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

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". Taxi rides were found to be overwhelmingly centered in Lower Manhattan, heavily used zones have higher tips and as expected, COVID-19 had a devesating impact on the Taxi industry in NYC. NOTE: This file format gave us some formatting issues with figures, and in some places we were left with large margins. This seems unique to this document template and did not appear when checked the document in other templates.

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. 1, 1 (August 2021), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The 20th century saw a monumental transition of human populations from rural settings into urban ones. Beginning in Europe and the United States in the 19th century, industrialization spread across the world and drove an enormous migration of people into city centers, forever altering the way we live. In 1950, 30% of the world population lived in an urban setting—today its over half [1]. Rapidly developing nations saw even more dramatic urban growth over comparatively shorter durations.

The 21st century will see more of the same. By 2030, the global population is expected to pass 8.5 billion, on its way to 11.2 billion by 2050. Driven in equal measures by this overall global population grow and the continued migration into urban areas, urban centers will continue to swell. India, China and Nigeria alone are projected to add 416 million, 255 million and 189 million people respectively to their cities by 2050. By then, it is expected that Earth's urban corridors will have added 2.5 billion people and that 7 out of 10 people on this planet will live in a city. [1]

With this increase in urban living, understanding the way cities work is an incredibly important question. How do energy, people and money move around in the urban environment?

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

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.

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 different types of cabs, the factors that predict tipping outcomes and the effects of COVID-19 on vehicle for hire commuting.

A large advantage of any 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 relatively few 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 and dependant on the task.

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 as 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 cabs and Green cabs. 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.

2 RELATED WORK

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[2]. 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[3] 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.[4] 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[5] 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 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 265 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

VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	congestion_surcharge
1	4/1/2019 0:04	4/1/2019 0:06	1	0.5	1	N	239	239	1	4	3	0.5	1	0	0.3	8.8	2.5
1	4/1/2019 0:22	4/1/2019 0:25	1	0.7	1	N	230	100	2	4.5	3	0.5	0	0	0.3	8.3	2.5
1	4/1/2019 0:39	4/1/2019 1:19	1	10.9	1	N	68	127	1	36	3	0.5	7.95	0	0.3	47.75	2.5
1	4/1/2019 0:35	4/1/2019 0:37	1	0.2	1	N	68	68	2	3.5	3	0.5	0	0	0.3	7.3	2.5
1	4/1/2019 0:44	4/1/2019 0:57	1	4.8	1	N	50	42	1	15.5	3	0.5	3.85	0	0.3	23.15	2.5
1	4/1/2019 0:29	4/1/2019 0:38	1	1.7	1	N	95	196	2	8.5	0.5	0.5	0	0	0.3	9.8	0
1	4/1/2019 0:06	4/1/2019 0:08	1	0	1	N	211	211	3	3	3	0.5	0	0	0.3	6.8	2.5
1	4/1/2019 0:52	4/1/2019 0:55	1	0.2	1	N	237	162	1	4	3	0.5	0	0	0.3	7.8	2.5
2	4/1/2019 0:52	4/1/2019 1:11	1	4.15	1	N	148	37	2	16.5	0.5	0.5	0	0	0.3	20.3	2.5
1	4/1/2019 0:02	4/1/2019 0:03	1	0	5	N	265	265	2	0.01	0	0	0	0	0.3	0.31	0
1	4/1/2019 0:03	4/1/2019 0:03	1	0	5	N	265	265	1	200	0	0	40.05	0	0.3	240.35	0
1	4/1/2019 0:13	4/1/2019 0:20	1	1.3	1	N	237	142	2	6.5	3	0.5	0	0	0.3	10.3	2.5
4	4/1/2019 0:25	4/1/2019 0:55	1	10.06	1	N	249	69	2	32.5	0.5	0.5	0	0	0.3	36.3	2.5
2	4/1/2019 0:14	4/1/2019 0:42	1	18.5	4	N	132	265	1	65.5	0.5	0.5	13.36	0	0.3	80.16	0

Fig. 1. A representative look at the typical raw format for ride data (as a CSV). Data is grouped into separate CSV files on a per mode (Yellow Cab, Green Cab, etc) and per month basis.

4 MAIN TECHNIQUES AND METHODS

Our primary analysis tool for these tasks will be Python, specifically the Numpy, Pandas, Scipy and Matplotlib libraries. Additional, specialized libraries will be used for specific tasks and introduced where relevant.

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.

4.1 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.

