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 THAT _____.

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

Goodness

What steps were taken to assess the 'goodness' of the analytic?
< Short description of why you believe the results of your analytic are correct and can be trusted

cleaning data

- Some of the taxi and crime data records had incorrectly entered/ambigious date/time columns. We tried to identify the correct dates and times for most of them so that we wouldn't have to throw away those records
- We had to make sure that the three data sets were related by zone

model testing

we made sure that –

Data Sources

Taxi rides data from TLC (*Link*)

- It covers years from 2009 to June 2017.
- Data Size: 2 TB
- 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.

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

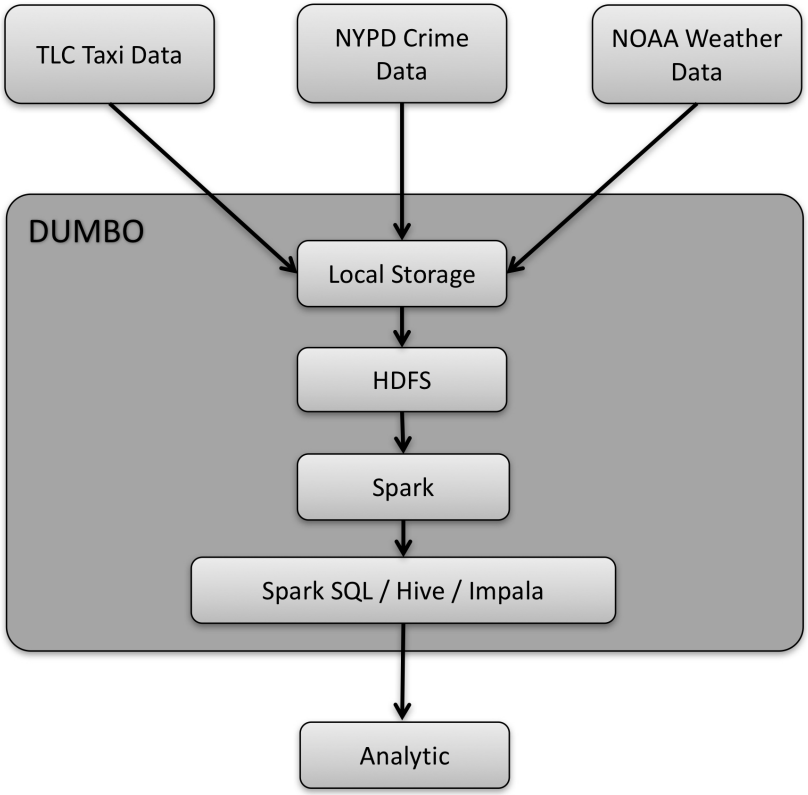
NOAA Weather stations data *(Link)*

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)

Results

3 results, insights, observations, outcomes

1. Result 1
2. Result 2
3. Result 3

Obstacles

1. Cleaning the data- The taxi data was the most challenging to clean due to its size and inconsistencies in columns for specific years. Meaningful interpretations of dates could not be made for certain records and these records had to be filtered.
2. Joining the data- Designated taxi zones had to be determined for taxi pickup locations and crime locations in order to join the crime and taxi data sets and an hourly_date column was added to join taxi and weather data sets.

Summary

- We collected NYC taxi trip data, NYC crime data and weather data from Central Park, JFK and LaGuardia
- We joined together the data sets through taxi zones (taxi_zone_id) and hourly_data
- Our main questions revolve around whether taxi usage depends on crime rate in a specific zone, how much weather plays a part in the relationship between taxi usage and crime and how the distance of the trip affects taxi usage in relationship to crime
- WE FOUND THAT _____

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!