Chicago Taxi Trips Study

Joshua Balson <[jbalson@mail.smu.edu](mailto:jbalson@mail.smu.edu)>

Hieu Nguyen <[hdnguyen@mail.smu.edu](mailto:hdnguyen@mail.smu.edu)>

Cam Lu <[ctlu@mail.smu.edu](mailto:ctlu@mail.smu.edu)>

Master of Science Data Science, Southern Methodist University

Course: MSDS7330, Instructor: Jongwook Woo

**Abstract:** The purpose of this research is to analyze taxi rides taken in Chicago from 2013 through 2017 to determine whether the city of Chicago is “centralized”. We will answer this question by analyzing the ratio of taxi rides where the destination was either to or from downtown compared to all taxi rides taken from 2013 to 2017.

A secondary area of focus will be to analyze whether the metro system in Chicago complements or detracts from the business of taxi drivers. In other words, “are taxis competing with the metro system by covering similar routes or supplementing metro routes by getting people to and from train stations?”.

**1. Related Work**

A data set was published by Kaggle, but contains only taxi trips taken in 2016. Kaggle’s data set replaces latitude and longitude with ‘community areas’. This makes it more difficult to perform detailed analysis around pick-up and drop-off locations. Kaggle consumers have used this data set to analyze many things outside the scope of this analysis – such as rush hour heat maps, fare versus ride time, etc. However, there is not a lot (if anything) posted that deals directly with the purpose of our research. This may be due to the way that Kaggle modified the “raw” data set that is provided by the City of Chicago, which may have discouraged researchers from engaging in the type of analysis we wish to perform.

Due to the issues encountered with Kaggle’s data set, we chose to instead focus on the more robust data set (which ranges from 2013 through 2017 and includes lat/lon) provided by the City of Chicago. This data set was last updated in August of 2017. The City of Chicago also maintains a dashboard of the data. This dashboard contains many views which show time duration histograms, drop-off and pick-up location volumes, and many more. It should be noted that the City of Chicago maintains many other data sets that would be of interest to researchers and analysts[[1]](#footnote-1).

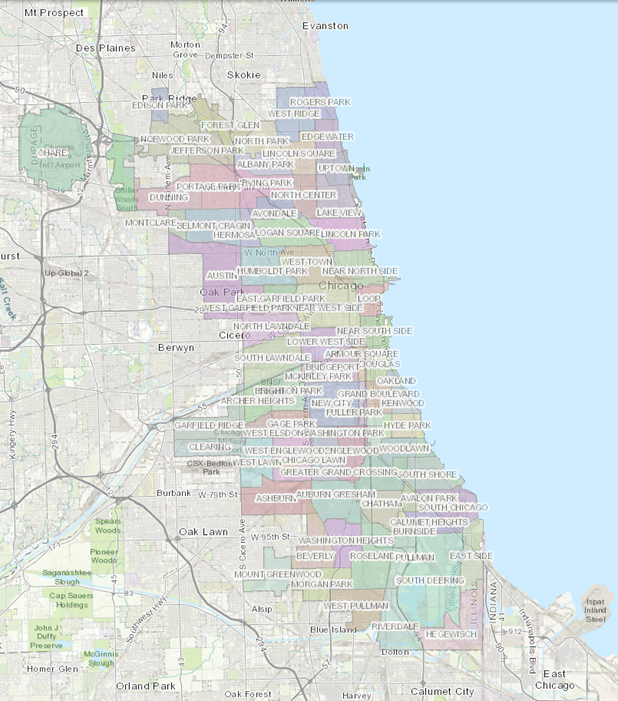
2. Introduction

At the time of this paper, the full City of Chicago data set contains 113 million rows of data and 23 colums of information, where each row represents one unique taxi trip. This makes it very difficult to download the data set, much less analyze it. MongoDB was used to upload the data and a mixture of SQL and Tableau was used for the analysis.

Before we get started on the analysis, it would be good to familiarize ourselves with the City of Chicago in general. It will also be necessary to know some basic facts about Chicago’s rapid transit system.

The city of Chicago is the third largest city in the U.S. with approximately 2.7 million residents and a metro population of over 9.5 million. Chicago’s downtown area is known as “the loop”[[2]](#footnote-2). This refers to the area encircled by the elevated (“L”) rapid transit tracks. The “L” has eight different routes, and average weekday ridership was over 750,000 in 2015[[3]](#footnote-3). Map 1 below shows the city of Chicago and its downtown “center”.

