NYPD Shooting Data Analysis

George B. Wofford

2025-01-22

Introduction

This report explores publicly available NYPD shooting incident data from New York City. The primary goal is to understand patterns and trends in shootings over time and across boroughs, and to investigate factors that may influence the likelihood of a shooting resulting in a murder classification.

Key Questions:

- How have shootings changed over the years captured in the dataset?
- Are there differences in the number of shootings among the boroughs of NYC?
- Is the timing of shootings (time of day) related to their frequency?
- Can basic variables (year, hour, borough) help us model the likelihood that a shooting is classified as a murder?

By addressing these questions, we aim to gain insights into temporal and spatial trends, and to experiment with a simple predictive model to see if basic factors correlate with shootings being deadly. Throughout, we will remain cautious about our interpretations and discuss possible sources of bias.

Data Source and Description

The data is sourced directly from the NYC Open Data Portal using the following URL provided as a parameter: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

This dataset includes information on shooting incidents in NYC, containing details such as the date, time, borough, victim and perpetrator demographics, and indicators of whether the incident was classified as a murder.

Note on Reproducibility: This R Markdown is fully reproducible. It imports data directly from the provided URL, ensuring that anyone knitting this document with the same code and environment can recreate the analysis and visualizations.

```
data_url <- params$data_url
nypd_shootings <- read_csv(data_url)
```

```
## Rows: 28562 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.
```

Data Preparation and Cleaning

Before analysis, we need to ensure the data is in a tidy and appropriate format. We will:

- Convert date and time columns into proper Date and time objects.
- Convert categorical variables such as borough and demographic groups into factors.
- Remove columns not needed for this analysis (e.g., unique incident keys, coordinate columns) to simplify the dataset.

```r
# Convert date from character to Date
nypd_shootings$OCCUR_DATE <- as.Date(nypd_shootings$OCCUR_DATE, format = "%m/%d/%Y")

# Convert time from character to a time object
nypd_shootings$OCCUR_TIME <- hms(nypd_shootings$OCCUR_TIME)

# Convert several categorical variables to factors
nypd_shootings$BORO <- as.factor(nypd_shootings$BORO)
nypd_shootings$LOC_OF_OCCUR_DESC <- as.factor(nypd_shootings$LOC_OF_OCCUR_DESC)
nypd_shootings$PERP_AGE_GROUP <- as.factor(nypd_shootings$PERP_AGE_GROUP)
nypd_shootings$PERP_SEX <- as.factor(nypd_shootings$PERP_SEX)
nypd_shootings$PERP_RACE <- as.factor(nypd_shootings$PERP_RACE)
nypd_shootings$VIC_AGE_GROUP <- as.factor(nypd_shootings$VIC_AGE_GROUP)
nypd_shootings$VIC_SEX <- as.factor(nypd_shootings$VIC_SEX)
nypd_shootings$VIC_RACE <- as.factor(nypd_shootings$VIC_RACE)
nypd_shootings$JURISDICTION_CODE <- as.factor(nypd_shootings$JURISDICTION_CODE)

# Remove unnecessary columns (coordinates and largely missing columns)
nypd_shootings <- subset(nypd_shootings,
                         select = -c(INCIDENT_KEY,
                                     LOC_OF_OCCUR_DESC,
                                     X_COORD_CD,
                                     Lon_Lat,
                                     Latitude,
                                     Longitude))

summary(nypd_shootings)
```

```
##    OCCUR_DATE           OCCUR_TIME                                  BORO
##  Min.   :2006-01-01   Min.   :0S                        BRONX        : 8376
##  1st Qu.:2009-09-04   1st Qu.:3H 30M 0S                 BROOKLYN     :11346
##  Median :2013-09-20   Median :15H 15M 0S                MANHATTAN    : 3762
##  Mean   :2014-06-07   Mean   :12H 44M 16.713115328057S   QUEENS       : 4271
##  3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S                STATEN ISLAND:  807
##  Max.   :2023-12-29   Max.   :23H 59M 0S
##
##     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC
```

```
##  Min.   :  1.0   0   :23923        Length:28562       Length:28562
##  1st Qu.: 44.0   1   :   81        Class :character   Class :character
##  Median : 67.0   2   : 4556        Mode  :character   Mode  :character
##  Mean   : 65.5   NA's:    2
##  3rd Qu.: 81.0
##  Max.   :123.0
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX             PERP_RACE
##  Mode :logical           18-24  :6438   (null): 1141   BLACK         :11903
##  FALSE:23036             25-44  :6041   F     :  444   WHITE HISPANIC: 2510
##  TRUE :5526              UNKNOWN:3148   M     :16168   UNKNOWN       : 1837
##                          <18    :1682   U     : 1499   BLACK HISPANIC: 1392
##                          (null) :1141   NA's  : 9310   (null)        : 1141
##                          (Other): 768                  (Other)       :  469
##                          NA's   :9344                  NA's          : 9310
##  VIC_AGE_GROUP   VIC_SEX                             VIC_RACE
##  <18    : 2954   F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  1022   :    1   M:25790   ASIAN / PACIFIC ISLANDER      :  440
##  18-24  :10384   U:   12   BLACK                         :20235
##  25-44  :12973             BLACK HISPANIC                : 2795
##  45-64  : 1981             UNKNOWN                       :   70
##  65+    :  205             WHITE                         :  728
##  UNKNOWN:   64             WHITE HISPANIC                : 4283
##    Y_COORD_CD
##  Min.   :125757
##  1st Qu.:182912
##  Median :194901
##  Mean   :208380
##  3rd Qu.:239814
##  Max.   :271128
##
```

