NYPD Shooting Incident Data Report

J. Anderson

2024-06-18

Abstract

This report was completed for CU Boulder's "Data Science as a Field" class. I analyzed "NYPD Shooting Incident Data (Historic) by first cleaning the data, creating some basic visualizations, further transforming the data, and then running linear regression models on variables of interest.

After a cursory exploration, I conclude that the data set supports the conclusion that a majority of the shooting incidents recorded by the NYPD in this data set were race-on-race, age-on-age incidents. This conclusion leads to a number of new questions about the social situations in which shootings, at least in NYC, most commonly occur.

Preparation

The following libraries will be loaded for data tidying and visualization. The libraries not discussed as part of CU Boulder's "Data Science as a Field" course are, I believe, "readr", "hms", and "patchwork". The former two are packages I will use to convert particularly tricky data types; the latter aggregates ggplots into one visualization.

```
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   3.5.1     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(hms)
```

```
##
## Attaching package: 'hms'
##
## The following object is masked from 'package:lubridate':
##
##     hms
```

```
library(ggplot2)
library(dplyr)
library(patchwork)
```

The NYPD Shooting Incident Data is loaded as a .csv file into R. You can find the .csv file I used in my GitHub profile's corresponding folder.

```
data <- read.csv(
  "/Users/jadanlynn/Documents/Data Science as a Field/NYPD_Shooting_Incident_Data__Historic_.csv")
```

Tidying Data

For this data set, the primary tidying task swill include removing unnecessary columns, renaming columns, changing data types (mostly from the character class into the factor class, although this data set also required date and boolean conversions), and the consolidation of missing data into "NA" types.

```
# remove unwanted columns
data <- data[, !names(data) %in% c("LOC_CLASSFCTN_DESC",
                                   "LOCATION_DESC",
                                   "Latitude", "Longitude",
                                   "XCOORD", "YCOORD", "Lon_Lat")]
```

```
# rename columns
data <- data %>%
  dplyr::rename(KEY = INCIDENT_KEY, DATE = OCCUR_DATE, TIME = OCCUR_TIME,
                IN_OUT = LOC_OF_OCCUR_DESC, JURIS = JURISDICTION_CODE,
                MURDER = STATISTICAL_MURDER_FLAG, P_AGE = PERP_AGE_GROUP,
                P_SEX = PERP_SEX, P_RACE = PERP_RACE, V_AGE = VIC_AGE_GROUP,
                V_SEX = VIC_SEX, V_RACE = VIC_RACE, XCOORD = X_COORD_CD,
                YCOORD = Y_COORD_CD)

names(data)
```

```
##  [1] "KEY"      "DATE"     "TIME"     "BORO"     "IN_OUT"   "PRECINCT"
##  [7] "JURIS"    "MURDER"   "P_AGE"    "P_SEX"    "P_RACE"   "V_AGE"
## [13] "V_SEX"    "V_RACE"   "XCOORD"   "YCOORD"
```

```
# data type conversion
data <- data %>%
  mutate(
    DATE = as.Date(DATE, format = "%m/%d/%Y"),
    TIME = hms::as_hms(TIME),
    BORO = as.factor(BORO),
    IN_OUT = as.factor(IN_OUT),
    PRECINCT = as.factor(PRECINCT),
    JURIS = as.factor(JURIS),
    MURDER = case_when(
      MURDER == "true" ~ TRUE,
      MURDER == "false" ~ FALSE),
    P_AGE = as.factor(P_AGE),
    P_SEX = as.factor(P_SEX),
    P_RACE = as.factor(P_RACE),
    V_AGE = as.factor(V_AGE),
    V_SEX = as.factor(V_SEX),
    V_RACE = as.factor(V_RACE)
    )

summary(data)
```

```
##       KEY                 DATE                 TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms
##  Median : 92711254   Median :2013-09-20   Class2:difftime
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##             BORO           IN_OUT         PRECINCT      JURIS
##  BRONX        : 8376          :25596   75     : 1628   0   :23923
##  BROOKLYN     :11346   INSIDE :  460   73     : 1500   1   :   81
##  MANHATTAN    : 3762   OUTSIDE: 2506   67     : 1259   2   : 4556
##  QUEENS       : 4271                   44     : 1076   NA’s:    2
##  STATEN ISLAND:  807                   79     : 1045
##                                        47     : 1006
##                                        (Other):21048
##    MURDER            P_AGE         P_SEX                P_RACE
##  Mode :logical          :9344         : 9310   BLACK         :11903
##  FALSE:23036     18-24  :6438   (null): 1141                 : 9310
##  TRUE :5526      25-44  :6041   F     :  444   WHITE HISPANIC: 2510
##                  UNKNOWN:3148   M     :16168   UNKNOWN       : 1837
##                  <18    :1682   U     : 1499   BLACK HISPANIC: 1392
##                  (null) :1141                  (null)        : 1141
##                  (Other): 768                  (Other)       :  469
##      V_AGE       V_SEX                                V_RACE
##  <18    : 2954   F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  1022   :    1   M:25790   ASIAN / PACIFIC ISLANDER      :  440
##  18-24  :10384   U:   12   BLACK                         :20235
##  25-44  :12973             BLACK HISPANIC                : 2795
##  45-64  : 1981             UNKNOWN                       :   70
##  65+    :  205             WHITE                         :  728
##  UNKNOWN:   64             WHITE HISPANIC                : 4283
##      XCOORD            YCOORD
##  Min.   : 914928   Min.   :125757
##  1st Qu.:1000068   1st Qu.:182912
##  Median :1007772   Median :194901
##  Mean   :1009424   Mean   :208380
##  3rd Qu.:1016807   3rd Qu.:239814
##  Max.   :1066815   Max.   :271128
##
```

```
# consolidation of missing data
data$IN_OUT[data$IN_OUT == ""] <- NA
data$P_AGE[data$P_AGE == ""] <- NA
data$P_AGE[data$P_AGE == "(null)"] <- NA
```

```r
data$P_AGE[data$P_AGE == "UNKNOWN"] <- NA
data$P_AGE[data$P_AGE == "1020"] <- NA
data$P_AGE[data$P_AGE == "1028"] <- NA
data$P_AGE[data$P_AGE == "224"] <- NA
data$P_AGE[data$P_AGE == "940"] <- NA
data$P_SEX[data$P_SEX == ""] <- NA
data$P_SEX[data$P_SEX == "(null)"] <- NA
data$P_SEX[data$P_SEX == "U"] <- NA
data$P_RACE[data$P_RACE == ""] <- NA
data$P_RACE[data$P_RACE == "(null)"] <- NA
data$P_RACE[data$P_RACE == "UNKNOWN"] <- NA
data$V_AGE[data$V_AGE == "1022"] <- NA
data$V_AGE[data$V_AGE == "UNKNOWN"] <- NA
data$V_SEX[data$V_SEX == "U"] <- NA
data$V_RACE[data$V_RACE == "UNKNOWN"] <- NA
```

```r
# removing factor values with 0 observations
data <- data %>%
  mutate(
    IN_OUT = droplevels(IN_OUT),
    P_SEX = droplevels(P_SEX),
    V_AGE = droplevels(V_AGE),
    V_SEX = droplevels(V_SEX),
    P_RACE = droplevels(P_RACE),
    P_AGE = droplevels(P_AGE)
    )

summary(data)
```

