NYPD Complaint Data HistoricPublic Safety

Team Members : Sanjana(lk2849),
Max Albrecht(MA5416),
Saketh Reddy(SP6322)

Introduction

Although New York City is arguably the most popular city in the United States, it nevertheless still experiences various types of criminal activity throughout the year. In this project we have analyzed historic crime data found in the NYPD Complaint Data dataset, identified possible data issues and implemented several cleaning strategies that are most suitable for this data set. This would enable a researcher to use this data more effectively and suggest preventive measures. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2019.
In the data-profiling part of this project, we have looked at each field in the dataset and identified what category it belongs to. All data-cleaning strategies have been recorded in detail before making changes to the dataset; these strategies have first been tested on a smaller sample of the data, and then used to clean all columns present.

2. Abstract

The goal of this project is to first clean the data in the original dataset, analyze the strategies used to clean the data and record the results. We have also implemented similar methods on several other datasets which have similar columns in them. Using this implementation we were able to compare the efficiency of the cleaned data from the first part. We have improved our methodology to make it applicable to a large number of datasets. In order to verify the efficiency of the cleaned data we have used the concepts of precision and recall. We began our data cleaning under an assumption that one particular strategy would work for the entire dataset, but after looking at a number of columns it became evident that our strategies needed to be revised to make them feasible for the majority of the data.

3. Initial Data analysis

- There are 35 columns (and 700K rows) in this NYPD dataset.
- This dataset had null, invalid, and empty fields.
- We found that the data type in some of the columns were not always uniform.
- Naming conventions in some of the columns made it difficult to tell its data type.

Table1 : Fields and Description

Field	Description
CMPLNT_NUM	Complaint Number
CMPLNT_FR_DT	Complaint From Date
CMPLNT_FR_TM	Complaint From Time
CMPLNT_TO_DT	Complaint To Date
CMPLNT_TO_TM	Complaint To Time
ADDR_PCT_CD	Code of Precinct in which the Incident Occured
RPT_DT	Report Date
KY_CD	“Key Code": Offense Classification Code (3 digits)
OFNS_DESC	Offense Description
PD_CD	PD Code of Offense. More granular than Key Code
PD_DESC	PD Description of Offense.
CRM_ATPT_CPTD_CD	Whether Crime was Attempted or Completed (values: 'COMPLETED', 'ATTEMPTED')
LAW_CAT_CD	Level of Offense (values: 'FELONY', 'VIOLATION', 'MISDEMEANOR')
BORO_NM	Name of Borough in which Incident Occurred
LOC_OF_OCCUR_DESC	Description of where the incident occurred with respect to the premises
PREM_TYP_DESC	Description of the type of premises in which

	the Incident Occurred
JURIS_DESC	Description of Jurisdiction in which Incident Occurred
JURISDICTION_CODE	Jurisdiction Code
PARKS_NM	Name of Park in which Incident Occurred

HADEVELOPT	Name of NYCHA Housing Development in which Incident Occurred, if Applicable
HOUSING_PSA	Housing PSA
X_COORD_CD	X-coordinate, New York State Plane Coordinate System
Y_COORD_CD	Y-coordinate, New York State Plane Coordinate System
SUSP_AGE_GROUP	Age Group of Suspect
SUSP_RACE	Race of Suspect
SUSP_SEX	Sex of Suspect
TRANSIT_DISTRICT	Transit-District code
Latitude	Global Latitude of Location where Incident Occurred
Longitude	Global Longitude of Location where Incident Occurred
Lat_Lon	'Latitude' and 'Longitude' together
PATROL_BORO	Patrol Borough
STATION_NAME	Station NameCRM_ATPT_CPTD_CD
VIC_AGE_GROUP	Age Group of Victim
VIC_RACE	Race of Victim
VIC_SEX	Sex of Victim

With the description for each field we now have a better idea of each column and its role in the dataset. Some columns are more important than the others, and implementing data-cleaning strategies on those columns should be prioritized.
In the following section, we will discuss several different data-cleaning strategies that will be used on the dataset. Before utilizing these strategies, we will first check the precision and recall of the raw data, and after data cleaning has been implemented we will check the data's precision and recall again to look for any improvements.

3.2 Data profiling

Data profiling allows us to identify and understand issues present in the dataset. Openclean is one of the tools we used for this analysis, and its output has given us a starting point as to where data anomalies are present. Using data profiling allowed us to find the following characteristics about the data:

- Find the minimum, maximum value in a particular column
- Find the total number of rows that are present in the data table
- Find the total number of empty values present in a specified column
- Find the potential data types present in the column specified
- Find the distinct values present in the column profiled
- Find the most frequent values in a column
- Find the entire profiling of the dataset which gives details about the data values.

These are a few salient points on the profiling of this data. We will discuss this further and show an implementation example based on this dataset in the companion Jupyter notebook

3.3 Problems with dataset

With the successful implementation of the dataset with the profiling techniques we were in a position to figure out some of the data issues which are addressed with the cleaning strategies. Let's discuss the initial problems with datasets that are found.

