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 the relation between crime rates and taxi usage in New York City at a granular level. We also want to see how peoples travel behavior is affected by certain weather conditions, such as rain. The idea is that people react according to how secure they feel, which extends to their travel preferences, and that more taxi usage can be expected in rainy weather conditions. 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 also aim to study our hypothesis under the influence of certain weather conditions such as rain. 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 our hypothesis.

1. MOTIVATION

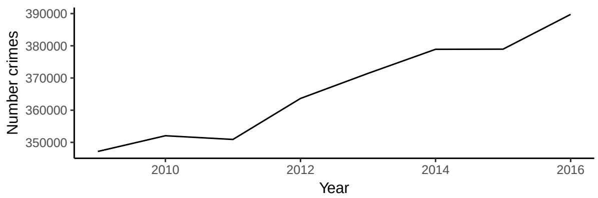
Understanding and predicting crime is a crucial task in any major city. The ability to analyze criminal activity and its effect on people's behavior in a very delimited area adds depth to the current understanding of crime.

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. This awareness can be observed by how comfortable residents feel walking as opposed to taking more immediate and 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.

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 223,000 for 2016 over cities like Los Angeles, Houston and Chicago, which have 156,000, 148,000 and 147,000 offenses respectively. This provides the largest crime dataset for this analysis.

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



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].

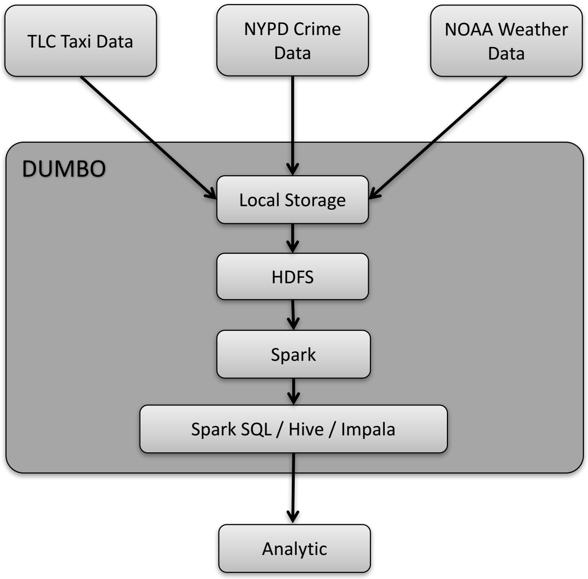
2. INTRODUCTION

It is easy to see that crime is a relevant factor that people take into consideration when commuting among the boroughs of New York City. What goes into an individuals subjective view of what is safe versus unsafe can be hard to gauge. However, a lot can be inferred from an individuals behavior patterns. One such pattern of behavior is transportation habits. It is inferred that people want to minimize the time spent in locations they consider unsafe, which affects taxi usage. Our idea is to correlate taxi pickup and drop off information to the perceived level of criminal activity in an area.

By matching the locations of criminal activity to geographical pickup/drop off locations of taxi trips provided in the open source taxi data, we can identify the areas in which people avoid walking in. Our analysis can also provide information into the development of areas that become more or less crime ridden.

To perform all the analyses, three main datasets are chosen. The Taxi & Limousine Commission (TLC) provide open source data for all the rides in New York City from 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 have crime reports from 2006 until present. Data collected from the National Oceanic and Atmospheric Administration (NOAA) [6] is used as a third

Figure 2: Data flow diagram



data source in order to provide a more controlled analysis. Hourly weather data captured from three stations, Central Park, JFK airport, and La Guardia airport, are used to explain taxi/crime patterns that do not correlate, seeing as weather is often a factor in the decision to take certain modes 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 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. 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 goal was to predict and explain crimes in urban areas through tweet volume where crime and tweets were related through time and location. Tweets and crime data were collect in hourly blocks at Market Street in San Fransisco during a duration of three months. 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. It was 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.

Urban crime has been correlated with different modes of communication data as well. Traunmueller et al. [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 studied by Chainey et al. [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

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

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 that give opportunity for street crimes to occur.

