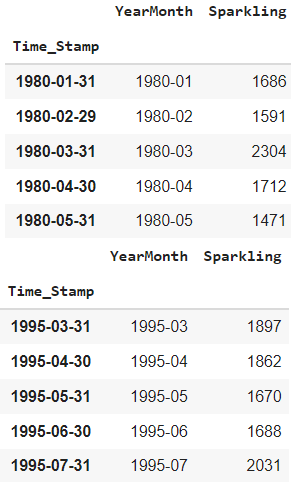
TIME SERIES FORECASTING PROJECT

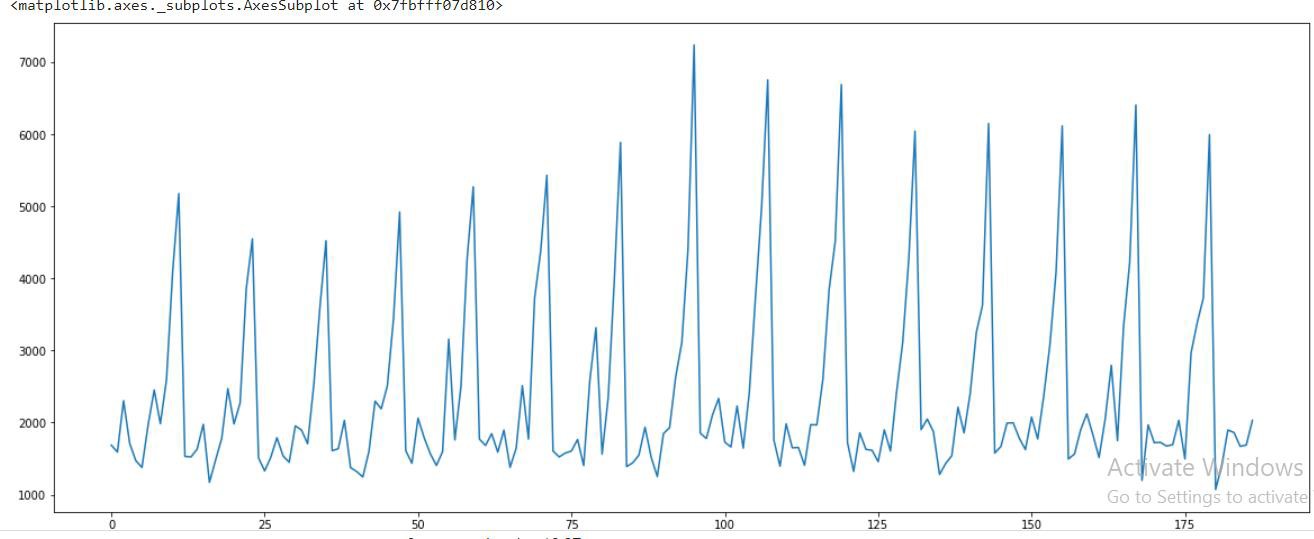
For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

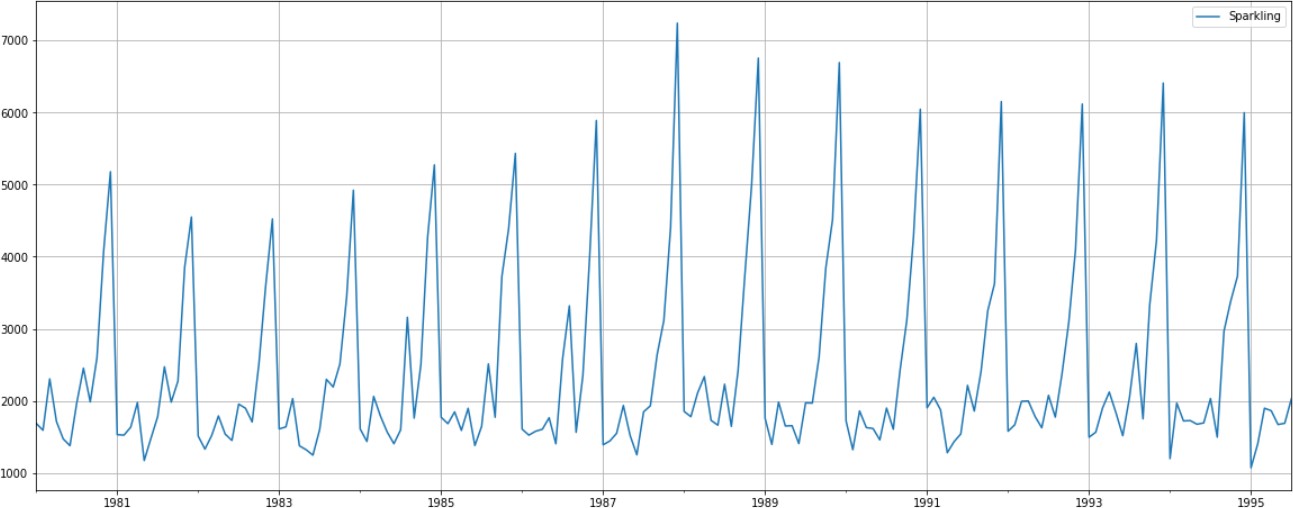
**SPARKLING.CSV**

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

Time Series is a sequence of observations recorded at regular time intervals.



