NYPD Shooting Incident Data Report

2025-07-15

Goal of project

The goal of this project is to determine if there are any factors within the NYPD shooting incident dataset that would allow us to predict whether an individual is murdered or not. We will look to discover interesting trends on how shootings change over years and by different groups such as location.

Load the data

This block of code is to load the city of New York shooting incident dataset into local memory. The New York shooting incident dataset is a dataset that contains information about shooting incidents in the New York area including Data, location, victim and perpetrator characteristics, and if they were murdered or not.

```
url = 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'
data = read.csv(url)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:29744       Length:29744       Length:29744
##  1st Qu.: 67321140   Class :character   Class :character   Class :character
##  Median :109291972   Mode  :character   Mode  :character   Mode  :character
##  Mean   :133850951
##  3rd Qu.:214741917
##  Max.   :299462478
##
##  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          X_COORD_CD         Y_COORD_CD           Latitude
##  Length:29744       Length:29744       Length:29744       Min.   :40.51
##  Class :character   Class :character   Class :character   1st Qu.:40.67
##  Mode  :character   Mode  :character   Mode  :character   Median :40.70
##                                                           Mean   :40.74
##                                                           3rd Qu.:40.83
##                                                           Max.   :40.91
##                                                           NA's   :97
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:29744
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :97
```

Clean The Data

This block of code will clean the data by converting character types into factors where applicable and removing columns with duplicated values such as lon_lat, x_coord, and y_coord because they are already defined in the lat and lon columns.

```
data$OCCUR_DATE = as.Date(data$OCCUR_DATE)
data$OCCUR_TIME = as.character(data$OCCUR_TIME)
data$BORO = as.factor(data$BORO)
data$LOC_OF_OCCUR_DESC = as.factor(data$LOC_OF_OCCUR_DESC)
data$LOC_CLASSFCTN_DESC = as.factor(data$LOC_CLASSFCTN_DESC)
data$LOCATION_DESC = as.factor(data$LOCATION_DESC)
data$STATISTICAL_MURDER_FLAG = as.factor(data$STATISTICAL_MURDER_FLAG)
data$PERP_AGE_GROUP = as.factor(data$PERP_AGE_GROUP)
data$PERP_SEX = as.factor(data$PERP_SEX)
data$PERP_RACE = as.factor(data$PERP_RACE)
data$VIC_AGE_GROUP = as.factor(data$VIC_AGE_GROUP)
data$VIC_SEX = as.factor(data$VIC_SEX)
data$VIC_RACE = as.factor(data$VIC_RACE)
data = subset(data, select = -c(Lon_Lat, X_COORD_CD, Y_COORD_CD)) # we will drop these columns because
```

Summary Check

This block of code will check that we have correctly converted all data types to the necessary types as well as the number of values in each including null values.

```
summary(data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :0001-01-20   Length:29744
```

