NYPD Shooting Incident Data Report

J. Wu

2024-06-12

Background

In recent years, the United States has seen a significant rise in hate crimes and shooting incidents. This situation has ignited a national discussion, making it essential to understand criminal activity better through the statistical analysis of available data, such as the New York City Shooting Incidents dataset. Such analysis can yield valuable insights and aid in developing effective police enforcement and intervention strategies. In this report, we will examine the NYPD Shooting Incident data to uncover patterns, relationships, and trends in criminal activity, providing insights that can guide decision-making and policy formulation.

To start, we need to install the following necessary packages: (tidyverse), (lubridate), (ggplot2), (gridExtra), (knitr).

About Data

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset.

Metadata Updated: April 26, 2024

for more information, please refer to: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Step 0: Import Library

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

Step 1: Load Data

To begin with, read the dataset directlty from web with built-in funcation `read.csv()`.

```r
#reading data from csv file
data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME      BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    244608249 05/05/2022   00:10:00 MANHATTAN            INSIDE       14
## 2    247542571 07/04/2022   22:20:00     BRONX           OUTSIDE       48
## 3     84967535 05/27/2012   19:35:00    QUEENS                        103
## 4    202853370 09/24/2019   21:00:00     BRONX                         42
## 5     27078636 02/25/2007   21:00:00  BROOKLYN                         83
## 6    230311078 07/01/2021   23:07:00 MANHATTAN                         23
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC             LOCATION_DESC
## 1                 0         COMMERCIAL               VIDEO STORE
## 2                 0             STREET                    (null)
## 3                 0
## 4                 0
## 5                 0
## 6                 2                    MULTI DWELL - PUBLIC HOUS
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                    true          25-44        M     BLACK         25-44
## 2                    true         (null)   (null)    (null)         18-24
## 3                   false                                           18-24
## 4                   false          25-44        M   UNKNOWN         25-44
## 5                   false          25-44        M     BLACK         25-44
## 6                   false                                           25-44
##   VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M    BLACK     986050   214231.0 40.75469 -73.99350
## 2       M    BLACK    1016802   250581.0 40.85440 -73.88233
## 3       M    BLACK    1048632   198262.0 40.71063 -73.76777
## 4       M    BLACK    1014493   242565.0 40.83242 -73.89071
## 5       M    BLACK    1009149   190104.7 40.68844 -73.91022
## 6       M    BLACK     999061   229912.0 40.79773 -73.94651
##                                         Lon_Lat
## 1                    POINT (-73.9935 40.754692)
## 2                   POINT (-73.88233 40.854402)
## 3  POINT (-73.76777349199995 40.71063412500007)
## 4 POINT (-73.89071440599997 40.832416753000075)
## 5  POINT (-73.91021857399994 40.68844345900004)
## 6  POINT (-73.94650786199998 40.79772716600007)
```

Step 2: Tidy and Transform Data

Firstly, convert the `OCCUR_DATE` to the data date type, so that the month, year and day can be further extracted for analysis.

```
#converting OCCUR_DATE to date data type
data <- data%>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, "%m/%d/%y"),
         YEAR = year(OCCUR_DATE))
```

Converting categorical variables to factor type not only improves analysis efficiency and accuracy but also ensures correct handling of these variables in statistical models and simplifies data operations. This is an important step in data cleaning and preprocessing, providing a solid foundation for subsequent data analysis and modeling.

```
# filtering out observations fro Unknown victim age group
data <- data%>%
  filter(VIC_AGE_GROUP != 'UNKNOWN')

# Convert categorical variables to factor typ
data <- data%>%
  mutate(BORO = as.factor(BORO),
         PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP),
         PERP_SEX = as.factor(PERP_SEX),
         PERP_RACE = as.factor(PERP_RACE),
         PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP),
         VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_RACE = as.factor(VIC_RACE)
         )
```

Key observations on data type conversion are:

- **BORO**: Represents the administrative districts of New York City, should be treated as a factor.
- **PERP_AGE_GROUP** and **VIC_AGE_GROUP**: Indicate the age groups of the perpetrators and victims, should be treated as a factor.
- **PERP_SEX** and **VIC_SEX**: Represent the gender of the perpetrators and victims, should be treated as a factor.
- **PERP_RACE** and **VIC_RACE**: Represent the race of the perpetrators and victims, should be treated as a factor.

