NYU | TANDON SCHOOL OF ENGINEERING

Project:

Cleaning and Analyzing
NYPD Complaint Data (Historic)

Team Members	NetId
Jain, Swathi	sp6180
Patil, Pruthviraj	prp7650
Satish, Nithya	nss9889

GitHub Link :

https://github.com/Aahbree/Crime-Data-Analysis

NYPD Complaint Data Historic

Nithya Satish
nss9899@nyu.edu

Pruthviraj Patil
prp7650@nyu.edu

Swathi Jain
sp6180@nyu.edu

ABSTRACT

The goal of this project is to clean the data in a dataset, analyze it and provide a descriptive summary. Also, we have done some data exploration with other data sets and generated our reference data. The data cleaned using PySpark(from the first part) is used to gain meaningful insights and develop a summary for the features and contents. Our techniques were further refined to make them applicable to sample data of other datasets. We have ensured the effectiveness of our technique by calculating precision and recall. The original dataset used (NYPD Complaint Data Historic) includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of the last year 2020.

INTRODUCTION

Our primary dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of the last year 2020. The dataset is available for download at NYC Open Data:https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
The downloadable file is 2.4 GB and contains 7,375,993 lines. Its sheer size motivated our use of big data tools for this project, including PySpark. In the data cleaning part, we have looked for data quality issues as well as anomalies. The methods used were improvised to fit other datasets with matching columns found using Auctus Dataset Search (https://auctus.vida-nyu.org/).

PROBLEM FORMULATION

The objective of this project is to do an analysis on crimes in New York city to allow police, citizens, and tourists to better maneuver around the city. Our insights are only as good as the data we are using to get them. So we need to make sure that the data we use is the quality data that we can make meaningful insights from.

The raw dataset includes values that can be incorrect, inaccurate, incomplete, incorrectly formatted, duplicated or even values that are irrelevant to the dataset. Data Profiling was undertaken to understand the inconsistencies in our data. We looked for data quality issues such as the ones mentioned below:

- Incorrect format of date and time
- Missing, Null and Unknown Values
- Mismatch between datatype and column values.
- Negative and other outlier values for the age group of both suspect and victim
- Unrefined race and gender values
- Invalid complaints where the crime is neither attempted nor completed
- Invalid Jurisdiction code and Precinct
- Invalid coordinates that do not belong to NYC

Any meaningful understanding from the features was recorded and plotted. The cleaning techniques should work on other similar datasets, so our techniques need to constantly evolve and improvise to generate reference data. In the data exploration, we wanted to find out whether the crime rates had any correlation with the climate, financial crisis, unemployment rates, and real-estate prices.

RELATED WORKS

There are many works available that relate to our dataset where crimes are plotted and analyzed based on boroughs, incident date, incident time, age group, race, and sex. Few others are used to forecast crimes using machine learning models. Few of them are listed below:

1. https://towardsdatascience.com/analysis-of-nyc-reported-crime-data-using-pandas-821753cd7e22
2. https://a1080211jeff.medium.com/exploring-nyc-analysis-of-crime-data-in-new-york-city-6134642b9833
3. https://towardsdatascience.com/what-triggers-crime-in-nyc-parks-3953c5df2be2

DESIGN AND ARCHITECTURE

1. Raw Dataset:

Raw data sometimes called source data has not been processed for use. Raw data processing can be a time-consuming task and it is not always easy to catch anomalies. Therefore simple checks should be run that are quite effective in eliminating the abnormalities. Statistical raw data processing needs to be carried out, in this case, to eliminate this data point in order to ensure the accuracy of the data. Data profiling helps us understand the raw datasets. There are 35 columns and 7 million lines of data in the NYPD Complaint dataset.

2. Data Profiling:

Using Openclean, we are profiling the data and as a preliminary check we found the columns with no null values: CMPLNT_NUM, RPT_DT, KY_CD, LAW_CAT_CD. Openclean is easy and intuitive, allowing users to compose and execute cleaning pipelines that are built using a variety of different tools. Using the DBSCANOutliers, we find the outliers in the complaint dates. We also find that the distinct values in many of the columns are higher than the valid data appropriate for that column. The unknown values are added in different formats in different columns. We are striving for data quality which is of utmost importance.

3. Pre-Screening:

After profiling, we found a few columns are redundant such as LatLon, X-coord, Y-coord. Also, columns that have very high Null values are not relevant. The impact of Null values can also be seen from the graph created from profiling using the Openclean package. We can see that this data (Fig 1) contributed quantitative reasons for us to drop the columns.

