NYPD Shooting Incident Data Report

Caden Zonnefeld

10/26/2022

Data Collection

```
library(tidyverse)
library(lubridate)

url <- 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'
nypd <- read_csv(url, show_col_types = FALSE)
```

Collects data pertaining to NYPD shooting incidents from data.gov. The data is accessed by utilizing a URL and reading the data in via a .csv file. The raw data consists of nearly 26k observations and 19 features. The data acquisition is reproducible since it is collected from a pubplic URL.

Tidying the Data

```
nypd$OCCUR_DATE <- mdy(nypd$OCCUR_DATE)

analysis_vars <- c('OCCUR_DATE', 'OCCUR_TIME', 'BORO', 'STATISTICAL_MURDER_FLAG',
                   'VIC_AGE_GROUP', 'VIC_SEX', 'VIC_RACE')
nypd <- nypd[analysis_vars]

nypd
```

```
## # A tibble: 25,596 x 7
##    OCCUR_DATE OCCUR_TIME BORO    STATISTICAL_MUR~ VIC_AGE_GROUP VIC_SEX VIC_RACE
##    <date>     <time>     <chr>   <lgl>            <chr>         <chr>   <chr>
##  1 2021-11-11 15:04      BROOKL~ FALSE            18-24         M       BLACK
##  2 2021-07-16 22:05      BROOKL~ FALSE            25-44         M       ASIAN /~
##  3 2021-07-11 01:09      BROOKL~ FALSE            25-44         M       BLACK
##  4 2021-12-11 13:42      BROOKL~ FALSE            25-44         M       BLACK
##  5 2021-02-16 20:00      QUEENS  FALSE            25-44         M       BLACK
##  6 2021-05-15 04:13      QUEENS  TRUE             25-44         M       BLACK
##  7 2021-04-14 21:08      BRONX   TRUE             18-24         M       BLACK
##  8 2021-12-10 19:30      BRONX   FALSE            25-44         M       BLACK
##  9 2021-02-22 00:18      MANHAT~ FALSE            25-44         M       BLACK H~
## 10 2021-03-07 06:15      BROOKL~ TRUE             25-44         M       WHITE H~
## # ... with 25,586 more rows
```

Inspecting the data and casting the OCCUR_DATE field to a date object to be properly treated as a date. Furthermore, reducing the dataset to the features I will use in an analysis that will focus on the victims of such incidents. The data is now tidy and ready for further analysis.

