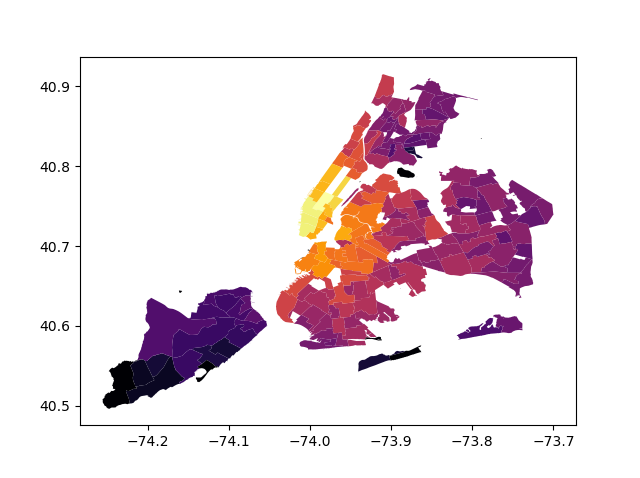
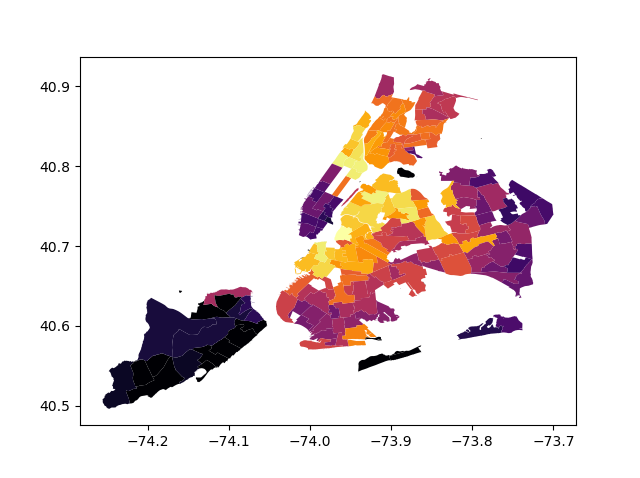
Question Statement

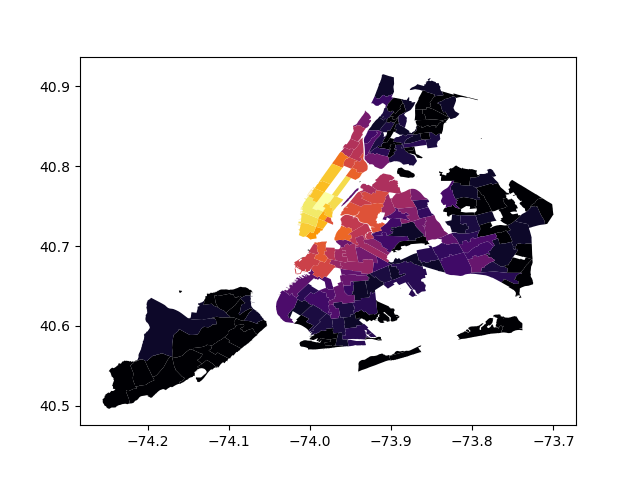
We want to examine the service patterns of the 3 service providers. The goal is to check if they have similar service patterns. The question is important because that’s the key of success of Uber -- People have taxi service in NYC for centuries, it’s a very mature industry -- which means it hard for newcomers to enter into the market. However, Uber successfully entered into the market and have a significant impact on the market. How can they do that? What is their key of success? Could they increase their market share in the future or they will be only able to increase their share to the a certain amount and hit the limit? We will try to answer the questions by starting looking at the services the three companies provided and identify the reasons of Uber’s success.

Geographical Location

We categorized the data points from 3 datasets (Uber 2014, green and yellow) and plotted them by regions into a map. We use color to represent the amount of drives in that region. Yellower the area is, means there is more pick-ups happened in that area. The data is randomly sampled from whole April for green and yellow taxi in April, the reason why we didn’t use the whole dataset is because we think since the amount of the dataset is huge, we could be sure that by randomly sampled them we can get a representative dataset of the two companies. The x-axis and y-axis is are latitude and longitude.

Map1: Uber

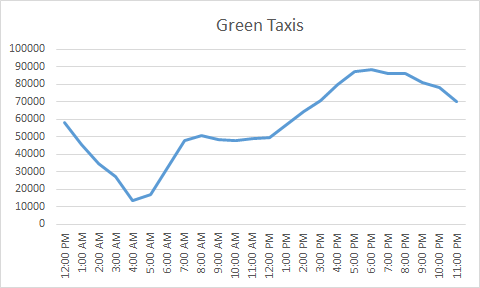
Map2: Green

Map 3: Yellow

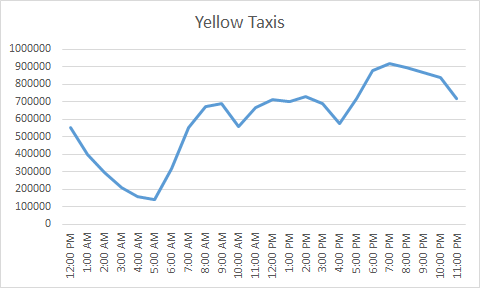
By looking at the geographically distributions of the three services, we assured one thing: yes, there is a balance between the existing service providers. The yellow company and green company operate their services in different areas -- most of pickups of yellow taxi happened at the left top area, and the pickups of green is mostly on the top area.

