Analysis on Citi Bike Ridership

Introduction

CITI Ride, an inspiring cycling share system, gives New Yorkers a quicker alternative means of transport than walking, more cheaply than a car, more affordable than a bus, and more fun than a subway. In May, the system began in Manhattan and Brooklyn with 6,000 bikes. Citi Bike has now operating docking stations in Manhattan, Brooklyn, Queens and Jersey City after years of growth, and the cycling fleet has expanded to over 13,000. It is common for both weekdays and weekends, and hence forms a key part of the transport network in New York City. We aimed at helping Citi Bike learn about the riding trends in NYC to recognize the most relevant factors that could affect its company and establish the best plan for expansion.

For any random positions in the city Citi Bikes cannot pick up and drop. Alternatively, riders can take and drop bicycles at countless stations in the area. In 2015, approximately 176 million trips in taxis, 36 million Uber trips and 10 million Citi Bike rides were possibly not reached but the motorcycle ratio is expected to start expanding over the next years.

For any trip in the network and the form of membership that each driver inscribes, Citi Bike makes data available. It can be extracted from the following site https://www.citibikenyc.com/system-data in form of Csv file, cleaned and transformed into a form that is suitable for our analysis. I have downloaded daily Central Park weather data and added it to Citi Bike data to better form the relationship between the usage of Citi Bike and weather assuming January is summer and August is winter of the years 2017, 2018 and 2019. Our data has the following variable:

1. Duration in each trip (seconds) -Shows the time it took to finish the bike ride
2. Start Time and Date
3. Stop Time and Date
4. Names of the station
5. Station locations for where the ride started- NYC addresses consisting of St and Ave
6. Station locations for where the ride ended- NYC addresses consisting of St and Ave
7. Station identifier
8. Latitude and longitude of the station- Coordinates
9. Membership type that is : Customer =short term that is 24 hours access to 3 days access ; Subscriber = Annual membership
10. The gender of the rider classified into three :Zero=unknown; 1=male; 2=female
11. Rider year of Birth
12. Unique ID of the bike being used.

We are dedicated to helping Citi Bike learn about the riding trends in New York city to recognize the most relevant factors that could affect its company and establish the best plan for expansion. This achievable by answering the following questions.

- Where do Citi bikers ride?

- When do they Ride that is in terms of time of the day; morning, afternoon or in the evening?
- How far do the riders go?
- Which stations can we say are the most popular ones?
- What days of the week are most rides taken on?
- Do seasons have any impact on the type of subscription a male or female rider take?

Data visualization techniques provides as an easy way of observing and interpreting trends, outliers and anomalies of data by means of visual elements such as maps, tables and charts. The digital view of information and statistics shall be the representation of our data. Therefore we choose Tableau as our statistical tool to help us in this.

Design

We intend to use pie chart to show the distribution (in terms of percentage) of the type of subscription an individual takes. We also use a bar graph to show rider membership type in correlation to Gender over the three years.

Another way in which we can use bar graph is to show the relationship between riders age and and also ridership by age and gender.

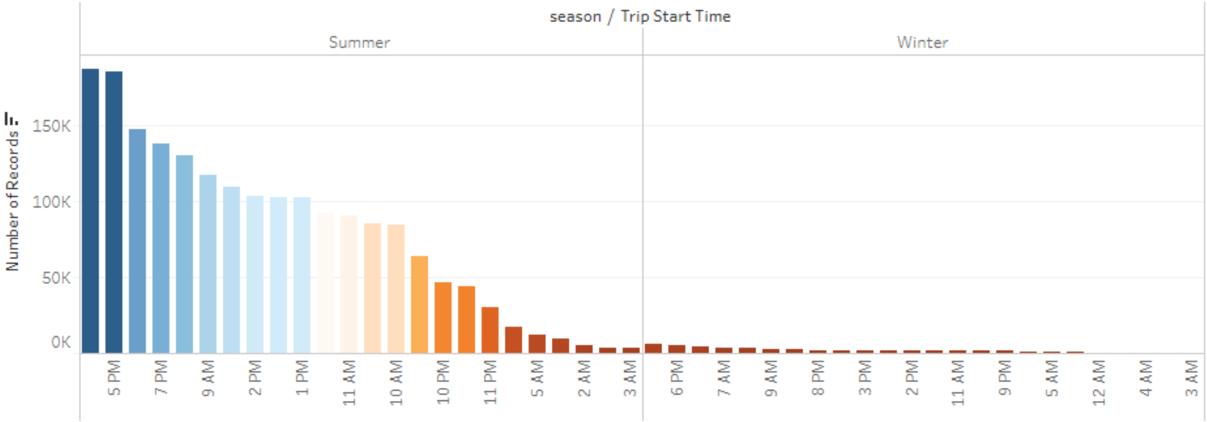
Using chart map of New York we will also be in a position to depict the average distance covered by an individual in each trip and also which stations are mostly riders likely to cycle to and fro.

