NYPD Shooting Incident Data (Historic) Analysis

M. Trammell

2024-07-05

Introduction

This document provides an analysis of the NYPD Shooting Incident Data (Historic) obtained from the NYC Open Data portal.

Loading the Data

First, load the necessary libraries, including readr, dplyr, and tidyverse. Next, read the data from the provided URL.

Data URL -> https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Data Overview

Here are the first few rows of the data followed by the column names.

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

```
##  [1] "INCIDENT_KEY"           "OCCUR_DATE"
##  [3] "OCCUR_TIME"             "BORO"
##  [5] "LOC_OF_OCCUR_DESC"      "PRECINCT"
##  [7] "JURISDICTION_CODE"      "LOC_CLASSFCTN_DESC"
##  [9] "LOCATION_DESC"          "STATISTICAL_MURDER_FLAG"
## [11] "PERP_AGE_GROUP"         "PERP_SEX"
## [13] "PERP_RACE"              "VIC_AGE_GROUP"
## [15] "VIC_SEX"                "VIC_RACE"
```

```
## [17] "X_COORD_CD"              "Y_COORD_CD"
## [19] "Latitude"                "Longitude"
## [21] "Lon_Lat"
```

Data Summary

Here the data is summarized to understand its structure and content.

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms         Class :character
##  Median : 92711254   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE            X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
```

