NYPD_Shooting_Incidents_Report

Daniel Mandel

2023-11-21

Introduction

Purpose

This assignment will show my ability to complete all steps in the data science process in a reproducible manner by producing a report on the NYPD Shooting Incident Data (Historic).

Question

I want to determine if a demographic (age or race) as well as location are good indicators to determine whether a shooting incident was fatal or not. Will my models be able to predict this?

Description of Data

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.

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

Step 1:

Import Libraries

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

Step 2:

Import the Dataset

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

# Read data from url
shootings_data <- read_csv(url_in)
```

```r
# Display every column
glimpse(shootings_data)
```

```
## Rows: 27,312
## Columns: 21
## $ INCIDENT_KEY            <dbl> 228798151, 137471050, 147998800, 146837977, 58~
## $ OCCUR_DATE              <chr> "05/27/2021", "06/27/2014", "11/21/2015", "10/~
## $ OCCUR_TIME              <time> 21:30:00, 17:40:00, 03:56:00, 18:30:00, 22:58~
## $ BORO                    <chr> "QUEENS", "BRONX", "QUEENS", "BRONX", "BRONX",~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PRECINCT                <dbl> 105, 40, 108, 44, 47, 81, 114, 81, 105, 101, 2~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "MULTI DWE~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, ~
## $ PERP_AGE_GROUP          <chr> NA, NA, NA, NA, "25-44", NA, NA, NA, NA, "25-4~
## $ PERP_SEX                <chr> NA, NA, NA, NA, "M", NA, NA, NA, NA, "M", NA, ~
## $ PERP_RACE               <chr> NA, NA, NA, NA, "BLACK", NA, NA, NA, NA, "BLAC~
## $ VIC_AGE_GROUP           <chr> "18-24", "18-24", "25-44", "<18", "45-64", "25~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "WHITE", "WHITE HISPANIC", "~
## $ X_COORD_CD              <dbl> 1058925.0, 1005028.0, 1007667.9, 1006537.4, 10~
## $ Y_COORD_CD              <dbl> 180924.0, 234516.0, 209836.5, 244511.1, 262189~
## $ Latitude                <dbl> 40.66296, 40.81035, 40.74261, 40.83778, 40.886~
## $ Longitude               <dbl> -73.73084, -73.92494, -73.91549, -73.91946, -7~
## $ Lon_Lat                 <chr> "POINT (-73.73083868899994 40.662964620000025)~
```

Step 3:

Tidying and Transforming Dataset

Remove any unnecessary columns from the dataset (anything not related to demographics, borough or if the shooting was fatal).

Create factors as they are used to work with categorical variables and regression later.

Remove NA values from the records.

Source https://r4ds.had.co.nz/factors.html

I do not these columns: INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, LOC_CLASSFCTN_DESC, LOC_OF_OCCUR_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, PERP_SEX, VIC_SEX.

```r
# Drop the columns
shootings_data <- shootings_data %>% select(-c(
                      INCIDENT_KEY,
                      OCCUR_TIME,
                      OCCUR_DATE,
                      PRECINCT,
                      JURISDICTION_CODE,
                      LOCATION_DESC,
                      LOC_CLASSFCTN_DESC,
                      LOC_OF_OCCUR_DESC,
                      X_COORD_CD,
                      Y_COORD_CD,
                      Latitude,
```

```
                    Longitude,
                    Lon_Lat,
                    PERP_SEX,
                    VIC_SEX
                  ))


# Treat categorical variables as factors to be used in regression analysis
shootings_data$BORO <- as.factor(shootings_data$BORO)
shootings_data$PERP_AGE_GROUP <- as.factor(shootings_data$PERP_AGE_GROUP)
shootings_data$PERP_RACE <- as.factor(shootings_data$PERP_RACE)
shootings_data$VIC_AGE_GROUP <- as.factor(shootings_data$VIC_AGE_GROUP)
shootings_data$VIC_RACE <- as.factor(shootings_data$VIC_RACE)
shootings_data$STATISTICAL_MURDER_FLAG <- as.factor(shootings_data$STATISTICAL_MURDER_FLAG)

# Lets remove any data that is incomplete (has NA) as that will not be useful for our analysis
shootings_data <- shootings_data %>% drop_na()