- For **CMPLNT_Num** we have to check weather the column is having integer values in
- The **KY_CD** field from the data profiling shows that it does not have any invalid or empty values (so there should be no problem with this column)

- The **OFNS_DESC** field contains empty values but this does not look to be a mandatory field and few descriptions tend to have the same meaning but with different names we can try to address this issue by grouping
- The **PD_CD** has a few empty values that are to be considered for cleaning.
- The **CRM_ATPT_CPTD_CD** has the Values as COMPLETED , ATTEMPTED and empty values , so changing the empty values field to UNKNOWN would address the major problem of deleting a huge amount of useful information .
- The **LAW_CAT_CD** Column has values , FELONY, VIOLATION, MISDEMEANOR values so Check if the offence column just consists of following mentioned offences and from profiling of data it is evident that there is no need of any transformations as the data seems to be perfectly alright
- The **JURIS_DESC** column has the data type of string and there are no invalid or empty values , so this is perfectly alright.
- The **LOC_OF_OCCUR_DESC** column consisted of the following discrete values in it FRONT OF, 'REAR OF' 'OUTSIDE' 'INSIDE' 'OPPOSITE OF. and the remaining empty values should be renamed to UNKNOWN as they we were so huge in count and removing the null values would lead to losing of valuable data.
- The **PREM_TYP_DESC** has few Empty values present so validate them and take appropriate action on them.
- **JURISDICTION_CODE** : This column seems to be least important in this data and has the highest number of discrepancy , by looking at the other part of the project we can make a point of leaving this column as it is.
- **PARKS_NM**: The highest NULL values are present in this column and there may be null values present in this column which do not have much impact in near future.
- **HADEVELOPT** : There may be null values present in this column which does not have much impact in the near future.
- **HOUSING_PSA** : There may be null values present in this column which does not have much impact in the near future.
- **'X_COORD_CD'**, **'Y_COORD_CD'**, **Latitude** , **Longitude** , **LatLon** : All these columns are representing the similar kind of the data so dropping 'X_COORD_CD', 'Y_COORD_CD' , LatLon would be a better option and which eradicates the redundancy of data.
- **'SUSP_AGE_GROUP'** : There are few values which are wrong of its type and removing those wrong values which are in less amount would be best practice
- **'SUSP_RACE'**, : There are null values so set the null values to UNKNOWN
- **'SUSP_SEX'** : There are null values so set the null values to UNKNOWN
- **VIC_AGE_GROUP**', '**VIC_RACE**','**VIC_SEX**' : These columns are validated and empty values are present which can be named as UNKNOWN as this is the best possible option to analyze the data further and helps in not losing other important data just because of these least important fields.

- **TRANSIT_DISTRICT** : This column has high number of empty values so cannot be used to get any insights so this channel can be ignored further
- **PATROL_BORO** :Empty values are present and these can be set to UnKnown as this change will not impact much on data analysis as the count is negligible when compared to total data.
- **BORO_NM :** Empty values are present in this column so validate the remaining values and change the empty fields to Unknown.

With the initial data profiling we have found to see some problems in the dataset and These anomalies will be cleared in the cleaning part of the work.

3.4 Finding the initial precision and Recall for the dataset

Precision and Recall are two great concepts which helps you in evaluating your work in scientific method, each individual has a different opinion of data cleaning and by using this methodology we can generate the accuracy of your strategy and make a point of how good the strategy worked out in the realtime dataset.

$$\text{Precision} = \frac{tp}{tp + fp}$$

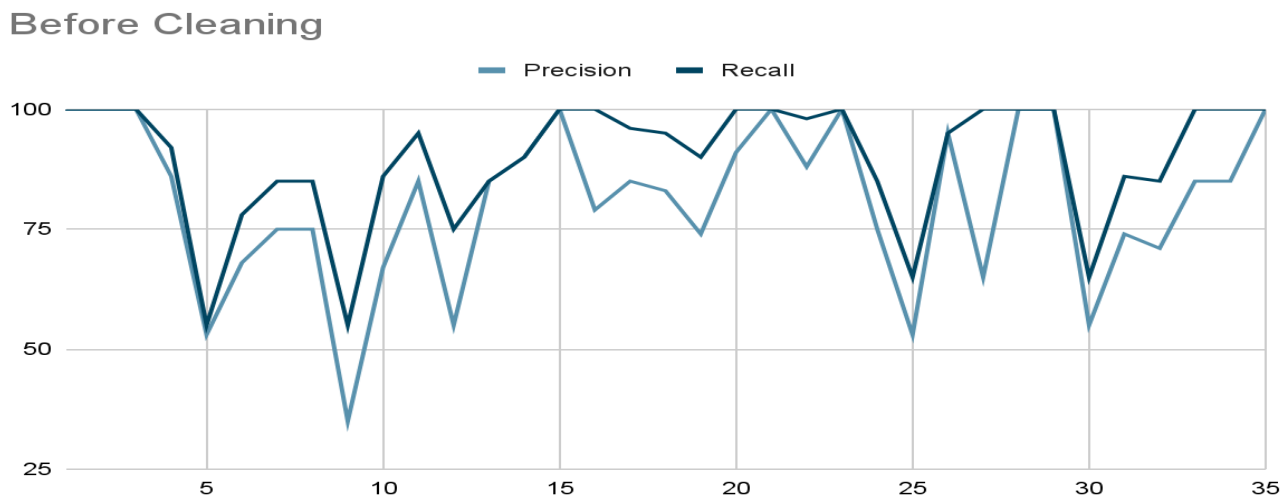
$$\text{Recall} = \frac{tp}{tp + fn}$$

Table2

Scenario	Precision	Recall
Check for Borough (Manhattan)	**83%**	**100%**
Check for "Rape" in Desc	**43%**	**57%**
Check for "REAR OF" in LOC_OF_OCCUR_DESC	**73%**	**100%**

In the above mentioned scenarios the data of empty values or values with similar kinds of names are present which makes the precision calculation to go back , cleaning the

data of that kind would possibly increase the precision and recall values.



