John Hui (jmh547)
Wentao Li (wl553)
Richard Quan (rq32)
INFO 5100

Project 2:
Taxi and Uber Visualization Across Prices, Durations, Geographies, and Uber Demographics

Overview

Our visualization compares Uber and Taxi data for prices across geographies in NYC and durations. The overall effects are that taxis are cheaper overall with greater effects on shorter rides as compared to longer rides. Additionally we include analysis and a visualization of the demographics of drivers in NYC. This is particularly relevant with today's current events as referenced below. It is all to help users visualize their choice in selecting either Uber or Taxi when deciding to travel in New York City.

Taxi Data

The taxi data was obtained from the NYC Taxi and Limousine Commission. The initial file we downloaded contained yellow taxi data for the month of April in NYC. This 1.9 GB csv file had over one million rows, and had columns for ride duration, ride distance, the fare charged to the customer, and longitude/latitude data for both the pickup and dropoff locations. We filtered down the rows to obtain a subset of the data for a three hour long interval during a single week (specifically April 20th to April 27th, 12PM to 3PM), ending up with around 50,000 rows.
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Uber Data

Uber data was very difficult to retrieve. As it is a private company, it doesn't release a lot of data online. Additionally, the only data that is out there does not include dropoff locations, only pickup locations. It would not have been possible to do our specific Uber vs. Taxi ride analysis and comparison with this. Thus, we used a different method and used Uber estimates for rides instead of actual rides. Uber has an API that returns ride price, duration, distance estimates given a pickup and dropoff location. We created a script that would use the pickup and dropoff longitudes/latitudes of our taxi data to get price estimates for corresponding Uber rides Due to data rate limits, it took the script around 30 hours total to query all the data that we had. The script (uberPredict.py) is included in the appendix.

Uber/Taxi Combination JSON

To limit the number of data files being loaded, we concatenated the Uber and Taxi data together into one CSV file, grouped by their matching pickup and dropoff locations. Data fields aside from those location latitudes/longitudes include ride duration, price, and distance per vehicle medium. The CSV file was then converted into a massive JSON file using any of the available online converters, allowing us to retrieve data objects quickly in exchange for one large file load in the beginning.

Demographic Data

The demographic data about uber and taxi drivers was retrieved from the following repo (https://github.com/apalbright/UbervTaxis/tree/master/raw_data). This data was well organized for analysis as compared to the taxi and uber data. This made it much easier to process and visualize. We changed the CSV format to JSON and replaced some variable names for their synonym for neat format. There are 4 JSON in total for this topic. (named: "attrAge.json", “attrEdu.json”, “attrRace.json”, “attrGender.json”). The age file contained the percentage total of the age categories 18-29, 30-39, etc. for both Taxi and Uber. The education file contained the percentage total of the education categories. This format was the same for the gender and ethnicity files.

Leaflet

We used Leaflet.js to load and overlay a map of New York City as our first graph’s background layer while we drew D3 elements on top of it. The NYC map itself was acquired from the Leaflet website. which was retrieved from OpenStreetMap’s datasets. The map, (http://www.openstreetmap.org/#map=5/50.667/-0.088), was originally created by MapBox. Leaflet was only used to load the bottom layer and create floating windows on top of the map. All the other functions, displays, and effects on the map are created using D3.

NYC Community Districts GeoJSON

The GeoJSON file, which holds geographical data for NYC’s community districts (nyccommunitydistricts.geojson), was retrieved from: https://github.com/dwillis/nyc-maps. The file itself holds longitude and latitude data for points along the border of each community district. We used these points to draw the SVG paths around the borders on our NYC Leaflet map.

Top Visualization Mapping

After overlaying a NYC map with Leaflet and sketching out the community districts using the geoJSON, we attach our ride data to the map. Circles are not drawn until a user clicks a region, for aesthetic reasons as well as computational (drawing and moving around 50k points lagged our animations too much). When a user first clicks a region, we lay out any data points that have pickup locations in that region. These appear as red circles. One of the harder parts was to calculate whether a point was within a region. We tried a lot of geometric and calculus approaches, and finally adopted the ray-casting algorithm to check if point is within a polygon (https://github.com/substack/point-in-polygon). When a user clicks the same region for a second time, the circles expand out to the dropoff locations of the rides they correspond to. Additionally, they are colored yellow when the taxi fare for the ride was cheaper than the uber price estimate, or black if vice-versa.

Donut charts on the right explain the proportion of the number of rides that would be cheaper for taking a taxi instead of an Uber in that region. There is also a percentage to show how many of NYC’s rides are in that region. The donut charts are separated by duration of the ride itself, by short rides (<= 10 minutes), long rides (> 10 minutes), and overall. The donuts on the floating window are created using the same technique as in the 2nd visualization which will be explained below.

Bottom Visualization Mapping

The bar charts were directly mapped from the demographic JSON files. It’s clear that each part represent a different attribute, like “age”, “gender”… and by hovering onto the bars, there will be percentage numbers show on the right side of the bars. These numbers are only stand for the percentage in a given attribute of a certain service (Uber/Taxi). These graphs serve as references as the donut charts are calculated.

The donut charts on the demographics page are calculated based on the user’s selection of demographic attributes. The attributes are equally weighted across ethnographic factors, and then recalculated based on individual historical taxi and uber percentages. The calculated percentages show the likelihood that the user would get a driver with the selected attributes.

Story

The main premise of the story compares Uber and Taxi data for prices across geographies in NYC and durations. This comes along with an analysis and a visualization of the demographics of drivers in NYC. It is all to help users visualize their choice in selecting either Uber or Taxi when deciding to travel in New York City.

There are several main conclusions observed from the demographic data of the Uber vs. Taxi drivers. The data is broken up in long and short durations, price, and regions. Contrary to many other major cities, in NYC, taxis are cheaper overall compared to Uber. Additionally in short durations there is a greater percentage of cheaper Taxis than in long durations. This is visualized across individual regions in NYC. The effects are additionally more pronounced in certain regions compared to others. In some regions, the effects of this construct are more evident such as in the Financial District. In areas that seem to have an abundance of of rides such as Midtown, the effects are less apparent. Additionally, each individual region's percentage of total, short, and long rides are shown compared with all of NYC. Through our research, we never found a price comparison between taxi and Uber as a data visualization. We completed this in addition to analyzing these effects across different regions in Manhattan.

Additionally we visualized the demographics of taxi vs. Uber data. Demographics pertaining to Uber drivers has recently been in the news due to the beginning of companies like "Chariot for Women" that only feature female drivers and female passengers. With the perception of safety and comfort in mind, it seems that passengers do not solely view Taxi or Uber drivers as conduits to facilitate going from place to place, but additionally with biased assumptions based on preconceived notions based on demographics.

Comparing Taxi and Uber drivers we see more female, higher education levels, and more drivers of white ethnicity and less of black ethnicity comparing Uber to Taxi. These effects are augmented through our data visualization. When combined, the visualization the effects of these trends are additionally amplified as demonstrated in the visualization. The percentages are calculated by multiplying the values for each of the attributes selected.