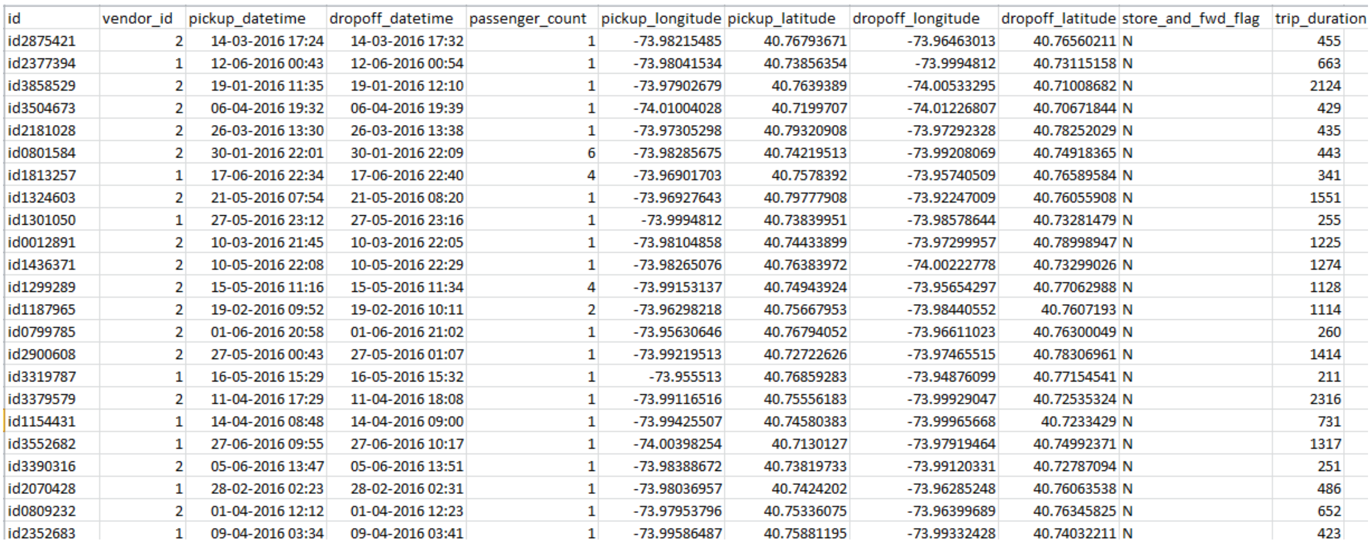
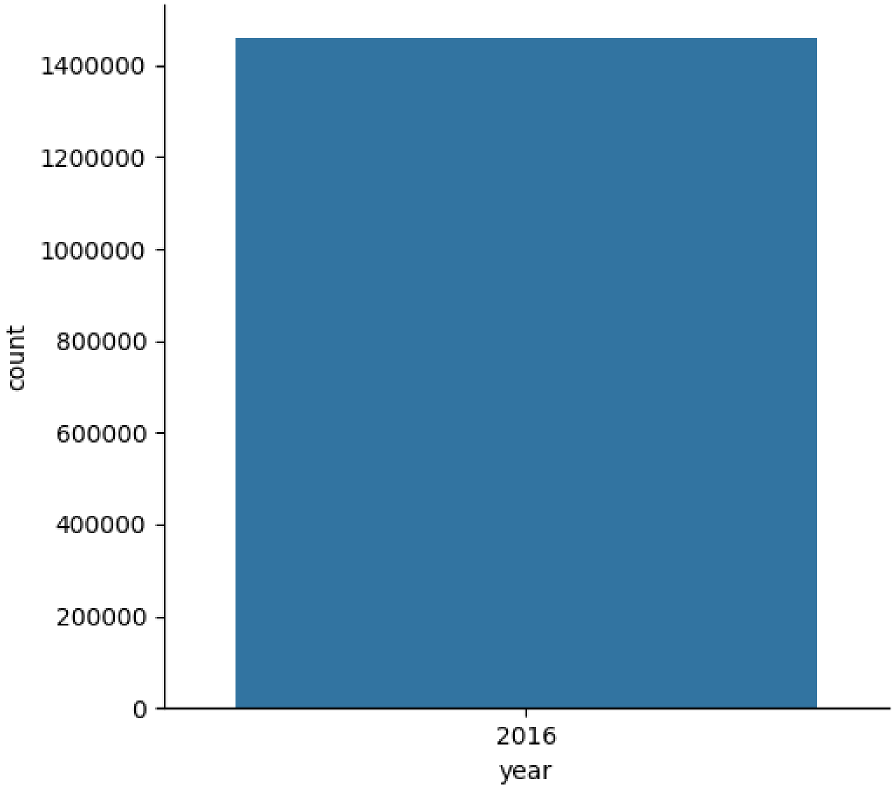
NYC Taxi Data Set



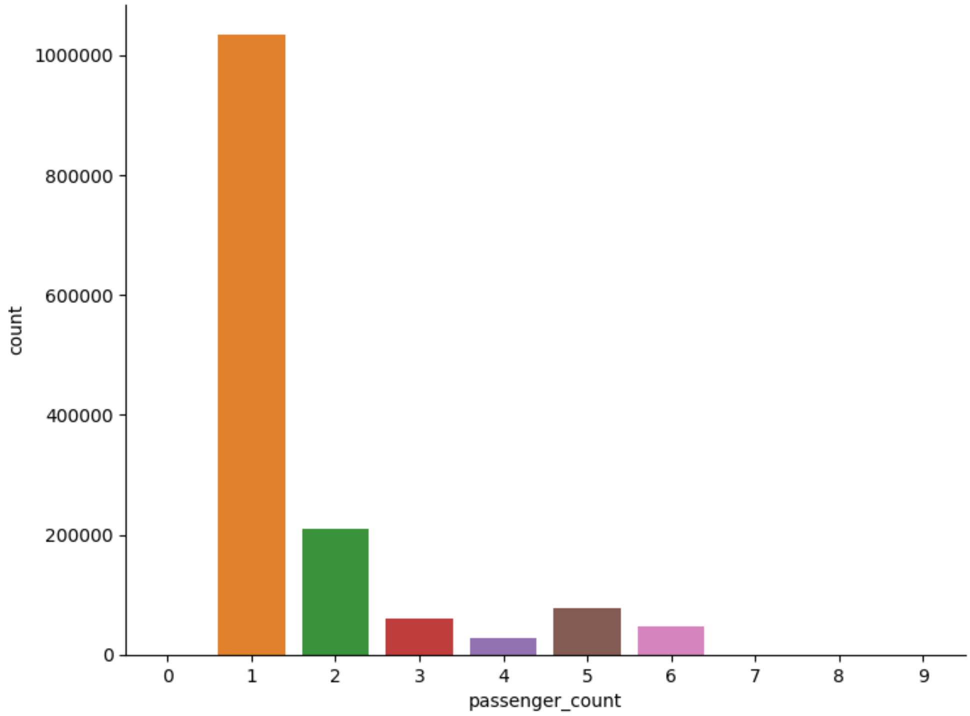
id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration
id2875421	2	14-03-2016 17:24	14-03-2016 17:32	1	-73.98215485	40.76793671	-73.96463013	40.76560211	N	455
id2377394	1	12-06-2016 00:43	12-06-2016 00:54	1	-73.98041534	40.73856354	-73.9994812	40.73115158	N	663
id3858529	2	19-01-2016 11:35	19-01-2016 12:10	1	-73.97902679	40.7639389	-74.00533295	40.71008682	N	2124
id3504673	2	06-04-2016 19:32	06-04-2016 19:39	1	-74.01004028	40.7199707	-74.01226807	40.70671844	N	429
id2181028	2	26-03-2016 13:30	26-03-2016 13:38	1	-73.97305298	40.79320908	-73.97292328	40.78252029	N	435
id0801584	2	30-01-2016 22:01	30-01-2016 22:09	6	-73.98285675	40.74219513	-73.99208069	40.74918365	N	443
id1813257	1	17-06-2016 22:34	17-06-2016 22:40	4	-73.96901703	40.7578392	-73.95740509	40.76589584	N	341
id1324603	2	21-05-2016 07:54	21-05-2016 08:20	1	-73.96927643	40.79777908	-73.92247009	40.76055908	N	1551
id1301050	1	27-05-2016 23:12	27-05-2016 23:16	1	-73.9994812	40.73839951	-73.98578644	40.73281479	N	255
id0012891	2	10-03-2016 21:45	10-03-2016 22:05	1	-73.98104858	40.74433899	-73.97299957	40.78998947	N	1225
id1436371	2	10-05-2016 22:08	10-05-2016 22:29	1	-73.98265076	40.76383972	-74.00222778	40.73299026	N	1274
id1299289	2	15-05-2016 11:16	15-05-2016 11:34	4	-73.99153137	40.74943924	-73.95654297	40.77062988	N	1128
id1187965	2	19-02-2016 09:52	19-02-2016 10:11	2	-73.96298218	40.75667953	-73.98440552	40.7607193	N	1114
id0799785	2	01-06-2016 20:58	01-06-2016 21:02	1	-73.95630646	40.76794052	-73.96611023	40.76300049	N	260
id2900608	2	27-05-2016 00:43	27-05-2016 01:07	1	-73.99219513	40.72722626	-73.97465515	40.78306961	N	1414
id3319787	1	16-05-2016 15:29	16-05-2016 15:32	1	-73.955513	40.76859283	-73.94876099	40.77154541	N	211
id3379579	2	11-04-2016 17:29	11-04-2016 18:08	1	-73.99116516	40.75556183	-73.99929047	40.72535324	N	2316
id1154431	1	14-04-2016 08:48	14-04-2016 09:00	1	-73.99425507	40.74580383	-73.99965668	40.7233429	N	731
id3552682	1	27-06-2016 09:55	27-06-2016 10:17	1	-74.00398254	40.7130127	-73.97919464	40.74992371	N	1317
id3390316	2	05-06-2016 13:47	05-06-2016 13:51	1	-73.98388672	40.73819733	-73.99120331	40.72787094	N	251
id2070428	1	28-02-2016 02:23	28-02-2016 02:31	1	-73.98036957	40.7424202	-73.96285248	40.76063538	N	486
id0809232	2	01-04-2016 12:12	01-04-2016 12:23	1	-73.97953796	40.75336075	-73.96399689	40.76345825	N	652
id2352683	1	09-04-2016 03:34	09-04-2016 03:41	1	-73.99586487	40.75881195	-73.99332428	40.74032211	N	423

The taxi data set contained information about the pickup and drop-off times, pickup and drop-off latitudes and longitudes, passenger counts and trip duration for taxi rides in New York city.

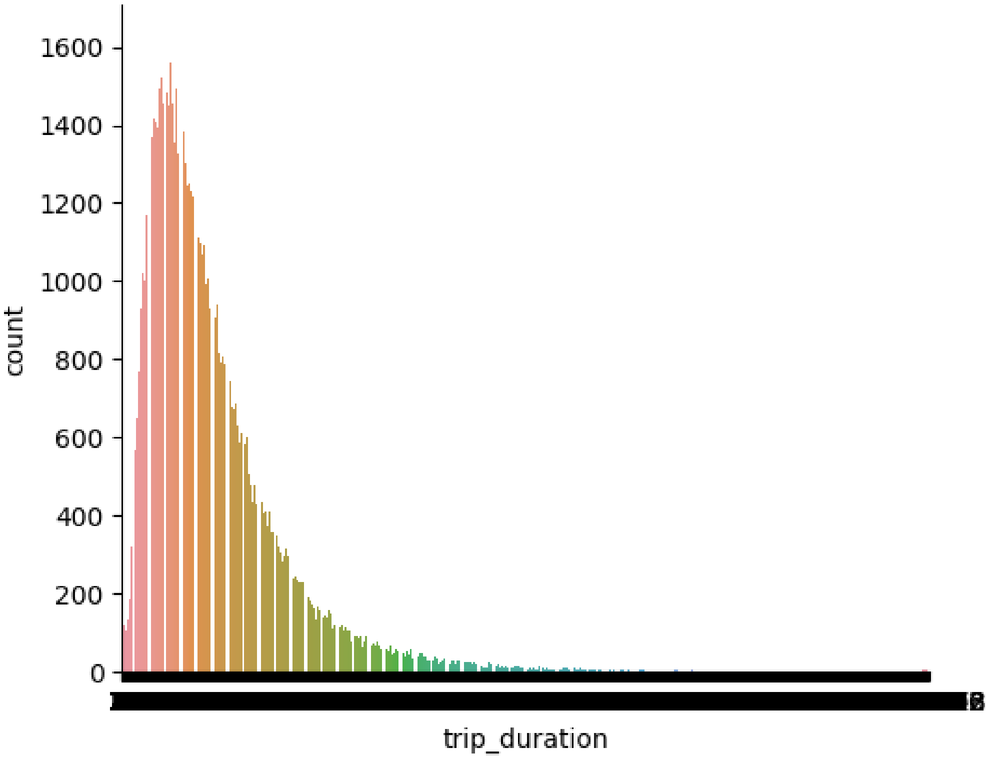
Analysing the data further: the data contained was only for the first 6 months of the year 2016



