STAT 689
Project
Report

Chicago Taxi Rides – To Tip or not to Tip?

Team Members:

Paul, Debapriyo (UIN: 525004584)

Tangadpalliwar, Abhilash (UIN: 825009793)

Table of Contents

Introduction and Data Description

The City of Chicago in November of 2016 released a public dataset [1] containing information over 100 million taxi rides since 2013. This public dataset does not include any data from the rideshare services like Uber and Lyft, but in 2015, the taxi-owners association of Chicago claimed that Uber and Lyft have caused them a loss of 30-40% in business [2].

The dataset contains the information about the trips taken in Chicago taxis. Each row in the dataset is representative of a distinct taxi trip and has information about the following fields: [1]

- Which taxi provided the trip
- What times the trip started and ended
- Length of the trip in both time and distance
- Starting and ending Community Area — plus Census Tract for many trips
- Fare amount and other components of the trip cost
- Type of payment — such as cash or credit card. (As an important note, cash tips are not included in the data because they do not go through the payment systems.)
- Taxi company

There are some assumptions and data masking in the dataset provided by the City of Chicago. They are as follows (as listed on their website):

Delay
Taxi trips are not reported in real time. Each trip appears long after the completion of the ride. Therefore, the dataset cannot be used to track trips in motion or even just-completed trips. By the nature of how the data are collected, reported, and processed, a minimum of a few days will pass between completion of a ride and its appearance in the dataset. More typically, the delay will be anywhere from a week to a month.

Masking of Taxi Medallion Number
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.

Masking of Time
It is anticipated that analysis of taxi trips by time will be a major use of this dataset and it will add significant value for understanding the taxi industry and travel in Chicago. However, there is minimal value and some potential privacy cost in making it possible to find a specific trip that someone knows departed at 10:13 am. To balance these issues, all start, and end times have been rounded to the nearest 15 minutes.

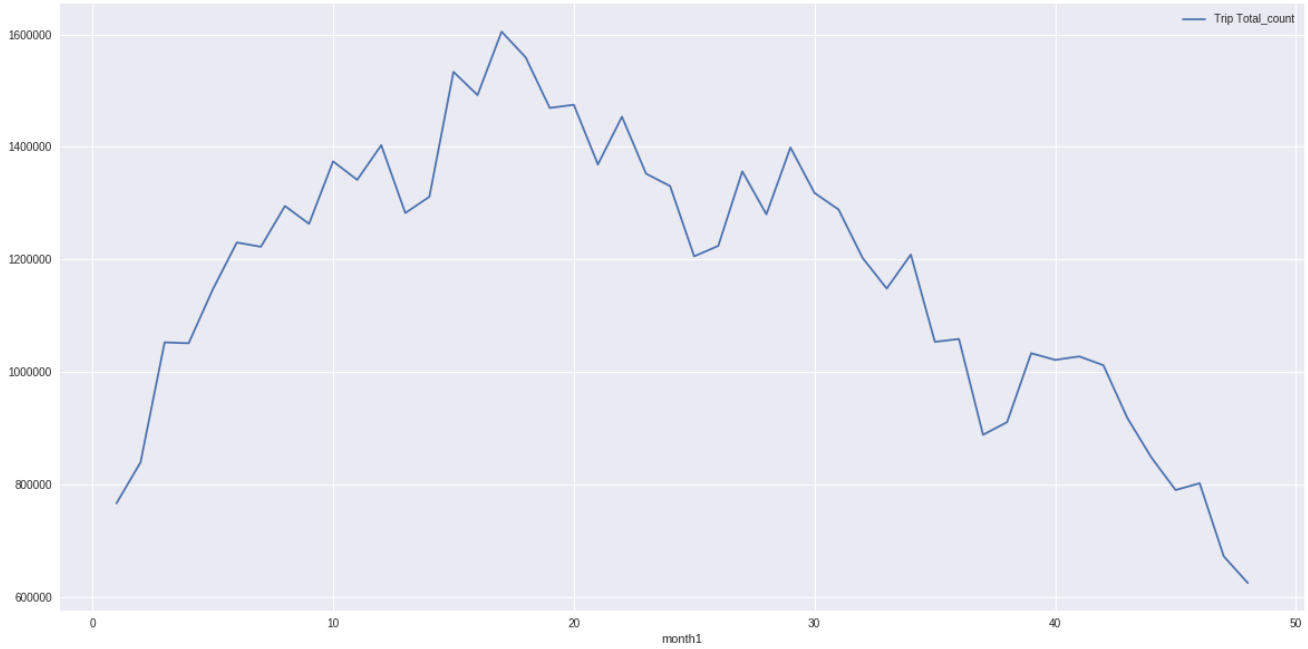
Masking of Location
From where and to where people travel is, of course, the most basic information about taxi trips and expected to be the topic of much analysis. However, as with exact time, exact location down to the street address could affect privacy. Therefore, the location is only provided at the Census Tract and Community Area levels.

Data Cleaning and Preprocessing

1. The data for the time period 2013-2017 used for the analysis is very large in size (~40 GB).
2. In order to store this large data, we created an SQLite3 database using chunks in Python. We loaded individual year's data and while loading each chunk, the data was processed. We then created a column "Month" in each year's data and using this column, we created 12 tables in SQL database for each year (representing 12 months of the year). Using IF function, we found the "month" information in the chunks of data, filtered and stored the data in respective table in the database. Thus, each database contained 12 tables representing 12 months of that year and there were 4 such databases for the years 2013-2016.
3. We then converted each table in SQL database into csv format by querying in Python.
4. We thus had 48 months of data available (for 4 years) and we used that to find trends from it, which has been described in the later stages of this report.
5. Further processed data for respective objectives (mentioned later)
6. Used Google Colaboratory to load the data and perform the analysis

