NYPD Shooting Data

SC

2022-10-29

Introduction

Using R Markdown, this study will examine each shooting in New York City from 2006 up until the end of 2021.

Each quarter, the New York Police Department (NYPD) website gets data entered manually by the Office of Management Analysis. Details of each shooting are given including the victims demographics, suspect demographics, and geographical locations of each shooting.

Import Libraries and Install Packages

In order to perform the analysis, libraries and packages need to be loaded in and installed. In order to move this analysis over to something readable like a pdf, tinytex will be installed. `#{r} #install.packages("tinytex") #tinytex::install_tinytex() #`

Next, the tidyverse and lubridate packages will be installed to help us parse through the data in a more user friendly way.

```
# install.packages("tidyverse")
library(tidyverse)
library(lubridate)
```

Load Data

Looking at the NYPD website, the data can be exported into a CSV file. The data can then be read using read_csv().

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

```
## Rows: 25596 Columns: 19

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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(data)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>       <dbl>             <dbl>
## 1    236168668 11/11/2021 15:04      BROOKLYN       79                 0
## 2    231008085 07/16/2021 22:05      BROOKLYN       72                 0
## 3    230717903 07/11/2021 01:09      BROOKLYN       79                 0
## 4    237712309 12/11/2021 13:42      BROOKLYN       81                 0
## 5    224465521 02/16/2021 20:00      QUEENS        113                 0
## 6    228252164 05/15/2021 04:13      QUEENS        113                 0
## # ... with 13 more variables: 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>
```

Clean and Transform Data

For the purposes of this study, certain information is not useful. These include Precinct, Juisdiction, X & Y Coordinates, and Longitude & Latitude. This can be done using the pipe operatore %>%.

```
data2 = data %>% select(INCIDENT_KEY,
                        OCCUR_DATE,
                        OCCUR_TIME,
                        BORO,
                        STATISTICAL_MURDER_FLAG,
                        PERP_AGE_GROUP,
                        PERP_SEX,
                        PERP_RACE,
                        VIC_AGE_GROUP,
                        VIC_SEX,
                        VIC_RACE)
#Find column(s) with confusing or missing values
lapply(data2, function(x) sum(is.na(x)))
```

```
## $INCIDENT_KEY
## [1] 0
##
## $OCCUR_DATE
## [1] 0
##
## $OCCUR_TIME
## [1] 0
##
## $BORO
## [1] 0
##
## $STATISTICAL_MURDER_FLAG
## [1] 0
##
## $PERP_AGE_GROUP
```

```
## [1] 9344
##
## $PERP_SEX
## [1] 9310
##
## $PERP_RACE
## [1] 9310
##
## $VIC_AGE_GROUP
## [1] 0
##
## $VIC_SEX
## [1] 0
##
## $VIC_RACE
## [1] 0
```

All of these data types are factors except for **INCIDENT_KEY**, which can be treated as a string.

Cleaning up the empty data spaces:

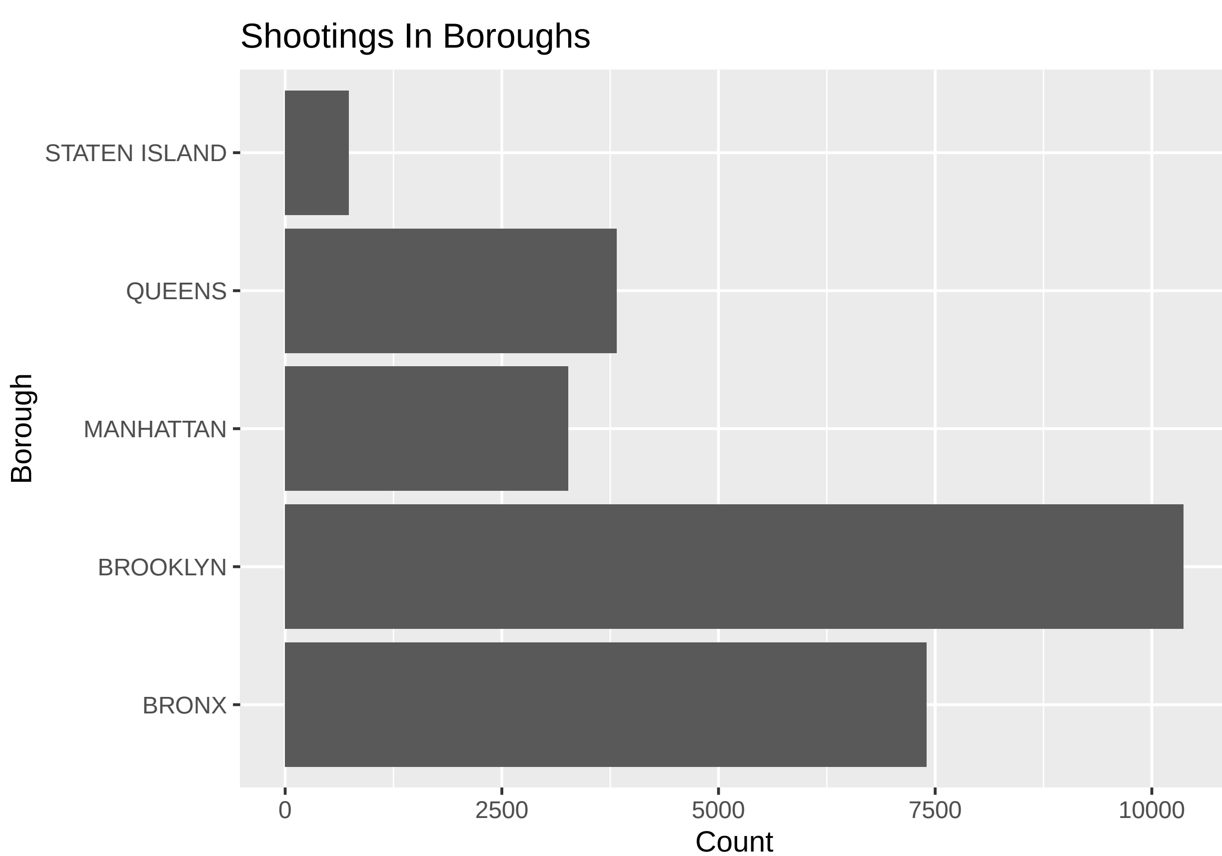
```
data2 = data2 %>%
  replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))

data2$PERP_AGE_GROUP = recode(data2$PERP_AGE_GROUP, UNKNOWN = "Unknown")
data2$PERP_SEX = recode(data2$PERP_SEX, U = "Unknown")
data2$PERP_RACE = recode(data2$PERP_RACE, UNKNOWN = "Unknown")
data2$VIC_SEX   = recode(data2$VIC_SEX, U = "Unknown")
data2$VIC_RACE   = recode(data2$VIC_RACE, UNKNOWN = "Unknown")
data2$BORO = as.factor(data2$BORO)
data2$PERP_AGE_GROUP = as.factor(data2$PERP_AGE_GROUP)
data2$PERP_SEX = as.factor(data2$PERP_SEX)
data2$PERP_RACE = as.factor(data2$PERP_RACE)
data2$VIC_AGE_GROUP = as.factor(data2$VIC_AGE_GROUP)
data2$VIC_SEX = as.factor(data2$VIC_SEX)
data2$VIC_RACE = as.factor(data2$VIC_RACE)

summary(data2)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                   BORO
##  Min.   :  9953245   Length:25596       Length:25596      BRONX        : 7402
##  1st Qu.: 61593633   Class :character   Class1:hms        BROOKLYN     :10365
##  Median : 86437258   Mode  :character   Class2:difftime   MANHATTAN    : 3265
##  Mean   :112382648                      Mode  :numeric    QUEENS       : 3828
##  3rd Qu.:166660833                                        STATEN ISLAND:  736
##  Max.   :238490103
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Mode :logical           Unknown:12492   F      :  371
##  FALSE:20668             18-24  : 5844   M      :14416
##  TRUE :4928              25-44  : 5202   Unknown:10809
##                          <18    : 1463
##                          45-64  :  535
##                          65+    :   57
```

```
##                          (Other):    3
##                           PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2681   F      : 2403
##  ASIAN / PACIFIC ISLANDER      :  141   18-24  : 9604   M      :23182
##  BLACK                         :10668   25-44  :11386   Unknown:   11
##  BLACK HISPANIC                : 1203   45-64  : 1698
##  Unknown                       :11146   65+    :  167
##  WHITE                         :  272   UNKNOWN:   60
##  WHITE HISPANIC                : 2164
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  354
##  BLACK                         :18281
##  BLACK HISPANIC                : 2485
##  Unknown                       :   65
##  WHITE                         :  660
##  WHITE HISPANIC                : 3742
```