4.2 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.
- "NaN" values for pickups and dropoffs are handled in data collection by assigning values of 264 and 265. Data cleaning is simplified then by first forming the summary matrix as defined above, and then removing the last two columns and last two rows.

4.3 Ride Associations

Most ride data for Yellow and Green cabs records the location of pickup and drop-off, binned into 265 zones to preserve sone anonymity. 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 and how can we efficiently them across the entire sample space? 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 of elements across sets, where order of elements does not 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.

In practice, for our 265 zones, there are nearly 70,000 unique pairs of drop-offs and pickups and the subsequent range of "Zone Associations" is quite broad. For an easier comparison, we generally report the LOG of associations and on this scale 0 represents no significant associations. Positive values indicate a stronger association than would be expected and negative values depict the opposite.

The structure of the summary matrix makes this value easy to calculate for every possible combination of destination and arrival.

4.4 Financial

The financial analysis required a multi-pronged approach. There were several basic tools that were used in conjunction. The first step was to calculate basic quantities such as the mean, median, and standard deviation. These provided a baseline when looking at the data. Furthermore, we constructed boxplots as part of the preliminary operations. Another part of this was a correlation matrix, although this ended up being less useful than we initially expected.

From that point, we began to dig more deeply into the spatial component, analyzing each pickup and dropoff location separately. During this process, we created a metric representing the ratio between tip amount and the fare. This is calculated using the simple expression:

$$tipRatio = \frac{tipAmount}{totalFare - tipAmount}$$

Most analysis was done using these tools; more detail is in the results section.

4.5 COVID-19 and Visualization

One of the primary goals of this analysis was to assess the impact of COVID-19 on yellow-cab rides in New York City. It was decided initially that this was best accomplished using a map visualizing the number of rides in each district through the months of March, April, and May. The intent of the maps is to show a gradient from before the COVID-19 pandemic lockdown in New York City, to after the lockdown was initiated.

To further examine the effects of the lockdown in New York City, it was decided to examine the number of rides at a more granular level. Thus a line plot showing the number of rides over time for the top 10 most visited taxi districts in

New York City. These top ten districts have thousands of rides each day, and it was determined that showing how the number of daily rides falls off would give some indication of the true effects of the pandemic lockdown.

5 KEY RESULTS AND APPLICATIONS

5.1 Pickup vs. Dropoffs

To investigate the associations between pickup zones and dropoff zones, we choose to analyze data over a 1 year period. A 1 year study period was chosen to maximize data collected, but limit temporal drift that these values no doubt experience. The change of these values over time would be an interesting follow up study. To avoid the effects of COVID prejudicing our results, we picked 2019.

We compare two major modes of public transit: Yellow (Medallion) Taxis and Green (Boro) Taxis. Yellow Taxis are licensed to pick up passengers anywhere in NYC, while Green Taxis are limited to the Outer Boroughs and Manhattan North of East 96th and West 110th. Because location data for For Hire Vehicles (FHVs) is highly incomplete, our investigation did not extend to ride share companies such as Uber and Lyft.

2019 saw 83 million Yellow Taxi rides and nearly 6 million Green Taxi rides. As seen in Table 1, Yellow Taxi rides are heavily concentrated to Manhattan, where as Green Taxi rides are more evenly distributed throughout the Boroughs. Perhaps unsurprisingly, Staten Island (as an island separate from the city) sees the smallest fraction of Taxi rides.

	Yellow Cabs	*Green Cabs*
Total Rides:	*83,219,960*	*5,991,544*
Manhattan	91.70%	34.7%
Queens	6.89%	29.5%
Brooklyn	1.21%	28.7%
Bronx	0.19%	7.0%
Staten Island	0.004%	0.05%
Newark	0.01%	0.002%

Table 1. Summary statistics of which boroughs taxi rides happen in.

In Fig. 2 we show the distribution of rides in 2019 for both types of Taxis across the city. 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 Pareto 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, when required plan to focus our presentation to a fraction of zones. To determine this, we sorted by ride volume and sought a clear decision boundary.