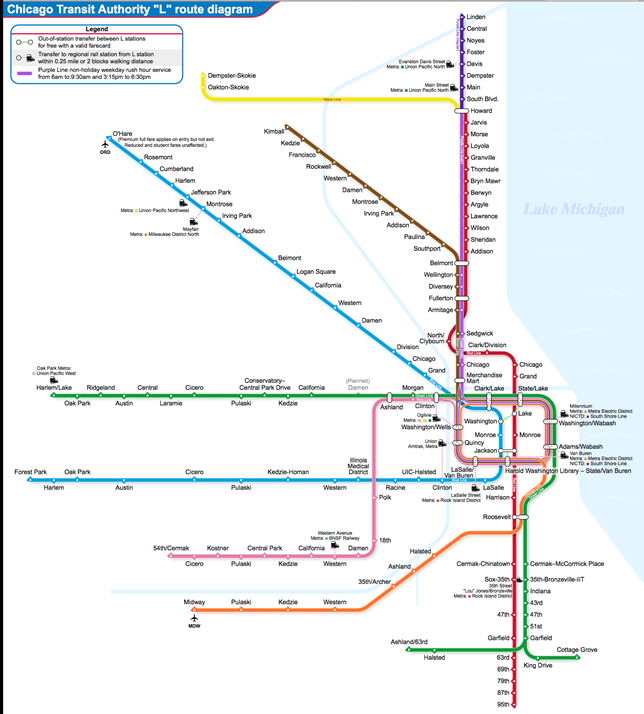
**Map 1:**



Chicago Downtown Center

Map 2 shows the city of Chicago’s “L” lines. The lines that form the loop in the map encircle Chicago’s downtown center.

**Map 2:**

****

**3. Analysis**

Figure 1-1 to the right displays the Chicago map with Chicago “L” railway line stations, trip origin, and trip destination. A green square represents drop-off locations, while a magenta x represents pick-up locations. Large blue dots represent main “L” terminals. If taxis were complementing “L” routes, one might expect to see a large concentration of both green (drop-offs) and magenta (pick-ups) corresponding to the blue dots (“L” terminals). In other words, the green and magenta would look ‘superimposed’ on top of the blue dots. However, Figure 1-1 shows the green and magenta in bands that are perpendicular to the “L” terminals. This suggests that taxis are detracting from “L” terminal business by both dropping off and picking up from locations far enough away from the blue dots to conclude that passengers are taking taxis to get to exact locations within the city and eschewing the “L” lines altogether.

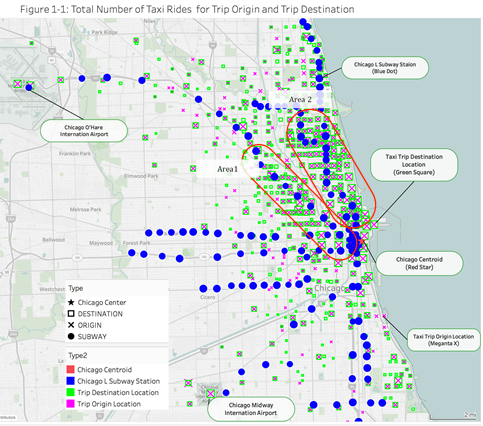
****

Figure 1-2 on the next page displays the zoom area for Chicago’s north-east region. Most locations for taxi pick-ups and drop-offs are a good distance away from the “L” line stations. The locations of taxi pick-ups and drop-offs are evenly distributed along the downtown area, especially the areas not covered by an “L” station. This is a further indication that taxis are not covering similar routes as the “L” stations, but are detracting from “L” lines by getting people directly to and from downtown

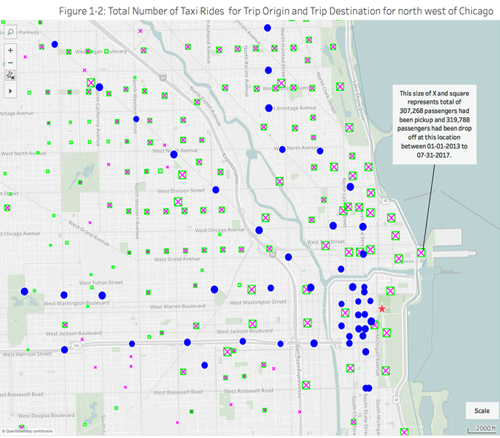


Figure 2-1 displays the average distance traveled at each location for each passenger origin and destination. A bigger x or circle represents a longer distance traveled. The data shows short trips around 2 miles average distance in the region of Chicago downtown. This is an indication that passengers use the taxi for short trips for the areas in downtown that are not covered by the subway, which also supports the evidence in the previous two figures that taxi and “L” line routes are competing with each other (i.e., detracting).

