NYPD Shooting Report

Kristin Kernler

2022-04-04

Import Libraries

```
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Introduction

Project Purpose

This project is an assignment for the course DTSA 5301: Data Science as a Field. We are demonstrating our ability to complete all steps in the data science process by producing a report on the NYPD Shooting Incident data.

Question of Interest

My objective is to determine whether any of the demographics of the victim (age, sex, or race) can be used to predict if a shooting was fatal.

Project Step 1: Describe and Import the Dataset

Data Description

NYPD Shooting Incident Data (Historic)

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.

Source https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Row Description

- Each **row** in this dataset is a **shooting incident**.

Column Description

- **INCIDENT_KEY**: Randomly generated persistent ID for each arrest
- **OCCUR_DATE**: Exact date of shooting incident
- **OCCUR_TIME**: Exact time of the shooting incident
- **BORO**: Borough where the shooting incident occurred
- **STATISTICAL_MURDER_FLAG**: Shooting resulted in the victim's death which would be counted as a murder
- **PERP_AGE_GROUP**: Perpetrator's age within a category
- **PERP_SEX**: Perpetrator's sex description
- **PERP_RACE**: Perpetrator's race description
- **VIC_AGE_GROUP**: Victim's age within a category
- **VIC_SEX**: Victim's sex description
- **VIC_RACE**: Victim's race description

Import Dataset

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"

shootings <- read_csv(url)

glimpse(shootings) # Glimpse prints every column in a data frame.
```

```
## Rows: 23,585
## Columns: 19
## $ INCIDENT_KEY           <dbl> 24050482, 77673979, 203350417, 80584527, 90843~
## $ OCCUR_DATE             <chr> "08/27/2006", "03/11/2011", "10/06/2019", "09/~
## $ OCCUR_TIME             <time> 05:35:00, 12:03:00, 01:09:00, 03:35:00, 21:16~
## $ BORO                   <chr> "BRONX", "QUEENS", "BROOKLYN", "BRONX", "QUEEN~
## $ PRECINCT               <dbl> 52, 106, 77, 40, 100, 67, 77, 81, 101, 106, 71~
## $ JURISDICTION_CODE      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ LOCATION_DESC          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE~
## $ PERP_AGE_GROUP         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PERP_SEX               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PERP_RACE              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ VIC_AGE_GROUP          <chr> "25-44", "65+", "18-24", "<18", "18-24", "<18"~
## $ VIC_SEX                <chr> "F", "M", "F", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE               <chr> "BLACK HISPANIC", "WHITE", "BLACK", "BLACK", "~
## $ X_COORD_CD             <dbl> 1017542, 1027543, 995325, 1007453, 1041267, 10~
## $ Y_COORD_CD             <dbl> 255918.9, 186095.0, 185155.0, 233952.0, 157133~
## $ Latitude               <dbl> 40.86906, 40.67737, 40.67489, 40.80880, 40.597~
## $ Longitude              <dbl> -73.87963, -73.84392, -73.96008, -73.91618, -7~
## $ Lon_Lat                <chr> "POINT (-73.87963173099996 40.86905819000003)"~
```

Step 2: Tidy and Transform Data

Remove Unnecessary Columns

The following columns are not needed: PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat.

```
shootings <- shootings %>% select(-c(
                      PRECINCT,
                      JURISDICTION_CODE,
                      LOCATION_DESC,
                      X_COORD_CD,
                      Y_COORD_CD,
                      Lon_Lat
                      ))
```

Convert Data Types

Convert OCCUR_DATE to **date object**.

```
shootings <- shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Factors are used to work with **categorical variables**.

The following variables should be treated as factors:

- BORO
- PERP_AGE_GROUP
- PERP_SEX
- PERP_RACE
- VIC_AGE_GROUP
- VIC_SEX
- VIC_RACE
- STATISTICAL_MURDER_FLAG

```
shootings$BORO <- factor(shootings$BORO)
shootings$PERP_AGE_GROUP <- factor(shootings$PERP_AGE_GROUP)
shootings$PERP_SEX <- factor(shootings$PERP_SEX)
shootings$PERP_RACE <- factor(shootings$PERP_RACE)
shootings$VIC_AGE_GROUP <- factor(shootings$VIC_AGE_GROUP)
shootings$VIC_SEX <- factor(shootings$VIC_SEX)
shootings$VIC_RACE <- factor(shootings$VIC_RACE)
shootings$STATISTICAL_MURDER_FLAG <- factor(shootings$STATISTICAL_MURDER_FLAG)
```

Summary of Data (Descriptive Statistics)

