NYU Crime data historic summary

Luying Yan
Chang Liu
Seong Woo Han

Abstract

This report presents statistics on NYU crime data based on NYC OpenData, which includes all kinds of crimes such as violations, misdemeanor and felony from 2006 to the end of 2015. This dataset contains 5.1 million rows of data with 24 columns. In the following section, we first give a brief description about each column and then analysis on the quality issue on this data, and finally summarize based on different aspect, such as time distribution, geographical distribution and so on. Finally, we generate a general idea about the crime in New York City.

Table of Contents

Introduction

This report presents statistics on NYU crime data based on NYC OpenData, which includes all kinds of crimes from 2006 to the end of 2015. The summary is organized as follows: In Section 1, we discuss the statistical consideration, in other words, the semantic meaning on each columns. In section 2, we elaborate in detail on how we manage to data cleaning. This part includes validation such as date time format check, date range check, crime code number check, borough code check and so on. In section 3, we generate data based on cleaned data and extract features in many aspects, and draw to conclusion on each feature.

The crime rate trend is similar among different years. Among all the boroughs in New York, Brooklyn is the top one who has the highest crime rate overall. The geographical distribution of different type of crime vary from each other, but there are some locations almost every type crime occurs frequently, such as Midtown of Manhattan, JFK airport and southern part of Bronx. And crimes are less likely to happen in southern west of Brooklyn and Queens. In addition, we listed frequent crime scenes as well as parks that crime rate are high.

This summary should give abundant information. People should be aware of these crimes, and make precautions such as avoid walking through parks during midnight. All in all, do not try to commit a crime. Make New York a better place to live.

Statistical Consideration

The data we use in this project is downloaded from NYC OpenData, and the information is accurate, and is considered to be a close approximation of current record.

The incident that covers multiple categories of crime is considered as the most serious one, so there won't be multiple records on the same event.

The crime date duration may meet the following three categories: 1. With both From Datetime as well as To Datetime, which means that the data is accurate and the time range is exactly the duration of the crime. 2. If the To Datetime column is empty, this means the exact time the event was reported to occur, however the finish time is unknown. 3. Only the To Datetime exists, this means that there is only a known endpoint to this crime. There is not relation between Report Date with To Datetime, which means the event may occurred years before it is reported to police. On the other hand, the event may finish after it is recorded in the NYPD database. The time records in this database are recorded on a 24-hour clock.

As for the GPS coordinates, not all the records have this information. The data can be categorized into two categories: 1. Crimes involve victims such as rape, sex offenses 2. Other misleading data. The reason that generate second category may due to various reasons, such as invalid street address, that cannot be able to pinpoint the location. However, information of the nearest police station is recorded.

There are some data inconsistency in this dataset, but only a little portion. The vast majority of the data is accurate and is filled in each column. The inaccuracy may result from manually transcription when people loading this data into the database. We will discuss in the next section about data cleaning.

Data Quality Issues

In this section, we describe how we clean the data from NYC OpenData, we aim to distinguish the type of data in csv file, i.e. whether they are valid, invalid, and null, respectively. The null data is easy to detect, so we focus on how to differentiate between valid and invalid data.

Through all 24 columns, we did semantic analysis on multiple columns, since not all columns could be tracked of its accuracy. The following columns are the portion that we did data clean on.

The first column is persistent ID for complaint, if the ID is unique, then it can be recognized as a valid data, otherwise, it is an invalid data. Duplicity is not allowed here, since each ID indicated only exactly one event.

The second column to the sixth column, stores date time information. In this case, we should not only check whether an entry is represented using a valid date or time format, but also check whether the exact date and time of occurrence is before the ending date and time of occurrence, and whether the exact date of occurrence is before the date of report. The example for the first circumstance: 12/04/1015, 24:00:00, we regard 00:00:00 as valid, and 24:00:00 as invalid. The example of the second circumstance: The report date is 05/14/2006 while the date of exact occurrence is 05/19/2006. In this case, we consider it an invalid record.

From the seventh column to the tenth column are offence codes corresponding with its descriptions. So there we only need to check if there is a one-to-one relationship between seventh and eighth row, as well as ninth and tenth row. As the data collected, the inaccuracy may result from three categories: 1. Redundancy records of description of the same code. 2. Inaccuracy of lacking description of certain offence code. 3. Similar representation of the same

offence code, but not identical. The example for circumstance 1: ['OFFENSES RELATED TO CHILDREN', 'ENDAN WELFARE INCOMP'], they both take up the code 345. In this case we cannot decide which is the correct relationship between the code and description. For circumstance 2: ['GAMBLING', ''], which means that there are rows that lack description, and we can fix this by replace the null value with the correct corresponding description. Finally, as for circumstance 3: ['KIDNAPPING & RELATED OFFENSES', 'KIDNAPPING', 'KIDNAPPING AND RELATED OFFENSES'], this kind of inconsistency could be solved by merging this cluster, choosing either name to be the representative name.

The date of eleventh column describes the indicator of whether was successfully completed or not. Data in this column are either valid or null because the value are only “COMPLETED”, “ATTEMPTED”, “FAILED” or “INTERRUPTED”. By omitting empty ones and the rest are cleaned.

The data of twelfth column which indicates the level of offence are all valid since the value are confined to “FELONY”, “MISDEMEANOR” or “VIOLATION”, with no invalid data.

The following seven columns are related to the area of occurrence, except the fifteenth column, which is about precincts, the data of other column are either valid or null. As for precincts, we should check whether the corresponding borough (i.e. fourteenth column) has this precinct, if not, then it can be known as an invalid data. As for the sixteenth column, it should only considered meaningful only when the seventeenth column’s value exists, otherwise, we mark this row as invalid.

