Big Data: Unveiling the optimal driving pattern and operation strategy for taxi drivers

Dongping Zhang[1], Huanye Liu[2], Ningyin Xu[3], Alice Mee Seon Chung[4]
Github Repository : <https://github.com/dpzhang/Project_AHDA>

June, 2016
(version 05.06.a)

I. Introduction

In the era of the sharing economy, taxi industry is heavily knocked out by ride-sharing companies like Uber and Lyft. The data presented in Figure 1 illustrates that taxi demands in Chicago has plummeted drastically since 2014. Before November 2016, taxi ridership was declining annually at a 35% rate, and had fallen by 55% cumulatively since its peak back in June 2014. While the market entrance of formidable competitors like Uber and Lyft make the decline of taxi ridership seem inevitable, there is still a group of taxi drivers who manage to survive through this negative demand shock and makes decent annual earnings out from the even more competitive market in Chicago. In our paper, we systematically study large-scale taxi driver's operation patterns by analyzing their continuous digital traces in a real and complex city context. Using big data approach, we identify a set of valuable features and use them to construct some simple and novel statistics, which could effectively quantify each unique taxi trips in greater detail. By comparing the difference in driving patterns between high-income drivers and ordinary-income drivers, we are interested to see how could this group of high-income taxi drivers stand out in the Chicago market.

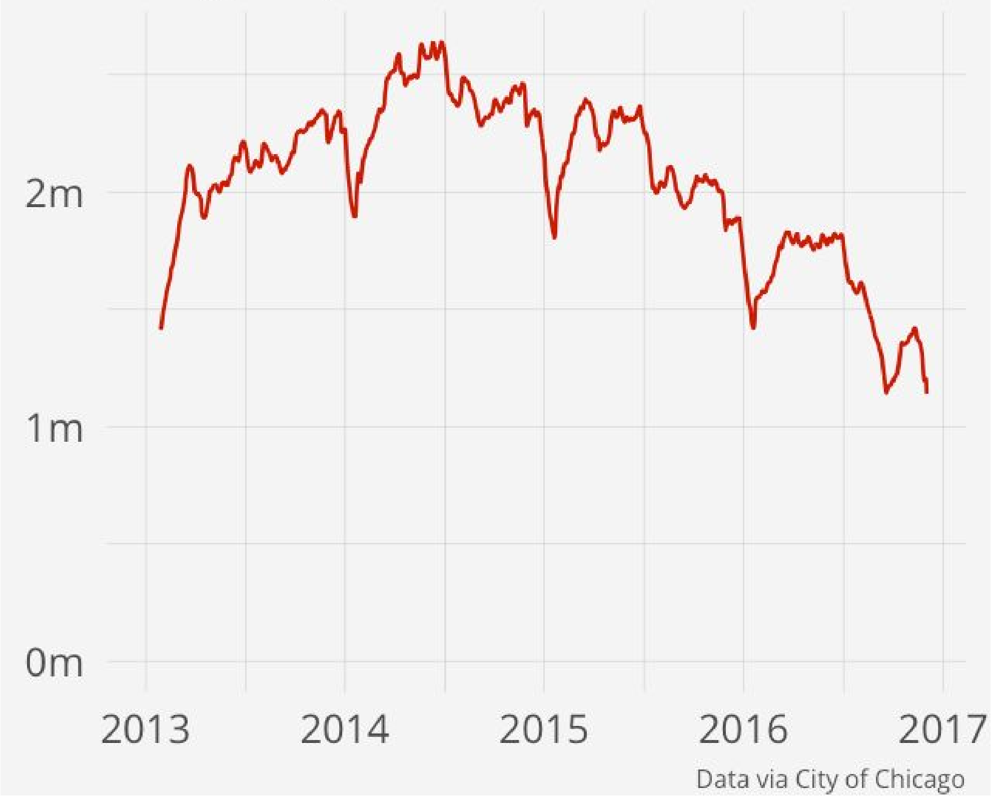
To refine our avenue of inquiry, we are interested in exploring: How do successful drivers optimize over the bounded resources of space and time? Are these successful drivers deliberately profiting by choosing longer routes that are more costly to the consumers? Are successfully drivers more hardworking or they are actually more efficient? In order to answer those questions, we would (1) systematically classify drivers into high-income and ordinary-income. (2) spatially and temporally quantify each trip's origin and destination. (3) use two effective and simple statistics to uncover route choices among two categories of drivers.

