NYPD Shooting Incident Data Report

2024-03-20

Introduction

This report presents an analysis of shooting incidents recorded by the NYPD from 2006 to the most recent full year. The dataset at hand is a comprehensive account detailing the locations, dates, and profiles of individuals involved in each incident.

The analysis is centered around a pivotal question: What are the factors that determine the outcome of a shooting incident, especially in terms of fatalities, in relation to the victims' demographics? By focusing on this question, we aim to uncover the elements that may affect the gravity of these incidents, thereby contributing to a deeper understanding of urban violence and its consequences.

Data Handling

Importing Libraries

```
library(tidyverse)
## -- Attaching core tidyverse packages -----
                                                       ----- tidyverse 2.0.0 --
## v dplyr
             1.1.4
                        v readr
                                     2.1.5
## v forcats
              1.0.0
                                     1.5.1
                        v stringr
## v ggplot2 3.5.0
                                     3.2.1
                        v tibble
                                     1.3.1
## v lubridate 1.9.3
                        v tidyr
## v purrr
              1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()
                    masks stats::lag()
## i Use the conflicted package (<a href="http://conflicted.r-lib.org/">http://conflicted.r-lib.org/</a>) to force all conflicts to become error
library(dplyr)
library(lubridate)
```

Importing Data

library(ggplot2)

```
# read the dataset from the url
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
df <- read.csv(url)
summary(df)</pre>
```

```
##
     INCIDENT_KEY
                           OCCUR_DATE
                                               OCCUR_TIME
                                                                       BORO
##
    Min.
            : 9953245
                         Length: 27312
                                              Length: 27312
                                                                   Length: 27312
                          Class : character
                                              Class : character
##
    1st Qu.: 63860880
                                                                   Class : character
    Median: 90372218
                         Mode :character
                                              Mode :character
                                                                  Mode :character
##
##
    Mean
            :120860536
    3rd Qu.:188810230
##
##
    Max.
            :261190187
##
##
    LOC_OF_OCCUR_DESC
                            PRECINCT
                                           JURISDICTION_CODE LOC_CLASSFCTN_DESC
                                                              Length: 27312
##
    Length: 27312
                        Min.
                                : 1.00
                                           Min.
                                                  :0.0000
##
    Class : character
                        1st Qu.: 44.00
                                           1st Qu.:0.0000
                                                              Class : character
##
                        Median: 68.00
                                           Median : 0.0000
                                                              Mode : character
    Mode :character
                                                  :0.3269
##
                        Mean
                                : 65.64
                                           Mean
                        3rd Qu.: 81.00
                                           3rd Qu.:0.0000
##
##
                                :123.00
                                           Max.
                                                  :2.0000
                        Max.
##
                                           NA's
                                                  :2
    LOCATION_DESC
                        STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##
##
    Length: 27312
                        Length: 27312
                                                  Length: 27312
    Class : character
                        Class : character
##
                                                  Class : character
##
    Mode :character
                        Mode :character
                                                  Mode : character
##
##
##
##
##
      PERP SEX
                         PERP RACE
                                             VIC_AGE_GROUP
                                                                    VIC SEX
##
    Length: 27312
                        Length: 27312
                                             Length: 27312
                                                                 Length: 27312
##
    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
                                                                      :40.51
##
    Length: 27312
                                : 914928
                                                   :125757
                        Min.
                                            Min.
                                                              Min.
##
    Class : character
                        1st Qu.:1000028
                                            1st Qu.:182834
                                                              1st Qu.:40.67
    Mode :character
##
                        Median :1007731
                                            Median :194487
                                                              Median :40.70
##
                        Mean
                                :1009449
                                            Mean
                                                   :208127
                                                              Mean
                                                                      :40.74
##
                        3rd Qu.:1016838
                                            3rd Qu.:239518
                                                              3rd Qu.:40.82
##
                        Max.
                                :1066815
                                                    :271128
                                                              Max.
                                                                      :40.91
                                            Max.
##
                                                              NA's
                                                                      :10
      Longitude
##
                        Lon Lat
           :-74.25
                      Length: 27312
##
    Min.
    1st Qu.:-73.94
##
                      Class : character
    Median :-73.92
                      Mode :character
##
            :-73.91
##
    Mean
##
    3rd Qu.:-73.88
##
    Max.
            :-73.70
    NA's
##
            :10
```

