NYPD Shooting Incident Data

Luis Hinostroza

2024-03-04

NYPD Shooting Incident Data Analysis

This data represent the NYPD Shooting Incident Data Report based on a dataset retrieved from https://catalog.data.gov/dataset covering 2006 to 2023 reported incidents. The purpose of this Analysis is to gain more insight into the shooting incidents and to find potential trends that can be used for crime prediction.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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
```

```
library(ggplot2)
library(dplyr)
library(lubridate)
library(chron)
```

```
##
## Attaching package: 'chron'
##
## The following objects are masked from 'package:lubridate':
##
##     days, hours, minutes, seconds, years
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(shiny)
```

```r
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
df <- read_csv(url_in[1])
```

```
## Rows: 27312 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.
```

```r
df[sapply(df, is.null)] <- NA
```

The Data has null values, non sense values, and empty cells. The following code will clean up the table and tidyup the results.

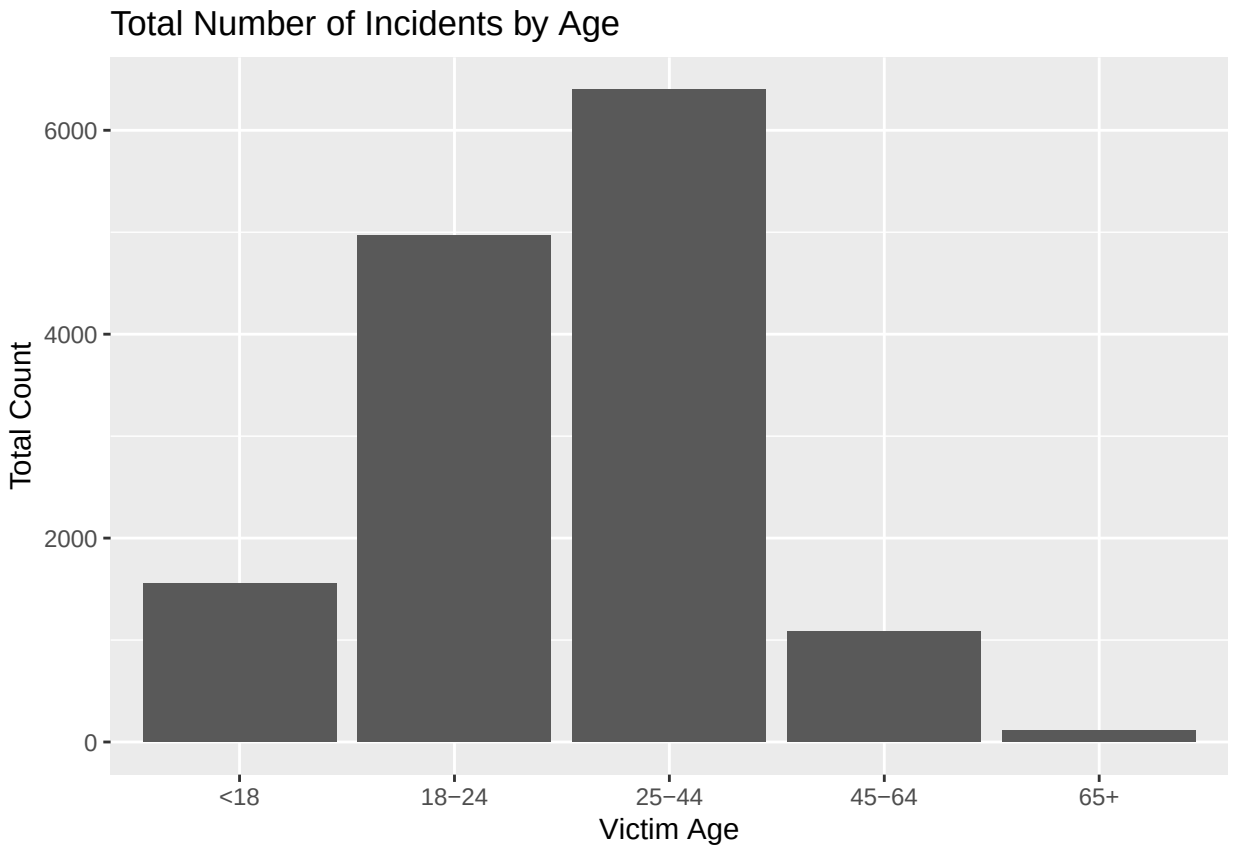
```
shooting_cases <- df %>%
  mutate(DATE = mdy(OCCUR_DATE)) %>%
  select(c(DATE, OCCUR_TIME, BORO, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE
  na.omit() %>%
  filter(PERP_AGE_GROUP != "UNKNOWN", VIC_AGE_GROUP != "UNKNOWN", PERP_AGE_GROUP != "(null)", VIC_AGE_GR