The paper is structured as follow. In the data section, we briefly describe the raw dataset. In the method section, we explain the methods we used and illustrate our big data approach and its advantages. In the classification section, we propose a way to systematically classify drivers into high-income and ordinary-income based on their annual income. In the

[1] The University of Chicago, MA in Computational Social Sciences, <dpzhang@uchicago.edu>
[2] The University of Chicago, MA in Computational Social Sciences, <huanyeliu@uchicago.edu>
[3] The University of Chicago, MA in Computational Social Sciences, <nyxu@uchicago.edu>
[4] The University of Chicago, MA in Computational Social Sciences, <alicearimchung1@uchicago.edu>

hypotheses section, we would state all hypotheses tested. In the result section, we compare those statistically quantified driving patterns between high-income and low-income drivers conditioning on different spatial and temporal levels. Finally in the conclusion, we would summarize our findings and discuss the challenges we faced and how we overcome those challenges.

Fig 1: Chicago Monthly taxi pickups (Trailing over 28 days)



II. Data Section

In this paper, we used Chicago Taxi Trips dataset to conduct our analysis. The dataset was recently released by the City of Chicago back in 2013 and could be downloaded from the Chicago Data Portal. The raw dataset is approximately 40 GBs, with each row to be one of 110 million unique taxi trips from January 1st, 2013 to April 1st, 2017. There are 23 columns containing 23 variables, which include some important identification variables such as trip ID and taxi ID, some trip-level statistics such as trip seconds and trip miles, detailed pickup and dropoff information such as pickup and dropoff areas with corresponding spatial coordinates, and some detailed records of trip fares and tips.

Although the number of observations in the data set is overwhelming, it is not clean and requires excessive processing so as to obtain the data set that is feasible to conduct our analysis. Trip-level variables that are crucial to our analysis are taxi ID, miles driven, trip duration, pickup and drop time, as well as pickup and dropoff spatial coordinates. In order to compute ratio of real path length over shortest path length (RRSL) and ratio of real path travel time over shortest path travel time (RRST) so as to use them for trip-level analysis, all variables mentioned are crucial to present in each observation (those two ratio statistics will be explained in Section III). Hence variables lacking any of those statistics are dropped from our analysis.

A brief summary statistics of observations missing spatial coordinates are presented in Table 1 below.

Table 1: Percentage of Missing Spatial Coordinates by Year

	2013	2014	2015	2016	Total
Total Trips	26,870,287	31,021,726	27,400,744	19,878,270	105,171,027
Spatial NA	4,646,758	4,527,054	4,226,875	2,757,663	16,158,350
Percentage	17.3%	14.6%	15.4%	13.9%	15.4%

III. Method

In order to obtain a general understanding of the data set, we subsetted 50,000 trips out of 110 million observations for testing purposes. It takes about 15 seconds on average to clean those 50,000 raw sample trips using MapReduce. This is highly costly if we run the data cleaning algorithm entirely on the local machine due to the fact that the algorithm needs to clean observations in a row-by-row basis. If cleaning 50,000 trips takes 15 seconds, it would cost 33,000 seconds, which is approximately 92 hours or 4 days, to complete the initial processing of the raw dataset. However, on Google Cloud - Dataproc, it only takes 2 hours for us to clean the 40 GB data, which is a huge efficiency improvement.

Drivers are classified into two categories: high-income and ordinary-income. The detailed classification scheme is covered in Section IV of the paper.

RRSL stands for ratio of real path length over shortest length. It is obtained by dividing actual trip miles over absolute trip distance, where absolute trip distance is defined as the straight-line distance connecting origin and destination coordinates in miles. The interpretation of RRSL is straightforward as it seems. The distance connecting two spatial coordinates could be assumed as the shortest distance possible for a driver to complete a trip from coordinate A to coordinate B. However, it is intuitive to realize such travel distance is infeasible to achieve in reality due to the complexity of urban plannings and traffic conditions. However, this shortest distance could be viewed as a benchmark to compare with the actual distance traveled by a driver. In short, RRSL could be interpreted as for every mile of the shortest straight-line distance between origin and destination coordinates, what is the number of extra miles a driver actually take. Thus, RRSL could provide us important information on route choices of a driver.

