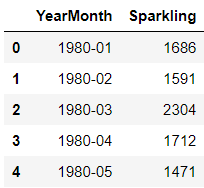
**Problem summary:**

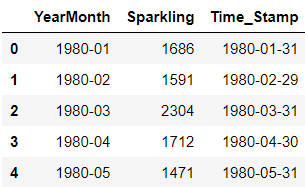
To forecast the Sparkling wine sales in the 20th Century using the time series data at our disposal. We are given the monthly sales figures of the Sparkling wine sales from January 1980 to July 1995.

**EDA:**

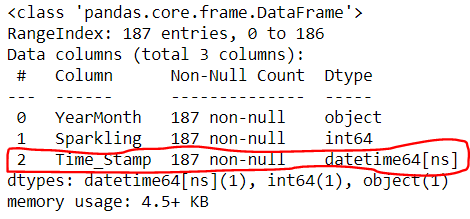
After importing the necessary libraries, we read the data and import the data into a data frame as follows:



The data has been read as a normal data and so we use the datetime library to read the data as a time series data using the start and end dates and the frequency as monthly.

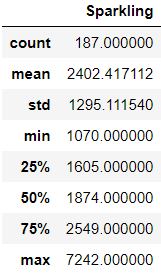


We then use the info command on the data to check if the time series data has been read and identified appropriately:



It can be seen that the required column is read correctly. We then remove the column: YearMonth from our dataframe.

We then use the describe function on the data to look at the description of the Sparkling wine sales



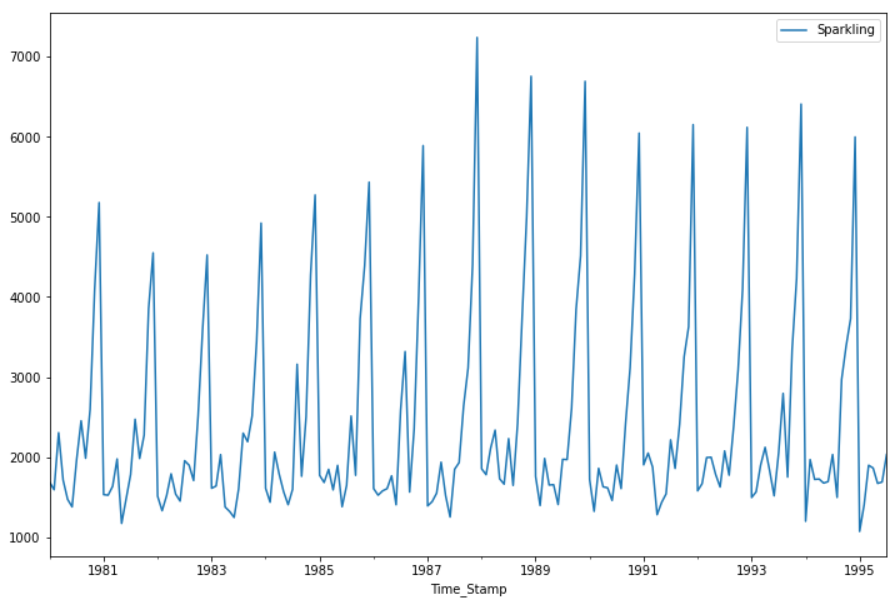
The mean and the median are not close to each other. The distance between the 25th percentile and the 50th percentile is less as compared to the distance between the 50th percentile and the 75th percentile. We see that the data has right skewness in it with many extreme values on the upper end of the spectrum. The mean is greater than the 50th percentile or the median which also suggests a right skewed distribution. It can be seen that the standard deviation is pretty high suggesting that there are may values between the 75th percentile and the maximum value. This also augments the statement earlier that we must have many values at the upper end. We will have to closely analyze these extreme values as they are of Sparkling wine sales for particular months.

Moving on, we check for the null values, na values, non-numerical object values in the data and find that there are no null values present in the data. The absence of any non-numerical object values in the data can be confirmed by the data type that is present. Since there are no object data types in our data this can be confirmed.

The shape of the data looks like this:

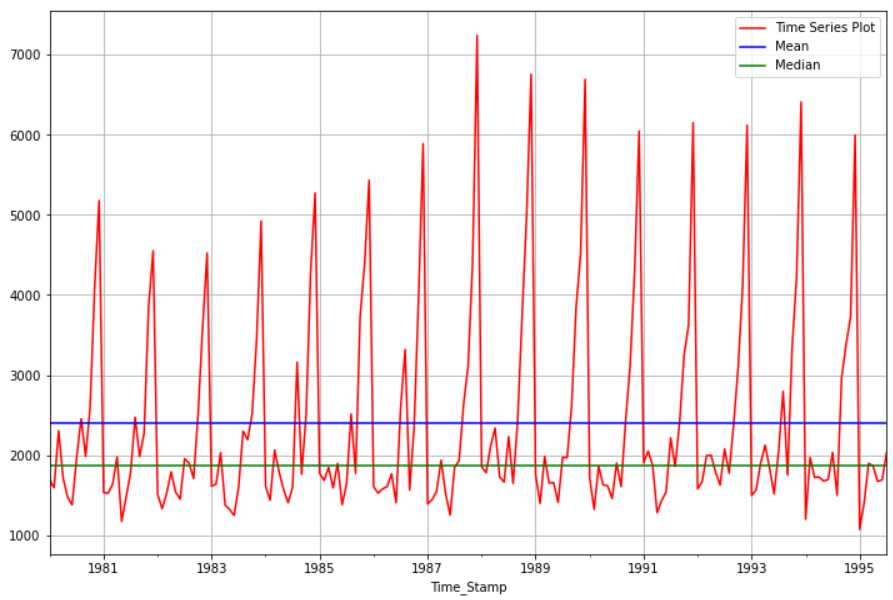


**Plotting Time series data:**



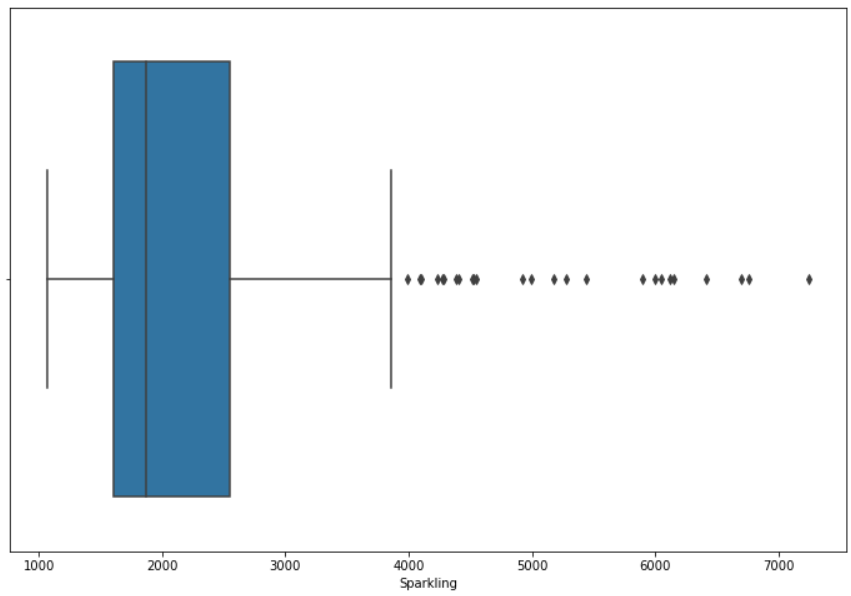
From the above plot it can be seen that the data has both trend and seasonality as well.

**Plotting Time series data along with Mean and Median:**



As stated earlier, the mean and median are not close to each other.

**Boxplot of Time series data:**



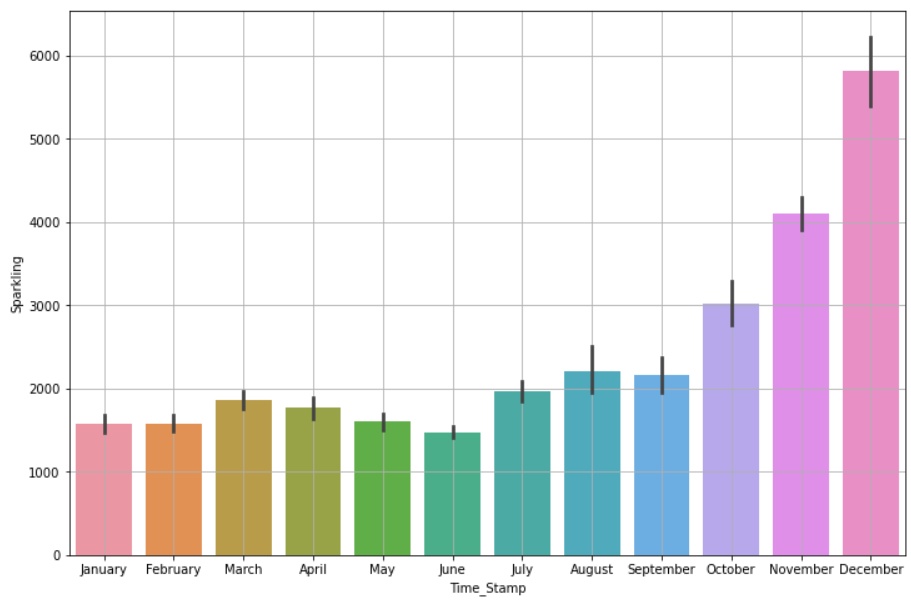
The plot above shows us that there are outliers present in the data beyond the upper whisker of the boxplot.