3.5 Data Cleaning (Methods, Design, Architecture)

In data cleaning, you fix or remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset. The likelihood of data being duplicated or mislabeled is high when combining multiple data sources. Even if the data looks correct, results and algorithms will be unreliable if the data is incorrect. There is no universal process to follow when cleaning data, as cleaning will vary from dataset to dataset.

How are we planning to clean the data —--
There are few generic methods that are usually followed inorder to make the data clean to an extent and we have also kept in mind that cleaning a dataset will mostly be different from other datasets.

Guidelines that are followed in cleaning the datasets:

- **Outliers are to be filtered :** usually data has some outliers which lead us to make false predictions , eradicating such a value in the cleaning process would be much helpful in getting proper outputs and visualizations
- **Identify and fix structural errors :** Sometimes we find data with values Na , N/A.. all these represent the same identity so these kinds of structural issues are to be addressed by making appropriate changes to those.
- **Missing data must be handled** : This is one of the most important aspect to be kept in mind, missing data makes us feel to remove it but by doing so we will be loosing the data integrity and huge amount of valuable information may be lost , so strategy to this mechanism is to be implemented based on the reliability of the

column with the other columns present , and see if the values in it are really important for analysis.

- **Get rid of redundant or irrelevant observations:** Fixing this would make us do further analysis without any abnormal outputs.
- **Validate the cleaning techniques :** All the cleaning techniques should be validated by asking a few questions such as Can the data be interpreted? Is there any pattern in the data that will assist you in forming your next theory? Does it support or refute your working hypothesis, or provide any new information?

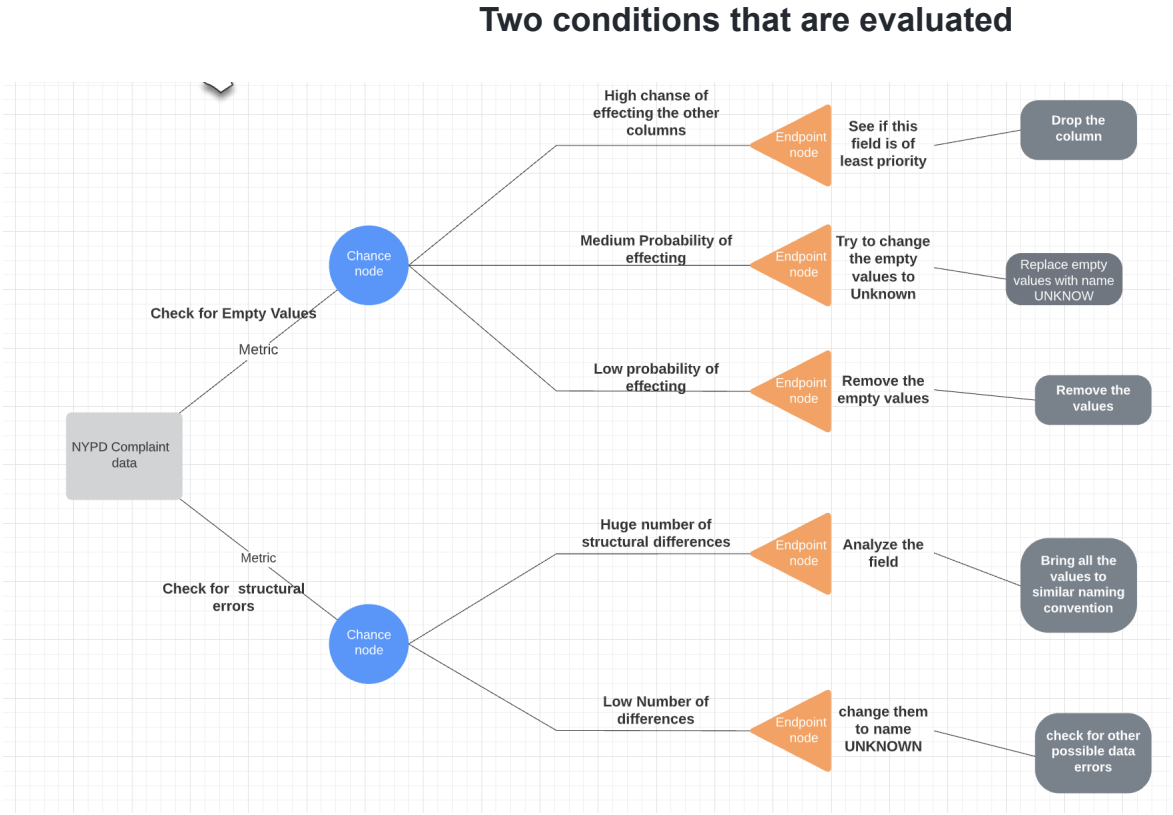
We have always kept these guidelines live when working on this data cleaning task and eventually the cleaned data obtained has satisfied all or majority of the questions we posed , All the techniques involved in the cleaning process are well documented in the jupyter notebook file present in the repository.

