

Shooting Crime in New York City

December 10, 2024

This project provides a descriptive analysis of the shooting crimes in New York City (NYC). It primarily focuses on the demographic and geographic characteristics of shooting incidents. The analysis is based on New York Police Department (NYPD) Shooting Incident data <https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic>.

Data preparation

A simple description of the data shows that it contains multiple variables.

```
## INCIDENT_KEY      OCCUR_DATE      OCCUR_TIME      BORO
## Min.      : 9953245  Length:28562    Length:28562    Length:28562
## 1st Qu.: 65439914  Class :character Class :character Class :character
## Median : 92711254  Mode  :character Mode  :character Mode  :character
## Mean      :127405824
## 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      Length:28562      Length:28562
## Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  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
```

We see that some of the variables have missing values or report unknowns although some of these variables are not key to the analysis. For the analysis, I excluded observation with missing or unknown race or age of the victim, as well as generated new variables. Note that if there are too many missing values on variable(s) that are key to the analysis, one has to do data imputation and/or look for alternative data sources or proxy variables.

```
nypd_shooting_clean <- nypd_shooting_incident %>%
  filter(VIC_RACE != "UNKNOWN", VIC_AGE_GROUP != "UNKNOWN") %>%
  mutate(
    OCCUR_DATE = as.Date(OCCUR_DATE, "%m/%d/%Y"),
    OCCUR_HOUR = lubridate::hour(as_hms(OCCUR_TIME)),
    Night = ifelse(OCCUR_HOUR >= 18 | OCCUR_HOUR < 6, 1, 0),
    CRIME = case_when(STATISTICAL_MURDER_FLAG == "true" ~ "Murder",
                      STATISTICAL_MURDER_FLAG == "false" ~ "Not Murder",
                      TRUE ~ STATISTICAL_MURDER_FLAG),
    Murder = ifelse(CRIME == "Murder", 1, 0),
    Male = ifelse(VIC_SEX == "M", 1, 0),
    Young = ifelse(VIC_AGE_GROUP == "18-24", 1, 0),
    Black = ifelse(VIC_RACE == "BLACK", 1, 0),
    Brooklyn = ifelse(BORO == "BROOKLYN", 1, 0),
    Bronx = ifelse(BORO == "BRONX", 1, 0),
    Manhattan = ifelse(BORO == "MANHATTAN", 1, 0),
    Queens = ifelse(BORO == "QUEENS", 1, 0),
    Staten_Island = ifelse(BORO == "STATEN ISLAND", 1, 0)) %>%
  select(-c(LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE,
