Planes, Trains, and Bicycles (Actually, Just Bicycles)

A Report on Citi Bike Ridership

By Ryan Fore

January 28, 2018

Interactive versions of most of the charts and maps found in this report can be located on my Tableau page

1. Introduction

Citi Bike first launched in May of 2013 with 332 stations and 6,000 bikes as a means of providing an alternative form of transportation to the public, and has been expanded and renovated several times since then. As of October 2017, the number of stations and bikes in service stood at 706 and 12,000, respectively, and a total of 50 million rides had been taken. Data is available for every ride that has been taken through the system, and will be used to explore different patterns in ridership with the aim of increasing ridership and operational efficiency. Amongst other things, it's found that:

- Women vastly underutilize the system compared to men
- A significant number of trips are taken to commute to work, which causes stress on the system
- Temperature and precipitation are very good predictors of daily ridership levels

2. The Data

The ride data is taken from Citi Bike's AWS databases, which has available data from July 2013 to September 2017, broken down by month. For every ride that has been taken, data such as trip duration, start and stop location and user type are available. For trips taken by annual subscribers, additional information such as birth year and gender is available. Due to the size of the dataset only every 10th trip was used for analysis, as this would have a miniscule effect on the metrics obtained while significantly increasing analysis speed.

Data regarding individual stations, specifically the total number of docks available, was taken from Citi Bike's Live Station Feed. Due to the fact that over the years, stations have been added and removed, as well as the fact that historical data is not available for station statuses, there are trips in the ride data that refer to stations for which data cannot be found. In these cases, the missing data was filled in using the average of the total docks available for the rest of the stations.

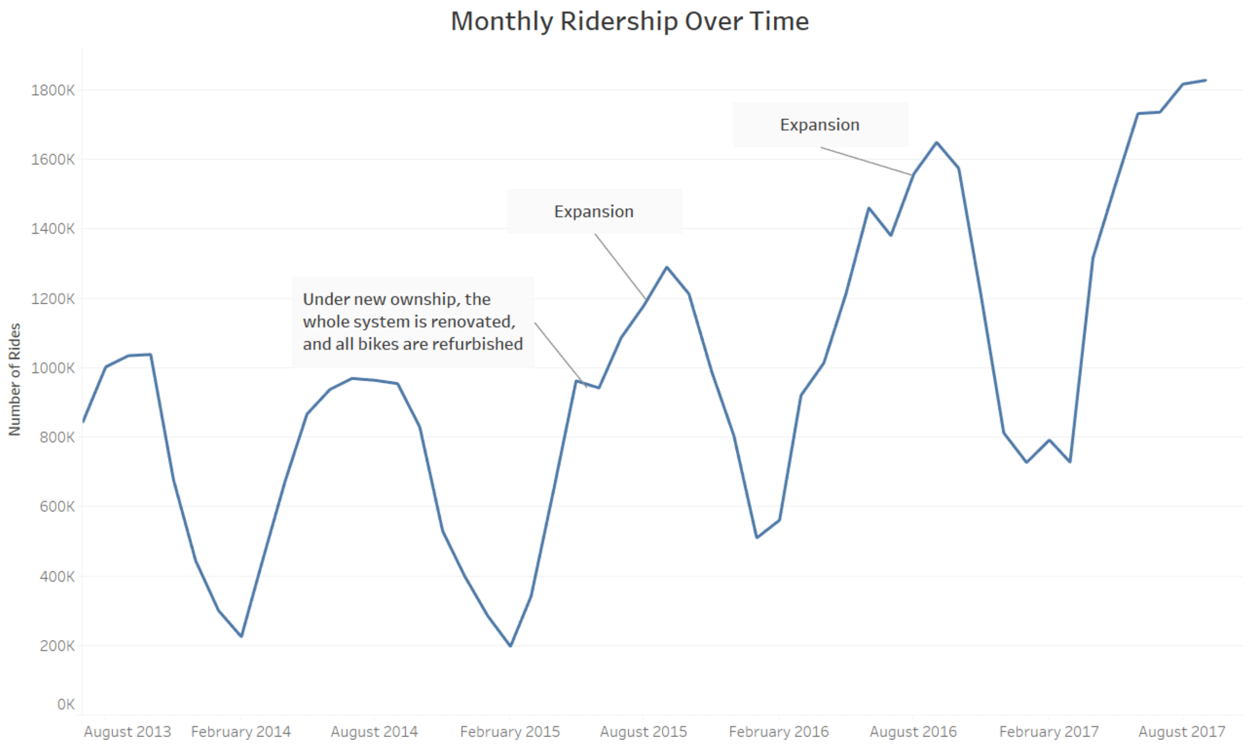
Weather data was obtained from the Dark Sky API. Dark Sky is a weather data aggregator that has a 'Time Machine' API that allows you to obtain historical hourly weather data based on latitude and longitude. This data includes information such as temperature, precipitation type and intensity, and even wind strength and humidity.

