nypd-shooting-submission

FB

2024-02-01

```
library(tidyverse) # for manipulating and data
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate) # for working with date series
```

```
# reference 1

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

nypd_df <- read_csv(url)
```

```
## 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.
```

Introduction

This project was completed for DTAS5301 as part of the MS-DS programme at the University of Colorado, Boulder. This topic is a matter of public safety and should be of interest to not only those who live in New York City, but for policy makers, law enforcement agencies, social advocacy groups, and community associations in other cities as well. Efforts to reduce gun violence would first begin with identifying patterns in the data, so that resources can be most efficiently allocated to make the largest impact for those most affected.

Questions of interest

In this analysis, I will be addressing the following questions:

- What is going on?
- Who is most affected?
- Where do the incidents occur?
- When do these incidents occur?
- Why is this happening?

The main focus of my analysis will be on when there incidents occur.

Data cleaning and exploratory analysis

What is going on?

To determine what is going on, some exploratory analysis is required to see what is included in the data set.

```
head(nypd_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>
```

```
summary(nypd_df)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
```

```
##                    3rd Qu.: 81.00   3rd Qu.:0.0000
##                    Max.   :123.00   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Notes The data set contains information about shooting incidents collected by the New York Police Department from, at time of writing, 2006 through 2022. Information contains dates, times, precints, jurisdictions, locations and their descriptions, perpetrator and victim demographics, and whether or not the victim survived(statistical murder flag). Some data cleaning is required.

```
head(nypd_df$LOCATION_DESC, 20)
```

```
##  [1] NA                          NA
##  [3] NA                          NA
##  [5] NA                          NA
##  [7] NA                          NA
##  [9] NA                          "MULTI DWELL - APT BUILD"
## [11] "MULTI DWELL - PUBLIC HOUS" "GROCERY/BODEGA"
## [13] NA                          NA
## [15] "MULTI DWELL - PUBLIC HOUS" "MULTI DWELL - PUBLIC HOUS"
## [17] NA                          "MULTI DWELL - APT BUILD"
## [19] NA                          "MULTI DWELL - APT BUILD"
```

```r
unique(nypd_df$LOCATION_DESC)
```

```
##  [1] NA                           "MULTI DWELL - APT BUILD"
##  [3] "MULTI DWELL - PUBLIC HOUS" "GROCERY/BODEGA"
##  [5] "JEWELRY STORE"              "CLOTHING BOUTIQUE"
##  [7] "GAS STATION"                "BAR/NIGHT CLUB"
##  [9] "PVT HOUSE"                  "NONE"
## [11] "COMMERCIAL BLDG"            "SMALL MERCHANT"
## [13] "BEAUTY/NAIL SALON"          "FAST FOOD"
## [15] "DRUG STORE"                 "TELECOMM. STORE"
## [17] "DRY CLEANER/LAUNDRY"        "RESTAURANT/DINER"
## [19] "HOTEL/MOTEL"                "SOCIAL CLUB/POLICY LOCATI"
## [21] "SUPERMARKET"                "CHAIN STORE"
## [23] "HOSPITAL"                   "LIQUOR STORE"
## [25] "STORE UNCLASSIFIED"         "(null)"
## [27] "FACTORY/WAREHOUSE"          "DEPT STORE"
## [29] "SHOE STORE"                 "VARIETY STORE"
## [31] "BANK"                       "ATM"
## [33] "DOCTOR/DENTIST"             "GYM/FITNESS FACILITY"
## [35] "CANDY STORE"                "VIDEO STORE"
## [37] "SCHOOL"                     "LOAN COMPANY"
## [39] "PHOTO/COPY STORE"           "CHECK CASH"
## [41] "STORAGE FACILITY"
```

Notes In the location description column, the NA values, '(null)', and 'NONE' need to be combined.

```r
# reference 2
nypd_df$LOCATION_DESC <- ifelse(is.na(nypd_df$LOCATION_DESC) | nypd_df$LOCATION_DESC == "(null)", "NONE
```

```r
loc_proportions <- as.data.frame(prop.table(table(nypd_df$LOCATION_DESC)))
  loc_proportions[order(loc_proportions$Freq, decreasing = TRUE), , drop = FALSE]
```

