Exploratory Analysis: Are 'Customer' More Likely to Use CitiBike during Working Hours than 'Subscriber'?

Lingyi Zhang[1]

[1]NYU Center for Urban Science & Progress

November 8, 2017

Abstract

CitiBike is a privately owned public bicycle sharing system. In this study, we analyzed whether the 'Customer' are more likely than 'Subscriber' to use Citibike during working hours (9:00 - 17:00). Using one-tailed Z-test and Chisq test, we found that the percentage of riders used CitiBike during working hours is significantly higher for 'Customer' than for 'Subscriber'.

1. Introduction

CitiBike is a privately owned public bicycle sharing system serving New York City and Jersey City, New Jersey.(wik) It is the nation's largest bike share program, with 10,000 bikes and 600 stations. There are two user types, 'Customer'(mainly visitors and tourists) and 'Subscriber' (mainly New York locals).(nyc)The question we want to answer in this article is that whether the 'Customer' are more likely than 'Subscriber' to use CitiBike during working hours (9:00 - 17:00). To know that can be helpful for CitiBike owner company (NYC Bike Share, LLC) to make their sales and operation strategies. For example, If we want to provide some targeted services or advertisement to the 'Customer'(mainly visitors and tourists), this analysis can give the best time span of providing.

2. Data

CitiBike Tripdata we used in this analysis is the dataset of January 2016. We converted the datatype of "starttime" column into datetime, which can be easily processed in pandas. We then aggregate the trip counts by hours, calculating the respective percentage of trips started during working hours (9:00 - 17:00) and resting hours (18:00 - 8:00) by "Subscriber" and "Customer" (Table 1).

	working hours	resting hours
Subscriber	0.55	0.45
Customer	0.79	0.21

Table 1: The percentage of trips using Citibike during working hours and resting hours for 'subscriber' and 'customer'.

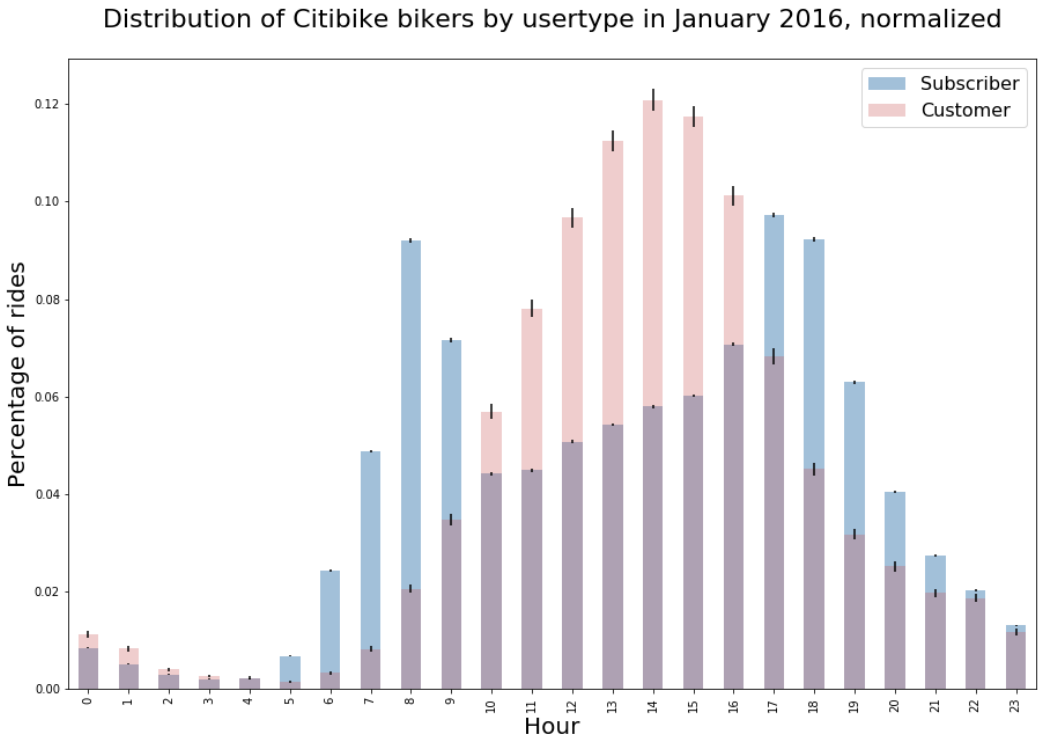

Figure 1: This is a caption

3. Methodology

In both tests below, we assign α=0.05. Here we incorporated Federica's suggestion using Chisq test for proportion because it is appropriate for testing hypotheses about proportions.

