NYPD Interratial vs Intraracial Shootings

Dillon Williams

8/17/2021

Bringing in the Data

First, the data will be loaded from https://catalog.data.gov. The specific URL to the NYPD Shooting Data file is https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD. The following code identifies the URL and reads the csv data, assigning the data to the variable NYPD_data.

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

```
NYPD_data <- read_csv(url_in, show_col_types = FALSE)
```

Tidying the Data

This project will be looking into race correlations between shooting perpetrators and victims. The following code will tidy the data to show perpetrator race verses the victim race in NYC shootings.

```
race_data <- NYPD_data[,c("PERP_RACE","VIC_RACE")]
head(race_data)
```

```
## # A tibble: 6 x 2
##   PERP_RACE      VIC_RACE
##   <chr>          <chr>
## 1 <NA>           BLACK
## 2 BLACK          BLACK
## 3 WHITE HISPANIC BLACK HISPANIC
## 4 BLACK          BLACK
## 5 BLACK HISPANIC BLACK
## 6 WHITE HISPANIC BLACK
```

From the head() command, an "NA" value can be seen in the PERP_RACE column. Upon further examination, there is a lot of "NA" values throughout the table. There are also "UNKNOWN" values in the data. The following code will remove them:

```
race_data <- na.omit(race_data)
race_data <- as.data.frame(race_data)
race_data <- race_data[!(race_data$PERP_RACE == "UNKNOWN" |
                           race_data$VIC_RACE == "UNKNOWN"), ]
head(race_data)
```

```
##        PERP_RACE       VIC_RACE
## 1          BLACK          BLACK
## 2 WHITE HISPANIC BLACK HISPANIC
## 3          BLACK          BLACK
## 4 BLACK HISPANIC          BLACK
## 5 WHITE HISPANIC          BLACK
## 6          BLACK          BLACK
```

Now that the "NA" and "UNKNOWN" values have been removed, the remaining subset of data must be compared to the original dataset.

```
perp_race_count = as.numeric(count(NYPD_data["PERP_RACE"]))
race_data_count = as.numeric(count(race_data["PERP_RACE"]))
race_sample_size = perp_race_count - race_data_count
race_sample_percentage = race_sample_size/perp_race_count*100
sprintf("Original sample size = %d", perp_race_count)
```

```
## [1] "Original sample size = 23568"
```

```
sprintf("race_sample_size = %d", race_sample_size)
```

```
## [1] "race_sample_size = 10350"
```

```
sprintf("The sample size is %f percent of the original dataset",
        race_sample_percentage)
```

```
## [1] "The sample size is 43.915479 percent of the original dataset"
```

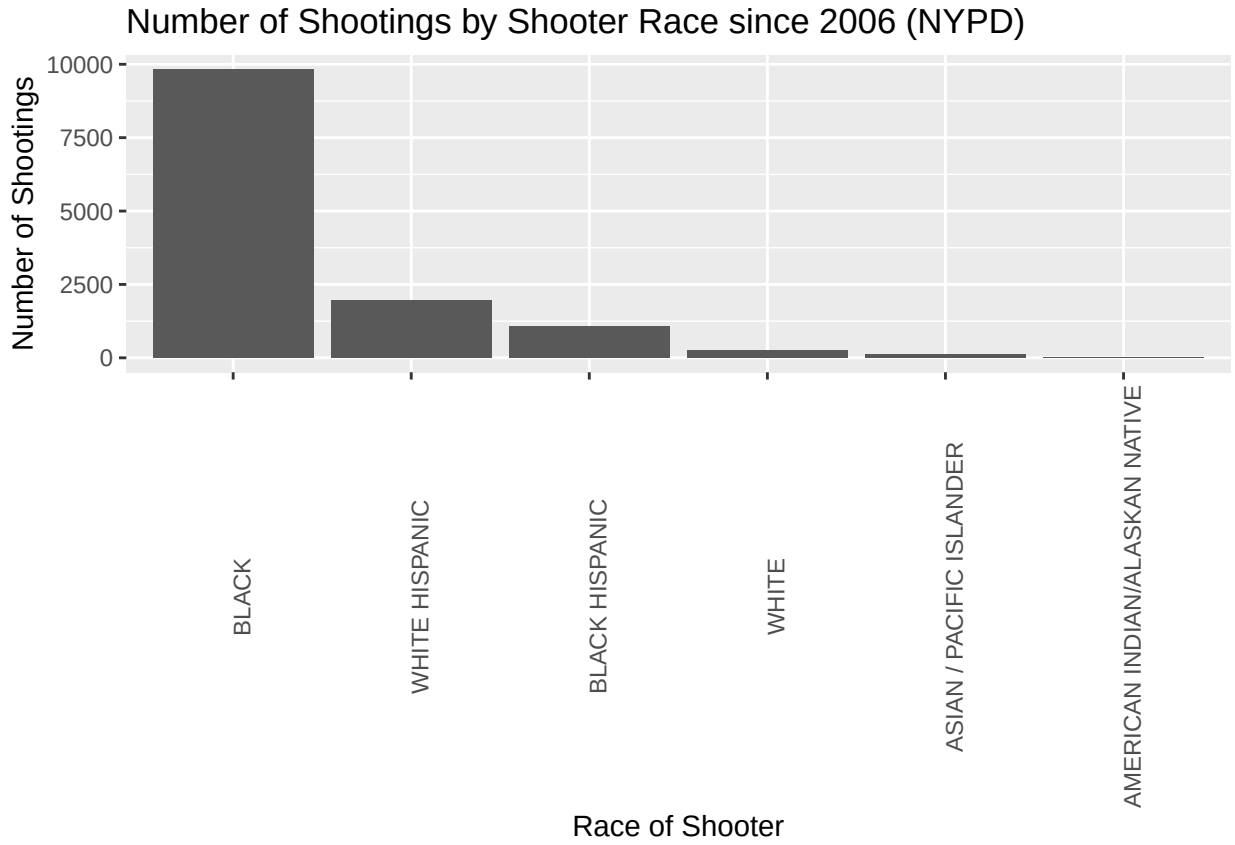
This is the first possible introduction of bias: race data is only present in ~44% of the data. The other 56% of data may tell a different story than what this subset shows. All conclusions must keep this in mind.

