NYPD Shooting Incidents (2006–2024): A Data-Driven Analysis

Eric F.

2025-04-29

Contents

1 Introduction

This project presents an exploratory and statistical analysis of the NYPD Shooting Incident Data (Historic), a public dataset provided by the City of New York and last updated in April 2025. The dataset is available online (https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic) and contains detailed information on every shooting incident that occurred in New York City from 2006 through the end of the previous calendar year. Each row represents a single shooting event and includes variables related to location, time, suspect and victim demographics, and whether the incident resulted in a murder.

The goal of this project is to import, clean, and analyze the dataset using the tools of data science in R. The analysis includes time-based trends, demographic distributions, and outcome-based visualizations. The project is intended to be reproducible, hence all processing steps are explicitly documented in code chunks.

In addition to generating insights, we reflect on potential biases in the dataset—such as underreporting, incomplete demographic information, and temporal gaps—and how these may impact the conclusions drawn from the analysis.

2 Extract, Transform and Load (ETL) data

First, we install packages and import the necessary libraries:

```r
packages <- c("tidyverse", "lubridate", "dplyr", "naniar",
              "hms", "scales", "patchwork", "forcats",
              "knitr", "kableExtra")

# Install packages
installed <- packages %in% rownames(installed.packages())
if (any(!installed)) {
  install.packages(packages[!installed])
}
```

```r
library(tidyverse)
library(lubridate)
library(dplyr)
library(naniar)
library(hms)
library(scales)
library(forcats)
```

Then, we copy the data url and create the R container for it. Now that let's us see the structure of the dataset.

```r
url_in <-
 ↪  "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

NY_shootings <- read.csv(url_in)

# Convert "" to NA so that vis_miss works
NY_shootings[NY_shootings == ""] <- NA


# Check the structure of the dataset
str(NY_shootings)
```

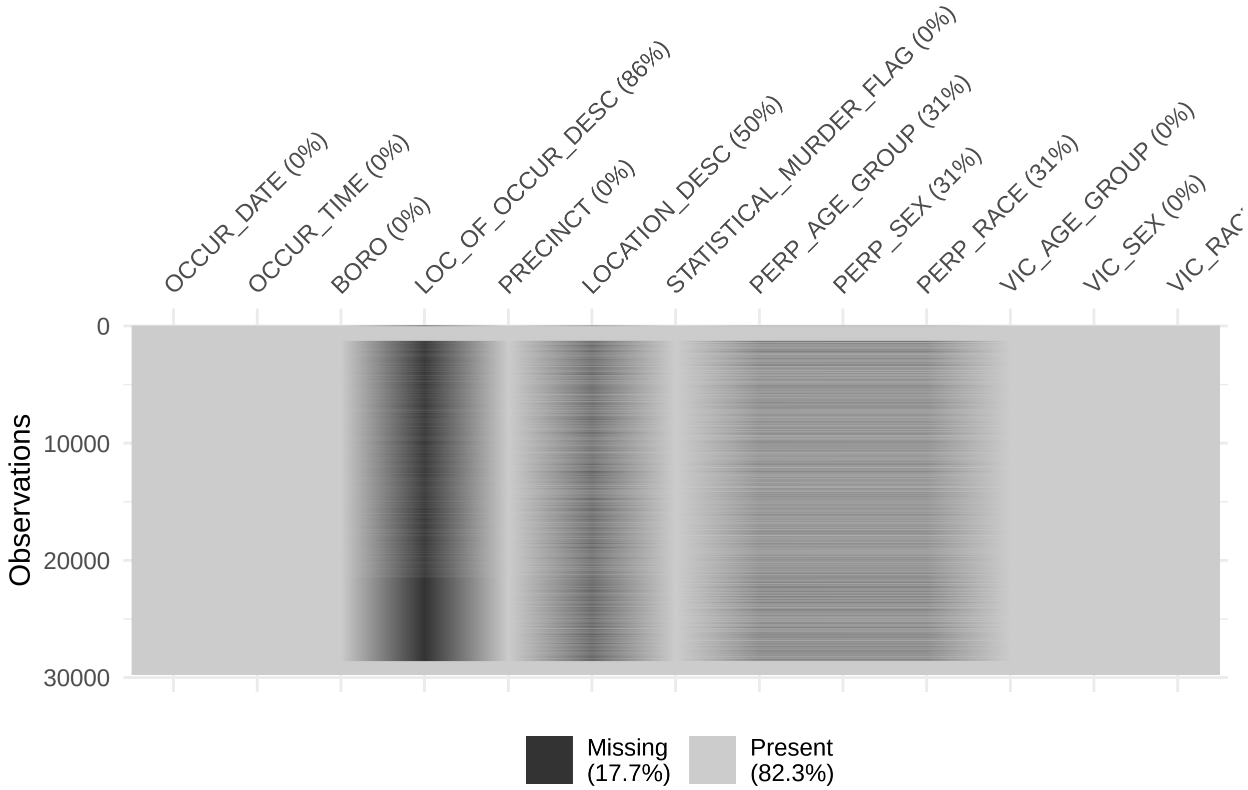
```
## 'data.frame':    29744 obs. of  21 variables:
##  $ INCIDENT_KEY           : int  231974218 177934247 255028563 25384540 72616285 8587
##  $ OCCUR_DATE             : chr  "08/09/2021" "04/07/2018" "12/02/2022" "11/19/2006"
##  $ OCCUR_TIME             : chr  "01:06:00" "19:48:00" "22:57:00" "01:50:00" ...
##  $ BORO                   : chr  "BRONX" "BROOKLYN" "BRONX" "BROOKLYN" ...
##  $ LOC_OF_OCCUR_DESC      : chr  NA NA "OUTSIDE" NA ...
##  $ PRECINCT               : int  40 79 47 66 46 42 71 69 75 69 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 2 0 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr  NA NA "STREET" NA ...
##  $ LOCATION_DESC          : chr  NA NA "GROCERY/BODEGA" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "true" "false" "true" ...
##  $ PERP_AGE_GROUP         : chr  NA "25-44" "(null)" "UNKNOWN" ...
##  $ PERP_SEX               : chr  NA "M" "(null)" "U" ...
##  $ PERP_RACE              : chr  NA "WHITE HISPANIC" "(null)" "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "25-44" "25-44" "18-24" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : chr  "1006343" "1000082.937500000000000" "1020691" "98510
##  $ Y_COORD_CD             : chr  "234270" "189064.671875000000000" "257125" "173349.7
##  $ Latitude               : num  40.8 40.7 40.9 40.6 40.8 ...
##  $ Longitude              : num  -73.9 -73.9 -73.9 -74 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.92019278899994 40.80967347200004)" "POIN
```

Since they do not give us useful information (for our analysis), we can get rid of the columns INCIDENT_KEY, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat. We also change the data type of OCCUR_DATE and OCCUR_TIME appropiately.

```r
# Change types and get rid of useless info
NY_shootings <- NY_shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(OCCUR_TIME = as_hms(OCCUR_TIME)) %>%
  select(-c(INCIDENT_KEY, JURISDICTION_CODE, LOC_CLASSFCTN_DESC,
   ↪  X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
```

Now, with vis_miss(dataset) we can visualize how many data points are missing for each attribute.

```
vis_miss(NY_shootings)
```

