NYC Shooting Data Analysis DTSA_5301

R. Dadmun

2024-01-22

Introduction

This is a breakdown of every shooting incident that occurred in NYC from 2006-2023. 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. Please refer to NYPD Shooting Incident Data (Historic) - CKAN for more information about this dataset.

Purpose

The purpose of this document is to show that I can create an Rmarkdown document to highlight my data analysis process, and provide a guide to reproduce the analysis I performed for DTSA 5301 - Data Science as a Field at the University of Colorado Boulder. This analysis will investigate which age groups, genders and race demographics are most likely to be the victim of a shooting incident in NYC, and which boroughs are most likely to have a shooting incident occur. Finally, we will develop a basic regression model to see if any statistic about the victim can be used to determine if the shooting will be fatal or not.

Import Libraries

The following code will install the tidyverse, lubridate, and the ggthemes library to brighten up our visualizations, and the second line will include it as a library within the Rmarkdown file:

```
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("ggthemes")
library(tidyverse)
library(lubridate)
library(ggthemes)
```

Next, we would like to ensure we have a connection point to our dataset. This dataset was pulled from https://catalog.data.gov as a CSV. We utilize the "read.csv" command to ensure our dataset is imported into R.

```
NYC_Shootings <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLO
head(NYC_Shootings)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
```

```
## 2    137471050 06/27/2014   17:40:00    BRONX                          40
## 3    147998800 11/21/2015   03:56:00   QUEENS                         108
## 4    146837977 10/09/2015   18:30:00    BRONX                          44
## 5     58921844 02/19/2009   22:58:00    BRONX                          47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                          81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                  false
## 2                 0                                                  false
## 3                 0                                                   true
## 4                 0                                                  false
## 5                 0                                                   true
## 6                 0                                                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                     Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

Cleaning the data

After the data has been loaded in, we want to take the time to clean the data. We have removed 9 fields from the data in order to slim the dataset down.

Removed Fields: **PRECINCT**, **JURISDICTION_CODE**, **LOCATION_DESC**, **X_COORD_CD**, **Y_COORD_CD**, **Lat**, **Long**, and **Lon_Lat**.

```
NYC_Shootings_2 <- NYC_Shootings %>%      select(INCIDENT_KEY,
                   OCCUR_DATE,
                   OCCUR_TIME,
                   BORO,
                   STATISTICAL_MURDER_FLAG,
                   PERP_AGE_GROUP,
                   PERP_SEX,
                   PERP_RACE,
                   VIC_AGE_GROUP,
                   VIC_SEX,
                   VIC_RACE)
```

```r
summary(NYC_Shootings_2)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Length:27312            Length:27312       Length:27312
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##   PERP_RACE          VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Observations

Key observations on data type conversion are:

- **INCIDENT_KEY** should be treated as a string.
- **BORO** should be treated as a factor.
- **PERP_AGE_GROUP** should be treated as a factor.
- **PERP_SEX** should be treated as a factor.
- **PERP_RACE** should be treated as a factor.
- **VIC_AGE_GROUP** should be treated as a factor.
- **VIC_SEX** should be treated as a factor.
- **VIC_RACE** should be treated as a factor.

