Predicting crime rates using taxi rides data

Carlos Petricioli
New York University
New York, USA
petricioli@nyu.edu

Valerie Angulo
New York University
New York, USA
vaa238@nyu.edu

Varsha Muralidharan
New York University
New York, USA
vm1370@nyu.edu

ABSTRACT

Understanding and predicting crime is a crucial task in any major city. The objective of this study is to understand crime rates and its effects on how the use of taxis in New York City at a granular level. The idea is that people react according to how secure they feel, which extends to their travel preferences. 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. New York City will be used in this study as it provides a large amount of public data on crimes throughout the five boroughs along with an extensive amount of data from NYC yellow cab usage. We found evidence that supports that our hypothesis holds.

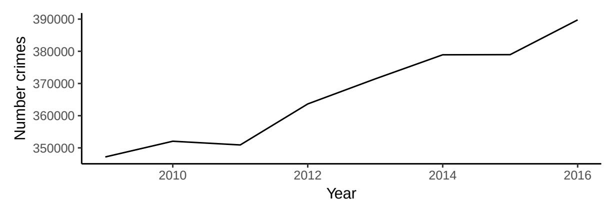
1. MOTIVATION

Understanding and predicting crime is a crucial task in any major city. The ability to understand criminal activity in a very delimited area adds depth to the current understanding of crime and it provides a baseline to understand its effect on people's behavior. This question is relevant because it has a direct impact on the families economies since taking a taxi is the most expensive public transportation option and might also have an impact on people's health because there might be avoiding to walk to avoid crime which directly reduces peoples physical activity and might impact their health in the long run. This fact could translate into a much bigger public health issue for the city in the long run.

This analytic can help law enforcement predict areas of crime based on New Yorkers transportation habits. Those 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. It is hypothesized that if someone feels unsafe in an area, they would be more likely to take a more direct and presumably safer mode of transportation such as a taxi.

By analyzing the patterns of taxi usage in New York City, along with current crime data, law enforcement officials may be able to predict which areas will have a higher rate of crime in the future. This information can also be used as open source data, which would benefit the community and tourists by revealing taxi usage in regards to crime, and perhaps comforting people in using less expensive means of transportation and allowing them to save money in an expensive city.

Figure 1: Total number of reported crimes (2008 - 2017)



Specifically, we picked New York City because it has the largest total number of offenses known to law enforcement by city crime according to the FBI's Uniform Crime Reporting [4], with a total of around 223 thousand for 2016 over cities like Los Angeles, Houston and Chicago, which have 156, 148 and 147 thousand offenses respectively. This provides the most vast crime dataset for this analysis.

Also, New York City has an Open Data Law which mandates that all public crime data be available online [7], which makes crime data collection easily accessible to the public. Figure 1 shows the growth on reported crimes in the New York Police Department (NYPD) data [9].

2. INTRODUCTION

It is hard to argue that violence and crime is not a relevant factor that people take into consideration when commuting among the boroughs of New York city. So, this work assumes that that people want to minimize the time spent in locations they consider unsafe which plays a roll in their transportation preferences affecting their taxi usage.

What goes into an individuals subjective view of what is safe versus unsafe can be hard to gage. However, a lot can be inferred from an individuals behavior patterns. One such pattern of behavior is transportation habits. The idea is to correlate taxi pickup and drop off information to the perceived level of crime activity in an area. It is inferred that people choose to not use public modes of transportation in areas where an individual feels unsafe or uncomfortable and will instead opt to use a more direct and safer source such as a designated taxi.

This study can give a look into what areas people avoid, allowing to geographically match the locations of crime activity to geographical pickup/drop off locations of taxi trips provided in the open source taxi data and can also provide information into the development of areas that become more or less crime ridden.

To perform all the analysis, three main datasets were chosen. The Taxi & Limousine Commission (TLC) provide open data for all the rides in New York City since 2009 up to the present [10]. The crime data collected for this study contains historic [9] and current [8] records from NYPD complaint data, which has crimes reported since 2006 until present. Data collected from the National Oceanic and Atmospheric Administration (NOAA) [6] is used as a third data source in order to provide a more controlled analysis. Specifically, hourly weather data captured from three stations, Central Park, JFK airport, and La Guardia airport stations, is used as a possible explanation for taxi/crime patterns that do not correlate, seeing as weather is often a factor in the decision to take certain forms of transportation.