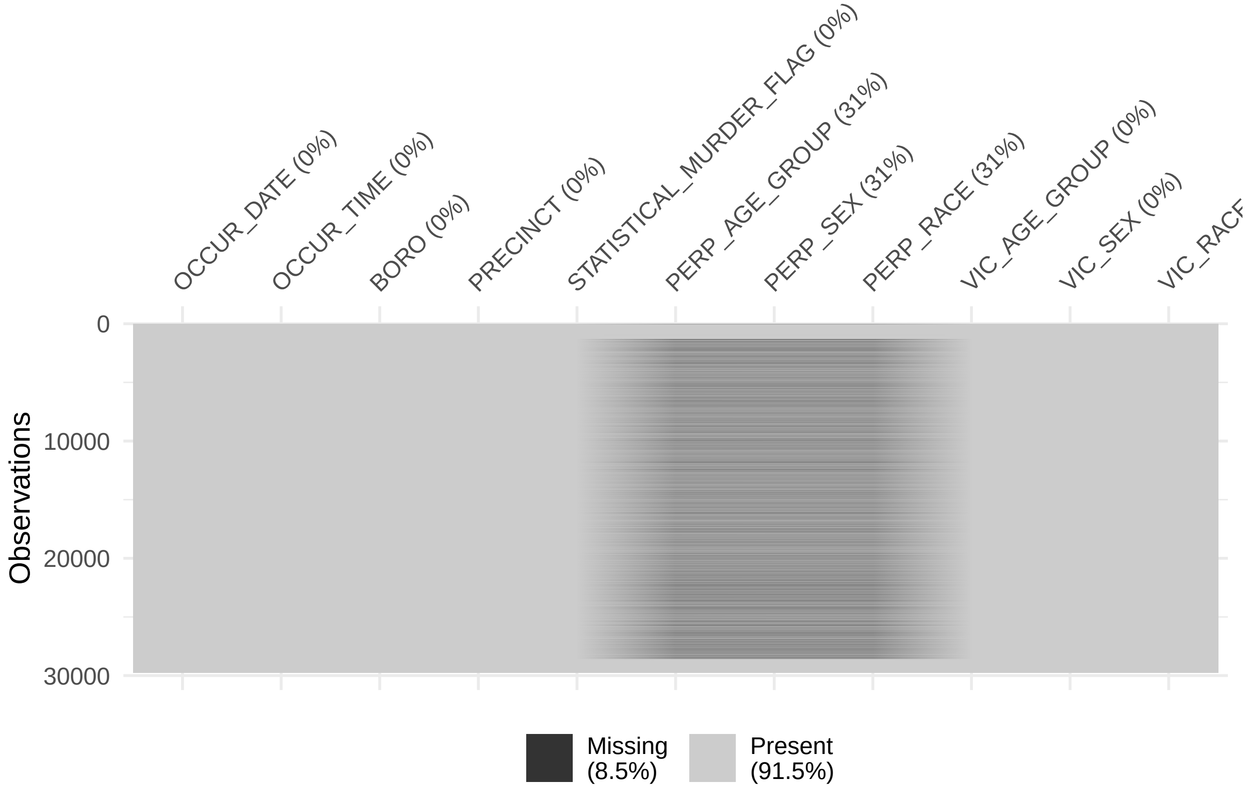


Since it is clear that LOCATION_DESC and LOC_OF_OCCUR_DESC have more than 50% of the data missing (86% in the case of the second one), we will get rid of them.

```
NY_shootings <- NY_shootings %>%
  select(-c(LOCATION_DESC, LOC_OF_OCCUR_DESC))
```

Finally, we can see the NAs of the final dataset:

```
vis_miss(NY_shootings)
```



As well as the summary of the final dataset:

```
summary(NY_shootings)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO               PRECINCT     
##  Min.   :2006-01-01   Length:29744      Length:29744       Min.   :  1.00  
##  1st Qu.:2009-10-29   Class1:hms        Class :character   1st Qu.: 44.00  
##  Median :2014-03-25   Class2:difftime   Mode  :character   Median : 67.00  
##  Mean   :2014-10-31   Mode  :numeric                       Mean   : 65.23  
##  3rd Qu.:2020-06-29                                        3rd Qu.: 81.00  
##  Max.   :2024-12-31                                        Max.   :123.00  
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX        
##  Length:29744            Length:29744       Length:29744      
##  Class :character        Class :character   Class :character  
##  Mode  :character        Mode  :character   Mode  :character  
##                                                               
##                                                               
##                                                               
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX             VIC_RACE        
##  Length:29744       Length:29744       Length:29744       Length:29744      
##  Class :character   Class :character   Class :character   Class :character  
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

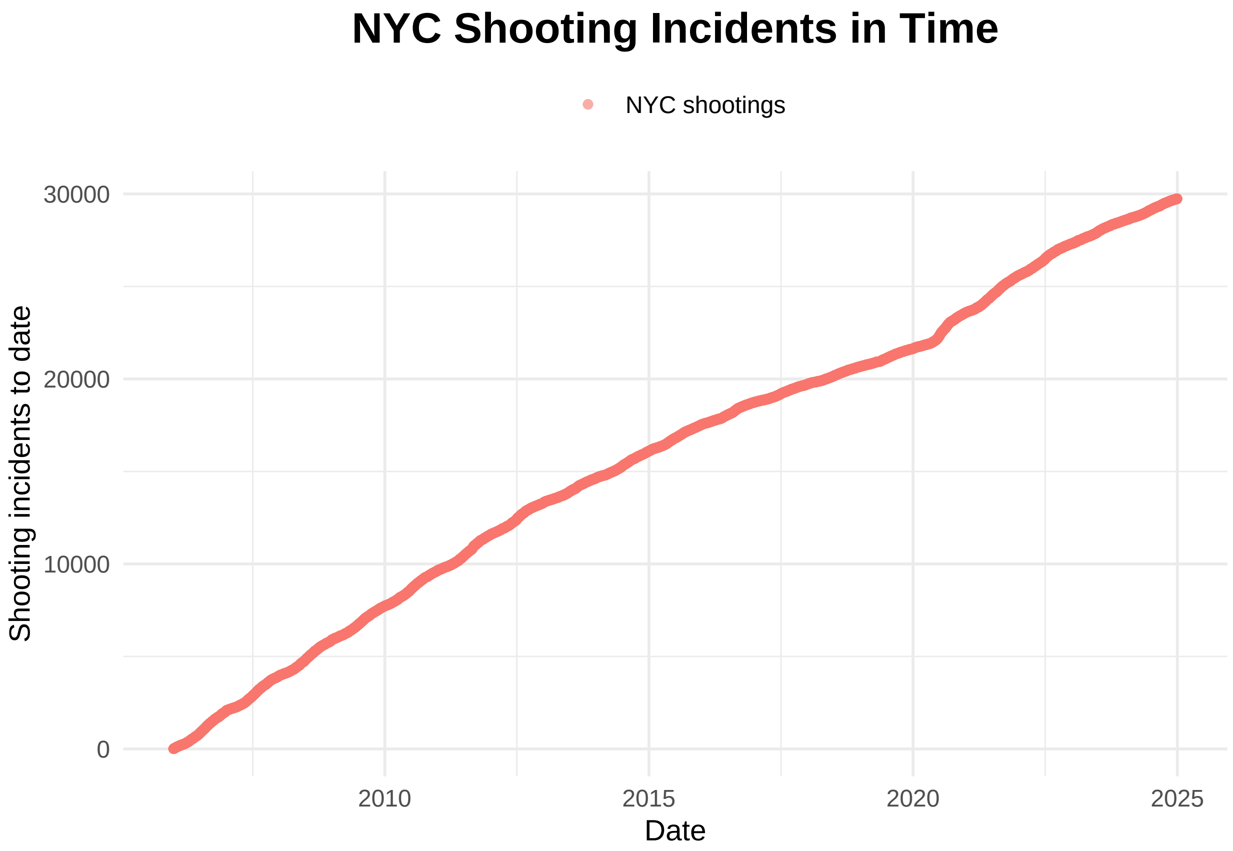
3 Exploratory Data Analysis (EDA)

3.1 Temporal analysis of NYC shooting incidents

3.1.1 Shooting incidents in time

Now we are able to explore the temporal trends of the shooting incidents in NYC. First, we can see that the cumulative plot below reveals a consistent increase in shooting incidents from 2006 to 2024, indicating that gun violence is a persistent issue directly proportional to population size, with no sustained long-term decline.

```r
#create table grouped by date
NY_shootings_by_date <- NY_shootings %>%
          group_by(OCCUR_DATE) %>% #group by date
          summarise(cases = n()) %>%
          ungroup() %>%
          arrange(OCCUR_DATE) %>%  # chronological order
          mutate(cumulative_cases = cumsum(cases)) #create column


NY_shootings_by_date_and_stat <- NY_shootings %>%
     group_by(OCCUR_DATE, STATISTICAL_MURDER_FLAG) %>%
     summarise(cases = n()) %>%
     ungroup()


