**Module 5 Project Intermediate Analytics**

**Author’s Name:** Group A(Shivani Sharma a d Rupali Somani)

**Course Number:** ALY6015

**Course Title:** Intermediate Analytics

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Joseph

**Assignment Completion Date:** 12-08-2019

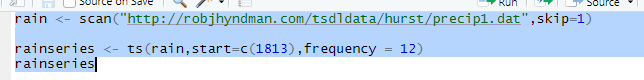


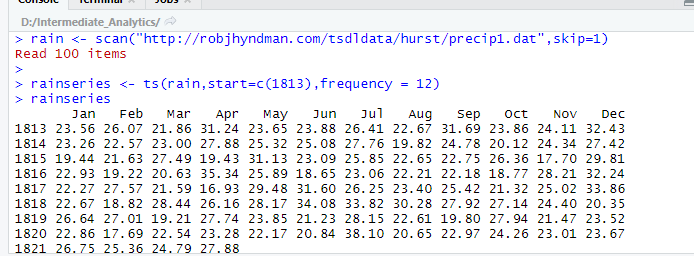
**Introduction**

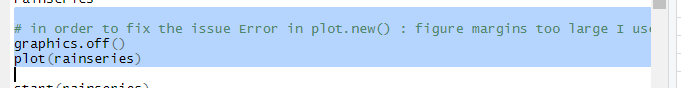
This project is based on time series analysis. I would like to give a brief about time series analysis. In layman’s term any metric which is measured for a continuous interval of time then it forms a time series. In order to explain the complete logic behind time series analysis I have selected a data set having yearly rainfall (in inches) statistics for a country from 1813-1912 for Part A of the assignment. And selected a data asset for Air passengers in order to explain the ARIMA model in part B of assignment. I would like to explain all the steps of the time series i.e reading the data, plotting, decomposing, diagnosis and forecasting in analysis section with the code and output.

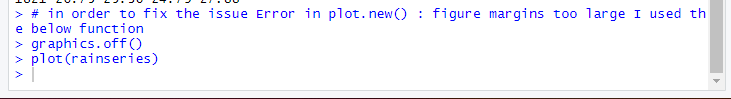
**Analysis**

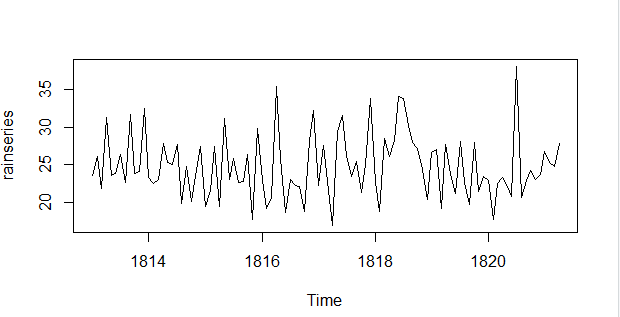
**Part A:** Now starting with the very first steps i.e to read the data of the data set and then feeding the dataset to the time series function in order to gain insights of time series pattern analysis of the rainfall data set and finally plotting the TS graph for visuals as visuals are always best to make someone understand the pattern. So, here is the logic and output.





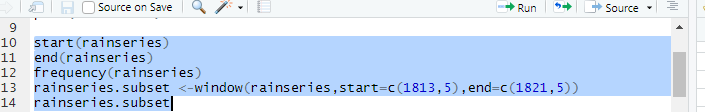


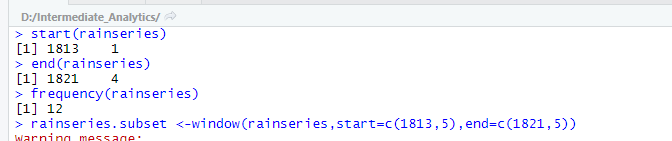


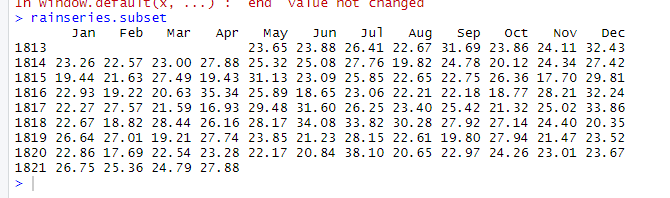


Here in the above code I have used a function graphics.off() before plotting the TS graph as I was facing a issue “**Error in plot.new() : figure margins too large**” for the plot function every time I was trying to plot a relation for a time series data set so in order to resolve the error I used this function.

