Data Science Challenge: Trips!

Submission By: Udit Anand
Program, School: MSFE, UIUC
Email: uanand2@illinois.edu
Phone: +1 217 9790599
Time: 1 day

This coding challenge is designed to test your skill and intuition about real world data. For the challenge, we will use data collected by the New York City Taxi and Limousine commission about "Green" Taxis. Green Taxis (as opposed to yellow ones) are taxis that are not allowed to pick up passengers inside of the densely populated areas of Manhattan. We will use the data from September 2015. We are using NYC Taxi and Limousine trip record data: (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml).
Required Questions: Please answer completely all four required questions.
Question 1

- **Programmatically download and load into your favorite analytical tool the trip data for September 2015.**

- For doing this challenge, I choose R and RStudio. Primarily because of its comprehensive availability of packages. The data was downloaded and loaded into the desired working directory.

- **Report how many rows and columns of data you have loaded.**

- Number of rows = 149426, Number of columns = 21

Question 2

- **Plot a histogram of the number of the trip distance ("Trip Distance").**

I wasn't sure about the meaning of the phrase "number of trip distance". Therefore, I have plotted two histogram. Figure1: Histogram of rip distance. Figure 2: The number (frequency) of trip distance. Furthermore, I also plot the pdf of the normal for the two graphs to get an idea about the distribution. This is further explored in part (b) of this question.

Note: I have assumed that the trip distance is in Miles as I could not find the units.