```
##                         Var1         Freq
## 26                      NONE 5.905463e-01
## 25 MULTI DWELL - PUBLIC HOUS 1.769186e-01
## 24   MULTI DWELL - APT BUILD 1.038005e-01
## 28                 PVT HOUSE 3.481986e-02
## 17            GROCERY/BODEGA 2.541008e-02
## 3             BAR/NIGHT CLUB 2.299356e-02
## 9            COMMERCIAL BLDG 1.069127e-02
## 29          RESTAURANT/DINER 7.469244e-03
## 4          BEAUTY/NAIL SALON 4.100762e-03
## 15                 FAST FOOD 3.807850e-03
## 33 SOCIAL CLUB/POLICY LOCATI 2.636204e-03
## 16               GAS STATION 2.599590e-03
## 19                  HOSPITAL 2.379906e-03
## 22              LIQUOR STORE 1.501172e-03
## 32            SMALL MERCHANT 1.354716e-03
## 35        STORE UNCLASSIFIED 1.318102e-03
## 20                HOTEL/MOTEL 1.281488e-03
## 13       DRY CLEANER/LAUNDRY 1.135032e-03
## 36                SUPERMARKET 7.688928e-04
```

```
## 8          CLOTHING BOUTIQUE 5.125952e-04
## 12               DRUG STORE 5.125952e-04
## 21            JEWELRY STORE 4.393673e-04
## 37          TELECOMM. STORE 4.027534e-04
## 38            VARIETY STORE 4.027534e-04
## 31               SHOE STORE 3.661394e-04
## 10               DEPT STORE 3.295255e-04
## 14         FACTORY/WAREHOUSE 2.929115e-04
## 39               VIDEO STORE 2.929115e-04
## 5                CANDY STORE 2.562976e-04
## 6                CHAIN STORE 1.830697e-04
## 2                      BANK 1.098418e-04
## 18      GYM/FITNESS FACILITY 1.098418e-04
## 1                       ATM 3.661394e-05
## 7                CHECK CASH 3.661394e-05
## 11            DOCTOR/DENTIST 3.661394e-05
## 23             LOAN COMPANY 3.661394e-05
## 27          PHOTO/COPY STORE 3.661394e-05
## 30                   SCHOOL 3.661394e-05
## 34          STORAGE FACILITY 3.661394e-05
```

There are a lot of different locations with only a tiny fraction of the population. Those can be combined into a single 'BUSINESS / OTHER' category for easier plotting.

```
loc_list <- c('NONE', 'MULTI DWELL - PUBLIC HOUS', 'MULTI DWELL - APT BUILD', 'PVT HOUSE')

nypd_df <- nypd_df %>%
  mutate(LOCATION_DESC = ifelse(LOCATION_DESC %in% loc_list, LOCATION_DESC, 'BUSINESS / OTHER'))
```

```
boro_proportions <- as.data.frame(prop.table(table(nypd_df$BORO)))
boro_proportions[order(boro_proportions$Freq, decreasing = TRUE), , drop = FALSE]
```

```
##            Var1       Freq
## 2      BROOKLYN 0.40030023
## 1         BRONX 0.29060486
## 4        QUEENS 0.14989748
## 3     MANHATTAN 0.13078500
## 5 STATEN ISLAND 0.02841242
```

Staten Island has a very small fraction (2.8%) of shooting incidents. For the sake of plotting, I will not use this subset in my analysis.

```
nypd_df <- subset(nypd_df, BORO != 'STATEN ISLAND')
```

```
age_group_proportions <- as.data.frame(prop.table(table(nypd_df$VIC_AGE_GROUP)))
age_group_proportions[order(age_group_proportions$Freq, decreasing = TRUE), , drop = FALSE]
```

```
##      Var1         Freq
## 4   25-44 4.502563e-01
## 3   18-24 3.693473e-01
## 1     <18 1.039720e-01
```

```
## 5   45-64 6.764396e-02
## 6     65+ 6.481761e-03
## 7 UNKNOWN 2.261079e-03
## 2    1022 3.768465e-05
```

In the age group column, 'UNKNOWN' and '1022' need to be removed.

```
nypd_df <- subset(nypd_df, !(VIC_AGE_GROUP %in% c('UNKNOWN', '1022')))
```

```
sex_proportions <- as.data.frame(prop.table(table(nypd_df$VIC_SEX)))
sex_proportions[order(sex_proportions$Freq, decreasing = TRUE), , drop = FALSE]
```

```
##   Var1         Freq
## 2    M 0.9045892351
## 1    F 0.0951841360
## 3    U 0.0002266289
```

In the victim sex column, unknown values of 'U' need to be removed. Note: the values contained in this column reflect how the data was collected and should not suggest, for example, that gender is binary.