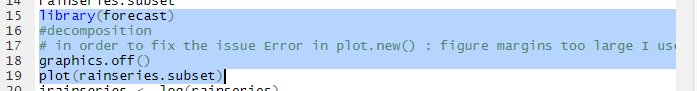
Now from the TS graph we can observe a approximately constant level for 25 inches as the mean value is looking constant too. Since there are random fluctuations so we need smoothening of the pattern. We will apply further calculations on the data now. First we need to find few parameters like start, end ,frequency etc

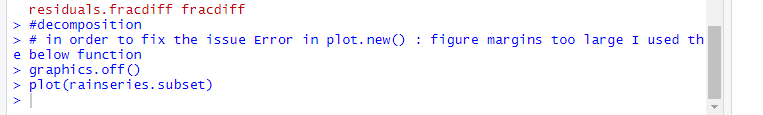


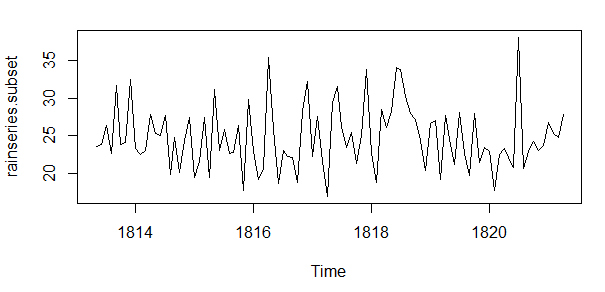




And then plotting it.

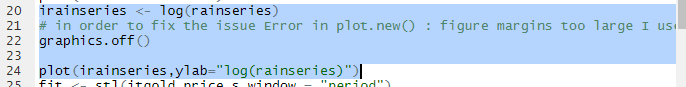




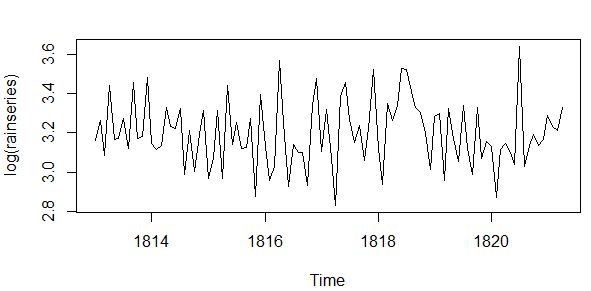


We can observe more clear plot in the subset graph.

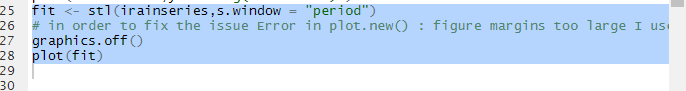
Now we are taking logarithmic of the subset data and plotting a TS for this in order to handle the skewness towards larger values

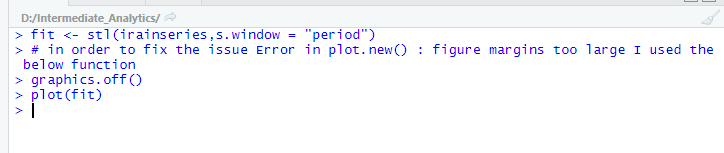


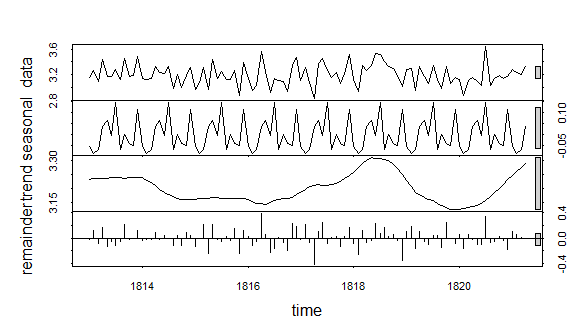




Now finally we will decompose the data into trend and seasonal data and before that we will try to fit the model using stl().

