NYPD Gun Incidence Data Analysis

2024-02-21

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## 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     1.3.1
## 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(readr)
library(knitr)

NYPD_Shooting_Incident_Data_Historic_ <- read_csv("NYPD_Shooting_Incident_Data__Historic_.csv")
```

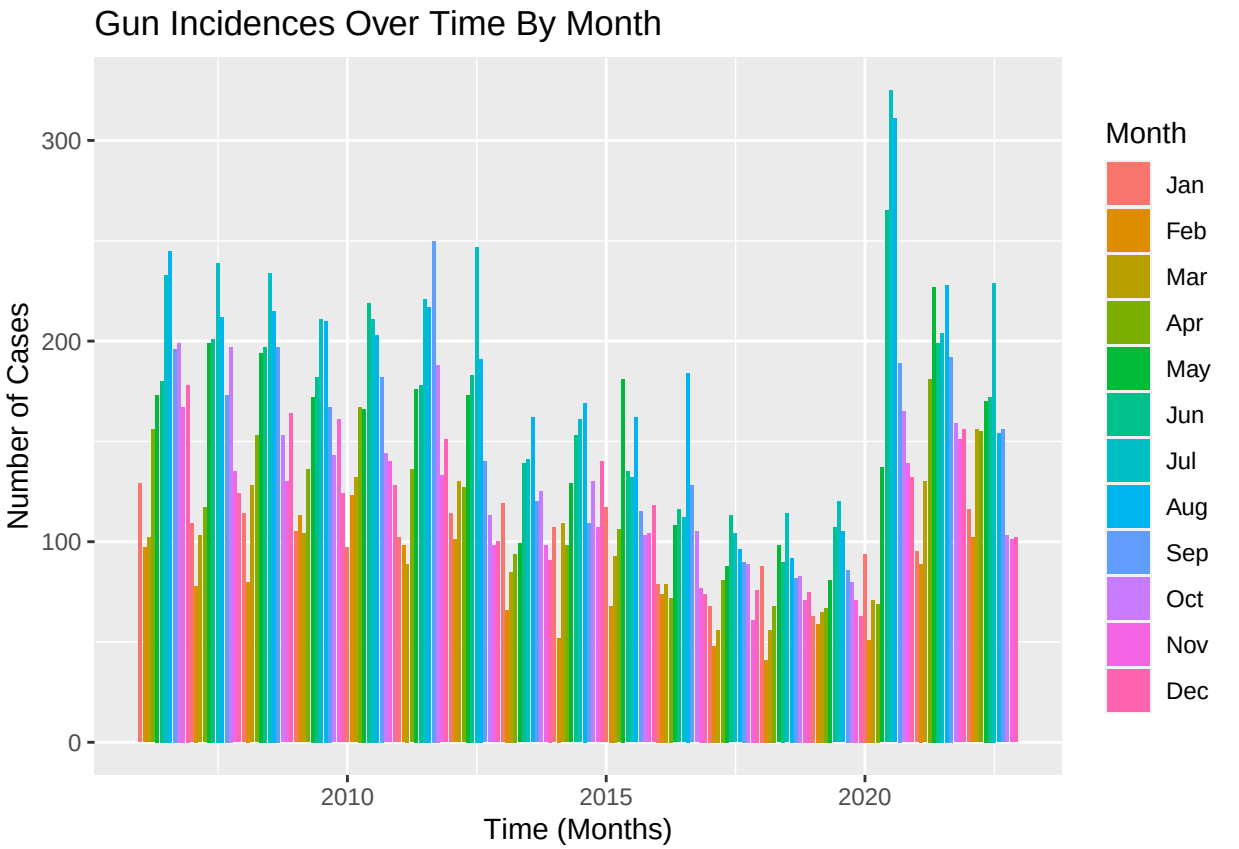
```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## 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.
```

