NYPDShooting

TP

4/15/2022

In this data report, I will import the NYPD Shooting Incident data, visualize and analyze that data, build a model, and identify different biases.

Question of interest

For this data set, I'm especially curious about the relationship between murder status and victims. Who are the most vulnerable during shooting incidents in New York based on this data set based on victims' gender, age, and race?

Importing Data

The data is downloaded from NYPD OpenData. The data file is in csv format.

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

```
shooting_cases <- read_csv(url_in)
head(shooting_cases)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>       <dbl>             <dbl>
## 1     24050482 08/27/2006 05:35      BRONX          52                 0
## 2     77673979 03/11/2011 12:03      QUEENS        106                 0
## 3    203350417 10/06/2019 01:09      BROOKLYN       77                 0
## 4     80584527 09/04/2011 03:35      BRONX          40                 0
## 5     90843766 05/27/2013 21:16      QUEENS        100                 0
## 6     92393427 09/01/2013 04:17      BROOKLYN       67                 0
## # ... with 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

```
summary(shooting_cases)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME          BORO
##  Min.   :  9953245   Length:23585       Length:23585      Length:23585
##  1st Qu.: 55322804   Class :character   Class1:hms        Class :character
##  Median : 83435362   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102280741                      Mode  :numeric
##  3rd Qu.:150911774
##  Max.   :230611229
```

```
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   FALSE:19085
##  Median : 69.00   Median :0.000     Mode  :character   TRUE :4500
##  Mean   : 66.21   Mean   :0.333
##  3rd Qu.: 81.00   3rd Qu.:0.000
##  Max.   :123.00   Max.   :2.000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:23585       Length:23585       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999925   1st Qu.:182539
##  Mode  :character   Mode  :character   Median :1007654   Median :193470
##                                        Mean   :1009379   Mean   :207300
##                                        3rd Qu.:1016782   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23585
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

There are 23,585 incidents reported in the data set. Each incident is associated with a incident key. The date, time, location, shooters' information (age, race, gender), victims' information (age, race, gender), precinct, jurisdiction code, statistical murder flag, x coordination, y coordination, latitude, longitude, lon_lat of each incident were reported. In total, there are 19 data features in the data set.

There are missing data in LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE. The missing data in LOCATION_DESC may be due to the locations of the incidents are not classified in the system and the missing shooters' information may be due to the fact that the shooters have not caught or died during the incidents.

```r
shooting_cases <- shooting_cases %>%
  select(-c(PRECINCT, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, JURISDICTION_CODE,
              PERP_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
shooting_cases
```

