NYPD Shootings

Loren Forrester

5/17/2022

Importing Data

Imports a dataset detailing NYPD shooting incidents between 2006 and 2021.

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

Cleaning Data

Focusing on borough and perpetrator/victim demographics, converts relevant variables to date/factor objects, omits irrelevant columns and incomplete cases including those with unknown and anomalous data entries. Displays a summary of the cleaned dataset.

```
nypd <- nypd %>%
  mutate(date = mdy(OCCUR_DATE)) %>%
  mutate(year = as.factor(year(date))) %>%
  mutate(month = as.factor(month(date))) %>%
  mutate(borough = as.factor(BORO)) %>%
  filter(PERP_AGE_GROUP != 'UNKNOWN') %>%
  filter(PERP_RACE != 'UNKNOWN') %>%
  filter(VIC_AGE_GROUP != 'UNKNOWN') %>%
  filter(VIC_RACE != 'UNKNOWN') %>%
  filter(PERP_SEX != 'U') %>%
  filter(VIC_SEX != 'U') %>%
  filter(PERP_AGE_GROUP != '1020') %>%
  filter(PERP_AGE_GROUP != '940') %>%
  filter(PERP_AGE_GROUP != '224') %>%
  filter(complete.cases(.)) %>%
  mutate(perp_age = as.factor(PERP_AGE_GROUP)) %>%
  mutate(perp_sex = as.factor(PERP_SEX)) %>%
  mutate(perp_race = as.factor(PERP_RACE)) %>%
  mutate(vic_age = as.factor(VIC_AGE_GROUP)) %>%
  mutate(vic_sex = as.factor(VIC_SEX)) %>%
  mutate(vic_race = as.factor(VIC_RACE)) %>%
  select(c(date, year, month, borough, perp_age, perp_sex, perp_race, vic_age, vic_sex, vic_race))

summary(nypd)
```

```
##       date                 year          month               borough
##  Min.   :2006-01-01   2006   : 625   8      : 594   BRONX       :1654
##  1st Qu.:2008-09-04   2008   : 529   6      : 549   BROOKLYN    :2190
##  Median :2011-09-02   2007   : 489   7      : 539   MANHATTAN   : 865
##  Mean   :2012-07-29   2011   : 489   5      : 535   QUEENS      : 829
##  3rd Qu.:2016-01-31   2009   : 467   1      : 506   STATEN ISLAND: 242
##  Max.   :2021-12-31   2010   : 466   9      : 501
##                       (Other):2715   (Other):2556
##   perp_age    perp_sex                          perp_race     vic_age
##  <18  : 593   F: 184   AMERICAN INDIAN/ALASKAN NATIVE:   1   <18  : 609
##  18-24:2551   M:5596   ASIAN / PACIFIC ISLANDER      :  59   18-24:2048
##  25-44:2351            BLACK                         :4337   25-44:2598
##  45-64: 247            BLACK HISPANIC                : 445   45-64: 471
##  65+  :  38            WHITE                         : 137   65+  :  54
##                        WHITE HISPANIC                : 801
##
##  vic_sex                          vic_race
##  F: 788   AMERICAN INDIAN/ALASKAN NATIVE:   2
##  M:4992   ASIAN / PACIFIC ISLANDER      :  95
##           BLACK                         :3972
##           BLACK HISPANIC                : 566
##           WHITE                         : 207
##           WHITE HISPANIC                : 938
##
```

