Chicago Taxi Rides

Final Report

Noah Leuthaeuser
University of Colorado - Boulder
nole1337@colorado.edu

Heesuk Jang
University of Colorado - Boulder
heesuk.jang@colorado.edu

Joe Alsko
University of Colorado - Boulder
joal2716@colorado.edu

Rei Isobe
University of Colorado - Boulder
reis9668@colorado.edu

ABSTRACT

This project looks at the Chicago Taxi Cab rides of 2016 to help taxi companies and drivers to be more profitable. Our first question we sought to answer was how do certain factors, such as weather, date/time, and income, affect ride frequency. The next question we sought to answer was which community areas gave the most generous tips? Following that question, which location had the most frequent pickups? And lastly, which routes were the most frequently taken between different areas? After mining through the data sets, we discovered that Near North had by far the most pickups of any area, followed by Loop and Near West. Weather, day of the week, and time correlated with pickup frequency in those areas. The areas with the higher pickup frequencies tended to have higher incomes as well, but those areas did not give out large tips amounts due to shorter rides.

1 INTRODUCTION

Many taxi cab companies are currently struggling to stay in business due to ride share companies, such as Uber and Lyft. The cost to keep the cab medallions and less business is causing taxi drivers to quit and switch to become Uber and Lyft drivers. To help taxi cab companies stay alive in these scenarios, we decided to research correlation between many different factors that would help taxi drivers be efficient. Time of day, day of week, and weather for each pickup location would let drivers know when it is the best time to pickup passengers. Looking into locations where there are higher tips would help the drivers earn more per ride and the areas with the most pickups would help them get the most rides per day. With this information, taxi cab drivers and companies would be able to plan routes and target locations to be efficient and profitable.

2 RELATED WORK

Some studies and surveys have indeed already been conducted on the Chicago taxi service, mostly by city transportation/commerce departments. Our group looked at three conducted by the City of Chicago Business Affairs and Consumer Protection office, the NYC Data Science Academy, and Todd W. Schneider, a Yale software engineer. These studies looked at pickup frequencies, rates, and pick up/drop off locations. Our study also aimed to study these attributes; however, unlike these studies, we were looking to find relations between them.

The Chicago BACP study (2014) looked at taxi rates in relation to value to the consumer as well as fairness of income for taxi drivers. The purpose was to develop a model that the city could use to determine the effect of fare changes on taxi drivers' income. This model accounted for different taxi ownerships (part time/full time, lease/ownership).

However, the study did not take into account locations of taxi service and looked at the Chicago taxi service as a whole. Also, the study combined fares and tips into revenue, not looking at each individually [1].

The NYC Data Science Academy study (2016) was more in line with our study (and used the same dataset). It looked at locations in Chicago and the average taxi pickup frequencies at those locations. Another part of this study mapped the ratio of pickups to drop offs in each community as well as average trip ranges. The study found that the airports and central section of the city receive the most frequent pickups and have the highest ratio of pickups to dropoffs. The trip ranges from the airports were consistently high, while the trip ranges from the city center were consistently low. So, in every location besides the airports, trip range seems to increase as pickup frequency decreases [2].

Todd W. Schneider conducted a survey on a similar dataset, but from 2013 to 2016 instead of just one year. His study seemed to be more aimed at trends in the Chicago taxi service as a whole. The study found that the taxi business in Chicago is declining faster than New York City's (55% decrease since 2013). However, his study did look at pickup frequency in certain locations of Chicago as well. Schneider mapped taxis' percent chance of a pickup within 30 minutes to each community in Chicago. Strangely, the study suggests that the airports were areas of very low pickup frequency, completely contradictory to the NYC Data Science Academy study. The rest of the data seems consistent, but this anomaly was something to focus on in our study. Schneider's study was built on his previous analysis of New York City taxis [3].

2.1 Difference from prior work

The Chicago BACP study focused on overall Chicago taxi service value and income balance. The NYC Data Science Academy study focused on the relationship between pickup frequency and trip range [2]. Todd Schneider's study looked at the trend of decline in the Chicago taxi business overall as well as pickup frequencies in different locations. Unlike the BACP study - as well as Schneider's - we focused on trends per community in Chicago [1,3]. Rather than just connecting pickup frequency and range per location, but we also looked at average tips and payment trends in each location as well. Alongside this, we also studied the correlations between these communities' taxi trends, weather, time and their poverty ratings.

3 DATA SET

The primary dataset for this project is the Chicago Taxi Rides dataset provided on Kaggle by the City of Chicago [4]. The dataset includes information about every taxi ride taken within the city for the year of 2016. The features of this dataset include a start and end timestamp, trip length in seconds and miles, pickup and dropoff locations in the form of census tract, community area, and geolocation, payment information including fare, tip, tolls, extras, and trip total, and the taxi company. It is divided into 12 separate files, one for each month of 2016. Each month contains around 1.7 million taxi trips.

We also used a secondary dataset containing socioeconomic information about the different areas in the city. This dataset is also provided by the City of Chicago [6], and is divided by the same community areas as the taxi rides dataset. The other features of this dataset include the community area name, percent of housing crowded, percent of households below poverty, unemployment information, per capita income, percent aged under 18 or over 64, and a hardship index.

Lastly, the weather dataset provided by AreaVibes was used [7]. The site contains monthly weather data for a single city or community area including average, maximum and minimum temperatures with precipitation, snow depth, and wind speed. The other features this site includes are air quality index and pollution index of the city or area as well as a list of nearby cities with good air quality and cities with similar population.

4 MAIN TECHNIQUES APPLIED

4.1 Preprocessing

The taxi rides dataset has already undergone some cursory preprocessing before being released to the public. All of the changes made are documented in the press release from the City of Chicago [5]. Major changes include masking time, location, and taxi medallion number for privacy. Pickup and dropoff time is rounded to the nearest 15 minutes, and location is given to the accuracy of the census tract. Implausible values were removed from the data, including negative lengths or costs, or extremely long trips. Some duplicates were also removed.

