NYPD Shooting Incident Data Report

Julia Hoglund

2022-07-19

Read CSV

The data is imported from the Data.gov data set provided to us.

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

Clean and Transform Data

The data is cleaned by removing unnecessary columns. For this data set I am interested in seeing if there are any trends or correlations between number of incidents verses murders per year depending on perpetrator sex or age. The only columns I need are OCCUR_DATE, STATISTICAL_MURDER_FLAG, PERP_SEX, and PERP_AGE_GROUP so I will remove the rest of the data. I will also transform the data by updating the OCCUR_DATE column to be in a date format and adding a YEAR column so I know in which year each incident took place. I removed any rows that include a NaN value to eliminate any missing data as well as remove some unknown values from PERP_SEX and PERP_AGE_GROUP.

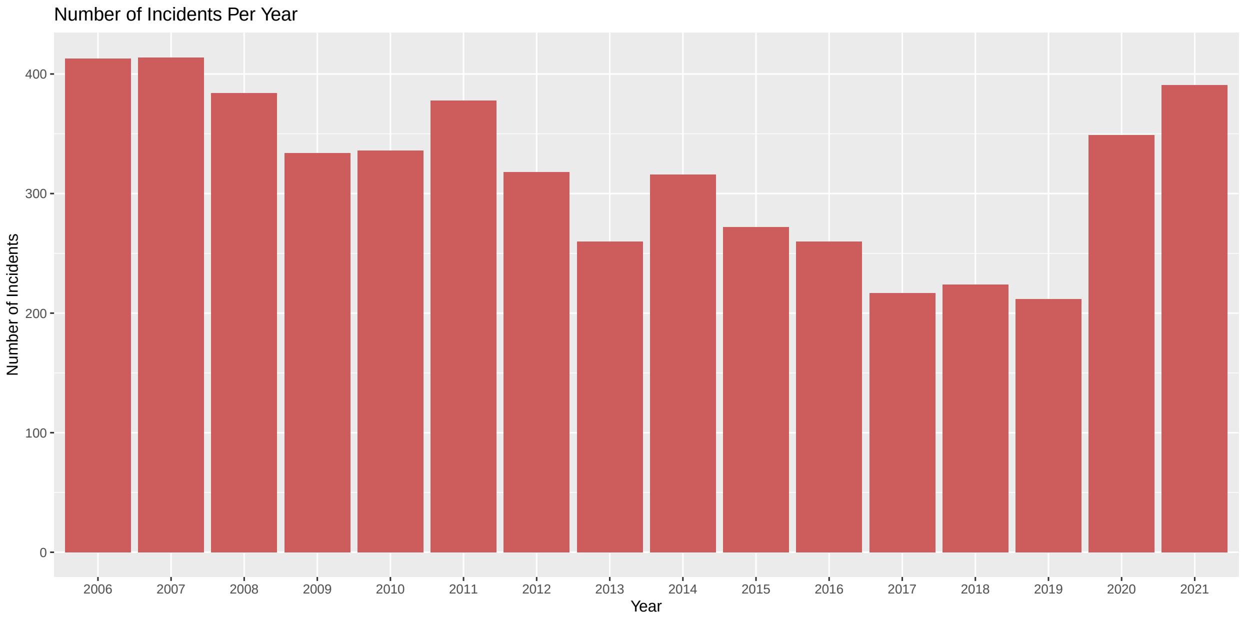
```
data_mod <- data %>%
            select(OCCUR_DATE, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SEX) %>%
            mutate(OCCUR_DATE = as.Date(data$OCCUR_DATE,format="%d/%m/%Y")) %>%