These three data sets were downloaded into a Hadoop cluster[1], as shown in Figure 2. Both crime and taxi data contain latitude and longitude data, which provide a very granular way of relating the datasets. This was achieved by a spatial join formulated as a Map/Reduce problem in order to exploit the benefits of the cluster. The problem was, how to join the longitude and latitude coordinates for crime activity with the coordinates for taxi usage to a specific area in New York City in order to have a commonly defined location between the two sources. Also, hourly weather data collected at three points, Central Park, LaGuardia and JFK, was assigned to each taxi ride by picking the data collected from the nearest station to the pickup location at a given time.

3. RELATED WORK

There have been some studies on big data sources used in conjunction with crime data in order to understand and predict crime patterns.

In one such study [12], the authors propose to complement the traditional ways of predicting crime rates by including the use points of interest (POI) and taxi flow data in Chicago. In that study, it is hypothesized that taxi flows are “hyper links” within a city that connect locations, where they may be a proxy for broader patterns of population routine activity and mobility, commuting flows, and other forms of social and economic exchanges between two communities over space. The authors use POI to enhance the demographics information and use taxi flow as hyper links to enhance the geographical proximity correlation however, the temporal dimension of crime is not considered in depth. The problem in this study is population-centric, where the crime rate for Chicago is profiled in community areas that are well-defined and stable geographical regions. The proposed POI features and taxi links provide new perspectives in profiling the crime rate across community areas and the crime data collected in Chicago contains detailed information about the time, location and type of crime committed.

Other studies utilize crowd sourced data to predict crime patterns. In a recent study [1]crowd sourcing of tweets was used as a virtual neighborhood watch in order to find crime patterns. The rough location for where the twitter post was sent can be determined by the social network provider or by geo-tags from the users phone. This study inferred that there was a correlation between an important event and the amount of tweets traced to a specific area, where an increase in the amount of tweets in a given area within a certain time span suggests an event was occurring at that time and place. The goal of this study was to predict and explain crimes in urban areas through tweet volume where crime and tweets were related through time and location. This study collected tweets and crime data in hourly blocks at Market Street in San Fransisco during a duration of three months.

Urban crime has been correlated with different modes of communication data as well. The authors of a recent study [11] presented a method to relate crime in London and people dynamics through the utilization of crime data records for the area of Greater London and data from a mobile telecommunication provider for details of people dynamics. Crime data was recorded with latitude/longitude coordinates whereas the telecommunication data was available as footfall in grids of varying sizes (smaller grids in central London as opposed to larger grids in less densely populated areas outside central London). While many people dynamics were looked at in depth in regards to crime, there were two major limitations in the study performed. One was that the crime data was recorded on a monthly basis whereas the telecommunications data recorded footfall on an hourly basis. This limitation is avoided in our study by grouping the data together by hour so that it is more cohesive. Because our study emphasizes time and location of taxi usage, crime activity and surrounding weather conditions, we have made sure that our three data sources share these conditions.

Various methods of mapping are explored in "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime" [3], which is of use to our study. Five different methods have been or are currently being utilized for crime mapping: point mapping, standard deviation spatial ellipses, thematic mapping of administrative units, grid thematic mapping and KDE (kernel density estimation). For this study, crime data was grouped by four types: burglary, street crime, vehicle theft and thefts from vehicles. Areas of high crime concentration (hotspots) in Central/North London were mapped using geocoded crime point data obtained from the Metropolitan Police covering the time period between January 2002 to December 2003. Two dates were chosen to be represented on the hotspot maps, one on Jan 1st as an unusual activity date and one on March 13th as a more ordinary activity date. The time data was sliced into 10 different time periods to avoid getting a map of a time range and perhaps having the map produce a strange result due to unusual activity day patterns. A Prediction Accuracy Index was used where the percentage of crime events for a specific time in a determined crime hotspot was divided by the percentage area of the hotspot compared to the total study area. The hotspot mapping techniques chosen for use in this study along with the methodologies being used were spatial ellipses (STAC: CrimeStat), thematic mapping of boundary areas (MapInfo), grid thematic mapping (MapInfo) and KDE (Hotspot Detective). After mapping the data, it was concluded that KDE is the best of the four methods for predicting spatial patterns of crime due to the accuracy in identifying the location, size, orientation and spatial distribution of the data. KDE uses point data along with two user defined parameters, search width and grid cell size. The street crime hotspot maps were best at predicting future street crime events compared to the other types of crime. This is because street crimes typically occur in areas where there are more shops, bars and other points of interest

[1]NYU's High Perfomance Computing Hadoop cluster, Dumbo

that give opportunity for street crimes to occur.

