The Weather Influences in Uber and Taxi Trips

SHIDA WU[1], MINQING ZHUANG[2], YUHAO ZHAO[3]
sw3276@nyu.edu, mz775@nyu.edu, yz3085@nyu.edu
[1] *Courant Institute of Mathematical Science, NYU.*
[23] *Center for Data Science, NYU*

Abstract

This project involves investigating how weather influences Uber and yellow taxi trips in New York City from various aspects, including pickup quantity, route, and people's preference under certain weather conditions. We access weather datasets and trip datasets of both yellow taxi and Uber, and conduct research using MapReduce and other tools such as Python, Hadoop, and R. After exploring the datasets, we have found various interactions between weather, Uber trips, and yellow taxi trips, such as how distributions of pickups vary under different weather conditions. Then we utilize multiple visualization tools to virtualize and illustrate our findings.

I. INTRODUCTION

You have to live in New York City for precisely one rainy day to appreciate that it's extremely difficult to find a taxi here when it rains. It's also extremely difficult to figure out how difficult it is compared to normal days. As an exponentially growing platform, Uber has galvanized the traditional taxi industry in an unprecedented way. Under extreme weathers, how Uber reacts and operates may be an interesting question. In this project, we will be exploring the impact of weather to Uber and the traditional taxi industry in New York City from multiple aspects. We are interested in figuring out how different weather conditions may impact Uber and yellow taxi trips. This report starts with a brief summary of the data sets used in our study, followed by the techniques we utilized in order to deal with the large amount of data. Afterward we analyzed several weather impacts on Uber trips and Taxi trips respectively with corresponding spatial visualization, impacts such as how Uber and taxi pick quantities and routines vary under different weather conditions.

II. DATA SUMMARY

In this project we assessed, combined, and manipulated three separate data sets :

- Uber Data[1]: 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. For each pickup, the pickup location and time was recorded.

- Taxi Data[2] : Taxi data from Jan 2014 to Dec 2015. For each month, the data contains around 10 million pickups including time, location and relevant fare information.

[1] Uber Data download from :*https://github.com/fivethirtyeight/uber-tlc-foil-response*
[2] Taxi Data download from :*http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml*

- Weather data[3]: Weather records from Global summary of the day from Jan 2014 to Dec 2015 in NYC. The data includes detailed daily weather summary such as temperature, rain and snow precipitation, visual range, and weather index.

III. Data Cleaning and joining

Since the total amount of taxi data is over 100GB, and it's impossible to compare them directly with Uber data. Therefore, big data infrastructure is needed. We first conducted data merging and cleaning using Mapreduce in Hadoop.

We first download NYC Neighborhood Tabulation Areas[4], which divide NYC into 195 Polygons.

Figure 1: *NYC Neighborhood plot*

To summarize all the taxi trips into one usable data set, we employed Map-Reduce:

Mapper:

For each trip, the key is a triple pair, (*pick_up polygon ID, drop_off polygon ID, date_hour*) and the values include *passenger_count, trip_distance, fare_amount, surcharge, tip_amount, tolls_amount,* and *total_amount*

In particular, the pick_up and drop_off polygon ID were generated by calculating the distance to the centers of each polygon and then selecting the one with smallest distance.

To summarize, the mapper maps each trip to an one-hour interval. By doing this, we can group all the trips traveling from the one location to another in any given date and hour.

Reducer:

For the values in the same key from the reducer, it prints some statistics:
sum of passenger_count, sum of trip_distance, sum of fare_amount, sum of surcharge, sum of tip_amount, sum of tolls_amount, sum of total_amount, tip_rate, surcharge_rate,and sum of trips.

The output of Reducer contains the overall summary of all the taxi trips from Jan 2014 to Dec 2015, each line records the summary statistics of trips from pick_up ID to drop_off ID at a certain day and time (hour).

Mapreduce configuration:

We used 1 mapper and 1 reducer in the jobs. The mapper took about 3:10 minutes to complete for a single month, and overall it took about 75 minutes to finish. In total, there are 27247865 lines in the reduce output.

For the weather data set, we first selected several features that might be relevant to transportations, including: temperature, rain/snow precipitation, and visual range. We also factorized the weather

[3] Data from *http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets*
[4] Shape file from *http://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page*

index into *Rain, Fog and Snow*. Then we joined the summarized Taxi data with weather data by date.

