Project:
NYPD COMPLAINTS DATA ANALYSIS

TEAMMATES:
Hrushi Patel(hp2307)
Hanmisha Voddineni(hv2085)

INTRODUCTION

Although New York City is the most popular city in the United States, it nevertheless still experiences various types of criminal activity throughout the year. In this project we will analyze historic crime data found in the NYPD Complaint Data dataset, identify possible data issues, and implement 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 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 goal of the project is to first clean the data in the original dataset, analyze the strategies used to clean the data and record the results. We want to implement different big data strategies to get the most efficient dataset making it easier for the researcher to use this data to take preventive steps and reduce the crime in NYC.

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 data type 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 or completed
- Invalid Jurisdiction code and Precinct
- Invalid coordinates that do not belong to NYC

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.
● Openclean- Python library for Data Profiling and cleaning
● Standard Matplotlib, Numpy, Scipy, Pandas libraries.
● 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.

MAIN STEPS

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 Preprocessing:

I. **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.

Profiling

```
dp=stream(final_dst)
profiles = dp.profile(default_profiler=DefaultColumnProfiler)
profiles.stats()
```

✓ 1m 58.7s

	total	empty	distinct	uniqueness	entropy
CMPLNT_NUM	7375993	0	7373143	9.996136e-01	22.813633
CMPLNT_FR_DT	7375993	655	8606	1.166862e-03	12.425578
CMPLNT_FR_TM	7375993	48	1441	1.953648e-04	8.136466
CMPLNT_TO_DT	7375993	1704204	6825	1.203324e-03	12.417984
CMPLNT_TO_TM	7375993	1699541	1441	2.538558e-04	8.862856
ADDR_PCT_CD	7375993	2166	77	1.044234e-05	6.148690
RPT_DT	7375993	0	5479	7.428152e-04	12.405384
KY_CD	7375993	0	74	1.003255e-05	4.170727
OFNS_DESC	7375993	18823	71	9.650450e-06	4.006583
PD_CD	7375993	6278	432	5.861828e-05	5.913459
PD_DESC	7375993	6278	422	5.726137e-05	5.904814
CRM_ATPT_CPTD_CD	7375993	7	2	2.711502e-07	0.124481
LAW_CAT_CD	7375993	0	3	4.067249e-07	1.373922
BORO_NM	7375993	11329	5	6.789176e-07	2.162442
LOC_OF_OCCUR_DESC	7375993	1543800	5	8.573104e-07	1.241785
PREM_TYP_DESC	7375993	40745	74	1.008828e-05	3.579655
JURIS_DESC	7375993	0	25	3.389374e-06	0.668693
JURISDICTION_CODE	7375993	6278	25	3.392261e-06	0.668795
PARKS_NM	7375993	3229345	1206	2.908373e-04	0.110850
HADEVELOPT	7375993	7029181	279	8.044704e-04	7.277703
HOUSING_PSA	7375993	1241432	5103	8.318444e-04	1.216803
X_COORD_CD	7375993	17339	71343	9.695115e-03	14.595751
Y_COORD_CD	7375993	17339	73933	1.004708e-02	14.744968
SUSP_AGE_GROUP	7375993	4795235	111	4.301062e-05	2.113260
SUSP_RACE	7375993	3426694	8	2.025676e-06	2.224526
SUSP_SEX	7375993	3560008	3	7.861666e-07	1.323637

II. Data Cleaning:

The main objective of this step is to identify the columns with empty values and extract the most meaningful ones. Created a scatter plot visualization with column names on the x axis and empty values count on the y axis. Transit_District, Station_Name has the most empty values. We have identified 23 columns that have the 0 empty values columns out of 35.

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_TM
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
ADDR_PCT_CD
The precinct in which the incident occurred

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
BORO_NM
The name of the borough 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.
JURIS_DESC
Description of the jurisdiction code
JURISDICTION_CODE
Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3), like Correction, Port Authority, etc.
PARKS_NM
Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT
Name of NYCHA housing development of occurrence, if applicable
HOUSING_PSA
Development Level Code
X_COORD_CD
X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD
Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
SUSP_AGE_GROUP
Suspect’s Age Group
SUSP_RACE
Suspect’s Race Description

SUSP_SEX
Suspect's Sex Description
TRANSIT_DISTRICT
Transit district in which the offense occurred.
Latitude
Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude
Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Lat_Lon
Geospatial Location Point (latitude and Longitude combined)
PATROL_BORO
The name of the patrol borough in which the incident occurred
STATION_NAME
Transit station name
VIC_AGE_GROUP
Victim's Age Group
VIC_RACE
Victim's Race Description
VIC_SEX
Victim's Sex Description

3. 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.