```
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Data Cleaning

This step cleans the data and handles any missing values.

```
##            INCIDENT_KEY             OCCUR_DATE             OCCUR_TIME
##                       0                      0                      0
##                    BORO      LOC_OF_OCCUR_DESC               PRECINCT
##                       0                  25596                      0
##       JURISDICTION_CODE     LOC_CLASSFCTN_DESC          LOCATION_DESC
##                       2                  25596                  14977
## STATISTICAL_MURDER_FLAG         PERP_AGE_GROUP               PERP_SEX
##                       0                   9344                   9310
##               PERP_RACE          VIC_AGE_GROUP                VIC_SEX
##                    9310                      0                      0
##                VIC_RACE             X_COORD_CD             Y_COORD_CD
##                       0                      0                      0
##                Latitude              Longitude                Lon_Lat
##                      59                     59                     59
```

```
## tibble [10,209 x 16] (S3: tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY           : num [1:10209] 2.45e+08 2.48e+08 3.34e+07 2.55e+08 1.11e+07 ...
##  $ OCCUR_DATE             : Date[1:10209], format: "2022-05-05" "2022-07-04" ...
##  $ OCCUR_TIME             : POSIXct[1:10209], format: "1970-01-01 00:10:00" "1970-01-01 22:20:00" ..
##  $ BORO                   : Factor w/ 5 levels "BRONX","BROOKLYN",..: 3 1 4 1 4 2 1 4 2 2 ...
##  $ PRECINCT               : Factor w/ 77 levels "1","5","6","7",..: 8 31 71 29 63 40 25 59 47 35 ...
##  $ JURISDICTION_CODE      : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 3 3 3 3 ...
##  $ LOCATION_DESC          : Factor w/ 40 levels "(null)","ATM",..: 40 1 4 1 29 27 26 26 26 26 ...
##  $ STATISTICAL_MURDER_FLAG: Factor w/ 2 levels "FALSE","TRUE": 2 2 1 2 1 1 1 2 1 1 ...
##  $ PERP_AGE_GROUP         : Factor w/ 11 levels "(null)","<18",..: 7 1 11 5 8 7 7 1 5 5 ...
##  $ PERP_SEX               : Factor w/ 4 levels "(null)","F","M",..: 3 1 3 3 3 3 3 1 3 3 ...
##  $ PERP_RACE              : Factor w/ 8 levels "(null)","AMERICAN INDIAN/ALASKAN NATIVE",..: 4 1 4 4
##  $ VIC_AGE_GROUP          : Factor w/ 7 levels "<18","1022","18-24",..: 4 3 4 1 5 4 3 3 1 3 ...
##  $ VIC_SEX                : Factor w/ 3 levels "F","M","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ VIC_RACE               : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 3 3 3 3 3 3 7 3 3
##  $ X_COORD_CD             : num [1:10209] 986050 1016802 1046405 1011263 1055659 ...
##  $ Y_COORD_CD             : num [1:10209] 214231 250581 187113 251671 201767 ...
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :1970-01-01 00:00:00.0000
##  1st Qu.: 55780317   1st Qu.:2009-01-18   1st Qu.:1970-01-01 03:58:00.0000
##  Median : 92152095   Median :2013-08-15   Median :1970-01-01 15:03:00.0000
##  Mean   :138290758   Mean   :2015-01-07   Mean   :1970-01-01 12:54:34.0346
##  3rd Qu.:243433247   3rd Qu.:2022-04-10   3rd Qu.:1970-01-01 20:23:00.0000
##  Max.   :279758069   Max.   :2023-12-29   Max.   :1970-01-01 23:59:00.0000
##
##             BORO         PRECINCT    JURISDICTION_CODE
##  BRONX        :2993   75     : 571   0:7339
##  BROOKLYN     :3821   73     : 516   1:  22
```

```
##  MANHATTAN    :1559   40     : 403   2:2848
##  QUEENS       :1498   47     : 373
##  STATEN ISLAND: 338   79     : 373
##                       42     : 355
##                       (Other):7618
##                    LOCATION_DESC  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  MULTI DWELL - PUBLIC HOUS:3256   FALSE:7963              25-44  :3227
##  MULTI DWELL - APT BUILD  :2319   TRUE :2246              18-24  :3182
##  (null)                   :1711                           UNKNOWN:1368
##  PVT HOUSE                : 691                           (null) :1141
##  GROCERY/BODEGA           : 576                           <18    : 819
##  BAR/NIGHT CLUB           : 532                           45-64  : 421
##  (Other)                  :1124                           (Other):  51
##    PERP_SEX             PERP_RACE    VIC_AGE_GROUP  VIC_SEX
##  (null):1141   BLACK          :6173   <18    :1041   F:1276
##  F     : 267   WHITE HISPANIC:1207   1022   :   1   M:8930
##  M     :8210   (null)         :1141   18-24  :3430   U:   3
##  U     : 591   UNKNOWN        : 740   25-44  :4764
##                BLACK HISPANIC: 682   45-64  : 842
##                WHITE          : 173   65+    : 100
##                (Other)        :  93   UNKNOWN:  31
##                             VIC_RACE      X_COORD_CD        Y_COORD_CD
##  AMERICAN INDIAN/ALASKAN NATIVE:   6   Min.   : 925480   Min.   :127539
##  ASIAN / PACIFIC ISLANDER      : 207   1st Qu.: 999484   1st Qu.:183445
##  BLACK                         :6954   Median :1007217   Median :198452
##  BLACK HISPANIC                :1012   Mean   :1008188   Mean   :209130
##  UNKNOWN                       :  23   3rd Qu.:1016159   3rd Qu.:239361
##  WHITE                         : 315   Max.   :1065474   Max.   :271128
##  WHITE HISPANIC                :1692
```

Handling Missing Data

After cleaning the data, we re-examine the missing values and ensure that there are no missing values in crucial columns.

