

ShootingProject

S.Smiley

2024-08-03

NYC Shooting Dataset Background

This report covers an analysis of the NYC shooting dataset found at “<https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD>”. The dataset lists every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

Clear Statement of the Question of Interest

The question of interest I have from this dataset is: **Does a victim’s age, race, and sex indicate who the perpetrator might be?**

Import Libraries

```
knitr::opts_chunk$set(echo = TRUE)
# Install necessary packages if they are not already installed
if (!requireNamespace("readr", quietly = TRUE)) {
  install.packages("readr")
}
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
if (!requireNamespace("VIM", quietly = TRUE)) {
  install.packages("VIM")
}
if (!requireNamespace("nnet", quietly = TRUE)) {
  install.packages("nnet")
}
if (!requireNamespace("pwr", quietly = TRUE)) {
```

```
install.packages("pwr")
}  
# import libraries  
library(readr)  
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(VIM)
```

```
## Loading required package: colorspace  
  
## Loading required package: grid  
  
## VIM is ready to use.  
  
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues  
  
##  
## Attaching package: 'VIM'  
  
## The following object is masked from 'package:datasets':  
##  
##   sleep
```

```
library(tidyr)  
library(nnet)  
library(pwr)
```

Data Loading

Get Shooting Data

Using the link of where the data comes from is a much more **reproducible** form of loading the data.

```
url_names <- c("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")  
  
shooting_data <- read_csv(url_names[1])
```

```
## 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.
```

```
head(shooting_data)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##   <dbl> <chr>      <time>    <chr>    <chr>              <dbl>
## 1   244608249 05/05/2022 00:10    MANHATTAN INSIDE              14
## 2   247542571 07/04/2022 22:20    BRONX     OUTSIDE             48
## 3    84967535 05/27/2012 19:35    QUEENS    <NA>               103
## 4   202853370 09/24/2019 21:00    BRONX     <NA>               42
## 5    27078636 02/25/2007 21:00    BROOKLYN  <NA>               83
## 6    230311078 07/01/2021 23:07    MANHATTAN <NA>               23
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Data Cleaning

Convert “(null)” strings to NA

```
shooting_data[shooting_data == "(null)"] <- NA
```

Remove unnecessary columns

```
columns_to_keep <- c("OCCUR_DATE", "BORO", "PRECINCT", "PERP_AGE_GROUP", "PERP_SEX", "PERP_RACE", "VIC_AGE_GROUP")
shooting_data <- shooting_data %>%
  select(all_of(columns_to_keep))
head(shooting_data)
```

```
## # A tibble: 6 x 10
##   OCCUR_DATE BORO      PRECINCT PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
##   <chr>      <chr>      <dbl> <chr>      <chr>    <chr>    <chr>
## 1 05/05/2022 MANHATTAN    14 25-44      M        BLACK    25-44
## 2 07/04/2022 BRONX         48 <NA>      <NA>      <NA>    18-24
## 3 05/27/2012 QUEENS       103 <NA>      <NA>      <NA>    18-24
## 4 09/24/2019 BRONX         42 25-44      M        UNKNOWN  25-44
```

```
## 5 02/25/2007 BROOKLYN      83 25-44      M      BLACK      25-44
## 6 07/01/2021 MANHATTAN    23 <NA>      <NA>    <NA>      25-44
## # i 3 more variables: VIC_SEX <chr>, VIC_RACE <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>
```

Summary w/o cleaning

```
summary(shooting_data)
```

```
##   OCCUR_DATE      BORO      PRECINCT  PERP_AGE_GROUP
## Length:28562    Length:28562    Min.   : 1.0    Length:28562
## Class :character Class :character 1st Qu.: 44.0    Class :character
## Mode  :character Mode  :character Median : 67.0    Mode  :character
##                                     Mean  : 65.5
##                                     3rd Qu.: 81.0
##                                     Max.   :123.0
##   PERP_SEX      PERP_RACE  VIC_AGE_GROUP  VIC_SEX
## Length:28562    Length:28562    Length:28562    Length:28562
## Class :character Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character Mode  :character
##
##
##   VIC_RACE      STATISTICAL_MURDER_FLAG
## Length:28562    Mode :logical
## Class :character FALSE:23036
## Mode  :character TRUE :5526
##
##
##
```

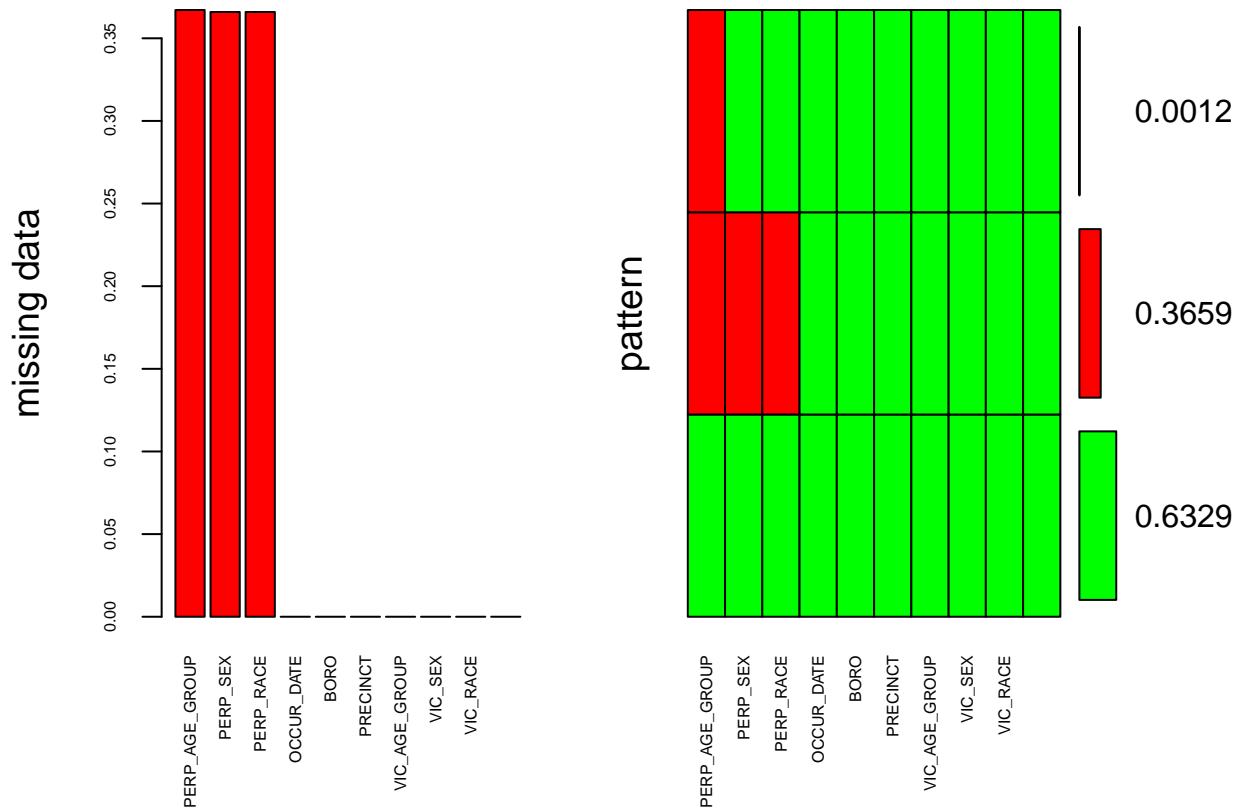
Total up missing data in each column

```
missing_counts <- colSums(is.na(shooting_data))
print(missing_counts)
```

```
##           OCCUR_DATE      BORO      PRECINCT
##              0              0              0
##   PERP_AGE_GROUP      PERP_SEX      PERP_RACE
##      10485      10451      10451
##   VIC_AGE_GROUP      VIC_SEX      VIC_RACE
##              0              0              0
## STATISTICAL_MURDER_FLAG
##              0
```

Visualize missing data

```
library(VIM)
aggr(shooting_data, col=c('green','red'), numbers=TRUE, sortVars=TRUE, labels=names(shooting_data), cex
```



```
##
## Variables sorted by number of missings:
##      Variable      Count
## PERP_AGE_GROUP 0.3670961
## PERP_SEX      0.3659057
## PERP_RACE      0.3659057
## OCCUR_DATE     0.0000000
## BORO           0.0000000
## PRECINCT       0.0000000
## VIC_AGE_GROUP  0.0000000
## VIC_SEX        0.0000000
## VIC_RACE       0.0000000
## STATISTICAL_MURDER_FLAG 0.0000000
```

Create an “UNKONWN” value for the missing data fields

It appears a significant amount of people might have gotten away with murder since over 30% of the missing data is from the perpetrator. Therefore, we don't want to omit this data. Instead, we want to just note that it is “UNKNOWN.” This should help us have less bias in conclusions on shootings and murders since we would have to make some major assumptions otherwise.

