Week 3 Assignment

2024-02-19

Hello Week 3!

NYPD Shooting Incident Data Report

Step 0: Load Required Packages + Import Data

```
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
# Read data from City of New York website
## Data validation showed that certain string values were blank instead of NA - this is corrected with
data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv", na.strings=c("",NA))
```

Step 1: Exploratory Data Analysis

```
# Dimension
dim(data)
```

```
## [1] 27312    21
```

```r
# Summary Statistics
summary(data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     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:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

The NYPD Shooting Incident dataset contains 21 columns and 27,312 rows. Of the 21 columns:

- 14 columns are character data types
- 7 columns are numeric data types
  - JURISDICTION_CODE, Latitude, and Longitude each contain limited null values

Here is a subset of the data:

```
head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS              <NA>      105
## 2    137471050 06/27/2014   17:40:00    BRONX              <NA>       40
## 3    147998800 11/21/2015   03:56:00   QUEENS              <NA>      108
## 4    146837977 10/09/2015   18:30:00    BRONX              <NA>       44
## 5     58921844 02/19/2009   22:58:00    BRONX              <NA>       47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN              <NA>       81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0               <NA>          <NA>                  false
## 2                 0               <NA>          <NA>                  false
## 3                 0               <NA>          <NA>                   true
## 4                 0               <NA>          <NA>                  false
## 5                 0               <NA>          <NA>                   true
## 6                 0               <NA>          <NA>                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1           <NA>     <NA>      <NA>         18-24       M          BLACK
## 2           <NA>     <NA>      <NA>         18-24       M          BLACK
## 3           <NA>     <NA>      <NA>         25-44       M          WHITE
## 4           <NA>     <NA>      <NA>           <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6           <NA>     <NA>      <NA>         25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                      Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

The following columns indicate that they could be treated as categorical values:
* BORO * PRECINCT * LOC_OF_OCCUR_DESC * JURISDICTION_CODE * STATISTICAL_MURDER_FLAG * PERP_AGE_GROUP * PERP_SEX * PERP_RACE * VIC_AGE_GROUP * VIC_SEX * VIC_RACE

The following prints the frequencies of each potentially categorical value:

```
count(data,BORO)
```

```
##            BORO     n
## 1         BRONX  7937
## 2      BROOKLYN 10933
## 3     MANHATTAN  3572
## 4        QUEENS  4094
## 5 STATEN ISLAND   776
```

```
count(data,PRECINCT)
```

```
##    PRECINCT    n
## 1         1   25
## 2         5   58
## 3         6   28
## 4         7  109
## 5         9  109
## 6        10   73
## 7        13   60
## 8        14   56
## 9        17   10
## 10       18   34
## 11       19   20
## 12       20   40
## 13       22    1
## 14       23  487
## 15       24  105
## 16       25  461
## 17       26  149
## 18       28  343
## 19       30  229
## 20       32  634
## 21       33  225
## 22       34  316
## 23       40  908
## 24       41  494
## 25       42  850
## 26       43  758
## 27       44 1020
## 28       45  182
## 29       46  895
## 30       47  953
## 31       48  787
## 32       49  353
## 33       50  154
## 34       52  583
## 35       60  372
## 36       61  153
## 37       62   70
## 38       63  282
## 39       66   46
## 40       67 1216
## 41       68   32
```

```
## 42       69  466
## 43       70  459
## 44       71  579
## 45       72  109
## 46       73 1452
## 47       75 1557
## 48       76  167
## 49       77  795
## 50       78   62
## 51       79 1012
## 52       81  799
## 53       83  500
## 54       84  124
## 55       88  280
## 56       90  315
## 57       94   86
## 58      100  170
## 59      101  489
## 60      102  210
## 61      103  593
## 62      104  102
## 63      105  479
## 64      106  224
## 65      107  101
## 66      108   67
## 67      109  115
## 68      110  160
## 69      111   11
## 70      112   23
## 71      113  802
## 72      114  369
## 73      115  179
## 74      120  572
## 75      121  112
## 76      122   61
## 77      123   31
```

```
count(data,LOC_OF_OCCUR_DESC)
```

```
##   LOC_OF_OCCUR_DESC     n
## 1            INSIDE   242
## 2           OUTSIDE  1474
## 3              <NA> 25596
```

```
count(data,JURISDICTION_CODE)
```

```
##   JURISDICTION_CODE     n
## 1                 0 22809
## 2                 1    74
## 3                 2  4427
## 4                NA     2
```

```
count(data,STATISTICAL_MURDER_FLAG)
```

```
##   STATISTICAL_MURDER_FLAG     n
## 1                   false 22046
## 2                    true  5266
```

```
count(data,PERP_AGE_GROUP)
```

```
##    PERP_AGE_GROUP    n
## 1          (null)  640
## 2            1020    1
## 3           18-24 6222
## 4             224    1
## 5           25-44 5687
## 6           45-64  617
## 7             65+   60
## 8             940    1
## 9             <18 1591
## 10        UNKNOWN 3148
## 11           <NA> 9344
```

```
count(data,PERP_SEX)
```

```
##   PERP_SEX     n
## 1   (null)   640
## 2        F   424
## 3        M 15439
## 4        U  1499
## 5     <NA>  9310
```

```
count(data,PERP_RACE)
```

```
##                         PERP_RACE     n
## 1                           (null)   640
## 2 AMERICAN INDIAN/ALASKAN NATIVE     2
## 3       ASIAN / PACIFIC ISLANDER   154
## 4                          BLACK 11432
## 5                 BLACK HISPANIC  1314
## 6                        UNKNOWN  1836
## 7                          WHITE   283
## 8                 WHITE HISPANIC  2341
## 9                           <NA>  9310
```

```
count(data,VIC_AGE_GROUP)
```

```
##   VIC_AGE_GROUP     n
## 1          1022     1
## 2         18-24 10086
## 3         25-44 12281
## 4         45-64  1863
## 5           65+   181
## 6           <18  2839
## 7       UNKNOWN    61
```

```r
count(data,VIC_SEX)
```

```
##   VIC_SEX     n
## 1       F  2615
## 2       M 24686
## 3       U    11
```

```r
count(data,VIC_RACE)
```

```
##                          VIC_RACE     n
## 1 AMERICAN INDIAN/ALASKAN NATIVE    10
## 2       ASIAN / PACIFIC ISLANDER   404
## 3                          BLACK 19439
## 4                 BLACK HISPANIC  2646
## 5                        UNKNOWN    66
## 6                          WHITE   698
## 7                 WHITE HISPANIC  4049
```

