Project Report

Tae Hyon Lee

October 26, 2018

1 Paper

Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips by Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Claudio T. Silva

2 Scope

This project will implement a scatter plot map of the origination and destination of taxi trips on January 1st, 2016 and a line graph of number of trips throughout the day in New York City. This project will not implement zone selection feature and heat map of average rides per hour as shown in the original paper.

3 Implementation

Steps:
1. Identified visualization to implement from the paper and found data from the author's GitHub page: "https://github.com/ViDA-NYU/TaxiVis".
2. Created a d3 file and used Leaftlet library to create map of New York City.
3. Transformed the data using Excel so that pickup columns and drop-off columns from trip data would be separated from a single row and merged into common columns with datetime, latitude, longitude and category of either "pickup" or "dropoff".
4. Plotted the data on the map with pickup data colored in blue and drop-off data colored in orange.
5. Tested out the code by running SimpleHTTPServer in the file directory and opening up the file by connecting to localhost using a web browser.
6. Added buttons that lets the user select the time range in a day (6AM-12PM, 12PM-6PM, 6PM-12AM, 12AM-6AM).
7. Found out that the data from the author's GitHub page is missing a lot of data. It only had data for two days and mostly from 12AM to 4AM.
8. Found a new set of data from Kaggle.com (https://www.kaggle.com/c/nyc-taxi-trip-duration/data) that has a more recent data ranging from January to

June in 2016.
9. Transformed the data using Python (Python script provided) in a similar manner as described in step 3. I used Python in this case because the data was too large to handle on Excel.
10. Tested out the new data on the code to make sure it works.
11. Google Chrome crashes with a timeout because the data is too large so I filtered the data so it only loads January 1st data.
12. Created a second visualization that compares the number of taxi trips with two different time range. In the paper, it compared 2011 and 2012 but for this data set, I only had 2016 data so I chose to compare between Jan-March and April - June in 2016.
13. Tested the code once again to make sure both the visualizations work as intended. 14. Organized the code in a more readable manner.