In the paper [5] the authors propose a spatial epidemiological analysis of crime to reveal uneven distributions of crime risks and spatial interaction between crime events. This article is intended to contribute to extending crime analyses from the spatial perspective to a spatiotemporal one, by proposing an exploratory method to comprehend space-time patterns of crime clusters using space-time statistics and 3D visualization techniques.

A number of studies have highlighted the importance of temporal aspects in crime concentrations is crucial for identifying appropriate crime reduction responses. For example, short-term or cyclic clusters would require a quick strategic action using policing resources, while stable clusters may require long-term efforts to modify social and built environments. However, less attention has been paid to the development of systematic analysis and representational methods of temporal dimensions, as compared to the geographic dimensions of crime epidemiology. Mapping crime at different time periods is probably the most common method to detect temporal changes in the distribution of crime clusters.

In one study [2], the authors proposed a method to predict crime in a geographic area using human behavioral data derived from a combination of mobile network activity and demographic information. Until the above method was presented, most existing research work had been from a people-centric perspective and made use of prior occurrences of crimes to identify patterns of crimes committed by the same offender/group of offenders, etc. A place-centric approach for crime hotspot detection and prediction as presented by the authors complements already existing methods and contributes to criminal studies and data-driven criminal studies. The datasets used are Criminal Cases Dataset (includes geo-location of all reported crimes with month and year tags, specific location of crime and type of crime, Smartsteps Dataset and London Borough Profiles Dataset (demographics). The problem is treated as a classification task to predict if a particular cell will be a crime hotspot in the next month. Since the Smartsteps cell IDs, crime locations and borough profiles are not spatially linked, each crime event was mapped to a Smartsteps cell which it occurred closest to. Features extracted from people who are 'at home' are found to be of high importance. This information can be used by the police departments to determine where to implement higher security.

4. DESIGN

Figure 2 shows our data flow diagram[2]. It shows that all three data sources were downloaded into a Hadoop cluster and cleaned in spark and later stored as Spark's SQL tables which can be read as Hive or Impala tables.

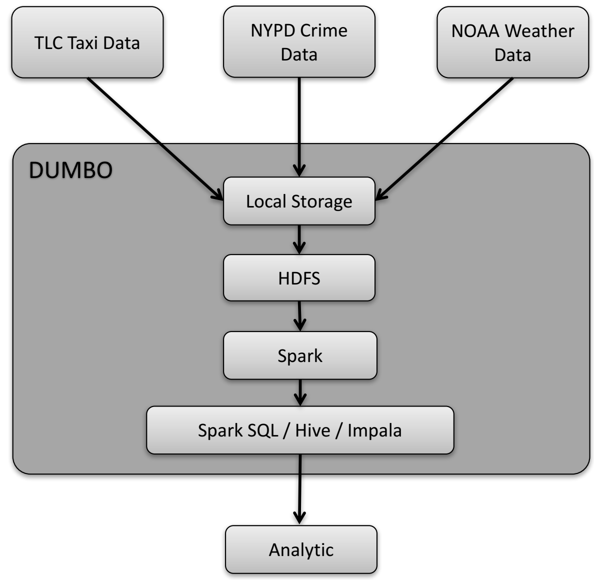
4.1 Getting and cleaning the data

Getting the data was fairly easy because all three sets have APIs which provide a reliable and easy access.

4.1.1 Crimes data

There were a few anomalies that were found in the NYPD crime datasets. The two datasets (historic and current crime) at hand were for crimes that were reported from 2006 till

Figure 2: Data flow diagram



present. However, certain records had incorrect or bad entries in the Crime Start and End Date-Time columns. For instance, there were several crimes reported to have occurred in the year 1016, which were assumed as a typological error for the year 2016. These records were fixed and used in the analytic. A few other records with year entries such as 1026 were dropped, as appropriate inferences could not be drawn. Certain records[3] had timestamps of `24:00:00`, and were corrected to `00:00:00`.

