# NYPD Shooting Incident Dataset Analysis

11/6/2021

## About

The following is an analysis of shooting incidents within New York City, New York from 2006 to 2020. The information is compiled quarterly and added to a csv file available on the City of New York's website. Attributes such as the victim's/perpetrator's gender, race, or age as well as the location of the incidents and other information can be found within the file.

## Code

### Step 1: Import Libraries

The first portion of this analysis is writing code to get needed information to generate models and visualizations. Initially, the correct libraries are called to help with pulling the data.

```
# Import Libraries
library(tidyverse)
library(tinytex)
```

The first library has useful tools for data analysis and the second is used to convert the R code output into a PDF.

### Step 2: Import Dataset

Next, the information from the website is extracted.

```
# Read csv file and save information
url_in<-"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
NYPD_cases<-read_csv(url_in)
```

This is done through reading the url and the information from the CSV file within it.

After, the columns being used are transformed into useful states.

### Step 3: Clean, Tidy, and Select Important Data

The column names are first obtained to have an insight into which columns might be useful for extracting. The first column has the perpetrators' age group which is put into a variable and the N/As (or Not Applicable data points) are removed since these were deemed to not add value to the graph or didn't have a perpetrator. There were typos or age ranges that couldn't exist which were also removed as well. Next the X and Y coordinates used by the police department are placed into a tibble for future use. Then the occurrence dates have their shooting incidents summed per month and are arranged in a monthly format.

```r
# Read Columns
colnames(NYPD_cases)
```

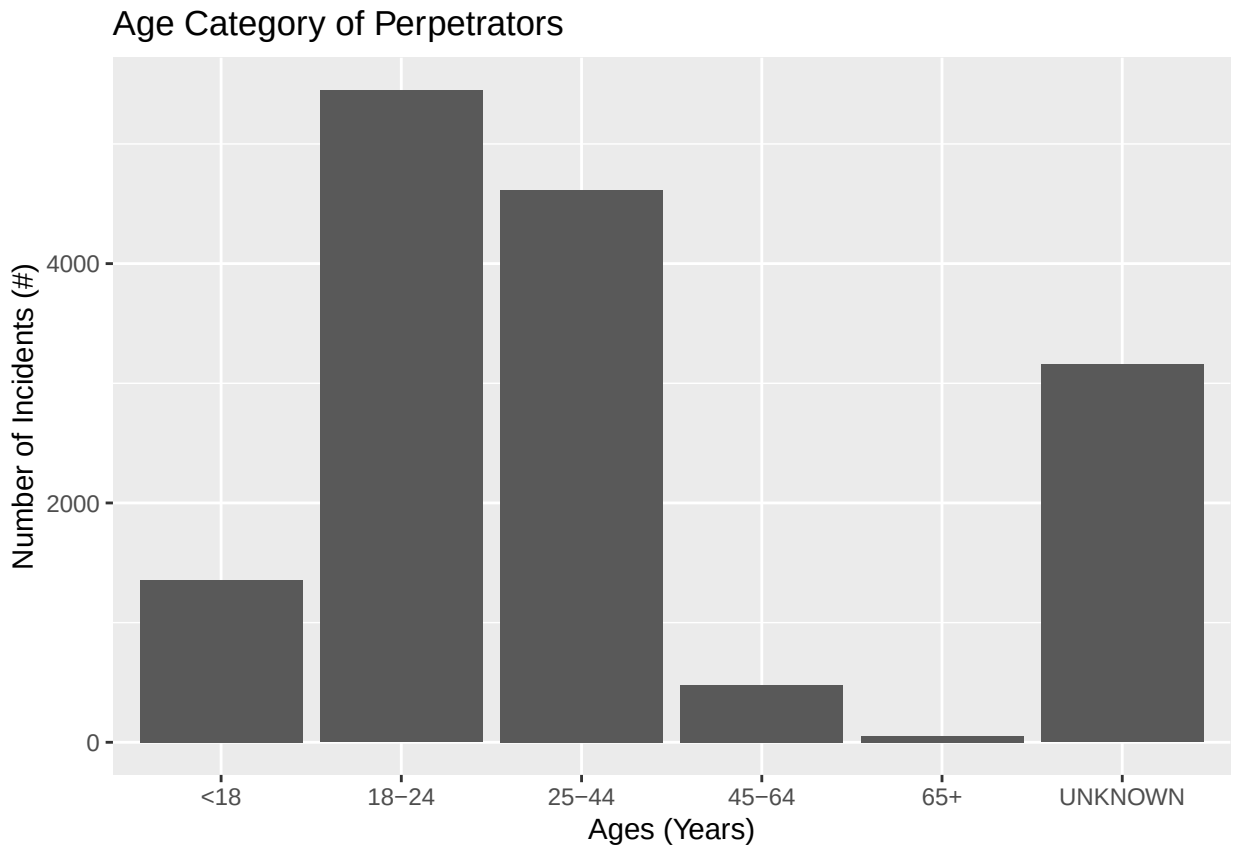
```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "PRECINCT"                "JURISDICTION_CODE"
##  [7] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
##  [9] "PERP_AGE_GROUP"          "PERP_SEX"
## [11] "PERP_RACE"               "VIC_AGE_GROUP"
## [13] "VIC_SEX"                 "VIC_RACE"
## [15] "X_COORD_CD"              "Y_COORD_CD"
## [17] "Latitude"                "Longitude"
## [19] "Lon_Lat"
```

```r
# Selecting "PERP_AGE_GROUP" column
NY1<-NYPD_cases %>%
  select(PERP_AGE_GROUP)
# Remove Typos or Entries that Happen Once
NY1<-NY1[NY1$PERP_AGE_GROUP %in%names(which(table(NY1$PERP_AGE_GROUP)>2)),]
# Remove N/As
NY1<-na.omit(NY1)

# Selecting "X_COORD_CD" and "Y_COORD_CD" into a tibble
NY2<-NYPD_cases%>%
  select(X_COORD_CD,Y_COORD_CD)

# Get a column of the months and a tally for each month of the incidents
NY4<-NYPD_cases%>%group_by(as.integer(substr(NYPD_cases$OCCUR_DATE,0,2))) %>% tally()
# Name new column
names(NY4)[1]='month'
```