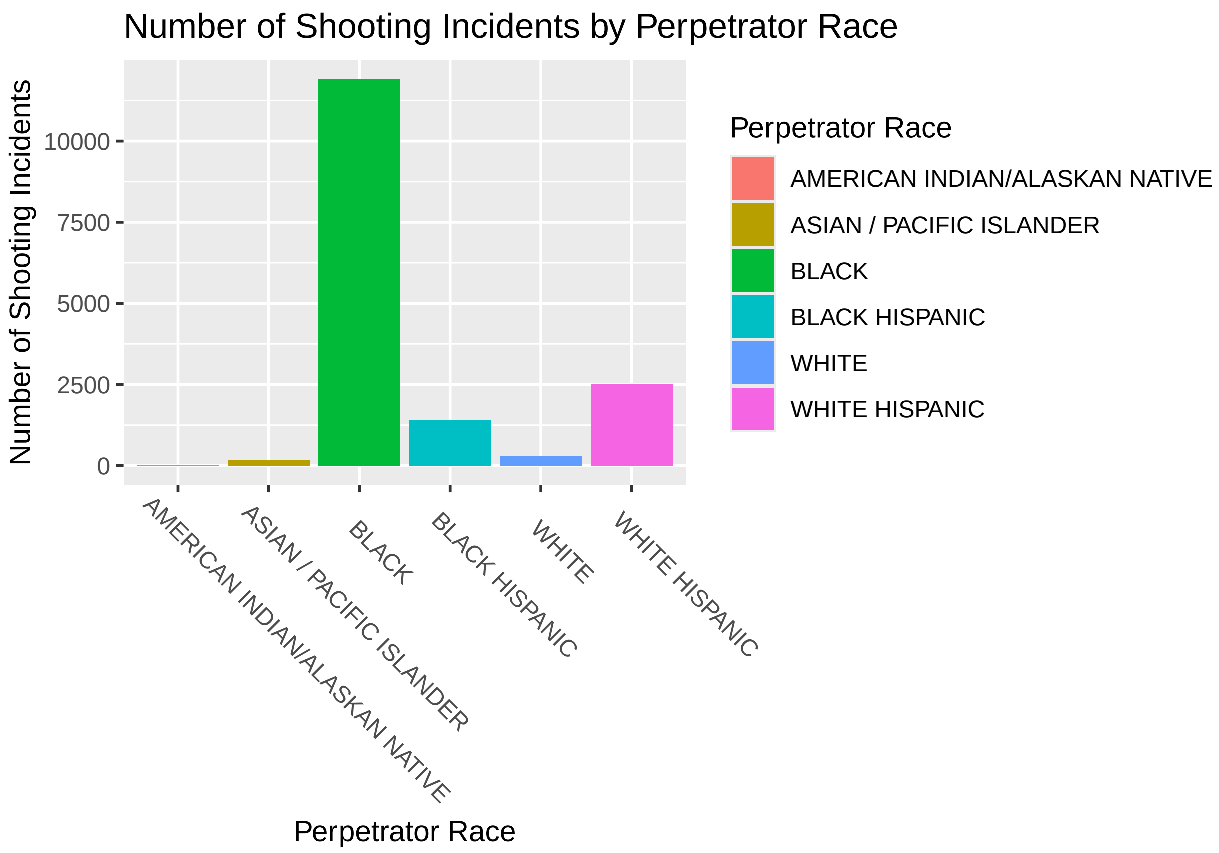
```
##       KEY                 DATE                TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms
##  Median : 92711254   Median :2013-09-20   Class2:difftime
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##             BORO           IN_OUT         PRECINCT      JURIS
##  BRONX        : 8376   INSIDE :  460   75     : 1628   0   :23923
##  BROOKLYN     :11346   OUTSIDE: 2506   73     : 1500   1   :   81
##  MANHATTAN    : 3762   NA’s   :25596   67     : 1259   2   : 4556
##  QUEENS       : 4271                   44     : 1076   NA’s:    2
##  STATEN ISLAND:  807                   79     : 1045
##                                        47     : 1006
##                                        (Other):21048
##    MURDER          P_AGE        P_SEX
##  Mode :logical   <18  : 1682   F    :  444
##  FALSE:23036     18-24: 6438   M    :16168
##  TRUE :5526      25-44: 6041   NA’s:11950
##                  45-64:  699
##                  65+  :   65
##                  NA’s :13637
##
```

```
##                              P_RACE        V_AGE        V_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18  : 2954   F   : 2760
##  ASIAN / PACIFIC ISLANDER      :  169   18-24:10384   M   :25790
##  BLACK                         :11903   25-44:12973   NA’s:   12
##  BLACK HISPANIC                : 1392   45-64: 1981
##  WHITE                         :  298   65+  :  205
##  WHITE HISPANIC                : 2510   NA’s :   65
##  NA’s                          :12288
##                        V_RACE          XCOORD             YCOORD
##  BLACK                   :20235   Min.   : 914928   Min.   :125757
##  WHITE HISPANIC          : 4283   1st Qu.:1000068   1st Qu.:182912
##  BLACK HISPANIC          : 2795   Median :1007772   Median :194901
##  WHITE                   :  728   Mean   :1009424   Mean   :208380
##  ASIAN / PACIFIC ISLANDER:  440   3rd Qu.:1016807   3rd Qu.:239814
##  (Other)                 :   11   Max.   :1066815   Max.   :271128
##  NA’s                    :   70
```

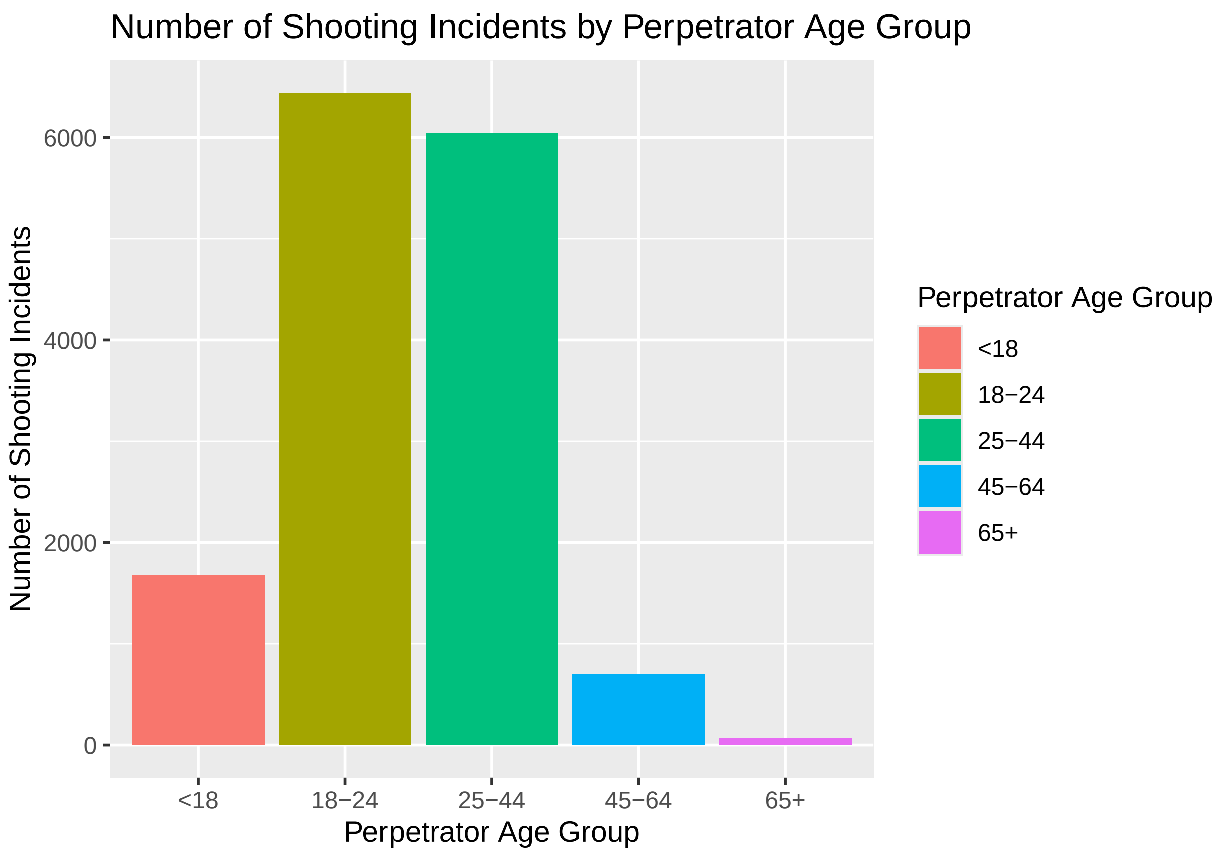
Exploration via Visualization

Exploration will begin by creating and analyzing basic plots of two key variables: perpetrator race and age group.

```r
perp_race_plot <- data %>%
  drop_na(P_RACE) %>%
  ggplot(mapping = aes(x = P_RACE, fill = P_RACE)) +
  geom_bar(stat = "count") +
  labs(
    x = "Perpetrator Race",
    y = "Number of Shooting Incidents",
    colour = "Perpetrator Race",
    title = "Number of Shooting Incidents by Perpetrator Race",
    fill = "Perpetrator Race") +
  theme(axis.text.x = element_text(angle = -45, vjust = 0.5, hjust = 0.1))

perp_race_plot
```



