Citi Bike mini project

ssb602[1]

[1]Affiliation not available

November 9, 2017

Abstract: As a part of the citi bike mini project, I aim to look at the citi bike usage for weekdays and weekends. After completing the analysis and using statistical tests to verify my results, I have concluded that the citibike usage is more for weekends as compared to weekdays.

Introduction:

Citi Bike is widely used in New York City. There is a lot of data collected for citi bikes. The data has many attributes, trip duration, start station, end station, type of users to mention a few. While there are many interesting trends that can be observed in this data, I have chosen to look at the average citi bike usage for weekdays and weekends. The usage will be gauged by using the trip duration parameter. I have assumed that trip duration is higher when the number of trips are more.The citi bike usage over weekdays and weekends can help to understand the traffic patterns. The trip duration total for weekdays or weekends will be more due to more number of users as well as the traffic.

Citi Bike data about is studied before for various purposes. I came across two papers where citi bike data is used for different analysis. One paper talks about the usage of citi bikes on weekdays and weekends by hour of the day. According to this study, the citi bike usage is more on weekdays.(sys) The another study talks about usage in summer and winter months. Their results show that usage during summer is more as people prefer riding bikes in summer than in winter.(Divya Singhvi, 2015) In the mini project, I am looking at usage of citi bike as well but with a little different perspective. Following are the details of data, methodology and tests used for the same.

Data Used:

Data used is citibike trip data for July 2017. A view of the data used for the analysis is shown in the figure below.

The analysis is finding the citibike usage for weekdays and weekends. But there is no such column that has the day of the week. The day of the week therefore is extracted from the start time column and a column named dayofweek is added to the existing dataframe. Data cleaning is performed to get rid of unwanted columns. The only columns that are retianed and which are relevant for analysis are tripduration and day of week.

Methodology:

I have started by defining my null and alternate hypothesis which are as follows:

1. Null and Alternate hypothesis

The null hypothesis for citi bike usage is:

	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
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
1	2142	2017-07-01 00:00:03	2017-07-01 00:35:46	293	Lafayette St & E 8 St	40.730207	-73.991026	3425	2 Ave & E 104 St	40.789210	-73.943708	19587	Subscriber
2	328	2017-07-01 00:00:08	2017-07-01 00:05:37	3242	Schermerhorn St & Court St	40.691029	-73.991834	3397	Court St & Nelson St	40.676395	-73.998699	27937	Subscriber

Figure 1: Citi Bike Data for July 2017

H0 : Average trip duration during weekends is same or less than weekdays

H0: $(\text{Avg. Trip Duration})_{\text{weekends}} <= (\text{ Avg. Trip Duration })_{\text{weekdays}}$

The alternate hypothesis is:

H1: Average trip duration during weekdays is more than weekends

H1: $(\text{ Avg. Trip Duration })_{\text{weekends}} > (\text{ Avg. Trip Duration })_{\text{weekdays}}$

The trip duration by weekday is as follows:

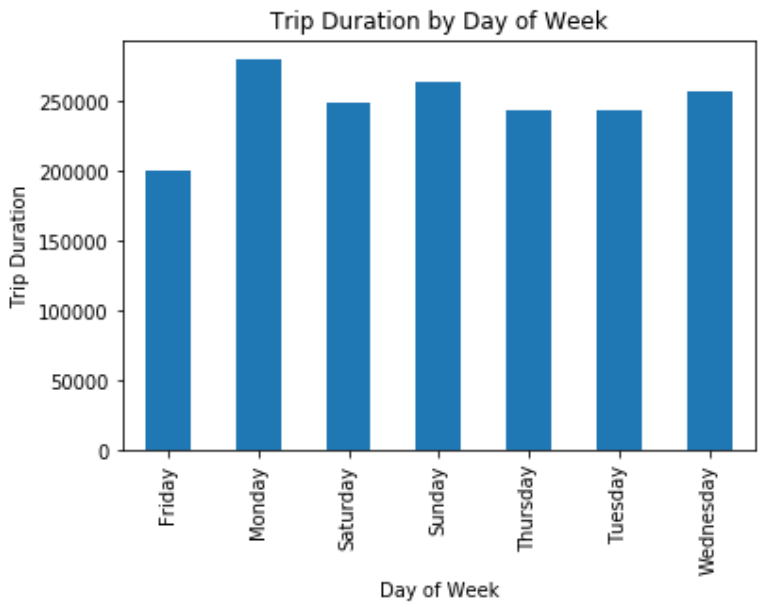

Figure 2: Trip Duration by day of week for the month of July 2017

Before proceeding with the test, I checked if average trip duration for weekends is more than weekdays. The code for the same is:

```
In [48]: weekend_trip_avg = weekend_trips.mean()
         weekday_trip_avg = weekday_trips.mean()

         ## Check if Avg_trip_duration(weekend) - Avg_trip_duration(weekday) > 0
         if(weekend_trip_avg - weekday_trip_avg) < 0:
             ##Null cannot be rejected
             print("Null hypothesis cannot be rejected")
         else:
             print("Further test need to be performed to test the hypothesis")

         Further test need to be performed to test the hypothesis
```

Figure 3: Check if further test is needed to reject the null hypothesis

Now that it is confirmed that further tests are needed to reject the null hypothesis, I had to chose a test that would be useful to compare averages as the hypothesis is tested for average trip duration. As per suggestion from Professor on my previous analysis, I have decided to use Mann Whitney U test and Moods Median Test. The tests and their results are discussed in the next part.

2. Test used and Test results:

i) Mann Whitney U Test:

Mann Whitney U Test can be used to check whether two samples from a population have a similar distribution. It is a non parametric test i.e. it does not make any assumption about the distribution of data at test. (wik)

For the mini project I have used the `scipy.stats` package. The result of Mann Whitney U test is as follows:

1. Mann Whitney U test

```
In [15]: result_mann_whitney = mannwhitneyu(weekday_trips,weekend_trips)

In [16]: result_mann_whitney

Out[16]: MannwhitneyuResult(statistic=276047462645.5, pvalue=0.0)
```

Figure 4: Mann Whitney U Test Result

ii) Moods Median Test:

Moods Median Test is used to compare medians of two or more populations. This is a non-parametric test as well.

For the mini project I have used the `scipy.stats` package. The result of Moods Median test is as follows:

2.Moods median test

```
In [19]: result_moods_median = median_test(weekday_trips,weekend_trips)

In [20]: result_moods_median

Out[20]: (11279.407917891802, 0.0, 676.0, array([[579360, 288033],
                [643759, 224447]]))
```

Figure 5: Moods Median Test Result

Looking at the test results, for both the tests p-value is 0.0 i.e less than chosen significance level of 0.05. Therefore Null hypothesis can be rejected with a significance of 0.05

Conclusion:

Traffic is a common issue in cities. The reason behind understanding citi bike usage is that it can be used to understand traffic patterns. May be analyzing trip duration during hour of the day can help in understanding the peak traffic hours and help in applying right measures to solve traffic issues. If average trip duration is more but number of trips in less then some preliminary conclusions can be drawn about traffic.

The tests have concluded that average trip duration for weekends is higher than weekdays according to results of Mann Whitney U and Moood's Median Test. The tests were chosen as per suggestions given to the previous work done on this data. The citi bike usage can be said to be more on weekends. But this is under the assumption that higher the trip duration higher the usage. There can be better way of analyzing this data by considering number of trips and trip duration both. Also considering data for more than one month can bring different results.

References

A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System. http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/. URL http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system. Accessed on Thu, November 09, 2017.

Mann–Whitney U test - Wikipedia. https://en.wikipedia.org/wiki/Mann URL https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test. Accessed on Thu, November 09, 2017.

Peter I. Frazier Shane G. Henderson Eoin O' Mahony David B. Shmoys Dawn B. Woodard Divya Singhvi, Somya Singhvi. Predicting Bike Usage for New York City's Bike Sharing System. 2015. URL https://people.orie.cornell.edu/woodard/SingSingFraz15.pdf. Accessed on Thu, November 09, 2017.