Analysis of NYPD Historic Shooting Incidents

Raheleh

2025-06-06

#Introduction

This project explores shooting incidents data in New York City using data from NYC Open Data. The goal is to explore trends, patterns and characteristics of shooting incidents in NYC. The following are the steps:

1. **Importing Data**: Import the dataset into R.
2. **Cleaning Data**: Correct data, transform variables, and handle missing values.
3. **Data Visualization**: Create plots to explore distributions and relationships.
4. **Analysis and Modeling**: Execute statistical evaluations and build a predictive model.
5. **Bias Discussion**: Highlight potential biases in the data and methodology.
6. **Conclusion**: Summarize key findings and discuss limitations.

Data Import

Data Import

I imported the shooting dataset using read_csv() from the URL below, I also installed the necessary R packages including: tidyverse, lubridate, Janitor.

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shooting_data <- read_csv(url)
dim(shooting_data)
```

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

The dataset contains 29,744 rows and 21 columns with various information such as occurrence date, location, victim and perpetrator demographics.

First Few rows of raw data:

```
shooting_data_raw <- read_csv(url)
head(shooting_data_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

I cleaned the dataset by standardizing column names, converting variables to appropriate types, and removing unnecessary columns.

```r
# Clean column names
shooting_data <- shooting_data %>%
  janitor::clean_names()

# Convert date and time
shooting_data <- shooting_data %>%
  mutate(
    occur_date = mdy(occur_date),
    occur_time = parse_time(as.character(occur_time), "%H:%M:%S"),
    perp_age_group = as.factor(perp_age_group),
    vic_age_group = as.factor(vic_age_group),
    perp_sex = as.factor(perp_sex),
    vic_sex = as.factor(vic_sex),
    perp_race = as.factor(perp_race),
    vic_race = as.factor(vic_race),
    boro = as.factor(boro),
    jurisdiction_code = as.factor(jurisdiction_code)
  )

# Remove columns mostly empty or irrelevant for this analysis
shooting_data <- shooting_data %>%
  select(-incident_key, -loc_of_occur_desc, -loc_classfctn_desc, -location_desc)

# Summary of cleaned data
summary(shooting_data)
```

```
##    occur_date           occur_time                        boro
##  Min.   :2006-01-01   Min.   :00:00:00.000000   BRONX        : 8834
##  1st Qu.:2009-10-29   1st Qu.:03:30:45.000000   BROOKLYN     :11685
##  Median :2014-03-25   Median :15:15:00.000000   MANHATTAN    : 3977
##  Mean   :2014-10-31   Mean   :12:46:10.874798   QUEENS       : 4426
##  3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000   STATEN ISLAND:  822
##  Max.   :2024-12-31   Max.   :23:59:00.000000
##
##     precinct      jurisdiction_code statistical_murder_flag perp_age_group
##  Min.   :  1.00   0   :24957        Mode :logical           18-24  :6630
```

```
##  1st Qu.: 44.00   1    :  109        FALSE:23979             25-44  :6342
##  Median : 67.00   2    : 4676        TRUE :5765              UNKNOWN:3148
##  Mean   : 65.23   NA's:    2                                 <18    :1805
##  3rd Qu.: 81.00                                             (null) :1628
##  Max.   :123.00                                             (Other): 847
##                                                             NA's   :9344
##    perp_sex              perp_race     vic_age_group   vic_sex
##  (null): 1628   BLACK         :12323   <18    : 3081   F: 2891
##  F     :  461   WHITE HISPANIC: 2667   1022   :    1   M:26841
##  M     :16845   UNKNOWN       : 1838   18-24  :10677   U:   12
##  U     : 1500   (null)        : 1628   25-44  :13563
##  NA's  : 9310   BLACK HISPANIC: 1487   45-64  : 2118
##                 (Other)       :  491   65+    :  236
##                 NA's          : 9310   UNKNOWN:   68
##                             vic_race       x_coord_cd        y_coord_cd
##  AMERICAN INDIAN/ALASKAN NATIVE:   13   Min.   : 914928   Min.   :125757
##  ASIAN / PACIFIC ISLANDER      :  478   1st Qu.:1000094   1st Qu.:183042
##  BLACK                         :20999   Median :1007826   Median :195506
##  BLACK HISPANIC                : 2930   Mean   :1009442   Mean   :208722
##  UNKNOWN                       :   72   3rd Qu.:1016739   3rd Qu.:239980
##  WHITE                         :  741   Max.   :1066815   Max.   :271128
##  WHITE HISPANIC                : 4511
##     latitude       longitude        lon_lat
##  Min.   :40.51   Min.   :-74.25   Length:29744
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.91   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.83   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :97      NA's   :97
```

