UC Berkeley MIDS, Python Fundamentals W200 Spring 2018

Project 2 Final Report

NYPD Complaint Data Historic Data Analysis

Haihui Cao, Kenneth Chen, Benjamin Silk

Summary

In accordance with New York City's open data law (Local Law 11 of 2012), a repository of government-produced, machine-readable data sets are available for free via the NYC Open Data portal on NYC.gov. NYC Open Data makes the wealth of public data generated by various New York City agencies and other City organizations available for public use. Anyone can use these data sets to participate in and improve government by conducting research and analysis or creating applications, thereby gaining a better understanding of the services provided by City agencies and improving the lives of citizens and the way in which government serves them. Thanks to NYC Open Data, we have data for NYPD Complaint Data Historic. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2016. With this data in hand we intend to pursue answers to the following questions

1. Do the data demonstrate improved government management and crime reduction over the documented time horizon?
2. How do crimes correlate geographically? Are there any hot spots for crimes?
3. Are there any areas in New York that crime cases take longer than usual or any discrepancy across crime cases in New York?
4. What are the most frequent crimes?
5. Does crime fluctuate over time? What are the crime trends from 2006-2016?
6. Which year or month has the most crimes? Are the crimes related to seasons or weather?
7. How are crimes correlated to demographic and population data geographically?

Data Source

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

Critical Parameters of the Dataset:

- "CMPLNT_FR_DT": Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
- "CMPLNT_TO_DT": Ending date of occurrence for the reported event, if exact time of occurrence is unknown
- "RPT_DT": Date event was reported to police
- "OFNS_DESC": Description of offense corresponding with key code
- "CRM_ATPT_CPTD_CD": Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
- "LAW_CAT_CD": Level of offense: felony, misdemeanor, violation
- "BORO_NM": The name of the borough in which the incident occurred
- "LOC_OF_OCCUR_DESC": Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
- "Latitude": Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- "Longitude": Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

Supplemental Datasets:

- NYC Demographic profile
- NYC total population, population change and density http://www1.nyc.gov/site/planning/data-maps/nyc-population/census-2010.page?tab=1

Data Sanitization and Engineering:

- **Approach 1 by Kenneth Chen**

We first checked the overall dataset. NYPD data have about 5.6 million crimes between 2006 and 2016. Each crime case has 24 features such as (1) case ID, (2) case reported date, (3) closed date, (4) crime description, (5) offense, (6) crime category, (7) city, (8) location, (9) latitude, (10) longitude so on and so forth. A preliminary check for all features led us to identify two features that are mostly empty values ['PARKS_NM'], and ['HADEVELOP']. We deleted the two features. The other features that we considered crucial in our data analysis are:

(1) ['CMPLNT_FR_DT'] = complaint from date, i.e., case reported date
(2) ['CMPLNT_FR_TM'] = complaint from time, i.e., case reported time
(3) ['CMPLNT_TO_DT'] = complaint to date, i.e., case final date
(4) ['CMPLNT_TO_TM'] = complaint to date, i.e., case final time

Before we merged date and time, we removed any complaints (rows) with missing value in any of the date and time above. We also removed the complaints with missing value in crime description, longitude and latitude. This left us with the crime data down to 3.9 millions rows with 22 features.

We merged complaint date and time to facilitate calculating the time taken to resolve the case. Since the majority of crime cases were resolved in a day, it would be necessary to check the time taken (hours or minutes) to close the case. Using pd.to_datetime, we merged columns[2] and [3], and converted them into [datetime]. Similar approach was also applied for columns[4] and [5].

After conversion, we observed that some complaints were filed 1906 and closed in 2006. We believed that the complaint should not have lasted for 100 years. We assumed that the complaint was opened and closed in 2006. Due to typo, it was recorded as 1906 or 1946 etc. For those situations, we removed the complaints from our analysis due to uncertainty surrounding the complaint case.

NYPD columns

0	CMPLNT_NUM
1	CMPLNT_FR_DT
2	CMPLNT_FR_TM
3	CMPLNT_TO_DT
4	CMPLNT_TO_TM
5	RPT_DT
6	KY_CD
7	OFNS_DESC
8	PD_CD
9	PD_DESC
10	CRM_ATPT_CPTD_CD
11	LAW_CAT_CD
12	JURIS_DESC
13	BORO_NM
14	ADDR_PCT_CD
15	LOC_OF_OCCUR_DESC
16	PREM_TYP_DESC
17	X_COORD_CD
18	Y_COORD_CD
19	Latitude
20	Longitude
21	Lat_Lon

- **Approach 2 by Haihui Cao**

Different ways of data sanitization should be explored by each group member. Original NYPD dataset has 5,580,035 rows and 24 columns. Another way to sanitize the data is to check the exact date of occurrence for the reported event (column 'CMPLNT_FR_DT'), not the exact time (i.e. hours:minutes:seconds). Some events have the exact date of occurrence, but don't have the exact time of occurrence. Different from Approach 1 that removed the events without either exact occurrence date or exact occurrence time, the events that have exact occurrence date but not exact occurrence time are included in Approach 2. The event report date (column 'RPT_DT') was also checked to cross-validate the dataset.

