NYPD Gun Violence Data

JJEA

2023-07-01

```
#We need the following libraries

knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
```

Evolution of Gun Violence in NYC

Using the public data from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD, we will try to show the development of violent events that involve the use of guns in New York City from 2006 to 2022.

```
#Import the data from the URL:

url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd <- read_csv(url)
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

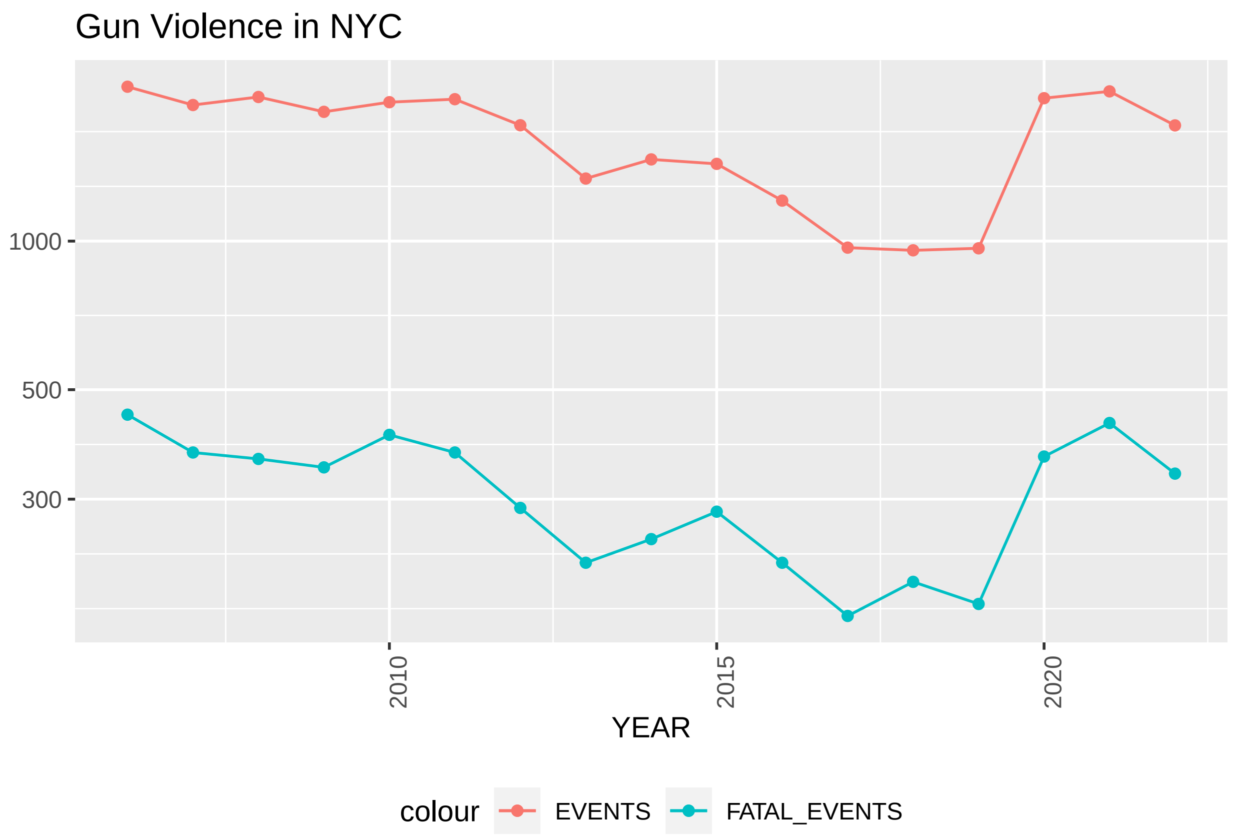
After importing the data we need to clean it up:

```
#Transform the date of the event to a type "DATE":
nypd_tidy <- mutate(nypd, OCCUR_DATE = as.Date(OCCUR_DATE, format= "%m/%d/%Y"))
#Delete columns that we are not going to use:
nypd_tidy <- nypd_tidy %>%
  select(-c(INCIDENT_KEY,JURISDICTION_CODE,LOCATION_DESC,LOC_CLASSFCTN_DESC,X_COORD_CD,
          Y_COORD_CD, Latitude, Longitude, Lon_Lat))
#Change a couple of column names for better understanding an easier coding:
nypd_tidy <- nypd_tidy %>% rename("LOCATION" = "LOC_OF_OCCUR_DESC",
                                  "MURDER" = "STATISTICAL_MURDER_FLAG")
#Create 3 new columns for the research:
nypd_tidy <- nypd_tidy %>% mutate(SHOOTINGS = 1) %>%
  mutate(DEATHS = case_when(MURDER == TRUE ~ 1, MURDER == FALSE ~ 0))
nypd_tidy <- nypd_tidy %>% mutate(YEAR = year(OCCUR_DATE))
```

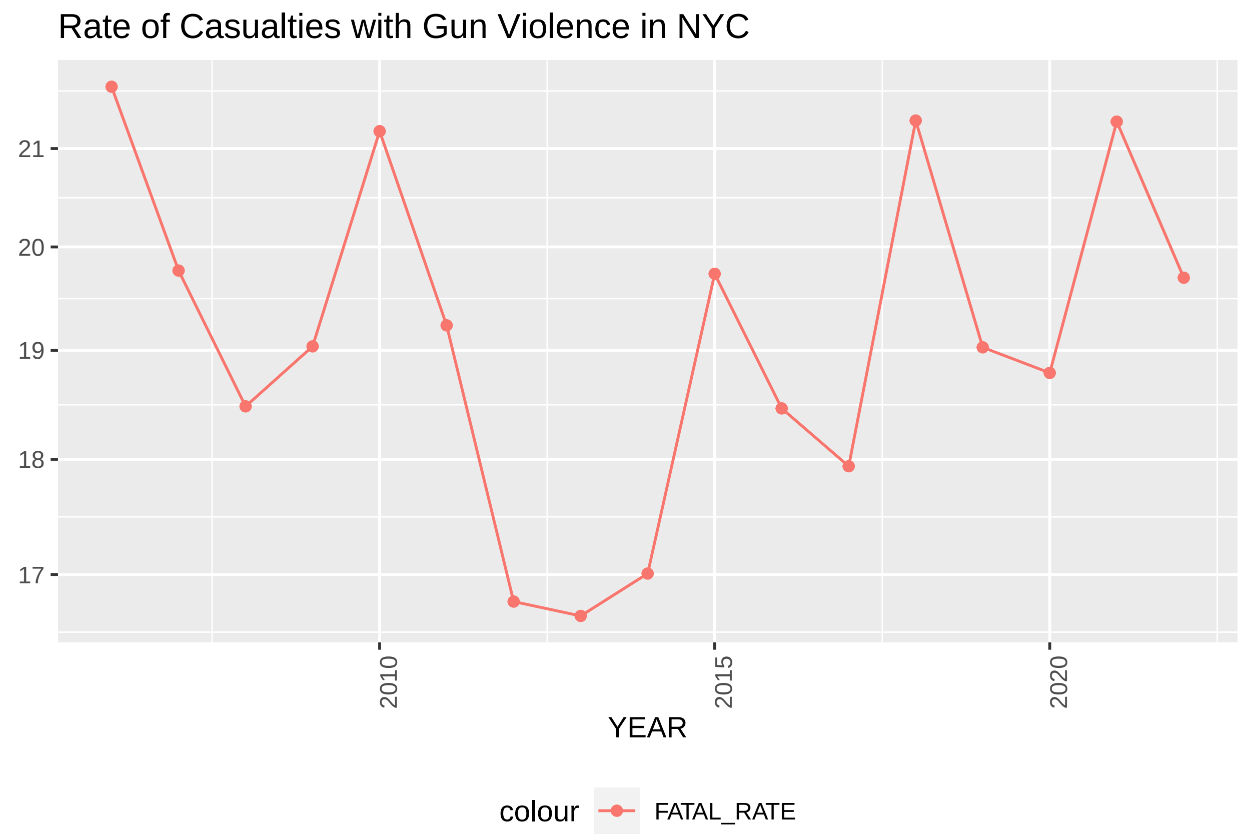
We plot the number of events and how many resulted on human casualties, we can see a downtrend from 2006 to 2019, apparently during the pandemic the was a surge of gun violence in New York City, one could argue that the isolation caused people to react more violently.

As we can see the rate of casualties does not follows the same trend as the number of events, there seems to be year where there are a lot more fatalities, like 2010, 2018 and 2021.

