CMU Citadel Hackathon

James Hahn, Daniel Zheng, Haichen Li, Mo Li

3/25/17

Question

When initially analyzing the data, we immediately noticed the table containing demographics data about customers for all FHVs (For-Hire Vehicles). It not only contains the median age and medium income, but also the population, number of households for a given income bracket, and people per acre, or density of an area in other words, of a given NTA (Neighborhood Tabulation Area). These are all important characteristics of data that help with the analysis of a service's market. Maybe Uber happens to draw in younger kids in college and taxis pick up business men and women from the financial district. Or perhaps it is the other way around; either way, this let us to an interesting question. *Since 2014, how do the neighborhood pickup locations of Uber and taxi users compare? Is there any significant difference among neighborhoods demographics and services (Uber, Yellow Taxi, Green Taxi)?*

Another question we're interested in is to learn about Uber customers' demands over certain periods of time in a day, based on which the company can optimize the distribution of the vehicles to improve profits. One approach is to count the number of Uber services during NYC rush hours (5am-10am, 4pm-8pm) and non-rush hours (10am-4pm, 8pm-4am) using the dataset 'uber_trips_2014' and 'uber_trips_2015' and higher number indicates greater demand. Similar information on demands can be obtained using dataset for yellow trips, green trips and other FHV trips. Furthermore, we can also figure out the similarity/differences in the demographics of people between rush hours and non-rush hours by using the demographics dataset.

Non-Technical Summary

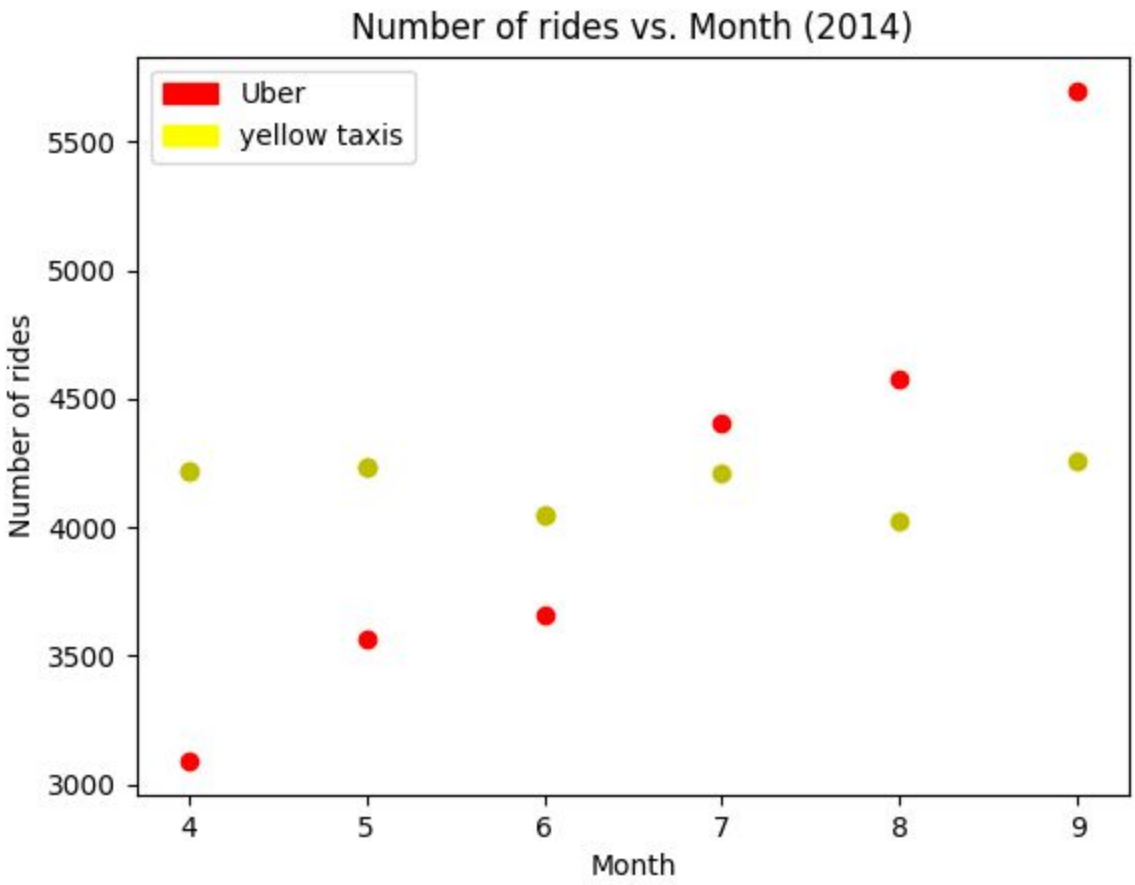
Most companies, whether it's Google, Facebook, Uber, Wal-Mart, or even a local pizza shop try their hardest to target their key demographics. For a clothes store on campus, their target demographic may be college-aged (18-22) women who are educated and do not have a job, so they may have less money. As for a company like Uber whose reach is city-wide, they must be keen as to not waste resources on sending their cars and drivers to obscure locations for one customer. For example, travelling long distances to catch a flight might make sense, but asking a driver to take a customer an hour away for a bowling alley may not be suitable. Because of this, the demographics data provided in this competition may be matched with the pickup locations of Ubers and the two major types of taxis in New York City.