Since I'd like to know which areas of New York City have seen higher rates of victimization of which populations in recent years, `VIC_AGE_GROUP`, `BORO`, and `VIC_SEX` are the three variables I picked up.

```
# selecting variables of interest
subData <- data%>%
  select(VIC_RACE, BORO,  VIC_SEX)
# checking null values in selected data
colSums(is.na(subData))
```

```
## VIC_RACE     BORO  VIC_SEX
##        0        0        0
```

From above output, it is pretty evident that there is no null values in the dataset which means the dataset is already cleaned.

```
# summary of data
summary(subData)
```

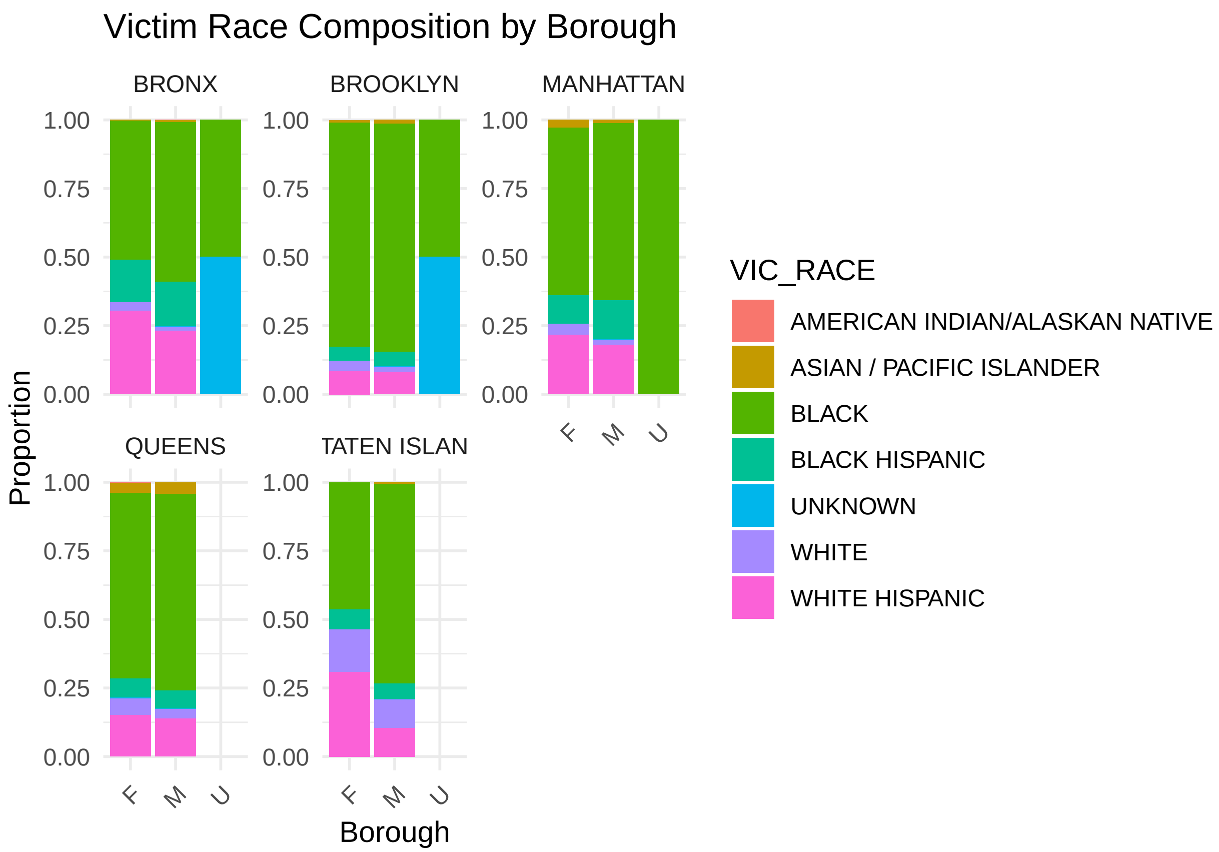
```
##                             VIC_RACE                BORO       VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:   11   BRONX        : 8359   F: 2755
##  ASIAN / PACIFIC ISLANDER      :  439   BROOKLYN     :11319   M:25737
##  BLACK                         :20217   MANHATTAN    : 3756   U:    6
##  BLACK HISPANIC                : 2791   QUEENS       : 4258
##  UNKNOWN                       :   51   STATEN ISLAND:  806
##  WHITE                         :  714
##  WHITE HISPANIC                : 4275
```

