NYPD Data

Venus Miskinyar

2023-12-01

```
# Install tidyverse package if not already installed
if (!requireNamespace("tidyverse", quietly = TRUE)) {
  install.packages("tidyverse")
}
```

NYPD Shooting Historic Incidents

The following is an analysis of NYPD Shooting incidents in recent history. The data was retrieved from City of New York: NYPD Shooting Incident Data (Historic)

Import Data

1. Set the url vector with the csv path and name:

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

2. Read the CSV file:

```
nypd_shooting_data <- read_csv(url[1])
```

```
## Rows: 27312 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 the data

1. View the raw data:

```
nypd_shooting_data
```

```
## # A tibble: 27,312 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
##  1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
##  2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
##  3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
##  4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
##  5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
##  6    219559682 10/21/2020 21:36      BROOKLYN <NA>                    81
##  7     85295722 06/17/2012 22:47      QUEENS   <NA>                   114
##  8     71662474 03/08/2010 19:41      BROOKLYN <NA>                    81
##  9     83002139 02/05/2012 05:45      QUEENS   <NA>                   105
## 10     86437261 08/26/2012 01:10      QUEENS   <NA>                   101
## # i 27,302 more rows
## # 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>
```

2. Remove data elements that pertain to exact locations or keys/codes that are specific to precincts etc. Also we will be focusing on victim data and therefore remove any perpetrator data:

```
nypd_shooting_data <- nypd_shooting_data[, !names(nypd_shooting_data) %in% c("INCIDENT_KEY", "OCCUR_TIM
nypd_shooting_data
```

```
## # A tibble: 27,312 x 6
##    OCCUR_DATE BORO     STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX VIC_RACE
##    <chr>      <chr>    <lgl>                   <chr>         <chr>   <chr>
##  1 05/27/2021 QUEENS   FALSE                   18-24         M       BLACK
##  2 06/27/2014 BRONX    FALSE                   18-24         M       BLACK
##  3 11/21/2015 QUEENS   TRUE                    25-44         M       WHITE
##  4 10/09/2015 BRONX    FALSE                   <18           M       WHITE HISP~
##  5 02/19/2009 BRONX    TRUE                    45-64         M       BLACK
##  6 10/21/2020 BROOKLYN TRUE                    25-44         M       BLACK
##  7 06/17/2012 QUEENS   FALSE                   25-44         M       BLACK
##  8 03/08/2010 BROOKLYN TRUE                    18-24         M       BLACK
##  9 02/05/2012 QUEENS   FALSE                   25-44         M       BLACK
## 10 08/26/2012 QUEENS   FALSE                   25-44         M       BLACK
## # i 27,302 more rows
```

3. Convert data type for OCCUR_DATE to Date Object:

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

```
## # A tibble: 27,312 x 6
##    OCCUR_DATE BORO     STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX VIC_RACE
##    <date>     <chr>    <lgl>                   <chr>         <chr>   <chr>
```

```
##  1 2021-05-27 QUEENS   FALSE                    18-24         M       BLACK
##  2 2014-06-27 BRONX    FALSE                    18-24         M       BLACK
##  3 2015-11-21 QUEENS   TRUE                     25-44         M       WHITE
##  4 2015-10-09 BRONX    FALSE                    <18           M       WHITE HISP~
##  5 2009-02-19 BRONX    TRUE                     45-64         M       BLACK
##  6 2020-10-21 BROOKLYN TRUE                     25-44         M       BLACK
##  7 2012-06-17 QUEENS   FALSE                    25-44         M       BLACK
##  8 2010-03-08 BROOKLYN TRUE                     18-24         M       BLACK
##  9 2012-02-05 QUEENS   FALSE                    25-44         M       BLACK
## 10 2012-08-26 QUEENS   FALSE                    25-44         M       BLACK
## # i 27,302 more rows
```

4. Show the minimum and maximum dates to determine the date range for the data

