NYC Shooting Data Analysis

Morgan I. MacKay

2024-09-07

Setup

To start, I loaded in the libraries I thought I would need to use for this project.

```
library(datasets)
library(lubridate)
library(ggplot2)
library(tidyverse)
```

Secondly, I loaded in the "NYC Shooting Incident Historic" data from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD (https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD). I also renamed it "dataset" to simplify things a bit. I executed the "View()" function to get a better idea of what the data looked like.

```
NYPD_Shooting_Incident_Data__Historic_ <- read.csv("~/Downloads/NYPD_Shooting_Incident_Data__Historic_.csv")

dataset <- NYPD_Shooting_Incident_Data__Historic_

View(dataset) #output not shown in markdown for simplicity
```

Cleaning

I knew that certain columns of the data set wouldn't be of interest to me for the analysis I was hoping to do. Using the "Dyplr" package in the "Tidyverse" I removed several columns.

```
data <- dataset %>%
  select(-OCCUR_TIME, -LOC_OF_OCCUR_DESC, -PRECINCT, -JURISDICTION_CODE,
         -LOC_CLASSFCTN_DESC, -STATISTICAL_MURDER_FLAG, -X_COORD_CD, -Y_COORD_CD,
         -Latitude, -Longitude, -Lon_Lat, -LOCATION_DESC)
```

I also changed the date variable from "chr" format to "Date" format. This operation relied on the "lubridate" package.

```
data$OCCUR_DATE <- mdy(data$OCCUR_DATE)
```

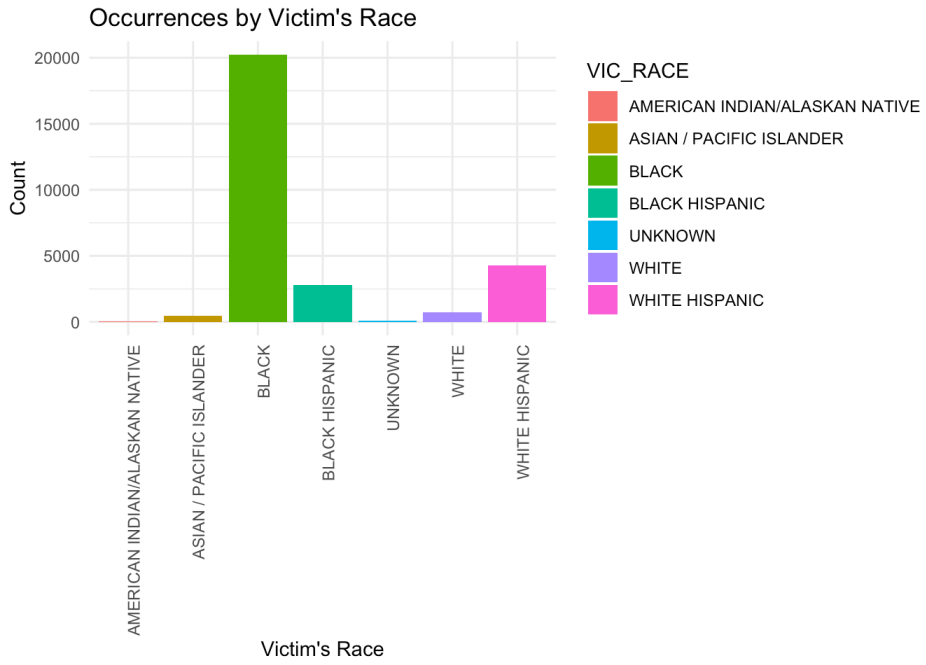
When reviewing the data plotted against age, I found an input error for one observation. One person's age was listed as 1022 years old. I decided to remove the observation entirely since there were so many observations in the data set and this was clearly an error.

```
data <- data %>%
  filter(VIC_AGE_GROUP != "1022")
```