Visualization and Analysis

```r
b <- ggplot(data2, aes(y = BORO)) + geom_bar() + labs(title = "Shootings In Boroughs",
                                                       x = "Count", y = "Borough")
b
```



Based on this chart, it is easy to see which Borough has seen the highest amount of shootings between 2006 and 2021. Let's do some further analysis with this information.

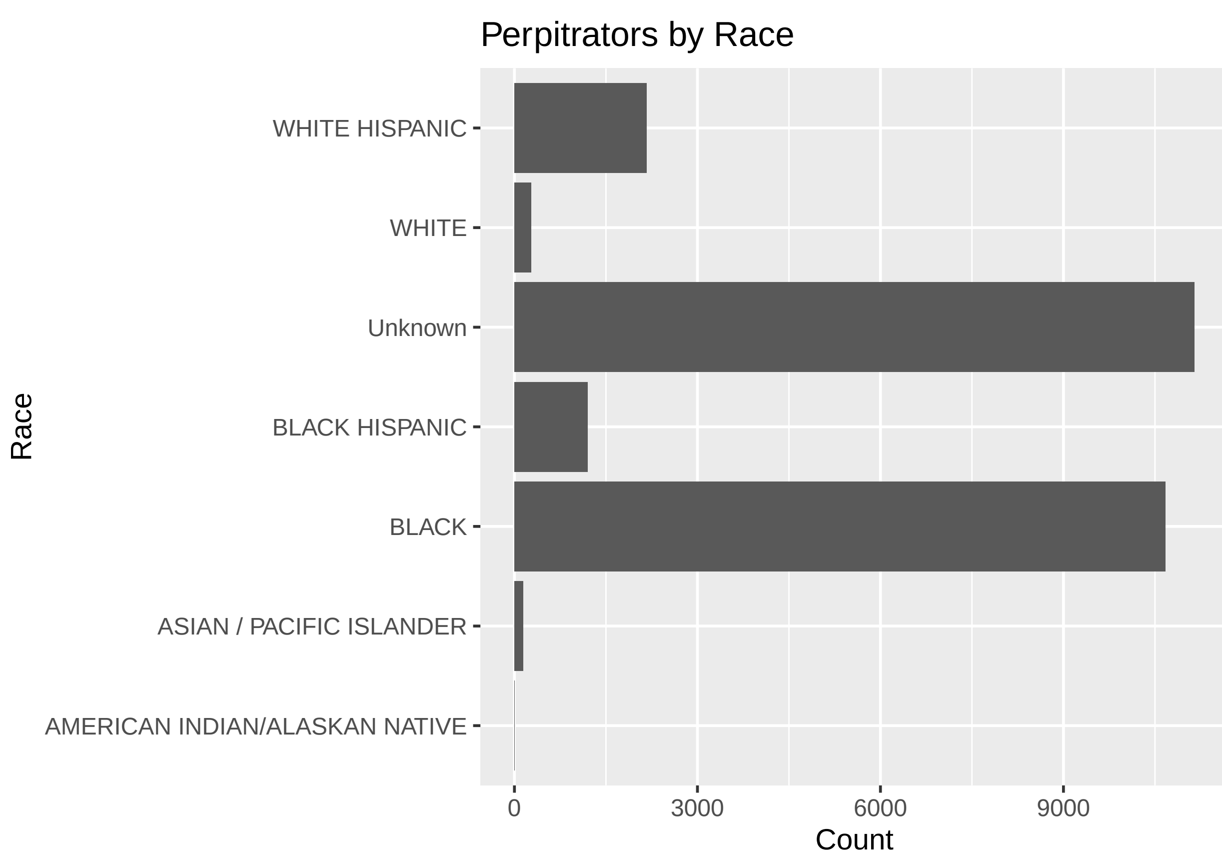
```
crime_num <- data2 %>%
  group_by(BORO) %>%
  count(name='crimes')
murder_num <- data2 %>%
  group_by(BORO) %>%
  summarize(murder = sum(STATISTICAL_MURDER_FLAG))
murder_data <- merge(crime_num, murder_num) %>%
  mutate(murder_rate = murder/crimes)
murder_data
```

```
##            BORO crimes murder murder_rate
## 1         BRONX   7402   1417   0.1914347
## 2      BROOKLYN  10365   2020   0.1948866
## 3     MANHATTAN   3265    574   0.1758040
## 4        QUEENS   3828    762   0.1990596
## 5 STATEN ISLAND    736    155   0.2105978
```

