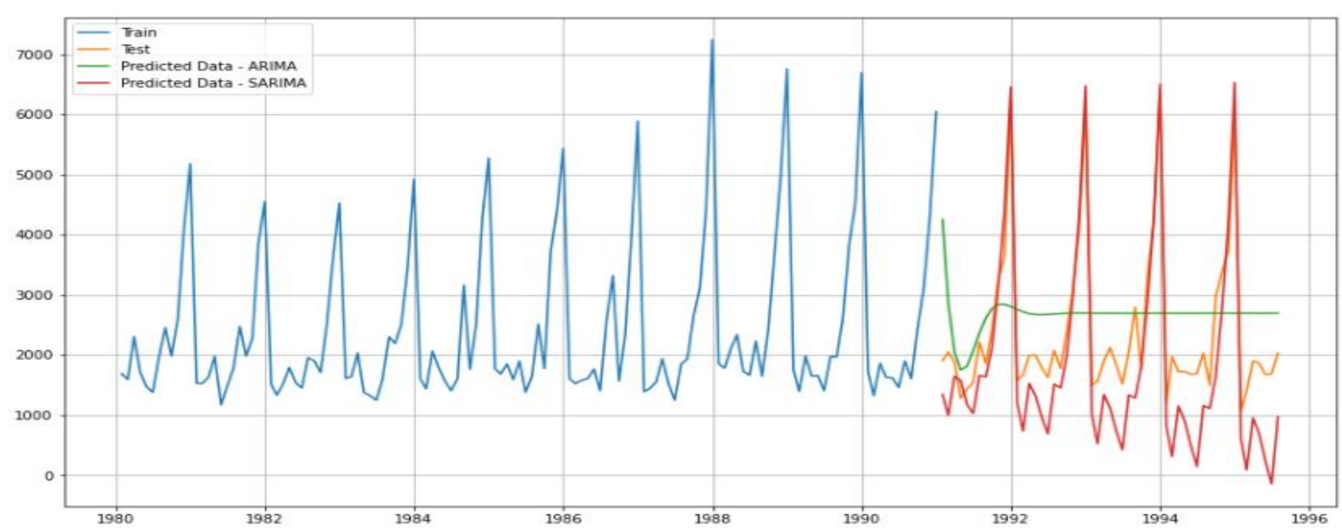
Inference:

1. Comparison Table of RMSE OF Different model

7]:

	RMSE
ARIMA(2,1,2)	1299.980869
SARIMA(2,1,3)(2,0,3,6)	820.623812

2. Comparison Table of RMSE OF Different model



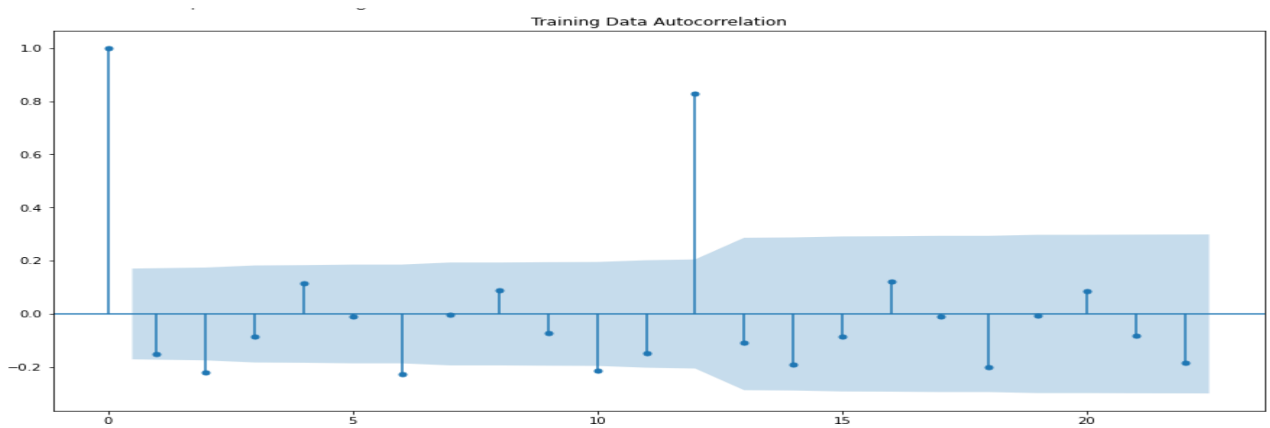
3. Insights:

RMSE has reduced in comparison to ARIMA when seasonality was introduced.

1.7) Building ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluating this model on the test data using RMSE.