```
clean_data <- shooting_data %>%
  mutate(
    PERP_AGE_GROUP = replace_na(PERP_AGE_GROUP, "UNKNOWN"),
    PERP_SEX = replace_na(PERP_SEX, "UNKNOWN"),
    PERP_RACE = replace_na(PERP_RACE, "UNKNOWN"),
    VIC_AGE_GROUP = replace_na(VIC_AGE_GROUP, "UNKNOWN"),
    VIC_SEX = replace_na(VIC_SEX, "UNKNOWN"),
    VIC_RACE = replace_na(VIC_RACE, "UNKNOWN")
  )
```

Total up missing data in each column again

```
missing_counts <- colSums(is.na(clean_data))
print(missing_counts)
```

```
##          OCCUR_DATE          BORO          PRECINCT
##              0              0              0
##    PERP_AGE_GROUP    PERP_SEX    PERP_RACE
##              0              0              0
##    VIC_AGE_GROUP    VIC_SEX    VIC_RACE
##              0              0              0
## STATISTICAL_MURDER_FLAG
##              0
```

Convert date to Date format

```
clean_data <- clean_data %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))
head(clean_data)
```

```
## # A tibble: 6 x 10
##   OCCUR_DATE BORO    PRECINCT PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
##   <date>     <chr>    <dbl> <chr>         <chr>    <chr>    <chr>
## 1 2022-05-05 MANHATTAN    14 25-44         M        BLACK    25-44
## 2 2022-07-04 BRONX        48 UNKNOWN      UNKNOWN   UNKNOWN   18-24
## 3 2012-05-27 QUEENS       103 UNKNOWN      UNKNOWN   UNKNOWN   18-24
## 4 2019-09-24 BRONX        42 25-44         M        UNKNOWN   25-44
## 5 2007-02-25 BROOKLYN    83 25-44         M        BLACK     25-44
## 6 2021-07-01 MANHATTAN    23 UNKNOWN      UNKNOWN   UNKNOWN   25-44
## # i 3 more variables: VIC_SEX <chr>, VIC_RACE <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>
```

Convert character columns to factors

```
clean_data <- clean_data %>%
  mutate(across(where(is.character), as.factor))
head(clean_data)
```

```
## # A tibble: 6 x 10
##   OCCUR_DATE BORO      PRECINCT PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
##   <date>      <fct>      <dbl> <fct>      <fct>    <fct>    <fct>
## 1 2022-05-05 MANHATTAN      14 25-44          M      BLACK    25-44
## 2 2022-07-04 BRONX          48 UNKNOWN        UNKNOWN UNKNOWN  18-24
## 3 2012-05-27 QUEENS        103 UNKNOWN        UNKNOWN UNKNOWN  18-24
## 4 2019-09-24 BRONX          42 25-44          M      UNKNOWN  25-44
## 5 2007-02-25 BROOKLYN      83 25-44          M      BLACK    25-44
## 6 2021-07-01 MANHATTAN      23 UNKNOWN        UNKNOWN UNKNOWN  25-44
## # i 3 more variables: VIC_SEX <fct>, VIC_RACE <fct>,
## #   STATISTICAL_MURDER_FLAG <lgl>
```

Summary w cleaning

```
summary(clean_data)
```

```
##   OCCUR_DATE          BORO      PRECINCT    PERP_AGE_GROUP
##   Min.   :2006-01-01  BRONX      : 8376   Min.   : 1.0   UNKNOWN:13633
##   1st Qu.:2009-09-04  BROOKLYN :11346   1st Qu.: 44.0  18-24 : 6438
##   Median :2013-09-20  MANHATTAN : 3762   Median : 67.0  25-44 : 6041
##   Mean   :2014-06-07  QUEENS    : 4271   Mean   : 65.5  <18   : 1682
##   3rd Qu.:2019-09-29  STATEN ISLAND: 807   3rd Qu.: 81.0  45-64 : 699
##   Max.   :2023-12-29                      Max.   :123.0  65+   : 65
##                                     (Other): 4
##   PERP_SEX          PERP_RACE    VIC_AGE_GROUP
##   F      : 444   AMERICAN INDIAN/ALASKAN NATIVE: 2  <18   : 2954
##   M      :16168  ASIAN / PACIFIC ISLANDER      : 169 1022 : 1
##   U      :1499   BLACK                          :11903 18-24 :10384
##   UNKNOWN:10451  BLACK HISPANIC                : 1392 25-44 :12973
##                                     UNKNOWN      :12288 45-64 : 1981
##                                     WHITE         : 298 65+   : 205
##                                     WHITE HISPANIC : 2510 UNKNOWN: 64
##   VIC_SEX          VIC_RACE    STATISTICAL_MURDER_FLAG
##   F: 2760   AMERICAN INDIAN/ALASKAN NATIVE: 11   Mode :logical
##   M:25790   ASIAN / PACIFIC ISLANDER      : 440  FALSE:23036
##   U: 12     BLACK                          :20235  TRUE :5526
##                                     BLACK HISPANIC : 2795
##                                     UNKNOWN        : 70
##                                     WHITE          : 728
##                                     WHITE HISPANIC : 4283
```

Data Analysis

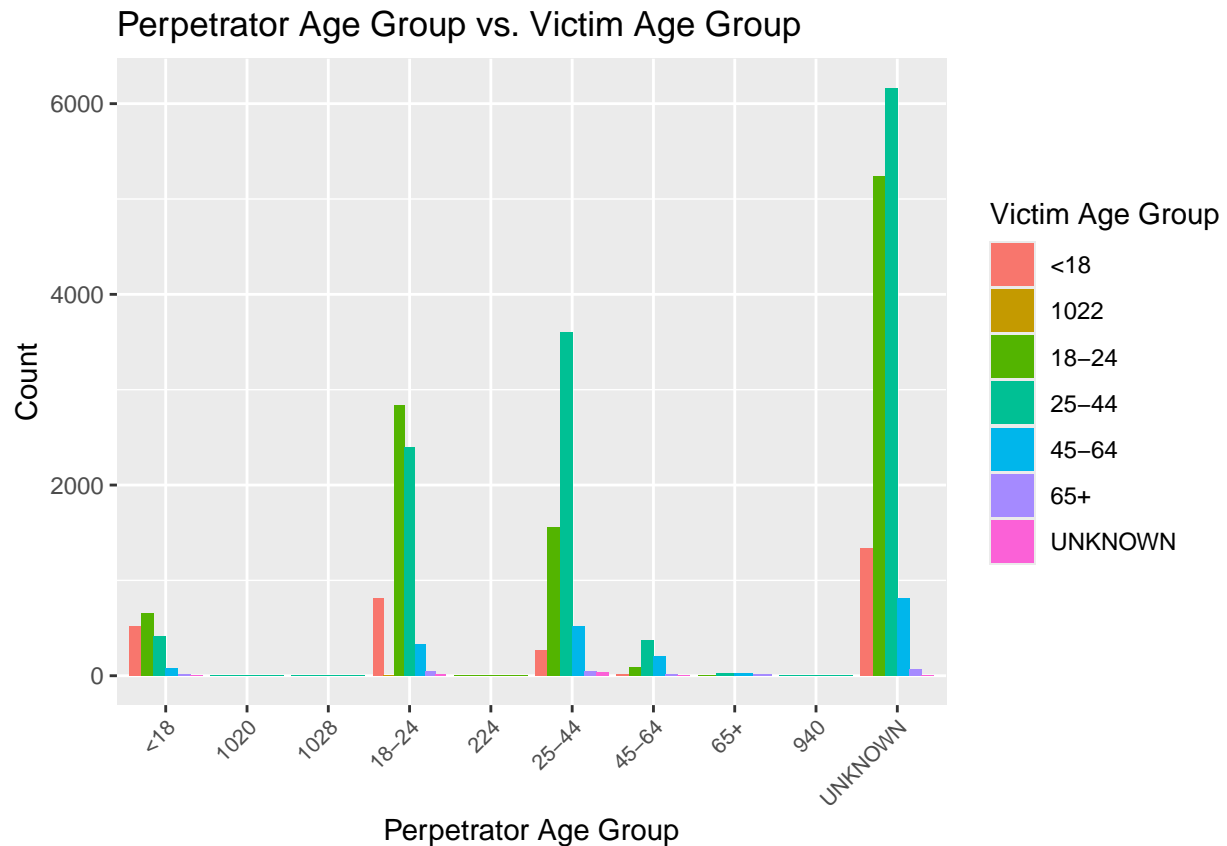
Bar plot of Perpetrator Age Group vs. Victim Age Group

```
ggplot(clean_data, aes(x = PERP_AGE_GROUP, fill = VIC_AGE_GROUP)) +
  geom_bar(position = "dodge") +
  labs(title = "Perpetrator Age Group vs. Victim Age Group",
       x = "Perpetrator Age Group",
```

```

y = "Count",
fill = "Victim Age Group") +
theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))

```



Notice odd age groups

Odd Perpetrator Age groups: 1020, 940, 224, 1028
 Odd Victim Age groups: 1022
 Replace with UNKNOWN for specific odd age groups for perpetrators and victims

```

odd_perp_age_groups <- c("1020", "940", "224", "1028")
odd_vic_age_groups <- c("1022")