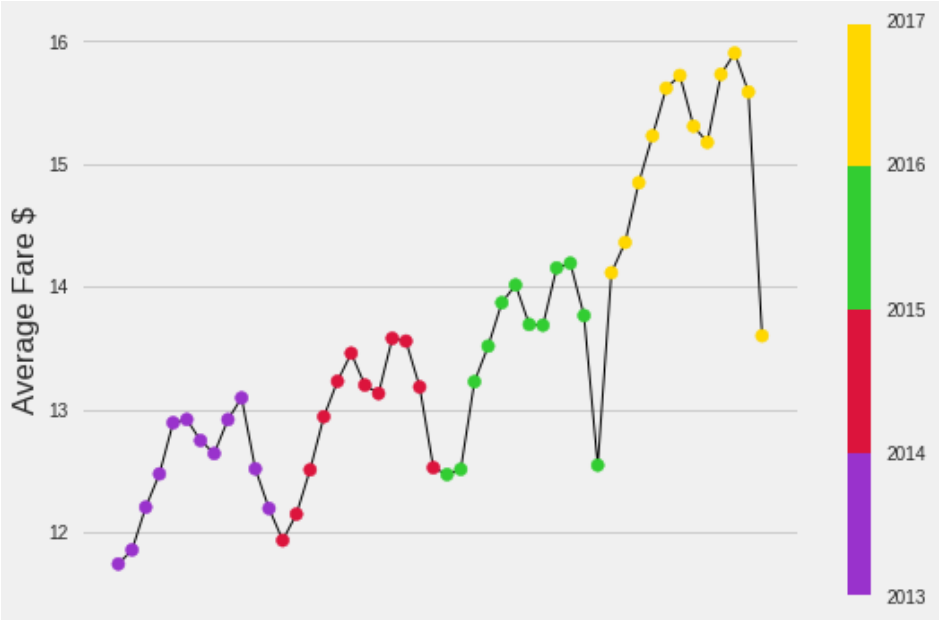
Data Exploration

Month-wise total trips for 2013-2016



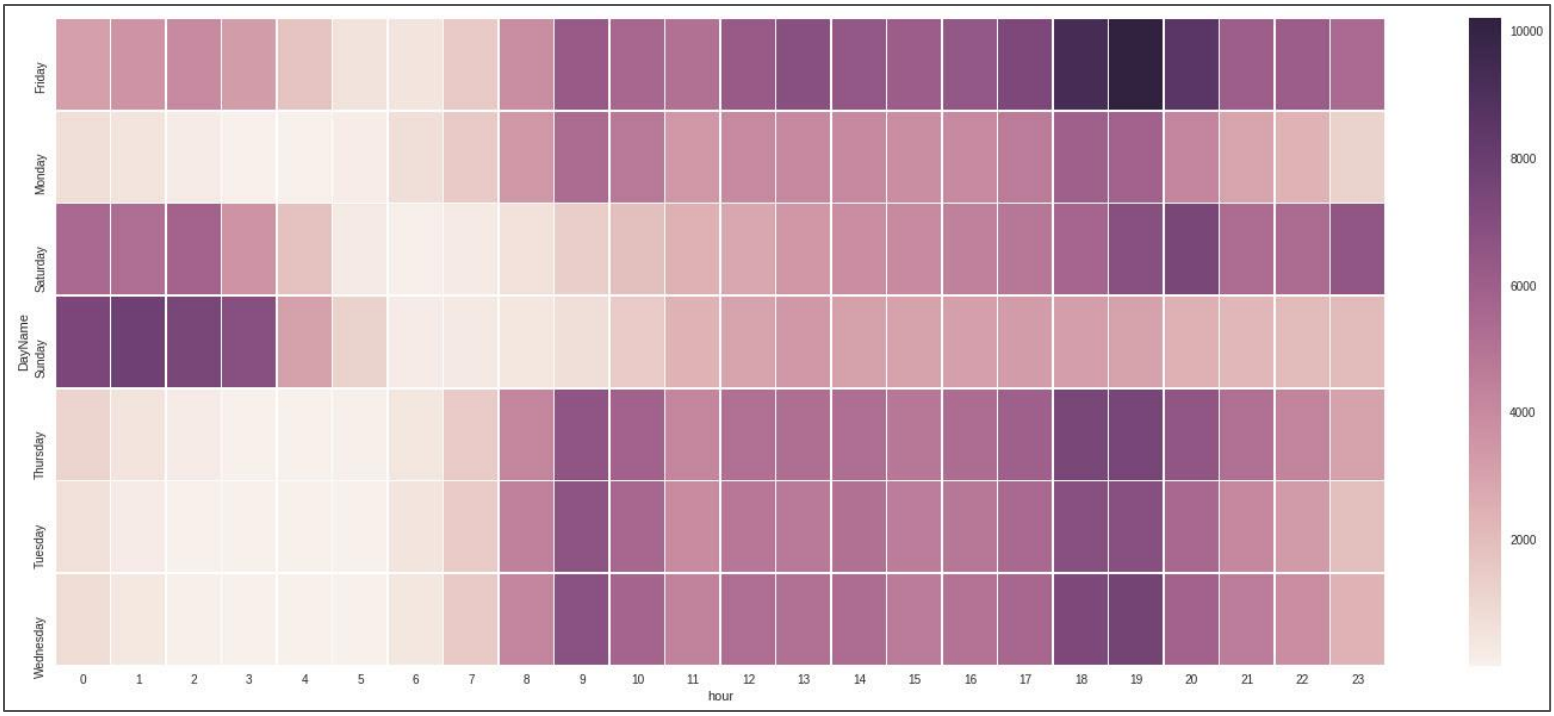
The above graph represents the total trips of the Chicago taxi industry over the period of 48 months (Jan 2013 – Dec 2016). The X-axis represents the month (1 is Jan 2013, 48 is Dec 2016). The Y-axis represents the number of trips. From 2014, we can see that the trips of the taxi industry have been declining which is mostly attributed to the emergence of Uber and Lyft in the Chicago area over the years. As mentioned earlier, Chicago taxi owners claim to have lost around 35%-40% of their business owing to Uber and Lyft, which is pretty much evident from the graph above.

Month-wise average trip fare from 2013-2016



The above graph represents the average fare of the Chicago taxi rides over the period of 48 months (Jan 2013 – Dec 2016). The X-axis represents the month. The Y-axis represents the average fare which is essentially sum of total fares/sum of trips. Since the total trips have decreased over the years, it is evident that the average fare has been showing an increasing trend to compensate with the decreasing rides. There is a seasonal and cyclic pattern that exists with the average fare. It drops during the summer months of June and July as well as December whereas in the months of April/May and October/November, it is relatively higher. One of the reasons behind this might be the fact that in the summer months and December holidays, the City of Chicago has a lot of tourists particularly in the downtown area who tend to use taxis for the shorter distances in and around the downtown thereby reducing the average trip cost.

