NYPD Shooting Incident Data Report

2023-05-02

Load These First

```
#install.packages("tidyverse")
#install.packages("magrittr")
#install.packages("dplyr")
#install.packages("readr")
#install.packages("lubridate")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(dplyr)
library(readr)
library(lubridate)
```

Data Source and Report Goals

I retrieved the New York Police Department's Shoot Incident Data Set from the city of New York's website. Below, I'll explain how I figured out which data fields I wanted to work with and which I decided to exclude.

But I ended up looking at the dates the shootings occurred compared against the shooting victims' age, race, and sex. To illustrate my findings, I have put together a few visualizations, a data analysis, and a couple of models to predict future trends.

```
## Importing the NYPD Shooting Incident Data (Historic) from the online .CSV file.
df <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
# df
```

Tidying the Data

As of today, the data set contains 27,312 rows - representing individual incidents.

Data represented

The data set is made up of 21 columns: INCIDENT KEY, OCCUR_DATE, OCCUR_TIME, BORO, LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat.

Based on the information presented in each column, it looks like I can look at either perpetrator or victim demographic information, at the borough or precinct the incident took place in, or at the date or time the incident took place at.

Data I'm excluding

LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC are empty, JURISDICTION_CODE appears to only contain 0s and 2s - so it doesn't appear useful, LOCATION_DESC seems to have notes about the location - not anything I can show right now in a statistical way, STATISTICAL_MURDER_FLAG contains only true or false - which isn't interesting to me right now, PERP_AGE_GROUP, PERP_SEX, and PERP_RACE appear to be incompletely filled, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, and Lon_Lat give me location data that I'm not interested in specifics of. For these reasons, I want to remove the previously mentioned columns.

Data I'm including

This leaves me with wanting to keep the following columns: INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, PRECINCT, VIC_AGE_GROUP, VIC_SEX, VIC_RACE. INCIDENT_KEY will serve as the incident identifier. OCCUR_DATE looks like it will be useful to model incidents in various demographics over time. VIC_AGE_GROUP, VIC_SEX, and VIC_RACE appear to be the best variables to create visualizations and analysis of.

So I'll first remove the potentially irrelevant columns.

```
# First I'm removing' the potentially irrelevant columns.
relevant_data <- df %>% select(-c(LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, JURISDICTION_CODE, LOCATION_DES

# Then I'll limit the remaining columns to the ones I'm looking at using for my visualizations, analysi
shooting_data <- relevant_data %>% select(-c(OCCUR_TIME, BORO, PRECINCT))
#shooting_data

# I'm parsing the columns I've selected to include so I can actually categorize them.
```

```
shooting_data$INCIDENT_KEY <- parse_factor(as.character(df$INCIDENT_KEY))
shooting_data$OCCUR_DATE <- parse_date_time(df$OCCUR_DATE, "%m%d%Y")
shooting_data$VIC_AGE_GROUP <- parse_factor(df$VIC_AGE_GROUP)
shooting_data$VIC_SEX <- parse_factor(df$VIC_SEX)
shooting_data$VIC_RACE <- parse_factor(df$VIC_RACE)

# I'm taking the first 10 rows to illustrate that I selected all the columns that I want for my report.
head(shooting_data, 10)
```

```
##    INCIDENT_KEY OCCUR_DATE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1     228798151 2021-05-27         18-24       M          BLACK
## 2     137471050 2014-06-27         18-24       M          BLACK
## 3     147998800 2015-11-21         25-44       M          WHITE
## 4     146837977 2015-10-09           <18       M WHITE HISPANIC
## 5      58921844 2009-02-19         45-64       M          BLACK
## 6     219559682 2020-10-21         25-44       M          BLACK
## 7      85295722 2012-06-17         25-44       M          BLACK
## 8      71662474 2010-03-08         18-24       M          BLACK
## 9      83002139 2012-02-05         25-44       M          BLACK
## 10     86437261 2012-08-26         25-44       M          BLACK
```

Also, this should show that the demographics I selected aren't missing any information.

```
sum(is.na(shooting_data))
```