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO                PRECINCT       JURISDICTION_CODE
##                       0                       0                       0
##           LOCATION_DESC STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                       0                       0                       0
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                       0                       0                       0
##                 VIC_SEX               VIC_RACE              X_COORD_CD
##                       0                       0                       0
##              Y_COORD_CD
##                       0
```

Explanation of Handling Missing Values

In this analysis, the missing values were handled as follows:

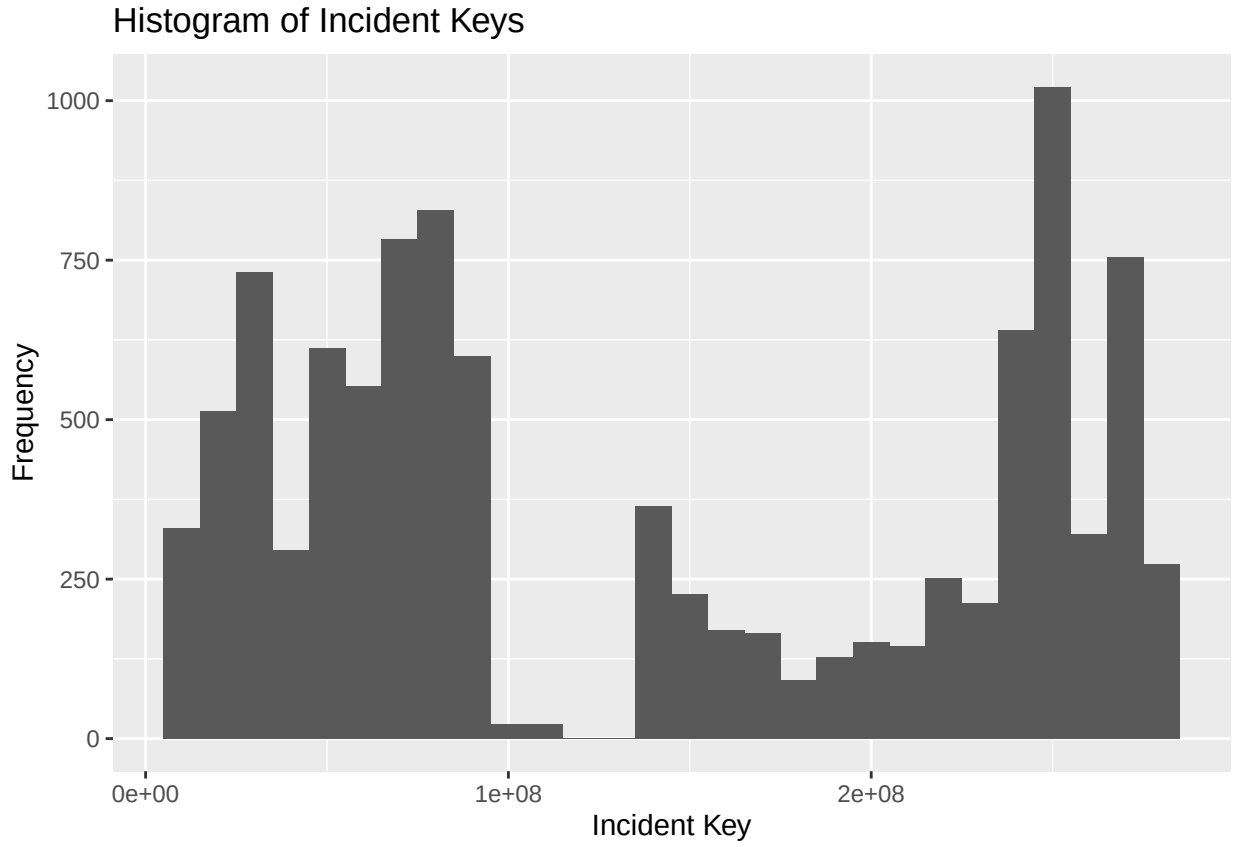
1. Removal of Columns with High Percentage of Missing Values: Columns such as Latitude, Longitude, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, and Lon_Lat were removed due to a high percentage of missing values. These columns were deemed non-essential for the analysis.

2. Removal of Rows with Missing Values in Crucial Columns: Rows with missing values in essential columns such as JURISDICTION_CODE, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, and PERP_RACE were removed. This ensures that the remaining dataset is complete and reliable for further analysis and modeling.

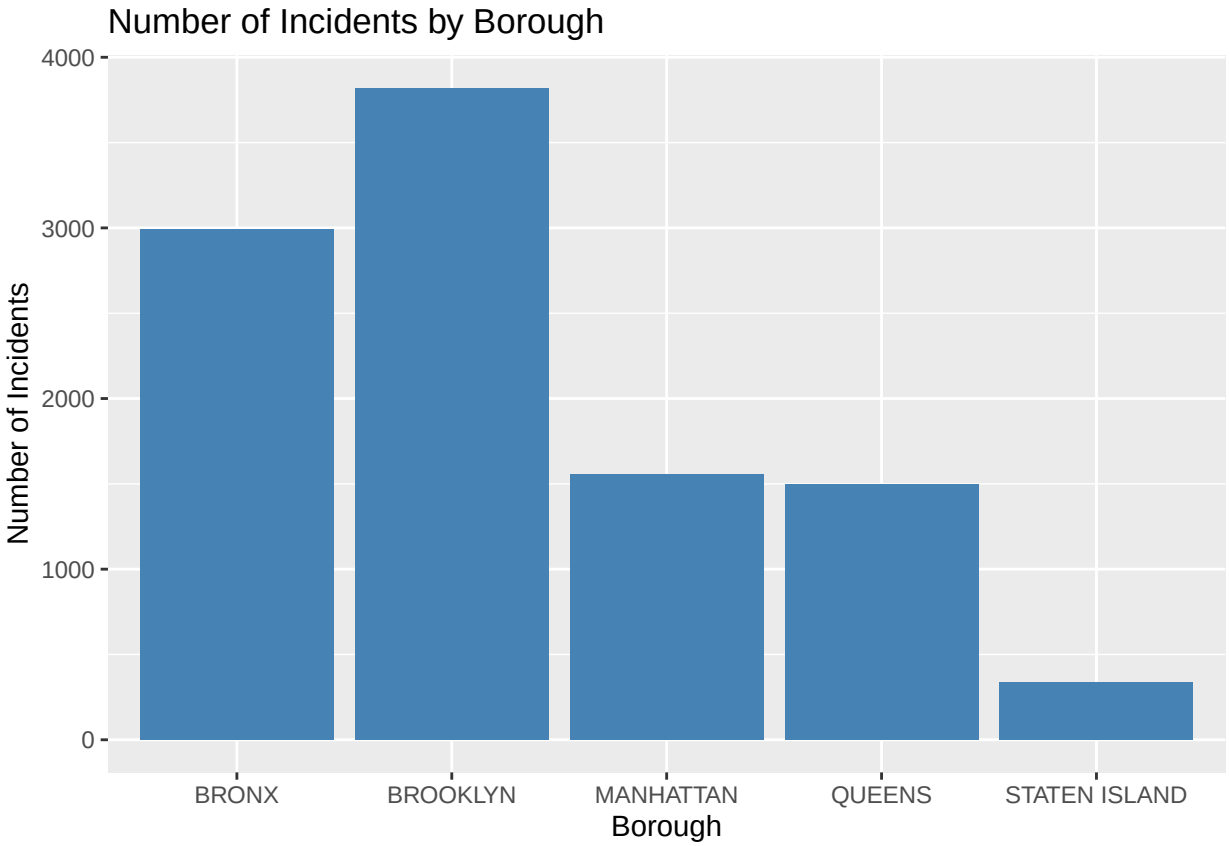
Data Analysis

