

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6015 20904 Intermediate Analytics**

**Professor Name: ChuanLi Jiang,**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: A. 2020**

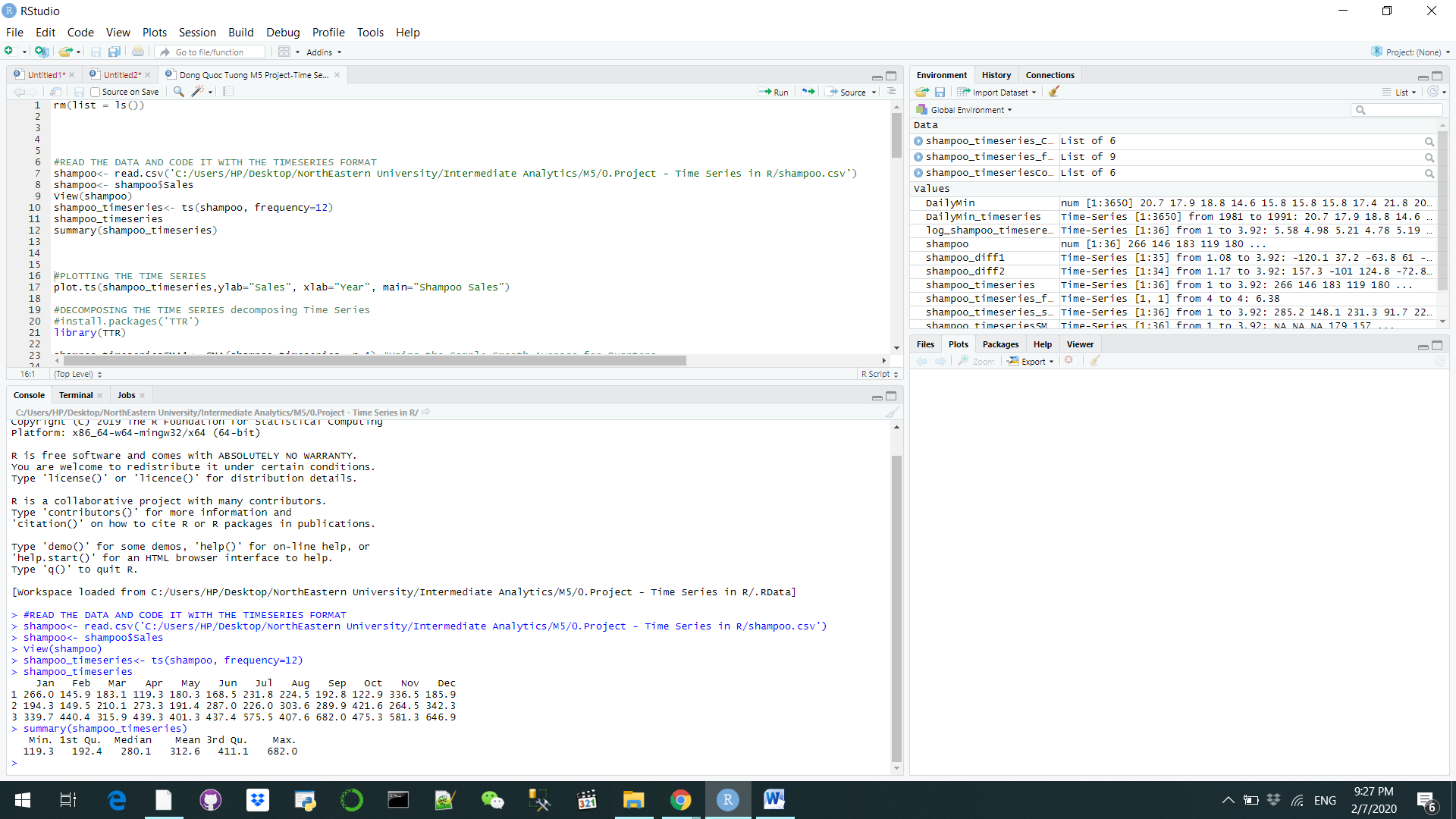
|  |
| --- |
| **Module 5: Time Series in R**  **Completion Date: February 9th Due Time:12:00am** |