We also did one-tailed Z-test. Because when the sample size is large enough (defined as both np and n(1-p) are greater than or equal to 5), the binomial distribution comes to resemble the normal distribution. (Boslaugh, 2012)

In this analysis, we finally didn't use the t-test suggested by Ian, because the sample size is large enough. We used Z-test instead.

3.1 Chisq test

Null hypothesis: The percentage of trips using Citibike during working hours is the same for 'subscriber' as for 'customer'. (Using Citibike during working hours and the user type are independent.)

$$H_0 : \frac{Cust_{\text{WorkingTime}}}{Cust_{\text{All}}} = \frac{Subs_{\text{WorkingTime}}}{Subs_{\text{All}}}$$

$$H_a : \frac{Cust_{\text{WorkingTime}}}{Cust_{\text{All}}} \neq \frac{Subs_{\text{WorkingTime}}}{Subs_{\text{All}}}$$

$\text{Cust}_{\text{WorkingTime}}$: The counts of 'Customer' using CitiBike during working hours.

$\text{Subs}_{\text{WorkingTime}}$: The counts of 'Subscriber' using CitiBike during working hours.

Cust_{All}: The counts of 'Customer' using CitiBike during the whole day.

Subs_{All}: The counts of 'Subscriber' using CitiBike during the whole day.

	working hours	resting hours	summary
subscriber	0.55*484935	0.45*484935	484935
customer	0.79*24543	0.21*24543	24543
total	286782	222696	509478

Table 2: Contingency table for the Chisq test.

Using scipy.stats.chi2_contingency for Chisq test, the chisq test statistic is 5198.869 with $Pvalue = 0.000$. We can reject the Null hypothesis. The percentage of trips using Citibike during working hours for 'subscriber' and 'customer' are not the same.

3.2 one-tailed Z test

Null hypothesis: The percentage of trips using CitiBike during working hours is the same or lower for 'Customer' than for 'Subscriber'.

$$H_0 : \frac{Cust_{\text{WorkingTime}}}{Cust_{\text{All}}} <= \frac{Subs_{\text{WorkingTime}}}{Subs_{\text{All}}}$$

$$H_a : \frac{Cust_{\text{WorkingTime}}}{Cust_{\text{All}}} > \frac{Subs_{\text{WorkingTime}}}{Subs_{\text{All}}}$$

Using statsmodels.stats.proportion.proportions_ztest for Z-test, the Z-test statistic is 72.386 with $Pvalue = 0.000$. We can reject the Null hypothesis. The percentage of trips using CitiBike during working hours is higher for 'Customer' than for 'Subscriber'. [Source code]

4. Conclusions

'Customer' are more likely than 'Subscriber' to use CitiBike during working hours. It might be because most of the 'Customer' are tourists, while most of the 'Subscriber' are local labor force. During working hours, a higher proportion of 'Subscriber' are staying in office, while 'Customer' are touring in the city.

The implication is that if we want to do advertisement targeting the 'Customer', the best time to display the advertisement would be the working hours because a higher proportion of 'Customer' can see the advertisement, while a lower portion of 'Subscriber' will be disturbed by useless information as for them. Regarding the types of advertisement, promotion for annual membership would not be a good choice, because most of the 'Customer' are temporally visiting. Instead, promotions about short time discount might be feasible, such as the one-day or three-day pass.

The weakness of this analysis is that we did not consider the seasonality. The conclusion we drew based on data of January might not able to apply to Summer.

References

Citi Bike: NYC's Most Popular Bike Rental Program — Citi Bike NYC. https://www.citibikenyc.com/how-it-works. URL http://www.citibikenyc.com/how-it-works. Accessed on Mon, November 06, 2017.

Citi Bike - Wikipedia. https://en.wikipedia.org/wiki/Citi$_B ike.URL$. Accessed on Mon, November 06, 2017.

Sarah Boslaugh. *Statistics in a nutshell: A desktop quick reference*. O'Reilly Media, Inc., 2012.