CitiBike Data Analysis Project

Marium Sultan[1] and Baoling[2]

[1]Affiliation not available
[2]New York University (NYU)

November 9, 2017

Abstract

Our idea, based on our own life experiences as females being warned to be careful about being out at night, was that females are less likely than males to choose biking at night versus in the day.

$H_0 : \frac{M_{\text{night}}}{M_{\text{total}}} <= \frac{F_{\text{night}}}{F_{\text{total}}}$

$H_1 : \frac{M_{\text{night}}}{M_{\text{total}}} > \frac{F_{\text{night}}}{F_{\text{total}}}$

We used Citibike's own data for the first 8 months of 2017 to do our analysis. Our research showed that we could reject the null hypothesis at an alpha value of 0.05.

Introduction

Citibike is a major bike-share system in New York City. People can rent it hourly, daily or even subscribe the membership for a longer period. Citibike has had over 50 million rides1. We wondered if gender (coded in the data as male, female or unidentified) caused a difference in biking patterns over a 24 hour span. We already know females bike less than males on average. Many females have heard warnings in their life not to go out late at night or be alone after dark. We decided to look at if there were less females riding at night than men on average. Night here, for lack of seasonal precision, is defined as 7pm - 5am, after and before morning and evening commutes. If females do ride less than males at night what is the reason behind this? Does it point to systemic problems with feeling unsafe during those darker and less busy hours? If so, what can be done in NYC as a whole to remedy this?

Data

Our dataset covers Citibike records from January to August 2017, taken from the Citibike website. As there were inconsistent column names across months before we could merge we had to rename them. At this point we decided to only put meaningful names on the columns we did not plan to drop, and so the rest were lettered 'a' to 'l'. After merging we sampled only 5% of the records to do the analysis or the dataset will be too large to fit code-running. Keeping only starttime, stoptime, and gender, we made a new column with month and one with type of rider. To get the type of rider (night vs day) we ran a datetime function on starttime and categorized all the riders whose start times that fell into our night hours as nightriders and the rest as dayriders.

Methodology

At the beginning, we calculated the ratio of male night riders to female night riders and determined there was enough uncertainty to go ahead with our test. We then ran a Z test with 95% confidence level to measure the significance of the difference. We chose the Z test instead t test because the sample size (518418 observations) is big enough to call it a population. Afterwards, we break down the ratio of night riders by