```
nypd_df <- subset(nypd_df, VIC_SEX != 'U')
```

```
race_proportions <- as.data.frame(prop.table(table(nypd_df$VIC_RACE)))
race_proportions[order(race_proportions$Freq, decreasing = TRUE), , drop = FALSE]
```

```
##                             Var1         Freq
## 3                          BLACK 0.7129850013
## 7                 WHITE HISPANIC 0.1488911557
## 4                 BLACK HISPANIC 0.0982658960
## 6                          WHITE 0.0226302467
## 2       ASIAN / PACIFIC ISLANDER 0.0151120178
## 5                        UNKNOWN 0.0017378821
## 1 AMERICAN INDIAN/ALASKAN NATIVE 0.0003778004
```

In the victim race column, 'UNKNOWN' values need to be removed. Note: the values contained in this column reflect how the data was collected and should not suggest, for example, that a person's ethnic identity can fit neatly into one of the specified categories.

```
nypd_df <- subset(nypd_df, VIC_RACE != 'UNKNOWN')
```

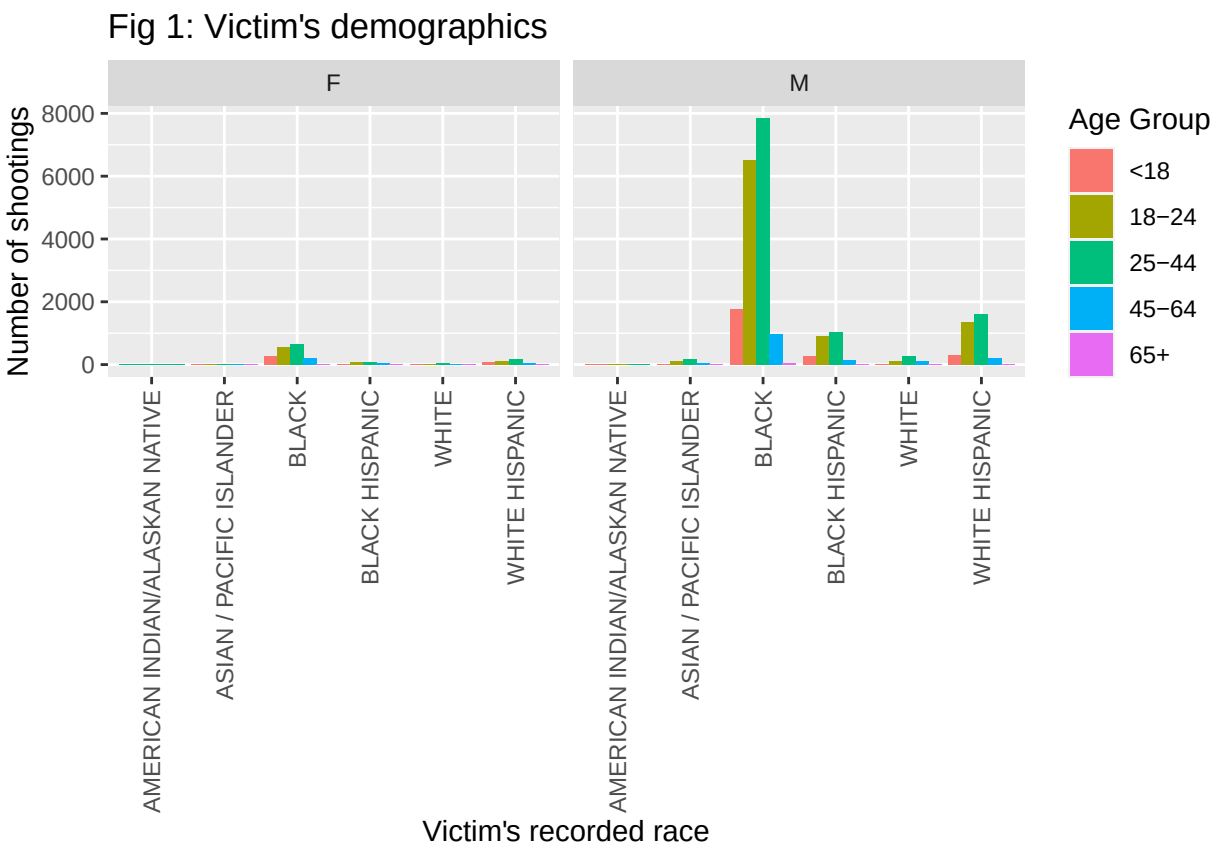
Visual analysis

Who is most affected?

