GROUP 5 - TLC TAXI DATASET: A comprehensive view of vehicle for hire data in NYC from 2009 until present

UPDATE - Changes from Part 2 include Section 7 (Results) and additions to Section 8 (Milestones)

JOE FROELICHER, CSCI Graduate Student, Denver

TOMMY GUESS, CSCI Graduate Student, Lafayette

MIKE HUFFMAN, APPM Graduate Student, Arvada

ABSTRACT: Group 5 has chosen to study the TLC Trip Data Record, a complete record of taxi usage from 2009 until present. We plan to mine this robust dataset, and learn about the geography, economics and effects of the COVID-19 pandemic on NYC's "Vehicles for Hire".

Additional Key Words and Phrases: data, taxis, transit, NYC, visualization

ACM Reference Format:

Joe Froelicher, Tommy Guess, and Mike Huffman. 2021. GROUP 5 - TLC TAXI DATASET: A comprehensive view of vehicle for hire data in NYC from 2009 until present: UPDATE - Changes from Part 2 include Section 7 (Results) and additions to Section 8 (Milestones). 1, 1 (July 2021), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

Today, over half of the world's citizens live in urban areas. As the global population grows to a projected 11.2 billion people by 2050, that figure is expected to increase to 70%. As our urban centers continue to grow exponentially, they will experience monumental energy, environmental and infrastructure challenges. Understanding the movement of human beings through metropolitan areas will be an important area of study for decades to come.

The New York City Taxi Limousine Commission, through partnership with authorized technology providers, has made available the entirety of taxi cab records for all rides in NYC since 2009. This staggeringly complete dataset of billions of rides provides data for each pickup including: pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

Our group choose this dataset primarily because of our shared interest in visualization. But in addition to the rich visualization opportunities, there are important facts to be learned about the differences between yellow cabs, green cabs and ride-sharing vehicles, the factors that predict tipping outcomes and the effects of COVID-19 on vehicle for hire commuting.

Authors' addresses: Joe Froelicher, CSCI Graduate Student, Denver, jofr1275@colorado.edu; Tommy Guess, CSCI Graduate Student, Lafayette, Tray.Guess@Colorado.edu; Mike Huffman, APPM Graduate Student, Arvada, michael.huffman@colorado.edu.

Manuscript submitted to ACM

2 LITERATURE SURVEY

The TLC Trip Data Record is a robust dataset of an important metro area spanning from 2009 until present, and so has been the subject of much previous work. We review some of that work here:

Kaggle, an online community of data scientists and machine learning practitioners, has made the TLC data set central to one of their popular competitions. In the challenge "New York City Taxi Trip Duration" which closed in 2017, participants were tasked with developing a predictive model to estimate total ride duration of taxi trips in NYC. The winning team produced a model with a RMSE of 0.28976.

Tseng and Chau used the TLC dataset to assess the viability of electric vehicles (EVs) as taxis. As the world moves towards a post-carbon future, EVs will play an increasingly important role in human transport. However, because of their limited range per charge, EVs have still not seen wide adoption in logistics and vehicle fleet roles. To study whether EVs would be feasible as taxis in NYC, Tseng and Chau used Markov decision processes to model the taxi service strategy. They found that EVs would be financially viable, and identified a minimum battery capacity to compete with internal combustion engines (45 kWh).

Wickramasinghe et al. applied supervised machine learning techniques to the TLC taxi data in order to predict the volume of taxi rides in any given hour. Their results suggest the random forest regression was a valuable tool across all zones in the city. They proposed additional research topics including: planning evacuation routes for possible disasters, getting general population counts in given locations at given times, and identifying 'hot spots' in a city.

In "Spatial Equilibrium, Search Frictions and Dynamic Efficiency in the Taxi Industry", Buchholz sought to identify inefficiencies (i.e. vehicle misallocation) in NYC by studying the TLC dataset. He imposed a dynamic model of spatial search and matching to identify mismatches in taxis and riders. His findings suggested large inefficiencies with significant economic implications.

3 PROPOSED WORK

A large advantage of a dataset collected automatically (such as this one) is that it typically requires little cleaning, as it's far less vulnerable to human error. Preliminary investigations have not yielded any examples of incomplete entries, incorrectly formatted dates/times, or other examples of dirty data. Thus, we expect the data cleaning phase to be fairly straightforward. Some simple computation ahead of time will be useful, such as determining the amount of time each trip took using subtraction. To facilitate efficiency in future computation, we plan on grouping the data by a timestep, such a minute or five minutes. This binning will make the data easier to utilize.

