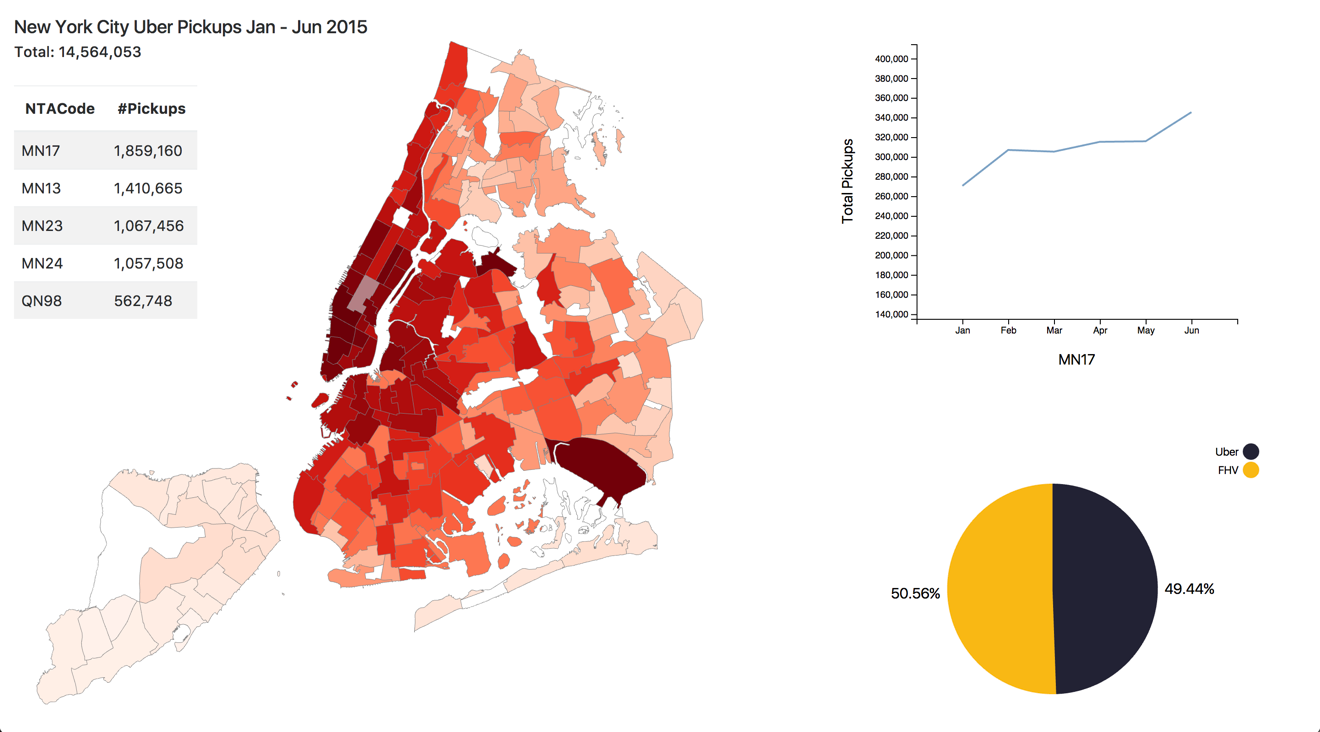
Project Report

CIS602-02 Data Visualization



Tu D. Luong

UMass Dartmouth - Spring 2017

TAble of  
Contents

Datasets 1

Tasks 3

Visualization evolution 3

Final Visualization 5

## Datasets

In this project, I will analyze about Uber trips data in New York city compared to other FHV (for-hire vehicles) companies. The data contains over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015, Uber data from January to June 2016. Data also contains 10 other for-hire vehicle (FHV) companies including Lyft company.   
  
The dataset’s URLs include:

1. Unified New York City Taxi and Uber data (<https://github.com/toddwschneider/nyc-taxi-data>)
2. New York Map Data (<https://raw.githubusercontent.com/se2/cis602-02-project/master/design/nyc.geo.2.json)>

The datasets of Uber and FHV (For-Hire Vehicles) from Todd Schneider are raw datasets which are approximately 1.3 billion taxi and Uber trips originating in New York City. Due to the limitation of the time and effort in this project, I will only analyze Uber and FHV data from Jan to Jun in 2015 (approx. 14.5 million Uber pickups and approx. 15.5 million pickups for other FHV types). Below is the format of original Uber dataset:

|  |  |  |
| --- | --- | --- |
| **Dispatching\_base\_num** | **Pickup\_date** | **locationID** |
| B02617 | 2015-05-17 09:47:00 | 141 |
| B02617 | 2015-05-17 09:47:00 | 65 |

The FHV dataset format is similar:

|  |  |  |
| --- | --- | --- |
| **Dispatching\_base\_num** | **Pickup\_date** | **locationID** |
| B00053 | 2015-01-01 01:05:00 | 45 |
| B00053 | 2015-01-01 01:30:00 | 141 |

