HW7_Assignment1_hh1827

Hongkai He[1]

[1]NYU Center for Urban Science & Progress

November 7, 2017

Time Series Analysis for the Different Citibike Usage Pattern Between Genders in New York

<Hongkai He, hh1827>

Abstract

this project aims to examine the Citibike ridership of different genders during different time periods (daytime vs night). The New York Citibike data set of March, 2017 is selected and processed to filter redundant information, and the Statistics Test, Z test, is used to verify whether to accept the null hypothesis that men are less likely to ride Citibikes than women during the night.

Introduction

Citibike is the largest bike sharing system in the U.S. which is located in New York City and funded by Citibank (nyc). A growing fleet of specially designed bikes are available anytime for residents in Manhattan, Brooklyn, Queens and Jersey City. People can unlock a bike from a docking station and return it to any stations at the end of the trip. Citibike has become a very popular choice for Commuting, city sightseeing, and any other short-to-middle distance trips in New York.

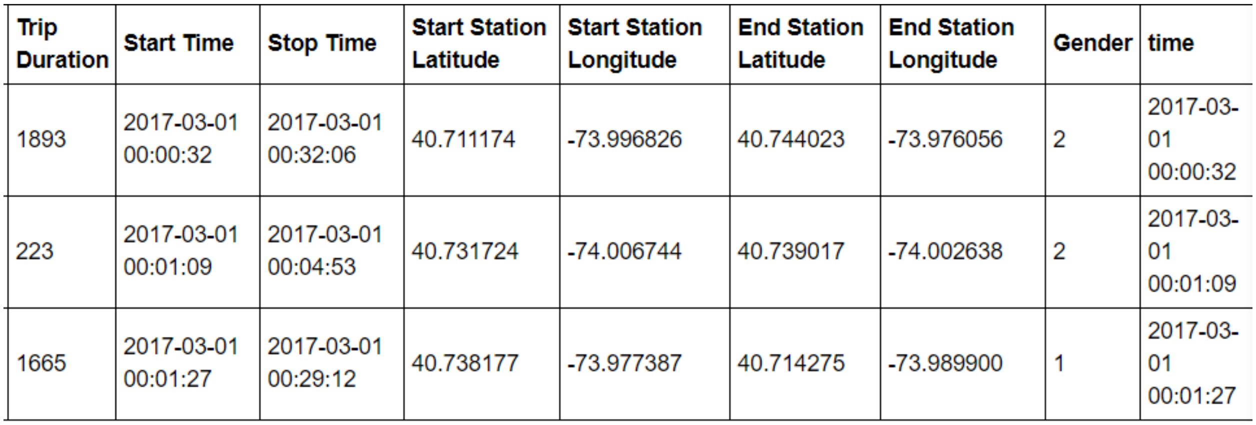
Safety issue is the primary concern for any forms of transportation. Some of the natures of Citibike, such as availability at night, slow speed, and easy blending into urban fabrics, etc. make the riders of Citibike more susceptible to street crime than other forms of transportation. In addition, it is widely accepted that street crimes pose large threats to women than men. This project focuses on exploring whether this conventional idea influences the Citibike usage between different genders, especially at night when people have higher chance to encounter crime.

Data

The data set used for this project is the ridership records of Citibike all over New York City. They are provided on the CUSP Data Facility Platform, and can also be obtained from the following website: https://s3.amazonaws.com/tripdata/. The data set is available as monthly subsets starting from July, 2013 to July, 2017. the Monthly data of March, 2017 is chosen for the project because the bike fleet and the number of membership of Citibke have been growing in the last several years so it is reasonable to deduce that the latest year has the largest data set and can be more representative. However, there are much redundant information that are irrelevant to our project, such as station names, station locations and user types, etc. Hence irrelevant columns are dropped and only gender and trip time are retained.

