Big Data Analytics Symposium - Fall 2019

Analytics Project:
Gaining Insights from NYC 311 Complaints Data and DOHMH Restaurants Data for Analyzing Crimes

Team Name:
311 Complaints Explorer

Team:
Jiyuan Lu, Iju Lee, Shih-Yao Chou

Abstract:
We help people better understand the city we live in, utilizing three data sources, 311 Complaint Data, NYC Restaurant Data, and NYPD Arrest Data. We made some analysis in order to explore the underlying correlations among the three datasets. The analytic result shows potential indicators of various types of crimes, and it helps further improve the safety and quality of life in NYC.

© 2013-2019 Suzanne McIntosh

Motivation

Who are the users of this analytic?
NYC government and NYC residents

Who will benefit from this analytic?
All the people who live or travel in NYC

Why is this analytic important?

1. It finds potential indicators of different types of crimes, and help the government improve the safety in NYC.
2. It explores the impact of different types of restaurants might have in an area, and supports the city planning or management in NYC.
3. It draws insights from the 311 complaint call data, and further enhances the quality of life in NYC.

© 2013-2019 Suzanne McIntosh

Goodness

What steps were taken to assess the 'goodness' of the analytic?

1. The analytic shows important correlations between restaurants and complaints, and between complaints and crimes, which can be further used to build up complaint or crime prediction models.

2. The insights drawn from the analytic can guide the NYC government to take actions that can decrease the complaint rate and the crime rate to make NYC a better city.

3. The analytic results support the broken window theory and several other findings in previous papers.

© 2013-2019 Suzanne McIntosh

Data Sources

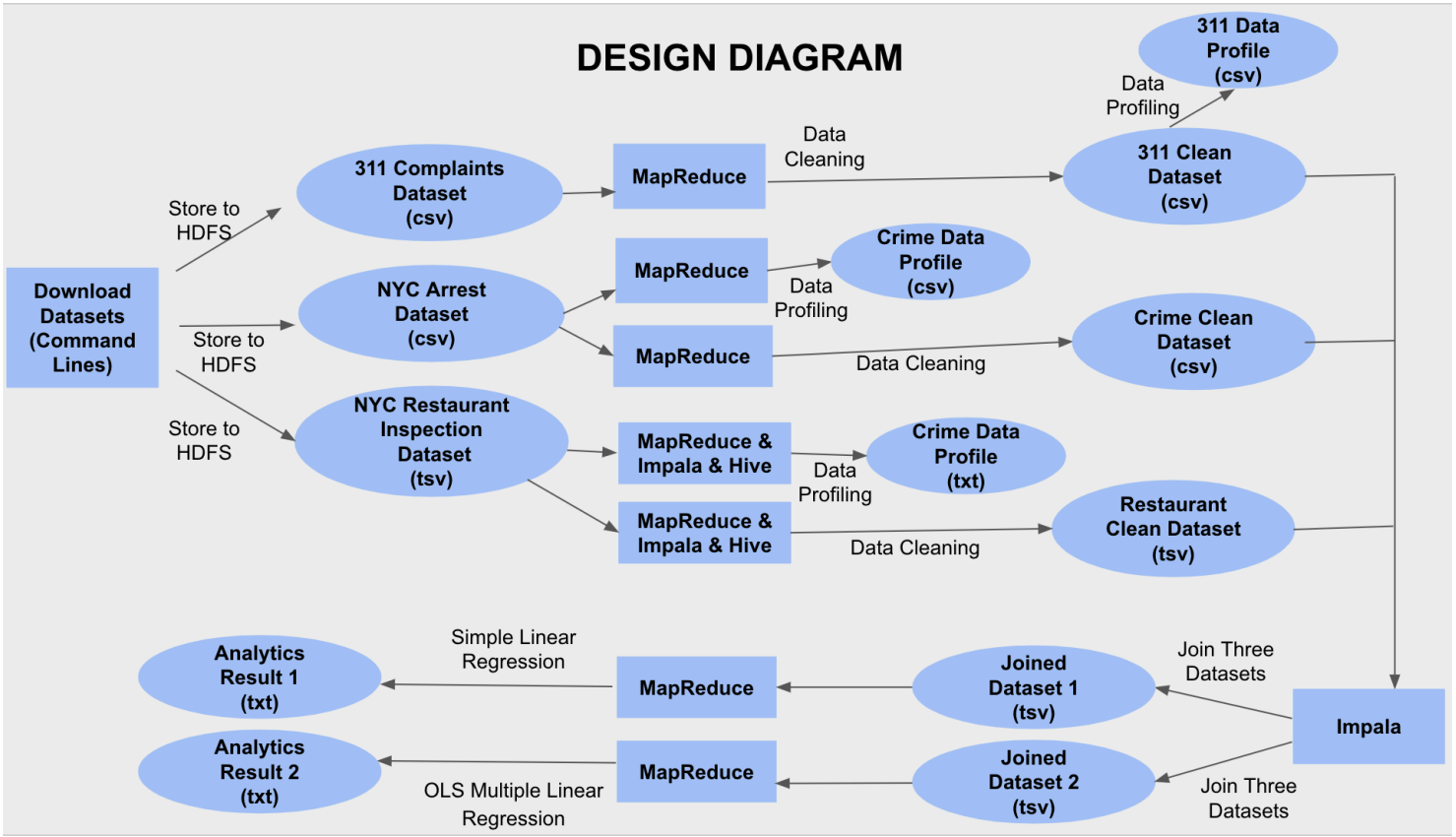
Name: 311 Service Requests from 2010 to Present
Description: 311 is a special telephone number supported in many communities in the US. The number provides access to non-emergency municipal services, and complaints like residential noise or illegal parking are captured in the dataset. This data source collects all the 311 Service Requests in NYC from 2010 to present, and the data is automatically updated daily.
Size of data: 12 GB

