NYPD Shooting Inicident Data Historic

Steven Hobbs

2024-10-12

Data Import and Wrangling

```
nypd <- read_csv("NYPD_Shooting_Incident_Data__Historic_.csv",
                 col_types = "ncnfnnclfffffffnnnnc")

names(nypd) = c('key', 'date', 'time', 'borough', 'precinct', 'jcode', 'loc_desc',
                'sm_flag', 'perp_age', 'perp_sex', 'perp_race', 'vic_age', 'vic_sex',
                'vic_race', 'xcoord', 'ycoord', 'lat', 'long', 'lon_lat')

nypd <-
    nypd %>%
    mutate(date = mdy(date),
                 day = day(date),
                 month = month(date),
                 year = year(date),
           vic_age = fct_relevel(vic_age, "<18", "18-24",
                                 "25-44", "45-64", "65+"),
           perp_age = fct_relevel(perp_age, "<18", "18-24",
                                  "25-44", "45-64", "65+")) %>%
        select(time, day, month, date, year, borough, precinct, perp_age:vic_race)

nypd |>
    select(1:5) |>
    summary() |> kable() |> kable_styling()
```

time	day	month	date	year
Min. : 0.00	Min. : 1.00	Min. : 1.000	Min. :2006-01-01	Min. :2006
1st Qu.: 3.00	1st Qu.: 8.00	1st Qu.: 5.000	1st Qu.:2009-05-10	1st Qu.:2009
Median :15.00	Median :16.00	Median : 7.000	Median :2012-08-26	Median :2012
Mean :12.19	Mean :15.95	Mean : 6.857	Mean :2013-06-13	Mean :2013
3rd Qu.:20.00	3rd Qu.:24.00	3rd Qu.: 9.000	3rd Qu.:2017-07-01	3rd Qu.:2017
Max. :23.00	Max. :31.00	Max. :12.000	Max. :2021-12-31	Max. :2021

```
nypd |>
    select(6:10) |>
    summary() |> kable() |> kable_styling()
```

borough	precinct	perp_age	perp_sex	perp_race
BROOKLYN :10365	Min. : 1.00	18-24 :5844	M :14416	BLACK :10668

