TIME SERIES ANALYSIS: CITIBIKE NYC

Cesar Rene Pabon Bernal
Professor Dana Sylvan
Stat 715
City University of New York, Hunter College

ABSTRACT

For the last six years, Citibike, a bicycle-sharing system (BSS), has become a vital feature of the transportation network in New York City.[1] However, due its increase in popularity, challenges have affected its performance, in particular, bicycle unavailability and dock shortages.[2] In this report, we study a time series analysis of the trips taken per day for all Citibike-NYC users and state an appropriate model: log-difference(1) AR(2,0,1).

INTRODUCTION

Established by activists in Amsterdam in the 1960s, the *Witte Fietsenplan* (white bicycle plan), was the first recorded BSS.[3] Although unsuccessful, it created the foundation to the popular system we have today. With a vast and complex network, Citibike is efficient, fun, healthy and positive for the environment.[4] As of May 2019, there are over 12,000 bikes and 750 docking stations that cover 60 neighborhoods in the tri-state area.[5] As ridership increases, so does the amount of information collected. Our objective is to provide an analysis that can alleviate data congestion by identifying a proper time series model, estimate the values of its parameters and explore the goodness of its fit via forecasting. Python and R were utilized interchangeably throughout sections of the project.

I. DATA

Data was collected from the Citibike System Data Repository. A big part of this report was cleaning the raw database and indexing appropriate features. A parser was created which combined and properly labeled quarterly outputs from May 27th, 2013 to February 15th, 2019. It's features are presented as:

1) Date
2) 24-Hour Passes Purchased : missing values = 92
3) 3-Day Passes Purchased : missing values = 1040
4) 7-Day Passes Purchased : missing values = 1057
5) Annual Member Sign-Ups : missing values = 1512
6) Total Annual Memberships Sold : missing values = 1638
7) Cumulative trips (since launch) : missing values = 872
8) Miles traveled to date : missing values = 2005
9) Miles traveled today : missing values = 0
10) Total Annual Members : missing values = 1787
11) Trips over the past 24-hours : missing values = 0

[1] Loaiza-Monsalve, D., and A. P. Riascos. "Human Mobility in Bike-Sharing Systems: Structure of Local and Non-Local Dynamics." *Plos One*, vol. 14, no. 3, 2019, doi:10.1371/journal.pone.0213106

[2] Chung, Hangil, et al. "Bike Angels." Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) - COMPASS 18, 2018, doi: 10.1145/3209811.3209866

[3] Demaio, Paul. "Bike-Sharing: History, Impacts, Models of Provision, and Future." *Journal of Public Transportation*, vol. 12, no. 4, 2009, pp. 41–56., doi: 10.5038/2375-0901.12.4.3.

[4] Heaney, A. K., Carrión, D., Burkart, K., Lesk, C., & Jack, D. (2019). Climate Change and Physical Activity: Estimated Impacts of Ambient Temperatures on Bikeshare Usage in New York City. *Environmental Health Perspectives, 127*(3), 037002. doi:10.1289/ehp4039

[5] Motivate International, Inc. "Citi Bike System Data." *Citi Bike NYC*, www.citibikenyc.com/system-data.

In this report, we focus in the study of *trips taken over the past 24-hours for all users* as a time series model and state a hypothesis:

H_O = Citibike Trips per day (for All Users) increase in the summer months in New York New York City

We propose exploration of H_O and ask:

1) Is seasonality present? Is the model additive or multiplicative?
2) Is the series stationary? Does the featured data need a numerical transformation?
3) Can we propose an acceptable ARIMA model to forecast a stationary time series of our hypothesis?