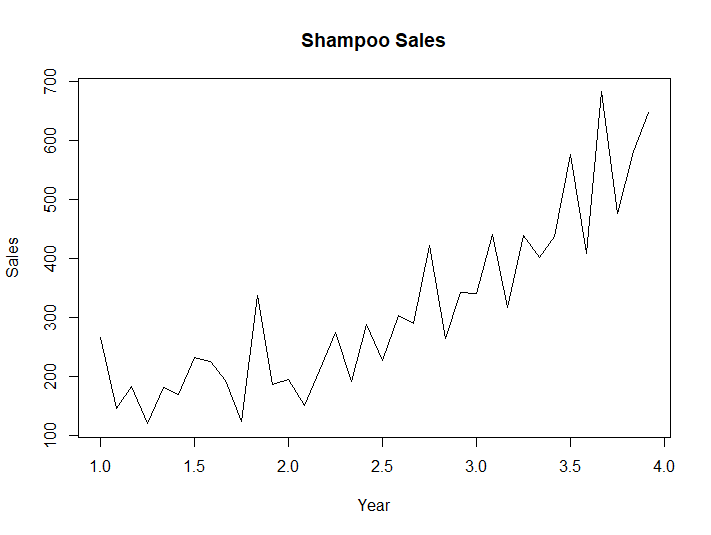
**Statement of Authorship**

*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

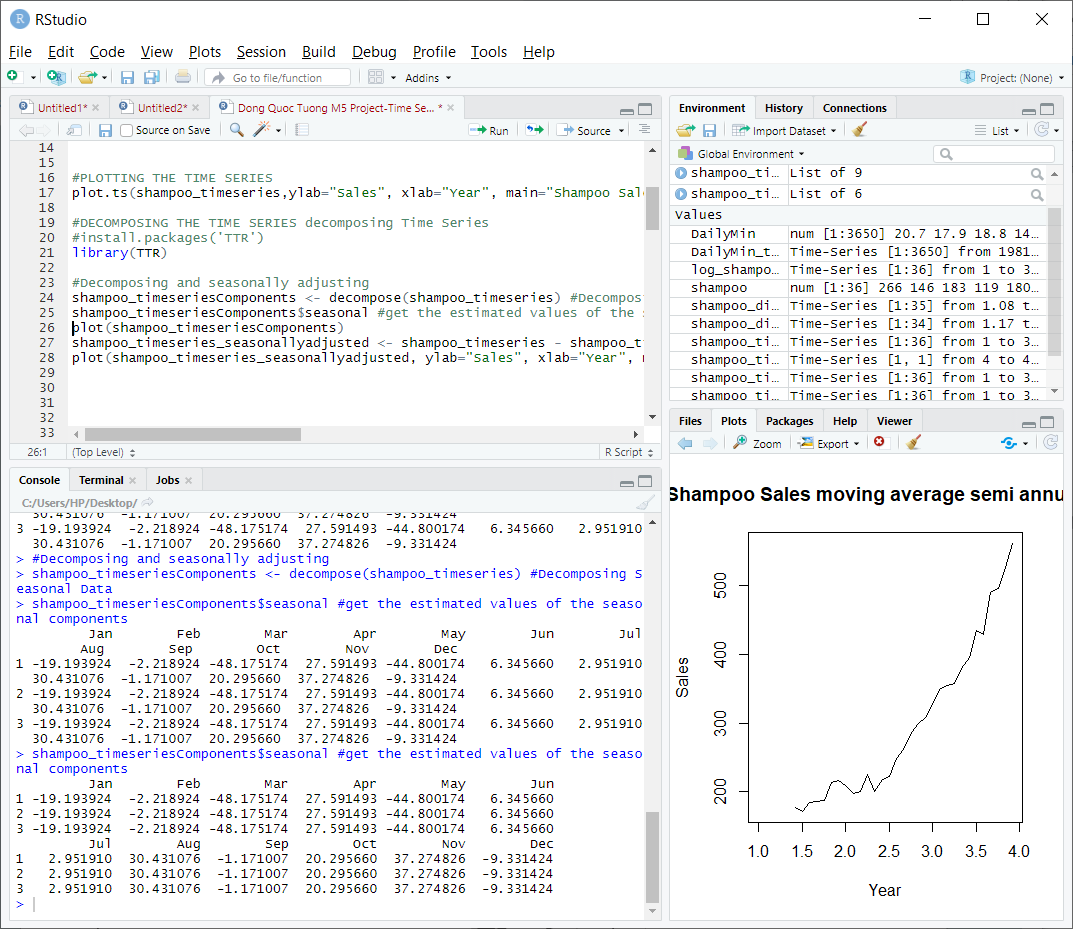
**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

### In the project, we will master the Time Series analysis with R to predict the dataset. It will include Decomposing Time Series, Forecasting Using Exponential Smoothing and ARIMA model at the end. The dataset has 36 observations that represent the sales of shampoo over a period of three years.

