Do Younger New Yorkers Bike Later? An Analysis of CitiBike Ridership by Age and Time of Day

sgo230[1]

[1]Affiliation not available

November 9, 2017

Abstract

I analyzed Citibike data in New York City to assess whether there is statistically significant difference in the proportion of young riders compared to old riders in daytime versus nighttime. A chi-squared test of proportions was performed on Citibike data for July 2017 consisting of 1,735,599 rides. The share of riders over 35 years old of all riders was compared for daytime rides (5 A.M. to 11:59 P.M.) versus nighttime rides (12:00 A.M. to 4:59 A.M.). The null hypothesis that older riders make up a greater or equal proportion of riders at night than during the day was rejected at the 95% confidence level with a chi-squared test statistic of 1471.5. The same test was performed on February 2016 Citibike data consisting of 560,874 rides, and the null hypothesis was rejected at the 95% confidence level with a chi-squared test statistic of 503.0. This suggests that the share of younger riders tends to be higher at night.

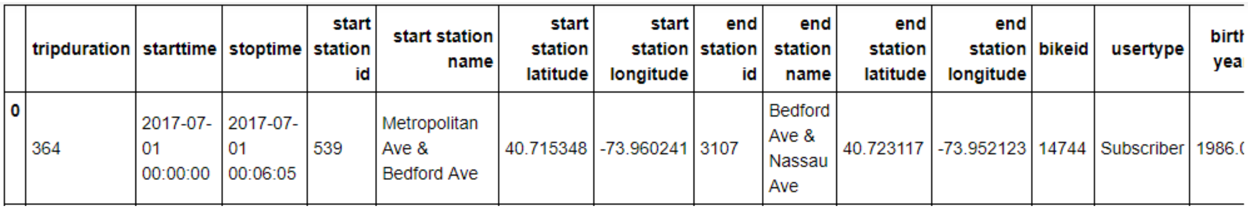
Introduction

New York City introduced the Citibike program in late 2013 as a public paid service enabling citizens to rent branded bikes for 30-minute ride sessions. Rich data exists for every historical ride in the system that includes the rider's starting and stopping station, age, gender, and duration. Like the New York City subway, the system is operational 24 hours a day, 365 days a year. I am interested in exploring whether there is a statistically significant difference in the age of Citibike riders in the daytime versus the nighttime. My intuition is that late-night riders will tend to be younger than daytime riders. This question may be a proxy for the effect of things like bedtime and sleep preferences, sense of safety late at night, use of Citibike for commuting vs. "going out", and the higher availability of alternative transportation methods for older bike riders.

Data

Citibike data is publicly available online and contains comprehensive ride records for one-month periods from July 2013 to present (nyc). I used the July 2017 Citibike data set, consisting of 1,735,599 rides. A ride record consists of the the following fields: trip duration, start time, stop start, start station ID, start station name, start station latitude, start station longitude, end station ID, end station name, end station latitude, end station longitude, bike ID, user type, user birth year, and user gender.

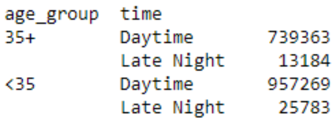
Using Python and the Pandas package, I created a data frame and reduced the fields to the hour and age of rider. For testing purposes, the time and age fields were divided into two groups. For ride times, they were the following: daytime rides, consisting of rides initiated between 5:00 AM and 11:59 P.M.; and nighttime



	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	birt yea
0	364	2017-07-01 00:00:00	2017-07-01 00:06:05	539	Metropolitan Ave & Bedford Ave	40.715348	-73.960241	3107	Bedford Ave & Nassau Ave	40.723117	-73.952123	14744	Subscriber	1986.(

Figure 1: Sample of a portion of one ride record from July 2017.

rides, consisting of rides initiated between 12:00 A.M. and 4:59 A.M. For age, the groups were the following: riders 35 years old or older ("older riders"); and riders younger than 35 ("younger riders").



```
age_group  time
35+        Daytime       739363
           Late Night     13184
<35        Daytime       957269
           Late Night     25783
```

Figure 2: Count of rides by age group and time group for July 2017

I used Matplotlib to create a histogram of ridership by age group and hour of day for July 2017(**Figure 3**). Please see caption below for more details on construction. Bars with red tops are hours with more young riders; bars with blue tops are hours with more older riders. Except for the early commuting hours of 5 A.M. - 8 A.M., younger riders outnumber older riders. That difference is especially apparent in the proportions for the early-morning hours (0-4, at left), though the overall number of trips is low.

