NYPD Shooting Data Report

T. Vo

2022-06-09

Setting up tidyverse package

```
library(tidyverse)
library(lubridate)
```

Importing in data from online website

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

Reading in the data imported

```
ny_shooting <- read_csv(url_in)
```

```
## 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.
```

Specify and list out all the columns

```
ny_shooting
```

```
## # A tibble: 25,596 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>        <dbl>              <dbl>
##  1     24050482 08/27/2006 05:35      BRONX           52                  0
##  2     77673979 03/11/2011 12:03      QUEENS         106                  0
```

```
##  3    226950018 04/14/2021 21:08      BRONX           42                  0
##  4    237710987 12/10/2021 19:30      BRONX           52                  0
##  5    224701998 02/22/2021 00:18      MANHATTAN       34                  0
##  6    225295736 03/07/2021 06:15      BROOKLYN        75                  0
##  7    231190175 07/21/2021 00:40      MANHATTAN       32                  0
##  8    233429421 09/11/2021 20:20      MANHATTAN       26                  2
##  9    227950661 05/09/2021 02:50      BRONX           41                  2
## 10    227344198 04/23/2021 13:25      BROOKLYN        67                  0
## # ... with 25,586 more rows, and 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>
```

Here I am selecting and removing all columns that I think won't serve a use for my analysis.

```
nyshoot <- ny_shooting %>%
  select(-c(INCIDENT_KEY,PRECINCT,JURISDICTION_CODE,X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat, L


nyshoot
```

```
## # A tibble: 25,596 x 10
##    OCCUR_DATE OCCUR_TIME BORO      STATISTICAL_MURDER_F~ PERP_AGE_GROUP PERP_SEX
##    <chr>      <time>     <chr>     <lgl>                 <chr>          <chr>
##  1 08/27/2006 05:35      BRONX     TRUE                  <NA>           <NA>
##  2 03/11/2011 12:03      QUEENS    FALSE                 <NA>           <NA>
##  3 04/14/2021 21:08      BRONX     TRUE                  <NA>           <NA>
##  4 12/10/2021 19:30      BRONX     FALSE                 <NA>           <NA>
##  5 02/22/2021 00:18      MANHATTAN FALSE                 <NA>           <NA>
##  6 03/07/2021 06:15      BROOKLYN  TRUE                  25-44          M
##  7 07/21/2021 00:40      MANHATTAN FALSE                 25-44          M
##  8 09/11/2021 20:20      MANHATTAN FALSE                 <NA>           <NA>
##  9 05/09/2021 02:50      BRONX     TRUE                  25-44          M
## 10 04/23/2021 13:25      BROOKLYN  FALSE                 <NA>           <NA>
## # ... with 25,586 more rows, and 4 more variables: PERP_RACE <chr>,
## #   VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>
```

To summarize the above, I imported the data and went through the columns and deleted the ones I think will not serve a purpose for my analysis, such as the latitude/longitude, x/y coordinates, jurisdiction codes, and so on. I changed the OCCUR_DATE column type to the appropriate date type.

Tidying and transforming the data

Here I saw that the perpetrators in terms of age, sex, and race had a large amount of missing data. Because of this huge amount of missing data, I've chosen to label them as unknown as part of my analysis.

```
nyshoot_2 <- nyshoot %>% select(everything())

# Returns column names and missing values

lapply(nyshoot_2, function(x) sum(is.na(x)))
```

```
## $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
```

