NYPD Shooting Incident Analysis

Anonymous

2025-04-20

Objectives

I will use the following data sets published by the City of New York to analyze shooting incident data. My objective is to answer the questions below.

1. Which borough had the highest shooting mortality rate?
2. Which borough experienced the most gun violence relative to population?
3. Can we predict 2023 shootings using an ARIMA model?

Data Overview

First, I will import the necessary libraries and read in the NYC shooting incident and population data from the below sources.

```
library("tidyverse")
library("dplyr")
library("ggrepel")
library("lubridate")
library("forecast")
library("tseries")
```

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

url2 <- "https://data.cityofnewyork.us/api/views/xywu-7bv9/rows.csv?accessType=DOWNLOAD"
pop_data <- read_csv(url2)

summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:29744       Length:29744      Length:29744
##  1st Qu.: 67321140   Class :character   Class1:hms        Class :character
##  Median :109291972   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :133850951                      Mode  :numeric
##  3rd Qu.:214741917
##  Max.   :299462478
##
##  LOC_OF_OCCUR_DESC      PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:29744       Min.   :  1.00   Min.   :0.0000    Length:29744
```

```
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.00   Median :0.0000    Mode  :character
##                     Mean   : 65.23   Mean   :0.3181
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:29744       Mode :logical           Length:29744
##  Class :character   FALSE:23979             Class :character
##  Mode  :character   TRUE :5765              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:29744       Length:29744       Length:29744       Length:29744
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:29744       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000094   1st Qu.:183042   1st Qu.:40.67
##  Mode  :character   Median :1007826   Median :195506   Median :40.70
##                     Mean   :1009442   Mean   :208722   Mean   :40.74
##                     3rd Qu.:1016739   3rd Qu.:239980   3rd Qu.:40.83
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :97
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:29744
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :97
```

```r
head(pop_data)
```

```
## # A tibble: 6 x 22
##   `Age Group`      Borough        `1950` `1950 - Boro share of NYC total` `1960`
##   <chr>            <chr>           <dbl>                            <dbl>  <dbl>
## 1 Total Population NYC Total     7891957                           100    7.78e6
## 2 Total Population Bronx         1451277                            18.4  1.42e6
## 3 Total Population Brooklyn      2738175                            34.7  2.63e6
## 4 Total Population Manhattan     1960101                            24.8  1.70e6
## 5 Total Population Queens        1550849                            19.6  1.81e6
## 6 Total Population Staten Island  191555                             2.43 2.22e5
## # i 17 more variables: `1960 - Boro share of NYC total` <dbl>, `1970` <dbl>,
## #   `1970 - Boro share of NYC total` <dbl>, `1980` <dbl>,
## #   `1980 - Boro share of NYC total` <dbl>, `1990` <dbl>,
## #   `1990 - Boro share of NYC total` <dbl>, `2000` <dbl>,
```

```
## #   ‘2000 - Boro share of NYC total‘ <dbl>, ‘2010‘ <dbl>,
## #   ‘2010 - Boro share of NYC total‘ <dbl>, ‘2020‘ <dbl>,
## #   ‘2020 - Boro share of NYC total‘ <dbl>, ‘2030‘ <dbl>, ...
```

The shooing incident data set contains columns for the various attributes of a shooting case, such as an incident key, date, location information, and information on the perpetrator, if available. The population data set consists of one row per borough, a grand total, and columns for each year and the percent share of the population for that year. For my analysis, I will only be using population data for the year 2020.

Tidy and Transform Data

After taking an initial look at the data, it is evident that some tidying and transformation needs to be done. To start, I will tidy up the shooting data by removing columns that are unnecessary for this analysis and creating new date and year columns. For the population data, I will remove all unnecessary columns, and estimate the populations for 2023 using a growth rate of 0.3158% per year. The shooting and population data will be merged into one final data frame.