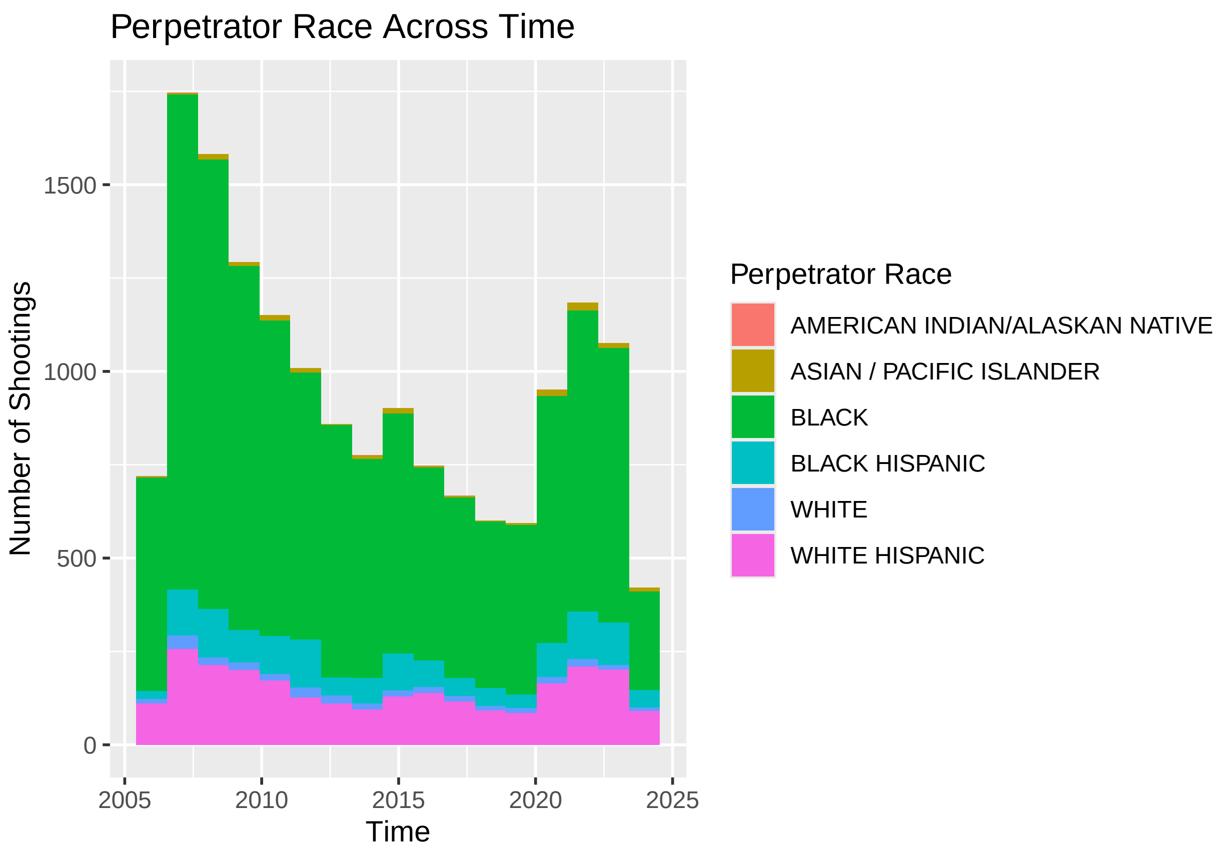
```
perp_age_graph <- data %>%
  drop_na(P_AGE) %>%
  ggplot(mapping = aes(x = P_AGE, fill = P_AGE)) +
  geom_bar(stat = "count") +
  labs(
    x = "Perpetrator Age Group",
    y = "Number of Shooting Incidents",
    title = "Number of Shooting Incidents by Perpetrator Age Group",
    fill = "Perpetrator Age Group")

perp_age_graph
```



From these basic visualizations we can easily see that, based off of our sampling of *known* data, majority of the shooting incident perpetrators are Black and between the ages of 18 and 44.

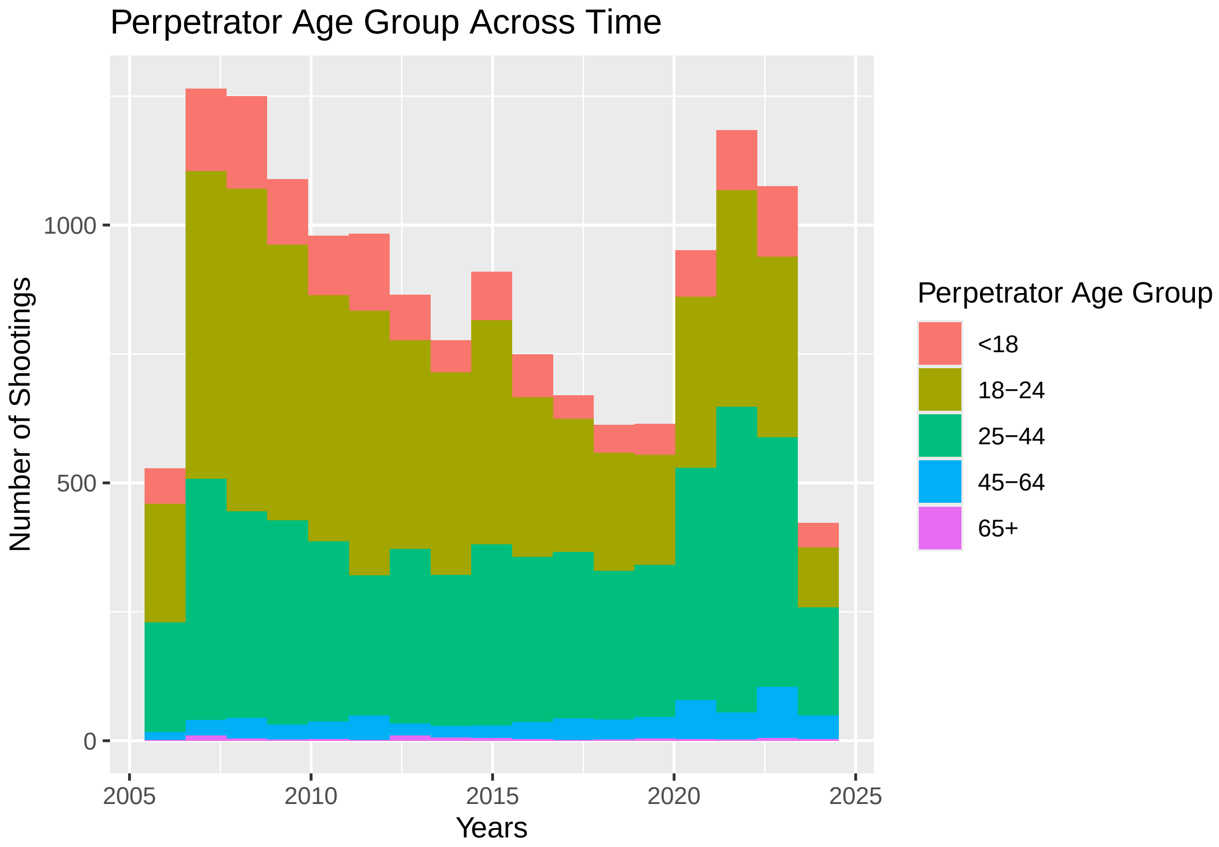
Now the visualizations will track these two metrics—perpetrator race and age group—over time.

```
perp_race_time <- data %>%
  drop_na(P_RACE) %>%
  ggplot(aes(x = DATE, fill = P_RACE)) +
  geom_histogram(bins = 17) +
  labs(
    x = "Time",
    y = "Number of Shootings",
    title = "Perpetrator Race Across Time",
    fill = "Perpetrator Race"
    )

perp_race_time
```



From an examination of the chart above, it does not seem like the racial composition of shooting perpetrators have changed over time.

```
perp_age_time <- data %>%
  drop_na(P_AGE) %>%
  ggplot(aes(x = DATE, fill = P_AGE)) +
  geom_histogram(bins = 17) +
  labs(
    x = "Years",
    y = "Number of Shootings",
    title = "Perpetrator Age Group Across Time",
    fill = "Perpetrator Age Group"
    )

perp_age_time
```



A careful look at perpetrator age group across time suggests that the ages of shooting incident perpetrators *have* changed over time. Specifically, whereas majority of perpetrators were from the 18-24 age group from 2006 to 2015, afterward it seems that the number of shooters from the 25-44 age group increased to equal, if not eclipse, the number of shooters from the 18-24 age group.