            LOC_CLASSFCTN_DESC, LOCATION_DESC, X_COORD_CD,
            Y_COORD_CD, Latitude, Longitude, Lon_Lat))
```

```
## INCIDENT_KEY      OCCUR_DATE      OCCUR_TIME      BORO
## Min.   : 9953245    Min.   :2006-01-01    Length:28447      Length:28447
## 1st Qu.: 65441446    1st Qu.:2009-09-05    Class :character   Class :character
## Median : 92728167    Median :2013-09-21    Mode  :character   Mode  :character
## Mean   :127490410    Mean   :2014-06-09
## 3rd Qu.:203517392    3rd Qu.:2019-10-07
## Max.   :279758069    Max.   :2023-12-29
## STATISTICAL_MURDER_FLAG PERP_AGE_GROUP      PERP_SEX
## Length:28447           Length:28447        Length:28447
## Class :character       Class :character    Class :character
```

```

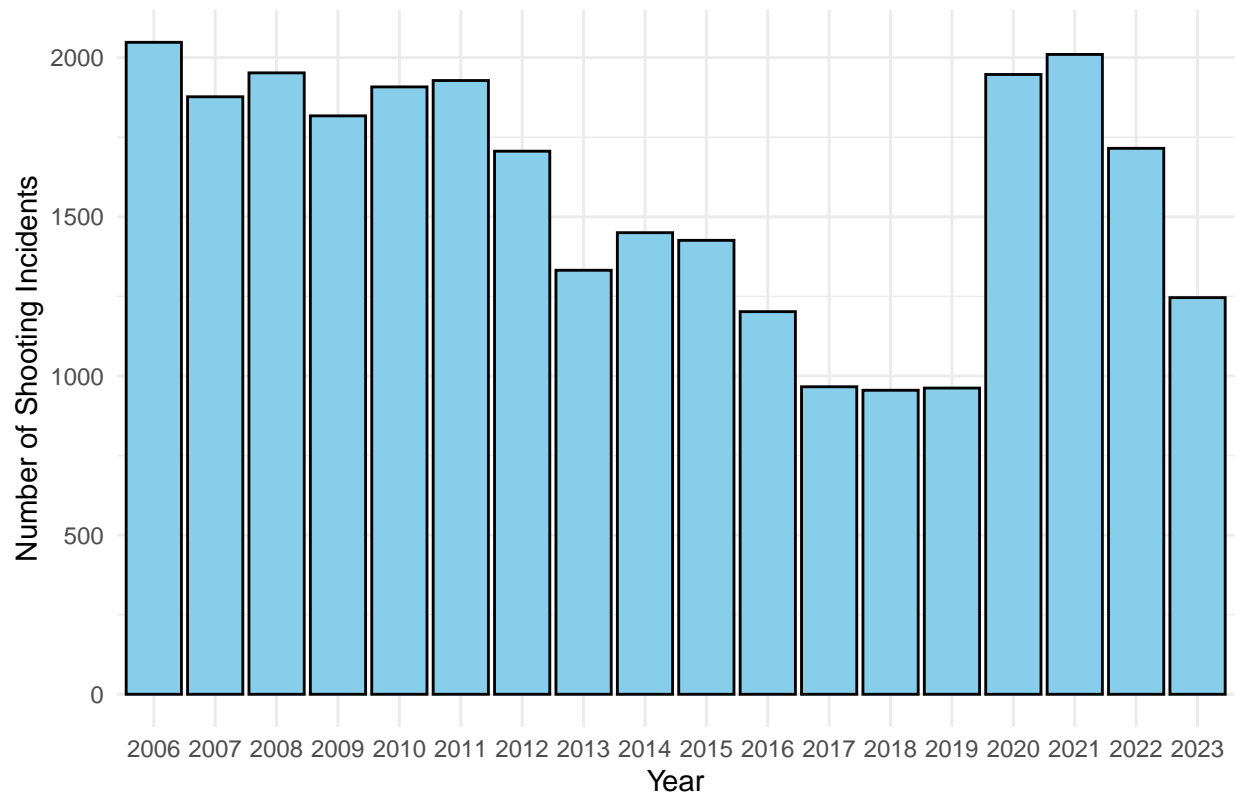
## Mode :character      Mode :character      Mode :character
##
##
##
## PERP_RACE      VIC_AGE_GROUP      VIC_SEX      VIC_RACE
## Length:28447    Length:28447    Length:28447    Length:28447
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
## OCCUR_HOUR      Night      CRIME      Murder
## Min. : 0.00      Min. :0.000      Length:28447    Min. :0.0000
## 1st Qu.: 3.00      1st Qu.:1.000      Class :character 1st Qu.:0.0000
## Median :15.00      Median :1.000      Mode :character  Median :0.0000
## Mean :12.27      Mean :0.754      Mean :0.1935
## 3rd Qu.:20.00      3rd Qu.:1.000      3rd Qu.:0.0000
## Max. :23.00      Max. :1.000      Max. :1.0000
## Male      Young      Black      Brooklyn
## Min. :0.000      Min. :0.0000      Min. :0.0000      Min. :0.0000
## 1st Qu.:1.000      1st Qu.:0.0000      1st Qu.:0.0000      1st Qu.:0.0000
## Median :1.000      Median :0.0000      Median :1.0000      Median :0.0000
## Mean :0.903      Mean :0.3643      Mean :0.7107      Mean :0.3972
## 3rd Qu.:1.000      3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:1.0000
## Max. :1.000      Max. :1.0000      Max. :1.0000      Max. :1.0000
## Bronx      Manhattan      Queens      Staten_Island
## Min. :0.0000      Min. :0.000      Min. :0.0000      Min. :0.00000
## 1st Qu.:0.0000      1st Qu.:0.000      1st Qu.:0.0000      1st Qu.:0.00000
## Median :0.0000      Median :0.000      Median :0.0000      Median :0.00000
## Mean :0.2932      Mean :0.132      Mean :0.1493      Mean :0.02826
## 3rd Qu.:1.0000      3rd Qu.:0.000      3rd Qu.:0.0000      3rd Qu.:0.00000
## Max. :1.0000      Max. :1.000      Max. :1.0000      Max. :1.00000

