Citi Bike: Trip time and gender

Sarah Schoengold[1]

[1]NYU Center for Urban Science & Progress

November 7, 2017

Abstract

We wanted to investigate the larger question regarding how citibike trip times related to the gender of the rider. Using the dataset from July 2017, we looked at the distributions of male and female riders. We found that the distributions of male and female riders are not statistically different from each other using the KS test. This knowledge is important when interpreting Citibike data in our future analysis.

Introduction

We started with a general question: Is the average trip time for Citibike affected by the gender of the rider? Our null hypothesis, then, is that the average trip time for male and female riders will be exactly the same. Our experiment will clean and visualize the data, and work to falsify the null using a statistical test that will tell us if the distributions of the two genders come from the same parent distributions. Understanding their relationship will be beneficial as context for future analyses using the Citibike data.

Data

The data we used was from the month of July 2017. It's open data available on the internet, but we chose to pull it from the NYU CUSP data facility gateway server. We read in the data with this format, so it's reproducible:

```
df = pd.read_csv("/gws/open/Student/citibike/201707-citibike-tripdata.csv.zip")
```

After reading in the data, we created a new data frame with only the columns relevant our question, the "tripduration" and "gender" features. From there, we cleaned out the outliers to help visualize our data. We assumed all trips greater than 5000 seconds were not relevant for our analysis.

After visualizing the data as a whole, we divided the data into separate distributions based on gender — one for male, one for female, and one for unknown riders. The subsequent distributions of each gender look like the following:

To understand the distributions a bit more, we can look at both the average trip time and standard deviations for each gender. By describing the distributions in python, we can quickly compare their moments.

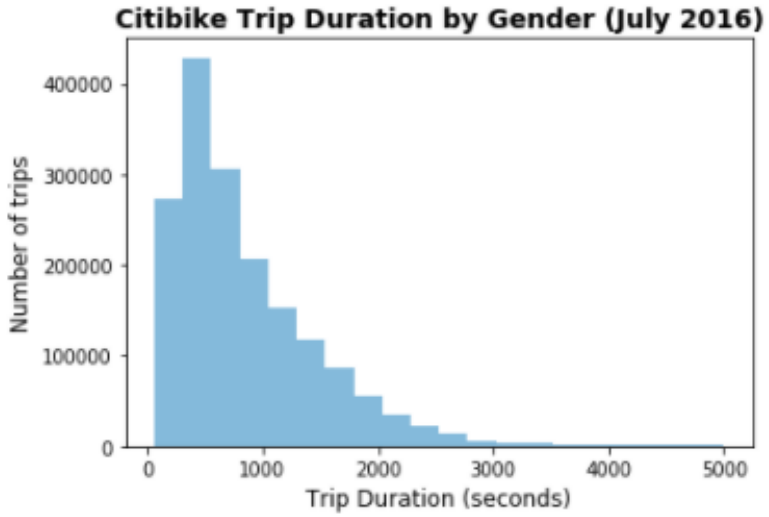

Figure 1: This shows the distribution of trip duration for all citibike rides in July of 2016, after outliers have been removed. We notice in this distribution that the average trip duration appears to be less than 900 seconds. This is for all riders, regardless of gender.

