NYPD Shooting Project

DAndre Tafoya

12-9-2025

Project Overview

This document presents a reproducible analysis of the NYPD Shooting Incident Data (Historic).

All analysis steps, from data import through visualization and interpretation, are contained in this R Markdown document. The raw data are read directly from CSV files using relative paths so that peers can download the GitHub repository and knit this document without modification.

Reproducibility and Setup

```r
# Libraries used in this project
library(readr)
library(dplyr)
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Data Import

```r
shootings_raw <- read_csv(
  "data/NYPD_Shooting_Incident_Data__Historic_.csv",
  show_col_types = FALSE
)

dim(shootings_raw)
```

```
## [1] 29744    21
```

```r
head(shootings_raw)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    231974218 08/09/2021 01:06      BRONX    <NA>                    40
## 2    177934247 04/07/2018 19:48      BROOKLYN <NA>                    79
## 3    255028563 12/02/2022 22:57       BRONX    OUTSIDE                 47
## 4     25384540 11/19/2006 01:50       BROOKLYN <NA>                    66
## 5     72616285 05/09/2010 01:58       BRONX    <NA>                    46
## 6     85875439 07/22/2012 21:35       BRONX    <NA>                    42
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Data Cleaning and Tidying

```
shootings_step1 <- shootings_raw %>%
  # Remove columns with excessive missingness
  select(-LOC_OF_OCCUR_DESC, -LOC_CLASSFCTN_DESC) %>%

  # Rename variables to be clean, simple, and consistent
  rename(
    incident_id   = INCIDENT_KEY,
    date          = OCCUR_DATE,
    time          = OCCUR_TIME,
    borough       = BORO,
    precinct      = PRECINCT,
    jurisdiction  = JURISDICTION_CODE,

    victim_age    = VIC_AGE_GROUP,
    victim_sex    = VIC_SEX,
    victim_race   = VIC_RACE,

    suspect_age   = PERP_AGE_GROUP,
    suspect_sex   = PERP_SEX,
    suspect_race  = PERP_RACE,

    murder_flag   = STATISTICAL_MURDER_FLAG,
    location_desc = LOCATION_DESC,

    x_coord       = X_COORD_CD,
    y_coord       = Y_COORD_CD,
    latitude      = Latitude,
    longitude     = Longitude
  )

# Inspect result
glimpse(shootings_step1)
```

```
## Rows: 29,744
## Columns: 19
## $ incident_id   <dbl> 231974218, 177934247, 255028563, 25384540, 72616285, 858~
## $ date          <chr> "08/09/2021", "04/07/2018", "12/02/2022", "11/19/2006", ~
## $ time          <time> 01:06:00, 19:48:00, 22:57:00, 01:50:00, 01:58:00, 21:35~
## $ borough       <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN", "BRONX", "BRON~
## $ precinct      <dbl> 40, 79, 47, 66, 46, 42, 71, 69, 75, 69, 40, 42, 71, 47, ~
## $ jurisdiction  <dbl> 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 2,~
## $ location_desc <chr> NA, NA, "GROCERY/BODEGA", "PVT HOUSE", "MULTI DWELL - AP~
## $ murder_flag   <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, TRUE, FALSE, FALS~
## $ suspect_age   <chr> NA, "25-44", "(null)", "UNKNOWN", "25-44", "18-24", NA, ~
## $ suspect_sex   <chr> NA, "M", "(null)", "U", "M", "M", NA, NA, "M", "M", "M",~
## $ suspect_race  <chr> NA, "WHITE HISPANIC", "(null)", "UNKNOWN", "BLACK", "BLA~
