NYPD Shooting Incident Data Analysis

Conrad Kleykamp

2023-01-30

NYPD Shooting Incident Data Analysis

This data set contains information regarding shooting incidents in NYC neighborhoods from 01/01/2006 to 12/31/2021. This data is extracted quarterly and is reviewed by the Office of Management Analysis and Planning before it is posted on the NYPD website. Each row of data represents a single shooting incident. This data set is available for public access and use.

Setup/Loading Libraries

These steps will ensure that all necessary packages and installed and loaded.

```
# Install and load tidyverse for future use
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/m7/bp8vnrh95sj669xxqwgwzh_c0000gn/T//Rtmpporhu0/downloaded_packages
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Load lubridate for future use
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
# Load dplyr for future use
library(dplyr)
```

```r
# Load ggplot2, ggthemes
library(ggplot2)
library(ggthemes)
```

Importing the Data

The data will be imported from the link available on the data.gov website. The link will be read in as 'NYPD_data' for ease of use.

```r
# Get data from online link
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

```r
# Read data from link
NYPD_data <- read_csv(url_in)
```

```
## 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.
```

Tidying and Transforming Data

The following steps will ensure that the data is ready for further analysis. For this particular assignment, I chose to focus my analysis on solely the number of shooting incidents and deaths across time. Because of this, I have opted to remove several nonessential columns from the data set. Please note that a few retained columns were not necessary for this particular analysis. I have not filtered these out as I plan to return to this project and conduct further analyses.

It should also be noted that some columns are empty for many shooting incidents, such as LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, and PERP_RACE. This is to be expected, as many shooting incidents are recorded after the fact and thus the identity of the perpetrator is unknown. If the reader of this project wished to analyze the demographics of the perpetrators, I would suggest to filter out all columns that are empty. I have not done this for my particular analysis, as I found that it would heavily skew the data and underrepresent the number of shooting incidents and deaths.

```r
# Many of the columns in the original data set will not be needed for this analysis.
# I have opted to remove INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, LOCATION_DESC,
# X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat
NYPD_data <- NYPD_data %>%
  select(-c(INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, LOCATION_DESC,
            X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
```

```r
# Format OCCUR_TIME to hms
NYPD_data <- NYPD_data %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME))
```