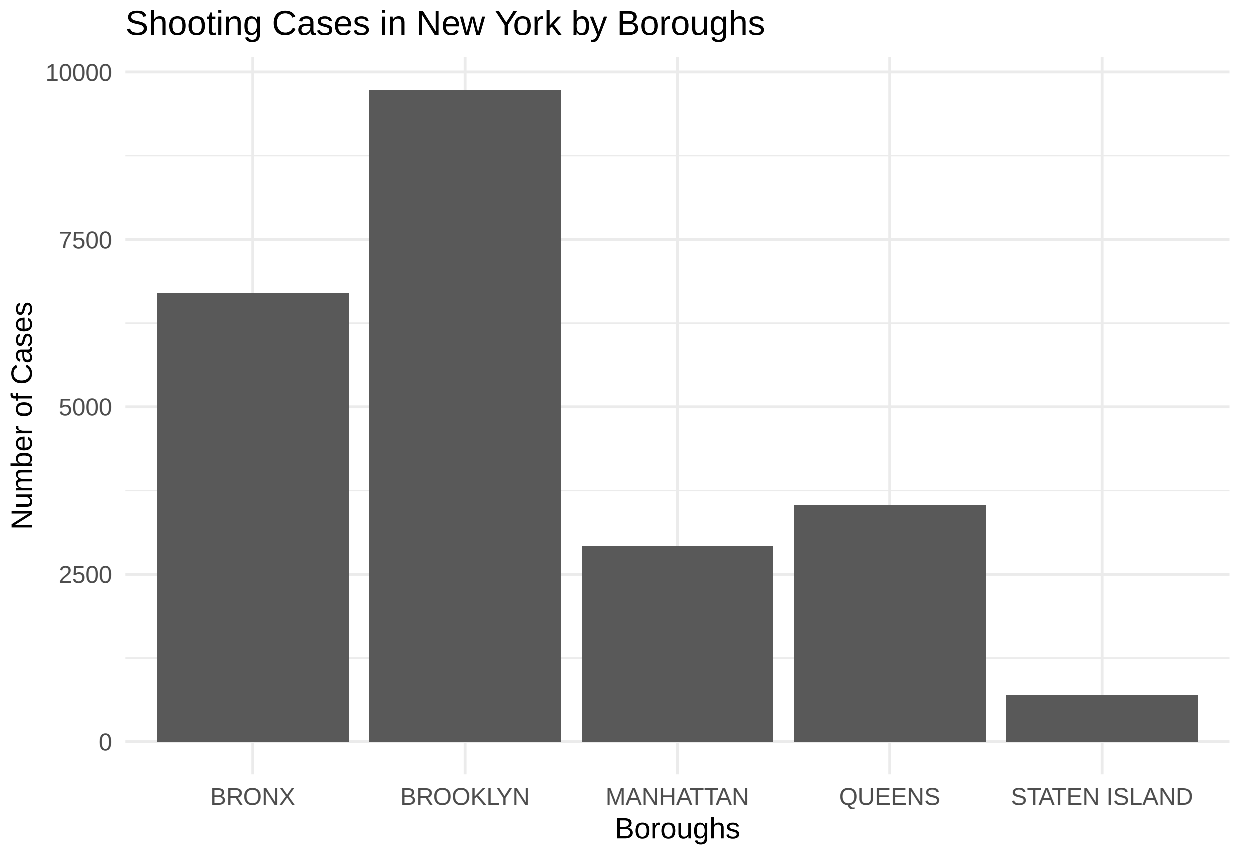
```
## # A tibble: 23,585 x 8
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     STATISTICAL_MURDER~ VIC_AGE_GROUP
##           <dbl> <date>     <time>     <chr>    <lgl>               <chr>
##  1     24050482 2006-08-27 05:35      BRONX    TRUE                25-44
##  2     77673979 2011-03-11 12:03      QUEENS   FALSE               65+
##  3    203350417 2019-10-06 01:09      BROOKLYN FALSE               18-24
##  4     80584527 2011-09-04 03:35      BRONX    FALSE               <18
##  5     90843766 2013-05-27 21:16      QUEENS   FALSE               18-24
##  6     92393427 2013-09-01 04:17      BROOKLYN FALSE               <18
##  7     73057167 2010-06-05 21:16      BROOKLYN FALSE               <18
##  8    211362213 2020-03-20 21:27      BROOKLYN FALSE               25-44
##  9    137564752 2014-07-04 00:25      QUEENS   FALSE               18-24
## 10    147024011 2015-10-18 01:33      QUEENS   FALSE               18-24
## # ... with 23,575 more rows, and 2 more variables: VIC_SEX <chr>,
## #   VIC_RACE <chr>
```

```
summary(shooting_cases)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:23585      Length:23585
##  1st Qu.: 55322804   1st Qu.:2008-12-31   Class1:hms        Class :character
##  Median : 83435362   Median :2012-02-27   Class2:difftime   Mode  :character
##  Mean   :102280741   Mean   :2012-10-05   Mode  :numeric
##  3rd Qu.:150911774   3rd Qu.:2016-03-02
##  Max.   :230611229   Max.   :2020-12-31
##  STATISTICAL_MURDER_FLAG VIC_AGE_GROUP        VIC_SEX
##  Mode :logical           Length:23585       Length:23585
##  FALSE:19085             Class :character   Class :character
##  TRUE :4500              Mode  :character   Mode  :character
##
##
##
##    VIC_RACE
##  Length:23585
##  Class :character
##  Mode  :character
##
##
##
```

Visualization and Analysis

```
# Visualization and analysis 1: Where has the most and the least number of shooting cases in New York
shooting_case_plot <- ggplot(shooting_cases, aes(x = BORO)) + geom_bar() +
  labs(title = "Shooting Cases in New York by Boroughs", x = "Boroughs", y = "Number of Cases") +
  theme_minimal()

shooting_case_plot
```

