TEAM 17：

Uber – the Success of Shared Economy

## Introduction:

America has been long blamed for its terrible public transportation, but there is one exception - New York. You can totally reliable on subway, buses and taxi rather than on your own car. With the rapid development of share economy, Uber has also become a popular choice. Having given tons of data about NYC Uber trip, we are able to analyze the demographics feature of Uber user, and Meanwhile, some prediction can also be made about the relationship between area’s features like population, income and Uber trip counts.

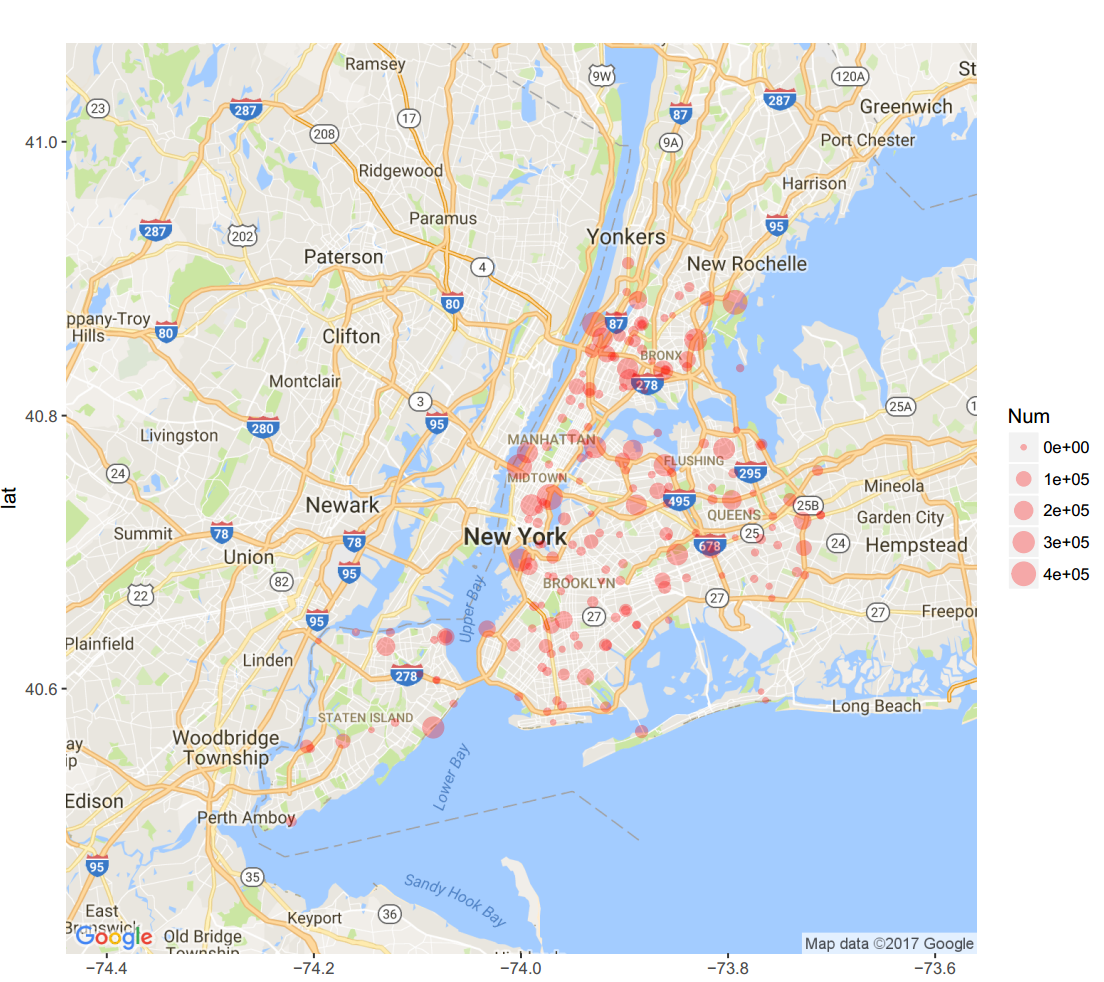
**First, we visualize the data with animations: (Shown in the mp4 file)**

Here are the questions we want to solve:

1. What is the demographics feature of Uber user?
2. Which’s the busiest area in NYC? How many people in total use Uber in that area in 2015?
3. Using linear regression, we try to predict customer behavior and predict the number of potential user in each area?

Guess what? We do find something interesting.

**Busiest Area**

Where is the busiest area in NYC? In other words, where can Uber driver pick up more potential rider? Using Uber trip count 2015 dataset, we have the answer. By sorting pickup station with zone ID, we visualize it by scatter plot, with the size of point indicating the total number of rider. As we can see, places near Center Park are the busiest area in NYC. Outside of Manhattan, Brooklyn, Flushing and Queens also have a buddle of Uber users.

