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


Shih-Yao Chou
NYU Courant CS
Taipei, Taiwan
syc574@nyu.edu

Iju Lee
NYU Courant CS
Taipei, Taiwan
ijl245@nyu.edu

Jiyuan Lu
NYU Courant CS
Nanjing, China
jl11046@nyu.edu



***Abstract*—**

***New York City is a very complex megacity. A lot of accidents and chaos happen in this city every day. To help us to better understand the city we live in, we utilized three data sources, 311 Services Requests Data, New York City Restaurant Data, and NYPD Arrest Data, and made some analysis in order to explore the underlying correlations. Our aim is to provide useful insights for people living, working or studying in New York. Specifically, we built several regression models combining relevant 311 Services Requests Data and New York City Restaurant data, hoping to find the potential indicators of various types of crimes. The coefficients of the linear regression models gave us some hints about what types of complaints and restaurants are positively or negatively correlated to crimes and to what extent they are correlated. By building linear regression models to find the potential relationships among data, we demonstrated a possible way to exploit open source datasets. The key findings of this analysis can be used by various parties to further improve the safety and quality of life in NYC.***



***Keywords—open data, NYC, 311 Services Requests, restaurants, crimes, analytics, regression model, correlation***


## I. INTRODUCTION

311 is a special telephone number supported in many communities in the US. The number provides access to non-emergency municipal services and allows people to report problems such as residential noise or illegal parking. Since everyone in the team lives in NYC, we are particularly interested in collecting and analyzing the 311 Service Requests in NYC from 2010 to present, aiming to understand the city better and to provide useful insights for potential users of our analysis.

To gain more comprehensive insights, we introduced two additional data sources, which would help us to learn the safety and the quality of life in NYC better. The first dataset is NYPD Arrest Data, which is a breakdown of every arrest happening in NYC. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning. Each record represents an arrest in NYC conducted by NYPD and includes information such as the crime type, the location, and the time of enforcement.

New York City Restaurant Inspection Results is the second dataset included in the analysis. This dataset shows the inspection results of every restaurant in NYC. Various information of restaurants are included in this dataset, including location (latitude and longitude), name, and cuisine type, etc.

In order to find the potential indicators of crimes, we built regression models that takes relevant 311 complaints counts and restaurants cuisine types counts as inputs, and see how well these features help model the crime counts as output. This helped us understand what types of complaints and restaurants are highly relevant to crimes.

By carrying out the above analysis, we show a possible way to exploit open source datasets and provide useful findings that can be used by various parties in NYC to improve safety and quality of life.

## II. MOTIVATION

New York is an attractive destination for people all over the world. The economic, political, and social aspects of the city capture everyone's attention. The 311 Service Request Data provides a unique perspective for people to better understand the city. By including two additional data sources, New York City Restaurant Inspections Results data and NYPD Arrest data, our analysis can help people better understand the relationship among complaints, restaurants, and crimes.

The insights from the analysis will benefit all the stakeholders in the area, including government agencies, business entities, individuals, and even visitors.

First of all, safety is important in any city, and knowing the key indicators of crime will help government agencies, such as police departments, to optimize their resource allocation and react to accidents more quickly and more effectively. Regulators can also establish rules or guidance to prevent criminal activities in certain locations. Another main user of the analysis is business entity. Location is the key for business operation, especially for companies in the food or entertainment industries. The rent in a neighborhood with high crime rate may be cheap, but the business has to bear the risks. With the knowledge provided by the analysis, business owners can make insightful decisions.

Moreover, the findings from the analysis will be useful for residents in NYC as well. People can incorporate the insights

in their daily life, for example, where to live or where to spend their spare time in the city. Finally, even visitors can benefit from the analysis. Many accidents happen when people first come to NYC and are not cautious enough when getting into unsafe areas. The insights from the analysis can depict a unique perspective of the city on the zip code level for visitors, which is more detailed and more informative than news or social media.

## III. Related Work

### #1 Spatio-temporal Prediction of Crimes Using Network Analytic Approach

This paper mainly focuses on analyzing crime dataset and demonstrating how to use network analytic techniques to predict criminal activities. More specifically, the authors leveraged these techniques to connect different areas of Chicago and merged them with various data types to predict potential crime in each area.

