CITIBIKE

ABSTRACT:

The study used Citi bike data for July 2016 in New York city to find out if the people under 30 is less likely to use citibike in daily life than elder people over 30 on weekends. The null hypothesis is that the proportion of people under 30 ride citbikes on weekends is the same or lower than the proportion of those aged over 30 in July,2016. Z-test was performed to test the null hypothesis. Through the analysis and calculating, I reject the null hypothesis. The result of z-test is 44.7, the pvalue under 0.05. The conclusion is that the proportion of people under 30 years old ride Citibikes on weekends is lower than the proportion of those over 30 on weekends in July, 2016.

Introduction:

Citi-Bike is becoming more and more popular in NYC daily public transportation. First launched in 2013, Citibike has now totals of 706 stations and 12,000 bikes which pushed itself to become the biggest bike-sharing system in the USA(Citi Bike – Wikipedia). In this study, I focus on the different age groups to see the differences in using citibikes because I want to know whether young people nowadays are more likely to use bike for commuting. Generally thinking, the bicycle is a cheap transit method and is convenient for younger people to use especially those under 30. For the people over 30, bikes may sometimes not a good choice for them in some district with low road safety. This project try to learn the patterns between two age groups and give suggestions for location of citibike station and number of bikes stored in stations.

Data

The datasets used for this study was July 2016 citibike dataset. I downloaded and then used ipython notebook to process the data.

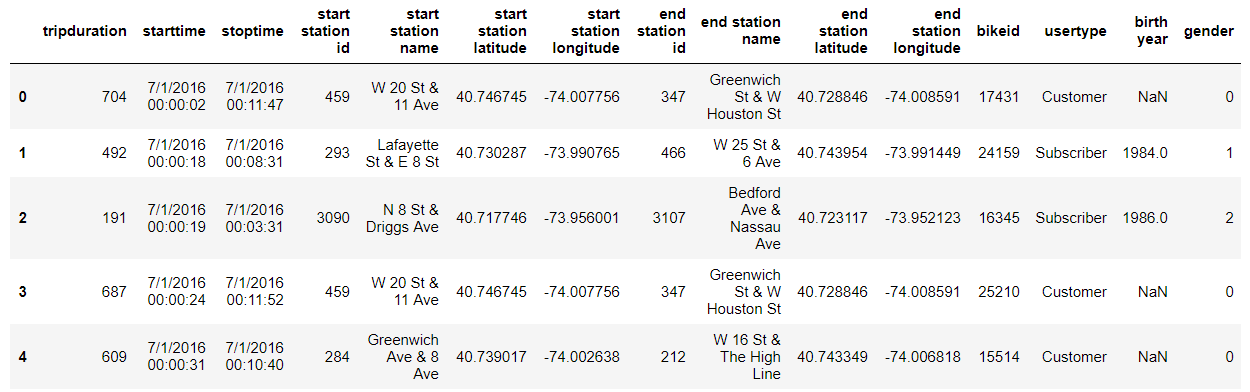


Figure1: July 2017 Citibike dataset

One important process is that I converted the “starttime” column which is string format into ‘date’ using the function ‘pd.to\_datetime’ and then I began to use counts to see the distribution of two age groups in specific weekdays.

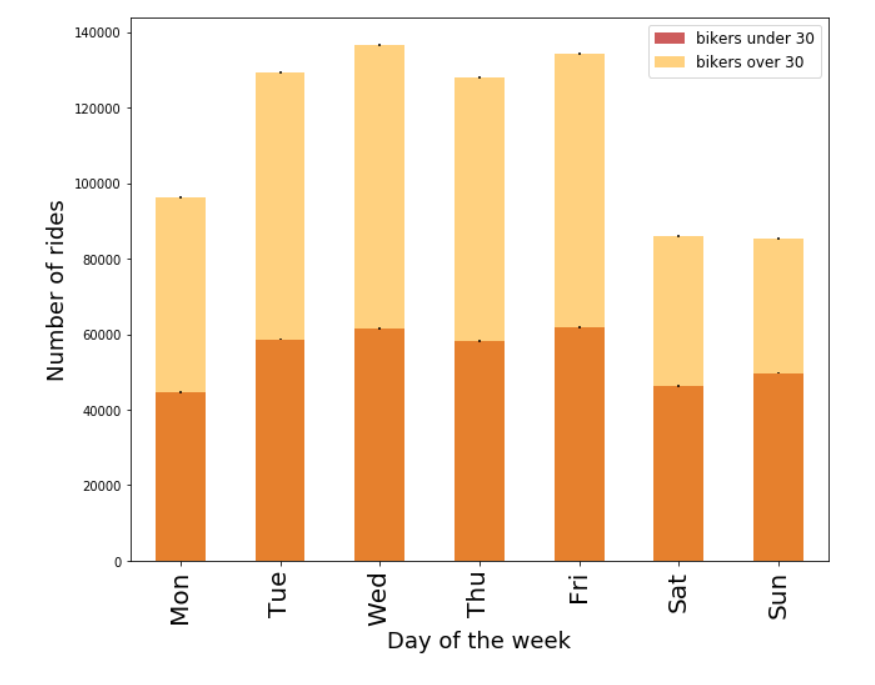


Figure2： The distribution of Citibike bikers by two age groups(under30 and over30) in July 2016, absolute counts, with statistical errors.