Implementation

1. Weekly and Hourly Trends

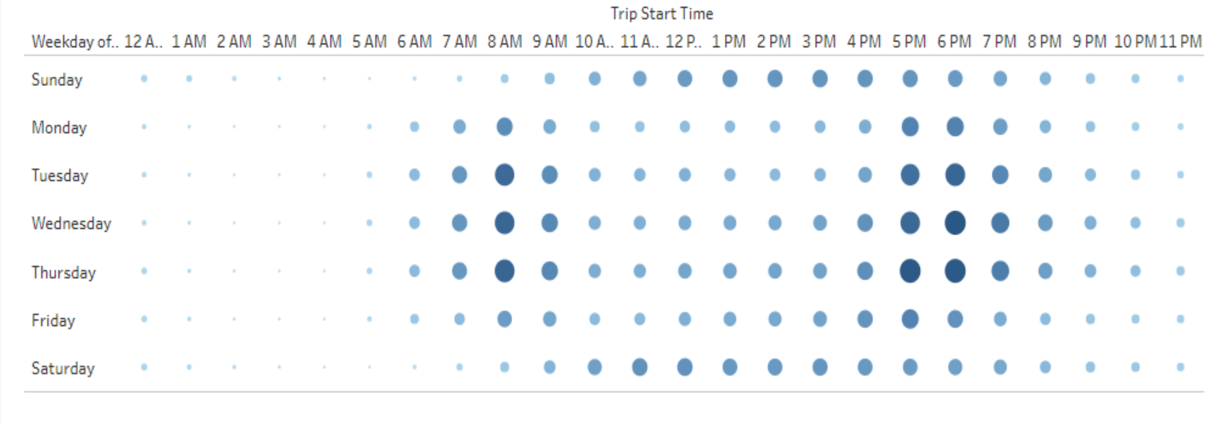
We were interested to hear about trends in riding for a week and a day after seeing how the weather influenced riding. In order to visualize peak hours, we first show the relationship between numbers of records in the two seasons against time. Housing the three measures we discussed earlier on the seasonal pattern are a graph showing contrasts of the Summer (left) and the Winter (right). On summer, there are more rides as most Citi Bike riders pedal to work.

Fig 1



We can also go further and build a chart to indicate each day of the week and the number of records are there in every hour of the day. On weekdays riders waste an average of 3 minutes on bike relative to weekends. This can be demonstrated by attempting to get to the destinations under time constraints as soon as possible. The average travel distance for both cases is 1.5 miles per trip.