The reasons that our team choose this paper are as follows:

1. The researchers are trying to investigate whether the crime prediction model accuracy can be improved by including datasets that do not have an obvious relationship with crime rate, which is similar to what our team want to achieve.
2. Simply predicting the overall crime rate in a city does not satisfy the goal of our analysis. What makes the prediction more meaningful is to dive deeper and predict criminal activity on the zip code level. This paper is a great demonstration of this method.
3. Criminal activities can be further divided into different categories, and they may be caused by different reasons. The research methods used in this paper are applicable in our analysis.
4. The researchers used different types of data in their analysis, including police stations, libraries, schools, sanitation, and social complaint calls. Similarly, our team include additional NYPD Arrest Data and NYC Restaurant Inspection Results datasets in the analysis. Our team can learn from the way how the researchers preprocessed and explored different datasets.
5. Furthermore, we can also refer to the way the researchers built and tuned their prediction model. For example, the researchers used polynomial regression, support vector regression, and auto-regressive model to make the prediction. Then, they compared the results and selected the optimal model. The researchers also defined additional features for each community for the prediction models, which might be useful for our team to learn and apply in our own analysis.

### #2. The Impact of Order-Maintenance Policing On New York City Homicide and Robbery

The study estimates the effects of arrest measures on NYC homicide and robbery trends between 1988 and 2001. The arrest measures are based on misdemeanor and ordinance-violation arrests, and the crime model incorporates multiple variables: controls of citizen complaints of disorder, felony arrests, imprisonment rates (a measure of drug involvement), police size, initial crime rates, and levels and changes in a broad range of demographic, social, and economic conditions. The research shows the direct effects of these factors on homicide and robbery trends and the indirect effects of selected features on Order-Maintenance Policing (OMP). It also includes controls for spatial autocorrelation in both the crime and the arrest data.

The authors used a hierarchical linear model (HLM) to assess the within-precinct effects of OMP conditional on other within- and between-precinct differences. The HLM can estimate the intercept as well as the linear trend for the data. More specifically, level 1 estimates individual precinct trends in violent crime rates, and level 2 models the between-precinct heterogeneity in level 1 parameter estimates.

The results shows that OMP has a significant negative effect on on robbery and homicide trends in NYC precincts between 1988 and 2001, which is consistent with the finding that serious crimes in NYC declined because of the strategy of targeting so-called quality-of-life offenses. The study also shows a significant and sizable effect of the prevalence of disorder on robbery and homicide trends in a precinct, represented by the volume of citizen complaints of misdemeanor and ordinance violations. Precincts where the level of disorder decreased the most experienced the greatest declines in robbery and homicide. Again, this supports the hypothesis that disorder and crime are directly related.

The paper also shows the associations between the base rate of each crime and each factor. The base rate of robbery has a negative association with the police per capita, a positive association with the level of immigration, but no significant association with the average level drug market activity or imprisonments per felony. The base rate of homicide has a positive association with the level of immigration, but no significant association with OMP, the police per capita, the average level of drug market activity, or imprisonments per felony arrests.

The paper relates to our project in the following ways:

1. The HLM model proposed in the paper is applicable to our analysis of the NYC 311 Service Request and the two additional datasets.
2. The paper shows a significant and sizable effect of the level of disorder on crime level, which is indicated by the volume of citizen complaints of misdemeanor and ordinance violations, on robbery and homicide trends. This implies that we may find some useful correlations between the 311 complaint data with the NYPD arrest data.
3. The researchers considered many other factors when analyzing the crime base rate and crime trend apart from the citizen complaints. Thus, we should take some additional factors (e.g. poverty level, level of immigration or police per capita) into account when building our prediction model.
4. The researchers included controls for spatial autocorrelation in both the crime and the arrest data, which we should also take care of when using our 311 complaint and NYPD arrest data.

### #3. Potential Benefits, and Limitations of Big Data Analytics: A Case Analysis of 311 Data from City of Miami

This paper focuses on finding patterns within 311 data from Miami and using big data analytics to provide insights to policymakers through exploring living conditions and socio-economic structures. Overall, the purpose of this research is to describe, not to predict.

This paper has several findings. Firstly, phone calls are the major methods for 311 requests. Even though we think smartphones has made significant impacts on our daily life, they only represent a tiny proportion of total service requests. Secondly, the frequency of service requests has been declining. Thirdly, service request types are indicative of socio-demographic factors within areas. For example, areas that have majority of people live under the poverty line tend to have frequent requests on Community Code Enforcement. All of these findings are beneficial to policymakers and can help them make better-informed decisions.

