WHERE DO YANKEES AND METS FANS LIVE IN NYC?

David Kaczmarkiewicz

TABLE OF CONTENTS

OBJECTIVE

Explore taxi trips originating from the zones including Yankee Stadium and Citi Field after home games and observe where passengers are dropped off.

Use a model based approach to filter out normal, every day, common taxi trips from these areas so the only trips being included should include novel taxi rides due to the games and the fans attending them.

DATA SOURCES

2019 Yellow Taxi Trip Data - NYC OpenData

2019 Green Taxi Trip Data - NYC OpenData

Taxi Zones and IDs - NYC OpenData

2019 MLB Regular Season Schedule - retrosheet.org

DATASET EVALUATION 1 – MLB GAME DATA

index	Date	Doubleheader	DayOfWeek	VisitingTeam	LeagueVisit	HomeTeam	Time	Postponement	MakeupDate
62	4/1/2019	0	Mon	DET	AL	NYA	n	None	None
73	4/2/2019	0	Tue	DET	AL	NYA	n	None	None
194	4/12/2019	0	Fri	CHA	AL	NYA	d	None	None

- Game times were not included in dataset only a field representing whether the game was played during the day or night.
  - Only night games were included in the dataset. They have a consistent start time of around 7pm, and my assumption is more people are heading home after those games as opposed to afternoon games.

- Games that were postponed or rained out and ended early were removed from the dataset which accounted for less than 5 instances for both the Yankees and the Mets.

Dataset evaluation II - Taxi Data

- Combine Datasets - Created SQL database from both Green and Yellow Taxi data due to combined size of datasets roughly 15gb.
  - Converted Pickup and Drop off dates and times to datetime.
  - Dropped two specific fields unique to Green Taxi Data.

- Created two queries to pull all trips originating from either Yankee Stadium or Citi Field in 2019.

Dataset evaluation II – Taxi Data

- Taking a closer look at the data... Besides the pick up and drop off location ID's and times these are the remaining fields available in the taxi data:

	Data Type	Description	Keep or Drop?
passenger_count	Integer, 1 - 6	Passenger count, entered by driver	Keep
trip_distance	Continuous	Trip distance in miles, reported by taximeter	Keep
RatecodeID	Categorical	Signifies special rates to airports, suburbs, etc.	Keep
payment_type	Categorical	How customer paid, cash, credit, etc.	Keep
fare_amount	Continuous	Fare calculated by time and distance	Keep
tip_amount	Continuous	Auto calculated by credit cards, not reported for cash payments	Keep
VendorID	Categorical	Represents company reporting data	Drop
store_and_fwd_flag	Categorical	Flags trips uploaded from car memory, car not connected to network at time of ride.	Drop
extra	Continuous	Extra surcharges, includes only rush hour and overnight charges.	Drop
mta_tax	Single Value	$0.50 automatically applied by taxi in use.	Drop
tolls_amount	Continuous	Amount of tolls	Drop
total_amount	Continuous	Final price of ride	Drop

Dataset evaluation II - missing data

Some of the fields I decided to keep, had missing data in some observations: passenger count, Ratecode ID, and payment type. The rows with missing data only accounted for roughly 1% of the data for each query.

- **Passenger Count** - 90% of the data had a single passenger, so missing values were replaced with a value of 1.
- **Ratecode ID** - Signifies a special rate to airport or suburb. This similarly was replaced with the value of 1 which was roughly 95% of the entries.
- **Payment Type** - Cash/Credit/etc. This field was much more diverse, not wanting to influence the model, I dropped these observations.

DATASET EVALUATION II - OUTLIER DETECTION

- 5 points in the dataset included negative distances, these points were removed.
- One passenger tipped $100 on their fare, this observation was removed.

MODEL BUILDING -

HOW DO WE FIND GAME DAY TRAFFIC?

- **Classification Problem** - For each trip after a game ends, we need to determine if that trip was a normal trip, or if it was caused by the event.
- **Labeled Data** - We have some labeled data, but only of one class. We can look at days when there was no game, or maybe even more representative, we can look at days the team is playing an away game. These can be labeled as normal, common traffic that we wish to filter out when looking at game day traffic.
- **One Class Support Vector Machines** - An unsupervised technique that finds similarities between training data, and can classify new data as similar or different from the training data.

MODEL BUILDING -

A LITTLE MORE DATA PREP

- Created Indicator Variables for all categorical data.
- Added drop off latitude and longitude to datasets from taxi zone ID's.

Split up training and testing datasets

- Train Data was filtered to include away game dates between 9PM - 12PM.
- Test Data was filtered to include home game dates between 9PM - 12PM.