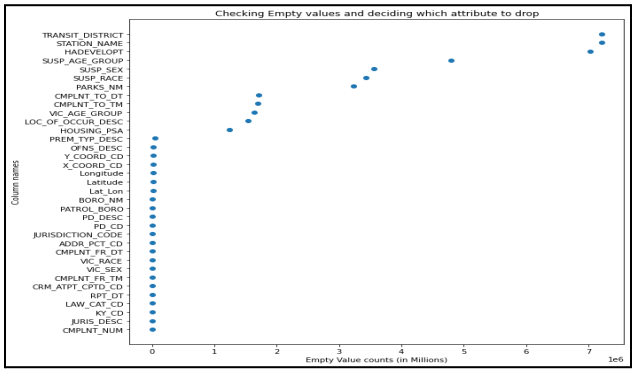

Fig 1: Empty values vs Attributes

So, we created a subset of columns that we are interested in cleaning and further analysis. The key columns we have considered for further processing are

1. CMPLNT_NUM
   Randomly generated persistent ID for each complaint
2. CMPLNT_FR_DT
   Exact date of occurrence for the reported event
3. CMPLNT_FR_TM
   Exact time of occurrence for the reported event
4. ADDR_PCT_CD
   The precinct in which the incident occurred
5. KY_CD
   Three-digit offense classification code
6. LAW_CAT_CD
   Level of offense: felony, misdemeanor, violation
7. BORO_NM
   The name of the borough in which the incident occurred
8. PREM_TYP_DESC
   Specific description of premises; grocery store, residence, street, etc.
9. VIC_AGE_GROUP
   Victim's Age Group
10. VIC_RACE
    Victim's Race Description
11. VIC_SEX
    Victim's Sex Description
12. Latitude
13. Longitude
14. SUSP_AGE_GROUP
    Suspect's Age Group
15. SUSP_RACE
    Suspect's Race
16. SUSP_SEX
    Suspect's Sex
17. JURISDICTION_CODE
18. PATROL_BORO
19. PD_CD
    Internal Classification Code
20. HOUSING_PSA
    Development Level Code

4. Data Processing:

Now we have a piece of detailed knowledge about the missing data, incorrect values, and mislabeled categories of the dataset. We will now see some of the techniques used for cleaning data. Finding data discrepancies is essential for further analysis because outliers can wildly cause misinterpretation of the analysis we make. We are explaining two parts of realization here, one is finding missing values and another is finding incorrect values. Processing and cleaning of data are done using PySpark.

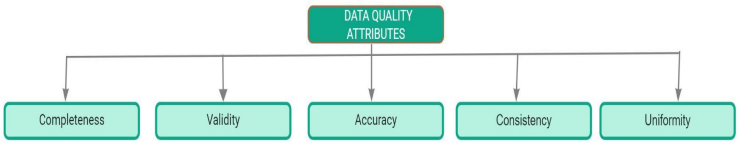

Fig 2: Data quality attributes

We have implemented modules to fix the below problems:

- **Invalid age group:** Few values in the raw dataset are negative or an incorrect integer(greater than 125) for the columns age group of suspect and victim. We replaced these incorrect values with "Unknown". This is followed by data imputation where even the Null values are filled with "Unknown"

- **Invalid date format:** After looking at the sample dataset we extracted from the original dataset, we noticed a bunch of

rows with dates in improper formatting. We used the combination of date and string manipulations to extract and format the values which are different from the original “mm-dd-yyyy” format. Also, in some datasets, the date format was different from strings. So, we added a condition in our method that could handle the timestamps with that format as well.

- **Invalid time format:** In the original dataset, we noticed a bunch of rows with time in improper formatting. Using the combination of time and string manipulation to extract values that did not conform to the 24-hour format and formatted it.

- **Invalid Race:** Null Values in the data can distort the analysis and validity of the results. When the values in column victim_race or suspect_race have empty/null values, we are replacing these rows with “Unknown”. Also, incorrect values are replaced with “Unknown”.

- **Precinct, Jurisdiction Code:** Here, we check if the values match the datatype of the column. If there are missing or null values, we remove the data since the Null values are of a very small range in these columns.

- **Bound the coordinates to NYC:** We found the bounding latitude and longitude values of New York City and removed values that do not fall within the city coordinates.

- **GeoSpatial Imputation:** There are a lot of geocoding libraries available in python. We chose geopy service as it can interact with many geocoding services like google, bing, etc. Using the API, we find the zip code with latitude and longitude values. Using the zip code mater dataset we imputed the borough values in the dataset.

- **Validate Borough Names:** Here, we check the validity of boroughs where the incident occurred as well as the patrol borough. We have replaced the abbreviated names with their full form. Also, the missing values are already imputed using reverse geocoding.

- **Validate Sex Column:** For this column, we have categorized it as M, F, E, D, and others. All the missing values are replaced with “Unknown”. This validation is done for the suspect and victim’s gender values.

- **Validate Level of Offense:** We check the validity of offense level and restrict it to Felony, Misdemeanor, and Violation. The offenses which are not classified into these three groups are deleted.

- **Validate Type of Offense:** There are two columns that indicate the category of offense committed. Both are of numeric data type. We check its validity and remove incorrect values.

Note: These validation techniques are later refined when applied over a similar dataset. For example, initially, we considered only the Borough Names in their full form. Later, we refined the method to accept abbreviated names or short forms and replace it with full names.