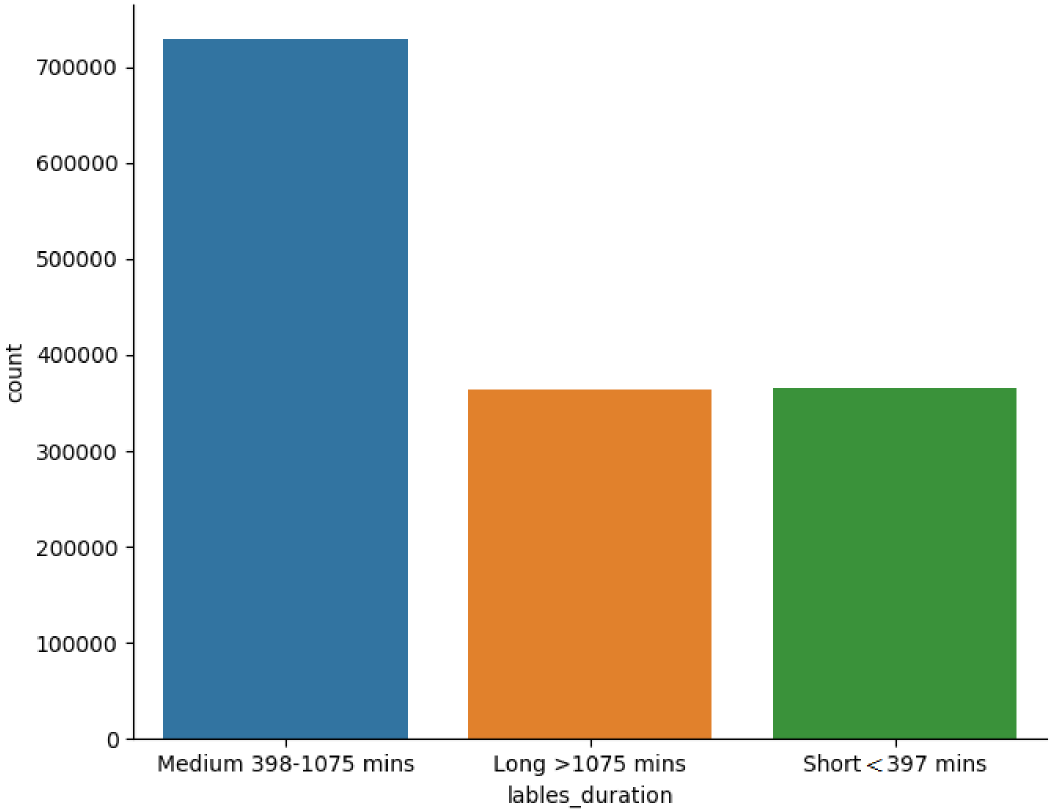
One was the highest passenger count



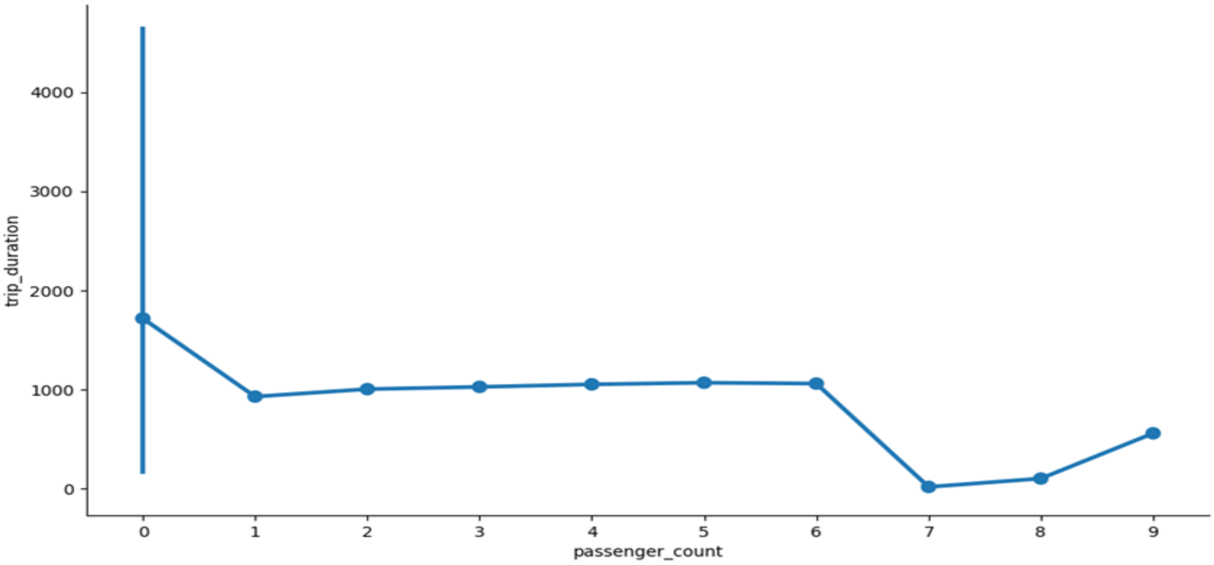
And there were a larger recorded number of comparatively shorter duration trips



Because of the large number of trip durations I classified them into 1) short : less than 397 mis, 2) intermediate: 398 - 1075 mins and 3) long : > 1075 mins, for easy processing

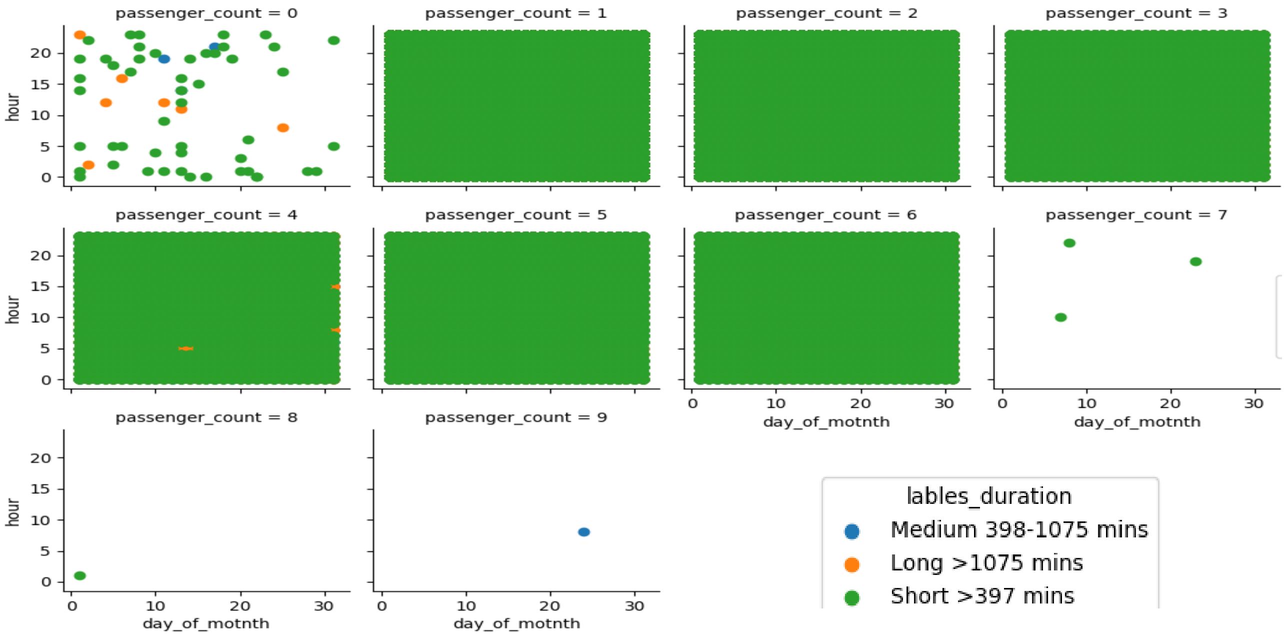


Passenger count and trip duration: most passenger counts have constant trip durations, except for empty cabs whose trip duration seems to vary a lot.

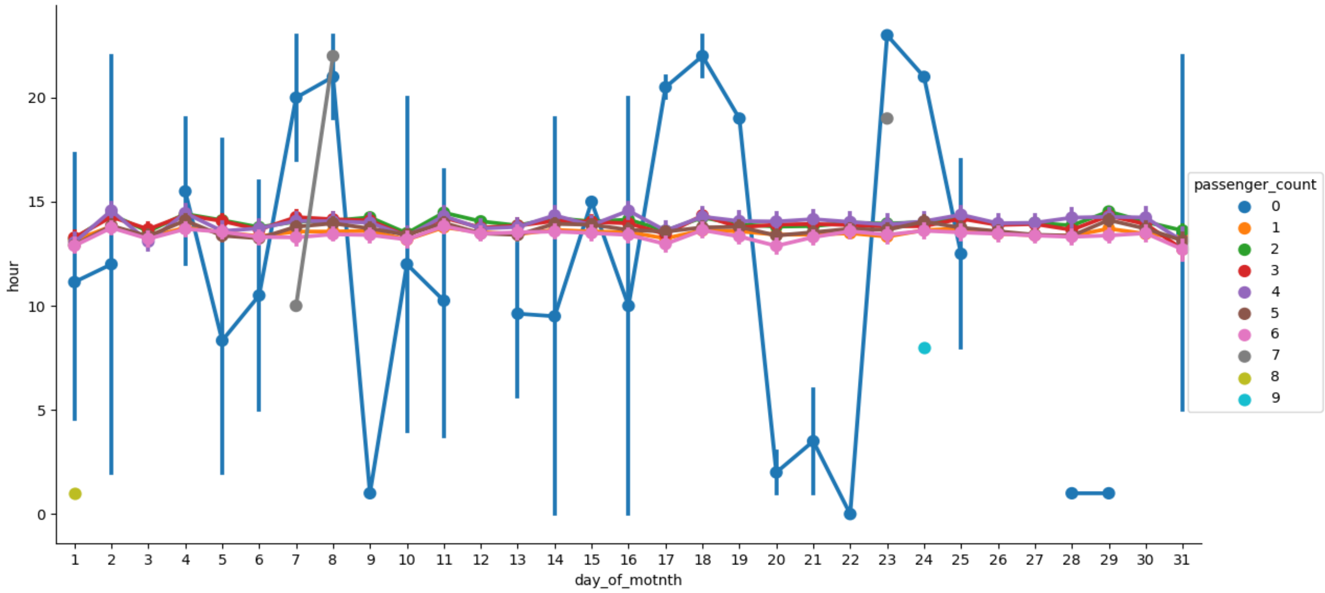


Time at which taxis were used

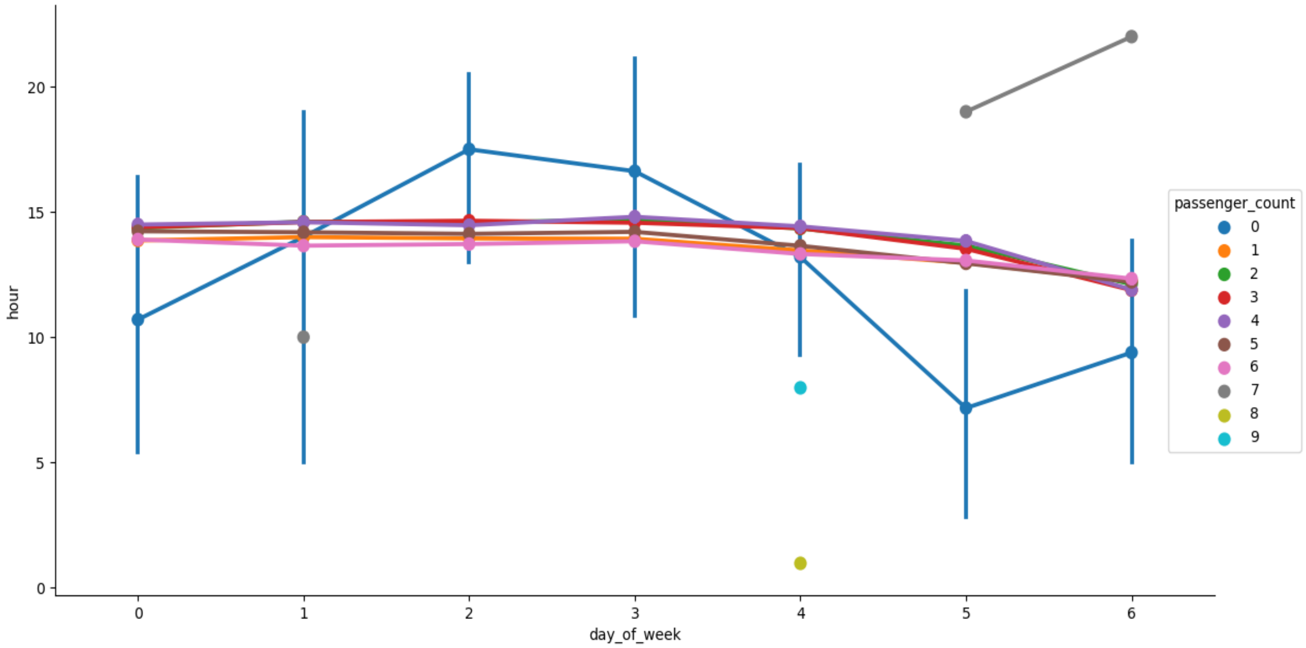
Since most of the data was time series and spatial data, I went on to make many plots with hours on the y axis, and time periods such as day, week, and month on the x axis, the seaborn library used to make these plots uses a mean function and a mean of the data through the day is plotted, this helps us get a trend of the data, the actual data when plotted looks like the plot below, and there is too much data to discern any good trends, **graph** : based on passenger count and trip duration over a month



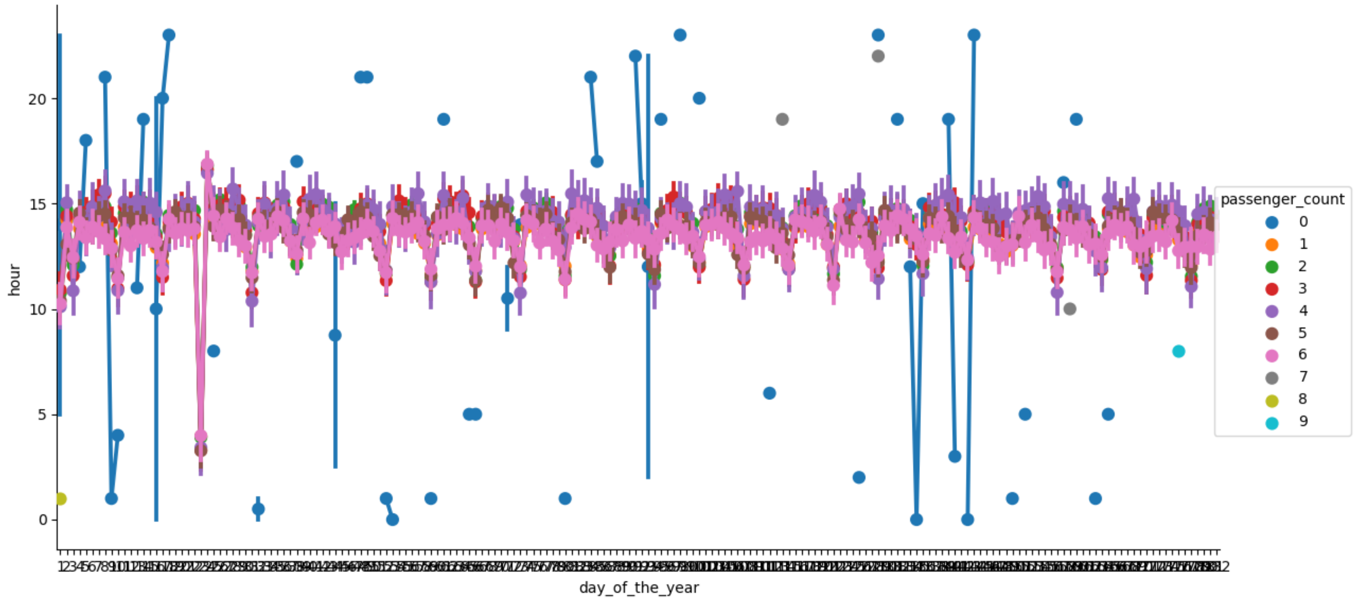
2) **Number of passengers taking taxies throughout the day over a month**: the data is spread through the day the graph represents the mean hour, all the passengers take taxis throughout the day, the time at which passengers take taxis decreases and increases throughout the month.Empty taxis are usually flying about either earlier or later