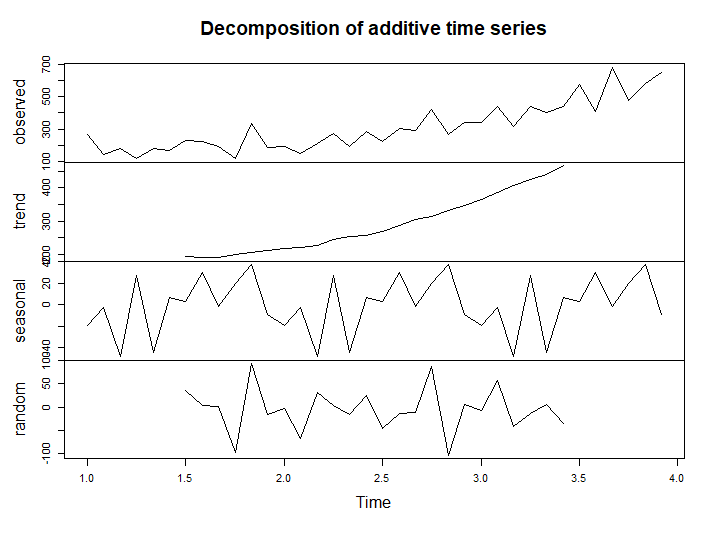




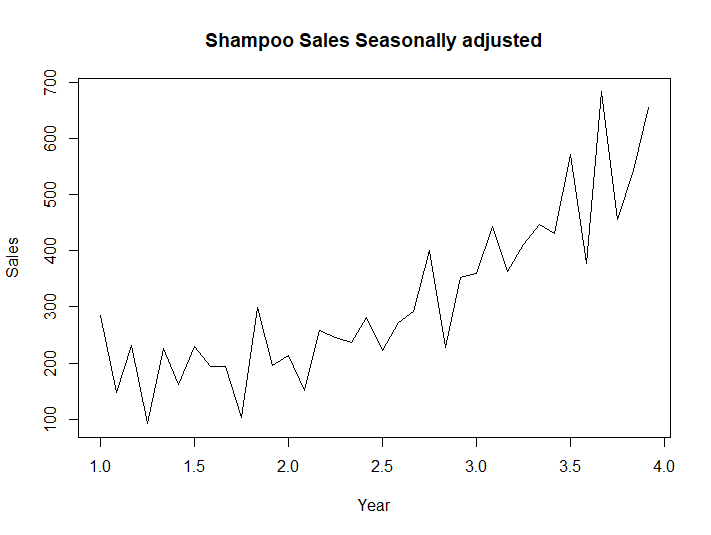
The data is then organized to the time series format through ts() function with the frequency of 12. AS we can see from the plot, shampoo sales rose substantially after 3 years and hit annual peaks near the end of every years, around November. It appears that an additive model might not be the best suitable model to describe this dataset since the size of seasonal fluctuations as well as random fluctuations rise as time went by. Therefore, we need to decompose the time series to fit it with an addictive model.



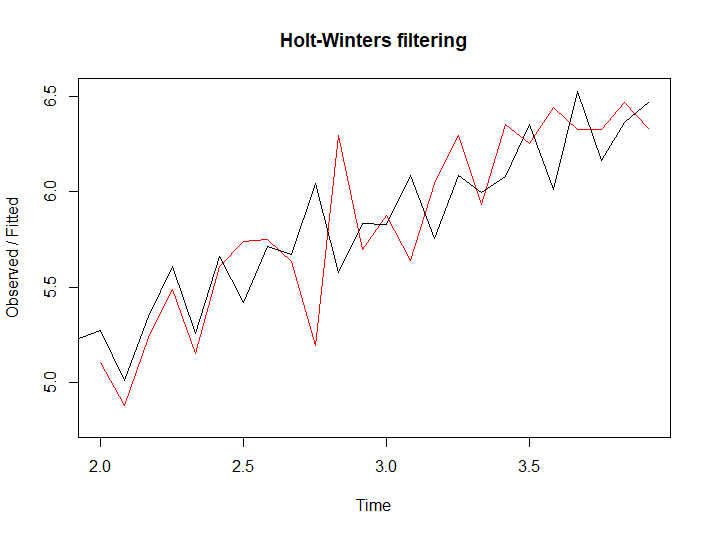
Decomposing a time series means we are extracting the data itself from the constituent components. Depending on the dataset, sometimes you can just decompose but other time you have to seasonally adjust it at the same time like this one.. We used the Decompose() function in R which will automatically estimates the trend, seasonal and irregular components of the shampoo sales for us. The table below illustrates the estimated seasonal components. Demand for shampoo soars in the summer-autumn with a positive seasonal factor, reaching a peak at 37.27 in November, but dwindle once winter comes (December-March).

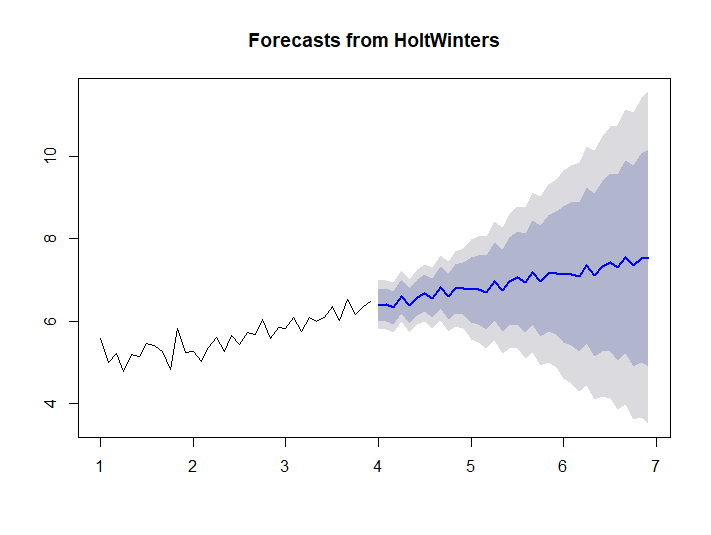


The plot above shows us the original time series at the top, followed by the estimation of simplified trend, seasonality, and the irregular components. The trend is a straight forward line from 100 to 400 within 3 years and no fluctuation. Then we subtract the seasonal component with the original time series shampoo sales dataset to have the final result. We see that there is not much difference between this plot and the plot of original time series we had.