            na.exclude
data_mod <- data_mod %>%
            mutate(YEAR = format(data_mod$OCCUR_DATE,"%Y"))
data_mod <- subset(data_mod, PERP_SEX !="U")
data_mod <- subset(data_mod, PERP_AGE_GROUP != 224)
data_mod <- subset(data_mod, PERP_AGE_GROUP != 940)
data_mod <- subset(data_mod, PERP_AGE_GROUP != "UNKNOWN")
summary(data_mod)
```

```
##    OCCUR_DATE         STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Min.   :2006-01-01   Mode :logical           Length:5078
##  1st Qu.:2009-02-10   FALSE:3822              Class :character
##  Median :2012-11-04   TRUE :1256              Mode  :character
##  Mean   :2013-06-02
##  3rd Qu.:2017-07-11
##  Max.   :2021-12-12
##    PERP_SEX             YEAR
##  Length:5078        Length:5078
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Visualize Data

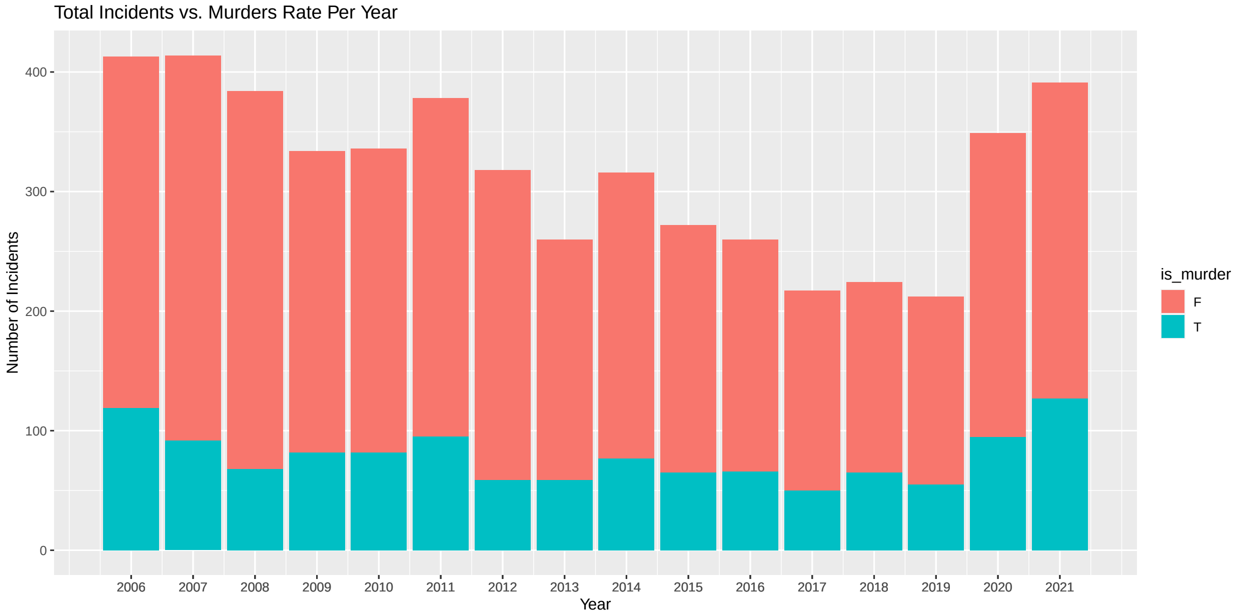
Visualization 1 For this first visualization, I was interested in seeing how many total incidents occurred for each year in the data set. This plot shows a fairly steady decrease in total incidents from 2006 to 2019 and then a significant increase in 2020 and 2021.

