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

Problem Statement

Our project will assess the recent impact of COVID-19 on taxi services in New York City. A multidimensional approach will be utilized to explore passenger volume, trip duration, and various other metrics from before and after the start of the pandemic to evaluate its effects on the industry. Though this analysis will be conducted primarily through the lens of COVID-19, general trends will also be explored such as revenue, traffic congestion by time and location, and the effect of competition from ride-sharing services. We hope that our findings will assist taxi providers and patrons alike in ensuring safe and efficient transportation. We also expect the information gained from this analysis to be useful to policymakers who are considering various pandemic mitigation strategies and business leaders hoping to learn from recent events.

1 Literature Survey

New York taxi services have been the subject of a wide range of studies. Our project will take the following research papers as a reference and starting point:

1.1 Taxi and Ride Hailing Usage in New York City

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

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.

1.2 Reducing Inefficiencies in Taxi Systems

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

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%.

1.3 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

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.

2 Data set

We will be making use of a large dataset of individual taxi trips made available to the public by the NYC Taxi and Limousine Commission on their website:
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

The website also contains data for ride-sharing services such as Uber and Lyft which we make extensive use of in this report. These services are referred to in NYC law as High-Volume For-Hire Services (HVFHS)

An accompanying index of taxi zone ID numbers and their corresponding physical locations will also be used: https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv

Finally, we will be leveraging data from the Port Authority of New York and New Jersey Airport Traffic Statistics https://www.panynj.gov/airports/en/statistics-general-info.html

Together, the combined database will contain over 100 million trips from the last several months. Each datum will include elements such as pickup/dropoff time and location, trip distance, passenger count, fare and tip amount, and congestion surcharge.

3 Summary of Proposed Work

3.1 Preliminary Tasks

Our original source data, while mostly accurate and complete, still contains a large number of inconsistencies that need to be addressed. In addition to these necessary cleaning steps, we also need to perform some basic sorting, labeling, and other preprocessing work before mining the data. The following is a short summary of these preliminary tasks:

3.1.1 Cleaning:

- Account for entry errors
- Correct for refunded trips, duplicate rows
- Control for extreme values such as very large trip lengths
- Dataset features a large number of negative values that must be either deleted or changed

3.1.2 Preprocessing:

- Append multiple flat files into a combined datasets for each data type, taxis and ride share
- Filter for yellow cabs only within taxi data
- Remove unnecessary data columns such as “Taxi - Store and Forward”
- Create monthly groupings to compare seasonal changes
- Classify data as categorical, ordinal, numerical, etc.
- Create categorical mapping to relevant fields such as “Location ID” to related location names
- Group and categorize variables to improve predictive modeling
- Random sampling to refine the dataset to a more reasonable size
- Generate derived columns such as “% tip per ride” to enable comparison between taxi and rideshare datasets

3.1.3 Integration:

- Aggregate data by date from datetime to reduce data size and complication while maintaining the full dataset for detailed analysis if needed
- Taxi, location, ride share datasets merged/appended by common date
- Separating pre and post COVID times

3.2 Data Analysis

After performing the aforementioned preparatory tasks, we began the actual mining process. Here a laundry list of all the mining techniques we will employ at some during the investigation:

- Perform basic statistical calculations – mean, median, range, mid-range, interquartile range, variance, standard deviation, etc.
- Visualize data as histograms, scatterplots, etc.
- Normalize data to account for long trips

- Data grouping such as merging location data into boroughs or other smaller buckets. Perform similar groupings based on time
- Chi-squared test – measure correlation for categorical data
- Correlation coefficient – measure correlation for numerical data
- Develop linear regression and multiple regression models
- Pattern discovery and association rules
- Identify frequent itemsets
- Calculate lift ratios
- Build decision tree
- Calculate information gain, information from each attribute, gain ratio
- Apply Bayes Theorem
- Make predictions using sample data and test against remaining data
- Calculate accuracy and coverage of rules

The primary difference between the proposed analysis in this project when compared to previous work is the division of pre and post COVID events. If time allows, we will search for similar data from other cities and compare these with those of the New York city impacts.

4 Evaluation Methods

In order to validate the conclusions of our research we intend to do the following:

- Legitimize our findings with confidence and support metrics
- Compare our results with existing research and established trends
- Model predicted values and compare to actual results

5 Tools

Here is a non-exhaustive list of tools we intend to use including our (tentative) platform for collaboration:

- Python – programming language
- Google Collaboration - Cloud ipython notebook
- Pandas – python library for data manipulation and analysis
- Numpy – python library with support for large arrays and matrices and mathematical functions
- Matplotlib – python library for creating plots
- SciPy – python library for linear algebra and other mathematical functions
- Seaborn – python library for creation of visualizations
- Sklearn – python library for clustering and regression algorithms
- Patsy – python library for creation of statistical models
- Microsoft Power BI - data integration and visualization software

