Analyzing current taxi usage patterns in NYC to detect local hot spots and dead zones

Nasser Zalmout
Tandon School of Engineering

Jacqueline Gutman
Center for Data Science

Abstract and dataset summary

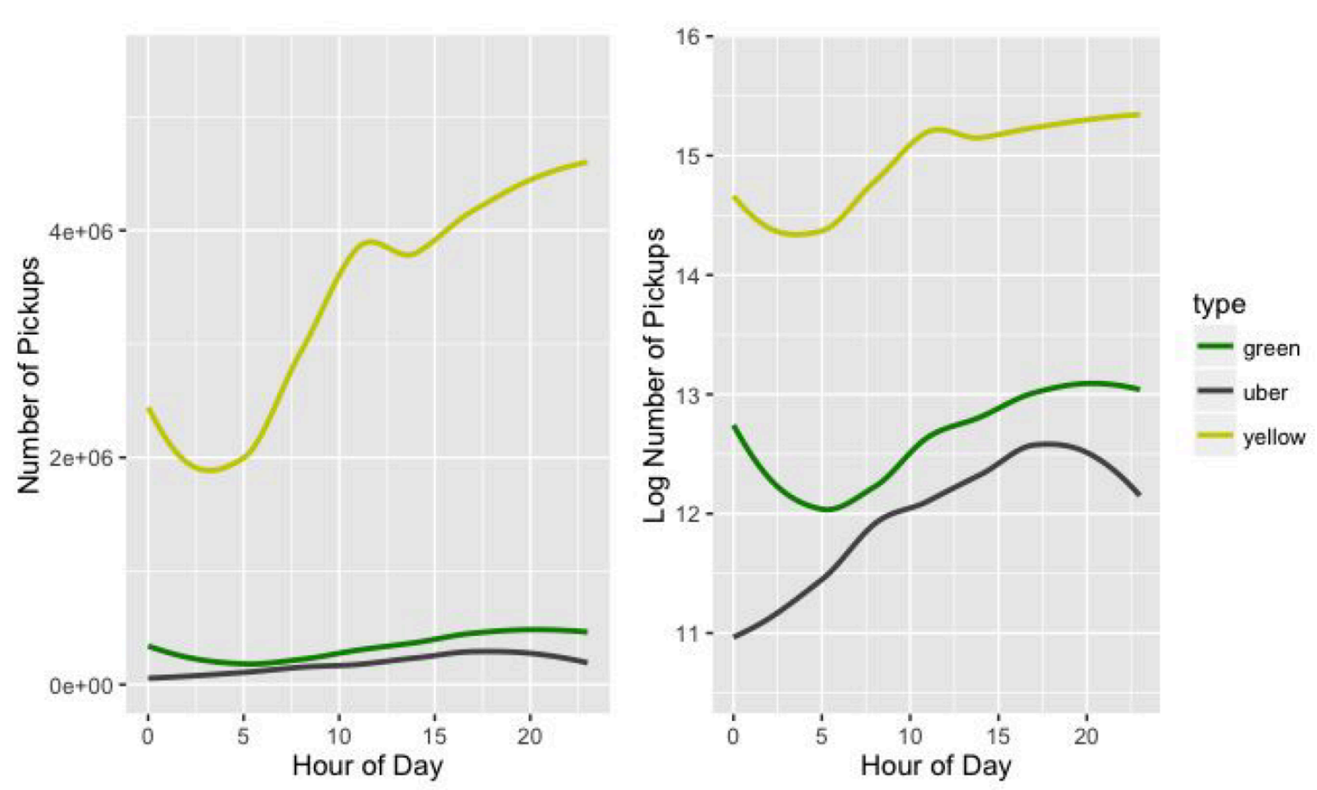
Abstract
The goal is to analyze the trends of the public taxi services in New York City, and investigating the dynamics of the market shares of each taxi service, including yellow cabs, green cabs and Uber.

Datasets
Taxi trip records from Uber, green and yellow cabs.
Other datasets: NYC shapefile, zip-code mapping and subway locations

Data preprocessing
- eliminating pickup locations outside NYC boundaries
- calculating/mapping locations to zip-codes/neighborhoods
- calculating distance to nearest subway station entrance

Day of the week	***Yellow cabs***	***Green cabs***	***Uber***
Sunday	10,602,423	1,237,063	470,988
Monday	10,307,896	919,565	525,777
Tuesday	11,699,072	992,684	648,668
Wednesday	11,702,437	1,023,029	681,205
Thursday	11,978,915	1,097,963	737,303
Friday	12,073,152	1,278,011	722,123
Saturday	11,948,311	1,473,385	625,254
total	80,312,206	8,021,700	4,411,318

