New York University

Citi Bike Usage and Weather in New York City

Xinyi Gong (xg555)

Lanyu Shang (ls3882)

Zihao Wang (zw1074)

Instructor: Juliana Freire

GitHub:

https://github.com/ShangLanyu/BigDataProject.git

Google Drive:
https://docs.google.com/a/nyu.edu/document/d/1tvIsw28NGjpESbex4KmSrXyIeZ_g8-3NaLnTF9N5IPw/edit?usp=sharing

May 13, 2016

Table of Contents

Abstract

With the heavier traffic in Manhattan, bikes are the alternatives for most people. Instead riding directly from outside Manhattan to Manhattan, people are more likely to choose to use Citi bikes, with which they can use motor vehicles to enter the Manhattan and move freely in the crowded traffic. As bikes do not have any shelters, the weather may play a key role in the usage of Citi bikes. This brings us the hypothesis that the usage of Citi bikes will decrease when the weather goes worse. Then we have gathered data from online sources of both Citi bikes and weather in 2015 to conduct the analysis for big data. We have applied multiple MapReduce functions to filter or edit dimension of data. Besides only testing the relation between weather and usage, we also added the dimension such as age and gender, in order to check if the results would vary for these groups. We have found out some interesting facts based on our results and the outside references. Finally, on one hand, the patterns we found partially agree on our hypothesis as the number of trips decreases when the weather turns worse. On the other hand, the duration of trips increases when the weather gets better, which negates our hypothesis.

Introduction

1. Business Understanding

According to the statistics from Department of Motor Vehicles of New York State, there are 2,107,321 registered vehicles in 2015 in New York City, the vehicle types including standard,

commercial, bus, taxi and ambulance (DMV, 2016). However, the road and transportation system are limited by space and time. That is, there is unlikely to conduct and finish any transportation-related projects in few months. This situation burdens the traffic in New York City so that the traffic jams occur very often. People and industrial have started to think of an alternative way to move around in the city freely. Then comes the Citi Bike. The parent company of Citi Bike's operator released a new statistic on May 5th, 2016, claiming that "the bike share service saw a 110% jump in rides in the first quarter of the year compared to the same period in 2015" (Pereira, 2016). This rapid increasing in rides raises our interests on exploring Citi bike. Additionally, when we think of bikes, the first idea comes to our mind is the weather. As bikes do not have rain shelter, air-conditioner or windshield comparing to motor vehicles, we are curious if there is any connection between the usage of Citi bikes and weather. We also bring the hypothesis of our project that the usage of Citi bikes is lower when the weather condition is worse.

2. Data Understanding

a. Weather Data

The weather data describing the weather condition of New York City contains 33 features and 13340 data entries from Jan 1st, 2015 to Dec 31st, 2015. Since there are too many missing values in some features, we will use the following features that having less missing values:

Feature Name	Type	Description
YR--MODAHRMN	int	E.g. datetime in GMT
SPD	int	Wind Speed
TEMP	int	Temperature in Fahrenheit
PCP01	float	1-hr liquid precipitation in inches and hundredths
PCP06	float	6-hr liquid precipitation in inches and hundredths
PCP24	float	24-hr liquid precipitation in inches and hundredths
SC	int	Snow depth in inches

b. Citi Bike Data

The monthly Citi bike trips data we acquired from its official website contains the following features:

Feature Name	Type	Description
Trip Duration	int	In seconds
Start Time and Date	datetime	e.g. “1/1/2015 0:01”.
Stop Time and Date	datetime	e.g. “1/1/2015 0:01”.
Start Station ID	int	An integer indicating station’s id number.
Start Station Name	string	Name of cross streets.
Start Station Latitude	float	Latitude

Start Station Longitude	float	Longitude
End Station ID	int	An integer indicating station's id number.
End Station Name	string	Name of cross streets.
End Station Latitude	float	Latitude
End Station Longitude	float	Longitude
Bike ID	int	An integer indicating bike's unique id number.
User Type	string	"customer" = 24-hour pass or 7-day pass user; "Subscriber" = Annual Member
Year of Birth	int	Year in 4 digits.
Gender	int	0=unknown; 1=male; 2=female

