Final Draft

Milica Cvetkovic(mcvetkovic) Keyu Hu(khu54) Steven Min(nmin)
Xintong Shi(xshi242) Tianrun Wang(twang494)

Group 11

December 12, 2020

1 Introduction

The data is called New York City Taxi data, and it is from 2013. In total, there are about 14,500,000 trips per month, and 174,000,000 trips in a year. The data is stored in CSV format, organized by month. For each month, there are two separate CSV files – one for the taxi trip data and one for the taxi trip fare. The taxi trip data files include information such as car ID, driver ID, pick-up and drop-off date/time, pick-up and drop-off location, trip time, and trip distance. The taxi fare data files include information such as car ID, driver ID, payment type, fare amount, surcharge, tip, and total amount.

2 Statistical Questions

How does speed vary with time of day? - To answer the question, we used the distance formula: $speed = \frac{distance}{time}$. We grouped the data by hour starting from 12:00 am to 11:00 pm and calculated the average speed at each hour. Then, we made a table and plotted the findings. From the plot, we found that from 8:00 am - 7:00 pm, the average speed is quite stable, and it is at its lowest. The average speed seems to increase overnight, peaking around 5:00 am before starting to decrease again.

Which area has the most taxi pick-ups? - We looked at the pick-up longitudes and latitudes of the taxi trip data. We counted how many times each coordinate appeared and made a list of the top 100 most common coordinates. This tells us the locations that the taxi-drivers are most likely to get customers. We made a table of the coordinates and created heat maps.

Can we predict trip time from trip distance, and pickup and drop-off locations? - We looked at the baseline model, linear model, and random forest model, using root mean square error (RMSE) to measure the performance of the models. Both the linear model and the random forest model suggest that trip distance is the only variable that impacts trip time.

medallion	hack_license	vendor_id	rate_code	store_and_fwd_flag	pickup_datetime	dropoff_datetime	passenger_count	trip_time_in_secs	trip_distance	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
2010000001	2010000001	VTS	1		2010-01-01 00:00:00	2010-01-01 00:34:00	1	34	14.05	-73.948418	40.72459	-73.92614	40.864761
2010000002	2010000002	VTS	1		2010-01-01 00:00:00	2010-01-01 00:33:00	1	33	9.65	-73.997414	40.736156	-73.997833	40.736168
2010000003	2010000003	VTS	1		2010-01-01 00:00:00	2010-01-01 00:07:00	1	7	1.63	-73.967171	40.764236	-73.956299	40.781261
2010000004	2010000004	VTS	1		2010-01-01 00:00:00	2010-01-01 00:33:00	1	33	26.61	-73.789757	40.646526	-74.136749	40.601543

Figure 1: A small subset of the New York City taxi trip data

3 Model Overview and Statistical Analysis

The data is from the University of Illinois at Urbana-Champaign databank, and the dataset was made publicly available under the CC0 license. The data was obtained through a Freedom of Information Law (FOIL) request from the New York City Taxi Limousine Commission (NYCTL). We used trip data from January 2013 to June 2013. These data are organized into 6 CSV files which are about 1.7 - 1.8 GB each. In total, they add up to around 11 GB.

How does speed vary with time of day?
To answer how speed varies with time of day, we looked at the trip data from January through June. We used R, specifically the dplyr package to efficiently manipulate our data. Dplyr and lubridate package allowed us to easily group the data by hour. For our calculation, first, we converted the trip time in seconds variable to trip time in hours. Then, we simply used the $\frac{distance}{time}$ to find speed in miles per hour. We wrote a speed.R file which takes one trip data CSV file as an input and creates a .txt file that contains information about the speed at each hour. Since we have six trip data CSV files, we utilized CHTC to run multiple jobs and created six .txt files. Each .txt file contains speed at each hour for their respective month. We ran two jobs at a time, and it took about 13 minutes to run two jobs. For two jobs, we requested 8 GB of memory and 1 GB of disk space. Then, we wrote an average_speed.sh file which reads 6 .txt files and calculates the average speed at each hour for January through June. Then, we created a plot.

