DTSA_5301: NYPD Shooting Data

9/10/2021

Introduction

Between the COVID-19 pandemic, high unemployment rates, contested national election results, and civil unrest ignited by the murder of George Floyd, 2020 undeniably disrupted the lives of many — perhaps all — Americans. Against this backdrop, politicians allege that urban communities in the U.S. are experiencing a sharp increase in violent crime — especially shootings. New York City, Chicago, Atlanta, Los Angeles, Portland, and other cities have all made national news for the alleged explosion in shooting incidents.

This report is a preliminary investigation into recent urban gun violence trends. Using the *NYPD Shooting Incident Data (Historic)*, we'll take a look at how 2020 compared to recent years, going back to 2006. This dataset contains all shooting events in New York City that are known to the NYPD.

Important Note: Each row in the dataset represents 1 victim. In our analysis, the data reflect the number of shooting *victims* rather than *events*

NYPD Shooting Incident Raw Data

```r
#libraries used in producing this report
library(tidyverse)
library(lubridate)
library(ggplot2)
library(data.table)
library(padr)
library(cowplot)
```

```r
#List of every shooting incident that occurred in NYC going back to 2006
#through the end of the previous calendar year (2020).
nypd_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
nypd_data <- read_csv(nypd_url, show_col_types = FALSE)
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME             BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class :character
##  Median : 83365370   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
```

```
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  Mean   : 66.21   Mean   :0.3323
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA’s   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:23568       Length:23568       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

This investigation will only make use of *INCIDENT_KEY*, *OCCUR_DATE*, and *BORO*. We can discard all other columns. Additionally, *OCCUR_DATE* should be transformed from 'character' to 'Date' objects.

```
nypd_data <- nypd_data %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    rename(DATE = OCCUR_DATE) %>%
    select(INCIDENT_KEY, DATE, BORO) %>%
    arrange(DATE)

