Assignment 4
Ruixuan Zhang

Question 1

What is the time scope of your analysis? (Did you use only the March 2016 data or did you us more?)

I used the March 2016 data and randomly sampled1% of the data from it as my analysis scope. The dataset that I worked on has 15764 observations in total. The reason I randomly sampled only 1% of the data is that's the maximal amount of data that my computer /python can handle.

Methodology and Introduction

I used DBSCAN for question 2 – question 6.

DBSCAN has great advantages over K-mean in that it doesn't not require the user to specify the number of clusters, it works great for geographical data, and it allows some observations to be grouped as outliers, as opposed to all categorize them into some clusters. To convert and compute longitude and latitude into distance, I used haversine function (metric); I specified epi in DBSCAN and applied haversine metric to it, so that the maximal distance between two points for them to be considered as in the same neighborhood is 0.12 km (for some questions I set it to 0.10 km. For more information, check my code). Since 120 m is about the length of 1 or 2 blocks, I think that is an appropriate distance to cluster two relative distant points into one cluster.

How to read this report?

To visualize clusters in each question clearly, I used ArcGIS mapping API and included six maps in each of the following questions: question2, 3, 6. I included one big map for question 5 because the result on the big map is already obvious, no detailed maps needed.

To generate each plot, I plotted a heat map using the original longitude and latitude (from the sample), and used that as base map. After I get the clustering results from DBSCAN, I added those top 5 clusters one by one on top of this base map layer. Each cluster is marked in different color. So the idea is that, if my clustering algorithm is good, then these top 5 locations should match with the high-density areas on the base map. As it turns out, my clustering result matches perfectly with the reality.

In the report part, I basically included 6 pictures for each question. The first one is a 'big picture', with all of the top 5 locations (clusters) on the map. It is followed by five detailed (zoomed in) maps, each map shows a detailed information of one of these top5 location.

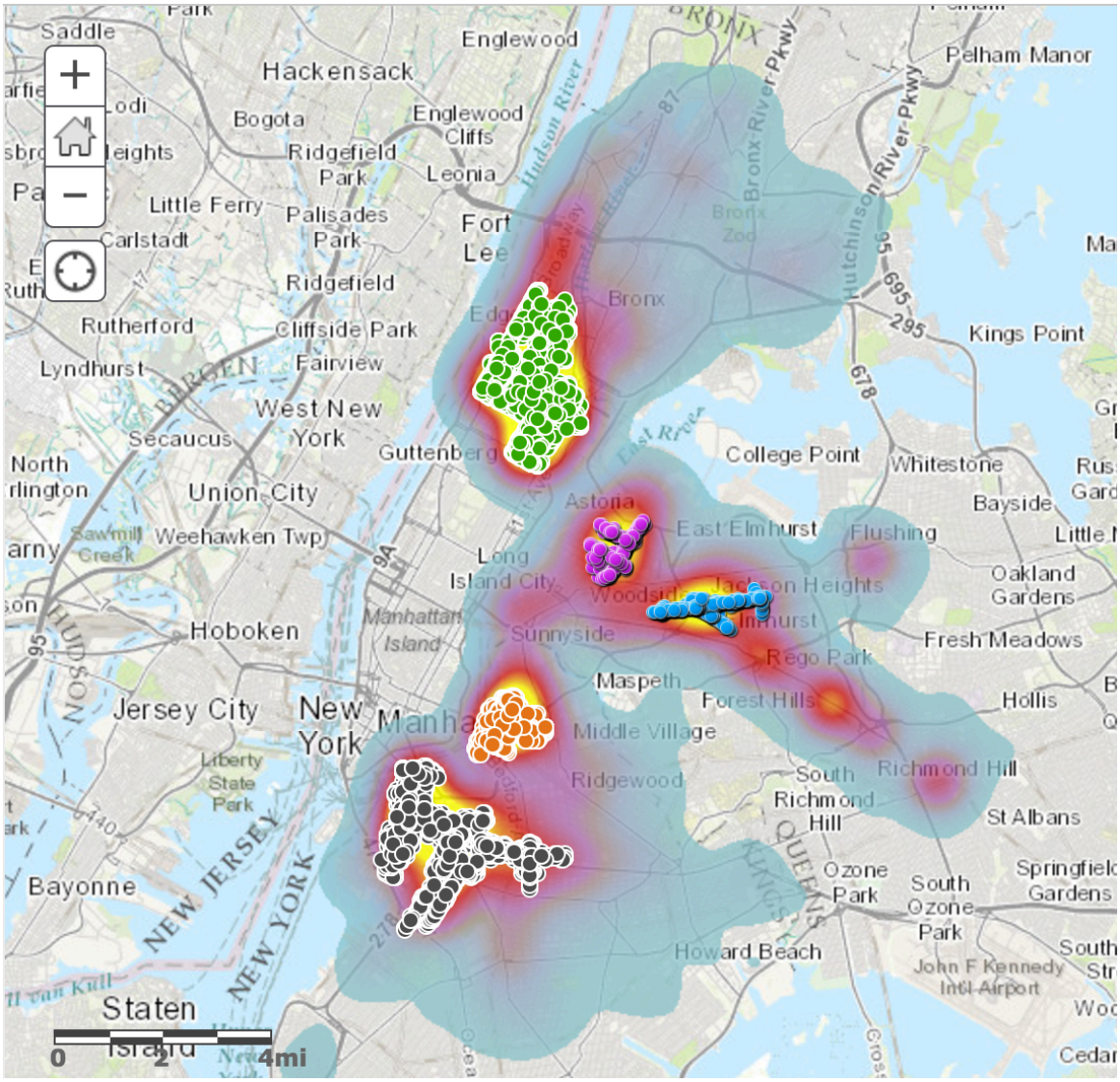
Question 2

What are the top 5 locations, which are the origin points of trips? (In other words, do certain areas generate more pick-ups than others?) What the percentage of trips originated from these locations?