```
#Create a query by year of accident, showing events and number of deaths, also
#create a new column with the rate of fatal events:
nypd_byyear <- nypd_tidy %>% group_by(YEAR) %>%
  summarise(EVENTS = sum(SHOOTINGS),FATAL_EVENTS = sum(DEATHS)) %>%
  mutate(FATAL_RATE = FATAL_EVENTS/EVENTS*100)
nypd_byyear %>% ggplot(aes( x = YEAR, y = EVENTS)) +
  geom_line(aes(color = "EVENTS")) +
  geom_point(aes(color = "EVENTS")) +
  geom_line(aes(y = FATAL_EVENTS, color = "FATAL_EVENTS")) +
  geom_point(aes(y = FATAL_EVENTS, color = "FATAL_EVENTS")) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Gun Violence in NYC", y = NULL)
```



```
#We plotted the new rate of fatal events by year:
nypd_byyear %>% ggplot(aes( x = YEAR, y = FATAL_RATE)) +
  geom_line(aes(color = "FATAL_RATE")) +
  geom_point(aes(color = "FATAL_RATE")) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Rate of Casualties with Gun Violence in NYC", y = NULL)
```



## NYC is divided in 5 boroughs, so its interesting to see which borough has the most events involving gun violence. They show a similar pattern but there are difference worth exploring further.

```
#Query showing how many events/deaths by borough
nypd_boro <- nypd_tidy %>% group_by(BORO) %>%
  summarise(EVENTS = sum(SHOOTINGS),FATAL_EVENTS = sum(DEATHS)) %>%
  mutate(FATAL_RATE = FATAL_EVENTS/EVENTS*100)
nypd_boro
```

```
## # A tibble: 5 x 4
##   BORO          EVENTS FATAL_EVENTS FATAL_RATE
##   <chr>          <dbl>        <dbl>      <dbl>
## 1 BRONX           7937         1542       19.4
## 2 BROOKLYN       10933         2122       19.4
## 3 MANHATTAN       3572          630       17.6
## 4 QUEENS          4094          810       19.8
## 5 STATEN ISLAND    776          162       20.9
```

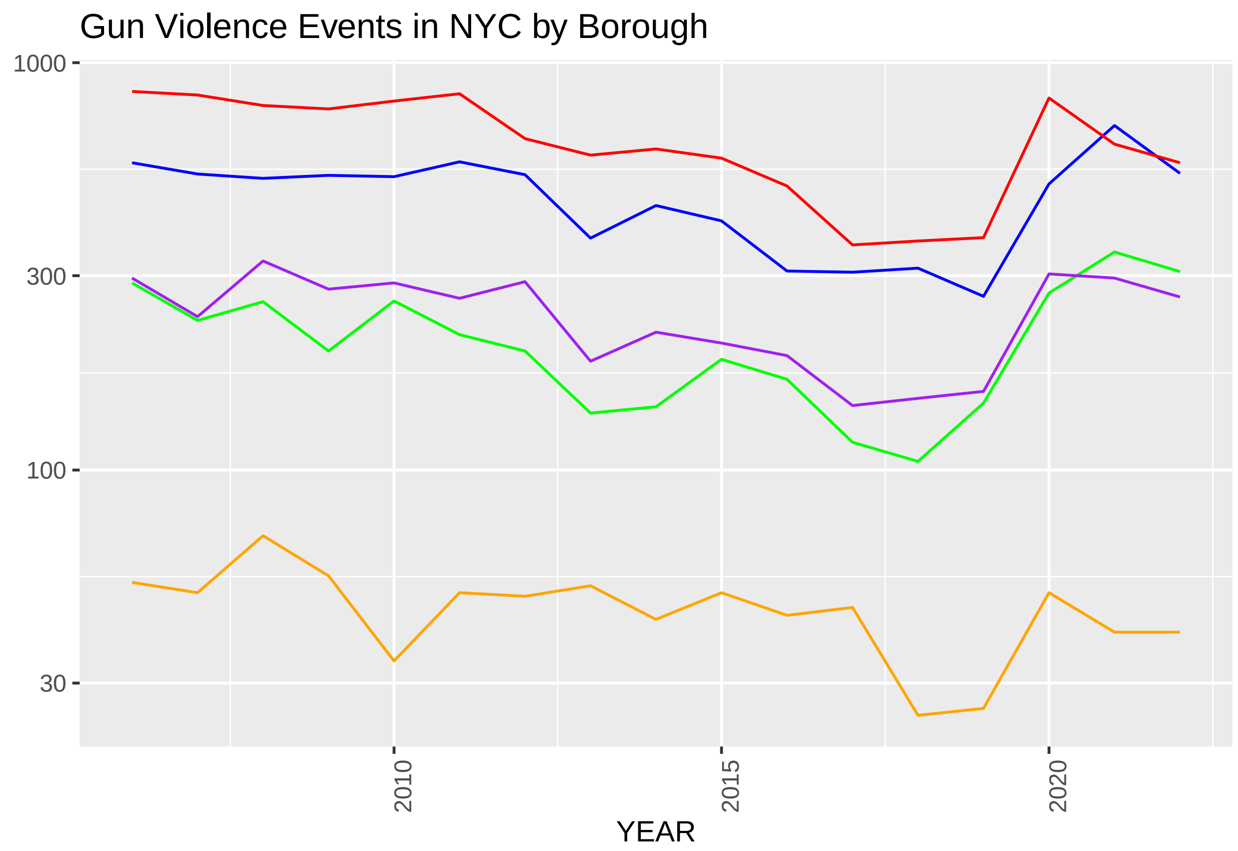
```
#Query showing the trends of events by borough, use pivot_wider for a better looking table
nypd_boro_year <- nypd_tidy %>% group_by(BORO,YEAR) %>%
  summarise(EVENTS = sum(SHOOTINGS)) %>%
  pivot_wider(names_from = BORO, values_from = EVENTS)
```

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