On the other hand, this paper also states that 311 data has several limitations. First of all, the dataset lacks individual-level data, making it impossible to answer why a particular pattern occur entirely. Furthermore, 311 data does not include contextual information like how the data is collected. Lastly, the quality and interoperability is not as good as people expect. The researcher spent about 75-80% of efforts cleaning and merging data.

This paper is significantly relevant to our project. Both of the research and our project share the common purpose — to find underlying patterns form 311 dataset with analytical skills of big data. Though the study is based on Miami dataset, and our team focus on NYC. There are still some methods in the paper that we can learn from and refer to when conducting our own analysis. For example, the research lists some pitfalls of the data. This will help us avoid making common mistakes and therefore improve the quality of our analysis.

### #4. Structure of 311 service requests as a signature of urban location

This paper utilized 311 service requests data to explore socioeconomic features across different cities, including New York City, Boston, and Chicago. It showed there was a consistent relationship between the features and 311 data.

This paper has several findings. Firstly, it showed that, with classifying locations by frequency vectors of types of requests, socioeconomic patterns could be found among the cities. Secondly, models could be built and used to explain certain economic features, such as education level, income, to name a few. Last but not least, house relatively average prices in each zip code could be predicted by 311 data in New York City.

The above findings showed that data similar to 311 service requests could be beneficial to policy-makers to have a beHer understanding of cities.

### #5. Entrepreneurship and crime: The case of new restaurant location decisions

The study presented in the paper explores the impact of various violent crimes such as burglaries, assaults, rapes and murders on restaurant location decisions in a single city, Memphis, Tennessee, from 2009 to 2014, using information on the locations of violent crimes and the locations of newly-opened restaurants.

The crime dataset is based on the Memphis Metropolitan Statistical Area (MSA) and is obtained from the Memphis Police Department's (MPD) Crime Report Map. The restaurant dataset is based on business licenses for new restaurants that were recently granted. From the location information on crimes and restaurant openings the authors created parcels, based on four parcel size choices: dividing the area into 1000 by 1000 grid, 100 by 100 grid, 50 by 50 grid, and 20 by 20 grid, respectively.

In an effort to include some "neighborhood" fixed effects, they constructed "neighborhoods" by grouping each 100 by 100 group of parcels into constructed "neighborhoods" resulting in a set of one hundred neighborhood dummy variables. Including these dummy variables helps averaging out differences in characteristics like population density, income, and education across neighborhoods.

They used both an OLS multiple linear regression model to model all crime variables and a simple linear regression model with each crime variable as the sole regressor to explore the presence of multicollinearity.

The experiment results suggest that new restaurants tend to open in areas with higher local rates of violent crime, regardless of the type of violent crime. This result is supportive of the hot spot effect discussed in criminology literature, in that restaurant districts are attractive targets for both restaurants and criminal activity, as customers and employees move to and from these locations both in large numbers and at night. The observed tendency is also consistent with the conclusion that residential and perhaps other sectors of the economy tend to outbid restaurant entrepreneurs in the market for relatively safe locations in the urban environment. The result also shows that decreasing the number of grids, thus increasing the area of each grid, helps significantly increase the R-squared score of the OLS regression model, while at the expense of the model granularity.

The paper relates to our project in the following ways:

1. We can apply the OLS multiple linear regression model proposed in the paper to model multiple dependent variables and apply the simple linear regression model with individual dependent variable in order to explore the presence of multicollinearity.
2. The paper found a positive relation between the number of crimes and the number of new restaurants in a parcel. This is an indicator that we might find some useful correlations between the NYC restaurant data with the NYPD arrest data.
3. The paper emphasized that neighborhoods effect should be considered for analyzing the relationship between restaurant location choices and past crimes. We can use the similar idea and construct "neighborhoods" by grouping nearby zip code level areas when exploring the relationship between restaurant locations and crime locations.

### #6. Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics

In this paper, it firstly tries to differentiate between two different use cases about big data, one is Personal Data Applications, where individuals' data are analyzed at the individual level to build computational models of each person.

The other one is Aggregate Data Applications, where aggregated and anonymized data of individuals are analyzed collectively so that it can infer large-scale human behavior. In the rest of the research, it only focuses on the second use case and is in the prospect of using aggregate data to influence policy and society positively.