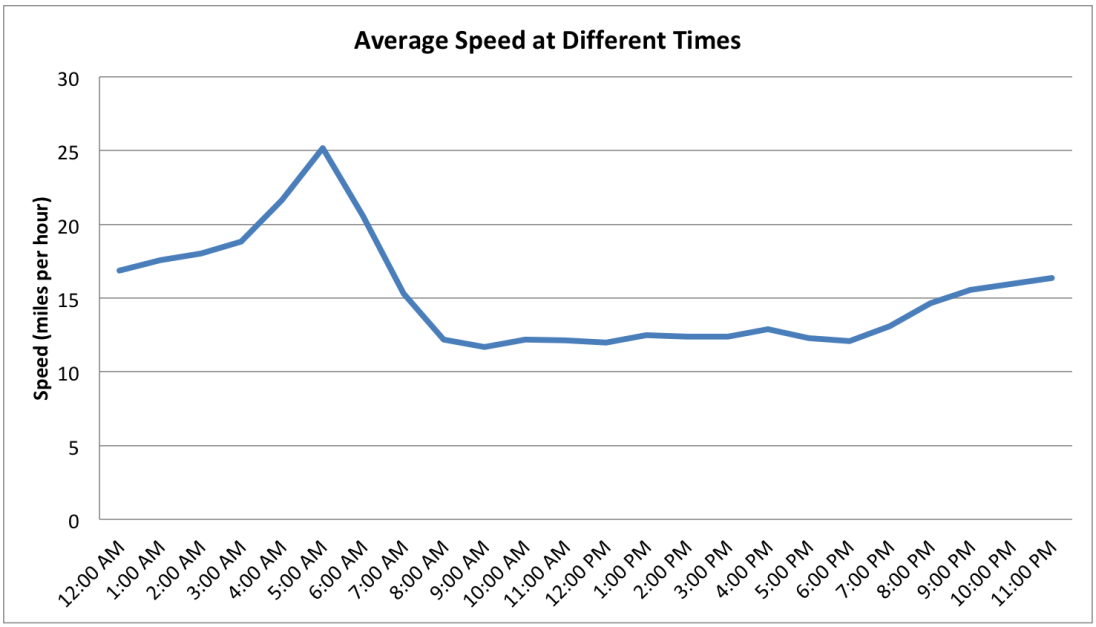

Figure 2: Speed at each hour

From Figure 2, an interesting finding is that the speed is highest at 5 am in the morning. To further investigate this, we decided to look at which hour is the busiest.
First, we wrote a count_trips.R file which finds the total number of trips at each hour. We used the dplyr package, along with the lubridate package, to group the data by hour. The file count_trips.R takes one trip data CSV file and creates a .txt file consisting of number of trips at each hour. We used CHTC to run six jobs for January through June. Once we have six .txt files consisting of count of number of trips for their respective month, we wrote a sum.sh to add up all the trips by hour. Then, we created a plot (Figure 3).
Looking at Figure 3, a possible explanation for why speed is highest at 5 am is the reduction in total number of trips. This suggests that less people are active during that time of day, and therefore, there is less traffic on the road. Another interesting finding is that the highest number of trips happen to be around night time, indicating that a lot of people are going out for dinner and nightlife.

