# SI 618 WN 2019 Final Project Report

# Sanghyun Lee (shleec)

## Motivation

In New York City, there is a bike share system called Citi Bike. As of October 2017, the system reached a total of 50 million rides and increased the number of bikes to 40,000 and stations to 706.[1] For the purpose of this project, I wanted to understand the Citi Bikers' pattern from 2017 to 2018. Every time I travelled to NYC, I was curious about the patterns of the bikers using the system. Therefore, I analyzed the differences between annual subscribers and customers, found out the most popular route and destination, and predicted the Citi Bike subscribers in 2019.

1. **How are rides different between Citi Bike customers and subscribers from 2017 to 2018?**
2. **Which was the most popular route (start to end station) for the customers and the subscribers for the last two years?**
3. **Which station was the most popular destination from 2017 to 2018?**
4. **Find out what features distinguish the Citi Bike subscribers and predict which user type it is using current data.**

## Data Source

*Citi Bike Trip History* is a public dataset produced by Citi Bike, which is New York City's current bike share system. The CSV data includes trip duration (seconds), date, station ID and latitude/longitude of both start and end station, as well as information about users' gender (0=unknown, 1=male, 2=female) and year of birth. Also, this dataset has user type information if the rider was a customer with a 24-hour or a 3-day pass, or a subscriber who is an Annual Member. The data is publicly provided and can be download from Citi Bike's official website (https://www.citibikenyc.com/system-data). Initially, I downloaded all 2017 and 2018 datasets, which are 24 monthly data. However, since each monthly dataset was over 50MB, I reduced the size and combined them into 2 dataset that are suitable for better and faster analysis. Overall, the dataset Citi Bike provided didn't have any missing or noisy data to handle.

Other than what we learned from SI618, I used GeoPandas, an open source Python module, to work with geospatial data. Also, for the purpose of plotting data on a map, I used open source shapefiles, which store location, shape and geographic features, from Jersey City Open Data website and New York City Open Data website.

Therefore, inside of *data* folder, you will see two CSV files of 2017 and 2018 *Citi Bike Trip History*, and shapefiles of Jersey City map and New York City map.

## Questions

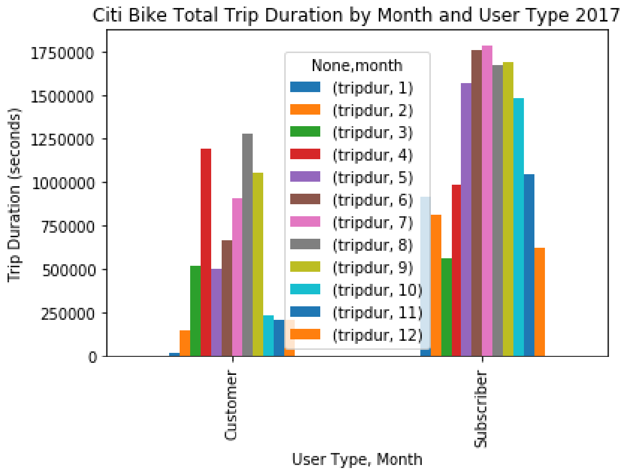
### 1. How are rides different between Citi Bike customers and subscribers from 2017 to 2018?

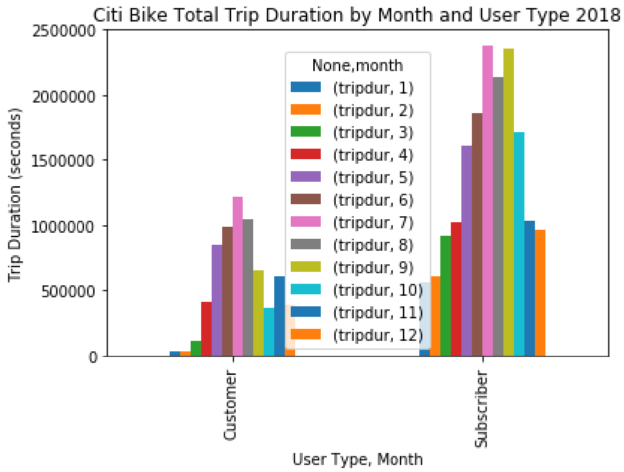
[1] "*Citi Bike*", Wikipedia https://en.wikipedia.org/wiki/Citi_Bike