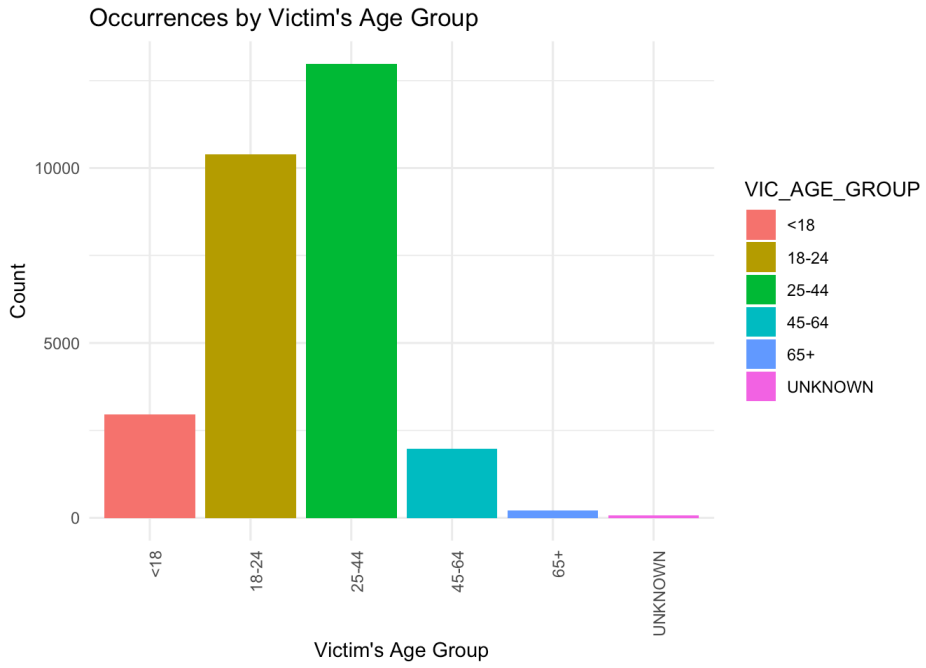
Analysis Part 1

For my analysis, I wanted to get a better sense of who the victims of gun violence were as far as NYC is concerned. I wanted to break this down in a few different ways: age, sex, and race. Using the "ggplot2" package I wrote the following code to visualize different parts of the data.

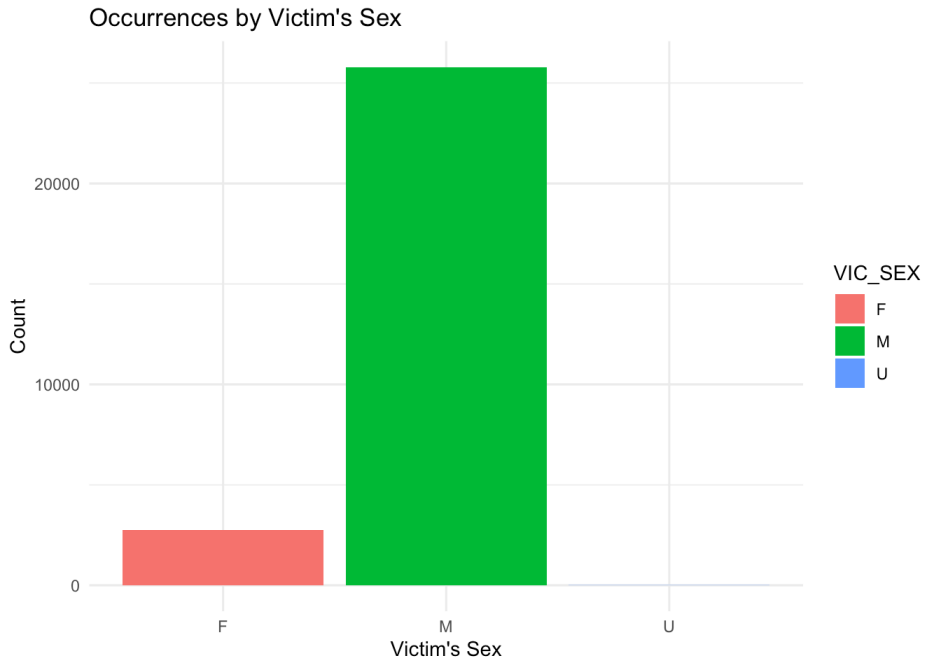
```
ggplot(data, aes(x = VIC_RACE, fill = VIC_RACE)) +
  geom_bar() +
  labs(title = "Occurrences by Victim's Race", x = "Victim's Race", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
ggplot(data, aes(x = VIC_AGE_GROUP, fill = VIC_AGE_GROUP)) +
  geom_bar() +
  labs(title = "Occurrences by Victim's Age Group", x = "Victim's Age Group", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
ggplot(data, aes(x = VIC_SEX, fill = VIC_SEX)) +
  geom_bar() +
  labs(title = "Occurrences by Victim's Sex", x = "Victim's Sex", y = "Count") +
  theme_minimal()
```



We can observe several trends based on these results. Each one deserves it's own investigation and shouldn't be taken at face value without additional research.

