NYPD Shooting Incident Data Report

5/05/2023

List of 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. Please refer to NYPD Shooting Incident Data (Historic) - CKAN for additional information about this dataset.

Step 0: Import Library

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

Step 1: Load Data

- `read_csv()` reads comma delimited files, read_csv2() reads semicolon separated files (common in countries where , is used as the decimal place), read_tsv() reads tab delimited files, and read_delim() reads in files with any delimiter.

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

```
## 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.
```

```
head(df)
```

```
## # A tibble: 6 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
## # 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>
```

Step 2: Tidy and Transform Data

Let's first eliminate the columns I do not need for this assignment, which are: **PRECINCT**,**JURISDICTION_CODE**,**LO**
X_COORD_CD, **Y_COORD_CD**, and **Lon_Lat**.

```
df_2 = df %>% 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,
                   Latitude,
                   Longitude)

# Return the column name along with the missing values
lapply(df_2, 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
##
## $Latitude
## [1] 10
##
## $Longitude
## [1] 10
```

Understanding the reasons why data are missing is important for handling the remaining data correctly. There's a fair amount of unidentifiable data on perpetrators (age, race, or sex.) Those cases are possibly still active and ongoing investigation. In fear of missing meaningful information, I handle this group of missing data by calling them as another group of "Unknown".

Key observations on data type conversion are:

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

```r
# Tidy and transform data
df_2 = df_2 %>%
  replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))

# Remove extreme values in data
df_2 = subset(df_2, PERP_AGE_GROUP!="1020" & PERP_AGE_GROUP!="224" & PERP_AGE_GROUP!="940")

df_2$PERP_AGE_GROUP = recode(df_2$PERP_AGE_GROUP, UNKNOWN = "Unknown")
df_2$PERP_SEX = recode(df_2$PERP_SEX, U = "Unknown")
df_2$PERP_RACE = recode(df_2$PERP_RACE, UNKNOWN = "Unknown")
df_2$VIC_SEX   = recode(df_2$VIC_SEX, U = "Unknown")
df_2$VIC_RACE   = recode(df_2$VIC_RACE, UNKNOWN = "Unknown")
df_2$INCIDENT_KEY = as.character(df_2$INCIDENT_KEY)
df_2$BORO = as.factor(df_2$BORO)
df_2$PERP_AGE_GROUP = as.factor(df_2$PERP_AGE_GROUP)
df_2$PERP_SEX = as.factor(df_2$PERP_SEX)
df_2$PERP_RACE = as.factor(df_2$PERP_RACE)
```

```r
df_2$VIC_AGE_GROUP = as.factor(df_2$VIC_AGE_GROUP)
df_2$VIC_SEX = as.factor(df_2$VIC_SEX)
df_2$VIC_RACE = as.factor(df_2$VIC_RACE)

# Return summary statistics
summary(df_2)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                   BORO     
##  Length:27309       Length:27309       Length:27309      BRONX        : 7935  
##  Class :character   Class :character   Class1:hms        BROOKLYN     :10932  
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    : 3572  
##                                        Mode  :numeric    QUEENS       : 4094  
##                                                          STATEN ISLAND:  776  
##                                                                               
##                                                                               
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX              PERP_RACE    
##  Mode :logical           (null) :  640   (null) :  640   BLACK         :11431  
##  FALSE:22043             <18    : 1591   F      :  424   Unknown       :11146  
##  TRUE :5266              18-24  : 6222   M      :15436   WHITE HISPANIC: 2339  
##                          25-44  : 5687   Unknown:10809   BLACK HISPANIC: 1314  
##                          45-64  :  617                   (null)        :  640  
##                          65+    :   60                   WHITE         :  283  
##                          Unknown:12492                   (Other)       :  156  
##  VIC_AGE_GROUP      VIC_SEX                                VIC_RACE    
##  <18    : 2839   F      : 2615   AMERICAN INDIAN/ALASKAN NATIVE:   10  
##  1022   :    1   M      :24683   ASIAN / PACIFIC ISLANDER      :  404  
##  18-24  :10085   Unknown:   11   BLACK                         :19438  
##  25-44  :12279                   BLACK HISPANIC                : 2646  
##  45-64  : 1863                   Unknown                       :   66  
##  65+    :  181                   WHITE                         :  698  
##  UNKNOWN:   61                   WHITE HISPANIC                : 4047  
##     Latitude       Longitude     
##  Min.   :40.51   Min.   :-74.25  
##  1st Qu.:40.67   1st Qu.:-73.94  
##  Median :40.70   Median :-73.92  
##  Mean   :40.74   Mean   :-73.91  
##  3rd Qu.:40.82   3rd Qu.:-73.88  
##  Max.   :40.91   Max.   :-73.70  
##  NA's   :10      NA's   :10      
```

