NY PD Shooting Analysis

P.Papa

2024-10-01

Load libraries

```r
install.packages('readr', dependencies = TRUE, repos='http://cran.rstudio.com/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/6d/dwy_4hn968schjl9qrfg0h5h0000gn/T//RtmpRMJsUq/downloaded_packages
```

```r
install.packages('tidyverse', dependencies = TRUE, repos='http://cran.rstudio.com/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/6d/dwy_4hn968schjl9qrfg0h5h0000gn/T//RtmpRMJsUq/downloaded_packages
```

```r
install.packages('lubridate', dependencies = TRUE, repos='http://cran.rstudio.com/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/6d/dwy_4hn968schjl9qrfg0h5h0000gn/T//RtmpRMJsUq/downloaded_packages
```

```r
install.packages('ggplot2', dependencies = TRUE, repos='http://cran.rstudio.com/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/6d/dwy_4hn968schjl9qrfg0h5h0000gn/T//RtmpRMJsUq/downloaded_packages
```

```r
install.packages('dplyr', dependencies = TRUE, repos='http://cran.rstudio.com/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/6d/dwy_4hn968schjl9qrfg0h5h0000gn/T//RtmpRMJsUq/downloaded_packages
```

```r
library(readr)
library(tidyverse)
library(lubridate)
library(ggplot2)
library(dplyr)
library(vcd)
```

Get current data

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

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Tidy and transform data

Remove unnecessary data

```
nyc_data <- nyc_data_orig %>% select(-PRECINCT, -JURISDICTION_CODE,-X_COORD_CD, -Y_COORD_CD, -Latitude,
```

Change data type

- **INCIDENT_KEY** to be treated as a string.
- **BORO** to be treated as a factor.
- **PERP_AGE_GROUP** to be treated as a factor.
- **PERP_SEX** to be treated as a factor.
- **PERP_RACE** to be treated as a factor.
- **VIC_AGE_GROUP** to be treated as a factor.
- **VIC_SEX** to be treated as a factor.
- **VIC_RACE** to be treated as a factor.

```
# Remove anomalies in the data values
nyc_data = subset(nyc_data, VIC_AGE_GROUP!="1022" & PERP_AGE_GROUP!="1020" & PERP_AGE_GROUP!="940" & PE

# Replace 'UNKNOWN' and 'U' with 'Unknown' and standardize missing values to NA
nyc_data$PERP_AGE_GROUP <- recode(nyc_data$PERP_AGE_GROUP, 'UNKNOWN' = 'Unknown')
nyc_data$PERP_SEX <- recode(nyc_data$PERP_SEX, 'U' = 'Unknown')
nyc_data$PERP_RACE <- recode(nyc_data$PERP_RACE, 'UNKNOWN' = 'Unknown')
nyc_data$VIC_SEX <- recode(nyc_data$VIC_SEX, 'U' = 'Unknown')
nyc_data$VIC_RACE <- recode(nyc_data$VIC_RACE, 'UNKNOWN' = 'Unknown')

# Convert variables to appropriate data types
nyc_data$INCIDENT_KEY <- as.character(nyc_data$INCIDENT_KEY)
nyc_data$BORO <- as.factor(nyc_data$BORO)
nyc_data$PERP_AGE_GROUP <- as.factor(nyc_data$PERP_AGE_GROUP)
nyc_data$PERP_SEX <- as.factor(nyc_data$PERP_SEX)
nyc_data$PERP_RACE <- as.factor(nyc_data$PERP_RACE)
```

```r
nyc_data$VIC_AGE_GROUP <- as.factor(nyc_data$VIC_AGE_GROUP)
nyc_data$VIC_SEX <- as.factor(nyc_data$VIC_SEX)
nyc_data$VIC_RACE <- as.factor(nyc_data$VIC_RACE)

# Remove unkwnown
nyc_data[nyc_data == 'Unknown'] <- NA
nyc_data[nyc_data == 'UNKNOWN'] <- NA
nyc_data <- na.omit(nyc_data)

# Drop unused factor levels
nyc_data <- nyc_data %>%
  mutate(across(where(is.factor), droplevels))

# Remove rows with missing values in key variables
nyc_data_clean <- nyc_data %>%
  filter(complete.cases(PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                        VIC_AGE_GROUP, VIC_SEX, VIC_RACE))

# Return summary statistics
summary(nyc_data_clean)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                      BORO
##  Length:1819        Length:1819        Length:1819       BRONX        :635
##  Class :character   Class :character   Class1:hms        BROOKLYN     :523
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    :358
##                                        Mode  :numeric    QUEENS       :257
##                                                          STATEN ISLAND: 46
##
##  LOC_OF_OCCUR_DESC  LOC_CLASSFCTN_DESC LOCATION_DESC
##  Length:1819        Length:1819        Length:1819
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  Mode :logical           <18  :219      F:  73
##  FALSE:1388              18-24:593      M:1746
##  TRUE :431               25-44:835
##                          45-64:164
##                          65+  :  8
##
##                    PERP_RACE    VIC_AGE_GROUP VIC_SEX
##  ASIAN / PACIFIC ISLANDER:  28   <18  :167     F: 234
##  BLACK                   :1232   18-24:469     M:1585
##  BLACK HISPANIC          : 188   25-44:967
##  WHITE                   :  26   45-64:187
##  WHITE HISPANIC          : 345   65+  : 29
##
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   1
##  ASIAN / PACIFIC ISLANDER      :  59
##  BLACK                         :1157
```

```
##  BLACK HISPANIC                 : 195
##  WHITE                          :  48
##  WHITE HISPANIC                 : 359
```

