NYPD Shooting Incidents Analysis

Alfonso Gutierrez

2024-03-31

#Introduction

This report utilizes historical information from the New York City police department from 2006 to 2023. The goal of the report is to visualize trends over time across the five boroughs of the city. The database includes very useful information such as the time and coordinates of the incident, the race, age, and sex of both the victim and the perpetrator, and whether the incident resulted in an arrest or not. The report covers the cleaning of the information, visualization, and the creation of a model that helps us predict the number of fatalities according to the number of incidents at a given time.

Packages and libraries

Here are all the packages and libraries I used for this analysis.

```
if (!require("tidyverse")) install.packages("tidyverse")
```

```
## Loading required package: 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   3.5.0     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
```

```
if (!require("tinytex")) install.packages("tinytex")
```

```
## Loading required package: tinytex
```

```
if (!require("lubridate")) install.packages("lubridate")
if (!require("scales")) install.packages("scales")
```

```
## Loading required package: scales
##
## Attaching package: 'scales'
```

```
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(lubridate)
library(dplyr)
```

Data Preparation

Our primary data source comprises the City of New York's dataset on shootings spanning from 2006 to 2023.

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

```
## 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.
```

```
print(nypd_shootings, n = 15, width = Inf)
```

```
## # A tibble: 28,562 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
##  7    229224142 06/07/2021 19:55      QUEENS    <NA>                   113
##  8    231246224 07/22/2021 01:47      BROOKLYN  <NA>                    77
##  9    228559720 05/22/2021 18:39      BRONX     <NA>                    48
## 10    238210279 12/22/2021 23:17      BRONX     <NA>                    49
## 11    233431365 09/10/2021 22:30      BROOKLYN  <NA>                    73
## 12    238238212 12/23/2021 04:05      QUEENS    <NA>                   114
## 13    227089385 04/18/2021 21:56      QUEENS    <NA>                   114
```

```
## 14    236818010 11/22/2021 21:40      MANHATTAN <NA>                    28
## 15    230713944 07/10/2021 01:40      BRONX     <NA>                    43
##    JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC
##                <dbl> <chr>              <chr>
##  1                 0 COMMERCIAL         VIDEO STORE
##  2                 0 STREET             (null)
##  3                 0 <NA>               <NA>
##  4                 0 <NA>               <NA>
##  5                 0 <NA>               <NA>
##  6                 2 <NA>               MULTI DWELL - PUBLIC HOUS
##  7                 0 <NA>               <NA>
##  8                 0 <NA>               MULTI DWELL - APT BUILD
##  9                 0 <NA>               <NA>
## 10                 0 <NA>               <NA>
## 11                 0 <NA>               <NA>
## 12                 0 <NA>               BAR/NIGHT CLUB
## 13                 0 <NA>               <NA>
## 14                 0 <NA>               <NA>
## 15                 0 <NA>               <NA>
##    STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE      VIC_AGE_GROUP
##    <lgl>                   <chr>          <chr>    <chr>          <chr>
##  1 TRUE                    25-44          M        BLACK          25-44
##  2 TRUE                    (null)         (null)   (null)         18-24
##  3 FALSE                   <NA>           <NA>     <NA>           18-24
##  4 FALSE                   25-44          M        UNKNOWN        25-44
##  5 FALSE                   25-44          M        BLACK          25-44
##  6 FALSE                   <NA>           <NA>     <NA>           25-44
##  7 TRUE                    <NA>           <NA>     <NA>           45-64
##  8 FALSE                   <NA>           <NA>     <NA>           25-44
##  9 FALSE                   <NA>           <NA>     <NA>           18-24
## 10 TRUE                    25-44          M        WHITE HISPANIC 25-44
## 11 FALSE                   <NA>           <NA>     <NA>           25-44
## 12 TRUE                    <NA>           <NA>     <NA>           25-44
## 13 TRUE                    <NA>           <NA>     <NA>           18-24
## 14 FALSE                   18-24          M        BLACK          25-44
## 15 FALSE                   <NA>           <NA>     <NA>           18-24
##    VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
##    <chr>   <chr>         <dbl>      <dbl>    <dbl>     <dbl>
##  1 M       BLACK        986050     214231     40.8     -74.0
##  2 M       BLACK       1016802     250581     40.9     -73.9
##  3 M       BLACK       1048632     198262     40.7     -73.8
##  4 M       BLACK       1014493     242565     40.8     -73.9
##  5 M       BLACK       1009149.    190105.    40.7     -73.9
##  6 M       BLACK        999061     229912     40.8     -73.9
##  7 M       BLACK       1042534     184647     40.7     -73.8
##  8 M       BLACK       1004507     182865     40.7     -73.9
##  9 M       BLACK       1016391     249523     40.9     -73.9
## 10 M       BLACK       1021438     252793     40.9     -73.9
## 11 M       BLACK       1009715     185349     40.7     -73.9
## 12 M       BLACK       1009215     219725     40.8     -73.9
## 13 M       BLACK       1007938     217508     40.8     -73.9
## 14 M       BLACK        997742     232494     40.8     -74.0
## 15 M       BLACK       1025108     243208     40.8     -73.9
##    Lon_Lat
```

```
##    <chr>
##  1 POINT (-73.9935 40.754692)
##  2 POINT (-73.88233 40.854402)
##  3 POINT (-73.76777349199995 40.71063412500007)
##  4 POINT (-73.89071440599997 40.832416753000075)
##  5 POINT (-73.91021857399994 40.68844345900004)
##  6 POINT (-73.94650786199998 40.79772716600007)
##  7 POINT (-73.78988688199998 40.673306465000046)
##  8 POINT (-73.92697993199994 40.66858395700007)
##  9 POINT (-73.88382239499998 40.851507714000036)
## 10 POINT (-73.86556104899995 40.86046306400005)
## 11 POINT (-73.90819699999997 40.675388531000074)
## 12 POINT (-73.90987205399993 40.769743719000076)
## 13 POINT (-73.91449012799995 40.76366214500007)
## 14 POINT (-73.95126656699995 40.804816163000055)
## 15 POINT (-73.85235185799998 40.83413887300003)
## # i 28,547 more rows
```