```
# Descriptive statistics.
summary(shootings)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:23585
##  1st Qu.: 55322804   1st Qu.:2008-12-31   Class1:hms
##  Median : 83435362   Median :2012-02-27   Class2:difftime
##  Mean   :102280741   Mean   :2012-10-05   Mode  :numeric
##  3rd Qu.:150911774   3rd Qu.:2016-03-02
##  Max.   :230611229   Max.   :2020-12-31
```

```
##
##             BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  BRONX        :6701   FALSE:19085             18-24  :5508   F   :  335
##  BROOKLYN     :9734   TRUE : 4500             25-44  :4714   M   :13490
##  MANHATTAN    :2922                           UNKNOWN:3148   U   : 1499
##  QUEENS       :3532                           <18    :1368   NA's: 8261
##  STATEN ISLAND: 696                           45-64  : 495
##                                               (Other):  57
##                                               NA's   :8295
##           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK         :10025   <18    : 2525   F: 2204
##  WHITE HISPANIC: 1988   18-24  : 9003   M:21370
##  UNKNOWN       : 1836   25-44  :10303   U:   11
##  BLACK HISPANIC: 1096   45-64  : 1541
##  WHITE         :  255   65+    :  154
##  (Other)       :  124   UNKNOWN:   59
##  NA's          : 8261
##                             VIC_RACE        Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   :40.51   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      :  327   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :16869   Median :40.70   Median :-73.92
##  BLACK HISPANIC                : 2245   Mean   :40.74   Mean   :-73.91
##  UNKNOWN                       :   65   3rd Qu.:40.82   3rd Qu.:-73.88
##  WHITE                         :  620   Max.   :40.91   Max.   :-73.70
##  WHITE HISPANIC                : 3450
```

Missing Data

```
# Identify columns with missing data and display the number of missing values per column.

colSums(is.na(shootings))
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                       0                        0                     8295
##                PERP_SEX              PERP_RACE           VIC_AGE_GROUP
##                    8261                     8261                        0
##                 VIC_SEX              VIC_RACE                 Latitude
##                       0                        0                        0
##               Longitude
##                       0
```

```
# Total number of missing values.

sum(is.na(shootings))
```

```
## [1] 24817
```

```
# Percentage of missing values.

mean(is.na(shootings))
```

```
## [1] 0.08094128
```

Plan for Missing Data: The columns PERP_SEX, PERP_AGE_GROUP, and PERP_RACE contain missing values, represented by `NA`. I will exclude missing values from mathematical analysis when relevant by including the `na.rm = True` argument. My research focuses on victim demographics so it is unlikely that I will use any columns containing missing values.

Step 3: Add Visualizations and Analysis

Fatal Shootings

The variable **STATISTICAL_MURDER_FLAG** tells whether a shooting was fatal (**FALSE** if non-fatal and **TRUE** if fatal).

I'll start my analysis by counting how many shootings were fatal and how many shootings were non-fatal in the dataset.