3. The System

There are two options for using the Citi Bike system. The first is through an annual membership, which allows a user to take an unlimited number of rides over one year after purchase. At first the plan cost $95/year, but in March 2015, shortly after ownership changed, the price was raised to $149/year, and now stands at almost $180/year. Furthermore, if you keep a bike out longer than 45 minutes, you have to pay $2.50 per additional 15 minutes.

Another option to using the Citi Bike system is to sign up as a customer. Customers have the option of buying a 1-day or 3-day pass allowing unlimited rides during that period, for $12 and $24 respectively. Like subscribers, customers have to pay extra if they have a bike out longer than a set period, in this case though its $4 per 15 minutes after only 30 minutes.

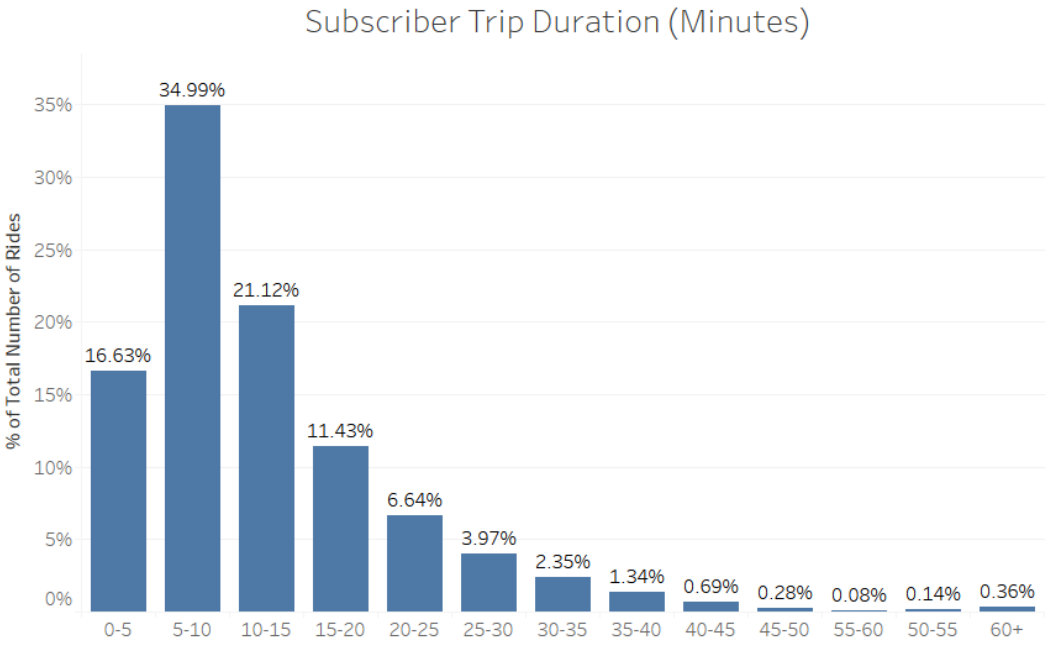
As mentioned before, since Citi Bike first launched in May of 2013 it has been expanded and renovated several times. Looking at a graph of monthly ridership over time, ridership shows both strong seasonality, and a general upward trend in use as the system has grown.



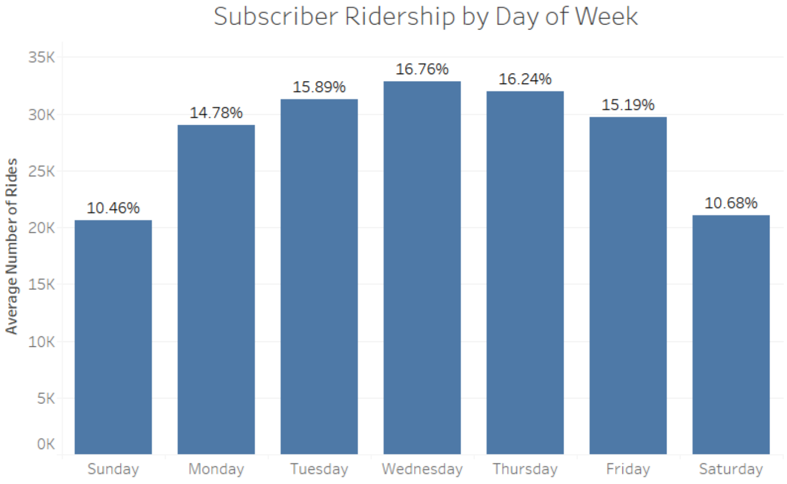
Perhaps unsurprisingly, the vast majority of rides are taken by subscribers, who take just over 88% of all rides. Since subscribers and customers presumably use the bikes for different reasons, moving forward analyses will focus on one or the other. Let’s start with subscribers.

