Understanding the Complex Interactions in New York City Taxi and Citi Bike

Shixin Li (sl3368)
Yichen Fan (yf511)
Zewei Liu (zl1446)

ABSTRACT

Based on deep studying and analyzing in behavior of Citi Bike riders and taxi passengers, give recommendation for both Citi Bike company and taxi drivers maximum their profit and limit cost with specific time interval and geographic region. Give suggestions to people to decrease waiting time as well.

INTRODUCTION

Business Understanding

The proposes of this paper is to study the triping behavior of Citi Bike rider and taxi passengers based on 2015 datasets. Further more, finding out the relationship between taxi and Citi Bike. After that, helping taxi drivers maximize their profit by giving them the locations and times that people are more willing to take taxis. At the same time, giving both taxi drivers and bike riders information to minimize their waiting time. Furthermore, we could make suggestions to Citi Bike Company to increase their revenue and limit cost by modifing bike stations and rental price accordingly.
Apart from the basic analysis on our data, we also tried to find insights in our data too. For example, we analyzed why only few people in a certain neiborhood use Citi Bike or Taxi.

Overview of Dataset & Understanding

The Table below is the Overview of Variables in Citi Bike Data

Variable	Type	Description
Trip Duration	Numeric	Time in second takes each trip
Start Time and Date	Numeric	Pick up time and date
Stop Time and Date	Numeric	Drop off time and Date
Start Station Name	Numeric	Pick up Station
End Station Name	Numeric	Drop off Station
Station ID	Categorical	Station ID Number
Station Lat/Long	Numeric	Station Location
Bike ID	Categorical	
User Type	Categorical	(Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)
Gender	Categorical	(Zero=unknown; 1=male; 2=female)

Year of Birth	Numeric	Bike Rider's Year of Birth

The Table below is the Overview of Variables in Citi Bike Data

Variable	Type	Description
VendorID	Categorical	A code indicating the TPEP provider that provided the record.
tpep_pickup_datetime	Categorical	The date and time when the meter was engaged.
tpep_dropoff_datetime	Categorical	The date and time when the meter was disengaged.
Passenger_count	Numeric	The number of passengers in the vehicle.
Trip_distance	Numeric	The elapsed trip distance in miles reported by the taximeter.
Pickup_longitude	Numeric	Longitude where the meter was engaged.
Pickup_latitude	Numeric	Latitude where the meter was engaged.
RateCodeID	Categorical	The final rate code in effect at the end of the trip
Store_and_fwd_flag	Categorical	Y= store and forward trip N= not a store and forward trip
Dropoff_longitude	Numeric	Longitude where the meter was disengaged.
Dropoff_latitude	Numeric	Latitude where the meter was disengaged.
Payment_type	Categorical	A numeric code signifying how the passenger paid for the trip.
Fare_amount	Numeric	The time-and-distance fare calculated by the meter

Extra	Numeric	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges
MTA_tax	Numeric	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	Numeric	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Numeric	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Numeric	Total amount of all tolls paid in trip.
Total_amount	Numeric	The total amount charged to passengers. Does not include cash tips.

Data Preparation

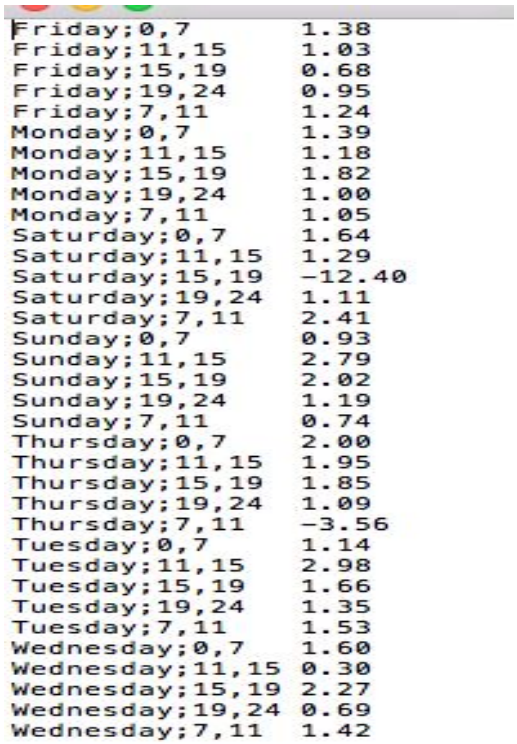
In this project, we mainly used two datasets, Year 2015's Citi Bike data and Year 2015's Yellow Cab Data. By using these two datasets, we were trying to find out relations between Citi Bike and Yellow Cab in given time interval and region. Firstly, we analyed the popularity of Citi Bike and Yellow Cab separately. In order to comprehensively analyze these two datasets, we assigned each data in the datasets into different months, days, five time intervals, and days of week. Secondly, we analyzed them separately by counting the number of trips. Below, we will talk about our analysis in detail. Besides that, in order to make our analysis or recommendation more precisely, we also involved in some weather data, for example maximum and minimum temperature. We were trying to find out as many factors as possible that affect people's choices between Citi Bike and Taxi.

The main method we used to prepare our data is MapReduce. For each pair of map and reduce, we have different key-value pair. For example, if we want to split the Citi Bike dataset up into different months, the map function will extract months as keys, and set 1s as values to present that each line of our dataset is a trip. In the reduce phase, the reduce function reads in each line from the output generated from map function and checks the month. Then the reduce function will sum the 1s together if they share the same month.

We used Amazon Web Service to run our MapReduce framework. We created a Hadoop streaming cluster on AWS and chose 1 reducer to carry out the jobs because we would like all the data in just one file to make it convenient for us to rank them by number of trips.

The size of the Citi Bike dataset is 1.7 GB, so we only used 1 master and the run time is 3 mins in average. The size of the yellow cab dataset is 21.3 GB, so in order to run it faster, we used 1 master and 9 cores. Thus, the run time is 15 mins in average.