Upon reviewing frequencies, certain variables contain blank values that should be converted to null. This will be fixed at time of import. Additionally, for modeling, we will also need to convert these categorical values to numeric representations.

Step 2: Data Cleaning

```r
# Convert categorical values to factors
data$BORO <- factor(data$BORO, exclude = NULL)
data$PRECINCT <- factor(data$PRECINCT, exclude = NULL)
data$LOC_OF_OCCUR_DESC <- factor(data$LOC_OF_OCCUR_DESC, exclude = NULL)
data$JURISDICTION_CODE <- factor(data$JURISDICTION_CODE, exclude = NULL)
data$STATISTICAL_MURDER_FLAG <- factor(data$STATISTICAL_MURDER_FLAG, exclude = NULL)
data$PERP_AGE_GROUP <- factor(data$PERP_AGE_GROUP, exclude = NULL)
data$PERP_SEX <- factor(data$PERP_SEX, exclude = NULL)
data$PERP_RACE <- factor(data$PERP_RACE, exclude = NULL)
data$VIC_AGE_GROUP <- factor(data$VIC_AGE_GROUP, exclude = NULL)
data$VIC_SEX <- factor(data$VIC_SEX, exclude = NULL)
data$VIC_RACE <- factor(data$VIC_RACE, exclude = NULL)

num_data <- model.matrix(~.-1,
                    data = data[,c("BORO",
                                   "PRECINCT",
                                   "LOC_OF_OCCUR_DESC",
                                   "JURISDICTION_CODE",
                                   "STATISTICAL_MURDER_FLAG",
                                   "PERP_AGE_GROUP",
                                   "PERP_SEX",
                                   "PERP_RACE",
                                   "VIC_AGE_GROUP",
                                   "VIC_SEX",
                                   "VIC_RACE")],
```

