NYPD Shooting Incident Data Report

2022-09-09

Import the necessary libraries

We will be using the tidyverse and ggplot2 libraries for this assignment.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

Import the data set

The data set is called “NYPD Shooting Incident Data (Historic)”, which lists every shooting incident occurred in New York City since 2006 until the most recent full year. It is published by the government of the City of New York and available at NYPD Shooting Incident Data (Historic) - Catalog. We will be using the Common Separated Values File.

Import the .csv format data set in the Data folder. We use a local source for this data set because the download link on the website is dynamic.

```
nypd <- read_csv("Data\\NYPD_Shooting_Incident_Data__Historic_.csv")
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, 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.
```

Cleaning the data set

The original data set contains 19 columns. We clean the data set by only retaining the information we want for the analysis.

When importing the data set, some columns are imported as characters. We convert them to factors for analysis. We also created two columns OCCUR_HOUR and OCCUR_YEAR. OCCUR_HOUR extracts information from OCCUR_TIME and is the hour at which the shooting incident happens. OCCUR_YEAR uses the information in OCCUR_DATE and is the year in which the shooting incident takes place.

```
nypd <- nypd %>% select(c(OCCUR_DATE, OCCUR_TIME, BORO,
                          STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP,
                          PERP_SEX, PERP_RACE, VIC_AGE_GROUP,
                          VIC_SEX, VIC_RACE))
```

```
nypd <- nypd %>% mutate(across(where(is.character), as_factor))
```

```
nypd$OCCUR_HOUR <- as.numeric(substr(as.character(nypd$OCCUR_TIME), 1, 2))
nypd$OCCUR_YEAR <- as.numeric(substr(as.character(nypd$OCCUR_DATE), 7, 10))

nypd <- nypd %>% select(-c(OCCUR_DATE, OCCUR_TIME))
```

```
summary(nypd)
```

