
Team Old Iron
Ray, Chelsea, Jayden, Yang

Background

Uber and other taxi generates large quantities of pick-up location data on a daily basis. For companies like Uber, new projects like Uber Elevate require a static of optimal pickup locations. For this project, we attempt to **decide on optimal skyport locations** in NYC for Uber Elevate, extracting and utilizing available data from Uber to build a model to determine the ideal location. Skyports are future trend in metropolitan areas, and such research highlights the potential of the transportation technology.

Plan

Scoring model for ideal skyport location (Multiple Facility location problem)

- Customer demand data in NYC
  - Customer traveling data:
    - NYC Taxi and Limousine Commission (TLC) Trip record data
      - Data format: Monthly taxi traveling .csv data in NYC
      - Data size: Around 0.5GB for each month
  - Traffic flow data:
    - NYC Real-Time Traffic Speed Data
      - Data format: Daily traffic flow .json API in NYC

Customer traveling data

	id	vendor_id	pickup_datetime	dropoff_datetime	\
0	id2875421	2	2016-03-14 17:24:55	2016-03-14 17:32:30	
1	id2377394	1	2016-06-12 00:43:35	2016-06-12 00:54:38	
2	id3858529	2	2016-01-19 11:35:24	2016-01-19 12:10:48	
3	id3504673	2	2016-04-06 19:32:31	2016-04-06 19:39:40	
4	id2181028	2	2016-03-26 13:30:55	2016-03-26 13:38:10	

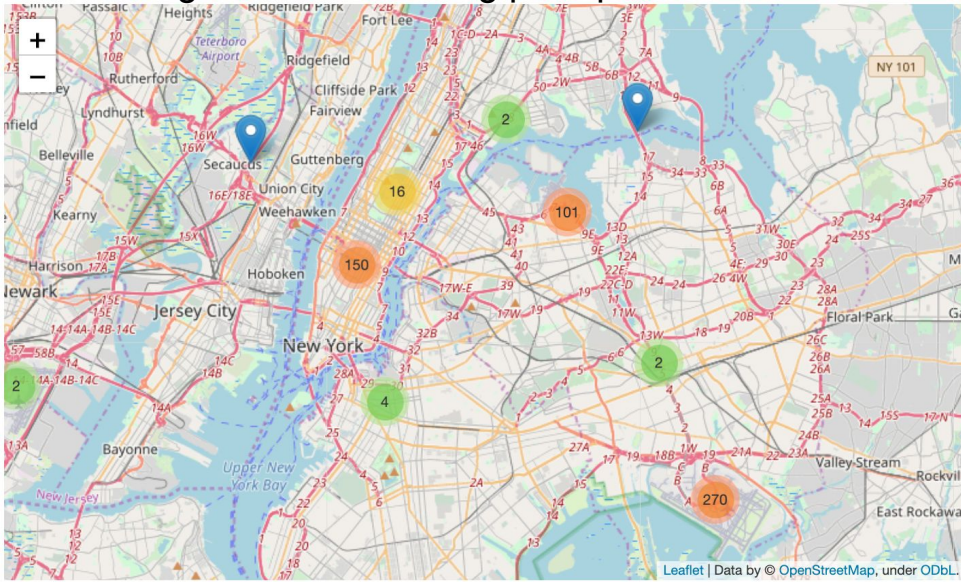
	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	\
0	1	-73.982155	40.767937	-73.964630	
1	1	-73.980415	40.738564	-73.999481	
2	1	-73.979027	40.763939	-74.005333	
3	1	-74.010040	40.719971	-74.012268	
4	1	-73.973053	40.793209	-73.972923	

	dropoff_latitude	store_and_fwd_flag	trip_duration
0	40.765602	N	455
1	40.731152	N	663
2	40.710087	N	2124
3	40.706718	N	429
4	40.782520	N	435

Data engineering

- Feature extraction:
  - Build the traffic flow data into graph
  - Filtering out data TLC trip record data that has long traveling distance
- Model building algorithms:
  - K-Means
  - Farthest-first Traversal
  - Integer programming

Long distance traveling pickup count in NYC



Validation Plan

- Metric:
  - Sum of all weighted costs between demand points and solution skyports
- Quantitative testing:
  - Random sample half of the month into validation data
  - Baseline comparison on metric
- Visualization:
  - Line chart for comparison on average for different demand points
  - Trade-off cost between setting more skyplots and sum of all weighted costs
  - Plot for destination and demand points using Folium

Plan B for Project

Topic: Getting Important Information from Social Network

In our team, we also discuss on a second possibility for our final project. Briefly, with the rapid growth of online social networks and IoT networks, mining valuable knowledge from the graph data become important. For example, people have been using graph mining strategies to enhance the performance of personal recommendation, price prediction, communication anomaly detection. However, it is challenging to extract network features from graph data conducted from company such as Facebook and Twitter. The crazy amount of raw data generated everyday makes it difficult to extract the real useful data from it. Therefore our second though for our final project will be focusing on **how to extract the real meaningful information from the real world noisy data** and make precisely predictions out of it.

Relevant Milestones

Milestone Name	Team members responsible	March 3	April 4	May 10
Data clearing and feature extraction	Yang, Chelsea			
Feature visualization	Ray, Jayden			
Finalize topic	All			
Hypothesis experiment	All			
Model building	All			
Fine tune parameters	Chelsea			

References

- NYC real time traffic speed data: https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq
- TLC trip record data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Facility location analysis: https://en.wikipedia.org/wiki/Facility_location_problem#Algorithms
- Uber elevate: https://www.uber.com/us/en/elevate/