Selecting relevant columns for the analysis

```
df_nypd <- select(nypd_shootings, INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO,
                  STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                  VIC_AGE_GROUP, VIC_SEX, VIC_RACE)

str(df_nypd)
```

```
## tibble [28,562 x 11] (S3: tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY           : num [1:28562] 2.45e+08 2.48e+08 8.50e+07 2.03e+08 2.71e+07 ...
##  $ OCCUR_DATE             : chr [1:28562] "05/05/2022" "07/04/2022" "05/27/2012" "09/24/2019" ...
##  $ OCCUR_TIME             : 'hms' num [1:28562] 00:10:00 22:20:00 19:35:00 21:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ BORO                   : chr [1:28562] "MANHATTAN" "BRONX" "QUEENS" "BRONX" ...
##  $ STATISTICAL_MURDER_FLAG: logi [1:28562] TRUE TRUE FALSE FALSE FALSE FALSE ...
##  $ PERP_AGE_GROUP         : chr [1:28562] "25-44" "(null)" NA "25-44" ...
##  $ PERP_SEX               : chr [1:28562] "M" "(null)" NA "M" ...
##  $ PERP_RACE              : chr [1:28562] "BLACK" "(null)" NA "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr [1:28562] "25-44" "18-24" "18-24" "25-44" ...
##  $ VIC_SEX                : chr [1:28562] "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr [1:28562] "BLACK" "BLACK" "BLACK" "BLACK" ...
```

```
#When checking types, I detected that OCCUR_DATE were saved as char
# Converting OCCUR_DATE from character to Date
df_nypd$OCCUR_DATE <- as.Date(df_nypd$OCCUR_DATE, format = "%m/%d/%Y")
#I'd like to know the last date we have
last_date <- max(df_nypd$OCCUR_DATE, na.rm = TRUE)
last_date
```

```
## [1] "2023-12-29"
```

```
#december 29th 2023
```

Creating a new column for perpetrator identification

```
df_nypd$PERP_ID <- ifelse(is.na(df_nypd$PERP_SEX), 0, 1)
```

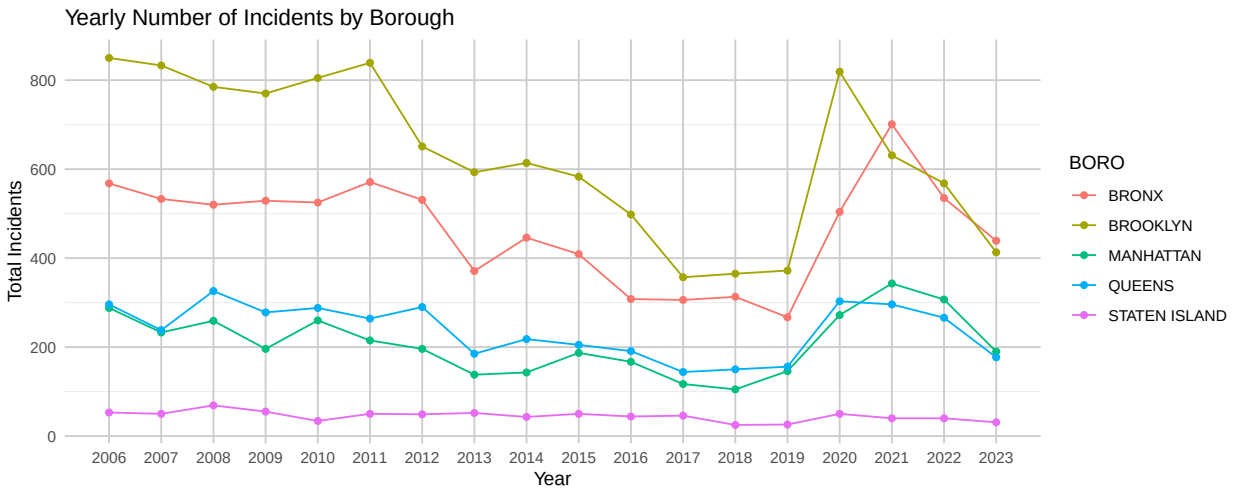
Summarizing the data by year and borough

```
df_nypd$OCCUR_YEAR <- year(as.Date(df_nypd$OCCUR_DATE, format = "%m/%d/%Y"))
yearly_incidents <- df_nypd %>%
  group_by(OCCUR_YEAR, BORO) %>%
  summarise(Total_Incidents = n(), .groups = 'drop')
```

Data Analysis/Visualization

Let's begin with a broad analysis and gradually narrow our focus, moving from the general to the specific.

```
ggplot(yearly_incidents, aes(x = OCCUR_YEAR, y = Total_Incidents, group = BORO, color = BORO)) +
  geom_line() +
  geom_point() +
  labs(title = "Yearly Number of Incidents by Borough", x = "Year", y = "Total Incidents") +
  theme_minimal() +
  theme(panel.grid.major.x = element_line(color = "grey80", linewidth = 0.5),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(color = "grey80", linewidth = 0.5)) +
  scale_x_continuous(breaks = unique(yearly_incidents$OCCUR_YEAR))
```

