NYPD Shooting Incidents

Artur Ayzenberg

2025-03-30

Step 1 importing the NYPD shooting data

```
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 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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(lubridate)
library(ggplot2)
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shooting <- read.csv(url_in)
```

Step 2 tidy the data

In the below step NYPD shooting data is collected. Next the data is further assigned to variables for step 3 visualization and analysis. One such point will be the year_frequency variable to determine if shootings are on the decline or on the rise. Next 2 separate variables are collected, age breakdown of perpetrators and victims.

```
nypd_shooting <- nypd_shooting %>% select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
nypd_shooting <- nypd_shooting %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
summary(nypd_shooting)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME            BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:29744       Length:29744
##  1st Qu.: 67321140   1st Qu.:2009-10-29   Class :character   Class :character
##  Median :109291972   Median :2014-03-25   Mode  :character   Mode  :character
##  Mean   :133850951   Mean   :2014-10-31
##  3rd Qu.:214741917   3rd Qu.:2020-06-29
##  Max.   :299462478   Max.   :2024-12-31
```

```
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:29744       Min.   :  1.00   Min.   :0.0000    Length:29744
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.00   Median :0.0000    Mode  :character
##                     Mean   : 65.23   Mean   :0.3181
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:29744       Length:29744            Length:29744
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:29744       Length:29744       Length:29744       Length:29744
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE
##  Length:29744
##  Class :character
##  Mode  :character
##
##
##
##
```

```
##### Isolate the yearly occurrences #####
yearly_totals <- format(as.Date(nypd_shooting$OCCUR_DATE, format="%Y/%m/%d"),"%Y")
yearly_totals <- table(yearly_totals)
year_frequency <- as.data.frame.table(yearly_totals)
year_frequency$yearly_totals <- as.numeric(as.character(year_frequency$yearly_totals))

##### Victims and Perpetrators by age #####
# Perp ages
ages_of_perp <- nypd_shooting %>% select(c(PERP_AGE_GROUP))
ages_of_perp <- as.data.frame.table(ages_of_perp)
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP %in% "UNKNOWN"),]
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP %in% "(null)"),]
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP == ""),]
table(ages_of_perp$PERP_AGE_GROUP) # Step is done to determine any ages/age groups that din't make sens
```

```
##
##   <18  1020  1028 18-24  2021   224 25-44 45-64   65+   940
##  1805     1     1  6630     1     1  6342   775    67     1
```

```
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP == "1020"),]
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP == "1028"),]
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP == "224"),]
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP == "940"),]
ages_of_perp <- ages_of_perp[!(ages_of_perp$PERP_AGE_GROUP == "2021"),]
ages_of_perp <- ages_of_perp[c("PERP_AGE_GROUP")]
ages_of_perp <- table(ages_of_perp)
ages_of_perp_groups <- as.data.frame.table(ages_of_perp)
# Victim ages
ages_of_vic <- nypd_shooting %>% select(c(VIC_AGE_GROUP))
ages_of_vic <- as.data.frame.table(ages_of_vic)
ages_of_vic <- ages_of_vic[!(ages_of_vic$VIC_AGE_GROUP %in% "UNKNOWN"),]
ages_of_vic <- ages_of_vic[!(ages_of_vic$VIC_AGE_GROUP %in% "(null)"),]
ages_of_vic <- ages_of_vic[!(ages_of_vic$VIC_AGE_GROUP == ""),]
table(ages_of_vic$VIC_AGE_GROUP) # Step is done to determine any ages/age groups that din't make sense
```

```
##
##   <18  1022 18-24 25-44 45-64   65+
##  3081     1 10677 13563  2118   236
```

