NYU Center for Data Science

DS-GA 1004 - Big Data - Spring 2017

Title: NYC CRIME COMPLAINTS 2006-2015

Authors: Luyu Jin (lj1035)

Siyuan Xiang (sx550)

Daniel Amaranto (da1933)

Abstract

Comprehensive data on criminal reports in New York City from 2006 to 2016 are available to the public as a result of the NYC Open Data Law. Given the size of the database, big data methods are required to manipulate the records into a form that can be analyzed. We took advantage of such capabilities via NYU's cluster computing services and, after summarizing important facts related to local crime, we developed hypotheses about what factors influence crime. Using external datasets we aimed to reassess factors that were determined to be important indicators of crime in the 1980s and 1990s when criminal activity in NYC was far greater. Although historical research showed connections between crime and commercial activity, our analysis did not clearly identify the same relationship. However, other factors that related to crime rates in historical research were again identified to be important in our analysis. Specifically, measurements of the urban environment (housing conditions and evidence of rodent populations) had positive, statistically significant relationships to crime levels. This research thus identifies some factors whose relationship to the crime rate has evolved, and others whose relationships to the crime rate has remained intact.

Introduction

New York City has been fortunate in the last several decades, witnessing a precipitous drop in all types of crime. Despite its size and diversity, it has a crime rate that is lower than the national average.(1) Given the drastic change in the composition of NYC's criminal activity, we believe it is important to determine if historic causes of crime are still relevant, or if they are no longer highly correlated to criminal activity. This question is important because it will help decide whether further reductions in crime will be achieved by continuing existing policing practices, or if brand new solutions are required to further reduce crime. Therefore we analyzed the crime complaints dataset from two different perspectives established by research from the 80s and 90s: first we examined 'crime attractors', places that provide specific opportunities for crime; (2)(3)(4) and second we considered *broken windows theory* and how indicators of urban disorder might be correlated with high crime rates. (5)

A crime attractor is a place that, for some particular reason, motivates criminal

activity. A crime attractor could be a place that is centered around illegality (for example a center of illegal gambling or prostitution) and therefore attracts criminals. But a crime attractor can also be any area that has legal concentrations of cash economies (i.e. businesses) or high numbers of people.(2)(3)(4) We chose to investigate whether legal crime attractors (concentrations of businesses) was related to crime in the data from 2006-2016.

Broken windows theory is well known, and many scholars attribute decreasing crime rates to police crackdowns on petty crime. The theory also applies to environmental indicators of urban disorder or decay. (5) We found public datasets with two types of disorder indicators: one was a record of different types of housing violations. We assumed that areas with lots of housing violations would likely correspond to a 'broken windows' setting. In addition we looked at records of rodent activity and also attempts to clean up rodent infestations. We assumed that large numbers of rodents correspond to areas identified by broken windows theory as being more susceptible to crime. If areas of the city with safer buildings and cleaner streets experience lower crime, then our findings would suggest that historic markers of criminal activity are still important today.

The use of Big Data

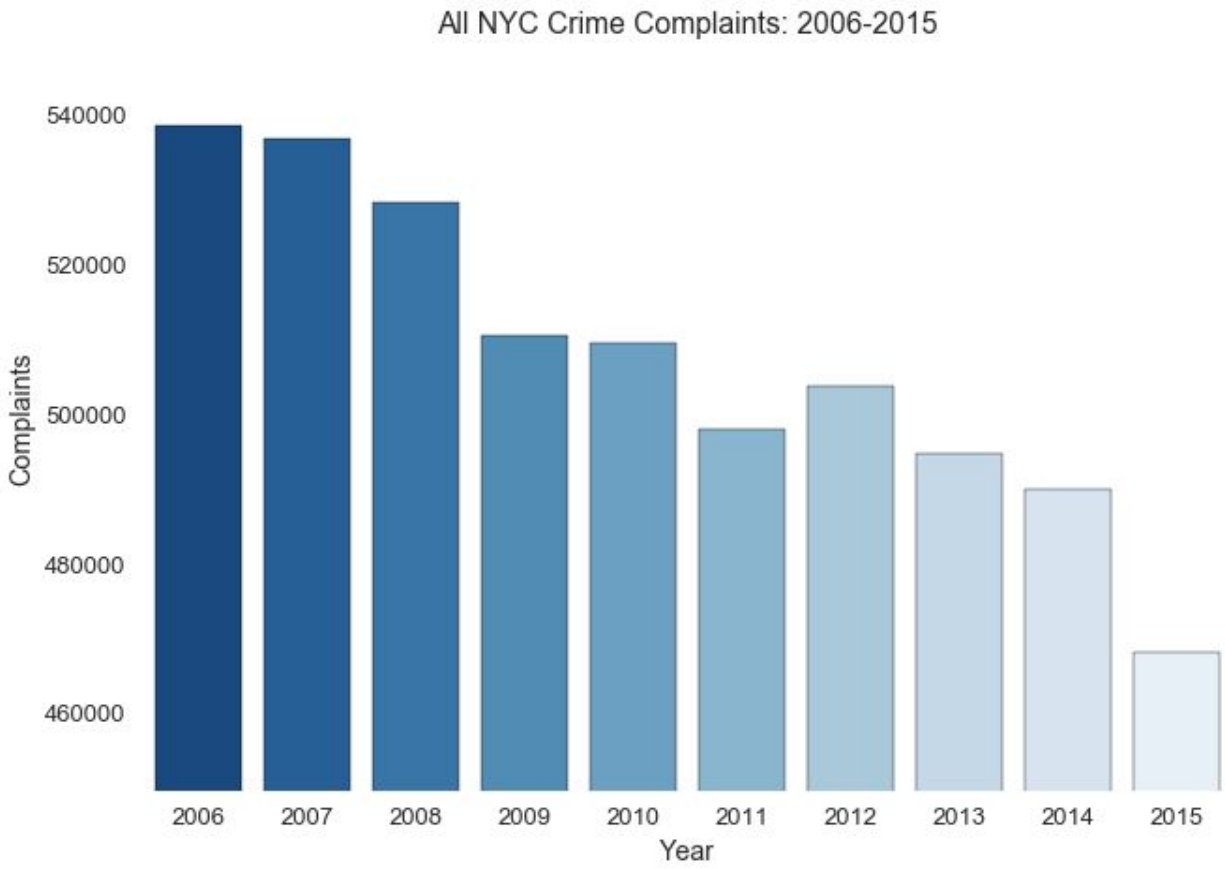
The key to our analysis is in taking advantage of large amounts of records that pertain to all areas of the city. We used gps and zip code data to link information about businesses and urban decay factors to crime rates witnessed in the crime data. While not every dataset we used required distributed file systems for ready analysis, the main target we were looking at (number of crimes) required using PySpark methods in order to distill it into useful information.

Part I

DATA SUMMARY

The *NYPD Complaint Data Historic* dataset…

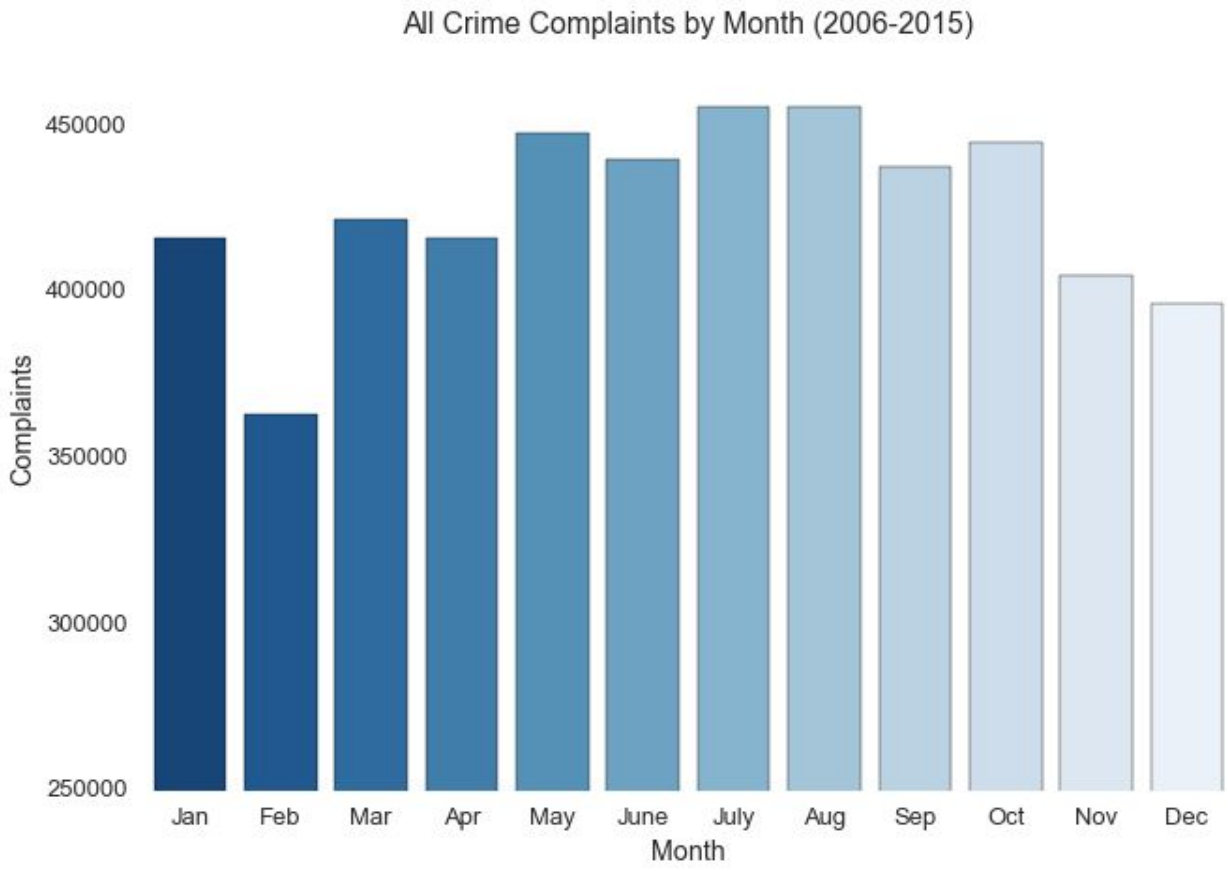
provides details of all crimes reported to the NYPD and it is available to the public under NYC's Open Data Law. It let's us take a look at important facts about where, when, and what kind of crimes get reported to the police.