To get the data into a workable state, we continued to preprocess and clean. First, the twelve distinct months was merged into a single dataset to evaluate the data for the entire year. Some features from the dataset were not needed for our purposes and were dropped. Our analysis of pickup and dropoff location was on the community area level, so the census tract and geolocation columns for pickup and dropoff location were removed. Any rows without the pickup or dropoff location in the community area level were removed. The payment extras column was sparse and did not specify us what the extra payment was for; therefore, it was removed. The taxi ID column did not provide us with any meaningful insights, so it was also removed.

Some trips recorded a 0 value for both trip length in miles and trip length in seconds. When the value was 0 for either columns, the row was removed. Tips were only recorded for credit card payments, so any analysis on tip amount would not include cash payments. Because tips were not recorded for cash payments, the fare column was used for cost analysis rather than the total column. Hence, the tolls and trip totals were removed. Any rows with negative values for trip seconds, trip miles, or fares were also removed.

The selected socioeconomic factors from our secondary dataset were merged to the taxi rides dataset using a key of community area. Analysis and evaluation methods are described in depth in section 4.5.1.

To determine what factors would affect the frequency of pickups, we used data reduction and focused on the top 3 community areas with the most pickups. We used dimensionality and numerosity reduction by sampling and selecting the trip timestamps of those areas. Afterwards, we did some data transformation by parsing and constructing month, day, hour, day of the week, and weekday/weekend from the timestamp. Finally, we performed data integration by merging the data set with the data sets of the socioeconomic data of Chicago and weather data set.

4.2 Central Tendency

4.2.1 Median Tip Amount by Community Area

Initially in the tips study, we analyzed the mean tips per community area, but this seemed to produce highly skewed results. High outliers dragged the average up quite a bit for many community areas. Instead, we then found the median for each area and went from there.

4.2.2 Median Tip / Fare Ratio

Only looking at the plain median tip amount per area also produced skewed results since taxi trips from certain areas were consistently much longer and gave out larger tips because of larger fares. So, we divided the tip amount by the fares for each transaction and found the median tip ratio for each community area instead (% tip).

4.3 Relim (Recursive Elimination) Algorithm

The Relim algorithm was used to find the frequent 2-itemsets for community area pickup and dropoff locations. The implementation of Relim used for this analysis was found in the PyMining library. Relim is inspired by the FP-Growth algorithm, but does not use prefix trees or other data structures, making it less efficient but simpler to implement. For our purposes,

we only needed to determine 2-itemsets from a limited number of community areas, so the efficiency of the algorithm was not a problem.

The Relim algorithm was also used to generate the frequency of the 2-itemset on pickup latitude and longitude for the top 3 community areas. Using the spatial attributes as a frequent itemset mining criteria helped pinpoint the exact hotspots of taxi pickup demand.

The Relim algorithm was also used to determine the specific days of the year and community areas with highest ride frequency.

4.4 K-Means Clustering with Elbow Method

As an unsupervised learning and partitioning-based algorithm, K-means clustering with k = 4 was used to find clusters on pickup latitude, pickup longitude, and frequency of the location. To minimize the risk of overfitting and thereby increasing the quality of clustering, the Elbow method was used to validate the elbow point, which represents the optimal number of clusters on the curve of the sum of squared errors. *KMeans* from *sklearn.cluster* was imported to implement distance calculation to group the data into four similar characteristics and centroid updates for each cluster.

K-means clustering was also used for the outlier or anomaly detection process. Using the Euclidian distance function that is built in *pdist* and *squareform* from *scipy.spatial.distance*, pairwise distances between two data objects are computed in 3-dimensional space. Then, the threshold value was defined by averaging the maximum and minimum distance. If a distance value is greater than the threshold, data was regarded as an outlier.

4.5 Correlation Study

4.5.1 Area Income vs. Ride Count

The counts of all rides starting or ending in a certain community areas were summed to get an idea of the total taxi traffic volume going in and out of certain community areas. With information about the income levels for each community area, we were able to compare the total number of rides for each area to the income in that area. To compute a correlation between the *scipy* Stats *linregress* function was used.

4.5.2 Monthly Pickup Count vs. Weather

After the extraction of the month data from the trip start timestamp dimension for the top 3 community areas, we computed the counts of pickups in each month. With this information, we analyzed how monthly pickup demand within these areas is correlated with monthly local weather conditions including temperature, precipitation, snow accumulation, and wind speed. For the correlation study, the *pandas' corr* function was used.

4.5.3 Daily Pickup Count of Weekends vs. Weekdays

From the extracted month data, we created a function to determine the day of the week for each date by counting off each day and looping back after the 7^{th} day. With the updated data set, we were able to analyze and see the correlation between the pickup frequency and week of the day as well as weekend vs. weekday ride count for the top 3 community areas.

4.5.4 Daily Pickup Count of Holidays vs. Non-Holidays

We analyzed holiday/event factors on ride frequency in two different ways. One, we looked at frequency based on days per month (e.g. 24^{th} day of every month) and two, we looked at most frequently rode days in individual community areas. For the day-per-month analysis, we simply looked at number of rides on each day number 1-31. For the most frequently rode days, we used the relim algorithm to find the most frequent community area-day combinations.

4.6 Time Series Modeling

The timestamp included in our primary dataset had to be parsed in order for us to use it effectively. We used the *strftime* function to retrieve the month, day, and hour for each row by using the *apply* function and created a new column for each attribute, making it easier and faster to use. We used the parsed time series to look for correlations between weekdays vs. weekends, monthly pickup frequency, hourly pickup demands, which day of the month had the most pickups, holidays vs. non-holiday pickup frequencies and pickup frequency by day of the week.