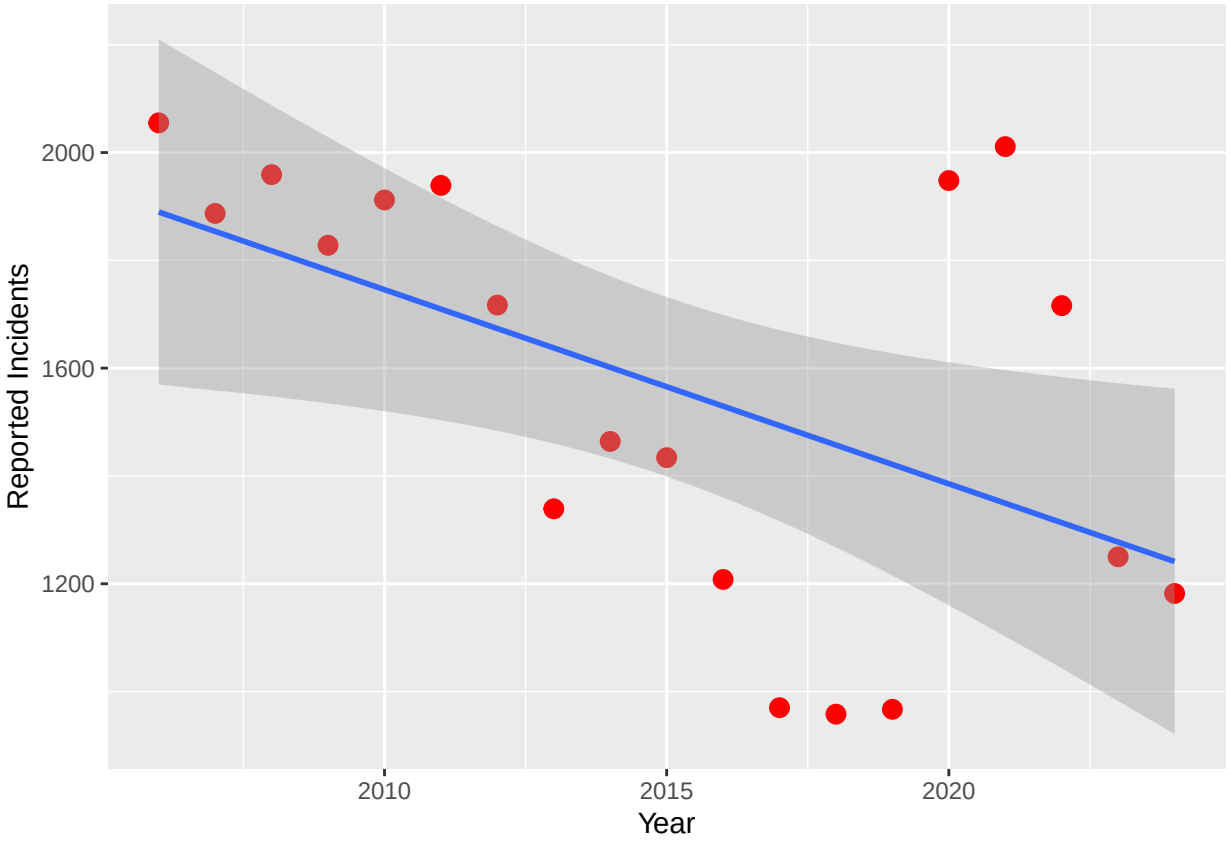
```
ages_of_vic <- ages_of_vic[!(ages_of_vic$VIC_AGE_GROUP == "1022"),]
ages_of_vic <- ages_of_vic[c("VIC_AGE_GROUP")]
ages_of_vic <- table(ages_of_vic)
ages_of_vic_groups <- as.data.frame.table(ages_of_vic)
```

