

Analysis of road-based shared transportation options in New York City

Anonymous Author(s)

ABSTRACT

To effectively design policy and aid researchers, New York City regularly publishes data from various city agencies. One agency of interest is the Taxi and Limousine Commission (TLC), who has published detailed data on taxi trips and over the last decade. From 2019 onwards, due to a city law concerning high-volume for-hire vehicle services such as Uber and Lyft, the commission has also published data from these services. These data items constitute a large and rich dataset, where many interesting analyses can be performed.

To this end, the authors performed a number of analyses on this data and identified interesting trends and patterns. Before performing data analysis, they cleaned up and standardized said data, ensuring all data adhered to a common set of columns. After storing this standardized data into a series of Parquet files, they performed exploratory data analysis to identify interesting trends and guide the direction of their project. Finally, they took the most interesting trends and created a Svelte-based visualization webapp, allowing end users to easily and interactively view selected trends and statistics. They conclude with an overview of legal and ethical considerations related to the analysis of this data.

KEYWORDS

taxi, for-hire vehicles, big data, Manhattan traffic pattern, NYC traffic analysis, TLC metadata analysis, Uber Traffic Congestion, TLC data privacy, ride-sharing, transportation optimization, NYC taxi parquet Files, urban movement analysis

1 MOTIVATION AND GOALS

The traffic situation in NYC has been getting worse in the past 20 years. One of the main reasons that is believed to be contributing to NYC traffic is the emergence of Uber, the app-based for-hire Vehicle Company (FHV) company [8]. Since 2011, Uber has added 18,000 affiliated cars to streets which many people believe has contributed to traffic patterns in NYC [8]. This project will look at the NYC taxi dataset which has 3.4 billion rows representing taxi rides. Each ride is characterized by 20 features such as location, payment methods, time and date. This report will look at February 2015 to May 2022, when a higher volume of data was collected compared to the previous years. The collection of such high-volume data provides a great opportunity to perform a comprehensive data analysis of the NYC Taxi Industry.

This report is organized as follows. In section 2, the report will discuss the work done by other researchers in the similar area, which looked at whether Uber rides contribute to congestion. Section 3 of the report will include the data engineering portion of the report. This section expands on the overall structure of the dataset which was originally sorted for Apache Hadoop and how DuckDB is used to query the data for the use for data analysis.

Section 4 presents the data analytics methods used for exploratory data analysis and data visualization.

The goal of the report is to build on the existing research using new data published since the other articles were published. Throughout this report, various analytics techniques will be used to extract statistical information that reveals specific traffic patterns using TLC ride metadata.

2 RELATED WORK

In 2015, in response to statements made by both the city of New York City and Uber on the impact of ride sharing platforms, FiveThirtyEight published a series of articles based on Uber ride data in New York City obtained via FOIA request. To verify whether Uber rides contribute to congestion, Carl Bialik and Mehta [6] compare Uber and taxi trip data. The authors find that Uber serves boroughs outside Manhattan more than yellow and green cabs. They also find that Uber's highest concentration of pickups occur South of 59th Street, the same as taxis, providing evidence for the congestion claim. Silver and Fischer-Baum [16]'s analysis concludes that both Ubers and taxis disproportionately serve wealthier communities in NYC, identifying a niche of primarily public transport users that would stand to benefit from using a ride hailing service over owning a car. Fischer-Baum and Bialik [11] base their analysis on Uber data from January to June 2015, a period when Uber was aggressively entering the market. They show during this period, Uber supplanted taxis both within and outside of Manhattan. Carl Bialik and Mehta [5] analyzes Uber data by time of day to see if pickup timing contributes to rush hour traffic. They find that in Manhattan, starting at 4PM, Uber adds slightly more rides than taxis lose, causing a very minor increase in traffic. However, outside Manhattan, during all hours of the day, Uber adds a substantial number of cars on the road, increasing congestion.

