Bike Stalking: Where Do The Centenarians Go?

Charles McKay and Kimberly Ly

The Dataset

http://www.citibikenyc.com/system-data

We chose to study data collected by CitiBike, a bike share program in New York. There are over a million entries in this dataset, which is a record of all bike rides in October 2015. The properties of the dataset include:

- length of the bike trip
- when the trip started
- start and end stations
- bike id
- type of user (day or yearly subscriber)
- gender of the user (sometimes)
- age of the user (sometimes)

Right off the bat, we noticed that there was no gender or age data for day subscribers, probably because an extensive profile is not needed for someone who rarely uses the service. In addition, there were some age points that indicated that there were about 300 riders over 100. We'll probably ignore these points as mistakes.

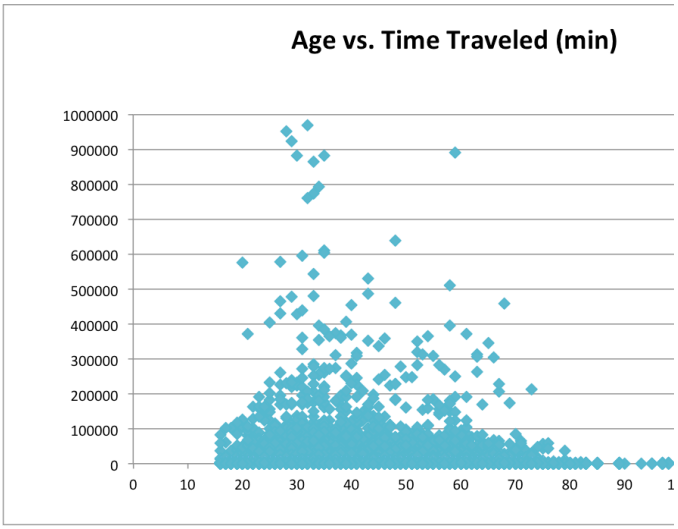
By doing an initial exploratory analysis, we came up with the following statistics:

- Average time on bike: 18 minutes
- Average age of user: 38 years old
- Most common rider age is in their 30s
- Most popular time to ride bike is at 5 p.m.
- Males ride bikes more often, but females ride for a longer period of time
- There is a 10 to 1.4 ratio between yearly subscribers and day subscribers
- The day with highest ridership is Thursday
- A total of 1.3 million miles were ridden by CitiBike during October 2015

Table 1 - *Age vs. Average Ride Time*

Age	Average Ride Time (min)
N/A	38.4604759337
10-19	12.0389869281
20-29	14.0541182753
30-39	15.1425404848
40-49	15.7198915348
50-59	16.1944139771
60-69	17.2471197991
70-79	17.2976800759
80-89	12.1398589065
90-99	13.266091954
100+	16.1013169846

Graph 1 - *Visual representation of Table 1 data*

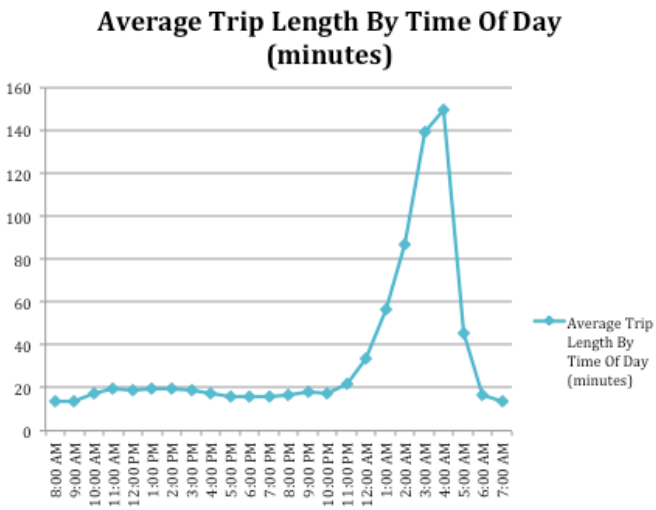


There seems to be more users in the 20-40 age range than any other category, and they also seem to ride their bikes for the longest periods of time. This graph is a little bit misleading because the colors do not show density; there is a very high density of users riding their bikes around the 15-30 minute range, which is why the average ride time is about 18 minutes, but some other users have skewed the graph to make it appear as though a lot of people are riding their bikes for hours at a time.

