NYPD Shooting Incident Analysis

Charles

2024-02-03

The NYPD Shooting Incident dataset provides a comprehensive overview of shooting incidents in New York City. This exploratory anlysis is designed to uncover the relationshiop between the time of day and borough location with the occurrence and severity of shootings. To faciliate my analyis I will employ a logistic regression to model the temporal (time of day) and spatial (borough) factors in assessing the probability of fatal shooting incidnetns. This approach will enhance our understanding of the dynamics influencing shootigns and their outcome across different times and areas within

New York City.

Load Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.4      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

```
library(dplyr)
library(ggplot2)
```

Import data

```
# used to read data
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read_csv(url)
```

Summarize data to provide conceptual understanding

```
summary_data <- summary(nypd_data)
summary_data
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms         Class :character
##  Median : 90372218   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
```

```
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                         NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

```
summary_data1 <- nypd_data %>%
  count(BORO,sort = TRUE)
summary_data1
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <chr>         <int>
## 1 BROOKLYN      10933
## 2 BRONX          7937
## 3 QUEENS         4094
## 4 MANHATTAN      3572
## 5 STATEN ISLAND   776
```

Looking at our summary data we see there's a few columns we need to tidy for analytic purposes

Tidy data

```
data <- nypd_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
  OCCUR_TIME = hms(OCCUR_TIME),
  YEAR = year(OCCUR_DATE),
  MONTH = month(OCCUR_DATE),
  MONTH_label = month(OCCUR_DATE, label = TRUE),
  HOUR = hour(OCCUR_TIME))

data <- data %>%
  dplyr::select(-LOC_OF_OCCUR_DESC, -LOC_CLASSFCTN_DESC) %>%
  mutate_if(is.character, ~replace(., is.na(.), "UNKNOWN")) %>%
  mutate(PRECINCT = as.factor(PRECINCT))
```

Analyze data

```
data %>%
  group_by(YEAR, BORO) %>%
  summarise(INCIDENTS = n_distinct(INCIDENT_KEY)) %>%
  ggplot(aes(x = YEAR, y = INCIDENTS, group = BORO, color = BORO)) +
  geom_line() +
  geom_point(size = 2, shape = 1) +
  geom_hline(aes(yintercept = mean(INCIDENTS)), color = "black", lty = "dashed") +
  scale_x_continuous(breaks = seq(2006, 2022, 2)) +
  theme_bw() +
  theme(
```