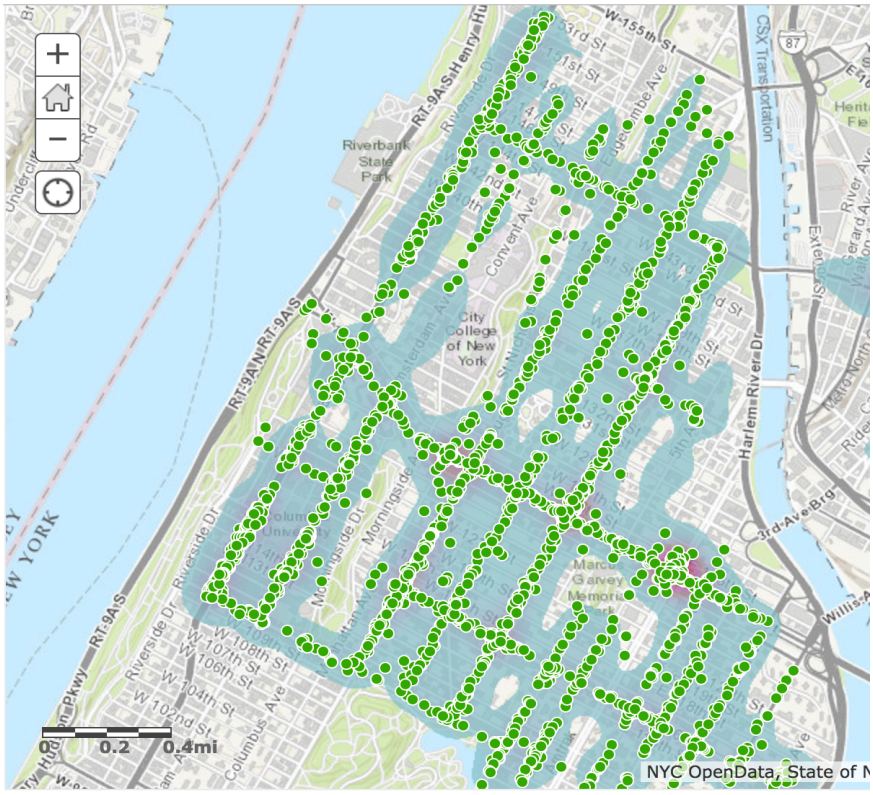
The top five clusters and the number of observations in each cluster:
{cluster #0: 3782, cluster #1: 2820, cluster #2: 1074, cluster #6: 659, cluster #15: 939}

Yes. Percentage of trips originated from those locations 58.83%

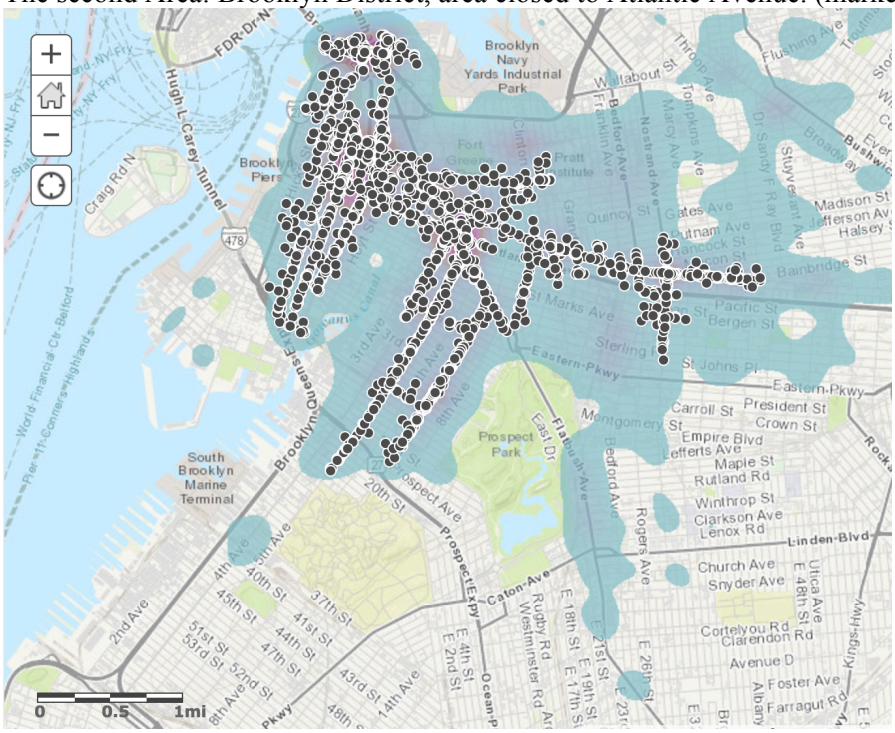
The top 5 locations that generate more pick-ups are:



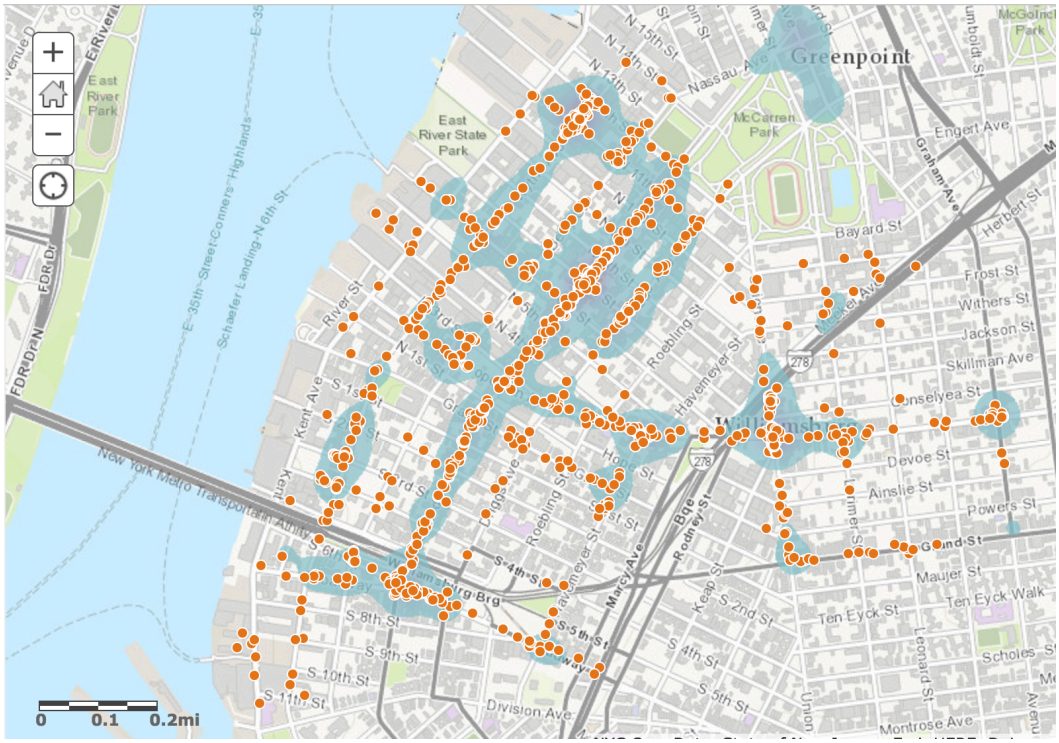
The first Area: East bank of Hudson River, North and northeast to the Central Park, around Columbia University (marked in green)