Surprisingly, the longest average trip times were at around 60 - 79 age group. It could indicate that the younger generation commute shorter distances, or use other transportation along with the bike shares. In addition, the 100+ group seems a little bit high, and although we are not discounting the very athletic centenarians, this will likely be ignored. In addition, the N/A group seems to be extremely high. Recall that day subscribers do not input their age or gender; this indicates that day subscribers use the bikes for longer than the yearly subscribers. This makes sense since day subscribers would want to make the most of their time and money on the bikes.

We also decided to investigate rider trends throughout the day on a 24-hour spectrum.

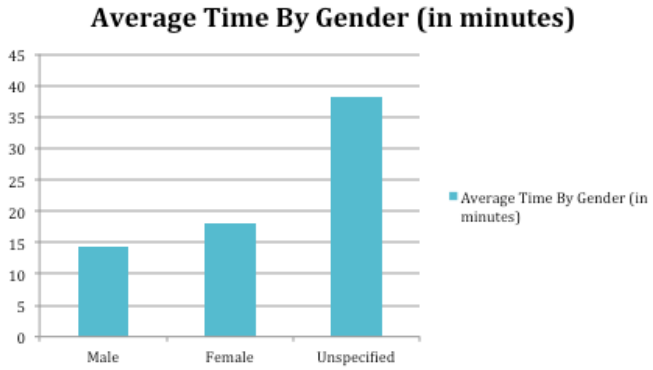
Graph 2 - *Time of Day vs. Average Trip Length*



The correlation in Graph 2 basically shows that people take their bikes out for much longer in the early morning, which might indicate that people use other transportation during the day, in conjunction with the bike system. However, we should also take into consideration the lesser amount of people riding bikes at early hours of the morning. Therefore, their bike times contribute more heavily to the average length at that time.

Another interesting statistic that we chose to investigate was how long users rode their bikes based on gender.

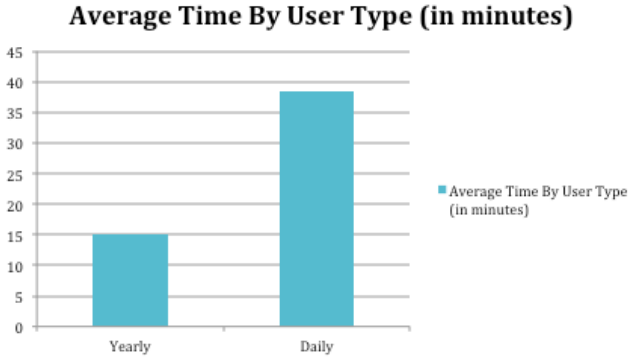
Graph 3 - *Gender vs. Average Trip Length*



This indicates that females ride 4 minutes longer on average than men, and that these unspecified riders ride twice as much. They could be the day subscribers.

We also conducted a statistical analysis on ride length by yearly versus day subscribers.

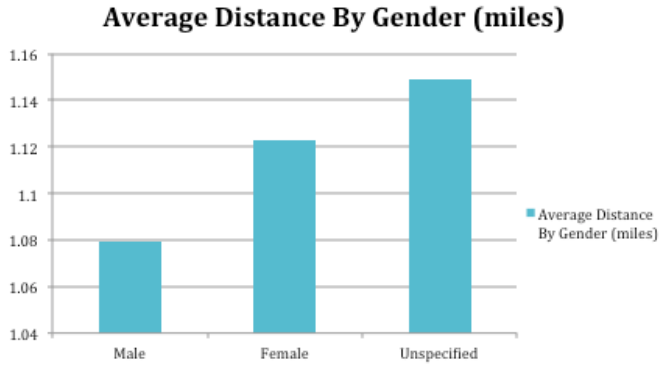
Graph 4 - *User Type vs. Average Time*



Graph 4 shows a strong correlation between day subscribers taking longer bike rides as opposed to their yearly subscriber counterparts. This strengthens our guess about the Unspecified genders being a part of the day subscribers.

We also studied the average distance traveled by different genders.

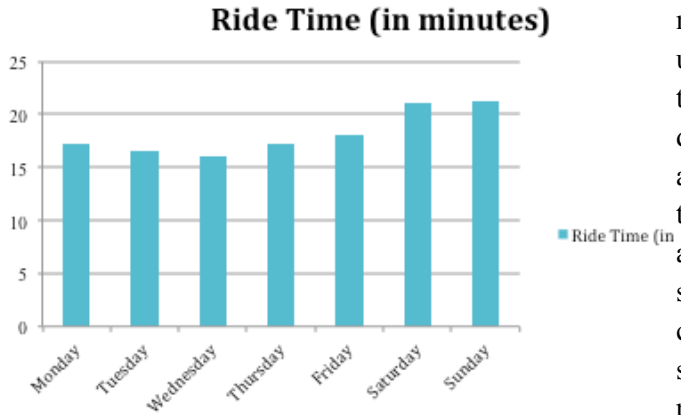
Graph 5 - *Average Distance by Gender*



For some reason, women ride for a longer distance than men do. Once again, the unspecified gender could indicate day users who chose not to enter this information.

