NYPD Report Analysis

Antonio J Rivera Lopez

```r
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     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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 (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
knitr::opts_chunk$set(echo = TRUE)
```

Reading the data:

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

```
## 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.
```

Clean out unnecessary columns:

```r
shooting_data <- shooting_data %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, LOC_OF_OCCUR_DESC
```

```r
library(lubridate)
```

Doing transformations to tidy up the data:

```r
# Join date and time variables
shooting_data <- shooting_data %>% mutate(OCCUR_DATETIME = mdy_hms(paste(OCCUR_DATE, OCCUR_TIME, sep =

# Add factors
shooting_data <- shooting_data %>% mutate(PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP), PERP_SEX = as.fact
```

```r
summary(shooting_data)
```

```
##   INCIDENT_KEY           BORO           LOC_OF_OCCUR_DESC  LOCATION_DESC
```

```
##  Min.   :  9953245   Length:29744       Length:29744       Length:29744      
##  1st Qu.: 67321140   Class :character   Class :character   Class :character  
##  Median :109291972   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :133850951                                                           
##  3rd Qu.:214741917                                                           
##  Max.   :299462478                                                           
##                                                                              
##  PERP_AGE_GROUP   PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX  
##  18-24  :6630   (null): 1628   BLACK         :12323   <18    : 3081   F: 2891  
##  25-44  :6342   F     :  461   WHITE HISPANIC: 2667   1022   :    1   M:26841  
##  UNKNOWN:3148   M     :16845   UNKNOWN       : 1838   18-24  :10677   U:   12  
##  <18    :1805   U     : 1500   (null)        : 1628   25-44  :13563            
##  (null) :1628   NA's  : 9310   BLACK HISPANIC: 1487   45-64  : 2118            
##  (Other): 847                  (Other)       :  491   65+    :  236            
##  NA's   :9344                  NA's          : 9310   UNKNOWN:   68            
##                            VIC_RACE     OCCUR_DATETIME                  
##  AMERICAN INDIAN/ALASKAN NATIVE:   13   Min.   :2006-01-01 02:00:00.00  
##  ASIAN / PACIFIC ISLANDER      :  478   1st Qu.:2009-10-29 21:05:30.00  
##  BLACK                         :20999   Median :2014-03-25 23:12:00.00  
##  BLACK HISPANIC                : 2930   Mean   :2014-11-01 01:44:39.85  
##  UNKNOWN                       :   72   3rd Qu.:2020-06-29 22:47:45.00  
##  WHITE                         :  741   Max.   :2024-12-31 19:16:00.00  
##  WHITE HISPANIC                : 4511
```

Consolidate all missing values to NA:

```
na_strings <- c("(null)", "NA's")

shooting_data <- shooting_data %>%
  mutate(across(everything(), ~ {
    x <- as.character(.x)
    x_clean <- reduce(na_strings, na_if, .init = x)
    factor(x_clean)  # Back to factor
  }))

shooting_data
```