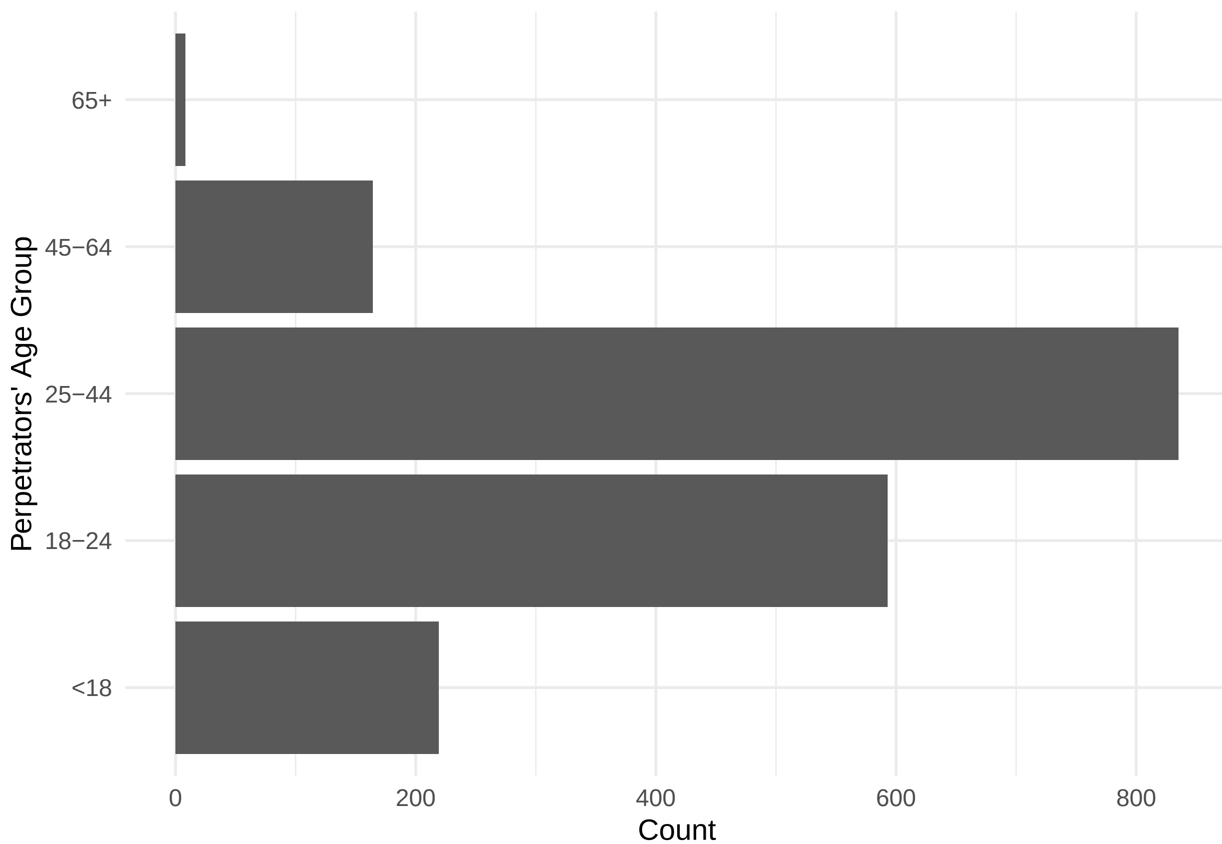
Data visualization

Perpetrators age distribution

```r
summary(nyc_data_clean$PERP_AGE_GROUP)
```

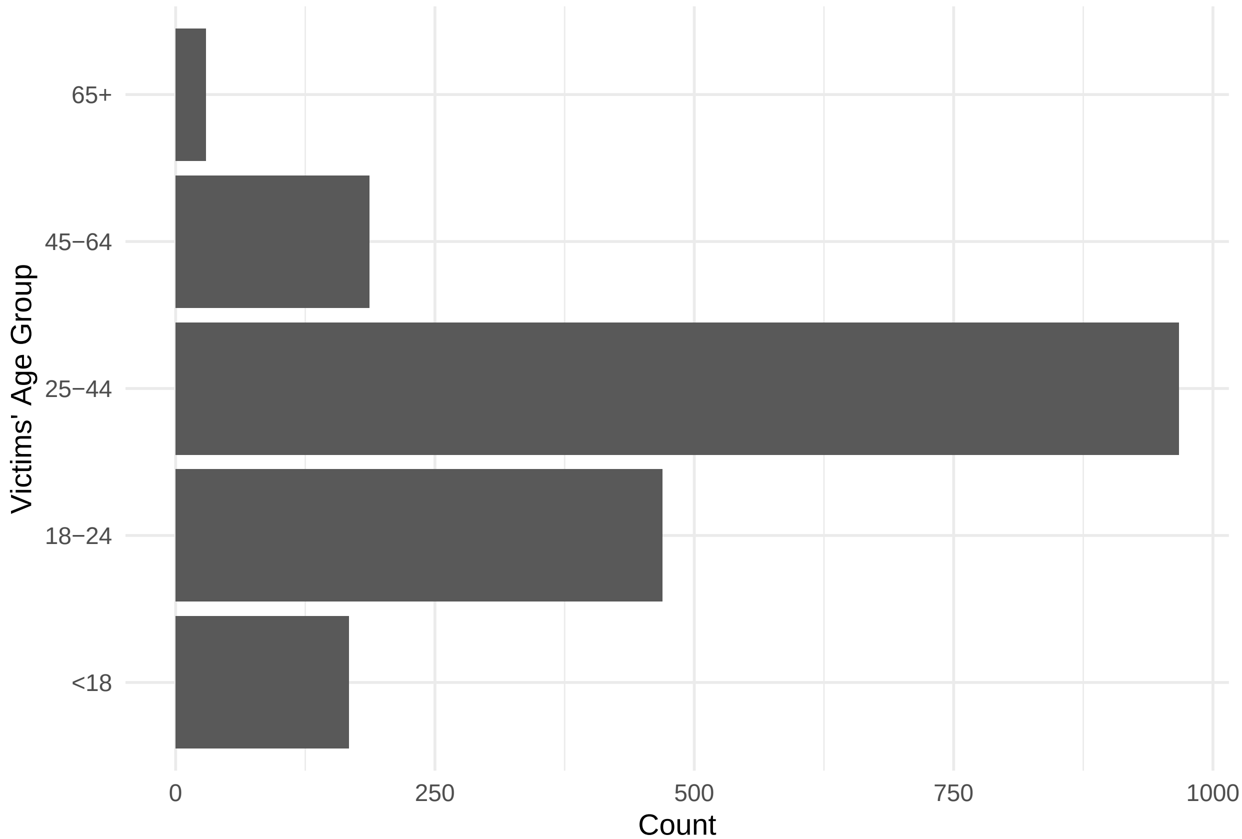
```
##   <18 18-24 25-44 45-64   65+
##   219   593   835   164     8
```

```r
ggplot(nyc_data_clean, aes(x = PERP_AGE_GROUP)) +
  geom_bar() +
  xlab("Perpetrators' Age Group") +
  ylab("Count") +
  theme_minimal() +
  coord_flip()
```



Victims age distribution

```
ggplot(nyc_data_clean, aes(x = VIC_AGE_GROUP)) +
  geom_bar() +
  xlab("Victims' Age Group") +
  ylab("Count") +
  theme_minimal() +
  coord_flip()
```