# Display the summary
summary(shootings_data)
```

```
##             BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  BRONX        :5425   FALSE:14408             18-24  :6222
##  BROOKLYN     :6642   TRUE : 3560             25-44  :5687
##  MANHATTAN    :2542                           UNKNOWN:3148
##  QUEENS       :2728                           <18    :1591
##  STATEN ISLAND: 631                           (null) : 640
##                                               45-64  : 617
##                                               (Other):  63
##           PERP_RACE     VIC_AGE_GROUP                              VIC_RACE
##  BLACK         :11432   <18    :2027   AMERICAN INDIAN/ALASKAN NATIVE:    8
##  WHITE HISPANIC: 2341   1022   :   1   ASIAN / PACIFIC ISLANDER      :  307
##  UNKNOWN       : 1802   18-24  :6518   BLACK                         :12252
##  BLACK HISPANIC: 1314   25-44  :7939   BLACK HISPANIC                : 1800
##  (null)        :  640   45-64  :1290   UNKNOWN                       :   48
##  WHITE         :  283   65+    : 137   WHITE                         :  552
##  (Other)       :  156   UNKNOWN:  56   WHITE HISPANIC                : 3001
```

Step 4:

Visualizing, Analyzing, and Modeling Data

Create tables to display the breakdown of shootings and fatal shootings by race and age.

```
# Breakdown the victims by race
table(shootings_data$VIC_RACE,
      shootings_data$STATISTICAL_MURDER_FLAG
)
```

```
##
##                                 FALSE TRUE
##   AMERICAN INDIAN/ALASKAN NATIVE     8    0
##   ASIAN / PACIFIC ISLANDER         228   79
##   BLACK                           9901 2351
##   BLACK HISPANIC                  1492  308
```

```
##   UNKNOWN                           42    6
##   WHITE                            398  154
##   WHITE HISPANIC                  2339  662
```

```
# Breakdown the perpetrators by race
table(shootings_data$PERP_RACE,
      shootings_data$STATISTICAL_MURDER_FLAG
)
```

```
##
##                                  FALSE TRUE
##   (null)                           545   95
##   AMERICAN INDIAN/ALASKAN NATIVE     2    0
##   ASIAN / PACIFIC ISLANDER         106   48
##   BLACK                           9053 2379
##   BLACK HISPANIC                  1054  260
##   UNKNOWN                         1699  103
##   WHITE                            173  110
##   WHITE HISPANIC                  1776  565
```

```
# Breakdown the victims by age group
table(shootings_data$VIC_AGE_GROUP,
      shootings_data$STATISTICAL_MURDER_FLAG
)
```