In addition, for further study of the weather impact on taxi trip counts and revenue, we did a second Map-Reduce. This Map-Reduce task simply eliminates the location information and summarized the data by day. In particular, the mapper generates date and trip information pairs. The reducer takes all the trips in the same day, and calculates summary statistics similar to the first reducer. We used one mapper and one reducer in this task, it took over 1 hour to finish.

IV. Results and Discussion

- How weather influences taxi and uber pickup quantity

In this section, we would like to explore relations between weather and pickup quantity by examining changes of pickup quantity under different levels of temperature and precipitation.

Firstly, we analyzed how temperature affects taxi and Uber pickup quantity by plotting distribution of taxi and uber along with temperature. As we can see in figure 2, the y axis is the number of pickup during one day divided by the total number of pickup. We transformed weather data into the same scale as percentage so that all the points and lines could be presented in a single graph.

The trend of taxi pickups seems relatively stable under different levels of temperatures compared to the trend of Uber pickups. There is apparent growth of Uber pickup quantity in Sep 2014, Jan 2015, and Feb 2015. The increase in Sep 2014 is paired with the highest temperature during the year, and the increase in 2015 is paired with the lowest temperature during the year. We conclude that temperature affects Uber pickup more, and under extreme temperatures more people tend to use Uber.

After we examined the interactions between temperatures and pickups, we wanted to see how levels of precipitation influences Uber and taxi pickup.