```
vic_demo_plot <- ggplot(nypd_df, aes(VIC_RACE, fill = VIC_AGE_GROUP)) +
  geom_bar(stat = 'count', position = 'dodge') +
  facet_wrap(~ VIC_SEX) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
```

```
  xlab("Victim's recorded race") +
  ylab('Number of shootings') +
  labs(fill = 'Age Group') +
  ggtitle("Fig 1: Victim's demographics")

vic_demo_plot
```



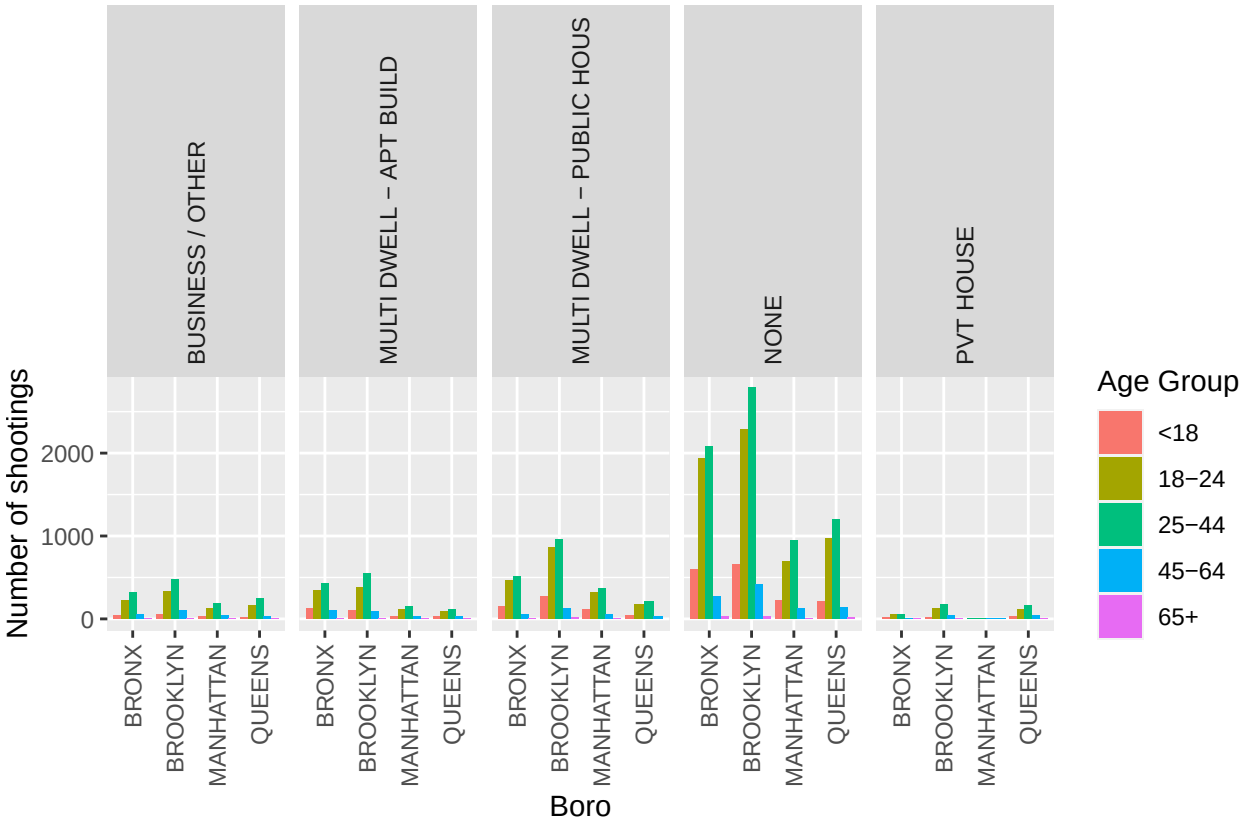
In figure 1, it appears that people who are male, between the ages of 18 and 44, and Black are the most frequent victims of the shootings.

Where do these incidents occur?