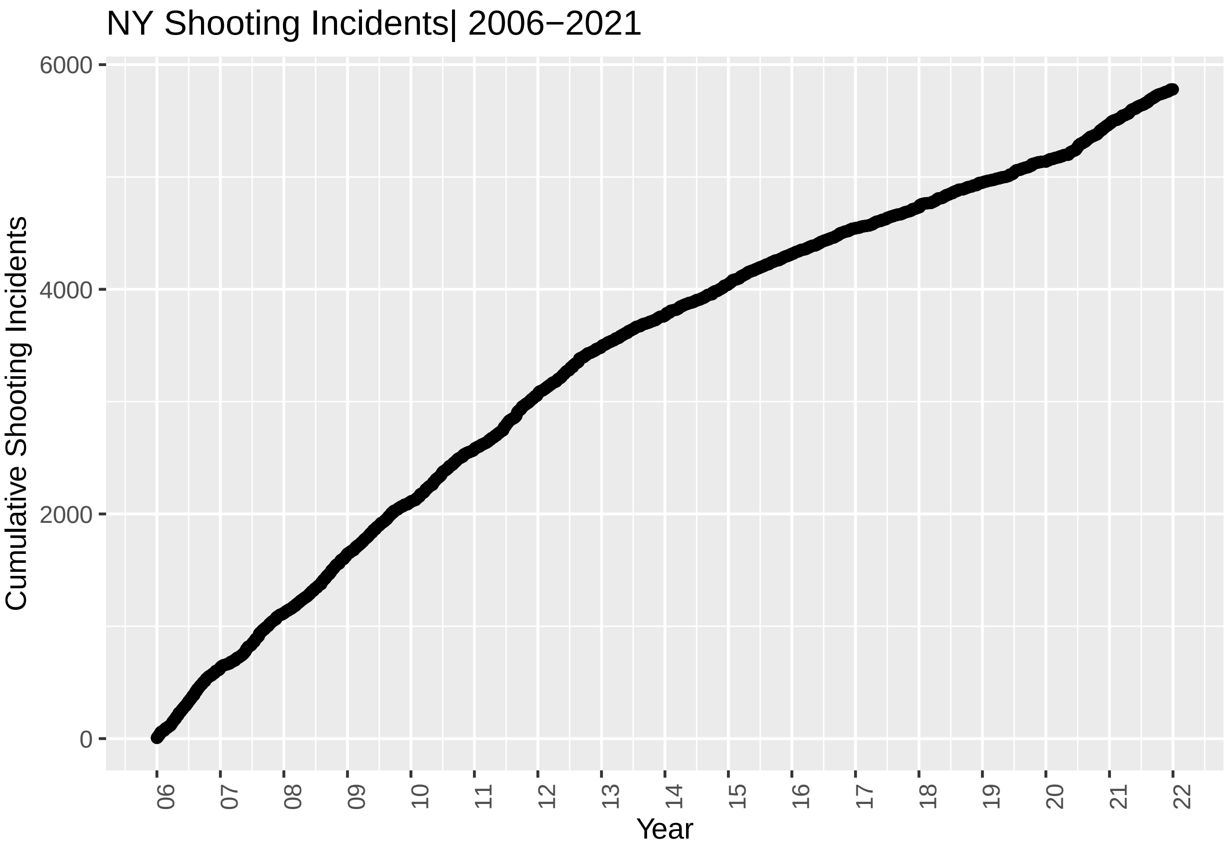
Visualizing Data

Here we'll take a look at a few visualizations of this dataset, first plotting all shootings cumulatively over time, then also examining shootings by borough, as well as by a couple of different demographics.

All Shootings

```
nypd_all <- nypd %>%
  group_by(date) %>%
  summarize(incidents = n()) %>%
  mutate(cumulative_incidents = cumsum(incidents)) %>%
  mutate(days_elapsed = as.numeric(difftime(date, date[[1]], units = 'days'))) %>%
  ungroup()
```

```
ggplot(nypd_all, aes(x = date, y = cumulative_incidents)) +
  geom_point() +
  theme(legend.position = 'right', axis.text.x = element_text(angle = 90)) +
  xlab('Year') +
  ylab('Cumulative Shooting Incidents') +
  labs(title = 'NY Shooting Incidents| 2006-2021') +
  scale_x_date(date_breaks = 'year', labels = date_format('%y'))
```