- The overwhelming majority of victims in this data set were labeled as black.
- The majority of victims in this data set were between the ages of 18 and 44. The 18-24 and 24-44 year old groups were much larger than the other age brackets.
- Almost all victims in the data set were labeled as male.

Analysis Part 2

I wanted to see how the quantity of shootings in NYC was changing over time. To do this, I wanted to plot those observations against time (or days since the start of the observation period).

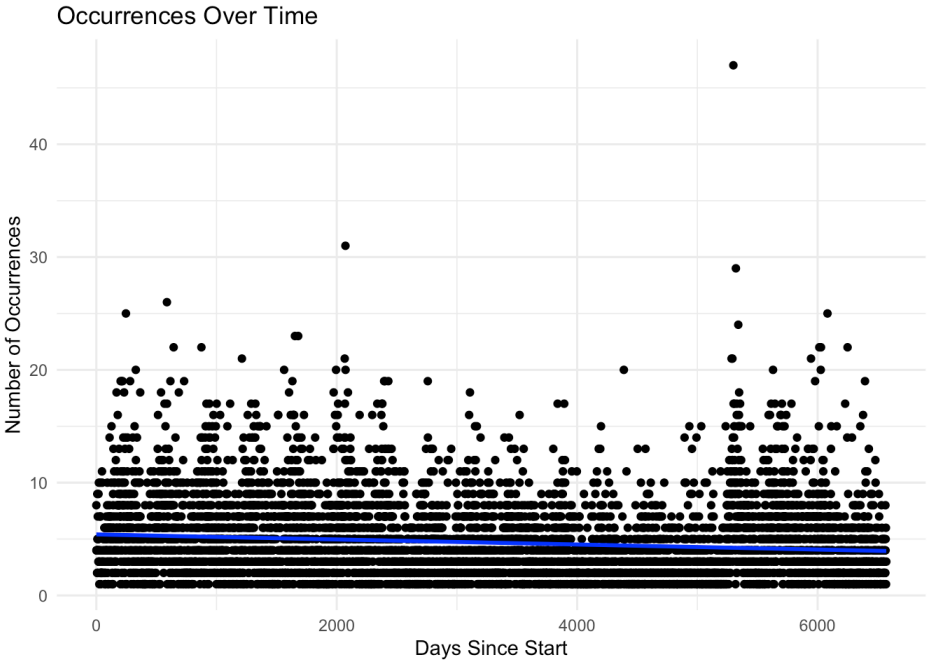
1. I started by modifying the existing data set to include a variable to track how many days into the observation period each data point was.
2. I got a sum of how many observations there were by each date.
3. I set up a linear model using the sum of observations by date and the days into the observation period.
4. I attempted to plot this data, but there were so many observations that it was a little disorienting.
5. I went back and grouped the data by week, which was much easier to understand when plotted.

```
data$days_since_start <- as.numeric(data$OCCUR_DATE - min(data$OCCUR_DATE))

daily_data <- data %>%  #grouping by date and getting sum
  group_by(OCCUR_DATE) %>%
  summarize(daily_count = n())

daily_data$days_since_start <- as.numeric(daily_data$OCCUR_DATE - min(daily_data$OCCUR_DATE))

model <- lm(daily_count ~ days_since_start, data = daily_data) #setting up model

summary(model) #checking model summary stats
```

```
##
## Call:
## lm(formula = daily_count ~ days_since_start, data = daily_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.408 -2.443 -0.977  1.645 42.771
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.414e+00  8.914e-02  60.734   <2e-16 ***
## days_since_start -2.236e-04  2.361e-05  -9.472   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.528 on 6093 degrees of freedom
## Multiple R-squared:  0.01451,    Adjusted R-squared:  0.01435
## F-statistic: 89.72 on 1 and 6093 DF,  p-value: < 2.2e-16
```

