George Mason University

CS 504

Spring 2024

NYC Crime Data
Project Paper
DL4 Project 1

Allison Forsyth
Harshini Kunjeti
Sadan Khan
Yashaswi Gurram
Nithish Kumar Amjolu
Vikram Datth Bolloju

Table of Contents

Table of Contents

Abstract

This project analyzes historical shooting data from NYC Open Data provided by the New York Police Department to gain insight on shooting trends and patterns with respect to when the shooing's happened, where they occurred, and who was involved. The work aimed to unveil seasonal trends, hot spots, and demographic profiles associated with the shooting events in New York City. The goal of this project is to get a better understanding of shootings in the city to help the police, city leaders, and communities make informed safety decisions and policy changes. Research methods included various statistical analyses and visualizations conducted in Python and R. Analyses covered various attributes of shooting records including victim and perpetrator demographics and location of incidents. Analyses and algorithms comprised of prediction algorithms (random forest trees), forecasting, hot spot analysis, and exploratory data analysis and visualizations of trends and attributes. Work on this project revealed higher shooting incidents in specific NYC boroughs (Bronx and Brooklyn) as well as the distributions of perpetrator demographics such as age, sex, and race. Data analysis revealed that ~43% of identified perpetrators were ages 18-24, ~73% were Black/African American, and ~97% were male. The project also revealed seasonal trends in the number of shooting incidents to be higher in the summer months. These insights emphasize the critical role of data-driven strategies in crafting effective public safety measures and policies. We recommend targeted interventions in identified high-crime areas and the adoption of evidence-based policies addressing the root causes.

Introduction

Background and Rationale

New York City is home to over 8 million people [1]. Composed of 5 boroughs, Manhattan, Queens, Staten Island, the Bronx, and Brooklyn, New York City is a densely populated urban area with cultural influences from all over the world [1]. NYC is also home to the largest police force with almost 36,000 officers [2]. In a city so large and densely populated, the New York Police Department (NYPD) is responsible for ensuring public safety, law enforcement, emergency response, and more [2]. Trusted with protecting a city of 8 million people, the NYPD must use its resources effectively and efficiently. Throughout this project, historical data in NYC shootings will be analyzed to better understand trends in shooting incidents and gather insights about factors such as where shootings are occurring, and who is involved.

Research

Since January last year, 2023, the NYPD has reported a decrease in most major crime categories [3]. They reported substantial decreases in murder, rape, burglary, and felony assault and only small increases in robbery and larceny [3]. This overall decrease in crimes follows a new trending decline since 2022 [4]. While the most recent years' crime rates are higher than in 2020 and 2021, they are lower than pre-pandemic, 2019, rates [4].

In the past, the NYPD has used data to implement tactics aimed at decreasing crime rates, such as patrolling areas with high crime saturation. In 2003, Operation Impact began, increasing the number of officers in high-crime concentration areas or "impact zones" [5]. The increased police presence and investigative stops showed an overall decrease in major crimes and an increase in arrests [5]. While investigative stops were encouraged, the city found that a majority of stops did not have an impact on crime reduction, suggesting that the increased police presence was the main factor in reducing crime [5]. It is unclear if Operation Impact is still in effect but, it is assumed that the NYPD still uses similar tactics to target high-crime areas and neighborhoods.

As indicated by the presence of "impact zones", the crime rates and statistics seen across the city differ in New York neighborhoods and boroughs. In an article written by Michael S. Barton in 2016 [6], he found that neighborhoods that experienced high rates of gentrification also experienced a significant decline in major crimes. This relationship between crime and gentrification suggests social and economic impacts on crime rates in neighborhoods. Further research from the Manhattan Institute [7] found that the most populated boroughs had the most crime hot spots. Continued analysis and research in recent crime statistics could give NYC stakeholders a new perspective on crime trends and create change in policy and policing to continue to drive down NYC crime.

Project Objectives

Throughout this project, the team aims to analyze various historical shooting data collected from NYC Open data [8] to gain insight on crime statistics including:

Demographic Profiling: Investigate the demographics of both victims and perpetrators to understand correlation between different attributes with shooting incidents.

Hot Spots: Identify areas with high shooting rates areas that may require increased police presence.

Trends: Conduct analysis of shooting data to identify trends yearly and seasonally.

Problem Space

In order to enhance public safety and inform decision-making processes there is a need for comprehensive analysis and understanding of crime in New York City. These datasets provide a detailed view of shooting incidents across different boroughs, including critical information such as incident dates, times, locations, descriptions, perpetrator details, victim information, and specific location coordinates. By analyzing this data, people can gain insights into shooting trends, demographic correlations, hot spots, and patterns of criminal activities within the city. The problem space involves utilizing this dataset to identify strategic resource allocation needs for law enforcement agencies, policymakers, and community leaders, and implement targeted interventions and evidence-based policies to combat crime effectively.

Primary User Story

Based on the user context and value proposition, we developed the following primary user story to guide our project:

As a policy maker in New York City, I want to access detailed analyses of shooting patterns, demographic profiles, and hot spots through an intuitive platform, enabling me to understand the underlying trends and allocate resources more effectively. This insight will help me devise targeted policies and interventions aimed at reducing crimes and improving public safety across the city.

Solution Space

Our research aims to deliver value to its readers by providing insightful analyses of shooting data sourced from NYC Open Data. This analysis will offer comprehensive and accurate assessments of trends, hot spots, and demographic correlations. Also, this research should equip our readers with valuable insights to address public safety concerns.

Readers will derive value from our analysis by gaining a deeper understanding of shooting patterns within New York city and its boroughs, which can then be utilized to make informed decisions. This will empower law enforcement agencies, policymakers, and community leaders to allocate resources strategically. Additionally, it will help with implementing targeted interventions, and devise evidence-based policies to combat shootings effectively.

Our analyses will also enhance community engagement by fostering transparency and accountability in addressing crime-related issues. This information will promote data driven discussions and collaboration between stakeholders. The findings should facilitate the initiatives of proactive strategies aimed at improving public safety and building trust between law enforcement agencies and local communities.

Overall, we expect our research and analysis to contribute to the reduction of crime rates and increase in public safety. We expect to create a positive impact on crime prevention efforts and increase community development initiatives to encourage a safer environment for all residents and visitors of New York City.

Product Vision - Sample scenarios

Scenario #1

A law enforcement official schedules patrol routes.

A law enforcement official in charge of organizing patrol routes around New York City is interested in optimizing patrol route efficacy to effectively deter crime. Using the data analysis tool created will allow the police to pinpoint regions with a high shooting rate and modify patrol routes appropriately by giving users access to analysis of past and present shooting data, including trends, hotspots, and demographic relationships. Unlike conventional techniques that depend on obsolete or general records our tool provides interactive mapping that shows shooting patterns and hotspots over time, enabling officers to more wisely deploy patrol resources. While the tool offers a unique way to prioritize police deployment, the precision and timeliness of the data entered into the system determine how useful the tool is.

Scenario #2

Helping policy makers make better city safety decisions.

This product could help those who make laws and policies for New York City, focusing on making it safer. It will allow users to use real numbers and facts to decide on the best ways to reduce shootings, rather than just going by what people say or political pressures. The tool is a high-tech system that looks at shooting numbers alongside other factors like demographics to figure out why crime happens where it does. It lets this policy maker see if what the city's been doing about shootings is working or not, and where new ideas might help. Unlike the old way of making safety rules, which might not always look at the full picture or use the latest data, this product gives a full overview of shooting factors to help come up with well-rounded solutions. While the product gives real time data and information on shootings, making good policies also involves thinking about the bigger picture, including politics and money, which our tool might not cover entirely.

Definition of Terms:

GIS – Geographical Information Systems
NYC – New York City
NYPD – New York Police Department

Data Acquisition

Overview:

The Historical NYPD Shooting Incident Data details 27,312 records about the shooting incidents in New York City. It details information of how many precincts (77 unique ones), boroughs (5), and diversified details for the location of the incident, timings, and the general and specific details for the victim and perpetrator. It shows the spatial location in the dataset on maps with latitude and longitude how the incidents fall and make categorical distinctions as to where the incident occurs spatially, to whom the jurisdiction of the incident belongs, and if it falls under statistical counts as murders. The demographics on age, sex, and race of the perpetrator and victims show an astonishing variety, reflecting urban crime's heterogeneous, complex socio-demographic aspects. It is against this background that it is within this diversified urban landscape like that of New York City that the use of this dataset becomes pertinent in telling the stories behind these numbers in terms of trends, risk factors, and informing strategies for the enhancement of public safety.