RRST stands for ratio of real path travel time over shortest path travel time. Similar to RRSL, RRST is obtained by dividing actual trip duration over absolute trip duration, where absolute trip duration could be computed by dividing absolute trip distance over "fastest-possible" travel speed. The "fastest-possible" travel speed in Chicago is 23.7 mph, and this statistic is obtained by averaging over all speed limit signs in Chicago. Thus, absolute trip duration could be viewed as hypothetically the shortest time required for a driver to complete a trip given spatial coordinates of pickup and dropoff locations. The interpretation of RRST is also intuitive and straightforward. As stated previously, because absolute trip duration is the hypothetical shortest-time possible for a driver to complete a trip given origin and destination. The ratio of actual trip path travel time over shortest path travel time could be interpreted as how many extra seconds does the driver actually take for every second if a driver is taking the shortest route and drive in fastest possible speed. Again, RRST is similar to RRSL as both of those two ratios provide intuitive and powerful information on the route choices of a driver and meanwhile they are both relatively easy and straightforward to compute.

Simply looking at one of the ratios of RRSL or RRST might not be that helpful and it would be difficult to formulate interpretation of drivers' route choices. However, reading RRSL together with RRST could provide profound implications on the operation pattern of a driver. For example, as explained previously, if both ratio equals to 1 for a trip, it means the driver is taking the shortest distance possible and using the shortest time possible, which could be defined as the most efficient trip.

Reading those two statistics on the same type of drivers is useless and not informative. Nevertheless, if we take a weighted average of RRSL and RRST of all high-income drivers and compare those two weighted averages of ratios with those of ordinary-income drivers while conditioning on some spatial or temporal units, the captured differences between those two weighted averages could systematically differentiate the driving behavior of high-income drivers from ordinary-income drivers. Detailed analyses and implementations of RRSL and RRST conditioning on different spatial and temporal units would be implemented in Section VI of the paper in greater detail.

IV. Algorithm

The natural way to process our trip data of such big size is to leverage the computational concurrency by applying the MapReduce paradigm on the Google Cloud Hadoop platform. The MapReduce paradigm perfectly meets the computational goal of our project: comparing the average RRSL and RRST between high-income drivers v.s. ordinary-income drivers and before v.s. after year 2016 conditioning on certain other variables such as the pickup and dropoff locations, the trip's time period of a day and the weekday trip v.s. the weekend trip.

The implementation is relatively straightforward. Each of the covariates we need to condition on is combined with the year of interest as a key of the yielded tuple, and the value of the yielded tuple consists of differences in average distance ratio, RRSL and RRST between the high-income drivers and the ordinary-income drivers. We obtain each result of interest by going through two steps:

Step 1: We first go through the whole data to extract the annual income information for each taxi driver. This is achieved by 1) the summing-up routine using both the taxi driver ID and year as the yielded key and the resulting total income as the yielded value, and then 2) for each driver, the yielded key, combine the year and the total income as the yielded value to figure out the average annual income. Finally, based on the distribution of the annual income, we classify all taxi drivers into two groups, the high-income group and ordinary income group, based on two selected thresholds. We implement the classification using the numpy library, and after that we write the classification to a dictionary and yield the resulting dictionary into a file for later use. The implementation is put in the reducer_final() function. For further information on the taxi driver income classification, please check Section IV.

Step 2: Based on the driver income classification, we first compare the difference in RRSL and RRST conditioning on pickup and dropoff region. First, we read in the income dictionary from the input file in mapper_init(), and later when we go through each line to extract the pickup and dropoff location in mapper(), we classify the taxi driver of interest into high income or ordinary income class in real time. Then we choose the pickup and dropoff region combined with the driver income class and the year as the yielded key, the tuple of average distance ratio combined with the RRST and RRSL as the yielded value, and feed them into the reducer for further aggregation. Once the reducer receives the yielded tuple, it extracts the driver income class and put it as one element of the yielded value(which is a tuple by itself) for later use, and meanwhile keeps the year and the pickup and dropoff region as the yielded key. Finally, a further reducer produces the results we need by computing the difference in average distance ratio, RRST and RRSL between the high income and ordinary income driver for each combination of year, pickup and dropoff region.

Similarly we compute the difference in average distance ratio, RRST and RRSL between the high income and ordinary income driver for each combination of year and the trip's time period of a day, and the difference for each combination of year and the weekday/weekend trip. Thanks to the MapReduce paradigm, we are able to do such kinds of conditional counting and averaging in a natural way by deliberately choosing different yielded keys of interest.