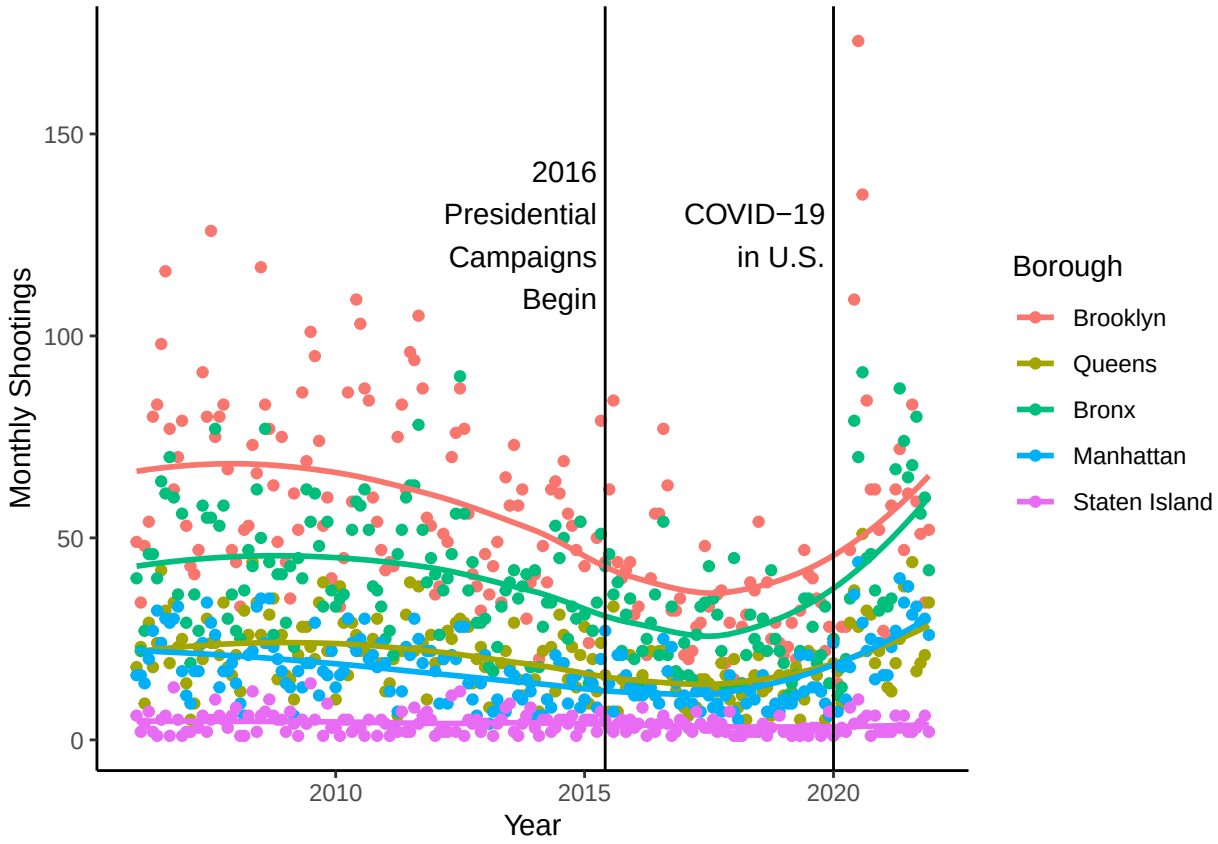
QUEENS : 3828	1st Qu.: 44.00	25-44 :5202	F : 371	WHITE HISPANIC: 2164
BRONX : 7402	Median : 69.00	UNKNOWN:3148	U : 1499	UNKNOWN : 1836
MANHATTAN : 3265	Mean : 65.87	<18 :1463	NA's: 9310	BLACK HISPANIC: 1203
STATEN ISLAND: 736	3rd Qu.: 81.00	45-64 : 535	NA	WHITE : 272
NA	Max. :123.00	(Other): 60	NA	(Other) : 143
NA	NA	NA's :9344	NA	NA's : 9310

```
nypd |>
    select(11:ncol(nypd)) |>
    summary() |> kable() |> kable_styling()
```

vic_age	vic_sex	vic_race
<18 : 2681	M:23182	BLACK :18281
18-24 : 9604	F: 2403	ASIAN / PACIFIC ISLANDER : 354
25-44 :11386	U: 11	BLACK HISPANIC : 2485
45-64 : 1698	NA	WHITE HISPANIC : 3742
65+ : 167	NA	WHITE : 660
UNKNOWN: 60	NA	AMERICAN INDIAN/ALASKAN NATIVE: 9
NA	NA	UNKNOWN : 65