```
data_year <- data_mod %>% count(YEAR)
ggplot(data=data_year, aes(x=YEAR, y=n)) +
       geom_bar(stat="identity", fill="indian red") +
       labs(title = str_c("Number of Incidents Per Year"),
            y = str_c("Number of Incidents"),
            x = str_c("Year"))
```



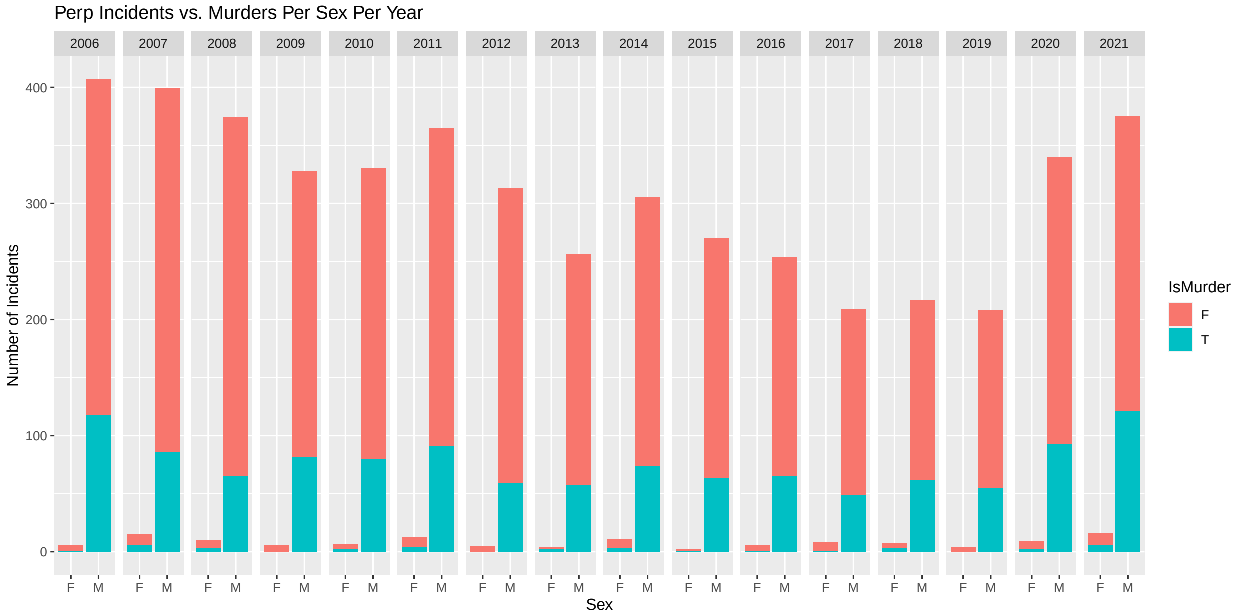
Visualization 2 The first visualization led me to compare the number of incidents verses the number of murders that occurred per year. The plot below shows the number of murders in blue and the number of non-fatal incidents in red. They are stacked on top of each other to give an easy comparison against the total number of incidents that occur. The number of murders appears to occur at a much lower rate than non-fatal incidents. The plot also shows that the rate of murders is roughly proportional to the rate of non-fatal incidents.

