# NYPD Shooting Incident Report

2022-10-04

This is an data analysis report based on shooting incident that happens in New York between 2006 and 2021, the data set is provided by NYPD. This report is focus on the incident happening time for individuals to be aware of the most risky time period.

## Step 0 Loading the data

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

## Step 1 Tidying and Transforming data

This step is for eliminating those columns which are not required for my analysis and adding columns for weekdays for analysis purpose. Also, the age value which is not reasonable is now replaced as "NA"

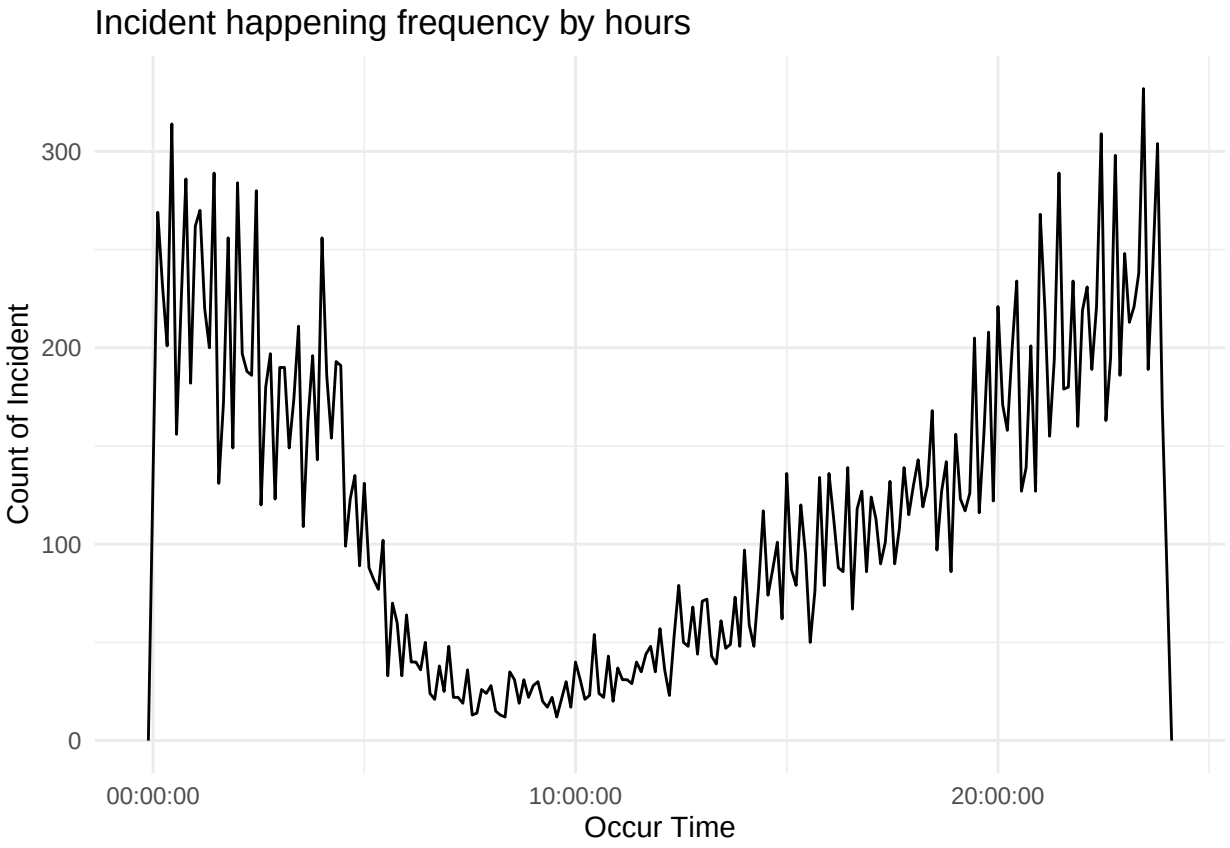
```
NYPD_tidy <- NYPD %>% select(-c(X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat,INCIDENT_KEY,JURISDICTI
  mutate(WEEKDAY = wday( mdy(OCCUR_DATE), week_start = 1)) %>%
  select(OCCUR_DATE,WEEKDAY,OCCUR_TIME,everything()) %>%
  mutate(PERP_AGE_GROUP =replace(PERP_AGE_GROUP,PERP_AGE_GROUP == "1020","NA")) %>%
  mutate(PERP_AGE_GROUP =replace(PERP_AGE_GROUP,PERP_AGE_GROUP == "940","NA")) %>%
  mutate(PERP_AGE_GROUP =replace(PERP_AGE_GROUP,PERP_AGE_GROUP == "224","NA")) %>%
  mutate(PERP_AGE_GROUP =replace(PERP_AGE_GROUP,PERP_AGE_GROUP == "UNKNOWN","NA"))
head(NYPD_tidy)
```

```
## # A tibble: 6 x 12
##   OCCUR_~1 WEEKDAY OCCUR~2 BORO  LOCAT~3 STATI~4 PERP_~5 PERP_~6 PERP_~7 VIC_A~8
##   <chr>      <dbl> <time>  <chr> <chr>   <lgl>   <chr>   <chr>   <chr>   <chr>
## 1 11/11/2~       4 15:04   BROO~ <NA>    FALSE   <NA>    <NA>    <NA>    18-24
## 2 07/16/2~       5 22:05   BROO~ <NA>    FALSE   45-64   M       ASIAN ~ 25-44
## 3 07/11/2~       7 01:09   BROO~ <NA>    FALSE   <18     M       BLACK   25-44
## 4 12/11/2~       6 13:42   BROO~ <NA>    FALSE   <NA>    <NA>    <NA>    25-44
## 5 02/16/2~       2 20:00   QUEE~ <NA>    FALSE   <NA>    <NA>    <NA>    25-44
## 6 05/15/2~       6 04:13   QUEE~ <NA>    TRUE    <NA>    <NA>    <NA>    25-44
## # ... with 2 more variables: VIC_SEX <chr>, VIC_RACE <chr>, and abbreviated
## #   variable names 1: OCCUR_DATE, 2: OCCUR_TIME, 3: LOCATION_DESC,
## #   4: STATISTICAL_MURDER_FLAG, 5: PERP_AGE_GROUP, 6: PERP_SEX, 7: PERP_RACE,
## #   8: VIC_AGE_GROUP
## # i Use `colnames()` to see all variable names
```

