Citi Bike New York City Data Analysis

About Citi Bike

Citi Bike is New York City's bike share system, and the largest in the nation. Citi Bike launched in May 2013 and has become an essential part of the city's transportation network. It's fun, efficient and affordable – not to mention healthy and good for the environment.

Citi Bike consists of a fleet of specially designed, sturdy and durable bikes that are locked into a network of docking stations throughout the city. The bikes can be unlocked from one station and returned to any other station in the system, making them ideal for one-way trips. People use bike share to commute to work or school, run errands, get to appointments or social engagements, and more.

Citi Bike is available for use 24 hours/day, 7 days/week, 365 days/year, and riders have access to thousands of bikes at hundreds of stations across Manhattan, Brooklyn, Queens and Jersey City. (www.citibikenyc.com)

Description of the Data

The Citi Bike Program has a robust infrastructure for collecting data on the program's utilization. Each month bike data is collected, organized, and made public on the Citi Bike Data webpage.

The data includes:

- 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 or 3-day pass; Subscriber = Annual Member)
- Gender (0=unknown; 1=male; 2=female)
- Year of Birth

Data Collection & Transformation

Source: Citi Bike Data

- Downloaded the individual .csv file for all months in 2018.
- Each file contained over a million rows of data (one row = one bike trip).

Jupyter Notebook

Created a Jupyter notebook to combine and condense the downloaded data into a sample set of data to use for this assignment.

- Loaded each .csv file
- Read the files and stored them into Pandas dataframe
- Selected a random sample of 350,000 records
- Exported to a new .csv file

Tableau

- Imported Sample data into Tableau
- Data Transformation & Exploration (calculated columns, converting numerical data to categorical, etc)
- Created quick tables and calculations to get a better understanding of the data and then decided which visualizations I wanted to execute

Data Visualizations

Created visualizations and dashboards to highlight trends and findings I felt were relevant and would be helpful/insightful to Citi Bike New York City.

Link to my Tableau Public Workbook:

https://public.tableau.com/views/Citi_Bike_NYC_Data_Analytics/CitiBikeDataAnalysisOverview?:display_count=y&publish=yes&:origin=viz_share_link

Citi Bike Data Collection Methodology

While exploring the data, I came across numbers that appeared fake and major outliers. I questioned some of the data in the following two columns:

- User Type (Customer = 24-hour or 3-day pass; Subscriber = Annual Member)
- Gender (0=unknown; 1=male; 2=female)

I did additional research to gain insight on how the data – specifically demographics – is collected. I had a feeling that "Customers" weren't required to provide Gender or Date of Birth. I was right.

Citi Bike New York Citi Plan Options

Customer Plan Options

Single Ride (one ride up to 30 minutes)
Day Pass (unlimited 30-minute rides in a 24-hour period)

Subscription Plan Option

Annual Membership (unlimited 45-minute rides)

I went through the online process of buying each type of plan to confirm the information collected at the time of purchase.

Customer plans (single & day pass): you are not required to enter your age or gender (only a credit card is necessary).

Subscription Plan: you are required to enter date of birth and gender (both fields are mandatory).

However, you can input any birthday possible and there is an "Other" option in the Gender drop down list. Just because these fields are required does not necessarily mean the data is accurate and valuable (in terms of data analytics).

This does explains why over 60% of the total Unkown Gender category are Customers.

Note: Our data set does contain gender and birth date information for Customers so there must be another way of collecting this information.

Findings

In addition to telling the story through my Tableau Workbook, I've included some notes I took along the way as I discovered different findings or trends.

User Type – Subscriber versus Customer

Based on the data, it's clear the typical "Subscriber" and "Customer" are two different people.

Subscriber
- Uses Citi Bike for business/work transportation (weekday user)
- Short average trip duration (12.9 minutes)
- More in depth data collected when subscribe (required to provided gender & DOB)

Customer
- Uses Citi Bike for pleasure/leisure activities (weekend user)
- Significantly longer average trip duration (45.5 minutes)
- Not required to provided gender & DOB
- Possible large portion = Tourists?

Subscribers vs. Customers
Subscribers = Most Active on Weekdays
Customers = Most Active on Weekends

Information helpful when it comes to marketing and advertising. You can change your message depending which type of user you are targeting.

Daily Activity / Volume
Saturdays & Sundays = lowest volume of bike trips per day (all 12 months)
Sundays = Lowest number of rides per day
Saturdays = 2nd Lowest number of rides per day

Saturdays & Sundays: Lowest in total volume per day, but the highest percentage of total activity for Customers. While Customers take the most trips these days, the total number of Customer trips on these days is still almost three times less than the number of Subscriber trips.

Bike Rentals & Weather
As expected, volume of bike rentals / average number of trips per month is relative to the average monthly temperature in NYC. As the weather gets nicer – the number of trips increases.

Findings Added to Gender Dashboard

Average Trip by User Type & Gender

Customers' average trip duration (45.5 minutes) is significantly longer than Subscribers' average trip duration (12.9 minutes).

Average Trip Duration by Gender

Unknown gender by far spends more time with the bike they rent because most of the Unknown gender group = Customers (vs Subscribers)

Customers are most active on the weekends which suggests they are renting bikes for leisure activities versus business transportation needs on the weekends - resulting in the longer bike rental times.

Future Considerations

- Include all data from 2018
- Explore trends related to peak times
- Create more in depth / interactive geo maps that also tie into peak hours at a given station
- Do they ever run out of bikes? Is one also available at all stations?
- How many incidents a year are reported? Is there a trend to where these incidents happen?
- Time permitting, could create helpful visuals for days with this data.
- Would love to incorporate the finances/revenues associated with Customers versus Subscribers.