In a similar vein, [?] use a combination of taxi trip data from the NYC Taxi and Limousine Commission (TLC) and New York Times articles to analyze how Uber has changed the NYC taxi industry. They confirm that Uber has led to a decrease in the use of taxis. Furthermore, the presence of Uber has led to an increase of complaints surrounding taxis, indicating the existence of the service has led to behavioral changes in customers.

More broadly, several papers have used taxi data to inform policy and decision making. Correa and Falcocchio [7] assess the effectiveness of an Adaptive Traffic Control System (ATCS) implemented in a congested area of Manhattan from 2012 to 2016. Using geolocation data provided by the TLC, they find that traffic speed increased in the year following the creation of the ATCS, but this increase was not sustained due to outside factors such as the installation of bike lanes and increases in tourism. Wickramasinghe et al. [17] use TLC taxi data from 2017 to 2018 to train a random forest model for predicting taxi rides at a given location and time. Yu [18] combines TLC taxi data with weather data from OpenWeatherMap in 2015

to analyze the effect of snowy weather on taxi prices. Through correlation analysis, they find that snowy weather may actually lower the cost of a taxi ride. Atkinson-Palombo et al. [3] use TLC for-hire vehicle data to analyze the outsized usage of ride sharing services in boroughs outside of Manhattan. Given that many poor neighborhoods are located outside of Manhattan, they identify problems with the primary mode of travel in these areas hinging on a for-profit company and recommend policy makers to take such areas into account when designing regulatory policy.

Prior end-user systems have been designed for reading and analyzing TLC taxi data. Aziz and Robila [4] describe a system for querying yellow and green taxi data in NYC using data from the TLC. Their system describes a dashboard to easily view data and provides a number of useful graphs, such as monthly trip counts.

3 DATA ENGINEERING

New York State has several different sets of data - one corresponding to each type of ride:

- (1) Yellow cab data (servicing Manhattan) is the only one that goes all the way back to 2009 - the first available data in the entire dataset.
- (2) Green cab data (servicing the bronx) were added in 2010-2011
- (3) "For hire vehicle" or FHV data was added after that, likely in response to a growing demand for rideshare apps. There is also a "high velocity" FHV dataset for those rideshare providers serving over 10,000 trips per day.

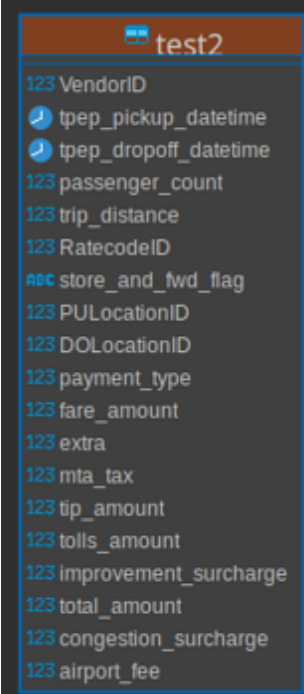
Each of these datasets consists only of a single table in a .parquet file format. Each type of ride has one parquet file per month to keep the individual files relatively small, however the entire data dump across all ride types as of late-September 2023 contains 3,377,845,468 rows in total, so we are quite confident that this qualifies as a "big data" dataset.

3.1 Data Architecture - The Parquet File Format

Upon further reading, the parquet file format [12], seems particularly well suited for querying big datasets such as this one because it was originally created "for use in Apache Hadoop... as a shared standard for high performance data IO." This origin in the high-performance data world, as well as its column-oriented storage scheme, means parquet files, are essentially already single-file databases, much like the SQLite format.

In addition, some Python libraries such as duckdb already support querying parquet files directly [14], while another blog post from influxdb [9] talks about how these queries can sometimes be made with millisecond latency by taking advantage of the design features of parquet file format.