As we can see, in 2015, complaints of crime decreased by over 10% from ten years earlier. This decrease in crime complaints continues a historic decrease in the actual crime rate that started well before 2006.

Seasonality

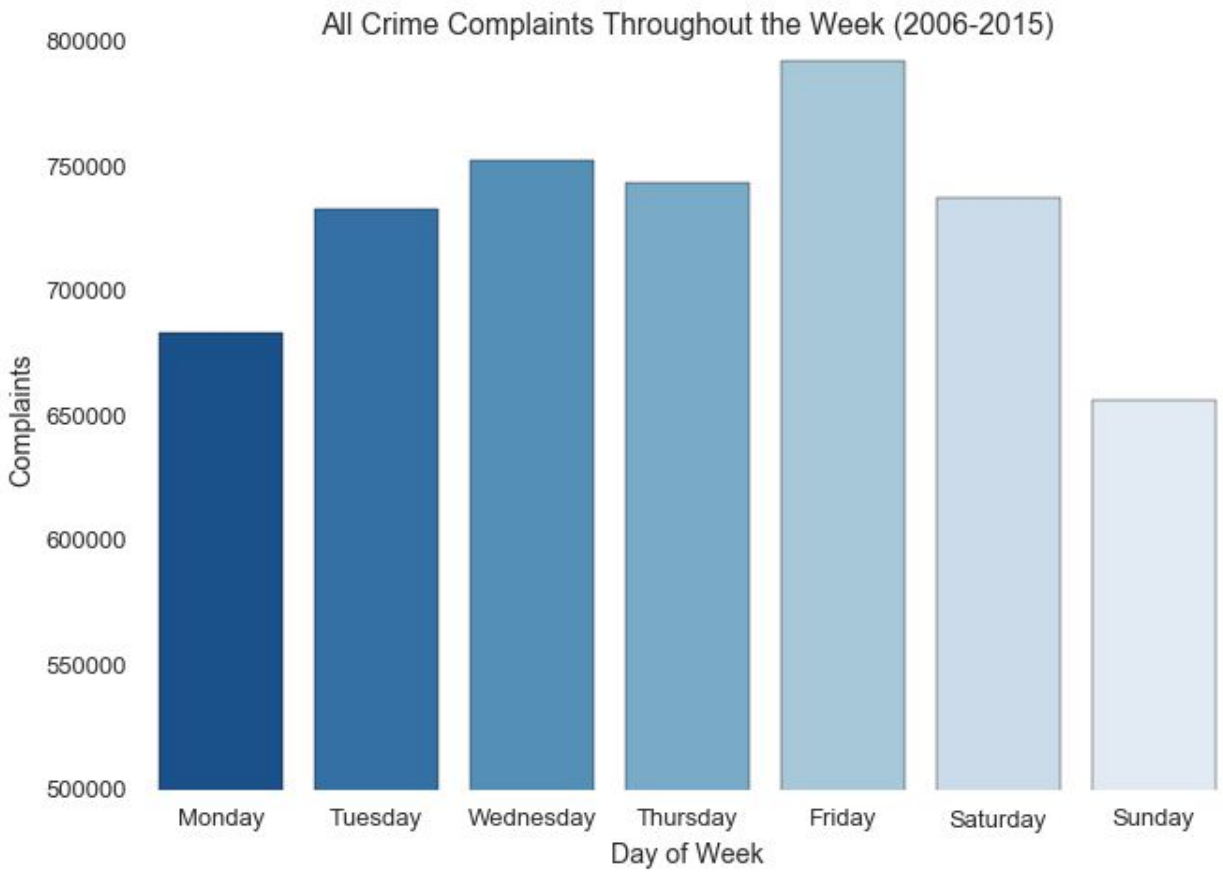
If you sum up all of the complaints that occurred in different months, a sense of the crime's seasonality beings to emerge. February, the shortest month, unsurprisingly has the lowest number of complaints.



With the exception of January, the cold-weather months have fewer crimes and the peak of complaints happen in July and August. This pattern will be worth exploring through the development of hypotheses and verification from other data sources. With more people out and about in the warm-weather months, opportunities for criminal activity most likely increase…

A Weekly Pattern

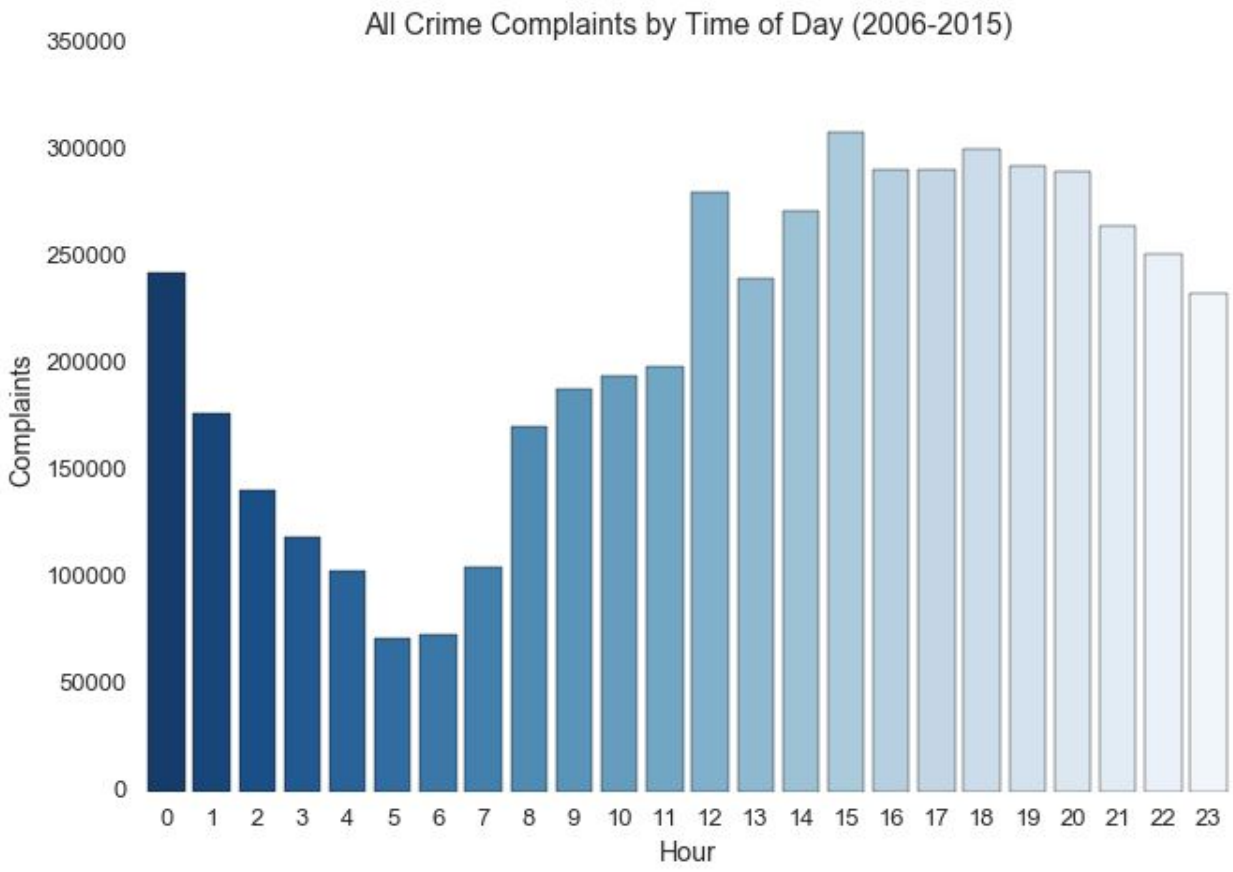
Crime complaints are not static throughout the week. They peak on Friday and hit a low on Sunday.



Are the city's packed bars, clubs, restaurants, and streets environments that invite criminal activity and complaints? Or is it that Friday is a common payday, with more people carrying more cash? Is Sunday a day of rest for crime, or is it the depressed weekend population that decreases opportunities for crime? Regional differences between the boroughs' criminal activity patterns might also explain this as well.

Crime's Circadian Rhythm

Here we see the number of crimes that start at a given hour. The sinusoidal shape is unmistakable, and invites many interpretations.



The uptick from 12:00-13:00 suggests that people should take care during their lunch hour. The early hours of the morning when most people sleep are, predictably, the safest in terms of crime.

Just what type of crimes are we looking at?

The dataset divides crimes into 3 levels.

A level represents a the severity of possible punishment

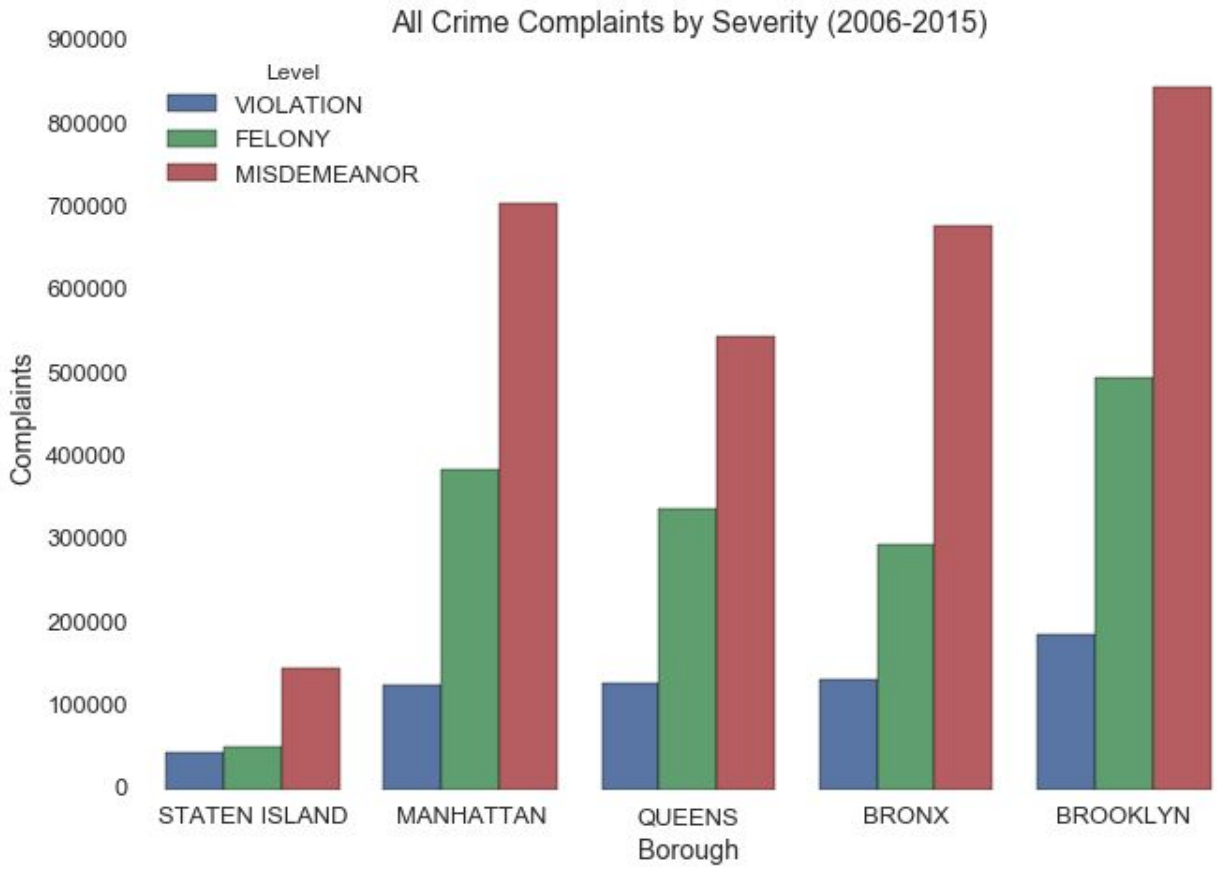
1) Violation – The potential sentence cannot be greater than fifteen days in jail.
2) Misdemeanor – The potential sentence for a misdemeanor between 15 days and one year in jail.
3) Felony – Punishment may exceed the one year.

