NYC Taxi Data Mining

Daniel Monahan

Background

New York City is famous for it's yellow taxi cabs that dominate the streets. Both a common way for tourists and residents to travel, as well as the means of making a living for thousands of people, taxis are an integral part of New York's infrastructure and economy. Taxis in New York City are licensed by the Taxi and Limousine Commission of New York, which issues medallions that allow an owner to operate a taxi cab. There are currently 13,437 medallions issued, which are operated by over 50,000 drivers and serve 600,000 passengers per year. (www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf).

The goal of this project is to analyze a dataset of taxi trips to look for trends that could improve the efficiency of the taxi service, such as whether certain locations often require larger vehicles to accommodate more passengers, or if certain locations or times of day have longer or shorter trips than average. We also wish to examine if the time the trip takes can be predicted accurately with knowledge of the pickup and dropoff locations as well as with the time of day.

The Data

The data used here was originally obtained from the NYC Taxi and Limousine Commission through a request made through the Freedom of Information Law by Chris Whong and made available online (http://chriswhong.com/open-data/foil_nyc_taxi/).

The data totals 11GB, and contains records for taxi trips made in 2013 stored in csv files. For each trip, information is recorded that includes identifying information about the taxi, including the medallion number and hack license number. It also contains information about the trip itself: The pickup coordinates, dropoff coordinates, distance, time spent, pickup time and date, and dropoff time and date.

Initial Data Analysis

The first step of working with the data involves filtering some of it out. I first filtered out any records that had a latitude or longitude out of the general range of NYC, in this case Latitudes between 39 and 42, and Longitudes between -72 and -76. This is to filter out any error codes

or other nonsensical values occasionally found in the dataset. We also only worked on a small subset of the data, ranging from January 1st, 2013 - January 20th, 2013, to enable fast computation. We wish to further examine with more data whether trends still hold.

The following were some basic statistics of the data:
Number of records: 13,714,176

Number of unique medallions: 9244

Max Trip Distance: 95.85 miles

Max Trip TIme: 10,800 seconds

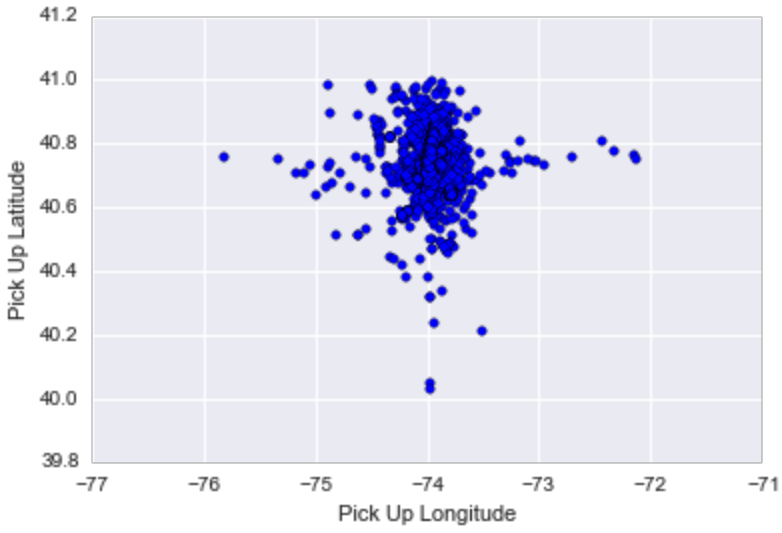
Mean # of passengers: 2.129

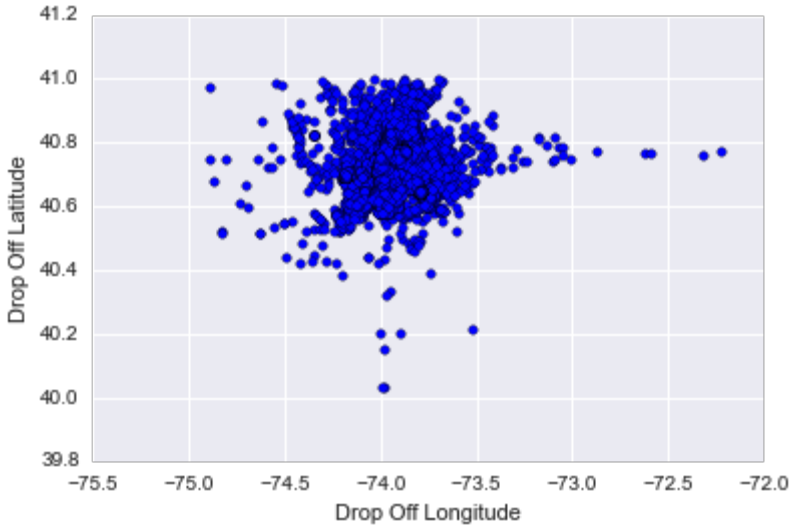
Mean distance: 2.895

Mean trip time: 691.155 seconds

Mean hour of the day (0-23): 12.996

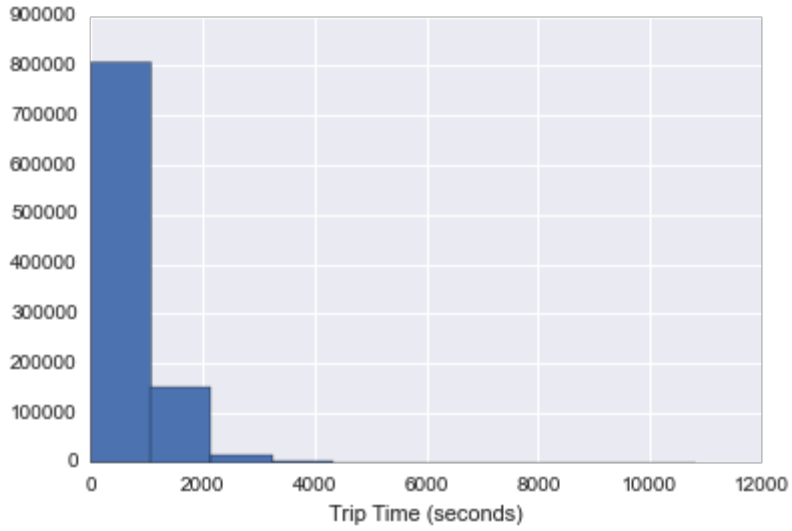
Here we plot the pick up and drop off latitude and longitude of each record:





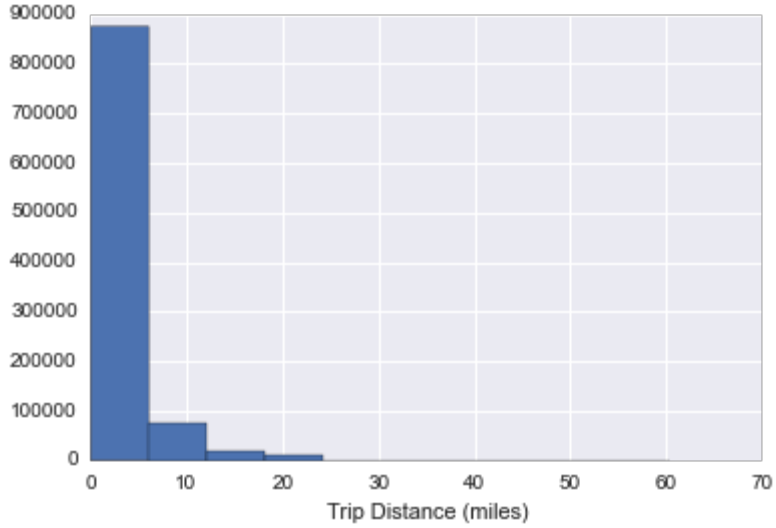
We see than the vast majority of pickups are concentrated within the dense center that is probably Manhattan and the immediate surrounding areas. Interestingly, we see that the drop off distribution is much less densely packed - suggesting that riders often take taxis from the dense center of the city to the outer parts, but not the other way around.

Here we plot the number of records with their trip time.

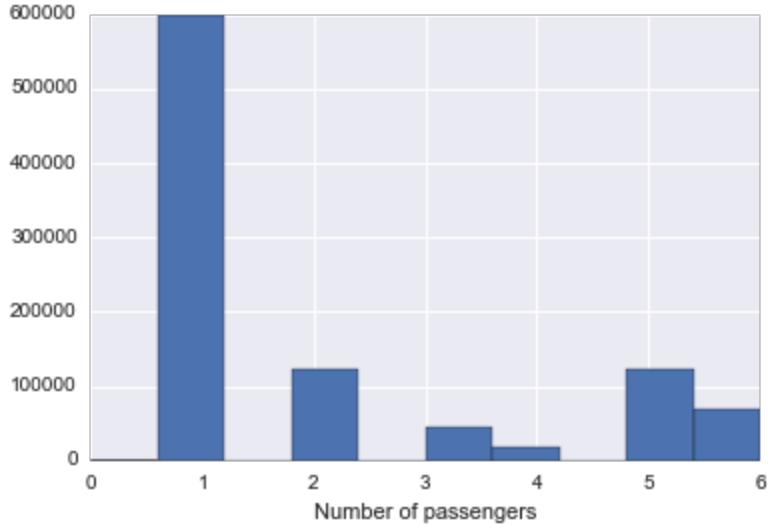


Here we see that the vast majority of trips are under 1000 seconds long, which is about 16 minutes and 40 seconds. We also see that there is a relatively steep drop off, with some between 1000 and 2000 seconds long, very few between 2000 and 3000, and then almost none longer than 3000 seconds.

The graph for trip distance looks very similar to the above time graph which follows logically that the time and distance of a trip should be closely related.

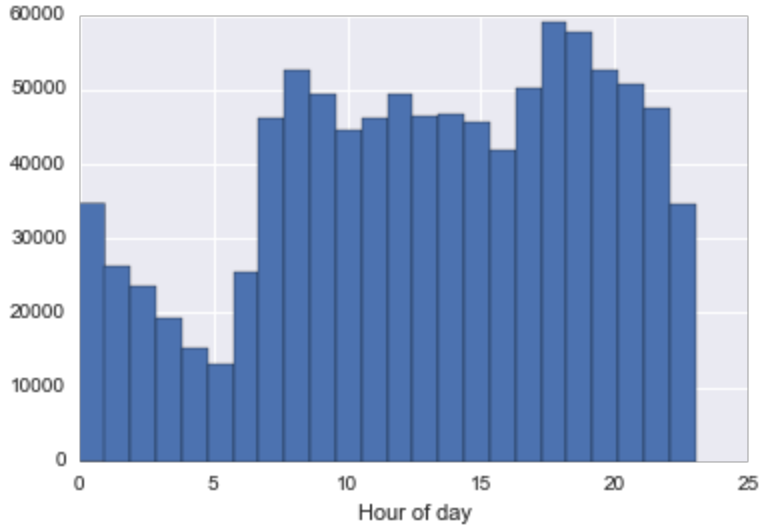


The graph for number of passengers:



Interestingly, while most trips appear to be for a single passenger, the number of trips transporting 2 passengers is approximately the same as the number of trips transporting 5 passengers, while there are relatively few trips carrying 3 or 4 passengers.

We also have the hour of day distribution of trips, ranging from 0(midnight) to 23(11 pm):



Note the slight dip at 16, indicating trips from 4pm to 5pm, which has been documented in other studies.