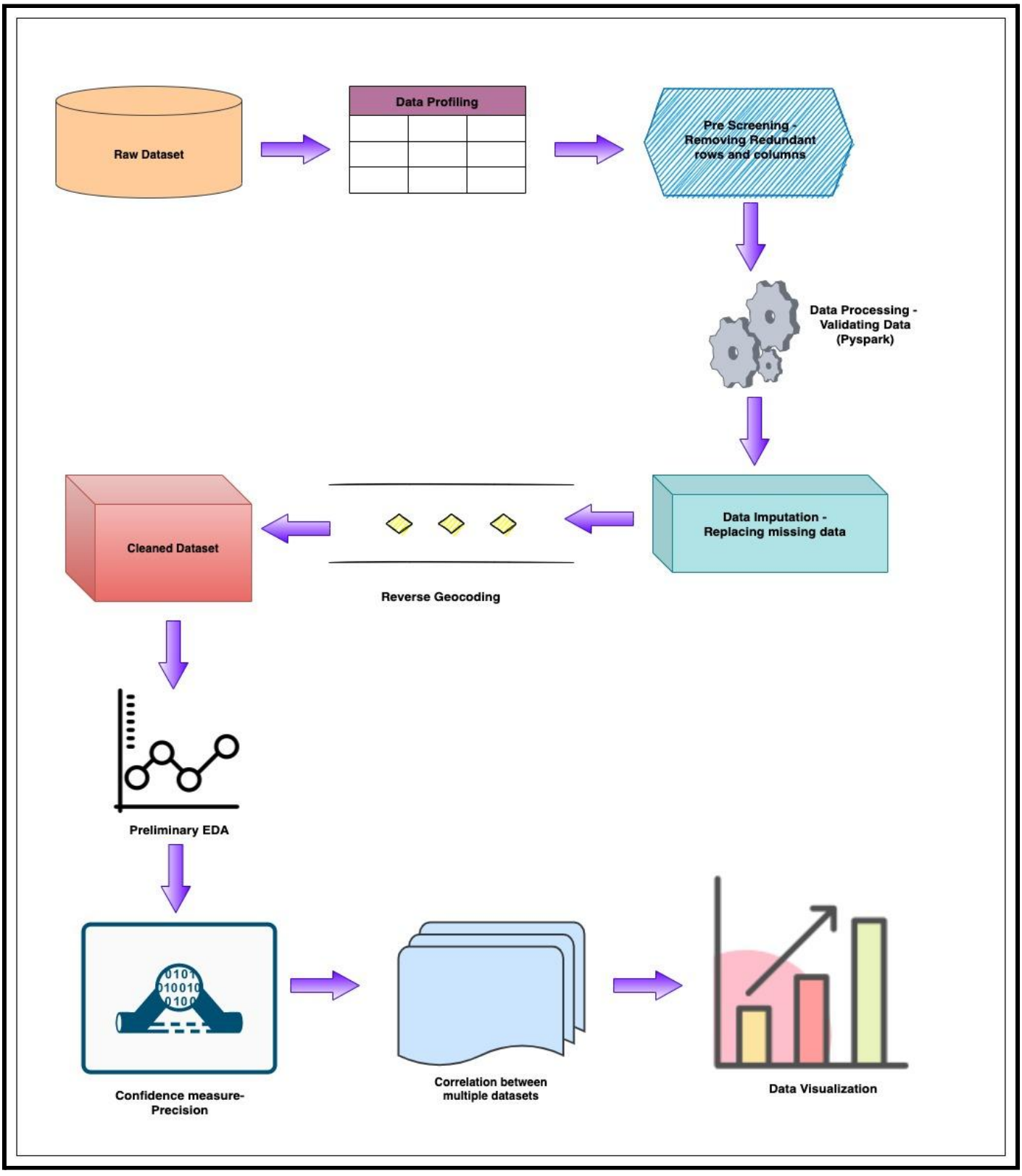

Figure 3: Architecture Diagram

5. Reverse Geo-Coding:

Reverse geocoding is the process of converting a location as described by geographic coordinates to a human-readable address or place name (Borough Name in our dataset). Our dataset has few Null values in the columns Incident Borough and Patrol Borough, we are using the reverse geo-coding method to find the borough name if the borough field has missing data.

With the help of the Latitude & Latitude coordinates, we can find the borough name using API Call. The Geopy package gives out the zip code using the latitude and longitude attributes. In turn, we use the Master dataset of zip codes vs boroughs and store them in a dictionary (to optimize the time complexity), and then use it to get Borough names.

6. Data Exploration:

Exploratory analysis of data is one of the best forms to gather the architecture and dependencies within the Data. This data may or may not be required for solving the problem at hand but will be very useful to grasp the structure of that data set. This can consist of various steps and charts that you can use to analyze data and explore connections and meanings between different data values present. The aim here should be to thoroughly understand the working of tabular columns and the values they hold.

From data profiling, we not only found the number of empty values which we took care of, but also came to know about some interesting statistics about the crime dataset. Such as the following:

a. As per Fig 4, We can see that the max crime rate is on the first day of every year surprisingly. This can be attributed to:

- ➔ different types of rent people have to pay on the first-month
- ➔ subscription charges people have to pay for services used.
- ➔ population density on roads to celebrate the new year (example: the very famous Ball Drop at Times sq NYC).

