# NYC Crime (Big Bang Data Group)

Duc Tran
dt2259@nyu.edu

Aigo Madakimova
am9634@nyu.edu

Bornita Das
bd1599@nyu.edu

## ABSTRACT

Crime in the U.S took a drastic change during 2020 at the event of COVID-19 pandemic, the movement of Black Live Matter and the presidential election. With interest to investigate the statistics and trends of crime activities during this unprecedented time, our project focus on data analysis of crime happens in New York City, one of the biggest city in the country. The primary purpose is to look for details within data set to extract meaningful information that can be use to prevent crime in case these events happen again in the future.

**ACM Reference Format:**
Duc Tran, Aigo Madakimova, and Bornita Das. 2021. NYC Crime (Big Bang Data Group). In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The goal of the project is to extract meaningful information from data set related to crime in New York City. To accomplish this, we perform various steps from data exploration and data study to data cleaning and data analysis. For data exploration, we searched through various data libraries. Our group performed data study and cleaning through tools such as OpenRefine by Google and OpenClean by NYU-VIDA. Our current report has only information related to data cleaning, our next step in the project is to perform data analysis to answer questions such as: How many crime based on different offense is committed in 2020? What are the statistic for age and race of the person being arrested for certain offense? Where are the areas that crime happens the most in NYC?

## 2 DATA OVERVIEW

The NYPD Arrest(Year_To_Date) Dataset is a breakdown of every arrest in NYC by NYPD in 2020. The data is manually extracted every quarter and reviewed by Office of Management Analysis and Planning. Each record has the 140413 rows and 19 columns which are listed below. The dataset can be found in https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/uip8-fykc

```
Schema
------
  'ARREST_KEY'
  'ARREST_DATE'
  'PD_CD'
  'PD_DESC'
  'KY_CD'
  'OFNS_DESC'
  'LAW_CODE'
  'LAW_CAT_CD'
  'ARREST_BORO'
  'ARREST_PRECINCT'
  'JURISDICTION_CODE'
  'AGE_GROUP'
  'PERP_SEX'
  'PERP_RACE'
  'X_COORD_CD'
  'Y_COORD_CD'
  'Latitude'
  'Longitude'
  'New Georeferenced Column'
```

Some columns that our project would like to focus on are: ARREST_DATE, PD_DESC, OFNS_DESC, ARREST_BORO, ARREST_PRECINCT, AGE_GROUP, PERP_SEX, PERP_RACE, Latitude and Longitude. These columns fit with the information we want to extract from the data set.

## 3 RESEARCH METHODS

### 3.1 Data exploration

To find the right data set, we searched through multiple data libraries such as: Kaggle, auctus.vida-nyu, Google data and opendatanetwork. The current data set is found in opendatanetwork.com and is also found in all these other data library platforms. Although there are many NYC crime related data, our team decided to work on NYPD Arrest data set because we are interested on severe crime that involves police.

### 3.2 Data study

Our first approach with understanding the data is through, OpenRefine which helps us cluster data with similarity in columns and provide some overview structure of the data. Later, we start using OpenClean which provided lots of interesting features to support our study. Some of the interesting details we found are:


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

| | total | empty | distinct | uniqueness | entropy |
|---|---|---|---|---|---|
| ARREST_KEY | 10000 | 0 | 10000 | 1.000000 | 13.287712 |
| ARREST_DATE | 10000 | 0 | 366 | 0.036600 | 8.396583 |
| PD_CD | 10000 | 1 | 166 | 0.016602 | 5.380796 |
| PD_DESC | 10000 | 2 | 158 | 0.015803 | 5.349247 |
| KY_CD | 10000 | 2 | 59 | 0.005901 | 4.501480 |
| OFNS_DESC | 10000 | 2 | 51 | 0.005101 | 4.290129 |
| LAW_CODE | 10000 | 0 | 421 | 0.042100 | 6.245291 |
| LAW_CAT_CD | 10000 | 108 | 4 | 0.000404 | 1.043776 |
| ARREST_BORO | 10000 | 0 | 5 | 0.000500 | 2.166298 |
| ARREST_PRECINCT | 10000 | 0 | 77 | 0.007700 | 6.113753 |
| JURISDICTION_CODE | 10000 | 0 | 21 | 0.002100 | 0.636239 |
| AGE_GROUP | 10000 | 0 | 5 | 0.000500 | 1.679094 |
| PERP_SEX | 10000 | 0 | 2 | 0.000200 | 0.643794 |
| PERP_RACE | 10000 | 0 | 7 | 0.000700 | 1.926892 |
| X_COORD_CD | 10000 | 0 | 5338 | 0.533800 | 11.349127 |
| Y_COORD_CD | 10000 | 0 | 5412 | 0.541200 | 11.367285 |
| Latitude | 10000 | 0 | 5574 | 0.557400 | 11.412375 |
| Longitude | 10000 | 0 | 5574 | 0.557400 | 11.412375 |
| New Georeferenced Column | 10000 | 0 | 5574 | 0.557400 | 11.412375 |