```
table(shootings$STATISTICAL_MURDER_FLAG)
```

```
##
## FALSE  TRUE
## 19085  4500
```

At the time of writing this report, there are 19,085 non-fatal shootings and 4,500 fatal shootings.

Victim Age

This frequency table counts the shootings in each age group based on whether a shooting was non-fatal (**False**) or fatal (**True**).

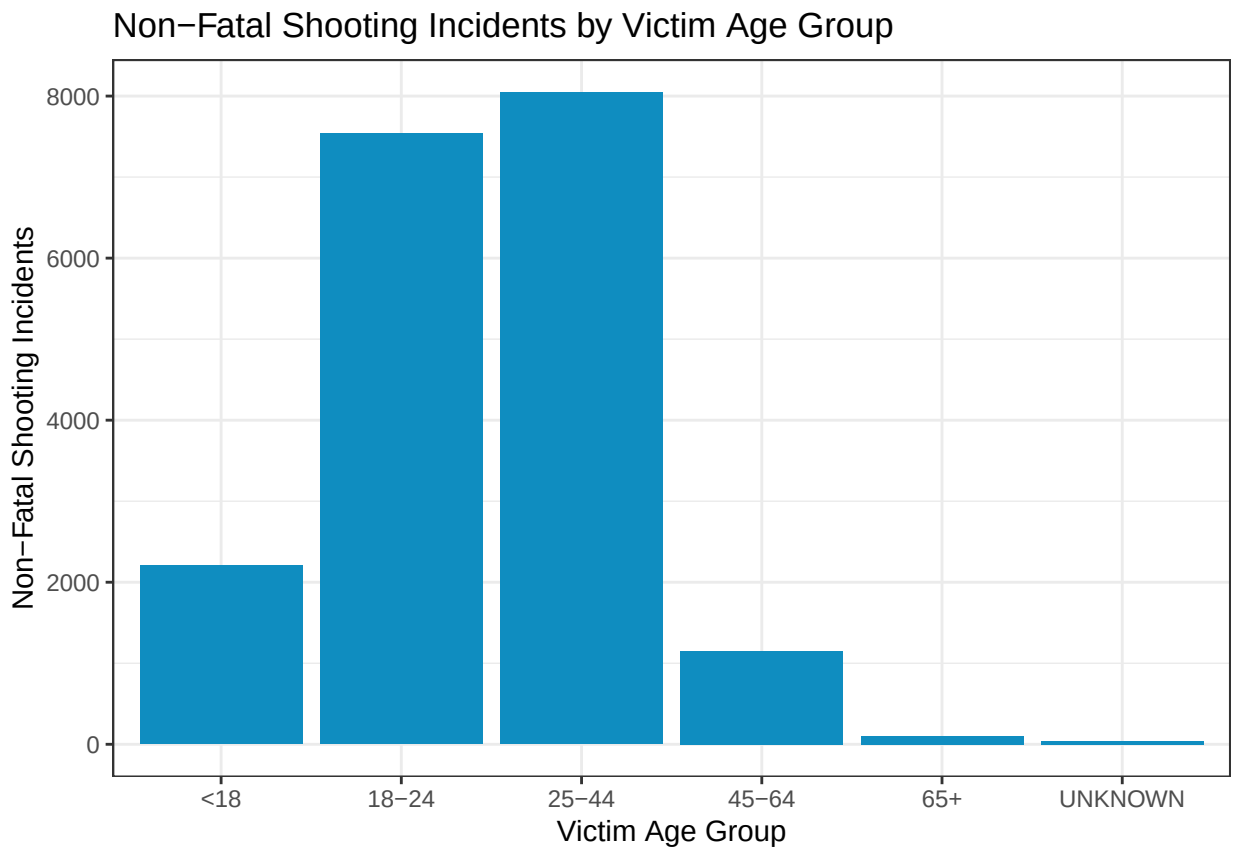
```
table(shootings$STATISTICAL_MURDER_FLAG, shootings$VIC_AGE_GROUP)
```

```
##
##          <18 18-24 25-44 45-64  65+ UNKNOWN
##   FALSE 2205  7537  8046  1151  102      44
##   TRUE   320  1466  2257   390   52      15
```

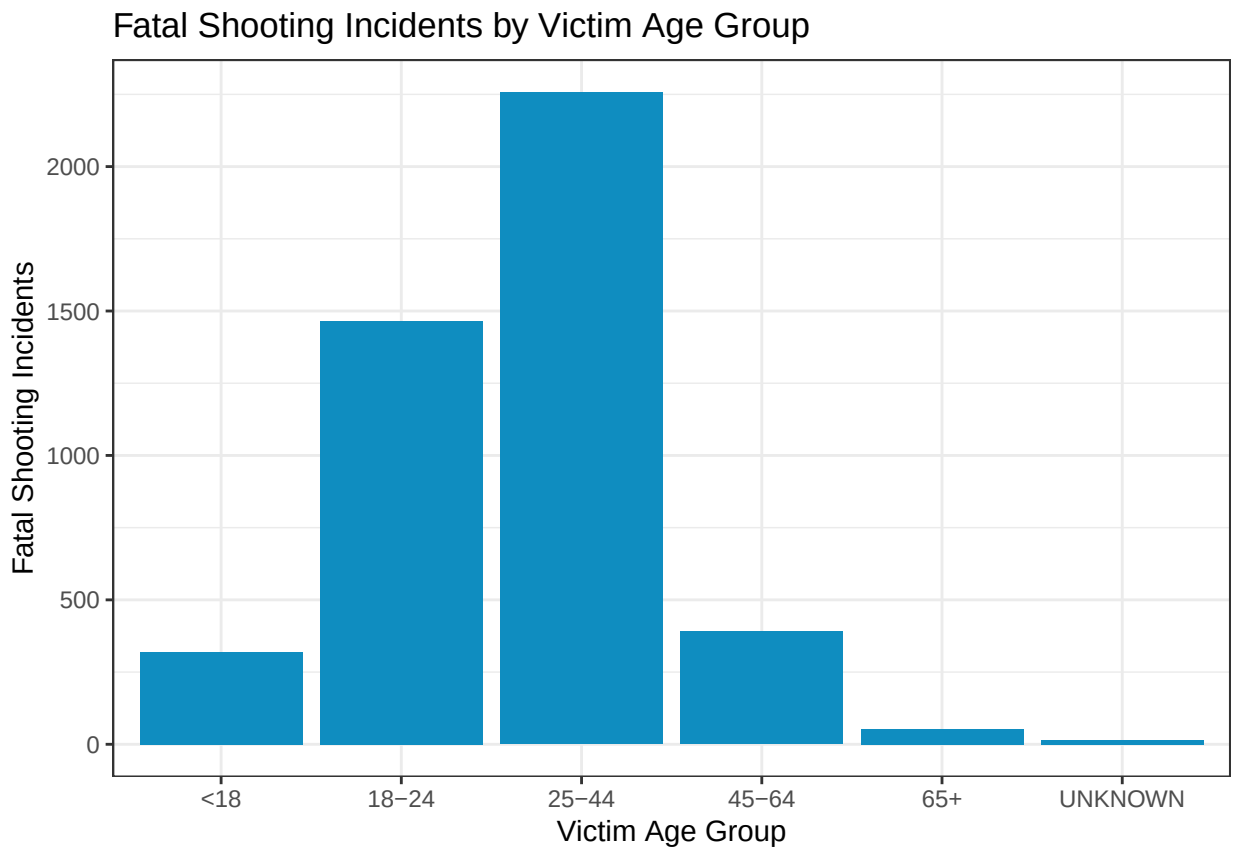
The majority of victims of both fatal and non-fatal shootings are in the 18-24 and 25-44 age groups.

We can use bar charts to compare the distribution of age groups for non-fatal and fatal shootings.