When running our MapRedue to get the average fare per mile for Yellow Cab data, the result file generated from AWS was not accurate in our common sense. Below Figure 1 is the screen shot of our result from AWS.



```
Friday;0,7        1.38
Friday;11,15      1.03
Friday;15,19      0.68
Friday;19,24      0.95
Friday;7,11       1.24
Monday;0,7        1.39
Monday;11,15      1.18
Monday;15,19      1.82
Monday;19,24      1.00
Monday;7,11       1.05
Saturday;0,7      1.64
Saturday;11,15    1.29
Saturday;15,19    -12.40
Saturday;19,24    1.11
Saturday;7,11     2.41
Sunday;0,7        0.93
Sunday;11,15      2.79
Sunday;15,19      2.02
Sunday;19,24      1.19
Sunday;7,11       0.74
Thursday;0,7      2.00
Thursday;11,15    1.95
Thursday;15,19    1.85
Thursday;19,24    1.09
Thursday;7,11     -3.56
Tuesday;0,7       1.14
Tuesday;11,15     2.98
Tuesday;15,19     1.66
Tuesday;19,24     1.35
Tuesday;7,11      1.53
Wednesday;0,7     1.60
Wednesday;11,15   0.30
Wednesday;15,19   2.27
Wednesday;19,24   0.69
Wednesday;7,11    1.42
```

Figure 1 AWS defualt results

From above, we can see a few odd numbers: negative and very small numbers are all strange. Since we cannot have negative cost, we decided to change our map function to fix this problem. The way we revised our function is to first restrict the range of the fare amounts and distance, since we only wanted to use positive fare amounts and distances. Second, we dropped NaN and inf as well. After revising our map function, the result looks much more reasonable.

Results and Discussion

- Citi Bike 2015 Pick Up Times v.s. Month

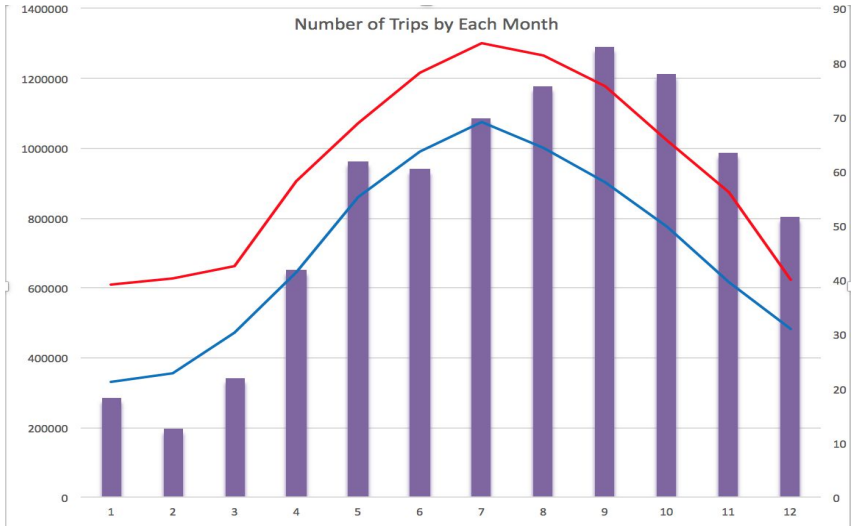


Figure 2 Number of Citi Bike trips by each month with maximum and minimum average temperature

Figure 2 exhibits number of trips by each month and the red line and blue line are maximum and minimum average temperature through each month. In September, we see a huge increase and in February a harsh drop in number of trips which has high correlation with temperature pattern. During winter, people would like to choose alternative ways like subway or taxi owing to lower temperature *(average temperature in February 2015 was 24F)* and poor weather conditions (snowing and windy). Contrarily, people may tend to ride Citi bike in summer and fall especially in September *(average temperature in February 2015 was 75F)*. Therefore, Citi Bike pick up times are relatively high during May to November. Therefore, number of trips is largely affected by temperature. Based on this result, we would encourage Citi preparing more bikes from May to November and provide maintain service and less bikes from Janurary to March. So that, bikes are used more efficiently and biks' useful life may be largely extended.

- Citi Bike 2015 number trip vs day of week

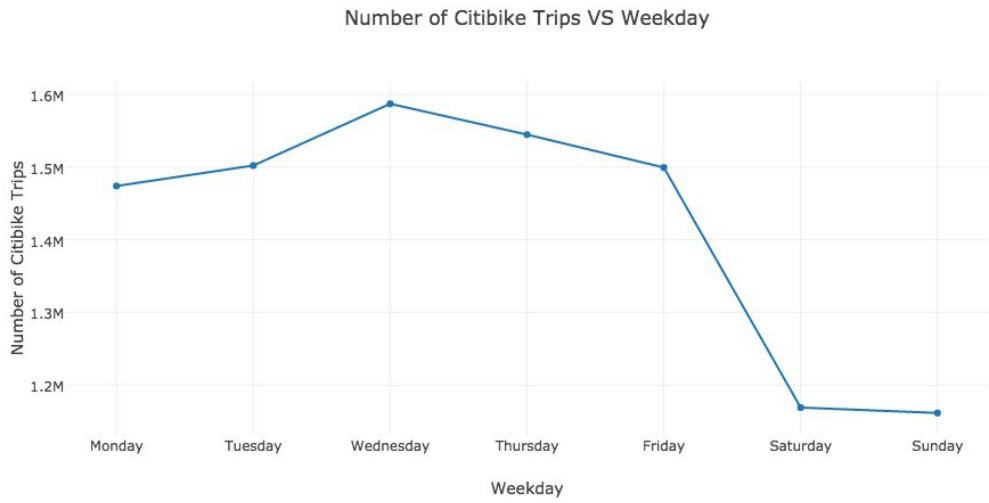


Figure 3 Number of Citi Bike trips by day of week

As we can see from the above figure 3, it is quite reasonable that week days have significantly higher Citi bike usage than weekends. Therefore, people riding Citi bike mainly for going to work or school rather than for leisure. As a result, Citi could build more stations around office buildings and schools but less around parks. Besides that, providng customized services like mudguard to prevent riders' business suits get stained would also benefit Citi bike riders. On the other hand, Citi could also pursue effors on advertising benefits of bicycle sports and family activities to remain fairly constant usage during weekends. For instance, two-seated-bikes and bicycle racing event.