```
loc_plot <- ggplot(nypd_df, aes(BORO, fill = VIC_AGE_GROUP)) +
  geom_bar(stat = 'count', position = 'dodge') +
  facet_grid(. ~ LOCATION_DESC) +
  theme(plot.margin = margin(l = 0, r = 0, unit = "pt"),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        strip.text.x = element_text(angle = 90, vjust = 0.5, hjust = 0)) +
  xlab('Boro') +
  ylab('Number of shootings') +
  labs(fill = 'Age Group') +
  ggtitle('Fig. 2: Locations of incidents')

loc_plot
```

Fig. 2: Locations of incidents



Here in figure 2, it appears that these incidents most frequently occur in the Bronx and Brooklyn, and do not have a more specific location description.

When do these incidents occur?

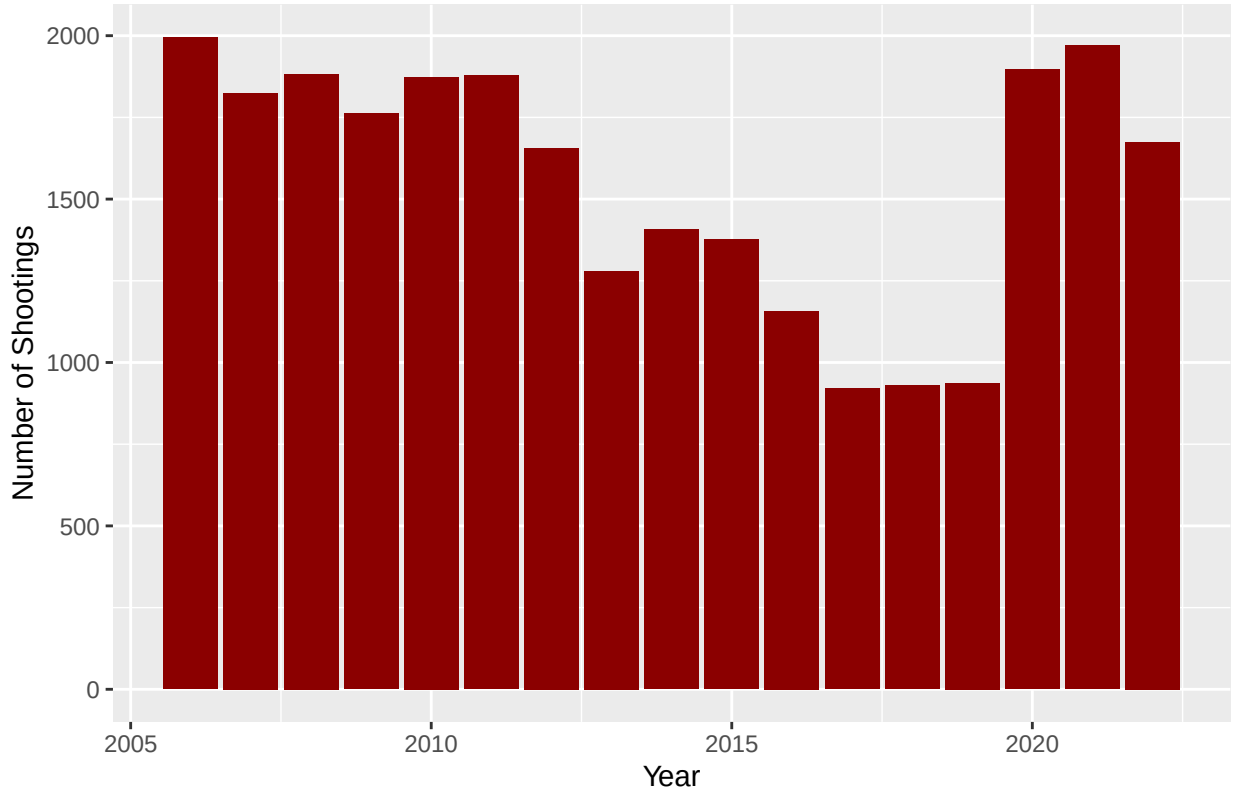
```
# transform data type
nypd_df$OCCUR_DATE <- mdy(nypd_df$OCCUR_DATE)

# create columns for year/month/day
nypd_df$Year <- lubridate::year(nypd_df$OCCUR_DATE)
nypd_df$Month <- lubridate::month(nypd_df$OCCUR_DATE, label = TRUE)
nypd_df$DayOfWeek <- lubridate::wday(nypd_df$OCCUR_DATE, label = TRUE)
```