Hour and Day-wise trips for a typical week



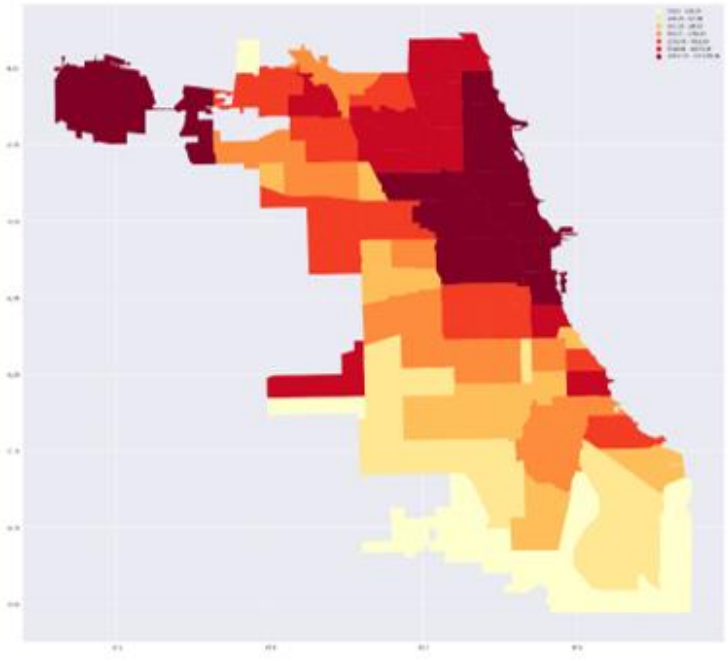
We have tried to create a heatmap of the typical number of rides based on the day of the week and the hour of the day. In the above graph, the Y-axis represents the days of the week, and the X-axis represents the hour of the day. The darker the shade in the heatmap, the higher is the number of rides in that particular hour.

Based on our conclusions from the hourly analysis and the daily analysis in the previous 2 graphs, this graph provides a combined picture of the 2 analyses. As we can see, the highest taxi rides take place between 6-8 pm on a Friday. On other weekdays as well, the number of taxi rides is high in the morning hours (8-10 am) as well as in the evening hours (6-8 pm) which is because of people taking taxis for workplaces.

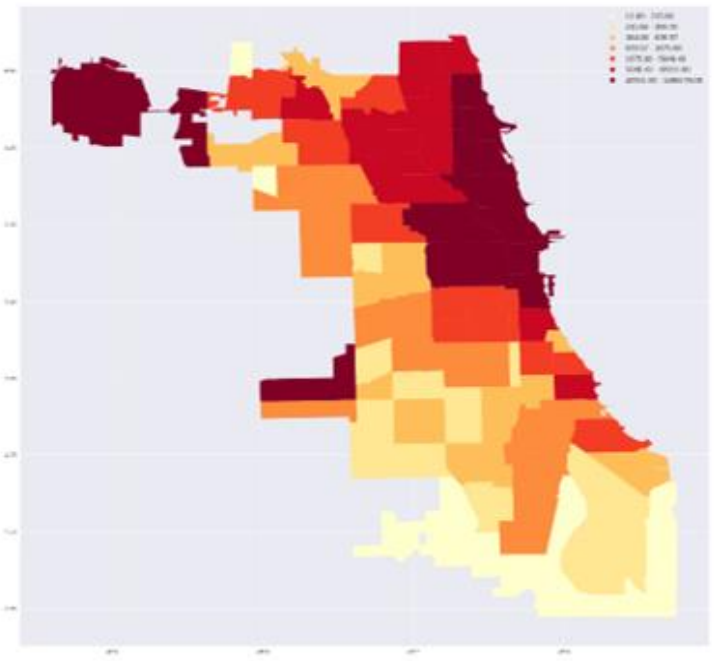
Another interesting observation comes from the weekends (Fridays, Saturdays, and Sundays) during the early hours (12 am - 3 am). The number of rides for these days during this particular time period is very high as compared to other weekdays. This can be attributed to the fact that many youngsters go out on weekends particularly to downtown/entertainment places for partying etc. and then take taxis back home late in the night.

Please refer to appendix for individual distributions.

Community area-wise pickup comparison (2013 v/s 2016)



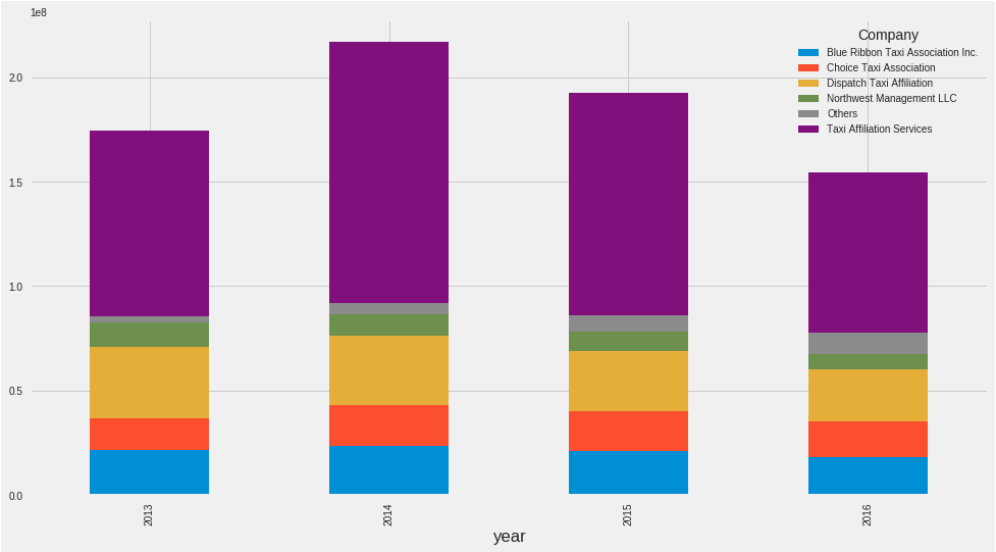
2013 Q1



2016 Q1

As we can see from the above two heatmaps, the overall pickups in the 2016 heatmap is less concentrated than the one in 2013 heatmap (lighter shades for many community areas). This corroborates our previous argument that the overall rides have decreased. We have included similar heatmaps for the other 3 quarters in the appendix.