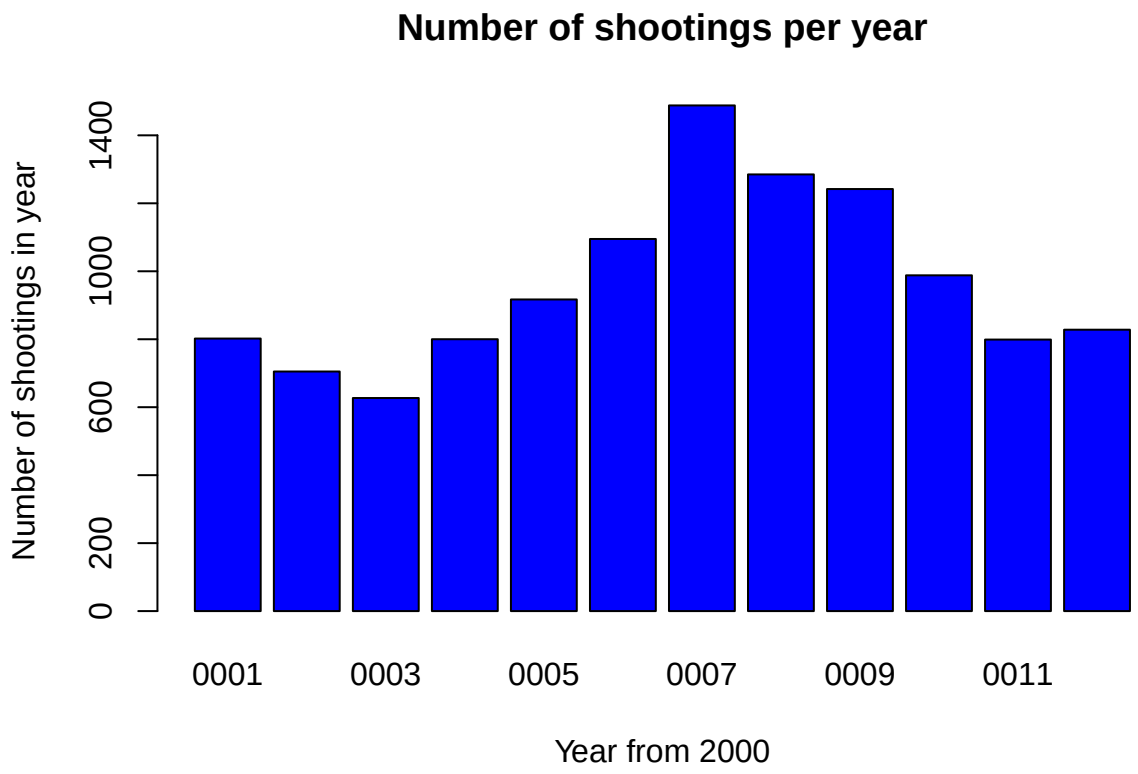
```
##  1st Qu.: 67321140   1st Qu.:0004-12-20   Class :character
##  Median :109291972   Median :0007-07-20   Mode  :character
##  Mean   :133850951   Mean   :0007-04-28
##  3rd Qu.:214741917   3rd Qu.:0009-09-20
##  Max.   :299462478   Max.   :0012-12-20
##                      NA's   :18168
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT      JURISDICTION_CODE
##  BRONX        : 8834          :25596     Min.   :  1.00   Min.   :0.0000
##  BROOKLYN     :11685   INSIDE :  682     1st Qu.: 44.00   1st Qu.:0.0000
##  MANHATTAN    : 3977   OUTSIDE: 3466     Median : 67.00   Median :0.0000
##  QUEENS       : 4426                     Mean   : 65.23   Mean   :0.3181
##  STATEN ISLAND:  822                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                                          Max.   :123.00   Max.   :2.0000
##                                                           NA's   :2
##   LOC_CLASSFCTN_DESC                   LOCATION_DESC   STATISTICAL_MURDER_FLAG
##            :25596                             :14977   false:23979
##  STREET    : 2639    MULTI DWELL - PUBLIC HOUS: 5188   true : 5765
##  HOUSING   :  643    MULTI DWELL - APT BUILD  : 3042
##  DWELLING  :  341    (null)                   : 2526
##  COMMERCIAL:  276    PVT HOUSE                : 1010
##  OTHER     :   74    GROCERY/BODEGA           :  775
##  (Other)   :  175    (Other)                  : 2226
##  PERP_AGE_GROUP   PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##         :9344          : 9310   BLACK         :12323   <18    : 3081   F: 2891
##  18-24  :6630   (null): 1628                 : 9310   1022   :    1   M:26841
##  25-44  :6342   F     :  461   WHITE HISPANIC: 2667   18-24  :10677   U:   12
##  UNKNOWN:3148   M     :16845   UNKNOWN       : 1838   25-44  :13563
##  <18    :1805   U     : 1500   (null)        : 1628   45-64  : 2118
##  (null) :1628                  BLACK HISPANIC: 1487   65+    :  236
##  (Other): 847                  (Other)       :  491   UNKNOWN:   68
##                             VIC_RACE        Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:   13   Min.   :40.51   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      :  478   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :20999   Median :40.70   Median :-73.91
##  BLACK HISPANIC                : 2930   Mean   :40.74   Mean   :-73.91
##  UNKNOWN                       :   72   3rd Qu.:40.83   3rd Qu.:-73.88
##  WHITE                         :  741   Max.   :40.91   Max.   :-73.70
##  WHITE HISPANIC                : 4511   NA's   :97      NA's   :97
```

```
# Because we are using this data for visualizations, we can treat the null values as their own value. I
```