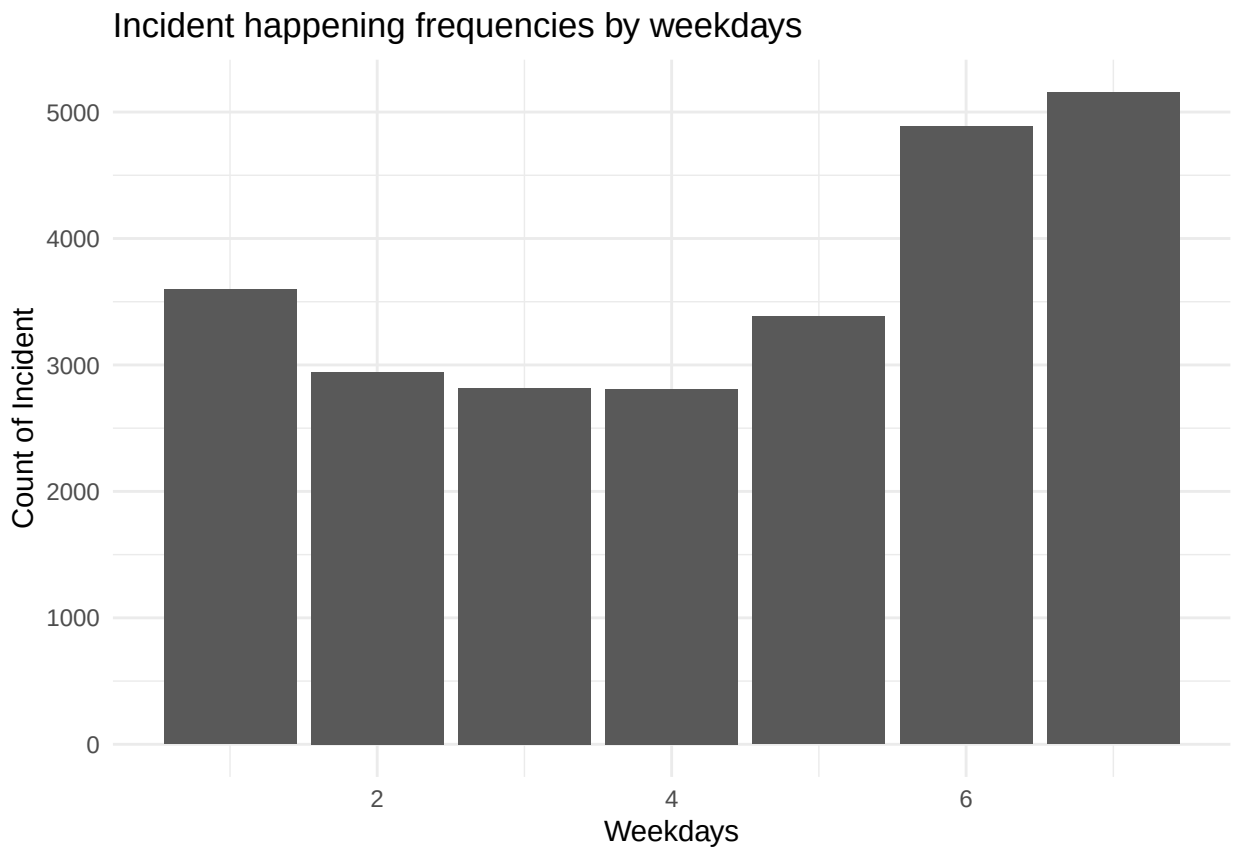
## Step 2 Data Visualization and analysis