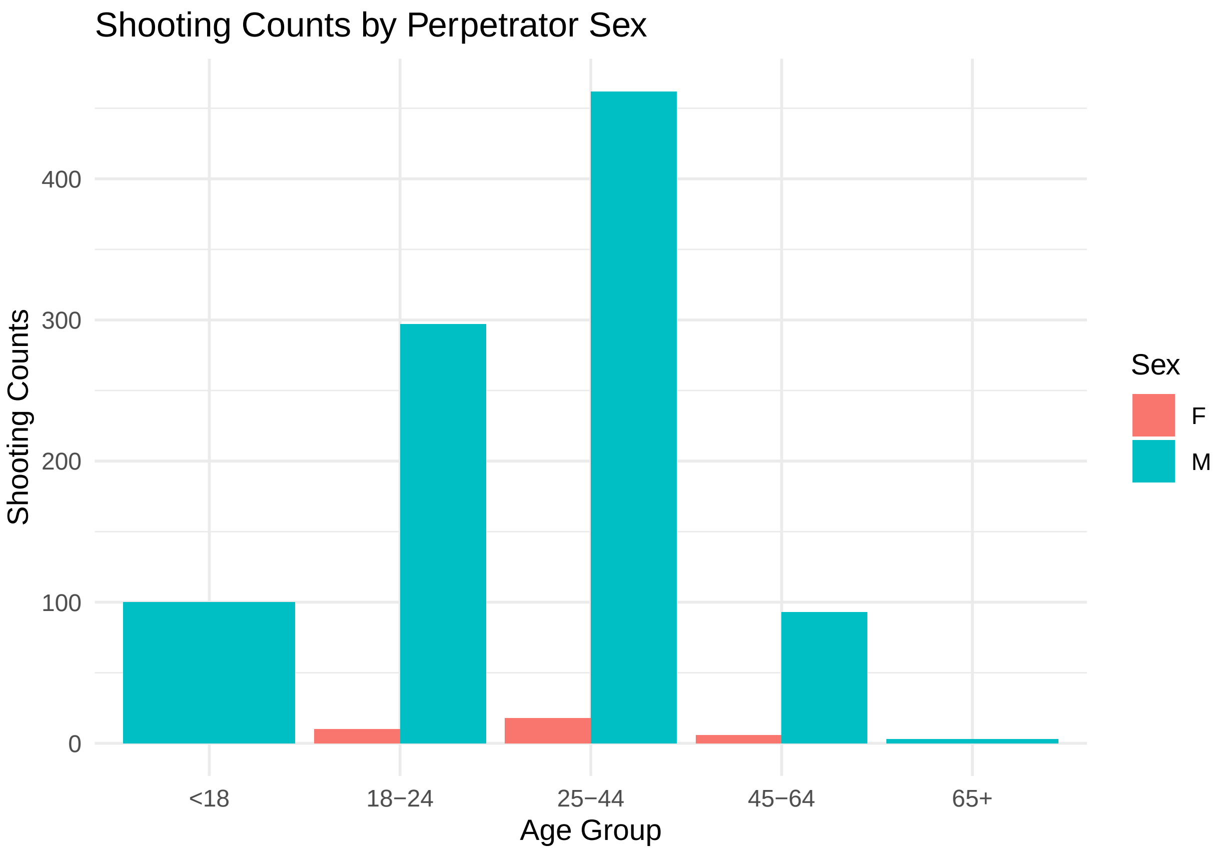
```
## # A tibble: 29,744 x 11
##    INCIDENT_KEY BORO     LOC_OF_OCCUR_DESC LOCATION_DESC PERP_AGE_GROUP PERP_SEX
##    <fct>        <fct>    <fct>             <fct>         <fct>          <fct>   
##  1 231974218    BRONX    <NA>              <NA>          <NA>           <NA>    
##  2 177934247    BROOKLYN <NA>              <NA>          25-44          M       
##  3 255028563    BRONX    OUTSIDE           GROCERY/BODE~ <NA>           <NA>    
##  4 25384540     BROOKLYN <NA>              PVT HOUSE     UNKNOWN        U       
##  5 72616285     BRONX    <NA>              MULTI DWELL ~ 25-44          M       
##  6 85875439     BRONX    <NA>              MULTI DWELL ~ 18-24          M       
##  7 79780323     BROOKLYN <NA>              <NA>          <NA>           <NA>    
##  8 85744504     BROOKLYN <NA>              MULTI DWELL ~ <NA>           <NA>    
##  9 142324890    BROOKLYN <NA>              MULTI DWELL ~ 25-44          M       
## 10 152868707    BROOKLYN <NA>              <NA>          18-24          M       
## # i 29,734 more rows
## # i 5 more variables: PERP_RACE <fct>, VIC_AGE_GROUP <fct>, VIC_SEX <fct>,
## #   VIC_RACE <fct>, OCCUR_DATETIME <fct>
```

There are plenty of NA values, I will leave them there for now, if some analysis requires them to be removed then I can remove them at that moment.

Visualizations

One simple but interesting visualization we can make is to see the most common age for the perpetrators along with their sex. Here I removed the NAs since they do not offer insight for this visualization:

```
ggplot(na.omit(shooting_data), aes(x = PERP_AGE_GROUP, fill = PERP_SEX)) +
  geom_bar(position = position_dodge()) +
  labs(title = "Shooting Counts by Perpetrator Sex", x = "Age Group", y = "Shooting Counts", fill = "Se
  theme_minimal()
```



We can see from this visualization that based on this data, men around 25-44 years old, are the ones that tend to commit more shootings.

```
M_perp <- shooting_data %>% filter(shooting_data$PERP_SEX == "M")
F_perp <- shooting_data %>% filter(shooting_data$PERP_SEX == "F")
sex_prop <- sum(!is.na(M_perp)) / sum(!is.na(F_perp))
print(sex_prop)
```

```
## [1] 36.14238
```

Based on this data, men in this study are 36 times more likely to commit shootings than women.

