Charley Ferrari

IS 607 Final Project Proposal

Kaggle Data: <http://www.kaggle.com/c/bike-sharing-demand>

Citibike Raw Data: <http://www.citibikenyc.com/system-data>

Example WUnderground hourly weather page: <http://www.wunderground.com/history/airport/KNYC/2014/11/19/DailyHistory.html?MR=1>

For my final project I’d like to recreate the Kaggle data I used for project 2 with the raw NYC bike share data. The Kaggle data contained the number of trips taken hourly in Washington DC, along with several explanatory variables of the weather for that hour. Citibike, New York’s bikeshare system, offers very granular trip data, showing the timestamped origin and departure data for each ride. WUnderground provides hourly summaries on a per day basis that can be scraped. Their URLs contain the date, and can therefore be looped through in a fairly straightforward way to obtain all the needed weather data.

I’ll need to play with the data to find out exactly how I’d like to store it. The trip data and hourly weather data seems like they could be pretty easily put into a relational database (just by taking the hour out and keying that with the hourly weather data.) I would love to put something like this in Neo4J. This lends itself more to a flights database model rather than my trade or travel time models (meaning it makes sense to have a separate “trip” node rather than just having a relationship between bike stations.)

The main statistical analysis will be to perform an LS regression model that can be used to predict future traffic. You asked during project 2 if I ended up submitting anything. I unfortunately wasn’t confident in anything I thought I was capable of creating and ended up just using this as a practice. My experience is in econometrics, so seeing that many variables presented in a categorical way took a bit to wrap my head around! I could have just thrown month and year into the model as their own categorical variables, but ideally wanted to find a way to extract the overall monthly trend and try to predict the deviation from that average. I’ll have to toy with this a bit to find out what I actually want to do. Having the raw trip data will also allow me to answer other questions (for example: from Neo4J it’ll be easy to answer the question of which two stations had the highest number of trips between them.)

For graphics I’ll at least be able to provide summary graphics, although I would like to toy around with some sort of mapping, or Neo4J visualizations. I could try to tag bike stations to neighborhoods for example and try to get an aggregated view of where people are travelling.