Above is a breakdown of the murder rates by each Boro in New York City.

Next, lets take a look at shooters by race and age.

```
r <- ggplot(data2, aes(y = PERP_RACE)) + geom_bar() + labs(title = "Perpitrators by Race",
                                                          x = "Count", y = "Race")
r
```



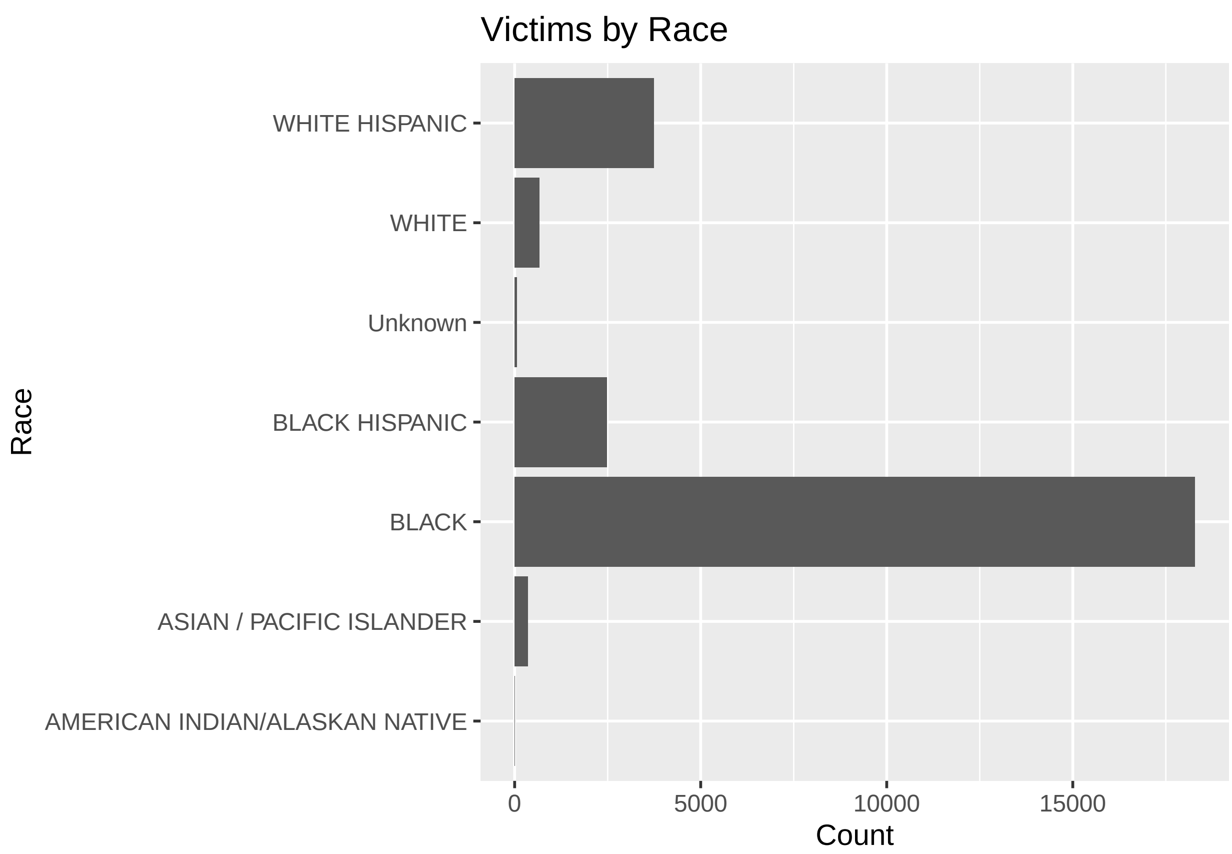
```r
crime_num <- data2 %>%
  group_by(PERP_RACE) %>%
  count(name='crimes')
murder_num <- data2 %>%
  group_by(PERP_RACE) %>%
  summarize(murder = sum(STATISTICAL_MURDER_FLAG))
murder_data <- merge(crime_num, murder_num) %>%
  mutate(murder_rate = murder/crimes)
murder_data
```

```
##                        PERP_RACE crimes murder murder_rate
## 1 AMERICAN INDIAN/ALASKAN NATIVE      2      0   0.0000000
## 2       ASIAN / PACIFIC ISLANDER    141     44   0.3120567
## 3                          BLACK  10668   2214   0.2075366
## 4                 BLACK HISPANIC   1203    230   0.1911887
## 5                        Unknown  11146   1809   0.1623004
## 6                          WHITE    272    108   0.3970588
## 7                 WHITE HISPANIC   2164    523   0.2416821
```

The table above shows murder rate by perpetrator race.

Now let's take a look at the victims.

```r
v <- ggplot(data2, aes(y = VIC_RACE)) + geom_bar() + labs(title = "Victims by Race",
                                                           x = "Count", y = "Race")
v
```

