CL Project W3: NYPD Shooting Incident Report

Catherine Lebastard

9/11/2021

About the data

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 the attached data footnotes for additional information about this dataset.

Data Source: https://catalog.data.gov/dataset

Step 1 - Import the project dataset

Imports the shooting project dataset in a reproducible manner.

1. Install the library tidyverse or load it

```
if (!require(tidyverse)) install.packages("tidyverse");
library(tidyverse)
library(lubridate)
```

2. Read the NYPD dataset

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

```
## Rows: 23568 Columns: 19

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): OCCUR_DATE, OCCUR_TIME, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_...
## dbl  (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## lgl  (1): STATISTICAL_MURDER_FLAG

##
## 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.
```

Step 2 - Tidy and Transform your data

Add a summary of the data and clean up the dataset by changing appropriate variables to factor and data types and getting rid of any columns not needed.

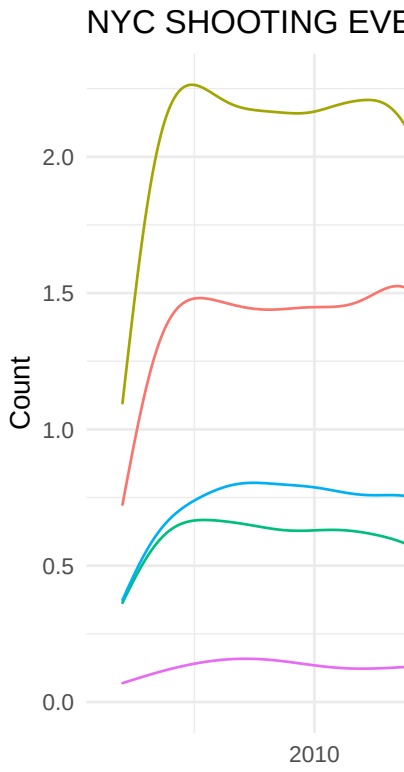
```
shooting_NY <- shooting_NY %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(BORO = fct_recode(BORO)) %>%
  mutate(PRECINCT = factor(PRECINCT)) %>%
  mutate(JURISDICTION_CODE = factor(JURISDICTION_CODE)) %>%
  mutate(PERP_AGE_GROUP = factor(PERP_AGE_GROUP)) %>%
  mutate(PERP_SEX = fct_recode(PERP_SEX)) %>%
  mutate(PERP_RACE = fct_recode(PERP_RACE)) %>%
  mutate(VIC_AGE_GROUP = fct_recode(VIC_AGE_GROUP)) %>%
  mutate(VIC_SEX = fct_recode(VIC_SEX)) %>%
  mutate(VIC_RACE = fct_recode(VIC_RACE)) %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Lon_Lat))

shooting_NY$PERP_RACE[shooting_NY$PERP_RACE == 'UNKNOWN'] <- NA
summary(shooting_NY)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:23568
##  1st Qu.: 55317014   1st Qu.:2008-12-30   Class :character
##  Median : 83365370   Median :2012-02-26   Mode  :character
##  Mean   :102218616   Mean   :2012-10-03
##  3rd Qu.:150772442   3rd Qu.:2016-02-28
##  Max.   :222473262   Max.   :2020-12-31
##
##             BORO         PRECINCT     JURISDICTION_CODE LOCATION_DESC
##  BRONX        :6700   75     : 1367   0   :19624        Length:23568
##  BROOKLYN     :9722   73     : 1282   1   :   54        Class :character
##  MANHATTAN    :2921   67     : 1102   2   : 3888        Mode  :character
##  QUEENS       :3527   79     :  920   NA's:    2
##  STATEN ISLAND: 698   44     :  842
##                       47     :  815
##                       (Other):17240
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  Mode :logical           18-24  :5448   F   :  334
##  FALSE:19080             25-44  :4613   M   :13305
##  TRUE :4488              UNKNOWN:3156   U   : 1504
##                          <18    :1354   NA's: 8425
##                          45-64  : 481
##                          (Other):  57
##                          NA's   :8459
##                      PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK                    : 9855   <18    : 2525   F: 2195
##  WHITE HISPANIC           : 1961   18-24  : 9000   M:21353
##  BLACK HISPANIC           : 1081   25-44  :10287   U:   20
##  WHITE                    :  255   45-64  : 1536
##  ASIAN / PACIFIC ISLANDER:  120   65+    :  155
##  (Other)                  :    2   UNKNOWN:   65
##  NA's                     :10294
##                            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                         :16846   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                : 3432
```

Step 3 - Visualizations and Analysis

Add at least two different visualizations and some analysis



1. Question1: Which boroughs are more unsafe than others in NYC from 2006 to 2020?
   With this visualization, we can answer that Brooklyn is the borough with the highest number of shootings over the years.