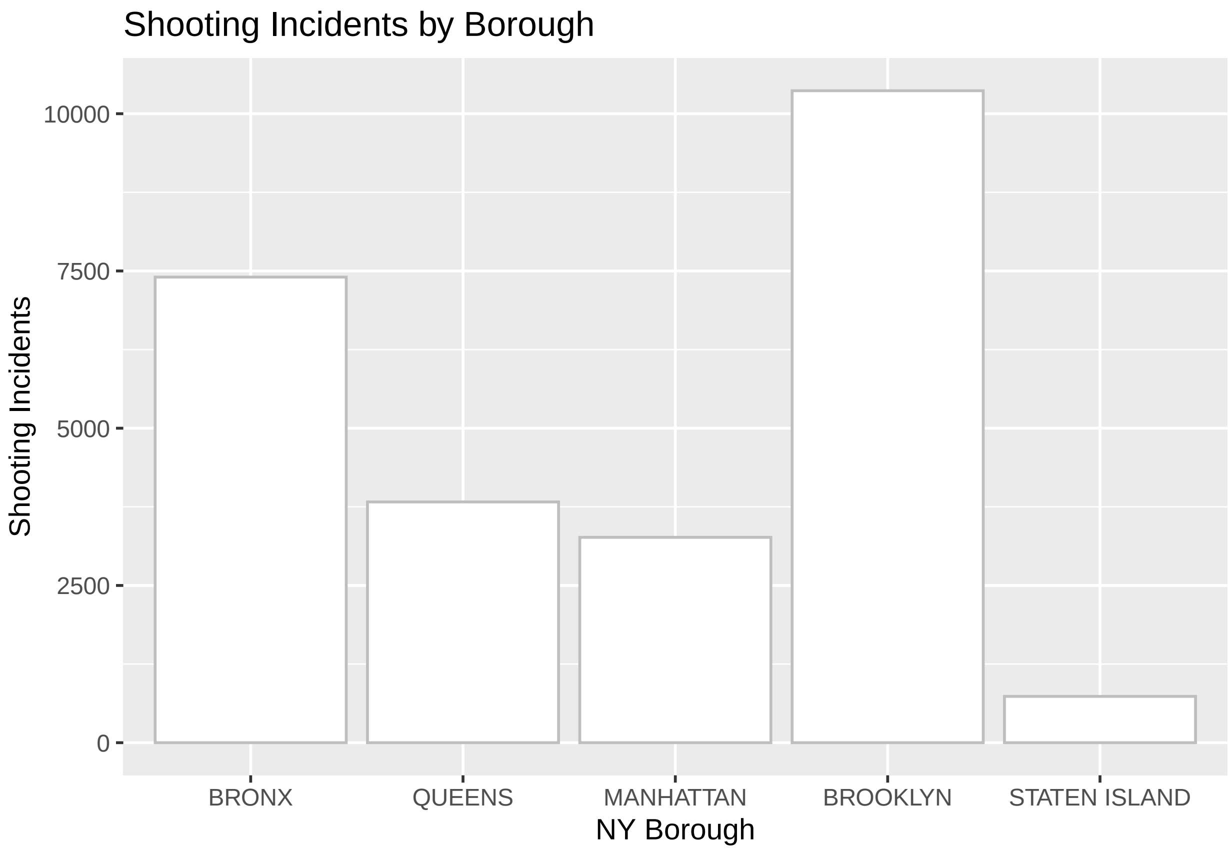
```
##             BORO       STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  BRONX        : 7402   Mode :logical           18-24  :5844   M   :14416
##  QUEENS       : 3828   FALSE:20668             25-44  :5202   F   :  371
##  MANHATTAN    : 3265   TRUE :4928              UNKNOWN:3148   U   : 1499
##  BROOKLYN     :10365                           <18    :1463   NA's: 9310
##  STATEN ISLAND:  736                           45-64  : 535
##                                                (Other):  60
##                                                NA's   :9344
##           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK         :10668   25-44  :11386   F: 2403
##  WHITE HISPANIC: 2164   65+    :  167   M:23182
##  UNKNOWN       : 1836   18-24  : 9604   U:   11
##  BLACK HISPANIC: 1203   <18    : 2681
##  WHITE         :  272   45-64  : 1698
##  (Other)       :  143   UNKNOWN:   60
##  NA's          : 9310
##                             VIC_RACE       OCCUR_HOUR      OCCUR_YEAR
##  BLACK HISPANIC                : 2485   Min.   : 0.00   Min.   :2006
##  WHITE                         :  660   1st Qu.: 3.00   1st Qu.:2009
##  BLACK                         :18281   Median :15.00   Median :2012
##  WHITE HISPANIC                : 3742   Mean   :12.19   Mean   :2013
##  ASIAN / PACIFIC ISLANDER      :  354   3rd Qu.:20.00   3rd Qu.:2017
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Max.   :23.00   Max.   :2021
##  UNKNOWN                       :   65
```

A summary of the data set shows NAs in some columns, particularly PERP_AGE_GROUP and PERP_RACE. This is probably due to the fact that in many cases the police did not have information on the perpetrator. I will remove these data entries with NAs in the modeling section. There are also some "UNKNOWN" in the data, which I would consider valid since they are entered by the police department and a very small portion of the data set.

Data Visualization

First we would like to take a look at the number of shooting incidents in each borough.

