cz1605_Citibike miniproject

Chenrui Zhang[1]

[1]Affiliation not available

November 8, 2017

Abstract:

The Citibike users have two user types, including subscribers and customers. Subscribers have long-term(several months or even a year) membership, using citibike very often and paying renting fee monthly while customers make one-off consumption, relatively seldom using Citibike and paying for ride every time they use the bike. Based on the above features of the two user types, I want to figure out whether there is difference between the two types in term of the using date.

This project specifically sets the null hypothesis that the ratio of subscribers biking on weekends over subscribers biking on the whole week is the same or higher than the ratio of customers biking over weekends to customers biking on the whole week. I use the proportion z-test to test the proportions of the two samples, getting z-score= -252.09274 and p-value=0(significance level=0.05). Since p-value is 0, the alternative hypothesis could be accepted that the ratio of subscribers biking on weekends over subscribers biking on the whole week is lower than the ratio of customers biking over weekends to customers biking on the whole week.

Introduction:

Citibike is a privately owned public bicycle sharing system serving New York City and Jersey City, New Jersey. As of March 2016, the total number of annual subscribers is 163,865. Citibike riders took an average of 38,491 rides per day in 2016. The system reached a total of 50 million rides in October 2017(Wikipedia). The user types of Citibike are subscribers and customers. I assume that subscribers relatively more rely on Citibike as their transportation tools while customers may usually take advantages of other tools and use Citibike for some special preference. For lots of people, they drive cars or take subways to get to the workplace on weekdays and go for a ride on weekends so that they do not need to be subscriders. So it is interesting to find whether the ratio of subscribers biking on weekends over subscribers biking on the whole week is lower.

Data:

I use the dataset "201705 Citibike" available from the CUSP data facility(DF). The dataset contains the information about all rides records in New York City on May,2017 and I process the data by jupyter ipython notebook.

Step 1 : Access to the data source and read it and get all the columns of the dataset.

```
In [2]: zipFile =zipfile.ZipFile("/gws/open/Student/citibike/201705-citibike-tripdata.csv.zip")

In [3]: zipFile.extractall(path=(os.getenv('PUIDATA')+'/citibike/'))

In [4]: df = pd.read_csv(os.getenv('PUIDATA')+'/citibike/201705-citibike-tripdata.csv')
```

Figure 1: Access to the data by zipfile function and read it

```
In [7]: df.columns

Out[7]: Index([u'tripduration', u'starttime', u'stoptime', u'start station id',
       u'start station name', u'start station latitude',
       u'start station longitude', u'end station id', u'end station name',
       u'end station latitude', u'end station longitude', u'bikeid',
       u'usertype', u'birth year', u'gender'],
      dtype='object')
```

Figure 2: Get all the column of the dataset

Step 2 : Filter the columns required for the test(user type and start time)and convert the time to datatime format for further data processing.

Step 3 : Calculate two usertypes' ratios of biking on each day over biking on the whole week and plot it.

Methodology :

The null hypothesis and significance level are as below:

Because I am comparing two ratios whose values come from two samples and range from 0 to 1, both proportion z-test and proportion chi-square test can be adopted. I finally use the proportion z-test method because I'm more familiar with it.

I use the function from statsmodels.stats.proportion to do the test in ipython notebook:

At first I calculate the ratio of biking on weekends over biking on the whole week and fill all needed data in the proportions_ztest function to get z-score and p-value.

Conclusions :

Under the significance level=0.05, we can see that p-value=0.0<0.5 so we can reject the null hypothesis that the ratio of subscribers biking on weekends over subscribers biking on the whole week is the same or higher than the ratio of customers biking over weekends to customers biking on the whole week. Then we can accept the alternative hypothesis that ratio of subscribers biking on weekends over subscribers biking on the whole week is lower. And it can tell us that customer prefer riding on weekends.

Weakness: (1). The sample size is small, the ratio may not represent the overall trend.

```
In [8]: df=df[['usertype','starttime']]

In [9]: df.head()
```

Out[9]:

	usertype	starttime
0	Subscriber	2017-05-01 00:00:13
1	Subscriber	2017-05-01 00:00:19
2	Subscriber	2017-05-01 00:00:19
3	Subscriber	2017-05-01 00:00:24
4	Subscriber	2017-05-01 00:00:29

```
In [10]: df= df.rename(columns={'starttime':'time'})
         df['time'] = pd.to_datetime(df['time'])
         df.head()
```

Out[10]:

	usertype	time
0	Subscriber	2017-05-01 00:00:13
1	Subscriber	2017-05-01 00:00:19
2	Subscriber	2017-05-01 00:00:19
3	Subscriber	2017-05-01 00:00:24
4	Subscriber	2017-05-01 00:00:29

Figure 3: Select two columns ['usertype'],['starttime'] and convert ['starttime] to datetime format and rename it to ['time']

(2). Ride duration is not considered. If some one use Citibike for 10 hours on weekdays and 1 hour on weekends, only counting ride times may not fully make sense in term of date distribution.

Strength: The test model fit the problem well and it can provide accurate conclusion based on the null hypothesis.