II PLOT OF TIME SERIES: SEASONALITY

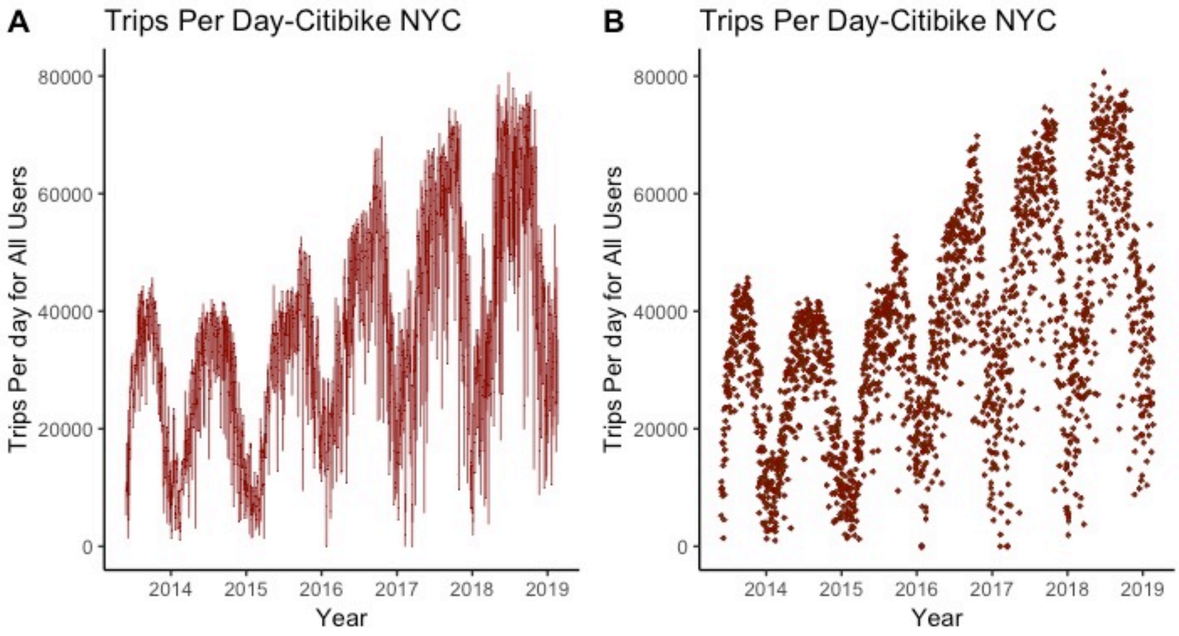

Figure 1 (A) Time series line plot with every value (B) Time series dot plot with every value

Figure 1 is a simple time series model of *trips per day for all users*. According to literature, averaged data produced cleaner results when analyzing noisy BSSs values.[6] Therefore, aggregated mean values were calculated, and for the remaining part of the report, average trips per month will be utilized for analysis. These values are presented in **Table A**.

Table A. MEAN OF AVERAGE TRIPS PER MONTH

Month	*Trips*
May	*33654*
June	*56932*
July	*54215*
August	*58644*
September	*62940*
October	*61864*

[6] Fuller, Daniel, et al. "The Impact of Public Transportation Strikes on Use of a Bicycle Share Program in London: Interrupted Time Series Design." *Preventive Medicine*, vol. 54, no. 1, 2012, pp. 74–76., doi:10.1016/j.ypmed.2011.09.021.

Figure 2 interprets these findings and we state that trips per month are highest between May-October. We extend our hypothesis where average trips per month continue to increase onto September and October; this is most likely due to the extension of warmer months in New York City and the commencement of cyclical seasonal jobs.[7]

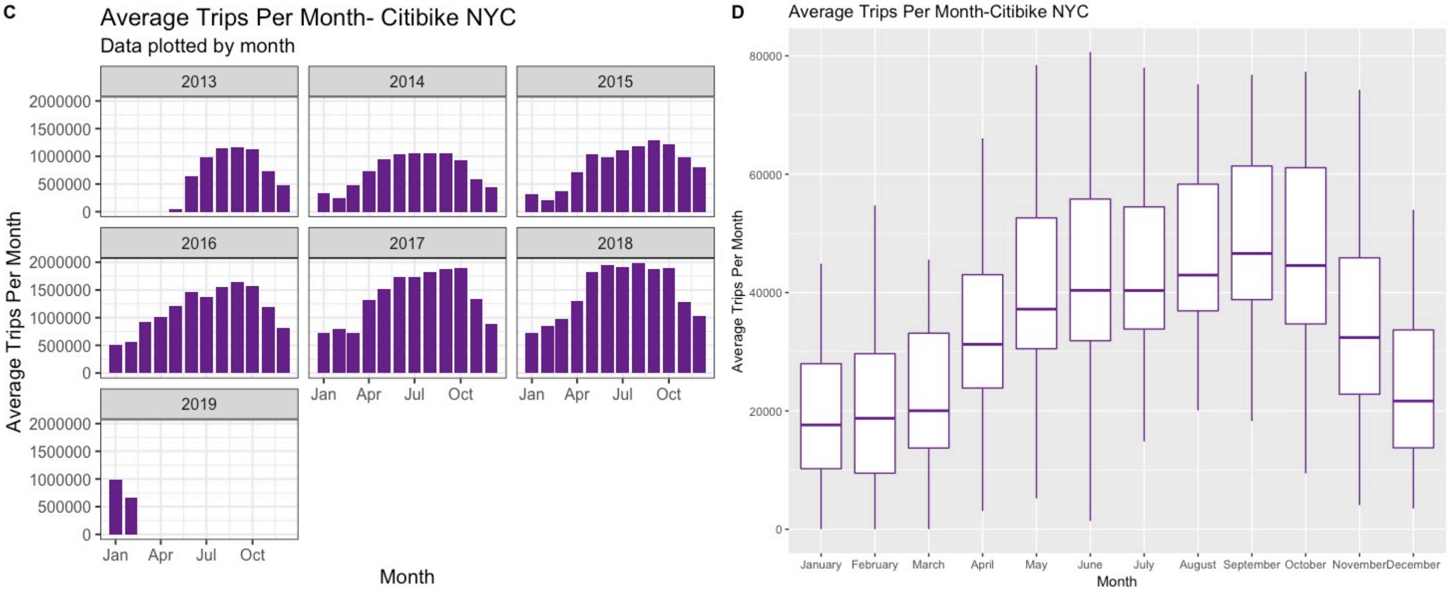

Figure 2 (C) Bar plot of Average Trips per month (D) Bar plot of Average Trips per month

According to **figures 1 and 2**, seasonality appears to be present, however, in order to definitely answer this question, we present a decomposition rendition plot of observed averaged trips, trend, seasonality and residuals in additive and multiplicative alternatives.