Fig 2

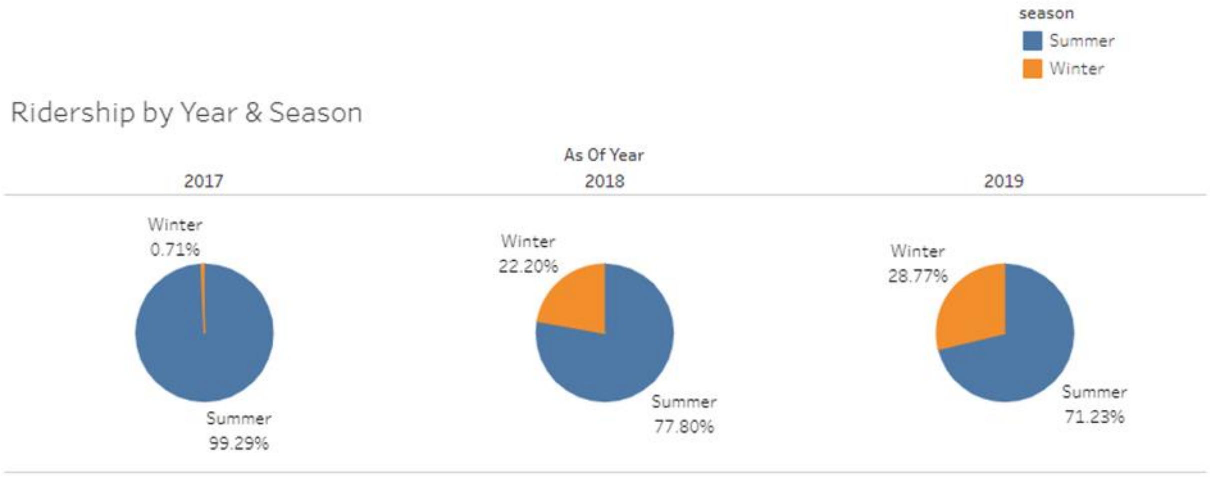


The next question we addressed was whether demand shifts between 5 a.m. and 10 p.m. during one day. The following animation helps us to imagine the hourly change in operation. The smaller the dot the less the demand the larger the dot the more the demand.

2. Seasonal Trend

Next, we wanted to understand has cycling changed over time and how does weather impact the total ridership. The graph below is a seasonal trend of Citi Bike trips during two season summer and winter from 2017 to 2019. The first graph below shows the ridership throughout the three years and the two seasons, and the last map shows the average trip duration in seconds and the last graph shows the average distance traveled per trip. It is obvious to observe a seasonality in all three measures. Trips increased not just in number but in duration and distance in warmer seasons, as seen in the season of summer, as compared to colder seasons, as winter. This is not surprising as people spend more time outdoors in warmer weather.