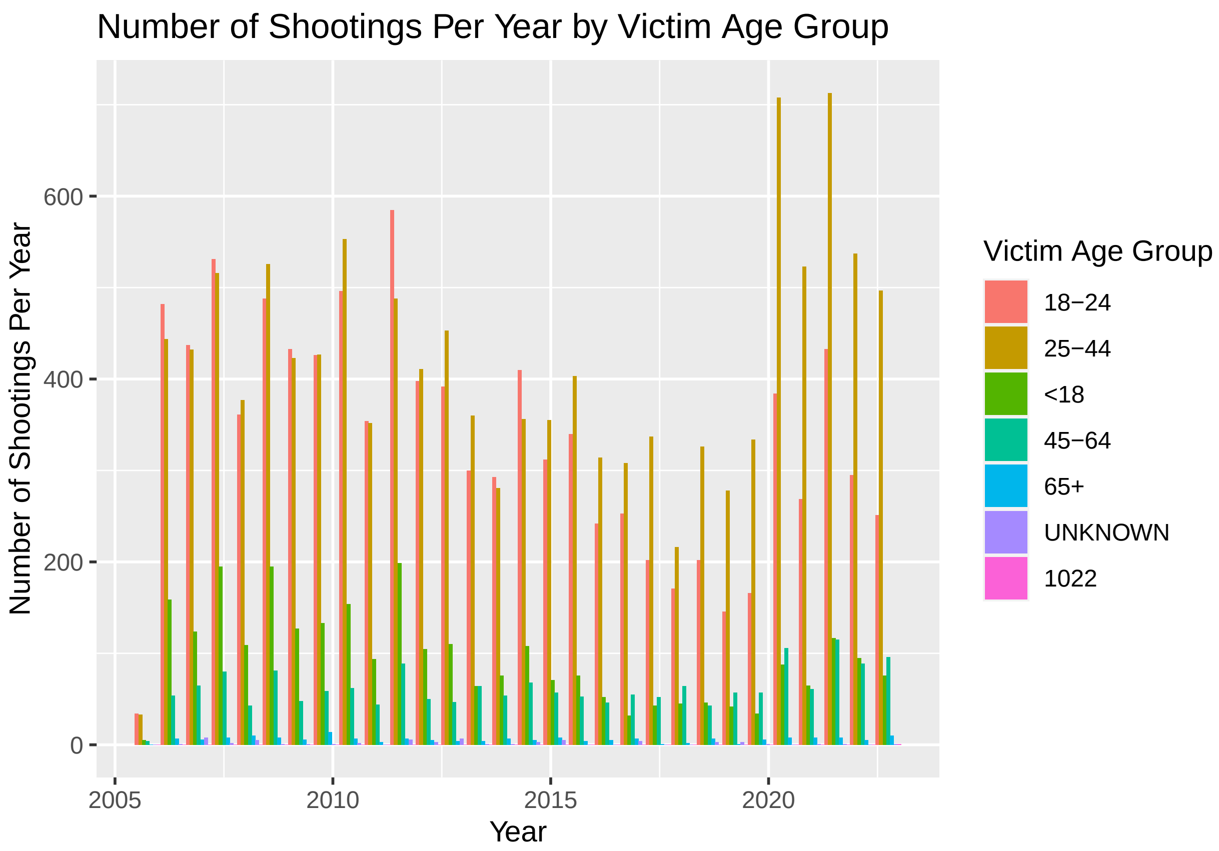
```
## [1] 0
```

Visualizations

Below, I'm including visualizations that depict each of my three chosen demographics against shootings by occurrence date. The X-axis is the Year and the Y-axis is the total in the specific demographic being looked at (Age/Race/Sex). There are multiple bars in the graph per year to illustrate the breakdown by category (Age Group/Race/Sex).