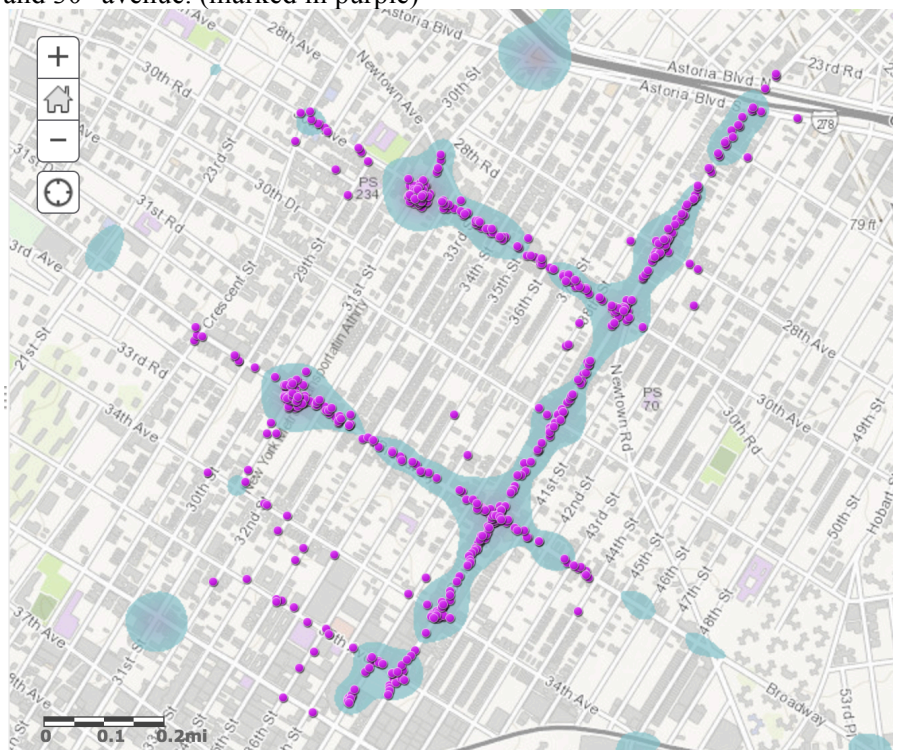
The second Area: Brooklyn District, area closed to Atlantic Avenue. (marked in black)



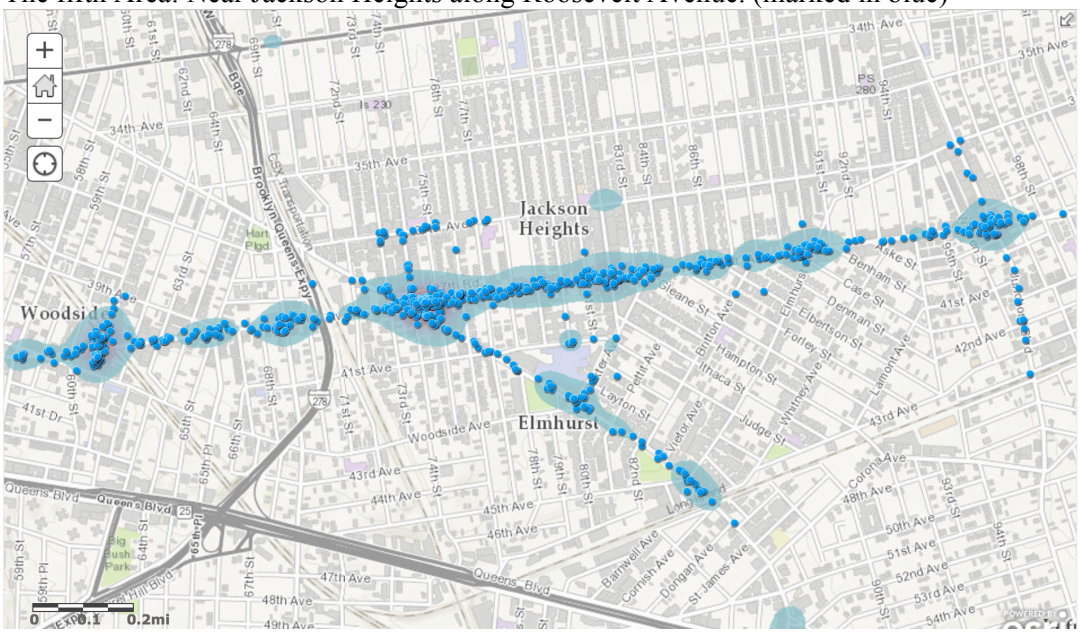
The third area: Area north to Williamsburg Brg and Brooklyn Queens Expy. (Marked in orange)



The fourth Area: Area south to Astoria and Steinway, basically along Broadway, Steinway Street and 30th avenue. (marked in purple)



The fifth Area: Near Jackson Heights along Roosevelt Avenue. (marked in blue)

