# NY Taxi Review


Ben Kuehner, Tim Papich, Steve Putt

# Project Overview

The main objective of our project will be to assess the recent impact of COVID-19 on taxi services in New York City. Our multidimensional approach will 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, we also expect to uncover general trends such as revenue, traffic congestion by time and location, and the effect of competition from ride-sharing services in addition to our stated goal.

# Existing Research and Secondary Materials

Our analysis will build on a wealth of previous work on the taxi industry. At the present, we are focusing on the following three papers for their relevance to our research.

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.

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%.

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.

# The Dataset

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

# Preliminary Tasks

## Cleaning:

- Account for entry errors, very large trip lengths
- Correct for refunded trips, duplicate rows
- Control for extreme values
- Dataset features a large number of negative values

## Preprocessing:

- Filter for yellow cabs only within taxi data
- Create monthly groupings to compare seasonal changes
- Group and categorize variables to improve predictive modeling
- Generate  derived columns to enable comparison between taxi and rideshare datasets

## Integration:

- Taxi, location, ride share datasets merged/appended
- Aggregation for monthly or periodic trends
- Separating pre and post COVID times

# Tools and Frameworks

- Python
- Pandas
- Numpy
- Matplotlib
- SciPy
- Seaborn
- Sklearn
- Patsy

# Evaluation

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