Both datasets normalized ((value - mean) / 2 * std dev) to the training data.

MODEL BUILDING -

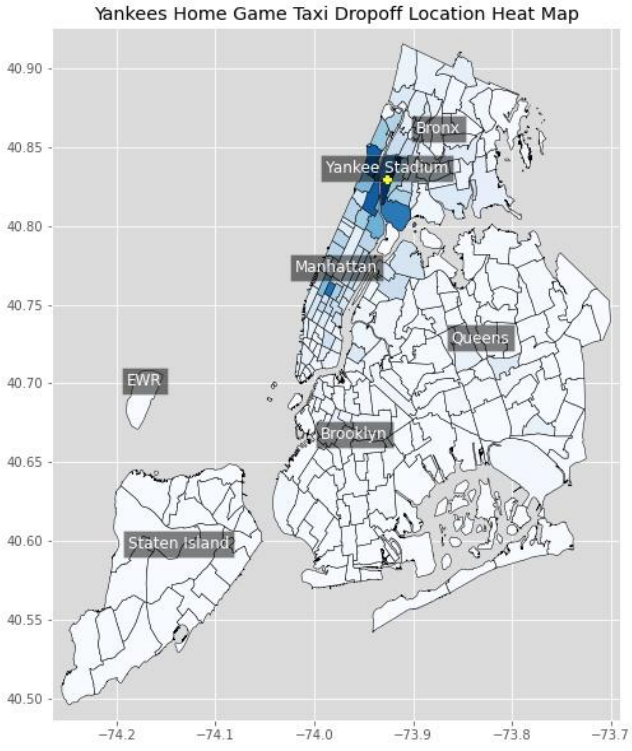
TRAIN THE MODEL

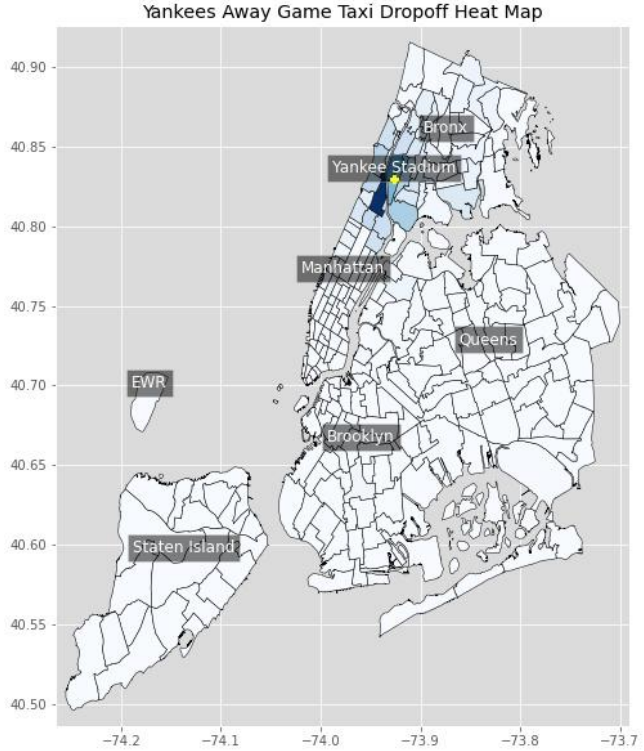
- Plug the training data into the model.
- Tune Hyperparameters
  - Kernel: Linear, Polynomial, Radial Basis Function
  - C Parameter, results in more or less regularization

Evaluate Results:

- All 3 kernels resulted in similar results. The C parameter tuning acts as a sliding scale for when to classify a point as similar to the training data or as an outlier, game day influenced taxi ride.
- I tuned the C parameter under the assumption that normal taxi traffic would be roughly the same on game days, so the increase in taxi traffic on those days should be classified as game day traffic.