```
year_plot <- ggplot(nypd_df, aes(x = Year)) +
  geom_bar(stat = 'count', fill = 'darkred') +
  labs(title = 'Fig 3: Yearly count of incidents, 2006 - 2022', x = 'Year', y = 'Number of Shootings')

year_plot
```

Fig 3: Yearly count of incidents, 2006 − 2022

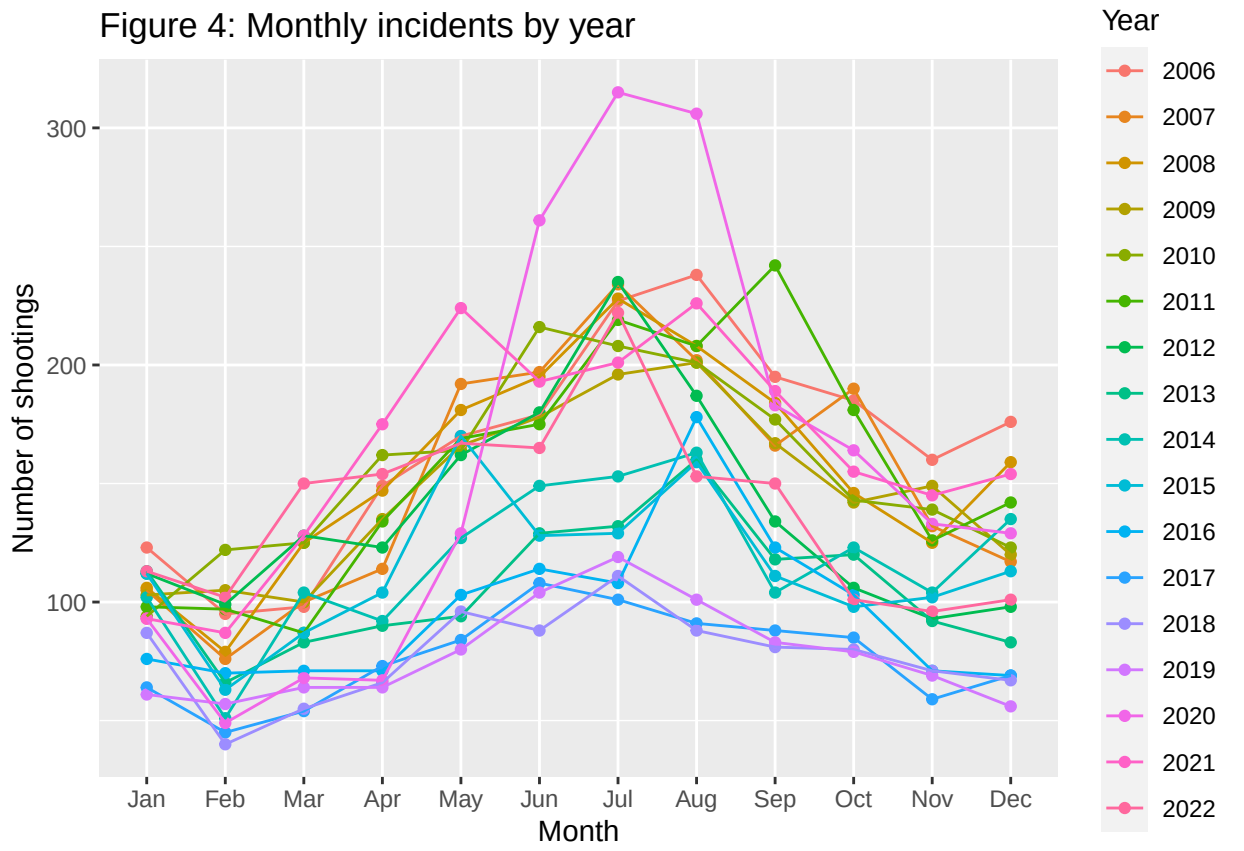


In figure 3, it appears that the yearly rate of shooting incidents was somewhat stable from the years 2006 through 2011, decreases between 2011 and 2019, then spikes in 2020 and 2021 to rates similar to those of the 2006 to 2011 period.

Now let's have a look at how many of these incidents occur each month.

```
year_month_plot <- ggplot(nypd_df, aes(x = Month, colour = factor(Year), group = factor(Year))) +
  geom_point(stat = 'count') +
  geom_line(stat = 'count') +
  labs(title = 'Figure 4: Monthly incidents by year',
       colour = 'Year',
       y = 'Number of shootings')

year_month_plot
```



In figure 4, there appears to be a fairly consistent seasonal pattern to the number of incidents. Incidents tend to be at the lowest during the month of February and peak around July through August.

Modelling

To make the creation of a model easier, I did some basic feature engineering of adding a column of 1's to the data set and aggregating the data by month, then created a linear model of a sinusoidal curve to reflect the seasonal 'wave' pattern.

