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.