The following columns/variables are checked by approach 2:

(1) [‘CMPLNT_FR_DT’] = Exact date of occurrence for the reported event
(2) ['RPT_DT'] = Date event was reported to police
(3) ['OFNS_DESC'] = Description of offense

Similar as Approach 1, two columns that are mostly empty values [‘PARKS_NM’], and [‘HADEVELOP’] are removed from the dataset. After checking the above 3 columns, some null data and event occurrence dates before 1/1/2006 were observed and removed from the dataset. This left us with the crime data down to 5,541,581 rows with 22 columns, which are different from 3.9 millions rows of Approach 1. The reason has been explained above.

- **Approach 3 by Ben Silk**

My approach was not remarkably different from the above. I imported ‘PARKS_NM’ data as np.str after receiving a memory warning. I tested for the uniqueness of each complaint number by verifying that the set of complaint numbers was the same length as the data set. I then eliminated rows with NaN in the following columns; [‘RPT_DT','KY_CD','CRM_ATPT_CPTD_CD','BORO_NM','LOC_OF_OCCUR_DESC', ‘CMPLNT_FR_DT’], and removed all complaint dates prior to the beginning of our sample (2006). I checked that no complaint dates post-dated our sample’s limit (2016); none of them did.

My analysis did not contemplate or operate on time of report. It seemed reasonable to me not to exclude reports with no time of day as long as the key variables were all accounted for. Later in the analysis, I added day, month, and year columns to the dataset in order to more easily count the number of crimes reported on specific dates and in each of the years of the sample.

Data Analysis Results:

Q1. Do the data demonstrate improved government management and crime reduction over the documented time horizon?

Complaints to NYPD gradually went up from 2006 to 2016. There were 331,754 complaints in 2006 and 385,539 complaints in 2016. 15.6% increase in complaint filed to NYPD within the span of 10 years.

number of complaints by year

2006.0	331754
2007.0	343727
2008.0	349474
2009.0	345874
2010.0	350686
2011.0	349737
2012.0	362387
2013.0	370114
2014.0	377513
2015.0	377502
2016.0	383539

Further look at the complaint increase by complaint category, we found that some complaints remained relatively the same whereas other complaints skyrocketed above 100%. For eg, complaint for ‘fraud’ was 1,537 cases in 2006 and went up to 2,320 cases in 2016, making up for 50.9% increase in ‘fraud’ cases. We also observed that ‘thief’ cases went up from 285 in 2006 to 1444 in 2016 (407% jump in 10 years). Interestingly and unfortunately, ‘sex crimes’ went up from 3 cases in 2006 to 35 cases in 2016 (1067% jump in 10 years).

Based on the statistics we gathered from the NYPD crime reports between 2006 and 2016, we can propose a feasible plan for New York Police Department to increase their awareness in some parts of the crime cases so as to better serve the New York city residents in coming years.

Q2. How does crimes correlate geographically? Are there any hot spots for crimes?

Analyzing the crime statistics between 2006 and 2016, we found that majority of crimes were concerned with 'petit larceny' or small theft, accounting for 688,005 cases. The second major crime was 'harrassment 2' category, counting for 431,275 cases within 10 years. Out of 3.9 million complaint cases, 'petit larceny' accounts for 17.45% followed by 'harrassment 2' for 11% (Figure 1).

The most frequent crimes between 2006 – 2016	
PETIT LARCENY	688005
HARRASSMENT 2	431275
CRIMINAL MISCHIEF & RELATED OF	428688
ASSAULT 3 & RELATED OFFENSES	398708
GRAND LARCENY	359811
DANGEROUS DRUGS	246411
OFF. AGNST PUB ORD SENSBLTY &	206066
BURGLARY	183337
ROBBERY	149707
FELONY ASSAULT	144699

