Citi Bike Project

Bianca Brusco[1]

[1]New York University (NYU)

November 6, 2017

Abstract

Bike sharing has gained in popularity as a mean of transportation in urban systems. In New York City, data from Citi-Bike usage is publicly available, so trends in riderships can be investigated. In this project, we use one month's data to examine riding trends for Citi-Bike subscribers and occasional users of the service, to understand whether there is a difference in the likelihood of taking shorter trips between the two groups. The results show that subscribers are indeed more likely to take shorter trips. A possible implication of this result is that Citi-Bike subscribers are choosing it as a mode of transport even when needing to cover shorter distances

Introduction

New York City has introduced a bike-sharing system in 2013: Citi-Bike. The service can be used either by subscribing with an yearly membership or by purchasing an occasional pass. The two groups of users are defined as subscribers and as consumers. In this project, we investigate whether the ratio of trips longer than average to trips shorter than average is smaller for subscribers than for consumers. Indeed, the first group might be more likely to use Citi-Bike for shorter trips, as there is no extra charge per trip, while consumers might decide to get a pass only to cover more substantial distances.

Data

Citi-Bike Data is publicly available at https://www.citibikenyc.com/system-data. For this project, we use ridership data from one single month: January 2015. Processing of the data has been completed with Pandas for Python 3.

The dataset includes 285552 observation, which are divided into 279924 for the subscriber group and 5628 for the customers group.

We observed distribution of length trip for the two groups:

Methodology

We compute the mean trip duration for the whole sample, which is of 10.96 minutes. Therefore, we define a long trip to be a trip with duration longer than 11 minutes, and a short trip to be a trip with duration shorter than 11 minutes.

We test the null hypothesis that the mean trip length for subscribers is equal or longer than the mean trip length for consumers, against the alternative hypothesis that the mean trip duration for subscribers is shorter than for consumers. We use a significance level of $\alpha = 0.05$

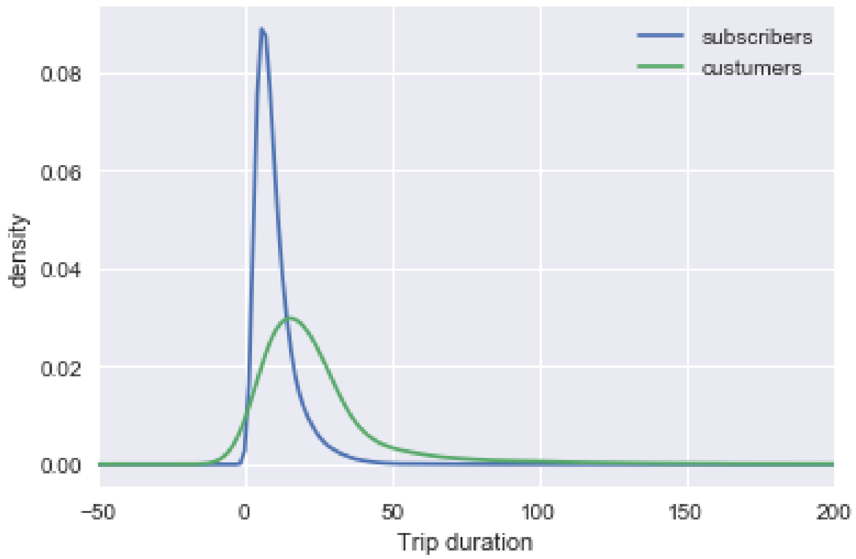

Figure 1: Distribution of trip duration for subscribers and customers.

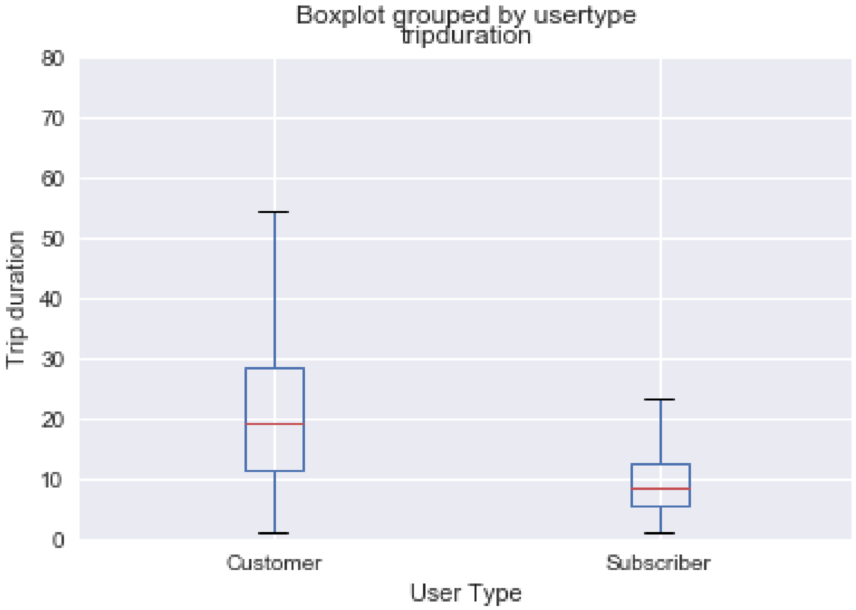

Figure 2: Box plot for trip duration, by user type.