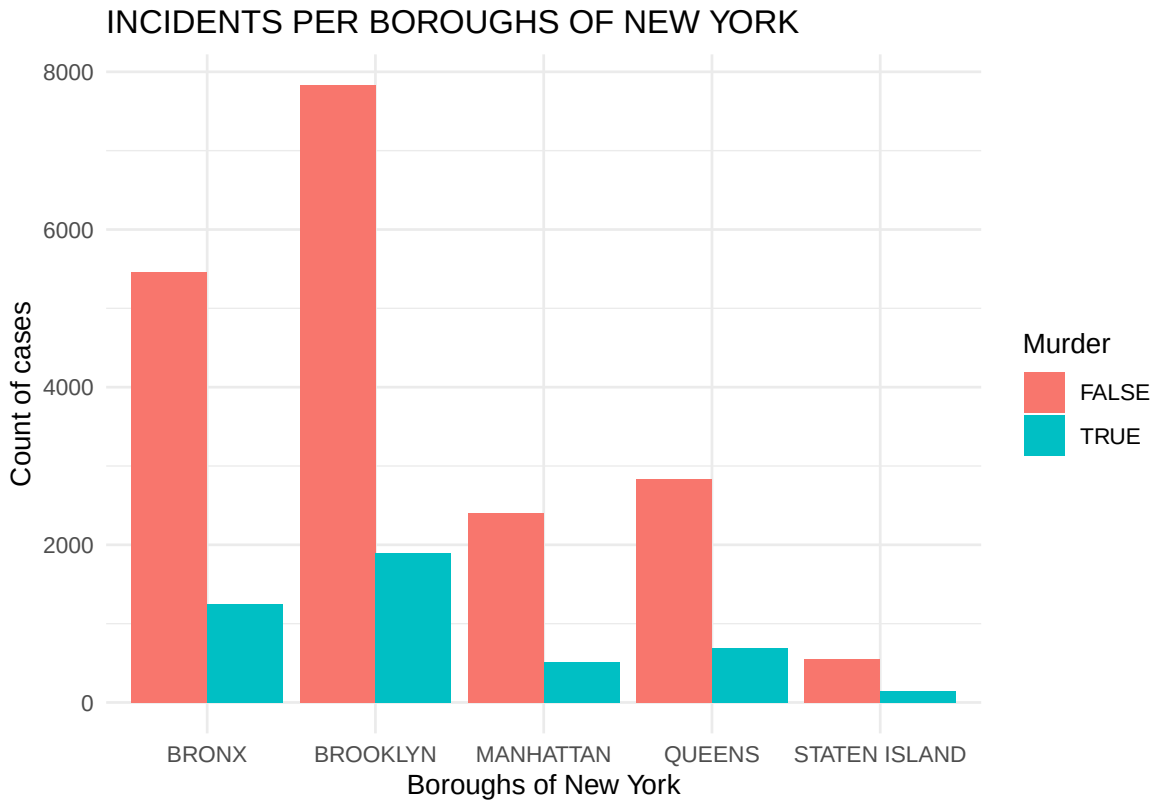
2. Question 2: Which boroughs of New York has the most number of shootings? From those shootings, how many are murder cases?

```
#Data preparation
shootings_NY_per_boro <- shooting_NY %>% group_by(BORO) %>% summarize(cases = n())
murders_NY_per_boro <- merge(shooting_NY %>% group_by(BORO, STATISTICAL_MURDER_FLAG) %>% summarize(
                             shootings_NY_per_boro, by='BORO', all.x = TRUE)
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the `.groups` argument.
```

```r
murders_NY_per_boro <- murders_NY_per_boro %>% rename(cases = cases.x, total_cases = cases.y)
murders_NY_per_boro <- murders_NY_per_boro %>% mutate(pct = round(cases / total_cases * 100, 2))

murders_NY_per_boro
```

```
##             BORO STATISTICAL_MURDER_FLAG cases total_cases   pct
## 1          BRONX                  FALSE  5456        6700 81.43
## 2          BRONX                   TRUE  1244        6700 18.57
## 3       BROOKLYN                  FALSE  7830        9722 80.54
## 4       BROOKLYN                   TRUE  1892        9722 19.46
## 5      MANHATTAN                  FALSE  2409        2921 82.47
## 6      MANHATTAN                   TRUE   512        2921 17.53
## 7         QUEENS                  FALSE  2830        3527 80.24
## 8         QUEENS                   TRUE   697        3527 19.76
## 9  STATEN ISLAND                  FALSE   555         698 79.51
## 10 STATEN ISLAND                   TRUE   143         698 20.49
```



With this prepared data and the visualization, we can answer that Brooklyn has the most number of shootings. It has 1,892 murder cases.