```

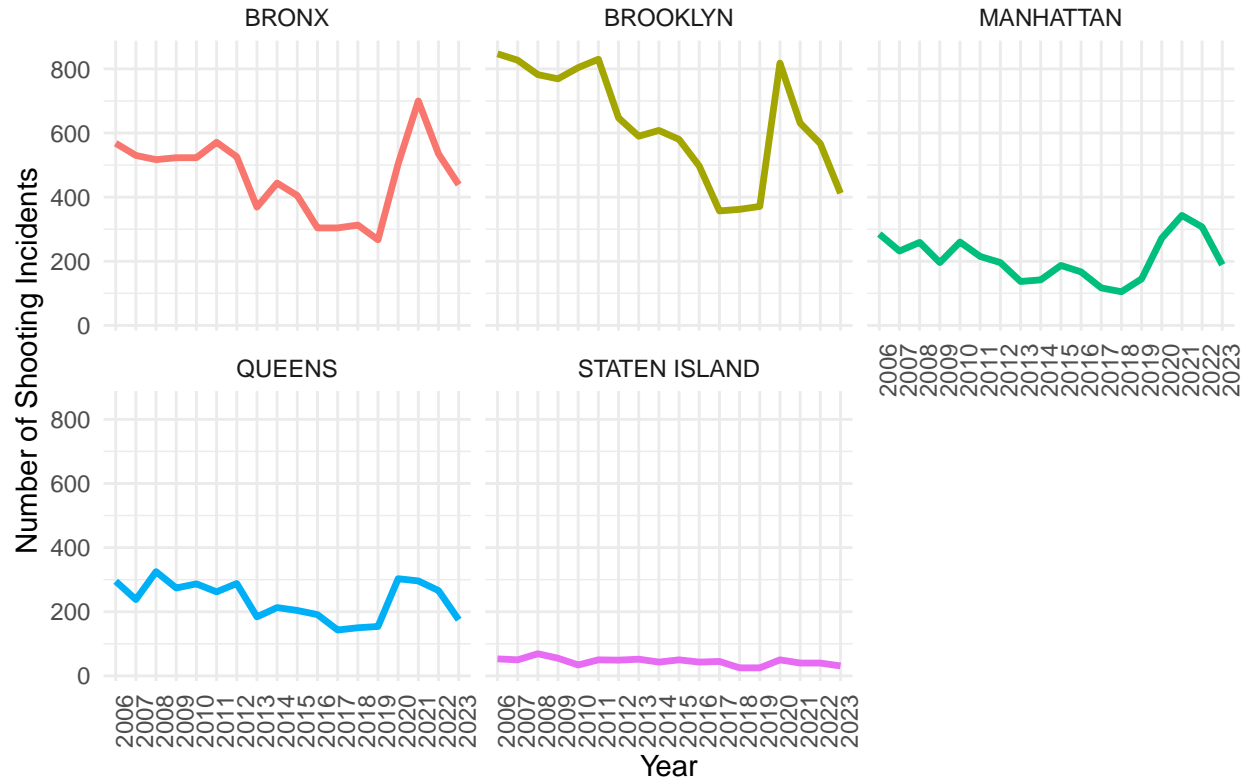
Data Visualization

The evolution of shooting incidents in NYC over time shows a general decline in such incidents. However, there was a significant increase during the COVID-19 pandemic years between 2020 and 2022, although the numbers later decreased to levels comparable to those in 2016. Additionally, the distribution of crime is uneven across the boroughs of NYC. The Bronx and Brooklyn appear to be the primary locations for most shooting incidents, whereas Staten Island and Manhattan experienced significantly fewer incidents.