### Quarterly comparison of shipments using barplot:

### 

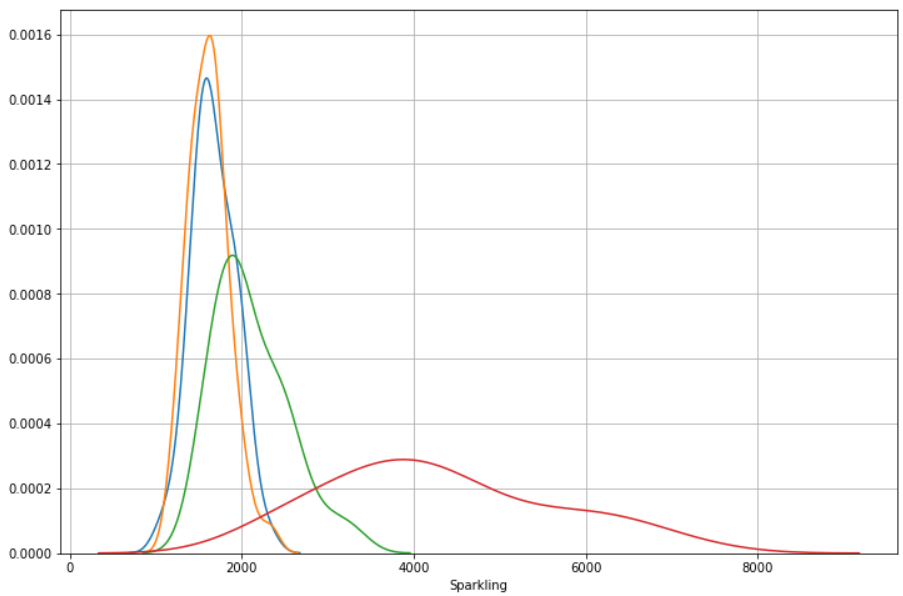
The above plot shows us that the sales were highest in the 4th quarter of each year. The sales figures are lowest in the second quarter each year and increases to more than 20% in the third quarter. The sales figures in the 4th quarter are almost twice as much as in the third quarter, every year.

**Monthly comparison of shipments using barplot:**



The Sparkling wine sales is the highest in the month of December and lowest in the month of June. In March the sales increase as compared to Jan and Feb only to see a gradual decline in the months of April May and June.

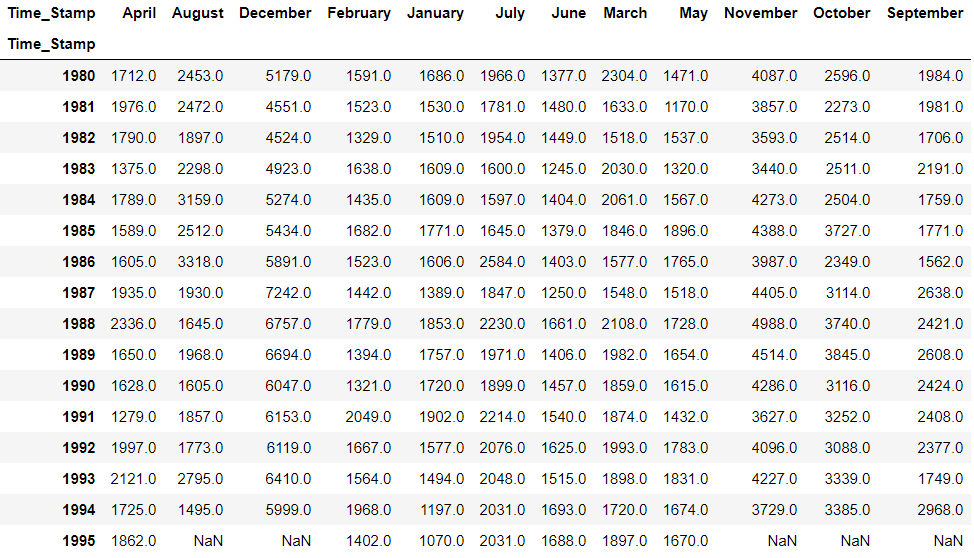
**Distribution plot for quarterly comparison of Sparkling wine sales:**



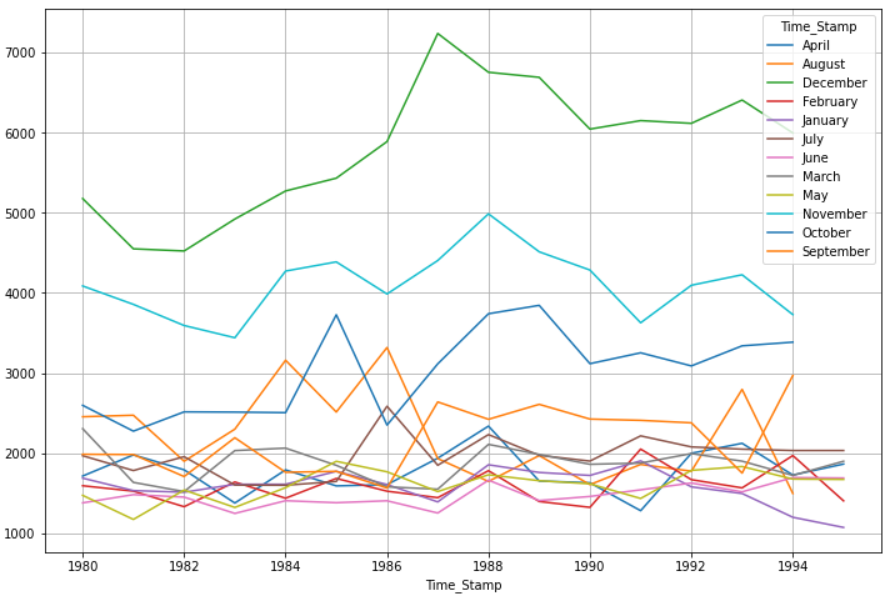
The quarterly distribution plot for the sales is shown above. The red line is the sales for the 4th quarter of years. Blue line corresponds to the first quarter, the orange line to the second quarter and the green line to the third quarter. The distribution in the second and first quarters of each year are distributed in a narrow price range as opposed to the sales in the third and fourth quarters.

**Monthly sales across years:**

We create a data frame to observe the monthly sales across each year and then plot the same.

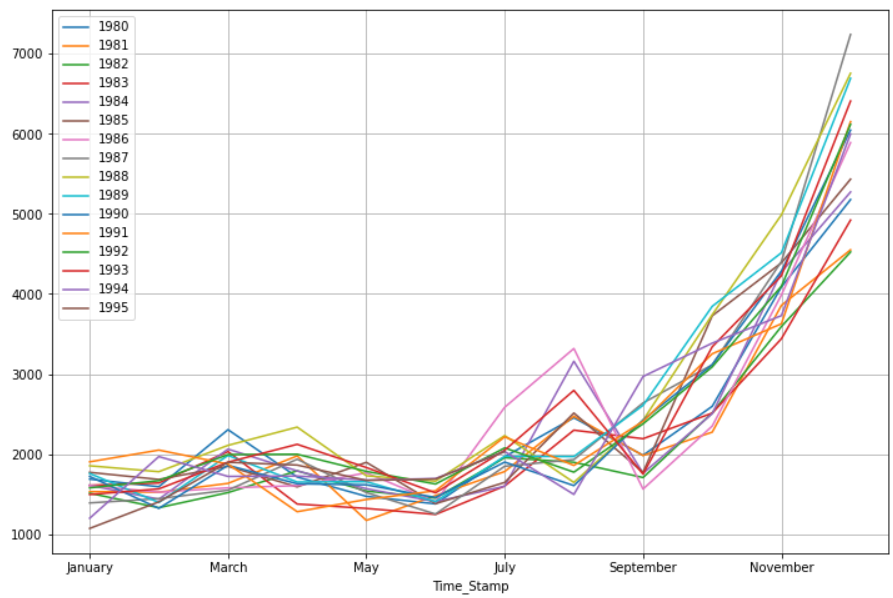


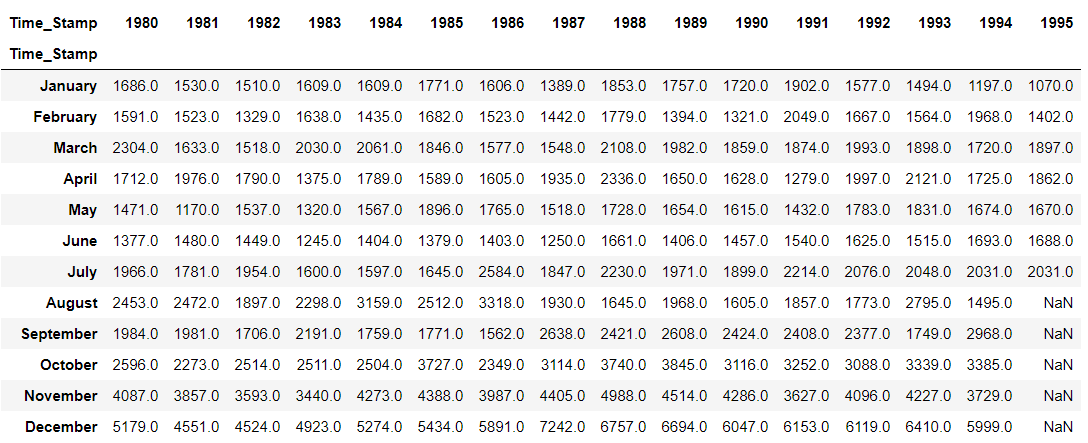
It is to note that ‘Time\_Stamp’ is the time index that we have assigned to the data frame.



