Big Data Project | DS 1003 Spring 2017

Joyce Wu, Alexandra Simonoff, Felipe Ducau

May 10, 2017

TABLE OF CONTENTS

Sections are clickable links

1. Abstract
2. Introduction
3. Dataset Description
4. Experimental Setup

PART I

5. Data Quality Summary
    - 5.1. Valid/Invalid/Null Counts Table
    - 5.2. Valid/Invalid/Null Justification and Reasoning
6. Other Data Quality Findings
7. Data Cleaning
8. Data Exploration
    - 8.1. Seasonality of Crime
    - 8.2. Violent Crimes
    - 8.3. Crime by Time of Day
    - 8.4. Crime by Day of Week
    - 8.5. Forecasting Crime Through 2020
    - 8.6. Drinking and Driving
    - 8.7. Observing Shoplifting Trends
    - 8.8. Difference Between Occurrence Date and Police Report Date
    - 8.9. Pre 2000 Crime Profile

8.10. Level of Offense by NYC Borough
8.11. Type of Offense by NYC Borough
8.12. Number of Crimes by Precinct

PART II

9. Hypotheses
9.1. Occurrence of crime is related to weather
9.1.1. Temperature
9.1.2. Snow
9.1.3. Rain
9.2. Crime per capita is related to population density
9.3. Potential jail sentence time compared to crime rates
9.4. Age is related to crime rates by borough
9.5. Minorities are targeted by the NYPD
9.6. There are more noise complaints when there are drunk driving offenses
9.7. Gentrification effect on crimes
10. Summary and Conclusions
10.1. Comments in geographical analysis
11. Contributions
12. References

1. Abstract

This project seeks to identify interesting trends that arise when looking at New York City police reports created between 2006 and 2015. We look to identify days with an unusually high or low amount of crime, the times of day that crime are more likely to happen and the precincts in New York where crime is more or less likely to happen. By looking into the police reports, we can get a feel for the landscape of crime in New York and inform decisions about when and where police should be available to mitigate crime.

Crime is most likely to happen when people are going through their daily routine, out and about town. Non-violent crime drops in the early hours of the morning and rises between 8am and 8pm, while all crime drops on days where there are massive storms or significant holidays. Similarly, more crime happens during the summer than the winter. Department and chain stores see an overwhelming majority of the shoplifting offenses committed, while jewelry stores see more grand larceny than petit larceny (as expected, given the price of fine jewelry). Rape, murder and fraud have the highest average delay between time of offense and report time. Regions of the Bronx, Brooklyn and Staten island see the most crime, while typically Manhattan sees fewer crime relative to the other boroughs.

2. Introduction

This report is attempting to do several things, namely, evaluate the quality of the dataset and identify possible flaws in the data and examine interesting trends and patterns that could produce actionable insights. We look to identify days or times of day when crime is less prevalent, regions of the city that might need more or less police support, determine the stores most likely to face problems with shoplifters, among other things. By exploring these trends we get a sense of the current state of affairs with respect to crime in the five boroughs and where civil defense infrastructure might need recalibrating.

This project required big data to uncover the findings we report on, as police activity requires ample memory. With over 5 million entries in less than 10 years of data, this project was made more efficient by the use of big data techniques, and as years progress the memory required for the data will only increase. Without big data architecture we will run into performance issues using something like SQL or Python, whereas Spark and Mapreduce on Hadoop will enable us to scale up without running into performance issues. As data increases, big data infrastructure becomes necessary for efficient data exploration.

3. Dataset description

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2015. There are 5,101,231 incidences of crime included in this dataset.

Column	Base Type	Description
CMPLNT_NUM	INT	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	DATE	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	TIME	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	DATE	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM	TIME	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
RPT_DT	DATE	Date event was reported to police
KY_CD	INT	Three digit offense classification code
OFNS_DESC	TEXT	Description of offense corresponding with key code
PD_CD	INT	Three digit internal classification code (more granular than Key Code)
PD_DESC	TEXT	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD	TEXT	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	TEXT	Level of offense: felony, misdemeanor, violation
JURIS_DESC	TEXT	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	TEXT	The name of the borough in which the incident occurred
ADDR_PCT_CD	INT	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	TEXT	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC	TEXT	Specific description of premises; grocery store, residence, street, etc.

PARKS_NM	TEXT	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT	TEXT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	FLOAT	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	FLOAT	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	FLOAT	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	FLOAT	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

Table 0. Data dictionary.

4. Experimental Setup

All the Hadoop and Pyspark scripts used in both Part I and Part II of this project were ran in NYU Cluster under HDFS (1.4 PB configured. 128MB block size) running in Linux (Centos 6.5). The Hadoop and Pyspark configurations can be found in conf/hadoop_conf.xml and conf/pyspark.txt respectively.
All the code provided is Python 2 and 3 compatible unless particularly stated.

PART I

5. Data Quality Summary

In the following section we will outline our findings regarding valid, invalid and null values in the columns of our dataset and discuss our reasoning in considering values invalid.

5.1. Valid/Invalid/Null Counts Table

Col #	Column Name	Valid count	Invalid Count	NULL Count
0	**CMPLNT_NUM**	5101231 (100%)	0 (0%)	0 (0%)
1	**CMPLNT_FR_DT**	5100538 (99.98%)	38 (<0.01%)	655 (0.012%)
2	**CMPLNT_FR_TM**	5100276 (99.98%)	907 (0.018%)	48 (<0.01%)
3	**CMPLNT_TO_DT**	3709721 (72.72%)	32 (<0.01%)	1391478 (27.27%)
4	**CMPLNT_TO_TM**	3712066 (72.77%)	1380 (0.027%)	1387785 (27.2%)
5	**RPT_DT**	5101231 (100%)	0 (0%)	0 (0%)
6	**KY_CD**	5101231 (100%)	0 (0%)	0 (0%)
7	**OFNS_DESC**	5082391 (99.63%)	0 (0%)	18840 (0.37%)
8	**PD_CD**	5096657 (99.91%)	0 (0%)	4574 (0.09%)
9	**PD_DESC**	5096657 (99.91%)	0 (0%)	4574 (0.089%)
10	**CRM_ATPT_CPTD_CD**	5101224 (>99.99%)	0 (0%)	7 (<0.01%)
11	**LAW_CAT_CD**	5101231 (100%)	0 (0%)	0 (0%)
12	**JURIS_DESC**	5101231 (100%)	0 (0%)	0 (0%)
13	**BORO_NM**	5100768 (>99.99%)	0 (0%)	463 (<0.01%)
14	**ADDR_PCT_CD**	5100841 (>99.99%)	0 (0%)	390 (<0.01%)
15	**LOC_OF_OCCUR_DESC**	3973890 (77.9%)	0 (0%)	1127341 (22.1%)
16	**PREM_TYP_DESC**	5067952 (99.34%)	0 (0%)	33279 (0.65%)
17	**PARKS_NM**	7599 (0.15%)	0 (0%)	5093632 (99.85%)
18	**HADEVELOPT**	253205 (4.96%)	0 (0%)	4848026 (95.04%)
19	**X_COORD_CD**	4913085 (96.31%)	0 (0%)	188146 (3.69%)
20	**Y_COORD_CD**	4913085 (96.31%)	0 (0%)	188146 (3.69%)
21	**Latitude**	4913085 (96.31%)	0 (0%)	188146 (3.69%)
22	**Longitude**	4913085 (96.31%)	0 (0%)	188146 (3.69%)

Table 1. Valid/Invalid/Null count per column in the datset.

A quick visualization of the table above can be seen in Figure 1 below.

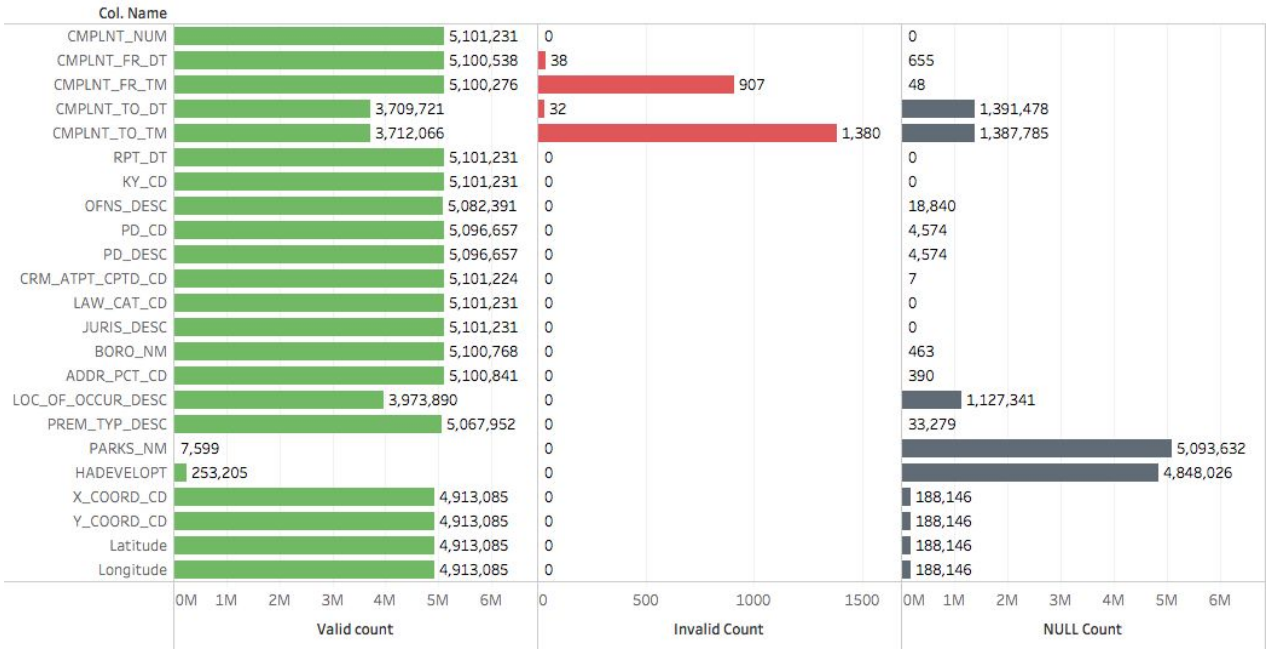

Figure 1: Valid, Invalid, and Null counts for each column in the data

5.2. Valid/Invalid/Null Justification and Reasoning

Validating columns in our dataset depended heavily on the type of data in that column. We will run through the validation logic and our reasoning behind it for each of the columns.

CMPLNT_NUM

This is the primary key of our table and as such is simply a list of integers. We checked that there were no duplicate keys (i.e. that is is a true primary key) using unique_primary_key_check.py and found that every value was in fact unique. We then check if the value is a digit and if it is we consider it valid. It will be invalid if a try: isdigit() fails.

CMPLNT_FR_DT, CMPLNT_TO_DT

For the date fields we first check if it is an empty string and assign a NULL value. If the field is not empty, then we check that it is a valid date between 1900 and 2017. If it meets this condition and the start date is before the end date, we assume it is valid; if not, it will be an invalid field.

CMPLNT_FR_TM, CMPLNT_TO_TM

For the time fields, we first check if the field is empty and assign a NULL value. If it is nonempty and can be casted to a datetime variable, then we check that the start date and time is before the end date and time to conclude the field is valid. Otherwise, if it can not be cast to a datetime variable or end date and time is before start end and time then the field is invalid.

RPT_DT
We assume that the date reported to the police is valid if it can be casted to a datetime variable and the year is between 2006 and 2017.

KY_CD, PD_CD
According to the description, this fields should contain three digit codes, therefore we just check if the fields are three digit integer numbers to conclude it is a valid number. Empty strings are mapped to NULL while anything else to invalid.

OFNS_DESC, PD_DESC, JURIS_DESC, LOC_OF_OCCUR_DESC, PREM_TYP_DESC, PARKS_NM, HADEVELOPT
These fields contains descriptions or names, so we assume the field is valid when it contains at least one character (a through z). If the field contains only spaces or it is an empty string it will be set to NULL, otherwise it is an invalid value.

CRM_ATPT_CPTD_CD
The only valid values for this field are 'ATTEMPTED' or 'COMPLETED' (after normalizing to uppercase). We check if it is one of both to be valid, if it is an empty string it will be NULL and otherwise invalid.

LAW_CAT_CD
The valid values for this column are 'MISDEMEANOR', 'FELONY' and 'VIOLATION'. If one of these is found (after normalizing to uppercase) it is considered a valid value, if it is an empty string it is a NULL value and otherwise invalid.

BORO_NM
This column stores the name of the borough in which the incident occurred. We check if it contains the strings 'BRONX', 'BROOKLYN', 'MANHATTAN', 'QUEENS', 'STATEN ISLAND' (after normalizing to uppercase) to decide if it is valid. If not, we check for empty strings to say that is a NULL value and otherwise we set it to invalid.

ADDR_PCT_CD
This field contains the precinct number in which the incident occurred. To be a valid value it has to be a string representing an integer between 1 and 123 (both included). If it is empty, it is considered NULL, and anything else is considered as an invalid value.

X_COORD_CD, Y_COORD_CD
These fields contain the X and Y coordinates for the New York State Plane Coordinate System. The values are assumed valid if it contains a string that can be casted to float and are within the projected bounds 909126.0155, 110626.2880, 1610215.3590, 424498.0529[1]. If it contains an empty string, then it is considered a NULL value, while anything else is considered invalid.

Latitude, Longitude
To valid, these fields must contain values which can be casted to float and are within the bounds 40<Latitude<42 and -75<Longitude<-73, which make up a bounding box on NYC. Empty strings are considered NULL values while anything else is invalid.

[1] See http://www.spatialreference.org/ref/epsg/2263/ for further reference.

6. Data Quality Findings

Times that are '24:00:00' that cannot be parsed by the datetime library

These times are mapped to INVALID due to how the datetime library is written.

CMPLNT_FR_TM
('NULL', 48)
('24:00:00', 903)
('VALID', 5100280)

CMPLNT_TO_TM
('NULL', 1387785)
('24:00:00', 1376)
('VALID', 3712070)

We investigated the data and saw that there were non-zero numbers of '00:00:00' in the data. This means that '24:00:00' is indeed misleading, and could either refer to midnight of that day or the next day. To check for what was intended, we counted the difference between the start date and the end date for all instances with '24:00:00' as the start time (**inspect24.py**). This could give us insight on whether '24:00:00' means midnight of the current day or of the next day. There were no negative values when we did this calculation, but there were 46 instances where there were 0 days of difference. We inspected these values (**inspect24_0.py**) and found that these were instances in start date time and end date time were exactly the same. This suggests means that these '24:00:00' should actually be '00:00:00' of the next day.

Categories that needed to be merged

We wrote a script to identify the counts of all the unique values in a column (**countuniques.py**). With the output of this, we could manually examine whether certain categories should be combined. For the column OFNS_DESC, we found that these categories should be merged into one category:

('ADMINISTRATIVE CODE', 11383)
('ADMINISTRATIVE CODES', 18)

('INTOXICATED/IMPAIRED DRIVING', 48)
('INTOXICATED & IMPAIRED DRIVING', 73730)

('KIDNAPPING', 2)
('KIDNAPPING AND RELATED OFFENSES', 2)
('KIDNAPPING & RELATED OFFENSES', 2300)

('OTHER STATE LAWS (NON PENAL LA', 5505)
('OTHER STATE LAWS (NON PENAL LAW)', 4)

We found that in LOC_OF_OCCUR_DESC (column 15) there were values with ' ' that should be the same as our standard NULL, ''.

('', 1127128)
(' ', 213)

Other columns did not seem to have any categories that should be merged.

Suspicious years

We noticed that there were a few years in our dataset that were mapped to invalid because they didn't fall within the proper range. This is because there were dates such as 1015, which are probably typos and are meant to be 2015.

Spikes of crimes at certain dates

Another interesting data quality issue is around dates on which crimes are reported (**daily_crime_counts.py**):

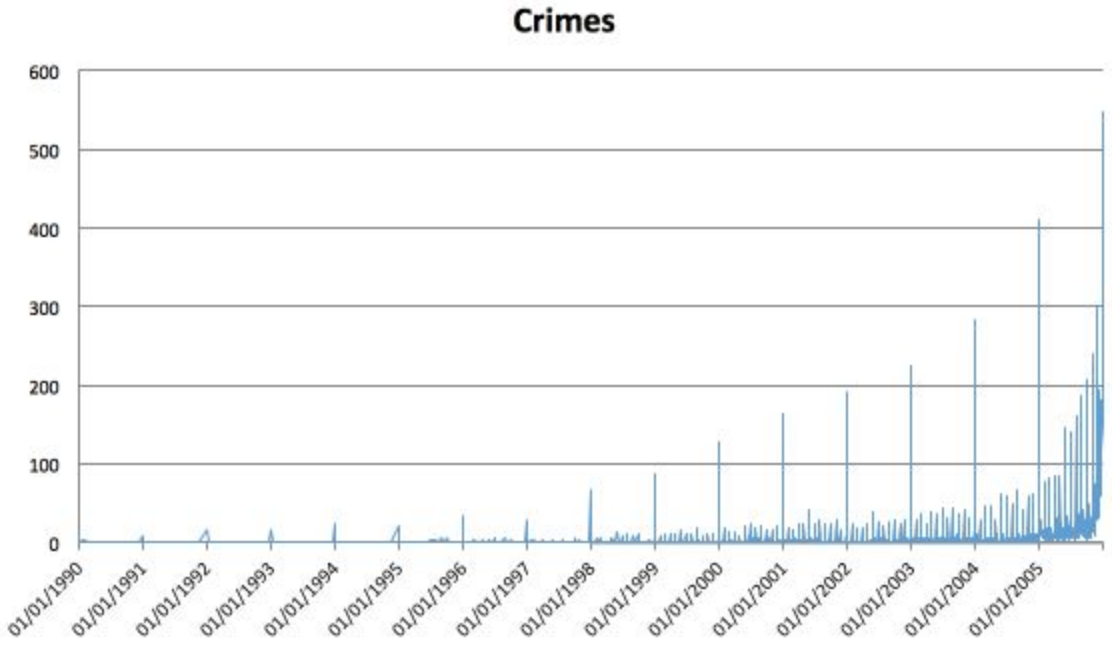

Figure 2: January 1 spikes between 1990 and 2006

We see very noticeable (and non-trivial) spikes on January 1 of every year pre-2006; we hypothesize that this is likely due to the fact that the NYPD included data that was reported between 2006 and 2015 and that if someone reports a crime years after it happened, they might only remember it happened in that year. If someone only remembers a year of a crime it is likely the NYPD would just code it as happening on Jan 1 of that year. We see smaller but still recognizable spikes between the January spikes and this is likely due to someone only remembering the month during which a crime occurred and a similar circumstance around recording dates may have persisted. We do not believe this only implies a ton of crime occurs on New Year's Eve/Day, as this trend does not continue (to this degree) in the data beginning January 1, 2006; While there is a high volume of crime on January 1, it is not nearly as pronounced as we would expect it to be if we were to extrapolate the above pattern.

