Mining For Rideshare Patterns

Saloni Sharma
CSPB Student
CU Boulder
Sash8251@colorado.edu

Ronald Thompson
CSPB Student
CU Boulder
Roth9236@colorado.edu

Jonathan Prindle
CSPB Student
CU Boulder
Jopr2193@colorado.edu

PROBLEM STATEMENT/MOTIVATION

The overall goal of our project is to utilize rideshare data to reveal interesting patterns that would improve the customer experience while simultaneously optimize profits for drivers. In particular, there are three major areas where we will focus our research.

First, we will perform a general analysis of rideshare patterns to better understand how customers use rideshare services. Specifically, we will concentrate on four central questions.

1. During peak traffic times, are there any notable patterns that can potentially ameliorate the customer experience?
2. What are the average durations of rideshare trips throughout the day and how does this affect the customer and driver experience?
3. What times are better and what times are worse to take a rideshare?
4. At what point do drivers hit diminishing returns?

In answering these questions, we will develop a comprehensive picture of the rideshare experience for both the customer and driver in aims of creating strategies to better their experiences.

Next, we will compare city specific rideshare patterns with general rideshare patterns. Principally, we will verify whether patterns with similar attribute values produce similar results. This examination will offer insights into the impact of rideshare on individual communities and countries as a whole.

Additionally, if we have extra time, we would like to explore questions in regard to census data. In particular we would like to examine if rideshare work and tips are related to the area in which a driver works. We would also like to examine which routes are frequented by customers and if they display some type of routinized pattern (e.g. ridesharing to work every day).

In summation, with the analysis of these questions, we hope to improve customer experiences and create a more enjoyable rideshare experience overall.

Update: Our overall goal since we have started has stayed the same. To reiterate, that goal is to use rideshare data to reveal interesting patterns that would improve customer experience while simultaneously optimizing profits for drivers.

So far, we have focused primarily on the four central questions originally outlined. The main deviation has come in the form of focusing more on individual cities. While we are still comparing data between New York City and Chicago, the data sets are divergent enough that we will not be able to compare the two cities to the level that we originally had planned.

Additionally, each member of our team is now diverging and focusing on different questions that align with our central questions. This has allowed us to dive deeper into each question in order to better analyze how to improve both the customer and driver experience.

LITERATURE SURVEY

As rideshare has become increasingly interconnected with our lives, so too has the amount of research in building more effective methods for better performance. The following are a list of sources that investigate data in various directions. Some pose questions similar to our own

while others venture to a different path. Each source has been provided with a summary.

1. *Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance* is an article by Todd Schneider that explores how rideshare has affected New York City. This article offers a thorough explanation of rideshare utilization by customers with topics including frequented areas of rideshare, comparisons between different types of rideshares, length of trip, and factors that affect rideshares such as weather.

2. *Uber pickups in New York City*. This is a data set that was explored by the site FiveThirtyEight. Broadly, FiveThirtyEight used the data for a variety of stories to describe how rideshare has been affecting the customer experience. In particular, most stories focused on whether or not rideshares were helping or hurting the community at large.

3. *New York City taxi fare prediction*. This competition had participants predict the taxi fare between two locations.

4. *An empirical analysis of on-demand ride-sharing and traffic congestion*. This is an in-depth study from Arizona State University that investigates the social welfare impact of rideshare, specifically Uber, on their home city. Essentially, it outlines the benefits and drawback of rideshare.

5. *The Economics of Ride Hailing: Driver Revenue, Expenses and Taxes.* This research article from MIT provides a cost analysis of rideshare. The authors look at the profit margins for drivers of rideshare and whether or not the profit gain outweighs the costs incurred.

6. *Uber Movement*. This website tracks Uber movements and speed. It provides anonymized data such as locations and insights into how fast Uber can get from point a to point b at different times of the day in an effort to improve urban-planning.

PROPOSED WORK

In order to efficiently utilize the raw data we have collected, we will need to preprocess it through steps such as cleaning, integration.

First, we will clean our data sets. All missing values and any unnecessary (outlier) values will be removed as long as they are not a large percentage of the data set. To smooth the data further, we will also bin by fare in 2.50 increments as the Chicago data set only includes such values. We plan to keep any duplicate values as they can demonstrate frequent patterns.