5 KEY RESULTS

5.1 Best Tips by Community Area

To begin, a simple association study between tip amounts and communities was performed on the cleaned dataset. Since tips weren't recorded for cash transactions, rows with cash transactions were omitted. Outliers made the mean an unreliable method for looking at the average tip in each community area, so median values were used. Data based on pickup location vs. dropoff location tended to be pretty similar, but dropoff data was more complete (since some areas don't seem to have regular taxi pickups), so it was the primary source for this data. Fig. 1 is a map of the plain tip amounts per area.

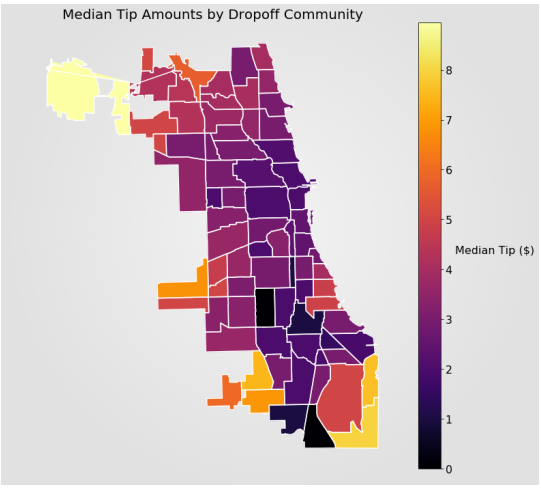

Figure 1: Map of the median tip amounts of each community area at dropoff

Based just on this, the area in the top left (O'Hare Airport) and the areas in the southwest corner (Hegewisch and East Side) hand out the biggest tips, while the downtown areas seem to hand out very little. This is because the trips from these "hot" areas are typically much longer than most. Instead, then, the fare to tip percentage was analyzed. Also appended is a map of dropoff frequencies per area. Obviously, taxi drivers can only choose where to pick up customers and not where to drop them off, so a map of pickup tip percentages and frequencies is also included. However, this data seems to be similar enough to draw conclusions from either.

What we can see is that O'Hare Airport turns out to be pretty average in tip giving, as well as the couple areas in the southwest. Most of the North and West Sides, in fact, look to be fairly average with tips. Austin and Garfield Park are subpar (Upper West Side) as well as a number of areas in the South Side. Obviously, the ride frequencies in the downtown areas (the Loop, Near North Side, Near West Side) are much higher than any other area. However, the median tip percentages also seem to peak in these areas compared to other adjacent areas. One last small thing to note is the somewhat anomalous area, Roseland, in the south which hits the peak tip percentage, unlike any other non-downtown community. The most frequently ridden areas are 6, 7, 8, 28, and 32. (See Fig. 2 and 3)

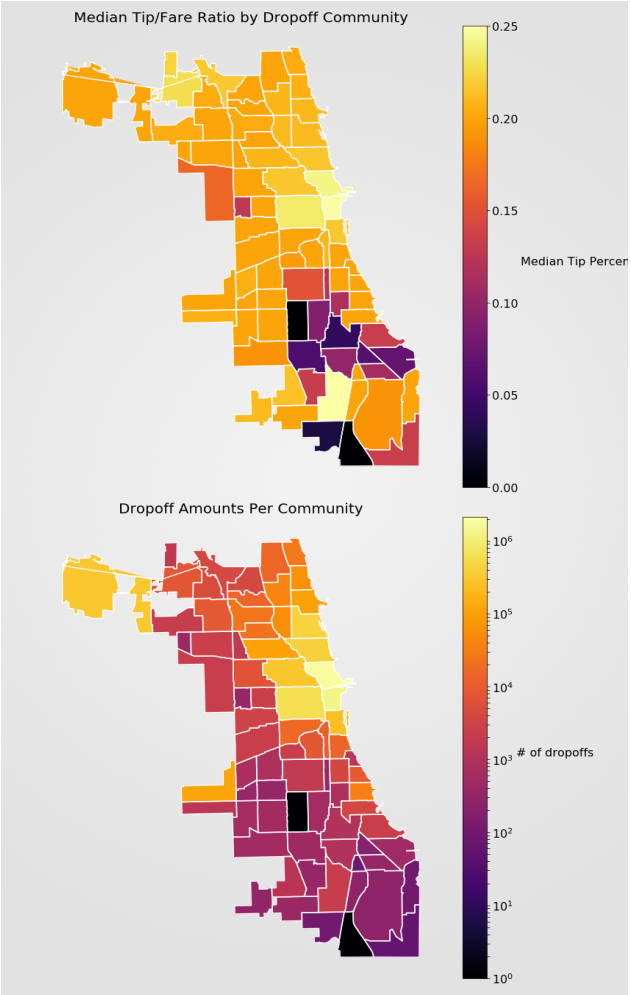

Figure 2: (Top) Map of the median tip/fare ratio by dropoff community area. (Bottom) Map of number of dropoffs at each community area.