Shifting patterns in taxi pickups by time of day

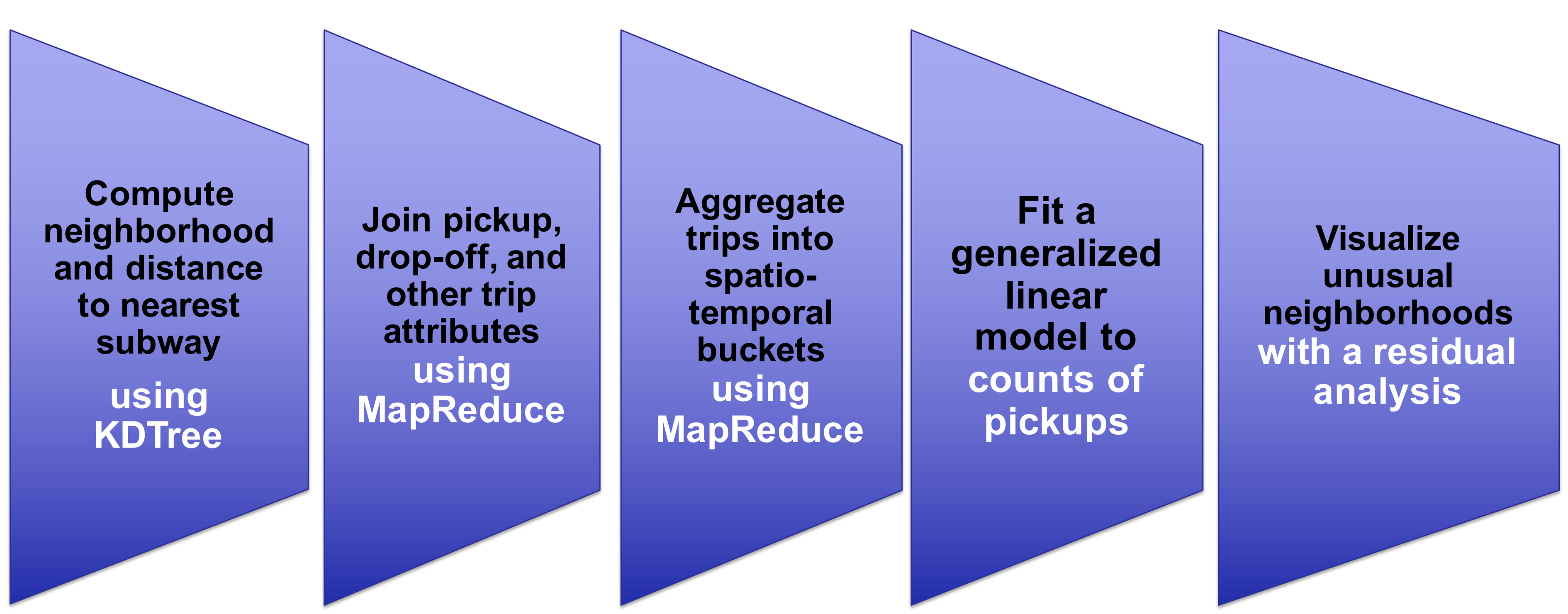


Patterns in number of taxi cab pickups for all taxi services from midnight to midnight

Original data 93 million trip records
Aggregated data 23,600 spatiotemporal buckets by taxi type

We can now analyze the aggregated data using either Spark, or local machines running Python/ R (small data!)

Data analysis pipeline



Poisson regressions and outlier detection

Outcome variable Number of pickups within a spatiotemporal bucket defined by the interaction of:
- taxi type (yellow, green, or Uber)
- neighborhood (e.g. Midtown, Bushwick/Williamsburg)
- day of the week
- hour of the day

These count variables have a conditional **Poisson** distribution
We can estimate expected counts $\left(\lambda_i \mid \mathbf{x_i}\right)$ using a generalized linear model

$$\log\left(\lambda_i \mid \mathbf{x_i}\right) = \mathbf{x_i}^T \beta + \varepsilon \qquad y_i \sim Poisson\left(\lambda_i\right)$$