6 Milestones

6.1 Timeline

10/11 - 10/18
Cleaning, preprocessing, and integration

10/19 - 11/14
Develop linear regression and multiple regression models, Pattern discovery, Make predictions using sample data and test against remaining data

11/15 - 11/20
Share initial results in Progress Report

11/21 - 11/28
More data mining over Thanksgiving break

11/29 - 12/5
Finish numerical analysis, clean up project code and descriptions, film project presentation video

12/6 - 12/10
Final Report, Peer evaluation and interview questions

6.2 Completed tasks:

- Preprocessing steps
  - Changed elements to appropriate data types (numerical data to float, categorical data to int, etc.)
  - Installed datetime parsing for better formatting
  - Joined 2019-2020 data
  - Matched location ID's with corresponding lookup names and added them as columns for both pick up and drop off
  - Created a COVID boolean column to quickly sort pre and post lockdown rows
  - Incorporated a corresponding dataset for Uber/Lyft trips during the same timeframe

- Cleaning
  - Removed all entries with invalid dates
  - Removed rows containing NaN or negative values

- Sampling
  - Sampled 1% of the original dataset and wrote it to a new csv file to use for analysis (this has saved us a huge amount of processing time).

- Basic Calculations
  - Filtered outliers (identified using influence plot with studentized residuals)
  - Computed simple figures such as total trips, median distance, etc.
  - Aggregated along datetime and derived averages for pre and post lockdown
  - Produced correlation tables and pair plots for each numerical attribute to establish some fundamental correlations

- Predictive Modeling
  - Developed a linear regression model to predict cab fare from relevant attributes
  - Created a predictive model for airport traffic at given times of year from cab data.

- Geographic data
  - Grouped and aggregated data by zone ID
  - Created heatmaps comparing pre and post covid averages
  - Created heatmaps comparing the various metrics such as average tip amount and passenger count between zone IDs and boroughs.

- Airport Data
  - Calculated aggregate data for drop offs and pickups for pre and post lockdown at NYC airports for both traditional cabs and ride sharing services
  - Plotted taxi traffic vs. rideshare traffic at all three airports in NYC

6.3 Tasks to Complete:

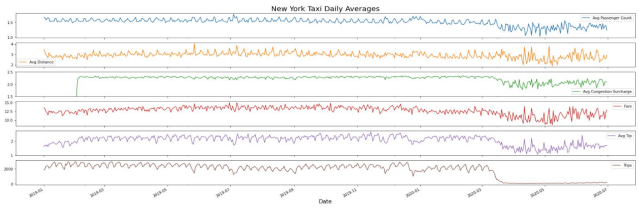
- Create regression models that can predict pre and post lockdown data separately and compare them.
- Incorporate geographic data of COVID cases (where available) into current models.
- Compare NYC data with data from other cities (where possible)

7 Initial Results

7.1 Baseline Statistics and Correlations

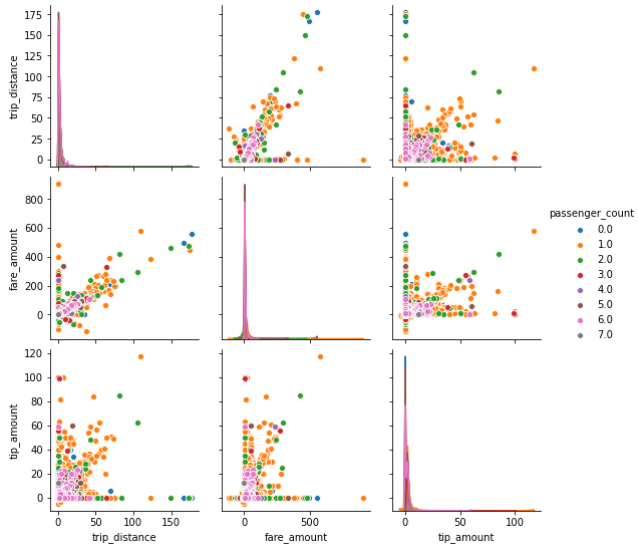
Our analysis begins by establishing basic trends from before and after the start of COVID lockdowns through some preliminary data visualizations. We produced these simple images not to prove banal assertions like “taxi use declined during lockdowns”, but instead to establish a frame of reference for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

Below is one such chart, which plots the daily averages for each of our numerical attributes:



There are a couple of things to be noted from this simple graph. First, 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 is expected - as the number of overall rides decreases, the effect of outliers will increase. What is not explained is how variables like fare and tip amount that were 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.



Why is it then that tips (as a function of fare and distance) become so much less predictable after the lockdown? Did peoples’ appetites for altruism change as a result of lockdown procedures? We will spend the next couple weeks digging deeper into this. One explanation of tip volatility might be varying levels of increased tip due to appreciation of front line workers.