Regarding the last five columns, which are all relevant to the concrete locations, and the data are either valid or null as well since the location is within the New York east area.

Data Summary

In this section, we summarize on the cleaned data, and try to see some trends and draw some conclusions based on the figures that we plot.

Borough Distribution

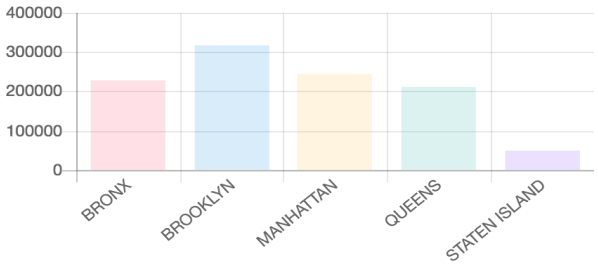


Figure 1 Borough Bar

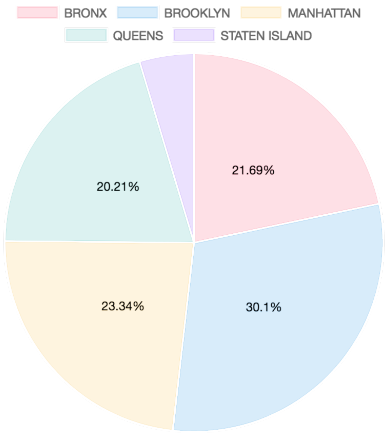


Figure2 1Borough Pie

As shown in the figures 1 and 2, crime happens most in Brooklyn, which takes up 30.1% of total crime. Manhattan comes second, occupies 23.34%. Bronx and Queens are quite similar, both take up around one fifth of total crime. Staten island has the lowest crime rate, and it is relatively the safest borough in New York.

Table 1Borough Table

Borough	Frequency
BRONX	227477
BROOKLYN	315648
MANHATTAN	244749
QUEENS	211957
STATEN ISLAND	48743

Geographical Distribution

here, we collect data of certain typical type of crime, and plot distribution of these crime with google maps. By using heat map, we can see the density variation among different types of crime as well as difference between different precinct for certain type of crime.

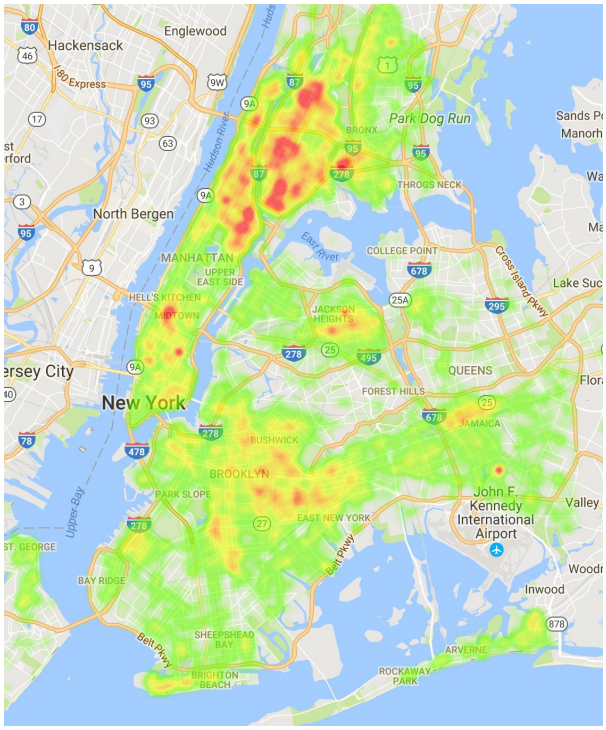


Figure 2 Robbery Heat Map

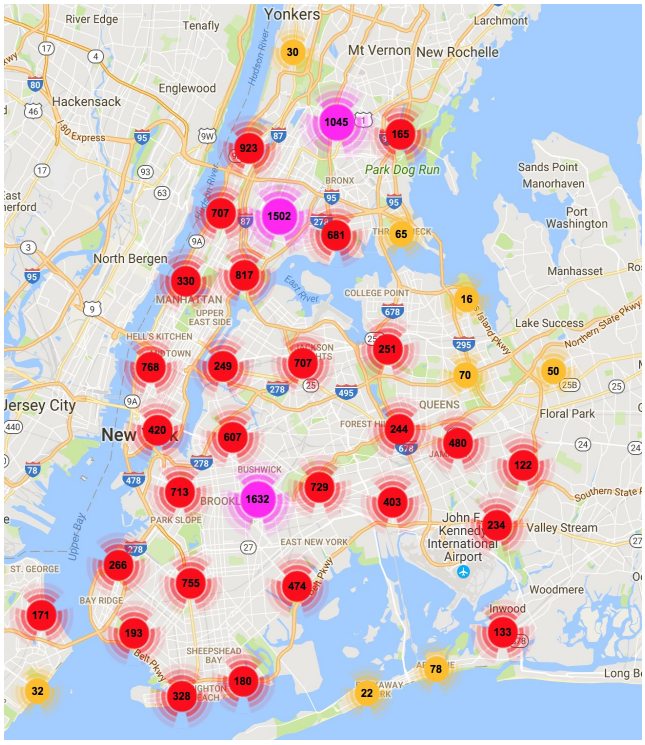


Figure 3 Robbery Heat Map

Here we draw a heat map as well as a cluster map to show the robbery distribution during year 2015. Robbery rate is relatively high in the east part of Brooklyn, and higher in Manhattan midtown, the highest in the west part of Bronx and upper east of Manhattan. Also, the density map shows a high rate in Jackson height as well as JFK airport.

