Are there some areas in NYC, traveling from which people pay more tips?

Eric Gleiß

February 21, 2016

1 Preprocessing

I focused my analysis on the data file *yellow_tripdata_2015-06.csv*. The upper panel of fig. 1 shows a histogram of all tips in the dataset. There are are some tip values that seem to be errors in the data set, e.g. values below zero. On the other hand, there are also extremely high tips occuring, which is not necessarily an error in the data set. However, I will not include negative tips nor extremely high tips, since tips above, say, 100$ occur only very rarely in the dataset and don't bear much statistical significance. Furthermore, the dataset includes only credit card tips and no cash tips. Therefore, only datapoints with "credit card" as payment type are considered.

It might also be interesting to look at the location of the pickup points in the data set. Since the area of interest is NYC, data points lying too far outside of this area are not considered (most likely they are either errors in the dataset or don't hold much statistical significance anyway). The following coordinate frame roughly matches the extension of NYC boundaries:

$$\text{latitude: } 40.48°N \text{ to } 40.93°N$$
$$\text{longitude: } 74.26°W \text{ to } 73.69°W$$

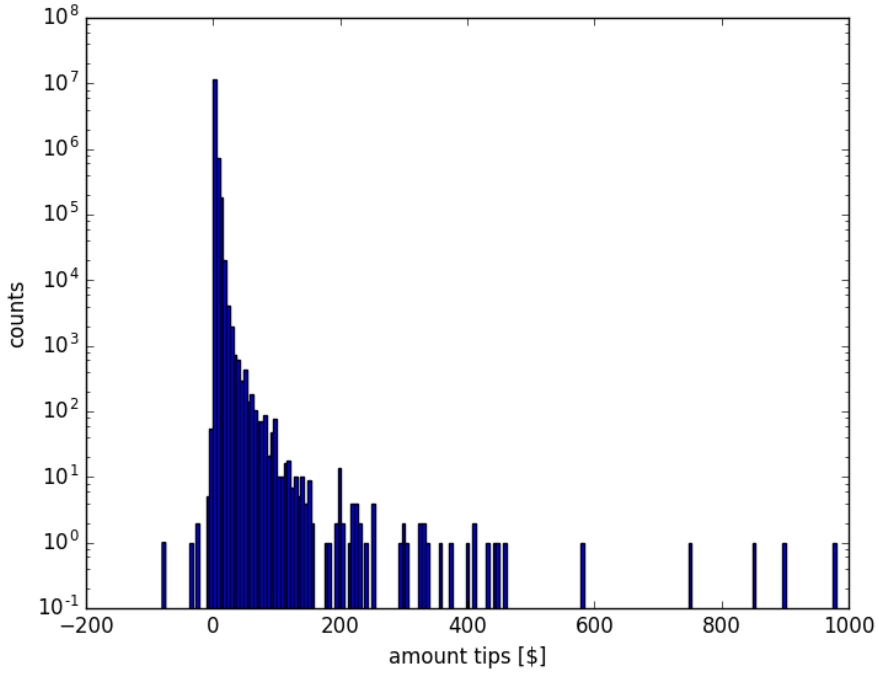
The lower panel of fig. 1 gives a first overview of all pickup points of the data set in the given coordinate frame. The layout is taken from `http://www.danielforsyth.me/mapping-nyc-taxi-data/`. The density of pickup points is by far the highest in Manhattan. The other boroughs of NYC - Bronx, Brooklyn, Queens and Staten Island - exhibit a much lower Taxi activity. Particularly high Taxi activity can be found around the major airports - JFK, La Guardia and to a lesser extent Newark Int. Airport/New Jersey. It could be interesting to investigate, if this heterogeneous distribution of Taxi activity correlates with the amount of tips paid. The preprocessing filters altogether lead to a dataset of 7604693 datapoints out of originally 12324935 datapoints.