```r
#remove unwanted columns
nypd_data <- nypd_data %>%
  select(-c(OCCUR_TIME, LOC_OF_OCCUR_DESC, JURISDICTION_CODE,LOC_CLASSFCTN_DESC, X_COORD_CD, Y_COORD_CD

  #create and format new date columns
  mutate(DATE = mdy(OCCUR_DATE)) %>%
  mutate(YEAR = year(DATE)) %>%
  #remove old date column
  select(-c(OCCUR_DATE))

#Display final data frame
head(nypd_data)
```

```
## # A tibble: 6 x 13
##   INCIDENT_KEY BORO     PRECINCT LOCATION_DESC            STATISTICAL_MURDER_F~1
##          <dbl> <chr>       <dbl> <chr>                    <lgl>
## 1    231974218 BRONX          40 <NA>                     FALSE
## 2    177934247 BROOKLYN       79 <NA>                     TRUE
## 3    255028563 BRONX          47 GROCERY/BODEGA           FALSE
## 4     25384540 BROOKLYN       66 PVT HOUSE                TRUE
## 5     72616285 BRONX          46 MULTI DWELL - APT BUILD  TRUE
## 6     85875439 BRONX          42 MULTI DWELL - PUBLIC HO~ FALSE
## # i abbreviated name: 1: STATISTICAL_MURDER_FLAG
## # i 8 more variables: PERP_AGE_GROUP <chr>, PERP_SEX <chr>, PERP_RACE <chr>,
## #   VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>, DATE <date>, YEAR <dbl>
```

```r
#remove unwanted columns
pop_data <- pop_data %>%
  select(c(Borough, `2020`))

#Rename the columns
colnames(pop_data) <- c('BORO', '2020')

#Remove the grand total row
boro_pop <- pop_data[-1, ]
```

```r
#Change column name to upper case
boro_pop$BORO <- toupper(boro_pop$BORO)

#Assumed annual population growth rate used to find 2021-2023 populations
annual_growth_rate = 1.003158

#calculate populations for years 2021-2023
boro_pop$`2021` <- round(boro_pop$`2020` * annual_growth_rate)
boro_pop$`2022` <- round(boro_pop$`2021` * annual_growth_rate)
boro_pop$`2023` <- round(boro_pop$`2022` * annual_growth_rate)

# Perform the pivot
boro_pop <- boro_pop %>%
  pivot_longer(
    cols = starts_with("20"),
    names_to = "YEAR",
    values_to = "Population"
  ) %>%
  mutate(YEAR = as.numeric(YEAR)) # Convert YEAR from character to numeric


#Display final data frame
head(boro_pop)
```

```
## # A tibble: 6 x 3
##   BORO      YEAR Population
##   <chr>    <dbl>      <dbl>
## 1 BRONX     2020    1446788
## 2 BRONX     2021    1451357
## 3 BRONX     2022    1455940
## 4 BRONX     2023    1460538
## 5 BROOKLYN  2020    2648452
## 6 BROOKLYN  2021    2656816
```

```r
#Create new dataframe to be used for analysis
shooting_data <- nypd_data %>%
  filter(YEAR > 2019) %>%
  group_by(DATE, YEAR, BORO) %>%
  summarize(shootings = n(), deaths = sum(STATISTICAL_MURDER_FLAG == TRUE)) %>%
  inner_join(boro_pop, by = c("BORO" = "BORO", "YEAR" = "YEAR")) %>%
  select(DATE, YEAR, BORO, shootings, deaths, Population) %>%
  ungroup()

head(shooting_data)
```

```
## # A tibble: 6 x 6
##   DATE        YEAR BORO      shootings deaths Population
##   <date>     <dbl> <chr>         <int>  <int>      <dbl>
## 1 2020-01-01  2020 BRONX             2      2    1446788
## 2 2020-01-01  2020 BROOKLYN          1      0    2648452
## 3 2020-01-01  2020 MANHATTAN         1      1    1638281
## 4 2020-01-02  2020 BROOKLYN          6      0    2648452
```