On the other hand, though the density near Jackson heights and JFK are high, the amount of crime is not as much as Bronx and Upper Manhattan. Instead, in Brooklyn Bushwick, the total amount of crime is relatively high, as is shown in the cluster map. Compared to the north east part of Brooklyn, southern west part is relatively safe and the crime is more sparsely separated. Also, downtown and is safer than midtown as well as uptown in Manhattan.

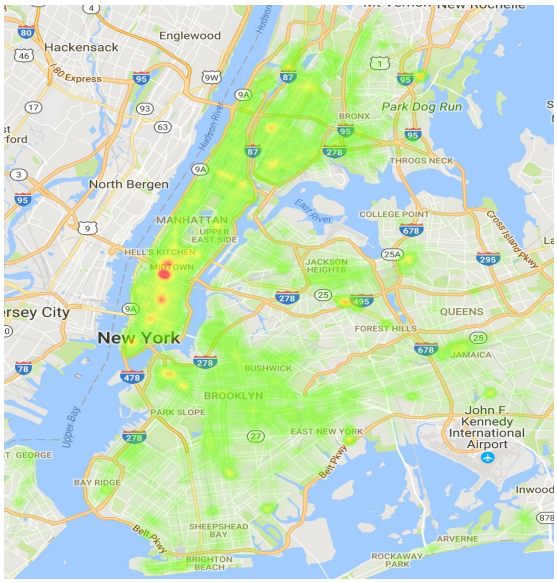


Figure 4Larceny Heat Map

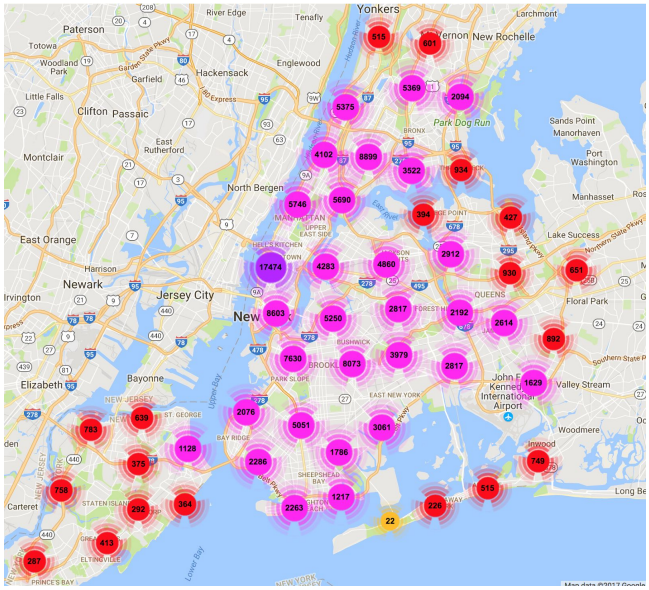
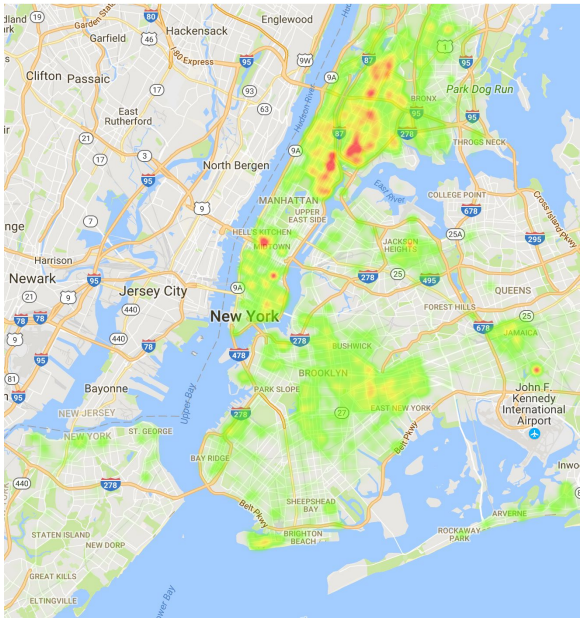


Figure 5 Larceny Cluster Map

As shown in the map, larceny is way much than the frequency of robbery and is mainly located in Manhattan and Brooklyn. At midtown of Manhattan reaches the highest (reaches 17474). It also remains high in upper of Brooklyn (over 5000 in each cluster, some reaches 8000). The larceny rate keep low in Staten island as well as the extended part of Queens. Also, the JFK remain dense and has 1629 larceny take place.

Weapon:



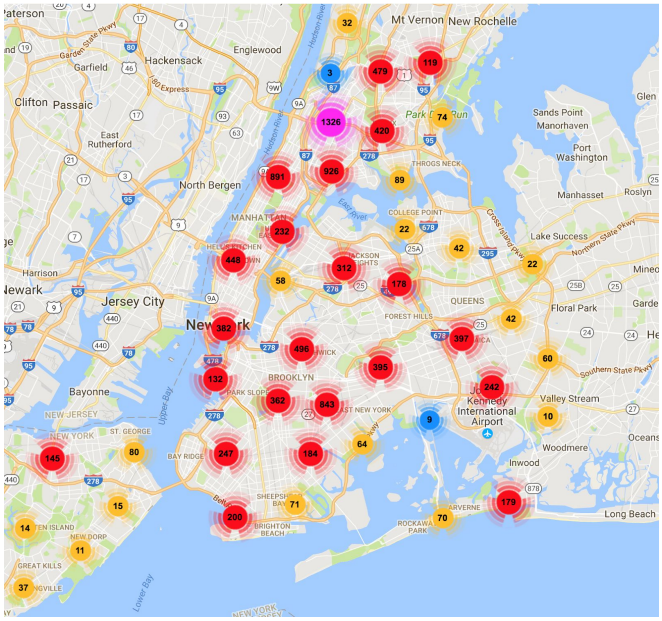


Figure 6 Weapon Heap Map

Figure 7 Weapon Cluster Map

As for the weapon category, it is high in the upper Manhattan and lower Bronx.