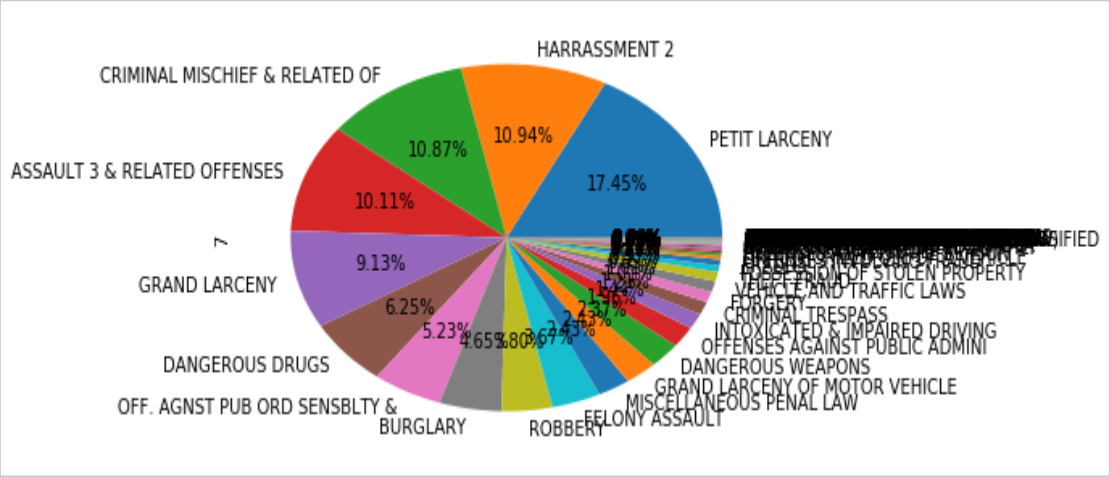
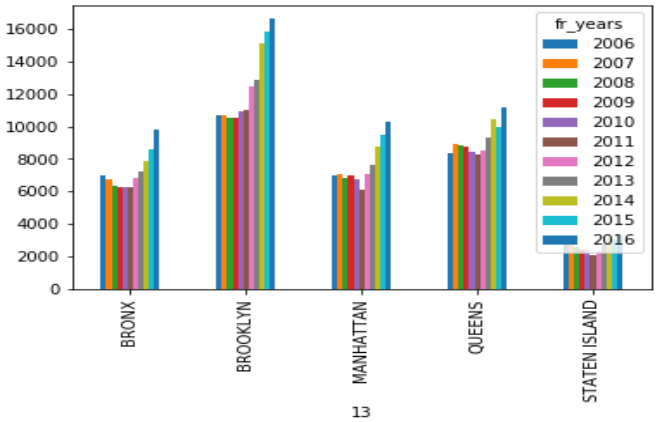


Figure 1. NYPD crimes statistics by each category between 2006 and 2016

'HARRASSMENT 2' statistics by city from 2006 to 2016	
BROOKLYN	137079
QUEENS	100964
MANHATTAN	84029
BRONX	79179
STATEN ISLAND	30024

We further looked at the 'Harrassment 2' statistics and found that majority of harassment occurred in Brooklyn and lowest number of cases from Staten Island. In 2006, 'Harrassment 2' cases were reported at 36,289 and the crime went up to 51,194 in 2016, accounting for 41% jump within 10 years. Since it was accounted for 2nd major crime in NYPD statistics, we further looked at the harassment cases by geography. We found that the crime was more prevalent in Brooklyn, accounting for 137,079 cases and least from Staten island. Although the crime went up by 41%, since it was within the span of 10 years, we further checked if there was any consistent increase in crimes over the years.



Interestingly, we found that 'Harrassment 2' crime consistently went up from 2011 to 2016 in all the city involved: Brooklyn, Queens, Manhattan, Bronx, Staten Island. Rather than a gradual increase in crime cases, 'harrassment 2' crime dramatically increased from 2011 onwards, which seems to suggest that there were more cases related to harassment or people became more sensitive from 2011 onwards. Based on this data, we could suggest NYPD to

further look into those cases, and employ more government personnel to resolve the case more efficiently.

Q3. Are there any area in New York that crime cases take longer than usual or any discrepancy across crime cases in New York?

We looked at the crime cases that took longer than usual (547 days = 1 and half year) from 2006 to 2016. There were 13 cases that took more than 547 days to close the complaint. Using the heatmap function in Python, we further checked if there were any discrepancy across region in New York or any lack of efficiency to close the case. Using the longitude and lagtitude information, we mapped all the crime cases with their duration. We found that there were no discrimination in crime cases and all cases were equally distributed across New York. 3 cases stood out in New York.

Crimes that takes 547 days to close

Complaint ID	Complaint Description
27435	BURGLARY
143061	HARRASSMENT 2
224522	OFF. AGNST PUB ORD SENSBLTY &
244856	THEFT-FRAUD
933785	GRAND LARCENY
1806860	GRAND LARCENY
1932161	HARRASSMENT 2
2027358	GRAND LARCENY
2334642	GRAND LARCENY
2635346	THEFT-FRAUD
2970653	GRAND LARCENY
3684636	OTHER OFFENSES RELATED TO THEF
3954257	OFF. AGNST PUB ORD SENSBLTY &

One case near the Central park at the upper east side and one at the upper west side. The other complaint was near the Times Square. Majority of cases were spread across the New York City. Based on this information, there were no discrimination against any kind of complaint or crimes cases in New York.

