NYPD Shooting Incidents RMD

G. Biehunko

9-5-2025

```
knitr::opts_chunk$set(echo = TRUE)
# included for grading and reproducibility.
```

```
library(tidyverse)
library(lubridate)
# packages used when creating this document
```

Import the Data set

I import and read in the CSV data set from data.gov to ensure reproducibility by including the URL and CSV file name.

```
# Import dataset directly from data.gov
url_in = "https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic"
file_name = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shootings <- read_csv(file_name)
```

Tidying the data

I tidy the data by transforming date and time for extraction in analysis and visualization. I exclude x and y coordinates, and longitudinal and latitudinal data.

```
# Clean data
shootings_clean <- shootings %>%
  mutate(
    OCCUR_DATE = lubridate::mdy(OCCUR_DATE),      # Convert to Date
    OCCUR_TIME = hms::as_hms(OCCUR_TIME),         # Convert to time
    BORO = as.factor(BORO),
    PERP_SEX = as.factor(PERP_SEX),
    VIC_SEX = as.factor(VIC_SEX)
  ) %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Lon_Lat, Longitude, Latitude))    # Drop unneeded cols
# I considered combining the 3 columns of location data and removing precinct and statistical murder fl
# Summary of cleaned data set
summary(shootings_clean)
```

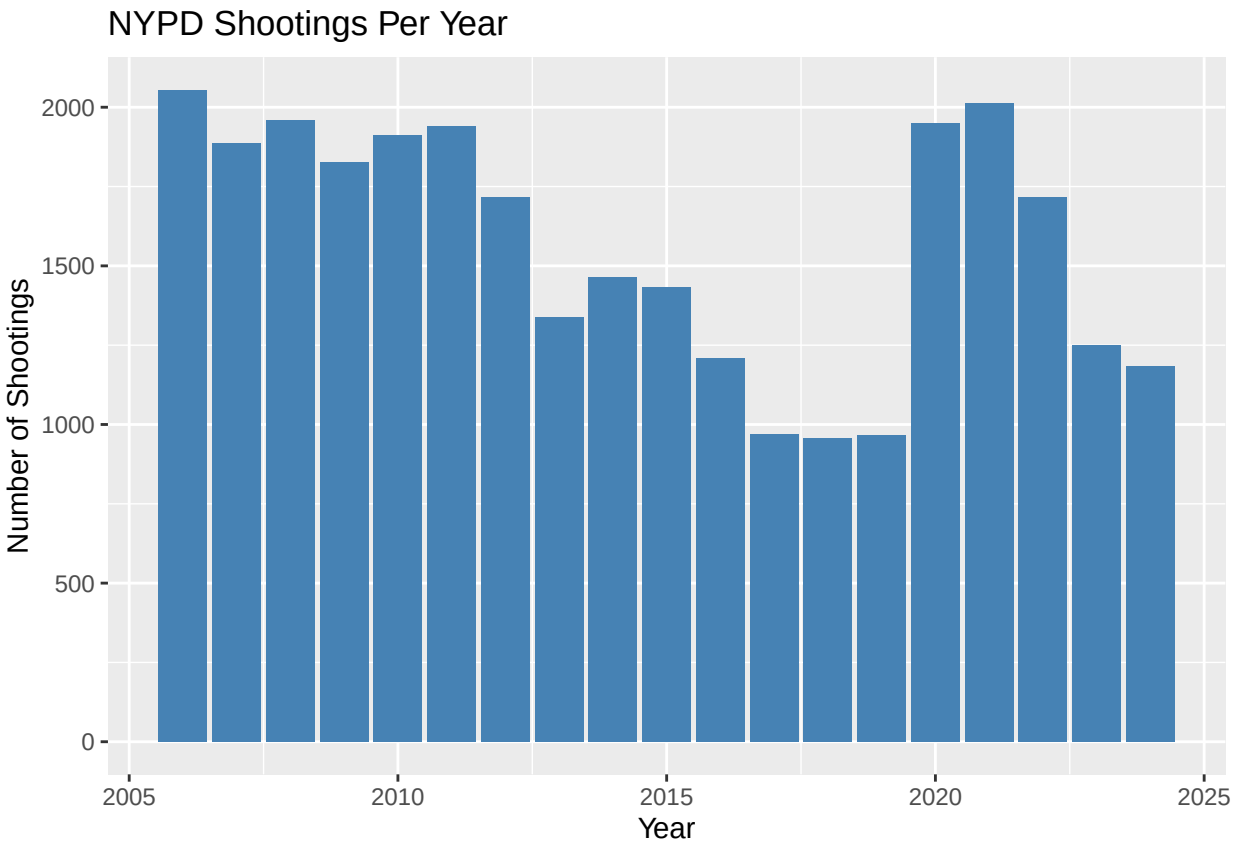
```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
```

```
