NYPD Shooting Incident Analysis

Jesse Navarrette

10/2/2024

Summary of Data

For this project, I am working with the NYPD Shooting Incident dataset, which is publicly available here.

My analysis focuses on examining shooting incidents by borough and fitting a logistic regression model to explore factors influencing whether a shooting results in a fatality.

The goal is to provide insights into shooting trends across different boroughs and to identify potential biases in the data and the analysis.

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## Warning: package 'lubridate' was built under R version 4.1.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```r
library(lubridate)
# update.packages(ask = FALSE, checkBuilt = TRUE)
```

```r
# Load the NYPD Shooting dataset
nypd_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd <- read_csv(nypd_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.
```

```r
glimpse(nypd) # See the columns (and types) of the data.
```

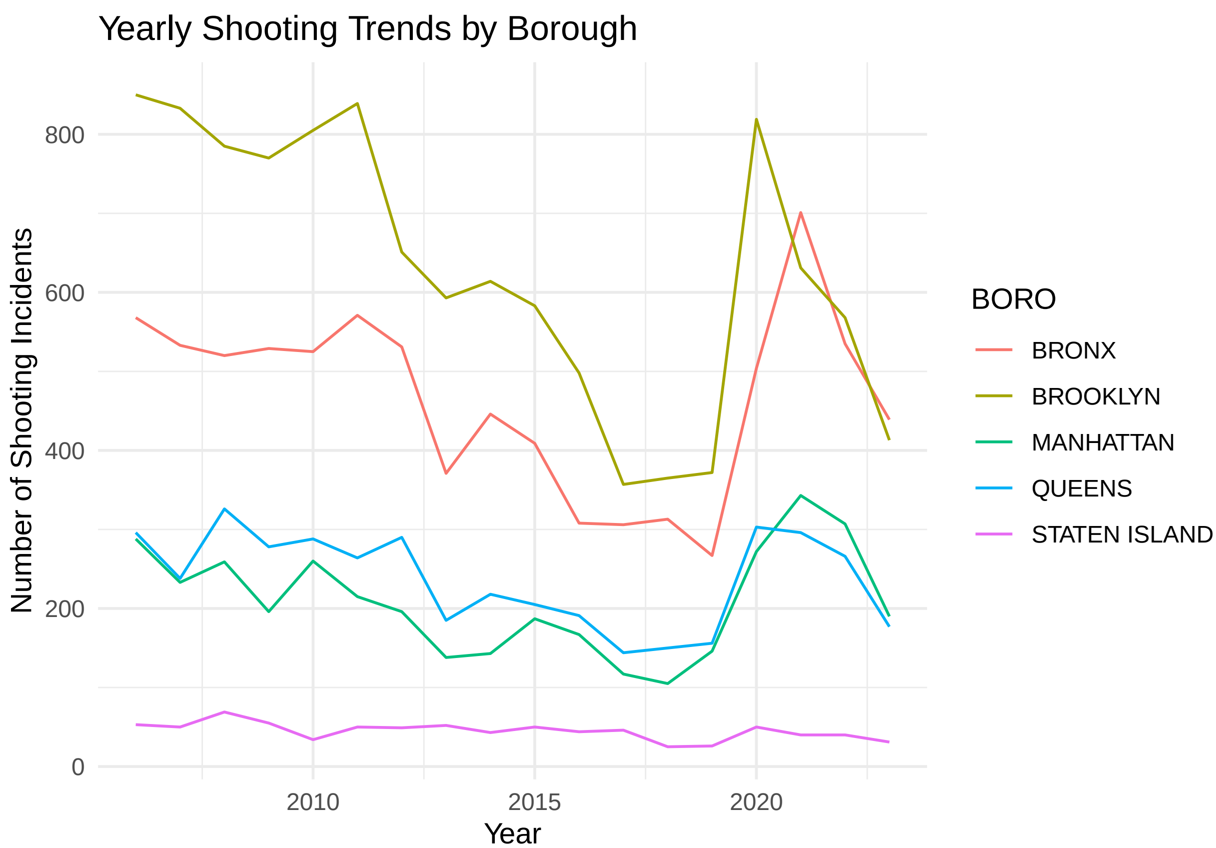
```
## Rows: 28,562
## Columns: 21
## $ INCIDENT_KEY            <dbl> 244608249, 247542571, 84967535, 202853370, 270~
## $ OCCUR_DATE              <chr> "05/05/2022", "07/04/2022", "05/27/2012", "09/~
## $ OCCUR_TIME              <time> 00:10:00, 22:20:00, 19:35:00, 