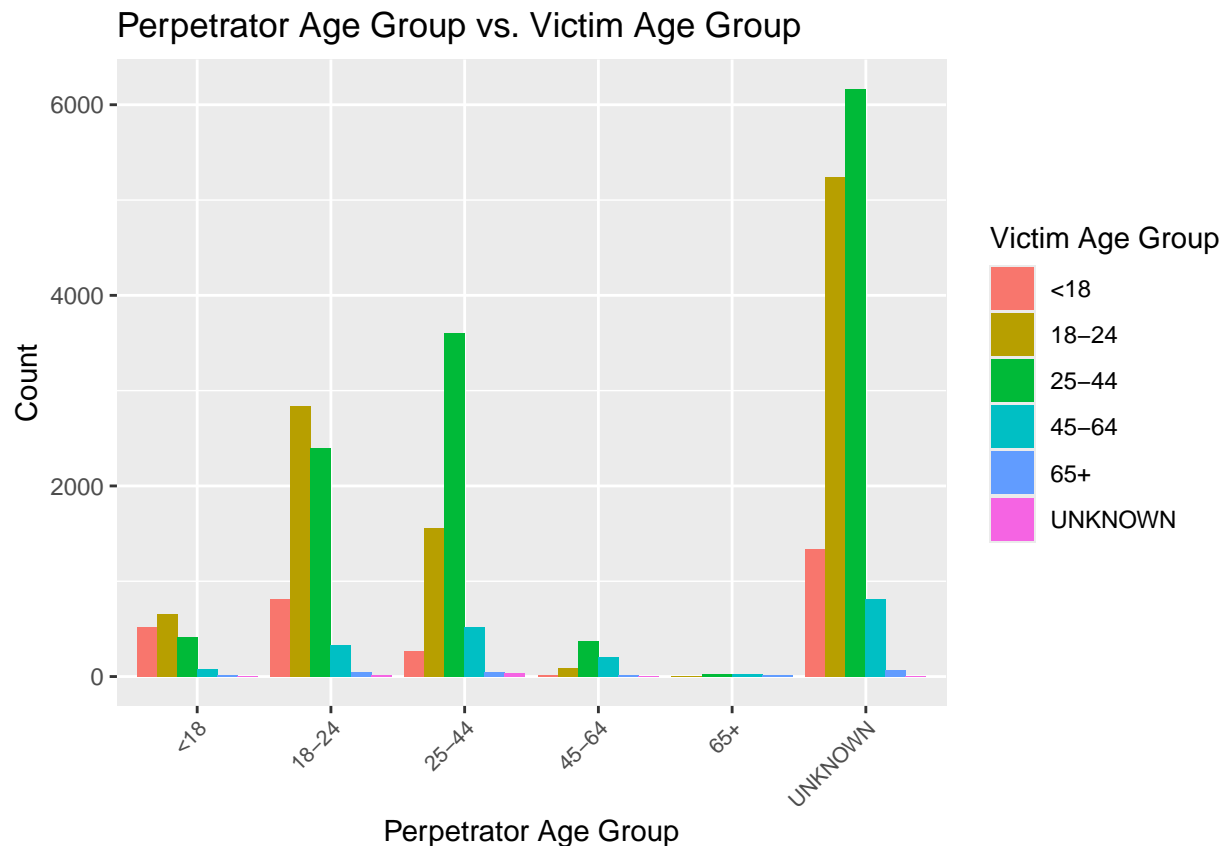
clean_data <- clean_data %>%
  mutate(
    PERP_AGE_GROUP = case_when(
      PERP_AGE_GROUP %in% odd_perp_age_groups ~ "UNKNOWN",
      TRUE ~ PERP_AGE_GROUP
    ),
    VIC_AGE_GROUP = case_when(
      VIC_AGE_GROUP %in% odd_vic_age_groups ~ "UNKNOWN",
      TRUE ~ VIC_AGE_GROUP
    )
  )

```


Bar plot of Perpetrator Age Group vs. Victim Age Group - Verify Age

Notice the victim and perpetrator age groups of 18-24 & 25-44 are the highest in these shooting of all known age groups. Again, perpetrator's UNKNOWN is significant relative to the age, and it appears reasonable to assume the UNKNOWN age is similar to their victim's age from this chart.

```
ggplot(clean_data, aes(x = PERP_AGE_GROUP, fill = VIC_AGE_GROUP)) +  
  geom_bar(position = "dodge") +  
  labs(title = "Perpetrator Age Group vs. Victim Age Group",  
        x = "Perpetrator Age Group",  
        y = "Count",  
        fill = "Victim Age Group") +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))
```



Bar plot by Perpetrator Sex

```
ggplot(clean_data, aes(x = PERP_AGE_GROUP, fill = VIC_SEX)) +  
  geom_bar(position = "dodge") +  
  facet_wrap(~ PERP_SEX) +  
  labs(title = "Perpetrator Age Group vs. Victim Sex Faceted by Perpetrator Sex",  
        x = "Perpetrator Age Group",  
        y = "Count",  
        fill = "Victim Sex") +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))
```

Perpetrator Age Group vs. Victim Sex Faceted by Perpetrator Sex



Replace sex “U” with “UNKNOWN”

```
# add levels to sex for UNKNOWN if not in it
levels(clean_data$PERP_SEX) <- c(levels(clean_data$PERP_SEX), "UNKNOWN")
levels(clean_data$PERP_AGE_GROUP) <- c(levels(clean_data$PERP_AGE_GROUP), "UNKNOWN")
levels(clean_data$PERP_RACE) <- c(levels(clean_data$PERP_RACE), "UNKNOWN")
levels(clean_data$VIC_SEX) <- c(levels(clean_data$VIC_SEX), "UNKNOWN")
levels(clean_data$VIC_AGE_GROUP) <- c(levels(clean_data$VIC_AGE_GROUP), "UNKNOWN")
levels(clean_data$VIC_RACE) <- c(levels(clean_data$VIC_RACE), "UNKNOWN")
clean_data <- clean_data %>%
  mutate(
    PERP_SEX = replace(PERP_SEX, PERP_SEX == "U", "UNKNOWN"),
    VIC_SEX = replace(VIC_SEX, VIC_SEX == "U", "UNKNOWN")
  )
```

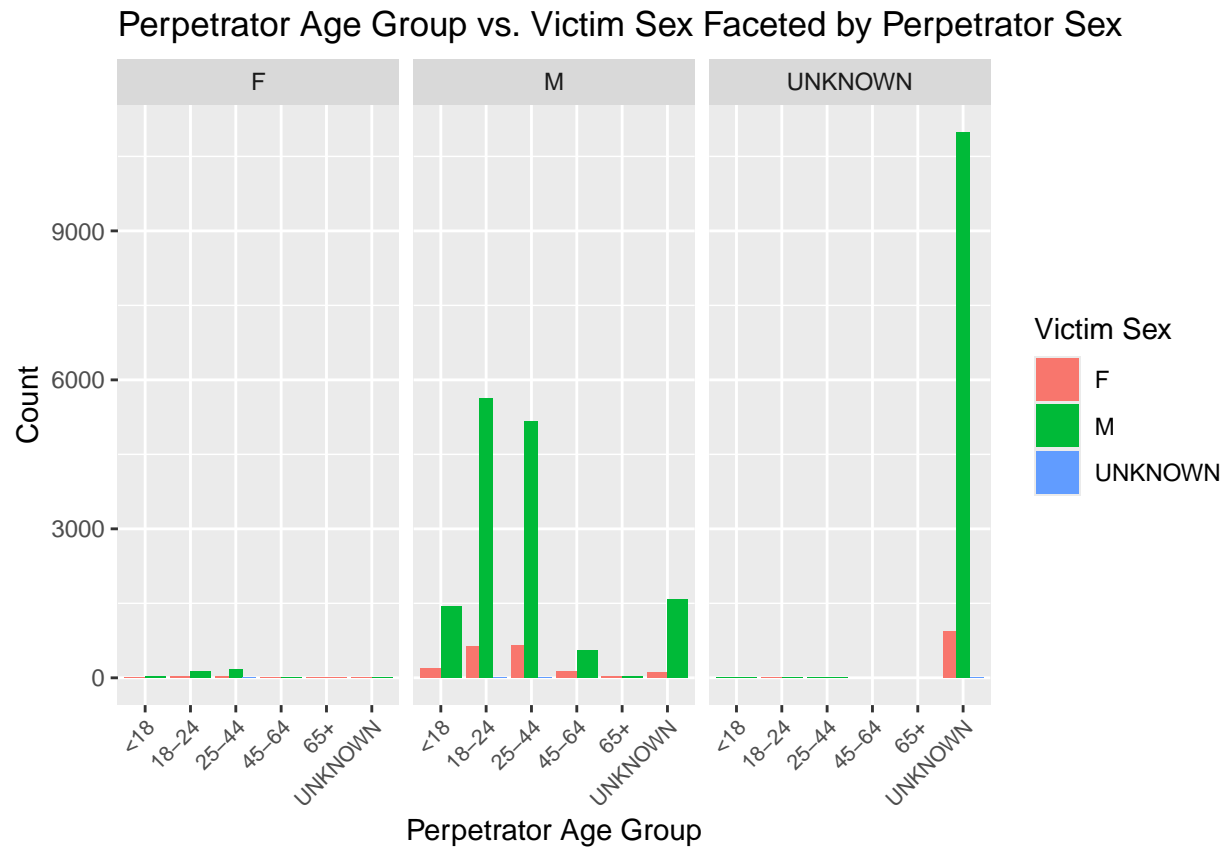
Bar plot by Perpetrator Sex Clean

```
ggplot(clean_data, aes(x = PERP_AGE_GROUP, fill = VIC_SEX)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ PERP_SEX) +
  labs(title = "Perpetrator Age Group vs. Victim Sex Faceted by Perpetrator Sex",
       x = "Perpetrator Age Group",
```

