**How would you assess the efficiency of aggregating rides within Brooklyn? From Brooklyn to Manhattan?**

Basic Algorithm

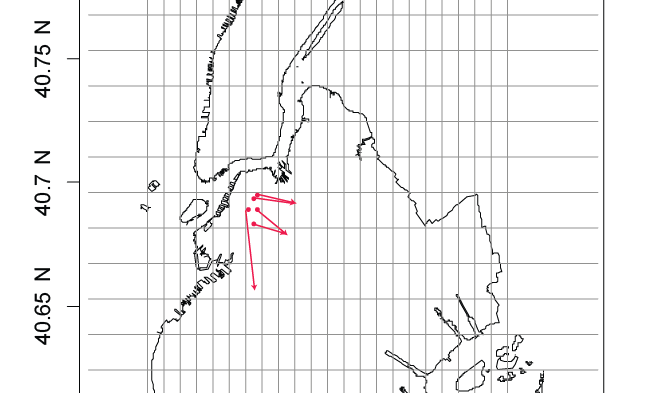
To assess the efficiency of aggregating rides, I have developed an algorithm to estimate the number of rides, at any given time, headed to similar destinations.

The base algorithm works as follows:

1. Divide Manhattan and Brooklyn into 2d bins.
2. Collect all departure points into their respective bins.
3. For each bin of departure points, track how many different bins the destinations wind up in. For example, below, there are 5 departures from Bin A. These 5 rides proceed to have destinations in 3 different bins.

This results in two measures: Number of trips from a bin (here 5), and number of unique bin destinations (here 3).

1. Given the Number of trips from a bin, and the number of unique bin destinations it’s possible to estimate the number of rides, from this departure bin, headed to similiar same destinations.   
   In the below example, 5 departures go to 3 destinations. The algorithm interprets that as an efficiency of 40% (1 - 3/5). And so, in this case, 2 trips could have been shared (40% of 5 total departures).
2. This metric is calculated for each bin, and summed, given a total number of trips that could have been shared.   
   Any bins with 0 trips are ignored.



Building in Time

Additionally there is a time component. A moving window is used to subset the taxi data, and the above algorithm is applied. A more detailed explanation of the process follows.

1. Across all the taxi departure times, rolling 15 minute windows are selected
2. The above algorithm is applied to the 15 minute subset of data.   
   This adds the restriction that the departures were in the same area, within 15 minutes of each other.   
   The result is, for a given 15 minute window, what is the number of rides that could have been shared.
3. The moving window advances in 5 minute steps. Producing an estimated number of rides that could have been shared for any 15 minute window, with a 5 minute resolution.

Data sources

The following addresses the data sources which were used, and the steps to preprocess which were required.   
The entire 2013 taxi data set (give here: <http://www.andresmh.com/nyctaxitrips/>) and the entire 2015 taxi data set (given here: <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>) was used.

The decision to analyze the 2013 and 2015 dataset was undertaken for two reasons. First, anecdotally, it seems many people (especially younger folks) are moving from Manhattan to Brooklyn. Secondly, between 2013 and 2015 ride sharing has come into much greater prominence. It will be interesting if either of these details are visible in the differences between 2013 and 2015.

For both the 2013 and 2015 dataset, it was required to identify trips within Brooklyn, and within Manhattan. To do this, a shape file of the NYC boroughs was download (<http://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page>) and used to label the pickup and dropoff borough for every ride in the dataset.

From these labels, 3 primary datasets, for the 2013 data, and the 2015 data, were created:

1. Within Brooklyn: Trips within Brooklyn (pickup and drop-off in Brooklyn)
2. Between Brooklyn and Manhattan: Trips between Brooklyn and Manhattan (1 pickup or drop-off in Brooklyn, and 1 pickup or drop-off in Manhattan)
3. Within Manhattan: Trips within Manhattan (pickup and drop-off in Brooklyn)