ggplot(NY_shootings_by_date, aes(x = OCCUR_DATE)) +
    geom_point(aes(y = cumulative_cases, color = "NYC shootings"),
               size = 1.2, alpha = 0.6) +
    labs(
        title = "NYC Shooting Incidents in Time",
        x = "Date",
        y = "Shooting incidents to date"
    ) +
    theme_minimal() +
    theme(
        plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
        legend.title = element_blank(),   # hide legend title
        legend.position = "top"
    )
```

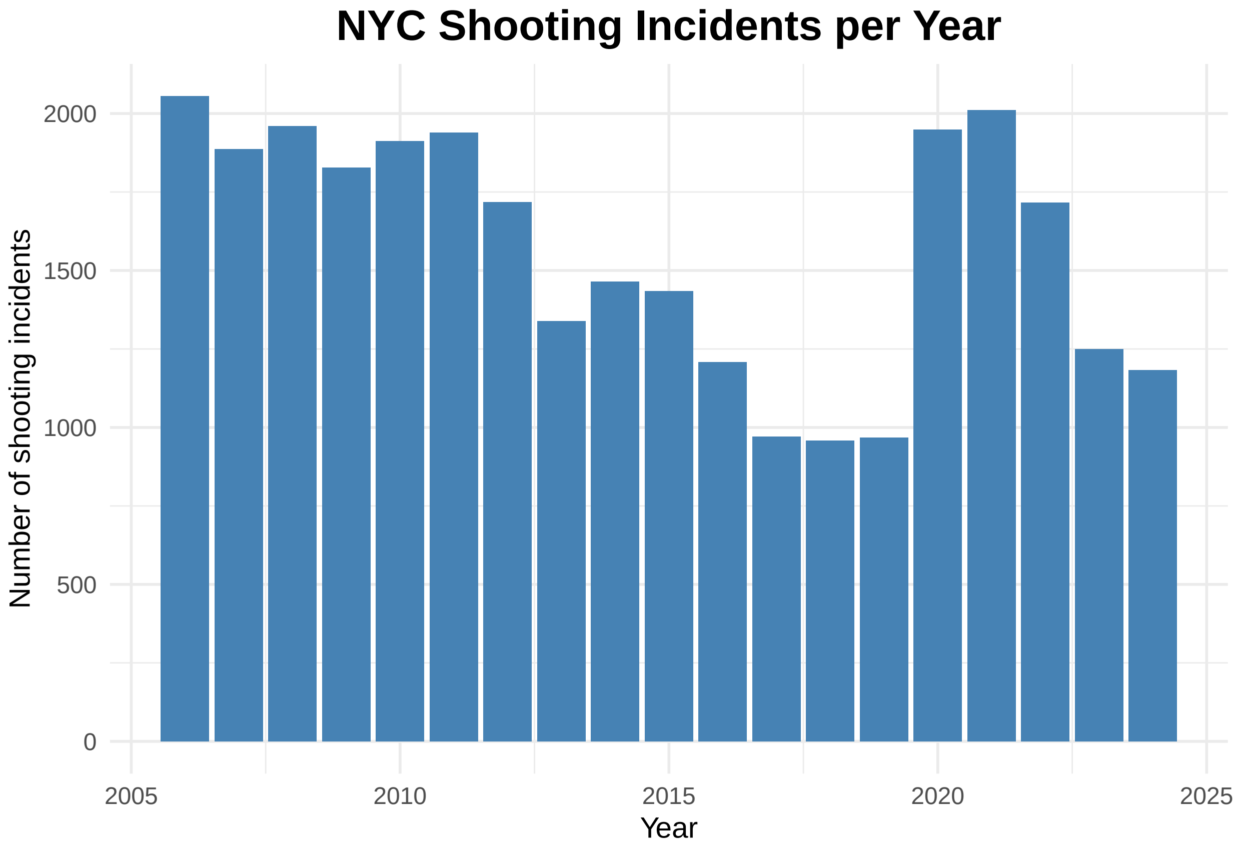


3.1.2 Shooting incidents per year

Now we can see that the bar chart of shooting incidents by year adds granularity. After a noticeable decline between 2011 and 2018, there was a sharp resurgence in 2020 and 2021—coinciding with the COVID-19 pandemic, social unrest, and reduced police presence in various neighborhoods. The trend again declines post-2021, signaling the post-COVID recovering. This shows that, although gun violence incidents in NYC will always grow due to population growth, the yearly rates of gun violence in NYC were actually decreasing and the city was actually becoming more peaceful, until the pandemic came and the number of incidents doubled from 2019 to 2020.

```
NY_shootings_by_year <- NY_shootings_by_date %>%
     mutate(year = year(OCCUR_DATE)) %>% #Extract year
     group_by(year) %>%
     summarise(total_cases = sum(cases)) %>%
     ungroup() %>%
     arrange(year)
```

```
ggplot(NY_shootings_by_year, aes(x=year, y=total_cases)) +
      geom_bar(stat="identity", fill="steelblue") +
      labs(title="NYC Shooting Incidents per Year", x="Year", y="Number of
       ↪  shooting incidents") +
  theme_minimal() +
  theme(
         plot.title = element_text(hjust = 0.5, face = "bold", size = 16))
```

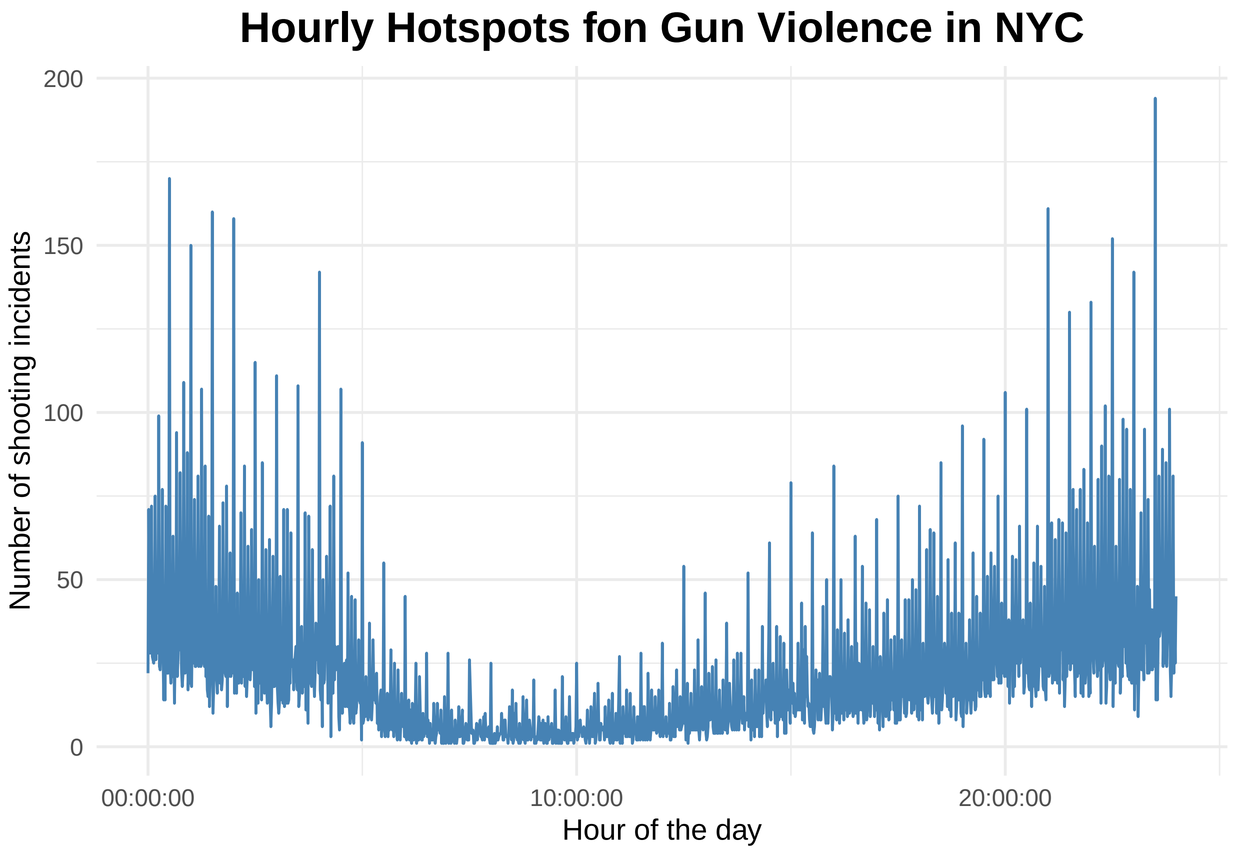


3.1.3 Shooting incidents by hour of the day

If we plot the number of shooting incidents against the hour of the day, there is a clear lesson: the safest time in NYC is around 08:00 in the morning, between 06:00 and 10:00, since there were very few shootings within that range; on the other hand, the most dangerous time is around midnight, between 20:00 and 02:00, since there are clear peaks at those hours.

These findings correlate strongly with social behavior patterns: nightlife, alcohol consumption, and late-night public activity appear as contributing factors. The pattern suggests that preventive policy efforts may be more effective if targeted at evening and early night hours.

