NYPD Shooting Incident Data Report

2023-02-14

```
library(tidyverse)
library(lubridate)
```

Getting the data

We first get the raw data from the following link: "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

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

Tidying and Transforming the Data

Several columns can be taken into account to draw any type of analysis. For my analysis, I will look at the shooting trend on a yearly basis while keeping the different NYC boroughs into consideration. Because we are only focusing on year of event and borough, we can remove most of the unnecessary columns for this analysis.

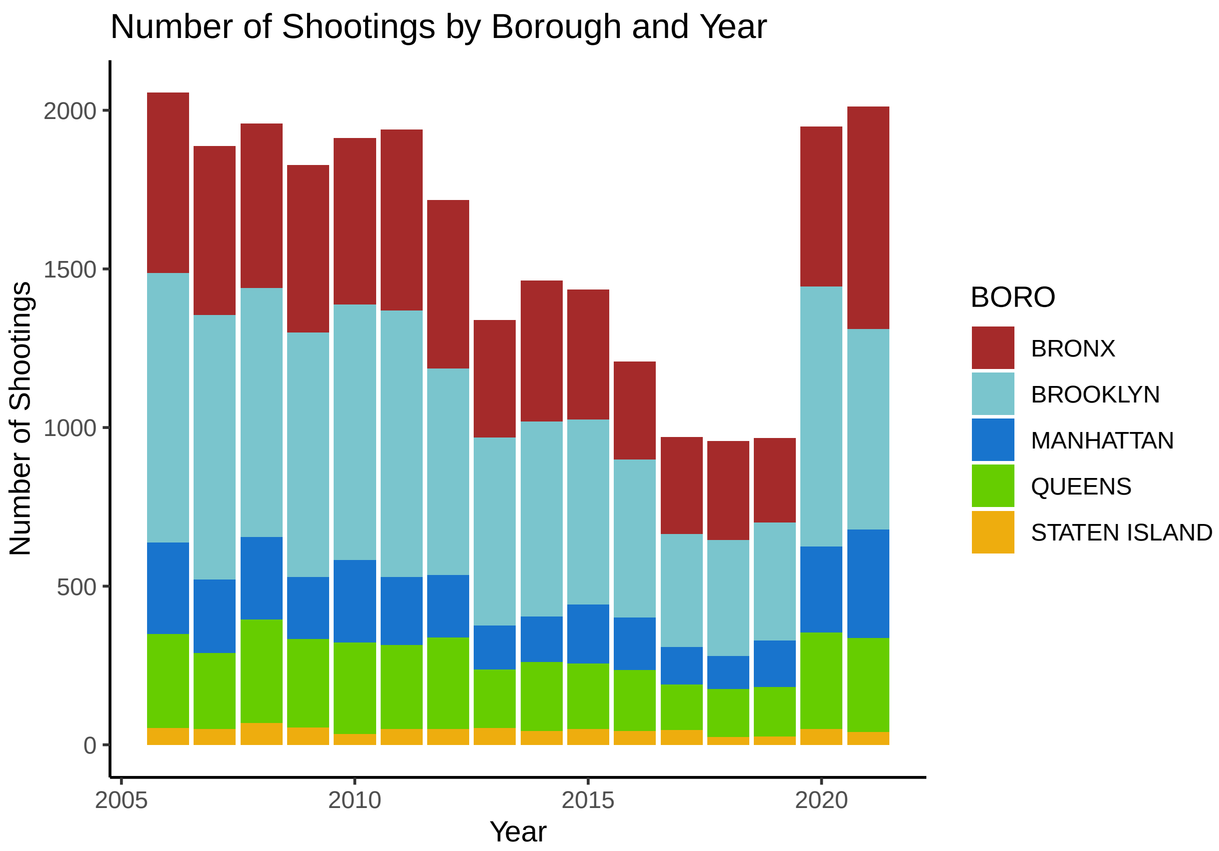
```
nypd_tr<- nypd_raw %>%
  mutate(Year = as.integer(substr(OCCUR_DATE,nchar(OCCUR_DATE)-3,nchar(OCCUR_DATE)))) %>%
  select(Year, BORO)

summarized_byBORO <- nypd_tr %>%
    group_by(Year,BORO) %>%
    summarize(shootings = n()) %>%
    pivot_wider(names_from = BORO,values_from = shootings)%>%
    ungroup() %>%
    mutate(TOTAL = rowSums(across(-Year)))
```