```r
# Format OCCUR_DATE to mdy
NYPD_data <- NYPD_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

```r
# Upon quick inspection, several values in PERP_AGE_GROUP seemed incorrect.
# These values have been removed to avoid potentially incorrect data.
NYPD_data <- NYPD_data %>%
  filter(PERP_AGE_GROUP != "940", PERP_AGE_GROUP != "1020",
         PERP_AGE_GROUP != "224")
```

```r
# Rename STATISTICAL_MURDER_FLAG column to "MURDER" for ease of use
NYPD_data <- NYPD_data %>%
  rename(MURDER_FLAG = "STATISTICAL_MURDER_FLAG")
```

```r
# View a quick summary of the transformed data
summary(NYPD_data)
```

```
##    OCCUR_DATE           OCCUR_TIME                          BORO
##  Min.   :2006-01-01   Min.   :0S                        Length:16249
##  1st Qu.:2008-06-01   1st Qu.:3H 46M 0S                 Class :character
##  Median :2011-01-16   Median :15H 22M 0S                Mode  :character
##  Mean   :2012-05-25   Mean   :12H 55M 16.3505446488853S
##  3rd Qu.:2016-03-06   3rd Qu.:20H 39M 0S
##  Max.   :2021-12-31   Max.   :23H 59M 0S
##  MURDER_FLAG     PERP_AGE_GROUP       PERP_SEX          PERP_RACE
##  Mode :logical   Length:16249       Length:16249       Length:16249
##  FALSE:13027     Class :character   Class :character   Class :character
##  TRUE :3222      Mode  :character   Mode  :character   Mode  :character
##
##
##
##  VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:16249       Length:16249       Length:16249
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Analysis

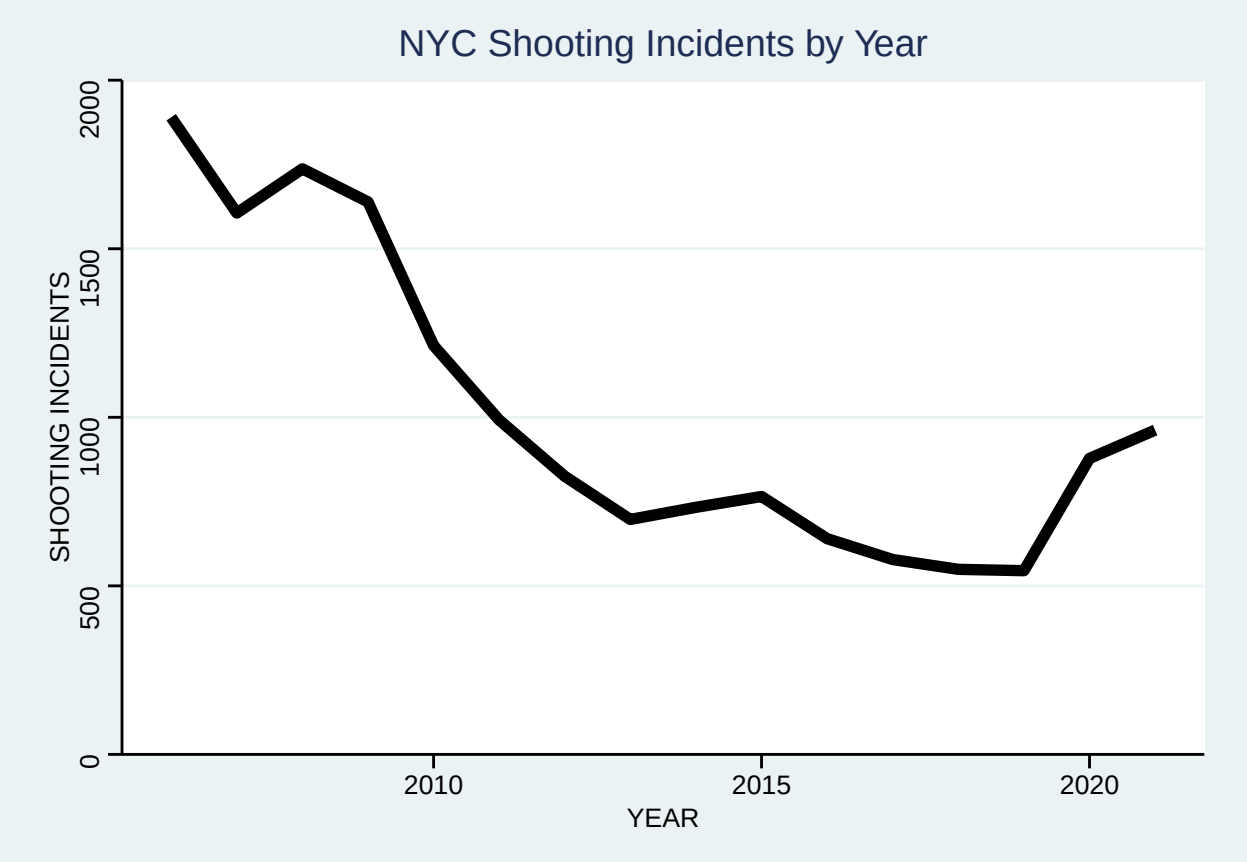
As mentioned above, I have opted to analyze the trend of shooting incidents and resulting deaths between the years of 2006 and 2021. The brief visualizations below will help provide rapid insight into any trends. Of course, one could also opt to analyze this data on a monthly or daily time frame for more refined trends.

```r
# Create YEAR column for analysis by year
NYPD_data$YEAR <- year(NYPD_data$OCCUR_DATE)
```

```r
# Group shooting incidents by year
NYPD_shootings_by_year <- NYPD_data %>%
  group_by(YEAR) %>%
  summarise(SHOOTING_INCIDENTS = n())
```