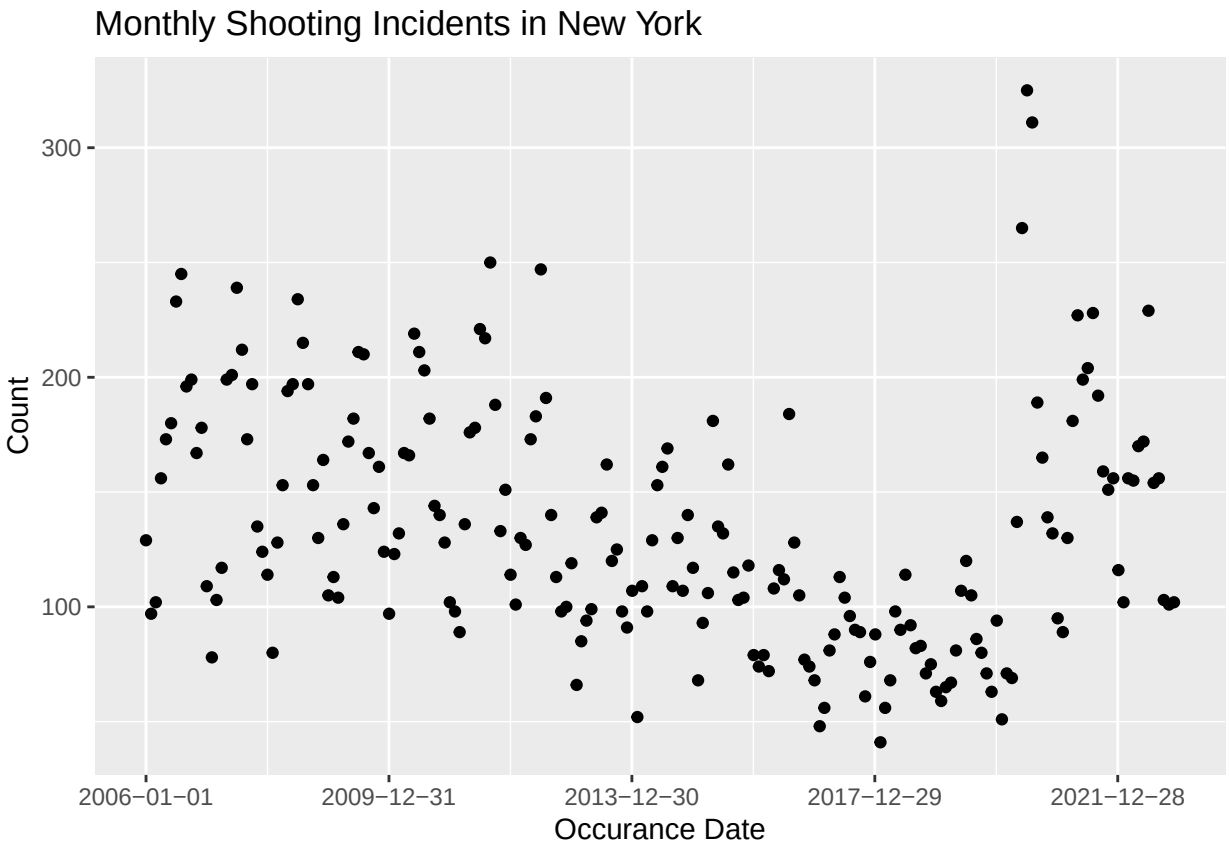
```
                      contrasts.arg = list(
                        BORO = contrasts(data$BORO, contrasts = FALSE),
                       PRECINCT = contrasts(data$PRECINCT, contrasts = FALSE),
                       LOC_OF_OCCUR_DESC = contrasts(data$LOC_OF_OCCUR_DESC, contrasts = FALSE),
                       JURISDICTION_CODE = contrasts(data$JURISDICTION_CODE, contrasts = FALSE),
                         STATISTICAL_MURDER_FLAG = contrasts(data$STATISTICAL_MURDER_FLAG, contrasts = F
                       PERP_AGE_GROUP = contrasts(data$PERP_AGE_GROUP, contrasts = FALSE),
                       PERP_SEX = contrasts(data$PERP_SEX, contrasts = FALSE),
                       PERP_RACE = contrasts(data$PERP_RACE, contrasts = FALSE),
                       VIC_AGE_GROUP = contrasts(data$VIC_AGE_GROUP, contrasts = FALSE),
                       VIC_SEX = contrasts(data$VIC_SEX, contrasts = FALSE),
                       VIC_RACE = contrasts(data$VIC_RACE, contrasts = FALSE)
                                       ))

model_data <- cbind(data,num_data)
```

Convert categorical values to numeric representations Next, as date values can be difficult to incorporate into models, it is important to understand if there are any trends/seasonality with dates that could influence a model.