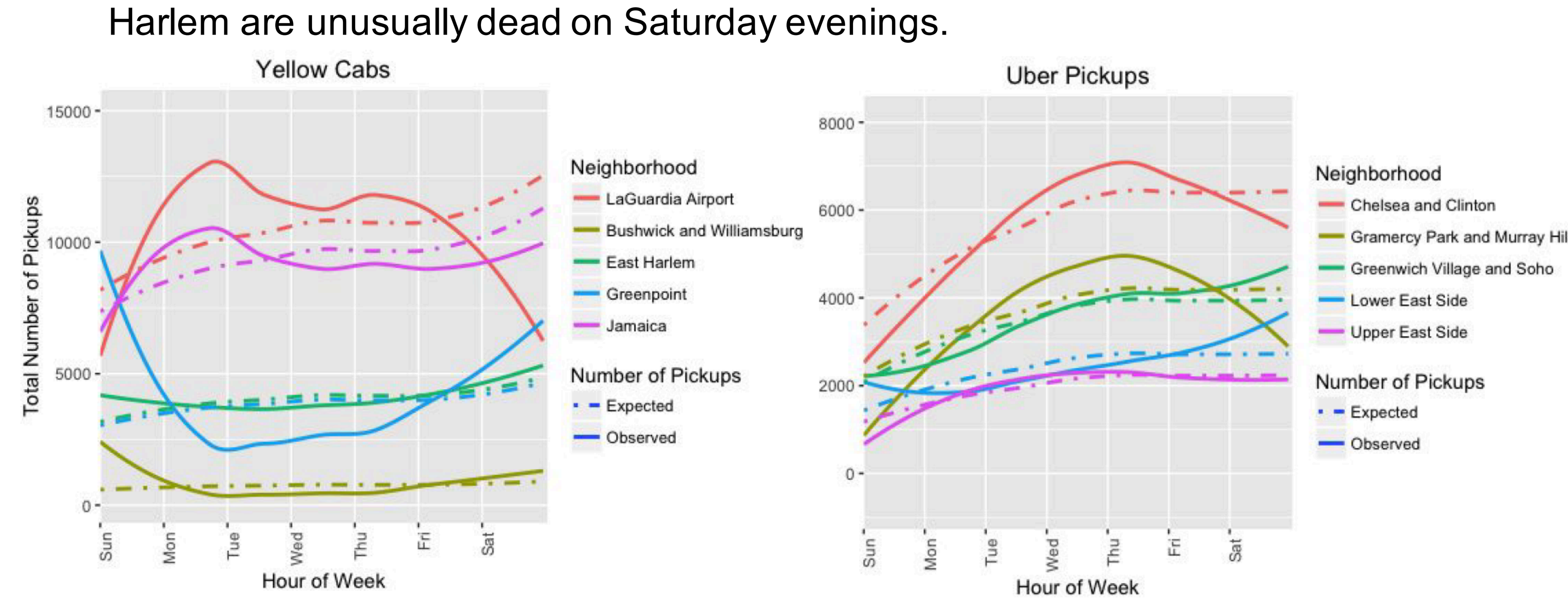
Predictors
- Day of the week (7 levels)
- Neighborhood (52 levels)
- Taxi type (3 levels)
- Weekend indicator (2 levels)
- Borough (5 levels)
- Average distance to subway at pickup location (continuous)
- Hour of the day Sine + Cosine transformation (continuous)
- Interaction terms by taxi type (130 interaction terms)

1 bias term + 3 continuous measures + 64 categorical measures + 130 interaction terms = **197 predictors + intercept**

➢ **After Lasso** 157 non-zero predictors

We considered two general procedures for estimating the conditional means (expected number of pickups within a particular spatiotemporal bucket)
➢ **stepwise regression** with AIC feature selection criterion
➢ **cross-validated Lasso** penalized regression (more robust predictions)

Residual analysis and outlier detection
We can identify neighborhoods that are *unusually **hot** or unusually **dead*** at particular times of day (e.g. LaGuardia Airport is unusually busy on Monday mornings, Bushwick/Williamsburg is unusually busy late Friday nights, Central Park and East Harlem are unusually dead on Saturday evenings.



We can visualize over the 168 week-hours, how selected neighborhoods deviate from typical weekly and hourly patterns of ebbs and flows in the number of pickups.

Does distance to subway matter? YES!

Yellow	**Model 1:** count ~ weekday + borough + hour.sin + hour.cos + avg.num.passengers + avg.distance.traveled + is.weekend	**P-value (models compared)**
	Model 2: count ~ weekday + borough + avg.pickup.dist.subway + hour.sin + hour.cos + avg.num.passengers + avg.distance.traveled + is.weekend	p < 2.2e-16 ***
	Model 3: count ~ weekday + borough + avg.pickup.dist.subway + avg.dropoff.dist.subway + hour.sin + hour.cos + avg.num.passengers + avg.distance.traveled + is.weekend	p < 2.2e-16 ***
Green	**Model 1:** count ~ weekday + borough + hour.sin + hour.cos + avg.num.passengers + avg.distance.traveled + is.weekend	**P-value (models compared)**
	Model 2: count ~ weekday + borough + avg.pickup.dist.subway + hour.sin + hour.cos + avg.num.passengers + avg.distance.traveled + is.weekend	p < 2.2e-16 ***
	Model 3: count ~ weekday + borough + avg.pickup.dist.subway + avg.dropoff.dist.subway + hour.sin + hour.cos + avg.num.passengers + avg.distance.traveled + is.weekend	p < 2.2e-16 ***
Uber	**Model 1:** count ~ weekday + borough + hour.sin + hour.cos + is.weekend	**P-value (models compared)**
	Model 2: count ~ weekday + borough + avg.pickup.dist.subway + hour.sin + hour.cos + is.weekend	p < 2.2e-16 ***

