NYC Taxi Review

The Effects of Covid-19 on New York Cab Services

Benjamin Kuehner
University of Colorado Boulder
benjamin.kuehner@colorado.edu

Tim Papich
University of Colorado Boulder
timothy.papich@colorado.edu

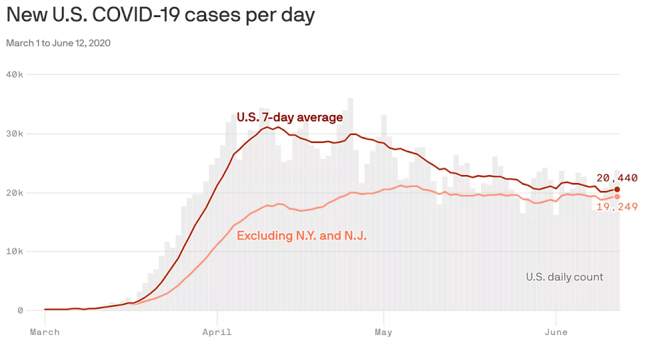
Steven Putt
University of Colorado Boulder
steven.putt@colorado.edu

Abstract

This paper assesses the impact of COVID-19 and various lockdown measures on taxi and ride-sharing services in New York City. The consequences of the pandemic for customer habits and the economics of the industry are evaluated across data from millions of taxi rides between 2019 and 2020. Passenger volume, trip duration, tip percentages, and a host of other metrics are explored in this multidimensional approach. While this analysis is performed primarily through the lens of COVID-19, general trends such as monthly revenue, average traffic congestion by time and location, and the increasing influence of ride-sharing services on the city's traditional medallion system are also investigated. The authors conclude that as a result of the pandemic (1) ridership decreased abruptly resulting in massive shortfalls, (2) overall volatility in all ride-related attributes increased, (3) customers began choosing to take longer trips to more remote locations, (4) tipping behavior became less predictable, (5) airport taxi traffic decreased significantly and became an accurate predictor for flight passenger volume.

1 Introduction

In the early months of 2020, the world was confronted by a highly infectious and dangerous disease. Sars-Cov-2, the novel coronavirus responsible for COVID-19, brought with it not only immense suffering and catastrophic mortality, but also a total upheaval of everyday life. Nowhere has this devastation been more visible than New York City, the cultural and economic capital of the United States. As one of the earliest and most prominent epicenters of the disease, as well as the site of some of the country's most stringent lockdowns, it has been the subject of much attention. The relative well-being of America's most populous city, after all, is seen by many as a benchmark for the health of the nation at large.[7]



For an economy as dynamic and complex as New York City's, measuring the full economic and social costs of both the virus and accompanying government interventions would be a tall order. A more practical approach may be to focus on a single sector, representative of the economy in totum.

As schools and businesses stay shuttered, and citizens remain confined to their homes, taxis and ride-share services have been hit hard by this crisis[9]. It is not far-fetched to say that the mobility of employees and customers is, in some sense, an analog for the liquidity and prosperity of an economy. Therefore, in this study we have chosen to look at New York's ride-hailing industry to gain a better insight into the ways in which COVID-19 has affected business and society as a whole.

Our goal was to mine through the millions of taxi and ride-share trips taken over the last two years and see if it could help answer some basic questions about the ways in which life has changed as a result of the virus. Our questions fell into two categories:

Economic:

- What is the total cost of COVID and lockdown measures on the cab industry?
- Has the decline in taxi services had cascading effects on other industries (i.e. airlines)?
- Have traditional taxi services or ride-share apps such as Uber and Lyft been more impacted by the pandemic?
- Has the industry been affected equally across geographic areas.

Sociological:

- Have people's habits changed? i.e. are more people going to parks and places of outdoor recreation?
- Has charitableness been affected by the pandemic? Are people tipping their drivers more or less?
- Has there been a significant, measurable exodus from the city to escape the virus?
- Have people's movements been more responsive to increasing case and death counts or to changes in government mandates?

It is important to establish a timeline of events as they unfolded in the city before we begin. Below is a list of key events from the early days of the pandemic[8].

January 21, 2020	First confirmed COVID-19 case in the U.S.
January 30, 2020	The W.H.O. declares a global health emergency
February 29, 2020	First reported COVID-19 death in the U.S.
March 1, 2020	**First COVID-19 case in New York State**
March 7, 2020	NY Governor Andrew Cuomo declares a state of emergency
March 12, 2020	Events with more than 500 people must be cancelled or postponed
March 14, 2020	**First two COVID-19 deaths in NYS**
March 16, 2020	**NYC public schools close**
March 17, 2020	**NYC bars and restaurants close, except for delivery**
March 22, 2020	**NYS on Pause Program begins, all non-essential workers must stay home**

Note how rapidly the situation progresses from the date of the first case to full-scale shutdowns. From the perspective of our data analysis, we can expect such an abrupt decline in ride numbers that pre- and post-Covid timeframes can be treated as entirely separate, with virtually no intermediate phase.