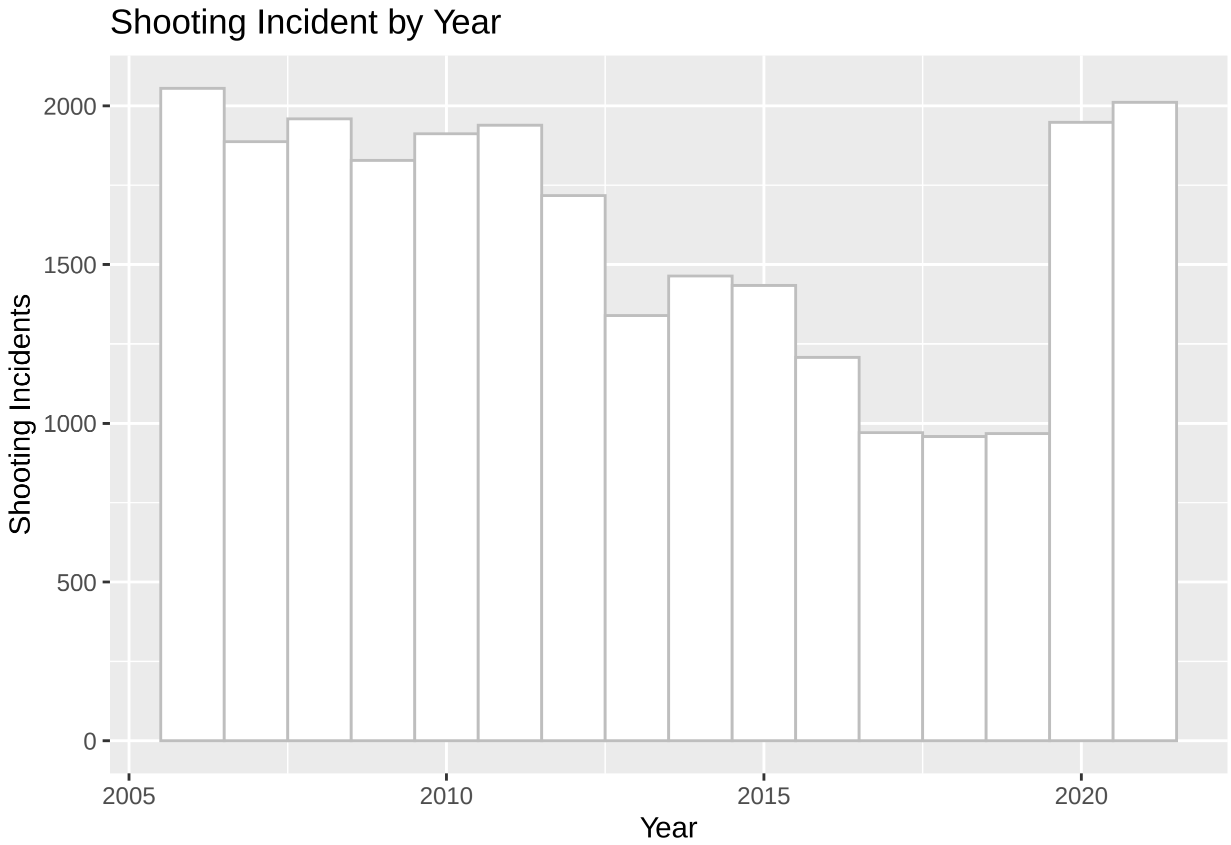
```
ggplot(nypd, aes(x=BORO)) +
  geom_bar(color='gray', fill='white') +
  ggtitle("Shooting Incidents by Borough") +
  xlab("NY Borough") +
  ylab("Shooting Incidents")
```



It is apparent from the bar chart that Brooklyn has by far the most cases of shooting incidents. Bronx has the second most cases, and significantly more than that of Queens or Manhattan. Staten Island has the lowest amount of shooting incidents out of the five boroughs. However, it is difficult to draw any reasonable conclusion without understanding of the New York demographics.

Next, we would like to take a look at the number of shooting incidents per year.

```
ggplot(nypd, aes(x=OCCUR_YEAR)) +
  geom_histogram(color='gray', fill='white', binwidth=1) +
  ggtitle("Shooting Incident by Year") +
  xlab("Year") +
  ylab("Shooting Incidents")
```