shooting_cases
```

```
## # A tibble: 14,122 x 9
##    DATE       OCCUR_TIME BORO    PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
##    <date>     <time>     <chr>   <chr>          <chr>    <chr>     <chr>
##  1 2009-02-19 22:58      BRONX   25-44          M        BLACK     45-64
##  2 2012-08-26 01:10      QUEENS  25-44          M        BLACK     25-44
##  3 2010-08-29 01:27      BROOKL~ 25-44          M        BLACK     25-44
##  4 2010-07-27 02:22      MANHAT~ 25-44          M        BLACK     25-44
##  5 2021-03-07 21:17      BROOKL~ 25-44          M        BLACK     25-44
##  6 2015-02-01 23:16      MANHAT~ 18-24          M        BLACK     18-24
##  7 2007-10-11 20:11      BRONX   45-64          M        BLACK     25-44
##  8 2016-03-28 19:10      BRONX   25-44          F        BLACK HI~ 45-64
##  9 2013-03-06 04:08      QUEENS  25-44          M        BLACK HI~ 25-44
## 10 2006-01-21 21:49      MANHAT~ 18-24          M        BLACK     18-24
## # i 14,112 more rows
## # i 2 more variables: VIC_SEX <chr>, VIC_RACE <chr>
```

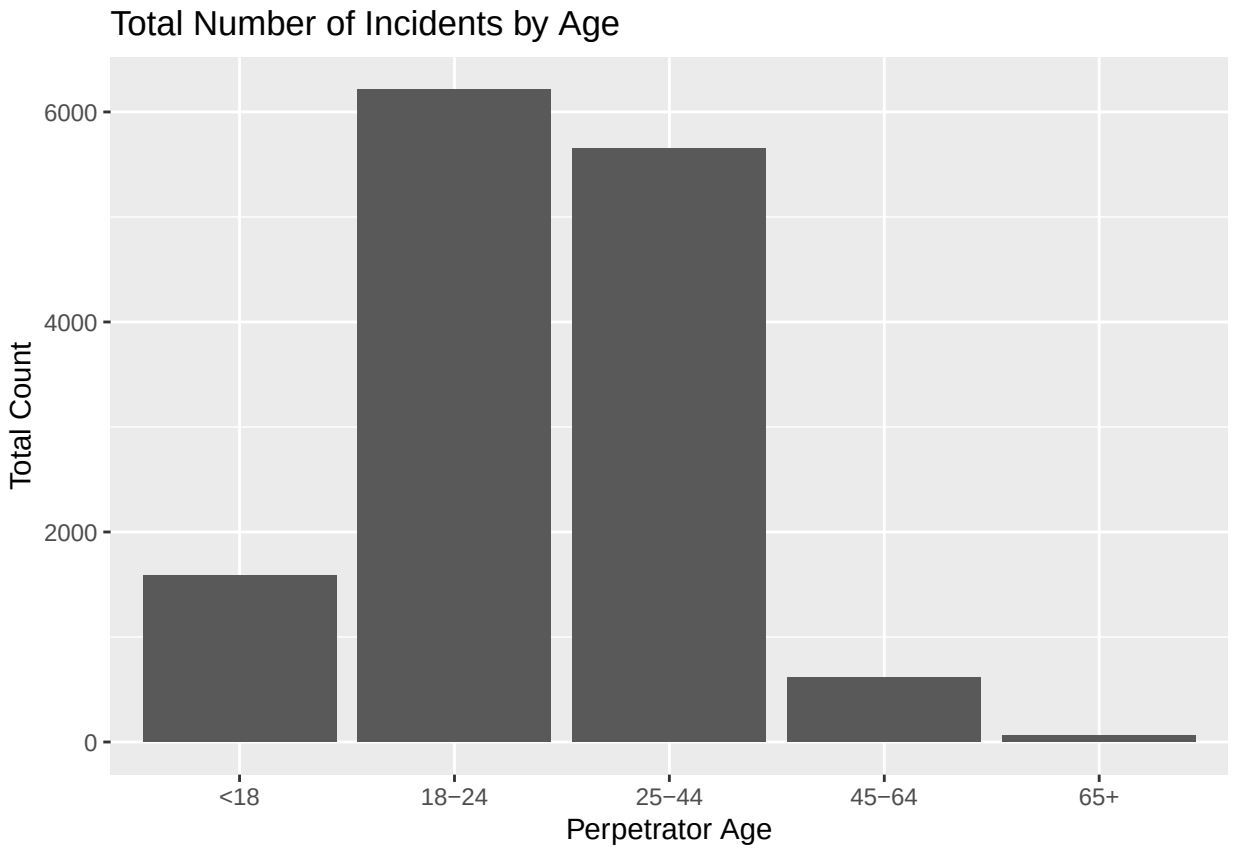
Visualization

Now we will visualize how age and race influence into the amount of shooting incidents. You will see from the graphs that young people between 