Visualize the Data

Now that the data is tidied, visualization begins. This visualization will begin by charting the amount of shootings per shooter's race.

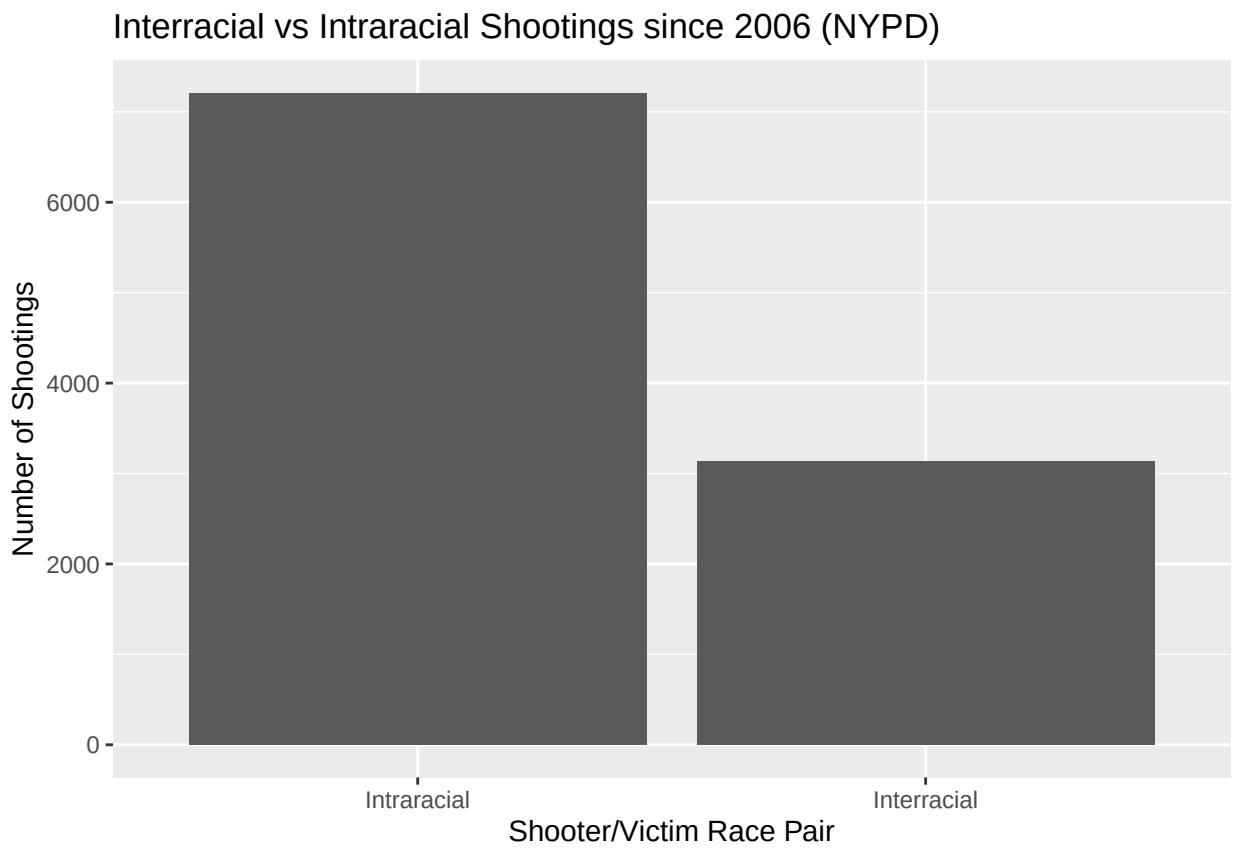
```
race_perp_table <- table(race_data["PERP_RACE"])
race_perp_table <- as.data.frame(race_perp_table)

native_perp_count = as.numeric(
  race_perp_table[race_perp_table$Var1 ==
                    "AMERICAN INDIAN/ALASKAN NATIVE", "Freq"])
asian_perp_count = as.numeric(
  race_perp_table[race_perp_table$Var1 == "ASIAN / PACIFIC ISLANDER",
                  "Freq"])
black_perp_count = as.numeric(race_perp_table[race_perp_table$Var1 ==
                                                "BLACK", "Freq"])
black_hispanic_perp_count = as.numeric(
  race_perp_table[race_perp_table$Var1 == "BLACK HISPANIC", "Freq"])
white_perp_count = as.numeric(
```

```
  race_perp_table[race_perp_table$Var1 == "WHITE", "Freq"])
white_hispanic_perp_count = as.numeric(
  race_perp_table[race_perp_table$Var1 == "WHITE HISPANIC", "Freq"])

race_perp_categories <- c("AMERICAN INDIAN/ALASKAN NATIVE",
                          "ASIAN / PACIFIC ISLANDER", "BLACK",
                          "BLACK HISPANIC", "WHITE", "WHITE HISPANIC")
race_perp_counts <- c(native_perp_count, asian_perp_count,
                      black_perp_count, black_hispanic_perp_count,
                      white_perp_count, white_hispanic_perp_count)

perp_race_table_new <- data.frame(race_perp_categories, race_perp_counts)

ggplot(perp_race_table_new, aes(x=reorder(race_perp_categories,
                                          -race_perp_counts),
                                y=race_perp_counts))+
  geom_bar(stat='identity')+
  theme(axis.text.x = element_text(angle = 90))+
  ggtitle("Number of Shootings by Shooter Race since 2006 (NYPD)")+
  xlab("Race of Shooter")+
  ylab("Number of Shootings")
```