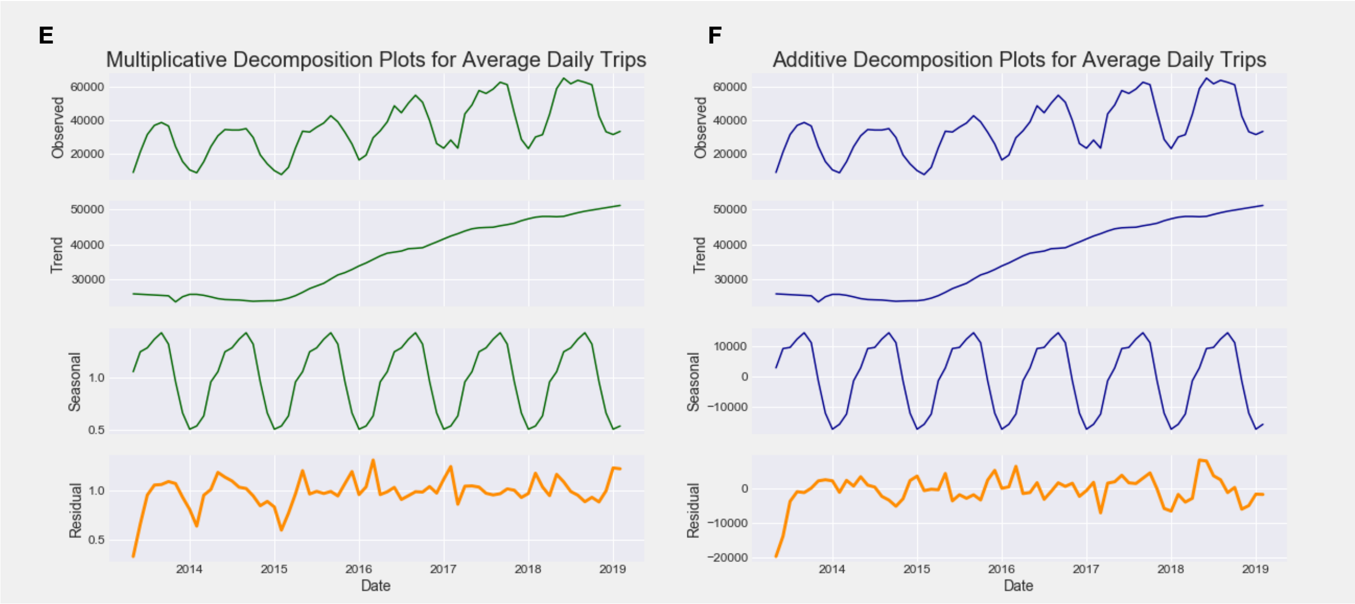

Figure 3 (E) Multiplicative Decomposition plot for Averaged Daily Trips (F) Multiplicative Decomposition plot for Averaged Daily Trips

[7] Norby, R. J., Rustad, L. E., Dukes, J. S., Ojima, D. S., Parton, W. J., Grosso, S. J., . . . Pepper, D. A. (2007). Ecosystem Responses to Warming and Interacting Global Change Factors. *Terrestrial Ecosystems in a Changing World Global Change — The IGBP Series,* 23-36. doi:10.1007/978-3-540-32730-1_3

Due to the small number of averaged observations, we are unable to categorize this series as either multiplicative or additive. Since they both demonstrate similarities in their rendition, and both appear to work in this series, I chose additive. Therefore, we summarize this section and propose:

The average daily trips series is seasonal with a positive trend

III STATIONARITY & TRANSFORMATIONS

In order for us to find an acceptable ARIMA model to forecast a stationary time series of average daily trips, we must transform (if needed) our seasonal data and make it stationary. There are 3 basic criterion for a series to be classified as stationary:

1) The mean of the series should not be a function of time but rather, be constant.
2) The variance should not be a function of time
3) The covariance should not be a function of time.

Figure 4 shows that the average daily trips series, without transformations, violate the three stationary criterion. The ACF and PACF plots have autocorrelation values beyond the 95% white noise C.I. and are thus statistically significantly different from zero. Applying a Dickey Fuller test, value =-3.2223, confirms non-stationarity providing a p-value= 0.08406 > 0.05.