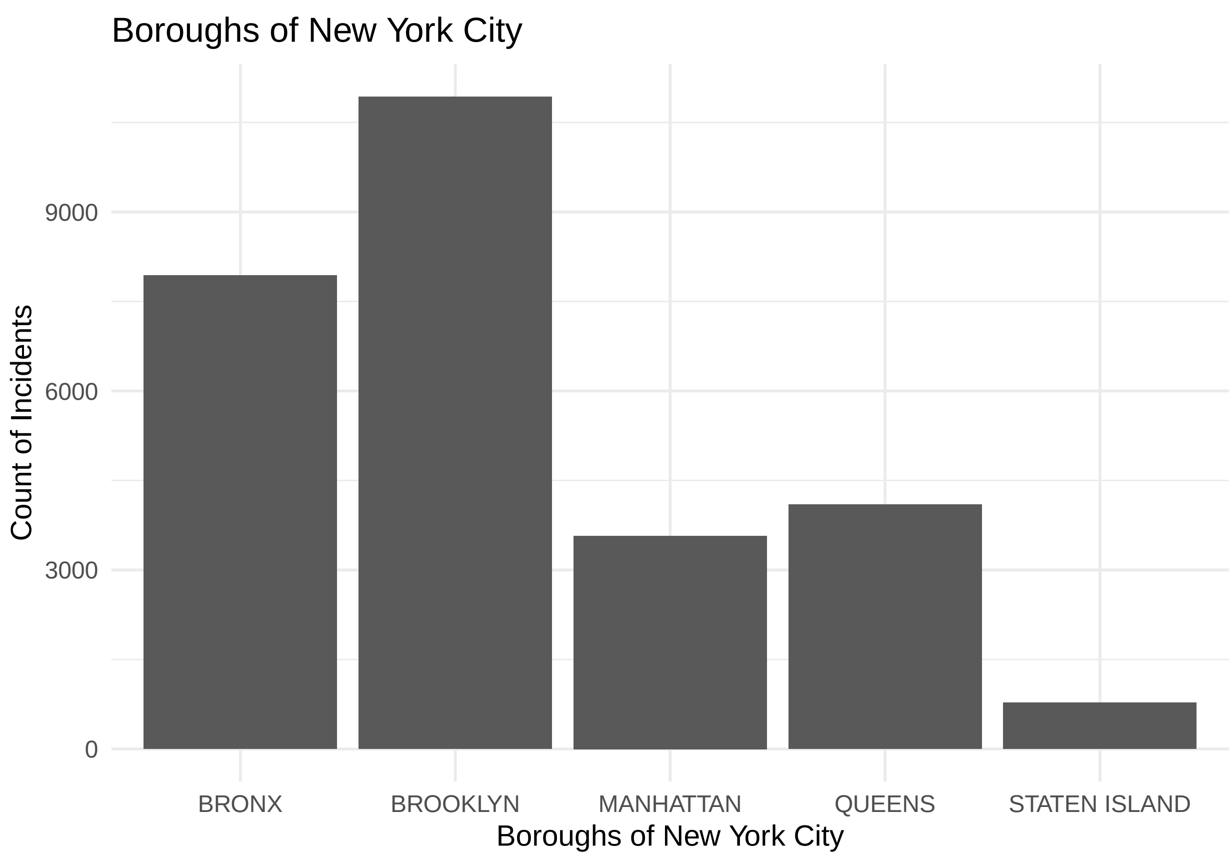
Step 3: Add Visualizations and Analysis

Research Question

1. Which part of New York has the most number of incidents? Of those incidents, how many are murder cases?

Brooklyn is the 1st in terms of the number of incidents, followed by Bronx and Queens respectively. Likewise, the number of murder cases follows the same pattern as that of incidents.

```r
g <- ggplot(df_2, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Boroughs of New York City",
       x = "Boroughs of New York City",
       y = "Count of Incidents") +
  theme_minimal()
g
```




```r
table(df_2$BORO, df_2$STATISTICAL_MURDER_FLAG)
```