This visual analysis will be verified by creating simpler graphs of the changes in number of perpetrators from individual age groups over time.

```
less_18_change <- data %>%
  drop_na(P_AGE) %>%
  ggplot(filter(data, P_AGE == "<18"),
         mapping = aes(x = DATE)) +
  geom_bar(stat = "count", width = 100, show.legend = FALSE, fill = "pink") +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Change in Age Group: <18 Perpetrators Across Time",
    )

# less_18_change
```

```
change_18_24 <- ggplot(data = filter(data, P_AGE == "18-24"),
                       mapping = aes(x = DATE)) +
  geom_bar(stat = "count", width = 100, show.legend = FALSE, fill = "darkgreen") +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Change in Age Group: 18-24 Perpetrators Across Time"
    )
```

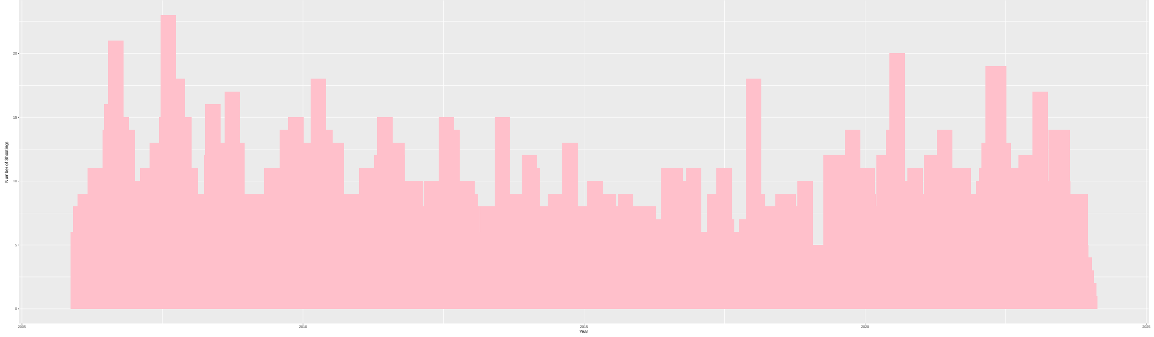
```
change_25_44 <- ggplot(data = filter(data, P_AGE == "25-44"),
                       mapping = aes(x = DATE)) +
  geom_bar(stat = "count", width = 100, show.legend = FALSE, fill = "turquoise") +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Change in Age Group: 25-44 Perpetrators Across Time"
    )
```

```
change_45_64 <- ggplot(data = filter(data, P_AGE == "45-64"),
                       mapping = aes(x = DATE)) +
  geom_bar(stat = "count", width = 100, show.legend = FALSE, fill = "lightblue") +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Change in Age Group: 45-64 Perpetrators Across Time")
```

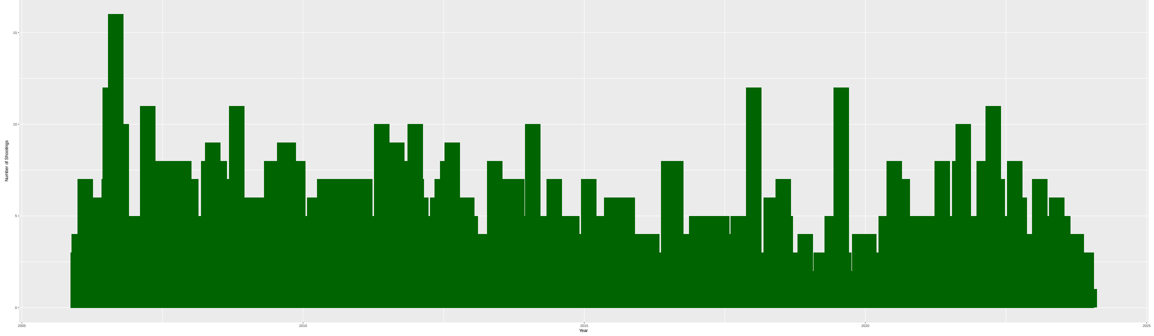
```
change_65 <- ggplot(data = filter(data, P_AGE == "65+"),
                    mapping = aes(x = DATE)) +
  geom_bar(stat = "count", width = 100, fill = "purple", show.legend = FALSE) +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Change in Age Group: 65+ Perpetrators Across Time"
    )

less_18_change/change_18_24/change_25_44/change_45_64/change_65
```

```
## Warning: ‘position_stack()‘ requires non-overlapping x intervals.
## ‘position_stack()‘ requires non-overlapping x intervals.
## ‘position_stack()‘ requires non-overlapping x intervals.
## ‘position_stack()‘ requires non-overlapping x intervals.
## ‘position_stack()‘ requires non-overlapping x intervals.
```

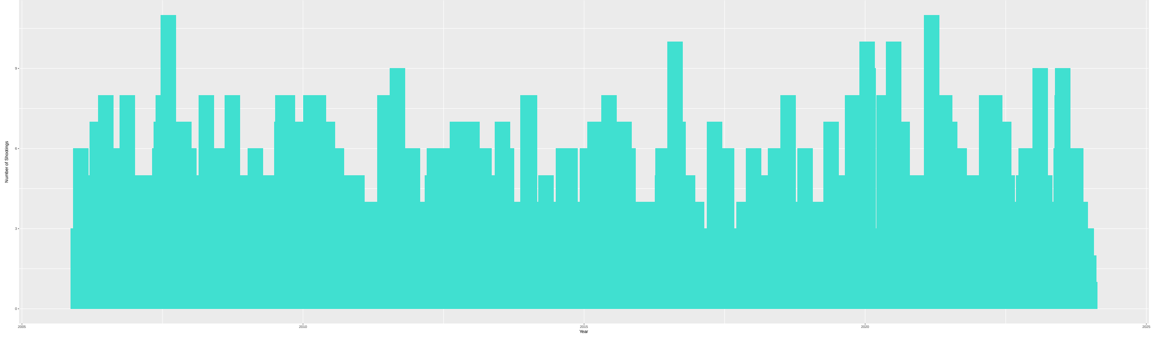
Change in Age Group: <18 Perpetrators Across Time



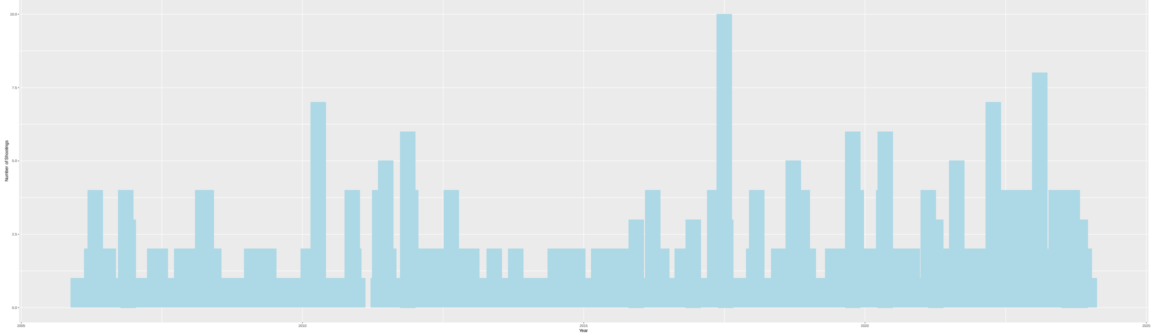
Change in Age Group: 18−24 Perpetrators Across Time



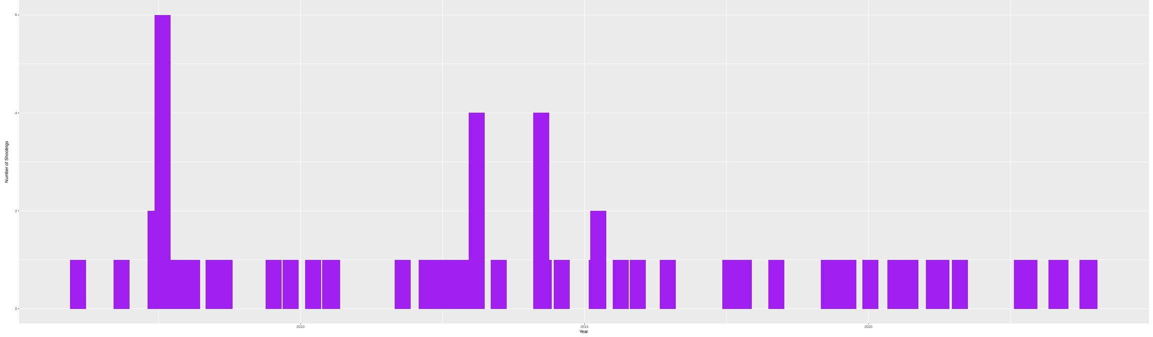
Change in Age Group: 25−44 Perpetrators Across Time



