CMSC 123 Project Written Proposal
Team Name: KASA

Team Members
Kei Irizawa, Abdallah Aboelela, Adam Oppenheimer, Swayam Sinha

Data Set
Taxi Trips (Chicago Data Portal)
URL: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

Dataset overview
113 million Taxi trips reported to the City of Chicago in its role as a regulatory agency. Each trip has an associated Trip id and Taxi id (anonymized medallion number), various variables for the time the trip was taken, the time it lasted, and the distance travelled. Pickup and dropoff locations in terms of community and census tract. Financial variables are also included (fare, tip - not incl. cash, method of payment, extras). Finally, there is more pickup and dropoff information, as well as the name of the taxi company.

Hypothesis/Questions
1. Most interested: Traffic levels (please look below at Traffic Metrics).
2. How often do taxi take shortest possible route from pickup to drop-off location? (compare shortest path using Dijkstra's Algorithm and trip miles)
3. Connection with crime data: do people take more taxi in areas with more crime (community area or location in terms of lat/lng) ?
4. Do drivers have specific location taxi in or random (can identify taxi by taxi id) ?
5. What kind of drivers most likely to receive tips?
6. Is there a relationship between the cost of the trip and payment type ?
7. Do certain companies have better drivers (have taxi company information) ? scale of how good driver is could be measured in terms of tip amount.

Traffic Metric
1. Find trips that have close starting and ending location (based on radius) and compare time it took to make that trip.
2. With the starting and ending location of taxi trips, get the shortest path nodes (which represent intersection between street and avenue). Count the number of times different taxi go through a node for a given time interval. Use this to measure traffic level of a specific intersection at a certain time interval.
3. Get the average speed of taxi given every combination of trip distance and time. Compare this average speed with a trip and evaluate traffic level. Issue is that there could be highways with higher speed. We could get a collection of highway nodes and eliminate these nodes when measuring traffic level.