NYPD Shooting Incident Report

L. Baldridge

2025-05-21

1. Introduction

This document showcases all the necessary steps within the data science process that exemplifies reproducibility. The report will be conducted on the `NYPD Shooting Incident Data (Historic)` dataset, taken from https://catalog.data.gov/dataset.

1.1 *Dataset Information*

This dataset contains every **recorded shooting incident** that occurred in **New York City** starting from **January 1st, 2006** to the last day of the previous year, **December 31st, 2024**. Each row represents a separate incident and includes information such as the date, time, borough where the incident occurred, coordinates, along with age, race, sex, and age group details for both perpetrator and victim.

2. Importing Libraries and Dataset

2.1 *Importing Tidyverse and Glue*

```r
library(tidyverse)
library(glue)
```

2.2 *Importing NYPD Shooting Dataset*

The dataset was downloaded from the link above and a copy was stored locally and subsequently imported into this R markdown document for further analysis.

```r
nypd_shooting_data <- readr::read_csv("Data\\NYPD_Shooting_Incident_Data__Historic_.csv")
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): 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.
```

3. Exploratory Data Analysis

3.1 *First Look at the Data*

First we transform our csv into a **dataframe** so that when we print it, *we can see entries for all available columns* and not just the first few. This will give us a better idea of what information is contained in each column.

```r
nypd_shooting_data <- as.data.frame(nypd_shooting_data)
print(head(nypd_shooting_data))
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    231974218 08/09/2021   01:06:00    BRONX              <NA>       40
## 2    177934247 04/07/2018   19:48:00 BROOKLYN              <NA>       79
## 3    255028563 12/02/2022   22:57:00    BRONX           OUTSIDE       47
## 4     25384540 11/19/2006   01:50:00 BROOKLYN              <NA>       66
## 5     72616285 05/09/2010   01:58:00    BRONX              <NA>       46
## 6     85875439 07/22/2012   21:35:00    BRONX              <NA>       42
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC             LOCATION_DESC
## 1                 0               <NA>                      <NA>
## 2                 0               <NA>                      <NA>
## 3                 0             STREET            GROCERY/BODEGA
## 4                 0               <NA>                 PVT HOUSE
## 5                 0               <NA>   MULTI DWELL - APT BUILD
## 6                 2               <NA> MULTI DWELL - PUBLIC HOUS
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX      PERP_RACE VIC_AGE_GROUP
## 1                   FALSE           <NA>     <NA>           <NA>         18-24
## 2                    TRUE          25-44        M WHITE HISPANIC         25-44
## 3                   FALSE         (null)   (null)         (null)         25-44
## 4                    TRUE        UNKNOWN        U        UNKNOWN         18-24
## 5                    TRUE          25-44        M          BLACK           <18
## 6                   FALSE          18-24        M          BLACK         18-24
##   VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M    BLACK  1006343.0   234270.0 40.80967 -73.92019
## 2       M    BLACK  1000082.9   189064.7 40.68561 -73.94291
## 3       M    BLACK  1020691.0   257125.0 40.87235 -73.86823
## 4       M    BLACK   985107.3   173349.8 40.64249 -73.99691
## 5       F    BLACK  1009853.5   247502.6 40.84598 -73.90746
## 6       M    BLACK  1011046.7   239814.2 40.82488 -73.90318
##                                          Lon_Lat
## 1  POINT (-73.92019278899994 40.80967347200004)
## 2 POINT (-73.94291302299996 40.685609672000055)
## 3                  POINT (-73.868233 40.872349)
## 4 POINT (-73.99691224999998 40.642489932000046)
## 5  POINT (-73.90746098599993 40.84598358900007)
## 6  POINT (-73.90317908399999 40.82487781900005)
```

3.2 *Displaying the Summary of the Dataset*

```r
summary(nypd_shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:29744       Length:29744      Length:29744
##  1st Qu.: 67321140   Class :character   Class1:hms        Class :character
##  Median :109291972   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :133850951                      Mode  :numeric
##  3rd Qu.:214741917
##  Max.   :299462478
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:29744       Min.   :  1.00   Min.   :0.0000    Length:29744
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.00   Median :0.0000    Mode  :character
##                     Mean   : 65.23   Mean   :0.3181
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:29744       Mode :logical           Length:29744
##  Class :character   FALSE:23979             Class :character
##  Mode  :character   TRUE :5765              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:29744       Length:29744       Length:29744       Length:29744
##  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:29744       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000094   1st Qu.:183042   1st Qu.:40.67
##  Mode  :character   Median :1007826   Median :195506   Median :40.70
##                     Mean   :1009442   Mean   :208722   Mean   :40.74
##                     3rd Qu.:1016739   3rd Qu.:239980   3rd Qu.:40.83
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :97
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:29744
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :97
```

3.3 *Checking for Missing Data*

```r
colSums(is.na(nypd_shooting_data))
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                       0                   25596                       0
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                       2                   25596                   14977
## STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP                PERP_SEX
##                       0                    9344                    9310
##               PERP_RACE           VIC_AGE_GROUP                 VIC_SEX
##                    9310                       0                       0
##                VIC_RACE              X_COORD_CD              Y_COORD_CD
##                       0                       0                       0
##                Latitude               Longitude                 Lon_Lat
##                      97                      97                      97
```

4. Tidying and Transforming Data

Displaying the first couple of rows shows us that we can filter down our columns as some of the information seems redundant or unnecessary. By checking for missing data, we can also exclude columns that contain a high amount of NAs.