### 1. Incident happening time analysis.

```
NYPD_tidy %>% ggplot(aes(x=OCCUR_TIME))+geom_freqpoly(binwidth = 400)+
  labs(title = "Incident happening frequency by hours", x = "Occur Time", y = "Count of Incident")+
  theme_minimal()
```



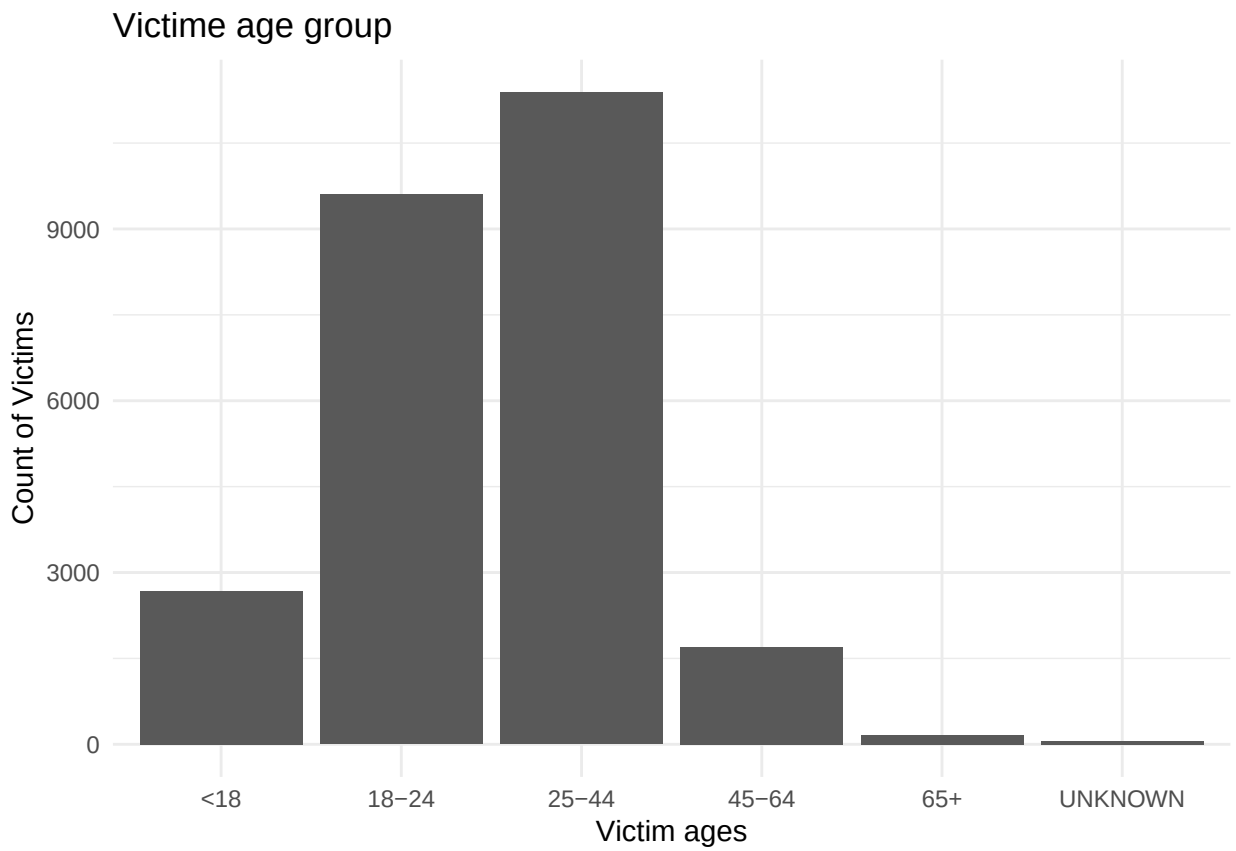

```
NYPD_tidy %>% ggplot(aes(x=WEEKDAY),margin (t = 0, r = 0, b = 0, l = 0, unit = "pt"))+
  geom_bar()+
  labs(title = "Incident happening frequencies by weekdays" , x = "Weekdays" , y = "Count of Incident")+
  theme_minimal()
```