Visualizations and Analysis

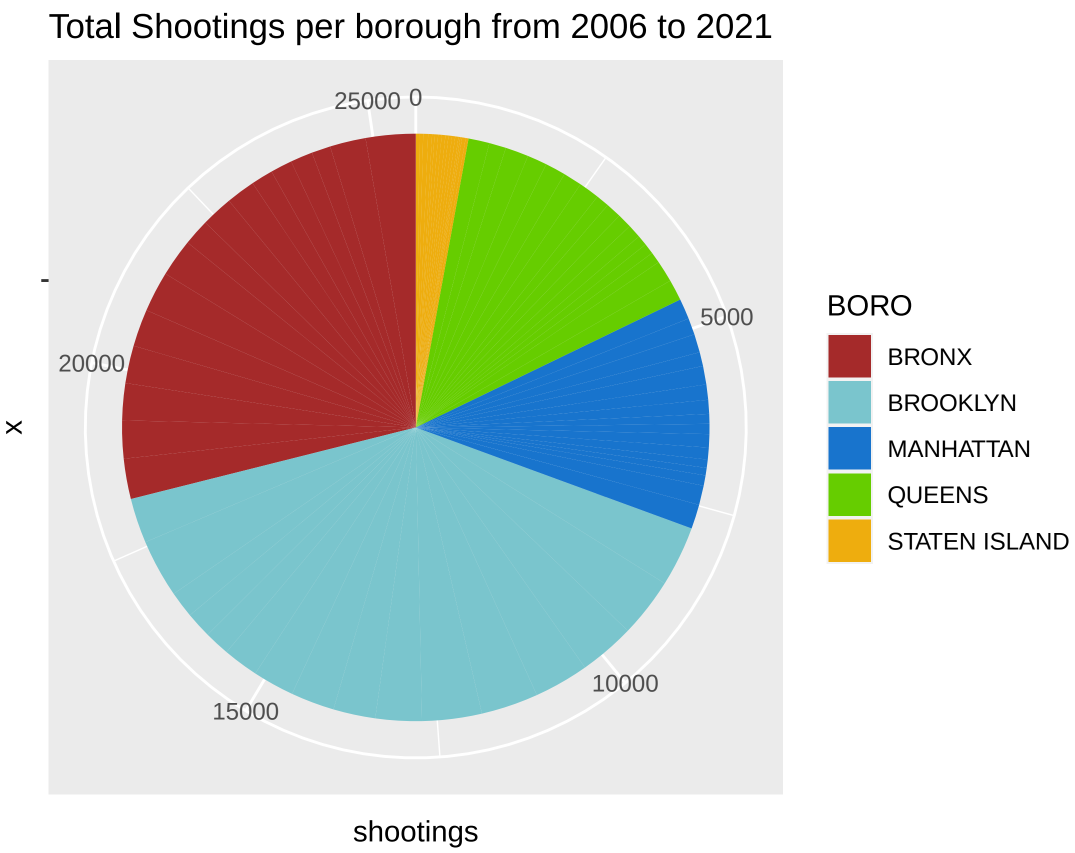
```
grouped_nypd_tr <- nypd_tr %>%
  group_by(Year,BORO) %>%
  summarize(shootings = n())

nypd_wide <- nypd_tr %>%
    group_by(Year,BORO) %>%
    summarize(shootings = n()) %>%
    pivot_wider(names_from = BORO,values_from = shootings)

nypd_wide
```

```
## # A tibble: 16 x 6
## # Groups:   Year [16]
##     Year BRONX BROOKLYN MANHATTAN QUEENS ‘STATEN ISLAND‘
##    <int> <int>    <int>     <int>  <int>           <int>
##  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
```

```
ggplot(grouped_nypd_tr, aes(x=Year, y=shootings, fill=BORO)) +
    geom_bar(stat="identity") +
    labs(title="Number of Shootings by Borough and Year") +
    xlab("Year") + ylab("Number of Shootings") +
    scale_fill_manual(values=c("brown", "cadetblue3", "dodgerblue3", "chartreuse3", "darkgoldenrod2"))
    theme_classic()
```