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".

- **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.

- **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.

4. Data Exploration and Visualization:

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 figure below most crimes are happening on the first day of every year.

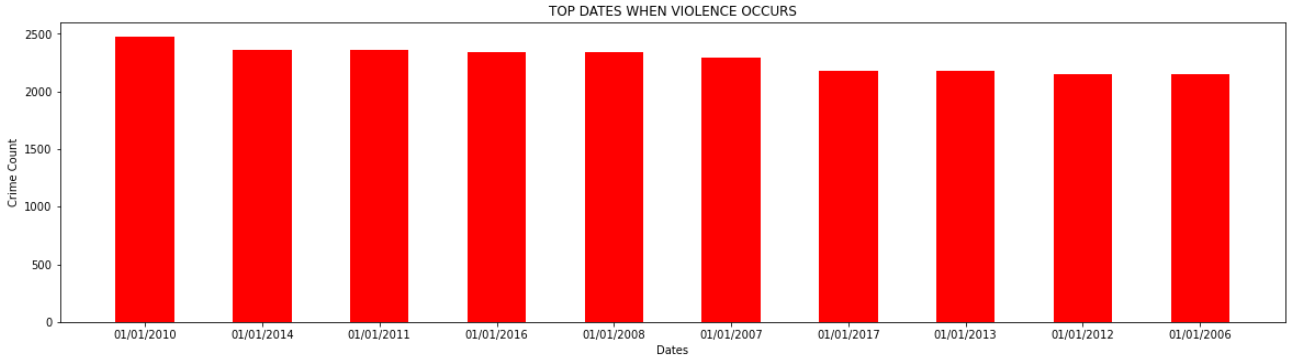

Fig. Max crime occurring dates

b) 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.

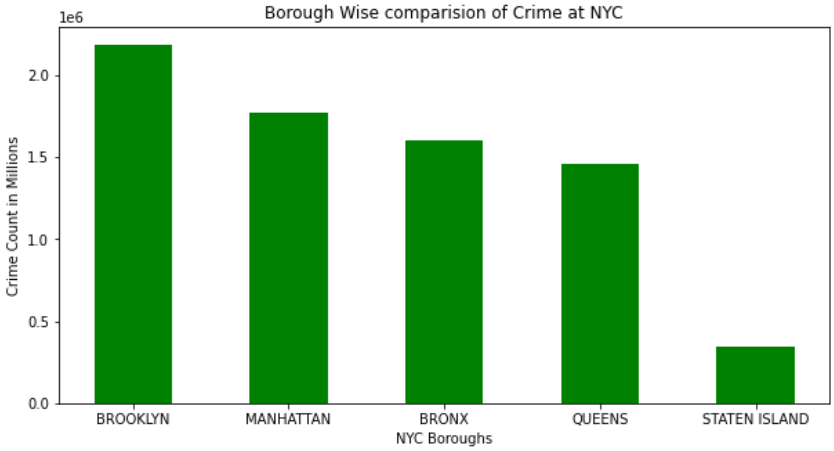

Fig. Boroughs vs Crime Count

c) As per figure below 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.

Fig. Max Violence points at NYC

d) As per below figure which is comparison of crime based on suspect sex, male outnumbers females when it comes to being suspect for the crime.