```
##
##           FALSE TRUE
##   <18      1746  281
##   1022        1    0
##   18-24    5356 1162
##   25-44    6195 1744
##   45-64     971  319
##   65+        95   42
##   UNKNOWN    44   12
```

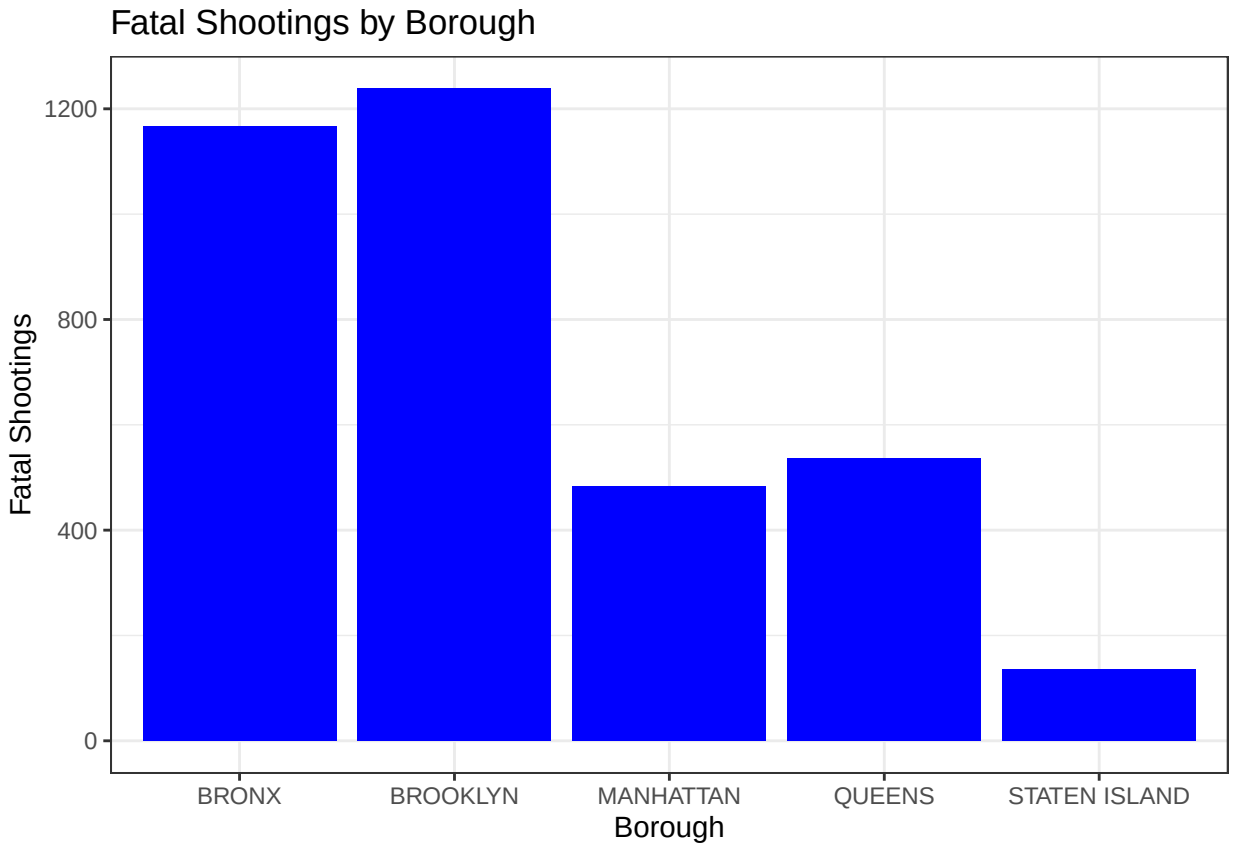
```
# Breakdown the perpetrators by age group
table(shootings_data$PERP_AGE_GROUP,
      shootings_data$STATISTICAL_MURDER_FLAG