- Citi Bike Five Time Intervals

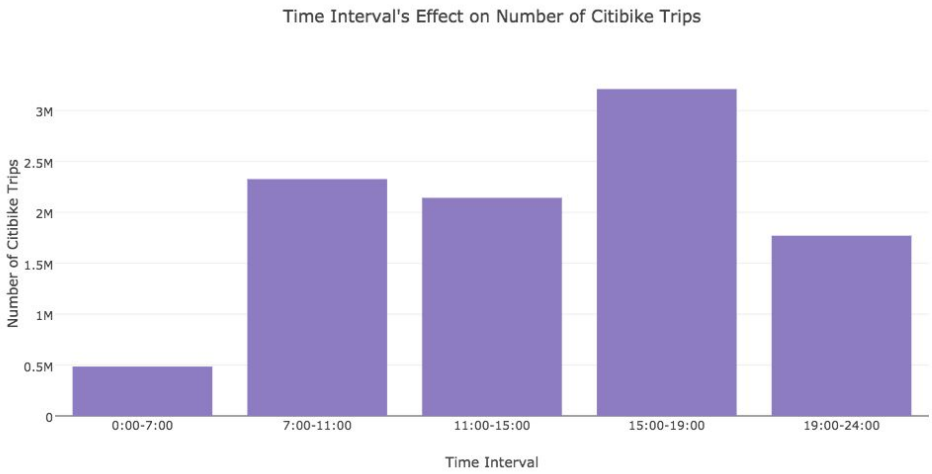


Figure 4 Number of Citi Bike trips by five time intervals

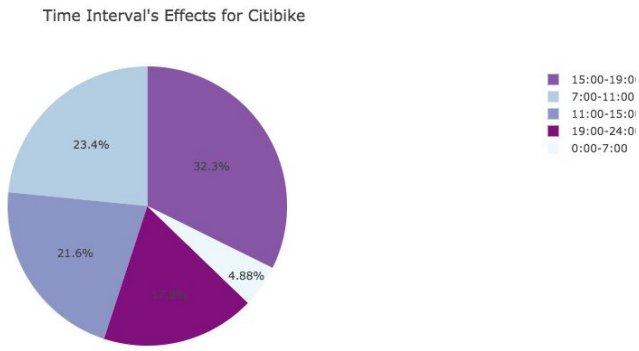


Figure 5 Number of Citi Bike trips by five time intervals for weekdays

In order to figure out when do people take Citi Bike in each day of week. We divided each day (24 hours) into five intervals, 00:00am~07:00am, 07:00am~11:00am, 11:00am~15:00pm, 15:00pm~19:00pm and 19:00:pm~24:00pm.
We found out that, for weekdays, only 23% Citi Bike were picked up in early morning but 32.3% of Citi Bike were picked up in afternoon (15:00 to 19:00). People prefer riding Citi Bike from riding bicycle home than going to work or school in morning in weekdays. People might take either taxi or subway going to work and school. We will discuss it in following part.

- Citi Bike 2015 Pick Ups vs Each Day

We wanted to figure out at what date people like riding Citi Bike the most and the least. We sorted pick up times by day and picked top 20 days and last 20 days to see if there is any relavance. The Table 1&2 shows Top and Last 20 Citi Bike pick up days.

Top 10 Days	Month/Day	Day of Week	Temperature ℉	Precipitation Inch	Snow Inch	Holiday
1.	9/24	Thursday	71	0.00	0.00	None
2.	9/25	Friday	70	0.00	0.00	Native American Day
3.	9/16	Wednesday	77	0.00	0.00	Stepfamily day
4.	9/17	Thursday	79	0.00	0.00	Citizenship Day
5.	9/18	Friday	78	0.00	0.00	None

6.	10/8	Thursday	66	0.00	0.00	None
7.	10/7	Wednesday	66	0.00	0.00	None
8.	10/14	Wednesday	63	0.00	0.00	None
9.	10/6	Friday	63	0.00	0.00	None
10	10/22	Thursday	67	0.00	0.00	None

Table 1 Top Twenty Citi Bike Trip Days

People who like to ride Citi Bikes the most during week days, especially Thursday (4 out of 10) with good weather conditions (neither rain nor snow and comfortable temperature). Mean temperature of top ten days was 70℉. September (5 out of 10) and October (5 out of 10) are months with more pick up times. Besides that, unlike our expectation, whether it is a holiday is not an effective factor for Citi Bike riding since we didn't see holidays in top ten pick up days.

Last 10 Days	Month/Day	Day of Week	Temperature ℉	Precipitation Inch	Snow Inch	Holiday
1.	3/28	Saturday	33	0.00	0.00	None
2.	1/27	Tuesday	24	0.36	0.00	None
3.	2/2	Monday	30	1.02	3.30	None
4.	3/5	Thursday	31	0.76	7.50	None
5.	1/18	Sunday	35	2.10	0.00	None
6.	2/15	Sunday	19	0.00	0.00	None
7.	3/1	Sunday	25	0.00	0.00	None
8.	2/21	Saturday	20	0.60	3.00	None
9.	3/14	Saturday	42	0.81	0.00	None
10	2/26	Thursday	27	0.00	0.00	None

Table 2 Last Twenty Citi Bike Trip Days

People who like to ride Citi Bikes least in weekends (6 out of 10) with poor weather conditions (rainy or snowy and chilling temperature). Mean temperature of last ten days was 28.6℉. Additionaly, whether it is a holiday is not an effective factor since we didn't see holidays in last ten pick up days as well.