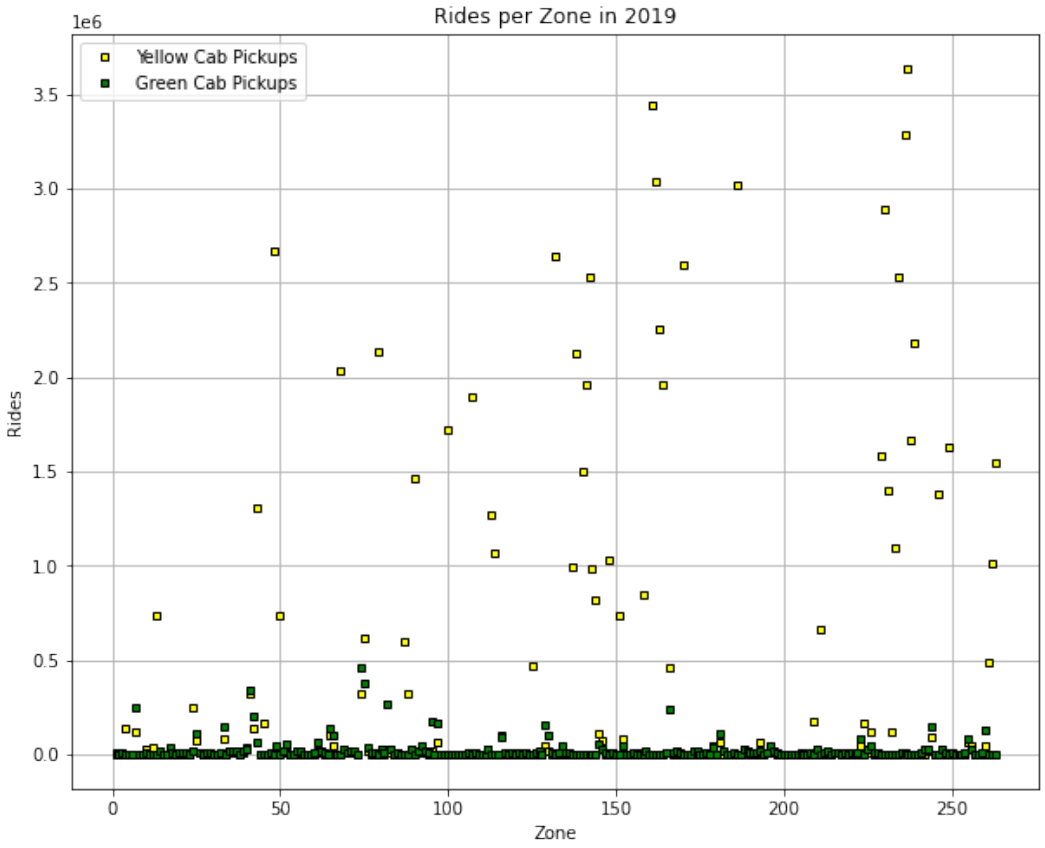

Fig. 2. Summary statistics from 83 million Yellow cab rides and 6 million Green cab rides in 2019 by zone.