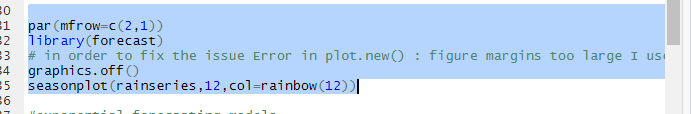


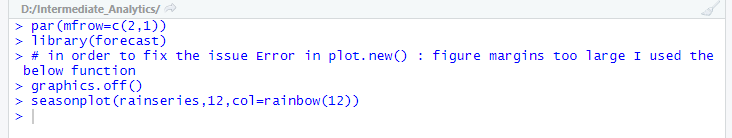


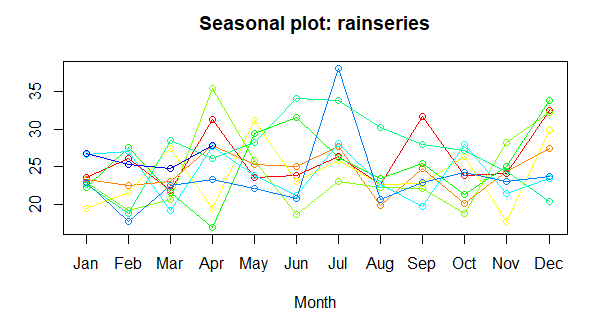


From the above plot we have a pattern for seasonal data , a trend for the behavior on the subset of data set after applying the logarithm**(though it is not needed but in order to skewness and observe the pattern we can do it additionally).**

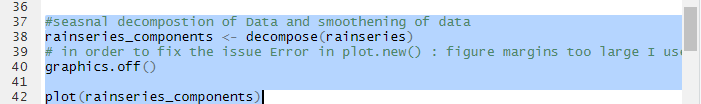
We can also observe the seasonal plot on logarithmic data as follows:

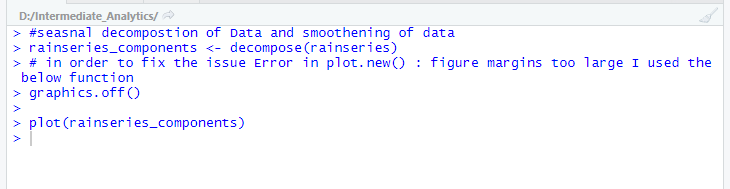


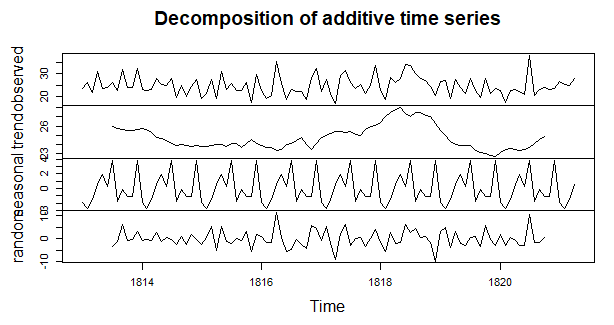




Now coming to the decomposition of the data set. Decomposition function-Time series decomposition is a procedure which creates multiple Time series out of a time series. The parent time series is split into three components such as seasonal, trend and random. Decomposition function is used to remove repeated pattern on fixed period of time from a time series.it helps us to understand trends. We can achieve it with following logic.

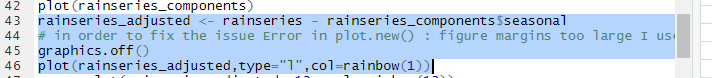


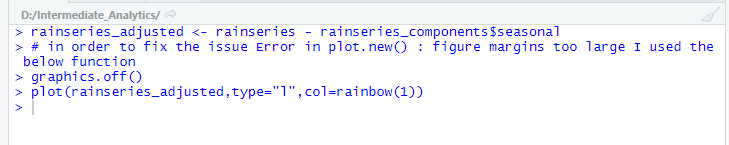


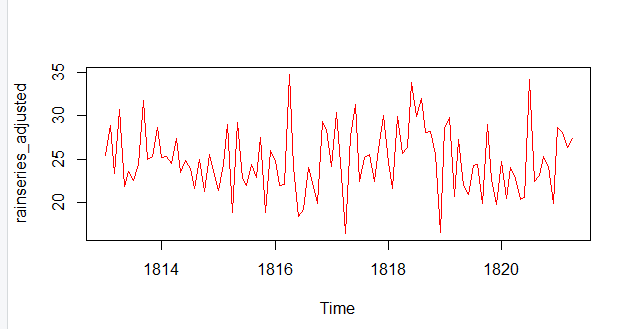


From the graph we can observe the pattern for observed data set which is based on the actual behavior of rainfall, a trending one which is based on historical data, seasonal one which is based on the measurement during seasonal time i.e rainy weather and the random one which is based on the randomly collected data.

Now we will calculate the adjusted value by subtracting components from the complete data set and then we can plot it.