## $ victim_age    <chr> "18-24", "25-44", "25-44", "18-24", "<18", "18-24", "25--~
## $ victim_sex    <chr> "M", "M", "M", "M", "F", "M", "M", "M", "M", "M", "M", "~
## $ victim_race   <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "B~
## $ x_coord       <dbl> 1006343.0, 1000082.9, 1020691.0, 985107.3, 1009853.5, 10~
## $ y_coord       <dbl> 234270.0, 189064.7, 257125.0, 173349.8, 247502.6, 239814~
## $ latitude      <dbl> 40.80967, 40.68561, 40.87235, 40.64249, 40.84598, 40.824~
## $ longitude     <dbl> -73.92019, -73.94291, -73.86823, -73.99691, -73.90746, --~
## $ Lon_Lat       <chr> "POINT (-73.92019278899994 40.80967347200004)", "POINT (~
```

```r
shootings_step2 <- shootings_step1 %>%
  mutate(
    # Convert date and time
    date = mdy(date),
    time = hm(time),

    # Convert categorical variables to factors
    borough      = factor(borough),
    precinct     = factor(precinct),
    jurisdiction = factor(jurisdiction),

    victim_age   = factor(victim_age),
    victim_sex   = factor(victim_sex),
    victim_race  = factor(victim_race),

    suspect_age  = factor(suspect_age),
    suspect_sex  = factor(suspect_sex),
    suspect_race = factor(suspect_race),

    # Convert murder flag to factor for clarity
    murder_flag = factor(
      murder_flag,
      levels = c(FALSE, TRUE),
      labels = c("No", "Yes")
    ),

    # Ensure coordinates are numeric
    x_coord   = as.numeric(x_coord),
    y_coord   = as.numeric(y_coord),
    latitude  = as.numeric(latitude),
    longitude = as.numeric(longitude)
  )

glimpse(shootings_step2)
```

```
## Rows: 29,744
## Columns: 19
## $ incident_id   <dbl> 231974218, 177934247, 255028563, 25384540, 72616285, 858~
## $ date           <date> 2021-08-09, 2018-04-07, 2022-12-02, 2006-11-19, 2010-05~
## $ time           <Period> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ borough        <fct> BRONX, BROOKLYN, BRONX, BROOKLYN, BRONX, BRONX, BROOKLYN~
## $ precinct       <fct> 40, 79, 47, 66, 46, 42, 71, 69, 75, 69, 40, 42, 71, 47, ~
## $ jurisdiction   <fct> 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 2,~
## $ location_desc  <chr> NA, NA, "GROCERY/BODEGA", "PVT HOUSE", "MULTI DWELL - AP~
## $ murder_flag    <fct> No, Yes, No, Yes, Yes, No, Yes, No, No, No, No, No, No, ~
## $ suspect_age    <fct> NA, 25-44, (null), UNKNOWN, 25-44, 18-24, NA, NA, 25-44,~
## $ suspect_sex    <fct> NA, M, (null), U, M, M, NA, NA, M, M, M, M, M, M, NA, M,~
## $ suspect_race   <fct> NA, WHITE HISPANIC, (null), UNKNOWN, BLACK, BLACK, NA, N~
## $ victim_age     <fct> 18-24, 25-44, 25-44, 18-24, <18, 18-24, 25-44, 25-44, 25~
## $ victim_sex     <fct> M, M, M, M, F, M, M, M, M, M, M, M, M, M, M, M, M, M, M,~
## $ victim_race    <fct> BLACK, BLACK, BLACK, BLACK, BLACK, BLACK, BLACK, WHITE H~
## $ x_coord        <dbl> 1006343.0, 1000082.9, 1020691.0, 985107.3, 1009853.5, 10~
## $ y_coord        <dbl> 234270.0, 189064.7, 257125.0, 173349.8, 247502.6, 239814~
## $ latitude       <dbl> 40.80967, 40.68561, 40.87235, 40.64249, 40.84598, 40.824~
## $ longitude      <dbl> -73.92019, -73.94291, -73.86823, -73.99691, -73.90746, -~
## $ Lon_Lat        <chr> "POINT (-73.92019278899994 40.80967347200004)", "POINT (~
```