* In the year 1993, the sales for Sparkling wine in the month of September has gone up to 2800 sales units whereas, the August month sales dropped sharply during the same time period.
* A similar trend has been observed in the year 1985 where the sales units of Sparkling wine has shot up to around 3700 units during October and in the month of September the sales figures have dropped by around 600 units.
* In the month of August between 1986 – 1987, the sales figures have increased by 1100.
* Between 1986 – 1988 August, the sales figures have dropped by around 1600 units.
* Between 1987 and 1988 March the sales figures have increased by more than 600 units.
* Between years 1990 and 1991 February, the sales figures have increased by more than 600 units.
* Between 1981 – 1982 September, the sales dropped by around 500 units.
* Between 1991 – 1992 April, the sales increased by around 600 units.

**Yearly sales across months:**





The trend is to have higher sales in the end of the year: November and December.

* Between June and August in the year 1986, the sales numbers are (1403 + 2584 + 3318) = 7305
* Between February and April in the year 1988, the sales numbers are (1779 + 2108 + 2336) = 6223

### Empirical Cumulative Distribution Function Plot:

### 

### The first 15% of the sales takes a slow start of 1200 sales units.

### Beyond this, there is a rapid increase until about 50% till 1800 – 1900 sales units.

### There is close to 40% of sales in the upper end of the spectrum and this needs to be looked into more carefully by the management.

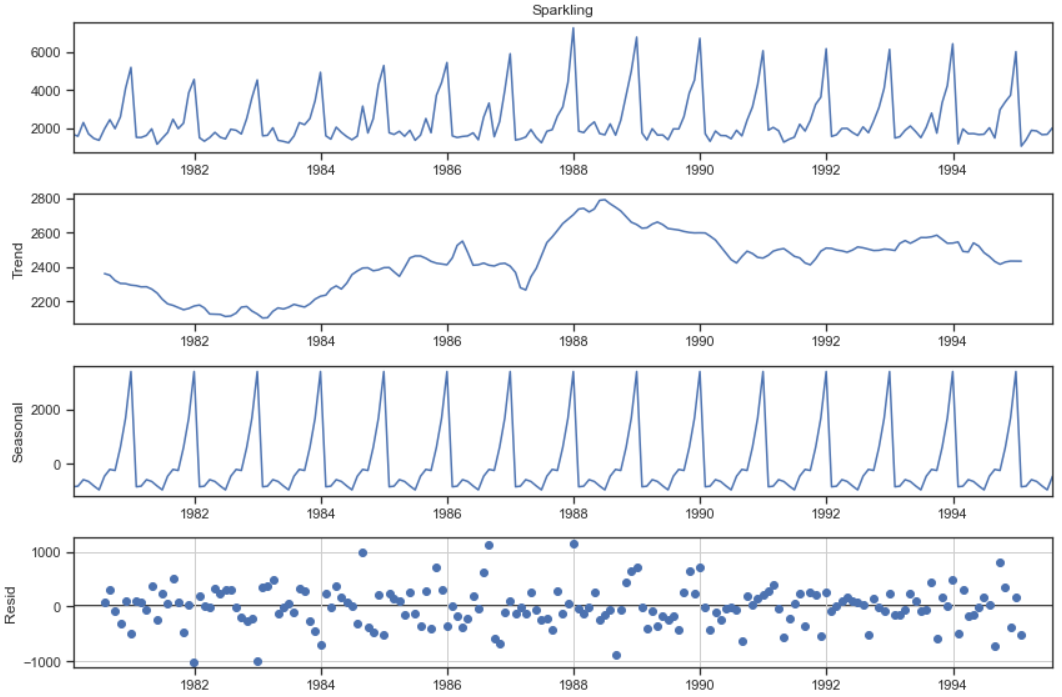
### Average sales and percentage change:

### 

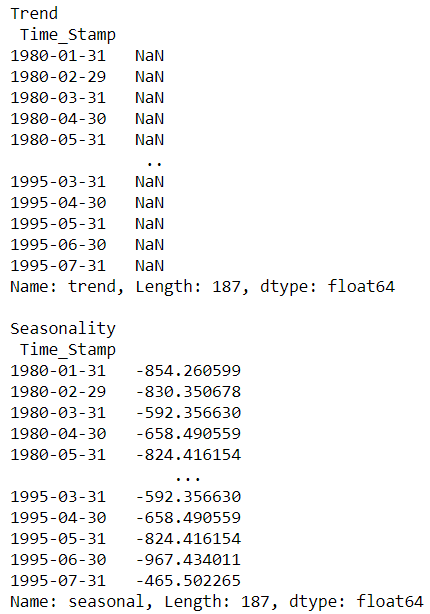
The average customers were low during the years 1982 – 1983. The average was highest in the year 1988. Beyond which the numbers have again reduced gradually towards the year 1991. And the numbers have again picked up beyond the year 1991. The high percentage increase of close to 140% has been observed in August to September in the year 1985. In the same year 1985, After September, there is 150% decline in sales until December of 1985. Between July and August in the year 1984 a 60% increase in sales units was observed. It is during the mid of the years is when the sales are low.

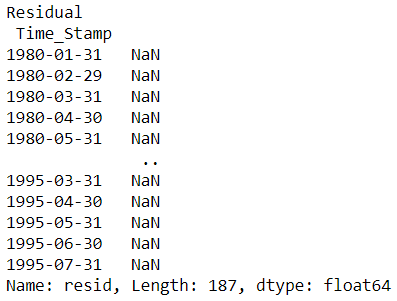
**Decomposition of Time Series:**

Let us decompose the time series in an additive way:

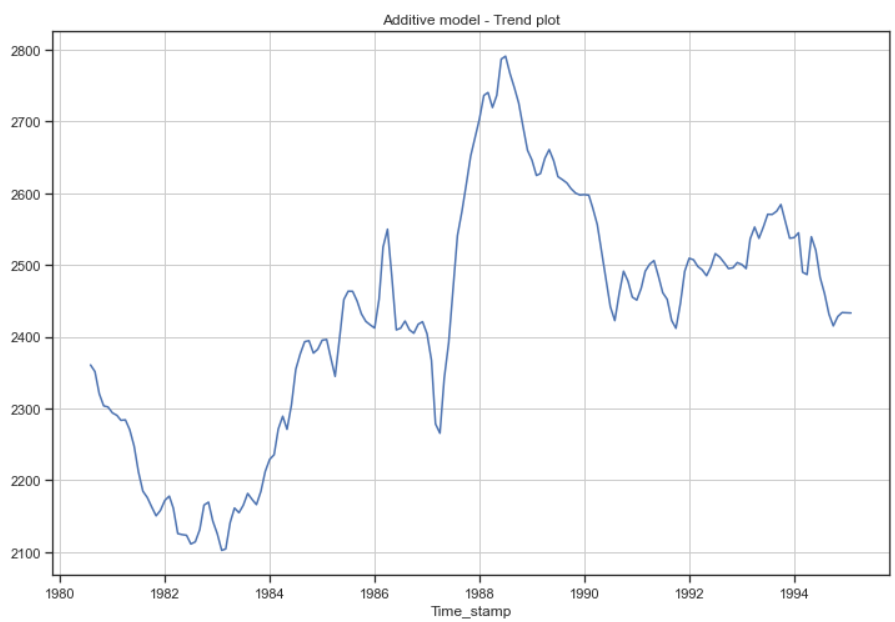


The trend is increasing from an all-time low in the year 1983 up to the year 1987. Between the years 1987 and the first quarter of 1988 the trend increases exponentially. The seasonality is constant over the years. The residual however, is distributed randomly away from the 0. The residual is highly scattered between the years 1985 and 1989. Head of the trend, seasonality and residual looks like this:

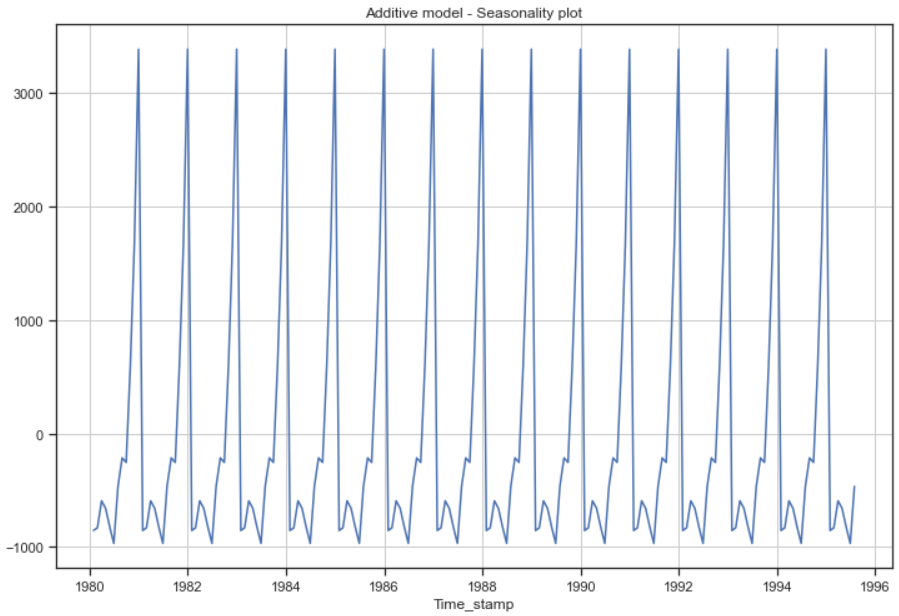




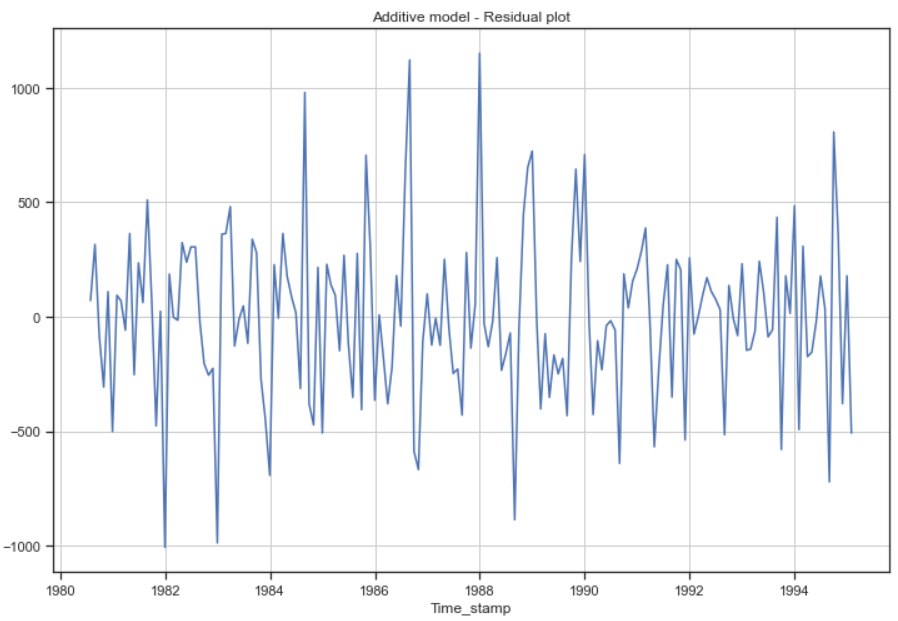
**Trend of the additive series:**



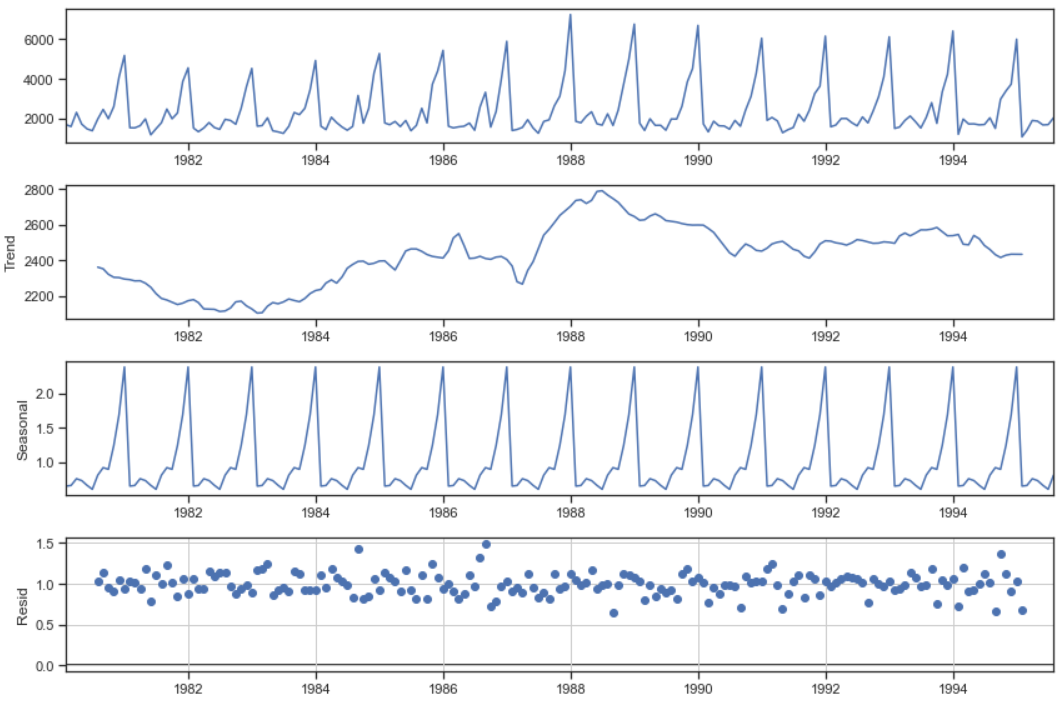
**Seasonality of the additive series:**



**Residual of the additive series:**



It looks like the residual of the additive series itself has a trend.

