**New York City - Crime Analysis**

Project Members:

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

Introduction:

We have chosen NYPD crime data set ranging from year 2006-2016. 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 some NULL/Invalid values in some columns. The downloadable file - (NYPD\_Complaint\_Data\_Historic.csv) is 1.3 GB and contains 5,101,232 lines. Its sheer size motivated our use of big data tools for this project, including PySpark.

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 had. We created a PySpark script that would check in all the columns, whether the values lie in the range in which they are supposed to lie, for e.g.- values in Latitude column should all correspond to the latitude of New York City. Similarly, the Borough names (BORO\_NM) should be one of the 5 boroughs of New York City. The script would then collect the sum of all kinds of values inside a column and give an output like this:

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

The following table contains the type of values that each column contains:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Valid** | **Invalid** | **Null** |
| CMPLNT\_NUM | 5580036 | 0 | 0 |
| CMPLNT\_FR\_DT | 5111061 | 468319 | 656 |
| CMPLNT\_FR\_DT | 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 |
| X\_COORD\_CD | 5384167 |  | 195869 |
| Latitude | 5384167 |  | 195869 |
| Latitude | 5384167 |  | 195869 |