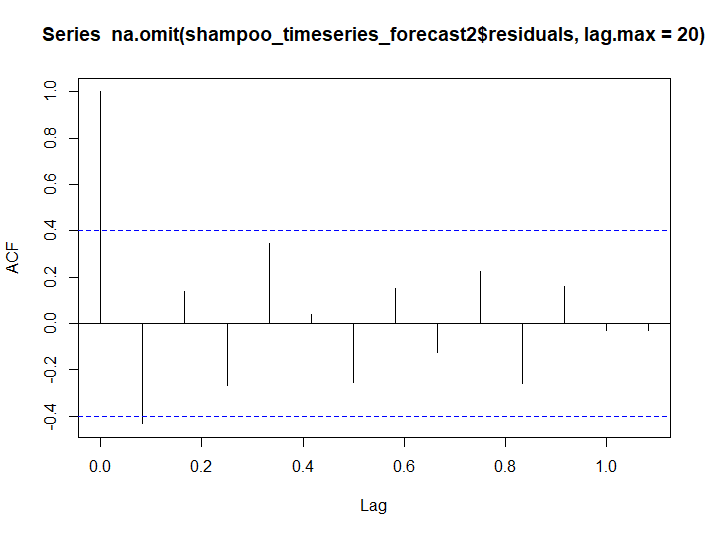


We will use the Holt Winters Exponential Smoothing to make short term we have the time series that can be described with an additive model with an upward trend and seasonality. (Hayes, 2020) Holts-Winters exponential smoothing predicts the level, slope and seasonality at the current time point. We first need to fit the predictive model for the log of the monthly shampoo sales. We will have the value of alpha 0.05, beta 1 and gamma 0.56. Since alpha has the closest value to 0, that means alpha has little weight on the shampoo sales when we plotting forecast in the future and the estimate of the level time rely heavily on both recent and past’ data. Value of beta of 1 means the estimate of slope b is updated as time goes by, on the contrary to the average gamma value of 0.56 indicating seasonal component is not updated all the time. We know that it is a relatively tight fit of the model to the data thanks to the small Sum of squared estimate of errors (2.153)

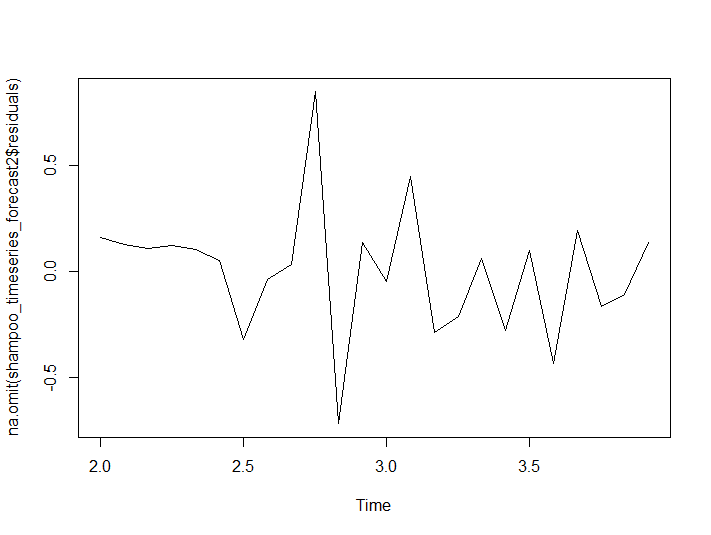


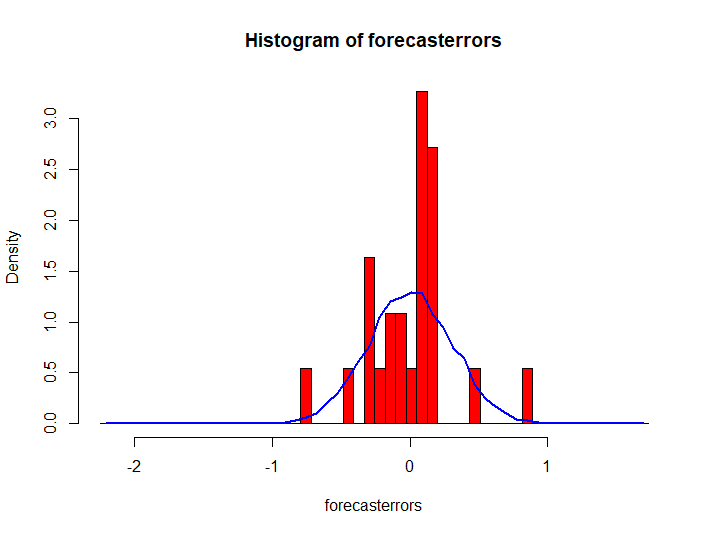


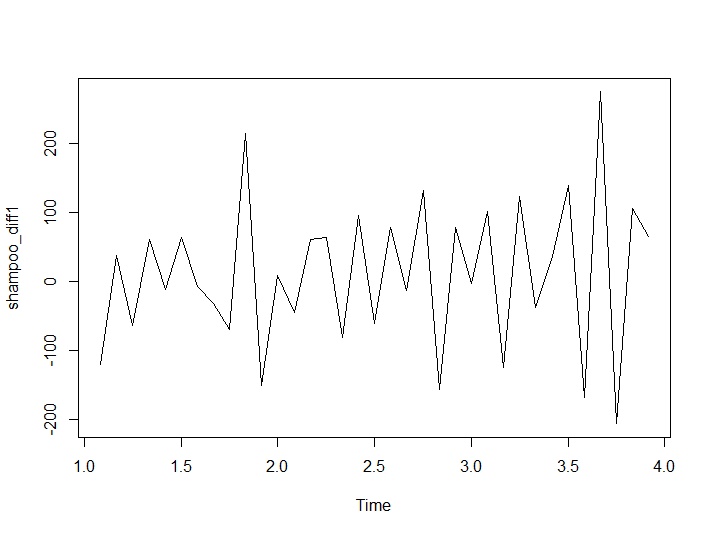
The black line indicates the real value while the red line indicates prediction. The plot from Holt Winters was not really successful as the red line does not correlate well with the black line by the end. But it is decent enough for us to make future estimation for the next 36 months or 3 years. The forecast is shown in the blue line below and the grey and purple shade area indicates the 80% and 95% prediction intervals respectively.

