NYPD Shooting Incident Data

Jordan Pelletier

2025-02-18

Data Overview

The dataset we are working with is the NYPD Shooting dataset, which contains detailed information about each shooting incident, including details like the location, date, time, and various demographic details about the perpetrators and victims.

```
# Loading necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# Read the dataset from the provided URL
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
NYPD_ShootingDF <- read.csv(url, stringsAsFactors = FALSE)

# View the structure of the dataset
str(NYPD_ShootingDF)
```

```
## 'data.frame':    28562 obs. of  21 variables:
##  $ INCIDENT_KEY           : int  231974218 177934247 255028563 25384540 72616285 85875439 79780323 85
##  $ OCCUR_DATE             : chr  "08/09/2021" "04/07/2018" "12/02/2022" "11/19/2006" ...
```

```
##  $ OCCUR_TIME             : chr  "01:06:00" "19:48:00" "22:57:00" "01:50:00" ...
##  $ BORO                   : chr  "BRONX" "BROOKLYN" "BRONX" "BROOKLYN" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "" "" "OUTSIDE" "" ...
##  $ PRECINCT               : int  40 79 47 66 46 42 71 69 75 69 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 2 0 2 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr  "" "" "STREET" "" ...
##  $ LOCATION_DESC          : chr  "" "" "GROCERY/BODEGA" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "true" "false" "true" ...
##  $ PERP_AGE_GROUP         : chr  "" "25-44" "(null)" "UNKNOWN" ...
##  $ PERP_SEX               : chr  "" "M" "(null)" "U" ...
##  $ PERP_RACE              : chr  "" "WHITE HISPANIC" "(null)" "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "25-44" "25-44" "18-24" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  1006343 1000083 1020691 985107 1009854 ...
##  $ Y_COORD_CD             : num  234270 189065 257125 173350 247503 ...
##  $ Latitude               : num  40.8 40.7 40.9 40.6 40.8 ...
##  $ Longitude              : num  -73.9 -73.9 -73.9 -74 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.92019278899994 40.80967347200004)" "POINT (-73.94291302
```

```
# Summary of the dataset
summary(NYPD_ShootingDF)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class :character   Class :character
##  Median : 92711254   Mode  :character   Mode  :character   Mode  :character
##  Mean   :127405824
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Length:28562            Length:28562
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

```
# Convert OCCUR_DATE to Date
NYPD_ShootingDF$OCCUR_DATE <- as.Date(NYPD_ShootingDF$OCCUR_DATE, format = "%m/%d/%Y")

# Convert OCCUR_TIME to POSIXct
NYPD_ShootingDF$OCCUR_TIME <- as.POSIXct(NYPD_ShootingDF$OCCUR_TIME, format = "%H:%M:%S")

# Convert categorical columns to factors
categorical_cols <- c("BORO", "LOC_OF_OCCUR_DESC", "LOC_CLASSFCTN_DESC", "LOCATION_DESC",
                      "STATISTICAL_MURDER_FLAG", "PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE",
                      "VIC_AGE_GROUP", "VIC_SEX", "VIC_RACE")

NYPD_ShootingDF[categorical_cols] <- lapply(NYPD_ShootingDF[categorical_cols], as.factor)

# Handle Missing Values in Categorical Columns
for (col in categorical_cols) {
  NYPD_ShootingDF[[col]] <- ifelse(is.na(NYPD_ShootingDF[[col]]) | NYPD_ShootingDF[[col]] == "", "Unknou
}

# Handle Missing Values in Numeric Columns (Replace with Median)
numeric_cols <- c("X_COORD_CD", "Y_COORD_CD", "Latitude", "Longitude")
for (col in numeric_cols) {
  median_value <- median(NYPD_ShootingDF[[col]], na.rm = TRUE)
  NYPD_ShootingDF[[col]] <- ifelse(is.na(NYPD_ShootingDF[[col]]), median_value, NYPD_ShootingDF[[col]])
}

# Remove columns that are not necessary for analysis
columns_to_remove <- c("Lon_Lat")
NYPD_ShootingDF <- NYPD_ShootingDF[, !(names(NYPD_ShootingDF) %in% columns_to_remove)]

# Replace NA values in JURISDICTION_CODE with 0 to handle missing values
NYPD_ShootingDF$JURISDICTION_CODE[is.na(NYPD_ShootingDF$JURISDICTION_CODE)] <- 0

