MS&E 231 HW2

Andrew (Foster) Docherty, Michael Spencer, Jorge Nam Song

I. Objective

Our investigation targets the classic behavioral economics question of "Why is it so hard to catch a cab in the rain?" Building off of the work of famous studies into this question such as Farber and Camerer et al., and their use of intuitive supply-demand modeling, target earners theory, and more, we seek to employ a modern approach to the question. That is, we use data on all trips from New York City taxicab trips from 2010-2013 to model and understand the behavior of taxicab drivers on rainy days.

II. Literature review

Our work primarily builds off the papers of Farber and Camerer et al., who offered strong yet contrasting theories of the behavior of New York City taxicab drivers on rainy days.

In the study *Labor supply of New York City cabdrivers: One day at a time*, Camerer et al. posit that cab drivers take their earnings one day at a time, setting targeted earnings goals for themselves. As such, during rainy days when the opportunity to earn is increased, due to increased demand, they will earn their target sooner and accordingly, go home earlier, limiting the availability of cabs and thus making it harder to "catch a cab in the rain"[1].

Farber on the other hand, replicated and extended the work of Camerer et al. by analyzing the complete driving records of all of New York City taxi drivers during 2009 and 2013. Unlike Camerer, Farber's study showed that "drivers tend to respond positively to unanticipated as well as anticipated increases in earning opportunities", which aligns with the neoclassical optimizing model of labor supply[2]. Farber claimed that taxi drivers are more likely to stop working for the day based on their accumulated work hours rather than their accumulated income, implying that any difficulties in catching a cab were due to an increase in supply that did not meet demand.

III. Data sources

Our data for this investigation comes from two sources. The first is fare and trip data from ~700M NYC taxicab rides from 2010 to 2013. This differs greatly from Farber's random subset of rides from 2009-2013, as our data is over 100GB and generally more comprehensive. Our second source of data comes from NOAA and allows us to determine the hourly precipitation in New York City for each hour in our trip dataset.

IV. Methodology

Given the massive size of our data, we decided to collect our aggregated fields of interest via three consecutive MapReduce operations, written in Python3 and executed via AWS Elastic MapReduce (EMR). Given that the relevant NYC taxicab data was initially stored in two separate tables, we first had to join the datasets into one. In the mapping step, we identify unique trips by using the driver's annually randomized hack license ID and the pickup date-time of the trip to create a unique key. The input value for the MapReduce was the full data from the original table, as well as a flag indicating which table, either trip or fare, that the row originated from. In the reducing step, we join the unique data from each row of "trip" and "fare" into one comprehensive row. In addition, we apply a number of cleaning operations, such as dropping rows with erroneous data (e.g. coordinates of zero) and correcting any potentially unreliable data (e.g. we re-calculated trip duration in seconds using the dropoff and pickup times).

---

[1] Camerer et al. *Labor Supply of New York City Cabdrivers: One day at a time.*

[2] Farber *Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers*

Using the joined data, we then performed a second MapReduce in which we group by date and hour for each driver to calculate fields of interest for our analysis. In the mapping step, we split trips spanning multiple hours into smaller trips to make downstream calculations easier. For instance, we convert a 3:30pm - 4:30pm trip into separate 3:30pm - 4pm and 4pm - 4:30pm trips, updating related trip metrics accordingly. In this step we also use a dual-key mapping approach by using both the hack license & year and the pickup date-time as keys. The hack license & year key is printed first, and used to map and sort the data. We then print a pickup-date time key to be used in the reducer for grouping by date and hour.

Finally, on this data grouped by date and hour, we conduct a third MapReduce which performs a simple summation across all fields of interest. In the mapping step, we map each row (e.g. trip) by using the date-hour as the key, and the fields as the value. In the reducing step we sum across the fields for each driver in the date-hour, resulting in aggregated values for each date-hour. After these three MapReduce operations, we joined the aggregated data with NOAA precipitation data, dropping one outlying row in which earnings seemed to be negative. This finalized data is then used in our analysis as below.