Feature Descriptions

The dataset provided contains detailed records of shooting incidents in New York City. Here's a breakdown of what each column represents:

- INCIDENT_KEY: A unique identifier for each shooting incident.
- OCCUR_DATE: The date on which the shooting incident occurred.
- OCCUR_TIME: The time at which the shooting incident occurred.
- BORO: The borough in New York City where the shooting incident took place.
- LOC_OF_OCCUR_DESC: A description of the location where the shooting occurred, which may include specific details like the type of building or area.
- PRECINCT: The NYPD precinct number that has jurisdiction over the area where the shooting occurred.
- **JURISDICTION_CODE**: A code representing the type of jurisdiction at the location of the incident (e.g., NYPD, federal).
- LOC_CLASSFCTN_DESC: A classification description of the location, which could provide additional context about the setting of the incident.
- LOCATION_DESC: A more detailed description of the location where the shooting occurred.
- STATISTICAL_MURDER_FLAG: A boolean flag indicating whether the shooting incident was classified as a murder for statistical purposes.
- PERP_AGE_GROUP: The age group of the perpetrator(s).
- **PERP_SEX**: The sex of the perpetrator(s).
- **PERP_RACE**: The race of the perpetrator(s).
- VIC_AGE_GROUP: The age group of the victim(s).
- VIC_SEX: The sex of the victim(s).
- VIC_RACE: The race of the victim(s).
- X_COORD_CD and Y_COORD_CD: The X and Y coordinates of the incident location, based on the New York State Plane Coordinate System.
- Latitude and Longitude: The geographical coordinates of the incident location.
- Lon Lat: A combined field of the longitude and latitude coordinates in a POINT format.

Feature Selection

```
# drop unnecessary columns
df <- select(df, -c(INCIDENT_KEY, OCCUR_TIME, JURISDICTION_CODE, LOCATION_DESC, PERP_AGE_GROUP, PERP_SE
```

Type Casting

```
# convert OCCUR_DATE to date type
df <- df %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))

# convert categorical columns to factor type
df$BORO <- factor(df$BORO)
df$VIC_AGE_GROUP <- factor(df$VIC_AGE_GROUP)
df$VIC_SEX <- factor(df$VIC_SEX)
df$VIC_RACE <- factor(df$VIC_RACE)
df$STATISTICAL_MURDER_FLAG <- factor(df$STATISTICAL_MURDER_FLAG)</pre>
```

Missing Values

```
# display missing values
colSums(is.na(df))
```

```
##
                 OCCUR_DATE
                                                   BORO
                                                               LOC_OF_OCCUR_DESC
##
                           0
                                                      0
##
                    PRECINCT
                                   LOC CLASSFCTN DESC STATISTICAL MURDER FLAG
##
                           Λ
              VIC_AGE_GROUP
##
                                                VIC SEX
                                                                         VIC_RACE
                                                                                 0
##
##
                    Latitude
                                             Longitude
##
                           10
                                                     10
# remove rows with missing values
df <- na.omit(df)</pre>
```

Descriptive Statistics

```
# show descriptive statistics of the cleaned data summary(df)
```

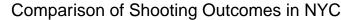
```
PRECINCT
##
      OCCUR DATE
                                      BORO
                                                  LOC_OF_OCCUR_DESC
##
    Min.
            :2006-01-01
                           BRONX
                                         : 7937
                                                  Length: 27302
                                                                       Min.
                                                                               : 1.00
##
                           BROOKLYN
    1st Qu.:2009-07-18
                                         :10929
                                                  Class : character
                                                                       1st Qu.: 44.00
##
    Median :2013-04-27
                           MANHATTAN
                                         : 3568
                                                  Mode : character
                                                                       Median: 68.00
##
    Mean
            :2014-01-05
                           QUEENS
                                          4092
                                                                       Mean
                                                                               : 65.64
    3rd Qu.:2018-10-08
                           STATEN ISLAND:
                                           776
                                                                       3rd Qu.: 81.00
##
##
    Max.
            :2022-12-31
                                                                       Max.
                                                                               :123.00
##
##
    LOC_CLASSFCTN_DESC_STATISTICAL_MURDER_FLAG_VIC_AGE_GROUP
                                                                    VIC_SEX
##
    Length: 27302
                        false:22036
                                                  <18
                                                          : 2839
                                                                   F: 2613
##
    Class : character
                        true: 5266
                                                  1022
                                                                    M:24678
##
    Mode :character
                                                  18-24
                                                          :10081
                                                                    U:
                                                                         11
                                                  25-44
                                                          :12276
##
                                                  45-64
                                                          : 1863
##
##
                                                  65+
                                                             181
##
                                                              61
                                                  UNKNOWN:
                                VIC_RACE
##
                                                 Latitude
                                                                 Longitude
##
    AMERICAN INDIAN/ALASKAN NATIVE:
                                                      :40.51
                                                                       :-74.25
                                         10
                                              Min.
                                                               Min.
##
    ASIAN / PACIFIC ISLANDER
                                        404
                                              1st Qu.:40.67
                                                               1st Qu.:-73.94
##
    BLACK
                                     :19433
                                              Median :40.70
                                                               Median :-73.92
##
    BLACK HISPANIC
                                      2644
                                              Mean
                                                      :40.74
                                                               Mean
                                                                       :-73.91
##
    UNKNOWN
                                         66
                                              3rd Qu.:40.82
                                                               3rd Qu.:-73.88
##
    WHITE
                                        698
                                                      :40.91
                                                                       :-73.70
                                              Max.
                                                               Max.
    WHITE HISPANIC
                                     : 4047
```