2 Related Work

Before we start digging into the data, it is worth acknowledging some of the prior work that has been done in this field. New York taxi services have been the subject of a wide range of studies. Our project has taken the following research papers as a reference and starting point:

2.1 Taxi and Ride Hailing Usage in New York City[3]

-A comprehensive set of charts displaying time-based trends of taxis and ride share companies, including trends like total trips, new drivers, trip length etc. per day or month.

2.2 Reducing Inefficiencies in Taxi Systems[4]

- Supply and demand analysis of New York taxis system which shows that predictive demand modeling and dispatching can reduce the required fleet size by 28% and decrease idle time by 32%.

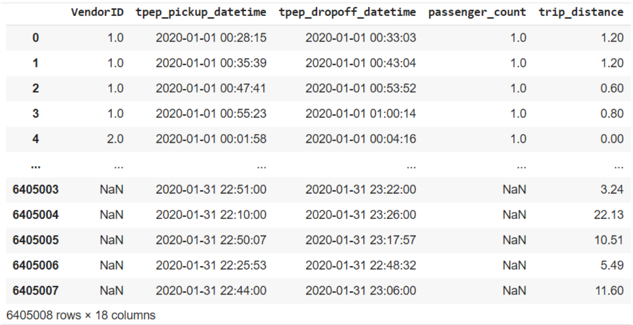
2.3 Spatio-temporal Pattern Analysis of Taxi Trips in New York City[5]

-Explores trends of New York taxi trips from different locations for trends including trip frequency and speed of travel during the week and on weekends. Uses binomial regression models to predict traffic to and from popular destinations such as airports.

As all of these studies took place pre-COVID, they are useful to us mostly as a point of reference for how the taxi industry *normally* operates.

3 Data Set

The primary data source utilized in this study comes from the New York City Taxi and Limousine Commission (TLC). Started in 1971, the agency is responsible for licensing and regulating both medallion taxis and for-hire vehicle services in the city. Since 2009, they have been collecting information on every taxi trip that takes place in their jurisdiction. The entirety of the data is now available on the agency website and freely accessible to the public. Below is an excerpt of this data from January 2020. In that month alone roughly 6.4 million trips took place.



	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance
0	1.0	2020-01-01 00:28:15	2020-01-01 00:33:03	1.0	1.20
1	1.0	2020-01-01 00:35:39	2020-01-01 00:43:04	1.0	1.20
2	1.0	2020-01-01 00:47:41	2020-01-01 00:53:52	1.0	0.60
3	1.0	2020-01-01 00:55:23	2020-01-01 01:00:14	1.0	0.80
4	2.0	2020-01-01 00:01:58	2020-01-01 00:04:16	1.0	0.00
...	...	...	...	...	...
6405003	NaN	2020-01-31 22:51:00	2020-01-31 23:22:00	NaN	3.24
6405004	NaN	2020-01-31 22:10:00	2020-01-31 23:26:00	NaN	22.13
6405005	NaN	2020-01-31 22:50:07	2020-01-31 23:17:57	NaN	10.51
6405006	NaN	2020-01-31 22:25:53	2020-01-31 22:48:32	NaN	5.49
6405007	NaN	2020-01-31 22:44:00	2020-01-31 23:06:00	NaN	11.60

6405008 rows × 18 columns

For the purposes of this study, we have limited ourselves to data from the period of January 2019 to June 2020 (the most recent available data). The combined database consists of over 100 million trips with each entry containing the following relevant attributes:

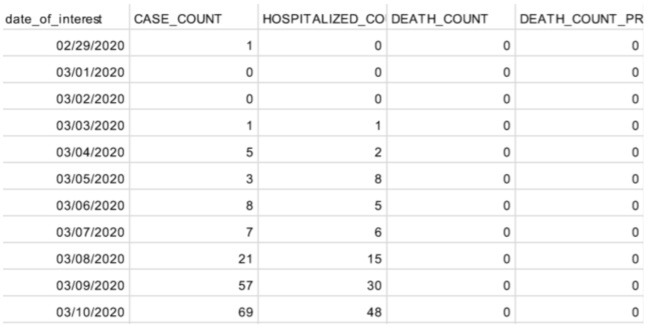
- pickup/dropoff datetime
- pickup/dropoff location ID
- passenger count
- payment type
- fare amount
- tip amount
- toll amount
- congestion surcharge

The TLC has similarly compiled data for ride-sharing services such as Uber and Lyft, referred to as High-Volume For-Hire Services (HVFHS) under New York law. However due to privacy constraints, only the attributes pickup/dropoff location and pickup/dropoff time are documented. Unfortunately, data for October

2019 is missing from their records. To account for this we chose to use averages from September and November of that year as stand-in data.

In addition to the TLC, we have leveraged data from the Port Authority of New York and New Jersey Airport Traffic Statistics to calculate flight passenger volume at each of New York's three major airports.

