Authorea (/) Beta HELP (https://intercom.help/authorea) EXPLORE C

Document share

PUI2017 HW7 Assignment 1

C *Ch3183* (/users/175552-ch3183)

Add Collaborator Manage

Who has more CitiBike usage on weekend,Age over 30 or under 30?

<NetID: ch3183>

Abstract:

CitiBike is a popular transportation alternative in New York City and is widely used by people across all ages. This project is designed to find out among those CitiBike riders, who have more usage of CitiBike on weekends over weekdays. The main idea is to divide the riders into two age groups:above 30 years old and under(includes equal to) 30 years old. By utlizing one month CitiBike data and Null Hypothesis Significance Test, we conclude younger generations who are under 30 years old are more prone to CitiBike on weekends.

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 is a quick and affordable way to get around town and very popular in NYC area. Analyzing CitiBike users' activities is one of the most important ways to understand the business and social behaviors. My project is trying to find out which group uses CitiBike more on weekend for transportation. We use 30 as divide age line because in general, people in the city below 30 years old are children, teenagers or singles , many of them are students or new starters in their careers. Meanwhile people over 30 years old might have families and stable jobs.

Data:

The data used for this project is CitiBike monthly ridership dataset . And specifically, the month of June 2016 dataset is used for analysis. It is provided by CitiBike Program, which can be accessed at their official website: https://www.citibikenyc.com/system-data (https://www.citibikenyc.com/system-data), and https://s3.amazonaws.com/tripdata/index.html (https://s3.amazonaws.com/tripdata/index.html). The dataset contains columns of trip duration, location information and riders information. To focus on our question mentioned above in introduction part. Only the column of the riders' birth year is kept, all the other columns are removed. Then riders who were born over 30 years ago are grouped together and summed up, same with riders who were born less than or equal to 30 years ago. Finally we plot these two groups' data into two figures, one is total quantity of each group's each week day's ridership, the other figure is each weekday's ridership fraction within their own groups. Note that for two groups data are plot into same figure for comparison.

	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	
0	1470	6/1/2016 00:00:18	6/1/2016 00:24:48	380	W 4 St & 7 Ave S	40.734011	-74.002939	3236	W 42 St & Dyer Ave	40.758985	-73.993800	19859	Subscriber	1972.0	1	00
1	229	6/1/2016 00:00:20	6/1/2016 00:04:09	3092	Berry St & N 8 St	40.719009	-73.958525	3103	N 11 St & Wythe Ave	40.721533	-73.957824	16233	Subscriber	1967.0	1	00
2	344	6/1/2016 00:00:21	6/1/2016 00:06:06	449	W 52 St & 9 Ave	40.764618	-73.987895	469	Broadway & W 53 St	40.763441	-73.982681	22397	Subscriber	1989.0	1	00
3	1120	6/1/2016 00:00:28	6/1/2016 00:19:09	522	E 51 St & Lexington Ave	40.757148	-73.972078	401	Allen St & Rivington St	40.720196	-73.989978	16231	Subscriber	1991.0	1	00
4	229	6/1/2016 00:00:53	6/1/2016 00:04:42	335	Washington Pl & Broadway	40.729039	-73.994046	285	Broadway & E 14 St	40.734546	-73.990741	15400	Subscriber	1989.0	1	00

Fig. 1

Frist everal Lines of the Original CitiBike Dataset

	date	birth year
0	2016-06-01 00:00:18	1972.0
1	2016-06-01 00:00:20	1967.0
2	2016-06-01 00:00:21	1989.0
3	2016-06-01 00:00:28	1991.0
4	2016-06-01 00:00:53	1989.0