One question has to do with the factors that influence tipping. Is it possible to predict with any degree of accuracy what a tip may be, given the conditions of a completed trip? During a trip, the passenger has control over some factors, such as when they hail the cab and where they plan to go. However, the time it takes, the skill of the driver, and various other factors are not known until after the trip. A potential way to interpret the tip is as a proxy for customer satisfaction. This project may be able to see if this is indeed the case by examining trips of a certain class and seeing if improved metrics (such as faster times) yield higher tips. If that's not the case, then tipping may prove even more interesting, as it will be less intuitively obvious what purpose it serves. Regression may be sufficient to answer this question. If not, we will use a simple, 3-4 input neural network for this particular problem.

Another question has to do with the customer's choice between a yellow cab, green cab, or a For-Hire Vehicle (FHV). Green cabs are also known as boro taxis, and they tend to be used in the neighborhoods outside of lower Manhattan. FHVs include ridesharing services such as Uber and Lyft. We hope to discover which factors play a role in a customer

choosing one of these vehicle types over another. This amounts to a comparison between datasets, and such datamining is best accomplished using multivariate analysis. This is an intriguing area that tends to be avoided in the existing literature, yet nonetheless will be a prevalent topic given the increase of ridesharing services. From the perspective of a company, we may be able to discover which factors are important to a customer choosing which type of ride to take.

4 DATA SET

This project uses official datasets from New York City taxicab rides. All taxicab journeys represent a connection from one point in space (the pickup location) to another point in space (the dropoff location). The dataset notates these as zones, which split up New York City into approximately 250 distinct areas. These areas also include the neighborhood through the use of a simple lookup table provided by the source. Each taxi trip also includes a timestamp for the pickup and dropoff, as well as various data regarding the payment, surcharges, and tips. This dataset seems to lend itself nicely to questions regarding the interaction between these factors.

The dataset is hosted at: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

5 EVALUATION METHODS

One of our evaluation tools for this project is data visualization. One of the questions we are most interested in is how the impact of COVID-19 affected taxi travel. Therefore, visualization of taxi data via correlations plots, and maps will be used to validate our analysis. We will be visualizing the volume of taxi traffic, as well as visualizing the time series of data for taxi travel.

In addition, we will be using our data to train a prediction model, and in turn the validation for that prediction model will be done with our data. The prediction model ideally will predict cost of ride, or ride destination based on a set of parameters. We will then use the error rate to evaluate our prediction model. We also will visualize the prediction model based on the data vs. the predictions.

6 TOOLS

The primary tool we will be using for this analysis is python 3.8, and multiple associated packages including: Scikit-Learn, Numpy, Pandas, Scipy, Plotly and Matplotlib. Numpy and Pandas are powerful tools that will allow us to efficiently read and perform analysis on large datasets. Those will be how we store and operate on our taxi data.

Scipy and Scikit-Learn are our primary analysis tools. Both of these packages contain a plethora of algorithms for statistical learning, optimization routines, and much more. In combination with Numpy and Pandas, this should cover the majority of the “backend” of this project.

In addition we will be using a few tools for visualization of our data. The first is Matplotlib, another package in python. And second is a package built for R, called ggmap. Depending on how this mapping goes, we may also employ Plotly. This is how we will visualize the map data. Finally, our project will use the industry standard "Github.com", and the associated git software to do version control for our project. Additional packages may be added as needed, particularly for use in R or python.

7 RESULTS (ON GOING)

7.1 Exploratory Work

For a number of our studies, we choose to control for time and picked a pre-COVID timeframe. For such areas of study, we decided on 2019. Below is the summary of pick-ups and drop-offs of Yellowcabs in 2019 (total: 8.3 million rides).