Model

We can use a logistic regression model to fit the data and see how well victim age groups relate to sex.

```
clean_sh_data <- shooting_data %>% filter(VIC_SEX %in% c("M", "F"), !is.na(VIC_AGE_GROUP), !is.na(VIC_SE
```

```r
clean_sh_data$VIC_SEX <- factor(clean_sh_data$VIC_SEX, levels = c("F", "M"))

clean_sh_data <- clean_sh_data %>% filter(VIC_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+"))

model <- glm(VIC_SEX ~ VIC_AGE_GROUP, data = clean_sh_data, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = VIC_SEX ~ VIC_AGE_GROUP, family = "binomial", data = clean_sh_data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.78949    0.05144  34.786  < 2e-16 ***
## VIC_AGE_GROUP18-24   0.64757    0.06256  10.351  < 2e-16 ***
## VIC_AGE_GROUP25-44   0.60656    0.06009  10.095  < 2e-16 ***
## VIC_AGE_GROUP45-64 -0.28512    0.07629  -3.737 0.000186 ***
## VIC_AGE_GROUP65+   -0.92600    0.15151  -6.112 9.86e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18932  on 29668  degrees of freedom
## Residual deviance: 18585  on 29664  degrees of freedom
## AIC: 18595
##
## Number of Fisher Scoring iterations: 5
```

Visualize model output

```r
# Get all levels of VIC_AGE_GROUP
age_levels <- levels(clean_sh_data$VIC_AGE_GROUP)

# Create a new data frame for prediction
newdata <- data.frame(VIC_AGE_GROUP = age_levels)

# Predict probabilities
newdata$predicted_prob <- predict(model, newdata = newdata, type = "response")

ggplot(newdata, aes(x = VIC_AGE_GROUP, y = predicted_prob)) +
  geom_col(fill = "steelblue") +
  labs(
    title = "Predicted Probability of Victim Being Male by Age Group",
    x = "Victim Age Group",
    y = "Predicted Probability (Male)"
  ) +
  ylim(0, 1) +
  theme_minimal()
```

