|  |
| --- |
| Time Series Forecasting - Group Assignment |
| **Concepts covered – Holt Winter and S/ARIMA** |

|  |  |
| --- | --- |
| **Group No:** | 3 |
| **Team Members** | 1. Divyalakshme Arangasubramanian 2. Meenu Sethuraman 3. Priya Srinivasan 4. Ravichandran Narayanan 5. Shiyam Sundar T K 6. Subashree Bhaskaran |

***Question:***

Quarterly beer sales data has been provided in the [beer.csv](https://olympus.greatlearning.in/courses/3768/files/352543/download?wrap=1)files.

**Part A)**

Using the Holt-Winters method, model the data and predict the sales for the next 2 years. Your submission should contain the complete modelling steps with explanations. Include pictures and R-code where applicable.

**Part B)**

Using the ARIMA method, model the data and predict the sales for the next 2 years. Your submissions should contain the complete modelling steps with explanations. Include pictures and R-code where applicable.

Dataset**:**[**beer.csv**](https://olympus.greatlearning.in/courses/3768/files/352544/download?wrap=1)

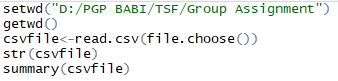
Info on Data:

**There are 72 observations and one variable names OzBeer. It is quarterly data. So, 72 observations mean 72/4=18 years data. OzBeer variable signifies the sales revenue in a thousand dollars. Here the data period is not known. Assume the data is for the past 18 years.**

**Solution:**

**Step1: Check the given problem statement read the CSV file into R.**

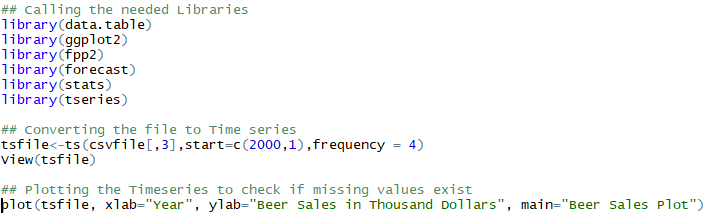
The csv file shared didn’t have time period reference. Hence for readability and referencing purpose, we randomly time tagged the beer sales values with starting period as 2000. The same is attached herewith for reference. This was done manually in Excel. Then read the modified CSV file into R. 



**Step2: Convert the data into Time Series format and check for missing data:**

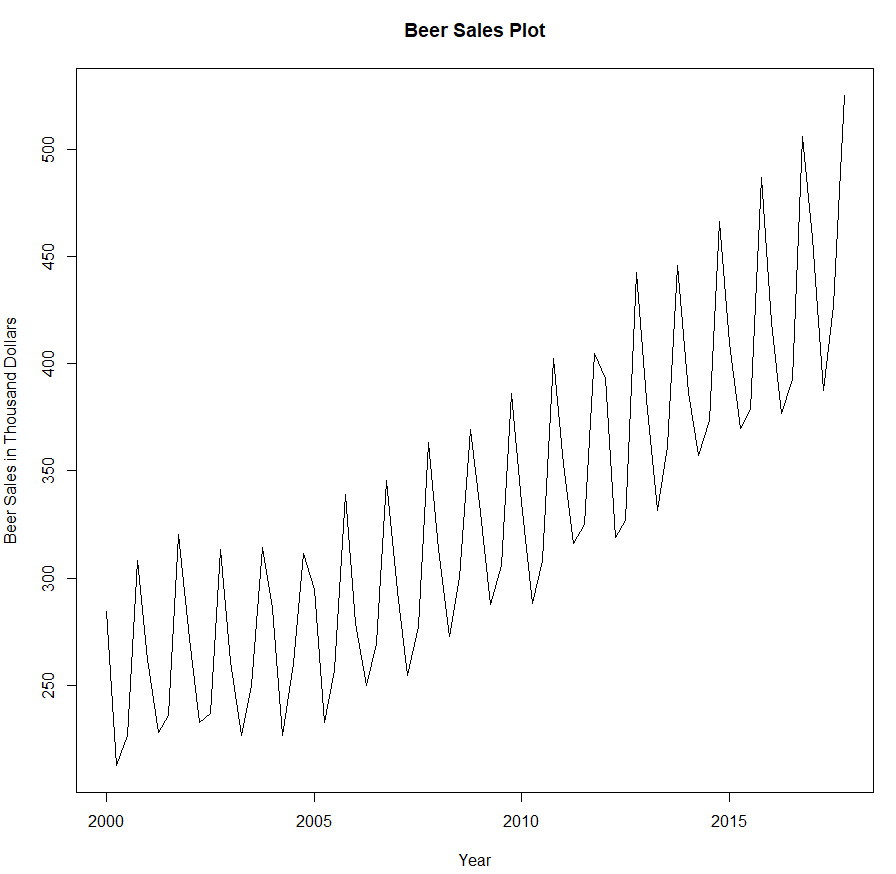
Convert the data into time series format and plot the time series to check if there exists any missing data for any time period.

Before doing this, load the necessary libraries needed for time series analysis.





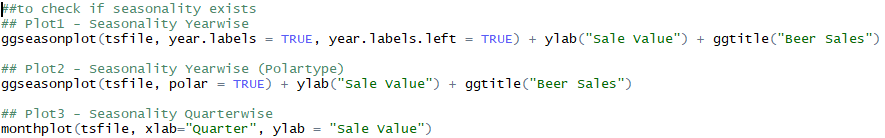
Sample output of time series format file is shown above for reference.



