NYPD Shooting Incident Data (Historic)

2024-04-15

```
#install.packages("tidyverse")
#install.packages("tidyverse", dependencies = TRUE)
#update.packages(ask = FALSE)
library(tidyverse)
library(lubridate)
```

Objective Statement

The primary goals of this project are to import, tidy, and analyze the NYPD Shooting Incident dataset. Key components of this project include:

- **Data Cleaning**: Ensuring the dataset is free of errors and formatted correctly for analysis.
- **Data Visualization**: Provide at least two visualizations that reveal trends, patterns, or insights in the data.
- **Statistical Modeling**: Develop at least one model to analyze or predict patterns within the data.
- **Bias Identification**: Discuss potential biases in the dataset or analysis process and their implications.

The following sections will address these objectives through detailed data exploration, visualization, and modeling.

```
# Data
csv_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

# read the csv
shootings <- read_csv(csv_url)
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(shootings)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
```

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

```
tail(shootings)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    270719378 07/02/2023 21:40      BRONX     OUTSIDE                 46
## 2    265354835 03/19/2023 23:48      BRONX     INSIDE                  47
## 3    272968931 08/16/2023 02:46      BRONX     OUTSIDE                 41
## 4    270489846 06/27/2023 12:27      BRONX     INSIDE                  41
## 5    271021661 07/08/2023 11:27      QUEENS    OUTSIDE                102
## 6    271818283 07/24/2023 23:38      MANHATTAN OUTSIDE                 28
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

```
summary(shootings)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms         Class :character
##  Median : 92711254   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
```

```
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Inspecting the Data

Now that we loaded in our CSV file, we need to inspect the data, to better understand what we are working with.

Data Cleaning

We now need to clean our data. We will do this by converting our data types, and by dropping any NA/null data, and any redundant rows/columns. Lets get to it.

```r
# Data Conversions
shootings <- shootings %>%
  mutate(
    OCCUR_DATE = mdy(OCCUR_DATE),
    OCCUR_TIME = hms(OCCUR_TIME),
    BORO = factor(BORO),
    LOC_OF_OCCUR_DESC = factor(LOC_OF_OCCUR_DESC),
    PERP_SEX = factor(PERP_SEX),
    VIC_SEX = factor(VIC_SEX),
    PERP_RACE = factor(PERP_RACE),
    VIC_RACE = factor(VIC_RACE)
  )

#Missing Data
```

```r
shootings <- shootings %>%
  mutate(
    Latitude = if_else(is.na(Latitude), median(Latitude, na.rm = TRUE), Latitude),
Longitude = if_else(is.na(Longitude), median(Longitude, na.rm = TRUE), Longitude)
  ) %>%
drop_na(JURISDICTION_CODE)

shootings <- select(shootings, -Lon_Lat)

#Lets  check the data again, now that we addressed some of the concerns.
summary(shootings)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 65439914   1st Qu.:2009-09-04   1st Qu.:3H 30M 0S
##  Median : 92711254   Median :2013-09-20   Median :15H 15M 0S
##  Mean   :127406776   Mean   :2014-06-07   Mean   :12H 44M 19.4390756302528S
##  3rd Qu.:203162840   3rd Qu.:2019-09-30   3rd Qu.:20H 45M 0S
##  Max.   :279758069   Max.   :2023-12-29   Max.   :23H 59M 0S
##
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT     JURISDICTION_CODE
##  BRONX        : 8376   INSIDE :  460     Min.   :  1.0   Min.   :0.0000
##  BROOKLYN     :11346   OUTSIDE: 2506     1st Qu.: 44.0   1st Qu.:0.0000
##  MANHATTAN    : 3761   NA's   :25594     Median : 67.0   Median :0.0000
##  QUEENS       : 4270                     Mean   : 65.5   Mean   :0.3219
##  STATEN ISLAND:  807                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                                          Max.   :123.0   Max.   :2.0000
##
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:28560       Length:28560       Mode :logical
##  Class :character   Class :character   FALSE:23034
##  Mode  :character   Mode  :character   TRUE :5526
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX              PERP_RACE     VIC_AGE_GROUP
##  Length:28560       (null): 1141   BLACK          :11902   Length:28560
##  Class :character   F     :  444   WHITE HISPANIC: 2509   Class :character
##  Mode  :character   M     :16166   UNKNOWN        : 1837   Mode  :character
##                     U     : 1499   BLACK HISPANIC: 1392
##                     NA's  : 9310   (null)         : 1141
##                                    (Other)        :  469
##                                    NA's           : 9310
##  VIC_SEX                              VIC_RACE       X_COORD_CD
##  F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11   Min.   : 914928
##  M:25788   ASIAN / PACIFIC ISLANDER      :  440   1st Qu.:1000068
##  U:   12   BLACK                         :20234   Median :1007772
##            BLACK HISPANIC                : 2795   Mean   :1009425
##            UNKNOWN                       :   70   3rd Qu.:1016807
##            WHITE                         :  728   Max.   :1066815
##            WHITE HISPANIC                : 4282
##    Y_COORD_CD        Latitude       Longitude
```

```
##  Min.   :125757   Min.   :40.51   Min.   :-74.25
##  1st Qu.:182910   1st Qu.:40.67   1st Qu.:-73.94
##  Median :194901   Median :40.70   Median :-73.92
##  Mean   :208380   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:239814   3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :271128   Max.   :40.91   Max.   :-73.70
##
```