Nakaya et al. [5] propose a spatial epidemiological analysis of crime to reveal uneven distributions of crime risks and spatial interaction between crime events. Their intention was to contribute to extending crime analyses from the spatial perspective to a spatiotemporal one by employing space-time statistics and 3D visualization techniques. A number of studies have highlighted the importance of temporal aspects in crime concentrations, which are 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.

The relationship between human behavior and crime is studied by Bogomolov et al. [2] using 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 is 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

4.1 Getting and cleaning the data

Figure 2 shows our design flow diagram[2]. All three data sources were downloaded into a Hadoop cluster, cleaned in Spark and later stored as Spark's SQL tables.

The first step in the 2 was getting the data. It was fairly easy because all three sets have APIs which provide reliable and easy access. We wrote a script to facilitate this process. It downloads everything from the different sources and uploads everything into our Hadoop cluster.

The next phase was cleaning everything. We used spark for this task and stored everything into partition tables divided with the following format: `year=YYYY/month=MM`. Tables are sparkSQL with schema for easy access in Hive and Impala.

4.1.1 Weather data

The weather dataset was the cleanest out of the three sets and contained an hour-by-hour collection of weather conditions at three station locations in New York City: Central Park, JFK airport and LaGuardia airport. This data set was 165 MB in size.

4.1.2 Crimes data

Both historic and current data sets were 1.5 GB in size. 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 until 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`.

4.1.3 TLC data

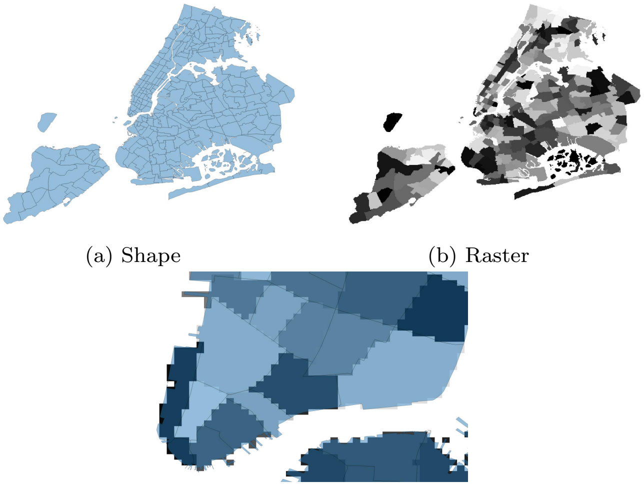
The taxi data covers the years from 2009 to June 2017 and is about 250 GB in size. The yellow taxi trip records include:

- 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. Some of the issues faced include inconsistencies in columns. For example, there were extra columns for some years records that contained more than one comma, which made parsing the data a complicated task. Another problem was that the column headers were not uniform over the years. The available dictionary defining the column headers pertain to the data from 2017, so

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

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



(a) Shape

(b) Raster

(c) Both

Figure 3: NYC Taxi zones file formats

we needed to figure out the exact column headers for the tables of the previous years. After cleaning the categorical variables, several numerical inconsistencies still had to be resolved. For example, longitude and latitude columns referred to places that were not in NY or were null values. The data also contained negative, yet consistent values for amounts. For example, a fare of -5, tip of -1 tax of -.5 still added up to -6.5. To correct this, we made everything positive for the cases where the addition was consistent. There were also values for improbable trip distances. For example, there were values greater than 1000 miles along with exorbitant total price amounts.

Another problem was that the data did not include longitude and latitude for the years 2016 and 2017. Instead they had zone id referring to the taxi zone id for the pickup and dropoff. 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 3(a). It is a simple task to perform a spatial join when working in a small dataset, but in order to do a spatial join for about 1.2 billion records, we needed to make some modifications. We transformed the shapefile 3(a) into a raster file Figure 3(b) and the raster into a csv with longitude, latitude, and zone id as a fine grid that covers the area as seen in Figure 3(c). This was a good solution because we were able to compute everything in our HDFS cluster with a map/reduce approach.