- Citi Bike 2015 Number of Trips vs Station

Figure 7 Top Ten Citi Bike Trip Stations

The above map with orange labels represent ten most popular Citi Bike Stations. Citi Bike could build more stations around these top ten popular stations to ensure every Citi Bike rider can at least get one bike in peak hours. In cluding Penn Station, Grand Central, Astor Plaza and Spring Street. On the other hand, there are two main reasons may cause these ten stations popular. Either because there are only few stations around these regions or because people who around these regions ride bikes a lot.

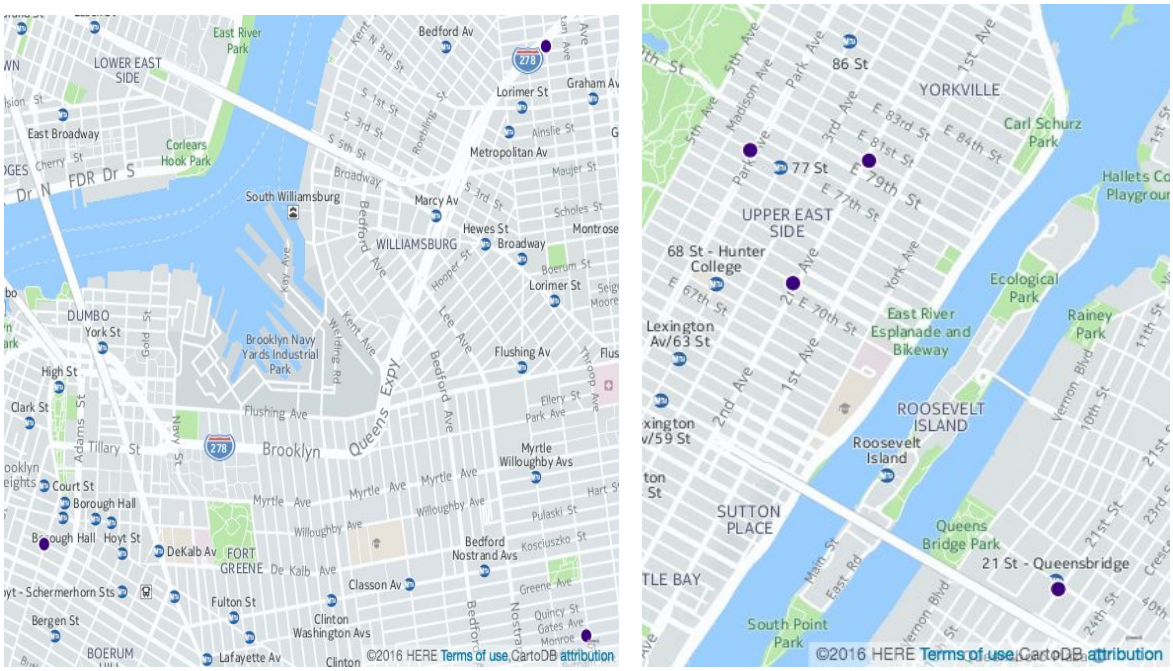


Figure 8 Last Eight Citi Bike Trip Stations

The above map with purple labels represent ten last popular Citi Bike Stations, including Upper East Side, Brooklyn and Long Island City. Citi Bike could somehow cut last ten popular Citi Bike Stations or less bikes in order to limit bike depreciation cost. Otherwise, Citi could move these bikes to bike shortage stations. So that, all bicycles are used more efficiently. On the other hand, there are two main reasons may cause these ten stations least popular. Either because there are a lot of stations around these regions or because people who around these regions ride bikes seldomly. We will figure it out in the following part.

