Project Book

Group H - Citi Bike Usage in NYC

# Overview and Motivation

Citi Bike is the nation’s largest bike share program, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens and Jersey City. It was designed for quick trips with convenience in mind, and it’s a fun and affordable way to get around town.

In this project, we’re going to conduct a case study about Citi Bike usage. We’ll be looking at the open source Citi Bike dataset as well as NYC traffic and transportation data and Twitter data, and try to visualize the growth of Citi Bike in terms of usage, expansion and membership, the daily using pattern, different factors affect bike usage, and the image of Citi Bike on the social media.

# Related Work

There have already been several visualizations and studies about Citi Bike on the Internet, since the data is open source encouraging everyone to use for study. Here are some previous works about Citi Bike on the Internet:

* <https://member.citibikenyc.com/map/>
* <http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/>
* <http://www.newyorker.com/news/news-desk/interactive-a-month-of-citi-bike>
* <http://radar.oreilly.com/2013/07/interactive-maps-bike-share-new-york-washington-citibike.html>
* <https://blog.nycdatascience.com/student-works/bikes-go-analysis-nyc-citi-bike-station-capacity/>

However most of the existed studies focus on the Citi Bike usage data itself and are on a daily basis, some of them also have efforts on the user profile. While what our study try to do, based on the existing ones, is studying Citi Bike in a historical growth perspective, combining other related dataset to look at Citi Bike as one of the elements in NYC environment, and looking into social media to study the social image of Citi Bike.

# Questions

The questions we want to address in this study are:

* What is the pattern (miles, time, routine, etc.) of Citi Bike usage in the daily, quarterly and yearly perspective?
* What is the history of Citi Bike since its establishment? How is the growth of numbers of users and members? How did it expand over the city?
* What is the role of Citi Bike in NYC daily life? What is the relationship between Citi Bike and other transportation?
* What is the image of Citi Bike on the social media? What are the related topics? Is it positive or negative? Why?
* Is riding Citi Bike safe in NYC?

# Data

We apply the Citi Bike System Data, which is an open source database provided by Citi Group, containing Citi Bike trip histories, ridership and membership data and other additional relevant resources.

The trip histories data include following variables:

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

The ridership and membership data include:

* Trips over the past 24-hours (midnight to 11:59pm)
* Miles traveled today (midnight to 11:59 pm)
* Total Annual Members (All Time)
* 24-Hour Passes Purchased (midnight to 11:59 pm)
* 3-Day Passes Purchased (midnight to 11:59 pm)

The City of New York’s bicycling data include:

* Locations of the City’s outdoor bicycle racks

In addition, we use Twitter data to do text analysis about Citibike. We search tweets that mention Citibike in the last two weeks of April via Twitter API and apply frequency analysis, sentiment analysis and word clouds to describe the data, analyze user’s opinions towards Citibike and look for other interesting facts.

All the datasets of Citibike can be downloaded in csv files in a clean form. Tweets can be downloaded in R via Twitter API. The potential obstacle in data processing may be the streaming and processing of tweets data since the tweets can be very messy with tags, abbreviations and unrelated information. And we want to get the geographic information of every tweet to get the location of that bike trip, which we are not sure if available.

# Exploratory Data Analysis:

*How did you start exploring your data? What did you find out? How did that help you structure the next steps?*

We started exploring data by …..

# Design Evolution:

*What were some earlier steps in the visualizations you considered? Which perceptual and design principles did you take into account when choosing the eventual 'winners' - i.e. the final designs you ended up with?*

At first we …..

# Implementation:

*Give some idea about the intent and functionality of the interactive visualizations you used in the final output.*

# Evaluation:

*What did you learn about the data by using your visualizations? How did you answer your questions?*

As for the relationship between Citi Bike and subway part, we found a strong correlation between the Citi Bike usage and subway station density, however due to the lack of controlling variables as well as time series analysis, the strong spatial relationship is what we conclude, and we cannot assert that there is any causal effect between the two factors.

As for the Citi Bike on social media, we get fully knowledge about the Twitter text data we got as it is rather small and not 100% ideal for sentiment analysis. We managed to give the answer for the questions we asked in this part, nevertheless it is not perfect and further study may be needed to polish the answer.

# Next Steps:

*How well does your visualization work, and how could you further improve it? Which things could you not do because of limitations of data, technical difficulties, time constraints etc.*

For the relationship between Citi Bike and subway part, it is more like a preliminary study because we only checked the correlation, but did not reveal the causal effect between the variables. Since midtown and downtown are the busiest place in NYC, the correlation may be due to some kind of other spuriousness. The next step of this study is to control more variables and consider the potential overlap of peak time of both the Citi Bike and subway.

For the Citi Bike on social media part, due to the restriction of Twitter Search API, our tweets sample size is comparatively small, we may look for larger and more sufficient data set for further study. The next steps for more findings include combining it with geographic data as well as looking into difference data source like Yelp review of Citi Bike racks/stations (if it exists).