Our Goal for the Project is to answer the following questions,

What strategies can NYC taxi drivers use to make more money?

We are going to look into NYC taxi data for two distinct seasons, (Summer 2016: June, July, August, Winter 2015: December, Winter 2016: January, February) to find out how NYC taxi drivers can make more money. Using data analytics, NYC taxi drivers can find different patterns in fares and using these patterns to maximize their profits. Also, using visualizations, we can easily explain the patterns and “Taxi Economics” in NYC. And now we done the date cleaning section, here is the obstacles we encountered; When choosing datasets, we ended up using data from June 2015 to August 2015, instead of June 2016 to August 2015, which was different from what we have claimed to use in our proposal, because the data of Summer 2015 do not have coordinates of pick up locations and drop off locations. Locations in those three datasets are presented as IDs. We excluded records that have 0 latitude or longitude, and also we eliminated records with less or equal to 0 amount of fare, and less than 0 amount of tips (including 0). It came to our notice that some records have unusual large amount of fare, which beyond our normal expectation. For example, one record in Dec 2015 dataset has a total fare amount of $825,998.61. We kept these records at this stage. We will take a further look into them later. We planned to transform coordinates into neighborhood in our proposal, so we can find travelers in which areas tip the most. However, due to API (geopy library) restrictions, we are going to tackle the problem in a different way. We will find the areas with most travelers, and then convert coordinates in that area to neighborhood. In this way, we are expecting far less calculations.

For the tools we use, we use Spark and SparkSQL to handle the huge datasets. Also geopy for converting the coordinates to the neighborhoods. AWS for the spark environment, S3 for storing the cleaned datasets. We are using Bitbucket for version control. In the following weeks we are going to use Bokeh, matplotlib and google chart for visualization. Lastly, the python package scikit-learn for Machine Learning.

Group Zymic:

Michael Chon, mc2153@georgetown.edu

Yuan-Yao Chang, [yc704@georgetown.edu](mailto:yc704@georgetown.edu)

Zheng Chai, zc104@georgetown.edu

Chong Zhang, cz211@georgetown.edu