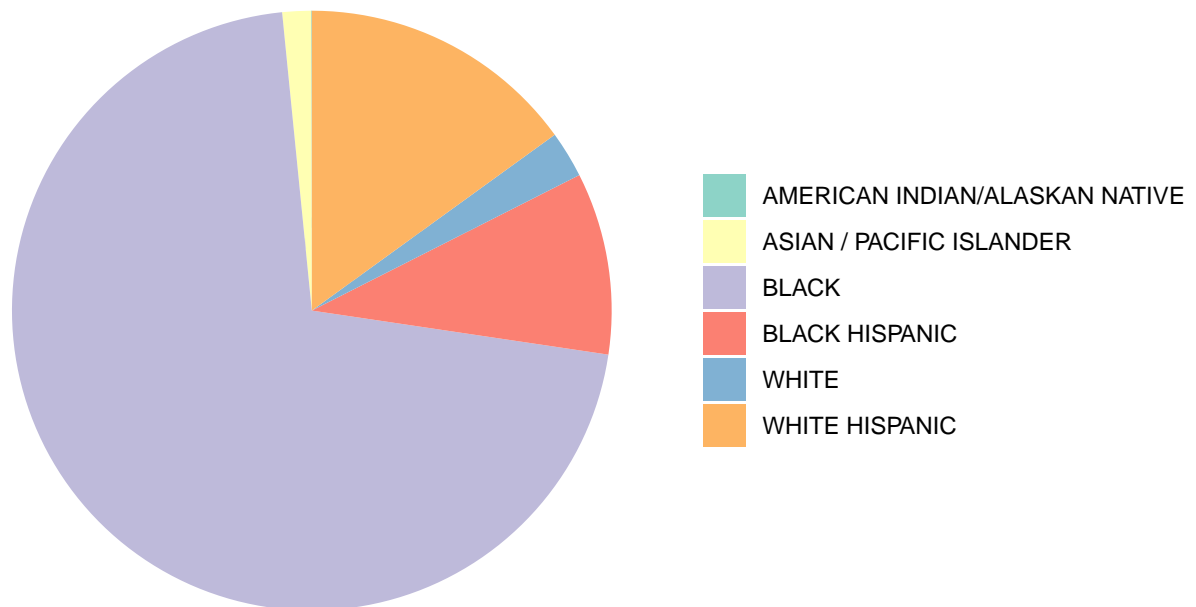
Total Shooting Incidents by Year



Number of Shooting Incidents by Borough by Year



Shooting Incidents by Race, 2006–23

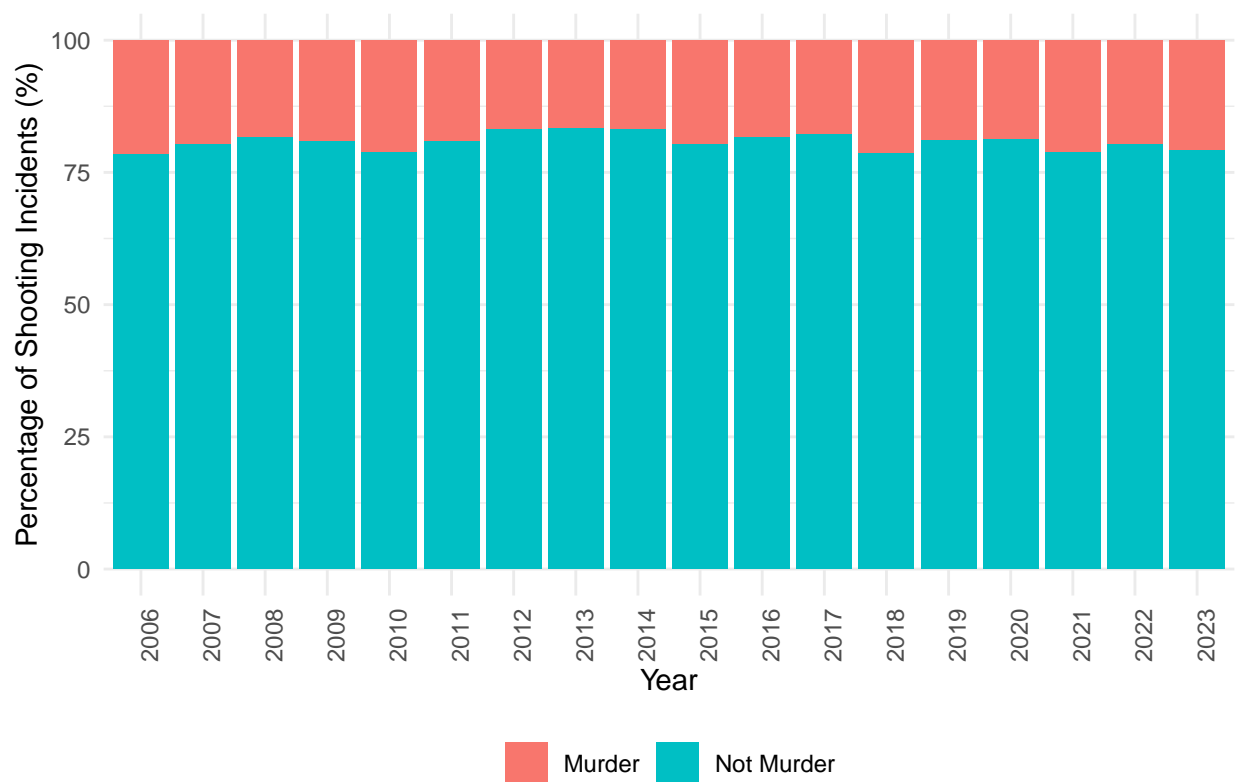


An analysis of the demographic characteristics of these crimes reveals that different communities are disproportionately affected by criminal activity. A significant number of victims were Black. White Hispanics and Black Hispanics also experienced higher levels of criminal activity compared to other racial groups.

Model of crime severity

Here, I examined the likelihood and severity of shooting incidents, differentiating between incidents that led to murder and those that did not. As shown by the proportion of shooting incidents by severity, the majority were non-fatal; however, the fatality rate was still relatively high. To explore potential factors contributing to this, I applied a simple yet effective modeling approach: a linear probability model.

Percentage of Shooting Incidents by Crime and Year



```
##
## Call:
## lm(formula = Murder ~ Male + Young + Black + Night + Brooklyn +
##       Bronx + Queens + Staten_Island, data = nypd_shooting_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.2539 -0.2033 -0.1841 -0.1600  0.8569
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.220297   0.011004  20.020 < 2e-16 ***
## Male          -0.005693   0.007917  -0.719  0.472077
## Young         -0.040992   0.004872  -8.414 < 2e-16 ***
## Black         -0.011087   0.005318  -2.085  0.037103 *
## Night         -0.019402   0.005442  -3.565  0.000364 ***
## Brooklyn      0.017736   0.007505   2.363  0.018123 *
## Bronx         0.016854   0.007763   2.171  0.029935 *
## Queens        0.019195   0.008846   2.170  0.030020 *
## Staten_Island 0.033638   0.015333   2.194  0.028252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3944 on 28438 degrees of freedom
## Multiple R-squared:  0.003528,    Adjusted R-squared:  0.003248
## F-statistic: 12.59 on 8 and 28438 DF,  p-value: < 2.2e-16
```