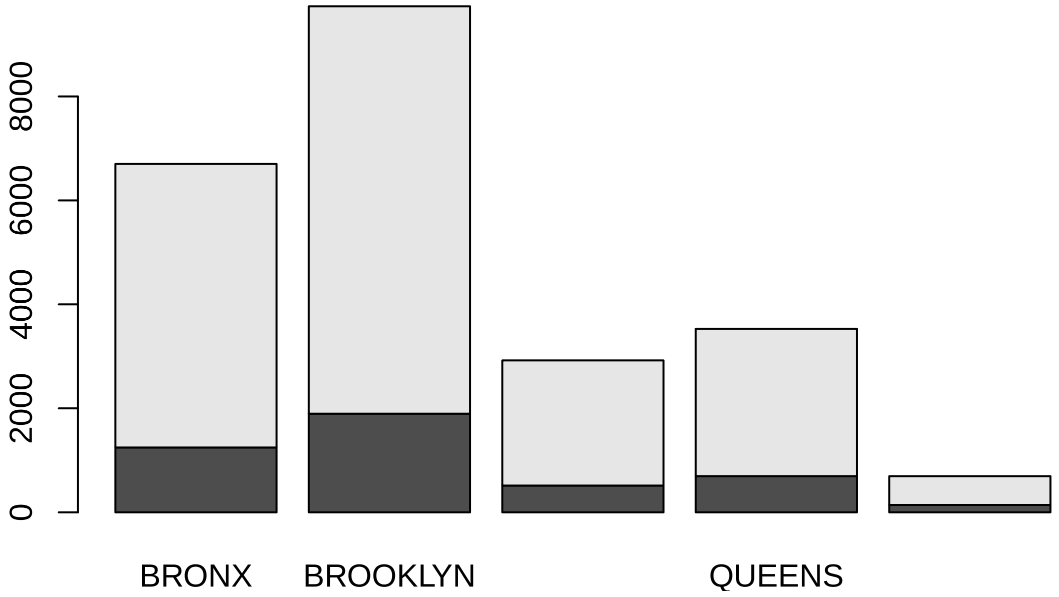


Analysis: From the plot above, we can see that Brooklyn has the highest number of shooting incidents (about 9000 cases) and Staten Island has the lowest number of incidents (about 600 cases).

```
murder_boolean <- shooting_cases$STATISTICAL_MURDER_FLAG
murders <- filter(shooting_cases, murder_boolean=='TRUE')
murder_by_boro <- table(t(murders$BORO))
not_murder<-filter(shooting_cases, murder_boolean=='FALSE')
not_murder_by_boro <- table(t(not_murder$BORO))
combined_murder_notMurder <- rbind(murder_by_boro, not_murder_by_boro)

table(shooting_cases$BORO, shooting_cases$STATISTICAL_MURDER_FLAG)
```