4.1.2 TLC data

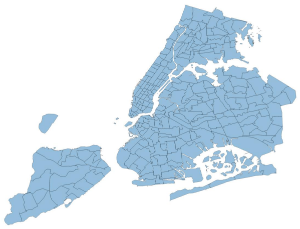
It covers years from 2009 to June 2017 it is about 250 GB. The yellow taxi trip records include:

- pick-up and drop-off dates/times,
- pick-up and drop-off locations,
- trip distance,
- itemized fares,
- rate types,
- payment type,
- passenger counts.

As expected there were complications while cleaning the Taxi rides data from TLC. Actually, the taxi data (250 GB) was the most challenging to clean. Some of the issues include: inconsistencies in columns where we had extra columns for some years rows with extra commas, which made parsing the data a complicated task; row values for each year were not that dirty but the data values were completely different for different years. A big problem was that the dictionary that defines the labels refers to the data from 2017, so we needed to figure out the meaning of labels

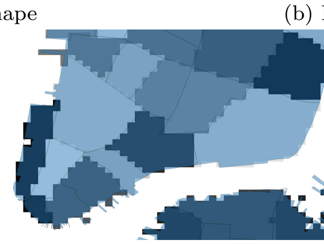
[2] DUMBO refers to NYU's High Performance Computing Hadoop cluster

[3] Which make us wonder how much should we trust this dataset, but, still, it is what we have.



(a) Shape

(b) Raster



(c) Both

Figure 3: NYC Taxi zones file formats

for previous years. 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. For example, longitude and latitude columns referred to places that are not even in NY or were simply null values. There are negative, but consistent values for amounts. For example, a fare of -5, tip of -1 tax of -.5 adding up to -6.5. We decided to made everything positive instead for the cases that the addition was consistent. There are also values for not feasible trip distances, values greater than 1000 miles, for example and exorbitant total amounts, which might not be a problem because most were negotiated fares.
Another big issue was that for the years 2016 and 2017 the data does not include longitude and latitude anymore. Instead they have zone id referring to the taxi zone id for the pickup and dropout. This became one of the greater obstacles, because we needed to assign a zone id to all the previous years.
In order to achieve this we had to perform a spatial join. On one side we have a shapefile for the taxi zones as seen in figure 8a. It is a simple task to perform a spatial join when working in a small dataset, but in order to a spatial join for about 1.2 billion records, we needed to make some modifications. We transformed the shapefile 8a into a raster file 8b and the raster into a csv that has longitude, latitude, and zone id as a fine grid that covers all the area as seen in Figure 3c. This was a good solution because we were able to compute everything in our HDFS cluster with a map/reduce approach.

4.2 Analytic

Crime occurrences will be divided into two levels: *high crime* and *low crime* with respect to the total number of observed number crimes in a given period of time which will be defined later. So crime level can be defined at every time t an location l as a binary variable for *high crime* and *low crime* when the total number of crime at is respectively higher and lower than the average for that time and location.

To simplify things, any location is considered to have a distance to a subway entrance equal to 0 if the location has any number available subway entrances, 1 if any of its neighbors location has a distance of 0 and 2 in every other case. **NOTE:** This assumption will be revised later to mark differences when a location has a lot vs just one entrance because it is important while making a very granular analysis.
Any taxi ride will be categorized as a *short ride* if the distance form the pickup location to drop off location is considered as: could be walked in in less than a given time (i.e. 10 minutes) for an average person. And *long ride* any other case.
With this implementation we can to answer this type of questions:

1. Is the average number of taxi pickups different in areas that have different levels of crime rates?

2. Is the average number of taxi pickups different in areas that have different levels of crime rates grouping by the proximity of an area to a **subway**?

3. Is the average number of taxi pickups different in areas that have different levels of crime rates when we compare at times with and without **rain** (or different weather variables)?

4. Does the average number of **short** rides hve a different average number of taxi pickups in areas that have different levels of crime rates compared to the average of **long** rides?

5. Does the answer of the previous questions change for special dates such as **holidays**?

6. When categorized by the **severity of crime**, is the average number of taxi pickups different in areas that have different levels of crime rates for a given level of severity in crimes?