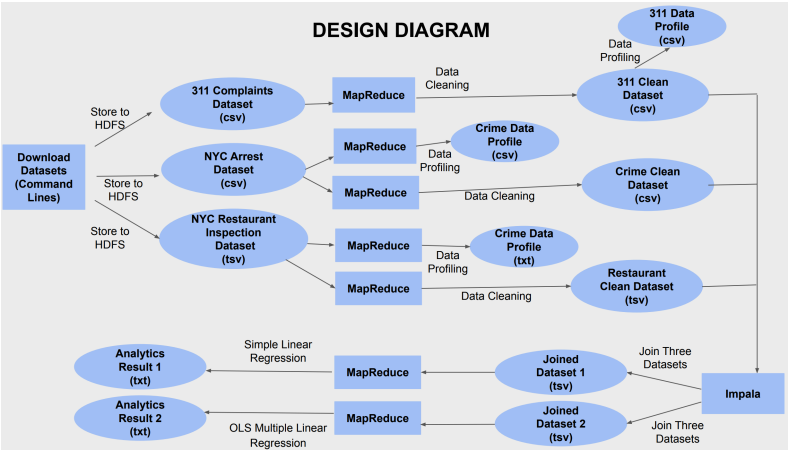
In the article, it also considers that even though crime hasn't been widely covered in the big data field for social good literature, it still can validate the power of place-based crime models built from anonymized and aggregated human behavioral data with limited to no privacy risks.

Consequently, this paper gives a big data method to handle the issue of crime hotspot classification based on past data, of geographic locations that are likely to become scene of a crime. To be more specific, they adopt a place-centric and data-driven approach for investigating the power of people dynamics, derived from a combination of mobile network activity and demographic information, for classifying and determine whether specific locations are more or less likely to become crime hotspots in the near future. Moreover, we also can notice that none of our data sources can be traced back to make inferences about individuals due to the aggregate data applications.

To sum up, this paper offers rich insights related to the big data analysis on crime topic, and is really appropriate to be as a part of our Related Works.

## IV. Design and Implementation

### A. Design Details




The first step of our analytics is to import three relevant data sources to NYU DUMBO, the Hadoop cluster where we run our analysis. The three datasets are: (1) 311 Service Requests from 2010 to Present, (2) DOHMH New York City Restaurant Inspection Results, and (3) Incident Level Arrest Data – 2013 through most recent full year from NYPD. And these large datasets are stored in HDFS.

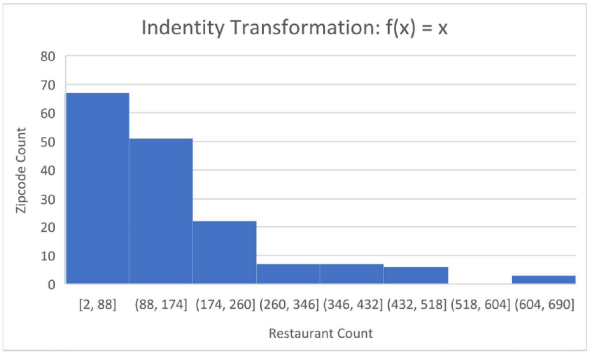
The second step is to profile and clean the datasets. For the 311 dataset, MapReduce tool was leveraged to clean and profile the data. After the process, information such as date, complaint type, zip code and location are kept. Moreover, the input of 311 data profiling is the output of 311 data cleaning. For the DOHMH restaurant dataset, after cleaning, only distinct restaurants are kept, along with their cuisine types and zip-codes. Tools such as Impala and Hive are used in data cleaning and profiling along with MapReduce. For the arrest dataset, we filtered the broken and unnecessary data through using MapReduce. We also produced profile for arrest data through MapReduce. Besides, we transformed the longitude and latitude into zip codes by leveraging a zip code matching table. At last, we produced a clean crime dataset remaining the arrested type, arrested date, location and zip code.

In the data preprocessing step, we met with an obstacle that we had to convert the longitude and latitude data into corresponding zip code. We attempted calling Google Geocoding API for converting geographical data into zip code. However, since Geocoding API can only convert one set of longitude and latitude at a time, thus we must call massive amount of API to convert them. This led to the following issues: 1. MapReduce timed out since calling API one by one is extremely slow. 2. We might be charged a lot of fee since Geocoding API offers limited free usage. Fortunately, we still found a way to convert after a long searching, and we used Haversine formula to approximate longitude and latitude into zip code.