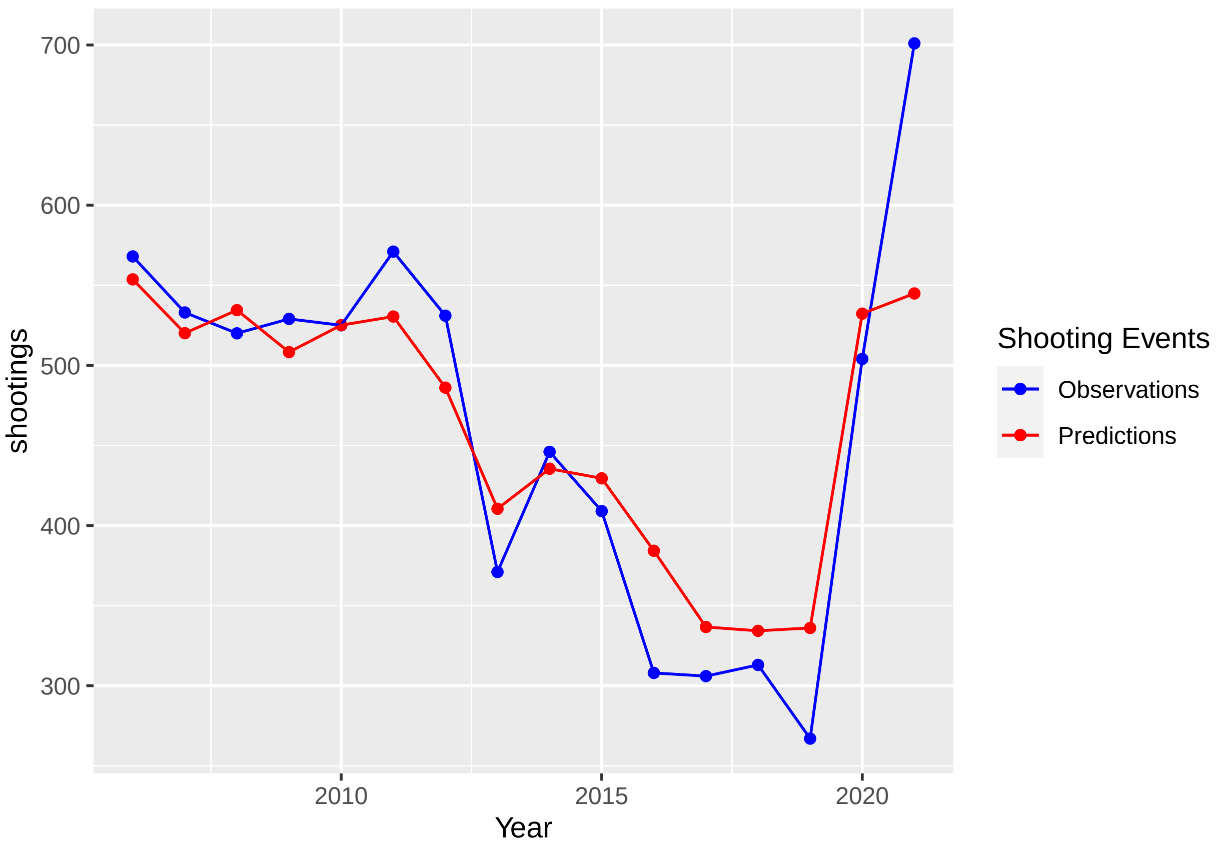
```
grouped_nypd_tr %>% ggplot(aes(x="", y=shootings, fill=BORO)) +
    geom_bar(stat="identity", width=1) +
    coord_polar("y", start=0) +
    labs(title="Total Shootings per borough from 2006 to 2021") +
    scale_fill_manual(values=c("brown", "cadetblue3", "dodgerblue3", "chartreuse3", "darkgoldenrod2"))
```



We can see clearly that some boroughs have more shooting incidents than others. The proportion of shooting incidents per borough over the years seems to be constant as well. We can also notice a steady decrease from 2006 to 2019 and a sharp increase in year 2020. The COVID pandemic and the lockdowns it caused might have played a role in this sudden increase of violence in 2020. If this is the case, we might experience a decrease from 2022 onward as it stabilizes around the shooting rates prior 2020. Since the proportions seem to stable, we will try to build a model that predicts the amount of shooting events per borough on a given year based on the year total.

Modelling Data

```
merged_df <- left_join(grouped_nypd_tr, summarized_byBORO, by="Year") %>%
              select(Year,BORO,shootings,TOTAL)
mod <- lm(shootings ~ BORO + TOTAL, data = merged_df)
predicted_nyc <- merged_df %>% ungroup() %>% mutate(pred = predict(mod))

bronx_filter = predicted_nyc %>% filter(BORO=="BRONX")

bronx_filter %>%
    ggplot() +
    geom_point(aes(x=`Year`, y = shootings, color ="Observations")) +
    geom_point(aes(x=`Year`, y = pred, color = "Predictions")) +
    geom_line(aes(x=`Year`, y = shootings, color ="Observations")) +
    geom_line(aes(x=`Year`, y = pred, color = "Predictions")) +
    scale_color_manual(values=c("blue", "red"), name = "Shooting Events")
```



Our initial intuition was very close to the observed data. We based our model on total events and we assumed a constant proportion of events per borough. We can notice that the prediction line is following the observation line closely up until the last recorded data of 2021.

Bias Identification

By looking at the different graphs and at the data, we could argue that the boroughs of Brooklyn and Bronx are the most dangerous and that Staten Island is the safest. However, we are not taking population into account in our study. Bronx and Brooklyn are the most populous boroughs in New York City. It would make sense that, everything else being equal, these regions have the highest number of incidents. A less biased and more informative study would've been to cross reference the populations of these boroughs and calculate a per capita rate instead. Moreover, the sudden increase in 2020 could lead to several misinterpretations. This period was greatly affected by the COVID pandemic and several social and societal issues were caused by it. These issues might have contributed to the rise of violence during that period of time.