**Sheng Liu (sl5924), Peimeng Sui (ps3336), Xiaoyu Wang(xw1435)**

**Instructor: Prof. Juliana Freire**

**DS-GA 1004 Big Data**

## NYC Incidents Dataset Summary Report

In this report, we will focus on data quality summary and our procedure of cleaning the NYC incidents dataset. This dataset, available on NYC Open Data website, 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 (2015). Our code for the procedures mentioned in this report is available here on our Github repo: https://github.com/peimengsui/BigDataProject .

## Part I Data Summary

We include information about name, description, base type, semantic type, missing label and validation for each of the 24 variables in the following data summary table. You can follow the instruction in our Github repo to run the code to automatically generate these information for each cell value of the dataset.

### Data Summary Table

| Field Name | Description | Base Type | Semantic Type | Missing Label | Invalid/Outlier |
|---|---|---|---|---|---|
| CMPLNT_NUM | Randomly generated persistent ID for each complaint | INT | Unique ID as Primary Key | No missing value | No invalid value or outlier detected. |
| CMPLNT_FR_DT | Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists) | Datetime | Date | 655 Missing values labeled as NaN | 7 Invalid Year less than 1800, 31 invalid from_date later than to_date, 2 invalid from_date later than report_date, 40 in total, excluded |
| CMPLNT_FR_ | Exact time of | Datetime | Time | 48 Missing | 903 invalid value |

| | | | | | |
|---|---|---|---|---|---|
| TM | occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_ TM exists) | | | values labeled as NaN | 24:00:00, excluded. |
| CMPLNT_TO_ DT | Ending date of occurrence for the reported event, if exact time of occurrence is unknown | Datetime | Date | 1391478 Missing values labeled as NaN | 1 Invalid Year greater than 2020 and 31 invalid to_date before from_date, 32 in total excluded. |
| CMPLNT_TO_ TM | Ending time of occurrence for the reported event, if exact time of occurrence is unknown | Datetime | Time | 1387785 Missing values labeled as NaN | 1376 invalid values 24:00:00, excluded |
| RPT_DT | Date event was reported to police | Datetime | Date | No missing value | 2 invalid value before from_date, excluded. |
| KY_CD | Three digit offense classification code | INT | 74 Different Classification Code | No missing value | No invalid value or outlier detected. |
| OFNS_DESC | Description of offense corresponding with key code | TEXT | 71 different descriptions corresponding to classification code | 18840 missing values labeled as NaN. | No invalid value or outlier detected |
| PD_CD | Three digit internal classification code (more granular than Key Code) | FLOAT | 416 more granular internal classification code | 4574 missing values labeled as NaN. | No invalid value or outlier detected. |
| PD_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) | TEXT | 404 different descriptions corresponding to more granular classification code | 4574 missing values labeled as NaN. | No invalid value or outlier detected. |

| | | | | | |
|---|---|---|---|---|---|
| CRM_ATPT_CPTD_CD | Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely | TEXT | Indicator of whether the crime completed | 7 missing values labeled as NaN | No invalid value or outlier detected. |
| LAW_CAT_CD | Level of offense: felony, misdemeanor, violation | TEXT | Category with 3 classes | No missing value | No invalid value or outlier detected. |
| JURIS_DESC | Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc. | TEXT | Category with 25 classes | No missing value | No invalid value or outlier detected. |
| BORO_NM | The name of the borough in which the incident occurred | TEXT | Category with 5 classes | 463 missing values | 17 inconsistent data detected, borough doesn’t correspond to precinct |
| ADDR_PCT_CD | The precinct in which the incident occurred | FLOAT | Category with 77 unique classes | 390 missing values | 17 inconsistent data detected, borough doesn’t correspond to precinct |
| LOC_OF_OCCUR_DESC | Specific location of occurrence in or around the premises | TEXT | Category with 5 unique classes | 1127128 missing values denoted NaN, 213 missing values denoted with whitespace | No outlier and invalid data detected |
| PREM_TYP_DESC | Specific description of premises; grocery store, residence, street | TEXT | 71 unique text description | 33279 Missing values denoted with NaN | No outlier and invalid data detected |
| HADEVELOPT | Name of NYCHA housing development of occurrence, if applicable | TEXT | 279 unique description | 4848026 Missing Values denoted NaN | No outlier and invalid data detected |