And lastly from our summary we can see that there are multiple columns such as the date column that needs to be transformed into a different type, and others that may be better suited to be transformed into categorical variables.

4.1 *Filtering and Transforming Data*

First we remove all the columns that we do not need. Afterwards, we transform OCCUR_DATE into a date column, INCIDENT_KEY into a character, and BORO as a factor.

```r
nypd_shooting_data_filtered <- nypd_shooting_data %>%
    dplyr::select(-c(LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat, JURISDICTI
    dplyr::mutate(
        OCCUR_DATE = lubridate::mdy(OCCUR_DATE),
        INCIDENT_KEY = as.character(INCIDENT_KEY),
        BORO = as.factor(BORO),
    )
```

4.2 *Identifying Unique Values*

By removing LOC_OF_OCCUR_DESC and LOC_CLASSFCTN_DESC, we have removed the columns with the most missing data, but we still have other columns that contain a substantial amount of NAs such as the location and perpetrator description columns. Since these columns contain important information, we want to find a way to keep them. In order to get some idea on these missing values, we will display all the unique values for most of the remaining columns.

```r
unique_values <- lapply(nypd_shooting_data_filtered[4:12], unique)
print(unique_values)
```

```
## $BORO
## [1] BRONX         BROOKLYN      MANHATTAN     QUEENS        STATEN ISLAND
## Levels: BRONX BROOKLYN MANHATTAN QUEENS STATEN ISLAND
##
## $LOCATION_DESC
##  [1] NA                          "GROCERY/BODEGA"
##  [3] "PVT HOUSE"                 "MULTI DWELL - APT BUILD"
##  [5] "MULTI DWELL - PUBLIC HOUS" "(null)"
##  [7] "BAR/NIGHT CLUB"            "COMMERCIAL BLDG"
##  [9] "FAST FOOD"                 "HOSPITAL"
## [11] "BEAUTY/NAIL SALON"         "LIQUOR STORE"
## [13] "CHAIN STORE"               "RESTAURANT/DINER"
## [15] "SMALL MERCHANT"            "GAS STATION"
## [17] "JEWELRY STORE"             "GYM/FITNESS FACILITY"
## [19] "STORE UNCLASSIFIED"        "SOCIAL CLUB/POLICY LOCATI"
## [21] "DRY CLEANER/LAUNDRY"       "NONE"
## [23] "VIDEO STORE"               "SUPERMARKET"
## [25] "VARIETY STORE"             "FACTORY/WAREHOUSE"
## [27] "CLOTHING BOUTIQUE"         "SHOE STORE"
## [29] "HOTEL/MOTEL"               "CANDY STORE"
## [31] "DEPT STORE"                "BANK"
## [33] "TELECOMM. STORE"           "DRUG STORE"
## [35] "LOAN COMPANY"              "CHECK CASH"
## [37] "SCHOOL"                    "STORAGE FACILITY"
## [39] "PHOTO/COPY STORE"          "ATM"
## [41] "DOCTOR/DENTIST"
##
## $STATISTICAL_MURDER_FLAG
## [1] FALSE  TRUE
##
## $PERP_AGE_GROUP
##  [1] NA        "25-44"   "(null)"  "UNKNOWN" "18-24"   "<18"     "45-64"
##  [8] "65+"     "1028"    "1020"    "940"     "224"     "2021"
##
## $PERP_SEX
## [1] NA       "M"      "(null)" "U"      "F"
##
## $PERP_RACE
## [1] NA                               "WHITE HISPANIC"
## [3] "(null)"                         "UNKNOWN"
## [5] "BLACK"                          "BLACK HISPANIC"
## [7] "ASIAN / PACIFIC ISLANDER"       "WHITE"
## [9] "AMERICAN INDIAN/ALASKAN NATIVE"
##
## $VIC_AGE_GROUP
## [1] "18-24"   "25-44"   "<18"     "45-64"   "65+"     "UNKNOWN" "1022"
##
## $VIC_SEX
## [1] "M" "F" "U"
##
```

```
## $VIC_RACE
## [1] "BLACK"                          "WHITE HISPANIC"
## [3] "BLACK HISPANIC"                 "ASIAN / PACIFIC ISLANDER"
## [5] "WHITE"                          "UNKNOWN"
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
```

