NYPD Data Science Project

Student

2025-09-07

Introduction

This analysis is to answer the question on how NYC shootings have been trending, and whether race is a factor for victims.

The source of data is historic NYPD shooting incident data from Data.gov (link: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic). This data contains a list of shooting incidents in NYC since 2006, including information about the incident time, location, as well as the perpetrator's and victim's race, gender, and age range.

Importing Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(dplyr)
library(ggplot2)
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read_csv(url_in)
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): X_COORD_CD, Y_COORD_CD
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

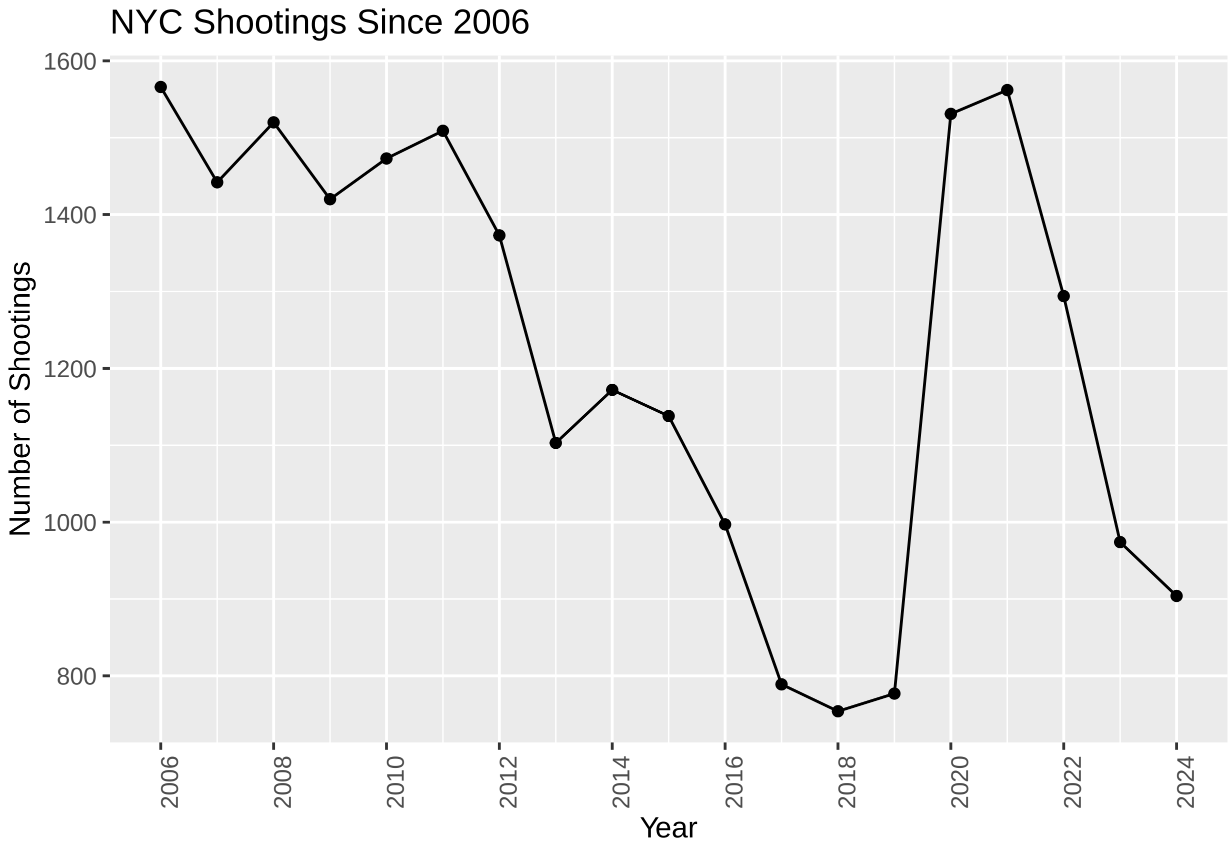
Transforming Data

```
# change factor columns to factor type
nypd_data <- nypd_data %>%
  mutate(BORO = factor(BORO),
         PERP_AGE_GROUP = factor(PERP_AGE_GROUP),
         PERP_SEX = factor(PERP_SEX),
         PERP_RACE = factor(PERP_RACE),
         VIC_AGE_GROUP = factor(VIC_AGE_GROUP),
         VIC_SEX = factor(VIC_SEX),
         VIC_RACE = factor(VIC_RACE)
         )

# change occur_date to date type
nypd_data <- nypd_data %>% mutate(OCCUR_DATE=mdy(OCCUR_DATE))

# remove coordinate and lat, long columns that are not needed
nypd_data <- nypd_data %>% select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

# show summary of transformed data
summary(nypd_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :00:00:00.000000
##  1st Qu.: 67321140   1st Qu.:2009-10-29   1st Qu.:03:30:45.000000
##  Median :109291972   Median :2014-03-25   Median :15:15:00.000000
##  Mean   :133850951   Mean   :2014-10-31   Mean   :12:46:10.874798
##  3rd Qu.:214741917   3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000
##  Max.   :299462478   Max.   :2024-12-31   Max.   :23:59:00.000000
##
##             BORO       LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  BRONX        : 8834   Length:29744       Min.   :  1.00   Min.   :0.0000
##  BROOKLYN     :11685   Class :character   1st Qu.: 44.00   1st Qu.:0.0000
##  MANHATTAN    : 3977   Mode  :character   Median : 67.00   Median :0.0000
##  QUEENS       : 4426                      Mean   : 65.23   Mean   :0.3181
##  STATEN ISLAND:  822                      3rd Qu.: 81.00   3rd Qu.:0.0000
##                                           Max.   :123.00   Max.   :2.0000
##                                                            NA’s   :2
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:29744       Length:29744       Mode :logical           18-24  :6630
##  Class :character   Class :character   FALSE:23979             25-44  :6342
##  Mode  :character   Mode  :character   TRUE :5765              UNKNOWN:3148
##                                                                <18    :1805
##                                                                (null) :1628
##                                                                (Other): 847
##                                                                NA’s   :9344
##    PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  (null): 1628   BLACK         :12323   <18    : 3081   F: 2891
##  F     :  461   WHITE HISPANIC: 2667   1022   :     1   M:26841
##  M     :16845   UNKNOWN       : 1838   18-24  :10677   U:   12
##  U     : 1500   (null)        : 1628   25-44  :13563
##  NA's  : 9310   BLACK HISPANIC: 1487   45-64  : 2118
##                 (Other)       :  491   65+    :  236
##                 NA's          : 9310   UNKNOWN:   68
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   13
##  ASIAN / PACIFIC ISLANDER      :  478
##  BLACK                         :20999
##  BLACK HISPANIC                : 2930
##  UNKNOWN                       :   72
##  WHITE                         :  741
##  WHITE HISPANIC                : 4511
```