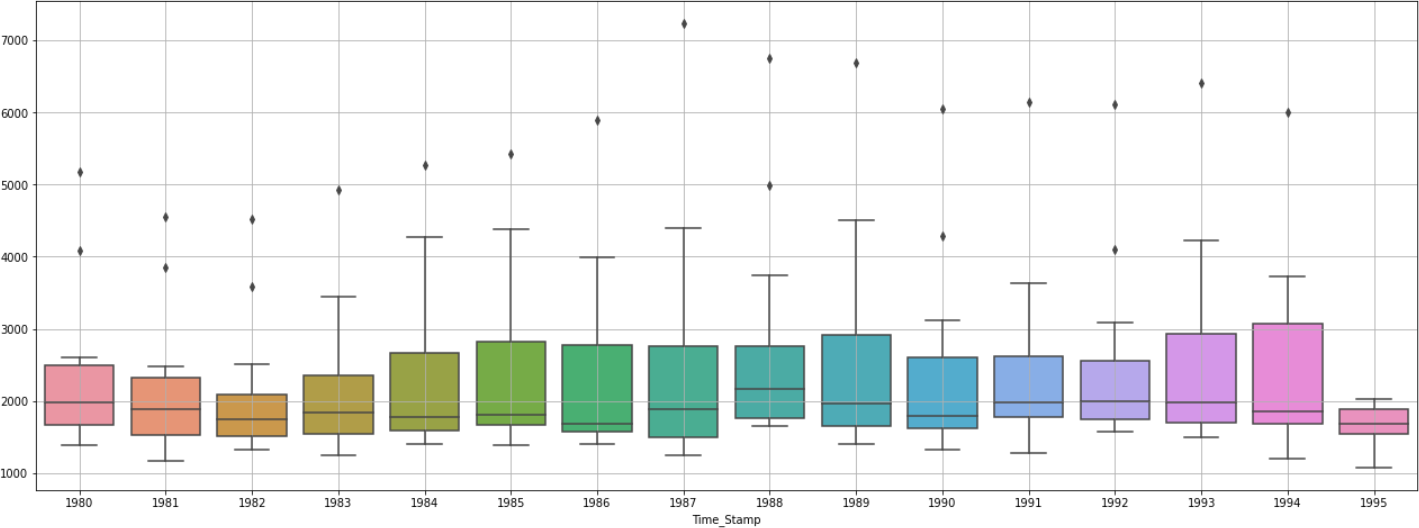


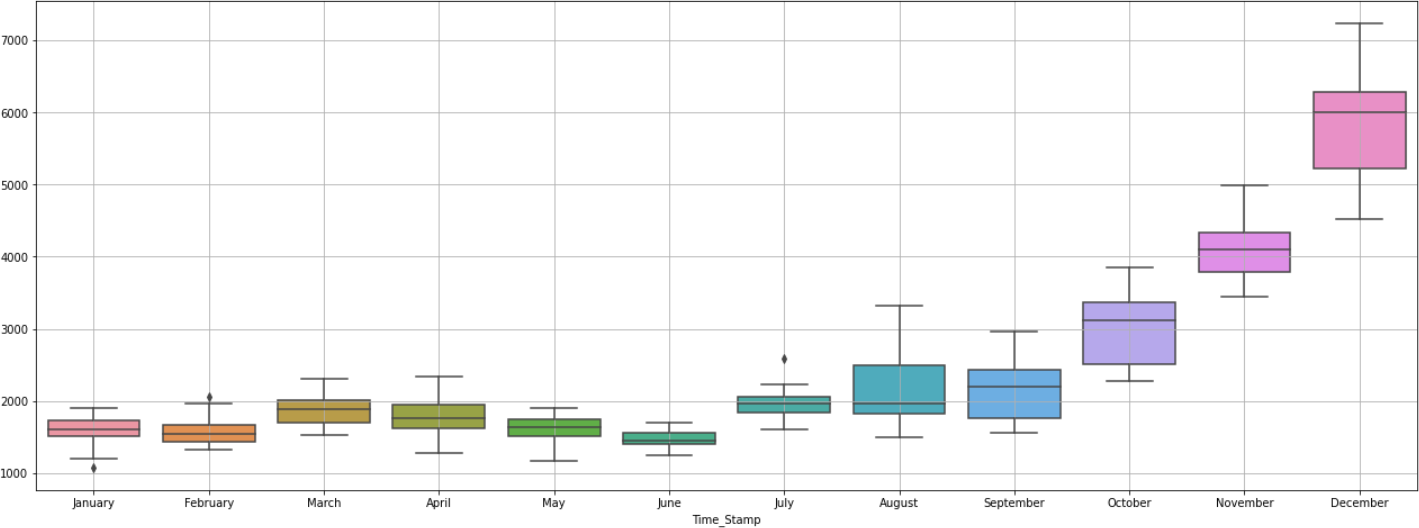
Given data is not time. So we parse the date range and create a timestamp. We also notice the increasing trend in the initial years.

* + Data consist of 187 data points
  + It seems to be contain seasonality

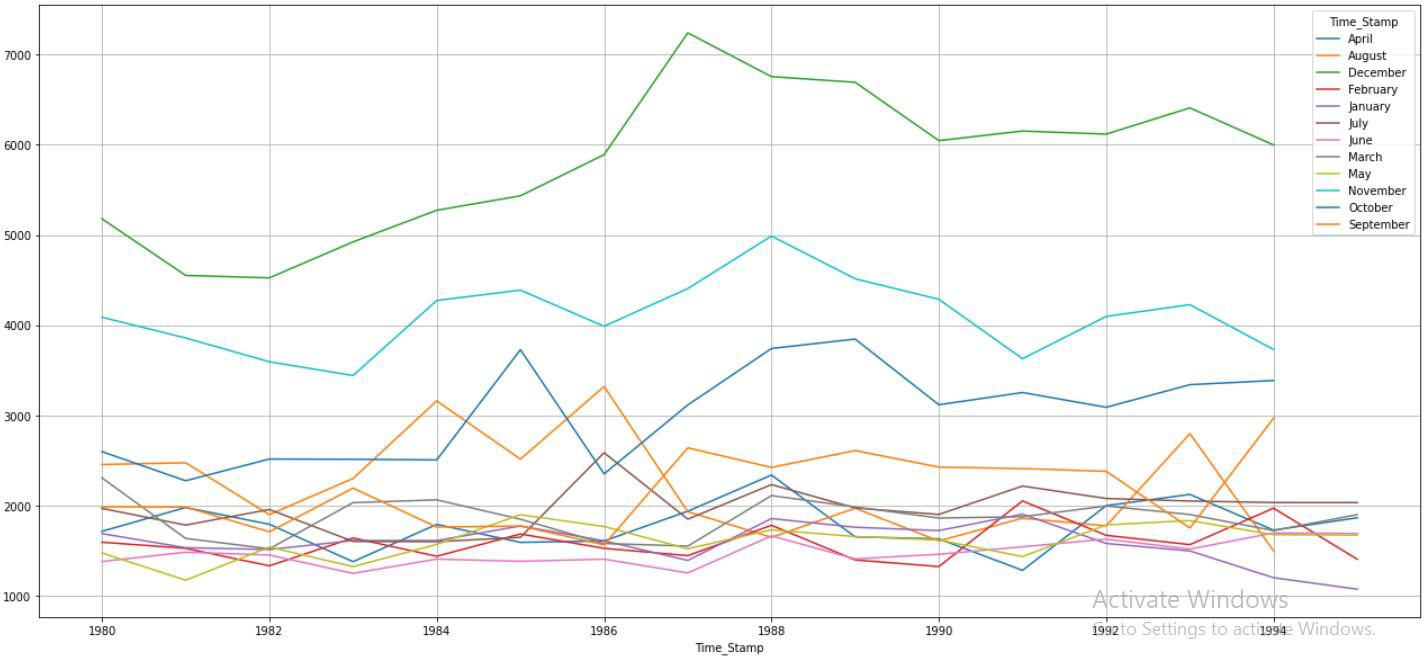
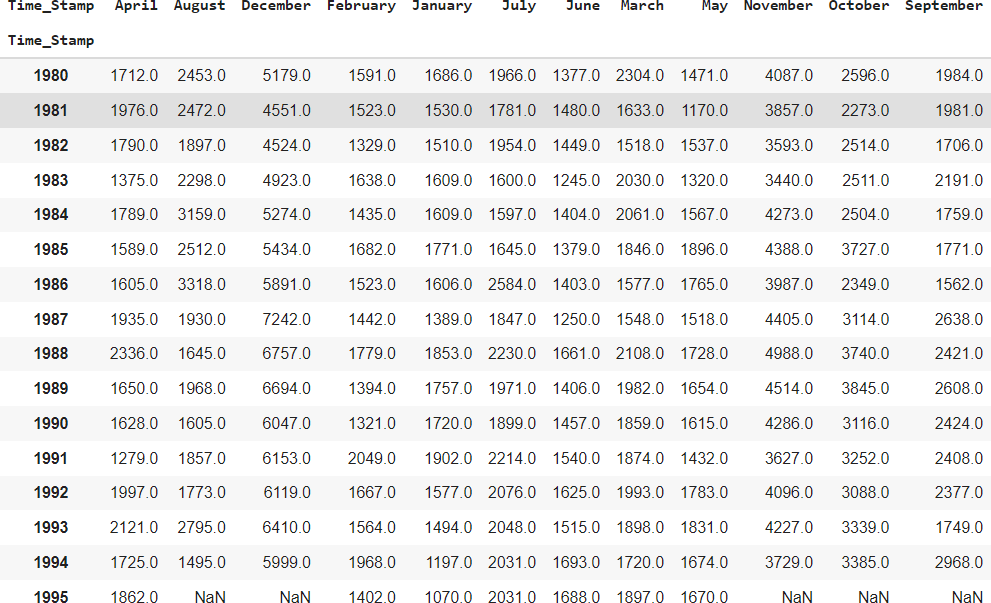
1. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

From 1981 to 1988 there is an increase in the sparkling data. After that, there is a decrease or fall. Seasonality is seen from the stable fluctuations repeating over the data.