Transforming the data

Here I have transformed all the data types to their respective types.

```
#Tidying it up and then transforming it

nyshoot_2 <- nyshoot_2 %>%
    replace_na(list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "UNKNOWN", PERP_RACE = "UNKNOWN"))

nyshoot_2 <- nyshoot_2 %>% mutate(
  PERP_AGE_GROUP=recode(PERP_AGE_GROUP, UNKNOWN="UNKNOWN"),
  PERP_SEX=recode(PERP_SEX, U="UNKNOWN"),
  PERP_RACE=recode(PERP_RACE, UNKNOWN="UNKNOWN"),
  VIC_AGE_GROUP=recode(VIC_AGE_GROUP, UNKNOWN="UNKNOWN"),
  VIC_SEX=recode(VIC_SEX, U="UNKNOWN"),
  VIC_RACE=recode(VIC_RACE, UNKNOWN="UNKNOWN"),
  PERP_AGE_GROUP=as.factor(PERP_AGE_GROUP),
  PERP_SEX=as.factor(PERP_SEX),
  PERP_RACE=as.factor(PERP_RACE),
  VIC_AGE_GROUP=as.factor(VIC_AGE_GROUP),
  VIC_SEX=as.factor(VIC_SEX),
  VIC_RACE=as.factor(VIC_RACE),
  BORO = as.factor(BORO),
```

```
  OCCUR_DATE = mdy(OCCUR_DATE)
)

#Summarization of data

summary(nyshoot_2)
```

```
##    OCCUR_DATE          OCCUR_TIME                  BORO
##  Min.   :2006-01-01   Length:25596      BRONX        : 7402
##  1st Qu.:2009-05-10   Class1:hms        BROOKLYN     :10365
##  Median :2012-08-26   Class2:difftime   MANHATTAN    : 3265
##  Mean   :2013-06-13   Mode  :numeric    QUEENS       : 3828
##  3rd Qu.:2017-07-01                     STATEN ISLAND:  736
##  Max.   :2021-12-31
##
##  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

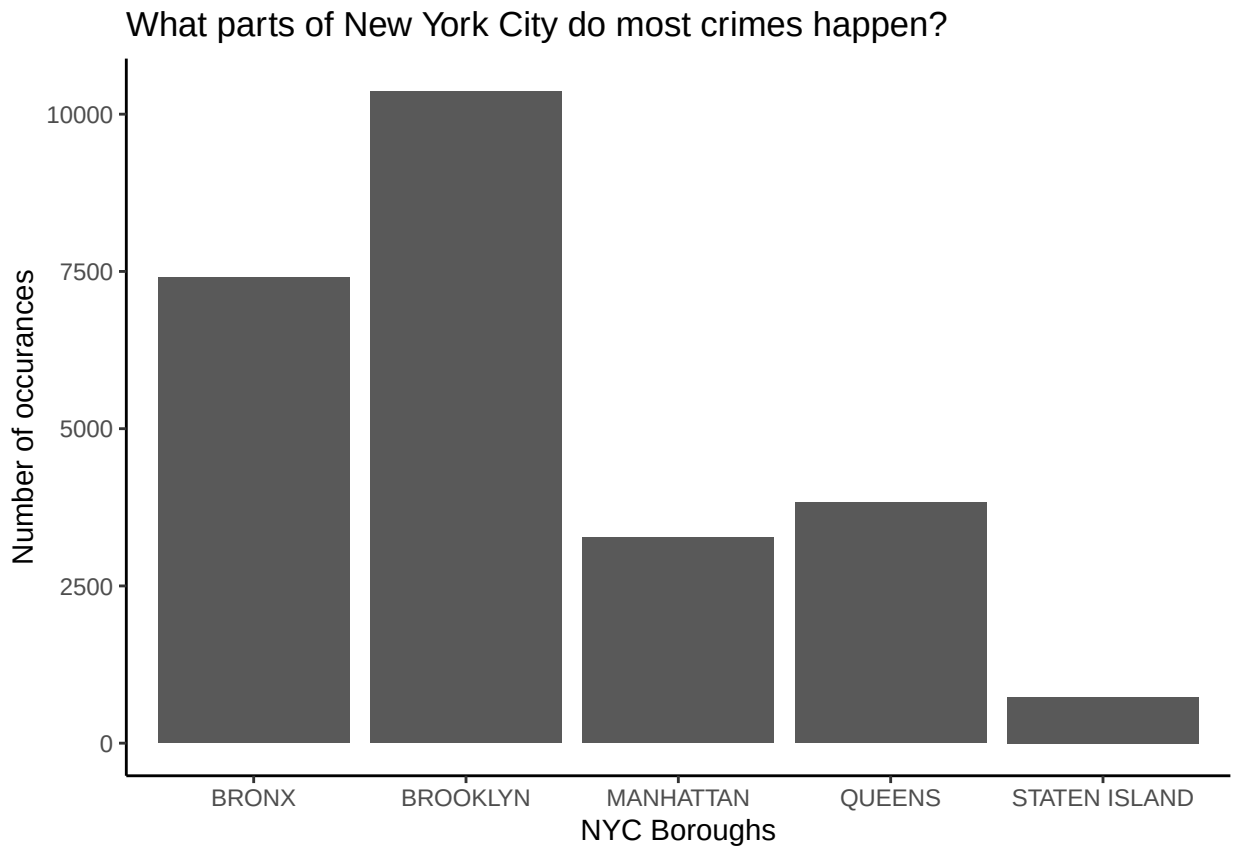
Bar Chart Visualization

Here I made a visualization showing which neighborhoods of New York city had the most occurrences of shooting incidents. As we can see Brooklyn is the top borough, with Staten Island all the way on the bottom.

