NYPD Shooting Incident Data Report

3/1/2024

Introduction

In this report, I will be importing and analyzing historic NYPD shooting incident data as reported by the City of New York.

Data description: List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

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.

Step 0: Import Library

The following libraries will be required:

```r
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
library(survival)
```

Step 1: Load Data

Import data from source.

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

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
```

```
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     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:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              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
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  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   Max.   :271128   Max.   :40.91
##                                                        NA’s   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA’s   :10
```

Step 2: Data Cleaning

I am going to do some data cleaning by changing variables to the appropriate formats and removing columns which are not needed for my analysis. There is also some missing data, which I will classify as “unknown”.

```
data_2 <- data
data_2 = subset(data_2, PERP_AGE_GROUP!="1020" & PERP_AGE_GROUP!="224"
                & PERP_AGE_GROUP!="940" & VIC_AGE_GROUP!="1022")
data_2["PERP_AGE_GROUP"][data_2["PERP_AGE_GROUP"] == "(null)"] <- "UNKNOWN"
data_2["PERP_SEX"][data_2["PERP_SEX"] == "(null)"] <- "U"
data_2["PERP_RACE"][data_2["PERP_RACE"] == "(null)"] <- "UNKNOWN"
```

```
data_2 <- data_2 %>%
  select(-c(LOC_OF_OCCUR_DESC,JURISDICTION_CODE,LOC_CLASSFCTN_DESC,LOCATION_DESC,
            X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat)) %>%
  replace_na(list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "U", PERP_RACE = "UNKNOWN")) %>%
  mutate(INCIDENT_KEY = as.character(INCIDENT_KEY),OCCUR_DATE = mdy(OCCUR_DATE),
         BORO = as.factor(BORO), PRECINCT = as.factor(PRECINCT),
         PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP), PERP_RACE = as.factor(PERP_RACE),
         PERP_SEX = as.factor(PERP_SEX), VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
         VIC_RACE = as.factor(VIC_RACE), VIC_SEX = as.factor(VIC_SEX))