	2006	2007	2008	2009	2010...
Felony	174,965	167,965	164,027	149,714	147,513
Misdemeanor	294,369	303,079	302,578	301,357	303,605
Violation	69,690	66,157	62,070	59,875	58,607

	...2011	2012	2013	2014	2015
Felony	148,082	153,697	153,730	150,716	146,137
Misdemeanor	294,697	294,697	282,483	276,770	260,133
Violation	55,419	58,172	58,745	62,877	62,306

Levels of crime across the boroughs

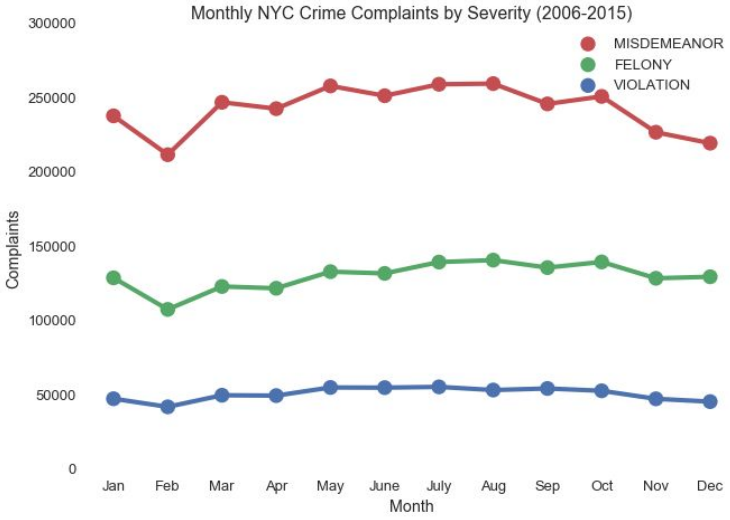
NYC's boroughs are cities unto themselves, and the way we look at crime in each borough must take this into account. While the distribution of violations is similar across the boroughs, we see changes in the amount of crime complaints in the last decade.

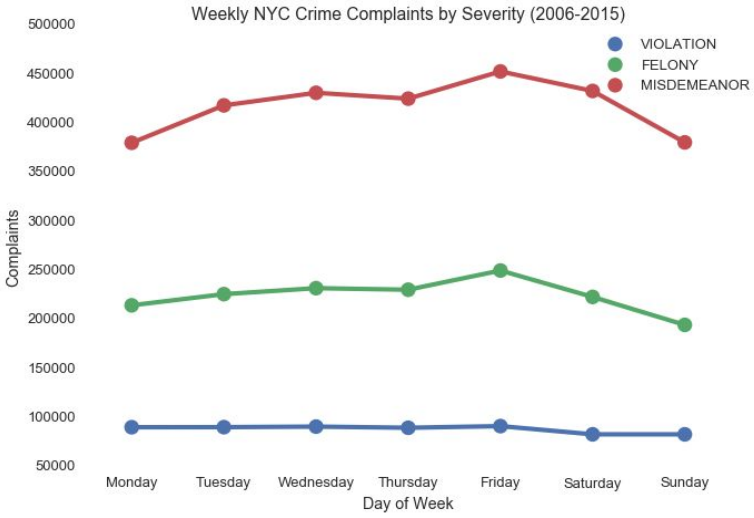


These volumes don't reflect populations. Queens has many more residents than Manhattan and the Bronx, yet it plainly has fewer felony and misdemeanor complaints. This is somewhat explained by Manhattan's booming weekday population, which dwarfs all the other boroughs.

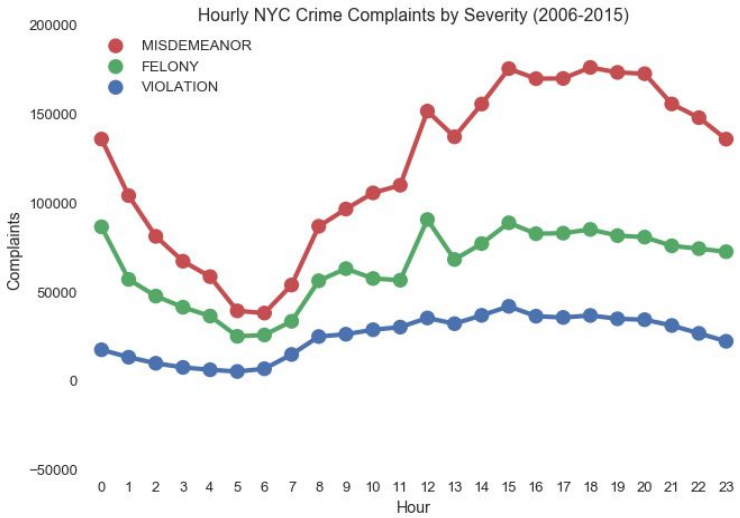
Another look at temporal relationships

After splitting levels of crime into the different categories, we can begin to zoom in on which crimes cause patterns.



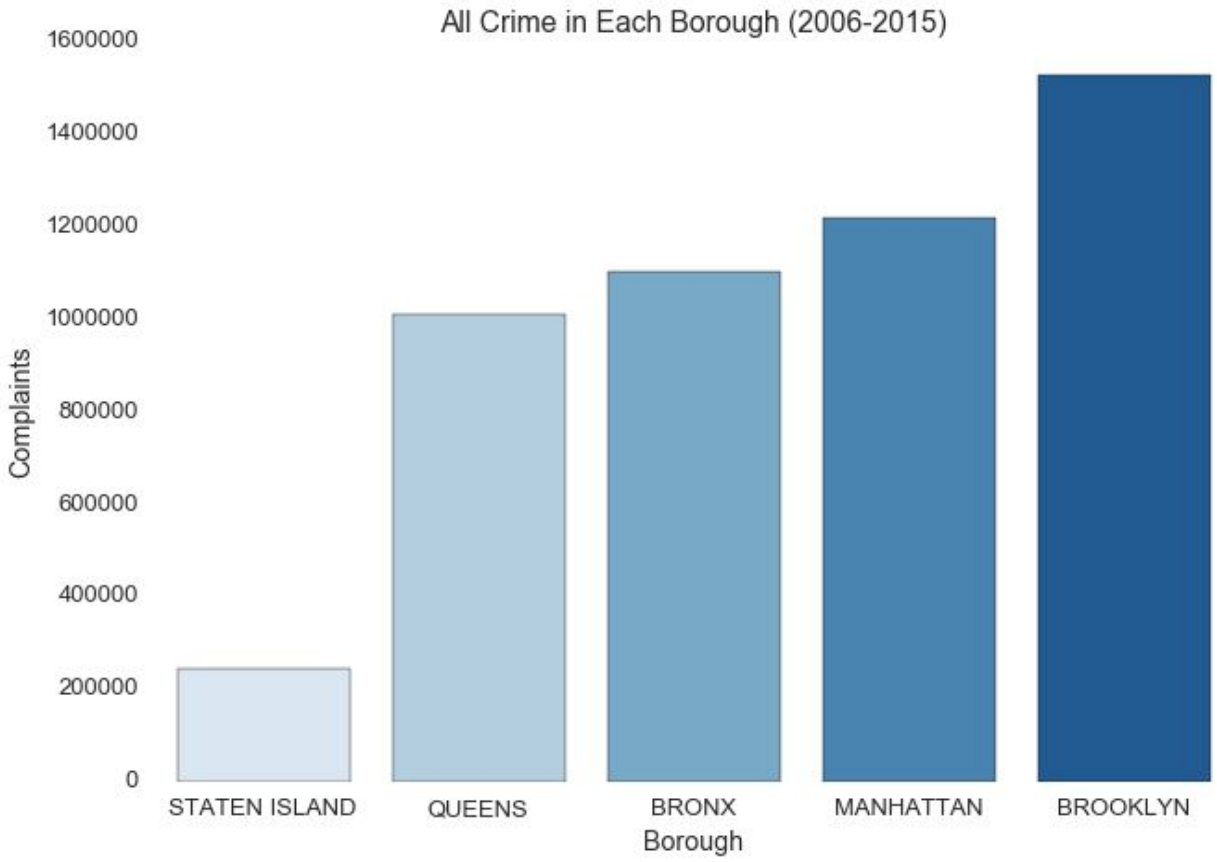


The least severe crime level, Violations, appear constant (or at least relatively so) throughout the year and week. Moreover, their impact on the number of crime complaints throughout the day is far less pronounced than the other two groups. From each temporal perspective, misdemeanors represent the bulk of the crime, and therefore influence the temporal patterns most strongly.