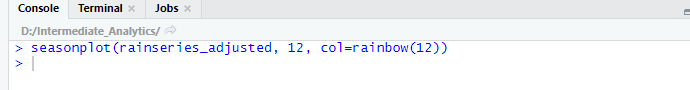


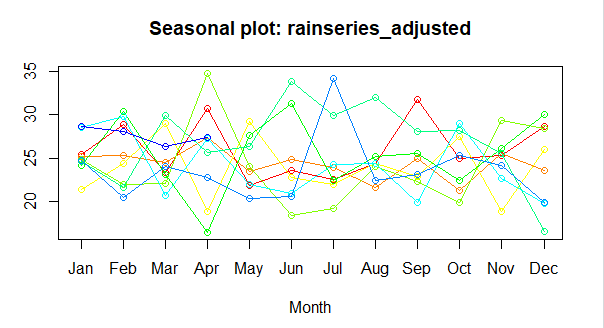
And above one is the plot for the adjusted values after decomposition

From the above plot we can see more clear insights of the rainfall now as compared to the one created with original data set and observes a more clear pattern of rainfall for 25 inches.

We can also observe the seasonal plot based out of seasonal data.

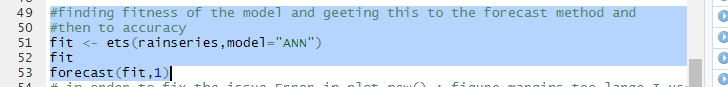


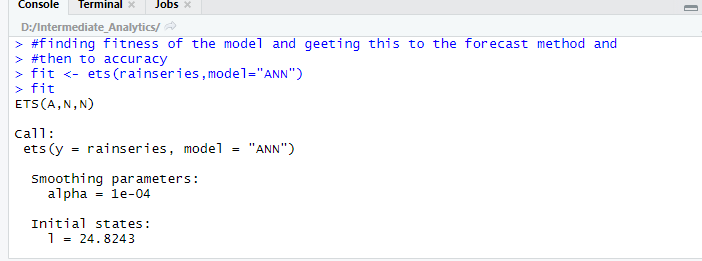


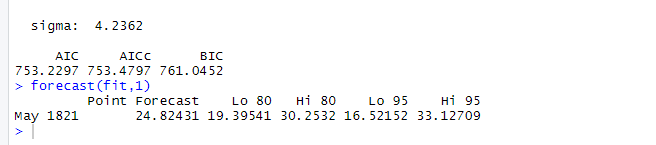


From the above plot we can see clearly the pattern for 25 inches value is also observed with a orange line.

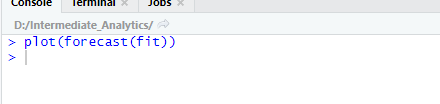
Now finally we are going to validate the accuracy of the model. Prior to which we are finding the fitness of the model using ets function, then forecasting and then finding the accuracy. ETS function is applied for the exponential smoothening of the data set(Exponential smoothening state space model).

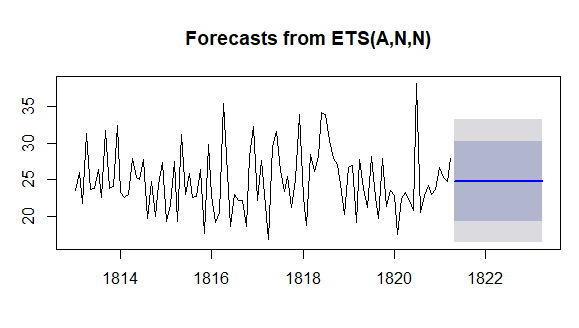






From the above results we can infer that the value l is 24.82 which is approximately 25 only which clearly indicates that the model we applied i.e TS is perfect fit for our data set and also the forecasted value is coming out as 24.82 which is ~25. We will plot the forecast against fit now.

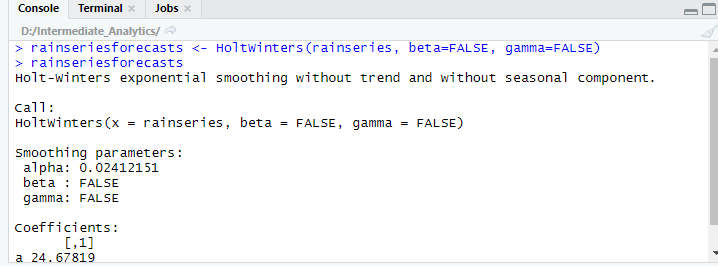




The above graph is also denoting the rainfall around 25 inches.