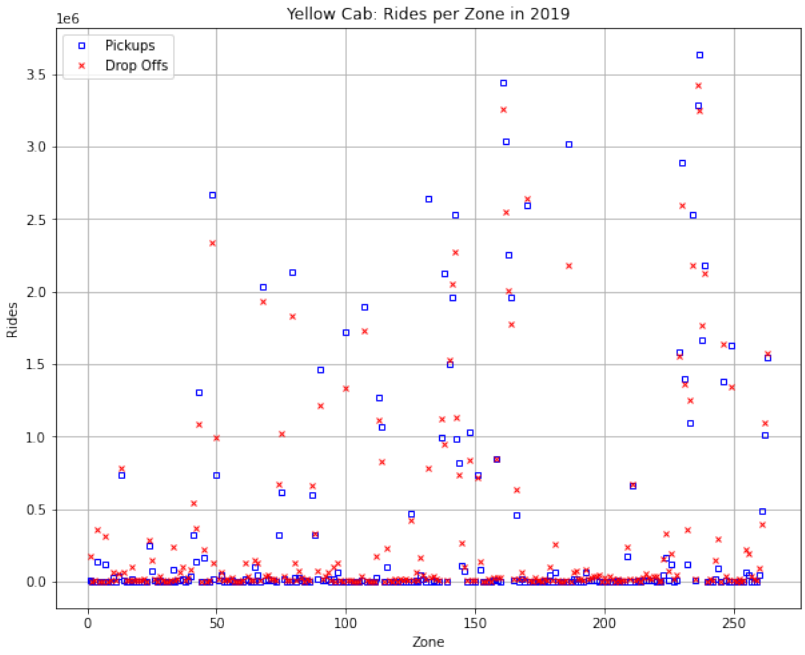

Fig. 1. Summary statistics from 8.3 million Yellow cab rides in 2019 by zone, with pick-ups and drop-offs information shown

One of the challenges of this data set is presenting results from 265 zones, especially when the distribution of data is so uneven. The rides per zone appear to follow a Parto distribution and while some zones may see millions of taxi rides in a year, others may see only hundreds. While this in and of itself is an interesting finding, to make our results easier to interpret we plan to focus our presentation to a fraction of zones. To determine this, we sorted by ride volume and sought a clear decision boundary.

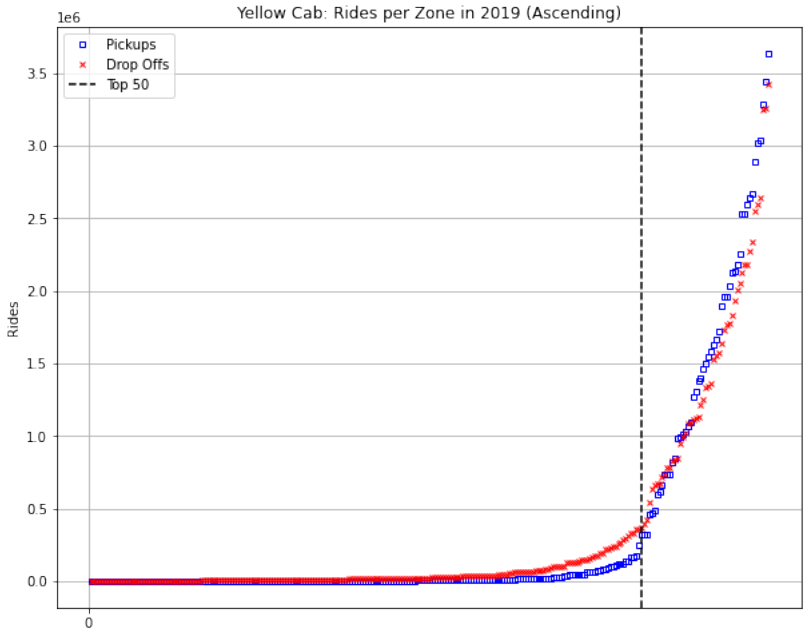

Fig. 2. SORTED: Summary statistics from 8.3 million Yellow cab rides in 2019 by zone, with pick-ups and drop-offs information shown

A clear decision boundary emerged around the top 50 rides. At this point (see Fig. 2), there's an inflection point and the top 50 zones represent 95% of all rides in 2019. Therefore, moving forward our discussion of results will focus on key findings from more popular zones. We found it interesting some zones are so sparse, and will explore reasons in the final report.