##  Min.   :  9953245   Min.   :2006-01-01   Length:29744
##  1st Qu.: 67321140   1st Qu.:2009-10-29   Class1:hms
##  Median :109291972   Median :2014-03-25   Class2:difftime
##  Mean   :133850951   Mean   :2014-10-31   Mode  :numeric
##  3rd Qu.:214741917   3rd Qu.:2020-06-29
##  Max.   :299462478   Max.   :2024-12-31
##
##            BORO       LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  BRONX        : 8834   Length:29744       Min.   :  1.00   Min.   :0.0000
##  BROOKLYN     :11685   Class :character   1st Qu.: 44.00   1st Qu.:0.0000
##  MANHATTAN    : 3977   Mode  :character   Median : 67.00   Median :0.0000
##  QUEENS       : 4426                      Mean   : 65.23   Mean   :0.3181
##  STATEN ISLAND:  822                      3rd Qu.: 81.00   3rd Qu.:0.0000
##                                           Max.   :123.00   Max.   :2.0000
##                                                            NA's   :2
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:29744       Length:29744       Mode :logical
##  Class :character   Class :character   FALSE:23979
##  Mode  :character   Mode  :character   TRUE :5765
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX      PERP_RACE         VIC_AGE_GROUP
##  Length:29744       (null): 1628   Length:29744       Length:29744
##  Class :character   F     :  461   Class :character   Class :character
##  Mode  :character   M     :16845   Mode  :character   Mode  :character
##                     U     : 1500
##                     NA's  : 9310
##
##
##  VIC_SEX     VIC_RACE
##  F: 2891   Length:29744
##  M:26841   Class :character
##  U:   12   Mode  :character
##
##
##
##
```