```

y = "Count",
fill = "Victim Sex") +
theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))

```



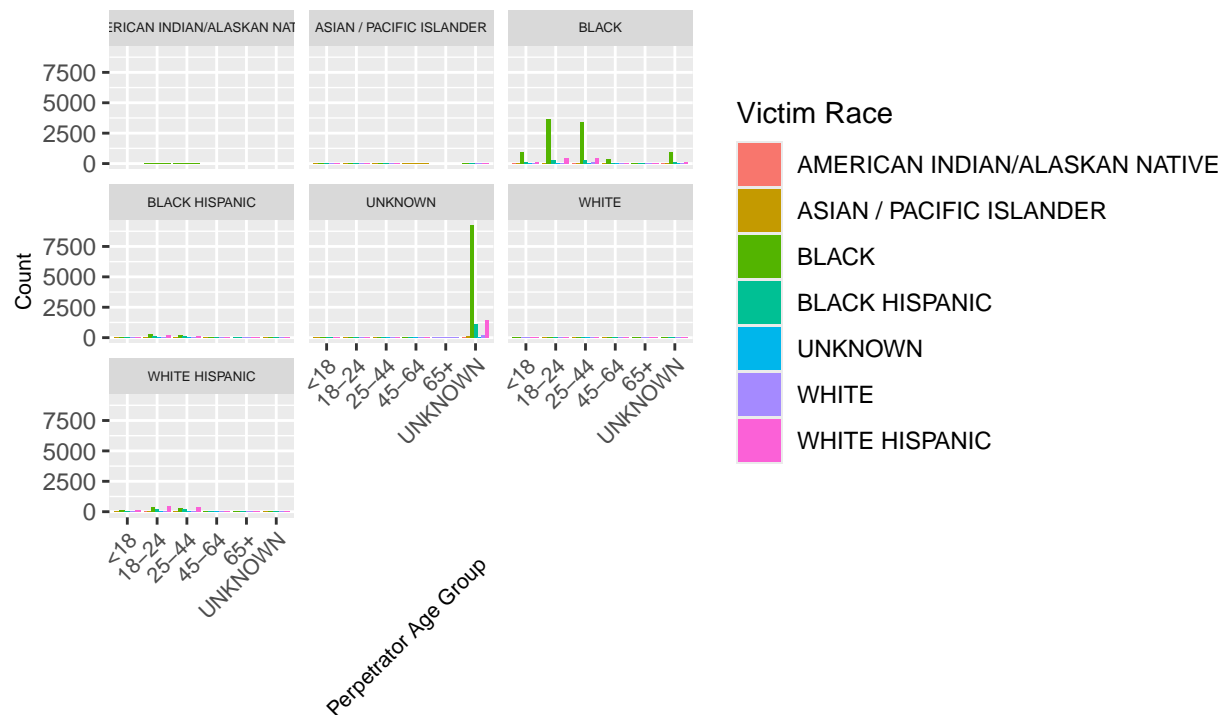
Bar plot by Perpetrator Race vs. Age Group

```

ggplot(clean_data, aes(x = PERP_AGE_GROUP, fill = VIC_RACE)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ PERP_RACE) +
  labs(title = "Perpetrator Age Group vs. Victim Race Faceted by Perpetrator Race",
        x = "Perpetrator Age Group",
        y = "Count",
        fill = "Victim Race") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8), axis.title = element_text(angle=45,

```

Perpetrator Age Group vs. Victim Race Faceted by Perpetrator Race

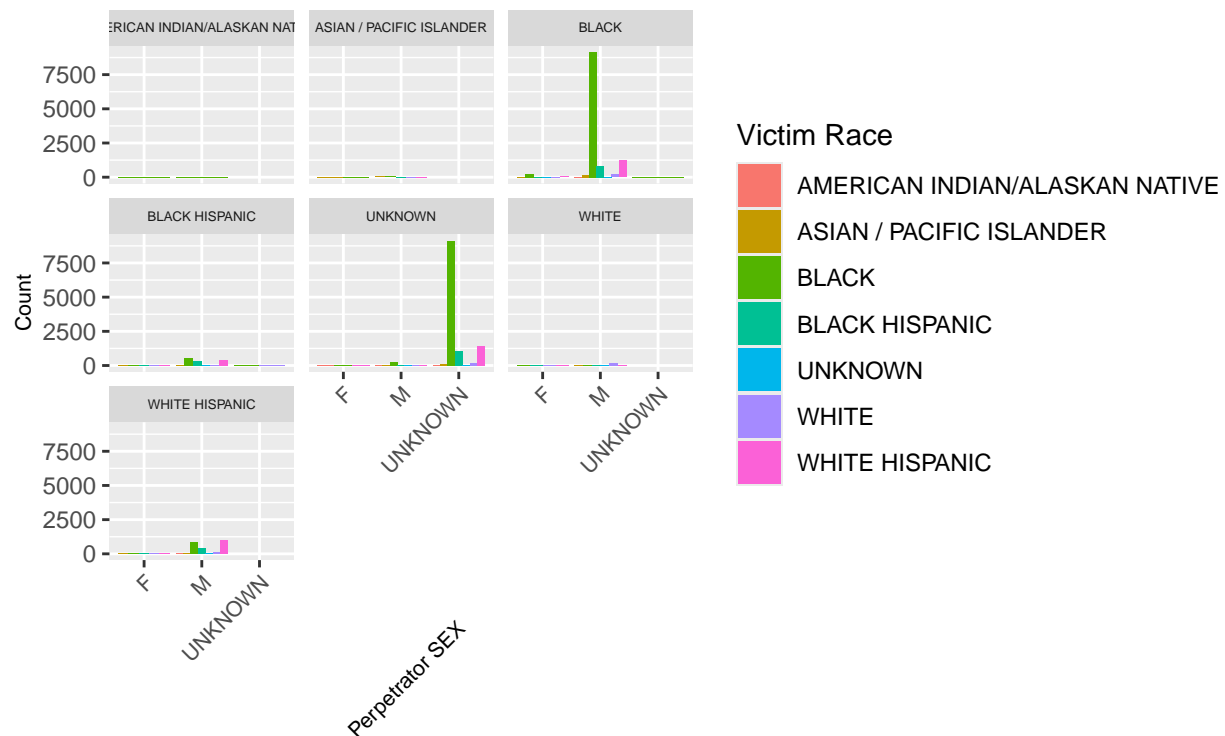


Bar plot by Perpetrator Race vs. Sex

It appears that Black Males are the majority of shooters among all races and genders.

```
ggplot(clean_data, aes(x = PERP_SEX, fill = VIC_RACE)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ PERP_RACE) +
  labs(title = "Perpetrator SEX vs. Victim Race Faceted by Perpetrator Race",
       x = "Perpetrator SEX",
       y = "Count",
       fill = "Victim Race") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8), axis.title = element_text(angle=45,
```

Perpetrator SEX vs. Victim Race Faceted by Perpetrator Race



Model Selection/training

Multinomial logistic regression model

Remove unused levels

When training a model, we need to ensure that the levels are consistent.

```
clean_data <- droplevels(clean_data)
```

```
# Contingency table for Victim Age Group vs. Perpetrator Age Group
table_vic_perp_age <- table(clean_data$VIC_AGE_GROUP, clean_data$PERP_AGE_GROUP)
print(table_vic_perp_age)
```

```
##
##      <18  18-24  25-44  45-64  65+  UNKNOWN
## <18    521   808   270    21    0   1334
## 18-24   652  2841  1560    85    2   5244
## 25-44   413  2394  3600   373   27   6166
## 45-64    79   335   524   202   24    817
## 65+     15    47    49    13   12     69
## UNKNOWN    2    13    38     5    0      7
```

```
# Contingency table for Victim Race vs. Perpetrator Race
table_vic_perp_race <- table(clean_data$VIC_RACE, clean_data$PERP_RACE)
print(table_vic_perp_race)
```

```
##
##              AMERICAN INDIAN/ALASKAN NATIVE
## AMERICAN INDIAN/ALASKAN NATIVE              0
## ASIAN / PACIFIC ISLANDER                    0
## BLACK                                         2
## BLACK HISPANIC                              0
## UNKNOWN                                      0
## WHITE                                        0
## WHITE HISPANIC                              0
##
##              ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
## AMERICAN INDIAN/ALASKAN NATIVE              0      4      0
## ASIAN / PACIFIC ISLANDER                   61    164     20
## BLACK                                       56   9411    561
## BLACK HISPANIC                           14    839    365
## UNKNOWN                                   0     25      6
## WHITE                                    12    205     36
## WHITE HISPANIC                           26   1255    404
##
##              UNKNOWN WHITE WHITE HISPANIC
## AMERICAN INDIAN/ALASKAN NATIVE              6      0      1
## ASIAN / PACIFIC ISLANDER                   140    13     42
## BLACK                                       9318    42    845
## BLACK HISPANIC                           1114    23    440
## UNKNOWN                                   26      1     12
## WHITE                                    207   165    103
## WHITE HISPANIC                           1477    54   1067
```

```
# Contingency table for Victim Sex vs. Perpetrator Sex
table_vic_perp_sex <- table(clean_data$VIC_SEX, clean_data$PERP_SEX)
print(table_vic_perp_sex)
```

```
##
##          F      M UNKNOWN
## F          77   1755     928
## M         366  14406   11018
## UNKNOWN      1      7      4
```

```
# Unique values for victim and perpetrator characteristics
unique_vic_race <- unique(clean_data$VIC_RACE)
unique_vic_sex <- unique(clean_data$VIC_SEX)
unique_vic_age_group <- unique(clean_data$VIC_AGE_GROUP)

unique_perp_race <- unique(clean_data$PERP_RACE)
unique_perp_sex <- unique(clean_data$PERP_SEX)
unique_perp_age_group <- unique(clean_data$PERP_AGE_GROUP)