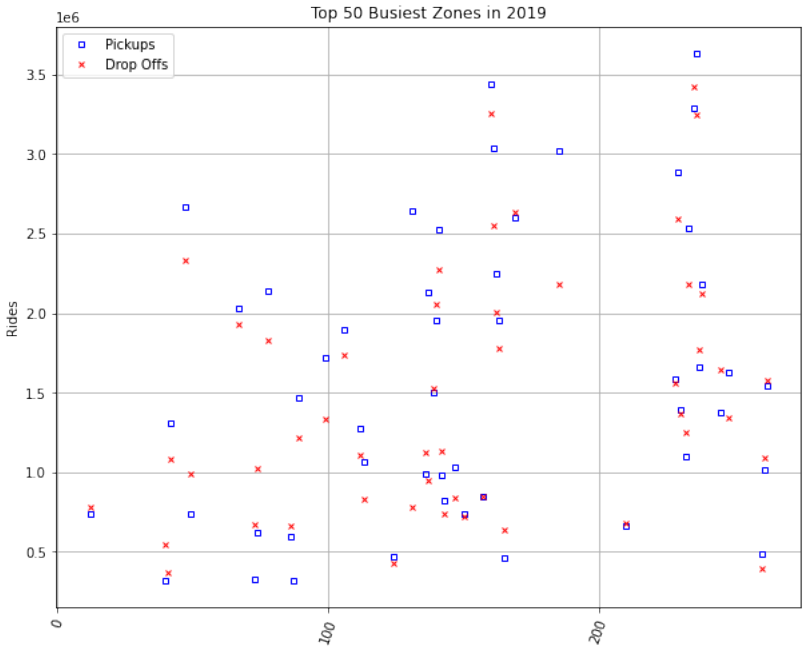

Fig. 3. The top 50 zones for Yellow cab rides in 2019

7.2 Summary Matrix

Many of our research questions required summary statistics of rides from the data set, and there was an immediate need for a compact representation of the data that we refer to as a "Summary Matrix", $S_{265x265}$. The *columns* of S correspond to the zone where a ride begins (pick-up) while the *rows* of S correlate to the zone where a ride terminates (drop-off). In this way, the entry $S_{i,j}$ gives the number of rides beginning in zone j and terminating in zone i over the summary period represented by the matrix.

This matrix is constructed as follows (see Alg. 1):

Algorithm 1 Building a Summary Matrix (S) from Rideshare Data

Require: X is an array of rideshare data where X_k is the row vector representing the kth ride and all of its associated statistics. DO is a feature of that set (and column of X) representing the drop-off zone, such that $DO_k \epsilon \{1, 2, ..., 265\}$ and represents the drop-off zone of the kth ride. PU is a feature of that set representing the pick-up zone, such that $PU_k \epsilon \{1, 2, ..., 265\}$ and represents the pick-up zone of the kth ride. S is initialized as the zero matrix $0_{265,265}$
for every ride X_k in X **do**
  $DO_k \rightarrow i$
  $PU_k \rightarrow j$
  $S_{i,j} += 1$
end for
return S

Once developed, Summary Matrices can be quickly found for each month of data. The Summary Matrices have a number of attractive features for data analysis:

- Many summary matrices can be added together to create summary matrix for the encompassing time period.
- the sum of all rows quickly and efficiently provides a vector describing summary pick-up information.
- the sum of all columns quickly and efficiently provides a vector describing summary drop-off information.

7.3 Ride Associations

One of our primary areas of interest entering this project was the "association" of different drop-off and pick-up zones. In other words, what are the correlations between given pick-up and drop-off zones? Our initial thought was to apply lift calculations and association rules to answer this question, but these required some modification. The lift computation considers associations across sets, where order of elements doesn't matter. For the question of pick-ups vs. drop-offs, there is an implicit order. A ride beginning in *Zone* 4 and ending in *Zone* 14 should not count the same as the reverse ride.

Instead, we let A be the event that a ride begins in a given zone and B be the event that a ride ends in a given zone, and consider the probability a ride ends in B given that it started in A: $P(B|A)$. Alone, this probability gives little insight, so we normalize by the probability of any ride ending in B, $P(B)$. This ride coefficient (r.c.) is found by:

$$r.c. = \frac{P(B|A)}{P(B)}$$

The law of conditional probability allows us to further substitute for this expression:

$$r.c. = \frac{P(A \bigcap B)}{P(A)P(B)}$$

This "ride coefficient" is analogous to lift but has a directional component. A ride coefficient value of 1 implies no correlation between destination and departure point, while a value > 1 means they are likely to be found. Unlike lift, its import to consider the $r.c._{x \to y}$ as well as $r.c._{y \to x}$, as they are not necessarily the same.

The structure of the summary matrix makes this value easy to calculate for every possible combination of destination and arrival. The full results of our origin/destination analysis will be included in the final report.