```
NY_shootings_by_hour <- NY_shootings %>%
     group_by(OCCUR_TIME) %>%
     summarise(cases = n()) %>%
     ungroup()

ggplot(NY_shootings_by_hour, aes(x = OCCUR_TIME, y = cases)) +
     geom_line(color = "steelblue") +
     labs(title = "Hourly Hotspots fon Gun Violence in NYC",
          x = "Hour of the day",
          y = "Number of shooting incidents") +
     theme_minimal() +
  theme(
        plot.title = element_text(hjust = 0.5, face = "bold", size = 16))
```



3.1.4 Daily hotspots of gun violence

The heatmap provides a detailed visualization of how shooting incidents in New York City vary across both the day of the week and the hour of the day.

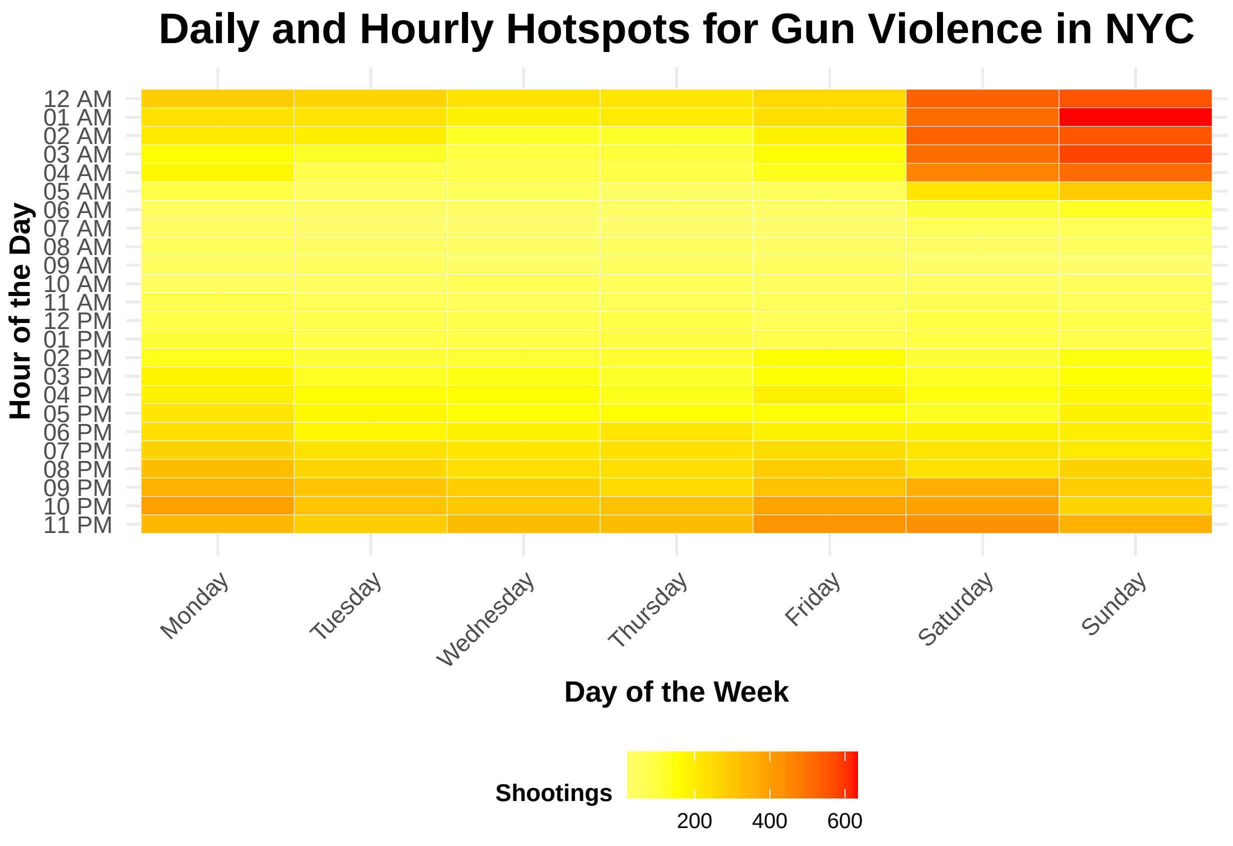
The most dangerous periods are clearly concentrated between midnight and 3 AM, especially on Sundays and Saturdays. This temporal clustering likely reflects increased late-night activity during weekends, including nightlife, alcohol consumption, and social gatherings, all of which can exacerbate the likelihood of violent encounters.

By contrast, the safest hours are observed between 6 AM and 10 AM across all days, aligning with reduced public activity during early morning hours and a general transition into work or school routines.

Interestingly, weekdays show a significantly lower incidence of shootings during nighttime hours compared to weekends, suggesting that work-week structures and reduced leisure activity during these days have a mitigating effect on violent crime.

Overall, the heatmap highlights that public safety initiatives could be most effectively targeted at weekend late-night hours, particularly around bars, clubs, and large gatherings. Tailoring police deployment and community outreach programs to these high-risk periods could be a strategic approach to reducing gun violence in New York City.

```r
invisible(Sys.setlocale("LC_TIME", "C"))  # Print dates in english


# Group by day and hour
NY_shootings_by_hour_day <- NY_shootings %>%
  mutate(
    DayOfWeek = wday(OCCUR_DATE, label = TRUE, abbr = FALSE),
    DayOfWeek = fct_relevel(DayOfWeek,
                            "Monday", "Tuesday", "Wednesday", "Thursday",
                            "Friday", "Saturday", "Sunday"),
    Hour = hour(OCCUR_TIME)
  ) %>%
  group_by(DayOfWeek, Hour) %>%
  summarise(Incidents = n(), .groups = "drop")

# Heatmap
ggplot(NY_shootings_by_hour_day, aes(x = DayOfWeek, y = Hour, fill =
 ↪  Incidents)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(
    low = "white", mid = "yellow", high = "red",
    midpoint = median(NY_shootings_by_hour_day$Incidents),
    name = "Shootings"
  ) +
  scale_y_reverse(
    breaks = seq(0, 23),
    labels = format(strptime(seq(0, 23), format = "%H"), "%I %p")
  ) +
  labs(
    title = "Daily and Hourly Hotspots for Gun Violence in NYC",
    x = "Day of the Week",
    y = "Hour of the Day"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    panel.grid.minor = element_blank(),
    legend.position = "bottom",
    legend.title = element_text(face = "bold", size = 9),
    legend.text = element_text(size = 8)
  )
```

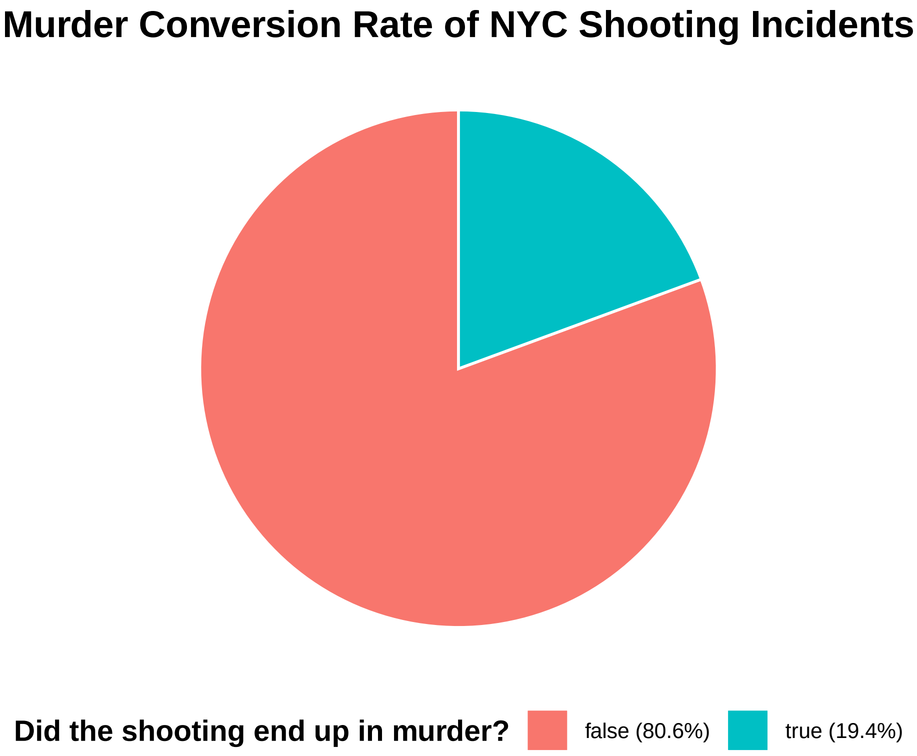


3.2 Murder Conversion Rate of NYC shooting incidents

We can also study the outcome of the shooting incidents, the murder conversion rate (whether someone died or not), as seen below. This pie chart shows that around 19.4% of shootings resulted in murder and that the vast majority (80.6%) did not result in fatality.

This is essential when discussing public perception versus reality: although all shootings are serious, not all result in homicide. However, the 1-in-5 fatality rate still highlights the lethality of gun violence in NYC.