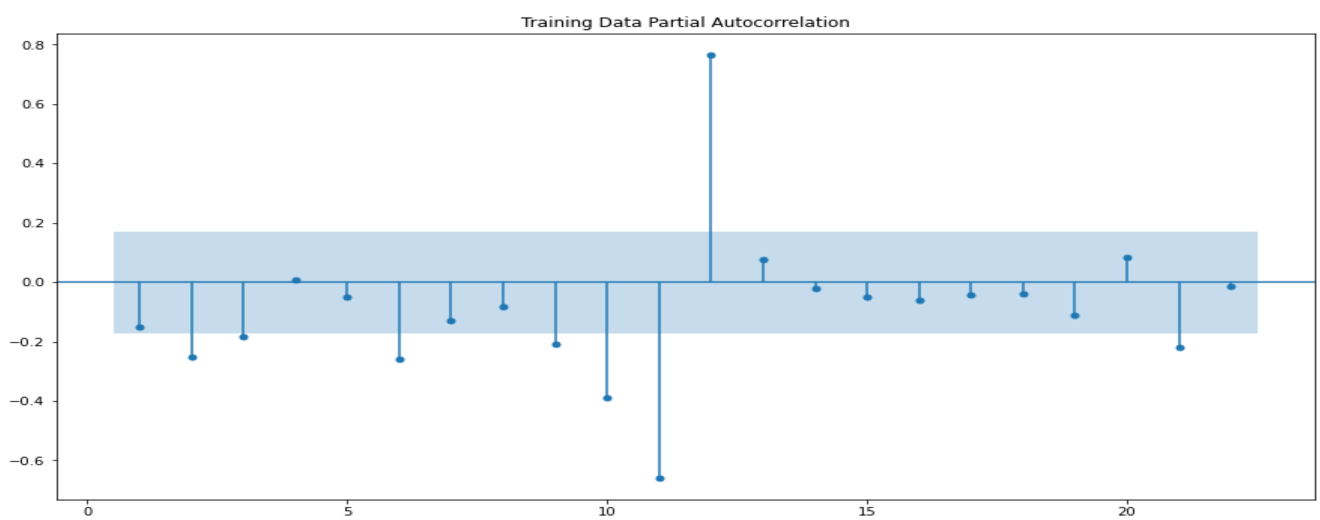
a.) ARIMA model method using cut-off points of ACF and PACF.

1.ACF plot of train data.



Based on ACF plot q value is 2. (i.e. 2nd lag is out of confidence level and next lag drops below confidence level)

2.PACF plot of train data



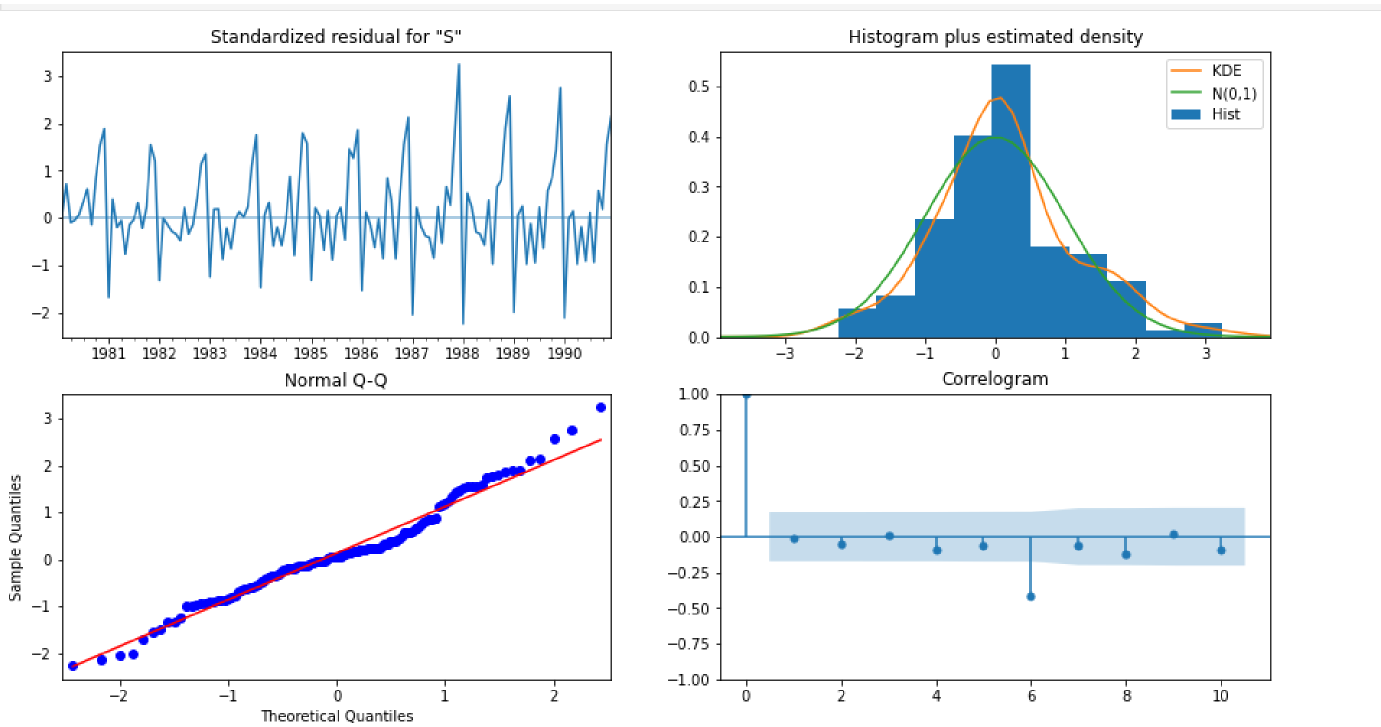
Based on PACF plot p value is 3. (i.e., 2nd and 3rd lag is out of confidence level and next lag drops below confidence level)

I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary. So, the parameter is (3,1,2)