V. Classification

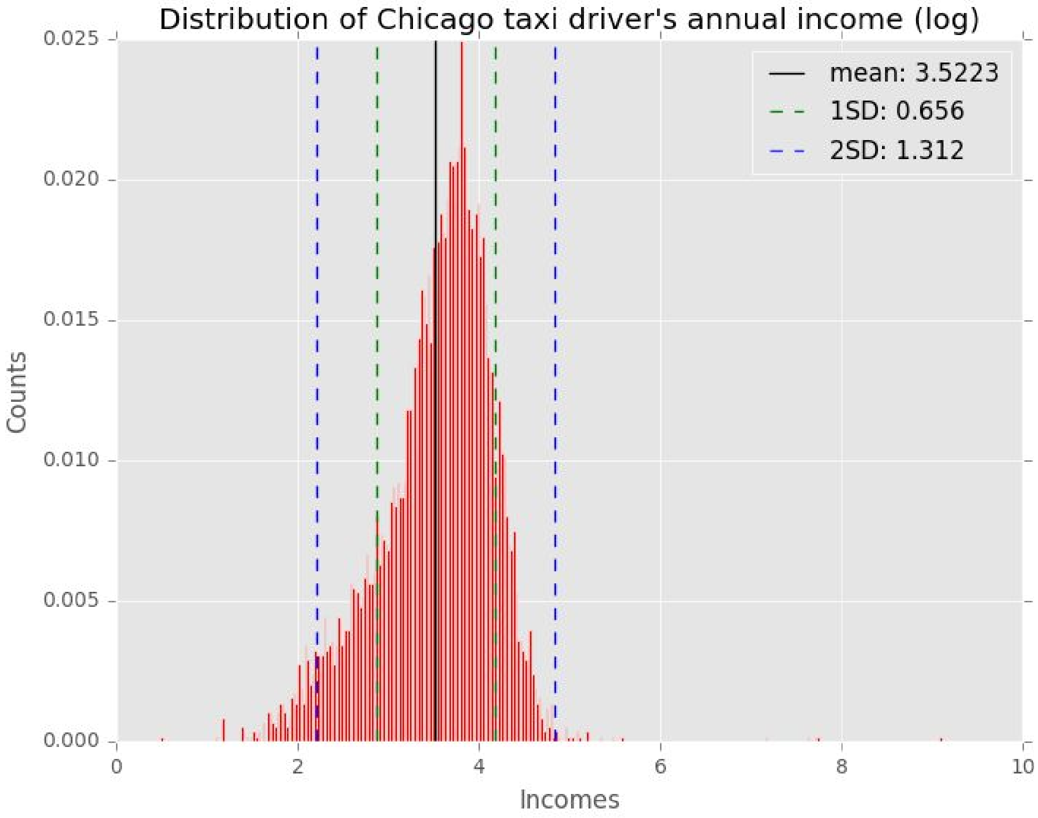
One unique variable in Chicago Taxi Trips dataset is the column specifies taxi ID. According to the codebook, each licensed Chicago taxi has a license number, indicated by the Illinois license plate number, a painted number on the body of the taxi, and the medallion on the taxi's hood. The Taxi ID in this dataset is not that license number. It is created specifically for this dataset, with no external meaning, to allow users to determine rides provided by the *same* taxi but not *which* taxi.

Using this variable, we are able to group all 110 million trips by unique taxi drivers. By extracting and summing fares of each trip from each group-by of taxi drivers and dividing the sum of fares of each taxi driver by corresponding number of years she has worked, average annual income could be computed so as to classify drivers by their average annual income.

In short, based on taxi ID, we identify all unique taxi drivers. Once all unique taxi drivers are identified, their average annual income is computed using MapReduce. The reason for our preference over the weighted annual income is because there could be drivers who drive more trips while other drivers who drive less trips within the same time-span. By looking at the weighted annual income, our interpretation could be less biased.

Similar to other wage variables, the raw annual incomes exhibit heavy right skewness. Thus, a log transformation is adapted so as to make the whole distribution look more normal. Figure 2 presents the distribution of annual income of every unique taxi drivers. After obtaining that normal-looking distribution, we classify and label high-income drivers to be all drivers located on the right of positive one standard deviation while ordinary-income drivers to be all drivers who fall between -2 standard deviation and below +1 standard deviation.