We have noticed that this solution scales relatively well on even moderately high performance machines such as queeg. The natural monthly partitioning of the individual parquet files helps keep them down to a reasonable size for processing by one machine (as opposed to trying to hold all 3.7 billion entries (50+ GB) of data in memory all at once). This, combined with the naturally tabular structure of parquet data, seems like quite a reasonable solution that, with some additional parallelism (more machines and/or multithreading), could become quite a performant solution for the relatively large queries we are making for this project. This



Column Name
123 VendorID
123 tpep_pickup_datetime
123 tpep_dropoff_datetime
123 passenger_count
123 trip_distance
123 RatecodeID
123 store_and_fwd_flag
123 PULocationID
123 DOLocationID
123 payment_type
123 fare_amount
123 extra
123 mta_tax
123 tip_amount
123 tolls_amount
123 improvement_surcharge
123 total_amount
123 congestion_surcharge
123 airport_fee

Figure 1: A listing of the columns in the schema for the yellow taxi table.

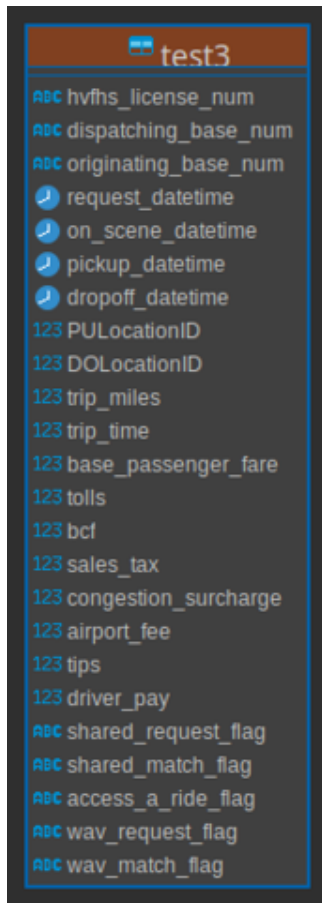
naturally partitioned file structure helps effectively force us to set up the data processing in a way that is scalable because the nature of having separate files makes it convenient to do them one at a time, closing each after it is no longer needed to free resources.

3.2 Schemas and Data Cleaning

Apart from the first few years of yellow cab data, the table schemas are quite consistent between each of the monthly files, requiring relatively little cleanup of the schemas. New York State loosely documents the schemas themselves in a set of PDF documents available on the same webpage [13] where the data files are available. These PDF files contain a table listing each of the column names with a description of what the data in that column represents (such as what units it is in). These documents made it easy to find the newest column names being used for the yellow cab data. An example of these schemas as pulled from the parquet files themselves can be seen in Figures 1 and 2.

Even just from the limited segments of the 3.7B rows of data that we have seen using data viewing tools such as Tad and DBeaver (in addition to our duckdb queries), we noticed that, while most of the data is largely consistent and already relatively clean, there are still some things we have noticed that need some cleanup. Below is an outline of the steps in our cleanup process:

- (1) As previously mentioned, the earliest data from the yellow cab dataset has some column names that don't match the column names used for the majority of the data. For most