Here lies the second possible bias: this chart paints black people in a negative light. Those with inherent racial bias against black people may use this chart to derive the wrong conclusion. To eleminate this bias, the data will now be sorted in a new way: interracial vs intraracial.

```
race_correlations <- data.frame("Racial_Combination" =
                                  c("Interracial", "Intraracial"),
                                "Incidents" = c(0,0))

for(i in 1:race_sample_size){
  perp_race = race_data[i, "PERP_RACE"]
  vic_race = race_data[i, "VIC_RACE"]
  if(perp_race != vic_race){
    race_correlations[1,2]=
      as.numeric(race_correlations[1,2]) + 1
  }
  else{
    race_correlations[2,2]=
      as.numeric(race_correlations[2,2]) + 1
  }
}

ggplot(race_correlations,
       aes(x=reorder(Racial_Combination, -as.numeric(Incidents)),
           y=as.numeric(Incidents)))+
  geom_bar(stat='identity')+
  ggtitle("Interracial vs Intraracial Shootings since 2006 (NYPD)")+
  xlab("Shooter/Victim Race Pair")+
  ylab("Number of Shootings")
```



This chart demonstrates how many more intraracial shootings there were than interracial. This could be

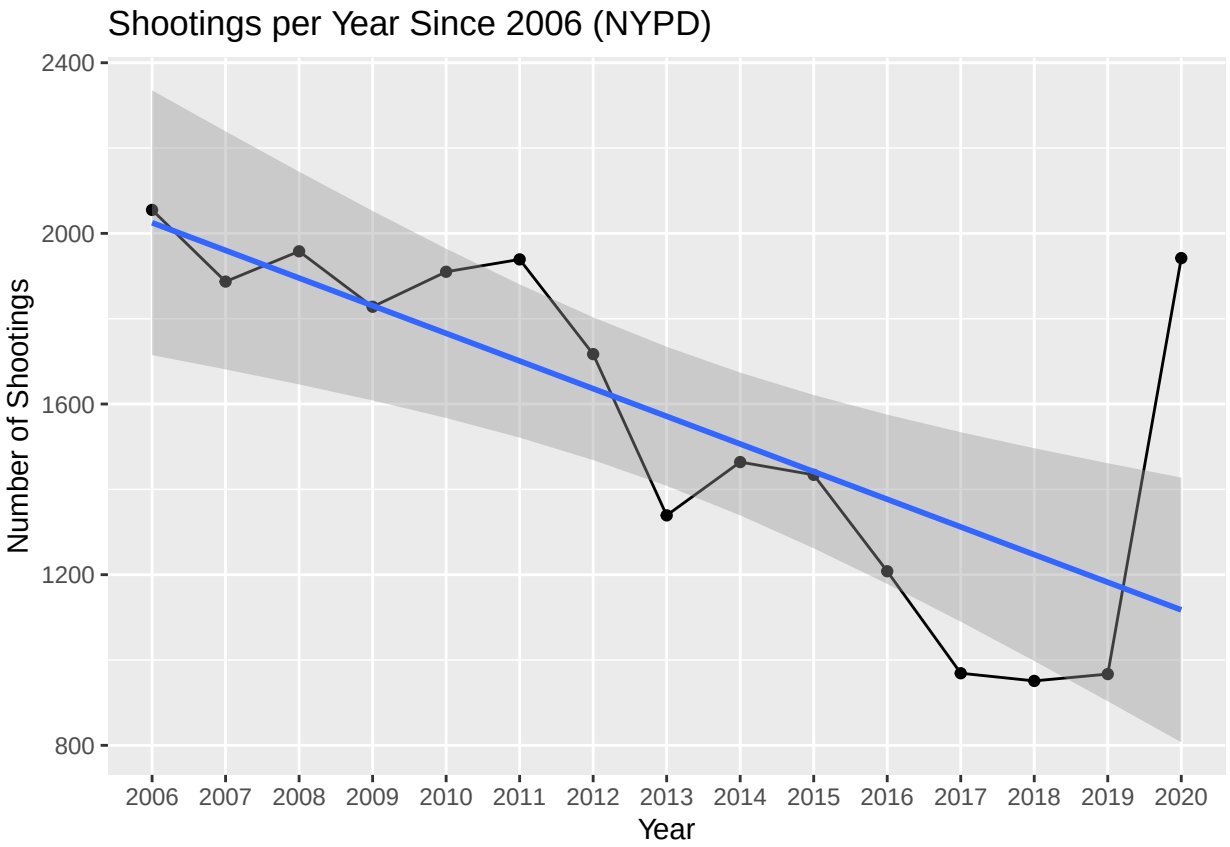
indicative of shootings happening within the neighborhood of the shooter/victim or most shootings being of a domestic nature, both of which are beyond the scope of this project; however, using the code provided, anyone could dig deeper to find why most shootings occur between people of the same race.

It is almost encouraging to see from the data that there is relatively little interracial gun violence. The belief that racial hatred is the primary motive in a lot of gun violence does not hold true in this NYPD data.