Step 3 Visualize the data

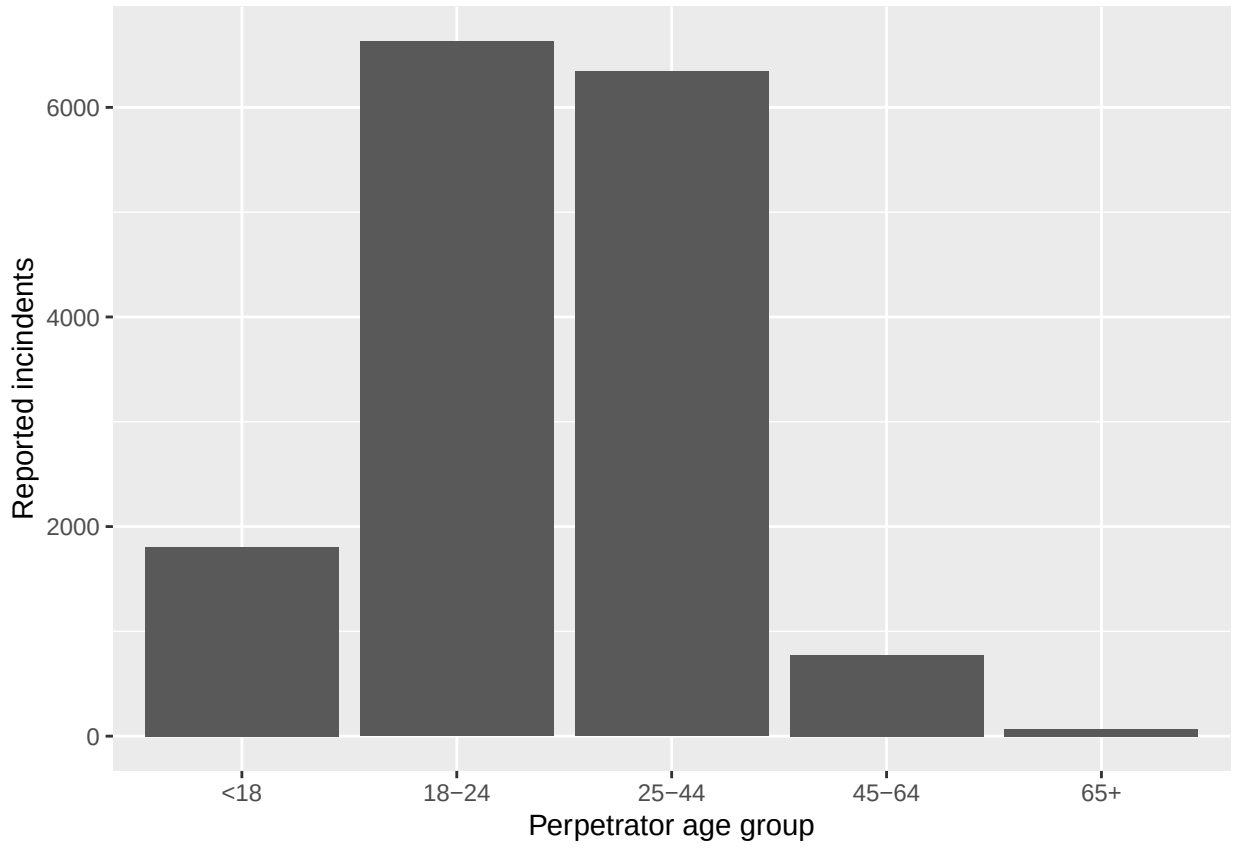
Based on the variable above data visualization will be presented as well as some respective analysis. Analysis will be based on R calculations or other possible tests

```
#Frequency visualization
ggplot(year_frequency, aes(x = yearly_totals, y = Freq)) + labs(x="Year", y="Reported Incidents") + geo
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Perpetrator ages
ggplot(ages_of_perp_groups, aes(x=PERP_AGE_GROUP, y=Freq)) + labs(x="Perpetrator age group", y="Reporte
```



