Examining New York City's Yellow Taxi Data Set

CS 516 Final Project - Midterm Report

Ziyi Wang, Timothy Blumberg

October 28, 2016

Abstract

In our final project, we analyze NYC's *very* public taxi dataset [1] for interesting and surprising results. Our analysis has been principally done through queries on a SQL database, but because of the geographic nature of the data, we were forced to visualize from a very early stage in our project's formation. During this preliminary stage of developing our project, we have established an efficient workflow and found areas to engage in a more prolonged analysis during the remainder of the semester.

1 The Data

The dataset is extremely large (there is about 1.6Gb of data produced every month at present date), so this creates many challenges as we attempt to gain insights from it. A powerful DBMS helps to cut down our query runtime considerably. The data is relatively clean given its size and complexity, and definitions for the coded portions of the data (such as the `payment_type` field) is given in the data dictionary [2]. The NYC Taxi & Limousine Commission (TLC) collects and reports data for three different kinds of vehicles in NYC: yellow taxis, green taxis, and for-hire vehicles (FHV). Yellow taxis provide street-hailing service in Manhattan, Green taxis are designed to be useful when getting around in the boroughs of NYC, and the FHVs are available only through pre-arranging the pickup (i.e. cannot provide service that was not pre-arranged). For our project, we focus exclusively on the yellow taxis.

Each row contains start and end time, pickup and drop-off coordinates, number of passengers (as reported by the driver), fare amount, tip amount, distance traveled and several others. We took a look at many of the fields individually as well as exploring relationships between several variables at a time.

2 Data Processing Workflow

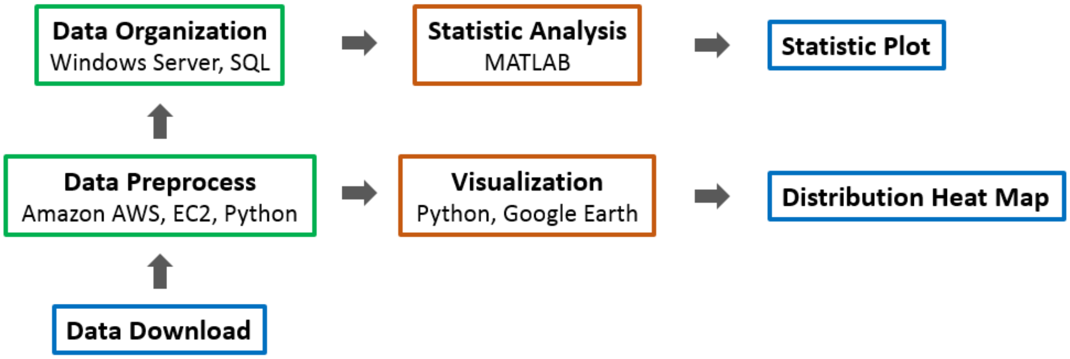
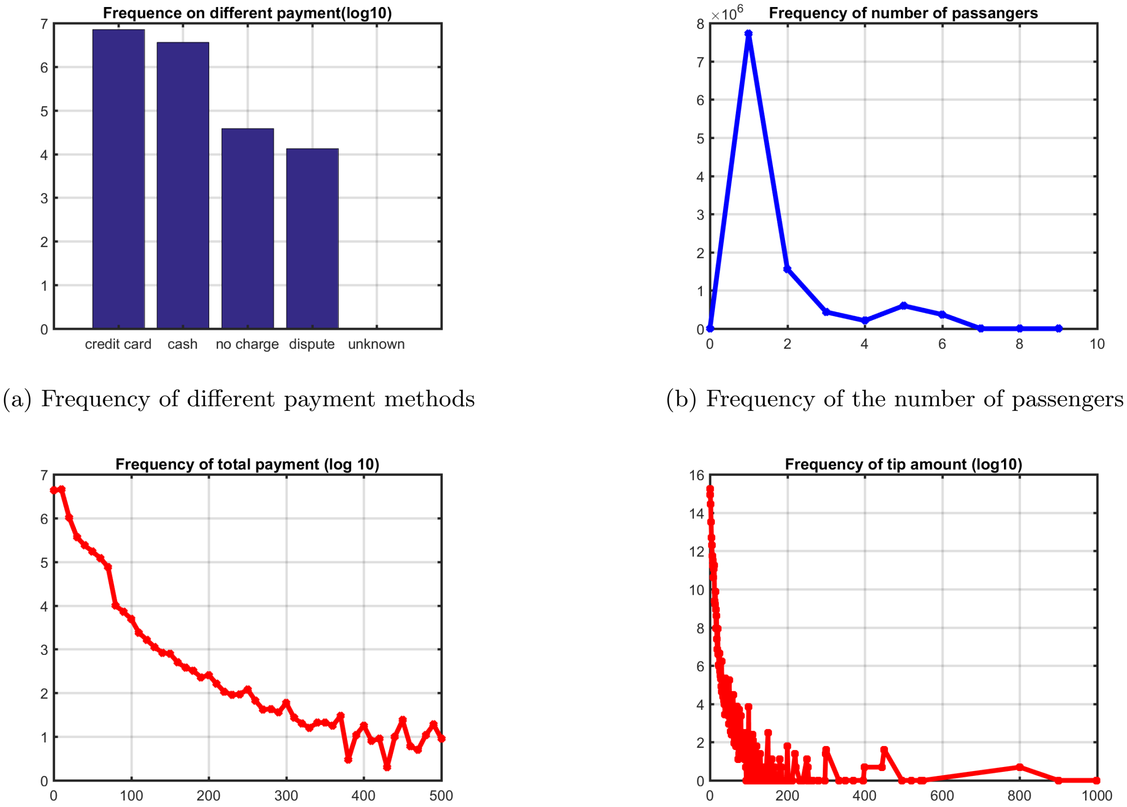