4. Subscribers

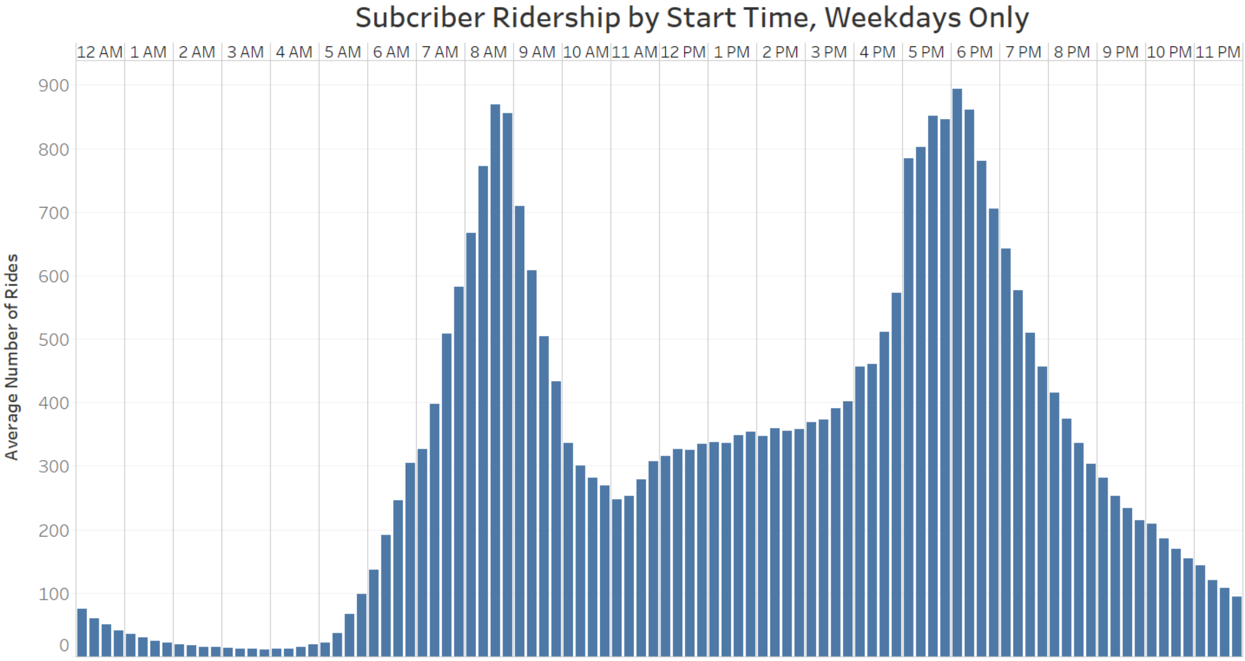
This first graph shows that most subscribers use the system for short jaunts rather than for longer leisurely rides, with over half of all trips lasting less than 10 minutes (the median trip length is 9.7 minutes). The average ride is 13.5 minutes long, while only 5% of rides last longer than half an hour. Less than 1% of rides go longer than 45 minutes, thus incurring a fee.



Next is to look at when these members are taking their trips. If we look at ridership broken down by day of the week, we see that weekdays see much more activity than weekends do.

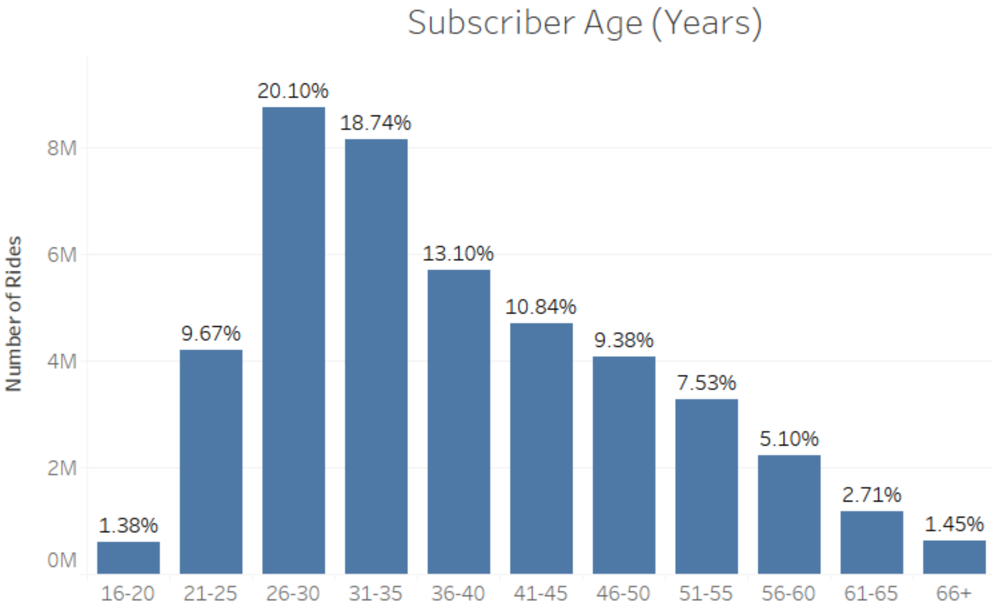


Short rides during the week seem to suggest that many rides are used for transportation to or from work. To check this, I charted the average number of trips that start during each 15-minute interval of the day, filtered to show only data from the workweek.