summary(data_2)
```

```
##  INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME                  BORO
##  Length:17964       Min.   :2006-01-01   Length:17964      BRONX        :5423
##  Class :character   1st Qu.:2008-08-05   Class1:hms        BROOKLYN     :6641
##  Mode  :character   Median :2011-11-18   Class2:difftime   MANHATTAN    :2541
##                     Mean   :2013-05-11   Mode  :numeric    QUEENS       :2728
##                     3rd Qu.:2018-04-26                     STATEN ISLAND: 631
##                     Max.   :2022-12-31
##
##     PRECINCT     STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  75     : 1001   Mode :logical           <18    :1591   F:  424
##  73     :  867   FALSE:14404             18-24  :6221   M:15435
##  47     :  693   TRUE :3560              25-44  :5687   U: 2105
##  44     :  690                           45-64  : 617
##  46     :  657                           65+    :   60
##  67     :  601                           UNKNOWN:3788
##  (Other):13455
##                            PERP_RACE     VIC_AGE_GROUP  VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    :2027   F: 1922
##  ASIAN / PACIFIC ISLANDER      :  154   18-24  :6517   M:16034
##  BLACK                         :11430   25-44  :7937   U:    8
##  BLACK HISPANIC                : 1314   45-64  :1290
##  UNKNOWN                       : 2442   65+    : 137
##  WHITE                         :  283   UNKNOWN:  56
##  WHITE HISPANIC                : 2339
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    8
##  ASIAN / PACIFIC ISLANDER      :  307
##  BLACK                         :12250
##  BLACK HISPANIC                : 1800
##  UNKNOWN                       :   48
##  WHITE                         :  552
##  WHITE HISPANIC                : 2999
```

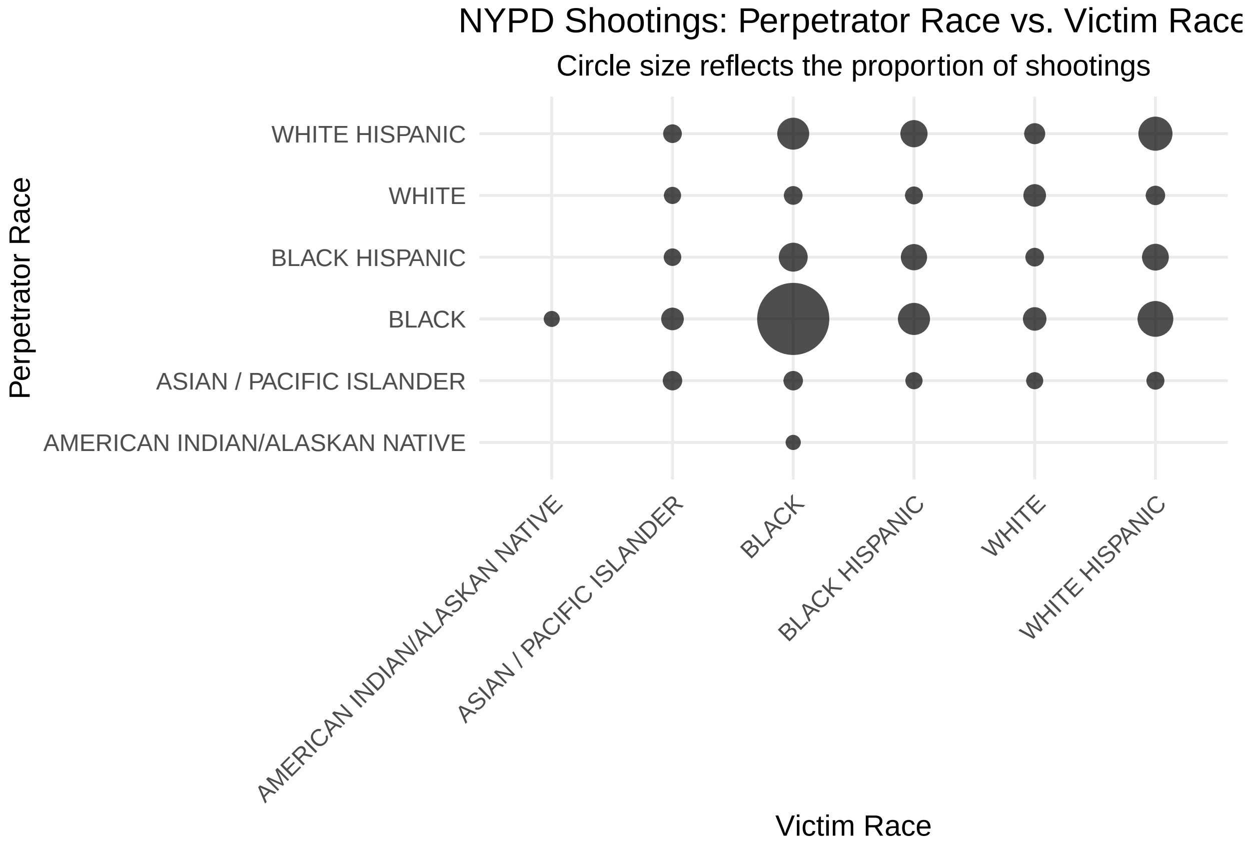
Step 3: Analysis & Visualization

1. My first question that I want to investigate further is if there is a relationship between the race of the victim and the race of the perpetrator.

```
race_combinations <- data_2 %>%
  filter(PERP_RACE!= "UNKNOWN", VIC_RACE!= "UNKNOWN") %>%
  group_by(PERP_RACE, VIC_RACE) %>%
  summarise(Count = n(), .groups = 'drop')

total_counts <- sum(race_combinations$Count)
race_combinations <- race_combinations %>%
  mutate(Proportion = Count / total_counts)