7. **Run away vs feel attracted** to crime. When there are "high profiled" crime incidents in a given location does the average number of drop offs is greater, lower or similar to the average?[4]

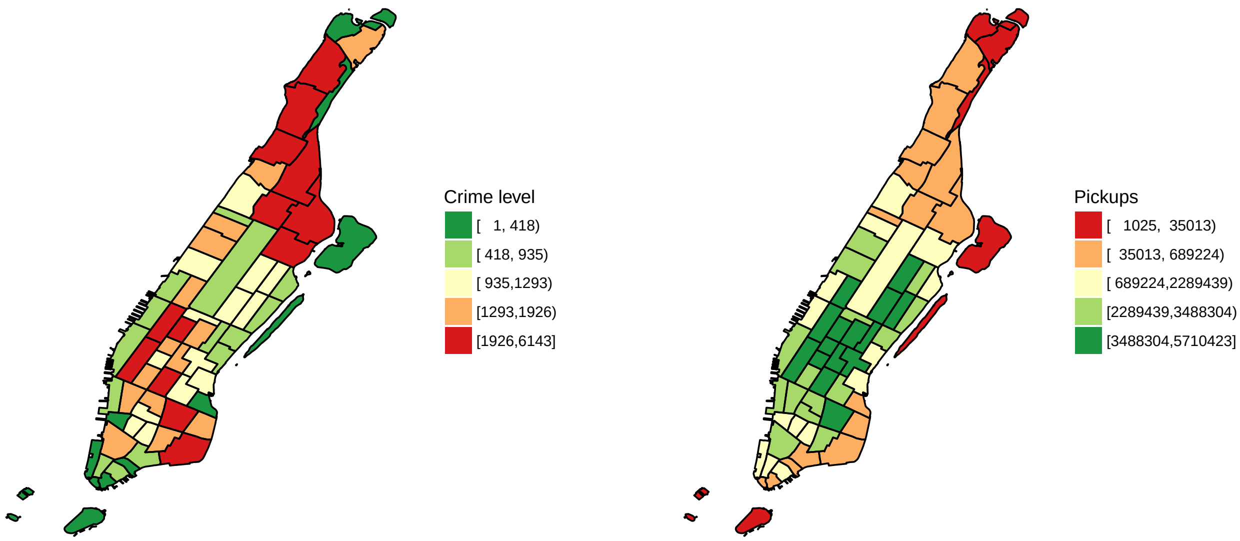
8. Given crime rates changes across time for a given area, is the average number of taxi pickups changing too? If so, same direction? With a time lag? does it lasts long? Does severity has an impact?

9. How does the error in the prediction of crime rates changes when modeling it with traditional demographics vs the taxi rides vs both as features?[5]

5. LIMITATIONS

Our proposed method to quantitatively study crime rates in the city of New York intends to test the appropriateness of employing New York City TLC data and NOAA weather data as a suitable indicator for criminal activity in different New York City neighborhoods. Although the results of our study produce a considerable degree of validity, our

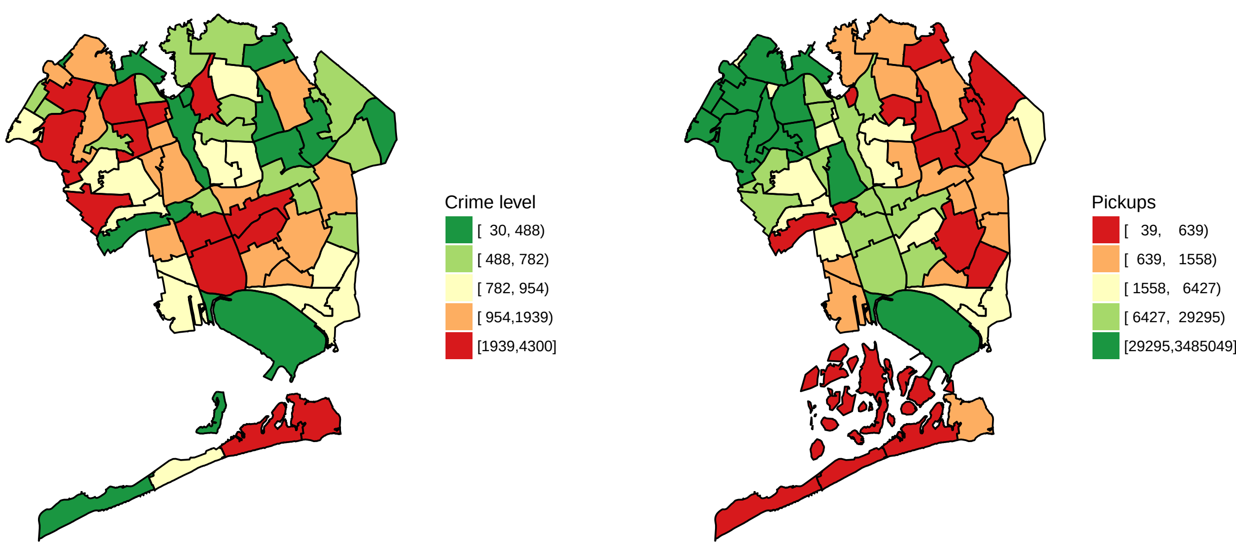
[4]**NOTE:** This is a similar question to [1].
[5]**NOTE:** This is the main question of the paper from KDD 2016 [12].