head(nypd_data, n=10)
```

```
## # A tibble: 10 x 3
##    INCIDENT_KEY DATE       BORO
##           <dbl> <date>     <chr>
##  1      9953250 2006-01-01 QUEENS
##  2      9953248 2006-01-01 QUEENS
##  3      9953250 2006-01-01 QUEENS
##  4      9953252 2006-01-01 MANHATTAN
##  5    139716503 2006-01-01 BROOKLYN
##  6      9953246 2006-01-01 BRONX
##  7      9953247 2006-01-01 BROOKLYN
##  8      9953245 2006-01-01 BRONX
##  9      9953249 2006-01-02 BROOKLYN
## 10      9953255 2006-01-02 BROOKLYN
```

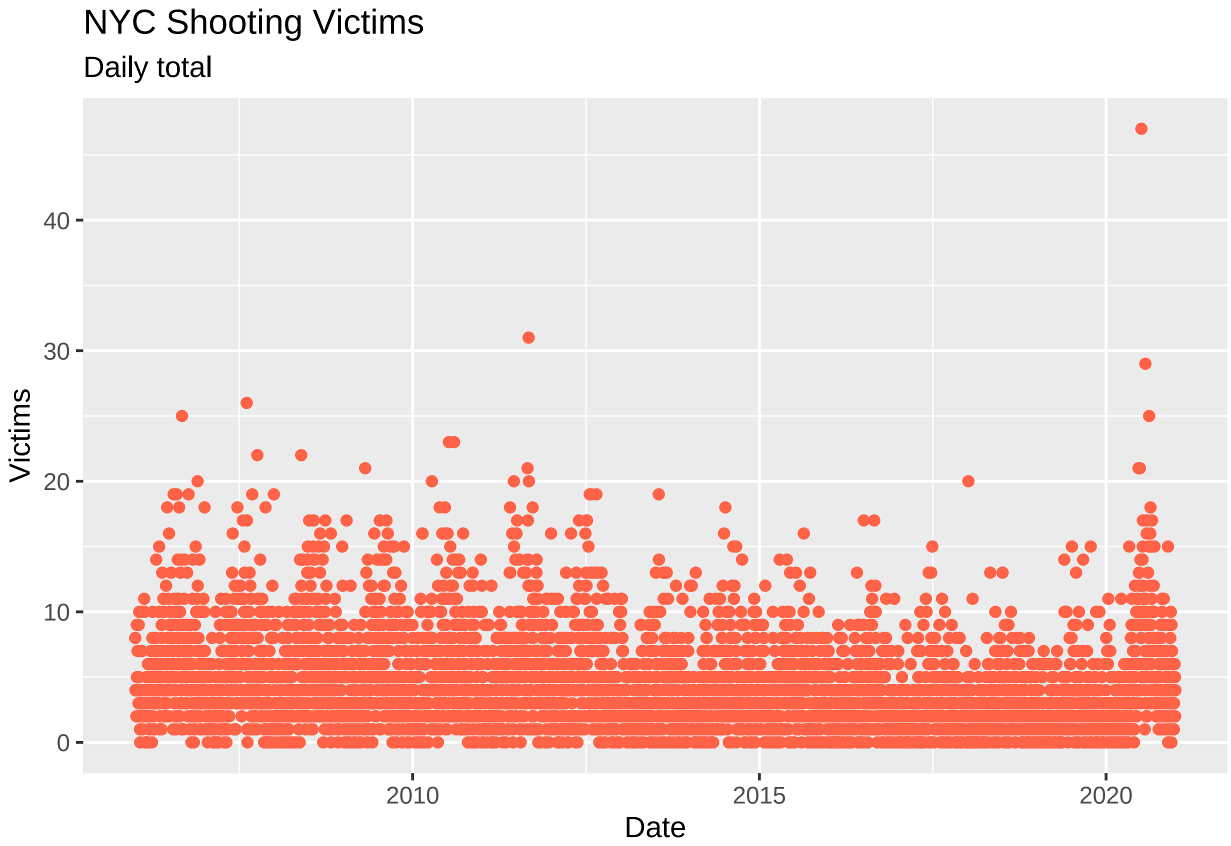
Note: The dataset is "missing" dates that had no shooting incidents.

After dropping extraneous variables, we'll pivot the data table to display total daily shooting victims by borough. Lastly, we need to add omitted dates when there were no shooting incidents.

```
daily_totals <- nypd_data %>%
    group_by(DATE,
             BORO) %>%
    arrange(DATE) %>%
    mutate(DAILY_BORO_TOTAL = length(INCIDENT_KEY)) %>% #get daily victim count
    pivot_wider(id_cols=DATE,
             names_from=BORO,
             values_from=DAILY_BORO_TOTAL,
             values_fn=mean,
             values_fill = 0) %>%
    rename(STATEN_ISLAND = `STATEN ISLAND`) %>%
    mutate(DAILY_TOTAL = sum(c_across(QUEENS:STATEN_ISLAND))) %>%
    ungroup() %>%
    pad() %>%   #automatically add omitted dates (days with 0 shootings)
    setnafill(type='const', fill=0) %>% #replace  'NA' with '0'
    mutate( WEEK = (year(DATE) - year(min(DATE)))*52
               + week(DATE) - week(min(DATE)),
           YEAR = year(DATE),
           MONTH = as.character(lubridate::month(DATE,label=TRUE)))

head(daily_totals)
```

```
## # A tibble: 6 x 10
##   DATE       QUEENS MANHATTAN BROOKLYN BRONX STATEN_ISLAND DAILY_TOTAL  WEEK
##   <date>      <dbl>     <dbl>    <dbl> <dbl>         <dbl>       <dbl> <dbl>
## 1 2006-01-01      3         1        2     2             0           8     0
## 2 2006-01-02      0         0        3     0             1           4     0
## 3 2006-01-03      2         0        2     0             0           4     0
## 4 2006-01-04      2         0        1     1             0           4     0
## 5 2006-01-05      0         0        2     2             0           4     0
## 6 2006-01-06      1         0        0     3             0           4     0
## # ... with 2 more variables: YEAR <int>, MONTH <chr>
```

