NYPD Shooting Incident Data Report

2025-09-05

1. Project Overview

This project uses publicly available data to analyze trends in gun violence in New York City. The primary objective is to practice fundamental data management and analysis skills, including data cleaning, transformation, and visualization, while interpreting key changes in gun violence rates over time.

2. Data Source

The analysis uses NYPD shooting incident data that can be accessed from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD.

3. Data management

3.1. Importing data

```
# Import tidyverse library and data into R
library(tidyverse)
url_of_shooting <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shooting <- read_csv(url_of_shooting)
shooting
```

```
## # A tibble: 29,744 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
##  7     79780323 07/12/2011 22:26      BROOKLYN <NA>                    71
##  8     85744504 07/14/2012 23:45      BROOKLYN <NA>                    69
##  9    142324890 04/21/2015 15:36      BROOKLYN <NA>                    75
## 10    152868707 05/07/2016 15:23      BROOKLYN <NA>                    69
## # i 29,734 more rows
## # 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>
```

From the summarized infomation above, we can see the data is composed by 29744 rows and 21 columns. Each row represents an incident, having a unique ID, occuring date and time, location information, perpetrator and victim information, and statistical murder flag (whether or not the victim is dead).

3.2. Checking missing values

Let's see how many missing values there are for each variable.

```
missing_values_summary <- shooting %>%
  summarise(across(everything(), ~ sum(is.na(.))))

print(missing_values_summary)
```

```
## # A tibble: 1 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME  BORO LOC_OF_OCCUR_DESC PRECINCT
##          <int>      <int>      <int> <int>             <int>    <int>
## 1            0          0          0     0             25596        0
## # i 15 more variables: JURISDICTION_CODE <int>, LOC_CLASSFCTN_DESC <int>,
## #   LOCATION_DESC <int>, STATISTICAL_MURDER_FLAG <int>, PERP_AGE_GROUP <int>,
## #   PERP_SEX <int>, PERP_RACE <int>, VIC_AGE_GROUP <int>, VIC_SEX <int>,
## #   VIC_RACE <int>, X_COORD_CD <int>, Y_COORD_CD <int>, Latitude <int>,
## #   Longitude <int>, Lon_Lat <int>
```

Several variables related with the locations of shooting incidents have large amount of missing values, as well as variables related with perpetrator age group, sex, and race. However, I would like to perform sevaral analysis, each with different subset of the original data. To fully utilize the available information, I would only remove the missing values in the specific subset of data before each analyses.

3.3 Data cleaning and transformation

- Selected 4 columns from original data: OCCUR_DATE, BORO, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP.
- Renamed STATISTICAL_MURDER_FLAG as "deaths"
- Extracted a new feature (year) from OCCUR_DATE and discarded OCCUR_DATE.
- Added a place holder column "Incidents" for future use.
- List category information for each remaining column.

```
shooting1 <- shooting %>%
          select(OCCUR_DATE, BORO, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP) %>%
          rename(Deaths = STATISTICAL_MURDER_FLAG) %>%
          mutate(date = mdy(OCCUR_DATE), Incidents = 1) %>%
          mutate(Year = year(date))  %>%
          select(-c(OCCUR_DATE,date))

shooting1 %>% map(table)
```

```
## $BORO
##
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          8834         11685          3977          4426           822
##
## $Deaths
##
## FALSE  TRUE
## 23979  5765
##
## $PERP_AGE_GROUP
##
##  (null)     <18    1020    1028   18-24    2021     224   25-44   45-64     65+
##    1628    1805       1       1    6630       1       1    6342     775      67
##     940 UNKNOWN
```

```
##       1    3148
##
## $Incidents
##
##     1
## 29744
##
## $Year
##
## 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
## 2055 1887 1959 1828 1912 1939 1717 1339 1464 1434 1208  970  958  967 1948 2011
## 2022 2023 2024
## 1716 1250 1182
```

From above, we see Brooklyn is the district with the most shooting incidents. We also see a lot of ambiguity and errors in the perpetrator age groups.