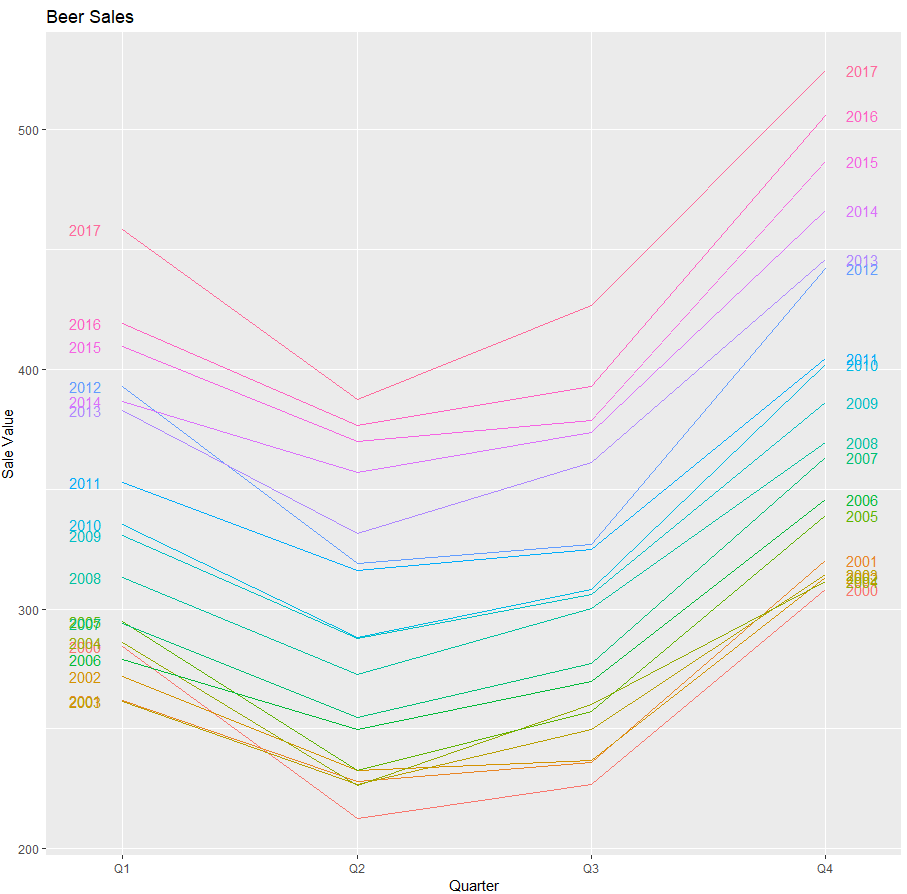
It can be observed from the above plot, data provided is intact (i.e no missing data exists) and the time series is ***additive******in nature***with the presence of trend and seasonality in it. The reason for calling the series as additive because based on above plot, it can be seen that the hikes in sales with respect to each successive year, tends to follow a common pattern, meaning, the hikes are uniform in nature (Following + / - format). Had it been multiplicative, the hikes would have been increasing / decreasing with each year rather following a uniform format.

**Step3: Identify seasonality pattern:**

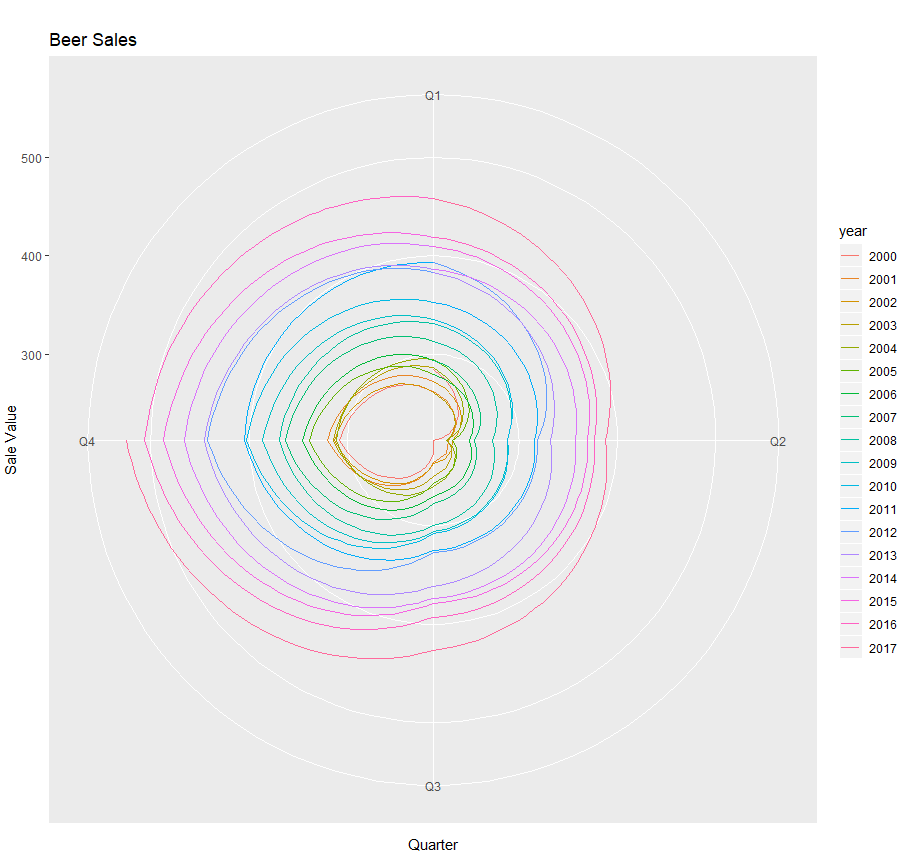
Plot the data quarter wise to understand the seasonality pattern.



