PUI Homework 7

Isha Chaturvedi[1]

[1]New York University (NYU)

November 9, 2017

Abstract

This article investigates into the bike usage data of Citi Bike, which is a NYC based - bike sharing company. The article studies whether the relative customer usage of bikes on the weekend is higher than that for the subscribers. The data of two months is used to avoid the skewness in the result. The test of robustness is performed by considering Friday as a weekend. A Chi-Squared test of proportions is used to compare the usage frequencies.

Introduction

CitiBike is a privately owned public bike sharing system that operates in New York City and Jersey City. The service can either be used via a 24 hour user pass or 3 day user pass or via an annual subscription service. The first two types of users are classed as customers whereas the users with annual subscription service are classed as subscribers. The research question is based on the idea that the customers tend to use Citi Bike more for leisure than commuting as compared to the subscribers, thus leading to high peak of bike usage by customers during the weekends.

Data

The Citi Bike data is given monthly. The data can be found at https://s3.amazonaws.com/tripdata/%Y%m-citibike-tripdata.zip, replacing the date format codes for the desired month and year. There is a possibility that the month of data used has large number of holidays and hence the result would be skewed if only one month data is used. To avoid skewness in the result, two months data - March and April 2015 is used to conduct this research. The data set consists of 15 columns, and has information such as tripduration, starttime, stoptime, usertype, gender etc. Each row in the dataset represents a single trip taken by a user. The relevant columns in the dataset are the date that the ride was taken and the user type, either customer or subscriber. The normalized distributions of Citi Bike users for each day of the week in March and April, 2015 by usertype is shown in Fig. 1 below.

The Fig. 1 shows the Citi Bike usage pattern of each user type for each day of the week. It can be seen that the customer usage shoots up over the weekend and is higher than the subscriber usage.