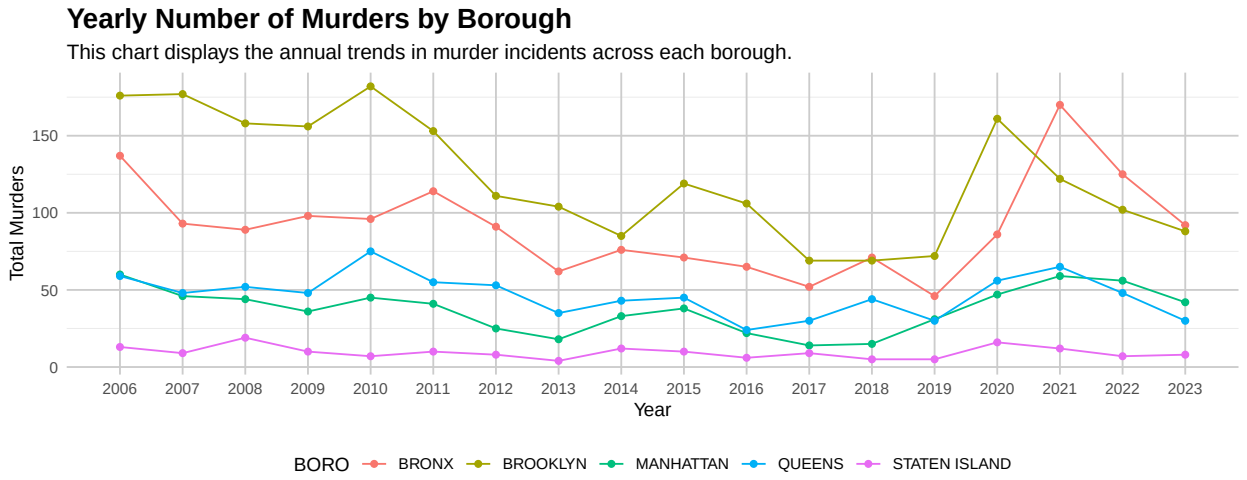


In this first graph, we can see a slight downward trend from 2006 to 2020, where the number of incidents surged, possibly due to COVID.Let's group by year and borough now and summarize total murders

```
murders_only <- df_nypd %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE)
yearly_murders <- murders_only %>%
  group_by(OCCUR_YEAR, BORO) %>%
  summarise(Total_Murders = n(), .groups = 'drop')
```

Create the line chart for murders

```r
ggplot(yearly_murders, aes(x = OCCUR_YEAR, y = Total_Murders, group = BORO, color = BORO)) +
  geom_line() +  # Connect points with lines
  geom_point() +  # Show individual points
  labs(title = "Yearly Number of Murders by Borough",
       subtitle = "This chart displays the annual trends in murder incidents across each borough.",
       x = "Year",
       y = "Total Murders") +
  theme_minimal() +
  theme(panel.grid.major.x = element_line(color = "grey80", linewidth = 0.5),
        panel.grid.minor.x = element_blank(),  # Remove the minor grid lines
        panel.grid.major.y = element_line(color = "grey80", linewidth = 0.5),
        legend.position = "bottom",  # Adjust legend positioning for better visibility
        plot.title = element_text(size = 16, face = "bold"),
        plot.subtitle = element_text(size = 12)) +
  scale_x_continuous(breaks = unique(yearly_murders$OCCUR_YEAR))  # One vertical line per year
```



Both the number of incidents and the number of murders share a similar trend, with a decline from 2006 to 2019, a surge in 2020, followed by a subsequent decrease, although still not returning to pre-pandemic levels.

Given that the boroughs of the city have different populations, it's important to include this variable. Due to the current trend, this analysis will only consider the years 2020 to 2023, and we will use the latest estimate from the US Government Census of 2022 to conduct an analysis per 100,000 inhabitants.

```r
#create population dataset
borough_data <- data.frame(
  BORO = c("MANHATTAN", "BROOKLYN", "QUEENS", "BRONX", "STATEN ISLAND"),
  population = c(1597451, 2561225, 2252196, 1356476, 490687))

#create a subset containing the years of interest
df_2020_2023<- df_nypd %>%
  filter(year(OCCUR_DATE) %in% c(2020, 2021, 2022, 2023))

#join both datasets
df_last_years <- df_2020_2023 %>%
  left_join(borough_data, by = c("BORO"))

# Count incidents by borough and year
```