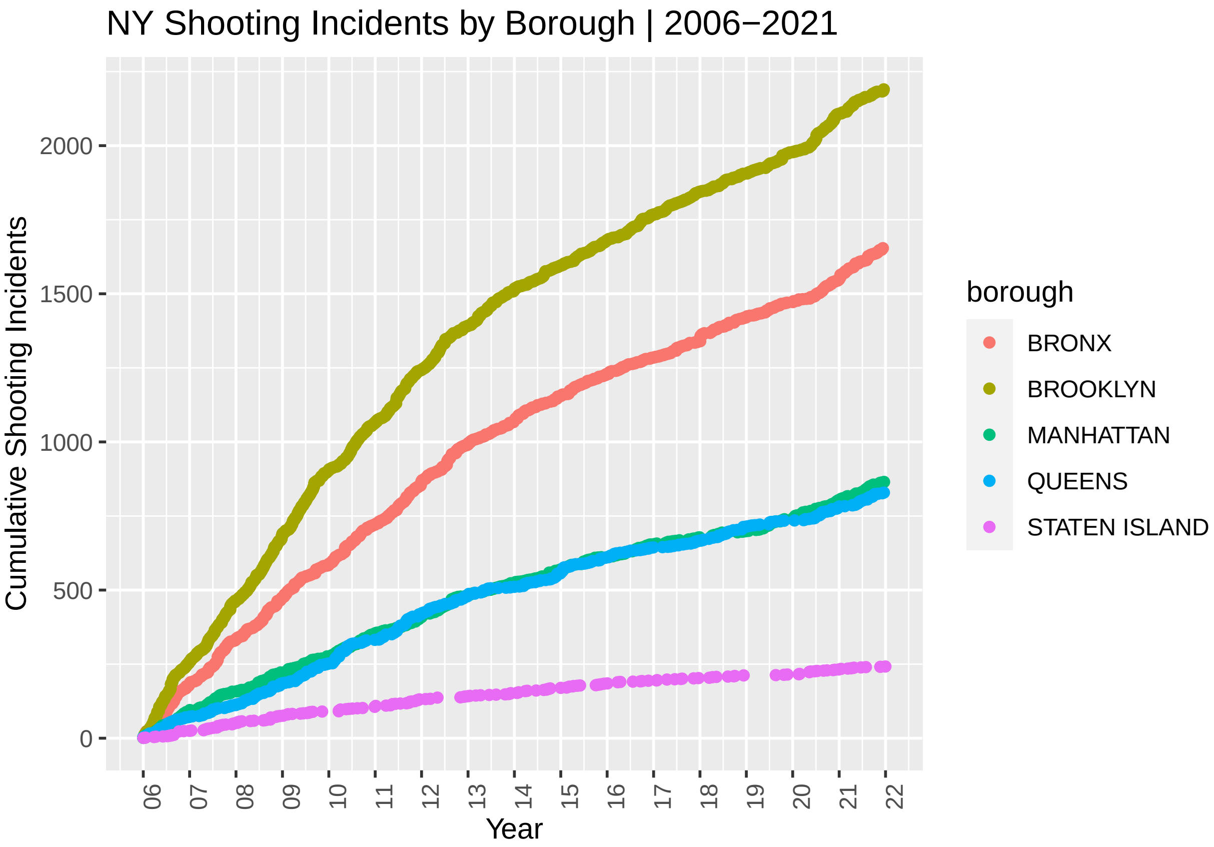


By Borough

Plots shootings in New York by borough over time.

```
nypd_by_borough <- nypd %>%
  group_by(borough, date) %>%
  summarize(incidents = n()) %>%
  mutate(cumulative_incidents = cumsum(incidents)) %>%
  ungroup()
```

```
ggplot(nypd_by_borough, aes(x = date, y = cumulative_incidents, color = borough)) +
  geom_point() +
  theme(legend.position = 'right', axis.text.x = element_text(angle = 90)) +
  xlab('Year') +
  ylab('Cumulative Shooting Incidents') +
  labs(title = 'NY Shooting Incidents by Borough | 2006-2021') +
  scale_x_date(date_breaks = 'year', labels = date_format('%y'))
```