We can also check the forecast using holtWinters function which will predict the value of alpha based on the data set and here is the logic:





From the results we are confirmed that the 25 inches is the predicted value.

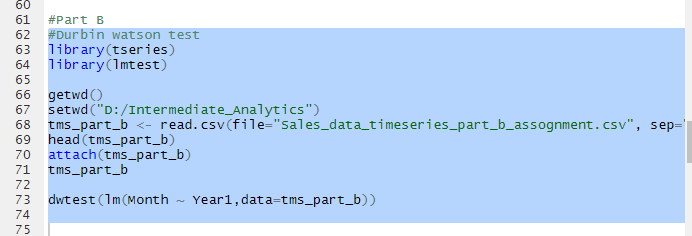
**Part B:** For addressing the issues of correlations between successive values of times series I have selected a data set of sales data for 12 months for one year. I am reading the csv file and using the data set to apply Durbin Watson test for verifying the autocorrelation. Durbin Watson test is used for detecting first level autocorrelation. Let’s assume that

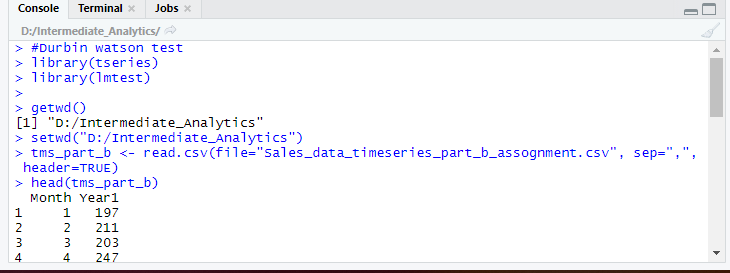
Ho(null hypothesis ) is that sales data is not autocorrelated and

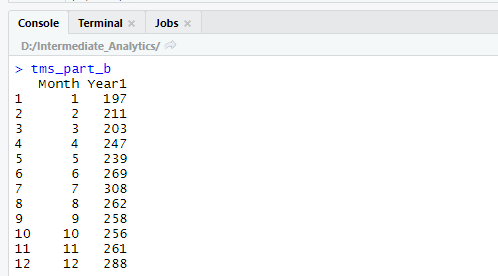
Ha stands for alternative hypothesis i.e sales data is positively autocorrelated.

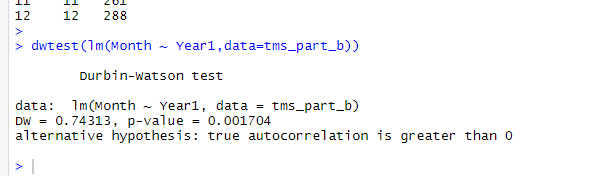
And based on the p-value we will reject or accept the values.

Here is the logic and results for the same.









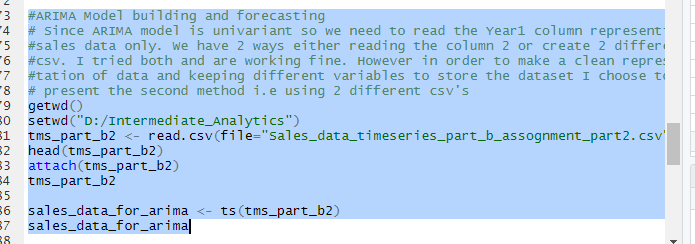
From the p-value we can clearly observe that it is very less and hence

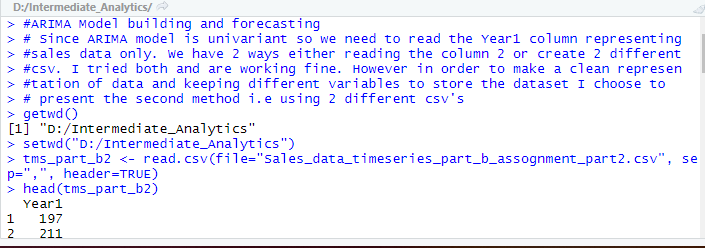
We have to reject the hypothesis which concludes that the sales data is having positive autocorrelation. Reason behind doing is that ARIMA model/Exponential models don’t make any assumptions about the correlations between consecutive values of any time series. Also, it allows non zero auto correlations in irregular components. The basic formula for this model is