Second, we will integrate our different data sets. Specifically, we will merge the 2014 and 2015 data from data set 2 together to create a single data set for Uber trips in New York City. We will also combine the monthly data files from 2015 from data set 3 to produce one data set. Then, we may integrate data set 2 with data set 3 to create one large set of data for NYC as well as standardize certain attribute values such that they can be better compared with values in dataset 1.

Third, we will transform and reduce the number of attributes in our data with dimension reducing algorithms such as PCA or with backward elimination. We will examine highly correlated values together to potentially remove or simply cluster attributes.

Overall, we will begin to work with smaller sample training sets and reserve test sets with data matching certain attributes such as month of year.

Our process will differ from the previous examples cited in the literature survey section in a variety of ways. We will be using more comprehensive data sets, which span over various regions in contrast with one specific area. The preprocessing and integration of our data will be more targeted to address our specific goals of customer experience improvement and profit optimization as well.

Update: A good portion of our proposed work has been completed. We are at a point where we have more specific questions that need to be answered.

This has led to additional work that now needs to be completed.

At this point we need to refine the tools that we have been using. Over the last few weeks, we have experimented and found what works best and what we needed to be discarded. We have transformed our data and cleaned it. After getting our initial results we have determined we need to do additional transformations on certain pieces of data in order to more accurately answer the questions we outlined in our motivation section.

Currently we have a better understanding of the tools we originally set out to use in order to complete this project. With the addition of other, more specialized tools, we have determined that we now have everything we need to thoroughly and accurately answer our original questions. Going forward, we will implement all our tools with our refined and further transformed data in order to reach conclusions and present our findings.

Additionally, we are looking for concise visualizations that will allow viewers to interact and interpret the results in an easy and straightforward way.

DATA SETS

1. **Dataset 1**: City of Chicago. This data set provides 129 million different data points with 21 categories. We have a variety of numeric attributes such as trip miles, trip seconds, binary attributes such as shared trip authorized, and interval attributes such as fares, which have been rounded to the nearest $2.50, and tips rounded to the nearest $1.00. Overall, this dataset provides comprehensive data and the relevant attributes in regard to the topics we are focusing on. City of Chicago Rideshare Data.

2. **Dataset 2:** Uber Pickups in New York City. This data set offers the nominal attribute of location (for pickups) in NYC and ratio interval attributes in the form of time. This dataset will allow us to track when and where Uber pickups occurred. This will allow us to measure the frequency of trips at different times. Uber Pickups

3. **Dataset 3:** NYC Taxi and Limousine Commission. This dataset offers various numeric attributes in regard to fare, tip, trip distance, etc. This dataset will allow us to examine rideshare rates, tips, and distances of the trip. In conjunction with dataset 2, we will be able to measure peak travel times and measure when travel times are less strenuous for passengers and more beneficial for drivers. NYC TLC Trip Data

EVALUATION METHODS

We will focus on a variety of ways to evaluate our data in order to ensure the best results.

First, we will compare our results with other existing analyses, specifically from those we cited in our literature survey, for aspects such as frequented locations and times of trips. We will use the following questions as a guide to evaluate our results. Does our analysis show a positive or negative correlation of high fares with certain locations and at specific times of day, and does this match the prior work done on rideshare data?

We will create hold outs of our data when running models to see initial sample performance. We will test to determine if we can use them to predict times or locations with high fares or tips. Additionally, we will test and see if the analysis of Chicago data can predict similar attributes in New York and vice versa.

Finally, we will try to match anecdotal experiences with our findings for both riders or passengers and drivers.

Update: We have successfully tested various hold outs of our data. While initially some of the resources we used to analyze our data proved insufficient, we have found various tools and methods to gain accurate results.

The biggest difference between our original evaluation method and our current evaluation

method is that we have shifted away from predictive models. Given the size of our two biggest data sets (146 million data points/22GB and 129 million data points/32 GB), we found that trying to model that much data was inconceivable given our current technological constraints. After researching sufficient hardware/software it was determined we would need around 64GB of ram to sufficiently be able to model on a personal computer or we would have to invest heavily elsewhere, which was not feasible for this project.

Currently, we are analyzing trends based on the data and comparing our results with those cited in the literature survey. We are still analyzing interesting correlations to see if we can glean any new insights.