Change in Age Group: 45−64 Perpetrators Across Time

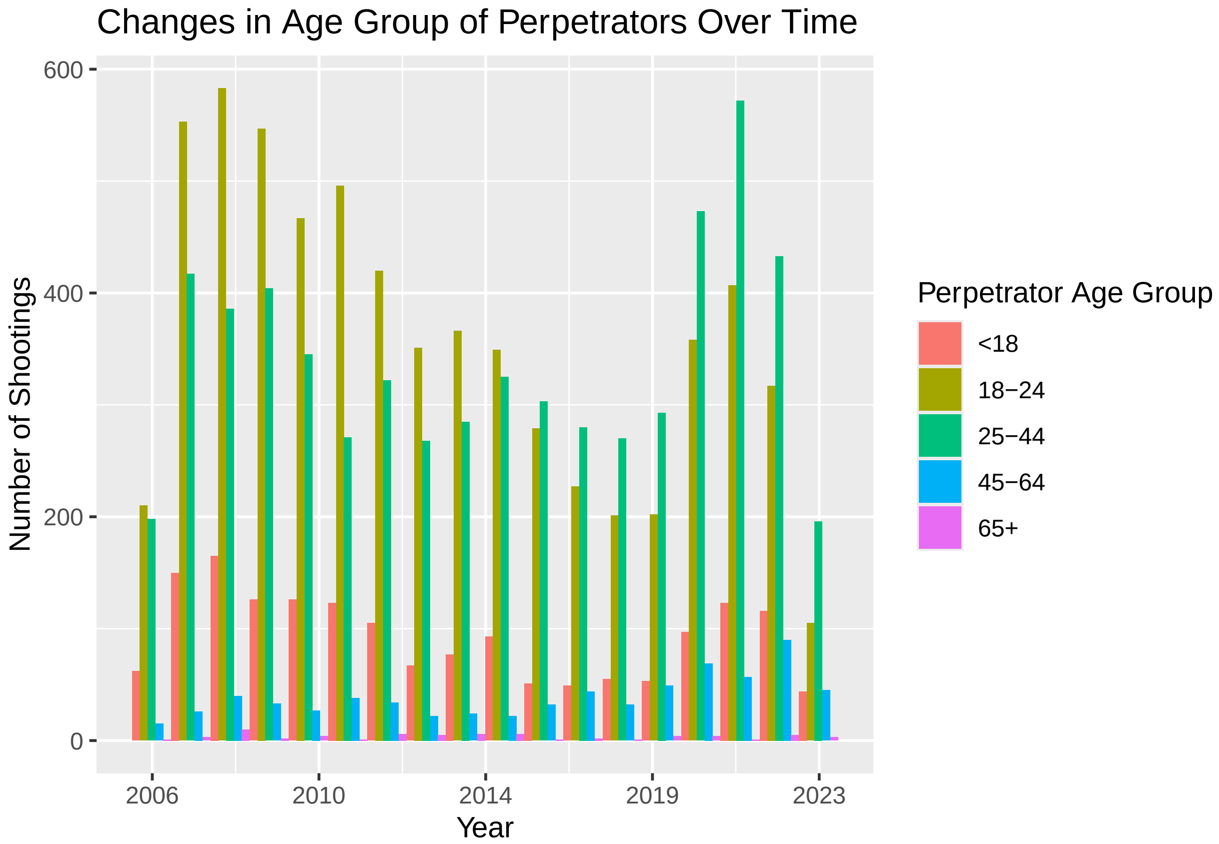


Change in Age Group: 65+ Perpetrators Across Time



It is difficult to see trends with the individual age group bar graphs. Instead, here is a grouped histogram of all perpetrator age groups over time.

```r
age_over_time_grouped <- data %>%
  drop_na(P_AGE) %>%
  ggplot(aes(x = DATE, fill = P_AGE)) +
  geom_histogram(position = "dodge", bins = 18) +
  scale_x_date(breaks = seq(min(data$DATE), max(data$DATE), length = 5),
               date_labels = "%Y") +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Changes in Age Group of Perpetrators Over Time",
    fill = "Perpetrator Age Group"
    )

age_over_time_grouped
```

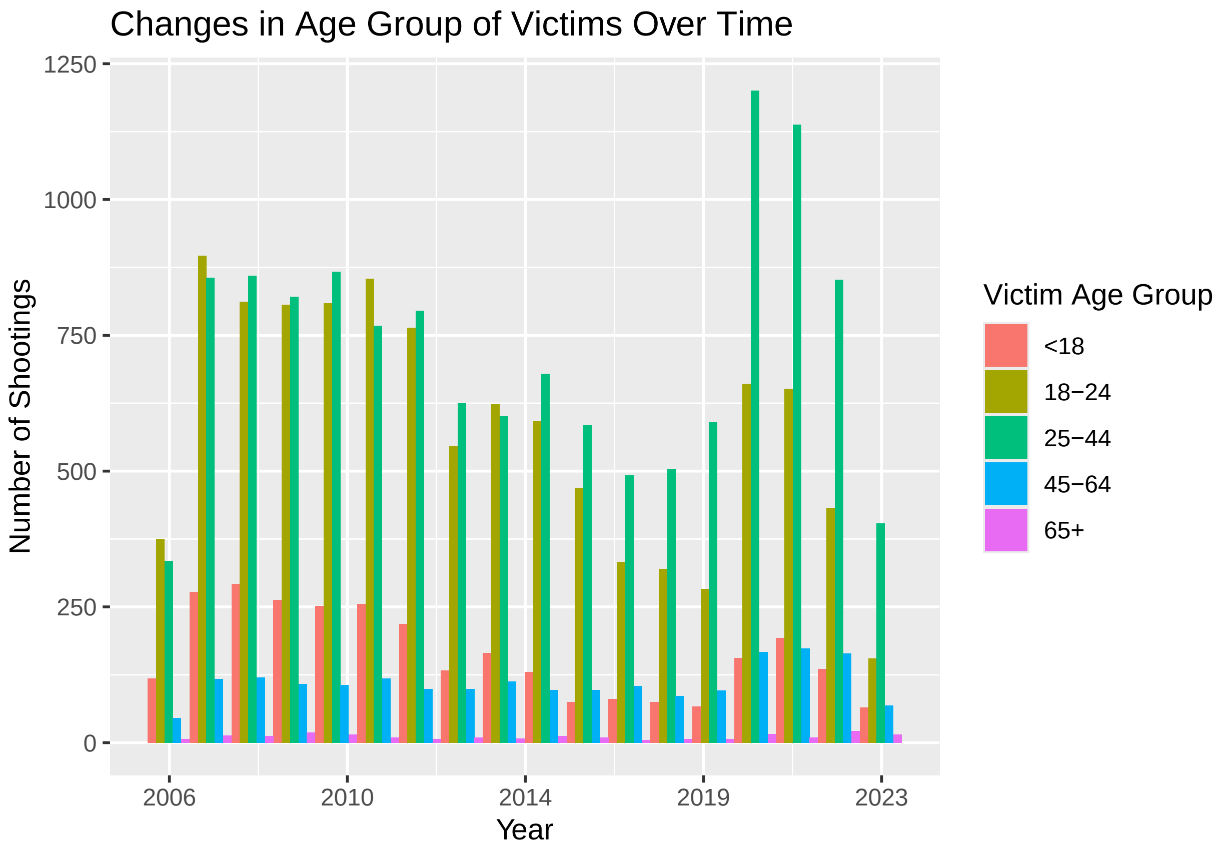


The “Changes in Age Group of Perpetrators Over Time” graph better displays the following trends:

1. The number of perpetrators from age group “65+” stays constant over time
2. The number of perpetrators from age group “45-64” slightly increases over time
3. The number of perpetrators from age group “25-44” increases over time
4. The number of perpetrators from age group “18-24” decreases over time
5. The number of perpetrators from age group “<18” stays constant over time

Consider victim age groups over time:

```r
victim_age_over_time <- data %>%
  drop_na(V_AGE) %>%
  ggplot(aes(x = DATE, fill = V_AGE)) +
  geom_histogram(position = "dodge", bins = 18) +
  scale_x_date(breaks = seq(min(data$DATE), max(data$DATE), length = 5),
               date_labels = "%Y") +
  labs(
    x = "Year",
    y = "Number of Shootings",
    title = "Changes in Age Group of Victims Over Time",
    fill = "Victim Age Group"
    )

victim_age_over_time
```