Suspicious number of counts in precinct 121

We note that there is a increase from a mean of around four crimes per year in precinct 121 before 2013 while the mean of reported crimes grows to 5792 for the period 2013-2016. The data for the first period will be considered as corrupted. Figure 3 (generated in Price-Crimes-Plots.ipynb) shows this behavior. Note the log scale in the y-axes.

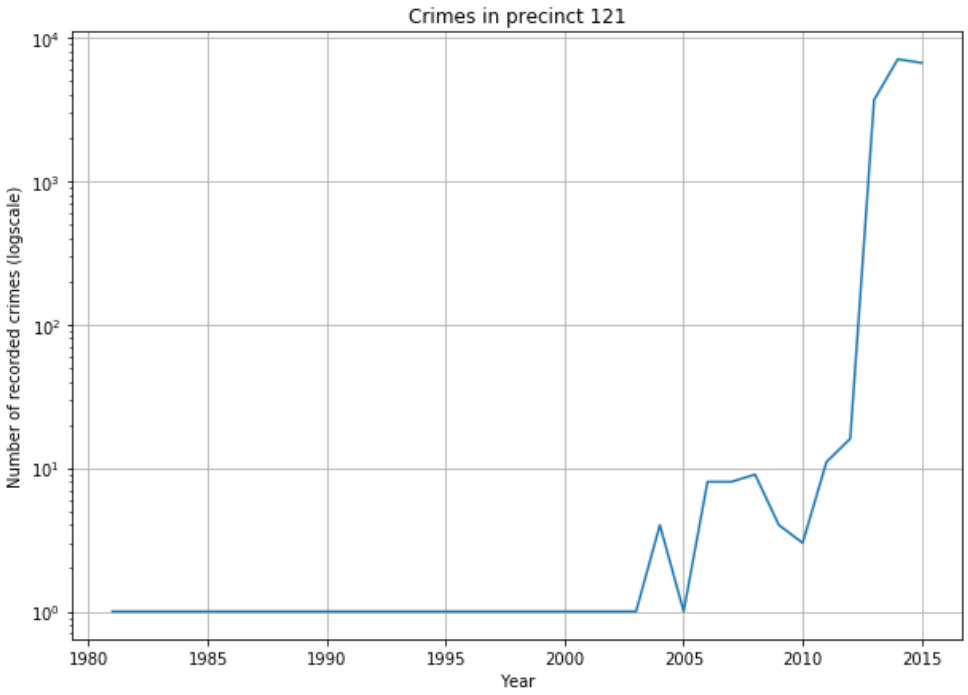

Figure 3. Yearly number of crimes in precinct 12. High increase noticed in the count. Y-axis in log scale.

7. Data Cleaning

Based on the data quality issues that we encountered in the previous sections, we created a data cleaning script using PySpark (**clean.py**) to clean the data prior to doing analysis. This script goes through each row and performs the following tasks sequentially for each row in the original dataset.

Removes the header

If the header with the column names is the first line of the file, then that row will be filtered out

Modifies all times with '24:00:00' as the value

If CMPLT_FR_TM or CMPLT_TO_TM are '24:00:00':

- Replace the value for that field to be '00:00:00'
- Add one day to the date to the corresponding CMPLT_FR_DT or CMPLT_TO_DT field

Modifies all dates with y0xx as the year, where y is not 2

If the year has 0 at the second position, but the first position is not a 2:

- Replace the value for that field to be 20xx instead

Merge specified categories for OFNS_DESC

A dictionary is defined for the keys that need to be mapped to another category.

If the value is contained within the dictionary:

- Replace the value with the dictionary lookup

Merges the ‘ ‘ into ‘’ for LOC_OF_OCCUR_DESC
If the value is ‘ ‘:

- Change value to ‘’ (NULL)

Replaces all invalid times with null
Attempts to combine CMPLT_FR_DT and CMPLT_FR_TM into a start datetime
Attempts to combine CMPLT_TO_DT and CMPLT_TO_TM into an end datetime

If both were successfully combined into a datetime:

- If the end datetime is after or equal to the start datetime:
    - Keep the values the same

Else if a start datetime exists but an end datetime does not exist:

- Return the start time as itself, but the end time with ‘’ (NULL)

Else if an end datetime exists but a start datetime does not exist:

- Return the end time as itself, but the start time with ‘’ (NULL)

Else:

- Replace both the start and end time with ‘’ (NULL)

Replaces all invalid dates with null
Attempts to convert CMPLT_FR_DT into a date
Attempts to convert CMPLT_TO_DT into a date

If both were successfully converted into dates:

- If the end date is after or equal to the start date:
    - Keep the values the same

Else if a start date exists but an end date does not exist:

- Return the start date as itself, but the end date with ‘’ (NULL)

Else if an end date exists but a start date does not exist:

- Return the end date as itself, but the start date with ‘’ (NULL)

Else:

- Replace both the start and end date with ‘’ (NULL)

8. Data Exploration

8.1 Seasonality of Crime

When we look at crime over time (**daily_crime_counts.py**) we see that, generally, fewer crimes are committed during the winter and more crimes happen during the summer:

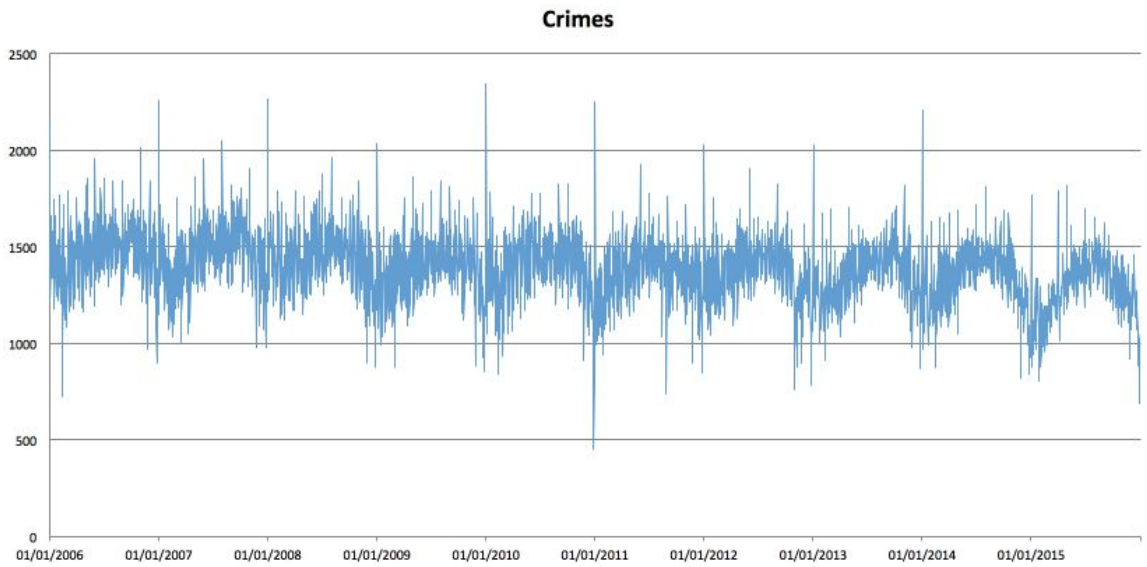

Figure 3: Crime over time (1/1/2006-12/31/2015)

In order to confirm this we also looked at averages by month and day (across all years beginning with 2006, **avg_day_crime.py**) and found that, while less pronounced, the trend is definitely apparent. Our hypothesis is that when it is cold outside or the weather is poor, people are less likely to want to go out and commit crimes. People tend to be outside more during the summer, both criminals and potential victims; It is particularly hard to find a victim to rob when there is nobody outside.

Figure 4: Average crime count by day of year

We see another visualization of the data (**daily_crime_counts.py**) in Figure 6 on the next page (light yellow corresponds to low count, dark purple corresponds to high count and everything in between is a gradient of low to high count). This visualization would make it very easy to view if there were weekday trends in the data by showing banding across the same weekday, though we see little evidence of that. It also highlights days or ranges of days when there was an unusually low or high count of crimes relative to the days around it or the prior year same day. For example, we see a rather distinct light band in late October, early November of 2012 which we know was around the time of Hurricane Sandy. It's not too large of a leap to infer that a major storm with flooding like Sandy either makes it harder for crime to happen or to be caught (either criminals are staying indoors, police are, or both).

There is also a distinct light box at the end of November every year, which corresponds to Thanksgiving. This follows from the same theory that more crime occurs when people are out and about. Even criminals can't escape Thanksgiving turkey with the family. January 1st remains the day with the highest counts of crime by a large margin, as it is the darkest colored box for every single year. This may be due to how the police codes complaints when the person doing the reporting only knows the year and not the exact date (as previously discussed), or that New Year's Day has an unusual amount of crime with to lots of people being outside. Looking at the lightest color, we see December 27th, 2010 has the lowest crime count of all the days in our range. There was a massive blizzard in New York in that time, with up to 2 feet of snow in places, resulting in a very low rate of crime.

In 2015, we see two very dark colors around March-May, the two days corresponding to these dark colors are 4/1/2015 and 5/1/2015. May 1st was the May Day parade in Manhattan to support worker's rights as well as a protest against police brutality as it coincided with the death of Freddie Gray in police custody just a couple weeks prior. Looking at the percent of crimes in Manhattan on 5/1 versus 5/2 and 5/3 we see a significant increase in the percent of crimes happening in Manhattan on May 1st (most of which being classified as criminal mischief or harassment), implying it isn't unreasonable to think the parade resulted in several more arrests than on the average day in Manhattan (and there appears to be some evidence that this is true if we trust Newsweek [15]). April 1st is a strange day as there doesn't appear to be any significant events in New York that day, and most of the crimes are criminal mischief; which is just ambiguous enough to make it challenging to understand why this might be. The city was in deep turmoil regarding police brutality between 2014 and 2015, so it is entirely possible there was a Black Lives Matter event or protest that resulted in high arrest counts this day as well.

As in Figure 5 we see a spike of crimes in the first day of each month, suggesting that, as hypothesized before, when the exact date is not available, the crime gets mapped to the first day of the month.

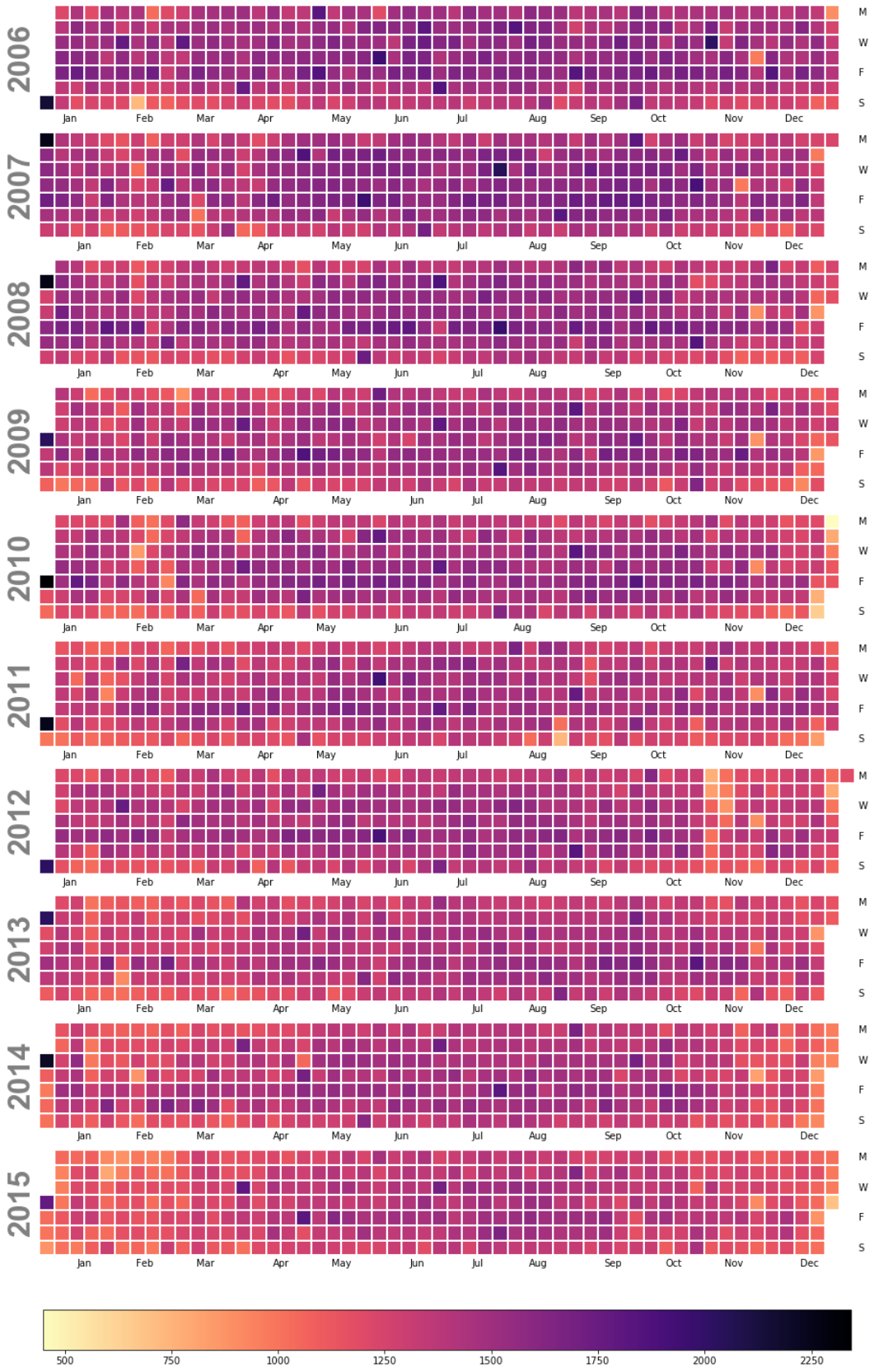

Figure 5: Crime over time gradients (1/1/2006-12/31/2015)

8.2 Violent Crimes

During the election this year, crime history and trends were highly investigated and we decided we would do that as well (obviously in a micro-environment as we don't have all US data). Many politicians were discussing the apparent rise in the 'violent crime' rate, though using our data (**daily_violent_crime.py**) we find that since 2006, violent crimes have remained relatively steady year over year (a trendline is shown on the graph with a slope of 0.0006, i.e. very low slope implying from day to day there is a trivial change in crime) with some fluctuations. Violent crimes are defined as robbery, rape, murder and assault by the FBI. We find that there has not been a serious rise in violent crime over the past 8 years (there is some growth, though given we are looking at absolute counts, the 'rate' is most likely stable given the population of New York is constantly growing).

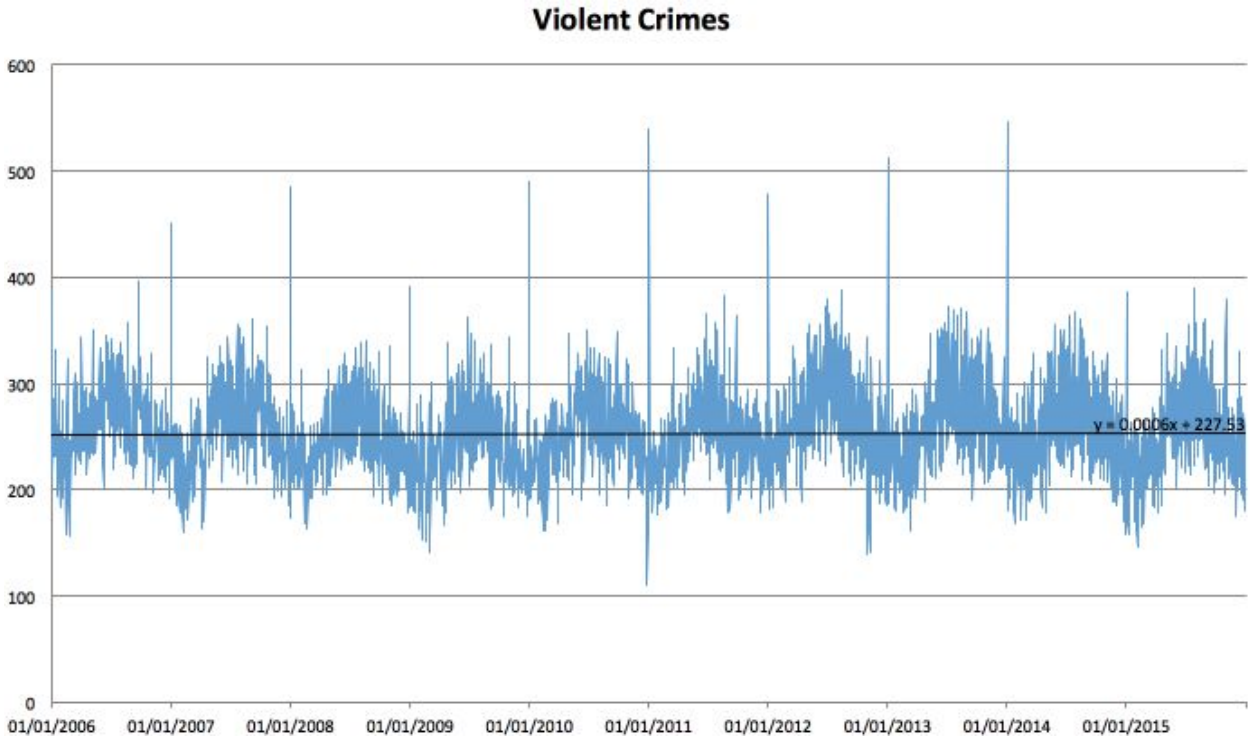

Figure 6: Violent crime over time (1/1/2006-12/31/2015)

8.3 Crime by Time of Day

When we look at crime by the hour of the day (**time_of_day_crime_rates.py**, **time_of_day_top_10.py**, **time_of_day_violent.py**), we find interesting, though unsurprising facts. Overall, and for the top 10 most common crimes we find that crime is very low in the early hours of the morning and picks up around 8am until about 8pm which also confirms the idea that crime happens (or is caught) when people are out and about. Another interesting finding is that violent crimes follow a different distribution than the others; Violent crimes tend to happen more in the early hours of the morning and late at night and less during the daylight hours. This makes sense as it seems violent crime tends to happen more when people are walking alone at night or people are intoxicated and these are generally more common during the nighttime hours.