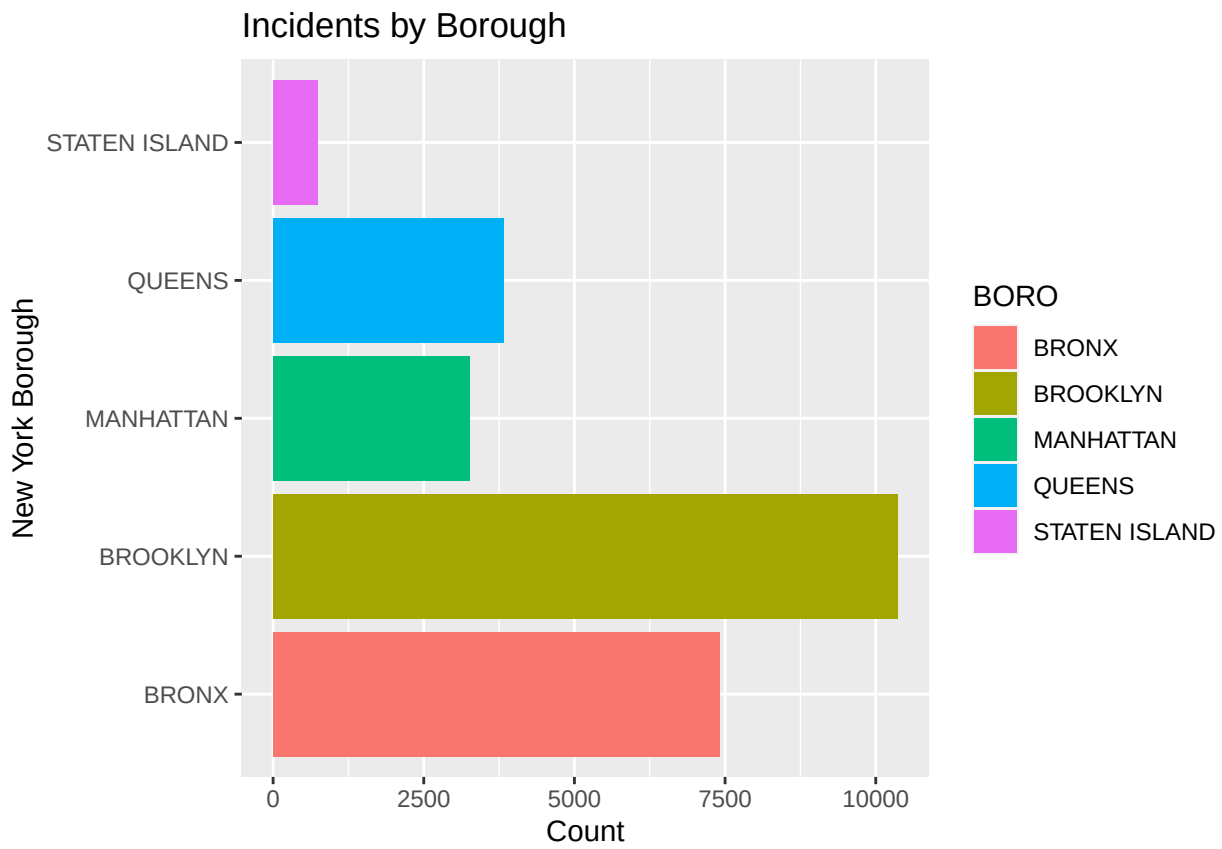
Data Visualization and Analysis

```
boro_numbers <- count(nypd %>% group_by(BORO))

boro_numbers$male <-count(nypd %>% group_by(BORO) %>% filter(VIC_SEX == 'M'))$n
boro_numbers$female <-count(nypd %>% group_by(BORO) %>% filter(VIC_SEX == 'F'))$n

boro_numbers %>% mutate(male_prop = male/n, female_prop = female/n)
```

```
## # A tibble: 5 x 6
## # Groups:   BORO [5]
##   BORO              n  male female male_prop female_prop
##   <chr>         <int> <int>  <int>     <dbl>       <dbl>
## 1 BRONX          7402  6753    646     0.912      0.0873
## 2 BROOKLYN      10365  9377    982     0.905      0.0947
## 3 MANHATTAN      3265  2952    311     0.904      0.0953
## 4 QUEENS         3828  3447    381     0.900      0.0995
## 5 STATEN ISLAND   736   653     83     0.887      0.113
```

