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

GROUP 5

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. TLC TAXI DATASET: A comprehensive view of vehicle for hire data in NYC from 2009 until present: GROUP 5. 1, 1 (July 2021), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

Today, slightly 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.

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:

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.

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 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