**Multiplicative decomposition:**

### In this the residual is centred around 1. Magnifying the trend, seasonal and residual components.

### The head of the multiplicative decomposition components: Trend, Seasonality and Residual looks like this:

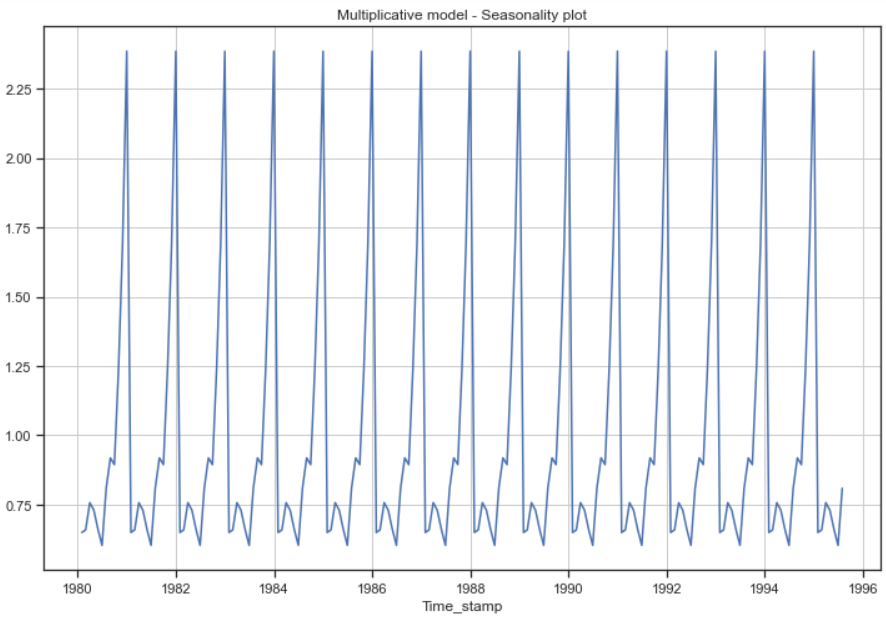
### 

### 

### Trend:

### 

**Seasonality:**



**Residual:**

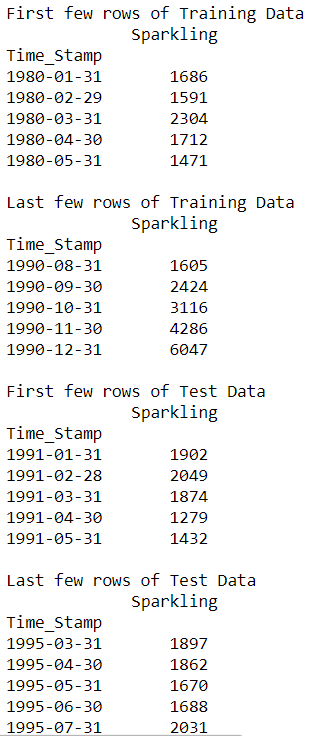


**Splitting the data into train and test:**

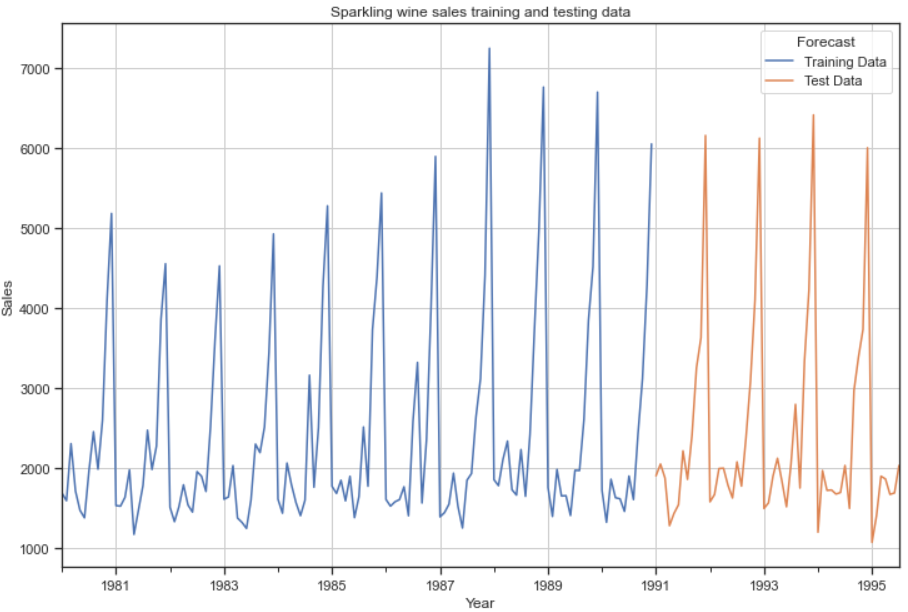
We have split the data into train and test such that the values before the year 1991 is considered as the train set and the values from the year 1991 is considered as the test set.

The shape of train and test sets look like this:

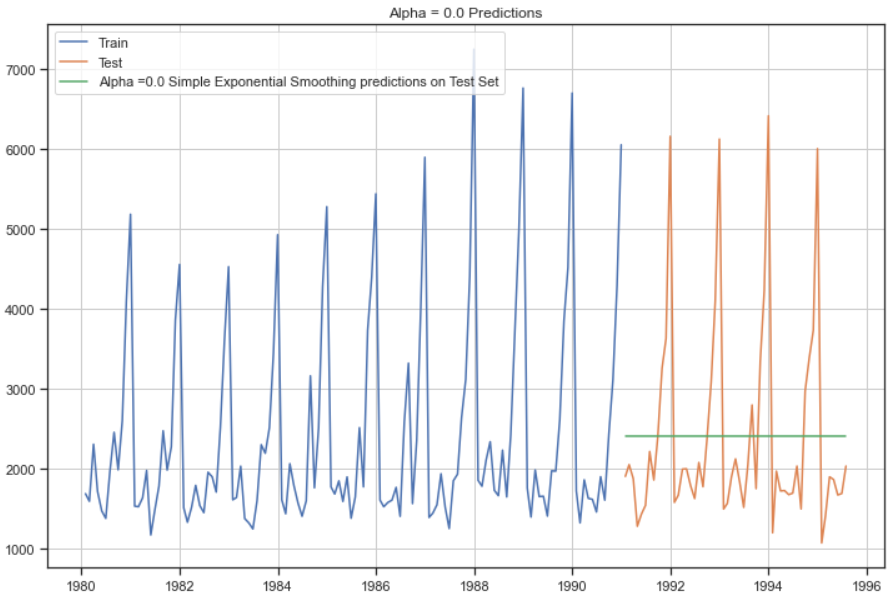




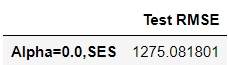
**Plot of Training and Testing data together:**



**Simple Exponential Smoothing model:**



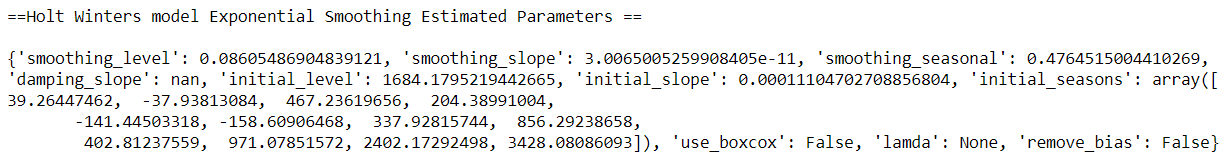
We see that the SES model gives us a straight line prediction. We check the RMSE value on the test data:



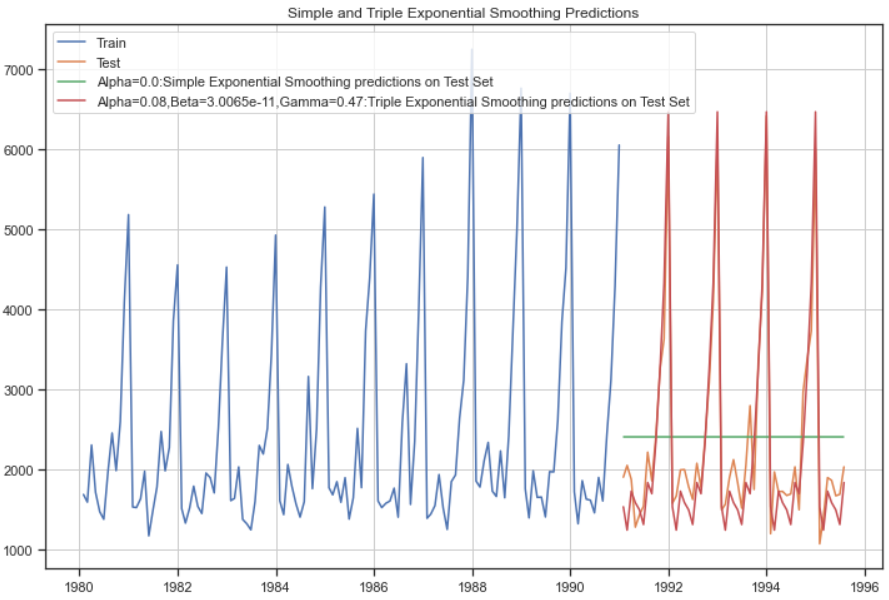
**Triple Exponential Smoothing model or Holt Winter's linear method:**

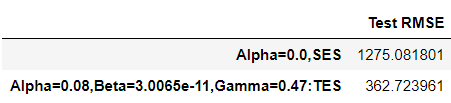
### ETS(A, A, A) : With additive trend and seasonality:

We get the following smoothing parameters when we fit the model into the training data:



We see that the Triple Exponential Smoothing is picking up the seasonal component as well.





### ETS(A, A, M) model : Additive trend and multiplicative seasonality

### We get the following smoothing parameters when we fit the model into the train data:

### 

### 

### 

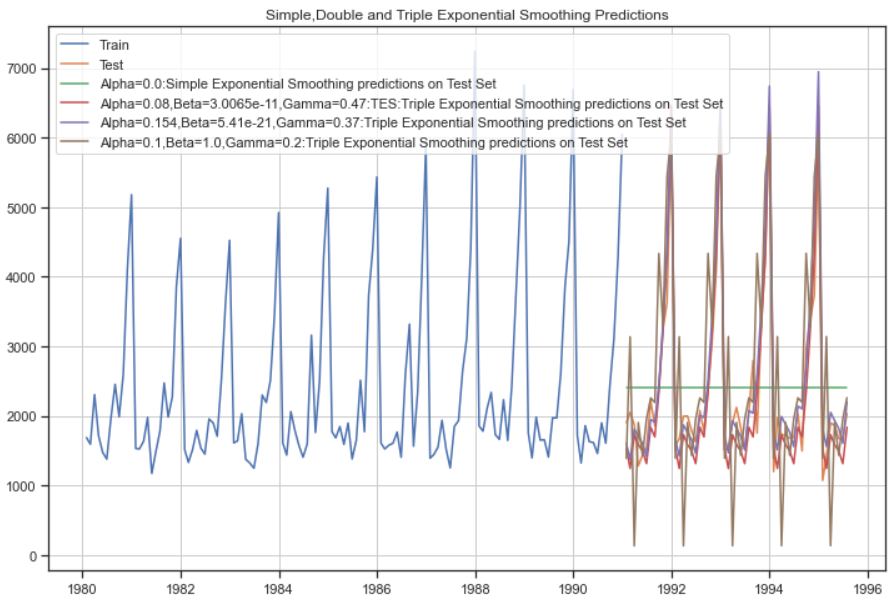
### Triple Exponential Smoothing model by taking the best alpha, beta and gamma [all in the range of 0.1 to 1 taking an interval of 0.1]

### We use the brute force in the TES model and also search the best value of the RMSE for different values of Alpha, Beta and Gamma in the given range.