)
```

```
##
##           FALSE TRUE
##   (null)    545   95
##   <18      1304  287
##   1020        1    0
##   18-24    4920 1302
##   224         1    0
##   25-44    4163 1524
##   45-64     399  218
##   65+        35   25
##   940         1    0
##   UNKNOWN  3039  109
```

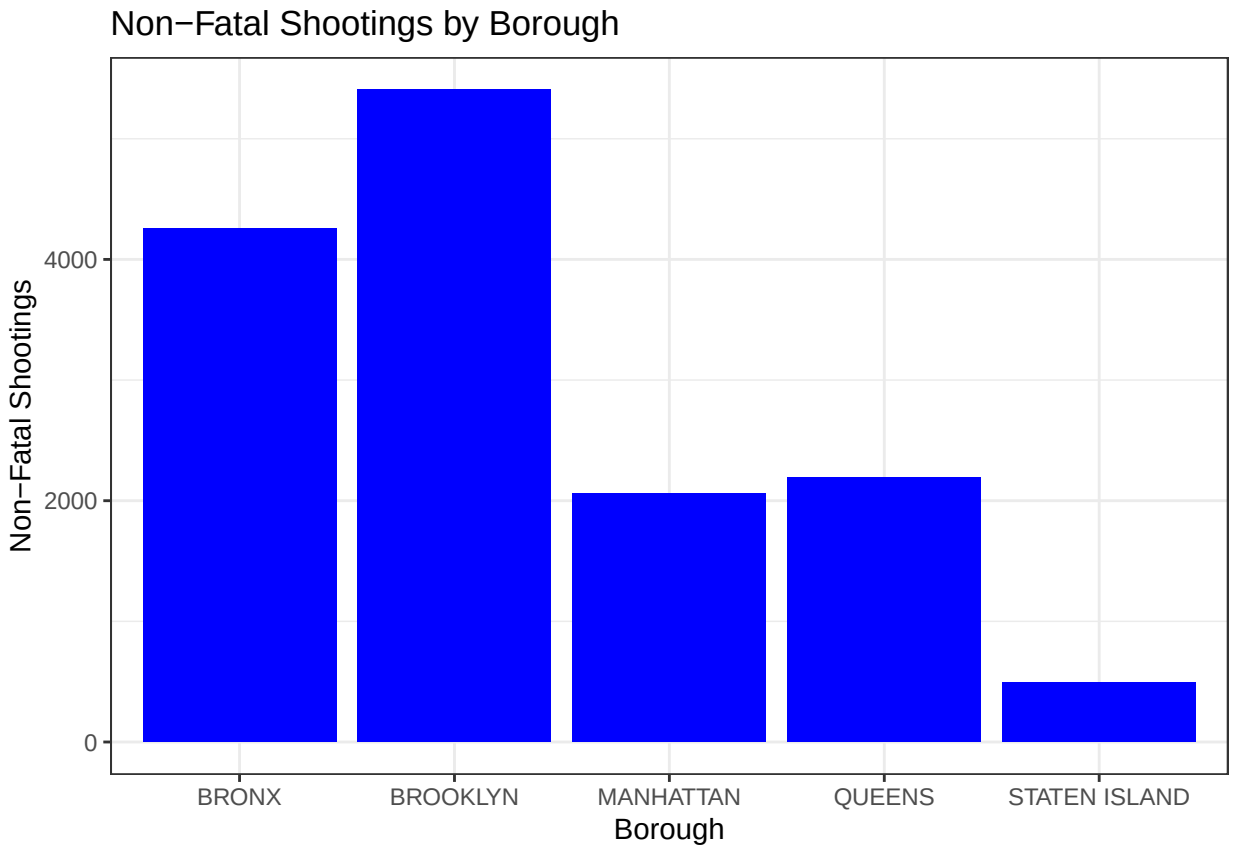
Create charts to compare as well