3. Analysis

```r
#Perpetrator per race
shootings_NY_per_perp_race <- shooting_NY %>% group_by(PERP_RACE) %>% summarize(cases = n())
shootings_NY_perp_race_vic_race <- merge(shooting_NY %>%
                                           group_by(PERP_RACE, VIC_RACE) %>%
                                           summarize(cases = n()),
                                         shootings_NY_per_perp_race, by='PERP_RACE', all.x = TRUE)
```

```
## `summarise()` has grouped output by 'PERP_RACE'. You can override using the `.groups` argument.
```

```
shootings_NY_perp_race_vic_race <- shootings_NY_perp_race_vic_race %>% rename(cases = cases.x, tota
shootings_NY_perp_race_vic_race <- shootings_NY_perp_race_vic_race %>% mutate(pct = round(cases / t
#Perpetrator per sex
shootings_NY_per_perp_sex <- shooting_NY %>% group_by(PERP_SEX) %>% summarize(cases = n())
shootings_NY_perp_race_vic_sex <- merge(shooting_NY %>%
                                          group_by(PERP_SEX, VIC_SEX) %>%
                                          summarize(cases = n()),
                                        shootings_NY_per_perp_sex, by='PERP_SEX', all.x = TRUE)
```

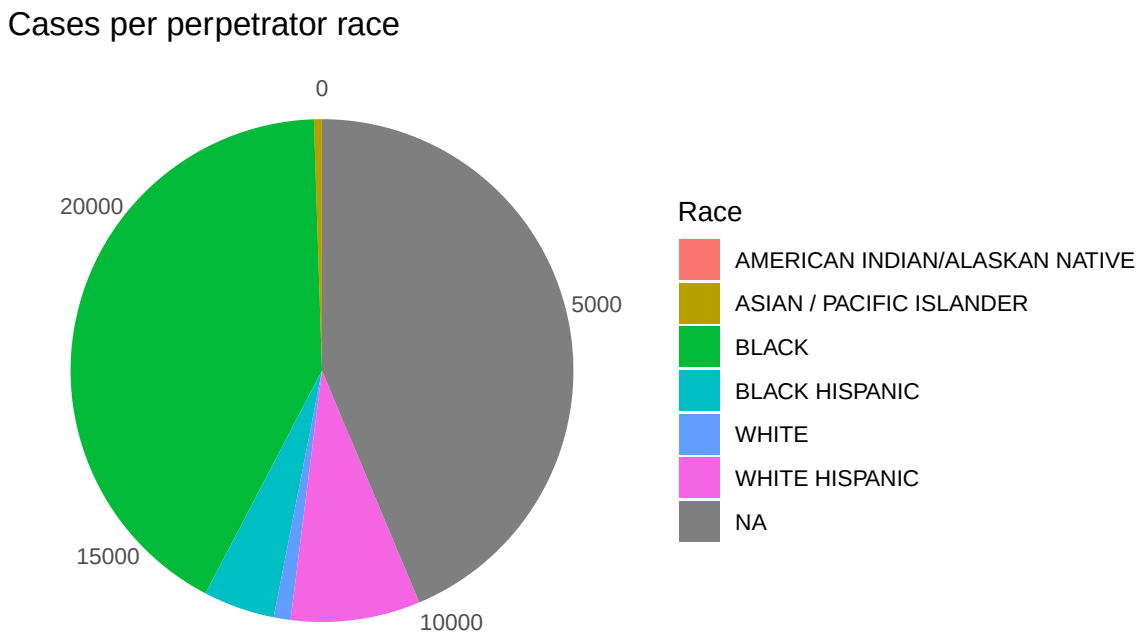
```
## `summarise()` has grouped output by 'PERP_SEX'. You can override using the `.groups` argument.
```

```
shootings_NY_perp_race_vic_sex <- shootings_NY_perp_race_vic_sex %>% rename(cases = cases.x, total_
shootings_NY_perp_race_vic_sex <- shootings_NY_perp_race_vic_sex %>% mutate(pct = round(cases / tot
```