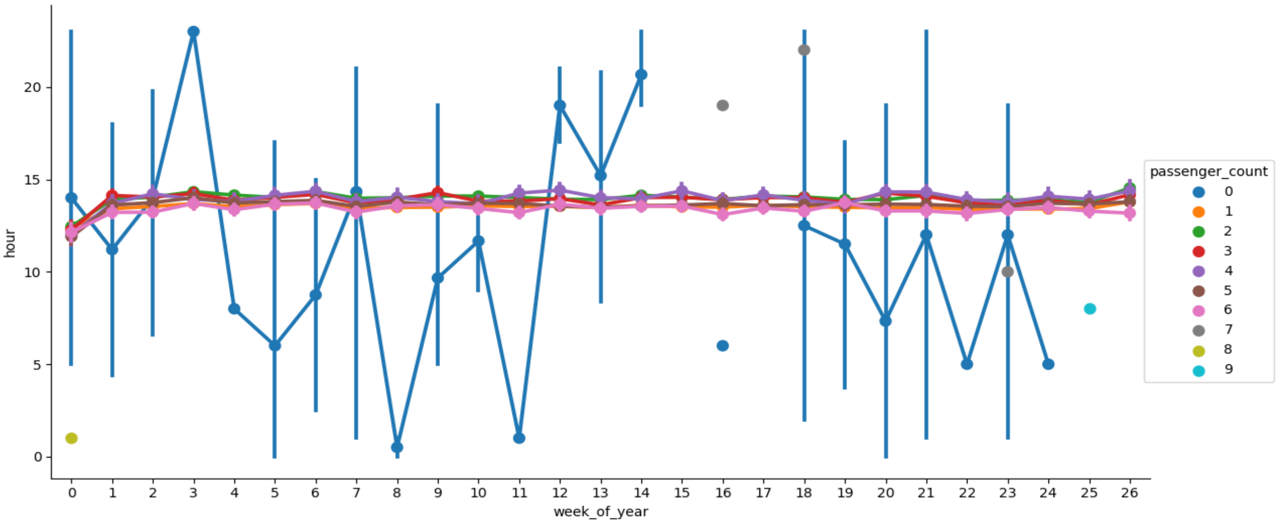
3) Passengers taking taxis throughout the week: the passengers usually take late taxis throughout the week, and tend to take earlier taxis around the weekend, probably because of an early end to their work day, while passengers in a group of 7 tend to take a late taxi around the week end, probably carpooling to parties.



3) **Passengers taking taxis throughout the 6 months**: The data contains data only for the first 182 days; the time at which passengers take taxis throughout this time follows the same trend as the trends over the month and the week, with early taxis being taken over the weekends, except for somewhere around the third week where there was a sharp rise in the number of early taxis followed by a sharp rise in the number of late taxis

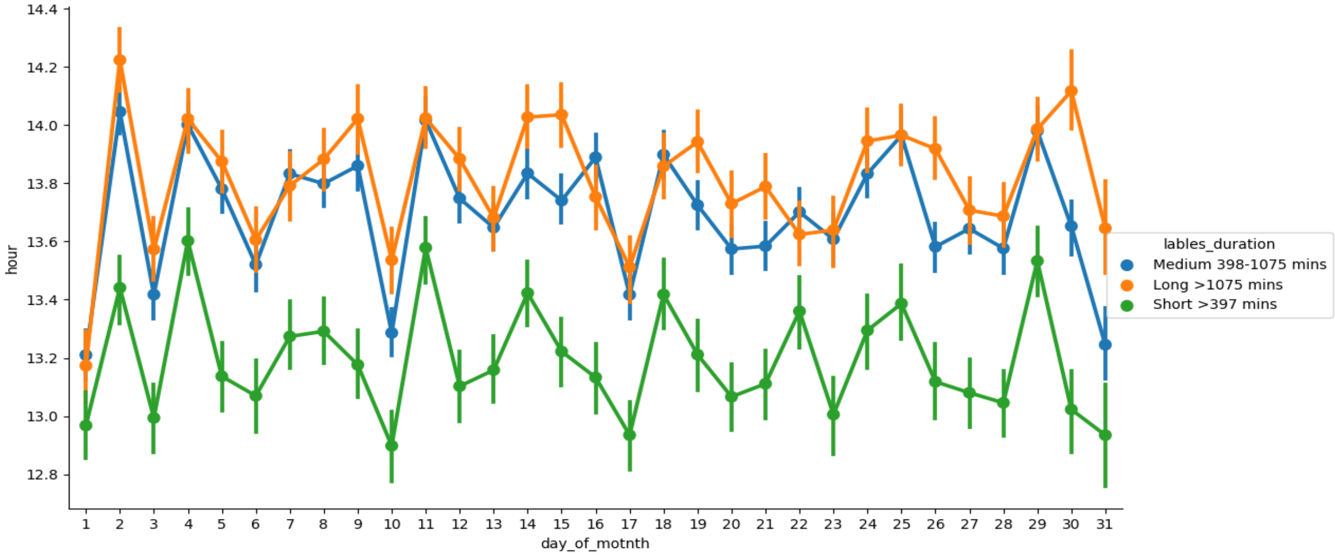


4) **Passengers taking taxis averaging over a week through the year**: There is data only for the first 26 weeks; the time at which taxis are taken follows the same trend, except in the beginning of the year where taxis are taken earlier.

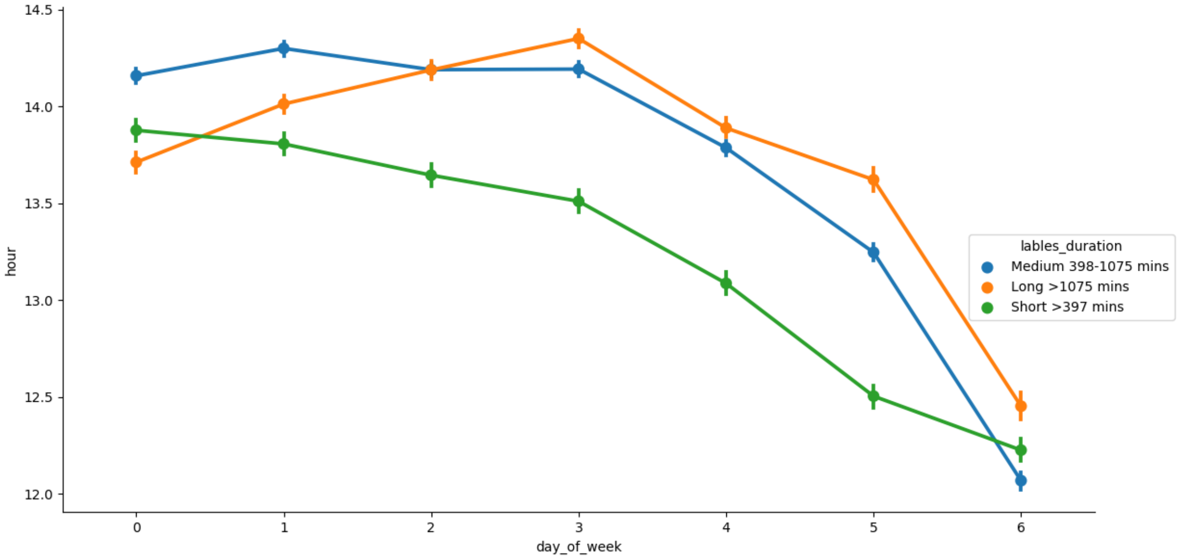


Trip Durations

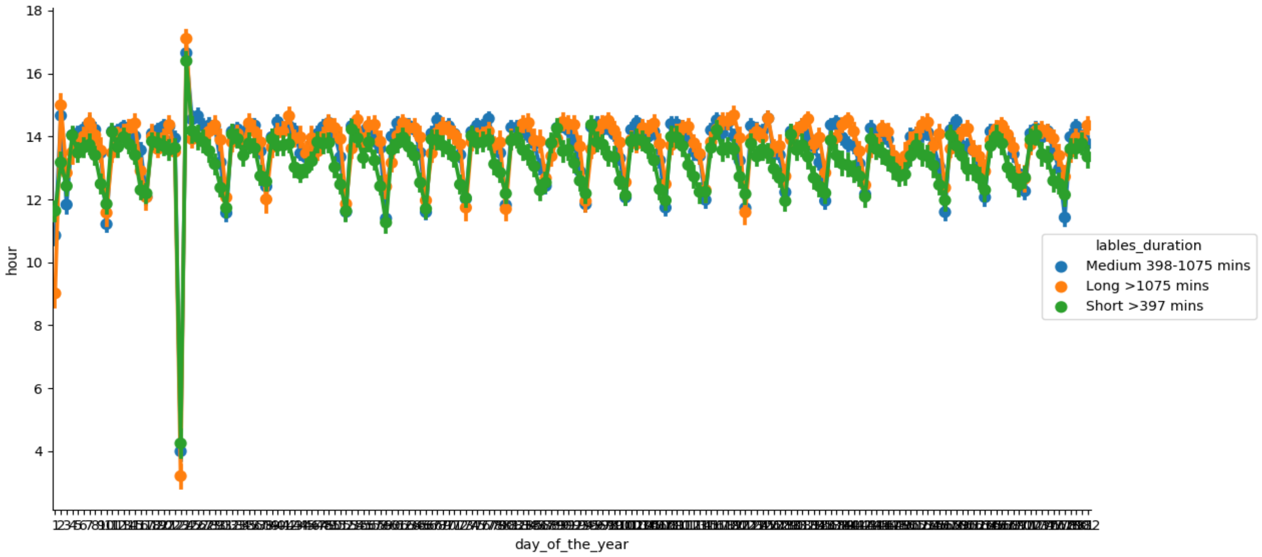
1) **Trip Duration Over a Month**: over a month the shorter trips took place earlier in the day time while the longer trips happened to take place later in the day



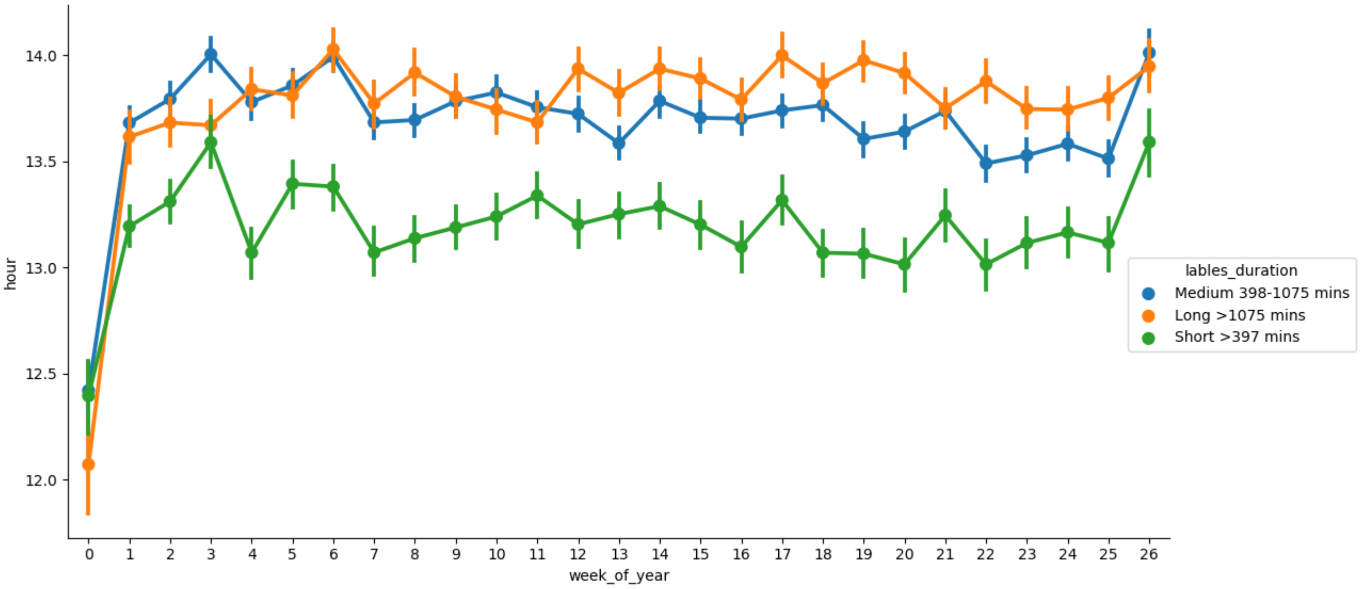
2) **Trip Durations over a week**: All trips started late in the day at the beginning of the week , towards the end of the week the trips started earlier in the day, the longer trips occurred later in the day compare to the shorter and intermediate trips, while longer trips start earlier at the beginning of the week



3) **Trip duration over a year:** the starting times and trend of the trips over the 182 days follow the same trends of trips over a month and the week



4) Trip Duration on average over the first six months: at the beginning of the year all trips started earlier in the day, and shifted towards later in the day as the year progressed.