```
NY_shootings_by_stat_flag <- NY_shootings %>%
          group_by(STATISTICAL_MURDER_FLAG) %>%
          summarise(cases = n()) %>%
          ungroup()


NY_shootings_by_stat_flag %>%
  mutate(percentage = cases / sum(cases) * 100,
         STATISTICAL_MURDER_FLAG = paste0(STATISTICAL_MURDER_FLAG, " (",
          ↪  round(percentage, 1), "%)")) %>%
  ggplot(aes(x = "", y = cases, fill = STATISTICAL_MURDER_FLAG)) +
  geom_col(width = 1, color = "white") +
  coord_polar(theta = "y") +
  labs(
    title = "Murder Conversion Rate of NYC Shooting Incidents",
    fill = "Did the shooting end up in murder?"
  ) +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    legend.title = element_text(face = "bold"),
    legend.position = "bottom",
    legend.text = element_text(size = 8),  # legend size
    legend.box = "horizontal",
    legend.justification = "center",
    plot.margin = margin(20, 20, 20, 20)
  )
```

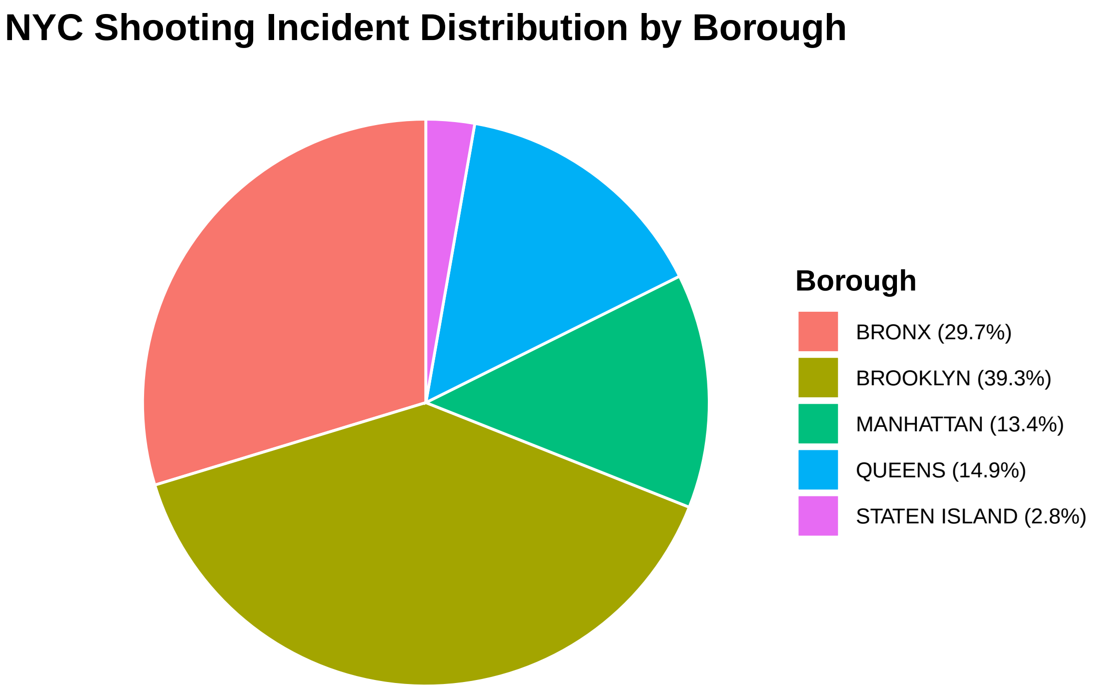


3.3 Spatial analysis of NYC shooting incidents

A geographical analysis can also be performed on the data. For example, the pie chart of borough distribution below shows a highly uneven spatial concentration: Brooklyn accounts for ~39% of all shootings and the Bronx follows with ~30%, while Manhattan, Queens, and Staten Island together make up the remaining third.

This suggests a spatial inequality in exposure to gun violence, likely tied to systemic socioeconomic disparities, varying levels of law enforcement presence, and neighborhood characteristics.

```r
NY_shootings_by_boro <- NY_shootings %>%
     group_by(BORO) %>%
     summarise(cases = n()) %>%
     ungroup()


NY_shootings_by_boro %>%
  mutate(
    percentage = cases / sum(cases) * 100,
    BORO = fct_reorder(BORO, cases)  # 1. Reordena BORO según cases
  ) %>%
  mutate(
    BORO_label = paste0(BORO, " (", round(percentage, 1), "%)")  # 2. Crea
     ↪  nueva columna para mostrar en gráfico
  ) %>%
  ggplot(aes(x = "", y = cases, fill = BORO_label)) +  # 3. Usa BORO_label
   ↪  como fill
  geom_col(width = 1, color = "white") +
  coord_polar(theta = "y") +
  labs(
    title = "NYC Shooting Incident Distribution by Borough",
    fill = "Borough"
  ) +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    legend.title = element_text(face = "bold"),
    legend.position = "right",
    legend.text = element_text(size = 8),
    legend.box = "horizontal",
    legend.justification = "center",
    plot.margin = margin(20, 20, 20, 20)
  )
```



3.4 Demographics of NYC shooting incidents

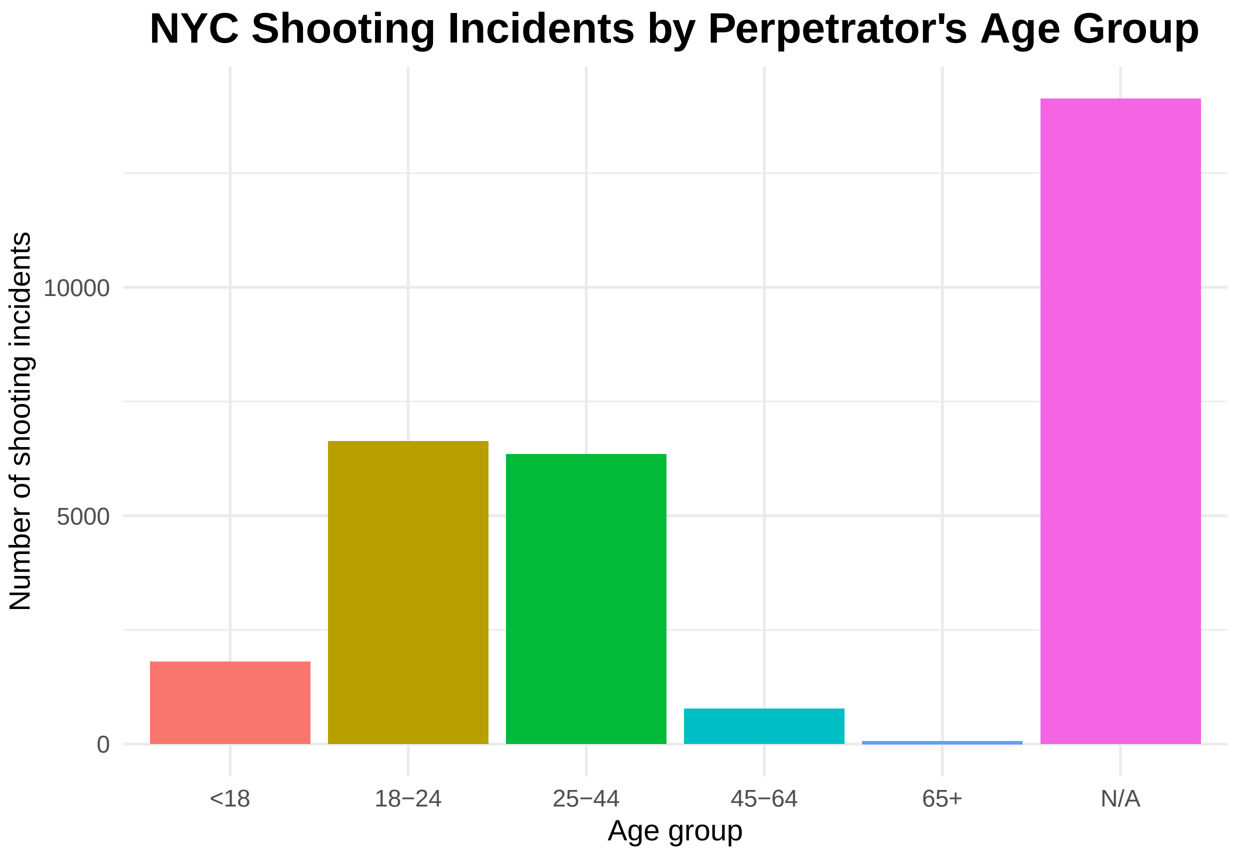
Finally, we can also analyze some demographics.

3.4.1 Age group of the perpetrators

First, the age data below reflects that most perpetrators are aged 18–44, with the 18–24 group slightly ahead.

A large number of perpetrators have missing or unknown ages (N/A), indicating a data quality concern that must be acknowledged.