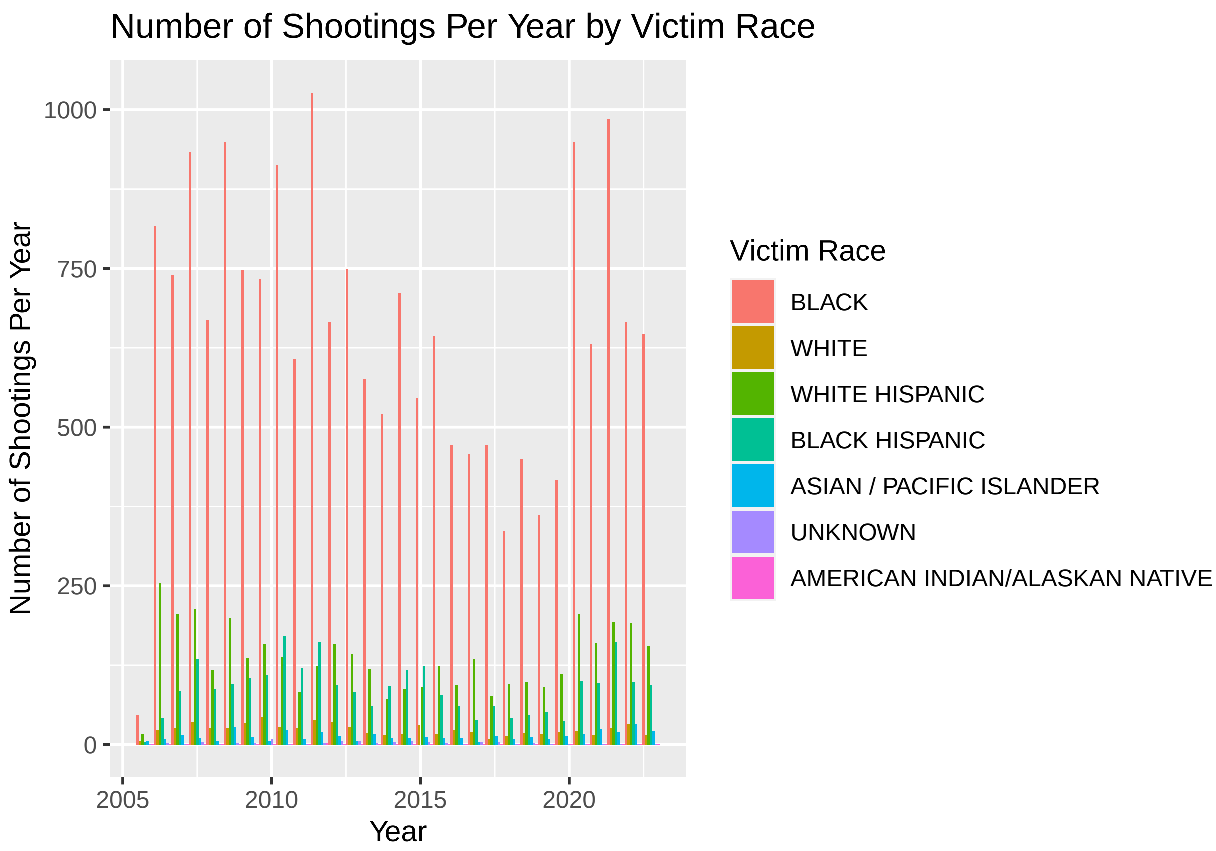
```
#ggplot(x = OCCUR_DATE, y = VIC_AGE_GROUP)
## Number of Shootings Per Year by Victim Age Group
shooting_data %>%
  ggplot(aes(x = OCCUR_DATE, fill = VIC_AGE_GROUP)) +
  geom_histogram(position = "dodge") +
  labs(x = "Year", y = "Number of Shootings Per Year", fill = "Victim Age Group", title = "Number of Sho
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



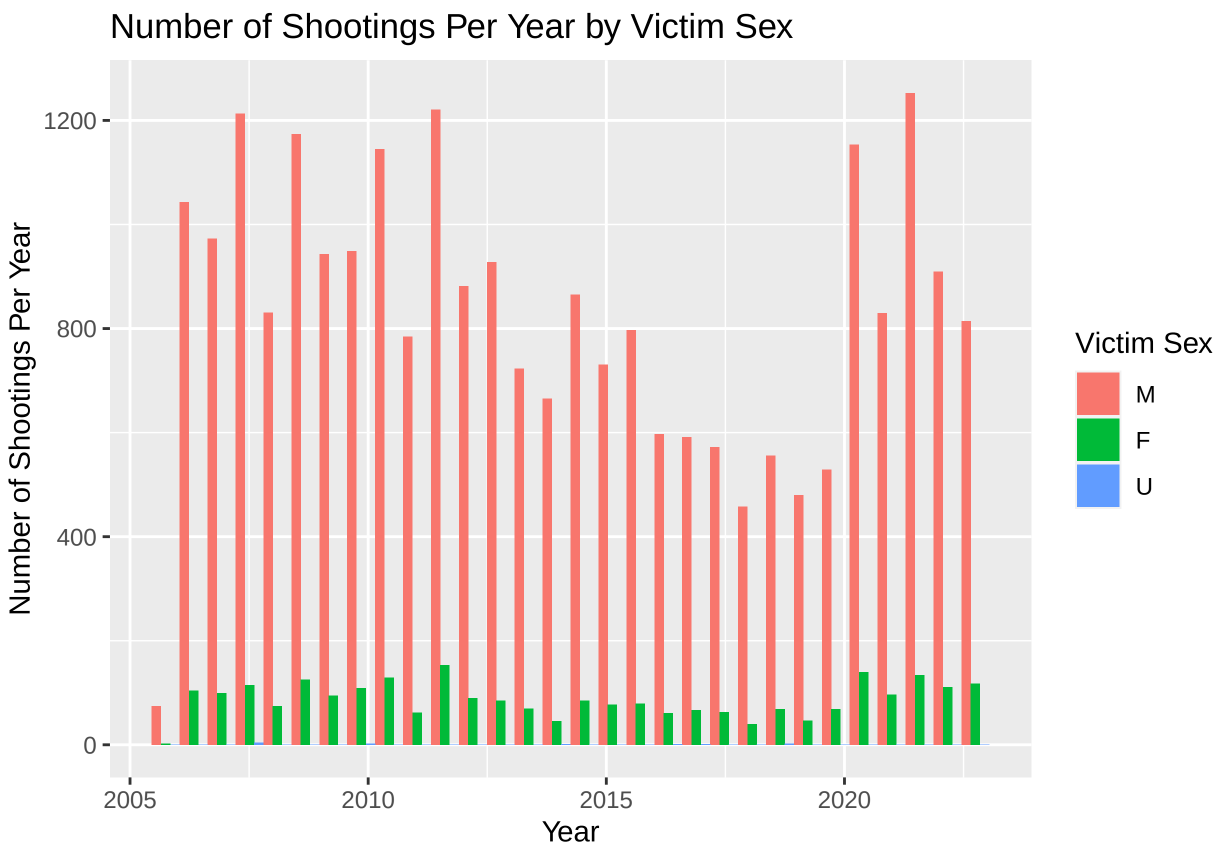
```
## Number of Shootings Per Year by Victim Race
shooting_data %>%
  ggplot(aes(x = OCCUR_DATE, fill = VIC_RACE)) +
  geom_histogram(position = "dodge") +
  labs(x = "Year", y = "Number of Shootings Per Year", fill = "Victim Race", title = "Number of Shootin
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
## Number of Shootings Per Year by Victim Age Group
shooting_data %>%
  ggplot(aes(x = OCCUR_DATE, fill = VIC_SEX)) +
  geom_histogram(position = "dodge") +
  labs(x = "Year", y = "Number of Shootings Per Year", fill = "Victim Sex", title = "Number of Shootings
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Analysis