```
shootings %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = VIC_AGE_GROUP)) +
  geom_bar(fill = "#0F8DC0")+
  theme_bw()+
  labs(x = "Victim Age Group",
       y = "Non-Fatal Shooting Incidents",
       title = "Non-Fatal Shooting Incidents by Victim Age Group")
```




```
shootings %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = VIC_AGE_GROUP)) +
  geom_bar(fill = "#0F8DC0")+
  theme_bw()+
  labs(x = "Victim Age Group",
       y = "Fatal Shooting Incidents",
       title = "Fatal Shooting Incidents by Victim Age Group")
```



There are significant differences in these two bar charts. I predict that the victim's age group can be used in determining whether a shooting incident is fatal.

Victim Sex

This frequency table counts the shootings for each sex based on whether a shooting was non-fatal (**False**) or fatal (**True**).

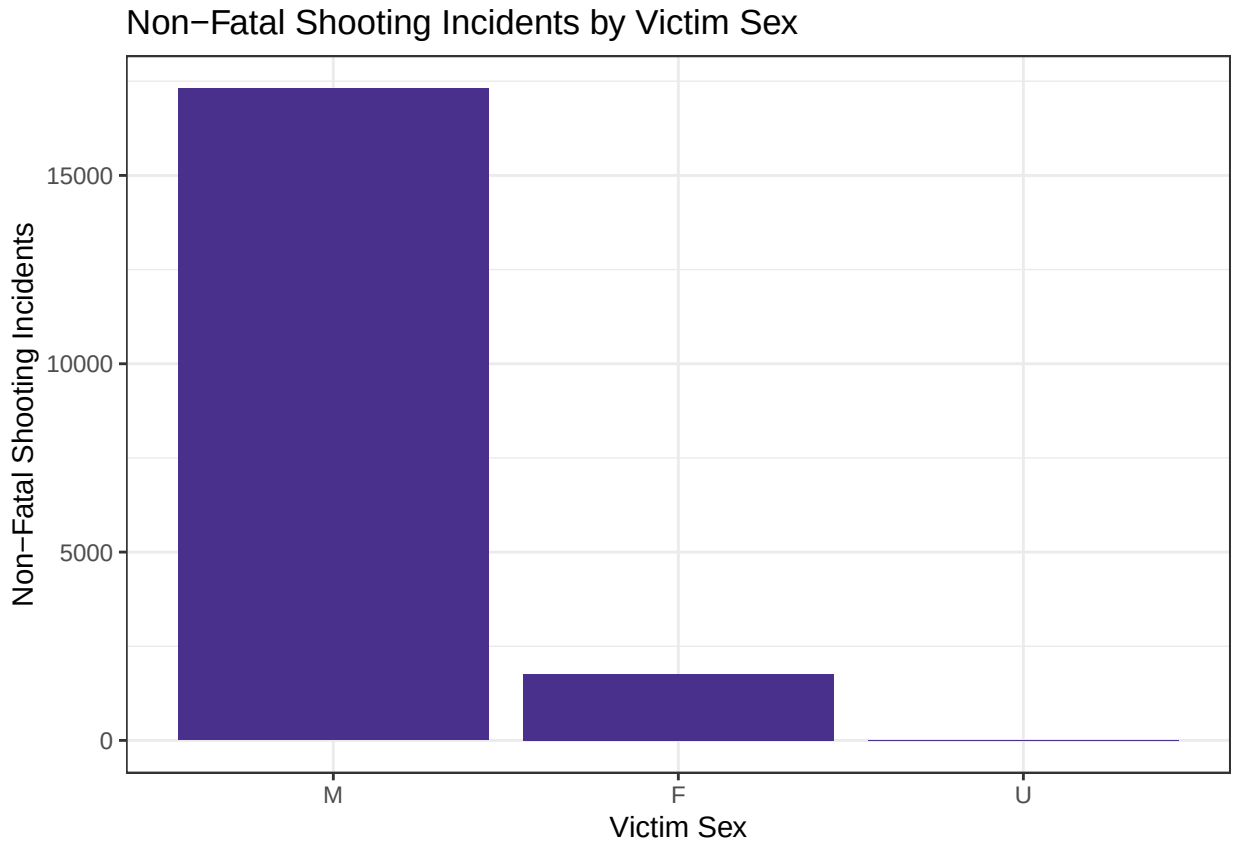
```
table(shootings$STATISTICAL_MURDER_FLAG, shootings$VIC_SEX)
```