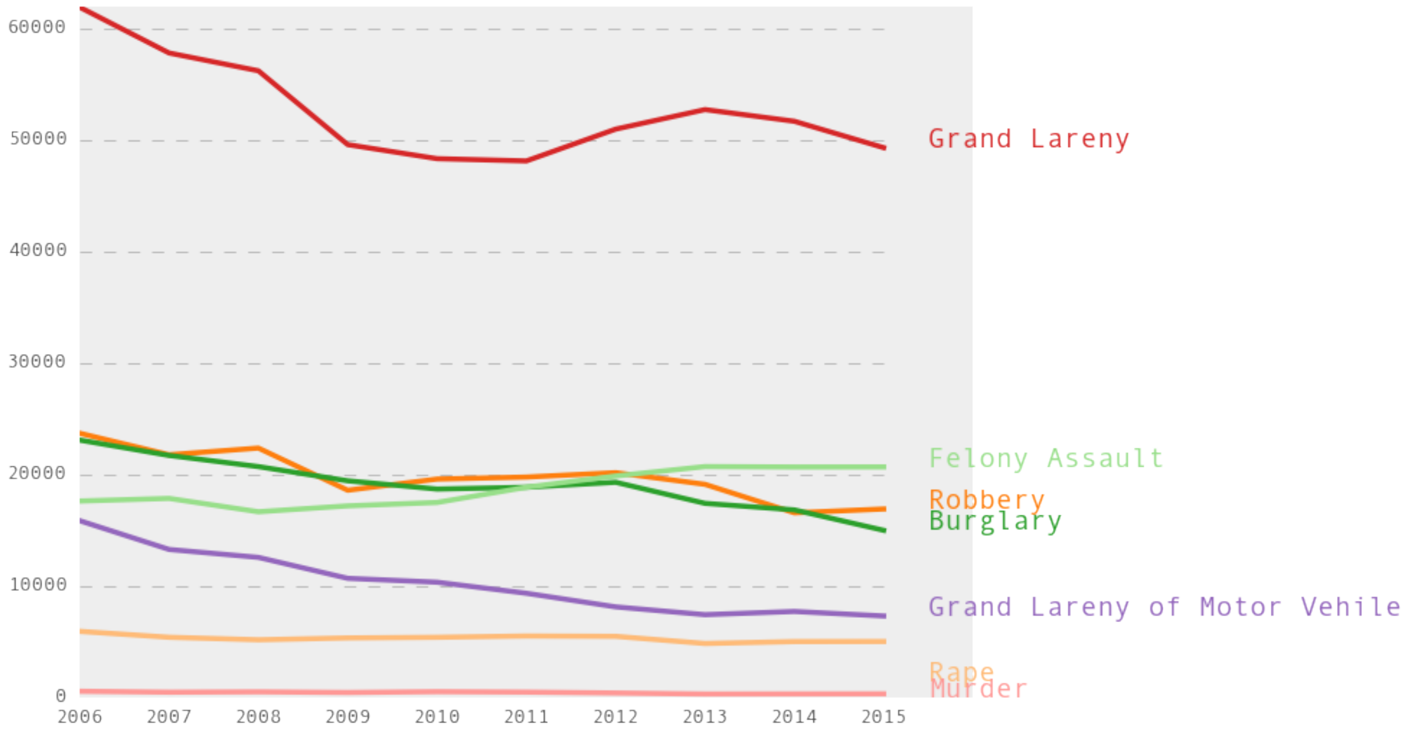
| | | | | | |
|---|---|---|---|---|---|
| PARKS_NM | Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included) | TEXT | 864 unique description | 5093632 Missing values denoted in NaN | No outlier and invalid data detected |
| X_COORD_CD | X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) | FLOAT | Coordinate | 188146 missing values denoted in NaN | No outliers, all coordinates are in NYC range |
| Y_COORD_CD | Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) | FLOAT | Coordinate | 188146 missing values denoted in NaN | No outliers, all coordinates are in NYC range |
| Latitude | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | FLOAT | Latitude coordinate | 188146 missing values denoted in NaN | No outliers, all coordinates are in NYC range |
| Longitude | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | FLOAT | Longitude coordinate | 188146 missing values denoted in NaN | No outliers, all coordinates are in NYC range |
| Lat-Lon | Latitude-Longitude Coordinate pair | FLOAT | Coordinate | 188146 missing values denoted in NaN | No outliers, all coordinates are in NYC range |

