Exploratory Data Analysis of New York City Taxi and Limousine Service Trip Record Data

Abstract

This paper conducts exploratory data analysis on the publicly available New York City Taxi and Limousine Commission (TLC) trip record data to identify trends in market share, travel times and congestion surcharge that provide directions for further research.

The Data

The official TLC trip record stores open-source trip data pertaining to licensed taxi and for-hire vehicle trips in New York since 2009.

This paper will focus on exploring trips in 2019 made by the New York's iconic yellow taxis, borough taxis (i.e. green taxis, street hail liveries), as well as rides completed by companies operating under the For Hire Vehicle High Volume license (Uber, Lyft, Via, and Juno). In totality, this subset of the data describes approximately 350 million trips, and was chosen for its recency, completeness, as well as the relevance of the largest ride operators. These traits make the data a good candidate for depicting the state of the New York ride-hail industry, and for making post-COVID inferences. The more recent 2020 trip records are incomplete due to relaxed data reporting requirements in recognition of the impact of COVID-19.[1] While this data—in its complete form—could also give insight into what the New York taxi industry might look like post-pandemic, it was not used for this analysis due to complexity in multiple confounding factors associated with it.

Scope and Stakeholders

The outcomes of this exploratory data analysis will be of interest to drivers, ride-hail companies, regulators, and travellers alike.

Patterns in market domination are of interest to the newer entrants to the ride-hail market—such as Lyft, Via and Juno—to inform their marketing and expansion strategy. Likewise, more experienced players such as Uber will be on the lookout to protect their market share, and traditional yellow and green taxi operators need this information to remain relevant in this rapidly changing industry.

[1] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Travel duration, especially between and within key economic areas (e.g. Manhattan and airports), are in regulator interests for improving economic outcomes for the region. These metrics are useful to assess the success of various regulatory incentives e.g. the congestion surcharge with aims of reducing congestion within Manhattan,[2] or the whether the JFK airport flat rate is harmful or beneficial to this. These travel duration metrics are also in the consumer interest of selecting the best travel option due to the different duration-dependent pricing structures by various operators.

Pre-processing

Missing data and data cleaning

Passenger count

The passenger_count attribute is a driver-entered value. Of the yellow taxi set, 0.2% of the 84,211,244 values were missing. For green taxi, 6% of 6,020,305 values was missing.

Additionally, it does not make sense to have no passengers on a taxi ride. The zero values were set to NaN. Ideally the zero values will be relabelled with valid passenger counts ($\geqslant 1$). One method that could be explored is using an expectation maximisation algorithm. Setting zeros to missing, the distribution can then be visualised (Figure 1).