5. ANALYTIC

5.1 Joining data

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

In the cleaning process we assigned taxi zones for taxi pickup and drop-off locations. As described in the previous section. The next logical step was to join with the weather data. Our approach was to assign a weather station to each of the taxi pickup longitude and latitude that corresponds to the closest station in distance and time to maintain accuracy. We repeated the process for assigning taxi zones but this time we assigned a station (JFK, La Guardia, Central Park) by computing the min distance from the pickup locations (long/lat) to the weather station and picking the record of the previous hour to the pickup time.

We had an extra problem assigning weather station to the most recent data in taxis. The taxi records for 2016 and 2017 do not include longitude and latitude just taxi zones. So we estimated the centroids for every taxi zones on the pick-up zones and then computed the min distance to the weather stations and assigned that data. Finally taxis were joined to crime by using time periods of one hour.

6. EXPERIMENTAL RESULTS

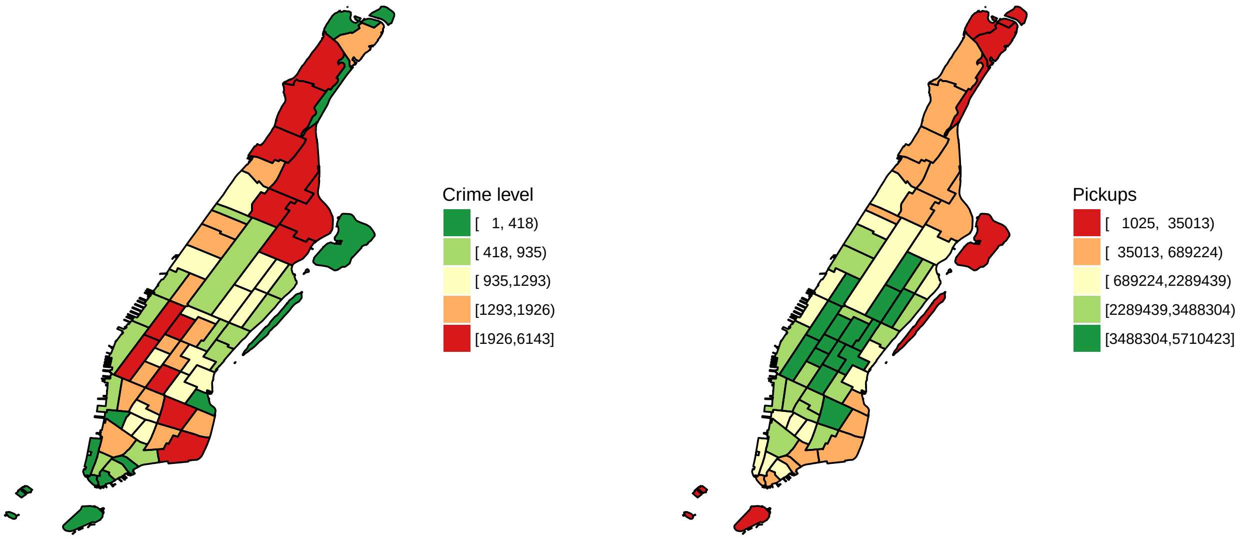
We consider the clean and joined dataset that covers all the years and relates the crime, hourly weather and taxi data as a first result by itself.

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

Our analysis shows that there is a considerable degree of correlation between crime rates and taxi pickups in different zones in the five boroughs of New York City. These can be observed in Figures 4, 5, 6, 7 and 8. In the figures, a high crime level is indicated by a red color and a high taxi pickup rate is indicated by a green color. It can be seen that a high number of red/orange zones on the crime map match to green zones on the pickups map. This suggests that people tend to use cabs more in areas that have higher crime rates. In Figure 4, it is seen that crime is highly prevalent in Midtown Manhattan as well as the northern zones of Manhattan. High pickup is seen in and around midtown. The relatively low value of pickups in the northern zones can be explained by the scarcity of taxis in low movement areas. The other boroughs, as shown in Figures 5, 6, 7 and 8, have a generally positive correlation between taxi pickups and crimes categorized by zone.