Finally, all of these taxi and ride-share analytics are put into context by coronavirus data available to us from the New York City Health Department's public github repository. From this trove of information, we chose only to incorporate a simple csv file of daily new cases, hospitalizations, and deaths in the city:



date_of_interest	CASE_COUNT	HOSPITALIZED_CO	DEATH_COUNT	DEATH_COUNT_PR
02/29/2020	1	0	0	0
03/01/2020	0	0	0	0
03/02/2020	0	0	0	0
03/03/2020	1	1	0	0
03/04/2020	5	2	0	0
03/05/2020	3	8	0	0
03/06/2020	8	5	0	0
03/07/2020	7	6	0	0
03/08/2020	21	15	0	0
03/09/2020	57	30	0	0
03/10/2020	69	48	0	0

4 Methodology and Results

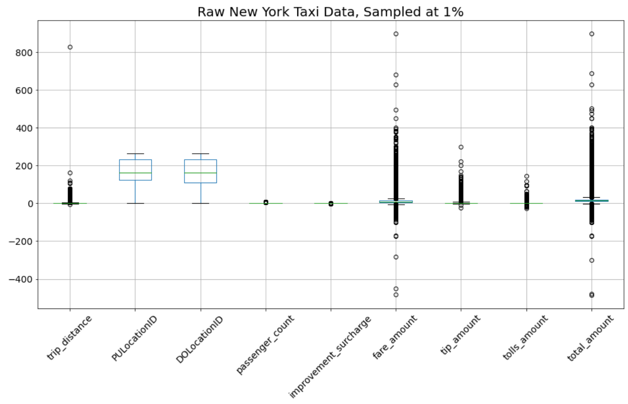
In this section, we present a detailed walkthrough of our data analysis process. All of our calculations and data manipulation were conducted through the use of Python and its standard supporting libraries - Pandas for dataframe maintenance, Numpy for larger mathematical functions, Matplotlib and Seaborn for data visualization, and Sklearn for clustering and regression algorithms.

4.1 Preprocessing and Cleaning

Given the sheer magnitude of the yellow cab dataset, our first task was to trim the data into a manageable and hardware-friendly size. To avoid compromising the integrity of the data, we chose to use uniform sampling at 1% resolution across each file in the database. As the taxi data comes in individual files for each month of the year, this process was as simple as converting each csv to a Pandas dataframe, extracting 1% of all rides for that month, and concatenating it to a master dataframe.

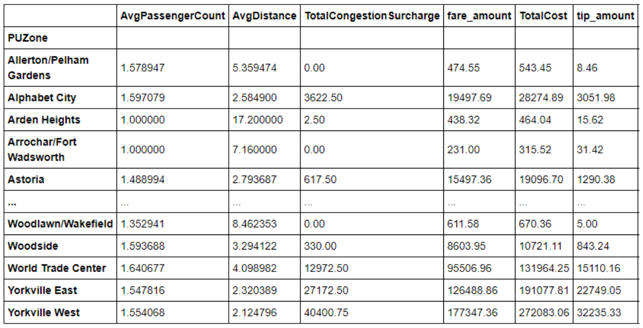
Due to the lengthy computation time involved in the sampling process, we also took this opportunity to define data types for each attribute in the set. While Pandas comes with automatic type recognition, it is not always entirely accurate. In particular, categorical data such as location ID's had to be redefined as integers and numerical data such as fare and tip amounts needed to be labeled as floats.

With the entirety of the trip data sampled, the total number of entries was reduced from over 100 million down to just 1 million - a far more workable figure. Any summary figures extracted from the sampled data, such as total tips, were multiplied by 100 to scale it back to the original data quantity.



Our next task was to incorporate the index files - mapping the zone ID, payment type, and rate code for each element to their corresponding string values. This was especially critical for the geographic data, given the large number of designated zones. Normally, it would be faster to manipulate index numbers alone, but having reduced the size of our data so dramatically, we felt the added processing time would not be too detrimental. In addition to indexing, we also created a derived column "from:to" which stores the combined pickup and dropoff location as a single unit. This extra column would later allow us to quickly determine the most common routes for any given timeframe. Lastly, we appended columns for pickup and dropoff borough to help identify trends on a more macro level.

Aggregated Figures for PU/DO locations



	AvgPassengerCount	AvgDistance	TotalCongestionSurcharge	fare_amount	TotalCost	tip_amount
PUZone						
Allerton/Pelham Gardens	1.578947	5.359474	0.00	474.55	543.45	8.46
Alphabet City	1.597079	2.584900	3622.50	19497.69	28274.89	3051.98
Arden Heights	1.000000	17.200000	2.50	438.32	464.04	15.62
Arrochar/Fort Wadsworth	1.000000	7.160000	0.00	231.00	315.52	31.42
Astoria	1.488994	2.793687	617.50	15497.36	19096.70	1290.38
...	...	...	...	...	...	...
Woodlawn/Wakefield	1.352941	8.462353	0.00	611.58	670.36	5.00
Woodside	1.593688	3.294122	330.00	8603.95	10721.11	843.24
World Trade Center	1.640677	4.098982	12972.50	95506.96	131964.25	15110.16
Yorkville East	1.547816	2.320389	27172.50	126488.86	191077.81	22749.05
Yorkville West	1.554068	2.124796	40400.75	177347.36	272083.06	32235.33

With the data fully sampled and merged, and with the various indices dereferenced, we began the cleaning process. The raw data, while mostly accurate and complete, still contained a fair number of inconsistencies.

Among the most glaring problems was the large number of invalid dates. Using Python's stock datetime parser, we were able to quickly separate year, month, day, and time into their own respective columns, and delete all entries with negative times, dates in the future, and dates in the remote past. In the process of scanning for these errors, we took the chance to append a column called "covid", storing a boolean value which held true for trips occurring after the start of city-wide lockdown measures.