In **Figure 4** , the effect of age by hour is more pronounced. Older riders outnumber younger riders for the bulk of the daytime, but that edge decreases dramatically by 8 P.M. From 9 P.M. to 4 A.M., younger riders slightly outnumber older riders.

Method

The test was performed using the chi-squared test of proportions, which suits the nature of the two proportions being compared in the null hypothesis and also aligns with the methodology introduced in the HW4 notebook reproducing the results from the recidivism study (rep). It is useful in cases of testing with a binary categorical variable (in this case, day vs. night) and two population groups (young vs. old) to compare the observed proportions to the expected proportions.

I tested the null hypothesis that the proportion of riders 35+ to total riders for trips starting at midnight-5 am is higher or equal to the proportion of riders 35+ to total riders for trips starting at 5 am - midnight. This null hypothesis and methodology received assistance from the Github review of HW3 by Hao Xi, who caught a typo in my original hypothesis formulation and added some additional clarity and rigor to my methodology (hw3).

I used Federica Bianco's function to determine the chi-squared test statistic for the above table, which is 1471.503. This statistic is evaluated against the test statistic of 3.84 for the 95% confidence level, and accordingly, the null hypothesis is rejected, suggesting that younger riders are are a greater proportion of late-night riders.

To validate these results against another month and season, I performed the same test on the February 2016 Citibike data, which yielded a chi-squared statistic of 502.98. The null hypothesis is again rejected at the 95% confidence level.