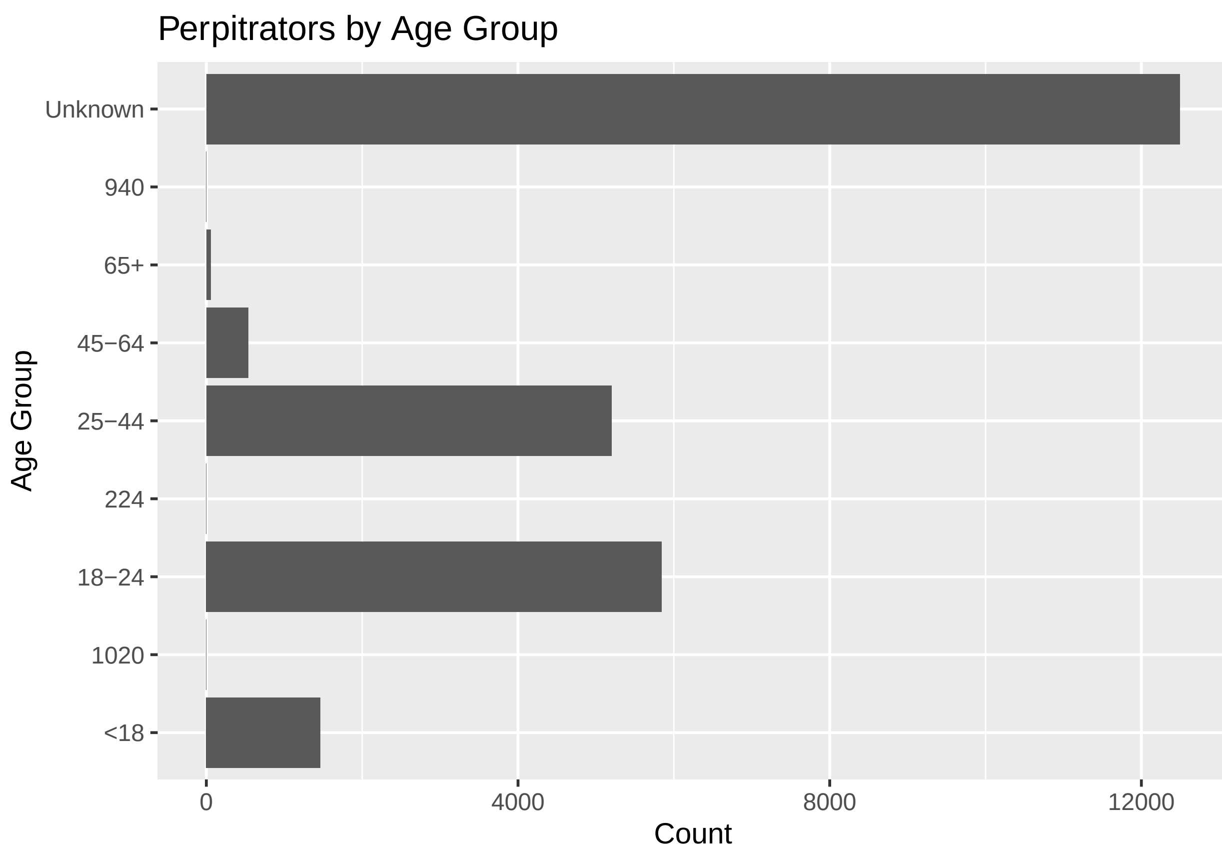


```r
crime_num <- data2 %>%
  group_by(VIC_RACE) %>%
  count(name='crimes')
murder_num <- data2 %>%
  group_by(VIC_RACE) %>%
  summarize(murder = sum(STATISTICAL_MURDER_FLAG))
murder_data <- merge(crime_num, murder_num) %>%
  mutate(murder_rate = murder/crimes)
murder_data
```

```
##                         VIC_RACE crimes murder murder_rate
## 1 AMERICAN INDIAN/ALASKAN NATIVE      9      0   0.0000000
## 2       ASIAN / PACIFIC ISLANDER    354     90   0.2542373
## 3                          BLACK  18281   3449   0.1886658
## 4                 BLACK HISPANIC   2485    404   0.1625755
## 5                        Unknown     65      7   0.1076923
## 6                          WHITE    660    186   0.2818182
## 7                 WHITE HISPANIC   3742    792   0.2116515
```