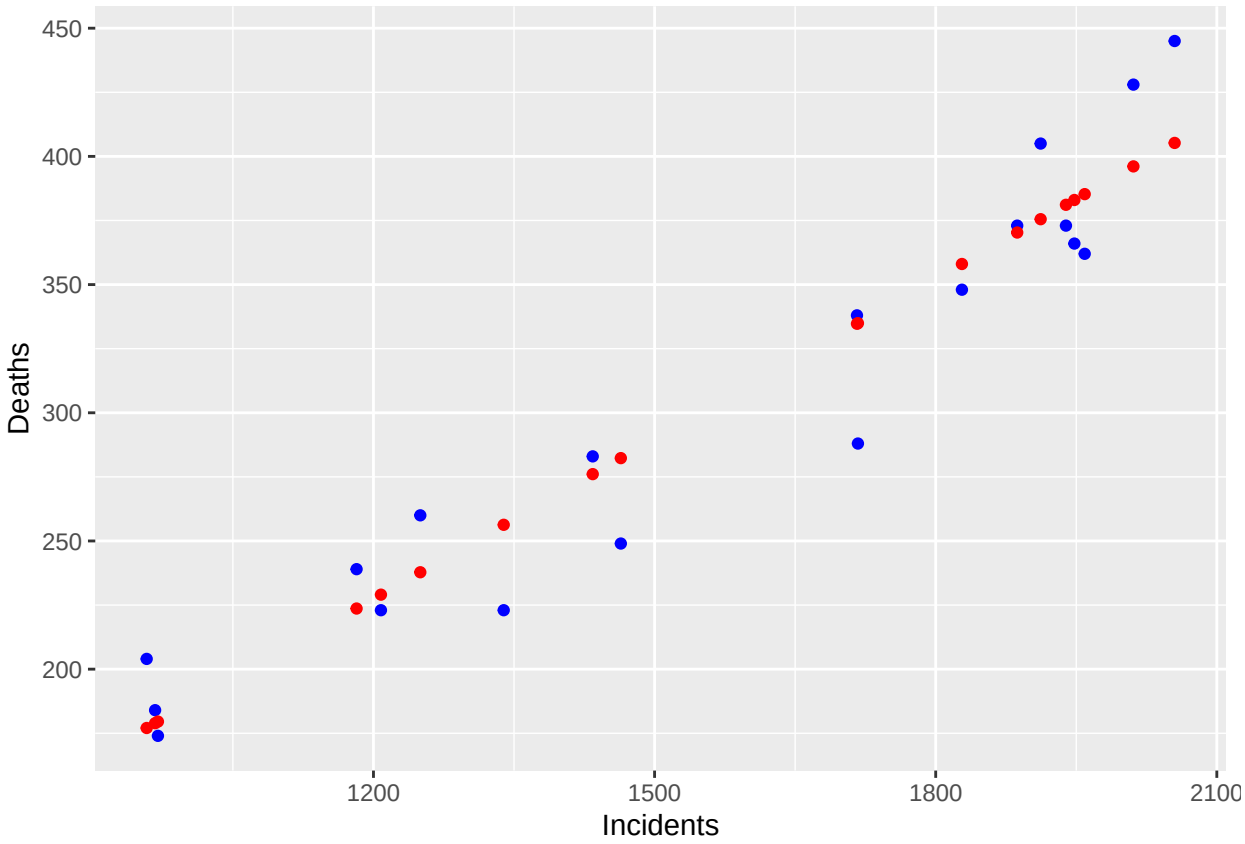
4. Analysis and Visualization

4.1. Modelling death count using the simple linear regression model

```
shooting2 <- shooting1 %>%
    select(Year, Incidents, Deaths, BORO) %>%
    group_by(Year) %>%
    summarise(Incidents = sum(Incidents),
              Deaths = sum(Deaths))
mod = lm(Deaths ~ Incidents, data = shooting2)
summary(mod)
```

```
##
## Call:
## lm(formula = Deaths ~ Incidents, data = shooting2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.941 -13.512   2.696  18.776  39.749
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.22569   24.13815  -0.921     0.37
## Incidents     0.20802    0.01498  13.884 1.05e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.85 on 17 degrees of freedom
## Multiple R-squared:  0.919,  Adjusted R-squared:  0.9142
## F-statistic: 192.8 on 1 and 17 DF,  p-value: 1.049e-10
```

```
shooting2_w_pred <- shooting2 %>% mutate(pred = predict(mod))
shooting2_w_pred %>%
    ggplot() +
    geom_point(aes(x = Incidents, y = Deaths), color = "blue") +
    geom_point(aes(x = Incidents, y = pred), color = "red")
```

