CS 3300 Project 1

Trevor McDonald (tdm67), Charles Byrnes (cbb74), Matt Hsu (mdh267)

2A.

The data we used we first found on Kaggle and then used from 538's Github Page. The entire dataset contains the times and coordinates of all for-hire vehicle and ride-sharing vehicle in New York City during the second half of 2014 and the first half of 2015. The data was in .csv files, where each row contained a timestamp of the pick-up, latitude, longitude, and the taxi zone number where the pick-up occurred.

Because this data was too large to use all at once, we decided to focus only on the Uber and Lyft sections of the data, the top two ride-sharing applications. We also chose to use only two months worth of data from August to September 2014, as this was just after Lyft launched in Brooklyn.

In addition to the .csv files, we had a few other files to help display the data. We used a shape file to show New York City taxi zones on the screen that we obtained from the New York City Taxi and Limousine Commission website. We converted this to a GeoJSON file. From the same website, we used another .csv to go from taxi zone numbers to names and locations of the zones for mapping.

After paring down the amount of data used, we reformatted it to better suit the graphs we wanted to make. We did this by "binning" our data appropriately for each graph we wanted to use. For our histogram, we binned our data into a new .csv file containing a Date object and the count for the respective company of the number of rides that day. For the matrix plot, we binned the data by hour of each day and each day in the range of interest, and associated each hour with the number of pick-ups that happened during that hour for each company. Finally, for the map we binned the data into number of pick-ups by taxi zone. In all of these cases, we used d3.js to generate these new .csv files, reducing our dataset down to these easier to work with files.

2B.

Histogram:

The histogram displays the total rides for each day, in separate bars for Uber and Lyft.

For the histogram, the x axis used the scaleBand() function to calculate where each bar of the chart would be placed. The dates were display and binned using the scaleTime() function between August 1st and the end of September 30, 2014, as well as the histogram() function for the binning. The y axis was a linear scale mapping the range of the data to the svg object to display the height of each bar in the histogram.

Matrix:

Each grid space of the matrix displays the difference in z-scores between Lyft and Uber rides for a specific hour. We normalized the values within each dataset and display their

difference as a color. This value is mapped to a pink - grey - black color space. A darker pink indicates more Lyft rides than usual, and a darker grey indicates more Uber rides than usual.

For the matrix, the x values for the positions of each rectangle were calculated using a linear scale, as were the y values. To actually display the axes, the y axis used a linear scale mapping numeric hours to the pixel space. The x axis used a time scale mapping days to the pixel space in the same way the histogram did. The actual color values of each grid space were calculated with a linear scale mapping the values to our pink - grey - black space.

Map:

The map uses the same intuition as the matrix as the display value, showing whether Lyft or Uber had a higher proportion than normal in terms of rides per zone. This value is obtained from the z-scores of each dataset, and mapped to the same color space.

The map is displayed with the geoAlbers() projection. The fill for each taxi zone is calculated with a similar linear color scale used in the matrix.

2C.

On July 8th, 2014 Lyft announced that it would launch its ridesharing service in Brooklyn and Queens after months of "unprecedented support" from city residents. In the two months that followed, we plotted the data to see what happened to Lyft and Uber as they competed in the most populous city in the US.

First in the map, we see that Lyft had a higher proportion of rides compared to the average in Brooklyn, while Uber had higher proportions than normal near airports and in Manhattan. This made sense because Lyft launched in Brooklyn, and also shows that travellers are most likely the ones most biased towards using Uber over Lyft.

Next, in the histogram we see just how many more rides Uber serves than Lyft. Especially as September starts, Uber gains more rides per day on average, while Lyft somewhat dwindles. The histogram also displays the cycle of the week, with peaks for both Uber and Lyft happening on weekends. What is surprising is that Lyft seems to peak about a day or two after Uber does, in general.

Finally, the histogram breaks down each day, and shows whether Lyft or Uber had a greater share of rides than normal by the hour. We see that Lyft was doing better in August than it did in September, presumably because of its big launch in the previous month. We see this recede over time, as Uber starts gaining again in September. The surprising thing about this graph is that it reveals Uber is used more than normal during the evening, and then eventually Lyft is used more during the very late night and early morning,

These three graphs help show how Lyft fared after its launch in New York City, and even just two months after this launch we see Uber reasserting its market dominance.