SI 618: Project Proposal

Dataset

The dataset that I have chosen for this project is the New York city taxi trip data. This dataset includes trip records from all the trips completed in Yellow taxis in January 2015. The records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types and driver-reported passenger counts. The data 1s approximately 2GB in size and is obtained from http://www.nyc.gov/html/tlc/html/about/trip_record.shmtl and is in CSV format. All the variables except pick-up and drop-off time are of numeric type. The pick-up and drop-off variables of type datetime. This data is made available by the Taxi and Limousine commission of New York city.

Questions to answer

Some of the questions that I hope to explore in this dataset but not limited to are:

1. Find out in which area is a heavy taxi activity (both pick-up and drop-off)
2. How much time does it take to travel from midtown, Manhattan to JFK, Newark and LaGuardia airports?
3. How many drop-offs have been made at firms like GoldmanSachs and Citi group which are located in Manhattan?
4. What is the tipping trend when the payment method is credit card and cash?

Exploratory Data analysis and visualization methods for the above questions

1. SqLite3 and ggmap
2. Sqlite3 and ggplot2
3. ddply and ggplot2
4. Sqlite3 and qplot