V. Analysis and Results

We begin our analysis by investigating the supply of taxicabs on rainy vs. non-rainy days. While the data does not directly decompose supply and demand, we can still look for indicators of higher or lower supply on each of rainy and non-rainy days. In looking at the average number of unique on-duty drivers for each hour (see Figure 1), we find that there are more drivers during working hours on rainy days than on dry days. Furthermore, we see in Figure 2 that on rainy days taxicab drivers spend a greater fraction of each hour on-duty, meaning they are most likely actively looking for rides for more of the day. This again suggests that supply is higher on these rainy days.

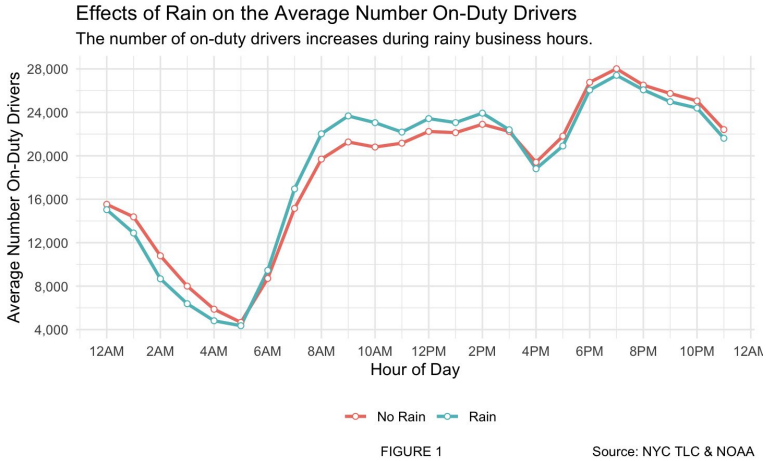

FIGURE 1 Source: NYC TLC & NOAA

On the demand side, we see in Figure 3 that the passenger utilization of taxis during their on-duty time is higher on rainy days. That is, when drivers are actively on-duty, they spend less time searching for passengers to drive and more time actively carrying them. We also see some indication that on rainy days passengers are more dispersed, and as such, demand more rides overall. This was demonstrated by the lower average number of passengers per trip on rainy days relative to non-rainy days.