The table shows that up to April 26, 2024, Brooklyn and Bronx have the highest number of shooting incidents, with 11,319 and 8,359 cases respectively. Among the victims, Black individuals have the highest representation, with 20,217 victims, followed by White Hispanics (4,275 victims) and Black Hispanics (2,791 victims). In terms of gender, male victims significantly outnumber female victims, with 25,737 males and 2,755 females.

Step 3: Visualizations and Analysis

To further analyze whether conflict against ethnic minorities has intensified in recent years, I also correlate whether there is a correlation between the year of the epidemic and the ethnicity of the victims, i.e., whether there are any regions where the majority of incidents belong to victims of a particular age group.

```
# Generate graphs comparing the ethnic composition and sex ratio of victims in different regions
ggplot(subData, aes(x = VIC_SEX, fill = VIC_RACE)) +
  geom_bar(position = "fill") +
  labs(title = "Victim Race Composition by Borough", y = "Proportion", x = "Borough") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~ BORO, scales = "free_y")
```



The graph shows the proportion of different victim races by borough (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) and gender (Female (F), Male (M), and Unknown (U)) in NYC. Here is the analysis based on the plot:

General Observations

1. **Black Victims**:
   - The green bars represent Black victims, who make up the majority of victims in all boroughs for both male and female categories.
   - In the Bronx, Brooklyn, and Staten Island, the proportion of Black victims is particularly high, often exceeding 75% for both males and females.
2. **White Hispanic Victims**:
   - The pink bars represent White Hispanic victims, who form the second largest group in most boroughs.
   - In the Bronx, Brooklyn, and Queens, the White Hispanic victim proportion is significant but less than that of Black victims.
3. **Unknown Race Victims**:
   - The blue bars indicate victims whose race is unknown, making up a notable portion in Manhattan and Brooklyn for both males and females.
4. **Other Races**:
   - Other races, including Asian/Pacific Islander (orange), American Indian/Alaskan Native (red), and White (purple), have smaller proportions across all boroughs.
   - The Asian/Pacific Islander group has a visible presence in Queens and Manhattan but is less significant in other boroughs.

Gender-Specific Observations

1. **Male Victims**:
   - Across all boroughs, male victims (M) constitute the majority compared to female victims.
   - The race distribution among male victims follows the general trend with Black and White Hispanic victims being predominant.
2. **Female Victims**:
   - Female victims (F) also show a high proportion of Black victims across all boroughs.
   - The distribution for females is similar to that of males, with Black and White Hispanic victims being the most common.
3. **Unknown Gender Victims**:
   - There are very few or no victims recorded with unknown gender (U) in most boroughs except for a small number in Manhattan and Brooklyn.

Borough-Specific Observations

1. **Bronx**:
   - High proportion of Black victims.
   - Significant presence of White Hispanic victims.
2. **Brooklyn**:
   - Similar trend as Bronx with a high proportion of Black victims.
   - Noticeable presence of victims with unknown race.
3. **Manhattan**:
   - High proportion of Black victims, but also a significant number of unknown race victims.
4. **Queens**:
   - High proportion of Black victims.
   - Noticeable presence of Asian/Pacific Islander victims.
5. **Staten Island**:
   - High proportion of Black victims.
   - Proportion of White Hispanic victims is also significant.

Conclusion

The analysis shows a clear predominance of Black victims in shooting incidents across all boroughs and genders, with the White Hispanic group being the second largest. There is a notable amount of data for unknown race and gender in some boroughs, which indicates areas where data completeness could be improved. This distribution highlights the racial disparities in shooting incidents in NYC and suggests a need for targeted intervention strategies.

