NYPD Shooting Analysis

Anonymous

2024-10-13

This report analyzes shooting incidents across New York City's boroughs, focusing on the number of incidents and the time of day they occur. The goal is to identify which boroughs experience the most gun violence and when incidents are most likely to happen. By understanding these patterns, we can highlight areas and time periods where targeted interventions may be most effective in reducing gun violence.

Load the library

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

Inport dataset

```
# Read the csv file
nypd_01 = read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
## 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.
```

```r
# Summary of dataset
head(nypd_01)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # 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>
```

Tidy and transform

We clean the data by removing unnecessary columns, transforming date and time columns, and dropping rows with missing values in key columns (such as Age Group, Gender, and Location).

```r
# Remove unnecessary columns
nypd_02 = nypd_01 %>%
  select(-c(X_COORD_CD:Lon_Lat, PRECINCT, JURISDICTION_CODE))

# Transform date and time columns
nypd_02 = nypd_02 %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME))

# Drop rows with missing values in key columns (Age Group, Gender, Location)
nypd_02 <- nypd_02 %>%
  filter(!is.na(VIC_AGE_GROUP), !is.na(VIC_SEX), !is.na(LOCATION_DESC))

# Check the cleaned data
summary(nypd_02)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 64893279   1st Qu.:2009-08-18   1st Qu.:3H 43M 0S
##  Median : 92606073   Median :2013-09-14   Median :15H 0M 0S
##  Mean   :135731057   Mean   :2014-12-09   Mean   :12H 45M 22.71917556128035
##  3rd Qu.:227344212   3rd Qu.:2021-04-24   3rd Qu.:20H 30M 0S
##  Max.   :279758069   Max.   :2023-12-29   Max.   :23H 59M 0S
##      BORO           LOC_OF_OCCUR_DESC  LOC_CLASSFCTN_DESC LOCATION_DESC
##  Length:13585       Length:13585       Length:13585       Length:13585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

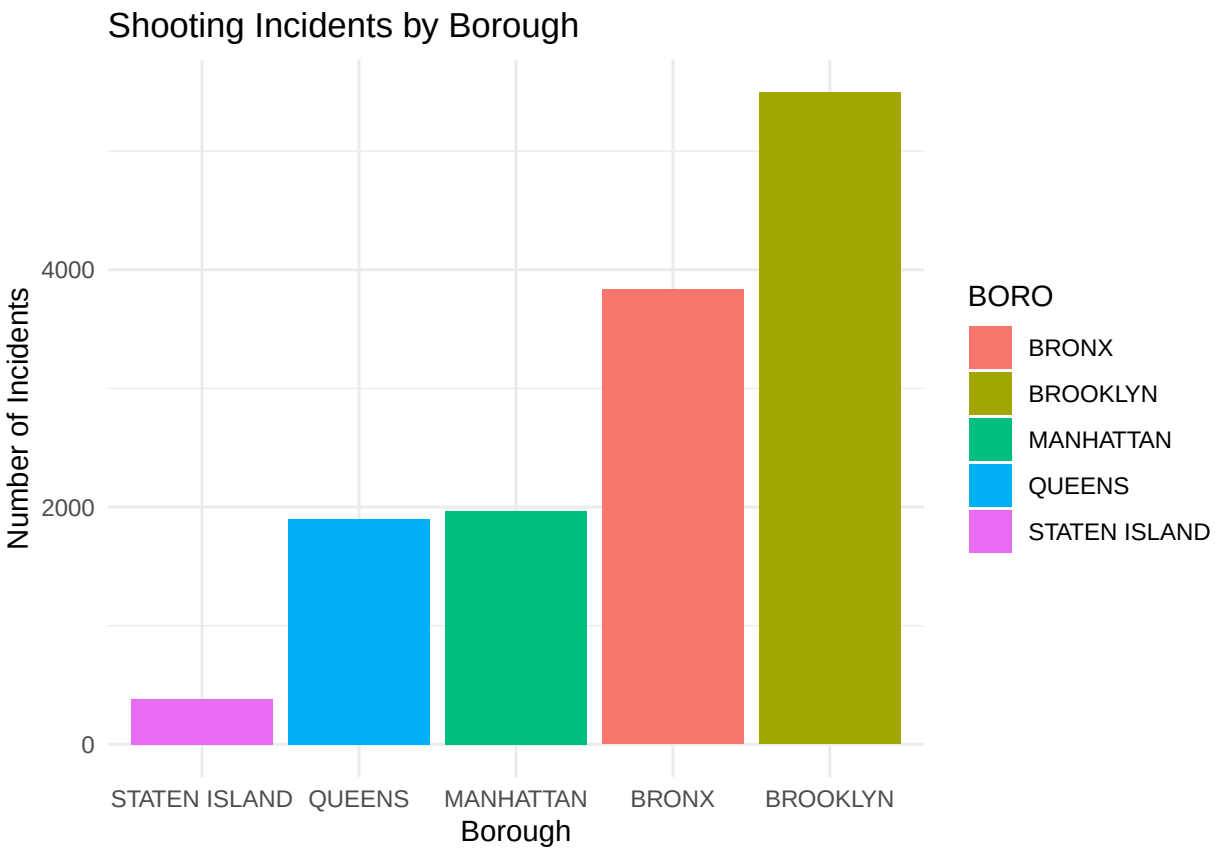
```
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:13585       Length:13585
##  FALSE:10568             Class :character   Class :character
##  TRUE :3017              Mode  :character   Mode  :character
##
##
##
##   PERP_RACE          VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:13585       Length:13585       Length:13585       Length:13585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Visualization : Shooting incidents by borough