Market Share for Taxi Companies over the years (2013-2016)



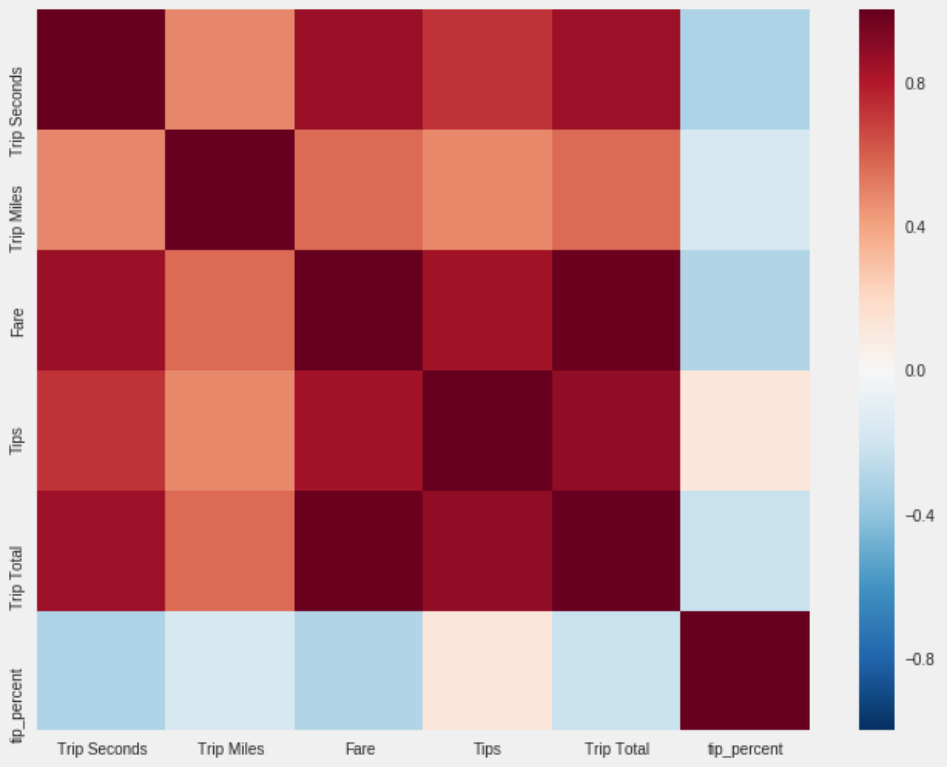
Model Building

Models for predicting Fares

1. Regression Model

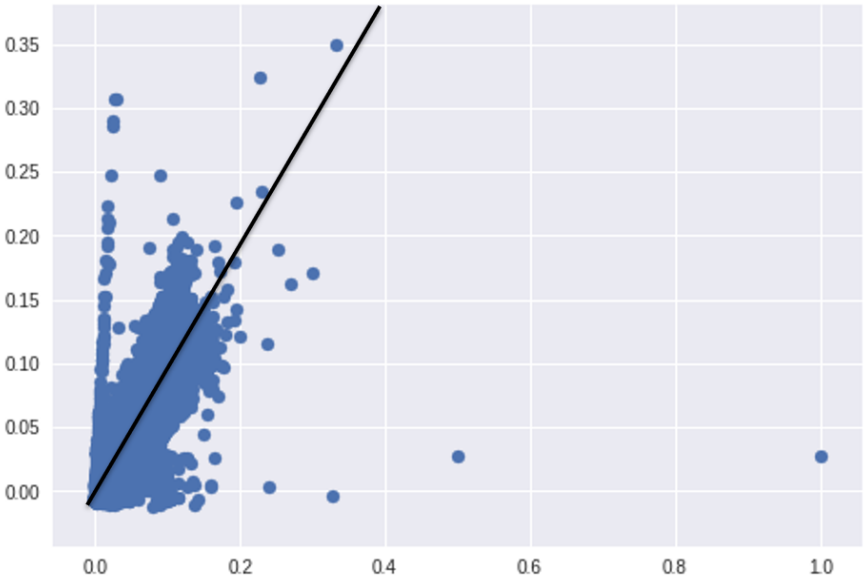
After exploring the data, we proceeded to build the prediction models. We started with the simple regression model to predict the trip fares in order to see the factors affecting the trip fares and their correlation.

The below heatmap shows the linear correlation of different fields:



As we can see, the fares seem to be linearly correlated with Trip Duration and Trip Miles which sounds obvious. But we also see how the Tip Percent has a very poor linear correlation with these fields and so we have used Random Forest classifier in the later stages to predict this “Tip %”.

As far as our Regression model for predicting fares is concerned, after running the model for 1-year worth of data, these were the results obtained:



Regression Equation

$y = 0.4678*(trip\_seconds) + 0.2904*(trip\_miles) + 0.0214*(pickup\_community) + 0.0152*(dropff\_community) + 0.0015*(day\_name) - 0.0026*(hour)$

$R^2 = 81\%$

2. Random Forest Model

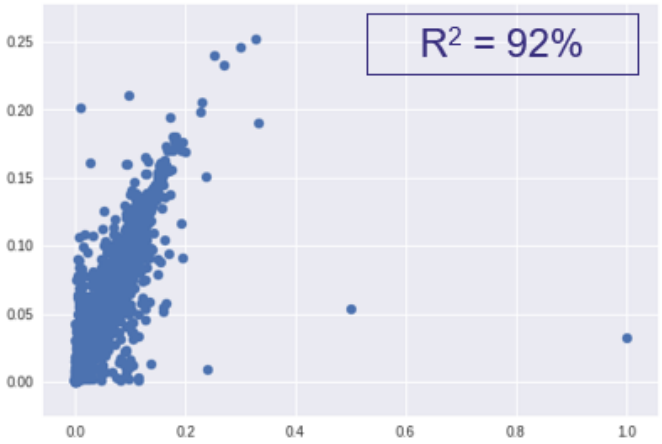
We then applied a tuned random forest regressor model to predict fares in order to see if we could improve the R^2 value of 0.81.

The parameters that we varied for tuning were as follows:

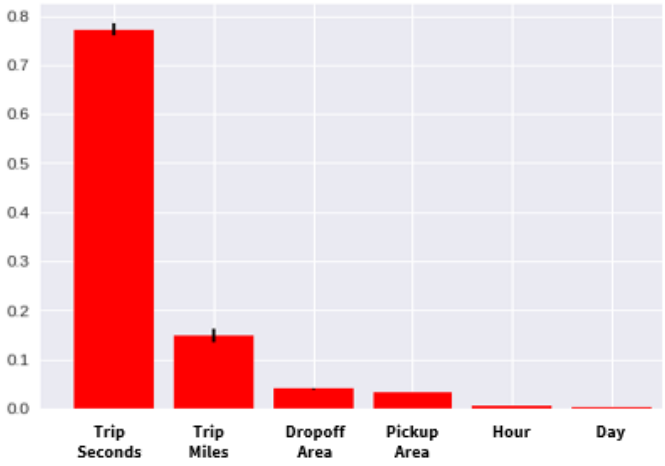
Parameters to vary:

- n_estimators = max no of trees
- Number of features to consider at every split
- Maximum number of levels in tree
- Minimum number of samples required to split a node
- Minimum number of samples required at each leaf node
- Method of selecting samples for training each tree
- Tuning performed on over 4000 settings