It is relatively low in southern west of Brooklyn as well as in Staten Island. The heat map indicates that midtown and uptown of Manhattan and southern part of Bronx have high ratio of weapon crimes, and southern west part of Bronx reaches highest frequency (1326). The east part of Brooklyn comes the second, it has the largest cluster among Brooklyn (843), however it is widely spread and not as dense as in upper Manhattan.

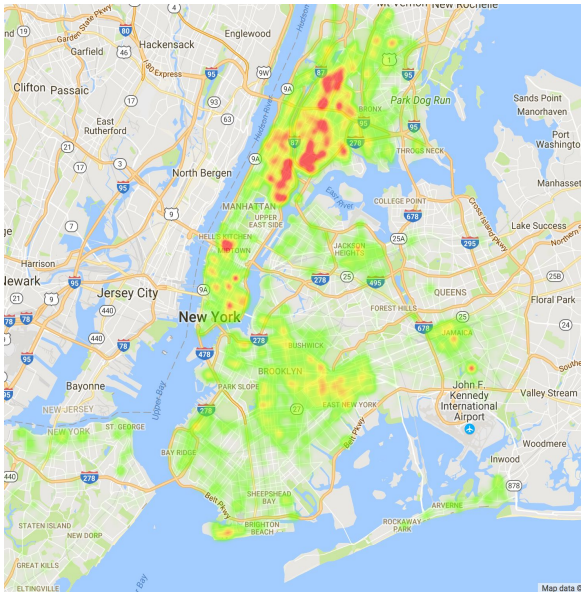


Figure 8 Drug Heat Map

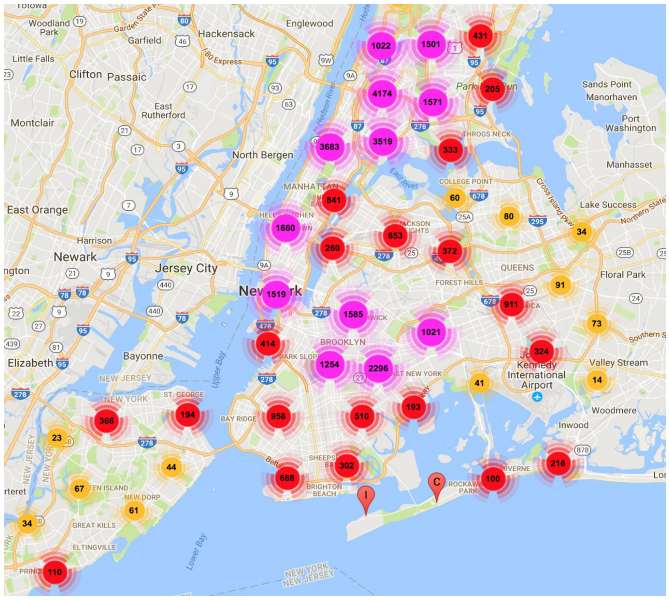


Figure 9 Drug Cluster Map

Finally, as for the drugs distribution, upper Manhattan would be the highest.

Conclusion, though midtown in Manhattan may not be the highest among all kinds of crime, however, it is one of the densest place where crimes take place. Other than that, we can safely conclude that upper Manhattan as well as southern Bronx are most dangerous of many kinds of crimes. Though Brooklyn has a very high crime frequency among all borough, however, crimes are committed more sparsely compared to Manhattan.

Overall, southern west of Brooklyn, and Queen may be considered as relatively safer place.