```
##
##              F     M     U
##   FALSE  1766 17309    10
##   TRUE    438  4061     1
```

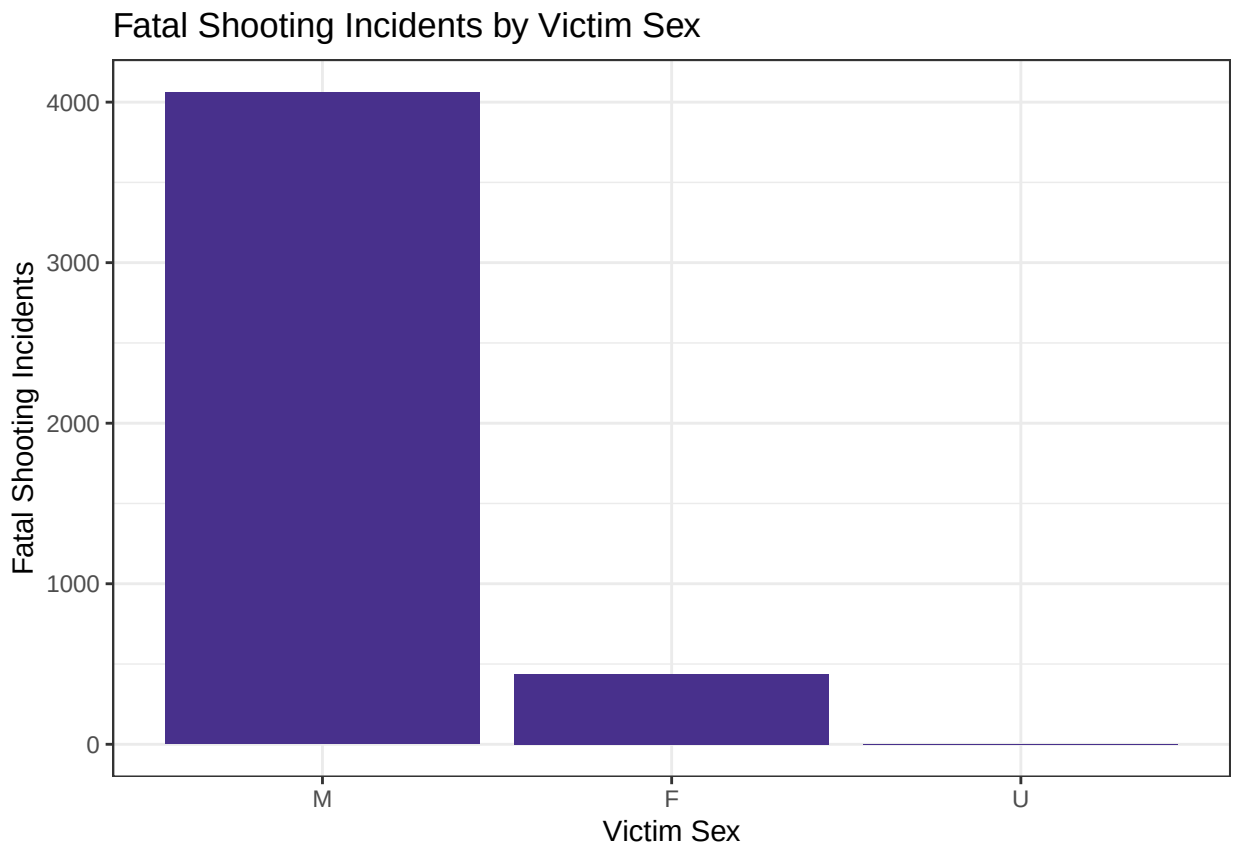
The majority of victims are male, but I would like to visualize this data to see if I am missing anything about female victims.