```
#Need to change a columns name for the plot
colnames(nypd_boro_year)[6] = "STATEN_ISLAND"
nypd_boro_year
```

```
## # A tibble: 17 x 6
##     YEAR BRONX BROOKLYN MANHATTAN QUEENS STATEN_ISLAND
##    <dbl> <dbl>    <dbl>     <dbl>  <dbl>         <dbl>
##  1  2006   568      850       288    296            53
##  2  2007   533      833       233    238            50
##  3  2008   520      785       259    326            69
##  4  2009   529      770       196    278            55
##  5  2010   525      805       260    288            34
##  6  2011   571      839       215    264            50
##  7  2012   531      651       196    290            49
##  8  2013   371      593       138    185            52
##  9  2014   446      614       143    218            43
## 10  2015   409      583       187    205            50
## 11  2016   308      498       167    191            44
## 12  2017   306      357       117    144            46
## 13  2018   313      365       105    150            25
## 14  2019   267      372       146    156            26
## 15  2020   504      819       272    303            50
## 16  2021   701      631       343    296            40
## 17  2022   535      568       307    266            40
```

```
#Plot the information by borough
nypd_boro_year %>% ggplot() +
  geom_line(aes(x = YEAR, y = BRONX), color = "blue") +
  geom_line(aes(x = YEAR, y = BROOKLYN), color = "red") +
  geom_line(aes(x = YEAR, y = MANHATTAN), color = "green") +
  geom_line(aes(x = YEAR, y = QUEENS), color = "purple") +
  geom_line(aes(x = YEAR, y = STATEN_ISLAND), color = "orange") +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Gun Violence Events in NYC by Borough", y = NULL)
```