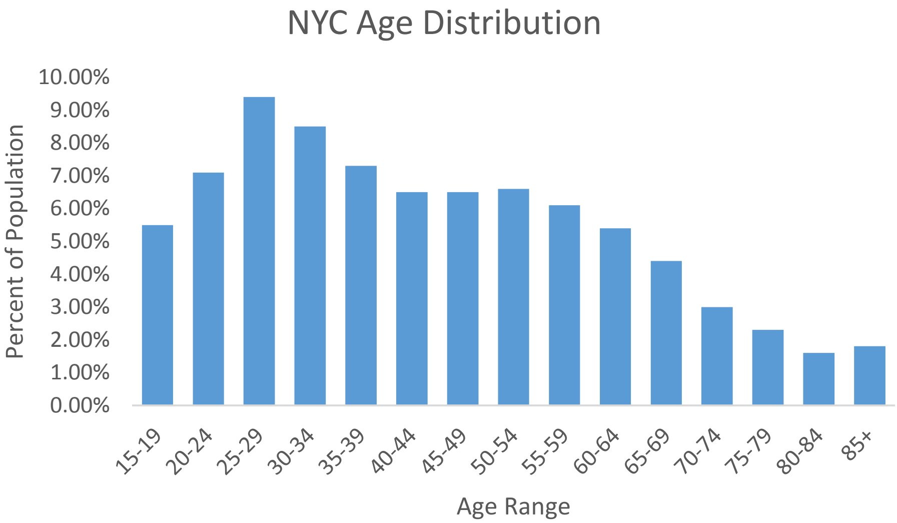


Looking at the data, we see clear peaks during both the morning and evening commuting hours. In fact, 21% of all subscriber trips during the week start between 5 and 7 p.m. These spikes in demand have such a profound impact on the system that they will be studied more in depth in a later section.

Moving on to the demographics of our subscribers, let's look at the age of the riders. Ages range from 16(the minimum age to obtain an annual membership), up to well over 65, although in general not many elderly people use the system.

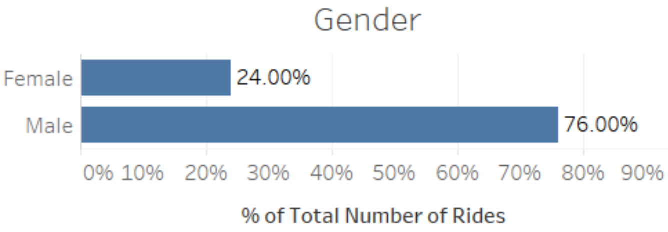


The 26-30 year old cohort clearly uses the Citi Bike system the most, accounting for over 1/5th of all subscriber rides. After that, as age increases, ridership gradually decreases. What stands out though is the sharp decline in ridership for ages younger than 25. A large part of this due to the population of the city, as shown below.



Since there are many more people in their late 20s in the city compared to people in their early 20s or late teens, it makes sense that they take the more trips. Other contributing factors to why younger people don't use the system as much are likely the high cost of an annual membership, and the fact that most younger people don't have daily jobs they have to commute to. It still however might be possible to increase ridership along the younger age groups to bring it more in line with the general population demographics

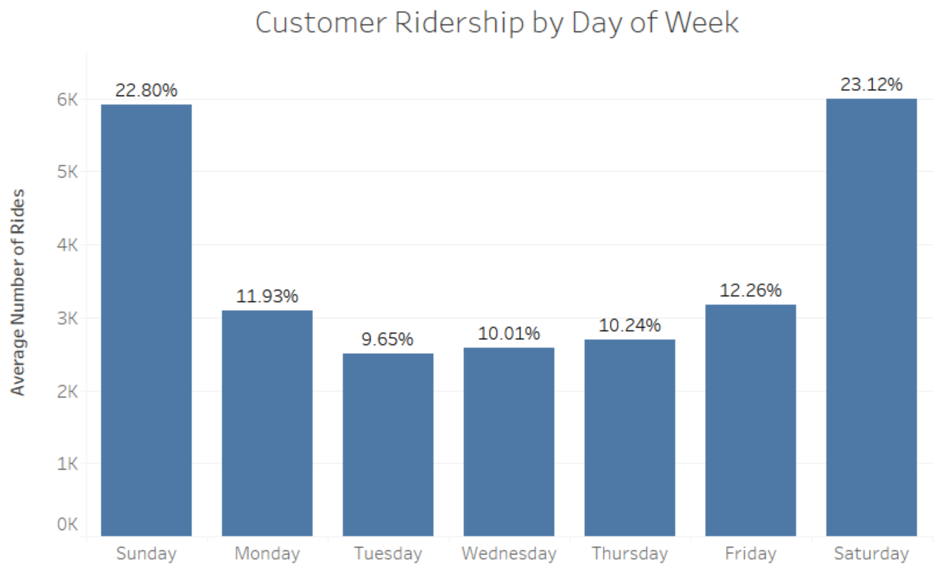
Looking next at the gender of the riders, we see that men take over 3 times as many trips as women.