```
    axis.text.x = element_text(size = 10, color = 'black'),
    axis.text.y = element_text(size = 10, color = 'black')
  ) +
  labs(
    title = "New York City Shooting Incidents per Year by Borough",
    x = "Year",
    y = "Count of Shooting Incidents"
  )
```

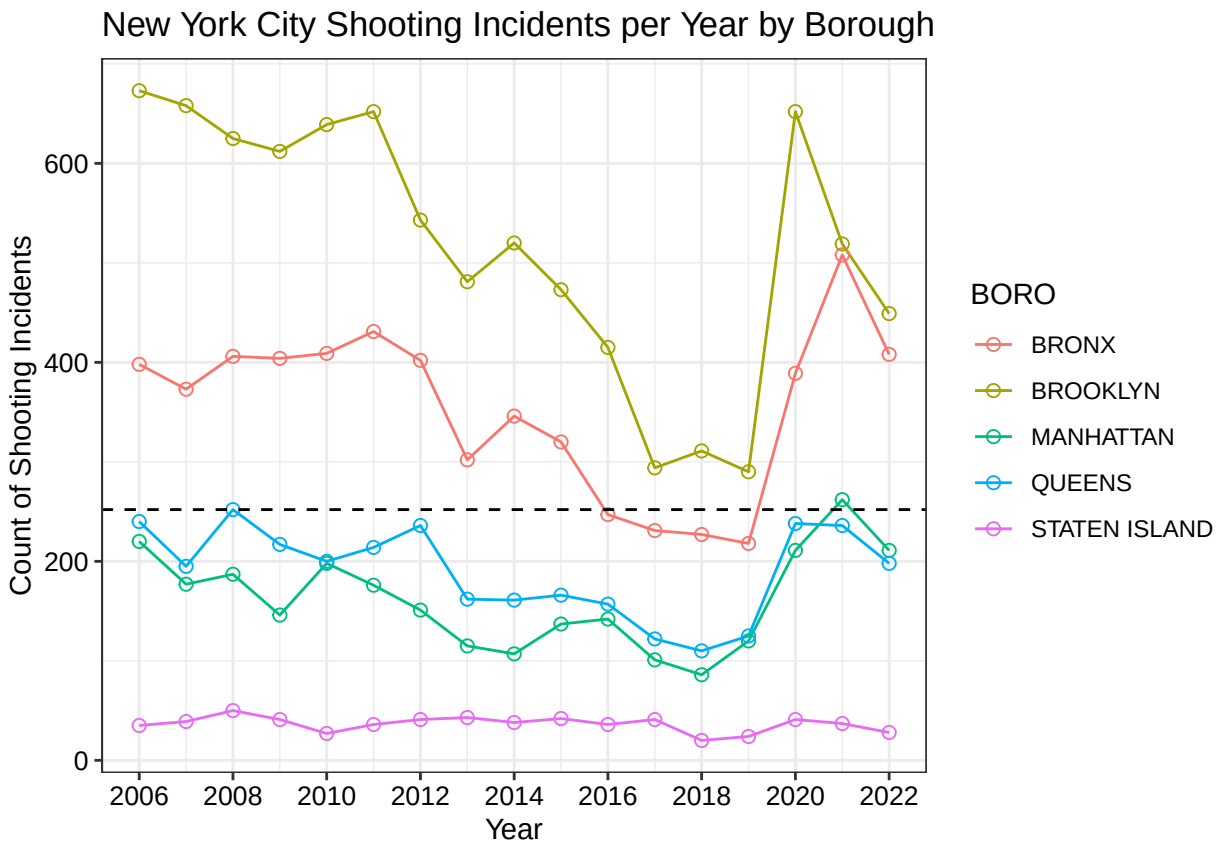


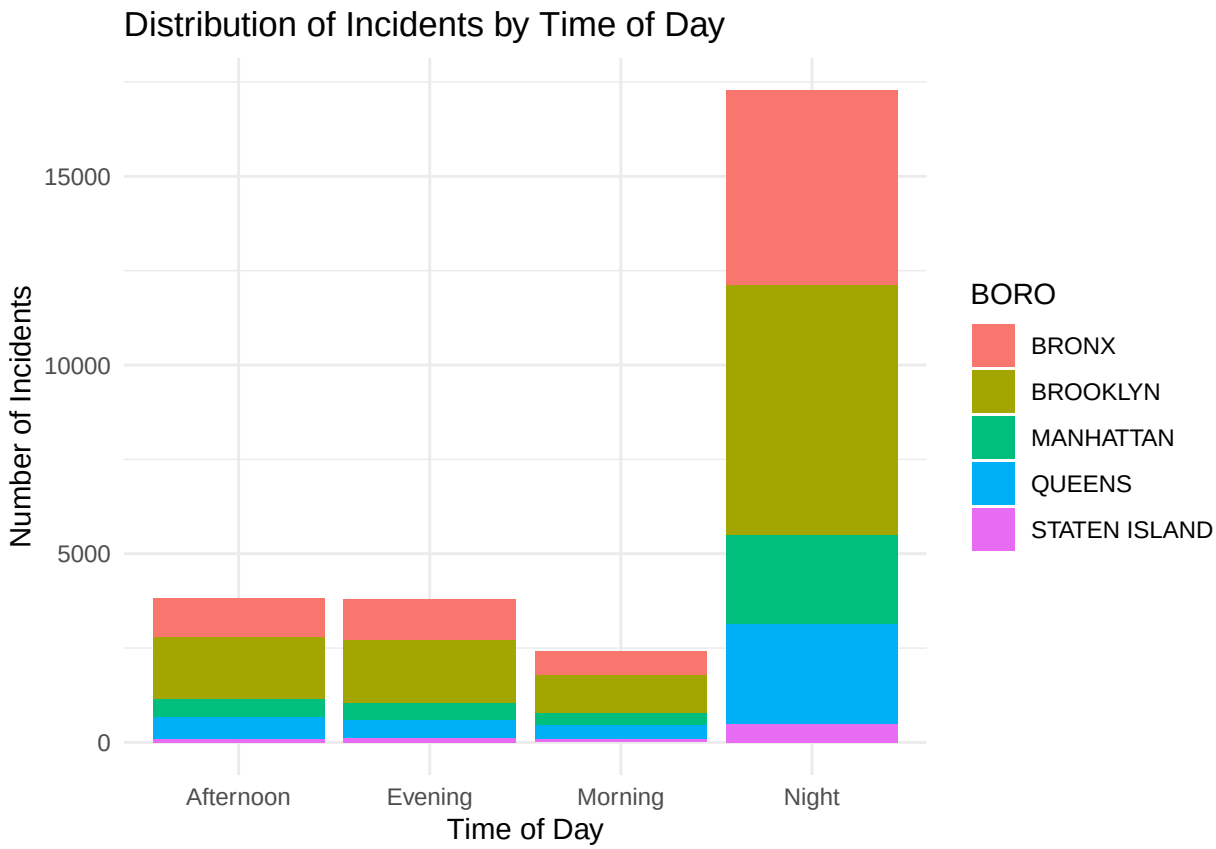
Figure 1 shows the temporal progression of shooting incidents recorded between 2006 and 2022. An analysis of the data reveals a consistent decline in the frequency of such incidents from 2006 until 2019. However, this descending trajectory underwent a reversal in 2019, marked by a notable upsurge in incidents, with the boroughs of the Bronx and Brooklyn experiencing the most significant escalations. The subsequent period, encompassing the years 2020 and 2021, was characterized by a precipitous decline in the frequency of shooting incidents. Notably, the incident frequency in Staten Island remained generally constant throughout the observed period from 2006 to 2022.

The next phase of our analysis will extend to include the time of day, thereby enriching our comprehension of the temporal and spatial dimensions in the distribution of incident frequency.

Time of day analysis