```r
# Plot shooting incidents by year
NYPD_shootings_by_year %>%
  ggplot(aes(x = YEAR, y = SHOOTING_INCIDENTS))+
  geom_line(size = 2)+
  theme_stata()+
  ggtitle("NYC Shooting Incidents by Year")+
  xlab("YEAR")+
  ylab("SHOOTING INCIDENTS")+
  scale_y_continuous(expand = c(0, 0), limits = c(0, 2000))
```

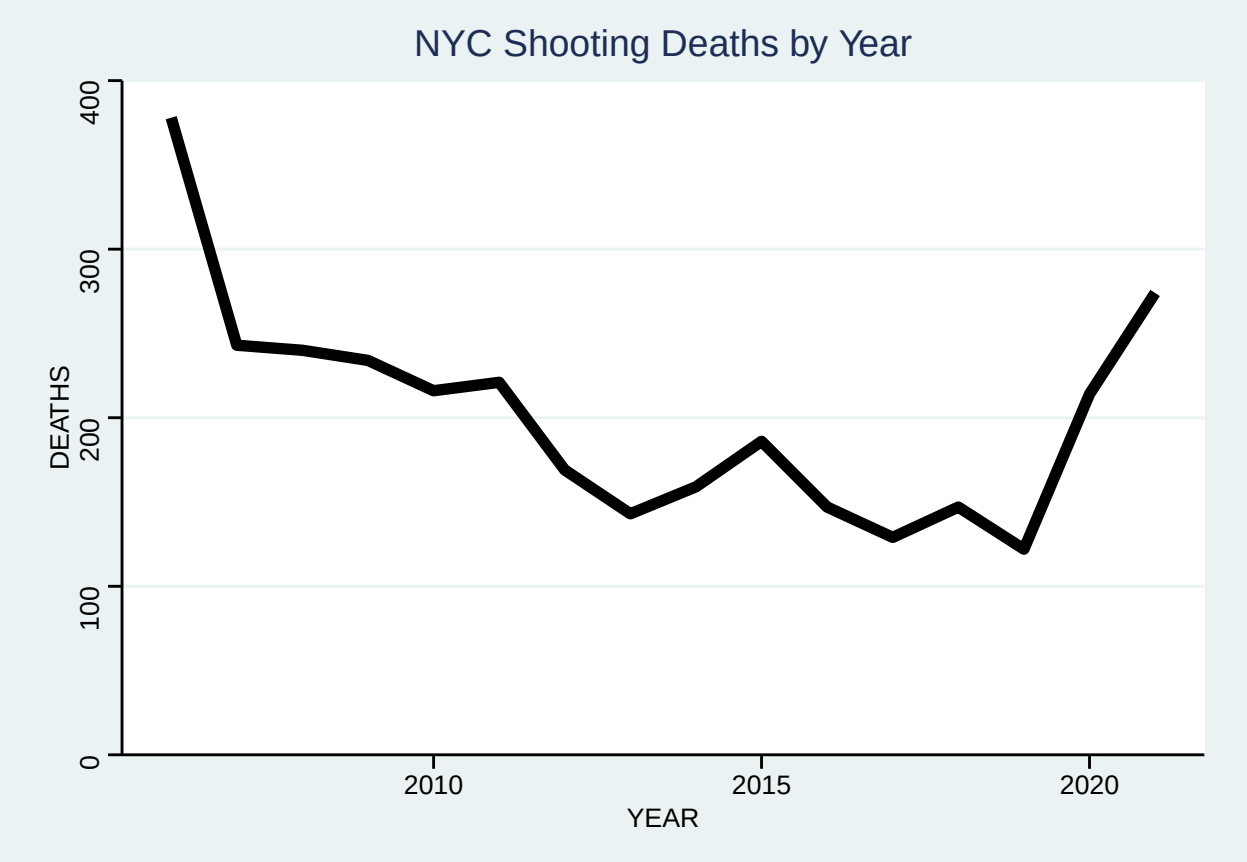
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```



The visualization above represents the trend of shooting incidents from 2006 to 2021. It is clear that there was a significant decrease in shooting incidents over this time period. This trend appeared to reverse before and during 2020. As I lack expertise in NYC law and crime policy, it would be intriguing to uncover whether or not particular policies or actions contributed to the significant decline in shooting incidents from 2006 onward. Moreover, I question whether or not the Covid-19 pandemic contributed to an increase in incidents during 2020 and beyond.