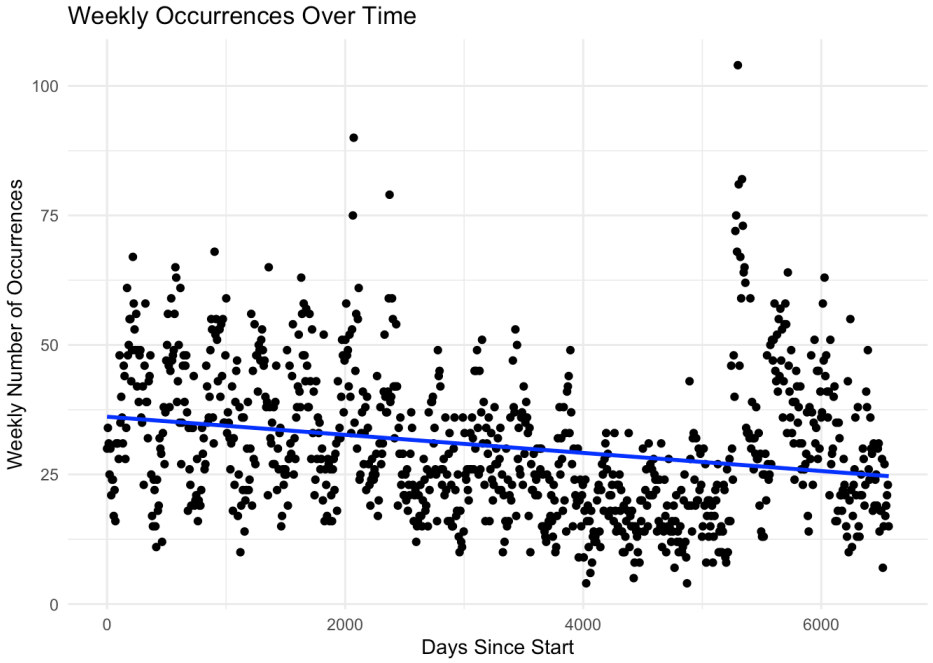
```
ggplot(daily_data, aes(x = days_since_start, y = daily_count)) +
  geom_point() +  #plotting data based on days
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add the linear model line
  labs(title = "Occurrences Over Time", x = "Days Since Start", y = "Number of Occurrences") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
daily_data_weekly <- data %>% #grouping by week and getting sum
  mutate(week = floor_date(OCCUR_DATE, "week")) %>%
  group_by(week) %>%
  summarize(weekly_count = n())

daily_data_weekly$days_since_start <- as.numeric(daily_data_weekly$week - min(daily_data_weekly$week))

ggplot(daily_data_weekly, aes(x = days_since_start, y = weekly_count)) +
  geom_point() + #plotting data based on days
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add the linear model line
  labs(title = "Weekly Occurrences Over Time", x = "Days Since Start", y = "Weekly Number of Occurrences") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Based on these plots, we can see that the number of observed shootings in NYC has gone down slightly over the observation period. One interesting aspect to the graphs is the apparent outliers. There seem to be surges in shooting activity, but also other points that seem more random or disconnected.

Conclusion & Bias

By evaluating this data set several notable findings were produced. Some of the more eyecatching ones include the number of shooting incidents decreasing over time, Black and Black-Hispanic populations being over represented in the victim statistics, and a good idea of what age groups are most at risk for being victims of these shootings. These findings are only cursory, and could serve as “jumping off” points for more detailed and informed investigations. Additional information would greatly benefit these findings as they would add context and start to paint a fuller picture. NYC general population statistics, geographical information on the areas considered for this data, and clearer explanations for what defines a “shooting incident” could all help in drawing better conclusions.

There are several different biases that should be considered when looking at this data set. Some of the more notable are the following:

1. **Measurement Bias** - due to the nature of shootings, getting accurate data can be extremely difficult. Unfortunately, there are probably a considerable amount of shootings that haven’t been recorded here because they were covered up, impossible to track, or were just never reported.
2. **Observer Bias -** This data set was compiled and put together by humans working for the City of New York. Anytime you have human involvement you get a certain level of bias. This can vary dramatically in its effect. It is difficult to say to what extent observer bias has touched this data without doing a much more detailed investigation.
3. **Confounding Bias** - Shootings and the factors that go into them are extremely complicated. There are a lot of different factors that go into shootings in NYC other than the variables that were covered in the data. This makes it hard to draw conclusions because while we may have the end result captured in the data, trying to see how we got to that end result is extremely difficult. In other words, we can confidently say the “what” but not the “why”.