Visualizing Data

```
incident_totals_by_year <- nypd_data %>% group_by(year(OCCUR_DATE)) %>% summarize(count_incidents=n_dis

incident_totals_by_year %>% ggplot(aes(x = `year(OCCUR_DATE)`, y = count_incidents)) +
  geom_line() +
  geom_point() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  scale_x_continuous( breaks = round(seq(min(incident_totals_by_year$`year(OCCUR_DATE)`),
    max(incident_totals_by_year$`year(OCCUR_DATE)`),
    by = 2),1)) +
  labs(title = "NYC Shootings Since 2006", y="Number of Shootings", x="Year")
```



```
victims_by_year_native <- nypd_data %>% filter(VIC_RACE == "AMERICAN INDIAN/ALASKAN NATIVE") %>% group_
victims_by_year_native$count_native[is.na(victims_by_year_native$count_native)] <- 0
victims_by_year_asian_pacific_islander <- nypd_data %>% filter(VIC_RACE == "ASIAN / PACIFIC ISLANDER") %
victims_by_year_black <- nypd_data %>% filter(VIC_RACE == "BLACK") %>% group_by(year(OCCUR_DATE)) %>% su
victims_by_year_black_hispanic <- nypd_data %>% filter(VIC_RACE == "BLACK HISPANIC") %>% group_by(year(
victims_by_year_white <- nypd_data %>% filter(VIC_RACE == "WHITE") %>% group_by(year(OCCUR_DATE)) %>% su
victims_by_year_white_hispanic<- nypd_data %>% filter(VIC_RACE == "WHITE HISPANIC") %>% group_by(year(O
victims_by_year_unknown <- nypd_data %>% filter(VIC_RACE == "UNKNOWN") %>% group_by(year(OCCUR_DATE)) %

victims_by_race <- victims_by_year_native %>%
  full_join(victims_by_year_asian_pacific_islander) %>%
  full_join(victims_by_year_black) %>%
  full_join(victims_by_year_black_hispanic) %>%
  full_join(victims_by_year_white) %>%
  full_join(victims_by_year_white_hispanic) %>%
  full_join(victims_by_year_unknown)
```