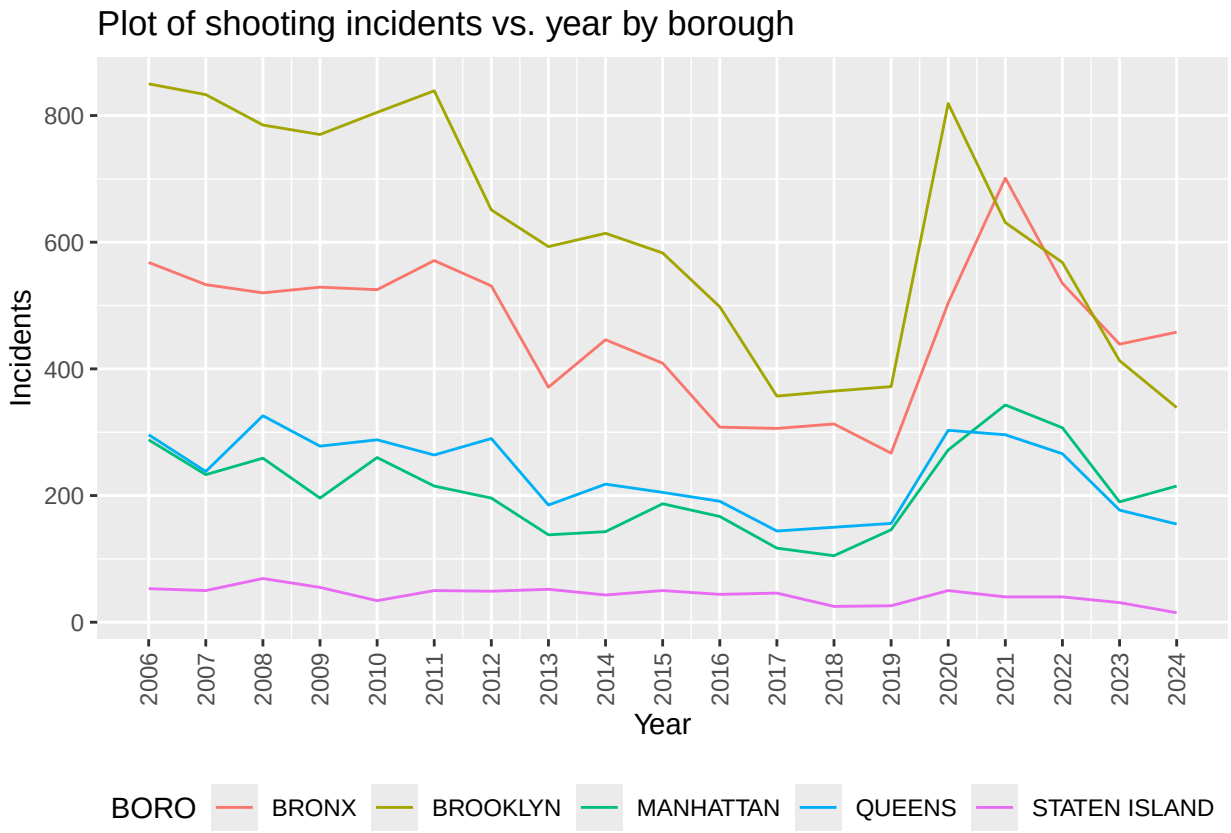


As expected, both the simple linear regression output summary and the ggplot figure showed that the number of deaths could be well explained and predicted by the number of shooting incidents. However, the overall distribution has a slightly convex shape, which implies that other factors might also be influencing the number of deaths.

4.2. Change of gun violence over years in different boroughs

```
# The data is aggregated by two variables: Year and BORO.
shooting1  %>%
    group_by(Year, BORO) %>%
    summarise(Incidents = sum(Incidents)) %>%
    ggplot(aes(x = Year, y = Incidents, group = BORO, color = BORO)) +
    geom_line() +
    scale_x_continuous(breaks = unique(shooting1$Year)) +
    theme(legend.position = "bottom",
          axis.text.x = element_text(angle = 90, vjust = 0.5)) +
    labs(title = "Plot of shooting incidents vs. year by borough")
```