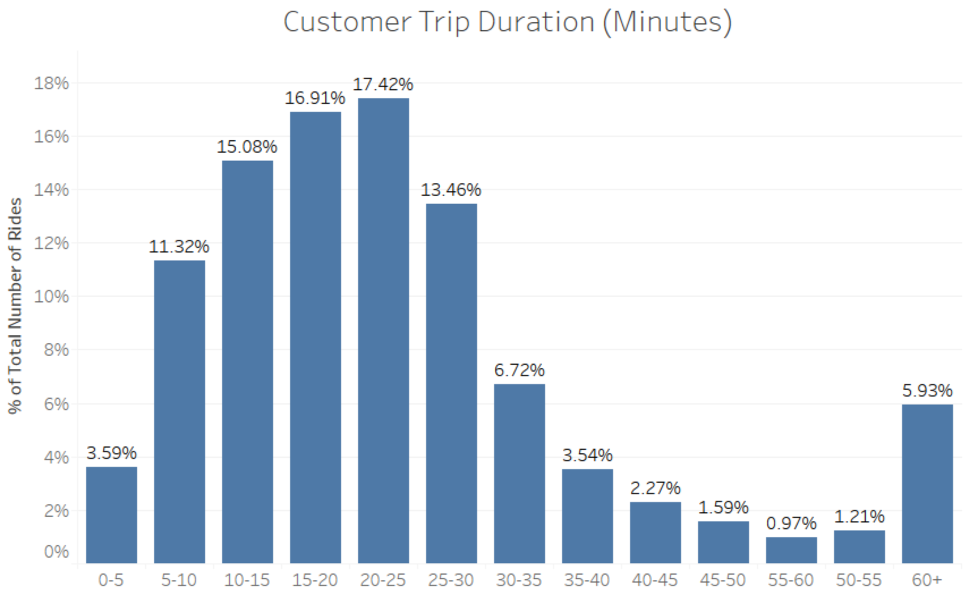
There are many potential reasons for this discrepancy. For example, women's attire (e.g. heels, skirts) is less conducive to riding a bike. Women are also less likely to ride bikes late at night compared to men, and also less likely to ride during adverse conditions such as rain. Despite all of this, with such a large disparity, women are a potential target market to greatly increase ridership.

5. Consumers

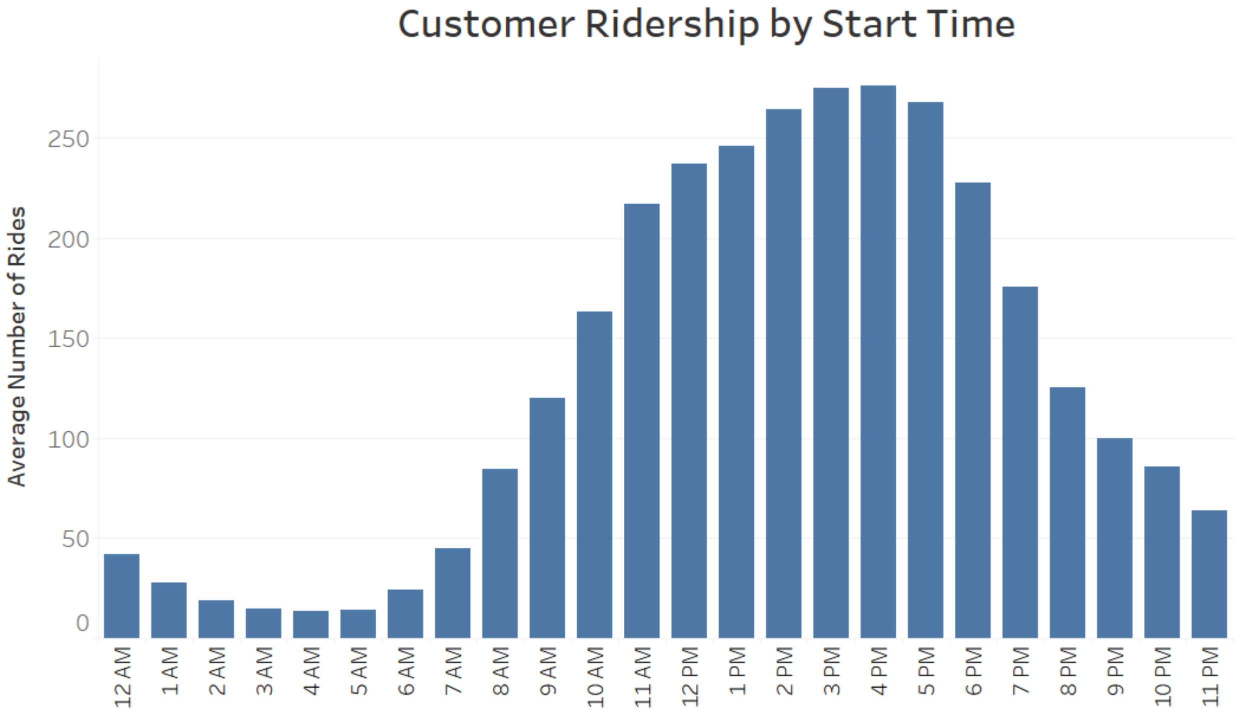
Looking first at their ridership by day of the week, we see that customers overwhelmingly use the system over the weekends, with over 45% of all trips occurring on Saturday and Sunday.