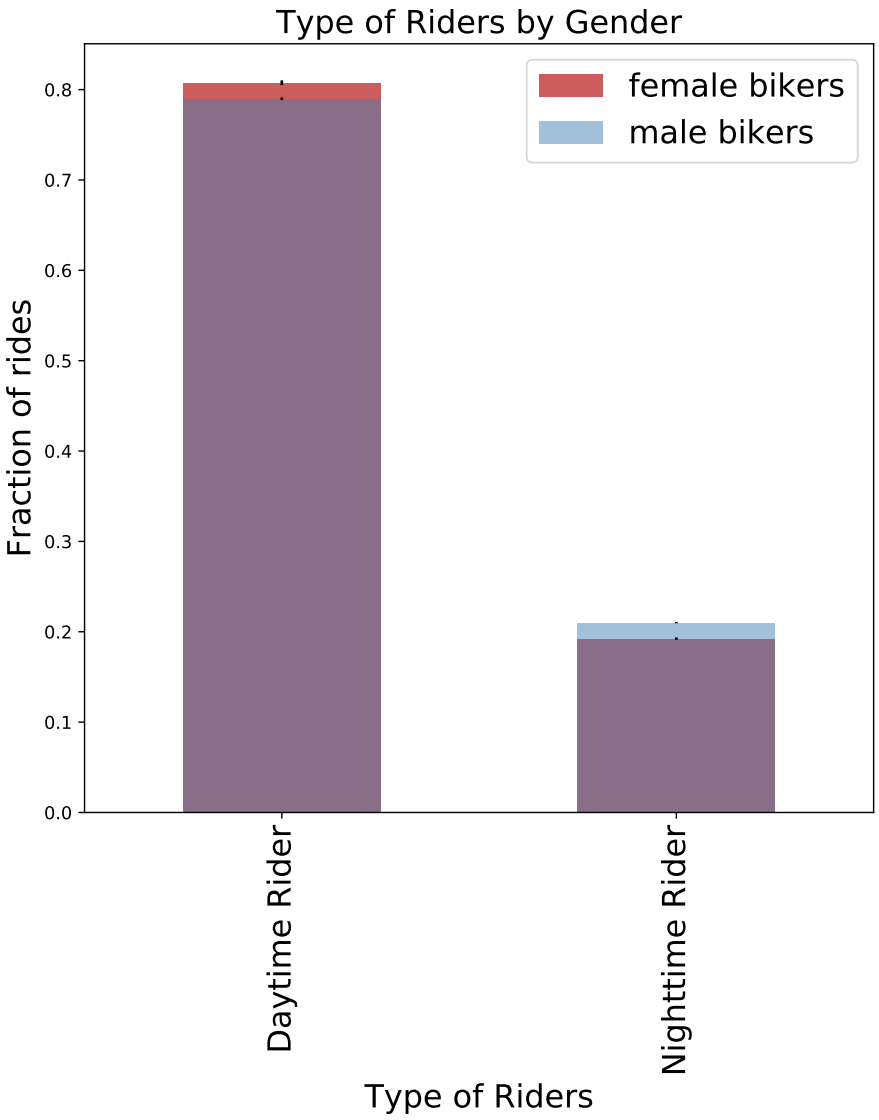

Figure 1: This histogram shows the fraction of night riders by gender. We can see that men are more likely to ride bike during nighttime than women do in the context of the first 8 months in 2017.

months and test the significance of the difference between male and female by month individually, producing a z test result for each month.

Conclusion

According to our Z test, it shows that the % of male nightriders is significantly higher compared to females in each month or in general in the context of Citibike records from Jan to August in 2017. This may indicate that women have saftey concerns about riding during nighttime. We wonder if they take alternate modes of transport during those hours and if we can run a study that discovers what those are.

https://www.citibikenyc.com/50million

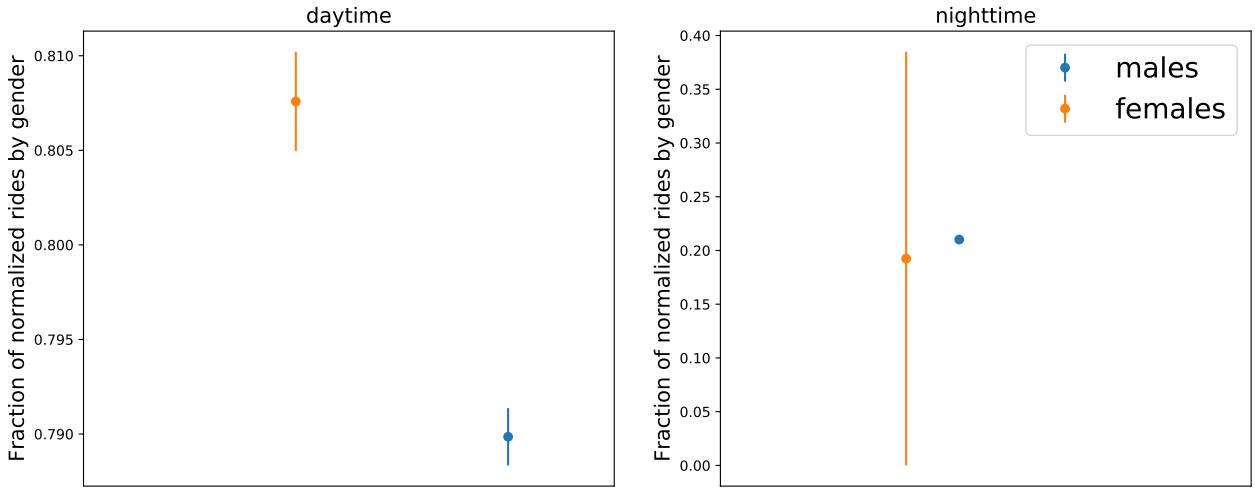

Figure 2: Fraction of normalized rides by gender from Jan to Aug in 2017 in terms of daytime (left) and nightime(right). It shows that the fraction of female nighttime riders has a very wide error bar so that it is necessary to normalize the value by using Z score.