21:00:00, 21:00~
## $ BORO                    <chr> "MANHATTAN", "BRONX", "QUEENS", "BRONX", "BROO~
## $ LOC_OF_OCCUR_DESC       <chr> "INSIDE", "OUTSIDE", NA, NA, NA, NA, NA, NA, N~
## $ PRECINCT                <dbl> 14, 48, 103, 42, 83, 23, 113, 77, 48, 49, 73, ~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ LOC_CLASSFCTN_DESC      <chr> "COMMERCIAL", "STREET", NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> "VIDEO STORE", "(null)", NA, NA, NA, "MULTI DW~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
## $ PERP_AGE_GROUP          <chr> "25-44", "(null)", NA, "25-44", "25-44", NA, N~
## $ PERP_SEX                <chr> "M", "(null)", NA, "M", "M", NA, NA, NA, NA, "~
## $ PERP_RACE               <chr> "BLACK", "(null)", NA, "UNKNOWN", "BLACK", NA,~
## $ VIC_AGE_GROUP           <chr> "25-44", "18-24", "18-24", "25-44", "25-44", "~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 986050, 1016802, 1048632, 1014493, 1009149, 99~
## $ Y_COORD_CD              <dbl> 214231.0, 250581.0, 198262.0, 242565.0, 190104~
## $ Latitude                <dbl> 40.75469, 40.85440, 40.71063, 40.83242, 40.688~
## $ Longitude               <dbl> -73.99350, -73.88233, -73.76777, -73.89071, -7~
## $ Lon_Lat                 <chr> "POINT (-73.9935 40.754692)", "POINT (-73.8823~
```