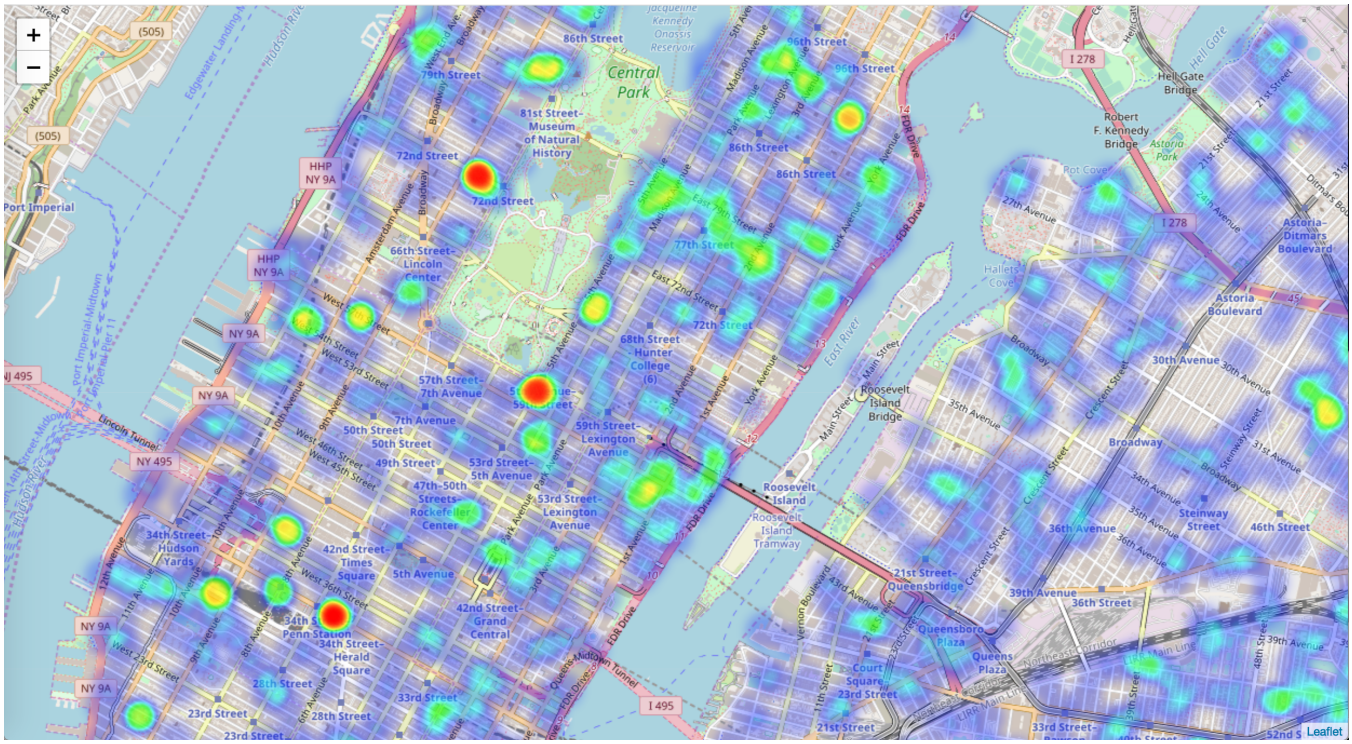


Fig 2. Heat map showing the duration of the crime cases in New York City.

Q4. What are the most frequent crimes?

The most frequent crimes city-wide are petit larceny, harassment 2, assault 3 & related offenses, criminal mischief and related offenses, and grand larceny.

The order of frequency of specific crimes was largely similar across boroughs, but not identical. Petit Larceny is the most common crime city-wide; it is also the most or second-most common crime in each borough. Harassment 2 is the most or second-most common crime in all boroughs except Manhattan, where it is third. Assault 3 and Criminal Mischief are all in the top 5 for each borough. We only really see variability at #5 (Grand Larceny) and lower down on the most common crimes list. The only crimes in some borough's top 5 but not in the city-wide top-5 are 1) Offenses against Public Order and 2) Dangerous Drugs.