I want to focus my analysis on shootings from 2010 to 2024.

```
shootings_2010s <- shootings_step2 %>%
  filter(date >= as.Date("2010-01-01"))

range(shootings_2010s$date, na.rm = TRUE)
```

```
## [1] "2010-01-01" "2024-12-31"
```

```
library(tidyr)

missing_2010s <- shootings_2010s %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(
    cols = everything(),
    names_to = "variable",
    values_to = "n_missing"
  ) %>%
  arrange(desc(n_missing))

missing_2010s
```

```
## # A tibble: 19 x 2
##    variable      n_missing
##    <chr>             <int>
##  1 time              22015
##  2 location_desc     10929
##  3 suspect_age        8487
##  4 suspect_sex        8487
##  5 suspect_race       8487
##  6 latitude             97
##  7 longitude            97
##  8 Lon_Lat              97
##  9 jurisdiction          1
## 10 incident_id           0
## 11 date                  0
## 12 borough               0
## 13 precinct              0
## 14 murder_flag           0
## 15 victim_age            0
## 16 victim_sex            0
## 17 victim_race           0
## 18 x_coord               0
## 19 y_coord               0
```

Exploratory Data Analysis and Visualizations

```r
#Create year
shootings_w_year <- shootings_2010s %>%
  mutate(year = year(date))

range(shootings_w_year$year, na.rm = TRUE)
```

```
## [1] 2010 2024
```

```r
analysis_dataset <- shootings_w_year %>%
  select(
    year,
    borough,
    murder_flag
  )

glimpse(analysis_dataset)
```

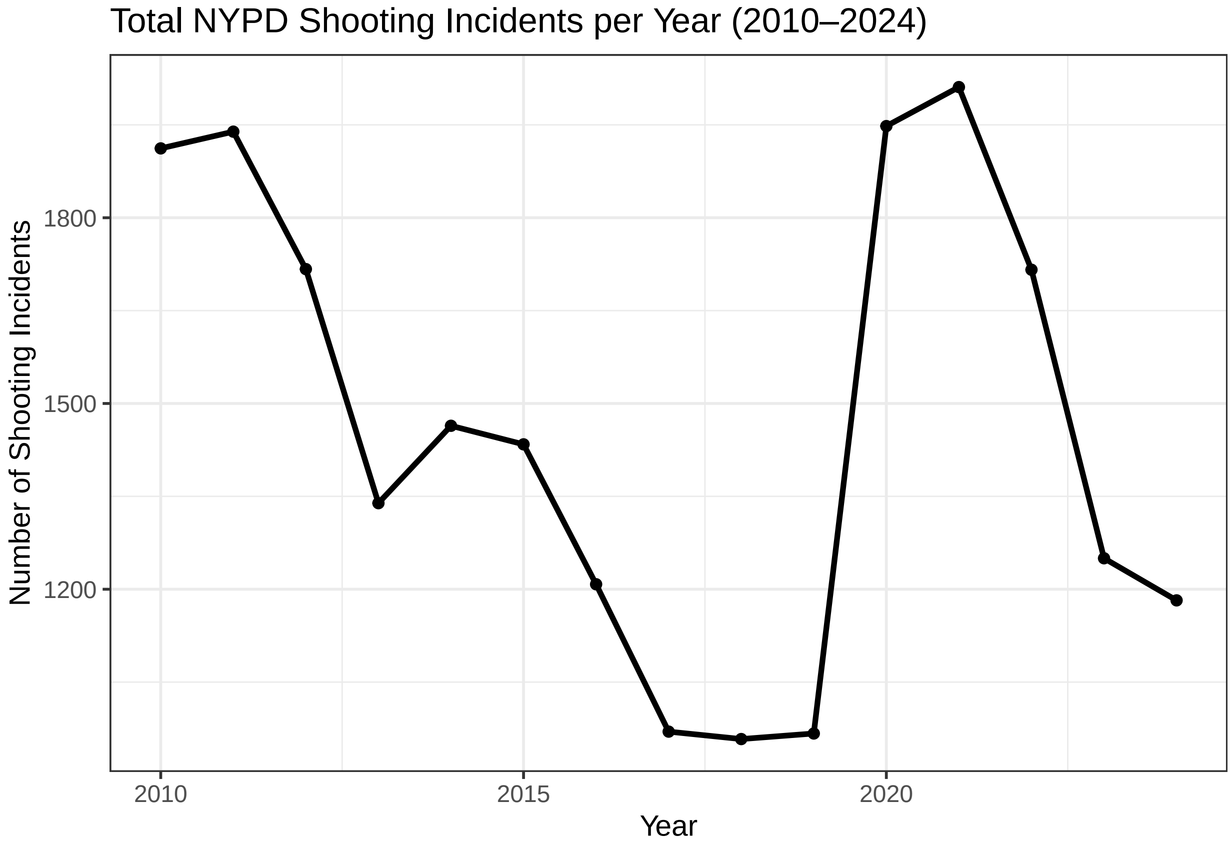
```
## Rows: 22,015
## Columns: 3
## $ year        <dbl> 2021, 2018, 2022, 2010, 2012, 2011, 2012, 2015, 2016, 2010~
## $ borough     <fct> BRONX, BROOKLYN, BRONX, BRONX, BRONX, BROOKLYN, BROOKLYN, ~
## $ murder_flag <fct> No, Yes, No, Yes, No, Yes, No, No, No, No, Yes, Yes, No, N~
```