1.1 Field Descriptions:

- INCIDENT_KEY (Type: Text) - Randomly generated persistent ID for each incident.
- OCCUR_DATE (Type: floating_timestamp) - Exact date of the shooting incident.
- OCCUR_TIME (Type: Text) - Exact time of the shooting incident.
- BORO (Type: Text) - Borough where the shooting incident occurred (Bronx, Brooklyn, Manhattan, Queens, Staten Island).
- LOC_OF_OCCURE_DESC (Type: Text) - displays if the shooting took place inside or outside.
- PRECINCT (Type: Number) - Precinct where the shooting incident occurred. New York City is divided into 77 police precincts.
- JURISDICTION_CODE (Type: Number) - Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non-NYPD jurisdictions.
- LOCATION_DESC (Type: Text) - Location of the shooting incident. This includes a variety of locations. Where location is not available, rows are blank, some contain “(null)” or “NONE”
- STATISTICAL_MURDER_FLAG (Type: checkbox) - Shooting resulted in the victim’s death which would be counted as a murder. This column contains True, or False.
- PERP_AGE_GROUP (Type: Text) - Perpetrator’s age within a category. The categories are, <18, 18-24, 25-44, 45-64, 65+, (null).
- PERP_SEX (Type: Text) - Perpetrator’s sex description.
- PERP_RACE (Type: Text) - Perpetrator’s race description.
- VIC_AGE_GROUP (Type: Text) - Victim’s age within a category.
- VIC_SEX (Type: Text) - Victim’s sex description.
- VIC_RACE (Type: Text) - Victim’s race description.
- X_COORD_CD (Type: Text) - Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104).
- Y_COORD_CD (Type: Text) - Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104).

- Latitude (Type: Number) - Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326).
- Longitude (Type: Number) - Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326).
- Lon-Lat (Type: Point) - Longitude and Latitude Coordinates for mapping.

Data Context:

This is a comprehensive dataset of shooting incidents reported to the New York Police Department (NYPD) from 2006 to 2024. The dataset contains detailed information about each incident, including the location, time, perpetrator and victim demographics, and whether the incident resulted in a murder.

The incidents are spread across the five boroughs of New York City, with the highest concentrations in the Bronx, Brooklyn, and Queens. The majority of the perpetrators and victims are young adult males, with the 25-44 age group being the most common. The race of the perpetrators and victims is predominantly Black, with some incidents involving White Hispanic, White, and Asian/Pacific Islander individuals. A significant portion of the incidents occurred in multi-dwelling apartment buildings, including public housing complexes. Several incidents resulted in murders, indicating the severity of the shooting events.

Data Conditioning:

We will be performing the following data conditioning steps:

Missing Values Handling: Our data's a bit like a jigsaw puzzle with some pieces missing. We will be dealing with it.

Data Type Conversion: We've got to ensure that dates, times, and places are all in formats we can understand and use. It's like making sure everyone at a global conference speaks a common language

Incident Categorization: It's helpful to categorize incidents so we can spot patterns. Are certain types of crimes happening more often in certain places? It's a bit like sorting your emails into folders.

Data Aggregation: We're pulling together all the data to get a broader view. This could mean looking at crime trends by neighborhood or over time. It's like stepping back to see a mural instead of focusing on each brushstroke.

Data Quality Assessment: Even after we've tidied things up, we'll keep checking to make sure everything stays accurate and consistent. It's a bit like proofreading your work to catch any typos.

By tackling these steps, we're setting the stage for some real insights into crime in New York City, making sure we're working with the clearest, most complete picture possible. It's all about laying a solid foundation before we start connecting the dots.

Data Quality Assessment:

To inspect the quality of the chosen data set, we will evaluate the data on its completeness, uniqueness, accuracy, atomicity, and conformity. The data is missing a total of 94165 values, ~16%, in the dataset. The majority of missing values come from the location of occurrence column and location classification column. These missing values occur in earlier data records indicating that the data points were likely not collected until recently, 2022. Other missing data occurs in the perpetrator age, sex, and race columns indicating that the perpetrator has not been identified. To ensure the completeness of the data we analyze, we will not be looking at the location of occurrence or location classification columns. Checking for duplicates and uniqueness of the data, there are 0 repeating records ensuring our data is unique. Because the source of the data comes from an established state government resource, we can assume the data is accurate in terms of its record keeping and data type structures. Checking the data set for atomicity, we use the is.atomic() function in R to check each of the columns. All the columns came out to be true in their atomicity with one value in each record field. Inspecting the data types, we need to change date and time of the shooting records to datetime format and the longitude latitude to Geo point format to ensure data conformity. Overall, the data set is of high quality with little cleaning needed to apply analytical techniques and algorithms.

2.6 Other Data Sources

We did consider other datasets, but we decided not to use them as we wanted to narrow down our research just to the shooting crime and decided not to take other crime datasets for our research analysis. We made sure that the dataset in consideration has all the columns needed for our analysis.

Analytics and Algorithms

The research studies the age and racial distributions of the victims and perpetrators of shooting incidents across New York City as well as the locations of shootings. This present study aims to bring forth some important patterns that can, in turn, help shape more specific interventions and better-informed policy changes related to gun violence.

Analytical Techniques and Tools

To perform comprehensive statistical analysis, Python libraries like Pandas for data manipulation and Matplotlib for data visualization were used. In addition to Python, R was also used for visualizations. We used multiple analytical techniques such as summary statistics, random forests, and forecasting to answer key research questions such as “What are the distributions of victim and perpetrator demographics?”, “Are there trends in the historical shooting data?”, and “What are the hot spots of shooting incidents?”. The code used in this analysis was all open source and includes small to medium size code that varies in difficulty level. If the NYPD were to implement the code created from the algorithms and analytics it would be suggested to store the code in a repository owned and supervised by the NYPD to update the analytics and algorithms with new data.

Demographic Distribution:

This work enabled us to identify which age groups are more affected by or involved in the shootings by providing a distribution of ages among victims and perpetrators. Preliminary analysis indicated that some age groups are more involved than others, with trends related to age that may produce focused prevention. The racial analysis points to the predominant racial demographics among victims and offenders. This takes a dimension of cultural implications and emphasizes how interventions need to have a cultural setting. Lastly, sex of victims and perpetrators was analyzed. Preliminary analysis indicated that males were most involved in shooting incidents as both perpetrators and victims.

Predicting Perpetrator Demographics:

Using summary statistics and filtering, we investigated the relationships between victim and perpetrator demographics. For each of the demographics, age, sex, and race, we filtered out null or "unknown" values, counted the number of occurrences where the victim demographic matched the perpetrator demographic, and divided by the number of records from the filtered data frame to get the percentage of interrace, intersex, and inter-age shootings. The statistics were performed in Python with the following results:

Demographic	Percent of Matching Perpetrator and Victim Demographics
Age	45.93%
Sex	83.89%
Race	68.55%

Table 1: Percent of Perpetrator and Victim Matching Demographics

To delve deeper into the common demographics between victims and perpetrators, we looked at the distribution of each group within sex and race. The following tables display the percentage of each subgroup that contributes to the percentage of matching perpetrator and victim demographics:

Sex	Percent of Matching Perpetrator and Victim Sex
Male	99.48%
Female	0.052%

Table 2: Percent of Perpetrator and Victim Matching Demographics-Sex

Race	Percent of Matching Perpetrator and Victim Race
Black	85.34%
White	1.48%
Black Hispanic	3.24%
White Hispanic	9.45%
Asian/Pacific Islander	0.49%
American Indian/Alaskan Native	0%

Table 3: Percent of Perpetrator and Victim Matching Demographics-Race

Given the correlation between victim and perpetrator demographics, we used classifier algorithms to test the prediction of perpetrator demographics. The historical shooting data from the NYPD currently has 27,312 records. Of those records, only 13,940 (51%) had all demographics perpetrator demographics suggesting many perpetrators in shooting incidents have not been caught. This emphasizes the need for an accurate prediction model that could help the NYPD create a profile of a possible suspect.

