FullStory Taxicab Challenge

Kyle Zimmerman

Summary

Based on my analysis, I would conclude that the best place to drive in order to maximize income is near the Newark Airport. The best time to drive would be in the early morning, especially around 4 to 6. If I were to choose to drive a taxi 10 hours a week and attempt to maximize income, I would try to primarily pick up fares from the airport, and I would drive between 4 and 6 in the morning for 5 days a week, although I might prefer evening hours to improve my quality of life without much loss of income. The choice of location seems to be considerably more important than the choice of time for maximizing income, with the most valuable pickup locations being worth over 10 times as much money per hour as others, while the most valuable time to drive was only worth ~50% more than the least valuable time.

The most significant factor missing from this analysis is some indication of how often a taxi would actually have a fare in the cab, which would allow me to more accurately estimate hourly income. One possible way to examine this factor would be an additional column in the dataset indicating the length of time between when the driver dropped of the previous fare and when they picked up the current one. The least useful column in the dataset is the “store and forward” flag; although the vendor ID, payment type, passenger count, and cost of taxes and tolls did not seem particularly relevant for my analysis either.

Method for Analysis and Results

In order to scope my analysis properly, I made the following assumptions. First, the driver’s income is calculated as fare plus tip, and any costs for leasing/fuel/maintenance were ignored. Second, the driver can choose where and when they want to drive, and will be significantly more likely to get fares in the zone in which they choose to drive. Third, I only focused on income while a passenger is in the cab and ignored time lost without a fare, due to the lack of data that would allow me to adjust for this loss (I could have counted the number of trips in each zone to help address this problem, but this method would still fail to account for the number of other cabs in the area). To improve data quality, all trips with a trip distance of 0 were removed, as well as all trips that were recorded as lasting fewer than 30 seconds (fewer than 1% of all trips). This adjustment was made to account for some trips that appear to have been logged incorrectly as lasting either no or very little time. It is also worth noting that this analysis is only based on data from June, and may not apply at other times of the year.

For the analysis itself, I chose to find which trips yielded the highest income per hour, with “income per hour” defined as fare plus tips divided by time between pickup and dropoff. I averaged the income per hour, weighted by how long the trip lasted, across all trips by pickup location, and separately by hour of the day and day of the week. For locations, I also ignored

any zones with fewer than 30 recorded fares (1 per day) in my analysis. I examined both fare per hour as well as tips per hour individually from each other to help improve confidence in the results. There was not much change to the top locations or times when separating out the sources of income, but this did provide a nice sanity check for my results.

From this analysis, the most valuable pickup location was the Newark Airport, with passengers picked up there generating fares worth around $400/hr (note that this rate only applies when a passenger is in the cab, the actual hourly rate for a driver will be far lower). Interestingly, the second most valuable zone code was one of the codes indicating the zone was unknown. It is possible this corresponds to rides from beyond the city, although this cannot be verified in the data. After that, the next most valuable zones were Richmond Hill and Kew Gardens, each of which generated fares worth a little over $110/hr. Figure 1 shows the zones of New York City, colored by their value per hour while a passenger is in the cab (grey indicates insufficient data to draw a meaningful conclusion). The map of the outlines of the zones was obtained from https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc.

