NYPD Shooting Incident Data Report

6/10/2021

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")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
```

```
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

```
head(df)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>            <dbl>              <dbl>
## 1    201575314 08/23/2019 22:10      QUEENS             103                  0
## 2    205748546 11/27/2019 15:54      BRONX               40                  0
## 3    193118596 02/02/2019 19:40      MANHATTAN           23                  0
## 4    204192600 10/24/2019 00:52      STATEN ISLAND      121                  0
## 5    201483468 08/22/2019 18:03      BRONX               46                  0
## 6    198255460 06/07/2019 17:50      BROOKLYN            73                  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>
```

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] 8459
##
## $PERP_SEX
## [1] 8425
##
## $PERP_RACE
## [1] 8425
##
## $VIC_AGE_GROUP
## [1] 0
##
## $VIC_SEX
## [1] 0
##
## $VIC_RACE
## [1] 0
##
## $Latitude
## [1] 0
##
## $Longitude
## [1] 0
```

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.

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

```r
# 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)
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:23565       Length:23565       Length:23565      BRONX        :6698
##  Class :character   Class :character   Class1:hms        BROOKLYN     :9721
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    :2921
##                                        Mode  :numeric    QUEENS       :3527
##                                                          STATEN ISLAND: 698
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Mode :logical           <18    : 1354   F       :  334
##  FALSE:19077             18-24  : 5448   M       :13302
##  TRUE :4488              25-44  : 4613   Unknown: 9929
##                          45-64  :  481
##                          65+    :   54
##                          Unknown:11615
##
##                            PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2525   F       : 2195
##  ASIAN / PACIFIC ISLANDER      :  120   18-24  : 8999   M       :21350
##  BLACK                         : 9854   25-44  :10285   Unknown:   20
##  BLACK HISPANIC                : 1081   45-64  : 1536
##  Unknown                       :10294   65+    :  155
##  WHITE                         :  255   UNKNOWN:   65
##  WHITE HISPANIC                : 1959
##                              VIC_RACE        Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   :40.51   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      :  320   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :16845   Median :40.70   Median :-73.92
##  BLACK HISPANIC                : 2244   Mean   :40.74   Mean   :-73.91
##  Unknown                       :  102   3rd Qu.:40.82   3rd Qu.:-73.88
##  WHITE                         :  615   Max.   :40.91   Max.   :-73.70
##  WHITE HISPANIC                : 3430
```

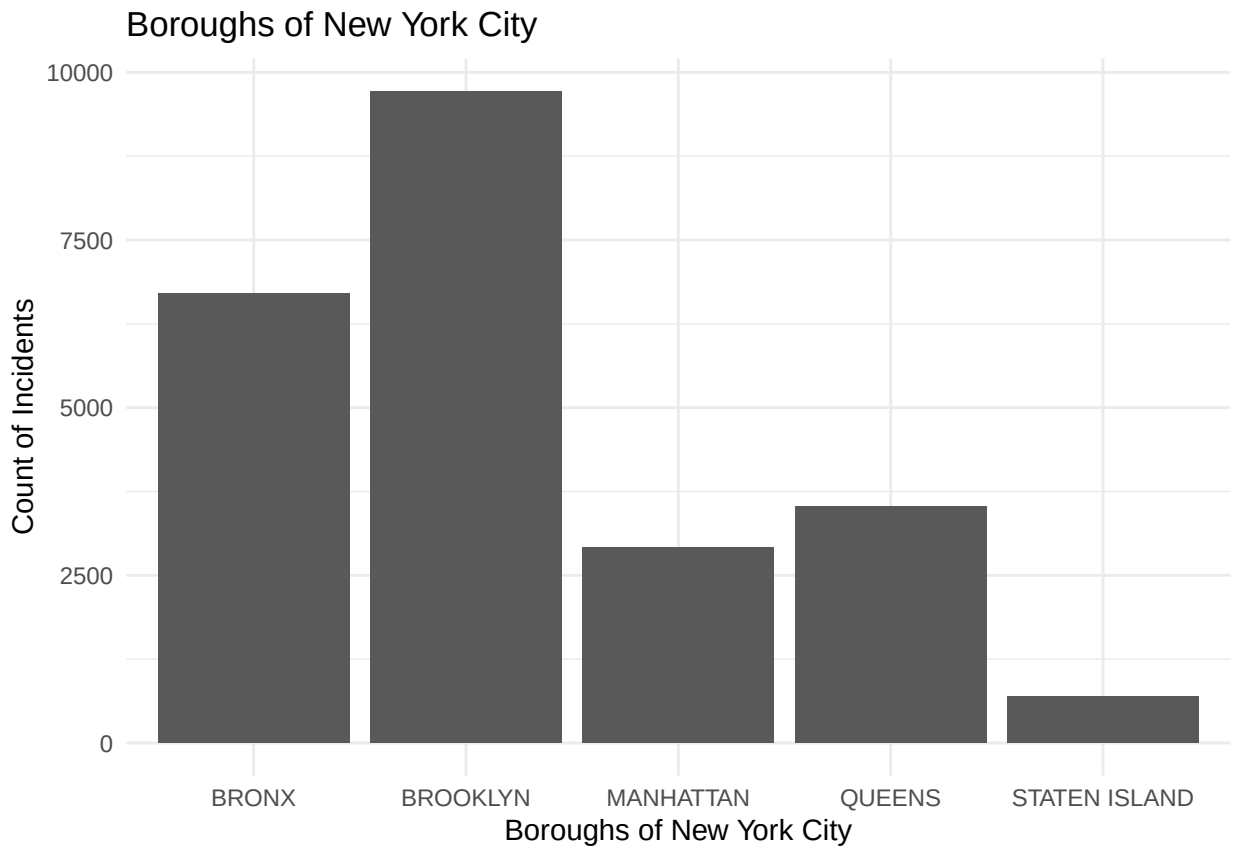
Step 3: Add Visualizations and Analysis

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.