```
murder <- factor(data_mod$STATISTICAL_MURDER_FLAG)
murder <- relevel(murder, "TRUE")
table_murder_rate <- table(murder,data_mod$YEAR)

vec_num_true <- unname(table_murder_rate[1,])
vec_num_false <- unname(table_murder_rate[2,])
num_incidents <- append(vec_num_true, vec_num_false)
is_murder <- c(rep("T", ncol(table_murder_rate)),
               rep("F", ncol(table_murder_rate)))
vec_year <- c(rep(min(data_mod$YEAR):max(data_mod$YEAR),2))
data_murders_rate <- data.frame(vec_year, is_murder, num_incidents)
ggplot(data_murders_rate, aes(x = vec_year,
                              y = num_incidents,
                              fill = is_murder)) +
       geom_bar(stat = "identity", position = "stack") +
       scale_x_continuous(labels=as.character(vec_year),breaks=vec_year) +
       labs(title = "Total Incidents vs. Murders Rate Per Year",
            y = "Number of Incidents",
            x = "Year")
```



Visualization 3 Now that I've compared the number of murders verses the number of non-fatal incidents, I was interested in seeing if the perpetrators sex or age had any impact on the number of incidents committed per year. The plot below compares the number of murders against the total number of incidents per year and per sex, male or female. This plot shows a very clear majority of male perpetrators to female perptetrators for each year.

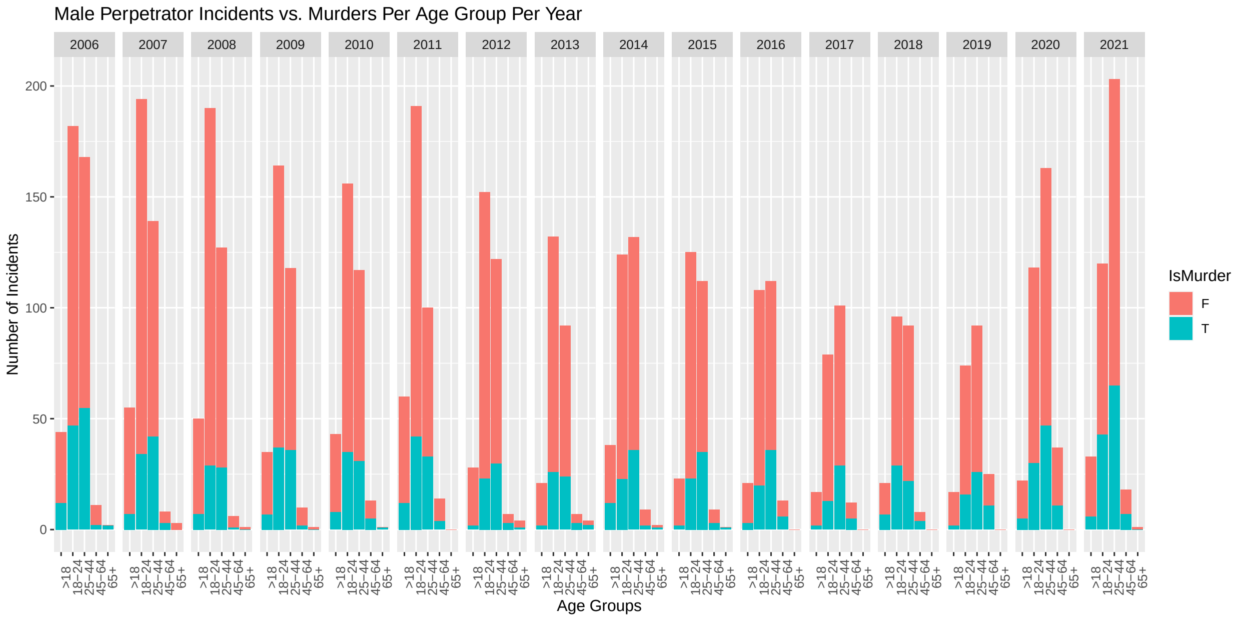
```
num_tf_per_sex <- table(data_mod$YEAR,
                        data_mod$STATISTICAL_MURDER_FLAG,
                        data_mod$PERP_SEX)
num_F_FALSE_per_year <- unname(num_tf_per_sex[,1,1])
num_F_TRUE_per_year <- unname(num_tf_per_sex[,2,1])
num_M_FALSE_per_year <- unname(num_tf_per_sex[,1,2])
num_M_TRUE_per_year <- unname(num_tf_per_sex[,2,2])
num_incidents_sex <- c(num_F_TRUE_per_year,
                       num_F_FALSE_per_year,
                       num_M_TRUE_per_year,
                       num_M_FALSE_per_year)
vec_year_sex <- c(rep(vec_year,2))
is_murder_sex <- rep(is_murder,2)
sex_flag <- c(rep(rep("F", ncol(table_murder_rate)),2),
              rep(rep("M", ncol(table_murder_rate)),2))
data_murders_rate_per_sex <- data.frame(vec_year_sex,
                                        is_murder_sex,
                                        sex_flag,
                                          num_incidents_sex)
IsMurder <- is_murder_sex
ggplot(data_murders_rate_per_sex, aes(x = sex_flag,
                                      y = num_incidents_sex,
                                      fill = IsMurder)) +
       geom_bar(stat = "identity", position = "stack") +
       facet_grid(~ vec_year_sex) +
       labs(title = "Perp Incidents vs. Murders Per Sex Per Year",
            y = "Number of Incidents",
            x = "Sex")
```



Visualization 4 For this visualization I wanted to continue this line of analysis to see if the perpetrators age group has any correlation to the rate of the incidents. Since the number of female perpetrators was so low compared to the number of male perpetrators, this plot focuses only on the age groups for the male perpetrators. The plot below shows several different trends. It is clear that the 18-24 and the 25-44 age group accounts for the most total number of incidents each year. Throughout the years, there appears to be a slight trend in the 25-44 age group committing more incidents in more recent years (2020, 2021) verses the 18-24 age group committing more incidents in earlier years (2006, 2007, etc.). There also appears to be a slight trend in which the 25-44 age group has a higher rate of murders verses non-fatal incidents compared to the 18-24 age group. This visualization leads me to question if there are other factors that impact the rate of incidents and murders, like which borough the incident took place in or the victim sex or age. If I were to continue this line of analysis, I would add the BORO or VIC_SEX or VIC_AGE_GROUP columns back into the data set and set up some visualizations of the number of incidents and murders against those factors.

