**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

**ABSTRACT**

From the outset of this project we had one overarching goal, which was to utilize rideshare data to reveal interesting patterns that would improve the customer experience while simultaneously optimizing profits for drivers.

After analyzing our data, we were able to be more targeted with our questions. We were able to answer the following:

1. During peak traffic times were there any notable patterns that can potentially ameliorate the customer experience?
2. What times are better, and what times are worse, to take a rideshare?
3. When do drivers maximize their profit and potential?

During the course of our analysis we uncovered some interesting patterns.

[insert additional results]

When analyzing the data, we also noticed that there was a correlation between distance, tips, and overall fare in the New York City area. Carrying a certain number of passengers typically yielded better tips and better fairs overall.

Additionally, trips to certain locations consistently yielded better tips and overall fare, making these locations far more lucrative for a rideshare driver.

**INTRODUCTION**

Taxi cabs and limousines have been around for decades. However, in March of 2009 uber was founded and changed the ridesharing landscape forever. Increasingly rideshares are proving to be a cheaper alternative to costly monthly vehicle/insurance payments. In urban areas, where it is hard to own a vehicle, rideshares have dominated and changed the public transportation landscape.

This project sought to better understand how rideshares are affecting both customers and drivers.

In particular this project is interested in finding trends and patterns that are beneficial to the customer experience while maximizing profits for the drivers.

One of the first questions that was addressed was “During peak traffic times were there any notable patterns that can potentially ameliorate the customer experience.” Connecting to our original goal, we wanted to explore if there were any lucrative patterns that could help the customer and driver experience during peak travel times. For instance, is there a more efficient path that would result in cheaper fare, larger tips, or a shorter travel time?

Another question we explored was “what times are better, and what times are worse, to take a rideshare.” Analyzing this data allowed us to track when a customer could reach their destination more efficiently and when a driver could maximize their own profits. Additionally, we were able to identify trends where certain times had more demand while little supply and vice versa. By identifying and addressing these times we are able to maximize the ridesharing service.

Finally, we explored the question of “when do drivers maximize their profits and potential.” For this question we looked at when drivers made the most tip and the most fare in relation to the distances they traveled. We also analyzed geographic data to examine where drivers were making the most profit.

These three central questions allowed us to explore our overall goal and assess trends that benefitted both customers and drivers. During the course of our examination we uncovered interesting and significant trends that we believe will improve the rideshare experience in both Chicago and New York City.

**RELATED WORK**

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.

**DATA SETS**

For this project we utilized three major data sets. These three data sets provided nearly three hundred million unique data points that we were able to drill down and assess.

The first dataset was on rideshare data from the City of Chicago. This particular data set provided 129 million data points with twenty-one unique attributes. It also provided 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. For each of analyses we were able to remove excess attributes and focus on a select few. This comprehensive data set was all that we needed in order to analyze Chicago.

For New York City we used two different data sets. The first data set provided information on pickups in New York City. In particular it provided the nominal attribute of location and the ratio interval attribute of time which allowed us to track when and where the majority of rideshares took place in New York City.

The last data set provided a more comprehensive picture of how rideshares were being utilized in New York City. This particular data set provided numeric attributes in the form of fare, tip, and trip distance. It also offered limited location data.

Using both of the New York City data sets together allowed us to properly examine how ride shares were being utilized, where they were being utilized, and what trends benefitted both customers and drivers the most.

The three data sets were pulled from the City of Chicago, Kaggle.com, and New York City’s transportation commission. Each source was verified and determined to be a legitimate source of information.

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](https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p).
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](https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city#other-Firstclass_B01536.csv)
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](https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)

**MAIN TECHNIQUES APPLIED**

[Feel free to add how you prepped/transformed your data]

For this project we each took an individualized approach to cleaning and preparing the data. Given the overall volume of the data this proved to be the most effective approach for each of our specific questions.

One of the steps in our preprocessing was combining our and data and ensuring that it was useable. The Chicago Data set and the first New York City data set were in good condition. The second New York City data set was broken into twelve different csv files, one for each month, that had to be combined before we could clean it. Through a windows command line prompt, we were able to copy all the different months of csv data into a single file which comprised of 146 million different data points.