Then Figure 9 shows a scatterplot of the total number of crimes against the number of pickups per zone id. It can be seen that there is a positive relation in general except for Brooklyn where it's not clear with that information. If you take a look into Manhattan you can see that there is a small decline in the relation for the places that have higher number of pickups. This was a red flag for us. So, Figure 10 shows each of the boroughs on days that encountered more than 10 inches of rain and days that did not. Each point on the graph represents the total number of pickups in an hour in correspondence to the total number of crimes in the previous year. The first thing to notice is that general there are more pickups when there is rain present. The next thing to notice is that the relationship is more clear when we separate by rain. Another thing to notice is that the same observation in Manhattan is present. There are many observations that have low levels of crime that have a lot of pickups. That was actually a good thing because when you go back to Figure 4 you can spot that this cases correspond mainly to zones that are more likely to be high income.

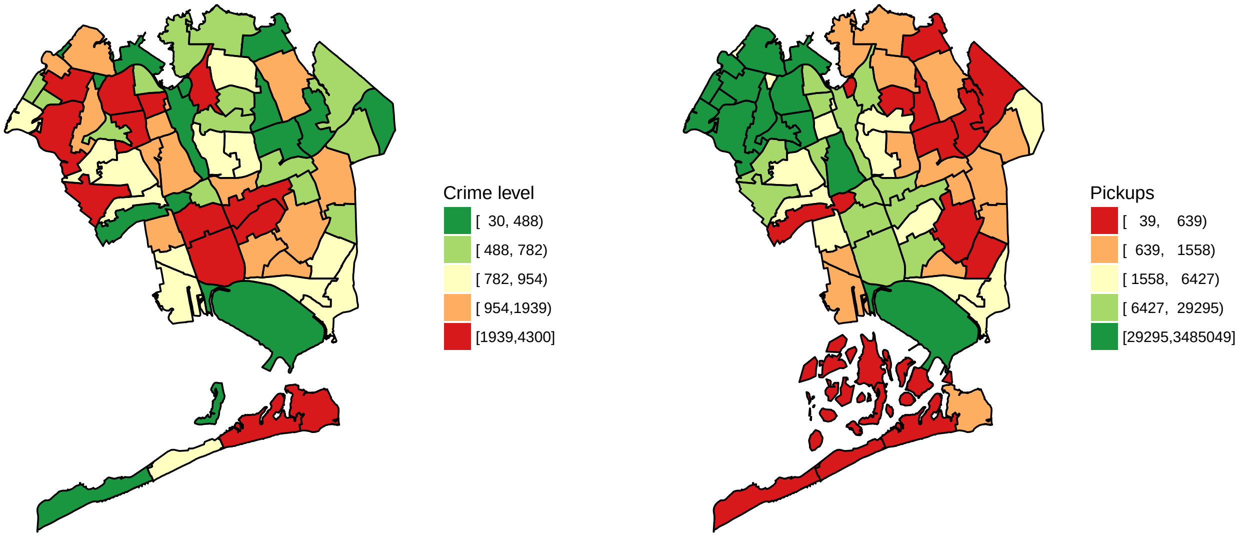
In general we can see that there is a strong relationship between taxi usage and crime, specially if you take into account other variables such as weather.