```
g <- ggplot(nyshoot_2, aes(x=BORO)) +
              geom_bar() +
              labs(
                title = "What parts of New York City do most crimes happen?",
```

```
                x = "NYC Boroughs",
                y = "Number of occurances") + theme_classic()

g
```



Line chart visualization

Here I visualized the number of incidents that happened at specific times during the day (in military time to account for time zone differences). As you can see, most of these crimes happen during dusk hours.

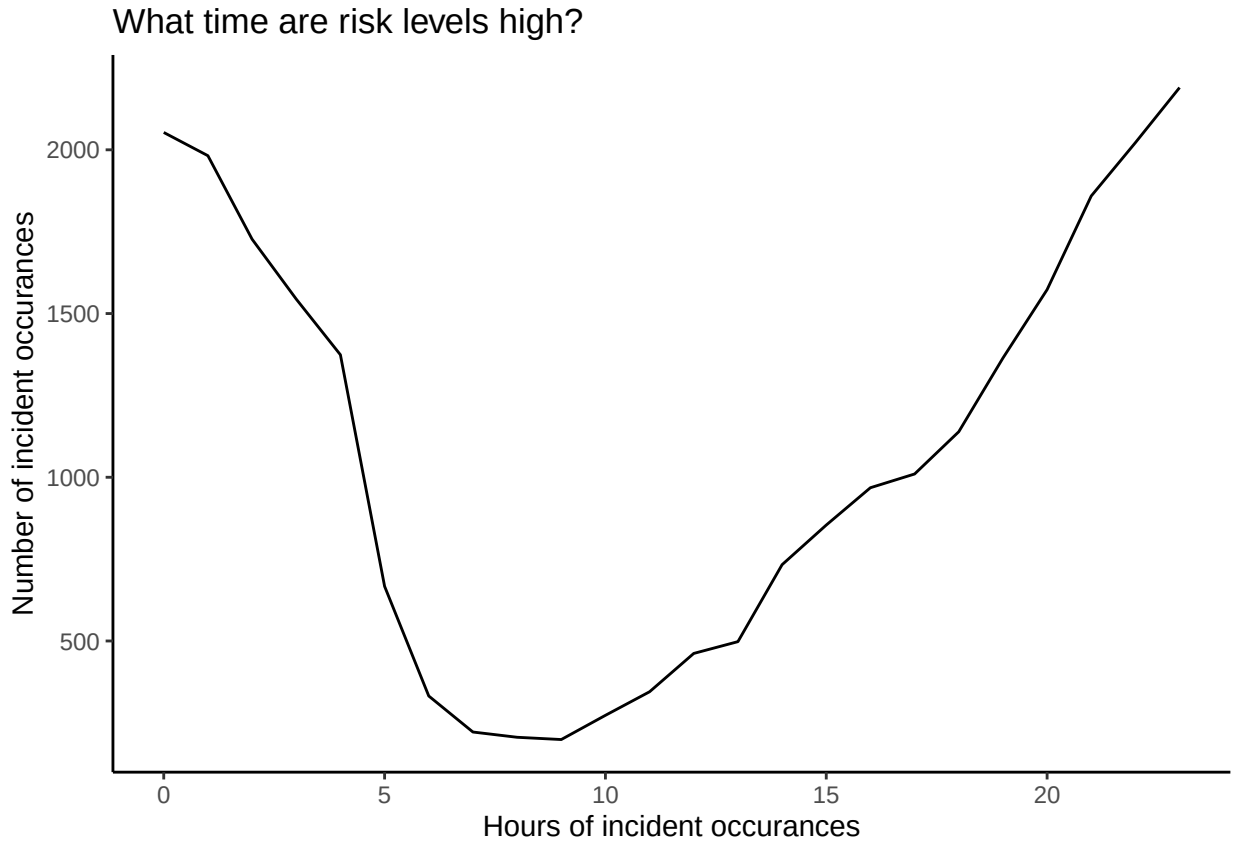
```
nyshoot_2 <- nyshoot_2 %>%
  mutate(OCCUR_HOUR = hour(hms(as.character(OCCUR_TIME))))

nyshoot_hr <- nyshoot_2 %>%
  group_by(OCCUR_HOUR) %>% count()

# Extracting hour time from OCCUR_DATE and making a seperate data variable for it
```

```
g <- ggplot(nyshoot_hr, aes(x = OCCUR_HOUR, y = n)) +
  geom_line() +
  labs(
    title = "What time are risk levels high?",
    x = "Hours of incident occurances",
    y = "Number of incident occurances"
```

```
  ) + theme_classic()
g
```



Linear model

Here I made a linear model based on these variables to make a prediction on how probable it is that the incident is also a case of murder as well based on the statistical murder flag data given. Based on the estimates given, the perpetrator whose race is white changes the likelihood of a murder related incident by about ten percent.