Addressing the changes

Data cleaning is a critical step in the data analysis process because it ensures the accuracy, completeness, and quality of the data. By cleaning the data, we ensure that it is correctly formatted and ready for analysis, which helps in making the analysis process more efficient and the results more reliable. The `OCCUR_DATE` was converted from a character string to a Date object. This allows for ore accurate data handleing/processing.

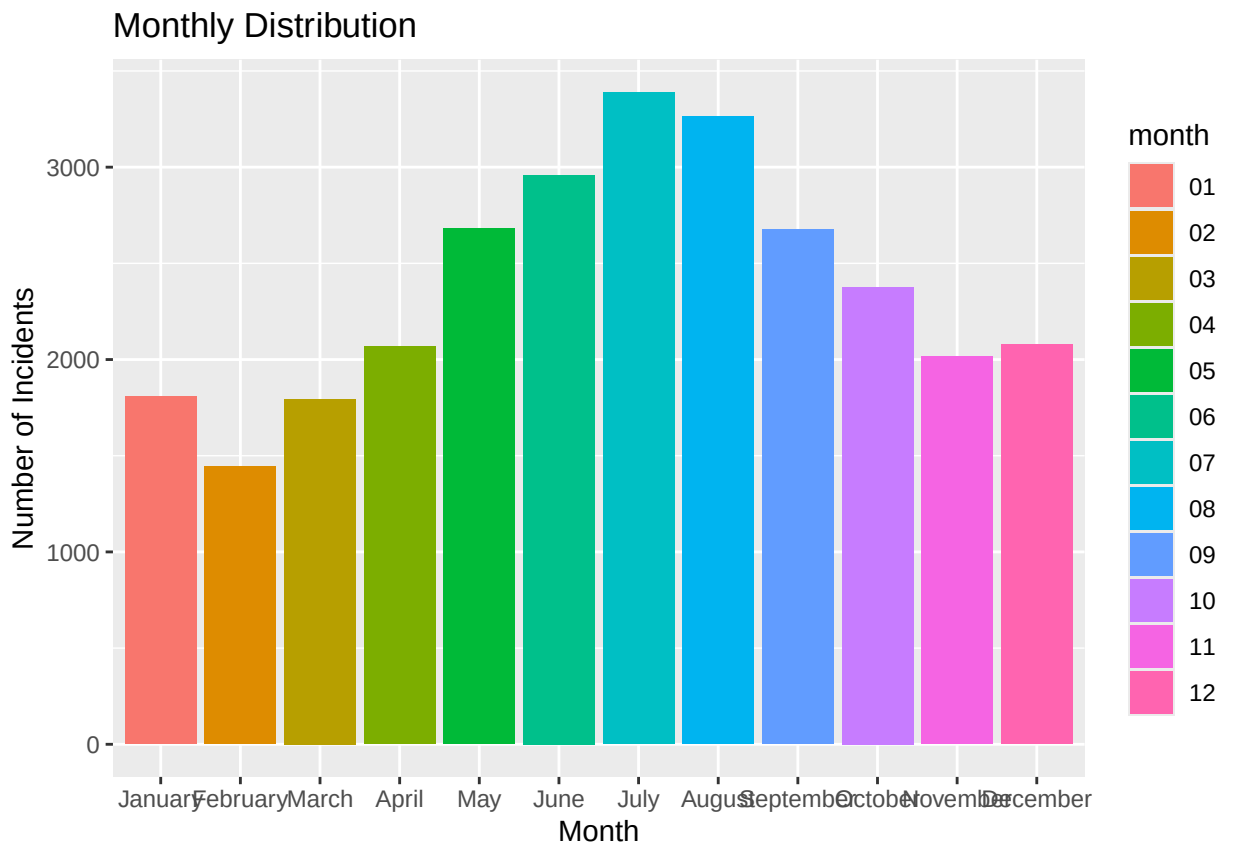
Data Analysis

To better understand the dynamics of shooting incidents, we will explore the following:

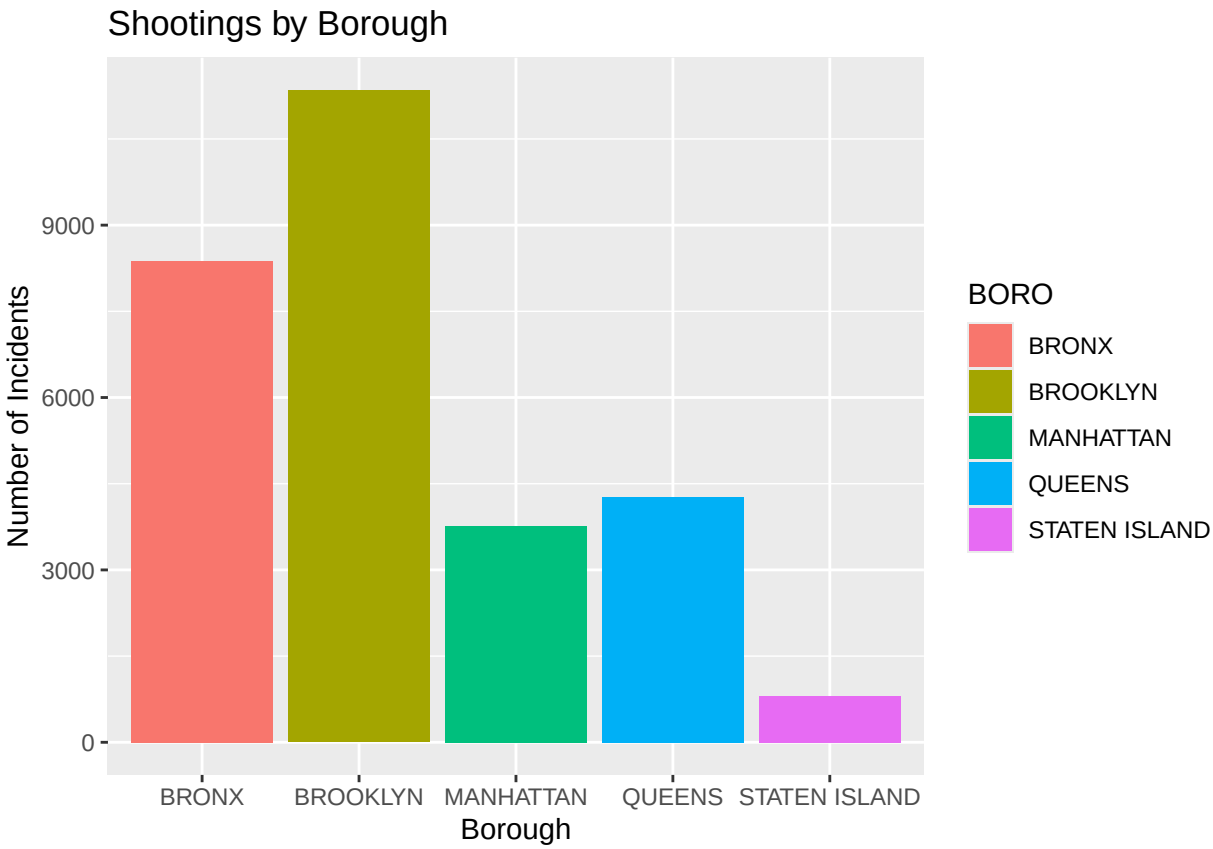
- **Seasonal Analysis**: How does time of year impact shooting data?
- **Borough Analysis**: Are some boroughs more prone to shootings than others?

```
# Extract months
shootings$month <- format(as.Date(shootings$OCCUR_DATE), "%m")

# Lets track our incidents by month in a barplot
ggplot(shootings, aes(x = month, fill = month)) +
    geom_bar() +
    scale_x_discrete(labels = month.name) +
    labs(title = "Monthly Distribution", x = "Month", y = "Number of Incidents")
```




```r
#Lets analyze shootings by Borough.
ggplot(shootings, aes(x = BORO, fill = BORO)) +
  geom_bar() +
  labs(title = "Shootings by Borough", x = "Borough", y = "Number of Incidents")
```