Data Preparation

1. Feature Definition

1.1 "Weather Index" Definition

A number of characteristics may define a weather condition to be good or not, such as temperature, wind speed, and liquid precipitation. To simplify our research and analysis, we generate an index from existing features to describe if a weather is "good" or "bad". First, we select *wind speed* and *hourly liquid precipitation* as the key features to make contribution to the index. When the *1-hour liquid precipitation* is not available, the average liquid precipitation of *6-hour liquid precipitation* and *24-hour liquid precipitation* will be used as a

substitution. Although temperature is a necessary feature in describing weather condition, it varies in different seasons. Thus, we do not take temperature as a component in the index. Next, we use min-max scale method to rescale *wind speed* and *hourly liquid precipitation* in range [0,1] and sum them together to be the index (i.e. index $\in$ [0,2]).

1.2 Age Group Definition

In our analysis, we divide target subscribers into three groups: young group (Group 1: age less than 25), middle-aged group (Group 2: age between 25 and 50) and elderly group (Group 3: older than 50).

2. MapReduce

2.1 Reduce Weather Data

As the weather index we defined above, we apply the MapReduce technique to achieve this variable from the existing weather data. First, we mapped each weather record with key *(YearMonDayHr)* and value *(Temperature, WindSpeed, PCP, SnowDepth)*. In the meantime, we would replace the PCP variable with 0 if all PCPs are missing. In the reducing part, we collected all records in the same hour of the same day and output their average value on temperature and snow depth, and the index we calculated from their average wind speed and PCP. Finally, the size of our output data was reduced to 223 KB and contained 8760 records.

2.2 Combine Weather Data and Citi bike Records

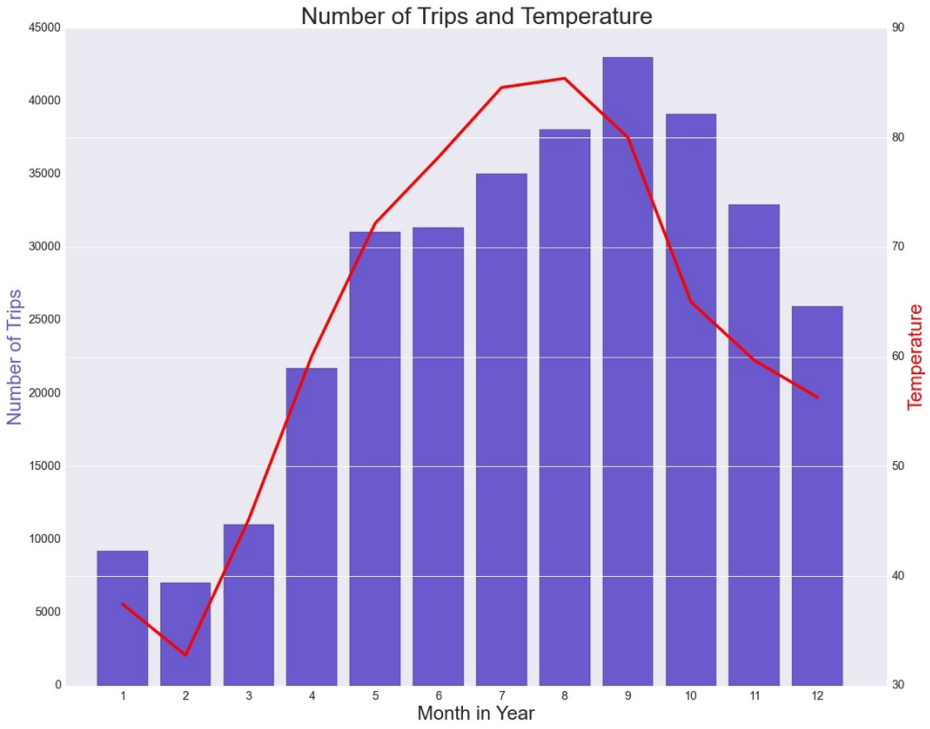
In order to get the useful information from the indexed weather data and Citi bike data conveniently, we designed a MapReduce method to combine these two datasets and most of our methods used in the analysis below were derived from this step. The main idea is to use the string composed of year, month, week (0-6), day and hour as the key. In the weather data, the value will start with digit "1" followed by variables we need in the analysis. Similarly, in the Citi bike data, the value will begin with digit "0" followed by variables we need in the analysis. Finally, we achieved several desired datasets containing the information we need for the analysis by implementing the method above.