(a) Crimes in Manhattan

(b) Pickups in Manhattan

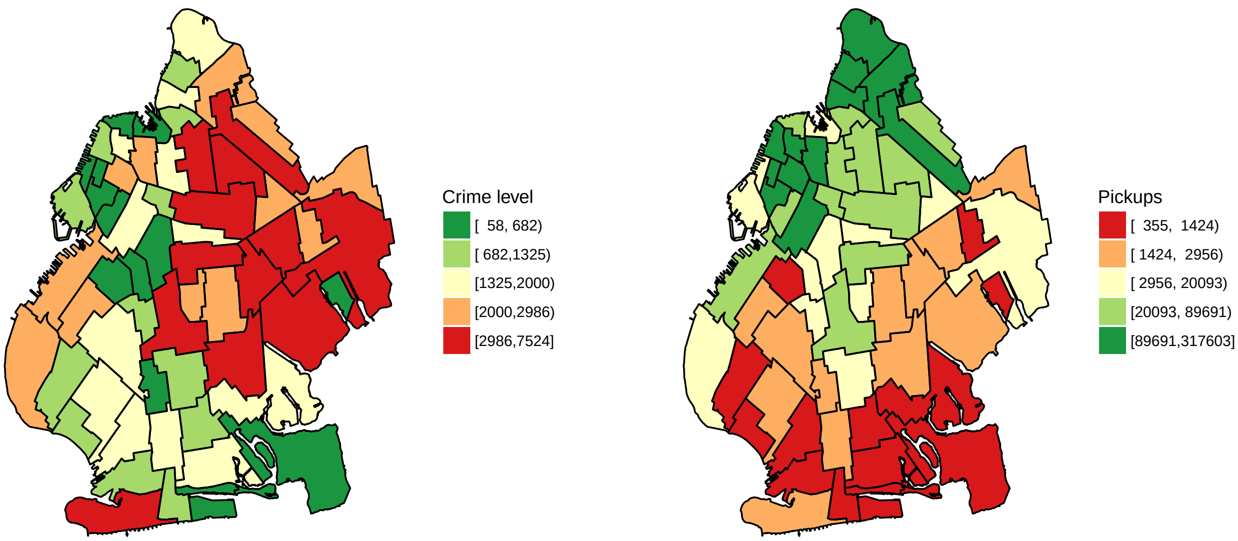
Figure 4: Crime and Pickups in Manhattan, 2015