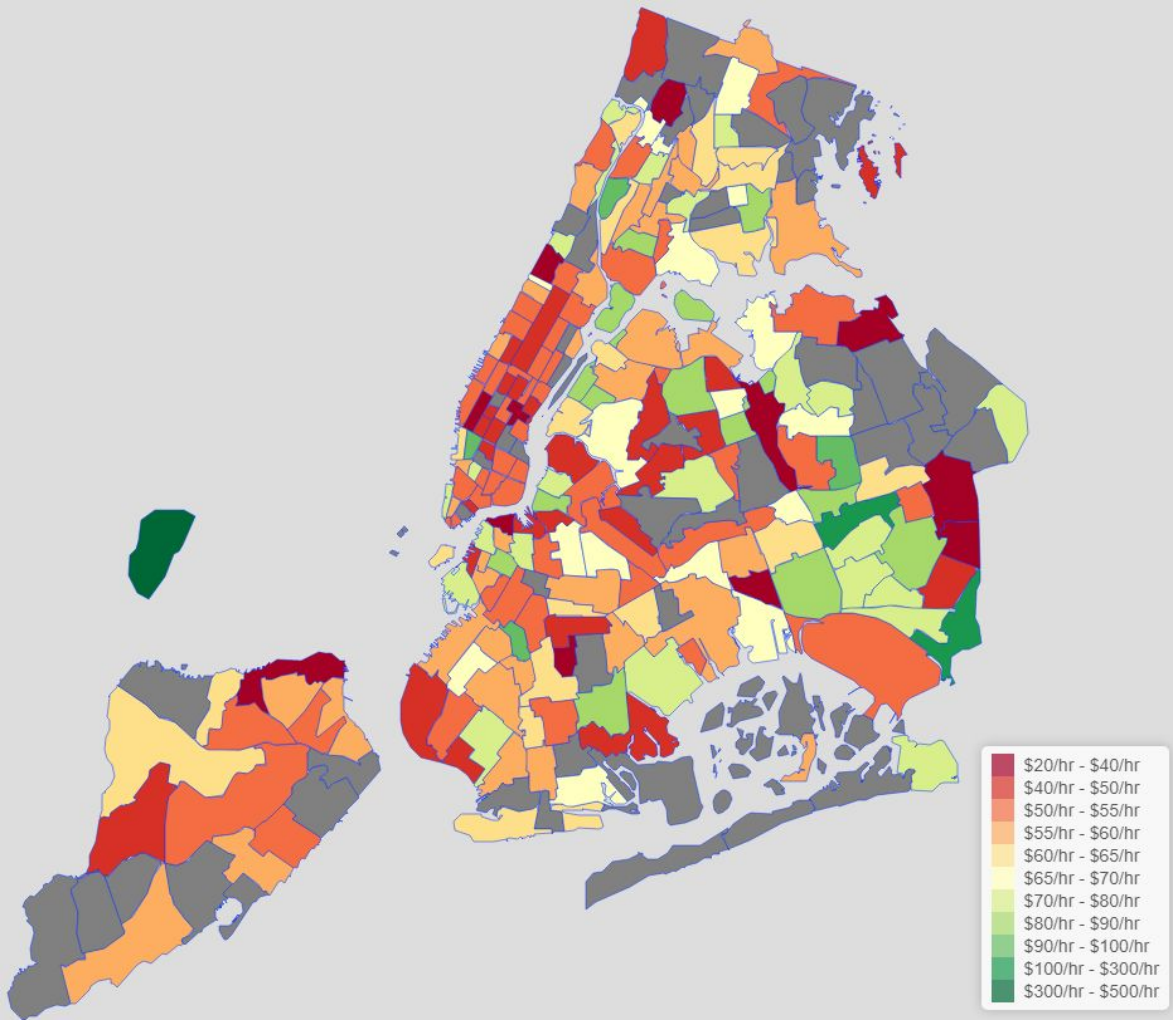

Figure 1: Average Hourly Rate of Taxis with Passenger in the Cab by NYC Taxi Zone

Additionally, the most valuable pickup times were between 5 and 6 in the morning, with fares picked up in that time being worth around $70/hr on average (again, note that this rate only applies when a passenger is in the cab, the actual hourly rate for a driver will be lower). All other early morning hours also had relatively high income per hour. Figure 2 shows income per hour by time of day.