This analysis looks at how shooting incidents are distributed across the different boroughs of New York City: Manhattan, Brooklyn, The Bronx, Queens, and Staten Island. By examining this, we can see which areas have the highest or lowest number of incidents.

Understanding these patterns helps focus efforts on the areas that need the most attention to reduce gun violence. The following bar chart shows the number of shooting incidents in each borough, giving a clear view of where incidents happen most often.

```r
# Group and summarize incidents by borough
nypd_02 %>%
  group_by(BORO) %>%
  summarize(Incidents = n()) %>%
  ggplot(aes(x = reorder(BORO, Incidents), y = Incidents, fill = BORO)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Shooting Incidents by Borough",
    x = "Borough",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```



The bar chart shows that Brooklyn has the highest number of shooting incidents, followed by The Bronx. Manhattan and Queens have moderate levels of incidents, while Staten Island reports the fewest.

This distribution highlights Brooklyn and The Bronx as the primary areas of concern for gun violence in New York City, indicating a need for focused resources and interventions in these boroughs.

Visualization : Shooting incidents by time of day for each borough

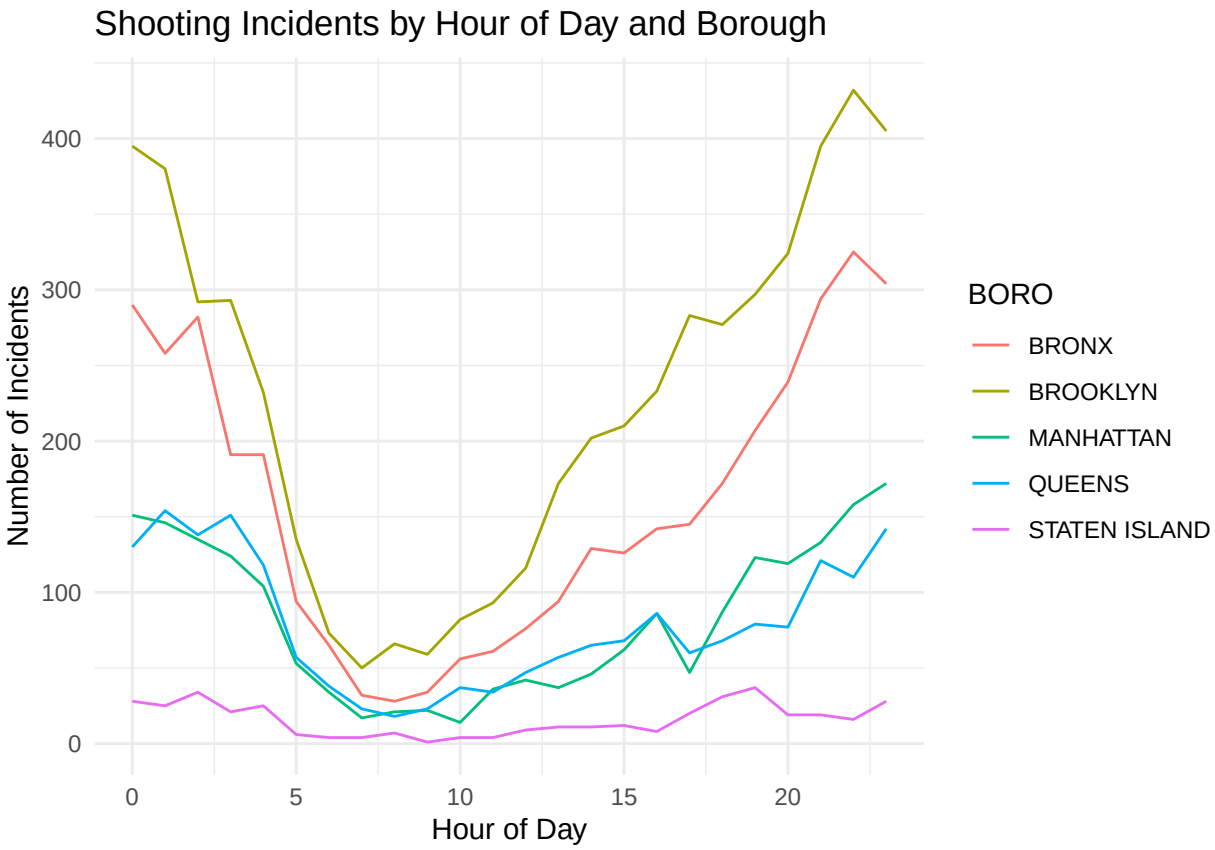
After analyzing shooting incidents by borough, it's important to see when these incidents happen. By looking at the time of day for each borough, we can identify key periods for gun violence and when prevention efforts may be most effective.

The following graph shows shooting incidents by time of day across each borough, highlighting the most active times for gun violence.

```
# Group and summarize incidents by location
# Group and summarize incidents by borough and time of day (hour)
nypd_02 %>%
  mutate(Hour = hour(OCCUR_TIME)) %>%
  group_by(BORO, Hour) %>%
  summarize(Incidents = n()) %>%
  ggplot(aes(x = Hour, y = Incidents, color = BORO)) +
  geom_line() +
  labs(
    title = "Shooting Incidents by Hour of Day and Borough",
    x = "Hour of Day",
```

```
    y = "Number of Incidents"
  ) +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```



The chart shows that shooting incidents peak in the evening, especially in Brooklyn and The Bronx. Manhattan and Queens follow similar patterns but with fewer incidents, while Staten Island consistently has the lowest numbers.