Premises Distribution

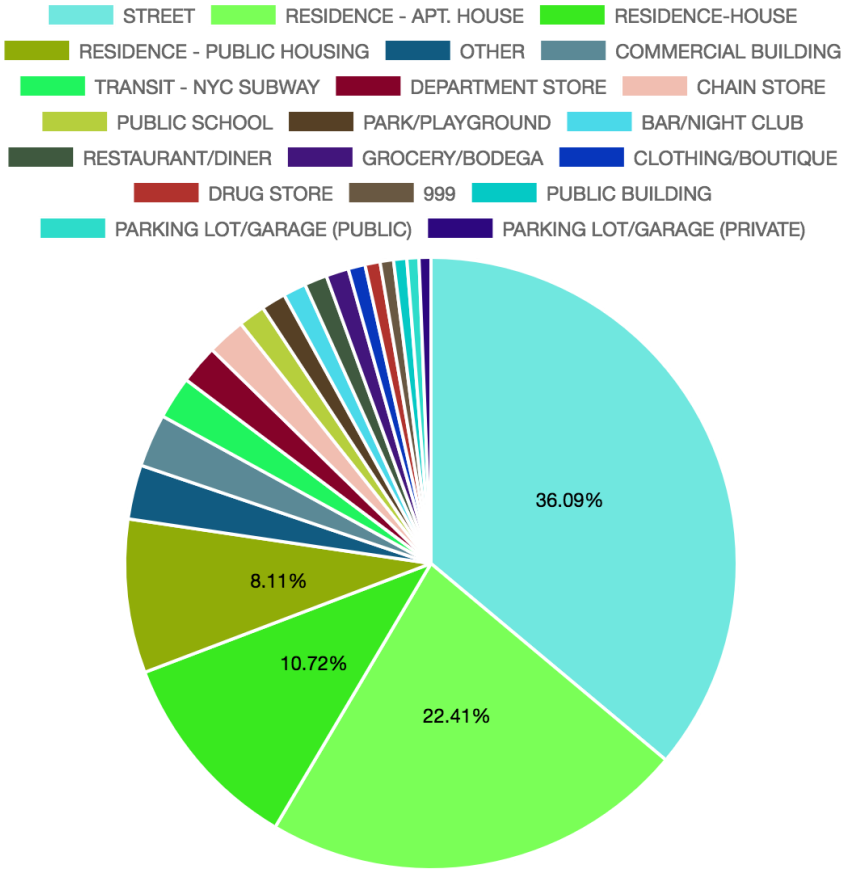


Figure 20 Premises Distribution

Figure 20 shows the top 20 frequent places where crime usually takes place. Outside crime that are committed on the street takes up around 36% of the top 20 crimes in total. Inside crime that related residence combine takes up around 40% of the data collected, which indicate that resident crime are relatively high, people should be aware of danger even at home.

Frequent Parks

Figure 19 Frequent Park Distribution

As shown in the above picture, crimes that take place in a park only occupies a tiny fraction in the total crime data. There listed the top ten frequent parks that crime is committed. Central Park is the most dangerous one, which is even larger than the combination of the second and third together. The high crime ratio may result from its geographical location. It is located in a high crime rate precinct near upper Manhattan. We come up this solution by referencing the criminal geographical distribution figures in the previous section.

The second highest is the Flushing Meadows Corona Park, which is located in Queens. The third is the Riverside Park located in Manhattan near Central Park.

The data seems randomly spread across New York, and generally there is more tendency that crime happens in the upper of New York, where crimes that take place in a park happens.

In addition, Washington Square park is the tenth highest, with around 100 records in total. Students from NYU should consider side walk instead of crossing the park in late night.

GPS Range

Table 2 NY SPCS Coordinate Table

New York State Plane Coordinate System	From	To
X_coordinates	913319.0	1067298.0
Y_coordinates	124364.0	208735.0

Table 3 GCS Coordinate Table

Global Coordinate System	From	To
Latitude	40.507750107	40.739224294
Longitude	74.255075543	73.700315857

We gather Geographical information by extracting the last five columns. The data collected are in the above coordinate range.

Average Hour Distribution

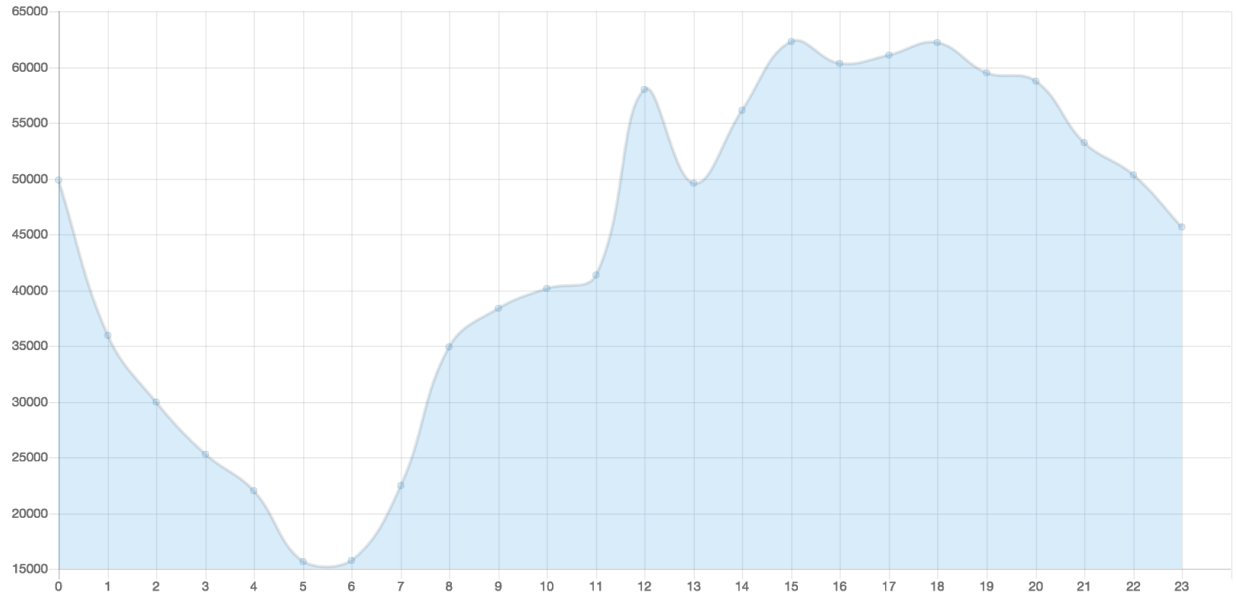


Figure 10 Crime Hourly Distribution

As shown in the figure above, the crime rate rises in the morning (from 6am to 12am), and it ascends sharply from 6am to 8am. From 8am to 11, the crime rate rises slowly, and it reaches a peak at noon, later slightly drop down at 1pm, and then start growing again. The crime rate keeps constant high from 3pm to 7pm, and later the number of crime drops down as the day goes dark. There is a sudden rise at 12am, then it continues to drop. At some time between 5am and 6 am, crime rate reaches the lowest point (around15000). This pattern usually follows daily pace. We assume that criminals follow similar pattern as normal people do during the day time. We can explain the drop at 1pm since that's the time when people usually take a nap in order to get productive during the afternoon, and the crime rate drop down from 12am to 5am can be interpreted as people are falling asleep during midnight.