Sex and Race distribution

```
  # Perpetrators' Sex
  print("Perpetrators' Sex Distribution:")
```

```
## [1] "Perpetrators' Sex Distribution:"
```

```
  table(nyc_data_clean$PERP_SEX)
```

```
##
##    F    M
##   73 1746
```

```r
# Victims' Sex
print("Victims' Sex Distribution:")
```

```
## [1] "Victims' Sex Distribution:"
```

```r
table(nyc_data_clean$VIC_SEX)
```

```
##
##    F    M
##  234 1585
```

```r
# Perpetrators' Race
print("Perpetrators' Race Distribution:")
```

```
## [1] "Perpetrators' Race Distribution:"
```

```r
table(nyc_data_clean$PERP_RACE)
```

```
##
## ASIAN / PACIFIC ISLANDER                    BLACK           BLACK HISPANIC
##                       28                     1232                      188
##                    WHITE           WHITE HISPANIC
##                       26                      345
```

```r
# Victims' Race
print("Victims' Race Distribution:")
```

```
## [1] "Victims' Race Distribution:"
```

```r
table(nyc_data_clean$VIC_RACE)
```

```
##
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                              1                             59
##                          BLACK                 BLACK HISPANIC
##                           1157                            195
##                          WHITE                 WHITE HISPANIC
##                             48                            359
```

Analysis of relationships between variables

```r
# Levels of variable VIC_SEX
levels(nyc_data_clean$VIC_SEX)
```

```
## [1] "F" "M"
```

```r
# Levels of variable PERP_SEX
levels(nyc_data_clean$PERP_SEX)
```

```
## [1] "F" "M"
```

```r
nyc_data_clean$VIC_SEX <- relevel(nyc_data_clean$VIC_SEX, ref = "M")

# Levels of variable VIC_SEX
levels(nyc_data_clean$VIC_SEX)
```

```
## [1] "M" "F"
```

```r
# Levels of variable PERP_SEX
levels(nyc_data_clean$PERP_SEX)
```

```
## [1] "F" "M"
```

Cross-tabulation of Perpetrators' and Victims' Sex

```r
sex_table_p <- table(nyc_data_clean$PERP_SEX, nyc_data_clean$VIC_SEX)
print("Cross-tabulation of Perpetrators' and Victims' Sex:")
```

```
## [1] "Cross-tabulation of Perpetrators' and Victims' Sex:"
```

```r
sex_table_p
```

```
##
##        M    F
##   F   54   19
##   M 1531  215
```

Cross-tabulation of Perpetrators' and Victims' Age Groups

```r
age_table <- table(nyc_data_clean$PERP_AGE_GROUP, nyc_data_clean$VIC_AGE_GROUP)
print("Cross-tabulation of Perpetrators' and Victims' Age Groups:")
```