### 

### We get the best values as alpha = 0.1, Beta = 1.0 and Gamma = 0.2

### 



### Evaluating most optimum model on the whole data:

### 

### Predicting 12 months into the future:

### 

### 

### 

### Linear Regression model: We fit the train time instance and test time instance into our model

### 

### 

### 

### Again, this model fails miserably on the prediction of the test data.

### Naive approach model:

### 

### 

### The Naïve model gives us a Test RMSE value of 3864.279 that is higher than any of the previous models so far.

### 

### This model as well, fails on the visual forecast of the test data.

### Simple Average model:

### 

### 

### 

### The Simple Average model fails on the Test RMSE and on the visual forecast of the test data.

### Checking for Stationarity of the whole time series data:

### 

### The test for stationarity is performed using the Dickey Fuller’s test:

### 

### 

### 

### 

### Automated version of an SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) and for a seasonality as 6:

### The loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2. We have kept the value of d as 1,2 as we need to take a difference once of the series to make it stationary. The seasonal differencing 'D' will be between 0,1 and 2 to check if it is needed or not.

### The best combinations are listed as follows with the lowest AIC – Akaike Information Criteria

### 

### We run the SARIMAX model with the train data, order as (1,2,3) , seasonal order as (1,2,3,6), enforce stationarity as false as the stationary series is already being considered in the parameters entered in the order, enforce invertibility as False as the effect on the y variable alone is required and not the inverse of it on the X variable (time variable). We choose the maximum iterations as 1000. The SARIMAX results are as follows:

### 

### Based on the p values below 0.05, we can see that the customer count Autoregressive model of order 1 is important. The seasonal moving average of second order is shown to be important. The AIC is 1459.793. The residual error of sihma2 is also shown to be important.

### Diagnostic plot:

### 

### Prediction on the test set with mean values and lower confidence intervals and upper confidence intervals:

### 

### 

### The test RMSE for this model is 2183.422.

### ACF and PACF on the time series data:

### 

### 

### 

### 

By looking at the above plots, we can say the following:

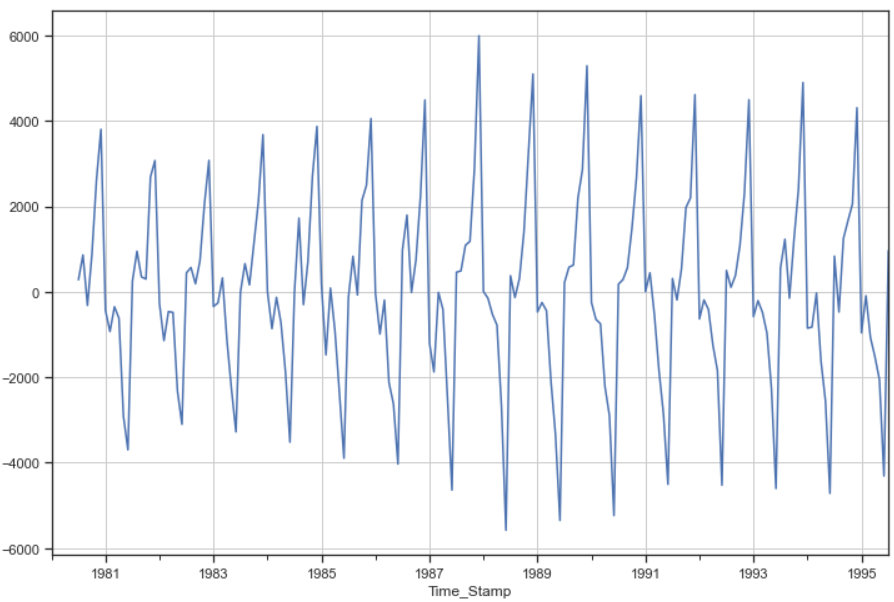
1. The ACF plot cuts off at lag 2. i.e. The Moving-Average parameter in an ARIMA model, q = 2
2. The PACF plot cuts off at lag 3. i.e. The Auto-Regressive parameter in an ARIMA mode, p = 4
3. We can also see there is a seasonality present and can be considered as either 6 or 12.

Let us plot the original data and the differenced series:



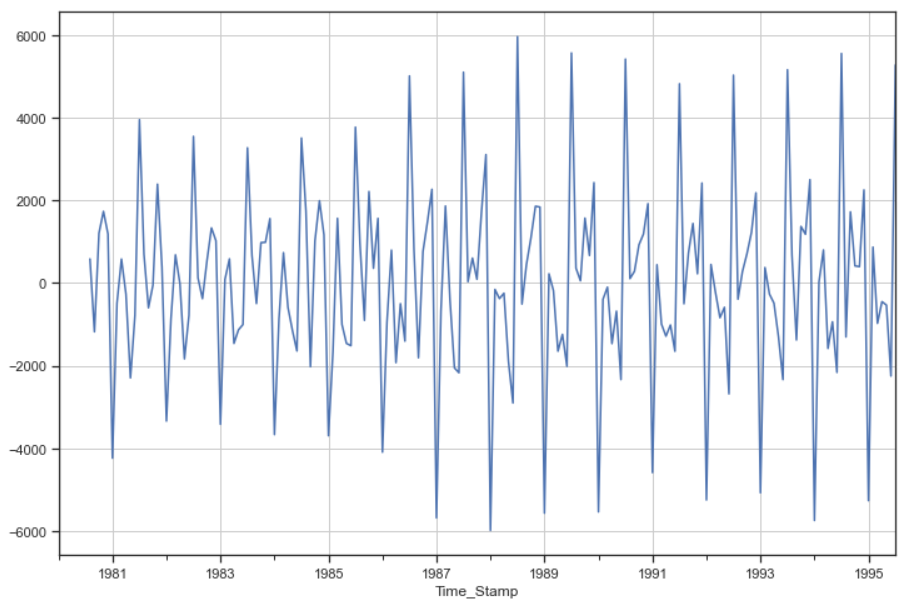
In the above plot, we can see there is both trend and seasonality.

Let us take a seasonal differencing and look at the series:

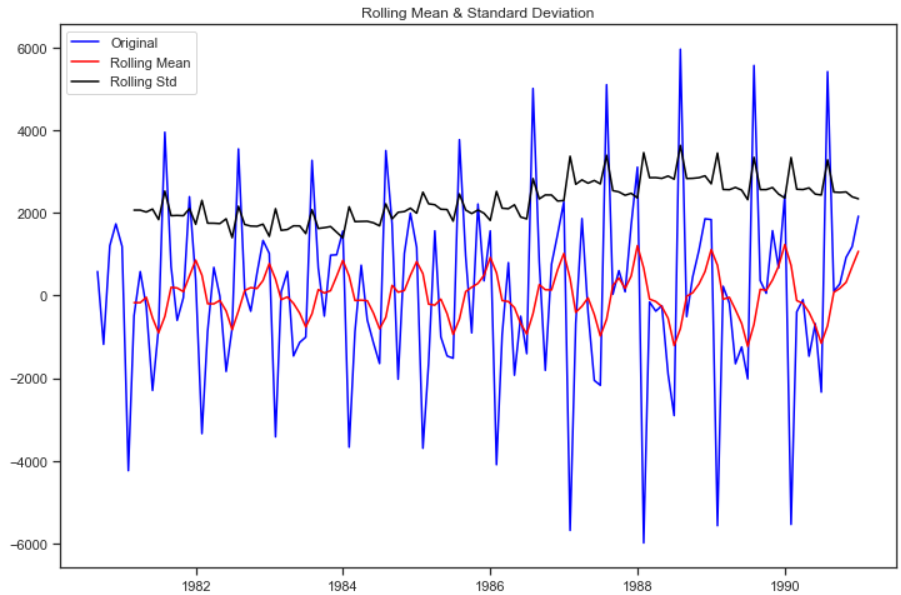


We can see there is a slight trend present in the data.

So, we take a differencing of first order on the seasonally differenced series.