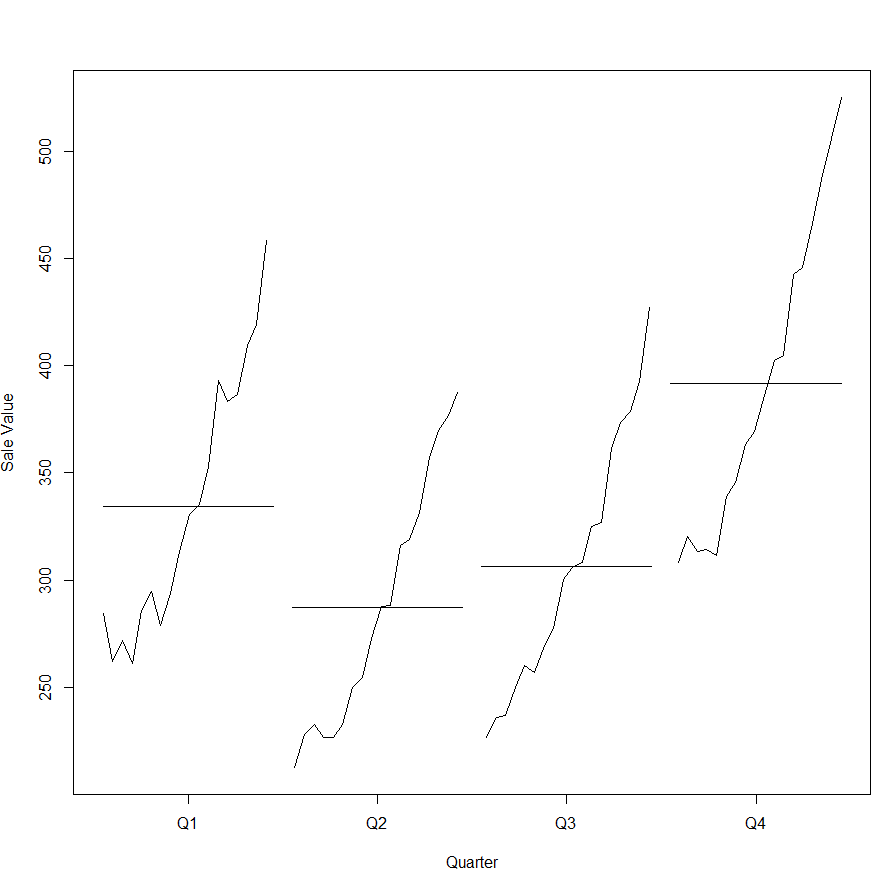
***Plot 1:***



***Plot 2:***

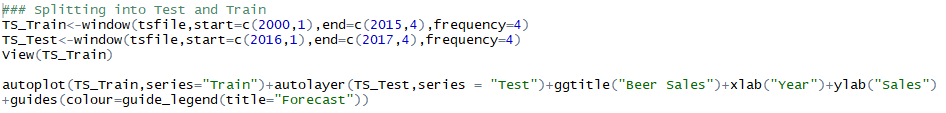


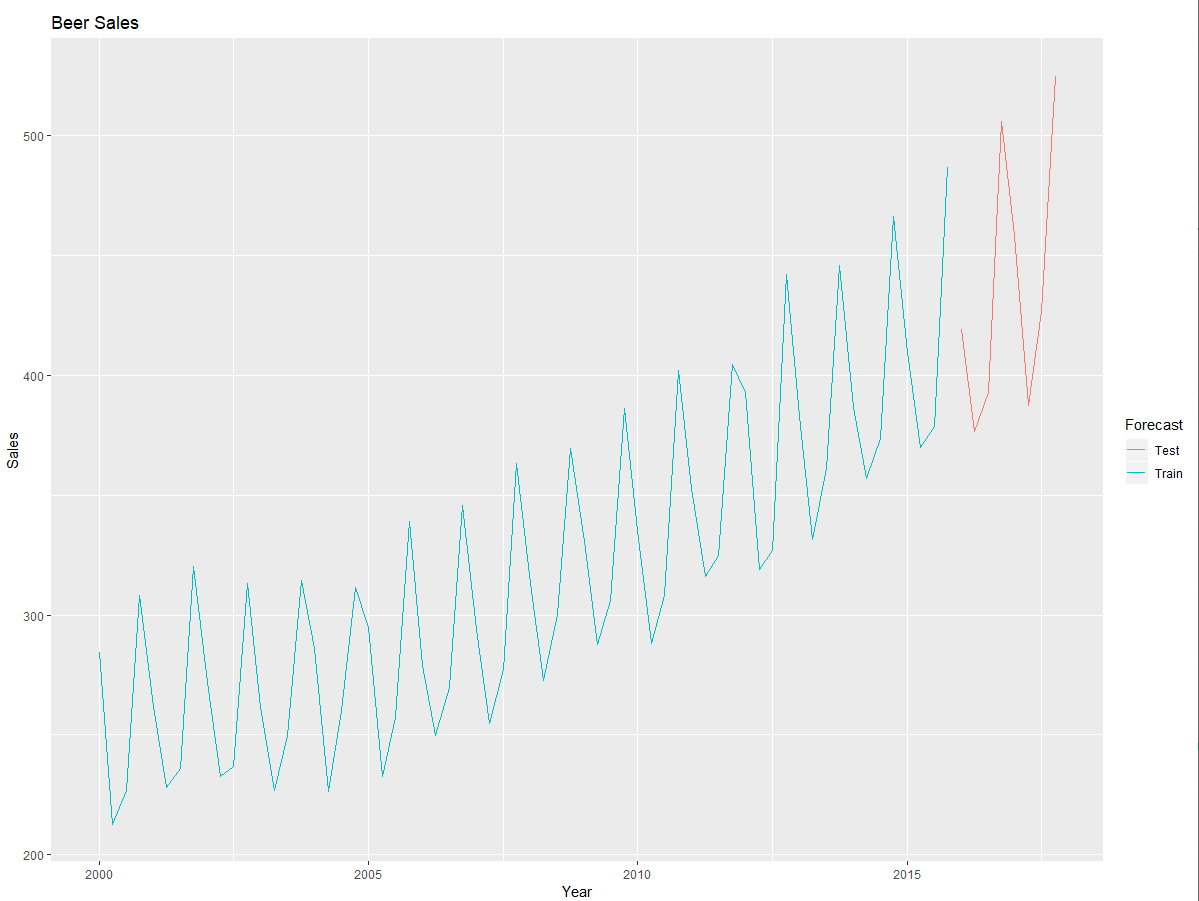
***Plot 3:***



From the above plots, it can inferred that seasonality exists between each quarter and the sales tend to lower at 2nd quarter and regains maximum sales at 4th quarter. This pattern applies to all years.

**Step4: Split the data into Train / Test data sets:**

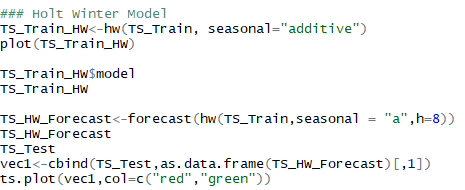


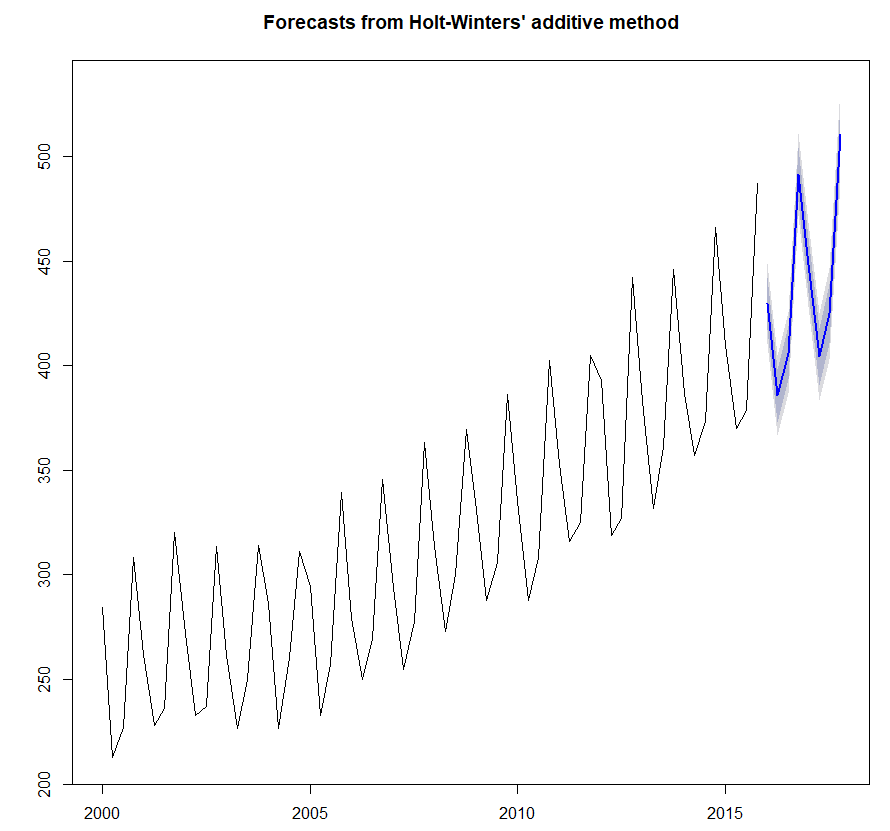


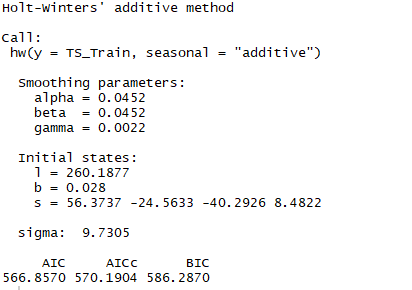
The train data is accounted from Quarter 1 of year 2000 till Quarter 4 of year 2015 and for test data, it is from Quarter 1 of year 2016 till Quarter 4 of year 2017. The same are plotted for reference.

**Step5: Build the Holt-Winter Model (Triple Exponential Smoothing):**

As stated in the problem as well based on the inference from the data set (the presence of level, trend and seasonality) we will go forward with Triple Exponential Smoothing aka Holt Winter Model.