Figure 7: Violent crime over time (1/1/2006-12/31/2015)

8.4 Crime by Day of Week

We were also interested in looking at the trend of crime occurrence based on the day of the week (**dayofweek.py**).

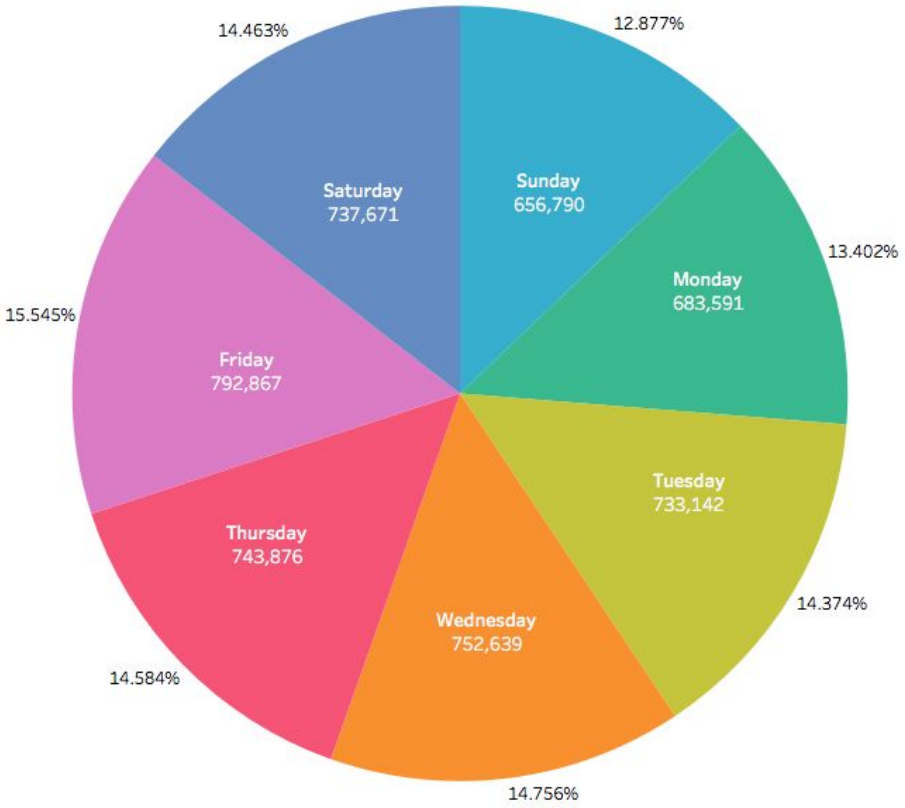

Figure 8: Crime occurrences by day of week

Similar to our observations on time of year, it looks like crime happens the most on days that more people are out and about, with Friday having the most occurrences of crime.

We also calculated the weekday and weekend average number of crimes for this data set (**weekday.py**). The average for weekdays was 741,223 while the average for weekends was 697,230.5. The weekday average is higher, though this is not that surprising after looking at this graph. This is probably because of the lower rate of crime on Sundays.

8.5 Forecasting Crime Through 2020

In addition to the violent crime rate, overall crime rates have been discussed and issue of a significant rise in crime year over year has been investigated. By looking at our 8 years of New York City data on a monthly level (**monthly_crime.py**) and running a seasonal ARIMA time series forecasting model (model can be found at SARIMA_Time_Series_Forecast_Crime.pdf) we found that there is very little evidence of systemic rise in crime and in fact the number appears relatively stable. Since we are looking at absolute crime counts, the stability in the model implies that rates would actually be flat or even decrease as the population grows. That being said, forecast models are only as good as the data they are fed and historical counts of crime aren not necessarily the best predictor of a rise in crime as these types of things can be brought on by national events that are unpredictable in nature.

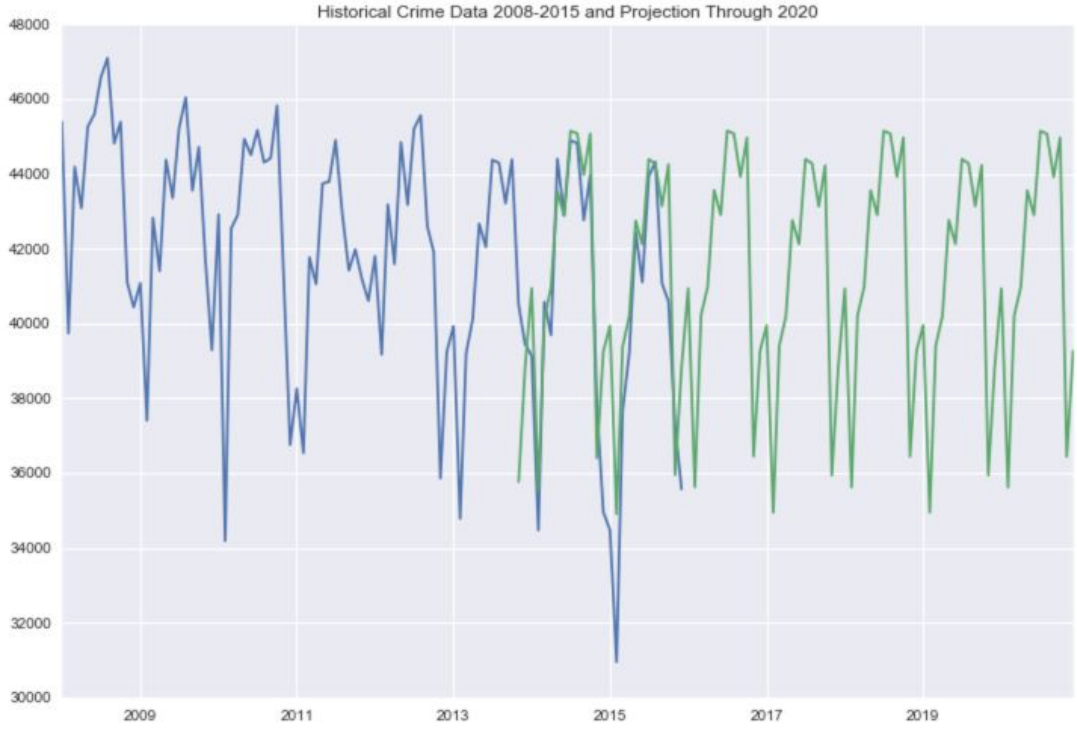

Figure 9: Crime history and projection (2008-2020)

8.6 Drinking and Driving

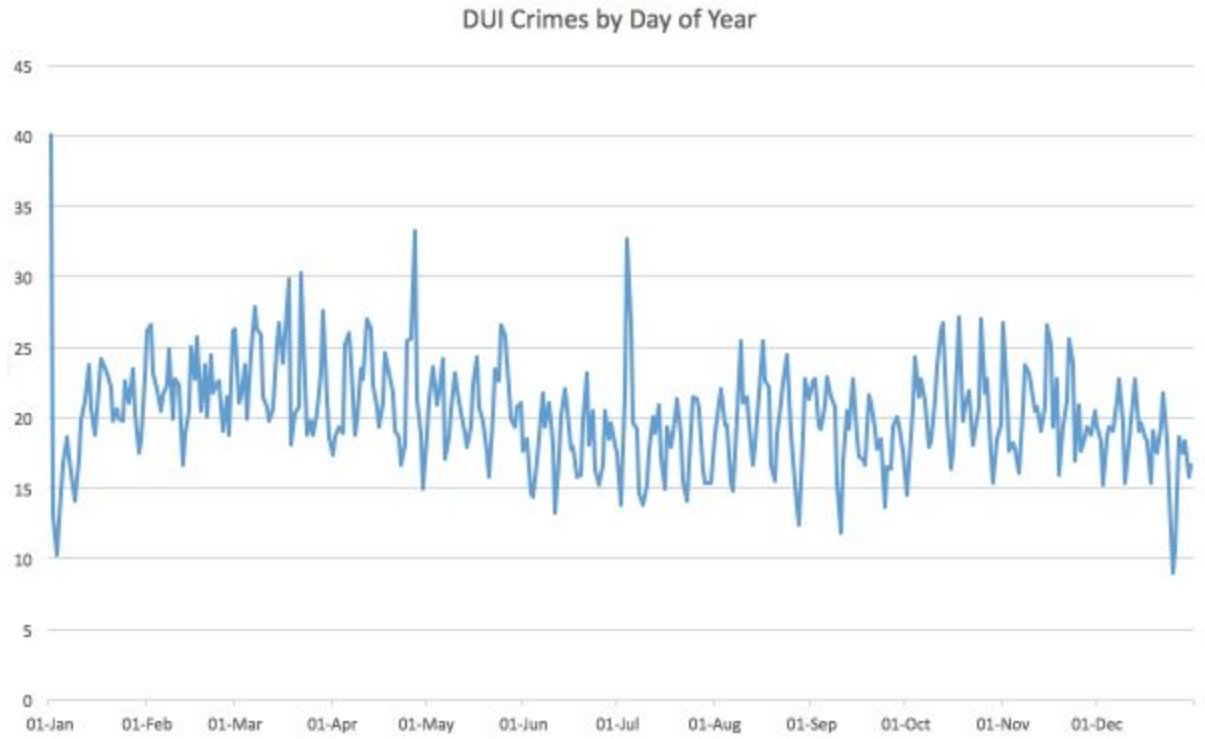

Figure 10: DUI crime over time (1/1/2006-12/31/2015)

Another interesting trend we wanted to look into was if certain days of the year see a higher average count of intoxicated driving counts. In the above plot looking at the average number of DUI crimes reported by day of year (between 2006 and 2015, **avg_day_DUI_crime.py**) we see three distinct peaks: January 1st, July 4th and April 27th. Two of these three make a lot of sense as New Year's Eve many people stay out late drinking and might leave parties while still impaired (in addition to our strange data finding around January 1 data, though I feel impaired driving is generally reported with very little delay as we will explore later), and the Fourth of July is another holiday centered around drinking. April 27th did not make much sense on first glance. Looking closer at the data we saw one day, April 27th 2013, where there were 147 impaired driving offenses. Seeing as this is a huge outlier, it skews the average up giving us that large bump. Doing a quick Google search, the only significant event in the New York area on that day was that it was the first day of TechCrunch's Disrupt 2013 conference, but that seems more coincidental and more research could be done into why this spike happened. Also of note, the two days in the year with the lowest number of DUI offenses on average are January 2nd and December 25th; December 25th is Christmas, so it is possible that people are indoors spending time with their family, include police officers, and as such there may just be less of a police presence on the streets as well as overall driver presence on the streets. On January 2nd, many people are still recovering from terrible New Year's Day hangovers and would probably vomit at the sight of a Bud Light. In addition, from the prior day peaks, police are probably bogged down with paperwork or see it unnecessary to patrol for impaired driver offenses as there was a huge influx the day before.

8.7 Observing Shoplifting Trends

We wanted to explore which kinds of stores were affected by shoplifting the most. We wrote a script (**shoplifting.py**) to count the instances of shoplifting by store type. Instances of shoplifting were further

separated into degree of theft: Petit Larceny, Grand Larceny, and Robbery (began as shoplifting). However, there were only 3 instances of Robbery (began as shoplifting), so it ended up being filtered out of the following chart. Stores types with less than 100 counts of shoplifting were also excluded for the chart.

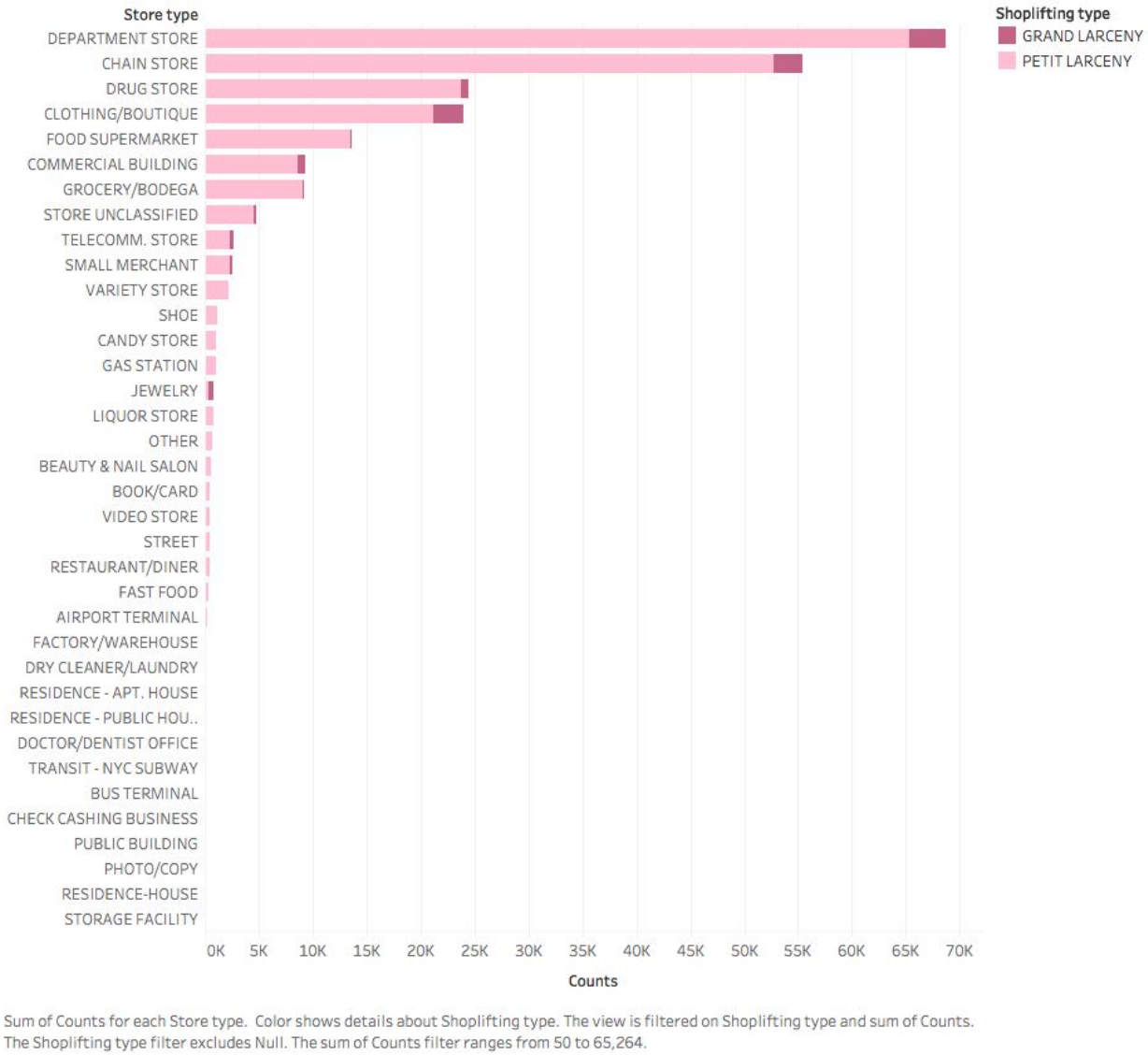

Figure 11: Counts of shoplifting by store type and degree of theft

Places with the most counts of shoplifting are large, common stores such as department stores, chain stores, drug stores, clothing stores, and supermarkets. This makes sense because people may think it is easier to get away with shoplifting, because there are not enough employees to watch them all the time. Also these places have common everyday items that people may need or want.

To further understand which stores were most affected by grand larceny, another visualization was created to show the percentage of petit larceny occurrences from the total of reported crimes for the store type.

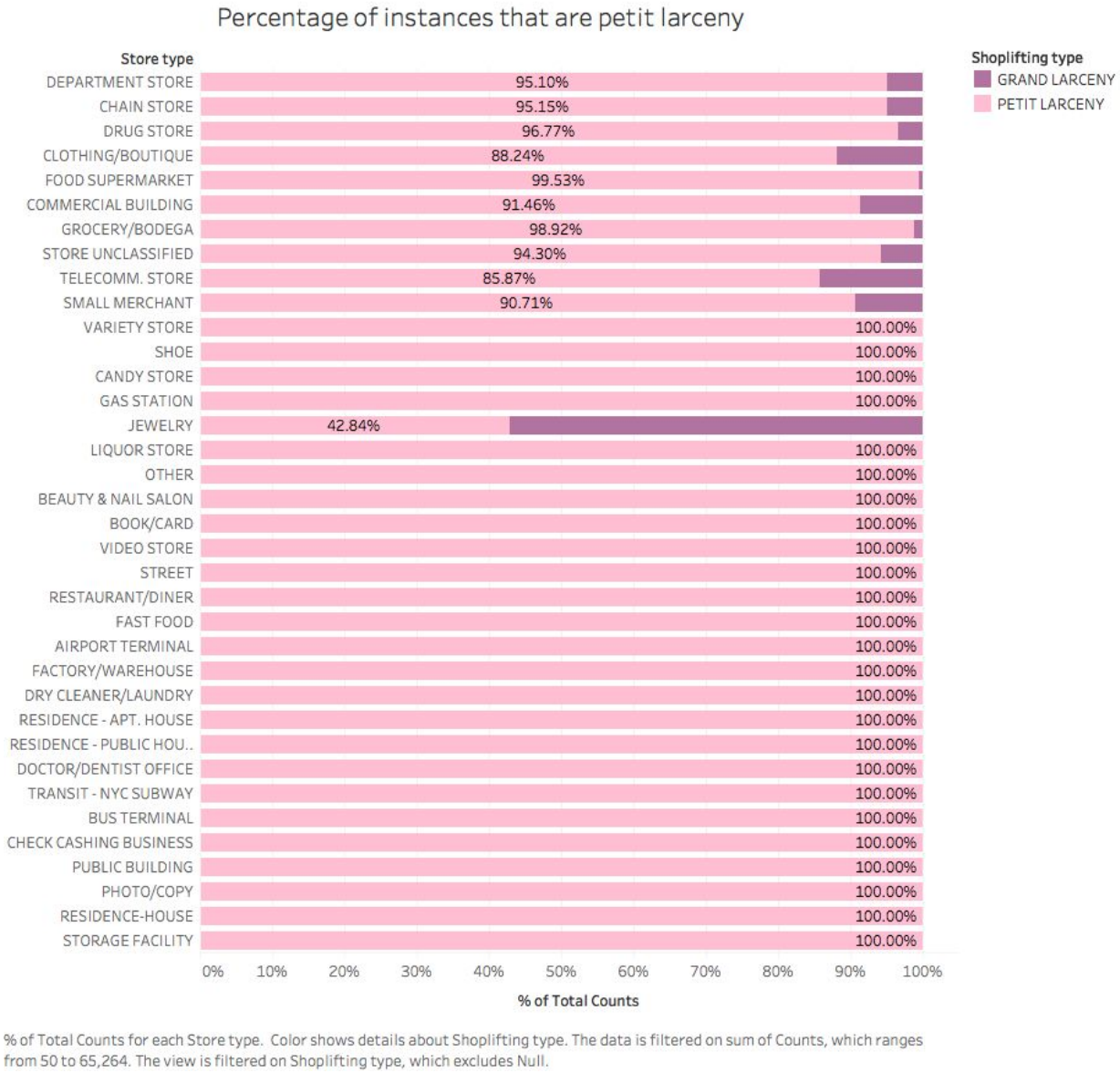

Figure 12: Percentage of petit larceny (shoplifting) by store type

Jewelry stores have the highest percentage of grand larceny, in fact, there are more grand larceny instances than petit larceny instances for these stores. Their products are certainly the most expensive on average compared to all the other stores in this list. The next highest is telecommunication stores, which stock expensive electronics like smartphones and tablets. It's interesting that the third most is clothing/boutique stores. Designer bags and clothes can get quite pricy!