Model

With a clear understanding of the distribution of shooting incidents by borough and time of day, the next step is to predict how these factors influence the frequency of incidents. Using a linear regression model, we aim to determine how borough, time of day, age group, and gender impact the number of shooting incidents. This model will help identify key predictors and provide insights into which factors contribute most significantly to gun violence patterns in New York City.

```
# Summarize incidents by borough and other factors to get count per group
nypd_summary <- nypd_02 %>%
  group_by(BORO, Hour = hour(OCCUR_TIME), VIC_AGE_GROUP, VIC_SEX) %>%
  summarize(Incidents = n(), .groups = 'drop')
```

```r
# Check the summarized data
head(nypd_summary)
```

```
## # A tibble: 6 x 5
##   BORO   Hour VIC_AGE_GROUP VIC_SEX Incidents
##   <chr> <dbl> <chr>         <chr>       <int>
## 1 BRONX     0 18-24         F               6
## 2 BRONX     0 18-24         M              93
## 3 BRONX     0 25-44         F              14
## 4 BRONX     0 25-44         M             120
## 5 BRONX     0 45-64         F               2
## 6 BRONX     0 45-64         M              15
```

```r
# Create a Poisson regression model to predict the number of shooting incidents
poisson_model <- glm(Incidents ~ BORO + Hour + VIC_AGE_GROUP + VIC_SEX,
                     data = nypd_summary, family = poisson)

# Display the summary of the Poisson regression model
summary(poisson_model)
```

```
##
## Call:
## glm(formula = Incidents ~ BORO + Hour + VIC_AGE_GROUP + VIC_SEX,
##     family = poisson, data = nypd_summary)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.599405   0.042071  14.247  < 2e-16 ***
## BOROBROOKLYN            0.343491   0.021051  16.317  < 2e-16 ***
## BOROMANHATTAN          -0.641704   0.027734 -23.138  < 2e-16 ***
## BOROQUEENS             -0.679397   0.028057 -24.215  < 2e-16 ***
## BOROSTATEN ISLAND      -2.075896   0.053567 -38.753  < 2e-16 ***
## Hour                    0.015208   0.001236  12.308  < 2e-16 ***
## VIC_AGE_GROUP1022      -2.104904   1.000605  -2.104   0.0354 *
## VIC_AGE_GROUP18-24      1.208900   0.031021  38.970  < 2e-16 ***
## VIC_AGE_GROUP25-44      1.489550   0.030128  49.441  < 2e-16 ***
## VIC_AGE_GROUP45-64     -0.242469   0.040904  -5.928 3.07e-09 ***
## VIC_AGE_GROUP65+       -1.870682   0.096850 -19.315  < 2e-16 ***
## VIC_AGE_GROUPUNKNOWN -1.950321   0.173703 -11.228  < 2e-16 ***
## VIC_SEXM                1.903877   0.027106  70.239  < 2e-16 ***
## VIC_SEXU               -1.954910   0.500877  -3.903 9.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 24765.4  on 880  degrees of freedom
## Residual deviance:  5732.8  on 867  degrees of freedom
## AIC: 8966.1
##
## Number of Fisher Scoring iterations: 6
```

What the Data Tells Us:

Location: Brooklyn experiences significantly more incidents compared to other boroughs like Manhattan, Queens, and especially Staten Island, which has the fewest incidents.