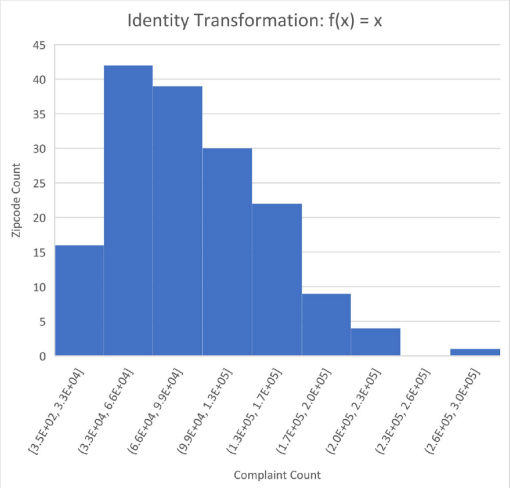
The third step is to build several linear regression models to characterize the data to explore potential correlations between the three datasets. Before this phase, we joined all three clean datasets of 311, arrest and restaurant data into one dataset by leveraging Impala. Throughout our three phases, we use R-squared and adjusted R-squared measure to evaluate the performance of our linear regression models.

In the first iteration, a simple linear regression model is used, in which there is a single independent variable X and a single dependent variable Y. We experimented with three different settings: 1. model the number of restaurants as X, and the number of complaints as Y, 2. Model the number of restaurants as X, and the number of arrests as Y. 3. Model the number of complaints as X, and the number of arrests as Y. These counts are zip-code based. Each count in a zip-code area makes up for an input datapoint for the regression model. In exploring the three datasets, we found that the distributions are quite skewed, especially for the restaurant and the arrest data. Therefore, log transformation is applied to the restaurant and the arrest data, and square root transformation is applied to the 311 complaint data, in order to make the data roughly follow a normal distribution, or more exactly a Poisson distribution for discrete values. These transformations are shown below:

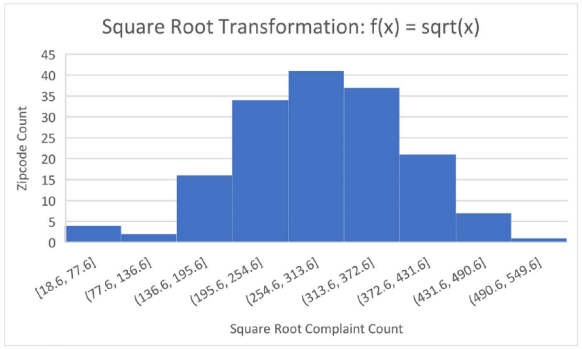
Restaurant Count:



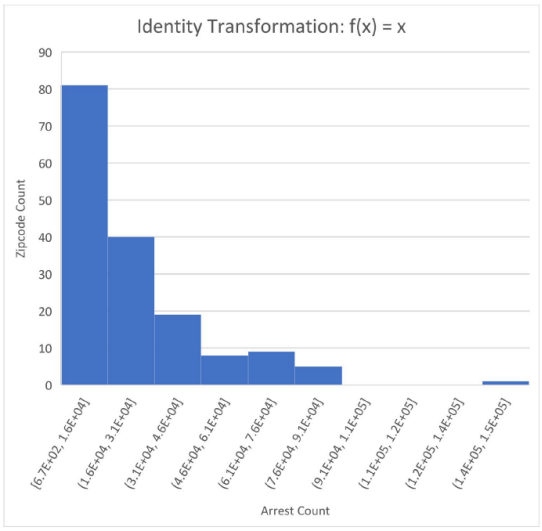
Complaint Count:



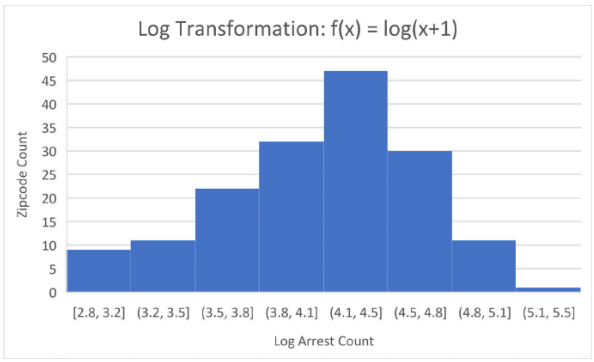

Square Root Complaint Count:



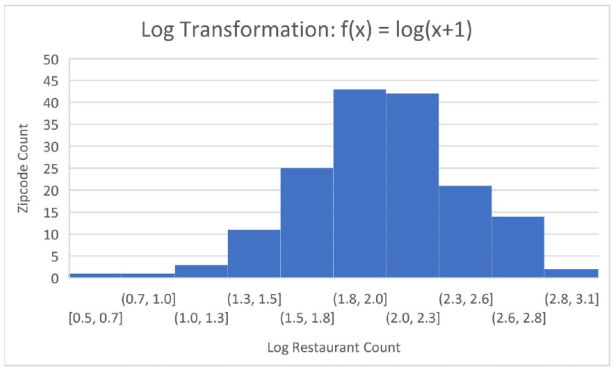

Arrest count:




Log Arrest Count:




Log Restaurant Count:




However, since we are only leveraging a single feature in the simple linear regression model, the performance of the model is far from satisfactory. As shown below, the three models have R-squared measure of 0.1779, 0.1074, 0.3206, respectively, which is an indication of a poor prediction model.

| 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

In the second iteration, an OLS multiple linear regression model is used, in which we have ten independent variables and one dependent variable. Five of the independent variables are chosen from the restaurant data: the number of American, Chinese, Mexican, Italian, and Japanese restaurants in a zip-code area. The other five independent variables are chosen from the 311 complaints data: the number of complaints about noise, homeless person assistance, safety, dead animal, and drug activity. The dependent variable is chosen from the arrest data. We experimented with five different arrest types: robbery, burglary, weapons, sex crimes and murder. All variables are transformed using the same transformation as

mentioned in the first iteration before they are fed to the linear regression model.

Using the OLS multiple linear regression model and taking into account 10 different features, we get reasonable R-squared measures, as shown below. However, it is still very difficult to use these models for prediction.

| 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

In the third iteration, an OLS multiple linear regression model is used with 160 independent variables, among which 38 are different restaurant cuisine counts, and the rest 122 are different complaint category counts. As in the second iteration, we experimented with the same five different arrest types as the dependent variable. All variables are transformed using the same transformation as mentioned in the first iteration before they are fed to the linear regression model.

Using the OLS multiple linear regression model and taking into account 160 different features, we get near-perfect R-squared measures and very promising Adjusted R-squared measures, as shown below.

| 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

In addition, we investigated pairwise correlations between the restaurant data, the complaint data, and the arrest data:

| Crime Type | adjustedRSqaured | RSquared |
|---|---|---|
| Burglary | 0.5023 | 0.8771 |
| Murder | 0.3878 | 0.8488 |
| Robbery | 0.5220 | 0.8820 |
| Weapons | 0.5495 | 0.8888 |
| sexCrimes | 0.4642 | 0.8677 |

Table 4: OLS multiple linear regression modeling complaint data & arrest data

| Crime Type | adjustedRSqaured | RSquared |
|---|---|---|
| Burglary | 0.3334 | 0.4898 |
| Murder | 0.3159 | 0.4763 |
| Robbery | 0.3250 | 0.4833 |
| Weapons | 0.3631 | 0.5125 |
| sexCrimes | 0.2900 | 0.4566 |

Table 5: OLS multiple linear regression modeling restaurant data & arrest data

| Complaint Type | Adjusted R-sqaured | R-squared |
|---|---|---|
| Commercial Noise | 0.7382 | 0.7996 |
| Homeless Encampment | 0.7925 | 0.8412 |
| Illegal Parking | 0.6059 | 0.6984 |
| Rodent | 0.6856 | 0.7593 |
| Unsanitary Condition | 0.7111 | 0.7789 |

Table 6: OLS multiple linear regression modeling restaurant data & complaint data

To see the effect of data transformation on the performance of the linear regression models, compare the table below for the same models using raw data. Notice that the performance of the Burglary, Robbery, and Sex Crimes models decrease significantly; the performance of the Murder model decreases slightly; and the performance of the Murder model increases slightly without data transformation.

| Crime Type | Raw Data | 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

The fourth step is to test and compare different models. In the case of linear regression, the R-squared score is used to evaluate the model. The R-squared score is a statistical measure between 0 and 1 that measures the proportion of the

variance for a dependent variable that is explained by an independent variable or variables in a regression model. In other words, a high R-squared value indicates a good linear regression model and a low R-squared value indicates a poor linear regression model. In the case of multiple linear regression, the R-squared score along with the adjusted R-squared score are used to evaluate the model. The adjusted R-squared score is a modified version of R-squared that has been adjusted for the number of predictors (independent variables, or features) in the model. The adjusted R-squared increases only if the new term improves the model more than would be by chance. The adjusted R-squared is always lower than the R-squared and it can be thought of as imposing a penalty on the number of features in the linear regression model. MapReduce is used to train and evaluate the linear regression models.