```
NY_shootings_by_perp_age <- NY_shootings %>%
     group_by(PERP_AGE_GROUP) %>%
               summarise(cases = n()) %>%
               ungroup()

NY_shootings_by_perp_age_grouped <- NY_shootings_by_perp_age %>%
     mutate(
         PERP_AGE_GROUP = if_else(
              PERP_AGE_GROUP %in% c("18-24", "25-44", "45-64",
↪  "65+","<18"),PERP_AGE_GROUP,"N/A")) %>%
     group_by(PERP_AGE_GROUP) %>%
     summarise(cases = sum(cases)) %>%
     ungroup()

ggplot(NY_shootings_by_perp_age_grouped, aes(x = PERP_AGE_GROUP, y = cases,
↪  fill = PERP_AGE_GROUP)) +
     geom_bar(stat = "identity", show.legend = FALSE) +
     labs(title = "NYC Shooting Incidents by Perpetrator's Age Group",
          x = "Age group",
          y = "Number of shooting incidents") +
     theme_minimal() +
  theme(
        plot.title = element_text(hjust = 0.5, face = "bold", size = 16))
```

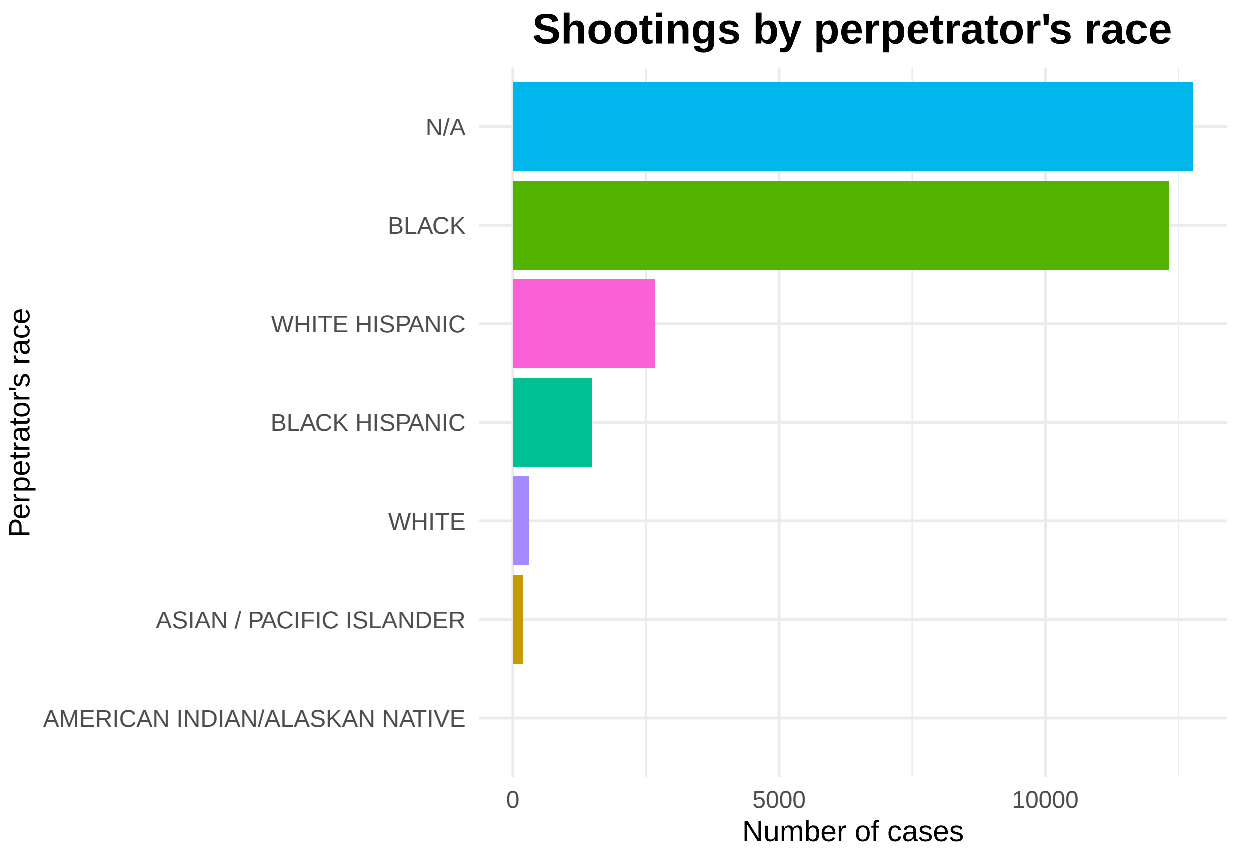


3.4.2 Race of the perpetrators

Second, race data indicates that black individuals are most frequently listed as perpetrators, followed by Black Hispanic, White Hispanic, and then White.

Again, N/A values are notably high and could bias interpretations, raising ethical concerns about over-representation and misclassification.

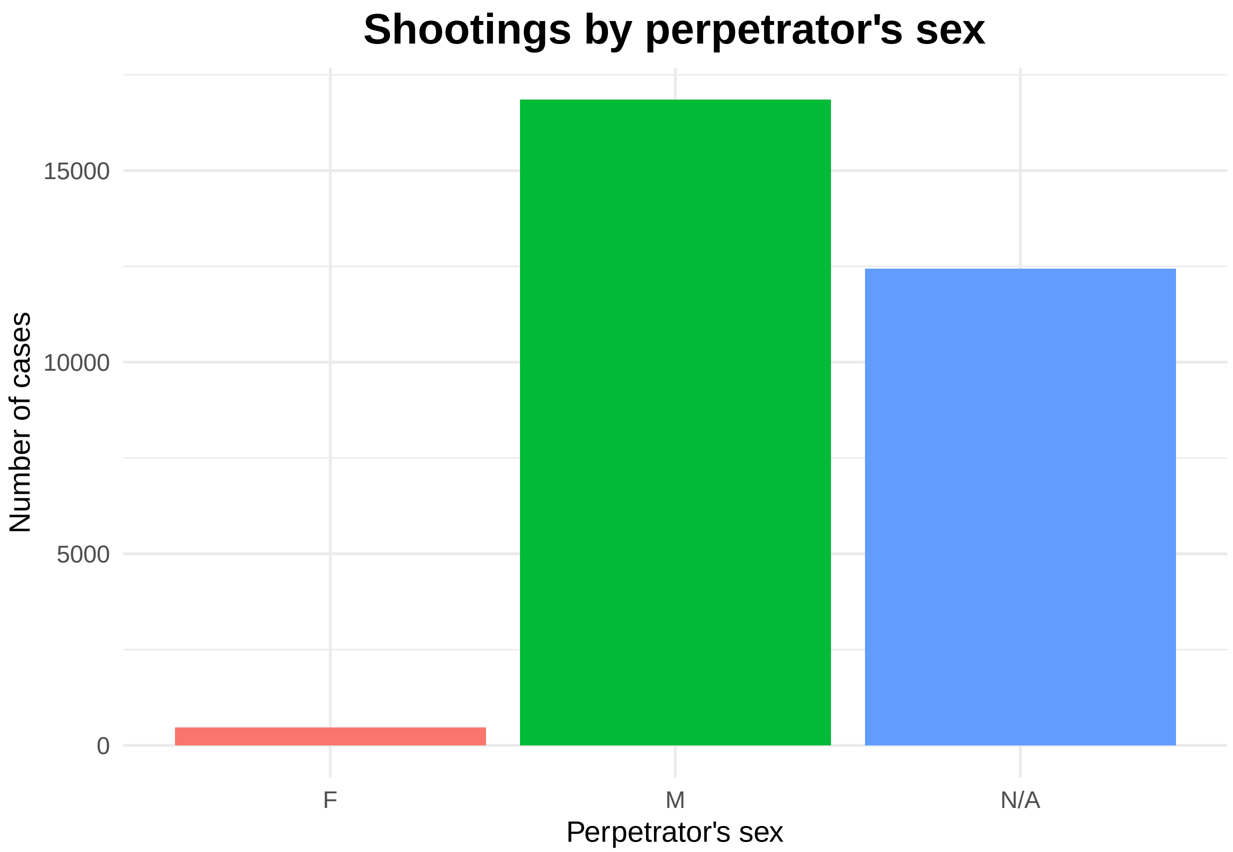
```
NY_shootings_perp_RACE <- NY_shootings %>%
          group_by(PERP_RACE) %>%
                    summarise(cases = n()) %>%
                    ungroup()


NY_shootings_perp_RACE_grouped <- NY_shootings_perp_RACE %>%
          mutate(
                  PERP_RACE = if_else(
                          PERP_RACE %in% c("AMERICAN INDIAN/ALASKAN
 ↪  NATIVE", "ASIAN / PACIFIC ISLANDER", "BLACK", "BLACK HISPANIC", "WHITE",
 ↪  "WHITE HISPANIC"),
                          PERP_RACE,
                          "N/A"
                      )
              ) %>%
          group_by(PERP_RACE) %>%
          summarise(cases = sum(cases)) %>%
          ungroup()


ggplot(NY_shootings_perp_RACE_grouped,
       aes(x = fct_reorder(PERP_RACE, cases), y = cases, fill = PERP_RACE))
        ↪  +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(title = "Shootings by perpetrator's race",
       x = "Perpetrator's race",
       y = "Number of cases") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16)
  ) +
  coord_flip()
```



3.4.3 Sex of the perpetrators

And third, sex data shows that over 16,000 shootings were committed by male individuals. Female perpetrators represent a tiny fraction, while a significant portion of cases are missing this data entirely.

