Data analysis of the impact of age on riding Citibike

Rufei sheng[1]

[1]New York University

November 7, 2018

Abstract

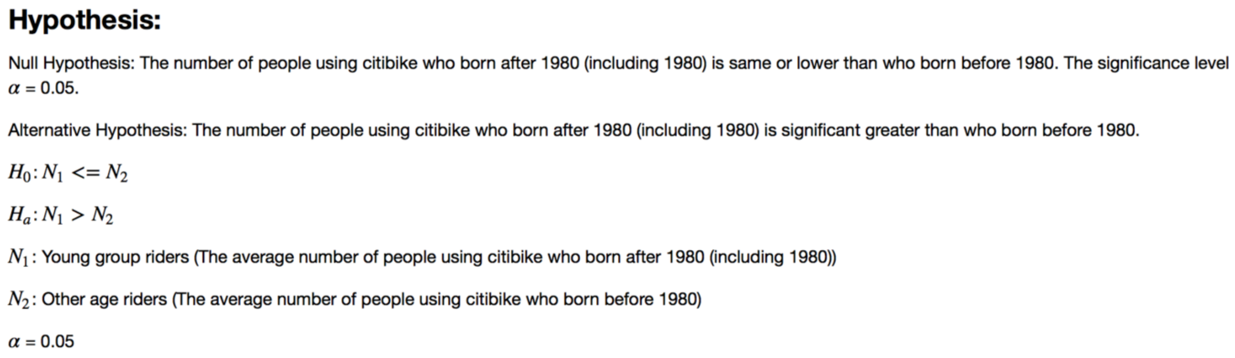
Citi Bike has successfully catch the first mover advantages in applying sharing economy in city's transportation system. This paper aims to examine the idea that young people tend to support and use Citi Bike more than other groups. In order to do so, first the hypotheses are established with users split into two group (young group and other age group). Then, the dataset of trip information is got from Amazon Web Services (AWS) with reproducible approach. After that, some data processing and t-test analyzing methods are applied. In conclusion, there is no statistic significant evidence show that young people ride more than other age groups. For further research, a larger dataset that more than one month can be used to increase the persuasiveness of the test. In addition, the interesting finding of seasonal effect also need further research.

Introduction

Sharing economy originally grew out of the open-source community, then gradually refers to peer-to-peer based sharing of access to goods and services(wik). Citi Bike is one of typical businesses benefited from sharing economy. Launched in May 2013, Citi Bike has become an essential part of transportation network through its bike sharing system, and also become the largest in the nation. The bikes can be unlocked from one station and returned to any other station in the system, making the ideal from one-way trip with efficient, affordable and environmentally friendly(nyc).

Young people are always regarded as the main supporters towards new things. According to Mithun (2012) research, consumers will be encouraged to engage in sharing economy because of rational reasons, such as saving and practicality, and emotional reasons, such as feel altruistic and part of community. Young people are especially driven by emotional motivations because they desire to feel smart and fashionable(Schmitt, 1999). Therefore, it is reasonable to assume that young people tend to support Citi Bike more than other groups. Similarly, Citi Bike also treat young group as their main target consumers. For example, they held the campaign named "Promoting Bike Share Through Youth Engagement" at April 2018, which clearly indicated the target participators are "Young adults aged 16-24 years old who live, work, or go to school in the Citi Bike service area".

Overall, it is interesting to test this assumption that whether young people tend to support more than other groups, which is also useful for Citi Bike in implementing effective promoting campaigns. Fig. 1 shows my null hypothesis and its corresponding math expression with significant level.



Hypothesis:

Null Hypothesis: The number of people using citibike who born after 1980 (including 1980) is same or lower than who born before 1980. The significance level $\alpha = 0.05$.

Alternative Hypothesis: The number of people using citibike who born after 1980 (including 1980) is significant greater than who born before 1980.