Fig. 2

First several lines of the Processed Dataset With Only Riders' Age Information

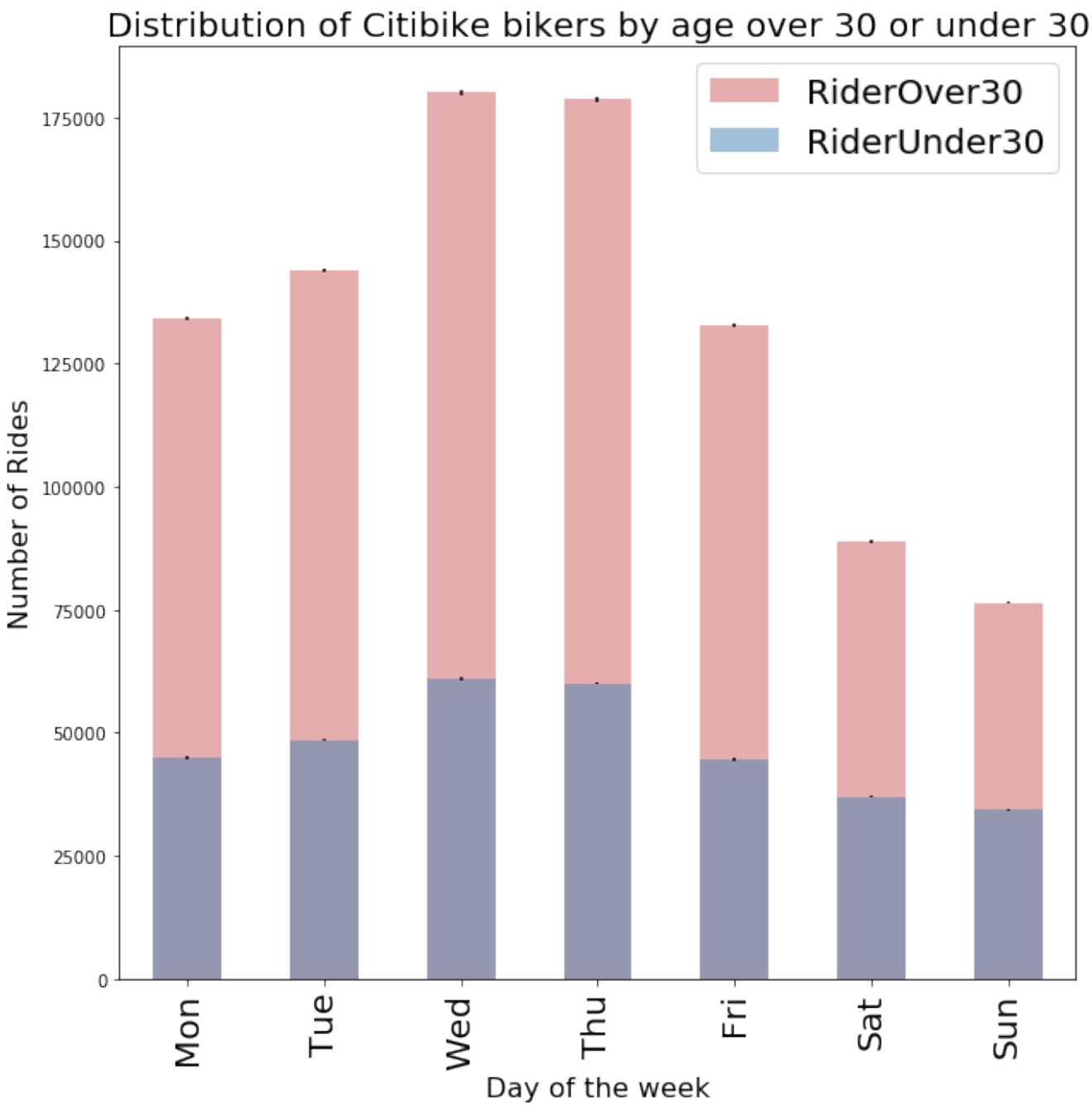

Fig. 3

Total quantity of each group's each week day's ridership

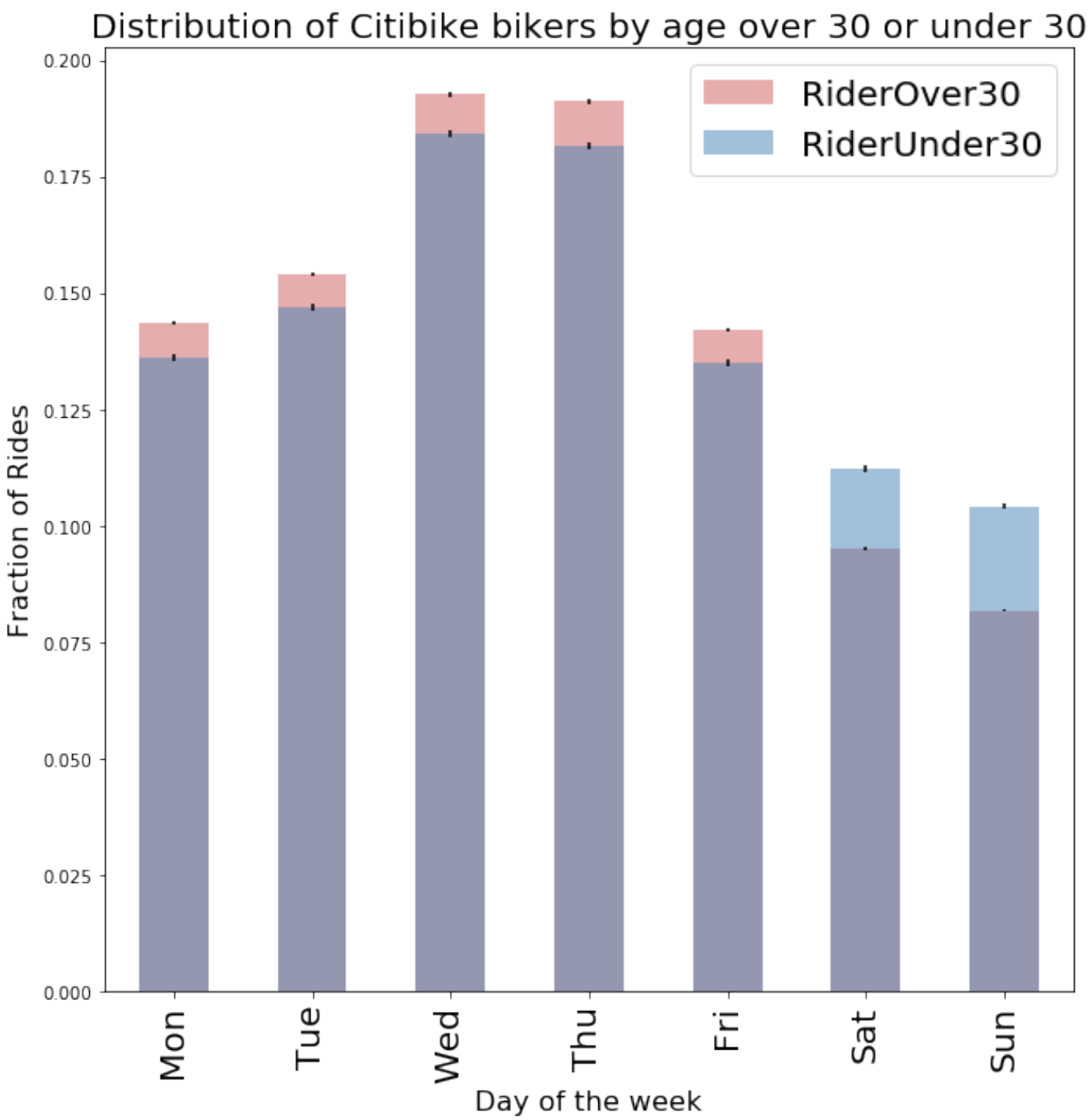

Fig. 4

Each weekday's ridership fraction within their own groups

Methodology:

To answer the project question which group is prone to CitiBike for weekend transportation, a Null Hypothesis is set up for the significance test. And we set the alpha as 0.05

NULL Hypothesis: The ratio of riders who are over 30 years old biking on weekends over the weekdays is the same or higher than the ratio of riders who are under 30 years old biking on weekends over biking on weekdays.

In formulas:

H_0: frac{{under30{weekend}}{under30{{week}}} <= frac{over30{weekend}}{over30{week}}

H_1: frac{under30{weekend}}{under30{week}}} > frac{over30{weekend}}{over30{week}}

Because the NHST is designed as to test a ratio with categorical endogenous variable. And the distributions are not parametrizable with a Gaussian. Appropriate test for such situation would be Fisher exact test, and chi sq test. But the Fisher exact test is suitable for small datasets which this one is obviously not, so the chi sq for proportion (contingency table)is a proper choice for the project.

A z-test is another option, it is simple, easy and quick. But it assumes simple random sampling from a normally distributed population, in this case, the riders population, it might be Normal and very likely to be, but we are not sure. Thus Chi Sq is better choice.

Conclusions:

rideship	on weekend	not on weekend	
age over 30	$0.177 * 934674$	$0.823 * 934674$	934674
age under 30	$0.216 * 330438$	$0.784 * 330438$	330438
total	236812	1028301	1265112

Fig. 5

TChi Square Test Contingency Table