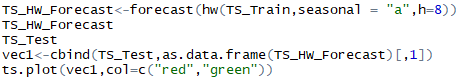


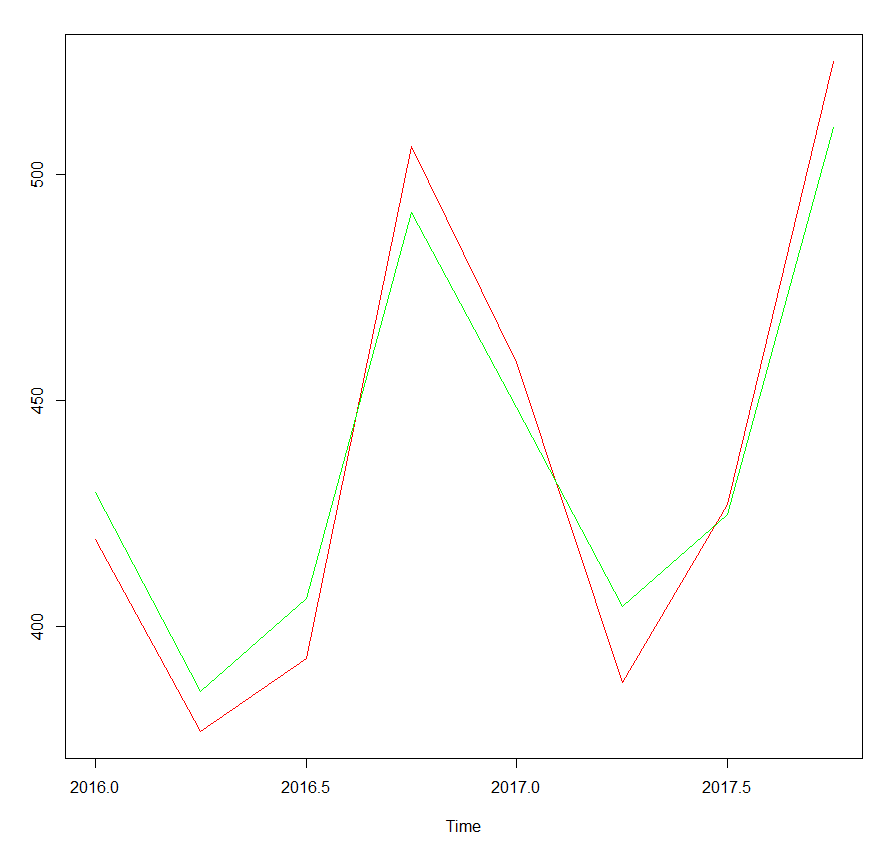


It can be observed from the above result, for level and trend, the model has chosen the value of alpha, beta as 0.0452, which signifies neither under smoothing or over smoothing. But for seasonality, the gamma value of 0.0022 is almost close to 0 which signifies under smoothing. Meaning, the model has correctly identified the variance in the seasonality pattern and chose to build a cumulative figure with most weightage given to recent observations.

**Step6: Forecast using the above model and check with test data:**

Forecast using the above model for another 8 years and check with the test data values and see how the model performs.



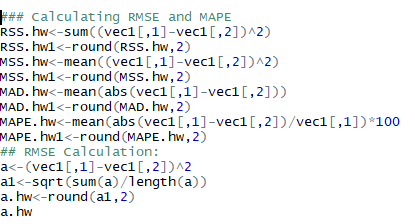


The time series plot shows the actual test values in red and the forecasted values in green. It can be observed that the model performs reasonably well. Meaning, the forecasts more or less follows the test data pattern except for some slight variations.

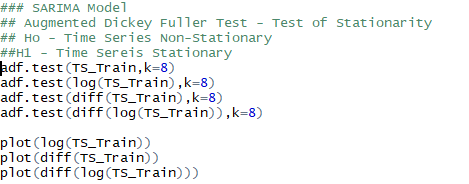
**Step7: Calculate the forecast accuracy measures:**

Following forecast accuracy measures are computed:

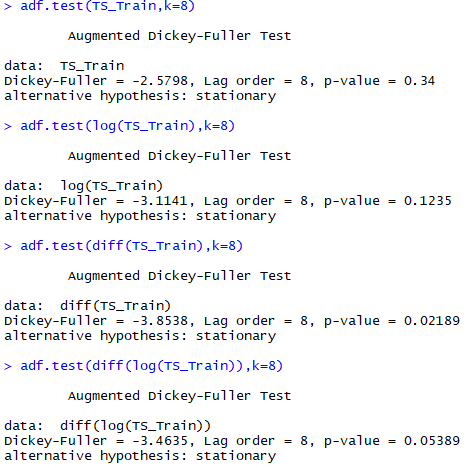
1. Residual Sum of Squares – 1178.66
2. Mean Residual Sum of Squares – 147.33
3. Mean Absolute Deviation – 11.34
4. Mean Absolute Percentage Error – 2.61%
5. Root Mean Square Error – 12.13