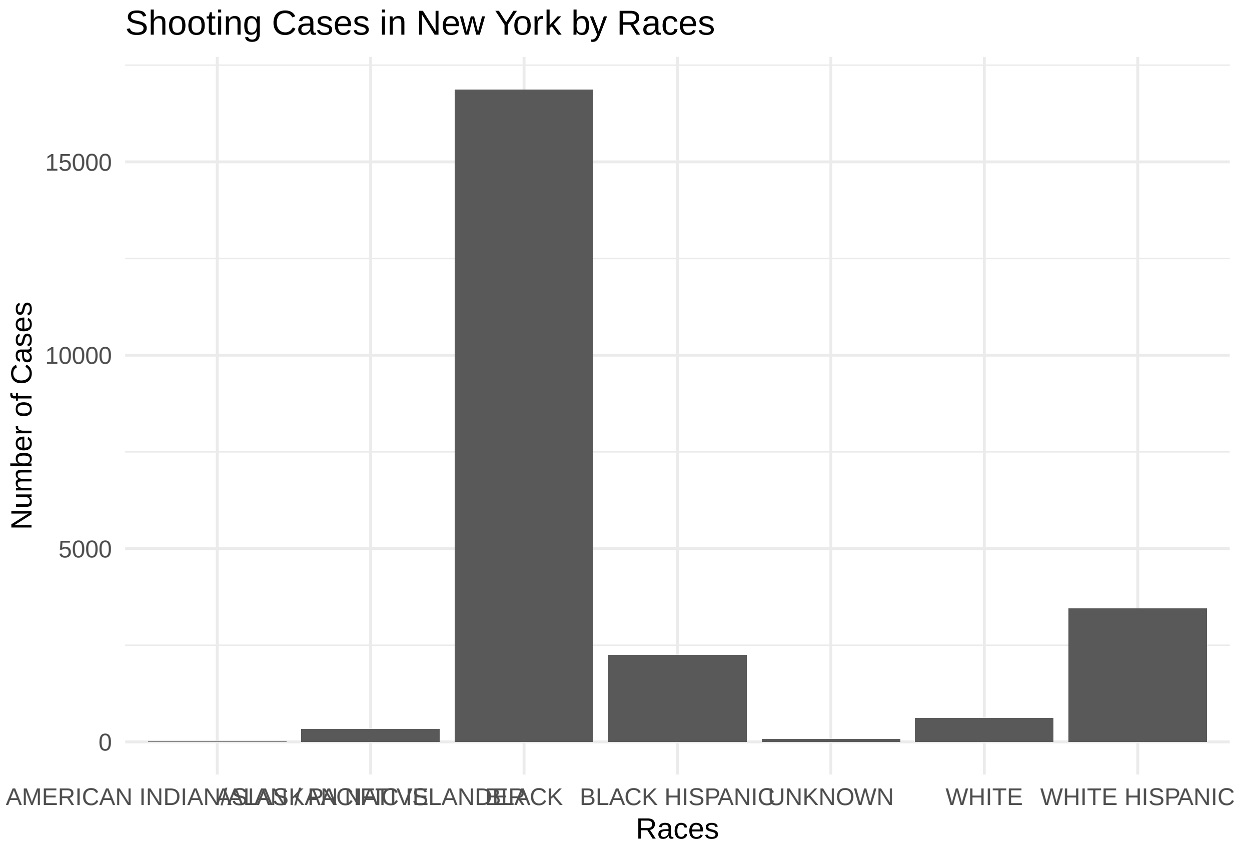
```
##
##                  FALSE TRUE
##   BRONX           5454 1247
##   BROOKLYN        7836 1898
##   MANHATTAN       2407  515
##   QUEENS          2835  697
##   STATEN ISLAND    553  143
```

```
barplot(combined_murder_notMurder)
```



Analysis: From the table and plot above, we can see that murder cases (darker part) take a small part of total shooting cases (lighter part).