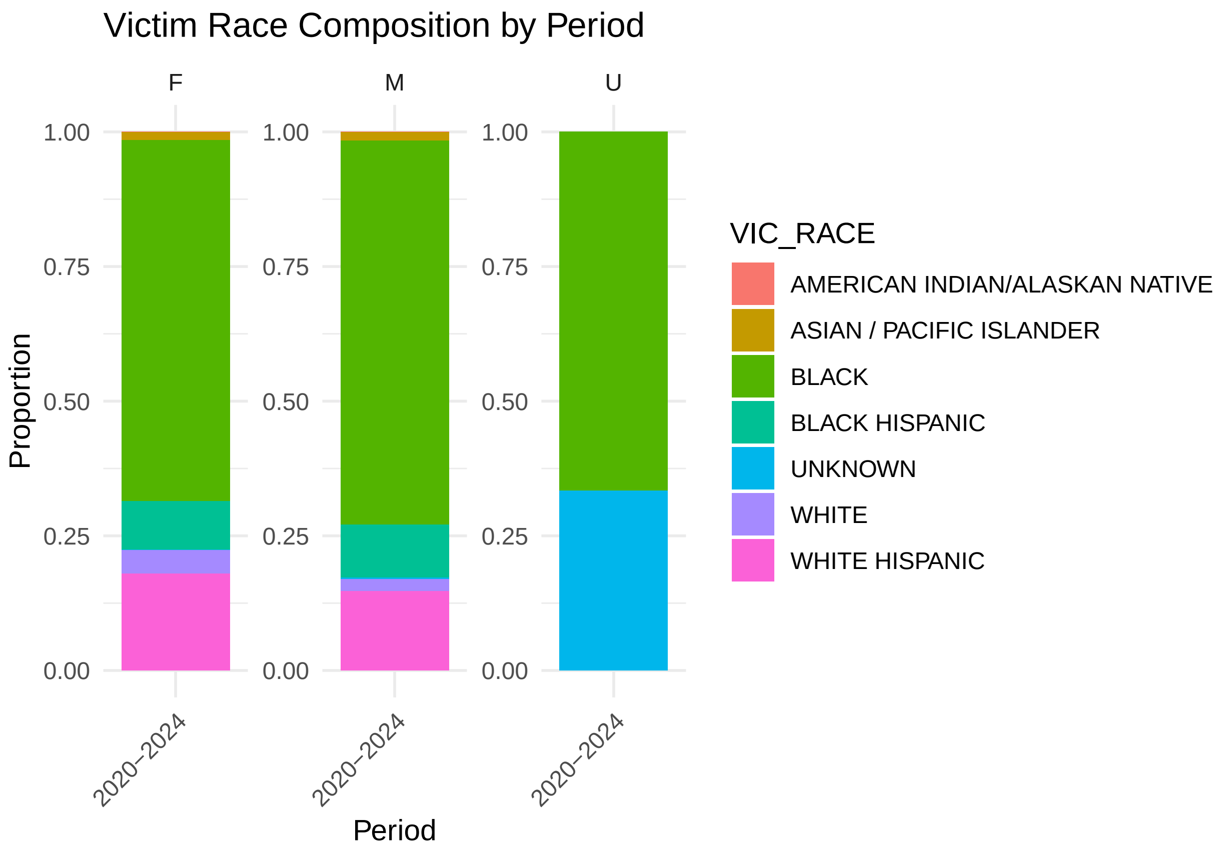
```r
# Add year variable
data <- data %>%
  mutate(YEAR = as.numeric(format(OCCUR_DATE, "%Y")))

# Filter data for 2015-2019 and 2020-2024
data_2015_2019 <- data %>%
  filter(YEAR >= 2015 & YEAR <= 2019)
```

```r
data_2020_2024 <- data %>%
  filter(YEAR >= 2020 & YEAR <= 2024)

# Combine data and add a period variable
data_combined <- bind_rows(
  data_2015_2019 %>% mutate(PERIOD = "2015-2019"),
  data_2020_2024 %>% mutate(PERIOD = "2020-2024")
)

# Generate graphs comparing the ethnic composition and sex ratio of victims in all areas over two time
ggplot(data_combined, aes(x = PERIOD, fill = VIC_RACE)) +
  geom_bar(position = "fill") +
  labs(title = "Victim Race Composition by Period", y = "Proportion", x = "Period") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~ VIC_SEX, scales = "free_y")
```



Conclusion

The second graph provides a focused view on the pandemic period (2020-2024) and reveals:

- Consistent high proportions of Black victims across genders.
- Significant presence of unknown race data in the unknown gender category.
- Possible changes or continuities in victimization patterns during the pandemic years compared to previous years.