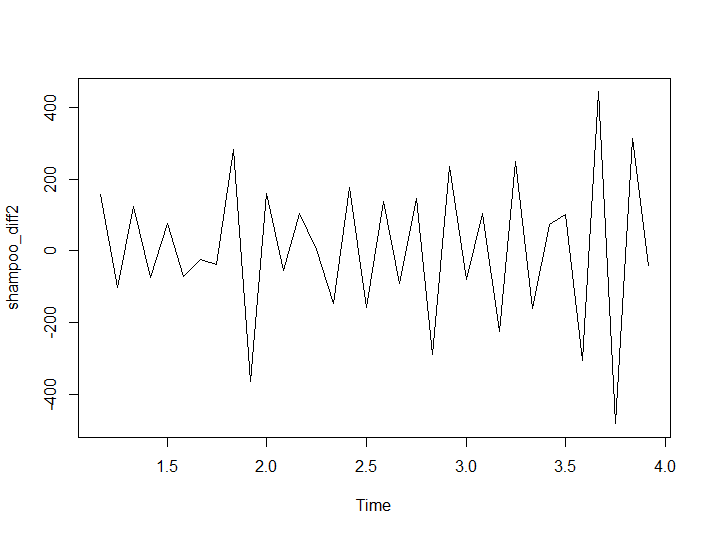


Now, we try to investigate if we can improve the model with the in sample forecast zero auto correlations at lag 1-20 through correlogram and Ljung-Box test. Almost non of the in-sample forecast errors exceed the significance bounds for lag 1-20. P-value for Ljung-Box test is 0.32, we do not reject the Null hypothesis that our model does not show the lack of fit. We then plot the time plot and we can tell that the forecast error does not have the constant variance. The histogram also tells us that the forecast does not have a normally distributed data but rather a left skewed.

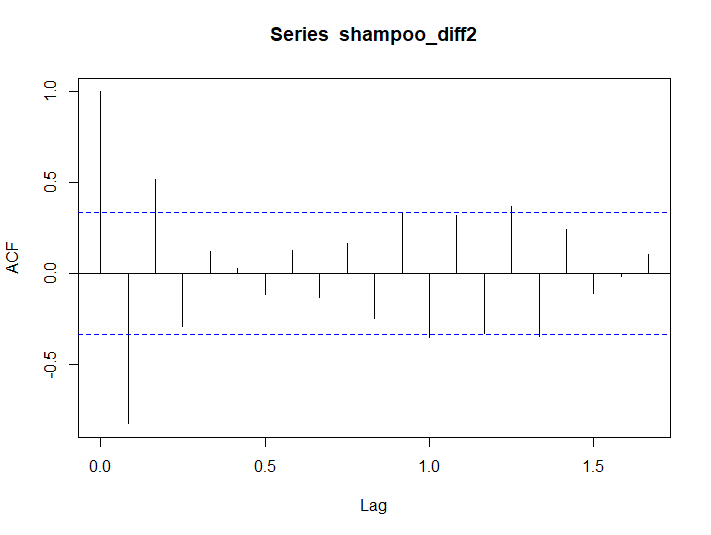
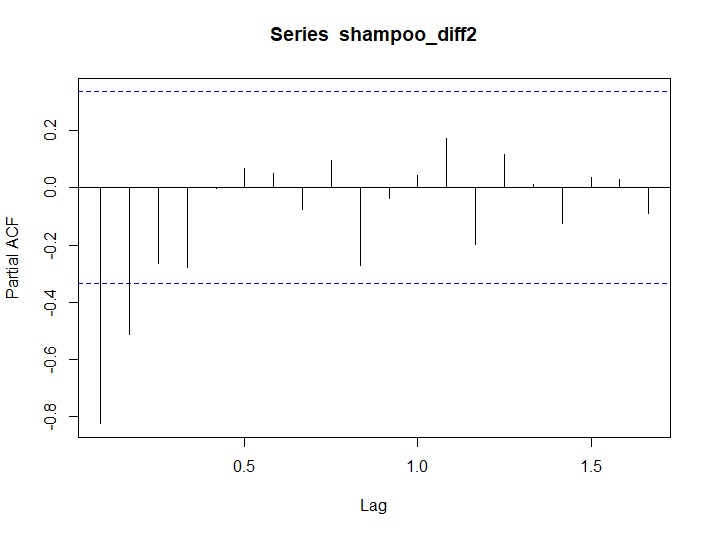




The evidences above tell us that Holt Winters exponential smoothing might not provide the adequate predictive model of the log of shampoo sales. It needs to be improved or changed completely and we can ignore the assumption upon the prediction intervals that were made earlier.