Finally, we scanned the dataframe to eliminate any entries with the following issues:

- fare or total payment amount with a negative or NaN value.
- trip distances either less than or equal to zero or exceeding 1000 miles
- payment types marked as "Disputed" or "No Charge"

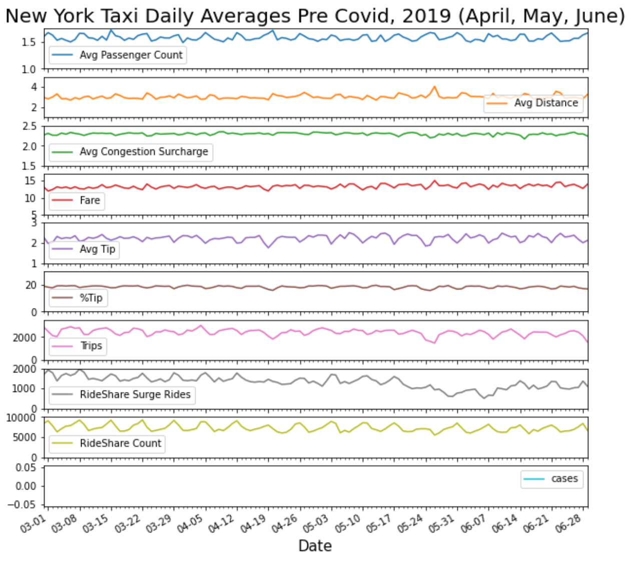
The sum of these removed entries came to just under 14,000, giving us a combined total of 997,966 taxi rides to use in analysis.

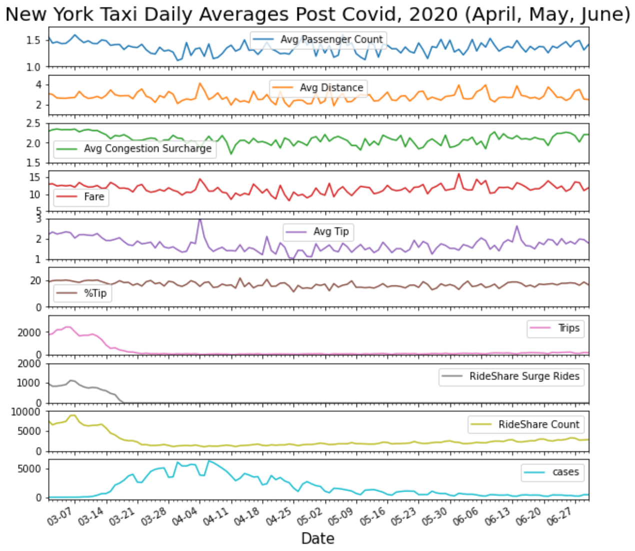
It is worth noting that each of these cleaning and preprocessing steps were performed identically on both the HVFHS data (Uber, Lyft, etc.) and the more robust yellow cab data. Because of the differing number of attributes, we chose to store them in separate files, invoking them individually according to the task at hand. Where ride-share and yellow cab data appear together in the analysis below, they are merely grouped together by date.

4.2 **Daily Averages and Changing Volatility**

Our analysis begins by establishing basic trends from before and after the start of COVID lockdowns through some preliminary data visualizations. We produced these simple images not to prove banal assertions like "taxi use declined during lockdowns", but instead to establish a frame of reference for the post-lockdown environment. We also expected that these visualizations could alert us to some features deserving of investigation.

Below are two such charts that plot the daily averages for each of our numerical attributes during equivalent months before and after COVID-19:



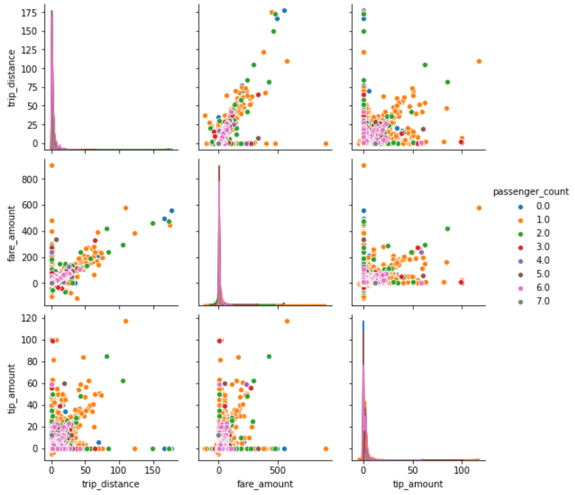


The most glaring aspect of these visualizations is the sharp decline in ridership between the pre- and post-Covid time frames. By summing the average total amounts for these two periods, we estimate that New York City lost nearly 150 million dollars in tax revenue from taxis alone. Drivers, likewise, lost a combined 128 million dollars *just in gratuities*. While there is little surprising about the weight of these figures, it is worth keeping in mind the sheer economic devastation that this pandemic has had on the industry.