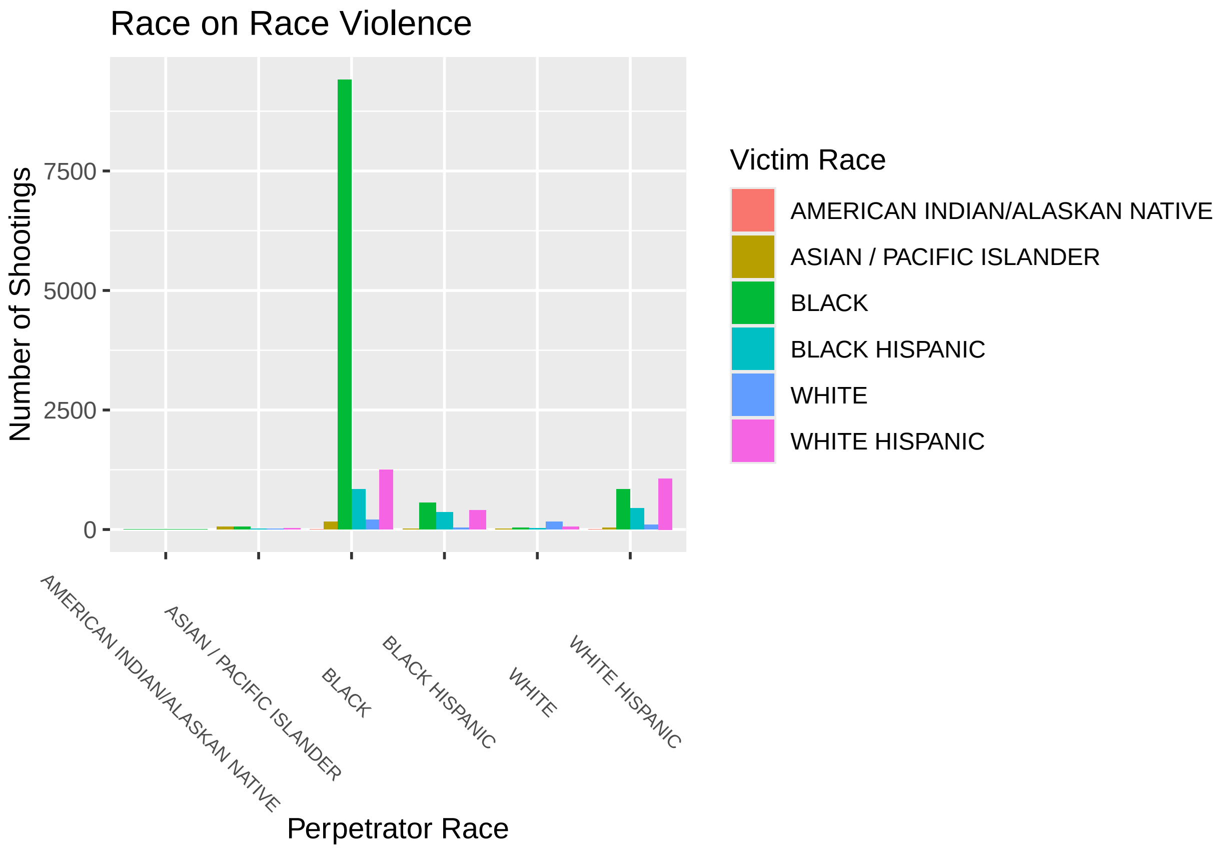
The “Changes in Age Group of Victims Over Time” graph displays the following trends:

1. The number of perpetrators from age group “65+” stays constant over time
2. The number of perpetrators from age group “45-64” stays constant over time
3. The number of perpetrators from age group “25-44” increases over time
4. The number of perpetrators from age group “18-24” decreases over time
5. The number of perpetrators from age group “<18” decreases over time

Analyzing

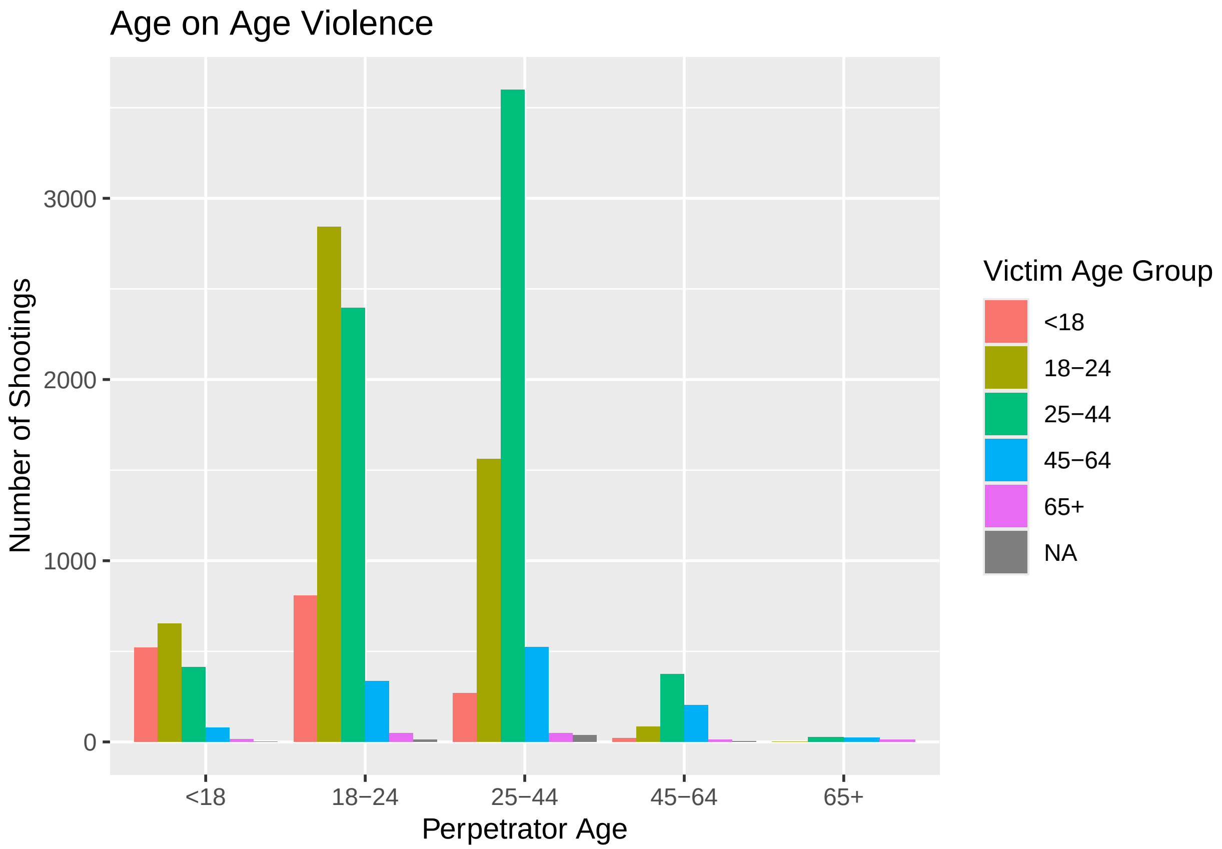
Seeing as how the racial breakdown of perpetrators and victims stays constant over time, and seeing the trend that the number of both “25-44” perpetrators and victims increases over time, lends to a hypothesis that gun violence from perpetrator to victim is often race on race, age on age.

```
race_on_race <- data %>%
  drop_na(V_RACE) %>%
  drop_na(P_RACE) %>%
  ggplot(aes(x = P_RACE, fill = V_RACE)) +
  geom_bar(position = "dodge") +
  labs(
    x = "Perpetrator Race",
    y = "Number of Shootings",
    title = "Race on Race Violence",
    fill = "Victim Race"
    ) +
  theme(axis.text.x = element_text(angle = -45, size = 7))

race_on_race
```



Two trends can be derived from the graph above. The first is the overwhelmingly the victims of Black perpetrators are Black. In fact, Black victims are the majority for Black, Black Hispanic, and White Hispanic perpetrators. The second is that the majority of White perpetrators’ victims are also White.

```r
age_on_age <- data %>%
  drop_na(P_AGE) %>%
  ggplot(aes(x = P_AGE, fill = V_AGE)) +
  geom_bar(position = "dodge") +
  labs(
    x = "Perpetrator Age",
    y = "Number of Shootings",
    title = "Age on Age Violence",
    fill = "Victim Age Group"
    )

age_on_age
```



From this graph the following trends can be derived:

1. For perpetrators between the ages of 18-24, majority of their victims are also ages 18-24
2. For perpetrators between the ages of 18-24, victims between 25-44 constitute the secondary majority
3. For perpetrators between the ages of 25-44, majority of their victims are also ages 25-44
4. For perpetrators between the ages of 25-44, victims between 18-24 constitute the secondary majority
5. For perpetrators below the age of 18, the primary and secondary majorities are age groups "18-24" and "<18", respectively
6. For perpetrators between the ages of 18-44, the primary and secondary majorities are some combination of age groups "18-24" and "24-44"
7. For perpetrators above 44, the primary and secondary majorities are age groups "25-44" and "45-64", respectively

