PUI2018 hw8

Yusu Qian[1]

[1]New York University

November 7, 2018

Abstract: For convenience, many people are choosing citibike to commute. Using the dataset provided by citibike, I confirmed that male riders have a longer average trip duration than female riders. Github link https://github.com/sueqian6/PUI2018_yq729/blob/master/HW8_yq729/assignment2_yq729.ipynb

Introduction: With the growing popularity of citibikes as a means of transportation, datasets on citibikes can give us a better idea of how people are travelling. In my study, I made an assumption that the average trip duration of male riders is longer than female riders. Then I used the dataset to confirm my assumption.

Data: I used the dataset downloaded here: https://www.citibikenyc.com/system-data. I chose entries of trips in August 2018. And then I selected only columns I need for this study, which are tripduration and gender. I noticed that other than 1 and 2 which represent male and female respectively, there was also 0, whose meaning I was not sure of, so I kept it in the plot too.

Methodology: I grouped the data by gender to calculate the average trip duration of each gender then drew a plot to show the difference. In addition, I calculated the standard deviation and drew a plot too. I followed the suggestion given in the peer review to do a Kruskal-Wallis test, which did not reject the null hypothesis.

Conclusion: It is understandable that female riders have a shorter trip duration in average. What surprised was the significance between trip durations of different genders. The trip duration of male riders is on average three times of female riders.

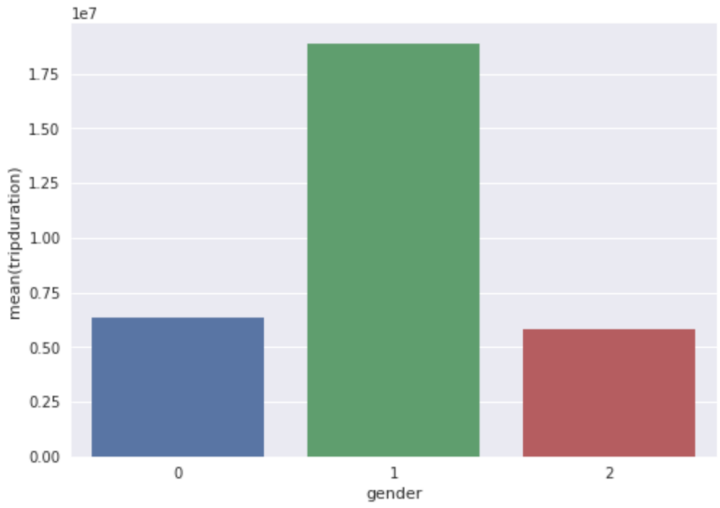

Figure 1: average trip duration of different gender

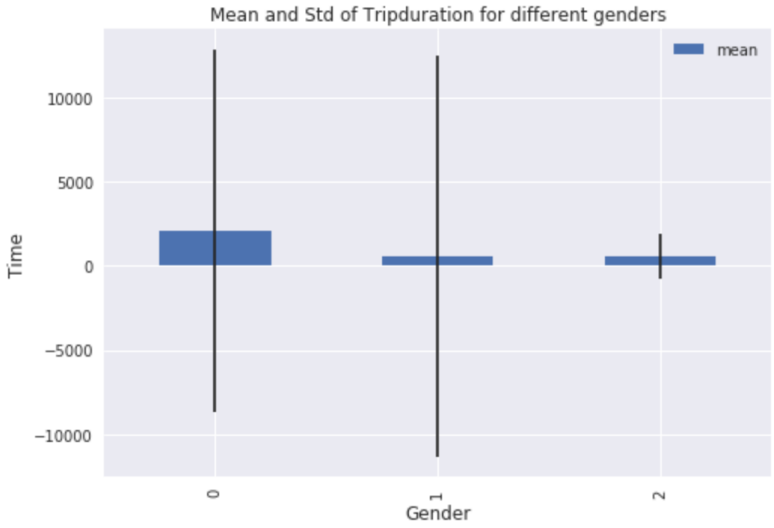

Figure 2: standard deviation of trip duration of different gender

```
In [16]: #i am following the suggestion given in the peer review to do a krustal wallis test
         x=df[df.gender==2].tripduration
         y=df[df.gender==1].tripduration

In [17]: from scipy import stats
         stats.kruskal(x, y)

Out[17]: KruskalResult(statistic=133.92698240471162, pvalue=5.6682232237788009e-31)
```

the p value is smaller than 0.05 so the null hypothesis is not rejected

Figure 3: krustal wallis test