

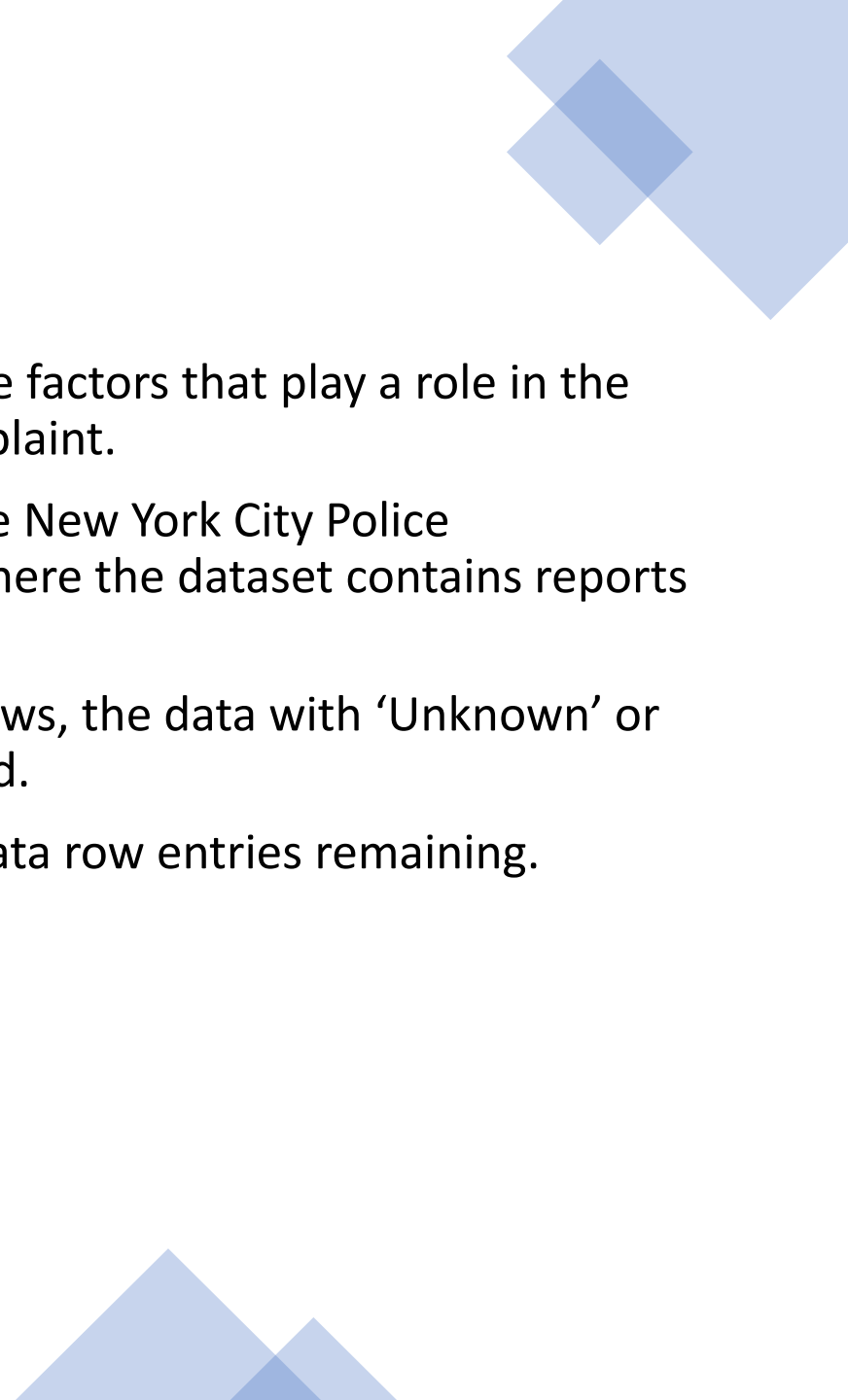


New York City Crime Complaints

IST 718: Big Data Analytics
By Shivani Sanjay Mahaddalkar

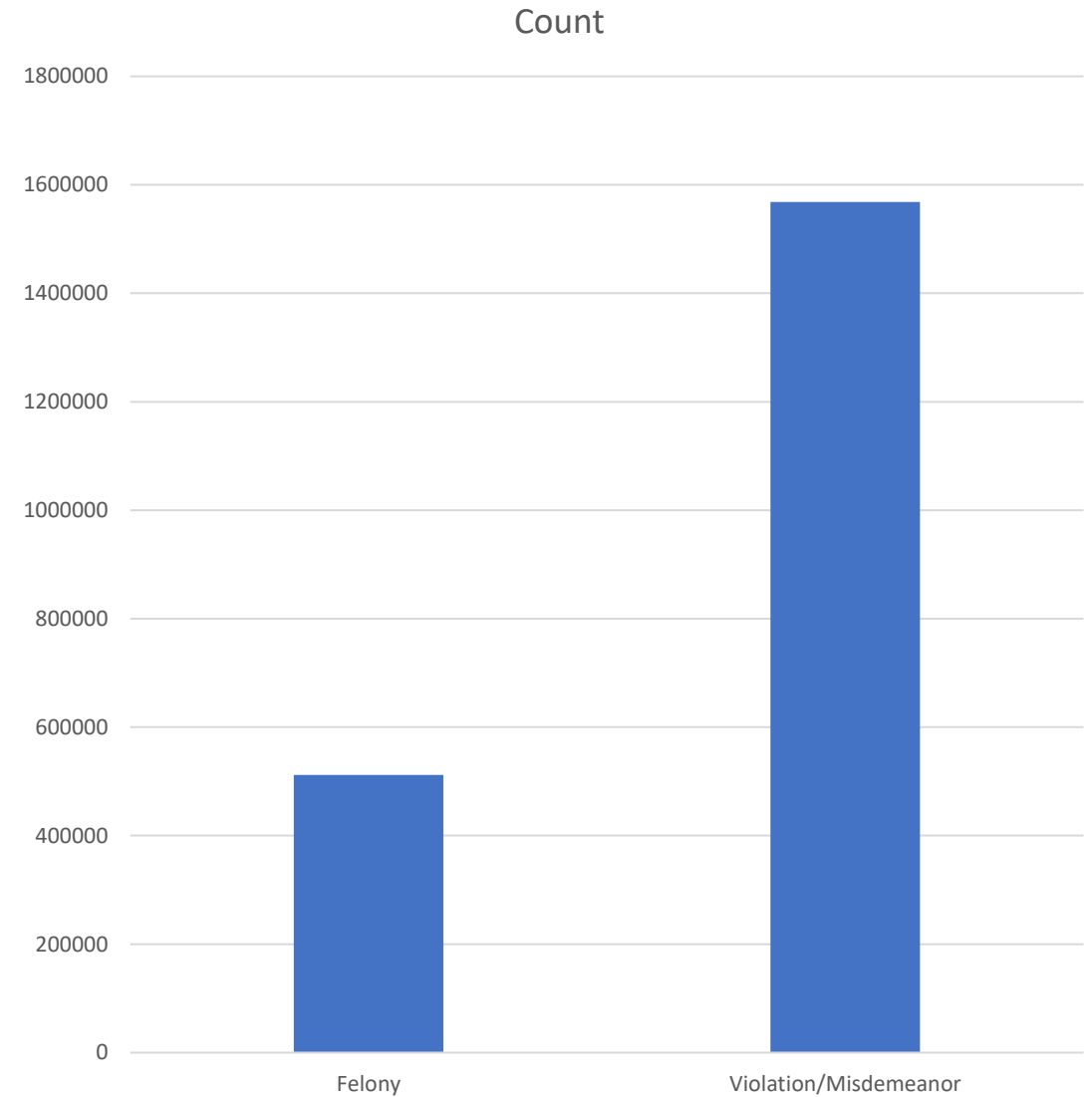


Problem Description

- Objective: Understand the factors that play a role in the level of severity of a complaint.
 - The data is taken from the New York City Police Department's website, where the dataset contains reports from 2006 to 2019.
 - It contained 6.9 million rows, the data with 'Unknown' or blank entries was dropped.
 - The data has 2 million+ data row entries remaining.
- 


Data Preprocessing

- The target variable: Level of offense are processed into 2 levels:
Felonies coded as 1
Violations and misdemeanors coded as 0





Data Preprocessing

- Columns that had more than 95% null values like 'Park Name', 'Station Name', etc. were dropped
 - Columns that gave direct information about the target variable like 'Key code of offense', 'Description of offense', were dropped
 - Columns that gave direct information about other independent variables like 'Jurisdiction description' about the 'Jurisdiction Code', 'Patrol Borough' about the 'Borough' were dropped
 - Redundant columns like the 'Complaint Index', 'Internal classification code' were dropped.
- 

Data Preprocessing

Columns that had more than 95% null values like 'Park Name', 'Station Name', etc.