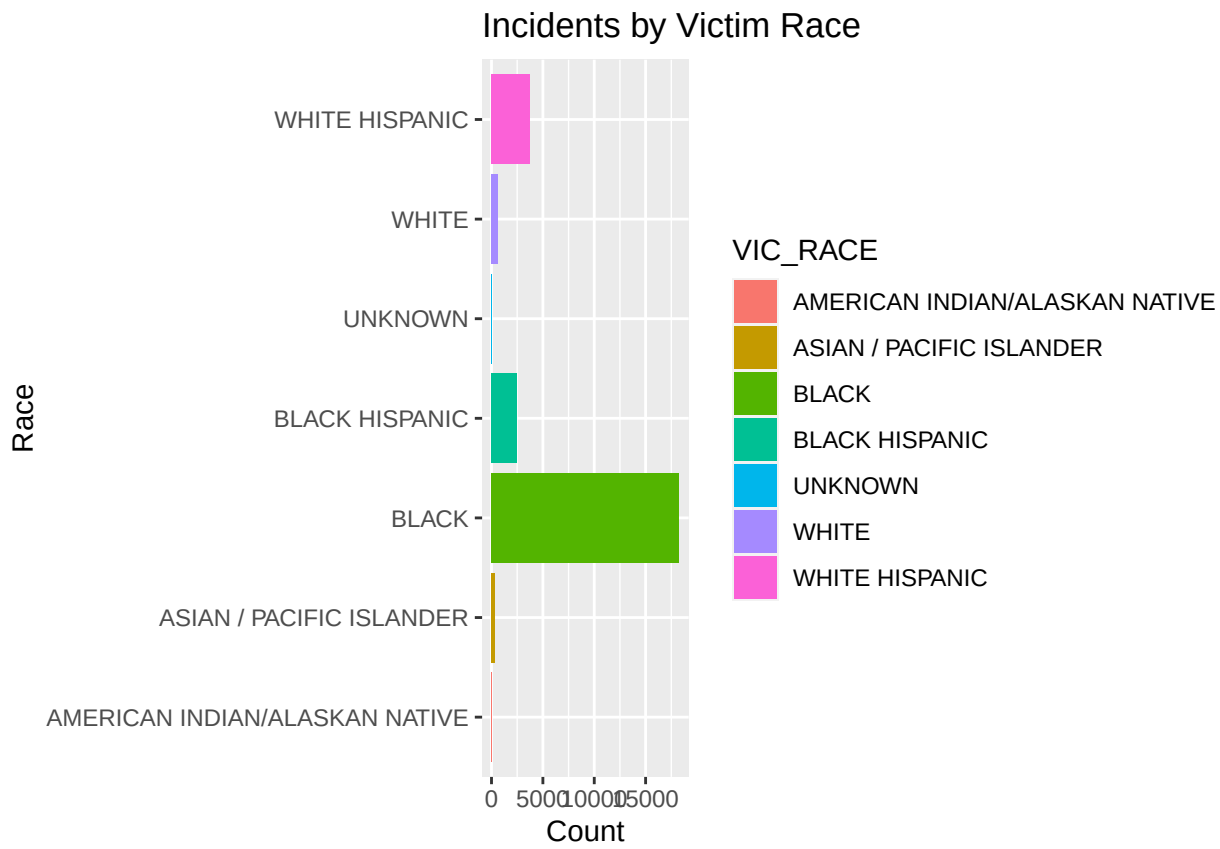
The table indicates the the victim count and proportion of victims by gender for each of New York's boroughs. The data reveals that there are far more males that have been involved in incidents than females.

```
ggplot(nypd) + geom_bar(aes(y=BORO, fill = BORO)) +
  ggtitle('Incidents by Borough') + xlab('Count') + ylab('New York Borough')
```



The first bar graph illustrates the number of incidents that occurred in each borough. We can gather that Staten Island has far fewer incidents while Brooklyn has the most recorded incidents. In fact, Brooklyn appears to have about 10x as many incidents as Staten Island.

```
ggplot(nypd) + geom_bar(aes(y=VIC_RACE, fill = VIC_RACE)) +
  ggtitle('Incidents by Victim Race') + xlab('Count') + ylab('Race')
```