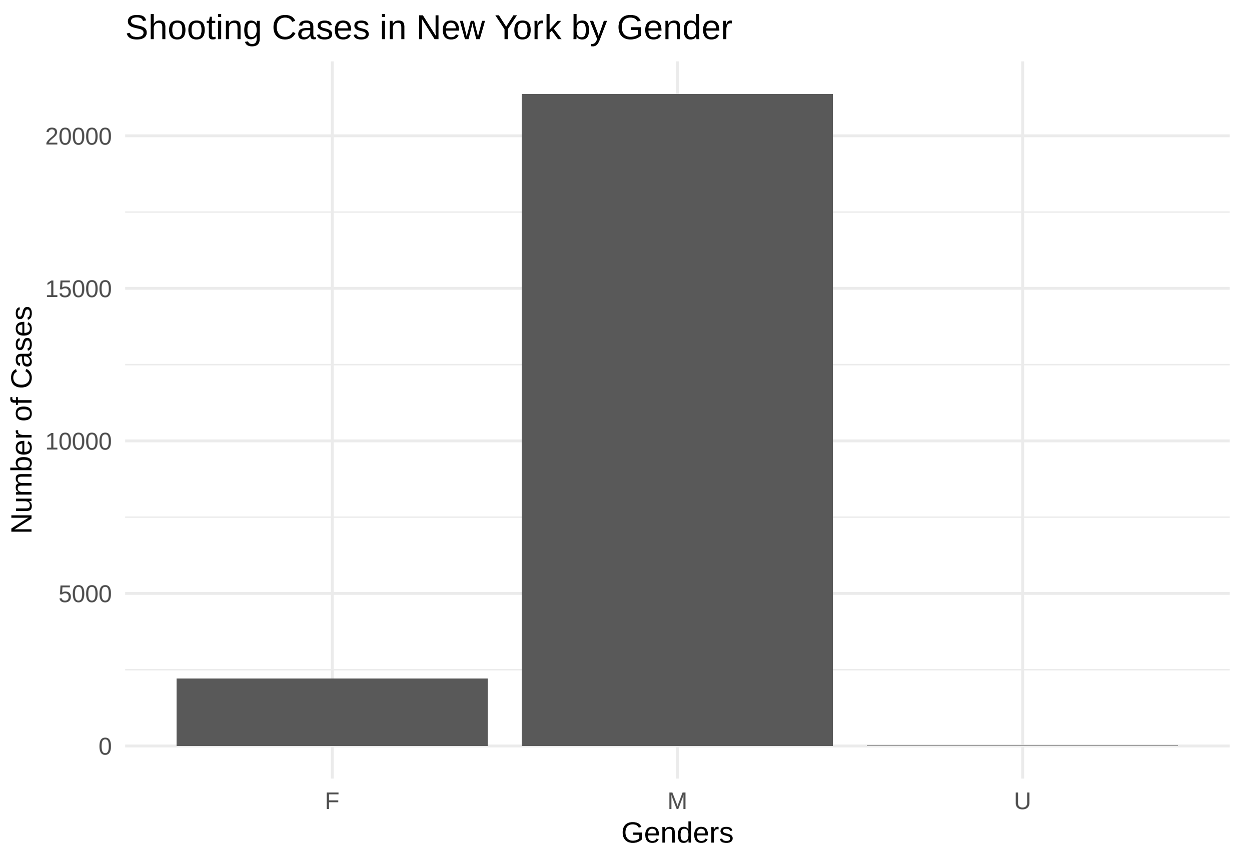
```
# Visualize race of victims
shooting_case_race_plot <- ggplot(shooting_cases, aes(x = VIC_RACE)) + geom_bar() +
  labs(title = "Shooting Cases in New York by Races", x = "Races", y = "Number of Cases") +
  theme_minimal()

shooting_case_race_plot
```



Analysis: From the plot above, we can see that the victims are mainly Black with more than half of the total cases. The second highest is White Hispanic. Pacific Islanders and American Indian/Alaskans make up a small number of cases.