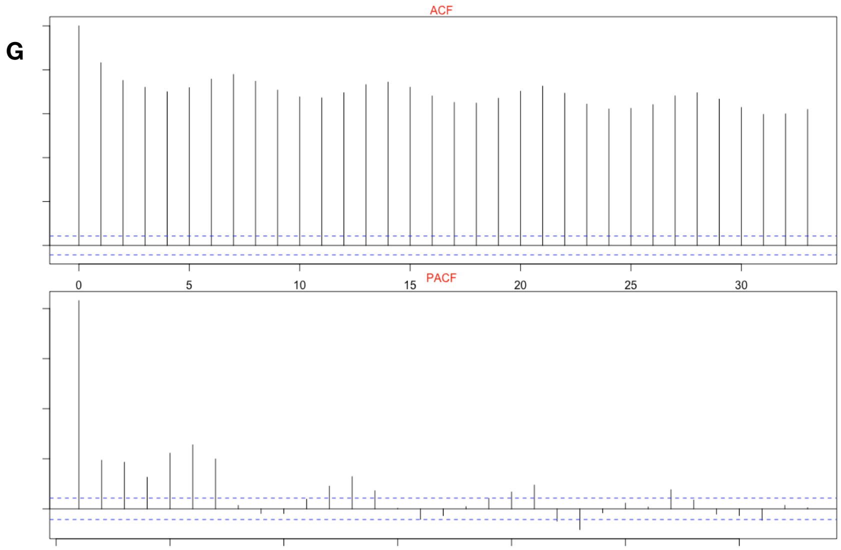

Figure 4 (G) ACF & PACF Plots of Average Daily Trips per Month without Transformations

To make our series stationary, we explore three types transformation: Log, Difference (1) and Log-Difference (1) . **Table B** has a summary of the Dickey fuller tests, **Figure 5** visualization of its outputs and **Figure 6** plots of its respective ACF/PACF results.

Table B. DICKEY-FULLER TEST RESULTS FOR TRANSFORMED SERIES

Transformation	*Trips*	*P-value*
No trasnformation	*-3.2223*	*0.08406*
Log	*-5.4888*	*0.01*
Difference (1)	*-5.731*	*0.01*
Log-Difference (1)	*-5.345*	*0.01*