Finally, we explored the average time on a bike according to the day of the week.

Graph 6 - *Average Time on Bike by Day*



It seems more people take longer bike rides during the weekends, with a dip in the middle of the week. This could indicate people have more free time or willingness to ride bikes near the weekends.

The Task

At first, we wanted to predict what type of subscriber (day or yearly) the user was by using the dataset, but found some features to be far too telling, such as the age and gender field. If either of these fields were blank, it was very likely the user was a day subscriber. Another telling feature was the distance; day subscribers tended to travel further distances than yearly subscribers by a significant amount. Also, the number of yearly subscribers far outweighed the number of day subscribers, so even if we randomly guessed "yearly subscriber" for all users, we would still have a pretty good accuracy; we would have an 88% accuracy to be precise.

So, our new, more challenging task is to predict the user's end location. We will take 50,000 data points as our training set, and then strip the ending location from the remaining 950,000+ data points to be our validation set to evaluate how well our predictive model performs. For our model, we are going to be using **linear regression** to see if there is a correlation between a set of features and the final end location. To quantify the final end location, we are going to be using latitude and longitude coordinates provided with the dataset, and then choosing the closest bike station as the end location.

The features that we are planning to use to train the model are the start location, the average distance a user travels, the day, and the time. The start location, the day, and the time are readily available in the dataset, so no further processing is needed. The average distance a user travels, however, is based on the statistics generated during our exploratory analysis. We believe that these four features hold the strongest correlation to the end location. We can combine the knowledge of the start location of a station and the average distance a user travels to predict where the user might end up.

Of course, getting the average distance is not enough. Distance is a scalar property, so we would not know the direction in which the user is traveling.

There could be many bike stations equidistant from the start location, and it would be inaccurate to just randomly pick one of them. Therefore, it is important to once again use the features to gain more context. If it is a weekday, it could be likely that users are headed towards large metropolitan areas. If it is around lunchtime, users are probably headed towards locations with a large number of restaurants. That's why features such as date and time are relevant for this linear regression model.

To evaluate our performance, we will have a baseline model. This baseline model will use an imaginary bike station located at the centroid of all the end station clusters. Then, it will take the average distance of all the trips combined and move towards the centroid. This baseline should perform pretty well because of the usage of clusters and centroids. Using the centroid will cause the imaginary center to be near the actual stations most people actually end at.

The Model

To predict where these bike users were going, we decided to use linear regression again. We chose not to use SVMs or logistic regression models because they are typically used for categorical data. We could have used a binary predictor if we only wanted to predict whether the user's destination was north or south and east or west of where they originally started.

First, we needed a training set, which we created by taking a random fifty thousand points. These points had to be random because the dataset was in chronological order; taking the first fifty thousand from the dataset would only get users from a limited time period. Randomization was the best way to achieve an even distribution. After we made our training set, we fit the data to a feature vector.

We optimized the feature vector by trying to minimize the average distance that we were off by in the model. This average distance is calculated by taking the coordinates of our prediction and finding the distance to the coordinate of the end station that the bike actually went to. The next thing we wanted to do was to find the closest station to our prediction. The only way we found to calculate this was to iterate over all the stations and find the closest station to our prediction. Unfortunately, because there were 500 stations, it became prohibitively expensive to do this. So, we no longer tried to find the closest station and instead relied on the average distance error to evaluate our performance. If we were able to, our solution could have been more accurate and made more sense.

As we began adding features to our vector, we found that there were a number of features in our dataset that didn't really help. We at first tried to use the bike id column of the dataset. Not only did this expand our feature vector to more than 500, but it also didn't make sense. In the real world, different people use the same bikes to get to different places. So, the bike ids did not correlate to the users themselves. This was a misread of the data, and our prediction showed it. In addition, we tried to categorize time into our findings.

We also tried to use days of the week, and whether it was a weekday or a weekend. Neither approach had any indication on how far a user was going to go in either direction. The hour was also another failing indication. However, we were able to adjust and make better use of this data by splitting it into six four-hour chunks instead of 24 one-hour chunks. This makes sense because people might be traveling inwards during some hour range of the day, maybe to work, and traveling outward during after-work hours. Since schedules vary vastly across users, it doesn't make sense to say that at 8:00 AM, everyone will be traveling to work. Rather, we treat it as a range of hours that could be interpreted as "going to work" time.