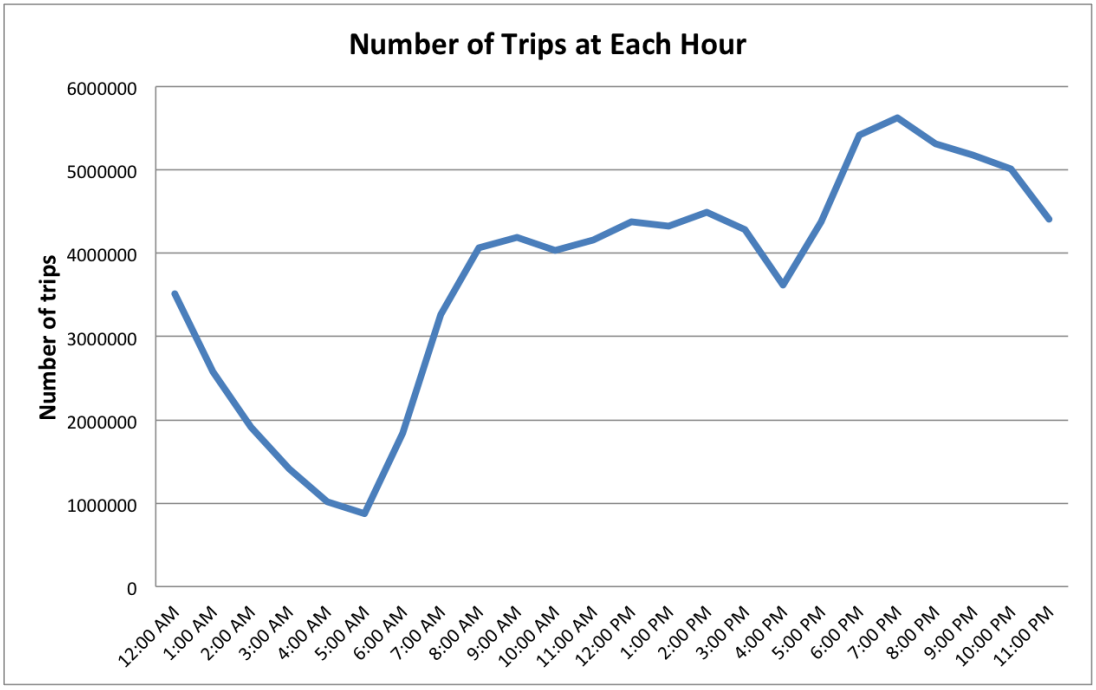

Figure 3: Trips at each hour

Which area has the most taxi pick-ups?

We investigated the busiest locations by using the pick-up longitude and pick-up latitude variables. First, we wrote a .sh file to generate a .csv file consisting of about 100,000 random trips from January to June and plotted the points.

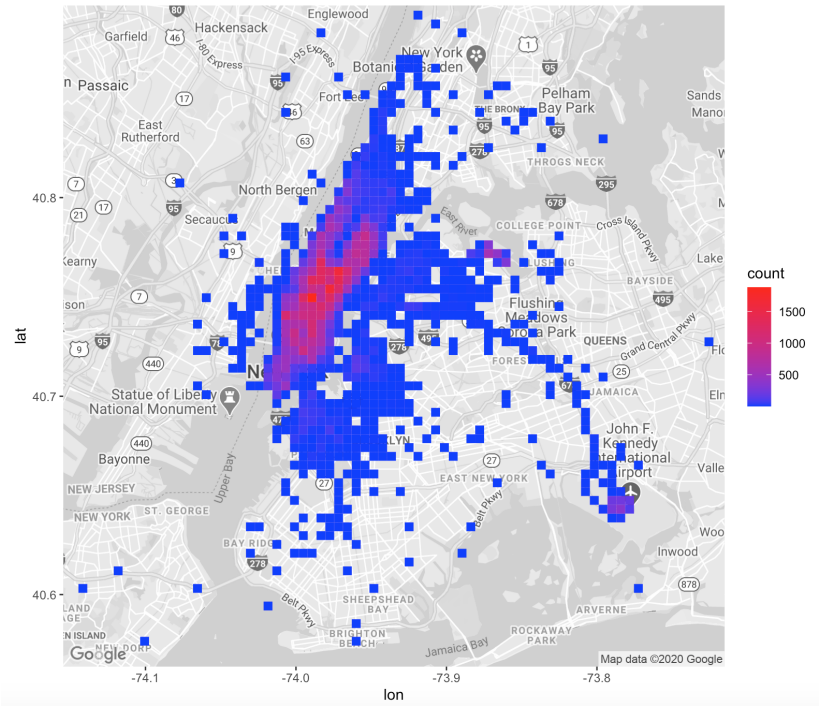

Figure 4: Heat map of pick up points

To further look into popular pick up locations, we rounded the coordinates to three decimal places since the original dataset has up to six decimal places. Rounding the decimal to three places makes the coordinates more general and reasonable. Then, we counted how many times the coordinates appear and made a list of top 100 coordinates. These top 100 coordinates and their respective counts are plotted in Figure 6. We removed any (0,0) coordinates.

Latitude	Longitude	Count
40.751	-73.994	434
40.750	-73.992	337
40.750	-73.991	332

Table 1: Top three coordinates

All the top 3 coordinates point to Penn Station and Madison Square Garden and its surroundings as shown in Figure 5. You can see three big circles around the middle of map (Figure 5). The rest of the top 100 most popular pick up spots that are not the airports are in Manhattan.