```
data_date_range <- nypd_shooting_data %>%
 summarize(min_date = min(OCCUR_DATE),
max_date = max(OCCUR_DATE))
data_date_range
```

```
## # A tibble: 1 x 2
##   min_date   max_date
##   <date>     <date>
## 1 2006-01-01 2022-12-31
```

5. Summarize data by year and total number of shooting incidents for each year sorted by number of shooting incidents

```
summarized_data <- nypd_shooting_data %>%
    filter(!is.na(OCCUR_DATE)) %>%
mutate(Year = lubridate::year(OCCUR_DATE)) %>%
group_by (Year) %>%
summarize(row_count = n(), .groups="drop")
summarized_data
```

```
## # A tibble: 17 x 2
##     Year row_count
##    <dbl>     <int>
##  1  2006      2055
##  2  2007      1887
##  3  2008      1959
##  4  2009      1828
##  5  2010      1912
##  6  2011      1939
##  7  2012      1717
##  8  2013      1339
##  9  2014      1464
## 10  2015      1434
## 11  2016      1208
## 12  2017       970
## 13  2018       958
## 14  2019       967
## 15  2020      1948
## 16  2021      2011
## 17  2022      1716
```

6. Show the minimum and maximum number of shootings and the year that those occurred on.

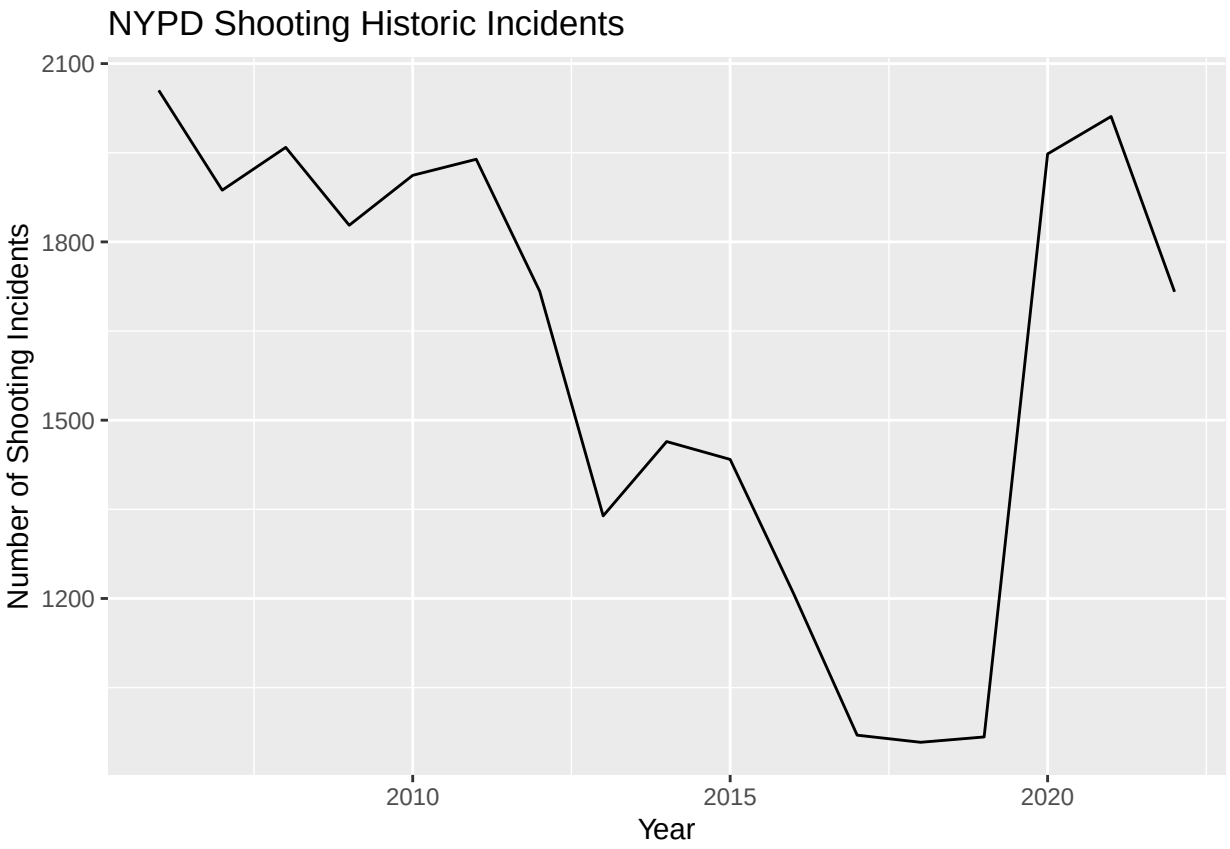
```
min_year <- summarized_data$Year[which.min(summarized_data$row_count)]
max_year <- summarized_data$Year[which.max(summarized_data$row_count)]