7.4 Financial Analysis

One of the elements we hope to learn about is the financial side of cab rides. There are several attributes which are relevant. Some are more interesting than others. Tipping is particularly interesting for it's supposed connection to the perceived quality of of a ride from the perspective of a passenger. We have accomplished some preliminary analysis with regards to this topic. The first step was to calculate basic quantities such as the mean, median, and standard deviation. While not particularly useful from a datamining perspective, they nonetheless provide a baseline when looking at the data. We constructed boxplots as well.

From there, we began to look more deeply into the data. Since this data has a spatial component, we hope to find connections between the pickup and dropoff location and tip amount. As a start, we wrote code which produces the statistical fundamentals for each individual pickup and dropoff location. Even with this basic analysis, some interesting trends are beginning to emerge. For example, the pickup location with the highest mean ($10.63) and second-highest median ($10.00) is Newark Liberty International Airport - a location outside of New York City. The same holds for this airport as a dropoff location, with a mean of $12.38 and a median of $15. However, these trips compose an extremely small part of the total dataset - about 0.18%. The trend of high airport tips does continue. John F. Kennedy International airport, located in Queens, is the pickup location for about 3% of all trips during this month. Interestingly, the mean of

$5.67 tip is substantially higher than the global mean of $2.22. The same holds for the third airport: LaGuardia's mean is $5.43 while accounting for slightly less than 2% of trips. Does this represent vacationers who are less clear on the expected tips? Or are these higher actually a function of some other variable correlated with the airports, such as trip length of congestion charge?

As part of this preliminary analysis, we constructed a correlation matrix. The factors with the strongest positive correlation to tip amount are the total amount (an obvious connection, as it includes the tip amount), the fare amount, and the tolls amount. The strongest correlation beyond this is the Ratecode ID. Upon further research, we discovered that the airports have a different fare than standard taxi rides - potentially a connection. However, the naive correlation analysis treats the ratecodes as continuous when they are better considered as ordinal, scaled appropriately to one another. This is an issue which we will address in the near future.

7.5 Ride Visualization

Captured below are maps for the Months of February, March, and April of 2020. The purpose of these maps is to visualize the impact of Covid-19 on Yellow cab travel in New York City. As you can see from left to right there was a significant decrease in the number of rides as a direct result of the Covid-19 pandemic. The bins for the count levels indicated by the colors on the map are based on the averaged quantiles from the months of February, March, and April for number of rides total in each location in New York City.

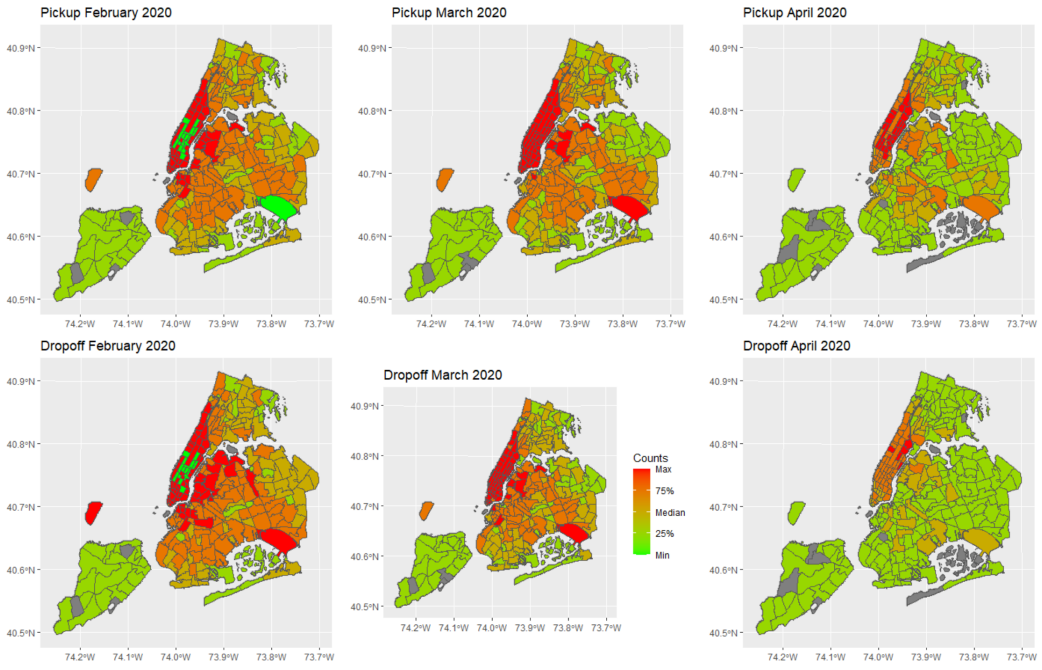

Fig. 4. Ride counts per month by New York City ride location

Note that decisions about binning and how to handle missing data for the purposes of visualization have not been finalized yet, and thus are reflected in figure 4 above.