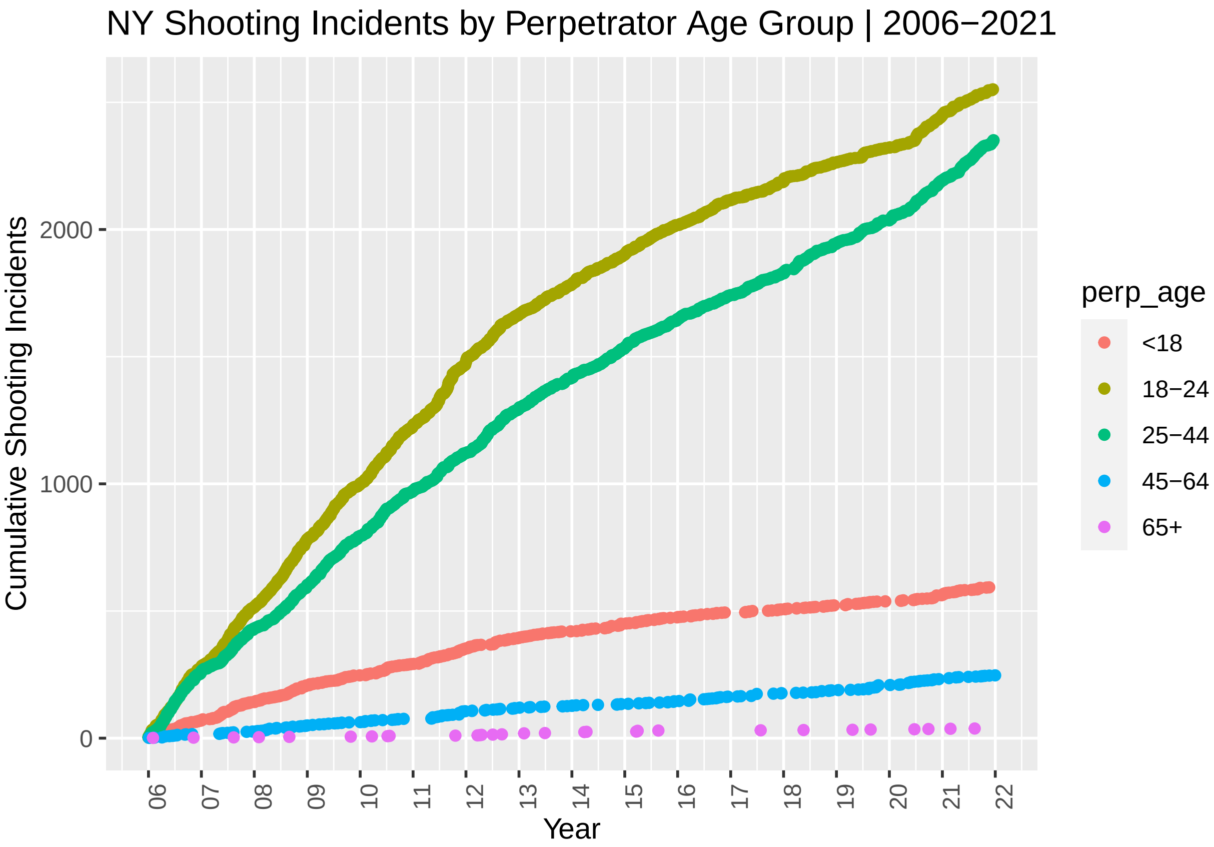


By Age Group

Plots shootings in New York by perpetrator age group over time.

```
nypd_by_perp_age <- nypd %>%
  group_by(perp_age, date) %>%
  summarize(incidents = n()) %>%
  mutate(cumulative_incidents = cumsum(incidents)) %>%
  ungroup()
```

```
ggplot(nypd_by_perp_age, aes(x = date, y = cumulative_incidents, color = perp_age)) +
  geom_point() +
  theme(legend.position = 'right', axis.text.x = element_text(angle = 90)) +
  xlab('Year') +
  ylab('Cumulative Shooting Incidents') +
  labs(title = 'NY Shooting Incidents by Perpetrator Age Group | 2006-2021') +
  scale_x_date(date_breaks = 'year', labels = date_format('%y'))
```

