UC Berkeley MIDS, Python Fundamentals W200
April 4, 2018

Project 2 Proposal

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. Did NYPD public access complaint data have a quantifiable impact on improving government management and reducing crimes?
2. How does crimes correlate geographically? Are there any hot spots for crimes?
3. What are the most frequent crimes?
4. Does crime fluctuate over time? What are the crime trends from 2006-2016?
5. Which year or month has the most crimes? Are the crimes related to seasons or weather?
6. How crimes correlate to demographic and population data geographically?

Data Source

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

Critical Parameters

- "RPT_DT": Date event was reported to police
- "KY_CD": Three digit offense classification code
- "CRM_ATPT_CPTD_CD": Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
- "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

Preliminary data cleaning will include removing empty data (NaN in start and closed dates, city, crimes, lat and long). In addition, dates and times will be reformatted.

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 Engineering

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

Q1. Did NYPD public access complaint data have a quantifiable impact on improving government management and reducing crimes?

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

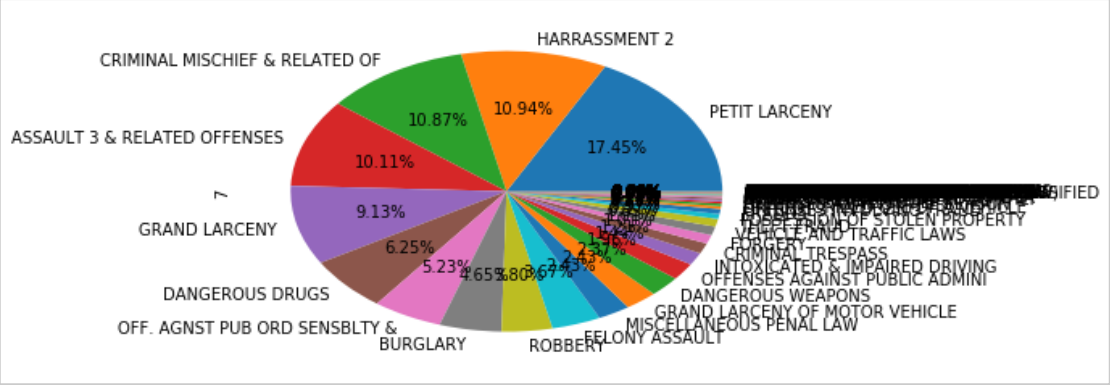
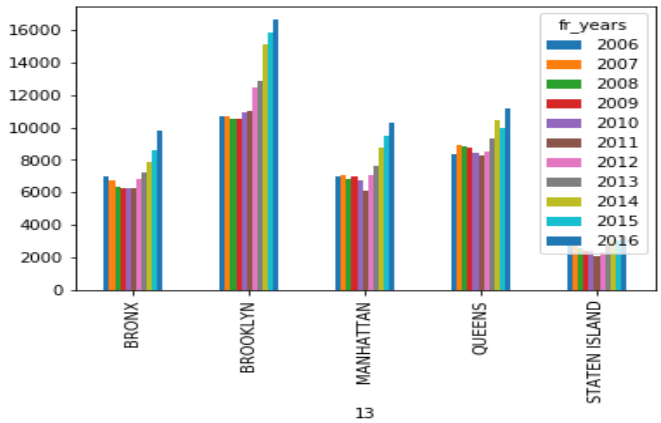


Figure 1. NYPD crimes statistics by each category between 2006 and 2016

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.

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

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. What are the most frequent crimes?
Q4. Does crime fluctuate over time? What are the crime trends from 2006-2016?

Q5. Which year or month has the most crimes? Are the crimes related to seasons or weather?
Q7. How crimes correlate to demographic and population data geographically?