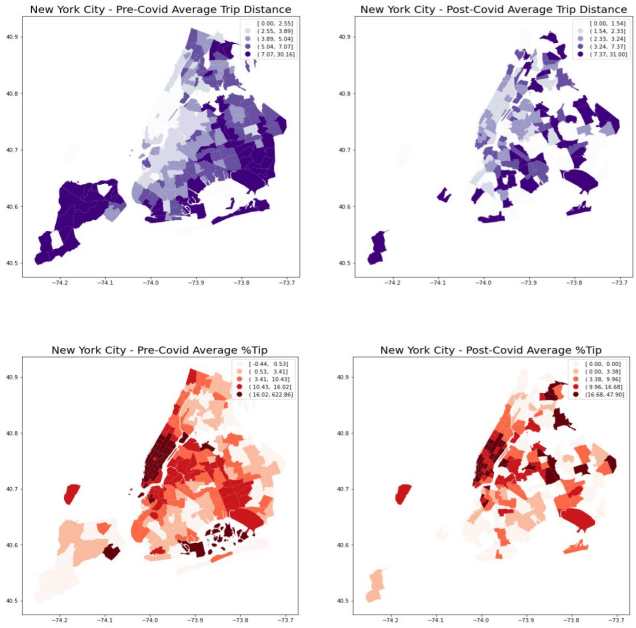
7.2 Geographical Observations

One of the major problems we set out to identify was the effect of COVID on specific areas of the city. Our first step in this process was to aggregate along each attribute for every pickup/dropoff location ID:

PUZone	AvgPassengerCount	AvgDistance	TotalCongestionSurcharge	fare_amount
Allerton/Pelham Gardens	1.850000	6.659000	0.00	571.90
Alphabet City	1.550383	2.611444	3750.00	20131.45
Arden Heights	1.000000	21.016667	0.00	392.00
Arrochar/Fort Wadsworth	1.000000	5.367500	0.00	73.00
Astoria	1.517592	2.625684	597.50	14382.90
...	...	...	...	...
Woodlawn/Wakefield	1.315789	3.803684	0.00	392.78
Woodside	1.581633	3.346204	342.50	8760.01
World Trade Center	1.656548	4.143918	12672.50	94419.59
Yorkville East	1.540060	2.315699	27355.75	128642.95
Yorkville West	1.538750	2.099601	40385.00	176743.57

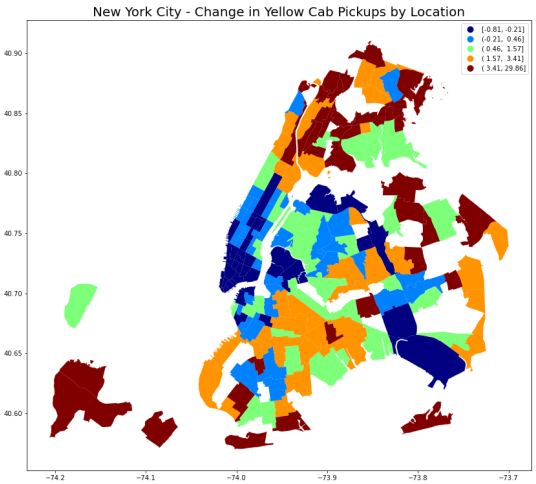
256 rows × 9 columns

Next, we split this derived data frame between pre- and post-lockdown and produced a heatmap for each attribute. Here are examples for average trip distance (above) and average %tip (below):



Just these two examples raise a couple of interesting questions: 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 (seemingly) so randomly in different locations?

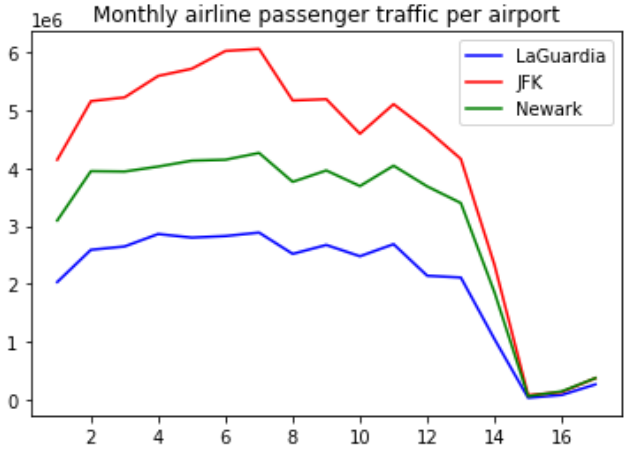
Finally we produced a color-coded map of the city indicating percent change between pre- and post-Covid for each location ID:



The map here indicates that in high density areas such as Brooklyn and Manhattan the percent change in ridership was far smaller than in less dense areas like Staten Island and Queens.

7.3 Airport Data

At the time of the lockdowns, many people chose to leave the city rather than stay in place. So naturally it was important for us to consider airport data alongside normal city traffic. We analyzed data from the three major airports in New York City:



From our examination of taxi and rideshare pickups and dropoffs at each airport, it is clear that rideshare was more affected than taxis in terms of airport business. It remains to be explained why that might be the case.

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