Exploratory Data Analysis and Visualizations

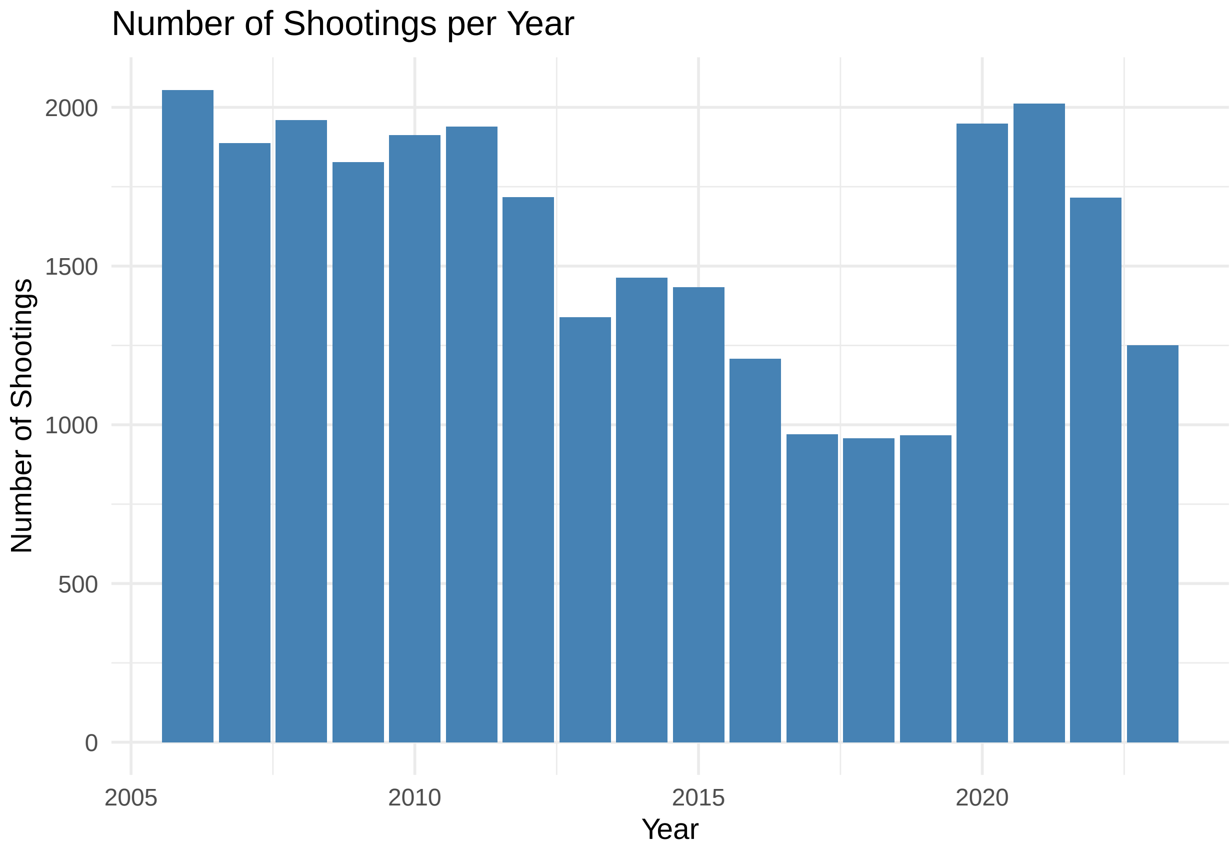
Shootings Over the Years

We first explore how shootings vary by year to detect any long-term trends.

```r
nypd_shootings <- nypd_shootings %>% mutate(YEAR = year(OCCUR_DATE))

yearly_counts <- nypd_shootings %>%
  group_by(YEAR) %>%
  summarize(Total_Shootings = n())

ggplot(yearly_counts, aes(x = YEAR, y = Total_Shootings)) +
