Group 5
TLC Taxi Dataset

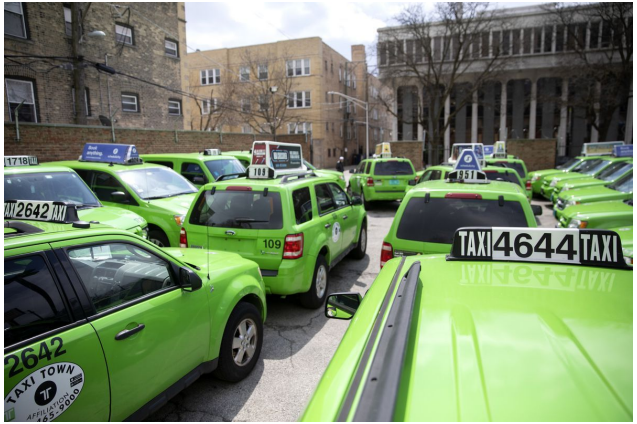
A comprehensive view of vehicle for hire data in NYC

Mike Huffman, Tommy Guess, Joe Froelicher

The TLC Dataset

- Trip records from Yellow Cab, Green Cab and Ride Share Taxis for each month
- Attributes include pickup/dropoff location (by zone), time, trip distance, number of passengers, fare amount, tipping, etc.
- The dataset is split into month-specific CSV files. Depending on the area of study, data preparation varies.

Cabs of NYC



	Yellow Cabs	*Green Cabs*
Total Rides:	*83,219,960*	*5,991,544*
Manhattan	91.70%	34.7%
Queens	6.89%	29.5%
Brooklyn	1.21%	28.7%
Bronx	0.19%	7.0%
Staten Island	0.004%	0.05%
Newark	0.01%	0.002%

Questions to Answer

- How are rides distributed, geographically?
- How can we figure out the relationship or “association” between pickup regions and drop-off regions?
- What factors influence tipping?
- How did COVID impact taxi rides in NYC?

Summary Matrix, S

Algorithm 1 Building a Summary Matrix (S) from Rideshare Data

Require: X is an array of rideshare data where X_k is the row vector representing the kth ride and all of its associated statistics. DO is a feature of that set (and column of X) representing the drop-off zone, such that $DO_k \epsilon \{1, 2, ..., 265\}$ and represents the drop-off zone of the kth ride. PU is a feature of that set representing the pick-up zone, such that $PU_k \epsilon \{1, 2, ..., 265\}$ and represents the pick-up zone of the kth ride. S is initialized as the zero matrix $0_{265,265}$