The results obtained from the tuned RF model are as follows:



Scatterplot for Predicted V/s Actual Responses (Normalized)



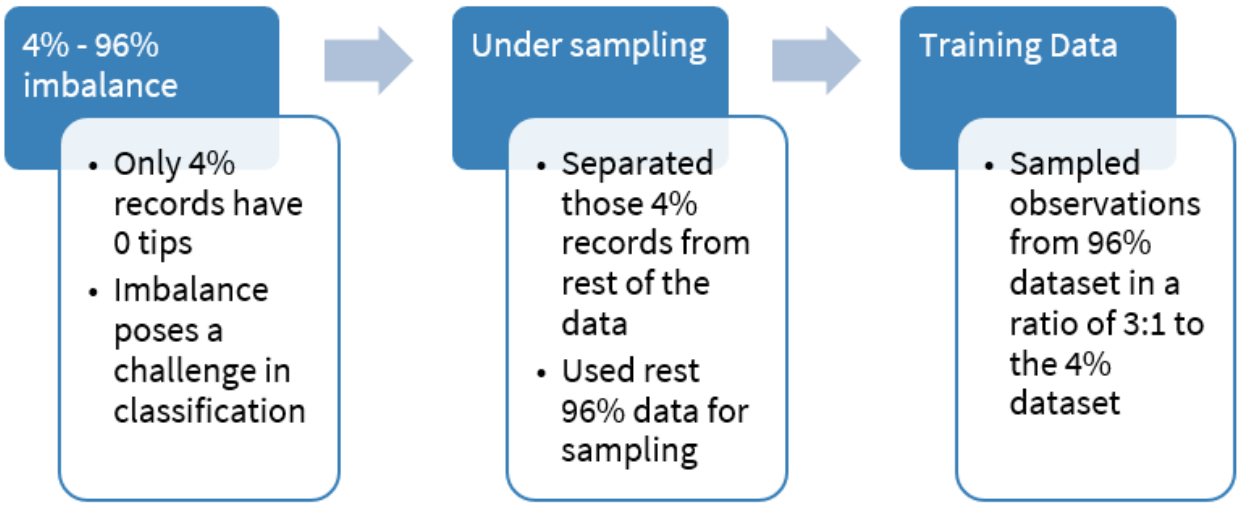
Feature Importance (Tuned Random Forest)

As we can see from the R^2 value, the latter model performs a lot better than the regression model and hence seems like a better choice for predicting fares. The result also includes the feature importance chart which tells us the factors that affect the fares with their relative importance. As mentioned earlier, the trip seconds and trip miles are the 2-biggest factors affecting the fares along with pickup and drop-off community areas. Along with that, surprisingly we have the hour of the day and the day of the week affecting the fares as well. For example, a person traveling from point A to B on two different days and two different times will pay different fares although the trip distance is similar.

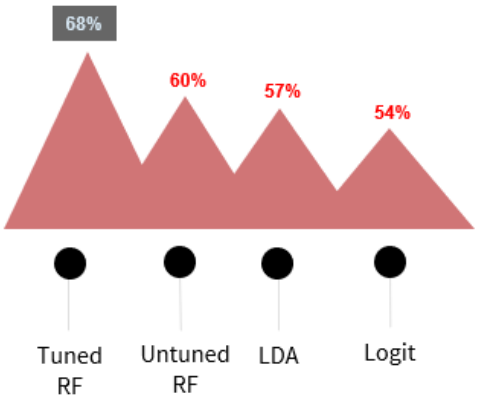
Models for predicting tip classifier

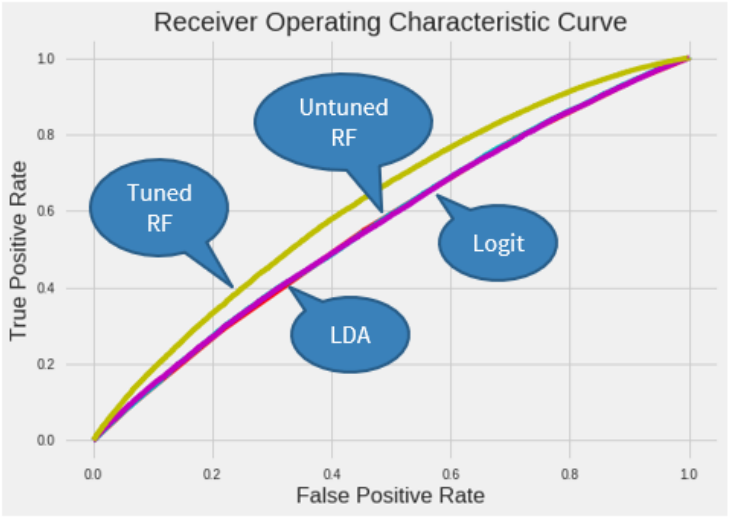
As mentioned in the previous section, the second type of models that we built was for predicting whether a person will tip or not; essentially a classification model.

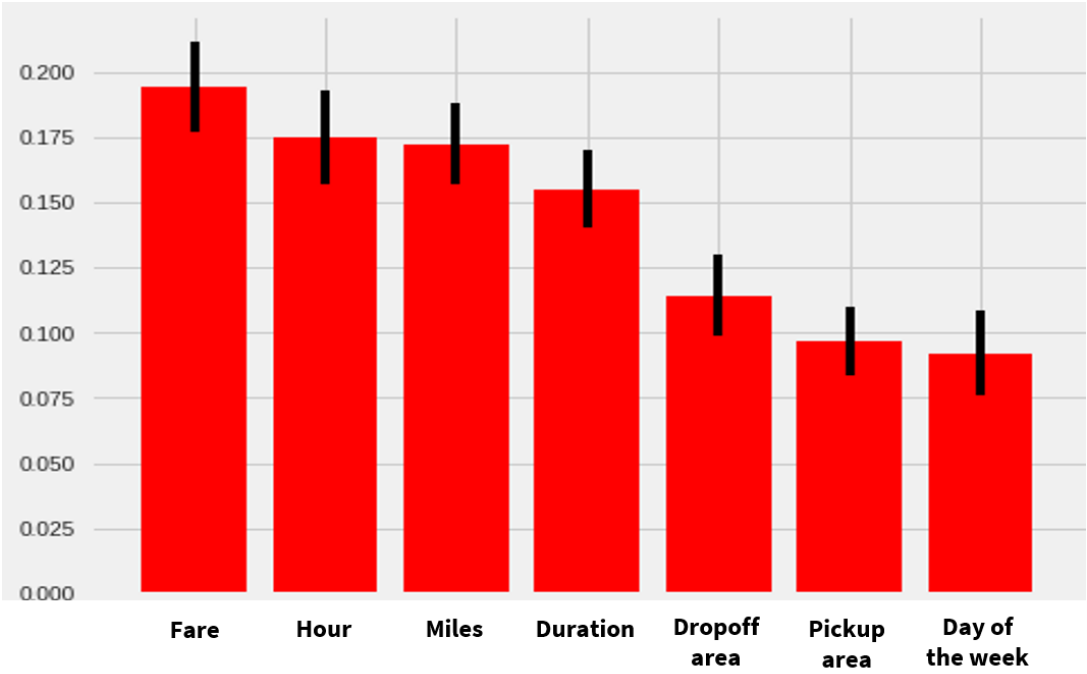
In order to do that, the biggest challenge that we faced was of a data imbalance. Here is how we eliminated the imbalance:



After doing so, we applied 4 different classification models in order to predict whether a person will tip or not, and if yes, what factors will that decision depend on. Here is the summary of results obtained:







As we can see in the above feature importance chart for the tuned random forest model which is our best classification model for the second objective, the decision to tip or not depends majorly on the fare. Surprisingly, hour of the day plays an important role in the decision as well. Other factors include miles, trip duration, drop-off and pickup area and one more surprising factor is the day of the week.

Thus, we can conclude our modeling with the aforementioned results and interpretations.

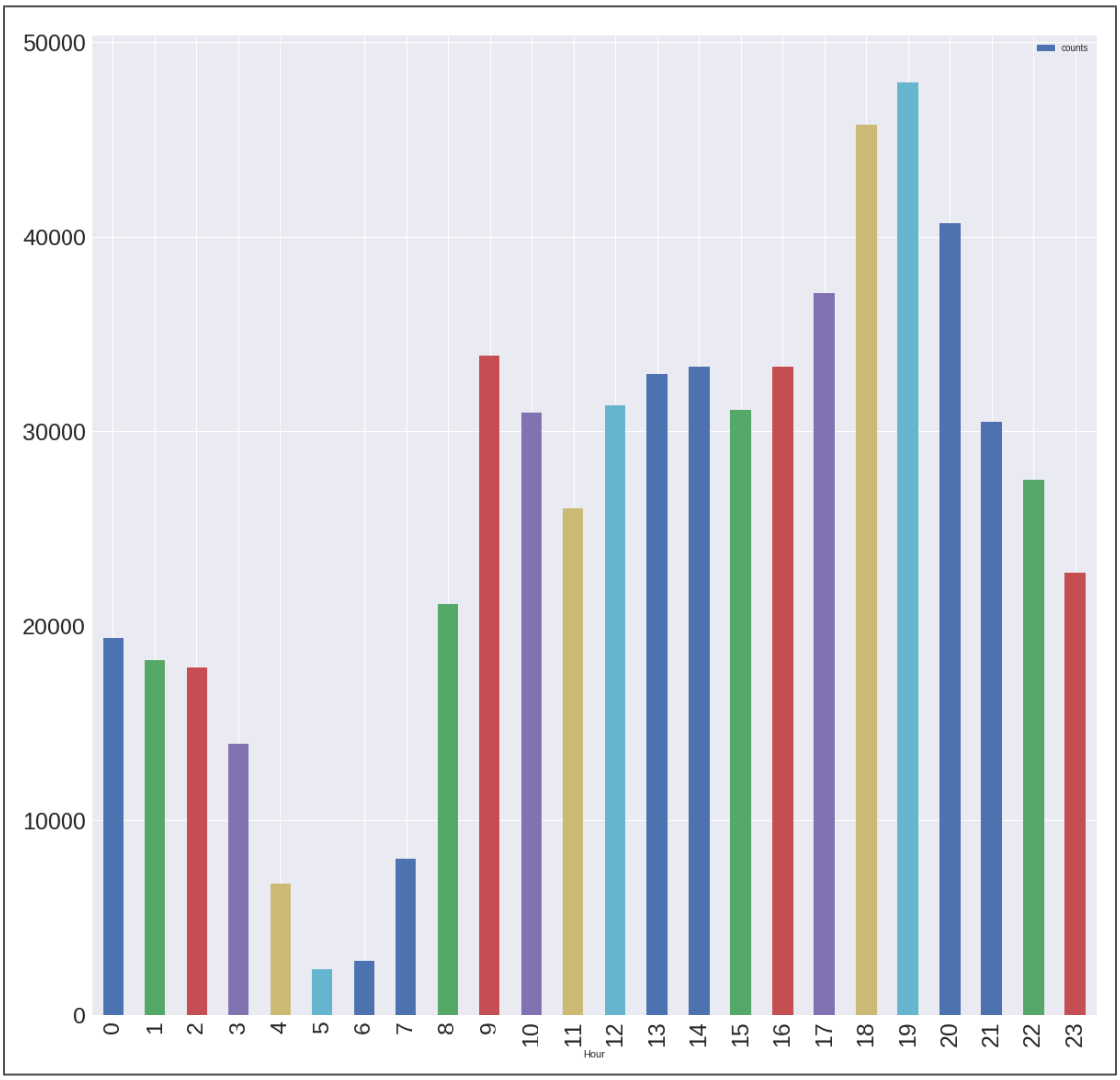
Please refer to appendix for more charts and information

Citations

1. https://digital.cityofchicago.org/index.php/chicago-taxi-data-released/
2. http://www.businessofapps.com/data/uber-statistics
3. http://chicagoist.com/2015/10/07/cab_drivers_plan_24_hour_strike.php
4. Ghoshal G., Paul D., Tangadpalliwar A. – TAMIDS Data Science Competition 2018 Report