The summary above gives me several different points of data that I an use for analysis. I'm going to keep things simple by looking at the shooting incidents over time by location instead of looking into demographics, to avoid as much bias as possible. With that, there is always bias in data analysis, but I will be focusing on the observable counts in order to predict counts in the future. Further analysis outside of this presentation could utilize the demographics to speculate on bias, and create a discussion on the qualitative side of the data.

Visualization 1: shootings per year

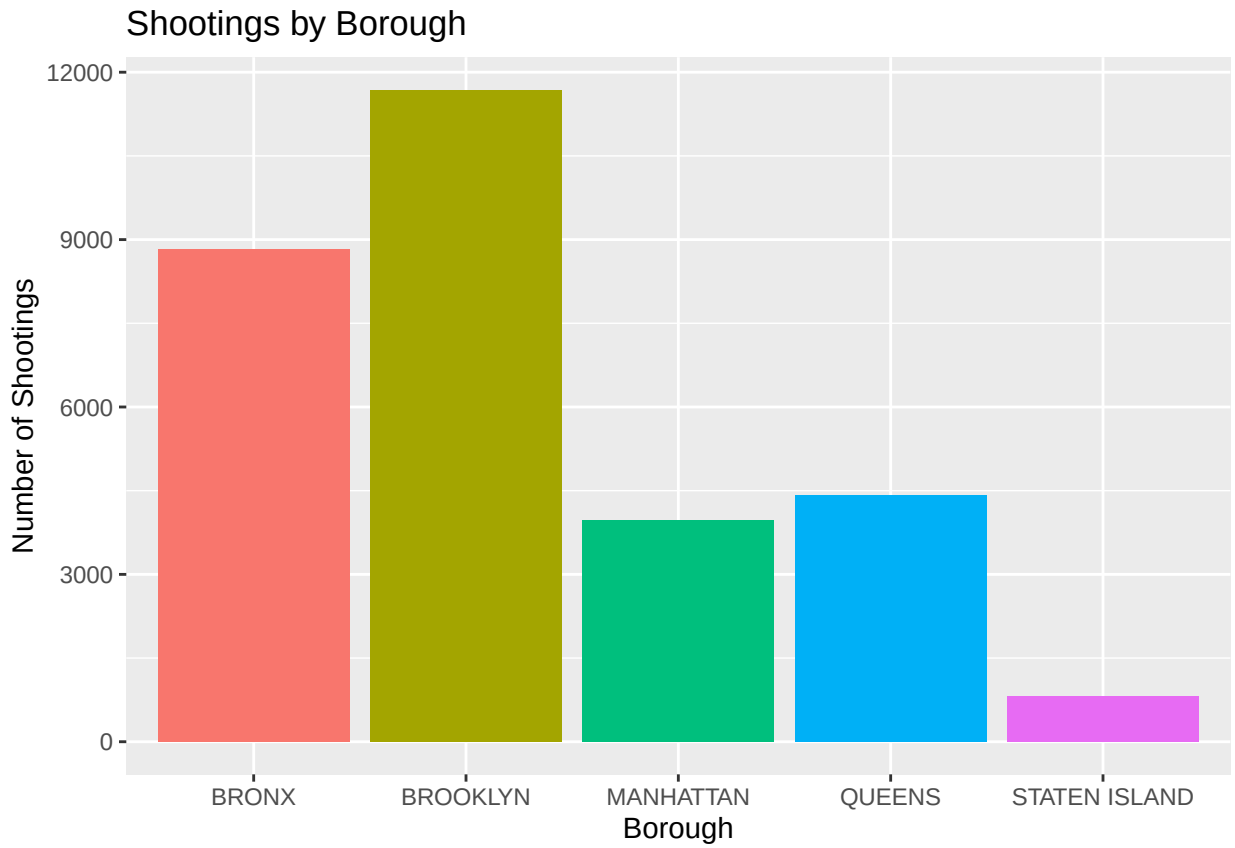
```
shootings_clean %>%
  count(year = year(OCCUR_DATE)) %>%
  ggplot(aes(x = year, y = n)) +
  geom_col(fill = "steelblue") +
  labs(
    title = "NYPD Shootings Per Year",
    x = "Year", y = "Number of Shootings"
  )
```



The bar graph above observes the overall shooting incidents per year from 2006 to 2024 in New York City. Observersers of this graph should note that shooting icidents showed a downward trend from 2012 to 2019, but shot back up in 2020. The actual contributing variables are complex, but it's rational to say that the global COVID-19 pandemic was a major contribution, and that the affirmative action in the procedures taken in response to COVID are a contributor in the decrease in shooting incidents post 2021.

Visualization 2: shootings by borough

```
shootings_clean %>%
  count(BORO) %>%
  ggplot(aes(x = BORO, y = n, fill = BORO)) +
  geom_col(show.legend = FALSE) +
  labs(
    title = "Shootings by Borough",
    x = "Borough", y = "Number of Shootings"
  )
```