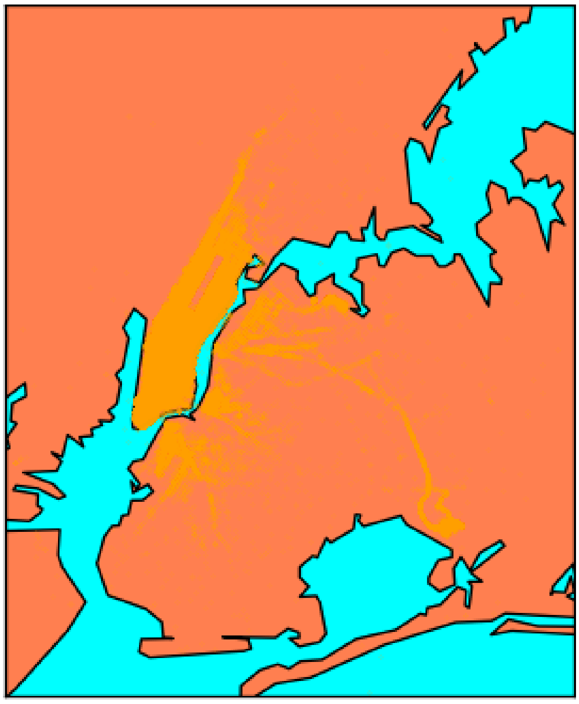


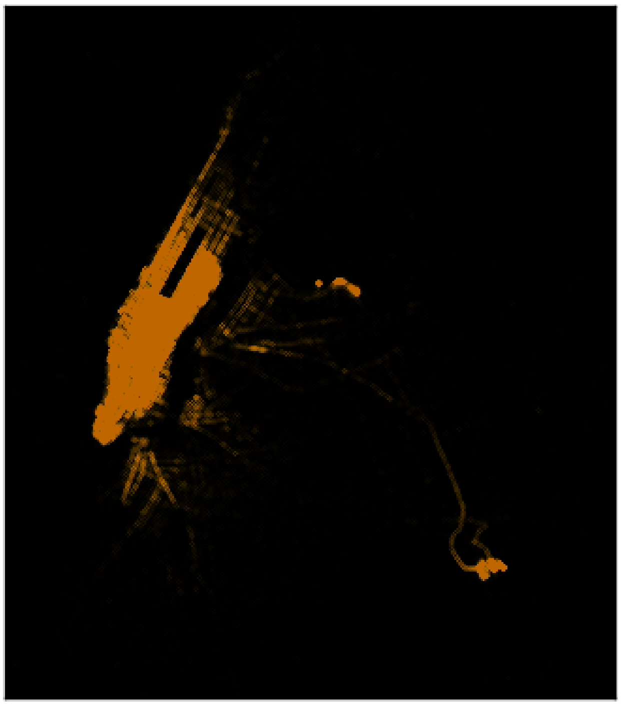
Spatial data

The data from the latitudes and longitudes were plotted on map, heat maps were drawn using an alpha of 0.1

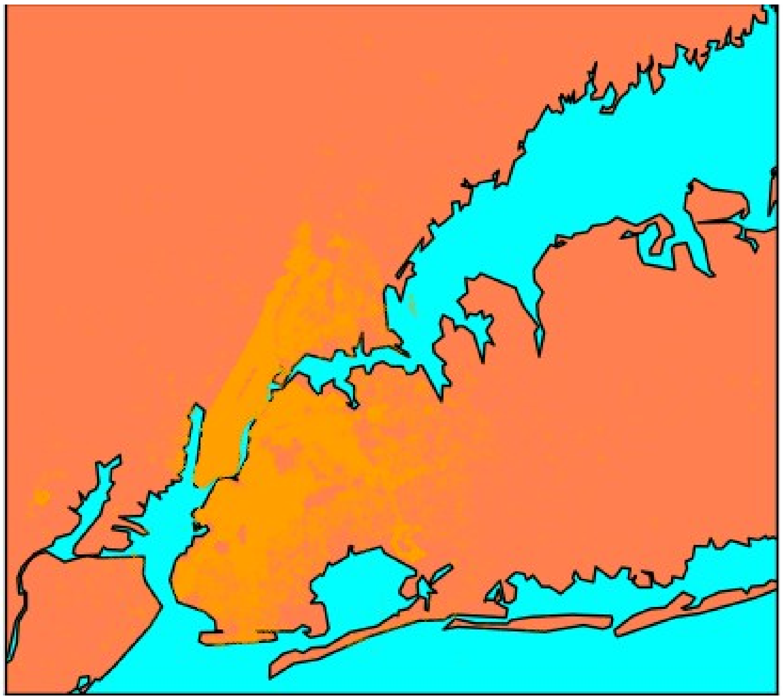
Trip Duration

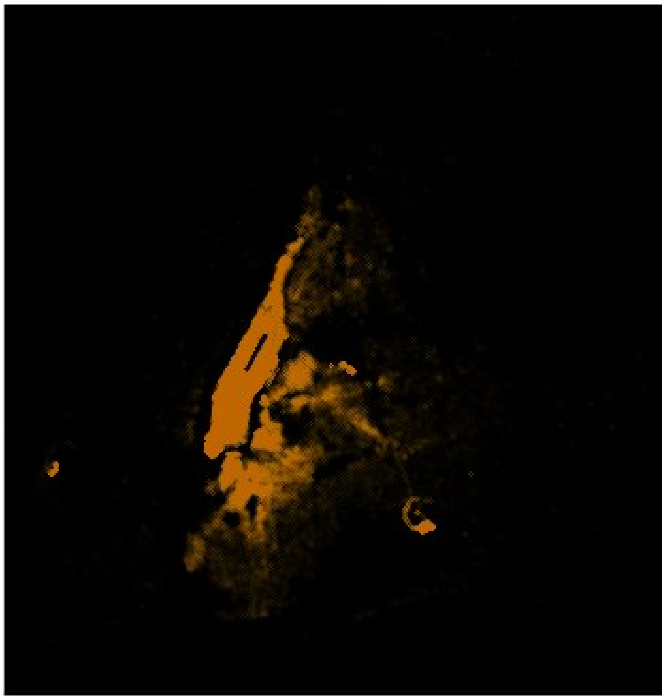
1) **Long Duration:** distribution of **pick up points** and its heatmap on the right



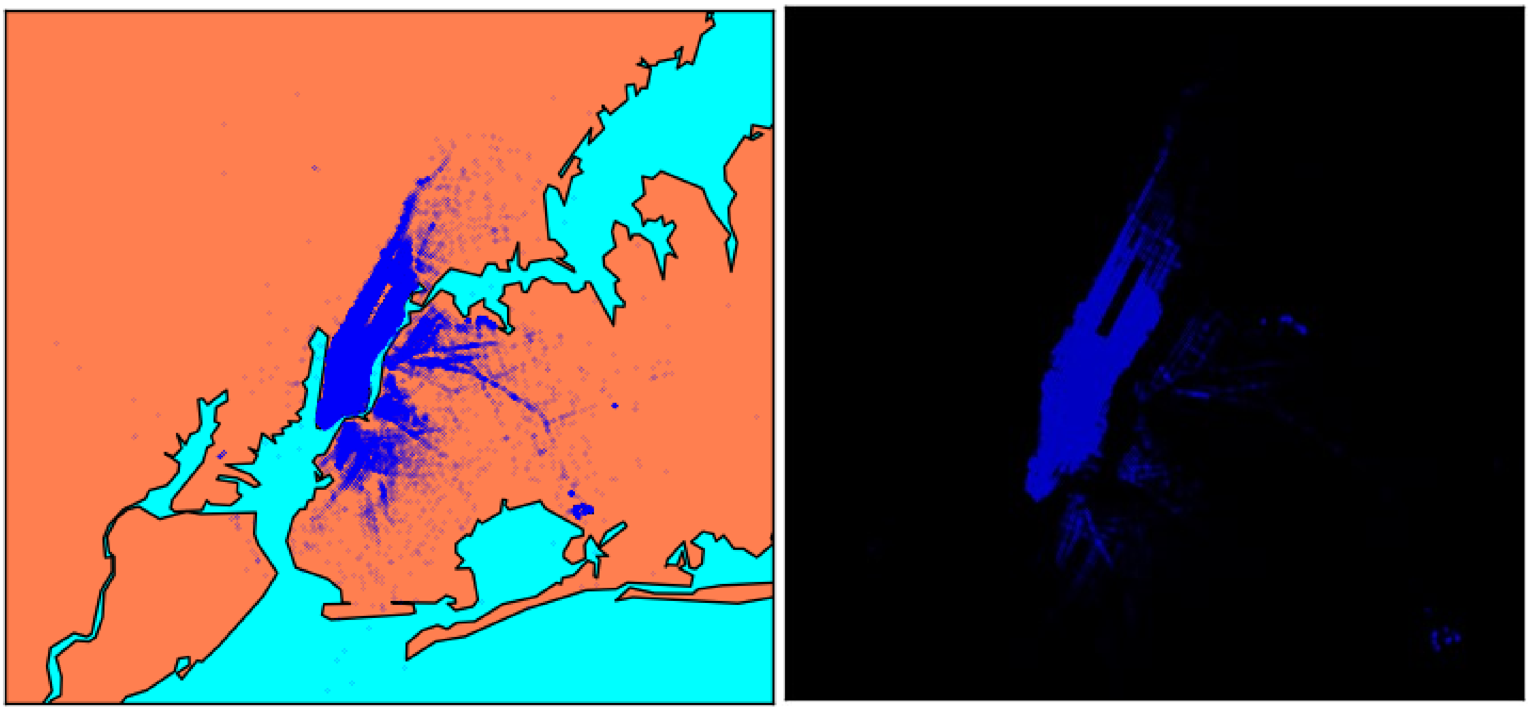


2) **Long Duration:** distribution of **Drop-off points** and its heat map on the right

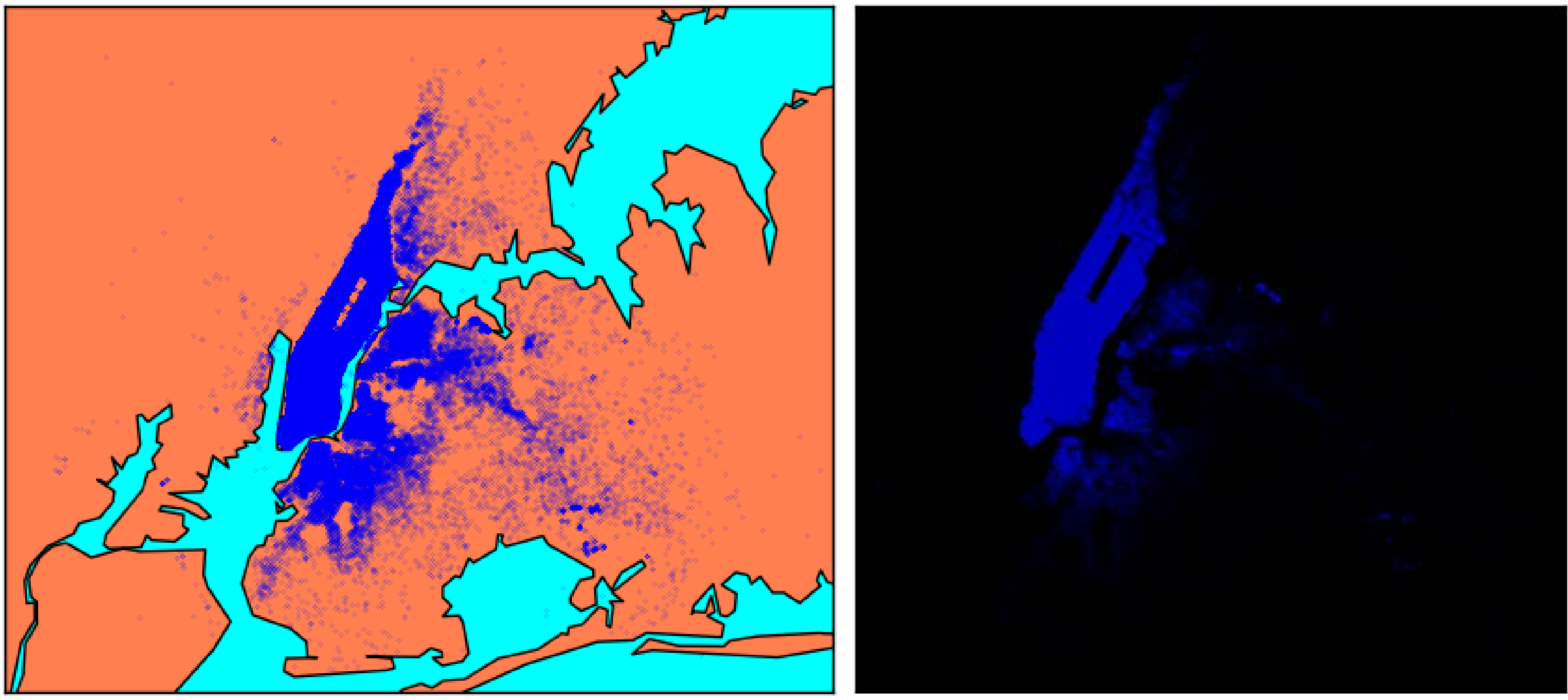




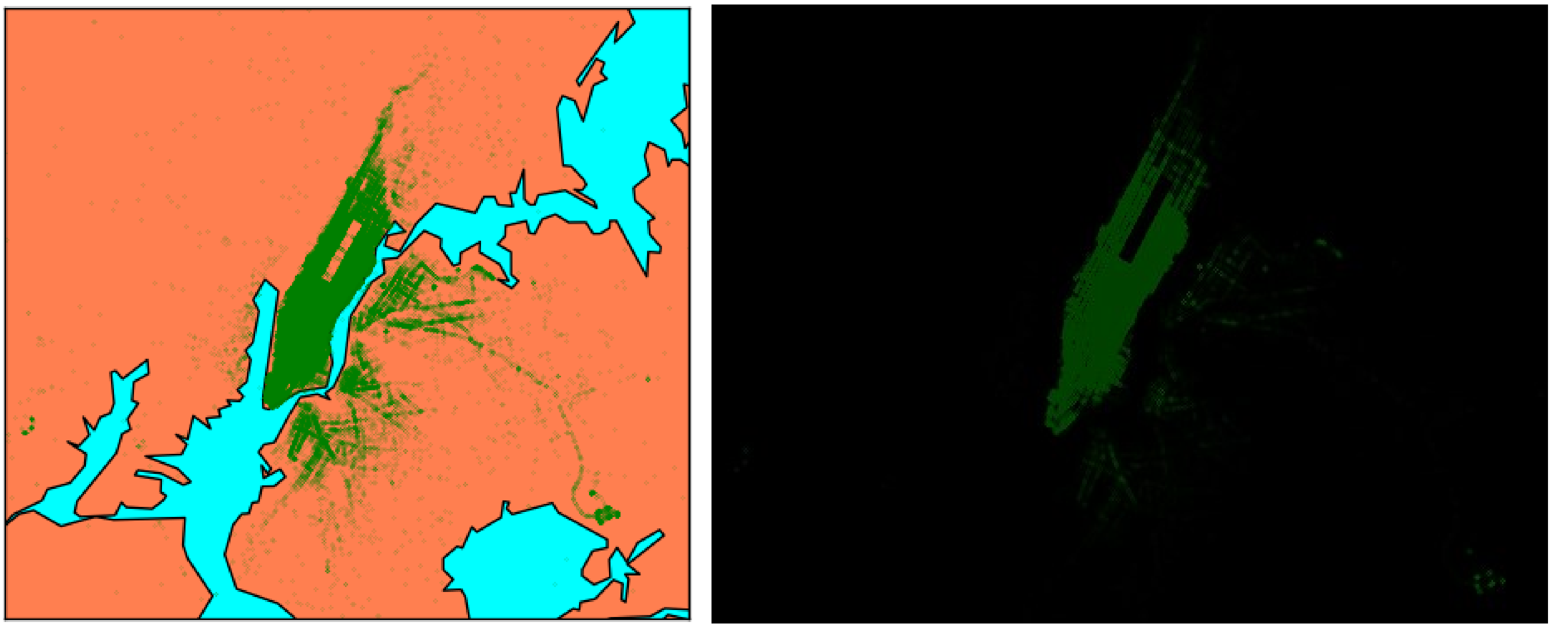
1) Intermediate duration: distribution of **pickup points** and its heatmap on the right