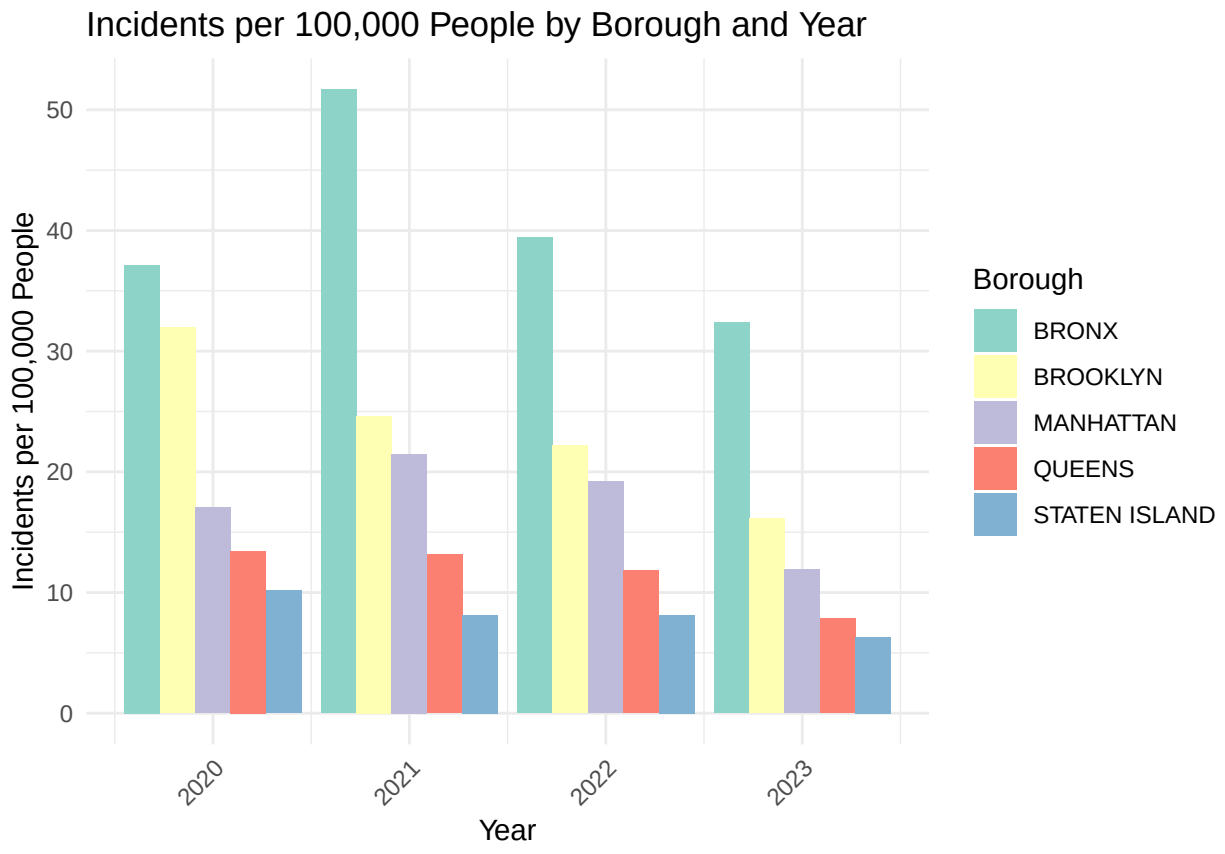
```r
incidents_per_borough_year <- df_last_years %>%
  group_by(BORO, year = year(OCCUR_DATE)) %>%
  summarise(incidents = n(), .groups = 'drop')

# Calculate incidents per 100,000 people
incidents_per_borough_year <- merge(incidents_per_borough_year, borough_data, by="BORO")
incidents_per_borough_year$incidents_per_100k <- (incidents_per_borough_year$incidents / incidents_per_

# Calculate the yearly average across all boroughs
yearly_avg <- incidents_per_borough_year %>%
  group_by(year) %>%
  summarise(nyc_avg_incidents_per_100k = mean(incidents_per_100k, na.rm = TRUE))

# Merge this average back with the original data frame
incidents_per_borough_year <- merge(incidents_per_borough_year, yearly_avg, by = "year")

ggplot(incidents_per_borough_year, aes(x = year, y = incidents_per_100k, fill = BORO)) +
  geom_col(position = "dodge") +
  labs(title = "Incidents per 100,000 People by Borough and Year",
       x = "Year",
       y = "Incidents per 100,000 People",
       fill = "Borough") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



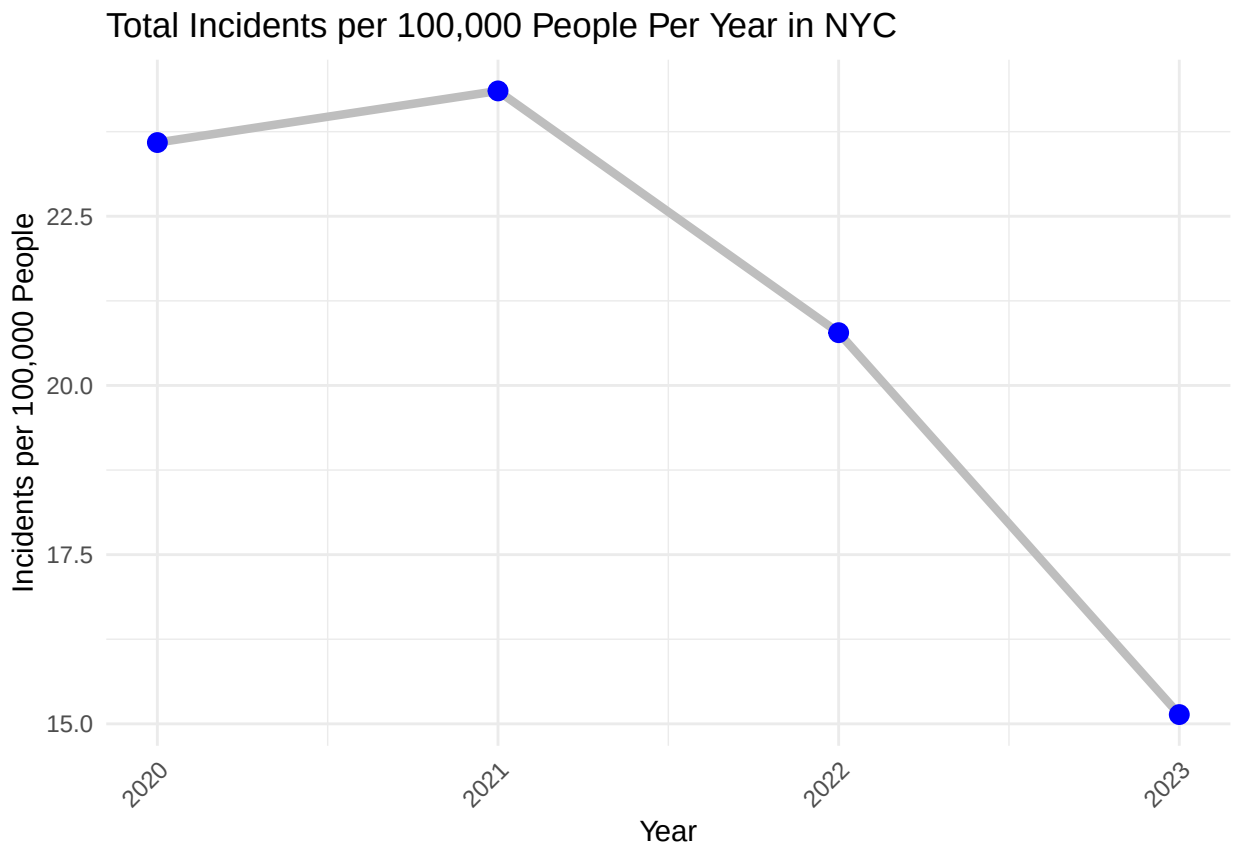
We can glean intriguing insights regarding incidents per borough. However, how does this fare compared to the city's overall situation? To contextualize this, let's compute the average per 100,000 inhabitants for each year.