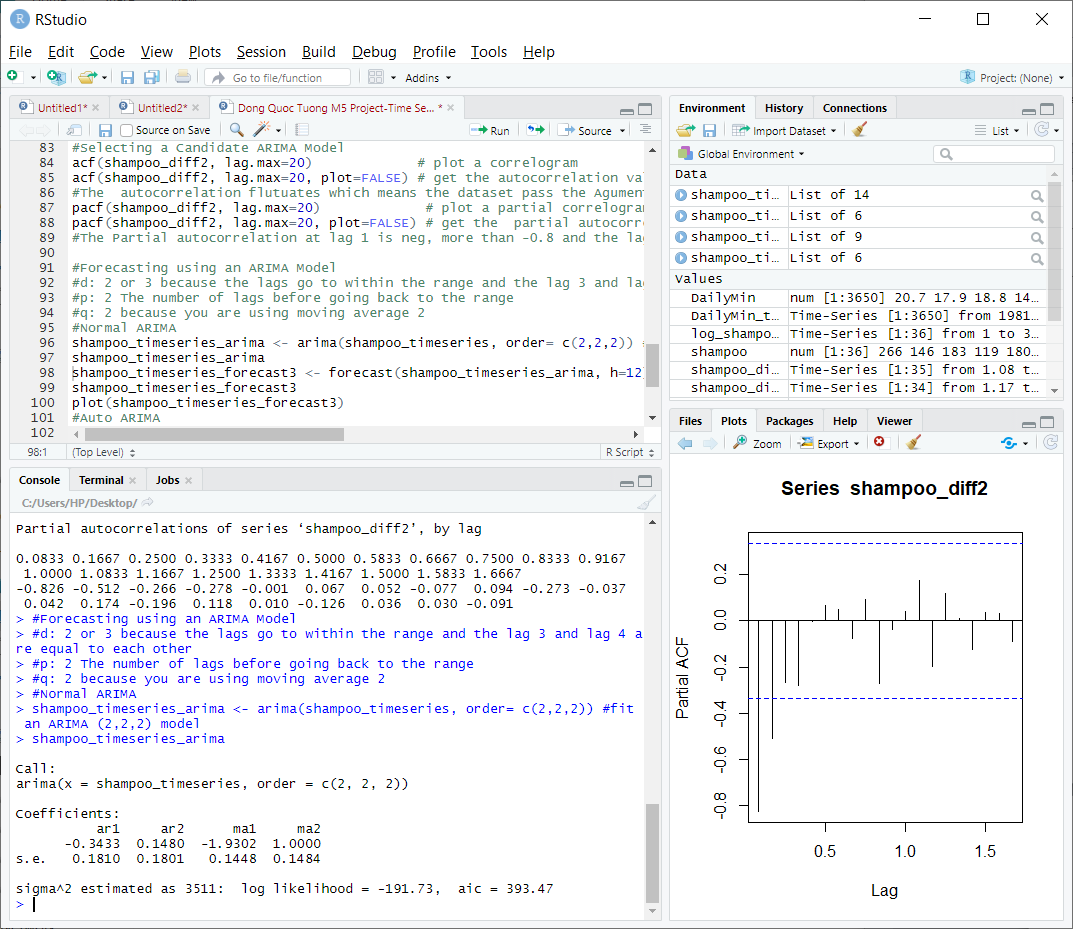


Keeping in mind that assumptions about successive values and time series’ relationship are not available when it comes to exponential smoothing, but we in this case we need a better predictive model like ARIMA. (Brownlee, 2017) We would first turn the dataset into stationary data, then plot the differences a time series through diff() function in R. If we have a look at the graphs above , the time series of second differences on the right looks more stationary in mean and variance compared to the one differences on the left. Hence, we will difference the time series in the diameter of skirts two times so as to get the stationary series.