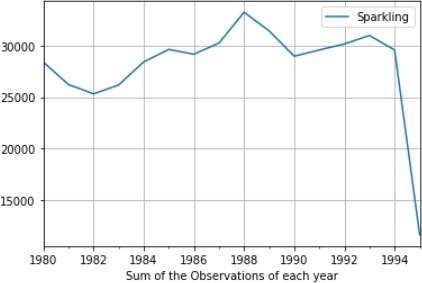
* + To understand the spread of the data, we use plotting.
  + Boxplot helps to check the outliers in each year and month.
  + Yearly plot
  + Boxplot indicates the trend being present in the data.
  + We can clearly see some of the outliers in the plot.
  + Monthly plot
  + The box plot for various months is plotted
  + Monthly plot contains outliers in the month of January, February and July.



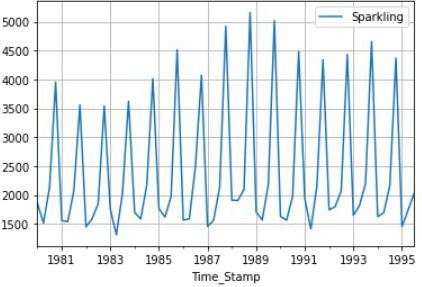
* + Plot for different months and different years



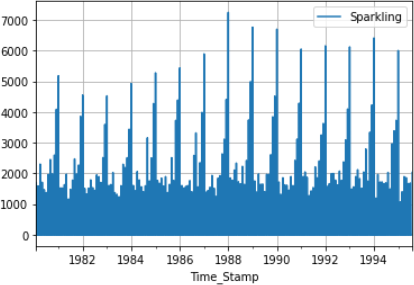
* + December records have the high number of wine sales followed by November and October.
  + May, June and July have low number of wine sales.
  + Yearly Plot – aggregate the time series from an annual perspective and summing up the observations

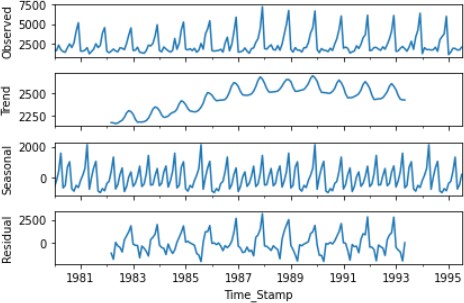
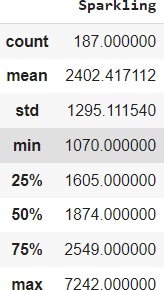
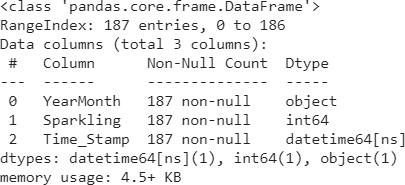


* + The plot shows that in 1982 there is a fall in the wine sales and a rise in 1984 and fall in 1986 followed by a maximum rise in 1988. In 1993 there is an increase and in 1994 there is a steep downfall is observed.
  + Quarterly plot – aggregate the time series from a quarterly perspective and sum the observations of each quarter.

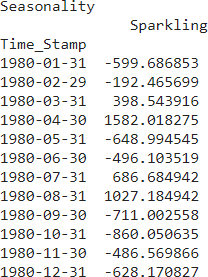


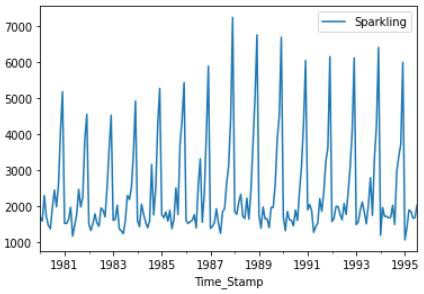
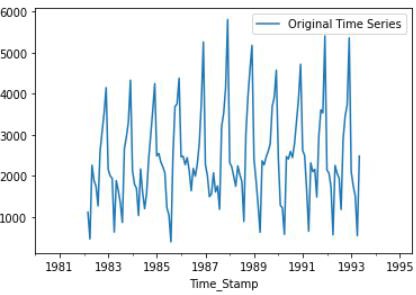
* + High rise is found in 1988
  + Daily plot –aggregate the data from a daily perspective
  + Resampling can also be used for better overview





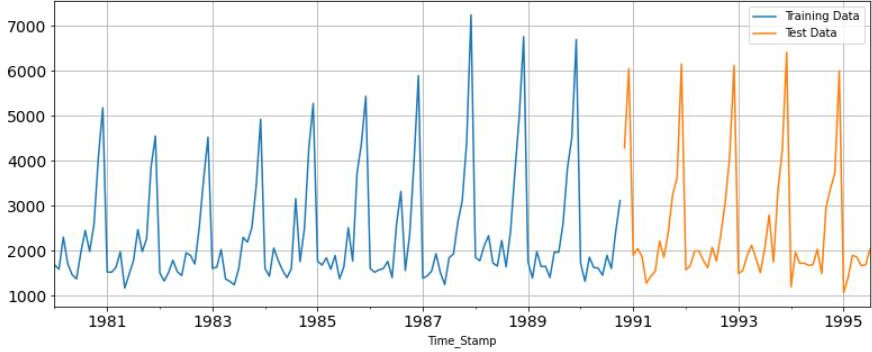
* + From the decomposition, there is seasonality in the data.



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

Train and test shapes

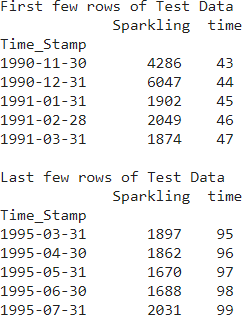
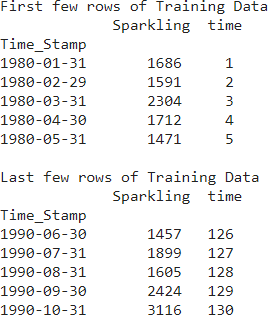


* + The test data starts from 1991
  + It is difficult to predict the future if the past is not happened. From the above split, we are predicting similar to the past data.

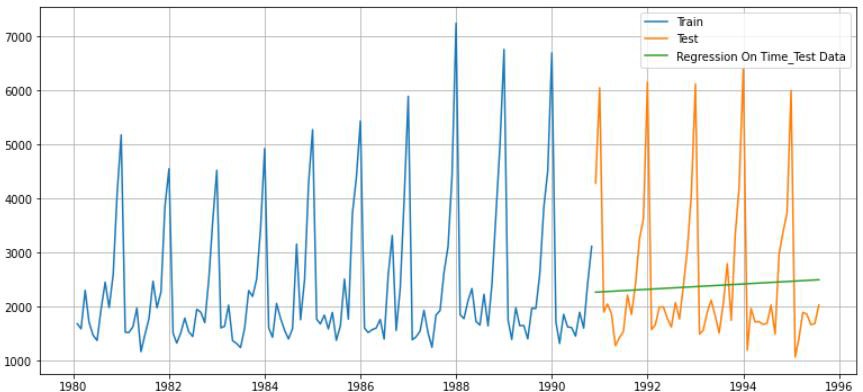
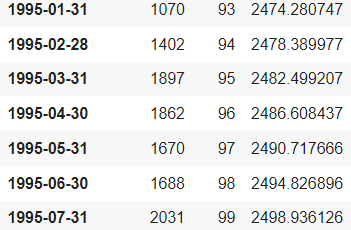
1. **Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.**