Multi-output Random Forest Classifier:

The label values in the model were PERP_SEX, PERP_RACE, and PERP_AGE_GROUP. The features used were BORO, VIC_SEX, VIC_RACE, and VIC_AGE_GROUP. The algorithm was created using a multi-class multi-output random forest tree model from Scikit Learn. This model is complex in that it can handle several joint classification tasks with two or more classifiers. It is a supervised machine learning algorithm that randomly selects features to build decision trees and averages the results. First, the data is cleaned, and we perform feature engineering on the data set (i.e. nominal classifiers get transformed into dummy variables). Then we split the data into training and test data sets, for this model we used 30% as the test data. Next, we specify our random forest model as a muti-output classifier and fit the model to the training data. Finally, we use our training data to predict the perpetrator demographics and evaluate our predictions to the true values. There are currently no metrics in Scikit Learn that support the multi-class multi-output classifier models so, we manually evaluated our output comparing the predictions to the true test data. The table below displays the percentage of demographics correctly classified with the model:

Perpetrator Demographic	Percent Accuracy
Age	52.25%
Sex	97.47%
Race	73.36%

Table 4: Prediction Accuracy by Correct Demographics

Additionally, of the three predictions being made from the model, we evaluated how many were correctly classified for each record seen in the table below:

Number of correct Labels	Percent of Labels Correctly Classified
All Labels	38.64%
Two Labels	46.20%
One Label	14.75%
None	0.41%

Table 5: Prediction Accuracy by Number of Correct Labels

From the above metrics, we assume our model is performing well. ~85% of our predicted classes have 2 or more labels correct. We can continue evaluating our model by exploring each class separately. For each classifier, we create a confusion matrix that displays the true and predicted labels and the percentage of the predicted value for each.

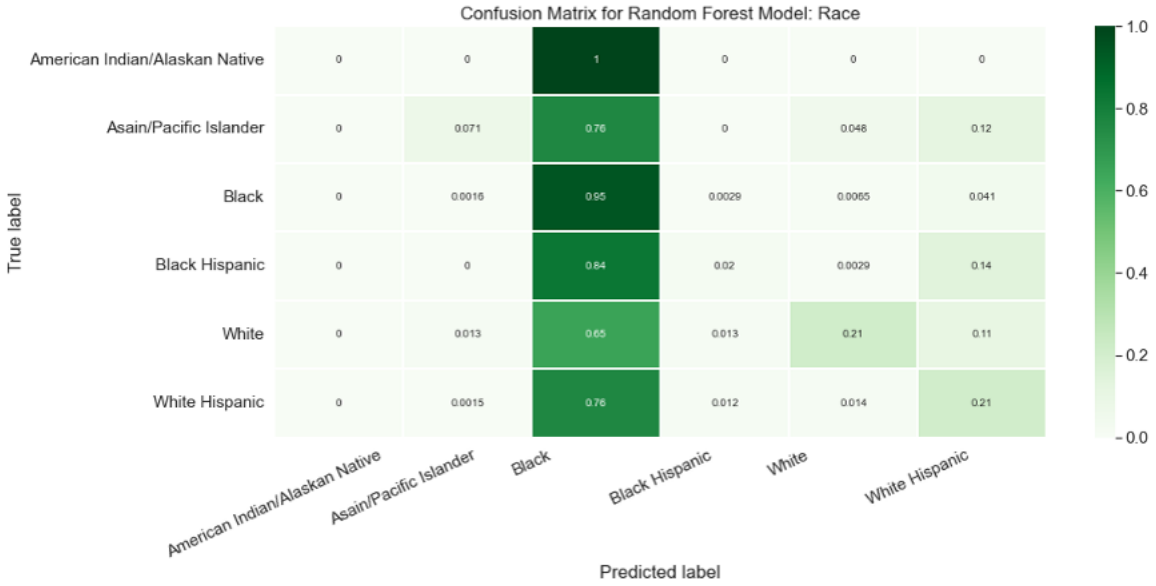

Figure 1: Confusion Matrix - Race

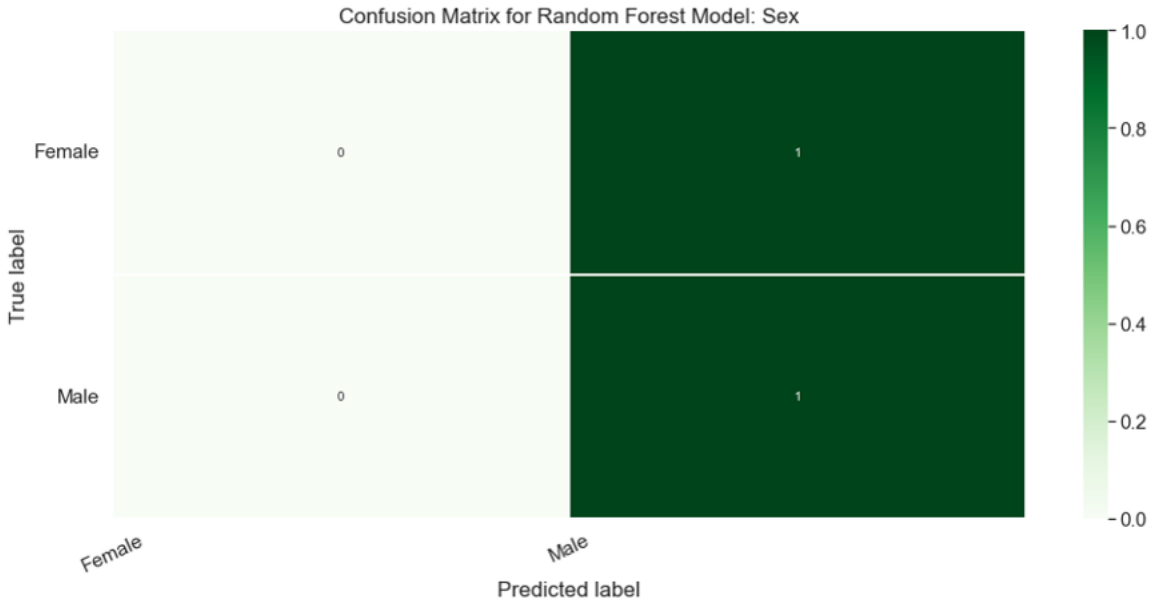

Figure 2: Confusion Matric - Sex

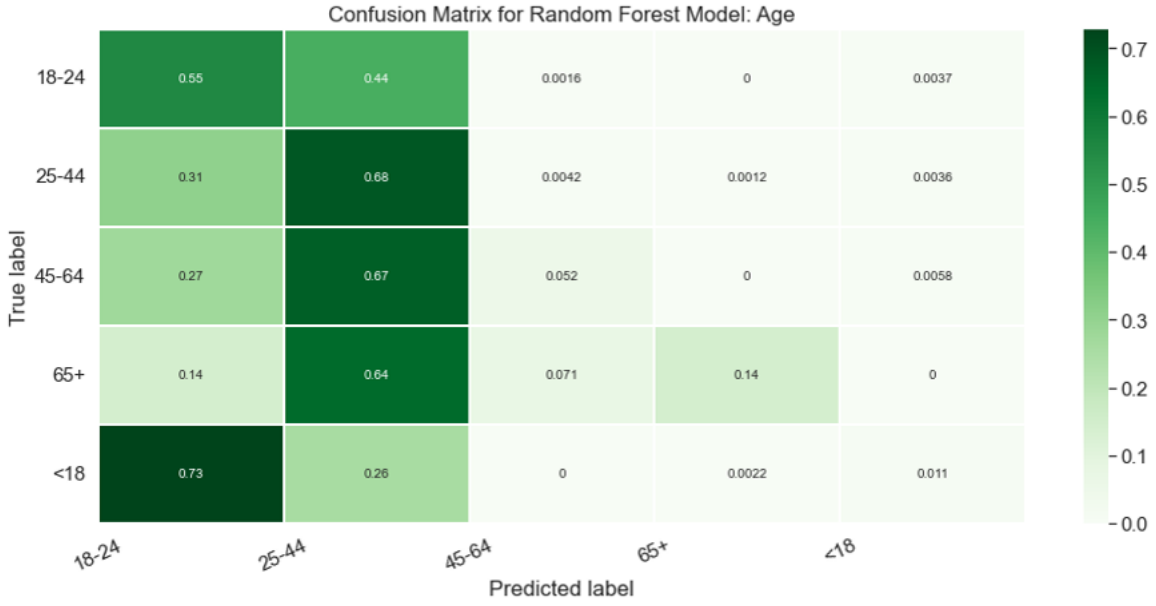

Figure 3: Confusion Matrix - Age

Looking at the confusion matrices, we can see that our predictions are very skewed. In the race predictor, the model predicts "black" most of the time. It predicts "male" 100% of the time in the sex category and primarily predicts age groups "18-24" and "25-44". These high percentage values and dark green blocks on the confusion matrix are caused by the uneven distribution of demographics in our historical shooting data set. Of Perpetrators with known race data, ~73% are black causing our model to have a bias towards the "black" race classifier. Additionally, ~97% are

male meaning if we classified the perpetrator as a male 100% of the time, we would have ~97% accuracy which is how the model behaved.