4.3 *Strategy for Remaining Missing Values*

As seen above, we will lose out on possibly valuable information if we simply dropped all columns that contained missing information. So we will instead keep them, and also transform all entries that signify missing information (values like "NONE", "(null)") to be the value NA as well.

```r
nypd_shooting_data_tidy <- nypd_shooting_data_filtered %>%
  mutate(
    across(
      c(LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE),
      ~ .x %>%
        dplyr::na_if("NONE") %>%
        dplyr::na_if("(null)") %>%
        dplyr::na_if("UNKNOWN") %>%
        dplyr::na_if("1020") %>%
        dplyr::na_if("1028") %>%
        dplyr::na_if("2021") %>%
        dplyr::na_if("224") %>%
        dplyr::na_if("940") %>%
        dplyr::na_if("U") %>%
        dplyr::na_if("1022")
    ),
)
```

Finally, let's verify to make sure that we have successfully converted all placeholders for missing information to actual NA values

```r
unique_values <- lapply(nypd_shooting_data_tidy[4:12], unique)
print(unique_values)
```

```
## $BORO
## [1] BRONX         BROOKLYN      MANHATTAN     QUEENS        STATEN ISLAND
## Levels: BRONX BROOKLYN MANHATTAN QUEENS STATEN ISLAND
##
## $LOCATION_DESC
##  [1] NA                          "GROCERY/BODEGA"
##  [3] "PVT HOUSE"                 "MULTI DWELL - APT BUILD"
##  [5] "MULTI DWELL - PUBLIC HOUS" "BAR/NIGHT CLUB"
##  [7] "COMMERCIAL BLDG"           "FAST FOOD"
##  [9] "HOSPITAL"                  "BEAUTY/NAIL SALON"
## [11] "LIQUOR STORE"              "CHAIN STORE"
## [13] "RESTAURANT/DINER"          "SMALL MERCHANT"
## [15] "GAS STATION"               "JEWELRY STORE"
## [17] "GYM/FITNESS FACILITY"      "STORE UNCLASSIFIED"
## [19] "SOCIAL CLUB/POLICY LOCATI" "DRY CLEANER/LAUNDRY"
## [21] "VIDEO STORE"               "SUPERMARKET"
## [23] "VARIETY STORE"            "FACTORY/WAREHOUSE"
## [25] "CLOTHING BOUTIQUE"        "SHOE STORE"
## [27] "HOTEL/MOTEL"              "CANDY STORE"
## [29] "DEPT STORE"               "BANK"
## [31] "TELECOMM. STORE"          "DRUG STORE"
## [33] "LOAN COMPANY"             "CHECK CASH"
## [35] "SCHOOL"                   "STORAGE FACILITY"
## [37] "PHOTO/COPY STORE"         "ATM"
## [39] "DOCTOR/DENTIST"
##
## $STATISTICAL_MURDER_FLAG
## [1] FALSE  TRUE
##
## $PERP_AGE_GROUP
## [1] NA      "25-44" "18-24" "<18"   "45-64" "65+"
##
## $PERP_SEX
## [1] NA  "M" "F"
##
## $PERP_RACE
## [1] NA                               "WHITE HISPANIC"
## [3] "BLACK"                          "BLACK HISPANIC"
## [5] "ASIAN / PACIFIC ISLANDER"       "WHITE"
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
##
## $VIC_AGE_GROUP
## [1] "18-24" "25-44" "<18"   "45-64" "65+"   NA
##
## $VIC_SEX
## [1] "M" "F" NA
##
## $VIC_RACE
## [1] "BLACK"                          "WHITE HISPANIC"
## [3] "BLACK HISPANIC"                 "ASIAN / PACIFIC ISLANDER"
## [5] "WHITE"                          NA
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
```