## Method

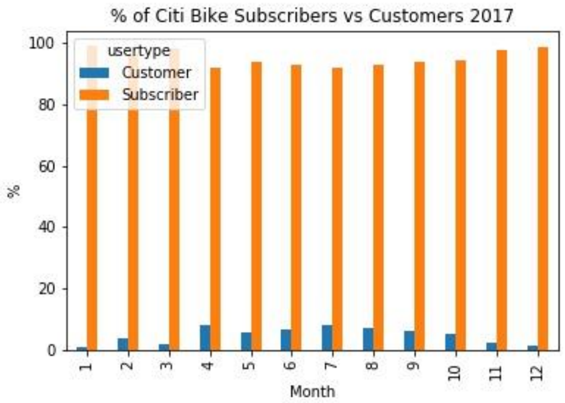
To analyze each 2017 and 2018 dataset, I dropped an unnecessary column, *Unnamed: 0,* which contained index numbers. Also, I had to change the column names that match to each dataset. Therefore, I lowercased and replaced the spaces with underscores. Then, I clustered customers and subscribers data under *usertype* column, and plotted boxplots using Seaborn to explore differences in monthly rides, yearly rides, total trip durations. The particular challenge I faced was to identify any missing or noisy data from in the datasets. However, after careful investigation, I couldn't find any data that could affect the results.

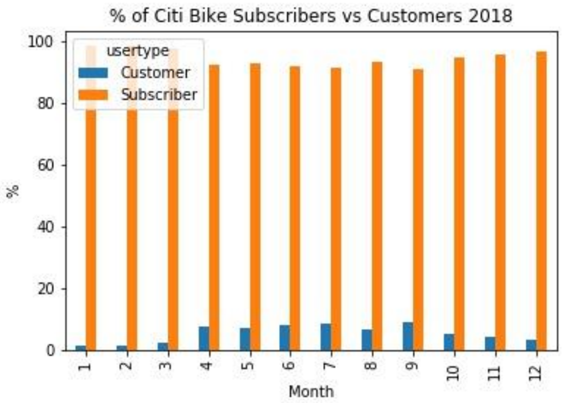
## Analysis & Results



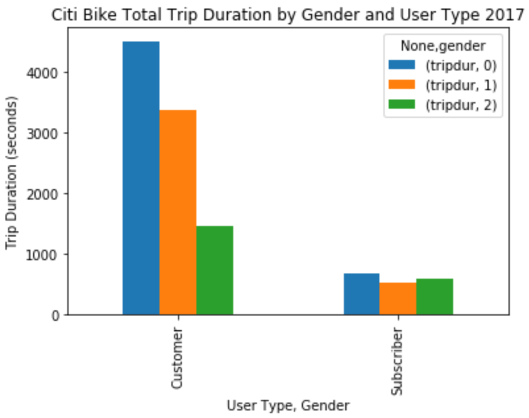


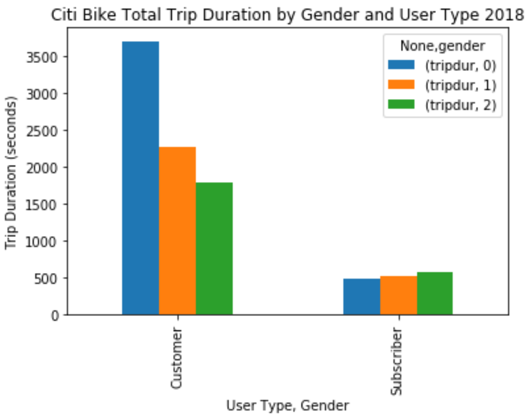
Comparing customers and subscribers from 2017 to 2018, there was significant increase in numbers of both customers and subscribers. In 2017, there were 27,846 subscribers and 1,638 customers. However, as of 2018, the numbers increased to 33,179 and 2,215, which are 19% and 35% increments respectively. Also, where a ratio of number of subscribers and customers in 2017 was 16.9:1, the ratio decreased to 14.9:1 in 2018.