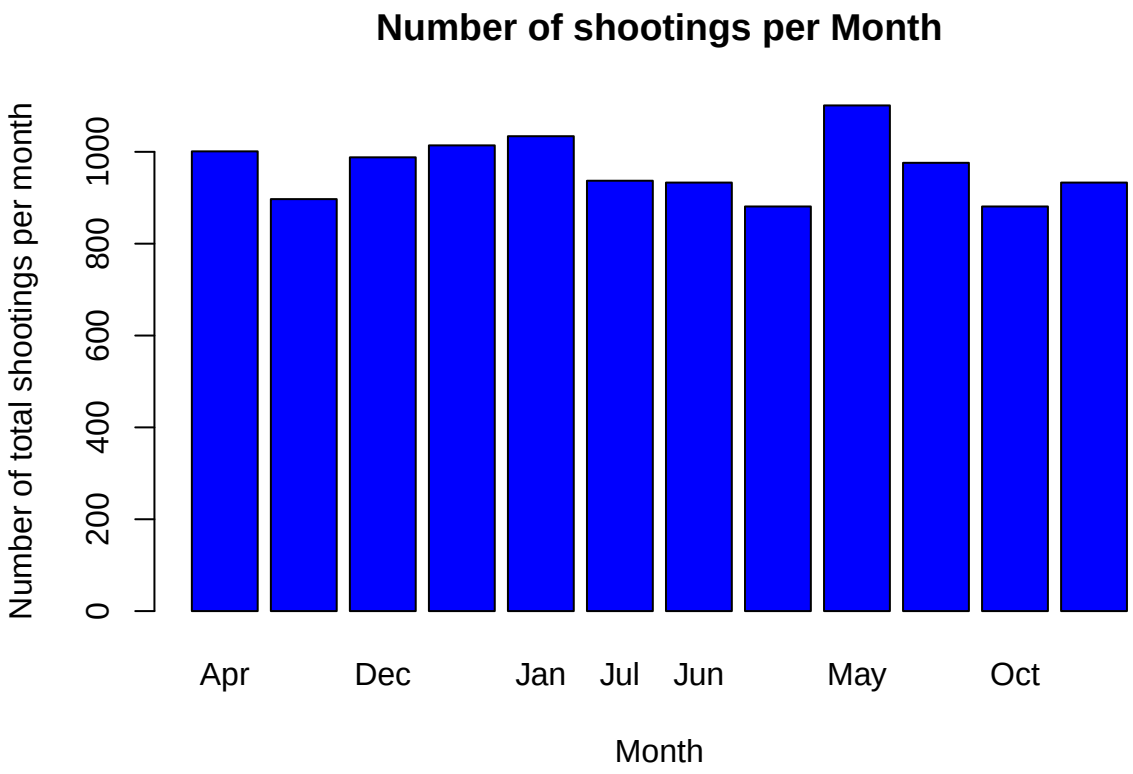
Shooting over Time Visualizations

We will be plotting the number of shooting incidents per year and month to determine if there is any abnormalities or increases throughout different periods of time.

```
barplot(table(format(data$OCCUR_DATE, "%Y")),
        main = "Number of shootings per year",
        xlab = "Year from 2000",
        ylab = "Number of shootings in year",
        col = "blue")
```




```r
barplot(table(format(data$OCCUR_DATE, "%b")),
        main = "Number of shootings per Month",
        xlab = "Month",
        ylab = "Number of total shootings per month",
        col = "blue")
```



Analysis

While the earlier visualizations were great for getting an overall understanding, the main problem that we would like to look at as well is if there are more murders over the years as well as in each BORO.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