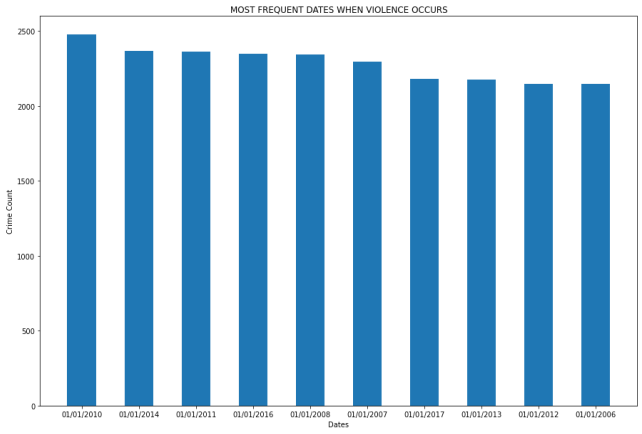

Figure 4: Max crime occurring dates

b. As per Fig 5. We can see that even though most of the crime happens mostly after dusk, the afternoon at 12:00 pm has the maximum crime rate

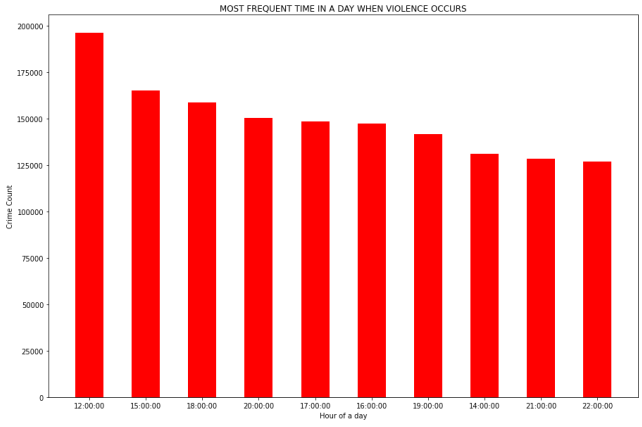

Figure 5: Hour of the day vs Crime Count

c. As per Fig 6. We can also say that Brooklyn has seen the most crime rate till now from 2006. This may be attributed to lesser development in the area, a large population density in the upper west side of Brooklyn, etc.

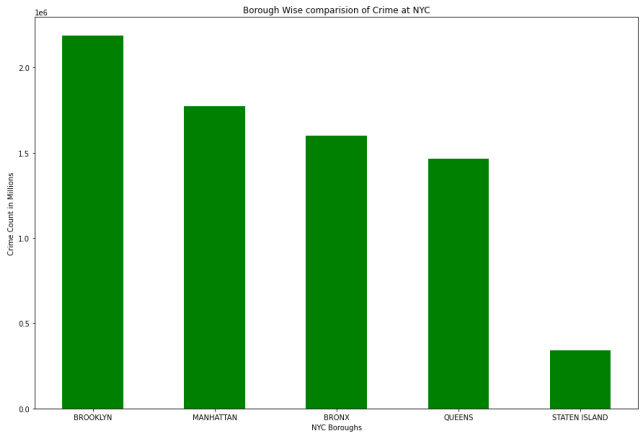

Figure 6: Boroughs vs Crime Count

d. As per fig.7. We can see the max crime occurring points in NYC. We can be sure that even though the max crime occurs in upper Brooklyn, Manhattan has max points in the top 10 violence-occurring places.

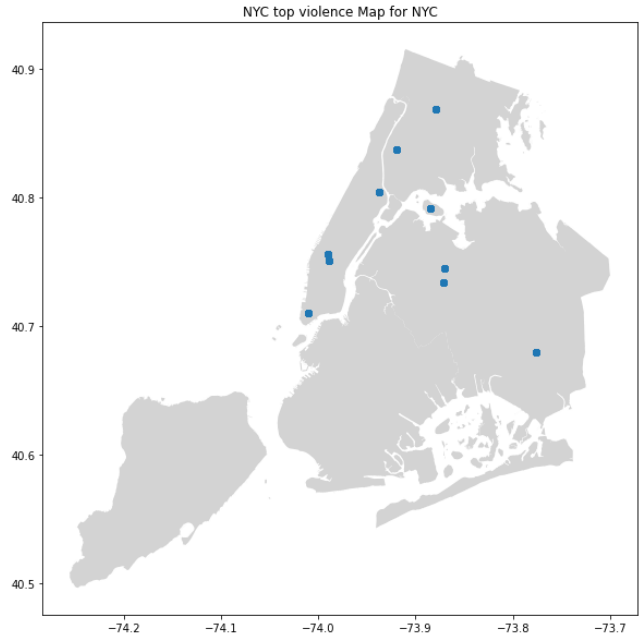

Figure 7. Max Violence points at NYC

7. Evaluating the data:

We calculated the accuracy score for our data manipulation and data cleaning and tested it out on ten sample datasets that had similar columns. Earlier, our approach did not satisfy the similar datasets we found using Auctus Dataset. We improvised our techniques to make it more comprehensive and holistic. By calculating Precision and Recall scores, we measured the effectiveness of our approach. Taking an average of these scores we can find the avg (precision and recall) score of our dataset. We have explained the entire calculation in the result section.