The model runs a regression of murder on indicators such as gender (male), age (youth, 18-24), race (Black), and location. As shown by the estimates, the model does not explain much of the variation in the likelihood of murder, so caution is necessary when drawing conclusions. However, the estimates reveal interesting and sometimes counterintuitive results: there is no significant difference between males and females, youth and Black individuals were less likely to be killed, nighttime shootings were negatively associated with the likelihood of murder, and compared with Manhattan, the other boroughs appeared to have a higher chance of a shooting incident being fatal.

Conclusion

The analysis showed a declining trend in shooting incidents over time until the COVID-19 pandemic, during which incidents significantly increased. It also highlighted the uneven distribution of shooting incidents across different locations and communities. Additionally, a simple model was applied to examine the relationship between victim characteristics, location of incidents, and the fatality of the incidents. However, the model lacked predictive power.

Despite the useful insights from the descriptive analysis, interpretation of the results requires caution. A main source of bias may be due to a premature conclusion from the descriptive statistics that shooting incidents are inherent to certain locations or demographic behaviors. However, this is erroneous because the model did not account for variables known to have a strong relationship with crime. For example, the negative significant coefficient for Black individuals, despite a high number of shooting incidents among Blacks, could be due to the failure to control for key socioeconomic factors, which may lead to biased estimates and conclusions. These factors include household income, age distribution, population composition and density, and the level of policing activity. The descriptive analysis provides key steppingstones for more rigorous and causal analysis. Therefore, it is necessary to collect more data and run additional robustness checks.

Additional material

GitHub: <https://github.com/kgmayd/NYC-Shooting-Incident>