Customer Ridership by Day of Week

Customers also generally use the bikes for much longer than subscribers, with an average trip duration of 33.2 minutes, compared to the subscribers' 13.5. Similarly, the median trip length is 20.9 minutes, compared to 9.7 for subscribers. Over 15 % of all customer rides last longer than 30 minute, thus incurring a fee, and almost 6% last longer than hour, which at a rate of $4 per additional 15 minutes means they are spending at least $12 in additional fees.


Customer Trip Duration (Minutes)

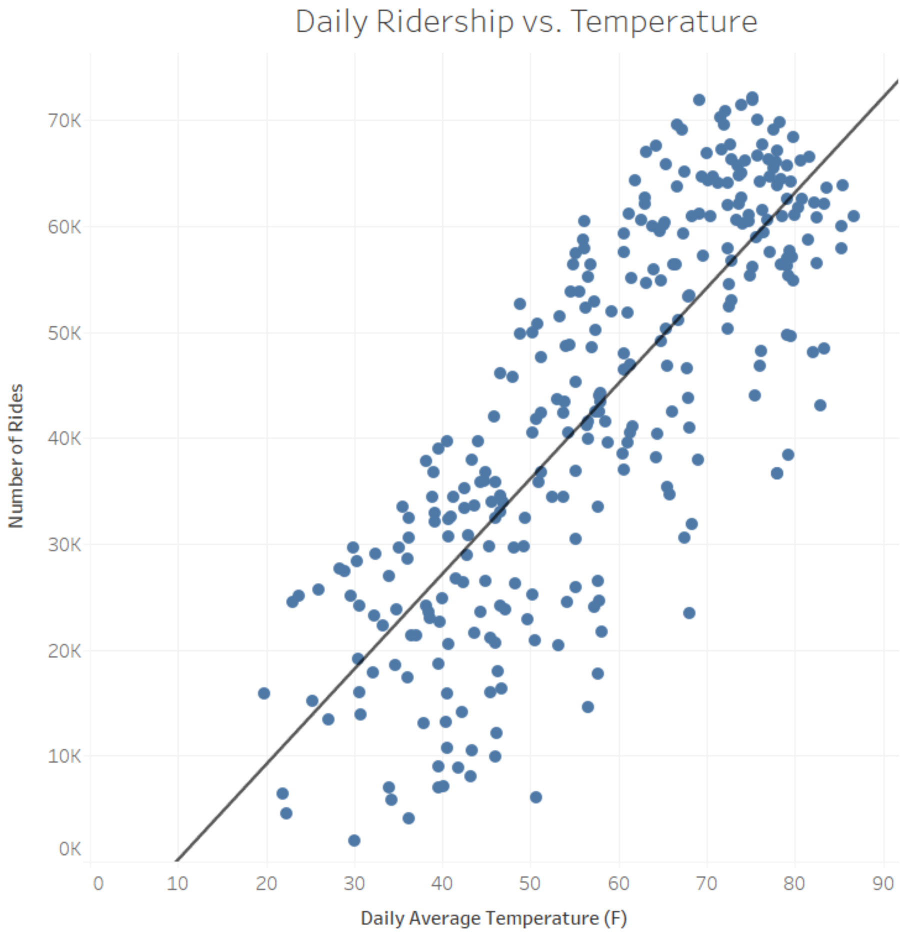
The graph for start times for customer trips is much smoother than the subscribers', showing that most customers use bikes in the afternoon, with usage spiking between 3 and 4 p.m.



Demographic data such as age and gender are unavailable for customers, so now we'll delve into the effects weather has on ridership.

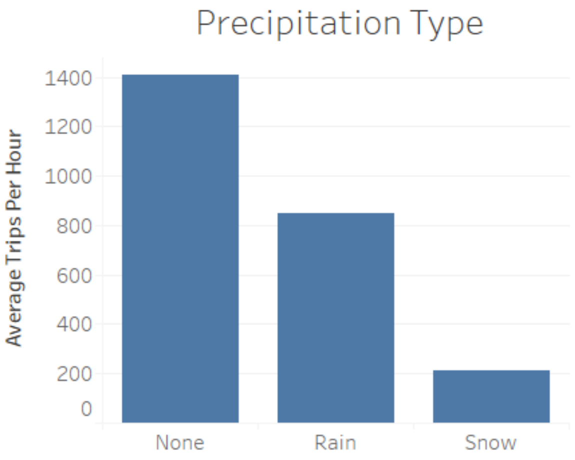
6. Weather

There are two main aspects of the weather that affect ridership: the temperature, and precipitation. To judge the effect temperature has, I calculated the average temperature for each day, and the amount of rides taken on that day, and plotted this data on a scatterplot with a fitted trend line. Only the last year of available data was used for this chart in order to remove the influence the size of the system has on ridership, since there were no major expansions during this time period.



