Citi Bike Data Analysis

Fangshu Lin[1]

[1]Center for Urban Science + Progress

November 9, 2017

Abstract

Citi Bike is a bike sharing system which is widely used by people in New York. It is important to analyze the riding patterns in order to provide better services. In this analysis, I investigated the age groups of night ridership. Using Z test and Chi-square test, the proportion of younger riders biking at night was compared with older riders. The tests results show that younger people are more likely to use the services at night hours. The analysis can facilitate the allocation of bikes at night hours.

Introduction

City Bike is a bike share system serving New York City and Jersey City which was launched in 2013. Riders can get access to Citi bikes at hundreds of stations in New York City. The system operates 24 hours a day, 365 days a year (https://en.wikipedia.org/wiki/Citi_Bike).

Since Citi Bike services are widely used by New Yorkers, it is important to understand the riding patterns in order to provide better services. For example, analyzing the user types will facilitate the allocation of Citi Bike stations based on the demographic information of different neighborhoods. In this Citi bike analysis, I will examine the day and night ridership. Specifically, I am interested to find out if older people (>=40) are less likely to use Citi bikes at night (19:00-5:00) compare to younger people.

Data

The data used is Citi Bike trip data which is published on the Citi Bike website (https://www.citibikenyc.com/system-data). Variables included in the trip history data are listed below.

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

In the analysis, Citi bike trip data of March 2015 was used. Data of July 2015 was selected to verify the analysis results. Data was cleaned to only include the variables needed which were the start time in datetime format and age of riders.

First, the distribution of the number of trips during 24 hours was plotted. The riders were separated into two age groups with the threshold of 40 years old. Figure 1 shows the absolute counts of the number of trips and Figure 2 shows the normalized results. Uncertainties were calculated using Poisson statistics. Figure 2 shows that the two groups of riders do have different patterns when using Citi Bikes. Younger riders tend to use the bikes more at night hours. We need to include statistical tests to test the hypotheses.

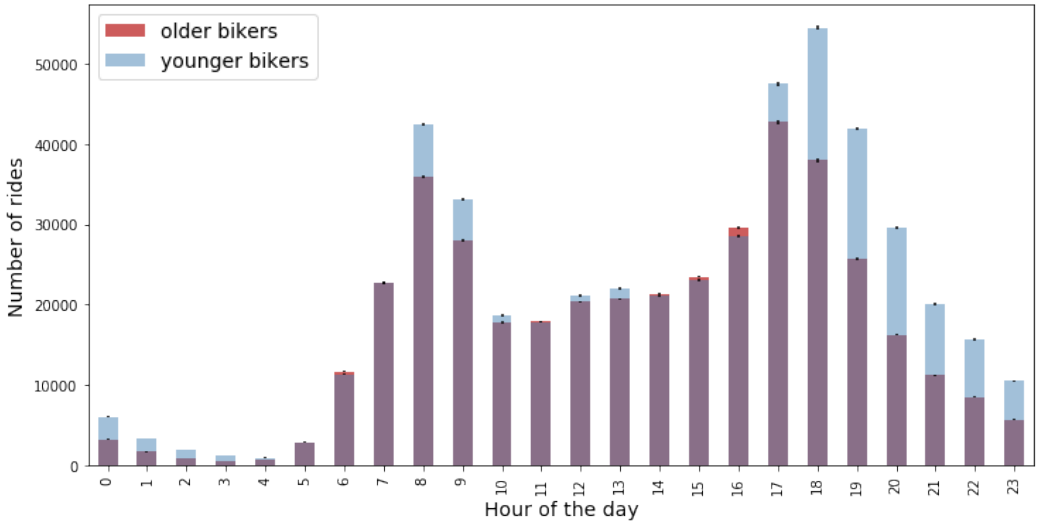

Figure 1: Distribution of Citi Bike riders by age in March 2015 (absolute counts with statistical errors)

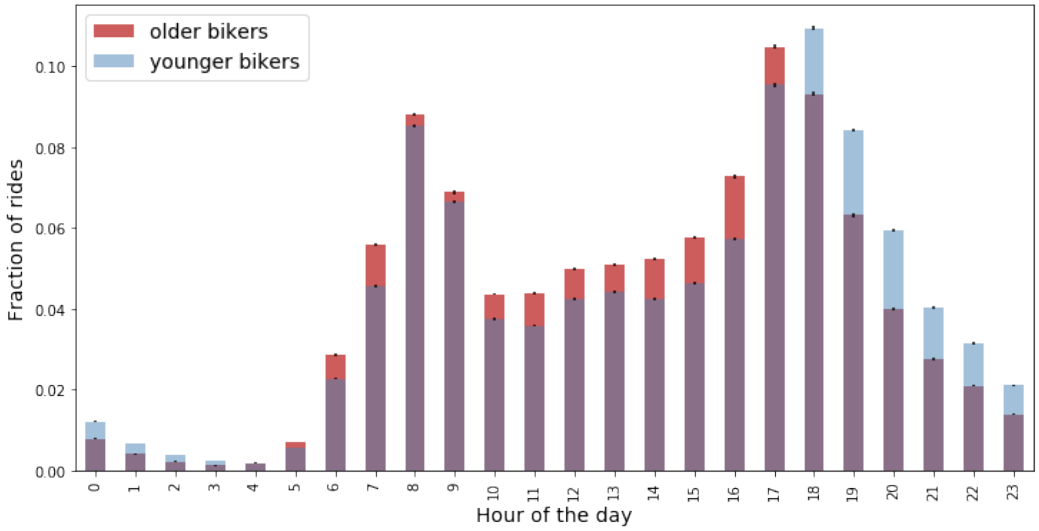

Figure 2: Distribution of Citi Bike riders by age in March 2015 (normalized)

Methodology

In order to investigate the night ridership by age groups, I need to compare the ratio of trips at night over the total number for two age groups, therefore I chose one-tail two-propotion Z test. In the review, Emily suggested a null hypothesis that the proportions equal to each other because of the lack of knowledge about the two proportions before testing. Therefore, in the second part, a chi-square test was also perfomed as suggested in Emily's review.