We will begin the exploratory analysis by examining the total number of shooting incidents in New York has changed from 2010 to 2024. This will help us understand overall trends before taking a closer look at fatal versus non fatal shootings and differences across boroughs.

```r
yearly_counts <- analysis_dataset %>%
  count(year)
yearly_counts
```

```
## # A tibble: 15 x 2
##     year     n
##    <dbl> <int>
##  1  2010  1912
##  2  2011  1939
##  3  2012  1717
##  4  2013  1339
##  5  2014  1464
##  6  2015  1434
##  7  2016  1208
##  8  2017   970
##  9  2018   958
## 10  2019   967
## 11  2020  1948
## 12  2021  2011
## 13  2022  1716
## 14  2023  1250
## 15  2024  1182
```

```
ggplot(yearly_counts, aes(x = year, y = n)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(
    title = "Total NYPD Shooting Incidents per Year (2010-2024)",
    x = "Year",
    y = "Number of Shooting Incidents"
  ) +
  theme_bw()
```



We can see that in the years 2010 and 2020 the number of shooting incidents is alarmingly high. This could be related to economic disruption. The high count in 2010 could possibly reflect lingering social and economic effects of the Great Recession. The surge in 2020 aligns with the COVID-19 recession. Although economic indicators are not included in this analysis both periods suggest that economic and social disruptions during these periods may coincide with an increase in violent crime.

Next, we will examine how fatal and nonfatal shootings have changed over time.

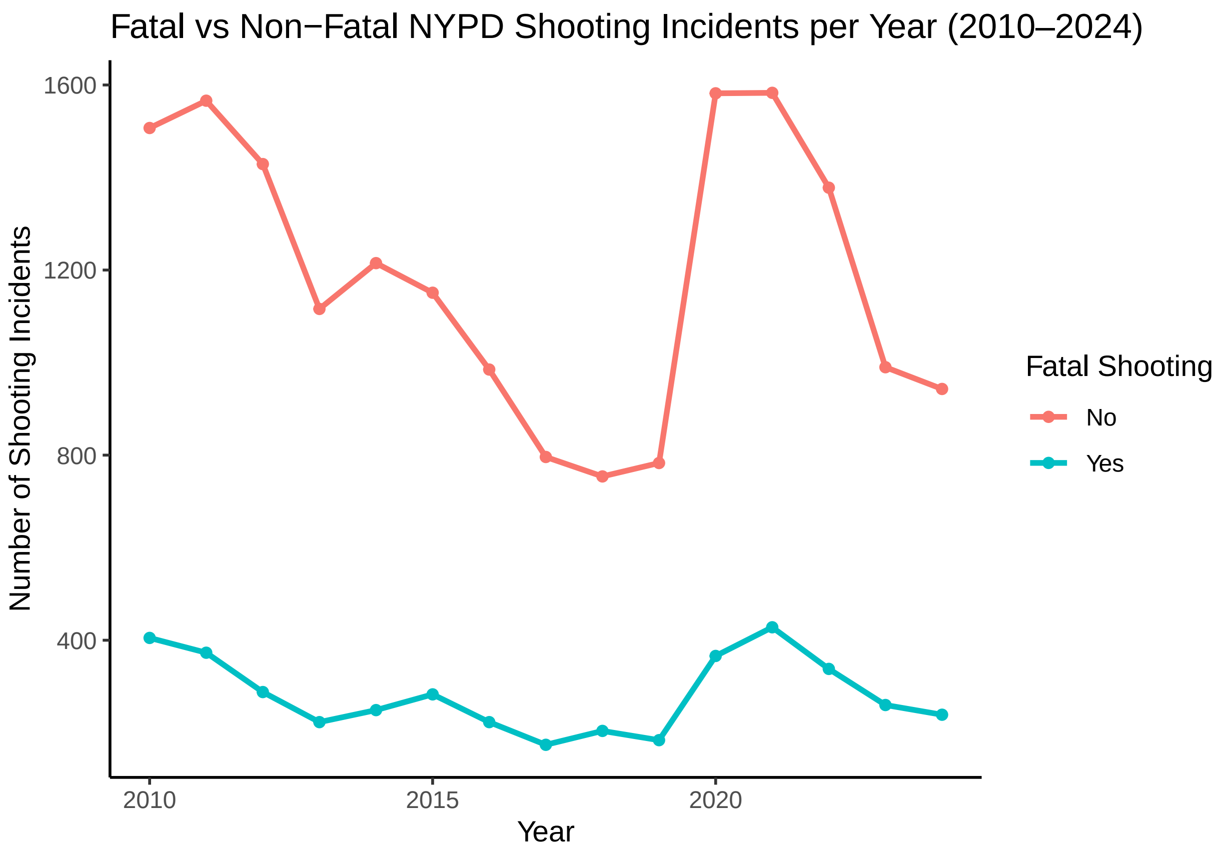
```r
yearly_fatal_counts <- analysis_dataset %>%
  count(year, murder_flag)

yearly_fatal_counts
```

```
## # A tibble: 30 x 3
##     year murder_flag     n
##    <dbl> <fct>       <int>
##  1  2010 No           1507
##  2  2010 Yes           405
##  3  2011 No           1566
##  4  2011 Yes           373
##  5  2012 No           1429
##  6  2012 Yes           288
##  7  2013 No           1116
##  8  2013 Yes           223
##  9  2014 No           1215
## 10  2014 Yes           249
## # i 20 more rows
```

```r
library(ggplot2)

ggplot(yearly_fatal_counts,
       aes(x = year, y = n, color = murder_flag)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(
    title = "Fatal vs Non-Fatal NYPD Shooting Incidents per Year (2010-2024)",
    x = "Year",
    y = "Number of Shooting Incidents",
    color = "Fatal Shooting"
  ) +
  theme_classic()
```