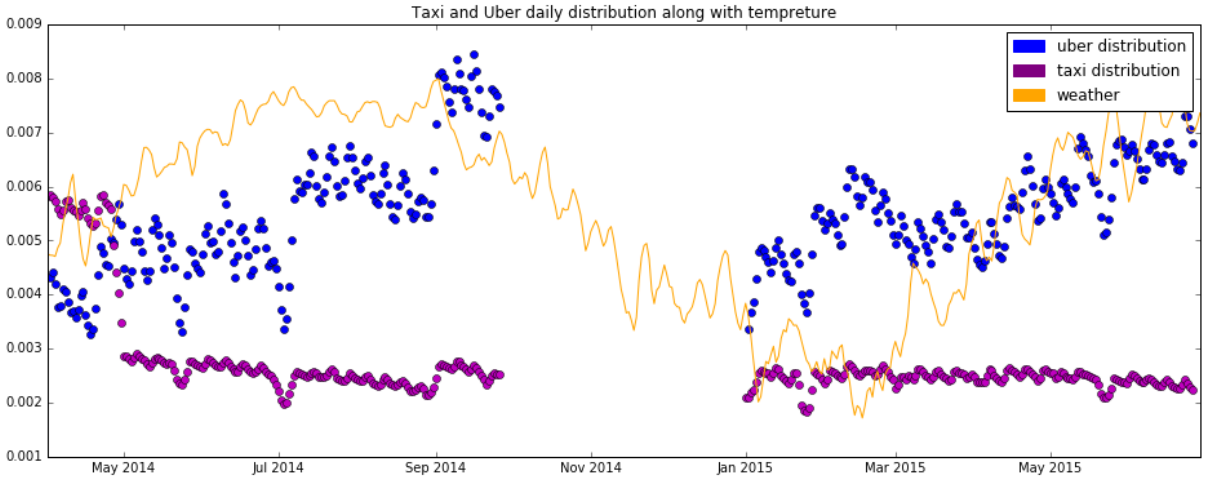

Figure 2: *Taxi and Uber distribution with temperature*

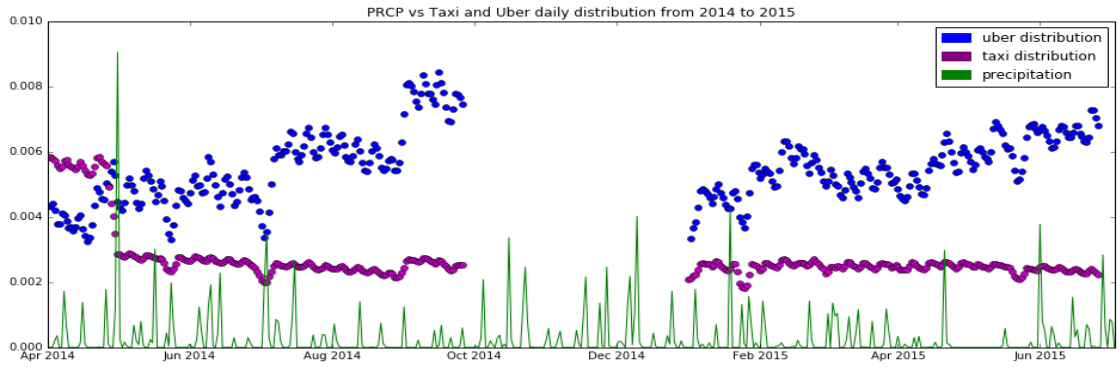

Figure 3: *Taxi and Uber distribution with precipitation from 2014 - 2015*

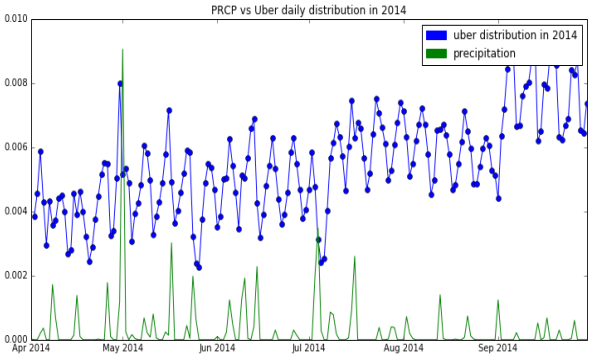

Figure 4: *Uber distribution with precipitation in 2014*

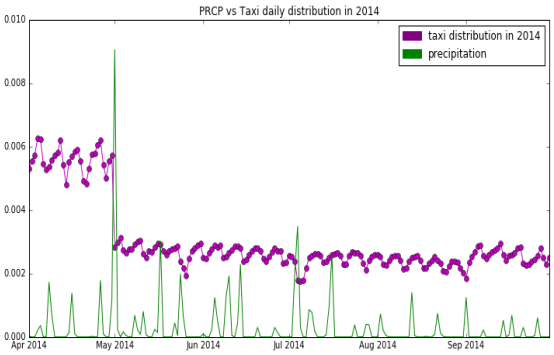

Figure 5: *Taxi distribution with precipitation in 2014*

Figures 3-5 show relations between precipitation and pickup quantity. The last two figures show taxi and Uber pick up distribution separately, which offers a closer look into the relations. Again we transformed the precipitation data in order to present all points and lines in one single plot. From the plots, we can see that for both Uber and taxi, each drop in pickup quantity is paired with a rise in precipitation. An obvious example of this is the two big drops of Uber pickups in May 2014 and July 2014, both paired with high levels precipitation. We can therefore infer that there is less pickup under poor weather and poor weather affects pickup quantity negatively.

❏ How weather influences taxi pickup distribution and routine

After we explored relations between weather and pickup quantity, we would like to examine how weather might affect distribution of taxi pickups across New York City. We did this by plotting density of taxi pickups across New York City under normal weather (sunny) and snow weather, with darker color indicating more pickups in the areas.

We used the summarized taxi trips data set in this problem which was further manipulated by python. And for visualization purposes we used several spatial plotter library in R such as *ggplot2, ggmap* and *sp*. We first queried and formatted the desired data into CSV using python, then we reloaded it using R.

Figure 6 demonstrates density of taxi pickup under normal weather. We can see that in Manhattan, pickups are most concentrated in Midtown and Flatiron areas, which makes sense in reality since there are many commercial activities going on in those areas. Outside Manhattan, the most pickup concentrated areas include LaGuardia Airport and JFK in Queens, and Dumbo and Williamsburg in Brooklyn. Despite of areas with obviously greater density, the whole distribution of taxi pickups across the entire city is relatively balanced.

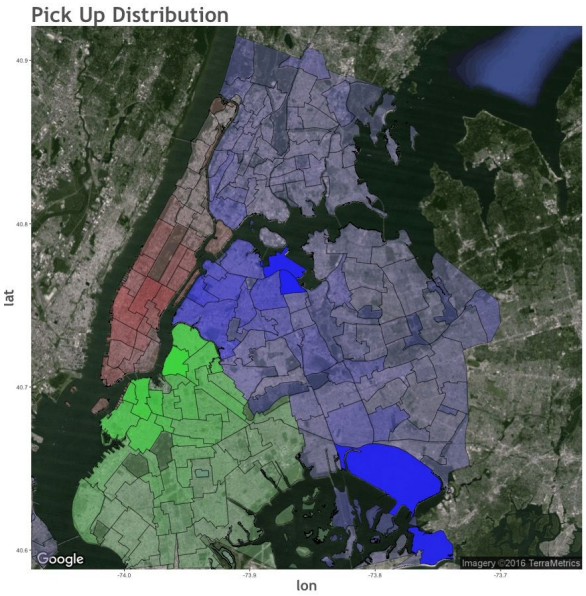

Figure 6: *Taxi pick up distribution in sunny days*

Figure 7 shows density of taxi pickup under extreme weathers such as heavy snow and rain. The most pickup concentrated areas are the same. However, we can see that the entire distribution across city is not as balanced as distribution under normal weather. There are way less pickups in areas other than the pickup concentrated areas such as JFK in Queens and Midtown in Manhattan. This suggests that poor weather affects taxi pickups negatively.