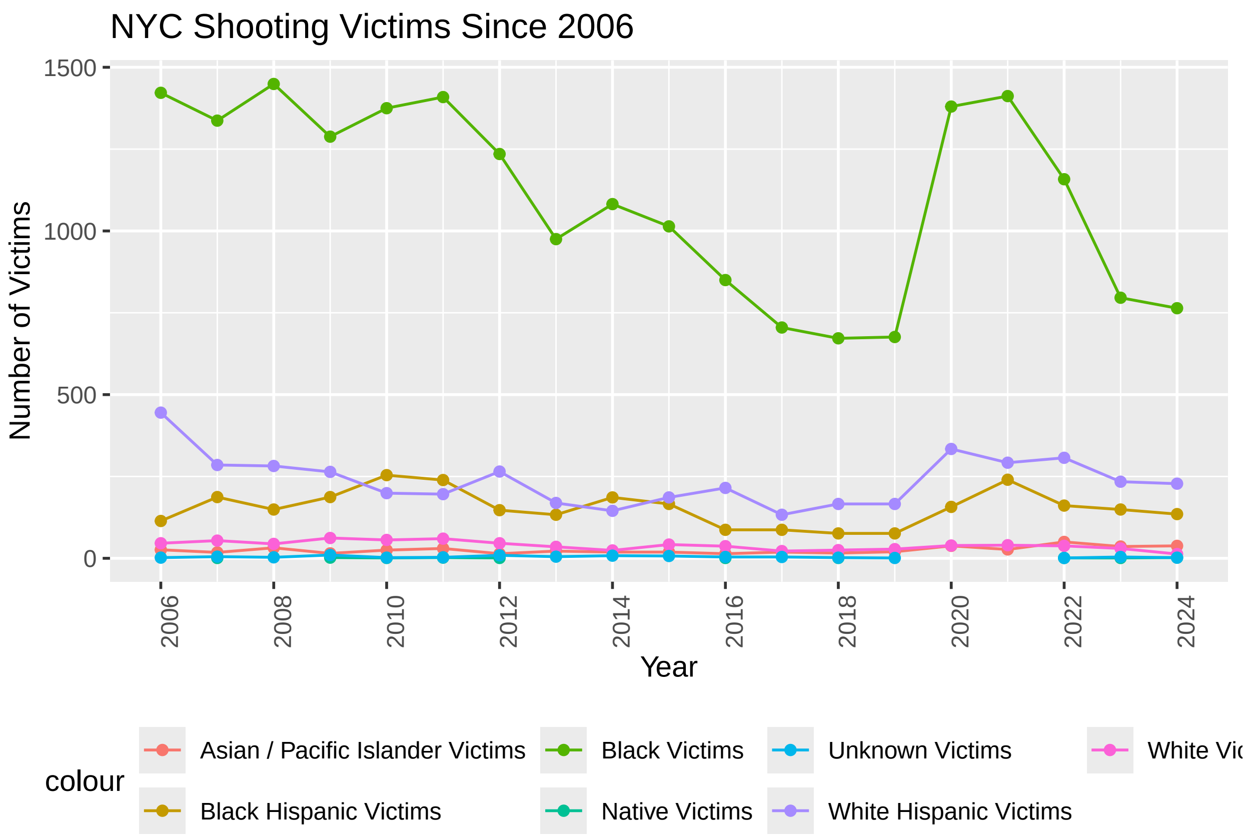
```
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
```

```r
victims_by_race %>% ggplot(aes(x = `year(OCCUR_DATE)`, y = count_native)) +
  geom_line(aes(color="Native Victims")) +
  geom_point(aes(color="Native Victims")) +
  geom_line(aes(y = count_asian, color="Asian / Pacific Islander Victims")) +
  geom_point(aes(y = count_asian, color="Asian / Pacific Islander Victims")) +
  geom_line(aes(y = count_black, color="Black Victims")) +
  geom_point(aes(y = count_black, color="Black Victims")) +
  geom_line(aes(y = count_black_hispanic, color="Black Hispanic Victims")) +
  geom_point(aes(y = count_black_hispanic, color="Black Hispanic Victims")) +
  geom_line(aes(y = count_white, color="White Victims")) +
  geom_point(aes(y = count_white, color="White Victims")) +
  geom_line(aes(y = count_white_hispanic, color="White Hispanic Victims")) +
  geom_point(aes(y = count_white_hispanic, color="White Hispanic Victims")) +
  geom_line(aes(y = count_unknown, color="Unknown Victims")) +
  geom_point(aes(y = count_unknown, color="Unknown Victims")) +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  scale_x_continuous( breaks = round(seq(min(incident_totals_by_year$`year(OCCUR_DATE)`),
    max(incident_totals_by_year$`year(OCCUR_DATE)`),
    by = 2),1)) +
  labs(title = "NYC Shooting Victims Since 2006", y="Number of Victims", x="Year")
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

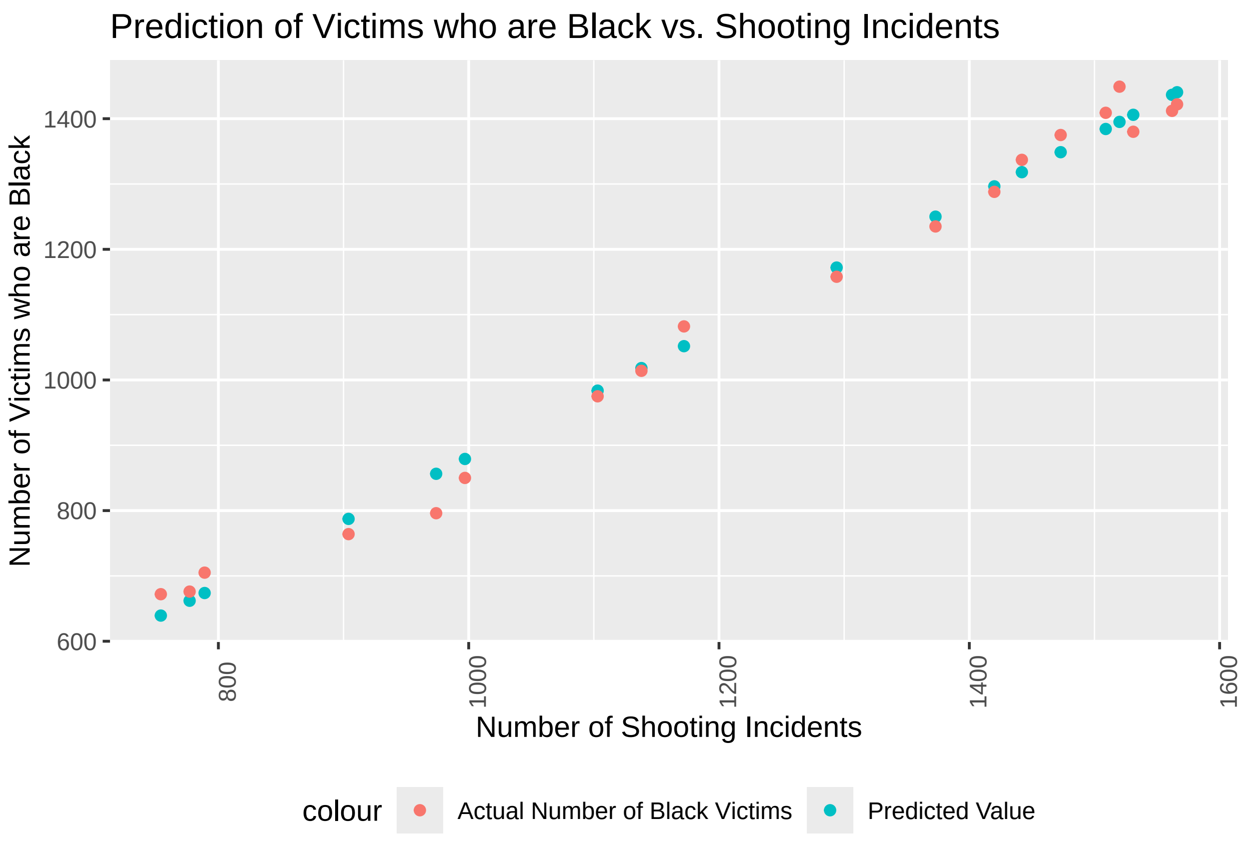


Modeling Prediction of Victims who are Black based on Number of Incidents