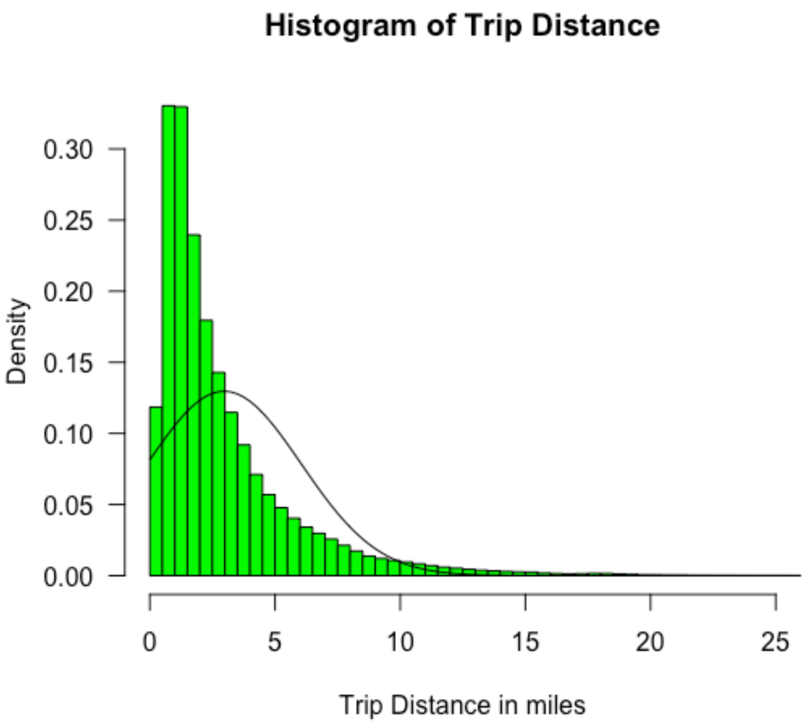

Figure 1: Histogram of Trip Distance

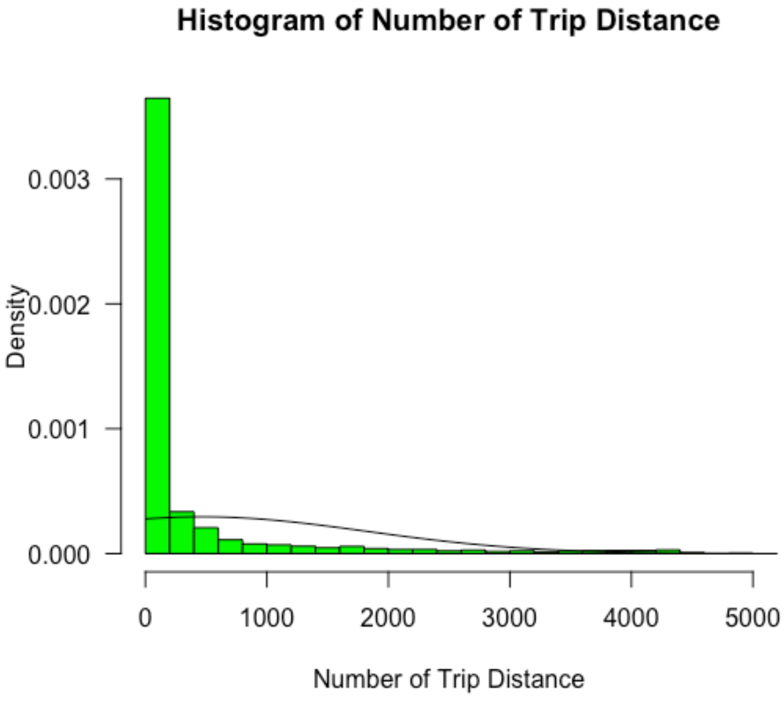

Figure 2: Number (Frequency) of trip distance

- **Report any structure you find and any hypotheses you have about that structure.**

  It seems that the data is right skewed. Furthermore, from Figure 1 we can infer that most of the trip distances are concentrated in the 0 to 5-mile range (highly dense). Similarly, for Figure 2, 0 -100 range seems to be the most dense.

  Something, that we can infer is that short travel distances are extremely popular and that as the trip distance increases the frequency of the trip decreases.

  Moreover, I wanted to find if the distribution follows a known distribution. However, both the trip distance and number of trip distance do not seem to follow any known distribution.

  This is further confirmed in Question (5)