Visualization and Insights

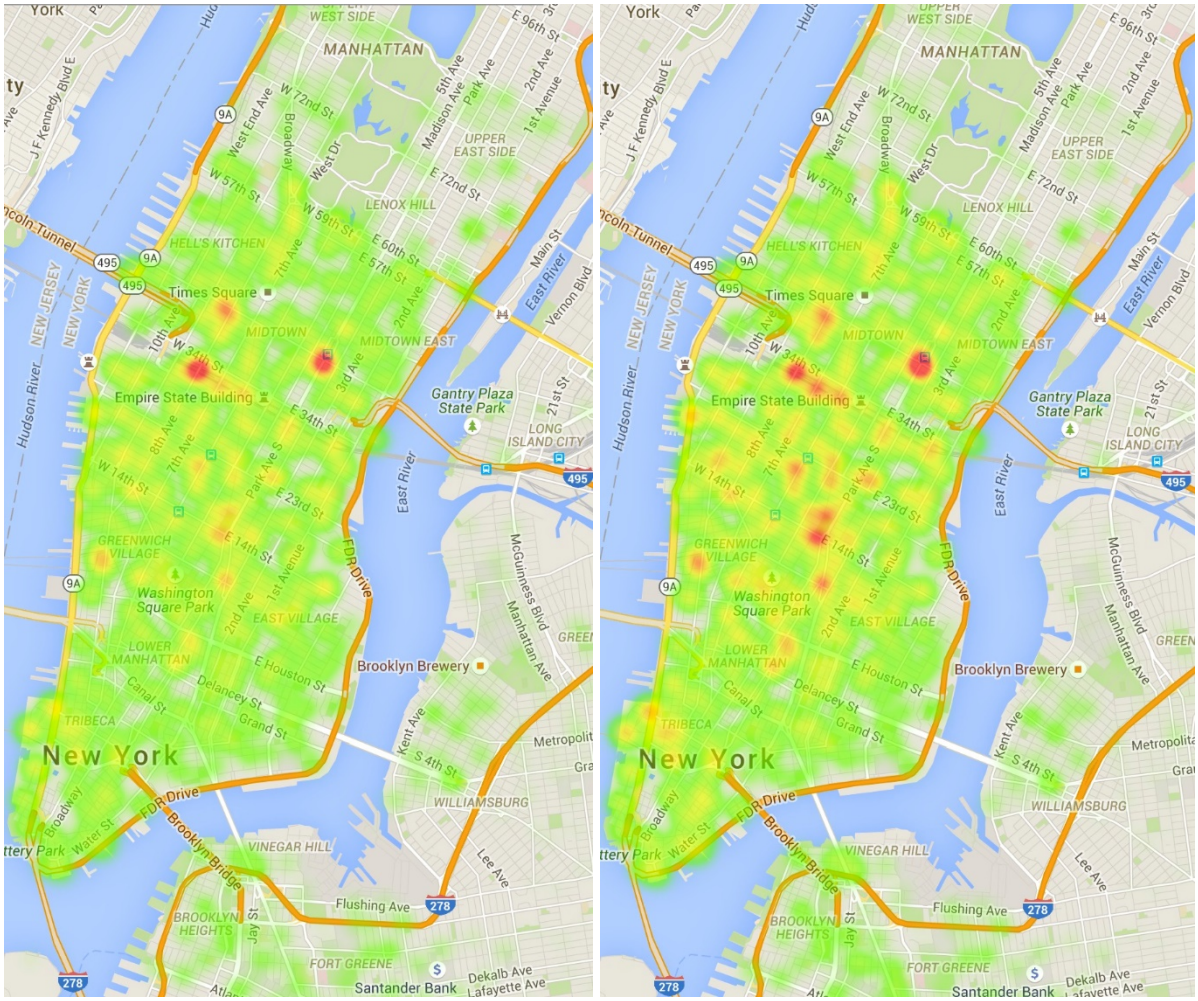
1. Number of Trips vs. Temperature

First of all, we tried to plot the relation between the number of trips and the average of maximum temperature of each month.

Given the plot as below, there seems when the temperature goes up, the usage of Citi bikes will also increase. However, the maximum temperature has no information about whether good or bad the weather it is. Thus, in order to excavate the deeper insight, we decided to apply the weather index on our data.