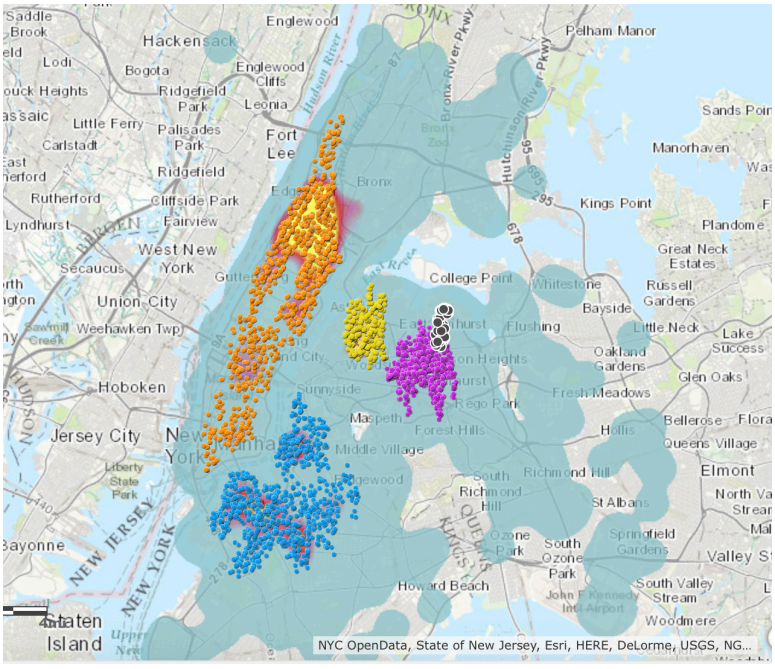


Question 3

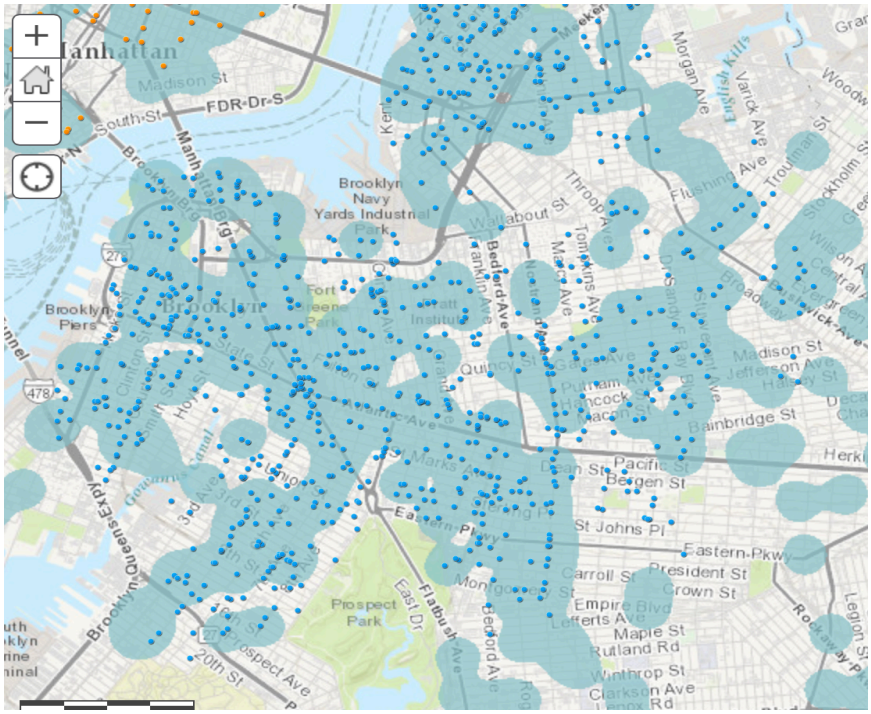
What are the top 5 locations, which are the termination points of trips? What percentage of trips terminated in these locations?

The top five clusters and the number of observations in each cluster
{cluster #0: 5198, cluster #1: 2562, cluster #2: 906, cluster #4: 828, cluster #5: 924}
Yes. Percentage of trips terminated in those locations 66.30%

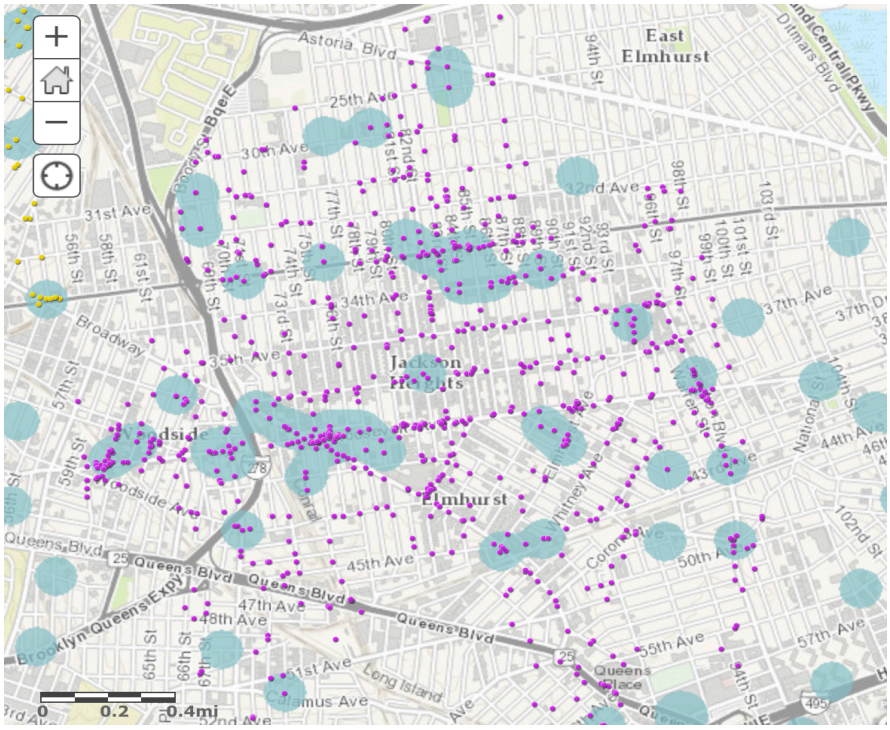
The top 5 locations that had more drop-offs are:



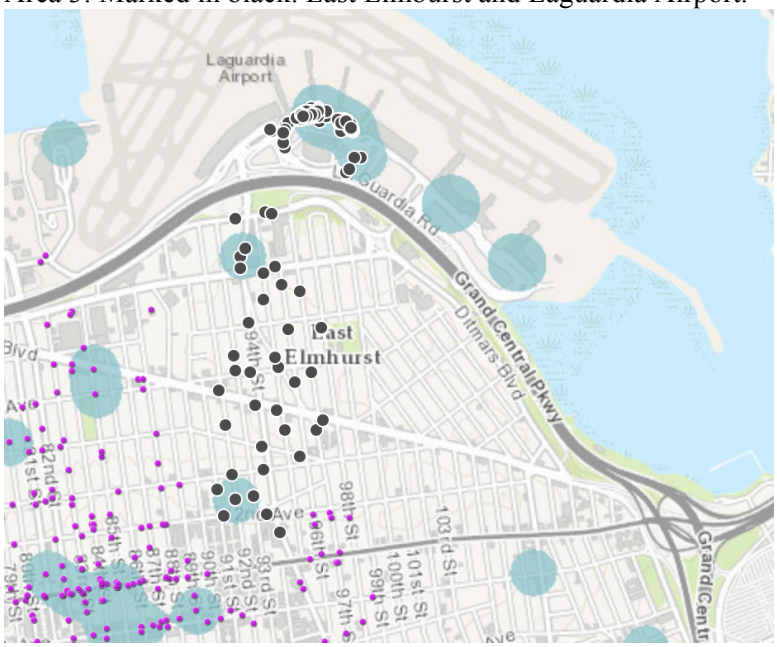
Area1: Marked in blue. Brooklyn District.



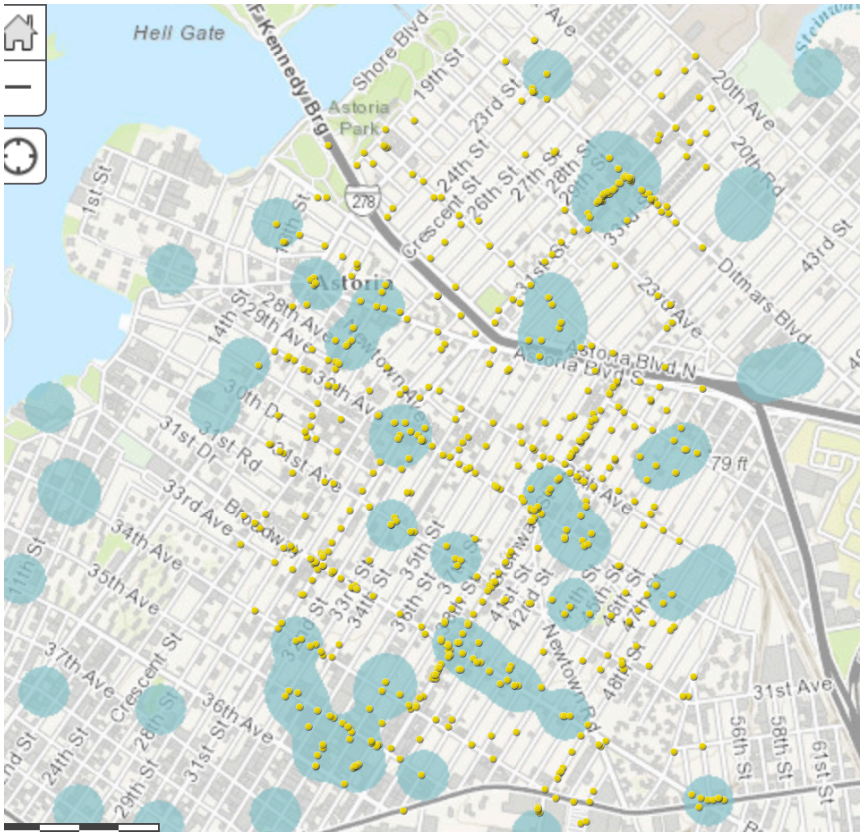
Area2: Marked in purple. Jackson Heights, Elmhurst, Woodside, north to Long Island Expy.



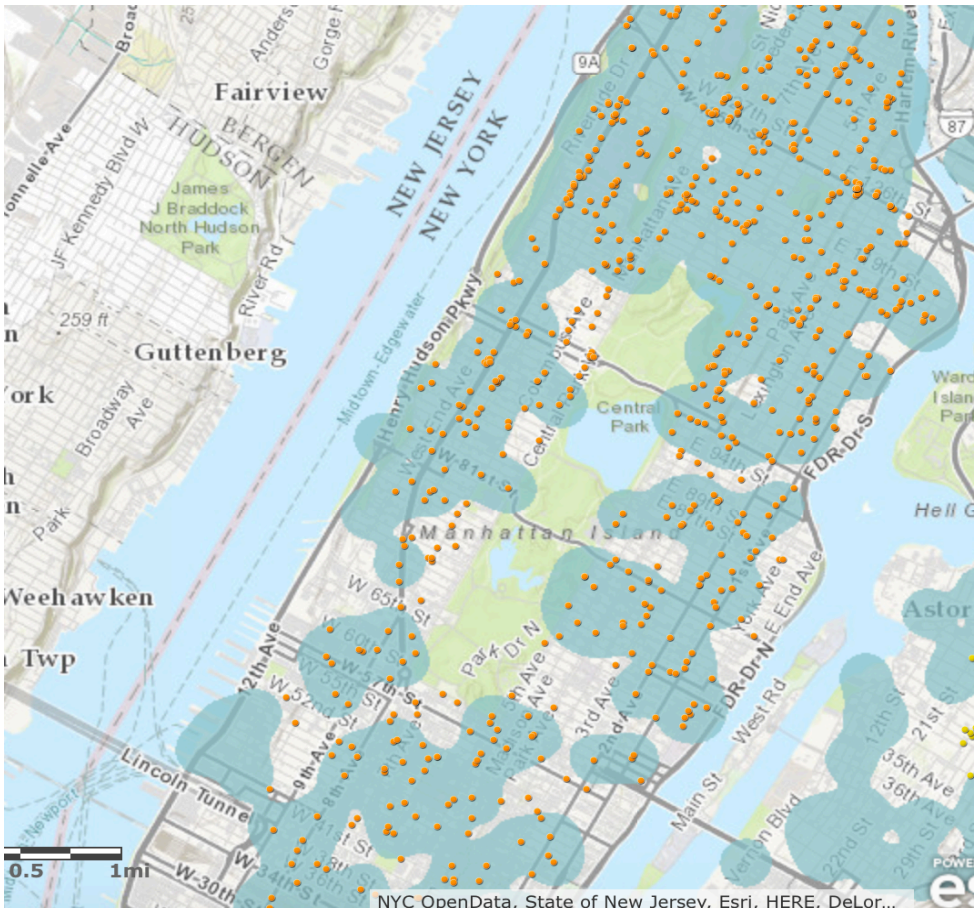
Area 3: Marked in black. East Elmburst and Laguardia Airport.