First four techniques have been implemented on the data set but in order to validate the approach we have chosen to consider 10 other data sets which have the similar columns overlapping from the original data source and implement the cleaning strategies on them and see if the techniques that we have defined are scalable for those datasets. And try to learn about any additional methods if needed to make the cleaning strategy scalable to a large number of datasets.

3.5.1 Code Repositories

Filtered data Repository	**https://github.com/sakethreddy997/BigData-Nyu-NYPD-Complaint-Data-HistoricPublic-Safety/tree/main/filtered_data**
Cleaned code repository (similar datasets)	**https://github.com/sakethreddy997/BigData-Nyu-NYPD-Complaint-Data-HistoricPublic-Safety/tree/main/jupyterNotebook_code**
NYPD Complaint dataset - Code	**https://github.com/sakethreddy997/BigData-Nyu-NYPD-Complaint-Data-HistoricPublic-Safety/blob/main/Bigdata_part1.ipynb**
Report	**https://github.com/sakethreddy997/BigData-Nyu-NYPD-Complaint-Data-HistoricPublic-Safety/tree/main/Report**

Figure 1:



3.6 . Datasets with similar fields

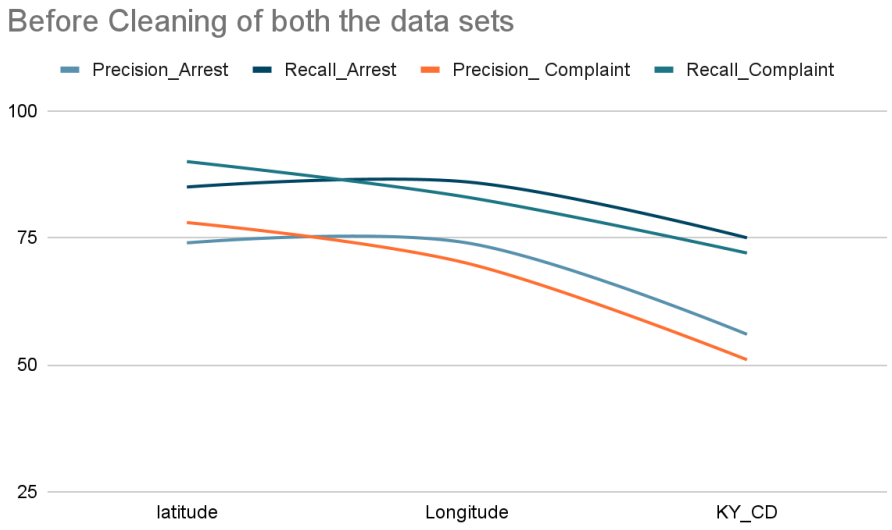
Here are the different datasets that we have used to implement the same strategies as that of the original NYPD dataset

Tabel 3 : Similar datasets

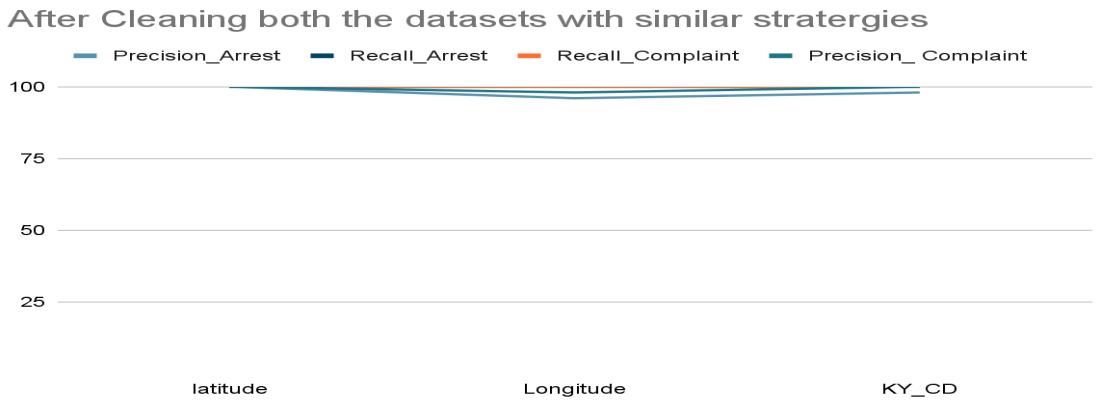
Datasets
NYPD Arrest Data (Year to Date) | NYC Open Data

NYPD Shooting Incident Data (Historic) | NYC Open Data
NYPD Complaint Map (Historic) | NYC Open Data
NYPD Criminal Court Summons (Historic) | NYC Open Data
NYPD B Summons (Year to Date) | NYC Open Data
NYPD-YTD-Criminal-Summons
NYPD Hate Crimes | NYC Open Data
NYPD YTD Arrests - Summary Dashboard | NYC Open Data
NYPD Shooting Incident Data (Year To Date) | NYC Open Data
Motor Vehicle Collisions - Crashes | NYC Open Data

- **NYPD Arrest Data** has shared the columns latitude , longitude ky_cd which are present in the original dataset of NYPD Compliant data.