Ideally, the data provided would cover every neighborhood of NYC uniformly. However, from further analysis, most of the FHV rides occur in Manhattan and a bit of Brooklyn. Luckily, the demographics table maps the demographics by NTA rather than borough, so the data is more precise to each location rather than taking the median income or median age for an entire borough like Manhattan.

As a result of analyzing the data, it was discovered that people with different levels of incomes show significant difference in their Uber ride patterns, especially in their pickup locations. Specifically, the pickup location is highly related with regions people live at. It has also been discovered that rich people tend to ride Uber much more often than people with lower incomes.

It’s somewhat counter-intuitive that the number of Uber service in rush hours and non-rush hours are not significantly different. That’s because the subway in NYC is very convenient, fast and cheap for commuters, which usually will not yield any delay for people who must go to work on time.

Technical Summary



In order to perform analysis on the NYC transportation data, we used a combination of Python and SQL. For Python, the libraries Pandas, Matplotlib, and GMPlot were utilized. We set up a Postgresql server and used Psycopg2 as a python interface.

As one can see, the typical pickup locations of yellow taxis and Ubers are similar. However, taxis dominate the market; the density of the heatmap in Figure 1 is higher than in Figure 2. To supplement this information, Figures 3-7 display the NTAs that specific income brackets belong to (0-30k, 30-60k, 60-90k, 90-120k, and 120k+). We generated these figures in the following way.

To reveal patterns of uber rides with people having different levels of incomes, we take advantage of the "pickup_location_id" column in the "uber_trips_2015" dataset, relate this column to the "location_id" column in the "zones" table, then relate the "nta_code" column in "zones" to the column having the same name in the "demographics" table. In this manner, we are able to figure out the NTA code for each Uber ride record in the "uber_trips_2015" dataset. Additionally, we convert the NTA code to latitude/longitude coordinates using information given in the "geographic" dataset. For each NTA code section, the "geographic" dataset gives us the coordinates of the vertices of the polygon

defining the NTA. We simply took the average point of the vertices and consider it as the center of the NTA, and assign the latitude/longitude coordinates of all Uber rides with pickup location in this NTA to the center. In this way, we can accurately estimate the pickup latitude/longitude coordinates of each 2015 Uber ride, with an error level of the size of the NTA.

By joining together tables “uber_trips_2015”, “zones”, “demographics”, and processed “geographic”, we were able to partition people into different income groups (<30000, 30000-60000, 60000-90000, 90000-120000, and 120000+), and visualize Uber pickup latitude/longitude heatmaps for each income group.