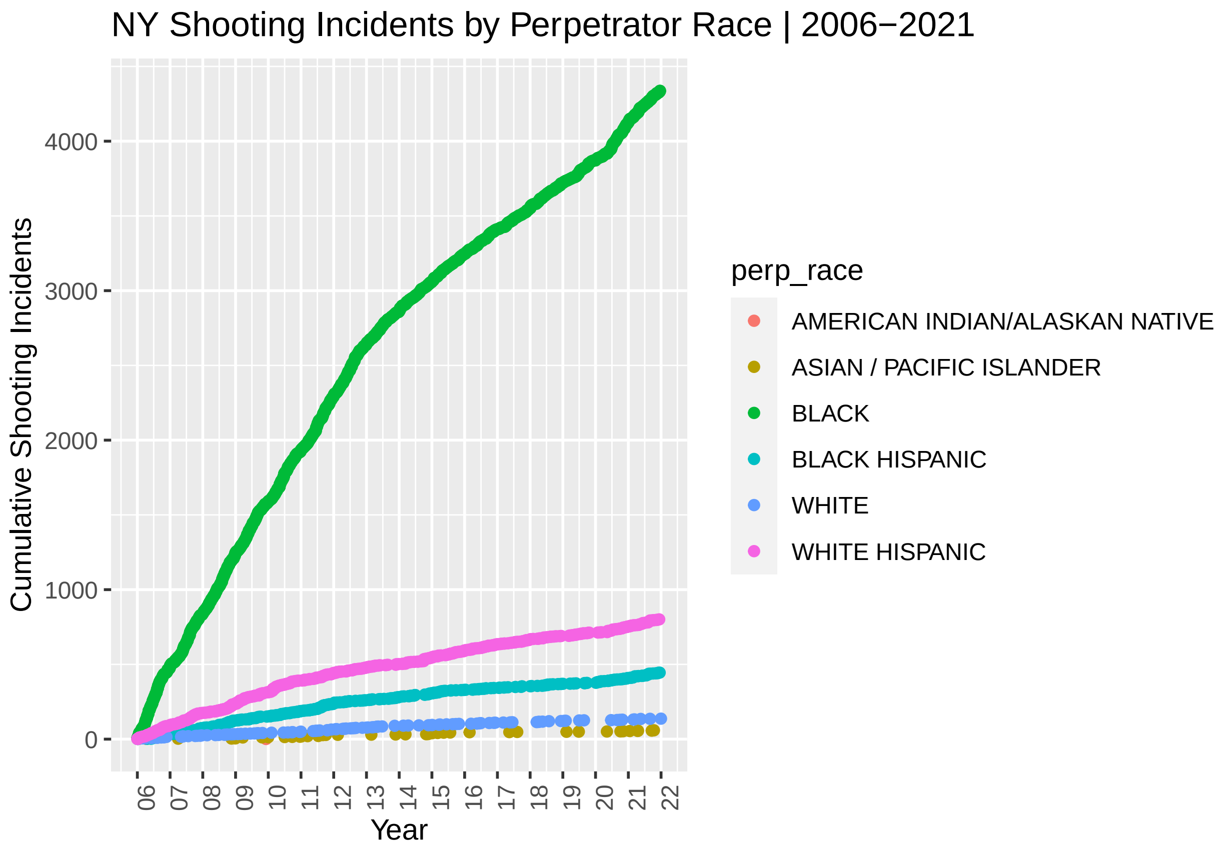


By Race

Plots shootings in New York by perpetrator race over time.

```
nypd_by_perp_race <- nypd %>%
  group_by(perp_race, date) %>%
  summarize(incidents = n()) %>%
  mutate(cumulative_incidents = cumsum(incidents)) %>%
  ungroup()
```

```
ggplot(nypd_by_perp_race, aes(x = date, y = cumulative_incidents, color = perp_race)) +
  geom_point() +
  theme(legend.position = 'right', axis.text.x = element_text(angle = 90)) +
  xlab('Year') +
  ylab('Cumulative Shooting Incidents') +
  labs(title = 'NY Shooting Incidents by Perpetrator Race | 2006-2021') +
  scale_x_date(date_breaks = 'year', labels = date_format('%y'))
```



Modeling Data

From the first of these plots, it appears shootings over time have been approximately linear, so let's examine that as a model.