for every ride X_k in X **do**

$DO_k \rightarrow i$

$PU_k \rightarrow j$

$S_{i,j} += 1$

end for

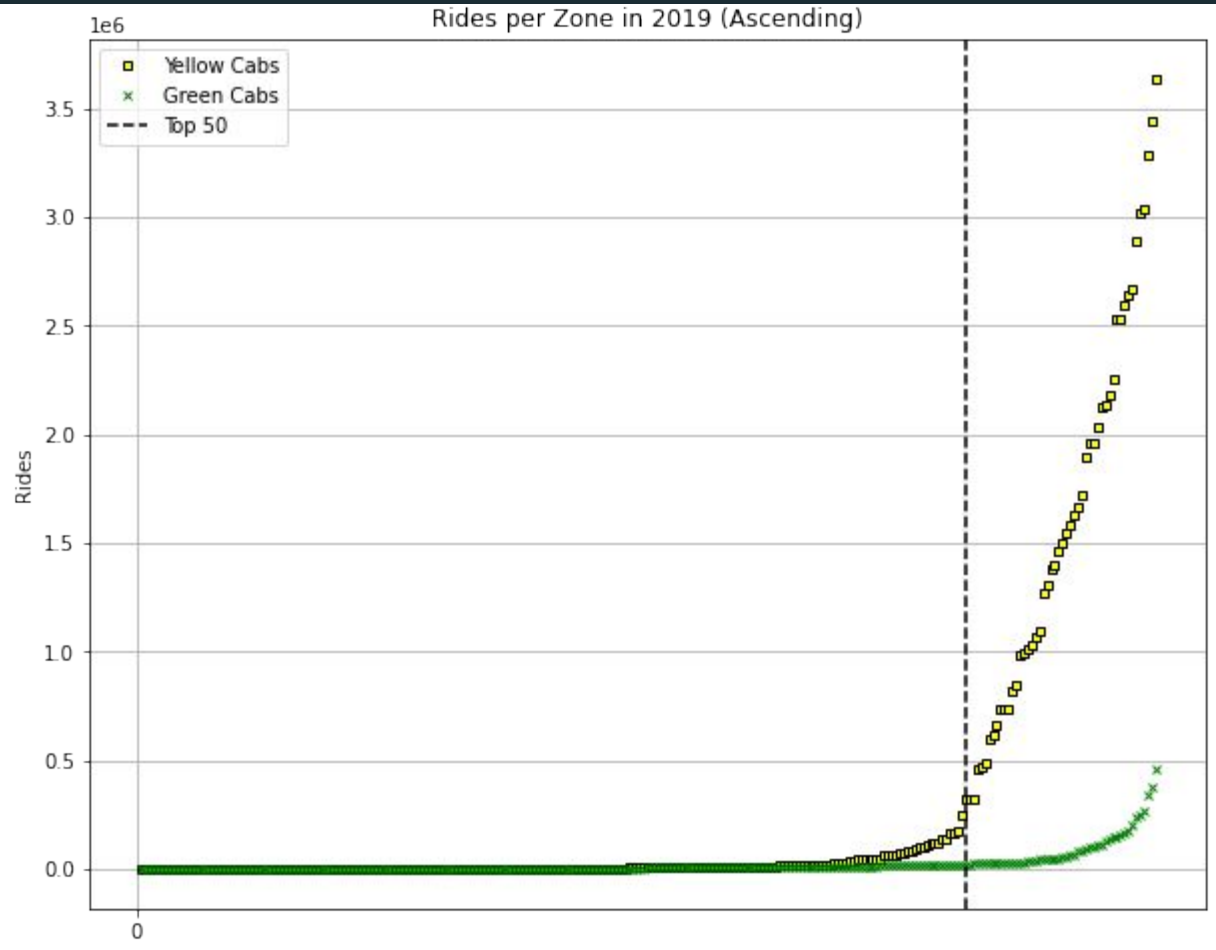
return S

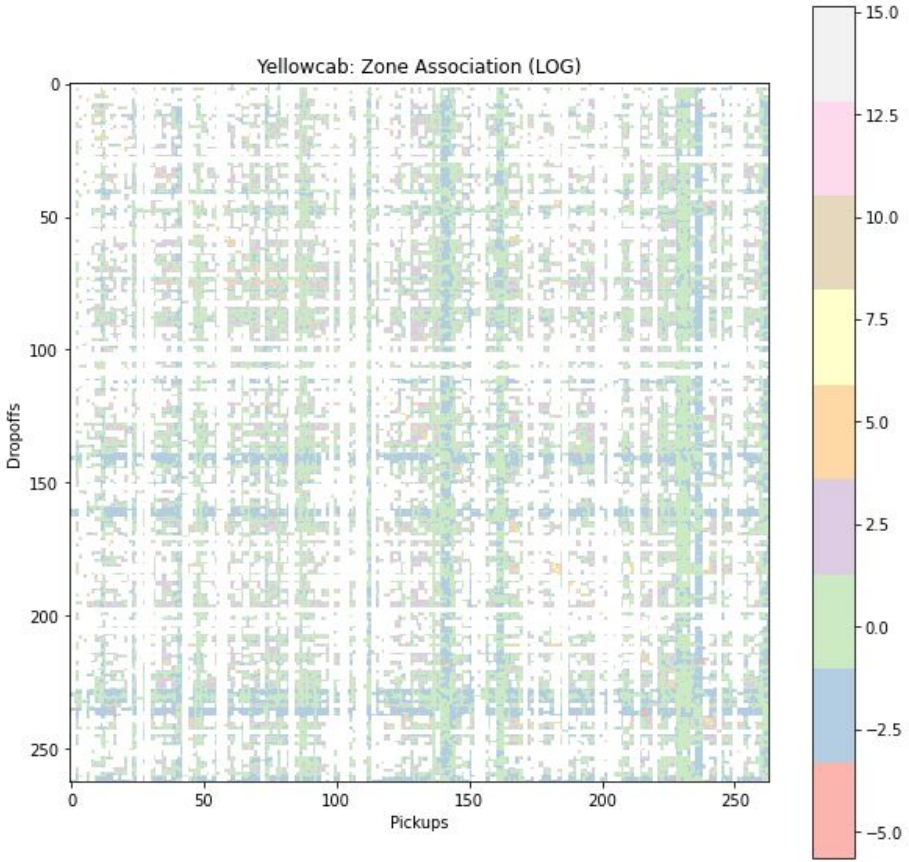
Ride Coefficient (r.c.)