8.8 Difference Between Occurrence Date and Police Report Date

We were interested in seeing the delay between when a crime occurred and when it was reported to police by crime type (**delay.py**). The delay was only calculated for rows that CMPLT_FR_DT and RPT_DT exist and are

valid. Because there was no report time column, the delay was only calculated in number of days. This was averaged across all crimes of the type, and the standard deviation was also calculated for good measure.

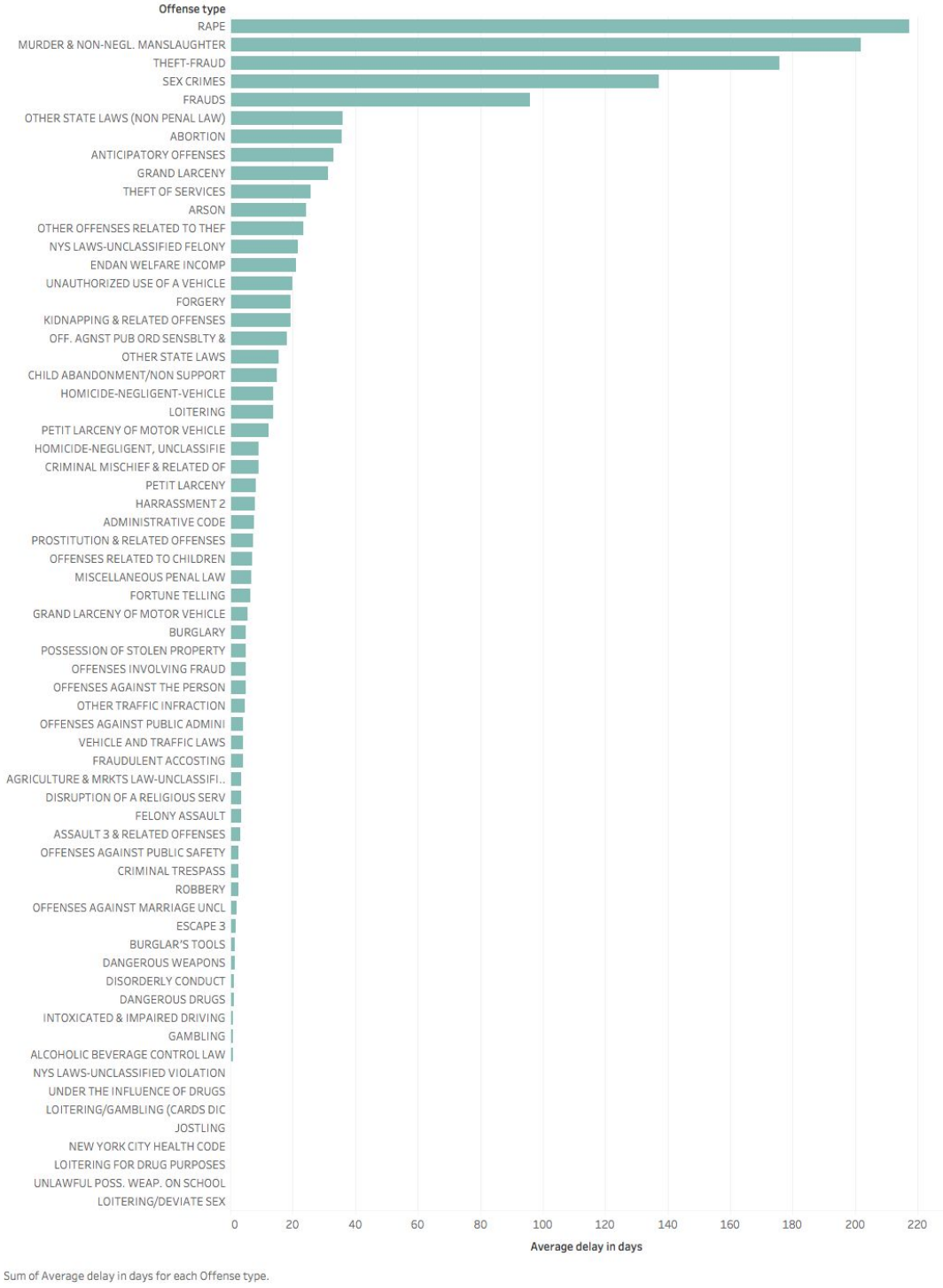

Figure 13: Average delay in days between the crime occurrence and report date.

Sadly, but understandably, rape has the longest average delay in days between when it occurred and when it gets reported to the police. In the same vein, sex crimes has the 4th longest average delay. Other crimes such

as murder, fraud, and abortion have a long delay because it takes a long time for other people to find out that these things have happened before they report it to the police. In the footnotes for the dataset, it also says that "If an incident eventually results in a victim's death, the incident is upgraded to a murder and the report date is recorded as the date of the victim's death, rather than the original report date of the incident." However the average delay still seems to be high for murder cases, so maybe whoever is supposed to be updating this dataset may not be doing the best job following this rule.

The instances with the shortest delay times are crimes in which people usually witness and report right away, such as loitering or unlawful possession of a weapon on school. Or, they are crimes that the perpetrators often get caught directly in the act by the police, such as "intoxicated & impaired driving", "disorderly conduct", "alcoholic beverage control" or "under the influence of drugs".

However, from the standard deviation calculated in the script, it seems that the standard deviation is huge for most of these crimes. It is likely that the data is non-normal and has long-tails. So the mean may not be a good measure. We also ran another script to calculate the median (**delay_median.py**). But actually the median for almost every category turned out to be 0 days. That means that more than half of crimes are reported within the same day. So this seems like an even worse measure. We decided to run a script to look at the distribution for one offense type (**delay_dist.py** RAPE).

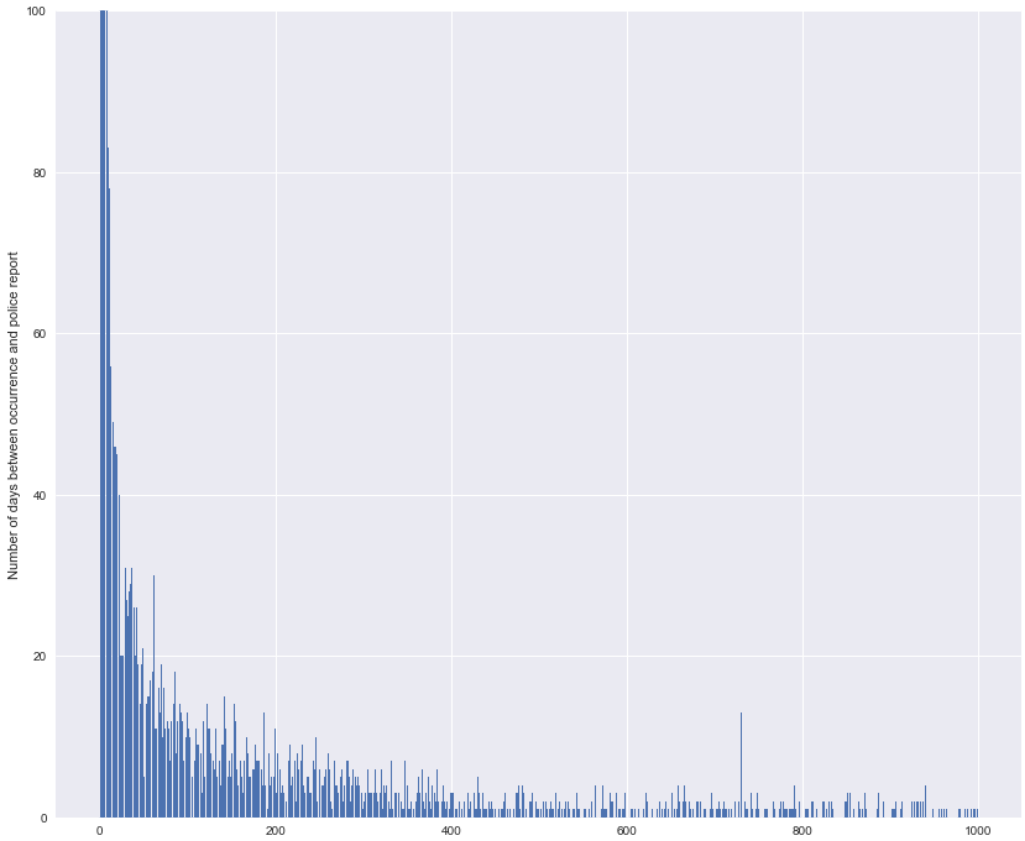

Figure 14: Counts of delay in days for offense type 'RAPE'

Indeed, the distribution is long-tailed and seems to be more like the geometric distribution. The axes have been shrunk to improve readability, it actually extends to one instance that was reported 14,708 days later. This explains the huge mean and standard deviation. But something interesting shows up, it seems that there

is a spike at 730 days, which is exactly 365 x 2 days, or two years later. Maybe some people don't remember the exact date it occurred so they just approximate it by saying it happened exactly two years ago. Is this is a trend with the other crimes too? We wrote another script to look at this (**delay730.py**).

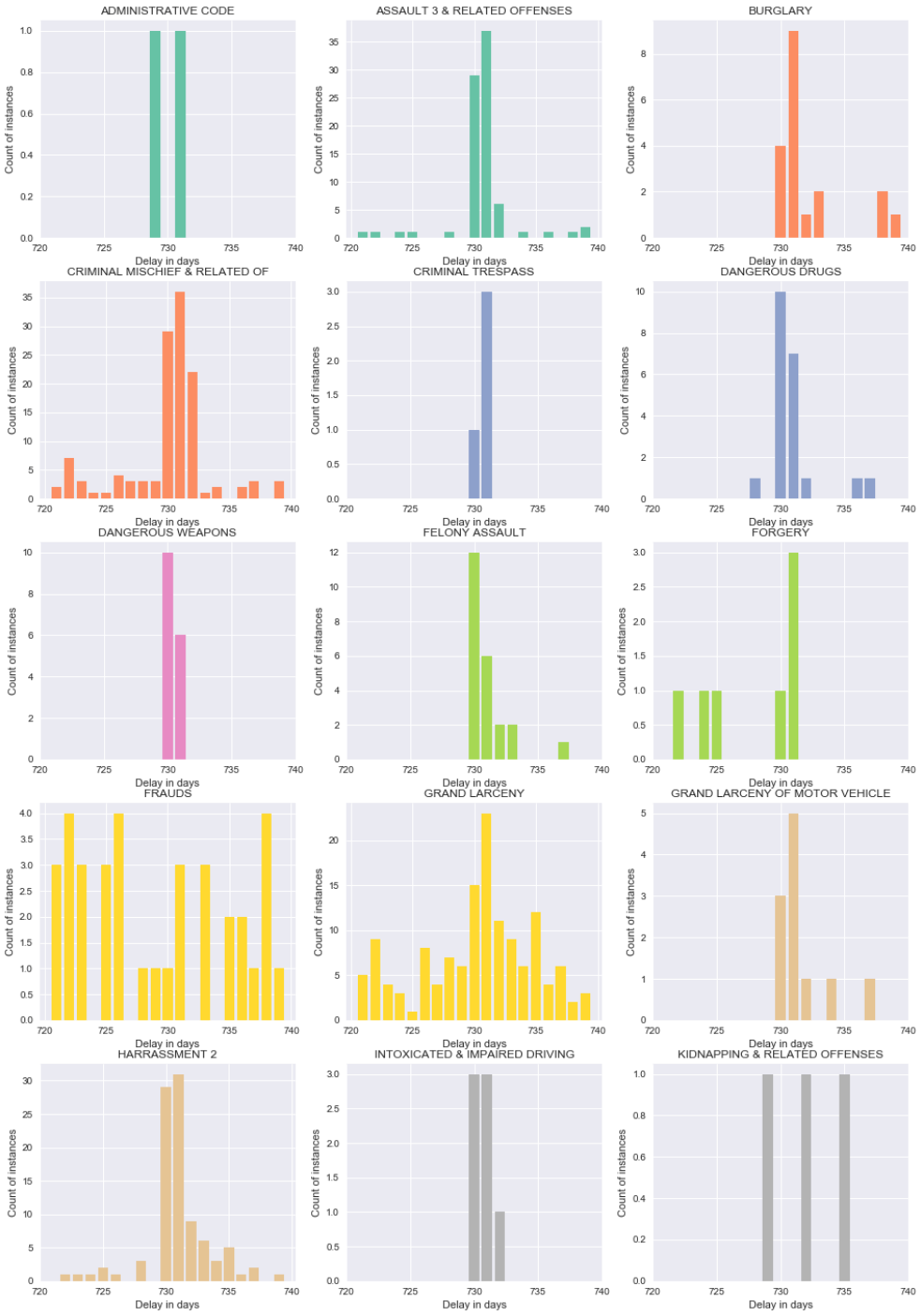

Figure 15.1: Counts of delay in days between 720 and 740 days, part 1

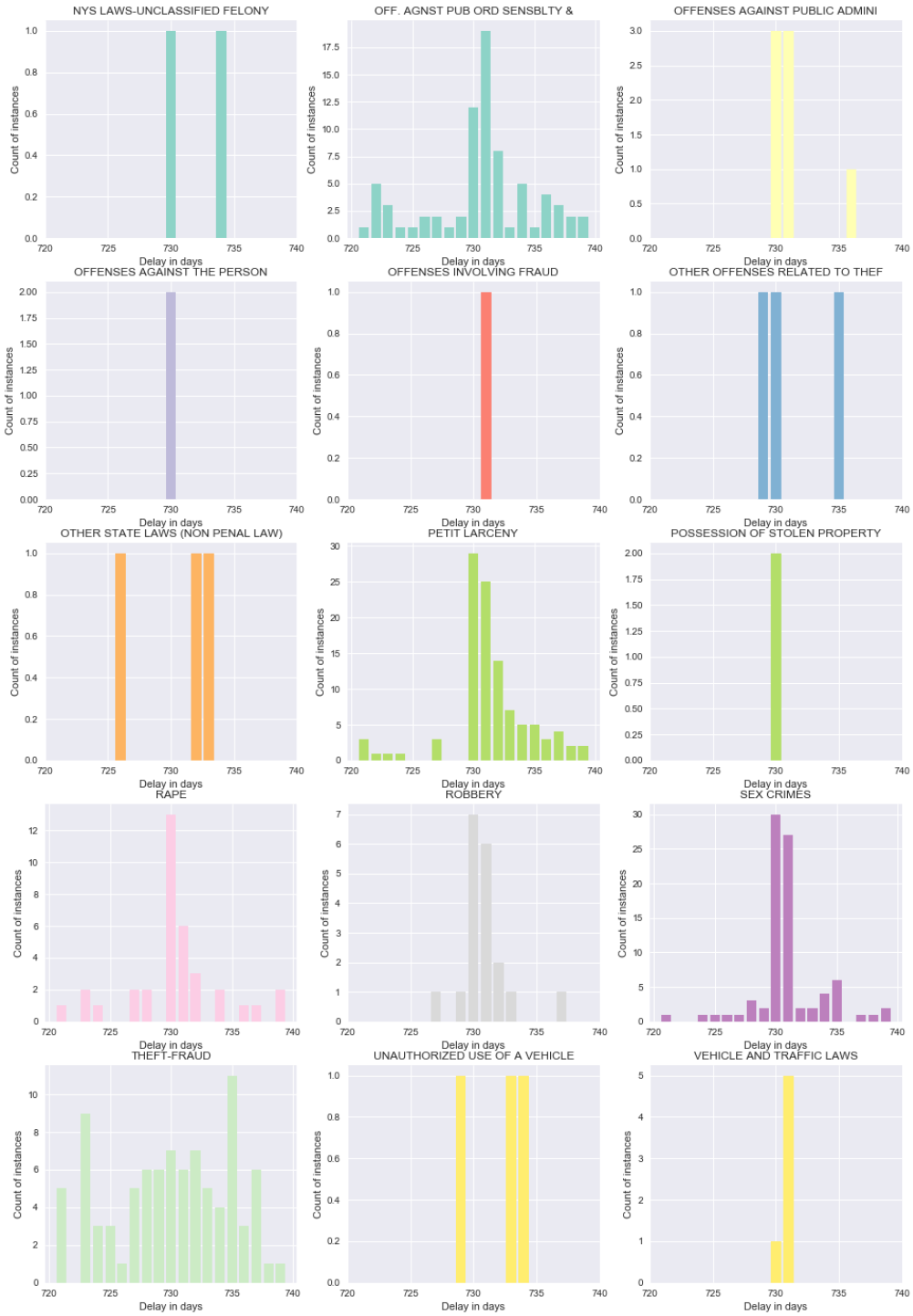

Figure 15.2: Counts of delay in days between 720 and 740 days, part 2

Either people are more likely to report these crimes exactly 2 years later, or they don't remember well so they just give out the date 2 years ago. The second option seems more likely. Leap years may explain the small spike at 731 as well. These seem to be related to crimes that require people to recall from memory when the incident occurred. For crimes like fraud that there may be a better paper trail for when it occurred, this does not seem to be the case.