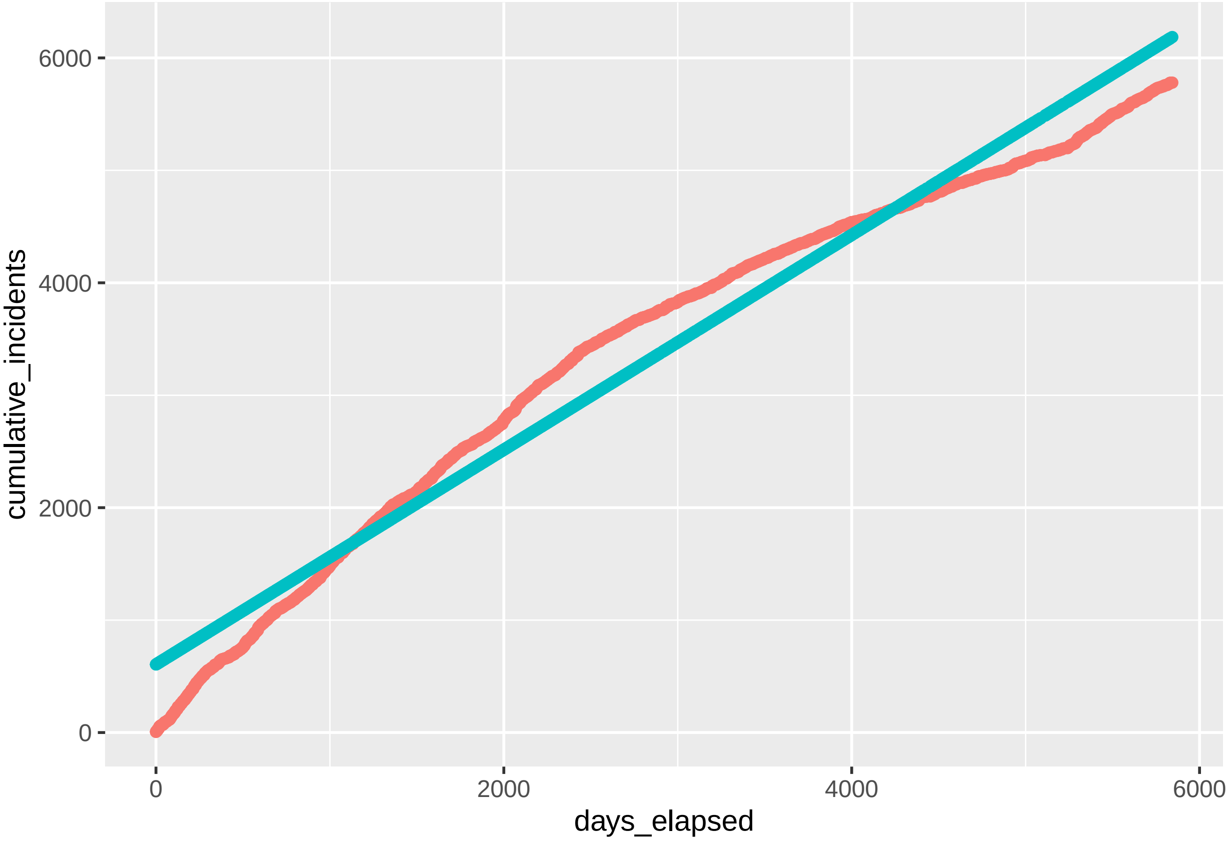
```
mod <- lm(cumulative_incidents ~ days_elapsed, data = nypd_all)
summary(mod)
```

```
##
## Call:
## lm(formula = cumulative_incidents ~ days_elapsed, data = nypd_all)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -599.91 -291.68   29.27  268.50  457.89
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 605.91482   10.23577    59.2   <2e-16 ***
## days_elapsed   0.95488    0.00331   288.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 299.1 on 2787 degrees of freedom
## Multiple R-squared:  0.9676, Adjusted R-squared:  0.9676
## F-statistic: 8.32e+04 on 1 and 2787 DF,  p-value: < 2.2e-16
```

These results do seem to indicate a fairly tight linear correlation, so let's go ahead and plot the linear model's expected results against the real values.

```
nypd_model <- nypd_all %>%
  mutate(pred = predict(mod))

nypd_model %>% ggplot() +
  geom_point(aes(x = days_elapsed, y = cumulative_incidents, color = 'blue')) +
  geom_point(aes(x = days_elapsed, y = pred, color = 'red')) +
  theme(legend.position = 'none')
```



This is actually quite interesting, as we can see that while most of the data does indeed fit close to the model, the datapoints themselves follow a distinct curve for most of the dataset before suddenly seeming to find a slope almost exactly parallel to the predicted slope only toward the very tail end of the dataset. This represents a significant deviation from how the data has performed relative to the model previously and is what I'll seek to examine in the next section.