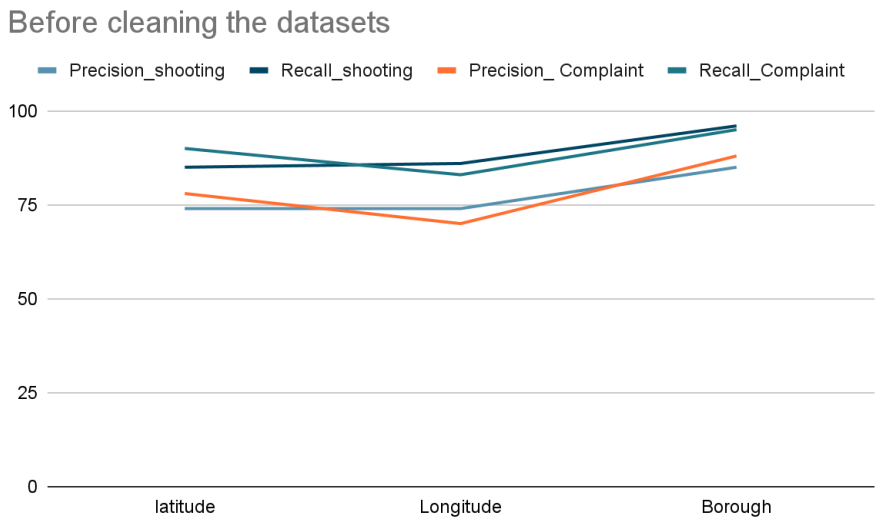


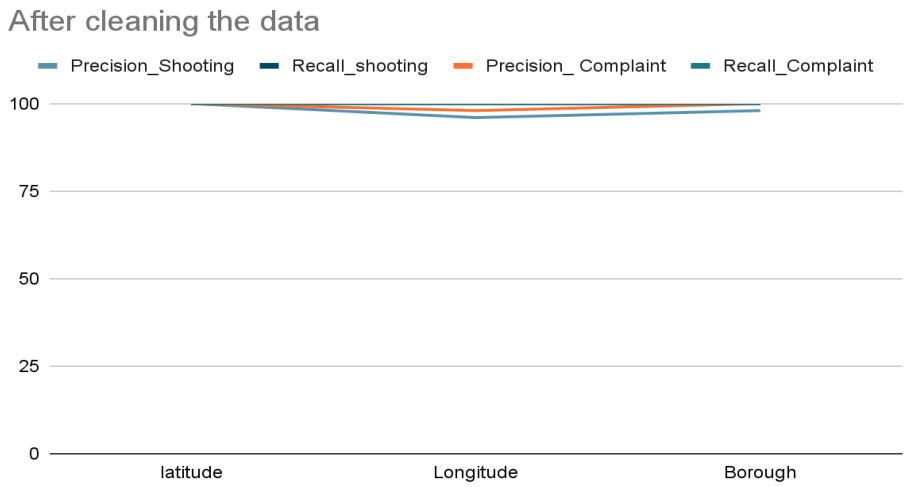
- After cleaning the datasets



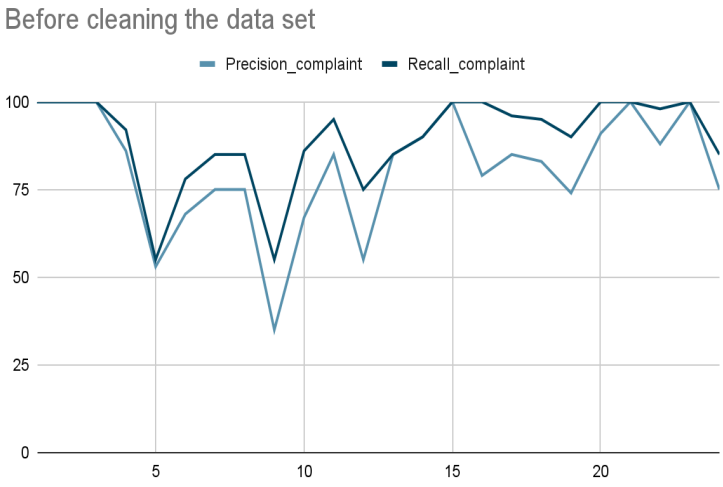
➔ **NYPD Shooting incident** dataset has overlapping columns such as latitude, longitude and borough with the original NYPD complaint dataset

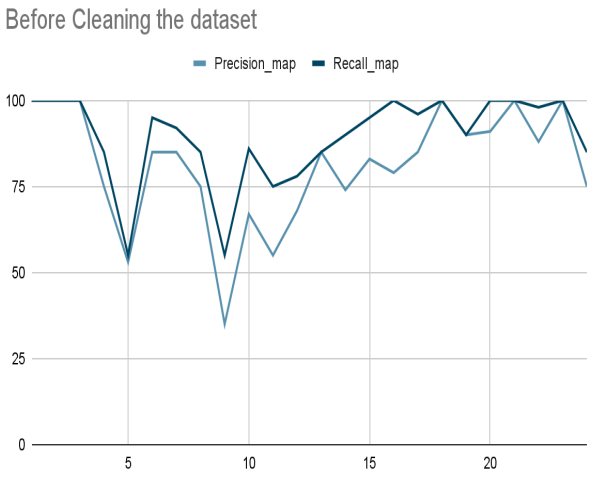
- Precision and recall values for these fields before cleaning are plotted below