```
# Create charts to show the distribution of fatal shootings by age, race and borough.
shootings_data %>%
```

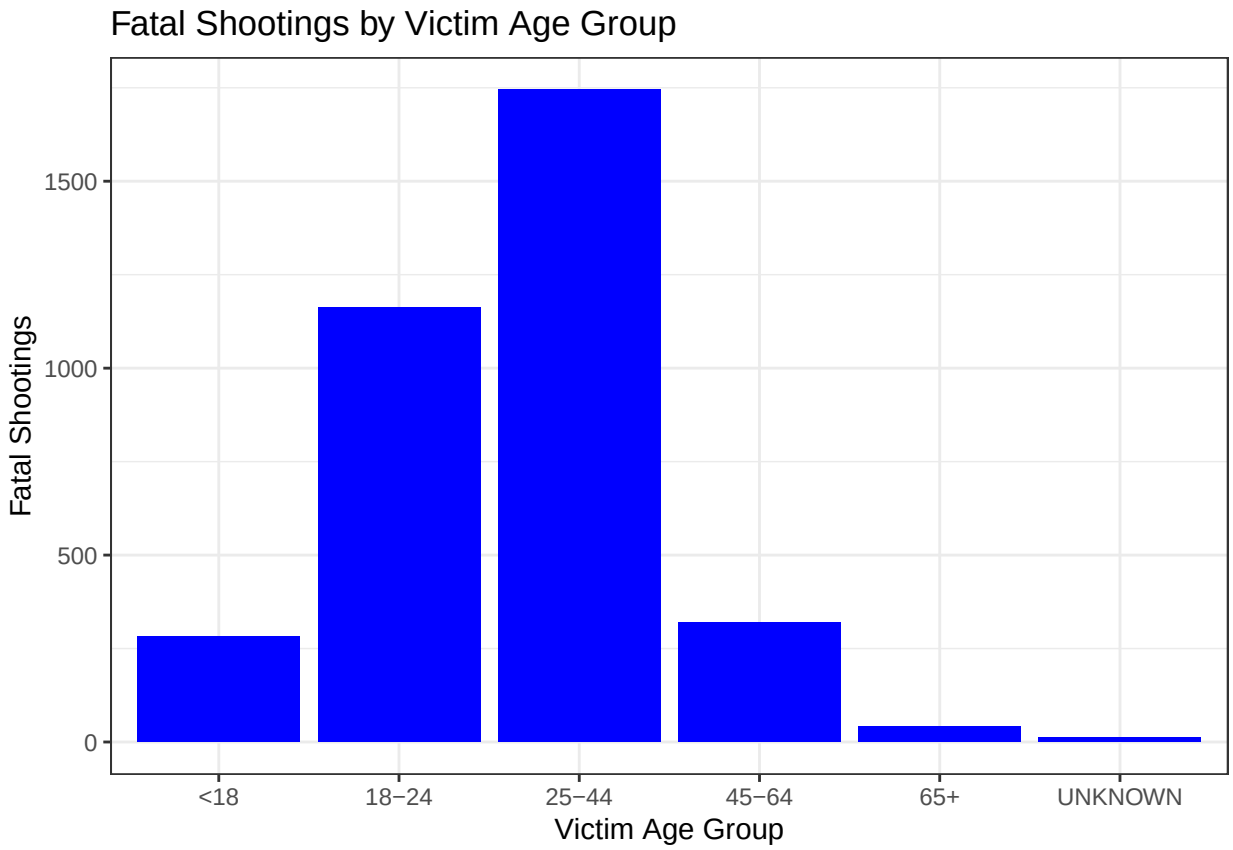
```
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = BORO)) +
  geom_bar(fill = "blue")+
  theme_bw()+
  labs(x = "Borough",
       y = "Fatal Shootings",
       title = "Fatal Shootings by Borough")
```



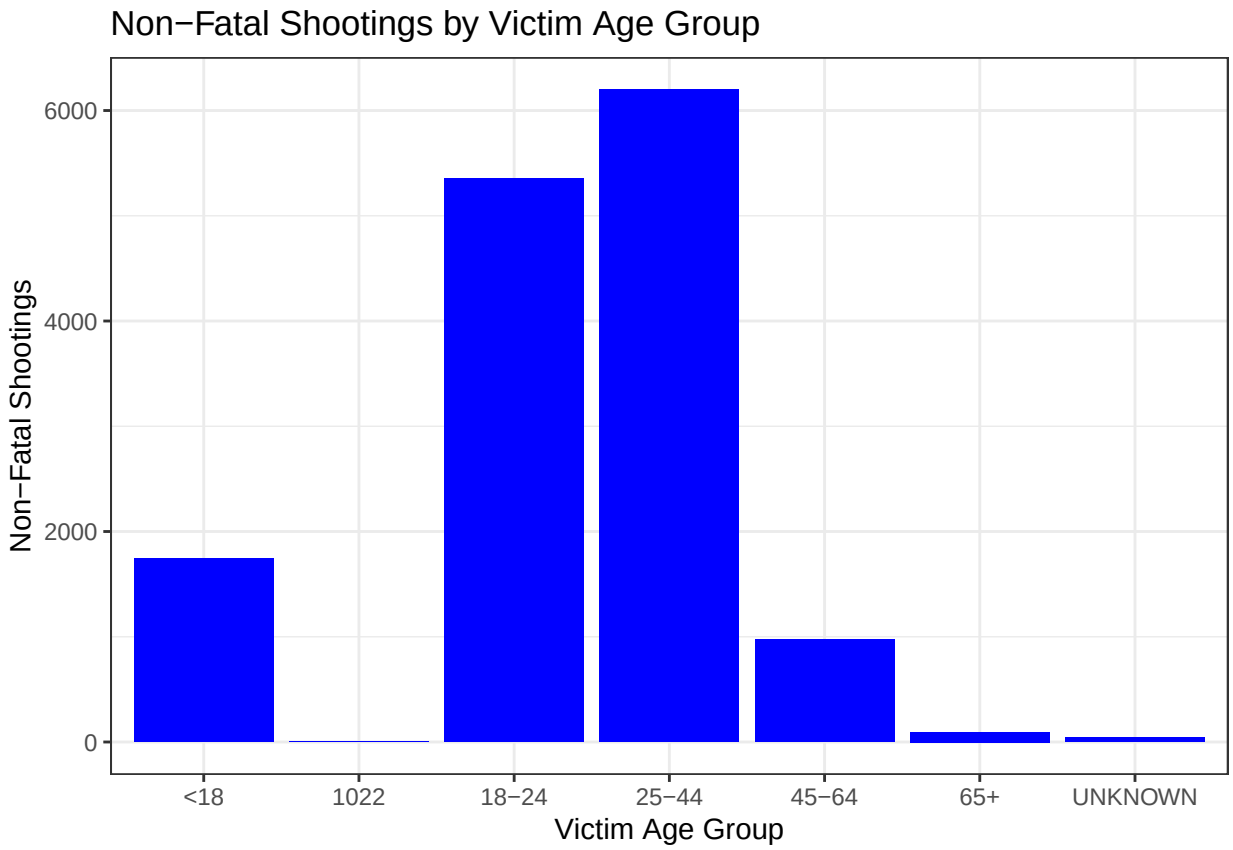