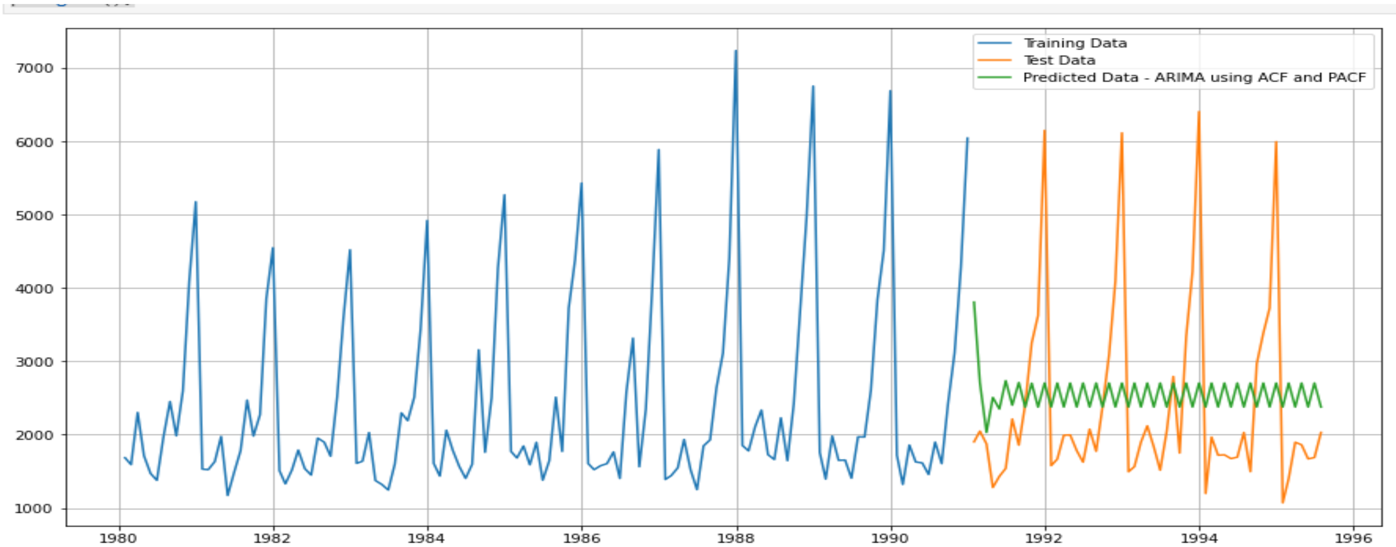
3.Summary of the model after fitting it with the best parameters (3,1,2):

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 1, 2)   Log Likelihood               -1109.377
Date:                Sat, 22 May 2021   AIC                           2230.755
Time:                        22:39:08   BIC                           2248.006
Sample:                    01-31-1980   HQIC                          2237.765
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4328      0.040    -10.743      0.000      -0.512      -0.354
ar.L2          0.3244      0.112      2.903      0.004       0.105       0.543
ar.L3         -0.2428      0.072     -3.395      0.001      -0.383      -0.103
ma.L1          0.0183      0.127      0.143      0.886      -0.232       0.268
ma.L2         -0.9815      0.136     -7.229      0.000      -1.248      -0.715
sigma2      1.274e+06   1.94e-07   6.57e+12      0.000    1.27e+06    1.27e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 4.70
Prob(Q):                              0.89   Prob(JB):                         0.10
Heteroskedasticity (H):               2.72   Skew:                             0.38
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.54
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 8.11e+28. Standard errors may be unstable.
```

4.Diagnostics Plot



5.ARIMA predication Plot on test data

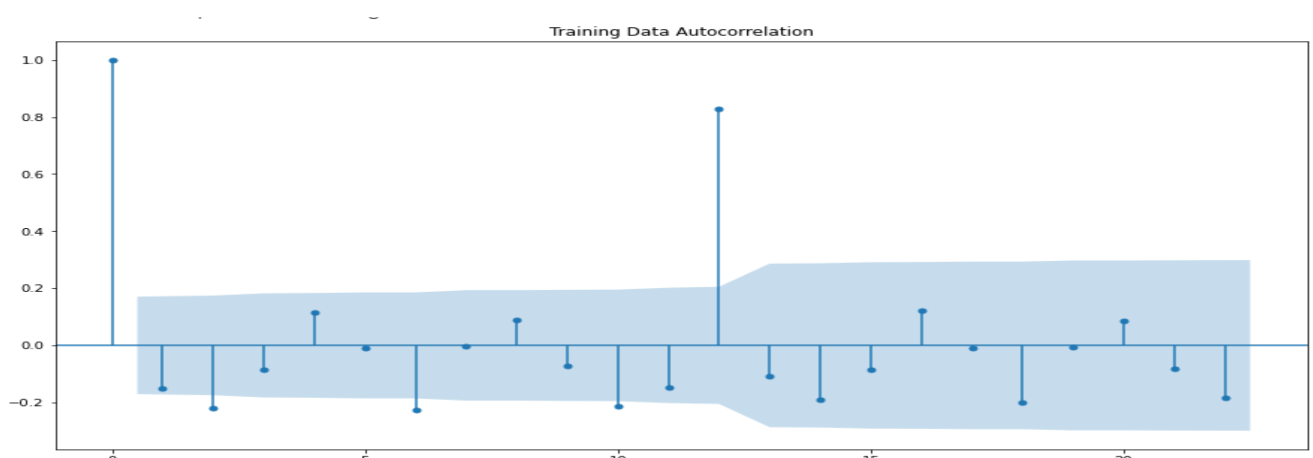


6.RMSE score on test data for Arima using ACF AND PACF (Manual)