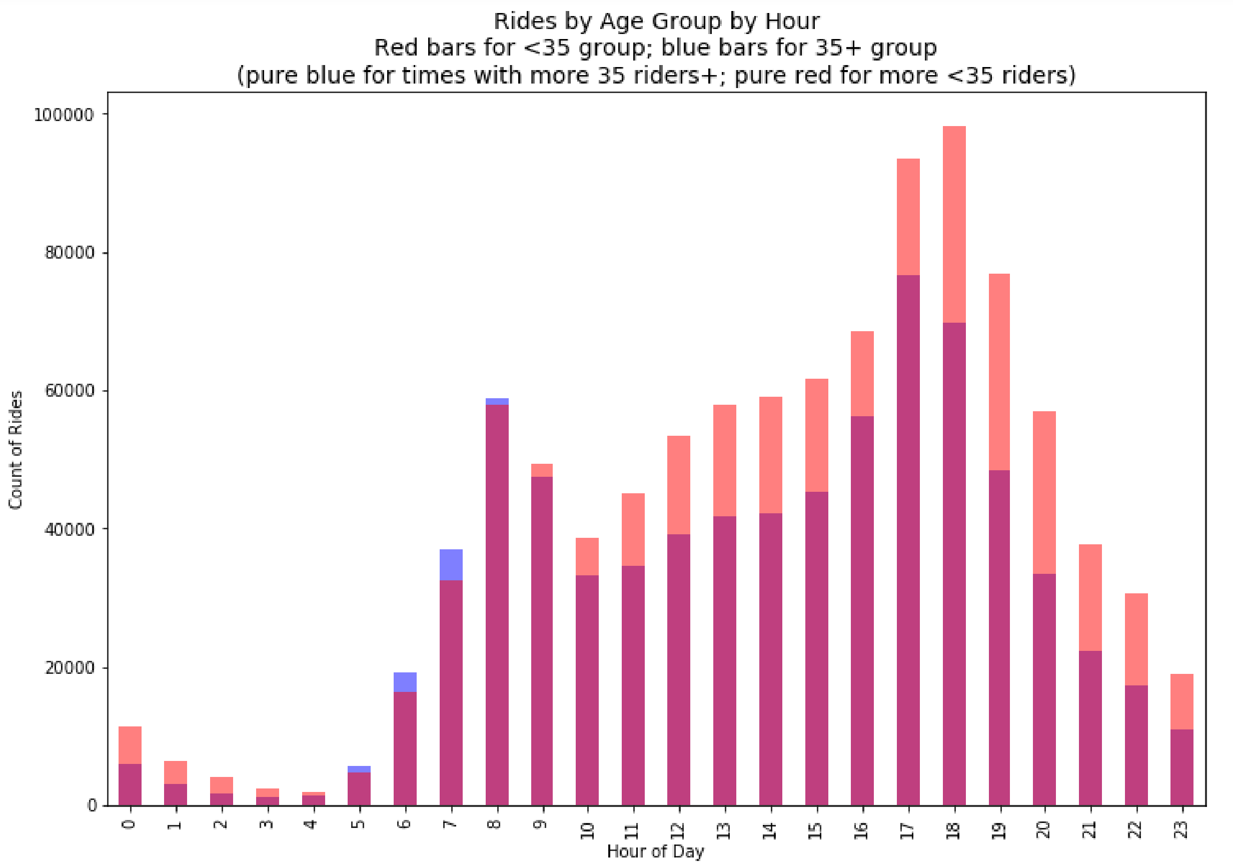


Figure 3: Histogram of rides by hour for July 2017. Age groups are overlaid on each x-axis hour. Blue represents older riders; red represents younger riders. Purple areas reflect the overlap of the two.

Please see Jupyter Notebook at end of this paper for the code used, additional figures, and the derivation of the test statistic.

Conclusion

The chi-squared test of proportions on Citibike data for July 2017 and February 2016 produced rejections of the null hypothesis at the 95% confidence level. This suggests that younger riders make a up a statistically significantly greater proportion of riders in late-night hours than older riders. Some limitations of the approach include the arbitrary division of “older” vs. “younger” riders based on age 35 and of daytime vs. nighttime based on midnight. A fuller analysis may use a Python package like Astral to use sunrise and sunset times to divide day and night more precisely. Additionally, as seen in **Figure 3**, the most interesting pattern in the data is that older riders emerge in strength in the early-morning commute hours between 5 A.M. and 8 A.M. I would like to explore that data further and to see whether the pattern holds as you divide the age brackets into decade bins rather than binary groups.