Model1: Linear Regression

* + Regress the “Sparkling” variable against the order of occurrence.
  + Modifying the training set
  + Generate the numerical instance order for both training and test set
  + Printing the head and tail of test and train data



* + Linear Regression is built on the training and test dataset



* + Defining the accuracy metrics
  + Evaluating the model

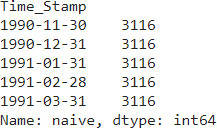


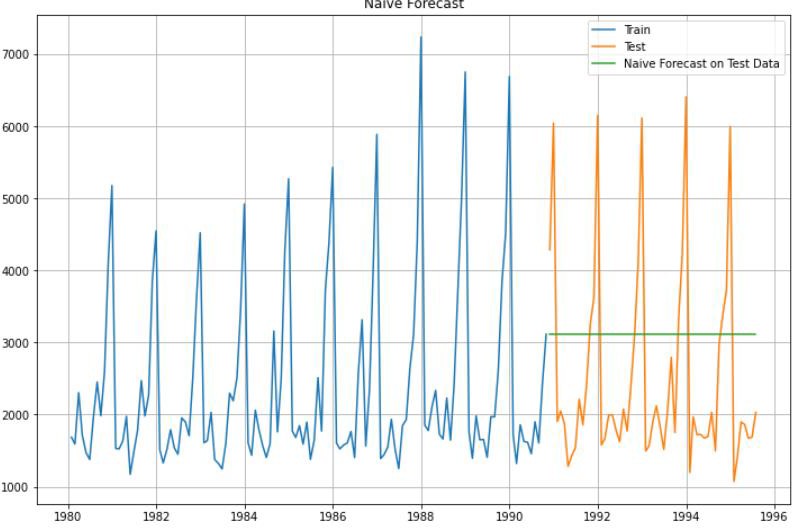


# Model2 – Naïve model

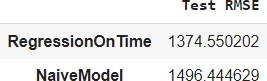
We say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow



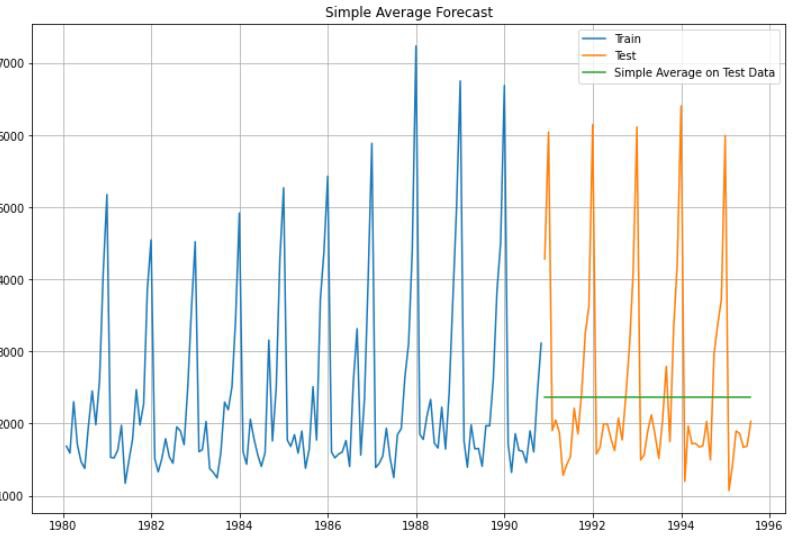
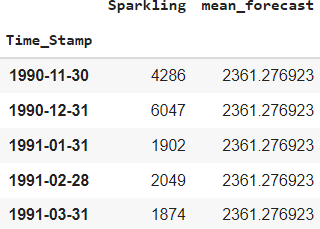




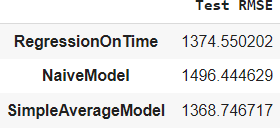




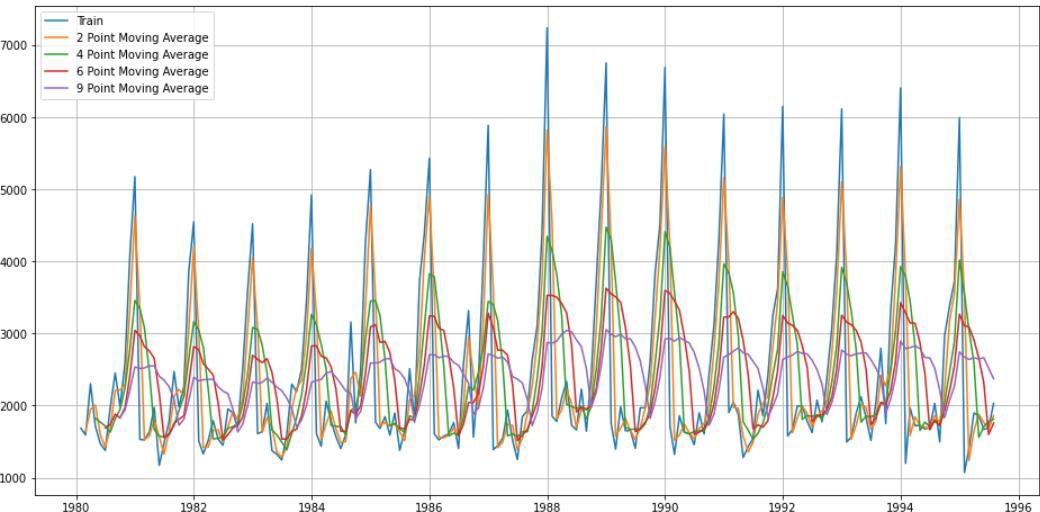
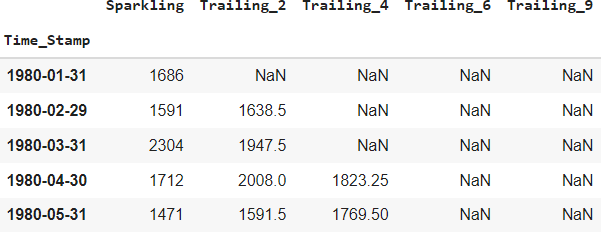
# Model3 – Simple Average – Forecast using the average of training values

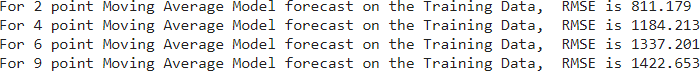




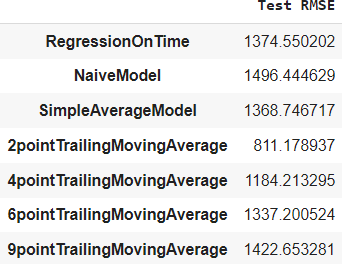
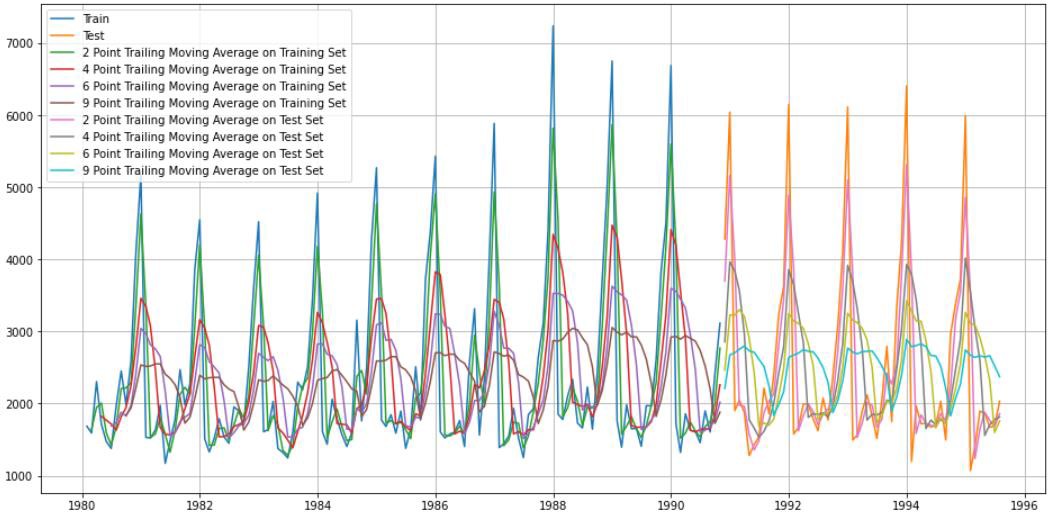