```
shootings %>%
    filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = fct_infreq(VIC_SEX))) + # Display by frequency.
  geom_bar(stat = 'count')+
  geom_bar(fill = "#48308C")+
  theme_bw()+
  labs(x = "Victim Sex",
```

```
      y = "Non-Fatal Shooting Incidents",
      title = "Non-Fatal Shooting Incidents by Victim Sex")
```




```
shootings %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = fct_infreq(VIC_SEX))) + # Display by frequency.
  geom_bar(stat = 'count')+
  geom_bar(fill = "#48308C")+
  theme_bw()+
  labs(x = "Victim Sex",
       y = "Fatal Shooting Incidents",
       title = "Fatal Shooting Incidents by Victim Sex")
```



There is little difference in the distribution of a victim's sex based on whether a shooting was fatal or non-fatal.

Victim Race

This frequency table counts the shootings for each race based on whether a shooting was non-fatal (**False**) or fatal (**True**).

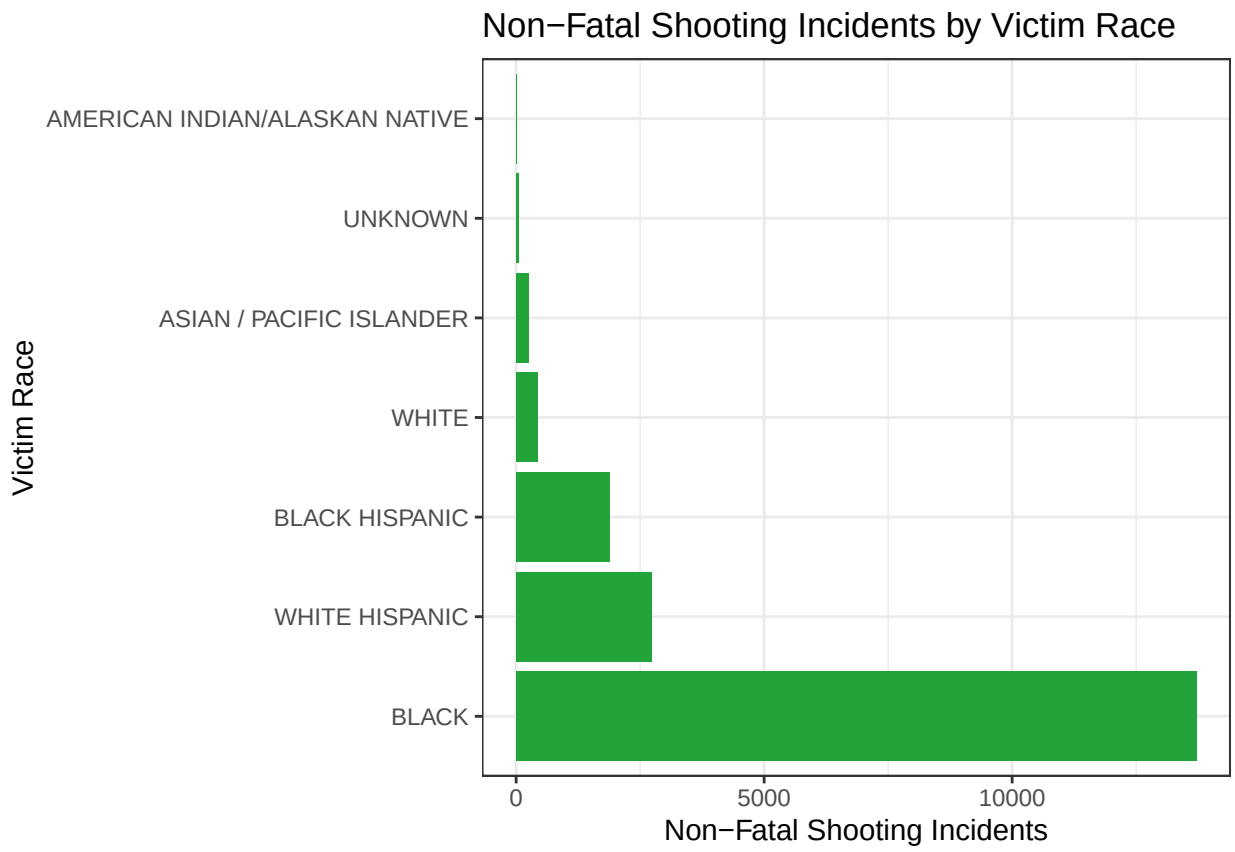
```
table(shootings$STATISTICAL_MURDER_FLAG, shootings$VIC_RACE)
```