# summary of clean data
summary(NYPD_ShootingDF)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
```

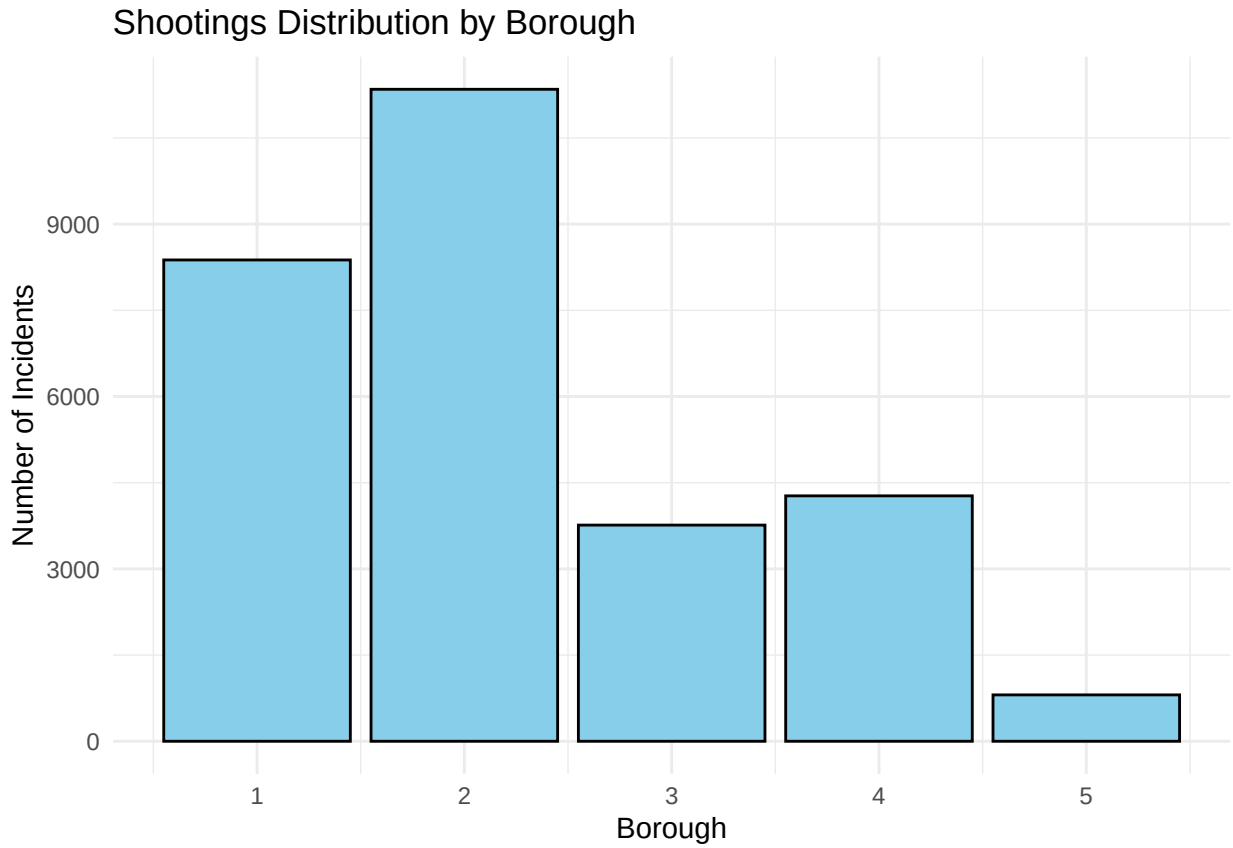
```
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :2025-02-18 00:00:00.00
##  1st Qu.: 65439914   1st Qu.:2009-09-04   1st Qu.:2025-02-18 03:30:00.00
##  Median : 92711254   Median :2013-09-20   Median :2025-02-18 15:15:00.00
##  Mean   :127405824   Mean   :2014-06-07   Mean   :2025-02-18 12:44:16.71
##  3rd Qu.:203131993   3rd Qu.:2019-09-29   3rd Qu.:2025-02-18 20:45:00.00
##  Max.   :279758069   Max.   :2023-12-29   Max.   :2025-02-18 23:59:00.00
##       BORO       LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE
##  Min.   :1.000   Length:28562       Min.   :  1.0   Min.   :0.0000
##  1st Qu.:1.000   Class :character   1st Qu.: 44.0   1st Qu.:0.0000
##  Median :2.000   Mode  :character   Median : 67.0   Median :0.0000
##  Mean   :2.222                      Mean   : 65.5   Mean   :0.3219
##  3rd Qu.:3.000                      3rd Qu.: 81.0   3rd Qu.:0.0000
##  Max.   :5.000                      Max.   :123.0   Max.   :2.0000
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:28562       Length:28562       Min.   :1.000
##  Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :1.000
##                                        Mean   :1.193
##                                        3rd Qu.:1.000
##                                        Max.   :2.000
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE          VIC_AGE_GROUP
##  Length:28562       Length:28562       Length:28562       Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:3.000
##  Mode  :character   Mode  :character   Mode  :character   Median :4.000
##                                                           Mean   :3.417
##                                                           3rd Qu.:4.000
##                                                           Max.   :7.000
##     VIC_SEX         VIC_RACE       X_COORD_CD        Y_COORD_CD
##  Min.   :1.000   Min.   :1.000   Min.   : 914928   Min.   :125757
##  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1000068   1st Qu.:182912
##  Median :2.000   Median :3.000   Median :1007772   Median :194901
##  Mean   :1.904   Mean   :3.763   Mean   :1009424   Mean   :208380
##  3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:1016807   3rd Qu.:239814
##  Max.   :3.000   Max.   :7.000   Max.   :1066815   Max.   :271128
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
```

Distribution of Shootings by Borough

```
# Load necessary libraries
library(ggplot2)