```
RMSE: 1281.7511774185691
```

b.) SARIMA model cut-off points of ACF and PACF

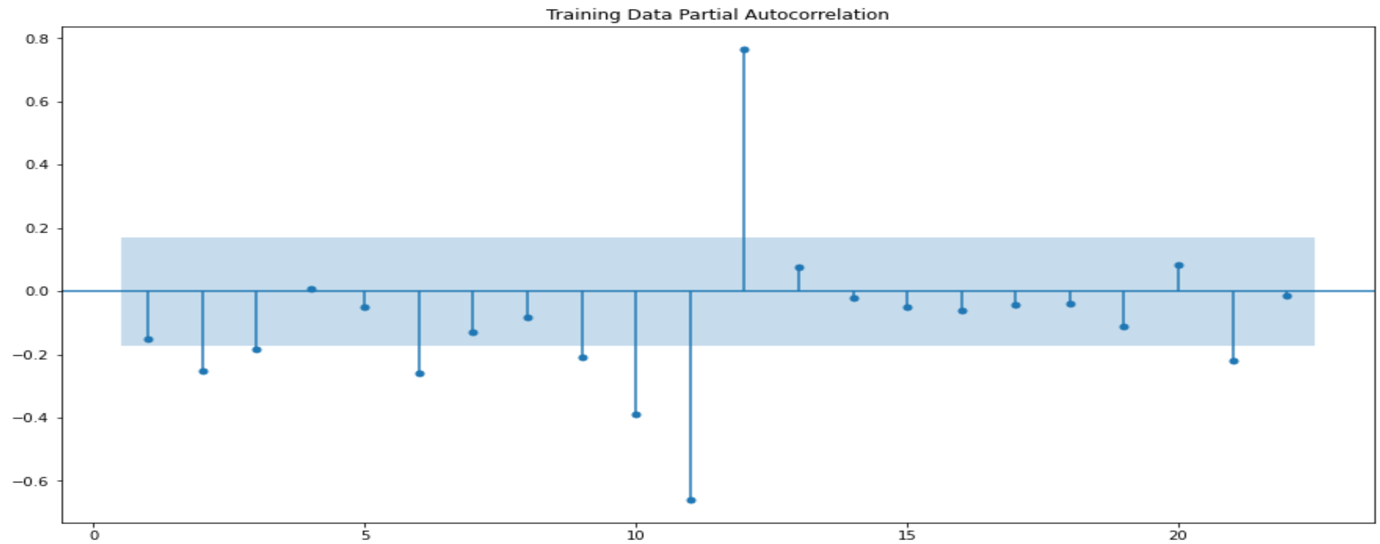
1.ACF plot of train data.



Based on ACF plot q value is 2. (ie 2nd lag is out of confidence level and next lag drops below confidence level) Q value is zero because there is no signification lag trend.

Example: if I take 2nd lag as significance, there 4th lag is not significant, and we have chosen seasonality to be 6 so we cannot have Q as 6 also. So, Q value is 0.

2.PACF plot of train data



Based on PACF plot p value is 3. (i.e. 2nd and 3rd lag is out of confidence level and next lag drops below confidence level) and P value is 3 cos there is a seasonality trend for every 3rd significant lags there is a cut off.

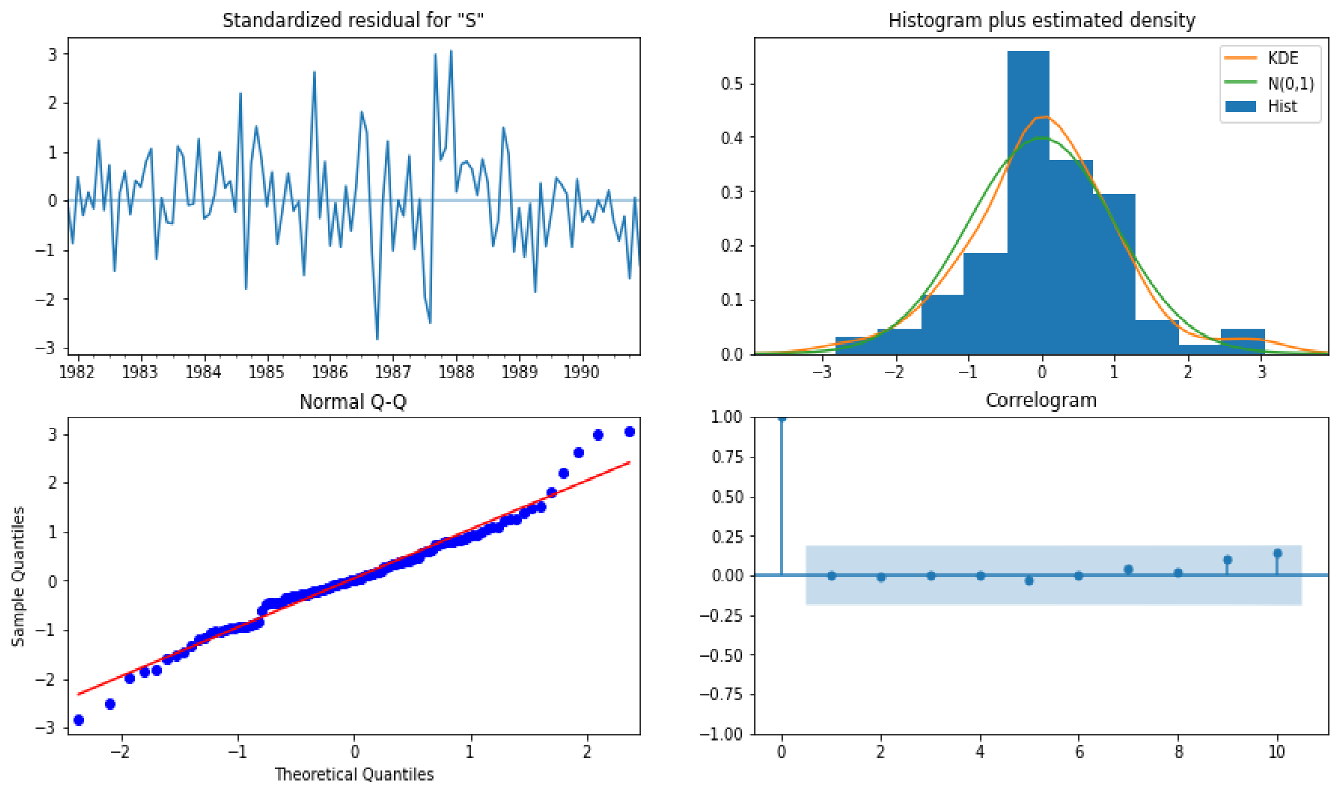
I have kept the d value as 1 as we have analyzed that we must differentiate it by 1 to make the data stationary. We do not differentiate again D is 0.

So, the parameter is (3,1,2) (3, 0, 0, 6)

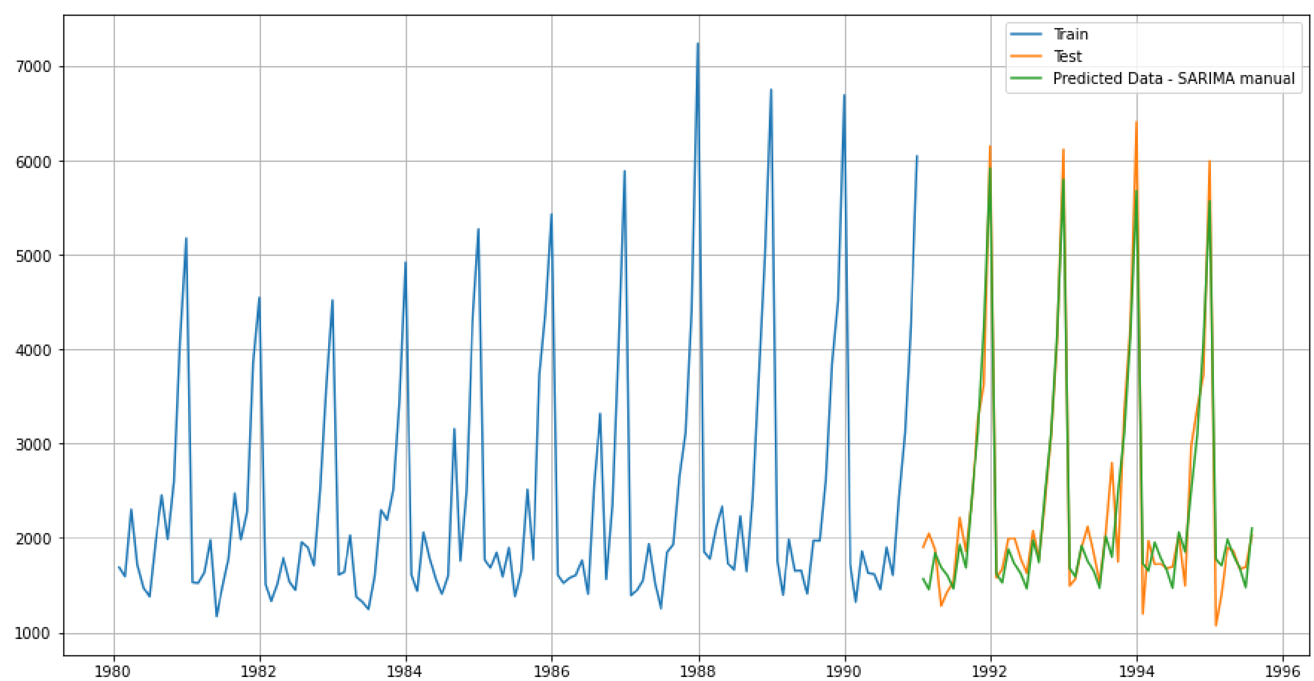
3.Summary of the model after fitting it with the best parameters (3,1,2) (3, 0, 0, 6)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                132
Model:             SARIMAX(3, 1, 2)x(3, 0, [], 6)   Log Likelihood                -822.494
Date:                              Sun, 23 May 2021   AIC                           1662.989
Time:                                      02:35:32   BIC                           1687.293
Sample:                                  01-31-1980   HQIC                          1672.847
                                       - 12-31-1990
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5675      0.189     -2.999      0.003      -0.938      -0.197
ar.L2          0.0952      0.108      0.883      0.377      -0.116       0.307
ar.L3         -0.0222      0.098     -0.226      0.821      -0.215       0.170
ma.L1         -0.1225      0.206     -0.595      0.552      -0.526       0.281
ma.L2         -0.8775      0.198     -4.439      0.000      -1.265      -0.490
ar.S.L6        0.0110      0.132      0.084      0.933      -0.247       0.269
ar.S.L12       0.9572      0.038     25.105      0.000       0.882       1.032
ar.S.L18      -0.0256      0.138     -0.185      0.853      -0.296       0.245
sigma2       1.76e+05   1.27e-06   1.39e+11      0.000    1.76e+05    1.76e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 5.86
Prob(Q):                              0.98   Prob(JB):                         0.05
Heteroskedasticity (H):               1.18   Skew:                             0.16
Prob(H) (two-sided):                  0.63   Kurtosis:                         4.08
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.94e+27. Standard errors may be unstable.
```