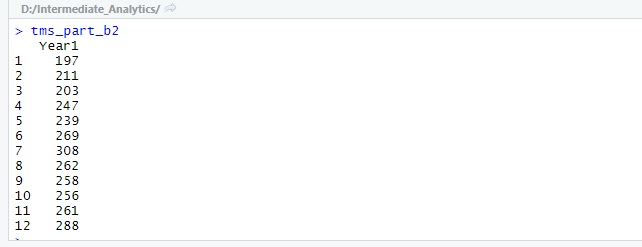
ARIMA(p,q,d), where d is the difference in the order and we need to find out p and q. and for that we need to find out correlogram and partial correlogram. Here is the logic for finding these values:

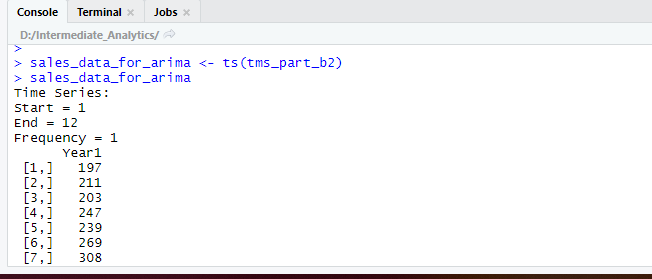
So, our data is valid to use for applying ARIMA model. This is a exponential smoothening method and used for making forecasts/predictions. Since we know that ARIMA model works on univariant data so we need to read the sales data only from the data set and storing in a new variable. Also, sales data is changing every month so we need to find out the order of difference too.

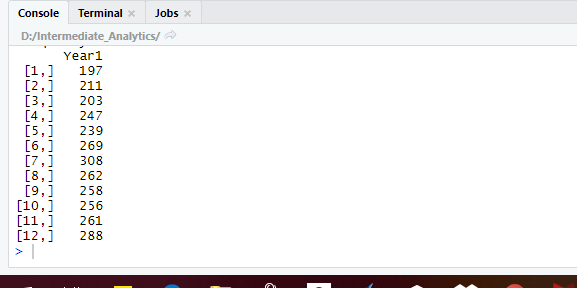
Here is the logic for reading the data and storing in a variable to feed for further calculations.





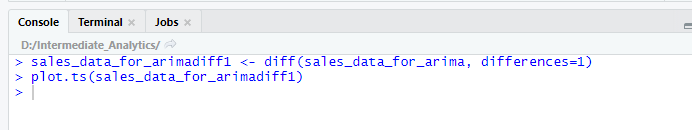


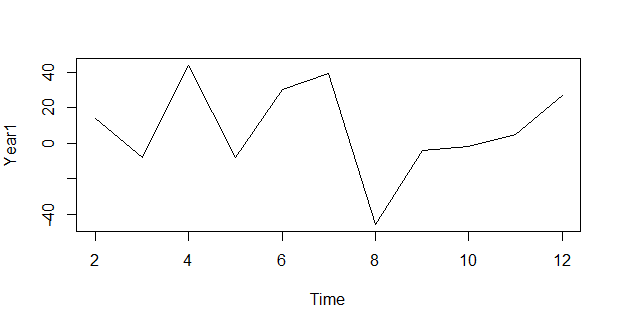




Now calculating the difference and plotting the difference time series.



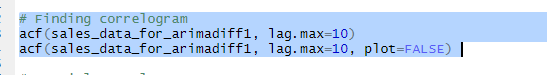


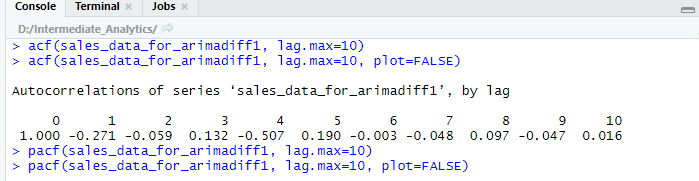


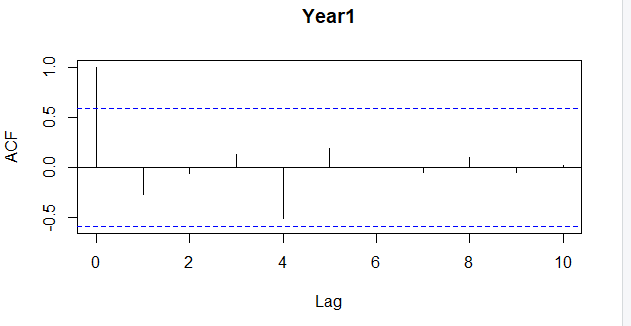
So from the graph we can see that the our data set is non stationary time series based and hence we have a ARIMA(p,q,d) model. So we have to difference the time series. “d” is the order of differencing and we need to find p and q.