**Step8: Build SARIMA Model starting with Test of Stationarity (Augmented Dickey Fuller Test):**

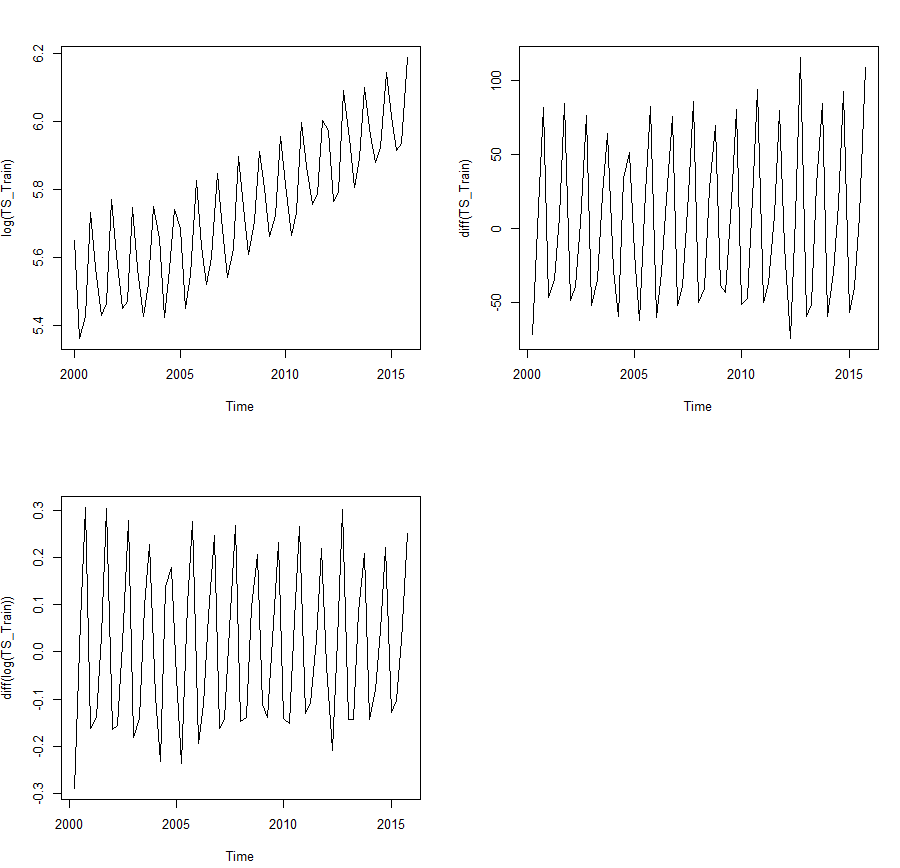


adf.test is done for the original series, log of original series, difference of original series and difference of log original series. Following are results:



It can be concluded from the above test result that difference of the original series is stationary and the others tests results are non-stationary.

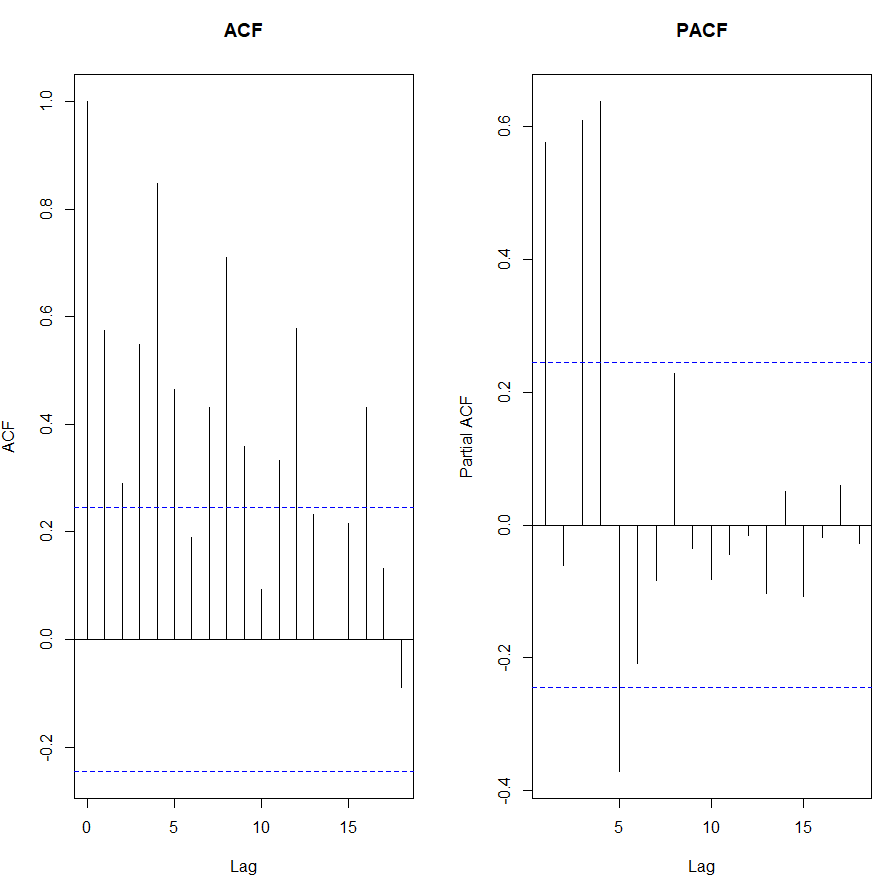
Plots of the above tests are shown below for reference:



**Step9: Analyse the ACF and PACF:**

Plot the Auto-correlation Functional plot to check for the presence of auto-correlation. As well plot the Partial Auto-correlation Function plot to roughly understand the number of lags to regress upon.



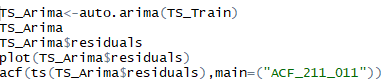


Presence of Auto-Correlation can be confirmed from the ACF plot, where the residuals are widely distributed outside the confidence interval with mean 0. From the PACF plot, it is seen that the second lag ends up being 0. Roughly thinking, the value of “p” in the (p,d,q) (P,D,Q)[4] can be taken as 2.