```
nypd <- NYPD_Shooting_Incident_Data_Historic_
nypd <- nypd %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         NUM_CASES = 1)
nypd[11:16][is.na(nypd[11:16]) |
            nypd[11:16] == "(null)" |
            nypd[11:16] == "U" |
            nypd[11:16] == 224 |
            nypd[11:16] == 940 |
            nypd[11:16] == 1020 |
            nypd[11:16] == 1022] <- "UNKNOWN"
```

Inital Cleaning and Observations

After loading the data and looking at the summary information I began cleaning the data. This was accomplished by addressing null values, poorly labeled categorical fields, and converting columns to the correct data types to enable further data analysis. Once the initial cleaning was done I looked at the available data and features to determine what routes for analysis were possible and where I wanted to focus my efforts. Aside from the expected features, I largely grouped the features into perpetrator, victim, and location, information. I started with the location data as I quickly noticed that the grouping for the locations was all clustered around Latitude 40 and Longitude -73. This observation was altogether uninteresting as after looking at this location on a population density map of New York it is clear that these gun incidents are simply occurring in the most populated areas. One other stat I found comforting is of the 27312 incidences only 5266 are confirmed murders, which was less than I anticipated at 19.28%. With the entire data set at my disposal, I followed my interest which was to analyze time data. Some questions and conclusions I would like to work towards answering: - Who is primarily at risk? - When do gun incidence become more prevalent? - Can a conclusion be discovered that will increase individual's safety? To conclude anything about the data I had to know more about it through a visualization or two and time was the particular asset I was interested in.

```
nypd %>%
  group_by(month=lubridate::floor_date(OCCUR_DATE, "month")) %>%
  summarize(num_cases=sum(NUM_CASES),
            .groups="drop_last") %>%
  mutate(Month=factor(month.abb[month(month)],
                          levels=c("Jan", "Feb", "Mar",
                                   "Apr", "May", "Jun",
                                   "Jul", "Aug", "Sep",
                                   "Oct", "Nov", "Dec"))) %>%
  ggplot(aes(fill=Month, y=num_cases, x=month)) +
  geom_bar(position="dodge", stat="identity") +
  ggtitle("Gun Incidences Over Time By Month") +
  theme_gray() +
  xlab("Time (Months)") +
  ylab("Number of Cases")
```



Clearly there is an interesting pattern here. There is an ebb and flow to gun incidents within New York with the high occurring around June, July, and August with it regularly declining again into the winter months from October to March. I would simply assume that the majority cause is the respective weather conditions of New York. When the weather is warmer there are more people outside ultimately leading to more interactions, confrontation, and thus shootings. This exact same effect works in reverse as in the colder months most people tend to stay inside and shy away from the outdoors thus leading to less potential for violence.

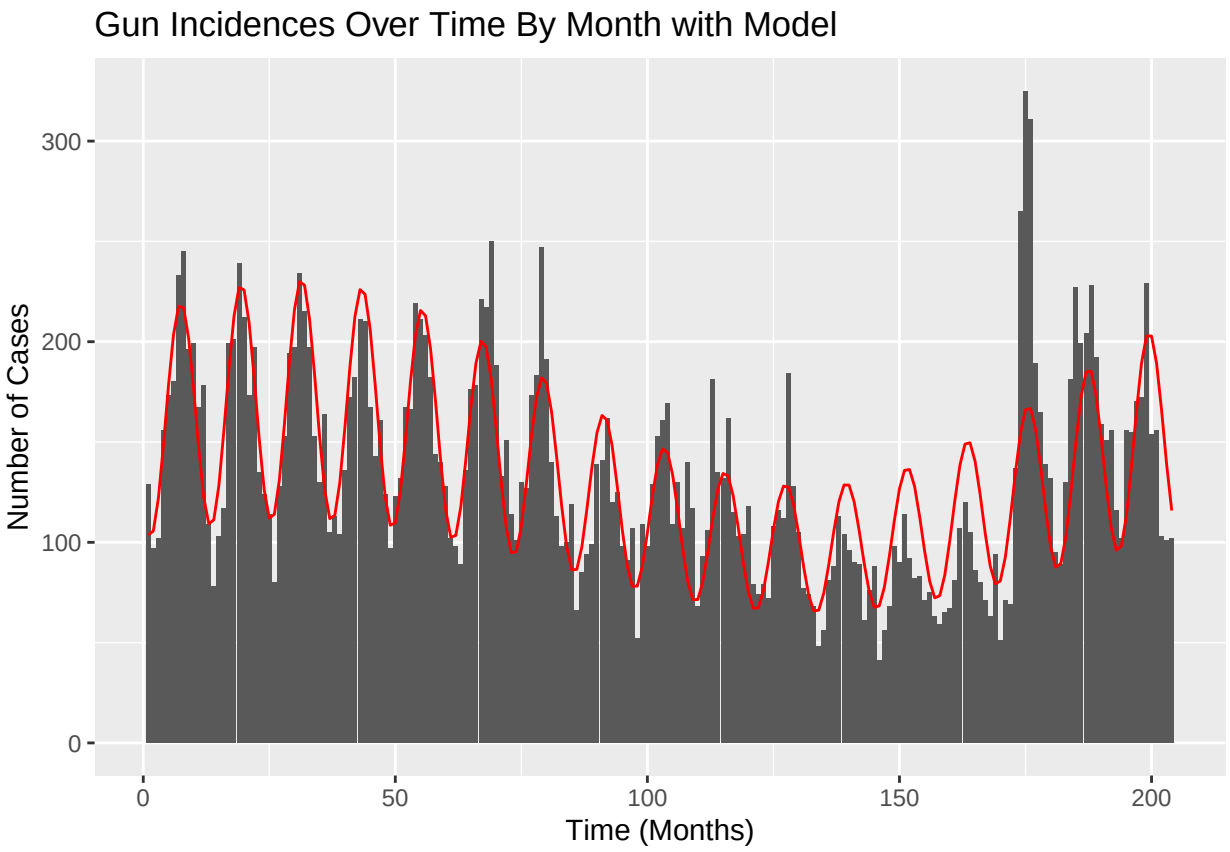
```
df <- nypd %>%
  group_by(month=lubridate::floor_date(OCCUR_DATE, "month")) %>%
  summarize(num_cases=sum(NUM_CASES),
            .groups="drop_last") %>%
  mutate(month=c(1:204))