# Generate all combinations of perpetrator race, gender, and age group
```

```
all_combinations <- expand.grid(
  VIC_RACE = unique_vic_race,
  VIC_SEX = unique_vic_sex,
  VIC_AGE_GROUP = unique_vic_age_group,
  PERP_RACE = unique_perp_race,
  PERP_SEX = unique_perp_sex,
  PERP_AGE_GROUP = unique_perp_age_group
)

# View the first few rows of the generated combinations
head(all_combinations)
```

```
##           VIC_RACE VIC_SEX VIC_AGE_GROUP PERP_RACE PERP_SEX
## 1           BLACK      M       25-44      BLACK      M
## 2           WHITE      M       25-44      BLACK      M
## 3      WHITE HISPANIC      M       25-44      BLACK      M
## 4      BLACK HISPANIC      M       25-44      BLACK      M
## 5 ASIAN / PACIFIC ISLANDER      M       25-44      BLACK      M
## 6           UNKNOWN      M       25-44      BLACK      M
## PERP_AGE_GROUP
## 1       25-44
## 2       25-44
## 3       25-44
## 4       25-44
## 5       25-44
## 6       25-44
```

Model for PERP_AGE_GROUP

```
age_model <- multinom(PERP_AGE_GROUP ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX, data = clean_data)
```

```
## # weights: 90 (70 variable)
## initial value 51176.233960
## iter 10 value 36943.478916
## iter 20 value 36639.281733
## iter 30 value 35831.734899
## iter 40 value 35553.028622
## iter 50 value 35402.069777
## iter 60 value 35346.495307
## iter 70 value 35332.548587
## iter 80 value 35331.933799
## iter 90 value 35331.795897
## final value 35331.795071
## converged
```

```
summary(age_model)
```

```
## Call:
## multinom(formula = PERP_AGE_GROUP ~ VIC_AGE_GROUP + VIC_RACE +
##           VIC_SEX, data = clean_data)
```

```

##
## Coefficients:
##      (Intercept) VIC_AGE_GROUP18-24 VIC_AGE_GROUP25-44 VIC_AGE_GROUP45-64
## 18-24      -0.9244584           1.024050           1.306704           0.988923
## 25-44      -0.9143144           1.531975           2.818131           2.504326
## 45-64     -11.2759284           1.194788           3.105194           3.989719
## 65+       -15.0471673           7.509251           10.517712          11.569059
## UNKNOWN    0.6008942           1.128902           1.750040           1.420436
##      VIC_AGE_GROUP65+ VIC_AGE_GROUPUNKNOWN VIC_RACEASIAN / PACIFIC ISLANDER
## 18-24           0.6784858           1.0650870           1.2047114
## 25-44           1.7496282           3.2306905           0.1443714
## 45-64           2.7314039           3.5097206           8.9483958
## 65+            12.1302456           0.7876541          -4.9339643
## UNKNOWN         0.6852055           0.0881566          -0.4948816
##      VIC_RACEBLACK VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN VIC_RACEWHITE
## 18-24           1.292163           1.2136979           1.6787301           1.7660871
## 25-44           0.330123           0.1939993           1.1263704           1.2168681
## 45-64           8.389152           8.1608342           9.8183919          10.0328058
## 65+            2.844613           3.0128648          -5.9050866           5.5233798
## UNKNOWN         0.127846          -0.2545571           0.4384378           0.2076029
##      VIC_RACEWHITE HISPANIC      VIC_SEXM VIC_SEXUNKNOWN
## 18-24           1.3170331  0.08724731      8.3148833
## 25-44           0.3402653 -0.08289398      6.6417726
## 45-64           8.6060745 -0.47288053     -5.2671203
## 65+            3.2299286 -1.63651798      0.2620697
## UNKNOWN        -0.2924216  0.36366174      8.0041405
##
## Std. Errors:
##      (Intercept) VIC_AGE_GROUP18-24 VIC_AGE_GROUP25-44 VIC_AGE_GROUP45-64
## 18-24      1.4241390      0.07134172      0.07781601      0.1376082
## 25-44      1.2555342      0.08863781      0.09161712      0.1426485
## 45-64      0.2520877      0.25120922      0.23442409      0.2603301
## 65+       6.4576772      2.63729741      2.58501416      2.5863620
## UNKNOWN   1.0744204      0.06671399      0.07294538      0.1292906
##      VIC_AGE_GROUP65+ VIC_AGE_GROUPUNKNOWN VIC_RACEASIAN / PACIFIC ISLANDER
## 18-24           0.3030489           0.78115897           1.4373525
## 25-44           0.3061938           0.74178614           1.2702315
## 45-64           0.4437977           0.89049296           0.2748167
## 65+            2.5987392           0.05328585          31.7160937
## UNKNOWN         0.2911826           0.81924368           1.0911413
##      VIC_RACEBLACK VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN VIC_RACEWHITE
## 18-24           1.4217499           1.4236526           1.61915463           1.4409611
## 25-44           1.2520243           1.2543536           1.46597173           1.2729373
## 45-64           0.1548337           0.1951875           0.77822244           0.2588753
## 65+            9.0308545           9.0361375           0.06113658           9.0335795
## UNKNOWN         1.0715904           1.0739195           1.30765261           1.0958392
##      VIC_RACEWHITE HISPANIC      VIC_SEXM VIC_SEXUNKNOWN
## 18-24           1.4229745  0.08697427      4.307801e-01
## 25-44           1.2534835  0.08847244      5.310482e-01
## 45-64           0.1720977  0.12777076      7.895726e-07
## 65+            9.0327763  0.27310627      1.150377e-05
## UNKNOWN         1.0731531  0.08344500      4.746731e-01
##
## Residual Deviance: 70663.59

```



```
## AIC: 70803.59
```

Model for PERP_RACE

```
race_model <- multinom(PERP_RACE ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX, data = clean_data)
```

```
## # weights: 105 (84 variable)
## initial value 55579.085677
## iter 10 value 34733.255277
## iter 20 value 33990.466735
## iter 30 value 32508.091332
## iter 40 value 31856.507428
## iter 50 value 31349.546100
## iter 60 value 31230.771414
## iter 70 value 31195.922943
## iter 80 value 31184.748810
## iter 90 value 31183.746796
## iter 100 value 31183.387797
## final value 31183.387797
## stopped after 100 iterations
```

```
summary(race_model)
```

```
## Call:
## multinom(formula = PERP_RACE ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX,
## data = clean_data)
##
## Coefficients:
## (Intercept) VIC_AGE_GROUP18-24 VIC_AGE_GROUP25-44
## ASIAN / PACIFIC ISLANDER 12.86887 6.553150 -6.856668
## BLACK 26.42123 6.454747 -7.307312
## BLACK HISPANIC 12.38022 6.383573 -7.454881
## UNKNOWN 26.30540 6.655952 -7.149366
## WHITE 13.95925 6.711761 -6.616384
## WHITE HISPANIC 24.80924 6.426062 -7.378855
## VIC_AGE_GROUP45-64 VIC_AGE_GROUP65+
## ASIAN / PACIFIC ISLANDER -9.229065 -0.1419325
## BLACK -9.318187 0.7932155
## BLACK HISPANIC -9.471450 1.0538642
## UNKNOWN -9.328787 0.5152134
## WHITE -8.414135 2.0447737
## WHITE HISPANIC -9.368347 0.2644328
## VIC_AGE_GROUPUNKNOWN VIC_RACEASIAN / PACIFIC ISLANDER
## ASIAN / PACIFIC ISLANDER -16.350507 11.301148
## BLACK 2.542143 -1.227556
## BLACK HISPANIC 2.263549 10.487086
## UNKNOWN 0.766080 -1.771431
## WHITE 1.346217 8.060077
## WHITE HISPANIC 2.614632 -1.157053
## VIC_RACEBLACK VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN
## ASIAN / PACIFIC ISLANDER 3.614348 8.7437562 -13.873902
```

## BLACK	-4.783889	-0.6884695	-1.428880
## BLACK HISPANIC	6.205898	12.2874676	11.216835
## UNKNOWN	-5.194031	-0.8016951	-1.389450
## WHITE	1.731458	7.6471896	7.790332
## WHITE HISPANIC	-5.776722	0.0844545	-0.568337
##	VIC_RACEWHITE	VIC_RACEWHITE HISPANIC	VIC_SEXM
## ASIAN / PACIFIC ISLANDER	10.4972433	9.5281663	-6.329724
## BLACK	-0.2047667	-0.1028563	-6.076347
## BLACK HISPANIC	11.9019990	12.5836231	-5.710140
## UNKNOWN	-0.5069703	-0.3270288	-5.666423
## WHITE	11.3829756	8.6630120	-6.055386
## WHITE HISPANIC	0.5584481	1.1635615	-5.807721
##	VIC_SEXUNKNOWN		
## ASIAN / PACIFIC ISLANDER	0.3506045		
## BLACK	5.3931481		
## BLACK HISPANIC	-5.4662530		
## UNKNOWN	6.0191027		
## WHITE	-2.7668843		
## WHITE HISPANIC	-6.5296137		
##			
## Std. Errors:			
##	(Intercept)	VIC_AGE_GROUP18-24	VIC_AGE_GROUP25-44
## ASIAN / PACIFIC ISLANDER	2.979879	0.29002094	1.697546
## BLACK	3.043836	0.08639871	1.676282
## BLACK HISPANIC	3.004487	0.11011950	1.677728
## UNKNOWN	3.039321	0.08654538	1.676293
## WHITE	3.013527	0.27688503	1.695363
## WHITE HISPANIC	3.086151	0.09968660	1.677056
##	VIC_AGE_GROUP45-64	VIC_AGE_GROUP65+	
## ASIAN / PACIFIC ISLANDER	1.702721	0.8892658	
## BLACK	1.666964	0.2240602	
## BLACK HISPANIC	1.670490	0.2895097	
## UNKNOWN	1.667015	0.2324469	
## WHITE	1.689797	0.4018563	
## WHITE HISPANIC	1.668729	0.2941476	
##	VIC_AGE_GROUPUNKNOWN	VIC_RACEASIAN / PACIFIC ISLANDER	
## ASIAN / PACIFIC ISLANDER	1.221622e-09		2.040555
## BLACK	2.980552e-01		2.065101
## BLACK HISPANIC	4.512807e-01		2.011849
## UNKNOWN	4.079231e-01		2.058566
## WHITE	8.647830e-01		2.018330
## WHITE HISPANIC	3.424749e-01		2.128921
##	VIC_RACEBLACK	VIC_RACEBLACK HISPANIC	VIC_RACEUNKNOWN
## ASIAN / PACIFIC ISLANDER	7.809175	4.506951	1.191130e-08
## BLACK	7.856455	4.496737	5.961155e-01
## BLACK HISPANIC	7.840666	4.470714	4.003810e-01
## UNKNOWN	7.854694	4.493640	5.682969e-01
## WHITE	7.842514	4.474437	7.263888e-01
## WHITE HISPANIC	7.872817	4.525237	7.965141e-01
##	VIC_RACEWHITE	VIC_RACEWHITE HISPANIC	VIC_SEXM
## ASIAN / PACIFIC ISLANDER	0.9400131	3.319119	3.314083
## BLACK	1.0138113	3.319604	3.308721
## BLACK HISPANIC	0.8939856	3.284060	3.309558
## UNKNOWN	1.0000588	3.315425	3.308748