18 to 24 years old are most of the perpetrator follow by 25 to 44 years old. The Victim ages are very similar making the 25 to 44 years old the most of the victims followed by 18 to 24. The last two graphs about race show black people as the majority race for the victim and perpetrator incidents, this could be due to the demographic predominance in the majority of these cities.

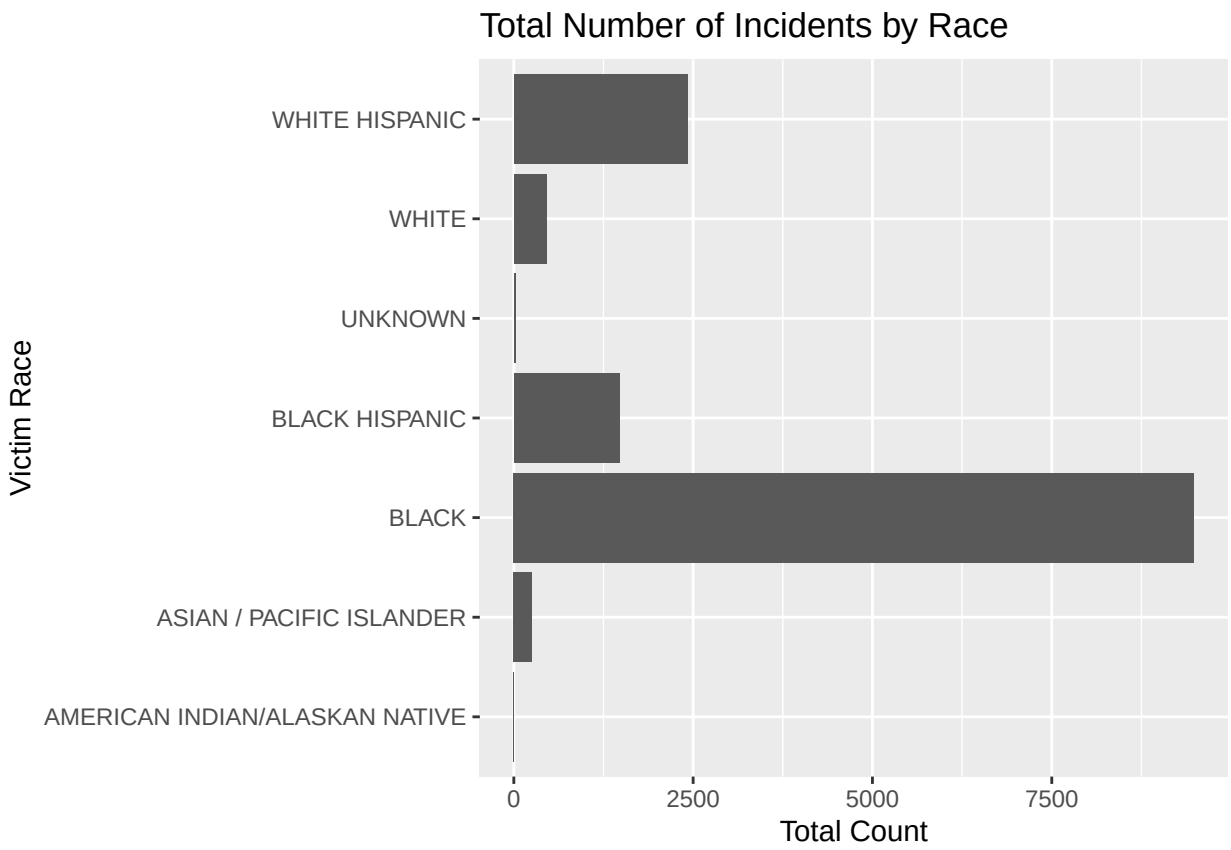
```
shooting_cases %>%
  ggplot(aes(x = VIC_AGE_GROUP)) +
  geom_bar() + labs(title = "Total Number of Incidents by Age", x = "Victim Age", y = "Total Count")
```



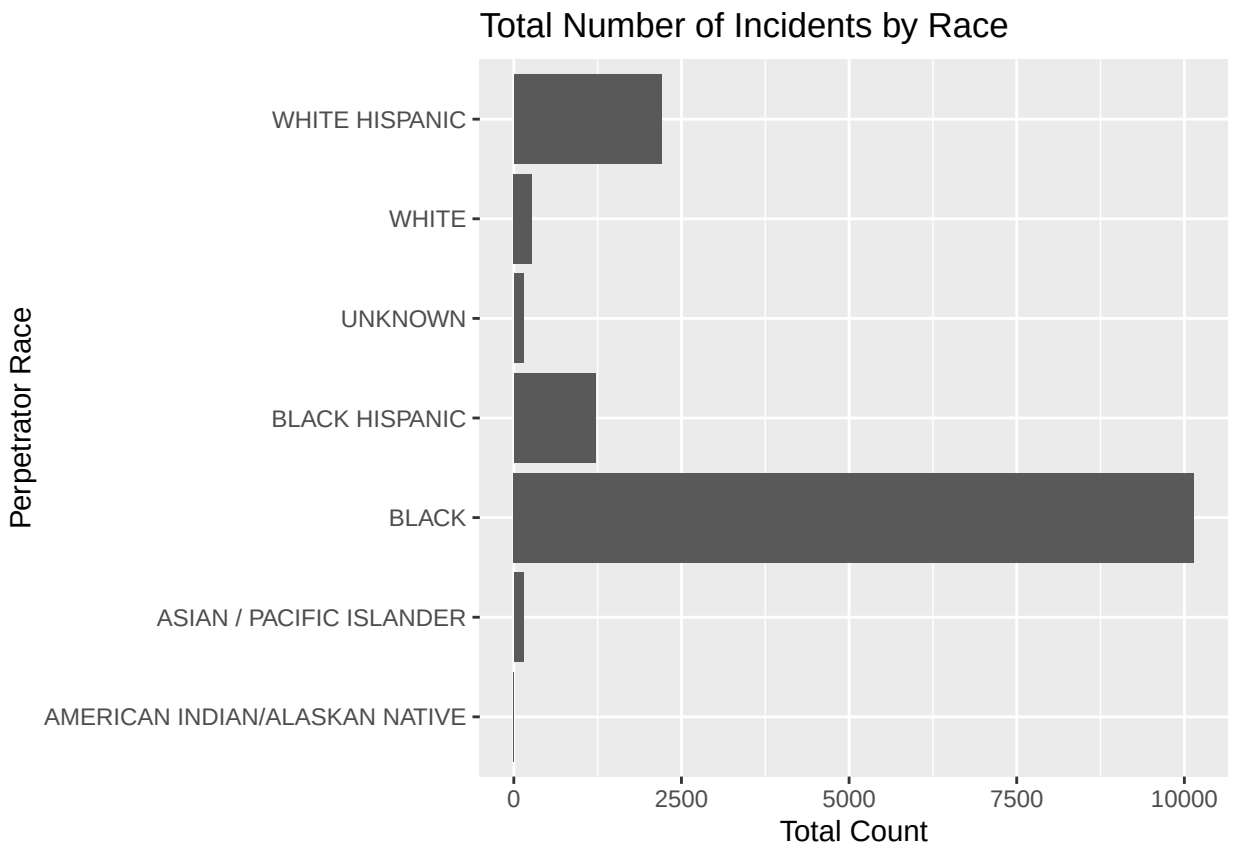