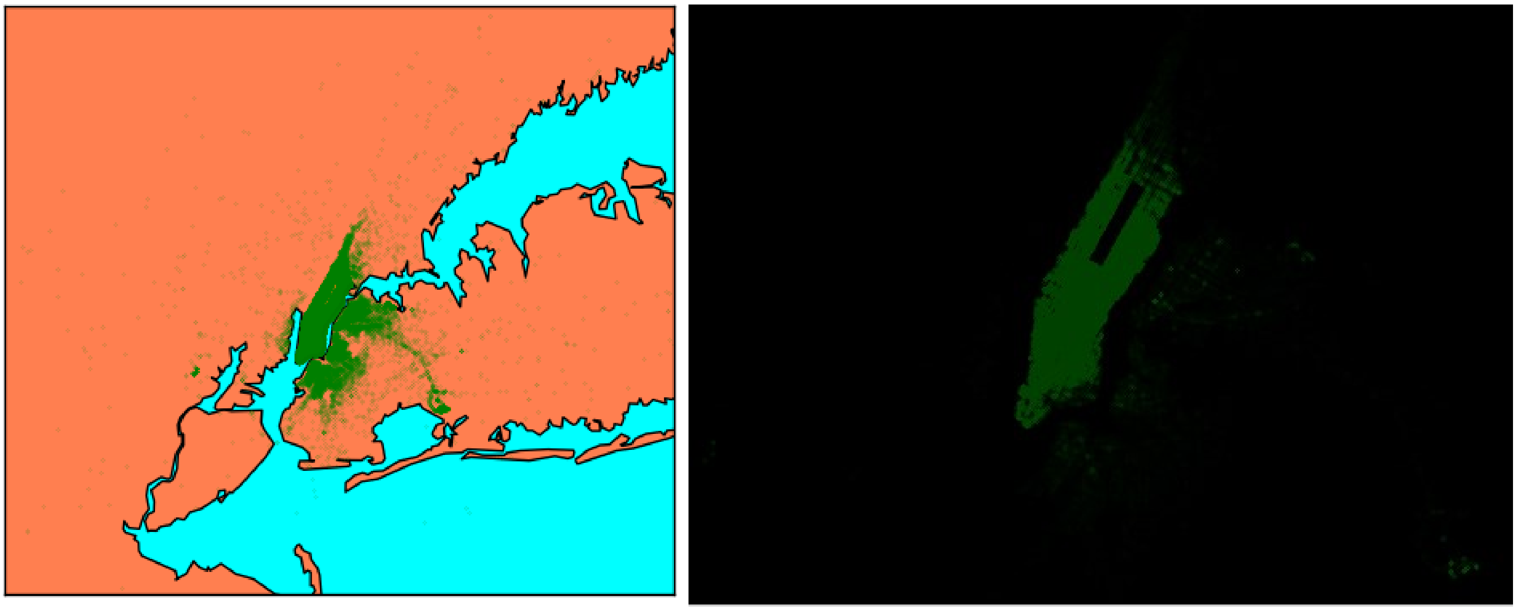
2) Intermediate Duration: Distribution of **drop-off** points and its heat map on the right



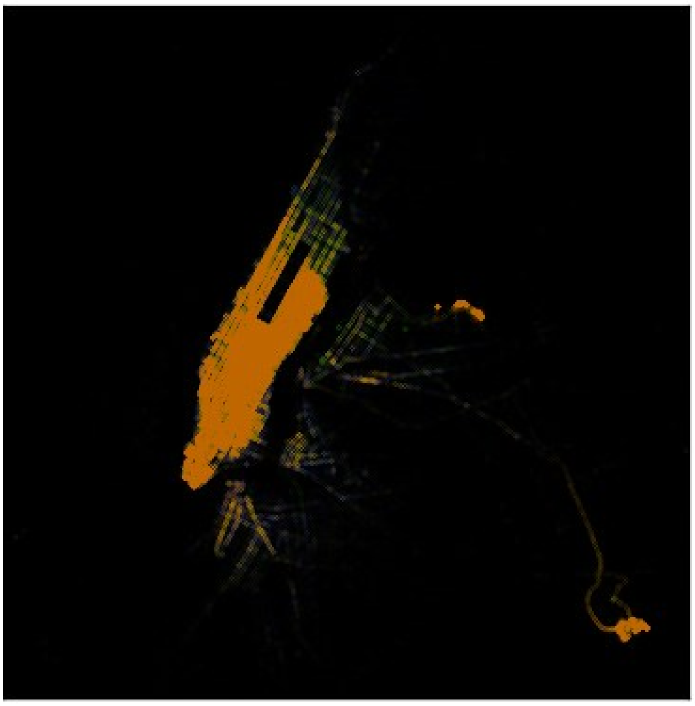
1) **Short Duration:** Distribution of **pickup** points and its heat map on the right

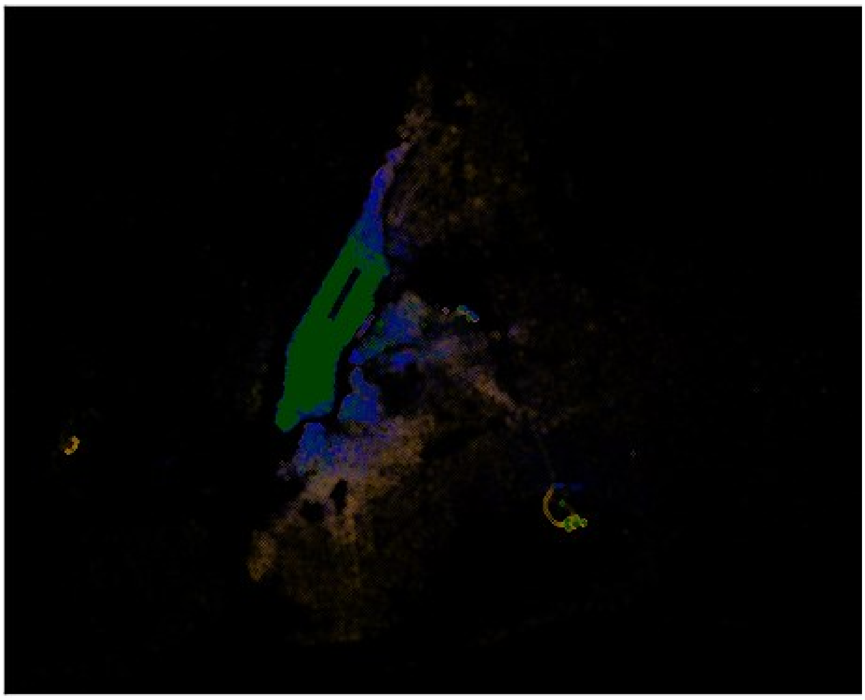


2) **Short Duration:** Distribution of **Drop-off** points and its heat map on the right



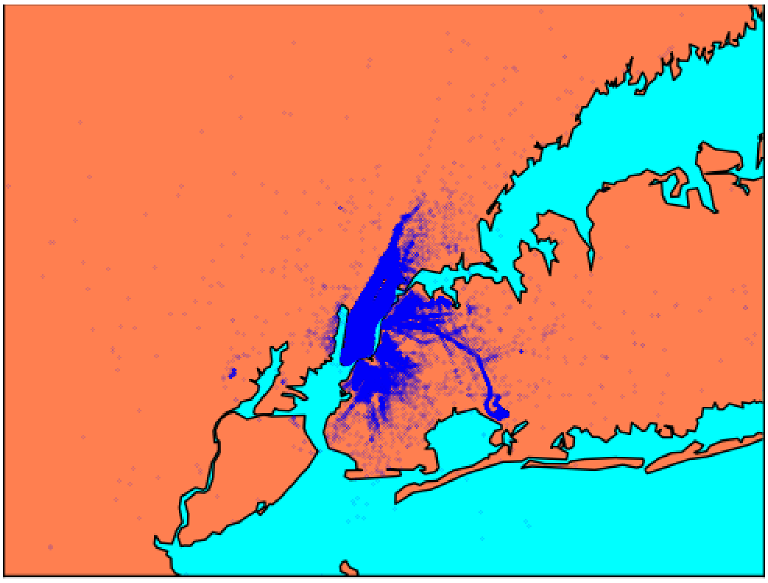
1) Combined heat maps of **pickup points** on the left and **Drop-off points** and on the right

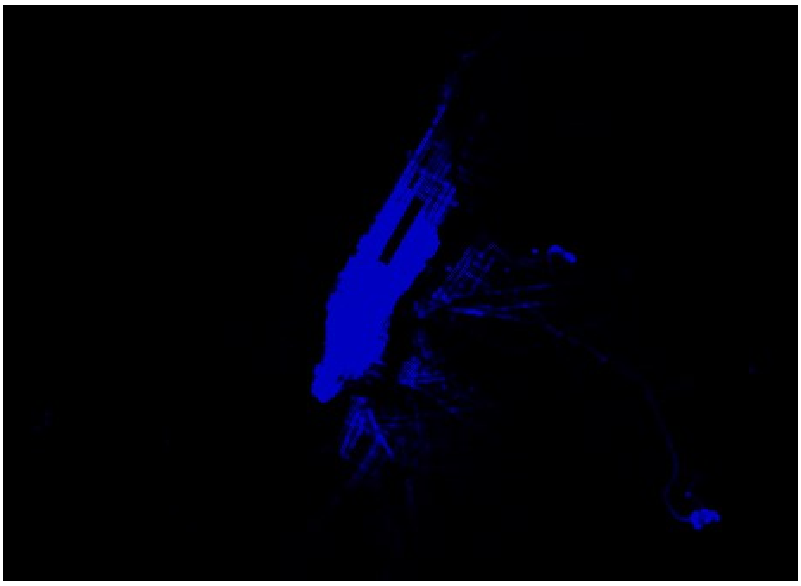




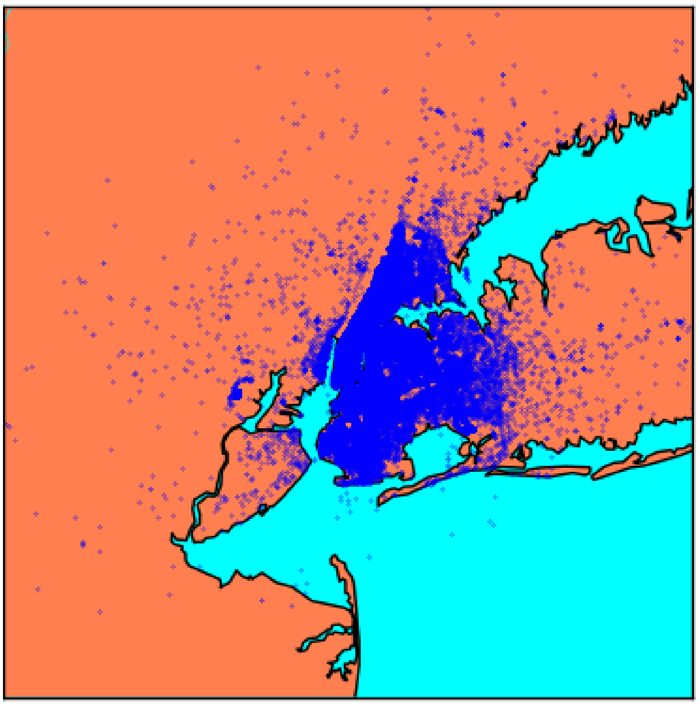
Passenger count heat maps

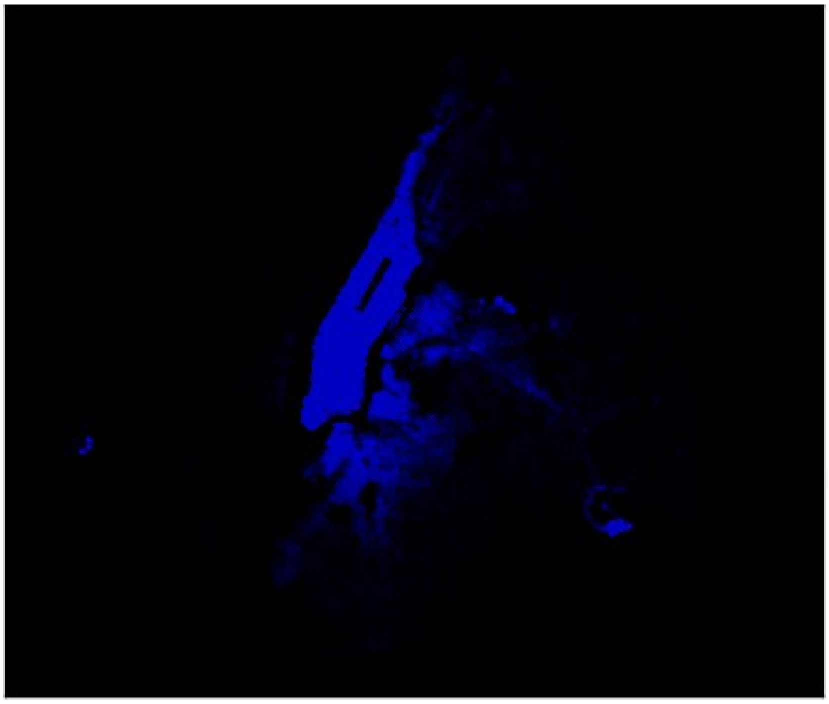
1) Pick up points for 1 passenger: Distribution and heat map to the right