## Step 4: Produce Visualization #1

Below the first visualization is made. This is a bar graph which shows the age ranges and shooting incidents that take place per group for the perpetrators.

```r
# Generate figure of age categories and shooting incidents
p1<-ggplot(NY1,aes(x=PERP_AGE_GROUP))+
  geom_bar()+
  labs(title="Age Category of Perpetrators",
  x="Ages (Years)",
  y="Number of Incidents (#)")
p1
```

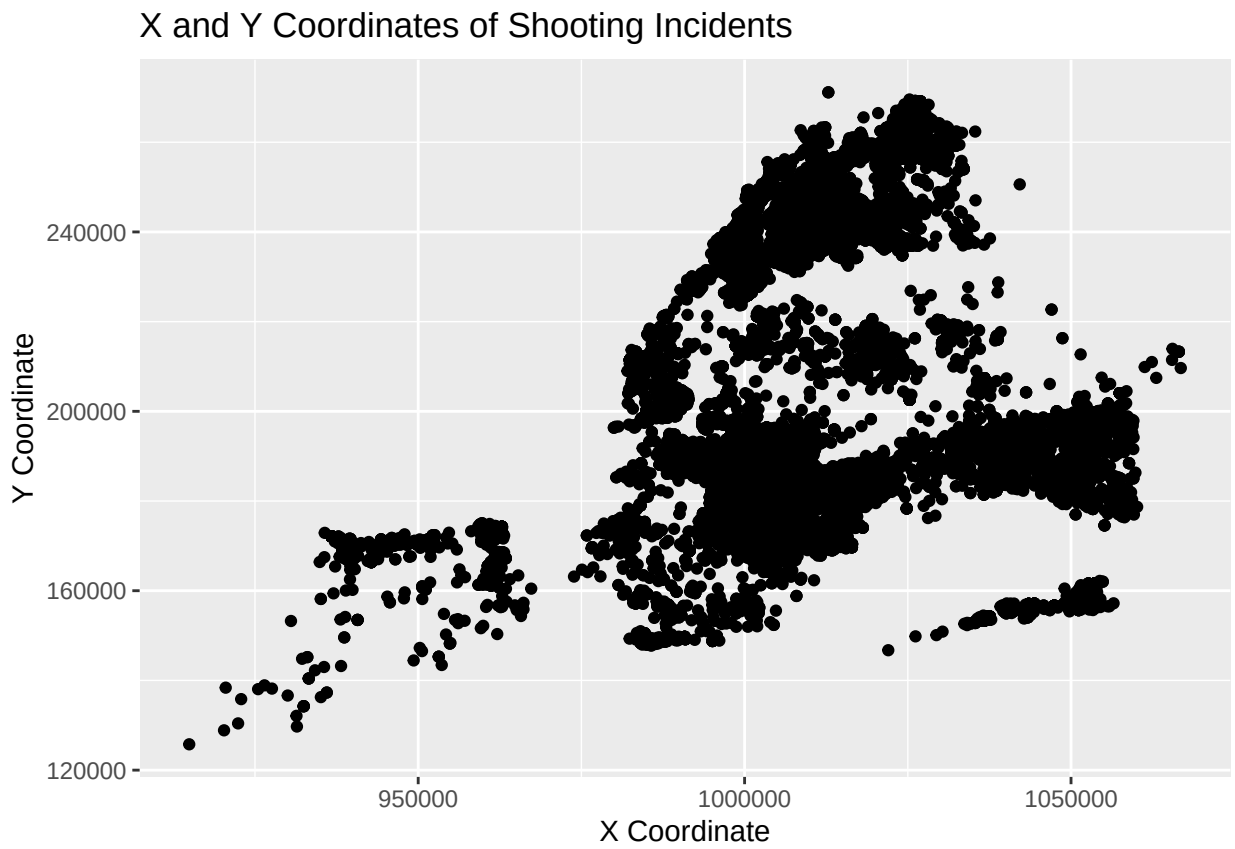



Above is the visualization of perpetrators' ages. There was a significant amount of unknowns. If the data points in this range were determined, there is a chance that another group's numbers would significantly raise. This is potentially important information that should be taken into consideration when determining laws when it comes to reducing crime.

### Step 5: Produce Visualization #2

Next, the second visualization of a map is created. This makes use of the X and Y coordinates provided to show pockets of high shooting activity.

```
# Generate figure of x and y coordinates
p2<-ggplot(NY2,aes(x=X_COORD_CD, y=Y_COORD_CD))+
    geom_point()+
    labs(title = "X and Y Coordinates of Shooting Incidents",
    x="X Coordinate",
    y="Y Coordinate")
p2
```