```
victims_total_by_race <- incident_totals_by_year %>% full_join(victims_by_race)
```

```
## Joining with ‘by = join_by(‘year(OCCUR_DATE)‘)‘
```

```
mod_black <- lm(count_black ~ count_incidents, data=victims_total_by_race)
victim_pred_b <- victims_total_by_race %>% mutate(pred = predict(mod_black))

victim_pred_b %>% ggplot(aes(x = count_incidents, y = pred)) +
  geom_point(aes(color="Predicted Value")) +
  geom_point(aes(y = count_black, color="Actual Number of Black Victims")) +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  labs(title = "Prediction of Victims who are Black vs. Shooting Incidents", y="Number of Victims who a
```



Conclusion

The analysis shows that the amount of shootings in NYC has been decreasing since 2006, with the exception of years 2020 and 2021. As 2020 and 2021 are outliers, further analysis is required to understand the factors related to the unusual spike in shootings these two years. Additionally, it shows that the ratio of victims who are Black is significantly higher than other victims of other races. The lowest number of victims have historically been people of White or Native race. Another source of bias could be missing information about the general population distribution among the different races. A low population of a particular race could be related to the low number among their race. Another possible sources of bias could be the way the original shooting data was collected. If there is higher police presence in predominantly non-Caucasian neighborhoods, it could make it more likely for shootings to be identified by police and recorded. To mitigate bias, I chose to include victims with unknown races in my analysis, to ensure the results were not skewed by missing race information for certain victims. I also used clearly labeled axes with consistent scales to avoid distorting the visualization of data.