Predicting crime rates using taxi rides in NYC

Carlos Petricioli Varsha Muralidharan Valerie Angulo

New York University
{cpa253,vm1370,vaa238}@nyu.edu

12/15/2017

Big Data Analytics Symposium - Fall 2017

Analytics Project

Predicting crime rates using taxi rides in NYC

Team

- Carlos Petricioli (cpa253)
- Varsha Muralidharan (vm1370)
- Valerie Angulo (vaa238)

Abstract

- Our study looks at the relationship between crime rates and taxi usage in New York City.
- Our hypothesis is that people are less likely to walk in areas subjectively deemed more dangerous and will instead opt to use more reliable and immediate transportation such as designated taxis.
- We found evidence that supports that our hypothesis holds.

Motivation

Typical users of this application

The scientific community, law enforcement, those in public transportation

Beneficiaries of this application

Members of the community and tourists, law enforcement

Importance of this analytic

- This analytic can help law enforcement predict areas of crime based on New Yorkers transportation habits. Law enforcement officials may be able to predict which areas will have a higher rate of crime in the future.
- People who live in an area are aware of the safety of their surroundings and this awareness can be represented by how comfortable residents may be in walking or taking the subway versus taking more immediate, more expensive, modes of transportation such as taxis.
- This analytic can benefit the community and tourists by influencing their current and future transportation behaviors

Data Sources

Taxi rides data from TLC (*Link*)

- It covers years from 2009 to June 2017.
- The yellow taxi trip records include:
  - pick-up and drop-off dates/times,
  - pick-up and drop-off locations,
  - trip distance,
  - itemized fares,
  - rate types,
  - payment type,
  - passenger counts.
- *Data Size*: 250 GB

NYPD Complaint Data (*Link 1, Link 2*)

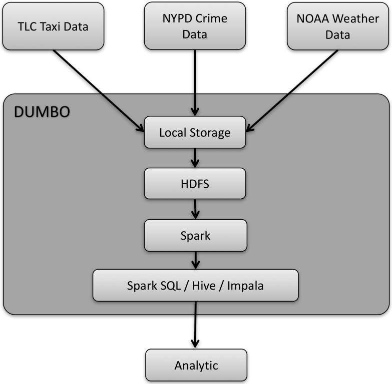
- This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to year to date data.
- *Data Size*: 1.5 GB

NOAA Weather stations data *(Link)*

The *Integrated Surface Database (ISD)* consists of global hourly ansynoptic observations compiled from numerous sources into a single common ASCII format and common data model.

- ISD's complete history of hour-by-hour readings for one user-specified weather stations
- We selected:
  - Central Park
  - JFK
  - Laguardia
- *Data Size*: 165 MB

Design Diagram



Platform:

- NYU HPC cluster (Dumbo)

Obstacles

Cleaning the data: Taxis

- The **taxi data** (250 GB) was the most challenging to clean.

Obstacles

Cleaning the data: Taxis

- The **taxi data** (250 GB) was the most challenging to clean.
- Inconsistencies in columns: extra columns for some years.

Obstacles

Cleaning the data: Taxis

- The **taxi data** (250 GB) was the most challenging to clean.
- Inconsistencies in columns: extra columns for some years.
- Rows with extra commas: avoiding an easy parse.

Obstacles

Cleaning the data: Taxis

- The **taxi data** (250 GB) was the most challenging to clean.
- Inconsistencies in columns: extra columns for some years.
- Rows with extra commas: avoiding an easy parse.
- Row values for each year were not that dirty but the data values were completely different for different years.

Obstacles

Cleaning the data: Taxis

- The **taxi data** (250 GB) was the most challenging to clean.
- Inconsistencies in columns: extra columns for some years.
- Rows with extra commas: avoiding an easy parse.
- Row values for each year were not that dirty but the data values were completely different for different years.
- The dictionary that defines the labels refers to the data from 2017, so we needed to figure out the meaning of labels for previous years.