```
daily_totals %>%
    ggplot(aes(x=DATE, y=DAILY_TOTAL)) +
    geom_point(color="tomato", aes(y=DAILY_TOTAL)) +
    labs(x="Date", y="Victims", title = "NYC Shooting Victims", subtitle ="Daily total" )
```

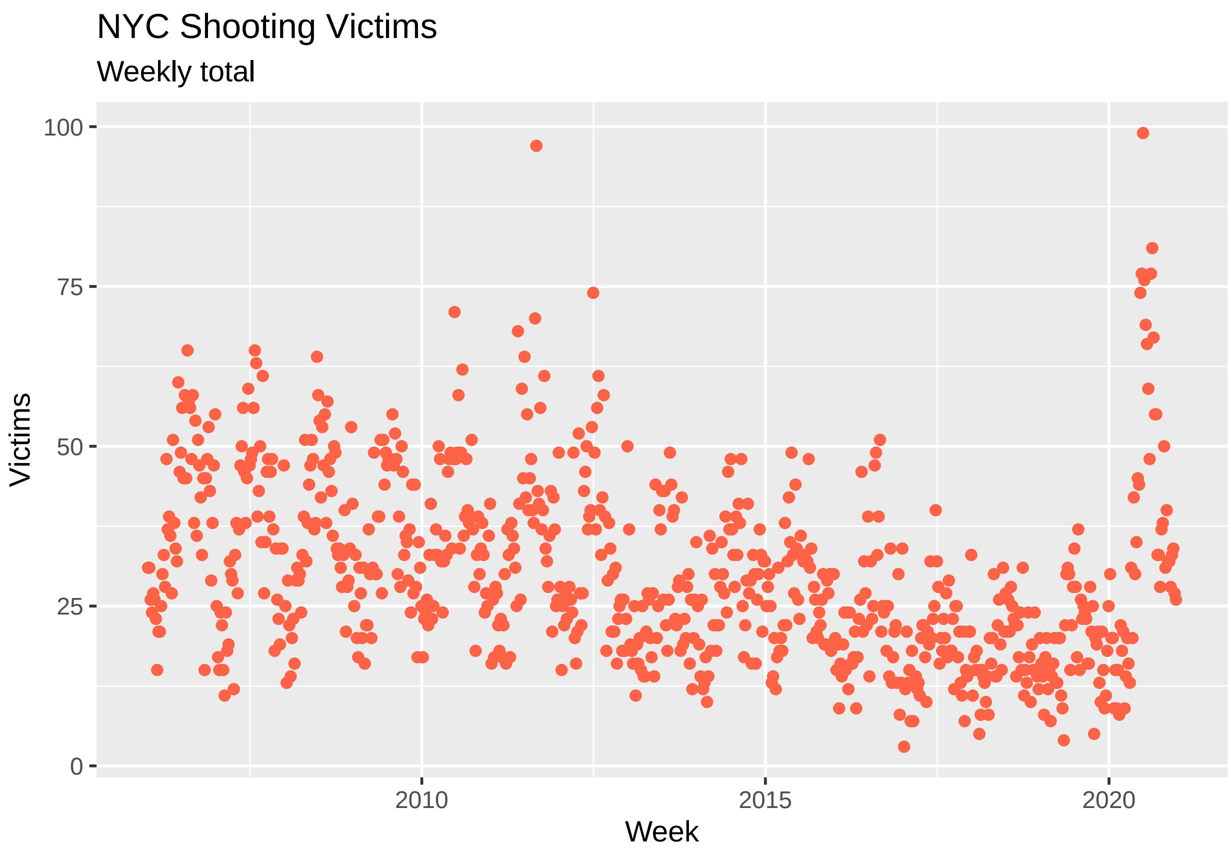


Weekly & Yearly Data

A mere glance at the daily victim data suggests that 2020 may have seen a notable spike in gun violence. However, much of the plot is completely saturated with data points so we will group our data by weekly totals.

When grouping the data by week, we must make a decision about how to handle partial weeks (first week of 2006 and