# Create a bar plot for the distribution of shootings by borough
ggplot(NYPD_ShootingDF, aes(x = BORO)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Shootings Distribution by Borough",
       x = "Borough",
```

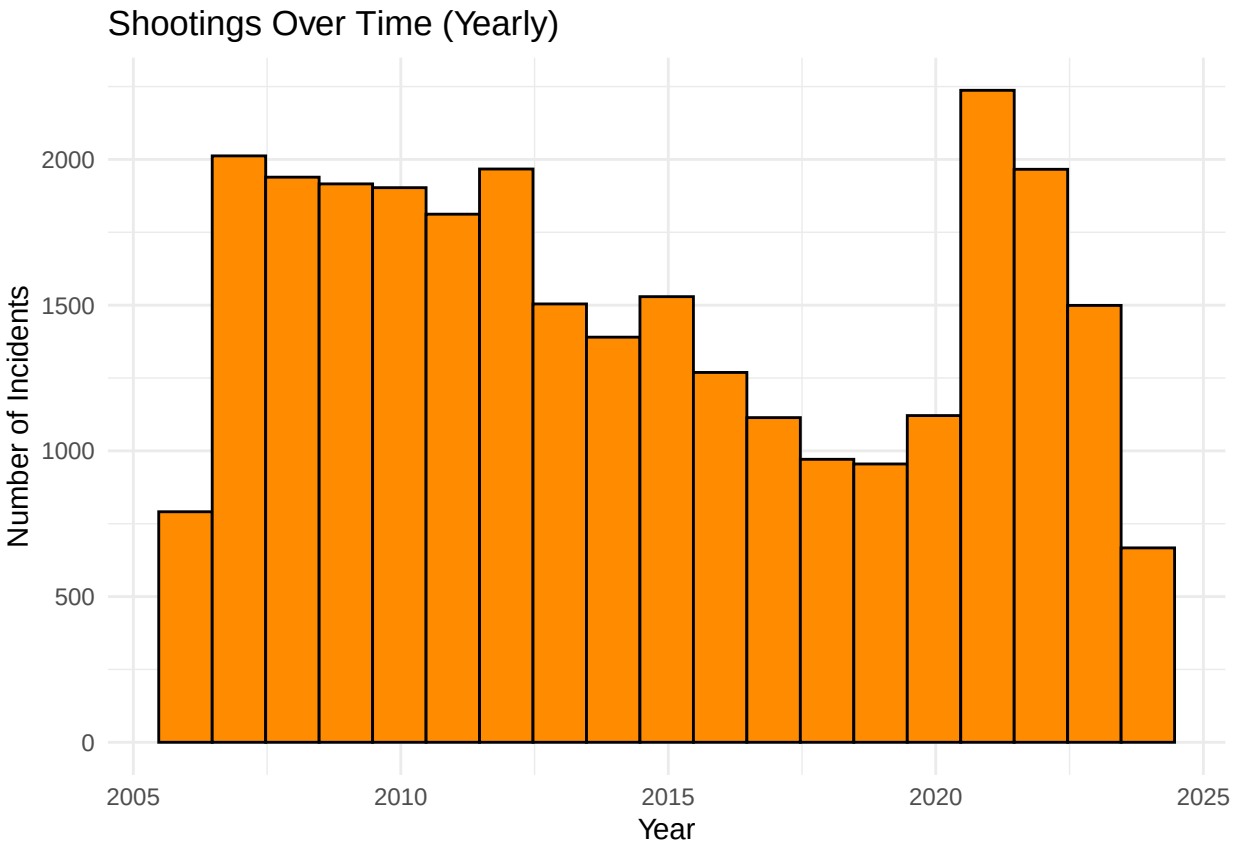
```
       y = "Number of Incidents") +
  theme_minimal()
```



Shootings by Year

```
# Convert OCCUR_DATE to Date format if not already
NYPD_ShootingDF$OCCUR_DATE <- as.Date(NYPD_ShootingDF$OCCUR_DATE)

# Create a time series plot for shooting incidents over time
ggplot(NYPD_ShootingDF, aes(x = OCCUR_DATE)) +
  geom_histogram(binwidth = 365, fill = "darkorange", color = "black") +
  labs(title = "Shootings Over Time (Yearly)",
       x = "Year",
       y = "Number of Incidents") +
  theme_minimal()
```



Conclusion

The project examines shooting incidents in New York City. Our analysis revealed patterns and trends including notable differences between boroughs. The visualization of these trends explored potential factors including the distribution of shootings such as location, age groups, and racial demographics.

The analysis raised further questions abaout the underlying factors contributing to these patterns. These could include population density, socio-economic factors, and perhaps law-enforcement practices that could potentially influence shooting incidents. A next step could be to include datsets with this information.

There are a few sources of bias that could affect our analysis such as Data Collection Bias, Geographical Bias, Data Imputation Bias, and Categorical Bias.

As far as personal bias, as a data analyst I acknowledge that personal bias can shape an approach to data cleaning, analysis, and result interpretation. My personal biases could arise from the framing of questions to focus on variables that seem most relevant to me. To mitigate bias, I must maintain objectivity, consider alternative perspectives, and ensure that this is reproducible.