Despite the bias in the data, our model has a fairly high accuracy rate. Of the training data, ~85% had 2 or more correctly classified labels. Because of the skewness of the data and the racial and sex bias in the model, we would not recommend this model or its predictions as a source of truth. The random tree classifier model could be used to help form an original profile of an unknown suspect with the caveat that there might be bias in the predictions. This model could improve in its accuracy with the addition of more historical data. As new data is collected the model can be retrained and retested in Python to create continuous improvement on the prediction metrics.

Exploring Trends and Patterns in New York City Shootings:

For data exploration, a couple of questions which we tried to analyze the data for; Are there any yearly trends in the number of shootings? Are there identifiable patterns in the escalation or decline of shootings? To answer these questions, one of the analyses done was visualizing the number of shootings per year. This will provide the authorities with an understanding of how well their previous techniques have worked for bringing down the shootings around New York City. Next, the yearly shooting data was analyzed by borough. This visual will provide the authorities with an understanding of patterns of each borough so they can decide if there are any specific hotspots or boroughs that need more attention. Monthly shooting data was also analyzed by creating a line graph of number of shootings over the years broken out by months. This analysis will provide the authorities with an understanding of patterns of the number of shootings over seasons.

Forecasting Shootings in NYC Boroughs:

To forecast shootings in NYC boroughs, we used an exponential smoothing forecast model with a seasonal component. Breaking down the count of monthly incidents per borough, the model uses past observations in a time series to make predictions on future incident numbers. This algorithm is implemented in Python's Statsmodels package using the exponential smoothing function. The code size is small but can be complex for new users and implemented using open-source knowledge. Forecasting future incidents per borough could allow the NYPD to prioritize resources in certain areas and during certain months.

A time series was trained using exponential smoothing with multiplicative seasonality for each borough except Staten Island where the number of shootings was lower and additive seasonality was used. The following table shows the time series model evaluations for each borough including root mean squared error (RMSE), mean forecast error (MFE), and the size of the 95% confidence interval for predictions (95% CI band). These results were calculated with a train set of data from 2006-2020 and test data from 2021-2022.

Borough	RMSE	MFE	95% CI band
Queens	7.290	23.785	14.288
Manhattan	7.941	22.345	15.566
Staten Island	2.166	3.543	4.245
Brooklyn	32.253	83.118	63.216
Bronx	16.507	45.615	32.355

Table 6: Forecast Metrics

Analyzing the above metrics, we have a low RMSE value for Queens, Manhattan, and Staten Island. The RMSE measures the average difference between predicted and actual values. RMSE values for Brooklyn and the Brox are much higher with RMSE values of 32.253 and 16.507. Similarly, the MFE, which measures the accuracy of a forecast prediction, is higher for Brooklyn and the Bronx. These higher RMSE and MFE values could be a result of the increased number of shootings from 2020-2021 compared to the previous years' data, before COVID-19, which showed a decrease in shootings.

Using the forecasting models created, we forecasted the next 12 months of shootings per borough. These forecasts are shown in the visualizations section of the report. We suggest updating the forecast models yearly to capture new trends in the data and increase accuracy in predicting future yearly data.

Seasonal Hot Spot Analysis (SHSA):

The seasonal hot spot analysis was implemented using heat maps of the location of shooting incidents. Using map plotting libraries in Python, we can see where higher concentrations of shootings occur in the NYC boroughs. The SHSA will map and depict the geographic distribution of gunshot incidents. We can examine how these hot spots' locations vary with the seasons to look for any relationships with local events, holidays, or weather variations. Additionally, we can examine the possible influence of socioeconomic variables on the seasonal fluctuation in gunshot events.

With an emphasis on seasonal and geographic variations, this systematic method describes how to use data analytics to investigate and comprehend the dynamics of shooting episodes in New York City. It depicts areas of high crime allowing the NYPD to allocate resources effectively to hot spot areas shown in this analysis and its visualizations.

Fatality Analysis:

The historical shooting incidents have an attribute indicating whether the incident resulted in the death of the victim. In this analysis, we looked at the proportion of shootings that resulted in deaths.

We found that approximately 19% or nearly 1 in 5 of the shooting incidents in the dataset were fatal. It is calculated by dividing the number of fatal incidents (5,266) by the total number of incidents (27,312). In summary, out of the 27,312 shooting incidents analyzed, 5,266 incidents resulted in fatalities. This proportion provides insight into the lethality of shooting incidents, with a significant portion leading to loss of life.

We filtered fatal incidents, and then visualized the distribution of these incidents across different categories such as boroughs, victim age groups, and victim races. The algorithm application is to understand patterns and distributions.

The algorithms for data analysis (Pandas) and visualization (Matplotlib) are already implemented and available as open-source libraries in Python. Algorithms are implemented using Python and the Code size estimate is Medium, Code complexity is Medium. In a team setting, testing responsibilities are shared among our team members

Visualizations

Demographic Analysis Visualizations:

Age Distribution of Perpetrators and Victims:

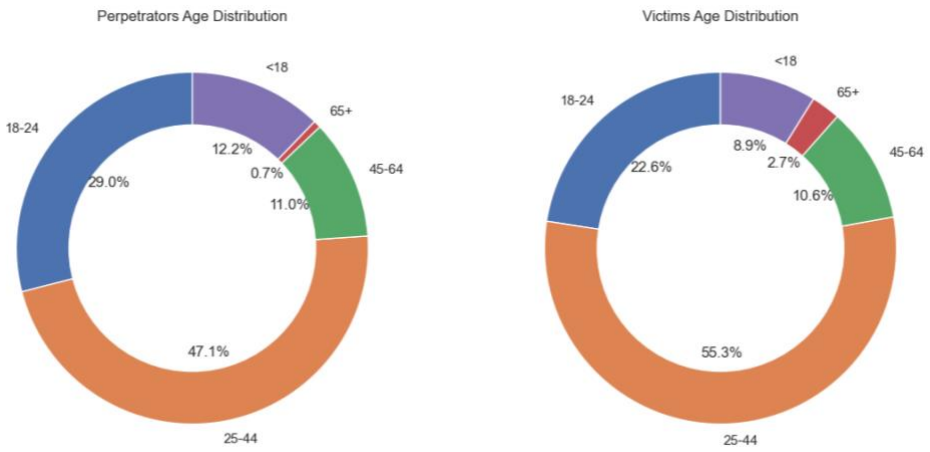

Figure 4: Age Distribution Chart

The above graph is a donut chart showing the age distribution for perpetrators and victims, where a similar trend between the two is more likely; 25-44 years of age account for the largest at 47.1% and 55.3%. Especially, it should be noted that one of the most vulnerable groups is the youth part, aged 18-24 years: 29% as for the offenders and 22.6% as for the victims of sexual violence. On the other hand, the lowest percentages are observed at the age category of 65+ of both charts, which is an indication that the elderly have a lower involvement ratio in the incidents. The implication for social services and crime prevention programs is that this shows a trend for the most active age group in society to be the one most commonly linked with involvements in incidents, either as perpetrator or victim.

Sex Distribution for Perpetrators and Victims:

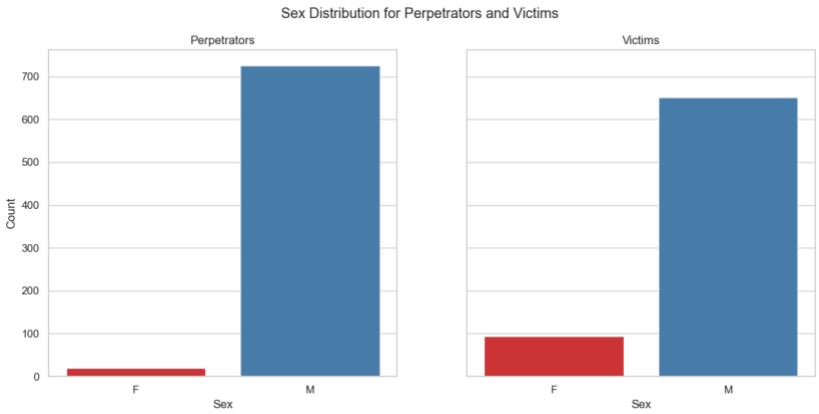

Figure 5: Sex Distribution Chart

From the bar charts, it is evident that there is a definite gender difference in the distribution of the two in question: perpetrators and victims. Males are way above females in the two categories, with the number of male perpetrators being far above females. The number of victims was correspondingly larger for males; the difference vis-à-vis the number of female victims much smaller compared to the distribution among the perpetrators. This points out to the fact that it is men who are involved most of the time as both the victims and the perpetrators in a scenario, thus bringing to light some of the social or behavioral factors that may bring about this skewered presentation of the genders. Data can be something useful, since information may help in understanding gender-specific trends and therefore guide intervention strategies.