Fig 3

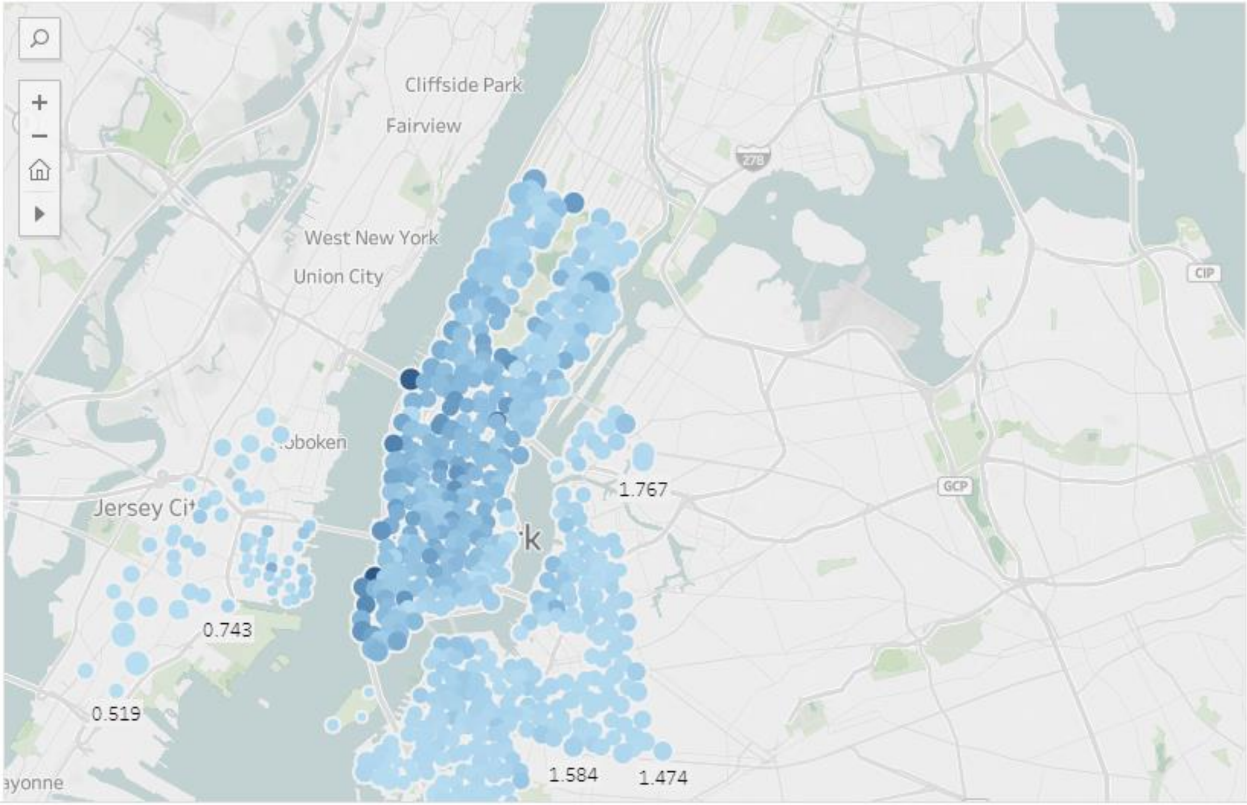


After seeing this seasonality in activity, we further investigated how temperature impacts the ridership to confirm what we learned from the seasonal trend. We observed a linear relationship between these two features. When we incorporate gender variable we can clearly see that male riders are more dominant than female riders regardless of the user type or the season.

3. The Area and Duration of Riding

The path to be taken according to the Google Maps involves each journey, and we vary from the time it takes to travel. Age and gender details can be paired with Google Maps' riding directions to answer a variety of interesting questions. It would illustrate the actual average riding pace for riders as they take the bike out of the station, change it, check for direction and other barriers, and return the bike to the station. This is a fantastic opportunity for drivers. It also ensures that the driver takes the Google Maps directions. When a rider followed longer than Google indicates, it will increase the distance and underestimate the average speed. If the driver has a smoother direction than Google suggests on the other side, the pace of the road could be overestimated.

Fig 4



We have no idea about any particular rider: Some cyclists are simply attempting to get as fast as possible from point A to point B, while some would prefer to take a scenic path from point A to point B. The following party would almost definitely not take a straight path and so, while the riders pedaled for the whole time, we would end up measuring an exceedingly slow average speed on these journeys. Therefore I limited myself to the following subset of trips for an overview of bike pace, which at least I weakly say would be more likely to include riders who want to rapidly get from station 1 to Station 2. We will see routes where riders want to follow the map that is built in our map

4. Rider performance based on gender

Several citizens were born before 1960 for the year of birth. I might assume that sixty years old should ride a bike, but this stretch is an anomaly and false facts, for everybody "born" before 1960 riding Citi Bike. Many disabled people may be riding a cycle, but certainly not. They will see the efficiency of the users with their age and sort of user.

5. Station Dynamics

To answer our question on which station is more popular we create a scatter plot of each stop and start station with their name against count of number of records. We expect top ten stations to appear in both the start and stop category.

6. Anomalies in the ride.

I know little about bike sharing network management. Nonetheless, I can think that one of the biggest challenges is to make sure there are enough bikes at the stations where people are required to pick them up. If station A ends with a lot of bicycles and people drive them to other stations and no one takes bicycles back to A, A is full of bikes, that's false.

The bike share worker can take more bikes to A in order to satisfy demand, but it takes time / money and thus the operator is more likely to do this. The stats demonstrate how much bikes move "magically," as though none of them travel, from one station to another. I took off the bike and measured the proportion of trips that I started on the next journey at a different station from where I left off the previous trip.

There is the question of sensitivity of data, how it is used and the intentions afterwards. Citi Bike offers demographic information on its users, such as gender, year of birth and status of subscriber. At first sight this does not seem to be insightful, but it turns out that several Citi Bike trips can be identified uniquely. If you know the following about a ride by Citi Bike

- Status of rider is an annual subscriber
- Rider gender
- Rider's birth year
- Station in which where they picked up a Citi Bike
- The date and time they picked up the bike, rounded to the nearest hour