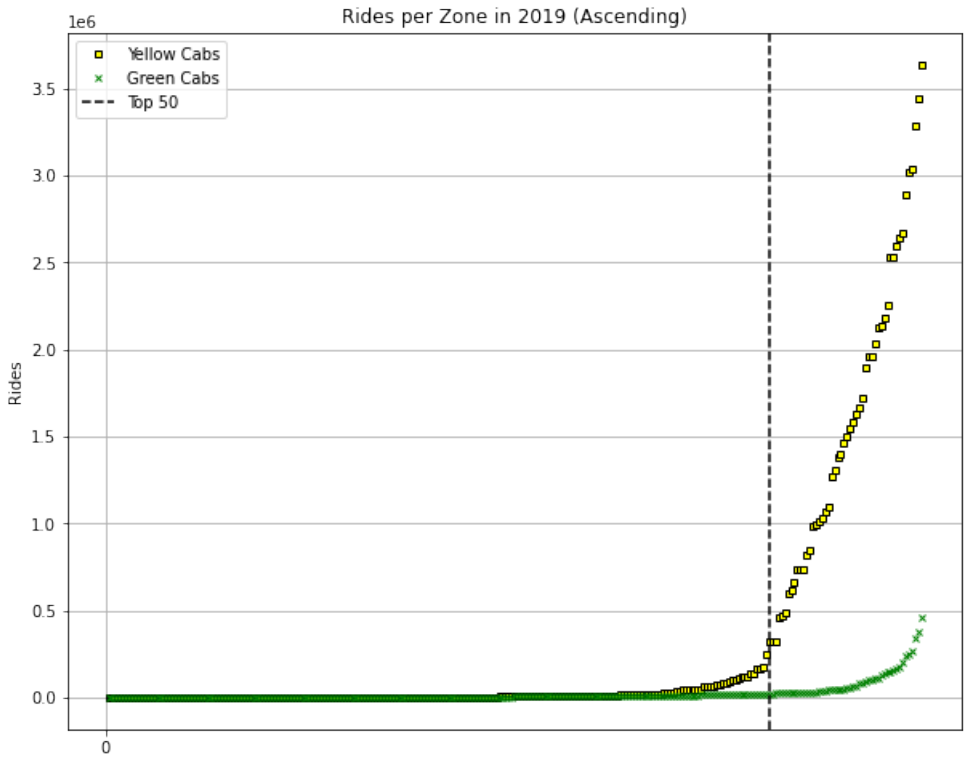

Fig. 3. Yellow and Green cab rides from 2019 by zone, arranged in ascending order. When displayed like this, the Pareto nature of this distribution is clear.

A clear decision boundary emerged around the top 50 rides. As seen in Fig. 3, 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. Fig. 4 shows the 50 busiest zones in NYC in 2019.

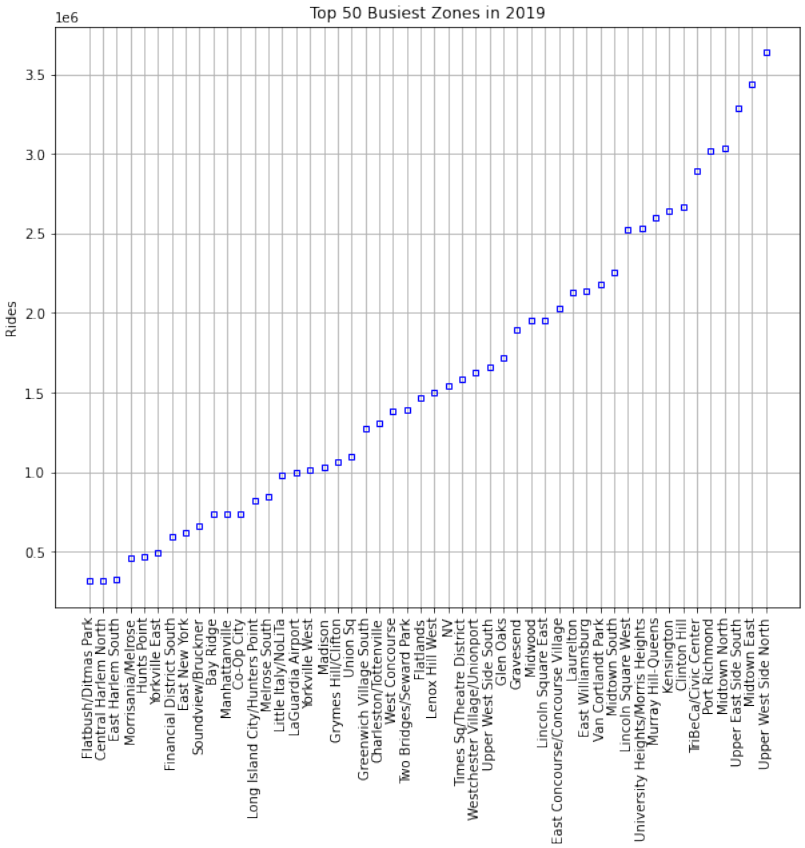

Fig. 4. The top 50 busiest taxi zones for pickups in NYC for 2019

We extended our Departure-Arrival Association analysis to both datasets in 2019 in order to learn the associations of unique pairs. Using the algorithm described previously, the ride coefficient (r.c.) for each combination of departure an arrival Understanding the results is a complex problem, are there are essentially 70,000 "answers". The total LOG results are given in Fig. 5, which is also shaded to aid in visualization. Positive values indicate a stronger association between a given pickup and drop-off pair, negative values indicate the opposite.

One interesting class of association pairs are the self-self departure arrival sets: rides which originate in and terminate in the same zone. For both Yellow Cabs and Green Cabs, the LOG(r.c.) was positive for *every* Zone. This is either an indication of the short-range, local nature of taxi rides in NYC or a systemic error in data collection. The top 5 zones for self-self association were Ellis Island, Eltingville, Rossville, Charleston and New Dorp: all located on islands. In this context, it makes sense they'd have high self association.