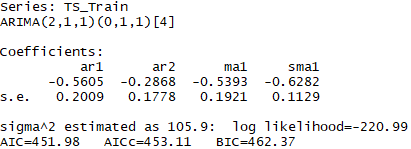
**Step10: Use R to build the initial SARIMA Model:**

The initial SARIMA model can be build using R, wherein a suitable model is automatically selected / built by R which takes into the account the following aspects:

1. No of lag to regress in the original series (“p”)
2. Level of difference in series to work with (“d”)
3. Extent of lags to be employed in Moving average computation (“q”)
4. No of lag to regress in the original series – applied only on seasonality (“P”)
5. Level of difference in seasonal series to work with (“D”)
6. Extent of lags to be employed in Moving average computation of seasonal series (“Q”)



Following is the model output from R:

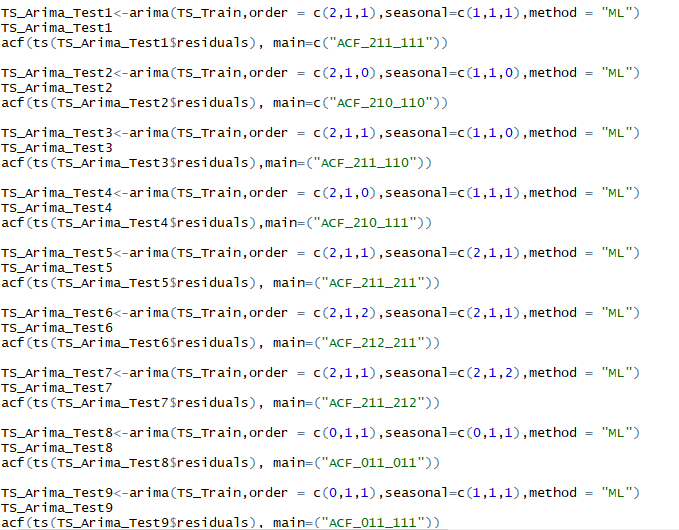


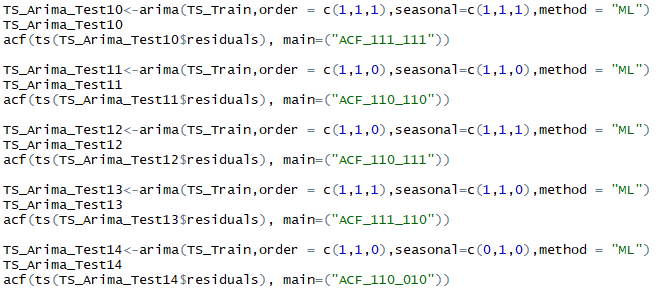
It can be seen that, the model has chosen a (2,1,1)(0,1,1)[4]. This exactly correlates with the ADF Test conducted in previous steps, where the difference of original series proved to be stationary and the same is reflected here.

**Step11: Check if the model can be fine-tuned:**

Iteratively, try different values of p,d,q and P,D,Q to check if a better model can be developed. This can be identified basis, the AIC and BIC values. If a new model renders the AIC and BIC values below 451.98 and 462.37, the it will be the best model for forecasting.

Different models try outs are shown below for reference:





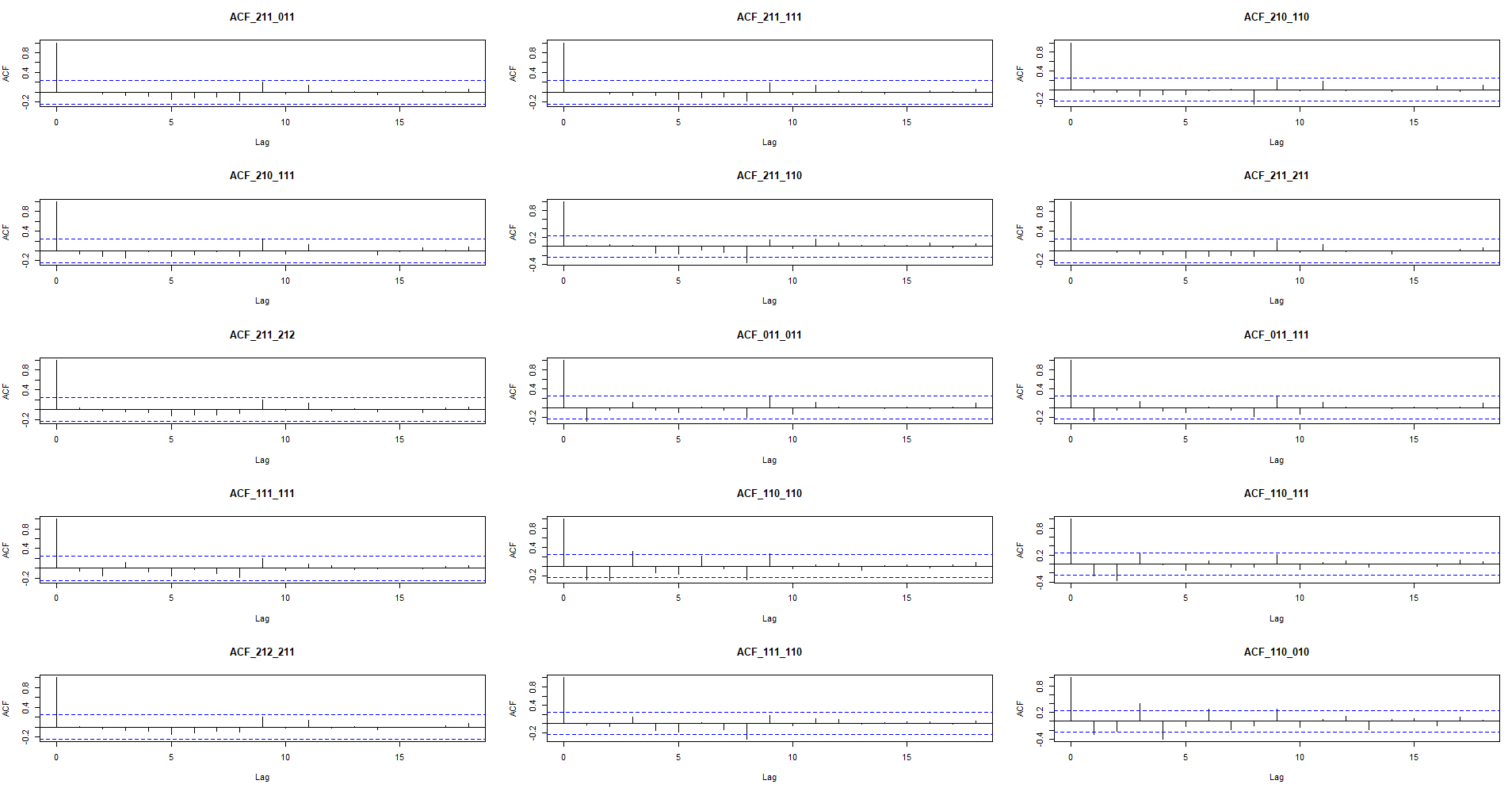
In total, 14 different models have been tried with the condition that “p+P” & “q+Q” not to exceed 3.