Data Modeling

Multiple linear regression models will be used to verify the trends derived from the two graphs above.

```
data <- data %>%
  mutate(
    P_RACE_NATIVE = ifelse(P_RACE == "AMERICAN INDIAN/ALASKAN NATIVE", 1, 0),
    P_RACE_API = ifelse(P_RACE == "ASIAN / PACIFIC ISLANDER", 1, 0),
    P_RACE_BLACK = ifelse(P_RACE == "BLACK", 1, 0),
    P_RACE_BLACKHIS = ifelse(P_RACE == "BLACK HISPANIC", 1, 0),
    P_RACE_WHITE = ifelse(P_RACE == "WHITE", 1, 0),
    P_RACE_WHITEHIS = ifelse(P_RACE == "WHITE HISPANIC", 1, 0)
    )
```

```
race_model_v_black <- lm(V_RACE == "BLACK" ~ P_RACE_BLACK + P_RACE_NATIVE + P_RACE_API +
                           P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS, data = data)
race_model_v_white <- lm(V_RACE == "WHITE" ~ P_RACE_BLACK + P_RACE_NATIVE + P_RACE_API +
                           P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS, data = data)
race_model_v_api <- lm(V_RACE == "ASIAN / PACIFIC ISLANDER" ~ P_RACE_BLACK + P_RACE_NATIVE
                       + P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS,
                       data = data)
race_model_v_native <- lm(V_RACE == "AMERICAN INDIAN/ALASKAN NATIVE" ~ P_RACE_BLACK +
                            P_RACE_NATIVE + P_RACE_API +
                            P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS, data = data)
race_model_v_blackhis <- lm(V_RACE == "BLACK HISPANIC" ~ P_RACE_BLACK + P_RACE_NATIVE +
                              P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE +
                              P_RACE_WHITEHIS, data = data)
race_model_v_whitehis <- lm(V_RACE == "WHITE HISPANIC" ~ P_RACE_BLACK + P_RACE_NATIVE +
                              P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE +
                              P_RACE_WHITEHIS, data = data)

summary(race_model_v_black) # [1] Black [2] White [3] White Hispanic
```

```
##
## Call:
## lm(formula = V_RACE == "BLACK" ~ P_RACE_BLACK + P_RACE_NATIVE +
##     P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7923 -0.3383  0.2077  0.2077  0.8586
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.33827    0.00849  39.844  < 2e-16 ***
## P_RACE_BLACK      0.45403    0.00934  48.611  < 2e-16 ***
## P_RACE_NATIVE     0.66173    0.30017   2.205   0.0275 *
## P_RACE_API       -0.00691    0.03373  -0.205   0.8377
## P_RACE_BLACKHIS   0.06649    0.01421   4.678 2.91e-06 ***
## P_RACE_WHITE     -0.19686    0.02604  -7.558 4.30e-14 ***
## P_RACE_WHITEHIS        NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4243 on 16224 degrees of freedom
##   (12332 observations deleted due to missingness)
## Multiple R-squared:  0.1826, Adjusted R-squared:  0.1823
## F-statistic: 724.8 on 5 and 16224 DF,  p-value: < 2.2e-16
```

```r
summary(race_model_v_white) # [1] White [2] Asian/Pacific Islander [3] White Hispanic
```

```
##
## Call:
## lm(formula = V_RACE == "WHITE" ~ P_RACE_BLACK + P_RACE_NATIVE +
##     P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55556 -0.01726 -0.01726 -0.01726  0.98274
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.041233   0.003218  12.812  < 2e-16 ***
## P_RACE_BLACK    -0.023974   0.003541  -6.771 1.32e-11 ***
## P_RACE_NATIVE   -0.041233   0.113784  -0.362  0.71707
## P_RACE_API       0.029773   0.012785   2.329  0.01988 *
## P_RACE_BLACKHIS -0.015259   0.005387  -2.832  0.00463 **
## P_RACE_WHITE     0.514323   0.009873  52.095  < 2e-16 ***
## P_RACE_WHITEHIS        NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1609 on 16224 degrees of freedom
##   (12332 observations deleted due to missingness)
## Multiple R-squared:  0.1676, Adjusted R-squared:  0.1673
## F-statistic: 653.3 on 5 and 16224 DF,  p-value: < 2.2e-16
```

```r
summary(race_model_v_api) # [1] Asian/Pacific Islander [2] White
```

```
##
## Call:
## lm(formula = V_RACE == "ASIAN / PACIFIC ISLANDER" ~ P_RACE_BLACK +
##     P_RACE_NATIVE + P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE +
##     P_RACE_WHITEHIS, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36095 -0.01443 -0.01381 -0.01381  0.98619
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.016813   0.002601   6.464 1.05e-10 ***
## P_RACE_BLACK    -0.003006   0.002861  -1.051  0.29341
## P_RACE_NATIVE   -0.016813   0.091955  -0.183  0.85492
## P_RACE_API       0.344133   0.010332  33.307  < 2e-16 ***
## P_RACE_BLACKHIS -0.002383   0.004354  -0.547  0.58410    
## P_RACE_WHITE     0.026958   0.007979   3.379  0.00073 ***
## P_RACE_WHITEHIS        NA         NA      NA       NA    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.13 on 16224 degrees of freedom
##   (12332 observations deleted due to missingness)
## Multiple R-squared:  0.06894,    Adjusted R-squared:  0.06866
## F-statistic: 240.3 on 5 and 16224 DF,  p-value: < 2.2e-16
```

```
summary(race_model_v_native) # No Significant Results
```

```
##
## Call:
## lm(formula = V_RACE == "AMERICAN INDIAN/ALASKAN NATIVE" ~ P_RACE_BLACK +
##     P_RACE_NATIVE + P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE +
##     P_RACE_WHITEHIS, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00040 -0.00034 -0.00034 -0.00034  0.99966
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.003e-04  3.512e-04   1.140    0.254
## P_RACE_BLACK    -6.356e-05  3.863e-04  -0.165    0.869
## P_RACE_NATIVE   -4.003e-04  1.242e-02  -0.032    0.974
## P_RACE_API      -4.003e-04  1.395e-03  -0.287    0.774
## P_RACE_BLACKHIS -4.003e-04  5.879e-04  -0.681    0.496
## P_RACE_WHITE    -4.003e-04  1.077e-03  -0.372    0.710
## P_RACE_WHITEHIS         NA         NA      NA       NA
##
## Residual standard error: 0.01755 on 16224 degrees of freedom
##   (12332 observations deleted due to missingness)
## Multiple R-squared:  4.141e-05,  Adjusted R-squared:  -0.0002668
## F-statistic: 0.1344 on 5 and 16224 DF,  p-value: 0.9845
```

```
summary(race_model_v_blackhis) # [1] Black [2] White [3] White Hispanic
```

```
##
## Call:
## lm(formula = V_RACE == "BLACK HISPANIC" ~ P_RACE_BLACK + P_RACE_NATIVE +
##     P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26335 -0.07063 -0.07063 -0.07063  0.92937
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.176141   0.005972  29.497  < 2e-16 ***
## P_RACE_BLACK    -0.105506   0.006569 -16.060  < 2e-16 ***
## P_RACE_NATIVE   -0.176141   0.211125  -0.834    0.404    
## P_RACE_API      -0.093301   0.023722  -3.933 8.42e-05 ***
## P_RACE_BLACKHIS  0.087207   0.009996   8.724  < 2e-16 ***
## P_RACE_WHITE    -0.098700   0.018319  -5.388 7.23e-08 ***
## P_RACE_WHITEHIS        NA         NA      NA       NA    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.2985 on 16224 degrees of freedom
##   (12332 observations deleted due to missingness)
## Multiple R-squared:  0.04096,    Adjusted R-squared:  0.04066 
## F-statistic: 138.6 on 5 and 16224 DF,  p-value: < 2.2e-16
```

```
summary(race_model_v_whitehis) # [1] Black [2] White Hispanic [3] White
```

```
## 
## Call:
## lm(formula = V_RACE == "WHITE HISPANIC" ~ P_RACE_BLACK + P_RACE_NATIVE + 
##     P_RACE_API + P_RACE_BLACKHIS + P_RACE_WHITE + P_RACE_WHITEHIS, 
##     data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.4271 -0.1057 -0.1057 -0.1057  0.8943 
## 
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      0.427142   0.007174  59.537   <2e-16 ***
## P_RACE_BLACK    -0.321484   0.007893 -40.731   <2e-16 ***
## P_RACE_NATIVE   -0.427142   0.253655  -1.684   0.0922 .  
## P_RACE_API      -0.273296   0.028501  -9.589   <2e-16 ***
## P_RACE_BLACKHIS -0.135655   0.012010 -11.295   <2e-16 ***
## P_RACE_WHITE    -0.245324   0.022009 -11.146   <2e-16 ***
## P_RACE_WHITEHIS        NA         NA      NA       NA    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.3586 on 16224 degrees of freedom
##   (12332 observations deleted due to missingness)
## Multiple R-squared:  0.1012, Adjusted R-squared:  0.1009 
## F-statistic: 365.2 on 5 and 16224 DF,  p-value: < 2.2e-16
```

