Citibike - analysis of subscriber and customer trip duration over 45 minutes

JKtours[1] and adn323[2]

[1]new york university
[2]Affiliation not available

November 9, 2017

Abstract

We wanted to investigate relationship between rides over 45 minutes for customers and subscribers, particularly considering the pricing models for citibike. We found that on average, the proportion of customers were more likely to ride for over 45 minutes than subscribers. Further to this, over 99% of rides by subscribers were less than 45 minutes (and for further analysis we might test this formally).

Introduction

Citibike provides a public bike share services with 10,000 bikes, 600 docking stations across 55 neighborhoods (in Manhattan, Queens, Brooklyn, and New Jersey). It is a public private partnership between NYC and Citi (the financial institution). It is a very common method for transportation used by thousands of New Yorkers and tourists daily to commute across the city. Similar bikesharing arrangements (and companies) operate in other cities. For the NYC service (Citibike) there are two approaches to payment, via day passes (known as Customers), or annual/monthly subscription (known as Subscribers). For subscribers, annual membership includes unlimited rides under 45 minutes with an additional charge of $2.50 per 15 minutes for rides over 45 minutes.

Data

We used the publicly available citibike data for the month of October 2016. The data provides information on all citibike trips with fields:

- 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, and
- Gender

We used the trip duration field to create an indicator variable of trips over 45 minutes (Over 45) and under 45 minutes (Under 45). Our finalized dataset summarized the trips neatly with three fields – the index, User Type and the Over 45 indicator (for trip over 45 minutes).

In order to test the hypothesis we used two different test of difference in proportions with a significance level $\alpha = 0.05$. The hypothesis is this:

	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
0	328	2016-10-01 00:00:07	2016-10-01 00:05:35	471	Grand St & Havemeyer St	40.712868	-73.956981	3077	Stagg St & Union Ave	40.708771	-73.950953	2525
1	398	2016-10-01 00:00:11	2016-10-01 00:06:49	3147	E 85 St & 3 Ave	40.778012	-73.954071	3140	1 Ave & E 78 St	40.771404	-73.953517	1781
2	430	2016-10-01 00:00:14	2016-10-01 00:07:25	345	W 13 St & 6 Ave	40.736494	-73.997044	470	W 20 St & 8 Ave	40.743453	-74.000040	2094

Figure 1: Header rows of extract from raw citibike data used for analysis

	User Type	Over45
0	Subscriber	Under 45
1	Subscriber	Under 45
2	Subscriber	Under 45
3	Subscriber	Under 45
4	Subscriber	Under 45
5	Subscriber	Under 45
6	Subscriber	Under 45
7	Subscriber	Under 45
8	Customer	Under 45
9	Subscriber	Under 45

Figure 2: Data manipulation to assemble relevant columns (and drop others) for analysis)

The first test we applied was the z-test for difference in proportions as we know the parameters for the whole population.

We also applied a Chi Squared test of proportions in groups of categorical variables to assess if there are any significance difference in categories of subscribers and customers who ride for over and under 45 minutes.

Conclusions

The tests both highlight that there is a significant difference between the proportion of customers and subscribers riding over 45 minutes (p-value = 1.652e-68 for Z-test and p-value very small (less than 0.01)

Over45	Over 45	Under 45
User Type		
Customer	7.968463	92.031537
Subscriber	0.776854	99.223146

Figure 3: Summary of data showing frequency percentage of User type over and under 45 minutes

$$H_0 : \frac{C_{\text{Over45}}}{C_{\text{Total}}} <= \frac{S_{\text{Over45}}}{S_{\text{Total}}}$$

$$H_1 : \frac{C_{\text{Over45}}}{C_{\text{total}}} > \frac{S_{\text{Over45}}}{S_{\text{Total}}}$$

for the Chi Squared test with a significance chosen of α=0.05). Therefore we reject the null hypothesis (for both tests)

One of the weaknesses with the analysis is that we have only looked at one month of data, while we may potentially have different results depending on the time of year.

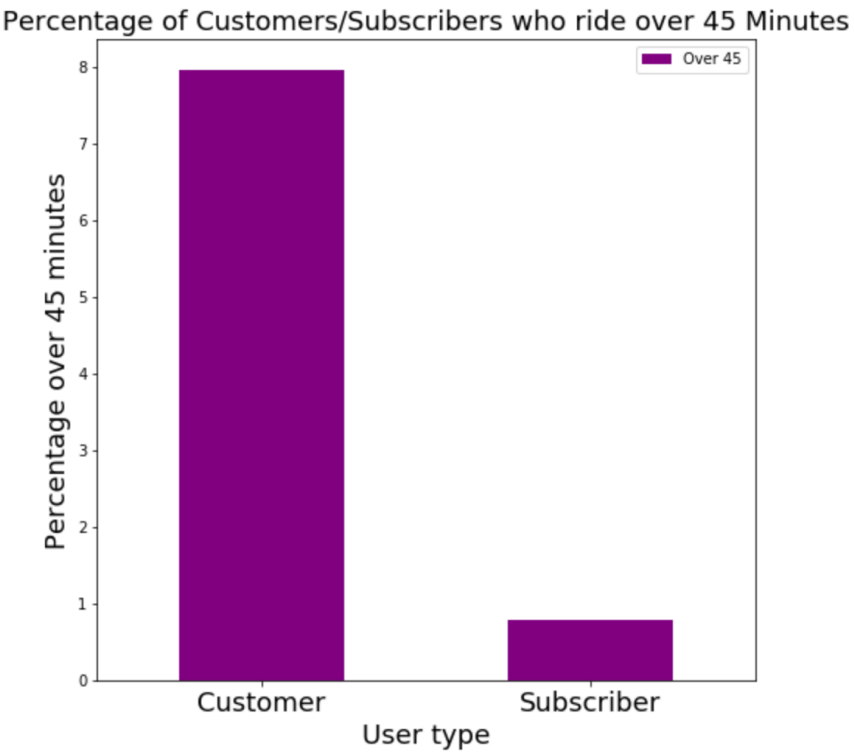

Figure 4: Percentage of Subscribers and Customers with rides over 45 minutes