- Citi Bike 2015 Station Distribution

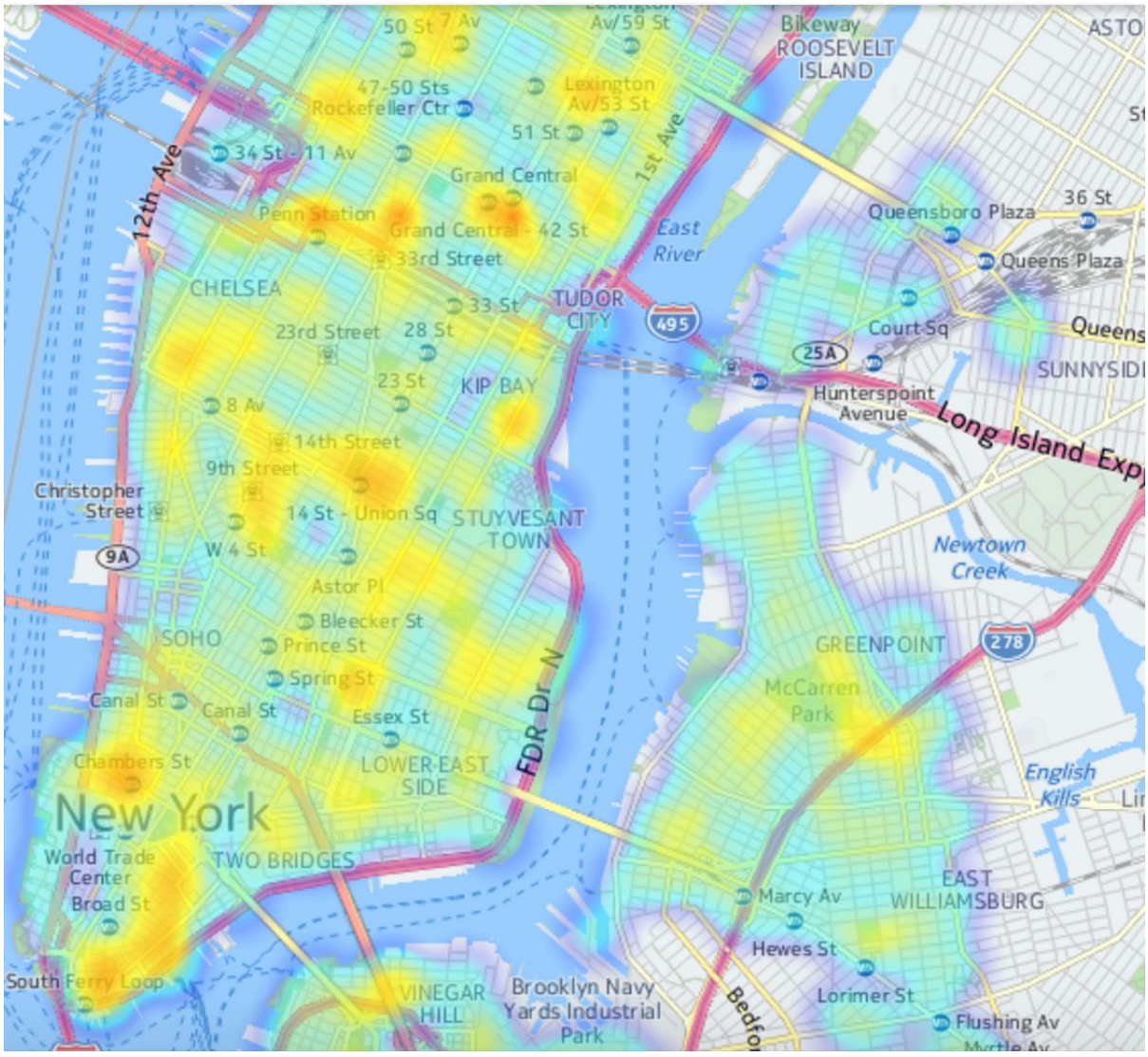


Figure 9 Citi Bike Station Heatmap

The aboved figure 9 shows the distribution of Citi Bike Stations in New York City. The region with more Citi Bike Stations, map around that region gets darker red color. The area with less Citi Bike Stations, map around that area gets lighter red color and turns blue.

Comparing to the top ten popular Citi Bike Stations, Grand Central, Penn Station, 14th Street Union Square, Spring Street and Chambers Street regions have both dense and popular Citi Bike Stations. We would recommend Citi Bike building more stations around these stations to ensure no shortage of Citi Bike in peak hours. Others, 23rd street, East Village (8th Avenue & 14th Street) and Astro Plaza has less amount of Citi Bike Stations. For last ten popular stations, same as we expected, there are less Citi Bike Stations around these area. Citi Bike could somehow cut these stations to limit cost.

- Taxi 2015 Number of Trip vs Month

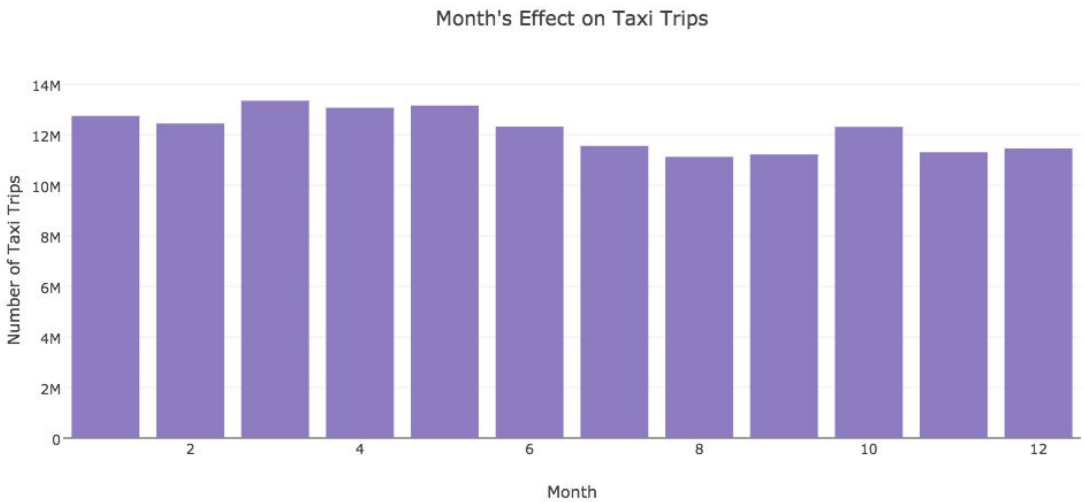


Figure 10 Taxi Trips by Month

Unlike Citi Bike, there is no significant difference between each month for taxi pick ups. Taxi trips are less relavent to temperature but only a few more trips in winter than summer. October has a little bit higher trips among fall. It pobably because October is one of tourist rush seasons.