  geom_col(fill = "steelblue") +
  labs(title = "Number of Shootings per Year",
       x = "Year",
       y = "Number of Shootings") +
  theme_minimal()
```



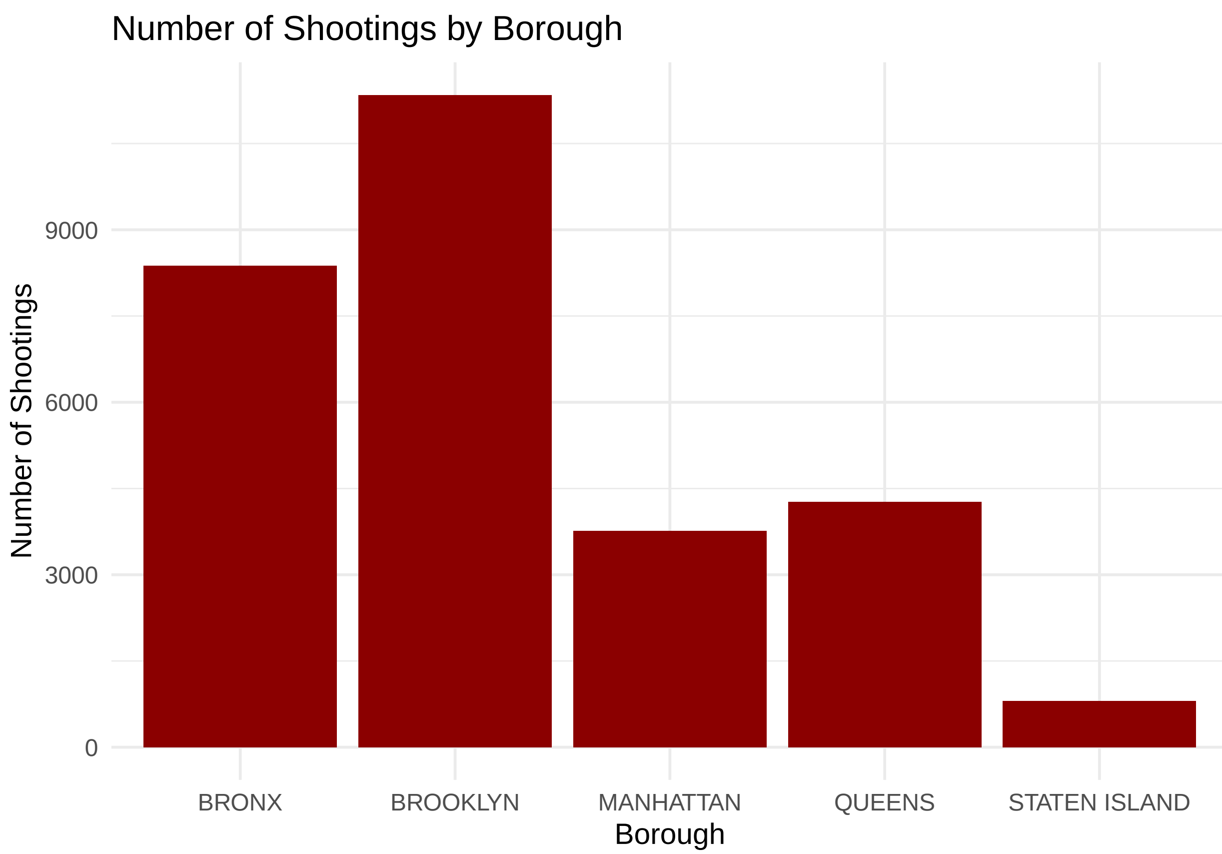
Interpretation: Some years may have more shootings than others, suggesting underlying shifts in policy, socioeconomic conditions, or law enforcement practices.