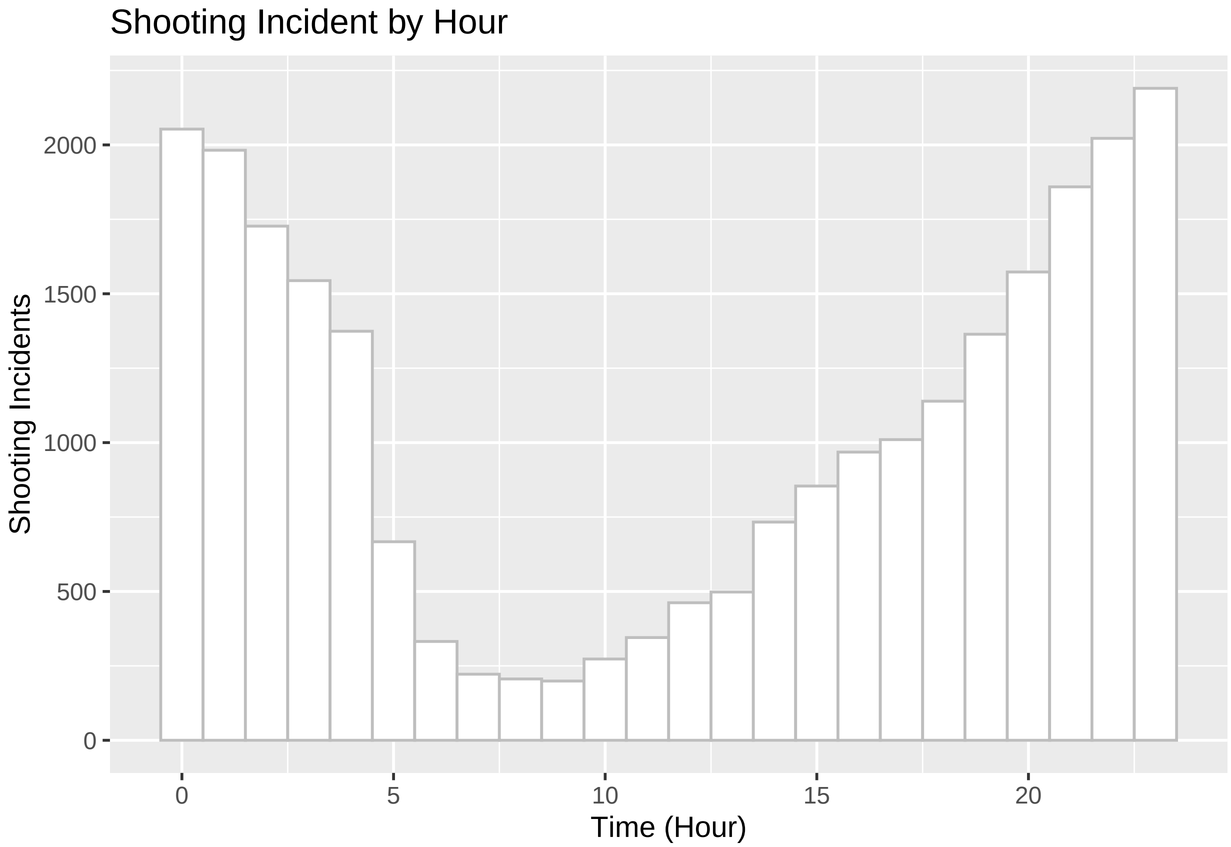


From the chart, we can clearly see a decreasing pattern until the year 2020. More police patrol, stricter gun control, and better education are examples of possible contributor to the decline in shooting incidents.

The COVID-19 pandemic likely contributed to the resurgence in shooting incidents starting from 2020. As we know, many people experienced financial difficulties during the pandemic and some may be inclined to use extreme force and commit crimes.

Next, we would like to take a look at the time during which the shooting incidents take place.