```
total_incidents_per_year <- df_last_years %>%
  group_by(year = year(OCCUR_DATE)) %>%
  summarise(total_incidents = n(), .groups = 'drop')

total_nyc_population <- sum(borough_data$population)
total_incidents_per_year$incidents_per_100k <- (total_incidents_per_year$total_incidents / total_nyc_po

ggplot(total_incidents_per_year, aes(x = year, y = incidents_per_100k)) +
  geom_line(group = 1, color = "grey", size = 1.5) +
  geom_point(color = "blue", size = 3) +
  labs(title = "Total Incidents per 100,000 People Per Year in NYC",
       x = "Year",
       y = "Incidents per 100,000 People") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = comma)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



```r
# Create a summary table for each borough with the total number of incidents and the percentage resulti
borough_summary <- df_last_years %>%
  group_by(BORO) %>%
  summarise(
    total_incidents = n(),
    murders = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE),
    murder_rate = (murders / total_incidents) * 100,
    .groups = 'drop'  # this removes the automatic grouping that summarise() creates
  )

# Print the summary table
print(borough_summary)
```

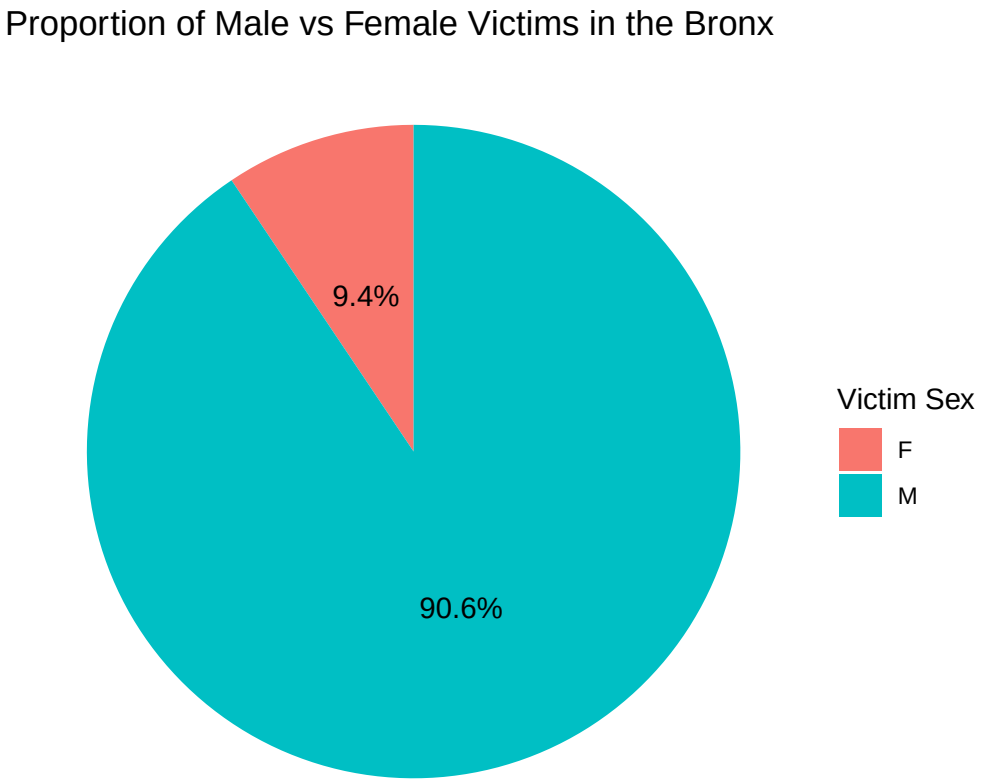
```
## # A tibble: 5 x 4
##   BORO          total_incidents murders murder_rate
##   <chr>                   <int>   <int>       <dbl>
## 1 BRONX                    2179     473        21.7
## 2 BROOKLYN                 2431     473        19.5
## 3 MANHATTAN                1112     204        18.3
## 4 QUEENS                   1042     199        19.1
## 5 STATEN ISLAND             161      43        26.7
```