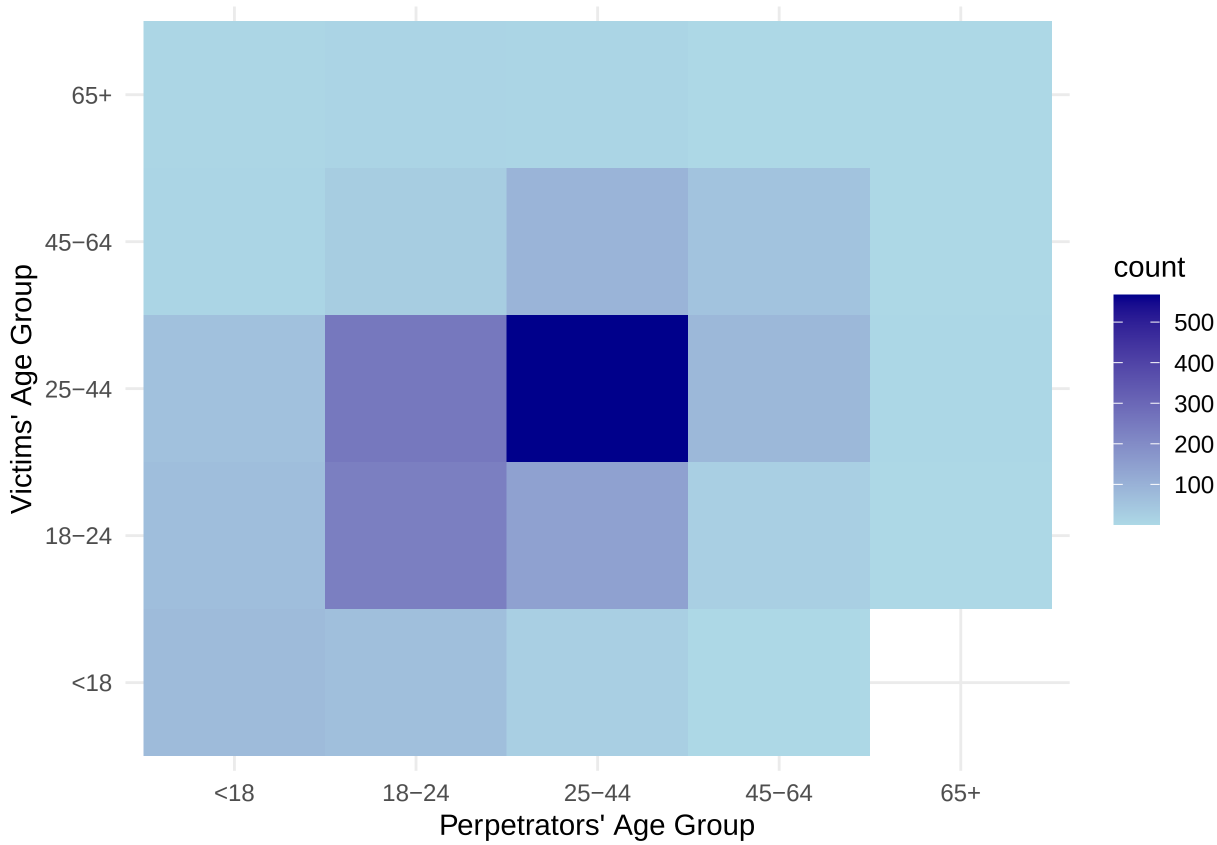
```
## [1] "Cross-tabulation of Perpetrators' and Victims' Age Groups:"
```

```r
age_table
```

```
##
##         <18 18-24 25-44 45-64 65+
##   <18    76    68    60     9   6
##   18-24  66   234   253    30  10
##   25-44  23   143   567    93   9
##   45-64   2    23    83    54   2
##   65+     0     1     4     1   2
```