In [13]:
```
fig = pl.figure(figsize(15,10))

norm_Sub = counts_Sub.sum()
error_Sub = np.sqrt(counts_Sub)
((counts_Sub) / norm_Sub).plot(kind="bar", color='IndianRed',
                           yerr=[((error_Sub) / norm_Sub, (error_Sub) / norm_Sub)],
                           label='Subscriber bikers')

norm_Cus = counts_Cus.sum()
ax = ((counts_Cus) / norm_Cus).plot(kind="bar", alpha=0.5,
                                yerr=[((error_Cus) / norm_Cus, (error_Cus) / norm_Cus)],
                                color='SteelBlue', label='Customer bikers')

ax.xaxis.set_ticklabels(['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun'], fontsize=20)
ax.set_ylabel ("Fraction of rides")
ax.set_xlabel ("Day of the week")

pl.legend(['Subscriber bikers','Customer bikers'],fontsize=20)
```

Out[13]:
```
<matplotlib.legend.Legend at 0x7fbffd464550>
```

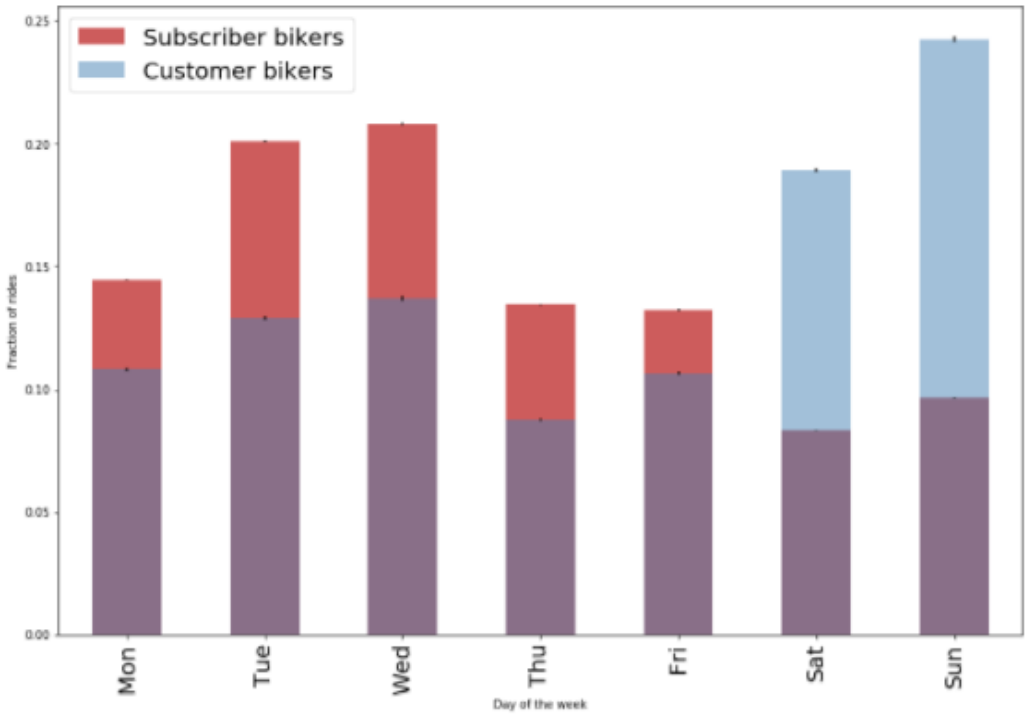

Figure 4: Distribution of Citibike bikers by usertype in May 2017, normalized

Reference :

Citibike - Wikipedia. Link

Null Hypothesis:

The ratio of Subscribers biking on weekends over Subscribers biking on the whole week is the same or higher than the ratio of Customers biking over weekends to Customers biking on the whole week

$$H_0: \frac{Sub\text{weekend}}{Sub\text{week}} >= \frac{Cus\text{weekend}}{Cus_{\text{week}}}$$

$$H_1: \frac{Sub\text{weekend}}{Sub\text{week}} < \frac{Cus\text{weekend}}{Cus_{\text{week}}}$$

I will use a significance level $\alpha = 0.05$

Figure 5: null hypothesis and significance level

```
In [34]: import statsmodels.stats.proportion as st

In [19]: counts_Sub

Out[19]: time
         0    192909
         1    267929
         2    277617
         3    179270
         4    176395
         5    110806
         6    128582
         Name: time, dtype: int64

In [36]: Sub_weekend = counts_Sub[5:].sum()
         Cus_weekend = counts_Cus[5:].sum()
         count = np.array([Sub_weekend,Cus_weekend])
         nobs = np.array([norm_Sub,norm_Cus])
         st.proportions_ztest(count, nobs, value=0, alternative='smaller', prop_var=False)

Out[36]: (-252.09274205968498, 0.0)
```

Figure 6: Apply proportion z-test for testing the null hypothesis

```
In [41]: Sub_weekend = counts_Sub[5:].sum()
         Cus_weekend = counts_Cus[5:].sum()
         count = np.array([Sub_weekend,Cus_weekend])
         nobs = np.array([norm_Sub,norm_Cus])
         result = st.proportions_ztest(count, nobs, value=0, alternative='smaller', prop_var=False)
         print('z-score is {} and p-value is {}'.format(result[0],result[1]))

         z-score is -252.09274206 and p-value is 0.0
```

Figure 7: Proportion z-test result