```r
# Sort shooting deaths by year
NYPD_deaths_by_year <- NYPD_data %>%
  group_by(YEAR) %>%
  summarize(DEATHS = sum(MURDER_FLAG))
```

```r
# Plot shooting deaths by year
NYPD_deaths_by_year %>%
  ggplot(aes(x = YEAR, y = DEATHS))+
  geom_line(size = 2)+
  theme_stata()+
  ggtitle("NYC Shooting Deaths by Year")+
  xlab("YEAR")+
  ylab("DEATHS")+
  scale_y_continuous(expand = c(0, 0), limits = c(0, 400))
```



The trend of deaths resulting from shooting incidents appears to closely resemble the prior visualization. It should be noted, however, that the number of deaths is far lesser than the number of shooting incidents. The similarity in trends suggests that shooting incidents and deaths may be linearly correlated. While this hypothesis may outwardly appear obvious, I will create models to support it.

Modeling

The first model will aim to model the number of deaths (monthly) as a function of total shootings.

```r
# Create month column for future modeling
NYPD_data$month <- NYPD_data$OCCUR_DATE %>%
  month()
```

```r
# Model the number of deaths in a month as a function of the total shootings
NYPD_total <- NYPD_data %>%
  group_by(YEAR, month) %>%
  summarise(SHOOTING_INCIDENTS = n(), DEATHS = sum(MURDER_FLAG))
```

```
## `summarise()` has grouped output by 'YEAR'. You can override using the
## `.groups` argument.
```

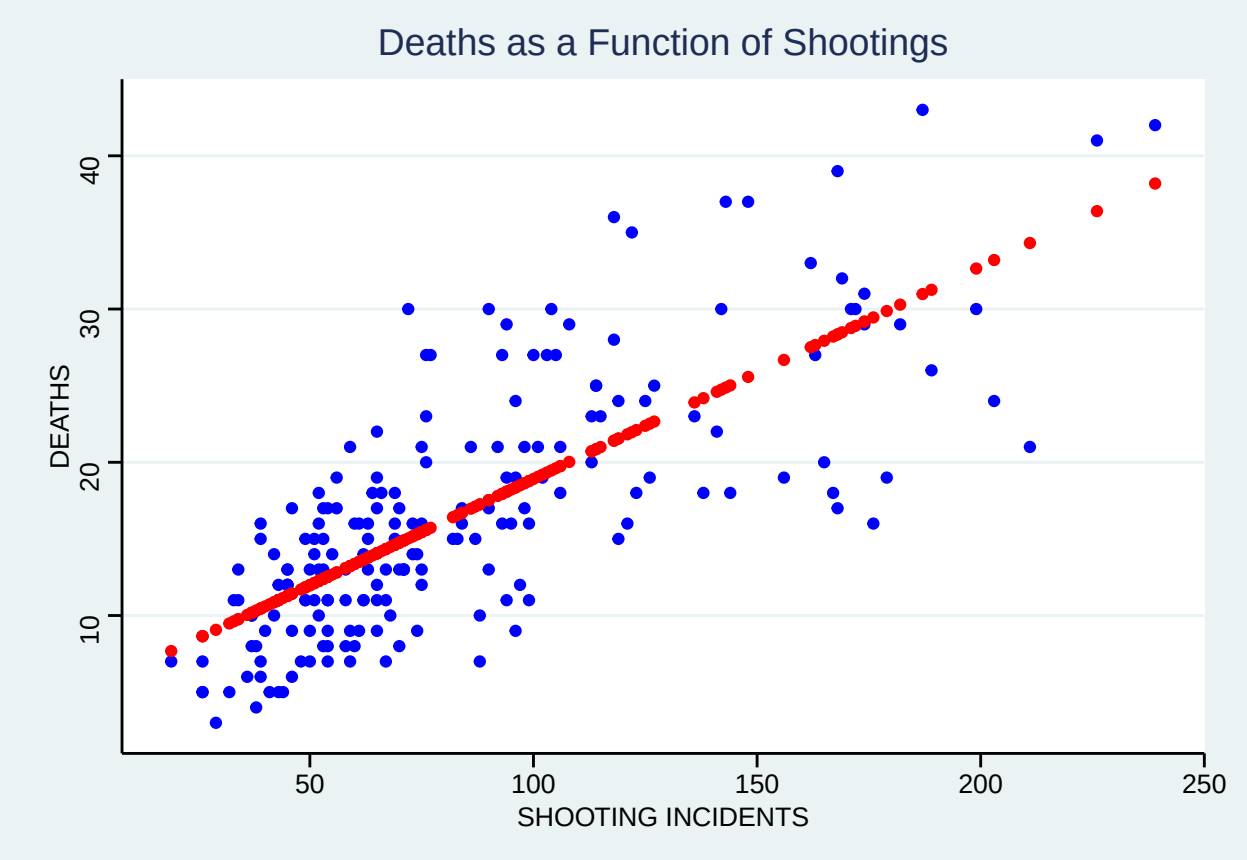
```r
X <- NYPD_total$SHOOTING_INCIDENTS
Y <- NYPD_total$DEATHS
```