Also, the average trip duration of customers was much longer than that of subscribers in both 2017 and 2018. And, the percentage of customers in total bike rides increased significantly from April to September. Therefore, from these boxplots, I was able to interpret it that the subscribers have a certain route they always go to and would take the most effective route to the destination, whereas the customers, most likely tourists, would use Citi Bike to explore around Manhattan or Jersey City.



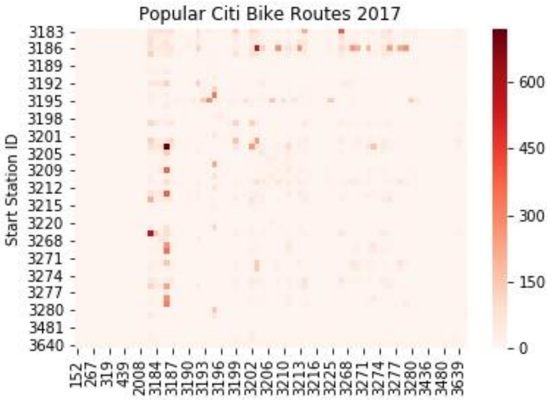


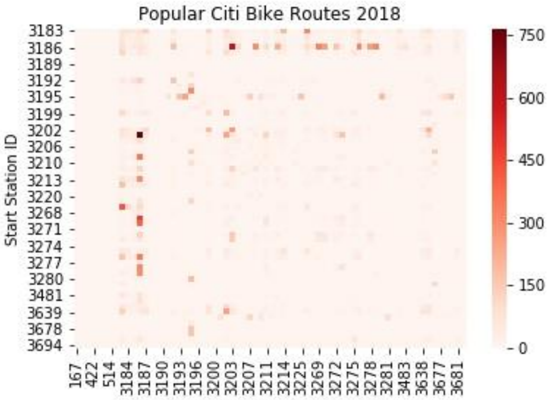
## 2. Which was the most popular route (start to end station) for the last two years?

### Method

Using station IDs, I clustered the start and end stations IDs and made heatmaps to figure out the most popular route. First, I joined the start station name and the end station name to make routes. Then, I counted the values of each path and figured out the most popular route. Then, I created tables with *start_st_id* and *end_st_id* using *crosstab* method to visualize the data in heatmaps. With heatmaps using Seaborn, I was able to understand the correlations between two different stations and analyze the most popular route. During this process, my biggest challenge was to visualize the heatmap with the most effective color palette. I experimented with many different colormaps, and I was able to identify that *Reds* was the most effective color to visualize the most popular route in heatmaps.

### Analysis & Results





As you can see from the heatmaps, there are just a number of distinguished routes that Citi Biker riders take in both 2017 and 2018. Also, these patterns don't really change as you can see the position of more red colored dots are in the similar location.

Also, when I counted the values of each route, the results of 2017 and 2018 were similar. As you can see from the tables, the most popular Citi Bike route for both years was from Hamilton Park station to Grove St PATH, which had 721 trips in 2017 and 767 trips in 2018. Similarly, the third most popular in 2017 and the second most popular in 2018 was from Grove St Path to Hamilton Park, which is the opposite direction of the most popular route. Therefore, I could interpret this result as that many Citi Bike riders from Hamilton Park, where my Jersey City resides make the best use of Citi Bike to get to Grove St PATH station.

| Popular Citi Bike Routes 2017 | |
|---|---|
| **Route** | **Trips** |
| Hamilton Park / Grove St PATH | 721 |
| Morris Canal / Exchange Place | 576 |
| Grove St PATH / Hamilton Park | 546 |
| Exchange Place / Morris Canal | 405 |
| Van Vorst Park / Grove St PATH | 390 |
| Brunswick St / Grove St PATH | 376 |
| Jersey & 6th St / Grove St PATH | 349 |

| Popular Citi Bike Routes 2018 | |
|---|---|
| **Route** | **Trips** |
| Hamilton Park / Grove St PATH | 767 |
| Grove St PATH / Hamilton Park | 597 |
| Brunswick & 6th / Grove St PATH | 471 |
| Morris Canal / Exchange Place | 422 |
| Jersey & 6th St / Grove St PATH | 395 |
| Brunswick St / Grove St PATH | 356 |
| Marin Light Rail / Grove St PATH | 343 |