8.9 Pre 2000 Crime Profile

In a similar vein, we also looked into the type of crimes that were most prevalent in our data pre 2000 (what type of crimes took many years to report, **crimes_pre_2000.py**, **countuniques.py**) and compared percents between the crime types in pre-2000 data and the overall data. We found that the pre-2000 data is much more likely to involve sex crimes, fraud, rape and murder; this is relatively consistent with what we saw earlier looking into the average delay between the offense and the day it was reported.

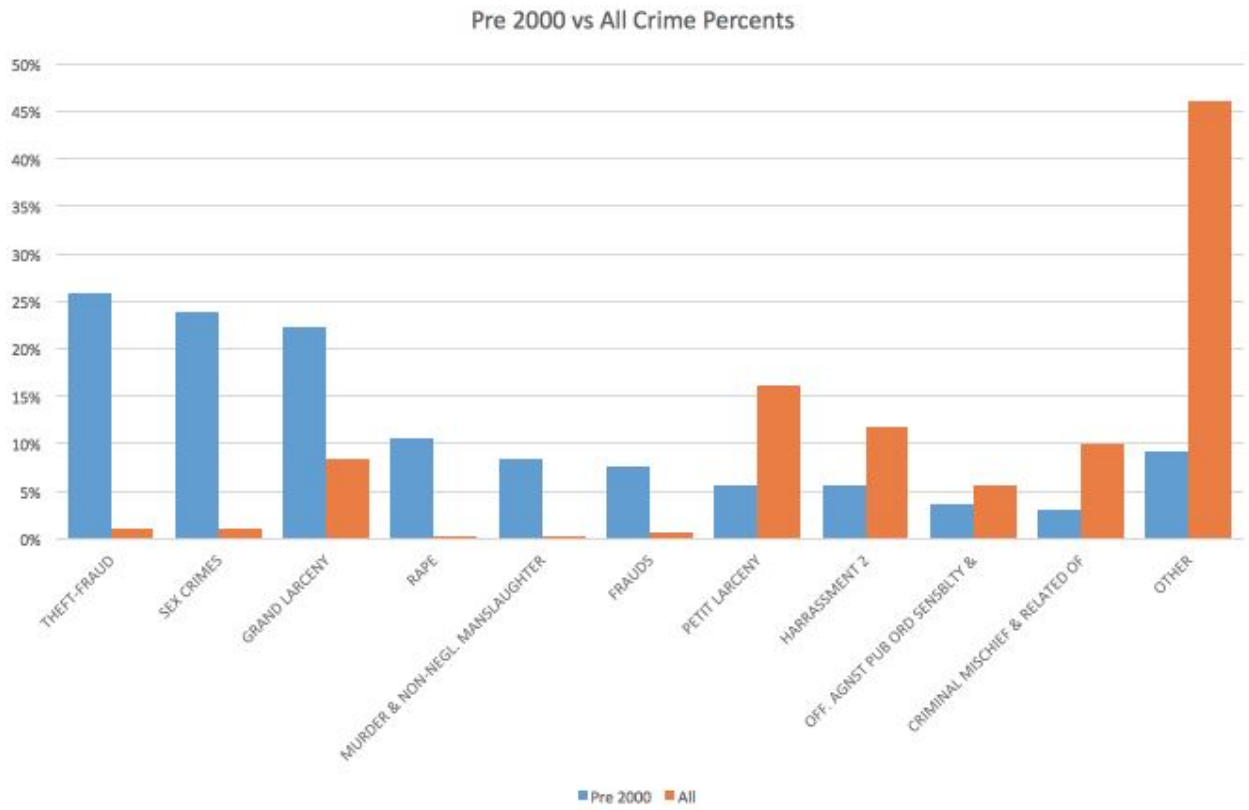

Figure 16: Distribution of crimes by type before 2000 and all time

8.10 Level of Offense by NYC Borough

We were interested in exploring the distribution of level of offense by each of the 5 NYC boroughs (**agg2cols.py** 13 11).

Figure 17: Distribution of crimes by NYC borough

Queens has the highest percentage of crimes in the area that are felonies. This may because it has more highly dense low-income neighborhoods. Brooklyn is second place - this may be because it has undergone some gentrification over the years. Bronx has the highest percentage of crimes that are misdemeanors, perhaps this is related to the age distribution in the area.

8.11 Type of Offense by NYC Borough

We were interested in what types of offenses were being committed the most in each of the NYC boroughs (**agg2cols.py** 13 7). This was done only using offenses that had more than 100 counts of instances. The following chart shows the percentage distribution for each of the boroughs.

Figure 18: Percentage of crimes by offense type by NYC borough

The majority of theft of services occur in Manhattan. This is probably because most service shops are in Manhattan. Manhattan also has the majority in crimes that involve large crowds (fraudulent accosting, jostling, offenses involving fraud). Staten Island has an unsurprisingly small percentage of all these crimes, likely because the population density is not as high there.

It looks like Brooklyn and Bronx are the hotspots for dangerous drugs. All of the child abandonment crimes also occur in Brooklyn and Bronx. These are probably correlated to the low-income neighborhoods in these

boroughs. Not only that, Brooklyn also has the highest percentage of assault, robbery, burglary, felony assault, arson, murder, and kidnapping. You should be extra alert when walking around Brooklyn at night!

Queens has the majority of the offenses related to children, so perhaps Child Protective Services should keep a sharper eye out on families in Queens.

It will be interesting to calculate the number of these offenses per capita for Part II of this project.

8.12 Number of Crimes by Precinct

Below we observe the geographical distribution on the total number of crimes divided by precincts. Figure 20 shows all the crimes while Figure 21 only focuses in violent crimes. Lighter colors on the map represent less crimes while darker colors indicate a larger number of crimes reported in that precinct. The data needed to generate these maps was generated with the scripts (**byprecinct.py** and **byprecinct_violent.py**) for figures 20 and 21 respectively.

By looking at the maps we can identify certain parts of Brooklyn, the Bronx, north Staten Island and the Bryant Park zone of Manhattan as the ones with highest number of crimes.

If we only focus on violent crimes, then Manhattan in general has a low count while certain parts of Brooklyn and the Bronx have the most occurrences.

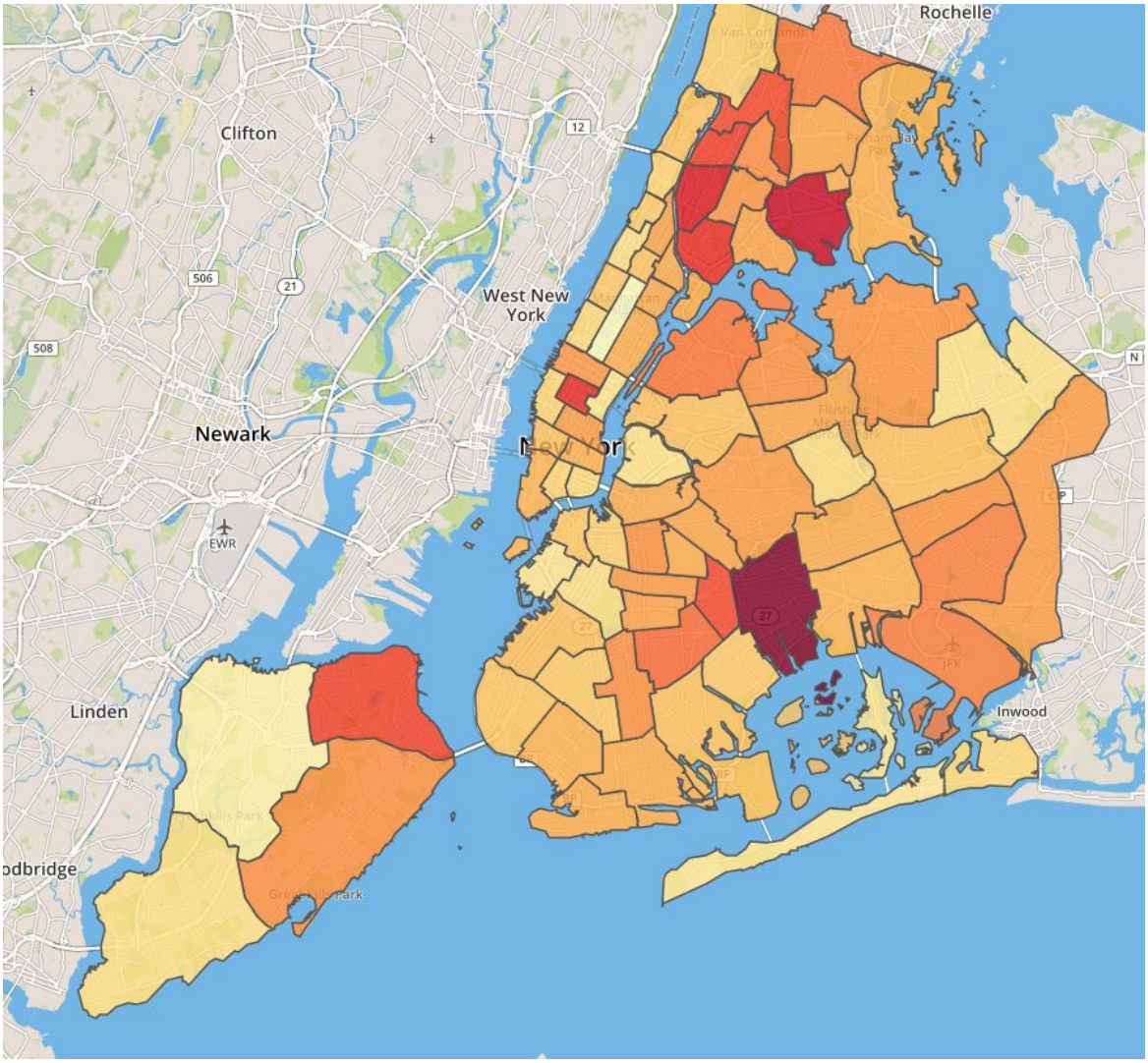

Figure 19: Map of total number of crimes by precinct.

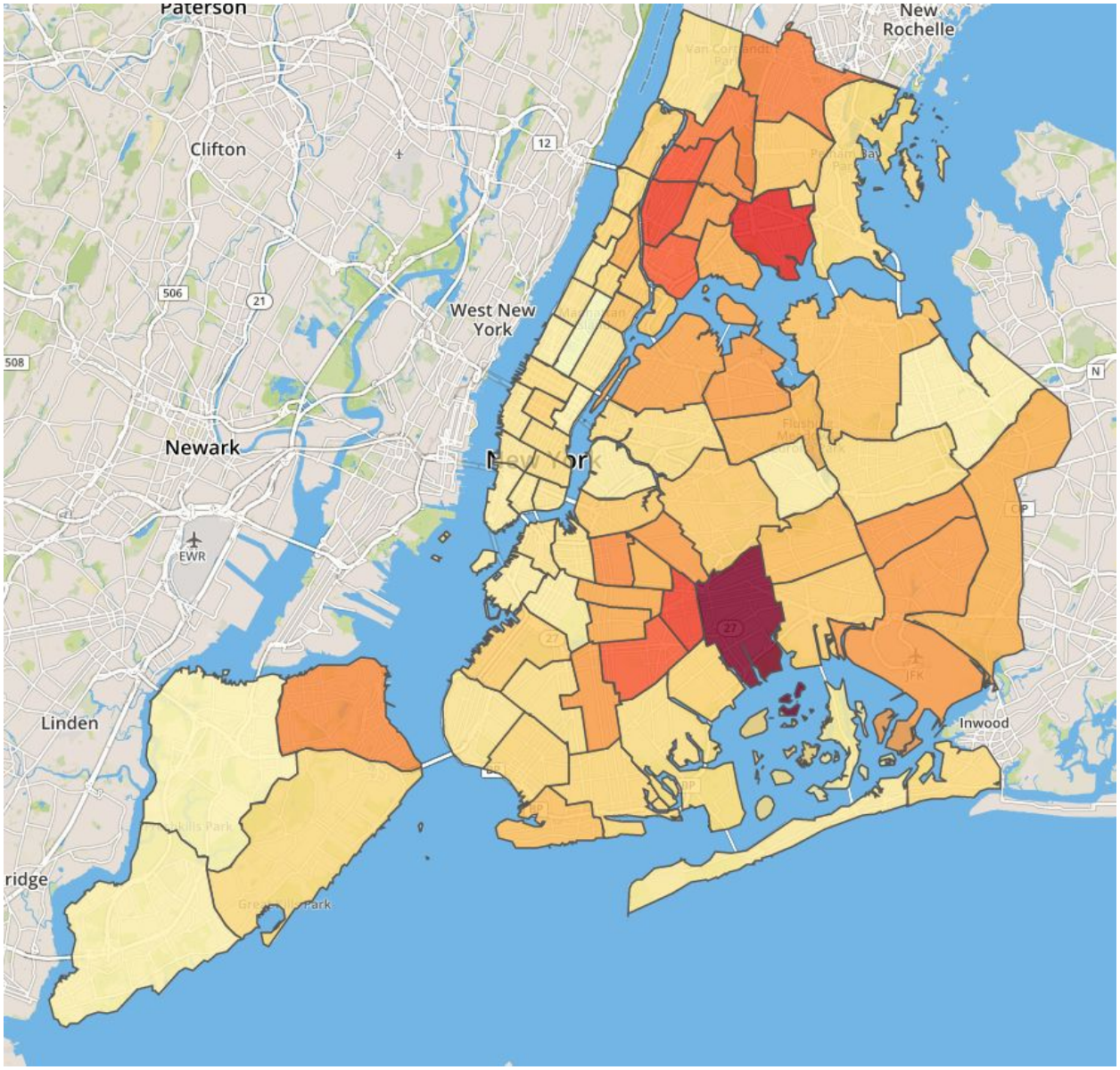

Figure 20: Map of total number of violent crimes by precinct.

PART II

9. Results and Discussion

All the results and analysis performed in this section can be reproduced with the tools provided in the GitHub repository of the project (https://github.com/florielle/bigdataproject). Accompanying our hypothesis, findings and visualizations, the necessary files or scripts for reproducibility are highlighted. The files related with visualizations, unless stated differently can be found in the data_visualization folder.

9.1 Is crime related to the weather?

One of the interesting things we found during data exploration is that when you plot the crime counts through time, it looks sinusoidal and implies seasonal effects. Weather data from Central Park was scraped from Weather Underground [1] from the beginning of 2006 to the end of 2015 (Wunderground scraper+parser.ipynb). This weather data includes information on temperature, wind chill, dew point, humidity, pressure, visibility, wind direction, wind speed, gust speed, precipitation, events, and conditions. We focused on 3 variables that were most likely to affect crime rates: **temperature**, **snow** and **rain**.

Temperature

Average temperature was obtained with a PySpark script (**temp.py**). The weather data includes temperature measurements in Fahrenheit that were made throughout the day at 15 to 30 minute intervals. These were used to calculate the average temperature for the day. Due to instances of extreme outliers in our data, we used the rolling 14 day average of both temperature and crime to plot the relation and calculate the correlation.

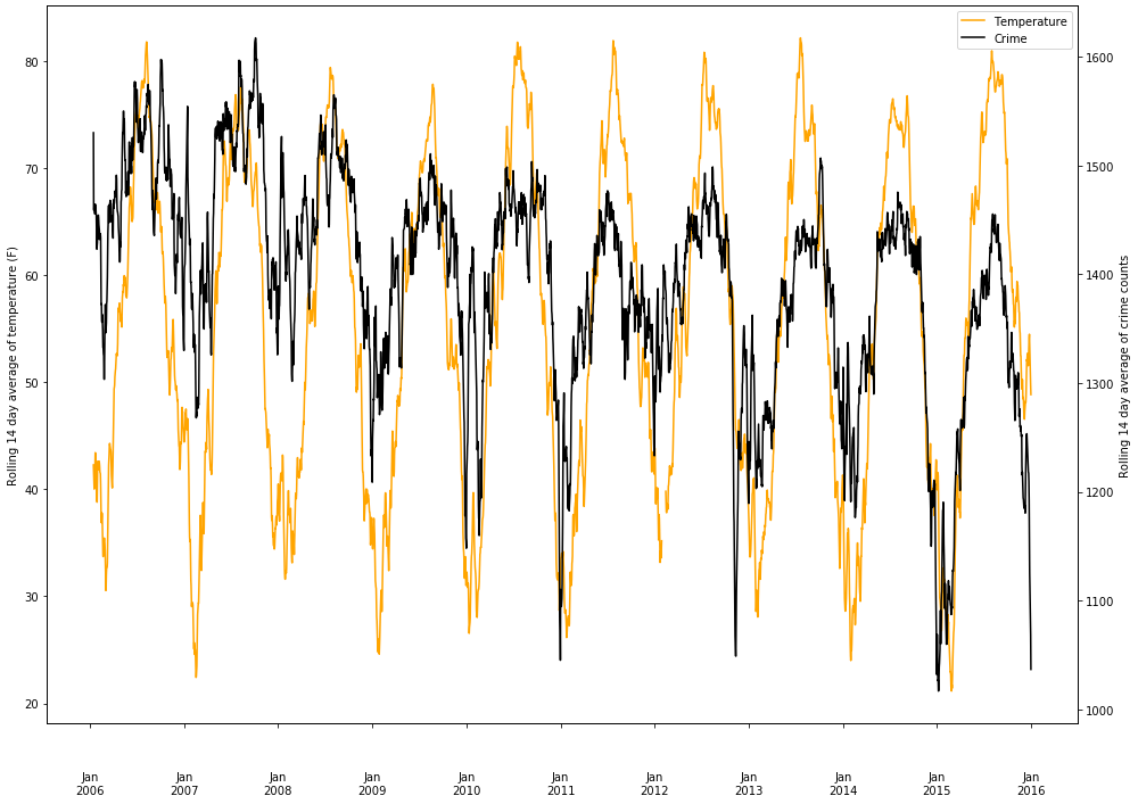

Figure 21: Crime counts over time vs. average temperature over time (rolling 14 day average).