4.Diagnostics Plot



5.SARIMA predication Plot on test data (manual)



6.RMSE score of Sarima using ACF AND PACF cutoff (Manual)

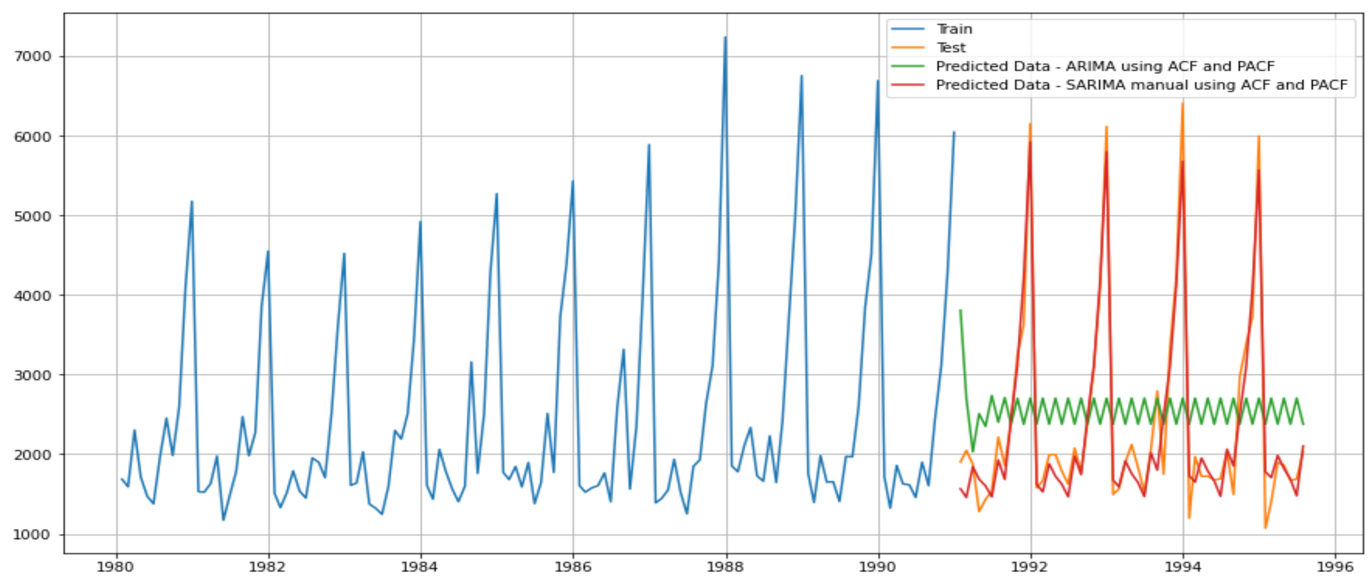
```
RMSE: 321.8681761919399
```

Inference:

1. Comparison Table of RMSE OF Different model

	RMSE
ARIMA M(3,1,2)	1281.751177
SARIMA M(3,1,2)(3,0,0,6)	321.868176

2. Comparison graph of Different model



3. Insights

RMSE has reduced in comparison to ARIMA when seasonality was introduced.

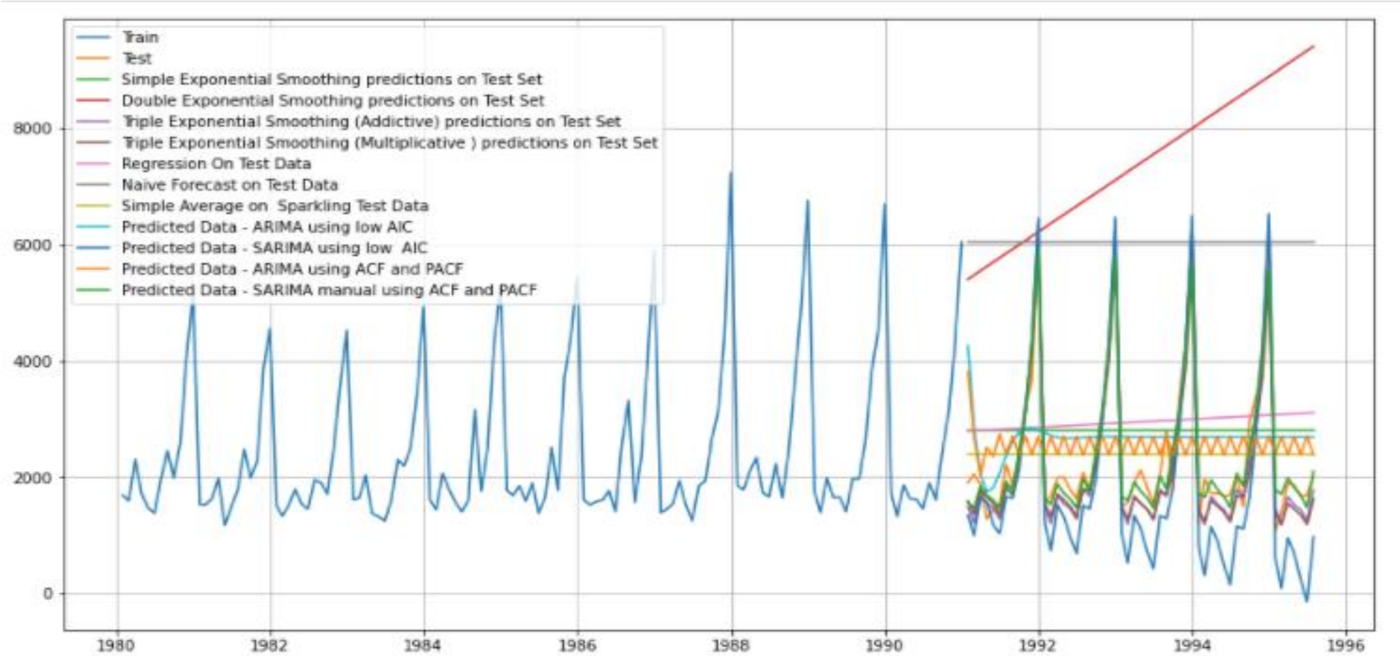
SARIMA with using ACF AND PACF cutoff has the least RMSE score. So, it can forecast with least error.