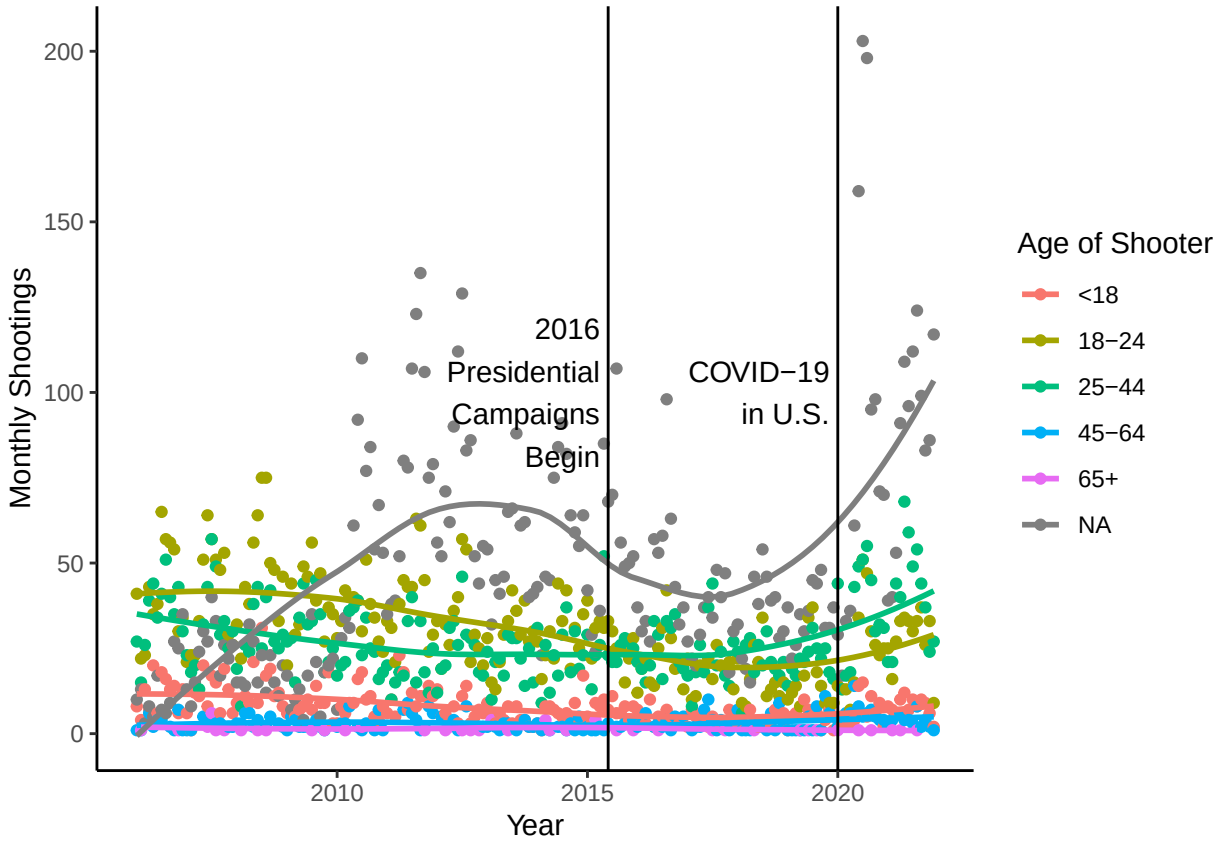
Shootings in New York by Borough

```
nypd %>%
    group_by(borough, year, month) %>%
    summarize(shootings = n()) %>%
    mutate(date2 = make_date(year, month)) %>%
    ggplot(aes(x = date2, y = shootings, color = borough)) +
    geom_point() +
    geom_smooth(method = "loess", formula = 'y~x', se = FALSE) +
    geom_vline(xintercept = as.Date("2020-01-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2020-01-01"), y = 125,
             label = "COVID-19 \n in U.S. ", hjust = "right") +
    geom_vline(xintercept = as.Date("2015-06-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2015-06-01"), y = 125,
             label = "2016 \n Presidential \n Campaigns \n Begin ",
             hjust = "right") +
    scale_color_discrete(name = "Borough",
                         labels = c("Brooklyn", "Queens", "Bronx",
                                    "Manhattan", "Staten Island")) +
    labs(x = "Year", y = "Monthly Shootings") +
    theme_classic()
```