(a) Crimes in Manhattan

(b) Pickups in Manhattan

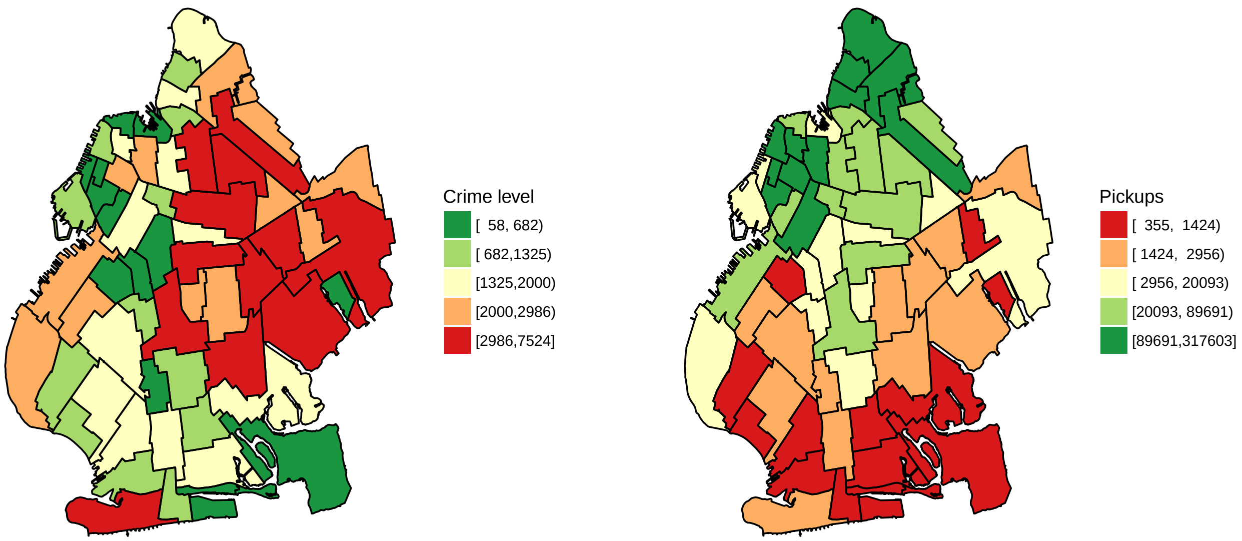
Figure 4: Crime and Pickups in Manhattan, 2015



(a) Crimes in Queens

(b) Pickups in Queens

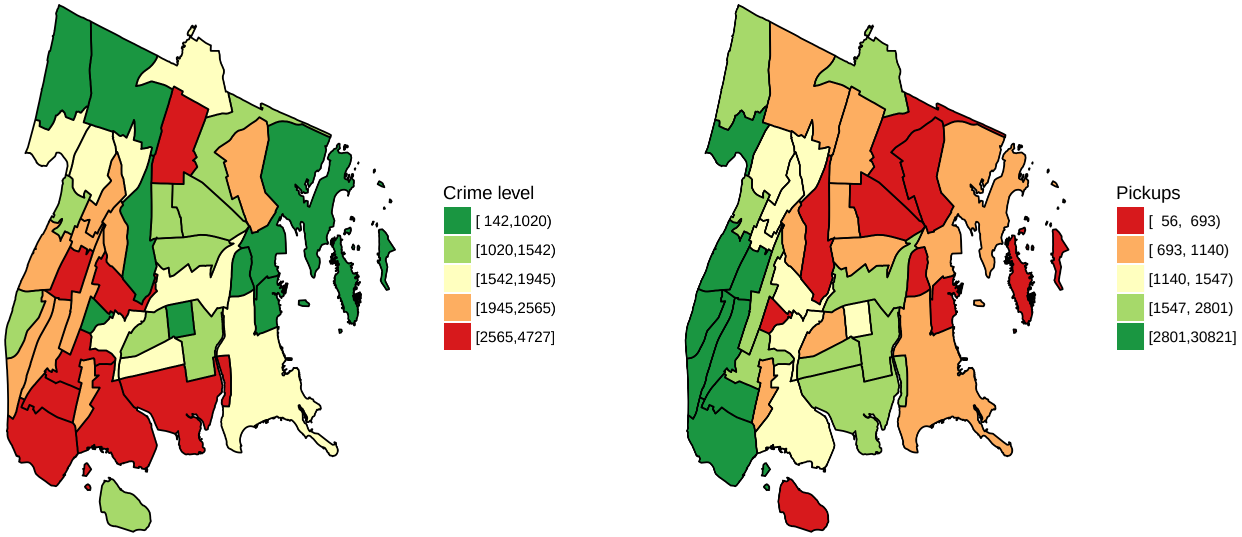
Figure 5: Crime and Pickups in Queens, 2015