Top 5 Race Distribution for Perpetrators and Victims:

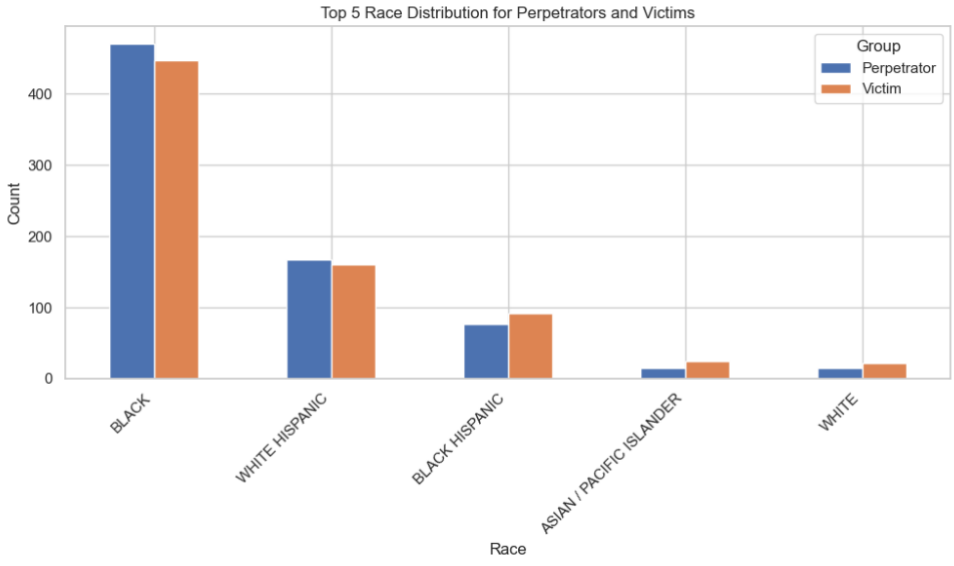

Figure 6: Race Distribution Chart

The bar chart summarizes the racial distribution of perpetrators and victims of the top five categories. It shows that the number of people is most numerous among both sets of people and

would probably mean that they are overrepresented more often than another racial group in both sets of people contained within this data set. Next is white Hispanic, black Hispanic, and black, all of which still have high counts of victims and tend to be greater in counts than the perpetrators. Notably, the difference is stark when it comes to the White count of people. Even the count of perpetrators comes down very sharply compared with other races; while in the numbers of victims, it is still fewer. Asian/Pacific Islanders have the least representation in both categories. The data may reflect socio-economic, demographic, or systemic factors influencing the representation of different races in these roles and can be pivotal in addressing racial disparities in social and criminal justice policies.

Trends and Patterns in New York City Shootings:

The following visuals were done to analyze the trends of historical shootings in NYC. As discussed in the analytics section the bar plot below shows the total number of shootings per year in NYC.

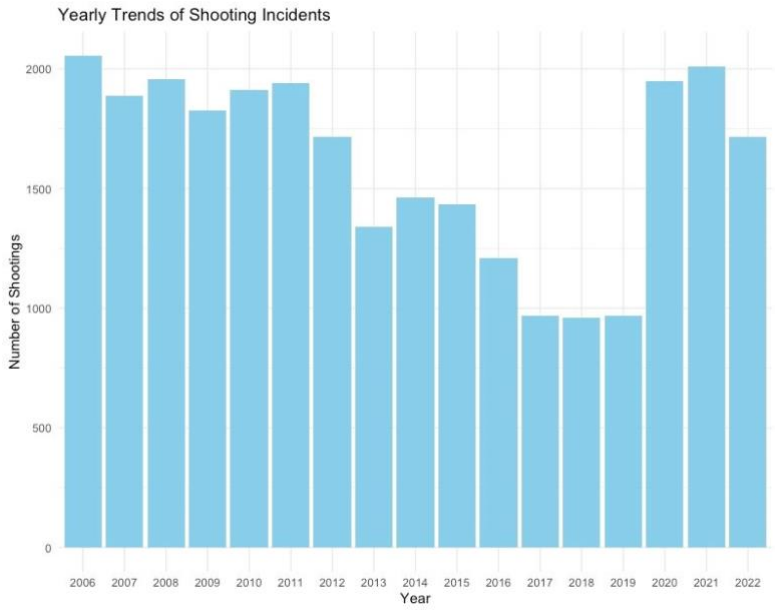

Figure 7: Yearly Shooting Trends Chart

This visual shows a continues declined in overall shooting from 2006 to 2019. Shooting rates increase again in 2020 back to shooting rates seen in 2006 but appear to be declining again. This increase could be related to the COVID-19 pandemic. The next visual discussed in the analytics of NYC shooting trends was yearly shootings by borough, shown below.

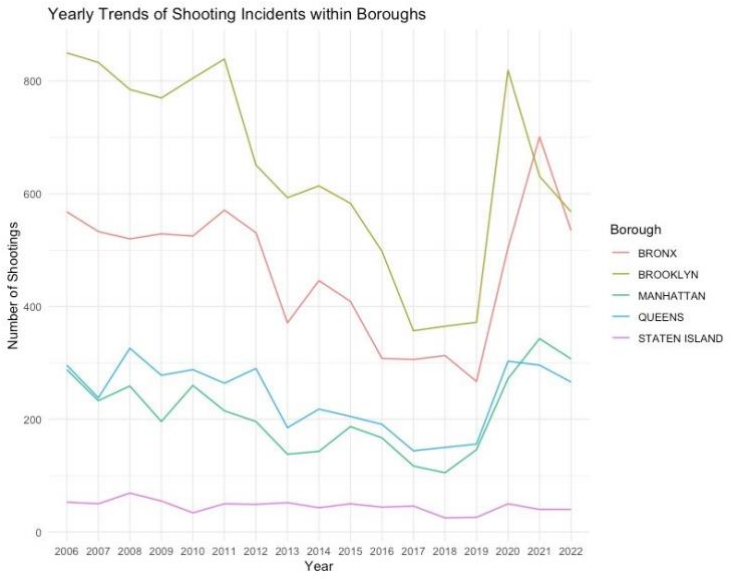

Figure 8: Yearly Shooting Trends by Borough Chart

The data showed a similar trend to the previous bar plot. The numbers of shooting were on the decline from 2006 to 2019 but jumped up in 2020 and continue to decline again. While most boroughs follow this trend, Staten Island has appeared to have a constant shooting rate from 2006-2022 and has the lowest number of yearly shootings. Other Boroughs could learn from Staten Island about how they are keeping their shooting low. This graph also shows that the Bronx and Brooklyn are the boroughs with the highest yearly shooting incidents suggesting more resources be allocated to these areas to combat crime effectively.

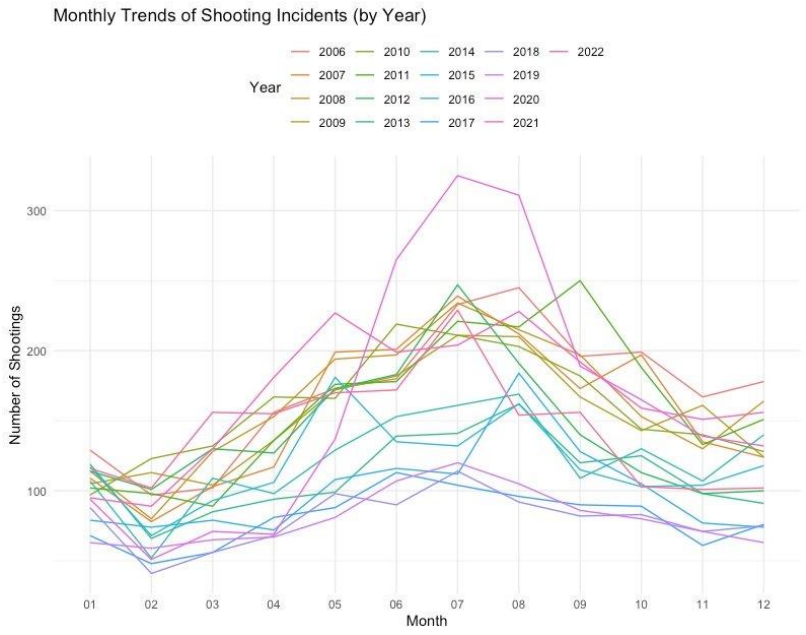

Figure 9: Yearly Monthly Trends Chart

The final trend visualization shown above displays an increase in shootings during the warmer months compared to colder months. Around June/July the numbers of shooting are the highest for almost every year but are lower towards the start and end of the year. One of the major differences between these months is the weather, which could be affecting these numbers.