As the heat map of pickup locations (4-8pm) plotted by different weather index (see below), there is a significant change on popular locations. A possible reason is that, during bad weather period, it is harder to get a Taxi in the city during peak hours and Citi bike may be used more frequent than usual. Thus, we are convinced that weather index is a descriptive characteristic of weather condition for Citi bikers.



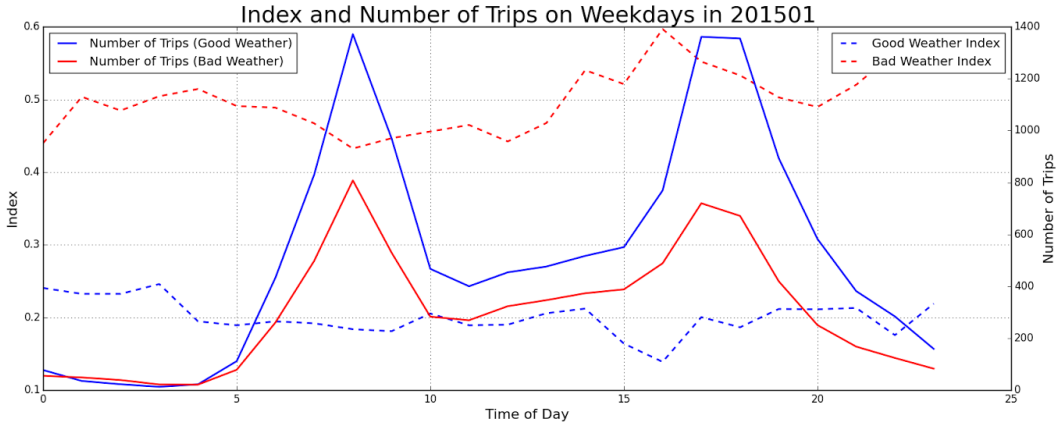
Left: Heat map of Pickup Location (Good Weather Index)

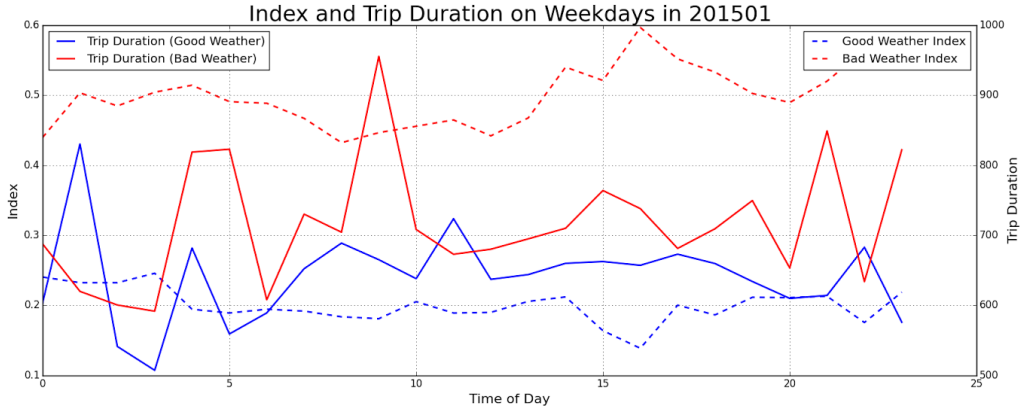
Right: Heat map of Pickup Location (Bad Weather Index)

2. Trip Information vs. Weather on Weekdays

With the assumption we made before, as the trip duration and number of trips will decrease when the weather is bad, we are going to investigate the trip information and weather first. As we have given an index that defines the levels of weather condition, we sorted the index of each day in January, 2015 and set the top 30% as good weather indexes and the bottom 30% as bad weather indexes. Then we picked the weekdays (Monday through Thursday) for analysis because during weekends there would be a more flexible pattern, meanwhile the

pattern during weekdays will be more regular and fixed. The following two plots indicate the relation among weather indexes, time of day and respectively trips duration and number of trips for all weekdays in January, 2015. Also, we used all average values for these plots.