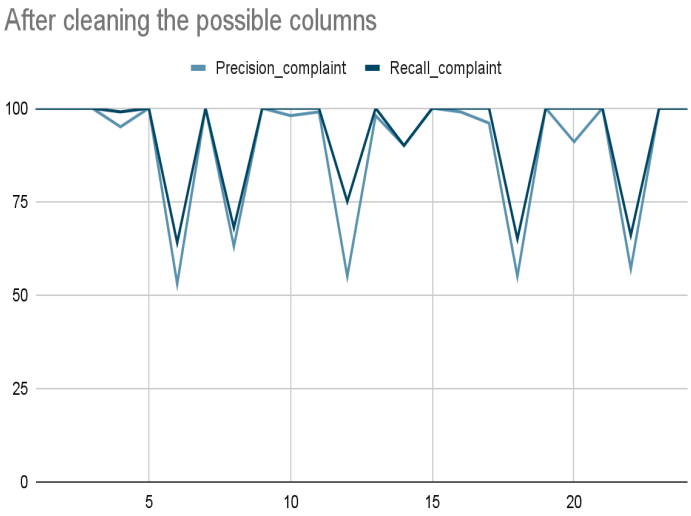


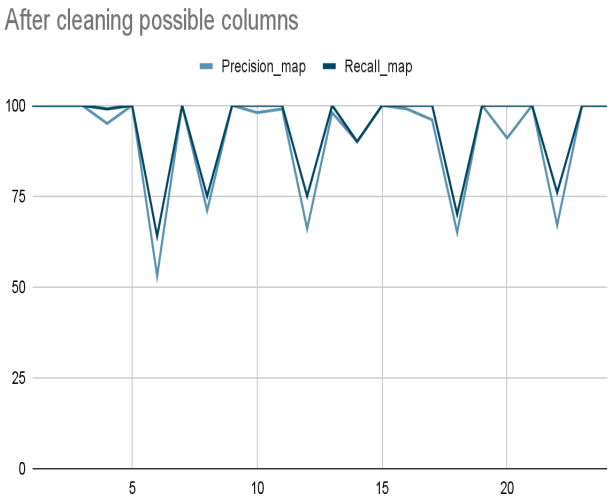


NYPD Complaint Map is sharing 24 columns when compared with the original dataset



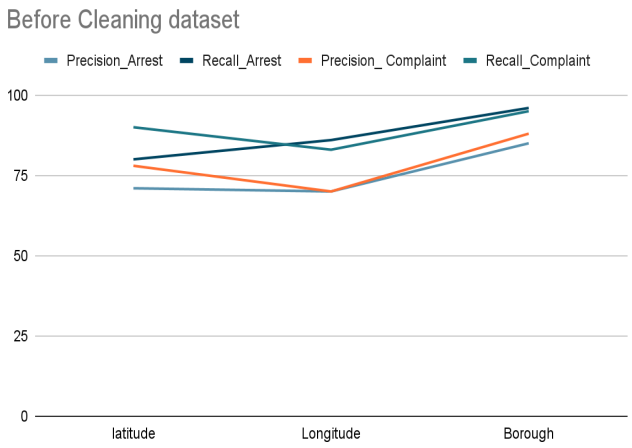


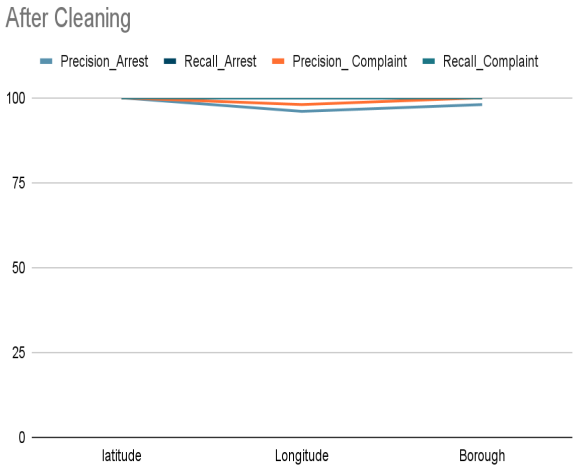




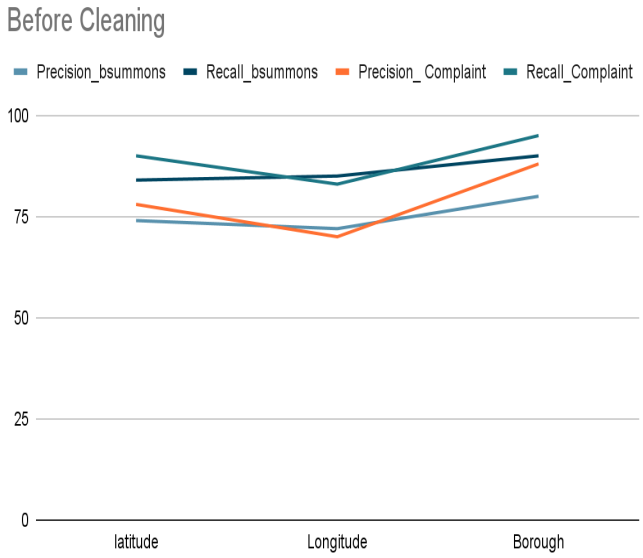
- After cleaning the possible columns in both the datasets we could observe that the 19 columns tend to show the best possible precision and recall values , these sharp decline in the graph are because those columns have least significance in the dataset and contains large number of empty values which have nothing to do with the remaining columns , so we have ignored those columns and because of this approach we have sharp decline in graph , if these columns are eliminated from the dataset then the graph would become a straight line graph and stays in the top 10 percentage range

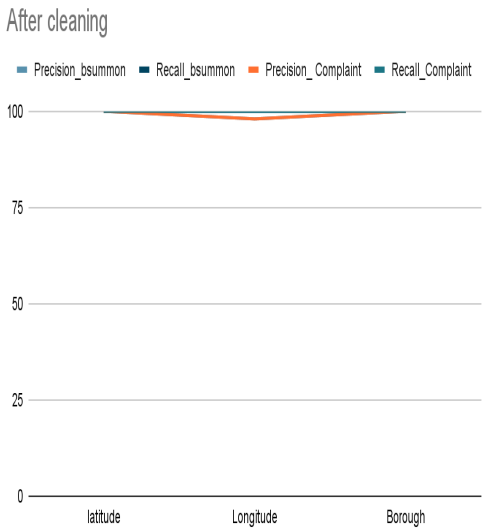