Forecasting Shootings in NYC Boroughs:

The visualizations in this section cover the previous work discussed in the forecasting algorithms. Using the algorithms discussed, a forecast for the next 12 months was created for each NYC borough shown below.

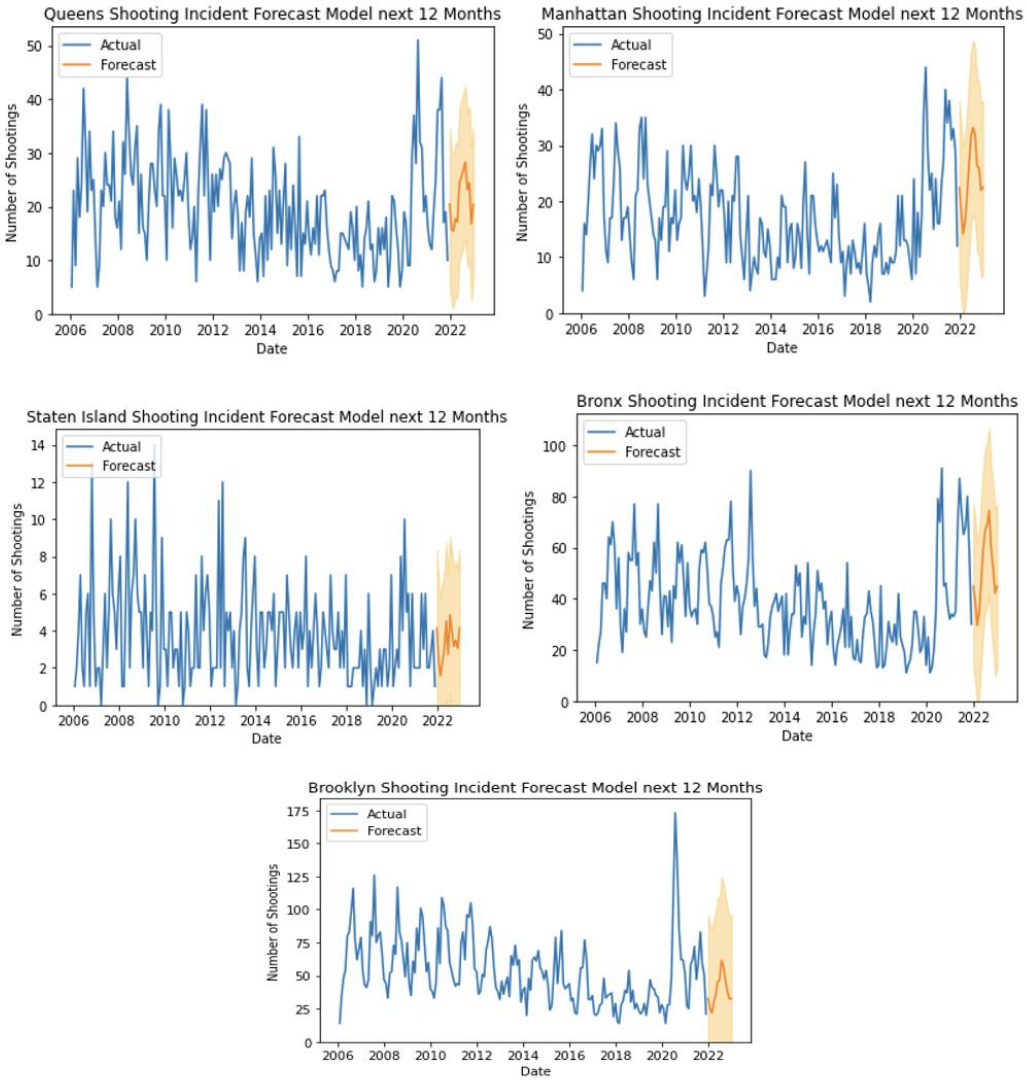

Figure 10: Borough Forecast Charts

Seasonal Hot Spot Analysis:

This section reviews the overall hot spots of shootings from 2006-2022 and the seasonal trends through heat maps.

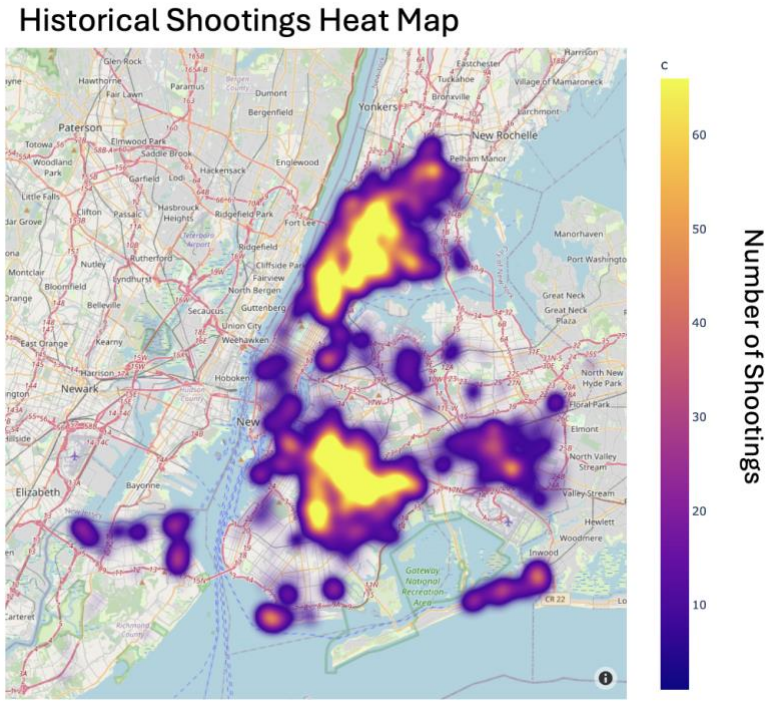

Figure 11: Historical Shooting Heat Map

The first graphic shows the hot spots from 2006-2022. As we can see from the graphic, The Bronx and Brooklyn appear to have major hot spots indicated by the light-yellow color. This is consistent with our previous analysis visuals of yearly trends per borough.

Next, we analyzed the seasonal trends of hot spots, spring, summer, winter, and fall.