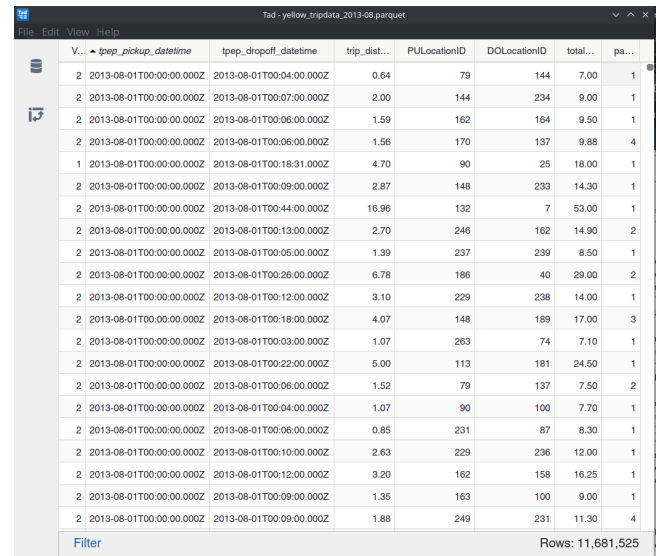


Column Name	Data Type / Notes
hvfhs_license_num	ADC
dispatching_base_num	ADC
originating_base_num	ADC
request_datetime	clock icon
on_scene_datetime	clock icon
pickup_datetime	clock icon
dropoff_datetime	clock icon
PULocationID	123
DOLocationID	123
trip_miles	123
trip_time	123
base_passenger_fare	123
tolls	123
bcf	123
sales_tax	123
congestion_surcharge	123
airport_fee	123
tips	123
driver_pay	123
shared_request_flag	ADC
shared_match_flag	ADC
access_a_ride_flag	ADC
wav_request_flag	ADC
wav_match_flag	ADC

Figure 2: A listing of the columns in the schema for the high volume for hire vehicles table.

(but not all) columns, these names can be fixed with a simple column rename operation.

- (2) Some of these modified columns above also contain data in a format that is inconsistent with the format currently in use. These include (but are not limited to):
 - (a) Older versions of the column representing the method by which people paid for their trip represent the data as a mixed-case text string containing values such as “CREDIT”, “Credit”, “Cre”, “CRE”. etc just for a credit card transaction. In more recent data from this dataset, this column is an integer (documented in the previously mentioned PDF) that represents the type of transaction
 - (b) Previously the locations of the pickup and dropoff locations were represented as latitudes and longitudes, but recent data represents this as an integer that identifies the taxi zone (as shown in Figure 5). We will need to convert these to the taxi zones in the older data.
 - (c) Many of the columns also contain empty values, presumably because either those columns were added to the data later, or because the empty values. We don't yet know whether documentation is available for these older



V...	tpep_pickup_datetime	tpep_dropoff_datetime	trip_dist...	PULocationID	DOLocationID	total...	pa...
2	2013-08-01T00:00:00.000Z	2013-08-01T00:04:00.000Z	0.64	79	144	7.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:07:00.000Z	2.00	144	234	9.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:06:00.000Z	1.59	162	164	9.50	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:06:00.000Z	1.56	170	137	9.88	4
1	2013-08-01T00:00:00.000Z	2013-08-01T00:18:31.000Z	4.70	90	25	18.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:09:00.000Z	2.87	148	233	14.30	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:44:00.000Z	16.96	132	7	53.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:13:00.000Z	2.70	246	162	14.90	2
2	2013-08-01T00:00:00.000Z	2013-08-01T00:05:00.000Z	1.39	237	239	8.50	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:26:00.000Z	6.78	186	40	29.00	2
2	2013-08-01T00:00:00.000Z	2013-08-01T00:12:00.000Z	3.10	229	238	14.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:18:00.000Z	4.07	148	189	17.00	3
2	2013-08-01T00:00:00.000Z	2013-08-01T00:03:00.000Z	1.07	263	74	7.10	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:22:00.000Z	5.00	113	181	24.50	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:06:00.000Z	1.52	79	137	7.50	2
2	2013-08-01T00:00:00.000Z	2013-08-01T00:04:00.000Z	1.07	90	100	7.70	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:06:00.000Z	0.85	231	87	8.30	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:10:00.000Z	2.63	229	236	12.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:12:00.000Z	3.20	162	158	16.25	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:09:00.000Z	1.35	163	100	9.00	1
2	2013-08-01T00:00:00.000Z	2013-08-01T00:09:00.000Z	1.88	249	231	11.30	4

Figure 3: A snapshot of a tabular view of the yellow taxi dataset from August 2013

schemas to help explain what the intended meaning of this empty data is. If we are unable to find it, we may need to look closer at the data and make some reasonable assumptions

While we spent significant amounts of time working on trying to clean the data, we ultimately ended up pivoting to using a more limited dataset that doesn't include this 2009-2010 data for the sake of time and completeness. This eliminated a large amount of the work that would have otherwise been needed to be done before we could query over all the data.