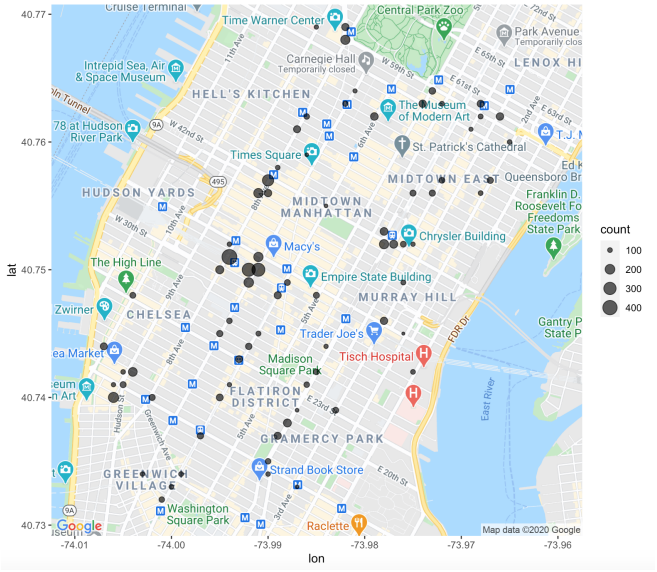

Figure 5: Popular pick up spots (without airports)

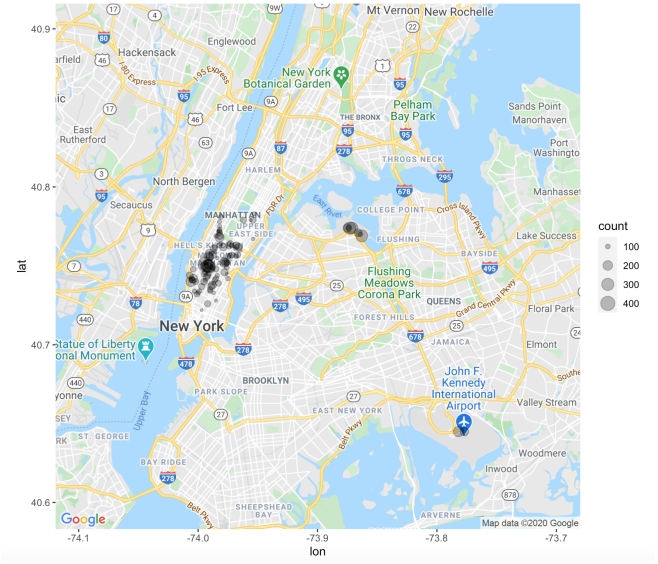

Figure 6: Popular pick up spots (with airports)

Figure 6 shows that John F. Kennedy airport and LaGuardia airport are extremely popular pick up spots. Since a lot of the traffic from the airports are likely to be long trips, we separated the data into long trips and short trips. Long trips are defined as trips which are more than 10 miles, and the short trips are defined as trips which are less than 2 miles.

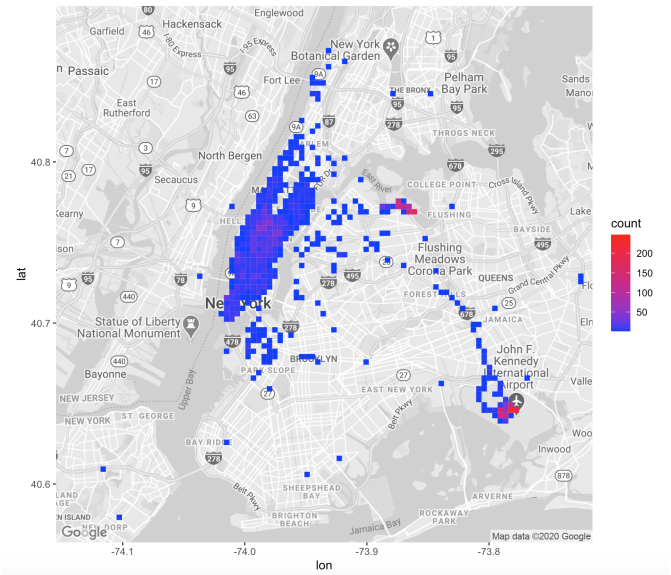

Figure 7: Long trips (trips more than 10 miles)