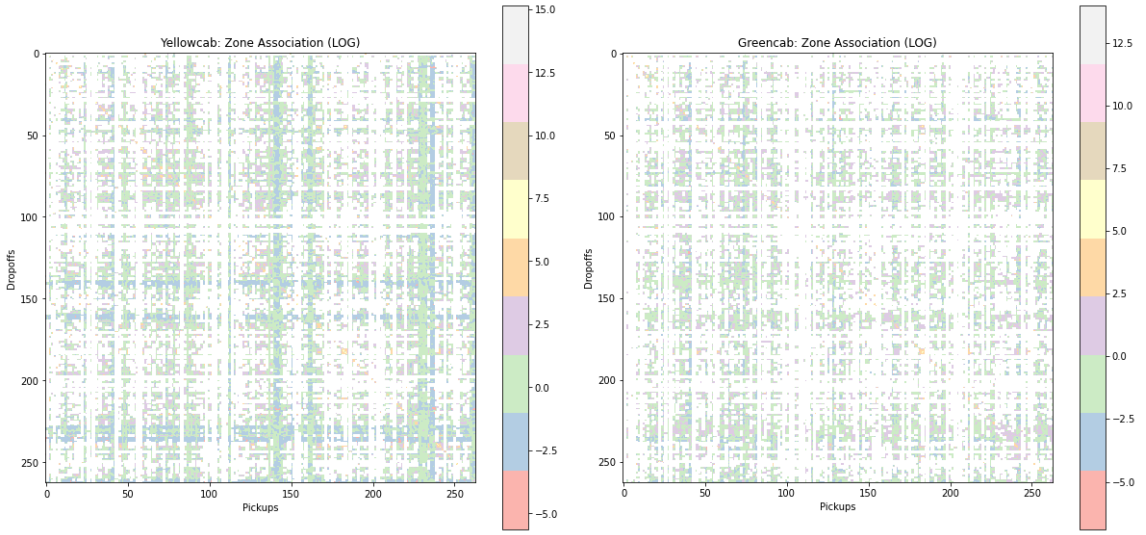

Fig. 5. The log values of ride coefficients (ZONE pickup/drop-off associations) for Green and Yellow cabs in 2019. A value of 0 indicates no particular association between a given pickup zone and drop-off zone. A positive value indicates that the indicated pickup zone and drop-off zone appear together more frequently than you'd expect, while a negative value indicates the opposite.

NYC is organized into subsections called "Boroughs", of which each zone belongs to one of five. By calculating the association values for Boroughs, the results are a little more straight forward to interpret. A modification was made to the algorithm above where zones were converted to Boroughs. The results of this study are given in Fig. 6.

These results demonstrate the intra-borough nature of taxi transit in NYC. For both Yellow and Green cabs, self association was very high for every borough. Yellow cabs generally seem to be more relied on for inter-borough travel.

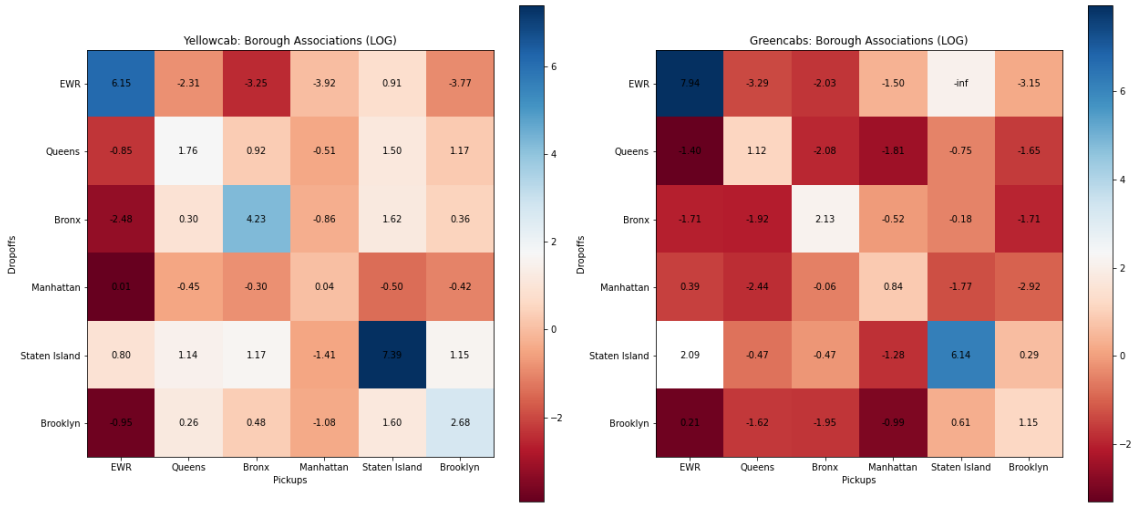

Fig. 6. The log values of ride coefficients (BOROUGH pickup/drop-off associations) for Green and Yellow cabs in 2019. A value of 0 indicates no particular association between a given pickup zone and drop-off zone. A positive value indicates that the indicated pickup zone and drop-off zone appear together more frequently than you'd expect, while a negative value indicates the opposite.