```
nypd_df$Month <- as.numeric(nypd_df$Month)
nypd_df$COUNT <- 1
```

```
# Reference 3
monthly_counts <- aggregate(COUNT ~ Month + Year, data = nypd_df, FUN = sum)

monthly_counts$Sin_Month <- sin(2 * pi * monthly_counts$Month / 12)
monthly_counts$Cos_Month <- cos(2 * pi * monthly_counts$Month / 12)

model <- lm(COUNT ~ Sin_Month + Cos_Month, data = monthly_counts)

summary(model)
```

```
##
## Call:
```

```
## lm(formula = COUNT ~ Sin_Month + Cos_Month, data = monthly_counts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.343 -31.757   5.259  27.050 141.479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  129.525      2.833  45.726  < 2e-16 ***
## Sin_Month    -30.168      4.006  -7.531 1.67e-12 ***
## Cos_Month    -33.386      4.006  -8.334 1.21e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.46 on 201 degrees of freedom
## Multiple R-squared:  0.3856, Adjusted R-squared:  0.3795
## F-statistic: 63.09 on 2 and 201 DF,  p-value: < 2.2e-16
```

Model interpretation

- Multiple R-squared value: 38.5% of the variance in the count of incidents can be explained by the month of the year.
- F-statistic = 63.09, meaning that the model as a whole is useful in predicting incident counts.

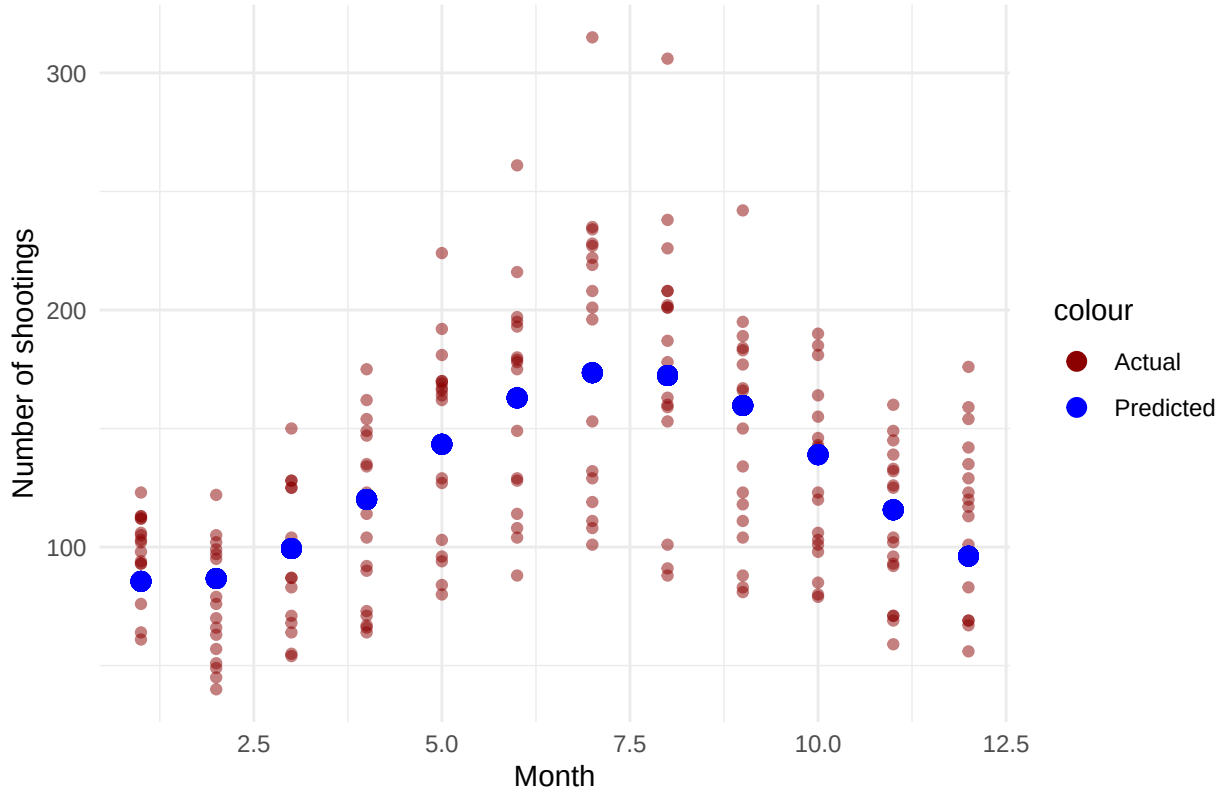
```
monthly_counts$PRED_COUNT <- predict(model)
```