(a) Crimes in Brooklyn

(b) Pickups in Brooklyn

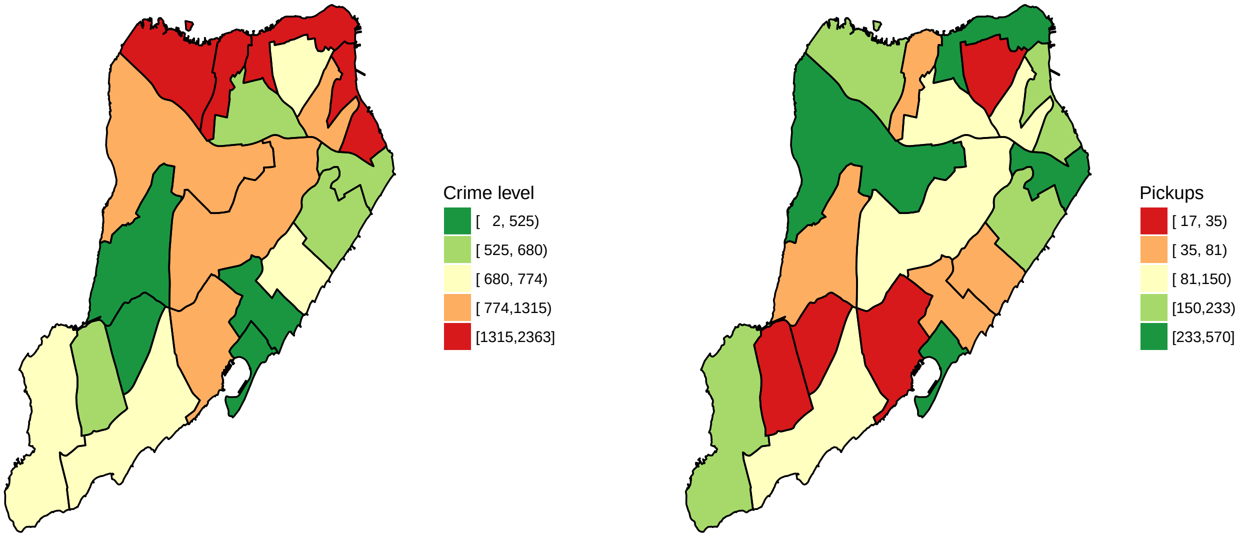
Figure 6: Crime and Pickups in Brooklyn, 2015



(a) Crimes in Bronx

(b) Pickups in Bronx

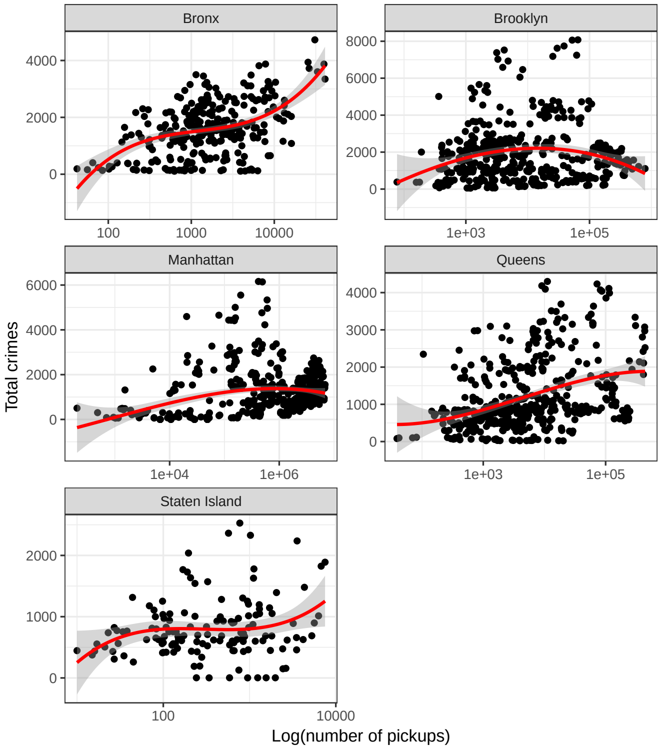
Figure 7: Crime and Pickups in Bronx, 2015