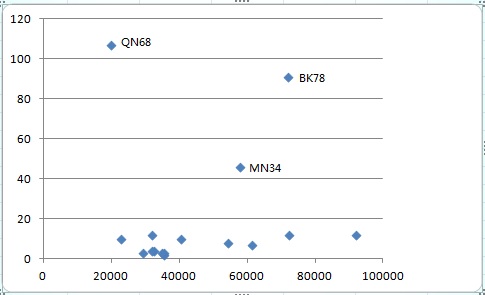
The table below give the top 10 busiest taxi zone with their approximate longitude and latitude.

**Table 1. Top Busiest Zone in NYC**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | | User counts | | longitude | | latitude | |
| MN20 | 460732 | | -73.97220 | | 40.74041 | |
| MN35 | | 420356 | | -73.92873 | | 40.86675 | |
| MN13 | | 419045 | | -74.00153 | | 40.76265 | |
| BX03 | | 407591 | | -73.79323 | | 40.88283 | |
| MN25 | | 323989 | | -74.00078 | | 40.69429 | |
| BX37 | | 315919 | | -73.83129 | | 40.85544 | |
| QN27 | | 299781 | | -73.86110 | | 40.76367 | |
| SI45 | | 296734 | | -74.08469 | | 40.57149 | |
| QN71 | | 290550 | | -73.92828 | | 40.77691 | |
| QN51 | | 285518 | | -73.80379 | | 40.77562 | |

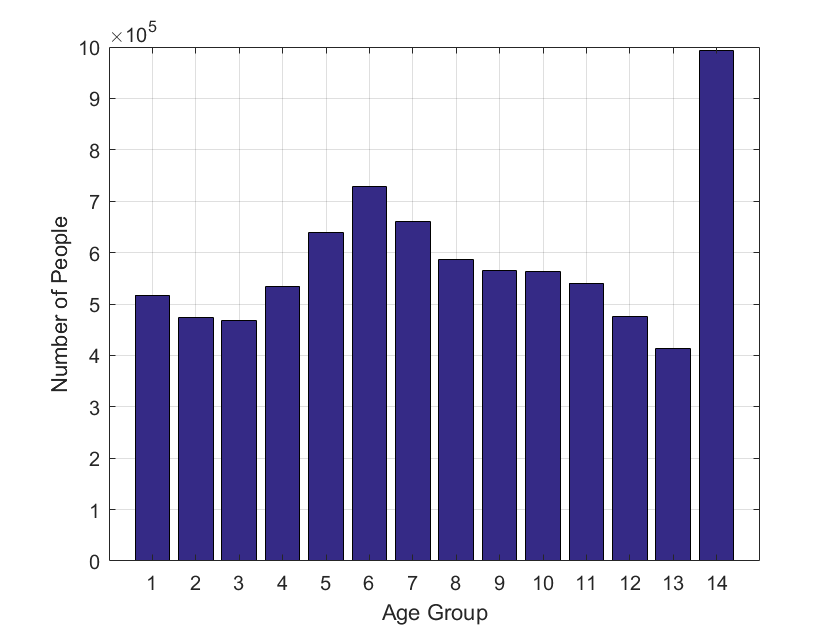
## Statistics Analysis:

To get the relationship between district’s features like population, income, age distribution, etc, we decided to do multi-linear regression on the relationship between these features and uber trip count. And according to cluster analysis, these relationships can vary a lot across populations with different income level, thus we divided the population according to their income level. We did regression on the population whose annual income is between 40000-50000 dollars. We looked into the first 1 million data and see how many people has used uber in this area. And we found there are three abnormal points. The result is shown as below:



From the results, we can conclude that people from rural area are more intended to take uber. This would pose a great effects to our regression results. So if we want to do regression and have a good result we need to exclude these points.

**Demographics Feature:**

**As the society is aging**, we can see that the old people in NYC take a large proportion in total population. The trend shows the great need to develop Uber

## Conclusion:

With the development of autonomous driving and aging population, it can be seen in the near future that the industry would have changed completely. We can have more efficient ways of travelling than ever before.

## Future work:

With more data available, we would be able to use universal function approximator like deep learning to do the regression. Besides, it is also interesting to discover how fast people are likely to adapt to new trends like uber or autonomous driving, this information is very important in trading as it might be challenging to decide when is the best time for long and short certain options.

## Appendix of failed attempts:

1. We looked to aggregate the whole trip count for 2015, but we failed to have the final useful data, we have uploaded what we have get in Github. We use the false data for regression and that takes us some time. Eventually we should exclude some abnormal points to have a better regression with population.

2. Relating uber\_2014, green\_trips, yellow\_trips dataset with demographic data:

With the current uber\_2014, green\_trips and yellow\_trips dataset, there is no relation to the NAT areas given, but only the latitute and longitude of the data. Therefore, in order to find the correlated NAT code, we tried to implement an algorithm, that takes a longitude and latitute value and checks if it is present in a NAT area. NAT areas are decribed in the geographic dataset, which give the area in longitude and latitute values.

With that algorithm, we would be able to include the datasets of uber\_2014, green\_trips, yellow\_trips in our other analysis.