$H_0: N_1 <= N_2$

$H_a: N_1 > N_2$

N_1 : Young group riders (The average number of people using citibike who born after 1980 (including 1980))

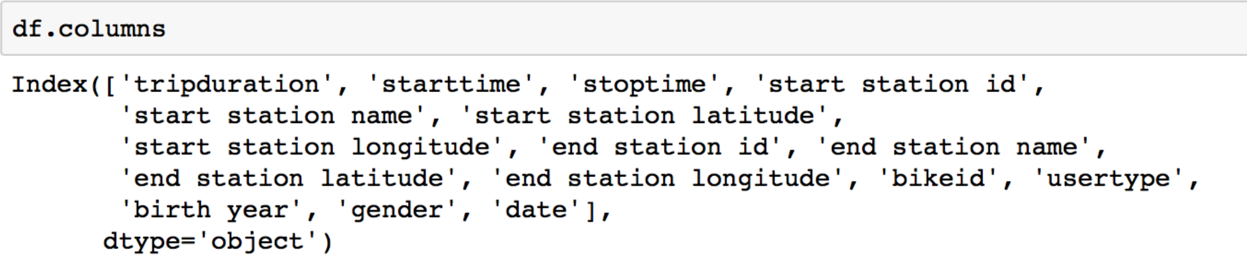
N_2 : Other age riders (The average number of people using citibike who born before 1980)

$\alpha = 0.05$

Figure 1: This Establish the hypotheses with significant level

Data

The dataset of trip information of Citi Bike is got from s3.amazonaws.com, which is an endpoint for a cloud file storage product offered by Amazon Web Services (AWS). In order to increase the reproducibility, the data acquiring process is defined by a function named “GetCitibikeCSV” with the parameter “DataString”. When the needed time period transfers to “DataString”, the corresponding data in that time period is downloaded as CSV format. This paper is based on a month data (July, 2015) to do the test. As Fig. 2 shown, the original dataset is really rich.



```
df.columns
```

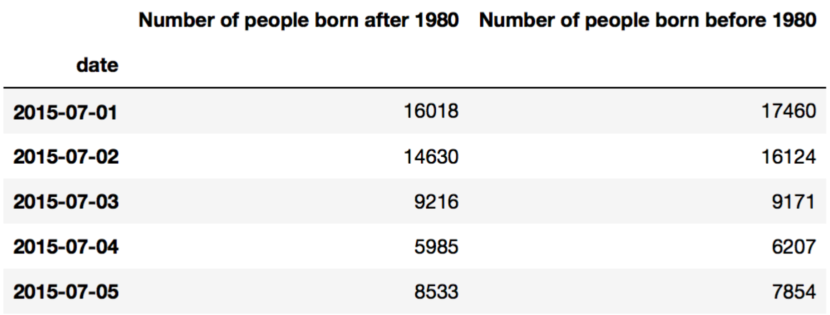
```
Index(['tripduration', 'starttime', 'stoptime', 'start station id',
       'start station name', 'start station latitude',
       'start station longitude', 'end station id', 'end station name',
       'end station latitude', 'end station longitude', 'bikeid', 'usertype',
       'birth year', 'gender', 'date'],
      dtype='object')
```

Figure 2: The columns from original dataframe

In order to test the hypotheses, unnecessary columns are dropped off and only “birth year” and “date” are left. Two sample group are split based on their birth year: users who born after 1980 (inclusing 1980) belongs to the young group, while users who born before 1980 is assigned to other age groups. By applying groupby() and count() function, finally the number of people in different group in each day is got as Fig. 3 shown.

Methodology

First, the data is visualized through bar chart as shown in Fig. 4. In order to showing the directly comparison, the young group and other age groups are separated as two bars in each day rather than combining in one bar. Surprisingly, in most day of a month, the number of riders who born before 1980 is more than who



	Number of people born after 1980	Number of people born before 1980
date		
2015-07-01	16018	17460
2015-07-02	14630	16124
2015-07-03	9216	9171
2015-07-04	5985	6207
2015-07-05	8533	7854

Figure 3: The number of people in different group in each day

born after 1980. Only 7 days (7/5, 7/11, 7/12, 7/18, 7/19, 7/25, 7/26) young group ride more, which is out of expectation. It also indicates that the hypotheses may need to reverse and then test again.