Looking at the visualizations, it is pretty easy to tell that the Race and the Sex of the shooting victims are extremely lopsided. I mean that Blacks and Hispanics are disproportionately the victims of shooting incidents compared to other races. The same can be said for Sex. Males are disproportionately more represented as shooting victims as compared to the female numbers.

However, the trends aren't quite as obvious in the Age Groups visualization. So I isolated the age groups for the years 2021 and 2022 to get a better look at what those numbers look like via their percentages of the whole group of shooting victims.

```
 age_analysis <- shooting_data %>% group_by(year = lubridate::floor_date(OCCUR_DATE, "year")) %>% mutat
                  group_by(year, VIC_AGE_GROUP) %>%
                  reframe(year = year, Percentage = 100*(n()/count)) %>%
                  arrange(-desc(year)) %>% distinct() %>% ungroup()
tail(age_analysis, n = 13)
```

```
## # A tibble: 13 x 3
##    year                VIC_AGE_GROUP Percentage
##    <dttm>              <fct>              <dbl>
##  1 2021-01-01 00:00:00 18-24             29.9
##  2 2021-01-01 00:00:00 25-44             53.9
##  3 2021-01-01 00:00:00 <18                7.76
##  4 2021-01-01 00:00:00 45-64              7.81
##  5 2021-01-01 00:00:00 65+                0.646
##  6 2021-01-01 00:00:00 UNKNOWN           0.0497
##  7 2022-01-01 00:00:00 18-24            28.1
##  8 2022-01-01 00:00:00 25-44            52.2
##  9 2022-01-01 00:00:00 <18               9.21
## 10 2022-01-01 00:00:00 45-64             9.62
## 11 2022-01-01 00:00:00 65+               0.816
## 12 2022-01-01 00:00:00 UNKNOWN           0.0583
## 13 2022-01-01 00:00:00 1022              0.0583
```

Now, it is easier to tell that the age group 25-44 is represented almost twice as much in shooting incidents compared to the 18-24 age group. Then, the <18 and 45-64 age groups are almost equally represented.

Linear Model

Now that I know the top two age groups being the victims of shooting incidents, I ran a linear model on both of them.

```
young_adults <- age_analysis %>% filter(VIC_AGE_GROUP == unique(age_analysis$VIC_AGE_GROUP[3]))
lmod.young_adults <- lm(Percentage ~ year, data = young_adults)
summary(lmod.young_adults)
```

```
##
## Call:
## lm(formula = Percentage ~ year, data = young_adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9644 -0.7776 -0.3625  0.6592  2.8388
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.995e+01  3.199e+00   9.363 1.18e-07 ***
## year        -1.437e-08  2.290e-09  -6.276 1.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 15 degrees of freedom
## Multiple R-squared:  0.7242, Adjusted R-squared:  0.7058
## F-statistic: 39.39 on 1 and 15 DF,  p-value: 1.485e-05
```

```
adults <- age_analysis %>% filter(VIC_AGE_GROUP == unique(age_analysis$VIC_AGE_GROUP)[1])
lmod.adults <- lm(Percentage ~ year, data = adults)
summary(lmod.adults)
```

```
##
## Call:
## lm(formula = Percentage ~ year, data = adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0718 -1.7585 -0.1311  2.2264  4.8957
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.499e+01  5.758e+00  13.023 1.40e-09 ***
## year        -2.767e-08  4.122e-09  -6.714 6.93e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.627 on 15 degrees of freedom
## Multiple R-squared:  0.7503, Adjusted R-squared:  0.7337
## F-statistic: 45.08 on 1 and 15 DF,  p-value: 6.935e-06
```

These linear models seem to indicate that the number of shootings for each age group should be decreasing. However, each of the visualizations indicate that the number of shootings in NYC spiked during the COVID-19 pandemic beginning in 2020. Before 2020, the general trend had been a decrease in shooting incidents lasting through 2019. So a linear model probably isn’t the most useful way to look at this information right now.

Bias Identification

Like I just mentioned, the COVID-19 pandemic coincided with a spike in the number of shooting incidents. So while you may be able to predict demographics of people involved in shooting incidents with this data, using the whole set of Occurrence Dates might not give you reliable information about the number of shooting incidents you can expect in a given year.

Otherwise, I did try to go into this investigation without considering personal biases. I looked at visualizations of the three different victim demographics first before deciding that the Age Groups were actually the most interesting and I wanted to further analyze and model that.

Though, there is also the possibility that the data that was recorded (or wasn’t recorded) was the subject of bias. Police shootings and perceptions of police trustworthiness are also unfortunately societal issues in the United States today. So it is possible that police bias went into recording the data that was made available in this data set. But it might also my bias that I’m too suspicious of the reliability of this data.