```
data1 <- data %>%
  mutate(
    TIME_CATEGORY = case_when(
      HOUR >= 5 & HOUR < 12  ~ "Morning",
      HOUR >= 12 & HOUR < 17 ~ "Afternoon",
      HOUR >= 17 & HOUR < 20 ~ "Evening",
```

```
      TRUE                    ~ "Night"
    )
  )
ggplot(data1, aes(x = TIME_CATEGORY, fill = BORO)) +
  geom_bar() +
  labs(
    title = "Distribution of Incidents by Time of Day",
    x = "Time of Day",
    y = "Number of Incidents"
  ) +
  theme_minimal()
```

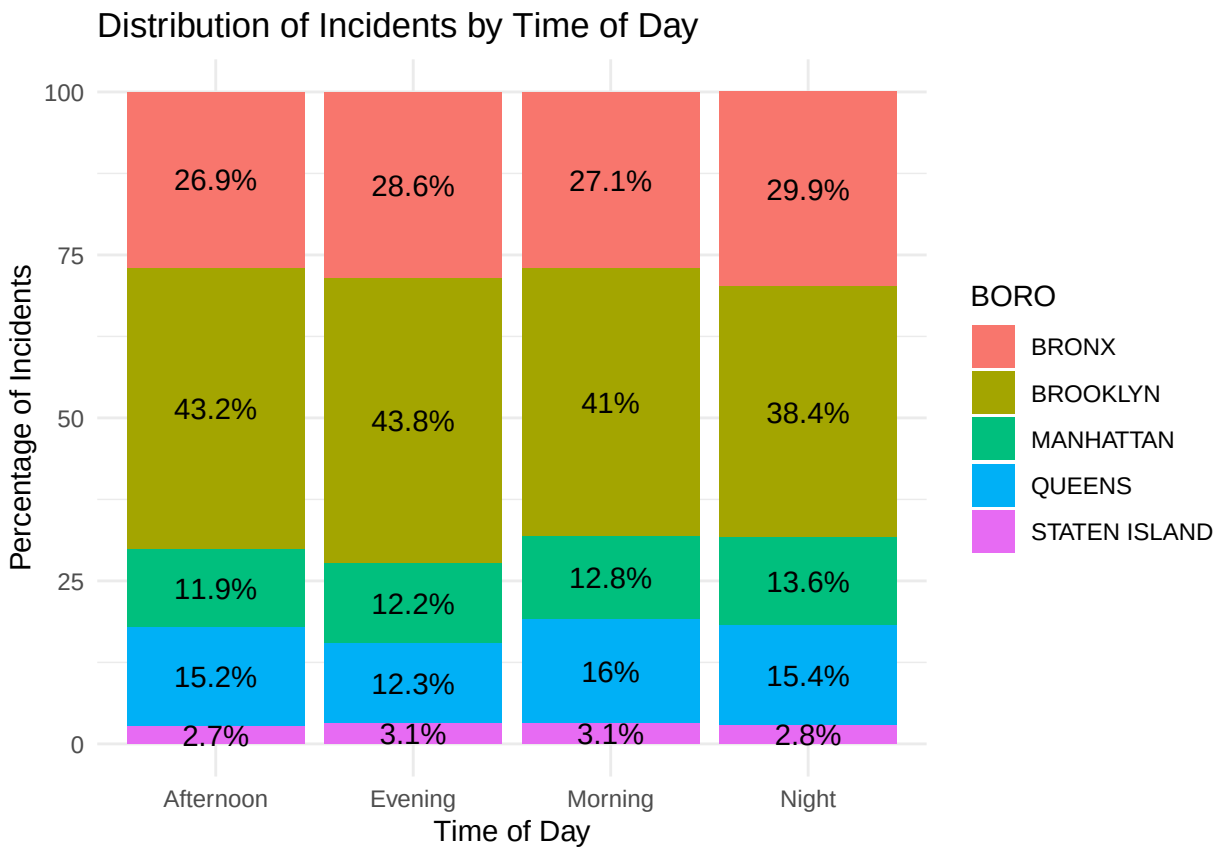


Next we'll use a stacked bar chart with percentages to enhance our understanding of the distribution of incidents by time of day.

```
# Calculate the percentages
data2 <- data1 %>%
  count(BORO, TIME_CATEGORY) %>%
  group_by(TIME_CATEGORY) %>%
  mutate(perc = n / sum(n) * 100)

# Create the stacked bar chart
ggplot(data2, aes(x = TIME_CATEGORY, y = perc, fill = BORO)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(perc, 1), "%")), position = position_stack(vjust = 0.5)) +
  labs(
    title = "Distribution of Incidents by Time of Day",
```

```
    x = "Time of Day",
    y = "Percentage of Incidents"
  ) +
  theme_minimal()
```



In our comprehensive analysis depicted in Figure 3, we quantitatively demonstrate that, across the evaluated time periods, Brooklyn and the Bronx collectively constitute approximately 70% of the total recorded shooting incidents. Meanwhile, Manhattan and Queens together account for roughly 27% of the incidents, with Staten Island comprising the remaining three percent.

```
summary(data1)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 63860880   1st Qu.:2009-07-18   1st Qu.:3H 27M 0S
##  Median : 90372218   Median :2013-04-29   Median :15H 11M 0S
##  Mean   :120860536   Mean   :2014-01-06   Mean   :12H 41M 31.7091388399567S
##  3rd Qu.:188810230   3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S
##  Max.   :261190187   Max.   :2022-12-31   Max.   :23H 59M 0S
##
##      BORO              PRECINCT     JURISDICTION_CODE LOCATION_DESC
##  Length:27312       75     : 1557   Min.   :0.0000    Length:27312
##  Class :character   73     : 1452   1st Qu.:0.0000    Class :character
##  Mode  :character   67     : 1216   Median :0.0000    Mode  :character
##                     44     : 1020   Mean   :0.3269
##                     79     : 1012   3rd Qu.:0.0000
```

