Transit Data Analysis: Part A

Students: Danielle Stacy, Jing Guo, Siyue Yang
Mentors: Candace Brakewood, Cheng Liu, and Kwai Wong

Introduction

Transit is an app used to collect and map real-time public transit data. People may use the app to determine which train or bus route to take, to plan a trip, or to search for the quickest form of transportation among other things. The data collected from the app includes the users’ locations, saved placemarks and routes, usage of the Uber features, usage of car sharing and bike sharing features, trip planning feature, and so on.

Process

Convert given json files to csv files and add unique codes to identify users
Isolate necessary data from tables of interest
Construct algorithms to perform analysis and add as needed to achieve optimal results

Research Questions

- Can Transit users’ home and work locations be inferred from the data collected from the users using the app?
- What is the difference between the Uber users from Transit app and the Uber users from Uber app? What is the connection between their travel time?
- What caused the users who want to use public transportation first to switch to request Uber through the Transit app? What are the characteristics of the user’s location and the distance from the subway or bus station? Does the weather factor affect the user’s choice?
- What is the relationship between the data characteristics of uber request and taxi travel?

Analysis of Uber Requests

Users who request Uber through the Transit app are signaling their intent to try to use public transportation first but are willing to move on to other modes when transit does not meet their needs.

We use the Uber request data and combined the public data on the network: the yellow taxi and the green taxi data for New York City and the time data for the Uber raw data. Here are aspects of this research:

1. For the Uber request data: Analyze the location characteristics and time characteristics of the users, research the relationship between location and subway distribution, and research the relationship between travel time and weather and traffic flow status.
2. Compare the time portion of the Uber raw data to research the time characteristics of different users choosing different modes of travel.
3. Combine the taxi data of New York City to research the relationship between taxi travel and Uber travel in Transit.

Home/Work Inferences of Users

Assumption: Users’ location data will have two well-distinguishable clusters, indicating their home and work places. The user will be at their home and work locations at distinct times of the day, times not overlapping.

Process: Specific users who have saved their home and/or work locations have been extracted. The data from these users has been run through a k-means algorithm specified to have two clusters.

Issues: In the beginning, the assumption was that users would open the app in the morning before work when they were at home and in the evening after work when they were getting ready to leave work. This assumption did not match the data. Then it was assumed the majority of data at the home would occur in the night and the majority of data at the workplace would occur in the daytime.

The data in this section is from November 1, 2016 to October 31, 2017.

This project concentrates on New York City.

The yellow and green taxi in New York City

Time trends of taking a taxi(one month)

Future work:
- Label centroids as “home” and “work”
- Expand dataset to include location data from the rest of the year
- Apply algorithm to users without saved locations
- Use Google Maps API to investigate more saved locations

Acknowledgements

This project was made possible by the funding of the National Science Foundation, facilities provided by the University of Tennessee and Joint Institute for Computational Sciences, guidance from our mentors, and data provided by Transit.