## Part II Other data quality issues:

1. Combined CMPLNT_FR_DT and CMPLNT_FR_TM, CMPLNT_TO_DT and CMPLNT_TO_TM together, for all of those valid values, we find that 1 from-to datetime pair is invalid, which contains from datetime after to datetime. We also excluded the data point in the cleaning procedure.
2. There is supposed to be a one-to-one mapping between the KY_CD and OFNS_DESC. However, some different OFNS_DESC correspond to the same KY_CD. In the labelling procedure, we label them as valid. In the data cleaning procedure, we merge them together to keep consistency.
3. As indicated by NYPD precinct map, each precinct belongs to a specific borough. We combined BOROUGH_NM and ADDR_PCT_CD together and find 17 inconsistent data values. For each invalid value, the recorded precinct does not correspond to its recorded borough. Cross check each invalid data entry with its coordinate information and make correction in cleaning procedure.

**Part III Data Cleaning:**

Once, we detect all the invalid values, we can further clean the dataset by excluding all rows containing invalid values. Furthermore, we merge values of OFNS_DESC corresponding to the same KY_CD key. The script to finish the cleaning task can be also run following the instruction on the github. For any further exploitation in our project, we will use the cleaned dataset instead of the original one so that we can trust more on our findings without worrying about data inconsistency.

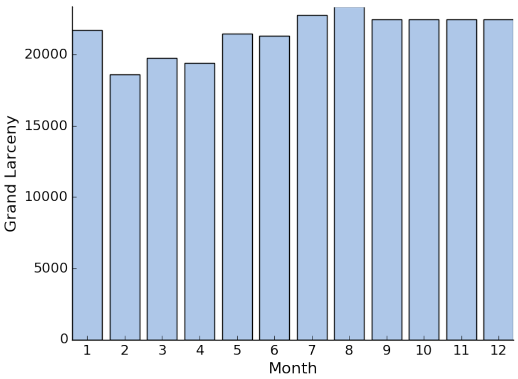


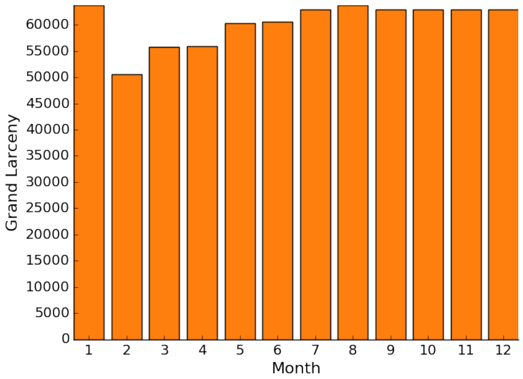
Data source: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic
Author: Police Department (NYPD)

**Part IV Data Visualization:**

The above figure shows the yearly trends of 7 main incidents in NYC, it is safe to draw that NYC is getting safer in recent years.

From the figure, we can see that Grand larceny is the most frequent offense of all 7 felonies. The number of incidents is more than twice that of the second most frequent one. The number of incidents for Robbery, Grand Laceny and Burglary decreased within 2006-2015, this may caused by the widely used technology in camera surveillance. Meanwhile, Murder and Rape stay relatively stable in past 10 years while the number of Felony Assault is slightly increasing.