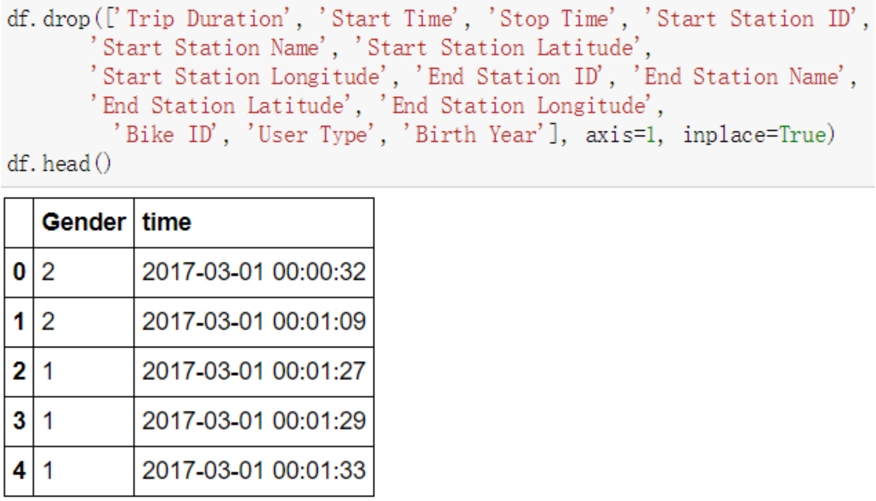
Methodology

The data set is first visualized to provide us a quick view on whether there exist expected different usage pattern of different genders during the night.



Trip Duration	Start Time	Stop Time	Start Station Latitude	Start Station Longitude	End Station Latitude	End Station Longitude	Gender	time
1893	2017-03-01 00:00:32	2017-03-01 00:32:06	40.711174	-73.996826	40.744023	-73.976056	2	2017-03-01 00:00:32
223	2017-03-01 00:01:09	2017-03-01 00:04:53	40.731724	-74.006744	40.739017	-74.002638	2	2017-03-01 00:01:09
1665	2017-03-01 00:01:27	2017-03-01 00:29:12	40.738177	-73.977387	40.714275	-73.989900	1	2017-03-01 00:01:27

Figure 1: Figure 1: Raw data set



```
df.drop(['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'], axis=1, inplace=True)
df.head()
```

	Gender	time
0	2	2017-03-01 00:00:32
1	2	2017-03-01 00:01:09
2	1	2017-03-01 00:01:27
3	1	2017-03-01 00:01:29
4	1	2017-03-01 00:01:33

Figure 2: Figure 2: Processed data set that contains only users' gender and trip start time.

A pair of Null and Alternative Hypothesis is formulated as follows for the statistical test:

H_0 : The ratio of man biking at night over man biking during the day is *the same* or *less* than the ratio of woman biking at night to woman biking during the day. $\frac{Wnight}{Wday} \geq \frac{Mnight}{Mday}$

H_1: The ratio of man biking at night over man biking during the day is higher than the ratio of woman biking at night to woman biking during the day. $\frac{Wnight}{Wday} < \frac{Mnight}{Mday}$

Since the number of observations in the sample is much larger than 30 and the null hypothesis is formulated in the form of proportion, the Z-test for two proportions is used to test the hypothesis. The mathematical formula for computing the statistics are listed as follows:

$Z\ score:\ Z = \frac{p0\ -\ p1}{SE}$

$Where:\ p0 = \frac{Wnight}{Wday},\ p1 = \frac{Mnight}{Mday},\ SE = \sqrt{p\,(1-p)\left(\frac{1}{n0} - \frac{1}{n1}\right)}\ ,\ p = \frac{p0n0\ +\ p1n1}{n0\ +\ n1} Z\ score:\ Z = \frac{p0\ -\ p1}{SE}$

Alternatively, the Chi Square goodness of fit test can also be used for this project.