Figure 1: Work Flow

The workflow of our analysis is shown in Figure 1.The data was automatically downloaded as a `.csv` file by a python script we wrote. After that, we created python code to select the columns of interest, map the location information(recorded by latitude and longitude) into a geohash string, create frequency table for each location, and translate the given timestamp into something more helpful for analysis (which we found to be Unix time). Of all the columns from the original data, we focused on pickup/drop off locations, pickup/drop off time, and the dollar amounts for fare/tip/total price.

3 Statistic Analysis

3.1 Statistics from the Data



(a) Frequency of different payment methods

(b) Frequency of the number of passengers

(c) Frequency of the total payment amount

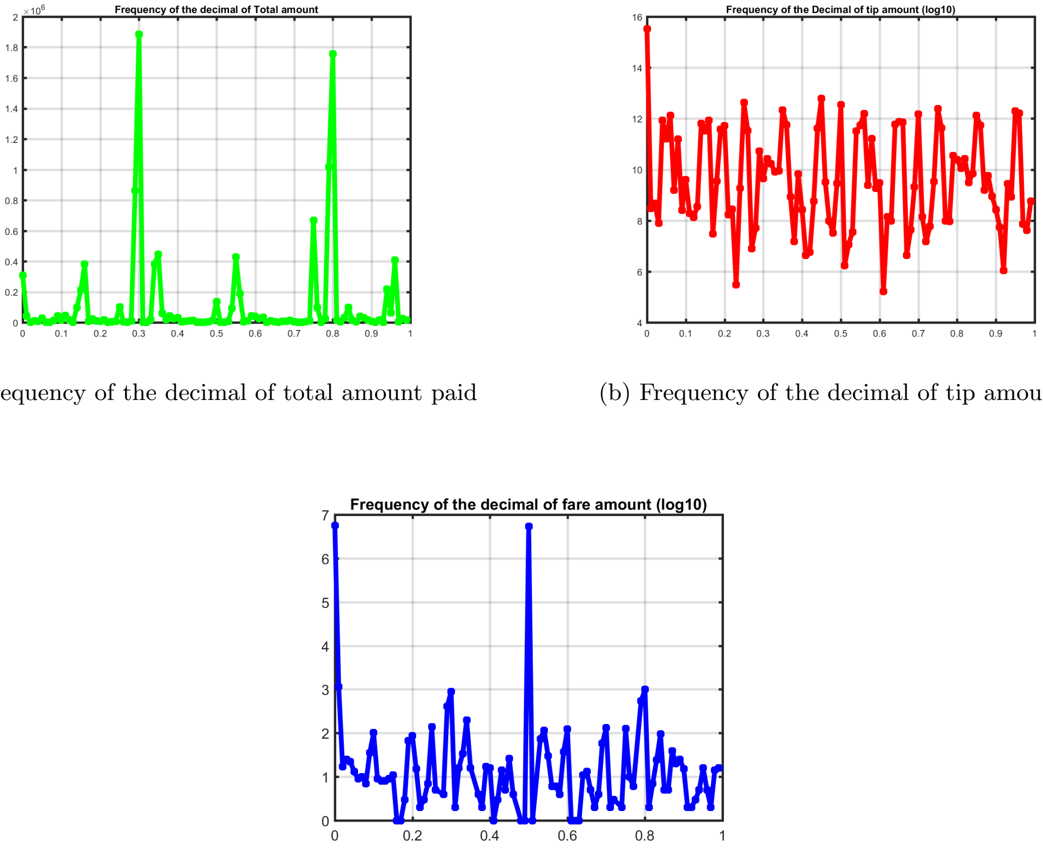
(d) Frequency of the tip amount

Figure 2: Basic ride statistics

Figure 2 shows some statistics from the Yellow Taxi Dataset [1], from which we have found some anomalous characteristics. 2a shows the frequency of payment methods used. Credit card was used the most frequently. No charge and disputed prices also occurred at a notable frequency. 2b shows the frequency of the number of passengers. From the subplot, we most of the case there was only one passenger in the taxi. 2c shows the frequency of the total amounts charged to passengers. It is clear that as the total amount paid increases, the frequency decreases.

3.2 Analysis of the Frequency of Certain Decimal Places in the Price

The prices of the taxi rides have peculiar distributions that show us a little about human psychology and a lot about the way that the taxi pricing scheme works. From Figure 3b we can see that the that tips are periodic fluctuations is frequency at numbers that are divisible by 5 and the fare amount (pre-tax / other charges) is divisible by $0.50 over 99.9% of the time (New Yorkers like round numbers). However, we noticed that the total price paid (Figure 3a) is either $0.30 and $0.80 in over 99.9% of taxi rides. We originally thought this must be the MTA Tax that started being levied in 2015, but 99.5% of passengers pay a $0.50 MTA tax. After a good bit of head scratching, we noticed that 99.95% of passengers pay a $0.30 ”improvement surcharge” and thus puts the final price at the odd $0.30 or $0.80.