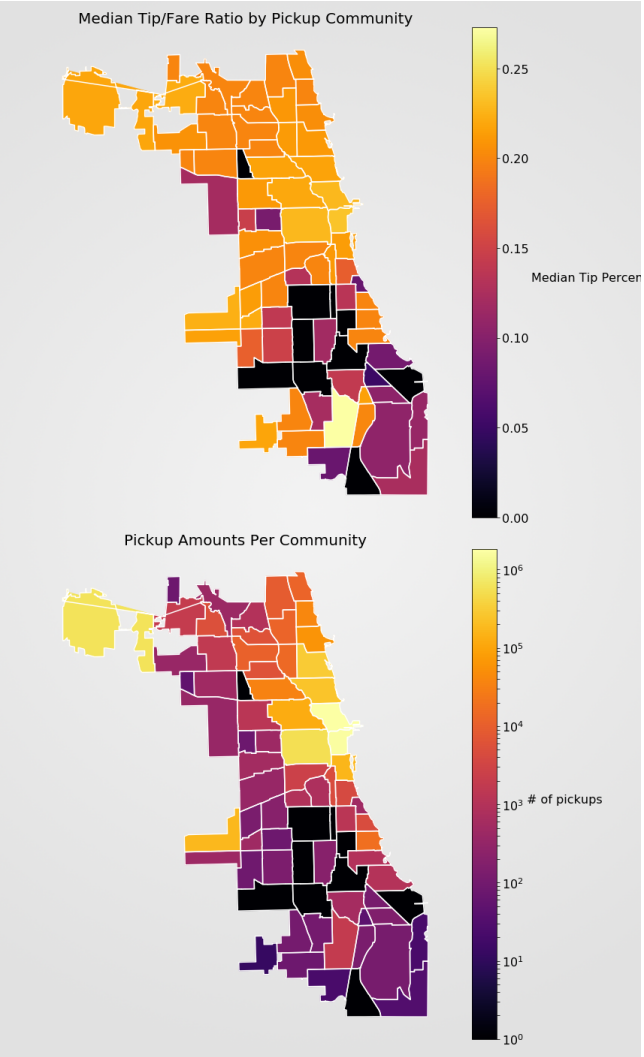

Figure 3: (Top) Map of the median tip/fare ratio by pickup community area. (Bottom) Map of number of pickups at each community area.

5.2 Frequently Traveled Routes

Table 1 shows ten of the most frequent community areas for pickup and dropoff. A ride is considered a member of the itemset when the pickup is in either area 1 or area 2, and the dropoff is in the other area out of the pair.

Area 1	Area 2	Count
Near North Side	Loop	2281451
Near North Side	Near West Side	899440
Near West Side	Loop	814120
Lincoln Park	Near North Side	475912
Near North Side	O'Hare	453597
Lake View	Near North Side	391833
Loop	O'Hare	342740
Near North Side	West Town	303845
Loop	Near South Side	274183
Near North Side	Near South Side	259442

Table 1: Five most frequently traveled community areas for dropoff and pickup

Most of the discovered frequently traveled community areas are around the downtown area of Chicago. Around 23.67% of rides pickup and drop off in the same community area. Those rides are not included in the frequent pattern analysis. The top four frequently traveled community areas are also adjacent to each other. The exception to most of the frequent community areas being downtown is O'Hare. There is also strong support for taxi rides going to or from the downtown area to the O'Hare airport. The trend of downtown areas with higher support counts continues down the full list of frequently traveled community areas, with a similar deviation for Midway airport. There are also high support counts for rides going to or from downtown and the Midway airport, with 139,746 rides between Midway and Near North Side. For reference, Fig. 2 displays a map of the Chicago community areas is included below.

Figure 4: Map of the Chicago Community Areas

5.3 Pickup Demand Hotspots

Exploring pickup latitude and longitude as a frequent 2-itemset, we determined the hot spot pickup locations in the top 3 community areas which demonstrated the highest taxi pickup demands among all 77 community areas in Chicago. For instance as shown in the Table 2, we created the resultant top 5 pickup hotspots for each area. A set of condensed historical pickup information like this can help us analyze the distribution patterns of pickup hot spots and predict the hot spots with the magnitude of pickup passenger demand.

The Most Frequent Pickup Locations in Top 3 Community Areas				
Community Area	Latitude	Longitude	Count	Approximate Location
Near North (8)	41.91293	-87.76142	555,344	1759 N Lotus Ave. Chicago, IL 60639
	41.89251	-87.62621	456,092	43 E Ohio St., Chicago, IL 60639
	41.78887	-87.61971	378,455	5806 S Prairie Ave., Chicago, IL 60637
Total Count : 3,215,696	41.75443	-87.66050	282,897	7646 S Bishop St., Chicago, IL 60620
No. of Pickup Loc : 21	41.66049	-87.63242	239,768	429 W 127th St., Chicago, IL 60628
Loop (32)	41.88099	-87.63275	1,097,844	10 S LaSalle St., Chicago, IL 60603 *(Downtown)*
	41.97883	-87.65379	628,970	5301 N Sheridan Rd, Chicago, IL 60640
Total Count : 2,338,884	41.87887	-87.62519	306,818	Theodore Thomas Orchestra Hall, Chicago, IL 60640
No. of Pickup Loc : 8	41.87741	-87.62197	205,170	Grant Park, Chicago, IL 60601
	41.87061	-87.62217	89,581	Grant Park, Chicago, IL 60601
Near West (28)	41.87926	-87.64265	328,605	621-601 Historic U.S. 66, Chicago, IL 60661
	41.88530	-87.64281	199,683	177-199 N Jefferson St., Chicago, IL 60661
Total Count : 862,424	41.87401	-87.66352	176,340	1445 W Harrison St., Chicago, IL 60607
No. of Pickup Loc : 18	41.99406	-87.59029	99,446	Near Lake Michigan, Chicago, IL 60611
	41.74249	-87.72261	27,498	4020 W 83rd St., Chicago, IL 60652

Note: An approximate pickup location is defined by its latitude and longitude.

Table 2: Top 3 pickup demand hotspots in top 3 community areas

In this instance, the summary makes it clear the top two locations in the Loop neighbourhood are the most frequent pickup spots in Chicago. It also visually shows the outliers. Especially the hot spot with demand of over 1 million pickups, which takes up to roughly 47% (=1,097,844/2,338,884) of the total demand in the Loop area, is easier to identify.