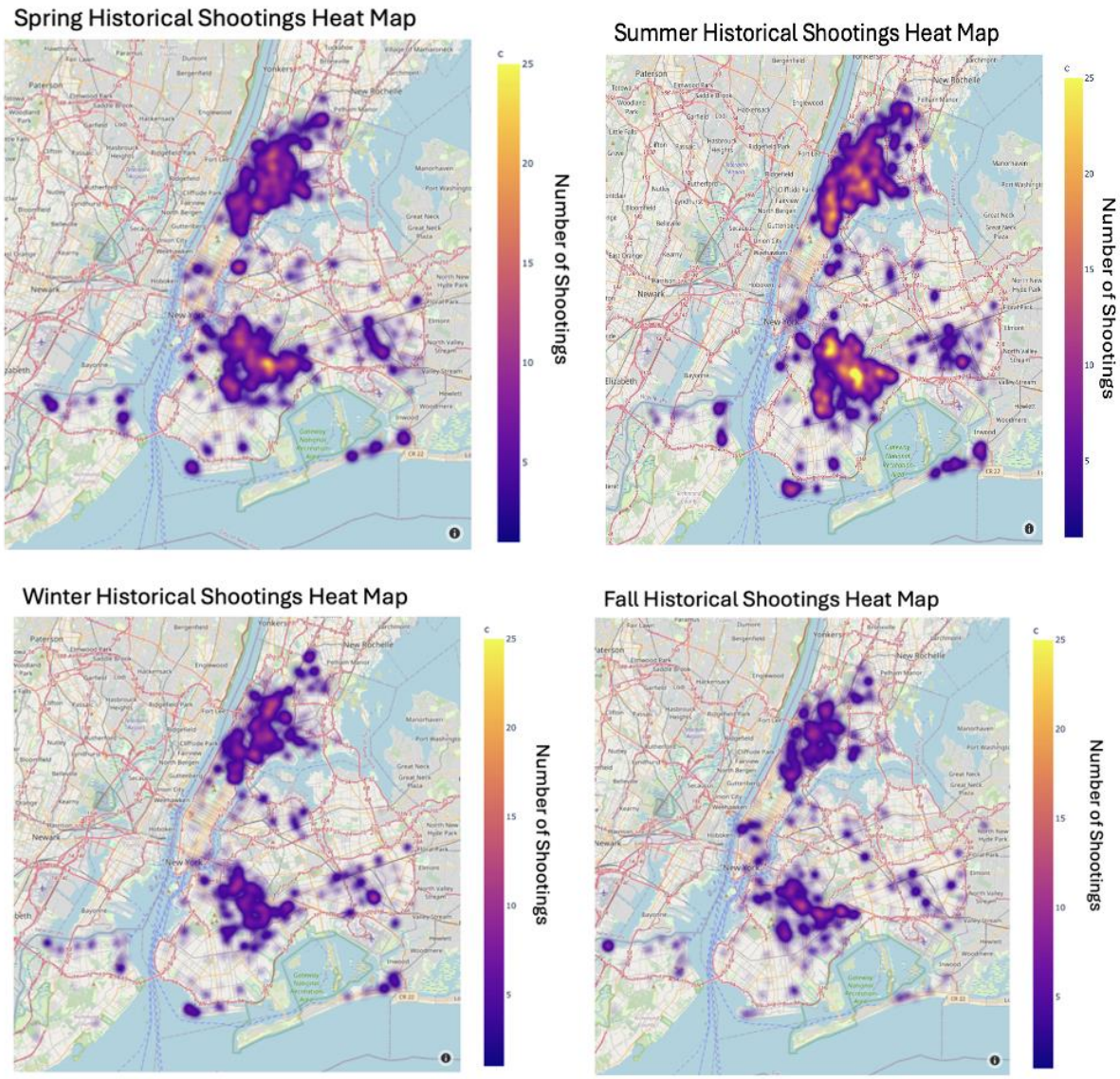

Figure 12: Seasonal Shooting Heat Maps

Similar to the previous graphic, The Bronx and Brooklyn appear to have the more shootings compared to the other three boroughs despite the season. As discussed in the monthly shooting trends we do see a change in the number of shooting incidents from season to season. Summer appears to have the most shooting incidents, depicted in the upper right graphic, with the most light-yellow hot spots. Summer shooting incidents are followed by the spring with the second most number of shooting incidents. Winter and fall have the least number of shooting incidents, fall less than winter. This finding could be related to weather but would require more analysis to uncover why these trends are occurring. These heat maps also indicate areas in other boroughs were shooting incidents occur most often (i.e. northern Staten Island, northern Manhattan/southern Bronx, northern Brooklyn, and southeastern Queens).

Fatality Visualizations:

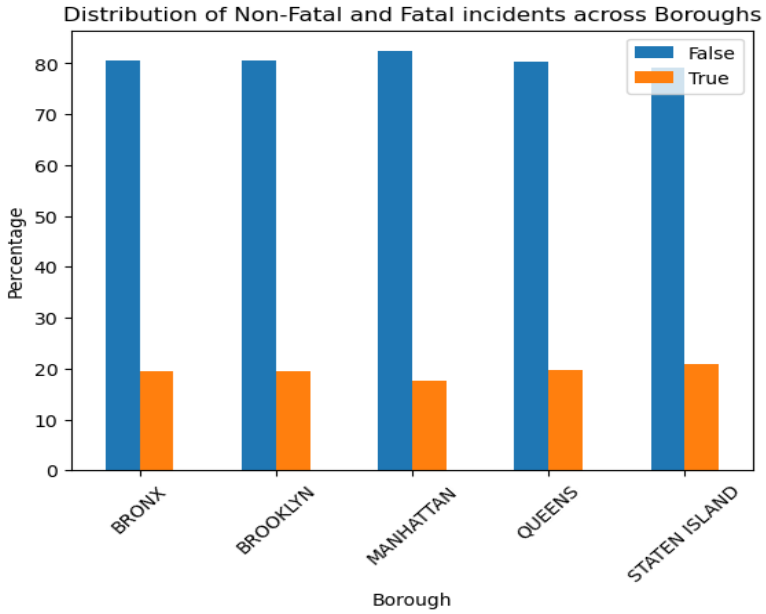

Figure 13: Borough Fatality Chart

The graph shows the distribution of non-fatal and fatal incidents across boroughs. It appears that non-fatal incidents are consistently higher than fatal incidents across all boroughs. This graph also shows that the distribution of fatal incidents to non-fatal incidents is similar in each borough as the fatalities are all around 20% per borough.

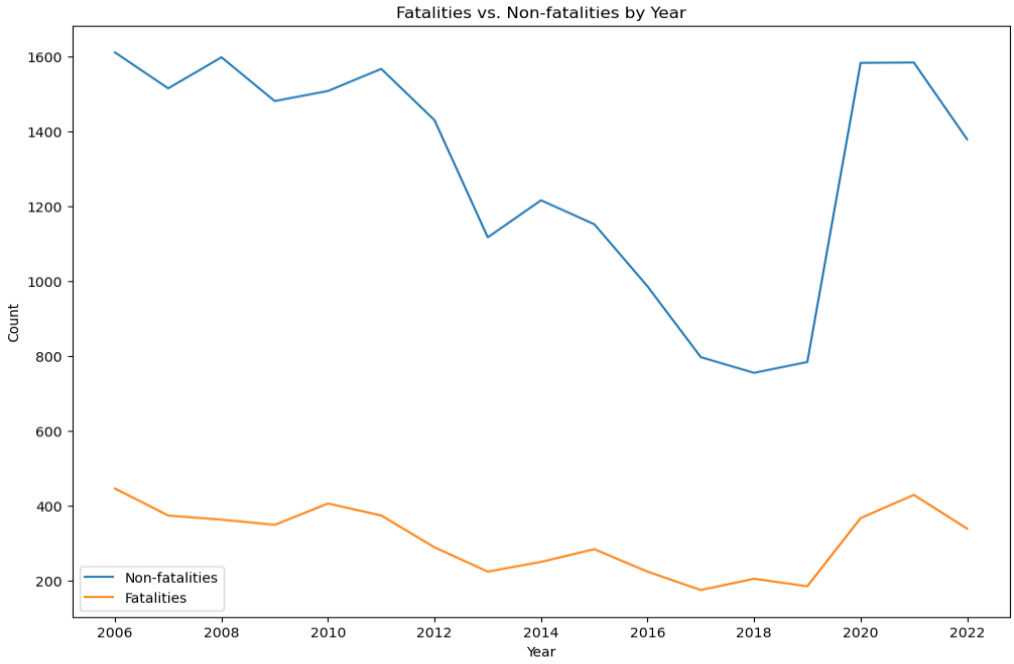

Figure 14: Yearly Fatality Trends Chart

From the graph, we can see the number of fatalities is consistently higher than the number of non-fatalities throughout the years shown. Overall, the number of incidents (both fatal and non-fatal) seems to have fluctuated somewhat over the years, with a possible peak around 2010 and a downward trend from 2018 to 2022.

Findings

Throughout this project, multiple algorithms, analytics, and visualizations were used to uncover information about historical shooting incidents in NYC. We first analyzed the demographic distribution of perpetrators and victims in shooting incidents. This analysis found that a majority of both perpetrators and victims identified as black or African American (>70%) followed by White Hispanics, Black Hispanics, Asian/Pacific Islander, and White. Most of the shootings occurred between victims and perpetrators ages 18-44 and almost all shootings involved male victims or perpetrators. Using the demographic information, we created a random forest algorithm to predict perpetrator demographics based on victim demographics and borough. This model was trained using records with complete victim and perpetrator demographics predicting correct age 52.25% of the time, sex 97.47% of the time, and race 73.36% of the time. Historical trends in shooting incidents were studied which allowed us to create forecasts predicting the number of shootings in each borough for the next 12 months. Historical trends showed decreasing incident numbers from 2006-2019 which then increased in 2020, we assume this is linked to the start of the COVID-19 pandemic. Trends also showed higher shooting numbers during certain seasons and in certain boroughs. These trends were also shown though the use of heat maps. From the hot spot and trend analysis, we found higher shooting numbers in the spring and summer months concentrated in the Bronx and Brooklyn. Finaly, we analyzed the shooting incident fatality rates. The shootings resulting in fatalities did not vary proportionally from borough to borough or from year to year, about 20% of shooting incidents resulted in fatalities in each borough and about 20% of yearly shootings were recorded as fatal with the lowest rates (<17%) in 2012 and 2013.