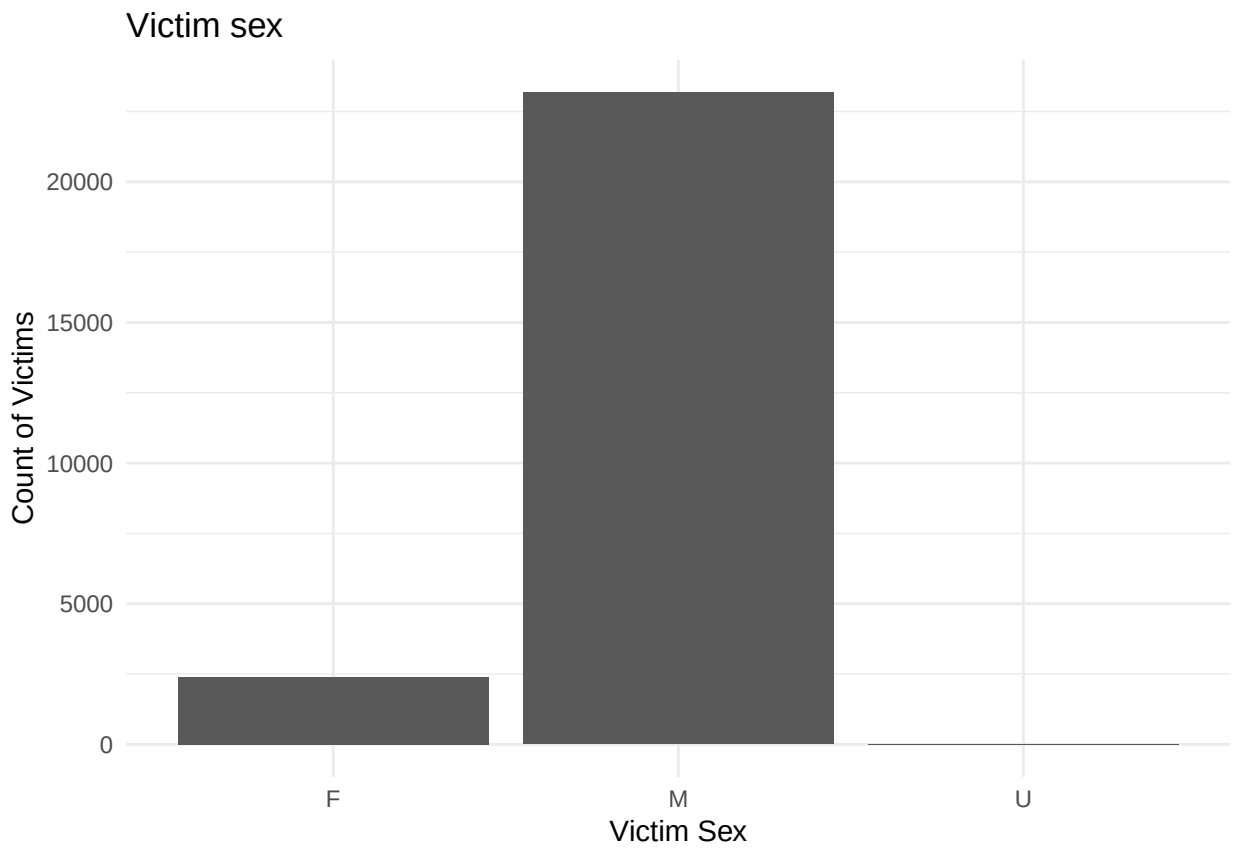

It can be observed that, the incidents peak hour is between 23:00 to 24:00 in a day time and lowest at 7:00 - 8:00. Also, weekends have higher incidents rate compared to weekdays.

