How does weather affect the use of Citi Bike in NYC?

Citi Bike is a bicycle rental service system around the state of New York since 2013. In order to optimize the Citi Bike business, we have to find out what are the factors influence the bike users. We found out that the bike users are highly influenced by the weather, so we analyzed the factors of bike users according to the weather.

The dataset of Citi Bike[1] is from Citi Bike main website[2]. The dataset of Citi Bike includes many factors, and among them, we used the time and date, and number of bikes per day for different user types (Customer=0, Subscriber=1). For subscriber's data, there provides gender information (1=male; 2=female; 0=unknown), which help in further analysis. The dataset is from September 2015 to October 2018. Then, we gathered weather dataset[3] from website National Oceanic and Atmospheric Administration (NOAA)[4]. From NOAA, the weather data is also obtained from September 2015 to October 2018 of New York State. The data includes various weather conditions, among them we chose AWND (Average wind speed in mph), PRCP (Precipitation in tenths of millimeters of water), and TAVG (Average temperature in Fahrenheit), which we considered those are most relevant factors in bicycle riding.

We used Python Jupyter notebook to combine the Citi Bike data with the weather data according to each date. Also, we created "Bike" column to show the total number of bikes used in each day. Then we used R to analyze statistical analysis. First, we used time series for each quarterly in 2017 for amount of bike used by hourly. Averaged the bike amount separated by weekdays and weekends for every 3 months. According to the time series of quarterly plot[5] of 2017, we can see weekends has always more people riding bikes than weekdays. The peak time would be between 7am-8am and 5pm-6pm, which can be considered is at the rush hour. Also, the total number of bikes at different quarters varies a lot. From this example, maximum of quarter 1

is at around 4000, quarter 2 is at around 12000, quarter 3 is at around 22000, and quarter 4 is at around 7000. From this result, we can assume that the weather affects the bike users a lot.

Now we use weather data to check how does weather indeed affect the Citi Bike users. We analyzed customers and subscriber separately to see how different types of users get affected by the weather. For both customers and subscribers used multiple linear regression to fit bike amount per day with weather predictors AWND, PRCP, and TAVG. However, customers residual plot[6] shows strong right tailed result, therefore we used log(bike) as response variable and refit the model. By doing so, model has been normalized, but could not avoid some heteroscedasticity in residual plot. For subscriber, their residual is more stable than the customer's residual, but still variance is not constant enough. It can be inferred as the subscriber uses bike more regularly. Therefore, we can draw the result: For customers, while PRCP and TAVG is held constant, as AWND (wind speed) increases by 1 unit, number of bikes multiplies by 0.94; while AWND and TAVG is held constant, as PRCP (precipitation percentage) increases by 1 unit, number of bikes multiplies by 0.55; while AWND and PRCP is held constant, as TAVG (average temperature) increases by 1 unit, number of bikes multiplies by 1.06. For subscribers, as other predictors kept constant, number of bike user will decrease by 26.3683 as wind speed increase by 1 unit; number of bike user will decrease by 187.5809 for every tenths of millimeters of water increase; number of bike user will increase by 13.3744 as the average of temperature increase by 1-degree Fahrenheit.

Since both customers and subscribers shows influence on weather, let's see which user type shows greater impact on weather change. we created an Indicator variable I: subscriber as 1 and customer as 0. We fit the model again as bike number as response variable and AWND, PRCP, TAVG as predictor variables, with I and interactive terms since there added an indicator variable.

After taking out not significant variables, variance would still be not constant enough. However, we still used this model to see the result. The final formula is: Bike = -59.133 + 1.934 * TAVG – 65.972 * I + 13.302 * TAVG*I. For subscriber (I=1), Bike = 125.105 + 15.236 * TAVG, which represents when every temperature increases by 1-degree Fahrenheit, number of bike users will increase by 15.236 per day. For customer (I=0), Bike = -57.776 + 1.934*TAVG, which represents when temperature increases by 1-degree Fahrenheit, number of bike users will increase by 1.934 per day. Since the total user number is larger in subscriber, the percentage of each users affected is: (15.236/749) *100 = 2.034% of subscribers (749 users per day) are affected; (1.934/54) *100 = 3.603% of customers (54 users per day) are affected. Therefore, customers have greater impact on the use of bikes at different weather, which make sense since subscribers ride Citi Bike more regularly compare to customers.