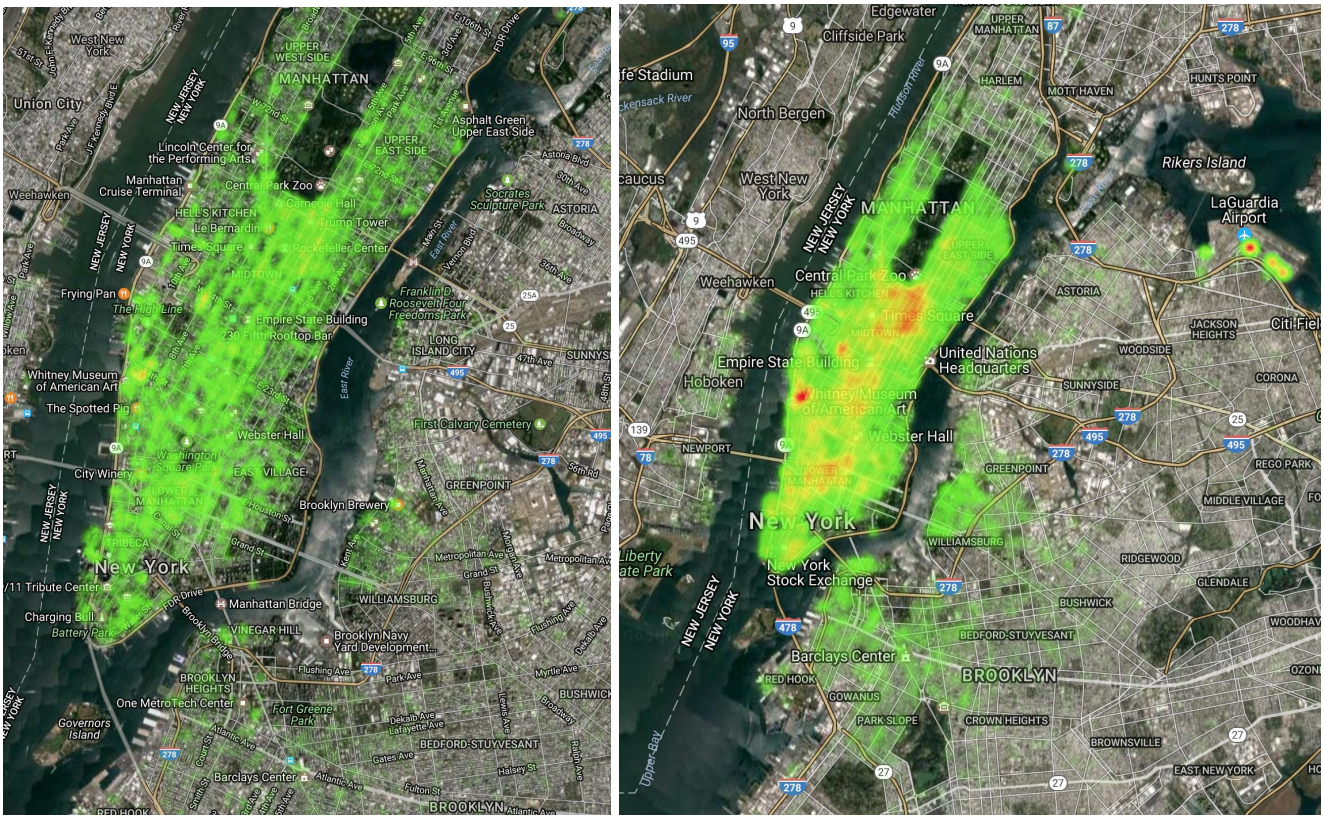


Figure 1. *2014 Yellow Taxi Pickup Destinations in NYC - Left is zoomed in, Right is zoomed out*