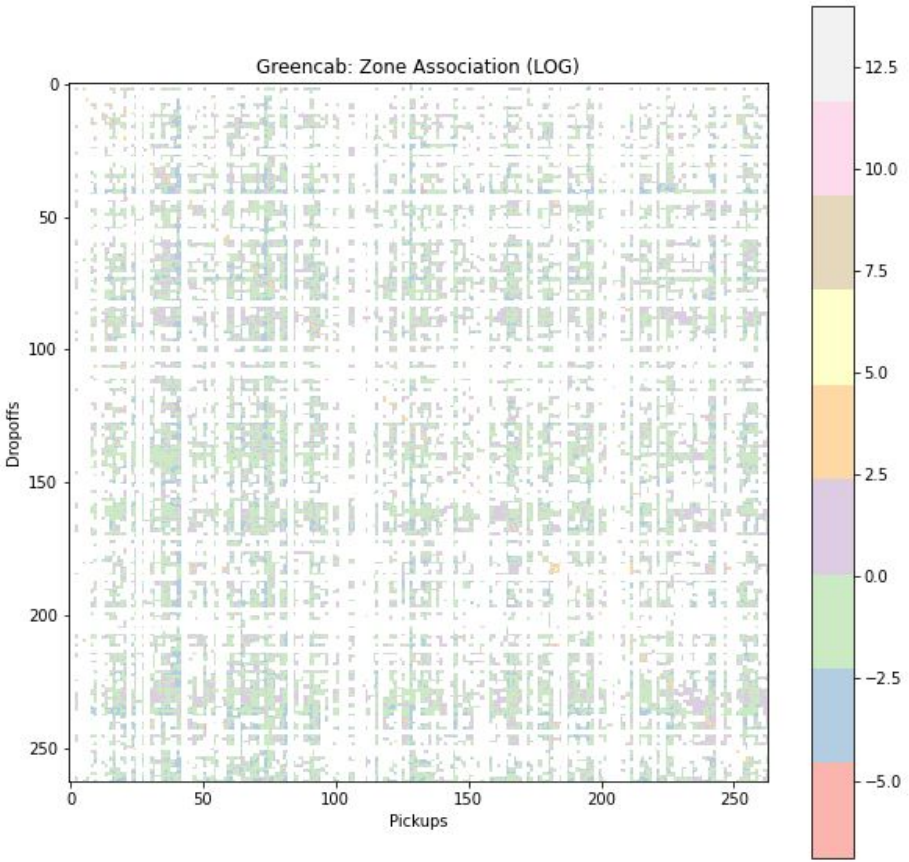
Let A be the event that ride pickups in zone *x*, and B be the event that the ride terminates in zone *y.*