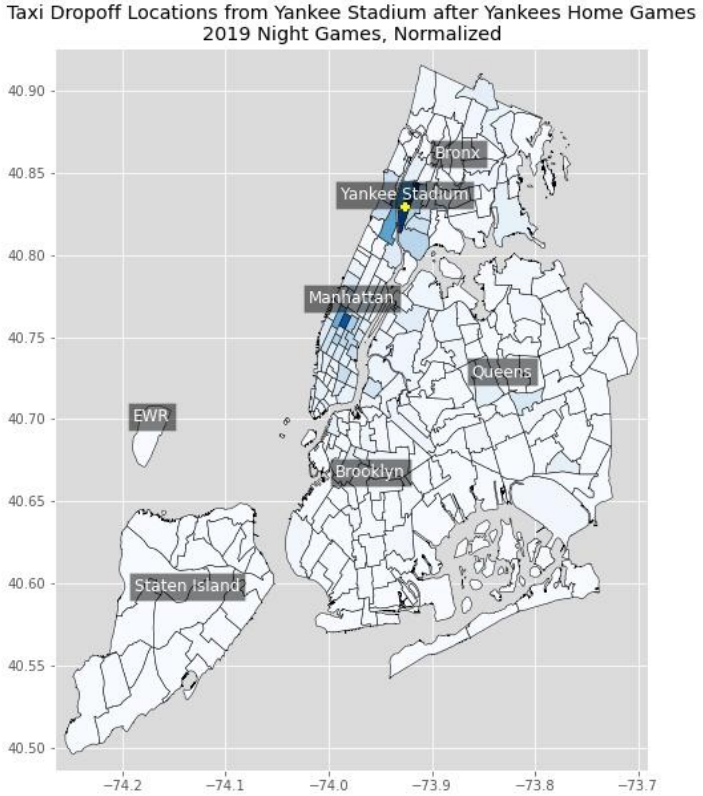
RESULTS – YANKEES



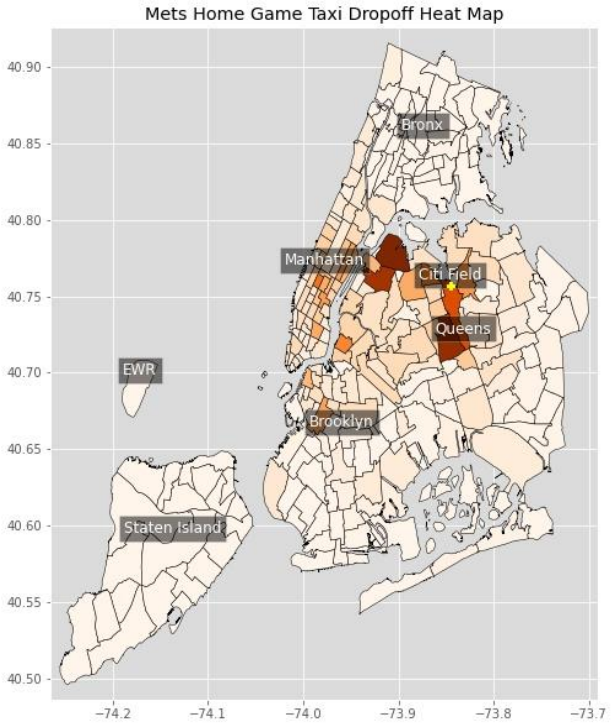


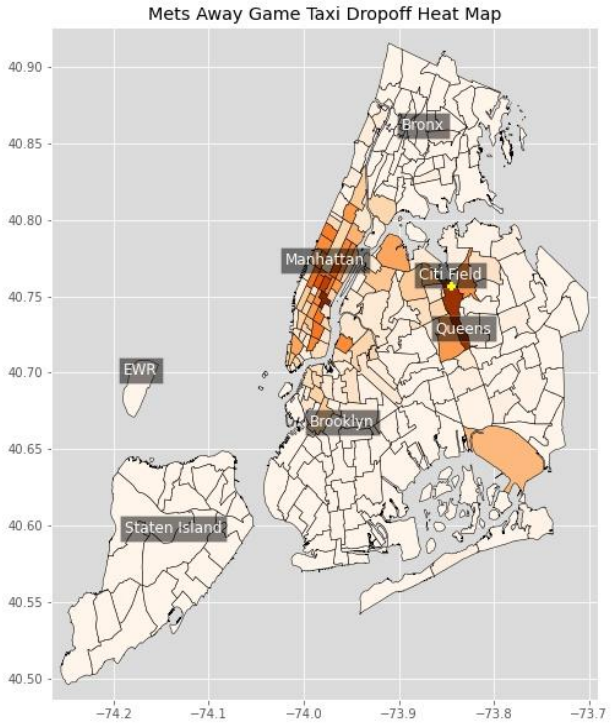
RESULTS – YANKEES

- High concentration around Yankee stadium.
- Times Square and surrounding zones also has a high concentration of drop offs.
- Smaller numbers through the northern and eastern parts of the Bronx as well as Queens.