### 2. Victim Analysis

```
NYPD %>% ggplot(aes(x=VIC_AGE_GROUP))+
  geom_bar()+
  theme_minimal()+
  labs(title="Victime age group", x = "Victim ages", y = "Count of Victims")
```



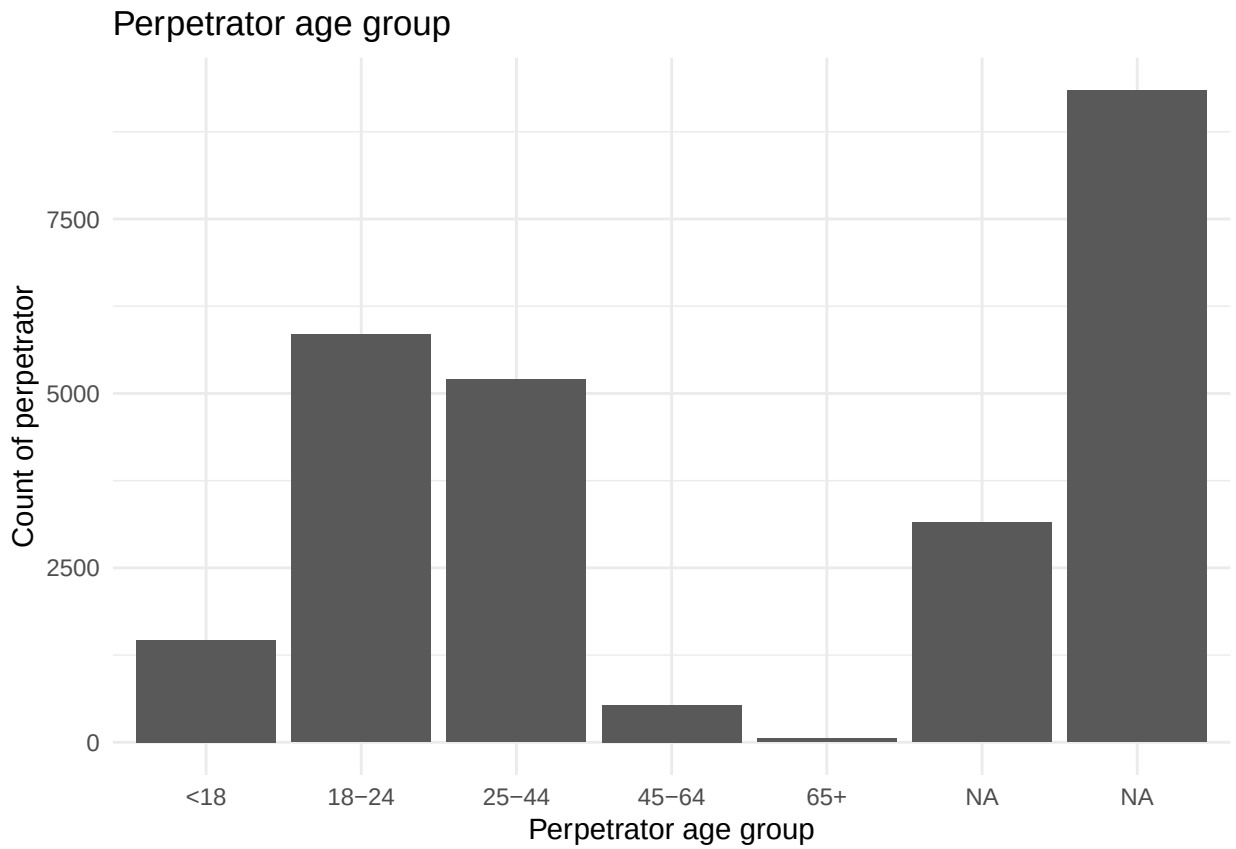

```
NYPD_tidy %>% ggplot(aes(x=VIC_SEX))+
  geom_bar()+
  theme_minimal()+
  labs(title = "Victim sex", x = "Victim Sex" , y = "Count of Victims" )
```




The Victims are mainly males in age group 25-44, followed by age group 18-24.

