DS702 Project Instructions

Release Date: 07 March 2022

- Cite your sources if you are taking help (papers, websites, students etc.).
- Plagiarism is strictly prohibited. Negative mark will be assigned for plagiarism.
- Remember to comment your code. And your answers should be detailed.
- You are allowed to use other datasets and come up with your own tasks.
- The project report should not exceed 20 pages.

New York City taxi and Limousine trips

We will use data collected by the New York City Taxi and Limousine commission about "Green" Taxis. Green Taxis (as opposed to yellow ones) are taxis that are not allowed to pick up passengers inside of the densely populated areas of Manhattan. We will use the data from September 2015. Here is the page for NYC Taxi and Limousine trip record data:
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

You will use the data for Green Taxis in September 2015 [Fig. 1].

The information for each column can be found from this page:
https://www1.nyc.gov/assets/tlc/downloads/pdf/data-dictionary-trip-records-green.pdf

Use Spark to finish following tasks. The answers can be quite flexible and creative. For open questions, remember to write your ideas and how you come up with them into your report, not just listing your conclusions.

You are also free to create any extra feature you need based on the data file.

Task 1.

Report how many rows and columns the data file have?

Task 2.

Collect all the trip distances for all trips. Plot a histogram of the trip distances. You can play with the number of bins or use the default.

Task 3.

Report mean trip distance grouped by pick-up hour of day. (you can do more than we ask – be creative)

Task 4.

We'd like to get a rough sense of identifying trips that terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.

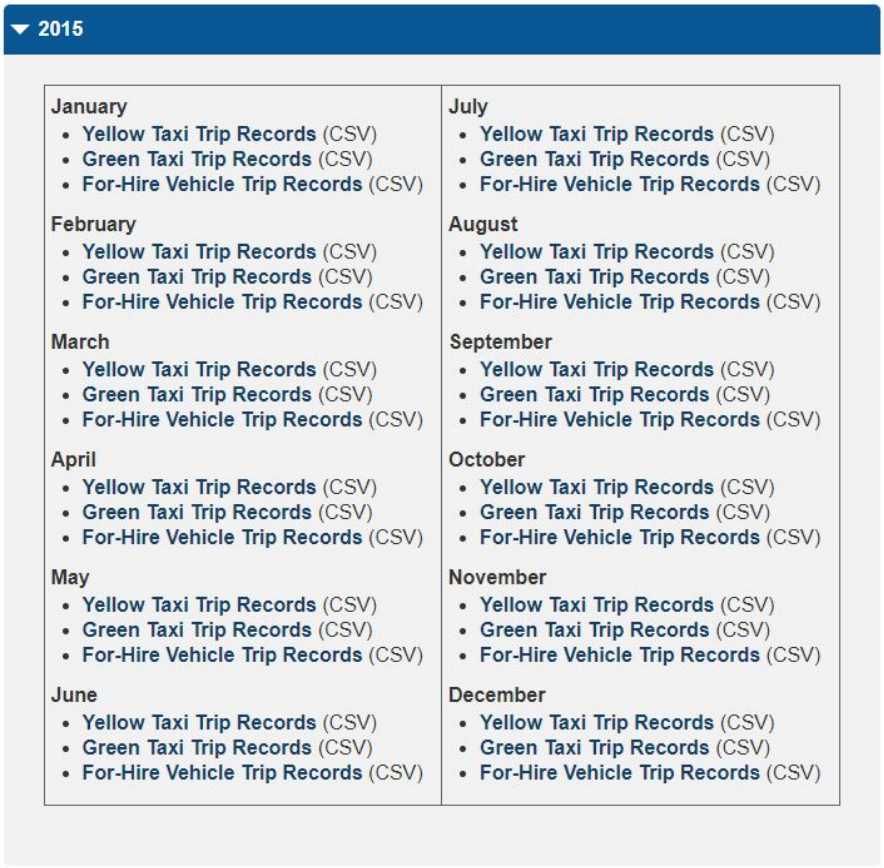

Figure 1: 2015. September. Green Taxi Trip Records

Task 5.

Do two clusterings on pick-up and drop-off locations respectively. Choose a proper number of clusters k and report the centroids. Is there much difference between centroids for pick-up and drop-off? Note: do not set the maximal k too large in your code. k from 2 to 10 - 20 should be enough to try.

Task 6.

Cluster pick-up and drop-off locations. Choose a proper number of clusters k and report the centroids (you can use google maps to show centroids). What is the percentage of trips for each cluster? What is percentage of trips where pick-up and drop-off are in the same cluster? What have you learnt from this?

Task 7.

Figure 2 is a visualization for a small number of trips. It gives us a rough sense where the trips happen.

Since the Hudson River, which is the boundary of New Jersey and NYC in this area, is quite straight, we can use a line to model this natural boundary. The approximated line in latitude (y) and longitude (x) can be represented as:

```
y = 1.323942 * x + 138.669195
```

If a location satisfies $y > 1.323942x + 138.669195$, it's in New Jersey. If a location satisfies $y < 1.323942x + 138.669195$, it's in Manhattan. After such processing, we can get Figure 2. In this way, we can better utilize the pick-up and drop-off latitude-longitude data.

In this task, you are expected to group all trips into the following four categories, NJ → NJ, NJ → NYC, NYC → NJ, NYC → NYC. Can you build some association rules on intra- vs. inter-borough traffic? What story does it tell about how New Yorkers use their green taxis? For example, in which hour of the day, there would be more inter-borough traffic than intra-borough traffic? For those NYC → NYC trips, people are more likely to take taxis from uptown to downtown or from downtown to uptown? There are many interesting stories in the data, feel free to discover some. Again, you can create new features as needed.