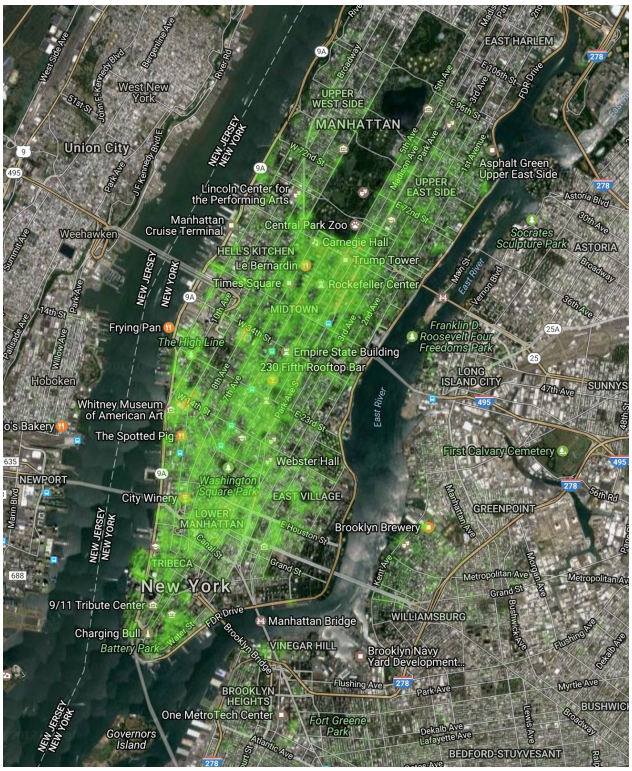


Figure 2. *2014 Uber Pickup Destinations in NYC - Left is zoomed in, Right is zoomed out*