As compared to the fairly distributed taxi pickup demand across the Near North Side area, the demands in the Loop and Near West Side areas are also highly concentrated within a couple locations. The top pickup demand spot in the Loop neighbourhood was pinpointed in the central downtown of Chicago. As the central business district, there is a high concentration of jobs and population along with many colleges and high schools [8]. Due to these factors, it was expected to observe that the taxi passenger demand was greater than that in any other area. The top pickup location in the Near West Side, which is south-west of the West Loop, also has a concentration of roughly 38% (=328,605/862,424) pickup demand of the total. This location is surrounded by the University of Illinois at Chicago, parks, museums, many restaurants and notably two major transportation stations, Ogilvie and Amtrak Station-CHI, which provide public transportation between surrounding cities and downtown Chicago. These surroundings and the transportation convenience appear to be the major factors for the highest pickup demand in this location. It would not be surprising to see people tend to use more taxis than driving themselves to get to the transportation station.

5.4 Pickup Demand Hotspot Prediction

We used the K-means clustering method with k = 4 to predict the taxi demand hotspots in the top 30 community areas. The final centroids for all four clusters are described in Table 3.

	Centroids : K-Means Clustering			
	Latitude	Longitude	Count	Approximate Location
Cluster 1	41.77643	-87.70121	4,909	6426 S. Albany Ave, Chicago IL 60629
Cluster 2	41.95340	-87.66419	449,089	3958-3932 N Southport Ave, Chicago IL 60613
Cluster 3	41.90838	-87.66128	186,301	1460 N Elston Ave, Chicago IL 60642
Cluster 4	41.88099	-87.63275	1,097,844	10 S. LaSalle St, Chicago IL 60603 *(Downtown)*

Table 3: Final centroids for four clusters in top 30 community areas

We observed that all four centroids indicated sparse locations in four different community areas: center of Chicago Lawn (66), north-west side of Lake View (6), north-west corner of Near North Side (8), and downtown Chicago in the Loop area (32). The wide spread is evident for the low inter-class similarity among the clusters, which is often regarded as one of the key factors to determine the quality of clustering. Although seeking the optimal value of k may not always be an unambiguous using the Elbow method, it appeared to be a better approach than applying a randomly chosen initial centroids for a higher quality of clustering.

It also should be noted that cluster 4 contains a single location with an extremely high pickup demand as shown in Figure 5. Understanding its geolocation in central downtown Chicago, the notably high pickup demand was somewhat expected.

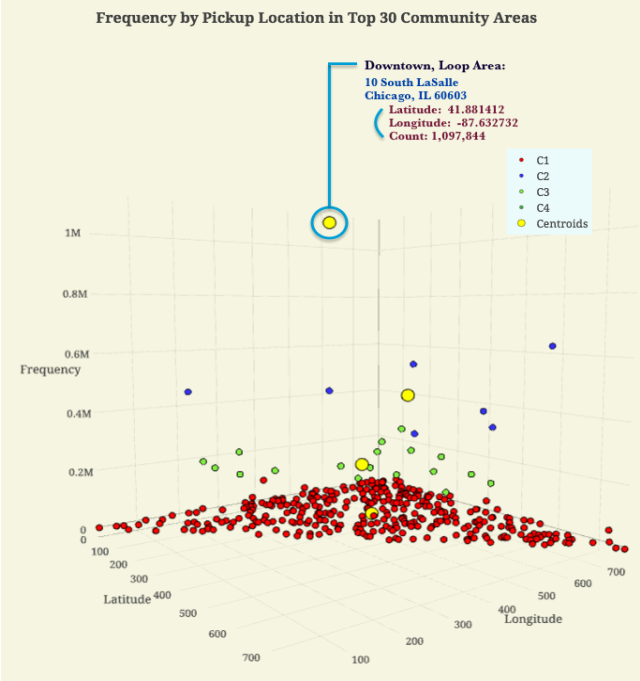

Figure 5: Frequency by pickup location in top 30 community areas (Link to a video demonstrating the interactive graph)

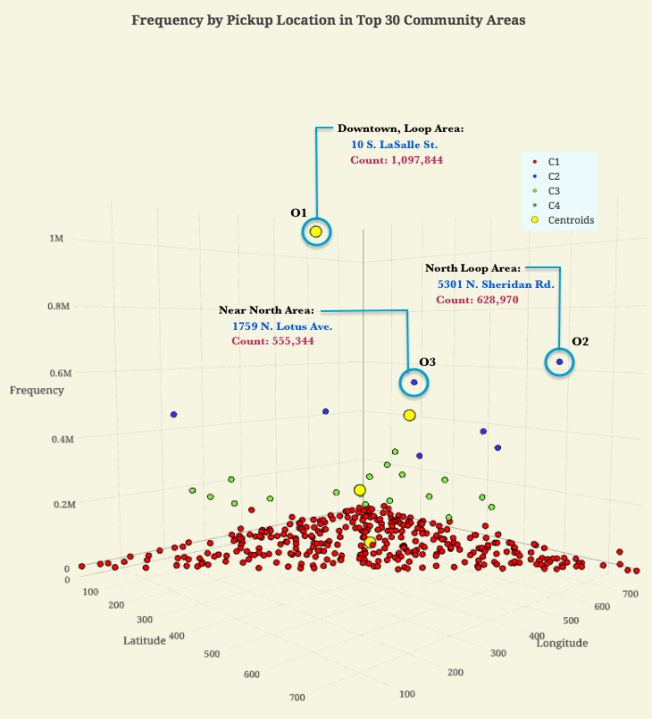

Figure 6: Detection of outliers using K-means clustering and Euclidian distance matrix

After observation of the extreme value, we applied the Euclidian distance function to calculate distances among data objects and defined a threshold at 548,923, which resulted from the average of maximum and minimum distance. We determined that it should be an outlier if a distance of a data instance is greater than the threshold. Applying this condition, the three outliers shown in the Figure 6 were detected. The first two highest pickup demand spots (O1 and O2) are located in the Loop area and the third instance (O3) is determined to be the hottest taxi passenger pickup spot in the Near North Side neighborhood.

