Datathon Report
Team member: Yujie Wei, Hengruo Zhang

Question:
Given the number of autonomous-driving vehicles, how many cars should be allocated to each region of the city at different time during the day in order to achieve the best efficiency while preserving the fairness of getting a ride?

Background:
The emergence of the autonomous-driving vehicles will definitely transform the industry. However, shifting to complete autonomous-driving transportation will be a long process during which taxis, FHVs, and autonomous-driving vehicles co-exist on roads. Our team believe that one possible advantage that the autonomous-driving vehicles could bring to us during the "hybrid" time is that they can act as a supplement of the conventional transportation. Therefore, it's critical for the autonomous-driving vehicle company and the government to understand the demand of cars for different region at different times to allocate the resources wisely.

Dataset selected:
We chose ***uber_trips_2014*** and ***yellow_trips_Q2/Q3*** for two main reasons: First, these two datasets were consistent to some extend. They both had similar entries and reflected the demand of the same time period (April 2014 - September 2014) which could reduce the possible interference due to variance of different periods; Second, the graph showed that they have different geographical and time distributions which we believed could reflect the difference of two typical service patterns - FHVs and taxis. Considering both the FHV and taxi datasets allowed us to achieve some balance between two service patterns in our suggestions on how to allocate the resources.

Approach:
- Basic analysis

We first randomly sampled 1% of the data and analyzed the time and geographical pattern of the car demand for uber cars and FHVs. As we can see from the graph, the FHVs and taxis had different demand distributions over time. By Figure 1, the taxi demand was more stable during the day while the uber demand was more centralized in the evening. From the geographical distribution we can see that they had different patterns as well. (Figure 2 and 3)

- Learning

To get the number of cars we should allocate to each region, we need to perform a regression over the time and region. In other words, given the time(hour) and the region(nta_code), our system should be able to show how many cars a specific region needs. We used a uber car distribution matrix with size 856440x219 as our training data set where each row represents a certain date/time and a region using the binary indicator with a demand number. The rows were corresponded with 24 hours and the columns were districts in the city.

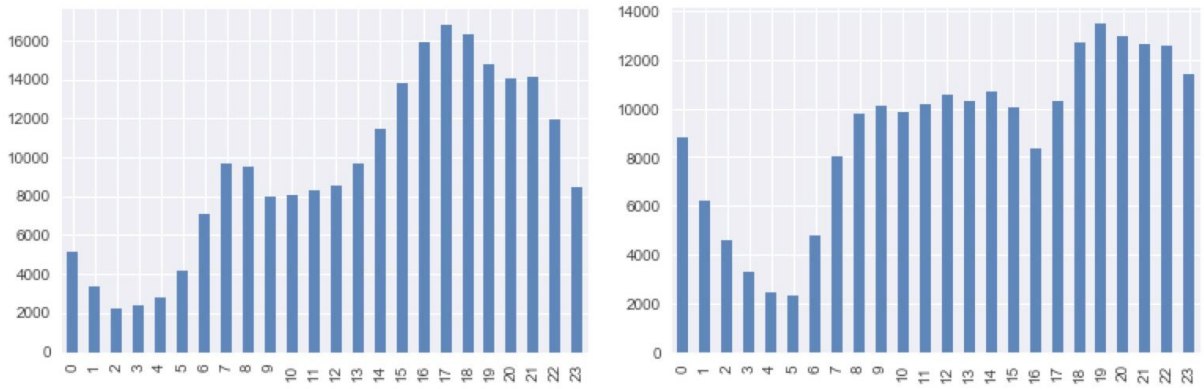


Figure 1. Demand distribution over time (Left: Uber 2014, Right: Yellow Trip 2014Q2)