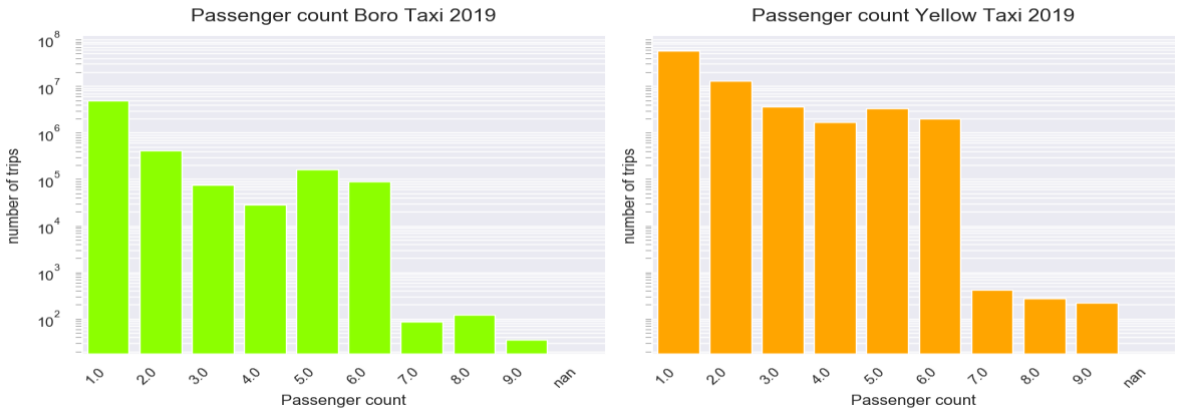


Figure 1—Taxi passenger distribution

Meter rate code

Many missing RatecodeID values were found. Similarly, an unspecified RatecodeID (99) was found. Further research provided no clarification into what rate code 99 represented, therefore it was treated as missing data.

[2] **New York will be the first US city to charge drivers to enter its busiest areas**
https://qz.com/1584287/new-yorks-congestion-pricing-will-make-it-more-expensive-to-drive-in-manhattan/

In yellow taxi data, the total proportion of unknown and missing values was 0.2%. In green taxi data, the total proportion missing or unknown was 6% (**Error! Reference source not found.**). Rate code 99 was combined with the missing data as NaNs.

Payment type

Similarly, this attribute has 0.2% (yellow) and 6% (green) of data missing or unknown when missing values were aggregated with the Unknown category. It is likely these missing values for both ratios were generated by the same mechanism.

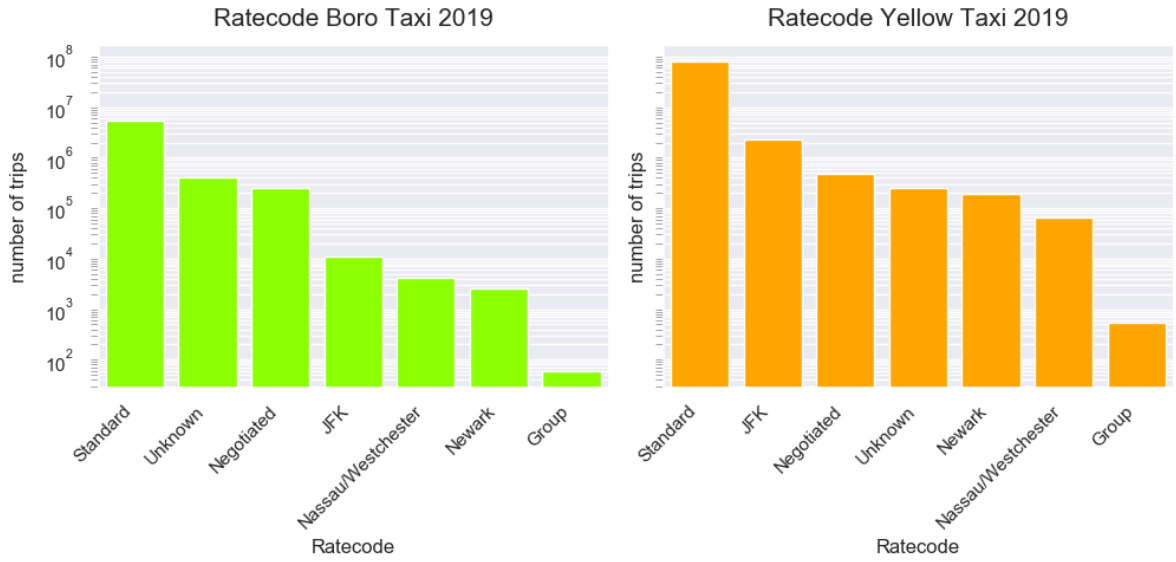


Figure 2—Distribution of meter rates

TPEP provider

A third TPEP-provider unspecified in the TLC Records User Guide[3] was identified with VendorID 4. As there are a limited few authorised TPEP-providers, light research makes it clear that this vendor is likely to be Flywheel Inc., whose technology was approved for operation in New York in July 2016.[4] This was relabelled as such in the data, although the effect is likely not too different from discarding the category altogether due the extremely low installation rate of VendorID 4's technology (**Error! Reference source not found.**).