```
##
##         AMERICAN INDIAN/ALASKAN NATIVE ASIAN / PACIFIC ISLANDER BLACK
##   FALSE                              9                      244 13714
##   TRUE                               0                       83  3155
##
##         BLACK HISPANIC UNKNOWN WHITE WHITE HISPANIC
##   FALSE           1893      58   442           2725
##   TRUE             352       7   178            725
```

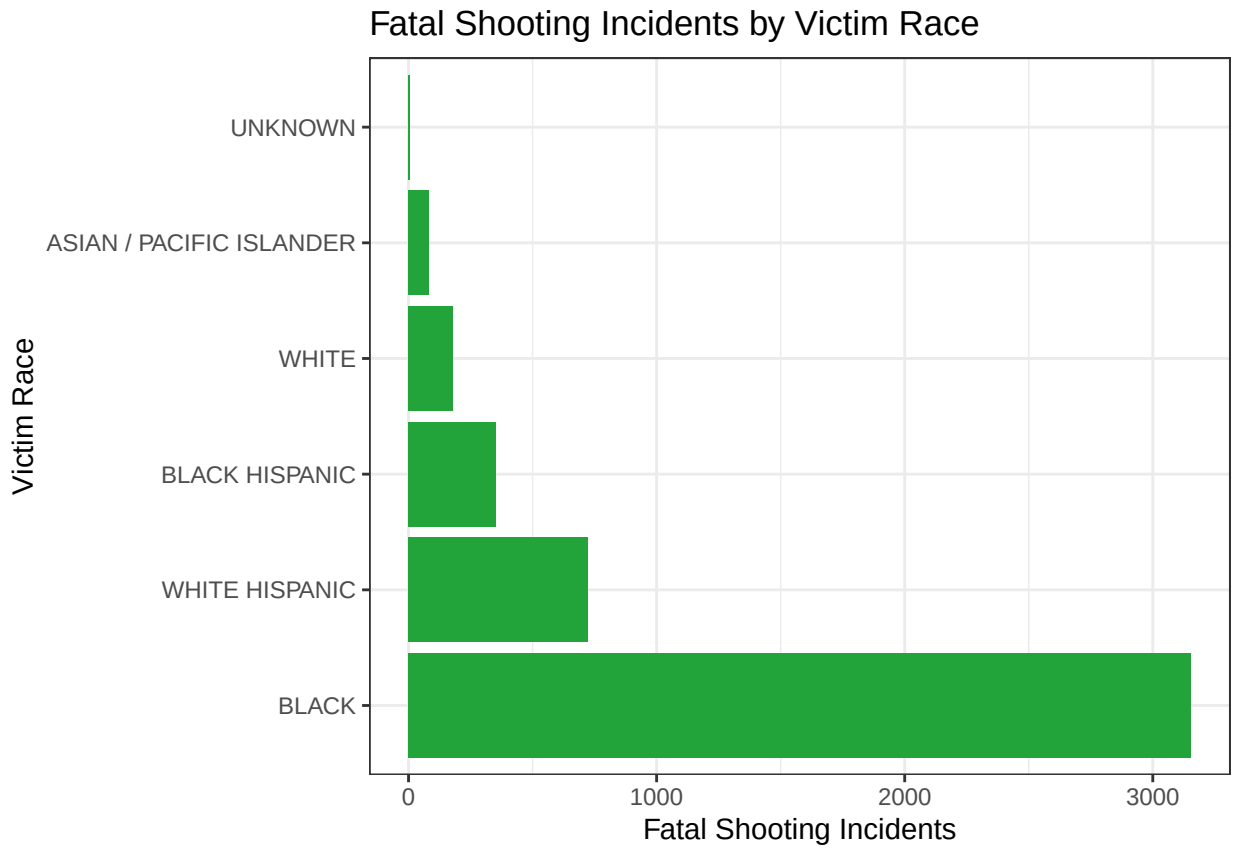
A horizontal bar chart represents this visually.