```
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
```




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

```
##
##                FALSE TRUE
##   BRONX         5454 1244
##   BROOKLYN      7829 1892
```

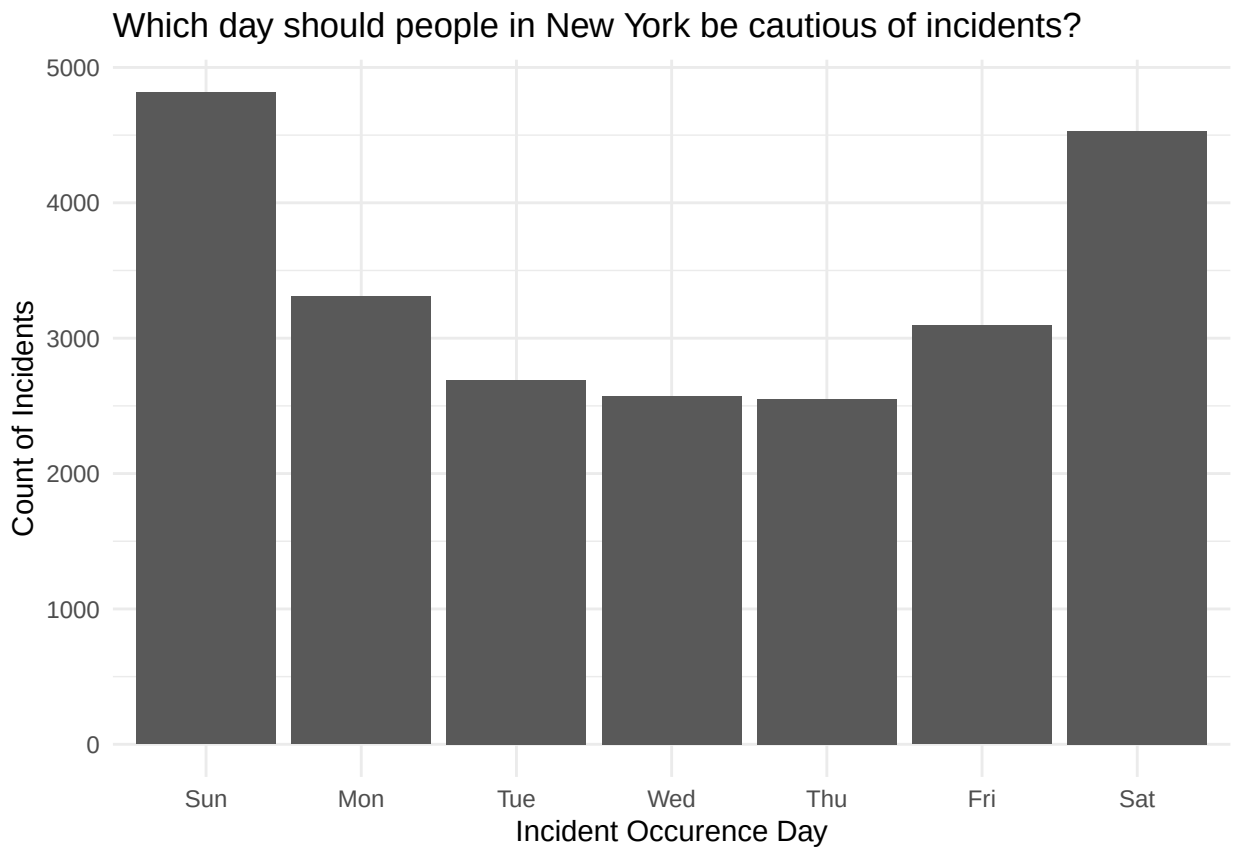
```
##   MANHATTAN      2409  512
##   QUEENS         2830  697
##   STATEN ISLAND   555  143
```