**Model4- Moving Average –** Calculating the rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

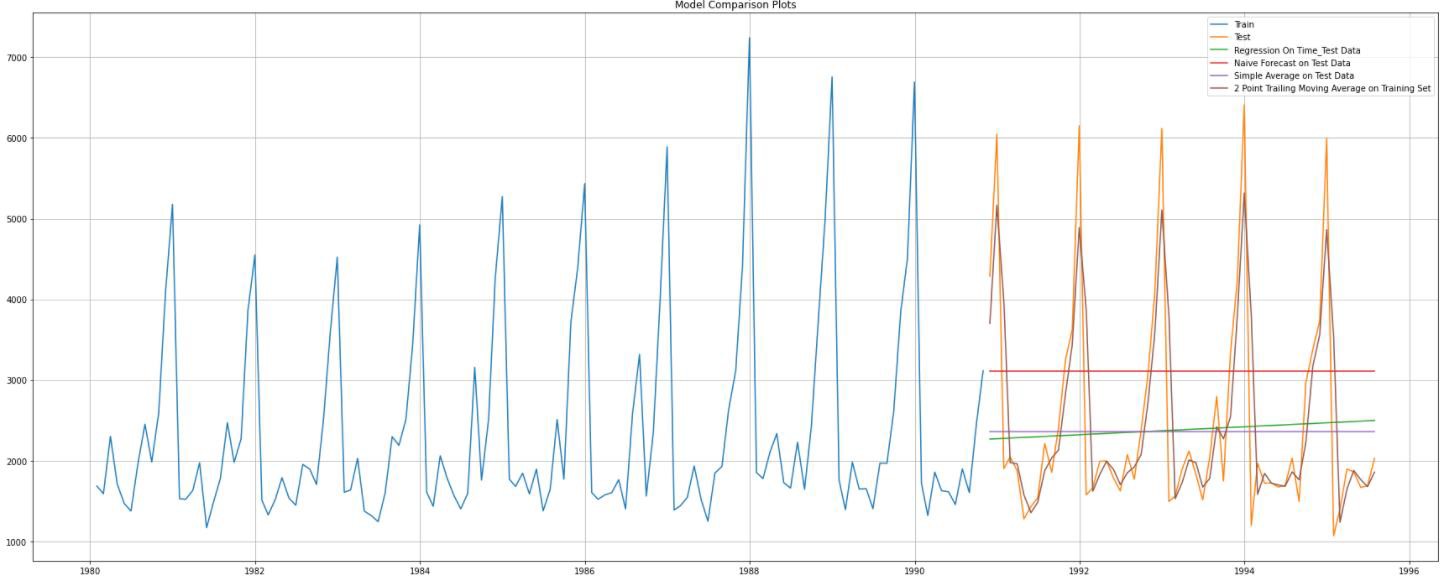




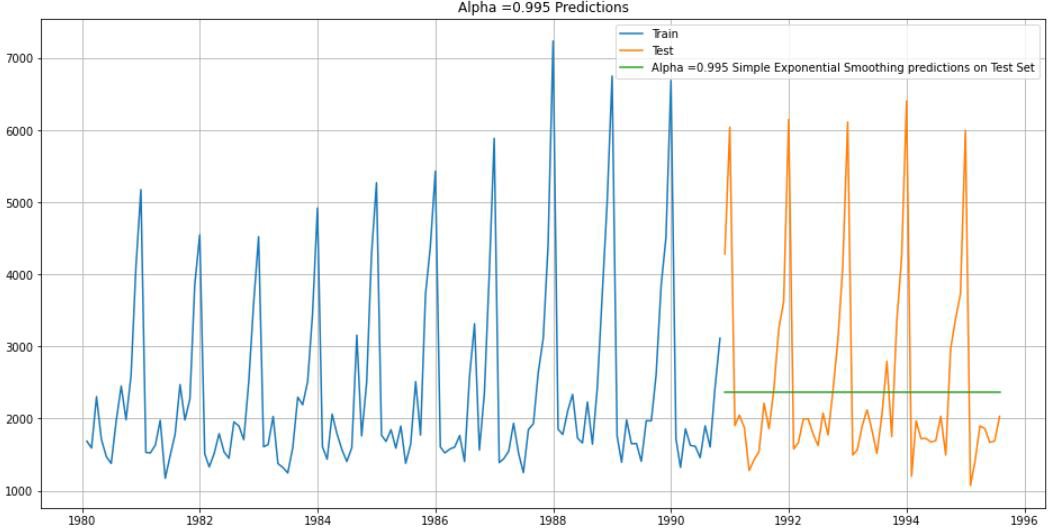
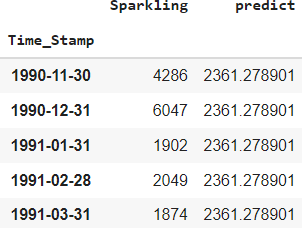
Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.



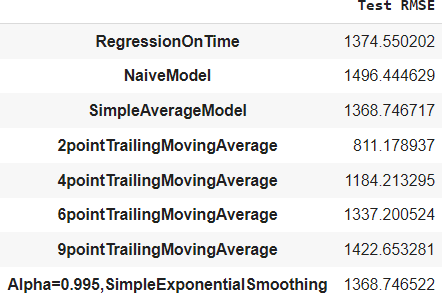
Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots



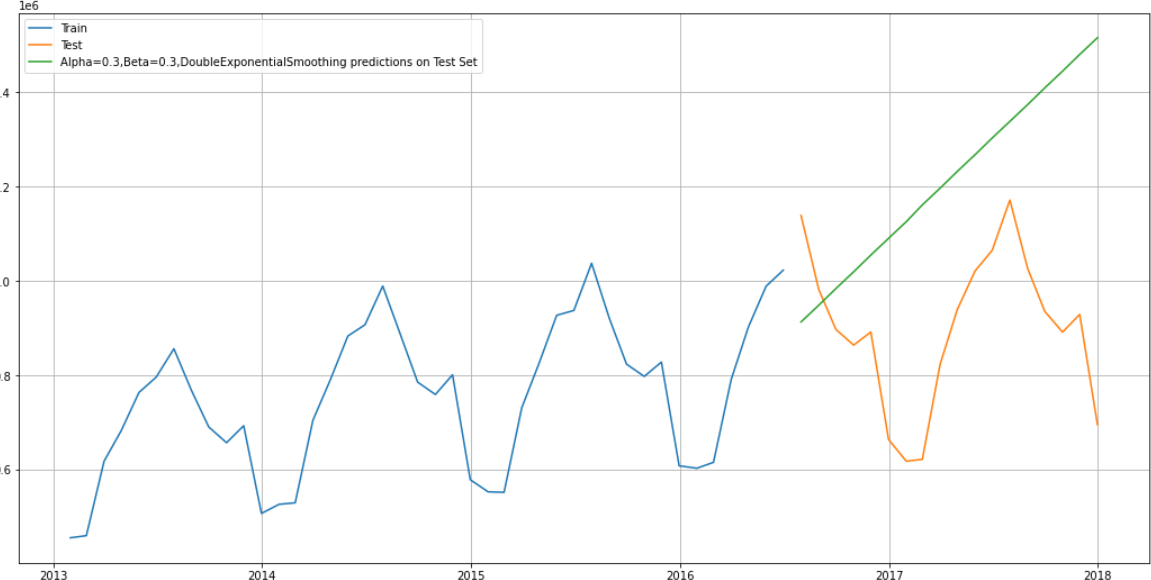
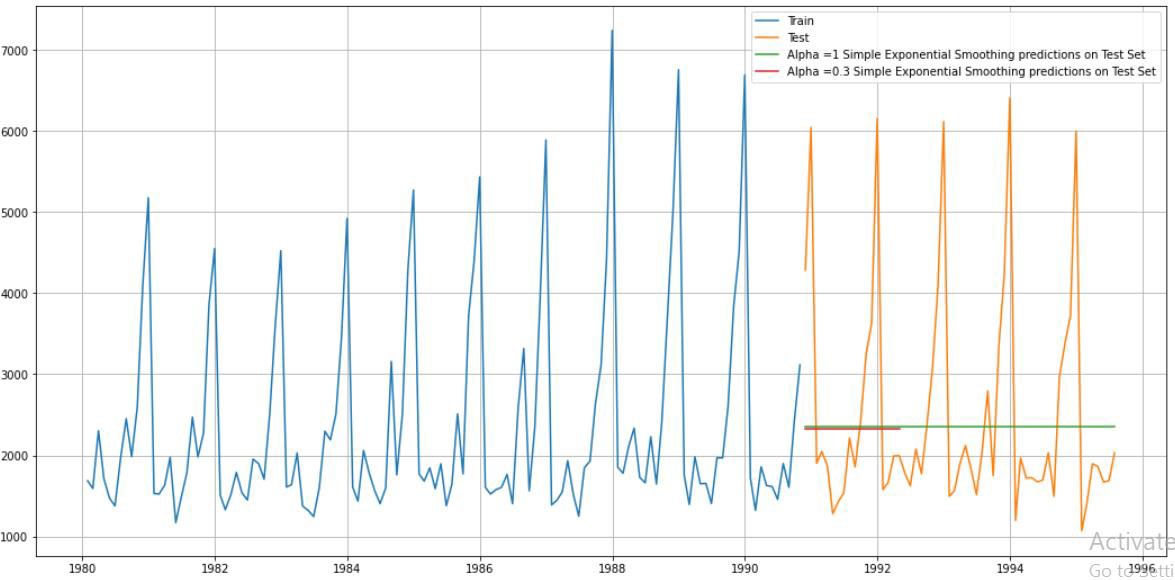
# Model -5- Exponential Smoothing







Setting different alpha values. Higher the alpha, the more weightage is given to more recent observation.



7000

|  |  |  |
| --- | --- | --- |
| -\_-..-.a,i,n..  - Alpha-0.676.Beta-0.088,Gamma-0.323.Tripl,1!£xponent1a1Smoothm9 pr@'diction5 on Tll!!:st Set  t- 1  ---+-- |  |  |
|  | | |

6000

5000

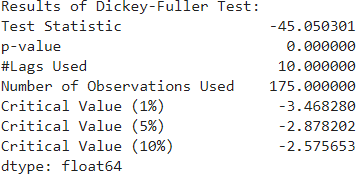
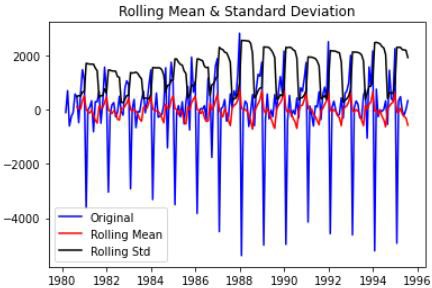
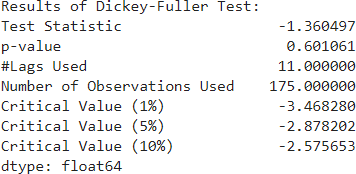
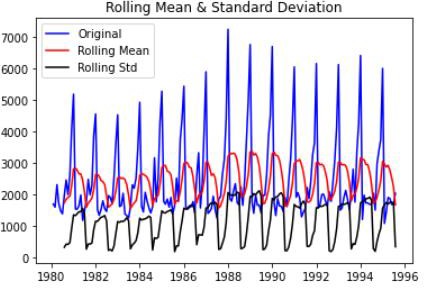
3000

JOOO

1980 1982 1984 1986 1988 1990 1992 1994 1996

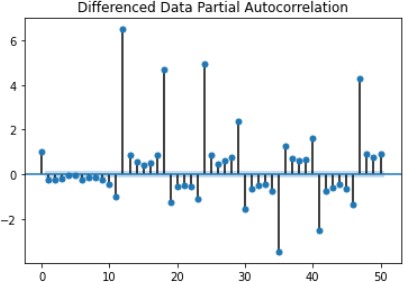
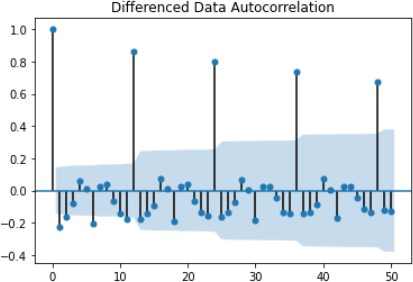
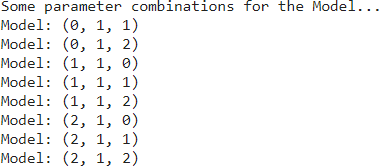
|  |  |
| --- | --- |
|  | **Test RMSE** |
| **RegressionOn** lime | 1374.550202 |
| **Naiiv,eModel** | 1496.44!4629 |
| **SimpleAverageModel** | 1368.74!6717 |
| **2poi11tTraiilingMovingAveraige** | 811.178937 |
| **4poi11tTraiilingMovingAveraige** | 1184.213295 |
| **6poi11tTraiilingMovingAverage** | 1337.200524 |
| **9p,oi11tTraiilingMovingAverage** | 1422.653281 |
| **Alpha=0.995,,SimpleExpo:nentiallSmoothing** | 1368.74!6522 |
| **Alpha=0.676,Beta=0..088,Gamma=0.323,**Trii1plleEx1pone11tiallSmoothing | 388.974278 |
|  | **Test RMSE** |
| Alpha=0..676,Beta=0.088,,Gamma=0..323,TriiplleEx1pone11tiallSmoothing | 388.974278 |
| 2po:intlirailingMovingAverage | 811.178937 |
| **4po:intlirailingMovingAverage** | 11184 213295 |
| **6po:intlirailingMovingAverage** | 1337.200524 |
| **Al1plha=0..995,Simp,leExp,one11tiallSmoothing** | 1368.746522 |
| **Al1plha=0..995,SimpleExpone11tiallSmoothing** | 1368.746522 |
| **SimpleAverageiModel** | 1368 746717 |
| **IRegres.sionOnlime** | 1374.5502()2 |
| **9po:intlirailingMovingAverage** | 1422653281 |
| **NaiveModell** | 1496444629 |

1. **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.**

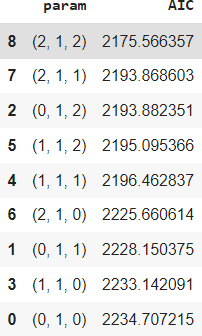
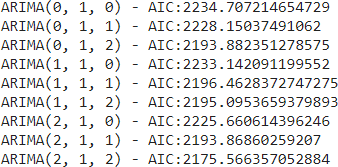


* + When the time series data is not stationary we need to convert it into stationary before applying models.
  + We use Augmented Dickey fuller test.
  + It determines how strongly a time series is defined by the trend.
  + From the null and alternate hypothesis, we define time series data is stationary or not.
  + We see that 5% significant level the time series is non-stationarity
  + P value >0.05 – Failed to reject null hypothesis - Stationary
  + Let us take a difference of order 1 and check whether the Time Series is stationary or not
  + At α = 0.05 the Time Series is indeed stationary.