Additionally, we are matching anecdotal experience to our findings. In particular, we have discovered some tip/distance and full fare/distance results that are being matched to anecdotal experiences from rideshare drivers. Our data presently shows that driving certain distances may result in higher returns and we're analyzing whether or not these accounts track with our data.

TOOLS

We plan to use Python as the main programming language for the project (via Jupyter Notebooks and VS Code). We will also utilize several resources and libraries provided by Python to implement our mining process.

The updated tools offer a list of tools that we have incorporated into our project. Over the course of the project we have supplemented our tools in order to better assess and analyze our data.

- SciPy, Pandas, Matplotlib, NumPy, GeoPy, GeoPandas, Tableau, Dropbox, SQLite Database.
- Updated Tools: Apache Spark, Qlik, RapidMiner.

We plan to use the previously stated datasets in their CSV format, which have been downloaded already. Our code will be stored on Github. Given our dataset sizes, we are exploring options for storage which include Dropbox or Google Cloud. Additionally, each member will retain their own personal copies.

Update: We are currently storing individual copies of the data and independently transforming and cleaning it in order for us to address specific questions in regard to our goals. All files have been merged into three major csv files, which are currently being stored and shared through Dropbox due to their size.

MILESTONES

We are planning to have the following tasks completed by the marked dates:

March 27th – Preprocessing: have data cleaned, integrated and transformed by this point.

April 10th – Select and test models: have tools functional and working on small sets of our data. Refine any questions or other aspects of the project that need to be refined.

April 17th – Progress report: have completed progress report submitted to Moodle.

April 24th – Finalize and Evaluate: have all analysis on data completed at this point. Write an initial README file for github and begin a rough draft for the final report.

May 1st – Submit project: have the project presentation, final report, README and other files in GitHub finished and submitted.

Milestones Completed

So far, we have completed the preprocessing portion of our data. While we have each taken individual approaches to the data, we have been able to sufficiently clean, integrate, and transform each data set to suit our specific needs.

Each of us has also done an initial analysis of the data sets and have some results. While these results will need to be refined and additional transformations of the data will be made, we are able to draw initial conclusions and correlations in

answer to the questions posed in the motivation section.

All tools have been tested with small sets of data before applying to the entire datasets. We feel confident in the tools that we have used thus far. While we have shifted away from predictive models, the tools we are using allow us to accurately assess trends and draw conclusions.

MILESTONES TODO

April 20th-24th – Final Transformations: We will continue to complete our individual transformations. Once transformations have been completed, we will analyze and cross reference with any anecdotal information we have.

April 24th – Finalize and Evaluate: Have all analysis on data completed at this point. Write an initial README file for github and begin a rough draft for the final report.

April 27th – Final Result Discussion: We will come together and discuss our final results. At this point we will analyze all our results and discuss our conclusions. We will also have completed our rough draft for the final project report at this time and will discuss any edits that need to occur.

April 29th – Project Presentation Creation: Determine what information we should include in our project presentation. Construct a presentation and have a finalized product ready to submit by the end of the day.

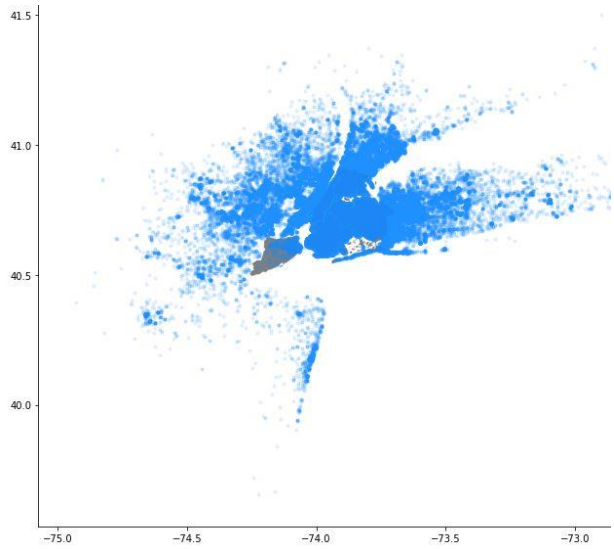
April 30th – Final Review: Review and proofread all project materials.

May 1st – Submit project: Have the project presentation, final report, README and other files in GitHub finished and submitted.