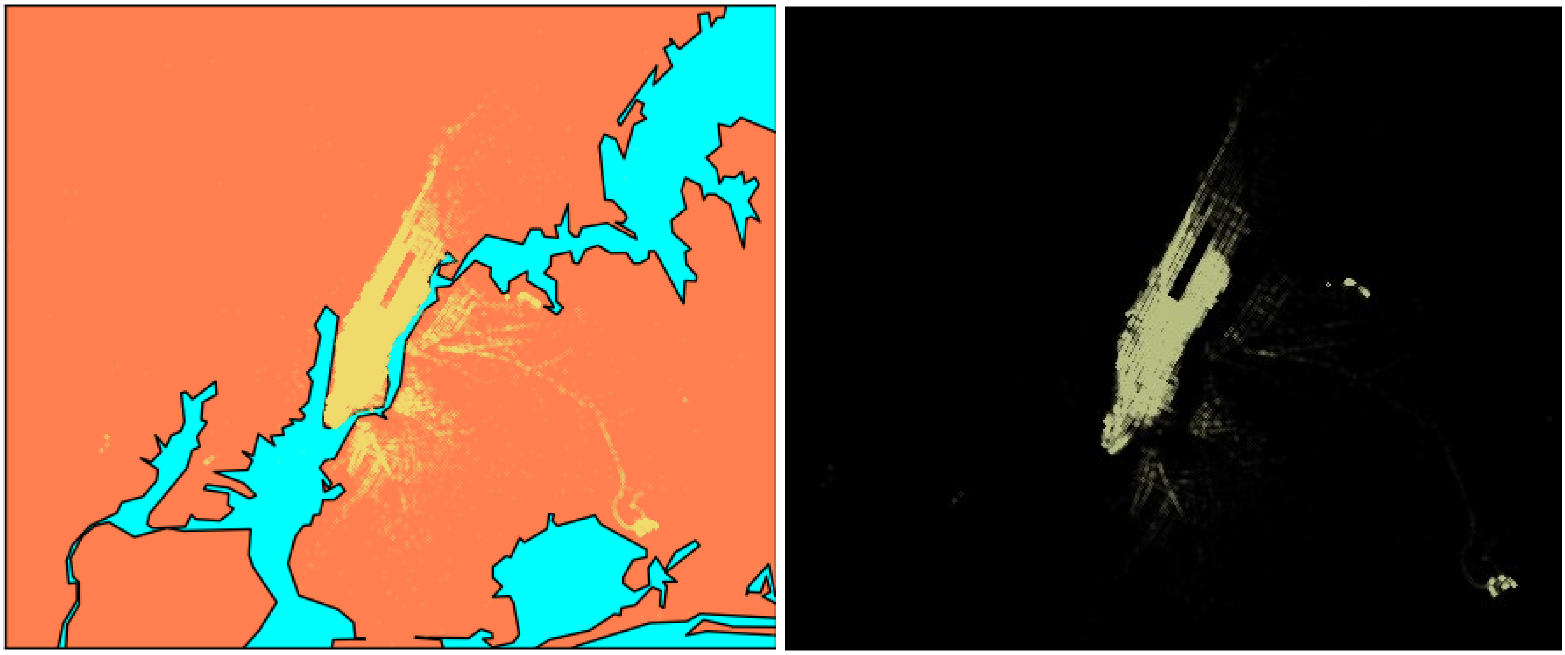


2) Drop-off points for 1 passenger: Distribution and heat map to the right

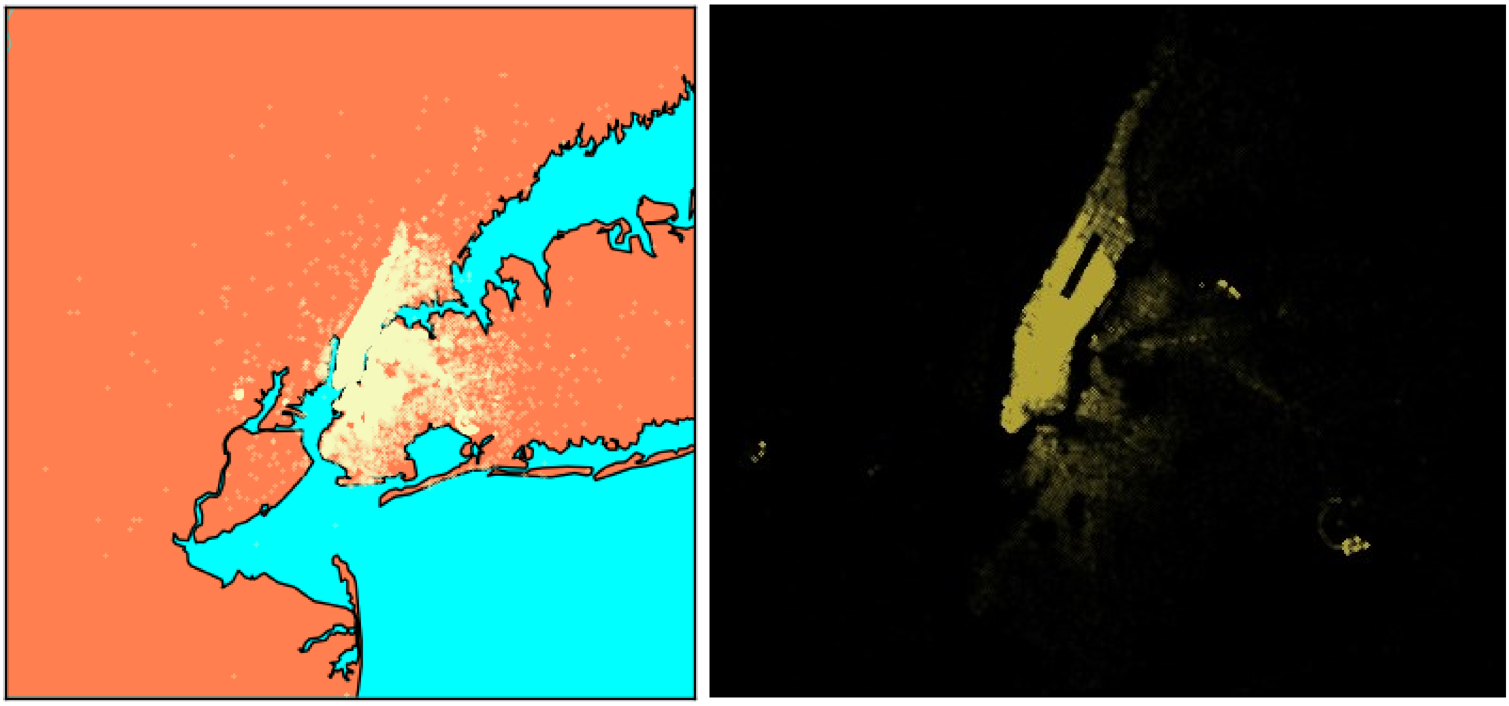




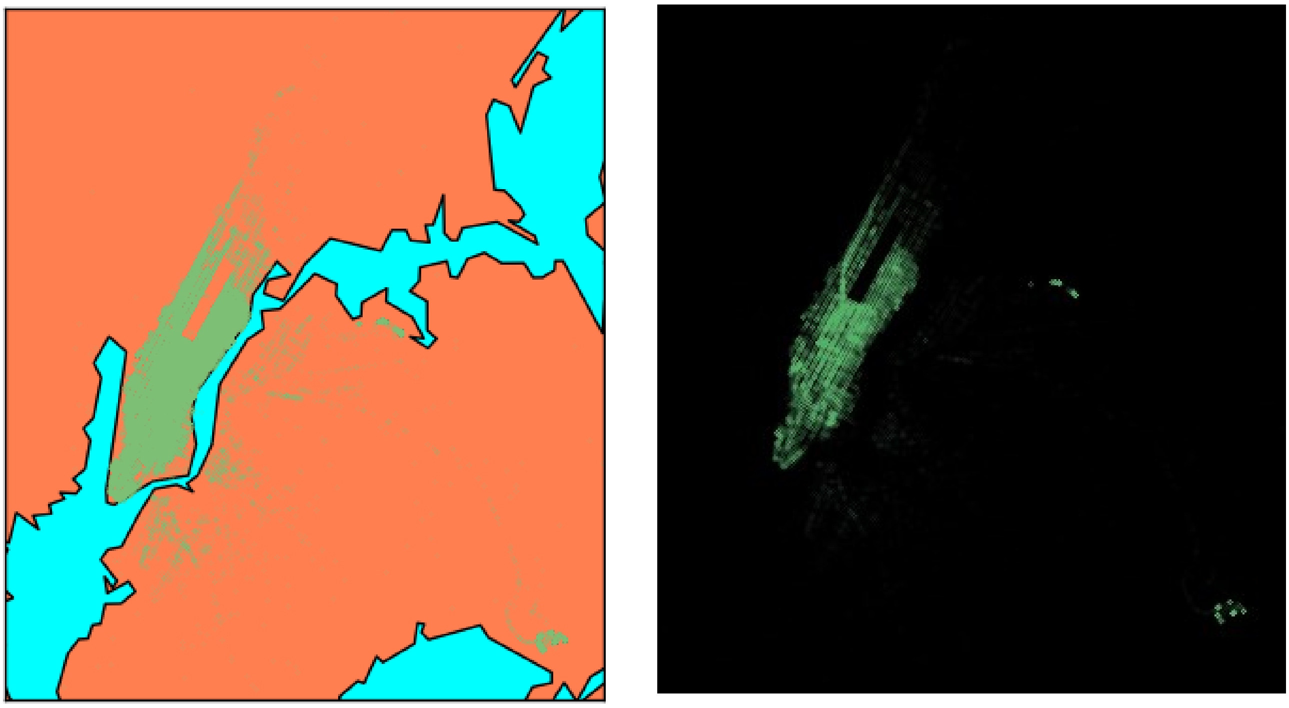
1) Pick up points for 2 passengers: Distribution and heat map to the right



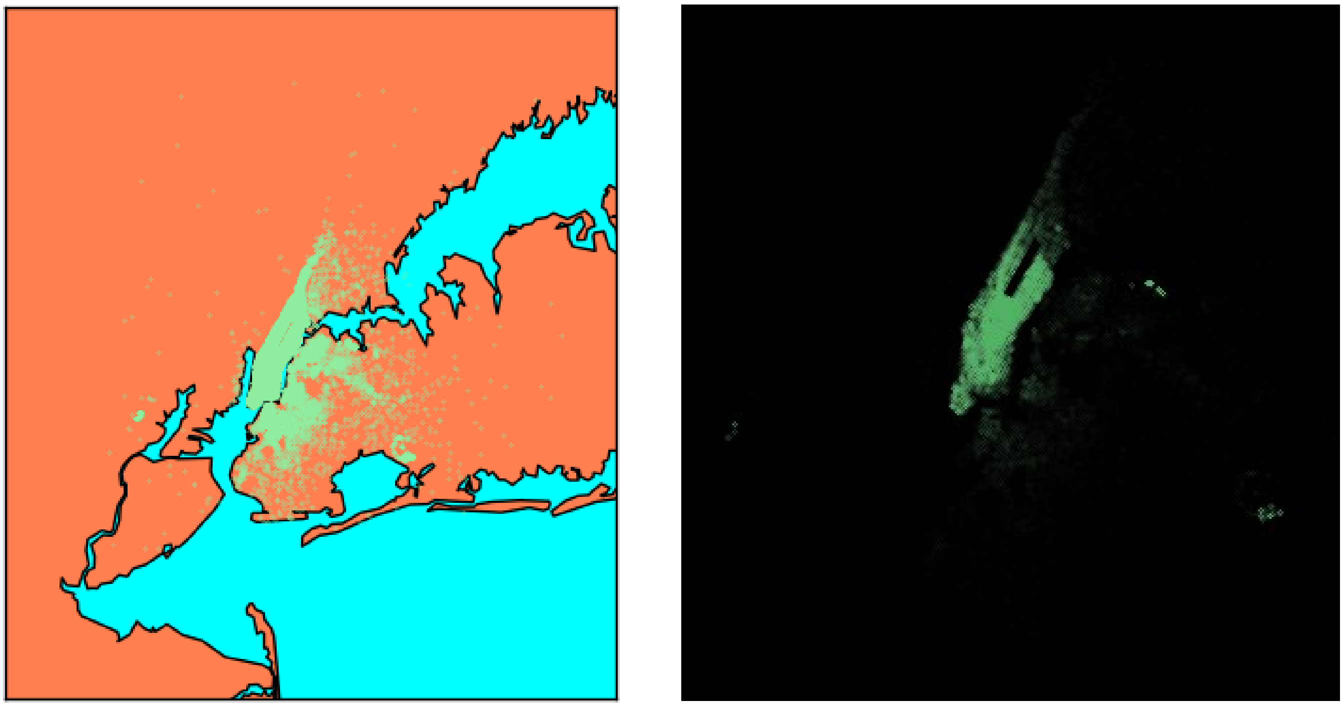
2) Drop-off points for 2 passengers: Distribution and heat map to the right