5. Visualization and Analysis Part 1

5.1 *Gun Incident Trend Over Time For Each Borough*

5.1.A *Creating a Year/Borough Group and Finding Counts*

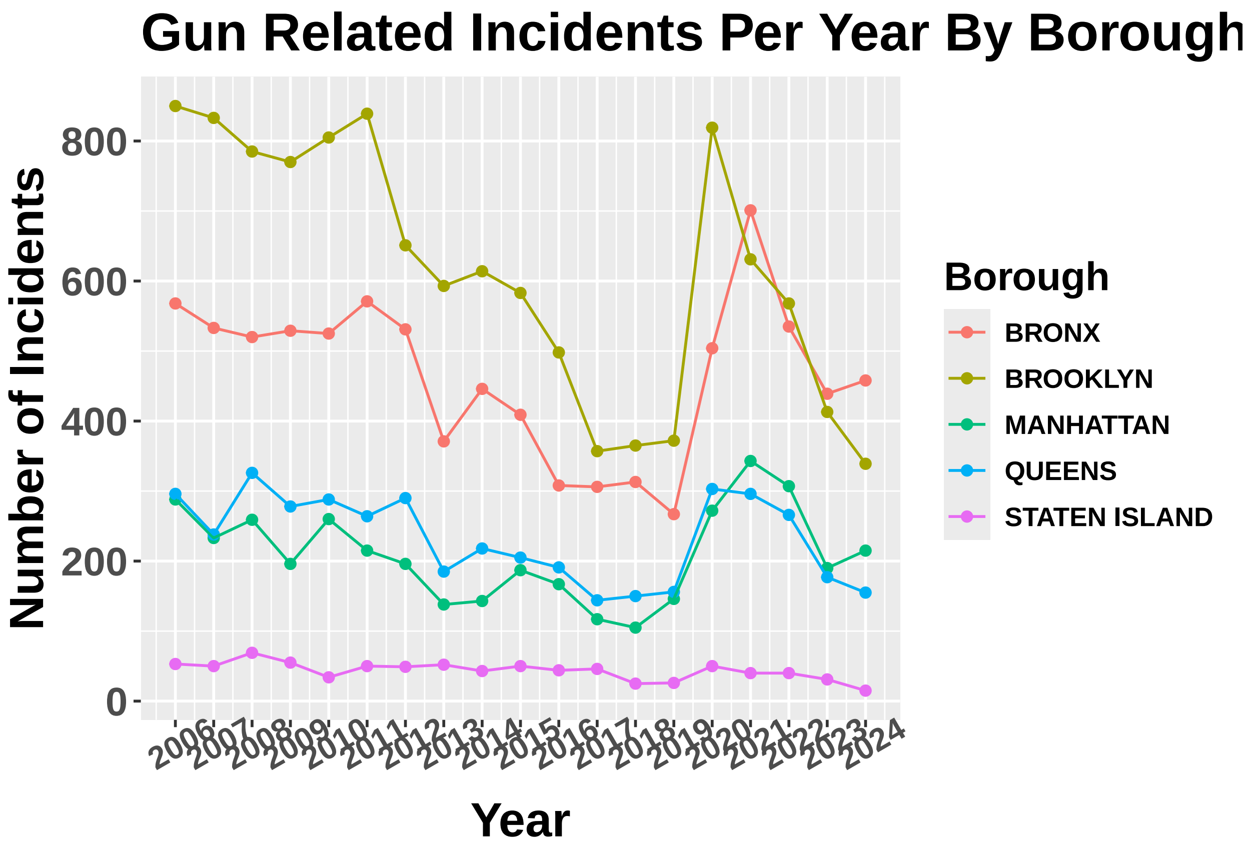
```r
nypd_shooting_data_yearly <- nypd_shooting_data_tidy %>%
    mutate(YEAR = year(OCCUR_DATE)) %>%
    group_by(YEAR, BORO) %>%
    summarize(year_count = n()) %>%
    ungroup()
```

```
## `summarise()` has grouped output by 'YEAR'. You can override using the
## `.groups` argument.
```

5.1.B *Plotting Trend of Gun Incidents for Each Borough*

```r
options(repr.plot.width = 12, repr.plot.height = 8)

ggplot(nypd_shooting_data_yearly, aes(x = YEAR, y = year_count, color = BORO)) +
    geom_point() +
    geom_line(aes(group = BORO)) +
    scale_x_continuous(breaks = unique(nypd_shooting_data_yearly$YEAR)) +
    labs(title = "Gun Related Incidents Per Year By Borough", x = "Year", y = "Number of Incidents", col
    theme(
        plot.title = element_text(size = 20, face = "bold"),
        axis.title.x = element_text(size = 18, face = "bold"),
        axis.title.y = element_text(size = 18, face = "bold"),
        axis.text.x = element_text(size = 12, face = "bold", angle = 30),
        axis.text.y = element_text(size = 15, face = "bold"),
        legend.title = element_text(size = 15, face = "bold"),
        legend.text = element_text(size = 10, face = "bold")
    )
```



From the graph above, we can clearly see that **Brooklyn** and the **Bronx** have the highest rates among all boroughs, with Brooklyn having more incidents early on while things have evened out in the most recent years. It is interesting to see that while some boroughs might be more dangerous than others, there is this overall trend that appears where the rate of incidents was *slowly decreasing over time* then suddenly spiking around the time when the pandemic was in full swing. Fortunately it seems that a return to lower incidents should be expected and if things continue, could improve to levels even lower than 2019.