Monthly Distribution

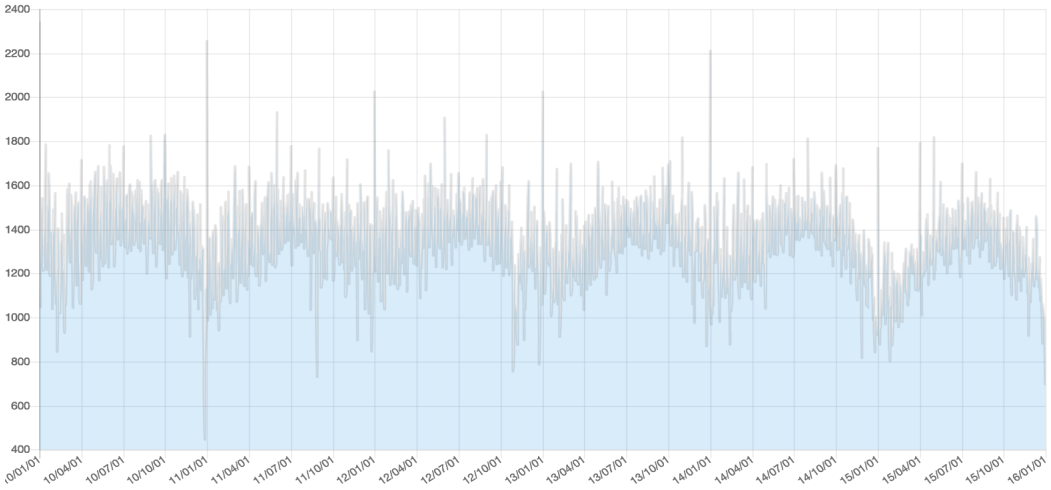


Figure 11 Crime Daily Distribution

The crime rate falls to a similar pattern in each year. The crime rate always reaches its peak at the first day of this year. Then it grows gradually until some time among June, and remain constant high until October, then it begins to drop to the end of this year.

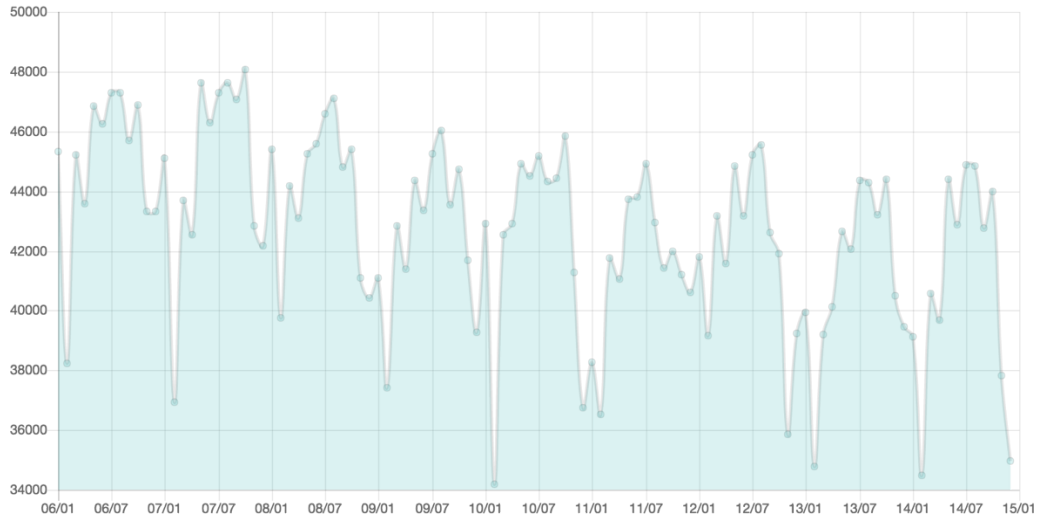


Figure 12 Crime Monthly Distribution

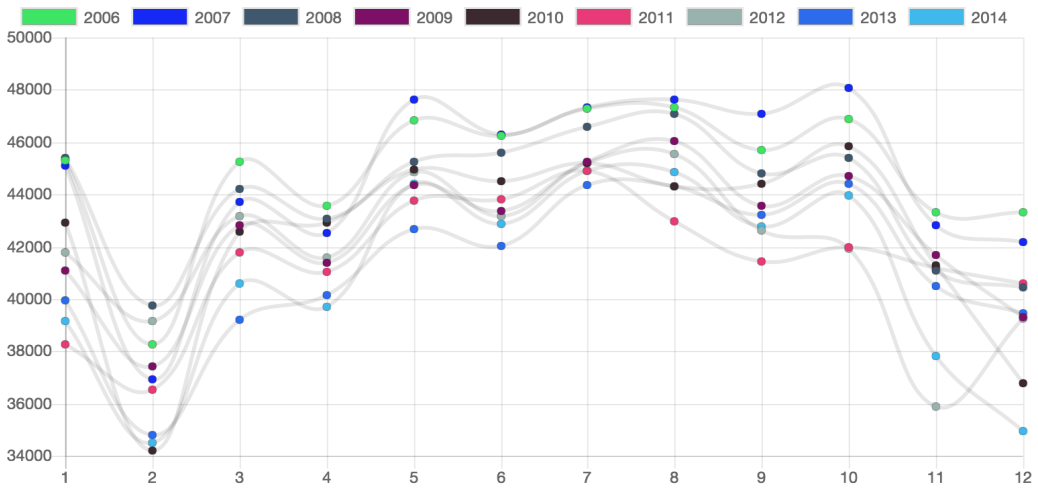


Figure 13 Crime Monthly Comparison

The data described crime rate trends from Jan 2007 to Dec 2015. We can see similar patterns in each year, despite that the amount of crime may vary.