4 DATA ANALYTICS

Our strategy for the exploratory data analysis portion of this section is a top-to-bottom approach. The first data exploration that we would do is to find the total number of rides per year from 2015-2022. This data exploration enables us to have general intuition about the shift in people's preferences for their primary form of transport. A closer data exploration that we plan on doing is the average speed per year. These two data explorations help find interesting patterns. To visualize our data exploration, we will be building a Svelte-based web application for end users to easily visualize trends we wish to highlight. We plan on using third-party libraries to easily add visualizations using bar graphs.

Even though Figure 5 might look like it is always a better idea to take a ride-share service to get to any destination in NYC since the average slowest speed is 12 miles/hour around 3-6 pm given that a person walking speed is 3 miles/hour. However, by looking at a different visualization form such as Figure 6, we can see that it could be viable to walk to a destination in certain regions in Manhattan. For example, in East Chelsea, the average speed of a car is about 6.6 miles per hour, making walking a much more attractive choice. The regional visualization 6 highlights the importance of having a different type of visualization in our data analysis.

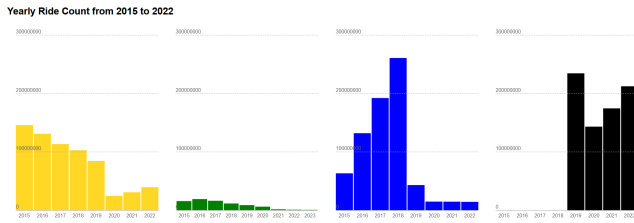


Figure 4: Ride counts per year for yellow taxi, green taxi, for-hire ride-share, high volume ride-share,

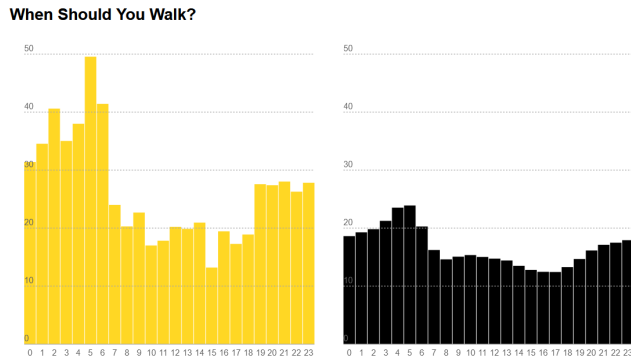


Figure 5: Average ride speed over 24 time period for yellow taxis and high volume ride-share



Figure 6: Darker regions correspond to slowest speed region, Lighter region corresponds to faster speed region

All of these visualizations fetch data from a static server. To reduce latency and increase efficiency, we plan on precomputing all statistics offline and passing them to the frontend as JSON files.

Svelte is a reactive frontend framework, making it simple to have end users interact with our data. We can leverage this by creating interactive visualizations such as showing how certain statistics change over time. By combining this with LayerCake,

a visualization library, we can create highly customizable visualizations for our users. The current application can be seen at <https://cs.rit.edu/~bsg8294/nyc-taxi/>.

5 LEGAL CONSIDERATIONS

In this section, we highlight relevant laws that apply to the creation and use of our dataset. Given that our data is produced by US-based entities (i.e. Uber, Lyft, private NYC taxi companies), distributed by a US government entity (i.e. the NYC TLC) and our product's intended audience resides in NYC (e.g. policy makers and NYC citizens), we focus primarily on US — particularly NYC — laws and legal issues.

To begin, it is useful to understand precisely *how* data is collected, and *which* entities store and process said data.

Taxis have historically been required to carry meters that charge passengers based on the distance travelled. After a 2016 decision by the United States Second Circuit Court of Appeals, which decided that mandating GPS installation did not violate the rights of taxi drivers, the city required all taxis to install said equipment [15]. Thus, all yellow and green taxis directly electronically send detailed trip data to the NYC TLC, such as pickup and dropoff locations, the price of a trip, and the distance travelled.