- Taxi 2015 Average Fare per Mile vs Month

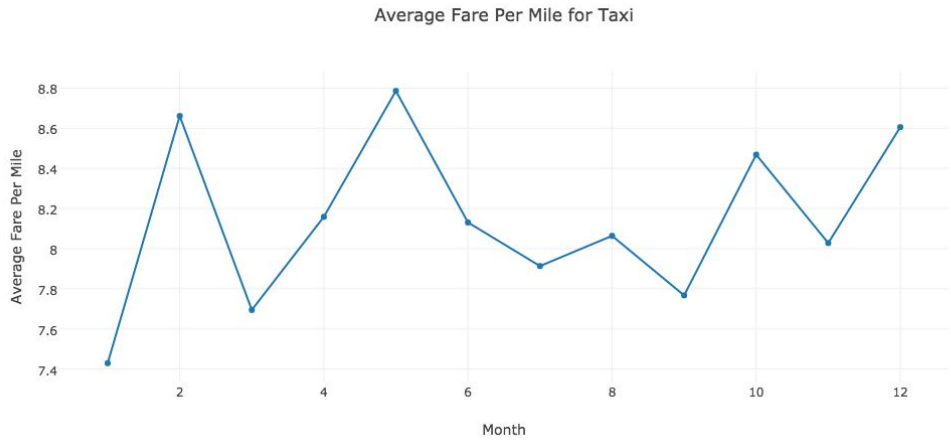


Figure 11 Taxi Fare per Mile by Month

The above plot shows average taxi fare per mile. Fare per mile is a high in Feburary, May, October and December which could illustrate that time spend on trip is longer than other months. Combined with previous figure, October and May has high amount of trips means that traffic is crowded and large amount of taxi are on the road. As for Feburary and December, number of taxi trips is relative low which means there are more private cars than taxis.

- Taxi 2015 Number trip csday of week

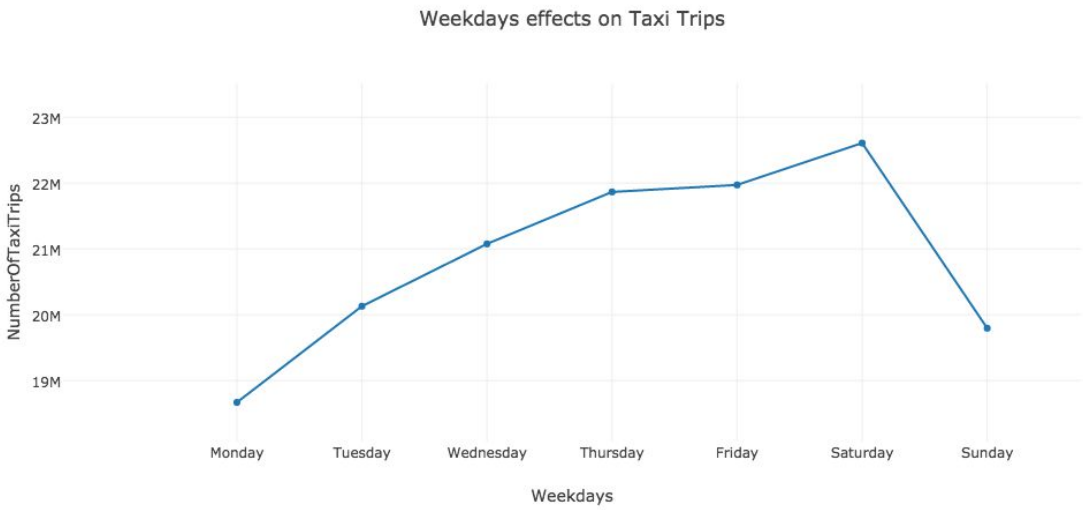


Figure 12 Number of Trips of Taxi by Day of Week

As we can see in the figure 12, Saturday has significantly higher trips of Taxi and increasing trend among week days. Monday has lowest number of trips. Same as we expected, people do not take Taxi as their main way going to work or school but Citi Bike or Subway alternatively. Sunday has a relative low positon of number of trips since people would like to stay at home to take a rest.

- Taxi number of time intervals for week days

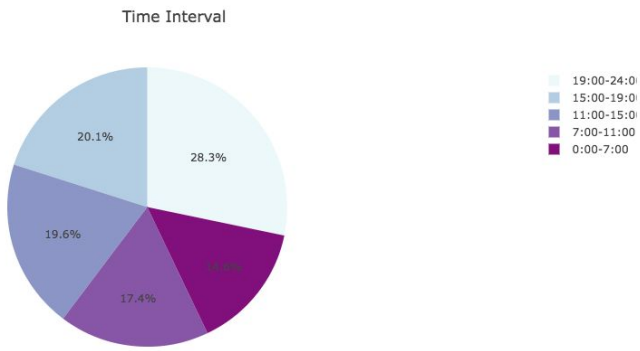


Figure 13 Taxi Trip by Five Time Intervals

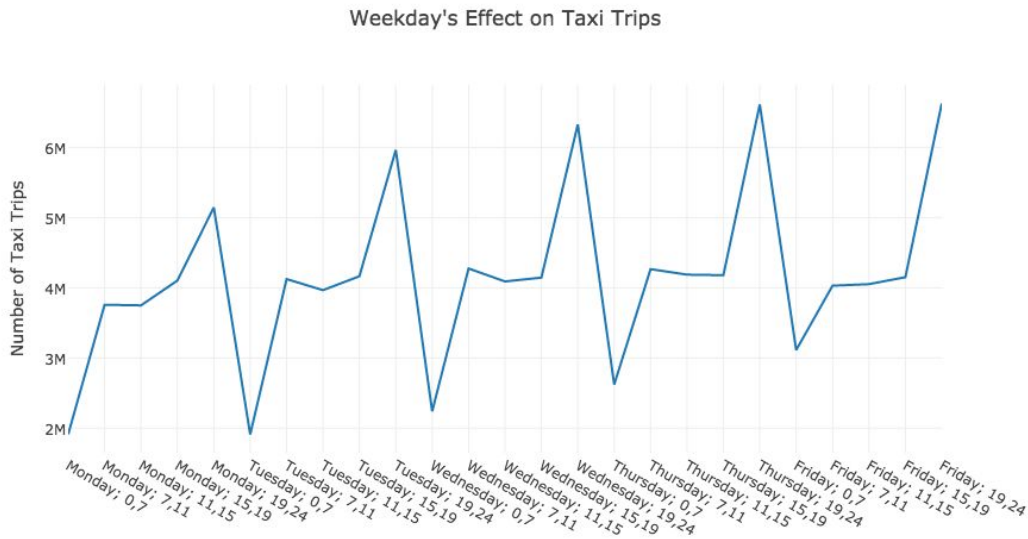


Figure 14 Taxi Trip by Five Time Intervals for Weekdays

It is a quite obvious pattern for taxi during weekdays. Different from Citi Bike, taxi trip has strong cyclical pattern.

