Homework 8 Assignment 2: Citibike with Subscribers and Their Trip Durations

Po-Yang Kang[1]

[1] Affiliation not available

November 6, 2018

Abstract

This paper is part of a university project in finding out whether there is or is not a difference between the average trip duration of subscribed citibike riders or non-subscribed (customer) ones. The result is that the null hypothesis that there is no difference has been rejected in the favor of the alternative that there is a difference between the trip duration of these two kinds of riders. A two-sided t-test has been used, as well as a K-S test. All are done and tested through the cusp.adrf environment.

Introduction

Citibike is a bike renting service in New York City that caters to any rider, including both a regular customer who can pay directly for their ride, and therefore pays for each ride, and a subscriber who subscribes for a certain time frame, and therefore can enjoy limitless rides within that time period for a specific price. Because of these two different classifications, I am especially interested whether there are any differences in trip duration between them: Do subscribers ride more or less than a regular customer, or about the same? Is the fact that the subscriber subscribes encourages them to ride more, or because their subscription is cheaper than if a customer rides everyday, convince them to be lazy and ride less? I therefore tested this, testing the null hypothesis that there is no difference between the average trip duration of customers and subscribers with the alternative hypothesis in that there is a difference between average trip duration of these two groups.

Data

I have used the citibike open data API, specifically from the month of June in 2013. The variables included includes the trip duration, start time, stop time, start station id, start station name, start station latitude, start station longitude, end station id, end station name, end station latitude, end station longitude, bike id, user type, birth year, gender, and date of riders. Since I am concerned with the trip duration and the user type that distinguishes customers and subscribers, I dropped all variables but those two. I first made a bar graph comparing these two by their trip counts:

However, I figured out it might be better to measure by average ridership instead of count, since there seems to be more riders for subscribers than customers.

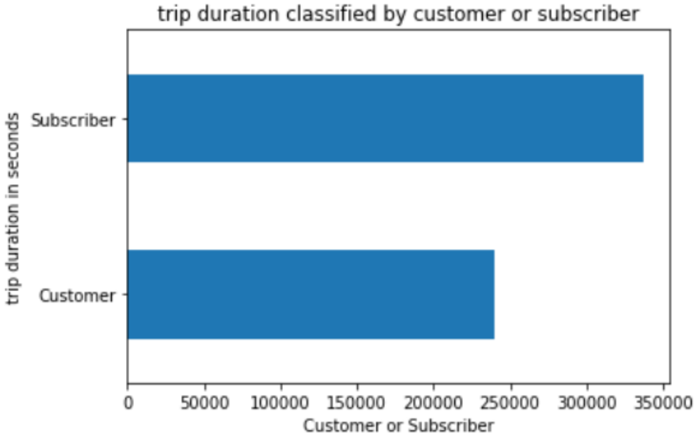

Figure 1: This is trip duration classified by whether the rider is a customer or subscriber. The subscriber's trip durations seems to be about 100,000 seconds more than the customer's. However, we need to continue on with the statistical analysis to confirm if we should reject the null hypothesis.

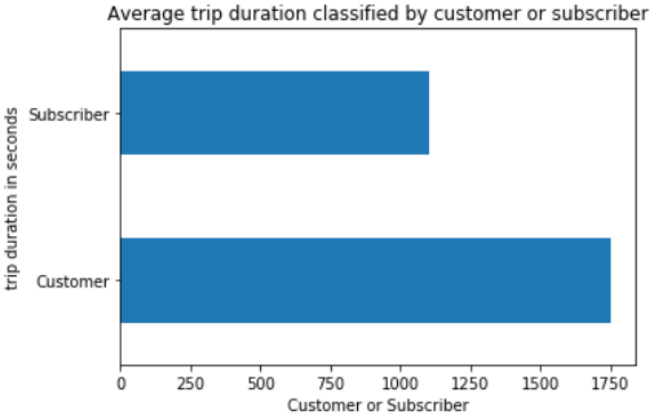

Figure 2: This is trip duration classified by whether the rider is a customer or subscriber, this time by their average trip durations. The customer's average trip durations seems to be about 700 seconds more than the customer's.

I therefore used the average trip duration in the test.

Methodology

Null Hypothesis:

The average of all subscriber's trip duration is about the same as the average customer's trip duration

Null Hypothesis: The average of all subscriber's trip duration is about the same as the average customer's trip duration

$\mu_{\text{S tripduration}} = \mu_{\text{C tripduration}}$

Alternative Hypothesis: The average of all subscriber's trip duration is longer than the average customer's trip duration

$\mu_{\text{S tripduration}}$ [?]$\mu_{\text{C tripduration}}$

testing with significance level

$\alpha = 0.05$

I was suggested to use a t-test by a classmate, and to focus only on the average trip durations, with the independent variable being the type of user, which is a dichotomous variable, being either customer or subscriber, and the dependent variable being the average trip duration. In this case, for the t-test, the degrees of freedom is 577701, and tested at the 0.05 significance level, the t-statistic has to be higher than the critical value of 1.96. A t-test is used over the other ones because of the dichotomous independent variable and the continuous dependent variable, testing only the difference between two groups, which is a classification to use the t-test over other tests.

I first tested it using the python scipy function, called scipy.stats.ttest_ind. The result that I got was a t-statistic of 27. It seems therefore that it is greater than the critical value of 1.96. However, this function is only accurate if the distribution and variance are the same in these two samples I am testing. I therefore did a K-S test to check the distribution sizes and see if they are similar.

If the K-S statistic is small or the p-value is high, then we cannot reject the hypothesis that the distributions of the two samples are the same. However, the K-S statistic is 0.27 with a p-value of 0.0, so therefore, since the p-value is low and the statistic is high, we reject the hypothesis that the distributions of the samples are the same.

In this case, we need to change the scipy function a little bit, by setting the equal_variance to False. The result that I got was a t-statistic of 23.821 and a p-value of 2.643818232977887e-125

Conclusions

It seems therefore that the t-statistic of 23.821 is greater than the critical value of 1.96. Also testing it with the function gave us a p-value of 2.643818232977887e-125, very much close to zero, and much less than the significance of 0.05. Thus, we can safely reject the null hypothesis that there is no difference of trip durations between customer and subscriber riders in favor of the alternative that there is a large difference.

However, there are some limitations using the t-test with the trip duration variable. It can tell us about the trip duration difference, but the variable trip duration tells nothing but how much time the rider is using the bike. For example, in this sense, we cannot conclude from this test, by using the trip duration variable, that the distance covered for each trip for a customer or subscriber is different from each other. None of the variables here gave us the distance. Trip duration gave us a specific idea for it, but so does the longitude and latitude of the docking stations. In this case, if one wish to calculate distance, one has to calculate the distance between these points and create a new variable for it.

Po-Yang Kang's github link