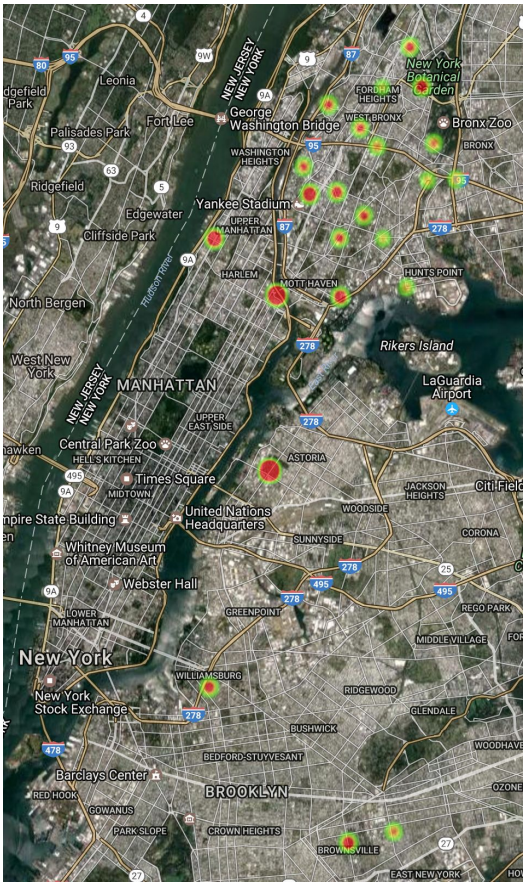


Figure 3. *Heatmap of people in 0-30k income bracket*

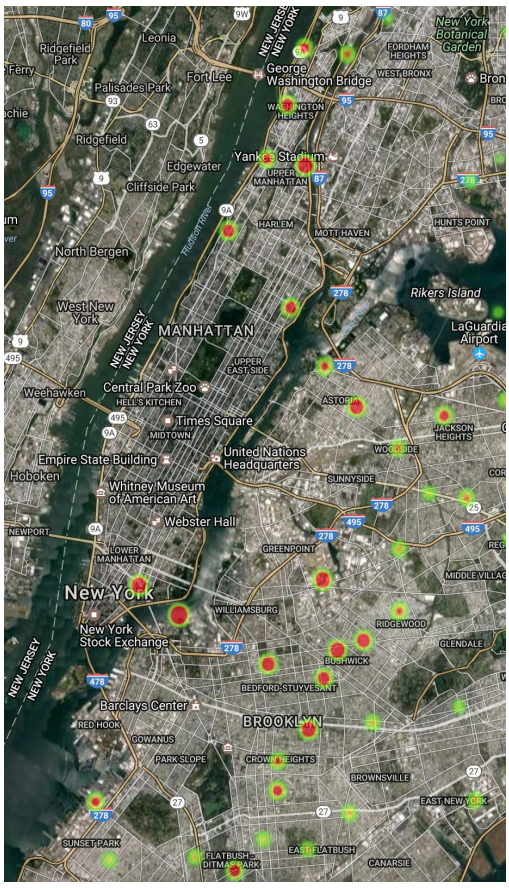


Figure 4. *Heatmap of people in 30-60k income bracket*

Figure 5. *Heatmap of people in 60-90k income bracket*

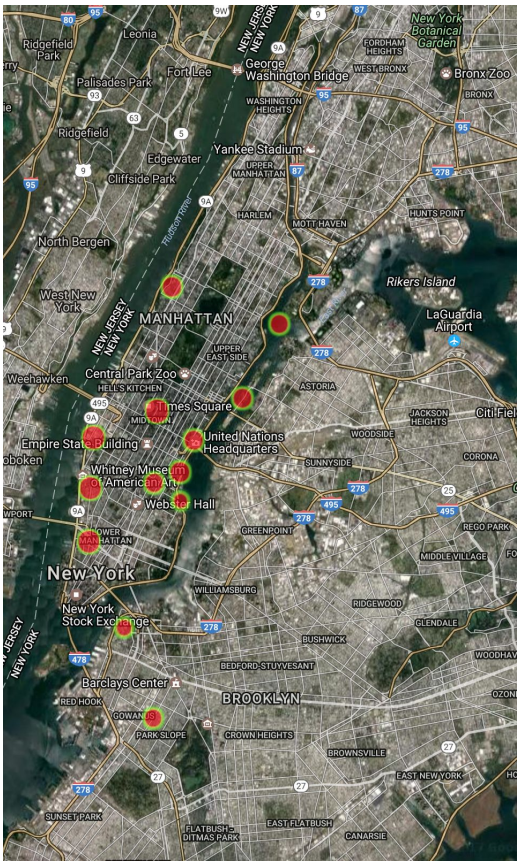


Figure 6. *Heatmap of people in 90-120k income bracket*

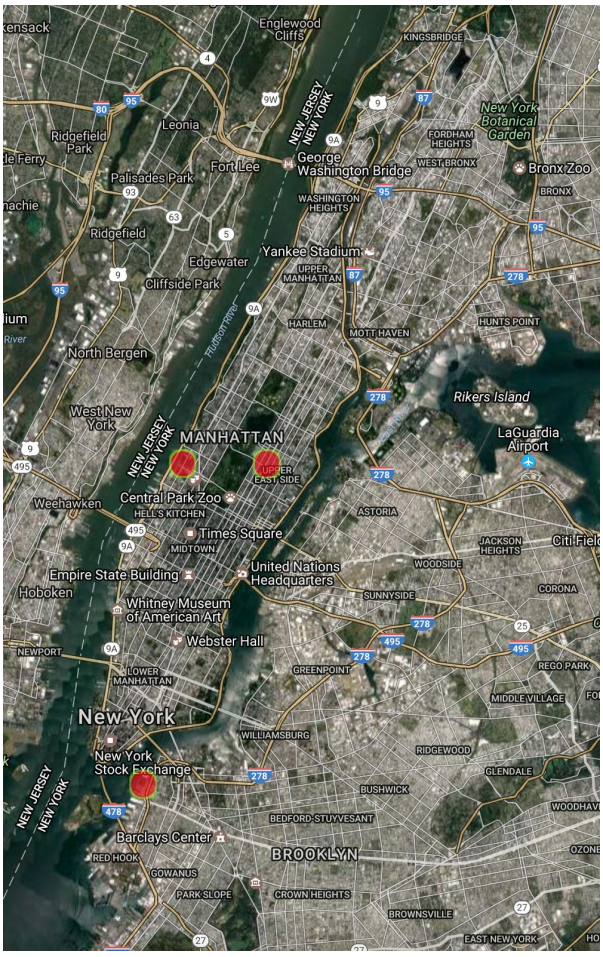


Figure 7. *Heatmap of 120k+ income bracket*