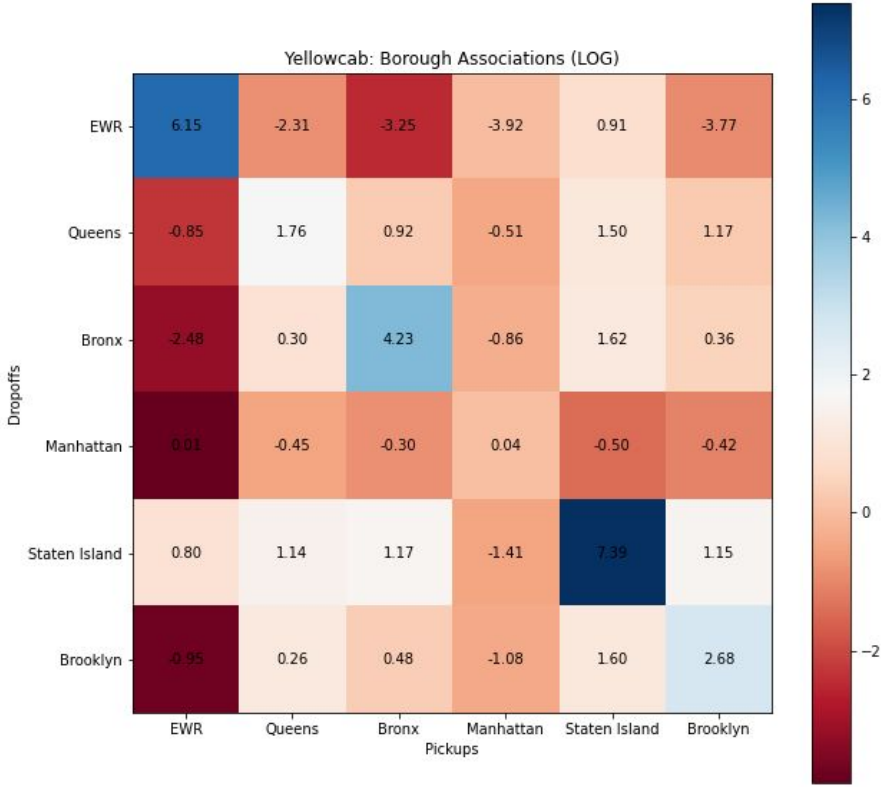
$$r.c. = \frac{P(B|A)}{P(B)}$$

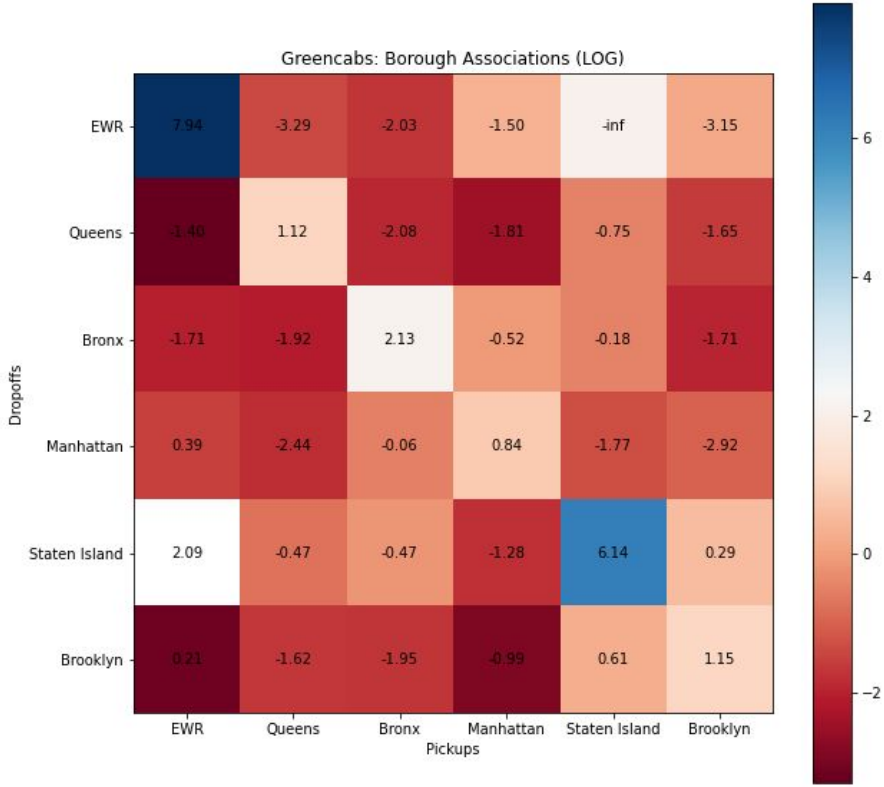
$$r.c. = \frac{P(A \cap B)}{P(A)P(B)}$$


Rides per Zone in 2019 (Ascending)
1e6
Yellow Cabs
Green Cabs
Top 50
Rides
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0


Yellowcab: Zone Association (LOG)
Dropoffs
Pickups
0
50
100
150
200
250
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0


Greencab: Zone Association (LOG)
Dropoffs
Pickups
0
50
100
150
200
250
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0


Yellowcab: Borough Associations (LOG)
EWR 6.15 -2.31 -3.25 -3.92 0.91 -3.77
Queens -0.85 1.76 0.92 -0.51 1.50 1.17
