New York City Taxi Data (2010-2013)

Brian Donovan and Dan Work
December, 2014

This dataset was obtained through a *Freedom of Information Law* (FOIL) request from the *New York City Taxi & Limousine Commission* (NYCT&L). It covers four years of taxi operations in New York City and includes 697,622,444 trips. Thanks to a generous hosting policy by the University of Illinois at Urbana Champaign, we are able to make this large dataset publicly available under the CC0 license.

You are free to use the data as you wish, we only kindly ask you to consider citing the following works if you plan to publish subsequent results using the dataset:

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.* preprint, source code.

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.

Download the data here: http://dx.doi.org/10.13012/J8PN93H8

The data is stored in CSV format, organized by year and month. In each file, each row represents a single taxi trip. Table 1 below gives a small sample of this data. As there are several entries per second for four years, the raw trip data takes up about 116GB in text CSV format. The data has been compressed (zip) to reduce download time.

The data is organized as follows:

medallion: a permit to operate a yellow taxi cab in New York City, it is effectively a (randomly assigned) car ID. See also medallions.
hack license: a license to drive the vehicle, it is effectively a (randomly assigned) driver ID. See also hack license.
vender id: e.g., *Verifone Transportation Systems* (VTS), or *Mobile Knowledge Systems Inc* (CMT), implemented as part of the Technology Passenger Enhancements Project.
rate_code: taximeter rate, see NYCT&L description.
store_and_fwd_flag: unknown attribute.
pickup datetime: start time of the trip, mm-dd-yyyy hh24:mm:ss EDT.
dropoff datetime: end time of the trip, mm-dd-yyyy hh24:mm:ss EDT.
passenger count: number of passengers on the trip, default value is one.
t**rip time in secs:** trip time measured by the taximeter in seconds.
trip distance: trip distance measured by the taximeter in miles.
pickup_longitude and pickup_latitude: GPS coordinates at the start of the trip.
dropoff longitude and dropoff latitude: GPS coordinates at the end of the trip.

The medallion and hack licenses are reassigned each year, so it is only possible to track drivers and vehicles within each year. This is necessary for to render the data pseudo-anonymous, since de-anonymized data from 2013 can be reconstructed from existing published datasets, see the note on anonymity below.

Please note that the dataset contains a large number of errors. For example, there are several trips where the reported meter distances are significantly shorter than the straight-line distance, violating Euclidean geometry. For some periods, the field trip_time_in_secs is reported in

seconds, in others it is reported in minutes (see the first record above). Generally the trip time can be safely computed by subtracting the pickup_datetime from the dropoff_datetime. Additionally, many trips report GPS coordinates of (0,0), or cover impossible distances, times, or velocities. All of these types of obvious trip errors should be discarded in any analysis. In our preliminary investigations, **these errors account for roughly 7.5% of all trips.** More details about these errors are available in the above article and corresponding open source code. Currently, only the raw data (no error filtering) is available for download via this site.

Fare data is also available from 2010-2014. The fare data takes about 75GB in raw text CSV format, and is also zipped to reduce download times. A sample of the fare data is shown in Table 2 below. The files are also organized by year and month, and contain the following attributes:

medallion: a permit to operate a yellow taxi cab in New York City, it is effectively a (randomly assigned) car ID. See also medallions.
hack license: a license to drive the vehicle, it is effectively a (randomly assigned) driver ID. See also hack license.
vender id: e.g., *Verifone Transportation Systems* (VTS), or *Mobile Knowledge Systems Inc* (CMT), implemented as part of the Technology Passenger Enhancements Project.
pickup datetime: start time of the trip, mm-dd-yyyy hh24:mm:ss EDT.
payment type: Cash or credit card.
fare amount: the meter fare, it should include the Newark surcharge, in USD.
surcharge: Extra fees, such as rush hour and overnight surcharges, in USD.
mta tax: Metropolitan commuter transportation mobility tax, in USD.
tip amount: tip amount, in USD.
tolls amount: total price paid for tolls, summed across all tolls for the trip, in USD.
total amount: all charges that are presented to the passenger at time of fare payment (includes tip for non-cash trips), in USD.

Again, note the medallion and hack licenses change each year.

Table 1. A small subset of the New York City taxi trip data. Each row corresponds to an occupied taxi trip.

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
2010000005	2010000005	VTS	1		2010-01-01 00:00:00	2010-01-01 00:28:00	2	28	3.15	-73.99955	40.731152	-73.977448	40.763031
2010000006	2010000006	VTS	1		2010-01-01 00:00:00	2010-01-01 00:27:00	1	27	11.15	-73.993698	40.736946	-73.861435	40.756256
2010000007	2010000007	VTS	1		2010-01-01 00:00:00	2010-01-01 00:18:00	3	18	4.30	-74.006058	40.739925	-73.957405	40.765686
2010000008	2010000008	VTS	1		2010-01-01 00:00:00	2010-01-01 00:27:00	1	27	9.83	-73.874245	40.773739	-74.0028	40.760498
2010000009	2010000009	CMT	1	0	2010-01-01 00:00:00	2010-01-01 00:18:13	1	18.219999999999999	3.40	-74.004868	40.751656	-73.988342	40.718399
2010000010	2010000010	CMT	1	0	2010-01-01 00:00:02	2010-01-01 00:36:27	2	36.420000000000002	12.40	-73.95546	40.787731	-73.961739	40.666935

Table 2. A small subset of the New York City taxi fare data. Each row corresponds to an occupied taxi trip.

medallion	hack_license	vendor_id	pickup_datetime	payment_type	fare_amount	surcharge	mta_tax	tip_amount	tolls_amount	total_amount
2010000001	2010000001	VTS	2010-01-01 00:00:00	CAS	34.1	0.5	0.5	0	0	35.1
2010000002	2010000002	VTS	2010-01-01 00:00:00	CAS	27.3	0.5	0.5	0	0	28.3
2010000003	2010000003	VTS	2010-01-01 00:00:00	CAS	6.9	0.5	0.5	0	0	7.9
2010000004	2010000004	VTS	2010-01-01 00:00:00	Cre	56.1	0.5	0.5	10	9.14	76.24
2010000005	2010000005	VTS	2010-01-01 00:00:00	CAS	14.5	0.5	0.5	0	0	15.5
2010000006	2010000006	VTS	2010-01-01 00:00:00	CAS	27.7	0.5	0.5	0	0	28.7
2010000007	2010000007	VTS	2010-01-01 00:00:00	CAS	13.3	0.5	0.5	0	0	14.3
2010000008	2010000008	VTS	2010-01-01 00:00:00	Cre	25.7	0.5	0.5	0	4.57	31.27
2010000009	2010000009	CMT	2010-01-01 00:00:00	Cas	12.5	0.5	0.5	0	0	13.5
2010000010	2010000010	CMT	2010-01-01 00:00:02	Cas	31.7	0.5	0.5	0	0	32.7