5.2 *Analysis on Brooklyn*

```
brooklyn_yearly_rate <- nypd_shooting_data_yearly %>%
    filter(BORO == "BROOKLYN") %>%
    mutate(year_change = (year_count - lag(year_count)))

brooklyn_negative_rate <- brooklyn_yearly_rate %>%
    filter(year_change < 0)

brooklyn_positive_rate <- brooklyn_yearly_rate %>%
    filter(year_change > 0)
```

5.2.A *Looking Closer At Consistent Decrease and Subsequent Highest Increase*

```
decreasing_years_sum <- sum(brooklyn_negative_rate$year_change)
decreasing_years_amount <- length(brooklyn_negative_rate$year_change)
glue("Since 2006, Brooklyn has had {decreasing_years_amount} years where gun related incidents were lowe
```

```
## Since 2006, Brooklyn has had 12 years where gun related incidents were lower than the previous perio
```

```
max_change <- max(brooklyn_positive_rate$year_change)
glue("This consistent decline, is made more shocking then when contrasted with the highest annual chang
```

```
## This consistent decline, is made more shocking then when contrasted with the highest annual change t
```

5.2.B *Current Promising Trends*

```
last_count <- brooklyn_yearly_rate[nrow(brooklyn_yearly_rate), ]$year_count
mean_count <- mean(brooklyn_yearly_rate$year_count)

glue("The good news is that since then it has once again subsided, with a continuous decline since then
```

```
## The good news is that since then it has once again subsided, with a continuous decline since then, c
```

6. Visualization and Analysis Part 2

6.1 *Analyzing Relationship Between Perpetrator and Victim Race*

6.1.A *Creating Perpetrator/Victim Group and Finding Counts*

```
nypd_shooting_data_race <- nypd_shooting_data_tidy %>%
    group_by(PERP_RACE, VIC_RACE) %>%
    summarize(count = n()) %>%
    ungroup()
```

```
## ‘summarise()‘ has grouped output by ’PERP_RACE’. You can override using the
## ‘.groups‘ argument.
```

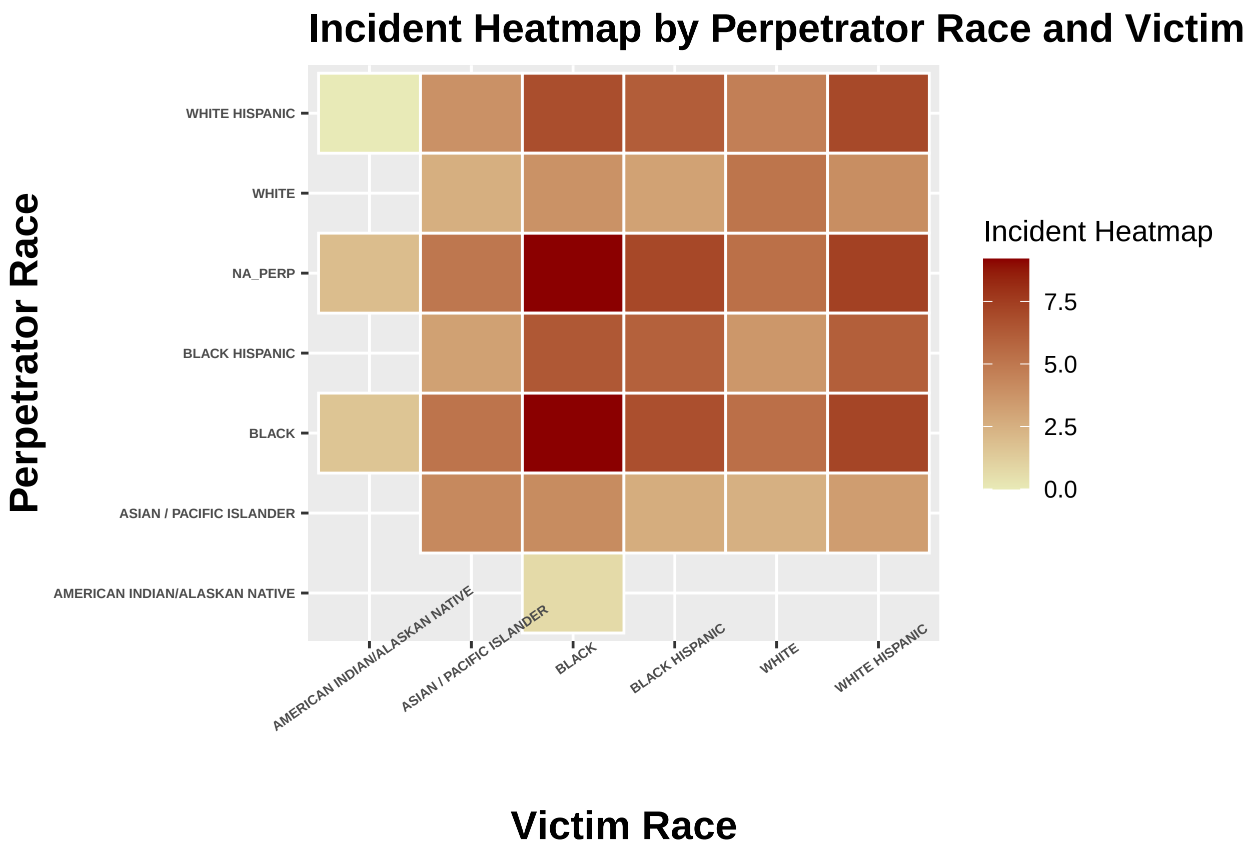
6.1.B *Creating Perpetrator/Victim Group and Finding Counts*