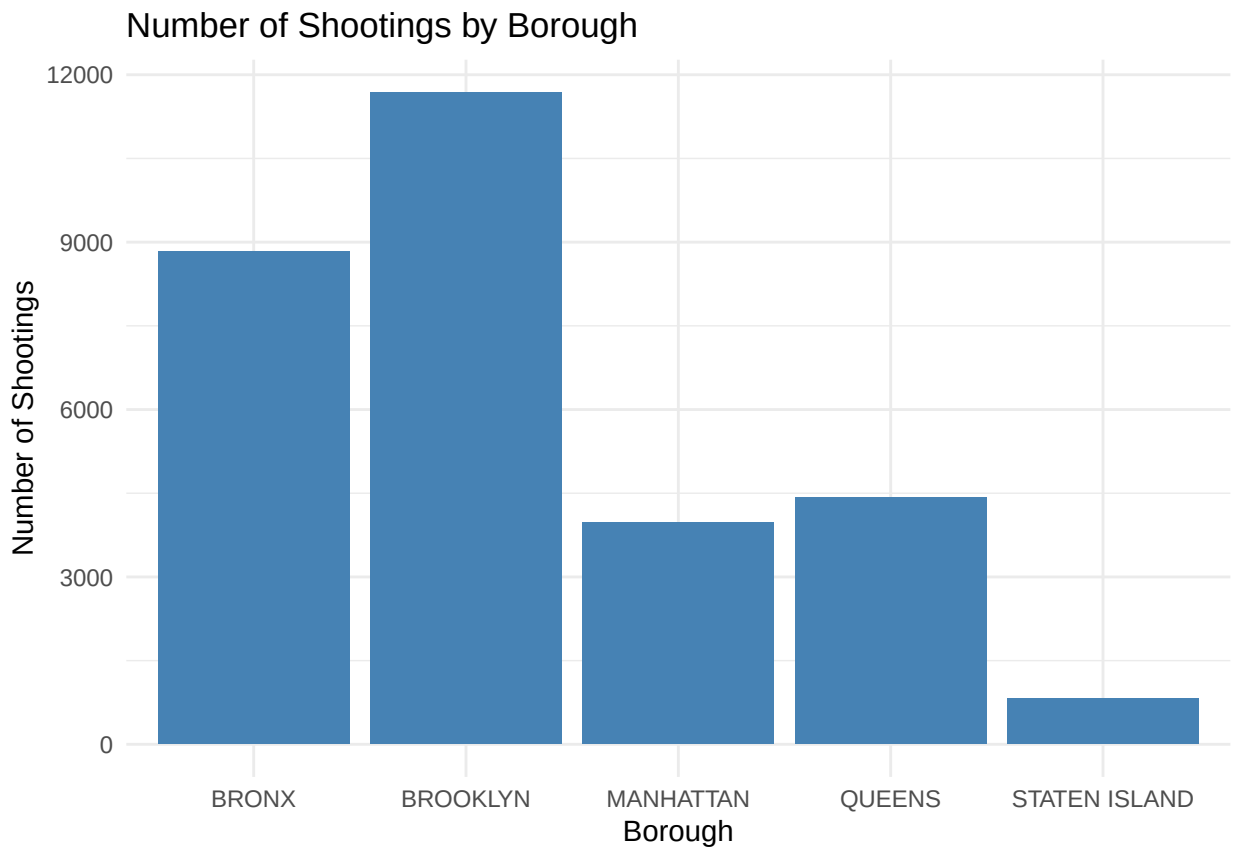
Some important fields, especially perpetrator demographics, contain many missing values. I replaced missing values in categorical variables with "Unknown" for clarity. I also remove columns that were mostly empty or not relevant to this analysis including: incident_key,loc_classfctn_desc, location_desc, loc_of_occur_desc.

Data Visualization

Number of shootings by borough

This bar chart shows the distribution of shooting incidents across NYC boroughs.

```
ggplot(shooting_data, aes(x = boro)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Number of Shootings by Borough", x = "Borough", y = "Number of Shootings") +
  theme_minimal()
```



Brooklyn has the highest number of shootings, while Staten Island has the lowest. This raises questions about demographic and social factors affecting crime rates.