This chart shows us that fatal and non-fatal shootings follow similar trends over time. This is expected because we can safely assume that as the number of shooting incidents increase the number of both fatal and non-fatal shootings also increase. The purpose of this chart is to assess whether shootings are becoming more deadly over time.

Before fitting a model, we should examine how shooting incidents have changed over time across New York City boroughs. We would like to see if trends are consistent across boroughs or driven by changes in specific areas.

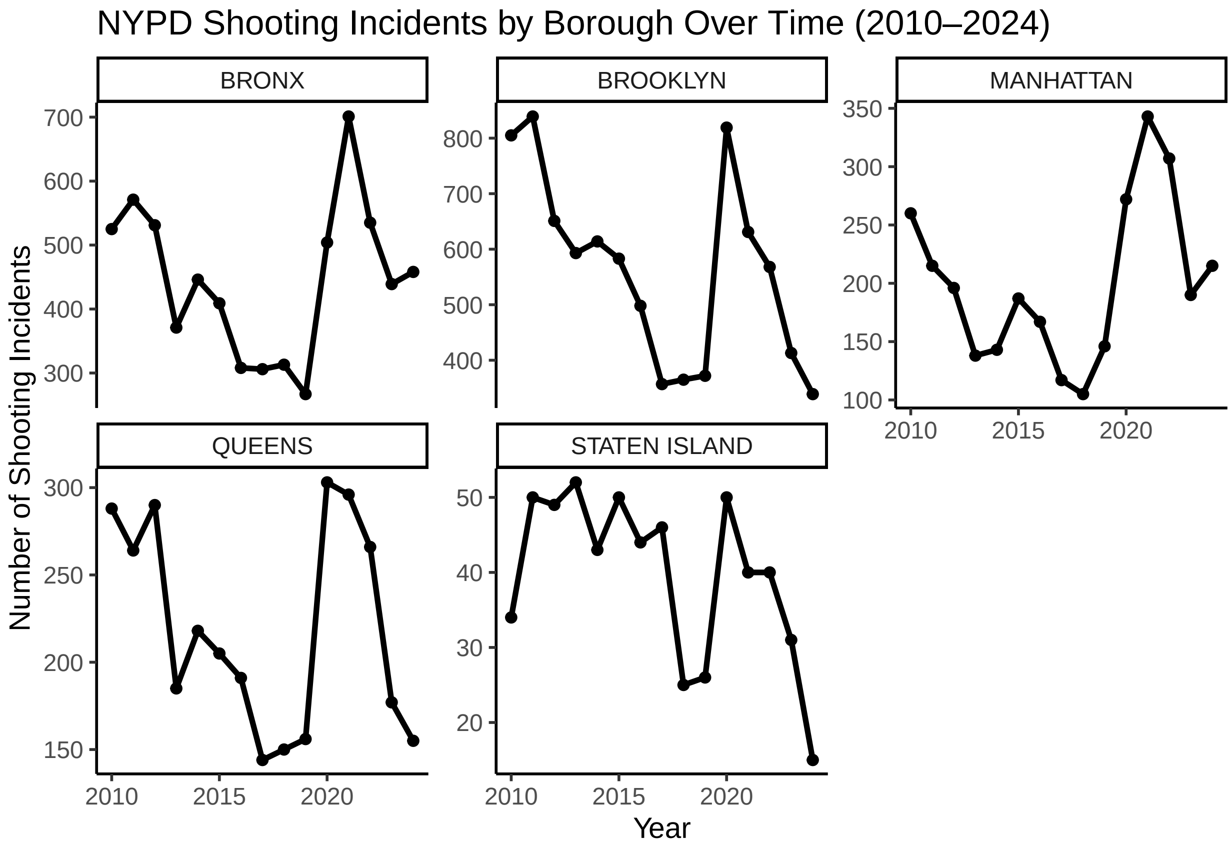
```
borough_yearly_counts <- analysis_dataset %>%
  count(year, borough)

borough_yearly_counts
```

```
## # A tibble: 75 x 3
##     year borough           n
##    <dbl> <fct>         <int>
##  1  2010 BRONX           525
##  2  2010 BROOKLYN        805
##  3  2010 MANHATTAN       260
##  4  2010 QUEENS          288
##  5  2010 STATEN ISLAND    34
##  6  2011 BRONX           571
##  7  2011 BROOKLYN        839
##  8  2011 MANHATTAN       215
##  9  2011 QUEENS          264
## 10  2011 STATEN ISLAND    50
## # i 65 more rows
```

```r
library(ggplot2)

ggplot(borough_yearly_counts,
       aes(x = year, y = n)) +
  geom_line(linewidth = 1) +
  geom_point() +
  facet_wrap(~ borough, scales = "free_y") +
  labs(
    title = "NYPD Shooting Incidents by Borough Over Time (2010-2024)",
    x = "Year",
    y = "Number of Shooting Incidents"
  ) +
  theme_classic()
```