y <- df$num_cases
t <- df$month

fit <- lm(y ~ sin(2*pi*t/204)+cos(2*pi*t/204))
a <- predict(fit, newdata=data.frame(t=t))
reslm <- lm(y ~ a*sin(2*pi/12*t)+a*cos(2*pi/12*t))
pred <- predict(reslm, newdata=data.frame(t=t))

df %>%
  mutate(t=t,
         pred=unname(pred)) %>%
  ggplot(aes(y=y, x=t)) +
```

```
  geom_bar(stat="identity") +
  geom_line(aes(y=pred, x=t), col="red") +
  ggtitle("Gun Incidences Over Time By Month with Model") +
  theme_gray() +
  xlab("Time (Months)") +
  ylab("Number of Cases")
```



I applied a sinusoidal model to the number of cases by month to better display the regular increase and decrease. As you can see the fit of the curve is telling. In the years coming up to 2020, there is a consistent decline across all cases in all months. This trend is shattered by an exceptional spike which can likely be explained by the increase in all crime during COVID. Fortunately, this increase in gun incidences returned to the pre 2010 levels after the spike and will hopefully continue their pre-COVID downward trajectory.

I liked the conclusions found in this graph so I decided to utilize it further in combination with some of the other perpetrator and victim data that was available.

Perpetrator versus Victim Data

Circling back temporarily to the cleaning work I did, there is a dramatic difference between the quality of the data/reporting between perpetrators and victims. For example, looking at the number of different categorical values for 'PERP_AGE_GROUP' and their respective occurrences in the data set compared to 'VIC_AGE_GROUP is quite stark:

```
vic <- nypd %>%
  group_by('Age Group'=VIC_AGE_GROUP) %>%
```

```
  summarize(Victim=sum(NUM_CASES),
            .groups="drop_last")
perp <- nypd %>%
  group_by('Age Group'=PERP_AGE_GROUP) %>%
  summarize(Perpetrator=sum(NUM_CASES),
            .groups="drop_last")
vic_v_perp_age <- full_join(vic, perp,
                            by=join_by('Age Group')) %>%
  select('Age Group', Victim, Perpetrator)
vic_v_perp_age[2:3][is.na(vic_v_perp_age[2:3])] <- 0
kable(vic_v_perp_age, caption="Number of Labeled Age Group Incidences for Victims and Perpetrators")
```

Table 1: Number of Labeled Age Group Incidences for Victims and Perpetrators

Age Group	Victim	Perpetrator
18-24	10086	6222
25-44	12281	5687
45-64	1863	617
65+	181	60
<18	2839	1591
UNKNOWN	62	13135

There were several different nonsensical fields before beginning cleaning such as 1020, 1022, 940, (null), NA, and UNKNOWN. As part of my cleaning efforts I simply took all the undecipherable fields and lumped them into UNKNOWN. Summing up these unusable categories for the victims results in a total of 62 which is entirely insignificant. Comparing that value to the 13135 of the 27312 total perpetrator instances which is nearly 50% of the available instances unreliable.

```
vic <- nypd %>%
  group_by(Sex=VIC_SEX) %>%
  summarize(Victim=sum(NUM_CASES),
            .groups="drop_last")
perp <- nypd %>%
  group_by(Sex=PERP_SEX) %>%
  summarize(Perpetrator=sum(NUM_CASES),
            .groups="drop_last")
vic_v_perp_sex <- full_join(vic, perp,
                            by=join_by(Sex)) %>%
  select(Sex, Victim, Perpetrator)
vic_v_perp_sex[2:3][is.na(vic_v_perp_sex[2:3])] <- 0
kable(vic_v_perp_sex, caption="Number of Labeled Sex Incidences for Victims and Perpetrators")
```

Table 2: Number of Labeled Sex Incidences for Victims and Perpetrators

Sex	Victim	Perpetrator
F	2615	424
M	24686	15439
UNKNOWN	11	11449

Unfortunately this trend continues with nearly 100% of victims sex data being labeled while less than 60% of the perpetrator sex data is labeled. Again, we have the victim's race unaccounted for in 66 incidences versus the 11786 unknown race incidences for perpetrators.

