Citi Bike: Gender Effect on Trip Duration

Quan Bui, Catherine Caruso, Patrick Goss, Hyesil Jung

Stats 501

Abstract

A study to see if there is a difference in travel time trips done by bikers who are male or female identified in New York City. We believe that there is a difference between the trips done by male and female riders. We performed a Two-Sample T test with a significance level of 0.05 and came to the conclusion that there is a difference. Data published by Citi Bike was examined and analyzed with Minitab.

1. Introduction

In that past few years, we saw the rise of new travel services. It started with ride sharing companies. Now we're getting bike sharing programs. A majority of these services are focused on the city-living demographic. Which is understandable because in cities like Boston and New York City, there are so many cars on the street. Going along with that, not everyone in the city owns a car due to space or money reasons so it only makes sense that these services arise. Citi Bike is a bike sharing program servicing New York City and has been running since 2013. Citi Bike collects and publishes data on monthly Citi Bike trips. They keep track of data like trip start and end points, trip length, gender, and more. New York is a perfect sample of a diverse and busy city that has heavy use of bike sharing services. We feel that this data holds valuable information about the travel, bikers, and other trends in New York City.

The question this paper seeks to answer is if there is a difference in the average trip time done by a male versus a female. This information is valuable as it allows us to examine the biking habits of men and women. Additionally, we will also take a look at the number of trips done by male and female riders within the data and examine if there is a significant difference between the two that may be affecting our data. This study is interesting and valuable because there gender equality is a hot topic for debate in today's society.

2. Data

2.1 Variables Of Interest

We are using the data collected for rides in October of 2018. There are 1,878,657 total rides in this data set. We chose the data from October of 2018 because it was one of the months with the most amount of ride data available. One unit of data corresponds to a single trip by a single rider. The variables of the data are shown below in Table 1. The variables that will be of interest to us are Trip Duration and Gender.

Table 1: *201810-citibike-tripdata.csv*

Variable	Unit	Description
Trip Duration	seconds	Duration of a single bike trip
Start Time and Date	MM\DD\YY HH:MM:SS	Time and date of when the trip started
Stop Time and Date	MM\DD\YY HH:MM:SS	Time and date of when the trip ended
Start Station Name		Name of the station the trip started
Start Station ID		ID of the station the trip started
Start Station Lat	degrees	Latitude of the station the trip started
Start Station Long	degrees	Longitude of the station the trip started
End Station Name		Name of the station the trip ended
End Station ID		ID of the station the trip started
End Station Lat	degrees	Latitude of the station the trip started
End Station Long	degrees	Longitude of the station the trip started
Bike ID		ID of the bike used for the trip
User Type		Type of user on the trip
Gender		Gender of user on the trip
Year of Birth	YYYY	Birth year of user on the trip

User Type: Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member
Gender: Zero=unknown; 1=male; 2=female

2.2 Descriptive Stats

Statistics

Variable	gender	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
tripduration	1	1243787	0	643.16	0.374	417.21	61.00	325.00	529.00	864.00	1934.00
	2	420103	0	726.29	0.671	434.65	61.00	385.00	619.00	987.00	1934.00

Figure 1: Descriptive Stats for the trips done by males and females after taking out outliers

The raw data set includes unreported user genders as 0. For analyzing purposes, we will not be including unknown genders. This is because the unreported gender data will not hold anything relevant to the study. We will also exclude outliers. Some of the rides last days long and do not make sense. They could be from people forgetting to return the bike to a station.

After filtering data, there are a total of 1,663,890 trips (1,243,787 males, 320,103 females). In the filtered data, we can see that there are more trips recorded for males in this sample. Along with females having a higher mean and median in these samples.

As we can see, the females have a higher mean and median in these samples. The standard deviations are close and the variances are almost the same. (20.42 and 20.83 respectively). In order to estimate if there is a difference in the true population means for men and women we will perform a 2-sample t test for a large sample and calculate a 95% confidence interval for the difference between the two means using Minitab.

2.3 Data Visualization

From Figure 2, we can see that the sample populations are not necessarily normal. But we are using a large number and will make assumptions about the normality of the total populations. As well as the fact that we have a much higher number of trips made by males.