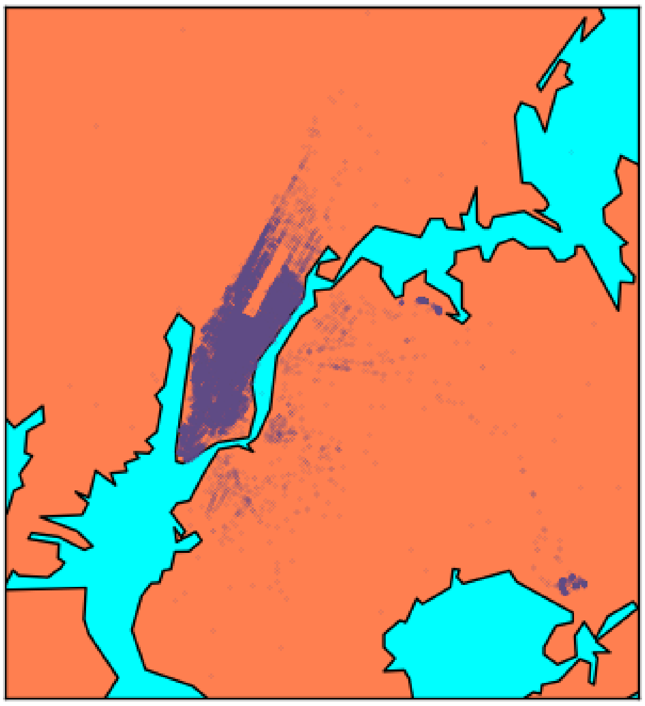
1) Pick up points for 3 passengers: Distribution and heat map to the right

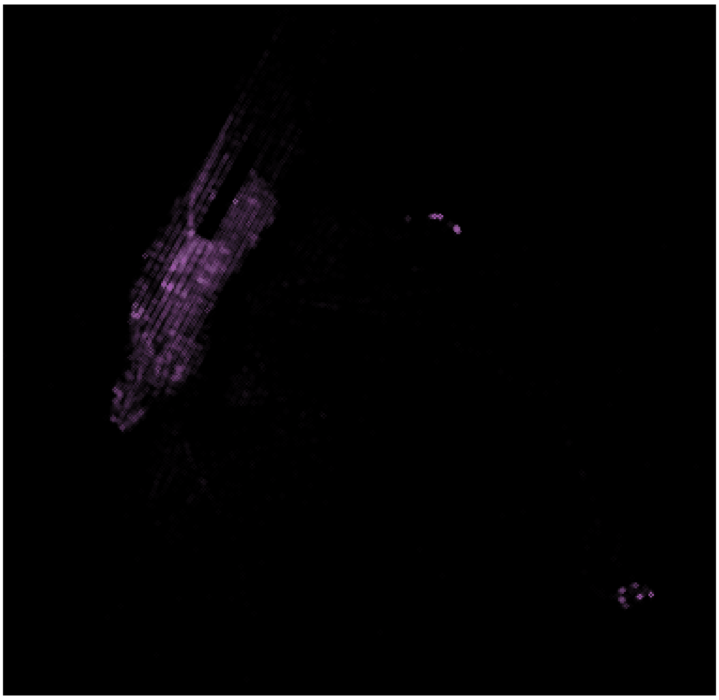


2) Dropoff points for 3 passengers: Distribution and heat map to the right

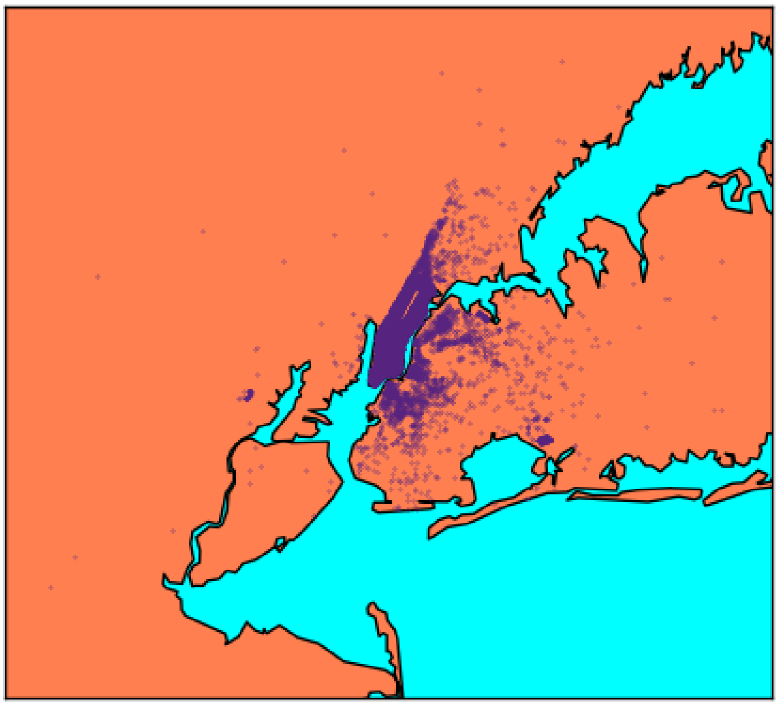


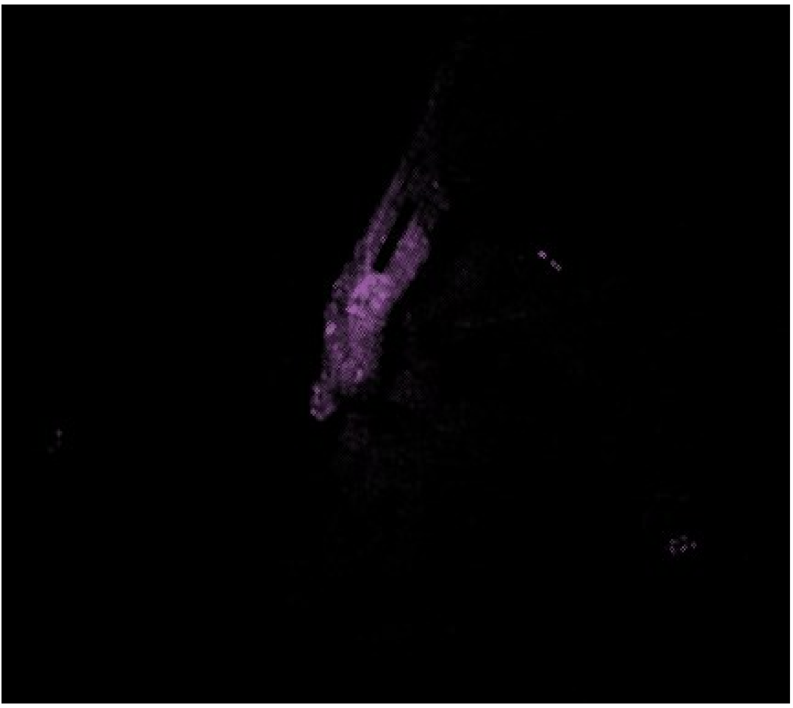
1) Pick up points for 4 passengers: Distribution and heat map to the right



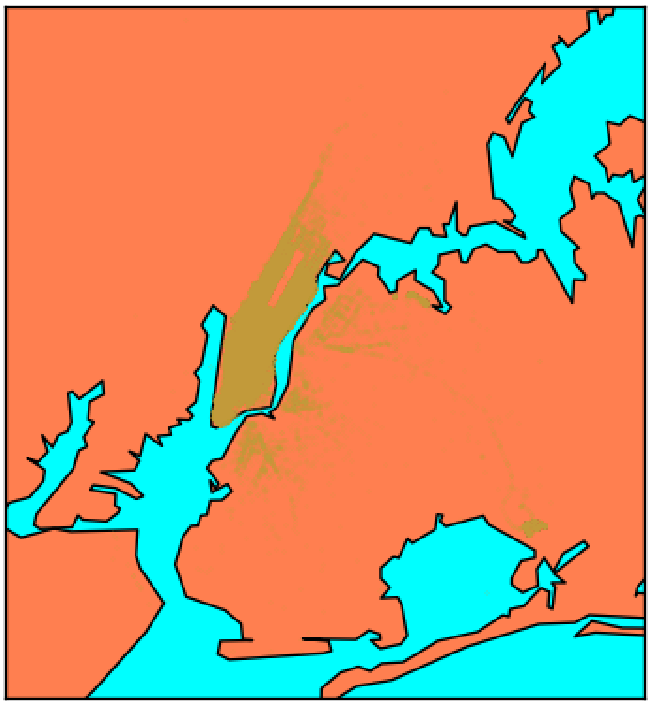


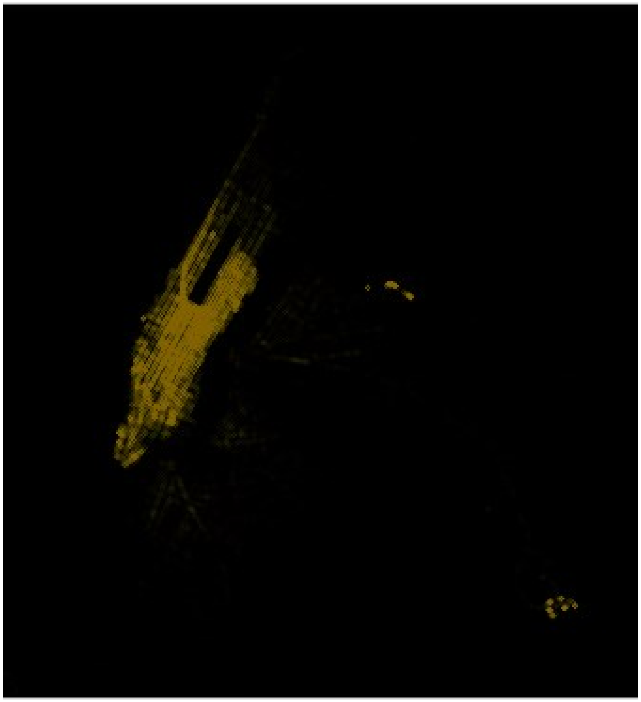
2) Dropoff points for 4 passengers: Distribution and heat map to the right



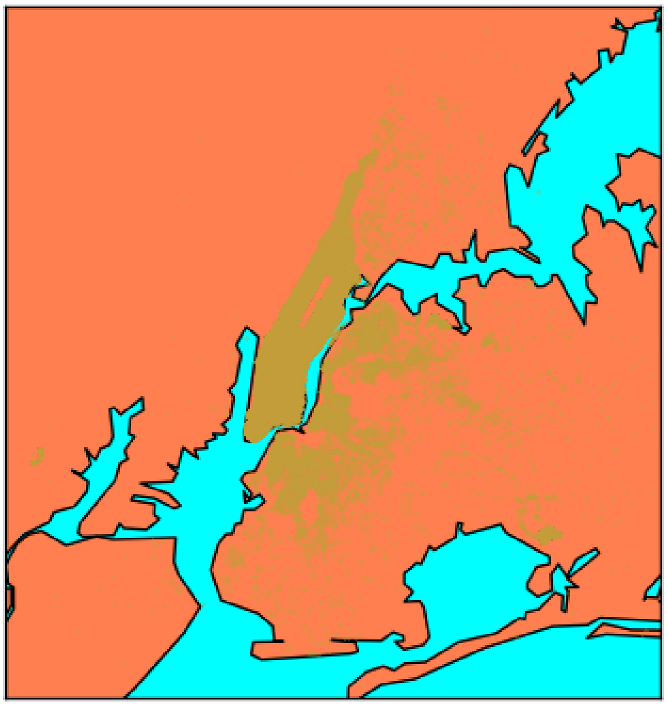


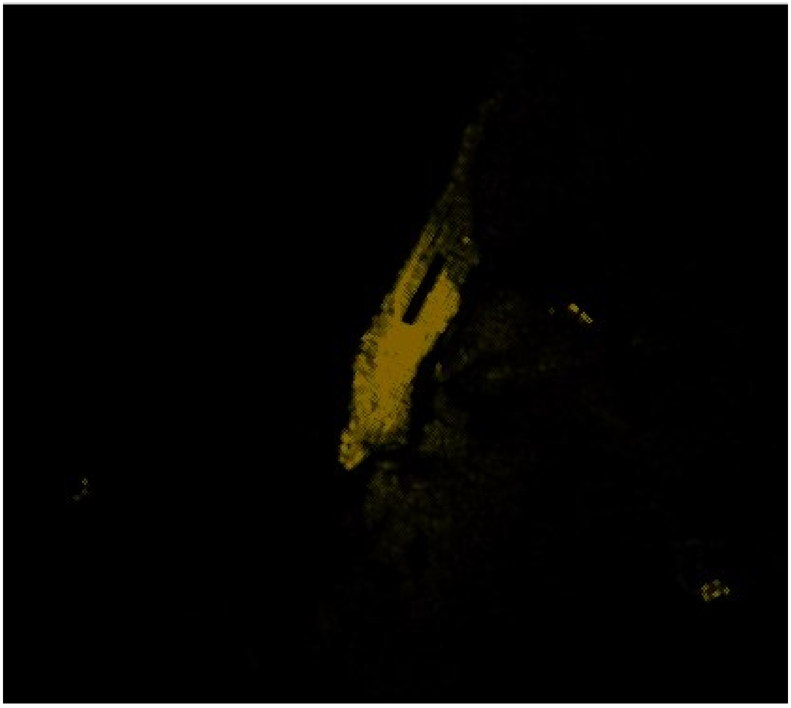
1) Pick up points for 5 passengers: Distribution and heat map to the right



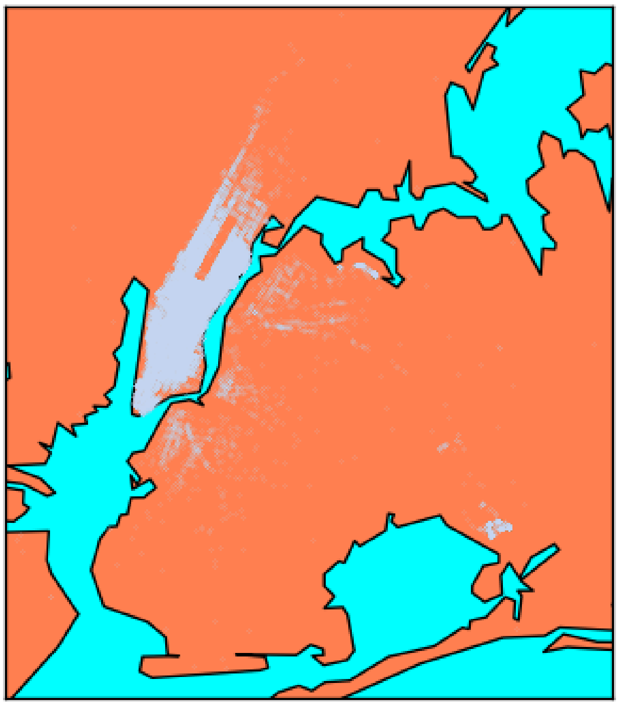


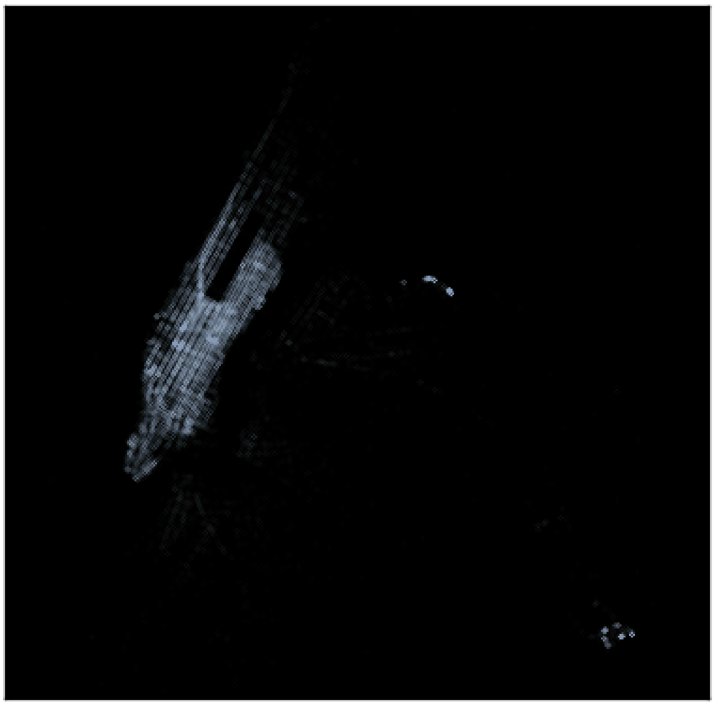
2) Dropoff points for 5 passengers: Distribution and heat map to the right