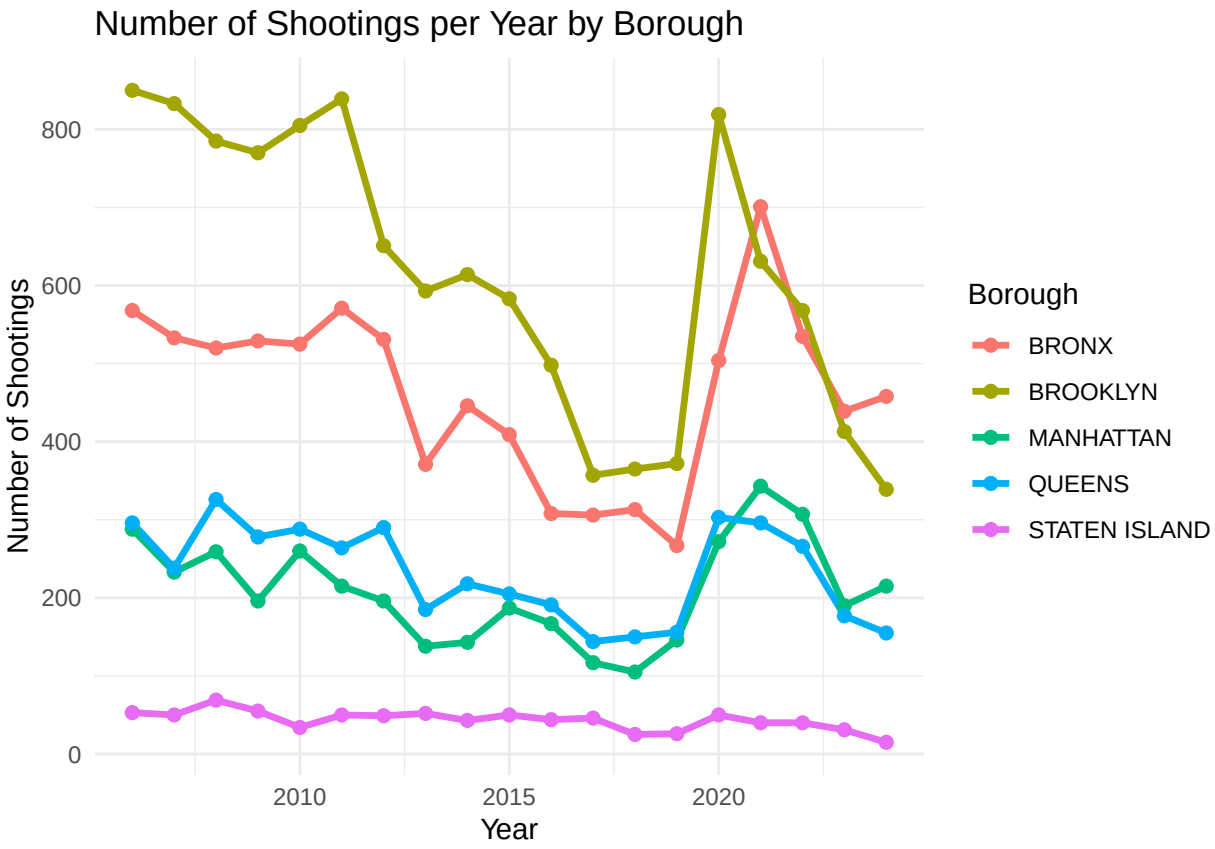
The bar graph above observes the number of shooting incidents by bourough in New York City. Observers can use the graph to speculate the difference in numbers by location. Population stats show that Staten Island has the lowest count and Brooklyn has the highest count, which makes the number of shooting incidents seem directly proportional. Another look at population stats shows something different; The shooting incidents from 2006 - 2024 in the Bronx are significantly higher than Queens even though Queens has a higher population. This proves that more variables than population influence the shooting incidents observed in each borough.

Trend observation for the Number of Shootings per Yer By Borough

Before fitting a model, I examine the raw data to see trends in shootings over time by borough.

```
shootings_year_boro <- shootings_clean %>%
  mutate(year = year(OCCUR_DATE)) %>%   # extract year
  count(year, BORO)                     # count shootings per year per borough

ggplot(shootings_year_boro, aes(x = year, y = n, color = BORO)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  labs(
    title = "Number of Shootings per Year by Borough",
    x = "Year",
    y = "Number of Shootings",
    color = "Borough"
  ) +
  theme_minimal()
```



The line plot above observes the shooting incidents per year by borough. Observers can see the actual observed points of data and should take note of the years that shooting incidents increased, decreased, peaked, and dipped. We can also see when boroughs switched positions and became a borough with less/more shooting incidents than another. This graph is used in the predictive model below.