Appendix

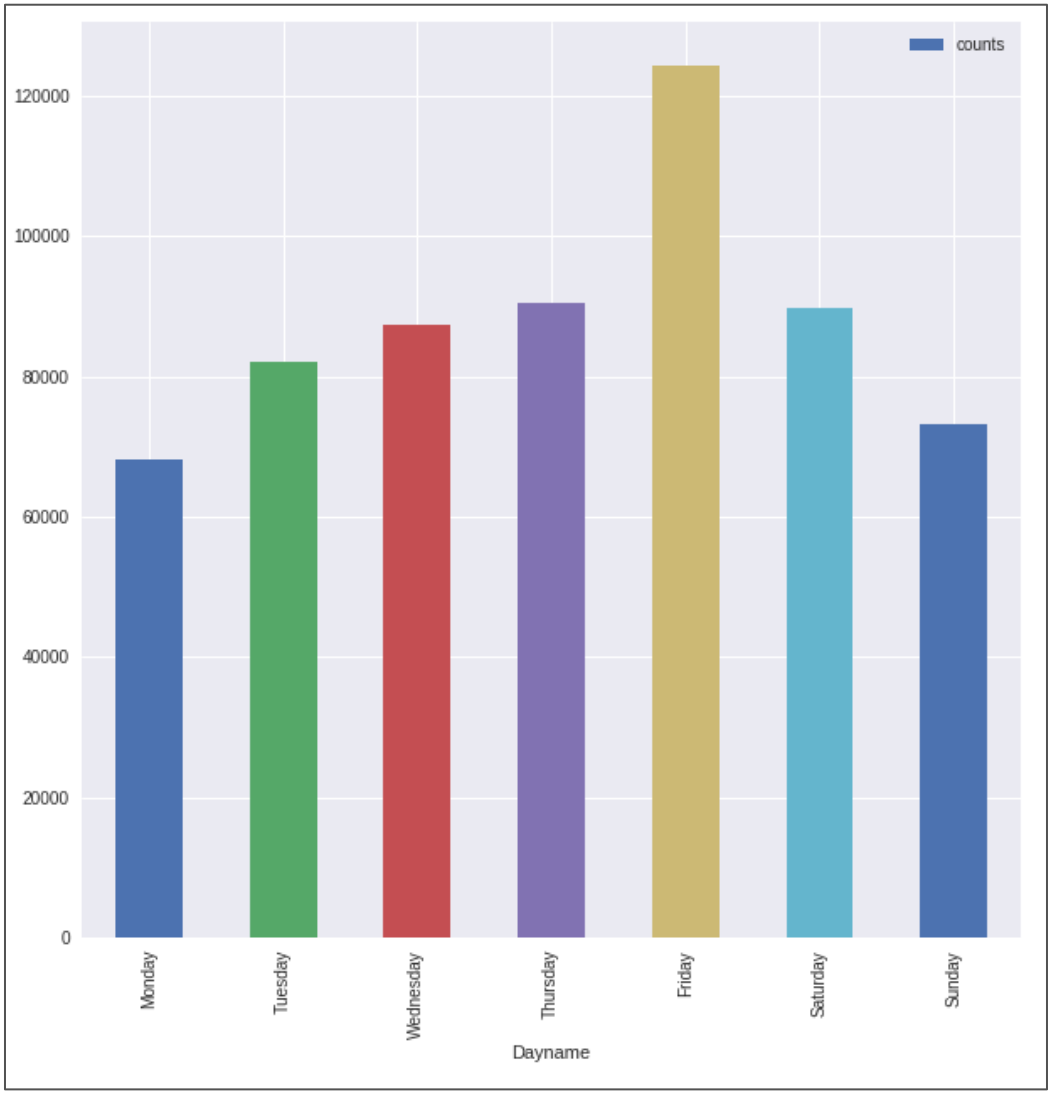
Hour-wise trips for a typical day



From the above graph, we can see how the hourly number of rides in Chicago look on a typical day. On the Y-axis, we have the total rides; and on the X-axis, we have the information about the hour of the day. We can clearly see that in the midnight, there aren't a lot of rides being taken and they drop till about 6 am, which is somewhat obvious since very few people take taxis at night. The number of rides start increasing from 6 am (which is when many people start traveling to work), and they keep on increasing till about 10 am. This is the time when majority of the office-going population gets to their workplaces.

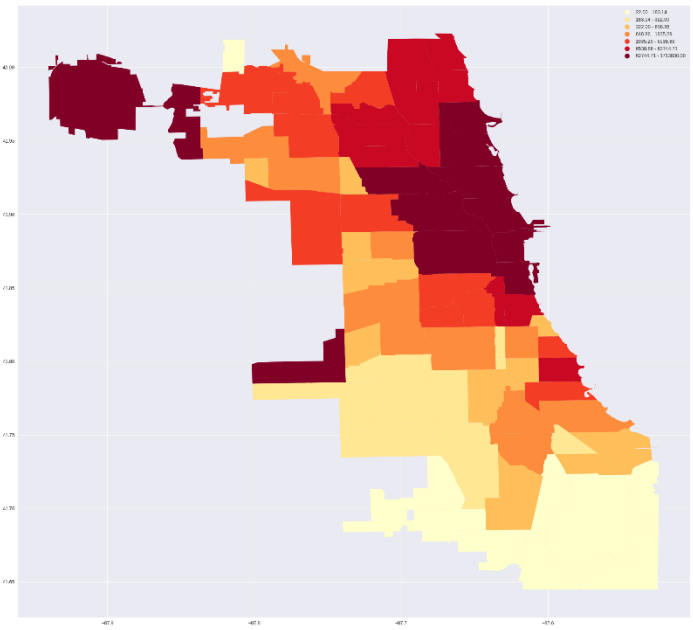
The number of rides is steady in the afternoon hours and it starts picking up again from evening (around 5 pm) and drastically increases and hits the peak between 6-8 pm. This is the time when many of the officegoers travel back home and also many youngsters and other citizens head out of their homes to take dinner or for leisure, which explains this peak. The number of rides then decreases and keeps on decreasing as the night progresses which is also self-explanatory as mentioned in the earlier paragraph.

Day-wise trips for a typical week

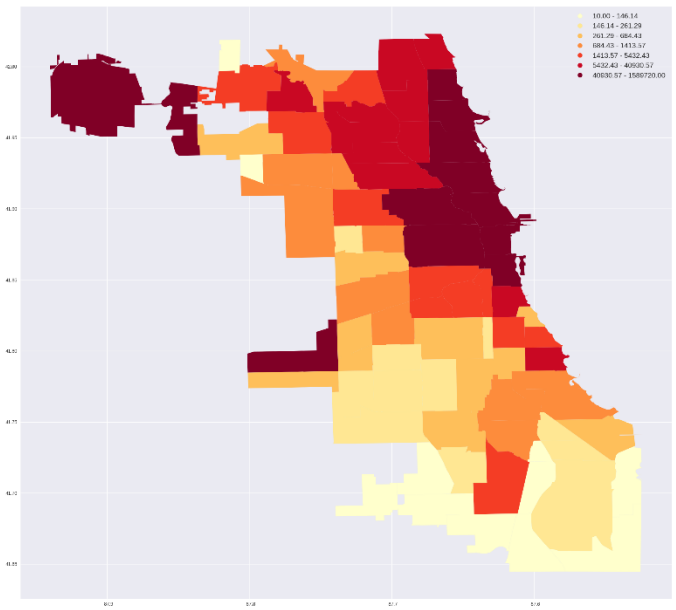


The above graph represents how a typical week looks like in terms of the number of rides. The X-axis represents the days of the week, and the Y-axis represents the total rides for that particular day. As the week progresses, the number of taxi rides increases steadily and there is a drastic jump in the rides on Friday. One of the possible reasons for this can be the increase in the number of people going out on Fridays either to Downtown, or for a movie, or entertainment, etc. There are many people who travel back home on Friday for the weekend, so there is increased number of rides to the airport/train stations/bus stations, etc. which also explains why there is a spike in the rides on Fridays typically.