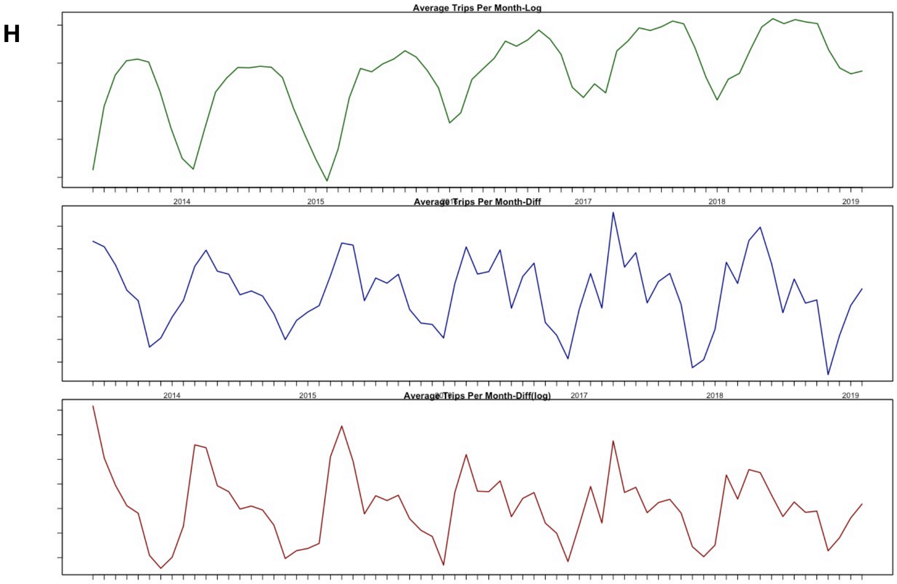

Figure 5 (H) Plots of Transformed Series for Average Daily Trips per Month

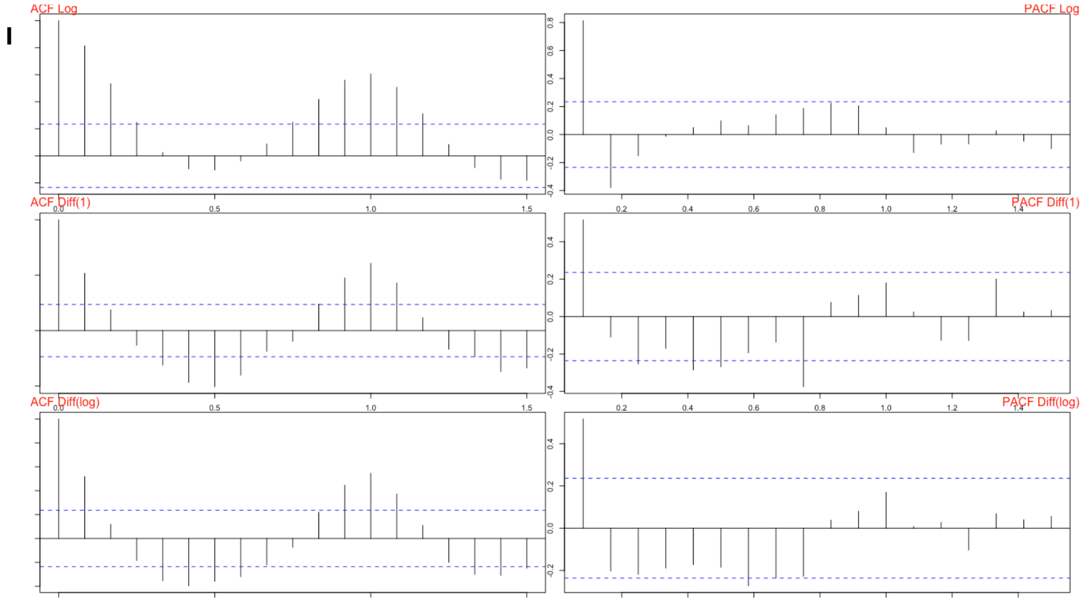

Figure 6 (I) ACF & PACF Plots of Average Daily Trips per Month With Transformations

Based on these results we state that transformation series are as followed:

1) Log: Our observed log series plot shows a slight positive trend but with a Dickey-Fuller score of -5.4888 and p-value= 0.01 < 0.05, its possible stationarity will be met if difference(1) is taken. Based on our ACF, we see that at AR lag 3, the autocorrelation values start to fall within our 95% interval.Based on our PACF, we see that at MA lag 1, the partial AC values start to fall within our 95% interval. Therefore, we propose an AR (3,1,1) model to be evaluated with *constant mean*

2) Difference (1): Our observed difference (1) series plot shows a constant mean with a Dickey-Fuller score of -5.731 and p-value= 0.01 < 0.05, and we confirm stationarity. Based on our ACF, we see that at AR lag 2, the autocorrelation values start to fall within our 95% interval. Based on our PACF, we see that at MA lag 1, the partial autocorrelation values start to fall within our 95% interval. Therefore, we propose an AR (2,0,1) model to be evaluated with constant mean

3) Log-Difference (1): Our observed log-difference (1) series plot shows a constant mean with a Dickey-Fuller score of --5.345 and p-value= 0.01 < 0.05, and we confirm stationarity. Based on our ACF, we see that at AR lag 2, the autocorrelation values start to fall within our 95% interval. Based on our PACF, we see that at MA lag 1, the partial autocorrelation values start to fall within our 95% interval. Therefore, we propose an AR (2,0,1) model to be evaluated with constant mean.

We summarize these finding and propose:

The best model should be a transformed series using Log-Difference (1). This is due to the series showing a constant mean in an observed plot, a competitive Dickey-Fuller score with a p-value <0.05 and the best ACF& PACF plots.

METHODS

IV. MODEL SPECIFICATION & FITTING USING RESIDUAL ANALYSIS

A function in R, auto.arima(), was utilized to fit an integrated auto regressive moving average to the three transformed series. This process is automated and once appropriately differenced, the auto regressive and moving average components are removed. Residuals with a variance is what's left and that's what is explored in this section.

1) Log: **Figure 7** displays the residual plots and an optimal model of AR(2,1,1). The AC values in the ACF are within the 95% interval but the PACF value at lag 9 and 12 have values slightly above the interval.