Finding Spatio-temporal Relationships I

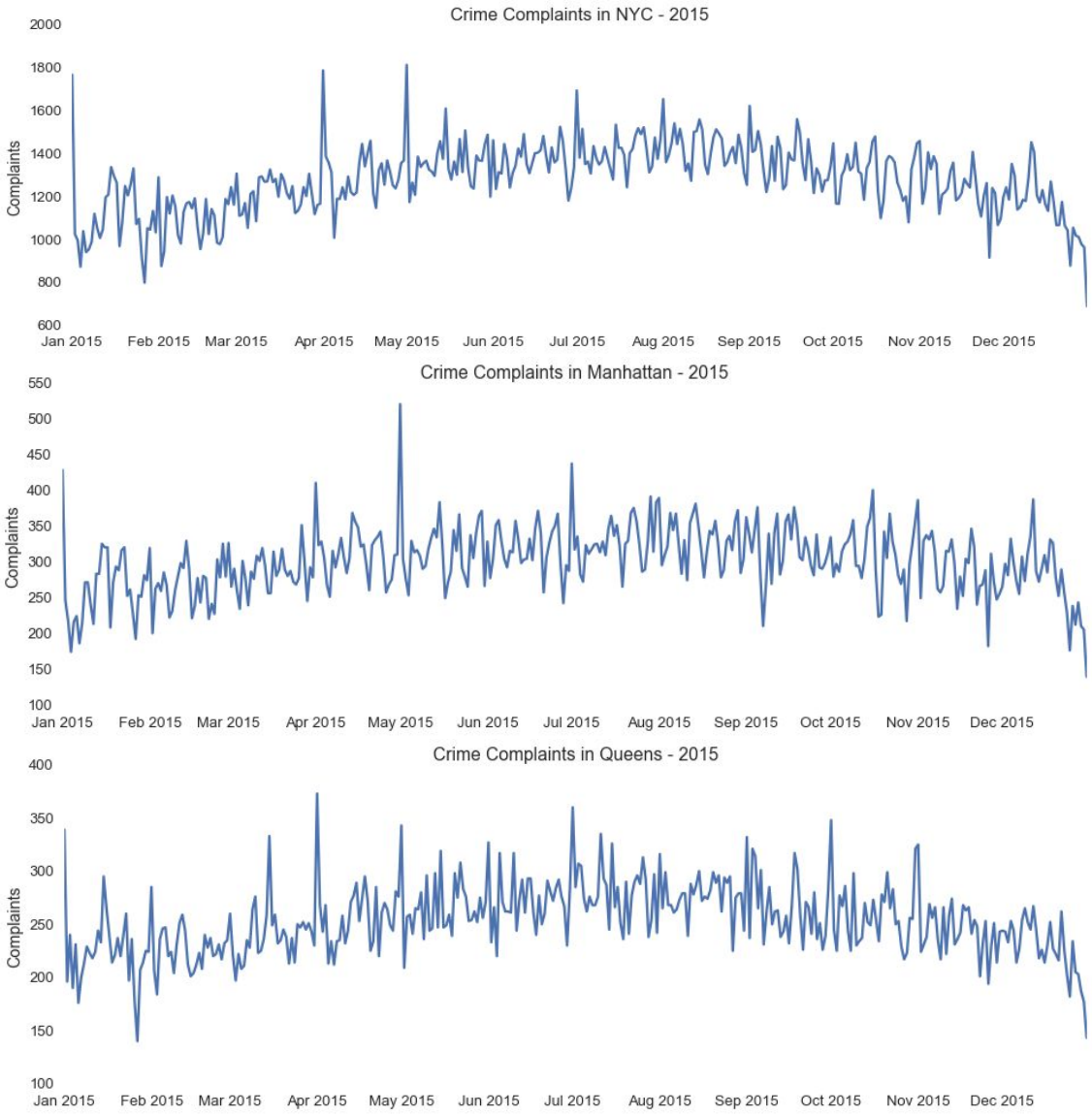
When it comes to predicting crimes and crafting policy that might reduce crime, it is necessary to finely slice the times and places where problems occur. To begin, we can look at where crimes happen a lot, and where they happen less often.



First, we will want to ascertain what impacts crime throughout all the boroughs, and then we can start to ask questions about what is responsible for the similarities and the differences. We will also be able to use GPS data and zip codes to characterize crime rates in smaller regions.

Finding Spatio-temporal Relationships II

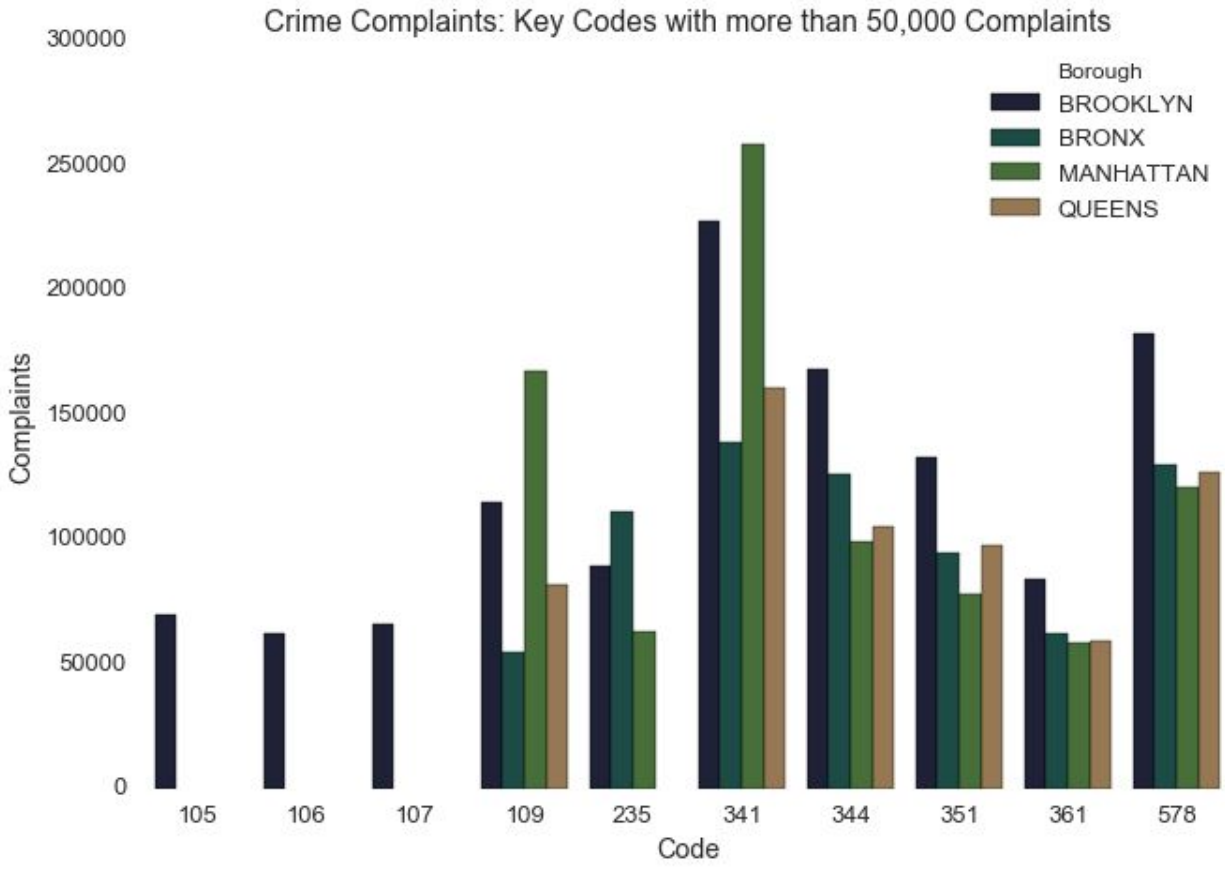
With so much data, we can take a bird's-eye view of crime and make conjectures about patterns and also irregularities. Perennial dips on Christmas and spikes on New Year's Day are witnessed throughout the boroughs and over all the years in the data.



These two days respectively correspond to times when people have decreased economic activity (most people stay home on Christmas) and increased economic activity (many people go out to bars or densely crowded areas on New Years Eve).

Finally - Specific Violations

Although we're looking at total complaints, we also have the opportunity to examine the individual crimes. Looking here at offenses with over 50,000 reports, we can see what takes the most resources from the people of NYC and the various jurisdictions that police them.



105 - ROBBERY
106 - FELONY ASSAULT
107 - BURGLARY
109 - GRAND LARCENY
235 - DANGEROUS DRUGS
341 - PETIT LARCENY
344 - ASSAULT 3 & RELATED OFFENSES
351 - CRIMINAL MISCHIEF
361 - LOITERING FOR DRUG PURPOSES
578 - HARASSMENT 2

DATA QUALITY ISSUES

Data Quality Summary: What follows is a description of data quality factors in each of the attributes of our main dataset. We ascertained the amount of missing, invalid, and suspicious data, and wherever possible we took steps to eliminate records that were misleading or unverifiable. Several columns were hard to verify, including ones that were qualitative descriptors of areas. Given the relatively large number of observations and the relatively small number of inaccuracies, we believe that information we gather from this dataset will be meaningful despite the existence of some invalid and suspicious entries.

5,101,231 rows, each representing a reported crime
24 columns containing various details of each reported crime
Columns

1. CMPLNT_NUM - Randomly generated persistent ID for each complaint
    1.1. There are 0 instances of missing data.
    1.2. The column contains entirely the base type of int, and there were no other data types present. This column serves as a unique identifier for each complaint in the dataset.
    1.3. There were no duplicate or missing values in this column, which means that the data were all valid.
    1.4. There were no suspicious events identified in this column, as it is simply a unique identifier.
    1.5. Script to assign base type, semantic type, and label

2. CMPLNT_FR_DT Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
    2.1. There are 655 instances of missing data.
    2.2. The column contains only string values that represent dates (i.e. in the form '01/01/2000'.
    2.3. The following instances of suspicious data were identified: The NYPD, who are the source of the data, stated that the dataset contains information on crimes from 2006-2015. However, there are 18,782 instances of crimes reported before 2006. Of these, 7 were reported in the year 1015, and the rest were reported in the 20th century. The 7 values are obviously invalid because NYC did not exist in 1015, however it is more difficult to ascertain the validity of dates reported from 1900. It is conceivable that someone reported crime that happened many years in the past. Given that there is no scientific way to choose a cutoff year, we will include all of crimes reported to have happened in the 20th century in our averages of crimes.
    2.4. The amount of crime reported on a given day did fluctuate in the dataset, as one would expect; but there were no days with suspiciously high or suspiciously low numbers of reported crime.
    2.5. Script to assign base type, semantic type, and label