Analyzing Data

From our visualizations, we can see that almost across the board (regardless of demographics and location), there seems to be a large uptick in the number of shooting deaths sometime during 2020, which could coincide with the nationwide protests and civil unrest that occurred following the body-cam footage of the death of George Floyd being released on May 25, 2020. To see if that bears out, we'll take a look at the maximum number of cases by day to see if we're correct in the assumption that the weeks immediately followed saw among the highest frequency of shooting incidents present in the dataset.

```
shooting_frequency <- nypd %>%
  group_by(year, month) %>%
  summarize(incidents = n()) %>%
  arrange(desc(incidents)) %>%
  ungroup()

shooting_frequency
```

```
## # A tibble: 192 x 3
##    year  month incidents
##    <fct> <fct>     <int>
##  1 2006  8            71
##  2 2007  8            67
##  3 2006  1            65
##  4 2006  5            64
##  5 2006  7            64
##  6 2006  4            62
##  7 2008  8            62
##  8 2011  9            62
##  9 2007  5            58
## 10 2008  9            58
## # ... with 182 more rows
```

Interestingly, no month in 2020 represented a high-water mark for shooting frequency, almost all which perhaps should have been evident based on the plots seeming to level off over time, especially after 2011-2012. In fact, each of the months which represented the top fifty shooting frequencies in New York occurred prior to 2013. However, the spike around the year 2020 still seems to be interesting, so let's look at just that portion of the data to see if the jump really was as drastic as the visualizations make it seem.

```
shooting_frequency_2020 <- shooting_frequency %>%
  filter(year == 2020) %>%
  arrange(month)

shooting_frequency_2020
```

```
## # A tibble: 12 x 3
##    year  month incidents
##    <fct> <fct>     <int>
##  1 2020  1            23
##  2 2020  2            12
##  3 2020  3            13
##  4 2020  4            12
##  5 2020  5            30
##  6 2020  6            37
##  7 2020  7            38
##  8 2020  8            38
##  9 2020  9            28
## 10 2020  10           30
## 11 2020  11           38
## 12 2020  12           34
```

Clearly, 2020 did see a marked increase in the number of reported shootings in New York (a >100% increase from April to May) and doesn't immediately see any kind of meaningful drop-off thereafter. So, does this represent the largest shift in the trend of shooting frequencies in the dataset? Let's take a look at how the months in question compare to overall to the six month averages of the months leading up to them.

```
roll <- function(x, n) {
  if (length(x) <= n) NA
```

```
   else rollapply(x, list(-seq(n)), mean, fill = NA)
}

frequency_vs_avg <- shooting_frequency %>%
  arrange(year, month) %>%
  mutate(lagging_6m_avg = roll(incidents, 6)) %>%
  mutate(relative_increase = incidents/lagging_6m_avg - 1) %>%
  arrange(desc(relative_increase))

frequency_vs_avg
```

```
## # A tibble: 192 x 5
##    year  month incidents lagging_6m_avg relative_increase
##    <fct> <fct>     <int>          <dbl>             <dbl>
##  1 2020  5            30           12.3             1.43
##  2 2020  6            37           15.7             1.36
##  3 2018  1            35           15.5             1.26
##  4 2019  6            28           13.2             1.13
##  5 2017  4            26           13.5             0.926
##  6 2007  5            58           30.3             0.912
##  7 2014  1            38           20.3             0.869
##  8 2020  7            38           21.2             0.795
##  9 2019  7            26           14.5             0.793
## 10 2010  5            58           33.5             0.731
## # ... with 182 more rows
```

Conclusion and Bias Identification

As the results here indicate, the months following the George Floyd incident represented not only *a* large uptick in shootings, but the highest (+143%) and second-highest (+136%) increases in shootings in New York relative to a six-month lagging-average, with the following month of July representing the eighth-highest (+79.5%) increase in such incidents, making it period of largest deviation from the overall trends presented in the dataset.

While these results are notable for their significant deviation from larger trends and uniquely trace back to a single nexus event, it is also imperative to acknowledge the inherent biases in the NYPD being the reporting entity of the raw data itself. The NYPD's issues with race-relations and misreporting are both longstanding and well-documented, so it is entirely possible the PD over-reported data during this period given the racial underpinnings of the events that preceded it.