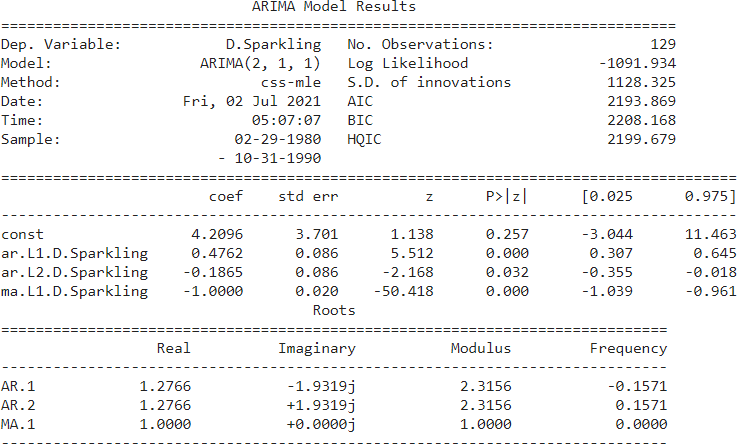
1. **Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**
2. **Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**



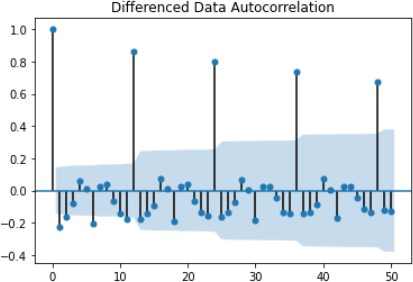
* + From the above plot, we see seasonality in the data.



* + If we have seasonality, then we should go for SARIMA model.
  + We are building ARIMA model by looking at minimum AIC values and ACF and PACF plots.
  + Sorting the AIC values to see the lower AIC value.







* + Again we plot ACF to see and understand the seasonal parameter of SARIMA model.
  + We see seasonality in 6 as well as 12.
  + We run SARIMA model by setting seasonality both as 6 and 12.
  + First iteration by setting 6 as the seasonality
  + We sort the AIC values to see the lowest of all vales.
  + Next predicting the data using the SARIMA model and evaluating the model.
  + We get the summary of the data

Examples of some paramet,er combinations for Model... Model: (0, 1, 1)(0, 0, 1, 6)

Model: (0, 1, 2)(0, 0, 2J 6)

Model: (1, 1, 0}(1, 0, 0, 6)

Model: (1, 1, 1)(1, 0, 1, 6)

Model: (1, 1, 2)(1, 0, 2J 6)

Model: (2, 1, 0}(2, 0, 0, 6)

Model: (2, 1, 1)(2, 0, 1, 6)

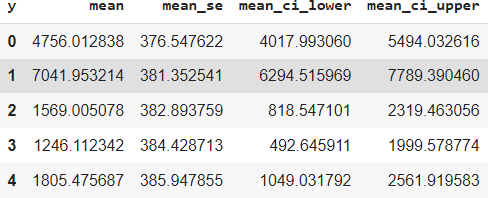
Model: (2, 1, 2}(2, 0, 2 **J** 6)

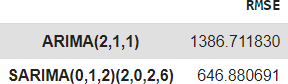
|  |  |
| --- | --- |
| -· · - *..-.-J*  SARIMA(2, **1,** l)x(2, 0, 0, 6) -  SARIMA(2, **1,** l)x(2, 0, 1, 6) -  SARIMA(2, **1,** l)x(2, 0, 2 **J** 6) -  SARIMA(2, **1,** 2)x(0, 0, 0, 6) -  SARIMA(2, **1,** 2)x(0, 0, 1, 6) -  SARIMA(2, **1,** 2)x(0, 0, 2 **J** 6) -  SARIMA(2, **1,** 2)x(l, 0, 0, 6) -  SARIMA(2, **1,** 2)x(l, 0, **1,** 6) -  SARIMA(2, **1,** 2)x(l, 0, 2J 6) - | ·--- ·-----. -. ---- -- - -- -  AIC:1731.8137132625177 AIC:1733.7098303257294 AIC:1710.4498904719412 AIC:2140.669395942271 AIC:2042.7000095525877 AIC:1850.8435518017636 AIC:2038.5071951549105 AIC:1927.69282782436 AIC:1796.3236930264027 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **53** | **p,aram**  (1,1,2) | **seasonal**  (2, 0, 2, 6) | **AIC**  1694.325453 |
| **26** | (0, **1,** 2) | (2, 0, 2, 6) | 11694 839212 |
| **80** | (2, **1,** 2) | (2, 0, 2, 6) | 1695.565322 |
| 17 | (0, **1,** 1) | (2, 0, 2, 6) | 1708.125767 |
| **44** | (1, **1,** 1) | (2, 0, 2, 6) | 1710045544 |
| Statespac,e Model Results | | | |

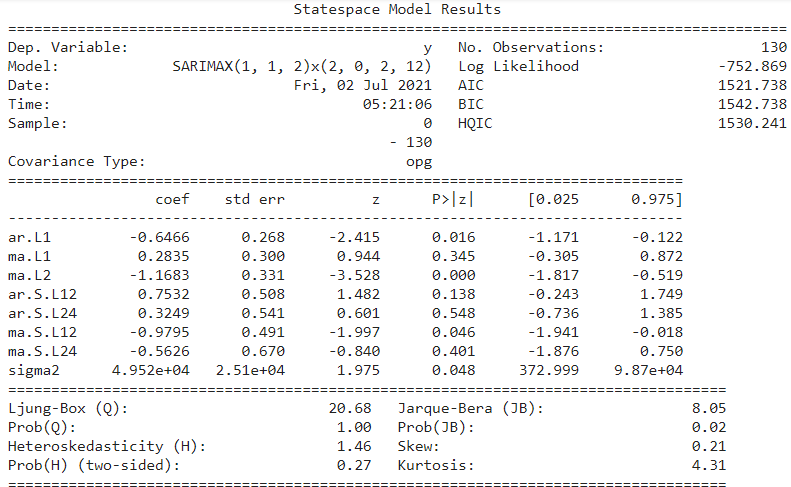
|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Variable:  Model: | y  SARIMAX(0, 1, 2)x(2, 0, 2, 6) | Nb. Observations::  Log Likelihood | B0  -840.420 |
| Date: | **Fri,** 02 Jul 2021 | AIC | 1694.839 |
| Time: | 05:11:33 | BIC | 1713.993 |
| Sample: | 0 | HQIC | 1702.612 |