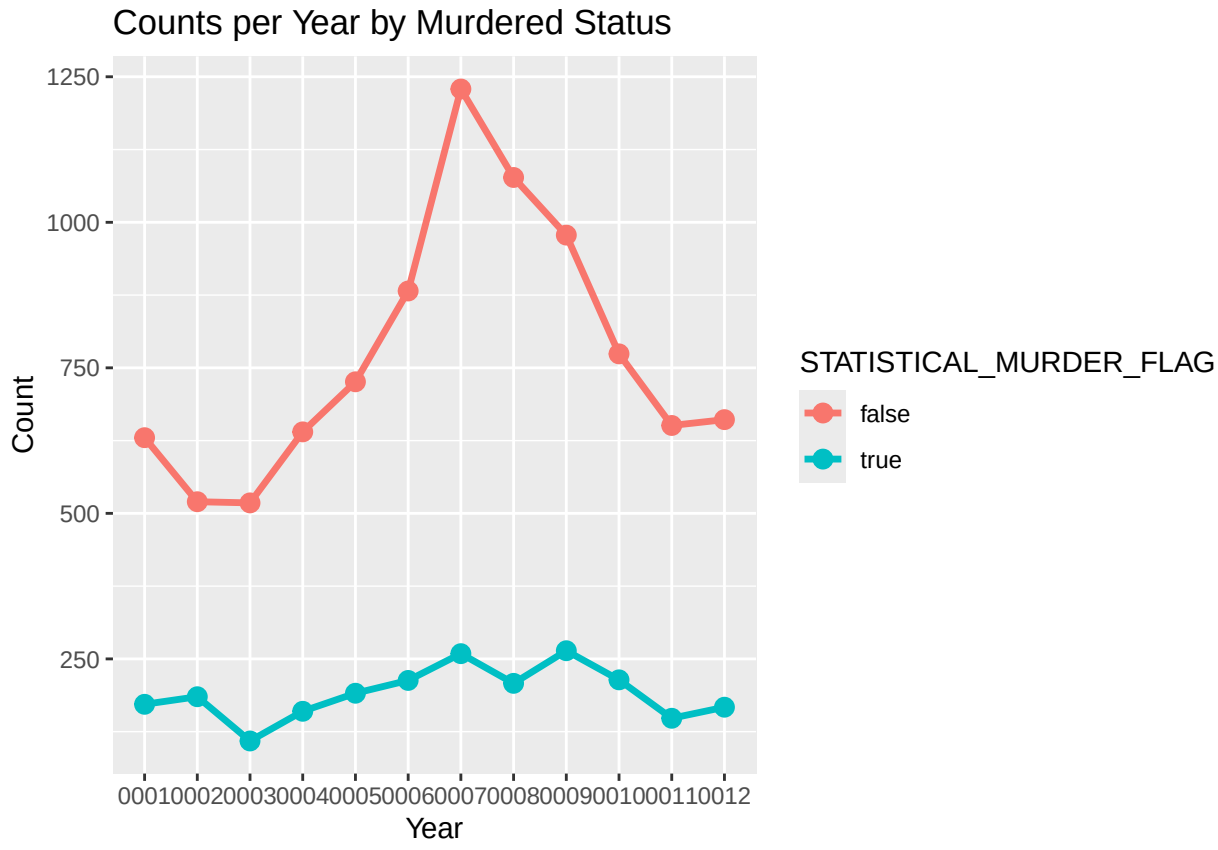
```r
temp = data
temp = temp %>% filter(!is.na(OCCUR_DATE))
temp$Year = format(temp$OCCUR_DATE, "%Y")

temp = temp %>%
  group_by(Year, STATISTICAL_MURDER_FLAG) %>%
  summarise(Count = n())
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```r
ggplot(temp, aes(x = Year, y = Count, color = STATISTICAL_MURDER_FLAG, group = STATISTICAL_MURDER_FLAG))
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  labs(title = "Counts per Year by Murdered Status", x = "Year", y = "Count")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

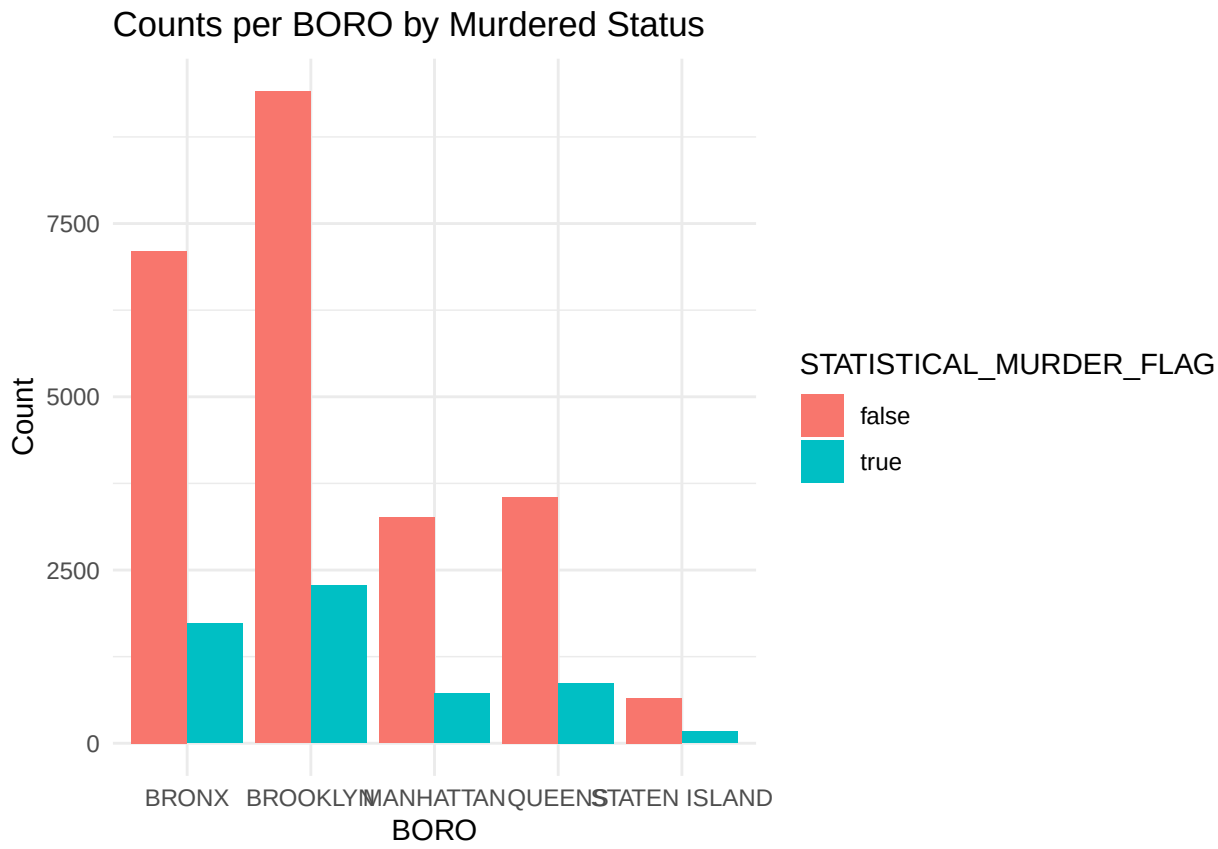