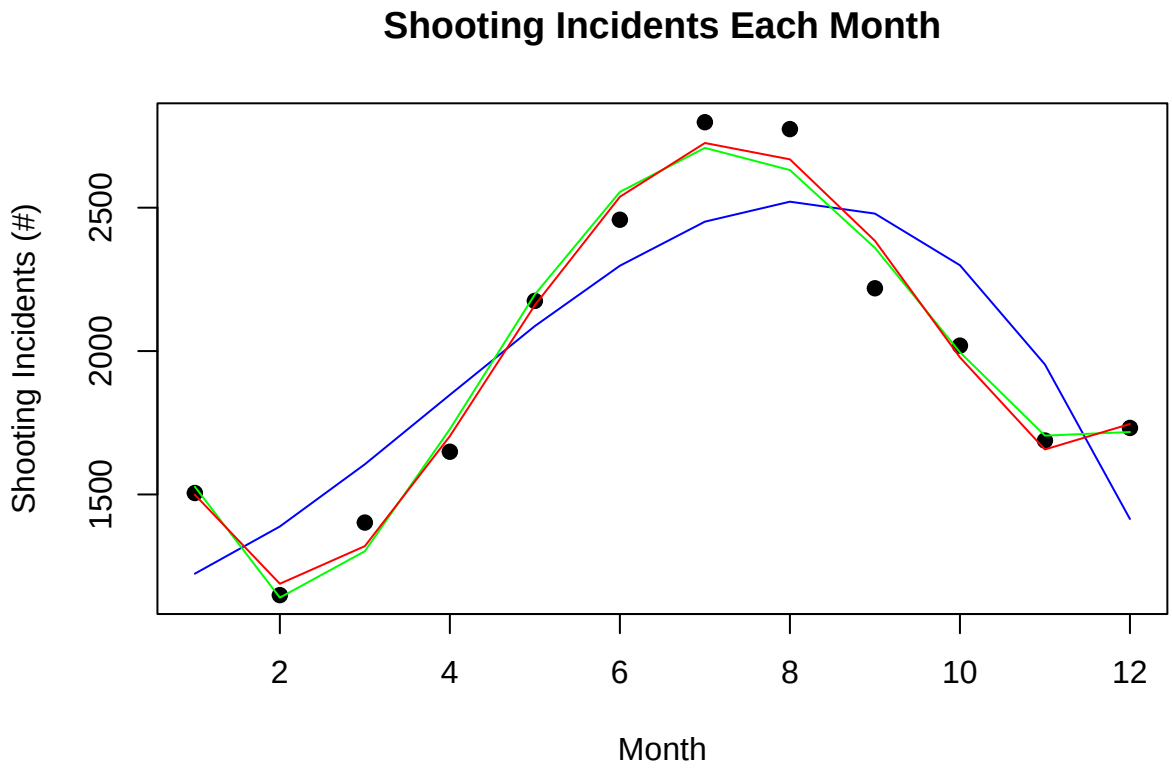

According to the website, the data points are the "midblock" X and Y-coordinates "for New York State Plane Coordinate System, Long Island Zone". Where the scatter plot heavily clusters indicates that there is a higher chance of shooting activity in those areas or neighborhoods. This figure locates parts that might need more patrolling or stricter laws applied.

### Step 6: Generate Model

Finally, the model of shooting incidents per month for New York City is generated. Data for the shooting incidents monthly is plotted with a 3rd (blue), 4th (green), and 5th (red) polynomial curve applied to them.

```
# Rename columns
names(NY4)[1]='x'
names(NY4)[2]='y'

# Generate plot
plot(NY4$x,NY4$y,pch=19, main='Shooting Incidents Each Month',
     xlab='Month',ylab='Shooting Incidents (#)')

# Create the polynomial values using x and y
f3<-lm(y~poly(x,3,raw=TRUE),data=NY4)
f4<-lm(y~poly(x,4,raw=TRUE),data=NY4)
f5<-lm(y~poly(x,5,raw=TRUE),data=NY4)

# Create x-axis values for month
x_axis<-seq(1,12,length=12)
```

```
# Show polynomial fits
lines(x_axis,predict(f3,data.frame(x=x_axis)),col='blue')
lines(x_axis,predict(f4,data.frame(x=x_axis)),col='green')
lines(x_axis,predict(f5,data.frame(x=x_axis)),col='red')
```




Above is the expected amount of shooting incidents to take place over a year based on information from 2006 to 2020. During the summer time, there is an increase in shooting incidents as the numbers peak in July and drop down going into fall. The least amount of shootings took place in the winter with February having the smallest amount. This however takes simplified information from the past 14 years (2006 - 2020) and doesn't evaluate whether there have been large changes in judicial infrastructure, policing policies, growth within the city, how much crime has changed yearly, and other factors that could invalidate this. Thus, this is a rough model open for feedback and to be modified with more information.

Below are the R squared values for the polynomial fits applied.

```
# Provides R squared of the polynomial fits applied
# 3rd degree polynomial
summary(f3)$adj.r.squared