## 3. Which station was the most popular destination from 2017 to 2018?

### Method

After exploring how I could plot geospatial data, which contains latitude and longitude coordinates of each Citi Bike stations, I decided to use GeoPandas. The first insight came from *GeoPandas 101: Plot any data with a latitude and longitude on a map* by Ryan Stewart. I first imported different shapefiles of maps that are suitable for plotting. Then, I dropped unnecessary information and only used trip duration, end station name, latitude and longitude to find out the most popular destination.
I used the latitude (*end_st_lat*) and longitude (*end_st_long*) coordinates of end stations to make a coordinate refence system using *Points* method from GeoPandas. However, this dataset contained some data with 0,0 coordinates, which was not correct. Therefore, I had to drop some of data for plotting. Also, I combined overlapping station names to find out the number of trips to the destination. Using the coordinate reference system, I plotted the exact point each end station on the maps of New York City and Jersey City. Then, I used the total number of trips data to visualize the most popular destination with sizes of each point on the scatterplot using Seaborn.

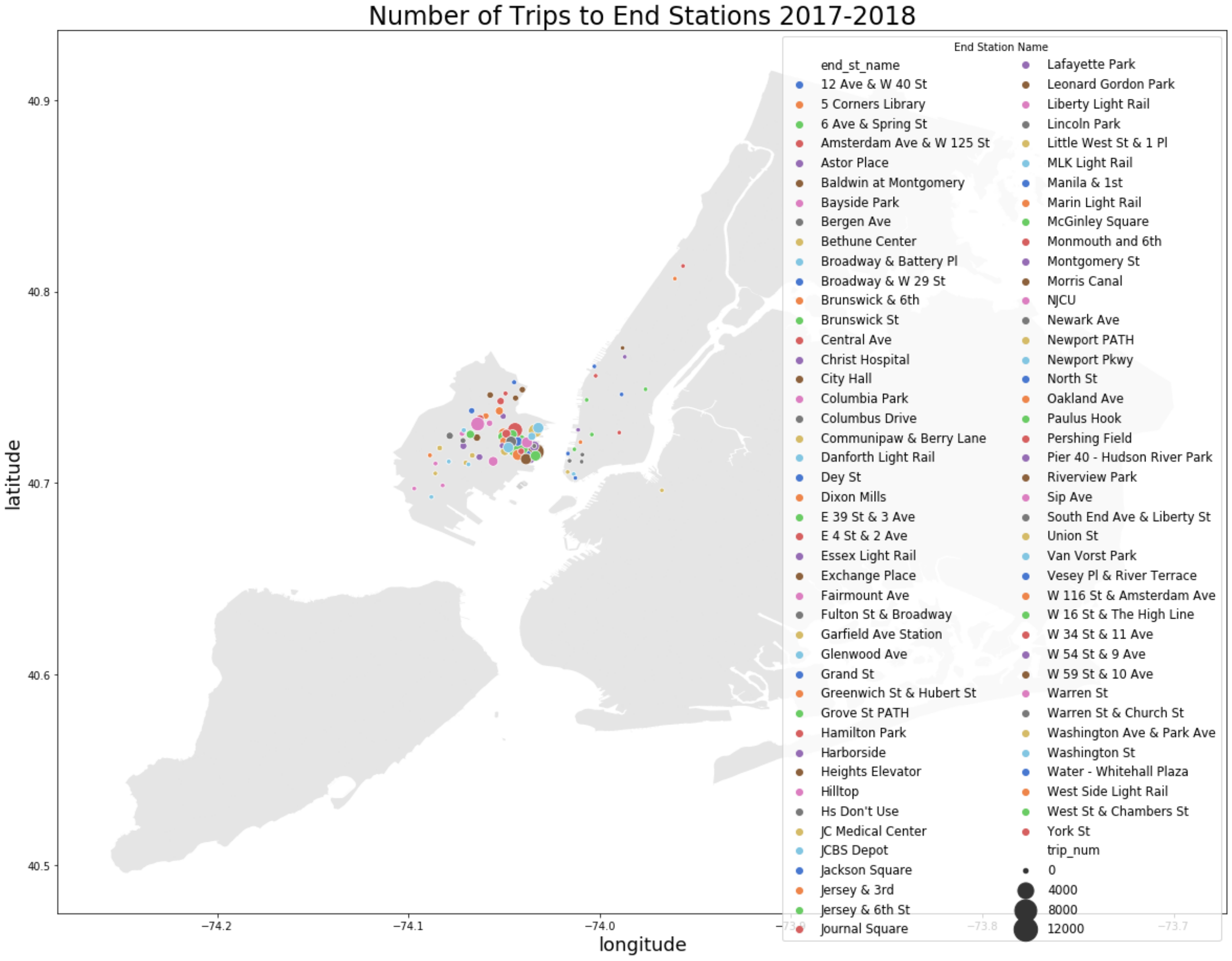
The biggest challenge I faced for this question was to research and find out the most effective tool to visualize the geospatial data. I spent most of my time researching than actually coding for this particular. However, from this experience I was able to find out a Medium post and able to learn about a new module, GeoPandas, which allow me to visualize and shapefiles, which contain location, shape and geographic features.