Dispatching base number is a unique number provided by TLC Company (<http://www.nyc.gov/html/tlc/html/industry/base_and_business.shtm)> for all FHV including Uber, format of FHV bases is as following:

|  |  |  |
| --- | --- | --- |
| **base\_number** | **base\_name** | **dba\_category** |
| B00001 | London Towners Inc. | Other |
| B02395 | Abatar Inc. | Uber |

New York map data is a **FeatureCollection** type with NTA Code (Neighborhood Tabulation Areas Code) for each small area (<https://raw.githubusercontent.com/se2/cis602-02-project/master/design/nyc.geo.2.json>), each NTA Code has a unique location ID which can be retrieved from <https://github.com/toddwschneider/nyc-taxi-data/blob/master/data/taxi-zone-lookup-with-ntacode.csv> with following format:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **location\_id** | **borough** | **zone** | **service\_zone** | **ntacode** |
| 1 | EWR | Newark Airport | EWR | NJ01 |
| 2 | Queens | Jamaica Bay | Boro Zone | QN61 |

We have a locationID attribute for each pickup from both Uber and FHV data above, this locationID can be used to retrieved ntacode, and therefore, we know how many pickups from each NTA area. In this project, due to the tremendous amount of data from both Uber and FHV (more than 1GB in total for both), I have calculated and filtered down the data to the minimal datasets including only attributes I need, this will enhance the performance of d3.js queue function. The final format of pre-processed dataset is as following, both Uber and FHV:

[

{

"key": "MN17",

"value": 315238

},

{

"key": "MN13",

"value": 237467

},

{

"key": "MN23",

"value": 182768

},

…

{

"key": "SI32",

"value": 2

},

]

Thus, base on the above dataset, I can create a choropleth map of NYC with density of Uber and FHV pickups for each NTA area.

## Tasks

There are some questions I want to find out when making the visualization:

1. Which places have the highest density of Uber and FHV pickups? And top 5 highest NTA areas of Uber and FHV?
2. The trending of Uber over first six months of 2015, represented by total number of pickups each month?
3. The sharing of Uber and FHV (in percentage) for each NTA area.

To answer those above questions, I need to create a colormap, a line chart, and a pie chart with the datasets I have finalized.