```r
# Model setup
model <- lm(Y ~ X)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4539  -3.5226  -0.3582   3.1387  14.9705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.04334    0.82893   6.084 6.32e-09 ***
## X            0.13870    0.00865  16.034  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.388 on 190 degrees of freedom
## Multiple R-squared:  0.575,  Adjusted R-squared:  0.5728
## F-statistic: 257.1 on 1 and 190 DF,  p-value: < 2.2e-16
```

```r
# Plotting the model
Y_predict <- predict(model)

ggplot()+
  geom_point(aes(x = X, y = Y), color = 'blue')+
  geom_point(aes(x = X, y = Y_predict), color = 'red', show.legend = TRUE)+
  scale_color_manual(name = 'Legend', values = c("Predicted Counts" = 'red'))+
  theme_stata()+
  ggtitle("Deaths as a Function of Shootings")+
  xlab("SHOOTING INCIDENTS")+
  ylab("DEATHS")
```

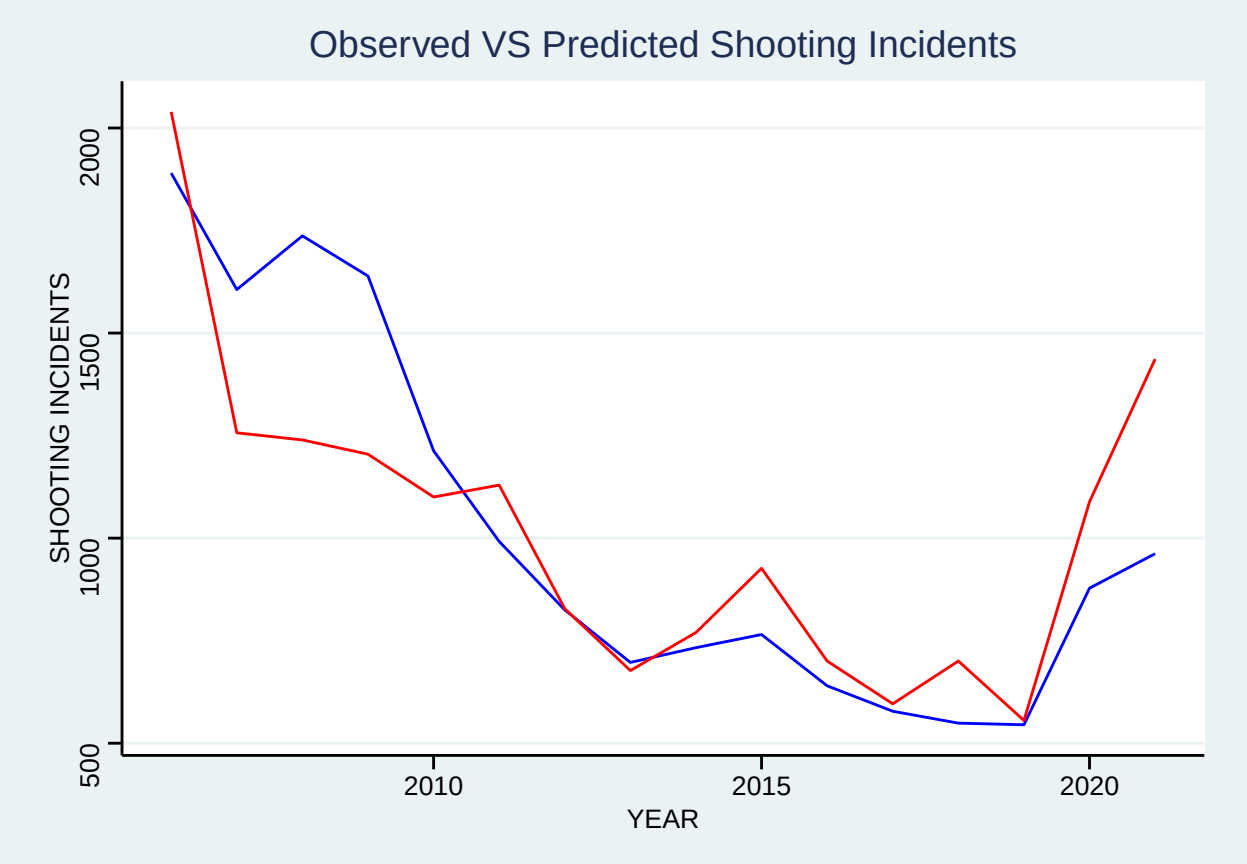


The model above shows a linear correlation between shooting incidents and deaths, i.e. as shooting incidents increase, so do the number of deaths. The blue circles represent actual data, while the red circles represent the projected linear trend. As mentioned above, the number of deaths is significantly lower than the number of shooting incidents.

The next model will aim to further prove the linear relationship between deaths and shooting incidents by deriving predicted shooting incidents from actual death counts.