Area 4 : Marked in yellow. Astoria and Sunnyside, east to Long Island City.



Area 5: Manhattan District

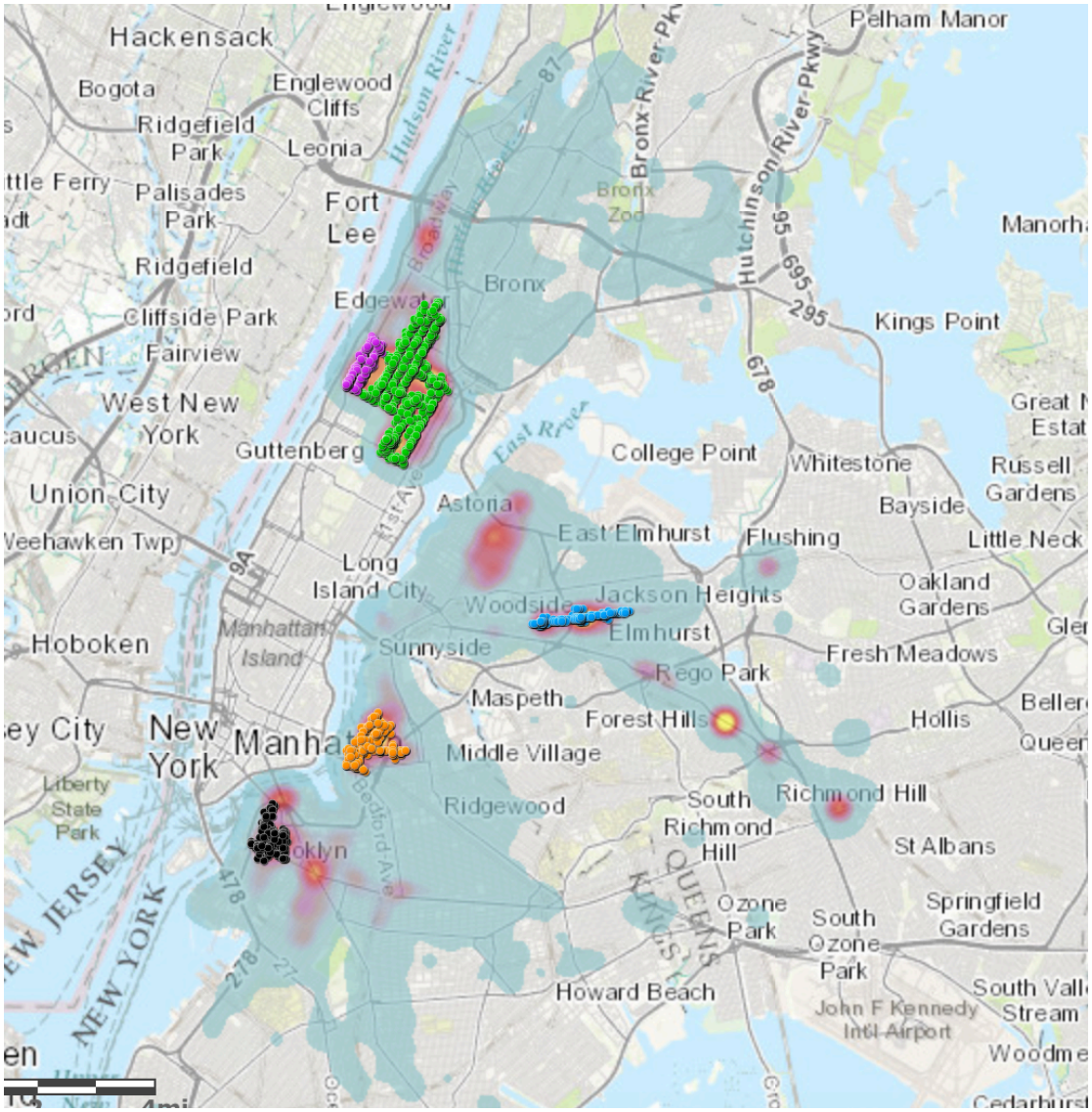


Question 4

The first cluster time interval is:
7:56:02 - 9:35:40
The second cluster time interval is:
14:42:51-16:44:18
The third cluster time interval is:
16:44:36 - 23:59:52

Question 5:

If we define lucrative trips as generating the highest fare for least amount time spent, what are the top 5 locations for the origin of the most lucrative trips?
If the fare per time spent of a trip is above the median of its peer, then it's defined as lucrative.