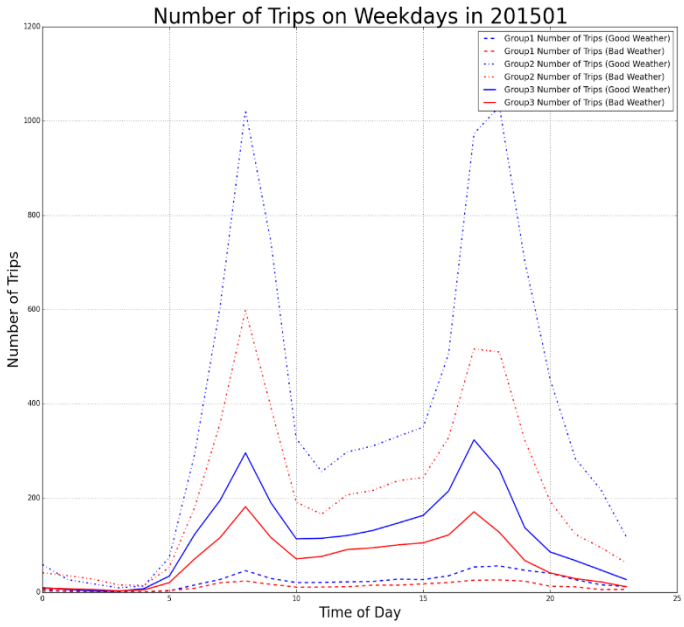


From these two plots, we can see that the trend of number of trips agrees on our assumption as the blue solid line is almost always above the red solid line. However, the trend of trip duration surprised us because the duration was longer when the weather is in bad condition. Moreover, from the plot for number of trips, there are two peaks during one day. The two time periods are around 7 to 8 and around 17 to 18. One reasonable explanation

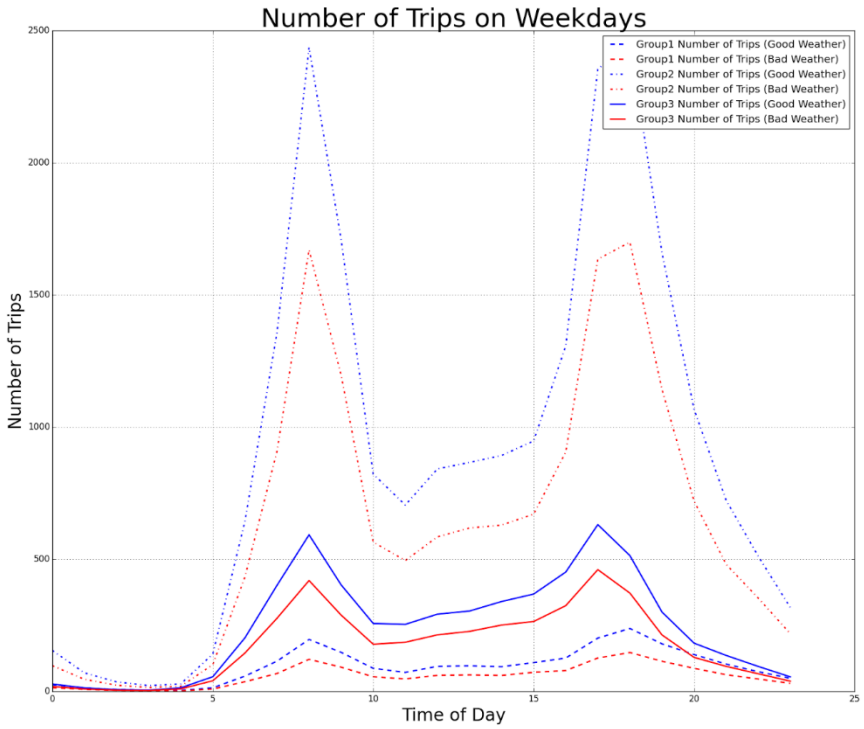
would be these two are rush hours and there are lots of heavy traffic in Manhattan and crowded in the public transportation, so riding a Citi bike will be a good alternative to move around Manhattan. For the trip duration, we did not expect that the duration would go up with the bad weather. As we compared all the data on weekdays, we can assume that the daily routines of riders are mostly the same, which means that they go to the same destinations but with longer time. The possible reasons would be either they ride slower due to the poor weather condition or they find some place as the shelter maybe for rain or snow. Next, we will add more dimension to the data and try to explore more insights under this interesting pattern.