last of 2020). Certain days of the week may be more likely to see gun violence, so I will drop the two incomplete weeks from the dataset. None of the days involved in the incomplete weeks had particularly high numbers of victims, so there should be minimal impact on the validity of any analysis.

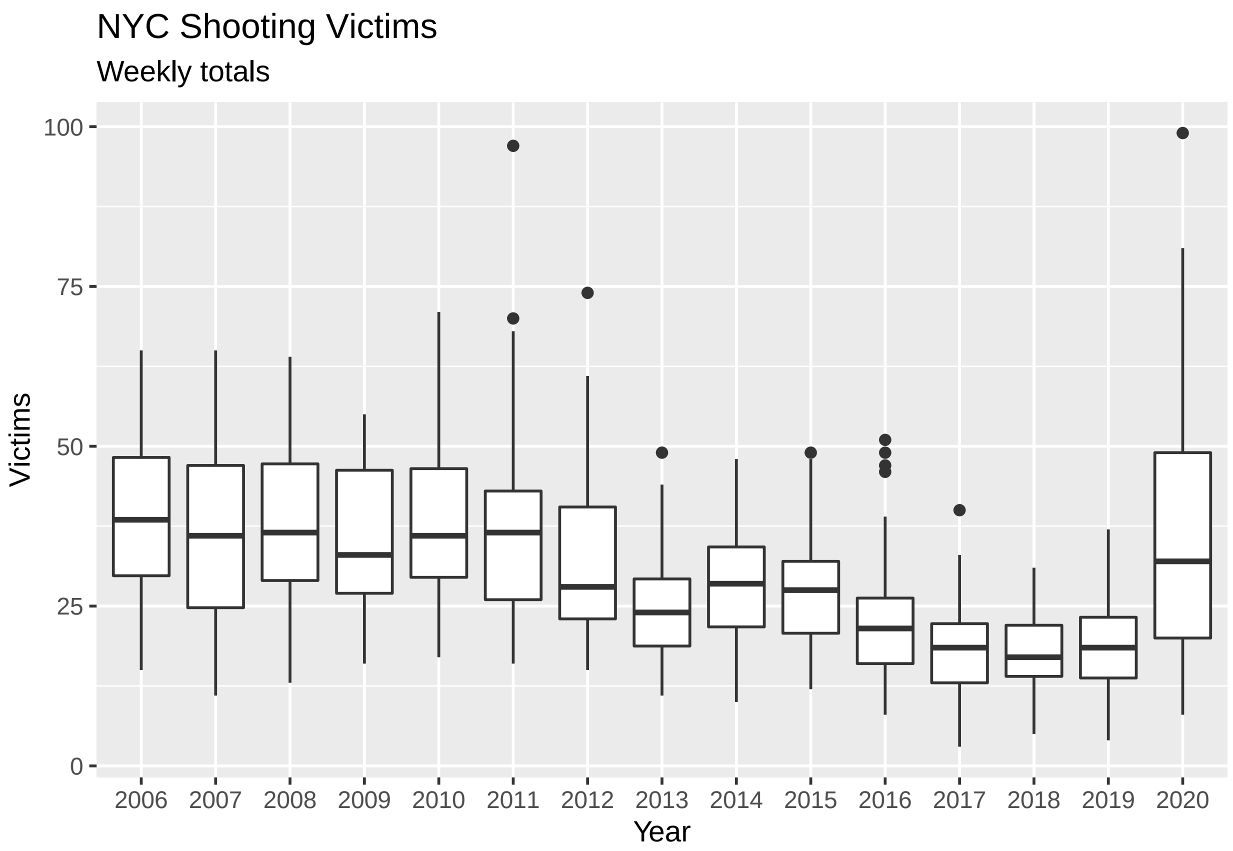
```r
weekly_totals <- daily_totals %>%
    select(WEEK,DATE, everything()) %>%
    filter(WEEK != 0 & WEEK != 780) %>%  #ignore partial weeks
    group_by(WEEK) %>%
    mutate(across(QUEENS:DAILY_TOTAL, sum)) %>%
    rename(WEEKLY_TOTAL = DAILY_TOTAL) %>%
    ungroup()