We also came up with a question if there shows a relationship between gender and weather condition in bike usage. For the subscribers each time they borrow a bicycle their gender is also recorded. The weather analysis is done with gender as Indicator variable: female as 1, male as 0, and also with the same predictors AWND, PRCP, TAVG with I and interactions, and same response variable, number of bikes. After removing not significant predictors, again with not constant variance, which should not be, the final equation is: Bike = -72.617 + 11.178 * TAVG + 26.497 * I – 7.533 * TAVG * I. Hence, for female (I=1), Bike = -46.12 + 3.645 TAVG; for male (I=0), Bike = -72.617 + 11.178 * TAVG. However, indicator variable “I” itself is not significant while all other interaction terms of “I” with other weather variables are highly significant. Therefore, we examined the fitted contrast of male and female bikes by having two different gender model fitted into the total bike plot[7] with weather TAVG as predictor variable. From this graph, we can see the male’s prediction line (blue) is above the female’s prediction line (red) with greater slope,

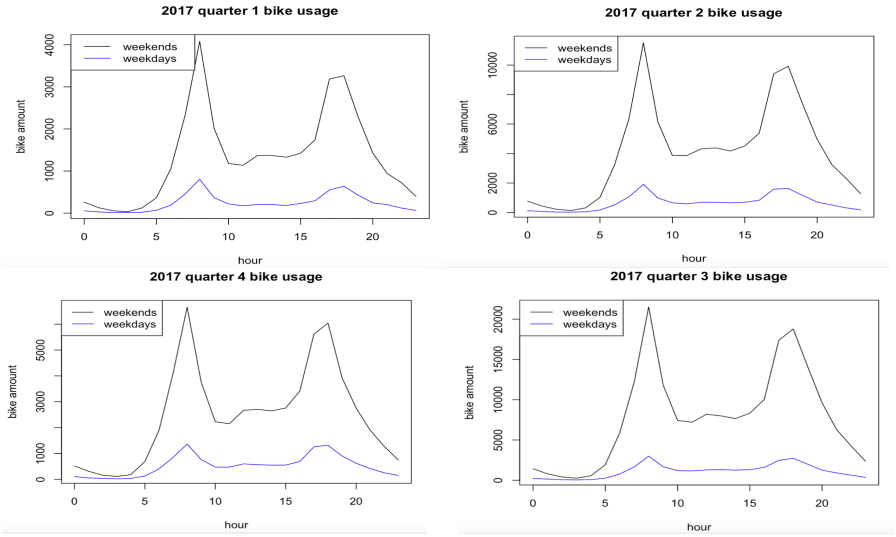
indicating that greater number of males gets more influenced by the weather. Gender itself does not show their significance, but since their interactive terms are significant, gender has roles when the weather comes in predictors. Therefore, when the weather gets hotter, the increasing rate of the bike users would be greater for males than for females.

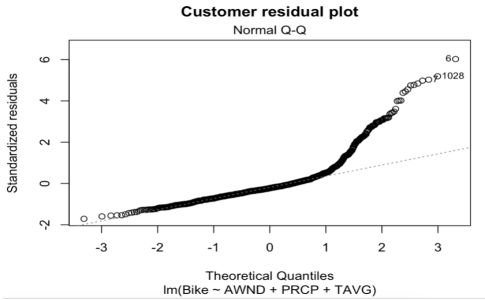
Consider how precipitation affects the gender in bike usage. In general, people will think that female would get greater influence on rainy day. Let's see if it really does. The model is fitted with number of bikes as response variable and PRCP, I, and PRCP*I as predictor variables. Variance is again not constant, but we ignored it. Since all of the variables are significant, we just used it to see the result. The result of female subscriber is shown as: Bike = 168.753 – 43.841 * PRCP; male subscriber is shown as: Bike = 585.898 – 132.795 * PRCP. Therefore, for female subscribers, the number of bike users will decrease 43.841 per day for every tenths of millimeters of water increase in precipitation. For male subscribers, the number of bike users will decrease 132.795 per day for every tenths of millimeters of water increase in precipitation. Since the male subscribers are greater than the female subscribers, we calculated the percentage of each gender get affected by precipitation: for female subscribers (163 users per day): 43.841/163*100=26.90%; for male subscribers (569 users per day): 132.795/569*100=23.34%. Therefore, female subscriber will get little more influence on the rainy day compare to the male subscriber in percentage, which agrees with our predictions. However, there violated the constant variance assumption in this model, so it could be wrong in some sense.

According to the analysis we made, weather temperature, wind speed, and amount of precipitation would give greater impact on Citi Bike users. In conclusion, customers get more affected by the weather comparing to subscribers. For gender comparisons, there are more male subscribers on the warmer weather. In addition, female subscribers get more affected by the

precipitation comparing to male subscribers. However, each analysis has somewhat heteroscedasticity in the residual plot. It would give better result probably using non-linear models. These analyses would be useful when optimizing Citi Bike services in NYC.

Reference

1. CSV files, JC-201509-citibike-tripdata.csv ~ JC-201810-citibike-tripdata.csv
2. https://www.citibikenyc.com/system-data
3. CSV file, 1558326.csv
4. https://www.ncdc.noaa.gov/cdo-web/search
5. 

6. 

7.