The second thing to note from these graphs is that the data is more or less static prior to the lockdowns, with very little seasonal variability. After the lockdowns, not only do the averages for each category decrease, but volatility greatly increases. This can partly be explained by mathematics - as the number of overall data points decreases, the effect of local outliers will increase. However, since our dataset contains well north of a million elements, one would expect this affect to be less apparent in a visualization. This overall change in volatility when viewed alongside geographic data we will present later, suggests that there was not just a decrease in the amount of trips taken, but a change in *type* of trips taken. Namely, people were more likely to take longer trips to more remote areas.

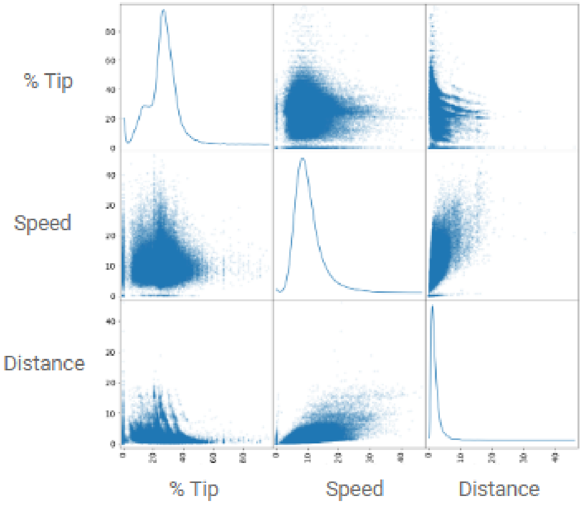
4.3 Customer Tipping Habits

Another notable aspect of the above graphs is that fare and average tip amount, attributes which were highly correlated before the lockdown, appear to be less so after. One of the questions we specifically set out to answer in this project was how tipping (altruism) might change during the pandemic. As both common sense and the pair plot below illustrates, tips (prior to the lockdown) are highly correlated with fare price and trip distance.



Interestingly, after Covid the opposite becomes true. Our data indicates that during the lockdown, shorter trips began to equate to larger tips as a percentage of fare. How could this be? Did peoples' appetites for altruism change as a result of lockdown procedures?

Speed of trip as a predictor for % tip:

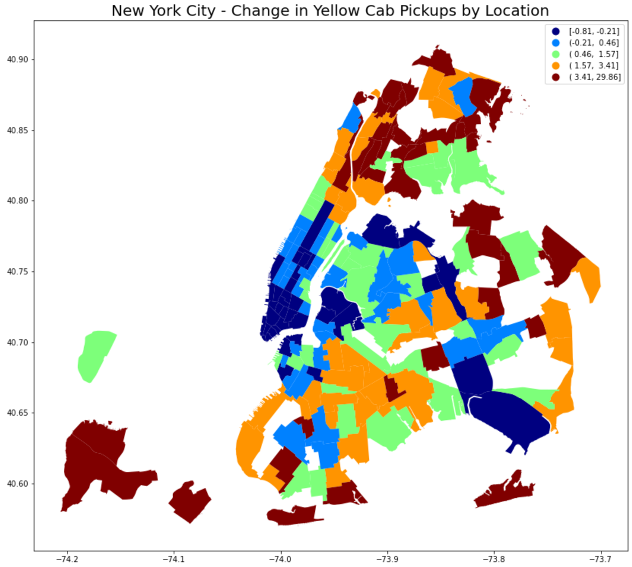


We initially hypothesized that as total traffic congestion decreased in the city, the length of rides would likewise decrease, and since customer satisfaction is ostensibly tied to the speed of service, the perceived increase in efficiency would be grounds for a more generous tip. In analyzing speed as a predictor of tipping, however, we found no such correlation.

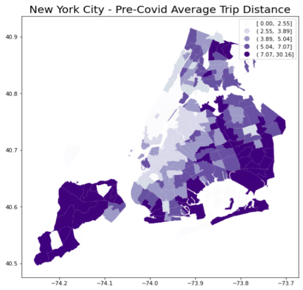
While we cannot say for certain why tipping habits have changed in this way, we can provide at least one speculative explanation. It could be that in the post-Covid world, since customers were far more likely to use other means of transportation such as walking or biking[9], they began to perceive small taxi rides as an indulgence. Coupled with an increased awareness and sympathy for "essential workers", could it be that customers were more inclined to feel such small trips deserved a larger tip? Perhaps this is a question better left to the psychologists.

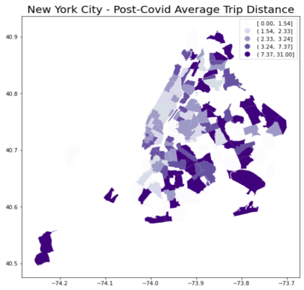
4.4 Geographical Observations

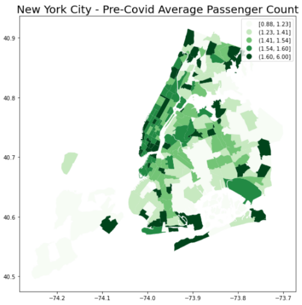
Another problem we were interested in solving was the effect of COVID-19 on transportation in specific areas of the city. To look into this, we started by producing a color-coded map of the city indicating percentage change between the number of trips pre- and post-Covid for each location ID:

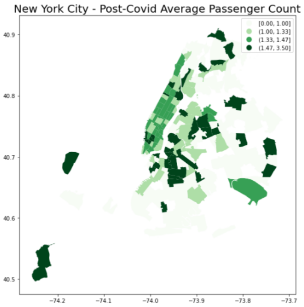


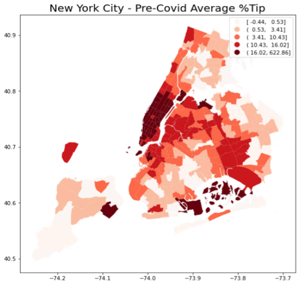
After aggregating along each attribute for every pickup/dropoff location, we derived similar heatmaps for each attribute. Here are examples for average trip distance (above), average passenger count (middle), and average %tip (below):

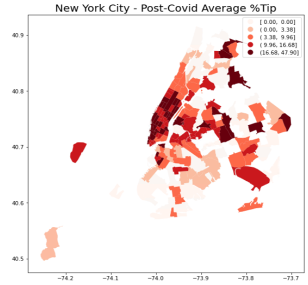








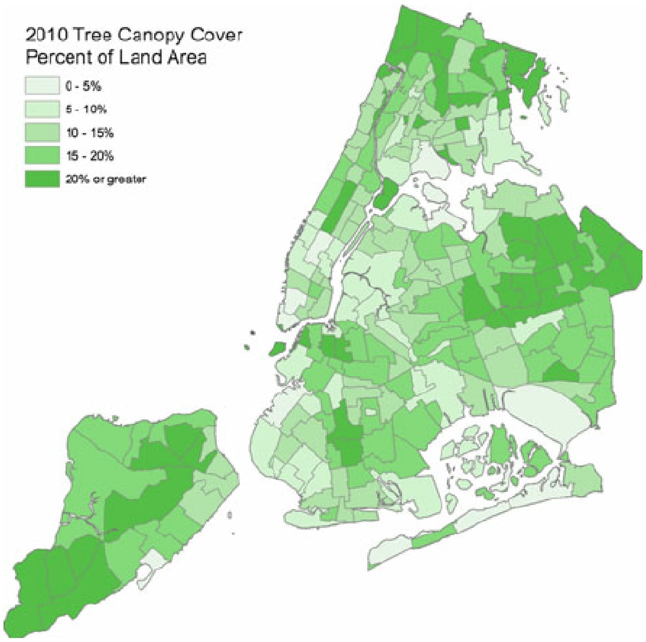


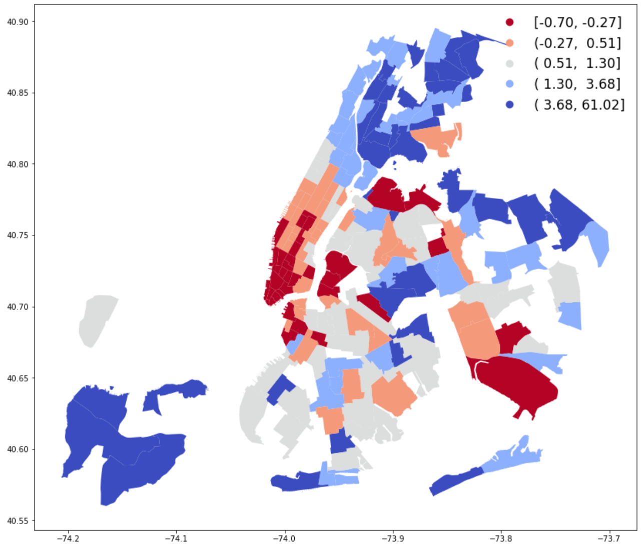


Looking at these heatmaps, a few peculiarities become immediately apparent: Why did the average trip distance increase in the most populous areas of New York, and decrease in the outer edges of the city? Similar to our earlier observations, why did average tips fluctuate so randomly in different locations? Why did certain areas of Queens and Brooklyn see a spike in passenger counts?

One explanation is that as coronavirus cases skyrocketed in the city, New Yorkers began seeking out less populous areas to pass the time. Likewise, as bars, restaurants, concert venues, and various other businesses closed, outdoor activities became one of the sole sources for recreation. Residents of downtown Manhattan would have to drive further distances than those living in Staten Island to find such places.

As expected, when comparing data from a 2010 environmental report on tree growth in New York City to our ride-hailing data, we found the corresponding heatmaps to be virtually identical. More trees (i.e. more “nature”) was correlated with the highest percent increase in pickups and dropoffs in the period post-Covid.



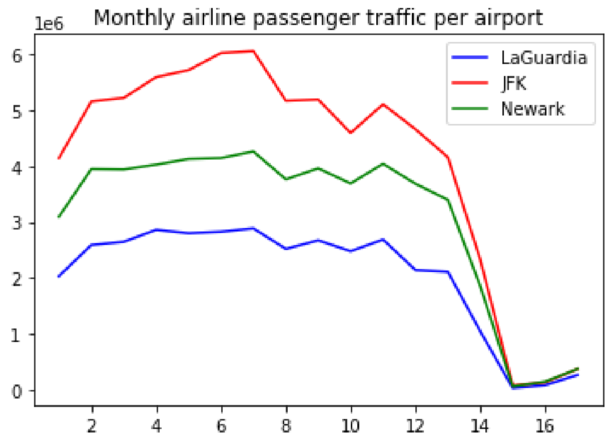


Certain questions remain unanswered. It is not clear, for instance, why parts of the city have had an increased average number of passengers per trip. To answer these types of questions, it may be necessary to poll the residents of these areas. What the geo-data does show, however, is that customer behavior changed significantly as a result of the pandemic. A variety of additional datasets would be useful for future spatial analysis. For example, the acreage of open space, number of bars and restaurants, demographics, etc. by borough could help us understand the behaviors before and after covid.