*In order to run my Jupyterlab, you need to install GeoPandas and Descartes.*

### Analysis & Results

The most interesting result I found from the scatterplot is that most of destinations were dispersed in Jersey City rather than in NYC. Also, the most popular destination from 2017 to 2018 was Grove St PATH station in Downtown Jersey City, which had total of 9,199 trips and 69,663 minutes to get to the station. This was a surprise, because Citi Bike has recently expanded the stations in Jersey City and there are much greater number of stations in New York City. Also, another interesting insight I got was from the result of the second most popular destination, Exchange Place. Since both Grove St PATH station and

Exchange Place station are near Red lane of Path train stations, I could assume that most riders made the trips to cross Hudson River to go to Manhattan or travel to Newark.

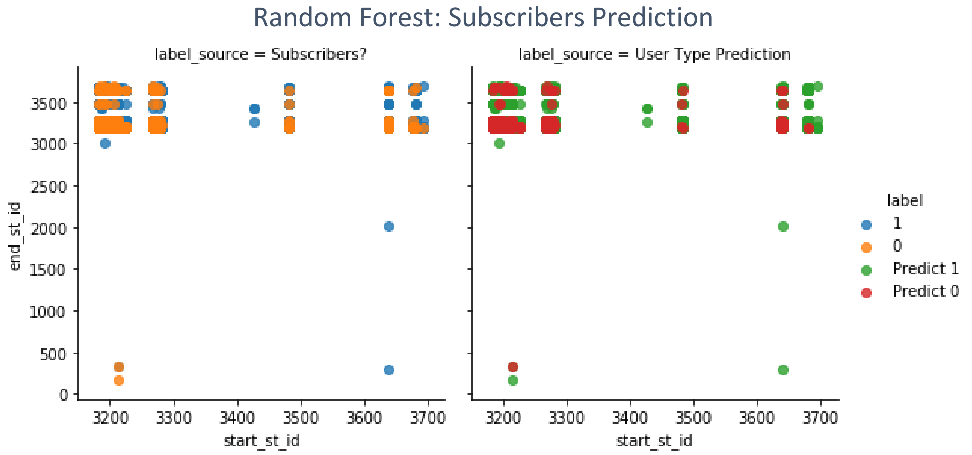


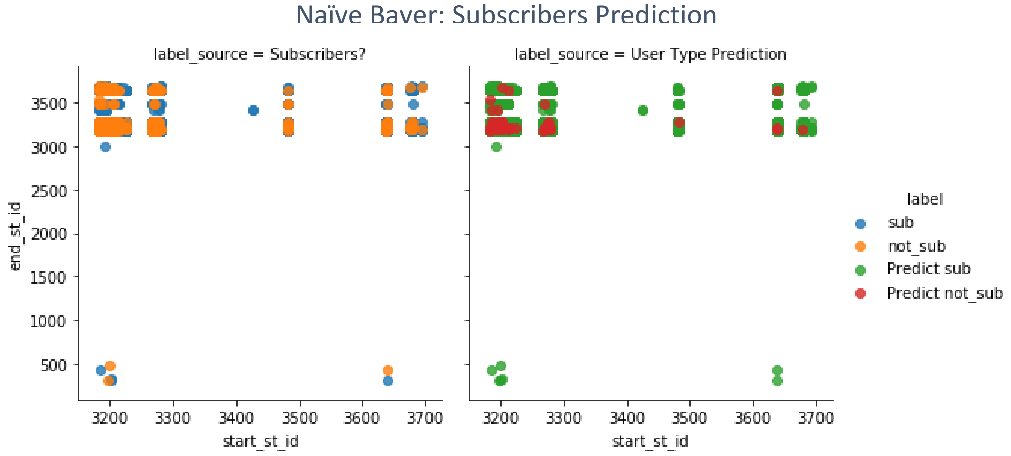
## 4. Find out what features distinguish the Citi Bike subscribers and predict which user type it is using current data.

### Method

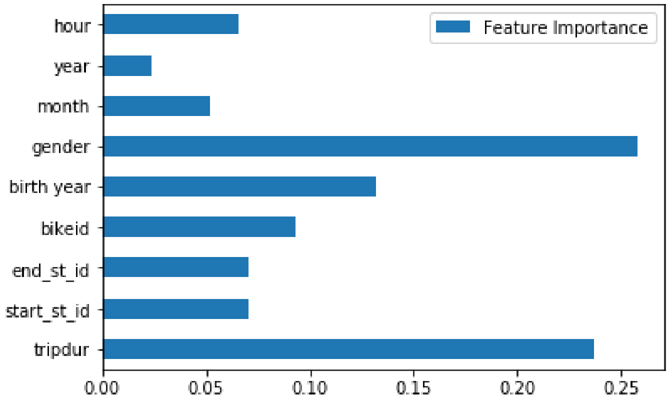
In order to found out the features that affect the prediction of Citi Bike subscribers, I concatenated 2017 and 2018 datasets into one. Then, I broke down the date and time into months, years and hours to create more variables for effective prediction. I dropped ids and coordinates of start station and end station, because they are pretty much the same data, and created dummy data to train and to test. With the dummy data, I build both Random Forest and Naïve Bayer classifiers to predict the