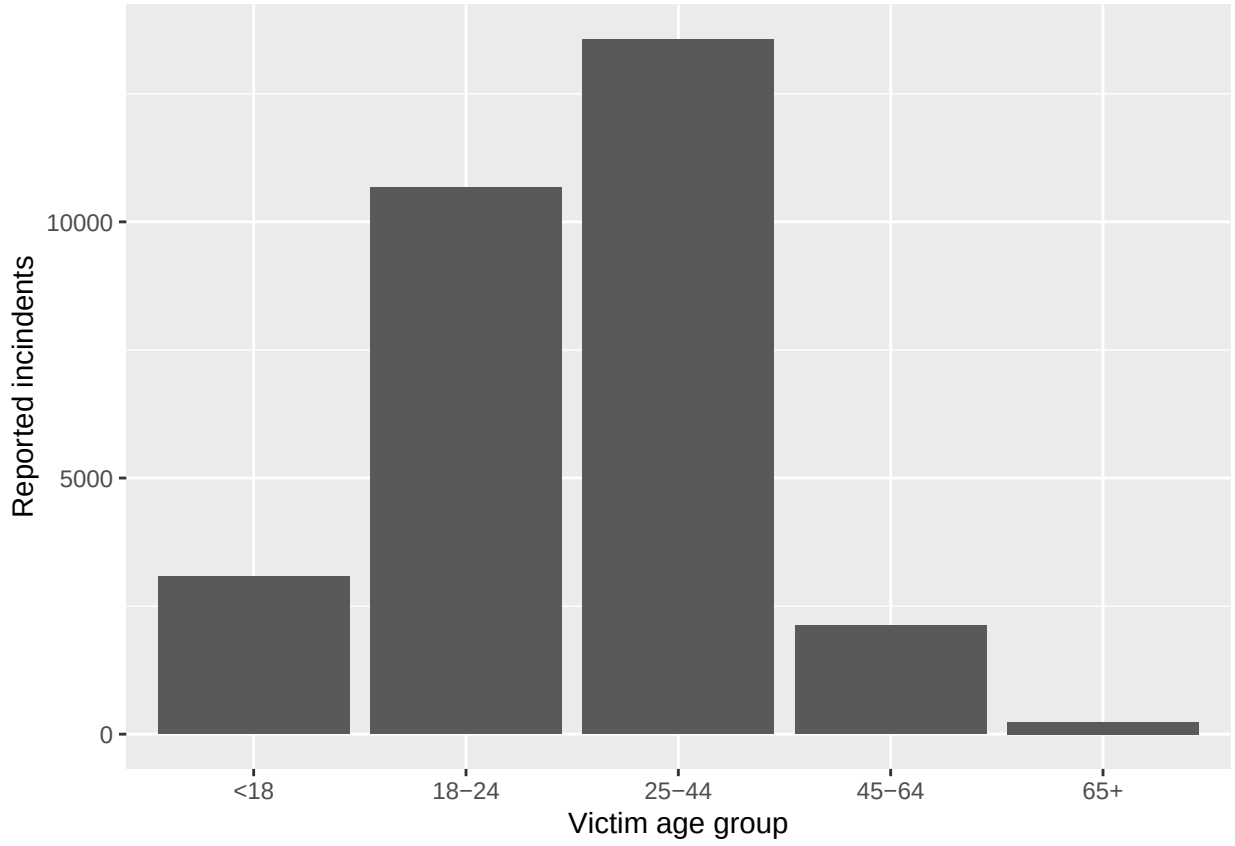

```
sum(ages_of_perp_groups$Freq)
```

```
## [1] 15619
```

```
ages_of_perp_groups
```

```
##   PERP_AGE_GROUP Freq
## 1            <18 1805
## 2          18-24 6630
## 3          25-44 6342
## 4          45-64  775
## 5            65+   67
```

```
#Victim ages
ggplot(ages_of_vic_groups, aes(x=VIC_AGE_GROUP, y=Freq)) + labs(x="Victim age group", y="Reported incin
```




```
sum(ages_of_vic_groups$Freq)
```

```
## [1] 29675
```

```
ages_of_vic_groups
```

```
##   VIC_AGE_GROUP  Freq
## 1           <18  3081
## 2         18-24 10677
## 3         25-44 13563
## 4         45-64  2118
## 5           65+   236
```

Step 3 Analyze the data

The following includes the analysis for items visualized above.

```
#Frequency analysis
year_model <- lm(year_frequency$Freq ~ year_frequency$yearly_totals, data = year_frequency)
summary(year_model)
```