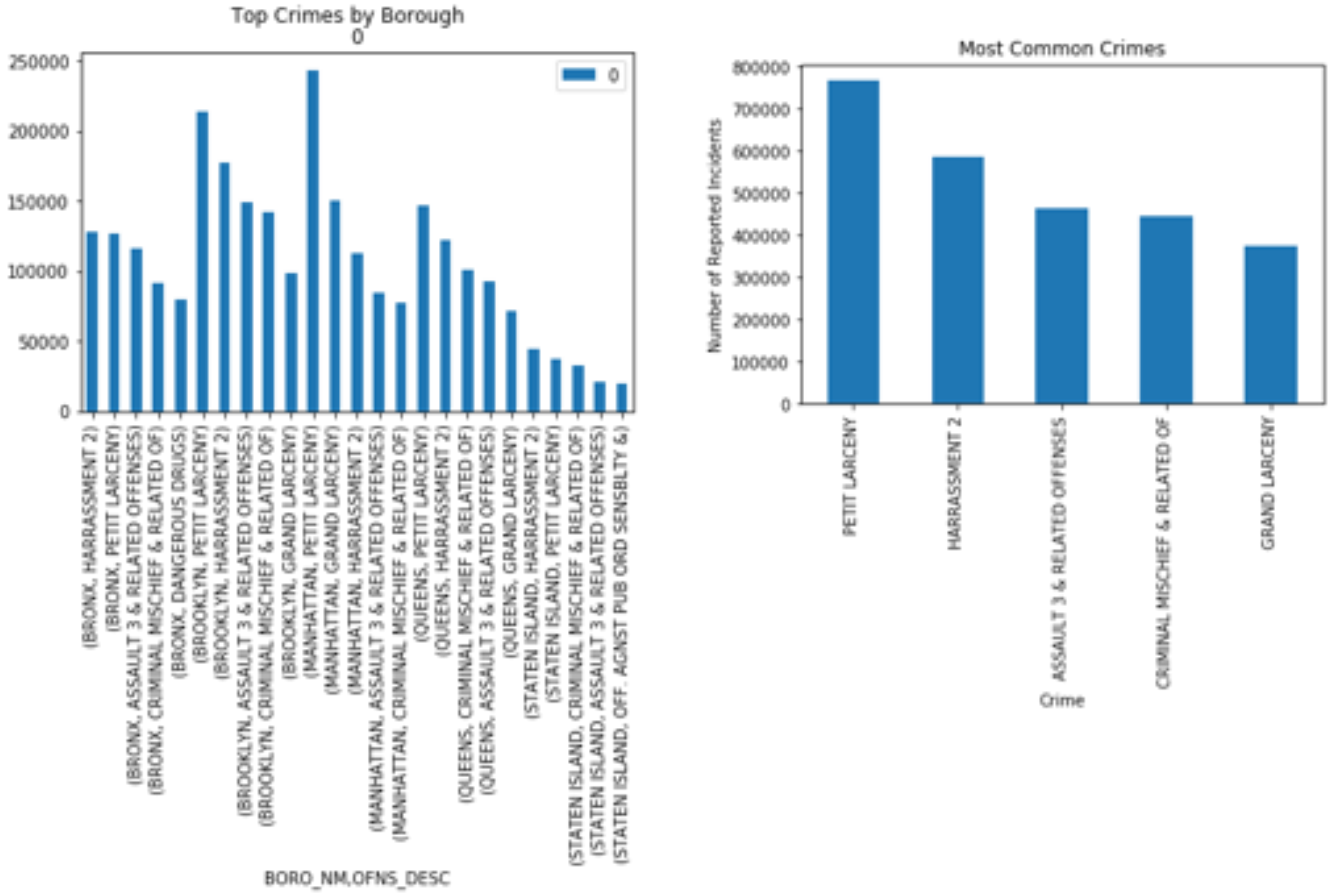
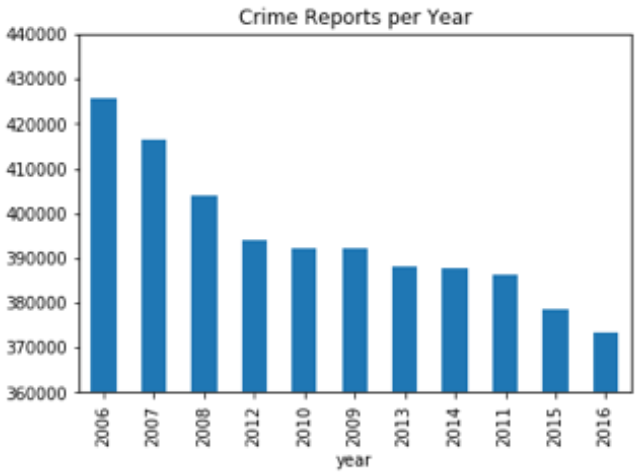


Figure 3. Top Crimes by Borough, and most common crimes from 2006 to 2016

Q5. Does crime fluctuate over time? What are the crime trends from 2006-2016?