Analyzing the charts, we observe a decrease in incidents per 100k people across the city. However, a significant concern persists in the Bronx, where last year's incidents per 100k people were more than double the city average. It's crucial to delve deeper into the data specific to the Bronx for a comprehensive analysis. When talking about the murder rate, the Bronx is the second place.

```r
# Summarize the data for victim sex
bronx_data <- df_last_years %>%
  filter(BORO == "BRONX")
bronx_victim_sex <- bronx_data %>%
  group_by(VIC_SEX) %>%
  summarise(count = n(), .groups = 'drop') %>%
  filter(VIC_SEX %in% c("M", "F"))  # Assuming the values are "M" and "F"

# Calculate percentages
bronx_victim_sex <- bronx_victim_sex %>%
  mutate(percentage = count / sum(count) * 100)

# Create a pie chart for victim sex with percentages
ggplot(bronx_victim_sex, aes(x = "", y = count, fill = VIC_SEX)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), position = position_stack(vjust = 0.5)) +
  labs(fill = "Victim Sex", title = "Proportion of Male vs Female Victims in the Bronx") +
  theme_void()
```




```
# Group by victim sex and calculate total incidents and murders
bronx_victim_sex <- bronx_data %>%
  group_by(VIC_SEX) %>%
  summarise(
    total_incidents = n(),
    murders = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE),
    murder_rate = (murders / total_incidents) * 100,
    .groups = 'drop'
  ) %>%
  filter(VIC_SEX %in% c("M", "F"))

# Print the table
print(bronx_victim_sex)
```

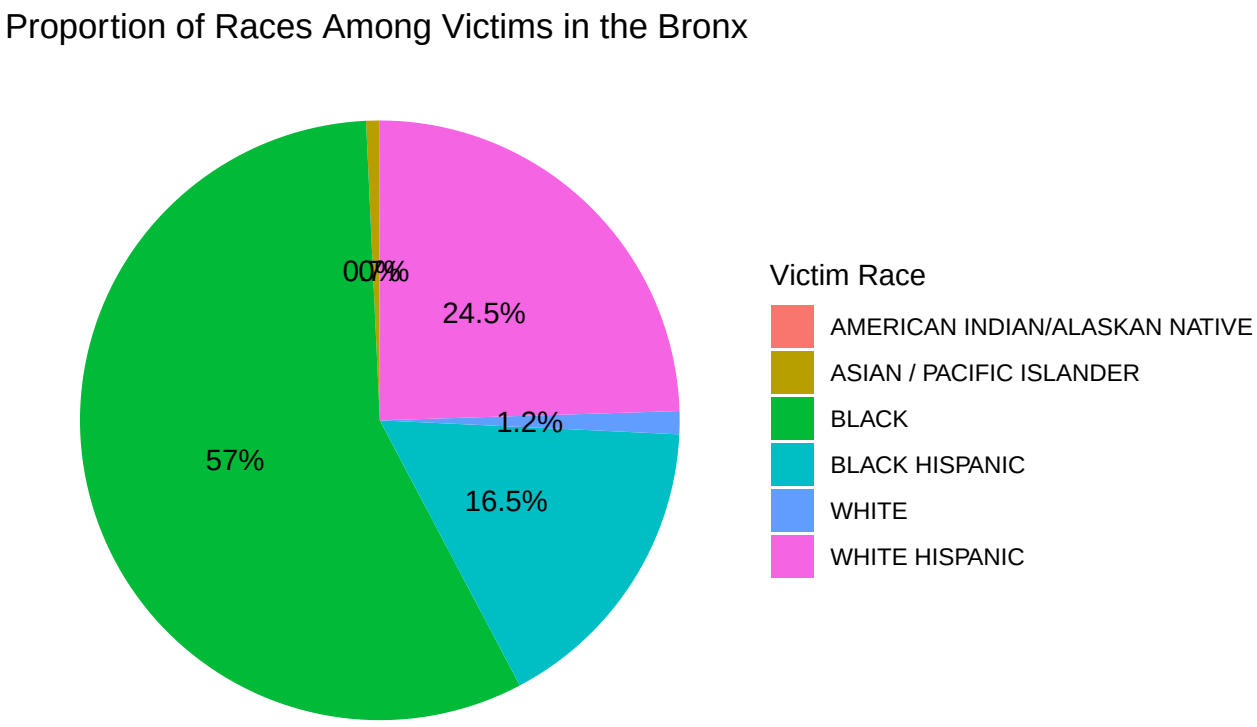
```
## # A tibble: 2 x 4
##   VIC_SEX total_incidents murders murder_rate
##   <chr>             <int>   <int>       <dbl>
## 1 F                   205      37        18.0
## 2 M                  1974     436        22.1
```

Ninety point six percent of the victims were men. It can also be seen that the murder rate in the case of men was 22%, compared to 18% in the case of women. With the information available and the analysis conducted so far, there is no evidence to suggest that there are hate crimes targeting any particular gender.

```r
# Summarize the data for victim race
bronx_victim_race <- bronx_data %>%
  group_by(VIC_RACE) %>%
  summarise(count = n(), .groups = 'drop')

# Calculate percentages
bronx_victim_race <- bronx_victim_race %>%
  mutate(percentage = count / sum(count) * 100)

# Create a pie chart for victim race with percentages
ggplot(bronx_victim_race, aes(x = "", y = count, fill = VIC_RACE)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), position = position_stack(vjust = 0.5)) +
  labs(fill = "Victim Race", title = "Proportion of Races Among Victims in the Bronx") +
  theme_void()
```




```r
# Summarize the data for victim race
bronx_victim_race <- bronx_data %>%
  group_by(VIC_RACE) %>%
  summarise(
    total_incidents = n(),
    murders = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE),
    murder_rate = (murders / total_incidents) * 100,
    .groups = 'drop'
```

```
  )

# Print the table
print(bronx_victim_race)
```

```
## # A tibble: 6 x 4
##   VIC_RACE                       total_incidents murders murder_rate
##   <chr>                                    <int>   <int>       <dbl>
## 1 AMERICAN INDIAN/ALASKAN NATIVE               1       0         0
## 2 ASIAN / PACIFIC ISLANDER                    15       2        13.3
## 3 BLACK                                     1242     278        22.4
## 4 BLACK HISPANIC                             360      67        18.6
## 5 WHITE                                       27       5        18.5
## 6 WHITE HISPANIC                             534     121        22.7
```

In the case of the proportion of incidents by race, we can see that both the percentage and the murder rate coincide, being higher in two groups: Black and White Hispanic, which together account for over 80% of the incidents.