From the race models above, the following results are derived at statistically significant levels:

1. In order of t-value, Black victims predict Black, White, and White Hispanic perpetrators
2. In order of t-value, White victims predict White, Asian/Pacific Islander, and White Hispanic perpetrators
3. In order of t-value, Asian/Pacific Islander victims predict Asian/Pacific Islander and White perpetrators
4. In order of t-value, Black Hispanic victims predict Black, White, and White Hispanic perpetrators
5. In order of t-value, White Hispanic victims predict Black, White Hispanic, and White perpetrators

```
data <- data %>%
  mutate(
    P_AGE_18 = ifelse(P_AGE == "<18", 1, 0),
    P_AGE_18_24 = ifelse(P_AGE == "18-24", 1, 0),
    P_AGE_25_44 = ifelse(P_AGE == "25-44", 1, 0),
    P_AGE_45_64 = ifelse(P_AGE == "45-64", 1, 0),
    P_AGE_65 = ifelse(P_AGE == "65+", 1, 0),
    )
```

```
age_model_18 <- lm(V_AGE == "<18" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
                     P_AGE_45_64 + P_AGE_65, data = data)
age_model_18_24 <- lm(V_AGE == "18-24" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
                        P_AGE_45_64 + P_AGE_65, data = data)
age_model_25_44 <- lm(V_AGE == "25-44" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
                        P_AGE_45_64 + P_AGE_65, data = data)
age_model_45_64 <- lm(V_AGE == "45-64" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
                        P_AGE_45_64 + P_AGE_65, data = data)
age_model_65 <- lm(V_AGE == "65+" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
                     P_AGE_45_64 + P_AGE_65, data = data)

summary(age_model_18) # [1] <18 [2] 18-24
```

```
##
## Call:
## lm(formula = V_AGE == "<18" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
##     P_AGE_45_64 + P_AGE_65, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31012 -0.12576 -0.04498 -0.04498  0.96974
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.974e-15  3.730e-02   0.000 1.000000
## P_AGE_18     3.101e-01  3.801e-02   8.158 3.68e-16 ***
## P_AGE_18_24  1.258e-01  3.749e-02   3.355 0.000797 ***
## P_AGE_25_44  4.498e-02  3.750e-02   1.199 0.230410
## P_AGE_45_64  3.026e-02  3.901e-02   0.776 0.437923
## P_AGE_65            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3007 on 14862 degrees of freedom
##   (13695 observations deleted due to missingness)
## Multiple R-squared:  0.06889,    Adjusted R-squared:  0.06864
## F-statistic: 274.9 on 4 and 14862 DF,  p-value: < 2.2e-16
```

```
summary(age_model_18_24) # [1] 18-24 [2] <18 [3] 24-44
```

```
##
## Call:
## lm(formula = V_AGE == "18-24" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
##     P_AGE_45_64 + P_AGE_65, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4422 -0.3881 -0.2599  0.5578  0.9692
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03077    0.05769   0.533    0.594
## P_AGE_18     0.35733    0.05879   6.078 1.25e-09 ***
## P_AGE_18_24  0.41141    0.05798   7.096 1.34e-12 ***
## P_AGE_25_44  0.22910    0.05800   3.950 7.85e-05 ***
## P_AGE_45_64  0.09171    0.06033   1.520    0.128
## P_AGE_65          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4651 on 14862 degrees of freedom
##   (13695 observations deleted due to missingness)
## Multiple R-squared:  0.04403,    Adjusted R-squared:  0.04377
## F-statistic: 171.1 on 4 and 14862 DF,  p-value: < 2.2e-16
```

```r
summary(age_model_25_44) # [1] 25-44 [2] <18
```

```
##
## Call:
## lm(formula = V_AGE == "25-44" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
##     P_AGE_45_64 + P_AGE_65, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5997 -0.3726 -0.2458  0.4003  0.7542
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.41538    0.05970   6.958 3.59e-12 ***
## P_AGE_18    -0.16955    0.06084  -2.787  0.00533 **
## P_AGE_18_24 -0.04278    0.06000  -0.713  0.47588
## P_AGE_25_44  0.18432    0.06002   3.071  0.00214 **
## P_AGE_45_64  0.12208    0.06243   1.955  0.05055 .
## P_AGE_65          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4813 on 14862 degrees of freedom
##   (13695 observations deleted due to missingness)
## Multiple R-squared:  0.06707,    Adjusted R-squared:  0.06682
## F-statistic: 267.1 on 4 and 14862 DF,  p-value: < 2.2e-16
```

```r
summary(age_model_45_64) # [1] <18 [2] 18-24 [3] 25-44
```

```
##
## Call:
```

```
## lm(formula = V_AGE == "45-64" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
##     P_AGE_45_64 + P_AGE_65, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36923 -0.08729 -0.05214 -0.05214  0.95298
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36923    0.03264  11.311   <2e-16 ***
## P_AGE_18    -0.32221    0.03327  -9.685   <2e-16 ***
## P_AGE_18_24 -0.31709    0.03281  -9.665   <2e-16 ***
## P_AGE_25_44 -0.28194    0.03282  -8.590   <2e-16 ***
## P_AGE_45_64 -0.07816    0.03414  -2.290   0.0221 *
## P_AGE_65          NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2632 on 14862 degrees of freedom
##   (13695 observations deleted due to missingness)
## Multiple R-squared:  0.04049,    Adjusted R-squared:  0.04024
## F-statistic: 156.8 on 4 and 14862 DF,  p-value: < 2.2e-16
```

```r
summary(age_model_65) # [1] 18-24 [2] 25-44 [3] <18
```

```
##
## Call:
## lm(formula = V_AGE == "65+" ~ P_AGE_18 + P_AGE_18_24 + P_AGE_25_44 +
##     P_AGE_45_64 + P_AGE_65, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18462 -0.00816 -0.00816 -0.00732  0.99268
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18462    0.01172   15.75   <2e-16 ***
## P_AGE_18    -0.17569    0.01194  -14.71   <2e-16 ***
## P_AGE_18_24 -0.17730    0.01178  -15.05   <2e-16 ***
## P_AGE_25_44 -0.17645    0.01178  -14.98   <2e-16 ***
## P_AGE_45_64 -0.16588    0.01226  -13.54   <2e-16 ***
## P_AGE_65          NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.09448 on 14862 degrees of freedom
##   (13695 observations deleted due to missingness)
## Multiple R-squared:  0.01553,    Adjusted R-squared:  0.01526
## F-statistic:  58.6 on 4 and 14862 DF,  p-value: < 2.2e-16
```

From the age models above the following results are derived at statistically significant levels:

1. In order of t-value, victims below age 18 predict “<18” and “18-24” perpetrators

2. In order of t-value, victims between 18-24 predict “18-24”, “<18” and “24-44” perpetrators
3. In order of t-value, victims between 25-44 predict “25-44” and “<18” perpetrators
4. In order of t-value, victims between 45-64 predict “<18”, “18-24” and “25-44” perpetrators
5. In order of t-value, victims older than 65 predict “18-24”, “25-44” and “<18” perpetrators

Conclusion

There is evidence that majority of the shooting incidents captured in this report’s data are race-on-race, age-on-age violence. Specifically, the data suggests that Black people shoot Black victims; White shooters shoot White victims; perpetrators from the ages 18-24 shoot 18-24 year olds; and perpetrators ages 25-44 shoot 25-44 year olds. The main implication to draw from race-on-race, age-on-age violence is that these shootings are not random, but likely violence between family members, acquaintances, and neighbors.

Points of entry for biases in this analysis start right at data clean up. One important prevention measure I took to avoid biases in my analysis is choosing to change all unknown values into NA-types, instead of imputing numbers based on inferred distributions. And yet, bias makes its way into a data set from the outset: the questions that I had to ask of the data, informed by preconceived notions about crime and shooting in New York City, impacted the data I counted as important. The absence of data will no doubt affect the kinds of results I can draw from the data set. The best way for me to avoid drawing biased conclusions from this data is gathering others’ conclusions from the same data set.