We have also taken the statistics of the NYPD Complaint Dataset and cross verified if there are any NULL or Empty values present after data cleaning.

From the below data, we can see that Data cleaning on NYPD Complaint is successful.

	total	empty	distinct	uniqueness	entropy
CMPLNT_NUM	60641	0	60641	1.000000	15.888006
CMPLNT_FR_DT	60641	0	5024	0.082848	12.109474
CMPLNT_FR_TM	60641	0	1421	0.023433	8.062748
ADDR_PCT_CD	60641	0	77	0.001270	6.141536
KY_CD	60641	0	59	0.000973	4.188476
LAW_CAT_CD	60641	0	3	0.000049	1.360670
LAW_CAT_CD	60641	0	3	0.000049	1.360670
BORO_NM	60641	0	5	0.000082	2.162510
PREM_TYP_DESC	60641	0	70	0.001154	3.544470
VIC_AGE_GROUP	60641	0	13	0.000214	2.202649
VIC_RACE	60641	0	8	0.000132	2.305399
VIC_SEX	60641	0	5	0.000082	1.848961
Latitude	60641	0	30255	0.498920	14.315671
Longitude	60641	0	30216	0.498277	14.312707
SUSP_AGE_GROUP	60641	0	6	0.000099	1.136848
SUSP_RACE	60641	0	7	0.000115	1.632319
SUSP_SEX	60641	0	4	0.000066	1.613879
JURISDICTION_CODE	60641	0	21	0.000346	0.721983
PATROL_BORO	60641	0	8	0.000132	2.896011
PD_CD	60641	0	308	0.005079	5.865226
HOUSING_PSA	60641	0	686	0.011312	1.212591

Figure 8. Cleaned Data Statistics

TOOLS AND TECHNOLOGIES:

- Pyspark (Apache Spark) for parsing and cleaning large datasets.
- Google Collab Notebooks for neat and clean data processing in Python. It helps in generating reproducible notebooks.
- Geopy Package for reverse geocoding.
- Openclean- Python library for Data Profiling and cleaning
- Standard Matplotlib, Numpy, Scipy, Pandas libraries.
- Sample Size Calculator to find the confidence level of our sample dataset.
- Auctus Dataset for searching similar column datasets.
- We cleaned the NYU complaint historic data and then parsed the useful columns for further analysis
- Crime (in the form of reported incidents) was analyzed in total, borough wise and their variation along with the variation in total reported incidents was analyzed.
- For further exploration, we tried collaborating with other datasets like climate, unemployment, and population.

RESULT ANALYSIS

To check if the data profiling and data cleaning methods used for our main dataset are accurate, we have applied the same method used for our main dataset to ten additional datasets. We have carefully chosen the additional datasets and made sure that some of the columns in these new data sets should be similar to some columns of the original data set. (Original Dataset - NYPD complaint historic dataset)

Similar Datasets:

NYPD Arrests Data

https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

NYPD Shooting Incident Data

https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

NYPD Criminal Court Summons Data

https://data.cityofnewyork.us/Public-Safety/NYPD-Criminal-Court-Summons-Historic-/sv2w-rv3k

NYPD Summons Historic Data

https://data.cityofnewyork.us/Public-Safety/NYPD-B-Summons-Historic-/bme5-7ty4

NYPD vehicle collision data

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

NYPD Service Calls Data

https://data.cityofnewyork.us/Public-Safety/NYPD-Calls-for-Service-Historic-/d6zx-ckhd

NYPD Incident Data

https://data.cityofnewyork.us/Public-Safety/NYPD-Use-of-Force-Incidents/f4tj-796d

Rodent Inspection Data

https://data.cityofnewyork.us/Health/Rodent-Inspection/p937-wjvj

Emergency Response Data

https://data.cityofnewyork.us/Public-Safety/Emergency-Response-Incidents/pas

Dataset name	Columns with common data types with respect to the main dataset
NYPD Shooting Incident Data	BORO, PRECINCT, JURISDICTION_CODE, PREP_AGE_GROUP, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, Latitude, Longitude, VIC_RACE, PERP_SEX, OCCUR_DATE, OCCUR_TIME
NYPD Criminal Court Summons Data	SUMMONS_DATE, AGE_GROUP, SEX, RACE,JURISDICTION_CODE,BORO, PRECINCT_OF_OCCUR, Latitude, Longitude
NYPD Arrest Data (Historic)	ARREST_DATE, ARREST_PRECINC, JURISDICTION_CODE,ARREST_BORO, Latitude, Longitude, Lon_Lat, PERP_SEX, PERP_RACE, LAW_CAT_CD, AGE_GROUP
Rodent Inspection	BOROUGH, Latitude, Longitude, Inspection_Date, INSPECTION_TYPE, JOB_TICKET_OR_WORK_ORDER_ID
NYPD Summons Historic Data	EVNT_KEY, VIOLATION_DATE,VIOLATION_TIME, CITY_NM, Latitude, Longitude

Table 1: Top datasets with similar columns with respect to the main dataset

We can see in Table 1 that there are multiple columns which are similar with respect to the main dataset. Same steps like Data profiling, Data Was preprocessing and Data Cleaning have been followed for additional datasets. A sample dataset with the right confidence measure alone is sufficient to analyze the data.