```
seasonal_plot <- ggplot(monthly_counts) +
  geom_point(aes(x = Month, y = COUNT, colour = 'Actual'), alpha = 0.5) +
  geom_point(aes(x = Month, y = PRED_COUNT, colour = 'Predicted'), size = 3) +
  labs(x = 'Month', y = 'Number of shootings',
       title = 'Fig. 5: Observed vs. predicted counts of monthly incidents') +
  scale_colour_manual(values = c('Actual' = 'darkred', 'Predicted' = 'blue'),
                      labels = c('Actual', 'Predicted')) +

  theme_minimal()

seasonal_plot
```

Fig. 5: Observed vs. predicted counts of monthly incidents



Discussion

In my analysis I have discovered that the neighbourhoods with the highest rate of shooting incidents are the Bronx and Brooklyn, and do not have a more specific location description. The people who are the most frequent victims of shooting incidents are those who are male, are in the age groups of 18-24 and 25-44, and are Black.

I decided to mainly focus on the analysis of the times and dates that these incidents occur. Taking a look at the year-to-year data in figure 3, the number of shooting incidents was somewhat stable from the years 2006 through 2011, decreases between 2011 and 2019, then rises in 2020 and 2021 to rates similar to those in the 2006 to 2011 period.

When looking at figure 4, I can see that there is a trend that seems similar to a seasonal average temperature in New York (4). The number of shootings are highest in the summer months of July and August, and lowest in the winter months of January and February.

Why do these incidents occur where, when, and to whom they do?

I don't know why. There would be numerous variables that could lead up to a person being shot, and consulting with subject experts from neighbourhood community associations, social advocacy groups, health care practitioners, sociologists, government administration and law enforcement would be needed in order to identify these factors. More demographic, educational and economic data would also be required.

I am not American, but have grown up consuming American media. If I were to offer a very biased opinion solely based on what I have seen in that media, I might say that the people who are often the most marginalized and economically vulnerable are those that are forced (in order to survive) to engage in questionable

dealings with questionable people that may get themselves shot, and America has a history of marginalizing people who are Black; maybe some of these people who were the victims of shootings were even shot by police because they were Black (5). As it happens, the neighbourhoods of the Bronx and Brooklyn, where these incidents most frequently occur, are listed among the most disadvantaged communities (6). Perhaps there may be some merit to my opinion, but further study would be required to determine its statistical significance.

The seasonal pattern in the data was quite interesting to me. My (unfounded) opinion is that maybe people like to keep warm (and safe) inside their own homes in the colder months, and spend more time outside in the warmer summer months where they are more exposed to receiving a gunshot wound. It is also possible that both the number of shootings and the seasonal temperatures are caused by a third unknown factor, and that weather itself does not affect the number of incidents.

Further study:

My analysis led me to consider some more questions:

- Was there a programme implemented to decrease the number of shooting incidents from 2011 onwards? Shooting incidents decreased from 2011 to 2017 and then remain at this relatively low level until May through September 2020 when incidents spiked to the highest levels within the data set. I feel that Covid-19 lock-downs may have been a factor in this spike, but again, consulting with subject experts and more data would be required.
- I had created a heatmap to show which hours of the week where the most shootings occured, but decided to cut it for brevity and because it required the installation of additional packages, which may cause issues during peer-review. According to the heatmap I created, shootings peaked on late Friday nights/Saturday mornings, and late Saturday night/Sunday Mornings. Given more time, I would have liked to use that heatmap and then do some feature engineering so that I could feed some hourly variables into a machine learning model (see next point).
- To further investigate this issue and improve the model I created, it would be interesting to use some machine learning to create another model where the number of shootings is a function of the average daily temperature, the date and time, the economic status of the neighbourhood in which the incident occurred, and the victim demographics.

References

1. DTSA 5301 course material: Nearly all code and methods in this project come from course material, printouts, lecture videos, etc.
2. ChatGPT: Not going to lie, I used AI to get some of my chunks of code working. I used this to edit my non-functioning code, to interpret error messages, to translate my ideas from my experience with Python/Pandas/Altair/SKLearn to R, and as a learning tool rather than a crutch. I know AI can spit out a lot of garbage and I did not 'copy & paste' anything blindly!
3. Modelling a sinusoidal curve: https://stats.stackexchange.com/questions/60500/how-to-find-a-good-fit-for-semi-sinusoidal-model-in-r
4. Average temperature by month in New York: https://en.climate-data.org/north-america/united-states-of-america/new-york/new-york-1091/
5. My own (theoretical) biased opinion.
6. New York Final Disadvantaged Communities: https://data.ny.gov/Energy-Environment/Final-Disadvantaged-Communities-DAC-2023-Map/6mn4-5vvz