Name: NYPD Arrest Data (Historic)
Description: This is a breakdown of every arrest happened in NYC conducted by the NYPD from 2006 through the end of the previous calendar year (2018). This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning. Each record represents an arrest, and variables include the type of crime, the location and time of enforcement, and so forth.
Size of data: 850 MB

Name: New York City Restaurant Inspection Results
Description: This dataset shows the inspection results for every restaurant in New York City. The columns include location (latitude and longitude), name of restaurant, type of restaurant, and et cetera.
Size of data: 600 MB

© 2013-2019 Suzanne McIntosh

Design Diagram



Platform(s) on which the analytic ran:

NYU HPC Dumbo cluster

© 2013-2019 Suzanne McIntosh

Results

1. Model performance for each analytic phase

Phase 1:

Independent Variable (X)	Dependent Variable (Y)	R-squared
Num_of_complaints	Num_of_crimes	0.1779
Num_of_restaurants	Num_of_crimes	0.1074
Num_of_restaurants	Num_of_complaints	0.3206

Table 1: Simple Linear Regression

© 2013-2019 Suzanne McIntosh

Results

1. Model performance for each analytic phase

Phase 2:

Crime Type	Adjusted R-squared	R-squared
Burglary	0.3561	0.3958
Murder	0.3351	0.3762
Robbery	0.3628	0.4021
Weapons	0.3868	0.4246
sexCrimes	0.3350	0.3761

Table 2: OLS multiple linear regression with 10 features

© 2013-2019 Suzanne McIntosh

Results

1. Model performance for each analytic phase

Phase 3:

Crime Type	Adjusted R-sqaured	R-squared
Burglary	0.9408	0.9993
Murder	0.8591	0.9983
Robbery	0.8778	0.9985
Weapons	0.7840	0.9973
sexCrimes	0.9573	0.9995

Table 3: OLS multiple linear regression with 160 features

© 2013-2019 Suzanne McIntosh

Results

2. Top restaurants positively correlated to crime:

Crime Type	Burglary	Murder	Robbery	Weapons	Sex Crimes
Restaurant_ Positive_1	Turkish	Thai	Thai	Asian	Australian
Restaurant_ Positive_2	Hawaiian	Asian	Asian	N/A	Thai
Restaurant_ Positive_3	Australian	Irish	Pakistani	N/A	Turkish
Restaurant_ Positive_4	Thai	Caribbean	Caribbean	N/A	Asian
Restaurant_ Positive_5	Asian	N/A	Irish	N/A	Mediterranean