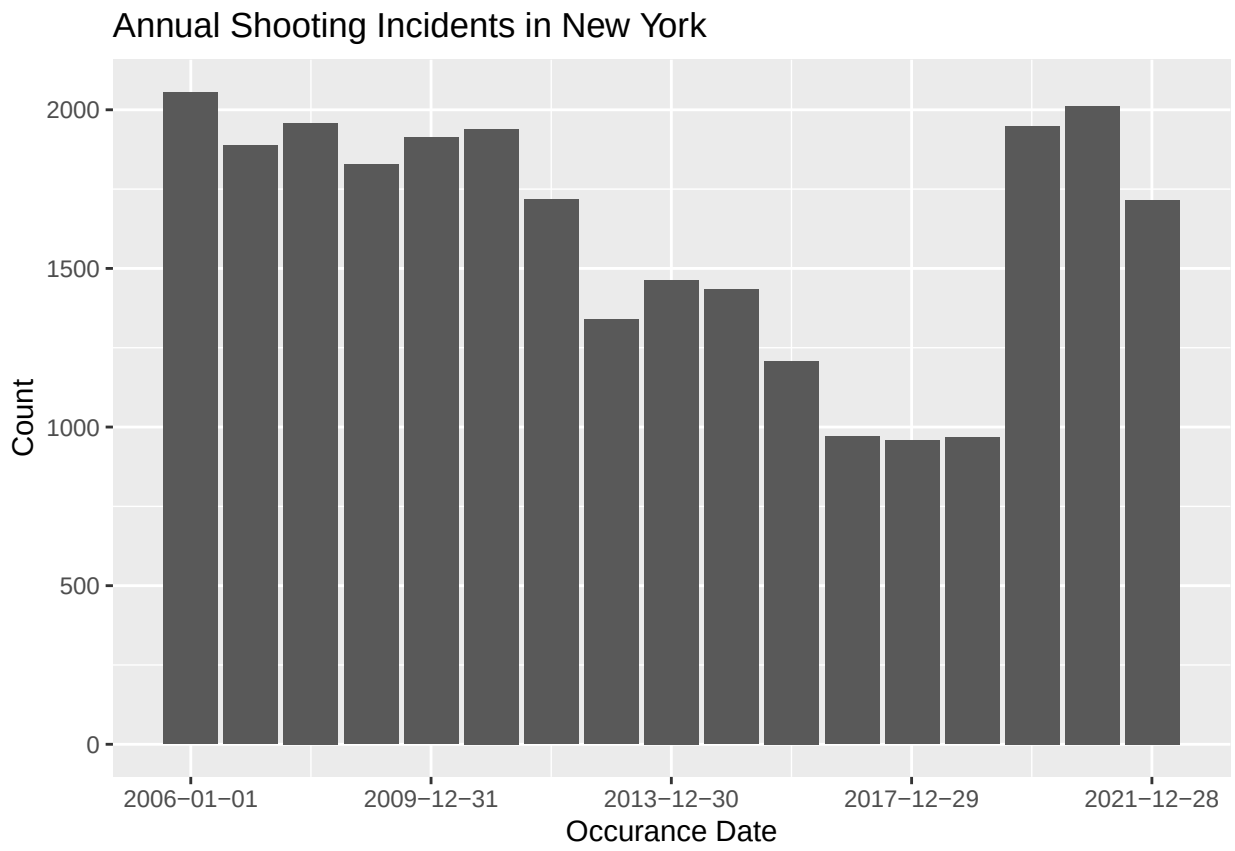
```
model_data$OCCUR_DATE <- as.Date(model_data$OCCUR_DATE,"%m/%d/%Y")

ggplot(model_data, aes(x=floor_date(OCCUR_DATE, "month"))) +
    geom_point(stat="count") +
  scale_x_continuous(breaks = round(seq(min(model_data$OCCUR_DATE), max(model_data$OCCUR_DATE), by = 365
  labs(x="Occurance Date",y="Count")+
  ggtitle(label="Monthly Shooting Incidents in New York")
```



Overall, shooting rates generally declined at a monthly rate between 2006 and 2018, but then increased again. As this data is so still granular at a monthly level, it is important to aggregate it to a slightly higher level to see if any additional trends appear.

```
ggplot(model_data, aes(x=floor_date(OCCUR_DATE, "year"))) +
    geom_bar(stat="count") +
  scale_x_continuous(breaks = round(seq(min(model_data$OCCUR_DATE), max(model_data$OCCUR_DATE), by = 365
  labs(x="Occurance Date",y="Count")+
  ggtitle(label="Annual Shooting Incidents in New York")
```



Viewing this data at an annual level reveals a more consumable trend - shootings in New York were relatively stable between 2006 to 2010, at which point shootings rapidly declined until rapidly increasing in 2019.