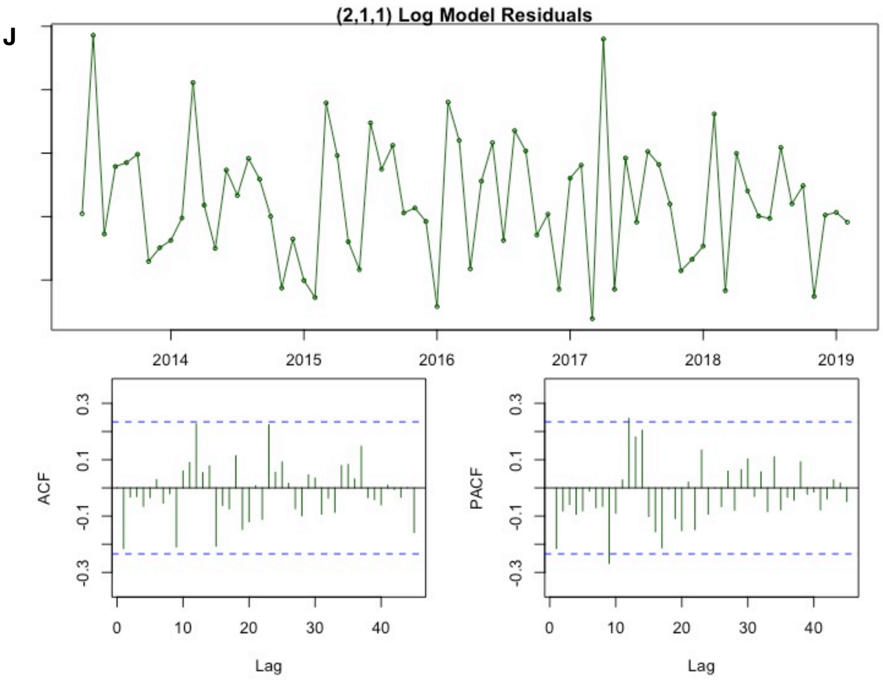

Figure 7(J) Series/ACF/PACF Residual plots for Log transformation

We present AR(2, 1, 1) as : $\nabla Y_t = Y_{t-1} + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} - \theta_1 \varepsilon_{t-1} + \varepsilon -> where : \nabla Y_t = Y_t - Y_{t-1}$ **eq.1**

Using the coefficients in R, we get: $\nabla Y_t = 1.399 Y_{t-1} - .7231 Y_{t-2} - 0.8854 \varepsilon_{t-1}$

2) Difference (1): **Figure 8** displays the residual plots and an optimal model of AR(3,0,1). The AC values in the ACF and PACF both have lag 2 values , respectively, outside within the 95% interval; lag 9 and lag 11.

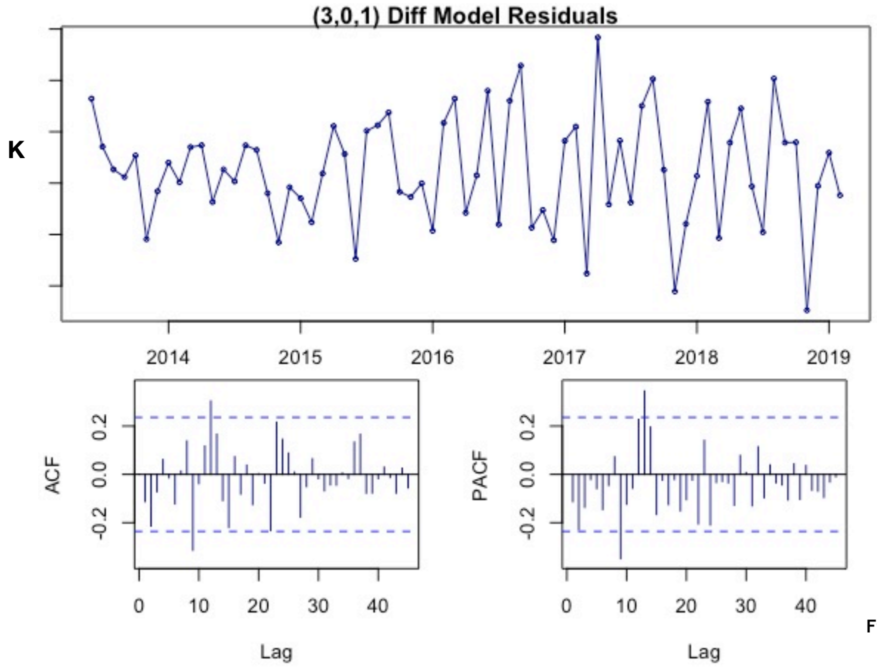

Figure 8 (K) Series/ACF/PACF Residual plots for difference (1) transformation

We present AR(3, 0, 1) as : $\nabla Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \Phi_3 Y_{t-3} - \theta_1 \varepsilon_{t-1} + \varepsilon -> where : \nabla Y_t = Y_t - Y_{t-1}$ **eq. 2**

Using the coefficients in R, we get: $\nabla Y_t = 1.1520 Y_{t-1} - 0.2535 Y_{t-2} - 0.306 Y_{t-3} - 0.857 \varepsilon_{t-1}$

3)Log-Difference (1): **Figure 9** displays the residual plots and an optimal model of AR(2,0,1). In the ACF plot, all of the AC values , woefully fall within the 95% C.I. where as in the PACF plot, 2 lag values slightly surpass the interval in (lag 9 & 11).