Congestion surcharge

The congestion surcharge is applied for journeys that start, end, or pass through central Manhattan, with the amount charged depending on the type of taxi license, as well as whether the ride was shared between multiple passengers or not. Trips with negative congestion surcharges are likely to be refund transactions and thus were discarded. The small number of trips with missing values (0.6%) were also discarded; setting them to zero would bias the data, and furthermore imputation would be difficult to do with good certainty given the exact route of each trip is unknown.

[3] https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf
[4] https://www.taxiintelligence.com/third-tpep-provider-in-new-york-taxis-flywheel/

Anomalies and outliers

Fare amount, plus extra

The collected data relating to this quantity was very dirty, with many non-sensical values such as negative fares and incredibly expensive (upwards of $1000) fares where the travel times were next to zero.

By summarising the data, it is clear that the classic method of classifying outliers as those data points further than 1.5 IQR from the median will cause the loss of sensible data. Applying this statistic to the green taxi data, $26.75 is a perfectly reasonable taxi fare, thus this should not be used as the outlier rule (Table 1).

Instead, this paper rejects the upper 0.01% quantile of data as outliers in effort to retain as much naturally generated data as possible. These are fares above approximately $200, which—assuming the longest trips are those that cross the geographical area under study—should be a fairly rare occurrence.

Table 1—Summary of unprocessed fare amounts

	Upper quanitles		*Data summary*	
			statistic	***fare_amount***
	quantile	***fare_amount***	count	6044050
	0.9	29.11	mean	14.4289326
	0.99	60	std	13.1612877
GREEN	0.999	90	min	-890
	0.9999	200	25%	6.5
	0.99999	727.23804	50%	10.5
	0.999999	1966.08153	75%	18
			max	4011.5
			statistic	***fare_amount***
	quantile	***fare_amount***	count	84399019
	0.9	26	mean	13.3439912
	0.99	52	std	174.374902
YELLOW	0.999	96	min	-1856
	0.9999	210	25%	6.5
	0.99999	400	50%	9.5
	0.999999	995.44475	75%	15

[redacted]		max	943274.8

Further validation is applied to identify data points likely generated by alternative mechanisms.

Negative values are discarded on the basis that these represent refunds. Supporting this, the majority of data are labelled as no-charge or disputes given a negative fare (Figure 3). These records represent only a small portion (0.2 %) of the data. Due to their nature as refunds, many will also already appear within the dataset, so their exclusion will improve the quality of data.

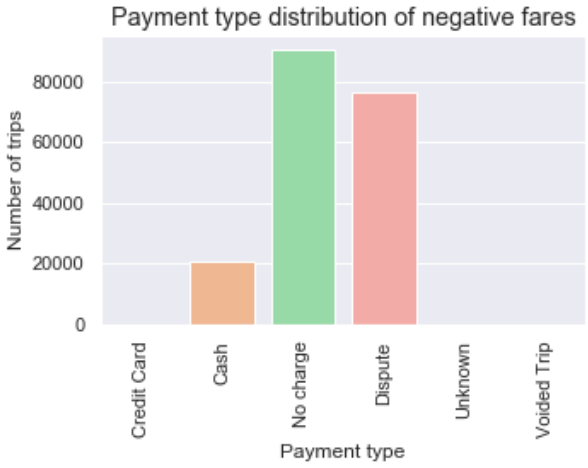


Figure 3—Payments are largely classified as no charge or disputes when fare is negative

Instances of zero fare are discarded only when the trip distance is not also zero to account for the possibility that only the fare was somehow not properly recorded.

Further rows are discarded based on analysis of the pricing structure of New York taxi fares. For example, trips between John F. Kennedy Airport and Manhattan operate on a flat rate of $52 plus surcharges. Instances where the fare is below this are almost certainly (91.3%) invalid trips and discarded if the travel distance is less than the straight-line distance between the two locations (14 miles). Google's mapping puts the route at roughly 18 miles—which is the mode distance of trips with the JFK rate code—but nobody should assume Google knows NYC better than an experienced cab driver.