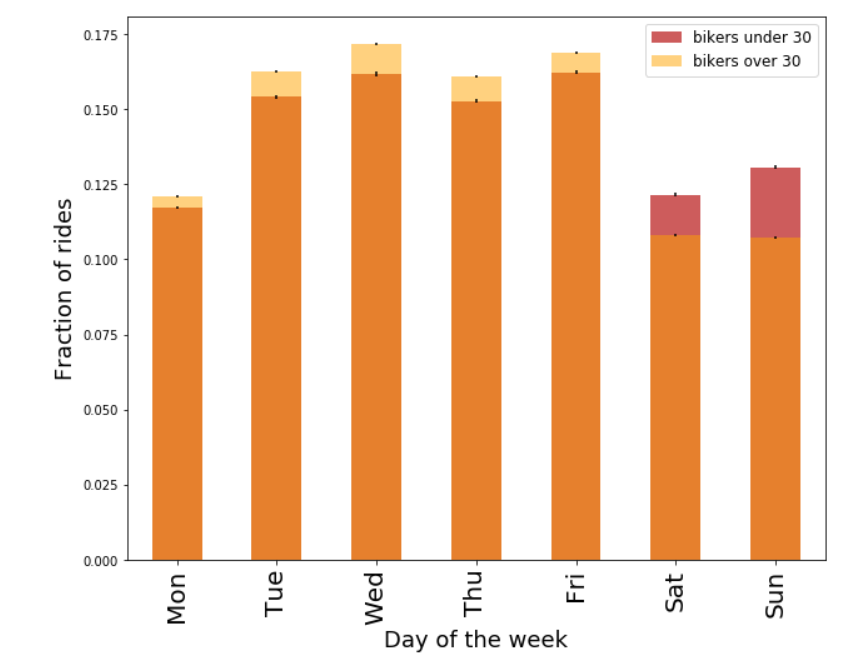
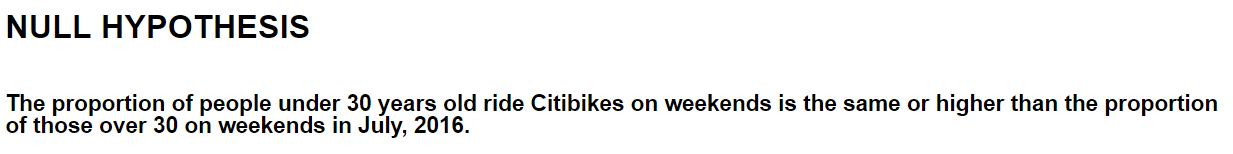


Figure3： Distribution of Citibike bikers by two age groups (under 30 and over 30) in July 2016, normalized

Methodology:



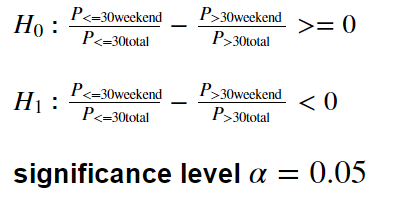
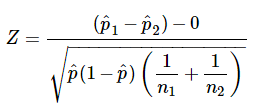


Figure4：The hypothesis and the significant level

After processing the data, I chose z-test to test the null hypothesis without following the peer suggestion from Kloe who suggested me to use t-test. Z test is proper and useful for this project to test H0 because the data for people under 30 and over 30 samples both from the same population, and it has one variable as usage quantity of the bike. And the number of sample is over 30. The alpha is 0.05.



The formula of ZTEST.

Conclusions：

The result of analysis is that the people under 30 are less likely than those over 30 to choose citi bikes for commuting on weekends.

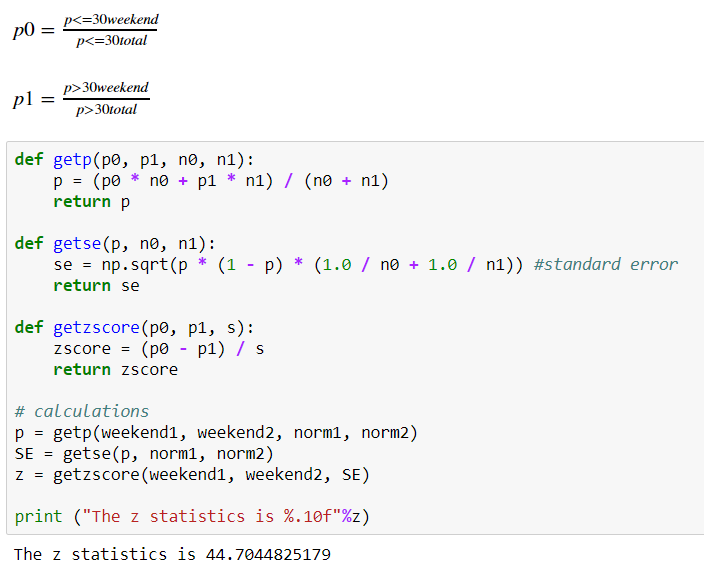


Figure3: the result of z-test

the z test results is 44.7, the p value is 0.00 which is smaller than the alpha 0.05 I chose. So I can reject the null hypothesis.

The strength of this study is that the conclusion is statistically significant. The weekness is that I just used one-month data to do the research. And I did not pick gender as a critical character given that different genders make different choices in different age groups.

**References**

Citi Bike - Wikipedia. [**Link**](https://en.wikipedia.org/wiki/Citi_Bike)