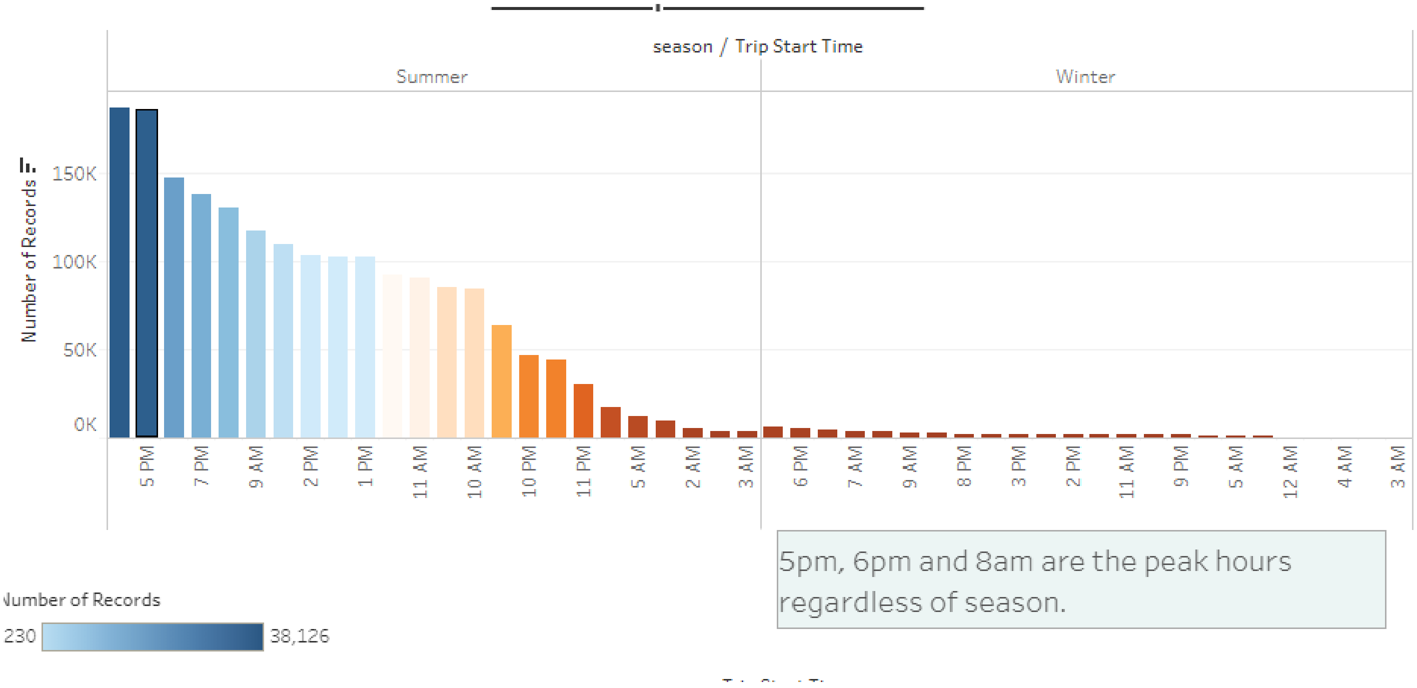
Therefore 84 per cent of the time you can recognize this single path! You will then figure out what confidential details could be, since the driver fell off the wheel. Because men make up 77% of all travels for users, it is also harder to classify women's trips uniquely: 92% of the journeys can be identified only if we are female passengers. Drivers that are significantly younger and older than the average can also clearly be identified

Evaluation

1. What days of the week are most rides Taken on?

On weekdays, there are more rides as most Citi Bike riders pedal to work. We also tracked the hourly pattern in cycling to help us identify the busy times of the day. We had a greater number of trips than the early morning, late night and middle time between 8-9 am and 4 – 7 pm. Those hours have been marked as hourly to be analyzed later. If this information is paired with Citi Bike maintenance details, it is possible to determine when maintenance and repairs are required for a bike. This will lead to a reduction in the number of damaged bikes in stations which is a concern for many passengers. This information can also be used to build a brand and for marketing purposes.