We present AR(2, 0, 1) as : $\nabla Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} - \theta_1 \varepsilon_{t-1} + \varepsilon -> where : \nabla Y_t = Y_t - Y_{t-1}$ **eq. 3**

Using the coefficients in R, we get: $\nabla Y_t = 1.399 Y_{t-1} - .7231 Y_{t-2} - 0.8854 \varepsilon_{t-1}$

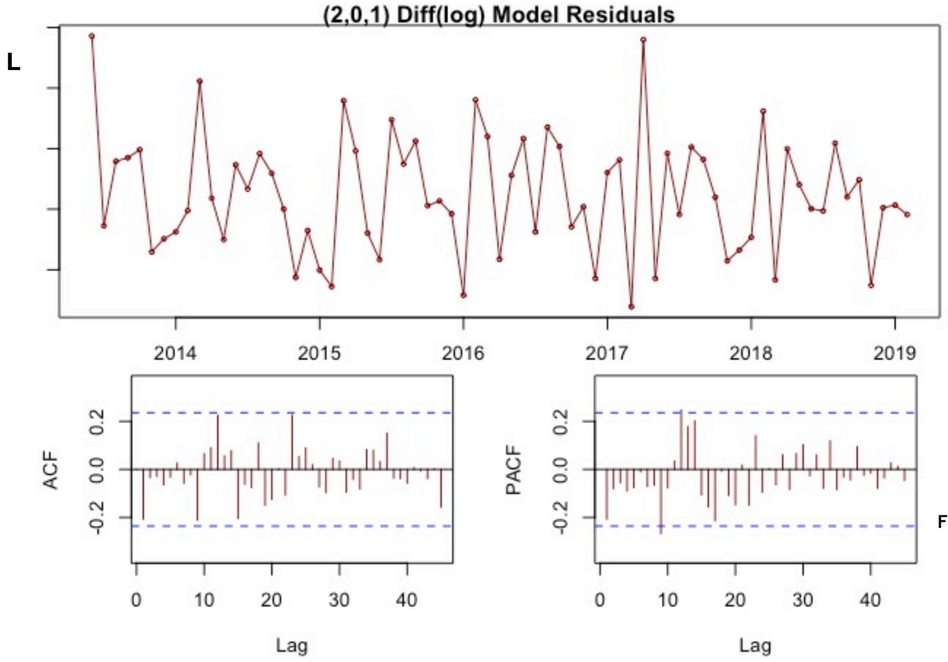

Figure 8 (L) Series/ACF/PACF Residual plots for difference (1) transformation

The values of interest for respective transformations of the auto.arima() function output are presented in **Table C.**

Table C. AUTO.ARIMA() OUPUT FOR RESIDUALS FOR TRANSFORMED SERIES

Transformation	*Models*	*P-value*	*AIC*	*BIC*	*RMSE*	*σ2*
Log	*AR (2,1,1)*	*0.01*	*-15.88*	*-6.94*	*0.1993*	*0.042*
Difference (1)	*AR(3,0,1)*	*0.01*	*1394.54*	*1405.71*	*5428.144*	*3.123E+07*
Log-Difference (1)	*AR (2,0,1)*	*0.01*	*-15.88*	*-6.94*	*0.2007*	*0.042*

We summarize the model specification findings and propose:

We identify an appropriate model for this series as AR(2,0,1) with constant mean. Overall, it had the best ACF plot and competitive low RMSE/σ2/AIC scores.

V. MODEL DIAGNOSTICS

An Ljung-Box test was evaluated on the AR(2,0,1) model. With a p-value = 0.246 > 0.05, it confirms that residuals are independent and therefore, solidifies the decision that this chosen model was the most appropriate. **Figure 9** is a QQ-plot of AR(2, 0, 1). It demonstrates a good fit with very little heaviness at its ends.

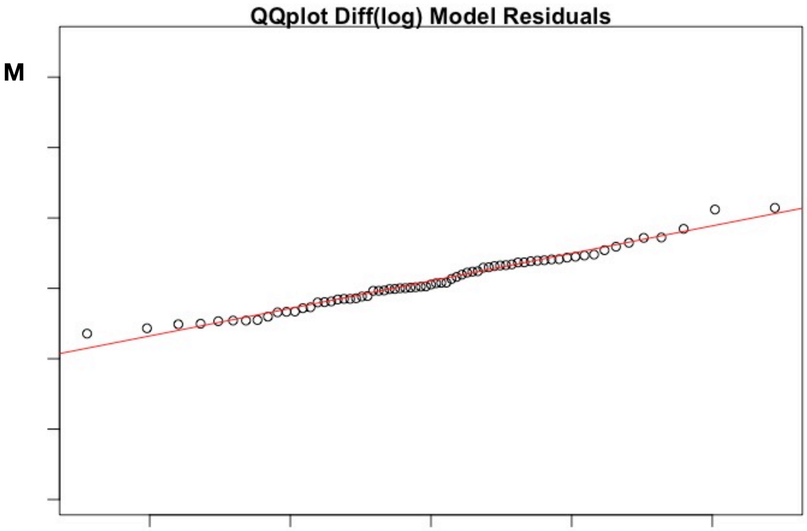

Figure 9 (M) QQplot of best model

VI. FORECAST

A training set was created with values from May 27th, 2013 to November 31st, 2018 and A test set was created with values from December 1st, 2018 to February 15th, 2019. **Table D** displays the actual and forecast values within the lower (80%) and upper limits (95%).

