Analysis of Citi Bike user type and average trip duration

Emily Hansen[1]

[1]NYU Center for Urban Science & Progress

November 9, 2017

Abstract

The Citi Bike program has revolutionized public transportation across New York City, with hundreds, if not thousands, of crowd-shared bicycle trips occurring daily. An analysis is performed on Citi Bike records to compare the mean travel times between users subscribed to the service and users employing the service as a one-time customer. It was found through a t-test for difference in means that one-time customers rode Citi Bikes for a statistically significantly longer period of time, on average, than subscribed users. This supports the idea of subscribers taking shorter trips for the purposes of neighborhood commute.

Introduction

The Citi Bike program is the largest bicycle sharing service in a United States major city, spanning several boroughs and providing over 12,000 bikes at over 700 stations to New Yorkers as of October 2017. Given its increasing popularity and the population density of New York City, it is possible to analyze large numbers of trips taken across the city over time and gleam insight from ridership activity. Citi Bike classifies its users as "subscribers"– with a monthly subscription– or "customers"– one-time service users. Knowing any difference between how long subscribers vs customers tend to spend on their routes will help Citi Bike improve their services to cater to clusters of office spaces or tourist destinations, for example.

Data

The data utilized are from an open, downloadable index of trip data published by Citi Bike. The data cover a single month– December 2015. The data were read in as a data frame and were cleaned to separate out user type and trip duration from the rest of the information. Their distributions are seen in Figures 1 and 2 below. The remaining data were grouped by user type, and the average trip duration for each group was calculated.

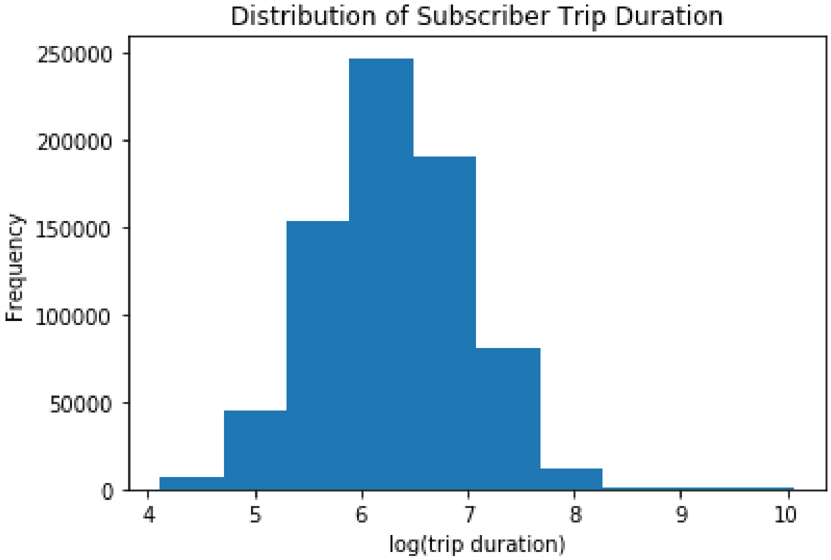

Figure 1: Distribution of subscriber trip duration at log scale. Note the approximately normal distribution.

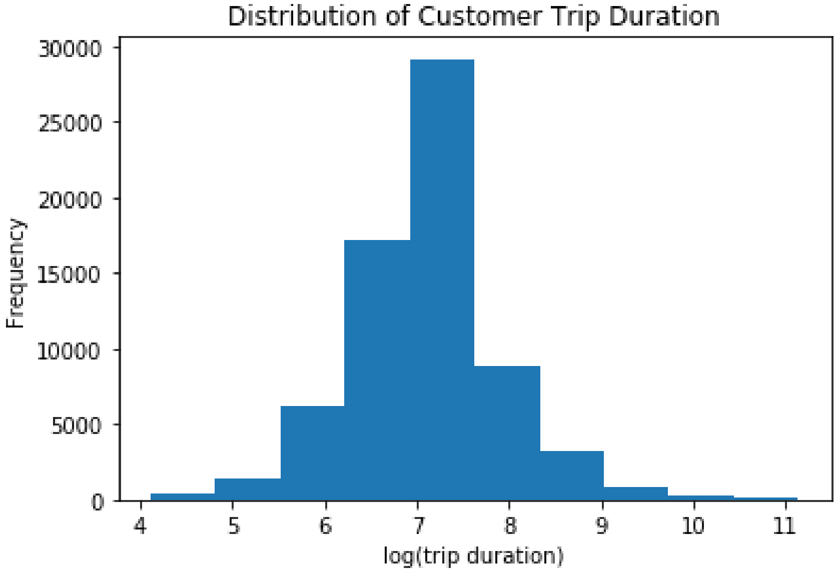

Figure 2: Distribution of customer trip duration at log scale. Note the approximately normal distribution.

Methodology

The test chosen to determine the significance of the difference between both groups' averages was suggested by a peer, Jon (jlk635), to be the t-test for difference in means, with the assumption that the trip times follow a normal/t-distribution as observed. This was tested under the null hypothesis that the average trip duration of subscribers is the same or greater than the average trip duration of customers. The bar plot of average trip duration is shown below in Figure 3, where it appears that the subscriber trip duration is on average much less than that of customers. With a significance level of 0.05, the standard error and degrees of freedom were used to calculate a p-value.

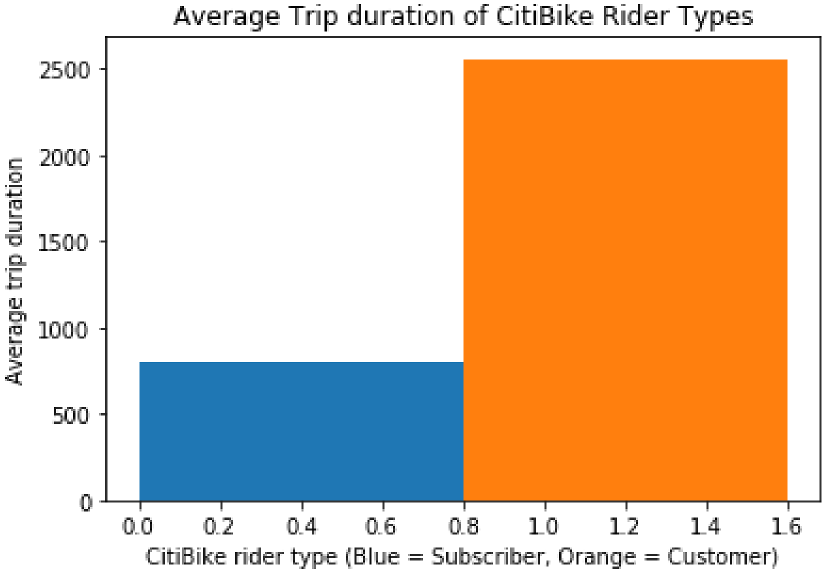

Figure 3: Bar plot showing the average trip duration of Citi Bike subscribers (left) versus Citi Bike customers (right). Customers show a longer average trip duration.

Conclusions

Given the plotted means, it was expected that the means of the two distributions were not equal. This was indeed the case. The difference in means was approximately 29.3 minutes, with the customers having a longer average trip duration. The t-statistic calculated was 20.52. This huge t-value is not surprising considering the large observable difference in means and the large sample size. The calculated p-value is extremely close to zero, with Python returning 0.0, indicating that the null hypothesis can be rejected and that the customers have a statistically significantly longer average trip duration.