Springboard Data Science Intensive  Winter 2016-2017

**36,000,000 Bike Rides: An Analysis of Several Key Metrics of New York City’s Bike Share System, Citi Bike**

Hilda Vargas

**Abstract**

*Citi Bike is the largest bike-sharing program in the United States, operating in many parts of New York City including Brooklyn, Manhattan, and Queens. Since it opened to the public in July of 2013, there have been over 30 million trips taken with Citi Bike and more than 600 stations added all over New York City. This paper serves as a guide to my project, in which I created an interactive dashboard using python to analyze specific key metrics related to Citi Bike trips. These metrics include the most popular stations, average trip duration per specific age groups, and the number of trips taken per gender, and my dashboard was created to help guide the advertising decisions made by my hypothetical client, Citi Bike’s marketing team.*

I. Introduction

**W**ith the rise of ridesharing companies like Uber and Lyft over the last couple of years and a public transportation system that many perceive as unreliable, Citi Bike’s bike share program offers a healthier, environmentally friendly alternative to New York City’s current methods of transportation. Citi Bike allows users to ride rented bicycles from one station to another, and it offers two different subscriptions for those who ride bicycles occasionally and those who rely on bicycles for transportation.

Despite the fact that Citi Bike publicly offers trip data on their website, there has been little public research done on the demographics of its users and other metrics integral to the company’s success. I believe that through the use of data wrangling and visualization techniques on Citi Bike’s data, there are patterns in these metrics that can guide decisions regarding the target audience and content of the company’s marketing campaigns.

Citi Bike’s dataset contains several fields that can be employed to answer specific questions about its ridership. The dataset includes detailed information for each bike ride taken from July of 2013 (the company’s opening to the public) to December of 2016, including the trip duration, start station name and coordinates, the gender of the user, and the birth year of the user. I will use these fields to create interactive plots in which my hypothetical client, Citi Bike’s marketing team, can choose a specific date range by month(s) and year(s) to examine a specific metric of their choice. I will seek to analyze the following metrics through a set of interactive Jupyter notebooks that together form a dashboard for Citi Bike:

1. Where are Citi Bike stations located in New York City?
2. What are Citi Bike’s 10 most popular stations?
3. What are the number of trips taken per age range and gender?
4. What is the average trip duration per age range and gender?
5. What are the most popular times of the day for ridership?

II. Related Work

Prior projects like “A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System” [1] have set great examples of applying data visualization techniques to Citi Bike’s data to observe popular routes taken by riders over time on a given day and machine learning techniques to measure weather’s impact on ridership. The author of “Analyzing NYC Biking Data with Google BigQuery” [2] utilized similar methods to visualize the most popular stations on a heat map and track the number of trips by age. However, neither of the authors created graphs that are interactive by giving the user choice over the time period in which they would like to study a specific metric. I believe that this is a big oversight as it relates to Citi Bike’s marketing department because there are patterns that may be observable over specific months and years that are not included in these prior projects. Additionally, the two aforementioned projects failed to analyze the average trip duration for user demographics, which could also be very useful to my client. It should be noted, however, that neither of the projects were intended for Citi Bike’s use while my project was created for the company itself.

III. Data

Citi Bike’s raw data contains 42 files, one for each month and year from July 2013 to December 2016, each with 15 fields in the following order:

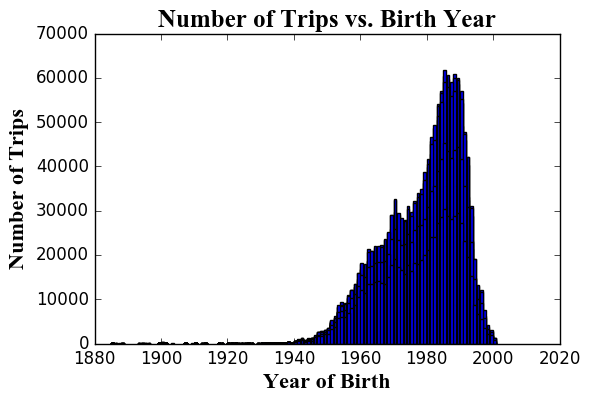
* Trip Duration (seconds)
* Start Time and Date
* Stop Time and Date
* Start Station ID
* Start Station Name
* Start Station Latitude
* Start Station Longitude
* End Station ID
* End Station Name
* End Station Latitude
* End Station Longitude
* Bike ID
* User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)
* Year of Birth
* Gender (Zero=unknown; 1=male; 2=female)

III.I. Data Wrangling and Cleaning

Because there have been 36,902,025 Citi Bike rides as of December 2016, the dataset was very large and difficult to work with. At first, my approach was to concatenate all of the files into one large pandas dataframe, but this proved to be largely inefficient and caused my computer to crash several times due to memory storage. After speaking with my mentor,

I made the decision to create 42 files with smaller subsets of the data from the original 42 files that would be easier and faster to work with. I then created one new CSV file for each question with the test data necessary to answer it. For example, for the question regarding the number of trips per age range and gender, I counted the number of trips per each unique pair of birth years and genders in each file and sent the results of all 42 files to one new file. This final CSV file with the consolidated data contained the month and year in which the trips were taken, the birth year and genders of the users, and the number of trips for each specific combination of the populated fields. This process was repeated with the large raw files for all of the remaining questions except that of the locations of the current Citi Bike stations, as the most recent data available (December 2016) was used to answer that question.