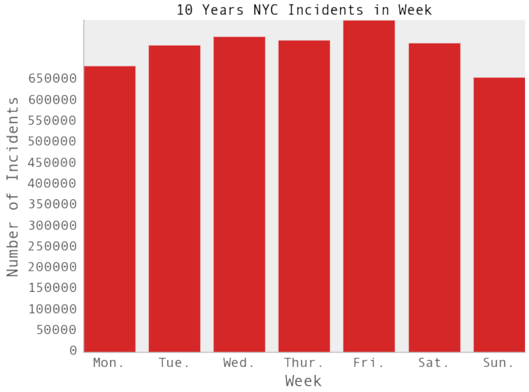


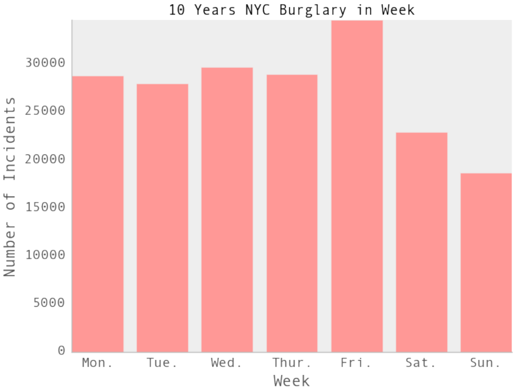
We defined NYC's seasons as follows:

Spring: March, April, May; Summer: June, July, August;

Fall: September, October, November; Winter: December, January, February

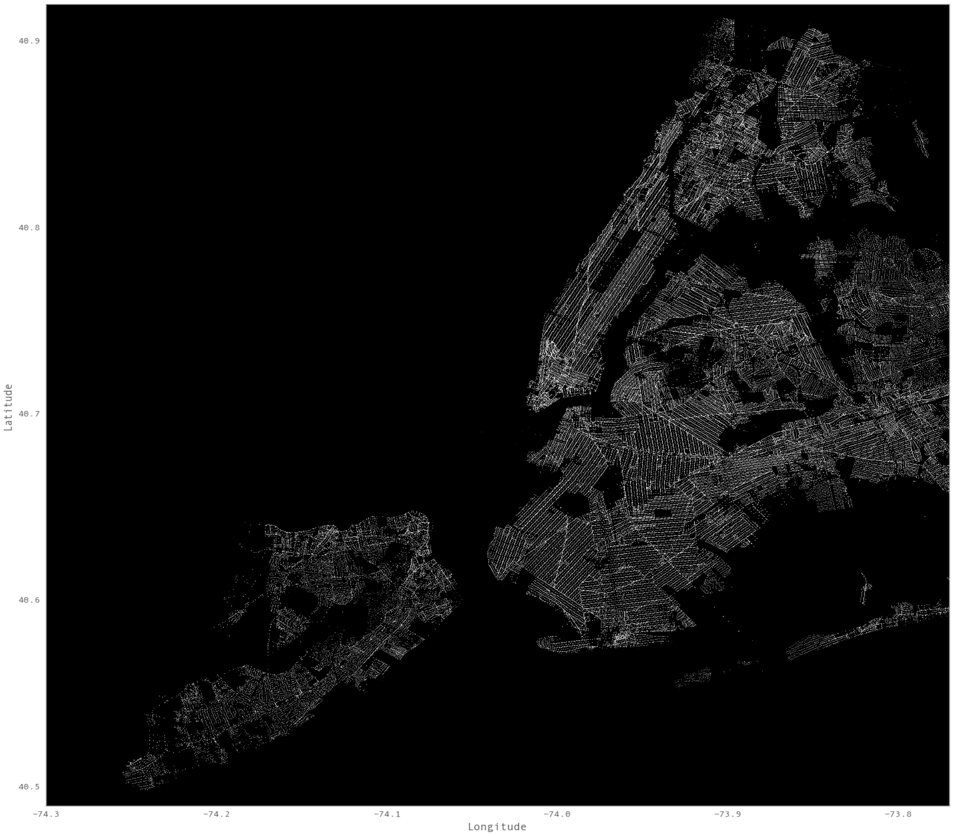
Late winter and early spring have the smallest number of incidents compared with all other seasons, hence could be considered as the safest seasons. This also make a lot of sense because who would like to go out in such snowy, windy and chilly days. In summer, more incidents happened, NYC becomes hot and humid, this may make some people feel febrile and agitated.





NYC incidents, according to 10 years of data, happened more often at Friday while less often at Sunday. But overall the frequency of incidents is uniformly distributed. For burglary happened in NYC, the frequency at Sunday is extremely lower than the other days, this may caused by the facts such as people may sleep later in the weekend than weekdays.

Beyond the time series analysis above, we transform our perspective into geometrical views.


40.9
40.8
40.7
40.6
40.5
Latitude
−74.3
−74.2
−74.1
−74.0
−73.9
−73.8
Longitude