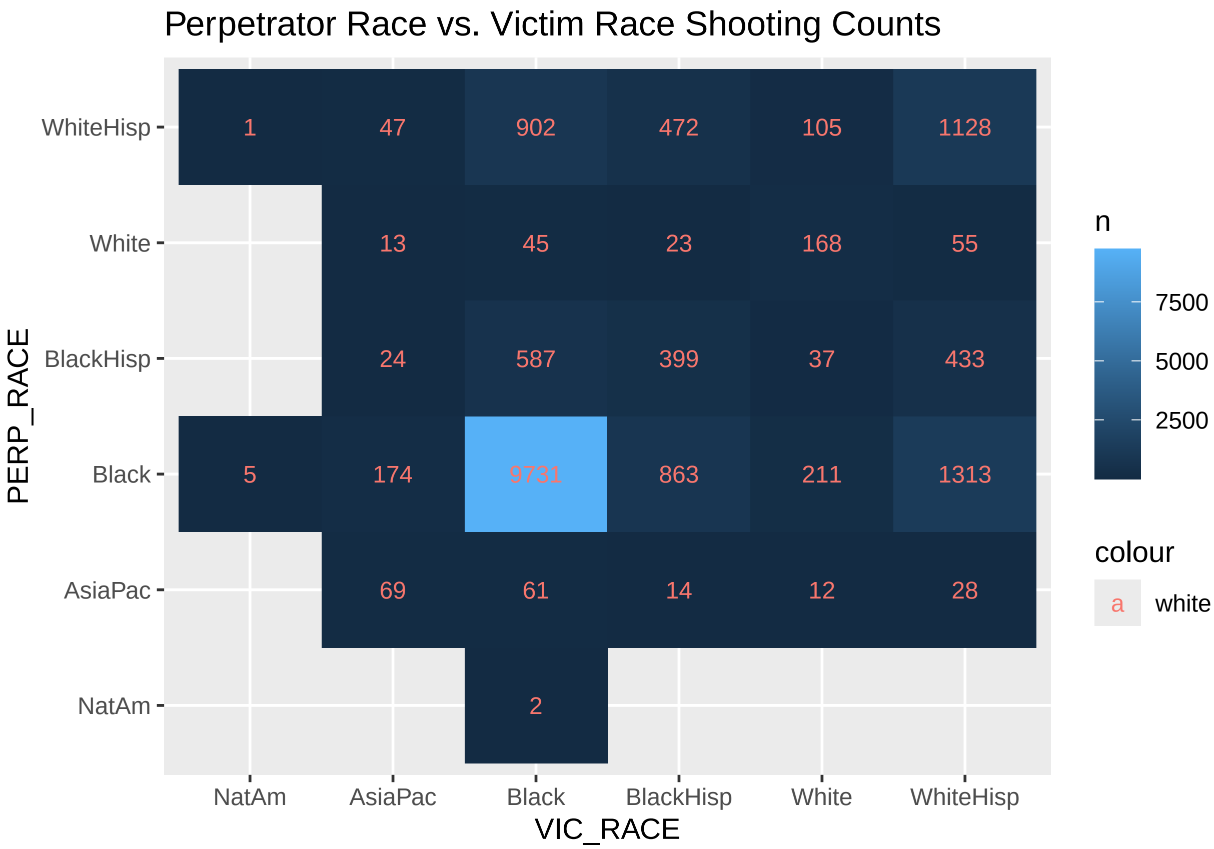
What we can see here is that the probability of a victim being male is the highest for the age groups: 18-24, 25-44. Using this information with the previous bar chart analysis, we can say that the victims and perpetrators are roughly the same age and are mostly male. This pattern is likely to repeat going forward.

Now we analyze how race affects the prevalence of shootings.

```
race_counts <- shooting_data %>% count(PERP_RACE, VIC_RACE) %>% filter(!is.na(PERP_RACE), !is.na(VIC_RA
    PERP_RACE,
    "NatAm" = "AMERICAN INDIAN/ALASKAN NATIVE",
    "AsiaPac"    = "ASIAN / PACIFIC ISLANDER",
    "Black"           = "BLACK",
    "BlackHisp"   = "BLACK HISPANIC",
    "White"           = "WHITE",
    "WhiteHisp"   = "WHITE HISPANIC"
  )) %>% mutate(VIC_RACE = fct_recode(
    VIC_RACE,
    "NatAm" = "AMERICAN INDIAN/ALASKAN NATIVE",
    "AsiaPac"    = "ASIAN / PACIFIC ISLANDER",
    "Black"           = "BLACK",
    "BlackHisp"   = "BLACK HISPANIC",
    "White"           = "WHITE",
    "WhiteHisp"   = "WHITE HISPANIC"
  ))


ggplot(race_counts, aes(VIC_RACE, PERP_RACE)) +
```

```r
geom_tile(aes(fill = n)) +
geom_text(aes(label = n, colour = "white"), size = 3) +
labs(title = "Perpetrator Race vs. Victim Race Shooting Counts")
```



We can see from the visualization that black on black shootings are the most common in this dataset. We need to be careful not to conclude that this implies that the race has an inherit correlation with violence, other variables such as income, poverty levels in the area, etc. could provide more insight and help us draw a more educated conclusion.

Conclusion

Using data wrangling, transformations and visualizations, we saw two main patterns in the data. The first one was that most shootings are executed by male subjects, female subjects have a much lower incidence count. Men of the same age group are also the most likely to shoot each other based on the logistic regression model we created.

The second pattern we found in the data is that race affects the prevalence of shootings. Certain race pairs have more shootings than others.

Personal Bias

I have a bias for thinking that traditionally unrepresented communities like hispanics, blacks, and women continue to be treated unfairly. But here I tried to be impartial by first analyzing a case where my bias is confirmed—men being the highest perpetrators—and then analyzing a case where we look at race and look for patterns without neglecting the effect it has on the shooting counts.