```
## WHITE 0.8954511 3.287120 3.311728
## WHITE HISPANIC 1.1346473 3.358018 3.309091
## VIC_SEXUNKNOWN
## ASIAN / PACIFIC ISLANDER 1.328690e-05
## BLACK 3.403456e-01
## BLACK HISPANIC 1.566520e-06
## UNKNOWN 3.403447e-01
## WHITE 2.413741e-05
## WHITE HISPANIC 1.312090e-06
##
## Residual Deviance: 62366.78
## AIC: 62534.78
```

Model for PERP_SEX

```
sex_model <- multinom(PERP_SEX ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX, data = clean_data)
```

```
## # weights: 45 (28 variable)
## initial value 31378.564189
## iter 10 value 23641.499119
## iter 20 value 21774.618198
## iter 30 value 21189.615282
## iter 40 value 21188.756473
## final value 21188.754411
## converged
```

```
summary(sex_model)
```

```
## Call:
## multinom(formula = PERP_SEX ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX,
## data = clean_data)
##
## Coefficients:
## (Intercept) VIC_AGE_GROUP18-24 VIC_AGE_GROUP25-44 VIC_AGE_GROUP45-64
## M 1.963873 -0.6087350 -0.8920714 -1.059258
## UNKNOWN 1.093870 -0.3844589 -0.6952135 -1.029213
## VIC_AGE_GROUP65+ VIC_AGE_GROUPUNKNOWN VIC_RACEASIAN / PACIFIC ISLANDER
## M -0.5532355 -0.8277576 1.496276
## UNKNOWN -0.7750257 -2.5351516 1.109400
## VIC_RACEBLACK VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN VIC_RACEWHITE
## M 1.931508 2.112784 2.950059 1.551779
## UNKNOWN 2.167252 2.070459 3.042859 1.069760
## VIC_RACEWHITE HISPANIC VIC_SEXM VIC_SEXUNKNOWN
## M 1.841210 0.5416342 -1.903586
## UNKNOWN 1.573242 0.8585556 -1.138234
##
## Std. Errors:
## (Intercept) VIC_AGE_GROUP18-24 VIC_AGE_GROUP25-44 VIC_AGE_GROUP45-64
## M 1.108439 0.2238559 0.2169580 0.2529777
## UNKNOWN 1.146922 0.2248092 0.2179334 0.2549733
## VIC_AGE_GROUP65+ VIC_AGE_GROUPUNKNOWN VIC_RACEASIAN / PACIFIC ISLANDER
```

```
## M          0.5510499          0.8419843          1.124848
## UNKNOWN    0.5613373          0.9316651          1.164241
##          VIC_RACEBLACK VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN VIC_RACEWHITE
## M          1.089591          1.100598          1.755197          1.110002
## UNKNOWN    1.128308          1.139088          1.785551          1.149461
##          VIC_RACEWHITE HISPANIC VIC_SEXM VIC_SEXUNKNOWN
## M          1.093701 0.1298153          1.386226
## UNKNOWN    1.132430 0.1319771          1.405314
##
## Residual Deviance: 42377.51
## AIC: 42433.51
```

Inference with Model on Sample Data Point

Define a single data point for victim characteristics

```
# Define a single data point for victim characteristics
single_data_point <- clean_data[1, c("VIC_AGE_GROUP", "VIC_RACE", "VIC_SEX")]
single_data_point[1, ] <- list("25-44", "WHITE", "M")
```

Predict the perpetrator's age group

```
predicted_age_group <- predict(age_model, newdata = single_data_point)
print(paste("Predicted Perpetrator Age Group:", predicted_age_group))
```

```
## [1] "Predicted Perpetrator Age Group: 25-44"
```

Predict the perpetrator's race

```
predicted_race <- predict(race_model, newdata = single_data_point)
print(paste("Predicted Perpetrator Race:", predicted_race))
```

```
## [1] "Predicted Perpetrator Race: UNKNOWN"
```

Predict the perpetrator's sex

```
predicted_sex <- predict(sex_model, newdata = single_data_point)
print(paste("Predicted Perpetrator Sex:", predicted_sex))
```

```
## [1] "Predicted Perpetrator Sex: M"
```

Look at predicted probabilities

Generate all combinations for a specific victim profile

```

specific_combinations <- expand.grid(
  VIC_RACE = "WHITE",
  VIC_SEX = "M",
  VIC_AGE_GROUP = "25-44",
  PERP_RACE = unique(clean_data$PERP_RACE),
  PERP_SEX = unique(clean_data$PERP_SEX),
  PERP_AGE_GROUP = unique(clean_data$PERP_AGE_GROUP)
)
# Predict probabilities for the specific victim profile
predicted_probabilities_age_model <- predict(age_model, newdata = specific_combinations, type = "probs")

# Combine the predictions with the specific combinations
results_with_probabilities_age_model <- cbind(specific_combinations, predicted_probabilities_age_model)

# Predict probabilities for the specific victim profile
predicted_probabilities_race_model <- predict(race_model, newdata = specific_combinations, type = "probs")

# Combine the predictions with the specific combinations
results_with_probabilities_race_model <- cbind(specific_combinations, predicted_probabilities_race_model)

# Predict probabilities for the specific victim profile
predicted_probabilities_sex_model <- predict(sex_model, newdata = specific_combinations, type = "probs")

# Combine the predictions with the specific combinations
results_with_probabilities_sex_model <- cbind(specific_combinations, predicted_probabilities_sex_model)

# View the results
head(results_with_probabilities_age_model)

```

```

##   VIC_RACE VIC_SEX VIC_AGE_GROUP          PERP_RACE PERP_SEX
## 1   WHITE      M      25-44                BLACK      M
## 2   WHITE      M      25-44                UNKNOWN     M
## 3   WHITE      M      25-44          WHITE HISPANIC     M
## 4   WHITE      M      25-44          BLACK HISPANIC     M
## 5   WHITE      M      25-44 ASIAN / PACIFIC ISLANDER     M
## 6   WHITE      M      25-44                WHITE      M
##   PERP_AGE_GROUP      <18    18-24    25-44    45-64    65+
## 1      25-44 0.01840582 0.1721295 0.3839289 0.07383683 0.009680106
## 2      25-44 0.01840582 0.1721295 0.3839289 0.07383683 0.009680106
## 3      25-44 0.01840582 0.1721295 0.3839289 0.07383683 0.009680106
## 4      25-44 0.01840582 0.1721295 0.3839289 0.07383683 0.009680106
## 5      25-44 0.01840582 0.1721295 0.3839289 0.07383683 0.009680106
## 6      25-44 0.01840582 0.1721295 0.3839289 0.07383683 0.009680106
##   UNKNOWN
## 1 0.3420188
## 2 0.3420188
## 3 0.3420188
## 4 0.3420188
## 5 0.3420188
## 6 0.3420188

```

