Average Age of Citi Bike Users Based on Gender

Matthew Sauter[1], Valeria Jimenez Schiavon[1], Kent Pan[1], and Tarek Arafat[1]

[1]NYU Center for Urban Science & Progress

November 5, 2017

Abstract

We set out to determine whether the average age of female Citibike users (based on self-reported gender identity) was significantly different from the average age of male Citibike users, in accordance with our hypothesis. Using a two-tailed independent t-Test, we determined that there **is** a significance in the difference between the average age of male and female users with a significance level of 0.05.

Introduction

Citi Bike, introduced to New York City in 2013, is a privately owned bike-share system expanding throughout the boroughs of the City. Since its inception, use of the bike share system has grown significantly, which has led to our inquiry into the average age of male and female riders and whether there is a significant difference between the two.

The marketing rationale worth exploring is the gender differences that may impact customer base. Citi Bike as a service wants to reconcile the reasons for restrictions in user uptake. If some users are not willing to use the service, there is a market expansion potential. For example if user disparities are based on perception of rider safety, proximal access, or income based preferences, these present potential market expansion opportunities for Citi Bike.

Data

We considered one month of data, July 2017, dropping extraneous observed variables so that the only two that remained were relevant to our research: birth year and gender. We subtracted birth year values from 2017 in order to determine the age of each rider. The data contained a few unrealistic observed ages that were likely incorrect or misreported (one rider was listed as being 143 years old), so we removed all far outlying data points with an observed age greater than 79.

Methodology

Figure 2 below shows the average age of male and female riders from our July 2017 data set. In order to test for significance, we used the two-tailed independent t-Test. Our independent variable (age) is categorical, and our dependent variable (age) is continuous. Based on this information, we determined that we could use either the t-Test or an ANOVA. Because the ANOVA is more complex, we decided that the simpler t-Test would be more appropriate.
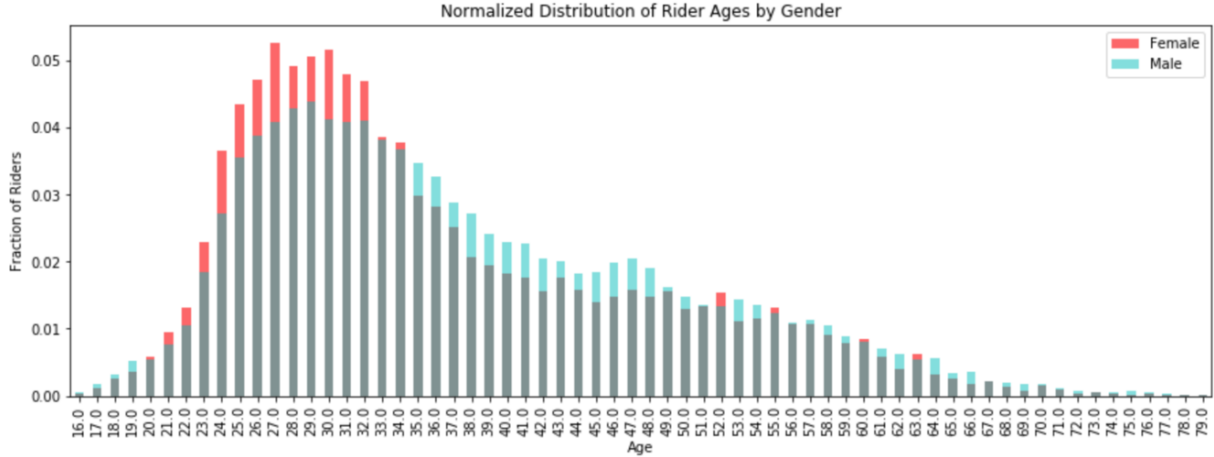
Figure 1: Fig. 1 shows the distribution of ages based on gender. Originally the spread went up to 143 years old, which was an obvious outlier. We decided to only consider reported ages below the age of 80. As you can see, there is a notable increase in female ridership between the ages of 20 and 34. This lends itself to our hypothesis that there will be a significant difference between the average ages of male and female riders.

	Gender	Average Age
1	Male	37.770700
2	Female	36.255908

Figure 2: Fig. 2 shows the calculated average age for each gender in July 2017 Citi Bike riders. The two-tailed t-test was used to determine if the difference between the average ages of male and female riders is statistically significant. At a significance level of 0.05, the t-test allowed us to reject our null hypothesis and assert significance.

Alternatively, we were encouraged to use the z-Test and the Chi-Squared test in conducting our statistical analysis. For this experiment, we cannot use Chi-Squared because age is not categorical and we are not testing a proportion, and we cannot use the z-Test because we do not have the population parameters. Thus, we concluded that the independent t-Test was our best fit. At first we were considering a one-tailed t-test to determine if female riders had a significantly lower average age than male riders, but eventually determined that we did not have a strong enough reason to pre-determine the direction of our significance, thus we used the two-tailed t-test.

Conclusion

Based on the outcome of the t-Test, we were able to reject the null hypothesis and assert that there is a significant difference between the average age of female and male users. To strengthen the analysis, we could increase the number of months that we consider in our dataset. To add additional value to this analysis, we could try to identify whether there is a significant difference between the ages of users using Citi Bike recreationally versus those using Citi Bike for commuting to work, based on gender. This could help to

better inform the user base and where future stations should be built.