Cleaning the data: Taxis

- To figure this out we needed to iterate through every row of the data because new things came up every time we thought we were done with the cleaning.

Cleaning the data: Taxis

- To figure this out we needed to iterate through every row of the data because new things came up every time we thought we were done with the cleaning.
- Even after we cleaned all the categorical variables and we thought we were done, many numerical inconsistencies appeared.

Cleaning the data: Taxis

- To figure this out we needed to iterate through every row of the data because new things came up every time we thought we were done with the cleaning.
- Even after we cleaned all the categorical variables and we thought we were done, many numerical inconsistencies appeared.
- Longitude/Latitude, like not even in NY or simply null.

Cleaning the data: Taxis

- To figure this out we needed to iterate through every row of the data because new things came up every time we thought we were done with the cleaning.
- Even after we cleaned all the categorical variables and we thought we were done, many numerical inconsistencies appeared.
- Longitude/Latitude, like not even in NY or simply null.
- Negative, but consistent values for amounts.

Cleaning the data: Taxis

- To figure this out we needed to iterate through every row of the data because new things came up every time we thought we were done with the cleaning.
- Even after we cleaned all the categorical variables and we thought we were done, many numerical inconsistencies appeared.
- Longitude/Latitude, like not even in NY or simply null.
- Negative, but consistent values for amounts.
- Weird trip distances, like greater than 1000 miles.

Cleaning the data: Taxis

- To figure this out we needed to iterate through every row of the data because new things came up every time we thought we were done with the cleaning.
- Even after we cleaned all the categorical variables and we thought we were done, many numerical inconsistencies appeared.
- Longitude/Latitude, like not even in NY or simply null.
- Negative, but consistent values for amounts.
- Weird trip distances, like greater than 1000 miles.
- Exorbitant total amounts (which might not be a problem because most were negotiated fares)

Cleaning the data: Taxis

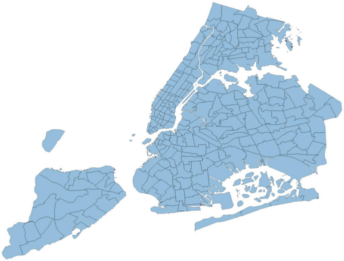
- 2016 and 2017 do not have longitude and latitude, just zone id.

Cleaning the data: Taxis

- 2016 and 2017 do not have longitude and latitude, just zone id.
- This became one of the greater obstacles, because we needed to assign a zone id to all the previous years.

Cleaning the data: Taxis

- 2016 and 2017 do not have longitude and latitude, just zone id.
- This became one of the greater obstacles, because we needed to assign a zone id to all the previous years.
- About 1.2 billion x 14 thousand $\approx$ 16,800 $\approx$ 2 1.6x10ˆ 13 distances computed (just for pick-up)



(a) Shape

(b) Raster

Figure 1: NYC Taxi zones file formats

(a) Both

Figure 2: NYC Taxi zones file formats

Cleaning the data: Crime

- Dates needed to be cleaned. (**24:00:00 vs 00:00:00**)
- Meaningful interpretations of other dates could not be made for certain records and these records had to be filtered for example:
  - 1016 $\rightarrow$ 2016.
  - 1026 $\rightarrow$ dropped.
- The most challenging factor here was that we had a lot of missing values for some columns so we needed to setup a schema that accepted this fact.

Cleaning the data: Weather

- Weather data was the most decent.
- We basically just checked that the data was clean.
- The only major issue was to figure out a way to assign weather data to the taxis.

Joining the data

- In the cleaning process we assigned taxi zones for taxi pickup and drop-off locations.

Joining the data

- In the cleaning process we assigned taxi zones for taxi pickup and drop-off locations.
- So we repeated this process but this time instead of assigning zones we assigned a station (JFK, La Guardia, Central Park) by computing the min distance from the pickup locations (long/lat) to the weather station.