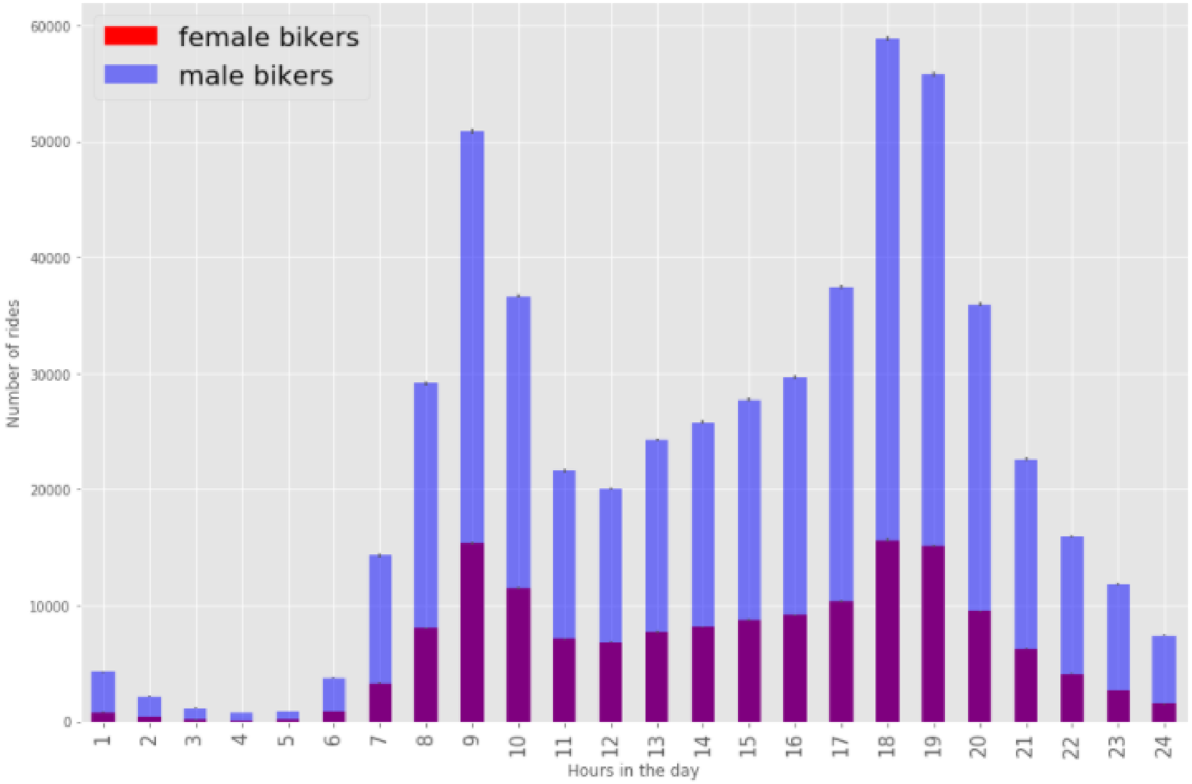


Figure 3: Figure 3: Hourly Distribution of Citibike bikers by gender in March 2017, absolute counts, with statistical errors

Conclusions

The final result of Z score is calculated as 138.93, which is much larger than the largest the largest number reported in Z stats tables, 3. In the Z stats table, Z=3 corresponds to a p value < 0.0002, hence the p value corresponding to the calculated result must be much smaller than the chosen significant level alpha = 0.05. Therefore, We can reject the null hypothesis that The ratio of man biking at night over man biking during the whole day is the same or less than the ratio of woman biking at night to woman biking during the day. We can conclude that women are less likely than men to choose biking during the night (9:00pm - 6:00am).

The weakness of this project is that the size of the data set chosen for the analysis is small, covering only one month in 2017. This is due to some objective constraints such as the limitation of computing capability of hardware used.