The NYC TLC also oversees and collects data from for-hire vehicles (FHV), which include livery vehicles (not to be confused with *street hail* livery vehicles, another name for green cabs), black cars, and luxury limousines. Companies that operate these vehicles are required to send to the TLC monthly reports regarding information such as pickup and dropoff times and locations [1]. This data does not have to be collected via electronic methods such as GPS, *per se*. If vehicles do opt to collect passenger personal information (e.g. names or credit card details) or geolocation data, they must file an information security and use of personal information policy with the TLC, which includes to what extent trip details are being collected and used, and a contingency plan in event of a data breach. If such a breach does occur, the company must notify the commission immediately.

Under the for-hire vehicle umbrella, but unique enough to be discussed in isolation, are high volume for hire vehicle (HVFH) services such as Uber or Lyft. Due to their high ride counts (HVFH services are defined as performing over 10,000 trips per day), and the massive amount of data they store and process, they are regulated differently from other FHV services. For one, not only is the timeframe for data reports shortened from one month to two weeks, but the granularity of data collected is far greater. As an example, the location of all available drivers must be collected at 60 second intervals [1]. This data is legally required to be collected via GPS. All this being said, the information security and use of personal information policy required by the TLC has not greatly increased in scope from standard FHV companies.

Due to the massive scale of data collected by the TLC and FVHF companies, it is critical for these organizations to ensure data breaches do not happen, and when they do, that people who have had their data breached have recourse. An example of a failure to do so is with Uber in 2016, when a data breach affected 57 million customers. The information breached included name, email

address, phone numbers, and even driver's license number. Surprisingly, Uber did not report the incident, which led Uber to enter a non-prosecution agreement with the Federal Trade Commission (FTC) [10].

6 ETHICAL CONSIDERATIONS

In this section, we discuss the ethical underpinnings of the work performed for this project. In particular, we highlight the struggle between respecting user privacy and acting in the public's interest.

Section 1.6 of the ACM Code of Ethics [2] states that computing professionals should respect privacy. For computing professionals, "privacy", i.e. the usage and processing of others' information, is a multifaceted concept that must be dealt with in different ways. To begin, data should not be collected unless there is a legitimate reason for doing so. If data is collected, it should be done with the consent of users. Of the data that is collected, it should be minimized such that a breach would reduce harm or embarrassment to users. Finally, when releasing data, it should be anonymized such that consumers cannot link sensitive information back to the originator of the data.

Starting with consent, it is debatable to what extent vehicle operators and passengers actively consent to their data being harvested by HVFH companies and the TLC. One could argue that by choosing to work as a taxi driver, or by choosing to hail down a taxi, drivers and passengers implicitly agree to give up this data. However, the act of needing to work as a driver or use such transportation options in the first place may be out of the hands of said people. If driving is the only work they can find, the only other option for drivers is to starve. And as pointed out in Silver and Fischer-Baum [16], for many people, rather than car ownership, the most economical choice is to primarily use public transport with the occasional taxi or Uber ride, rather than owning a car outright.

For the most part, the TLC appears to handle data anonymization well. In particular, they maintain usefulness of the data while reducing identifiability by using taxi zones, which are coarse and are sized roughly proportionally to the density of the area, instead of longitude and latitude for pickups and dropoffs, at least for most years. In 2009, the TLC violated the principle of respecting privacy by releasing precise coordinates instead. By combining this data with other datasets, one could reconstruct information like residence addresses. While this was changed in later years, it is odd that they continue to keep the 2009 dataset around in its original form, without performing any retroactive anonymization of the data.

Moving on, Section 3.1 of the ACM Code states that one should ensure the public good is central to their work. Related, 3.7 states that special care should be taken for systems that have become integrated with society. Taken together, the Code advises computing professionals to develop systems that use data to help the public interest.