The fifth step is to visualize the insights gained from our data. Microsoft Excel tables are used to visualize the relations among the three datasets. Raw / transformed data distributions are also shown using Microsoft Excel histogram plots to help us better understand our data

The sixth step is to compare our results with previous research works to see if our finding matches or contradicts with their conclusions. We compared our findings with the work done in the paper *A Spatial Analysis of Non-Emergency Requests for Service & Violent Crime in St. Louis, Missouri* and our findings strongly support their theory.

## V. DATASETS

### A. 311 Service Requests from 2010 to Present

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. The data size is around 12 GB.

Link: https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

### B. NYPD Arrest Data (Historic)

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. The data size is about 850 MB.

Link: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

### C. New York City Restaurant Inspection Results

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. The data size is around 600 MB.

Link: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

## VI. RESULTS

| 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 | Mediterran ean |
| 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 |
| Complaint_ Positive_1 | Bike Rack Condition | Bike Rack Condition | City Vehicle Placard Complaint | Investigati ons 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 | Investigati ons and Discipline | Indoor Air Quality | Indoor Air Quality | Electric | Investigatio ns and Discipline |
| Complaint_ Positive_4 | Safety | Mosquitoes | Heating | Sustainabil ity Enforceme nt | Bike Rack Condition |
| Complaint_ Positive_5 | Sweeping Missed | Heating | Investigations and Discipline | Missed Collection | Illegal Tree Damage |
| 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 Encampm ent | Food Establishm ent |

Table 8: Top restaurant features and complaint features for analyzing crimes

Result 1:

Table 8 demonstrates for each crime category, the top five restaurant/complaint categories that have positively/negative correlations to the crime (N/A indicates that there are fewer than five relevant features).

For Burglary, the top five restaurant features that are positively correlated are Turkish, Hawaiian, Australian, Thai, and Asian; the top five restaurant features that are negatively correlated are Polish, Russian, Ethiopian, French, and American; the top five complaint features that are positively correlated are Bike Rack Condition, City Vehicle Placard Complaint, Investigations and Discipline, Safety, and Sweeping Missed; the top five complaint features that are negatively correlated are Window Guard, Day Care, Illegal Fireworks, OEM Disabled Vehicle, and Bus Stop Shelter.

For Murder, the four restaurant features that are positively correlated are Thai, Asian, Irish, Caribbean; the top five restaurant features that are negatively correlated are Polish, American, Armenian, Middle Eastern, and Korean; the top five complaint features that are positively correlated are Bike Rack Condition, Safety, Indoor Air Quality, Mosquitoes, and Heating; the top five complaint features that are negatively correlated are Illegal Fireworks, Food Establishment, Quality of Life, Poison Ivy, and Bus Stop Shelter.

For Robbery, the top five restaurant features that are positively correlated are Thai, Asian, Pakistani, Caribbean, and Irish; the top five restaurant features that are negatively correlated are Polish, French, Korean, Middle Eastern, and American; the top five complaint features that are positively correlated are City Vehicle Placard Complaint, Request Christmas Tree, Indoor Air Quality, Heating, and Investigations and Discipline; the top five complaint features that are negatively correlated are Poison Ivy, Illegal Fireworks, Bus Stop Shelter, Food Establishment, and Quality of Life.

For Dangerous Weapons, the only restaurant feature that is positively correlated is Asian; the top three restaurant features that are negatively correlated are Polish, Middle Eastern, and American; the top five complaint features that are positively correlated are Investigations and Discipline, Indoor Air Quality, Electric, Sustainability Enforcement and Missed Collection; the top five complaint features that are negatively correlated are Illegal Fireworks, Bus Stop Shelter, Poison Ivy, Plant and Homeless Encampment.

For Sex Crimes, the top five restaurant features that are positively correlated are Australian, Thai, Turkish, Asian and Mediterranean; the top five restaurant features that are negatively correlated are Ethiopian, Polish, American, French and Korean; the top five complaint features that are positively correlated are City Vehicle Placard Complaint, Request Xmas Tree Collection, Bike Rack Condition, and Illegal Tree Damage; the top five complaint features that are negatively correlated are Standing Water, Poison Ivy, Window Guard, OEM Disabled Vehicle, and Food Establishment.

| 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 |
| 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 |

Table 9: Top restaurant features for analyzing complaints

Result 2:

Table 9 demonstrates for each compliant category, the top five restaurant categories that have positively/negative correlations to the complaint (N/A indicates that there are fewer than five relevant features).