Indeed, we are testing

$H_0 : \frac{S_{long}}{S_{short}} \geq \frac{C_{long}}{C_{short}}$

$H_A : \frac{S_{long}}{S_{short}} < \frac{C_{long}}{C_{short}}$

To investigate the difference in means, we can use a Z-test, as we are investigating whether two sample proportions appear to come from the same population or not.

An alternative would be to use a Chi-square test. But since we are only testing two proportions, the results should be the same for both tests.

Conclusions

The samples we have used from the population have different proportions of longer to shorter trips.

We can observe the distribution of shorter and longer trips by user group with the bar plots below, first for absolute counts and then normalized.

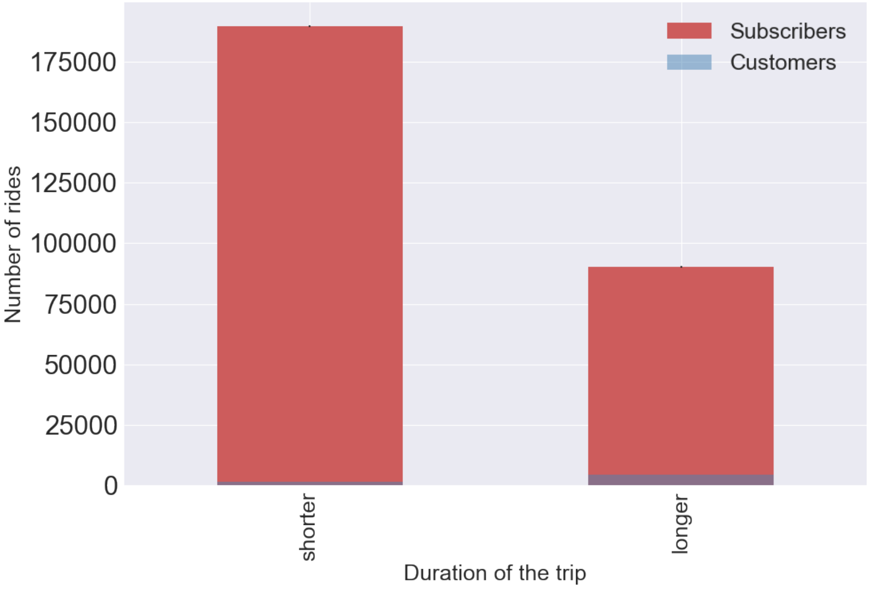

Figure 3: Distribution of Citibike bikers by user type in January 2015, absolute counts, with statistical errors

We conduct a Z test comparing the two proportions, and obtain a Z-score with absolute value of 414.45. This Z-score is significantly larger than the Z-score for 0.05 significance level, which is 2.96. Therefore, we reject our null hypothesis, and conclude that the ratio of longer/shorter trips is smaller for subscribers than for customers.

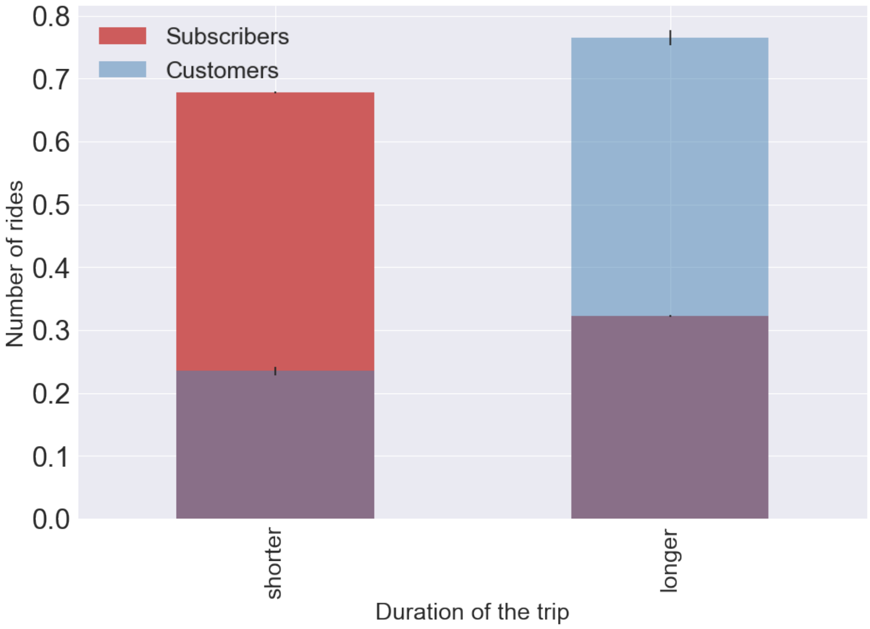

Figure 4: Distribution of Citibike bikers by user type in January 2015, normalized. We see that while subscribers take more shorter trips, consumers take more longer trips.