As demonstrated in our related works section, the TLC transportation data is primarily used to drive policy decisions rooted in hard data. Opening this data up to the public allows even private citizens to make important observations and contribute to discussions in a meaningful way. Our project itself is designed to help people in NYC make useful decisions in how they should structure

their commute in order to save time and money. Furthermore, it could be used to convince policymakers to adopt certain positions on bike lane zoning and taxi fare structuring.

Clearly, with greater *granularity* and greater *amounts* of data, better, more informed decisions can be made, both by policymakers and ordinary citizens. However, scaling the data must be weighed against the privacy cost. As mentioned in the previous section, the TLC has extremely detailed data on most trips under their jurisdiction, including minute by minute longitude and latitude coordinates of yellow/green cabs and HVFH vehicles. While this data could certainly be used to make useful policy decisions, it does start to feel invasive.

An example of an instance where the line becomes blurred is the 2016 decision that led to GPS being installed in all NYC taxi cabs. In 2004, the city began integrating GPS systems into cabs to identify if drivers were overcharging passengers. By 2010, the TLC found that thousands of drivers were indeed misleading passengers into how much they should be paying, placing charges on investigated drivers. The 2016 case was an effort by one of the charged drivers to claim that the investigation was unlawful; however, the judges ruled against him, leading to NYC being able to mandate GPS in all taxi cabs.

In regards to this case, on one hand, it is in the public interest to ensure passengers are not scammed by drivers. To prevent scamming, it is not enough for the city to simply use less granular forms of data; moving just a couple extra blocks within the same taxi zone will increase the fare price. On the other hand, this comes at the cost of the city knowing exactly where a person in a taxi is at all times, which, coupled with other forms of city data, could allow reconstructing a person's behavior in an intrusive way. Different people may choose to make different tradeoffs between privacy and social/personal good. Ultimately, this is an area that will always be contentious.

7 CONCLUSIONS

This writeup demonstrates our progress in analyzing and visualizing taxi and for-hire vehicle data for New York City. We were able to generate some interesting surface level analyses of the data with our current system. We hope to find some time to potentially improve it going forward in a few key ways.

This includes more interactive graphs such as ones that allow the user to select individual taxi zones as the pickup and drop-off locations and analyze traffic flow between them as well as potentially having granular control over the time ranges in the analysis. However, these analytical goals bring with them some data engineering and time management challenges. For example, in order to allow for the level of interactivity and detail that we had initially planned for our graphs and analyses, the data transmission between our databases, backend, and front end will likely need to be significantly more performant than they currently are. We may also have to dramatically re-engineer the back end to avoid using our pre-computed metrics approach that we have been using thus far. One potential solution we had to mitigate part of this problem was to use some existing solutions such as Google BigQuery to import and query our data. However, with such a large database (in excess of 50 gigabytes), we suspect that we will rapidly exceed

the \$300 free tier of this Google service, as well as the time we have remaining in the semester with which to perform this amount of re-engineering.

Overall, we have learned much about large real-world data sets, as demonstrated by the New York City taxi dataset. Additionally, we have made a small but not completely meaningless contribution to the body of ride sharing industry research by conducting our analysis using NYC taxi data that is more recent.