Jupyter Notebook

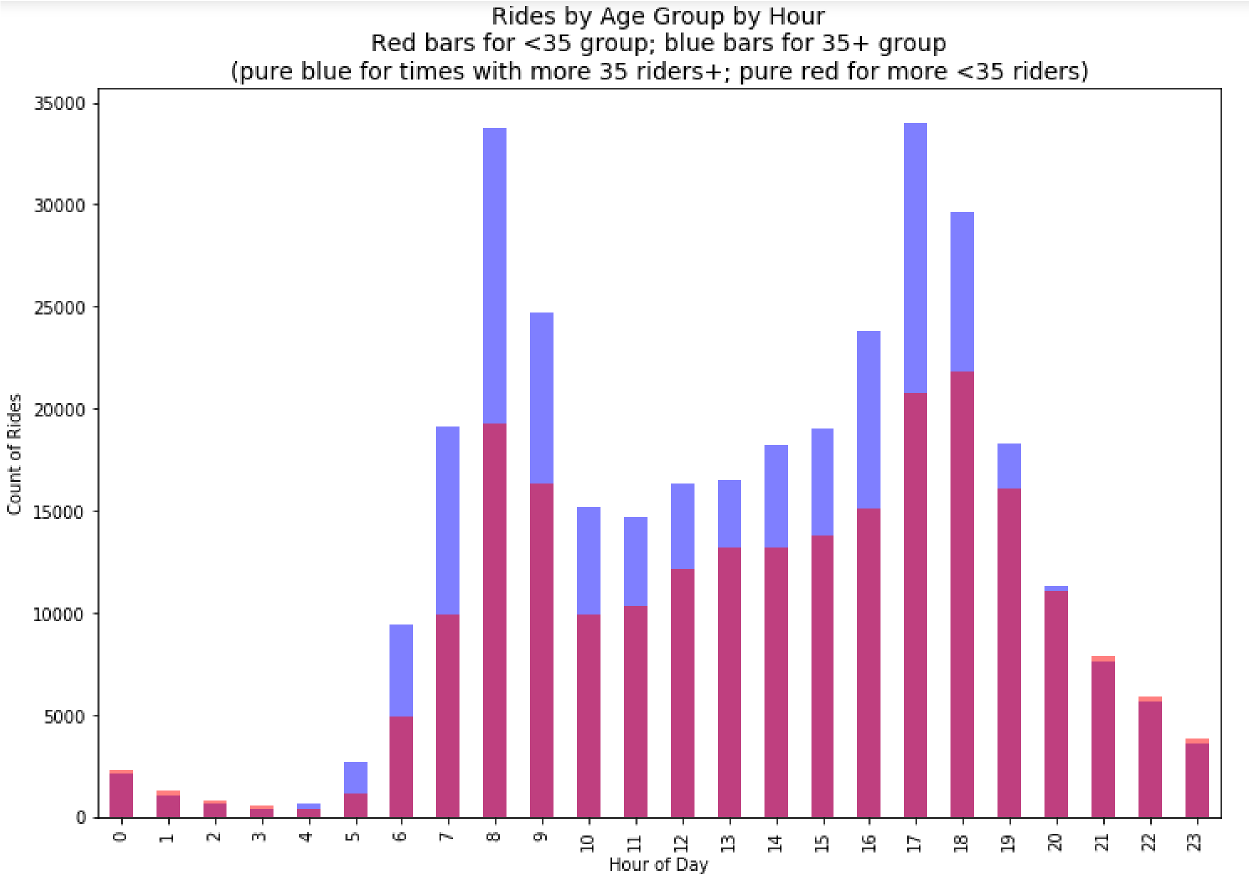
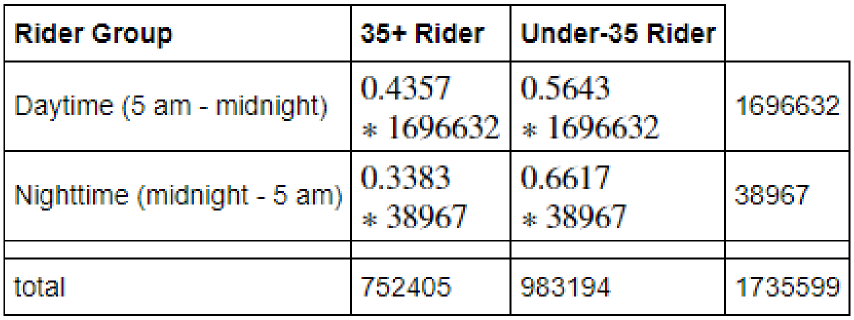


Figure 4: Histogram with same structure as above with ride data for February 2016.



Rider Group	35+ Rider	Under-35 Rider	
Daytime (5 am - midnight)	$0.4357 * 1696632$	$0.5643 * 1696632$	1696632
Nighttime (midnight - 5 am)	$0.3383 * 38967$	$0.6617 * 38967$	38967
total	752405	983194	1735599

Figure 5: Chi-squared frequency table for July 2017 data.

jupyter

References

Hao Xi Review of HW3. https://github.com/sgo230/PUI2017$_{s}go230/blob/master/HW3_{s}go230/CitibikeReview_{h}x517.md.URL$ Accessed on Mon, November 06, 2017.

Citi Bike System Data — Citi Bike NYC. https://www.citibikenyc.com/system-data. URL http://www.citibikenyc.com/system-data. Accessed on Mon, November 06, 2017.

Federica Biano - HW4 Repository. https://github.com/fedhere/PUI2017$_{f}b55/blob/master/HW4_{f}b55/effectivenesofNYC$ $PrisonEmploymentPrograms_{s}olution.ipynb.URL.$ Accessed on Mon, November 06, 2017.