(a) Crimes in Queens

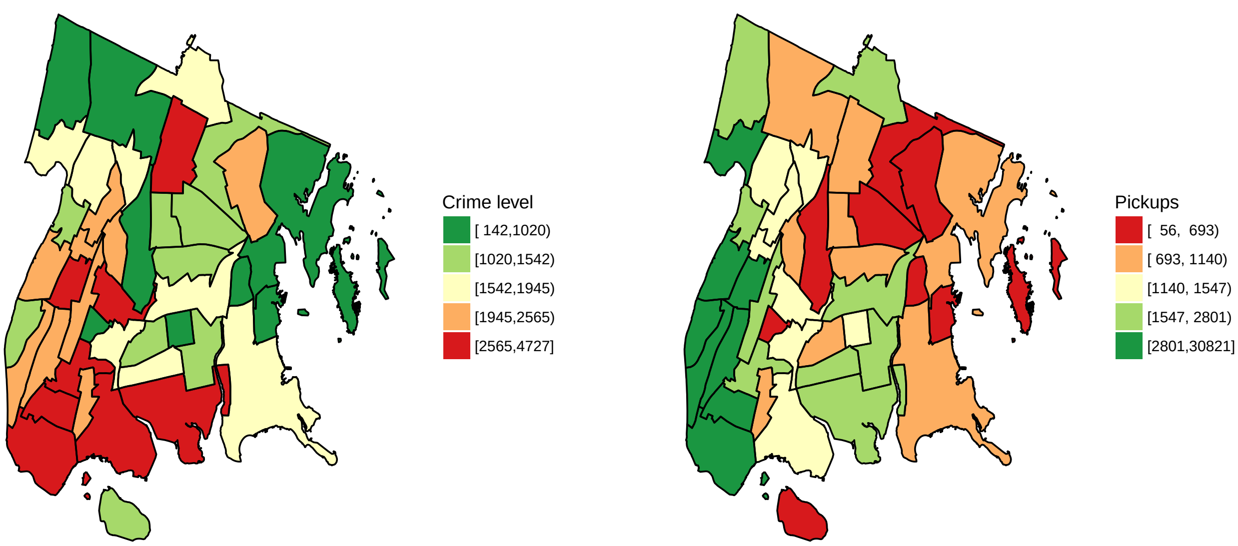
(b) Pickups in Queens

Figure 5: Crime and Pickups in Queens, 2015



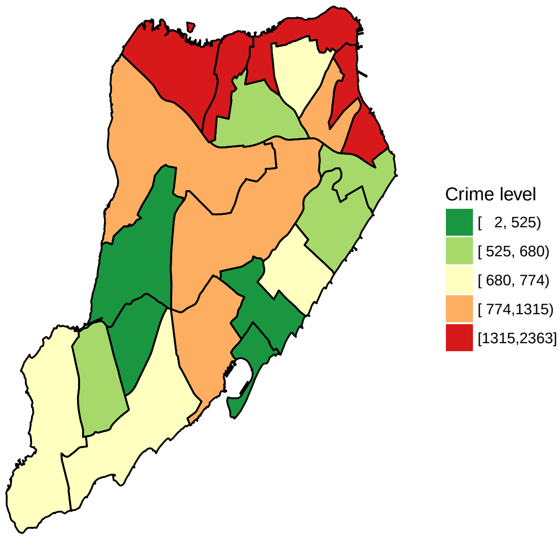
(a) Crimes in Brooklyn (b) Pickups in Brooklyn