It’s apparent that trends across boroughs are very similar indicating that temporal factors plays a larger role than location in shaping shooting patterns. This consistency could also further support the idea that economic and social disruptions affecting the city as a whole may leads to an increase in the number of shootings.

Modeling

We have observed differences in shooting patterns over time and across boroughs. For further examination of these patterns, we will fit a logistic regression model to estimate the probability that a shooting results in a fatality of as a function of year and borough.

Model specification

- **Outcome:** `murder_flag`

- **Predictors:** `year`, `borough`

- **Family:** binomial (logistic regression)

```r
fatal_model <- glm(
  murder_flag ~ year + borough,
  data = analysis_dataset,
  family = binomial
)

summary(fatal_model)
```

```
##
## Call:
## glm(formula = murder_flag ~ year + borough, family = binomial,
##     data = analysis_dataset)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -18.816089   7.691738  -2.446   0.0144 *
## year                  0.008629   0.003813   2.263   0.0236 *
## boroughBROOKLYN      -0.029375   0.041522  -0.707   0.4793
## boroughMANHATTAN     -0.125424   0.056862  -2.206   0.0274 *
## boroughQUEENS         0.039082   0.053274   0.734   0.4632
## boroughSTATEN ISLAND  0.029966   0.107060   0.280   0.7796
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21565  on 22014  degrees of freedom
## Residual deviance: 21552  on 22009  degrees of freedom
## AIC: 21564
##
## Number of Fisher Scoring iterations: 4
```

Summary Analysis - The model is a logistic regression predicting the probability that a shooting results in a fatality (`murder_flag`) using year and borough as predictors.