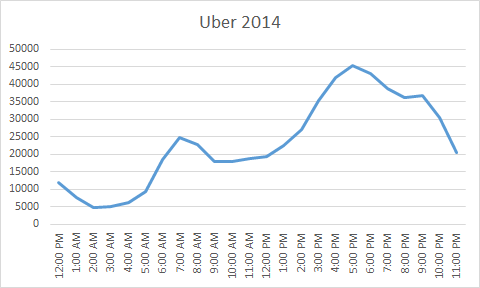
If we consider that fact of that, we can see, we can immediately realized that Uber is a balance-breaker --- Uber provided their service across the whole city! Although the hottest area of Uber’s heat map is still on the top area, but in other areas, the color of each region is always yellower than the other two companies -- which means, Uber broke the geographical separation of the companies that lasting for a long time! Additionally, Uber provides their services on the area where existing companies tend to not provide the service! That means, in this mature market, Uber found the “blue sea” -- new potential buyers and sellers(drivers).

People might be curious that why the existing companies can’t identify the share and provide the service? The possible reason is that, the area could be very dangerous, so drivers don’t want to drive that area. But the Uber drivers might live in the neighborhood, so they are more confident to avoid the risk in the area, provide the service and earn the money.



We can see the green taxis are mostly popular around 6:00pm.



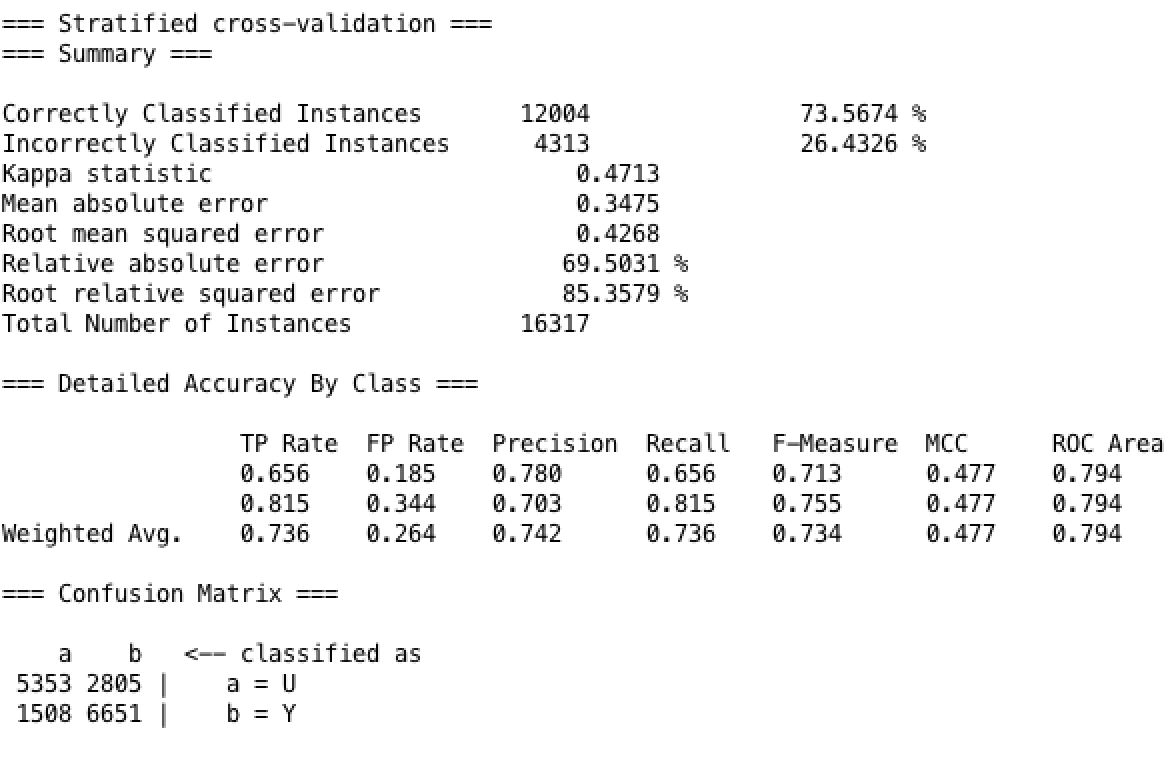


Compared with Green and Yellow Cabs, Uber are also most popular in around 6:00pm, but the user numbers are much less than the other two in arou

Classification on Uber Or Yellow Cab

Based on the statistical analysis above, we can see there is a big difference between the area that Yellow cab and Uber were used. And it is also a good guess that the time period is related to user choices. So a classification problem is built up: based on feature of Hour period and Location\_ID, can we predict that if user prefer Uber or a Yellow Cab? This problem will be helpful for company to assign the number of cars in certain area.

Based on 16317 instances, in which 8158 are labeled as Yellow, 8159 are labeled as Uber, I built up a random forest classifier, and 10-folds cross-validation, the result is as follows:



Basically the classifier shows a good performance on accuracy and confusion matrix.