**New York City - Crime Analysis**

**Big Data Project**

Project Members:

Harshit Srivastava (hs3500), Joby Joy (jj2196) and Pravar Singh (ps3309)

Introduction:

We have chosen NYPD crime data set ranging from 2006-2016 to apply big-data tools on & generate insights. The dataset is available for download at <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>. We found that the data set was relatively clean, however there were several NULL/Invalid values in almost all columns. The downloadable file - (NYPD\_Complaint\_Data\_Historic.csv) is 1.3 GB and contains 5,580,036 lines. Its sheer size motivated us to use big data tools such as PySpark for performing analysis.

Part I - Data Cleaning:

The following table contains the details about each of the 23 columns and the corresponding description for it. This description is provided by NYPD and NYC Open Data and can also be downloaded from the page where the data set is found, the description file is named - NYPD\_Incident\_Level\_Data\_Column\_Descriptions.csv.

|  |  |
| --- | --- |
| **Column** | **Column Description** |
| CMPLNT\_NUM | Randomly generated persistent ID for each complaint |
| CMPLNT\_FR\_DT | Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT\_TO\_DT exists) |
| CMPLNT\_FR\_DT | Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT\_TO\_TM exists) |
| CMPLNT\_TO\_DT | Ending date of occurrence for the reported event, if exact time of occurrence is unknown |
| CMPLNT\_TO\_TM | Ending time of occurrence for the reported event, if exact time of occurrence is unknown |
| RPT\_DT | Date event was reported to police |
| KY\_CD | Three-digit offense classification code |
| OFNS\_DESC | Description of offense corresponding with key code |
| PD\_CD | Three-digit internal classification code (more granular than Key Code) |
| PD\_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) |
| CRM\_ATPT\_CPTD\_CD | Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely |
| LAW\_CAT\_CD | Level of offense: felony, misdemeanor, violation |
| JURIS\_DESC | Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc. |
| BORO\_NM | The name of the borough in which the incident occurred |
| ADDR\_PCT\_CD | The precinct 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 |
| PREM\_TYP\_DESC | Specific description of premises; grocery store, residence, street, etc. |
| PARKS\_NM | Name of NYC park, playground or green space of occurrence, if applicable (state parks are not included) |
| HADEVELOPT | Name of NYCHA housing development of occurrence, if applicable |
| X\_COORD\_CD | X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| X\_COORD\_CD | Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Latitude | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |
| Latitude | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |

For part I of the project, we used PySpark to analyze the data set . We created a PySpark script that checks all the columns & identifies whether the values adhere to the constraints of their respective columns, for e.g.- values in Latitude column should all correspond to the latitude range of New York City. Similarly, the Borough names (BORO\_NM) should be one of the 5 boroughs of New York City. The script then collects the sum of Valid, Invalid & Null values in every column and gives an output like this:

[('VALID', 5111061), ('INVALID', 468319), ('NULL', 656)]

The following table depicts the number of valid, invalid & null values for each column:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Valid** | **Invalid** | **Null** |
| CMPLNT\_NUM | 5580036 | 0 | 0 |
| CMPLNT\_FR\_DT | 5111061 | 468319 | 656 |
| CMPLNT\_FR\_TM | 5579084 | 903 | 49 |
| CMPLNT\_TO\_DT | 3713616 | 393633 | 1472787 |
| CMPLNT\_TO\_TM | 4109777 | 1376 | 1468883 |
| RPT\_DT | 5101229 | 478806 | 1 |
| KY\_CD | 5580035 |  | 1 |
| OFNS\_DESC | 5561144 |  | 18892 |
| PD\_CD | 5575126 |  | 4910 |
| PD\_DESC | 5575127 |  | 4909 |
| CRM\_ATPT\_CPTD\_CD | 5580028 |  | 8 |
| LAW\_CAT\_CD | 5580035 |  | 1 |
| JURIS\_DESC | 5580036 |  |  |
| BORO\_NM | 5579572 |  | 464 |
| ADDR\_PCT\_CD | 5579645 | 1 | 390 |
| LOC\_OF\_OCCUR\_DESC | 4356430 |  | 1223606 |
| PREM\_TYP\_DESC | 5544838 |  | 35198 |
| PARKS\_NM | 12539 |  | 5567497 |
| HADEVELOPT | 277818 |  | 5302218 |
| X\_COORD\_CD | 5384167 |  | 195869 |
| Y\_COORD\_CD | 5384167 |  | 195869 |
| Latitude | 5384167 |  | 195869 |
| Longitude | 5384167 |  | 195869 |

**Steps taken to clean the data based on the above analysis:**

* Removed all the rows containing invalid/null values for the following columns :

CMPLNT\_FR\_DT, CMPLNT\_FR\_TM, RPT\_DT, KY\_CD, OFNS\_DESC, PD\_CD, PD\_DESC, CRM\_ATPT\_CPTD\_CD, LAW\_CAT\_CD, BORO\_NM, ADDR\_PCT\_CD, LOC\_OF\_OCCUR\_DESC, PREM\_TYP\_DESC, X\_COORD\_CD, Y\_COORD\_CD, Latitude, Longitude

since having a Valid non-empty value in these columns is mandatory and crucial for performing analysis in the next phase of the project.

* For CMPLNT\_TO\_DT, CMPLNT\_TO\_TM columns : Retained all the rows with null values in both the columns & copied over the values from CMPLNT\_FR\_DT, CMPLNT\_FR\_TM columns respectively since the exact date & time of occurrence of the event is known & to date, to time columns are not applicable. Deleted all other rows having Invalid values for either of the 2 columns or having null for just 1 of the columns.
* For PARKS\_NM, HADEVELOPT columns : Retained all the rows containing null values for these columns, since they are optional and have values for only a small fraction of the rows.

After performing calculations, we reached a conclusion that < 10% of the total records were pruned from the dataset which also justifies that pruning was the most efficient method rather than trying to convert these values to valid one’s since the size of the affected data set is small.