Applications

Understanding the association between a pickup and drop-off location is extremely valuable. Sometimes, as with the case of many of the islands, geography imposes a strong self-association to a zone, but in other cases the cause is not apparent. Subject matter experts would be critical for interpreting these results and explaining the associations we see in the data.

5.2 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, and some are more interesting than others. Tipping is particularly useful for it's supposed connection to the perceived quality of of a ride from the perspective of a passenger.

Preliminary analysis included creating the correlation matrix as a way of deciding where to go next. The factors with the strongest positive correlation to tip amount are the total amount (a trivial connection, as it includes the tip amount), the fare amount, and the tolls amount. The strongest correlation outside of fares is the Ratecode ID. Upon further research, we discovered that the airports have a different fare than standard taxi rides - potentially a connection. This helped guide the rest of our research.

From there, we began to look more deeply into the data. Since this data has a spatial component, we hoped 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 were beginning to emerge. For example, the pickup location with the highest mean ($10.63) and second-highest median ($10.00) was Newark Liberty International Airport - a location outside of New York City. The same held for this airport as a dropoff location, with a mean of $12.38 and a median of $15. However, these trips composed an extremely small part of the total dataset - about 0.18%. The trend

of high airport tips did continue, however. John F. Kennedy International airport, located in Queens, was the pickup location for about 3% of all trips during this month. Interestingly, the mean of $5.67 tip was substantially higher than the global mean of $2.22. The same held for the third airport: LaGuardia's mean was $5.43 while accounting for slightly less than 2% of trips. Did this represent vacationers who are less clear on the expected tips? Or were these higher tips actually a function of some other variable correlated with the airport?

It was here that we considered a simple answer - NYC charges a much higher ratecode for trips to the airport. This meant that for any given airport trip, the total cost was going to be higher - thus explaining the strong positive correlation between total fare and Ratecode ID. So it seemed clear that there was a connection between higher total fare and higher tips.

However, there was another way to approach the problem, one which was, in many ways, more intriguing. What about the amount of tip as a percentage of the fare? This is where the tipping ratio came into play. Below is a chart with the tip ratio sorted from highest to lowest:

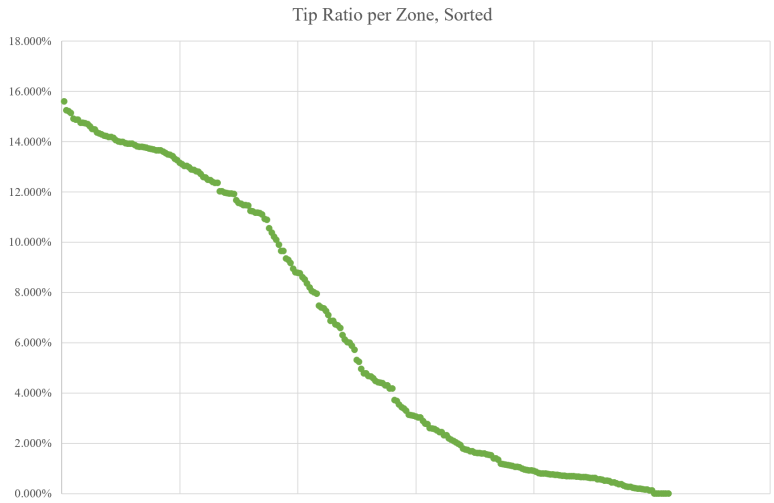

Fig. 7. Distribution of tip ratio versus pickup zone, sorted from highest to lowest. Two high outliers with a collective support of 0.00045% were removed.

Despite the high ratecode of the airports, they still had relatively high average tip ratios - LaGuardia's tips were 15.142% of the total fare, Newark's were 12.456%, and JFK had 11.926%. This seemed to imply that customers were willing to tip highly for being picked up at the airport, even though they would charged more overall. There were several more interesting nuggets to find beyond just the airports, however. Another question was whether there was any connection between support and tip ratio. In other words, do the more heavily utilized zones have higher tips?