We also ensure our NA factors are removed from the dataset, and every remaining Unknown variable matches the same format. There is a single value in the **VIC_AGE_GROUP** field marked as 1022, which we will move and merge with Unknown.

```r
# Remove NA Values in data
NYC_Shootings_2 <- NYC_Shootings_2 %>%
  replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))

#Change the coding and element types per field
NYC_Shootings_2$VIC_SEX = recode(NYC_Shootings_2$VIC_SEX, U = "Unknown")
NYC_Shootings_2$VIC_RACE = recode(NYC_Shootings_2$VIC_RACE, UNKNOWN = "Unknown")
NYC_Shootings_2$VIC_AGE_GROUP = recode(NYC_Shootings_2$VIC_AGE_GROUP, "1022" = "Unknown")
NYC_Shootings_2$VIC_AGE_GROUP = recode(NYC_Shootings_2$VIC_AGE_GROUP, UNKNOWN = "Unknown")
NYC_Shootings_2$INCIDENT_KEY = as.character(NYC_Shootings_2$INCIDENT_KEY)
```

```
NYC_Shootings_2$BORO = as.factor(NYC_Shootings_2$BORO)
NYC_Shootings_2$PERP_AGE_GROUP = as.factor(NYC_Shootings_2$PERP_AGE_GROUP)
NYC_Shootings_2$PERP_SEX = as.factor(NYC_Shootings_2$PERP_SEX)
NYC_Shootings_2$PERP_RACE = as.factor(NYC_Shootings_2$PERP_RACE)
NYC_Shootings_2$VIC_AGE_GROUP = as.factor(NYC_Shootings_2$VIC_AGE_GROUP)
NYC_Shootings_2$VIC_SEX = as.factor(NYC_Shootings_2$VIC_SEX)
NYC_Shootings_2$VIC_RACE = as.factor(NYC_Shootings_2$VIC_RACE)
NYC_Shootings_2$STATISTICAL_MURDER_FLAG <- factor(NYC_Shootings_2$STATISTICAL_MURDER_FLAG)

# Return summary statistics
summary(NYC_Shootings_2)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                    BORO
##  Length:27312       Length:27312       Length:27312       BRONX        : 7937
##  Class :character   Class :character   Class :character   BROOKLYN     :10933
##  Mode  :character   Mode  :character   Mode  :character   MANHATTAN    : 3572
##                                                           QUEENS       : 4094
##                                                           STATEN ISLAND:  776
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX              PERP_RACE
##  false:22046                    :9344         : 9310   BLACK         :11432
##  true : 5266             18-24  :6222   (null):  640                 : 9310
##                          25-44  :5687   F     :  424   WHITE HISPANIC: 2341
##                          UNKNOWN:3148   M     :15439   UNKNOWN       : 1836
##                          <18    :1591   U     : 1499   BLACK HISPANIC: 1314
##                          (null) : 640                  (null)        :  640
##                          (Other): 680                  (Other)       :  439
##  VIC_AGE_GROUP      VIC_SEX                                VIC_RACE
##  <18    : 2839   F       : 2615   AMERICAN INDIAN/ALASKAN NATIVE:   10
##  18-24  :10086   M       :24686   ASIAN / PACIFIC ISLANDER      :  404
##  25-44  :12281   Unknown:   11   BLACK                         :19439
##  45-64  : 1863                  BLACK HISPANIC                : 2646
##  65+    :  181                  Unknown                       :   66
##  Unknown:   62                  WHITE                         :  698
##                                 WHITE HISPANIC                : 4049
```

Checking for additional missing data:

We will locate where data is missing within the data using the colSums function. Next, we will check the total number of missing data elements (which are marked as NA), and the percent of the data which is missing:

```
colSums(is.na(NYC_Shootings_2))
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                       0                        0                        0
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                       0                        0                        0
##                 VIC_SEX               VIC_RACE
```

```
##                         0                         0
```

As we can see, our data is clean with no NA entries.

Research Questions:

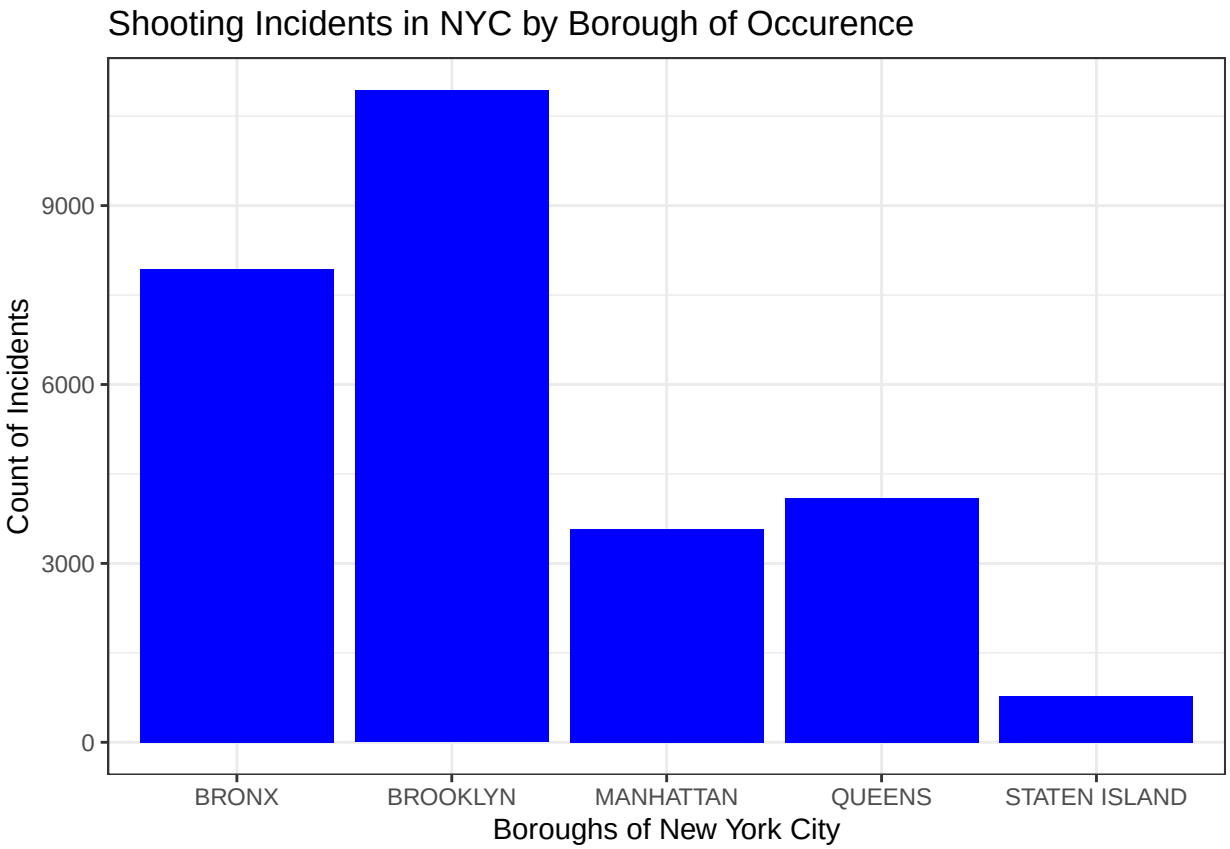
1. Which part of New York has the most number of incidents? Of those incidents, how many are murder cases?
2. Which groups of people were most likely to be victims of a shooting in NYC?

Visualizing and Analyzing the Data

1. Which part of New York has the most number of incidents? Of those incidents, how many are murder cases?

```
##
##                 false true
##   BRONX          6395 1542
##   BROOKLYN       8811 2122
##   MANHATTAN      2942  630
##   QUEENS         3284  810
##   STATEN ISLAND   614  162
```

```
##
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          7937         10933          3572          4094           776
```



Based on the graph, we can see that Brooklyn is 1st in the number of incidents, with 10,932 incidents, 2,122 of which were flagged as murders. The Bronx is significantly lower than this, with 7,935 incidents, of which 1,542 were flagged as murders. Queens then follows, with 4,094 incidents. Both Manhattan and Staten Island both have a significantly smaller number of incidents than Brooklyn, with Staten Island having both their number of incidents and the flags for murder numbering in the triple digits.

2. Which groups of people were most likely to be victims of a fatal shooting in NYC?

Which age group is most likely to be the victim of a shooting incident in NYC? First, we want to see how many of our shootings were fatal, which is summarized in the STATISTICAL_MURDER_FLAG field:

```
Fatal_Tbl <- table(NYC_Shootings_2$STATISTICAL_MURDER_FLAG)
Fatal_Tbl
```

```
##
## false  true
## 22046  5266
```

At the time of this analysis, we can see that there are 22,046 non-fatal shootings in the dataset, and 5,266 fatal shootings.