Here are some basic analysis and visualizations of the data.

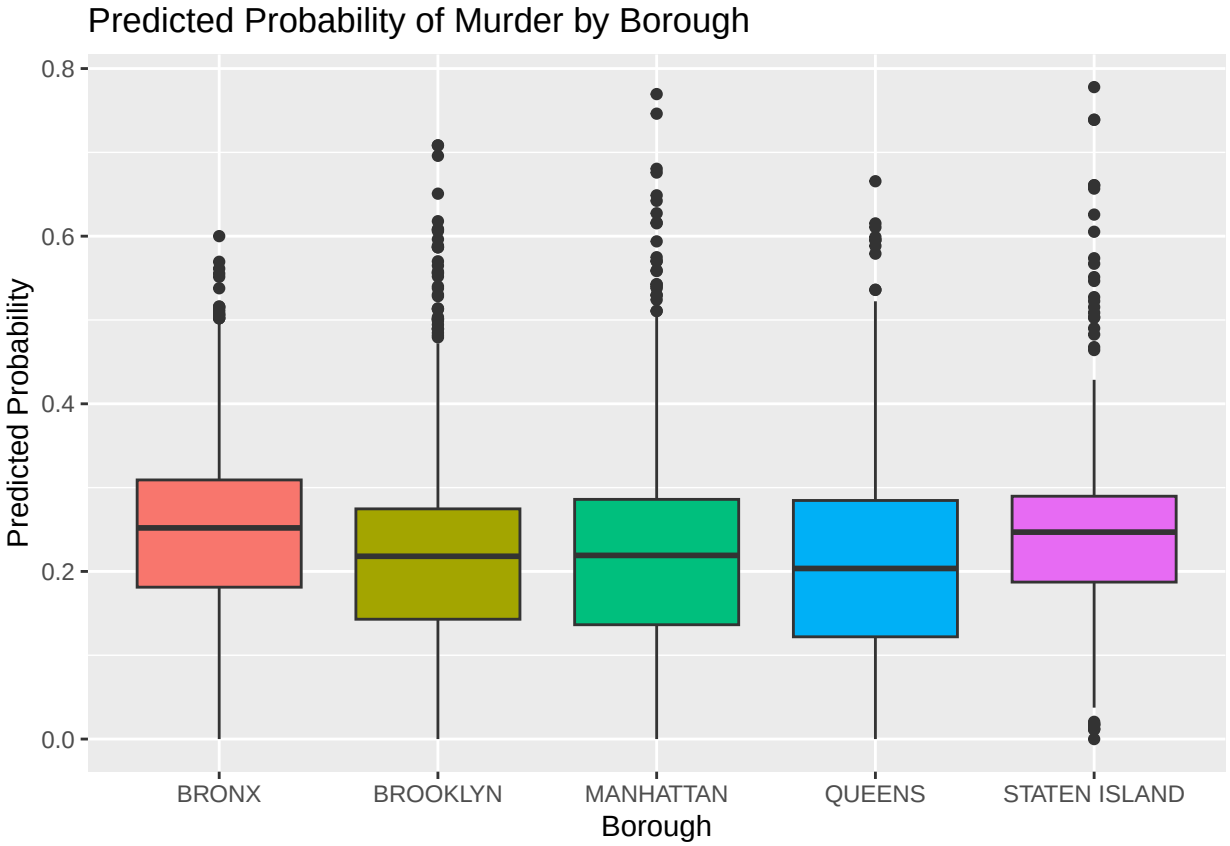
Visualization 1: Histogram of Incident Keys



Visualization 2: Bar Plot of Incidents by Borough



Modeling: Logistic Regression to Predict Murder Flag



Analysis

From the histogram of incident keys, we can observe the distribution of incidents over time.

The bar plot of incidents by borough reveals the number of incidents that occurred in each borough. The bar plot reveals that Brooklyn has the highest number of incidents, followed by the Bronx.

Modeling Insights: The logistic regression model identifies significant predictors of whether an incident is flagged as a murder. Factors such as the borough, perpetrator's age group, and sex appear to be significant. The visualization of predicted probabilities by borough shows the likelihood of an incident being flagged as a murder based on the borough.

Further Questions

1. Temporal Analysis: How do the number of incidents vary over different years and months? Are there specific periods with higher incident rates?
2. Demographic Analysis: What are the characteristics of perpetrators and victims in terms of age, sex, and race? Are there any noticeable patterns?
3. Geospatial Analysis: Are there specific precincts or locations within boroughs that have higher incident rates?

These questions can guide further investigation to uncover deeper insights into the data.

Conclusion

The analysis of the NYPD Shooting Incident Data (Historic) has provided valuable insights into the distribution of incidents over time and across different boroughs. The visualizations highlighted the concentration of incidents in specific boroughs and the overall distribution trend.

Possible Sources of Bias

1. Data Collection Bias: The data might be incomplete or inaccurate due to reporting discrepancies or missing information from certain periods or precincts.
2. Survivorship Bias: Focusing only on reported incidents may overlook unreported cases, leading to skewed results.
3. Selection Bias: The dataset might represent only certain types of incidents that meet specific criteria, excluding others that are relevant.

Personal Bias and Mitigation

As a data analyst, I acknowledge the potential for personal bias in interpreting data. To mitigate this, I have:

1. Used Standardized Methods: Followed standardized data cleaning and analysis methods to ensure consistency and reduce subjective influence.
2. Cross-Validated Findings: Compared results with existing literature and data sources to validate findings and ensure they are not isolated or anomalous.
3. Transparent Reporting: Provided clear documentation of the methods and steps taken, allowing for reproducibility and external verification.

The insights gained from this analysis can inform policy decisions, resource allocation, and targeted interventions to address and reduce shooting incidents in New York City.