Fig 2: Distribution of Log Annual Income



VI. Hypotheses

Based on our research question, we have following major hypotheses that we like to testify:

1. High income drivers are able to make extra income from similar trips at the same time of the day.

2. High income drivers are willing to go to the outskirt neighborhoods where trips are more likely to be longer while ordinary-income drivers might just think they will earn more by staying in downtown, or staying in places where is more populous.
3. There are communities/regions that high income drivers are more likely to visit. Theses neighborhoods have certain characteristics in common.
4. There are differences in driving patterns, defined by RRSL and RRST.

VII. Analysis

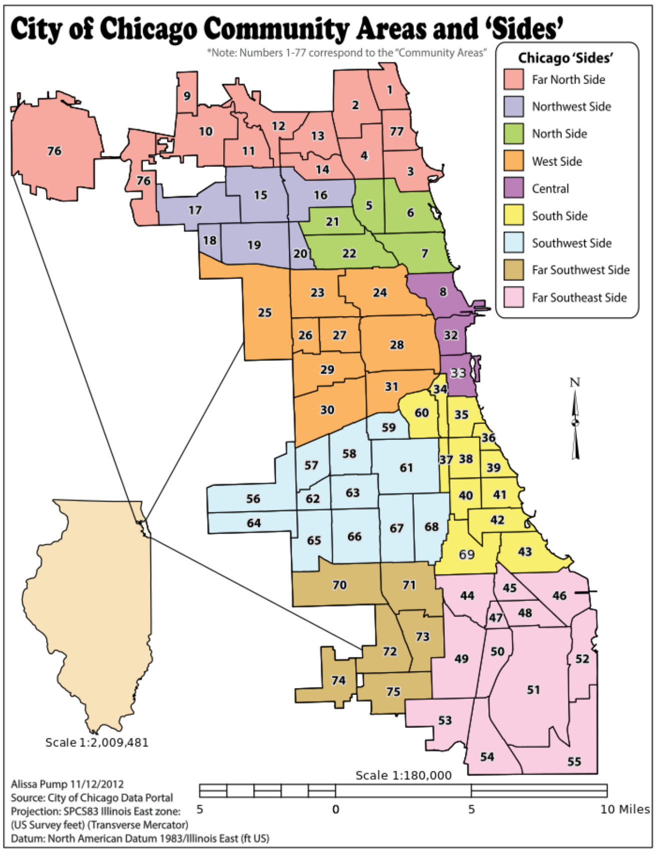
Part 1 : Conditions

We compare driving patterns between high-income and low-income drivers conditioning on different spatial and temporal levels. The spatial levels have a total of 9 regions introduced by the work of the Social Science Research Committee at UChicago, which is officially recognized by the City of Chicago. Table 2 and Figure 3 provides comprehensive regional information.

Table 2: Regional Classification

Direction	**Region**
North	Far North Side, Northwest Side, North Side
West	West side
East	East side
South	Southwest Side, South Side, Far Southwest Side, Far Southeast Side

Figure 3: City of Chicago Community Areas and Sides


City of Chicago Community Areas and 'Sides'
*Note: Numbers 1-77 correspond to the "Community Areas"
Chicago 'Sides'
Far North Side
Northwest Side
North Side
West Side
Central
South Side
Southwest Side
Far Southwest Side
Far Southeast Side
N
Scale 1:2,009,481
Scale 1:180,000
5
0
5
10 Miles
Alissa Pump 11/12/2012
Source: City of Chicago Data Portal
Projection: SPCS83 Illinois East zone:
(US Survey feet) (Transverse Mercator)
Datum: North American Datum 1983/Illinois East (ft US)

Next, we divided 24 hours into 8 time intervals and each interval has 3 hours. For example, 6AM to 9AM is time period 1 (morning rush hours) and 3PM to 6PM is time period 4 (evening rush hours). Using these eight temporal intervals, we are able to internalize traffic conditions in our analysis.

The other temporal level we used in our analysis is weekdays or weekend based on the assumption that taxi demands will vary in weekdays comparing to weekdays.

Part 2 : Analyze with three perspectives

We conduct our analysis from three perspectives : pickups and dropoffs analysis, flow analysis, and temporal analysis

Pickups and dropoffs Analysis

We aggregated the data based on pickup point since dropoff locations are completely random once a passenger got picked up. By counting the number of pickups based on 77 community areas we can create the density plot to see the locations where high income drivers usually pick up passengers and see if there is any difference between high and low income drivers regarding pickup locations.