```
data_mod_male <- subset(data_mod, PERP_SEX !="F")
num_tf_per_age <- table(data_mod_male$YEAR,
                        data_mod_male$STATISTICAL_MURDER_FLAG,
                        data_mod_male$PERP_AGE_GROUP)
num_18_FALSE_per_year <- unname(num_tf_per_age[,1,1])
num_18_TRUE_per_year <- unname(num_tf_per_age[,2,1])
num_18_24_FALSE_per_year <- unname(num_tf_per_age[,1,2])
```

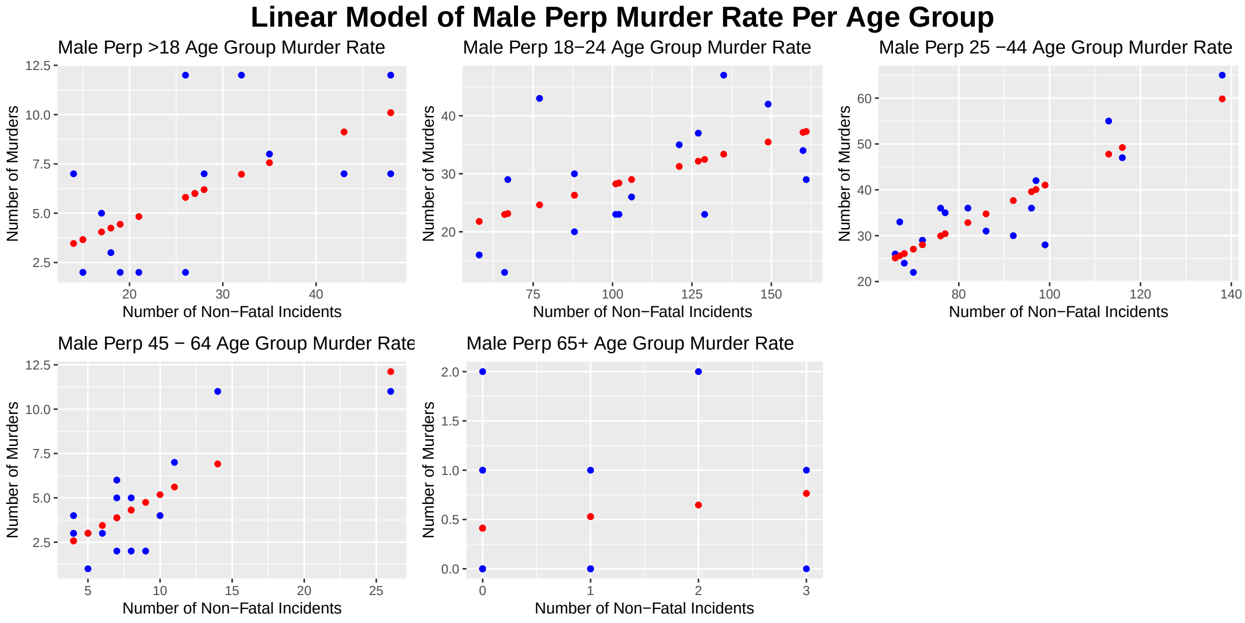
```r
num_18_24_TRUE_per_year <- unname(num_tf_per_age[,2,2])
num_25_44_FALSE_per_year <- unname(num_tf_per_age[,1,3])
num_25_44_TRUE_per_year <- unname(num_tf_per_age[,2,3])
num_45_64_FALSE_per_year <- unname(num_tf_per_age[,1,4])
num_45_64_TRUE_per_year <- unname(num_tf_per_age[,2,4])
num_65_FALSE_per_year <- unname(num_tf_per_age[,1,5])
num_65_TRUE_per_year <- unname(num_tf_per_age[,2,5])
num_incidents_male <- c(num_18_TRUE_per_year,
                        num_18_FALSE_per_year,
                        num_18_24_TRUE_per_year,
                        num_18_24_FALSE_per_year,
                        num_25_44_TRUE_per_year,
                        num_25_44_FALSE_per_year,
                        num_45_64_TRUE_per_year,
                        num_45_64_FALSE_per_year,
                        num_65_TRUE_per_year,
                        num_65_FALSE_per_year)
vec_year_male <- c(rep(vec_year,5))
is_murder_male <- rep(is_murder,5)
age_group_male <- c(rep(rep(">18", ncol(table_murder_rate)),2),
                   rep(rep("18-24", ncol(table_murder_rate)),2),
                   rep(rep("25-44", ncol(table_murder_rate)),2),
                   rep(rep("45-64", ncol(table_murder_rate)),2),
                   rep(rep("65+", ncol(table_murder_rate)),2))
data_murders_rate_male_per_age <- data.frame(vec_year_male,
                                             is_murder_male,
                                             age_group_male,
                                             num_incidents_male)
IsMurder <- is_murder_male
ggplot(data_murders_rate_male_per_age, aes(x = age_group_male,
                                           y = num_incidents_male,
                                           fill = IsMurder)) +
       geom_bar(stat = "identity", position = "stack") +
       facet_grid(~ vec_year_male) +
       theme(axis.text.x=element_text(angle=90)) +
       labs(title = "Male Perpetrator Incidents vs. Murders Per Age Group Per Year",
            y = "Number of Incidents",
            x = "Age Groups")
```



Analysis and Modeling

For my analysis and modeling section, I was interested in looking at the linear relationship between the rate of murders verses the number of non-fatal incidents for each age group for male perpetrators. The group of plots below is the comparison of the number of non-fatal incidents against the number of murders which is plotted in blue and the linear model of the two variables is plotted in red. There is a separate plot for each age group. The plots show that there is a pretty linear relationship between murders and incidents for the 25-44 age group as well as the 18-24 age group, but with a few more outliers in the latter group. The relationships for the other age groups also appear to be roughly linear but it is difficult to be certain due to so few data points for those groups. If I were to do other modeling, I would be interested in determining if another type of distribution would be a more appropriate model for this data.