Summary

This analysis revealed insights to the trends and common factors of shooting related incidents in the boroughs of New York City. Using algorithms and visualization methods in R and Python, the team worked to add value to the recorded data for the NYPD. Analyzing the demographics of perpetrators and victims, we created a model to predict perpetrator demographics. While the demographics of both perpetrators and victims was highly skewed, the model predicted all three demographics, age, sex, and race, correctly 38.64% of the time and at least one demographic 99% of the time. Because of the non-uniform distribution of perpetrator and victim demographics we recommend this model be used only as a preliminary profile not the ultimate source if truth. Through analyzing trends and hot spots, we discovered seasons with higher shooting incidents, spring and summer, as well as boroughs/areas in boroughs with higher shooting rates, The Bronx and Brooklyn. Insight on seasonal trends and hot spots in boroughs could allow for more patrols in certain areas at certain times of the year and increased intervention in specific borough neighborhoods. Identifying hot spots, trends, predicting shooter demographics, and forecasting shooting incidents gives the NYPD analytical insight to

increasing effectiveness of current policing methods as well as lay the groundwork for policy intervention. While analyzing historical shooting records gave critical insight to the landscape of gun violence in NYC, more research behind the causes of trends and patterns is needed to effectively enact policy and change.

Future Work

There are multiple directions this project can be driven toward with future work. One proposal of future step would be to start working and integrating with law enforcement systems. There is a need to find an opportunity to share the insights found from this project into existing law enforcement systems and workflows. This can lead to providing law enforcement agencies with user friendly dashboards, alerts, and reports which can support data driven decision making and could lead to proactive crime prevention in future.

Another proposal of future step would be to conduct in-depth geospatial analysis to identify environmental and spatial factors which may be associated with crime hotspots. If we can explore the relationship between urban infrastructure, socioeconomic indicators, and crime rates, that can allow urban planning and development boards/teams to aim to reduce crime.

Other future work with this project could be to perform community based participatory research. Local communities know best what is happening in their area and what may be helpful to bring the change wanted. We could engage local communities in brainstorming for crime prevention strategies. Collaboration between researchers, law enforcement agencies, policymakers, and community members to develop community driven interventions.

We can create technologies that enable law enforcement to monitor crime patterns in real time, we can help prevent shootings before they occur. To assist develop safer cities, we may also investigate how neighborhood and urban area architecture influences crime. Finally, collaborating directly with communities to develop shooting prevention techniques can result in tactics that are more successful and favorably received by people who will be most impacted.

Appendix

Appendix A: Code References

One Drive Code Folder

Appendix B: Risks

- Time Management: Risk of timely project delivery and project completion. This risk has a low to medium chance of occurring due to the speed of the course and will have a high impact on project success. The team will mitigate this risk by having weekly meetings and plan and track project deliverables in YouTrack. At this time, the status of this risk is low.
- Project value add: Risk of project not adding new value or insights to stakeholders. This is a medium risk. If this risk occurs, there will be little impact as we are still learning about the landscape of shootings in New York City. To mitigate this risk, we will review analyses and project goals as we proceed to ensure we are adding value. The risk is currently low.
- Data Quality: The dataset may contain inconsistencies, errors, or missing values which could impact the accuracy and reliability of analytical results. It has a medium probability of occurring. Poor data quality will have a high impact on analysis quality. It can lead to flawed conclusions and ineffective decision-making. To mitigate this risk, we will perform preprocessing steps addressing data cleaning, missing values, outliers or any other inconsistencies observed in the data. The status of this risk is currently low.
- Data Bias: The project could lead to a biased outcome giving a skewed view of crime in New York City based on certain demographics. The probability of this occurring is medium with a high impact on the project if it does occur. To mitigate this risk, we will ensure the entire scope is being studied and all the available data is being used to create a complete story. The risk is low and being monitored as we complete the project.
- Fatality analysis risk: Risks associated with the algorithms are the common risks in data analysis projects include data quality issues, incorrect assumptions, and misinterpretation of results. Risks associated with data analysis projects include incorrect data interpretation, biased analysis, and scalability issues with larger datasets. The Probability of Occurring is Low. The impact of these risks can range from minor inaccuracies in analysis to significant misinterpretation of results, leading to incorrect decisions or conclusions. Mitigation strategies include thorough data validation and cleaning, sensitivity analysis to test the robustness of results, and involving domain experts to validate assumptions and interpretations. The risk status would depend on the specific context of the project, including the data quality, team expertise, and complexity of the analysis. Regular risk assessments and mitigation efforts can help manage and monitor the risk status throughout the project lifecycle.

- Hot Spot analysis: Inaccurate hot spot identification may result in ineffective community safety and police efforts, which could expose high-incident regions to unnecessary hazards. Law enforcement agencies' operational preparedness and strategic planning may be impacted by incorrect interpretations of seasonal patterns. Mitigation strategy is to guarantee the accuracy and dependability of the data utilized for analysis, implement strong data cleaning and validation procedures. To validate results, compare results from various clustering approaches and parameter settings using a multi-model approach. Collaborate with neighborhood associations and law enforcement to confirm hotspots and seasonal patterns. In order to preserve privacy and adhere to legal requirements, make sure that extensive data protection mechanisms are in place. Law enforcement agencies' operational preparedness and strategic planning may be impacted by incorrect interpretations of seasonal trends.
- Updating data to keep analysis current: In order to adjust to new data and developing patterns, active monitoring and ongoing improvement of analysis techniques will be necessary. Working together with data scientists, law enforcement professionals, and community leaders will be essential to successfully controlling project risks.
- Bias in perpetrator predictions: Because of the distribution of current data in perpetrator demographics, the model tends to predict one race or sex more often than others. If the NYPD uses this model, we need to ensure the model is only used for a predictive profile of a perpetrator not the ultimate source of truth.

Appendix C: Agile Development

Agile Development was conducted using the YouTrack platform. Sprint materials were updated in a shared one drive folder containing notes, code, project paper documents, and project presentation documents.

- Sprint 1:
  - Project definition
- Sprint 2:
  - Collaborate on ideas for analytics
  - Identify risks and mitigation plans
  - Finalize data set
- Sprint 3:
  - Begin initial data processing
  - Define research questions
  - Create algorithms
  - Compile work
- Sprint 4:

    - Create visualizations based on analytics
    - Add visuals to project document and presentation
    - Finalize draft summary
- Sprint 5:
    - Finalize visualizations
    - Findings
    - Summary
    - Future work
    - Turn in final project presentation and project paper

Appendix D: References

[1] "About New York City Government," *City of New York*. https://www.nyc.gov/nyc-resources/about-the-city-of-new-york.page

[2] "About NYPD - NYPD." https://www.nyc.gov/site/nypd/about/about-nypd/about-nypd-landing.page

[3] "NYPD announces January 2024 citywide crime statistics," *The Official Website of the City of New York*, Feb. 06, 2024. https://www.nyc.gov/site/nypd/news/p00099/nypd-january-2024-citywide-crime-statistics

[4] J. Flanagan, "NYC crime stats 2023: Are crime rates up or down in the city?," *FOX 5 New York*, Jan. 03, 2024. [Online]. Available: https://www.fox5ny.com/news/nyc-crime-rate-2023-statistics

[5] J. M. MacDonald, J. A. Fagan, and A. Geller, "The effects of local police surges on crime and arrests in New York City," *PLOS ONE*, vol. 11, no. 6, p. e0157223, Jun. 2016, doi: 10.1371/journal.pone.0157223.

[6] M. S. Barton, "Gentrification and violent crime in New York City," *Crime & Delinquency*, vol. 62, no. 9, pp. 1180–1202, Jul. 2016, doi: 10.1177/0011128714549652.

[7] Manhattan Institute, "Crime Hot Spots: A study of New York City streets in 2010, 2015, and 2020," *Manhattan Institute*, Mar. 03, 2023. https://manhattan.institute/article/crime-hot-spots-a-study-of-new-york-city-streets-in-2010-2015-and-2020

[8] NYC Open Data. (2023). *NYPD Shooting Incident Data (Historic)* [Dataset]. New York City Police Department. https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/about_data