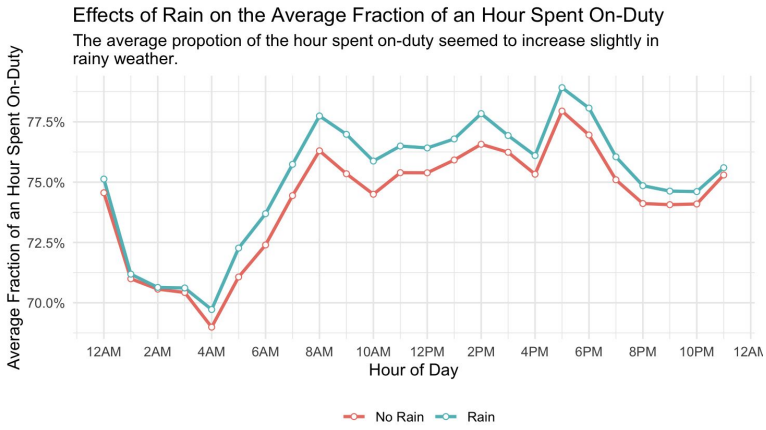

FIGURE 2 Source: NYC TLC & NOAA

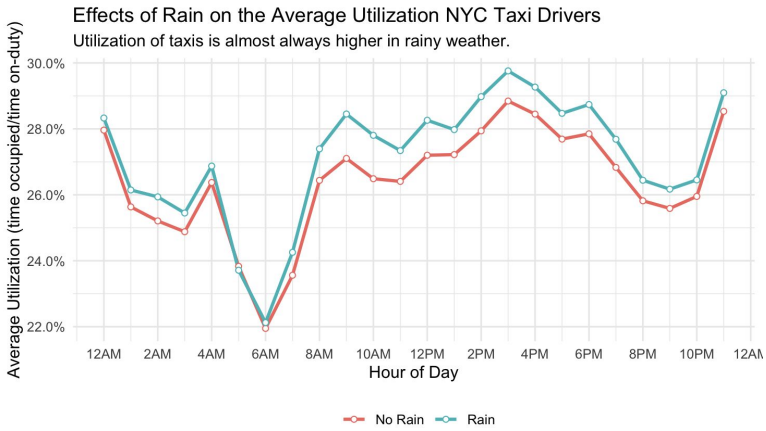

FIGURE 3 Source: NYC TLC & NOAA

Finally, we were interested to find that average wages for rainy days were lower for almost all hours of the day, as seen in Figure 4. These results are discussed further in the following section.

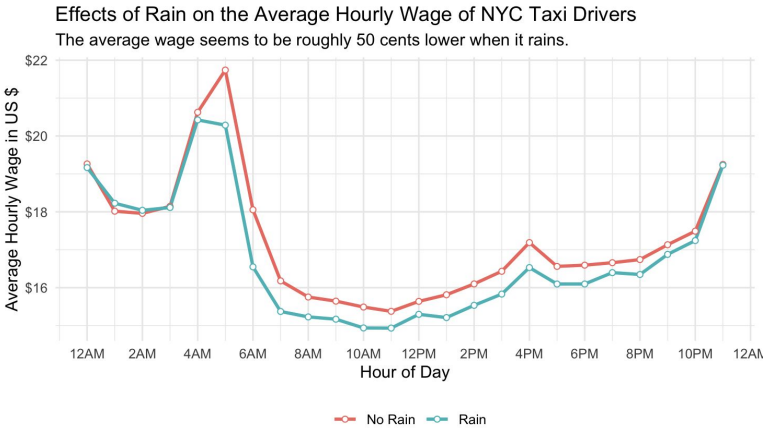

Figure 4 Source: NYC TLC & NOAA

VI. Conclusions and Limitations

Given the results of our analysis, we are able to engage with Farber and Camerer et al. on whether target earners theory explains why it is harder to catch a cab in the rain. Our analysis suggests that there is an increase in supply of taxicab drivers on rainy days, and that this could be a response to the increased demand for taxicabs on these rainy days. This would align well with Farber's neoclassical argument, and is not entirely unexpected given his analysis was conducted on 2009-2013 NYC data. Target earner's theory is not apparent in the data, as we see no indication of a significant decline in supply relative to the status quo in dry hours. However, what is interesting about the neoclassical approach is that the average wages we see are actually lower on rainy days. This could be because on rainy days, despite the increase in overall trips from greater demand and supply, taxicab drivers face more traffic and lower fares on trips than they would on non-rainy days due to the method by which fares are calculated (e.g. the relative weighting of distance vs. stopped time and how each is charged). Nonetheless, it appears driver's expectations of higher wages from higher demand persist and supply pours into the city streets.

It is important to note some of the limitations of this work. Namely, the integrity of the data may not be the best. In our own work and in Farber's work, data integrity was an issue and some data had to be dropped. It cannot necessarily be known whether all issues were fixed. In addition, we had to make assumptions about what it means to be on-duty vs. off-duty. Our assumption that thirty minutes or more of non-occupied time meant a driver was off-duty could be wrong, and it could adversely affect our results. In addition, both of the theories this work is based on as well as our own work are attempting to identify individual behavior. In reality, this analysis would be most robust if done at an individual level and at a massive scale. If so, it would be much easier to draw conclusions about the behavior of taxicab drivers and how they react to rain. As it stands, our analysis as done at an aggregated level and drawing conclusions about driver behavior is difficult.