### 3. Perpetrator analysis

```
NYPD_tidy %>% ggplot(aes(x=PERP_AGE_GROUP))+
  geom_bar()+
  theme_minimal()+
  labs(title = "Perpetrator age group", x = "Perpetrator age group", y = "Count of perpetrator")
```




Majority of the perpetrators are males. But unlike the victims, the top age group is now ages 18-24, followed by ages 25-44

### Step 3 Modelling Data

We are using logistic regression to get a qualitative response (whether it is a murder or not). Based on the p-value in the summary below, the perpetrator and victim ages groups are statistical significant.

```
glm<- glm(STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + WEEKDAY + OCCUR_TIME + VIC_
summary(glm)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     PERP_AGE_GROUP + WEEKDAY + OCCUR_TIME + VIC_RACE + VIC_SEX +
##     VIC_AGE_GROUP, family = binomial, data = NYPD_tidy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5080  -0.7455  -0.6454  -0.1936   3.1198
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -2.342e+01  2.544e+02  -0.092 0.926655
## PERP_RACEASIAN / PACIFIC ISLANDER  1.190e+01  2.295e+02   0.052 0.958660
```

```
## PERP_RACEBLACK                     1.153e+01  2.295e+02   0.050 0.959939
## PERP_RACEBLACK HISPANIC            1.140e+01  2.295e+02   0.050 0.960388
## PERP_RACEUNKNOWN                   1.088e+01  2.295e+02   0.047 0.962186
## PERP_RACEWHITE                     1.207e+01  2.295e+02   0.053 0.958079
## PERP_RACEWHITE HISPANIC            1.165e+01  2.295e+02   0.051 0.959538
## PERP_SEXM                         -1.606e-01  1.212e-01  -1.324 0.185360
## PERP_SEXU                          1.555e+00  2.891e-01   5.379 7.49e-08 ***
## PERP_AGE_GROUP18-24                1.125e-01  7.632e-02   1.474 0.140439
## PERP_AGE_GROUP25-44                3.957e-01  7.772e-02   5.092 3.55e-07 ***
## PERP_AGE_GROUP45-64                6.815e-01  1.181e-01   5.772 7.82e-09 ***
## PERP_AGE_GROUP65+                  7.717e-01  2.896e-01   2.665 0.007708 **
## PERP_AGE_GROUPNA                  -2.468e+00  1.811e-01 -13.630  < 2e-16 ***
## WEEKDAY                           -2.449e-03  1.003e-02  -0.244 0.807087
## OCCUR_TIME                        -2.675e-07  7.041e-07  -0.380 0.704030
## VIC_RACEASIAN / PACIFIC ISLANDER   1.061e+01  1.098e+02   0.097 0.923032
## VIC_RACEBLACK                      1.039e+01  1.098e+02   0.095 0.924608
## VIC_RACEBLACK HISPANIC             1.019e+01  1.098e+02   0.093 0.926010
## VIC_RACEUNKNOWN                    9.720e+00  1.098e+02   0.089 0.929437
## VIC_RACEWHITE                      1.047e+01  1.098e+02   0.095 0.923996
## VIC_RACEWHITE HISPANIC             1.050e+01  1.098e+02   0.096 0.923791
## VIC_SEXM                          -5.285e-02  6.447e-02  -0.820 0.412351
## VIC_SEXU                          -1.388e-01  1.133e+00  -0.123 0.902485
## VIC_AGE_GROUP18-24                 2.358e-01  7.768e-02   3.036 0.002401 **
## VIC_AGE_GROUP25-44                 3.637e-01  7.720e-02   4.712 2.46e-06 ***
## VIC_AGE_GROUP45-64                 3.743e-01  1.022e-01   3.662 0.000251 ***
## VIC_AGE_GROUP65+                   7.391e-01  2.162e-01   3.418 0.000630 ***
## VIC_AGE_GROUPUNKNOWN               2.304e-01  3.550e-01   0.649 0.516299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16186  on 16251  degrees of freedom
## Residual deviance: 15037  on 16223  degrees of freedom
##   (9344 observations deleted due to missingness)
## AIC: 15095
##
## Number of Fisher Scoring iterations: 11
```

## Step 4 Bias

For perpetrator analysis section, since there are a lot of open cases, so the data set is not complete, the conclusion we draw from the data may not be accurate. (eg. A particular age group may be more likely to escape from the police )