➔ **NYPD criminal court Summons dataset** has shared 3 columns from the original dataset from which we are getting the cleaning strategies. Cleaning is done on all possible columns of these datasets but we are trying to compare the precision and recall values for the columns that are overlapping with the original dataset so that we can possibly understand how well our strategy has worked on both the datasets.



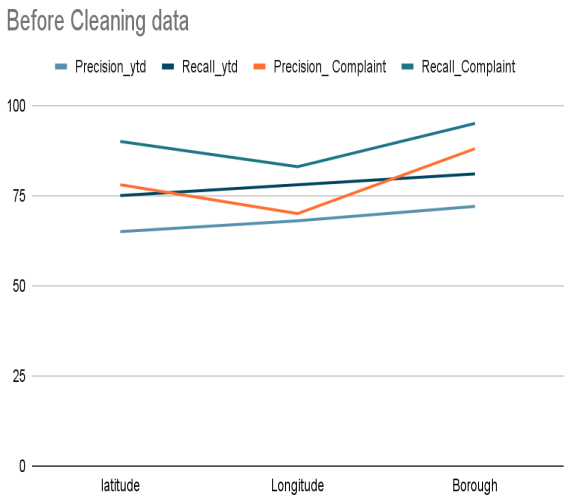


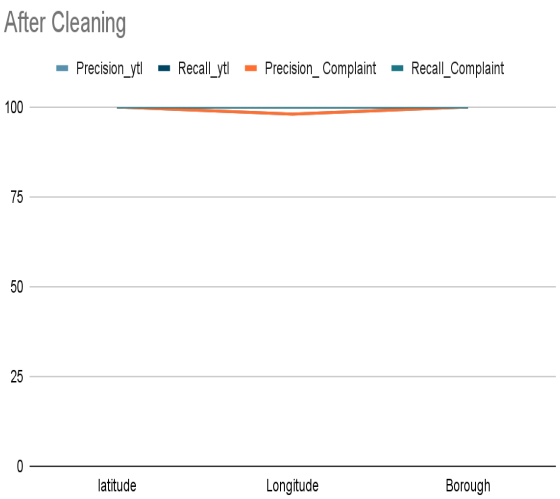
➔ **NYPD NYPD B Summons dataset** has shared 3 columns from the original dataset from which we are getting the cleaning strategies. Cleaning is done on all possible columns of these datasets but we are trying to compare the precision and recall values for the columns that are overlapping with the original dataset so that we can possibly understand how well our strategy has worked on both the datasets.



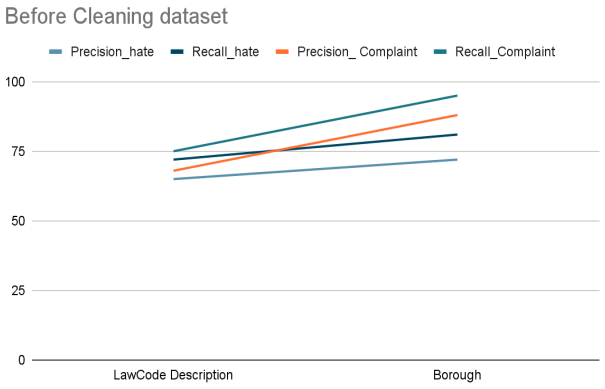


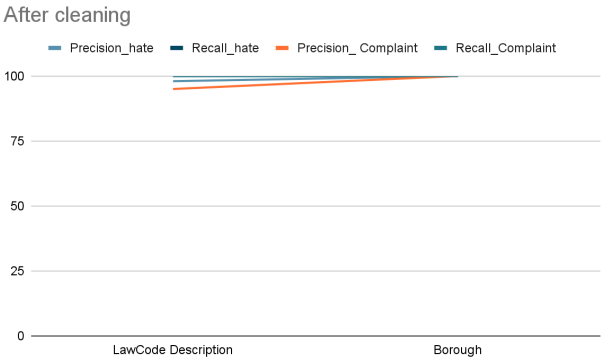
➔ **NYPD YTD Criminal Summons** has shared 3 columns with the original dataset nypd complaints (latitude , longitude , borough)