```r
# Tidying the data: remove irrelevant columns and handle missing values
nypd_clean <- nypd %>%
  select(OCCUR_DATE, BORO, PRECINCT, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_RACE, VIC_AGE_GROUP,
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  filter(!is.na(OCCUR_DATE), !is.na(BORO), !is.na(PRECINCT))
```

Step 2: Visualizing Shooting Trends by Borough

To understand trends over time, I will calculate the number of incidents by year for each borough and create a line plot.

```
nypd_by_year_boro <- nypd_clean %>%
  group_by(Year = year(OCCUR_DATE), BORO) %>%
  summarise(Incident_Count = n(), .groups = 'keep')

# Visualization: Shooting incidents by borough over time
ggplot(nypd_by_year_boro, aes(x = Year, y = Incident_Count, color = BORO)) +
  geom_line() +
  labs(title = "Yearly Shooting Trends by Borough",
       x = "Year",
       y = "Number of Shooting Incidents") +
  theme_minimal()
```

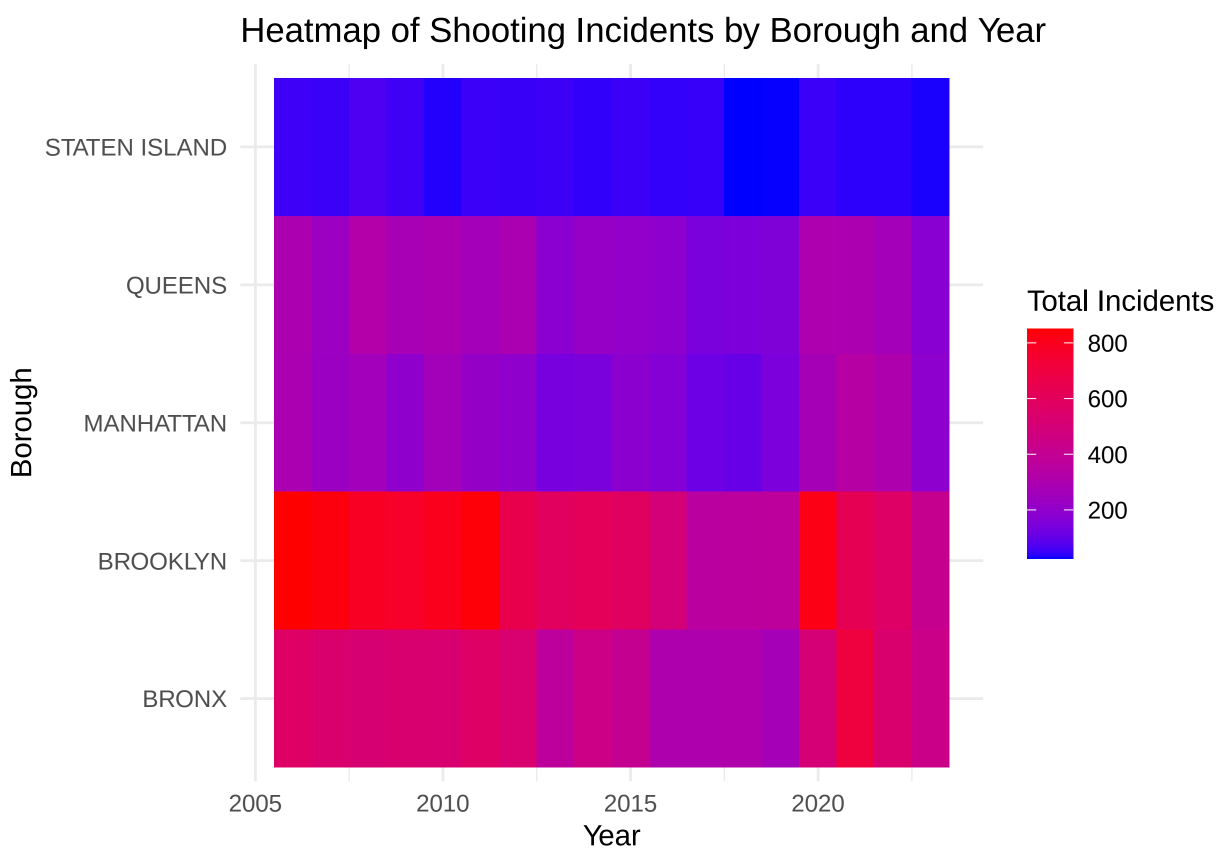



```
# This heatmap shows the distribution of shooting incidents across boroughs and over years, providing a

library(ggplot2)

heatmap_data <- nypd_by_year_boro %>%
  group_by(Year, BORO) %>%
  summarise(Total_Incidents = sum(Incident_Count))
```

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

```
## `.groups` argument.
```

```
ggplot(heatmap_data, aes(x = Year, y = BORO, fill = Total_Incidents)) +
  geom_tile() +  # Creates the heatmap tiles
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Heatmap of Shooting Incidents by Borough and Year",
       x = "Year",
       y = "Borough",
       fill = "Total Incidents") +
  theme_minimal()
```



Step 4: Logistic Regression Model

To explore the likelihood of a shooting resulting in a fatality, I will fit a logistic regression model. The dependent variable is whether the incident resulted in a murder, and the independent variables include borough and the perpetrator's age group.

```
# Convert the murder flag to binary
nypd_clean <- nypd_clean %>%
  mutate(is_murder = as.integer(STATISTICAL_MURDER_FLAG))

# Fit a logistic regression model
murder_model <- glm(is_murder ~ BORO + PERP_AGE_GROUP + PERP_RACE, data = nypd_clean, family = binomial)
```