5.5 Possible Causes of Pickup Frequency

5.5.1 Area Income

A positive correlation was discovered between the number of rides and the income in a community area with the correlation coefficient 0.708. This relationship is shown in Fig. 1. The graph is dominated primarily by the extreme high values for ride count and income for the central Chicago community areas, specifically Near North Side, the Loop, Near West Side, and Lincoln Park.

Figure 7: Comparison between number of rides and the income in a community area with the correlation coefficient.

5.5.2 Weather

Figure 8 below shows the general trend in pickup frequency per month for the top 3 highest frequency areas. The frequency is highest for the period between March and June, with lower frequencies during winter months.

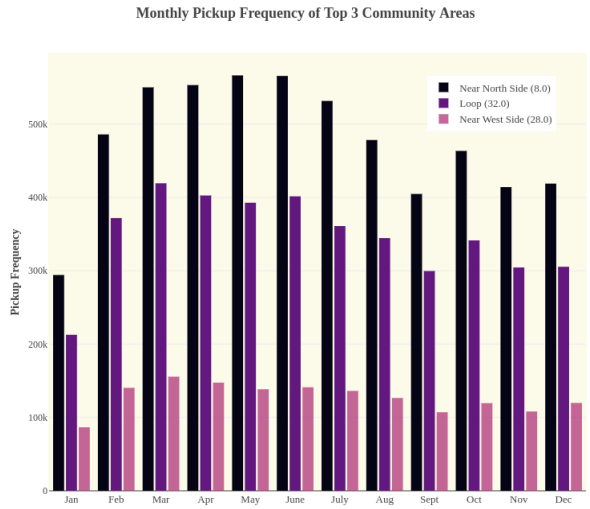

Figure 8: Monthly pickup frequency for top 3 community areas

Figure 9 shows the snow depth and precipitation each month for those same community areas. Interestingly, the pickup frequency for higher snow level precipitation levels is lower than when snow is not common.

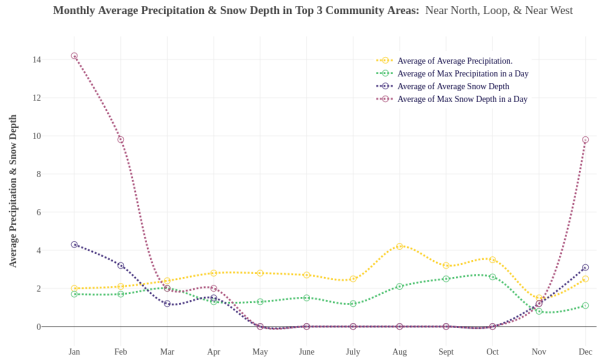

Figure 9: Monthly precipitation and snow depth for top 3 community areas

Table 4 shows the correlations between several weather conditions and the pickup counts for the top three community areas, with NN indicating Near North Side, LP indicating Loop, and NW indicating Near West Side. As indicated by the figures above, there is a negative correlation between average snowfall and pickup frequencies.

	Month	avg_temp	avg_prec	avg_snow_acc	avg_wind_spd	count_NN	count_LP	count_NW
Month	1.000000	0.273255	0.243468	-0.386599	-0.507837	-0.139851	-0.178340	-0.267171
avg_temp	0.273255	1.000000	0.621772	-0.897922	-0.860164	0.461307	0.336745	0.214132
avg_prec	0.243468	0.621772	1.000000	-0.576374	-0.580073	0.226852	0.199322	0.131690
avg_snow_acc	-0.386599	-0.897922	-0.576374	1.000000	0.756764	-0.578900	-0.502565	-0.341703
avg_wind_spd	-0.507837	-0.860164	-0.580073	0.756764	1.000000	-0.147367	-0.056334	0.049625
count_NN	-0.139851	0.461307	0.226852	-0.578900	-0.147367	1.000000	0.976156	0.940942
count_LP	-0.178340	0.336745	0.199322	-0.502565	-0.056334	0.976156	1.000000	0.972879
count_NW	-0.267171	0.214132	0.131690	-0.341703	0.049625	0.940942	0.972879	1.000000

Table 4: Correlations between precipitation and ride frequency in top 3 community areas

5.5.3 Weekends vs. Weekdays

Figure 10 below shows the average pickup frequency per day divided into weekdays and weekends for the top three community areas. Community area 8, the Loop, is the only community area that shows an increase in pickup frequency on the weekend. Community area 28 shows a sharp decline on the

weekend, and 32 shows a slight decline in pickups for the weekend.

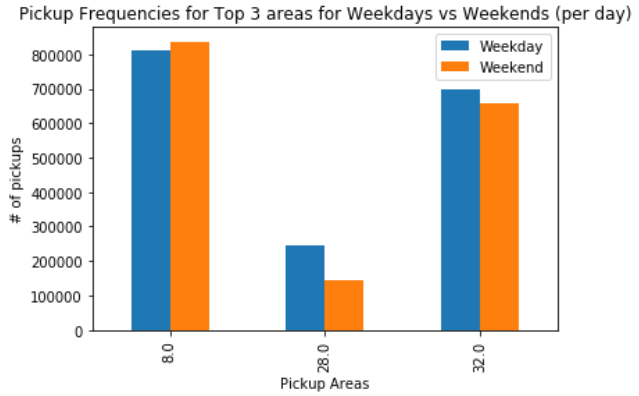

Figure 10: Pickup frequencies for top 3 community areas for weekdays and weekends

Table 5 in the following section shows the correlation between weekend vs. weekday and the ride count in the top 3 neighborhoods. No meaningful correlation was found between weekend vs weekday and the pickup frequency of taxi rides.