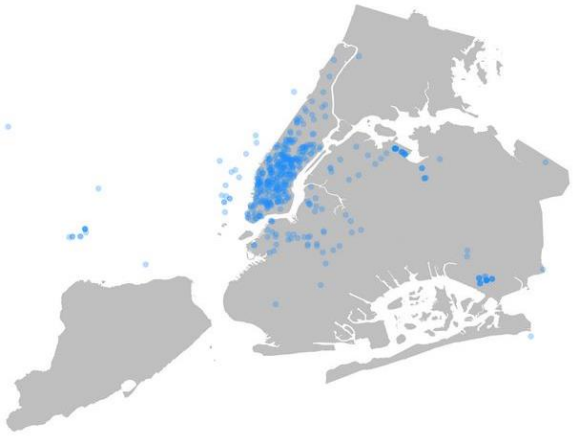
RESULTS SO FAR

Using GeoPandas for geographic data has allowed us to map some of the information presented in the data sets.

The following figure shows the NYC Uber dataset from 2014 (April - September) with 4.5 million points plotted over a map of the city. This is a stepping stone that will allow us to accurately chart where trips originated from and where they are going.

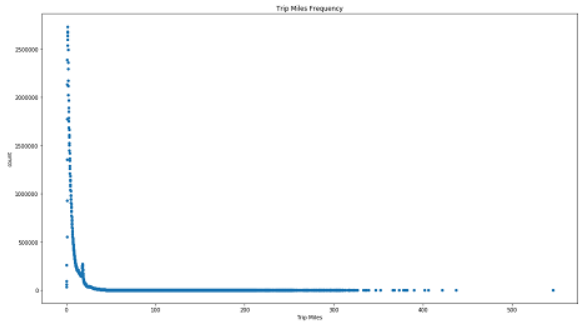


The map below is a less noisy version with a reduced number of data points, which shows Manhattan as the most frequented area for pickups. For the entire dataset, we will consider grouping the location data, such as by borough or census tracts, in order to better discern popular pickup locations overall.



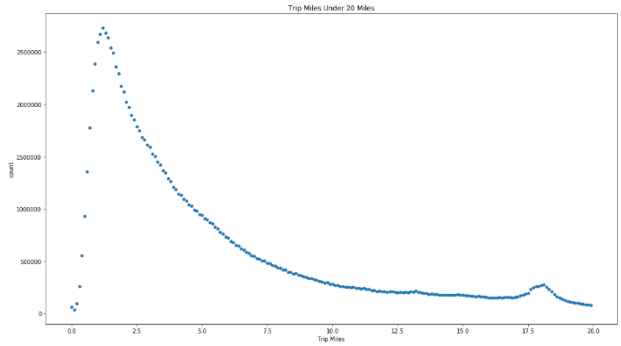
We have also been able to do some analysis on trip distance frequency. The following graph shows a major spike around the 18 mile mark for our data. After researching it, this distance resembles the

average trip from downtown Chicago to O'Hare airport.

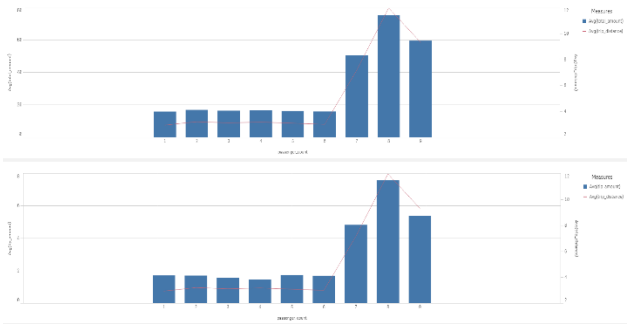


As a general trend we have also noticed that most trips in both Chicago and New York City are shorter and tend to be within 1 mile and 3 miles.

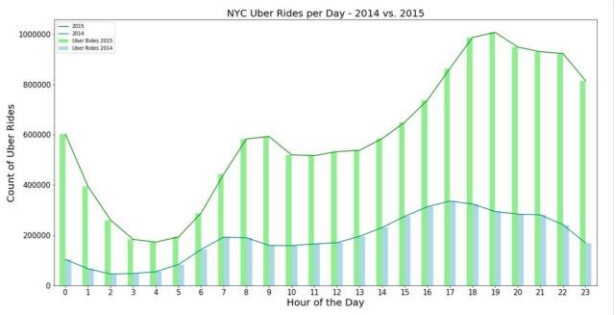
The below diagram focuses on Chicago trips between 0 and 20 miles. Other than the airport bump, most trips are between various areas in the city of Chicago (rather than the wider area), most are from one neighborhood to the one next door rather than intra-neighborhood trips.

