Catching a Ride in the Big Apple

Team Members:
Spencer Milbrandt
Rasheeq Jahan
Peter Wang

Picking Up and Dropping Off

Using the NYC Taxi Data Sheet, we will determine whether or not there is a correlation between pick up location, drop off location, and total amount. With this information we plan to be able to predict the most likely destination based on pick up location which would help determine the viability of carpooling based on fare amounts. Additionally, studying the likelihood of finding an available taxi at any given location. Another possibility will be to predict if weather has an effect on pick up and drop off locations as well as taxi fares.

What Has Already Been Done?

Related Work:

- A study of spatial data from taxi trips found the source of uneven location distribution due to “land use, function of major establishments, and urban layout” (Zhang, Zheng, and Xiang He 1234).
- Research on time series forecasting “present[ed] a model for predicting the number of services that will emerge at a given taxi stand” (Moreira-Matias, Luis, et al 1394).

Datasets In Use:

- http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml (Taxi Data NYC 2016)
- https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016 (Weather Data NYC 2016)
- https://data.cityofnewyork.us/Transportation/Uber-Trips-NYC-2016/gt3n-7ri6/data (Uber Data NYC 2016)

Proposed Work and Tools

The datasets are already fairly cleaned therefore we will mostly just need to do simple preprocessing steps, data mining steps, and post-processing steps.

Preprocessing: Some Data Cleaning (Removal of unnecessary attributes and NULL values), Data Integration, Normalization, Dimension Reduction

Data Mining: Pattern Discovery and Clustering (Important), Classification, and Outlier Analysis

Post-Processing: Pattern Evaluation, Selection, Interpretation, Visualization (Spatial Data Map)

Tools: SQL, Python, WEKA, Python Libraries: numpy, sci-kit, pandas

Evaluation of Taxi Data

Be able to determine pick up and drop off hotspots to give suggestions on where one would likely be able to find a taxi ride at any given time. This suggestion would adjust based on the information presented from weather conditions that may change the average distance, taxi fares, and influences from relevance of Uber and Lyft business. Given the hotspot data and taxi rates, individuals would be able to decide whether carpooling is viable at that given time or when to expect a taxi at a given time.

- Spatial data mapping
- Time series analysis over seasons or days of the week

Works Cited

- Moreira-Matias, Luis, et al. "Predicting Taxi-Passenger Demand using Streaming Data." *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, 2013, pp. 1393-1402.

- Zhang, Zheng, and Xiang He. "Analysis and Application of Spatial Distribution of Taxi Service in City Subareas Based on Taxi GPS Data." .