Joining the data

- In the cleaning process we assigned taxi zones for taxi pickup and drop-off locations.
- So we repeated this process but this time instead of assigning zones we assigned a station (JFK, La Guardia, Central Park) by computing the min distance from the pickup locations (long/lat) to the weather station.
- We had another problem here because we did not have (long/lat) for the recent data.

Joining the data

- In the cleaning process we assigned taxi zones for taxi pickup and drop-off locations.
- So we repeated this process but this time instead of assigning zones we assigned a station (JFK, La Guardia, Central Park) by computing the min distance from the pickup locations (long/lat) to the weather station.
- We had another problem here because we did not have (long/lat) for the recent data.
- So, we estimated the centroids on the pick-up zones and then computed the min distance to the weather stations.

Joining the data

- In the cleaning process we assigned taxi zones for taxi pickup and drop-off locations.
- So we repeated this process but this time instead of assigning zones we assigned a station (JFK, La Guardia, Central Park) by computing the min distance from the pickup locations (long/lat) to the weather station.
- We had another problem here because we did not have (long/lat) for the recent data.
- So, we estimated the centroids on the pick-up zones and then computed the min distance to the weather stations.
- Finally taxis were joined to crime by using time periods of one hour.

Goodness

Consistent data

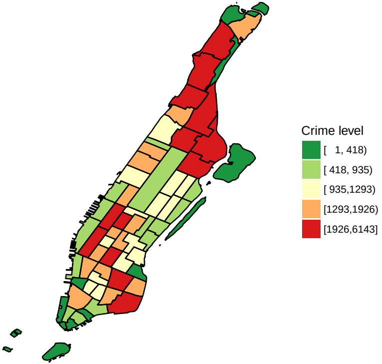
- One of our main concerns was the consistency of the data through time and among the different sources, so we made a lot of effort to keep all variables, even the ones we ended up not using.

Empirical observations not causality

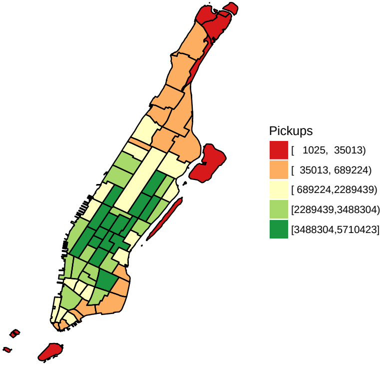
- We are not trying to explain causality so our observations should be interpreted as empirical correlations and raw insight obtained from a very long cleaning data phase

Results

- Opposite colors support our hypothesis.