```
NY_shootings_perp_sex <- NY_shootings %>%
               group_by(PERP_SEX) %>%
                         summarise(cases = n()) %>%
                         ungroup()


NY_shootings_perp_sex <- NY_shootings_perp_sex %>%
               mutate(
                           PERP_SEX = if_else(
                                        PERP_SEX %in% c("M", "F"),
                                        PERP_SEX,
                                        "N/A")) %>%
               group_by(PERP_SEX) %>%
               summarise(cases = sum(cases)) %>%
               ungroup()

ggplot(NY_shootings_perp_sex, aes(x = PERP_SEX, y = cases, fill =
 ↪  PERP_SEX)) +
          geom_bar(stat = "identity", show.legend = FALSE) +
          labs(title = "Shootings by perpetrator's sex",
                          x = "Perpetrator's sex",
                          y = "Number of cases") +
          theme_minimal() +
  theme(
        plot.title = element_text(hjust = 0.5, face = "bold", size = 16))
```



These demographic results, while statistically insightful, must be interpreted cautiously. They reflect patterns in arrests and investigations—not necessarily the actual distribution of criminal activity. Furthermore, systemic bias, underreporting, and socioeconomic context may skew these patterns.

4 Predictive or statistical model

To deepen our understanding of temporal shooting patterns in New York City, two statistical models were used to explore and describe the data: one focusing on the evolution of shooting incidents over time, and the other on the daily cycle of violence based on the hour of the day.

4.1 Linear Regression on Cumulative Shootings Over Time

The first graph illustrates the cumulative number of shooting incidents in NYC from 2006 to 2024. A linear regression model was fitted to this cumulative trend to determine whether the growth in total shootings follows a consistent pattern over the years.

Surprisingly, the linear fit aligns closely with the actual trajectory of the data, suggesting that — despite local fluctuations and temporary surges — the long-term accumulation of shootings has progressed at a relatively steady rate. Notably, the model maintains a strong fit even around the period of social disruption during 2020–2021, when shooting rates sharply increased. While a linear model is inherently simple, its performance in this context implies that the underlying drivers of gun violence may be relatively persistent over time.

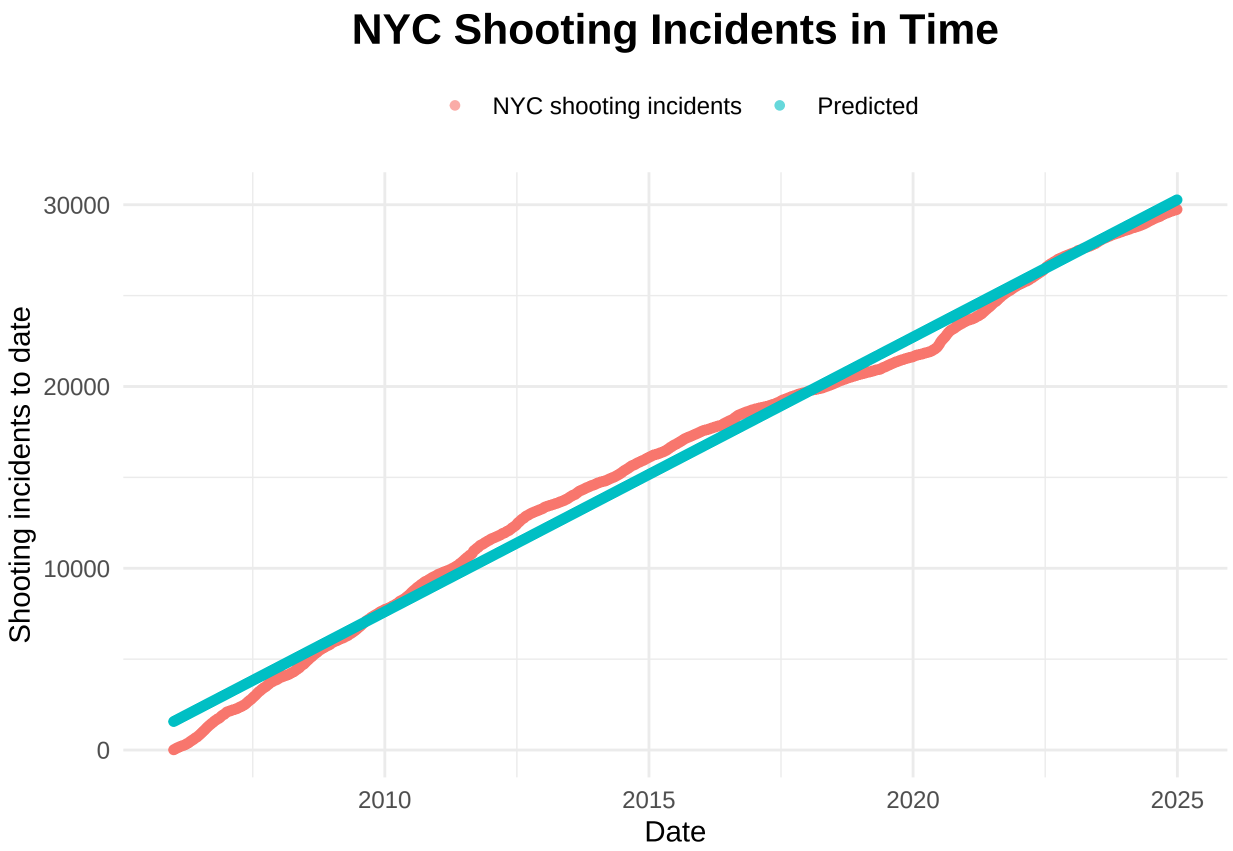
This model is particularly useful for forecasting future incidents under the assumption that historical patterns continue. However, this simplicity is also a limitation: the linear model is unable to capture short-term disruptions or structural shifts. For example, it may under- or overestimate periods of unusual crime behavior triggered by economic downturns, political protests, or public health crises. As such, the model should be interpreted as a baseline approximation of growth, not a complete picture of complex social dynamics.

```r
#create the linear model
mod <- lm(cumulative_cases ~ OCCUR_DATE, data = NY_shootings_by_date)

#append preds to the table
NY_shootings_by_date <- NY_shootings_by_date %>% mutate(pred =
 ↪  predict(mod))

#show summary of fitted model
summary(mod)
```

```
##
## Call:
## lm(formula = cumulative_cases ~ OCCUR_DATE, data = NY_shootings_by_date)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1606.23  -584.85   -40.61   684.73  1229.66
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.284e+04  7.626e+01  -692.9   <2e-16 ***
## OCCUR_DATE   4.137e+00  4.564e-03   906.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 739.4 on 6429 degrees of freedom
## Multiple R-squared:  0.9922, Adjusted R-squared:  0.9922
## F-statistic: 8.217e+05 on 1 and 6429 DF,  p-value: < 2.2e-16
```