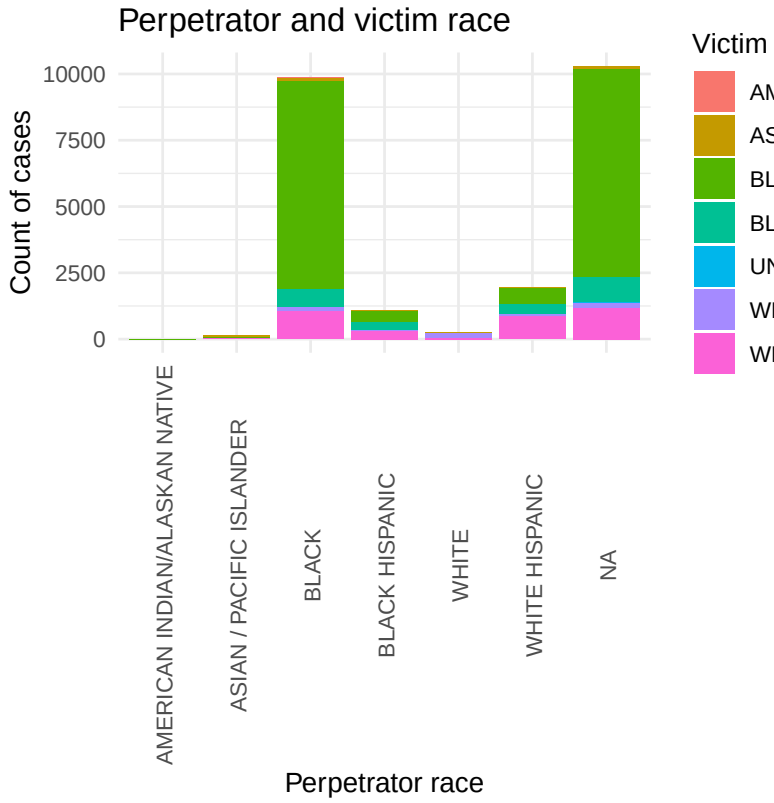
Analyzing the shootings per race

```
shootings_NY_per_perp_race %>% arrange(desc(cases))
```

```
## # A tibble: 7 x 2
##   PERP_RACE                      cases
##   <fct>                          <int>
## 1 <NA>                           10294
## 2 BLACK                           9855
## 3 WHITE HISPANIC                  1961
## 4 BLACK HISPANIC                  1081
## 5 WHITE                            255
## 6 ASIAN / PACIFIC ISLANDER         120
## 7 AMERICAN INDIAN/ALASKAN NATIVE     2
```



Looking at the race of perpetrators it is immediately visible that there is a huge proportion of unknown values. The 2nd largest group is black, while the smallest one is American Indian/Alaskan native. Ana-



lyzing the shootings per race perpetrator and victim

Looking at the perpetrator and victim race it is visible that the black race is predominant for being perpetrator and a victim. Also, there is no American Indian/Alaskan native being perpetrator and victim. Analyzing the shootings per sex

```
shootings_NY_perp_race_vic_sex %>% arrange(desc(total_cases), desc(cases))
```

```
##    PERP_SEX VIC_SEX cases total_cases   pct
## 1         M       M 11881       13305 89.30
## 2         M       F  1414       13305 10.63
## 3         M       U    10       13305  0.08
## 4      <NA>       M  7798        8425 92.56
## 5      <NA>       F   619        8425  7.35
## 6      <NA>       U     8        8425  0.09
## 7         U       M  1390        1504 92.42
## 8         U       F   113        1504  7.51
## 9         U       U     1        1504  0.07
## 10        F       M   284         334 85.03
## 11        F       F    49         334 14.67
## 12        F       U     1         334  0.30
```

Looking at the gender is immediately visible that the perpetrator and the victim are males. Very few females are perpetrators attacking another female.

4. Model Linear regression is used to estimate the relationships between the number or murders and not murders per boroughs