3. Number of Trips vs. Age Group under Different Weather Condition

Now we added the dimension of age for subscribers. We wanted to test if different age groups would react differently on weather index. As mentioned before, we grouped the riders by their ages, and then we made some plots based on different months and whole year and tried to get some insights of the plots. We picked the same month as the previous part for analyzing first.



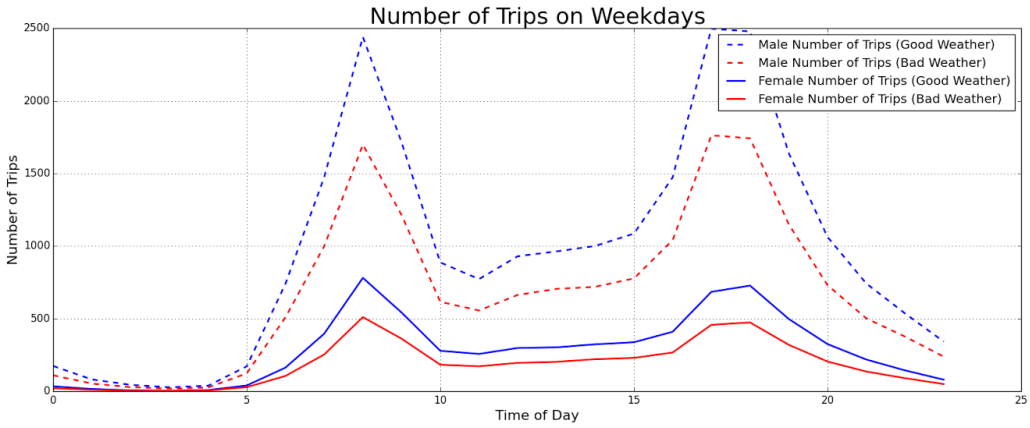
As all the information based on different groups are plotted under the same weather indexes as before, the only difference is instead of plotting regarding all riders, we have three age groups now. From this plot, obviously, people in group 2 use Citi bikes more often than people in group 3, and people in group 3 use Citi bikes more often than people in group 1. As for the age between 25 and 50, people are the mainly workforce in Manhattan, during weekdays, people in group 2 are more likely to use the Citi bikes. Additionally, for the lowest number of trips of group 1, the possible explanation is that under 25, more people are students in school, and they do not need to ride Citi bikes that often. Some of them will have school bus for the people before undergraduate, and for undergraduate or graduate students, they might live in dormitory or close to the campus. Furthermore, the trend of the number of trips regarding on weather index agrees with the previous result: there are more trips in good weather.

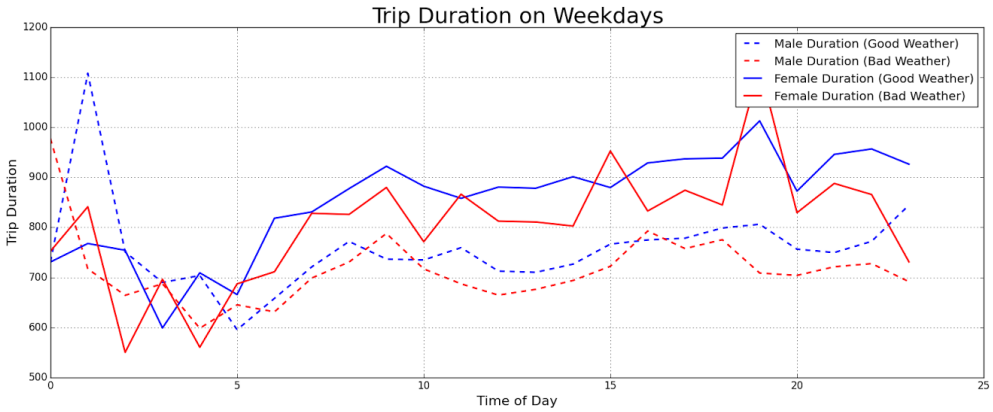


Just for curious, we want to test whether the trend of 3 groups would be different for different period during the year, so we made the plot above. It turns out that the proportion of three age group is consistent throughout the whole year. Also, it matches the results we found before that the two-time period should be the rush hours because the age group of workforce contributes most to those periods.