And for that we need to figure out correlogram and partial correlogram using acf and pacf functions. Here is the logic and output:

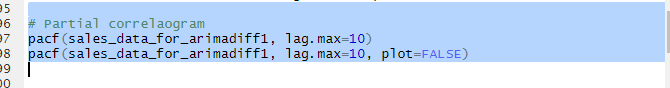
NOTE: Plot function is set as false so as to get the actual values autocorrelation and partial autocorrelation.

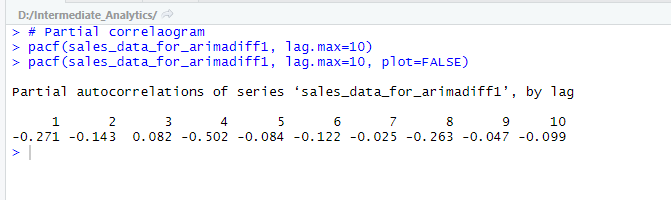


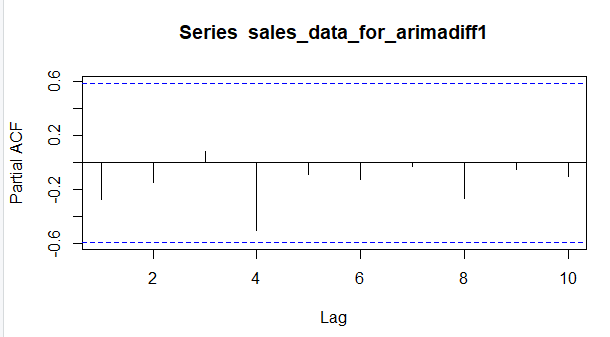




We can observe that autocorrelation at lag 0 is exceeding the significance boundary and no other value is exceeding it.



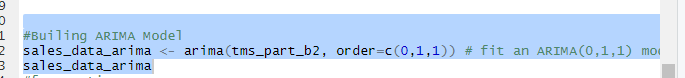


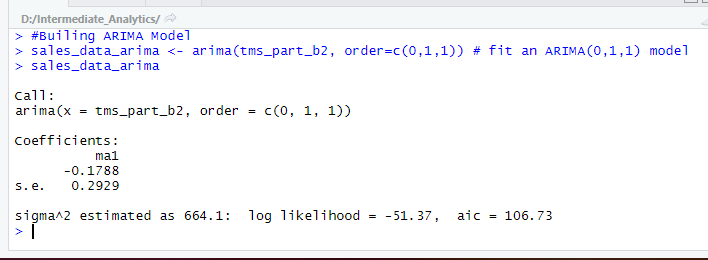


In partial auto correlation no value is exceeding or crossing the significant boundaries

So we can conclude that a ARMA(0,1) model i.e moving average model for order q=1 can be build based on the values obtained which are exceeding the significant boundary for auto correlation and partial autocorrelation.

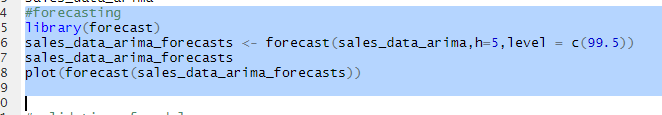
So now we will build the model as follows:

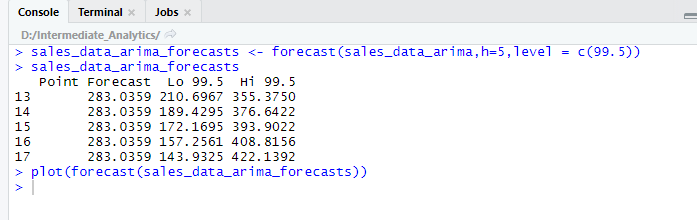


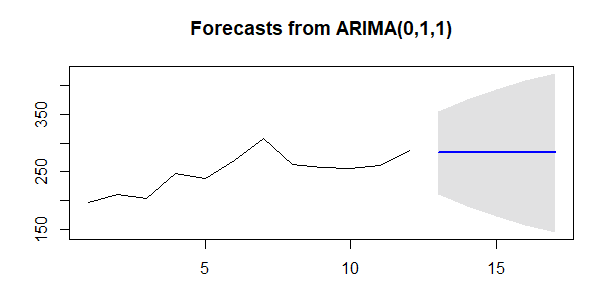


So we are fitting ARIMA(0,1,1) model to our time series of sales data for a particular year. The value of ma1 is coming as - 0.1788.