Choosing the dataset size:

We are using a sample size data set calculator as shown in Fig 9 to find the minimum number of rows needed to find the discrepancies in the dataset. By Setting the confidence level as 95% and confidence Interval as "5".

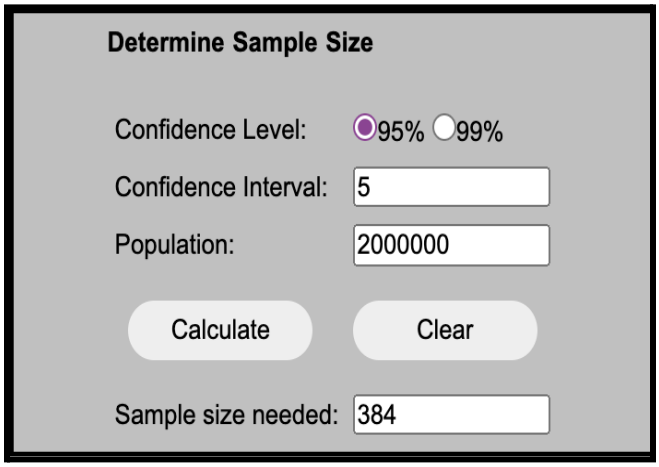

Figure 9: Sample Size Calculator

For Example, for the dataset of size 2 million Rows we need a minimum of 384 rows to analyze the data. Similarly, we have calculated the sample size for all the 10 additional datasets.

Data cleaning and Data Preprocessing techniques have been applied on the 10 additional datasets. We have checked the statistics of the data to make sure the dataset has no empty values.

Now, for each additional dataset, we have the data before cleaning and data after cleaning. By inspecting the data manually we find out the below parameters for data analysis.

After cleaning the dataset:

	total	empty	distinct	uniqueness	entropy
EVNT_KEY	2000	0	2000	1.000	10.965784
VIOLATION_DATE	2000	0	338	0.169	8.030701
VIOLATION_TIME	2000	0	1650	0.825	10.546927
CITY_NM	2000	0	6	0.003	2.304341
Latitude	2000	0	1250	0.625	9.911207
Longitude	2000	0	1250	0.625	9.911207

Figure 10: Dataset profile example after preprocessing

- **True Positive (TP):** True positive represents the value of correct predictions of positives out of actual positive cases.
- **False Positive (FP)**: False-positive represents the value of incorrect positive predictions. This value represents the number of negatives that get falsely predicted as positive.
- **True Negative (TN):** True negative represents the value of correct predictions of negatives out of actual negative cases.
- **False Negative (FN)**: False-negative represents the value of incorrect negative predictions.

The precision score is a useful measure of the success of prediction when the classes are very imbalanced. Mathematically, it represents the ratio of true positive to the sum of true positive and false positive. Recall score helps us measure how many predictions made are actually positive out of all positive predictions made. It identifies all actual positives out of all positives that exist within a dataset. The higher the recall score, the better the system is at identifying both positive and negative examples. Accuracy score is used to measure the system performance in terms of measuring the ratio of the sum of true positives and true negatives out of all the predictions made. And correctly predict an outcome out of the total number of times it made predictions.

Using the True Positive, True Negative False Positive and False Negative values, we calculate the Precision Score and recall Score using the below formula

Precision Score = TP / (FP + TP) – eqn 1
Recall Score = TP / (FN + TP) – eqn 2

Also Accuracy Score can be calculated using the below formula:
Accuracy Score =
(TP + TN)/ (TP + FN + TN + FP) - eqn 3

For example, consider NYPD arrest dataset
The precision score for the "borough" Column was zero before data cleaning as the "borough name" was in an abbreviated format and had some empty values as well. **After cleaning the data, we got the precision score of 0.98 and recall score of 1".** By cleaning the data, we are able to improve the Precision and Recall score. The process of how Precision and Recall are calculated using the equations above is illustrated in Figure 11 below.

```
Earlier, precision and recall were both zero. Because borough names were abbreviated
True Positive = 0
selected elements = 1553
Relevant elements = 1534

Later, when the technique was modified to handle abbreviated borough names
True Positive = 1534
selected elements = 1553
Relevant elements = 1534

precision= 1534/1553
recall = 1534/1534
```

Figure 11: Example of how we calculated Precision and recall for the Arrest Dataset

Precision and Recall score of some of the datasets can be seen in figures 12, 13, 14. The recall values are always greater than 98% in almost all the cases whereas the precision values varied across the datasets.