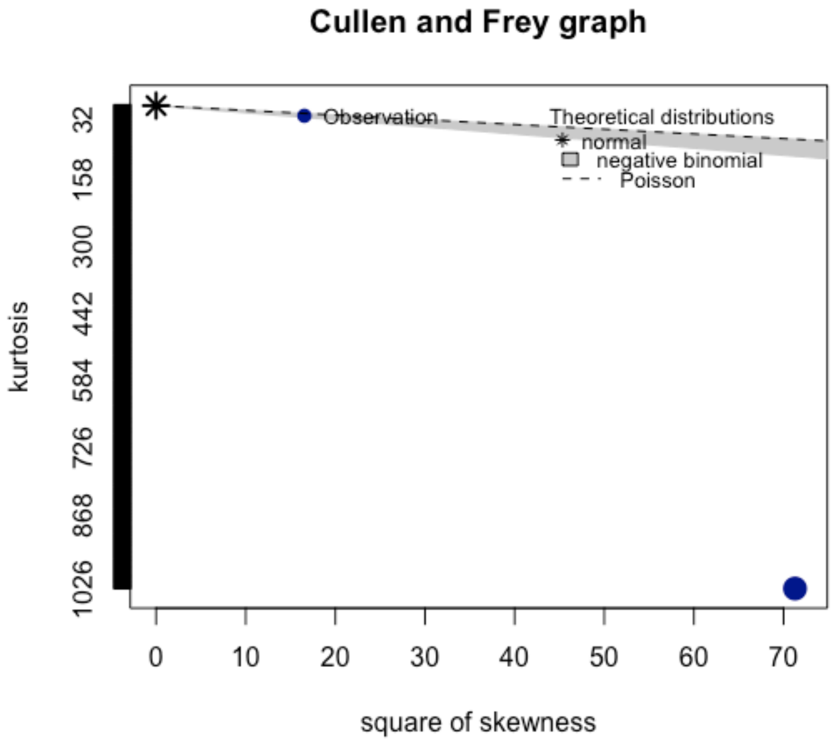

Figure 3: Describe Distribution of trip distance

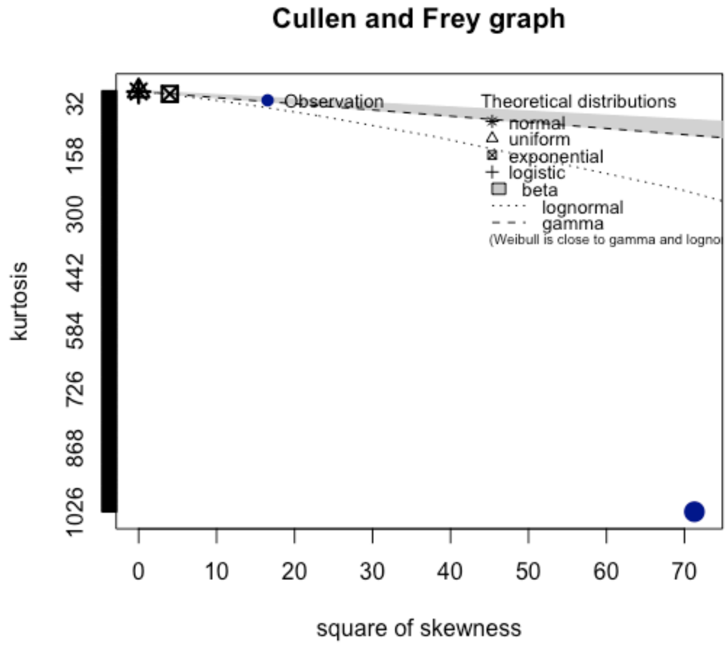

Figure 4: Describe distribution of number of trip distance

Question 3

- **Report mean and median trip distance grouped by hour of day.**
- Mean: 62288.58, Median: 60496
- **We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.**

  For this part, I compare two airports: JFK and LaGuardia. I compute the average fare for pickup and dropoff, and the number of pickup and drop off. Results are plotted below.

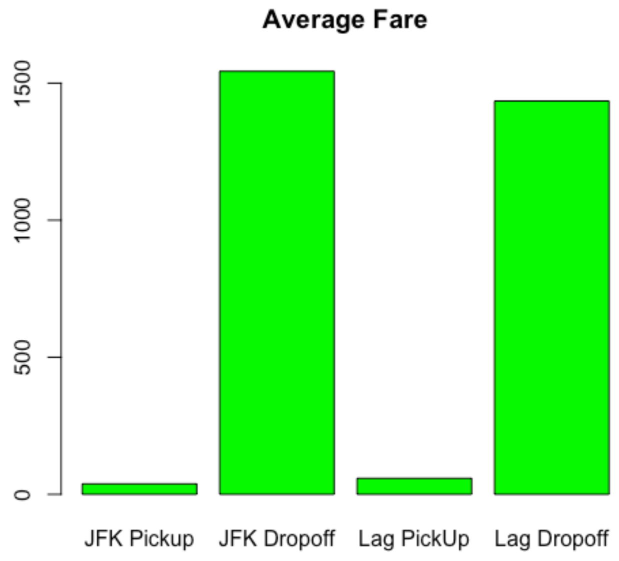

Figure 5: [Airport Statistics] Average Fare

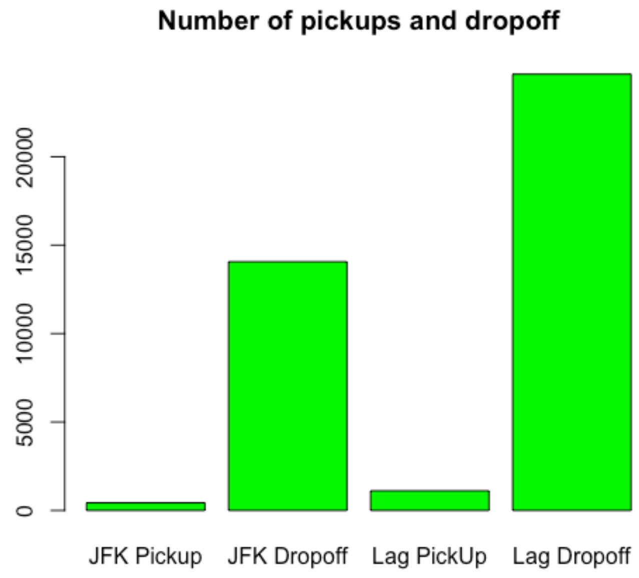

Figure 6:[Airport Statistics] Number of Pickup and Dropoff

Question 4

- **Build a derived variable for tip as a percentage of the total fare**

  The following R code was used to achieve the above task:

```
total_cost_except_tip  = trainData$Fare_amount + trainData$Extra +
trainData$MTA_tax + trainData$Tolls_amount + trainData$improvement_surcharge

tip_amount = trainData$Total_amount - total_cost_except_tip

trainData$Tip_as_perecentage_of_total_amount =
(tip_amount/trainData$Total_amount)*100
```

- **Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). We will validate a sample.**

- For the purpose of this assignment I built a predictive model using Random Forest. The code is in the R script attached. Another plausible option, that I could have used is the XGBoost in R. Inherently, the two are different as the former focusses more on bagging and the latter on boosting. Random Forest reduce error by reducing variance and XGBoost reduce error by reducing bias. XGBoost oftern outperforms Random Forest but I chose to build a Random Forest Model because: (a) Less parameters need to be configured as compared to a boosting algorithm (b) It is relatively more robust to overfitting.