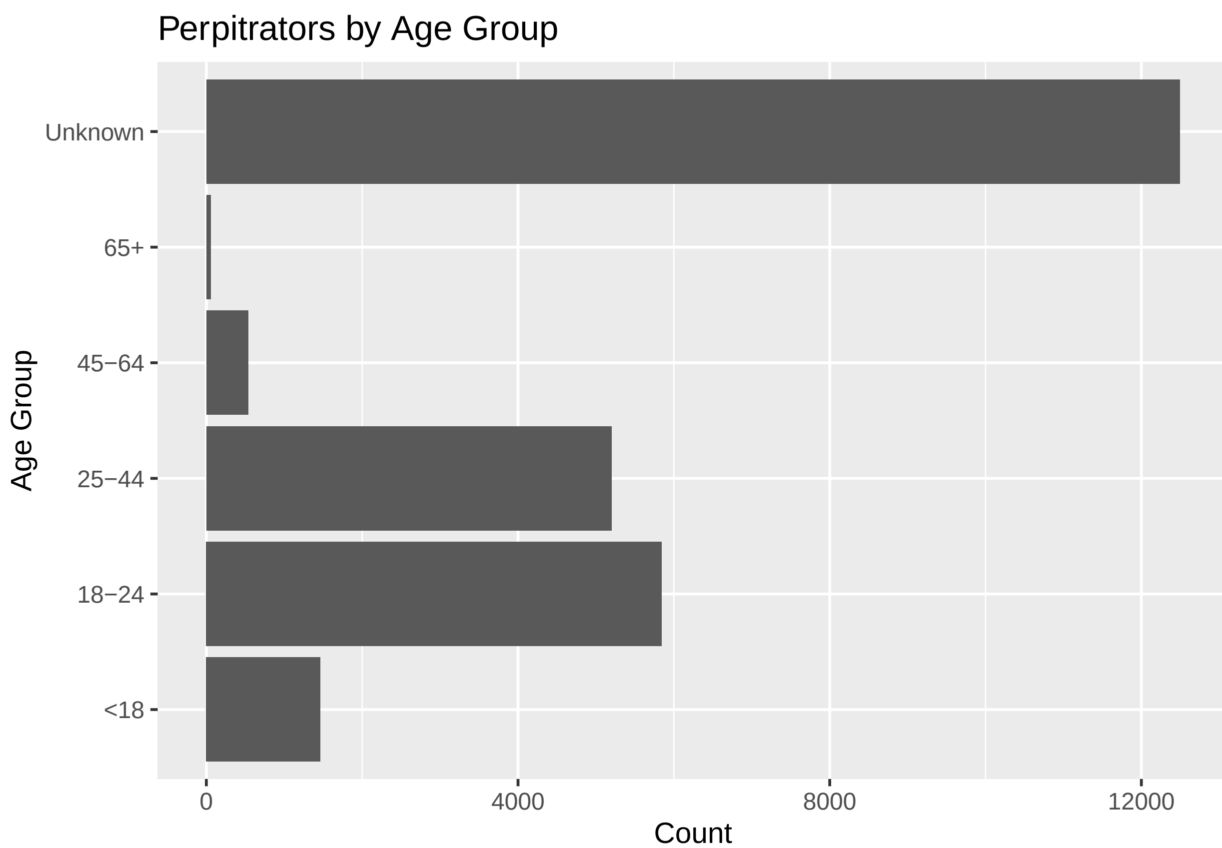
Looking at the graphs above, there is a breakdown of perpetrators and victims by race. You can also see a glaring issue in the first graph when attempting to do any kind of analysis on the people who committed the shootings: the shooters race wasn't able to be identified. While there was still a breakdown of murder rates by perpetrator and victim, it must be iterated again that there's an issue with not knowing the race of a large number of shooters. Let's examine if the same thing happens when breaking it down by age.

```r
age <- ggplot(data2, aes(y = PERP_AGE_GROUP)) + geom_bar() + labs(title = "Perpitrators by Age Group",
                                                                  x = "Count", y = "Age Group")
age
```