ggplot(race_combinations, aes(x = VIC_RACE, y = PERP_RACE, size = Proportion)) +
  geom_point(alpha = 0.7) +
  scale_size_continuous(range = c(2,12)) +
  theme_minimal() +
  labs(title = "NYPD Shootings: Perpetrator Race vs. Victim Race",
       subtitle = "Circle size reflects the proportion of shootings",
       x = "Victim Race",
       y = "Perpetrator Race",
       size = "Proportion") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))+
  guides(size = "none")
```



There does appear to be some correlation between the race of the victim and the race of the perpetrator. Additional analysis would need to be completed to determine if these findings are statistically significant. It would also help to know the overall demographics of New York City and the NYPD.

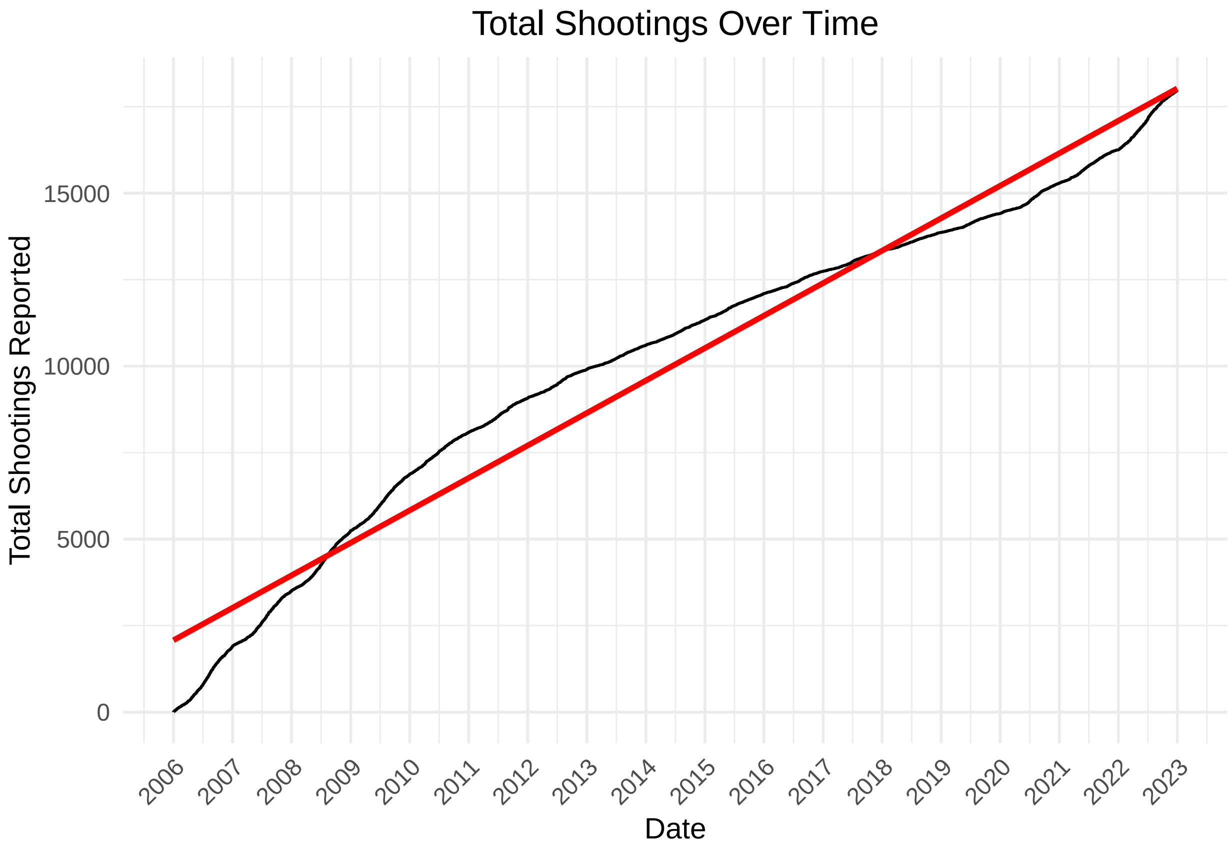
2. Violence committed by police officers is a topic that is frequently in the news. I would like to see if it

appears that the rate of shootings is increasing over time.

```
data_graph_2 <- data_2 %>%
  arrange(OCCUR_DATE)%>%
  mutate(TotalShootingsToDate = cumsum(!is.na(OCCUR_DATE))) %>%
  mutate(DailyRateOfChange = c(0, diff(TotalShootingsToDate)))

ggplot(data_graph_2, aes(x = OCCUR_DATE, y = TotalShootingsToDate)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
  theme_minimal() +
  labs(title = "Total Shootings Over Time",
       x = "Date",
       y = "Total Shootings Reported") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Looking at cumulative shootings over time, it does not appear that the rate of shootings is necessarily increasing or decreasing significantly in New York City. I think that it would help to have population data to put shootings in terms of “per 100,000”, for example.