Figure 4: Pickup Density Plot

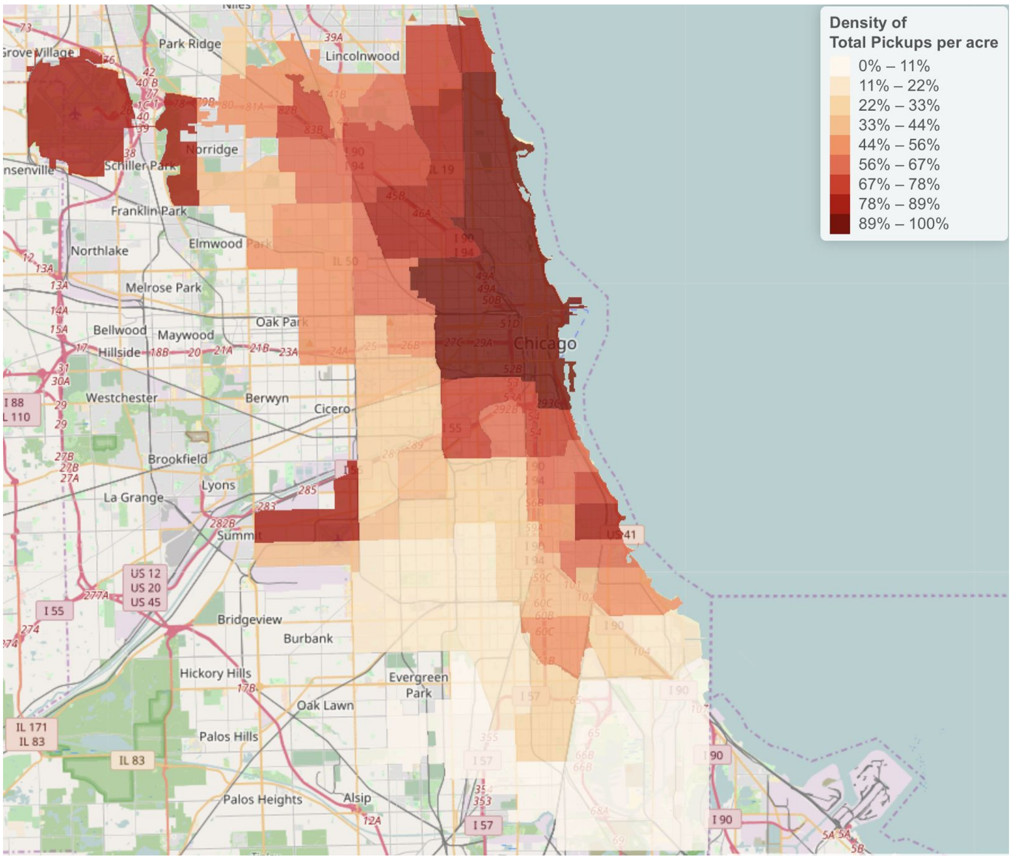
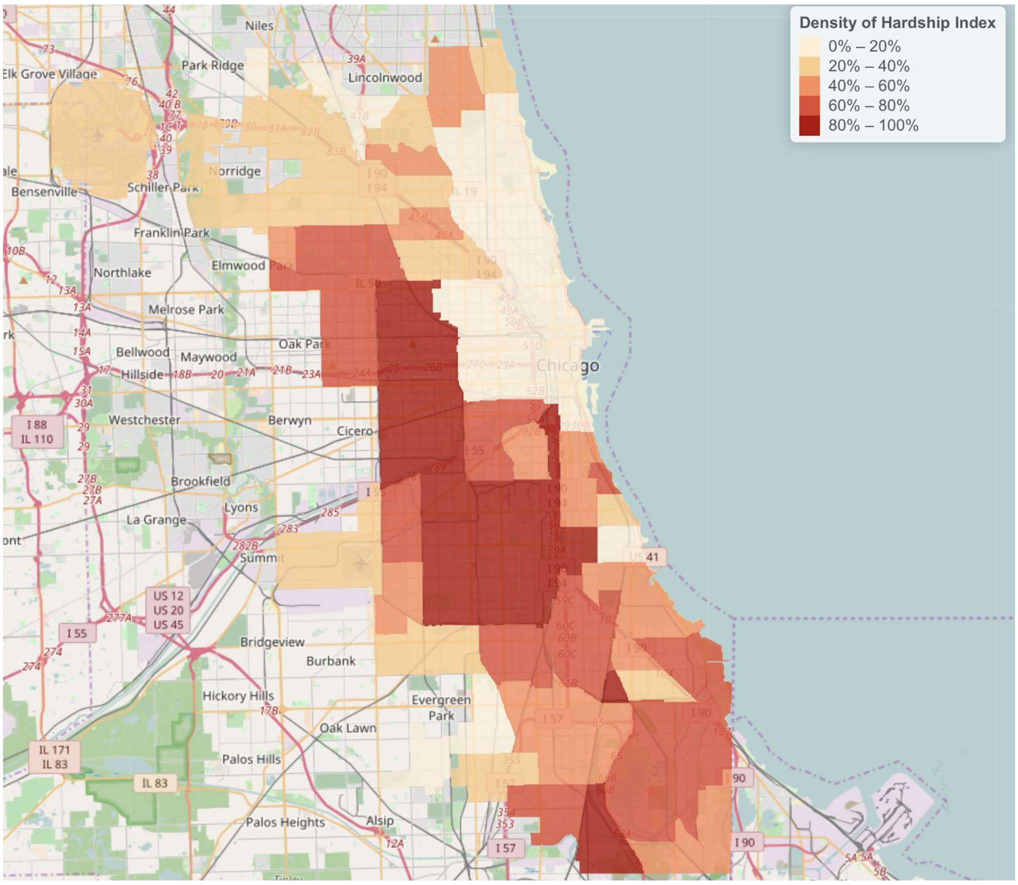