- Please Note: Due to the large volume of data, I did not perform a Kfold cross validation.

Question 5
Choose only one of these options to answer for Question 5. There is no preference as to which one you choose. Please select the question that you feel your particular skills and/or expertise are best suited to. If you answer more than one, only the first will be scored.

- ***Option E*: Your own curiosity!**

- **If the data leaps out and screams some question of you that we haven't asked, ask it and answer it! Use this as an opportunity to highlight your special skills and philosophies.**

- For this task, I chose to go with Option E: Your own curiosity. I had an intuition that I could use massage this data to get some extra information. Some initial ideas that I had were:

    - Plot a list of popular weekend night spots: Since we are given the dropoff latittudes and longitudes. We could group them by hour to get the popular night hours.

    - Get a list of most popular routes. We could use the pairs of latitude and longitude to generate this.

    - Since the FinTech is such a big thing these days. We could get the payment methods on most popular routes wanted to see what payment method are used on which routes. Using the payment types, we could correlate it with user demographics to get an idea about the area.

    - I decided to find the most popular route based on two characteristics: total number of trips and average fare. These two were derived using the given data set. Since the data showed that short trips were most common, which means that trips with low average fare and high number of trips would be most popular.

- The data points were then plotted using ggplot on a map of New York. This data set was then joined with the original given data set to get other conclusions like most popular payment type for that fare, most popular area to get that fare.

- Plot of locations with average fare > $14 and Total Number of Trips > 100. The rationale behind using these numbers is explained in the R script.

- And, I also plotted a map of NYC where Number of Trips is greater than 10.