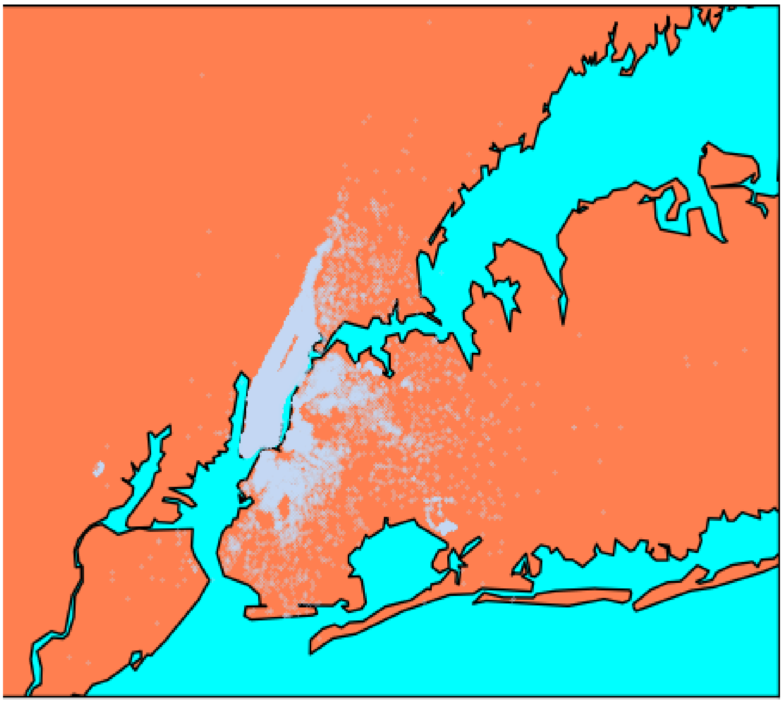


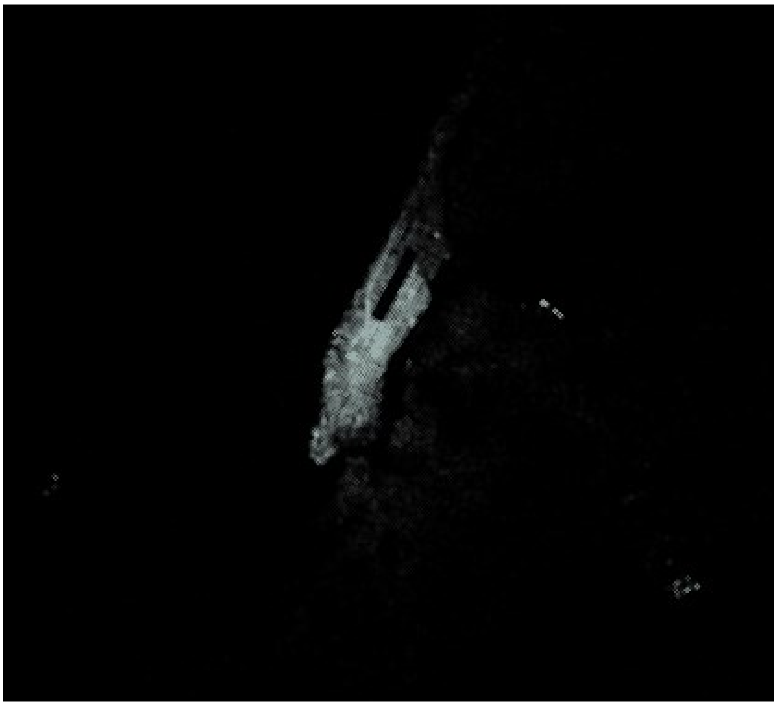
1) Pick up points for 6 passengers: Distribution and heat map to the right





2) Dropoff points for 6 passengers: Distribution and heat map to the right





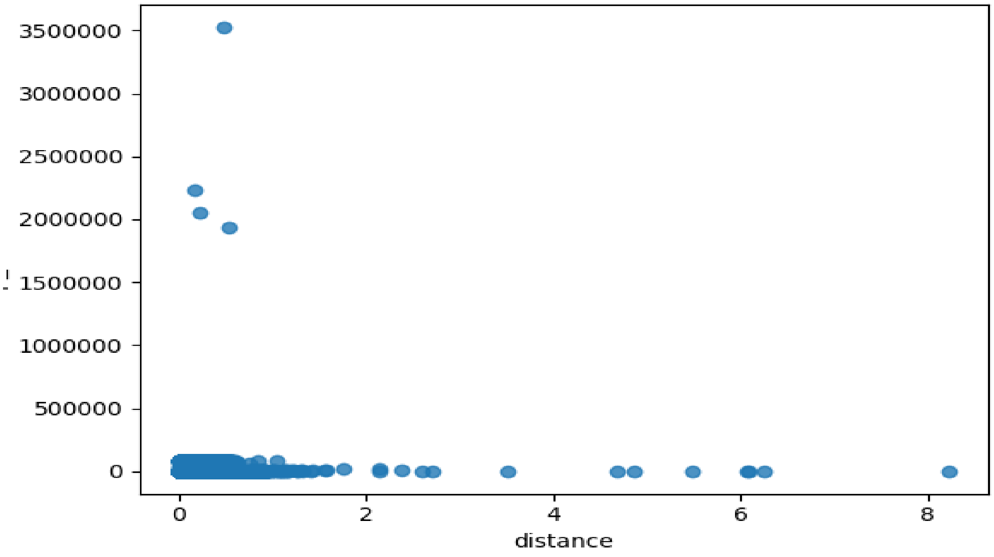
***Passengers 0, 7, 8 and 9 do not have enough data to plot distributions and heat maps**

Distance using latitude and longitudinal data

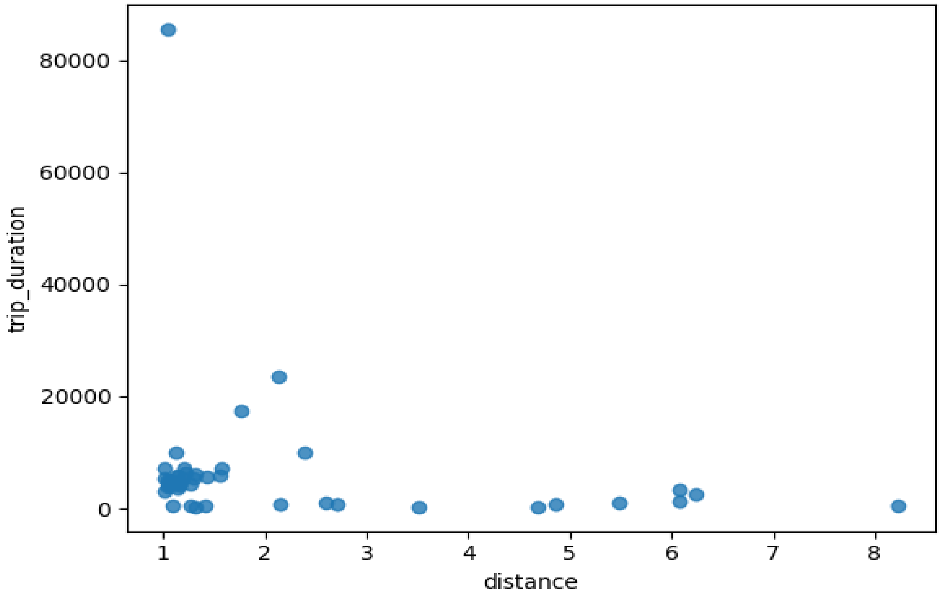
Using the Pythagoras theorem I calculated the distance between the pickup and the drop off locations, where the distance was the hypotenuse.

Plotting distance vs trip duration

1) The original plot



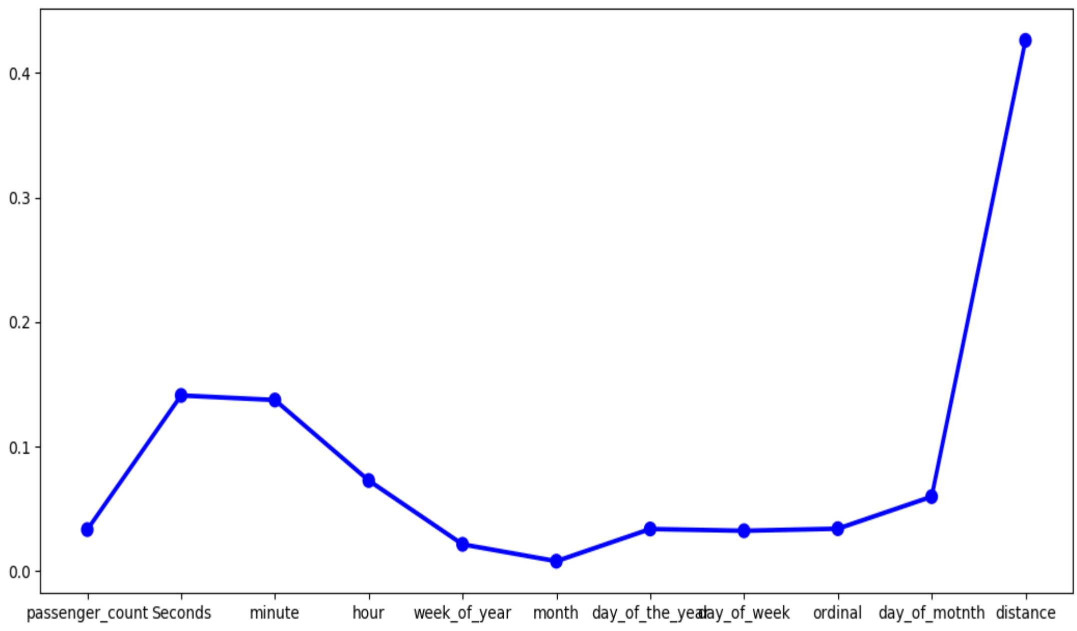
2) Plot after outliers were dropped



There appears to be some correlation at smaller distances but this doesn't hold as distance grows larger, this could probably because of the use of free ways to get to further locations compared to city driving for shorter distances, correlation wasn't expected to be strong as the distance calculated was a straight line while the taxi would have taken a curving path.

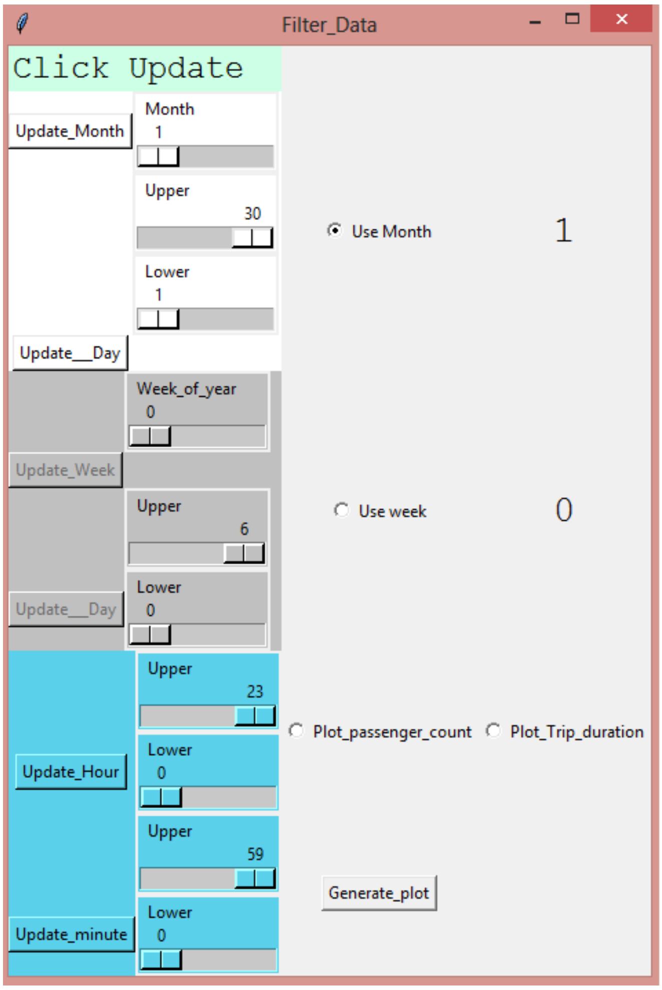
Using a Decision Tree on the data

Feature Importance as given by the Decision tree is



The class labels used in the decision tree were **Trip Duration: short, intermediate and long.**

Using the GUI



Month 1

Upper 30

Lower 1

Week_of_year 0

Upper 6

Lower 0

Upper 23

Lower 0

Upper 59

Lower 0

● Use Month 1

○ Use week 0

○ Plot_passenger_count ○ Plot_Trip_duration

To use the GUI, elect weather you would like to use month or week then filter the data according to day, **after every choice Click the update button on the left,** then click generate plot