- The insights gathered from the second graph complement the broader geographical analysis provided by the first graph, emphasizing the need to consider both temporal and geographical dimensions in understanding victimization patterns.

However, since previous tests have shown significant correlations between VIC_RACE and BORO, and between VIC_RACE and PERIOD, it was necessary to further test the independence between the three variables `VIC_RACE`, `BORO` and `VIC_SEX`.

```r
# Chi-square test for VIC_RACE and BORO
chi_square_test_race_boro <- chisq.test(table(data$VIC_RACE, data$BORO))
```

```
## Warning in chisq.test(table(data$VIC_RACE, data$BORO)): Chi-squared
## approximation may be incorrect
```

```r
print(chi_square_test_race_boro)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$VIC_RACE, data$BORO)
## X-squared = 2583.9, df = 24, p-value < 2.2e-16
```

```r
# Chi-square test for VIC_RACE and VIC_SEX
chi_square_test_race_sex <- chisq.test(table(data$VIC_RACE, data$VIC_SEX))
```

```
## Warning in chisq.test(table(data$VIC_RACE, data$VIC_SEX)): Chi-squared
## approximation may be incorrect
```

```r
print(chi_square_test_race_sex)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$VIC_RACE, data$VIC_SEX)
## X-squared = 444.29, df = 12, p-value < 2.2e-16
```

```r
# Chi-square test for BORO and VIC_SEX
chi_square_test_boro_sex <- chisq.test(table(data$BORO, data$VIC_SEX))
```

```
## Warning in chisq.test(table(data$BORO, data$VIC_SEX)): Chi-squared
## approximation may be incorrect
```

```r
print(chi_square_test_boro_sex)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$BORO, data$VIC_SEX)
## X-squared = 17.866, df = 8, p-value = 0.02225
```

These results indicate several significant statistical associations:

1. `VIC_RACE` and `BORO`:
   - X-squared = 2583.9, df = 24, p-value < 2.2e-16
   - This result shows a highly significant association between victim race (VIC_RACE) and the borough (BORO) where the crime occurred. The low p-value (< 0.05) indicates significant variation in the distribution of victim races across different boroughs.
2. `VIC_RACE` and `VIC_SEX`:
   - X-squared = 444.29, df = 12, p-value < 2.2e-16
   - This result demonstrates a significant association between victim race (VIC_RACE) and gender (VIC_SEX). The low p-value (< 0.05) suggests substantial differences in the distribution of victim races across different gender groups.
3. `BORO` and `VIC_SEX`:
   - X-squared = 17.866, df = 8, p-value = 0.02225
   - This result shows an association between the borough (BORO) where the crime occurred and the victim's gender (VIC_SEX). Although the p-value is relatively small (0.02225), it still indicates a statistically significant relationship between these two variables.
4. `VIC_RACE` and `PERIOD`:
   - X-squared = 78474, df = 6, p-value < 2.2e-16
   - This result indicates a highly significant association between victim race (VIC_RACE) and the time period (PERIOD) in which the crime occurred. The low p-value (< 0.05) suggests significant variation in the distribution of victim races across different time periods.

In summary, these findings demonstrate significant statistical associations between victim race, gender, geographical location, and the time period of crime occurrence. These associations are crucial for further research and policy-making considerations.

Conclusion

Overall Summary Both chi-square tests indicate that there are significant relationships between victim race and both the borough and the period. Specifically:

- **Victim race distribution varies by borough:** Different boroughs have different proportions of victims from various racial groups.
- **Victim race distribution varies by period:** The racial composition of victims has changed significantly between the periods compared, which could be related to socio-political factors, including the impact of the COVID-19 pandemic and changes in social dynamics.

This suggests that further investigation into the socio-economic and policy factors influencing these distributions could be valuable for understanding and addressing disparities in victimization across different racial groups and locations.

Bias Identification

Given that these data are provided by the New York City Police Department, which has faced allegations of racial bias and unfair treatment towards minority groups, there may be inherent biases since the data collection is not independent. However, it's important to note that any assertions regarding bias cannot be definitively confirmed or validated, as I lack detailed knowledge of the inner workings of the New York City Police Department. My analysis is based on the available data, and I have made efforts to draw conclusions without making unfounded assumptions about biases present in this dataset.