min_count <- min(summarized_data$row_count)
max_count <- max(summarized_data$row_count)

min_max_count_year <- data.frame("Shooting Incident Count" = c(min_count, max_count), 'Year' = c(min_yea
min_max_count_year
```

```
##   Shooting.Incident.Count Year
## 1                     958 2018
## 2                    2055 2006
```

Plot the summarized data

1. Create a line plot to visualize the data

```
x_axis = summarized_data$Year
y_axis = summarized_data$row_count

ggplot(summarized_data, aes(x = x_axis, y = y_axis)) +
    geom_line()  +
    labs(x = "Year", y = "Number of Shooting Incidents", title = "NYPD Shooting Historic Incidents")
```



Modeling the number of male and female shooting incidents

1. Summarize and display the number of male shootings and female shootings by year

```
victim_data_by_year <- nypd_shooting_data %>%
filter(!is.na(OCCUR_DATE)) %>%
mutate(Year = lubridate::year(OCCUR_DATE)) %>%
group_by (Year, VIC_SEX) %>%
summarize(row_count = n(), .groups="drop") %>%
arrange(desc(Year))
victim_data_by_year
```

```
## # A tibble: 39 x 3
##     Year VIC_SEX row_count
##    <dbl> <chr>       <int>
##  1  2022 F             212
##  2  2022 M            1504
##  3  2021 F             199
##  4  2021 M            1812
##  5  2020 F             201
##  6  2020 M            1747
##  7  2019 F             102
##  8  2019 M             865
##  9  2018 F              99
## 10  2018 M             857
## # i 29 more rows
```

2. Model the number of male and female shooting incidents

```
victim_data_by_year <- victim_data_by_year %>%
mutate(Male = ifelse(VIC_SEX == "M", 1, 0), Female = ifelse(VIC_SEX == "F", 1, 0))
model = lm(row_count ~ Male + Female, data=victim_data_by_year)
summary(model)
```