Crime rate is lowest in February, and rise March, then drop down a little bit in April. Then the crime comes to its prosperity, at May it reaches to a peak, and during the following month remain relatively high. The crime rate drops in November and remains low in December as well.

The peak among those years take place at October 2006, and drops to its lowest level at February 2015.

Severity Distribution

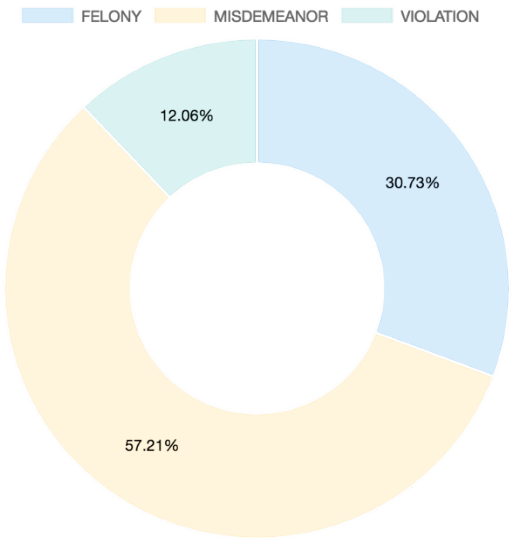


Figure 14 Severity Distribution

As seen on the figure, over 50% of crime are misdemeanor, which involve crimes like: mischief, assault, petit larceny, and light offences. Felony takes up around 30%, which involve crimes like: grand larceny, burglary, dangerous weapons robbery and all. Only around 10% of crimes are violations, where most of it being harassment.

Not Geocoded Data Distribution

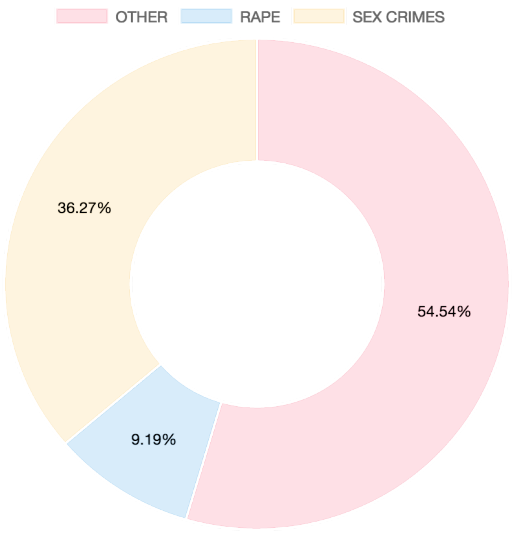


Figure 15 Not Geocoded Distribution

As shown in the figure above, over half of the record that are not geocoded are due to inaccurate address information, as described in a related document, which means there still need improvement of accuracy when filling crime information. Other than that, sex crime occupies around 36% of not geocoded data, and rape takes up less than 10%.

Crime Status Distribution

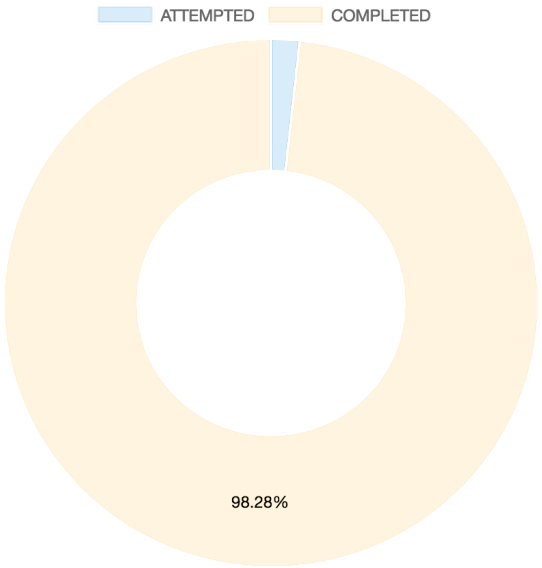


Figure 16 Crime Status Distribution

As shown in the figure above, over 98% of data are labeled completed, and the rest are labeled as attempted. There is no other category in the dataset. The assumption is that people tend to report completed crime to the police station instead of uncompleted or failed ones.