4 Result

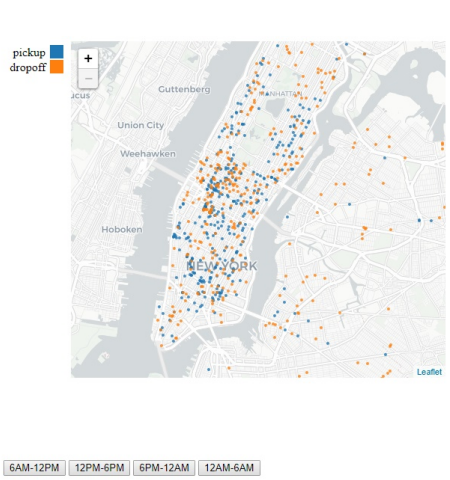

Figure 1: 12AM to 6AM

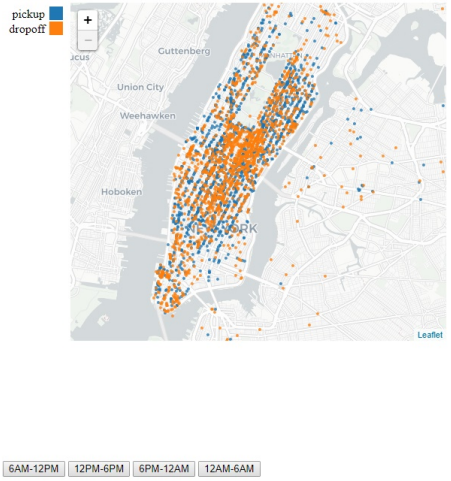

Figure 2: 6AM to 12PM

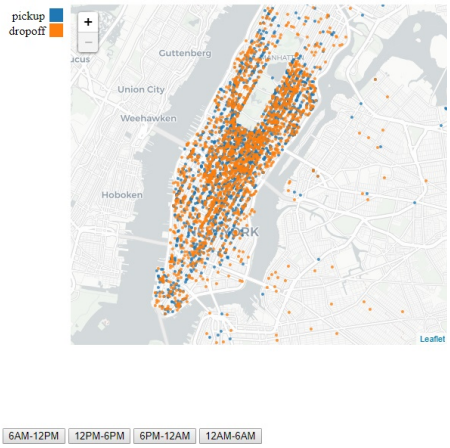

Figure 3: 12PM to 6PM

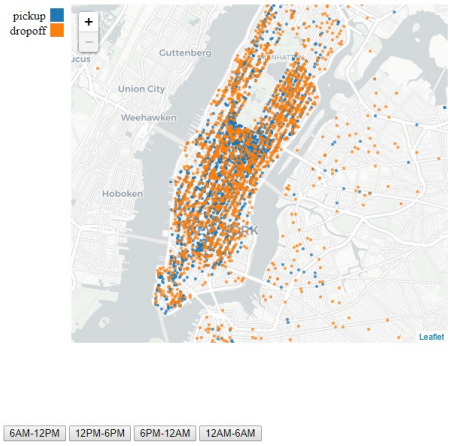

Figure 4: 6PM to 12AM

Figure 5: Maps of New York City with origination and destination of taxi trips plotted on the map. Each Figures above have different time range (12AM to 6AM, 6AM to 12PM, 12PM to 6PM and 6PM to 12AM). With these visualizations, we can get a general sense of where people go during each time ranges. For example, in figure two (6AM to 12PM), we can see people being picked up in the outer regions of the peninsula and getting dropped off in central area of the peninsula, hinting that industrial or business buildings are in the central area of the peninsula.

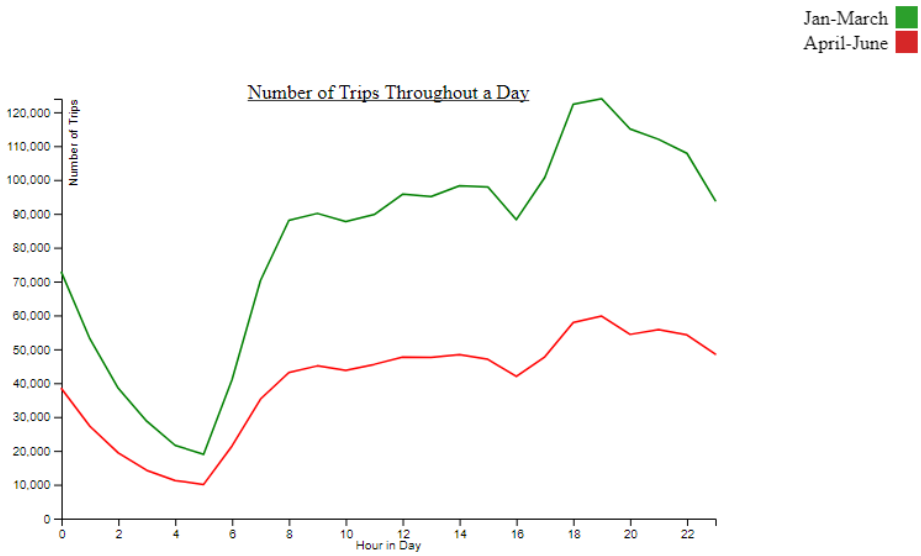

Figure 6: The number of trips throughout a day for Jan-March vs April-June in New York City. We can see that there is a significantly higher number of taxi trips for Jan-March range, suggesting that people went out more during that time range in 2016.

5 Discussion

This implementation will not work well if you want to take a closer look at smaller range of time for the map because the range of time is fixed. Also if the user wants to examine the visualization with more data, it will not work because the browser will crash. Because of the huge size of the data, there is already a significant delay when switching among different time ranges for a single day.

6 Future Work

For future works, I would recommend implementing an interactive feature where the user can select a zone and get more detailed information for that zone.