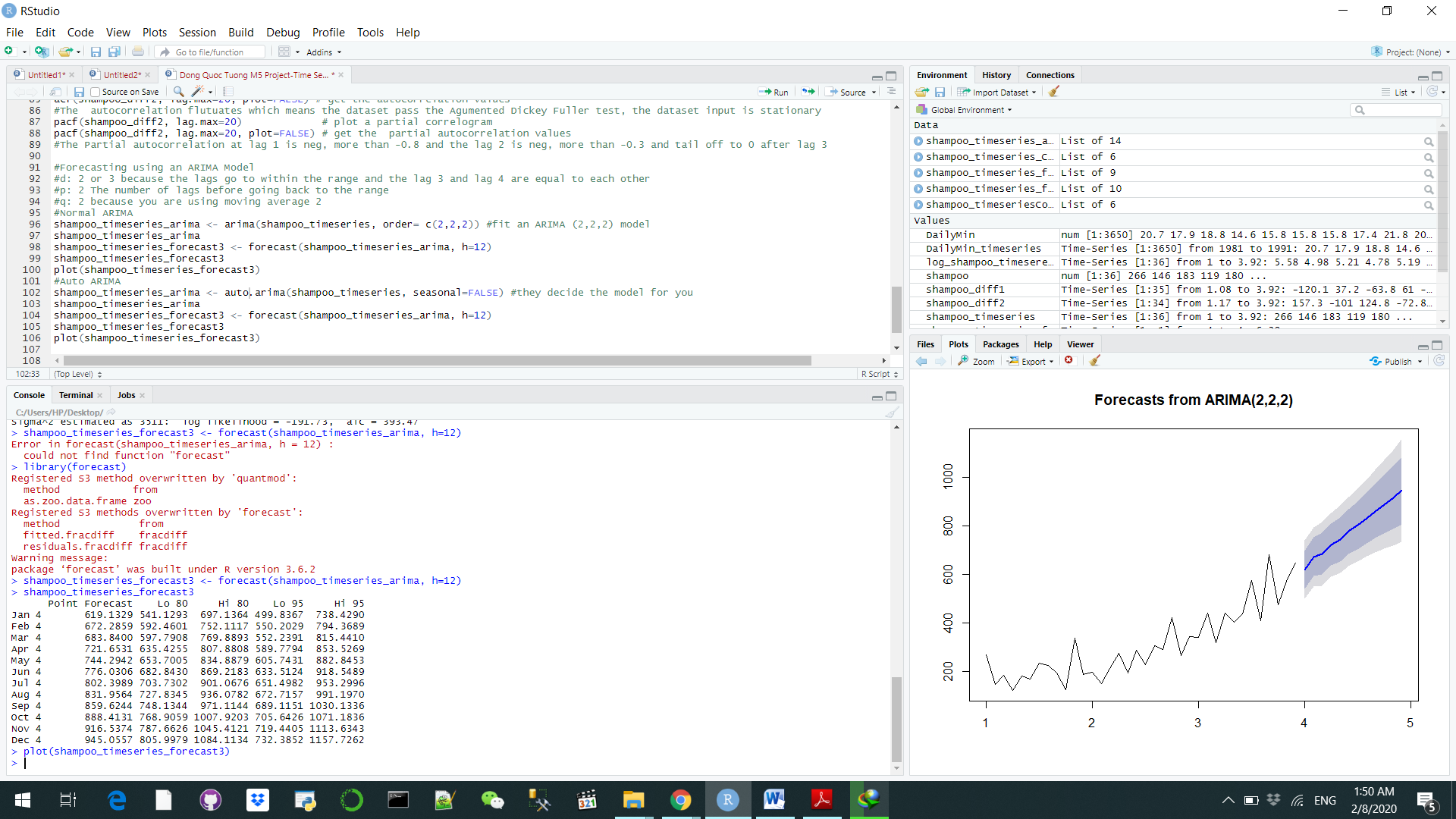


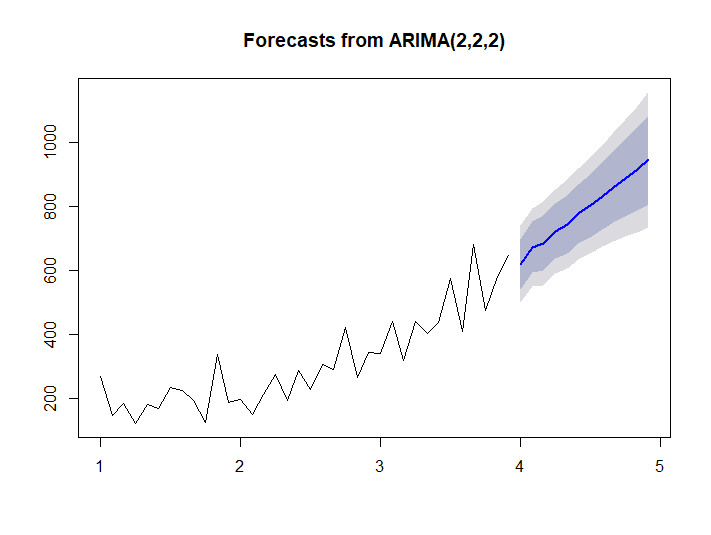
Next we will plot the ACF and the PACF plot. The autocorrelation on the left fluctuates, means that the dataset passes the Augmented Dicket Fuller Test, the dataset input is stationary. Additionally, the Partial autocorrelation on the right at lag 1 is negative, more than -0.8 and the lag 2 is negative, more than boundary and tail off to 0 after lag 3. From these two tables we can pick out the 3 number for our ARIMA model.