```
## 5 2020-01-02  2020 MANHATTAN         2      0    1638281
## 6 2020-01-03  2020 BRONX             1      0    1446788
```

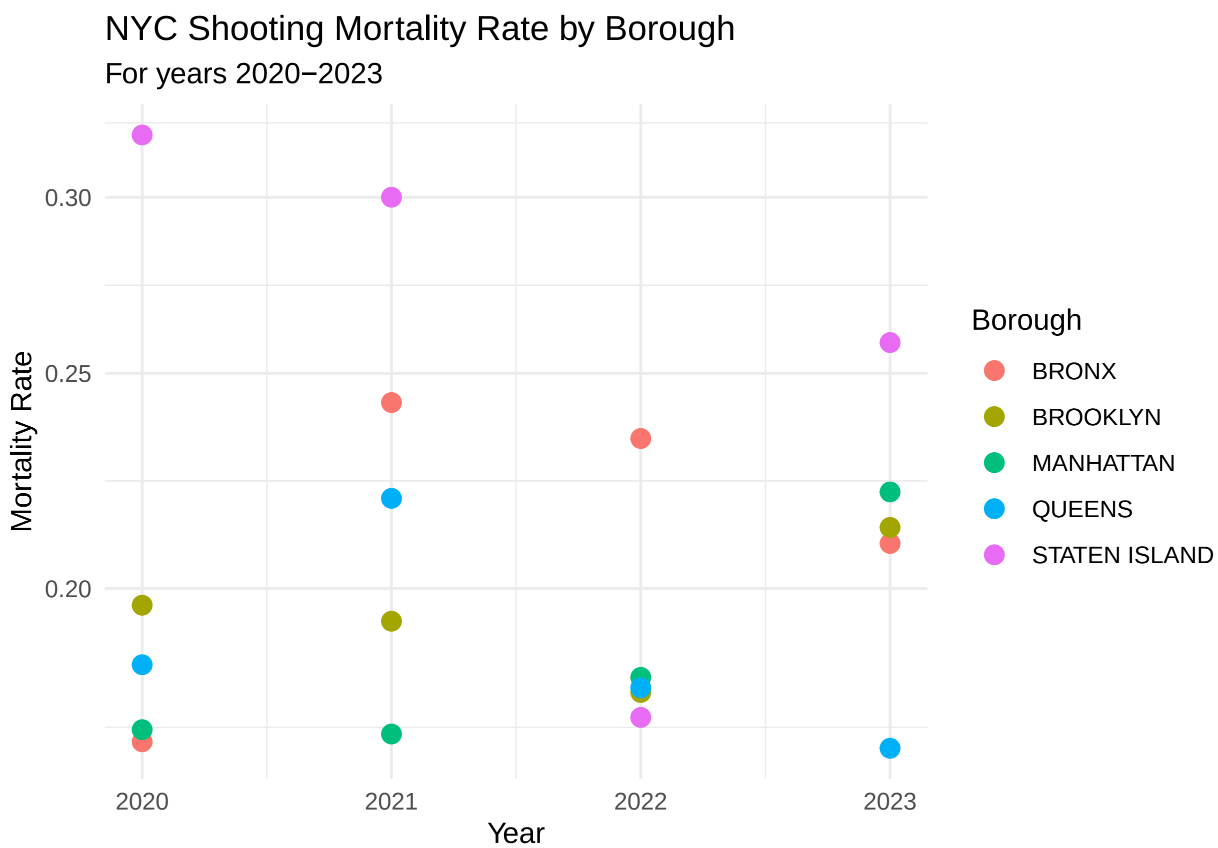
Data Analysis

Objective 1: Which borough had the highest shooting mortality rate?

For my first objective, I will summarize shootings, deaths, and population by each borough. I will also create variables for shootings per one hundred thousand, deaths per one hundred thousand, and mortality rate. It is easier to compare the impact of shootings and deaths across each borough shootings and deaths per one hundred resident to adjust for population differences.