Modeling Future Gun Violence in NYC

Here, the original NYPD_data will be broken down into occurrences per year. Then a trend line forecast will be used to predict future gun violence.

```
date_data <- NYPD_data["OCCUR_DATE"]
date_data <- na.omit(date_data)
dates <- as.POSIXct(date_data$OCCUR_DATE, format = "%m/%d/%Y")
dates <- format(dates, format="%Y")

occurence_data = data.frame("Year"=c("2006","2007","2008","2009",
                                     "2010","2011","2012","2013",
                                     "2014","2015","2016","2017",
                                     "2018","2019","2020"),
                            "Incidents"=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
num_incidents = as.numeric(count(date_data["OCCUR_DATE"]))

for(i in 1:num_incidents){
  occurence_data[occurence_data$Year == dates[i], "Incidents"] =
    as.numeric(occurence_data[occurence_data$Year == dates[i], "Incidents"])+1
}

ggplot(occurence_data,
       aes(x=reorder(Year, +as.numeric(Year)),
           y=as.numeric(Incidents), group=1))+
  geom_line()+
  geom_point()+
  geom_smooth(formula = y ~ x, method="lm")+
  ggtitle("Shootings per Year Since 2006 (NYPD)")+
  xlab("Year")+
  ylab("Number of Shootings")
```



The trend line models works great during the mid 2000's, but begins to fall apart toward the end. 2020 was a tough year for many different reasons and hopefully remains an anomaly for gun violence in the future. It is safe to say that, 2020 removed, gun violence is on a downward trend in NYC.