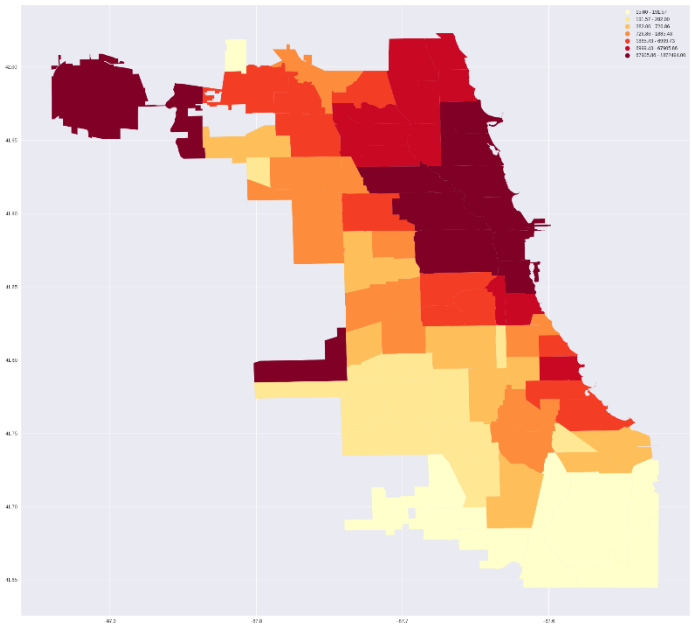
Community area-wise pickup comparison (2013 v/s 2016 – Other 3 Quarters)



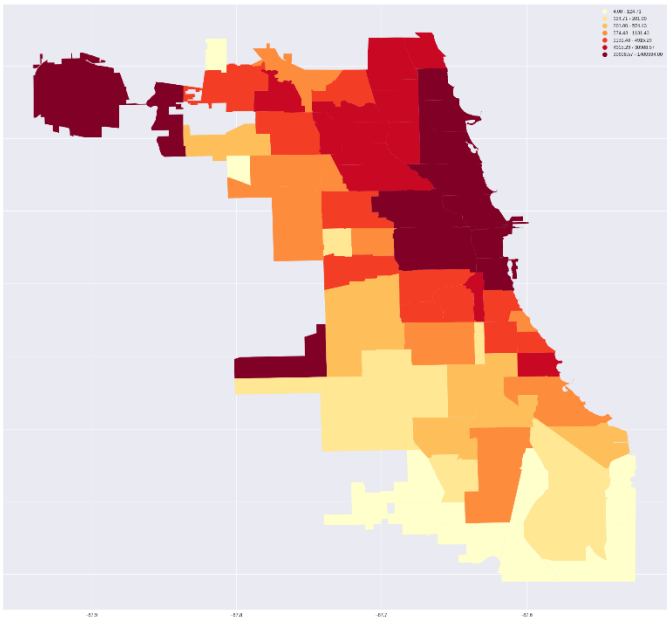
Q2 – 2013



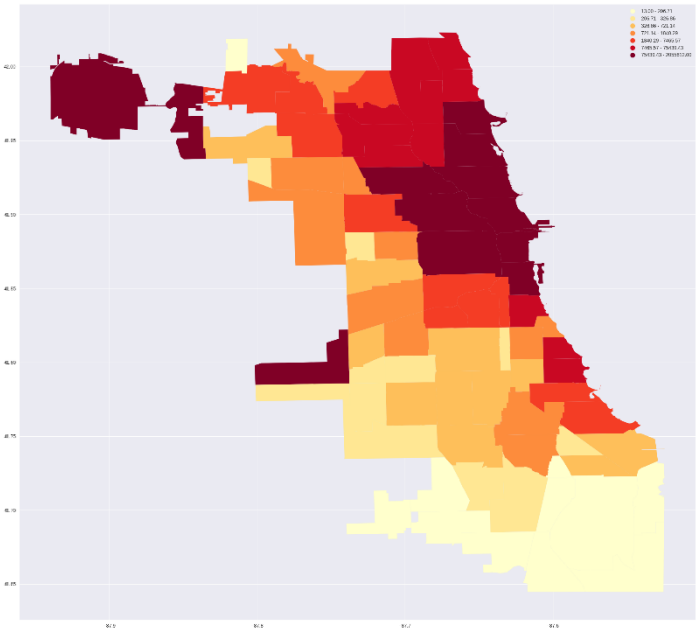
Q2 - 2016



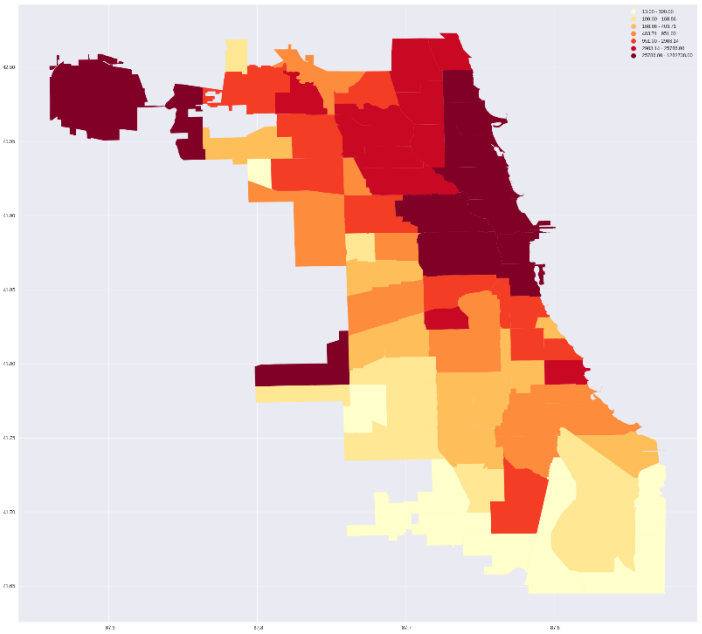
Q3 – 2013



Q3 - 2016



Q4 – 2013



Q4 - 2016

Histogram of Tip % values