Applying these cleaning rules, the distribution of fares can be much better characterised (Figure 4).

As shown summarised in the previous table, the median fare of a trip is quite low. However, that statistic doesn't summarise the higher fares when travelling to airports, Nassau and when fares are negotiated. Additionally, there is little difference in fare when riding a Boro taxi vs. Yellow taxi on the same kind of fare, though riders may prefer to negotiate with Green taxi drivers over Yellow ones.

Trip distance

After cleaning the fare_amount, trip_distance was likewise found to have significant anomalies in the data. Negative values are discarded, as this is nonsensical. Furthermore, values in the top 0.0001% percentile are also discarded, which are trips that cover more than 53 miles of ground. This is roughly the length of New York City at its widest point, and trips that are longer than this are outside the scope of analysis.

Trip duration

Some trip durations were negative. While some of these records were due to trips that wrapped around the beginning of the year, month or day, the majority of negative durations had pickup time being recorded as after the drop-off time. All instances of negative trip duration were discarded instead of attempting to amend them.

New features

Trip durations were calculated by the difference in the pickup and drop-off datetimes.

To help characterise the market share, the trip count is aggregated by zones and per company (Uber, Lyft, Via, Juno, as well as Yellow and Green will be defined as companies).

Data formatting

Attributes in the supplied data were reformatted to numpy data types described in Table 2.

