Analysis on Citi Bike trip durations in weekdays

YukunWan[1]

[1]New York University (NYU)

November 7, 2017

Abstract

Citi Bike has become a major choice of NYC public transportation. This research focuses on the distribution of trip durations in weekdays and see whether most of Citi Bike trips are lower than 15 minutes, which is defined as short trips. This result helps to verify whether the Citi Bike is the tool of the "last mile" commuter. The research shows that the fraction of long(>15min) trips over total number of trips on weekdays is less than the fraction of short(<15min) trips over total number of trips on weekdays.

Introduction

Citi Bike is the nation's largest bike share program, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens and Jersey City. It was designed for quick trips with convenience in mind, and it's a fun and affordable way to get around town. As increasingly people choose Citi Bike as their major vehicles, it is worthy to investigate the riding behaviors. If Citi Bike is usually used to complete the "last mile" of public transportation, we can assume that the trip durations are often under an estimated number. Meanwhile, some people may regard riding Citi Bike as a way of exercise, which may leads to longer trip durations. It is interesting to find out the real pattern of the trip durations on weekdays and weekends. The result can help to optimize the resource planning of Citi Bike and maximum the advantage of it. This research tries to answer the question of whether there are more long trips of Citi Bike then short trips of it on weekdays.

Data

The data used in this research is retrieved from Citi Bike's open data. The analysis is based on all the trips in December 2016. The research try to figure out the riding behavior on weekdays. Thus I transfer the column "start time", which is in string format, to column "date" with function "to_datetime". After that, it is easy to find out whether the day is weekday or weekend with function "dt.weekday". Meanwhile, since only trip durations and weekdays are required for this research, other features are all dropped.

There are some outliers in this dataset. Fig.1 shows the original data plot and Fig.2 shows the data after removing outliers with 3-sigma rules.

Methodology

Chi-square test is chosen for this research. It is because the research works on two unpaired group - long trips and short trips. At the same time, although the trip durations are in numeric types, I put them into two

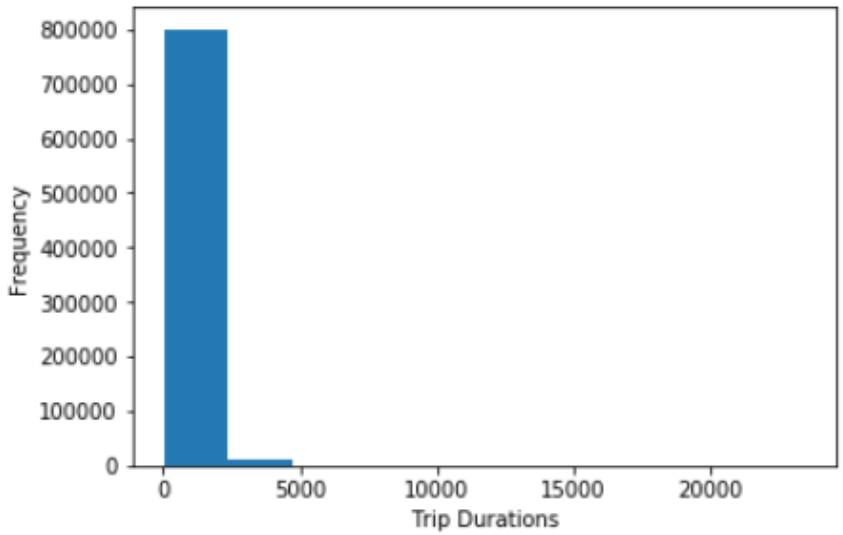

Figure 1: Frequency of Trip durations in December 2016 before removing outliers.

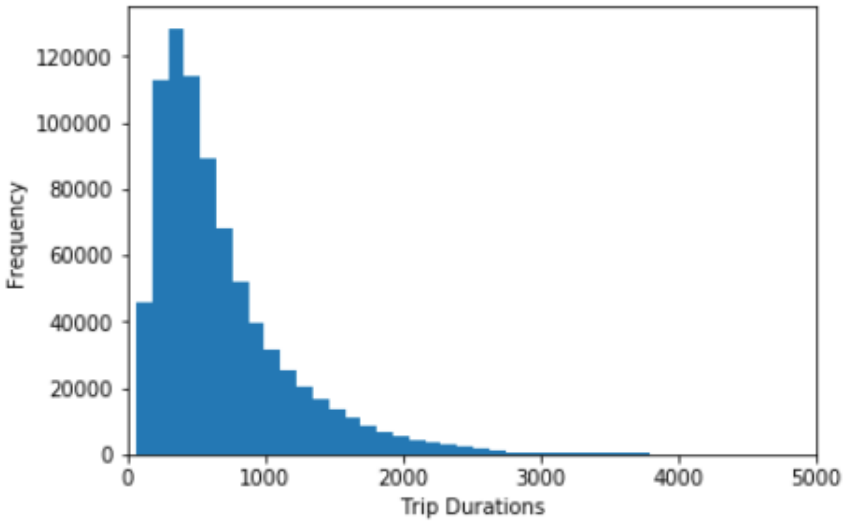

Figure 2: Frequency of Trip durations in December 2016 after removing outliers with 3-sigma rules.