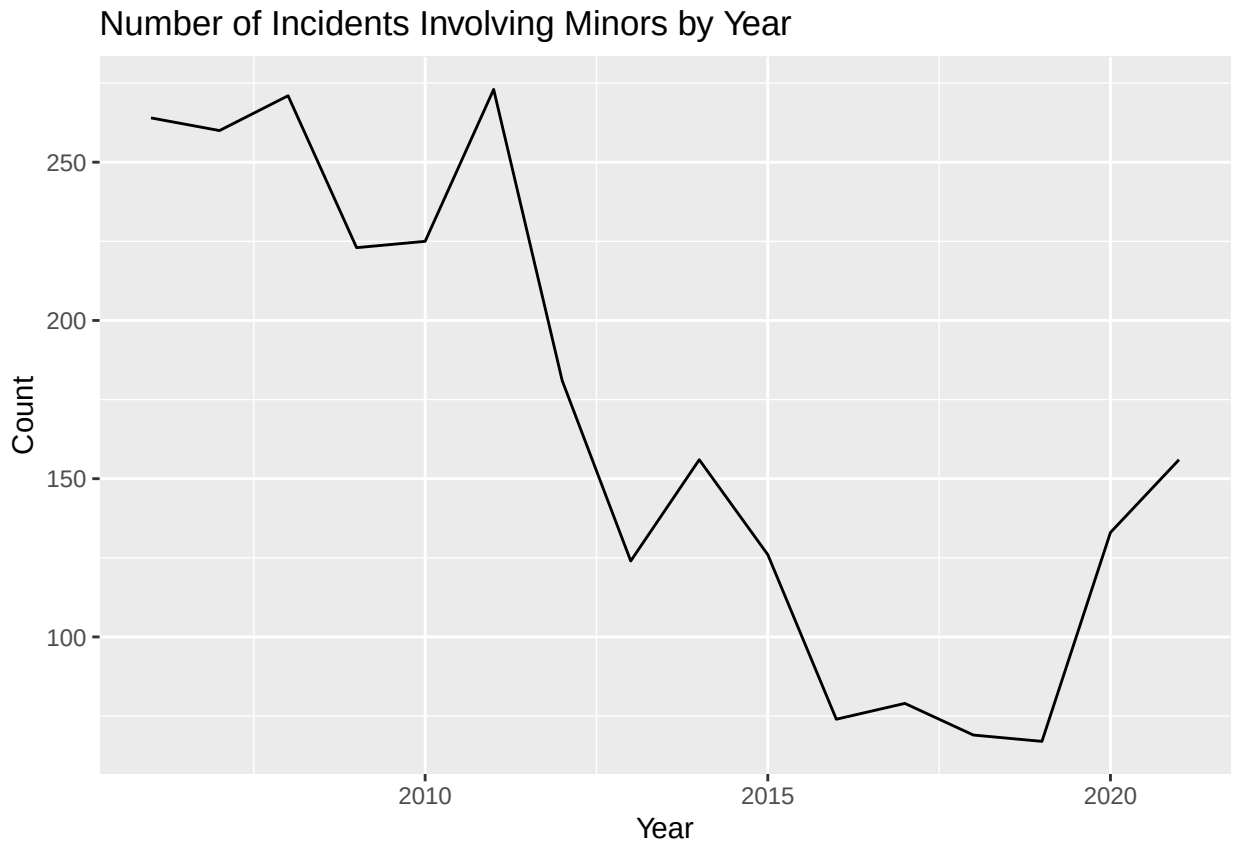
The next bar graph breaks down the race of the victim in each incident. Note that black and to a lesser extent Hispanic individuals account for far more incidents than other races. It appears that black individuals are cited in incidents at disproportionate rates to their portion of the population.

```
nypd_date <- nypd %>% filter(VIC_AGE_GROUP != 'UNKNOWN') %>%
  group_by(YEAR = year(OCCUR_DATE)) %>% count(VIC_AGE_GROUP)
nypd_date
```

```
## # A tibble: 80 x 3
## # Groups:   YEAR [16]
##     YEAR VIC_AGE_GROUP     n
##    <dbl> <chr>         <int>
##  1  2006 <18             264
##  2  2006 18-24           849
##  3  2006 25-44           813
##  4  2006 45-64           111
##  5  2006 65+              13
##  6  2007 <18             260
##  7  2007 18-24           749
##  8  2007 25-44           751
##  9  2007 45-64           107
## 10  2007 65+              12
## # ... with 70 more rows
```

```
minors <- nypd_date %>% filter(VIC_AGE_GROUP == '<18')
```

```
minors %>% ggplot(aes(x=YEAR, y=n)) + geom_line() + ylab('Count') + xlab('Year') +
  ggtitle('Number of Incidents Involving Minors by Year')
```