Lastly, we tried using both the gender and user type as part of our feature vector. User types is a binary category of either being a day customer or a yearly customer. However, it appears that day customers

aren't required to give their gender and age, and often don't. By using both of these features into our feature vector, we ran into another overfitting issue. We resolved this by removing the gender feature from our feature vector altogether. It's justified because our subscriber feature was indicative enough of the end location, while whether a user was female or male gave no indicator on which direction this person would travel.

Literature

The data is collected and published by CitiBike themselves and it seems that they started collecting the data to help put them back on financial track. The company was plagued by faulty check in/check out stations, and glitchy bad software that drove people away. Additionally, sometimes people would have to ride around looking for a spot, and they would get charged for it, or there would be empty stations, and no bikes. They hired a new CEO to help address these problems, and at first, he didn't care about operating in the red. His focus was improving the system, and we're sure they kept data to justify the expenditures.

Other transportation system datasets are generally used in the same way. In San Francisco, a bike sharing system's data was analyzed to see financial patterns for their business. For example, one of the things they looked for was a correlation between users using the bike share and whether the tram station lowered people's use of it. In addition, like us, they looked at whether or not a particular user was a 1 day user, a 3 day user, or a long term subscriber. However, in general, these datasets were used for predicting financial trends, and not to guess where a user was going.

However, one could generalize our findings and other people's findings to have similar qualities. How far someone goes indicates where bikes should be rerouted to, and whether stations needed to have more bikes, or another station needs to be built there. How long the person travels for, or what time of day it is, helps indicate these features. Additionally, revenue projections could be extracted from our use because you could see potential revenue loss or gain from switching to a length of use to the distance the bike traveled as a basis of cost. In that sense, the findings would be similar because there is a significant overlap in features used to calculate these desired outcomes.

The Results

The features that worked the best and that we included in our final model were the duration, the type of user (yearly or day subscriber), where the user was located with respect to the centroid station, which hour chunk it was in (as listed above), and the station id the user started from. By using these features, we found that day subscribers tended to go farther than their yearly subscriber counterparts. They tended to do this probably to maximize the dollars they spent for the day. These users also tended to be more recreational, and not people headed to work, which indicated a certain amount of leisure. In addition, using the duration in our feature vector greatly decreased our error, because there is a strong correlation between being out longer and traveling farther.

The next biggest indicator was where users were located with respect to the map. Users that were on the edge of the map tended to not go outwards, and instead went inwards. We quantified this by using the difference from their starting location and the centroid of the stations. The centroid of the stations was calculated by finding the center of the cluster of bike stations.

Lastly, we used which station the user started from because there were noticeable trends from start station to end station. This suggests that users leaving from the same area tend to go to the same area.

The model we chose had an average accuracy rate of within a mile, which was better than what the baseline predicted, which was within 1.2 miles. It's a good indicator because of the fact that, on average, the distance between stations was 0.15 miles away. Our model would have been improved even more if we

could have calculated the closest station to our predicted point, instead of just leaving it as a point. This is also better than trying to predict the station directly. That is, if given a list of 500 stations, we would predict whether or not the bike ended there. This means we would have had to train on every single station as a binary choice. After that, we would have chosen the one that had the greatest choice. However, we would have to run this calculation five hundred times instead of just twice, which was what we did for our model to calculate latitude and longitude.

While we were able to beat the baseline, it was only by about 17%. The baseline performed pretty well since we were taking the average distance traveled across all trips, and it seemed to even out, despite some groups (such as day subscribers) going for much longer trips. However, overall our model beat the baseline in almost all cases because it had more context about the direction to head towards, while the baseline just moved the average distance towards the centroid of the bike stations.

Ultimately, we think the data given to us wasn't enough to give a much better indication of where the user was going. We couldn't harness the power of the bike ids, nor could we really categorize the stations themselves. We wished for more telling features that would indicate what the user was doing. For example, if the user was in a heavy food district, they would probably return to a work place, or maybe return home. Even a user id would have been very helpful to analyze user trends, because it was likely that if a user has visited a place before, they would visit it again. The only trends we were able to harness were types of user, what time of day it was, and where the station was located. We were able to get an improved version of the baseline, but the task we wanted to predict probably didn't fit the data that was given. However, it was still a very interesting exploration of the data, and is a lot less trivial than predicting a user type.