```r
temp = data
temp = temp %>% filter(!is.na(BORO))

temp = temp %>%
  group_by(BORO, STATISTICAL_MURDER_FLAG) %>%
  summarise(Count = n())
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```r
ggplot(temp, aes(x = BORO, y = Count, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Counts per BORO by Murdered Status", x = "BORO", y = "Count") +
  theme_minimal()
```



We can see that Most of the shootings occur in the Bronx and Brooklyn BOROs as well as the most murders.

Model

This block of code will attempt to model the relationship between predictors such as Boro, Victim age group, Victim Sex, and the year and our response variable murder flag.We will be using a logistic regression model because the murder flag predictor is a binary variable.

```r
temp = data
temp = temp %>% filter(!is.na(OCCUR_DATE))
temp$Year = as.numeric(format(temp$OCCUR_DATE, "%Y"))
model = glm(STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP + VIC_SEX + Year, data = temp, na.action=na.
summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP +
##     VIC_SEX + Year, family = binomial(link = "logit"), data = temp,
##     na.action = na.omit)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.764870   0.124045 -14.228  < 2e-16 ***
## BOROBROOKLYN           0.013993   0.056944   0.246  0.80589
## BOROMANHATTAN         -0.070887   0.079700  -0.889  0.37378
## BOROQUEENS             0.087735   0.074885   1.172  0.24136
## BOROSTATEN ISLAND     -0.002349   0.144230  -0.016  0.98701
## VIC_AGE_GROUP18-24     0.235062   0.097253   2.417  0.01565 *
## VIC_AGE_GROUP25-44     0.678995   0.093469   7.264 3.75e-13 ***
## VIC_AGE_GROUP45-64     0.849110   0.117529   7.225 5.02e-13 ***
## VIC_AGE_GROUP65+       1.252727   0.232779   5.382 7.38e-08 ***
## VIC_AGE_GROUPUNKNOWN   1.255045   0.421788   2.976  0.00292 **
## VIC_SEXM              -0.050656   0.077239  -0.656  0.51193
## VIC_SEXU             -11.738275 131.219581  -0.089  0.92872
## Year                  -0.013581   0.007370  -1.843  0.06537 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11515  on 11575  degrees of freedom
## Residual deviance: 11365  on 11563  degrees of freedom
## AIC: 11391
##
## Number of Fisher Scoring iterations: 11
```

Conclusion

We can see that the significant p values are mostly in the victim age group which leads us to believe that this plays the biggest part in understanding if someone is murdered or not.

BIAS

In our model I excluded some predictors that others may use due to reporting bias. I excluded predictors that included perpetrator characteristics to avoid generalized false assumptions, however bias will still be present due to the many null values in our dataset. The values may be null due to individuals not wanting to report full details.