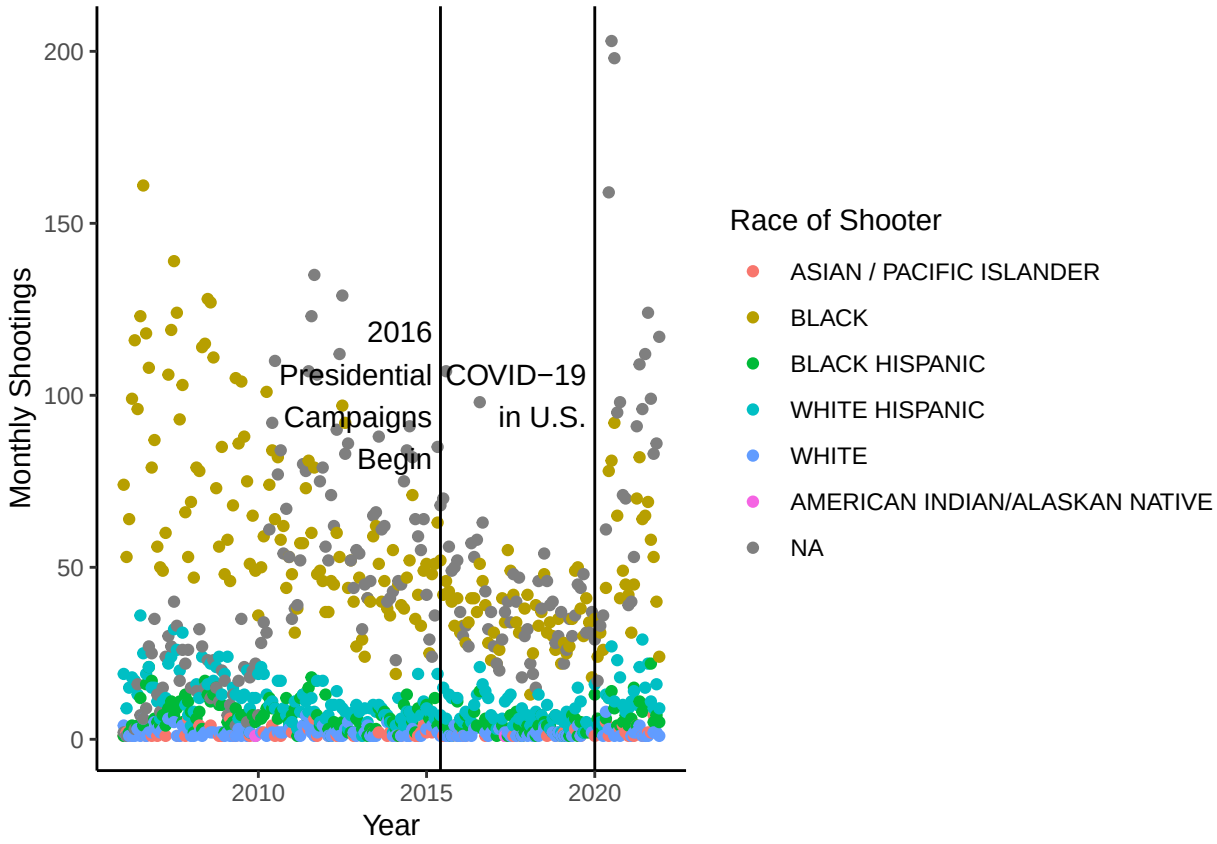
Shootings in New York Boroughs by Age

```
nypd %>%
    filter(!(perp_age %in% c("", "UNKNOWN", "1020", "940", "224"))) %>%
    droplevels %>%
    group_by(year, month, perp_age) %>%
    summarize(shootings = n()) %>%
    mutate(date2 = make_date(year, month)) %>%
    ggplot(aes(x = date2, y = shootings, color = perp_age)) +
    geom_point() +
    geom_smooth(method = "loess", formula = 'y~x', se = FALSE) +
    geom_vline(xintercept = as.Date("2020-01-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2020-01-01"), y = 100,
             label = "COVID-19 \n in U.S. ", hjust = "right") +
    geom_vline(xintercept = as.Date("2015-06-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2015-06-01"), y = 100,
             label = "2016 \n Presidential \n Campaigns \n Begin ",
             hjust = "right") +
    scale_color_discrete(name = "Age of Shooter") +
    labs(x = "Year", y = "Monthly Shootings") +
    theme_classic()
```



Shootings in New York Boroughs by Shooter Race

```
nypd %>%
    filter(!(perp_race %in% c("", "UNKNOWN"))) %>%
    droplevels %>%
    group_by(year, month, perp_race) %>%
    summarize(shootings = n()) %>%
    mutate(date2 = make_date(year, month)) %>%
    ggplot(aes(x = date2, y = shootings, color = perp_race)) +
    geom_point() +
    # geom_smooth(method = "loess", formula = 'y~x', se = FALSE) +
    # Insufficient data for smooth curves in most groups.
    geom_vline(xintercept = as.Date("2020-01-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2020-01-01"), y = 100,
             label = "COVID-19 \n in U.S. ", hjust = "right") +
    geom_vline(xintercept = as.Date("2015-06-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2015-06-01"), y = 100,
             label = "2016 \n Presidential \n Campaigns \n Begin ",
             hjust = "right") +
    scale_color_discrete(name = "Race of Shooter") +
    labs(x = "Year", y = "Monthly Shootings") +
    theme_classic()
```