```
shooting_cases %>%
  ggplot(aes(x = PERP_AGE_GROUP)) +
  geom_bar() + labs(title = "Total Number of Incidents by Age", x = "Perpetrator Age", y = "Total Count
```




```
shooting_cases %>%
  ggplot(aes(y = VIC_RACE)) +
  geom_bar() + labs(title = "Total Number of Incidents by Race", x = "Total Count", y = "Victim Race")
```




```
shooting_cases %>%
  ggplot(aes(y = PERP_RACE)) +
  geom_bar() + labs(title = "Total Number of Incidents by Race", x = "Total Count", y = "Perpetrator Ra
```



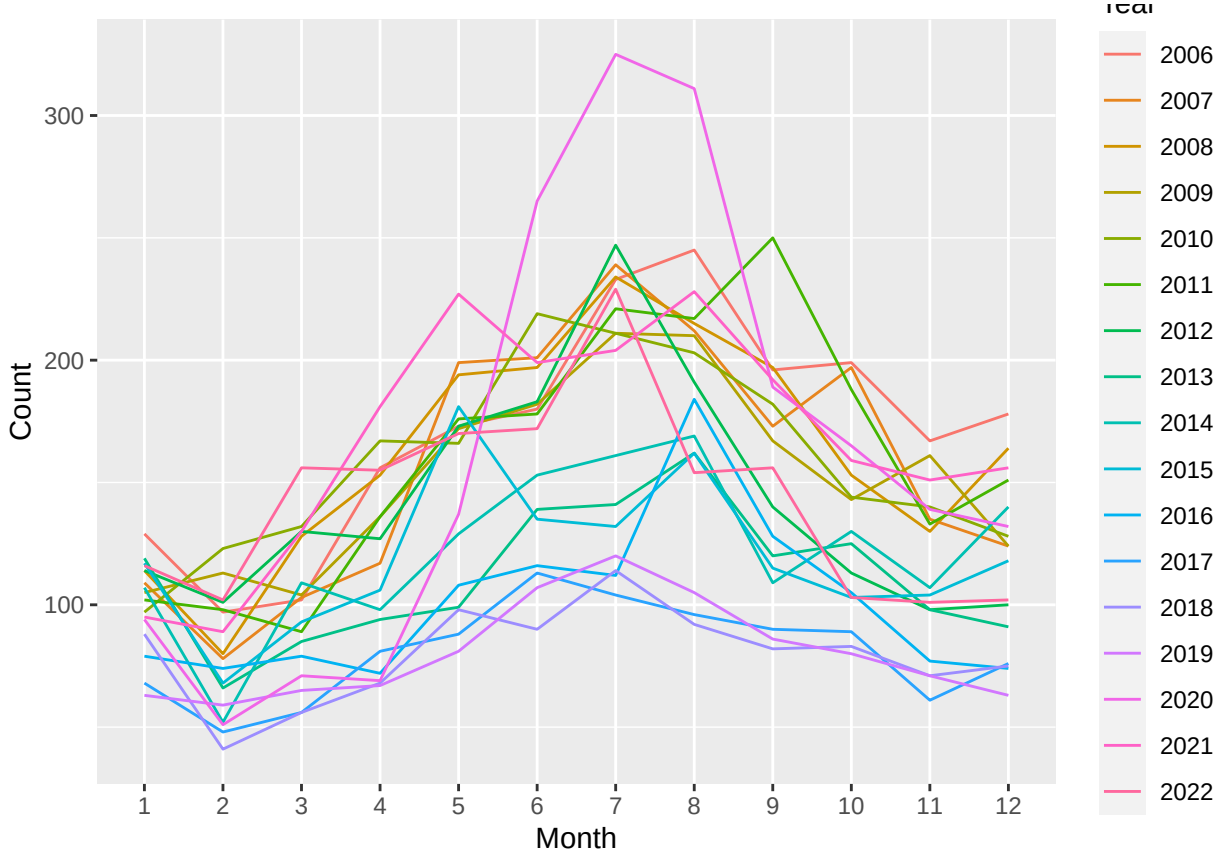
Analysis

My following analysis is to find a pattern on the incidents for every year. The graph below shows the the peak tendency is during the summer. We can also see the surge in shootings in the summer of 2020 after the killing of George Floyd

```
df$OCCUR_DATE = as.Date(df$OCCUR_DATE, format = "%m/%d/%Y")
df$Year = df$OCCUR_DATE %>% format("%Y") %>% as.integer()
df$Month = df$OCCUR_DATE %>% format("%m") %>% as.integer()
yearmon_df = df %>% group_by(Year, Month) %>% dplyr::summarise(Count = n()) %>% as.data.frame()
```

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

```
yearmon_df$Year = yearmon_df$Year %>% as.factor()
yearmon_df$Month = yearmon_df$Month %>% as.factor()
ggplot(yearmon_df, aes(x = Month, y = Count, group = Year, col = Year)) + geom_line()
```



Data Bias and Conclusion

Possible bias that may have been introduced into this data set may come from areas that are more heavily policed. Accordingly, other boroughs that do not have as large of a law-enforcement presence may be under- reporting incidents. In other words, more police officers in an area means more availability to respond to and report shooting incidents. It is not clear whether this data includes instances where people literally were hit with a bullet or if there are also incidents where a victim was just shot at; either way there are presumably more 'shots fired' incidents not included in this data set which have different feature distributions from this dataset. Perpetrator description columns may be subject to direct bias as they may be garnered from witness statements which can be faulty. Victim description columns should be better since it is easier to actually locate and confirm a shooting victim