Deviations from expected pickup counts

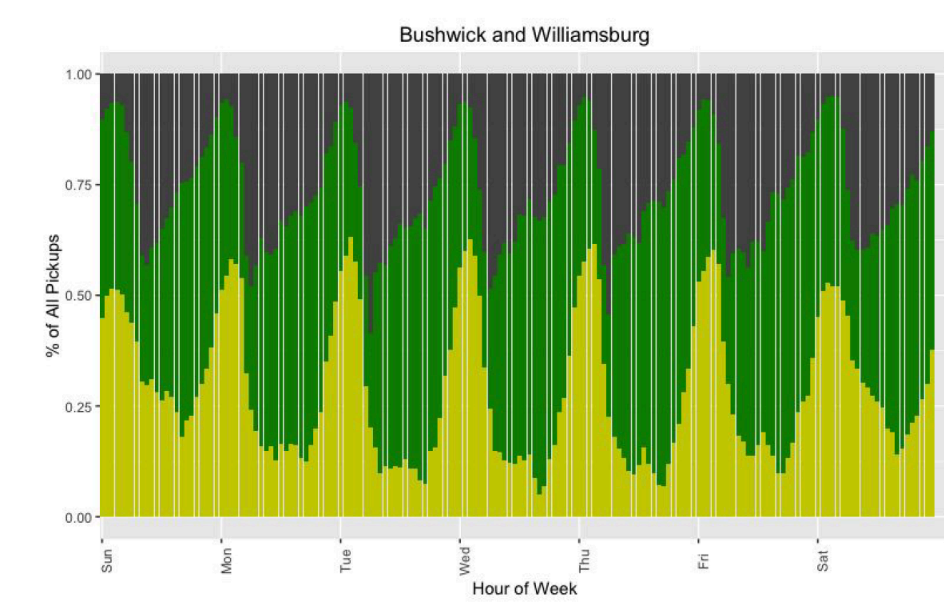
How do we detect outliers?
- Calculate *standardized Pearson residual* : $\frac{O-E}{sd\left(E\right)}$
- For each taxi type, examine the spatiotemporal buckets with residuals in the *top or bottom half percent of all data*

➢ Large positive residuals: **hot spots**
➢ Large negative residuals: **dead zones**

Neighborhood	Time	Day	Type	Observed	Expected
Bushwick and Williamsburg	Midnight - 3 am	Sun	green	12am: 4497, 1am: 4803, 2am: 4482	12am: 1001, 1am: 889, 2am: 796
Gramercy Park	8am	Tue, Wed	yellow	Tue: 134368, Wed: 134459	Tue: 66320, Wed: 66333
Gramercy Park	5pm	Tue, Wed	Uber	Tue: 12791, Wed: 12581	Tue: 6350, Wed: 6668
Greenpoint	11pm - 3am	Fri - Sat	green	11pm: 18279, 12am: 19749, 1am: 20337, 2am: 19620	11pm: 8531, 12am: 6348, 1am: 5593, 2am: 4925
Upper East Side	6am - 9am	Mon, Tue, Wed	Uber	Tue 6am: 3769, Tue 7am: 5502, Tue 8am: 4547	Tue 6am: 1055, Tue 7am: 1141, Tue 8am: 1276

Neighborhood hot spots A selection of neighborhoods that were overall busier than expected during selected hours of the week

Peak service hours for each taxi type



Relative frequency of pickups in Bushwick / Williamsburg, Brooklyn

X-axis begins at midnight for each day

We can also use our tool to visualize when the relative prevalence of different taxi services within a neighborhood tends to shift. For example, in Williamsburg, yellow cab availability peaks around midnight each day, while green cab pickups really dominate during the afternoons and early evenings. Uber is less popular, but it does best in the morning, when yellow cabs are relatively scarce.

Summary

- Using **KD Trees and K Nearest Neighbors**, we can quickly identify the borough and neighborhood of each taxicab pickup, as well as the **distance from the pickup (and dropoff) location to the nearest subway station**
- Using MapReduce, we can aggregate all trip records into spatiotemporal buckets defined by hour of the day, day of the week, and neighborhood to drastically ***reduce the size of the data to 2.5*10⁻⁴ times its original size***
- Patterns of taxi pickups vary drastically by hour of the day and day of the week, and these **ebbs and flows in taxi pickups are *not always identical*** **across neighborhoods**
- **Average distance to subway at pickup location is an important feature** for predicting pickup counts for all taxicab services
- Using residual analysis and outlier detection, we can **identify neighborhoods that are unusually busy** at certain times of the week (for example, Saturday nights, or Wednesday morning rush hour) –can query all neighborhoods in Manhattan that are busier than usual on Fridays after 10 pm, for example