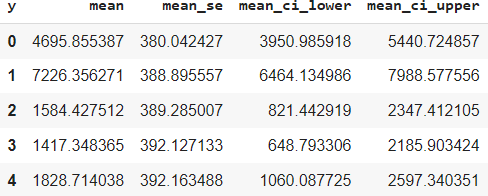
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Covariance | Type: |  |  | - 130  opg |  | | |
|  | coef | std err | z |  | P> I z. I | [0.025 | 0.975] |
| ma.Ll | -1.0171 | 0.153 | -6.663 |  | 0.000 | -1. 316 | -0.718 |
| ma.L2 | -10.0825 | 0.120 | -0.685 |  | 0.493 | -0.318 | 0.153 |
| ar.S.L6 | 10.0073 | 0.022 | 10.326 |  | 0.744 | -0.036 | 0.051 |
| ar.S.L12 | 1.0571 | 0.017 | 62.698 |  | 0.000 | 1.024 | 1.090 |
| ma.S.L6 | 10.0334 | 0.142 | 10.235 |  | 0.815 | -0.245 | 0.312 |
| ma.S.L12 | -10.6723 | 0.086 | -7.819 |  | 0.000 | -0.841 | -0.5,04 |
| sigma2 | l,.187e+05 | 1. 7e+04 | 6.990 |  | 0.000 | 8.,55e+04 | l.52e-f<05 |
| Ljung-Box | (Q): | 25.24 | | Jlarque-Bera (JB): | | | 30.2.5 |
| Prob(Q): |  | 0.97 | | Prob(JB): | | | 0.00 |
| Heteroskedasticity (H): | | 2.99 | | Skew: | | 0.44 | |
| Prob(H) (two-sided): | | 10.00 | | Kurtosis: | | 5.37 | |



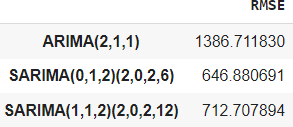


* + There is a huge gain in the RMSE value by including seasonal parameters
  + Keeping 12 as seasonal parameter for second iteration



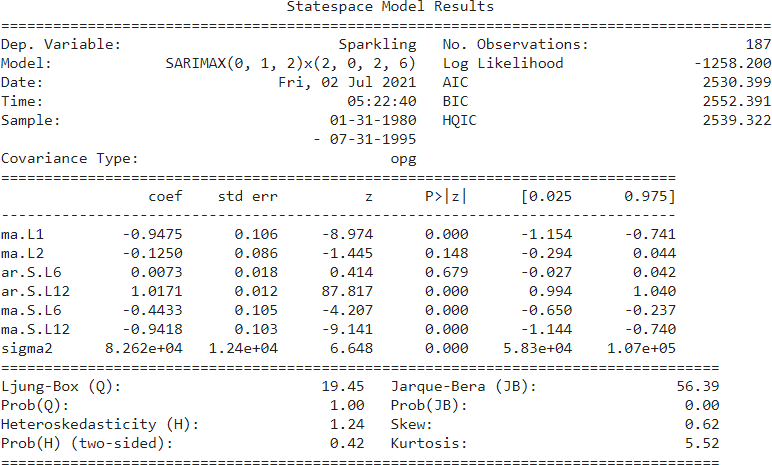


1. **Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**



* + It is clear that SARIMA(0,1,2)(2,0,2,6) has the lower RMSE and ARIMA(2,1,1) has the higher value.

1. **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.**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Spa,1rkling** | **me,an** | **mea,1n\_se** | **mean\_ci\_loNer** | **m.ean\_ci\_upper** |
| **1995-08-31** | 1863.656191 | 372.567001 | 1133.438287 | 2593.874095 |
| **1995-09-30,** | 2393.459684 | 378.443212 | 1651.724617 | 3135.194!750 |
| **1995-10-31** | 3285.461780 | 379.292653 | 2542061841 | 4028.861719 |
| **1'995-11-30** | 40117.185571 | 38!)\_140202 | 3272.124465 | 4762.246676 |
| **1995-12-31** | 6286.4!73411 | 380.985873 | 5539 754821 | 7033.192001 |
| **1996-01-31** | 1220.495767 | 38L829677 | 472.123352 | 1968.868183 |
| **1996-02-291** | 1544.236886 | 381.967774 | 795 593805 | 2292.879967 |
| **1996-03-31** | 1777.91522!) | 382.6404!94 | 1027.953633 | 2527.876807 |
| **1996-04-30** | 1781.404294 | 383.413983 | 1029 926696 | 2532.881892 |
| **1996-05-31** | 1665.476283 | 384.185923 | 912.485711 | 2418.466855 |
| **1996-06-30** | 1637.321580 | 384.956318 | 882.821062 | 2391.822098 |
| **1996-07-31** | 1980.909543 | 385.725179 | 1224 902085 | 2736.917002 |

RMSE of the Full odel 530.8502242211852

**StandardiZJed residua,!**

4

2

o **NlflPtllWWl-'McdMMMfl** 0.2

**-2**

**-2 0 2**

**TheoreticalQuantiles**

**1.,0**

**0.0 2.5 s.o 7.5 10.0**

**7000**

* **6000 5000**

**4000**

i 3000

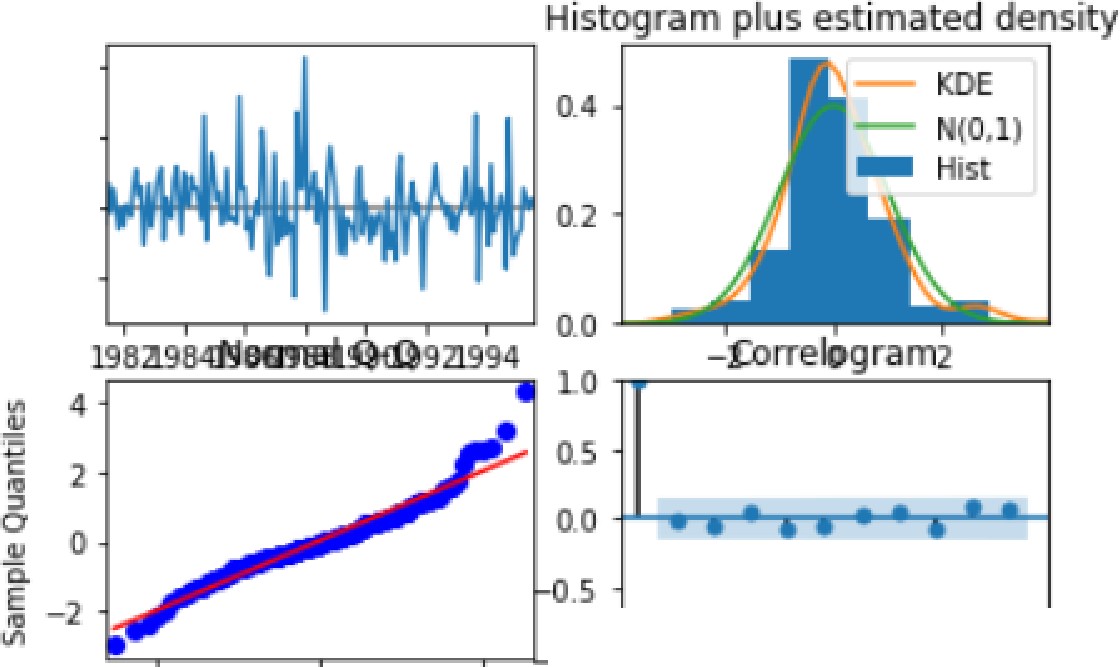
**2000**

'

R

**1000**

- **Cb-served F c:ast**



**1984 1989 1994 1999 2004 2009 2014**

**ar MOJ'llf'ls**

1. **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**
   * To find the most optimum model, we run the model on the full data
   * Correlogram, histogram, residual and quartiles are shown.
   * We predict for the next 12 months for next years.
   * We get forecast.
   * RMSE of the full complete data is 530.8
   * Plotting the forecast with the confidence band
   * It is clear that SARIMA(0,1,2)(2,0,2,6) has the lower RMSE and ARIMA(2,1,1) has the higher value.