[I’m guessing you probably cleaned your data in different ways then I did. I primarily utilized the software I was working with]

Once the data sets were combined, we started cleaning our data. Most of the software that we utilized allowed us to transform the data from the outset. Qlik and RapidMiner in particular allowed the replacement or removal of nulled values. Additionally, they also allowed the filtering of data which allowed us to eliminate outlier data and data that had been compromised.

Various techniques were applied to analyze the data. We utilized a variety of tools including Qlik, RapidMiner, Apache Spark, and GeoPandas.

Qlik provided basic statistically descriptions of the data such as the mean, median, mode, quartiles, and general dispersion of the data. Additionally, Qlik had a custom extension that allowed for kmean’s clustering which helped highlight areas of profitability for drivers. Qlik was also able to provide geographic density maps allowing us to examine areas where the best tips and best fares are.

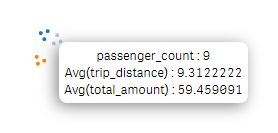
RapidMiner was able to detect some outlier data for both data sets. While no noticeable trends were noted as a result of the outlier data, RapidMiner provided information that backed up other results.

[Expand]

**KEY RESULTS**

Our analysis of the data had yielded various trends that can benefit both customers and rideshare drivers.

For New York City, we were able to analyze the best value for rideshare drivers in terms of passenger count. Through k-means clustering we were able to separate passengers into essentially one of three major groups. Each group had variable profit margins in terms of fare and tip.



Group 1: The Good

This cluster was composed of rideshares where there were 7, 8, or 9 passengers for the trip. On average this cluster represented the best return in terms of fare per distance and tip per distance. Below is a breakdown of the profitability of these ranges

7 passengers: $7.008/per mile. Tip: $.666/per mile.

8 passengers: $6.629/per mile. Tip: $.633/per mile.

9 passengers: $6.39/per mile. Tip: $.580/per mile

Group 2: The Bad

This cluster composed of rideshares where there were 1, 5, or 6 passengers. On average this group underperformed in comparison to group one. The fair per mile distance is comparable to group three, however, this cluster typically tipped better then group three.

1 passenger: $5.44/per mile. Tip: $.586/per mile.

5 passengers: $5.22/per mile. Tip: $.548/per mile.

6 passengers: $5.27/per mile. Tip: $.566/per mile.

Group 3: The Ugly

This cluster was composed of rideshares where there were 2, 3, or 4 passengers. Typically, this group had similar fare per distance in relation to group two, however, the amount of tips per distance was far lower.

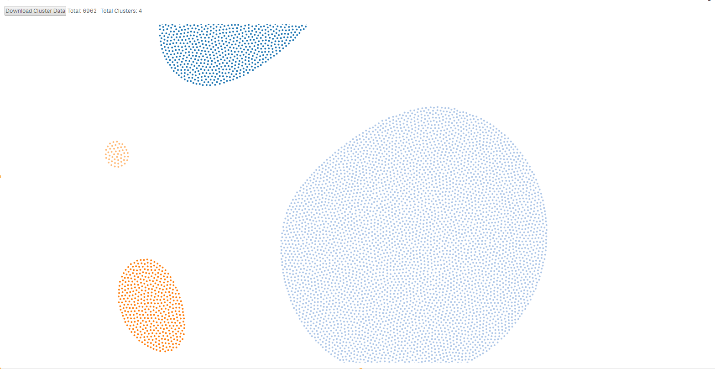
2 passengers: $.5.24/per mile. Tip: $.531/per mile

3 passengers: $.5.26/per mile. Tip: $.516/per mile

4 passengers: $5.15/per mile. Tip: $.468/per mile

It is fair to point out that the number of large passenger rideshares numbered fewer than the number of rideshares with fewer passengers. However, even after random sampling was completed for the data for fewer rideshare passengers and comparing them to the results for larger groups of rideshare passengers these clusters and results held.

We also conducted a K-means clustering analysis of the tips in NYC area which revealed some interesting trends.



The light blue represents tips that average between $1 and $10. As you can see from the graph a majority of trips tend to tip between this range despite distance traveled. In other words, even if a person traveled 50 miles, there was still a high chance of being tipped between this range.

Dark blue represents the next range of tips which falls between $11 and $35. This is the second biggest cluster. Orange represents the third cluster with ranges between $36 and $90. The red cluster represents the extreme with tips in the range of $91 and above. Arguably these results aren’t surprising as the majority of trips are short distance. Perhaps the biggest surprise of this cluster analysis is that distance didn’t necessarily indicate a large tip. In other words, you could travel 100 miles and get a tip of $5 and conversely you could travel 5 miles and get a tip of $30 dollars. Typically, however, no matter what distance you travel you’re far more likely to end up in the giant light blue cluster and consequently get a lower tip.