Total reported crimes appear to be on a downward trend in New York over the period from 2006-2016. With the exception of 2011->2012, each year had a few number of reported crimes than the year before it. Month-by-month data will be discussed in a subsequent question.



In addition, there appeared to be a regular increase in crimes reported on New Year's day. Indeed, over the 11-year period of the sample, there were 19512 crimes reported on New Year's day, compared to 15718 on the second-highest day, and 11,853 on the average day.

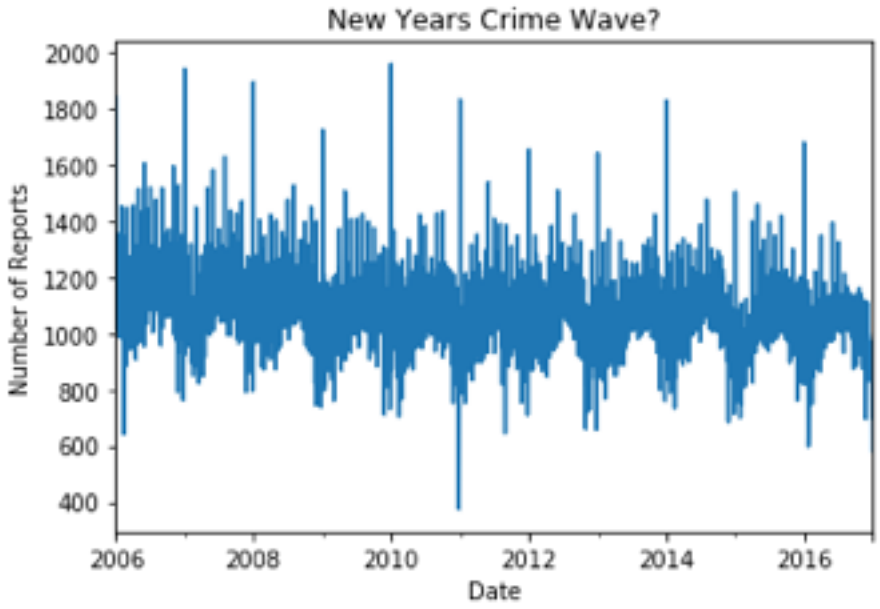


Figure 4. New Year's Crimes from 2006 to 2016

Q6. Which year or month has the most crimes? Are the crimes related to seasons or weather?

We looked at the monthly average of crimes from 2006 to 2016, and the trends are shown in figure 5. It seems that there is some correlation between the monthly average of crimes to seasons over the years. The general trend is that the crimes increase from February to May, keep high during May to October, then decrease from October to December.

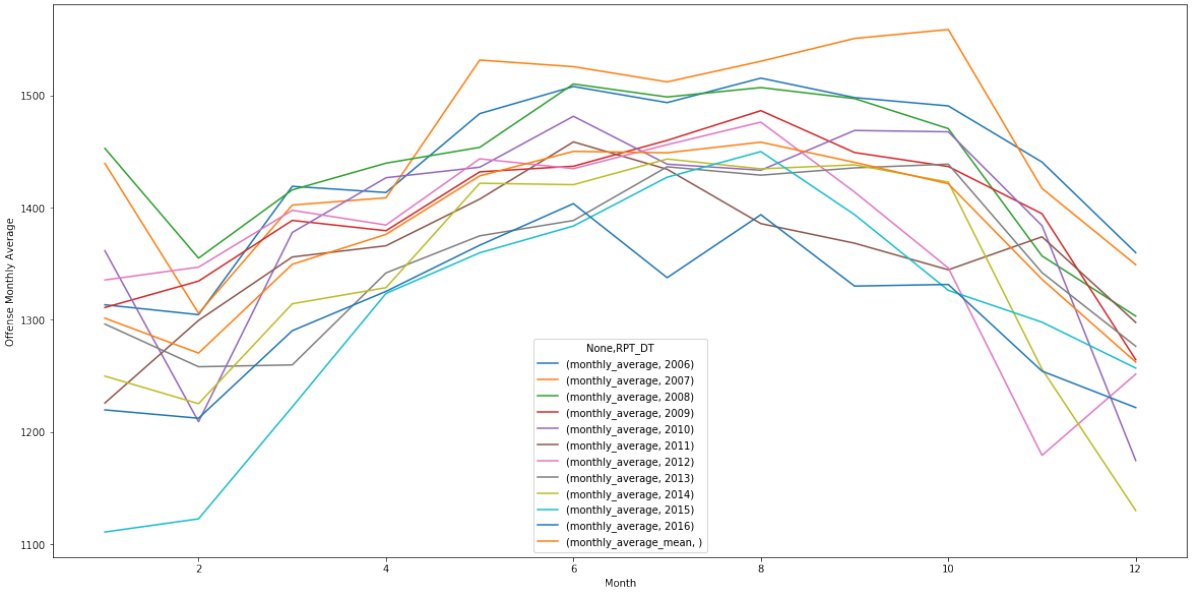


Figure 5. The monthly average of crimes from 2006 to 2016.