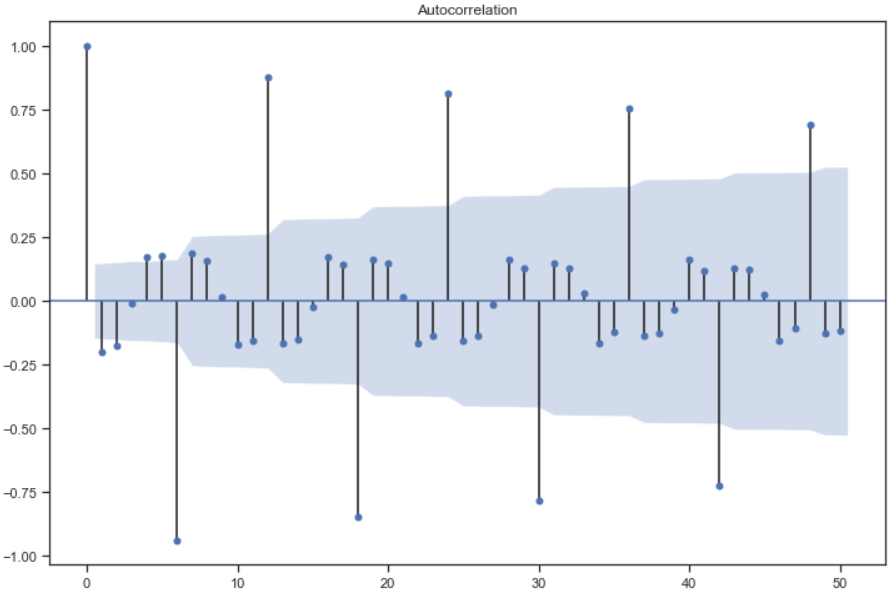


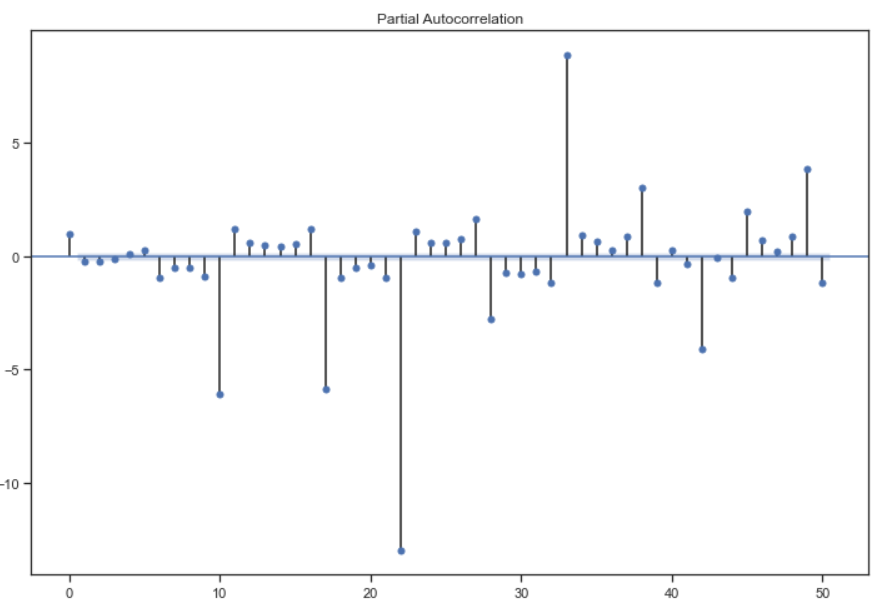
Checking for the stationarity of the above series:





Checking the new ACF and PACF plots for the new modified time series:

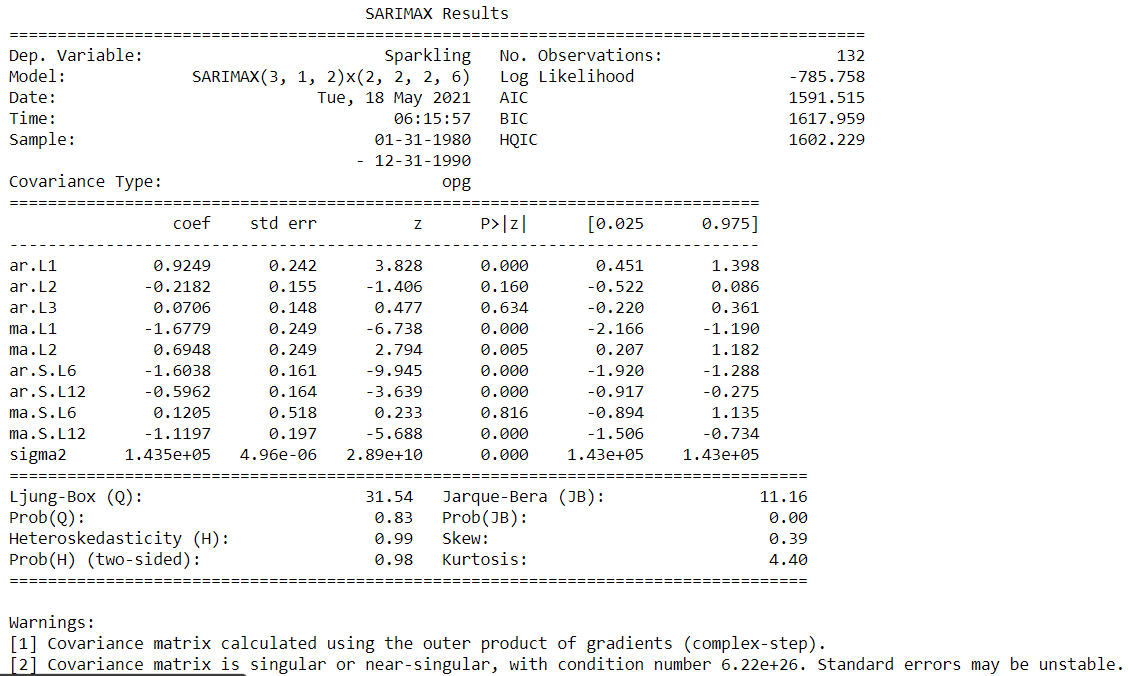




Here both P and Q are taken as 2 as the differenced data ACF and PACF plots indicate the same.

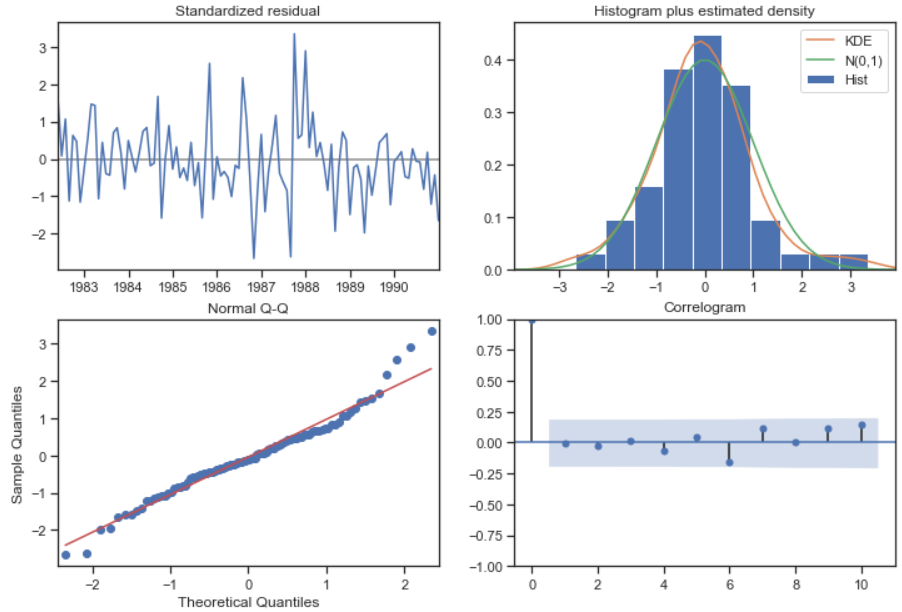
Seasonal differencing is taken as 2 as we have differenced the series twice.

q = 2 and p = 3 and d as 1



The AIC values have actually decreased as compared to the previous iteration. This time the Auto regressive 1 interval of period count and one interval of moving average count is important. So is two intervals of moving average period count is important. The Auto Regressive seasonal count as 6 and 12 is important. So is the moving average seasonal count of 12. The residual sigma2 is also important.

Plot diagnostics:



### Predict on the Test Set using this model and evaluate the model:

### 

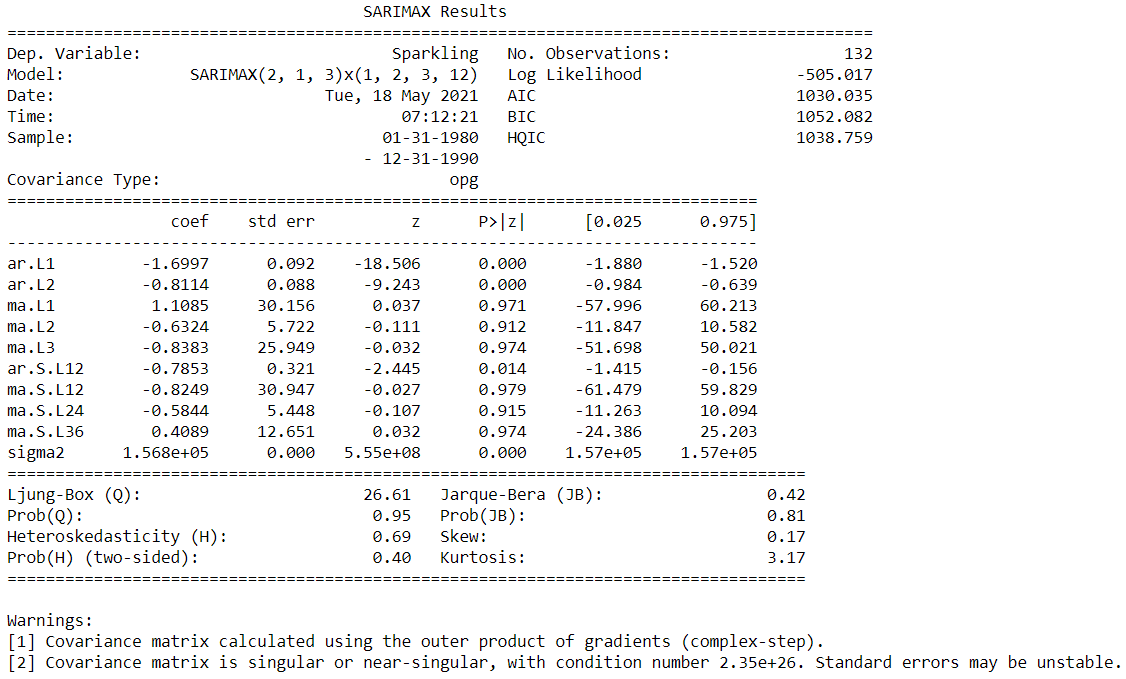
### Automated version of an SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) and for a seasonality as 12:

### The following loop helps us in getting a combination of different parameters of p and q in the range of 0,1,2,3. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary. The seasonal differencing 'D' will be between 0,1 and 2.