Here we drop all the NAs from Victim Race but keep the NAs from Perpetrator Race as it might be able to provide us with some information on what the baseline/general relationship is between an unknown perpetrator and an identified victim. We also mutate the count column to its log form in order to decrease the gap between counts which will help with the visualization.

```r
nypd_shooting_data_race_tidy <- nypd_shooting_data_race %>%
    arrange(desc(count)) %>%
    filter(!is.na(VIC_RACE)) %>%
    mutate(
        PERP_RACE = ifelse(is.na(PERP_RACE), "NA_PERP", PERP_RACE)
    ) %>%
    mutate(
        count = log(count),
        PERP_RACE = as.factor(PERP_RACE),
        VIC_RACE = as.factor(VIC_RACE)
    )
```

6.1.C *Heatmap of Perpetrator Race/ Victim Race*

```r
options(repr.plot.width = 25, repr.plot.height = 10)
ggplot(
    nypd_shooting_data_race_tidy,
    aes(x = VIC_RACE, y = PERP_RACE, fill = count)
) +
    geom_tile(
        color = "white",
        linewidth = 0.5
    ) +
    scale_fill_gradient(
        low = "#e8eab7",
        high = "darkred",
        name = "Incident Heatmap"
    ) +
    labs(
        title = "Incident Heatmap by Perpetrator Race and Victim Race",
        y = "Perpetrator Race",
        x = "Victim Race"
    ) +
    theme(
        plot.title = element_text(size = 15, face = "bold"),
        axis.title.x = element_text(size = 15, face = "bold"),
        axis.title.y = element_text(size = 15, face = "bold"),
        axis.text.x = element_text(size = 5, face = "bold", angle = 35),
        axis.text.y = element_text(size = 5, face = "bold")
    )
```



6.1.D *Analysis on Heatmap*

The main interesting part to note for me is the fact that the *NA_PERP heatmap closely resembles the heatmaps of other higher aggressors**. It could mean that this does represent the general relationship of gun incidents in New York, or that the true makeup of the population of this group is very similar to other high incident aggressors, but without more information, we cannot definitively say.

The other interesting part to note is that while it might be easy to say that African Americans are the most aggressive perpetrators, if we look closely, even if that may be the case, they still also end up being victimzed at a disproportionately higher rate in comparison to other races.

6.2 *Analysis on Other Races*

Ultimately what I find most interesting is the fact that in almost all instances, a specific race will be the aggressor at the highest rate towards someone that is from their own race. This can be seen clearly by filtering the perpetrator/victim grouped dataset and displaying each race one by one.

```
filter(nypd_shooting_data_race_tidy, PERP_RACE == "ASIAN / PACIFIC ISLANDER")
```

```
## # A tibble: 5 x 3
##   PERP_RACE                VIC_RACE                 count
##   <fct>                    <fct>                    <dbl>
## 1 ASIAN / PACIFIC ISLANDER ASIAN / PACIFIC ISLANDER  4.23
## 2 ASIAN / PACIFIC ISLANDER BLACK                     4.11
## 3 ASIAN / PACIFIC ISLANDER WHITE HISPANIC            3.33
## 4 ASIAN / PACIFIC ISLANDER BLACK HISPANIC            2.64
## 5 ASIAN / PACIFIC ISLANDER WHITE                     2.48
```

```
filter(nypd_shooting_data_race_tidy, PERP_RACE == "WHITE")
```

```
## # A tibble: 5 x 3
##   PERP_RACE VIC_RACE                 count
##   <fct>     <fct>                    <dbl>
## 1 WHITE     WHITE                     5.12
## 2 WHITE     WHITE HISPANIC            4.01
## 3 WHITE     BLACK                     3.81
## 4 WHITE     BLACK HISPANIC            3.14
## 5 WHITE     ASIAN / PACIFIC ISLANDER  2.56
```

```
filter(nypd_shooting_data_race_tidy, PERP_RACE == "WHITE HISPANIC")
```

```
## # A tibble: 6 x 3
##   PERP_RACE      VIC_RACE                       count
##   <fct>          <fct>                          <dbl>
## 1 WHITE HISPANIC WHITE HISPANIC                  7.03
## 2 WHITE HISPANIC BLACK                           6.80
## 3 WHITE HISPANIC BLACK HISPANIC                  6.16
## 4 WHITE HISPANIC WHITE                           4.65
## 5 WHITE HISPANIC ASIAN / PACIFIC ISLANDER        3.85
## 6 WHITE HISPANIC AMERICAN INDIAN/ALASKAN NATIVE  0
```

7. Modeling

In this section, we will attempt to fit a linear model and find out if the amount of statistical murders are a good predictor of the total annual amount of gun incidents