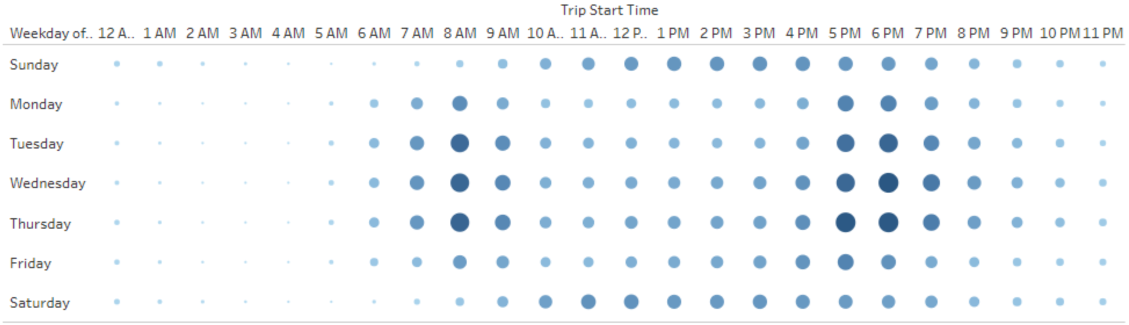
Fig 6



5pm, 6pm and 8am are the peak hours regardless of season.

The most commonly bike, prominently marketed as a sign to promote the bicycle use in NYC, and would be cool and a very smart idea for the business.

Fig 7

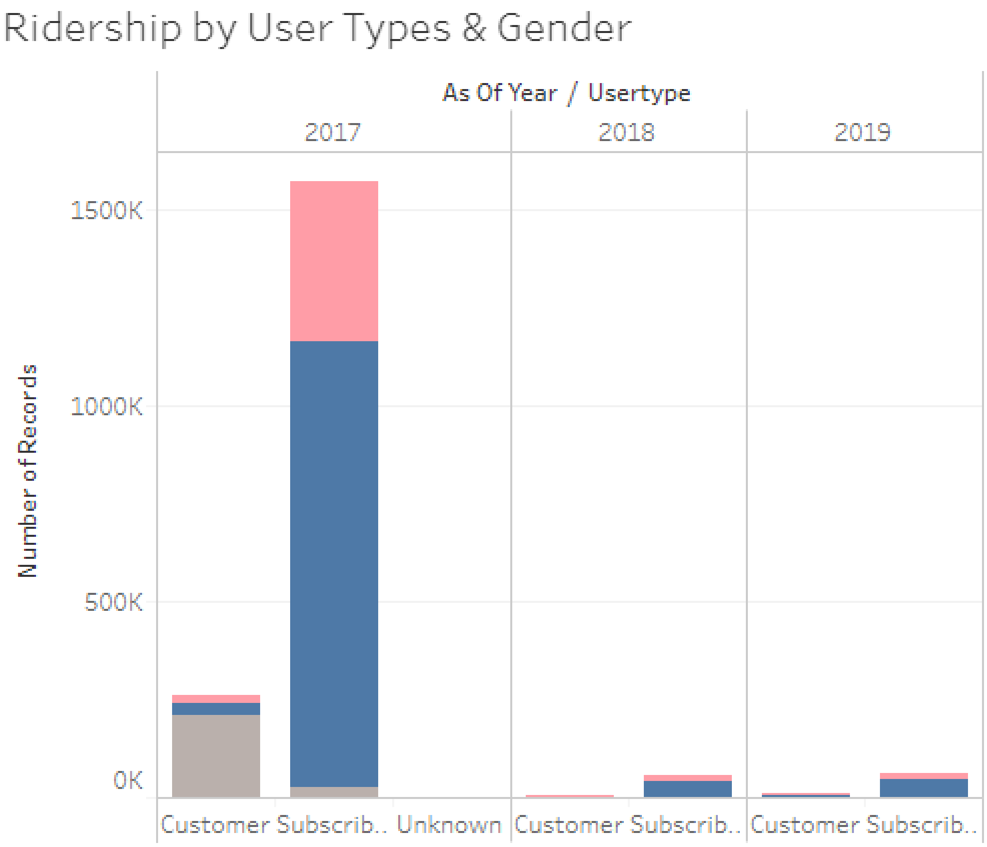


2. Rider’s performance by age and gender

There is no credible way to find out a bike path because we do not know the route that a rider took on each bike without GPS info. We can use longitude co-ordinates to define cycling path distance with Google maps and latitude. However, more than the daily cap on API calls would be needed. This trips are usually include some trips are uphill, others are downhill. Some paths, including once in a while, require heavy traffic, which relies on intuition. According to the results, the explanation for user type customers typically drives slower with regular stops than a subscriber.