(a) Frequency of the decimal of total amount paid

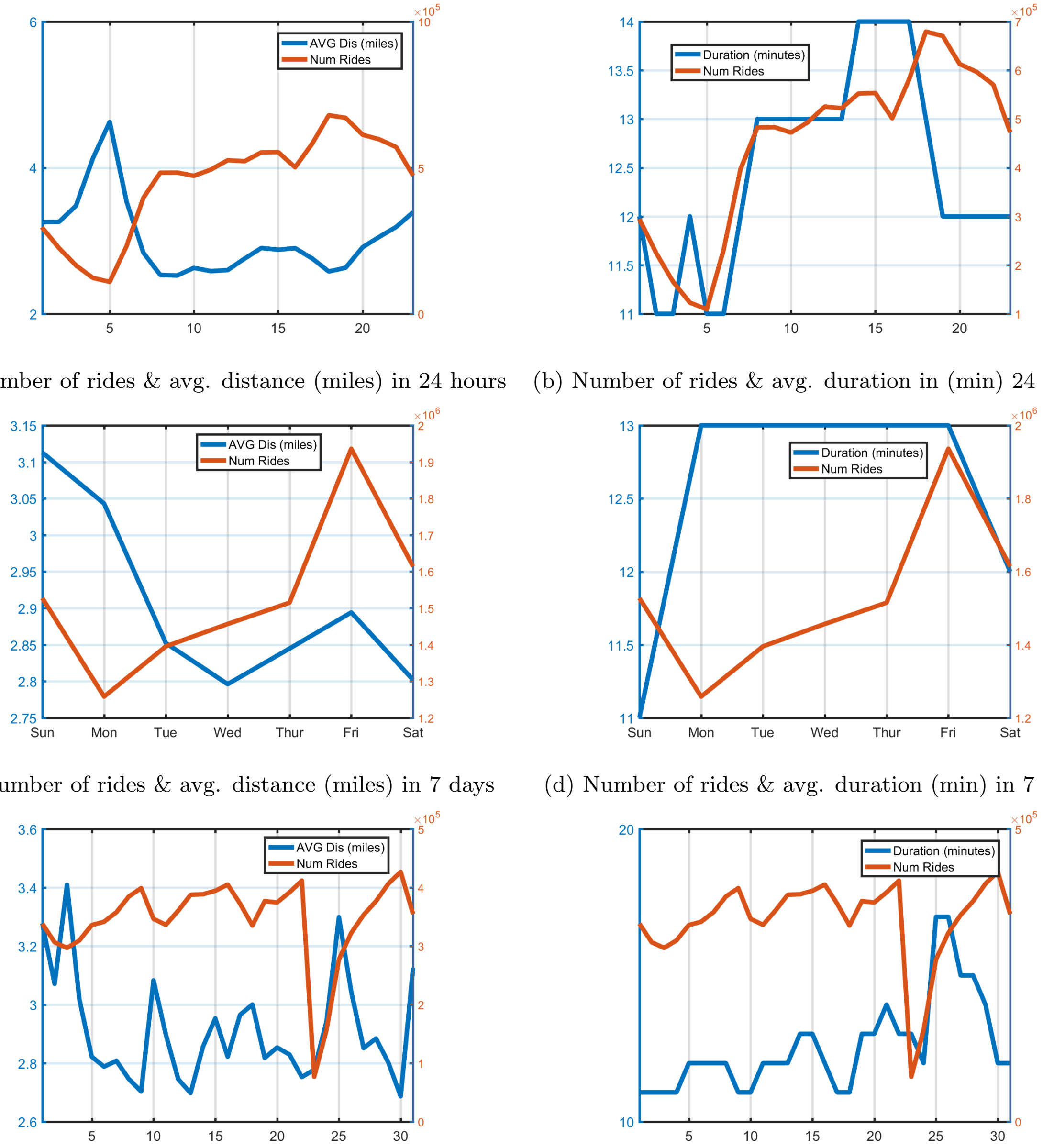
(b) Frequency of the decimal of tip amount

(c) Frequency of fare amount decimal

Figure 3: Frequency of decimal amounts in various parts of the fare

3.3 Changes in Ride Characters based on Time of Ride

Figure 4 shows the fluctuation of number of rides, average distance of rides and average duration, with respect to the 24 hours in a day, 7 days in a week and 31 days in a month. Our data is from January 2016, so the monthly view shouldn’t be taken as an aggregate representation about rides at certain times of the month, but rather a case study of the rides done in January, 2015. In the figure, each subfigure contains two lines with different Y scales. The left column is the numbers of rides or average distance, whereas the right column is number of rides or average duration. The three rows correspond to hours in a day, days in a week and days in a month.