People take taxi during late night (19:00pm~24:00pm) the most. During weekdays, late night (19:00pm~24:00pm) has highest trips and keep increasing from Monday to Friday. And for early morning, (00:00pm~7:00am), has lowest trips but also rising around weekdays. Other times, like morning (7:00am ~11:00am) and afternoon (11:00am~15:00pm) are quite stable. Maybe after hang out and party, taxi is a main way going back home. It also illustrated that, people do not take Taxi as their main method for work or school. Additionally, as we

mentioned previously, Citi Bike only have 23% of pick up trips in early morning. Therefore, we can say that, people are more likely going to work and school by subway.

- Taxi number of time intervals for weekends

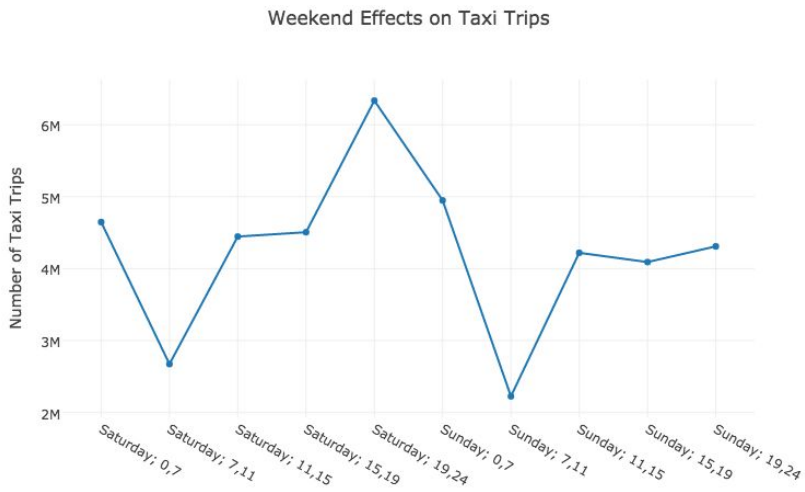


Figure 15 Taxi Trip by Five Time Intervals for Weekends

Different from week days, people take taxi a lot during midnight (24:00pm~7:00am) in weekends. Maybe after work overtime or hangout, instead of Citi Bike, taxi is a main way going back home. Secondly, morning (7:00am ~11:00am) has lowest trips in weekends, it is easy to understand that, people prefer sleep late and stay at home. Thirdly, (19:00pm~24:00pm) on Saturday has high trips but low on Sunday. It might because next day is weekday, people intend to stay at home and take a rest.

- Regions

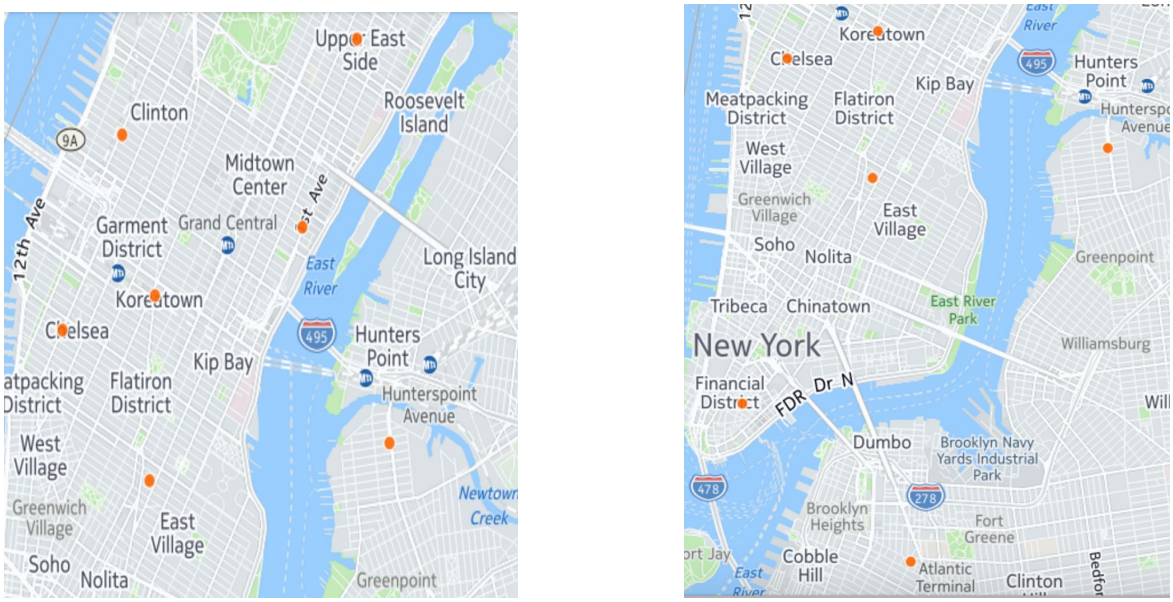


Figure 15 Ten Manhattan regions

We wanted to figure out where people like to take taxi the most and the least. Therefore, we divided pick up address into ten characteristical regions and sorted by number of trips for whole year of 2015.

	Place	Number of Trips	District Type
1	East Village	53948	Living
2	Brooklyn	49957	Living
3	Financial District	46063	Commercial
4	Chelsea	26433	Leisure and Entertainment
5	Clinton	21196	Living
6	Turtle Bay	18287	Commercial
7	Korean Town	14908	Leisure and Entertainment
8	Long Island City	6688	Living
9	Lincoln Square	1637	Leisure and Entertainment
10	Upper East Side	883	Living

Table 3 Ten Manhattan Regions with Number of Taxi Trips for 2015

It is not surprising that East Village has highest number of trips come with Brooklyn and Financial District. Since East Village and Brooklyn are living neighborhood with various restaurants and shops. Besides that, East Village is close to NYU as well. Financial District is a busy commercial area and Upper East Side has lowest trips because it is a luxury living neighborhood. People who live are more likely having a car. Therefore, only a few taxi trips over there.