© 2013-2019 Suzanne McIntosh

Results

2. Top complaints positively correlated to crime:

Crime Type	Burglary	Murder	Robbery	Weapons	Sex Crimes
Complaint_Positive_1	Bike Rack Condition	Bike Rack Condition	City Vehicle Placard Complaint	Investigations and Discipline	City Vehicle Placard Complaint
Complaint_Positive_2	City Vehicle Placard Complaint	Safety	Request Xmas Tree Collection	Indoor Air Quality	Request Xmas Tree Collection
Complaint_Positive_3	Investigations and Discipline	Indoor Air Quality	Indoor Air Quality	Electric	Investigations and Discipline
Complaint_Positive_4	Safety	Mosquitoes	Heating	Sustainability Enforcement	Bike Rack Condition
Complaint_Positive_5	Sweeping Missed	Heating	Investigations and Discipline	Missed Collection	Illegal Tree Damage

© 2013-2019 Suzanne McIntosh

Results

2. Top restaurants negatively correlated to crime:

Crime Type	Burglary	Murder	Robbery	Weapons	Sex Crimes
Restaurant_ Negative_1	Polish	Polish	Polish	Polish	Ethiopian
Restaurant_ Negative_2	Russian	American	French	Middle Eastern	Polish
Restaurant_ Negative_3	Ethiopian	Armenian	Korean	American	American
Restaurant_ Negative_4	French	Middle Eastern	Middle Eastern	N/A	French
Restaurant_ Negative_5	American	Korean	American	N/A	Korean

© 2013-2019 Suzanne McIntosh

Results

2. Top complaints negatively correlated to crime:

Crime Type	Burglary	Murder	Robbery	Weapons	Sex Crimes
Complaint_Negative_1	Window Guard	Illegal Fireworks	Poison Ivy	Illegal Fireworks	Standing Water
Complaint_Negative_2	Day Care	Food Establishment	Illegal Fireworks	Bus Stop Shelter	Poison Ivy
Complaint_Negative_3	Illegal Fireworks	Quality of Life	Bus Stop Shelter	Poison Ivy	Window Guard
Complaint_Negative_4	OEM Disabled Vehicle	Poison Ivy	Food Establishment	Plant	OEM Disabled Vehicle
Complaint_Negative_5	Bus Stop Shelter	Bus Stop Shelter	Quality of Life	Homeless Encampment	Food Establishment

© 2013-2019 Suzanne McIntosh

Results

3. Top restaurants positively correlated to complaints:

Complaint Type	Commercial Noise	Homeless Encampment	Illegal Parking	Rodent	Unsanitary Condition
Restaurant _Positive_1	Eastern European	Afghan	Chinese	Polish	African
Restaurant _Positive_2	Spanish	Mediterranean	Middle Eastern	Caribbean	Polish
Restaurant _Positive_3	Russian	Chinese	Armenian	French	Latin
Restaurant _Positive_4	Polish	Asian	American	American	N/A
Restaurant _Positive_5	N/A	N/A	Eastern European	Spanish	N/A

© 2013-2019 Suzanne McIntosh

Results

3. Top restaurants negatively correlated to complaints:

Complaint Type	Commercial Noise	Homeless Encampment	Illegal Parking	Rodent	Unsanitary Condition
Restaurant _Negative_ 1	Bangladeshi	German	Vietnamese/ Cambodian/ Malaysia	Turkish	German
Restaurant _Negative_ 2	Afghan	Greek	Bangladeshi	Bangladeshi	Portuguese
Restaurant _Negative_ 3	Asian	Bangladeshi	Irish	Mexican	French
Restaurant _Negative_ 4	Italian	Latin	Mexican	N/A	Turkish
Restaurant _Negative_ 5	Korean	N/A	N/A	N/A	Asian