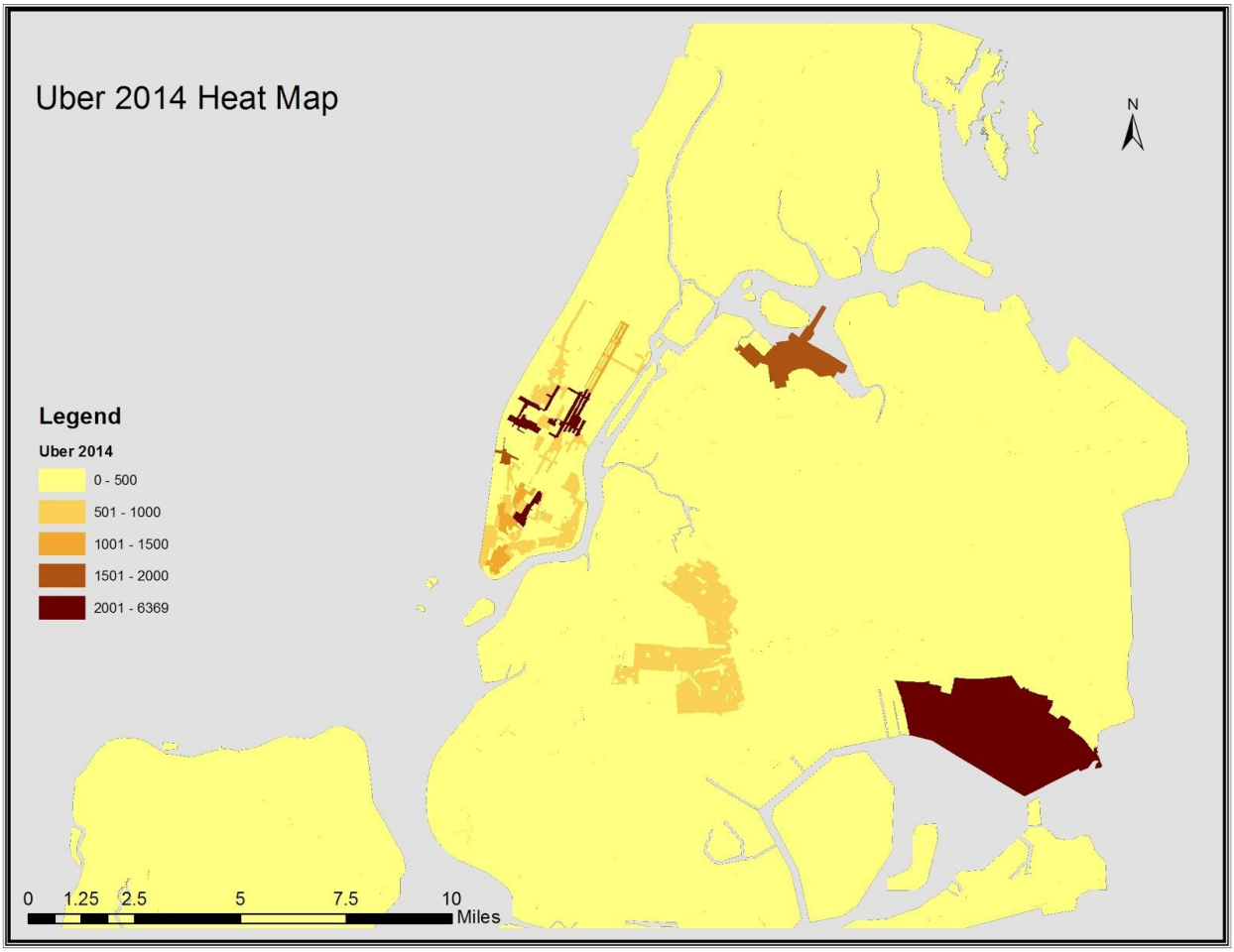


Figure 2. Region distribution over different regions (Uber)