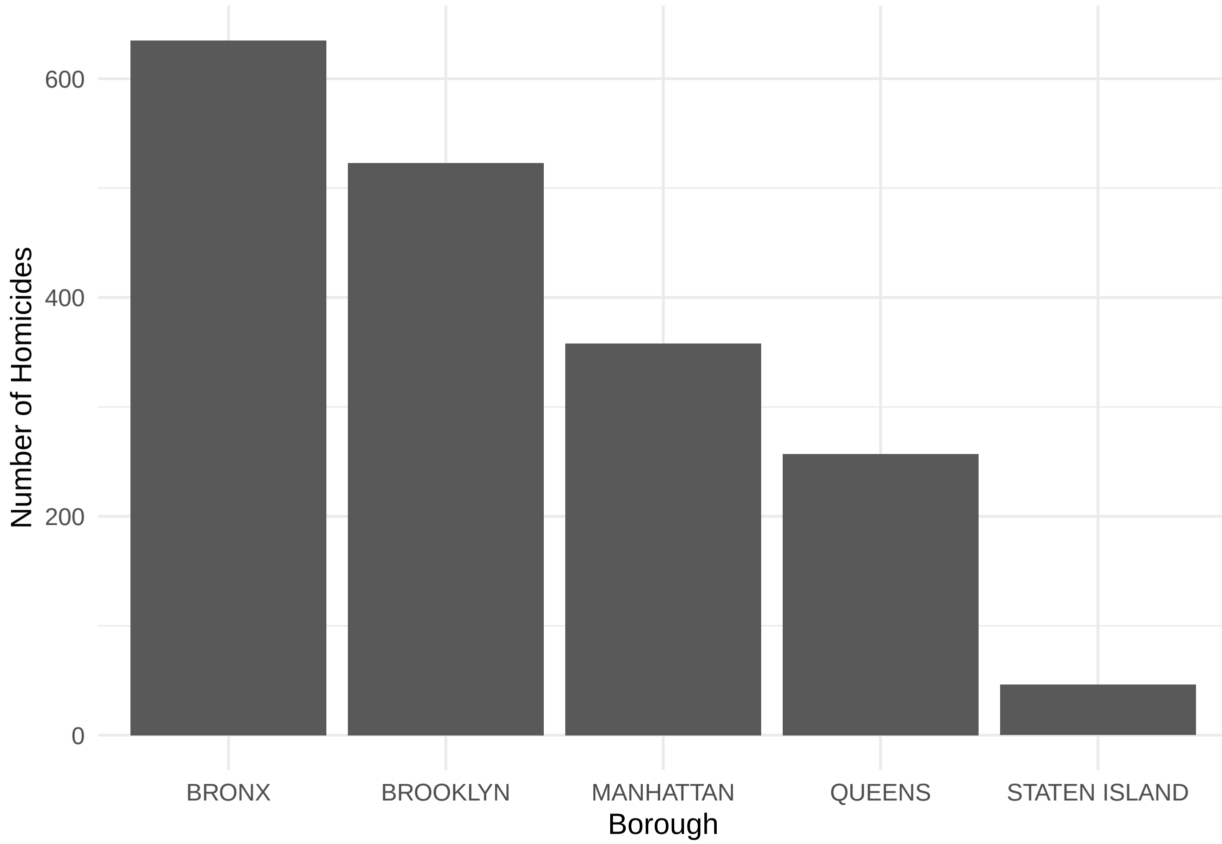
Heatmap of Age Group Interactions

```
ggplot(nyc_data_clean, aes(x = PERP_AGE_GROUP, y = VIC_AGE_GROUP)) +
  geom_bin2d() +
  xlab("Perpetrators' Age Group") +
  ylab("Victims' Age Group") +
  theme_minimal() +
  scale_fill_gradient(low = "lightblue", high = "darkblue")
```