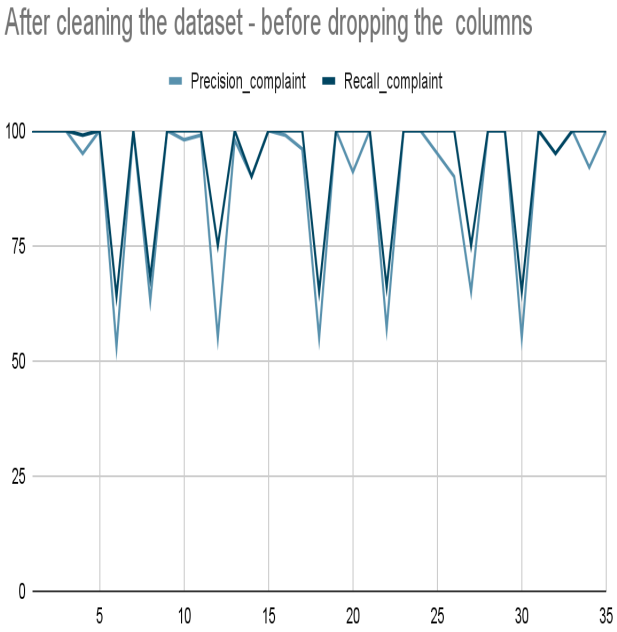


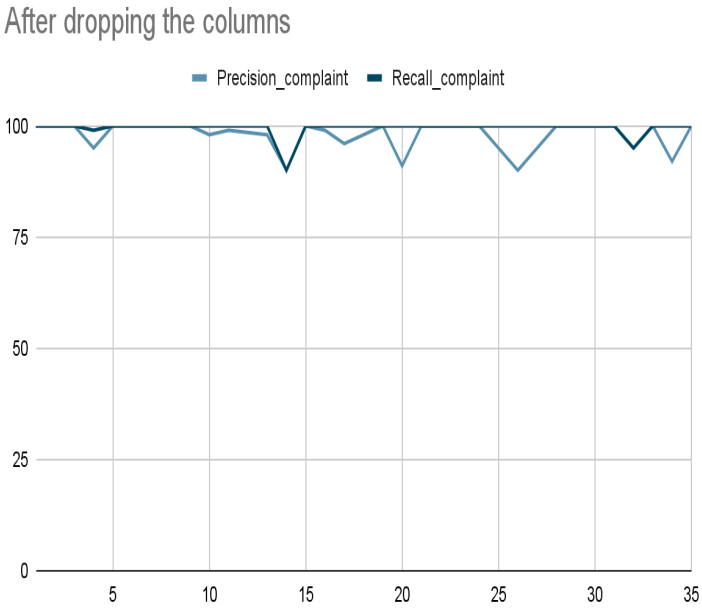
→ **NYPD Hate Crimes** has shared 2 columns which are similar to that of the original dataset.





It has been observed that all the datasets are performing best with the approaches that are present with the strategies but cleaning each dataset differs from every other. When we were working on the similar datasets we have found a few new strategies which will help to clean the original dataset.





4. Improvements

Following are the improvements that we have identified while working with the other similar datasets.

- Check the date field and make sure it is in the range.
- Check for the date format , and make sure entire column has the same kind of format
- Similar case is for the time fields present , make sure the format is same in all the values.
- PD Description field can be reduced together into a few categories and rename the field values in such a way that those come under their category this will help in analyzing the categories easily.
- Borough field has proper values in the original dataset but in the other datasets it is with the shortcut names , so bringing all of them to a similar format would be the best working model.
- In general it is clear and evident that the guidelines that are defined in section 3.5 made the cleaning process much easy and accurate.

5. References

Precision and recall

Never Forget Again! // Precision vs Recall with a Clear Example of Precision and R…

Clean and Shape Data in Tableau Prep - Tableau

Data cleansing.

Classification: Precision and Recall | Machine Learning Crash Course

When Accuracy Isn't Enough, Use Precision and Recall to Evaluate Your Classification Model

Precision vs Recall | Precision and Recall Machine Learning

The Case of Precision v. Recall. Allow me to attempt to un-confuse you | by Robert Alterman

VIDA-NYU/openclean: openclean - Data Cleaning and data profiling library for Python

What is Data Profiling? Data Profiling Tools and Examples

Data Profile - User Guide

Project Jupyter | About Us

Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data

8 Techniques for Efficient Data Cleaning

openclean/Parking Violations - Profiling and Cleaning Example.ipynb at master · VIDA-NYU/openclean