We observe quite a strong positive correlation, in fact. The correlation coefficient was computed to be **0.686146.**

It looks like although temperature is a good predictor, crime seems to be going down over time. We decided to calculate a new function that subtracted 4% of each value per year after 2006. The corrected prediction function looks like *temperature - 0.04*temperature*(year-2006)*.

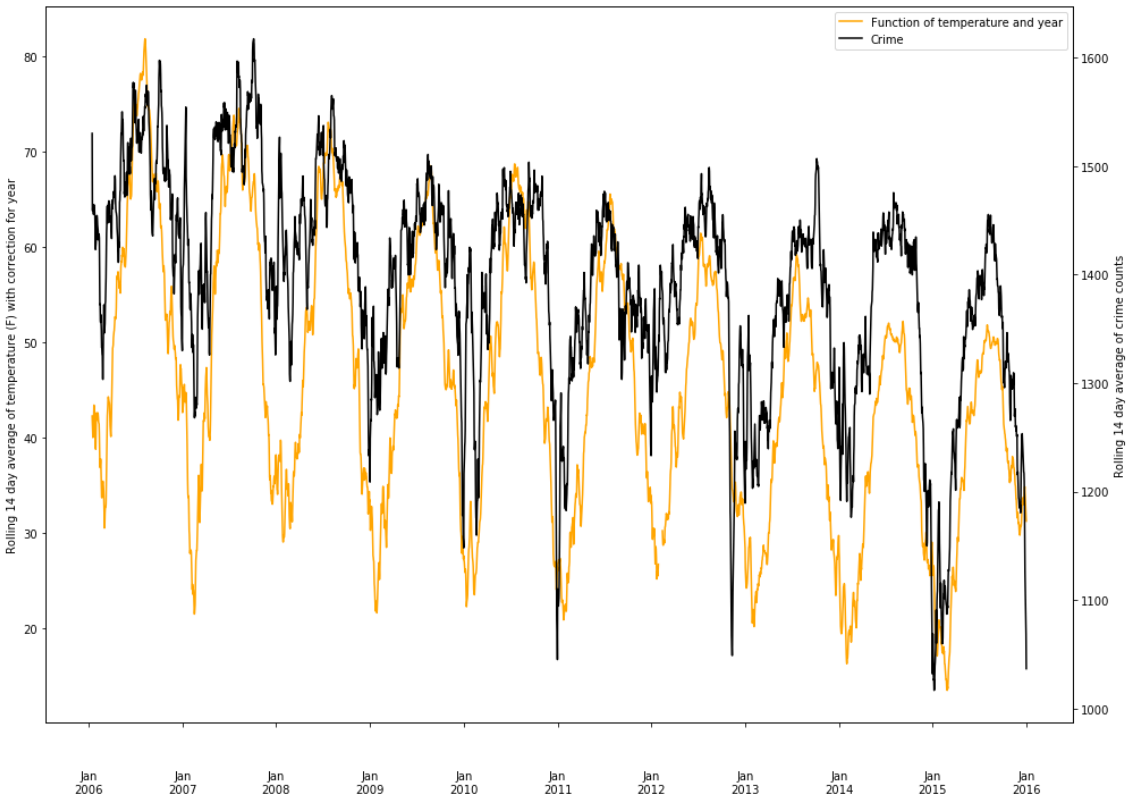

Figure 22: Crime counts over time vs function of temperature and year (rolling 14 day average)

With our prediction function based on year and temperature, we obtain a correlation of **0.839443**. It seems that temperature is very predictive of the amount of crime that occurs and/or gets reported, and crime has been going down slowly in the last 10 years. When is cold outside, people are less motivated to commit crime and/or there are less targets for crime. Perhaps warmer weather causes increased blood flow that makes people more impulsive. It could be that the police are more diligent when the weather is nicer out. The relationship with time could be because the younger generation is committing less crimes according to a Washington Post article [2].

Snow

To go along with the hypothesis that less crimes are committed when the weather is not nice outside, we also looked at the correlation between snow and crime. In the weather underground data, the occurrence of snow is recorded in the *Events* column. We counted the percent of logged points for each day that recorded a snow instance in the Events column with a PySpark script (**snow.py**).

In Figure 23, the whiter dots represent less percentage of snow, while the darker black spots represent more percentage of snow recorded throughout a given day. Days with no snow instances do not have a dot on top of them. One can see that there are several black dots in February 2010 which correspond to lighter squares. However, the correlation does not seem to be particularly strong. It could just be that snow is correlated with colder weather which is a better indicator of crime rates. Still, when there is snow, there does seem to be less crime on average. The correlation was calculated to be **-0.247516**.

Rain

How about rain, then? We also looked at the correlation between rain and crime. In wunderground weather data, the amount of rain is recorded in the *Precipitation* column. We counted total inches of rain that fell each day with the PySpark script (**rain.py**).

In Figure 24, the lighter blue dots correspond to less inches of rain, while the dark blue dots represent more inches of rain. It's hard to see much of a correlation, but it can be noted that few of the very dark purple squares have a rain dot on them. There are some light blue dots around Hurricane Sandy in 2012, but surprisingly it is not very dark and throughout the light area correlated with Hurricane Sandy. The correlation between crime counts and rain was computed to be -**0.113636**, not as strong as the correlation with snow.

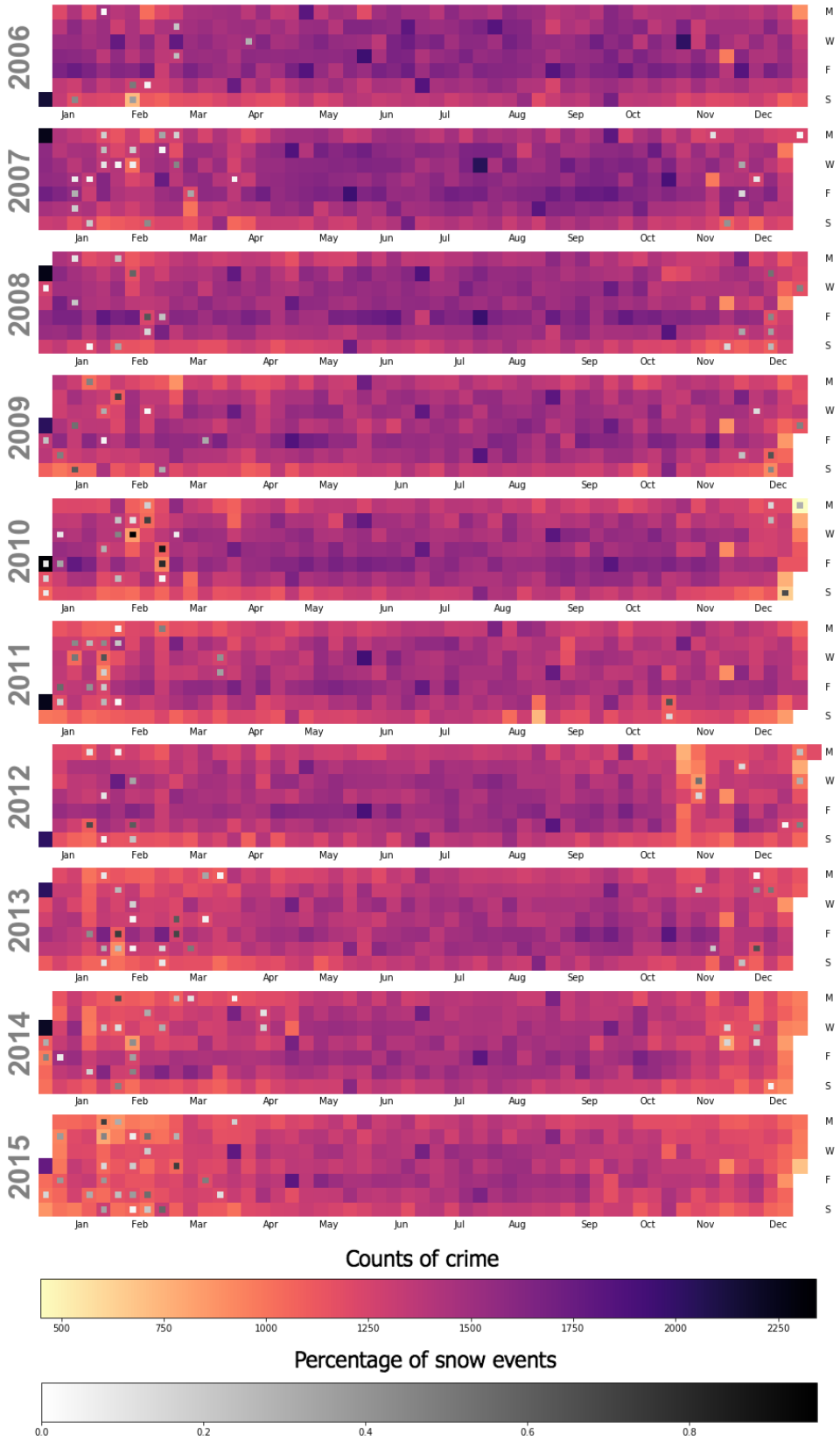

Figure 23: Snow vs. crime counts over time.

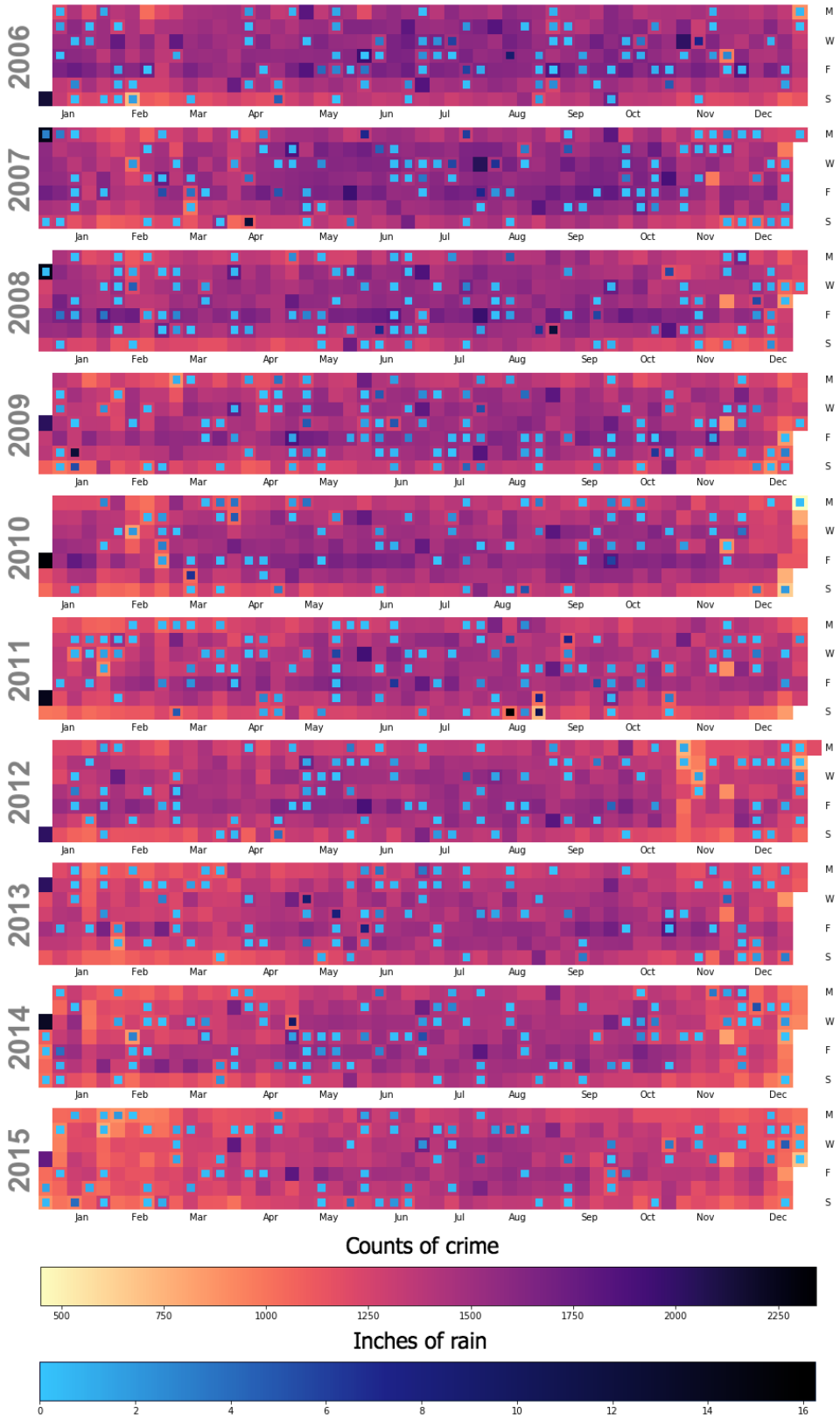

Figure 24: Inches of rain vs. crime counts over time.

9.2 Is crime per capita related to the population density?

It is not that hard to imagine that crimes go up as the population goes up for certain area. More people, more criminals. But does the crowdedness of an area encourage people to commit more crimes? To examine this, we decided to look at the number of crimes per capita and compare it to the population density for a given borough. Because good estimates for population by borough are only available with the US Census of 2010 and 2015, we decided to use the population numbers for 2015 as the estimate for population [3]. Population density was computed using the land area in square miles [4]. Crime per capita was also calculated by dividing the crimes by borough (**countuniques.py** 13) with the US Census 2015 population numbers.

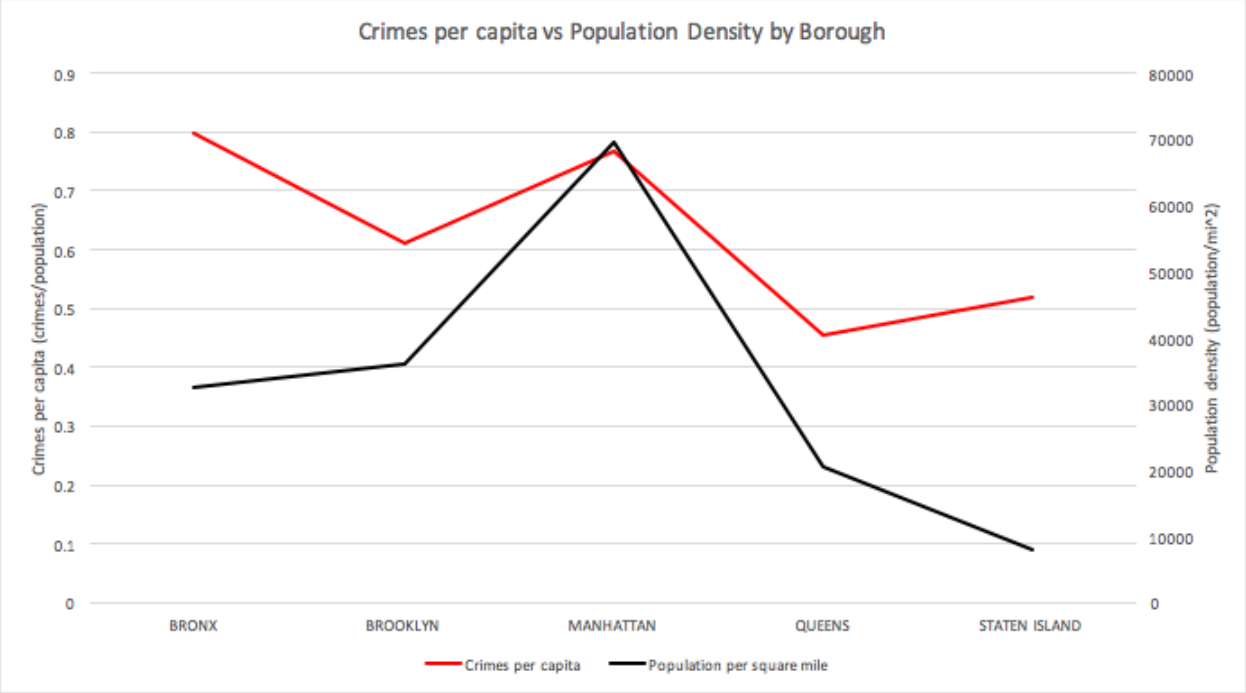

Figure 25: Crime per capita vs Population Density by Borough (plot generated in data_visualization/borough_pop.xlsx)

It does seem like there is a positive correlation between crime per capita and population density. The correlation was calculated to be **0.708895.** Some theories are that the more people that around in a given area, the more targets there are for a criminal operation. It is easier for criminals just to stay in one location and carry out the same crime, rather than having to search for targets in areas that only have the occasional passerby. It could also be that criminals feel that it is easier to blend into the crowd and get away with crimes when there are more people around, at least for many common crimes like pickpocketing, shoplifting, and purse snatching. Crimes such as murder are probably more prevalent in areas with less people around. Another explanation is that certain crimes tend to be committed in large public areas where all the fun things are going on, such as public indecency, drunkenness, gambling, and drugs.

It is interesting to note Bronx seems to be an exception to the trend. It has relatively low population density compared to the crimes per capita that occur there.

9.3 Do long sentence lengths discourage crime?

The populace is often divided on whether we should be "tough on crime" by throwing more criminals in jail with long jail sentences, rather than trying to rehabilitate them and reintegrate them back into society. But are sentence lengths correlated to crime rates? We decided to map each crime type to a sentence length to find out. Using our previously written PySpark script (**agg2cols.py** 7 11), we used the offense description and felony/misdemeanor/violation information to map each offense type to a penal level as defined in the NY Penal Code. We used a guide to NY Penal Law Criminal Offenses [5] to perform this mapping manually, as it requires some discretion. A few offense descriptions could not be found with this data source. Therefore, felonies that could not be mapped to a level were assumed to be the lowest class of felony, E. Misdemeanors that could not be mapped to a level were assumed to be the most common type of misdemeanor, A. Because certain crimes can have various degrees of severity, we mapped each offense to the minimum degree of the crime possible. Then, we mapped each of these degrees of crime to the maximum jail time you can receive for each penalty using a sentencing chart for New York State [6]. Most misdemeanors do not result in jail time, therefore to make it a good comparison, we mapped each to the worst possible outcome.

