Citibike and Gender Trip Duration

Alia Kasem [1]

[1]New York University Shanghai

February 16, 2022

Abstract

This analysis will be focusing on Citibike ridership in terms of total trip duration based on male and female usage per miles and minutes. The study data period for this project from January 2016 to December 2016. The data explore the trip duration for citibike. The Null Hypothesis, the female duration is higher or equal to the male trip. The analysis relied on uses the T-testing as a statistical method to test the two means of samples.

A reproducible notebook can be accessed through Github here.

Introduction

CitiBike is a public bike sharing system that could be rented in different stations across New York City. CitiBike serves both New York City and New Jersy City. The CitiBike serves multiple type users mainly in the borough of Manhattan and Brooklyn. New York City was introduced to Citibike in 2013, shortly the Citibike stations became an iconic feature throughout the city as one of the major transportation modes. This analysis mainly examines the diversity of trip duration in terms of male and female use; the analyzed data provide an understanding of customer trends and behaviors. Based on the results of data analysis Citibike can increase business productivity based on usage of the total trip duration. (This analysis indicates some factors that contribute to differences in bike usage).

Converting the hypothesis to formulas: Null Hypothesis Male rider's trip duration is not significantly longer than or is equal to the trip duration of female riders.

Statement Normally, dressing code would affect the choice of biker riders for commute.

Null Hypothesis Male rider's trip duration is not significantly longer than or is equal to the trip duration of female riders.

Alternative Hypothesis Male rider's trip duration is longer or than the trip duration of female riders.

$H_0 : T\text{Man} <= T_{\text{Woman}}$ $H_1 : T\text{Man} > T_{\text{Woman}}$ or identically:

$H_0 : T\text{Man} - T_{\text{Woman}} <= 0$ $H_1 : T\text{Man} - T_{\text{Woman}} > 0$ The significance level chosen is $\alpha = 0.05$¶

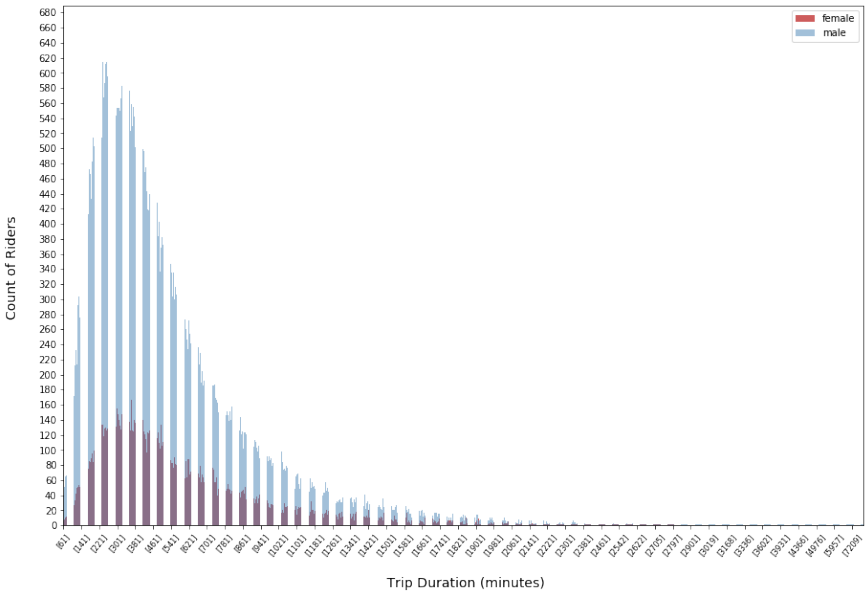


Figure 1: The plot indicates Male rides' trip duration is not significantly longer or equal to female trip duration.

Data

The data used for this analysis driven from: https://www.citibikenyc.com/system-data. Citi Bike Trip Data.

* Trip Duration (in seconds)

* Start Time and Date

* Stop Time and Date

* Start Station Name

* End Station Name

* Station ID

* Station Lat/Long

* Bike ID

* User Type (Customer = 24-hour pass or 7-day pass

* User; Subscriber = Annual Member)

* Gender and Year of Birth

The specific time frame for the two genders has been identified as Jan-2016 to Dec-2016. Throughout the year, we see male users higher than female users. Please note that there are multiple factors for trip duration. Data was then separated by gender (Male == 1 and Female == 2) to run the analysis .