7.1 *Create Yearly Grouping and Column for Counts of Statistical Murder*

```
nypd_shooting_model_yearly <- nypd_shooting_data_tidy %>%
    mutate(YEAR = year(OCCUR_DATE))

yearly_summary <- nypd_shooting_model_yearly %>%
    group_by(YEAR) %>%
    summarise(
        total_incidents_per_year = n(),
        total_murders_per_year = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE)
    ) %>%
    ungroup()

print(head(yearly_summary))
```

```
## # A tibble: 6 x 3
##    YEAR total_incidents_per_year total_murders_per_year
##   <dbl>                    <int>                  <int>
## 1  2006                     2055                    445
## 2  2007                     1887                    373
## 3  2008                     1959                    362
## 4  2009                     1828                    348
## 5  2010                     1912                    405
## 6  2011                     1939                    373
```

7.2 *Fit the Linear Model, Print the Summary and Coefficients*

```
nypd_shooting_murder_model <- lm(total_incidents_per_year ~ total_murders_per_year, data = yearly_summar
summary(nypd_shooting_murder_model)
```

```
##
## Call:
## lm(formula = total_incidents_per_year ~ total_murders_per_year,
##     data = yearly_summary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.264 -100.549   -2.232   86.111  219.652
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            225.0583   100.0567   2.249    0.038 *
## total_murders_per_year   4.4177     0.3182  13.884 1.05e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.5 on 17 degrees of freedom
## Multiple R-squared:  0.919,  Adjusted R-squared:  0.9142
## F-statistic: 192.8 on 1 and 17 DF,  p-value: 1.049e-10
```

```
coefficients(nypd_shooting_murder_model)
```

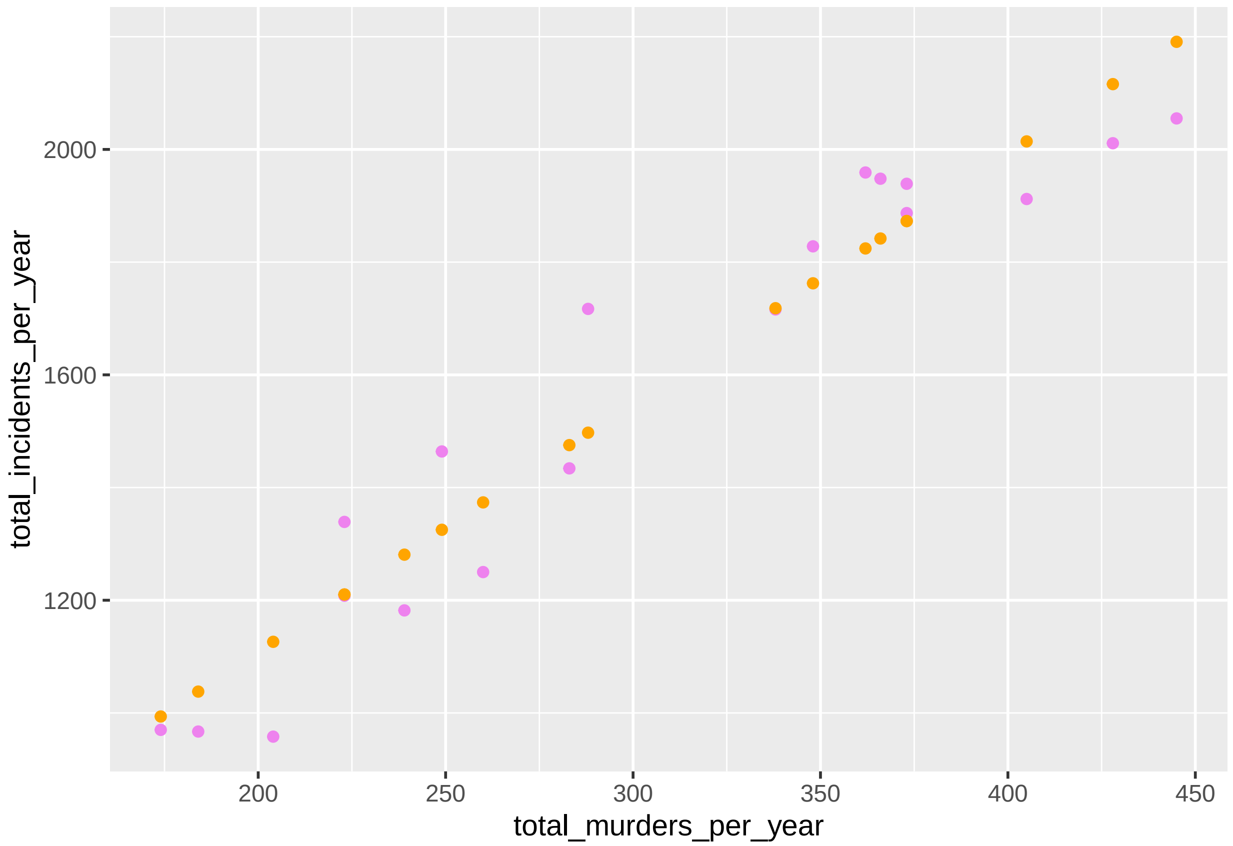
```
##            (Intercept) total_murders_per_year
##             225.058251               4.417674
```

Looking at the summary of our model, we see that our median residuals are close to zero, both of our p-values are very low suggesting that this relationship is statistically significant, and a high R-squared of 91% suggest that much of the variance in yearly gun incidents can be explained by the variation in statistical murders. All of these suggest that our model should be a good fit for these variables.