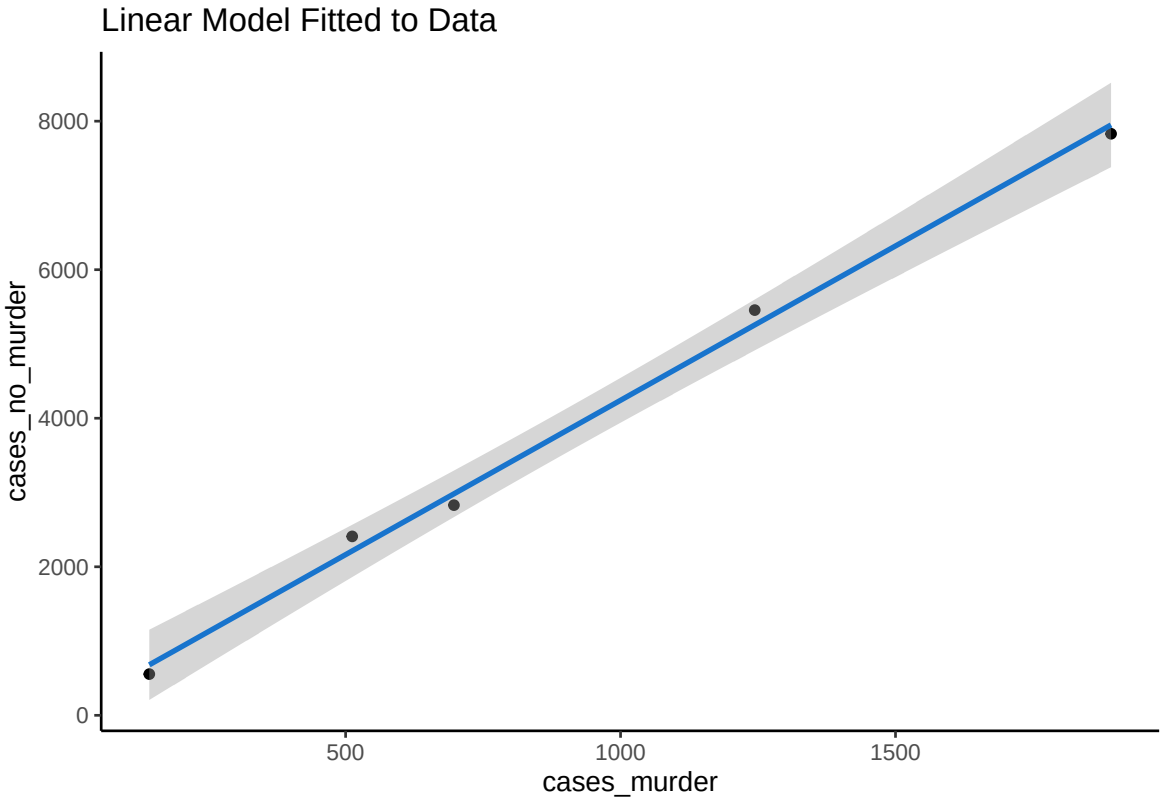
```
# Create the murders_NY_per_boro_total dataset
murders_NY_per_boro_Y <- murders_NY_per_boro %>%
filter(STATISTICAL_MURDER_FLAG==TRUE)
murders_NY_per_boro_N <- murders_NY_per_boro %>%
filter(STATISTICAL_MURDER_FLAG==FALSE)
murders_NY_per_boro_total <- left_join(murders_NY_per_boro_Y,murders_NY_per_boro_N, by='BORO') %>%
select(-c(STATISTICAL_MURDER_FLAG.x,STATISTICAL_MURDER_FLAG.y,total_cases.y,pct.x,pct.y)) %>%
rename("cases_murder"="cases.x","cases_no_murder"="cases.y","total_case"="total_cases.x")
```

```
# Create the Linear regression
mod <- lm(cases ~ total_cases, data = shootings_NY_perp_race_vic_sex)
summary(mod)
```

```
##
## Call:
## lm(formula = cases ~ total_cases, data = shootings_NY_perp_race_vic_sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4425.0 -2342.1  -249.3   351.7  7446.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.313e-12  1.517e+03   0.000    1.000
## total_cases  3.333e-01  1.917e-01   1.739    0.113
##
## Residual standard error: 3506 on 10 degrees of freedom
## Multiple R-squared:  0.2322,    Adjusted R-squared:  0.1554
## F-statistic: 3.024 on 1 and 10 DF,  p-value: 0.1127
```

Look at our model fitted to our data for murder and no murder cases

```
## `geom_smooth()` using formula 'y ~ x'
```



Step 4 - Add Bias Identification

Write the conclusion to your project report and include any possible sources of bias.

In conclusion, this is a challenging data set because of the large number of missing values as well as some incorrect data found on the PER_AGE column. However, the data is very interesting because there are many possible analysis to make according to the questions to solve. For this report, I focused on the boroughs, and perpetrators and its victims per sex and gender.

After reading and watching some articles about Bronx, I believed 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 because I watched shows and ads to stop men killing women on the TV. I mitigated this bias by doing this assignment which uses factual data.