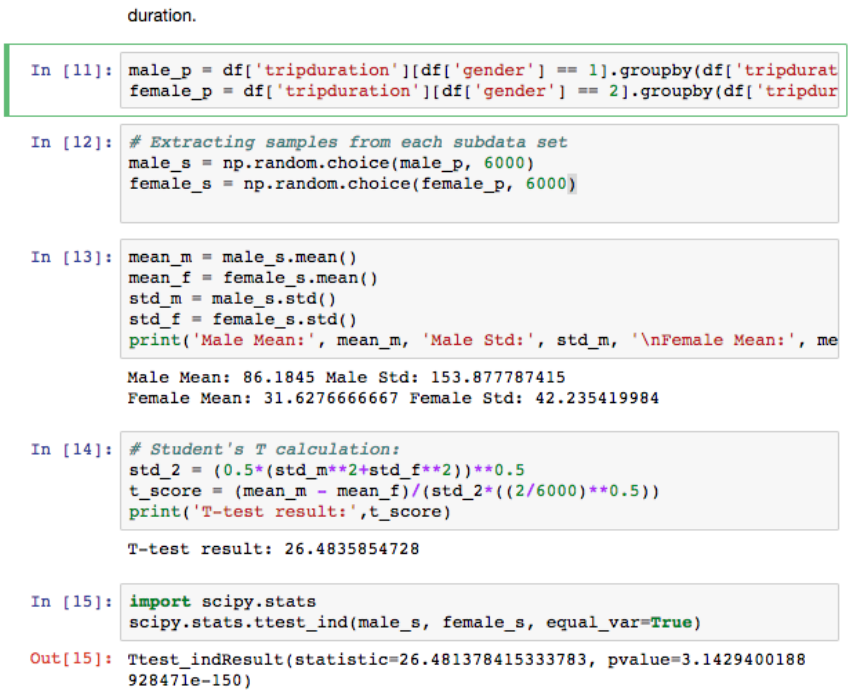
Out[16]:

	tripduration	gender	date
0	923	1	2016-01-01 00:00:41
1	379	1	2016-01-01 00:00:45
2	589	2	2016-01-01 00:00:48
3	889	2	2016-01-01 00:01:06
4	1480	1	2016-01-01 00:01:12

Figure 2: Male vs. Female ridership based on trip duration and gender

Methodolgy

The average trip duration for females is higher than males across all the days of January 2016. Frequently, dressing code would affect the choice of bike riders for a commute. Another factor is street geometry, for example, NYC does not provide a designated bike lane. However, shared bike lanes are available for users; keeping in mind that the shared bike lane might be a local truck route or heavily dominating the roads.



duration.

```
In [11]: male_p = df['tripduration'][df['gender'] == 1].groupby(df['tripdurat
         female_p = df['tripduration'][df['gender'] == 2].groupby(df['tripdur

In [12]: # Extracting samples from each subdata set
         male_s = np.random.choice(male_p, 6000)
         female_s = np.random.choice(female_p, 6000)

In [13]: mean_m = male_s.mean()
         mean_f = female_s.mean()
         std_m = male_s.std()
         std_f = female_s.std()
         print('Male Mean:', mean_m, 'Male Std:', std_m, '\nFemale Mean:', me

         Male Mean: 86.1845 Male Std: 153.877787415
         Female Mean: 31.6276666667 Female Std: 42.235419984

In [14]: # Student's T calculation:
         std_2 = (0.5*(std_m**2+std_f**2))**0.5
         t_score = (mean_m - mean_f)/(std_2*((2/6000)**0.5))
         print('T-test result:',t_score)

         T-test result: 26.4835854728

In [15]: import scipy.stats
         scipy.stats.ttest_ind(male_s, female_s, equal_var=True)

Out[15]: Ttest_indResult(statistic=26.481378415333783, pvalue=3.1429400188
         928471e-150)
```

Figure 3: T-testing results

Conclusion

The calculation of Student's T-test shows that the P-value is really low less than 0.05, which is the selected significance level. Therefore, we can reject the null hypothesis. Instead, the alternative hypothesis is valid in this case.

In 2016 January, the average trip duration of male riders is higher than the average trip duration of female riders.