Columns that gave direct information about the target variable like 'Key code of offense', 'Description of offense'

Columns that gave direct information about other independent variables like 'Jurisdiction description' about the 'Jurisdiction Code', 'Patrol Borough' about the 'Borough'

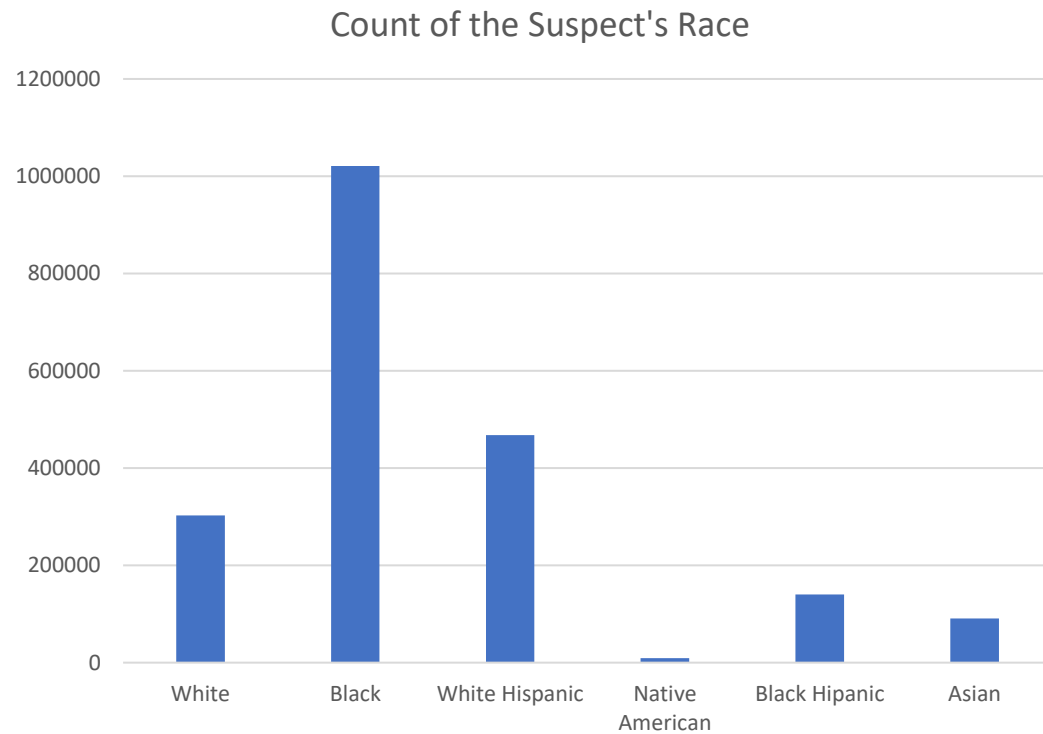
Redundant columns like the 'Complaint Index', 'Internal classification code'