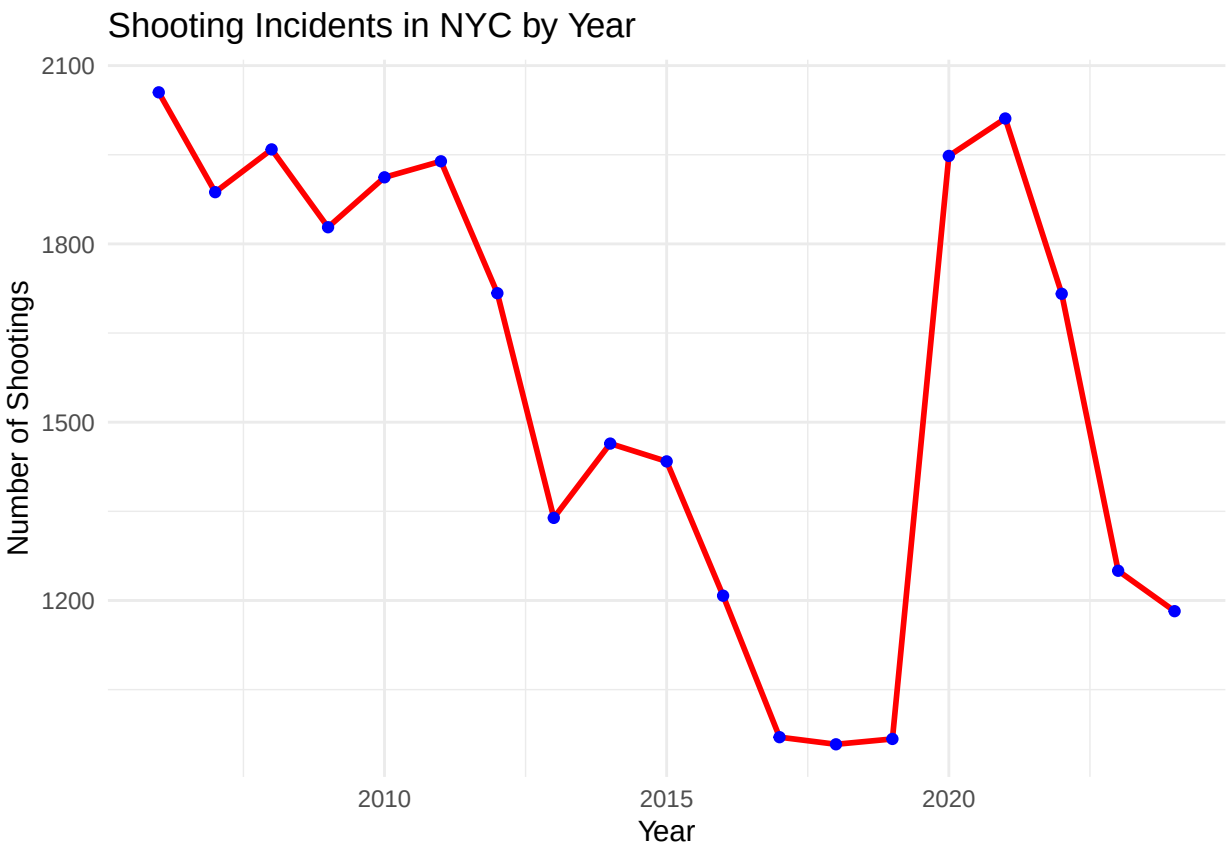
Shooting trends over time

This plot shows the trend of shooting incidents annually.

```
shooting_data <- shooting_data %>%
  mutate(year = year(occur_date))

yearly_shooting <- shooting_data %>%
  count(year)

ggplot(yearly_shooting, aes(x = year, y = n)) +
  geom_line(color = "red", linewidth = 1) +
  geom_point(color = "blue") +
  labs(title = "Shooting Incidents in NYC by Year", x = "Year", y = "Number of Shootings") +
  theme_minimal()
```