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



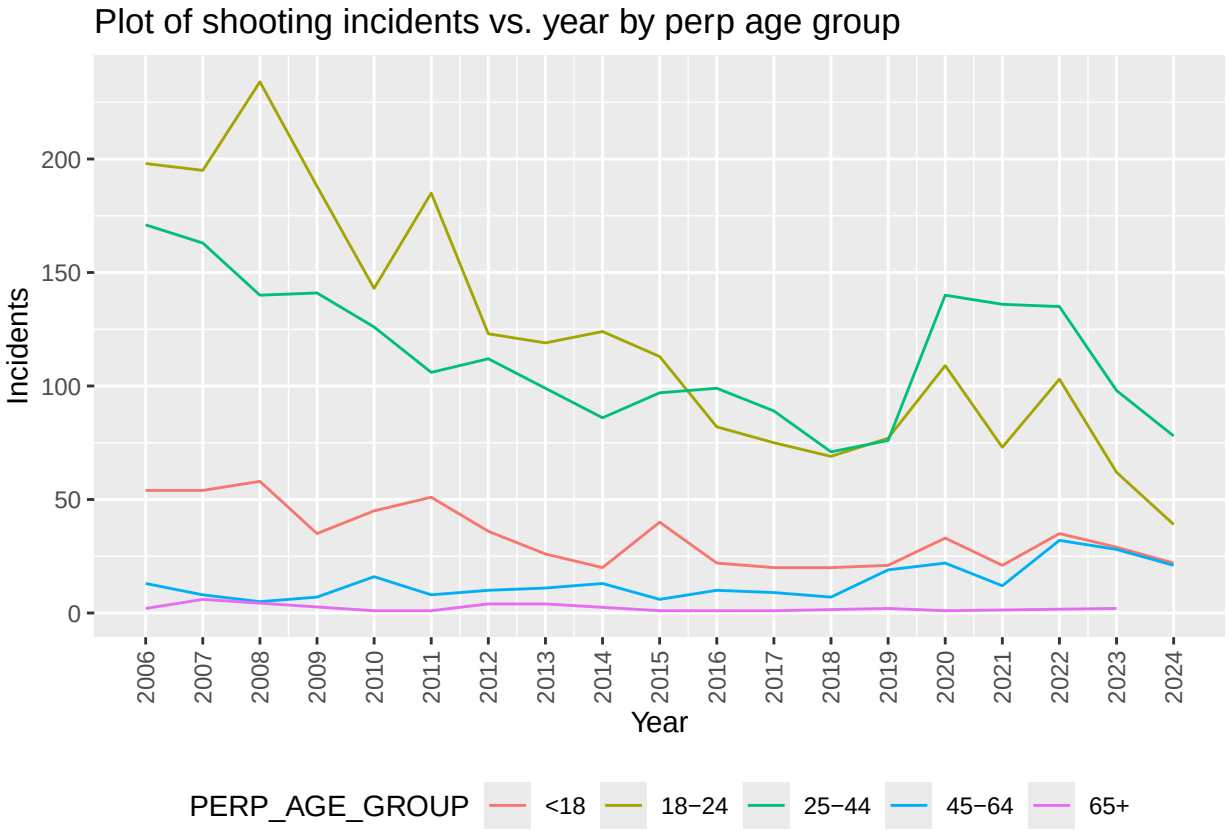
From the figure, the largest number of shootings occurred in Brooklyn and the Bronx compared to the other boroughs. The number of shootings decreased from 2006 until 2019, then increased. Since 2021, the number of shootings in the Bronx has begun to reach the level of Brooklyn. What caused the number of shootings to decrease until 2019 and then rise is an interesting question to explore.

4.3. Change of gun violence over years by different perpetrator age groups in Brooklyn

Here, we zoom in on Brooklyn, to see the change of the number of shooting incidents over years by different perpetrator age groups.

```
#The data is subset to only include shooting incidents
#in Brooklyn and with clear perptrator age groups
#It is then aggregated by two variables: Year and PERP_AGE_GROUP.
shooting1 %>%
    filter(BORO == "BROOKLYN",PERP_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+")) %>%
    group_by(Year, PERP_AGE_GROUP) %>%
    summarise(Incidents = sum(Incidents)) %>%
    ggplot(aes(x = Year, y = Incidents, group = PERP_AGE_GROUP, color = PERP_AGE_GROUP)) +
    geom_line() +
    scale_x_continuous(breaks = unique(shooting1$Year)) +
    theme(legend.position = "bottom",
          axis.text.x = element_text(angle = 90, vjust = 0.5)) +
    labs(title = "Plot of shooting incidents vs. year by perp age group")
```

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



From the figure, we can observe that the number of shootings was lowest around 2018-2019, after which it began to rise again. The "18-24" and "25-44" age groups consistently account for the highest numbers of shootings. The "18-24" group committed more shootings until approximately 2015. Subsequently, the "25-44" age group became responsible for a greater number of incidents. It would be insightful to investigate the reasons behind this shift.

5. Conclusion

- The number of deaths was well predicted by the number of shooting incidents.
- The number of shootings decreased from 2006 to 2019, then increased afterward.
- More shooting incidents occurred in Brooklyn and the Bronx than in other boroughs.
- Perpetrators in the "18-24" and "25-44" age groups were responsible for more shootings. Interestingly, before 2016, the "18-24" group committed more shootings than the "25-44" group; after 2016, the "25-44" group committed more.

6. Possible Bias

- The data does not include suicidal and accidental shootings, which makes it difficult to paint a full picture of gun violence.
- The population size for each borough and age group is unknown. The conclusions would be more accurate if we could calculate standardized rates (e.g., incidents per 100,000 people) for each group.
- I acknowledge that I have personal biases regarding different New York districts and age groups. I tried to maintain an open mind and analyze the data from all angles, regardless of my personal views, and obtained some interesting results.