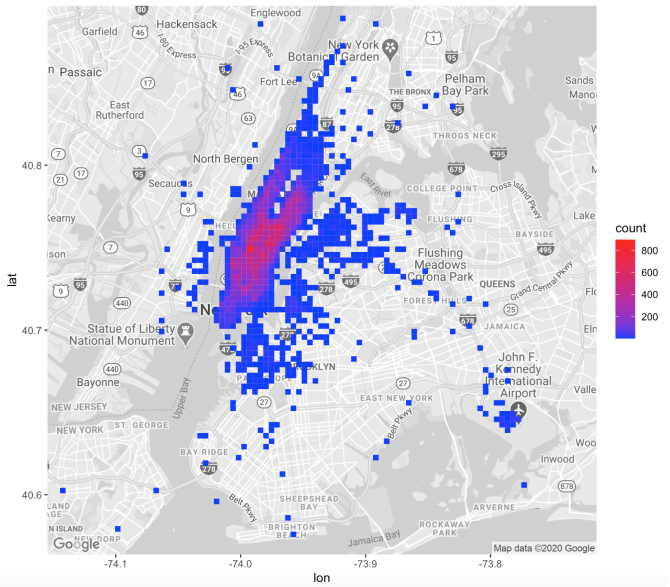

Figure 8: Short trips (trips less than 2 miles)

As expected, a lot of the longer trips start from pick up spots at the airports.

Can we predict trip time from trip distance, and pickup and drop-off locations?
We looked at linear regression and random forest models. We tried to build regression models to predict trip time in seconds using trip distance, pick up longitude, pick up latitude, drop-off longitude, and drop-off latitude. Since we are using root mean square error (RMSE) to measure the performance of our models, we ran a R file on CHTC that takes in input data and produces three RMSE.txt files for baseline model, linear model, and random forest model. Additionally, it creates two model.rds file - one for the linear model and one for the random forest. Running this job on CHTC took about 50 minutes after requesting 15 GB of memory and 1 GB of disk space.
In our computation, we split the data into 75% train data, and 25% test data. Then, we created baseline value using the mean of trip time in seconds and calculated the root mean square error (RMSE). Below is the table of the models' RMSE.

Model	RMSE
Baseline	536
Linear	342
Random forest	291

Table 2: RMSE of models

We see that the linear model (Figure 9) is an improvement from the baseline as the linear model has lower RMSE. Linear model also shows that trip distance is the only significant variable. The longitude and latitude of pick up and drop-off locations are not significant.

```
Call:
lm(formula = trip_time_in_secs ~ trip_distance + pickup_longitude +
    pickup_latitude + dropoff_longitude + dropoff_latitude, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7987.6  -204.5   -66.8   137.7  8188.2

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       341.72745   10.08220  33.894   <2e-16 ***
trip_distance     125.42876    0.40568 309.178   <2e-16 ***
pickup_longitude    0.03726    0.97833   0.038   0.9696
pickup_latitude    -3.37198    1.86728  -1.806   0.0710 .
dropoff_longitude  -1.66585    0.96537  -1.726   0.0844 .
dropoff_latitude    1.26423    1.85951   0.680   0.4966
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 351.2 on 66849 degrees of freedom
Multiple R-squared:  0.5888,    Adjusted R-squared:  0.5888
F-statistic: 1.915e+04 on 5 and 66849 DF,  p-value: < 2.2e-16
```

Figure 9: Linear model

Next, we built a random forest model (Figure 10). In our random forest model, we set the "mtry", which is the number of features bagged in each independent tree to 2, roughly the half of the total number of features. We observe that as the number of trees increases, the RMSE decreases as shown in Figure 12. At its lowest, RMSE of random forest is 291 and it is lower than that of linear model and baseline. Therefore, we can say that random forest model outperforms the linear model in predicting trip time. Furthermore, from the importance of variables produced by random forest (Figure 11), we observe that trip distance impacts the trip time the most compared to other variables. This agrees with the result of linear regression.

```
Call:
 randomForest(formula = trip_time_in_secs ~ trip_distance + pickup_longitude +
pickup_latitude + dropoff_longitude + dropoff_latitude, data = train,      mtry = 2,
importance = TRUE, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 90469.2
                    % Var explained: 69.85
```