```
# model predicted shooting incidents
NYPD_total_counts <- cbind(NYPD_shootings_by_year, NYPD_deaths_by_year[,"DEATHS"])
model2 = lm(SHOOTING_INCIDENTS ~ DEATHS, data = NYPD_total_counts)
NYPD_total_counts <- cbind(NYPD_total_counts, predCounts = predict(model2))
```

```
# Plotting the model
NYPD_total_counts[,c('YEAR', 'SHOOTING_INCIDENTS', 'predCounts')] %>%
  group_by(YEAR) %>%
  summarize(SHOOTING_INCIDENTS = sum(SHOOTING_INCIDENTS),
            predCounts = sum(predCounts)) %>%
  ggplot(aes(group = 1))+
  geom_line(aes(x = YEAR, y = SHOOTING_INCIDENTS), color = 'blue')+
  geom_line(aes(x = YEAR, y = predCounts), color = 'red', show.legend = TRUE)+
  scale_color_manual(name = 'Legend', values = c("Predicted Deaths" = 'red'))+
  theme_stata()+
  ggtitle("Observed VS Predicted Shooting Incidents")+
  xlab("YEAR")+
  ylab("SHOOTING INCIDENTS")
```



The model above shows a close relationship between the predicted shooting incidents (red) and the actual shooting incidents (blue) over time. As the predicted incidents were derived from actual death counts, this further demonstrates a linear relationship between shooting incidents and deaths.

Bias Sources

1) Filtering out odd data During my transformation steps, I opted to remove potentially incorrect data. These data had improbable values for the perpetrator's age, such as 940, 224, and 1020. My bias led me to believe that this information may have been inputted incorrectly. While only a handful of incidents were filtered out, this ultimately must cause a small difference in my analysis.
2) Crime in the US As a US citizen who follows current events, I am aware that this country has a difficult history with gun violence. Throughout this project, I was sure to not let any preconceived knowledge affect my analysis.

Conclusion

This brief analysis demonstrates similar trends between shooting incidents and shooting deaths from 2006 to 2021. More specifically, both shooting incidents and deaths have decreased significantly since 2006. However, there has been an uptick from 2019 onward. Furthermore, this analysis also suggests a positive, linear relationship between shooting incidents and deaths. In other words, as the number of shooting incidents increases, so does the number of deaths.

Upon returning to this project, one could analyze trends across months or days. Furthermore, one could attempt to analyze the demographics of both the perpetrators and of the victims. Lastly, I believe it would be interesting to view the number of incidents across each NYC neighborhood. I welcome whoever views my project to piggyback off of my work and make further discoveries!