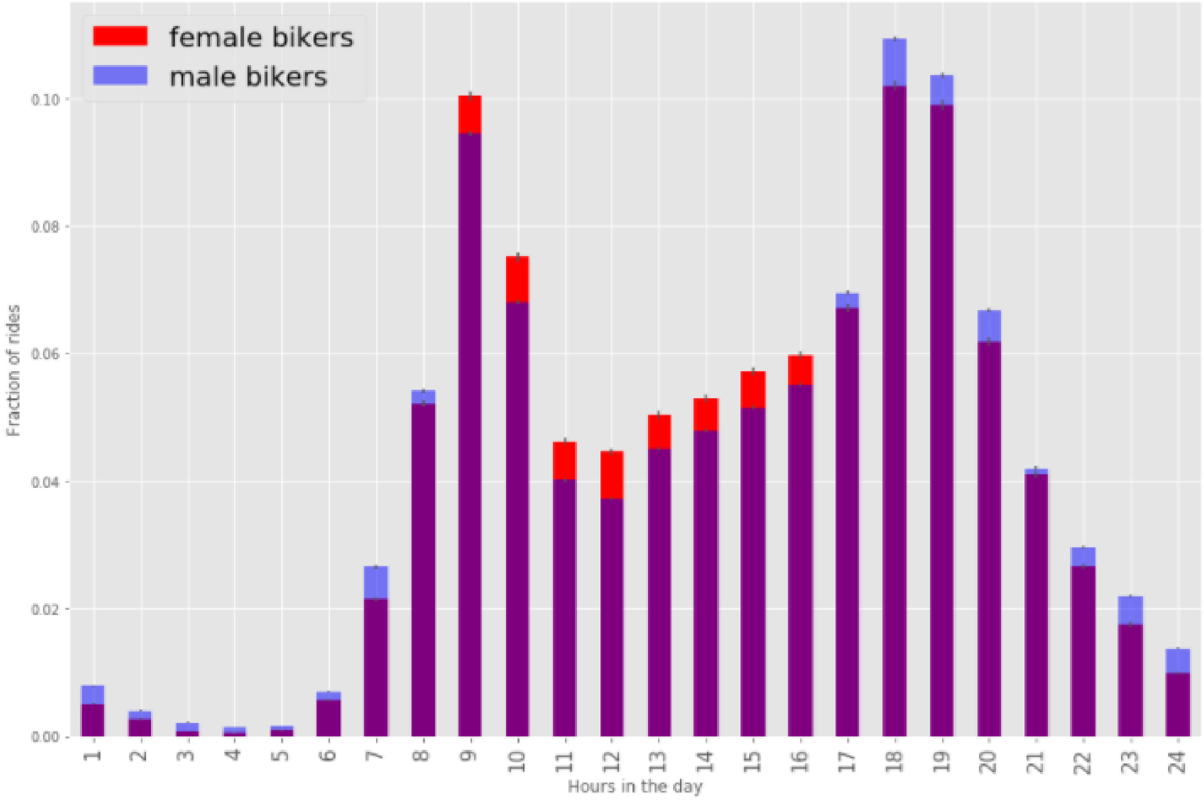
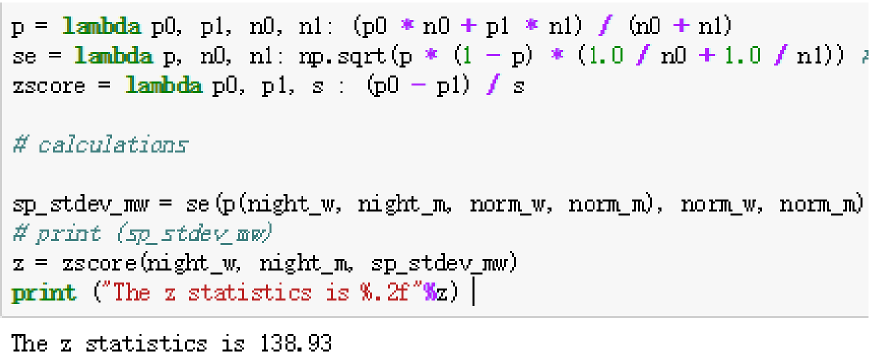


Figure 4: Figure 4: Hourly Distribution of Citibike bikers by gender in March 2017, normalized



```
p = lambda p0, p1, n0, n1: (p0 * n0 + p1 * n1) / (n0 + n1)
se = lambda p, n0, n1: np.sqrt(p * (1 - p) * (1.0 / n0 + 1.0 / n1)) #
zscore = lambda p0, p1, s : (p0 - p1) / s

# calculations

sp_stdev_mw = se(p(night_w, night_m, norm_w, norm_m), norm_w, norm_m)
# print (sp_stdev_mw)
z = zscore(night_w, night_m, sp_stdev_mw)
print ("The z statistics is %.2f"%z)
```

```
The z statistics is 138.93
```

Figure 5: Figure 5: Code for calculating the statistics required by Z test for two proportions and the Z score result

References

About Citi Bike: Company, History, Motivate — Citi Bike NYC. https://www.citibikenyc.com/about. URL http://www.citibikenyc.com/about. Accessed on Tue, November 07, 2017.