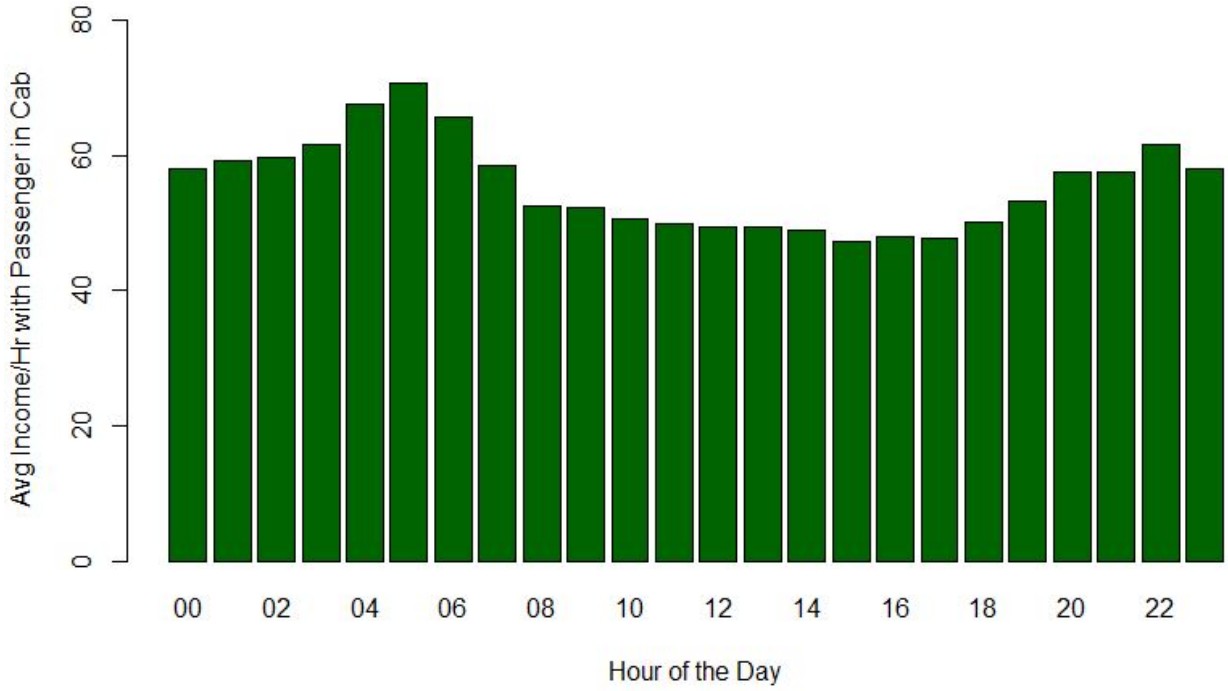

Figure 2: Average Hourly Rate of Taxis with Passenger in the Cab by Hour of the Day

For the most part, further separating out the data by the day of the week shows comparatively little differences. Sunday mornings (around 6 to 8) show relatively good hourly rates, around $82 to $85/hr. Most strangely, Tuesday between 10PM and 11PM is the single hour of the week that is most valuable, at $93/hr on average. The next most valuable evening hour in the week is Monday between 11PM and midnight, with a value of only $66/hr. It seems unlikely to me that one particular hour on Tuesday evenings breaks the overall pattern, although there may be a reasonable explanation as to why this happened to be true in June 2017.

Potential Enrichment for Dataset and Unnecessary Variables

The most significant factor missing from this analysis is some indication of how often a taxi would actually have a fare in the cab. The simplest way to incorporate this data into the existing data would be the addition of a column indicating the time of the previous dropoff. It could also be derived if taxi IDs were added to the database. That being said, either of these options would decrease anonymity in the database, which may be of greater concern.

The data also lacks any data on related costs for drivers; such as fuel, leasing costs, or insurance. However, some of these may be constants and effectively ignored for maximizing income, while others may be well correlated with time or distance driven and therefore can be derived with little additional data. There is also no data for any cash tips received, but that is unlikely to change.

Some of the provided data was unnecessary for this analysis. The “store and forward” flag is particularly unhelpful when trying to maximize income. Vendor ID, payment type, passenger count, and the fields for various taxes and surcharges also do not appear helpful, although analyses involving them could yield interesting insights. This analysis did not use the dropoff location or trip distance, although they could be used to try to find chains between high value zones between which passengers tend to ride, allowing a driver to avoid wasted return trips of lower value.

Potential Next Steps

The clear next step in this analysis is to cross the searches for the best time and for the best place together, and try to find how the best place to drive changes over time. It would also be useful to control for any accidental correlation between time and place that would throw off a basic analysis.

Further steps could include more detailed analysis involving the unused fields to estimate the level of competition in a given zone or the additional costs associated with driving a taxi. Searching for other data sources to integrate with the data provided by the NYC Taxi and Limousine Commission could also help with these tasks, and provide a better context in which to make a plan for maximizing taxi driver income.