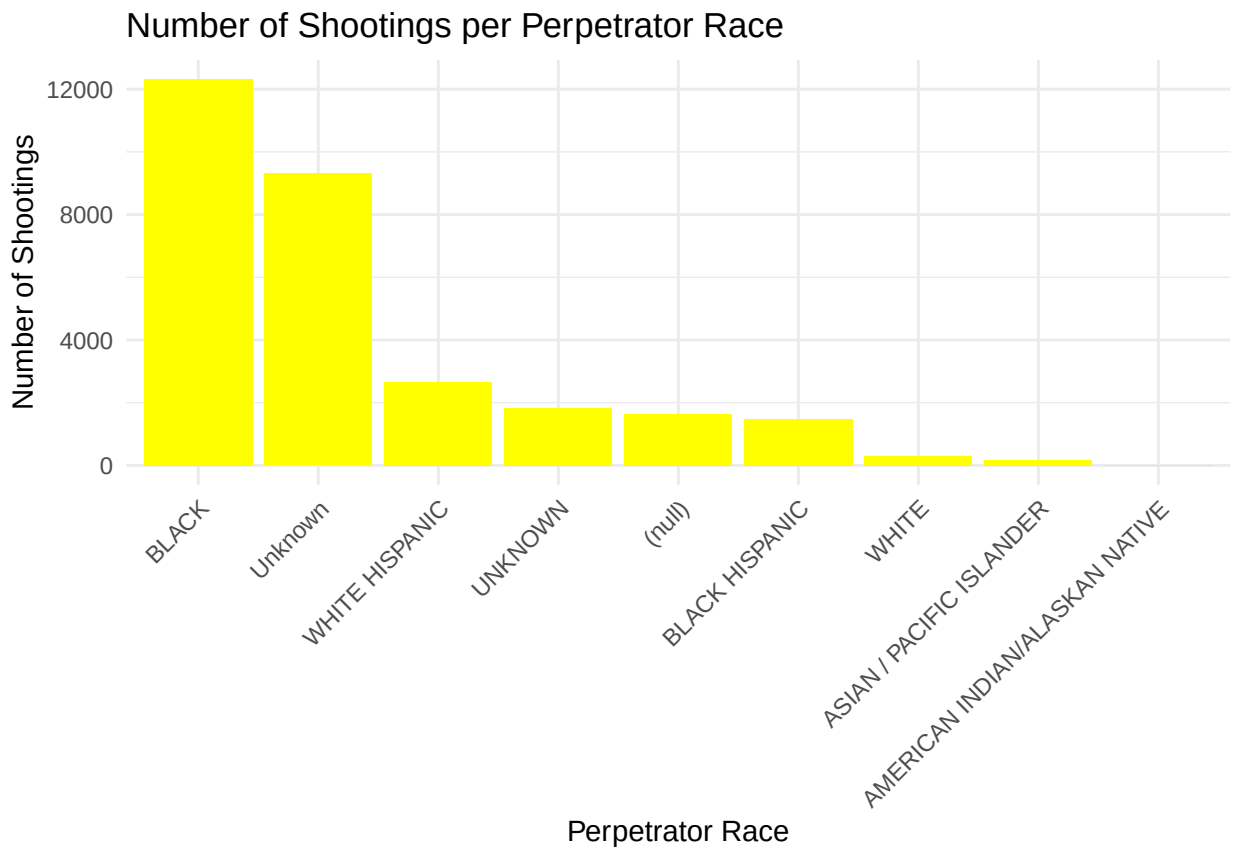
Shootings were at their lowest between 2015 and 2019 but rose significantly during the COVID-19 pandemic period (2019–2022). Further analysis is needed to understand this relationship.

Number of shootings by perpetrator race

This bar chart shows the distribution of shooting incidents among different races

```
race_count <- shooting_data %>%
  count(perp_race) %>%
  arrange(desc(n))

ggplot(race_count, aes(x = reorder(perp_race, -n), y = n)) +
  geom_bar(stat = "identity", fill = "yellow") +
  labs(title = "Number of Shootings per Perpetrator Race", x = "Perpetrator Race", y = "Number of Shooti
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The majority of known perpetrators are recorded as Black or Hispanic. This observation raises important questions about population proportions and other socioeconomic factors.

Analysis and Modeling

Predictive Modeling: Linear Regression of Shooting by Year

I fit a linear regression model to examine the trend of shootings over the years.

```
linear_model <- lm(n ~ year, data = yearly_shooting)
summary(linear_model)
```

```
##
## Call:
## lm(formula = n ~ year, data = yearly_shooting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523.42 -218.01   33.32  165.84  661.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74158.50   29035.88   2.554   0.0205 *
## year          -36.03      14.41  -2.500   0.0229 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 344 on 17 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2258
## F-statistic: 6.251 on 1 and 17 DF,  p-value: 0.02294
```

Interpretation:

-The coefficient for year is negative, indicating a decreasing trend in shootings over time.

- The p-value is 0.0229, which is less than 0.05, meaning this decreasing trend is statistically significant.

Bias Discussion

Sources of Bias

- **Surveillance Bia**s: Some neighborhoods may be policed more heavily, increasing reported incidents there.
- **Reporting Bias**: Racial and demographic data contain many missing values.
- **Structural/Systemic Bias**: Social and economic inequalities influence crime patterns.
- **Data Gaps**: Many unknown or missing values weaken the reliability of some conclusions.

Personal Bias and Mitigation

As a student analyst, I acknowledge that my interpretation may be influenced by personal and cultural perspectives. To mitigate this:

- I highlighted dataset limitations and biases.
- I used neutral language to avoid stigmatization.
- I interpreted demographic data cautiously without overgeneralization.

Conclusion

- Brooklyn had the highest number of shooting incidents, while Staten Island had the lowest.
- Shootings peaked during the COVID-19 pandemic (2019–2022).
- Among known cases, a disproportionately high number of reported perpetrators were categorized as Black or Hispanic.