2 Regions with Higher Tips

The upper panel of fig. 2 shows a two-dimensional histogram depicting the average amount of tips paid in the area of each of its hexagonal bins. In order to increase the statistical significance, bins with fewer than 50 counts are not depicted. Mean and standard deviation of all tips in the dataset have been determined:

$$\text{average tip: } 2.76\$$$
$$\text{standard deviation: } 2.79\$$$

The plain histogram shows that tips are particularly high around the areas of the major airports - JFK, La Guardia and to some extent Newark Int. In order to define, what counts as significantly higher tip compared to the average tip, one could set



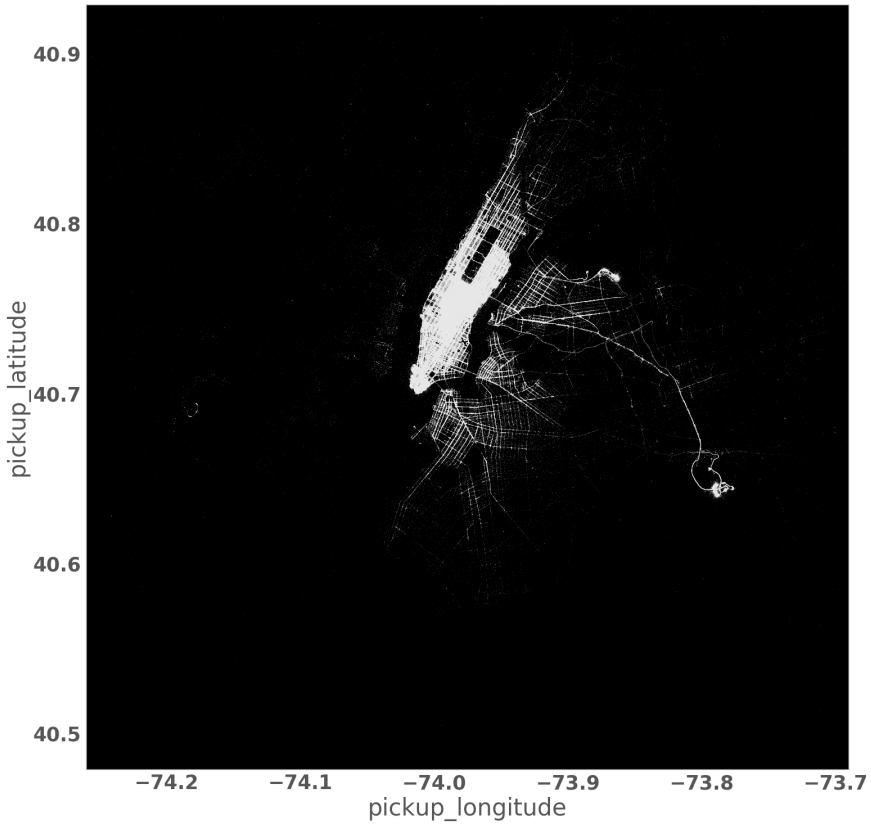


Figure 1: The upper panel shows a 1d histogrm of tip values in the data set. The lower panel shows a map of pick-up points within the data set. The highest densities of pick-up points are in Manhattan and around airports.

a threshold. A natural threshold would be the standard deviation of the tips paid. The lower panel of fig. 2 highlights all bins, where the tip amount is higher than the average tip of the dataset plus one standard deviation. Besides the airports, there are some other locations where tips are above the threshold.

3 Possibilities for Further Investigation

The simple threshold method was able to reveal some areas where significantly higher than average tips are paid. This was most significant around the airports. It is not obvious to me, why people travelling from the airport would pay more tips (especially since these tips are paid by card and not cash). One possibility could be that more passengers share a taxi and thereby pay more tips. Using the dataset, it could be investigated if taxis with more passengers receive more tips and if taxis starting from the airports carry more passengers simultaneously. Another possible explanation could be that people travelling by plane generally plan to spend more money on the trip anyway and thus pay more tips to taxi drivers. Furthermore, taxi rides from the airports may be longer and therefore more tips may be paid. Besides the airports, there are a few other spots where the amount of tips is higher. It could be interesting to investigate if these geographic locations are in some way special, similar to the airports.

Finally, the dataset also offers a time resolution, which was not used at all in this analysis. Neither made I use of additional datasets from the other months or years. Using these could give additional insights.

Further limitations of my results may be due to the fact that only one specific taxi line is considered, which does not necessarily distribute its coverage homogeneously, i.e. the taxi line is already optimised in some way.