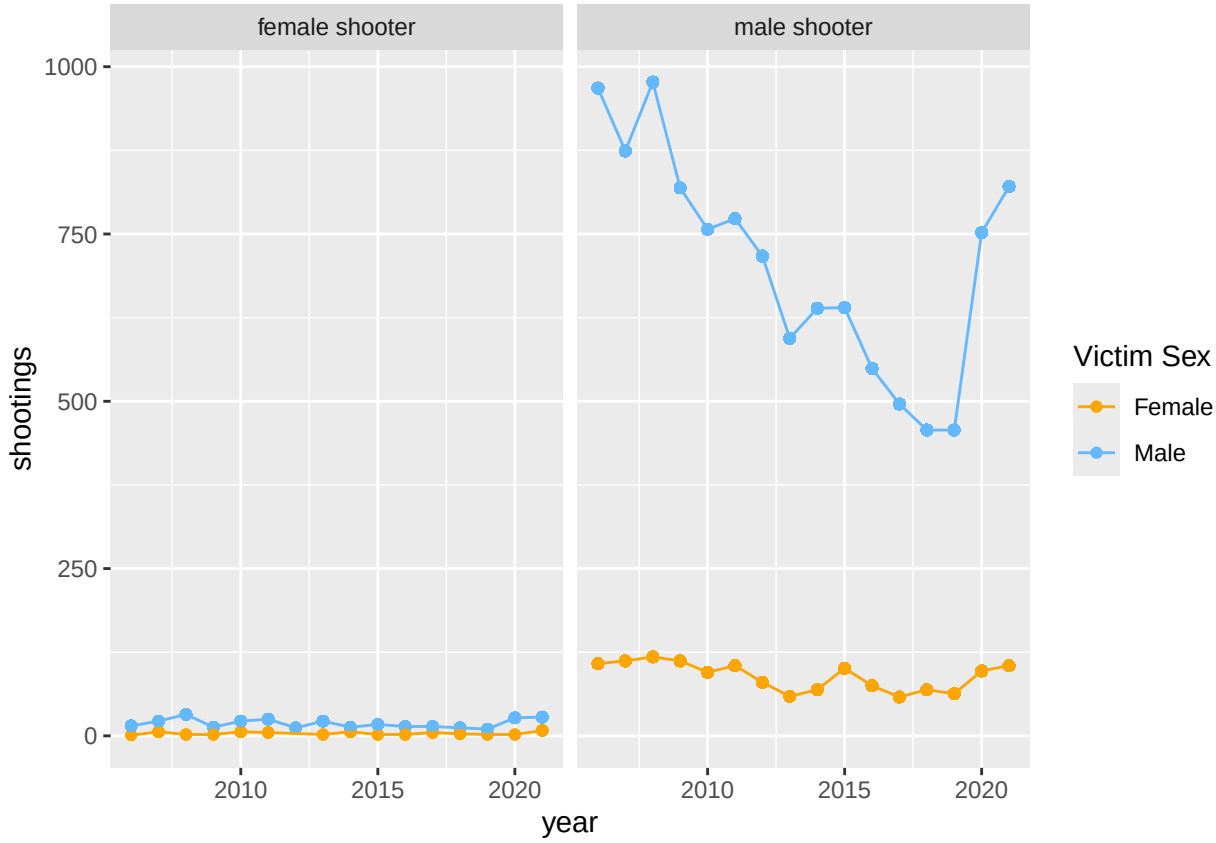


Shootings in New York Boroughs by Sex

```
nypd_sex <-
    nypd %>%
    filter(vic_sex != "U", perp_sex == "M" | perp_sex == "F",
           !(perp_age %in% c("1020", "940", "224", "UNKNOWN", ""))) %>%
    mutate(perp_sex = fct_recode(perp_sex,
                                 "male shooter" = "M",
                                 "female shooter" = "F"),
           perp_sex = fct_relevel(perp_sex, 'female shooter'),
           vic_sex = fct_relevel(vic_sex, 'F')) %>%
    droplevels %>%
    group_by(year, perp_sex, vic_sex) %>%
    mutate(shootings = n()) %>%
    ungroup %>%
    group_by(year) %>%
    mutate(year_total = sum(shootings)) %>%
    ungroup %>%
    select(year, perp_sex, vic_sex, perp_age, vic_age, shootings, year_total)

nypd_sex %>%
    ggplot(aes(x = year)) +
    geom_point(aes(y = shootings, color = vic_sex)) +
    geom_line(aes(y = shootings, color = vic_sex)) +
    scale_color_manual(name = "Victim Sex",
```