Distribution by Borough

```
ggplot(nyc_data_clean, aes(x = BORO)) +
  geom_bar() +
  xlab("Borough") +
  ylab("Number of Homicides") +
  theme_minimal()
```



Logistic Regression model

I first ensure that VIC_SEX (victim's sex) is correctly formatted as a factor variable suitable for logistic regression. This variable is binary, representing two categories (e.g., "M" for male and "F" for female). Then I use glm function for modeling the probability of the victim being of a certain sex based on perpetrator characteristics.

```r
# Ensure VIC_SEX is a binary factor for logistic regression
nyc_data_clean$VIC_SEX <- factor(nyc_data_clean$VIC_SEX)

# Fit the model
model <- glm(VIC_SEX ~ PERP_SEX + PERP_AGE_GROUP + PERP_RACE,
             data = nyc_data_clean, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = VIC_SEX ~ PERP_SEX + PERP_AGE_GROUP + PERP_RACE,
##     family = "binomial", data = nyc_data_clean)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.3751     1.0862  -2.187  0.02878 *
## PERP_SEXM                  -0.9115     0.2849  -3.199  0.00138 **
## PERP_AGE_GROUP18-24        -0.1163     0.2360  -0.493  0.62208
```

```
## PERP_AGE_GROUP25-44      -0.1739     0.2276  -0.764  0.44503
## PERP_AGE_GROUP45-64       0.7721     0.2752   2.806  0.00502 **
## PERP_AGE_GROUP65+       -13.4407   505.5238  -0.027  0.97879
## PERP_RACEBLACK            1.5588     1.0250   1.521  0.12832
## PERP_RACEBLACK HISPANIC   1.0796     1.0525   1.026  0.30503
## PERP_RACEWHITE            0.6088     1.2644   0.482  0.63016
## PERP_RACEWHITE HISPANIC   0.6281     1.0435   0.602  0.54720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1396.3  on 1818  degrees of freedom
## Residual deviance: 1342.3  on 1809  degrees of freedom
## AIC: 1362.3
##
## Number of Fisher Scoring iterations: 14
```