According to the table through sample of 10000 data, the ARREST_KEY is in correct structure with 100% uniqueness, as well as the distribution of X and Y_COORD_CD, Latitude and Longitude columns. There some blank values in LAW_CAT_CD, which need further study to see if it makes a strong impact on the analysis.
Overall the data looks fine but when we dig deeper, we found lots of spelling mistake across multiple columns. There are many more but we only listed a few here. Please check our Google Colab files for full details:

```
KIDNAPPING & RELATED OFFENSES
KIDNAPPING AND RELATED OFFENSES
KIDNAPPING

INTOXICATED & IMPAIRED DRIVING
INTOXICATED/IMPAIRED DRIVING
```

In another study, we found some columns use abbreviation that results in ambiguity. Values in columns LAW_CAT_CD showns:

```
F
M

I
V
```

We also ran a test on ARREST_DATE to check if the data is in correct format and we didn't find any issue. The rest of our study focus on some visualization to better understand the data and what it can offer. For example to deal with complex data that we don't understand fully such as longitude and latitude, we created graph HEATMAP of all geolocation data points to see if there are any weird location that is not in NYC.

### 3.3 Data cleaning

For data cleaning, we work with OpenClean, an open-source data cleaning tools and visualization created by NYU-VIDA. Majority of our clean related to spelling that located in multiple columns across the data. Here are some changes that we have done for the data:

- In order to fix the different representations discovered through clustering of same item in OFNS_DESC column we updated the spelling.
  - Changed KIDNAPPING and KIDNAPPING RELATED OFFENSES to KIDNAPPING AND RELATED OFFENSES
  - Changed ADMINISTRATIVE CODES to ADMINISTRATIVE CODE.
  - Changed INTOXICATED IMPAIRED DRIVING, INTOXICATED/IMPAIRED DRIVING to INTOXICATED AND IMPAIRED DRIVING.
- To make the data more interpretive we modified abbreviation of some columns
  - Changed ARREST_BORO to full name of boroughs.
  - Changed PERP_SEX to Male, Female
  - Changed LAW_CAT_CD to full descriptions.
- In order to fix the different representations, we remove space between for PREP_RACE column to ASIAN / PACIFIC ISLANDER.
- In order to fix the different representations discovered through clustering of same item in PD_DESC column we updated the spelling.
  - Standardized all variants of POSSESSION spelling.
  - Removed extra spaces in SALE
  - Standardized all variants of DRUG spelling.

## 4 CHALLENGE

There are many challenges that we faced while working with the data set. The size of the data sets being massive our personal laptop could not handle the computation, therefore picking the right data set to fit our needs can be difficult. After obtaining the data, we spend tremendous amount of time to study, ask ourselves what we would like to learn from the data so that our data cleaning step fits with our goals and purposes. Through the introduction of OpenClean, we could use this powerful tool to study the data through set of features it provides to us such as profiling, sampling and visualizing portions of the data. The data cleaning step becomes a bit easier, however there are some gray areas we still don't know much about such as the relationship between key code (KY_CD) and offense description (OFNS_DESC), these 2 columns have data that does not connect. For example: with offense ROBBERY key code can be listed as 109 or 105 and not unique only to the description. Although this is an issue, we decide to ignore the KY_CD as we can't find official data on what they are represented. Our goal is to categorize OFNS_DESC as it tell us the most what the crime is about. The biggest challenge we face is to learn about OpenClean and what it can offer. Sometimes it can be frustrating to understand the package library and syntax. Luckily the documentation of functionalities is very well written which helped us execute our queries.

## 5 ONLINE RESOURCES

**Data set:**
https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/uip8-fykc

**Open clean documents:** https://openclean.readthedocs.io/

## 6 GITHUB

https://github.com/duketran1996/NYC-Crime