```r
# Model summary
summary(murder_model)
```

```
##
## Call:
## glm(formula = is_murder ~ BORO + PERP_AGE_GROUP + PERP_RACE,
##     family = binomial, data = nypd_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1834  -0.7553  -0.6605  -0.2468   2.6516
##
## Coefficients: (1 not defined because of singularities)
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -1.68132    0.08830 -19.042  < 2e-16
## BOROBROOKLYN                             -0.10165    0.04639  -2.191 0.028448
## BOROMANHATTAN                            -0.14975    0.05944  -2.519 0.011759
## BOROQUEENS                               -0.13234    0.05878  -2.252 0.024348
## BOROSTATEN ISLAND                        -0.15218    0.10275  -1.481 0.138598
## PERP_AGE_GROUP<18                         0.34045    0.11419   2.982 0.002868
## PERP_AGE_GROUP1020                      -10.76824  324.74371  -0.033 0.973548
## PERP_AGE_GROUP1028                      -10.61607  324.74373  -0.033 0.973921
## PERP_AGE_GROUP18-24                       0.53591    0.09923   5.400 6.65e-08
## PERP_AGE_GROUP224                       -10.88474  324.74371  -0.034 0.973262
## PERP_AGE_GROUP25-44                       0.84158    0.09906   8.495  < 2e-16
## PERP_AGE_GROUP45-64                       1.20282    0.12342   9.746  < 2e-16
## PERP_AGE_GROUP65+                         1.25447    0.27576   4.549 5.39e-06
## PERP_AGE_GROUP940                       -10.78309  324.74371  -0.033 0.973511
## PERP_AGE_GROUPUNKNOWN                    -1.57646    0.15748 -10.011  < 2e-16
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -11.50875  229.36271  -0.050 0.959981
## PERP_RACEASIAN / PACIFIC ISLANDER         0.29404    0.17766   1.655 0.097912
## PERP_RACEBLACK                           -0.11650    0.05332  -2.185 0.028893
## PERP_RACEBLACK HISPANIC                  -0.22761    0.08245  -2.761 0.005769
## PERP_RACEUNKNOWN                          0.11011    0.14331   0.768 0.442311
## PERP_RACEWHITE                            0.44096    0.13370   3.298 0.000973
## PERP_RACEWHITE HISPANIC                        NA         NA      NA       NA
##
## (Intercept)                             ***
## BOROBROOKLYN                            *
## BOROMANHATTAN                           *
## BOROQUEENS                              *
## BOROSTATEN ISLAND
## PERP_AGE_GROUP<18                       **
## PERP_AGE_GROUP1020
## PERP_AGE_GROUP1028
## PERP_AGE_GROUP18-24                     ***
## PERP_AGE_GROUP224
## PERP_AGE_GROUP25-44                     ***
## PERP_AGE_GROUP45-64                     ***
## PERP_AGE_GROUP65+                       ***
## PERP_AGE_GROUP940
## PERP_AGE_GROUPUNKNOWN                   ***
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
```

```
## PERP_RACEASIAN / PACIFIC ISLANDER      .
## PERP_RACEBLACK                         *
## PERP_RACEBLACK HISPANIC                **
## PERP_RACEUNKNOWN
## PERP_RACEWHITE                         ***
## PERP_RACEWHITE HISPANIC
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19168  on 19217  degrees of freedom
## Residual deviance: 18071  on 19197  degrees of freedom
##   (9344 observations deleted due to missingness)
## AIC: 18113
##
## Number of Fisher Scoring iterations: 11
```

Step 4: Identifying Biases

Data Biases:

Missing or Incomplete Data: Many rows have missing information on the perpetrator's age, race, and sex. This could bias the model and analysis if certain demographics are underrepresented.

Location Bias: Incidents are categorized by borough, but the data does not include the specific neighborhoods within each borough. This could hide disparities at a more granular level.

Analytical Bias:

Preconceived Notions About Safety: I had an assumption that more densely populated areas like Brooklyn would have higher incidents. However, the data shows that Bronx often has higher incidents relative to other boroughs.

Modeling Assumptions: The logistic regression model assumes a linear relationship between the independent variables and the log-odds of an incident being a murder. However, this assumption may not hold true across all boroughs and demographic groups.

Conclusion

This project analyzed NYPD shooting incidents and explored factors that influence the likelihood of a shooting resulting in a fatality. By examining trends across boroughs, I identified differences in the frequency of incidents. Finally, I addressed potential biases in the data and my analysis, including missing demographic information and the use of borough-level data instead of more granular neighborhood data.