Traffic is one aspect that will have a major effect on travel time. Since we never know the path the driver took, though, the use of this detail is difficult. Finally, the use of the Google Maps AS caps such that traffic trends cannot be readily detected.

Fig 8



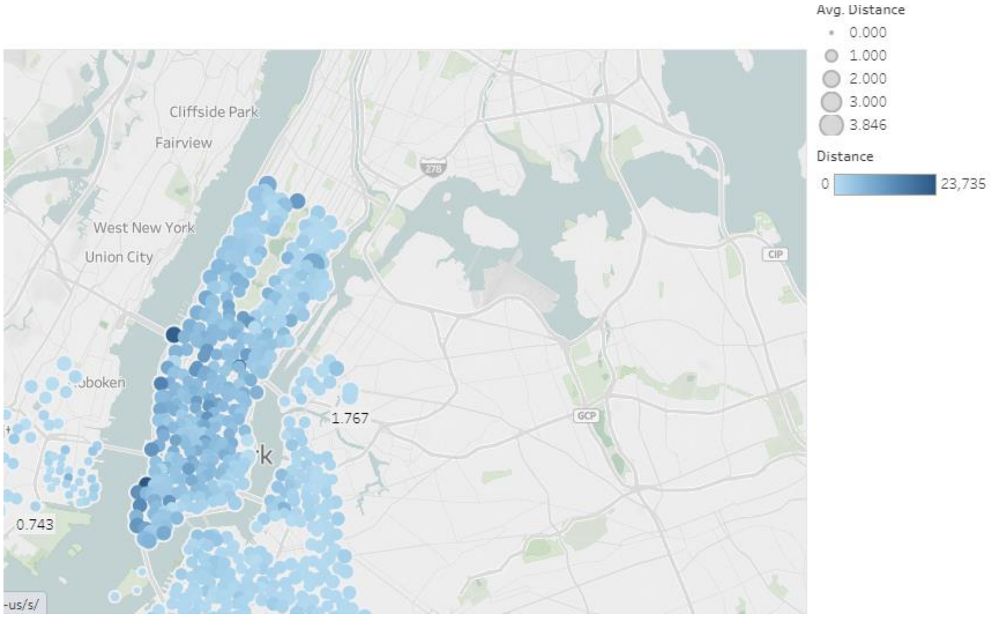
When we consider trend of riders based on their age we see that most riders are between 23 and 38.However I tend to believe there is an anomaly considering there are records of 77 year old riders. This is because normally individuals at such an age can hardly perform physical activities such as cycling. Maybe the next time we do our analysis we can eliminate this group of data.

3. Trip duration

It is fair to assume that the user’s status could really be a good indicator of trip duration. This is a point to remember for the time being and we will return to it later. Customers are eligible for 30 minutes by bike and 45 minutes by bike for subscribers. The evidence shows clearly that consumers prefer to use bikes longer. The new time limit makes sense if the key problem is to have bikes at the docks. But to allow consumers more time per bicycle is worth pursuing if Citi Bike wishes to be user-centered. There is a major concern of having "circular" trips

Circular trips are those beginning and finishing at the same place. But this is not the case. The distance for such trips is 0. The details and graphics would be blurred. The only person who knows how long a trip would take is the rider, if he / she has the experience.