Data Preprocessing

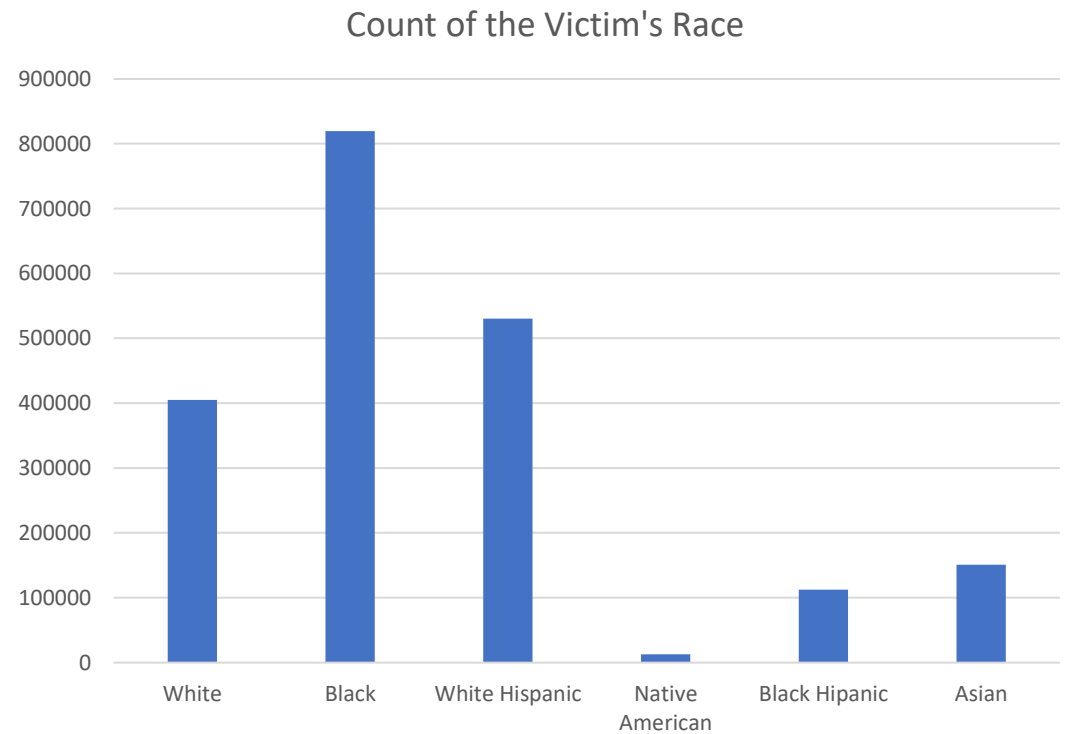
Attempt	Borough	Jurisdiction	Premises	Location	Vic's Age	Race	Sex	Time	Season
<ul style="list-style-type: none">• Completed• Attempted	<ul style="list-style-type: none">• Queens• Brooklyn• Manhattan• Bronx• Staten Island	<ul style="list-style-type: none">• Police• Others	<ul style="list-style-type: none">• Residential• Public	<ul style="list-style-type: none">• Inside• Outside	<ul style="list-style-type: none">• <18• 18-24• 25-44• 44-65• 65+	<ul style="list-style-type: none">• White• Black• Native American• Black Hispanic• White Hispanic• Asian• Other	<ul style="list-style-type: none">• Male• Female	<ul style="list-style-type: none">• Day• Night	<ul style="list-style-type: none">• Winter• Spring• Summer• Autumn

Data Visualization

Suspect's race




Victim's race





Logistic Regression Model


- Each independent column was transformed to a vector with one hot encoding.
 - Logistic regression was performed, where the target variable was the severity of the crime, with 0 being a violation and 1 being a misdemeanor/felony.
 - The area under the curve for the model was 0.6735
 - However, since this is an inherently biased data set we need to look at the weights of the independent variables.
- 

Results

- Initial hypothesis was to assess whether the race of a person impacts the severity of the crime they complain about and are complained against, the model shows that a Black or a Hispanic person is complained against for more severe crimes, whereas a White, Native American or an Asian is complained against for pettier crimes.
- We see the trend reversed slightly in case of the race of the victim. A black person is likelier to report pettier crimes, whereas a White, White Hispanic or an Asian person are more likelier to report a felony.
- Women are more likely to report pettier crimes, whereas men are more likely to report a serious crime.
- Brooklyn and Queens are likelier to report a felony.
- A felony is likelier to be reported that has happened in a public setting.
- People in the age category 65+ are likelier to report a felony.



Caveats to Note

- As this is data that is taken from the real world that relies heavily on the human instinct, the data is inherently biased.
 - To gauge if there was influence due to the race of the people involved, in one of the initial steps, that data was cleaned to include complaints that have actual values, it could have made the data even more biased.
 - One way to further check would be to sample data points from the previously dropped rows and include in an analysis like this.
 - While cleaning the data, some groups that are marginalized minorities have not been considered due to the lack of representation in the data. Which continues to cyclically enhance the problem of the lack of representation.
- 



Thank you