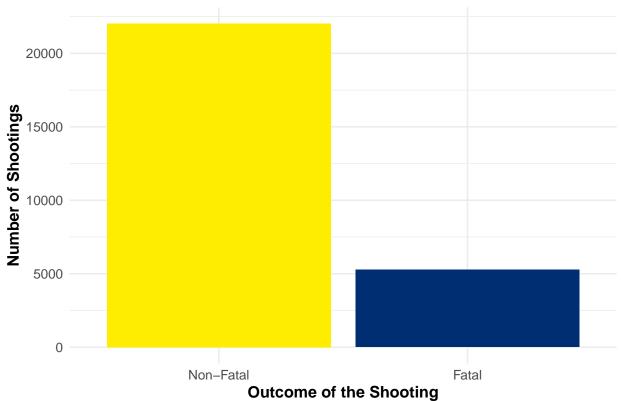
Data Visualization

In the forthcoming section, we will present a series of visual analyses focusing on the outcomes of shooting incidents, categorized as either fatal or non-fatal. These visualizations are designed to elucidate the relationship between the outcome of the incident and various demographic factors pertaining to the victims, such as *sex*, *age group* and *race*. By employing graphical representations, we aim to highlight any discernible patterns or trends that emerge from the data, providing a clearer understanding of how these factors may correlate with the likelihood of fatality in shooting incidents. This visual exploration is intended to support

a deeper comprehension of the dynamics at play in such critical events and to aid in the development of informed strategies for prevention and intervention.

```
df %>%
  ggplot(aes(x = STATISTICAL_MURDER_FLAG, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(show.legend = FALSE) +
  scale_fill_manual(values = c("false" = "#ffed00", "true" = "#002e72")) +
  scale_x_discrete(labels = c("false" = "Non-Fatal", "true" = "Fatal")) +
  labs(
    title = "Comparison of Shooting Outcomes in NYC",
    x = "Outcome of the Shooting",
    y = "Number of Shootings"
) +
  theme_minimal() +
  theme(
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    plot.title = element_text(hjust = 0.5)
)
```