weekly_totals <- data.table(weekly_totals) %>%
    unique(by=1)

weekly_totals %>%
    ggplot(aes(x=DATE, y=WEEKLY_TOTAL)) +
    geom_point(color="tomato", aes(y=WEEKLY_TOTAL)) +
    labs(x="Week", y="Victims", title="NYC Shooting Victims", subtitle="Weekly total")
```



The scatterplot of weekly shooting victim totals shows some interesting trends.

We can see a sawtooth pattern of shooting incidents increasing during the spring, peaking in the summer, and then decreasing to minimal levels by winter. Additionally, it is apparent that there was an overall decrease in shootings from 2013-2019. That trend came to an abrupt end in 2020, which saw a sharp increase in shooting victims. 2020 contains eye popping weeks, but the beginning and end of the year saw shooting rates inline with previous years. Box plots of each year may provide clarity.

```
weekly_totals %>%
    select(YEAR, WEEKLY_TOTAL) %>%
    mutate(YEAR = as.factor(YEAR)) %>%
    ggplot(aes(x=YEAR, y=WEEKLY_TOTAL)) +
    geom_boxplot() +
    labs(x="Year", y="Victims", title="NYC Shooting Victims", subtitle="Weekly totals")
```



The box plot clearly shows that 2020 deviates from the preceding years 2016-2019. Interestingly, 2020 appears much more similar to the trend seen from 2006-2011, albeit statistical analysis is needed to justify any such claim. Still, it should be uncontroversial to conclude that 2020 saw an alarming rise in gun violence when compared to 2016-2019.

Analysis & Modeling

2020's rise in gun violence appears to coincide with COVID-19 beginning its spread throughout NYC.

COVID-19 doesn't seem to be a sufficient explanation; after all, NYC had higher levels of violence in 2006-2011 without a corresponding pandemic. Perhaps the ensuing spike in unemployment can explain 2020's increase in violence. Mass unemployment could result in desperation to the point of resorting to violence to meet one's needs.

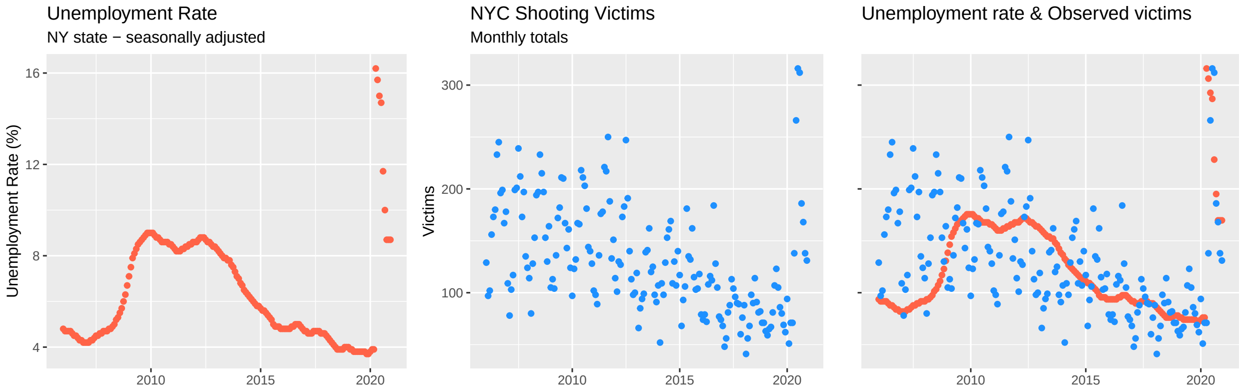
To investigate the possible relationship between unemployment and gun violence, I'll use the U.S. Bureau of Labor Statistics' dataset for New York state. The dataset contains monthly, seasonally corrected unemployment rates. Before creating a model, we'll view the unemployment rates overlaid onto the monthly shooting scatterplot.

