NYC Taxi Review

The Effects of Covid-19 on New York Cab Services

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

Benjamin Kuehner
University of Colorado Boulder
benjamin.kuehner@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

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

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 Proposed Work

3.1 Preliminary Tasks

The robust data available to us, while mostly accurate and complete, still contains a large number of inconsistencies that must be addressed. In addition to these necessary cleaning steps, we will also need to perform some basic sorting, labeling, and other preprocessing work before we can begin 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

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
- Generate derived columns to enable comparison between taxi and rideshare datasets

3.1.3 Integration:

- Aggregate data by date from datetime to reduce datasize 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 can begin the actual mining process. Since it is too early to make any predictions about what we will uncover during our analysis, the exact order of these steps cannot be given with any certainty. Therefore, we will instead present here a laundry list of all the mining techniques we expect to employ at some point 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

6 Milestones

Week 1 Cleaning, preprocessing, and integration

Week 2-4 Develop linear regression and multiple regression models, Pattern discovery, Make predictions using sample data and test against remaining data

Week 5 - Finish numerical analysis, Part 3 due - Results so far, graphs, correlations, etc.

Week 6 Thanksgiving break

Week 7 Finish final report, project code and descriptions, project presentation video

Week 8 Final Report, Peer evaluation and interview questions

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