Fig 5: Hardship Index



Using MapReduce, we can get the counts of pickup locations conditioning on spatial units as well as driver labels. We further calculate the proportion of pickups on each 9 regions grouped on 77 Chicago communities and are able to produce a Pickup Density Plot in Figure 5.

According to Figure 5, the higher density location is Central, North Side and two International Airport, O'hare and Midway Airport. Combining Figure 5 together with Figure 6 of hardship index, an intuitive pattern could be seen that community with higher hardship index would have lower pickups.

Temporal Analysis

We aggregated the data based on 8 temporal units and weekdays/weekends since temporal units contains traffic on roads. In this perspective we can compare each temporal units considering rush hour and see the drivers' driving patterns. We can also compare the driving patterns between weekdays and weekends block on high-income driver and low-income driver.

Flow Analysis

To further conduct our analysis on flow of the trips and characteristics of route choices using RRSL and RRST from regions to regions. Some regional transportation with obvious patterns are presented below.

Flow Analysis Results 1

Table 3: RRSL contrast between High-income driver and low-income driver condition on Central to Central by year

Year	Origin	Destination	High - Low
2013	Central	Central	-0.8472
2014	Central	Central	-0.6522
2015	Central	Central	-0.4629
2016	Central	Central	-0.5124

Table 3 is obtained by computing the weighted average of RRSL of all trips classified by drivers in 4 years, from 2013 to 2016, where pickup location is central and drop-off location is central. Top-income drivers have lower RRSL consistently and -0.618 on average. Based on these statistics, we are able to claim that on average top-income drivers drive 0.618 miles less per absolute mile than low-income drivers. We can interpret the results as top-income drivers are more familiar with city road condition and all possible "shortcuts" to drive less miles.

Flow Analysis Results 2

Table 4: RRSL contrast of High-income driver and low-income driver condition on West to West Side by year

Year	Origin	Destination	High - Low
2013	West Side	West Side	-0.5325
2014	West Side	West Side	0.4342
2015	West Side	West Side	0.4986
2016	West Side	West Side	0.5565

Table 3 is obtained by computing the weighted average of RRSL of all trips classified by drivers in 4 years, from 2013 to 2016, where pickup location is central and drop-off location is central. Except 2013, high income drivers drive 0.499 miles more per absolute mile than low-income drivers in all other 3 years. Here, we have to note that west side is low-income area by Hardship Index. We are able to assume that high-drivers drive more because it is hard to find next passenger in that area.