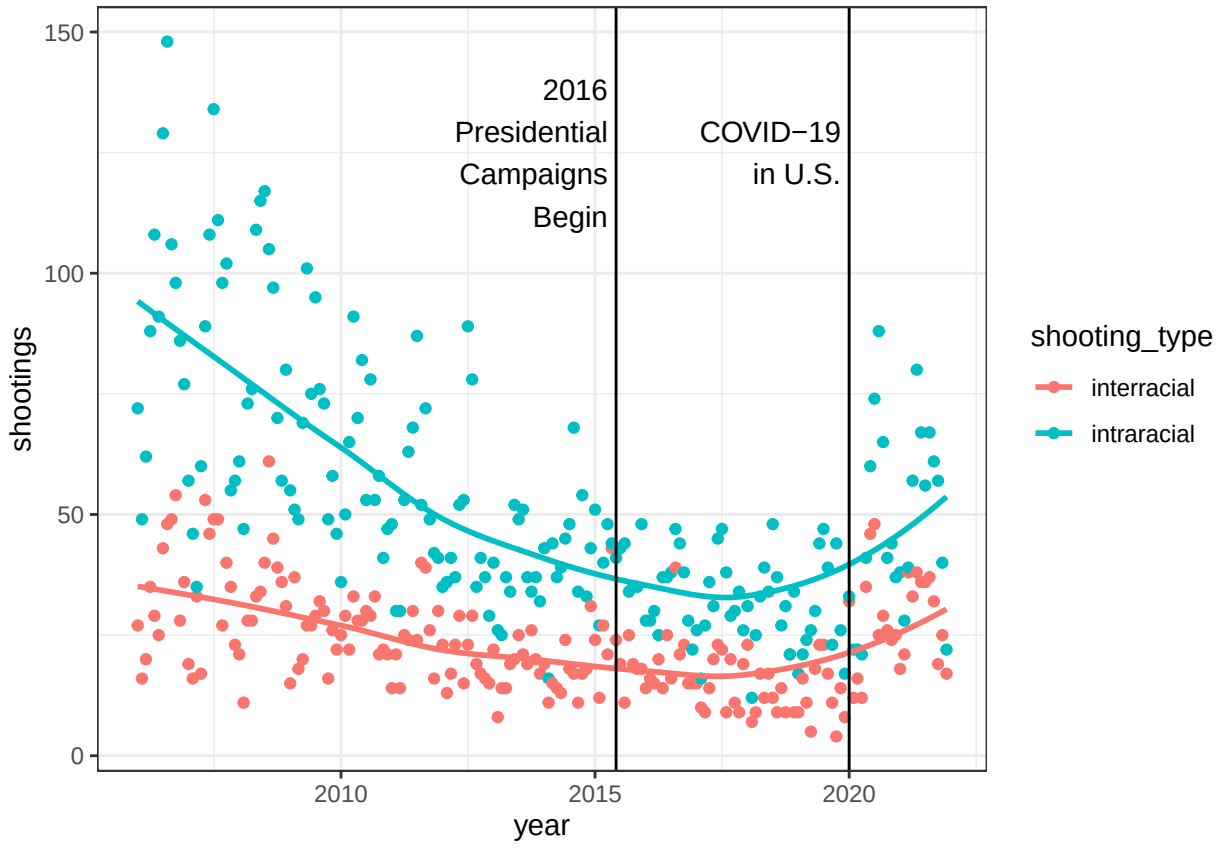
```
                      values = c("orange", "steelblue1"),
                      labels = c("Female", "Male")) +
    facet_wrap(~ perp_sex)
```