```
vic <- nypd %>%
  group_by(Race=VIC_RACE) %>%
  summarize(Victim=sum(NUM_CASES),
            .groups="drop_last")
perp <- nypd %>%
  group_by(Race=PERP_RACE) %>%
  summarize(Perpetrator=sum(NUM_CASES),
            .groups="drop_last")
vic_v_perp_race <- full_join(vic, perp,
                             by=join_by(Race)) %>%
  select(Race, Victim, Perpetrator)
vic_v_perp_race[2:3][is.na(vic_v_perp_race[2:3])] <- 0
kable(vic_v_perp_race, caption="Number of Labeled Race Incidences for Victims and Perpetrators")
```

Table 3: Number of Labeled Race Incidences for Victims and Perpetrators

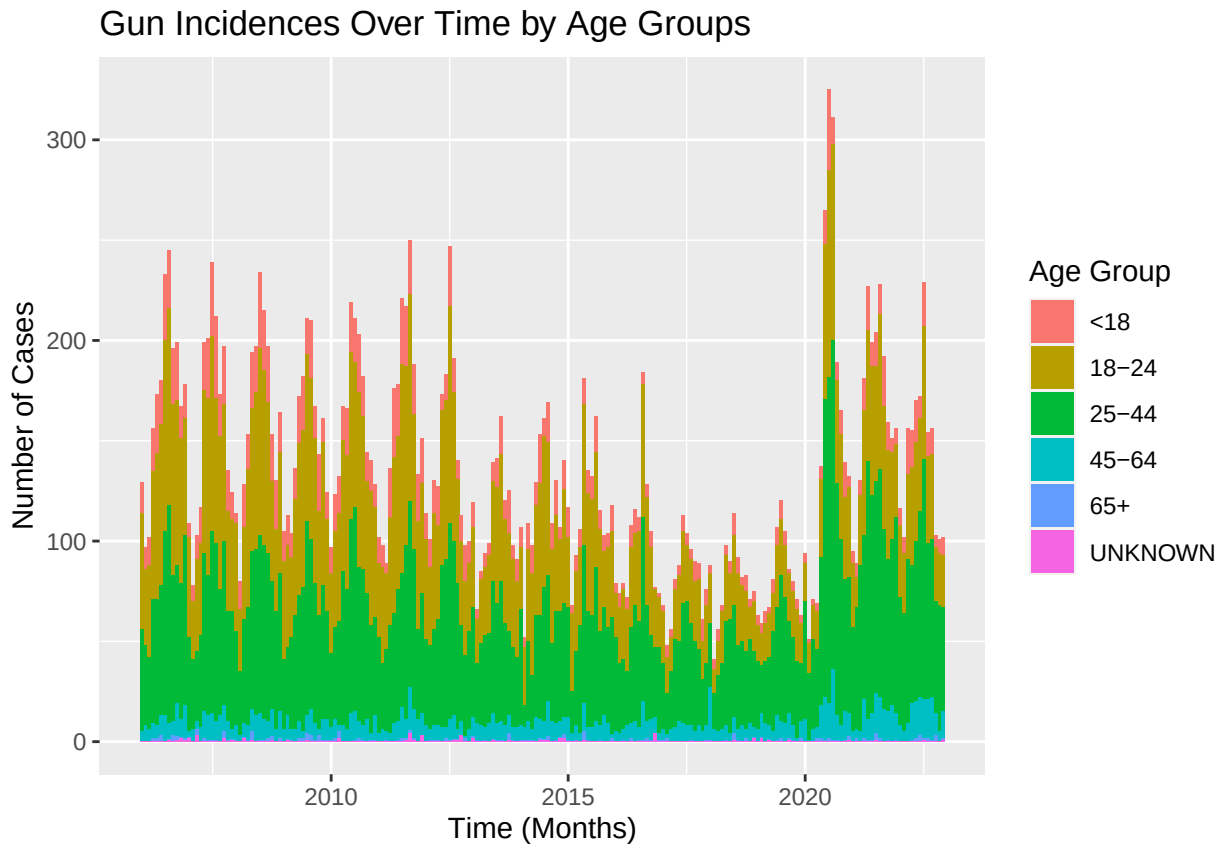
Race	Victim	Perpetrator
AMERICAN INDIAN/ALASKAN NATIVE	10	2
ASIAN / PACIFIC ISLANDER	404	154
BLACK	19439	11432
BLACK HISPANIC	2646	1314
UNKNOWN	66	11786
WHITE	698	283
WHITE HISPANIC	4049	2341

This trend is likely explained by the fact that perpetrators of gun incidences are attempting to get away with the crime committed and thus do their best to conceal their age, sex, and race. While the vast majority of these incidents are likely criminal the respective authorities are not always able to