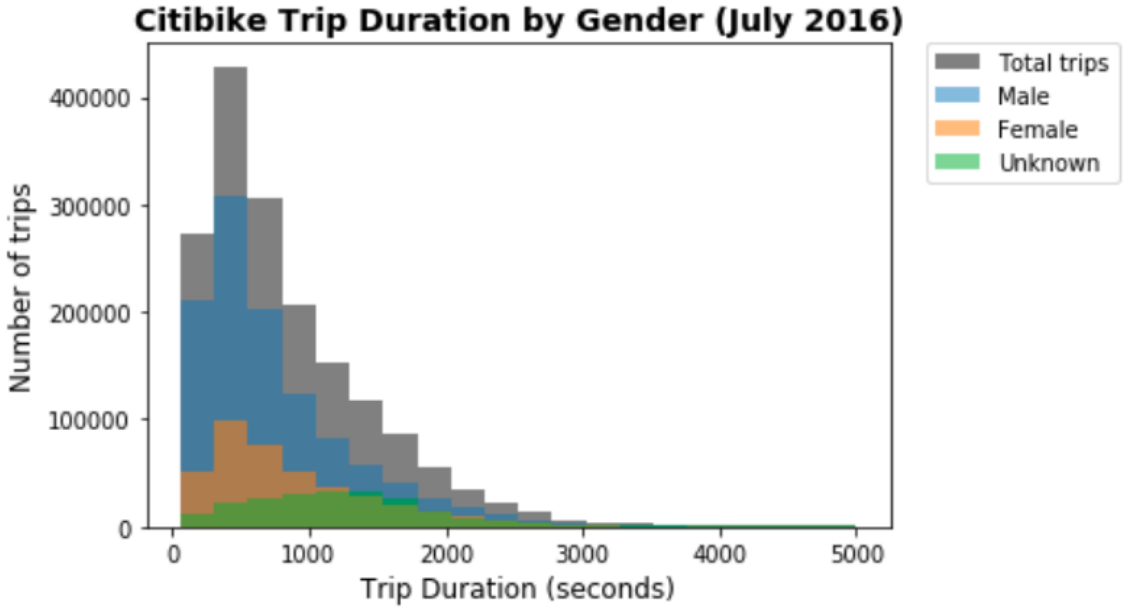

Figure 2: The figure shows the distribution of trip durations by the number of trips for July 2016 Citibike data. In gray we see the total numbers. In blue, we see the distribution of male riders. In orange, we see the distribution of female riders. While it appears that the number of trips is much less for females than for males, the shape of the distributions themselves look similar. In order to understand if their distributions are significantly different, we'll have to do testing.

Methodology

Since we aim to understand if there's a difference between these two distributions, I chose the kolmogorov-smirnov (KS) test for two samples. The KS test is a two-sided test for the null hypothesis that the two

```
dfM.describe()
```

	tripduration
count	1.095725e+06
mean	7.585743e+02
std	5.717199e+02
min	6.100000e+01
25%	3.550000e+02
50%	5.840000e+02
75%	9.960000e+02
max	4.999000e+03

```
dfF.describe()
```

	tripduration
count	399225.000000
mean	875.134360
std	617.219874
min	61.000000
25%	424.000000
50%	701.000000
75%	1171.000000
max	4997.000000

Figure 3: Here we can see the moments for the distribution of male riders (left) and female riders (right). Notice that the means are 758 and 875, respectively. The standard deviations of the two distributions are 572 and 617, respectively. With these discriptive statistics, we can better understand the relationship between these distributions.

samples come from the same distribution. This test compares the two samples to generate a KS statistic and a p-value.

Other tests that compare two samples are the Chi-Square test, which also was an option to test these distributions' relationships. This other test was recommended by my peer reviewers, including Praveen and Colin. Although Chi-Square would certainly work, I chose to run the KS test because it's easier to execute using Python, my statistical tool of choice.

In order to run the test, I first created a sample of the distribution I created for male riders so it would be the same size as that of the smaller, female distribution. For this, I employed the random.choice function from the numpy package, using (14) as my random seed to ensure reproducibility.

```
Msample = np.random.choice(dfM.tripduration, size=798450, replace=False)
```

After making sure I was comparing distributions with the same size, I ran the KS test with a single line of code:

```
scipy.stats.ks_2samp(dfF.tripduration, Msample)
```

The output of the test displays both the KS statistic and a p-value. **The KS statistic is: 0.09456 and the p-value is: 0.0.**

Conclusions

The results of this test show that there is no statistical difference for the trip duration from female and male riders. Since the p-value is zero we can reject the hypothesis that the distributions of the two samples are the same. This result is harmonious with our descriptive statistics which showed similar means and standard deviations. It also makes sense with the data visualization we generated. Although there is a large difference in the number of riders, the trip distribution follows the same shape.

This test could be pushed farther, utilizing other statistical tests such as Chi-Squared, or factoring in errors in the samples. We also could expand this to look at more months of the year — perhaps the distributions differ more in different seasons.