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

Model Performance comparison table

	Test RMSE	MAPE
SES (Alpha=0.0702)	1338.004623	53.879778
DES (Alpha=0.6649,Beta=0.0001)	5291.879833	268.912388
TES (Alpha=0.1112,Beta=0.01236,Gamma=0.4607):Addictive	378.626241	53.618933
TES (Alpha=0.1111,Beta=0.0494,Gamma=0.3620):Multiplicative	403.706228	48.365465
Regression	1389.135175	59.410392
NaiveModel	3864.279352	201.327650
SimpleAverage	1275.081804	39.157336
Arima (2,1,2) : Low AIC	1299.980869	47.100060
Sarima (2, 1, 3)(2, 0, 3,6), :Low AIC	820.623812	36.126725
Arima (3,1,2) : cut-off points of ACF and PACF	1281.751177	44.067516
Sarima (3, 1, 2)(3, 0, 0,6), :cut-off points of ACF and PACF	321.868176	11.391015

Forecasting comparison plot.



Inference:

Based on the above table, we can see that SARIMA(3,1,2)(3,0,0,6) with using ACF AND PACF cutoff has the least RMSE score. It MAPE of around 11.3 percent. So, it will be able to forecast the future with least error.

1.9) Based on the model-building exercise, building the most optimum model on the complete data and predicting 12 months into the future with appropriate confidence intervals/bands.

Best model is SARIMA with using ACF AND PACF cutoff with parameter of (3,1,2) (3,0,0,6)

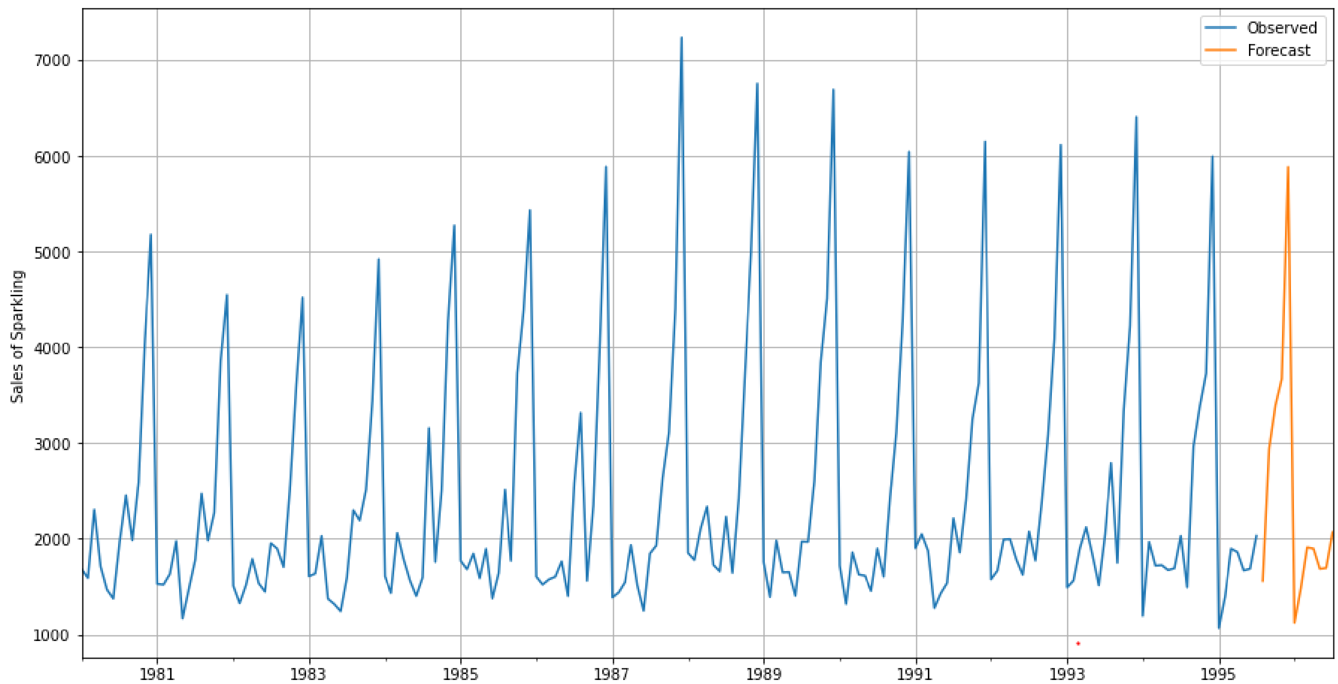
Summary of optimized model on complete data