REFERENCES

- [1] 2023. In *THE RULES OF THE CITY OF NEW YORK*. American Legal Publishing, Chapter 59.
- [2] Association for Computing Machinery. 2018. ACM Code of Ethics and Professional Conduct. ACM, New York. <https://www.acm.org/code-of-ethics>.
- [3] Carol Atkinson-Palombo, Lorenzo Varone, and Norman W. Garrick. 2019. Understanding the Surprising and Oversized Use of Ridesourcing Services in Poor Neighborhoods in New York City. *Transportation Research Record* 2673, 11 (2019), 185–194. <https://doi.org/10.1177/0361198119835809> arXiv:<https://doi.org/10.1177/0361198119835809>
- [4] Zahid Aziz and Stefan Robila. 2019. Interface for Querying and Data Mining for NYC Yellow and Green Taxi Trip Data. In *2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*. IEEE, Long Island, New York, 1–7. <https://doi.org/10.1109/LISAT.2019.8817347> ISSN: 2642-8873.
- [5] Reuben Fischer-Baum Carl Bialik and Dhruvil Mehta. 2015. Is Uber Making NYC Rush-Hour Traffic Worse? | FiveThirtyEight. <https://fivethirtyeight.com/features/is-uber-making-nyc-rush-hour-traffic-worse/>
- [6] Reuben Fischer-Baum Carl Bialik, Andrew Flowers and Dhruvil Mehta. 2015. Uber Is Serving New York's Outer Boroughs More Than Taxis Are | FiveThirtyEight. <https://fivethirtyeight.com/features/uber-is-serving-new-yorks-outer-boroughs-more-than-taxis-are/>
- [7] Diego Correa and John C. Falcocchio. 2022. A Data-Driven Case Study Following the Implementation of an Adaptive Traffic Control System in Midtown Manhattan. *Journal of Transportation Engineering, Part A: Systems* 148, 4 (2022), 05022001. <https://doi.org/10.1061/JTEPBS.0000645> arXiv:<https://doi.org/10.1061/JTEPBS.0000645>
- [8] American Safety Council. [n.d.]. Is Uber Affecting New York Traffic? <https://blog.americansafetycouncil.com/is-uber-affecting-new-york-traffic/#:~:text=Many%20people%2C%20including%20New%20York's,Uber%20to%20the%20city's%20congestion>
- [9] Raphael Taylor-Davies Andrew Lamb / Dec 07, 2022 / InfluxDB IOx, and Community. 2022. Querying Parquet with Millisecond Latency. <https://www.influxdata.com/blog/querying-parquet-millisecond-latency/>
- [10] Lesley Fair. 2022. FTC addresses Uber's undisclosed data breach in new proposed order. <https://www.ftc.gov/business-guidance/blog/2018/04/ftc-addresses-ubers-undisclosed-data-breach-new-proposed-order>
- [11] Reuben Fischer-Baum and Carl Bialik. 2015. Uber Is Taking Millions Of Manhattan Rides Away From Taxis | FiveThirtyEight. <https://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/>
- [12] Apache Software Foundation. 2016-2023. Reading and Writing the Apache Parquet Format — Apache Arrow v13.0.0. <https://arrow.apache.org/docs/python/parquet.html>
- [13] New York City Taxi and Limousine Commission. [n.d.]. TLC Trip Record Data - TLC. <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
- [14] Hannes Mühleisen and Mark Raasveldt. 2021. Querying Parquet with Precision using DuckDB. <https://duckdb.org/2021/06/25/querying-parquet.html>
- [15] Christopher F. Droney Rosemary S. Pooler, Debra Ann Livingston. 2016. Hassan El-Nahal, individually and on behalf of all others similarly situated, Plaintiff-Appellant, v. David Yassky, Commissioner Matthew Daus, Michael Bloomberg, The City of New York, Defendants-Appellees. United State Court of Appeals, Second Circuit.
- [16] Nate Silver and Reuben Fischer-Baum. 2015. Public Transit Should Be Uber's New Best Friend | FiveThirtyEight. <https://fivethirtyeight.com/features/public-transit-should-be-ubers-new-best-friend/>
- [17] Chathurika S. Wickramasinghe, Daniel Marino, Fatih Yucel, Eyuphan Bulut, and Milos Manic. 2019. Data Driven Hourly Taxi Drop-offs Prediction using TLC Trip Record Data. In *2019 12th International Conference on Human System Interaction (HSI)*. IEEE, Richmond, USA, 168–173. <https://doi.org/10.1109/HSI47298.2019.8942633> ISSN: 2158-2254.
- [18] Xuqiao Yu. 2021. Big Data Driven Model for New York Taxi Trips Analysis. In *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*. IEEE, online, 1–4. <https://doi.org/10.1109/ICBDA51983.2021.9403054>