Frequent Offence Classification

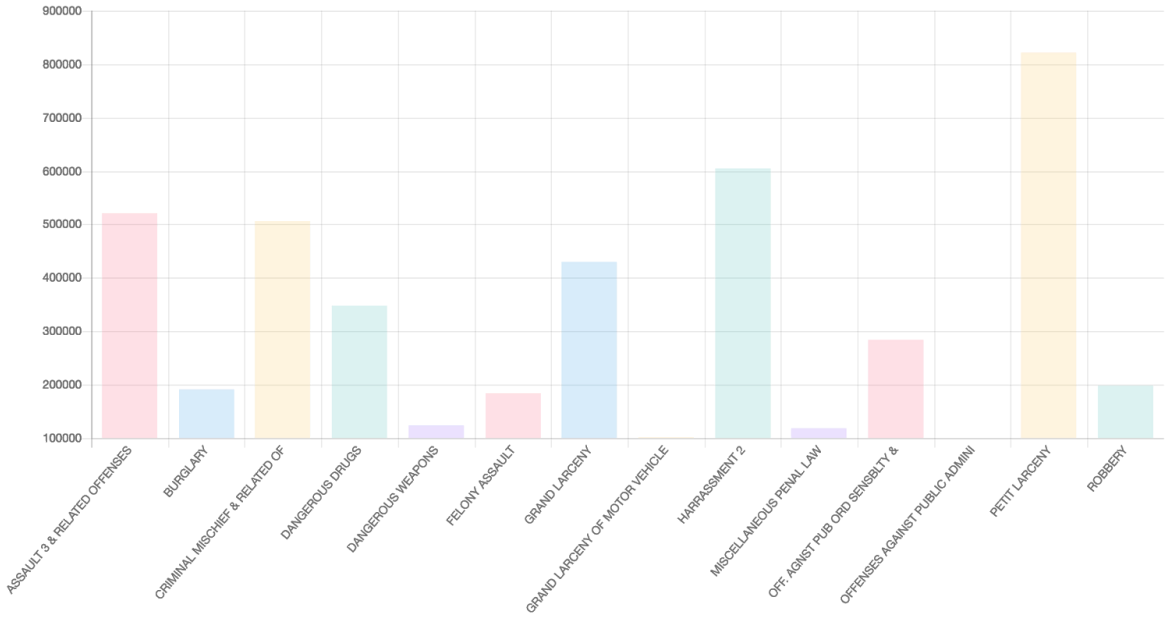


Figure 17 Frequent Offence Classification Bar

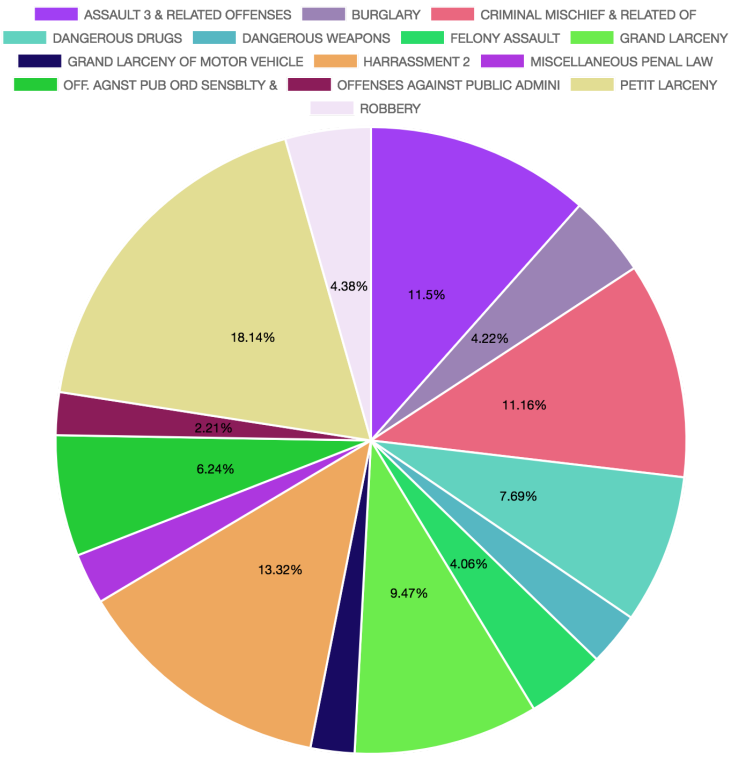


Figure 18 Frequent Offence Classification Pie

Figure above list frequent crime types whose frequency is higher than certain threshold based on whole dataset, which we treat as frequent if the frequency is above 100000. As shown in the figure, Petit larceny comes the first, take up around 18% of crime data collected, and its frequency is over 800000. Harassment comes the second, occupying around 13%, and its frequency is over 600000. Assault 3 and criminal mischief comes the third and fourth, which each take up around 10%, around 500000 each.

Summary

In this project, we did data cleaning and analysis on the NYC crime dataset. By digging into this dataset, we generate the trend of crimes in New York and it is meaningful for future reference, and give predictions for the coming year. The data is already cleaned before it is recorded into

the database, however there are still inaccuracy by human mistake. We generate data cleaning strategies on each column which are mentioned in Data Quality Issue section.

Based on the cleaned data, we extracted several features in terms of geographical distribution, date and time, and other aspects such as crime type. The geographical part is relatively intuitive and give more information of the distribution of certain type of crime throughout New York. By checking the heat map on selected type of crimes, we can conclude that Midtown Manhattan, uptown Manhattan, southern part of Bronx, as well as east part of Brooklyn are considered as more dangerous place compared to other places. Queens and southern west part of Brooklyn are considered as relatively safer, in terms of crime frequency as well as density. In addition, though crime frequency at JFK airport is not high, but the density is high, where we consider also a dangerous place.

Also, crimes are most committed on the street, but residence related crime combine is equal as high as on the street, which mean people should take care of themselves even when at home.

This project provide hands on experiment on real data, and we encounter inaccuracy and discussed how to properly handle them. This is a great opportunity to hone programming skills on Big data programming.

Reference

1. Year End 2010 Enforcement Report – NYC
   http://www.nyc.gov/html/nypd/downloads/pdf/analysis_and_planning/yearend2010enforcementreport.pdf
2. Spark Programing Guide
   http://spark.apache.org/docs/latest/programming-guide.html
3. Chartjs API Documentation
   http://www.chartjs.org/docs/