```
data_murders_rate <- data.frame(vec_year,
                                num_18_FALSE_per_year,
                                num_18_TRUE_per_year,
                                num_18_24_FALSE_per_year,
                                num_18_24_TRUE_per_year,
                                num_25_44_FALSE_per_year,
                                num_25_44_TRUE_per_year,
                                num_45_64_FALSE_per_year,
                                num_45_64_TRUE_per_year,
                                num_65_FALSE_per_year,
                                num_65_TRUE_per_year)

mod_18 <- lm(num_18_TRUE_per_year ~ num_18_FALSE_per_year, data_murders_rate)
data_murders_rate_18_pred <- data_murders_rate %>%
                   mutate(pred = predict(mod_18))
fig_18 <- data_murders_rate_18_pred %>%
         ggplot() +
         geom_point(aes(x = num_18_FALSE_per_year,
                        y = num_18_TRUE_per_year), color = "blue") +
         geom_point(aes(x = num_18_FALSE_per_year,
                         y = pred), color = "red") +
         theme(legend.position = "bottom") +
         labs(title = "Male Perp >18 Age Group Murder Rate",
              y = "Number of Murders",
              x = "Number of Non-Fatal Incidents")

mod_18_24 <- lm(num_18_24_TRUE_per_year ~ num_18_24_FALSE_per_year, data_murders_rate)
data_murders_rate_18_24_pred <- data_murders_rate %>%
                   mutate(pred = predict(mod_18_24))
fig_18_24 <- data_murders_rate_18_24_pred %>%
             ggplot() +
             geom_point(aes(x = num_18_24_FALSE_per_year,
                            y = num_18_24_TRUE_per_year), color = "blue") +
             geom_point(aes(x = num_18_24_FALSE_per_year,
                            y = pred), color = "red") +
             theme(legend.position = "bottom") +
             labs(title = "Male Perp 18-24 Age Group Murder Rate",
                  y = "Number of Murders",
                  x = "Number of Non-Fatal Incidents")

mod_25_44 <- lm(num_25_44_TRUE_per_year ~ num_25_44_FALSE_per_year, data_murders_rate)
data_murders_rate_25_44_pred <- data_murders_rate %>%
                   mutate(pred = predict(mod_25_44))
fig_25_44 <- data_murders_rate_25_44_pred %>%
             ggplot() +
             geom_point(aes(x = num_25_44_FALSE_per_year,
                            y = num_25_44_TRUE_per_year), color = "blue") +
             geom_point(aes(x = num_25_44_FALSE_per_year,
                            y = pred), color = "red") +
             theme(legend.position = "bottom") +
             labs(title = "Male Perp 25 -44 Age Group Murder Rate",
                  y = "Number of Murders",
                  x = "Number of Non-Fatal Incidents")