```
shootings_by_year <- shooting_data %>%
  group_by(YEAR, BORO, Population) %>%
  summarize(shootings = sum(shootings), deaths = sum(deaths)) %>%
  mutate(shootings_per_100k = shootings * 1000000 / Population,
         deaths_per_100k = deaths * 1000000 / Population,
         mortality_rate = deaths / shootings) %>%
  select(YEAR, BORO, shootings, shootings_per_100k, deaths, deaths_per_100k, mortality_rate, Population)
  ungroup()
```

```
ggplot(shootings_by_year, aes(x = YEAR, y = mortality_rate, color = BORO)) +
  geom_point(size = 3) +
  scale_y_log10() +
  theme(legend.position="bottom",
      axis.text.x = element_text(angle = 90)) +
  labs(title = "NYC Shooting Mortality Rate by Borough",
     subtitle = "For years 2020-2023",
     x = "Year",
     y = "Mortality Rate",
     color = "Borough") +
  theme_minimal()
```



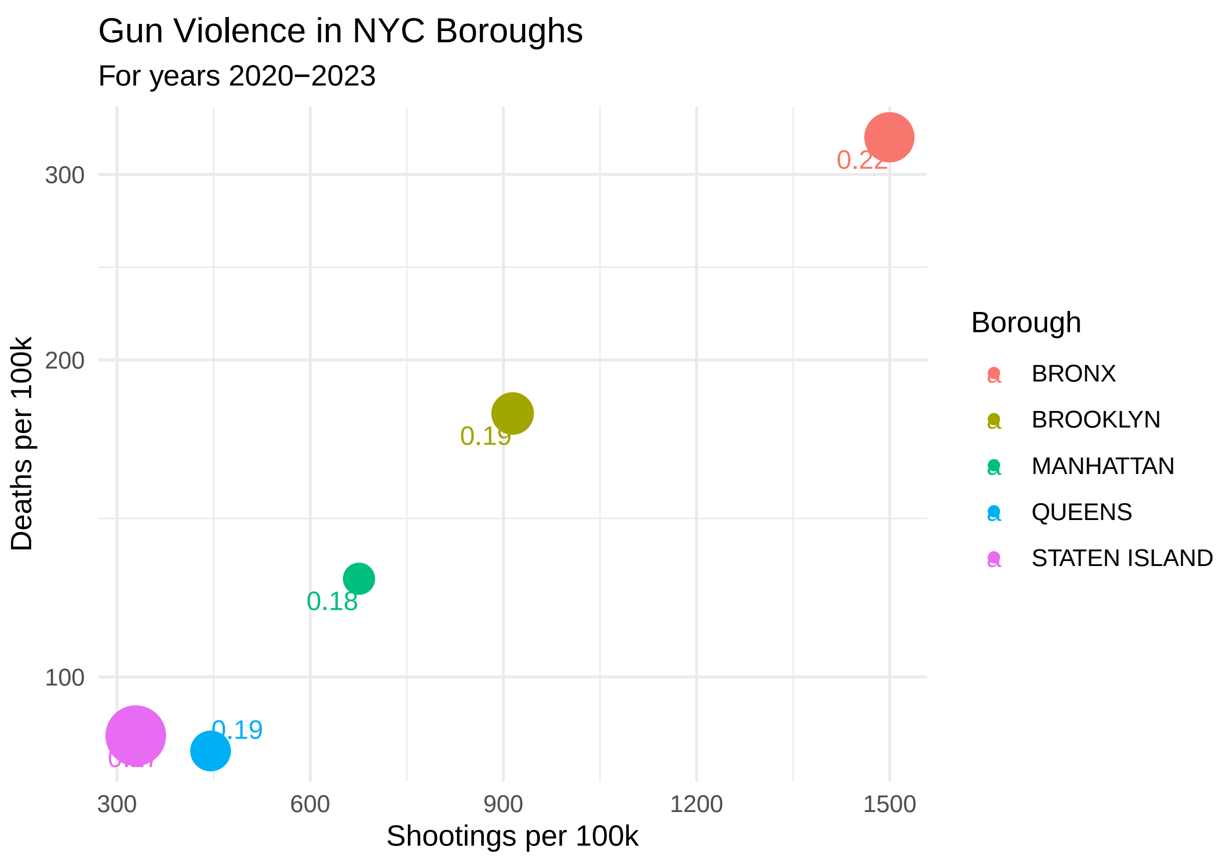
The plot shows that for the period 2020-2023, Staten Island had the highest shooting mortality rate three out of four years. Due to its small population and low number of total shooting incidents, this could be a misleading statistic without the proper context.

Objective 2: Which borough experienced the most gun violence relative to population?