Penal level	Maximum sentence time	Average days	Counts
Violation	No jail time	0	615188
B misdemeanor	6 months	182	395290
A misdemeanor	1 year	365	2505034
E felony	1.33-4 years	972.725	885795
D felony	3-7 years	1825	439574
D violent felony	7 years	2555	234632
B violent felony	25 years	9125	2304
A-I felony	100 years	36500	4574

Table 2. Max sentence length by penal level vs counts of crime

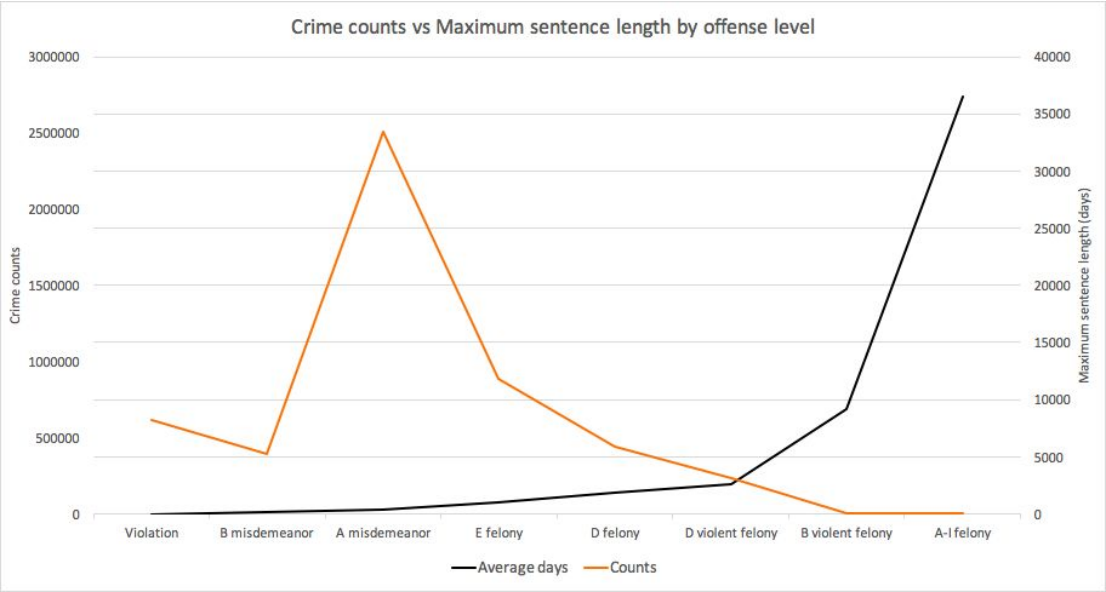

Figure 26: Max sentence length by penal level vs counts of crime (plot generated in data_visualization/sentence_length.xlsx)

Aside from violations and B misdemeanors, it does seem like there is a negative correlation between sentence length and counts of crime. Violations and B misdemeanors do not follow the trend, probably because most police officers do not bother catching and reporting these criminals most of the time, as they have bigger fish to fry. They may let off these people with a warning without filing a report. Most people probably also do not bother to report these incidents to the police. So it could be that they occur much more often than is shown in the graph, they are just not accurately represented in our dataset.

With violations and B misdemeanors removed, the linear correlation is **-0.469974** (though the relationship does not look very linear). Besides long sentences actively discouraging crime, a more significant explanation for the trend could be that people are less willing to commit crimes of higher calibers due to personal ethics. Longer jail sentences are associated with crimes that are truly horrible, like rape and murder. More people are willing to commit small crimes that they can justify to themselves and they think they can get away with.

9.4 Are younger people committing more crime?

In 8.1 we noted findings from an article that implied that the younger generation is committing fewer crimes looking at the trend between 1993 and 2013 while the older generation appears to have maintained the same rate of crime over time. However, in that article, they also mention that the per capita crime rate for 18-39 year olds is substantially higher than the rates for 40 year olds and above. Older people may be tied down with families and more settled down, while younger people could be more deviant with their young and wild spirits. We wanted to see if this phenomenon exhibited itself in our data, and if perhaps it would explain some of the differences in crime rate between boroughs. We pulled the distribution of age groups by borough as determined by the 2015 census [7] and created a 'borough average age'. We compared that indicator to the crime per capita in each borough (**countuniques.py** 13); This index approximates an 'age' for each borough by weighting the median age in ranges provided by census data by the percent of the population in each borough that fall into that age range: (*percent of borough in age group) x (median age of age group for borough)*.

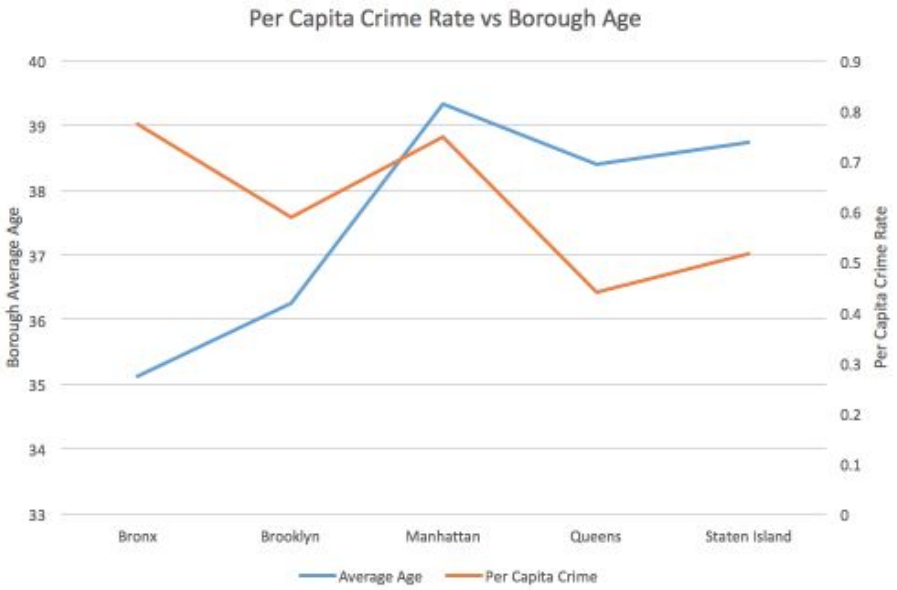

Figure 27: Average age plotted against per capita crime rate by borough

Looking at the 5 data points (one for each borough) we see that there appears to be an inverse relationship and we get a correlation of **-0.3679**. While this is not a strong correlation, is does suggest a non-trivial relationship between per capita crime and age: As the age of a population increases, according to our indicator, the per capita crime rate decreases. If we exclude the Bronx from the correlation calculation, we actually get a positive correlation of 0.2, though it would be irresponsible to exclude a borough for being a leverage point if in fact it is identifying the true trend. We also note that there is a spike in age and per capita crime rate in Manhattan. As travel between the boroughs is incredibly easy, it is tough to entirely attribute the crime per capita of a borough to the population residing in that borough. That said, we do see a non-trivial correlation between younger population in a borough and higher per capita crime.

9.5 Do police officers target minorities?

In a piece released by the NYTimes in 2016 [8], a black NYPD officer details systemic racism and targeting practices in the NYPD that have existed for decades. All over the country, the Black Lives Matter movement is fighting the perceived unfair targeting of black and African American people by the police. In trying to explain differences in crime rate by borough, we hope to examine if this targeting appears in our data.

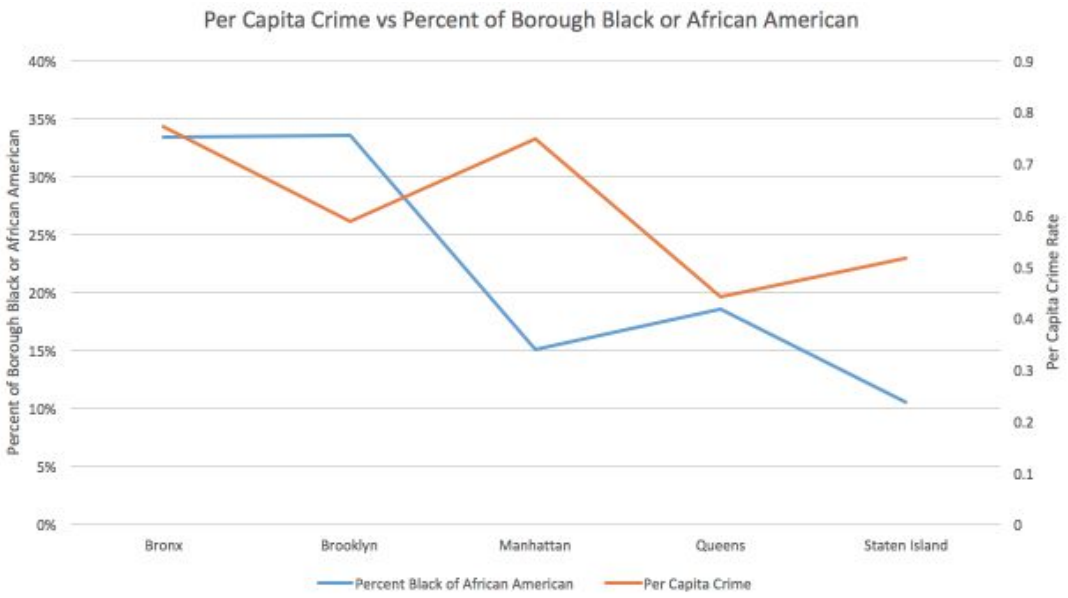

Figure 28: Percent of borough black or African American plotted against per capita crime rate

We do see what appears to be a correlation between per capita crime rate and the percent of a borough that is black or African American; The correlation is measured as **0.375** which might not be a high correlation but it is non-trivial. While some may interpret this to mean that African Americans are committing more crime, we believe this might show a tendency of the NYPD to target those individuals. We wondered if this trend would extend to the percent of non-white people in a borough.

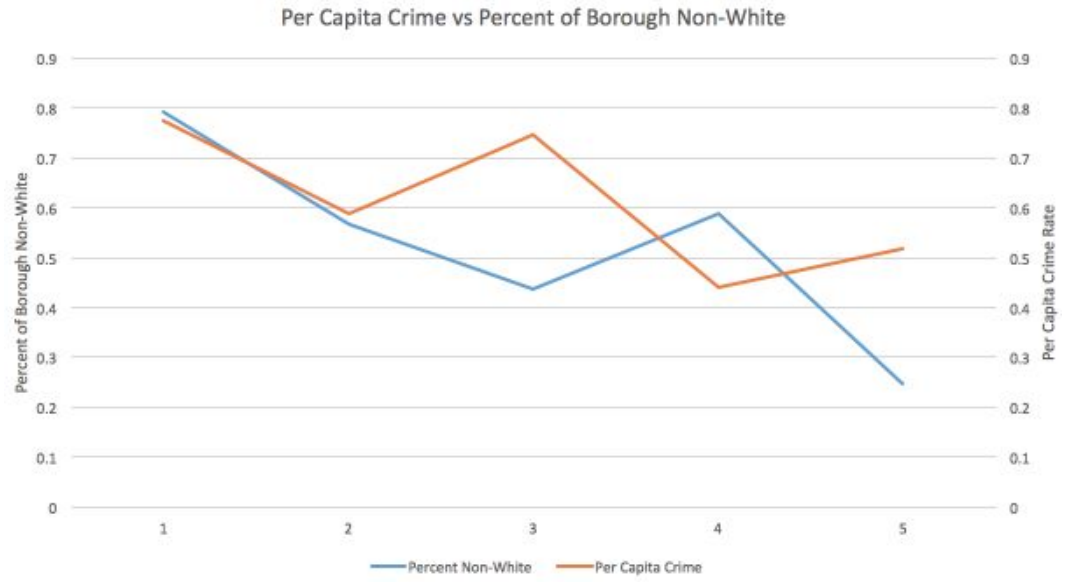

Figure 29: Percent of borough non-white plotted against per capita crime rate

In Figure 29, we see a marginally stronger correlation of **0.39** between per capita crime rate and the percent of a borough that is non-white. This either means that police officers in heavily non-white regions are quicker

to arrest people and more vigilant about petty crime or that police staffs these areas with more officers relative to other areas/boroughs. A piece by the Huffington Post [9] argues that blacks are more heavily policed (and this may extend to all minorities) which confirms our second hypothesis above. Given the NYPD does not release staffing numbers by precinct, we cannot confirm nor refute this, but the data suggests this might be the case.

9.6 Are noise complaints and drunk driving correlated?

In our eyes, when people are driving drunk in large amounts, they are probably also partying in large amounts and making a lot of noise. By this logic we expect to see a high correlation between DUI arrests and noise complaints filed. With that in mind, we mined the 311 Service Requests [10] dataset looking for noise complaints between 2010 and 2015 and calculated the average daily noise complaints both by the day of the year (**avg_day_noise_complaints.py**) as well as the total on every day in that range (**daily_noise_complaints.py**). Then we compared these numbers with the comparable DUI arrest trend (average daily: **avg_day_DUI_crime.py** , every day: **daily_DUI_crimes.py**).

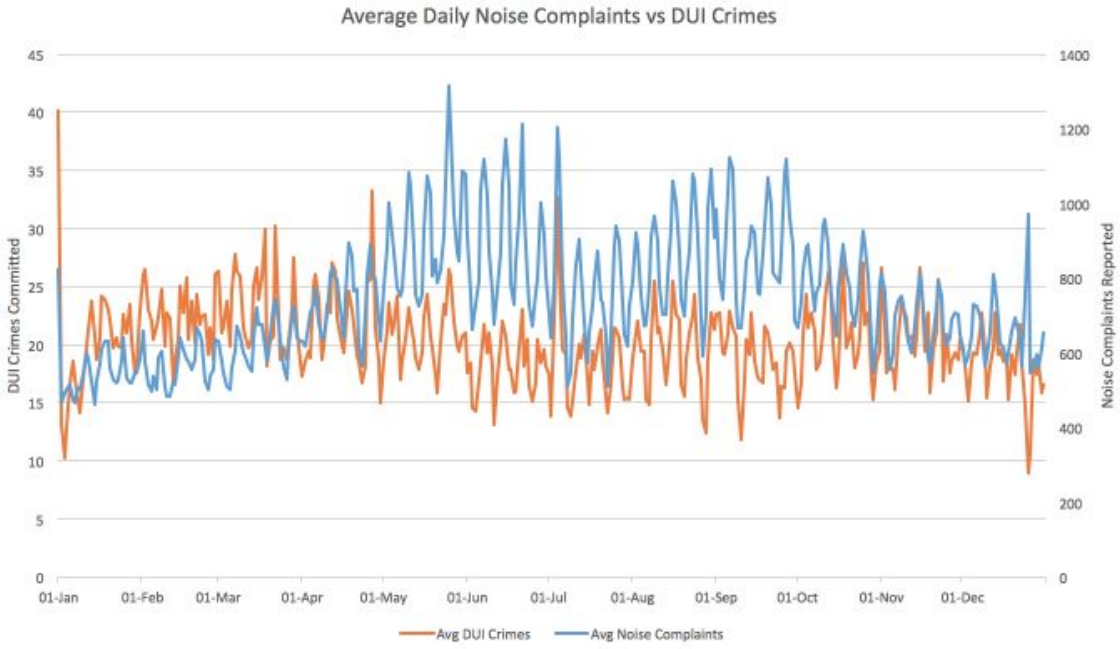

Figure 30: Average noise complaints per day of year (2010-2015) plotted against average number of DUI crimes (2006-2015)

We see a few interesting trends in the noise complaint data itself; it appears that noise complaints increase between April and July, then drop a bit after the 4th of July but increase again until late September when an overall decrease then occurs. As noise complaints are not only in apartment buildings (can also be outside) it's likely that as the weather warms up in April and the early to mid summer, more people are getting out and meeting up with friends and making noise, hence an increase in complaints. In mid to late summer it gets seasonably hot outside and it's possible fewer people are outside making a ruckus until the school year begins again when the complaint rate goes back up. A few interesting dates with spikes in 311 noise complaints (relative to dates around them) are January 1, July 4 and December 25; Similarly to the DUI data, we found relatively higher noise complaint incidence on New Year's Day and the 4th of July as these days large groups of people tend to celebrate and drink together, but unlike the DUI data, we find another spike on Christmas Day. While the DUI crime rate drops on Christmas Day, noise complaints are relatively high compared to days

around Christmas. Perhaps carollers are bothering some, but more likely, families are gathering in small apartments and neighbors on the other side of paper thin walls are being disturbed by the noise naturally generated by a large amount of people in the apartment next door.

We find the correlation between these two series to be **0.26** which, while not strong, does imply a weak relationship between DUI crimes and noise complaints. That said, we pose the hypothesis that the noise complaints might be understated. There is at least one noise complaint filed we know about because of a certain Halloween party which is not recorded in the 311 dataset. It is possible that on days with many noise complaints that 311 does not record every single one given there are so many. If that is true (and this Halloween party did not fall into a data black hole) then the spikes we see on January 1, July 4 and December 25 might be slightly muted, causing a reduction in the correlation, as we see the spikes are more pronounced on January 1 and July 4 in the DUI data. Next, we will look at the total daily counts of these data and test if the series of individual days have strong correlation.