```
shootings_data %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = BORO)) +
  geom_bar(fill = "blue")+
  theme_bw()+
  labs(x = "Borough",
       y = "Non-Fatal Shootings",
       title = "Non-Fatal Shootings by Borough")
```



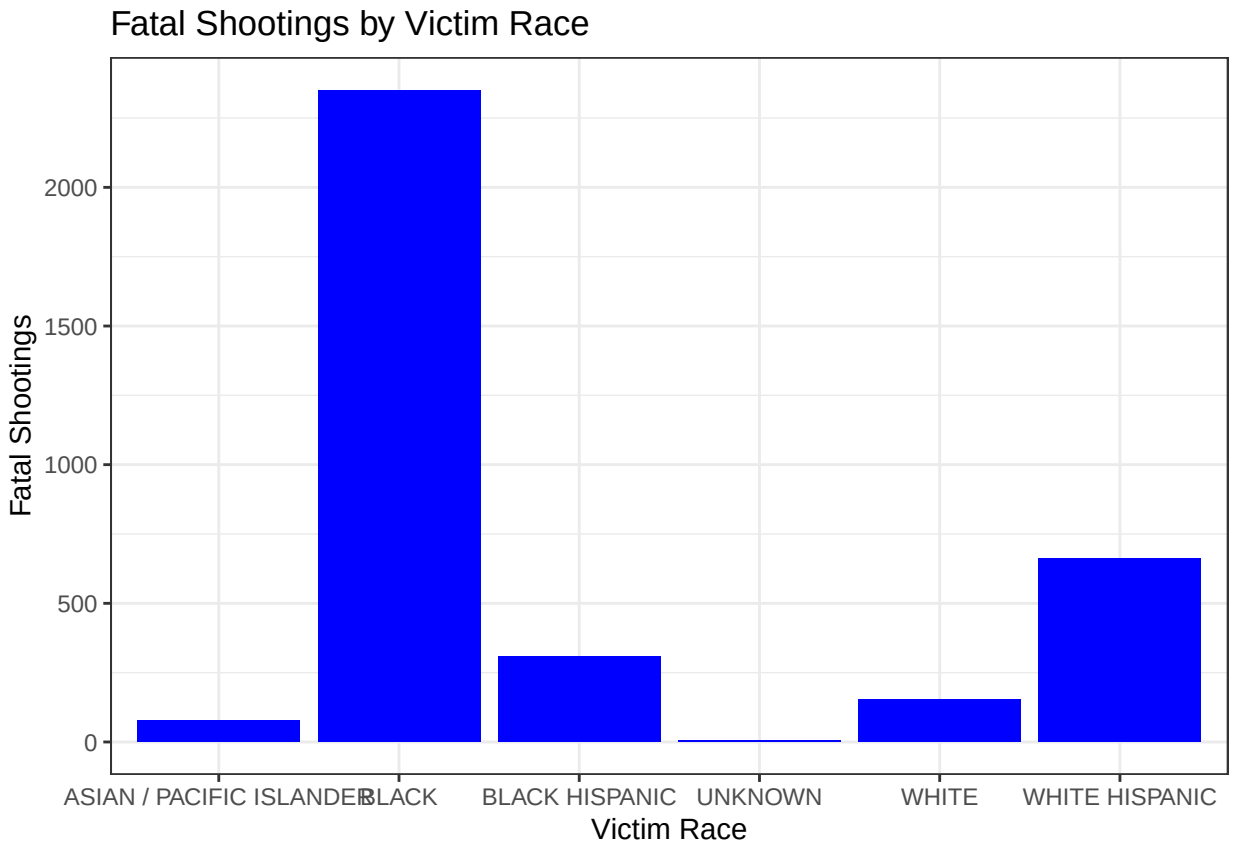