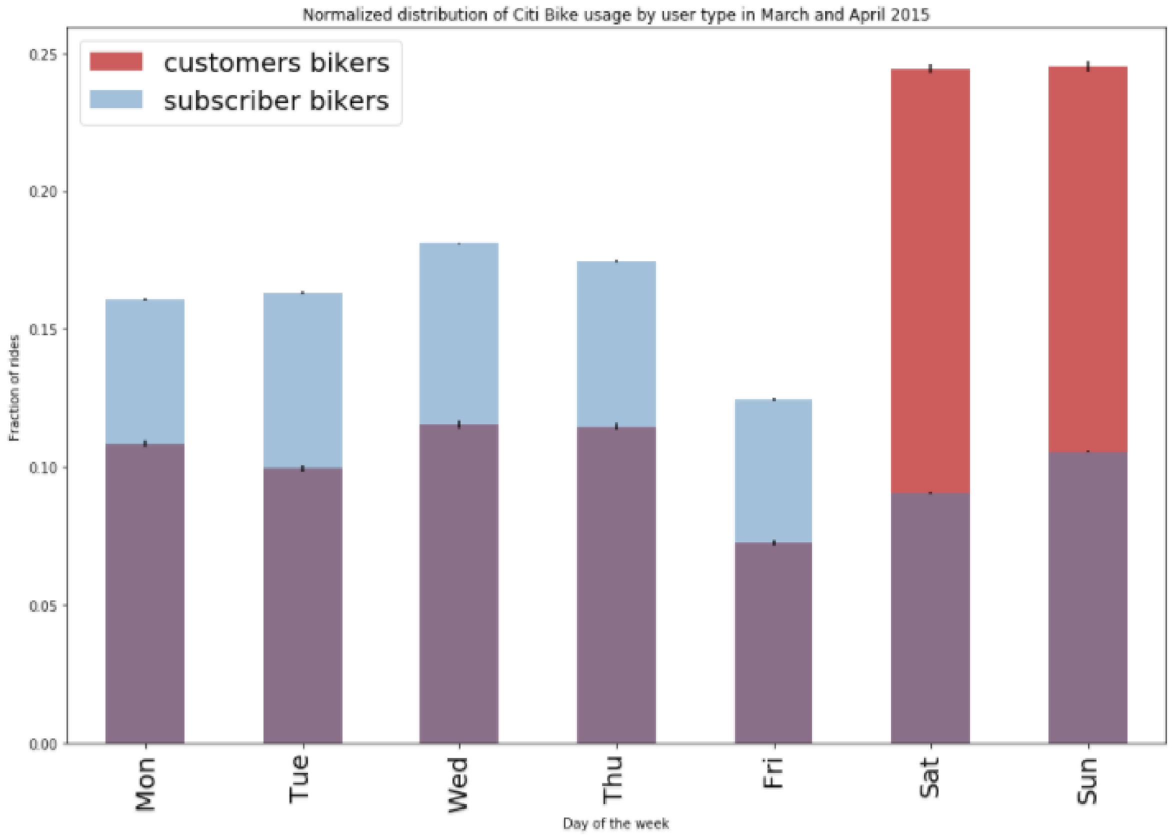


Figure 1: Normalized distribution of Citi Bike usage by user type in March and April 2015

Methodology

The overall counts for customers and subscribers were taken for each day of the week and were subsequently aggregated into counts for the week and the weekend. The statistical errors equal to the square root of the counts of customers or subscribers were added to the counts and counts were then normalized by their overall count (Fig. 1). The counts were normalized as the overall magnitude of subscriber usage were higher than the customer usage. The null and alternative hypotheses are as follows:

$H_0 : \frac{P_{customer,weekend}}{P_{customer,total}} <= \frac{P_{subscriber,weekend}}{P_{subscriber,total}}$

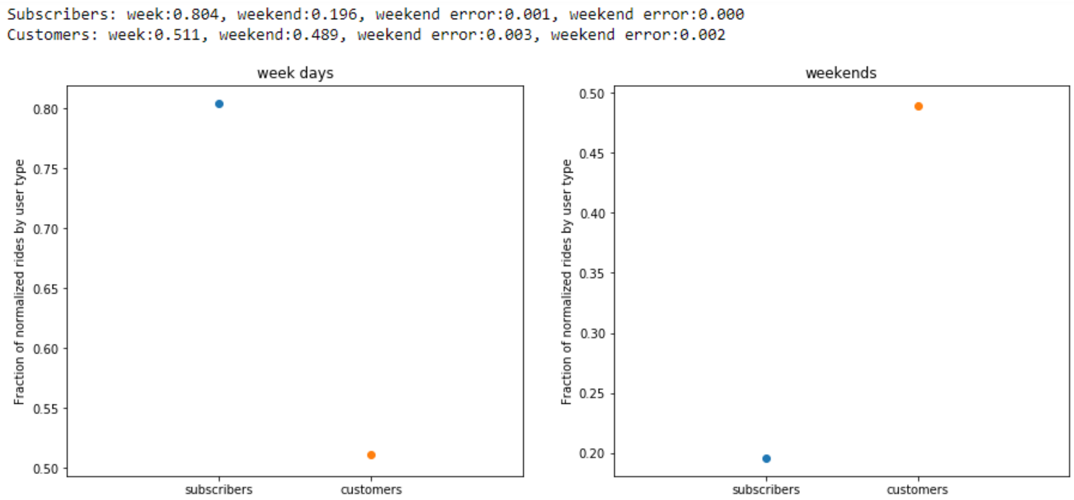
$H_A : \frac{P_{customer,weekend}}{P_{customer,total}} > \frac{P_{subscriber,weekend}}{P_{subsciber,total}}$

where P is the count for a user type using the bike either during the week or weekend.

The above hypothesis was formulated to include the total number of customers or subscribers using the bike both on the week and the weekend to have proper normalization. Chi-Squared test of proportions was chosen to test the significance of the difference between the frequencies of two categories. A significance level of $\alpha = 0.05$ was used. A test for robustness was performed by associating fridays with weekends, instead of weeks.