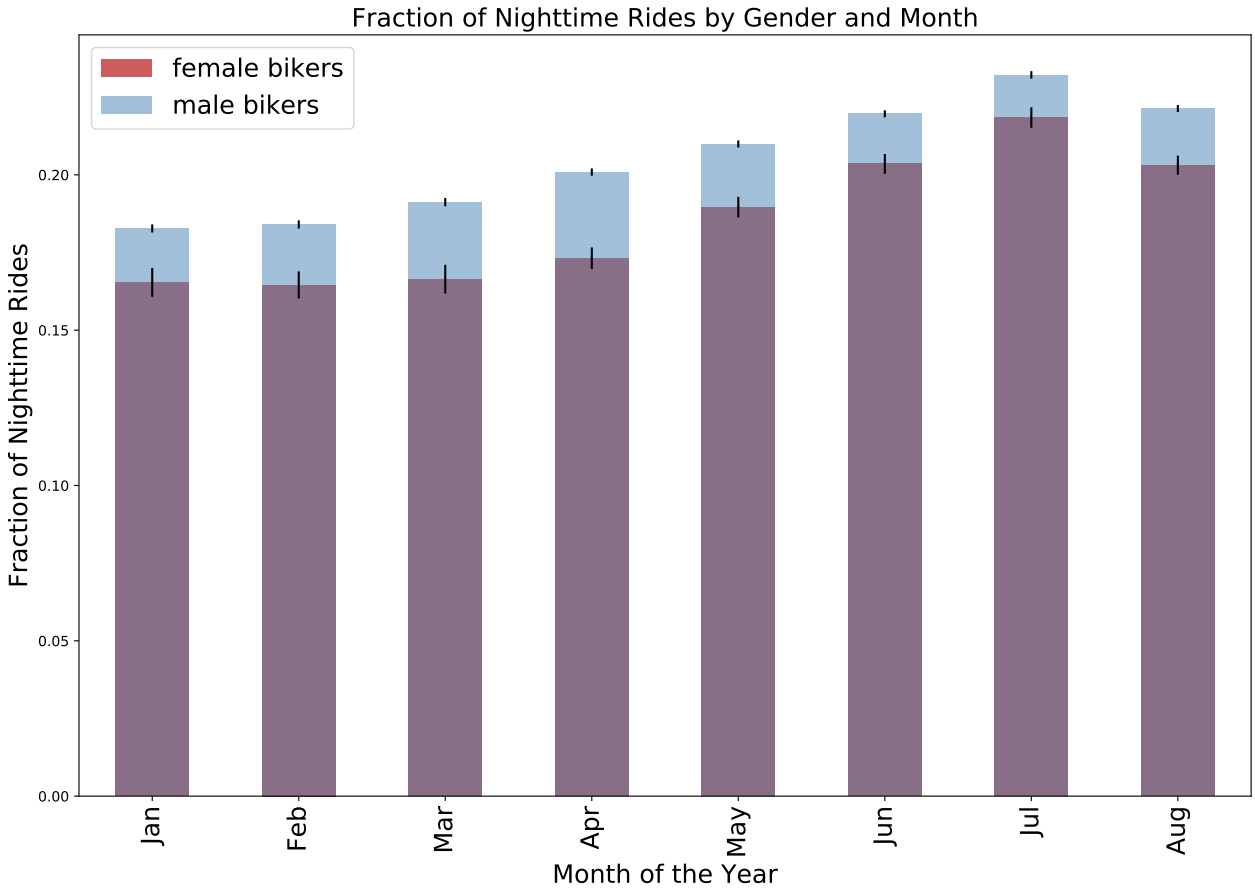

Figure 3: Fraction of Nighttime Riders by Gender in each month. According to the chart, it seems that male are more likely to ride at nighttime compared to female during any of the months from Jan to August in 2017. We will run a series of z tests to measure the corresponding significance.

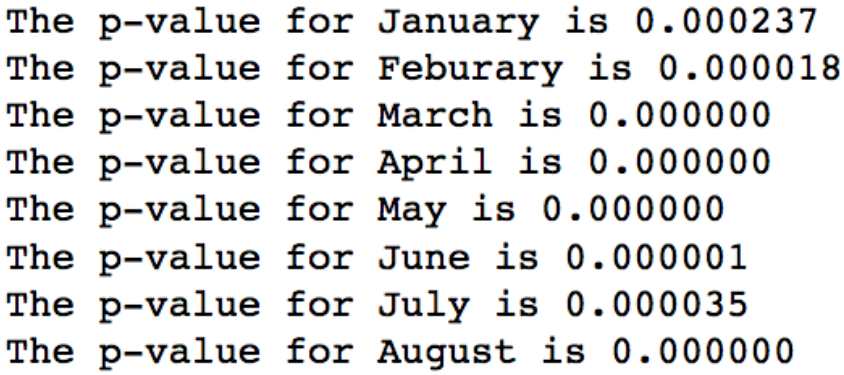

Figure 4: These are the p values based on the Z test we generated for each month.