```
shootings_data %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = VIC_AGE_GROUP)) +
  geom_bar(fill = "blue")+
  theme_bw()+
  labs(x = "Victim Age Group",
       y = "Fatal Shootings",
       title = "Fatal Shootings by Victim Age Group")
```



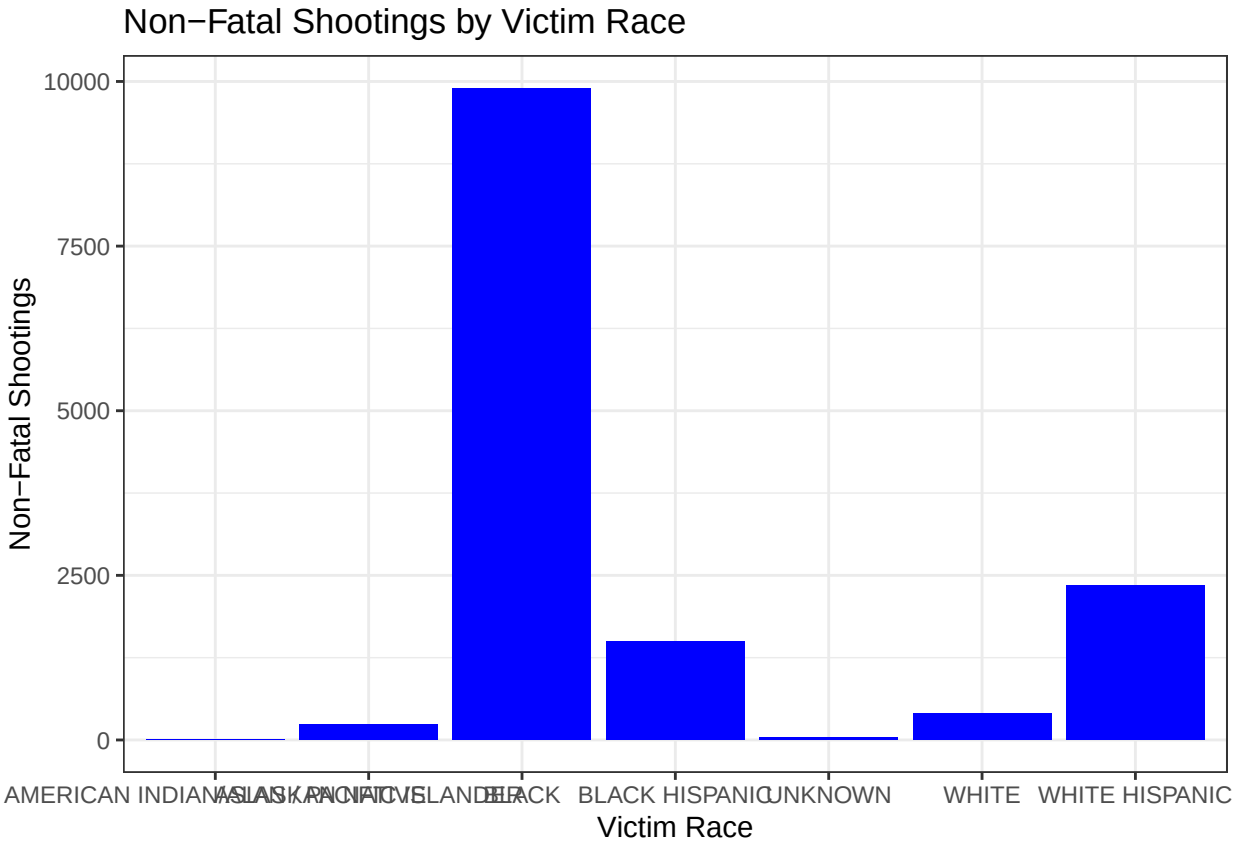