Shootings in New York Boroughs Within (intra) and Across (inter) Racial Groups.

```
nypd %>%
    filter(perp_race != "", perp_race != "UNKNOWN", vic_race != "UNKNOWN") %>%
    droplevels %>%
    mutate(shooting_type = ifelse(perp_race == vic_race,
                                  "intraracial",
                                  "interracial"),
           date2 = make_date(year, month)) %>%
    group_by(date2, shooting_type) %>%
    summarize(shootings = n()) %>%
    ggplot(aes(x = date2, y = shootings, color = shooting_type)) +
    geom_point() +
    geom_smooth(method = 'loess', formula = y~x, se = FALSE) +
    geom_vline(xintercept = as.Date("2020-01-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2020-01-01"), y = 125,
             label = "COVID-19 \n in U.S. ", hjust = "right") +
    geom_vline(xintercept = as.Date("2015-06-01"), color = "black") +
    annotate(geom = "text", x = as.Date("2015-06-01"), y = 125,
             label = "2016 \n Presidential \n Campaigns \n Begin ",
```

```
            hjust = "right") +
    labs(x = "year", y = "shootings") +
    theme_bw()
```



Logistic regression model predicting victim sex from shooter and victim ages.

```
model <-
    nypd_sex %>%
    mutate(vic_sex = fct_relevel(vic_sex, "M")) %>% # Make "F" the predicted sex
    filter(vic_age != "UNKNOWN") %$%
    glm(vic_sex ~ perp_sex + perp_age + vic_age,
        family = binomial(logit))

summary(model)
```

```
##
## Call:
## glm(formula = vic_sex ~ perp_sex + perp_age + vic_age, family = binomial(logit))
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.20776    0.17727  -6.813 9.56e-12 ***
## perp_sexmale shooter -0.35429    0.15374  -2.305 0.021192 *
## perp_age18-24        -0.06856    0.09442  -0.726 0.467812
## perp_age25-44         0.11525    0.09768   1.180 0.238059
```

```
## perp_age45-64         0.37430    0.14812   2.527 0.011503 *
## perp_age65+           1.93107    0.29445   6.558 5.44e-11 ***
## vic_age18-24         -0.86550    0.08929  -9.693  < 2e-16 ***
## vic_age25-44         -0.71477    0.08722  -8.195 2.51e-16 ***
## vic_age45-64          0.13611    0.11019   1.235 0.216719
## vic_age65+            0.83304    0.22043   3.779 0.000157 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9209  on 13019  degrees of freedom
## Residual deviance: 8900  on 13010  degrees of freedom
## AIC: 8920
##
## Number of Fisher Scoring iterations: 5
```

```
# Predict the probability that the victim is male for perp_sex = "male shooter" across all levels of pe
male_shooters = data.frame(perp_sex = rep("male shooter", 9),
                           perp_age = rep(levels(nypd_sex$vic_age)[1:3]), 3,
                           vic_age = c(rep("<18", 3), rep("18-24", 3), rep("25-44", 3)))

temp_m = 100 * predict(model, newdata = male_shooters, type = "response") |> round(2)
temp_m = matrix(temp_m, nrow = 3, ncol = 3, byrow = TRUE)
colnames(temp_m) = paste("Perp Age", levels(nypd_sex$perp_age)[1:3])
rownames(temp_m) = paste("Vic Age", levels(nypd_sex$vic_age)[1:3])
temp_m |>
    kable(caption = "The Probability of a Female Victim for Male Shooters by age") |>
    kable_classic(full_width = F)
```

Table 4: The Probability of a Female Victim for Male Shooters by age

	Perp Age <18	Perp Age 18-24	Perp Age 25-44
Vic Age <18	17	16	19
Vic Age 18-24	8	8	9
Vic Age 25-44	9	9	10

```
# Predict the probability that the victim is female for perp_sex = "female shooter" across all levels o
female_shooters = data.frame(perp_sex = rep("female shooter", 9),
                             perp_age = rep(levels(nypd_sex$vic_age)[1:3]), 3,
                             vic_age = c(rep("<18", 3), rep("18-24", 3), rep("25-44", 3)))

temp_f = 100 * predict(model, newdata = female_shooters, type = "response") |> round(2)
temp_f = matrix(temp_f, nrow = 3, ncol = 3, byrow = TRUE)
colnames(temp_f) = paste("Perp Age", levels(nypd_sex$perp_age)[1:3])
rownames(temp_f) = paste("Vic Age", levels(nypd_sex$vic_age)[1:3])
temp_f |> kable(caption = "The Probability of a Female Victim for Female Shooters by age") |> kable_cla
```