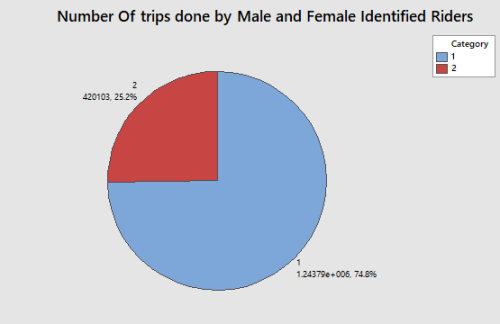

Figure 2: A pie chart the number of trips done by males vs. females (1=Male, 2= Female) There are 1243787 males, 320103 females

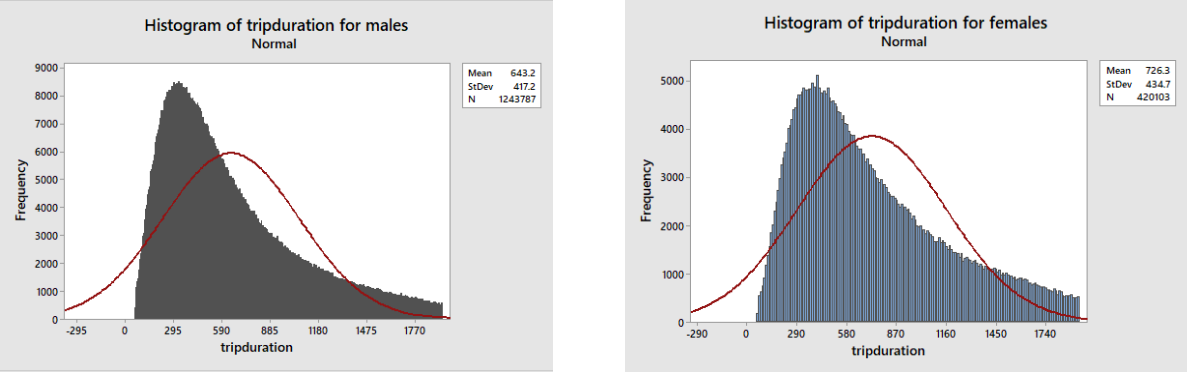

Figure 3: Histograms for trip duration for both males and females

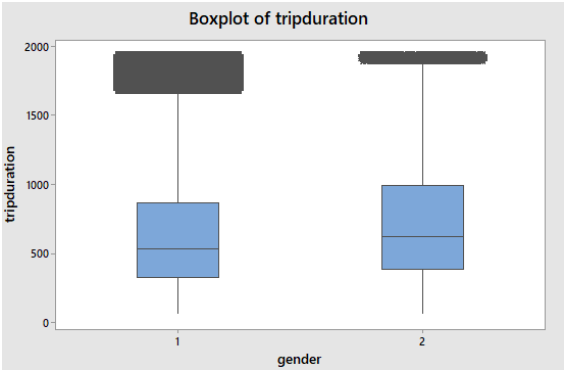

Figure 4: A box plot of the two samples

2.4 Other

Assumptions

- That one month's trip data is a random sample
- That every trip made is independent
- The populations (trips made by either of the genders we are looking) are normal. (Because these are large samples)

3. Analysis

Two-Sample T-Test and CI: tripduration, gender

Method

μ_1: mean of tripduration when gender = 1
μ_2: mean of tripduration when gender = 2
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics: tripduration

gender	N	Mean	StDev	SE Mean
1	1243787	643	417	0.37
2	420103	726	435	0.67

Estimation for Difference

Difference	95% CI for Difference
-83.128	(-84.633, -81.623)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-108.26	699382	0.000

Figure 5: Two-Sample T-Test and Confidence Interval

As is clear from all the statistical tests, there is in fact a difference between the average duration in trips made by males and females. The two sample t-test and confidence interval

show this clearly. The null hypothesis is that there is no difference between the means, and the alternative hypothesis is that there is a difference.

If we were to use the critical value approach with $\alpha = 0.05$, we would check whether or not the T-value is greater than 1.96 or less than -1.96. The T-value from the test is -108.26 which is much less than -1.96, so we can reject the null hypothesis.

Using the p-value approach, we check if the p-value is is less than 0.05. The p-value from the test is 0.000 , so we again can reject the null hypothesis and say these results are very highly significant.

The 95% confidence interval for male average-female average is (-84.633, -81.623). This interval does not include 0 so we can reject the null hypothesis .

4. Conclusions

Our findings appear to be significant in proving that there is a difference between the average trip duration of males and females in New York City. Real world implications of this is that there is something attributing in the realm of sex and gender to the difference in duration in trips made by males and females. Whether it be biological or concerns of bike safety in the city.

However, we do have a few concerns about our data and how it affected the results. One concern is how extreme our t-value is of -108.26. We believe this large t-value may have to do with the large difference between the sample sizes of males and females in our data. If we were to do our testing again, we would hope to use data where the number of trips for males and females is more similar.

Also, the samples are right skewed. Which isn't desirable, but we worked with large samples to combat this.

Another concern is that this study would probably be best done if we had average trip duration per user by gender. Unfortunately, the data isn't tied to any specific user key or ID. So this study does not account for a handful of users who take repeated bike rides.

Our data could have benefited from the use of more than one month of data, however for this reports purpose the data from October worked. If we were to collect our own data we would likely record the same variables but not allow a rider to be listed with an unknown gender. The large number of trips with the unknown gender could have been female riders that could have balanced our data. Since we omitted the unknown gender trips in our testing, we could have altered the accuracy of the results.