The following is a K-means cluster analysis of distance in relation to fare.

A close up of a logo

Description automatically generated

This clustering is more even then the in regards to the tip. I.E. that more distance you travel, the higher the fair. The clustering is proportional to the number of trips traveled at that distance, for instance there are more shorter trips to light blue reflects more small trips at a smaller fare.

By contrasting this k-mean cluster analysis with the tips analysis, we can see that the gratuity system is not always beneficial to drivers, as lower tips tend to dominate.

The geographic data included with the New York City data set also allowed us to see areas where larger total fares were the most dense.

A picture containing map

Description automatically generated

There are four major areas that this map highlights. First Manhattan in general is incredibly dense. There are numerous businesses and corporations located in this area and suggests that a fair amount of people commute into the city which drivers up the fare density. The three other areas that were the densest were the three airports located around Manhattan, which are JFK, Newark International, and LaGuardia.

In terms of large tip amounts, there is a similar trend where larger tips are located near Manhattan, however, more notably tips aren’t as lucrative at the airports.

A close up of a map

Description automatically generated

In general, one of the more curious insights while examining both the New York City data sets and the Chicago data sets is the lack of tipping in general for ride share drivers. In the Chicago data set there were over 100 million entries of 0 or more tips. For the New York City data set many tip amounts were also missing. Based off anecdotal accounts this can be explained in one of two ways. The first, and arguably more prominent explanation, is that many people do not feel obligated to tip rideshare drivers. This is backed up by various studies. Andrew Hawkins in particular explored this topic and found around 66% of rideshare passengers do not tip their drivers. The second explanation for this missing data is that tips are taxable income. By not reporting cash tips, drivers are able to net more profits and so that may be a cause of the lack of data.

**APPLICATIONS**

Our results have yielded a few potential applications. The overall goal of trying to improve the customer experience while maximizing driver profits could be achievable.

Tipping seems to be an area of need based off both the Chicago data set and the New York City data sets. In both instances, there was a severe lack of tipping. Uber’s original business model discouraged tipping as it thought that it would cause issues for both passenger and driver. It wasn’t until 2017 that tipping was officially introduced as part of Uber’s app. In fairness, even after analyzing other forms of rideshare, for instance taxi’s and limousine, there was still an over abundance of data that showed people were not tipping.

Our analysis revealed areas and conditions where rideshare drivers are more likely to make higher tip rates. For instance, the New York City data strongly suggests that if you transport more than seven passengers and frequent either downtown area or the airports you are more likely to maximize your profits.

The data also opens the debate to larger questions. For instance, should tips be factored into the fare for rideshares? Or perhaps should tips be eliminated altogether, and higher rates of pay be given to drivers? The only question that we can answer definitively at this point is that rideshare drivers are not benefitting off the gratuity system.

**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](https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/)

[3] 538, 2017. Uber pickups in New York City.. [Kaggle.com/UberPickups](https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city)

[4] Kaggle Competition, 2019. New York City Taxi Fare Prediction. [Kaggle.com/TaxiPrediction](https://www.kaggle.com/c/new-york-city-taxi-fare-prediction)

[5] Ziri Li and Yili Hong and Zhongju Zhang. 2018. An empirical analysis of on-demand ride-sharing and traffic congestion. [Scholarspace.edu](https://scholarspace.manoa.hawaii.edu/bitstream/10125/41152/1/paper0003.pdf)

[6] Stephen Zoepf, Stella Chen, Paa Adu, and Gonzalo Pozo, 2018. T*he Economics of Ride**Hailing: Driver Revenue,**Expenses and Taxes.* [*Princeton.edu*](https://orfe.princeton.edu/~alaink/SmartDrivingCars/PDFs/Zoepf_The%20Economics%20of%20RideHialing_OriginalPdfFeb2018.pdf)

*[7] David Schnurr. 2019. Uber Movement* [*https://medium.com*](https://medium.com/uber-movement)

[8] Andrew J Hawkins, Oct 2019, *Nearly two-thirds of Uber customers don’t tip their drivers, study says.* Theverge.com