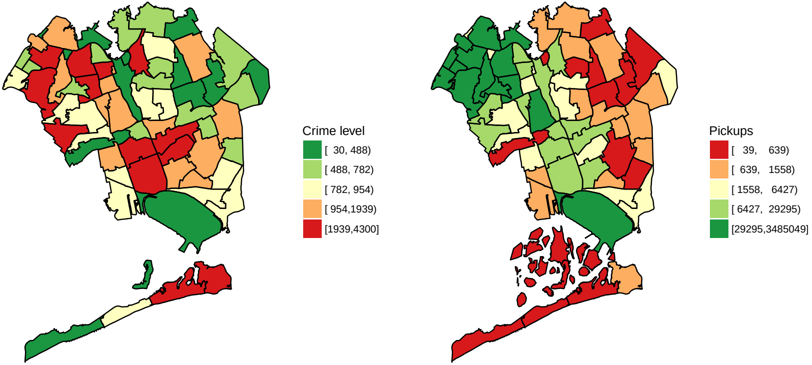


(a) Crimes in Manhattan



(b) Pickups in Manhattan

Figure 3: Crime and Pickups in Manhattan, 2015



(a) Crimes in Queens

(b) Pickups in Queens

Figure 4: Crime and Pickups in Queens, 2015

Crimes and taxis

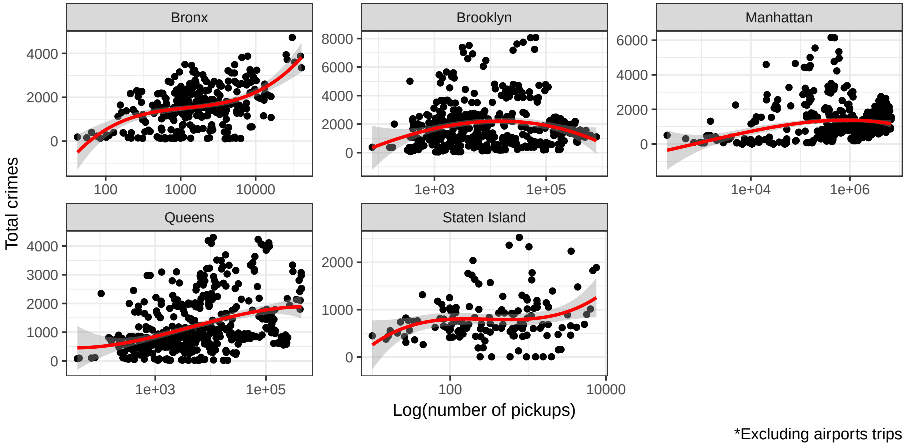


Figure 5: Crimes and pickups per zone

Results when considering Rain

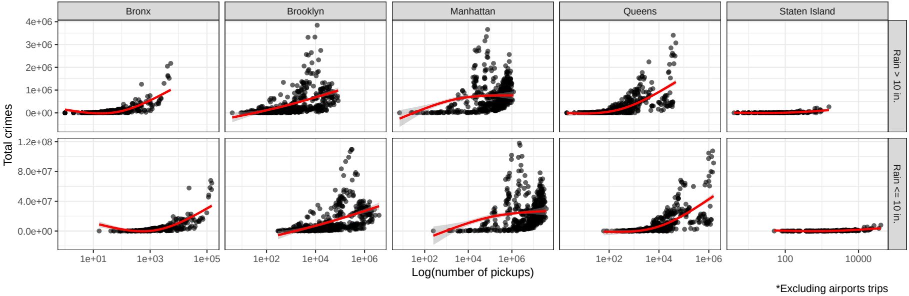


Figure 6: Crimes and hourly pickups per zone in rain

Summary

- We collected NYC taxi trip data, NYC crime data and weather data from Central Park, JFK and LaGuardia and we were able to join everything at a very granular level.
- We found evidence that suggests that our hypothesis might be true, places that have higher levels of crime showed evidence of having a higher number of pickups, especially when taking rain into account.

Acknowledgements

Thank you to everyone at NYU HPC Support for helping us with questions and problems we encountered during this project. Special thanks to Santhosh Konda hpc@nyu.edu for responding so quickly to our e-mails!

References

J. Bendler, T. Brandt, S. Wagner, and D. Neumann.

Investigating crime-to-twitter relationships in urban environments - facilitating a virtual neighborhood watch.

In M. Avital, J. M. Leimeister, and U. Schultze, editors, *ECIS*, 2014.

H. Wang, D. Kifer, C. Graif, and Z. Li.

Crime rate inference with big data.

In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 635–644, New York, NY, USA, 2016. ACM.

M. Traunmueller, G. Quattrone, and L. Capra.

Mining Mobile Phone Data to Investigate Urban Crime Theories at Scale, pages 396–411.

Springer International Publishing, Cham, 2014.

A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland.

Once upon a crime: Towards crime prediction from demographics and mobile data, Sep 2014.

S. Chainey, L. Tompson, and S. Uhlig.

The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime.

Security Journal, 21(1-2):4–28, Feb 2008.

T. Nakaya and K. Yano.

Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics.

Transactions in GIS, 14(3):223–239, 2010.

Thank you!