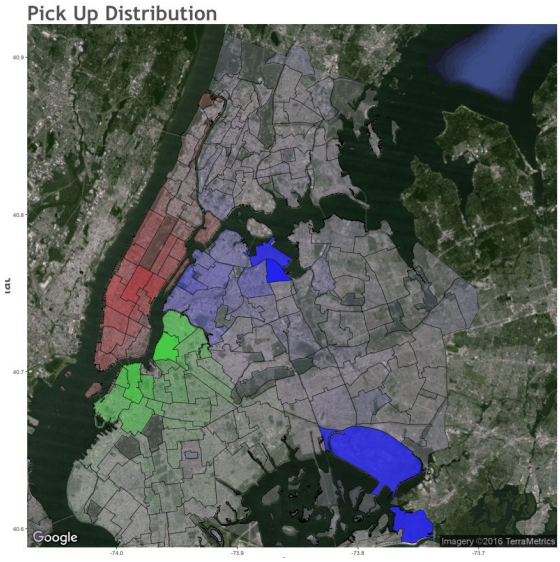

Figure 7: *Taxi pick up distribution in extreme weather*

We also studied the trip route in Manhattan and Brooklyn during normal weather and extreme weather. We selected the top 20 frequent pickup counts. Figure 8 and 9 compared the routes in Manhattan. The thickness of arrow represents the trip frequency, and the heatmap describes the distribution. It's clear that the most frequent trips are between midtown south and midtown east.