Model Fitting

Fitting a model: Total Deaths as a function of Total Incidents

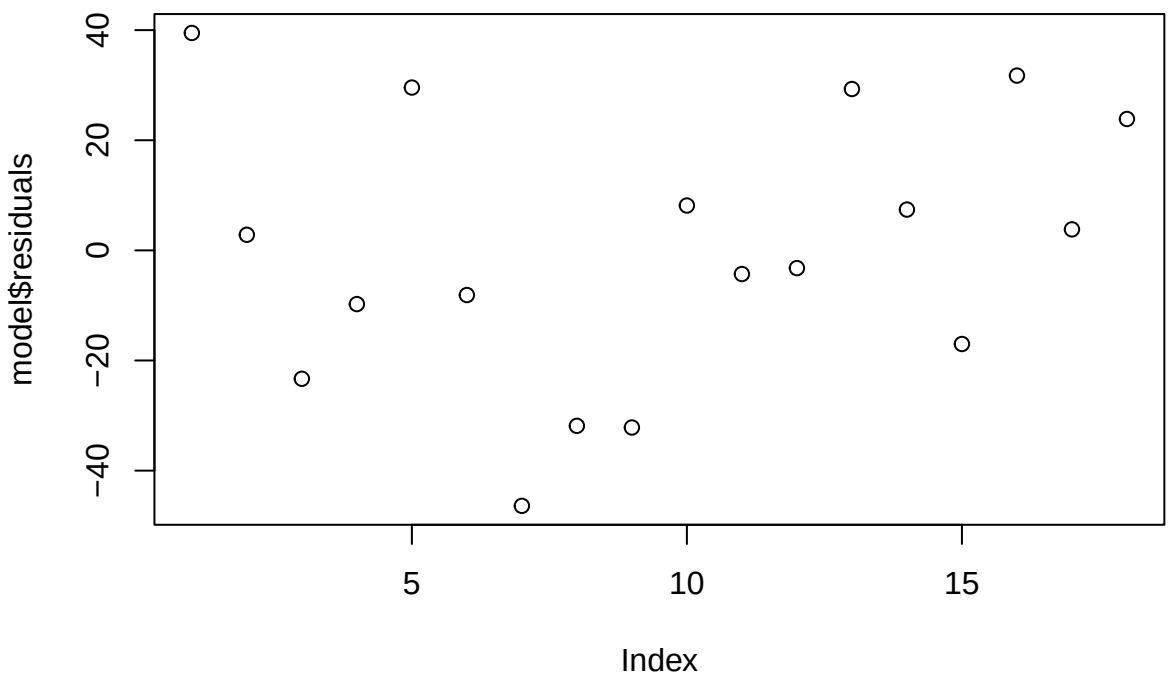
```
aggregated_data <- df_nypd %>%
  group_by(OCCUR_YEAR) %>%
  summarise(Total_Incidents = n(), Total_Deaths = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE))

model <- lm(Total_Deaths ~ Total_Incidents, data = aggregated_data)
summary(model)
```

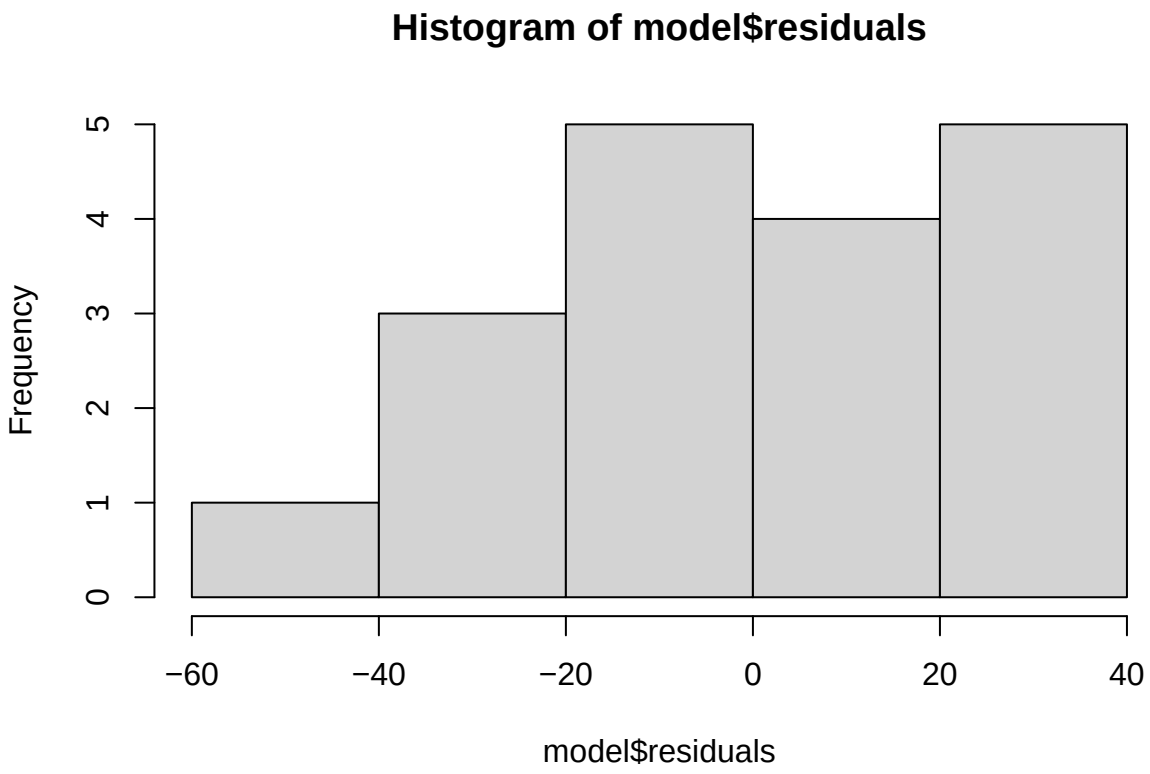
```
##
## Call:
## lm(formula = Total_Deaths ~ Total_Incidents, data = aggregated_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.400 -15.193  -0.197  19.933  39.481
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -26.8749    25.6128  -1.049     0.31
## Total_Incidents   0.2104     0.0157  13.404 4.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.3 on 16 degrees of freedom
## Multiple R-squared:  0.9182, Adjusted R-squared:  0.9131
## F-statistic: 179.7 on 1 and 16 DF,  p-value: 4.078e-10
```

