Citibike usage by gender

Nicholas Jones[1]

[1]NYU Center for Urban Science & Progress

November 9, 2017

Abstract

This project investigated the usage of the Citibike bikeshare system in New York City. Through analysis of large-scale data on rider demographics and characteristics of individual rides, it was possible to explore how uptake of the Citibike system differs by demographic group. We tested for difference in ride length between male and female users. The study found a statistically significant different in mean ride length, with men taking longer rides than women on average during the test period June-July 2017. The finding is of policy interest given New York City's commitment to equal opportunities, gender empowerment and mobility for all, and may support the need for outreach and awareness campaigns aimed at narrowing the gender gap in Citibike usage.

Introduction

Many cities around the world have introduced bike-share schemes in the past decade. City governments see these schemes as a means to promote several policy goals, including improved connectivity, better public health, reduced pollution, and enjoyment of residents and tourists (Fishman et al., 2013). New York joined this trend in May 2003 when it launched Citibike with an initial 6,000 cycles at more than 100 stations in Manhattan.

The scheme subsequently expanded to 10,000 cycles. The distance travelled on Citibikes has now surpassed 3,300,000 miles, equating to just over 13 times the distance between the earth and the moon (www.citibikeblog.tumblr.com).

A key consideration for city government is to ensure equitable provision of services such as Citibike to all citizens. The degree to which Citibike by different demographic groups can be influenced by behavioral, physical and economic factors, but these havereceived little independent study (Faghih-Imani and Eluru, 2016). We therefore examined usage of Citibike between the genders to determine whether take-up of this service differs systematically on a gender basis. This question is of policy relevance to New York City government and civic groups, and of operational relevance to the Citibike franchise operator, given the shared interest of these actors in ensuring Citibike is a service for all New Yorkers and visitors (not just a subset of them).

Data

The data utilized was Citibike usage data for June and July 2017. The data were downloaded from `https://www.citibikenyc.com/system-data` on November 6, 2017.

The dataset comprises trip duration, start and stop times and locations, and information on customers making the journeys - including gender, year of birth and subscription type.

A key data issue for our analysis is the nature of the gender field, which records values of '1' for male, '2' for female, and '0' for unknown. The data was provided in CSV form.

Several data cleaning tasks were carried out. Fields other than gender and trip duration were dropped. The dataset initially comprised 3,293,678 rows. A descriptive summary of the statistics was generated (Figure 1) and found to suggest the presence of significant outliers, indicated by extremely high maximum values which exceeded 3,000,000 seconds (or 800 hours). On further manual inspection of the data, it was determined that these high outlying values likely derive from errors such as docking stations not working properly, or from lost cycles.

	count	mean	std	min	25%	50%	75%	max
Male	2231314.0	915.674561	11655.232424	61.0	356.0	586.0	997.0	3765047.0
Female	802274.0	1021.775620	10430.924660	61.0	426.0	702.0	1175.0	3050849.0

Figure 1: Descriptive statistics for Citibike dataset: before removal of outliers

Given the research's interest in actual rides carried out by New Yorkers, it was determined that such outlying values should be dropped. The 90% quantile was calculated and trips with duration above this level were dropped from the dataset, reducing its size to 3,120,128.

The data was cleaned to remove fields other than gender and trip duration, and to remove outliers. The 90% quantile of trip duration was calculated and values above this level were dropped. The cleaned dataset had much more realistic mean and maximum values, allowing the analysis to proceed.

	count	mean	std	min	25%	50%	75%	max
Male	2070291.0	646.286923	391.400315	61.0	341.0	547.0	877.0	1738.0
Female	721359.0	725.304064	405.057347	61.0	401.0	637.0	994.0	1738.0

Figure 2: Descriptive statistics for Citibike dataset: after removal of outliers

Methodology

The research hypothesis was that males make longer rides on average than females. This was based on anecdotal evidence from observing Citibike riders and reading blogs and civic group's reports on bicycling in New York. In particular, the literature examined indicated that females often have greater safety concerns than males, which may act as a behavioral barriers to riding for long periods.

The following hypotheses were investigated:

H0: Mean ride length (males) <= mean ride length (females).

H1: Mean ride length (males) > mean ride length (females).

Significance level: alpha = 0.05.

The test adopted was a two-sample T-test. The T-test is required because the variance of the two samples is not known, therefore the sample follows a t-distribution rather than a normal distribution.

The decision rule was to reject the null hypothesis if the p-value / 2 was less than if alpha. This is because we are conducting a one-sided test, and the t-test produces a T-test produced a p-value of less than alpha.