catch the perpetrator and thus their information would never get logged. Since our perpetrator data is so incomplete I feel uncomfortable extrapolating to fill such large gaps. That being said I'm impressed with the quality of our victim data which enables the few unknown features to simply be ignored or dropped. As a result of the irreparable state of the data, I will only be analyzing the victim data.

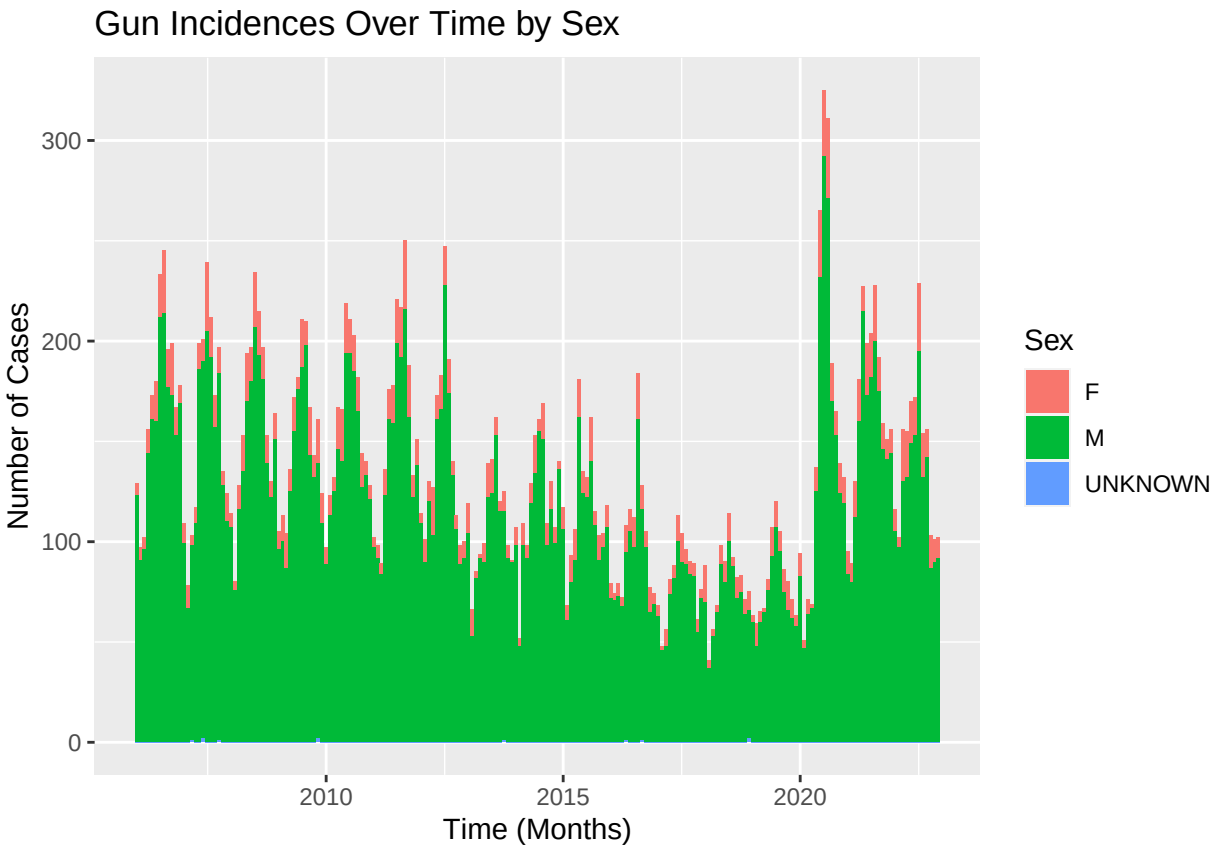
```
stack_plot_month_var <- function(title, y_lab, x_lab, var_name, col){
  nypd %>%
    group_by(month_=lubridate::floor_date(OCCUR_DATE, 'month'),
             var_=col) %>%
    summarize(num_cases_ = sum(NUM_CASES),
              .groups="drop_last") %>%
    tibble(
      month=month_,
      var=var_,
      num_cases=num_cases_
    ) %>%
    complete(month, var,
             fill = list(var=NA, num_cases=0)) %>%
    ggplot(aes(fill=var, y=num_cases, x=month)) +
```

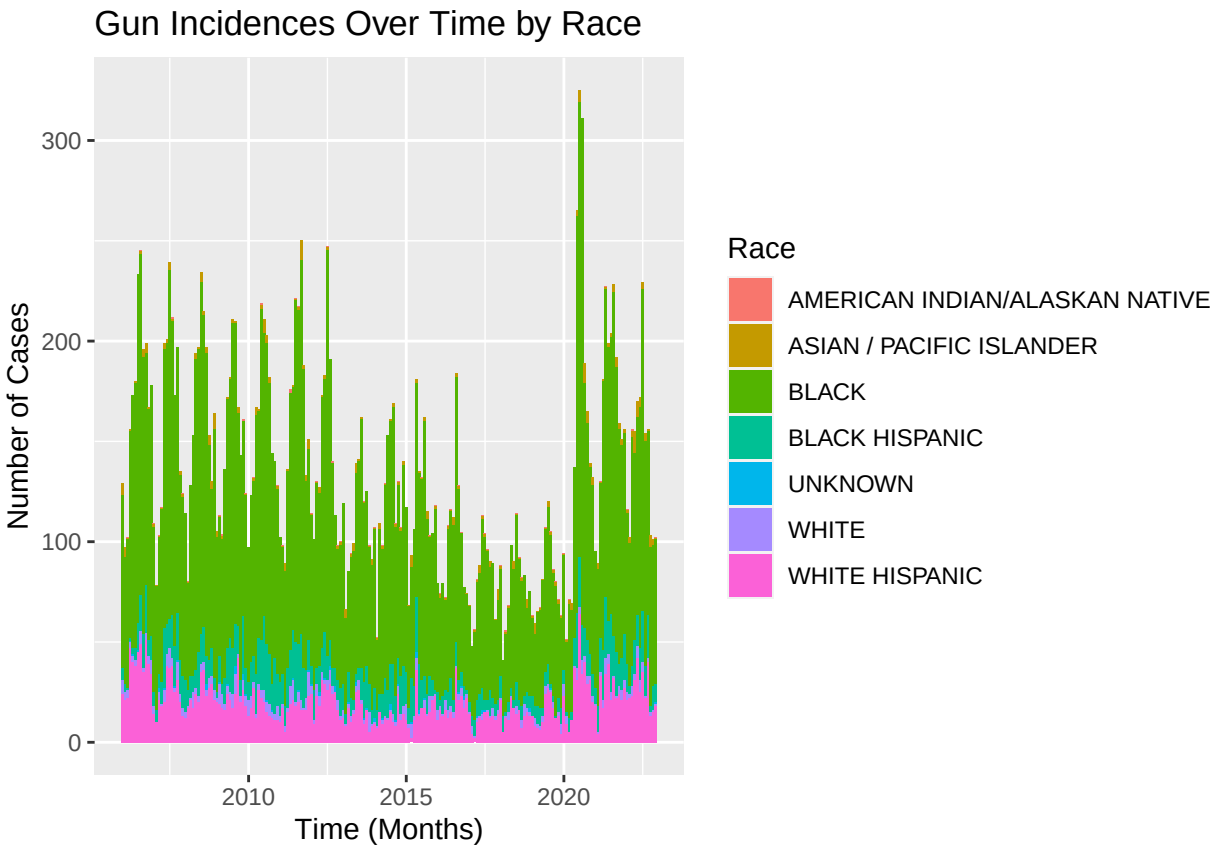
```
    geom_bar(position="stack", stat="identity") +
    ggtitle(title) +
    theme_gray() +
    ylab(y_lab) +
    xlab(x_lab) +
    scale_fill_discrete(name=var_name)
}
```