categories - long trips (>15 minutes) and short trips (<= 15 minutes). This clustering makes the parameters categorical data.

My peer Emily suggests an alternative option - ANOVA test. However, according to the flow chart in article "How to choose the right statistical test", ANOVA is more suitable for parametric issue.

Thus, I conduct one-tail chi-square test under the null hypothesis that the fraction of long(>15min) trips over total number of trips on weekdays is more or the same as the fraction of short(<15min) trips over total number of trips on weekdays.

Conclusions

Both Fig.3 and Fig.4 show the distribution of Citi Bike trips on riding durations. Obviously, the number of short trips is about triple the number of long trips when we separate the trips by the day of week. Chi-square test also gives the same result that the null hypothesis is rejected on significance level 0.05. The chi-square statistics is 714.68, which is far higher than the threshold of 3.84 (one-tail, alpha = 0.05, degree of freedom = 1).

The result matches the assumption of the "last mile" theory that most of people regard Citi Bike as a commuter for short distance/time. However, there are some weaknesses in the analysis. The data I used is the trips in December 2016. It's very likely that people chose not to ride for too long time in cold winter. The better way is to choose the data in each season in order to ensure that there is no bias in data selection.

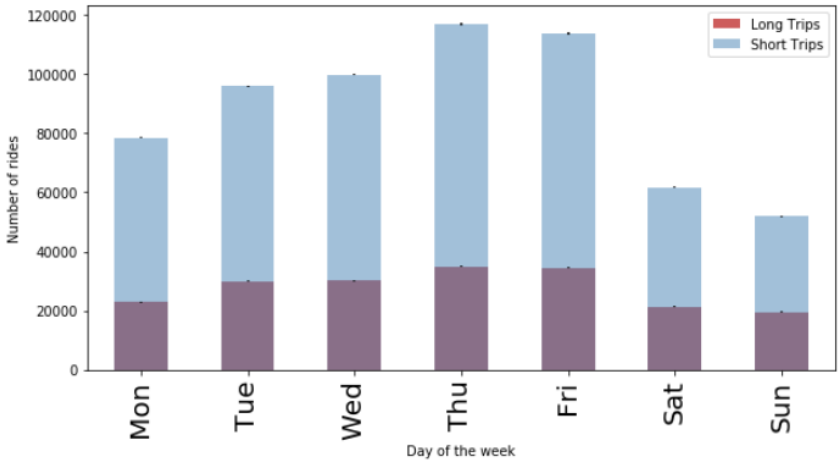

Figure 3: Distribution of Citi Bike trips by ride durations in December 2016, absolute counts

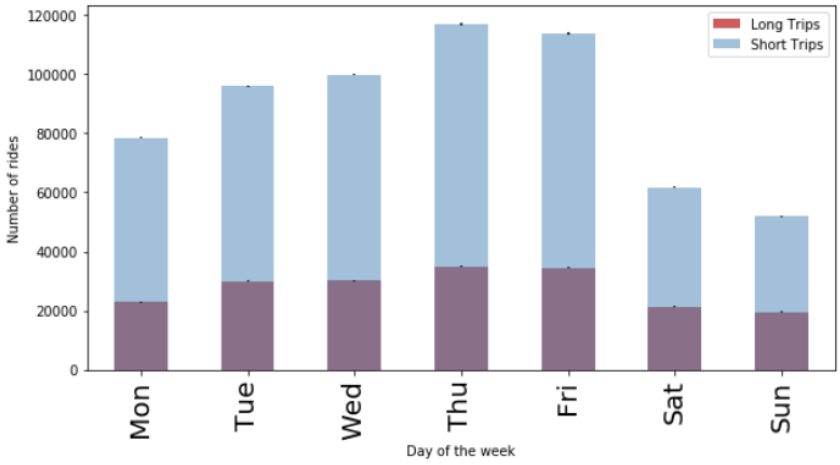

Figure 4: Distribution of Citi Bike trips by ride durations in December 2016, absolute counts, with statistical errors