Temporal Analysis

Temporal Analysis Results 1

Table 5: RRSL and RRST on WeekDays and Weekend by year

	Week Days		Weekend	
	H-L RRSL[5]	H-L RRST	H-L RRSL	H-LRRST
2013	-0.8244	-0.8780	-0.7169	-0.8274
2014	0.8599	1.6364	-0.0712	-0.7261
2015	-0.1251	-0.6555	-0.8627	-0.4001
2016	-0.4062	-1.1469	-0.6438	-0.0713

From above table 5, first we can see that overall high-income driver tend to drive less miles per absolute miles than low-income driver and high-income drivers take less travel time per shortest path travel time compared to low-income drivers. It means that high-income drivers are able to make more trips than low-income drivers in the same temporal interval by finishing the current trip efficiently and promptly. However when we compare with weekdays-weekends wide, still high-income drivers are faster and shorter drivers on

[5] H-L denotes the contrast between high-income driver and low-income driver.

weekdays. Even though the difference is smaller, this is because of increasing traffic on weekends. We are able to conclude that this is the main key driving patterns for high-income drivers to become so-called "high-income drivers".

Temporal Analysis Results 2

Table 6 : RRSL and RRST on Rush Hour by year

	2013		2014		2015		2016	
	H-L RRSL	H-L RRST	H-L RRSL	H-L RRST	H-L RRSL	H-L RRST	H-L RRSL	H-L RRST
AM6-AM9	-1.0711	-0.8512	-0.7169	-0.8274	-0.0357	-0.0194	0.0796	0.4156
PM3-PM6	-1.2816	-1.7092	-0.2271	-0.5970	-0.4519	-1.3661	-0.8280	-0.3711
PM6-PM9	-0.5376	-0.4789	-0.0146	-1.3761	-0.4500	-0.5971	-0.6248	-1.5474

We only compare RRSL and RRST during rush hour (AM6-AM9, PM3-PM6 and PM6-PM9). We can see clear driving pattern from table 6 that overall high-income driver tend to drive less miles per absolute miles than low-income driver and high-income drivers take less travel time per shortest path travel time compared to low-income drivers. It means that high-income drivers tend to drive fast and shorter in the rush hour and this implies high-income drivers can have more trips even in rush hour.

VIII. Conclusion

Summary

To sum up, we can generate three major results from our project. First, as an initial analysis of our dataset, we're able to classify and label 110 million drivers into top-income drivers and ordinary-income drivers, and see how much are top-income drivers' wage higher than ordinary-income drivers. Second, from pickup and dropoff analysis, we are able to conclude high-income drivers tend to visit high-income neighborhood most frequently as high-income neighborhoods have the highest probability of taxi demands according to Figure 5 Pickup Density Plot and Figure 6 Hardship Index. Third, from flow analysis, we observed that for drivers who move from central to central, high-income drivers drive less miles than low-income drivers. Our interpretation is that high-income drivers are familiar with the city roads and shortcuts, or they are more educated and is able to fully leverage technologies such as GPS. Interestingly this trends turns opposite when drivers drive west to west, that is, within low-income area, high-income drivers drive more than low-income drivers. Our interpretation is that those areas are where high-income drivers do not typically visit while those pickups are mainly done by low-income drivers who would typically drive longer miles due to unfamiliarity of road conditions or unskilled driving abilities. Third, through conducting temporal analysis, we observed that high-income drivers drives less miles and drives faster than low-income driver on both weekdays and weekends. This trend that high-income drivers tend to drive shorter routes using shorter time becomes weaker on weekends. Our interpretation is that traffic is likely to increases on weekends. Conditioning on rush hours, we observed same trends in previous temporal analysis. In rush hours, high-income drivers drive faster and less miles. It means high-income drivers work more efficient and have higher chance to ride more trips than ordinary-income drivers. To sum up, high-income drivers are familiar with city roads and layouts or more skilled in leveraging technologies such as GPS, which allows them to drive faster and less miles than low-income drivers. We can generalize this driving pattern of high-income driver as being “economical” drivers.

Challenges & Findings

MapReduce is a helpful tool and we leveraged it from either local machine or the cloud. Our major challenges in this project came from Google Cloud Dataproc, including the setup, bootstrap, uploading big dataset, and data processing. We met problems such as:

1. It took a lot of time to use our dataset in the disk when we’re running Dataproc, so we upload it to google cloud storage, which was a lot faster.
2. During bootstrap, it’s difficult for us to figure out which version we should install, whether we should upgrade pip3 before we install those packages. Tons of changes were added to our code and config file.
3. The quota from Google Cloud did not help too much either.