The map pictured in figure 4 indicates some interesting findings, and suggests that more granular time periods would indicate more interesting trends. As a part of this investigation, included in the final draft will be investigations into the effect of time on yellow cab travel.

8 MILESTONES

Below is a description of major project milestones:

1. [July 13th] - Data Preparation – the data must be clean and ready to use. This includes solving issues like NA or NaN values, improbable values (ie. negative values for ride cost, or 0 distance values). The data will clean and be stored in numpy array's for subsequent work. This will be the most efficient way to work with the data, and pass it into analysis tools.

2. [July 16th] - Exploratory analysis – after the data is clean, we can begin to read the story that the data tells us. Some of our initial intuition about analyses maybe be confirmed, or we may need to change directions a bit. This will be a good time for reflection on the project, and the group can make necessary updates to our projections. It will be crucial that this is accomplished before our progress report is due, in the case where we do need to change directions.

3. [July 23rd] - Primary analysis – For the primary analysis, we will be examining correlations between districts, running regressions for various outcomes, and building a prediction model using a simple neural net. Additional analyses may be added during the exploratory analysis.

4. [July 28rd] - Data Visualization – There are two primary elements to our data visualization step. The first and most important is map making. This will be the most effective tool in conveying our analysis. The second is an addendum to the primary analysis, visualizations of the results of the primary analyses will be necessary. This should include results tables, diagnostic plots, etc.

5. [July 30th] - PART 3: PROGRESS REPORT. For our progress report, we hope to be finished with all data analysis and visualization. This will hopefully give us plenty of time to perform additional analysis depending on our initial results.

6. [August 2nd] - Secondary analyses – this is precautionary, we may add additional analyses during the exploration stage.

7. [August 5] - Report and presentation rough drafts

8. [August 11] - Final presentation and report

8.1 Updates and Milestones Completed

In summary, Milestones [1] and [2] are completely finished. Milestone [3] is largely completed, with the major exception that we haven't deployed our methods to the Green cab and ride-share data sets yet. We expect this to be straightforward and look forward to interpreting the results. Milestone [4] is mostly complete, and the remainder of our energy here will be in developing animations. More details follow below:

We made major progress in many areas of this project. The data required more cleaning than we initially anticipated. Some rows in the data contained values that were either undefined or abnormal. For example, there existed many trips with zero passengers, a negative tip, a negative total fare, or a negative trip distance. Most likely these were generated due to errors. Regardless, we needed to decide how to approach these values.

Perhaps the simplest to explain is the trips with zero passengers. There are some taxis which use trips for deliveries - in fact, the New York City Taxi and Limousine Commission has a program for food deliveries. Out of 6,299,354 trips in the February 2020 dataset, 123,583 are trips without a passenger. This equates to 1.96% of trips in the dataset. This

number, while small, is far from insignificant. However, since the aim of this project is to understand the patterns of passengers, these zero-person trips are not relevant to the task at hand. Thus, the clearest option is to ignore these trips for our analytical purposes.

The reasoning for other aberrations are less clear - namely, trips with zero or negative tips and fare. These trips count for 0.0028% and 0.338% of the total trips, respectively. These percentages are approaching negligible; we therefore decided to remove these trips from the set.

Some preprocessing amounted to calculations. The clearest place where this occurs is the conversion of starting- and stopping-times to a single number representing trip duration. For temporal analysis, this will be much more useful. Additionally, we enacted a simple conversion from the binary "Y" or "N" to 0 and 1.

8.2 Milestones: To Do

We still intended to deploy our analysis to the Greencab and Ride-share data sets. We still intended to animate our maps.

To this point, much of the financial analysis has been fairly simple. Moving forward, we hope to consider more complex factors, including the number of riders as well as the temporal element. Such analysis will require more complex methods. Regression analysis should provide even more of a baseline, and if necessary we will move into more advanced analysis such as a neural network. The purpose of the mining so far has been to help us discover which questions may be interesting to discover, and it has been successful in that regard.

The next major milestone is completing the rough draft presentation and report by the end of next week.