```
# View the results
```

```
head(results_with_probabilities_race_model)
```

```
##   VIC_RACE VIC_SEX VIC_AGE_GROUP          PERP_RACE PERP_SEX
## 1   WHITE      M      25-44          BLACK        M
## 2   WHITE      M      25-44          UNKNOWN       M
## 3   WHITE      M      25-44          WHITE HISPANIC  M
## 4   WHITE      M      25-44          BLACK HISPANIC  M
## 5   WHITE      M      25-44 ASIAN / PACIFIC ISLANDER M
## 6   WHITE      M      25-44          WHITE        M
##   PERP_AGE_GROUP AMERICAN INDIAN/ALASKAN NATIVE ASIAN / PACIFIC ISLANDER
## 1              25-44              7.061867e-07              0.01861733
## 2              25-44              7.061867e-07              0.01861733
## 3              25-44              7.061867e-07              0.01861733
## 4              25-44              7.061867e-07              0.01861733
## 5              25-44              7.061867e-07              0.01861733
## 6              25-44              7.061867e-07              0.01861733
##   BLACK BLACK HISPANIC   UNKNOWN   WHITE WHITE HISPANIC
## 1 0.2643238   0.04753975 0.3070499 0.2247123   0.1377562
## 2 0.2643238   0.04753975 0.3070499 0.2247123   0.1377562
## 3 0.2643238   0.04753975 0.3070499 0.2247123   0.1377562
## 4 0.2643238   0.04753975 0.3070499 0.2247123   0.1377562
## 5 0.2643238   0.04753975 0.3070499 0.2247123   0.1377562
## 6 0.2643238   0.04753975 0.3070499 0.2247123   0.1377562
```

```
# View the results
```

```
head(results_with_probabilities_sex_model)
```

```
##   VIC_RACE VIC_SEX VIC_AGE_GROUP          PERP_RACE PERP_SEX
## 1   WHITE      M      25-44          BLACK        M
## 2   WHITE      M      25-44          UNKNOWN       M
## 3   WHITE      M      25-44          WHITE HISPANIC  M
## 4   WHITE      M      25-44          BLACK HISPANIC  M
## 5   WHITE      M      25-44 ASIAN / PACIFIC ISLANDER M
## 6   WHITE      M      25-44          WHITE        M
##   PERP_AGE_GROUP          F          M   UNKNOWN
## 1              25-44 0.0286199 0.6781156 0.2932645
## 2              25-44 0.0286199 0.6781156 0.2932645
## 3              25-44 0.0286199 0.6781156 0.2932645
## 4              25-44 0.0286199 0.6781156 0.2932645
## 5              25-44 0.0286199 0.6781156 0.2932645
## 6              25-44 0.0286199 0.6781156 0.2932645
```

```
# Initialize an empty data frame to store the results
```

```
results_combined <- data.frame(
  PERP_Sex = character(),
  PERP_Age_Group = character(),
  PERP_Race_Group = character(),
  Combined_Probability = numeric(),
  stringsAsFactors = FALSE
)
```

```

for (sex in unique_perp_sex) {
  for (age_group in unique_perp_age_group) {
    for (race_group in unique_perp_race) {

      # Extract the corresponding probabilities from each model
      age_prob <- results_with_probabilities_age_model[[age_group]][results_with_probabilities_age_model$PERP_Sex == sex]
      age_prob <- age_prob[1]

      race_prob <- results_with_probabilities_race_model[[race_group]][results_with_probabilities_race_model$PERP_Age_Group == age_group]
      race_prob <- race_prob[1]

      sex_prob <- results_with_probabilities_sex_model[[sex]][results_with_probabilities_sex_model$PERP_Age_Group == age_group]
      sex_prob <- sex_prob[1]

      # Ensure there's only one value returned, if not handle accordingly
      if (length(age_prob) == 1 && length(race_prob) == 1 && length(sex_prob) == 1) {

        # Calculate the combined probability
        combined_prob <- age_prob * race_prob * sex_prob

        # Store the result
        results_combined <- rbind(results_combined, data.frame(PERP_Sex = sex, PERP_Age_Group = age_group, PERP_Race_Group = race_group, Combined_Probability = combined_prob))
      }
    }
  }
}

# View the results
# Sort the data frame by the Combined_Probability column in descending order
results_combined_sorted <- results_combined[order(-results_combined$Combined_Probability), ]
head(results_combined_sorted)

```

```

##      PERP_Sex PERP_Age_Group PERP_Race_Group Combined_Probability
## 2          M      25-44          UNKNOWN          0.07993990
## 9          M      UNKNOWN          UNKNOWN          0.07121356
## 1          M      25-44          BLACK          0.06881622
## 8          M      UNKNOWN          BLACK          0.06130416
## 6          M      25-44          WHITE          0.05850345
## 13         M      UNKNOWN          WHITE          0.05211715

```

All vic combinations

```

suppressWarnings({# Initialize an empty data frame to store the results
results_combined_ALL <- data.frame(
  VIC_Sex = character(),
  VIC_Age_Group = character(),
  VIC_Race_Group = character(),
  PERP_Sex = character(),

```

```

    PERP_Age_Group = character(),
    PERP_Race_Group = character(),
    Combined_Probability = numeric(),
    stringsAsFactors = FALSE
  )
results_combined_ALL_filtered <- data.frame(
  VIC_Sex = character(),
  VIC_Age_Group = character(),
  VIC_Race_Group = character(),
  PERP_Sex = character(),
  PERP_Age_Group = character(),
  PERP_Race_Group = character(),
  Count = integer(),
  Probability = numeric(),
  stringsAsFactors = FALSE
)

# Loop through each combination of sex, age group, and race group
for (sex_vic in unique_vic_sex){
  for (age_group_vic in unique_vic_age_group){
    for (race_group_vic in unique_vic_race){
      specific_combinations <- expand.grid(
        VIC_RACE = race_group_vic,
        VIC_SEX = sex_vic,
        VIC_AGE_GROUP = age_group_vic,
        PERP_RACE = unique(clean_data$PERP_RACE),
        PERP_SEX = unique(clean_data$PERP_SEX),
        PERP_AGE_GROUP = unique(clean_data$PERP_AGE_GROUP))
      # Predict probabilities for the specific victim profile
      predicted_probabilities_age_model <- predict(age_model, newdata = specific_combinations, type = "j")

      # Combine the predictions with the specific combinations
      results_with_probabilities_age_model <- cbind(specific_combinations, predicted_probabilities_age_model)

      # Predict probabilities for the specific victim profile
      predicted_probabilities_race_model <- predict(race_model, newdata = specific_combinations, type = "j")

      # Combine the predictions with the specific combinations
      results_with_probabilities_race_model <- cbind(specific_combinations, predicted_probabilities_race_model)

      # Predict probabilities for the specific victim profile
      predicted_probabilities_sex_model <- predict(sex_model, newdata = specific_combinations, type = "j")

      # Combine the predictions with the specific combinations
      results_with_probabilities_sex_model <- cbind(specific_combinations, predicted_probabilities_sex_model)

      filtered_data <- clean_data %>%
        filter(VIC_RACE == race_group_vic, VIC_SEX == sex_vic, VIC_AGE_GROUP == age_group_vic)

      # Calculate the frequency of each perpetrator combination
      perp_combinations <- filtered_data %>%
        group_by(PERP_RACE, PERP_SEX, PERP_AGE_GROUP) %>%
        summarise(Count = n(), .groups = 'drop') %>%

```



```

    mutate(Probability = Count / sum(Count)) # Calculate observed probabilities

# Ensure that the columns match between the two data frames before rbind
perp_comb_binded <- data.frame(
  VIC_Sex = rep(sex_vic, nrow(perp_combinations)),
  VIC_Age_Group = rep(age_group_vic, nrow(perp_combinations)),
  VIC_Race_Group = rep(race_group_vic, nrow(perp_combinations)),
  PERP_Sex = perp_combinations$PERP_SEX,
  PERP_Age_Group = perp_combinations$PERP_AGE_GROUP,
  PERP_Race_Group = perp_combinations$PERP_RACE,
  Count = perp_combinations$Count,
  Probability = perp_combinations$Probability,
  stringsAsFactors = FALSE
)

# Store the result
results_combined_ALL_filtered <- rbind(results_combined_ALL_filtered, perp_comb_binded)

for (sex in unique_perp_sex) {
  for (age_group in unique_perp_age_group) {
    for (race_group in unique_perp_race) {

      # Extract the corresponding probabilities from each model
      age_prob <- results_with_probabilities_age_model[[age_group]][results_with_probabilities_age_model$PERP_SEX == sex,]
      age_prob <- age_prob[1]

      race_prob <- results_with_probabilities_race_model[[race_group]][results_with_probabilities_race_model$PERP_SEX == sex,]
      race_prob <- race_prob[1]

      sex_prob <- results_with_probabilities_sex_model[[sex]][results_with_probabilities_sex_model$PERP_AGE_GROUP == age_group,]
      sex_prob <- sex_prob[1]

      # Ensure there's only one value returned, if not handle accordingly
      if (length(age_prob) == 1 && length(race_prob) == 1 && length(sex_prob) == 1) {

        # Calculate the combined probability
        combined_prob <- age_prob * race_prob * sex_prob

        # Store the result
        results_combined_ALL <- rbind(results_combined_ALL, data.frame(
          VIC_Sex = sex_vic,
          VIC_Age_Group = age_group_vic,
          VIC_Race_Group = race_group_vic,
          PERP_Sex = sex,
          PERP_Age_Group = age_group,
          PERP_Race_Group = race_group,
          Combined_Probability = combined_prob,
          stringsAsFactors = FALSE
        ))
      }
    }
  }
}

```

```

    }
  }
}
}
}

# View the results
# Sort the data frame by the Combined_Probability column in descending order
results_combined_sorted_ALL <- results_combined_ALL[order(-results_combined_ALL$Combined_Probability), ]
head(results_combined_sorted_ALL)

results_combined_sorted_ALL_filtered <- results_combined_ALL_filtered[order(-results_combined_ALL_filtered$
head(results_combined_sorted_ALL_filtered))}

```