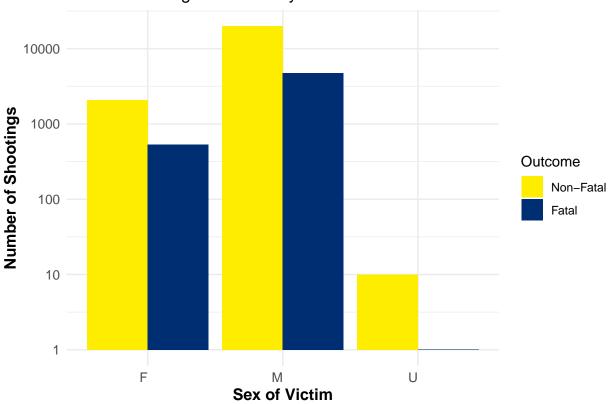
It is evident that the majority of the dataset, approximately 76.1%, consists of non-fatal shootings.

```
df %>%
  ggplot(aes(x = VIC_SEX, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("false" = "#ffed00", "true" = "#002e72"),
```

```
labels = c("false" = "Non-Fatal", "true" = "Fatal")) +

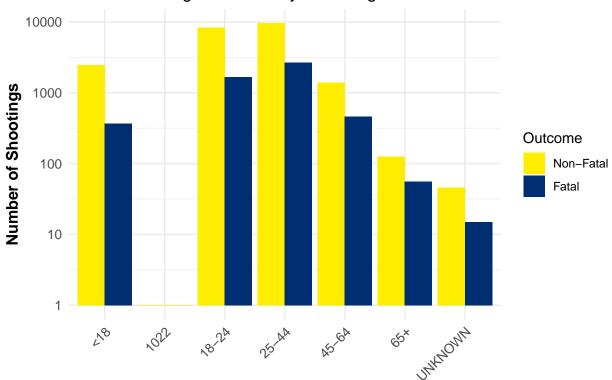
labs(
   title = "Shooting Outcomes by Victim Sex in NYC",
   x = "Sex of Victim",
   y = "Number of Shootings",
   fill = "Outcome"
) +
theme_minimal() +
theme(
   axis.title = element_text(size = 12, face = "bold"),
   axis.text.x = element_text(hjust = 1),
   axis.text = element_text(size = 10),
   plot.title = element_text(hjust = 0.5)
) +
scale_y_log10()
```

Shooting Outcomes by Victim Sex in NYC



```
fill = "Outcome"
) +
theme minimal() +
theme(
  axis.title = element_text(size = 12, face = "bold"),
  axis.text.x = element_text(angle = 45, hjust = 1),
 axis.text = element_text(size = 10),
 plot.title = element_text(hjust = 0.5)
) +
scale_y_log10()
```

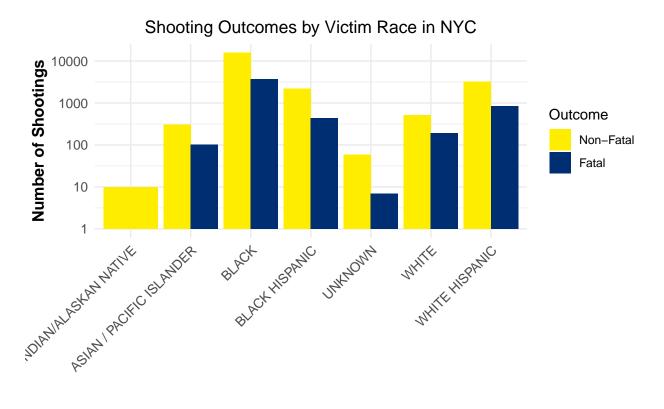
Shooting Outcomes by Victim Age in NYC



Age of Victim

```
df %>%
  ggplot(aes(x = VIC_RACE, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("false" = "#ffed00", "true" = "#002e72"),
                    labels = c("false" = "Non-Fatal", "true" = "Fatal")) +
  labs(
   title = "Shooting Outcomes by Victim Race in NYC",
    x = "Race of Victim",
    y = "Number of Shootings",
    fill = "Outcome"
  theme_minimal() +
  theme(
    axis.title = element_text(size = 12, face = "bold"),
```

```
axis.text.x = element_text(angle = 45, hjust = 1),
axis.text = element_text(size = 10),
plot.title = element_text(hjust = 0.5)
) +
scale_y_log10()
```



Race of Victim

The analysis of the data reveals several key insights into the nature of shooting incidents in New York City. Firstly, non-fatal shootings significantly outnumber fatal ones, indicating that while gun violence is prevalent, it does not always result in loss of life. This could be attributed to prompt medical care or the circumstances of the shootings themselves.

When considering the *sex* of the victims, it is evident that males are more frequently involved in both fatal and non-fatal shootings than females. This suggests that males may be more likely to be in situations where shootings occur or may be more frequently targeted.

Age also plays a role in shooting incidents, with a noticeable concentration among young adults, particularly those in the 18-25 age bracket. This age group's heightened involvement might reflect social or economic factors that increase their exposure to violent environments.

Lastly, the *racial* analysis of victims shows that Black and Hispanic individuals are disproportionately affected by gun violence, facing higher rates of both fatal and non-fatal shootings. This disparity points towards systemic issues that may influence the risk of gun violence faced by these communities.

Overall, the data underscores the need for interventions that address the specific risks associated with different demographics to effectively reduce and prevent shootings in the city.