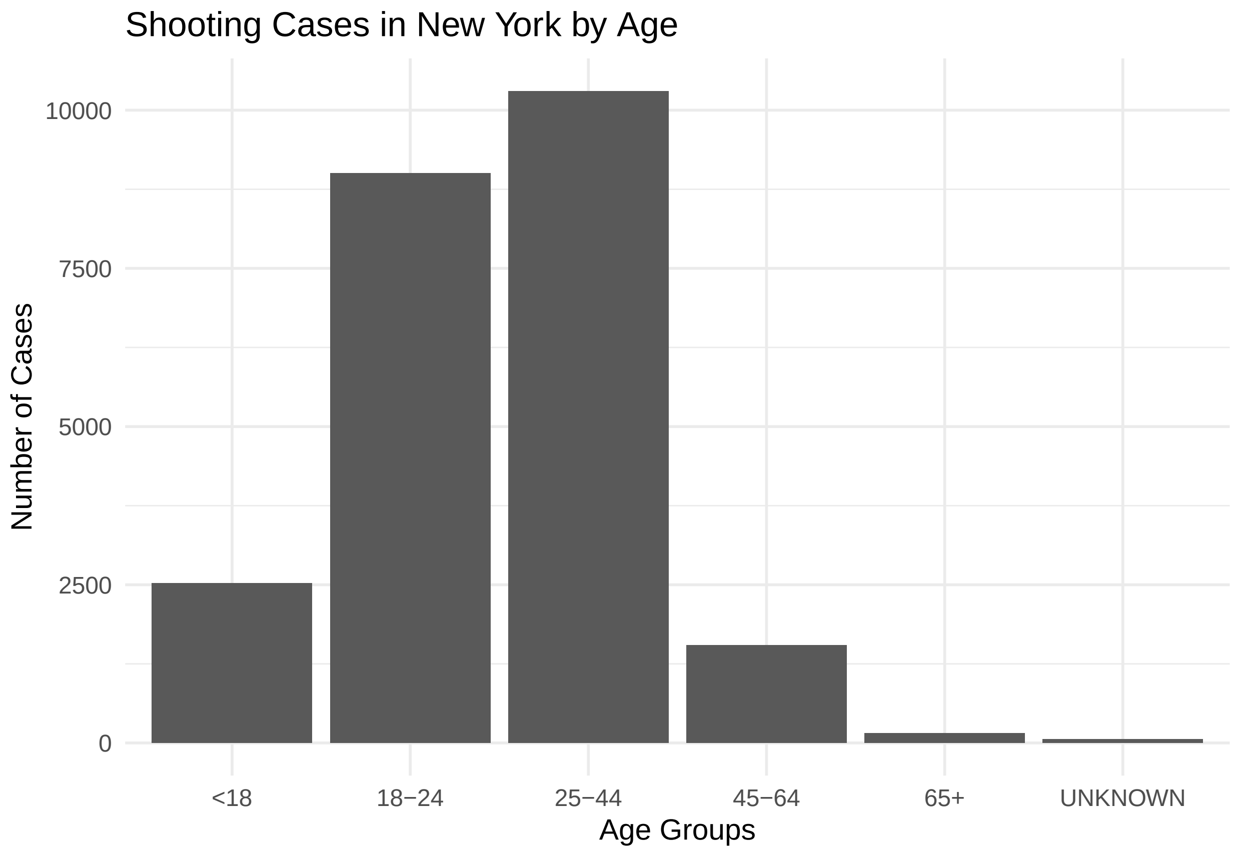
This raises the question of why Black people make up so many cases of shooting. Does Black have the highest number of population in New York? Do the shootings usually happen where Black people live?

```r
# Visualize genders of victims
shooting_case_gender_plot <- ggplot(shooting_cases, aes(x = VIC_SEX)) + geom_bar() +
  labs(title = "Shooting Cases in New York by Gender", x = "Genders", y = "Number of Cases") +
  theme_minimal()

shooting_case_gender_plot
```



Analysis: From the gender plot above, we can see that about 90% of victims are male and about 10% of victims are female. Unknown gender makes up a very small number.

```
# Visualize age groups of victims
shooting_case_age_plot <- ggplot(shooting_cases, aes(x = VIC_AGE_GROUP)) + geom_bar() +
  labs(title = "Shooting Cases in New York by Age", x = "Age Groups", y = "Number of Cases") +
  theme_minimal()

shooting_case_age_plot
```



Analysis: Top two victim age groups are 18-24 and 25-44. A very small number of victims are older than 65.

Model

In this logistic regression model, the independent variables will be VIC_AGE_GROUP, VIC_SEX, VIC_RACE, and dependent variable will be STATISTICAL_MURDER_FLAG. I will to see if the age, gender, and race of the victims affect the murder status.

```
mod <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX + VIC_RACE, data = shooting_cases, family=
summary(mod)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_SEX +
##     VIC_RACE, family = "binomial", data = shooting_cases)
##
## 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  107.58066  -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  107.58071   0.105  0.91647
## VIC_RACEBLACK                   10.99264  107.58064   0.102  0.91861
## VIC_RACEBLACK HISPANIC          10.78012  107.58065   0.100  0.92018
## VIC_RACEUNKNOWN                 10.27115  107.58146   0.095  0.92394
## VIC_RACEWHITE                   11.39679  107.58068   0.106  0.91563
## VIC_RACEWHITE HISPANIC          11.12689  107.58065   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
```

The model summary shows that victims in younger ages (< 25 years old) are more likely to survive after the shooting. The probability of surviving is decreasing as the ages get increased. And victims in older ages (65+) are less likely to survive.

Conclusion and Bias

The data set shows us that Brooklyn has the highest number of shooting incidents and Staten Island has the lowest number of incidents in New York. Victims are mainly Black and male, between the age of 18-44. The model shows that victims' ages affect the murder status or surviving rate. There is potential for biases occurring here in the data set and model. The given data set are highly specific in one area of New York which is Brooklyn and the victim are mainly Black. The data set is highly imbalanced.

sessionInfo()