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

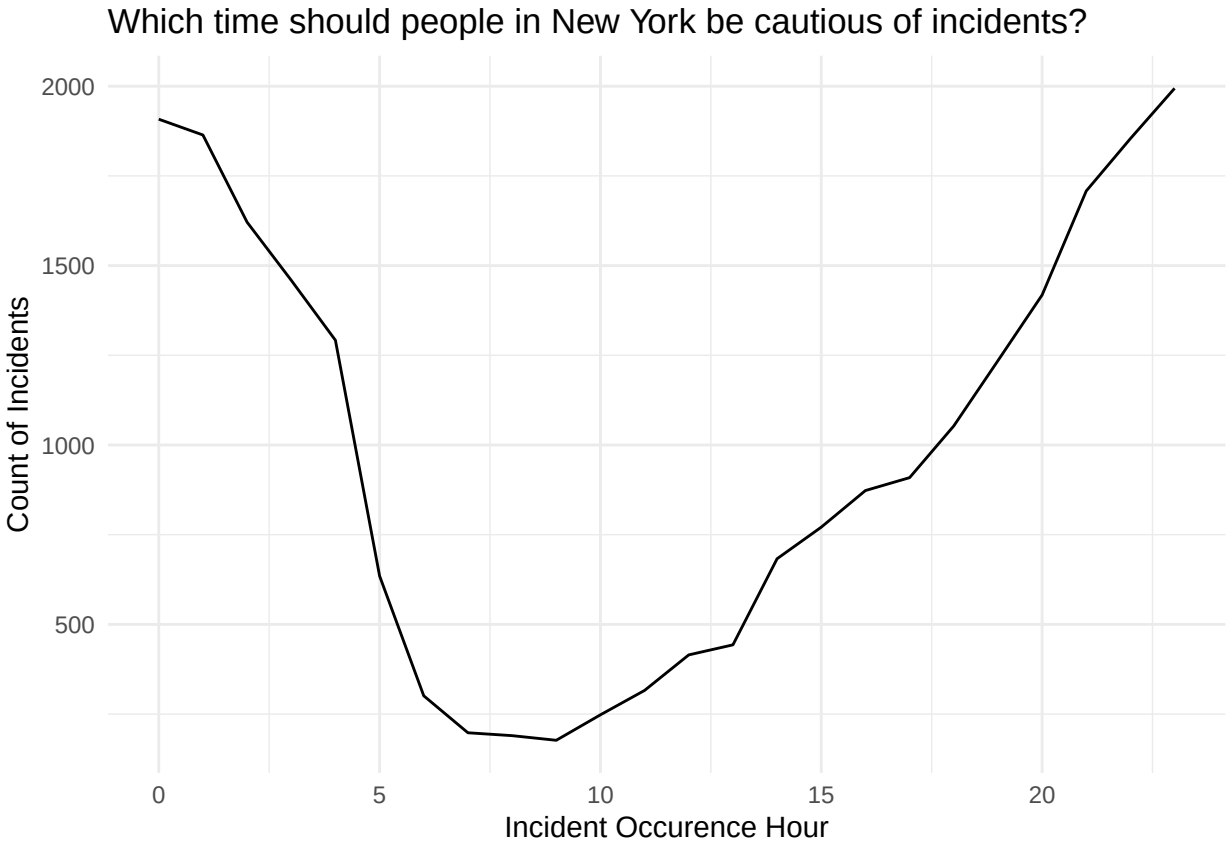
- Weekends in NYC have the most chances of incidents. Be cautious!
- Incidents historically happen in the evening and night time. If there's nothing urgent, recommend people staying at home!