5.5.4 Holidays vs. Non-Holidays

Table 5 displays the correlation coefficients between the count of rides in the top 3 neighborhoods and holiday events, with NN indicating Near North Side, LP indicating Loop, and NW indicating Near West Side. The results did not indicate a correlation between general holiday events and an increase or decrease in pickup frequency in the top 3 community areas.

	day	weekend_count	weekday_count	holiday_count	count_NN	count_LP	count_NW
day	1.000000	0.007445	-0.007445	-0.067028	-0.431329	-0.429429	-0.446415
weekend_count	0.007445	1.000000	-1.000000	-0.196957	0.056967	-0.251339	-0.218524
weekday_count	-0.007445	-1.000000	1.000000	0.196957	-0.056967	0.251339	0.218524
holiday_count	-0.067028	-0.196957	0.196957	1.000000	-0.209802	-0.139587	-0.117693
count_NN	-0.431329	0.056967	-0.056967	-0.209802	1.000000	0.918059	0.928940
count_LP	-0.429429	-0.251339	0.251339	-0.139587	0.918059	1.000000	0.988334
count_NW	-0.446415	-0.218524	0.218524	-0.117693	0.928940	0.988334	1.000000

Table 5: Correlation between ride count, weekday, weekend, and holiday in top 3 community areas

5.5.5 Day of Week

Figure 11 below displays the pickup frequencies for each day of the week in the top three community areas. Frequency increases steadily throughout the week, peaking Thursday for area 32 and or Friday for 8 and 28. The weekend decline is sharp for area 32 especially, with around half of the pickups on a typical weekday.

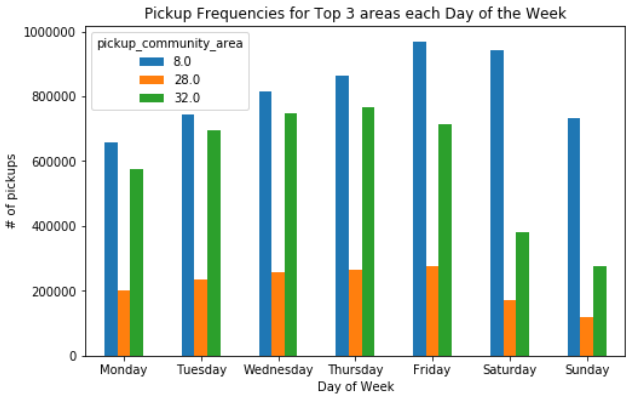

Figure 11: Pickup frequencies for top 3 community areas for each day of the week

5.5.6 Time Series Trends

5.5.6.1 Daily and holidays

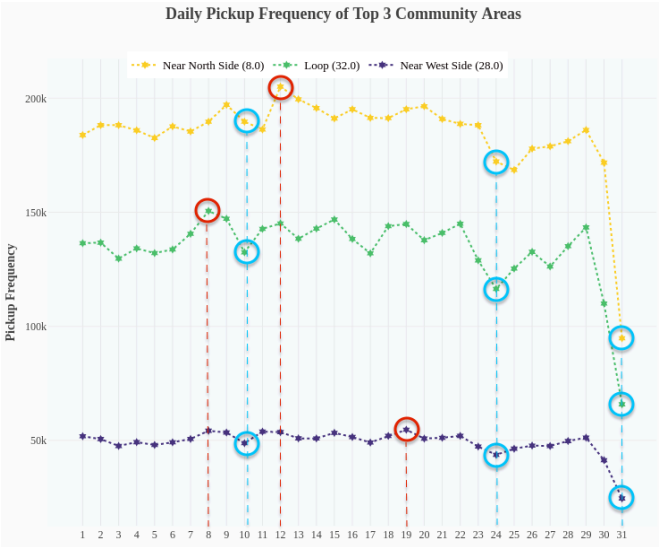

Figure 12: Time series analysis of ride frequency per day of the month

Holidays did seem to have some effect on ride frequency, but not much. The St. Patrick's Day

Parade on March 12 made the most obvious spike (at least in Near North Side), while other holidays like Thanksgiving and Christmas Eve seemed to coincide with a dip in rides. Certain events also correlated with increased ride frequencies, mainly in the spring.

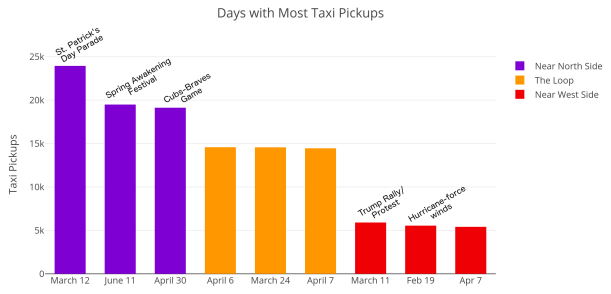

Figure 13: Days with highest ride frequencies in 2016

Events like the Spring Awakening Festival and the Trump Protest saw significantly high frequencies of taxi rides. Other things like Cubs games and weather events seemed to coincide with more rides as well. Aside from the St. Patrick's Day Parade, these events don't seem to have a huge effect on ride frequencies outside of the seasonal tendencies.

5.5.6.2 Hourly

To begin, a simple association study between tip amounts and communities was performed on the cleaned dataset. Since tips weren't recorded for cash transactions, rows with cash transactions were omitted. Outliers made the mean an unreliable method for looking at the average tip in each community area, so median values were used. Data based on pickup location vs. dropoff location tended to be pretty similar, but dropoff data was more complete (since some areas don't seem to have regular taxi pickups), so it was the primary source for this data.