Modeling

To predict the outcome of fatality in shooting incidents, we will construct a logistic regression model. Logistic regression is suitable for binary outcomes, such as fatal vs. non-fatal shootings. The model will estimate the probability of a shooting being fatal based on various predictors, which could include the victim's demographics and other relevant factors. By analyzing the coefficients produced by the model, we can interpret the influence of each predictor on the likelihood of a fatal outcome.

```
# convert the outcome variable to factor
df$STATISTICAL_MURDER_FLAG <- factor(df$STATISTICAL_MURDER_FLAG, levels = c("false", "true"), labels =
# build the logistic regression model
model <- glm(STATISTICAL_MURDER_FLAG ~ VIC_SEX + VIC_AGE_GROUP + VIC_RACE, data = df, family = "binomia
# display summary of the model to view coefficients and statistics
summary(model)
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_SEX + VIC_AGE_GROUP +
       VIC_RACE, family = "binomial", data = df)
##
## Coefficients:
##
                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)
                                    -12.86401 102.15907 -0.126 0.89979
## VIC_SEXM
                                                         -0.930 0.35244
                                     -0.04841
                                                 0.05207
## VIC_SEXU
                                     -0.59029
                                                 1.08279
                                                         -0.545 0.58565
## VIC_AGE_GROUP1022
                                              324.74370
                                                         -0.033 0.97382
                                    -10.65719
## VIC_AGE_GROUP18-24
                                     0.28620
                                                 0.06197
                                                          4.618 3.87e-06 ***
## VIC_AGE_GROUP25-44
                                      0.61318
                                                 0.06006
                                                         10.210 < 2e-16 ***
## VIC_AGE_GROUP45-64
                                     0.75941
                                                 0.07781
                                                          9.760 < 2e-16 ***
## VIC_AGE_GROUP65+
                                     1.01916
                                                 0.17146
                                                          5.944 2.78e-09 ***
## VIC_AGE_GROUPUNKNOWN
                                     0.87559
                                                 0.31660
                                                           2.766 0.00568 **
## VIC_RACEASIAN / PACIFIC ISLANDER 11.28117 102.15912
                                                          0.110 0.91207
## VIC_RACEBLACK
                                     11.00355 102.15905
                                                          0.108 0.91423
## VIC_RACEBLACK HISPANIC
                                     10.82285 102.15906
                                                           0.106 0.91563
## VIC_RACEUNKNOWN
                                     10.25893 102.15991
                                                           0.100 0.92001
## VIC_RACEWHITE
                                     11.34238
                                              102.15909
                                                           0.111 0.91160
## VIC_RACEWHITE HISPANIC
                                     11.12498
                                              102.15906
                                                           0.109 0.91328
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
       Null deviance: 26776
##
                            on 27301
                                      degrees of freedom
## Residual deviance: 26500
                            on 27287
                                      degrees of freedom
## AIC: 26530
##
## Number of Fisher Scoring iterations: 11
```

The logistic regression model output indicates that the victim's age group is a significant predictor of the fatality outcome in shooting incidents. Specifically, individuals in the age groups 18-24, 25-44, 45-64, and 65+ have increasingly higher odds of a shooting being fatal compared to the reference group, which is likely

under 18. The coefficients for these age groups are positive and statistically significant, suggesting a clear trend where the likelihood of a shooting being fatal increases with the victim's age. The sex and race of the victim, however, do not appear to be significant predictors in this model, as their coefficients are not statistically significant.

Conclusions

The analysis suggests that age is a crucial factor in the fatality of shootings. The increasing coefficients for the older age groups imply that as victims get older, the probability of a shooting resulting in death rises. This could be due to a variety of reasons, including potential delays in receiving medical care, the severity of injuries, or pre-existing health conditions that make recovery less likely. The lack of significance in sex and race coefficients may indicate that, once controlling for age, these factors do not independently predict the fatality of shootings. However, it's important to consider the broader context and other variables not included in the model that could interact with these demographics.

Possible Sources of Bias

Several sources of bias could affect the results of this model. First, there may be selection bias in the data collection process, as not all shooting incidents are reported or recorded with equal diligence. Second, there could be measurement bias if the categorization of age, sex, or race is inaccurate or inconsistent. Additionally, omitted variable bias is a concern, as other factors not included in the model, such as the location of the incident, the type of firearm used, or the response time of emergency services, could influence the outcome. Lastly, there may be survivorship bias, as the data only includes those who survived long enough to be reported as a victim. Each of these biases could skew the results and should be carefully considered when interpreting the findings.