After exploring the dataset, there were a few more issues to address. One problem was that some users declined to provide their year of birth. For questions involving age groups, these missing birth years were converted to nulls and subsequently dropped from the dataset. Another problem was that some users provided birth years that were clearly not real (birth years in the 1800’s), presumably because they had privacy concerns or made human errors. In an effort to address this, I created a histogram of the birth years to see the distribution of the variable and find outliers. The plot depicted birth years of 1940 and below as outliers and those below 1920 as extreme outliers. I chose to drop the trips of people with birth years that were extreme outliers (<1920) and analyze the remaining observations.



**Figure 1:** Histogram of birth years employed to find outliers in the variable.

Another data wrangling challenge I faced occurred with parsing dates. There were at least three different formats in the 42 files for the time at which the trips were started. Some of the dates positioned the year first, others positioned the month first, and others had different formats for the hour, minutes, and seconds. I was able to split the times so that I was only left with the hour, minutes, and seconds at which the trips were taken, and use dateutil’s parse function to correctly extract the hour from each of the different time formats and solve this issue.

Finally, in order to make my graphs less cluttered, I had to create ranges for users’ ages because there were about 80 unique years of birth. In order to do this, I used panda’s cut function with a list of specific bins I created that encompassed the full range of ages observed in the data.

IV. Methods

To create the dashboard, I used specific libraries depending on the question at hand. In this section, I will highlight the specialized python packages that were used for the data visualization aspect of this project.

IV.I. Folium

For the questions pertaining to Citi Bike’s stations, I experimented with several python packages, namely Bokeh. Although Bokeh is very user-friendly and has thorough documentation on its website, it is a very young library that was started only a couple of years ago and still has many bugs in its system. In search for an alternative with sleek maps and intrinsic interactive widgets, I landed upon Folium and decided to use it due to its simple syntax. When utilizing Folium, I simply iterated through the rows of the relevant dataframe and created tuples for the coordinates and the pop-up text for each marker on the map, the station name.

IV.II. Creating Plots with Seaborn

The remaining questions I answered for this project concerned statistical metrics for Citi Bike, namely the average trip duration per age group and gender, the most popular times of the day for ridership per gender, and the number of trips per age group and gender. For each of these questions, I created two plots with one of python’s statistical data visualization libraries, Seaborn —one point plot with the ungrouped data and one bar plot with the data broken up into the relevant categories needed to answer the question. Each bar in the bar plot is a different color, representing each gender available in the data (unknown, female, male).

IV.III. Adding Interactors with ipywidgets

In order to give the user choice over the time period under which a given metric is measured, I used python’s ipywidgets package, which contains interactive HTML widgets for Jupyter notebooks and the IPython kernel. Interactive widgets allow one to visualize the changes in data, and it seemed necessary to incorporate them into my dashboard.

For each plot and the one map illustrating the 10 most popular stations, I created two widgets, one for every month (January-December) and one for every year in the dataset (2013-2016). I then defined a function to command the graphs to solely plot data from trips that occurred during the user’s chosen month(s) and year(s). The final steps in this process were to style the graphs and instruct the widgets to interact with my defined function through ipywidgets’ “interact” function.

I chose to use the multiselect widget for all the graphs in my interactive dashboard because it allows the user to select one or more of the provided options in each of the widgets.

V. Results and Recommendations

V.I. Current Citi Bike Stations and the 10 Most Popular Stations

From the map visualizing the locations of current Citi Bike stations, it is quite easy to see that there are many Citi Bike stations, over 600, dispersed all over the city. However, if one plays around with the widgets on the map of the 10 most popular Citi Bike stations, it is obvious that there are several stations that have been consistently popular since Citi Bike’s inception. These stations are generally located near Grand Central (42nd street), Chelsea, the East Village, and Tribeca, which intuitively makes sense because these are very popular areas in New York City for renting apartments and office spaces and nightlife.

However, curiously, in all of the month/year combinations that I experimented with, none of the top 10 stations were located in Brooklyn. This is surprising to me because Brooklyn is generally less congested than Manhattan, and it is also a popular borough for apartment and office rentals. There is a chance that the reason none of the stations are very popular has to do with the level of bike owners in Brooklyn compared to Manhattan, and I think that would be a great extension to this research. Moreover, I found it interesting that the popularity in the station located at 60th street and Broadway decreased in popularity, while the station located on 40th street and 12th avenue increased in popularity between 2013-2016. There might be many reasons for these findings, including an increase/decrease in ownership of bicycles or bike-renters moving to other parts of the city, but again, the causes of these patterns should be investigated in another project.

Based on my findings, I would suggest that my client focus on marketing and advertising in Brooklyn and upper parts of Manhattan. Moreover, I would recommend that they consider creating a promotional campaign to increase ridership in the less popular areas.