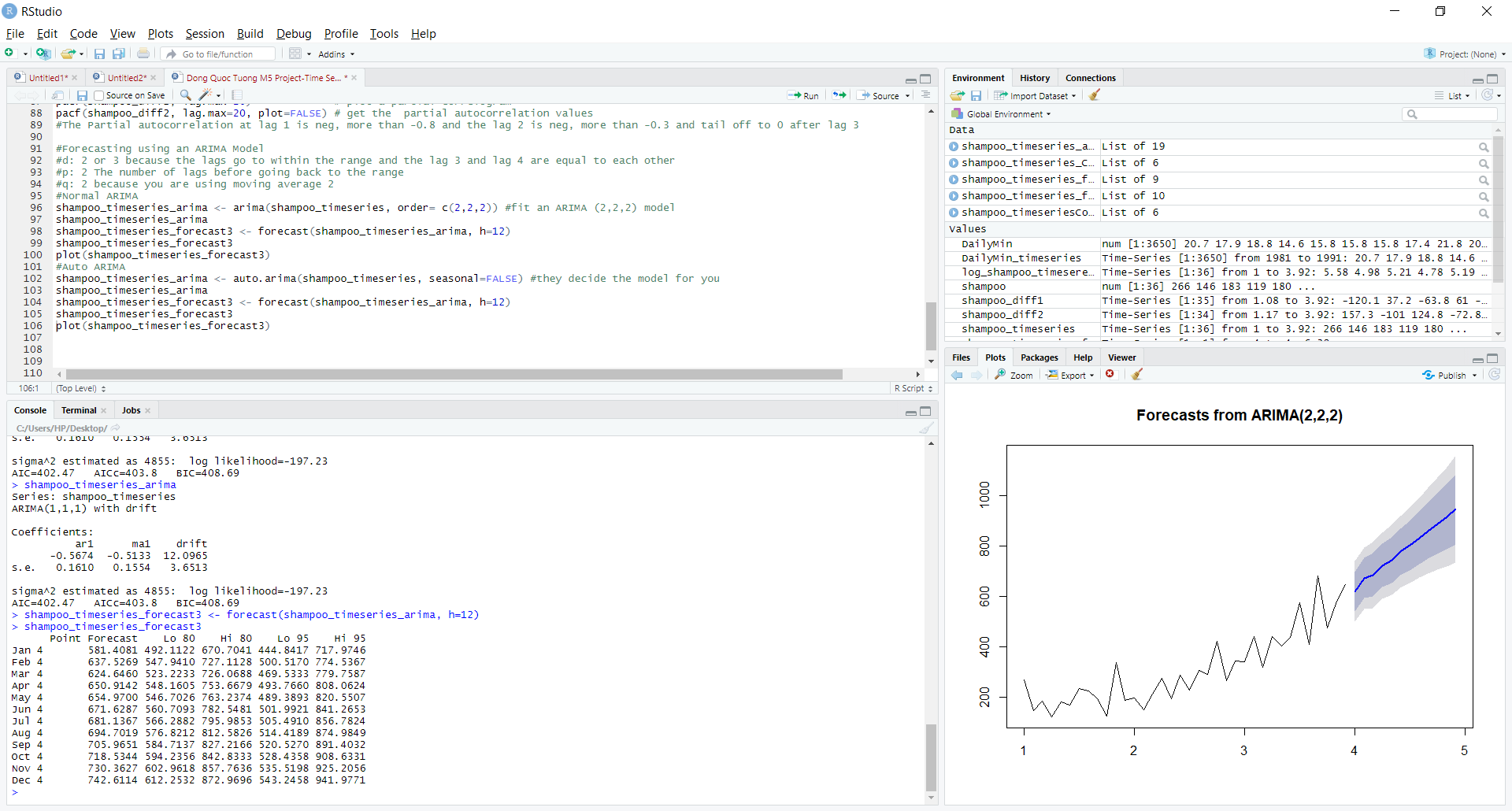
* d= 2 or 3: because the lags go to within the range and the lag 3 and lag 4 are equal to each other.
* p= 2: The number of lags before going back to the range
* q= 2: Because you are using moving average 2

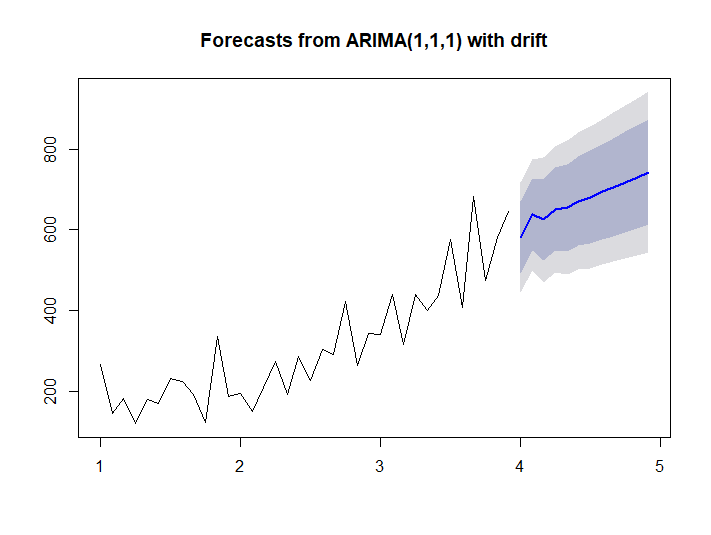


We will now we can build our prediction upon the ARIMA model of (2,2,2). The estimated value of theta is -0.34 in the case of ARIMA(2,2,2) model fitted to the time series of shampoo sales. The forecast of shampoo sales through ARIMA(2,2,2) is demonstrated below with the blue line indicating prediction, purple and grey are 80% and 95% prediction intervals, respectively.





Let’s try this with the ARIMA automation with the auto.arima() function. We saw that the plot and prediction result is pretty much indifferent to what we had before so we can conclude that our method yielded fruitful result



To sum up, the exercise has shown us that time series data requires a lot of effort in analyzing and interpreting it. Even if the requirements match with the methods we choose, the results will indicate differently.

**References**

Brownlee, J. (2017). *How to create an ARIMA model for time series forecasting in python*. Machine Learning Mastery. Retrieved from: <https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/>

Hayes, A. (2020). *Exponential moving average - EMA*. Investopedia. Retrieved from : https://www.investopedia.com/terms/e/ema.asp