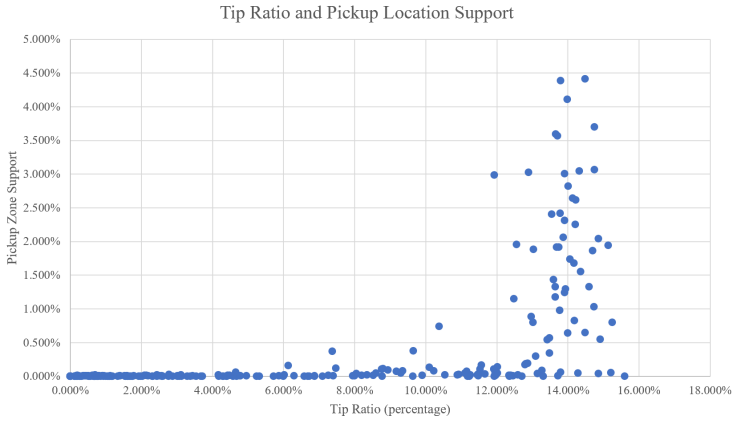

Fig. 8. Scatterplot comparing the tip ratio and the support of trips to this zone. Two high outliers with a collective support of 0.00045% were removed.

The answer was yes. It does appear that the tip ratio is more a function of popularity as opposed to total fare. In fact, the top 50 pickup zones all hovered in 12-14% range of the total fare. There were a few notable exceptions to this: East Harlem South was just above 10%. Furthermore, Central Harlem and East Harlem North were both below 10%, with the latter dropping all the way down to 7.37% - substantially lower than anything else in the top 50. The simple reasoning would be that the Harlem neighborhood just happens to have lower tips. However, while Harlem is low compared to the top 50, it is not low compared to other, less-supported locations.

The reasoning for this phenomenon is not immediately clear - why is it that more popular locations tend to tip higher? One possible idea is that in these locations people are more likely to know what the appropriate tipping amount is. It would make sense that an area with heavy cab usage would have a better sense of the etiquette of tipping. Another option is that, for some reason, the cab service is substantially worse in the less-popular areas. If this were the case, then low tipping would make sense, as it would then serve as a proxy for customer satisfaction. We were not able to find a way to determine if either of these were the case using this available data.

Applications

The potential applications of this are quite rich, especially if tips do in fact serve as a proxy for rider satisfaction. A taxicab company could look at this data to see what areas/factor cause customers to be happier with their rides. Another way this could be applied would be to look for trends over time. Such an extension which would be interesting to do in the future. While we mined some interesting financial data, an expert in the field may be necessary for us to truly interpret what this data means.

5.3 Ride Visualization and COVID

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.