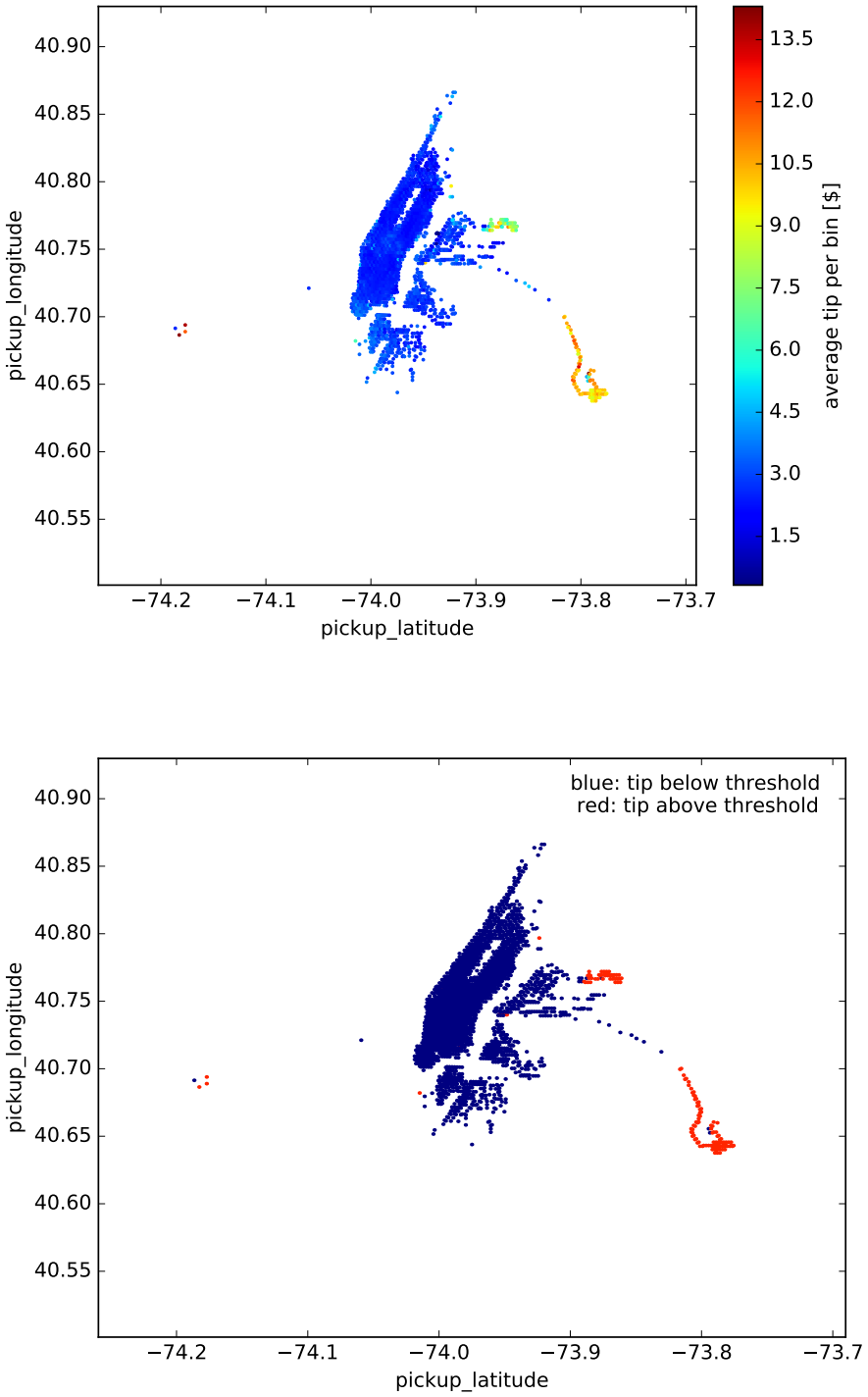


Figure 2: The upper panel shows a 2d histogram of tips. Each hexagonal bin depicts the average amount of tips within its area. The lower panel shows the same histogram where all bins with average tips below a threshold (mean tip value of the whole dataset plus one standard deviation) are depicted in blue and bins with average tips above the threshold are depicted in red.