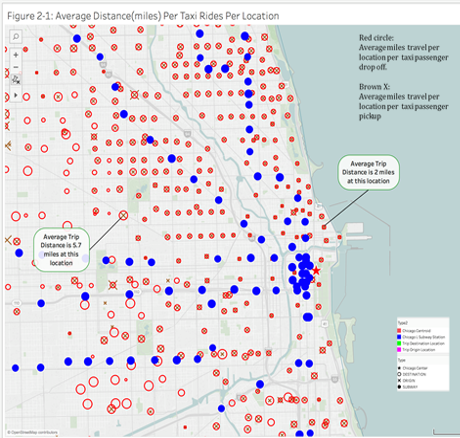
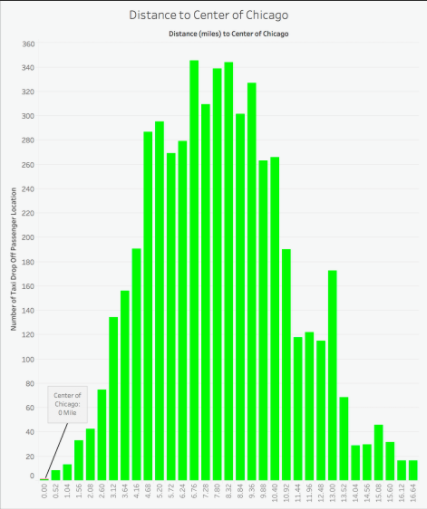
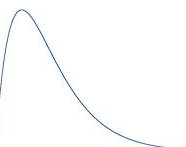


Figure 3-1a to the right shows a histogram of average distances from Chicago centroid in ½ mile increments along the x-axis and frequency of passenger drop-offs (in thousands) along the y-axis. If passenger drop-offs were “centralized”, then we would expect to see a heavy right skew in Figure 3-1a similar in shape to Figure 3-1b. However, Figure 3-1a appears to have a normal distribution. This suggests that the city of Chicago is not “centralized” when looking at taxi drop-offs.

**Figure 3-1a**



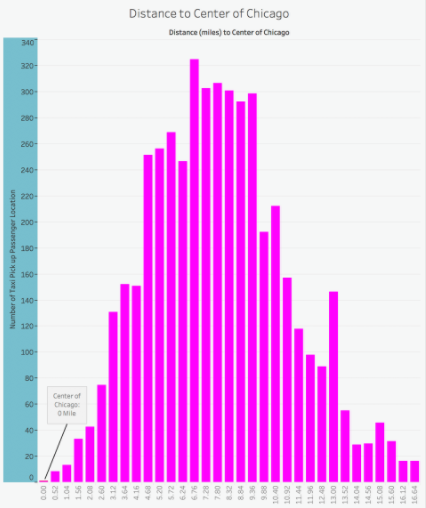
**Figure 3-1b**



Expected right skew if “centralized”

Figure 3-2 on the next page shows the same histogram, but for taxi passenger pick-ups instead of drop-offs. As with the drop-offs, we would expect to see a heavy right skew to the pick-up histogram in order to be considered “centralized”. Again we see a normal distribution in Figure 3-2 and not the expected right skew as shown in Figure 3-1b.

**Figure 3-2**



**4. Conclusion**

This analysis leads to conclusions that were not expected. Chicago, above all other U.S. cities (with the exception of perhaps New York), prides itself on the notion that its downtown is the hub of social and economic activity. Sometimes referred to as the “city of skyscrapers”, it boasts an impressive array of shopping, dining, and tourist attractions (Navy Pier, guided boat tours on Lake Michigan, one of the nation’s only free zoos, etc.) which are all densely packed around the Chicago centriod area.

Therefore, one would expect that Chicago would be “centralized” with respect to taxi pick-ups and drop-offs. However, from the analysis presented and the histograms shown in figure series 3, this does not appear to be the case.

Faced with this contradiction, it is reasonable to conclude that further analysis is warranted. For example, it would be interesting to perform this same analysis with extra tranches added to account for things like: seasonality, year, and weekdays versus weekends – just to name a few. Researches who wished to exand upon this analysis might also consider how the histograms would look if Chicago’s two major airports were excluded altogether from the analysis.

Above all, the biggest underlying factor that may be confounding this analysis may be socio-economic circumstance – where those living in underserved areas are not represented in this data because they neither drive nor take taxis, but simply take only the “L” and then walk to their final destinations without ever (or rarely) taking a taxi. For example, the northern parts of the city tend to be more affluent while the southern parts of the city tend to be more underserved. Referring back to Figure 1-1 we can see that the vast majority of taxi rides occur in the more affluent northern areas, while the southern underserved areas show much less taxi activity. Therefore, considering only taxi rides as a proxy to determine how “centralized” the city is may be only one component of a series of other missing factors that would be necessary to exhaustively analyze the main question that we seek an answer to in this analysis. A detailed step-by-step outline of how this analysis was performed is available at the GitHub link shown in the Reference section below. The authors of this paper would welcome anyone who wishes to further this research.

**References**

Kaggle link to data set: <https://www.kaggle.com/chicago/chicago-taxi-rides-2016/data>

City of Chicago Data Portal: <https://data.cityofchicago.org/d/wrvz-psew>

GitHub Link: <https://github.com/hdnguyen88/MSDS7330_GroupProject>

1. <https://data.cityofchicago.org> [↑](#footnote-ref-1)
2. https://www.choosechicago.com/plan-your-trip/chicago-fun-facts/ [↑](#footnote-ref-2)
3. http://www.cnn.com/2013/11/01/us/chicago-transit-authority-fast-facts/index.html [↑](#footnote-ref-3)