INFO 3300 Project 2

Alex Fernandez/af394
Jacob Rauch/jer322
Jatin Bharwani/jsb399

Work Done By Each Team Member

Alex started the project with getting the data and trimming it down to a reasonable size and to relevant portions. He also created the time label, weather label and tooltip like pop up information near the top of manhattan. Jatin focused on the graphs at the bottom, and also created the interactive radio button system. Jacob created the slider element, the dropoff dot feature and heavily improved the data parsing and queue functions.

Description of the Data

Our main data set was about taxis in NYC during the month of June 2016. We got this data set from Kaggle. We filtered this down to a couple of select days in June. The variables we found useful were pickup and dropoff time, pickup and drop off location (latitude and longitude), fare and distance traveled. There were other variables like number of passengers and how they paid, but we decided those were irrelevant to us in the end. Since this file was so big (11 million plus rows), we had to use Delimit to narrow it down to a workable size. We bounced a couple supplementary data sets including an uber in NYC data set, a yelp data set, a business inspection data set and finally a small weather dataset. We were going to use the uber dataset to compare the relative popularities of taxis and uber, but we realized that these two datasets were much better suited to other comparisons. Next the Yelp dataset was going to be used to see where people were going and what the reviews were for that place. This failed because the data did not have enough relevant observations, there were only 17 reviews in NY and exactly none in NYC. It also wasn’t ideal for the same reason as the inspection data set. The inspection data set was very comprehensive. It had tons of observations and variables, including longitude and latitude which we were going to use when comparing with data from the taxi dataset. However, we realized (fairly late in the project) that we could not really assume where a person was going based on their taxi drop off location. There are many businesses very close together in NYC (within feet of each other), and eventually we concluded that we could not determine which business to use (or even if they passenger went to a business at all as opposed to a private residence), so we had to scrap that data set. We ended up with a smaller weather data set from the National Oceanic and Atmospheric Administration that includes several variables related to temperature and precipitation as well as other less relevant data for each day in June. We are using this now to compare the differences in the taxi data for several types of days and conditions; specifically weekday vs weekend and rain vs sun. We also used a json file of NYC for the maps.

Mapping the Data to Visual Elements

For the visual elements, we have a silhouette of manhattan that we rotated and scaled up so it’s nice and big and that the dots we place on it later are distinguishable. On the map of manhattan, we plotted the location of each pickup point with a small red circle. When you hover your mouse over the circle some extra information (trip distance and fare) becomes available, and a new blue circle appears at the dropoff location with a yellow line linking them. All the other dots also become opaque to really highlight the dot your mouse is on. We also have some supplementary charts that analyze the data at a higher level. These show at each time the average number of taxi rides, the average fare and the average distance travelled in each ride.

The Story

The visualization compares taxi pickups on a sunny and a rainy Saturday and on a sunny and rainy Wednesday, which allows us to gain an intuition behind which factor, weather or day of week, is more important when considering number of rides, average fare and average distance traveled. Judging by the number of dots on the map and the graph at the bottom, we can see that there were more rides early in the morning on Saturday (but probably late Friday night to the passengers). This is not apparently dependent on weather, but solely on the fact that it is a weekend. Comparatively, you can see rush hour traffic at around 7-9am Wednesday mornings, as compared to a much more relaxed saturday morning commute. On Wednesday, there was a sharp dip in taxi usage just before 5 pm, which did not happen on Saturday. For fares, the average again does not seem to depend on the weather. The data shows that prices are higher very early in the morning on both Wednesdays and Saturday, and then during rush hour on Wednesdays. People seem to travel farthest at around 6:00-7:30 am every morning, with no discernable difference between days or weather. We were very surprised by the lack of affect weather had on any of the data we analyzed. The only difference in the data seemed to be at 8pm each day. There was a spike in number of rides but the size of the spike depended on weather and day, with rainy Saturday having around 150 more rides than sunny saturday. Any other differences in the data seemed too small and random to attribute to weather. However, with the lack of variable, it we can attribute more of the changes to time of day. This can be useful for a taxi driver who wants to pick up shift. We can see by the graphs at the bottom, for instance, that a taxi driver working around midnight to 3 in the morning on Saturdays will probably have a relatively higher number of rides, with a relatively higher average fair, and a slightly higher than average distance. This might be good for making a little more money than on a different shift. Rush Hour on Wednesdays also have a lot of rides, with a very high average fair with a decently low ride distant. A taxi driver might have to make a lot of short trips that are more expensive each.