Plotting residuals to check for model assumptions

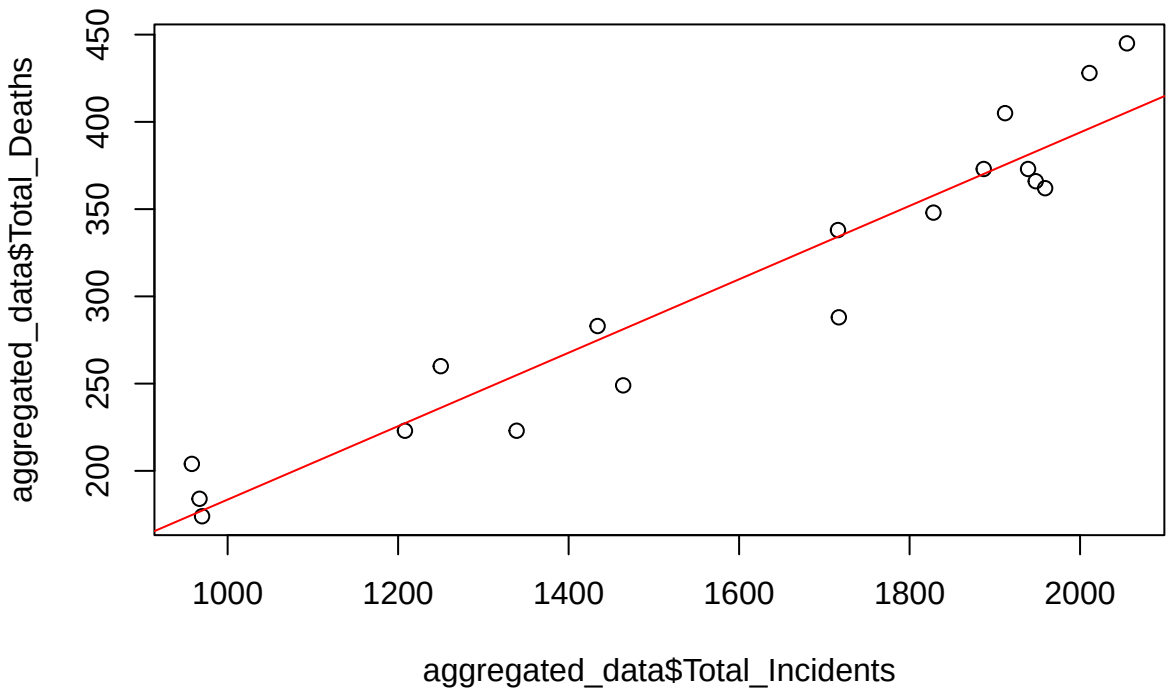
```r
plot(model$residuals)
```




```r
hist(model$residuals)
```




```
plot(aggregated_data$Total_Incidents, aggregated_data$Total_Deaths)
abline(model, col = "red")
```



The model has statistical significance, according to the p and R values, there are a strongh relation between number of incidents and murders. Homoscedasticity and normality assumptions are also true.

Conclusion and Bias consideration

Analyzing such an extensive database from a city as large and diverse as New York represents a significant challenge. During this analysis, we progressively delved from the most general to the most specific.

After identifying with data that the area with the most shootings in the city is the Bronx, we were able to gain a broader understanding of the victims of these incidents. Of course, the information could be further scrutinized, leading to new questions such as why these races are the most affected? Why is there such a marked variation in murder rates?

Additionally, utilizing the complete dataset, an analysis could be conducted on specific areas within the borough and the times when historically more incidents have been recorded to bolster surveillance in those areas. While no internal bias was detected in the database, it cannot be ruled out that there may be bias in data collection or in the behavior of residents in each borough. For example, we do not know the number of incidents that were never reported, or if in some boroughs there is greater trust in the police, leading to the reporting of all incidents. Following this analysis, a definitive conclusion about the incidents in the city was not reached, but a deeper understanding was attained, prompting new questions for further analysis aimed at deriving actionable insights.