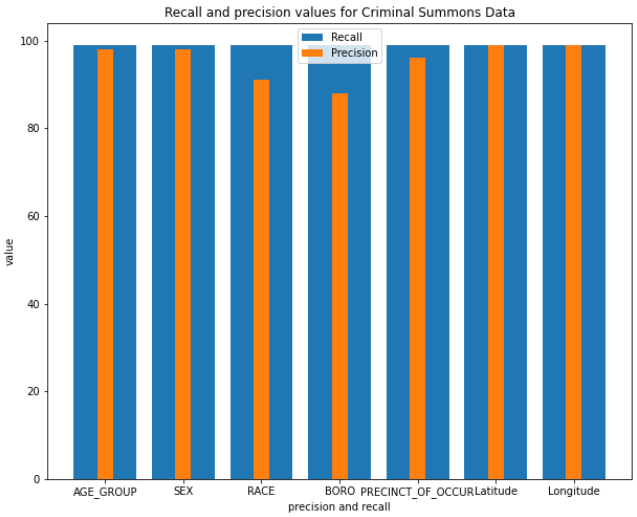

Figure 12: NYPD Criminal Court Summons Data Columns' Precision and recall after cleaning

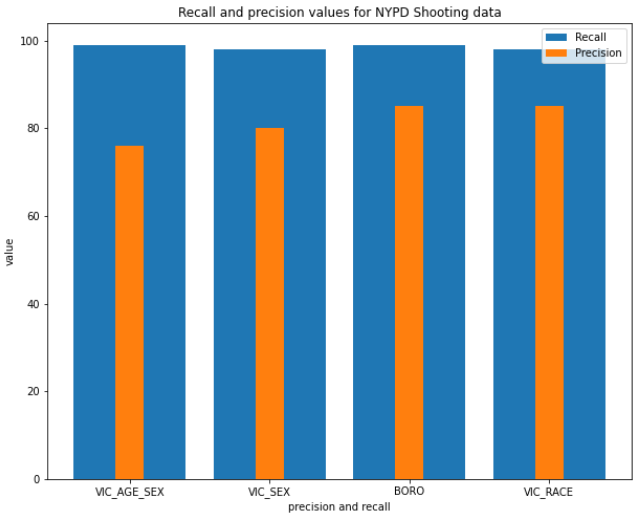

Figure 13: NYPD Shooting Incident Data Columns' Precision and recall after cleaning

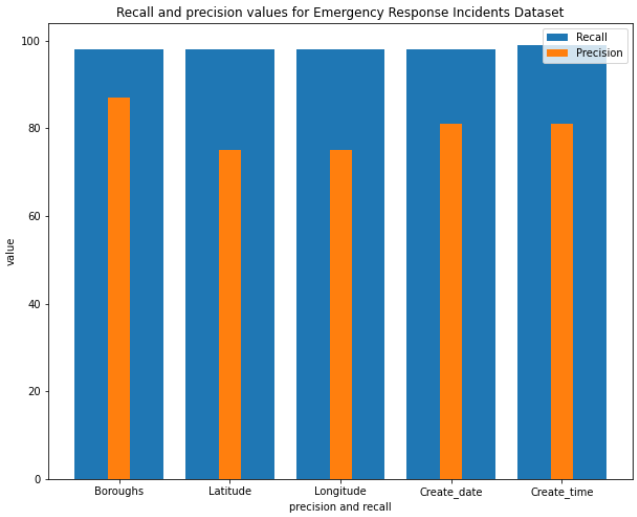

Figure 14: Emergency Response Data Columns' Precision and Recall after cleaning

CORRELATION WITH OTHER DATASETS

1. UNEMPLOYMENT DATA: While observing the graphs for the crime rate in New York City, saw a decrease in total criminal offenses after 2011. This behavior could be noted because of the following analogies we gathered: There was a steady increment in the employment rate after 2010. Data(NYC-employment.csv) was collected from the Economic Research Department of the Federal Reserve.

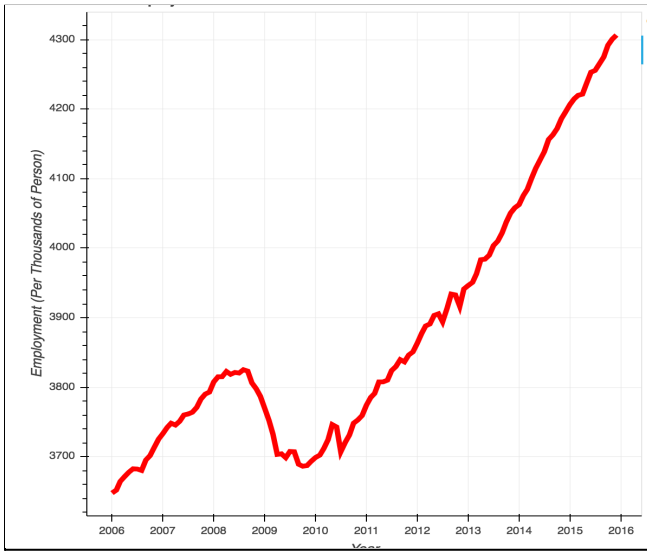

Figure 15: Relation between Employment and Crime data

2. POPULATION DATA: In further exploration, we tried to establish the relationship between *crime rate* and population size and could observe that there was a decrease in population between 2009-2010 and an increase in population size after 2010. We observed in the below graphs crime did not increase, with the increase in population. Estimate of reported crime numbers for each borough.