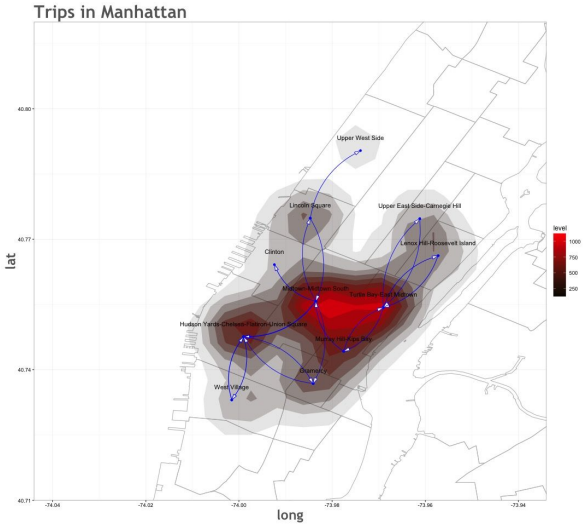

Figure 8: *Taxi route in Manhattan in normal weather*

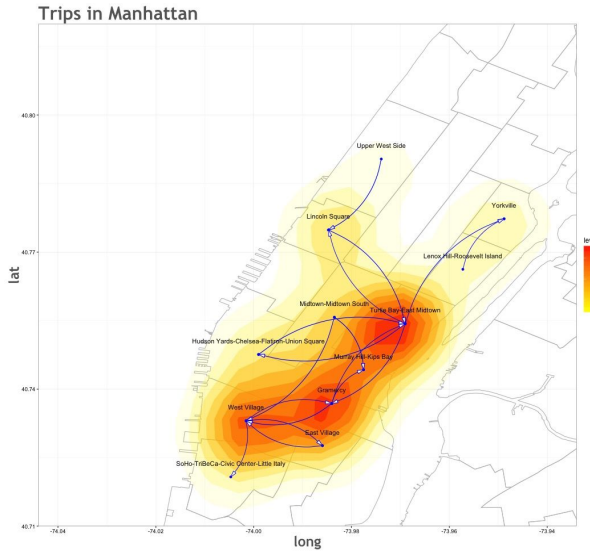

Figure 9: *Taxi route in Manhattan in extreme weather*

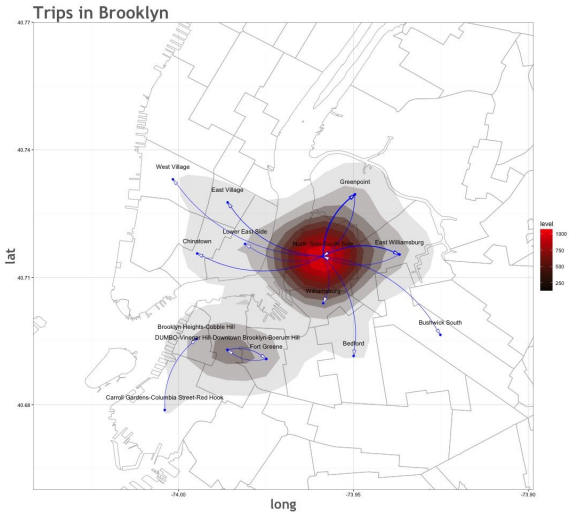

Figure 10: *Taxi route in Brooklyn in normal weather*

Under extreme days, the pickup distribution is less concentrated in midtown. We observed more trips starting from West Village.

For trips in Brooklyn, as in figure 10, during normal days, most of the pick-ups are around Williamsburg. However, during extreme weather, we observed more trips to Manhattan. Also, there is another cluster of pick-ups near Brooklyn Heights-Cobble Hill

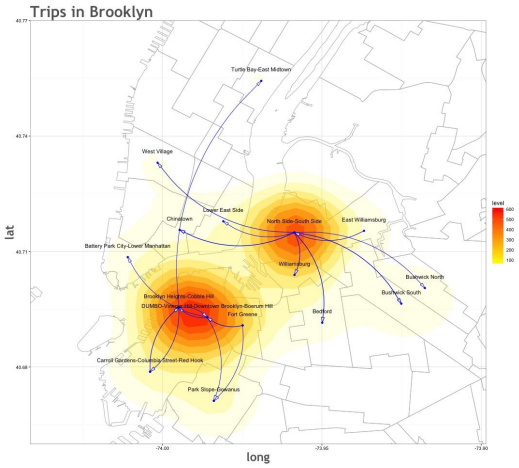

Figure 12: *Taxi route in Brooklyn in extreme weather*

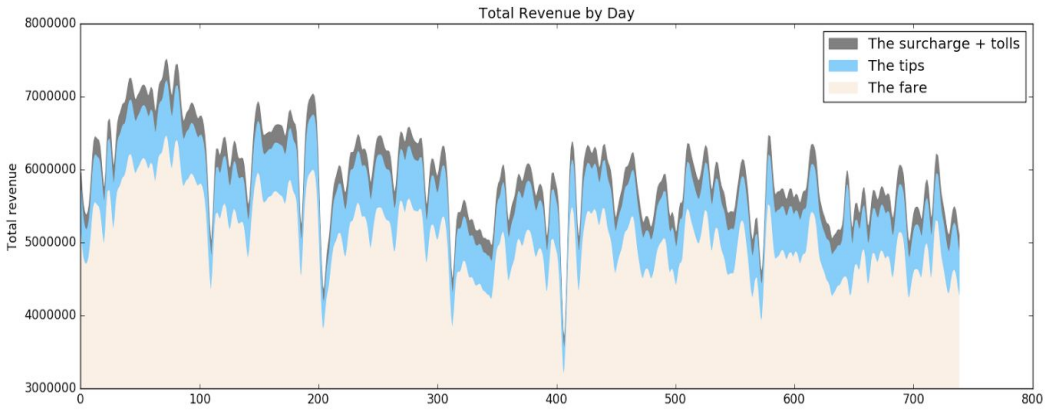

Figure 13: *Taxi Revenue by Day*

❏ How weather influences the tips and revenue

One of the most interesting questions is how does the weather condition affect the overall taxi income. Regarding this problem, we used the daily summarized data from Jan 2014 to Dec 2015. Figure 13 summarizes taxi daily revenue. There is a clear trend that the revenue is decreasing. What's more, there is some evidence shows that the revenue has seasonality patterns.

To study the weather impact, we found out that it's more reasonable to compare within corresponding weekdays.