Looking at this graph, there was an obvious oversight when cleaning the data. On the y-axis "940, 224, 1020" mean nothing and need to be cleaned up.

```
data2 = subset(data2, PERP_AGE_GROUP!='1020' & PERP_AGE_GROUP != '224' & PERP_AGE_GROUP != '940')
age2 <- ggplot(data2, aes(y = PERP_AGE_GROUP)) + geom_bar() + labs(title = "Perpitrators by Age Group",
                                                                    x = "Count", y = "Age Group")
age2
```




```
crime_num <- data2 %>%
  group_by(PERP_AGE_GROUP) %>%
  count(name='crimes')
murder_num <- data2 %>%
  group_by(PERP_AGE_GROUP) %>%
  summarize(murder = sum(STATISTICAL_MURDER_FLAG))
murder_data <- merge(crime_num, murder_num) %>%
  mutate(murder_rate = murder/crimes)
murder_data
```

```
##   PERP_AGE_GROUP crimes murder murder_rate
## 1            <18   1463    266   0.1818182
## 2          18-24   5844   1221   0.2089322
## 3          25-44   5202   1414   0.2718185
## 4          45-64    535    188   0.3514019
## 5            65+     57     24   0.4210526
## 6        Unknown  12492   1815   0.1452930
```

Much better!

Looking at this graph, there is the same issue. There's a lack of detail about a large number of the shooters. While this data has been clean enough to visualize, there's certain pitfalls to be careful about. The same can be said about the perpetrator murder rate. It should also be noted that there is a very small number of people in the 65+ group.

Model

Let's examine the ways it can be predicted if a shooting incident is a murder case or not. In order to do this, the best tool that can be used is logistical regression. The variables I'll take a look into are: BORO, PERP_Race, PERP_AGE

```
glm.fit = glm(STATISTICAL_MURDER_FLAG ~ BORO + PERP_RACE + PERP_AGE_GROUP, family = binomial, data = da
summary(glm.fit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + PERP_RACE + PERP_AGE_GROUP,
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4832  -0.6807  -0.6005  -0.4493   2.4883
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -10.93640   84.09023  -0.130  0.89652
## BOROBROOKLYN                       0.08778    0.03989   2.201  0.02776 *
## BOROMANHATTAN                     -0.11076    0.05547  -1.997  0.04583 *
## BOROQUEENS                         0.04646    0.05134   0.905  0.36547
## BOROSTATEN ISLAND                 -0.03709    0.09762  -0.380  0.70400
## PERP_RACEASIAN / PACIFIC ISLANDER   9.83090   84.09040   0.117  0.90693
## PERP_RACEBLACK                     9.36455   84.09020   0.111  0.91133
## PERP_RACEBLACK HISPANIC            9.25732   84.09023   0.110  0.91234
## PERP_RACEUnknown                  10.59730   84.09026   0.126  0.89971
## PERP_RACEWHITE                    10.01804   84.09030   0.119  0.90517
## PERP_RACEWHITE HISPANIC            9.53103   84.09021   0.113  0.90976
## PERP_AGE_GROUP18-24                0.17168    0.07533   2.279  0.02266 *
## PERP_AGE_GROUP25-44                0.50509    0.07498   6.737 1.62e-11 ***
## PERP_AGE_GROUP45-64                0.82937    0.11464   7.235 4.67e-13 ***
## PERP_AGE_GROUP65+                  0.98782    0.28335   3.486  0.00049 ***
## PERP_AGE_GROUPUnknown             -1.37037    0.11629 -11.784  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25076  on 25592  degrees of freedom
## Residual deviance: 24343  on 25577  degrees of freedom
## AIC: 24375
##
## Number of Fisher Scoring iterations: 9
```

Pitfalls and Bias

As previously discussed, there's a number of issues with the data when trying to come to a conclusion about the shooters: a large number of shooter's race and age were not able to be identified. That means any takeaways anybody might have about "who commits the most shootings" has to include the massive caveat that there is a large portion of shooters that cannot be identified at all.

When looking at bias I may have had doing this analysis, I only looked for the total number of shootings committed by race and age, which is itself bias. I put a lot of emphasis on my analysis that there's a large number of shootings where the age and race of the shooter was not identified to mitigate this.

I also did not break this down per capita. This would be an excellent next point to study but we would need more population information about New York City.