Bronx -2.48 0.30 4.23 -0.86 1.62 0.36
Manhattan 0.01 -0.45 -0.30 0.04 -0.50 -0.42
Staten Island 0.80 1.14 1.17 -1.41 7.39 1.15
Brooklyn -0.95 0.26 0.48 -1.08 1.60 2.68
Dropoffs
EWR Queens Bronx Manhattan Staten Island Brooklyn
Pickups
6 4 2 0 −2

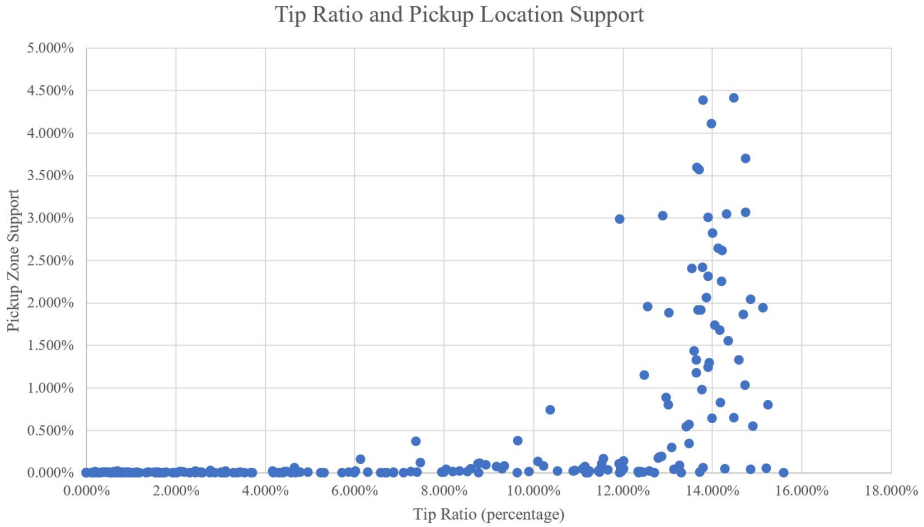

Greencabs: Borough Associations (LOG)
EWR 7.94 -3.29 -2.03 -1.50 -inf -3.15
Queens -1.40 1.12 -2.08 -1.81 -0.75 -1.65
Bronx -1.71 -1.92 2.13 -0.52 -0.18 -1.71
Manhattan 0.39 -2.44 -0.06 0.84 -1.77 -2.92
Staten Island 2.09 -0.47 -0.47 -1.28 6.14 0.29
Brooklyn 0.21 -1.62 -1.95 -0.99 0.61 1.15
Dropoffs
EWR Queens Bronx Manhattan Staten Island Brooklyn
Pickups
6 4 2 0 −2