From the Chi square test contingency table above, we can get the Chi square statistics as 2440.53 which is way more than the 3.84. It means the P <<0.05, so we can reject the NULL hypothesis that 'the ratio of riders who are over 30 years old biking on weekends over the weekdays is the same or higher than the ratio of riders who are under 30 years old biking on weekends over biking on weekdays'. It also means we can say the riders under or equal to 30 years old are more prone to CitiBike on weekends than riders over 30 years old.

Some interpretation:

In the dataset, generally, riders with birth year information are subscribers and it has a great chance that they are city residents. Older(>30) people might have family and stable jobs, during weekends, they probably spend more time at home with family or choose to go outside by driving together, and normally places for family activities are too far for biking. Younger people in the meantime, might do more social activity in town at some places close by, or use bikes to commute in college campus.

The weakness and potential further studies of this project are:

- (https://www.authorea.com/users/106033/articles/144161-pui2016-extra-credit-project/comments)Data limitation. Only use one month data might not enough to demonstrate a trend. We can improve the experiment by using more data, maybe a month from winter since this is a summer one.
- New York is such an international city that many young riders on weekend can also be visitors from other cities or even countries. Further study can look into how many of them are subscribers.
- Further study of these two age group ridership by areas, this can be done by using the location information to group them into different boroughs or even zip code areas.

Links:

https://github.com/hcpenguin/PUI2017_ch3183/blob/master/HW7_ch3183/HW7_assignment1.ipynb
(https://github.com/hcpenguin/PUI2017_ch3183/blob/master/HW7_ch3183/HW7_assignment1.ipynb)