The AIC and residuals of different models are compressed onto two data frames and the same are attached herewith for easy reference:



The ACF plots of the different try out models is attached as in the image form for easy reference.

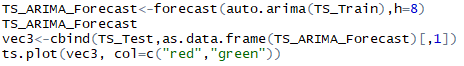


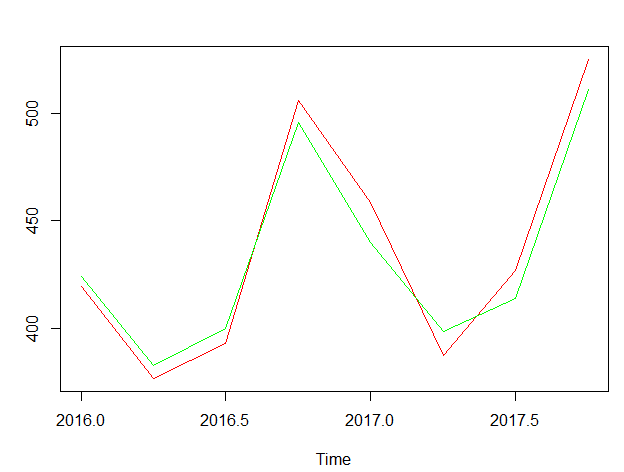


From the above plot, it is evident, for most of the models, at least for one of the lags, the ACF is well outside the confidence band (meaning, they are non-zero). But SARIMA Model (2,1,1)(0,1,1) has an ACF of zero. Meaning, no auto-correlation exists.

**Step12: Forecast using the best SARIMA Model and check with Test data values:**

Forecast using the initial SARIMA Model (2,1,1)(0,1,1) which proved to be best of all the models prepared.



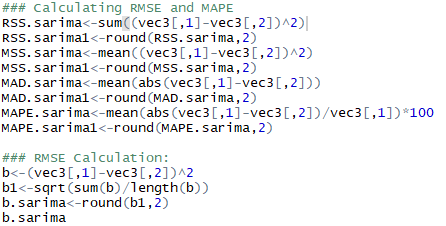


It can be seen from the above time series plot, the forecasted values (green colour) is almost matching with the test values (red colour).

**Step13: Calculate the forecast accuracy measures:**

Following forecast accuracy measures are computed:

1. Residual Sum of Squares – 1042.65
2. Mean Residual Sum of Squares – 130.33
3. Mean Absolute Deviation – 10.57
4. Mean Absolute Percentage Error – 2.39%
5. Root Mean Square Error – 11.42



**Step14: Compare the forecast accuracy measures of both Holt-Winter and SARIMA model:**

For easy reference, the forecast accuracy measures of both models are shown next to each other.



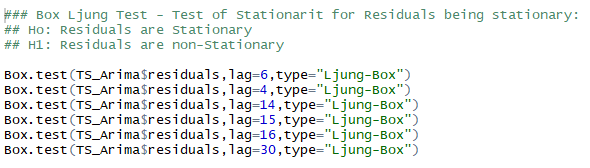


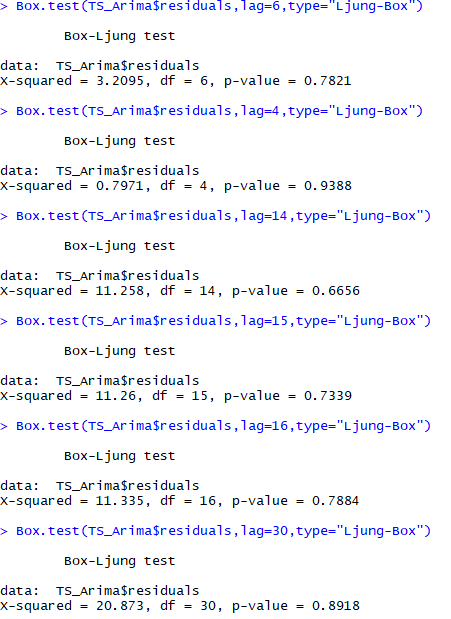
Since, SARIMA model has error parameters lower than Holt-Winter model, it can be inferred that, SARIMA model better performs when compared to Holt-Winter model.

Hence for forecasting purpose, SARIMA Model should be used.

**Step14: Check for Stationarity in residuals:**

As a final check, test the residuals of the (2,1,1)(0,1,1)[4] SARIMA model for stationarity. Use the Box Ljung test at different lags of the model.





It can be inferred from the above results that the residuals are indeed stationary. Hence, the SARIMA Model (2,1,1)(0,1,1)[4] is good for forecasting purpose.

**Step15: Forecast for next two years:**

Using the above SARIMA Model, forecast for the next years of Beer sales. In below figure, the values highlighted in “Green Colour” is the forecasts for next two years. The figures mentioned are in Thousand Dollars.