```
##
## Call:
## lm(formula = row_count ~ Male + Female, data = victim_data_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595.12  -49.32    1.80   55.03  420.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.2      106.9   0.021    0.984
## Male          1449.9      121.6  11.923 4.64e-14 ***
## Female         151.6      121.6   1.247    0.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 239 on 36 degrees of freedom
## Multiple R-squared:  0.8927, Adjusted R-squared:  0.8868
## F-statistic: 149.8 on 2 and 36 DF,  p-value: < 2.2e-16
```

The model summary indicates that model isn't well-fitted since the residuals and the standard error values are higher than expected.

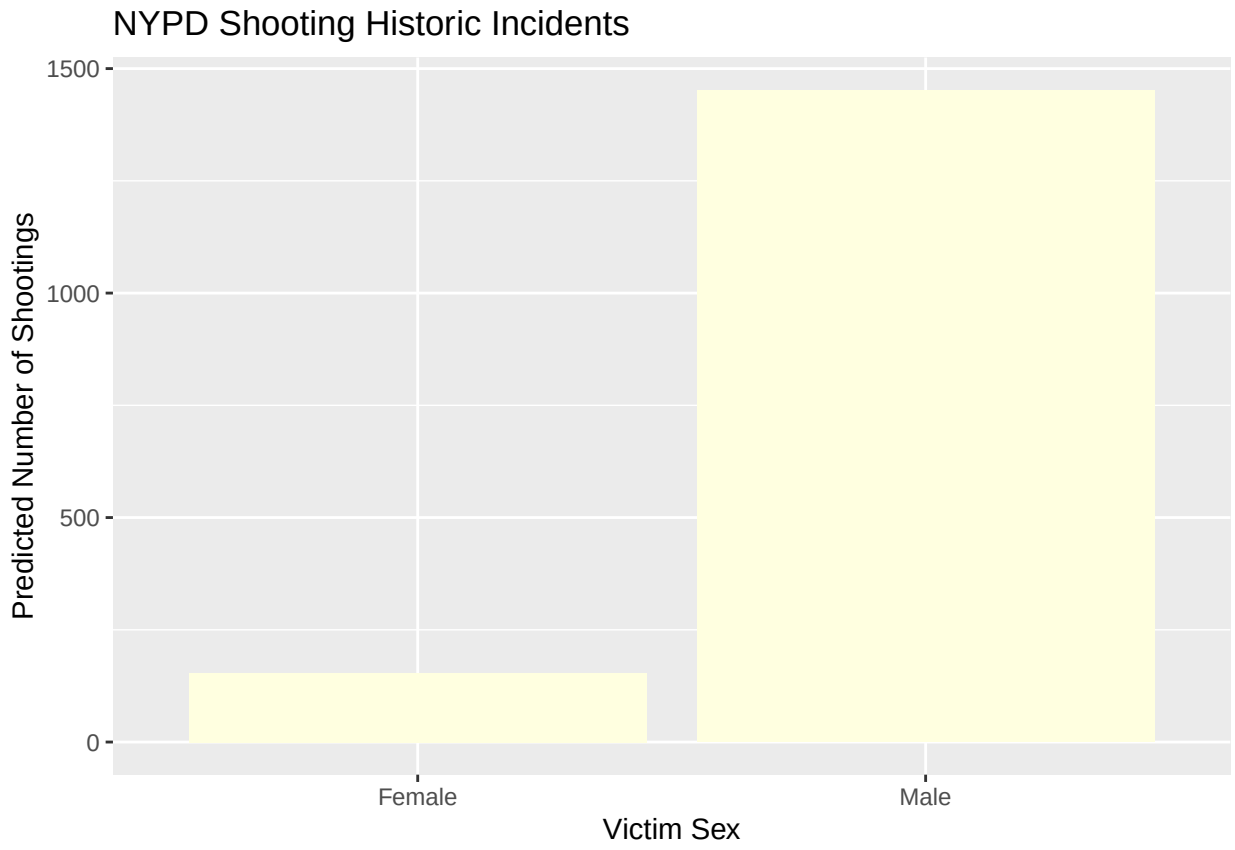
3. Create a prediction model for the number of male and female shooting incidents

```
male_data <- data.frame(Male = 1, Female = 0)
male_victim_prediction <- predict(model, newdata = male_data)

female_data <- data.frame(Male = 0, Female = 1)
female_victim_prediction <- predict(model, newdata = female_data)

predicted_counts <- data.frame(VIC_SEX = c('Male', 'Female'), predictedCounts = c(male_victim_predictio
predicted_counts
```

```
##   VIC_SEX predictedCounts
## 1    Male       1452.1176
## 2  Female        153.8235
```

4. Plot the prediction model

```
ggplot(predicted_counts, aes(x = VIC_SEX, y = predictedCounts, fill = VIC_SEX)) +
geom_bar(stat = "identity", fill = "lightyellow") +
labs(x = "Victim Sex", y = "Predicted Number of Shootings", title = "NYPD Shooting Historic Incidents")
```



High Level Data Analysis

1. Now we will focus on analyzing the data at the borough level and will not need any dates. The purpose of my analysis is to allow people, who are considering moving into a borough, to be able to gauge the level of violence. I understand that race plays a vital role in violence, but I want to remove victim race from my analysis and see if leaving race out, will impact the decision making process.