Figure 6: Crime and Pickups in Brooklyn, 2015

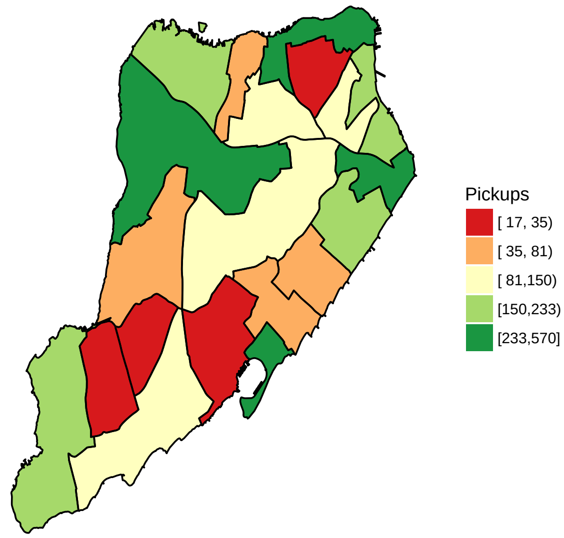


(a) Crimes in Bronx (b) Pickups in Bronx

Figure 7: Crime and Pickups in Bronx, 2015



(a) Crimes in Staten Island



(b) Pickups in Staten Island

Figure 8: Crime and Pickups in Staten Island, 2015

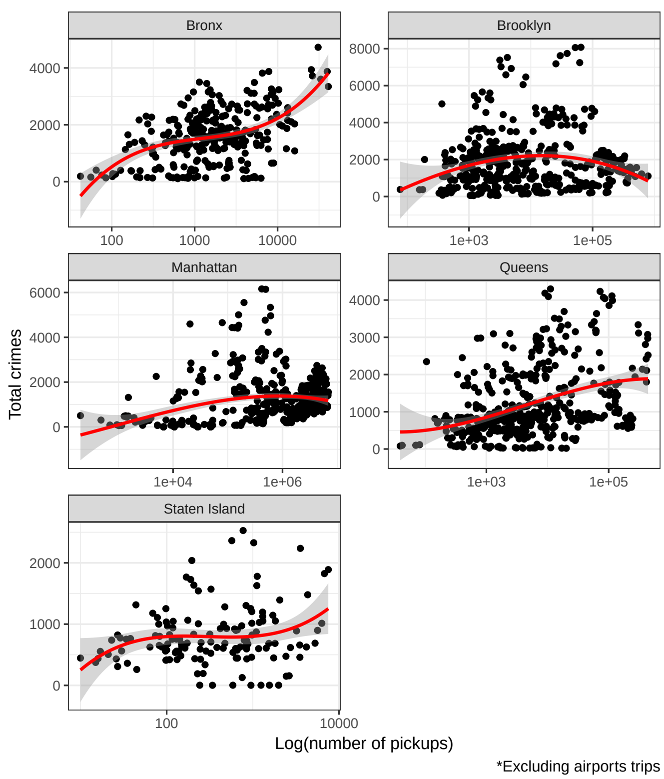


Figure 9: Crime and Pickups in Staten Island, 2015

work suffers from a few limitations. In the age of app-based transportation technology, companies such as Uber and Lyft are challenging the age-old taxi industry. A significant portion of the city's population makes use of these options over New York City taxis because of numerous reasons. For one, Uber's new strategy of making their services more available in the outer boroughs of New York City where taxis are scarce and access to public transport is not easy. Nevertheless, the TLC industry is still strong and serves roughly 50% of city-wide taxi users.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Federal Bureau of Investigation, Uniform Crime Reporting. Offenses Known to Law Enforcement by State by City. 2016.

[5] 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.

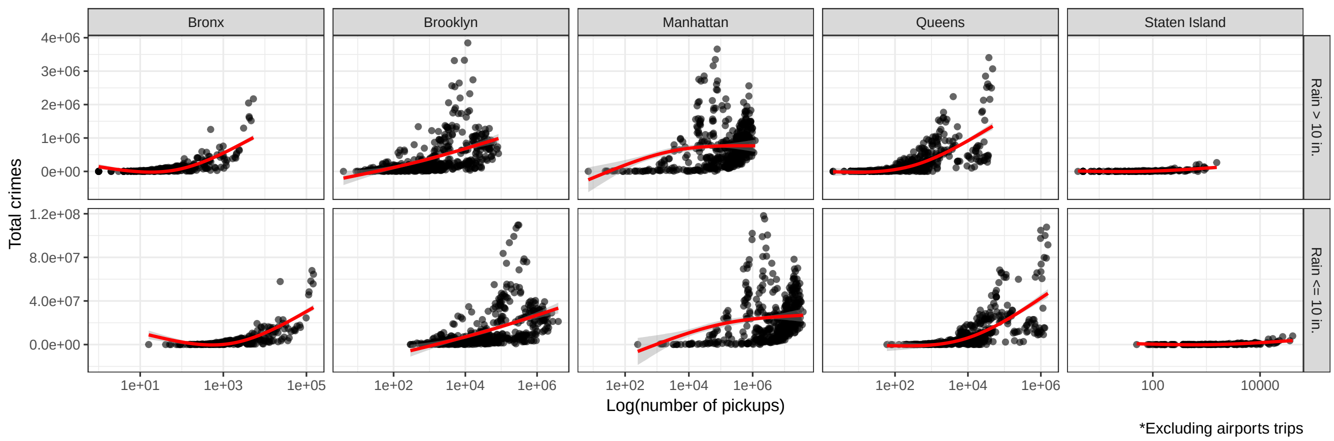


Figure 10: Crime and Pickups in Staten Island, 2015

[6] National Oceanic and Atmospheric Administration - Integrated Surface Database. National weather service.

[7] Open Data City of New York. New York City Open Data Law. 2017.

[8] Open Data City of New York. NYPD Complaint Data Current. 2017.

[9] Open Data City of New York. NYPD Complaint Data Historic. 2017.

[10] Taxi & Limousine Commission. Trip Record Data, NYC. 2017.

[11] 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.

[12] 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.