As expected, we see a strong positive correlation between temperature and the number of trips taken, with the r value standing at .81.

To look at the effects precipitation has on ridership, I plotted the average number of rides taken during hours when it's raining vs. when it's snowing vs. when there's no precipitation.

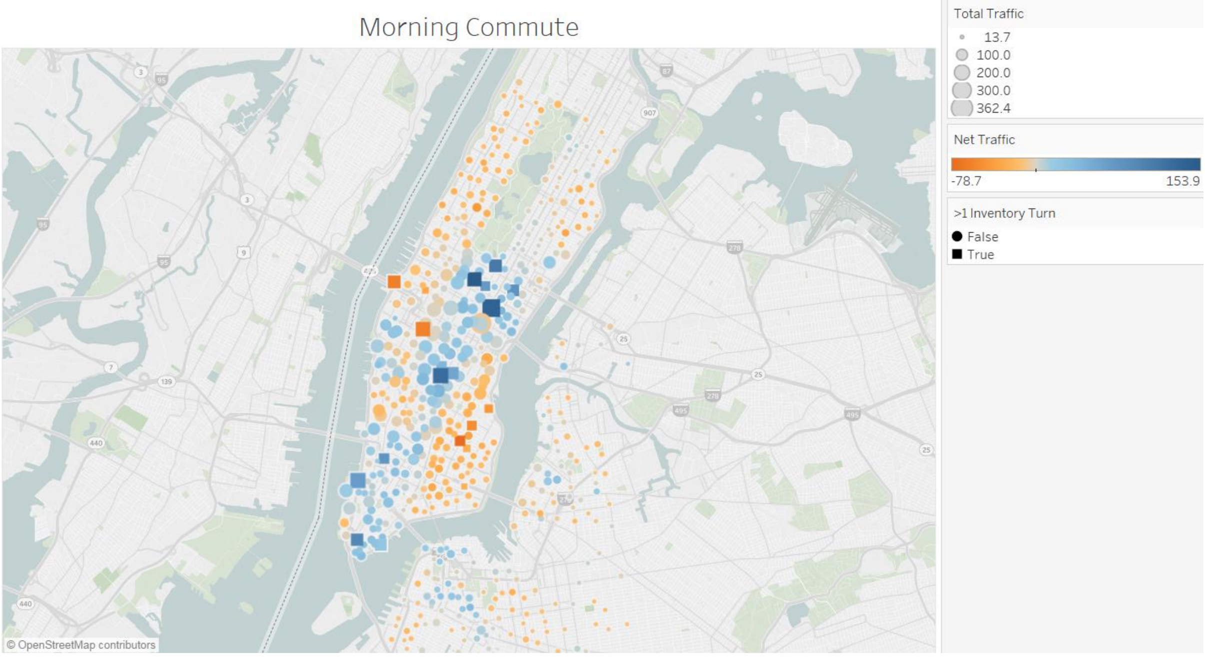


Unsurprisingly, less rides are taken when it raining versus when there's no precipitation, and even less when it's snowing. It should be noted though that snow is heavily correlated with temperature, while rain is not, so a large part of the low ridership while it's snowing is likely due to the temperature rather than the snow itself.