© 2013-2019 Suzanne McIntosh

Obstacles

1. Convert the longitude and latitude data into corresponding zip code

2. Skewed Data Distribution

© 2013-2019 Suzanne McIntosh

Convert the longitude and latitude data into corresponding zip code

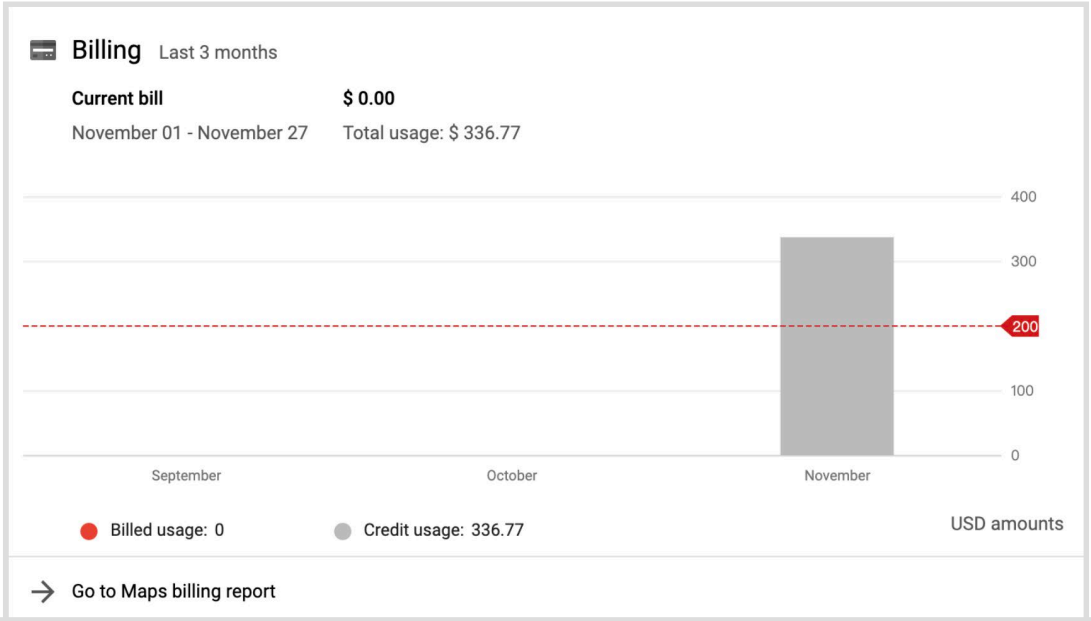
Problem: We attempted calling Google Geocoding API for converting geographical data into zip code.

- Geocoding API can only convert one set of longitude and latitude at a time
- MapReduce timed out since calling API one by one is extremely slow.
- We might be charged a lot of fee since Geocoding API offers limited free usage.

Solution: Using Haversine formula to approximate longitude and latitude into zip code

© 2013-2019 Suzanne McIntosh

Convert the longitude and latitude data into corresponding zip code



Project	Project ID	Cost	One time credits	Discounts	Subtotal
My Project 1	theta-yen-259216	$336.77	-$136.77	-$200.00	$0.00

Subtotal	$0.00
Tax	–
Filtered total	$0.00

© 2013-2019 Suzanne McIntosh

Convert the longitude and latitude data into corresponding zip code

Haversine formula to find distance between two points on a sphere

The **Haversine** formula calculates the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface. It is important for use in navigation.
The haversine can be expressed in trignometric function as:

$$haversine(\theta) = sin^2\left(\frac{\theta}{2}\right)$$

The haversine of the central angle (which is d/r) is calculated by the following formula:

$$\left(\frac{d}{r}\right) = haversine(\Phi_2 - \Phi_1) + cos(\Phi_1)cos(\Phi_2)haversine(\lambda_2 - \lambda_1)$$

where r is the radius of earth(6371 km), d is the distance between two points, ϕ_1, ϕ_2 is latitude of the two points and λ_1, λ_2 is longitude of the two points respectively.

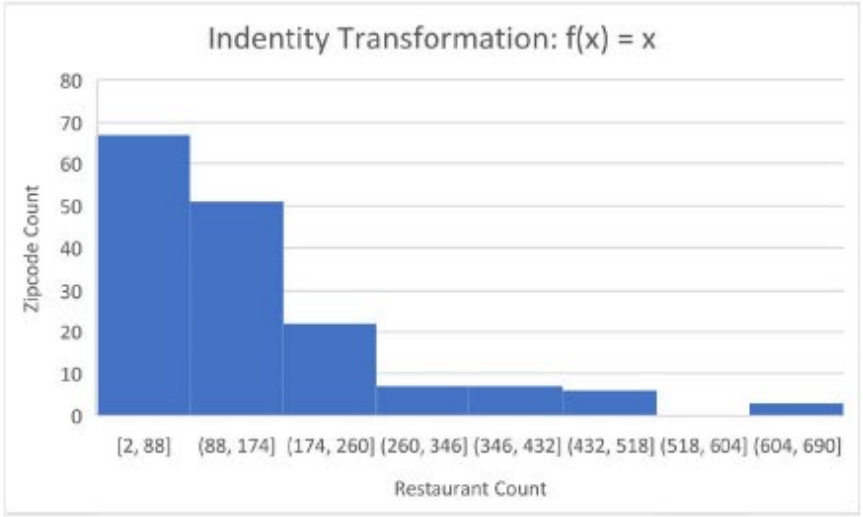
Reference: https://www.geeksforgeeks.org/haversine-formula-to-find-distance-between-two-points-on-a-sphere/

