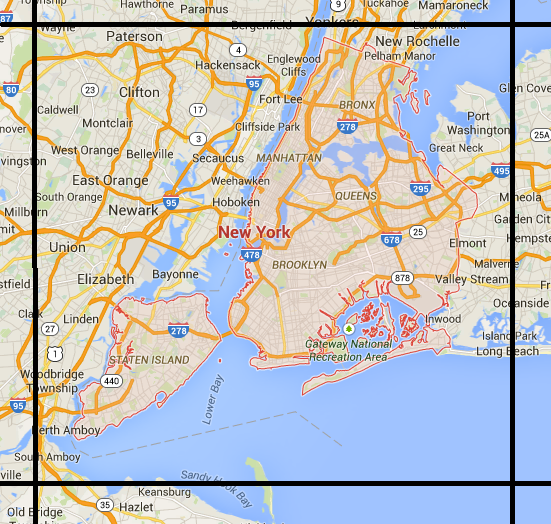
**Geo Coding:**

We ignored all the points that are way outside NYC. The area under consideration:



It is easier to GeoCode the distinct latitude and longitude points instead of using the whole trip data table.

/\*

Trip data contains 173179759 pickup points and 173179759 drop off points. (Approx. 340 million)

But there are only 80 million distinct points (pickup and drop off combined).

\*/

**Pig Script to identify distinct points:**

raw = LOAD ‘/user/dg1937/trips\_2013/trip\_data’ USING PigStorage(‘,’);

pickups = FOREACH raw GENERATE $10,$11;

dropoffs = FOREACH raw GENERATE $12,$13;

points = UNION pickups, drop-offs;

distinct\_latlong = DISTINCT(points);

STORE distinct\_latlong INTO ‘/user/dg1937/distinct\_latlong/’ USING PigStorage(‘,’);

**Framework for Geocoding:**

|  |  |
| --- | --- |
| Latitude | Longitude |

ArcGIS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Latitude | Longitude | Boro\_Code | Boro\_Name | Neighborcood\_Code | Neighborhood\_Name |

**GeoCoding**