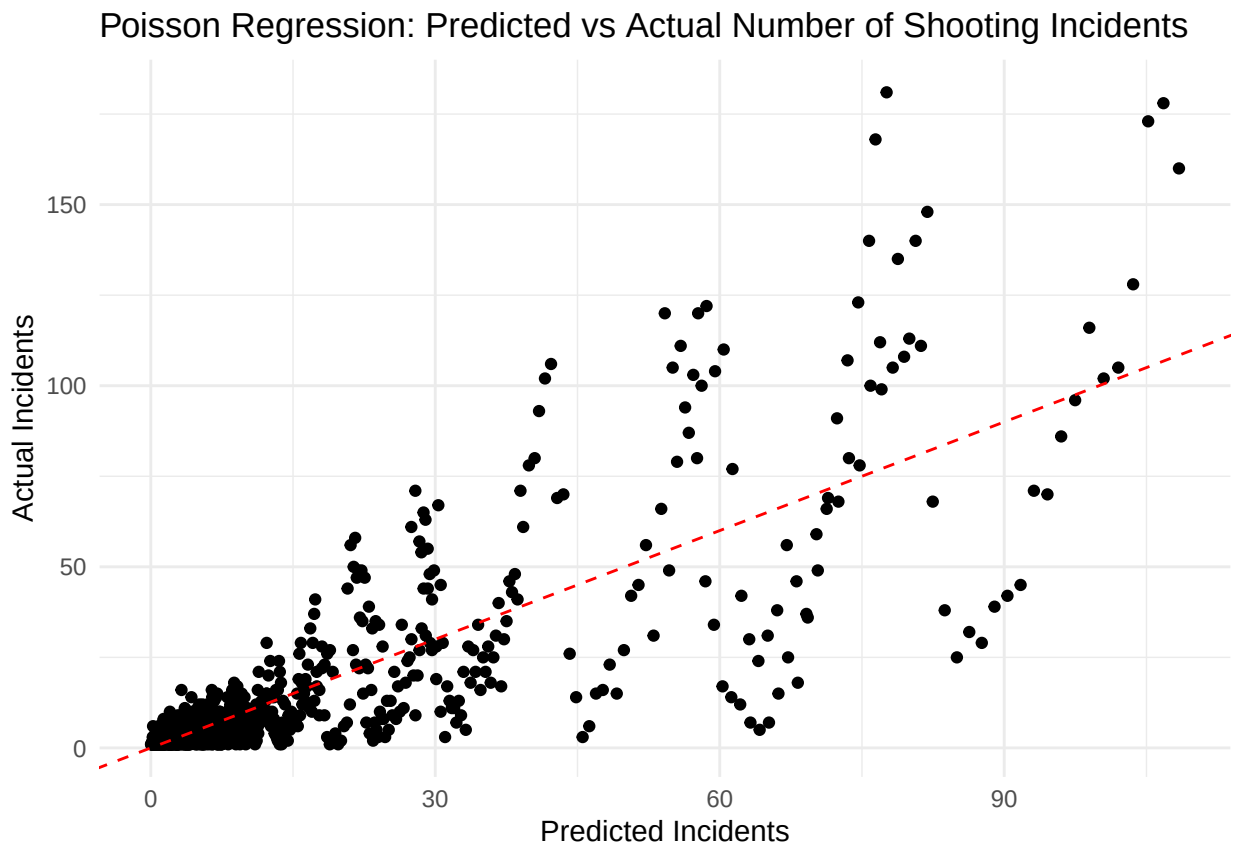
Time of Day: Shooting incidents tend to increase as the day progresses, with a higher likelihood later in the day.

Demographics: The age group 25-44 is most associated with shooting incidents, while younger (10-24) and older (65+) groups experience fewer incidents. Male victims are involved in fewer incidents than female victims.

Using key factors like borough, time of day, and demographics, we apply a Poisson regression model to predict the number of shooting incidents. This model helps us understand how well these variables explain the occurrence of incidents. The following visualization compares the model's predicted values to the actual incident counts.

```r
# Get the predicted values from the Poisson regression model
nypd_summary$Poisson_Predicted <- predict(poisson_model, type = "response")

# Create a scatter plot of actual vs predicted incidents for the Poisson model
ggplot(nypd_summary, aes(x = Poisson_Predicted, y = Incidents)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Poisson Regression: Predicted vs Actual Number of Shooting Incidents",
    x = "Predicted Incidents",
    y = "Actual Incidents"
  ) +
  theme_minimal()
```



The scatter plot compares predicted vs. actual shooting incidents. Points near the red line indicate accurate predictions. The model performs well for lower incident counts (0-30) but shows more variability and less accuracy as the number of incidents increases, suggesting room for improvement with higher counts.

Conclusion

This analysis reveals key patterns in shooting incidents across New York City:

Borough Distribution: Brooklyn and The Bronx have the highest number of shooting incidents, while Staten Island has the fewest. Efforts to reduce gun violence should focus on these higher-risk areas.

Time of Day: Shootings peak in the evening, especially in Brooklyn and The Bronx, highlighting the need for increased prevention during these hours.

Demographics: Individuals aged 25-44 are most affected, with fewer incidents involving younger and older groups. Female victims are slightly more frequent than male.

Model Insights: The Poisson regression model shows that borough, time of day, age group, and gender are key predictors of shootings. However, it struggles with higher incident counts, suggesting room for improvement.

Possible bias

Borough Focus: Since Brooklyn has many incidents, the model might focus too much on it, making predictions for smaller boroughs like Staten Island less accurate.

Missing Factors: The model doesn't include other important factors like income level or neighborhood safety, which could affect shooting rates but aren't part of the data.

Underestimating High Counts: The model may not handle very high incident numbers well, leading to predictions that underestimate areas or times with a lot of shootings.