V.II. Number of Trips per Gender and Age Group

The plot depicting the number of trips per age range consistently shows that users who are between 31-35 years old take the overwhelming majority of rides. People who are 26-30 years old take the second most trips, and those who are 36-50 years old take the third most trips and consistently have for the last couple of years. As expected, the number of trips increased between the months of March-August, and decreased from November-February.

Between 2013 and 2015, few rides were taken by millennials aged between 21-25. In fact, users between the ages of 36 and 50 years old took 2-3 times more rides than riders aged 21-25. However, in 2016, the number of rides taken by millenials nearly equaled the rides taken by those who are 36-50 years old.

Interestingly, the 16-20 age group accounts for very few rides. In the three and a half years that the dataset encompasses, people in this age range took approximately 300,000 rides. This may be due to parents’ safety concerns and alternatives that may seemingly be safer for the younger demographic. However, this observation surprised me.

The bar plot illustrating the number of rides taken per age group, grouped by gender depicted a large disparity in the rides taken by men and women. In every month of every year, men accounted for at least two times as many rides as women. As a matter of fact, women have only taken approximately 8 million rides out of the 36,902,025 total trips taken in July 2013–December 2016. This is less than 25% of the total Citi Bike trips taken, assuming that people told the truth about their gender. Because so few users’ genders are unknown, we can easily estimate that the other ~75% of rides were taken by men.

With regards to the client, I think that the takeaways from this analysis are very clear. First, the content in the advertising in Citi Bike’s campaigns should be targeted to more millenials and kids in their late teens. The number of trips for people aged 36-50 is strong, but it would not hurt to direct more advertisements to people in that age range or above. Finally, Citi Bike’s marketing team should focus on increasing ridership for women and understanding why men take so many more rides than women. Once the root of this cause is understood, the marketing team can make better decisions of how to market to female bike riders.

V.III. Average Trip Duration per Gender and Age Group

The plot for average trip duration per age, ungrouped is noteworthy because the age groups with the highest average trip duration were 16-20 and 75+. This is most likely due to the fact that there were so few rides taken by people in these age ranges. If one person claiming to in their 80’s took a trip that was 80 minutes long (March 2016), and there were no other trips taken by people in that age range, then the average at that time for that age range would remain at 80 minutes.

As expected, the average trip duration for all age groups increased over the spring and summer months, but only by 1-3 minutes, which was not expected. I hypothesize that this might be due to the fact that it gets very warm in the summer months, and biking may not be ideal. Further research should be done on the impact of weather on ridership to truly know the cause of this.

I was quite surprised that on average, Citi Bike trips have only been 12-16 minutes long, historically. This might be because most people use the bikes to go to work and not for leisurely activities, but again, another project should be done on that question.

With regards to average trip duration per gender, I was fascinated to learn that women’s trip durations are longer than men for almost every age group. Although the difference is quite small, I was surprised because of how many more rides have been taken by men than women.

I believe that Citi Bike could use this data to create promotional campaigns with special subscriptions for women. For example, now that we know that most trips taken by women are about 15 minutes long, the marketing team could attempt to increase ridership by offering a certain amount of minutes for free for every 15-minute bike ride. This decision would have to be made depending on other business metrics, but it is just one application for the information available in the plot.

V.IV. Most Popular Times of the Day for Citi Bike Use per Gender

My final plot concerns the most popular times of the day for Citi Bike use. The company could use this information to increase ridership during times of the day that are less popular for riders by providing discounted prices for bike rentals during these time periods.

Unsurprisingly, most Citi Bike trips are taken between 6 AM – 10 AM, corresponding to the times at which many arrive to work, and 4 PM – 9 PM, the interval at which many go home from work. Moreover, as expected, more rides were taken at later times of the day in the warmer months, while this number decreased in the colder months. These observations held true for both the ungrouped point plot and the barplot grouped by gender.

VI. Conclusion

Through the comparisons of the

various plots included in my dashboard project, it is clear that Citi Bike’s marketing team could benefit through its use by examining the patterns that exist in the metrics I explored and creating market campaigns according to these patterns. Specifically, Citi Bike should aim to increase ridership among women and millennials, which they can do by providing incentives based on the trip durations of these groups and creating advertisements that target them. They should also aim to increase ridership in less popular areas and during less popular times of the day, which they can also attempt to do through promotional campaigns. However, in order to make those decisions, more data is necessary to ensure that the promotions are viable and likely to increase ridership.

Some future work that can be done to benefit my client includes:

* Predicting the impact of weather on ridership per gender and/or age group
* Gaining a better understanding of the demographics of Citi Bike users by tying in census data such as income, occupations etc.
* A look at the most popular routes amongst different demographics of Citi Bike users

**References**

[1] T. Schneider. “A Tale of Twenty-Two-Million Citi Bike Rides: Analyzing the NYC Bike Share System.” Internet:

<http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/>, Jan. 13, 2016 [Mar. 28, 2017].

[2] S. Robinson. “Analyzing NYC Biking Data with Google BigQuery.” Internet: <https://cloud.google.com/blog/big-data/2016/12/analyzing-nyc-biking-data-with-google-bigquery>, Dec. 7, 2016.

[Mar. 28, 2017].