Step 3: Modeling

For this assignment, I will be creating a model to evaluate if a shooting perpetrator's age and sex correlates with a shooting victim's sex.

```
# Create new df with variables of interest
model_data_subset <- model_data[,c("PERP_AGE_GROUP(null)"
                                   ,"PERP_AGE_GROUP<18"
                                   ,"PERP_AGE_GROUP1020"
                                   ,"PERP_AGE_GROUP18-24"
                                   ,"PERP_AGE_GROUP224"
                                   ,"PERP_AGE_GROUP25-44"
                                   ,"PERP_AGE_GROUP45-64","PERP_AGE_GROUP65+"
                                   ,"PERP_AGE_GROUP940"
                                   ,"PERP_AGE_GROUPUNKNOWN"
                                   ,"PERP_SEX(null)"
                                   ,"PERP_SEXF"
                                   ,"PERP_SEXM"
                                   ,"PERP_SEXU"
                                   ,"PERP_SEXNA"
                                   ,"VIC_SEXF"
                                   ,"VIC_SEXM"
                                   ,"VIC_SEXU"   )]
```

```r
# Remove null/missing/invalid perp ages
model_data_subset <- subset(model_data_subset,"PERP_AGE_GROUP(null)" != 1 &
                                        PERP_AGE_GROUP1020 != 1 &
                                       PERP_AGE_GROUP224 != 1 &
                                      PERP_AGE_GROUP940 != 1 &
                                       PERP_AGE_GROUPUNKNOWN != 1 )

# Remove null/missing/invalid perp sex
model_data_subset <- subset(model_data_subset,"PERP_SEX(null)" != 1 & PERP_SEXNA != 1)

# Remove null/missing/invalid vic sex
model_data_subset <- subset(model_data_subset,VIC_SEXU != 1)
```

Now that missing and invalid values have been removed, it is possible to compile a linear model.

```r
vicsex <- lm(VIC_SEXF~model_data_subset$"PERP_AGE_GROUP18-24"
             +model_data_subset$"PERP_AGE_GROUP25-44"
             +model_data_subset$"PERP_AGE_GROUP45-64"
             +model_data_subset$"PERP_AGE_GROUP65+"
             +model_data_subset$"PERP_SEXF", data = model_data_subset)
summary(vicsex)
```

```
##
## Call:
## lm(formula = VIC_SEXF ~ model_data_subset$"PERP_AGE_GROUP18-24" +
##     model_data_subset$"PERP_AGE_GROUP25-44" + model_data_subset$"PERP_AGE_GROUP45-64" +
##     model_data_subset$"PERP_AGE_GROUP65+" + model_data_subset$PERP_SEXF,
##     data = model_data_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53242 -0.11373 -0.11373 -0.09932  0.90068
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           0.116049   0.006651  17.449  < 2e-16
## model_data_subset$"PERP_AGE_GROUP18-24" -0.016731   0.007763  -2.155 0.031156
## model_data_subset$"PERP_AGE_GROUP25-44" -0.002323   0.007862  -0.295 0.767637
## model_data_subset$"PERP_AGE_GROUP45-64"  0.065309   0.014364   4.547 5.49e-06
## model_data_subset$"PERP_AGE_GROUP65+"    0.416366   0.041366  10.065  < 2e-16
## model_data_subset$PERP_SEXF              0.055092   0.015944   3.455 0.000551
##
## (Intercept)                           ***
## model_data_subset$"PERP_AGE_GROUP18-24" *
## model_data_subset$"PERP_AGE_GROUP25-44"
## model_data_subset$"PERP_AGE_GROUP45-64" ***
## model_data_subset$"PERP_AGE_GROUP65+"   ***
## model_data_subset$PERP_SEXF             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3163 on 14838 degrees of freedom
```

```
## Multiple R-squared:  0.01055,    Adjusted R-squared:  0.01022
## F-statistic: 31.65 on 5 and 14838 DF,  p-value: < 2.2e-16
```

The output of this model using the NYPD Shooting Incident Data Report data indicates that a shooting perpetrator's age and sex does not effectively predict a shooting victim's sex as indicated by a low adjusted R-Square (0.01). However, both perpetrator age and sex p-values indicate statistical significance, therefore there is a statistical relationship that exists in this model.

Step 4: Bias

There are a few instances in which bias may appear in this dataset. In particular, this dataset only includes reported shootings - if a shooting was not reported, it would not appear in this dataset. This may exclude individuals who are unable to report a shooting for reasons such as fear of additional harm or legal reasons for not being able to report a crime.