3. CMPLNT_FR_TM Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
    3.1. There are 48 instances of missing data.
    3.2. The column contains only strings in 24-hour time format, e.g. 15:30:00. There were no other types of data present in this column.
    3.3. 903 times were reported as happening at hour 24 (e.g. 24:30:00), which is a time that does not exist
    3.4. The amount of crime reported at a given time did fluctuate in the dataset, as one would expect; but there were no times with suspiciously high or suspiciously low numbers of reported crime.
    3.5. Script to assign base type, semantic type, and label

4. CMPLNT_TO_DT Ending date of occurrence for the reported event, if exact time of occurrence is unknown
    4.1. There are 1,391,478 instances of missing data. These should not all necessarily be interpreted as incorrectly missing because most crime reported occurred on a single day, as opposed to occurring over a range of days.
    4.2. The column contains only string values that represent dates (i.e. in the form '01/01/2000'.
    4.3. The following instances of suspicious data were identified: The NYPD, who are the source of the data, stated that the dataset contains information on crimes from 2006-2015. 3 were crimes reported to have ended in 2016, and they were after the reported date.. These represent a typo: in each case the day and month were the same, but year was one year after the crime starting date. It is likely that there are many other instances of this type of inaccuracy. There 4,883 instances of crimes reported to have ended between 1912 and the end of 2005. These are not necessarily invalid because crimes could have been reported during the 2006-2015 range that occurred before 2006. (NOTE: Compare from and to date)1 crime reported to end in 2090, which is obviously invalid.
    4.4. The amount of crime reported on a given day did fluctuate in the dataset, as one would expect; but there were no days with suspiciously high or suspiciously low numbers of reported crime.
    4.5. Script to assign base type, semantic type, and label

5. CMPLNT_TO_TM Ending time of occurrence for the reported event, if exact time of occurrence is unknown
    5.1. There are 1,387,785 instances of missing data. These should not all necessarily be interpreted as incorrectly missing because most crime reported had a single report time, and not a range.
    5.2. The column contains only strings in 24-hour time format, e.g. 15:30:00. There were no other types of data present in this column.
    5.3. 1,376 times were reported as happening at hour 24 (e.g. 24:30:00), which is a time

that does not exist. Rather than invalidate and exclude data that had inaccuracies, we chose to rely on CMPLNT_FR_TM for correct time.

5.4. The amount of crime reported at a given time did fluctuate in the dataset, as one would expect; but there were no times with suspiciously high or suspiciously low numbers of reported crime.

5.5. Script to assign base type, semantic type, and label

6. RPT_DT Date event was reported to police

6.1. There are no instances of missing data.

6.2. The column contains only string values that represent dates (i.e. in the form '01/01/2000'.

6.3. All of the data fell into the correct date range of 2006-2015 and is not suspicious.

6.4. Similar to the dates of crimes, there were not suspiciously high or low dates. In the case of an assault that results in a murder after the report date, the report date is updated to the victim's date of death (see footnote 5).

6.5. Script to assign base type, semantic type, and label

7. KY_CD Three digit offense classification code

7.1. There are no instances of missing data.

7.2. The column contains only 3 digit ints, each of which represents a broad category of crimes. There were no other types of data in this column.

7.3. All of the key codes were in the correct format and identified a specific crime in the description column (even though sometimes a code was reported without the designated description), therefore none of the data is deemed suspicious.

7.4. All of the key codes fell into a designated category, so none of them were suspicious.

7.5. Script to assign base type, semantic type, and label

8. OFNS_DESC Description of offense corresponding with key code

8.1. There are 18,840 instances of missing data.

8.2. The column contains only strings that are descriptions of crimes. They match the key codes in the previous column. All data is of this type.

8.3. There are some instances of suspicious data. For example, 'PROSTITUTION AND RELATED OFFENSES' is the description applied to two different key codes: 115 and 234. While these seemed suspicious, further corroboration with the more specific PD_DESC column confirmed that these were actually different sub categories of related crimes.

8.4. There were no data that seemed suspicious, as they all fell into the predetermined categories.

8.5. Script to assign base type, semantic type, and label

9. PD_CD Three digit internal classification code (more granular than Key Code)

9.1. There are 4,574 instances of missing data.

9.2. The column contains only 3 digit ints, each of which represents a specific category of

crimes that are a subset of the broader Key Code data. There were no other types of data in this column.

9.3. All of the internal codes were in the correct format and identified a specific crime in the description column. Furthermore, the specific crime described in the internal code matched the broader key code designation. Therefore none of the data is deemed suspicious.

9.4. None of the data seemed suspicious.

9.5. Script to assign base type, semantic type, and label

10. PD_DESC Description of internal classification corresponding with PD code (more granular than Offense Description)

10.1. There are 4,574 instances of missing data.

10.2. The column contains only strings that are descriptions of crimes. They match the key codes in the previous column. All data is of this type.

10.3. There are no instances of suspicious data. All of the columns that were not missing corresponded to PD Codes, Key Codes, and Key Code Descriptions that were in the same general category of crime. The PD description was at least as specific as the Key Code description, and generally more so.o

10.4. There were no data that seemed suspicious, as they all fell into the predetermined categories.

10.5. Script to assign base type, semantic type, and label

11. CRM_ATPT_CPTD_CD Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely

11.1. There are 7 instances of missing data.

11.2. This is a binary column that indicates whether a flag was completed or only attempted. Other than the 7 missing values, the only instances of data are 87,913 ATTEMPTED complaints and 5,013,311 COMPLETED complaints.

11.3. None of this data seemed suspicious: It was very complete and only fell into the appropriate two categories.

11.4. There was no indication that this corresponded to a suspicious event.

11.5. Script to assign base type, semantic type, and label

12. LAW_CAT_CD Level of offense: felony, misdemeanor, violation

12.1. There are no instances of missing data.

12.2. The column contains only a categorical string data type. Each value is one of the strings: 'FELONY,' MISDEMEANOR,' 'VIOLATION.'

12.3. No suspicious data were found. There are several reasons to assume that none of the data are invalid. The person reporting a crime would not indicate this

12.4. information; rather it would be filled in by the police department because each class of crime corresponds to a particular key code.

12.5. No suspicious events were found. Similar to the above, this data field corresponds to the type of offense reported and for analysis purposes we assume that it was

filled in correctly.

12.6. Script to assign base type, semantic type, and label

13. JURIS_DESC Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
    13.1. There are no instances of missing data.
    13.2. The column contains only strings that identify one of 25 jurisdictions where the given complaint occurred (e.g. Subway, City Parks, or Metro North Railroad).
    13.3. This is categorical data with a limited number of categories and they are all consistent with actual NYC departments, so they are assumed to be correct.
    13.4. This information is assumed to be correct as long as it was reported accurately and entered into the dataset correctly.
    13.5. Script to assign base type, semantic type, and label

14. BORO_NM The name of the borough in which the incident occurred
    14.1. There are 463 instances of missing data.
    14.2. The column contains only 5 different text categories corresponding to NYC's different boros.
    14.3. No suspicious data was found. Only the 5 boros of NYC were found in this column and no other values.
    14.4. There were no suspicious events found. Subsequent analysis revealed levels of crime that were roughly proportionate to the size of each boro.
    14.5. Script to assign base type, semantic type, and label

15. ADDR_PCT_CD The precinct in which the incident occurred
    15.1. There are 390 instances of missing data.
    15.2. The column contains either 1-digit, 2-digit, or 3-digit int describing the precinct in which the crime occurred. There are no other data types in this column.
    15.3. All of the predicts were in the correct format , therefore none of the data is deemed suspicious.
    15.4. NYC has 123 police precincts in its 5 boros and they are named for those numbers. The only reported values in this column were numbers between 1 and 123, which does not seem suspicious.
    15.5. Script to assign base type, semantic type, and label

16. LOC_OF_OCCUR_DESC Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
    16.1. There are 1,127,341 instances of missing data. For some crime reports, although this description is missing, there is information provided about the name of a park or housing development where the crime occurred.
    16.2. The column contains strings that are descriptions of the specific location of occurrence. Each of the string fell into one of these categories: 'FRONT OF', 'INSIDE', 'OPPOSITE OF', 'OUTSIDE', and 'REAR OF'.

16.3. Each data in this column specifies one exact location description (inside, opposite of, front of , outside, rear of), so there were no suspicious data found.
16.4. There were no instances of suspicious events.
16.5. Script to assign base type, semantic type, and label

17. PREM_TYP_DESC Specific description of premises; grocery store, residence, street, etc.
17.1. There are 33, 279 instances of missing data.
17.2. The column contains strings that are description of premises. There are no other data types in this column.
17.3. All of the data describe some type of premises: grocery store, residence, street, etc. So none of them were suspicious.
17.4. None of data seemed suspicious.
17.5. Script to assign base type, semantic type, and label

18. PARKS_NM Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
18.1. There are 5,093,632 instances of blank data, but these are not necessarily invalid or missing. This column is only filled when a crime is reported to have taken place in a park. There are 54,572 instances of missing data, where in a Park was listed under the PREM_TYP_DESC, but no park name was provided.
18.2. The column contains only strings that identify different parks or playground in NYC.
18.3. We assume that all the park names are valid and that the data was accurately recorded.
18.4. In order to corroborate validity, we analyzed whether PREM_TYP_DESC matched the park name. In many cases crimes happened on premises such as STREET or RESIDENCE, and a park was listed in the vicinity. Since a park can be used as a general reference of location, it is not possible to strictly verify whether or not it is valid.
18.5. Script to assign base type, semantic type, and label

19. HADEVELOPT Name of NYCHA housing development of occurrence, if applicable
19.1. There are 4,848,026 instances of blank data, but similar to the parks, this does not necessarily represent missing data because the majority of crimes did not happen in public housing residences. There are 142,302 instances of missing data, wherein 'RESIDENCE - PUBLIC HOUSING' was listed under the PREM_TYP_DESC, but no Public Housing Name name was provided.
19.2. The column contains only strings that identify different public housing developments in NYC.
19.3. There were no suspicious instances of data. We assume that all the public housing development names are valid and that the data was accurately recorded.
19.4. In order to corroborate validity, we analyzed whether PREM_TYP_DESC matched the park name. In many cases crimes happened on premises such as STREET or RESIDENCE, and a housing development was listed in the vicinity. Since a park can

be used as a general reference of location, it is not possible to strictly verify whether or not it is valid.