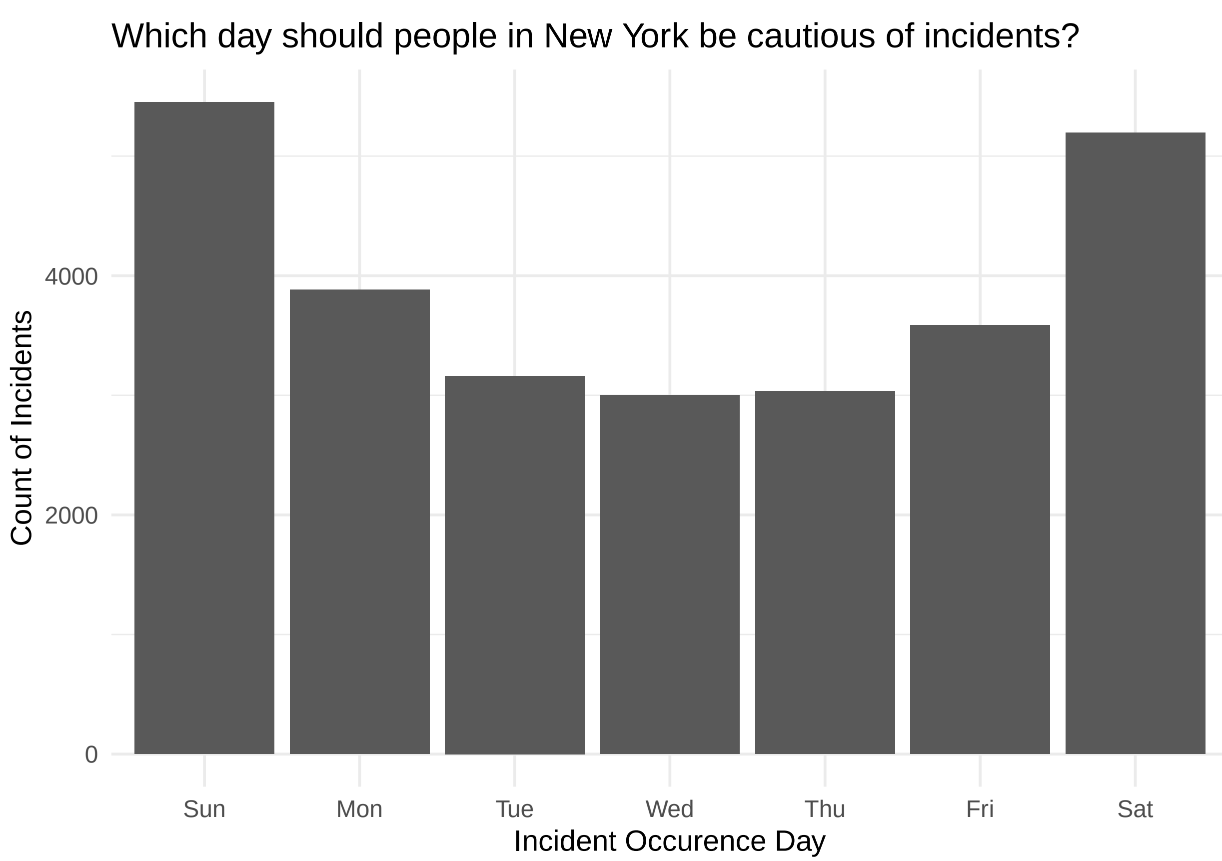
```
##
##                 FALSE TRUE
##   BRONX          6393 1542
##   BROOKLYN       8810 2122
##   MANHATTAN      2942  630
##   QUEENS         3284  810
##   STATEN ISLAND   614  162
```

2. Which day and time should people in New York be cautious of falling into victims of crime?

- Weekends in NYC have the most chances of incidents. Be cautious!
- Incidents historically happen in the evening and night time. If there's nothing urgent, recommend people staying at home!