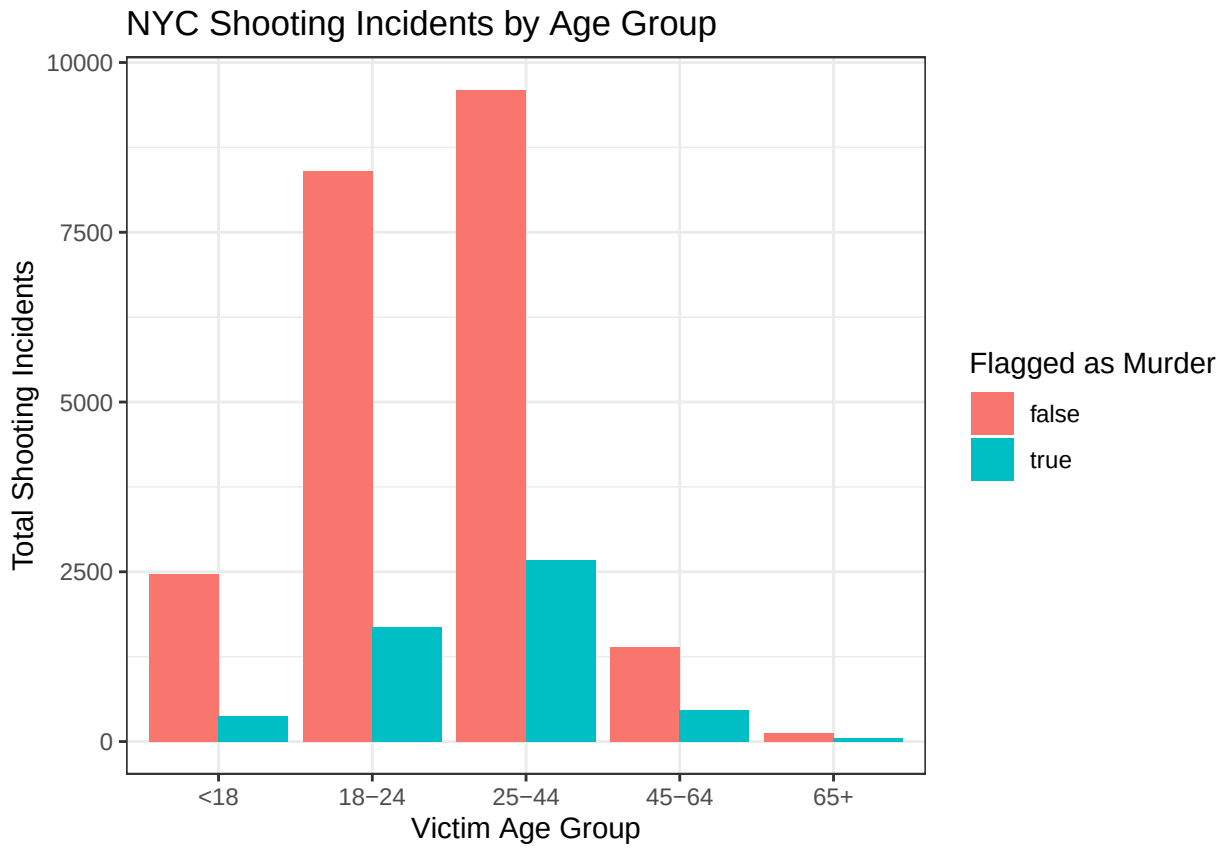
Next, we would like to check how our fatal shootings are distributed by victim age:

```
Fatal_Age_Tbl <- table(NYC_Shootings_2$STATISTICAL_MURDER_FLAG, NYC_Shootings_2$VIC_AGE_GROUP)
Fatal_Age_Tbl
```

```
##
##           <18 18-24 25-44 45-64  65+ Unknown
##   false 2469  8406  9601  1398  125      47
##   true   370  1680  2680   465   56      15
```

From the table above, we can see that the vast majority of victims of shooting incidents were in the 18-24 and 25-44 age groups. We will utilize a bar chart to compare this distributions in a more visual way:

```
NYC_Shootings_2 %>%
  filter(VIC_AGE_GROUP != "Unknown") %>%
  ggplot(aes(x = VIC_AGE_GROUP, fill = (STATISTICAL_MURDER_FLAG))) +
  geom_bar(position = "dodge") +
  theme_bw() +
  labs(title = "NYC Shooting Incidents by Age Group",
      x = "Victim Age Group",
      y = "Total Shooting Incidents",
      fill = "Flagged as Murder")
```



Based on the table above, I hypothesize that age group is correlated with the STATISTICAL_MURDER_FLAG field.