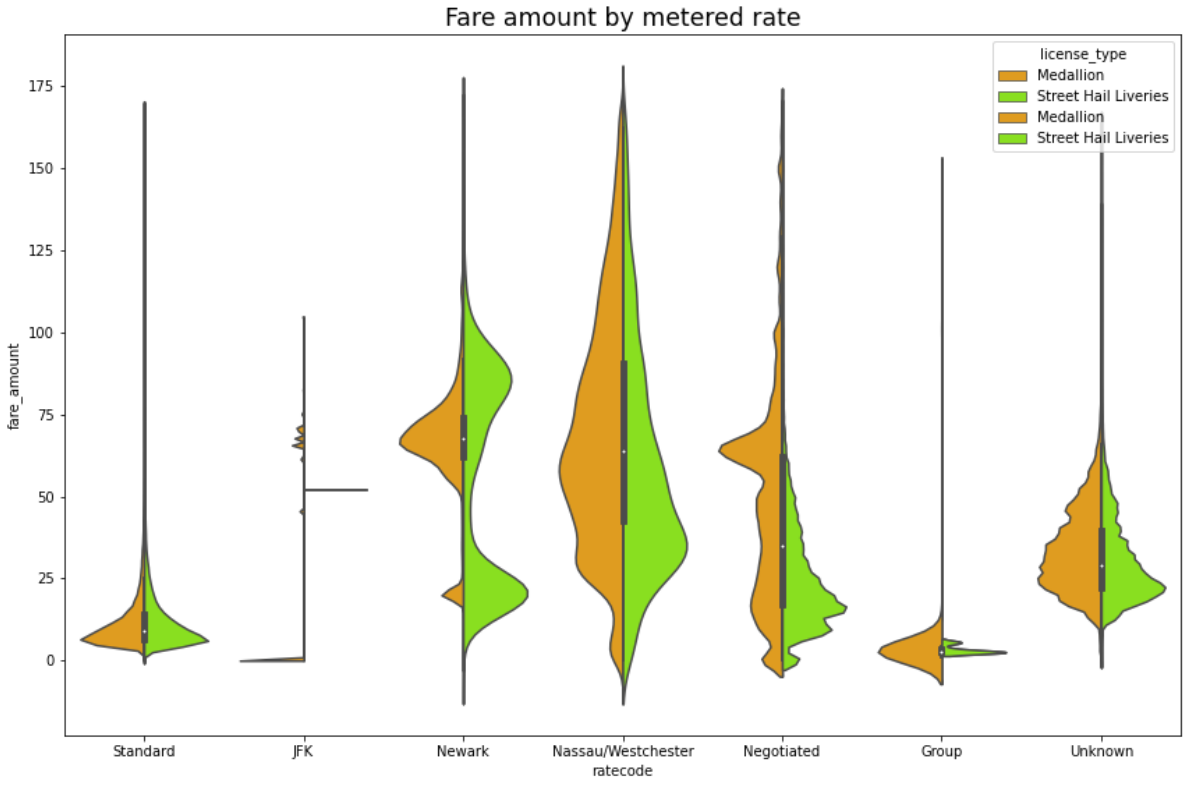


Figure 4—Fare distribution after cleaning

Table 2—Loaded attribute categories

Yellow & Green Taxi		For Hire Vehicle (Uber, Lyft, Juno, Via)	
Attribute	**Type**	**Attribute**	**Type**
VendorID	category	hvfhs_license_num	category
store_and_fwd_flag	bool	dispatching_base_num	category
{.}pep_pickup_datetime	datetime64	pickup_datetime	datetime64
{.}pep_dropoff_datetime	datetime64	dropoff_datetime	datetime64
trip_duration	timedelta64	trip_duration	timedelta64
passenger_count	int64	PULocationID	category
ratecode	category	DOLocationID	category
DOLocationID	category	SR_flag	bool
PULocationID	category		
payment_type	category		
fare_amount	float64		
extra	float64		
congestion_surcharge	float64		