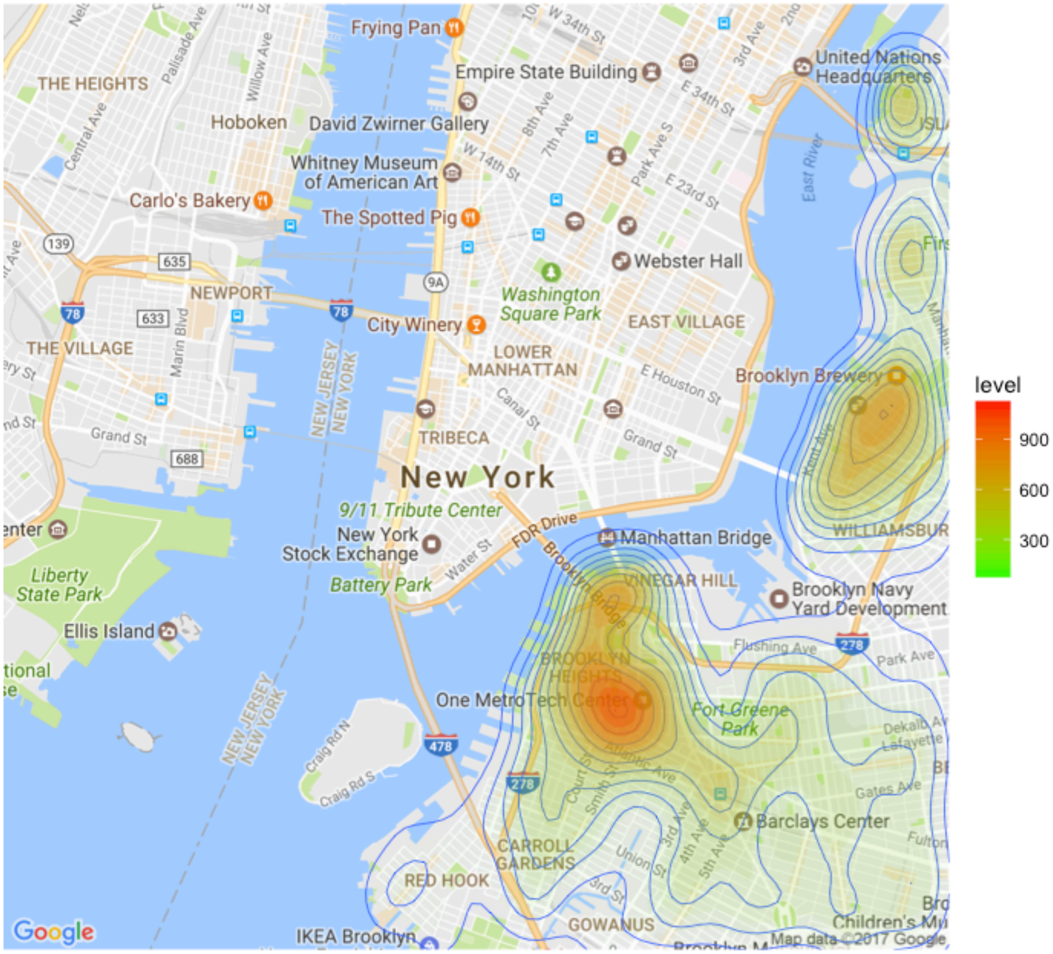

Figure 7: Map of locations with Average Fare > $ 14 and Number of Trips > 100

Most popular location: Area around Brooklyn bridge

Latitude	Longitude	Frequency
40.808	-73.964	16
40.718	-73.958	11
40.722	-73.958	23
40.714	-73.952	11
40.79	-73.952	11
40.734	-73.871	12
40.755	-73.845	13
40.721	-73.844	13
40.76	-73.83	14
40.7	-73.808	15