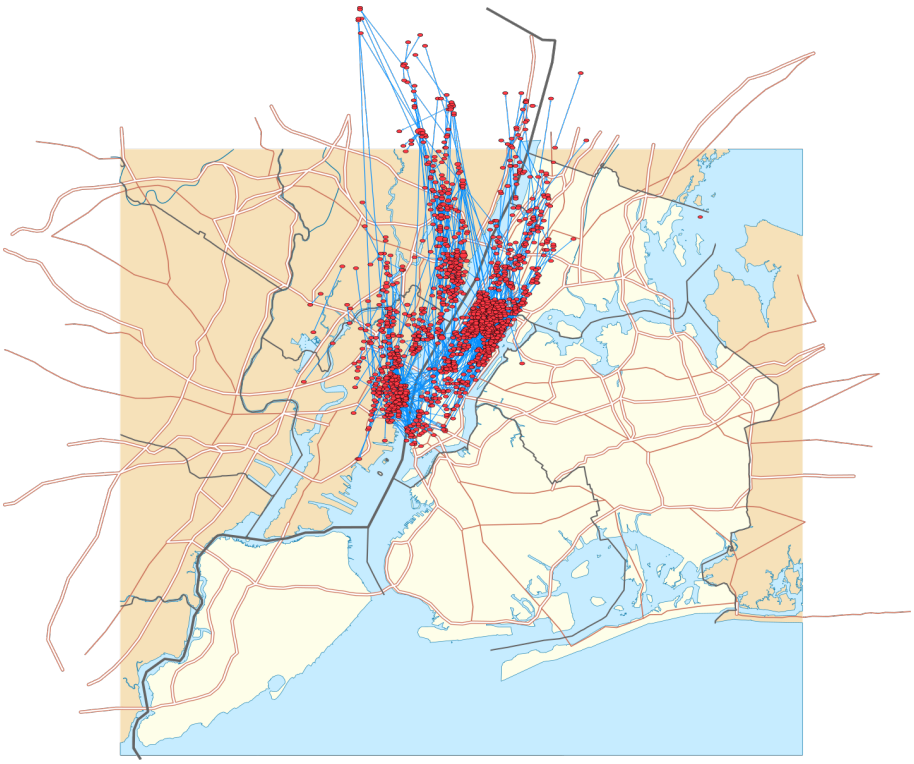

Figure 2: 500 trips

Task 8.

Could you suggest to Green Taxi how to improve their business? Think about ways they could decrease cost, get more customers, improve quality of service etc.

In all tasks, try to use the algorithms that you learnt in class, if you didn't succeed, you can come up with other tasks yourself using this data set.

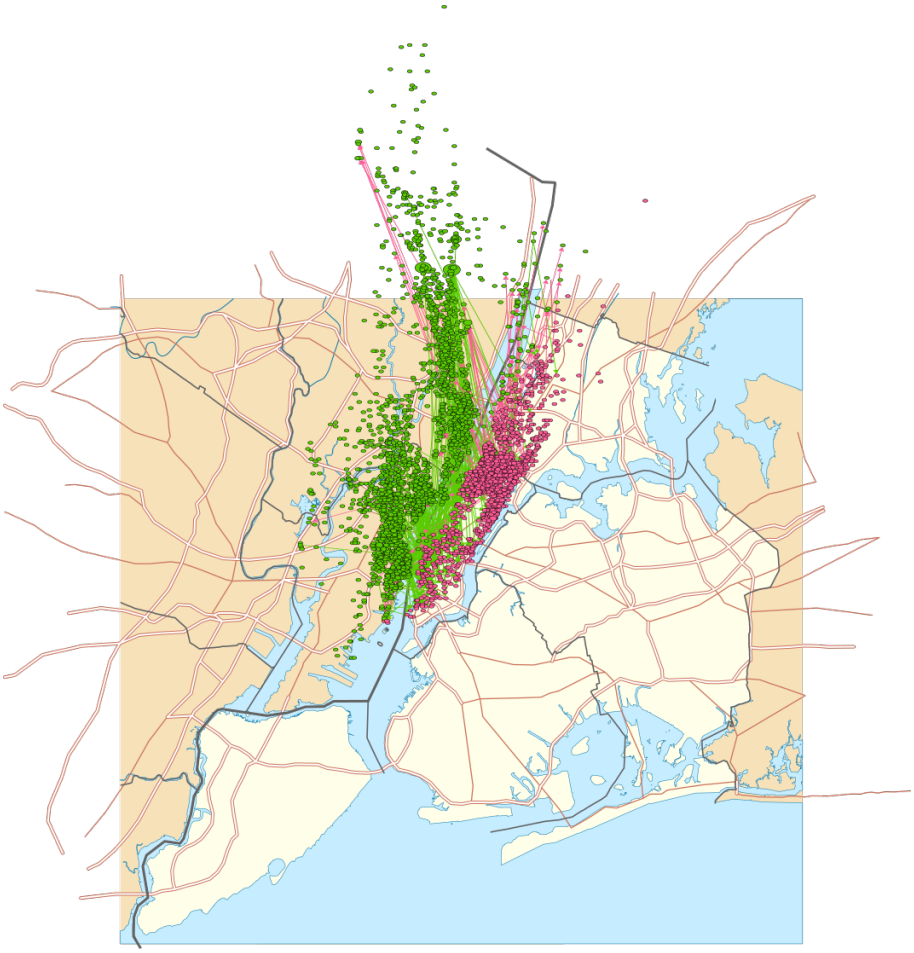

Figure 3: Red points - Manhattan, Green points - New Jersey