Table 5: The Probability of a Female Victim for Female Shooters by age

	Perp Age <18	Perp Age 18-24	Perp Age 25-44

Vic Age <18	23	22	25
Vic Age 18-24	11	11	12
Vic Age 25-44	13	12	14

```
# Find the Odds ratio comparing the odds of a female victim for female perps to the odds of a female vi
OR = exp(coef(model)["perp_sexmale shooter"])^-1 |> round(2)
names(OR) = "Female perpetrator / Male perpetrator"
OR |> kable(col.names = "OR", caption = "Female Victim Odds Ratio") |> kable_classic(full_width = F)
```

Table 6: Female Victim Odds Ratio

	OR
Female perpetrator / Male perpetrator	1.43

Conclusions

From March, 2006 to November, 2021 monthly shootings in New York were consistently highest in Brooklyn, followed by Bronx, and then Queens and Manhattan, which were consistently similar, and lastly by Staten Island. All but Staten Island show a clear increasingly downward trend in monthly shootings going into mid 2015, at which point the direction of rate change changed, leveling off by 2017 and then rising into 2019 and 2020.

While the COVID-19 pandemic lock downs have been associated with various mental health challenges, such as depression and anxiety, these data show that increased shootings preceded the COVID-19 pandemic. The beginning of 2016 presidential campaign align closely with an inflection point in shootings and could be a source of aggravating factors and mental health problems connected to shootings. However, these data are correlative and offer few unequivocal conclusions. Due to the timing of events though, the COVID-19 pandemic could not possibly have caused the reversal in shooting rates. Other societal causes should be investigate, particularly in 2015 and 2016 when the direction of rate change switched from negative to positive.

Most shooting perpetrators were in the 18-24 and 25-44 age groups. In the late 2000's and early 2010's perpetrator rates were decreasing in both of these groups, but more dramatically for the youngest age group, 18-24. In the late 2010's and early 2020's shooting perpetrator rates increased in both groups, but this trend was stronger for the 25-44 age group, which took over as the highest rate group in late 2015.

Most shootings in New York boroughs were perpetrated by black men. However, these data do not contain population demographic or economic data, meaning we can not determine if, for example, the shooting rate per 1000 persons is comparable or different across the races, and we cannot determine how poverty and prosperity connect with shooting rates.

No obvious linear relationships were present, so I created a logistic regression model to explore if victim sex could be predicted from perpetrator sex, perpetrator age, and victim age. I created a prediction matrix for male and female perpetrators and the first three age groups. The last two age groups were ignored due to low sample size.

The results show that females are less likely to be victims of shootings at all ages and regardless of the perpetrator sex. However, female victims are about twice as common in the youngest age group, <18, compared to the older age groups. This trend is consistent across all perpetrator ages and for both male and female perpetrators. Additionally, the odds of a female victim are about 1.4 times higher for female perpetrators compared to male perpetrators.

A limitation imposed by the dataset is that suicides are not identifiable. In some cases the perpetrator and the victim may be the same individual. Additionally, the sample sizes in some groups may be very small,

such that even very small P-values may represent spurious false positive relationships. While further analysis, could verify sample sizes, it's time to wrap up this analysis and move on in the course!

Upon beginning this analysis, I was expecting to see an increase in shootings after the COVID-19 pandemic took hold in the spring of 2020 and I was expecting the 18-24 year old age group to be responsible for the uptick. I mitigated these biases by retaining all the data from all New York boroughs. The only data not included in any given figure are data with missing or erroneous values directly related to the plot. Also, the data have been minimally processed, and everything I've done to the data is fully transparent in this document.

I was surprised to learn that shootings began to increase well before the COVID-19 pandemic onset. Nearly all the observable changes in shooting rates appear to reflect a "volume" adjustment, rather than a "radio station" change, meaning the changes were more quantitative than qualitative in nature. Continued analysis could bring in demographic and economic population data, changes in police tactics, and other factors not present in the data to explore potential causes of the trend reversal that appears in mid 2015.