The proportions of biking at night and the sample sizes for two age groups are listed below, where Y represents younger riders and O represents older riders.

$P_0 = \frac{Y_{night}}{Y_{total}} = 0.219, \quad n_0 = 160822$

$P_1 = \frac{O_{night}}{O_{total}} = 0.141, \quad n_1 = 168151$

For both tests, the significance level $\alpha = 0.05$ is chosen.

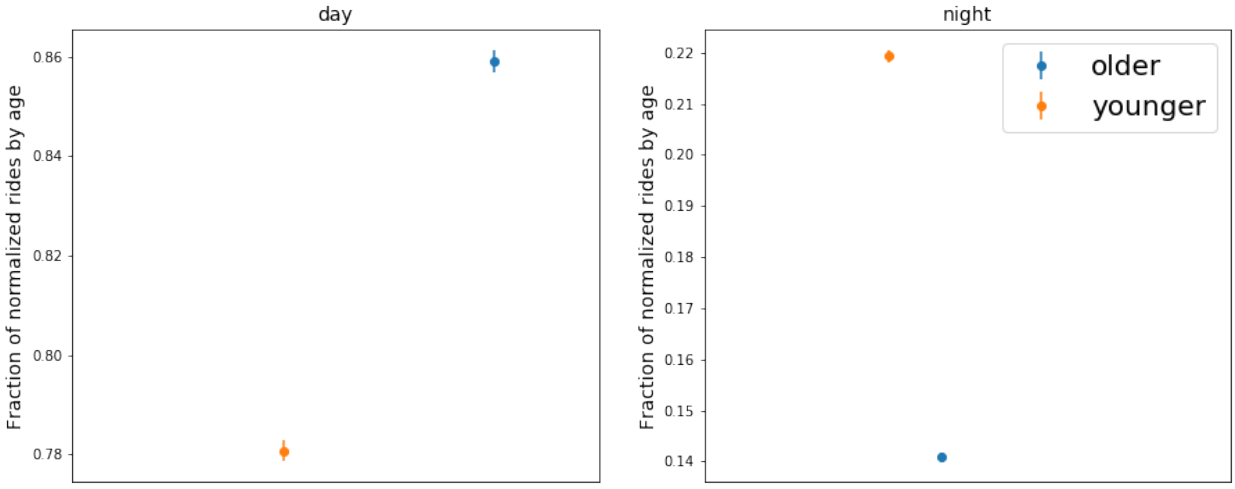

Figure 3: Fraction of Citibike bikers by age in March 2015 for day time (left) and night time (right)

1. Z test

Null Hypothesis:

The ratio of older people (>=40) biking at night over all older people biking during a day is the same or higher than the ratio of younger people (< 40) biking at night to all the younger people biking during a day.

$H_0 : \frac{Y_{night}}{Y_{total}} \leq \frac{O_{night}}{O_{total}}$

$H_a : \frac{Y_{night}}{Y_{total}} > \frac{O_{night}}{O_{total}}$

The Z statistics is:

$z = \frac{(p_0 - p_1)}{SE}$

$p = \frac{p_0 n_0 + p_1 n_1}{n_0 + n_1}$

$SE = \sqrt{p(1-p)(\frac{1}{n_0} + \frac{1}{n_1})}$

2. Chi-square test

Null Hypothesis:

The ratio of older people (>=40) biking at night over all older people biking during a day is the same as the ratio of younger people (< 40) biking at night to all the younger people biking during a day.

$H_0: \ \frac{Y_{night}}{Y_{total}} = \frac{O_{night}}{O_{total}}$

$H_a: \ \frac{Y_{night}}{Y_{total}} \neq \frac{O_{night}}{O_{total}}$

The chi-square statistics is:

$\chi^2 = \sum_i \frac{(observation_i - \exp ectation_i)^2}{\exp ectation_i}$

The contingency table is listed below.

Bike at night	Bike at night time	Bike at day time	Total
Older riders	0.14×168151	0.86×168151	168151
Younger riders	0.22×160822	0.78×160822	160822
Total	58929	270044	328973

Table 1: Contingency table

Conclusions

The calculated Z statistics is 58.61. The largest number reported in Z table is 3.5 that gives a p value of 0.0002, which is smaller than 0.05. The results show that we can reject null hypothesis with a p value close to 0. The proportion of older riders biking at night is significantly smaller than the proportion of younger riders biking at night.

The calculated chi-square statistics is 3434.787, which is much larger than the threshold 3.84. We can reject the null hypothesis. The proportion of older riders biking at night is significantly different from the proportion of younger riders biking at night.

Then I used the dataset from July 2015 to verify the results. Null hypothesizes were rejected with Z statistics and chi-square statistics equal to 91.31 and 8338.16, respectively. The results are robust that younger riders are more likely to use Citi bikes at night in warmer weathers as well.

The test results show that younger riders are much more likely to use Citi Bike services at night hours than older riders. This information is useful when allocating the number of bikes. For example, we can expand the number of available bikes at the stations near the neighborhoods where have more young people at night hours, such as the stations near universities in New York City.