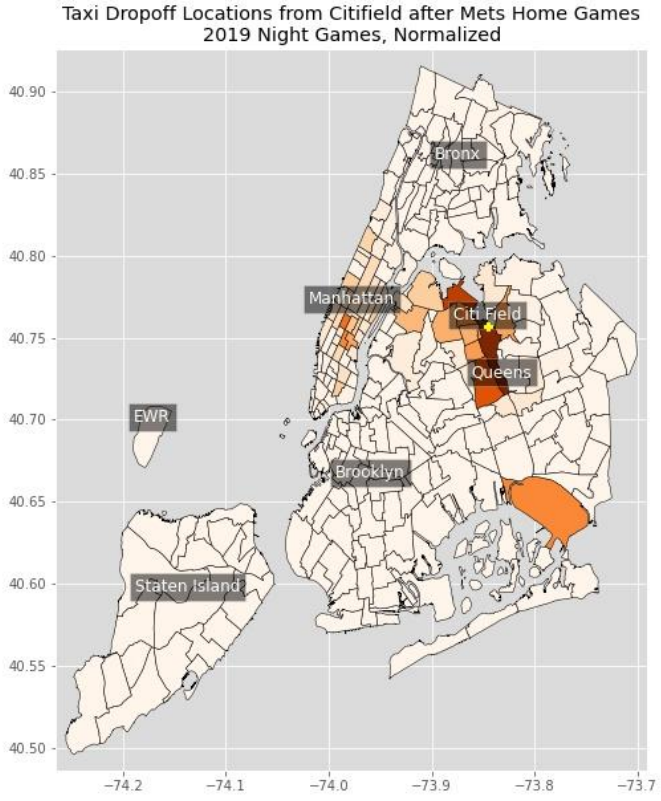
Results – Mets



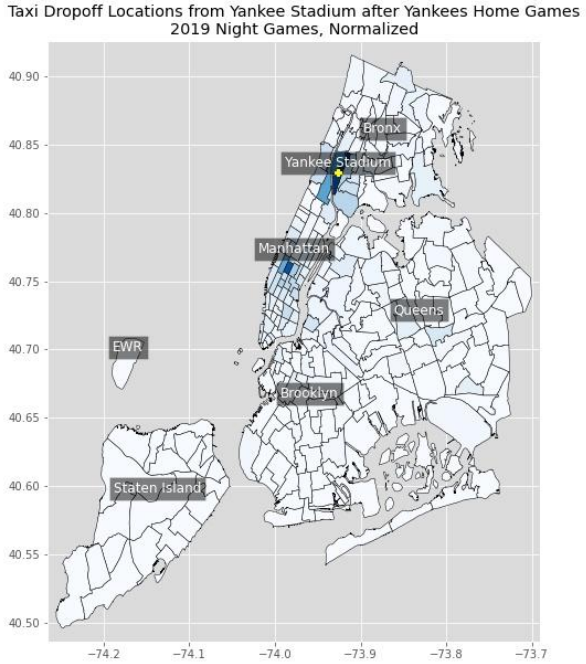


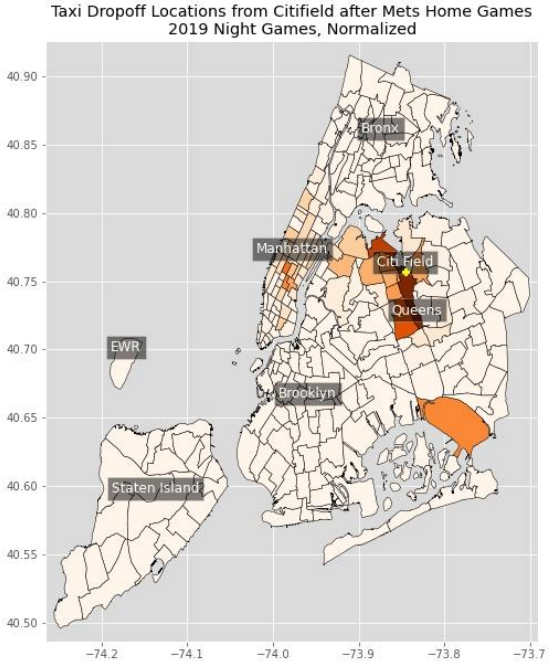
RESULTS – YANKEES

- High concentration around Citi Field.
- Times Square and surrounding zones also has a high concentration of drop offs.
- Smaller numbers throughout the rest of the city.



END RESULTS


Taxi Dropoff Locations from Yankee Stadium after Yankees Home Games
2019 Night Games, Normalized
Bronx
Yankee Stadium
Manhattan
Queens
EWR
Brooklyn
Staten Island
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7


Taxi Dropoff Locations from Citifield after Mets Home Games
2019 Night Games, Normalized
Bronx
Manhattan
Citi Field
Queens
EWR
Brooklyn
Staten Island
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7

THANK YOU

Do you have any questions?

dave.kacz@gmail.com
917-817-6872

(Brooklyn and Staten Island are so under represented in this analysis, they need the Dodgers back.)