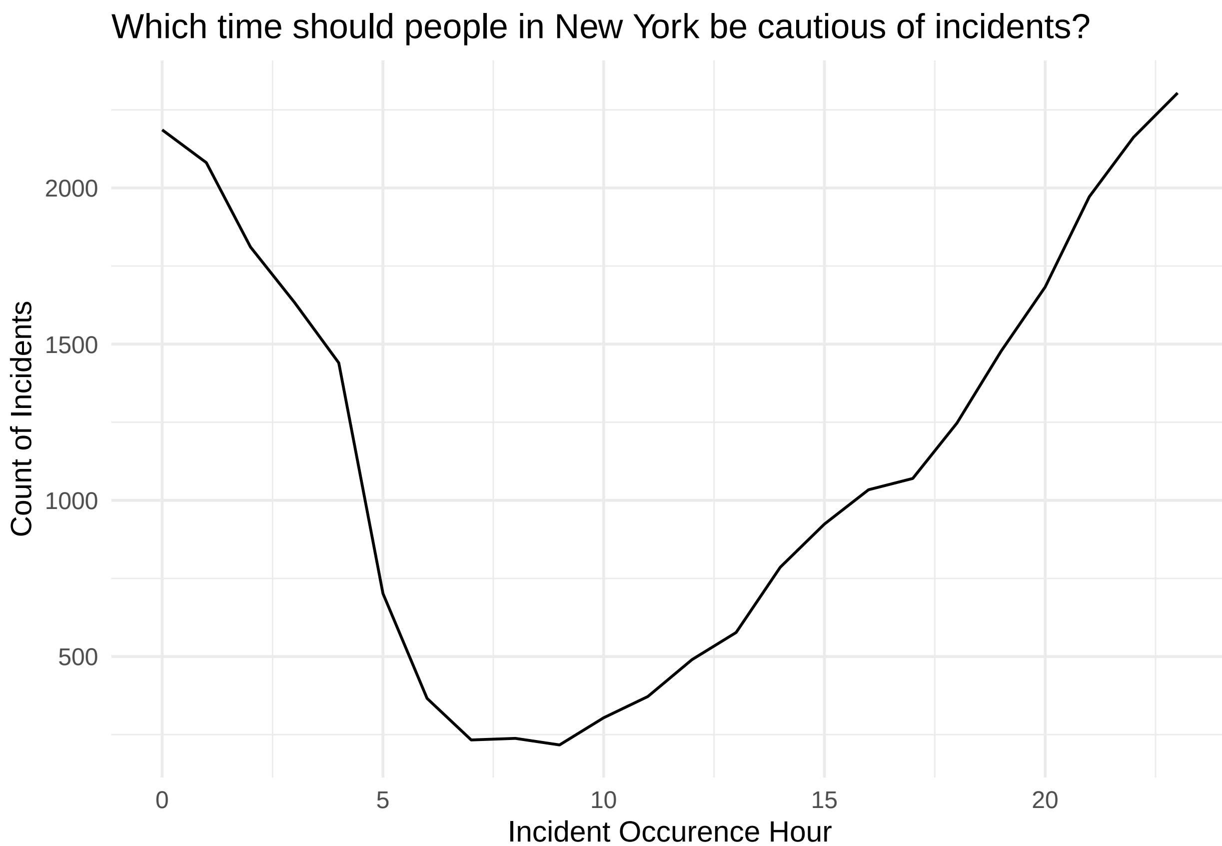
```r
df_2$OCCUR_DAY = mdy(df_2$OCCUR_DATE)
df_2$OCCUR_DAY = wday(df_2$OCCUR_DAY, label = TRUE)
df_2$OCCUR_HOUR = hour(hms(as.character(df_2$OCCUR_TIME)))

df_3 = df_2 %>%
  group_by(OCCUR_DAY) %>%
  count()

df_4 = df_2 %>%
  group_by(OCCUR_HOUR) %>%
  count()
```

```r
g <- ggplot(df_3, aes(x = OCCUR_DAY, y = n)) +
  geom_col() +
  labs(title = "Which day should people in New York be cautious of incidents?",
       x = "Incident Occurence Day",
       y = "Count of Incidents") +
  theme_minimal()
g
```




```r
g <- ggplot(df_4, aes(x = OCCUR_HOUR, y = n)) +
  geom_line() +
  labs(title = "Which time should people in New York be cautious of incidents?",
       x = "Incident Occurence Hour",
       y = "Count of Incidents") +
```

```
  theme_minimal()
g
```