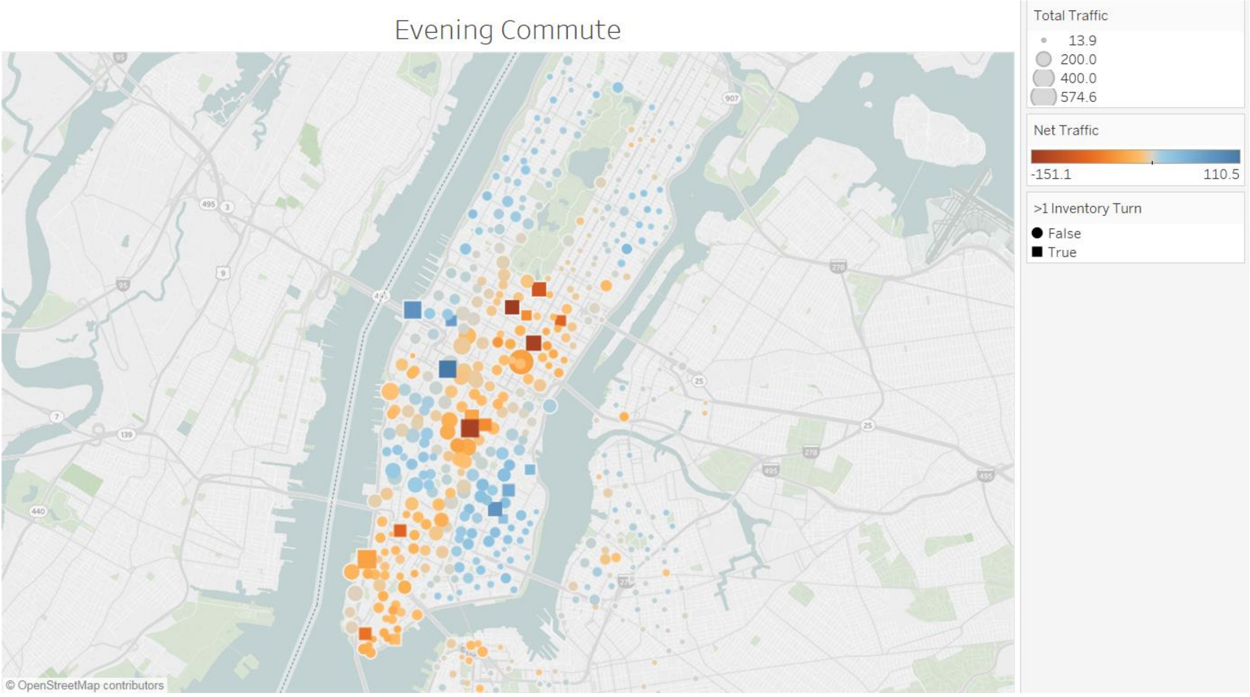
7. Commuter Hours

To understand the effect the commuter spikes have on the system, it's important to not look at the just the number of rides taken, but where they are taken. To this end, I created two maps, one for each commuting period, showing the traffic at each bike station. The morning commute encompasses all rides started between 7 a.m. and 10 a.m., while the evening commute is comprised of rides starting between 4 p.m. and 8 p.m. The maps show daily averages from June 2017 through September 2017, as there were no major expansions during this time, and the summer months see the highest ridership.

The size of each mark on the maps below is based on the total traffic at each individual station (# of bikes in + # of bikes out), while the color is based on the net traffic flow (bikes in – bikes out). Blue marks indicate more bikes were taken to the station than out of the station, and orange marks mean the opposite. I also found the number of inventory turns for each station during the commute, calculated as (Net Traffic)/(Number of Docks). If a station has at least 1 full inventory turn, that means that at some point during the time period bikes were rebalanced to or from the station. The stations that have at least one inventory turn during a commute are marked as squares instead of circles.



The morning commute map shows clear patterns as to where riders are heading. In general bike traffic flows from more residential areas, such as East Village and the upper west side, into areas with lots of office buildings, such flatiron, the financial district, and especially midtown. For example, on average the station at W 52nd Street and 6th Ave sees 153.9 more bikes deposited than taken out during the morning commute. Since there are only 39 docks at the station, this means there are almost 4 full inventory turns, and so bikes have to be rebalanced from the station at least 4 times during this time period.



In the evening commute, we see an exact reversal of the patterns found during the morning commute. Riders are overwhelmingly going from places of work to more residential areas. Again we see the west 52nd Street and 6th Ave station with the biggest net change, with on average 151 more bikes taken from there than returned there (3.9 turns).

Given that these maps use data from both when it's raining and not raining, it's safe to say that these numbers regarding flow and inventory turns would be even higher if we looked only at when it wasn't raining. Since there are stations that go through multiple inventory turns while there are stations a block away that don't even go through half a turn, it might be beneficial to incentivize riders to take bikes from the latter stations.

8. Closing Thoughts and Actionable Ideas

After a bit of a rocky start due to reliability issues, the Citi Bike has grown to be extremely popular, especially amongst people looking for alternative ways to commute to work. The program however is not equally popular amongst different sectors of the population, with women notably taken less than a third of the number of trips men do. It is therefore recommended to conduct more research as to why this is, and perhaps to find ways to market the program to women.

Another subsection of the population where ridership remains low is amongst people in their teens and low 20s. If ridership could be increased at younger ages, it could create a pipeline effect where they continue to use the system as they get older. Since price is likely a contributing factor towards low usage, a lower cost plan could be offered to people under a certain age limit around 24 years that could get these younger people riding. These riders also will generally not be using the bikes for work, so additional limitations could be placed on it, for example limiting the number of rides taken per month, or what stations they can use during peak operating hours. This could also potentially help with the rebalancing issues caused by commuters.

Other potential solutions to the rebalancing issue could be incentivizing users to use loss popular stations, adding stations in high-demand areas, or expanding the use of valet stations, which are mobile, temporary stations put up in high-demand areas during peak-usage, a system that so far been almost exclusively used near Penn Station.

Another area for further analysis would be to create a predictive model that predicts the usage at each station throughout the day. To make this model accurate, more detailed information regarding the number of bikes and stations in circulation would have to be scraped from Citi Bike's monthly operating reports. It would also be important to have more detailed information regarding each station, such as regular updates about how many bikes are available. Historical station data however is not available to the public, so data would have to be collected through the station data api over a very extended period of time.