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

```
## lm(formula = year_frequency$Freq ~ year_frequency$yearly_totals,
##     data = year_frequency)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523.42 -218.01   33.32  165.84  661.68
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  74158.50   29035.88   2.554   0.0205 *
## year_frequency$yearly_totals   -36.03      14.41  -2.500   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 344 on 17 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2258
## F-statistic: 6.251 on 1 and 17 DF,  p-value: 0.02294
```

```
#Perpetrator ages
ratio_perp <- 1682/sum(ages_of_perp_groups$Freq)
ratio_victim <- 2954/sum(ages_of_vic_groups$Freq)
std_err <- sqrt(((ratio_perp * (1 - ratio_perp)) / sum(ages_of_perp_groups$Freq)) + ((ratio_victim * (1
z_score <- (ratio_perp - ratio_victim) / std_err
p_value <- pnorm(q=z_score, lower.tail=FALSE)
ratio_perp
```

```
## [1] 0.1076894
```

```
ratio_victim
```

```
## [1] 0.09954507
```

```
std_err
```

```
## [1] 0.003028673
```

```
z_score
```

```
## [1] 2.68906
```

```
p_value
```

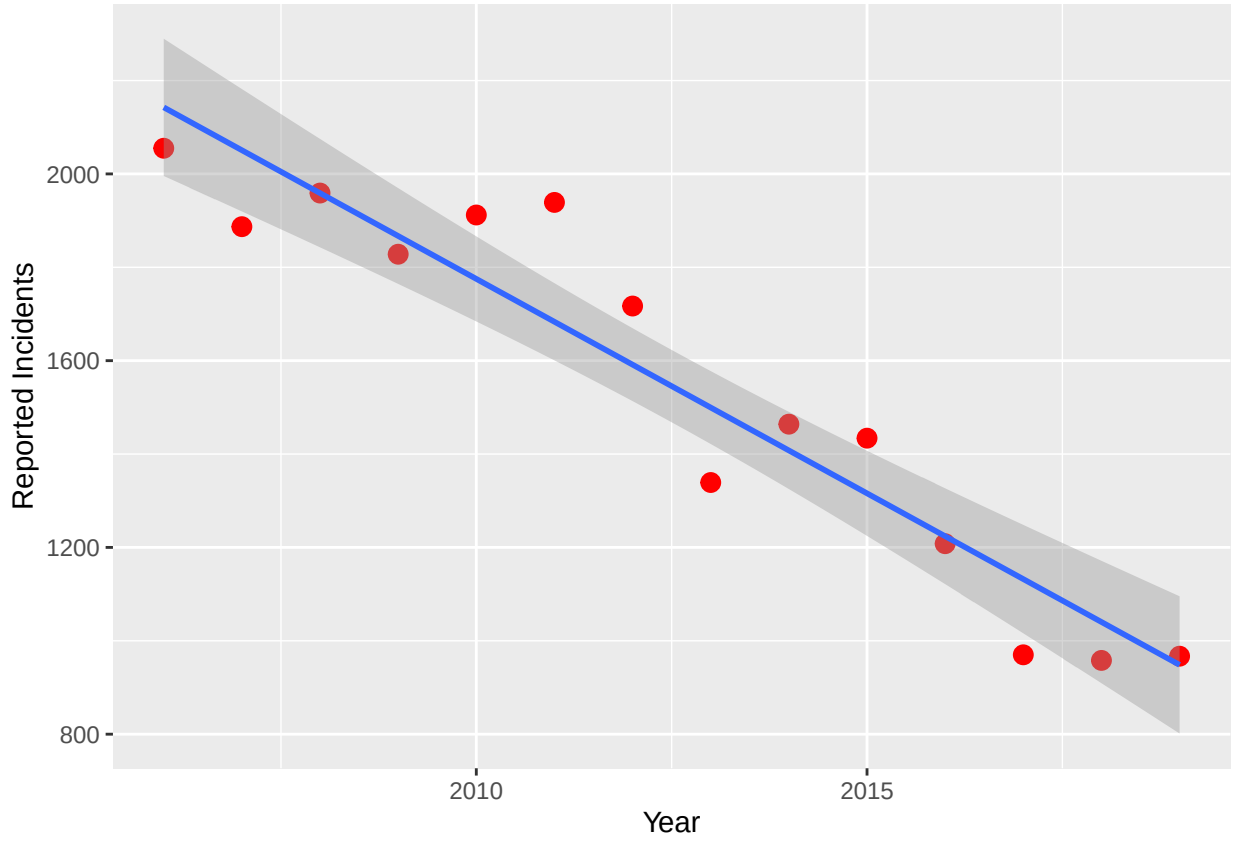
```
## [1] 0.003582682
```

Step 3 Extras

Here is a thought experiment to do with yearly shooting data. When looking at the graph we almost see 2 separate groups. Before covid there appears to be a clear trend downwards which takes a rapid spike in 2020, after which we see it quickly dropping down to pre covid levels. As a thought experiment below is the data separated into 2 separate groups, visualized and trend lines drawn and shown.

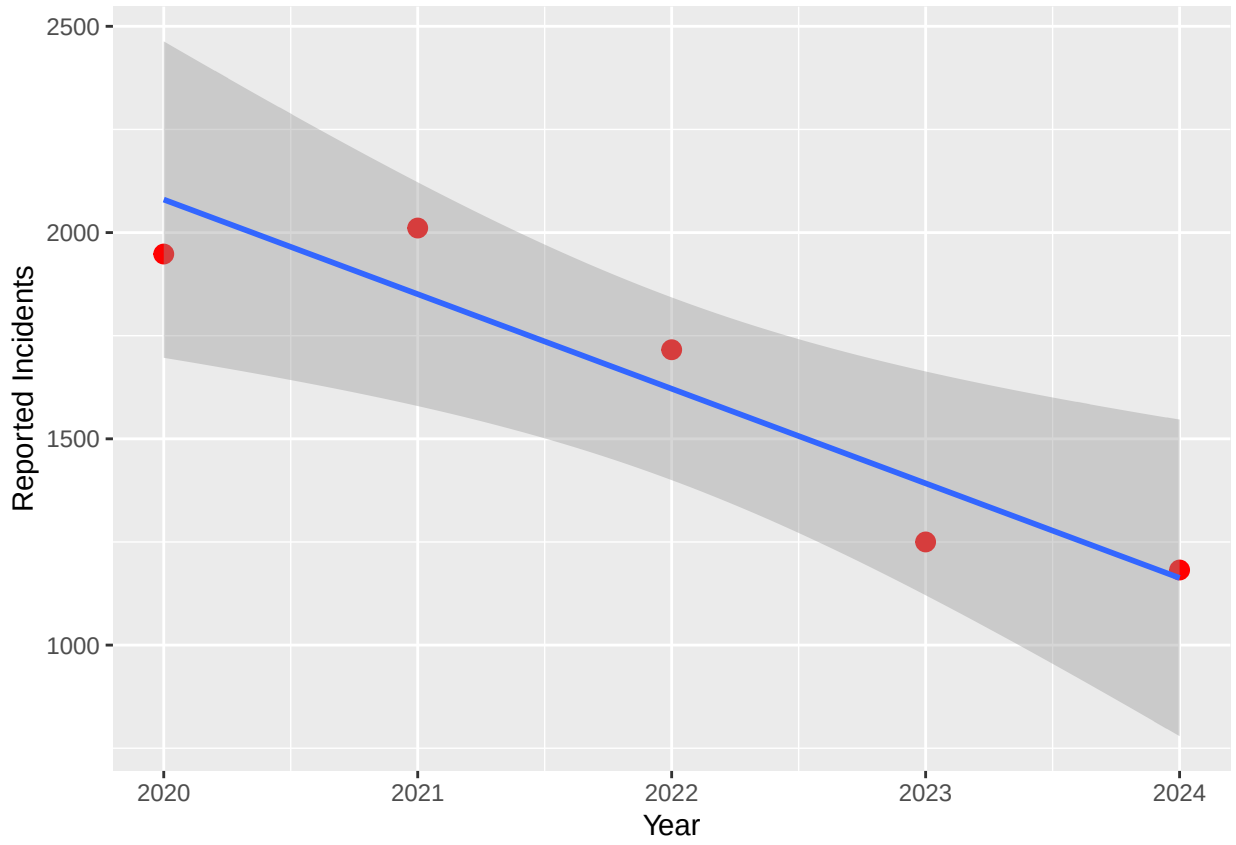
```r
#Get 2 separate populations
pre_covid <- year_frequency[year_frequency$yearly_totals < 2020, ]
post_covid <- year_frequency[year_frequency$yearly_totals > 2019, ]