3. The Profile of Perpetrators and Victims

- There's a striking number of incidents in the age group of 25-44 and 18-24.
- Black and White Hispanic stood out in the number of incidents in Boroughs of New York City.
- There are significantly more incidents with Male than those of Female.

```
table(df_2$PERP_AGE_GROUP, df_2$VIC_AGE_GROUP)
```

```
##
##            <18 1022 18-24 25-44 45-64  65+ UNKNOWN
##   (null)    57    0   181   340    57    5       0
##   <18      484    0   621   397    77   10       2
##   18-24    788    1  2758  2294   329   40      12
##   25-44    262    0  1516  3352   479   43      35
##   45-64     20    0    76   327   177   12       5
##   65+        0    0     1    25    23   11       0
##   Unknown 1228    0  4932  5544   721   60       7
```

```
table(df_2$PERP_SEX, df_2$VIC_SEX)
```

```
##
##              F     M Unknown
##   (null)     72   568       0
##   F          72   351       1
##   M        1666 13764       6
##   Unknown   805 10000       4
```

```
table(df_2$PERP_RACE, df_2$VIC_RACE)
```

```
##
##                                  AMERICAN INDIAN/ALASKAN NATIVE
##   (null)                                                      1
##   AMERICAN INDIAN/ALASKAN NATIVE                              0
##   ASIAN / PACIFIC ISLANDER                                    0
##   BLACK                                                       4
##   BLACK HISPANIC                                              0
##   Unknown                                                     5
##   WHITE                                                       0
##   WHITE HISPANIC                                              0
##
##                                  ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   (null)                                               15   446             58
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     2              0
##   ASIAN / PACIFIC ISLANDER                             52    53             13
##   BLACK                                               157  9058            803
##   BLACK HISPANIC                                       18   531            344
##   Unknown                                             113  8523            999
##   WHITE                                                13    37             23
##   WHITE HISPANIC                                       36   788            406
##
##                                  Unknown WHITE WHITE HISPANIC
##   (null)                               0    12            108
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0    12             24
##   BLACK                               25   197           1187
##   BLACK HISPANIC                       5    36            380
##   Unknown                             24   187           1295
##   WHITE                                1   157             52
##   WHITE HISPANIC                      11    97           1001
```

4. Building logistic regression model to predict if the incident is likely a murder case or not?

Logistic regression is an instance of classification technique that you can use to predict a qualitative response. I will use logistic regression models to estimate the probability that a murder case belongs to a particular profile, location, or date & time.

The output shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. **PERP_SEXUnknown**, **PERP_AGE_GROUP45-64**, **PERP_AGE_GROUP65+**, **PERP_AGE_GROUPUnknown**, and **PERP_AGE_GROUP25-44** are statistically significant, as are the **latitude** and **longitude**. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