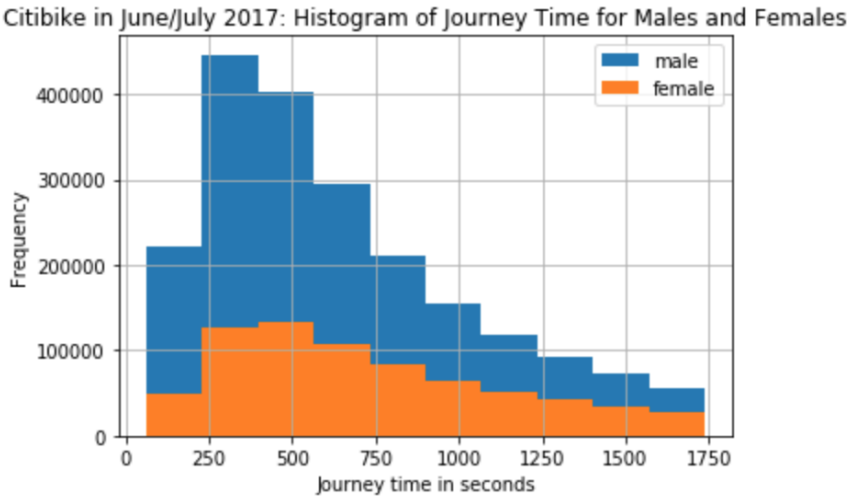

Figure 3: Histogram of Journey Times

The T-test produced output of:

- t statistic: -146.323
- p value: 0.00.

Accordingly, the null hypothesis was rejected.

The statistical test supported the data visualization shown in Figure 4, providing a robust support for the conclusion.

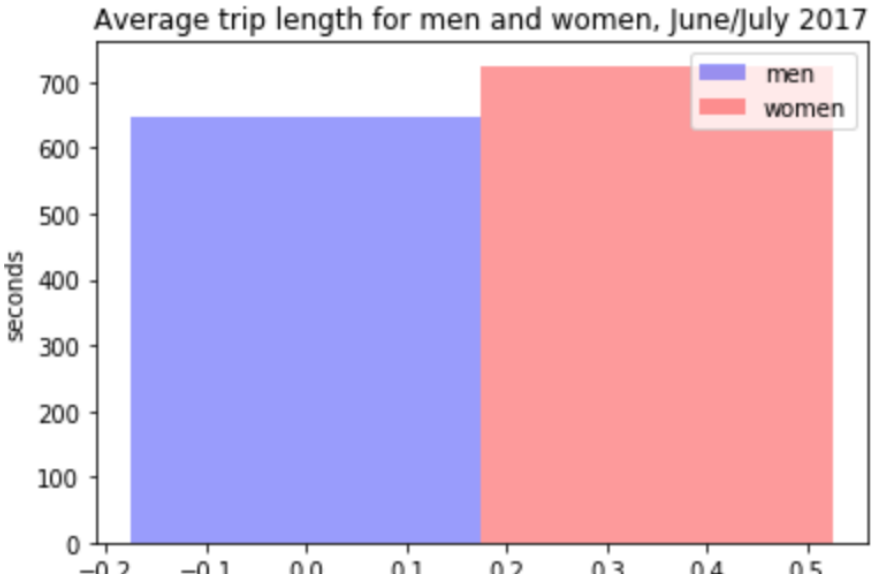

Figure 4: Mean ride length for males and females

Conclusion

The project investigated average length of Citibike rides by males and females. We used a two-sample T-test to compare male and female ride length, using a dataset of some 3 million rides from June and July 2017. At significance level 0.05, we rejected the null hypothesis that males make rides shorter than or equal to those made by females. This finding is supportive of our original research idea. It appears likely that there is indeed a statistically significant difference. The finding provides motivation for further research. Additional research questions could include whether males make longer rides than females systematically during all times of the week - including during commuting hours and weekends - and whether the lower propensity for long rides is shared by females at younger and older age ranges. Such an extension of the research could help identify specific reasons for shorter rides by females, such as whether this is simply due to less weekday commuting or whether safety concerns on the part of some of the female population is impeding their Citibike use. Based on this analysis, policy measures could be adopted to make Citibike more attractive to females and alleviate the difference in uptake suggested by this study.

Acknowledgements

I worked on the project by myself. Reuben provided peer review input. Prince gave me a review of t-statistics versus the possible alternative approach of a two-way Anova test, supporting my decision to choose a t-test.

References

Ahmadreza Faghih-Imani and Naveen Eluru. Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system. *Journal of Transport Geography*, 54:218–227, jun 2016. doi: 10.1016/j.jtrangeo.2016.06.008. URL https://doi.org/10.1016%2Fj.jtrangeo.2016.06.008.

Elliot Fishman, Simon Washington, and Narelle Haworth. Bike Share: A Synthesis of the Literature. *Transport Reviews*, 33(2):148–165, mar 2013. doi: 10.1080/01441647.2013.775612. URL https://doi.org/10.1080%2F01441647.2013.775612.