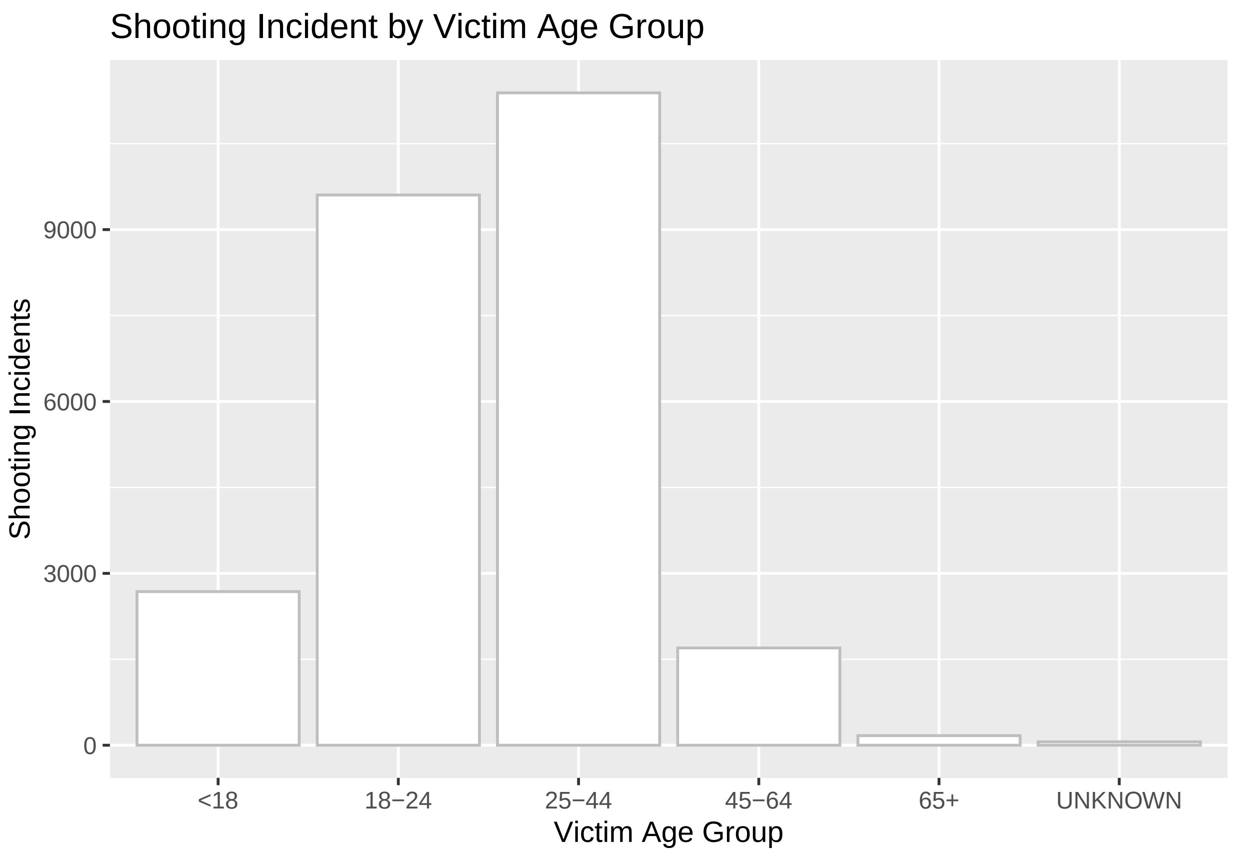
```
ggplot(nypd, aes(x=OCCUR_HOUR)) +
  geom_histogram(color='gray', fill='white', bins=24) +
  ggtitle("Shooting Incident by Hour") +
  xlab("Time (Hour)") +
  ylab("Shooting Incidents")
```



From the chart, it is clear that a significant portion of shooting incidents happened during the night. We can see the lowest point of the chart is at 8 o'clock. Shooting Incidents continuously go up until midnight, at which the number of shooting incidents starts to drop. This is consistent with the common sense that criminals are less likely to commit crimes during the day.

We would then like to take a look at the shooting incidents victims, more specifically their ages.

```
nypd$VIC_AGE_GROUP <- factor(nypd$VIC_AGE_GROUP,
                             levels=c("<18","18-24","25-44",
                                      "45-64", "65+", "UNKNOWN"))
ggplot(nypd, aes(x=VIC_AGE_GROUP)) +
  geom_bar(color='gray', fill='white') +
  ggtitle("Shooting Incident by Victim Age Group") +
  xlab("Victim Age Group") +
  ylab("Shooting Incidents")
```



As shown in the chart, adults from 18 to 44 make up the majority of victims in shooting incidents. It appears that young adults aged from 18 to 24 are most likely to be the victims of shooting incidents. Despite more cases in the "25-44" group, there are likely more people in this group given the larger age span.