```
shootings %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = fct_infreq(VIC_RACE))) + # Display by frequency.
  geom_bar(stat = 'count')+
```

```
  geom_bar(fill = "#23A43B")+
  coord_flip()+ # Display race on the y-axis because it looks cluttered on the x-axis.
  theme_bw()+
  labs(x = "Victim Race",
       y = "Non-Fatal Shooting Incidents",
       title = "Non-Fatal Shooting Incidents by Victim Race")
```




```
shootings %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = fct_infreq(VIC_RACE))) + # Display by frequency.
  geom_bar(stat = 'count')+
  geom_bar(fill = "#23A43B")+
  coord_flip()+ # Display race on the y-axis because it looks cluttered on the x-axis.
  theme_bw()+
  labs(x = "Victim Race",
       y = "Fatal Shooting Incidents",
       title = "Fatal Shooting Incidents by Victim Race")
```



Most victims are black, but there are still a significant amount of incidents involving White Hispanic and Black Hispanic victims.

Multivariable Logistic Regression Model

Logistic Regression works well with categorical variables, such as the variables in this dataset that represent a victim's age, sex and race.

My objective is to determine whether any of the demographics of the victim (age, sex, or race) can be used to predict if a shooting is fatal.

Independent Variable: STATISTICAL_MURDER_FLAG

Dependent Variables: VIC_AGE_GROUP, VIC_SEX, VIC_RACE

The variable **STATISTICAL_MURDER_FLAG** indicates whether a shooting was fatal (True is represented by **1** and False is represented by **0**).

```
glm_model <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX + VIC_RACE, data = shootings, family

summary(glm_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX +
##     VIC_RACE, family = "binomial", data = shootings)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0338  -0.6972  -0.5931  -0.5190   2.3350
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -12.90709  107.58068  -0.120  0.90450
## VIC_AGE_GROUP18-24                0.28840    0.06647   4.339 1.43e-05 ***
## VIC_AGE_GROUP25-44                0.64643    0.06460  10.006  < 2e-16 ***
## VIC_AGE_GROUP45-64                0.79971    0.08446   9.468  < 2e-16 ***
## VIC_AGE_GROUP65+                  1.16279    0.18224   6.381 1.76e-10 ***
## VIC_AGE_GROUPUNKNOWN              0.92970    0.31915   2.913  0.00358 **
## VIC_SEXM                         -0.02251    0.05725  -0.393  0.69417
## VIC_SEXU                         -0.58048    1.08474  -0.535  0.59256
## VIC_RACEASIAN / PACIFIC ISLANDER 11.28270  107.58074   0.105  0.91647
## VIC_RACEBLACK                    10.99264  107.58066   0.102  0.91861
## VIC_RACEBLACK HISPANIC           10.78012  107.58068   0.100  0.92018
## VIC_RACEUNKNOWN                  10.27115  107.58148   0.095  0.92394
## VIC_RACEWHITE                    11.39679  107.58070   0.106  0.91563
## VIC_RACEWHITE HISPANIC           11.12689  107.58067   0.103  0.91762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22990  on 23584  degrees of freedom
## Residual deviance: 22706  on 23571  degrees of freedom
## AIC: 22734
##
## Number of Fisher Scoring iterations: 11
```

Observations: The **victim's age group** seems to be a determining factor in whether the victim is likely to survive a shooting. Specifically, a victim is **most likely to survive** a shooting if they are in the **< 18 and 18-24 age groups**. The likelihood of survival **decreases** with each subsequent age group. Most shootings in the **65+ age group** appear to be **fatal**.

Does this raise additional questions that you should investigate?

- Are there other variables that can be used to determine if a shooting is fatal, such as location?
- Are there variables that could be added to this dataset?

Step 4: Report Conclusion and Sources of Bias

Conclusion

I wanted to see whether any of the demographics (age, sex, or race) of a victim could predict whether a shooting was fatal. By modeling the data using Logistic Regression, I discovered that the **victim's age group is significant** in determining whether a victim survived a shooting incident.

Sources of Bias

My political stance on gun control and my ethnicity are sources of bias. When I explored the data, I tried to mitigate my bias by remaining as objective as possible. I didn't make assumptions prior to my analysis.

I let the data speak for itself during the analysis. I found it fairly easy to be objective since we were not provided with much context about the individual incidents.

Resources

- https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic
- https://r4ds.had.co.nz/index.html
- https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8
- https://uc-r.github.io/missing_values
- https://www.geeksforgeeks.org/regression-with-categorical-variables-in-r-programming/