```
shootings_total <- shootings_by_year %>%
  group_by(BORO) %>%
  summarize(shootings = sum(shootings), deaths = sum(deaths), shootings_per_100k = sum(shootings_per_10
  mutate(mortality_rate = deaths / shootings) %>%
  select(BORO, shootings, shootings_per_100k, deaths, deaths_per_100k, mortality_rate) %>%
  ungroup()
```

```
ggplot(shootings_total, aes(x = shootings_per_100k, y = deaths_per_100k, size = mortality_rate, color =
  geom_point() +
  geom_text_repel(aes(label = round(mortality_rate, 2)),
                  size = 3.5,
                  max.overlaps = Inf) +
  scale_y_log10() +
  scale_size_continuous(range = c(5, 10), name = "Mortality Rate") +
  labs(title = "Gun Violence in NYC Boroughs",
       subtitle = "For years 2020-2023",
       x = "Shootings per 100k",
       y = "Deaths per 100k",
```

```
      color = "Borough") +
  theme_minimal() +
  guides(size = "none")
```



My analysis has found that the Bronx was the deadliest borough for the years 2020-2023. For this period, it had the most shootings and deaths per one hundred thousand residents, while also having the second highest shooting mortality rate.

Modeling

Objective 3: Can we predict 2023 shootings using an ARIMA model?

For my third and final objective, I will use an ARIMA model to predict shooting incidents for the year 2023. The model will be trained using the data from 2020-2022, and the predicted shooting incidents will be compared to the actual shooting incidents.

```
# Filter and aggregate model
model_data <- nypd_data %>%
  filter(YEAR > 2019) %>%
  group_by(DATE) %>%
  summarize(shootings = n(), deaths = sum(STATISTICAL_MURDER_FLAG == TRUE)) %>%
  select(DATE, shootings, deaths) %>%
  filter(shootings > 0) %>%
  ungroup()
```

```r
# Aggregate data weekly
weekly_data <- model_data %>%
  mutate(week = floor_date(DATE, "week")) %>%
  group_by(week) %>%
  summarize(shootings = sum(shootings)) %>%
  filter(week < as.Date("2024-01-01"))

# Convert weekly data to time series (starting from the first week of 2020)
ts_weekly <- ts(weekly_data$shootings, start = c(2020, 1), frequency = 52)

# Split the data into training (2020-2022) and testing (2023) sets
train_data <- window(ts_weekly, end = c(2022, 52))
test_data <- window(ts_weekly, start = c(2023, 1), end = c(2023, 52))

