Course: Big Data Management and Analysis
Group members: Ding Chao Liao, Weifan Lin, Minghui Wang, Satya Gupta
Date: April 6, 2016

Citi Bike - Taxi Analysis

Motivation & Objective

Traffic has been a major problem in the city of New York, making the life of a commuter, tedious and time consuming. New York is one of the few cities that have the longest commutes. Because of this heavy traffic, we want to possibly reduce taxi usage involving trips within a mile in New York City. On the other hand, we want to increase the coverage of Citi Bike station throughout the city in order to increase accessibility.

The goal of our project is to find the correlation between Citi Bike trip and taxi trip data sets in the city. In case if there exists a correlation between these data sets, we can use our analysis result to determine certain area where Citi Bike stations are demanding.

Data Source

Two primary data sets are:

1. Citi Bike's trip data from July, 2013 to February, 2016
   Data source: https://s3.amazonaws.com/tripdata/index.html
2. NYC taxi trip data from 2009 to 2015
   Data source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Methodology

In order to find correlation between citi bike and taxi trip data, we have to metastasize the data. Steps are as follows :

- Filter the taxi trip data to get the trips within 1 mile
- P1: Analyze taxi trip data, which is from 2009 to 2013, before the existence of Citi Bike,which is from 2013 to current.
- P2: Analyze taxi trip data, which is from 2013 to current, with the existence of Citi Bike.
- Compare p1 and p2
- Analyze Citi Bike trip data

Then, we need to compare the data and find the possible correlation.

- Correlate the pickups and dropoffs of p2 with Citi Bike start_station and end_station

Finally, we will analyze and visualize the data and produce a report with respect to certain area if it demands more citi bikes.

Big Data Challenges

Taxi trip data is relatively larger than Citi Bike trip data. The way to metastasize them is going to be a big challenges for us. The way to plot the data point into a city map is going to be hard as well. Although the data will be filtered to a smaller dataset, how we are going to use some visualization to represent the relation between taxi and citi bike will be a obstacle we would need to tackle.

Deliverables

A report will be provided. It will consist of all the data we have generated for the project, which is informative and analytic at same time.

A visual representation of relation between citi bike and taxi which may contain potential Citi Bike station locations for CitiBike Corp. With these locations CitiBike Corp can make Citi Bike more accessible to the public.