The grouped table shows the number of incidents that involved people from each age group band in the specified year. Lastly, the line graph shows the number of incidents involving minors on a yearly basis. Incidents involving minors have been greatly reduced since the beginning of data collection; however, there have been an uptick in such incidents over the last few years.

```
age_model <- lm(n ~ VIC_AGE_GROUP, data = nypd_date)
summary(age_model)
```

```
##
## Call:
## lm(formula = n ~ VIC_AGE_GROUP, data = nypd_date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -328.25  -28.98   -0.44   42.25  371.38
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           167.56      30.35   5.520 4.64e-07 ***
## VIC_AGE_GROUP18-24    432.69      42.93  10.080 1.35e-15 ***
## VIC_AGE_GROUP25-44    544.06      42.93  12.674  < 2e-16 ***
## VIC_AGE_GROUP45-64    -61.44      42.93  -1.431 0.156515
```

```
## VIC_AGE_GROUP65+    -157.12      42.93  -3.660 0.000466 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.4 on 75 degrees of freedom
## Multiple R-squared:  0.8517, Adjusted R-squared:  0.8437
## F-statistic: 107.6 on 4 and 75 DF,  p-value: < 2.2e-16
```

A simple linear model used to predict the number of shooting incidents in a given year by using only victim age group band indicates that age group 65+ has a significant effect that lowers the expected number of incidents. Conversely, age groups 18-24 and 25-44 have significant effects that increase the expected number of shooting incidents. Interestingly, victim age group alone accounts for 84% of the variation in the data.

```
nypd_race <- nypd %>% filter(VIC_AGE_GROUP != 'UNKNOWN') %>%
  group_by(YEAR = year(OCCUR_DATE)) %>% count(VIC_RACE)

race_model <- lm(n ~ VIC_RACE, data = nypd_race)
summary(race_model)
```

```
##
## Call:
## lm(formula = n ~ VIC_RACE, data = nypd_race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -469.44  -15.38   -0.29   14.95  306.56
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.286     46.486   0.028  0.97799
## VIC_RACEASIAN / PACIFIC ISLANDER  20.777     55.735   0.373  0.71016
## VIC_RACEBLACK                   1140.152     55.735  20.457  < 2e-16 ***
## VIC_RACEBLACK HISPANIC           153.777     55.735   2.759  0.00698 **
## VIC_RACEUNKNOWN                    2.560     57.659   0.044  0.96468
## VIC_RACEWHITE                     39.089     55.735   0.701  0.48484
## VIC_RACEWHITE HISPANIC           232.089     55.735   4.164 6.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 123 on 93 degrees of freedom
## Multiple R-squared:  0.9174, Adjusted R-squared:  0.912
## F-statistic: 172.1 on 6 and 93 DF,  p-value: < 2.2e-16
```

Once again I used a simple linear regression to predict the number of shooting incidents in a given year using victim race as a predictor. We can see that a race of Hispanic or black have a significant positive effect on the predicted number of incidents involved in for a given year. This model has good explanatory power since it accounts for 91% of the variation in the data.

Bias Recognition and Conclusion

For full disclosure, I am a white male which may have implicitly biased my approach to considering the data, particularly the race and gender features.

The data indicates that non-white individuals, particularly those who are black, are at a far greater risk of being victimized by a shooting incident. The stark contrast makes me wonder what biases against black individuals put them at such an elevated risk to be involved in a shooting incident. Furthermore, we can see that men are more prone to being victimized as there is about an 8:1 ratio for men to women victimized buy shootings. One encouraging trend from the data is the decrease of minors being involved in shootings; however, there is some cause for concern as the number of victimized minors has increased each year since 2019. Though we do not have the population of each borough, we can see that by number of incidents alone Staten Island appears to be the safest as it has less than a third of the incidents of the next lowest borough.