Shootings by Borough

Next, we look at the distribution of shootings across NYC boroughs to see which areas have higher frequencies.

```
boro_counts <- nypd_shootings %>%
  group_by(BORO) %>%
  summarize(Total_Shootings = n())

ggplot(boro_counts, aes(x = BORO, y = Total_Shootings)) +
  geom_col(fill = "darkred") +
  labs(title = "Number of Shootings by Borough",
       x = "Borough",
       y = "Number of Shootings") +
  theme_minimal()
```



Interpretation: Certain boroughs show consistently higher shooting counts. This might raise questions about neighborhood-level differences or other structural factors.

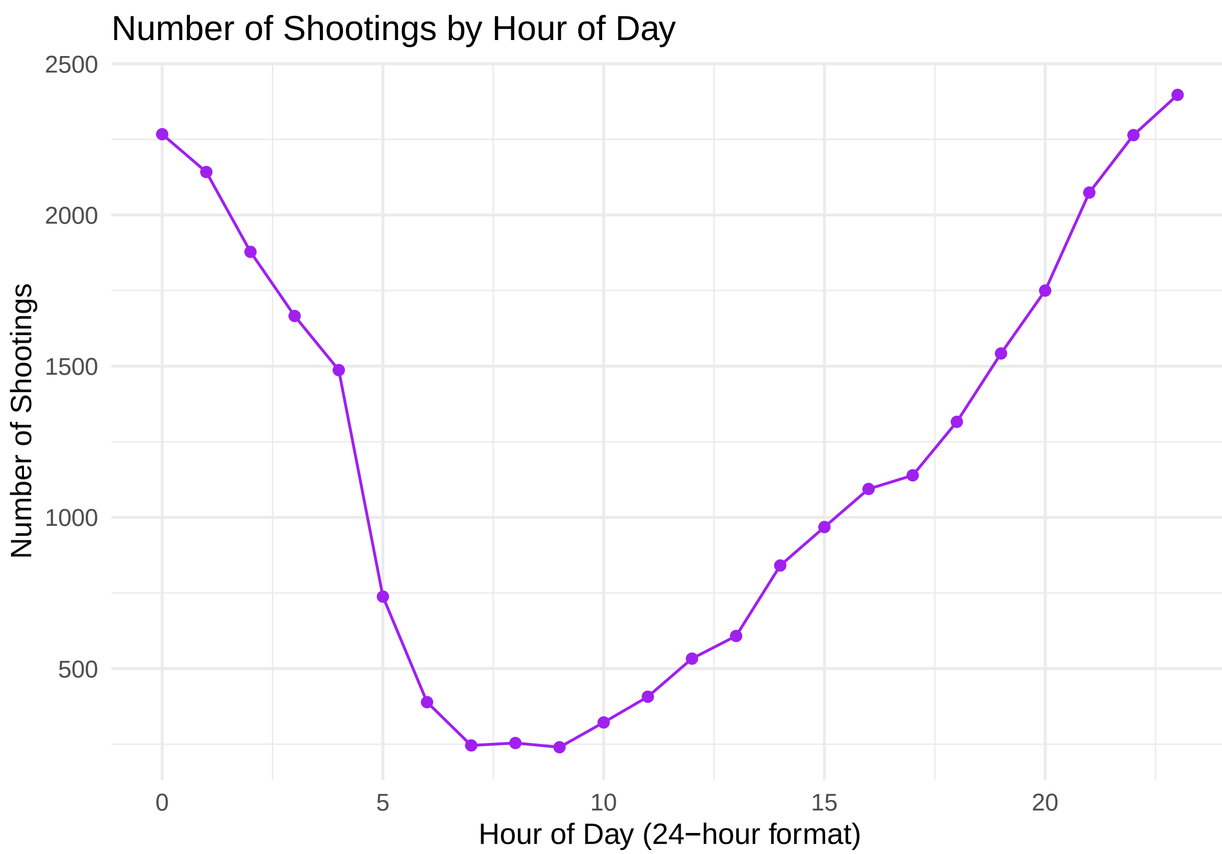
Shootings by Time of Day

We also consider the time of day. Do shootings peak at certain hours?

```
nypd_shootings <- nypd_shootings %>% mutate(HOUR = hour(OCCUR_TIME))

hour_counts <- nypd_shootings %>%
  group_by(HOUR) %>%
  summarize(Total_Shootings = n())

ggplot(hour_counts, aes(x = HOUR, y = Total_Shootings)) +
  geom_line(color = "purple") +
  geom_point(color = "purple") +
  labs(title = "Number of Shootings by Hour of Day",
       x = "Hour of Day (24-hour format)",
       y = "Number of Shootings") +
  theme_minimal()
```