Which gender is more likely to be a victim of a shooting incident in NYC? Next, we will run a quick analysis on which gender is more likely toi be the victim of a shooting incident in NYC. In our data,

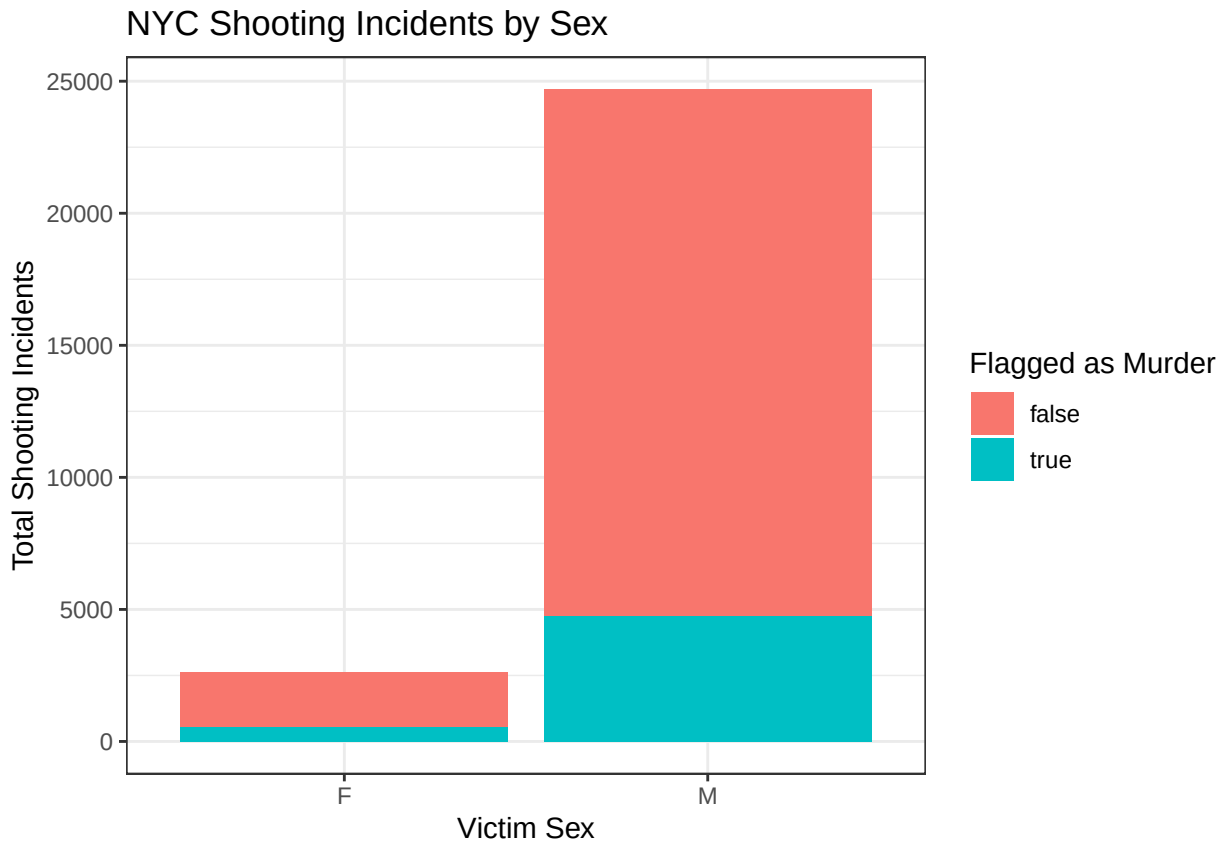
genbder is marked by the VIC_SEX field, and is given values of Male, Female or Unknown. We will make a table to list out the exact count of each, and then create a graph to better visualize this data. The graph will also show the value differences between fatal and non-fatal shooting incidents.

```
table(NYC_Shootings_2$VIC_SEX)
```

```
##
##        F        M Unknown
##     2615    24686       11
```

```
Gender_Graph_Stack <- NYC_Shootings_2 %>%
  filter(VIC_SEX != "Unknown") %>%
  ggplot(aes(x = VIC_SEX, fill = (STATISTICAL_MURDER_FLAG))) +
  geom_bar() +
  theme_bw() +
  labs(title = "NYC Shooting Incidents by Sex",
       x = "Victim Sex",
       y = "Total Shooting Incidents",
       fill = "Flagged as Murder")

Gender_Graph_Stack
```



Based on the above graphs, we can see that Men are unilaterally more likely to be the victim of a shooting even in NYC than Women.