3. Next I want to build a model using logistic regression to determine if race, sex, or age are a predictor of whether a shooting victim will survive.

```r
logistic_model <- glm(STATISTICAL_MURDER_FLAG ~ VIC_RACE + VIC_SEX + VIC_AGE_GROUP,
                      data_2, family = "binomial")
summary(logistic_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_RACE + VIC_SEX +
##     VIC_AGE_GROUP, family = "binomial", data = data_2)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -12.76308  114.10229  -0.112  0.91094
## VIC_RACEASIAN / PACIFIC ISLANDER  11.36796  114.10234   0.100  0.92064
## VIC_RACEBLACK                     11.05406  114.10227   0.097  0.92282
## VIC_RACEBLACK HISPANIC            10.90933  114.10228   0.096  0.92383
## VIC_RACEUNKNOWN                   10.49549  114.10321   0.092  0.92671
## VIC_RACEWHITE                     11.41747  114.10230   0.100  0.92029
## VIC_RACEWHITE HISPANIC            11.21451  114.10227   0.098  0.92171
## VIC_SEXM                          -0.16254    0.05909  -2.751  0.00595 **
## VIC_SEXU                          -0.32960    1.12749  -0.292  0.77003
## VIC_AGE_GROUP18-24                 0.30495    0.07224   4.221 2.43e-05 ***
## VIC_AGE_GROUP25-44                 0.55537    0.07006   7.927 2.25e-15 ***
## VIC_AGE_GROUP45-64                 0.66478    0.09183   7.239 4.51e-13 ***
## VIC_AGE_GROUP65+                   0.90917    0.19774   4.598 4.27e-06 ***
## VIC_AGE_GROUPUNKNOWN               0.57580    0.34918   1.649  0.09915 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17887  on 17963  degrees of freedom
## Residual deviance: 17723  on 17950  degrees of freedom
## AIC: 17751
##
## Number of Fisher Scoring iterations: 11
```

The logistic regression model predicts the log-odds of the event STATISTICAL_MURDER_FLAG based on race, sex, and age group variables. Significant coefficients and their associated significance codes indicate the direction and strength of the relationships. From the results, it appears that race and gender do not have a significant impact on the log-odds of fatality, but as may have been expected, fatality is more likely for individuals who are older.

Step 4: Identifying Bias

One source of bias could be in the way the data is reported and reviewed. Is the reporting police officer responsible for filling out the incident report? Is the report reviewed by an unbiased individual? This could impact the data that is reported versus what is omitted. For example, I noticed than the race of the perpetrator was unreported or “unknown” for 2,2442 observations, which was the case for only 48 of the victims reported. Is that information omitted by simple oversight or could it be intentional?

Because of all the media attention on this topic, I was also likely biased in my analysis. I tried to mitigate this by simply asking myself what I was most curious to learn from the data, as opposed to setting out to prove a specific point. I wrote the code not knowing what I would find.

Appendix: Session Info

```r
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Big Sur 11.7.4
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib 
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK v
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: America/New_York
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] survival_3.5-7  lubridate_1.9.3 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_3.4.4   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    stringi_1.8.3     lattice_0.21-9   
##  [5] hms_1.1.3         digest_0.6.34     magrittr_2.0.3    evaluate_0.23    
##  [9] grid_4.3.2        timechange_0.3.0  fastmap_1.1.1     Matrix_1.6-1.1   
## [13] mgcv_1.9-0        fansi_1.0.6       scales_1.3.0      cli_3.6.2        
## [17] rlang_1.1.3       crayon_1.5.2      bit64_4.0.5       munsell_0.5.0    
## [21] splines_4.3.2     withr_3.0.0       yaml_2.3.8        tools_4.3.2      
## [25] parallel_4.3.2    tzdb_0.4.0        colorspace_2.1-0  curl_5.2.0       
## [29] vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4   bit_4.0.5        
## [33] vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0      gtable_0.3.4     
## [37] glue_1.7.0        highr_0.10        xfun_0.42         tidyselect_1.2.0 
## [41] rstudioapi_0.15.0 knitr_1.45        farver_2.1.1      nlme_3.1-163     
## [45] htmltools_0.5.7   labeling_0.4.3    rmarkdown_2.25    compiler_4.3.2
```