In order to see the dependence of crimes on months more clearly, the mean crimes of each month from 2006 to 2016 was plotted and shown in figure 6. The average temperature of New York City was also shown in figure. The crimes are highest during Summer months of June to August with the highest crimes in August. Crimes decrease from August to December when temperature gets lower. From February to June, the crimes go up with the weather's getting warmer. The correlation between Crimes and temperature is 0.95 indicating the high correlation. The exception is January. With the lowest average temperature during the year, the crimes in January is not the lowest. There must have some other factors having effect on the crimes in January and the temperature may not be the only factor. Figure 4 also shows the New Year's crimes waves. Based on the data, we could suggest NYPD to allocate more resources during Summer to decrease the crimes.

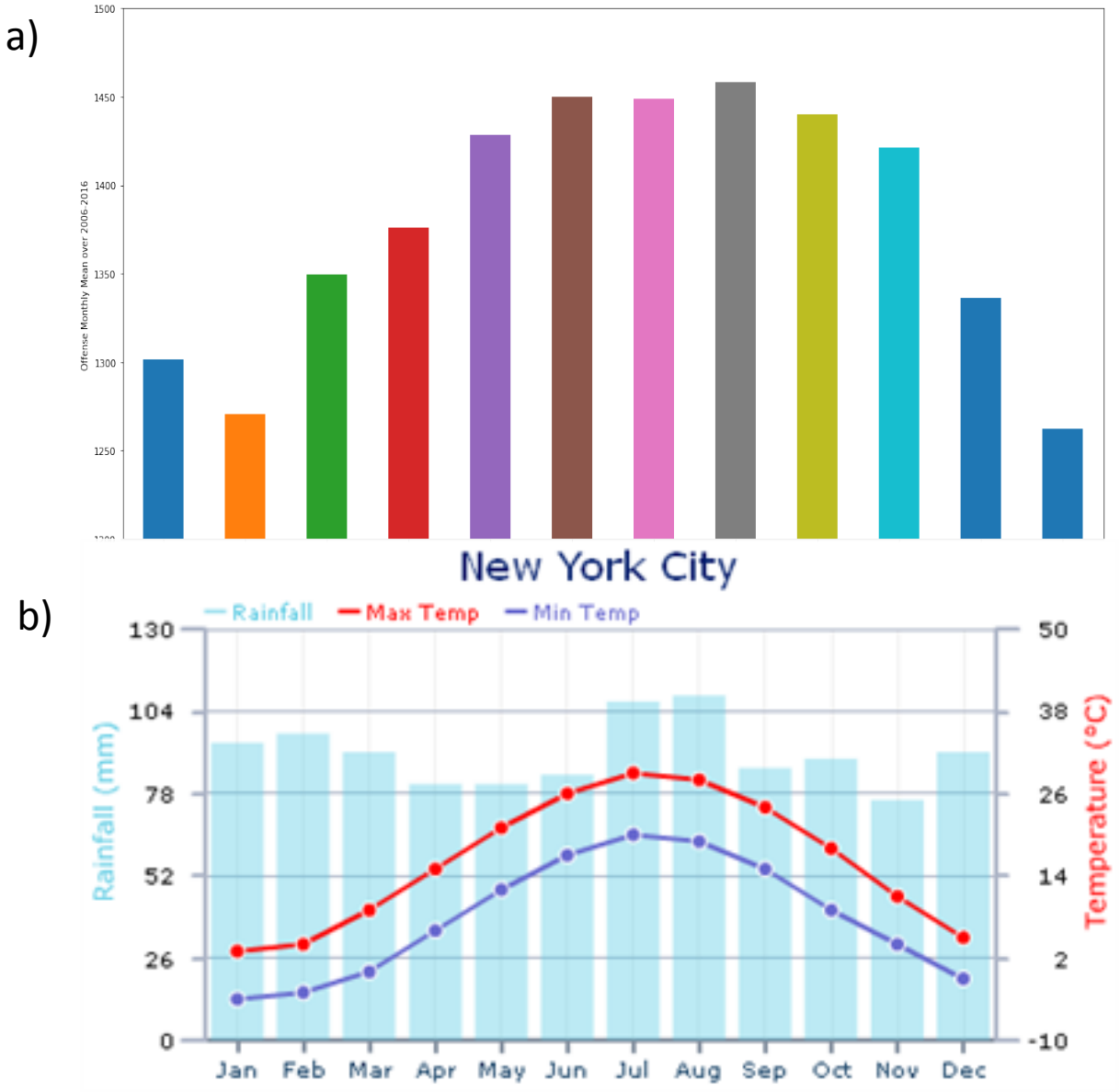


Figure 6. a) The mean crimes of each month from 2006 to 2016. b) The average temperature of New York City each month.