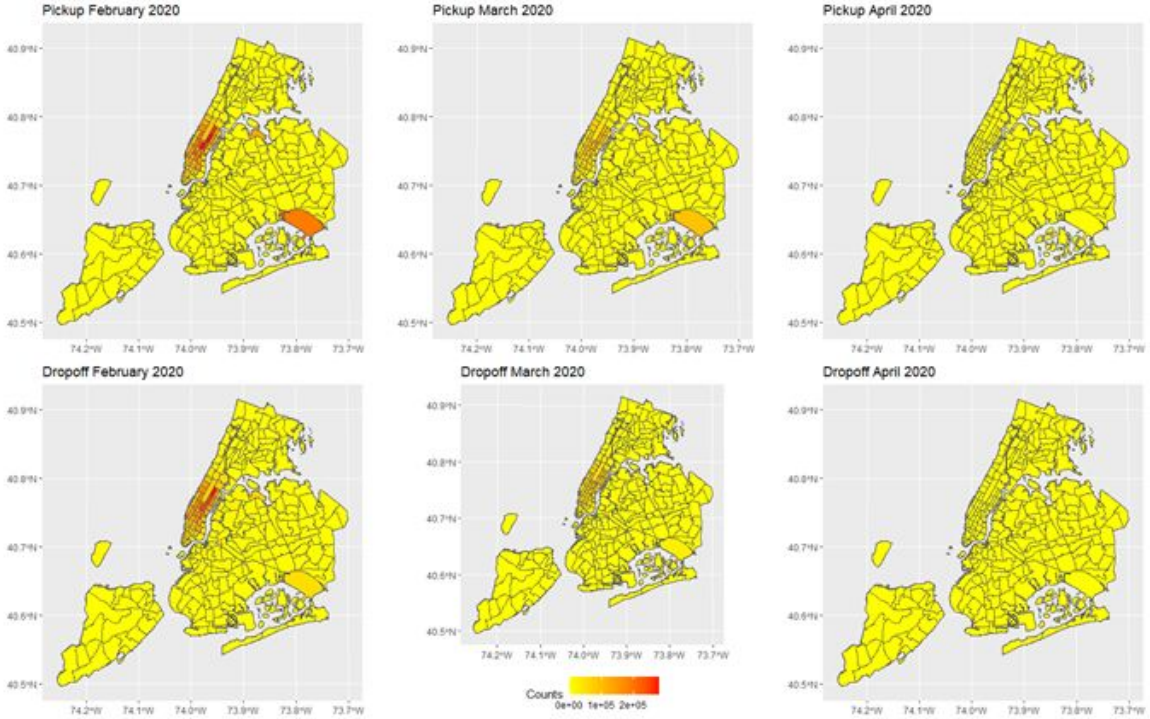
Financial Analysis

- Several attributes:
  - Tolls amount, improvement surcharge, MTA tax, etc.
- Most interesting: tips
- Rider has control over the tips
- Could potentially be a proxy for satisfaction
- Preliminary analysis included correlation matrix

Tip Ratio

- Tip as a percentage of fare





Maps: Normal cab patterns vs. abnormal patterns

What can we learn from the maps?

- Normal yellow cab traffic is primarily in Manhattan, and at La Guardia
- Color gradient from before the lockdown to during the lockdown
- Number of rides was negligible in those high trafficked districts

Covid-19 Effects on Yellow Taxi Rides

- Startling (but unsurprising) drop in the number of daily rides
- Normal weekly fluctuations diminished

What is the economic impact?

- Thousands of dollars change hands as a result of daily yellow cab rides

What other economic impacts might the fall of in yellow cab activity indicate?

- Where are people riding cabs going?
- Will they be spending money where they go?