- The person in the age group of 65+, versus a person whose age < 18, changes the log odds of murder by 1.03.

```r
# Logistics Regression
glm.fit <- glm(STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + OCCUR_HOUR + OCCUR_DAY
summary(glm.fit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     PERP_AGE_GROUP + OCCUR_HOUR + OCCUR_DAY + Latitude + Longitude,
##     family = binomial, data = df_2)
##
## Coefficients: (2 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         49.825253  19.849788   2.510 0.012069
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -8.915610  84.241402  -0.106 0.915714
## PERP_RACEASIAN / PACIFIC ISLANDER     1.027692   0.295457   3.478 0.000505
## PERP_RACEBLACK                        0.583282   0.236967   2.461 0.013838
## PERP_RACEBLACK HISPANIC               0.500464   0.246258   2.032 0.042125
## PERP_RACEUnknown                      0.114060   0.114303   0.998 0.318340
## PERP_RACEWHITE                        1.192839   0.268215   4.447  8.7e-06
## PERP_RACEWHITE HISPANIC               0.732434   0.241341   3.035 0.002406
## PERP_SEXF                            -2.459168   0.264949  -9.282  < 2e-16
## PERP_SEXM                            -2.615159   0.239331 -10.927  < 2e-16
## PERP_SEXUnknown                             NA         NA      NA       NA
## PERP_AGE_GROUP<18                     2.232264   0.170345  13.104  < 2e-16
## PERP_AGE_GROUP18-24                   2.413127   0.160286  15.055  < 2e-16
## PERP_AGE_GROUP25-44                   2.726829   0.160268  17.014  < 2e-16
## PERP_AGE_GROUP45-64                   3.091787   0.179314  17.242  < 2e-16
## PERP_AGE_GROUP65+                     3.243423   0.310185  10.456  < 2e-16
## PERP_AGE_GROUPUnknown                       NA         NA      NA       NA
## OCCUR_HOUR                           -0.002167   0.001916  -1.131 0.257959
## OCCUR_DAY.L                          -0.040648   0.038500  -1.056 0.291074
## OCCUR_DAY.Q                          -0.079104   0.041301  -1.915 0.055455
## OCCUR_DAY.C                          -0.058826   0.041569  -1.415 0.157029
## OCCUR_DAY^4                          -0.012408   0.042343  -0.293 0.769489
## OCCUR_DAY^5                           0.017122   0.044427   0.385 0.699941
## OCCUR_DAY^6                          -0.075924   0.045700  -1.661 0.096645
## Latitude                             -0.383301   0.183827  -2.085 0.037058
## Longitude                             0.485996   0.234079   2.076 0.037875
##
## (Intercept)                          *
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER    ***
## PERP_RACEBLACK                       *
## PERP_RACEBLACK HISPANIC              *
## PERP_RACEUnknown
## PERP_RACEWHITE                       ***
## PERP_RACEWHITE HISPANIC              **
## PERP_SEXF                            ***
## PERP_SEXM                            ***
## PERP_SEXUnknown
## PERP_AGE_GROUP<18                    ***
## PERP_AGE_GROUP18-24                  ***
## PERP_AGE_GROUP25-44                  ***
```

```
## PERP_AGE_GROUP45-64                    ***
## PERP_AGE_GROUP65+                      ***
## PERP_AGE_GROUPUnknown
## OCCUR_HOUR
## OCCUR_DAY.L
## OCCUR_DAY.Q                            .
## OCCUR_DAY.C
## OCCUR_DAY^4
## OCCUR_DAY^5
## OCCUR_DAY^6                            .
## Latitude                               *
## Longitude                              *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26775  on 27298  degrees of freedom
## Residual deviance: 25831  on 27275  degrees of freedom
##   (10 observations deleted due to missingness)
## AIC: 25879
##
## Number of Fisher Scoring iterations: 9
```

Step 4: Identify Bias

In this topic, it can spur discrimination and implicit bias unbeknownst among individuals. If I based my judgement on prior experience after living near New York City for a while, I would personally believe that Bronx must have had the most number of incidents. I might make an assumption that the incidents are more likely to occur with women than those of men. However, I must validate all the conviction with data, so I can make a better, well-informed decision. It's intriguing to find out that Brooklyn is the 1st in terms of the number of incidents, followed by Bronx and Queens respectively. Likewise, the number of murder cases follows the same pattern as that of incidents. In addition, there are significantly more incidents with Male than those of Female. It's best to test and validate the assumption in a data-driven way rather than believing in your experience it all, which may be seriously wrong and biased towards a certain group and population. My finding is consistent with CNN's report on "Hate crimes, shooting incidents in New York City have surged since last year", especially that "shooting incidents in NYC increase by 73% for May 2021 vs. May 2020."

Additional Resources

- NYPD Shooting Incident Data (Historic) - CKAN
- NYC, Chicago see another wave of weekend gun violence
- Hate crimes, shooting incidents in New York City have surged since last year, NYPD data show - CNN