Table 1: 10 Most Popular Locations satisfying the above constraint

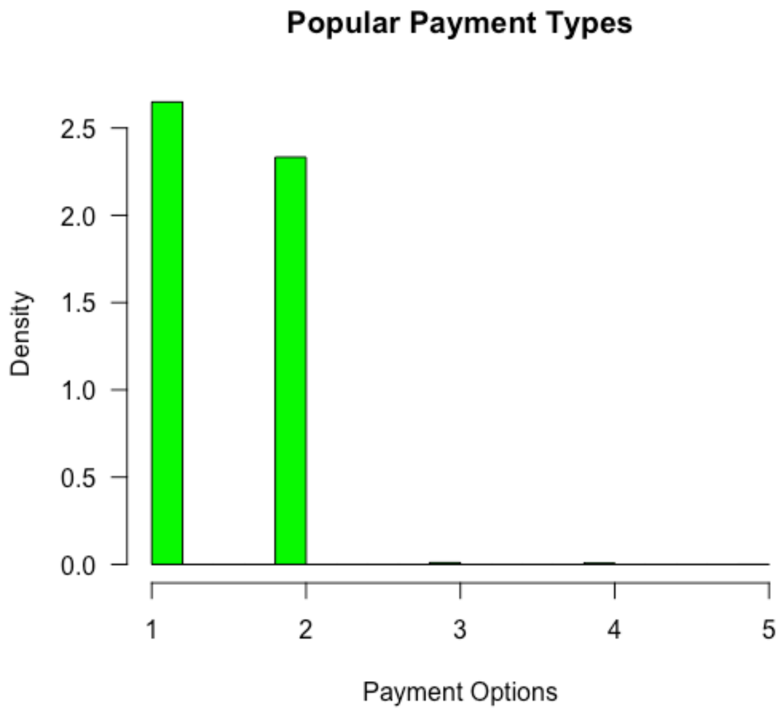

Figure 8: Popular payment types for the given constraint

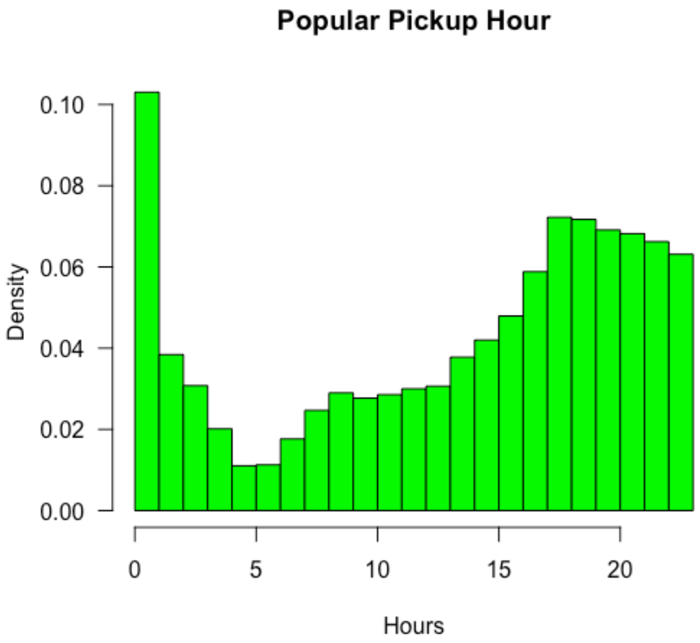

Figure 9: Popular Pickup Hour for given constraint

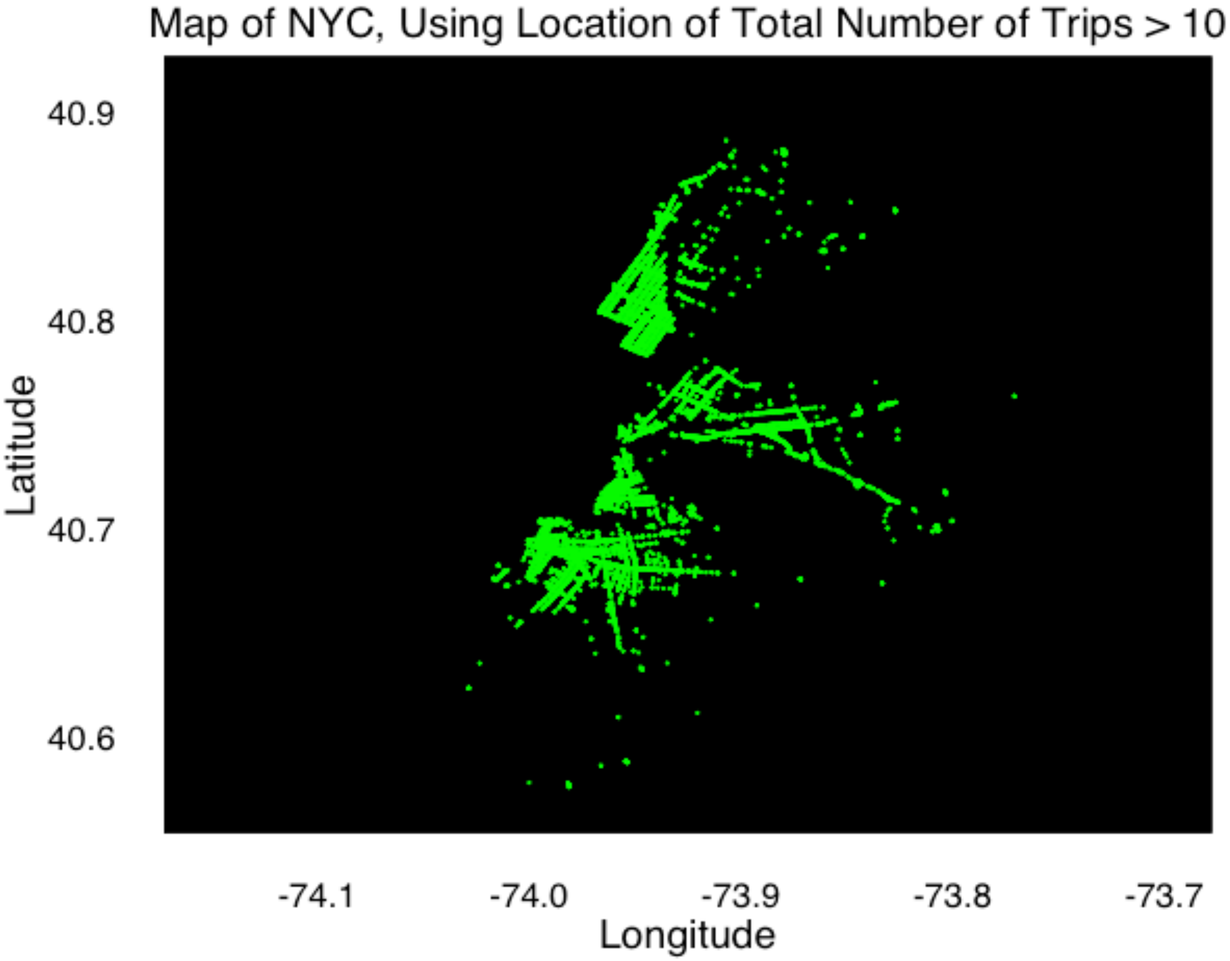

Figure 10: Map of NYC using location of Total number of trips greater than 10