```
nypd_shooting_data <- nypd_shooting_data[, !names(nypd_shooting_data) %in% c("OCCUR_DATE","VIC_RACE")]
```

2. Filter any null values from our dataset

```
nypd_shooting_data <- nypd_shooting_data %>%
filter_all(all_vars(!is_null(.)))
nypd_shooting_data
```

```
## # A tibble: 27,312 x 4
##    BORO     STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX
##    <chr>    <lgl>                   <chr>         <chr>
##  1 QUEENS   FALSE                   18-24         M
##  2 BRONX    FALSE                   18-24         M
##  3 QUEENS   TRUE                    25-44         M
##  4 BRONX    FALSE                   <18           M
##  5 BRONX    TRUE                    45-64         M
##  6 BROOKLYN TRUE                    25-44         M
##  7 QUEENS   FALSE                   25-44         M
##  8 BROOKLYN TRUE                    18-24         M
##  9 QUEENS   FALSE                   25-44         M
## 10 QUEENS   FALSE                   25-44         M
## # i 27,302 more rows
```

3. View cleaned up shooting data:

```
nypd_shooting_data
```

```
## # A tibble: 27,312 x 4
##    BORO     STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX
##    <chr>    <lgl>                   <chr>         <chr>
##  1 QUEENS   FALSE                   18-24         M
##  2 BRONX    FALSE                   18-24         M
##  3 QUEENS   TRUE                    25-44         M
##  4 BRONX    FALSE                   <18           M
##  5 BRONX    TRUE                    45-64         M
##  6 BROOKLYN TRUE                    25-44         M
##  7 QUEENS   FALSE                   25-44         M
##  8 BROOKLYN TRUE                    18-24         M
##  9 QUEENS   FALSE                   25-44         M
## 10 QUEENS   FALSE                   25-44         M
## # i 27,302 more rows
```

4. Group data by Borough, Victim Age Group and Victim Sex. The idea of this analysis is to show the number of shootings for males and females in different age groups in various boroughs:

```
grouped_by_victim_data <- nypd_shooting_data %>% filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  group_by(BORO, VIC_AGE_GROUP, VIC_SEX) %>%
mutate(BORO_AGE_SEX = paste(BORO, VIC_AGE_GROUP, VIC_SEX, sep = "_")) %>%
group_by(BORO_AGE_SEX) %>%
summarise(count = n())
```

5. Sort data by highest statistical murder flag. We want to focus on the most violent boroughs based on victim counts:

```
sorted_victim_data <- grouped_by_victim_data %>%
arrange(desc(count))
```

6. View the sorted data:

```
sorted_victim_data
```

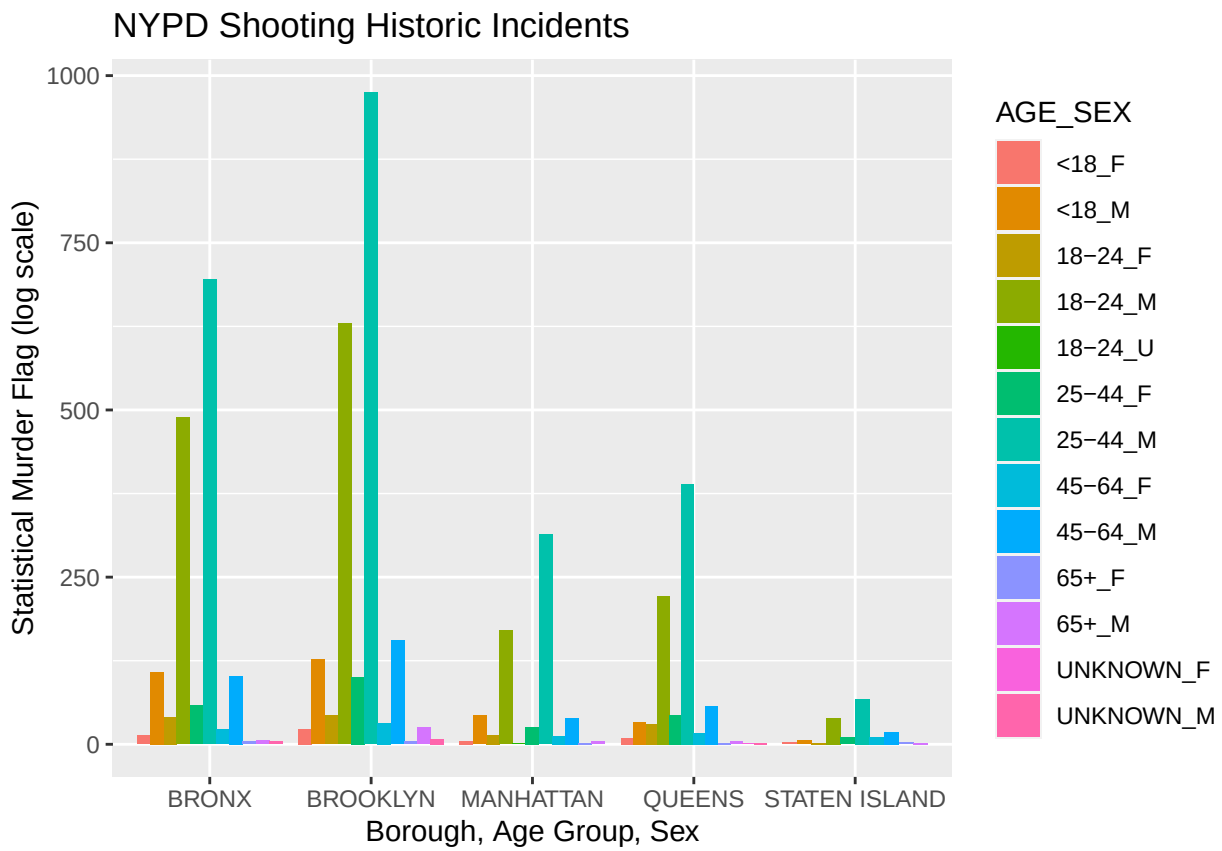
```
## # A tibble: 55 x 2
##    BORO_AGE_SEX      count
##    <chr>             <int>
##  1 BROOKLYN_25-44_M    975
##  2 BRONX_25-44_M       695
##  3 BROOKLYN_18-24_M    629
##  4 BRONX_18-24_M       489
##  5 QUEENS_25-44_M      389
##  6 MANHATTAN_25-44_M   314
##  7 QUEENS_18-24_M      222
##  8 MANHATTAN_18-24_M   170
##  9 BROOKLYN_45-64_M    155
## 10 BROOKLYN_<18_M      127
## # i 45 more rows
```

Plot NYPD Shooting Historic Incidents

1. We want to group the number of statistical murders by borough, age group, and sex.

```
grouped_victim_data <- nypd_shooting_data %>%
filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
group_by(BORO, VIC_AGE_GROUP, VIC_SEX) %>%
mutate(AGE_SEX = paste(VIC_AGE_GROUP, VIC_SEX, sep = "_")) %>%
group_by(BORO, AGE_SEX) %>%
summarise(count = n(), .groups="drop")
```

2. We want to plot the number of statistical murders by borough, age group, and sex



Bias in the data

The data in NYPD shooting statistics could be biased if the analysis is done to determine whether certain neighborhoods are safer than others. The reason is because the demographic data only includes race. It doesn't have socio-economic data, drug addiction rates, and incarcerations. All of these factors contribute to the likelihood of violence. Therefore omitting these factors and only including race could lead to some biased conclusions when doing analysis. To mitigate that, I have completely excluded the race columns so that we would not draw any conclusions based on race and its impact on borough violence.