Interpreting the Graphs

As we can see from our first graph: Monthly Distribution, the number of incidents peaks during the summer months, with the highest rates occuring in July, and the lowest rates occurring in February. We further break down our data in our second visual: Shootings by Borough, where we see Brooklyn having the highest number of incidents, and Staten Island having the lowest number of incidents.

```
# Convert STATISTICAL_MURDER_FLAG to a binary numeric variable
shootings$STATISTICAL_MURDER_FLAG <- as.numeric(shootings$STATISTICAL_MURDER_FLAG)

# Lets fit our model
model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + month + PRECINCT + PERP_RACE, data = shootings, family = b

summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + month + PRECINCT +
##     PERP_RACE, family = binomial(), data = shootings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0744  -0.7056  -0.6531  -0.3691   2.4013
##
## Coefficients:
```

```
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -1.6417367  0.1760025  -9.328  < 2e-16
## BOROBROOKLYN                            -0.1403700  0.1009987  -1.390  0.16458
## BOROMANHATTAN                           -0.1419345  0.0855379  -1.659  0.09705
## BOROQUEENS                              -0.1587765  0.1986175  -0.799  0.42405
## BOROSTATEN ISLAND                       -0.1406878  0.2519845  -0.558  0.57663
## month02                                  0.0481894  0.1017114   0.474  0.63565
## month03                                 -0.1264014  0.0985382  -1.283  0.19957
## month04                                 -0.0122714  0.0938629  -0.131  0.89598
## month05                                  0.0743564  0.0887598   0.838  0.40218
## month06                                 -0.1565600  0.0903045  -1.734  0.08297
## month07                                 -0.1551232  0.0883061  -1.757  0.07898
## month08                                 -0.1730728  0.0895007  -1.934  0.05314
## month09                                  0.0463337  0.0902907   0.513  0.60784
## month10                                 -0.1018154  0.0943803  -1.079  0.28069
## month11                                 -0.0615631  0.0976703  -0.630  0.52849
## month12                                  0.1276686  0.0954905   1.337  0.18123
## PRECINCT                                 0.0004381  0.0030462   0.144  0.88565
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -8.9815836 84.3933720  -0.106  0.91524
## PERP_RACEASIAN / PACIFIC ISLANDER        0.9219742  0.1883185   4.896 9.79e-07
## PERP_RACEBLACK                           0.4402550  0.0869241   5.065 4.09e-07
## PERP_RACEBLACK HISPANIC                  0.3510646  0.1079082   3.253  0.00114
## PERP_RACEUNKNOWN                        -0.8961031  0.1263202  -7.094 1.30e-12
## PERP_RACEWHITE                           1.2452997  0.1475590   8.439  < 2e-16
## PERP_RACEWHITE HISPANIC                  0.6177552  0.0963519   6.411 1.44e-10
##
## (Intercept)                             ***
## BOROBROOKLYN
## BOROMANHATTAN                           .
## BOROQUEENS
## BOROSTATEN ISLAND
## month02
## month03
## month04
## month05
## month06                                 .
## month07                                 .
## month08                                 .
## month09
## month10
## month11
## month12
## PRECINCT
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER       ***
## PERP_RACEBLACK                          ***
## PERP_RACEBLACK HISPANIC                 **
## PERP_RACEUNKNOWN                        ***
## PERP_RACEWHITE                          ***
## PERP_RACEWHITE HISPANIC                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 19229  on 19249  degrees of freedom
## Residual deviance: 18804  on 19226  degrees of freedom
##   (9310 observations deleted due to missingness)
## AIC: 18852
##
## Number of Fisher Scoring iterations: 9
```