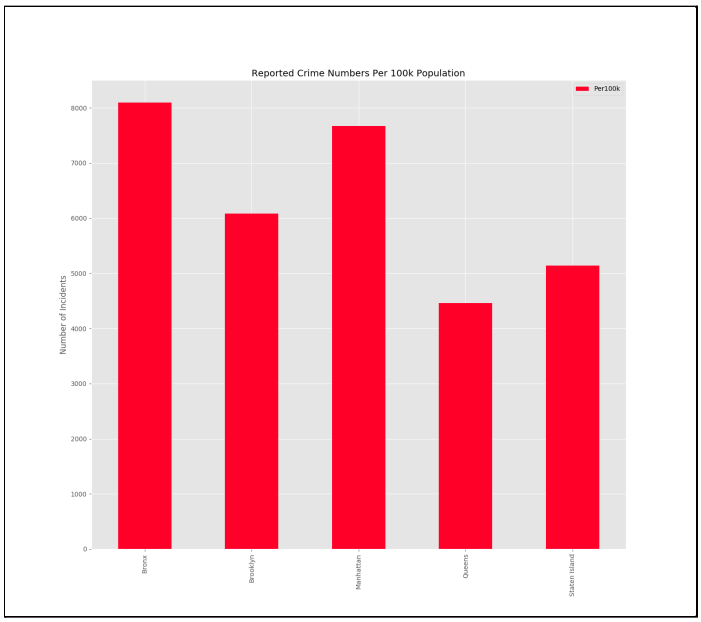

Figure 16: Relation between Population and Crime data

3. WEATHER DATA: We suspected there was a relationship between weather and crime. All of the highest spikes occur on the first of a month. New Year's has the highest number of incidents. The lower spikes, however, seem to be mainly related to external factors. Most of them lie either on a holiday or on a day of bad weather.

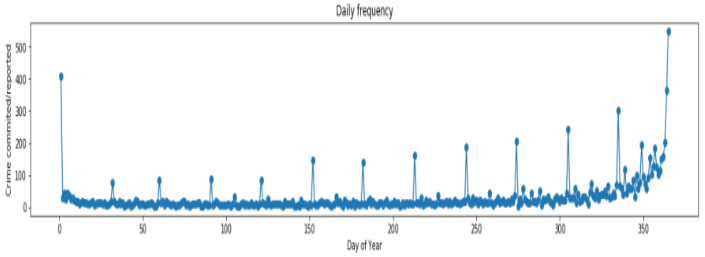

Figure 17: Crime count by day of the year

CONCLUSION

Data cleaning and data analysis plays an indispensable role in the knowledge discovery process of extracting interesting patterns or knowledge for understanding various phenomenons. By finding and understanding the discrepancies in the data we can suggest techniques that will help prospective researchers in analysing and exploring complaints filed.

Cleaned data can be used for civic action and policy change responsibly to expose the hidden patterns and ideologies. Precision-Recall score is a useful measure of success of prediction when the data are very imbalanced.

LIMITATIONS & FUTURE WORKS

We can use the data to visualize the data further and come up with reliable conclusions regarding which neighborhoods have the worst casualties, which months or time of the day had the most incidents, the relationship between the boroughs and crimes, etc.

NYPD crime has been the center of media attention for the last few years and the problem has never been solved. Due to the heated debate on crimes in NYC, we feel compelled to look into the data on violence in NYC. We focus on this because this is what we are familiar with and most care about. By utilizing the cleaned data sets, one can suggest sensible patterns, visualizations about crimes in NYC, certain projections, and even relevant recommendations for policy-makers.

REFERENCES

[1] S. Garc´ıa, S. Ram´ırez-Gallego, J. Luengo, J. M. Ben´ıtez, and F. Herrera, "Big data preprocessing: methods and prospects," Big Data Analytics, vol. 1, no. 1, p. 9, 2016.

[2] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3–13, 2000.

[3] "NYPD Complaint Data Historic (August 2017 updated); Police Department (NYPD); available from: https://data.cityofnewyork.us/public-safety/nypdcomplaint-data-historic/qgea-i56i,"

[4] "City Record Online (Oct 2017 Created; Department of Citywide Administrative Services (DCAS); available from:https://data.cityofnewyork.us/citygovernment/city-record-online/dg92-zbpx),"

[5] Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy

[6] "NYPD Complaint Data Historic (August 2017 updated); Police Department (NYPD); available from: https://towardsdatascience.com/analysis-of-nyc-reported-crime-data-using-pandas-821753cd7e2

[7] "City Record Online (Oct 2017 Created; Department of Citywide Administrative Services (DCAS); available https://a1080211jeff.medium.com/exploring-nyc-analysis-of-crime-data-in-new-york-city-6134642b9833