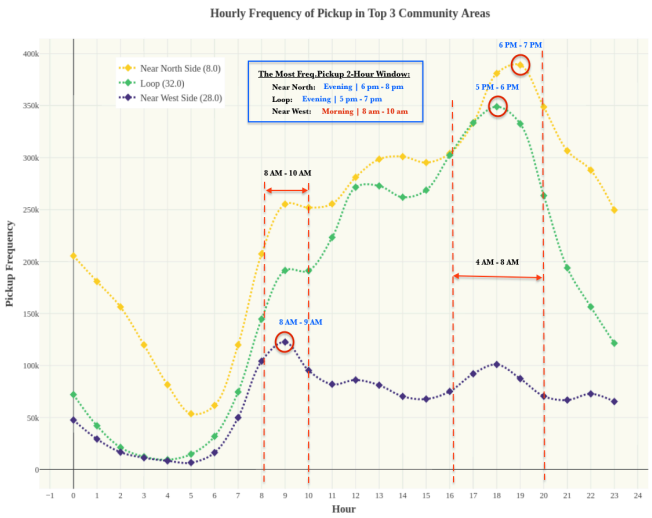

Figure 14: Chart of hourly time series analysis of top 3 community areas' ride frequencies

Looking at the top 3 community areas, we analyzed the taxi ride frequencies per hour (in a 24-hour period). All 3 areas have relative maximums at 8AM-10AM and 4PM-8PM. The two higher frequency areas (Near North Side and The Loop) peak in the evening and fall off until the morning. The Near West Side doesn't have as much of a peak, but does also fall off at night and comes back up in the morning.

6 APPLICATIONS

6.1 Most lucrative areas in Chicago

Determining the best areas in a city to serve as a taxi depends on many different factors and sits on multiple levels. Good areas for picking up customers can depend on things like general demand, customer generosity, and guaranteed fare. The term "area" can also mean different things. In the context of our research, it meant either Chicago's community areas or specific addresses in the city. We looked at most factors (best tips, predicted routes, etc.) at the community area level as this seemed the most reliable for generalization. We did, however, look at things like ride frequency at the specific address level as well.

As a whole, the results aren't unexpected – the downtown area has the highest ride frequencies.

There are some useful takeaways from these results that could give certain taxi drivers an edge over others. Obviously taxi trips to and from the O'Hare airport will yield high fares because of the relatively long ride distance to the downtown area. However, the tips for these rides, on average, are pretty moderate. The tips for rides in the downtown area seem to be generally higher than any other areas in Chicago, so staying in that area could be beneficial. Looking at the frequent item sets for routes, you can see that, of the 3 main downtown areas, Near North Side rides often go to O'Hare, whereas Near West Side rides most often go to The Loop and rides from The Loop frequently go to Near West Side. Since fares are a fixed value based on route, tips are the main variable in choosing where to work. Since the downtown area overall gives the best tips, it would be beneficial for a taxi driver to stay there as long as possible. So, working mainly in Near West Side and The Loop would be the best way to do this, whereas working in Near North Side would result in more trips to the airport, but also be more frequent.

As previously mentioned, we also looked at specific addresses' ride frequencies. The one location in The Loop, 10 S LaSalle St., averages more than 2,000 rides per day. It's also located in The Loop, so rides are more likely to remain in The Loop/Near West Side area. Since rides are so frequent in this area, however, it could become congested and taxi drivers might elect to target more spread out areas. Near North Side has a few highly frequent ride areas that we identified, but they're not as concentrated as The Loop, so they could be more effective in that regard.

6.2 Best times to target taxi pickups

Another important, and perhaps less controllable, factor in frequency of taxi rides is time, both hourly and monthly. Obviously, taxis will operate year round and all day, knowing peak times and days could give certain taxi drivers an edge over others. We looked at the relation of time to ride frequency in the top 3 community areas.

On a monthly scale, spring is the clear peak period for taxi rides with a semi gradual decline after June. On a weekly scale, rides peak on Friday and drop off on the weekend. On a daily scale, Festivals, Cubs games, and the St. Patrick's Day Parade seem to coincide with more frequent rides, while holidays later in the year (Thanksgiving, Christmas) seem to coincide with less frequent rides. On an hourly scale, rides tend to peak in the 4PM-8PM block with a consistent bump in the 8AM-10AM block.

Taxi drivers could use this information to prioritize working certain days or even months. Tracking popular spring events and prioritizing taxi service around them seems to be a very beneficial practice. Avoiding working 11PM-6AM because of the huge dip in frequency also seems important. Weekly ride frequencies don't seem to fluctuate much, but prioritizing Friday rides could have some benefit.

REFERENCES

[1] Nelson/Nygaard Consulting. "Taxi Fare Rate Study." City of Chicago, www.cityofchicago.org/content/dam/city/depts/bacp/publicvehicleinfo/publicchauffer/chicagotaxifaresstudyaug2014.pdf.

[2] Wu, Yiming. " 2016 Chicago Cabs Analysis." NYC Data Science Academy, 18 Sept. 2017, nycdatascience.com/blog/r/2016-chicago-cabs-analysis/.

[3] Schneider, Todd W. "Chicago's Public Taxi Data." Todd W. Schneider, toddwschneider.com/posts/chicago-taxi-data

[4] City of Chicago, "Chicago Taxi Rides 2016." City of Chicago, https://www.kaggle.com/chicago/chicago-taxi-rides-2016

[5] City of Chicago, "Chicago Taxi Dataset Released." City of Chicago, 16 Nov. 2016, https://digital.cityofchicago.org/index.php/chicago-taxi-data-released/

[6] City of Chicago, "Census Data – Selected socioeconomic indicators in Chicago." City of Chicago, https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2/data

[7] AreaVibes, "Near North Side Average Weather", http://www.areavibes.com/chicago-il/near+north+side/weather/
"Loop Average Weather", http://www.areavibes.com/chicago-il/loop/weather/

"Near West Side Average Weather",
http://www.areavibes.com/chicago-il/near+west+side/weather/

[8] Wikipedias, "Chicago Loop",
https://en.wikipedia.org/wiki/Chicago_Loop