subscribers variable. I also experimented with changing the fold in my cross validator for Random Forest classifiers for more precise prediction results. Then, based on my Random Forest classifier, I looked at the features that have affected the most. In this particular question, my biggest concern was to tune the  k-folds and depth numbers for the best result. However, after experimenting with *best_params* method, I was able to find the best hyperparamers.

## Analysis & Results


Random Forest: Subscribers Prediction


Naïve Bayer: Subscribers Prediction

After experimenting with k-folds and max-depth using cross-validator for my Random Forest classifiers, I used *max-depth* of 9 and *n_estimators* of 20. With Random Forest and Naïve Bayer classifiers, I was able to predict the subscribers at an accuracy of 97.6% and 96.0% respectively. The importance features that affected the predictions were gender and trip duration. After understanding the patterns of both customers and subscribers from previous questions, I was able to understand why two features are important. In the question 1, the boxplots of "*Citi Bike Total Trip Duration by Gender and User Type"* shows that the trip duration of male Citi Bike customers traveled as twice more than female customers. However, the trip duration was as identical for both genders in subscribers. Therefore, among variables like *hour*, *year*, *birth year* and stations, I could understand that *gender* and *tripdur* variables are the most important features to predict the user types of Citi Bike riders.


Feature Importance
hour
year
month
gender
birth year
bikeid
end_st_id
start_st_id
tripdur