4. Trips vs. Gender under Different Weather Condition

Next, we took a step back and add another dimension to the data -- gender, in order to check if the pattern we found before toward the duration of time and number of trips would vary from male to female.



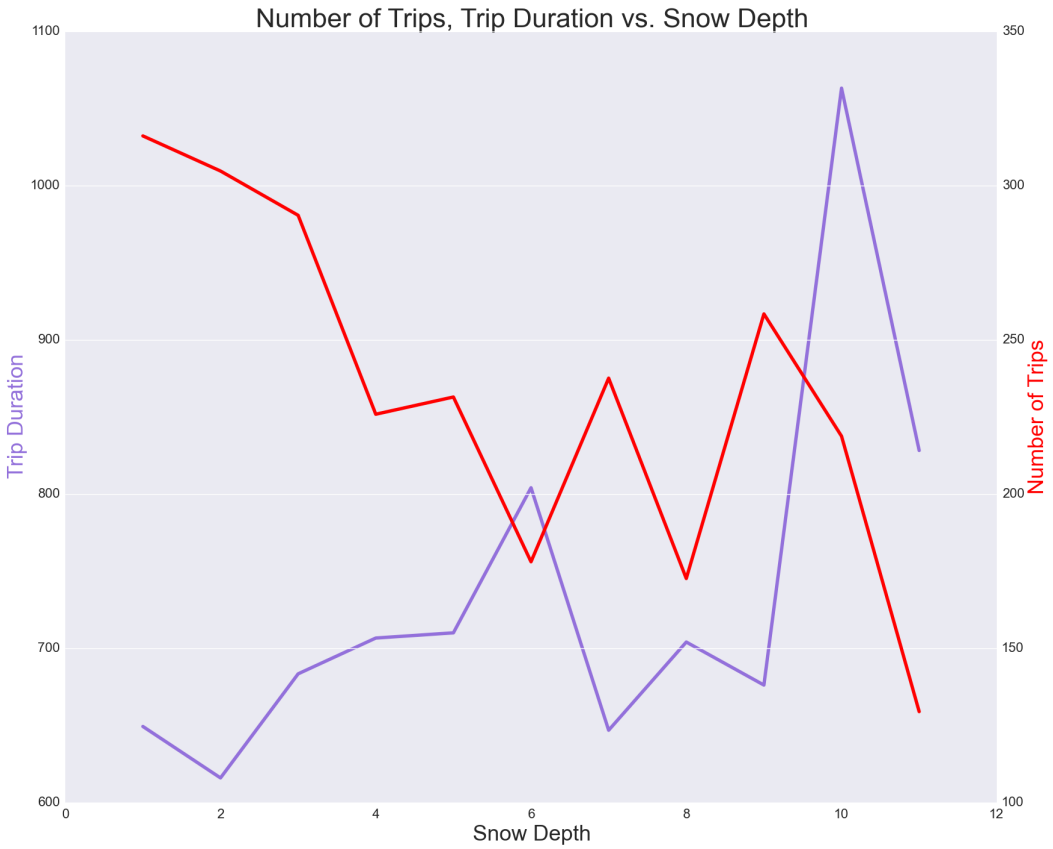


For the plots regarding number of trips on weekdays, males use Citi bike more often than females. *New York Times* (2015) once conducted an interview few females that rides Citi bike, one of them said that "(a)s a female bicyclist, you not only have to deal with the usual hollering from men on the street, but you also feel much more vulnerable to the incredibly aggressive traffic culture — where your body is constantly at peril" and other opinion stated that "mostly male bicyclists" are "tend to break more rules more and put you in harm's way". One also saw a couple biking during when "the man run the red light and the woman sit at it awkwardly, not sure if she should chase after him and put her life at risk". We think the less female Citi bike users is not only the problem for Citi bike, but also for the bike riders. Women are more worried about the safety both mentally and physically when they ride a bike

on the streets in Manhattan, where the traffic is crowded as well as fast-paced. Then we are going to discuss the plots of trip durations. Both the female and male trend of number of Citi bike match the trend for all riders -- the duration in bad weather are longer than that in good weather. There are some interesting patterns in this plot: males contribute most of the trips during 0 to 5 o'clock, and the duration of females are longer than that of males. This matches the results of the analysis for the number of trips, i.e. women care more about their safety so that females rarely use the Citi bikes during midnight and ride much slower than males.