Now we will do forecasting based on the arima model.

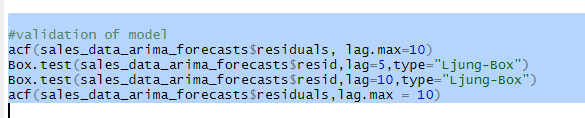


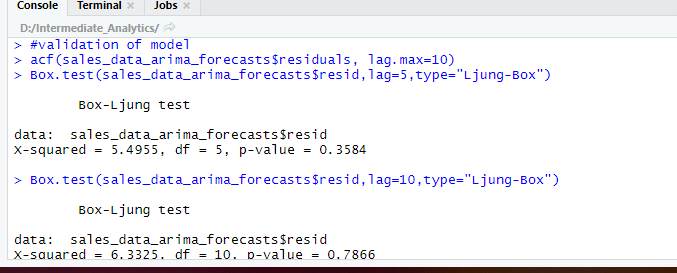


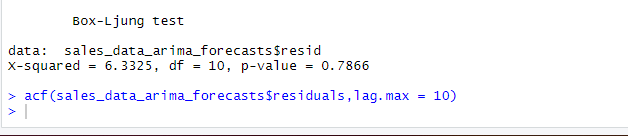


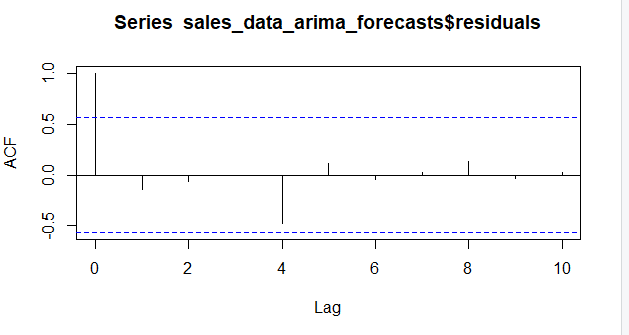
We are making the forecast for 99.5% prediction level. The original time series includes the forecast of sales based on the data provided. This forecast is giving us the data for low and high values for next coming 5 months.

Now coming to the validation of the test. We will use Ljung-Box test.









Since the p-value is 0.7 which indicates that there are chances of correlation based on the test results which is already known to us via Watson test. Also from the acf graph we can observe that only one value is exceeding the significance boundary.

**Conclusion**

For part A of this assignment I would like to conclude that it is highly predicted to have a 25 inches rainfall based on all the predictions, models and forecast we applied.(detailed explanation with proper facts and figures are already explained in the analysis.)

For Part B of this assignment I would like to conclude that autocorrelation existed in the data set (calculated an proved using the Durbin Watson test) and the ARIMA model results are also good and the estimated value of theta is -0.1788 and also based on the value of p,q and d we build ARIMA (0,1,1) model which can be numerically represented as : X\_t - mu = Z\_t - (theta \* Z\_t-1). Also the 5 months forecast is done (explained in the analysis with proper logic facts and figures.)

# References

1. “Using R for Time Series Analysis¶.” *Using R for Time Series Analysis - Time Series 0.2 Documentation*, <https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html#decomposing-time-series>.
2. Dalinina, Ruslana. “Introduction to Forecasting with ARIMA in R.” *Oracle Data Science*, <https://blogs.oracle.com/datascience/introduction-to-forecasting-with-arima-in-r>.
3. Fox, John, et al. “DurbinWatsonTest: Durbin-Watson Test for Autocorrelated Errors in Car: Companion to Applied Regression.” *DurbinWatsonTest: Durbin-Watson Test for Autocorrelated Errors in Car: Companion to Applied Regression*, 15 Nov. 2019, <https://rdrr.io/cran/car/man/durbinWatsonTest.html>.
4. Bowerman, Bruce L., and Richard T. OConnell. *Forecasting and Time Series: an Applied Approach 3rd Ed.* Duxbury Press, 1993.
5. <https://www.youtube.com/watch?v=Y5T3ZEMZZKs>
6. Chatterjee, Subhasree. “Time Series Analysis Using ARIMA Model In R.” *DataScience*, 5 Feb. 2018, <https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/>.