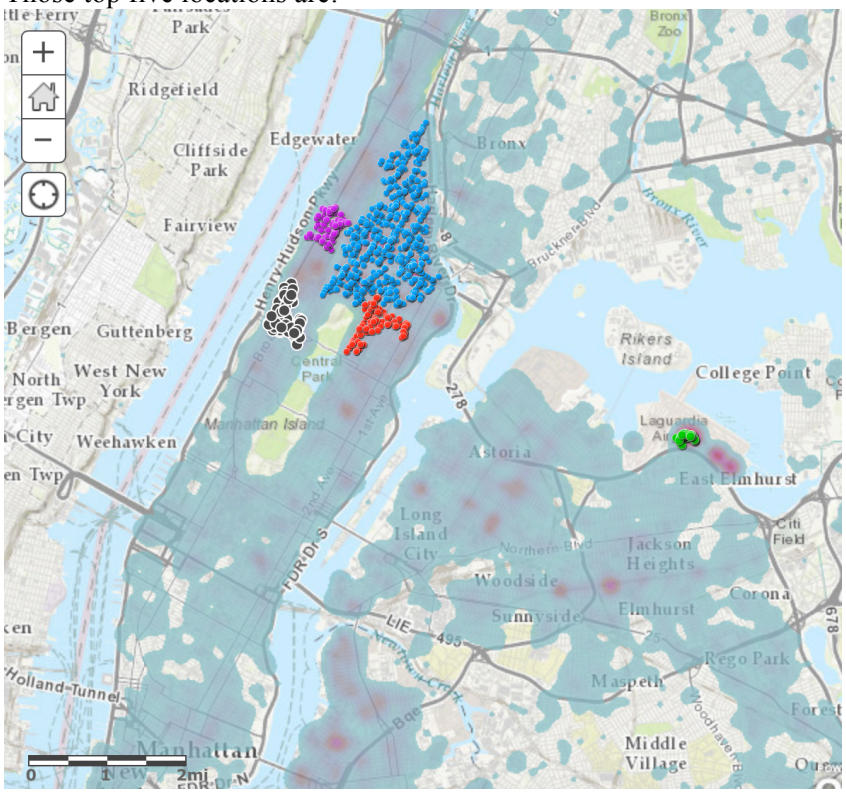


Area 1: couple of streets in Brooklyn Height and Brooklyn: Clinton St, Count st, Atlantic Avenue, North of Henry street Transit Bureau Hp.Insp, emergency medical dispatch (marked in black)
Area 2: Columbia University (marked in purple)
Area3: couple of streets near Williamsburg: Metropolitan Ave, Bedford Ave, Wythne Ave, Dnggs Ave (marked in orange)
Area 4: Roosevelt Ave. Especially around MTA –Jackson heights-Roosevelt Ave, and MTA woodside 61st (marked in blue)
Area 5: other parts of upper Manhattan. Starting from 97th street to north, end until W145th street (marked in green)

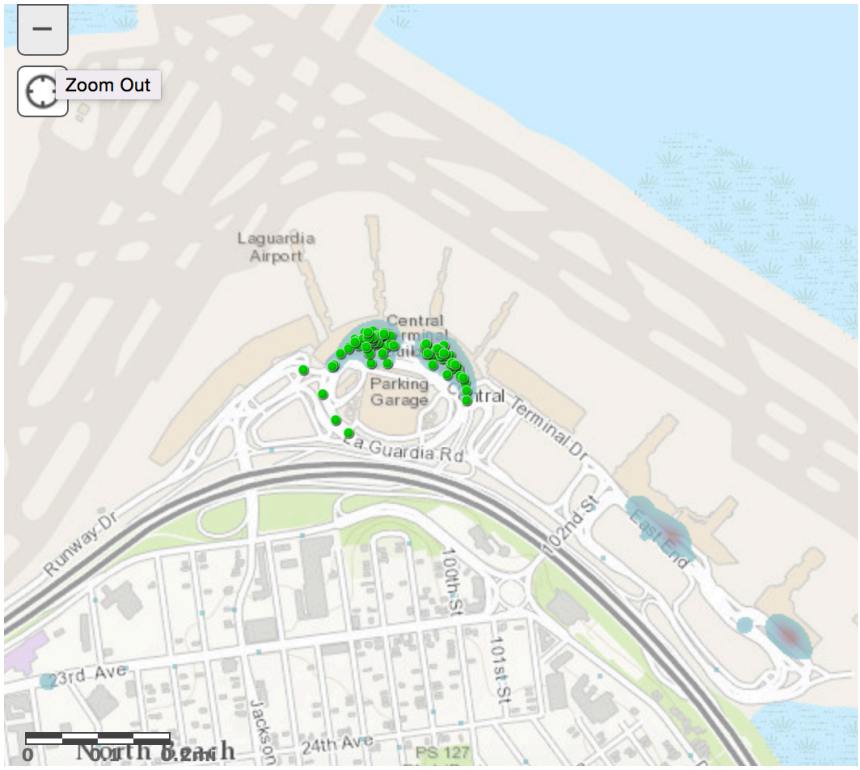
Question 6:

If we define lucrative trips as generating the highest fare for least amount time spent, what are the top 5 locations for the termination of trips?

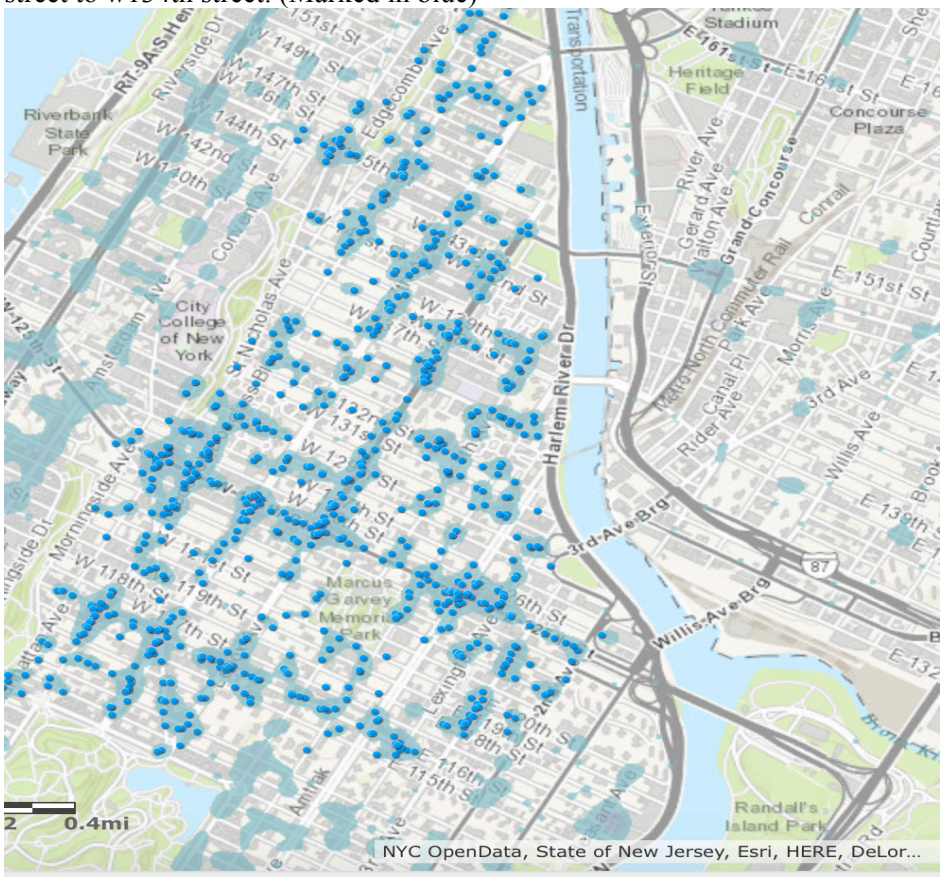
Those top five locations are:



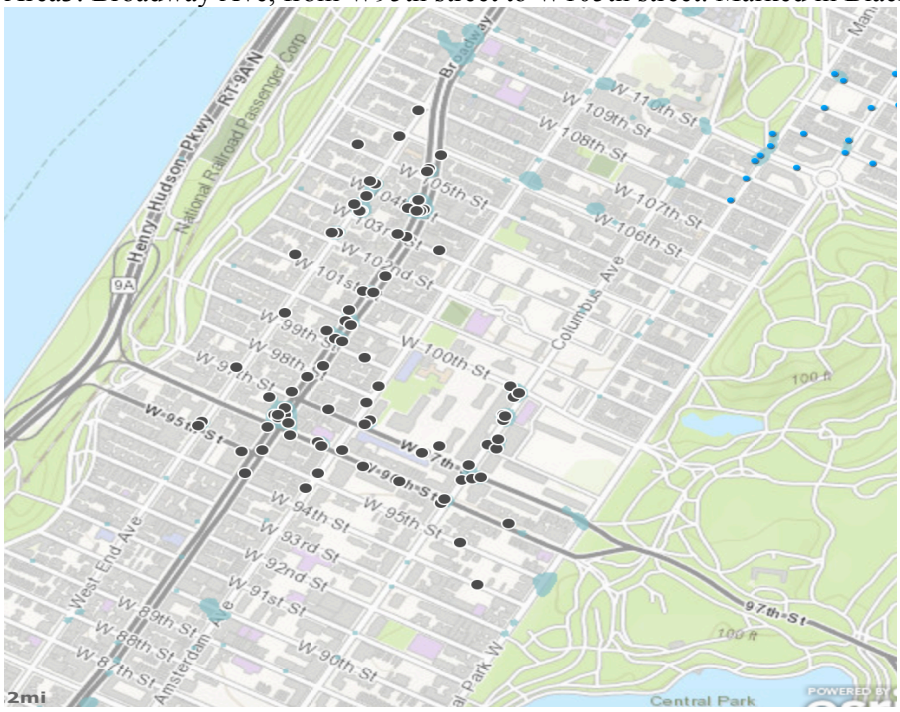
Area1: La Guardia Airport. Marked in green color.



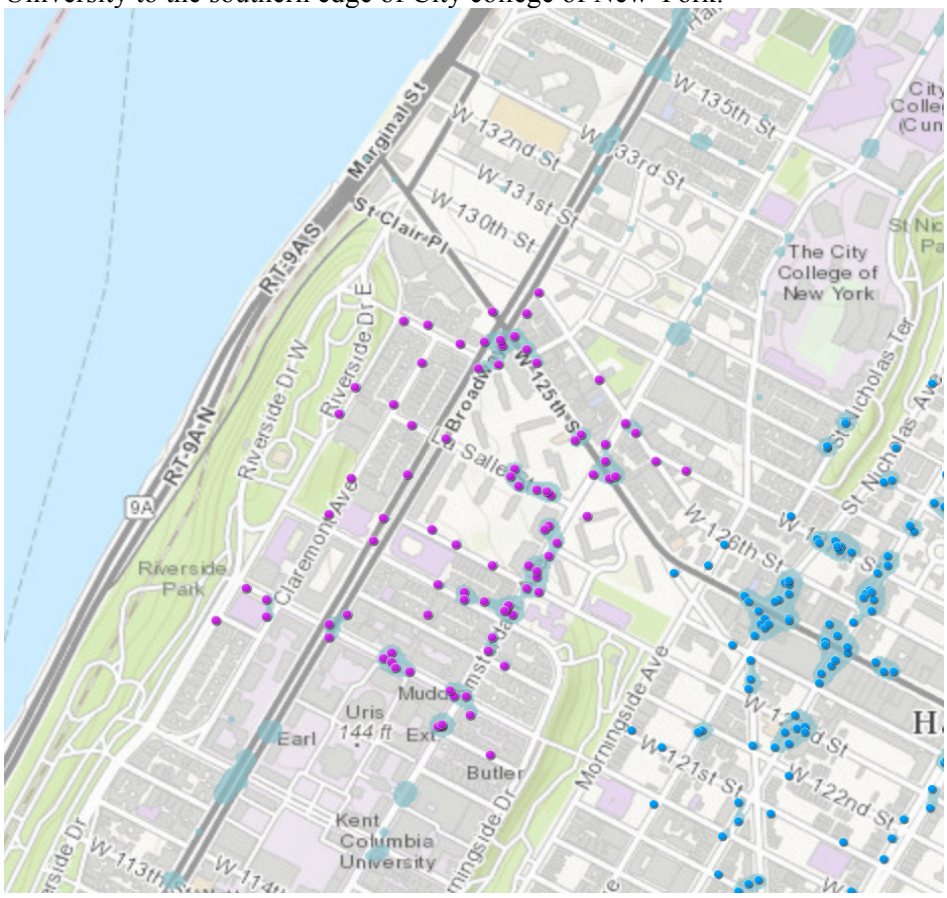
Area 2: Upper Manhattan, marked in blue color. Especially the area around Harlem, from w111th street to w134th street. (Marked in blue)



Area3: Broadway Ave, from W95th street to W105th street. Marked in Black



Area 4: Marked in purple: northern part to Columbia University, from north of Columbia University to the southern edge of City college of New York.



Area5: Northeastern corner of the Central park, especially E110th street to E112th street. (Marked in Red)