Conclusions

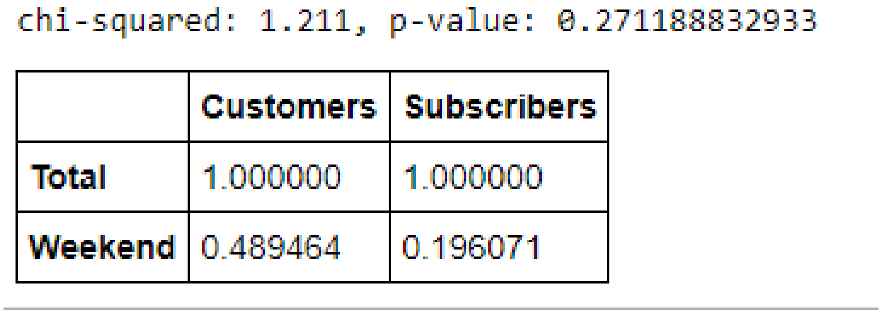
The fraction of Citibike bikers per user type in March and April 2015 for weekdays and weekend along with the errors is shown in Fig. 2.



```
Subscribers: week:0.804, weekend:0.196, weekend error:0.001, weekend error:0.000
Customers: week:0.511, weekend:0.489, weekend error:0.003, weekend error:0.002
```

Figure 2: Fraction of Citibike users per user type in March and April 2015 for week days (left) and weekends (right)

This plot shows the comparison of normalized counts for user type over the week and the weekend respectively. The plot shows that the distribution of customers weighs more than that of the subscribers towards the weekend. A chi-squared test of proportions was performed to test the significance of this difference. The results of the test is shown in the FIg. 3 below.



```
chi-squared: 1.211, p-value: 0.271188832933
```

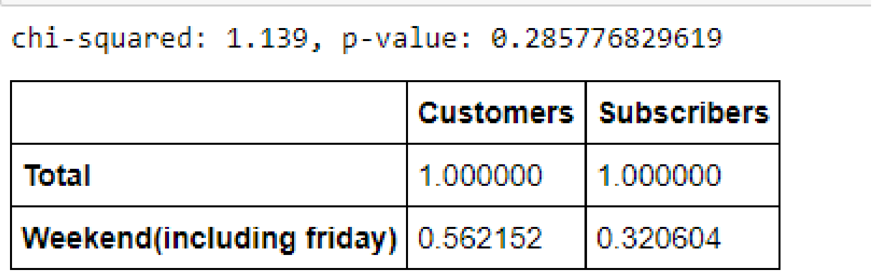
	Customers	Subscribers
Total	1.000000	1.000000
Weekend	0.489464	0.196071

Figure 3: Results of Chi-Squared Test and the Contingency Table

As seen from the Fig.3 the p value (0.27) is greater than 0.05 (significance level of $\alpha = 0.05$), and chi-squared statistic of 1.211 is lower than the chi-squared critical value for a significance level of 0.05, $\chi^2_{critical} = 3.84$. This means that the null hypothesis cannot be rejected.

The results of test of robustness is shown in the Fig. 4 below.

The frequencies of customers increased after including Fridays as weekend from 0.49 to 0.56 and chi-squared statistic improved from 1.211 to 1.139 (from Fig. 3 and Fig. 4). But as seen from the Fig.4 the p value



```
chi-squared: 1.139, p-value: 0.285776829619
```

	Customers	Subscribers
Total	1.000000	1.000000
Weekend(including friday)	0.562152	0.320604

Figure 4: Results of Chi-Squared Test and the Contingency Table (After including Fridays as weekend)

(0.29) is greater than 0.05 (significance level of $\alpha = 0.05$), and chi-squared statistic of 1.139 is still lower than the chi-squared critical value for a significance level of 0.05, $\chi^2_{critical} = 3.84$. This means that the null hypothesis cannot be rejected. The test could be improved by including data from more number of months, specifically from different times of the year to account for bias from the seasonality.