```r
#Data from U.S. Bureau of Labor Statistics website.  CSV of data hosted on github
#Includes monthly employment data for New York state from 2006-2020
labor_url <- "https://raw.githubusercontent.com/r-quillen/DTSA-5301/main/BLS_NY_employment.csv"
labor_data <- read_csv(labor_url, show_col_types = FALSE)

labor_data <- labor_data %>%
    rename(YEAR = Year,
           MONTH = Period,
           UNEMP_RATE = `unemployment rate`) %>%
    select(YEAR, MONTH, UNEMP_RATE) %>%
    mutate(MONTH = as.character(MONTH),
           YEAR = as.integer(YEAR))

head(labor_data)
```

```
## # A tibble: 6 x 3
##    YEAR MONTH UNEMP_RATE
##   <int> <chr>      <dbl>
## 1  2006 Jan          4.8
## 2  2006 Feb          4.7
## 3  2006 Mar          4.7
## 4  2006 Apr          4.7
## 5  2006 May          4.7
## 6  2006 Jun          4.6
```

```r
monthly_totals <- daily_totals %>%
    select(DATE, YEAR, MONTH, DAILY_TOTAL) %>%
    group_by(YEAR, MONTH) %>%
    mutate(DAILY_TOTAL = sum(DAILY_TOTAL)) %>%
    rename(MONTHLY_TOTAL = DAILY_TOTAL)

monthly_totals <- data.table(monthly_totals) %>%
    unique(by=c(2,3)) %>%
    full_join(labor_data)

pl1 <- monthly_totals %>%
    ggplot(aes(DATE)) +
    geom_point(color="tomato", aes(y=UNEMP_RATE)) +
    labs(x="", y="Unemployment Rate (%)",
         title="Unemployment Rate", subtitle="NY state - seasonally adjusted")
```

```
pl2 <- monthly_totals %>%
    ggplot(aes(DATE)) +
    geom_point(color="dodgerblue", aes(y=MONTHLY_TOTAL)) +
    labs(x="", y="Victims",
         title="NYC Shooting Victims", subtitle="Monthly totals", )

pl3 <- monthly_totals %>%
    ggplot(aes(DATE)) +
    geom_point(color="tomato", aes(y=(max(MONTHLY_TOTAL)*UNEMP_RATE/max(UNEMP_RATE)))) +
    geom_point(color="dodgerblue", aes(y=MONTHLY_TOTAL)) +
    labs(x="", y="", title="Unemployment rate & Observed victims", subtitle="") +
    theme(axis.text.y = element_blank())

plot_grid(pl1, pl2, pl3, nrow=1)
```



The unemployment data appear to follow the same overall trend as the shooting data, particularly from 2010-2020.

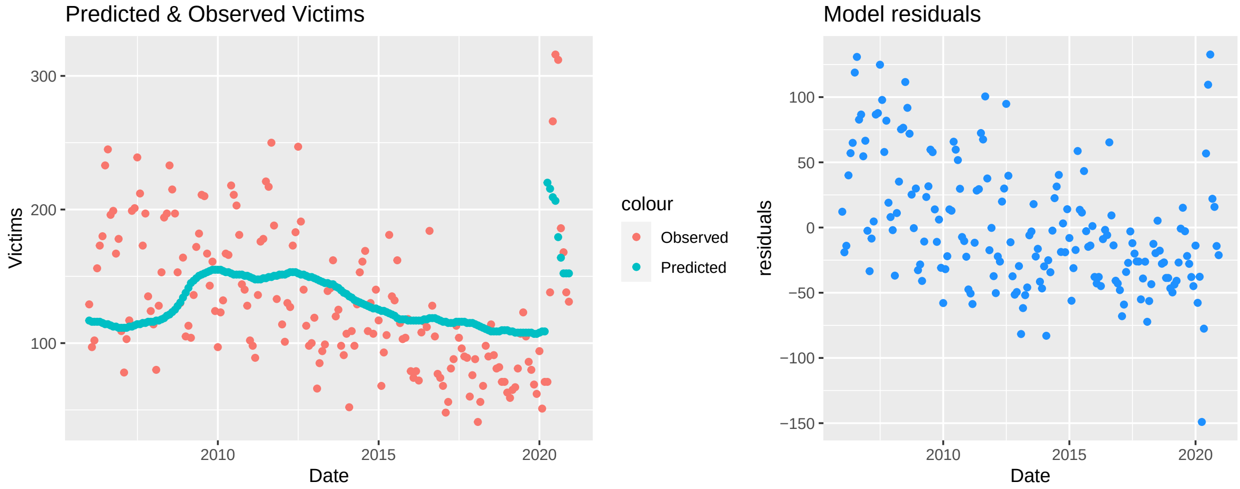
As a preliminary step in investigating the link between unemployment rates and shooting victims, we'll see how well a linear regression model fits our data.