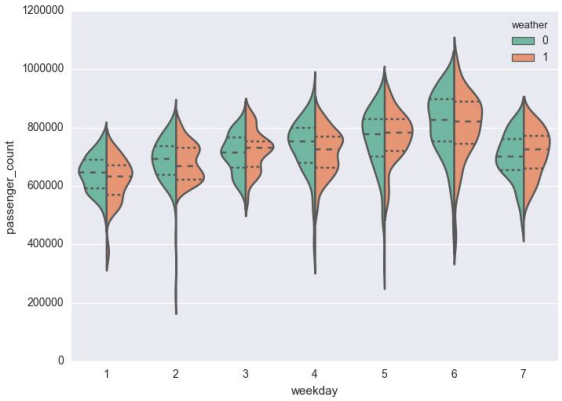

Figure 14: *Passenger Count Distribution*

In figure 14, for each weekday, the left distribution approximates the counts in normal days, and the right distribution approximates the counts in raining days or snowy days. Inside distribution, the dot shows the quantile of the corresponding distribution. We can see that Saturday has more trips in average than other days, but the impact of weather is unclear since the means doesn't show obvious divergence.
An interesting fact is that the distribution of trips in normal days are more bell-shaped while the distribution for rain/snow days are quite random.

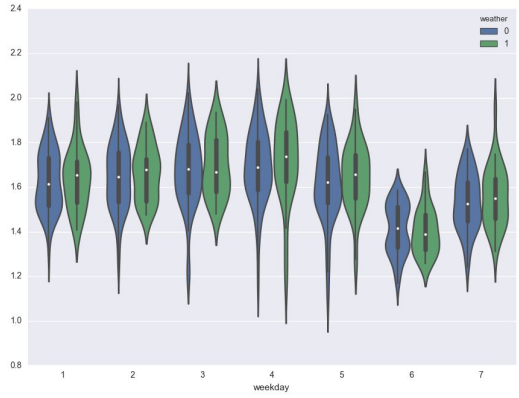

Figure 15: *Average tips*

Figure 15 shows the average tips by weekday and weather. We see that people in rain/snow days tend to give more tips except in Saturday. What's more, the average tip amount in normal days has larger variance, which suggests people leave tips differently by weather. In Saturday, even though there is a larger number of trips, people give less tips on average.

We also explored the relationship of average trip distance with weather and passenger per trip with weather, no interesting pattern was observed.

❏ Do people prefer Uber or Taxi during peak hour (or off-peak hour)

As seen from previous sections stating that how weather influences Uber and taxi based on different subject, we would like to know how people prefer using Uber or Taxi in general. One particular question we are interested in is: whether people prefer Uber than taxi during peak hour or vice versa.

First of all, we have to set the range of peak hour. We define peak hours are from 5am to 10am or 3:30pm to 10pm and other hours will be off-peak hours. Hence, we get the statistics below:

	Peak Hours	Off-peak hours
Taxi	37747460	58440656
Uber	2183991	2350336

Table 1: Amount of people taking taxi or uber during peak hours or off-peak hours

Notice that the number shown above is the amount of vehicles that already pick up someone during rush hour or not. To make this statistics reasonable, we would like to consider the change of ratio between taxi and uber from peak hours to off-peak hours. During off-peak hours, the ratio is approximately about **24.86** (58440656/2350336), whereas the ratio drop to **17.28** (37747460 / 2183991) during peak hours. This situation demonstrates that people tends to use Uber than Taxi during peak hours. Therefore, if you are an Uber driver and have limited time, then peak hours may be a better time for you to work.

V. Summary

Through our study of the three data sets, we found out that even though the data is public, some of them have low quality or are stored in inconvenient format. Therefore, we believe that data quality is an important sector before publishing the data. What's more, even though the data is huge and dirty, using mapreduce allows us to reshape the data into our desired format, meanwhile keep efficiency and accuracy.

Uber and taxi has their unique properties regarding the weather effect. Uber trip quantities are more sensible to the weather, due to the flexibility of Uber drivers. In contrast, the number of taxi trips experiences less influence, since taxi drivers still have to work for their company even under severe weather conditions. However, the weather indeed affects taxi routines and the effect varies for different neighborhoods throughout NYC.

In conclusion, the impact to Uber and taxis is very complex. When suffering unusual weathers, we expect very different reactions of the two service due to their own nature.

VI. Contribution

Yuhao Zhao:
Merging datasets; Generating plots; Report Writing
Minqing Zhuang:
Report Writing; Running Map-Reduce Tasks
Shida Wu:
Generating plots; Running Map-Reduce Tasks

VII. References

1. *https://github.com/fivethirtyeight/uber-tlc-foil-response*

2. *Veloso, Marco, Santi Phithakkitnukoon, and Carlos Bento. "Sensing urban mobility with taxi flow." Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM, 2011.*

3. *Rajaraman, Anand, and Jeffrey D. Ullman. Mining of massive datasets. Vol. 1. Cambridge: Cambridge University Press, 2012.*