```
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()
```

```
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
```




```
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 18-24 25-44 45-64  65+ UNKNOWN
##   <18       410   548   324    62    8       2
##   18-24     712  2447  1959   283   34      13
##   25-44     232  1291  2632   386   39      33
##   45-64      18    58   255   133   10       7
##   65+         0     1    22    21   10       0
##   Unknown  1153  4654  5093   651   54      10
```

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

```
##
##                F     M Unknown
##   F           49   284       1
##   M         1414 11878      10
##   Unknown    732  9188       9
```

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

```
##
##                                  AMERICAN INDIAN/ALASKAN NATIVE
##   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
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     2              0
##   ASIAN / PACIFIC ISLANDER                             38    37             12
##   BLACK                                               124  7825            676
##   BLACK HISPANIC                                       17   444            276
##   Unknown                                              99  7878            912
##   WHITE                                                11    29             18
##   WHITE HISPANIC                                       31   630            350
##
##                                  Unknown WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             2    11             20
##   BLACK                               34   160           1031
##   BLACK HISPANIC                       6    31            307
##   Unknown                             46   179           1175
##   WHITE                                1   151             45
##   WHITE HISPANIC                      13    83            852
```

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.

```
# 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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9895  -0.6692  -0.6156  -0.2267   2.9730
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     46.6487856 86.8848399   0.537   0.5913
## PERP_RACEASIAN / PACIFIC ISLANDER  9.9583265 84.2371629   0.118   0.9059
## PERP_RACEBLACK                     9.4739726 84.2369224   0.112   0.9105
## PERP_RACEBLACK HISPANIC            9.3665415 84.2369569   0.111   0.9115
## PERP_RACEUnknown                   8.8306675 84.2371713   0.105   0.9165
## PERP_RACEWHITE                    10.1798523 84.2370262   0.121   0.9038
## PERP_RACEWHITE HISPANIC            9.6533960 84.2369353   0.115   0.9088
## PERP_SEXM                         -0.1624763  0.1294760  -1.255   0.2095
## PERP_SEXUnknown                    2.6324936  0.2724963   9.661  < 2e-16 ***
## PERP_AGE_GROUP18-24                0.1507956  0.0788415   1.913   0.0558 .
## PERP_AGE_GROUP25-44                0.4889669  0.0788390   6.202 5.57e-10 ***
## PERP_AGE_GROUP45-64                0.8269393  0.1207340   6.849 7.42e-12 ***
## PERP_AGE_GROUP65+                  1.0304833  0.2910766   3.540   0.0004 ***
## PERP_AGE_GROUPUnknown             -2.1879192  0.1705836 -12.826  < 2e-16 ***
## OCCUR_HOUR                        -0.0028675  0.0020679  -1.387   0.1655
## OCCUR_DAY.L                       -0.0501244  0.0415798  -1.205   0.2280
## OCCUR_DAY.Q                       -0.1178332  0.0449146  -2.623   0.0087 **
## OCCUR_DAY.C                       -0.0459558  0.0449878  -1.022   0.3070
## OCCUR_DAY^4                       -0.0466743  0.0459089  -1.017   0.3093
## OCCUR_DAY^5                       -0.0008623  0.0481348  -0.018   0.9857
## OCCUR_DAY^6                       -0.0410963  0.0497679  -0.826   0.4089
## Latitude                          -0.4202647  0.1988285  -2.114   0.0345 *
## Longitude                          0.5459242  0.2506784   2.178   0.0294 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22947  on 23564  degrees of freedom
## Residual deviance: 22095  on 23542  degrees of freedom
## AIC: 22141
##
## 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