```
model <- lm(MONTHLY_TOTAL ~ UNEMP_RATE,monthly_totals)

monthly_totals <- monthly_totals %>%
    mutate(PRED = predict(model),
           RESID = resid(model))

p_mod1 <- monthly_totals %>%
    ggplot(aes(x=DATE)) +
    geom_point(aes(color="Observed", y=MONTHLY_TOTAL)) +
    geom_point(aes(color = "Predicted", y=PRED)) +
    labs(x="Date", y="Victims", title="Predicted & Observed Victims")

p_mod2 <- monthly_totals %>%
    ggplot(aes(x=DATE)) +
    geom_point(color="dodgerblue", aes(y=RESID)) +
    labs(x="Date", y="residuals", title="Model residuals")

summary(model)
```

```
##
## Call:
```

```
## lm(formula = MONTHLY_TOTAL ~ UNEMP_RATE, data = monthly_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149.07  -34.93  -11.14   28.61  132.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.443     10.486   7.004 4.92e-11 ***
## UNEMP_RATE     9.051      1.549   5.841 2.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.54 on 178 degrees of freedom
## Multiple R-squared:  0.1609, Adjusted R-squared:  0.1561
## F-statistic: 34.12 on 1 and 178 DF,  p-value: 2.412e-08
```

```
plot_grid(p_mod1, p_mod2, rel_widths=c(3,2))
```



Despite the similarity in overall trend between the unemployment rate and shooting victims, our linear model is not very good. The coefficient of determination = .1561, which is much lower than we would have liked. It may be worth additional time/effort to try more advanced modeling techniques.

Conclusion

In this report, we saw ample evidence that during 2020, NYC experienced a spike in gun violence in comparison to recent years. Looking back to 2006-2011, we saw higher levels of gun violence which seem more similar to that of 2020.

We then investigated a hypothesis that unemployment is a catalyst for gun violence. We were unable to convincingly show such a link with a linear regression model despite similar overall trends in unemployment and gun violence from 2006-2020. Still, the hypothesis may warrant further investigation with more sophisticated modeling techniques and larger datasets covering more cities.

Sources of Bias

Personal:

Readers should be aware that I believe that modern, macro trends in urban crime are best explained by economic circumstances. My suspicion that there is a causal link between unemployment rates and shooting violence – despite not producing a model that demonstrates a strong correlation – may be a manifestation of this bias.

Bias also played a role in my decision to not investigate any link between shooting incidents and the civil unrest ignited by the murder of George Floyd. I recognize my bias in the belief that the demonstrators were not associated with gun violence, despite any other crimes that may be attributable to the groups. I avoided the topic altogether in my analysis to prevent my bias from impacting this report.

Data:

The unemployment data used in this report was for the state of New York and was seasonally adjusted. Unemployment rates in NYC may differ substantially from that of the rest of the state. Seasonal adjustment of unemployment data is standard, but may be inappropriate for this analysis.