```

```
## [1] 0.6692876
```

```
# 4th degree polynomial
summary(f4)$adj.r.squared
```

```
## [1] 0.9618545
```

```
# 5th degree polynomial
summary(f5)$adj.r.squared
```

```
## [1] 0.9622065
```

The 5th fit has the highest value and thus a 5th degree polynomial is best for simulating the cyclical nature of shootings within New York.

## Conclusions

**Visualization #1:** If the unknown data points are determined to belong to a group, that group should receive more attention and be targeted further to reduce shooting incidents more.

**Visualization #2:** Pockets of New York are more heavily impacted by shootings and can belong to certain neighborhoods.

**Model:** A 5th degree polynomial correlates to a more cyclical (than say logarithmic) nature of shootings throughout the year when taken on a 14 year scale from 2006 to 2020.

## Potential Forms of Biases

The second largest group in the first figure had an age range that went from 25 to 44 years old. That age range isn't exactly what first comes to my mind when imagining perpetrators of shooting incidents, but is significant and rivals the dominating one in numbers.

When it came to the image of the X and Y coordinates of shooting incidents, I thought that certain neighborhoods must have associated biases with them due to having a relatively large amount of shootings compared to others. This would feed a bias when it comes to people who live in those areas due to surrounding shooting crime rates. However, correlation doesn't always equal causation and not everyone who lives in those neighborhoods should be associated with these incidents.

These are all reminders that data should always be taken carefully and that there aren't absolutes when it comes to statistics. Biases are self-propagating. If an analyst has a bias inherently then they are more likely to generate data that fulfills these biases. Then others will see these biases, which will confirm their biases or change their mind into believing them, and be more likely to generate data that fulfills these biases and so forth. Thus, it is pertinent to analyze one's data critically as well as objectively.