mod_45_64 <- lm(num_45_64_TRUE_per_year ~ num_45_64_FALSE_per_year, data_murders_rate)
data_murders_rate_45_64_pred <- data_murders_rate %>%
                   mutate(pred = predict(mod_45_64))
fig_45_64 <- data_murders_rate_45_64_pred %>%
             ggplot() +
             geom_point(aes(x = num_45_64_FALSE_per_year,
                            y = num_45_64_TRUE_per_year), color = "blue") +
             geom_point(aes(x = num_45_64_FALSE_per_year,
                            y = pred), color = "red") +
             theme(legend.position = "bottom") +
             labs(title = "Male Perp 45 - 64 Age Group Murder Rate",
                  y = "Number of Murders",
                  x = "Number of Non-Fatal Incidents")

mod_65 <- lm(num_65_TRUE_per_year ~ num_65_FALSE_per_year, data_murders_rate)
data_murders_rate_65_pred <- data_murders_rate %>%
                   mutate(pred = predict(mod_65))
fig_65 <- data_murders_rate_65_pred %>%
          ggplot() +
          geom_point(aes(x = num_65_FALSE_per_year,
                         y = num_65_TRUE_per_year), color = "blue") +
          geom_point(aes(x = num_65_FALSE_per_year,
                         y = pred), color = "red") +
          theme(legend.position = "bottom") +
          labs(title = "Male Perp 65+ Age Group Murder Rate",
               y = "Number of Murders",
               x = "Number of Non-Fatal Incidents")

figure <- ggarrange(fig_18, fig_18_24, fig_25_44, fig_45_64, fig_65,
                    ncol = 3, nrow = 2)
annotate_figure(figure, top = text_grob("Linear Model of Male Perp Murder Rate Per Age Group",
                face = "bold", size = 20))
```



Bias and Conclusion

To conclude this report, it was very interesting to visualize the data and it led to some interesting results, like showing clear trends in the number of incidents throughout the years and determining that the male sex and the 18-24 and 25-44 age groups are more likely to commit the most incidents. A linear model also showed which age groups have the most linear relationship between the number of murders and the number of non-fatal incidents.

Possible sources of bias include bias against which certain perpetrator sexes or races makes it more likely for them to be involved in an incident or not. I have an assumption that males are more likely to be involved in incidents than females. I also have an assumption that there is a higher rate of Black and Latino perpetrators involved in an incident, due to systemic and geopolitical differences in neighborhoods. I mitigated these potential biases by doing the minimal amount of transforming of the data, only to the formats that I needed, and letting the results from the plots inform my analysis instead of the other way around. I also focused on drawing comparisons against factors that are more likely to be unbiased like the date or year.

Session Info

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggpubr_0.4.0    lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0
##  [5] dplyr_1.0.9     purrr_0.3.4     readr_2.1.2     tidyr_1.2.0
##  [9] tibble_3.1.7    ggplot2_3.3.6   tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] assertthat_0.2.1 digest_0.6.29    utf8_1.2.2       R6_2.5.1
##  [5] cellranger_1.1.0 backports_1.4.1  reprex_2.0.1     evaluate_0.15
##  [9] highr_0.9        httr_1.4.3       pillar_1.7.0     rlang_1.0.3
## [13] curl_4.3.2       readxl_1.4.0     rstudioapi_0.13  car_3.1-0
## [17] rmarkdown_2.14   labeling_0.4.2   bit_4.0.4        munsell_0.5.0
## [21] broom_1.0.0      compiler_4.2.1   modelr_0.1.8     xfun_0.31
## [25] pkgconfig_2.0.3  htmltools_0.5.2  tidyselect_1.1.2 gridExtra_2.3
## [29] fansi_1.0.3      crayon_1.5.1     tzdb_0.3.0       dbplyr_2.2.1
## [33] withr_2.5.0      grid_4.2.1       jsonlite_1.8.0   gtable_0.3.0
## [37] lifecycle_1.0.1  DBI_1.1.3        magrittr_2.0.3   scales_1.2.0
## [41] cli_3.3.0        stringi_1.7.6    vroom_1.5.7      carData_3.0-5
## [45] farver_2.1.0     ggsignif_0.6.3   fs_1.5.2         xml2_1.3.3
## [49] ellipsis_0.3.2   generics_0.1.3   vctrs_0.4.1      cowplot_1.1.1
## [53] tools_4.2.1      bit64_4.0.5      glue_1.6.2       hms_1.1.1
## [57] parallel_4.2.1   abind_1.4-5      fastmap_1.1.0    yaml_2.3.5
## [61] colorspace_2.0-3 rstatix_0.7.0    rvest_1.0.2      knitr_1.39
## [65] haven_2.5.0
```