(a) Crimes in Staten Island

(b) Pickups in Staten Island

Figure 8: Crime and Pickups in Staten Island, 2015



*Excluding airports trips

Figure 9: Crimes and Pickups in NYC boroughs without rain conditions

This work was limited by time and we did not have enough for a more formal modeling phase, but we have a very good sense of the data and we can conclude that it makes a lot of sense to explore this in more detail.

7. LIMITATIONS AND FUTURE WORK

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 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 makes 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.

For future work, it would be very beneficial to use Uber and Lyft data sets along with our taxi data set to produce enhanced results. We would also like to bring in a subway station location dataset. It would be interesting to note peoples preference of using a subway versus a taxi in high crime areas. Another question of interest is how high profile crime locations affect people, by using a media dataset. High profile crimes are crimes that reach news sources, therefore reaching a much wider audience than crimes that aren't publicized. This can affect peoples perceptions of dangerous areas, so perhaps there would be a higher observed correlation between crime rates and taxi pickups.

8. CONCLUSION

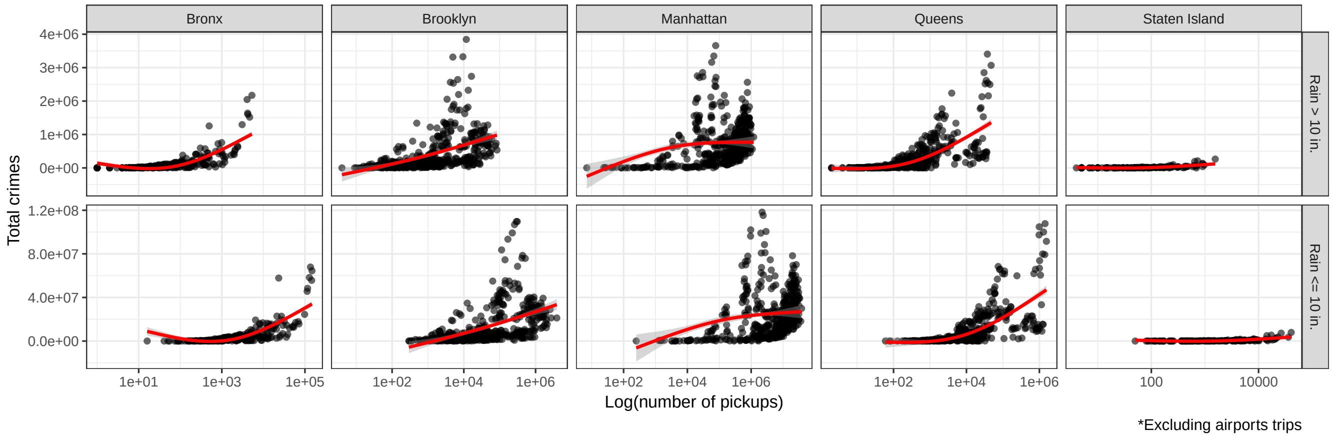


Figure 10: Crimes and Pickups in NYC boroughs with rain conditions

NYC taxi trip data from TLC, NYC crime data and NOAA weather data from stations at Central Park, JFK and LaGuardia were collected and analyzed. The data was loaded into NYUs Hadoop cluster and cleaned and analyzed using Spark. The datasets were joined based on date, time and zone IDs. Areas that had a higher level of crime showed evidence of a higher number of pickups and these results were more pronounced when taking rain conditions into account. Our results suggest that our hypothesis holds true.

9. 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.

[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.