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                  187
Model:             SARIMAX(3, 1, 2)x(3, 0, [], 6)   Log Likelihood               -1232.468
Date:                              Sat, 22 May 2021   AIC                           2482.937
Time:                                      22:39:13   BIC                           2510.890
Sample:                                  01-31-1980   HQIC                          2494.284
                                       - 07-31-1995
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7886      0.111     -7.110      0.000      -1.006      -0.571
ar.L2          0.0090      0.085      0.106      0.916      -0.157       0.175
ar.L3         -0.0476      0.078     -0.614      0.539      -0.200       0.104
ma.L1         -0.0694      0.141     -0.492      0.623      -0.346       0.207
ma.L2         -0.9306      0.128     -7.271      0.000      -1.181      -0.680
ar.S.L6        0.0441      0.100      0.439      0.661      -0.153       0.241
ar.S.L12       0.9588      0.030     31.689      0.000       0.900       1.018
ar.S.L18      -0.0543      0.104     -0.522      0.601      -0.258       0.149
sigma2      1.747e+05   1.05e-06   1.66e+11      0.000    1.75e+05    1.75e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                16.21
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               1.32   Skew:                             0.24
Prob(H) (two-sided):                  0.30   Kurtosis:                         4.46
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.6e+26. Standard errors may be unstable.
```

Forecasted value

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1561.095664	419.265586	739.350216	2382.841112
1995-09-30	2941.474690	423.818058	2110.806561	3772.142820
1995-10-31	3395.394296	425.725937	2560.986793	4229.801799
1995-11-30	3671.463183	426.065904	2836.389357	4506.537010
1995-12-31	5889.376449	426.229438	5053.982102	6724.770796
1996-01-31	1124.170746	426.537103	288.173385	1960.168107
1996-02-29	1481.877984	426.632242	645.694155	2318.061812
1996-03-31	1910.606815	426.901454	1073.895340	2747.318291
1996-04-30	1897.471966	426.964542	1060.636840	2734.307091
1996-05-31	1686.959291	427.094318	849.869809	2524.048773
1996-06-30	1695.860447	427.107912	858.744321	2532.976573
1996-07-31	2067.591788	427.192424	1230.310023	2904.873553

Plotting the forecast (mean value) of the whole data.



Forecasted value description.

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
count	12.000000	12.000000	12.000000	12.000000
mean	2443.611943	425.794576	1609.069908	3278.153978
std	1342.818304	2.256646	1342.687748	1342.963414
min	1124.170746	419.265586	288.173385	1960.168107
25%	1655.493384	425.980912	822.239911	2488.746857
50%	1904.039390	426.584673	1067.266090	2740.812691
75%	3054.954592	426.996986	2223.351619	3886.557565
max	5889.376449	427.192424	5053.982102	6724.770796

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

Analytical Insights:

- I have formatted the data into time series data by creating time stamp instead of YearMonth date field. It is a monthly data from January 1980 to July 1995.

- Did exploratory analysis and found out that there is no missing data, and more than 75 percent of data has sales count of 2549 and below with average sales count being 2402.

- performed decomposition to understand that there is no uniform trend. It keeps changing over time.

- There is a seasonality in the data. Peak sales are in December and least sales are in June. this suggest that the sales of Sparkling wine follow a festive seasonality.

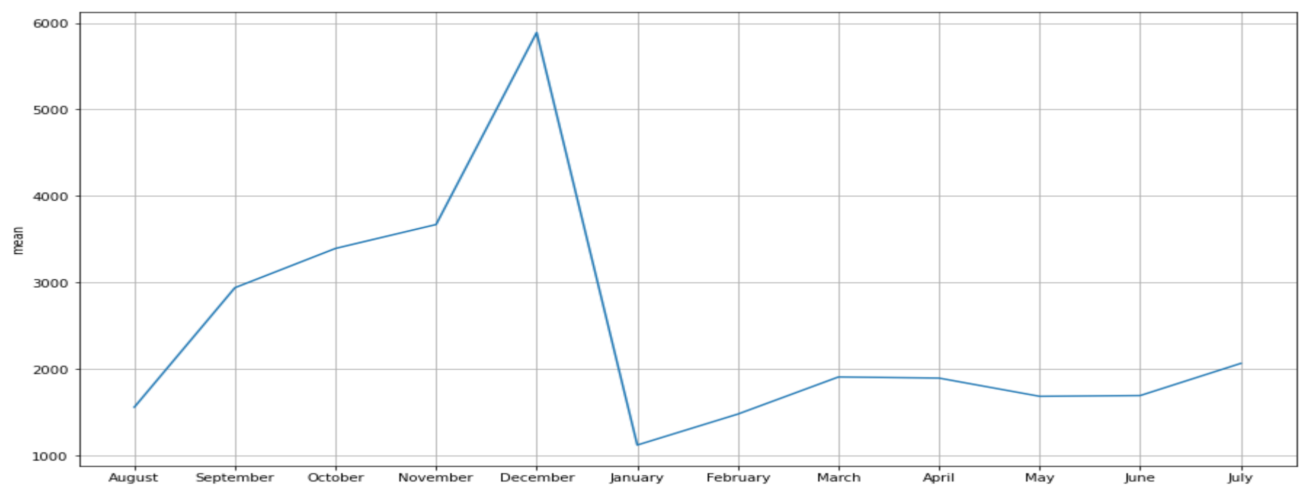
- Have split the data into train and test. Train data is from January 1980 December 1990. Test data is from Jan 1991 till July 1995 (end of the data set).

- Using Stationary test, we have found that the dataset is not stationary. After one differencing It becomes stationary.

- Have analyzed the data using various time series models and found out that SARIMA (3,1,2) (3,0,0,6) (using ACF AND PACF cutoff) is the most optimized model with least RMSE score.

- Have predicted the data for next 12 months using the SARIMA model with parameters of (3,1,2) (3,0,0,6) (The most optimized model). The maximum forecasted average sales will be in December 1995 with the count of 5889. The average sales count will be 2443.

- Forecasting for 12 months (Month wise sales count :



Business Recommendations:

- I suggest the "ABC Estate Wines "to have average stock up a minimum of 2450 sparkling wine every month.

- I suggest them to increase the stock by 10 percent every month for the time July to October and increase the stock by 25 percent each month for the month of November and December.

- I recommend them to have a stock of 6724 sparkling wine during December which is the forecasted maximum average sale count.

- The difference between forecasted maximum sale count and minimum sale count for month of December is around 1500 which is around the forecasted sales for January .So we can sell the remining stock by January and February .So risk is minimum for having additional stock during December .

- As the Sparkling wine sales has festive seasonality (the sales are maximum in December so it may be due to Christmas and new year) I suggest them to spend more on marketing during the month of October to December. It may help them to increase sales.

- As the sales are least during January till June, I suggest them to give discount or voucher during these months. It may increase the sales.