19.5. Script to assign base type, semantic type, and label

NOTE FOR BELOW: COORDINATES USED: E:-73.69 N:40.93 S:40.47 W:-74.26

20. X_COORD_CD X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
    20.1. There are 188,146 instances of missing data.
    20.2. The column contains only ints that represent horizontal values in a localized coordinate system for NYC. There are no other data types in this column.
    20.3. All the data was in the correct latitude format.
    20.4. We verified that all points were within a square perimeter around NYC and were not suspicious.
    20.5. Script to assign base type, semantic type, and label

21. Y_COORD_CD Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
    21.1. There are 188,146 instances of missing data.
    21.2. The column contains only ints that represent vertical values in a localized coordinate system for NYC. There are no other data types in this column.
    21.3. The following instances of suspicious data were identified:
    21.4. The following instances of suspicious events were identified:
    21.5. Script to assign base type, semantic type, and label

22. Latitude Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
    22.1. There are 188,146 instances of missing data.
    22.2. The column contains only floats that represent latitude values. There are no other data types in this column.
    22.3. All the data was in the correct longitude format.
    22.4. We verified that all points were within a square perimeter around NYC and were not suspicious.
    22.5. Script to assign base type, semantic type, and label

23. Longitude Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
    23.1. There are 188,146 instances of missing data.
    23.2. The column contains only floats that represent longitude values. There are no other data types in this column.
    23.3. All the data was in the correct latitude format.
    23.4. We verified that all points were within a square perimeter around NYC and were not suspicious.

23.5. Script to assign base type, semantic type, and label

24. Lat_long
    24.1. There are 188,146 instances of missing data.
    24.2. This column is a concatenation of columns 22 and 23 that form a full coordinate pair. The details with respect to data quality are the same as for those two columns.
    24.3. All the data was in the correct format.
    24.4. We verified that all points were within a square perimeter around NYC and were not suspicious.
    24.5. Script to assign base type, semantic type, and label

Part II - Analysis

Experimental techniques and methods

The basis of our analysis relied on manipulation of the crime dataset using PySpark on NYU's dumbo cluster. We used default settings and were able to efficiently run dozens of different scripts that extracted key-value pairs corresponding to given features and targets of interest. For example, when our goal was to analyze the number of robberies that occurred at a given time of day, our key was data from the COMPLAINT_TIME_FROM column and our value was the count of robberies indicated in the KEY_CODE column by code 105. Our code consistently produced results in around one minute, so we did not pursue further optimization. With this method we were able to analyze the following targets from the crime dataset: counts of total crime; counts of felonies, misdemeanors, and violations; counts of crime by key codes. We applied standard tools from the pandas, numpy, and seaborn libraries in python to produce the data description in the Data Summary.

A similar method was used on additional datasets identified below. We used PySpark in dumbo with default settings in order to generate key-value pairs that corresponded to factors provided in a dataset and a count. We also included zip codes as part of every key so that we could organize our analysis around small regions of the city.

In order to assign zip code for every instance, we adopted the idea of 1-NN (1-nearest neighborhood): we randomly sampled GPS information of 5,000 instances from our dataset and queried their exact zip codes through Google Map API.(5) Then for each instance in the whole dataset, we computed its distances to all samples by GPS location and then assigned the corresponding zip code to this instance (the zip code of its nearest sample).

For most of our PySpark programs (except for one that assigns zip code), runtime was around 1 minute with default settings, so we didn't do extra optimization. When assigning zip code, we first used a small sample of 100 instances to measure the timing performance with default settings, and it only took around 1 minute. Since the computing time of distances is linear to the number of samples, we believe that the time using 5,000 instances would be within an hour, which was confirmed by our actual runtime (around 30 minutes). So the computing time is acceptable without further optimization.

Additional Data

In order to refine our general research questions about historic crime attractors and broken window theory into specific hypotheses, we first surveyed available public datasets with two requirements in mind: first, the data had to include zip codes from throughout NYC, and second, the data had to tell us something about crime attractors or urban disorder

and decay factors. We identified the following datasets:

Licensed businesses by license type
Housing maintenance violations by severity
Rodent inspection data

Features from additional datasets

Features extracted from the licensed business dataset:

businesses - the total number of businesses in a zip code. This variable was chosen as an overall indicator of economic activity.

cafe - The number of sidewalk cafes in a zip code. Outdoor businesses were an example of crime attractors in early research. (3)

pawn - The number of pawnbrokers in a zip code. This

debt - The number of debt collection agencies in a zip code

cig - the number of cigarette retailers in a zip code.

Features extracted from the housing maintenance violations dataset:

Note: all of these metrics were interpreted as indicators of urban decay (4)

hpdA - Type A housing maintenance violations (lowest severity)

hpdB - Type B housing maintenance violations (moderate severity)

hpdC - Type C housing maintenance violations (highest severity)

hpdTotal - All housing maintenance violations in a zip code

Features extracted from rodent inspection dataset:

Note: all of these metrics were interpreted as indicators of urban decay (4)

negativeRodent - the sum of Active Rat Signs and Problem Conditions reported in a zip code.

positiveRodent - this variables was named positive because administering poison would presumable mitigate rodent problems. However, we interpreted the poison administration as an additional indicator of rodent activity.

Data Preprocessing and Correlation Table

While we had ample data that assessed the overall economic activity from licensed businesses, there were not enough examples of specific types of businesses in each zip code to warrant using those features. We retained the overall number of businesses as a feature. In addition, we consider the possibility that higher population might always imply more crimes. Thus, before doing all the regression analysis described later, we preprocessed each feature via dividing the counts in different zip code regions by the population in that region to exclude such possibility.

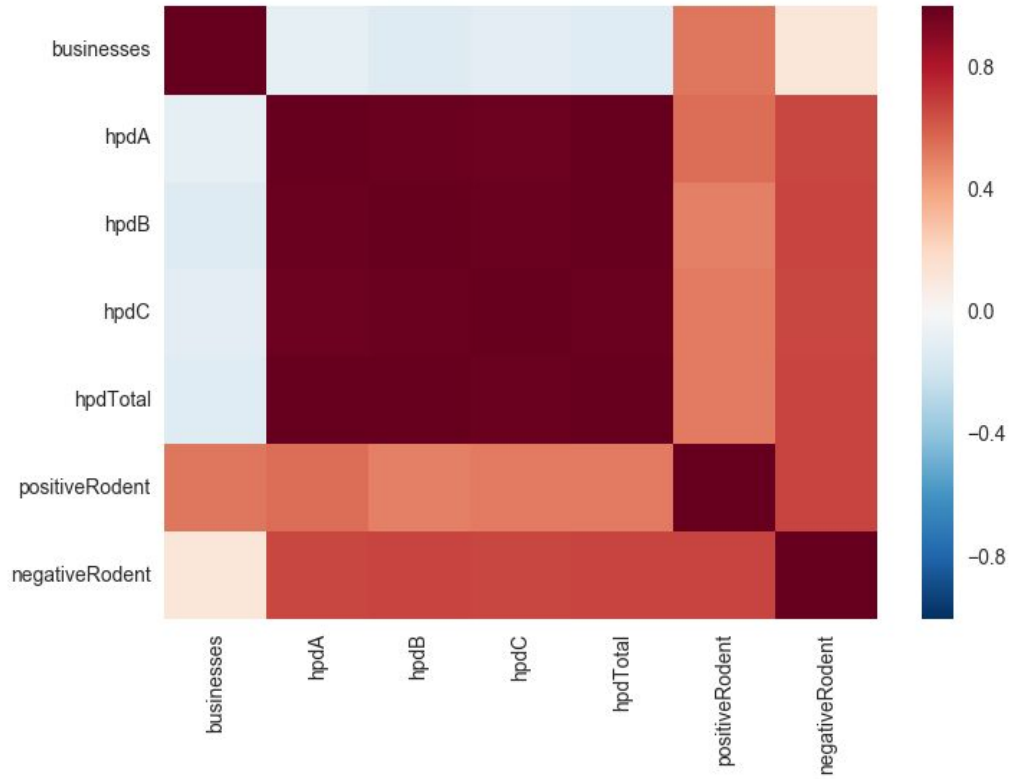
In order to select features, we computed the correlation matrix adjusted for population. Given the high correlation (see Table 1) among all housing violation features,

we removed hpdA, hpdB, and hpdC and only focused on the total number of housing violations. A better visualization of the correlation between features is shown in Figure 1. Therefore, we narrowed down to only four features: businesses, hpdTotal, positiveRodent, negativeRodent.

Table 1: Correlation between features

	businesses	hpdA	hpdB	hpdC	hpdTotal	positiveRodent	negativeRodent
businesses	1.000	-0.092	-0.129	-0.108	-0.118	0.539	0.117
hpdA	-0.092	1.000	0.990	0.980	0.994	0.555	0.668
hpdB	-0.129	0.990	1.000	0.988	0.999	0.506	0.676
hpdC	-0.108	0.980	0.988	1.000	0.992	0.516	0.665
hpdTotal	-0.118	0.994	0.999	0.992	1.000	0.520	0.675
positiveRodent	0.539	0.555	0.506	0.516	0.520	1.000	0.678
negativeRodent	0.117	0.668	0.676	0.665	0.675	0.678	1.000

Figure 1: Correlation Heatmap of Features



Final Hypotheses

From the selected attributes in these datasets we were able to propose the following hypotheses:

With respect to crime attractors:

1. If an area (zip code) has a high number of centers of economic activity (licensed businesses), then it will also have a high crime rate.
    a. Sub-hypotheses: this relationship extends beyond a high overall crime rate to the following specific targets: high rates of felony, misdemeanor, violation, theft, harassment, and assault.

With respect to broken window theory:

1. If an area exhibits high numbers of housing maintenance violations, then it also has a high crime rate.
    a. Sub-hypotheses: this relationship extends beyond a high overall crime rate to the following specific targets: high rates of felony, misdemeanor, violation, theft, harassment, and assault.
2. If an area exhibits high rodent activity, then it also has a high crime rate:
    a. Sub-hypotheses: this relationship extends beyond a high overall crime rate to the following specific targets: high rates of felony, misdemeanor, violation, theft, harassment, and assault.

We looked specifically at theft, harassment, and assault because we identified them as the most common individual types of crime in our data summary.

Ordinary Least Squares (OLS) Regression

Using zip codes as unique identifiers, we merged in variables from the business, housing violation, and rodent datasets with our target variables of various types of crime counts. We also merged in population data from each zip code. An ordinary least squares (OLS) regression analysis was set up with zip codes as observations and output from the auxiliary datasets as features. We took advantage of the sklearn and statsmodel libraries in python to perform the analysis and present detailed summary reports.