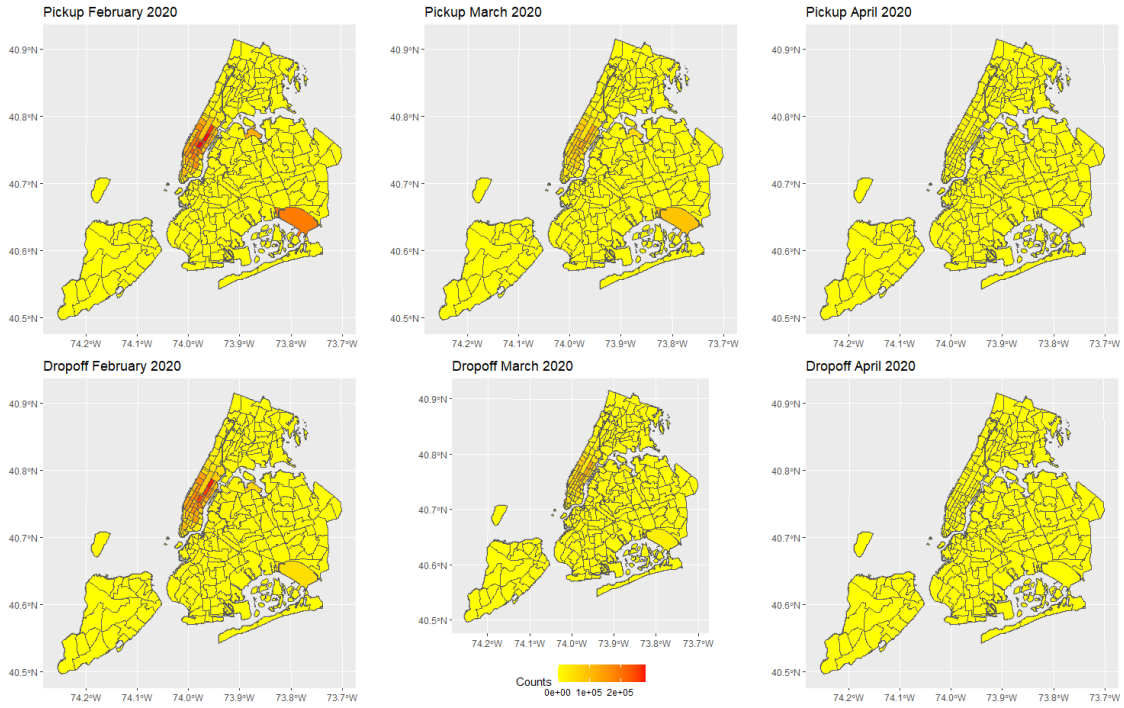

Fig. 9. Number of Taxi rides total per month in each of the taxi zones in NYC for the months of February, March, and April 2020

The map pictured in figure 4 indicates some interesting findings, and suggests that more granular time periods would indicate more interesting trends.

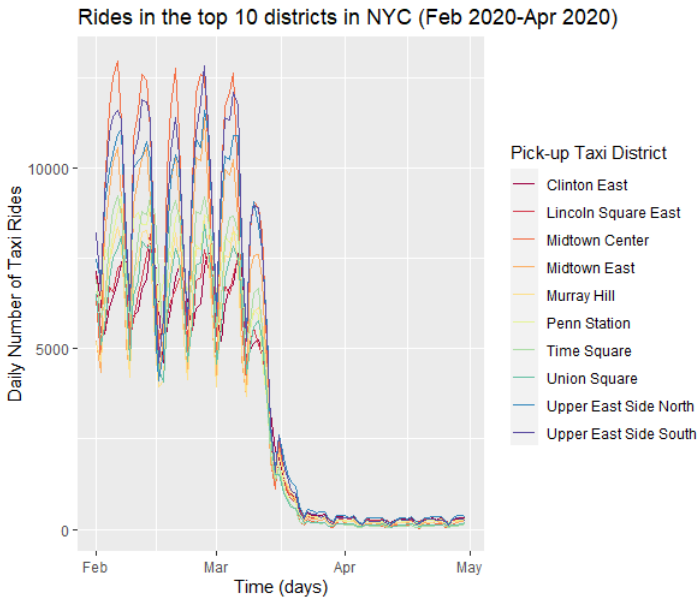

Fig. 10. Number of daily rides in the 10 busiest taxi districts during the initialization of the lockdown in New York City.

At a more granular time period (daily) two things can be drawn. The normal daily rides throughout the week typically fluctuate in similar patterns for each of the busiest 10 taxi districts. The second is the significant effect that the lockdown due to COVID-19 had on yellow-cab travel. Districts with well over 10,000 rides per day were down below 100 rides per day. COVID-19 left a devastating impact on the New York City economy, and yellow-cab data from the early pandemic certainly confirms the impact that it had on every aspect of life in New York City.

Applications

As a result of the visualization it was learned that 1) there was a significant impact from the COVID-19 lockdown, and 2) we should begin to ask questions about using yellow cab data as an indicator of greater economic impacts and fluctuations in New York City. There may be answers to questions about economic prediction and stability that can be mined in the yellow cab data.

REFERENCES

[1] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420). New York: United Nations.

[2] https://www.kaggle.com/c/nyc-taxi-trip-duration

[3] Chien-Ming Tseng and Chi-Kin Chau. 2017. Viability Analysis of Electric Taxis Using New York City Dataset. In Proceedings of the Eighth International Conference on Future Energy Systems (e-Energy '17). Association for Computing Machinery, New York, NY, USA, 328–333. DOI:https://doi.org/10.1145/3077839.3078463

[4] Wickramasinghe, Chathurika S. et al. "Data Driven Hourly Taxi Drop-offs Prediction using TLC Trip Record Data." 2019 12th International Conference on Human System Interaction (HSI) (2019): 168-173.

[5] Buchholz, Nicholas. "Spatial Equilibrium, Search Frictions and Dynamic Efficiency in the Taxi Industry." (2019).