## Since there's a linear relationship between number of events and casualties we can create a linear model to project number or deaths by gun violence.

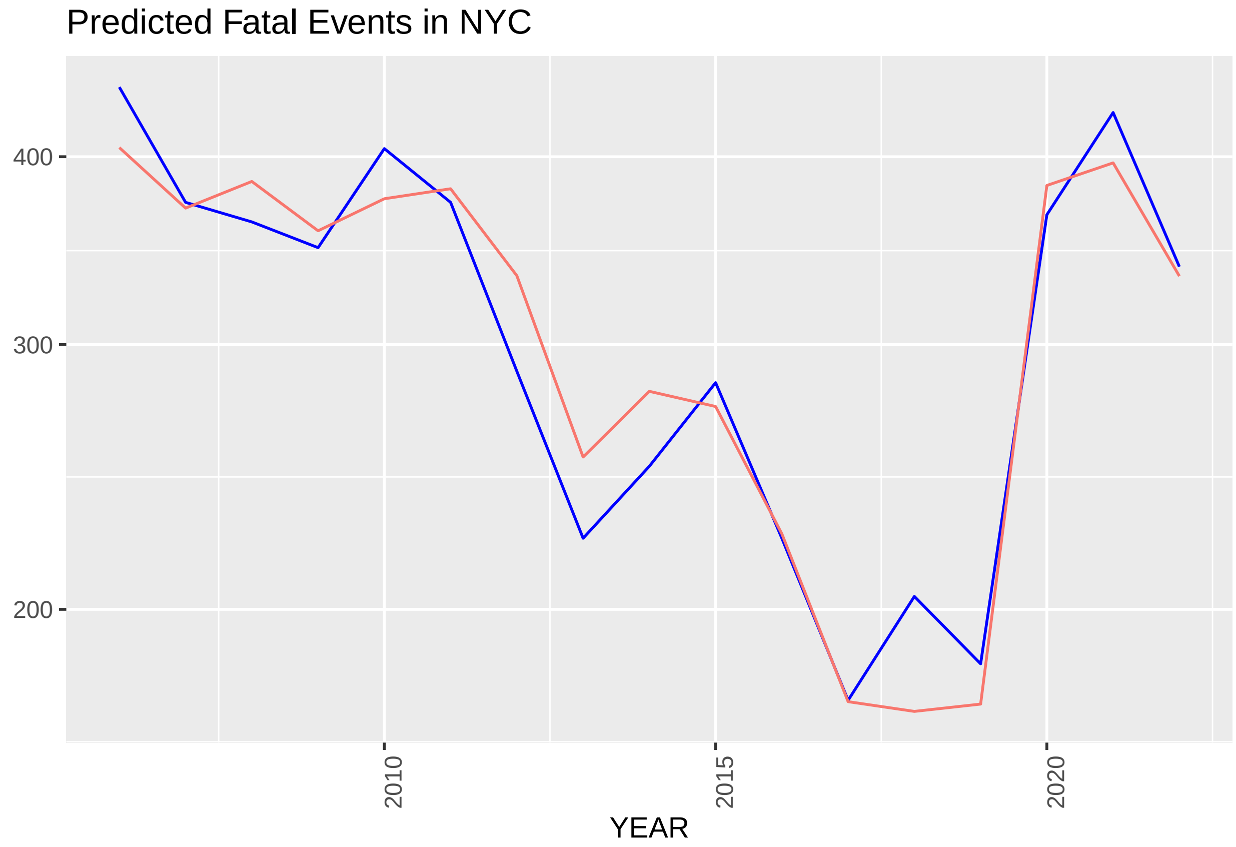
On the summary of the model we get an R^2 of 92.18 so it a fairly good prediction.

On the graph you can see how the actual number of deaths compares to our model.

```
#Create linear model, get the summary
model <- lm(FATAL_EVENTS ~ EVENTS, data = nypd_byyear)
summary(model)
```

```
##
## Call:
## lm(formula = FATAL_EVENTS ~ EVENTS, data = nypd_byyear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.376 -16.774   0.367  11.009  39.344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.79833   26.55391  -1.273    0.222
## EVENTS        0.21385    0.01608  13.298 1.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.31 on 15 degrees of freedom
## Multiple R-squared:  0.9218, Adjusted R-squared:  0.9166
## F-statistic: 176.8 on 1 and 15 DF,  p-value: 1.049e-09
```

```r
#Plot the actual number of casualties vs the predicted
nypd_model <- nypd_byyear %>% mutate(PREDICTED_FATAL_EVENTS = predict(model))
nypd_model %>% ggplot() +
  geom_line(aes(x = YEAR, y = FATAL_EVENTS), color = "blue") +
  geom_line(aes(x = YEAR, y = PREDICTED_FATAL_EVENTS, color = "red"), show.legend = FALSE) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Predicted Fatal Events in NYC", y = NULL)
```



Bias on the database and analysis:

I think there is bias in the database because it includes race of the perp and the victims; one of the sad parts of looking at the database is that it shows that african americans are more likely to be involved in gun violence.

As for the analysis, I am biased because I don't like guns and I was hoping to show that gun violence has become more scarce, which was true up until the pandemic, but the also sad part is that gun violence is still common in NYC.