7.3 *Create Prediction Column and Plot Graph of Actual vs Predicted*

```
yearly_predictions <- yearly_summary %>%
  mutate(predictions = predict(nypd_shooting_murder_model))
```

```
yearly_predictions %>% ggplot() +
  geom_point(aes(x = total_murders_per_year, y = total_incidents_per_year), color = "violet") +
  geom_point(aes(x = total_murders_per_year, y = predictions), color = "orange")
```



8. Conclusion

Overall, we were able to gain significant insights surrounding the problem of gun violence within New York city. It seems as if currently, the city is possibly reaching a turning point, where if it keeps incidents low and sustains that level - it could further open up the possibility of even higher levels of safety with regards to gun related incidents but we have also seen that its population was most impacted during a time of great uncertainty when Covid was at the forefront and today, while we might not have a pandemic, many things are still seemingly in flux which could in turn produce chaos reminiscent of what had happened years before. We were able to see how incidents have risen and decline, but much of why it does seems to stem from factors outside of guns themselves but rather might be more closely tied to the population, which makes it hard to guarantee further improvements.

New York's gun problem being tied to each individual that makes up the city is also displayed on the analysis of the relationship between perpetrator and victim race. The ethnic groups that have been marginalized the most in our society should not just be seen as the highest aggressors, but more importantly, are also primary victims as well. Even further, we have seen that same-race tensions tend to be higher for any ethnic group, and as such, this problem is one that affects all of us and not just a selected few.

As with any analysis on a topic that has a deep societal impact, we are usually left with more questions, especially on what we can do next to improve the situation. How can we ensure the continuation of the current decline in overall incidents? While the most violent boroughs have seen improvement, there has been an uptick in previously safer ones - is there anything that can prevent that from progressing further? What can be done about the disproportionate effects of gun violence in different ethnic groups? Why are same-race incidents predominantly at the top of perpetrator-victim relationships? We might not be able to answer these questions at the moment, but we can hope that by investing further into learning more and finding ways to positively influence the situation, we might at least help create a safer future for New Yorkers.

9. Bias

The most important factor to consider when thinking about Bias is of course, how the data is sourced and gathered. Due to how inherent racial biases are ingrained in our society - especially in the judiciary and law enforcement space, it can be easy to think of many ways in how incident reports can be manipulated, consciously or subconsciously, to over represent certain neighborhoods or groups of people that might have lead to the results we saw in our analysis. As I do not have much knowledge in the finest details in how these incidents are logged, we can only trust that our data is kept and maintained under a fair and objective point of view.

As for my own personal biases, when I had first thought of the idea to tackle the racial aspect of the relationship between perpetrators and victims, I became a bit hesitant as I did not want the results to cloud my judgement on individuals that I encounter in my day to day. But after objectively going through the analysis, finding the disproportionate level of victimization, and the fact that same-race tension is a matter that affects most of us that I did not even consider previously, I was able to reserve my initial judgments which allowed me to be much more factually informed on the matter at hand.

```r
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26100)
## 
## Matrix products: default
## 
## 
## locale:
## [1] LC_COLLATE=English_United States.utf8 
## [2] LC_CTYPE=English_United States.utf8   
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.utf8    
## 
## time zone: America/Los_Angeles
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] glue_1.8.0      lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1  
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1    
##  [9] tibble_3.2.1    ggplot2_3.5.2   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
```

```
##  [1] bit_4.6.0          gtable_0.3.6       crayon_1.5.3       compiler_4.4.2    
##  [5] tidyselect_1.2.1   parallel_4.4.2     scales_1.4.0       yaml_2.3.10       
##  [9] fastmap_1.2.0      R6_2.5.1           labeling_0.4.3     generics_0.1.4    
## [13] knitr_1.49         pillar_1.10.0      RColorBrewer_1.1-3 tzdb_0.5.0        
## [17] rlang_1.1.4        utf8_1.2.4         stringi_1.8.4      xfun_0.49         
## [21] bit64_4.6.0-1      timechange_0.3.0   cli_3.6.3          withr_3.0.2       
## [25] magrittr_2.0.3     digest_0.6.37      grid_4.4.2         vroom_1.6.5       
## [29] rstudioapi_0.17.1  hms_1.1.3          lifecycle_1.0.4    vctrs_0.6.5       
## [33] evaluate_1.0.1     farver_2.1.2       rmarkdown_2.29     tools_4.4.2       
## [37] pkgconfig_2.0.3    htmltools_0.5.8.1
```