#Visualize the pre-covid trend
ggplot(pre_covid, aes(x = yearly_totals, y = Freq)) + labs(x="Year", y="Reported Incidents") + geom_poi
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
#Visualize the post-covid data
ggplot(post_covid, aes(x = yearly_totals, y = Freq)) + labs(x="Year", y="Reported Incidents") + geom_po
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Analysis of pre-covid trend
pre_year_model <- lm(pre_covid$Freq ~ pre_covid$yearly_totals, data = pre_covid)
summary(pre_year_model)
```

```
##
## Call:
## lm(formula = pre_covid$Freq ~ pre_covid$yearly_totals, data = pre_covid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.738  -86.265   -7.931  102.688  255.708
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              186416.846  17753.540   10.50 2.11e-07 ***
## pre_covid$yearly_totals     -91.862      8.822  -10.41 2.31e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.1 on 12 degrees of freedom
## Multiple R-squared:  0.9004, Adjusted R-squared:  0.8921
## F-statistic: 108.4 on 1 and 12 DF,  p-value: 2.308e-07
```

```
#Analysis of post-covid trend
post_year_model <- lm(post_covid$Freq ~ post_covid$yearly_totals, data = post_covid)
summary(post_year_model)
```

```
##
## Call:
## lm(formula = post_covid$Freq ~ post_covid$yearly_totals, data = post_covid)
##
## Residuals:
##     15     16     17     18     19
## -132.0  160.3   94.6 -142.1   19.2
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               465266.0    99490.0   4.677   0.0185 *
## post_covid$yearly_totals   -229.3       49.2  -4.660   0.0186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 155.6 on 3 degrees of freedom
## Multiple R-squared:  0.8786, Adjusted R-squared:  0.8382
## F-statistic: 21.72 on 1 and 3 DF,  p-value: 0.01865
```