Conclusion

Citi Bike has strong correlation with weather temperature and condition. As for Taxi dataset, there is not an obvious cyclical pattern among each month but only a small difference between spring and fall. Next step, we could add more weather features, for instance, wind speed. We could also put more previous year data in to see if they are similarly correlated as Year 2015.

According to this fact, we would highly recommend Citi preparing more bikes from May to November. Provide maintain service and provide less bikes in order to increase bike's useful life during January and February.

Although a lot of people takes Citi Bike going home after work or school, subway is their main way going to work or school. As for late night, people prefer taking taxi home. Weekends are different from week days, people don't ride Citi Bike but taking taxis especially at mid-night. Further more, there is only few people take either Citi Bike or Taxi from 7:00am to 11:00am. People might stay at home and sleep late. Besides that, people stay home rather than go out at Sunday night.

Based on the result, we would suggest Citi Bike adjust their fare policies. Give some discounts during weekend. Secondly, limit bikes in some less busy stations and cut unnecessary ones. Besides that, providing customized services like mudguard to prevent riders' business suits get stained would also benefit Citi bike riders. On the other hand, Citi could also pursue efforts on advertising benefits of bicycle sports and family activities to remain fairly constant usage during weekends. For instance, two-seated-bikes and bicycle racing event.

Citi Bike Stations, Grand Central, Penn Station, 14th street union square, Spring street and Chambers Street regions have both dense and popular Citi Bike Stations. Citi Bike can build more stations around these popular stations to ensure every Citi Bike rider can get one bike during peak hours. Others, 23rd street, East Village (8th Avenue & 14th Street) and Astor Plaza has less amount of Citi Bike Stations around the area and only a few trips in whole year. Citi Bike could somehow cut these stations to limit cost or intense advertisement among these regions. East Village has both low amount of trips for Taxi and City Bike.

As for taxi drivers, we would suggest them going to East Village and Financial District when they are unloaded. Because there are relatively more trips picked up there. For people who want to get a taxi,we recommend them wait in other area to shorten waiting time.On the other hand, people who live around Upper East Side are more likely having own cars. So that, we would expect taxi drivers spend less waiting time there.

We would also suggest taxi drivers work from 15:00pm to 24:00am during week days since most of people taking taxis during this time period. As for weekends, Saturday 15:00pm to Sunday 7:00am would expect a lot of people taking cabs. On the other hand, passengers could try alternative ways like Citi Bike and Subway to reduce waiting time.

Individual Contributions

Yichen Fan

- Task assigning
- Report editing
- Result analyzing

Shixin Li

- Data cleaning
- MapReduce
- Report editing

Zewei Liu

- Data visualizaion including both plots and maps
- Data cleaning
- Report revising

References

Datasets:

- 2015 Yellow Taxi Dataset

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Collect all the data files from 2015 (yellow taxi only)

Metadata available at:

http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Temporal data is in EST (second resolution).

Spatial data is in GPS.

- Citi Bike Trip Histories

https://www.citibikenyc.com/system-data ("Citi Bike Trip Histories" section)

Collect all the data from 2015.

Temporal data is in EST (second resolution).

Spatial data is in GPS.

Big Data Proposal

Choice for the project

Exploring Citi Bike and Taxi Interactions

Project Purpose

We are going to use the two groups of dataset to find out the relationship between taxi and Citi Bike in given area and time. After finding out the relationships, we could help taxi drivers maximize their profit by giving them the locations and times that people are more willing to take taxis. At the same time, giving both taxi drivers and bike riders information to minimize their waiting time. Furthermore, we could make suggestions to citi bike company to increase their revenue by modifying bike stations and rental price accordingly.

Apart from the basic analysis on our data, we also tried to find insights in our data too. For example, we analyzed why only few people in a certain area use Citi Bike or taxi. The reason might be that there are other more convenient transportations such as subway in this certain area, which lead people to choose subway or other transportations instead of Citi Bike or taxi.

Data
- 2015 Yellow Taxi Dataset
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
Collect all the data files from 2015 (yellow taxi only)
http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

- Citi Bike Trip Histories
https://www.citibikenyc.com/system-data ("Citi Bike Trip Histories" section)

Timeline

04/10/2016-04/17/2016 Gather the taxi and Citi Bike data
04/18/2016-04/25/2016 Clean the taxi and Citibike data
04/26/2016-05/02/2016 Write map reduce program to analyse data
05/03/2016-05/10/2016 Visualize the data and find insight from our result
05/11/2016-05/16/2016 Write project report and make poster for our project

Status Report - 04/17/2016

We have cleaned the taxi and Citi Bike datasets by python.

We also have clarified the goal of our project, which is to make a recommendation to the consumers on the choice of the means of transportation. We would give people recommendations on the following two ways: The first is the time needed for each person to travel from one place to another by taxi or bike at certain time; The second is to help consumers estimate their costs on the trip.

Status Report - 05/01/2016

We wrote map-reduce program to analyse taxi and Citi Bike data and applied many visualization tools such as plotly and cartoDB to visualize the data.

We have got many interesting result from our initial analysis. For example, we find that different day in a week such as Monday, Tuesday, etc. could affect the number of taxi and Citi Bike trips. Different month in a year could also affect the numebr of taxi and Citi Bike trips. According to our analysis, we have made some recommendations such as recommending Citi Bike company to build more stations in certain area or recommend taxi companies to increase their charge in busy hour in order to meet the needs of taxi.