```

##      VIC_Sex VIC_Age_Group          VIC_Race_Group PERP_Sex
## 631      M      <18              UNKNOWN      UNKNOWN
## 652      M      UNKNOWN      ASIAN / PACIFIC ISLANDER      M
## 754      F      25-44              UNKNOWN      M
## 755      F      25-44 AMERICAN INDIAN/ALASKAN NATIVE      M
## 915      F      65+        ASIAN / PACIFIC ISLANDER      UNKNOWN
## 969      F      UNKNOWN      WHITE HISPANIC      M
##      PERP_Age_Group PERP_Race_Group Count Probability
## 631      UNKNOWN      UNKNOWN      4      1
## 652      <NA>      BLACK      1      1
## 754      <NA>      WHITE HISPANIC      1      1
## 755      <NA>      BLACK      1      1
## 915      UNKNOWN      UNKNOWN      2      1
## 969      <NA>      BLACK      1      1

```

```

# Set minimum count threshold, lets look at the power of all counts
min_count_threshold <- 1

# Filter the results by this threshold
results_significant <- results_combined_sorted_ALL_filtered %>%
  filter(Count >= min_count_threshold)

# View the filtered significant results
head(results_significant)

```

```

##      VIC_Sex VIC_Age_Group          VIC_Race_Group PERP_Sex PERP_Age_Group
## 1      M      <18              UNKNOWN      UNKNOWN      UNKNOWN
## 2      M      UNKNOWN      ASIAN / PACIFIC ISLANDER      M      <NA>
## 3      F      25-44              UNKNOWN      M      <NA>
## 4      F      25-44 AMERICAN INDIAN/ALASKAN NATIVE      M      <NA>
## 5      F      65+        ASIAN / PACIFIC ISLANDER      UNKNOWN      UNKNOWN
## 6      F      UNKNOWN      WHITE HISPANIC      M      <NA>
##      PERP_Race_Group Count Probability
## 1      UNKNOWN      4      1
## 2      BLACK      1      1
## 3      WHITE HISPANIC      1      1
## 4      BLACK      1      1
## 5      UNKNOWN      2      1

```

```
## 6          BLACK      1          1
```

```
# Number of unique levels in each category (n)
n_race <- length(unique(results_significant$PERP_Race_Group))
n_sex <- length(unique(results_significant$PERP_Sex))
n_age_group <- length(unique(results_significant$PERP_Age_Group))

# Calculate degrees of freedom (DOF)
df <- (n_race - 1) * (n_sex - 1) * (n_age_group - 1)
calculate_power <- function(row,df) {
  effect_size <- 0.3
  alpha <- 0.05
  group_sample_size <- as.numeric(row['Count'])
  df_group <- df
  power_result <- pwr.chisq.test(w = effect_size, N = group_sample_size, df = df_group, sig.level = a
  return(power_result$power)
}

results_significantwpower <- results_significant %>%
  rowwise() %>%
  mutate(Power = calculate_power(cur_data(),df))
```

```
## Warning: There was 1 warning in 'mutate()'.
## i In argument: 'Power = calculate_power(cur_data(), df)'.
## i In row 1.
## Caused by warning:
## ! 'cur_data()' was deprecated in dplyr 1.1.0.
## i Please use 'pick()' instead.
```

```
head(results_significantwpower)
```

```
## # A tibble: 6 x 9
## # Rowwise:
##   VIC_Sex VIC_Age_Group VIC_Race_Group   PERP_Sex PERP_Age_Group PERP_Race_Group
##   <chr>   <chr>         <chr>         <fct>   <fct>         <fct>
## 1 M      <18          UNKNOWN       UNKNOWN UNKNOWN       UNKNOWN
## 2 M      UNKNOWN     ASIAN / PACIFIC~ M      <NA>         BLACK
## 3 F      25-44        UNKNOWN       M      <NA>         WHITE HISPANIC
## 4 F      25-44        AMERICAN INDIAN~ M      <NA>         BLACK
## 5 F      65+        ASIAN / PACIFIC~ UNKNOWN UNKNOWN       UNKNOWN
## 6 F      UNKNOWN     WHITE HISPANIC M      <NA>         BLACK
## # i 3 more variables: Count <int>, Probability <dbl>, Power <dbl>
```

```
# A good power level cutoff is 80%. This would mean that there is an 80% chance of detecting an effect
# So lets look at values that are above the cutoff of 80%.
head(results_significantwpower %>% filter(Power>0.8))
```

```
## # A tibble: 6 x 9
## # Rowwise:
##   VIC_Sex VIC_Age_Group VIC_Race_Group PERP_Sex PERP_Age_Group PERP_Race_Group
##   <chr>   <chr>         <chr>         <fct>   <fct>         <fct>
## 1 M      25-44        BLACK         UNKNOWN UNKNOWN       UNKNOWN
```

```
## 2 M      18-24      BLACK      UNKNOWN UNKNOWN UNKNOWN
## 3 M      <18       BLACK      UNKNOWN UNKNOWN UNKNOWN
## 4 F      18-24      BLACK      UNKNOWN UNKNOWN UNKNOWN
## 5 M      45-64      BLACK      UNKNOWN UNKNOWN UNKNOWN
## 6 M      18-24      BLACK HISPANIC UNKNOWN UNKNOWN UNKNOWN
## # i 3 more variables: Count <int>, Probability <dbl>, Power <dbl>
```

This shows only 14 of the ~967 combinations have enough power to say there is this 80% chance. We wo

Look at filtered_data on myself

We saw earlier the model predicted that I (MALE, 34 years old, and WHITE) would most likely be shot by someone who was UNKNOWN or BLACK, similar age and race. However, we now can see how much power there is behind that prediction. And it appears not much. The probabilities aren't much different, but the power is less than 20%. This means that we can't say there is a 20% chance of detecting an effect if one truly exists here.

```
results_significantpower[results_significantpower$VIC_Sex=="M" & results_significantpower$VIC_Age_Group=="34-44"]
```

```
## # A tibble: 32 x 9
## # Rowwise:
##   VIC_Sex VIC_Age_Group VIC_Race_Group PERP_Sex PERP_Age_Group PERP_Race_Group
##   <chr>   <chr>           <chr>      <fct>   <fct>         <fct>
## 1 M      25-44       WHITE      UNKNOWN UNKNOWN        UNKNOWN
## 2 M      25-44       WHITE      M        <NA>          WHITE
## 3 M      25-44       WHITE      M        <NA>          BLACK
## 4 M      25-44       WHITE      M        <NA>          BLACK
## 5 M      25-44       WHITE      M        <NA>          WHITE HISPANIC
## 6 M      25-44       WHITE      M        <NA>          WHITE
## 7 M      25-44       WHITE      M        <NA>          BLACK HISPANIC
## 8 M      25-44       WHITE      M        <NA>          WHITE HISPANIC
## 9 M      25-44       WHITE      M        UNKNOWN        BLACK
## 10 M     25-44       WHITE      M        <NA>          WHITE
## # i 22 more rows
## # i 3 more variables: Count <int>, Probability <dbl>, Power <dbl>
```

Bias Identification/Conclusion

This dataset is strictly for New York City. So the data found here might not apply outside of this city in other parts of America or the world at large. Since I live in New York, but outside of New York City (source of this dataset), my **personal bias** might be that there isn't as many shootings. However, if I were to spend time in New York City, my **personal bias** might be more corrected by what I am able to observe, **mitigating** its overall effect on my analysis/conclusions.

From this dataset, there is a significant amount of UNKNOWN data. However, with the UNKNOWN data aside, it might appear that Black Males have the largest correlation with shooting. Therefore, there could be bias built into this dataset to conclude UNKNOWN or Black Males are the number one perpetrators in shooting cases. The power analysis I did showed this because only 14 of the ~967 combinations had a power level greater than 80% chance of detecting an effect if one truly exists. Thus, the models I made would carry the bias of the data into them as they did. We would need more data in the other missing groups to have more statistical power. It could be that another race/gender/age has been getting away with shooting or murder much easier. Therefore, garbage in, garbage out might apply here to the model's prediction.

When we look at the models fit to the data we observe for a victim of my age, gender, and race [34,M,White] that the perpetrator is most likely a Male, between 25-44, and of an UNKNOWN race. Excluding the UNKNOWN, the perpetrator is most likely a Male, between 25-44, and Black with less than a 7% chance of this combination. And interesting enough, at only a 6% chance, the same categories as myself would predict a perpetrator. This tells me that there is a positive correlation between the victim and the perpetrator in terms of their race, sex, and age. We further investigated the statistical power of these predictions of the model's by looking at the filtered data without the model and performing a statistical power test. We found that the probabilities aren't much different, but the statistical power is less than 20%. This means that we can't say there is a 20% chance of detecting an effect if one truly exists here.

In conclusion, all we know is that there is a significant amount of UNKNOWN shooters out there, and from what we do know, a large number of them appear to be Black Males from this dataset. Using a model based on this dataset that predicts the perpetrator given the victim's race, age, and gender likely will give very biased results as shown in this analysis because of this missing data and also because of the lack of diverse data. Hopefully with time, we can solve more of the mystery cases out there to fill in the missing data. With time, there might be more data that adds to statistical significance of potential models or analysis.