We started with a preliminary linear OLS model by regressing total crime on businesses, hpdTotal, positiveRodent, and negativeRodent. According to the results shown in Figure 2, the p-value of negativeRodent in this regression model is 0.710, which is not statistically significant at almost any level of significance. The adjusted R-squared value associated with this model is 0.147. If we remove negativeRodent, the regression results in Figure 3 indicates that the adjusted R-squared value increases to 0.152, which implies the feature negativeRodent negatively impacts the crimes. Therefore, we continued doing all the later regression analysis without taking negativeRodent as a feature.

Figure 2: Regression Results of Total Crime Versus businesses, hpdTotal, positiveRodent, and negativeRodent

OLS Regression Results

Dep. Variable:	total	R-squared:	0.168
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	8.119
Date:	Tue, 09 May 2017	Prob (F-statistic):	5.47e-06
Time:	21:51:02	Log-Likelihood:	-220.29
No. Observations:	166	AIC:	450.6
Df Residuals:	161	BIC:	466.1
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.633e-17	0.072	1.06e-15	1.000	-0.142	0.142
businesses	-0.1999	0.097	-2.053	0.042	-0.392	-0.008
hpdTotal	0.0135	0.110	0.123	0.902	-0.203	0.230
positiveRodent	0.4933	0.131	3.759	0.000	0.234	0.752
negativeRodent	-0.0430	0.115	-0.373	0.710	-0.271	0.185

Omnibus:	41.790	Durbin-Watson:	1.523
Prob(Omnibus):	0.000	Jarque-Bera (JB):	78.587
Skew:	1.187	Prob(JB):	8.61e-18
Kurtosis:	5.393	Cond. No.	3.53

Figure 3: Regression Results of Total Crime Versus businesses, hpdTotal, positiveRodent

OLS Regression Results

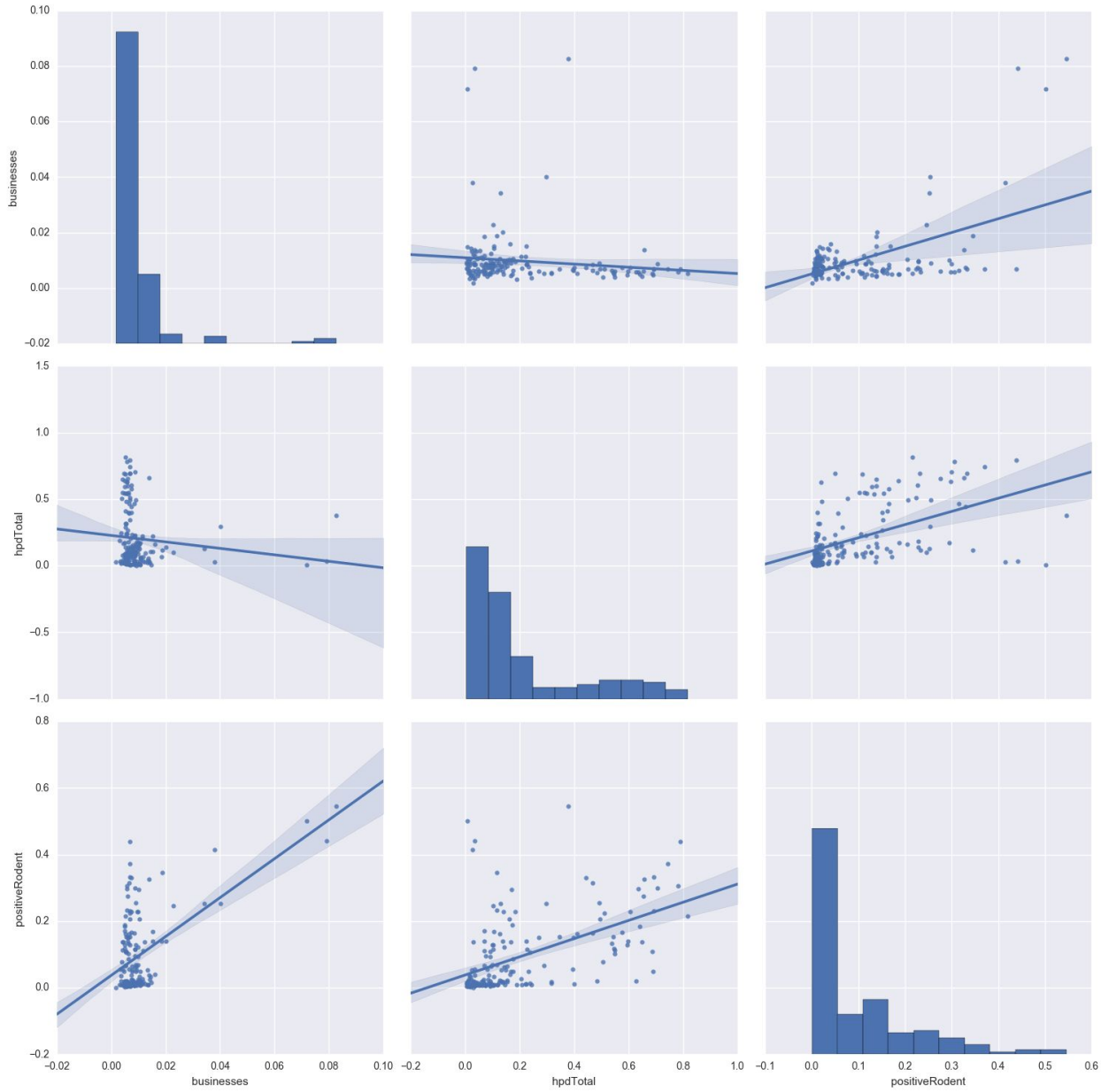
Dep. Variable:	total	R-squared:	0.167
Model:	OLS	Adj. R-squared:	0.152
Method:	Least Squares	F-statistic:	10.84
Date:	Tue, 09 May 2017	Prob (F-statistic):	1.58e-06
Time:	20:42:51	Log-Likelihood:	-220.36
No. Observations:	166	AIC:	448.7
Df Residuals:	162	BIC:	461.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.633e-17	0.072	1.06e-15	1.000	-0.142	0.142
businesses	-0.1957	0.096	-2.029	0.044	-0.386	-0.005
hpdTotal	-0.0023	0.101	-0.023	0.982	-0.201	0.197
positiveRodent	0.4701	0.115	4.079	0.000	0.242	0.698

Omnibus:	42.390	Durbin-Watson:	1.517
Prob(Omnibus):	0.000	Jarque-Bera (JB):	80.105
Skew:	1.202	Prob(JB):	4.03e-18
Kurtosis:	5.410	Cond. No.	2.87

After ending up with three features (businesses, hpdTotal, and positiveRodent), we used a pairwise scatterplot to visualize the pairwise relationships between the features businesses, hpdTotal, and positiveRodent. Since all these features are continuous variables, we believe that using scatterplot is a recommendable way to detect correlations. As indicated in Figure 4, our results suggest that there is a negative relationship between businesses and hpdTotal, while the other two relationships are positive.

Figure 4: Pairwise Scatterplots and Histograms for Features in OLS



To further investigate the multicollinearity between features, we computed the Variance Inflation Factor (VIF) (summarized in Table 2), which is a measure of the severity of multicollinearity in an OLS regression. It is obvious that all the VIF values are below 5, which implies that the multicollinearity is not severe at all by the rule of thumb. So we ultimately decided to use the specification of linear regression with businesses, hpdTotal, and positiveRodents as regressors.

Table 2: Variance Inflation Factor for Coefficients of Each Feature in OLS

	businesses	hpdTotal	positiveRodent
VIF	2.399	2.328	4.280

Results and Discussion

The results of the regression analysis revealed both conclusive and inconclusive factors that relate to crime rates. With respect to the impact of the presence of businesses per capita in a zip code on various crime targets, the results were often statistically significant for a p-value of 0.05, but the direction of the coefficient for the business feature changed in contradictory ways. Results (summarized in table 3) show that for a given unit increase in the number of licensed businesses per capita, total crime decreases by -0.1957 and the 95% confidence interval includes only negative values; however, a unit increase in the same variable with respect to the per capita number of violations, felonies, and misdemeanors results in an increase in those respective levels of crime. According to these results we cannot reject the null hypothesis or suggest that a high prevalence of businesses increases the number of crimes.

Also illustrated in Table 3 are the OLS results for housing complaints and rodent mitigation efforts, which we are treating as measures of broken window environmental factors. Here the results support the rejection of the null hypothesis, given that there were statistically significant positive associations between crime and housing violations, as well as between crime and signs of rodents.

Table 3 : OLS results for features per capita and various crime targets in NYC zip codes

Variable:	Businesses	Housing Violations	Rodent Activity
Target:	Coefficient, (p-value) [95% Confidence Interval]		
Total Crime	-0.1957, (0.044) [-0.386, -0.005]	-0.0023, (0.982) [-0.201, 0.197]	0.4701, (<0.001) [0.242, 0.698]
Felony	0.8086, (<0.001) [0.723, 0.895]	0.1937, (<0.001) [0.104, 0.284]	0.1472, (0.005) [0.044, 0.250]
Misdemeanor	0.6936, (<0.001) [0.603, 0.784]	0.2228, (<0.001) [0.128, 0.318]	0.2610, (<0.001) [0.152, 0.370]
Violation	0.6489, (<0.001) [0.519, 0.778]	0.2701, (<0.001) [0.135, 0.406]	0.1440, (0.068) [-0.011, 0.299]
Theft	-0.1231, (0.105) [-0.272, 0.026]	0.6512, (<0.001) [0.495, 0.807]	0.0523, (<0.001) [-0.126, 0.230]
Assault	-0.1555, (0.067) [-0.322, 0.011]	0.5417, (<0.001) [0.368, 0.715]	0.0600, (0.552) [-0.139, 0.259]
Harassment	-0.4835, (<0.001) [-0.654, -0.313]	0.0705, (0.436) [-0.108, 0.249]	0.5564, (<0.001) [0.353, 0.760]