Exploratory Data Analysis

Market share

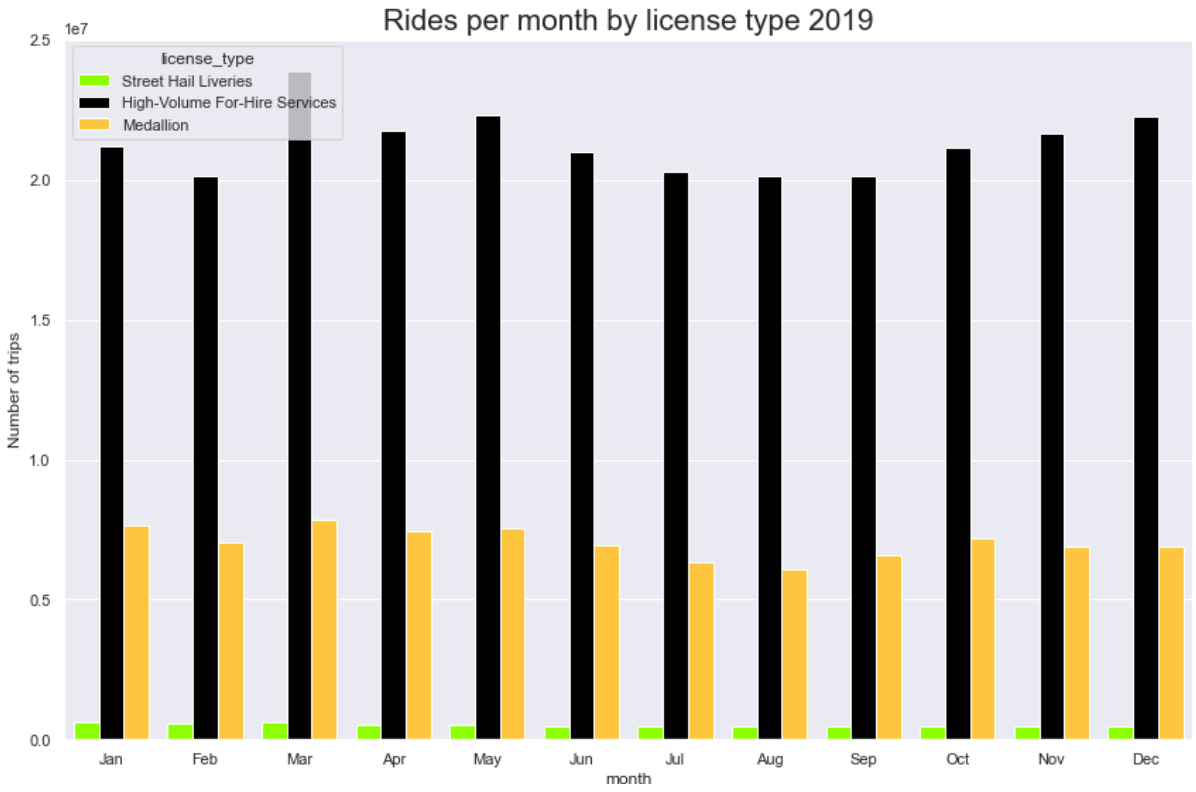


Figure 5—Monthly trips by license type

It is clear that ride-hail companies have the largest slice of the market and far outnumber the trips completed by yellow taxi's by at least triple (Figure 5).

However, further breaking down yearly pickups by companies, a few patterns emerge. Uber by far dominates the ride sharing industry in NYC, and is extremely popular amongst New Yorkers in Brooklyn and lower Manhattan, as shown by the lighter black colours. While