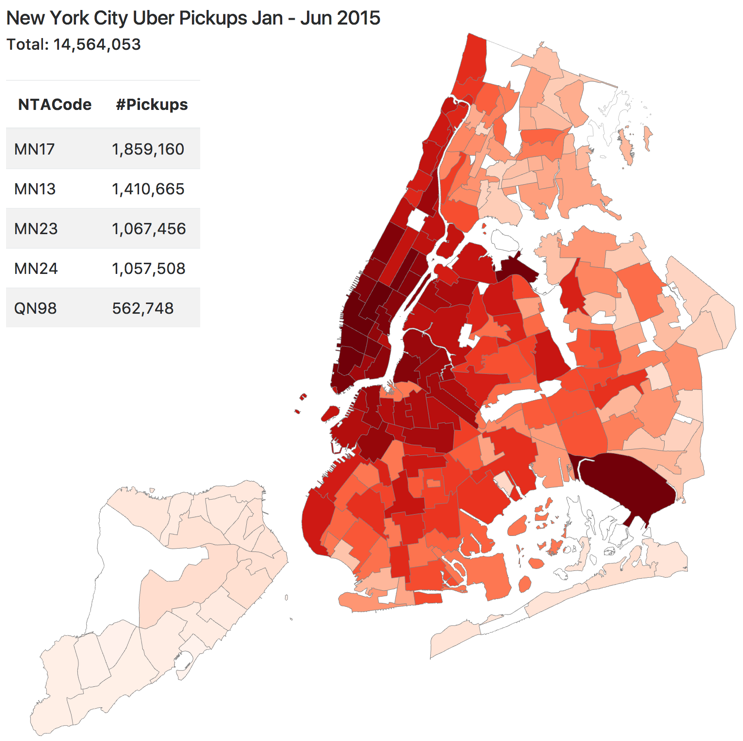
## Visualization evolution

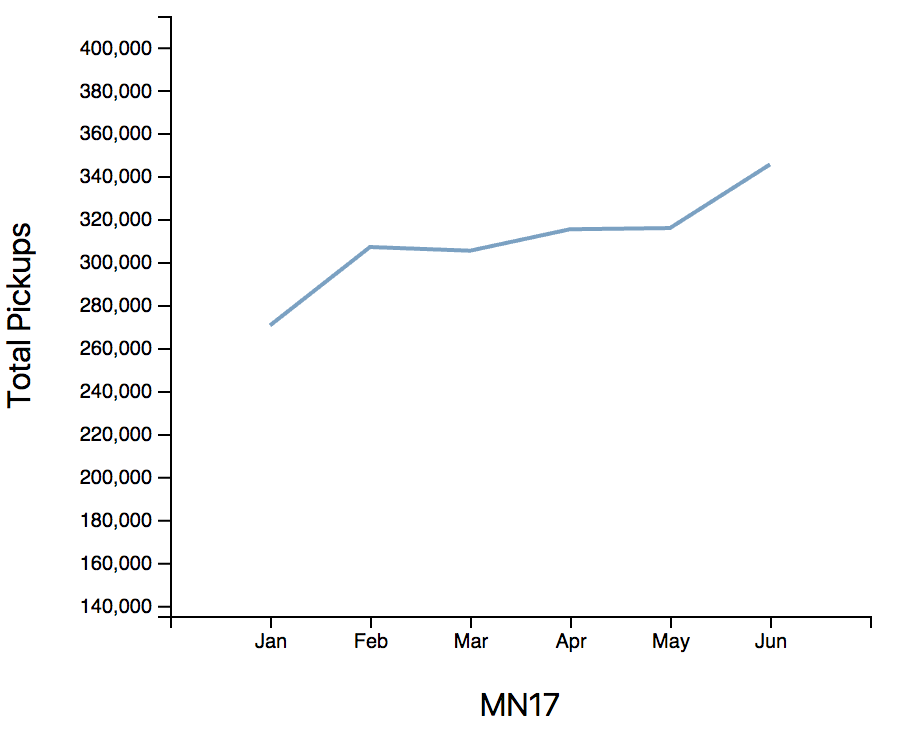
My visualization focuses on plotting density of Uber pickups via choropleth map of NYC, emphasizing the trending of Uber using via the consideration of the increase of total pickups over months, and weighing the share of Uber rides in percentage. At first, I intended to plot data base on latitude and longitude of each individual pickup and use some sort of programming technique to count how many dots in a polygon in the colormap, however, this technique is quite performance-costly and not quite correct, so I decided to use another dataset with location ID including in each pickup (as above dataset description). My first design looked like as following  
  


Those red “dots” in the map represent the position of pickups encoded by Latitude/Longitude attributes.

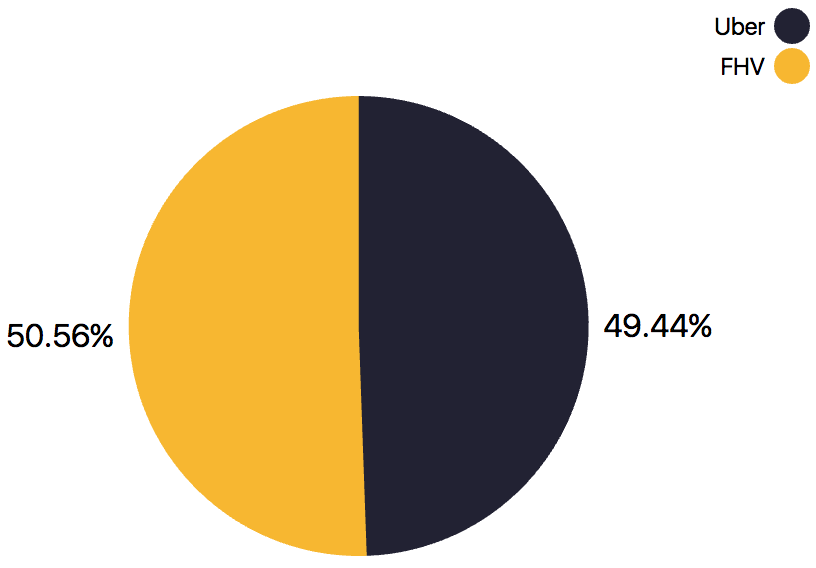
## Final Visualization

As you can see, using Latitude/Longitude to plot pickups is quite messy and overlap in visualization, so I decided not using this, instead, counting number of pickups based on location ID and display by color saturation in my final visualization.

Density plot of Uber pickups represented in a choropleth map  


Trending in using Uber can be seen for each NTA area, when we click on that area on the map, a simple line chart displays a trending of Uber using over six months from January to June 2015  


Another attribute I want to see is that how many percentage of Uber comparing to FHV in the same NTA area, this could be analyzed via a simple pie chart



Marks & Channels:

|  |  |  |
| --- | --- | --- |
|  | Marks | Channels |
| Choropleth map | Area  Polygon | Color saturation |
| Line chart | Line | Position |
| Pie chart | Area | Size  Categorical color |

My visualization is based on multiple view principal, with main part is the choropleth map, and supporting views are line chart and pie chart. In future work, I want to analyze more about number of pickups broken down in weekdays & weekends; moreover, comparison between NTA areas or between boroughs is also my wish to accomplish in this visualization.