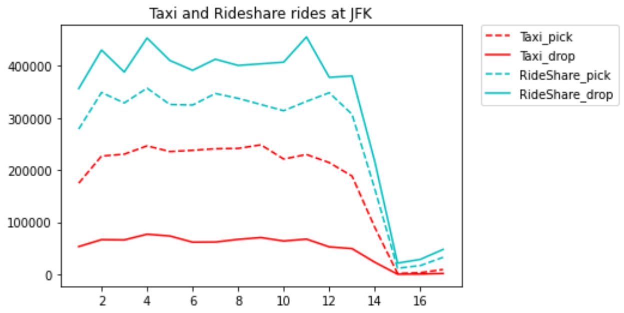
4.5 Airport Traffic Statistics

New York City has three major airports - LaGuardia and JFK which are both in Queens and Newark in New Jersey. New York City's three major airports account for a significant portion of taxi business in the metropolitan area. As it is widely known that the airline industry has faced significant challenges as a result of the pandemic, we chose to also examine changes in taxi and ride-sharing traffic from passengers at JFK, LaGuardia, and Newark airports. Our goal was to see whether taxi traffic itself could be used as a model to predict the number of airline passengers flying in and out of each airport.

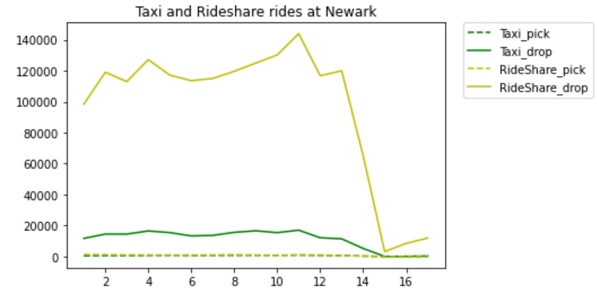
Not surprisingly, flight passengers at these airports declined significantly after the start of the pandemic:



Accordingly, taxi passenger traffic at each airport also slowed. One surprising feature was that ride-sharing services were far more negatively affected than traditional taxis at all three airports.



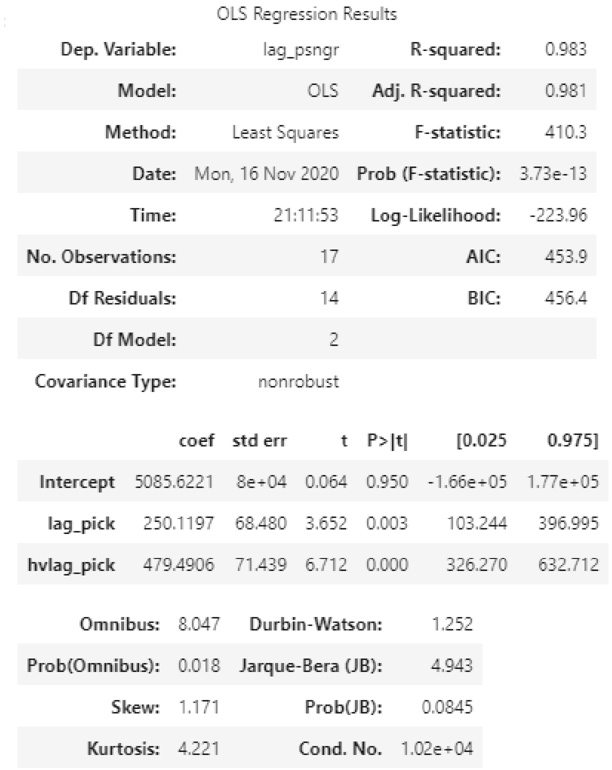
The trends at LaGuardia were similar to JFK. However, the dataset contained far more drop offs than pickups at Newark. This is probably because Newark is located in another state. New Jersey may have rules that make it more difficult for New York-based taxis to pick up passengers. It might also be because New York-based taxis spend almost all their time in New York so they don't go to Newark on speculation that they may pick up a passenger.



We used ordinary least squares regression to build models that could predict the number of air passengers at each airport based on taxi data. The p-values for every variable used in each model is zero or close to zero and the confidence interval for each variable does not include zero. This presents strong evidence that the effect of each of the variables is statistically significant.

Here is an example of our model for LaGuardia based on taxi and rideshare pickups at the airport. This model has an R squared of 0.983 which is very high.