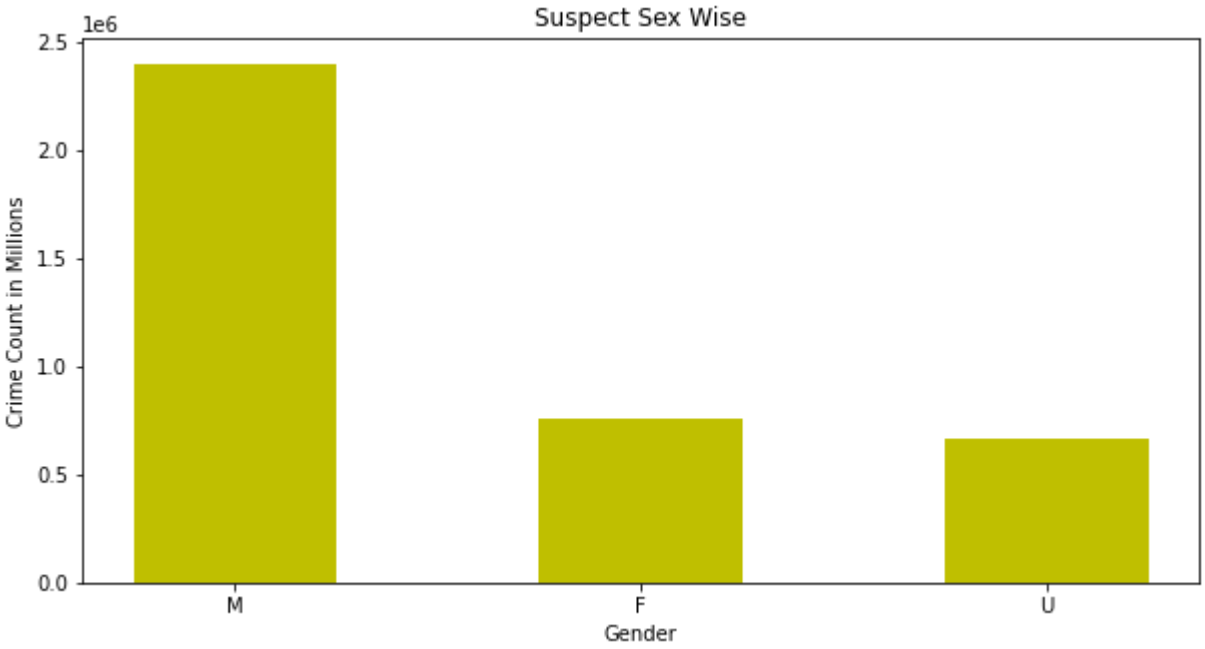

Fig. Suspect Sex Wise at NYC

e) As per below figures which is comparison of victims sex in the crime. In a crime, females have been victimized more than the males.

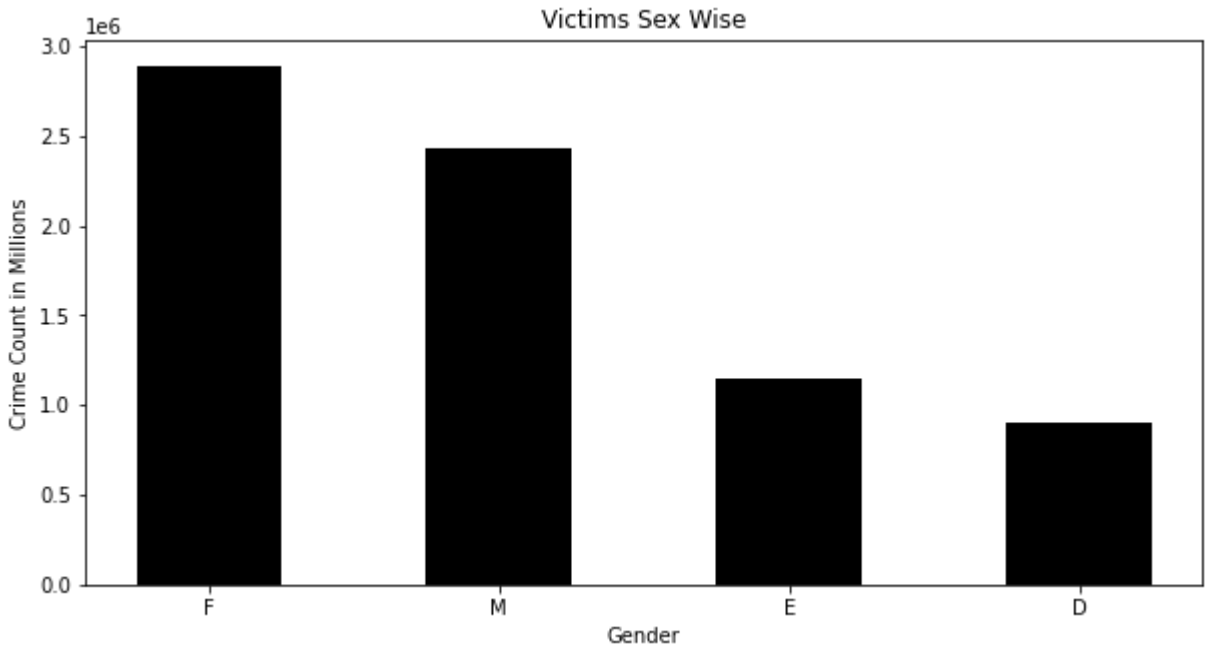

Fig. Victim Sex Wise at NYC

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

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 analyzing 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

The end data is saved and this can be used in different ways. This data can be used to visualize the data further and come up with reliable conclusions regarding which neighborhoods have the worst casualties, which dates had the most incidents, the relationship between the boroughs and crimes, etc. 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/nypdc complaint-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-zbox),”

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

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