Interpretation: If we see a pattern, perhaps increased shootings late at night, it might suggest that certain social activities or reduced police presence align with higher incident counts.

Statistical Modeling

We will fit a simple logistic regression model to investigate if YEAR, HOUR, and BORO can help predict whether a shooting is classified as a murder (STATISTICAL_MURDER_FLAG).

```
# Ensure murder flag is a binary factor
nypd_shootings$STATISTICAL_MURDER_FLAG <- factor(nypd_shootings$STATISTICAL_MURDER_FLAG,
                                                 levels = c(FALSE, TRUE))

# Fit a logistic regression model
murder_model <- glm(STATISTICAL_MURDER_FLAG ~ YEAR + HOUR + BORO,
                    data = nypd_shootings,
                    family = binomial(link = "logit"))

summary(murder_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ YEAR + HOUR + BORO, family = binomial(link = "logit"),
##     data = nypd_shootings)
##
## Coefficients:
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.4158662  5.6132849  -0.609   0.5428
## YEAR               0.0009853  0.0027874   0.353   0.7237
## HOUR               0.0011272  0.0017777   0.634   0.5260
## BOROBROOKLYN      -0.0015169  0.0364090  -0.042   0.9668
## BOROMANHATTAN     -0.1080222  0.0507284  -2.129   0.0332 *
## BOROQUEENS         0.0115520  0.0473956   0.244   0.8074
## BOROSTATEN ISLAND  0.0976385  0.0906412   1.077   0.2814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28061  on 28561  degrees of freedom
## Residual deviance: 28053  on 28555  degrees of freedom
## AIC: 28067
##
## Number of Fisher Scoring iterations: 4
```

Interpretation: The coefficients indicate which factors might be associated with higher odds of a shooting being classified as a murder. A positive coefficient for a given borough, for instance, might suggest that shootings in that borough have a higher probability of being deadly, controlling for year and hour. This is a simplistic model and should not be taken as definitive. Nonetheless, it demonstrates how data can be used to attempt predictive or explanatory modeling.

Discussion and Potential Biases

Reporting and Collection Bias:

The dataset only includes reported shootings, so it may not reflect all incidents. Reporting practices vary over time and between communities, influencing the apparent trends.

Temporal Bias:

Over multiple years, changes in policies, economic conditions, and societal trends can affect both the frequency and classification of shootings. Without context, raw trends could be misinterpreted.

Categorization and Missing Data:

Many variables were turned into factors, and some categories contain unknown or null values. Missing or imprecise data may skew patterns, especially if some groups are less frequently identified.

Analyst Bias:

The selection of which variables to visualize and the simplicity of the chosen model reflect certain assumptions. A more thorough analysis might incorporate additional demographic details, spatial analysis using coordinates (if retained), or more robust modeling techniques.

Conclusion

This exploratory analysis of NYPD shooting data provides a high-level look at how shootings vary over time, by borough, and by time of day. We introduced a simple logistic model to see if a few variables predict the classification of a shooting as a murder. The findings highlight temporal trends, geographic differences, and potential time-of-day patterns.

However, the analysis is only a starting point. It raises further questions about the underlying drivers of these patterns, and any interpretation must be cautious due to the potential biases in data collection and categorization. A deeper dive with more sophisticated modeling, external data sources, and domain expertise would be needed to draw stronger, more actionable conclusions.

Next Steps:

- Incorporate socioeconomic or law enforcement policy data for richer context.
- Consider more advanced models or machine learning approaches.
- Conduct sensitivity analyses to understand the impact of missing or uncertain data.

Overall, while we have visualized key trends and fit a basic model, we recognize that this analysis merely scratches the surface of understanding complex social phenomena like shootings in a large, diverse city.