```
model <- glm.fit <- glm( STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + OCCUR_HOUR, 

summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     PERP_AGE_GROUP + OCCUR_HOUR, data = nyshoot_2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.51595  -0.20515  -0.16761  -0.02327   1.02958
##
## Coefficients:
```

```
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -0.0356475  0.2750576  -0.130 0.896884
## PERP_RACEASIAN / PACIFIC ISLANDER  0.3401742  0.2767328   1.229 0.218989
## PERP_RACEBLACK                     0.2527505  0.2748272   0.920 0.357754
## PERP_RACEBLACK HISPANIC            0.2300557  0.2750312   0.836 0.402898
## PERP_RACEUNKNOWN                   0.2056396  0.2756046   0.746 0.455590
## PERP_RACEWHITE                     0.3840598  0.2758241   1.392 0.163811
## PERP_RACEWHITE HISPANIC            0.2751276  0.2749177   1.001 0.316951
## PERP_SEXM                         -0.0356969  0.0204664  -1.744 0.081141 .
## PERP_SEXUNKNOWN                    0.1567518  0.0289440   5.416 6.16e-08 ***
## PERP_AGE_GROUP1020                -0.1809068  0.3885286  -0.466 0.641491
## PERP_AGE_GROUP18-24                0.0267349  0.0113627   2.353 0.018637 *
## PERP_AGE_GROUP224                 -0.2027846  0.3885947  -0.522 0.601786
## PERP_AGE_GROUP25-44                0.0864979  0.0115199   7.509 6.17e-14 ***
## PERP_AGE_GROUP45-64                0.1566608  0.0197597   7.928 2.31e-15 ***
## PERP_AGE_GROUP65+                  0.2032352  0.0529160   3.841 0.000123 ***
## PERP_AGE_GROUP940                 -0.1987905  0.3885884  -0.512 0.608956
## PERP_AGE_GROUPUNKNOWN             -0.1581333  0.0140694 -11.239  < 2e-16 ***
## OCCUR_HOUR                        -0.0002496  0.0002861  -0.872 0.382952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1508354)
##
##     Null deviance: 3979.2  on 25595  degrees of freedom
## Residual deviance: 3858.1  on 25578  degrees of freedom
## AIC: 24242
##
## Number of Fisher Scoring iterations: 2
```

Analysis of data

After going through the data, there are some interesting points that stood out. Most the perpetrators as well as victims were male, Black and White Hispanic make up a majority of the victims, and although a large chunk of the sexes of the perpetrators are unknown, a majority of it is made up of males. A majority of these victims and perpetrators were also from ages 44 to <18.

Some questions this might raise to me would be why are Brooklyn and the Bronx leading in terms of crime? Why is Staten Island so low? Is there any other links between all these variables that can be made?

```
table(
  nyshoot_2 %>% select(VIC_SEX, PERP_SEX)
)
```

```
##          PERP_SEX
## VIC_SEX       F     M UNKNOWN
##   F          58  1540     805
##   M         312 12870   10000
##   UNKNOWN     1     6       4
```

```
table(
  nyshoot_2 %>% select(PERP_AGE_GROUP, VIC_AGE_GROUP)
)
```

```
##                VIC_AGE_GROUP
## PERP_AGE_GROUP  <18 18-24 25-44 45-64  65+ UNKNOWN
##        <18      445   584   353    70    9       2
##        1020       0     0     1     0    0       0
##        18-24    742  2607  2141   305   37      12
##        224        0     1     0     0    0       0
##        25-44    247  1417  3033   431   40      34
##        45-64     19    62   290   148   11       5
##        65+        0     1    23    23   10       0
##        940        0     0     1     0    0       0
##        UNKNOWN 1228  4932  5544   721   60       7
```

```
table(
  nyshoot_2 %>% select(PERP_RACE, VIC_RACE)
)
```

```
##                                VIC_RACE
## PERP_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
##                                VIC_RACE
## PERP_RACE                       ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     2              0
##   ASIAN / PACIFIC ISLANDER                             43    51             13
##   BLACK                                               135  8471            749
##   BLACK HISPANIC                                       17   481            320
##   UNKNOWN                                             113  8523            999
##   WHITE                                                11    34             21
##   WHITE HISPANIC                                       35   719            383
##                                VIC_RACE
## PERP_RACE                       UNKNOWN WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0    11             23
##   BLACK                               24   183           1102
##   BLACK HISPANIC                       5    34            346
##   UNKNOWN                             24   187           1295
##   WHITE                                1   156             49
##   WHITE HISPANIC                      11    89            927
```

Bias identification

On the topic of crime in America, which is something that a lot of people have implicit bias already in the present day. With things like social media and the internet in this day and age, it is incredibly easy and also hard for people to develop bias towards this topic. With so much information, it can be overwhelming.

My personal bias I would say coming into this data analysis, even though I've never visited New York City, is that I had some innate feelings regarding New York and crime, to me it seemed like the two went hand in hand somewhat. Even growing up my parents always told me to not go there because of their fear of

crime in that city (even though they've never visited either). Although I did have these bias regarding this topic on crime in New York City, when analyzing data it is of utmost importance that you look at things objectively, which I focused on doing while reading through and analyzing the data.