```
##                    47     :  953   Max.   :2.0000
##                    (Other):20102   NA's   :2
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:27312       Length:27312
##  FALSE:22046             Class :character   Class :character
##  TRUE :5266              Mode  :character   Mode  :character
##
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    X_COORD_CD        Y_COORD_CD        Latitude       Longitude
##  Min.   : 914928   Min.   :125757   Min.   :40.51   Min.   :-74.25
##  1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67   1st Qu.:-73.94
##  Median :1007731   Median :194487   Median :40.70   Median :-73.92
##  Mean   :1009449   Mean   :208127   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :1066815   Max.   :271128   Max.   :40.91   Max.   :-73.70
##                                     NA's   :10      NA's   :10
##    Lon_Lat                YEAR          MONTH         MONTH_label
##  Length:27312       Min.   :2006   Min.   : 1.000   Jul    : 3238
##  Class :character   1st Qu.:2009   1st Qu.: 5.000   Aug    : 3156
##  Mode  :character   Median :2013   Median : 7.000   Jun    : 2829
##                     Mean   :2013   Mean   : 6.825   Sep    : 2572
##                     3rd Qu.:2018   3rd Qu.: 9.000   May    : 2571
##                     Max.   :2022   Max.   :12.000   Oct    : 2279
##                                                     (Other):10667
##       HOUR       TIME_CATEGORY
##  Min.   : 0.00   Length:27312
##  1st Qu.: 3.00   Class :character
##  Median :15.00   Mode  :character
##  Mean   :12.22
##  3rd Qu.:20.00
##  Max.   :23.00
##
```

```
data3 <- data2 %>%
  mutate(
    TIME_CATEGORY = as.factor(TIME_CATEGORY),
    BORO = as.factor(BORO)
  )


model <- glm(STATISTICAL_MURDER_FLAG ~ TIME_CATEGORY + BORO, data = data1, family = "binomial")


summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ TIME_CATEGORY + BORO,
##     family = "binomial", data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7906  -0.6508  -0.6388  -0.6050   1.8912
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.436504   0.048062 -29.888  < 2e-16 ***
## TIME_CATEGORYEvening  0.094035   0.057646   1.631   0.1028
## TIME_CATEGORYMorning  0.352418   0.062417   5.646 1.64e-08 ***
## TIME_CATEGORYNight   -0.049182   0.045764  -1.075   0.2825
## BOROBROOKLYN         -0.008036   0.037351  -0.215   0.8296
## BOROMANHATTAN        -0.119590   0.052331  -2.285   0.0223 *
## BOROQUEENS            0.020197   0.048492   0.416   0.6770
## BOROSTATEN ISLAND     0.081338   0.092905   0.875   0.3813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26781  on 27311  degrees of freedom
## Residual deviance: 26708  on 27304  degrees of freedom
## AIC: 26724
##
## Number of Fisher Scoring iterations: 4
```

In the logistic regression analysis, we identified two predictors with statistically significant associations with the outcome variable. The time category 'Morning' has a p-value substantially below 0.001, its level of significance, and a positive coefficient, 0.35, indicating a robust association with the incidence of murders. Additionally, the variable representing Manhattan exhibits a p-value below 0.05, and a negative coefficient, -0.119, denoting a significant but lesser likelihood of shootings being fatal compared to the reference boroughs. In contrast, the non-significant p-values for other time categories suggest no substantial deviation from the baseline in terms of their association with murder outcomes.

Bias 1. Upon initiating the analysis of the data, I became aware of a potential bias, particularly as it pertains to the predominance of minority groups among both perpetrators and victims. As a member of a minority community, this observation elicited slight discomfort and highlighted the potential risk that such biases pose in deterring comprehensive demographic analysis. To address and mitigate these biases it is essential to actively acknowledge their presence and engage with them through a process of reflection and adjustment. It is important for data professionals to recognize the existence of conscious and unconscious biases within ourselves and undertake measures to counteract their influence on our work. This commitment to bias mitigation is crucial to ensuring the integrity and objectivity of our analyses.