Minors who are under the age of 18 also account for a portion of the victims in shooting incidents. I am intrigued to see if shooting incidents involving minors have any distinct characteristic that can be used to predict if the victim is a minor. Before creating a model, we first create a new column "VIC_UNDERAGE" that is TRUE if the victim is a minor.

```
nypd$VIC_UNDERAGE = FALSE
nypd$VIC_UNDERAGE[nypd$VIC_AGE_GROUP=="<18"] = TRUE
```

Modeling and Prediction

Since there are NAs in the data set, we need to remove these. We then apply a binomial logistic regression model on every columns except VIC_AGE_GROUP to predict VIC_UNDERAGE. Using this model, we are able to obtain a vector of possibilities that VIC_UNDERAGE is TRUE. Using a threshold of p=0.5, we compare the prediction with actual data in a table.

```
nypd <- nypd %>% drop_na()
model <- glm(VIC_UNDERAGE ~ . -VIC_AGE_GROUP, data=nypd, family=binomial)
pre1 <- predict(model, nypd, type="response")
pre1.result <- ifelse(pre1>0.5, TRUE, FALSE)
result <- table(Prediction=pre1.result, Actual=nypd$VIC_UNDERAGE)
result
```

```
##            Actual
## Prediction FALSE  TRUE
##      FALSE 14360  1848
##      TRUE     23    21
```

```
result/length(pre1.result)
```

```
##            Actual
## Prediction       FALSE        TRUE
##      FALSE 0.883583559 0.113709082
##      TRUE  0.001415210 0.001292149
```

```
sum(diag(result))/sum(result)
```

```
## [1] 0.8848757
```

Looking at the table above, the model did not predict many TRUE cases. Out of the 44 cases the model predicts to be TRUE, 23 are Type I errors or false positives. Despite predicting 14360 or 88.36% true negative, the model also produces 1848 or 11.37% Type II errors or false negatives. In total, the model predicts approximately 88.49% of the data set correctly.

Next, we would like to see if a lower threshold p value would help. For the second prediction, we use p=0.3 to make it easier to predict TRUE.

```
pre1.result2 <- ifelse(pre1>0.3, TRUE, FALSE)
result2 <- table(Prediction=pre1.result2, Actual=nypd$VIC_UNDERAGE)
result2
```

```
##            Actual
## Prediction FALSE  TRUE
##      FALSE 13898  1604
##      TRUE    485   265
```

```
result2/length(pre1.result2)
```

```
##            Actual
## Prediction      FALSE       TRUE
##      FALSE 0.85515629 0.09869555
##      TRUE  0.02984248 0.01630569
```

```
sum(diag(result2))/sum(result2)
```

```
## [1] 0.871462
```

As shown in the results above, we observe an increasing amount of false positives and the accuracy of the prediction drops to 87.15%.

Despite getting relatively high accuracy, the classification model isn't very effective since it doesn't do a good job detecting the TRUE cases.

One of the possible reason is that shooting incidents are very complicated issues and there is simply not enough information in the data set to predict if the victim is a minor.

It is also possible that a binary logistic regression classification isn't suited for the task here. I believe that a decision tree model may have been a better choice.

Discussion and Conclusion

In Data Visualization, we take a very brief look at the shooting incidents in New York City. We discovered the number of shooting incidents in each borough in New York City. We discussed the trend of shooting incidents over the past 16 years. We also discussed the more likely time that shooting incidents took place.

There are certain limitations that prevent us from getting useful conclusion from this analysis.

One of the key limitation is that the data set does not contain any information about the demographics, which makes the numbers difficult to compare. For example, in our first visualization, we discovered that Brooklyn has the most cases of shooting incidents. But Brooklyn also has the largest population out of the five boroughs. A more useful statistics would be the number of shooting accidents per million of population, for instance, which would require external source of data.

Another limitation is that the data set does not give us the full picture of the nature of these shooting incidents. There are also missing information because sometimes the perpetrators get away.

An example would be that we only know if the victim is under the age of 18. However, as one approaches the age of 18, they behave more like adults rather than children. When we are trying to learn things about young teens, our analysis could have been contaminated by those who are 16 or 17.

A source of bias is that there could potentially be unreported shooting incidents that the Police Department has no information on. The Police may sometime rely on the information from witnesses, which is not always accurate.

As a foreigner, I am also a source of bias because I lack the background knowledge and understanding of social problems in the United States. I purposely avoid the discussion of race, which could have been a huge part of the analysis.