Table D. FORECAST OF LOG-DIFF(1) AR (2,0,1)

Point	*Lo 80*	*Hi 95*	*Train*	*Forecast*
2018.833	*-1.035*	*0.491*	*-0.432*	*-0.391*
2019.000	*-0.674*	*0.978*	*-0.021*	*-0.0765*
2019.139	*-0.878*	*1.19*	*0.159*	*0.03633*

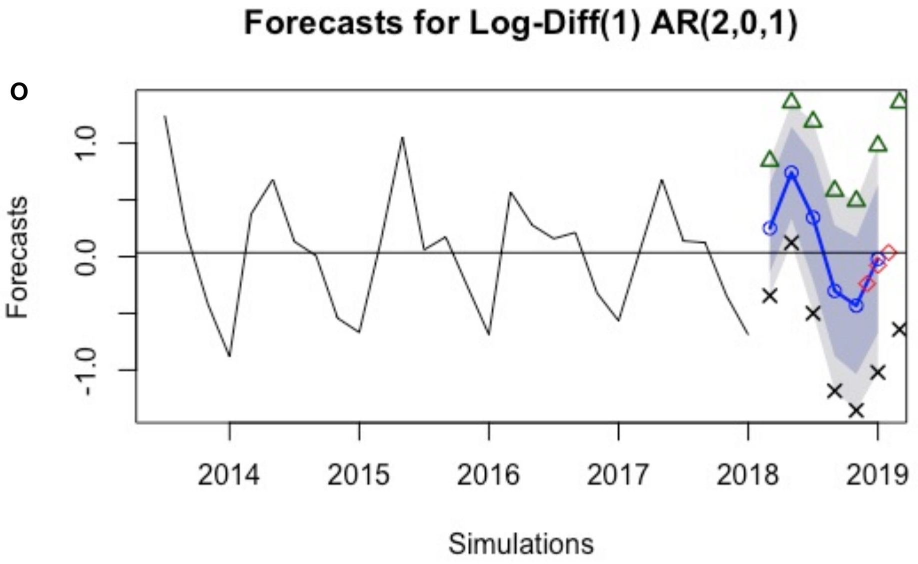

Figure 10 (O) Forecast using AR(2,0,1))

We summarize the forecast with the chosen model as:

Using the appropriate model AR(2,0,1), the forecast value can be found in between the low 80% and high 95% limits and thus confirm an appropriate choice for model and fit.

RESULTS AND CONCLUSION

In the introductory section, we were able to categorize the series, Citibike Trips per day (for All Users), as seasonal, with high values between May-October. In order to eliminate the positive seasonal trend and obtain stationarity, a transformed log-difference (1) was explored. Using optimization of the auto.arima () function, an appropriate model was chosen for the series, AR (2,0,1). It demonstrated the best ACF residual plot and competitive RMSE/σ2/AIC scores. An Ljung-Box test and QQ-plot confirmed its fit. A forecast was done on training and test data with the chosen model, and predicted values fell in between the low 80% and high 95% limits . This finding confirmed an appropriate choice for model and fit.

In conclusion, we state a successful analysis of a time series model and forecast. For future studies, we recommend a super-imposition using temperature between May 27th, 2013 to February 15th, 2019 and explore our findings to determine if temperature affects trips per day for all users.

References

1) Loaiza-Monsalve, D., and A. P. Riascos. "Human Mobility in Bike-Sharing Systems: Structure of Local and Non-Local Dynamics." Plos One, vol. 14, no. 3, 2019, doi:10.1371/journal.pone.0213106

2) Chung, Hangil, et al. "Bike Angels." Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) - COMPASS 18, 2018, doi:10.1145/3209811.3209866

3) Demaio, Paul. "Bike-Sharing: History, Impacts, Models of Provision, and Future." *Journal of Public Transportation*, vol. 12, no. 4, 2009, pp. 41–56., doi:10.5038/2375-0901.12.4.3.

4) Heaney, A. K., Carrión, D., Burkart, K., Lesk, C., & Jack, D. (2019). Climate Change and Physical Activity: Estimated Impacts of Ambient Temperatures on Bikeshare Usage in New York City. *Environmental Health Perspectives, 127*(3), 037002. doi: 10.1289/ehp4039.

5) Motivate International, Inc. "Citi Bike System Data." Citi Bike NYC, www.citibikenyc.com/system-data.

6) Fuller, Daniel, et al. "The Impact of Public Transportation Strikes on Use of a Bicycle Share Program in London: Interrupted Time Series Design." *Preventive Medicine*, vol. 54, no. 1, 2012, pp. 74–76., doi:10.1016/j.ypmed.2011.09.021.

7) Norby, R. J., Rustad, L. E., Dukes, J. S., Ojima, D. S., Parton, W. J., Grosso, S. J., . . . Pepper, D. A. (2007). Ecosystem Responses to Warming and Interacting Global Change Factors. *Terrestrial Ecosystems in a Changing World Global Change — The IGBP Series,* 23-36. doi:10.1007/978-3-540-32730-1_3