The final model is: LaGuardia air passengers = 250.1197 * LaG taxi pickups + 479.4906 * LaG rideshare pickups + 5085.6221



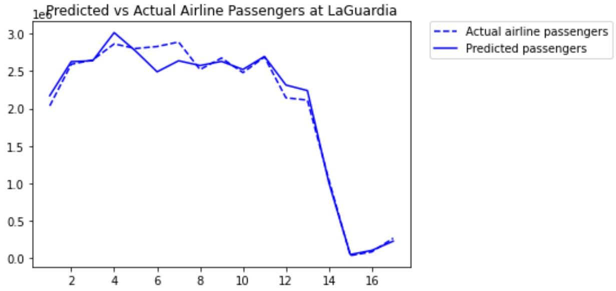
OLS Regression Results

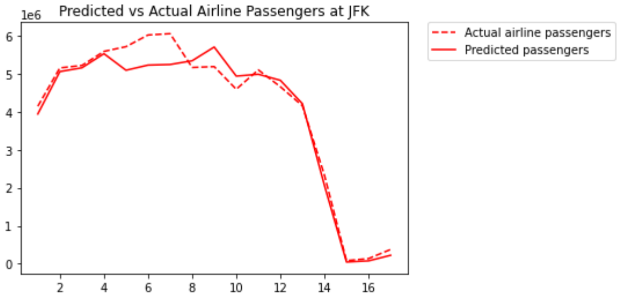
Dep. Variable:	lag_psngr	**R-squared:**	0.983
Model:	OLS	**Adj. R-squared:**	0.981
Method:	Least Squares	**F-statistic:**	410.3
Date:	Mon, 16 Nov 2020	**Prob (F-statistic):**	3.73e-13
Time:	21:11:53	**Log-Likelihood:**	-223.96
No. Observations:	17	**AIC:**	453.9
Df Residuals:	14	**BIC:**	456.4
Df Model:	2		
Covariance Type:	nonrobust		

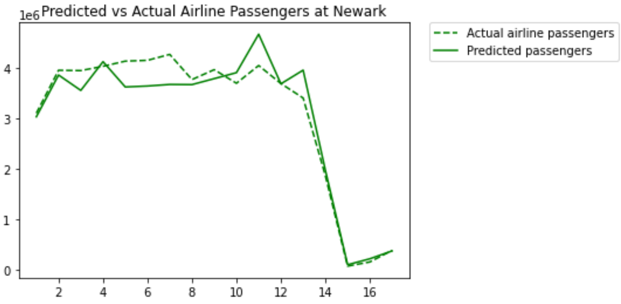
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5085.6221	8e+04	0.064	0.950	-1.66e+05	1.77e+05
lag_pick	250.1197	68.480	3.652	0.003	103.244	396.995
hvlag_pick	479.4906	71.439	6.712	0.000	326.270	632.712

Omnibus:	8.047	**Durbin-Watson:**	1.252
Prob(Omnibus):	0.018	**Jarque-Bera (JB):**	4.943
Skew:	1.171	**Prob(JB):**	0.0845
Kurtosis:	4.221	**Cond. No.**	1.02e+04

Plotting predicted versus actual values based on our models for each airport gives us the following:







A key difference between the models is the Newark model uses rideshare drop offs instead of pickups. The number of pickups was too low at Newark to build a meaningful predictive model.

Another difference is the LaGuardia model uses data from taxi and rideshare services while the JFK model only uses taxi data. Taxi and rideshare volume shows higher collinearity than comparable volume at LaGuardia. Therefore, it is not necessary to include both types of services in the JFK model.

One useful application of these models is that taxi companies will be able to better schedule their airport presence based on predicted airline passenger volume.

6 Conclusion

COVID-19 has had a considerable and catastrophic effect on our society and our way of life. In addition to the truly regrettable loss of life, the economic consequences have been nothing short of devastating. Nowhere has this been more visible than in the transportation industry, where taxi and ride-share services have seen staggering declines. While virtually every aspect of this crisis has been a net negative, this study has shown that at least some room for optimism remains. People are taking more trips to parks and more natural locations. Taxi customers (in some cases) are exhibiting a willingness to give bigger tips. Last of all, our most recent data appears to show a gradual return to normalcy. Many questions remain unanswered, but we hope ride-hailing industries will be able learn something from this report as they continue to adapt to the challenges posed by COVID-19.

REFERENCES

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

[2] https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv

[3] https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/

[4] http://www2.cs.uic.edu/~urbcomp2013/urbcomp2016/papers/Reducing.pdf

[5] Spatio-temporal Pattern Analysis of Taxi Trips in New York City https://flrec.ifas.ufl.edu/geomatics/hochmair/pubs/NYC_Taxi_Hochmair_2016_TRR_DraftLayout.pdf

[6] https://www.panynj.gov/airports/en/statistics-general-info.html

[7]https://www.axios.com/us-coronavirus-new-cases-second-wave-new-york-b6eda2dc-ef39-4b61-9385-d02a2fd3c494.html

[8]https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986

[9] https://www.nytimes.com/2020/11/12/nyregion/nyc-taxi-drivers-coronavirus.html

[10] McPhearson T., Maddox D., Gunther B., Bragdon D. (2013) Local Assessment of New York City: Biodiversity, Green Space, and Ecosystem Services. In: Elmqvist T. et al. (eds) Urbanization, Biodiversity and Ecosystem Services: Challenges and Opportunities. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7088-1_19