Step 3 Final write ups

Trend in shooting incidents

For the first explored issues involving determining the trend for shooting incidents there is an optimistic outlook that shooting incidents are decreasing. Based on the data from either pre-covid, post-covid or the general trend. It is clear that the number of shooting incidents is trending downwards.

Important: This is an observation of the trend, not an exploration of the cause for the spike

It is referred to as a covid spike due to 2020 being the year the spike occurs and 2020 being the year know for the covid outbreak When a thought experiment was done to look at the trend before the 2020 spike (years 2006 through 2019) the decrease trend is even more definitive. The downward trend was -91.862 or a decrease of almost 92 incidents per year with a p-value when rounded to 4 digits being 0, leaving no room for doubt that there was a decrease. Starting from the 2020 data there is a suggestion of a decrease, but it is impossible to reject the null as the p-value for that data is 0.106 or greater than the 0.05 significance needed. An issue with that data is the small sample size (4 years only) and it would be worth exploring when the sample size increases.

Perpetrator and Victims

While the sample ratio of underage perpetrator and victim are slightly less than 1% apart (0.1127 compared to 0.1037). Due to the large sample size there is enough evidence to suggest that minors are more likely to be perpetrators than victims of gun violence. With a p-value of 0.0021 the null is rejected.

Step 4 Biases

While the data might app0ear as raw data, there is a lot of room for potential biases to come in. For example one thing that might stick out is that the number of victims and perpetrators have a nearly 10,000

gap between them. Depending on the situations and who the victims are, but mainly due to socio-economic reasons, perpetrators can sometimes not be found. Additionally if this data was used to train an AI on predicting potential criminals, the information known about the perpetrators is their age, race and sex, all things out of a person's control, and not other contributing factors including but not limited to economic status, mental health issues as well as previous history. Additionally, another way biases can be used is when presenting conclusions regarding this data, it is important to be careful and not present the data in ways that can be seen as potentially harmful without investigation root causes behind this data.