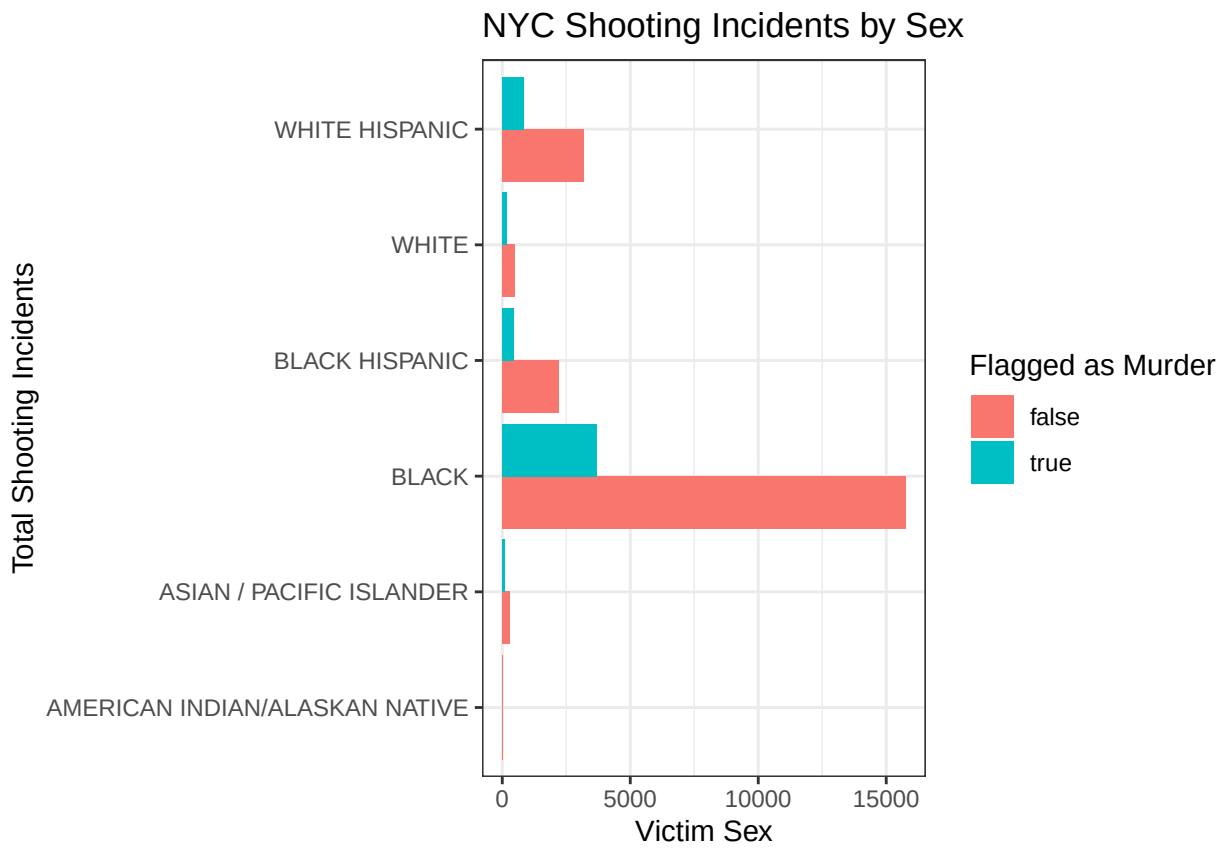
Which demographic race is most likely to be the victim of a shooting event in NYC? Finally, we would like to explore the rates of shooting incidents in NYC by victim race.

```
table(NYC_Shootings_2$VIC_RACE)
```

```
##
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                             10                            404
##                          BLACK                 BLACK HISPANIC
##                          19439                           2646
##                        Unknown                          WHITE
##                             66                            698
##                 WHITE HISPANIC
##                           4049
```

```
Race_Graph <- NYC_Shootings_2 %>%
  filter(VIC_RACE != "Unknown") %>%
  ggplot(aes(y = VIC_RACE, fill = (STATISTICAL_MURDER_FLAG))) +
  geom_bar(position = "dodge") +
  theme_bw() +
  labs(title = "NYC Shooting Incidents by Sex",
      x = "Victim Sex",
      y = "Total Shooting Incidents",
      fill = "Flagged as Murder")

Race_Graph
```



Based on the graphs above, we have a few major takeaways to flag for future analysis:
1. Brooklyn and the Bronx are by far the most likely boroughs in which a shooting incident may occur.
2. Men are far more likely to be the victim of a shooting incident than women.
3. Black individuals are at a much higher likelihood to be the victim of a shooting incident in NYC. Both White Hispanic and Black Hispanic individuals follow in second and third place as the most likely demographic to be the victim of a shooting event.