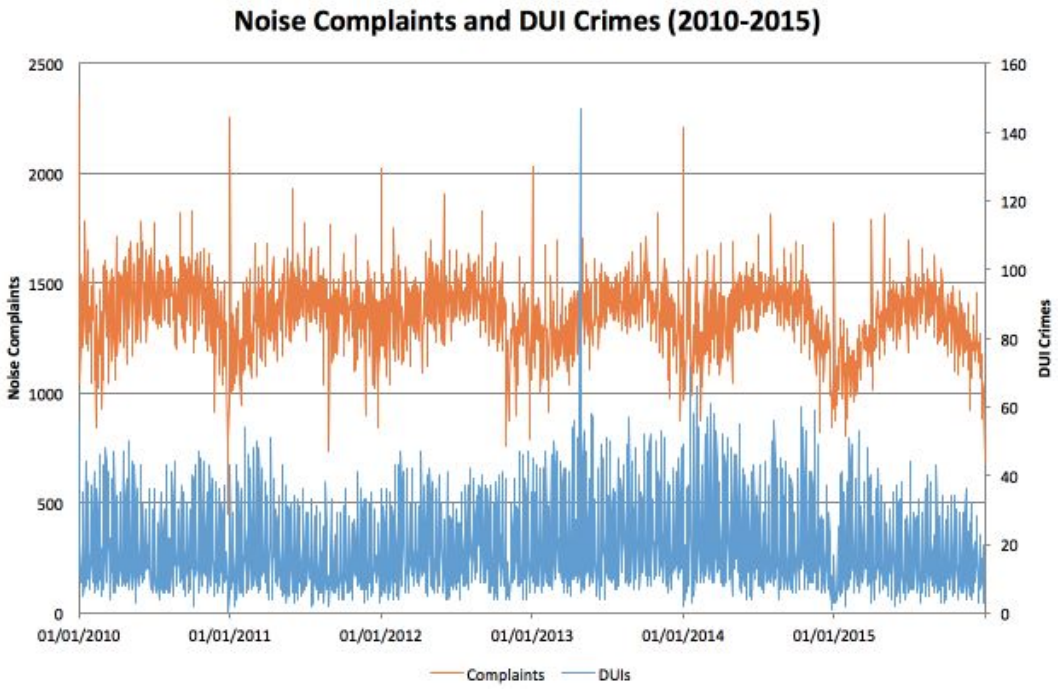

Figure 31: Daily noise complaints (2010-2015) plotted against average number of DUI crimes

We find the correlation here to be much lower, at **0.058** signifying there is no material correlation between the number of DUI crimes on a given day and the number of noise complaints. As drunk driving is not an activity exclusive to noisy parties and noise complaints are not exclusive to party-related noise complaints, we find the January 1, July 4 and December 25 findings to be interesting and insightful, though the overall trend implies that en masse there is no relationship between these incidents.

When we only look at values that can be classified as outliers (not falling between the range [mean - 2*stdev, mean + 2*stdev]) we find an ever lower correlation, at **0.0074**. This only furthers our inclination that noise complaints and DUI crimes are on the whole uncorrelated. Both happen in higher quantities on January 1 and July 4, but looking across days we don't see a direct daily correlation.

9.7 Gentrification Effect

We attempted to understand how the process of gentrification relates with crimes. Gentrification is defined according to Wikipedia as "the process of renovation of deteriorated urban neighborhoods by means of the influx of more affluent residents" [11]. One of the main characteristics of gentrification is the increase in the properties' value because of the mentioned renovation process. We use the time series of price per square feet of the properties within each police precinct as a proxy for the degree of gentrification.

By mining the PLUTO dataset [12] we obtain the mean value per square feet of properties (**analyze_pluto.py**) for each of the police precincts and for each of the years in the range 2006-2015. Particularly the columns used from the dataset are *AssessTot*, *PolicePrct* and *LotArea* . The PLUTO dataset does not report any information for the year 2008 so we do not take this year into account in the analysis. On the other hand, we create a time series of the yearly number of recorded crimes for each of the police precincts and compute the correlation between the two time series.

The result of this analysis is shown in Figure 32 where we show the correlation for each of the police precincts grouped by borough. In order to better identify the scenario in each of the precincts we also plot the mean yearly percentage change in number of crimes for each of the precincts.

The overall trend shows a decrease in the number of crimes as previously identified in Section 9.1. It is also the case that the values of the properties in most of the precincts is consistently increasing. The correlation analysis indicates a strong overall negative correlation between this two variables, which points to the conclusion that more wealth tend to reduce the number of crimes.

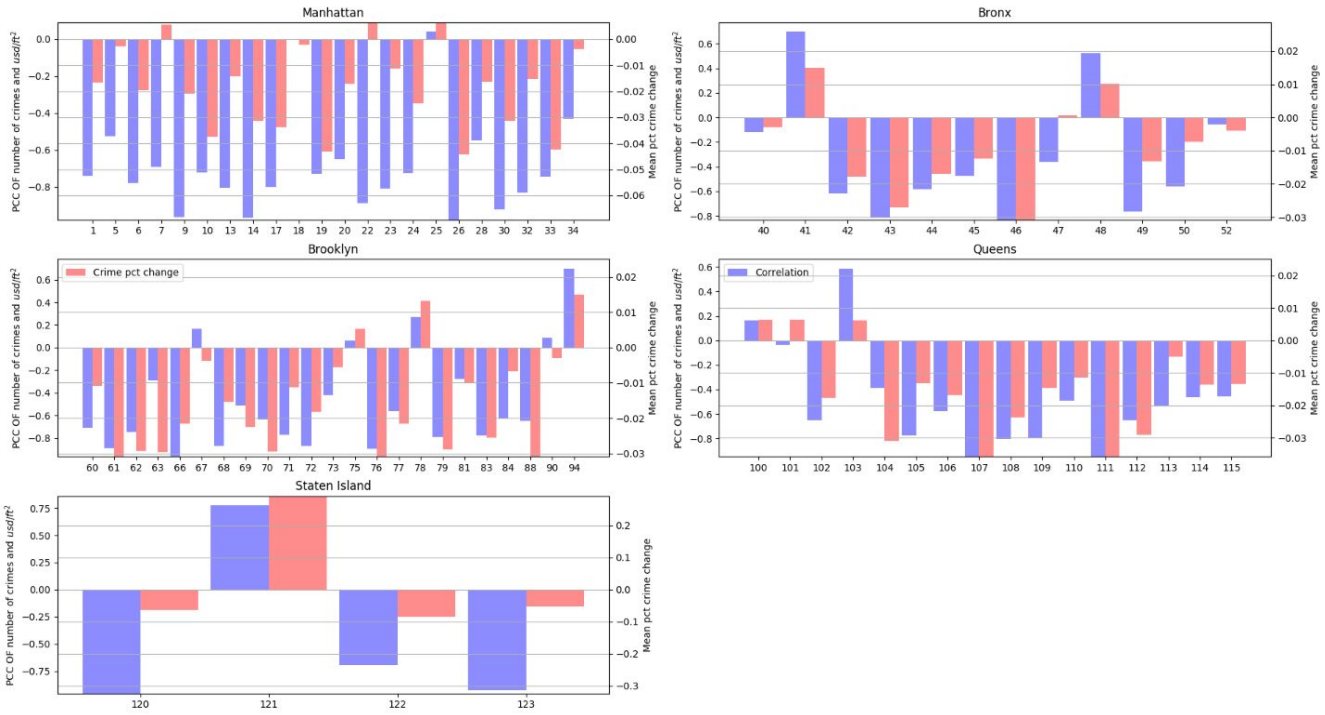

Figure 32: Correlation of the number of crimes and the mean value of the properties per square feet for 2005-2015 (year 2008 excluded from the analysis) Code to reproduce this plot in data_visualization/Price-Crimes-Plots.ipynb.

The highest negative correlation (-0.97) was found to be for precinct 26 which is in the neighborhood of Morningside Heights / East Harlem, and was also identified as one of the most gentrified neighborhoods by NYU Furman Center [13].

There are a few cases in which the correlation is positive, meaning more crime in zones in which the land price is increasing as well. The precinct with the highest positive correlation is in Staten Island. After closer inspection, there seems to be a problem with the data in this precinct, since the number of crimes jumps from 16 in 2012 to 3669 in 2013. If we disregard this precinct, the second highest positive correlation (0.697) is observed in precinct 94 corresponding to the east side of Greenpoint.

We summarize this information in the map of Figure 33 showing the correlation for each of the police precincts, where bluer values represent a more negative correlation while redder ones represent positive correlations. As in section, the geojson file for the map was generated with **heatmap.py**.

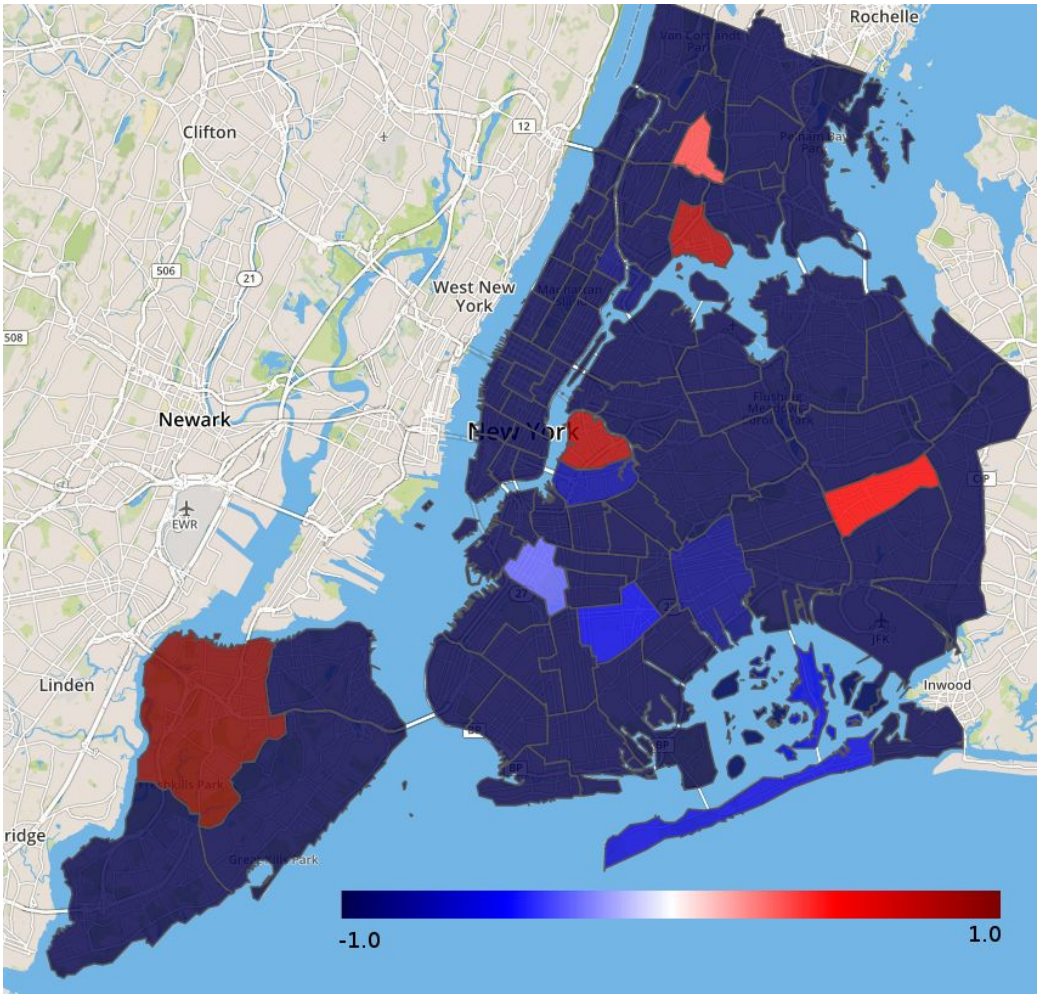

Figure 33: Correlation of the number of crimes and the mean value of the properties per square feet for 2005-2015 (year 2008 excluded from the analysis).

10. Summary/Conclusions

In this project, we first analyzed the NYPD Complaint Data Historic dataset and evaluated the quality and consistency of the dataset. We identified various data quality issues and problems with the data, including, but not limited to: missing data, unnatural spikes on certain dates, time inconsistencies with values at both 00:00:0 and 24:00:00, end dates that occurred after start dates, years that were recorded as 1015, and categories that required merging.

The second main part of the project involved Data Exploration, in which we analyzed different aspects of the data such as seasonality trends, time differences in the reporting and occurrence dates of crimes. We also explored relationships between different attributes of our data: level of offense per borough, shoplifting trends, drinking and driving, among others. We identified two spatial resolutions for the geographical analysis performed, boroughs and police precincts and mapped the crimes of the city over them.

In the last part of the project, we posed several hypotheses and evaluated them using the original dataset alongside several different open data resources. Some of the mined datasets include PLUTO for real estate data, historical weather data, 311 complaints, and the US Census data. The topics we explored along this part were the seasonality of the crimes, how the population density affects the crimes per capita, the relationship between age and crime rates and the effect of gentrification in crime, among others. We were able to disprove some of our hypotheses, while some others held. For example, the correlation between crime and average temperature was quite strong. On the other hand, no significant relationship was found between 311 noise complaints and drunk driving offenses. For most of the cases, we found that the available data was not necessarily conclusive, since this is a very complex topic in which unobserved factors can play an important role (and of course, correlation does not imply causation). Table 3 on the next page summarizes the relationships found throughout our study.

All of our work can be reproduced and replicated, with all of our scripts and data visualization methods available on Github (https://github.com/florielle/bigdataproject) with explicit README files.

10.1 Note on the geographical analyses

We performed several analyses relating crimes in a certain location with some characteristics of the demographics at that location, particularly in sections 9.2, 9.4 and 9.7. We acknowledge that a crime in certain neighborhood or precinct was not necessarily committed by a resident of it, and thus our results are potentially biased, although still relevant. Manhattan likely is the most biased borough, given the percent of daily 'foreign' visitors it hosts.

Crime variable	Variable	Pearson correlation
Rolling 14 day average Crime counts	Rolling 14 day average Temperature	0.686146
Rolling 14 day average Crime counts	Rolling 14 day average Temperature with correction factor for year	0.839443
Crime counts (per day)	Percent of reported snow instances (per day)	-0.247516
Crime counts (per day)	Inches of precipitation (per day)	-0.113636
Crime per capita (per borough)	Population density (population / mi^2)	0.708895
Crime counts (per penal level without violations and B misdemeanors)	Maximum sentence length in days	-0.469974
Crime per capita (per borough)	Average age	-0.3679
Crime per capita (per borough)	Percent black/African American	0.375
Crime per capita (per borough)	Percent non-white	0.39
Average number of DUI crimes (per day of year)	Average number of noise complaints (per day of year)	0.26
DUI crimes (per day)	Noise complaints (per day)	0.058
Yearly number of crimes	Yearly mean property value per square foot (mean across all precincts)	-0.53
Yearly number of crimes in precinct 26	Yearly mean property value per square foot in precinct 26 (maximum negative correlation across all precincts)	-0.97
Yearly number of crimes in precinct 94	Yearly mean property value per square foot in precinct 94 (maximum positive correlation across all precincts)	0.697

Table 3. Variable relationships.

11. Contributions

Section	Joyce Wu	Alexandra Simonoff	Felipe Ducau
Part I Data quality and cleaning	- Count uniques script - Inspect 24:00:00 script - Spark validation scripts for numeric, date, and time columns - Data cleaning script	- Spark validation scripts for categorical columns and location coordinates - Spark script to check for unique primary key	- Shell scripts to run all column valid/invalid/null scripts - Hadoop script to count the result of validation scripts - Spark validation script to ensure end date is after start date - Data quality issues update
Part I Exploration	- Crime by Day of Week - Shoplifting Trends - Difference Between Occurrence Date and Police Report Date - Level of Offense by Borough - Type of Offense by Borough	- Seasonality of Crime - Violent Crimes - Crime by Time of Day - Forecasting Crime Through 2020 - Drinking and Driving - Pre 2000 Crime Profile	- Number of Crimes by Precinct - Heat map script
Part II Hypotheses	- Crime vs Weather (Temperature, Snow, Rain) - Crime per capita vs population density by borough - Crime counts vs potential jail sentence time	- Crime rates vs. age by borough - Crime per capita by race - Drunk driving offenses vs 311 noise complaints	- Experimental Setup - PLUTO dataset mining - Gentrification analysis
Report & misc.	- Valid/Invalid/Null Counts Justification - Data cleaning	- Introduction - Abstract	- Valid/Invalid/Null Counts Table - Conclusion - Sources compilation

Table 4. Project contributions.

All the participants of the group contributed equally to the following tasks:

- Code review.
- Creating the README files.
- Report review/corrections.

12. References

[1] Weather Underground - https://www.wunderground.com/

[2] Keith Humphreys. “Young people are committing much less crime. Older people are still behaving as badly as before”. The Washington Post. September 7, 2016. https://www.washingtonpost.com/news/wonk/wp/2016/09/07/young-people-are-committing-much-less-crime-older-people-are-still-behaving-as-badly-as-before/?utm_term=.3dac88ea02d3

[3] Current and Projected Populations. NYC Planning. http://www1.nyc.gov/site/planning/data-maps/nyc-population/current-future-populations.page

[4] Boroughs of New York City. Wikipedia, the free encyclopedia. May 9, 2017. https://en.wikipedia.org/wiki/Boroughs_of_New_York_City

[5] New York State Law Penal Law: Guide to NY Penal Law Criminal Offenses. http://ypdcrime.com/penallawlist.php

[6] Sentencing Chart. Galluzzo & Arnone LLP. http://www.gjllp.com/Resources/Sentencing-Chart.aspx

[7] American Fact Finder. United States Census Bureau. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

[8] A Black Police Officer’s Fight Against the N.Y.P.D. Saki Knafo. The New York Times Magazine. Feb 18 2016. https://www.nytimes.com/2016/02/21/magazine/a-black-police-officers-fight-against-the-nypd.html

[9] Black Crime Rates: What Happens When Numbers Aren’t Neutral. Kim Farbota. Huff Post, THE BLOG. 09/02/2015. http://www.huffingtonpost.com/kim-farbota/black-crime-rates-your-st_b_8078586.html

[10] 311 Service Requests from 2010 to Present - NYC Open Data. https://data.cityofnewyork.us/Social-Services/31With that in mind, we looked into 311 Service Requ1/wpe2-h2i5

[11] Gentrification. Wikipedia, the free encyclopedia. May 9, 2017. https://en.wikipedia.org/wiki/Gentrification

[12] PLUTO and MapPLUTO. NYC Planning. https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

[13] Tanay Warerkar. Behold, NYC's 15 Most Rapidly Gentrifying Neighborhoods. Curbed. May 9, 2016.https://ny.curbed.com/2016/5/9/11641588/nyc-top-15-gentrifying-neighborhoods-williamsburg-harlem-bushwick

[14] NYPD Complaint Data Historic. New York City Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[15] Polly Mosendz, Victoria Bekiempis and Michele Gorman. In New York, mixing May day and Freddie Gray. Newsweek. 5/1/15. http://www.newsweek.com/may-day-freddie-gray-new-york-327775