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 the end of last year (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 do 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.

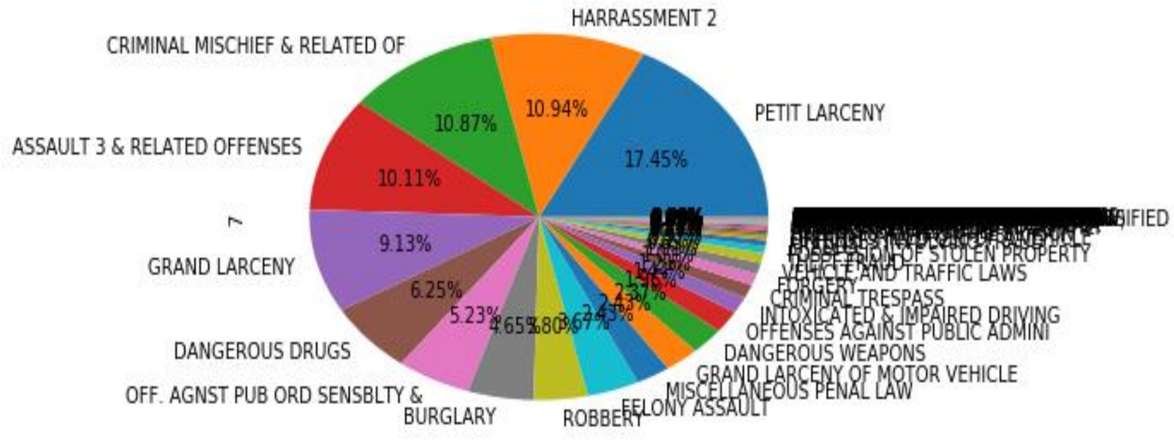
Supplemental Datasets:

- 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

Approach:

We’ll perform row engineering and feature engineering and clean the data first. The preliminary analysis on the crime types gave us the frequent crimes over the time.

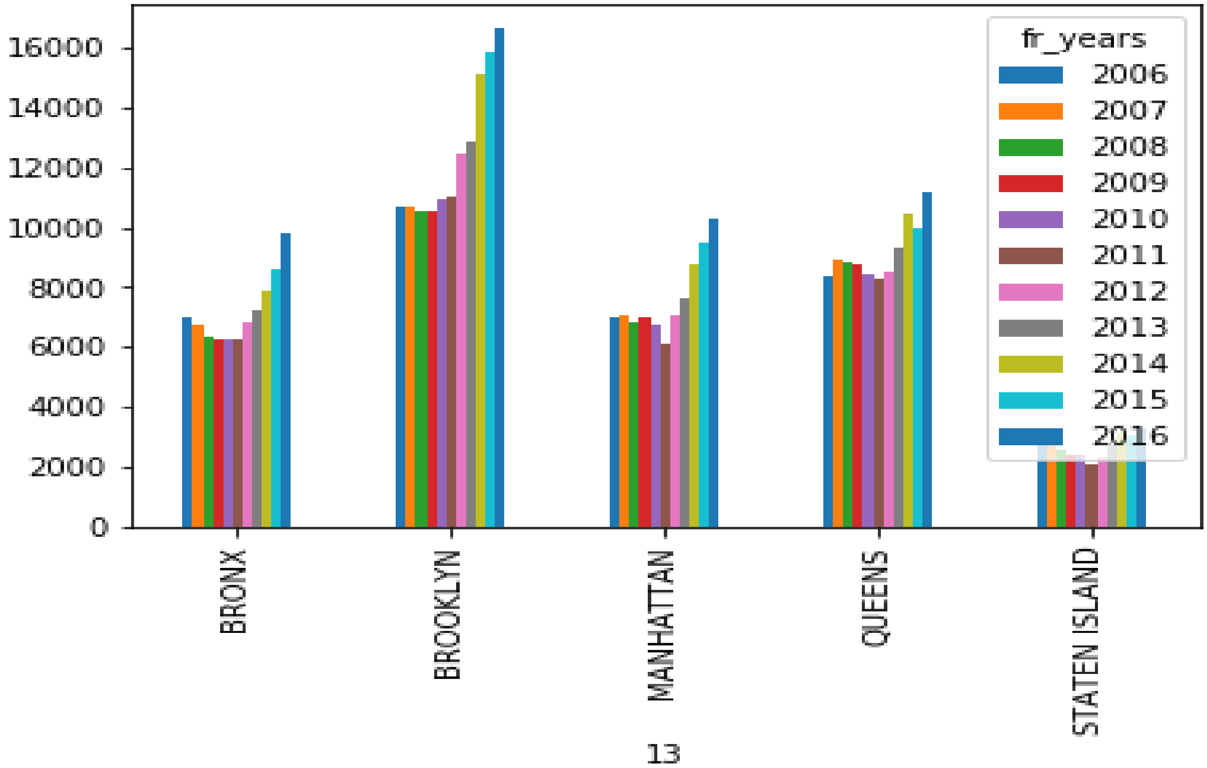
```
The most frequent crimes between 2006 - 2016
PETIT LARCENY                               688005
HARRASSMENT 2                               431275
CRIMINAL MISCHIEF & RELATED OF              428688
ASSAULT 3 & RELATED OFFENSES                398708
GRAND LARCENY                               359811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



We also performed analysis on the most frequent crimes over the years and regions to find any correlations.

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

```
 'HARRASSMENT 2' crimes in each city with respective year
```



Specific crime analysis results will be integrated into the demographic and population data over regions to determine if there is any significant correlation. More analysis will be coming and we expect to gain some insights and answer our questions. We are looking forward to final reports.