(a) Number of rides & avg. distance (miles) in 24 hours

(b) Number of rides & avg. duration in (min) 24 hours

(c) Number of rides & avg. distance (miles) in 7 days

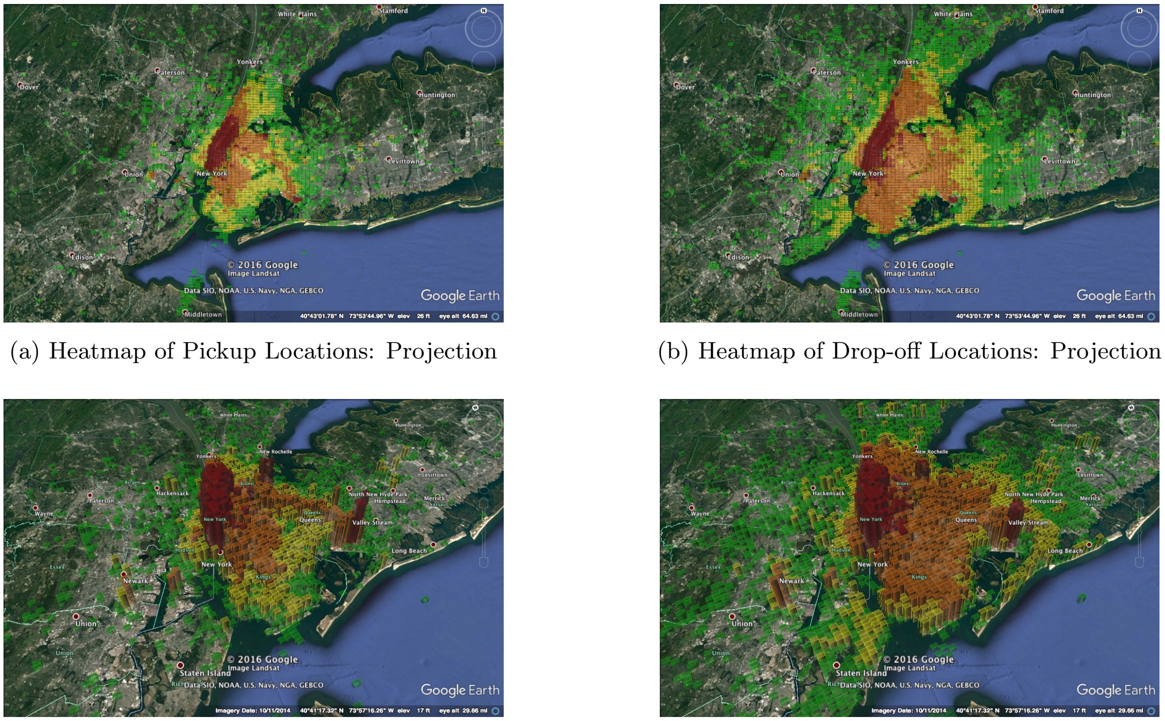
(d) Number of rides & avg. duration (min) in 7 days

(e) Number of rides & avg. distance (miles) in 31 days

(f) number of ride & average duration (min) in 31 days

Figure 4: Statistics from number of rides, distance and duration

Column-wise, the average distance shows negative correlation with rides numbers whereas the average duration shows positive relation with number of rides. It is intuitive that with more rides, the traffic is heavier thus increases average duration. However, it is hard to interpret the negative correlation between number of rides and the average distance. Another interesting fact is the plunge on riders around Jan. 23, in the plot of days in month (e,f). This was due to the blizzard that happened on January 22nd, 2016. It took a couple of days for the number to recover. The average distance and average duration also spiked in the days after the blizzard. We hypothesize that some people were displaced / stranded because of the snow and were finally able able to return home, thus increasing average utilization of the taxis. Where utilization is the ratio of the time that all taxis are spent with and without passengers in them.



(a) Heatmap of Pickup Locations: Projection

(b) Heatmap of Drop-off Locations: Projection

(c) Heightmap of Pickup Locations: 3D

(d) Heightmap of Dropoff Locations: 3D

Figure 5: Maps of Activity by type

4 Geographical Visualization

Given that our dataset is describing how New Yorkers got around, we thought it proper to analyze the frequency of rides as they relate to each other geographically. Figure 5 shows heat maps of the pickup and drop-off locations for the yellow taxi rides from January 2016. Figure 5a and Figure 5c are heatmap of pickup frequencies in the locations, generated from different view angle. And Figure 5b and 5d are for dropoff locations. From the figures, we see that rides around Manhattan are significantly more popular, as most yellow taxis will immediately return after making a drop-off outside the zone. However, as you can see from Figure 5b, the drop-off area for rides is far wider than that of pickups because people can only get a yellow taxi while down-town.

There are four notable zones of high ride volume: Manhattan and JFK / Laguardia / Newark Airports. Newark airport sees considerably more drop-offs than pickups, while the numbers are fairly similar for Laguardia and JFK. It is unclear from this dataset whether fewer people take a taxi back from Newark than take a taxi to be dropped off or whether there are only fewer yellow taxis that pickup passengers there (for reasons discussed in above paragraph).

5 Future Plans

For next step, we will try to combine multiple attributes and analyze the relationship between them. For example, it will be really interesting to combine the location information with time, to investigate the taxi popularity in different regions with regard to time. We could also look into the tip amount with respect to travel duration, distance and time. These combination of different attributes will help us understand human behaviors. Moreover, with the heatmap visualization, the temporal change of distribution could be easily presented and perceived. A .gif image could be used to show the longitudinal evolution of the changes.

References

[1] TLC [Taxi and Limousine Commission] Trip Record Data. `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`, 2009.

[2] Data Dictionary —Yellow Taxi Trip Records . `http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf`, 2015.