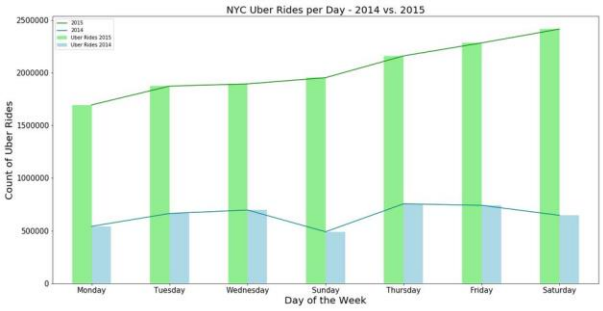


For New York City, we were able to glean that when you have four passengers you gain the least amount overall for mile traveled and you gain the least tip for mile traveled. On the flip side, you gain the most per mile when you have seven passengers and you get the biggest tip when you have seven passengers.



We have also been able to infer the frequency of Uber rides during the time of day and the day of the week through comparisons of the 2014 (Apr - Sept) and 2015 (Jan - June) Uber datasets with 4 million and 14 million data points, respectively.





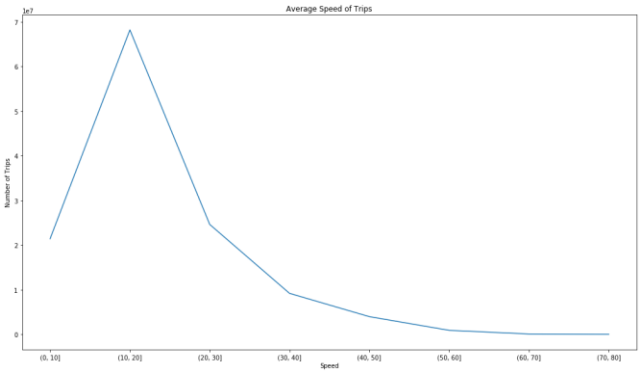
With this knowledge, we can surmise when an Uber driver can maximize the supply of his service in relation to the demand of trips. According to these charts, we can see that weekends and evenings generally have the largest amount of Uber rides.

The first chart with 'hour of the day' shows a similar pattern in both datasets. However, the second plot with 'day of the week' has different patterns. For the 2015 data, Saturday is the highest

and Monday has the lowest, but for 2014, the rides dip on Saturday while Thursday is the highest and Sunday is the lowest.

The reasons for the difference in the results may include the different seasons in which the data was collected (i.e. there are fewer Uber rides for non-essential trips on weekends during snowy weather). For further analysis, we will break down the dataset to compare the overlapping months for a better correlation.

If an Uber driver plans for the right time of day, the right day of the week, the right location and the right number of passengers, they can maximize their revenue and profits while providing a regular supply to the consumer. Our analysis can help drivers know their customers' needs before they do.



Most of the trips in Chicago are between 10 and 20 mph. This shows when an Uber driver will be stuck in more traffic and so combined with understanding supply and demand they can make a better call if it's "worth it" to be driving at that time. This also can change the margin for the driver as fuel costs change with the efficiency of driving.

REFERENCES

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: https://doi.org/10.1145/1188913.1188915.

[2] Todd Schneider, 2015. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. toddwschneider.com

[3] 538, 2017. Uber pickups in New York City.. Kaggle.com/UberPickups

[4] Kaggle Competition, 2019. New York City Taxi Fare Prediction. Kaggle.com/TaxiPrediction

[5] Ziri Li and Yili Hong and Zhongju Zhang. 2018. An empirical analysis of on-demand ride-sharing and traffic congestion. Scholarspace.edu

[6] Stephen Zoepf, Stella Chen, Paa Adu, and Gonzalo Pozo, 2018. T*he Economics of Ride Hailing: Driver Revenue, Expenses and Taxes.* Princeton.edu

[7] David Schnurr. 2019. Uber Movement https://medium.com