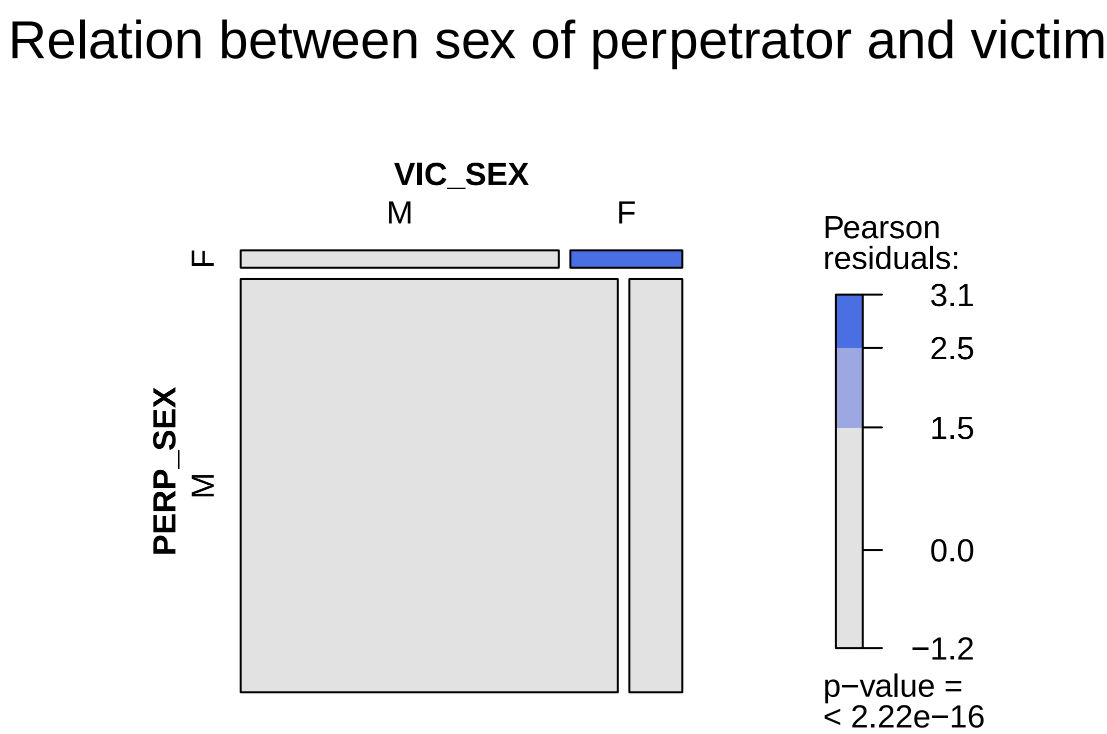
Bias reduction

Since in my country there are many crimes committed by men targeting women, I have eliminated the potential bias and adjusted the model to correctly calculate the relationship between the perpetrator's characteristics and the victim's sex.

An additional investigation

Since my initial hypothesis is not confirmed by the data, I would like to also explore the opposite case to see if there is any inverse relationship.

```
mosaic(~ PERP_SEX + VIC_SEX, data = nyc_data_clean, shade = TRUE,
       legend = TRUE, gp = shading_max,
       main = "Relation between sex of perpetrator and victim",
       xlab = "Sex of Perpetrator",
       ylab = "Sex of Victim")
```




```r
# Set "M" (male) as reference level for VIC_SEX
nyc_data_clean$VIC_SEX <- relevel(nyc_data_clean$VIC_SEX, ref = "M")

# Check that PERP_SEX has "F" (Female) as reference level
nyc_data_clean$PERP_SEX <- relevel(nyc_data_clean$PERP_SEX, ref = "F")

# Regression model
model <- glm(VIC_SEX ~ PERP_SEX, data = nyc_data_clean, family = binomial)

# Visualize the model
summary(model)
```

```
##
## Call:
## glm(formula = VIC_SEX ~ PERP_SEX, family = binomial, data = nyc_data_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0445     0.2667  -3.916    9e-05 ***
## PERP_SEXM    -0.9185     0.2765  -3.322 0.000894 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1396.3  on 1818  degrees of freedom
## Residual deviance: 1386.7  on 1817  degrees of freedom
## AIC: 1390.7
##
## Number of Fisher Scoring iterations: 4
```

Conclusions

Final Summary

- Female Perpetrators: ** More likely to have female victims
- Perpetrators Aged 45-64: ** More likely to have female victims compared to younger perpetrators.
- Perpetrator's Race: ** Did not show a significant effect on the victim's sex in the current model.

The results suggest significant differences in the victim's sex based on the perpetrator's sex and age. This information can be useful for:

- Developing Prevention Programs: Targeted at specific demographic groups.
- Informing Law Enforcement: To better understand crime dynamics and allocate resources.
- Promoting Further Research: Exploring other variables and delving deeper into the underlying causes of these relationships.

Resources

- https://data.cityofnewyork.us/