Q7. How are crimes correlated to demographic and population data geographically?

The crimes of the five Boroughs of New York city are plotted in figure 7. Brooklyn has the highest crimes compared to other Boroughs. The population and population per acre of the five Boroughs were shown in Table 1. From the data, we didn't see apparent correlation of the crimes to the populations and population per acre in these areas. Brooklyn has the largest population and highest crimes, while the population per acre is lower than Manhattan. Queen has the second largest population, while it has second lowest crimes. It is probably due to its low population density. Staten Island has the lowest crimes, lowest population and population per acre in the five Boroughs. The correlation between crimes and population is 0.85, and the correlation between crimes and population per acre is 0.64. These correlation numbers indicate the weak correlations. The population is not the only factor playing role in crimes.

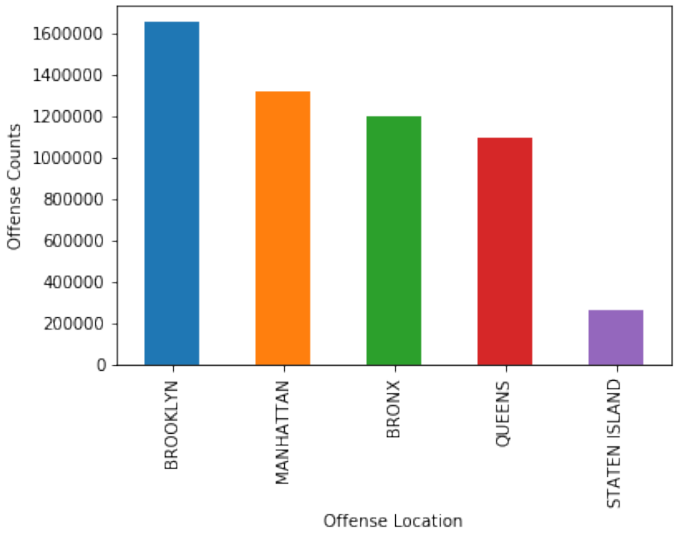


Figure 7. Crimes of the five Boroughs of New York City.

Table 1. The crimes, population, and population per acre of the five Boroughs of New York city in 2010.

Area	Crimes	population_2010	population_per_acre
BROOKLYN	151919	2504700	55.3
MANHATTAN	120960	1585873	108.5
BRONX	111824	1385108	51.4
QUEENS	99082	2230722	32.1
STATEN ISLAND	24063	468730	12.5

Conclusions:

1) Based on the statistics we gathered from the NYPD crime reports between 2006 and 2016, we can propose a feasible plan for New York Police Department to increase their awareness in some parts of the crime cases so as to better serve the New York city residents in coming years.

2) We found that the crime was more prevalent in Brooklyn. Interestingly, we found that 'Harrassment 2' crime consistently went up from 2011 to 2016 in all the city involved: Brooklyn, Queens, Manhattan, Bronx, Staten Island. Based on this data, we could suggest NYPD to further look into those cases, and employ more government personnel to resolve the case more efficiently.

3) There were 13 cases that took more than 547 days to close the complaint. Using the heatmap function in Python, we found that there is no discrimination against any kind of complaint or crimes cases in New York.

4) The most frequent crimes city-wide are petit larceny, harassment 2, assault 3 & related offenses, criminal mischief and related offenses, and grand larceny. The order of frequency of specific crimes was largely similar across boroughs, but not identical.

5) Total reported crimes appear to be on a downward trend in New York over the period from 2006-2016. In addition, there appeared to be a regular increase in crimes reported on New Year's day.

6) The crimes are highest during Summer months of June to August with the highest crimes in August. Crimes decrease from August to December when temperature gets lower. From February to June, the crimes go up with the weather's getting warmer. The exception is January with New Year's Crime wave. Based on the data, we could suggest NYPD to allocate more resources during Summer to decrease the crimes.

7) We didn't see apparent correlation of the crimes to the populations and population per acre in Boroughs of New York City. The population is not the only factor playing role in crimes.