```
shootings_data %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = VIC_AGE_GROUP)) +
  geom_bar(fill = "blue")+
  theme_bw()+
  labs(x = "Victim Age Group",
       y = "Non-Fatal Shootings",
       title = "Non-Fatal Shootings by Victim Age Group")
```




```
shootings_data %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = VIC_RACE)) +
  geom_bar(fill = "blue")+
  theme_bw()+
  labs(x = "Victim Race",
       y = "Fatal Shootings",
       title = "Fatal Shootings by Victim Race")
```




```
shootings_data %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = VIC_RACE)) +
  geom_bar(fill = "blue")+
  theme_bw()+
  labs(x = "Victim Race",
       y = "Non-Fatal Shootings",
       title = "Non-Fatal Shootings by Victim Race")
```



The majority of victims as well as perpetrators of shootings both fatal and non-fatal are Black. The majority of victims as well as perpetrators of shootings both fatal and non-fatal are between the age groups 18-24 and 25-44 at the time of writing. Brooklyn and the Bronx experience the most shooting incidents.

Analysis

In order to analyze the data further, I am going to use regression analysis as this is the best use case because we have categorical data. I used this **Source**: https://www.geeksforgeeks.org/regression-with-categorical-variables-in-r-programming/ as a reference.

I want to determine what demographic / location data is a good predictor on the outcome of a shooting incident. To answer my question from above. I selected as my independent variable as STATISTICAL_MURDER_FLAG, and I chose my dependent variables as VIC_AGE_GROUP, VIC_RACE, BORO.

```
# Regression is a multi-step process for estimating the relationships between a dependent variable and

glm_model <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_RACE + BORO, data = shootings_data, famil

summary(glm_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_RACE +
##     BORO, family = "binomial", data = shootings_data)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -12.82092  114.57014  -0.112  0.91090
## VIC_AGE_GROUP1022             -10.67194  324.74371  -0.033  0.97378
```

```
## VIC_AGE_GROUP18-24                 0.29513    0.07209   4.094 4.24e-05 ***
## VIC_AGE_GROUP25-44                 0.55040    0.06996   7.867 3.62e-15 ***
## VIC_AGE_GROUP45-64                 0.67944    0.09185   7.397 1.39e-13 ***
## VIC_AGE_GROUP65+                   0.94459    0.19750   4.783 1.73e-06 ***
## VIC_AGE_GROUPUNKNOWN               0.57366    0.34746   1.651  0.09874 .
## VIC_RACEASIAN / PACIFIC ISLANDER  11.43038  114.57020   0.100  0.92053
## VIC_RACEBLACK                     11.10313  114.57013   0.097  0.92280
## VIC_RACEBLACK HISPANIC            10.91468  114.57014   0.095  0.92410
## VIC_RACEUNKNOWN                   10.48014  114.57101   0.091  0.92712
## VIC_RACEWHITE                     11.47184  114.57017   0.100  0.92024
## VIC_RACEWHITE HISPANIC            11.22623  114.57013   0.098  0.92194
## BOROBROOKLYN                      -0.19593    0.04727  -4.145 3.40e-05 ***
## BOROMANHATTAN                     -0.17633    0.06079  -2.901  0.00372 **
## BOROQUEENS                        -0.15469    0.05954  -2.598  0.00938 **
## BOROSTATEN ISLAND                 -0.06473    0.10411  -0.622  0.53411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17889  on 17967  degrees of freedom
## Residual deviance: 17712  on 17951  degrees of freedom
## AIC: 17746
##
## Number of Fisher Scoring iterations: 11
```

Given that the p-value for the predictor variable for VIC_AGE_GROUP18-24, VIC_AGE_GROUP25-44, VIC_AGE_GROUP45-64, VIC_AGE_GROUP65+, BOROBROOKLYN, BOROMANHATTAN, and BOROQUEENS is less than 0.05, means that they have a statistically significant relationship with the response variable in the model. **Source**: https://www.statology.org/interpret-prz-logistic-regression-output-r/

After reviewing the evidence in my regression analysis and table data. It appears that the age and location are determining factors in the outcome of a shooting incident. There is no statistical evidence that race has is a determining factor in the outcome of a shooting incident.

Bias

I tried to limit my own personal bias by staying true to the data and not jumping to conclusions. I let the data speak for itself when it comes to the statistical significance of a determining factor. I really had no dog in the fight it was fun to try and figure out which factors mattered the most.