Figure 10: Random forest model

trip_distance	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
205.78610	67.70196	75.29090	74.71308	62.86790

Figure 11: Random forest importance

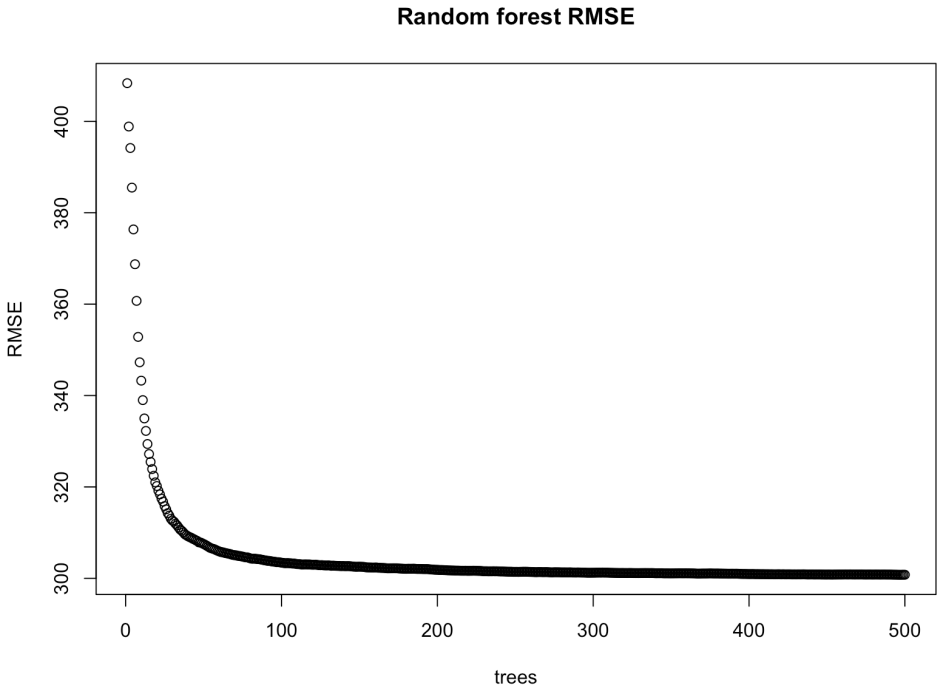

Figure 12: Random forest RMSE

Difficulties

In general, a difficulty we encountered was figuring out bad data points. From skimming through the first few lines of data, it is hard to find out which type of errors could be present in our dataset. Many trips reported the coordinate (0,0), or covered unreasonable distances or times. We discarded the (0,0) coordinates, and we checked the accuracy of the trip time variable by making sure it aligns with the number we get from subtracting pickup time from drop-off time.

Another difficulty was computing on CHTC. We encountered a lot of "Job has gone over memory limit" problems. We had to request more memory and run the job again. When we requested more memory, our job priority got lowered, and therefore, it took quite a while for the job to start running.

4 Conclusion

The results from investigating speed at each hour seem reasonable with the fastest speed being at 5 am, and then, it decreases as the day progresses due to traffic. The map of popular pick-up locations shows that the airports are popular pick-up locations, and trips beginning from the airports are usually long distances. More work can be done by picking a specific pick-up location and looking at the number of trips at each hour for that location. This may tell us whether the taxi trips are mostly from local people taking taxis for daily work commute, or from tourists. Furthermore, more interactive maps showing taxis moving from their pick-up locations to drop-off locations on the map throughout the day can be interesting. Our results from the linear model and random forest model show similar outcomes, indicating that trip distance is the only variable significant in predicting trip time. Pick-up and drop-off locations are not significant. These models can possibly be improved by adding other variables such as fare and time of pick-up.

5 References

Brian Donovan and Daniel B. Work. "Using coarse GPS data to quantify city-scale transportation system resilience to extreme events." presented at the Transportation Research Board 94th Annual Meeting, January 2015.
Brian Donovan and Daniel B. Work "New York City Taxi Trip Data (2010-2013)". 1.0. University of Illinois at Urbana-Champaign. Dataset. http://dx.doi.org/10.13012/J8PN93H8, 2014.