Summarized in a sentence: Black or Hispanic males in Brooklyn or the Bronx have an elevated chance to be the victim of a shooting event as opposed to other population demographics or locations in NYC

Multivariable Logistic Regression Model

Logistic regression models work well with a plethora of categorical variables, of which this data set has many. In our model, we will be utilizing the victim's age, gender and race.

The objective of the model is to determine if any of the aforementioned variables can be used to predict if a shooting will be fatal or not.

Independent Variable STATISTICAL_MURDER_FLAG This variable indicates whether a shooting was fatal or not. True (1) signals that the shooting was fatal, while False (0) indicates that the shooting was not fatal. **Dependent Variable** VIC_AGE_GROUP, VIC_RACE, VIC_SEX

```
Reg_Model <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX + VIC_RACE, data = NYC_Shootings_2, f

summary(Reg_Model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX +
##     VIC_RACE, family = "binomial", data = NYC_Shootings_2)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -12.86405  102.16037  -0.126  0.89980
## VIC_AGE_GROUP18-24                0.28557    0.06197   4.608 4.06e-06 ***
## VIC_AGE_GROUP25-44                0.61258    0.06005  10.200  < 2e-16 ***
## VIC_AGE_GROUP45-64                0.75933    0.07781   9.759  < 2e-16 ***
## VIC_AGE_GROUP65+                  1.01911    0.17146   5.944 2.79e-09 ***
## VIC_AGE_GROUPUnknown              0.85023    0.31531   2.696  0.00701 **
## VIC_SEXM                         -0.04778    0.05206  -0.918  0.35869
## VIC_SEXUnknown                   -0.58211    1.08249  -0.538  0.59075
## VIC_RACEASIAN / PACIFIC ISLANDER  11.28121  102.16041   0.110  0.91207
## VIC_RACEBLACK                     11.00307  102.16035   0.108  0.91423
## VIC_RACEBLACK HISPANIC            10.82209  102.16036   0.106  0.91564
## VIC_RACEUnknown                   10.26523  102.16120   0.100  0.91996
## VIC_RACEWHITE                     11.34289  102.16038   0.111  0.91159
## VIC_RACEWHITE HISPANIC            11.12441  102.16035   0.109  0.91329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26781  on 27311  degrees of freedom
```

```
## Residual deviance: 26504  on 27298  degrees of freedom
## AIC: 26532
##
## Number of Fisher Scoring iterations: 11
```

Observations: The **victim's age group** seems to be the most significant variable for determining how likely a victim is to survive a shooting incident in NYC. More specifically, a victim is **most likely to survive the incident** if they are in the **< 18 or 18-24 age groups**. As the age increases, the liklihood of survival diminished. With a coefficient greater than 1 for the **65+ age group**, it appears that most shooting events in this age group are fatal.

Both **Victim Age and Victim Race** appear to be uncorrelated in the determination of whether a shooting event in NYC is fatal to the victim. No singular race demographic had a significant correlation with the determination as to if shooting incidents in NYC ended in murder; nor did the sex of the victim have aa significant correlation value to whether or not a murder occured.

Identifing Bias

The topics of gun violence and gender are both sources of bias for me. I assumed initially that women would more likely be the incident of crime, and tried to ensure I approached this topic as neutrally as possible to ensure I was ready for the takeaways the data showed me. Additionally, my political stances on gun ownership and gun violence would have lead me to believe most gun crimes end in death. Overall, I found it relitively easy to remain objective in my analysis as there was little context or discussion surrounding this dataset prior to exploration.

##Resources - https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic
- https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/about_data
- https://www.vitalcitynyc.org/vital_signs/gun-violence-in-new-york-city-the-data
- https://uc-r.github.io/missing_values
- https://www.codecademy.com/learn/learn-linear-regression-in-r/modules/linear-regression-in-r/cheatsheet
- https://r4ds.had.co.nz/index.html