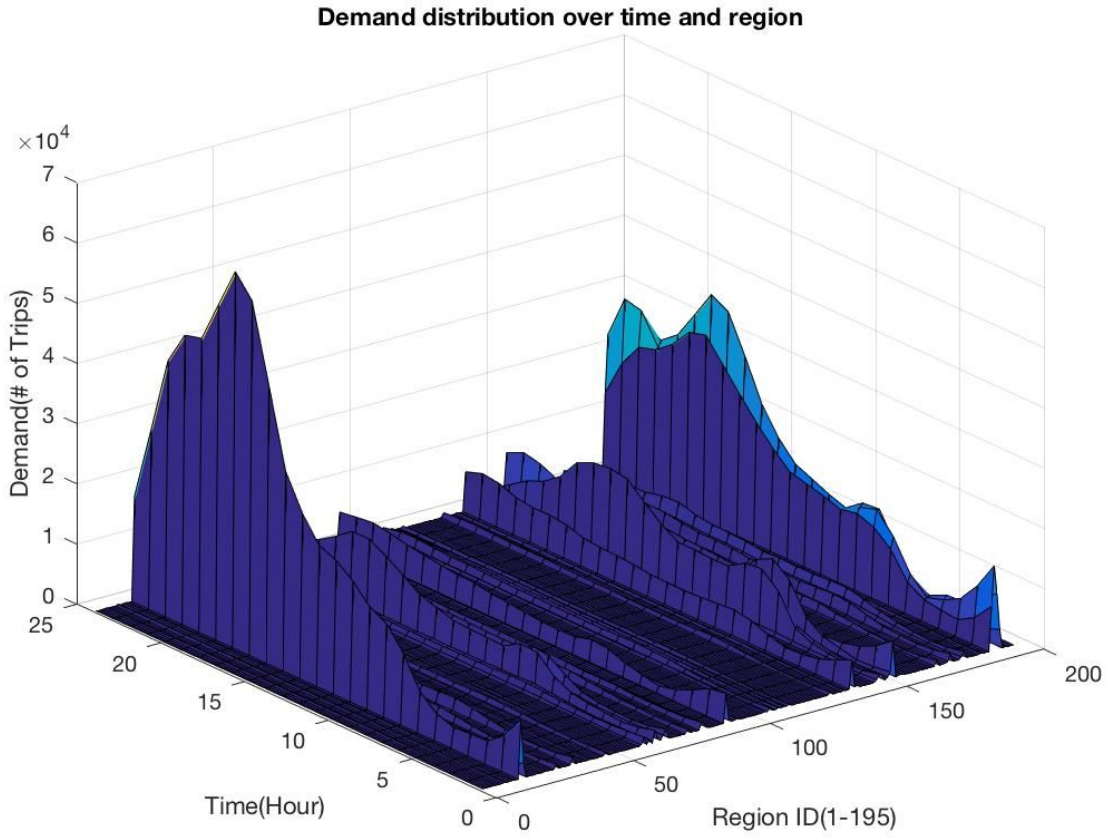
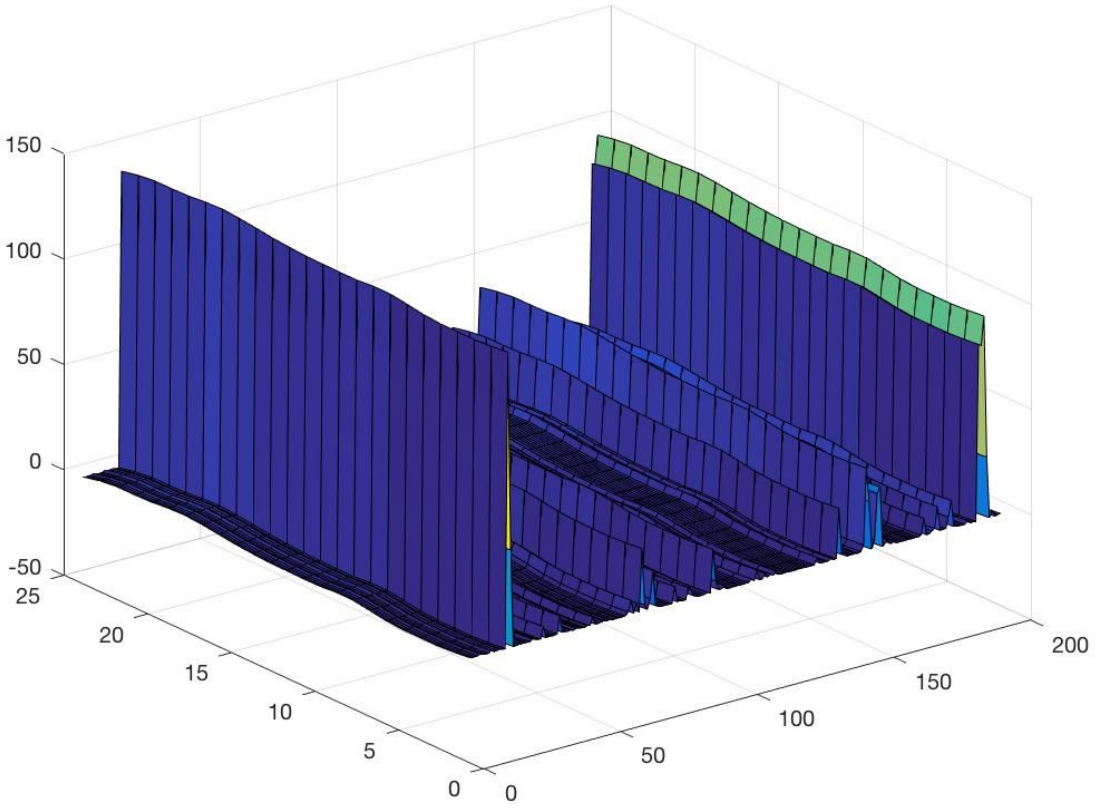


Figure 3. Demand distribution over time and region

Result:
After linear regression learning, our model could fit well uber car demand distribution over region. However, due to the limit of time, we didn't finish the whole learning process and therefore the model couldn't reflect the changes over time.


150
100
50
0
-50
25
20
15
10
5
0
0
50
100
150
200