```
stack_plot_month_var("Gun Incidences Over Time by Age Groups",
                     "Number of Cases",
                     "Time (Months)",
                     "Age Group",
                     nypd$VIC_AGE_GROUP)
```




```
stack_plot_month_var("Gun Incidences Over Time by Sex",
                     "Number of Cases",
                     "Time (Months)",
                     "Sex",
                     nypd$VIC_SEX)
```


Gun Incidences Over Time by Sex
Number of Cases
300
200
100
0
2010
2015
2020
Time (Months)
Sex
F
M
UNKNOWN



Taking an elementary look at all of these plots, we can see the clear majority victim groups. For age group the vast majority of cases occur between ages 25 and 44 with a distant second place going to 18 to 24. To be expected by global historical trends, men drastically overshadow women on the victim graph. And for the race victims of gun incidences, the bulk are Black followed up by White Hispanics. Interestingly enough all of the sub groups for all of the graphs follow the initial time wave we noticed in the first graph. The reason I referred to these observations as "elementary" is since they are misleading without factoring in their statistics proportional to their population.

After searching for a source for this data on the internet for a while I was able to find the data for a few separate years, but not able to find the data for the all of the years in our existing data set. Nevertheless, I believe this data will at least give use some context to better analyze the prior graphs. I would like to draw lots of attention to this step as it is working off a lot of unverified assumptions that I will list now: 1. Age, race, and sex demographics didn't dramatically shift in New York (enough to invalidate the graphs above) between 2006 and 2022. 2. The data I have found is accurate for the respective year. 3. Even if the age, race, and sex demographics did change, more importantly the ratios between the respective classes didn't change in a significant enough manner to disrupt our analysis process.

The data I found is as follows:

```
age_ratio <- tibble('Age Group'=c("<19",
                                  "20-24",
                                  "25-44",
                                  "45-64",
                                  "65+"),
                    Percentage=c(22.99,
                                 6.15,
                                 31.16,
                                 24.63,
```

```
                                  15.09))
sex_ratio <- tibble(Sex=c("M",
                          "F"),
                    Percentage=c(48.51,
                                 51.49))
race_ratio <- tibble(Race=c("White",
                            "Black",
                            "Other",
                            "Asian",
                            "Two or more",
                            "Native American",
                            "Native Hawaiian/Pacific Islander"),
                     Percentage=c(60.73,
                                  15.21,
                                  8.99,
                                  8.65,
                                  5.97,
                                  0.42,
                                  0.05))
kable(age_ratio, caption="Percentage of New York by Age Group")
```

Table 4: Percentage of New York by Age Group

Age Group	Percentage
<19	22.99
20-24	6.15
25-44	31.16
45-64	24.63
65+	15.09

```
kable(sex_ratio, caption="Percentage of New York by Sex")
```

Table 5: Percentage of New York by Sex

Sex	Percentage
M	48.51
F	51.49

```
kable(race_ratio, caption="Percentage of New York by Race")
```

Table 6: Percentage of New York by Race

Race	Percentage
White	60.73
Black	15.21
Other	8.99
Asian	8.65
Two or more	5.97

Race	Percentage
Native American	0.42
Native Hawaiian/Pacific Islander	0.05

Now these elementary observations become more meaningful and begin to hold more weight. As now for age the statistic becomes less surprising. The majority age group for victims corresponds with the majority age group in total making the statistic more proportional. This also reveals a shift that into the 45-65 age group even though it still represents a sizable chunk of the total population, the amount of victims within this age group is smaller than the victims within the ~18-24 age group which is responsible for a much smaller share of the total population. This likely makes sense as older people commit less crime for a variety of reasons and are therefore less likely to be involved in crime in general. The observations in the sex chart become incredibly profound with the external context of the total population. Women and men are almost split for the total population with women actually holding a slight majority, meanwhile men dominate the victims graph. I again lean towards the correlation that men commit more crime and are thus more likely to be involved in a gun incident (as either the perpetrator or the victim). Analyzing the racial gun incident victims graph with this new context show that the majority class of Black individuals is the vast majority of victims while they are nowhere near the majority population in New York. That total majority holder is White which is hardly even shown on the graph.

Conclusion and Bias

I am pleased with the results of the analysis achieved. I believe individuals can gleam substantive information that would enable they to live more safely within New York. First off, it is incredibly rare for you to be involved in a gun incident if you are not located near New York City. Additionally, the primary demographic that should exercise extra precautions to avoid potential gun incidence are Black males between the ages of 25 to 44. All individuals should be more wary during the summer months and everyone should take comfort in the fact that your likelihood of being murdered in one of these gun incidences, or being involved in one at all given the total population of New York is small.

As far as bias is concerned, I attempted to minimize my personal bias and the bias of the data set by not coming to unreasonable conclusions about the perpetrators of these gun incidences and acknowledging the assumptions I was working off of. It should go without saying that I am not the one who collected this data and I have not, nor do I have any way, of verifying the validity of the data which I analyzed. This also applies to the data I brought in from additional external sources that may also not be accurate and which I rely on crucial assumptions to build the latter part of my analysis. More bias that could be exist in the data which I analyzed is the few unknown entries within the victim data which I believe is negligible, but again I would like to emphasize this is an assumption.

Citation

- New York 2019 Sex Data
- New York 2022 Race Data
- New York 2021 Age Data