For Commercial Noise, the top four restaurant features that are positively correlated are Eastern European, Spanish, Russian and Polish; the top five restaurant features that are negatively correlated are Bangladeshi, Afghan, Asian, Italian and Korean.

For Homeless Encampment, the top four restaurant features that are positively correlated are Afghan, Mediterranean, Chinese, and Asian; the top four restaurant features that are negatively correlated are German, Greek, Bangladeshi, and Latin.

For Illegal Parking, the top five restaurant features that are positively correlated are Chinese, Middle Eastern, Armenian, American and Eastern European; the top four restaurant features that are negatively correlated are Vietnamese/ Cambodian/Malaysia, Bangladeshi, Irish, and Mexican.

For Rodent, the top five restaurant features that are positively correlated are Polish, Caribbean, French, American, and Spanish; the top three restaurant features that are negatively correlated are Turkish, Bangladeshi, and Mexican.

For Unsanitary Condition, the top three restaurant features that are positively correlated are African, Polish and Latin; the top five restaurant features that are negatively correlated are German, Portuguese, French, Turkish and Asian.

## VII. Future Work

Although we've already gained some insights of the relevancy between the complaints call types and the crime arrested types and the correlation between the complaints call types and the restaurant types, there's still much spaces could exploit on the 311 complaint datasets.

1. We could keep iterating on our model by removing redundant or irrelevant features or try more complicated

architectures such as hierarchical linear regression (HLM) models.

2. In addition to examining relevancy, we could also step further to implement a crime prediction model through using machine learning.

Apart from analyzing the crime and restaurant datasets, the model can also be applied to wider variety of datasets. Specifically, we could check how the complaints are relevant to other statistics. For instance, we can extract the relationship between health data and complaint data to see if the complaints-dense areas might have unhealthy state of residence. Then action such as arranging medical centers in these areas having high complaint calls might be taken.

## VIII. Conclusion

In the paper *A Spatial Analysis of Non-Emergency Requests for Service & Violent Crime in St. Louis, Missouri*, the author concluded that requests for service ("311" calls) are positively and significantly related to violent crimes in the area. And their finding suggests that requests for non-emergency services are an indicator of neighborhood disorder.

Our experiment results strongly support the findings in the above paper, suggesting significant correlations among restaurant data, 311 complaint data, and NYPD arrest data. 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. When restaurant cuisine counts are used solely as features, the performance of the model decreases dramatically. This is an indication that the restaurant cuisine counts might not be directly correlated with the crime counts. When complaint counts are used solely as features, the performance of the model decreases slightly. This shows that the complaint counts have strong correlation with the crime counts. In addition, our model also does a good job when modeling restaurant cuisine counts as features and complaint counts as output variables.

## Acknowledgment

Firstly, we would really appreciate the supports from NYU HPC staffs. They maintained the cluster and made us able to run heavy jobs on it. Secondly, we would like to thank NYC Big Data Platform for sharing the open data. Thirdly, we also wanted to thank Professor McIntosh for giving us advices as well as helping us solve technical problems throughout the research.

## References

1. S. K. Dash, I. Safro, R. S. Srinivasamurthy. Spatio-temporal Prediction of Crimes using Network Analytic approach. 2018 IEEE International Conference on Big Data.
2. R. Rosenfeld, R. Fornango, A. F. Rengifo. The Impact of Order-maintenance Policing on New York City Homicide and Robbery.
3. 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.
4. J. Bendler, A. Ratku. Crime Mapping through Geo-Spatial Social Media Activity.
5. A. L. Gatens. A Spatial Analysis of Non-Emergency Requests for Service & Violent Crime in St. Louis, Missouri.
6. M. Traunmueller, G. Quattrone, L. Capra. Mining Mobile Phone Data to Investigate Urban Crime Theories at Scale.
7. A. P. Wheeler. The Effect of 311 Calls for Service on Crime in D.C. at Microplaces.
8. T. Law, J. Legewie. Urban Data Science.
9. J. Legewie. Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict.
10. Y. Wang, Y. Zheng, T. Liu. A noise map of New York city.
11. Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, E. Chang. Diagnosing New York City's Noises with Ubiquitous Data.
12. B. Chandar and O. Dean. The Effect of a 311 Vacant Building Call on Crime Rate.
13. S. L. Minkoff. NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City.
14. 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.
15. 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.