# Fit ARIMA model to training data only
arima_model <- auto.arima(train_data)
summary(arima_model)
```

```
## Series: train_data
## ARIMA(0,1,2)(1,1,0)[52]
##
## Coefficients:
##           ma1     ma2     sar1
##       -0.7548  0.1854  -0.3412
## s.e.   0.0919  0.0921   0.1253
##
## sigma^2 = 189.2:  log likelihood = -418.15
## AIC=844.3   AICc=844.7   BIC=854.84
##
## Training set error measures:
##                      ME     RMSE      MAE       MPE     MAPE      MASE
## Training set -0.3301177 11.01318 7.148463 -4.680756 21.93377 0.5242878
##                     ACF1
## Training set -0.03324573
```

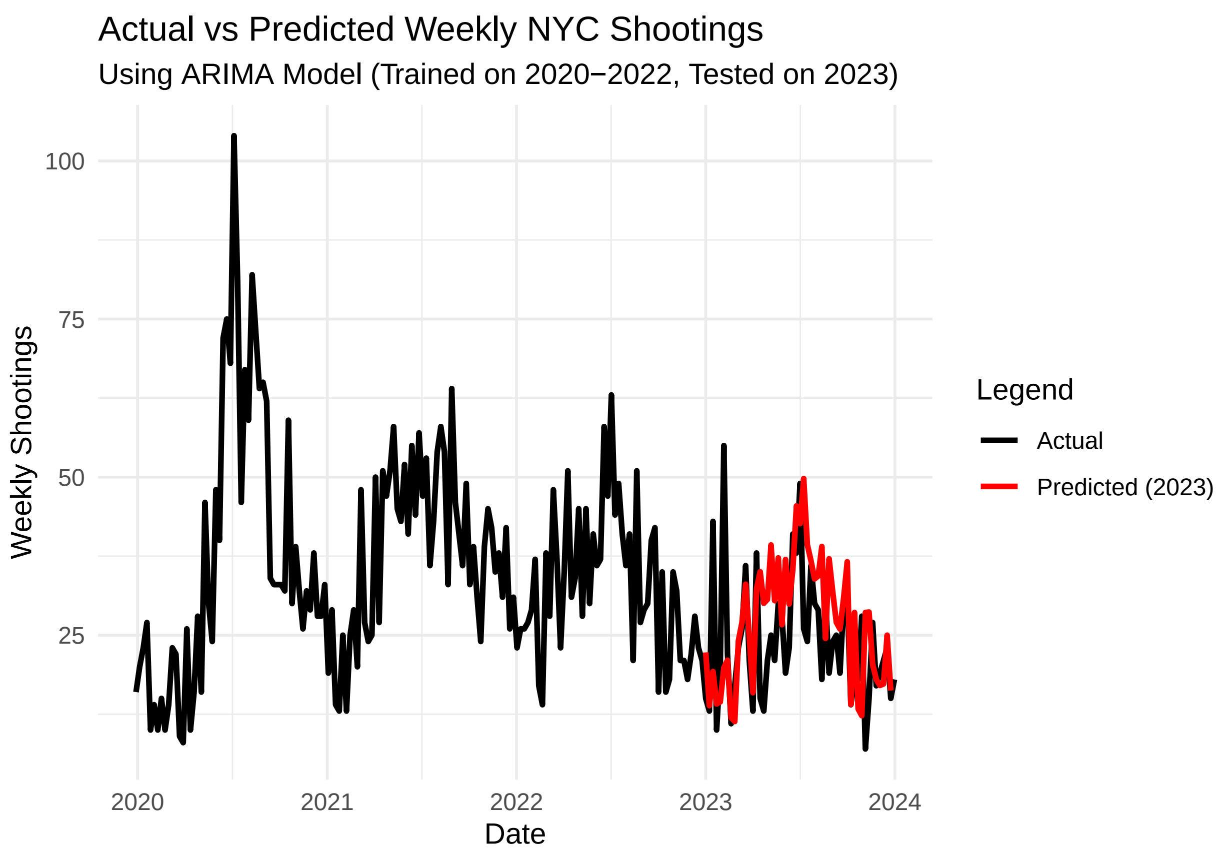
```r
# Forecast for the testing period
forecasted <- forecast(arima_model, h = length(test_data))

# Create dates for predicted results
predicted_dates <- seq(
  from = as.Date("2023-01-01"),
  by = "week",
  length.out = length(test_data)
)

# Plot predicted vs actual shootings
ggplot() +
  geom_line(data = weekly_data, aes(x = week, y = shootings, color = "Actual"), size = 1) +
  geom_line(aes(x = predicted_dates, y = as.numeric(forecasted$mean), color = "Predicted (2023)"), size
  scale_color_manual(values = c("Actual" = "black", "Predicted (2023)" = "red")) +
  labs(
    title = "Actual vs Predicted Weekly NYC Shootings",
    subtitle = "Using ARIMA Model (Trained on 2020-2022, Tested on 2023)",
```

```r
    x = "Date",
    y = "Weekly Shootings",
    color = "Legend"
  ) +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



Bias Identification

A potential bias in my analysis is that additional variables used to predict shootings incidents were not used. The original data set included different variables, such as location, time of day, and gender, however these variables were not consistently reported. Another potential variable that could influence the number of shooting incidents, would be the law enforcement population. If the number of police officers fluctuates year to year, the crime rate could be expected to reasonably fluctuate as well. For example, there were lock downs and social distancing measures introduced during 2020 and 2021 that could have reduced the law enforcement population. Incident reporting tendencies could have also changed during that time period, and are liable to change year to year. There are numerous factors that can directly contribute to a shooting incident and how/when it is reported.