5. Trip Information vs. Snow Depth

In our last step of the project analysis, we focused on the weather condition in winter season. Snowing in a common but serious weather condition especially for bikers. According to the result we plotted below, we discovered that, other than the severe storm weather, trip duration and number of trips are negative correlated. As the depth of snow increasing, number of bike trips will decrease while the trip duration increase. Winter is the off-peak season for Citi bikes as shown in our previous analysis. Therefore, it will a good time for maintenances and recall of Citi bikes while keeping the condition of existing Citi bike to guarantee customer experiences.



Conclusion and Future Works

The results described above are the insights we gain while analyzing the weather data and Citi bike trip records in New York City. Total number of bike trips is highly correlated with temperature in such time period. Also, the daily average number of trips is various depending on time and traffic. Among all Citi bike subscribers, we have found typical patterns on Citi bike usage for different gender and age groups which can be further investigated for marketing purpose.

For future analysis, we may combine other datasets that describe the transportation and population in the city, such as the traffic statistics and tourism information. Moreover, we may also adopt the same dataset but in earlier years together with the most current data to discover more patterns.

Appendix

1. Data Source:

a. Weather Data: https://nyu.box.com/s/4lkrxs9rdsfjzpu1gh9nwen89jxtc9dd

b. Citi Bike Data: https://www.citibikenyc.com/system-data

2. MapReduce:

All MapReduce works are done on NYU HPC.

a. Task 1: Weather Index

   Number of Reducer: 3

   Cluster Elapsed Time: 9 sec

b. Task 2: Number of Trips, Duration and Index

   Number of Reducer: 3

   Cluster Elapsed Time: 38 sec

c. Task 3: Number of Trips, Duration and Index (Membership Only)

   Number of Reducer: 3

   Cluster Elapsed Time: 30 sec

d. Task 4: Number of Trips, Duration, Index and Gender (Monthly)

   Number of Reducer: 3

   Cluster Elapsed Time: 39 sec

e. Task 5: Number of Trips, Duration, Index and Gender (in Year)

   Number of Reducer: 1

Cluster Elapsed Time: 54 sec

f. Task 6: Number of Trips, Duration, Index and Age (Monthly)

   Number of Reducer: 3

   Cluster Elapsed Time: 43 sec

g. Task 7: Number of Trips, Duration, Index and Age (in Year)

   Number of Reducer: 1

   Cluster Elapsed Time: 1 min 4 sec

h. Task 8: Extract Temperature

   Number of Reducer: 3

   Cluster Elapsed Time: 9 sec

i. Task 9: Number of Trips and Temperature (in Year)

   Number of Reducer: 1

   Cluster Elapsed Time: 55 sec

j. Task 10: Number of Trips and Maximum Temperature (Monthly)

   Number of Reducer: 3

   Cluster Elapsed Time: 45 sec

k. Task 11: Pickup Location and Index (in Year)

   Number of Reducer: 1

   Cluster Elapsed Time: 52 sec

3. GitHub Repository:

https://github.com/ShangLanyu/BigDataProject.git

4. **Google Drive:**

https://docs.google.com/a/nyu.edu/document/d/19xqsxH35lSwBupA8KhmPDOWreSm90yKsoJe5ajutT4o/edit?usp=sharing

5. **Contribution:**

Each member in the group contributed equally in this project:

Xinyi Gong (xg555): Background research, MapReduce task 4-6 and report write up.

Lanyu Shang (ls3882): MapReduce task 1-3, plotting result and report write up

Zihao Wang (zw1074): MapReduce task 7-11, plotting result and report write up.

References

Department of Motor Vehicle New York. (2016). *Vehicle Registrations in Force (2015)*. Retrieved May 11, 2016, from https://dmv.ny.gov/statistic/2015reginforce-web.pdf

Ivan, P. (2016, May 5). Citi Bike ridership up 110% in New York City. *AmNEWYORK*. Retrieved May 11, 2016, from http://www.amny.com/transit/citi-bike-ridership-up-110-in-new-york-city-1.11766306

Times Readers React to Citi Bike's Gender Gap. (2015, July 10). New York Times. Retrieved May 12, 2016, from http://www.nytimes.com/2015/07/11/nyregion/times-readers-react-to-citi-bikes-gender-gap.html?_r=0