Linear model for the Observed vs Predicted Shootings per Year by Borough

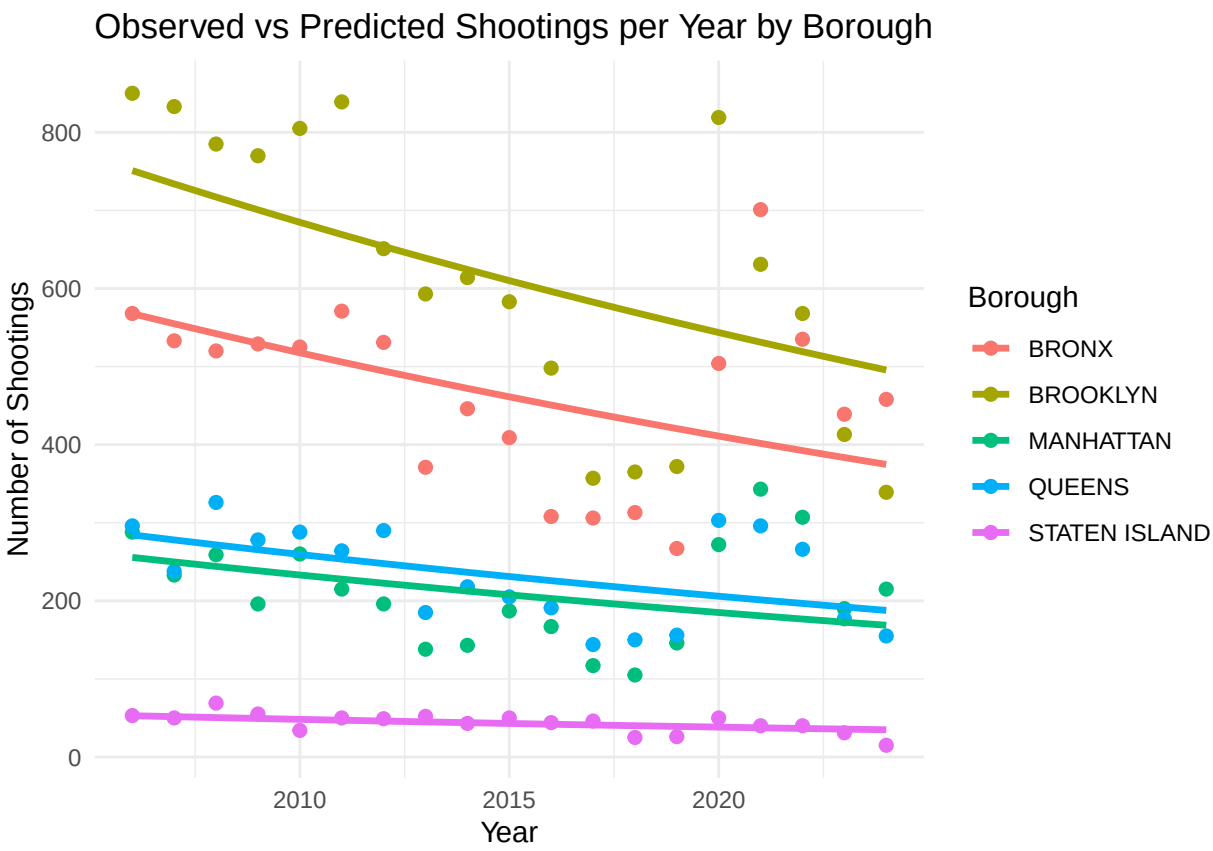
I then fit a Poisson regression model to the data. This allows me to compare actual counts of shootings (points) with predicted values from the model (lines). This approach also provides the basis for forecasting future shooting trends.

```
shootings_model <- glm(n ~ year + BORO, data = shootings_year_boro, family = poisson)

shootings_year_boro <- shootings_year_boro %>%
  mutate(predictions = predict(shootings_model, type = "response"))

ggplot(shootings_year_boro, aes(x = year, color = BORO)) +
  geom_point(aes(y = n), size = 2) +                # observed counts
  geom_line(aes(y = predictions), linewidth = 1.2) +    # model predictions
  labs(
```

```
    title = "Observed vs Predicted Shootings per Year by Borough",
    x = "Year",
    y = "Number of Shootings",
    color = "Borough"
  ) +
  theme_minimal()
```



The linear model above takes the actual observed points of data and creates a linear model to show a trend line in the shooting incidents per borough. Since we don't actually know the future, we can roughly predict the continuation of declining shooting incidents in each borough and a very rough estimate of the number of each shooting. We can also speculate on what might change the current trend lines, such as another global event like COVID-19 which was unpredictable and never-before-seen.

Conclusion and Bias

This project examined historical NYPD shooting incident data to identify patterns across years and boroughs. The visualizations showed that shootings have fluctuated over time, with some boroughs consistently experiencing higher counts than others. A Poisson regression model was used to compare observed and predicted values, and it provided a framework for forecasting future shooting trends. While the model captured general patterns, it is important to note that it simplifies the dynamics of shooting incidents and cannot account for all social, economic, or policy factors that may influence shootings in New York City. Some examples would be areas with higher and lower tourism, and overall population of each borough.

Bias is an important limitation of this project. The data set reflects only reported incidents and may be influenced by under reporting, police practices, or errors in record keeping. My own bias as the analyst also shapes the results. By choosing to focus primarily on borough and year, I emphasized geographic differences while leaving aside other possible dimensions such as demographics or contextual factors. I used a reproducible workflow, and avoided altering or excluding data without clear justification. The findings should therefore be interpreted as a partial view of a complex issue, not a definitive explanation of shooting trends in New York City.