© 2013-2019 Suzanne McIntosh

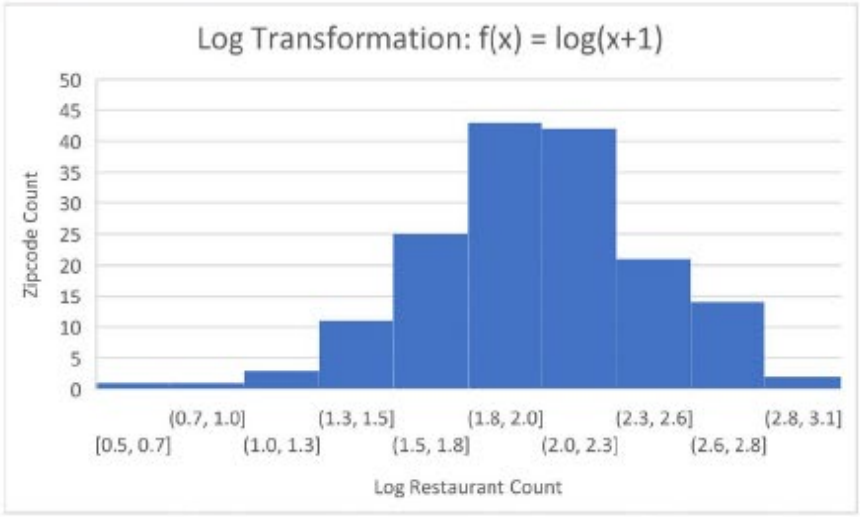
Gaining Insights from NYC 311 Complaints Data and DOHMH Restaurants Data for Analyzing Crimes

Skewed Data Distribution

Restaurant Count:



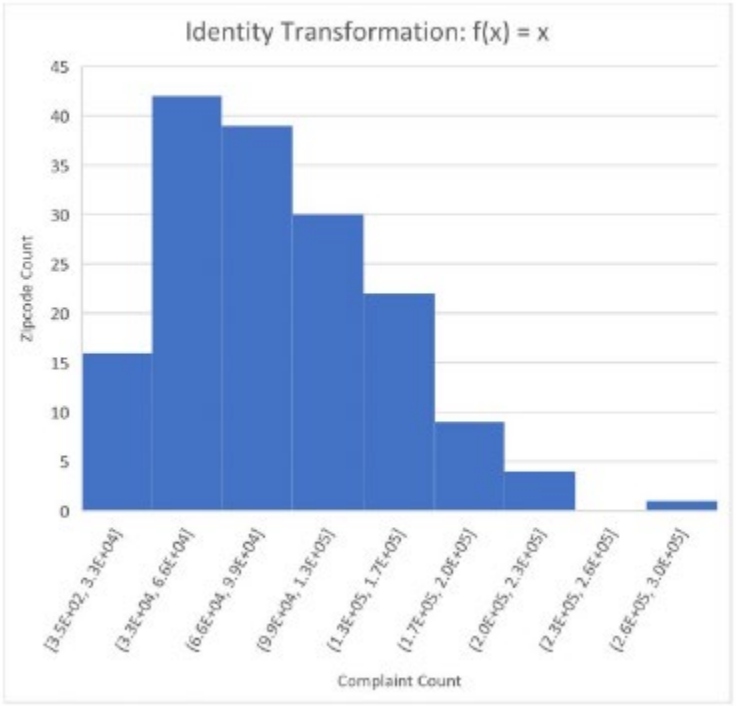
Log Restaurant Count:



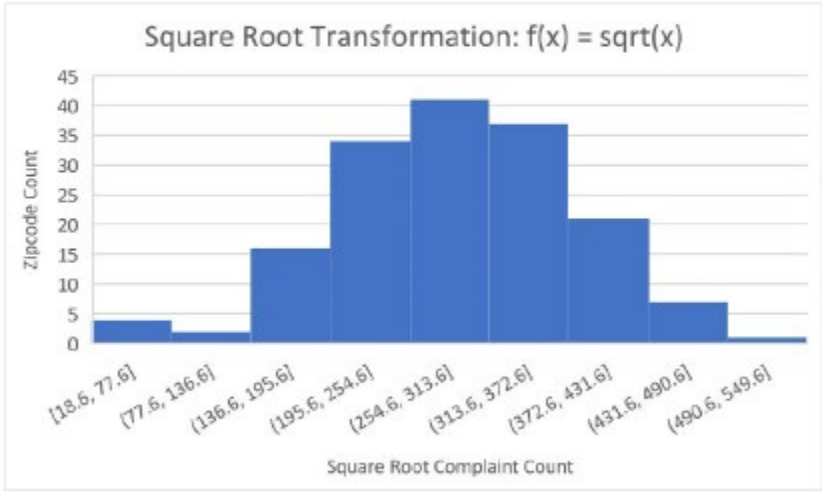
© 2013-2019 Suzanne McIntosh

Skewed Data Distribution

Complaint Count:

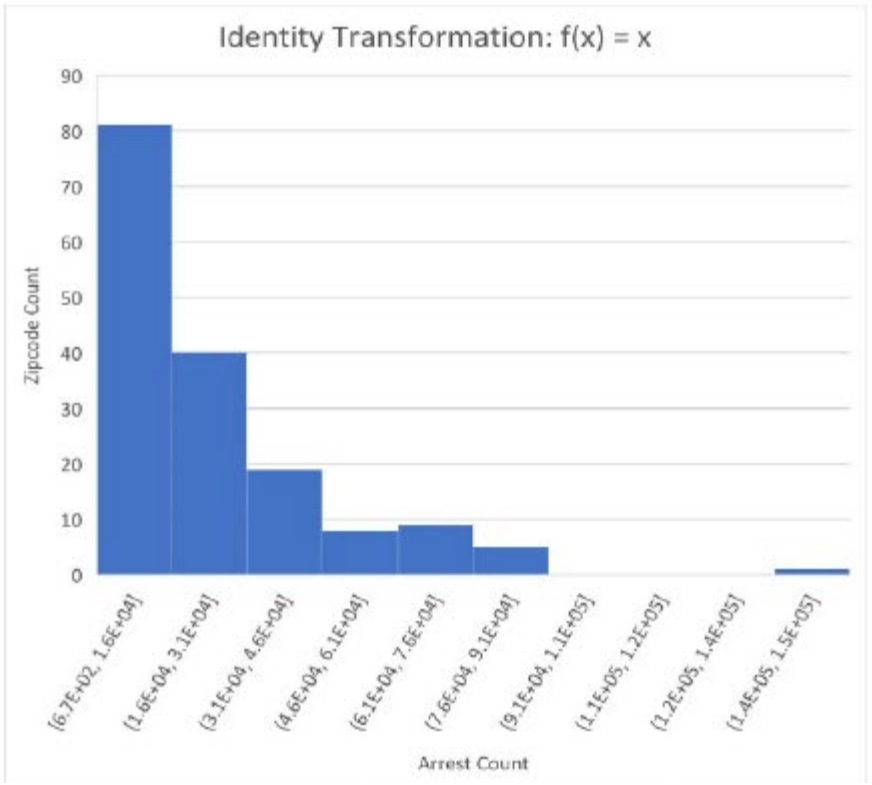


Square Root Complaint Count:

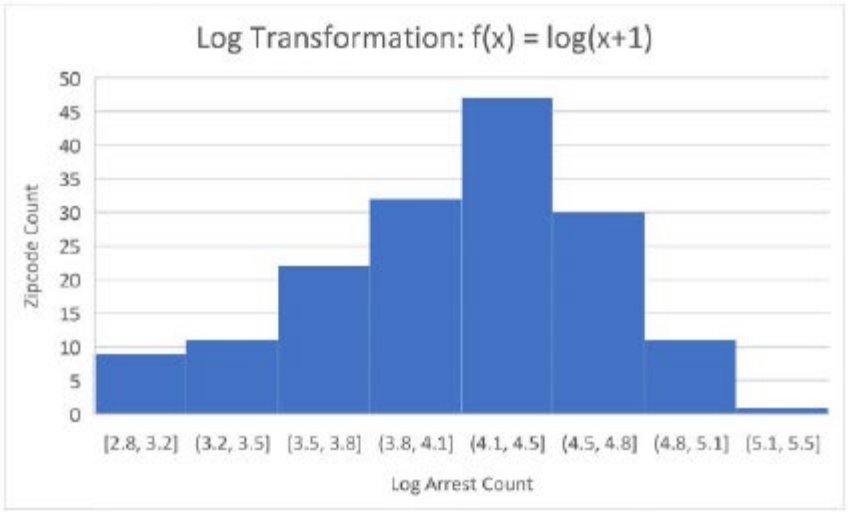


© 2013-2019 Suzanne McIntosh

Skewed Data Distribution



Log Arrest Count:



© 2013-2019 Suzanne McIntosh

Skewed Data Distribution

Crime Type	Adjusted R-squared using Raw Data	Adjusted R-squared using Transformed Data
Burglary	0.8238	0.9408
Murder	0.8460	0.8591
Robbery	0.7978	0.8778
Weapons	0.7998	0.7840
sexCrimes	0.8612	0.9573

Table 7: Comparison between raw data and transformed data

© 2013-2019 Suzanne McIntosh

Summary

- The restaurant cuisine counts are not strongly correlated with the crime counts.
- The complaint counts have strong correlations with the crime counts.
- Our linear regression models are best when combining the restaurant cuisine counts and 311 complaint type counts as features to model the number of arrest counts.
- Our model also does a good job when modeling restaurant cuisine counts as features and complaint counts as output variables.
- Our experiment results strongly support the findings in the previous related researches suggesting significant correlations among restaurant data, 311 complaint data, and NYPD arrest data.

Acknowledgements

- Thanks to NYU HPC staffs.
- Thanks to NYC Big Data Platform.

© 2013-2019 Suzanne McIntosh

References

- S. K. Dash, I. Safro, R. S. Srinivasamurthy. Spatio-temporal Prediction of Crimes using Network Analytic approach. 2018 IEEE International Conference on Big Data.
- R. Rosenfeld, R. Fornango, A. F. Rengifo. The Impact of Order-maintenance Policing on New York City Homicide and Robbery.
- L. Hagen, H. S. Yi, S. Pietri, T. E. Keller. Potential Benefits, and Limitations of Big Data Analytics: A Case Analysis of 311 Data from City of Miami.
- J. Bendler, A. Ratku. Crime Mapping through Geo-Spatial Social Media Activity.
- A. L. Gatens. A Spatial Analysis of Non-Emergency Requests for Service & Violent Crime in St. Louis, Missouri.
- M. Traunmueller, G. Quattrone, L. Capra. Mining Mobile Phone Data to Investigate Urban Crime Theories at Scale.
- A. P. Wheeler. The Effect of 311 Calls for Service on Crime in D.C. at Microplaces.
- T. Law, J. Legewie. Urban Data Science.

© 2013-2019 Suzanne McIntosh

References

- J. Legewie. Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict.
- Y. Wang, Y. Zheng, T. Liu. A noise map of New York city.
- Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, E. Chang. Diagnosing New York City's Noises with Ubiquitous Data.
- B. Chandar and O. Dean. The Effect of a 311 Vacant Building Call on Crime Rate.
- S. L. Minkoff. NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City.
- K. Mulligan, Ph.D., C. Cuevas, B.S., E. Grimsley, B.A., P. Chauhan, Ph.D., & E. Bond, J.D. Justice Data Brief: Understanding New York City's 311 Data.
- A. Bogomolov, B. Lepri, J. Staiano, E. Letouze, N. Oliver, F. Pianesi, A. Pentland. Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics.

© 2013-2019 Suzanne McIntosh

Thank you!

© 2013-2019 Suzanne McIntosh