# Understanding Citi Bike: Data Visualization and Exploration in Tableau and Python

BY C. DANIEL GUETTA*

## Introduction

Citi Bike is a privately owned public bicycle sharing system in New York City and its surroundings. Over 700 stations located in Manhattan, Brooklyn, and Jersey City play host to over 12,000 bikes, making it the largest bike sharing system in the United States. Users sign up for long- or short-term memberships, and use these membership to unlock bikes at a Citi Bike station. Riders can purchase one- or three-day passes to the system, or can subscribe to the system annually. Trips below a certain duration are free (45 minutes for annual subscribers, and 30 minutes for others), and longer trips incur additional fees.

The first documented bike sharing system began in Europe in 1965. Provo was a Dutch counterculture movement that focused on provoking violent responses from authorities using non-violent bait. One of their schemes was to paint 50 bikes white and leave them on the streets of Amsterdam for public use. The bikes were quickly impounded by the police; they violated local laws prohibiting the leaving of bikes on streets without locking them. Provo quickly returned them to the streets equipped with combination locks, and painted the combinations on each bike.

Since then, bike sharing technology has evolved considerably. The first mainstream, large-scale bike sharing system was launched in Copenhagen in 1995. It comprised a system of 300 bicycles that could be unlocked with coins, as with shopping carts in supermarkets. Copenhagen's system has grown considerably since its launch, and many other cities have adopted bike sharing systems of their own. As of 2016, almost 1,000 cities worldwide had bike sharing systems.

In 2008, the New York City Department of Transportation released a strategic plan on alternative transportation methods, in an effort to reduce emissions, road wear, collisions, and road and transit congestion. One of the report's observations was that 56% of all automobile trips within the city were under 3 miles, with 22% under 1 mile and 10% under 0.5 miles – all distances easily ridden on a bicycle. The city later decided to create a bicycle share program,

**Author affiliation**
* *Columbia Business School*

following other cities' lead, and the system was eventually launched in March of 2013. The system is not funded by any subsidies from the city; Citigroup spent $41M to be the system's lead sponsor for six years (and in return was allowed to put its name on the bikes). In 2014, Citigroup injected an additional $70.5M to extend its sponsorship to 2024.

Since its establishment, Citi Bike has grown tremendously, and become part of the fabric of New York City. In October 2017, the 50 millionth trip of all time on a Citi Bike was recorded. Despite some criticism of the system (mostly around the impact on local bike shops, the look of the bikes, and the physical space occupied by the stations), it is has continued to grow – most recently introducing electric pedal-assist bikes, and dockless bikes in the Bronx.

The system's growing popularity and exponential growth has come as a welcome development to those who run it. It has, however, also led to some increasingly difficult logistical challenges, as well as a strong need to better understand Citi Bike's ridership and the way they use the system. The data underlying these trips, however, is very large (comprising tens of thousands of trips every *day*), and this makes it very difficult to efficiently analyze. In this case, we will use modern data visualization and exploration tools to address the challenge and better understand the Citi Bike system.

## The Data

We will be basing our exploration on data that Citi Bike graciously makes available on its website. In particular, the data provided contains one row per trip taken on the Citi Bike platform, and the following columns

- The trip duration, in seconds
- The start and end date and time of the trip
- The start and end station name, ID, and latitude and longitude
- The bike ID
- The user type; either "customer", referring to a rider who has purchased a one-day or three-day pass on the platform, or a "subscriber", referring to a rider who has purchased an annual Citi Bike subscription
- Gender (0 for unknown, 1 for male, 2 for female)
- The year of birth of the rider, if known

The files provided separate trips taken in Jersey City and in the rest of the system. Given the scale of trips taken on the platform, the scale of the data is very large. We will focus on data from a single month (May 2018), which provides dataset that is manageable but large enough to be difficult to analyze and to provide key insights into the system's operations.

## Citi Bike's Ridership

Our first challenge will center around Citi Bike's ridership. The goal of the system, first and foremost, is to provide a useful service to its riders. To that end, it is important for the operators of Citi Bike to understand their ridership to be able to tailor the system.

Specifically, based on this data, consider the following questions:

- What is the gender balance in the system? Does this differ among subscribers and customers?
- How long, on average, are the trips on the platform? Does this differ among subscribers and customers? And by gender? On the week and weekends?
- Does age have an appreciable effect on ride length? How reliable would you say these results are?

Based on your investigations, what patterns do you observe? Are there any changes you might make to the policies or operations or a platform as a result of those observations?

## The System's Use Throughout the Day

We will next analyze the use of Citi Bikes throughout the day. This analysis is key to understanding how the system is used and therefore to correctly allocating resources throughout the system.

Specifically, based on these data, consider the following questions

- How is system usage (as measured by number of rides) distributed throughout the day? If you had to take the system offline for a short period of time, when should you do it?
- Are these patterns different on weekdays and weekends?
- Are these patterns different for subscribers and customers?

## The System's Use Around the City

Finally, we will look at the system from a geographical perspective. One of the most difficult decisions for any bike-share system operator is how to expand the system – where to put new stations, for example. A geographical analysis is essential to answering these questions.

Specifically, based on these data, consider the following questions

- What are the most popular stations in the city for pick-up? What about drop-offs?
- Are some stations more popular at different times of day? With subscribers vs. customers?
- How often is Citi Bike used to take trips across boroughs? What are the most popular borough combinations?

- Citi Bike, like most other bike sharing systems, engage in rebalancing. Trucks are used to carry bikes manually from areas of high bike density to areas of low bike density. What stations are most involved in rebalancing? Which stations receive the most bikes, and which stations have the most bikes taken from them?