### 

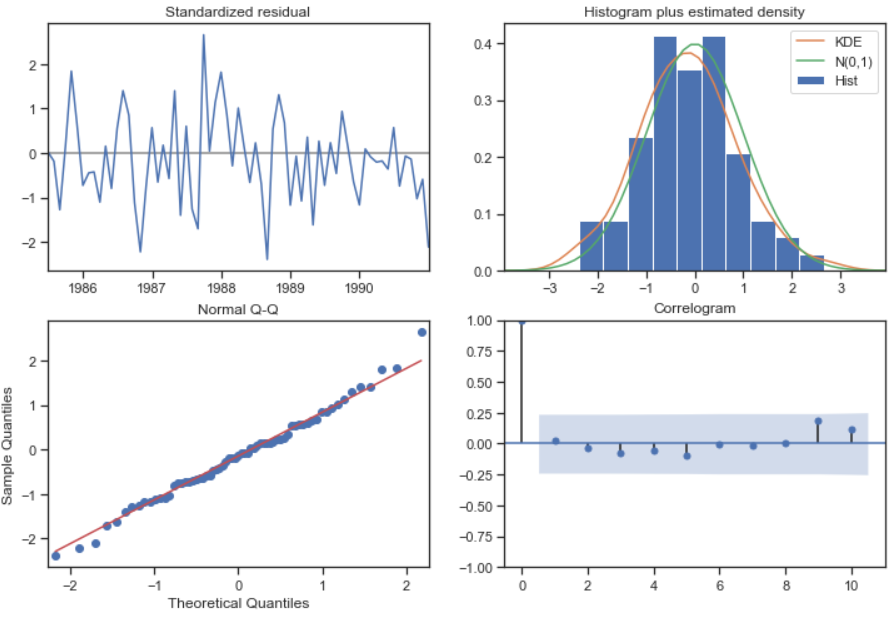
We then list the best combinations of the parameters based on the lowest AIC:



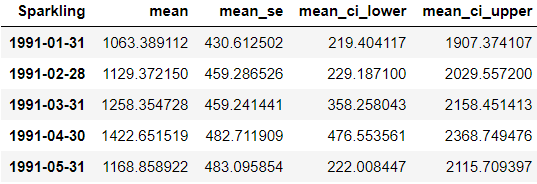


Predict on the Test Set using this model and evaluate the model:

Plot diagnostics:



The head of the mean predicted values, the upper confidence intervals and the lower confidence intervals are as follows:





This is the lowest RMSE we have achieved so far.

### ACF and PACF plot of the time series data with the seasonality as 12:

### Seasonally differenced time series:

### 

### We can see there is a slight trend present in the data.

### So, we take a differencing of first order on the seasonally differenced series.

### 

### Checking for the stationarity of the above series:

### 

### 

### Checking the new ACF and PACF plots for the new modified time series:

### 

### 

### We run the SARIMAX model with the following parameters: Here, P = 2 and Q = 1 , p = 4 and q = 1

### 

### The AIC is 1706.423 in this combination which is higher than that of the previous model. In this model, the moving average period count and two intervals of period count are shown to be important. The Auto regressive seasonal one period count and two intervals of period count are shown to be important. The sigma2 is also shown to be important in this case which is the estimate of the variance of the error term. Also, the Jarque-Bera (JB) test for the normality is shown, the hypothesis of the test is as follows:

### JB(P Value>0.05)= Accept H0 (Normal Distribution) – Null Hypothesis

### JB(P Value<0.05)= Reject H0 (Non-Normal Distribution) – Alternate Hypothesis

### In this case the value obtained from the above model is 0 which is corresponding to the non-Normal distribution.

### The Ljung-Box (Q) test for the independence of residuals.

### LB(P value > 0.05) = Accept H0 ( Residuals are independant) – Null Hypothesis

### LB (P value < 0.05) = Reject H0 (Residuals are not independent) – Alternate hypothesis

### In our case the value obtained is more than 0.05 and thus suggests that the residuals are independent.

### The test for Heteroskedasticity (H):

### H (P value > 0.05) = Accept H0 (The null hypothesis states that the error does not form a relationship with the values of x or with the predicted values of y.)

### H (P value < 0.05) = Reject H0 (The Alternate hypothesis states that the error does form a relationship with the values of x or with the predicted values of y)

### In our case the p value obtained is 0.02 so according to the model we assume that the error does form a relationship with the predicted values of y. This is also in line with the presence of sigma2 being shown as an important variable in predicting the y values.

### Plot diagnostics:

### 

### Standardized residual: Residual mean is not so constant. It should be around 0. But as we had observed before the residual component in this data is pretty high so this is okay.

### Histogram plus estimated density: Residuals have an almost perfectly normal distribution and so this is okay.

### Normal Q-Q: All the data points should fall in the line of the Q-Q. They should be linear. The above plot satisfies the beforementioned conditions and hence looks okay.

### Correlogram: The residuals themselves should not have any autocorrelation. They should not have any significant lags. As observed, they do not.

### If any of this is violated, then we cannot claim that the point forecast: upper confidence level and lower confidence level is true 95% of the time.

### Predict on the Test Set using this model and evaluate the model

### 

### 

### Building the secondmost optimum model on the full data:

### 

### The AIC values have shot up to 1896 when forecasting for 12 months into the future. The Autoregressive period count 1 and 2, Moving average period counts 1, 2 and 3 and the residual sigma2 is also important.

### Plot diagnostics:

### 

### Evaluate the model on the whole and predict 12 months into the future:

### Following is the head of the forecasted values with the mean values, mean standard error, mean lower CI and mean upper CI values:

### 

### 

### The forecast of the model for 12 months into the future along with the confidence intervals are shown in the following plot as follows:

### 

### Table listing of all the models with their respective RMSE values on the test data:

### 

**Observations and suggestions:**

* The probability of the Jarque-Bera test is 0.0 thus indicating that the distribution is normal.
* The standardized residual plot has a high error variation. The histogram plus estimated density shows a perfectly normal distribution. The normal Q-Q plot is okay as all the data points are lying in the line of the Q-Q. They are linear. The correlogram shows us that the residuals do not have any autocorrelation. They do not have any significant lags.
* The skewness is 0 and the Kurtosis is 3 for a normal distribution. In our data, the skewness is observed to be 0.21 and Kurtosis is 4.50. Sometimes, the values of these tests do not correspond to the results thus obtained.
* Prob(H) (Two-sided) Heteroskedasticity is 0.02 and hence there is heteroskedasticity present in the time series. This means that the residuals are not distributed with equal variance. The null hypothesis of this test shows that there is no heteroskedasticity and the alternate hypothesis shows us that there is. At 95% confidence level, we can say that there is heteroskedasticity present. This does not come as a surprise as the sigma2 value is shown to be important with p value of 0 and thus does indicate that the variance of the error term is significant. This means that the model thus built is able to explain some pattern in the response ‘y’ variable which in this case is our Sparkling wine sales. Reference: Datascienceplus, towardssciencedirect.
* The inferences from the EDA have been mentioned previously in the appropriate sections and are to be taken a closer look at to identify any reason for the spike in sales. If possible, the same methods can be adopted with a view to increase the sales.
* This calls for careful investigation on the periods where the spike in the Sparkling wine sales was observed. The spike could have been the result of any one or many of the following factors, which is to be considered in no particular order: The composition of the Sparkling wine during that period, the offer price with respect to the economic conditions prevailing during that period, the promotional strategies during that period, the location of the stores during that period and the sales in each of those stores, the medium of advertising and the ambassador for endorsements, government policies during that period, the management team during that period, etc.
* The Sparkling wine needs to be offered at a discounted price especially during the last quarter of every year to increase sales even more than what they already are. Another way to look at this would be to increase the Sparkling wine price during the fourth quarter of every year as there is surely an increase in sales during this period, thus the income would increase.
* Some important variables to be included in such an analysis are: Location of the store, sales with respect to each location / region / store. The economic condition of the people in that location also need to be analysed for drawing out a pattern.