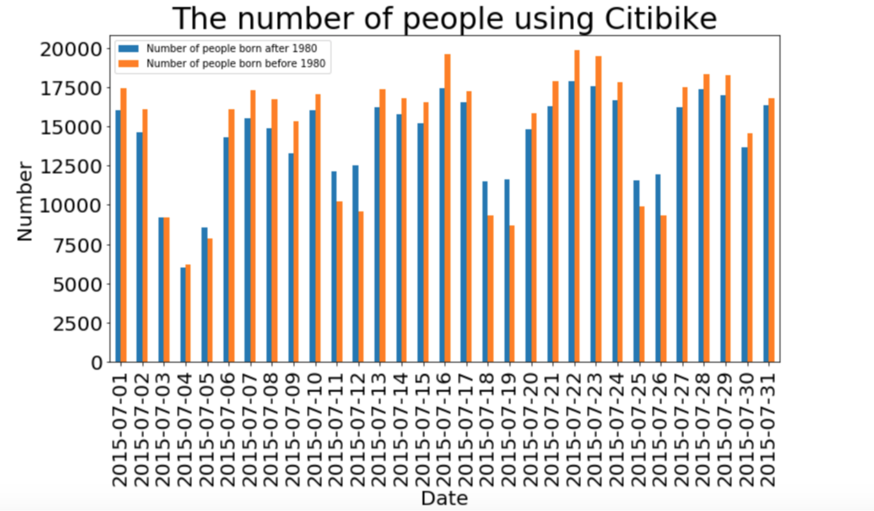
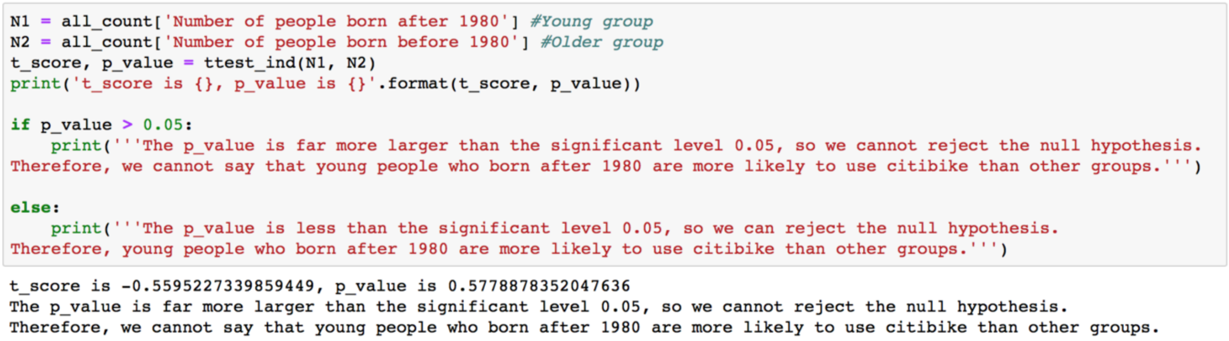


Figure 4: The distribution of number of users by age group in each day

Second is to choose a test to test the hypotheses. Through peer-review (thanks peer cd2682 and ab8131), it is suggested that z-test should be used. However, giving the hypotheses and dataset, the t-test may be more suitable. T-test is always used for testing the null hypothesis that 2 independent samples have identical average (expected) values. The 2 samples here are two types of group draw from total population, with unknown population mean and standard deviation. Therefore, the objective of this test is not testing the sample drawn belongs to the same population but comparing the mean of two given samples. In addition, the sample size just around 30 so it's not a large sample size. Therefore, the t-test should be applied.

Conclusions

The ttest_ind method imported from scipy.stats is used for calculating the t-score and p-value. Results are shown in Fig.5, since the p-value 0.58 is far larger than the significant level 0.05, we cannot reject the null hypothesis that the number of rider who born after 1980 (including 1980) is same or lower than who born before 1980. The idea that young people tend to ride more than other age groups cannot be proved.