Significant Positive Relationship Significant Negative Relationship

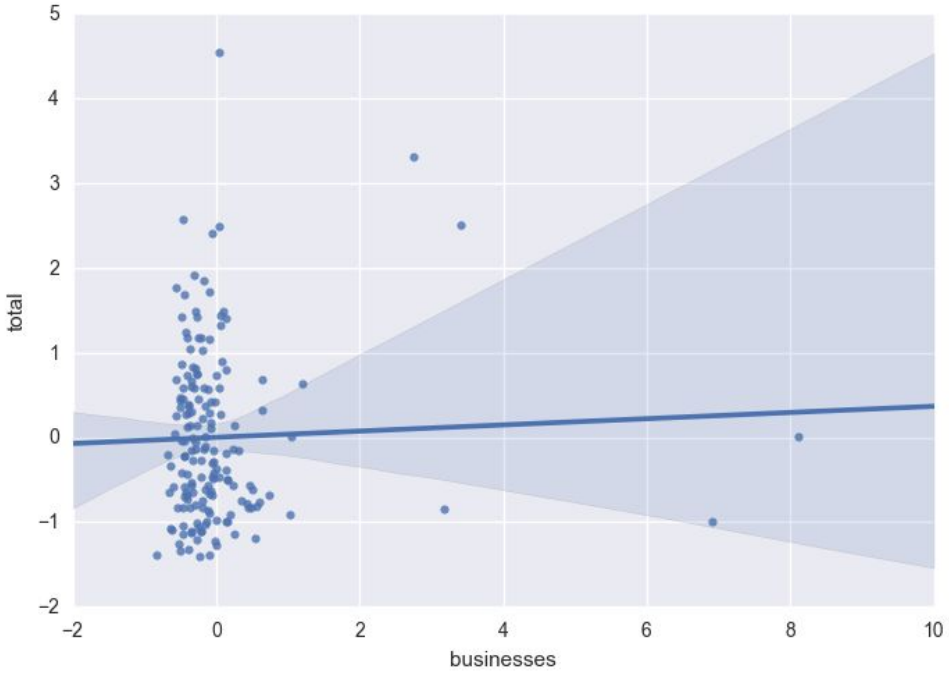
As mentioned above, the conclusion derived from the business efforts is inconsistent. For the three different levels of crime, businesses impact crimes negatively while for the total crime businesses contribute positively. To investigate the possible explanation for this inconsistency, we compared the adjusted R-squared value of the corresponding regressions.

Table 3 : R-squared of the OLS results for the businesses per capita

	Total Crime	Felony	Misdemeanor	Violation
Adjusted R-squared	0.152	0.827	0.807	0.607

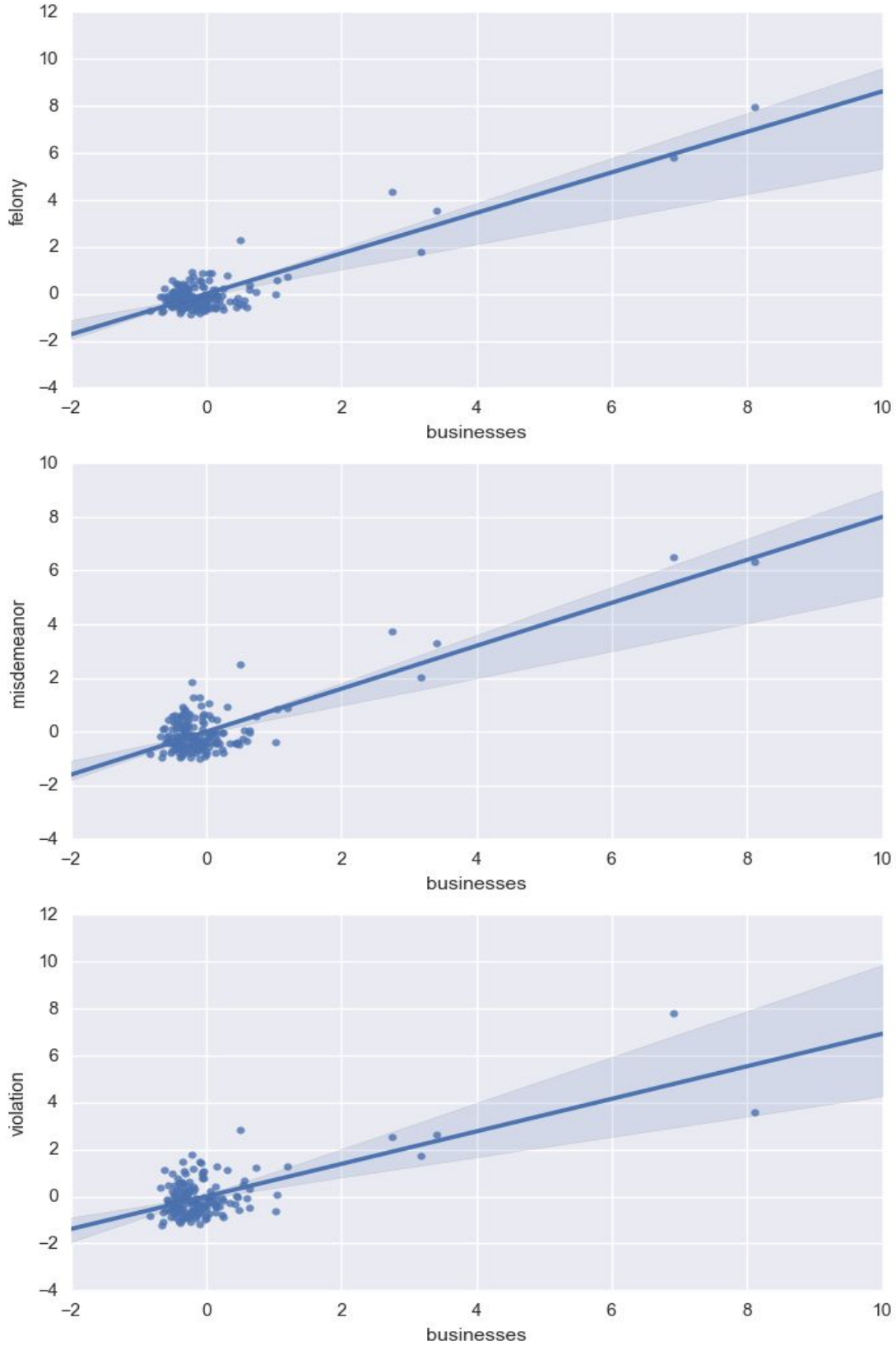
As indicated in Table 3, the adjusted R-squared value is much lower when regressing on the total crime than when regressing on each of the three different levels of crime (i.e. felony, misdemeanor, and violation). To examine why this happens, we continued our investigation by plotting total crime versus businesses.

Figure 5: Plot of Total Crime Versus Businesses



We can easily see from Figure 5 that most data points are clustered in one place and the shaded area is quite spread out, which indicates that this linearity discovered probably happens because a few influential data points determine the slope of the line. We can imagine that if you take out these few data and plotted the rest of the data versus businesses, a totally different straight line can be fit in. However, when plotting felony, misdemeanor, and violation versus businesses (shown in Figure 6), the shared areas are relatively small and all the data points lie close to the fitted line. This finding implies that the fitted lines in these three plots are better than the one in Figure 5. Also, it explains why inconsistency of the conclusion derived from business efforts happens to a great extent.

Figure 6: Plot of Felony, Misdemeanor, Violation Versus Businesses



From these results we can address our original question and suggest that the factors that foster crime in NYC today are somewhat different than the factors that fostered crime during high-crime eras several decades ago. Past research that we have already cited suggested that areas with lots of of business activity tend to also be areas with higher levels of crime rates. Our analysis is not conclusive about that relationship persisting in the last decade. It is possible that a more extensive analysis, one that analyzes relationships between specific types of businesses and specific types of crime, would yield more conclusive results. Looking at two factors related to broken window theory suggests that indicators of urban disorder are still important indicators of crime.

Limitations and Future Works

The most significant limitation that we faced was difficulty in identifying enough datasets with a temporal dimension that exactly matched that of the main crime dataset (2006-2016). The datasets we found were very recent, but did not overlap completely with the crime data timeframe. If we had been able to look at all of our additional features from a temporal perspective, we could have extended our analysis to see how recent fluctuations in (e.g. housing violations), affect crime. Given this limitation, we centered our analysis around spatial comparisons. With respect to our main research question (whether factors from NYC's era of high crime are still important today), excluding temporal data from 2006-2016 was not a hindrance. This is because taking important crime indicators from the 1980s and 1990s and looking at them over the last decade still suffices to answer our hypotheses.

One aspect we can explore further is incorporating seasonality into the models. For instance, we might investigate why some crimes happen more frequently during weekends. Also, considering seasonality might extend our hypotheses from discovering correlation to examining causality, which can be another interesting topic. Some other future explorations might be investigating the relationship between drug markets and crimes, analyzing how demographics might impact different crimes, and probing into specific criminal acts such as computer crimes and medical fraud.

Individual Contributions

Daniel Amaranto (da1933) - Visualizations for data summary, Spark scripts for daily crime rate, literature review of crime research to inform hypothesis choices, spark scripts for licensed business dataset, data merging, analysis and write-up.
Luyu Jin (lj1035) - Correlation and regression analysis, data cleaning and preprocessing, PySpark scripts for the rodent dataset and housing development dataset, PySpark scripts and visualization for crime datasets, and report write-up.

Siyuan Xiang (sx550) - PySpark scripts for data summary and data quality issues, PySpark scripts for preparation of data analysis, Python scripts for Google Map API, PySpark scripts for computing zip codes and report write-up.

Summary/Conclusion

Using big data methods, we were able to efficiently analyze over 5 million records of crime complaints that covered all of NYC for ten years. We found external datasets allowed us to corroborate whether or not certain economic and urban-environmental factors were related to the crime rate. We conclude that even though the crime rate of present-day New York City is a small fraction of what it once was, some of the essential factors that contribute to the crime rate today are the same as those of many years ago. Crime attractors such as areas of economic activity were not conclusively linked to crime in the last decade, but metrics of poor urban environments (prevalent housing violations and rodent activity) were consistently related to higher crime levels. Therefore continued vigilance and awareness of these important factors should be maintained.

References

1) "What 'broken windows' policing is." *The Economist*. 27 January 2015. Web. 26 April 2017 Accessed.
2) Bernasco, Wim and Richard Block. "Robberies in Chicago: A Block-Level Analysis of the Influence of Crime Generators, Crime Attractors, and Offender Anchor Points." *Journal of Research in Crime and Delinquency*. 48(I) (2011) p33-57.
3) Brantingham, Paul J. and Patricia L. Brantingham. "Criminality of Place: Crime Generators and Crime Attractors." *European Journal of Criminal Policy & Research*. 3 (1995) p5-26.
4) Eck, John E. and David L. Weisburd. "Crime Places in Crime Theory." *Hebrew Univeristy of Jerusalem Legal Research Paper*. 4 (2015) p1-33.
5) Kelling, George L. and James Q. Wilson. "Broken Windows: The police and neighborhood safety." *The Atlantic*. March 1982. 26 April 2017 Accessed.
6) "Geocoding API." Google Maps APIs. Google Inc. 8 May 2017 Updated. Web.