- The estimated coefficient for year is positive and statistically significant ($p = 0.0236$), This means that, holding borough constant, the odds of a shooting being fatal increase slightly each year (about 0.9% higher odds per year, since exp(0.0086) approximately equal to 1.009).

- Manhattan has a statistically significant negative coefficient (p = 0.0274), suggesting lower odds of fatal outcomes compared to the reference borough.
- Coefficients for Brooklyn, Queens, and Staten Island are not statistically significant, indicating no detectable differences in fatality odds relative to the reference category for these boroughs.
- The intercept is statistically significant and reflects the baseline log-odds of a fatal shooting for the reference borough at year zero.
- The reduction from null deviance (21,565) to residual deviance (21,552) indicates that the model explains some variance in fatal outcomes, though the overall improvement is modest.
- The AIC value (21,564) reflects model fit and can be used for comparison with alternative specifications.
- The model converged successfully in four Fisher scoring iterations, indicating stable estimation.

```r
analysis_dataset %>%
  count(murder_flag) %>%
  mutate(
    proportion = n / sum(n)
  )
```

```
## # A tibble: 2 x 3
##   murder_flag     n proportion
##   <fct>       <int>      <dbl>
## 1 No          17778      0.808
## 2 Yes          4237      0.192
```

```r
analysis_dataset %>%
  group_by(borough) %>%
  summarise(
    total_incidents = n(),
    fatal_incidents = sum(murder_flag == "Yes"),
    fatal_rate = mean(murder_flag == "Yes")
  ) %>%
  arrange(desc(fatal_rate))
```

```
## # A tibble: 5 x 4
##   borough       total_incidents fatal_incidents fatal_rate
##   <fct>                   <int>           <int>      <dbl>
## 1 QUEENS                   3288             664      0.202
## 2 STATEN ISLAND             595             119      0.2
## 3 BRONX                    6684            1311      0.196
## 4 BROOKLYN                 8447            1610      0.191
## 5 MANHATTAN                3001             533      0.178
```

```r
analysis_dataset %>%
  group_by(year) %>%
  summarise(
    total_incidents = n(),
    fatal_incidents = sum(murder_flag == "Yes"),
    fatal_rate = mean(murder_flag == "Yes")
  )
```

```
## # A tibble: 15 x 4
##     year total_incidents fatal_incidents fatal_rate
##    <dbl>           <int>           <int>      <dbl>
##  1  2010            1912             405      0.212
##  2  2011            1939             373      0.192
##  3  2012            1717             288      0.168
##  4  2013            1339             223      0.167
##  5  2014            1464             249      0.170
##  6  2015            1434             283      0.197
##  7  2016            1208             223      0.185
##  8  2017             970             174      0.179
##  9  2018             958             204      0.213
## 10  2019             967             184      0.190
## 11  2020            1948             366      0.188
## 12  2021            2011             428      0.213
## 13  2022            1716             338      0.197
## 14  2023            1250             260      0.208
## 15  2024            1182             239      0.202
```

- **Overall fatality rate (2010–2024):**
  - Approximately 19% of shooting incidents are fatal, while about 81% are non-fatal.
  - Fatal shootings represent a substantial but smaller share of incidents.
  - The relative stability of fatality proportions over time suggests that changes in gun violence are driven more by incident frequency than by increased lethality.
- **Borough-level fatality patterns:**
  - Fatality rates vary only modestly across boroughs, ranging from roughly 18% to just over 20%.
  - Despite large differences in total shooting counts, geographic variation in lethality is limited.
  - This aligns with the regression results, where only Manhattan shows a statistically significant difference and other borough effects are small.
- **Temporal trends in fatality rates:**
  - Fatality rates decline slightly through the mid-2010s and increase again in the late 2010s and early 2020s.
  - These changes are modest in size but persist over time.
  - The pattern helps explain the statistically significant positive effect of year in the logistic regression model.

Limitations and Pitfalls

The logistic regression model is purely descriptive. The observed associations between year, borough, and the likelihood that a shooting was fatal does not imply causal relationships. There are many unobserved factors that our model does not capture which introduces omitted variable bias. Some examples of these unobserved factors could be emergency response time, incident context, and perhaps weapon type. We must also note that there are likely many shootings that were not reported to law enforcement due to the public's willingness to report a crime.