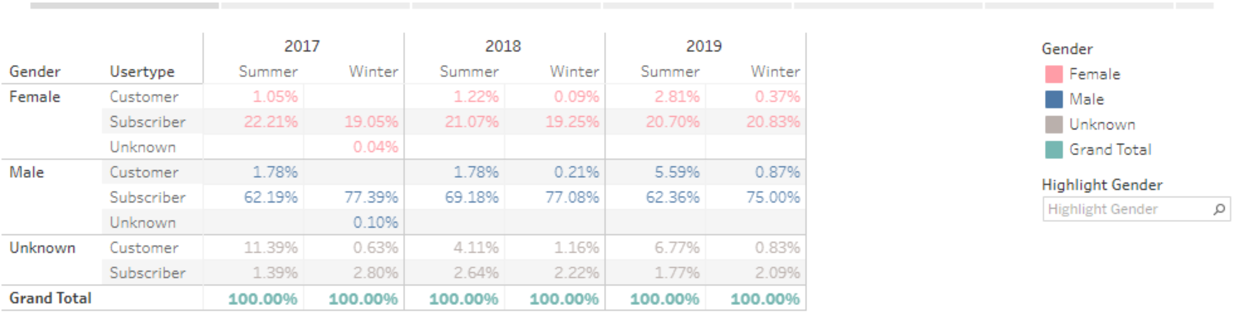
Fig 9



4. Season Trends

We've noticed before that in summer there's a lot more Citi Cycle trips than in winter. This isn't surprising: anyone with the slightest sense knows that when it's cold it's not very fun to cycle. Similarly, bicycling on rainy and snowy days is possibly less common. This prompted me to wonder: how accurate is the daily weather outlook of Citi Bike.

Fig 10



		2017		2018		2019	
Gender	Usertype	Summer	Winter	Summer	Winter	Summer	Winter
Female	Customer	1.05%		1.22%	0.09%	2.81%	0.37%
	Subscriber	22.21%	19.05%	21.07%	19.25%	20.70%	20.83%
	Unknown		0.04%				
Male	Customer	1.78%		1.78%	0.21%	5.59%	0.87%
	Subscriber	62.19%	77.39%	69.18%	77.08%	62.36%	75.00%
	Unknown		0.10%				
Unknown	Customer	11.39%	0.63%	4.11%	1.16%	6.77%	0.83%
	Subscriber	1.39%	2.80%	2.64%	2.22%	1.77%	2.09%
Grand Total		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

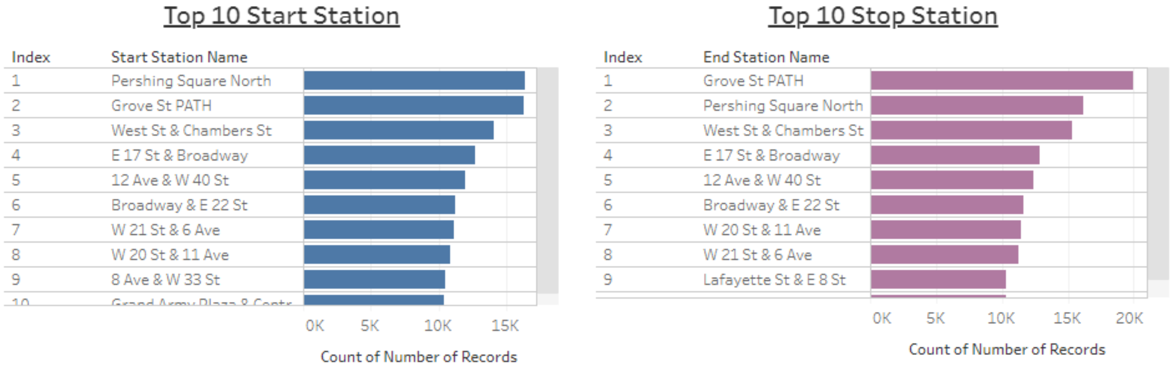
Gender
- 🟥 Female
- 🟦 Male
- ⬜ Unknown
- 🟩 Grand Total

Highlight Gender

Highlight Gender

5. Which stations are most popular?

Fig 11



Clearly we can see that Pershing Square north and Grove St Path are the most popular stations. It is also safe to say that a station that is closer to a transit or tourist hub will have a significantly higher demand than a station that is in downtown.

Discussion

In addition, when stations are introduced in the future I wonder about medialization at the station stage. Opening a new station will have a positive or negative effect on riding in existing stations. A new station could cannibalize trips from other nearby stations, so that overall riding is not significantly increased. We have shown that there are likely to be a shortage at rush-hour stations in certain places, i.e., the outbound figure is significantly higher than the inbound figure. By designing a more effective rebalancing approach we see a potential to change this scenario, but this involves a possible analysis of how often the bikes need to be rebalanced so they can satisfy the demand and reducing the cost of transport. We would also like to evaluate the efficacy of the current system of re-equaling based on rewards and to propose our potential initiatives. When we compare using tableau and another visualization tool like Qlikview we see that in tableau data integration is exceptional while in Qlikview it basically good. When we look visual drilldown is very good in tableau as compared to Qlikview but since scalability is limited by RAM in tableau we would rather use tableau.