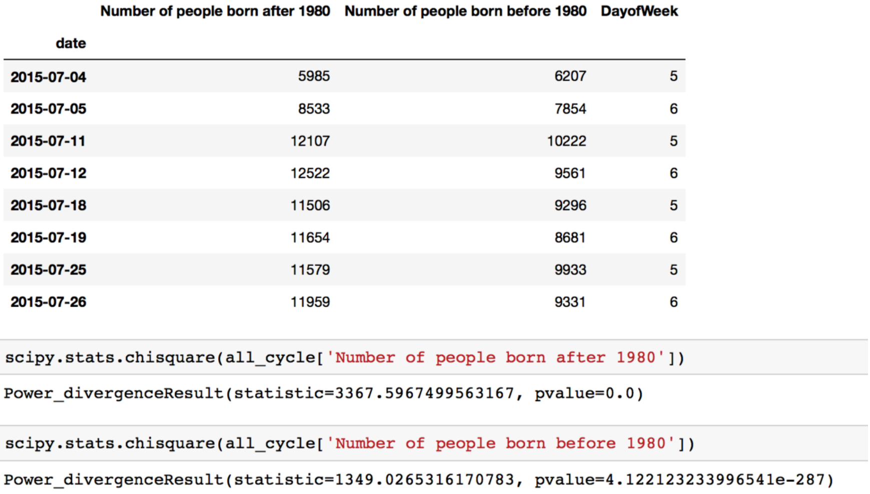


```
N1 = all_count['Number of people born after 1980'] #Young group
N2 = all_count['Number of people born before 1980'] #Older group
t_score, p_value = ttest_ind(N1, N2)
print('t_score is {}, p_value is {}'.format(t_score, p_value))

if p_value > 0.05:
    print('''The p_value is far more larger than the significant level 0.05, so we cannot reject the null hypothesis.
Therefore, we cannot say that young people who born after 1980 are more likely to use citibike than other groups.''')

else:
    print('''The p_value is less than the significant level 0.05, so we can reject the null hypothesis.
Therefore, young people who born after 1980 are more likely to use citibike than other groups.''')
```

```
t_score is -0.5595227339859449, p_value is 0.5778878352047636
The p_value is far more larger than the significant level 0.05, so we cannot reject the null hypothesis.
Therefore, we cannot say that young people who born after 1980 are more likely to use citibike than other groups.
```

Figure 5: The results of t-tes

Another interesting point is that the number of users seems to show the seasonal effect in both group. After applying the dt.dayofweek function as shown in FiG. 6, it is interesting that the number of riders in Saturday and Sunday tend to less than that in weekdays for both two groups. In order to test whether there exist periodic time series, the Chi-Squared is applied. Under the null hypothesis that there is no seasonal effect (i.e., F0(x) is a uniform distribution) and significant level as 0.05, we get the p-value roughly 0.0 for both two samples, which is less than significant level 0.05. Therefore, the null hypothesis can be rejected and there exist seasonal effect in this dataset.



date	Number of people born after 1980	Number of people born before 1980	DayofWeek
2015-07-04	5985	6207	5
2015-07-05	8533	7854	6
2015-07-11	12107	10222	5
2015-07-12	12522	9561	6
2015-07-18	11506	9296	5
2015-07-19	11654	8681	6
2015-07-25	11579	9933	5
2015-07-26	11959	9331	6

```
scipy.stats.chisquare(all_cycle['Number of people born after 1980'])
```

```
Power_divergenceResult(statistic=3367.5967499563167, pvalue=0.0)
```

```
scipy.stats.chisquare(all_cycle['Number of people born before 1980'])
```

```
Power_divergenceResult(statistic=1349.0265316170783, pvalue=4.122123233996541e-287)
```

Figure 6: Detect seasonal effect and test by Chi-Squared test

In conclusion, there is no statistical significance can prove that the younger people ride more than other age groups. Therefore, Citi Bike may need to reconsider the target consumers of their promotional activities. Some weakness in this mini-project such as only one-month dataset is insufficient need to be improved in the future. In addition, the interesting findings of seasonal effect in this time-series data is also worth to do further reaserch.

References

About Citi Bike: Company, History, Motivate — Citi Bike NYC. https://www.citibikenyc.com/about. URL `http://www.citibikenyc.com/about`. Accessed on Wed, November 07, 2018.

Sharing economy - Wikipedia. https://en.wikipedia.org/wiki/Sharing$_{e}conomy. URL.$ Accessed on Wed, November 07, 2018.

Campbell Mithun. National study quantifies reality of the “sharing economy” movement. 2012.

Schmitt. Experiential marketing. *Journal of Marketing Management*, 15(1-3), 53–67., 1999.