```r
ggplot(NY_shootings_by_date, aes(x = OCCUR_DATE)) +
     geom_point(aes(y = cumulative_cases, color = "NYC shooting
      ↪  incidents"),
                size = 1.2, alpha = 0.6) +
     geom_point(aes(y = pred, color = "Predicted"),
                size = 1.2, alpha = 0.6) +
     labs(
         title = "NYC Shooting Incidents in Time",
         x = "Date",
         y = "Shooting incidents to date"
     ) +
     theme_minimal() +
     theme(
         plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
         legend.title = element_blank(),   #hide legend title
         legend.position = "top"
     )
```



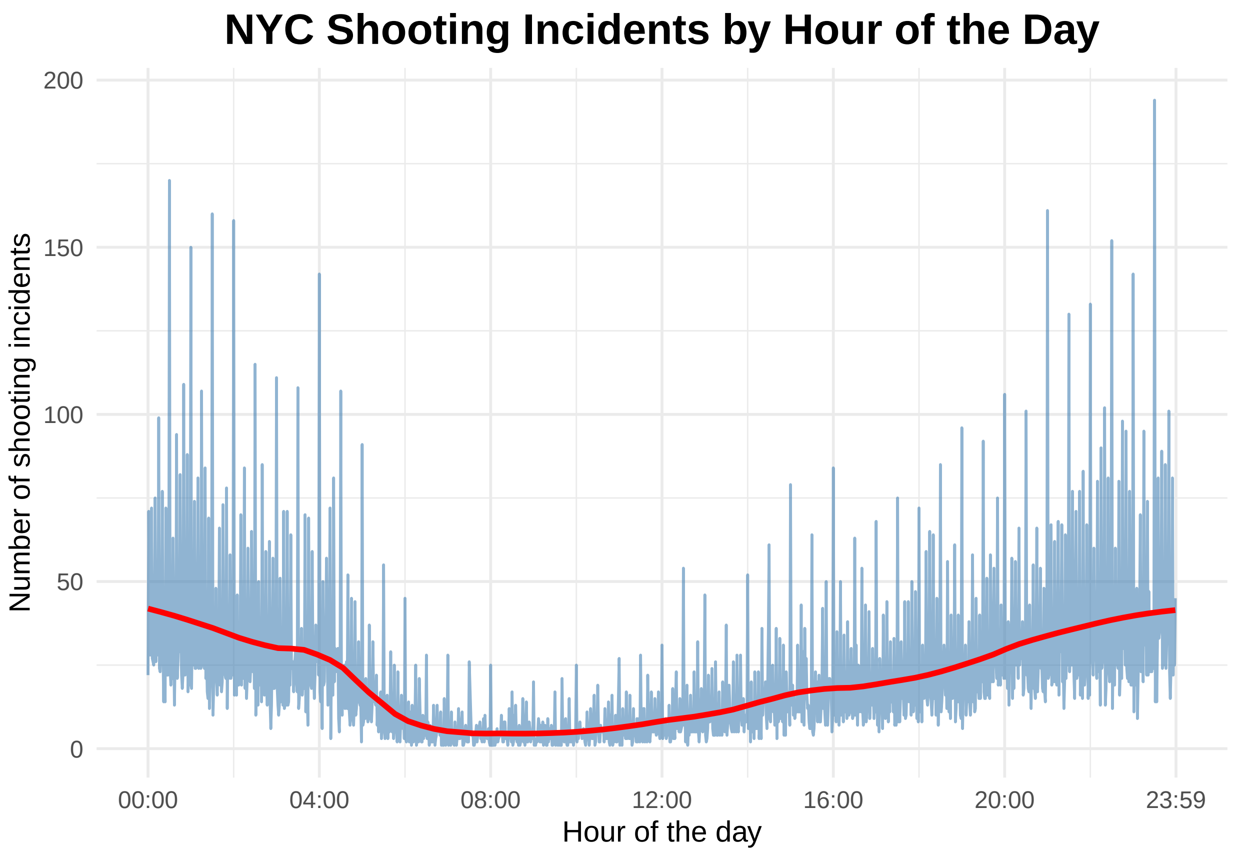
4.2 LOESS Smoothing of Hourly Shooting Frequencies

The second model focuses on daily behavioral patterns, visualizing how shooting incidents vary by the hour of the day. Here, a LOESS (Locally Estimated Scatterplot Smoothing) technique was applied — a non-parametric regression method that adapts to the local structure of the data without assuming any specific functional form.

The resulting red curve captures two prominent peaks: one shortly after midnight and another during the late evening hours. The early morning (around 6 to 9 AM) shows the lowest levels of gun violence. This bimodal pattern is consistent with sociological expectations: nightlife activity, social tensions, and alcohol consumption tend to increase during late hours, while most people are indoors and inactive during the morning.

LOESS is particularly well suited for this type of descriptive analysis, as it allows the analyst to uncover organic behavioral cycles hidden within noisy data. Unlike linear models, LOESS does not enforce a global trend, instead providing a flexible, localized view of the data. However, it should be noted that LOESS is not designed for prediction — its primary function is to highlight existing structure rather than forecast future outcomes.

```r
ggplot(NY_shootings_by_hour, aes(x = OCCUR_TIME, y = cases)) +
  geom_line(color = "steelblue", alpha = 0.6) +
  #LOESS model fitting
  geom_smooth(method = "loess", span = 0.2, color = "red", se = FALSE) +
  scale_x_time(
    limits = c(as_hms("00:00:00"), as_hms("23:59:59")),
    breaks = as_hms(c("00:00:00", "04:00:00", "08:00:00", "12:00:00",
     ↪  "16:00:00", "20:00:00", "23:59:59")),
    labels = time_format("%H:%M")
  ) +
  labs(
    title = "NYC Shooting Incidents by Hour of the Day",
    x = "Hour of the day",
    y = "Number of shooting incidents"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    legend.title = element_blank(),    #hide legend title
         legend.position = "top"
  )
```



4.3 Comparison and Purpose

Together, these two models serve complementary roles in the analysis. The linear model gives a macro-level view of the persistence of shootings across nearly two decades and offers predictive potential. In contrast, LOESS smoothing provides a micro-level look into the rhythms of urban violence within a single day, uncovering cyclical behaviors that are invisible in yearly summaries.

Both models are valuable tools for understanding the dataset, but they must be interpreted within their proper scope. Neither accounts for external variables like socioeconomic status, law enforcement activity, or policy changes — future work could incorporate these dimensions through multivariate models or time series decomposition.

5 Conclusions

The analysis of NYPD shooting incidents from 2006 to 2024 provides a multifaceted view into the temporal, spatial, and demographic dynamics of gun violence in New York City. Several key insights emerge from the data:

- **Temporal Trends:**
  The cumulative number of shootings has increased in a roughly linear fashion over the years, despite temporary decreases and sharp disruptions such as the 2020–2021 spike. This pattern suggests that the underlying social conditions fostering violence have remained consistently unresolved.

- **Time-of-Day Behavior:**
  Shootings follow a clear daily rhythm, with highest frequencies during late-night and early-evening hours. This highlights the social component of gun violence, often linked to nightlife, disputes, and street activity.

- **Geographic Disparities:**
  Gun violence is not evenly distributed across boroughs. Brooklyn and the Bronx bear a disproportionate burden of incidents, indicating the presence of long-standing structural inequalities that concentrate violence in specific areas.

- **Demographic Patterns:**
  Perpetrators are predominantly male and between the ages of 18 and 44. While racial data shows a majority of perpetrators identified as Black or Hispanic, a significant portion of records lack complete demographic data – a factor that introduces bias and limits conclusive interpretation.

- **Lethality:**
  Around 19.4% of shootings result in murder. This fatality rate underscores the urgency of prevention efforts, even as the majority of cases do not lead to death.

Overall, this project provides a data-driven foundation for understanding the scope and structure of gun violence in NYC. Future work could involve: - Integrating socioeconomic indicators, - Building predictive machine learning models, or - Conducting spatial correlation analyses with external datasets (e.g., income, police presence, housing policy).

Additionally, addressing missing and inconsistent data remains a crucial step in improving the reliability of future research.

6 Bias discussion

While the NYPD Shooting Incident dataset provides a valuable and detailed account of gun violence across New York City, several sources of bias must be acknowledged and critically assessed to interpret the results with the appropriate level of caution:

- **Missing and Incomplete Data:**
  A significant proportion of entries in demographic fields such as `PERP_AGE_GROUP`, `PERP_SEX`, and `PERP_RACE` are either missing, coded as `UNKNOWN`, or contain invalid entries. This limits the reliability of subgroup analysis and introduces bias, especially if the missingness is non-random and associated with case severity or location.
- **Underreporting and Classification Ambiguity:**
  Only shootings officially recorded by the NYPD are included, which excludes incidents that go unreported or misclassified. Inconsistencies in how "shooting incidents" are defined and recorded over time or across precincts can create temporal and geographic biases in the dataset.
- **Racial and Socioeconomic Bias in Policing:**
  Law enforcement practices influence what gets recorded. Heavily policed neighborhoods may show higher incident counts not necessarily due to more violence, but due to higher detection. This **detection bias** can overrepresent lower-income or minority communities.
- **Binary Categorization of Outcomes:**
  The variable `STATISTICAL_MURDER_FLAG` simplifies the outcome of incidents to a binary classification (murder or not), omitting crucial context such as intent, injury severity, or medical outcomes. This simplification can distort the understanding of lethality patterns.
- **Interpretation Bias in Aggregated Data:**
  Aggregating incidents by year, borough, or demographic group can obscure local patterns or context-specific spikes (e.g., gang activity, policy changes, public events). This may lead to misleading conclusions if not paired with contextual analysis.

Overall, while the dataset provides a rich foundation for analysis, it reflects both the realities of urban gun violence and the limitations of institutional data collection. Careful interpretation, combined with supplementary data and awareness of potential blind spots, is essential for producing fair and insightful conclusions.