Analyzing Model Outcomes

The results indicate that perpetrator race significantly affects the likelihood of a shooting being fatal. These insights emphasize the complex interplay of demographic and temporal factors in the dynamics of shooting incidents.

Effects of Predictive Variables

The model's coefficients offer insights into the potential fatality of a shooting, while controlling for the influence of other variables in the model.

- **Intercept (`-1.6417367`)**: This value indicates the baseline log odds of a fatality in shootings when all other variables are at their reference levels. The negative coefficient suggests a generally low probability of fatality under these conditions.
- **Borough**:
  - **Brooklyn (`-0.1403700`)**: This borough shows a slightly lower likelihood of fatalities compared to the Bronx, although the difference is not statistically significant.
  - **Manhattan (`-0.1419345`)**: Similarly, shootings in Manhattan are marginally less likely to be fatal compared to those in the Bronx, but this effect does not reach statistical significance.
  - **Queens (`-0.1587765`)**: Shows a trend of lower fatality rates compared to the Bronx, though this is also not statistically significant.
  - **Staten Island (`-0.1406878`)**: Like the other boroughs mentioned, Staten Island demonstrates a negative coefficient, suggesting a lower probability of shooting fatalities compared to the Bronx, albeit not statistically significant.
- **Month**:
  - The coefficients for months such as June, July, and August are slightly negative but only marginally significant, hinting at possible seasonal influences on shooting fatalities.
- **Perpetrator Race**:
  - **Asian / Pacific Islander (`0.9219742`)** and **White (`1.2452997`)**: Shootings involving perpetrators from these racial groups show a significantly higher probability of being fatal.
  - **Black (`0.4402550`)** and **White Hispanic (`0.6177552`)**: These groups also demonstrate an increased likelihood of fatality, with positive coefficients that are statistically significant.
  - **Unknown (`-0.8961031`)**: This category shows a significantly lower likelihood of fatality, suggesting that shootings involving perpetrators of unknown racial background are less likely to be fatal.

Potential Sources of Bias

With any dataset, there are potential sources of bias that must be considered. In this analysis, several key biases could affect the reliability and interpretation of our findings:

Data Reliability

The accuracy and completeness of the data are foundational to any analysis. Errors in data collection, entry, or processing can lead to unreliable results.

Reporting Biases

Variations in how incidents are documented by different officers or precincts can introduce inconsistencies. Factors such as the perceived severity of an incident or the demographics of the individuals involved may influence how, or even if, details are reported.