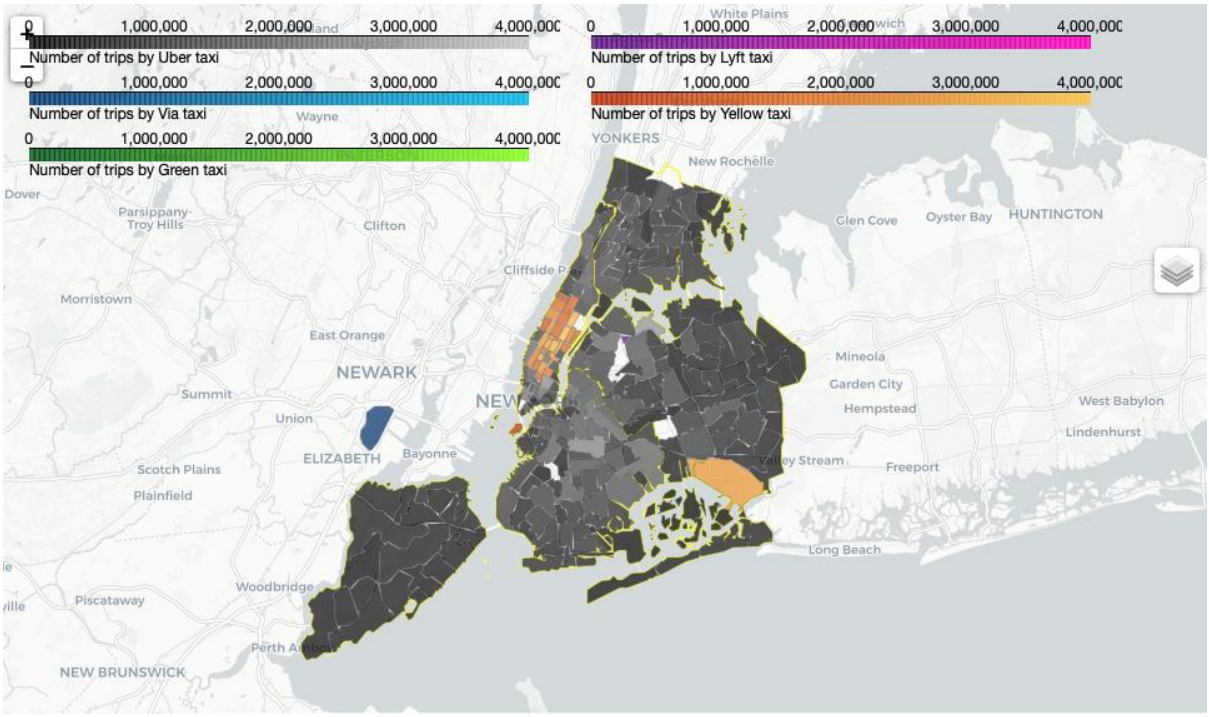
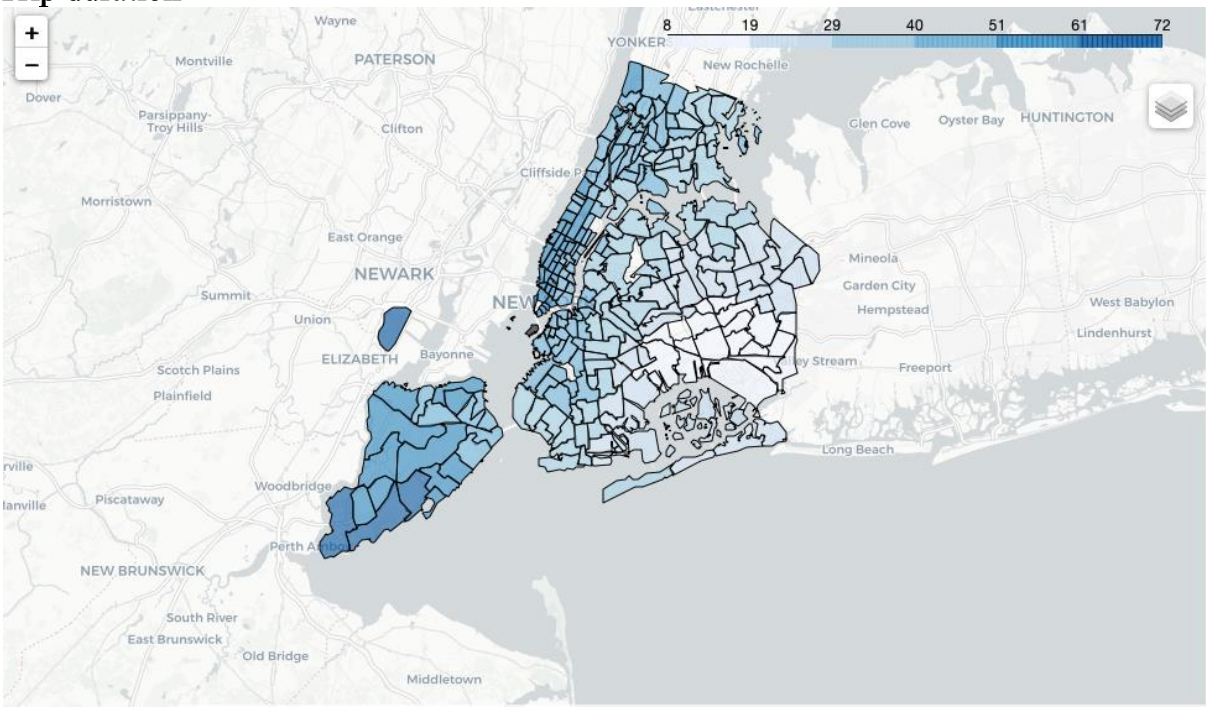


Figure 6—Dominant players and pickup volume by taxi zone. Download interactive HTML here.

yellow taxis are absolutely outnumbered, they still remain the dominant player in Midtown Manhattan as well as at JFK Airport. Interestingly, only two zones are dominated by other players: Newark Airport is dominated by Via and Lyft is most popular for rides originating at Saint Michaels Cemetery.

Trip duration



This paper also visualised the travel time between to JFK Airport. As expected, the average travel time generally increases with distance from the airport. It could be more interesting fix a point of original and explore how travel times vary according to the times of day, as well as look at travel durations before and after the congestion surcharge became mandated.