3-2-Shooting-Data

Anonymous

2024-01-08

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     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 error
```

The Data Set

For this project, we're looking at data covering NYPD shooting incidents.

"Start an Rmd document that describes and imports the shooting project dataset in a reproducible manner."

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

shooting_data <- 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.
```

The output of the previous code block tells us that there are $27,312$ observations, each composed of 21 columns.
Let's get an idea of what's in the data set.

```
shooting_data
```

```
## # A tibble: 27,312 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
##  7     85295722 06/17/2012 22:47      QUEENS   <NA>                   114
##  8     71662474 03/08/2010 19:41      BROOKLYN <NA>                    81
##  9     83002139 02/05/2012 05:45      QUEENS   <NA>                   105
## 10     86437261 08/26/2012 01:10      QUEENS   <NA>                   101
## # i 27,302 more rows
## # 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>
```

So this appears to be a list of crimes that involve shooting including details about the perpetrators and the victims as well as when and where the shooting occurred.

Summary

“Add to your Rmd document a summary of the data and clean up your dataset by changing appropriate variables to factor and date types and getting rid of any columns not needed. Show the summary of your data to be sure there is no missing data. If there is missing data, describe how you plan to handle it.”

Let’s see what those columns are and a summary of their contents.

```
summary(shooting_data)
```

```
##   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.:1000028   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
```

From the summary, we can see that the date column is not a date data-type, so we can fix that.
As with the COVID data, precise locations are probably not interesting, so we will drop the coordinates and longitude/latitude columns.
Similarly, the incident key is more of a logging element, so not useful for analysis.

Before we do anything, what else can we work out?
We probably want to check out how many NA values exist in each column, which will give us an idea of what columns are useless (even if they are theoretically interesting) as well as some idea of what could be done with other columns.

It might also be useful to see how many unique values occur in each column. This can give us an idea if a column has meaningful information. If there are as many unique values as there are rows when talking about a location, for example, then there's probably nothing useful in that column.

```r
# Stolen from somewhere online - count how many values in each column are NA
shooting_data %>%
  summarise(across(everything(),
                  ~sum(is.na(.))))
```

```
## # A tibble: 1 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME  BORO LOC_OF_OCCUR_DESC PRECINCT
##          <int>      <int>      <int> <int>             <int>    <int>
## 1            0          0          0     0             25596        0
## # i 15 more variables: JURISDICTION_CODE <int>, LOC_CLASSFCTN_DESC <int>,
## #   LOCATION_DESC <int>, STATISTICAL_MURDER_FLAG <int>, PERP_AGE_GROUP <int>,
## #   PERP_SEX <int>, PERP_RACE <int>, VIC_AGE_GROUP <int>, VIC_SEX <int>,
## #   VIC_RACE <int>, X_COORD_CD <int>, Y_COORD_CD <int>, Latitude <int>,
## #   Longitude <int>, Lon_Lat <int>
```

```r
# Counting unique values - apply the function to each column with sapply
sapply(shooting_data, function(x) n_distinct(x))
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                   21420                    5761                    1421
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                       5                       3                      77
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                       4                      10                      41
## STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP                PERP_SEX
##                       2                      11                       5
##               PERP_RACE           VIC_AGE_GROUP                 VIC_SEX
##                       9                       7                       3
##                VIC_RACE              X_COORD_CD              Y_COORD_CD
##                       7                   12088                   12283
##                Latitude               Longitude                 Lon_Lat
##                   12628                   12618                   12646
```

Looking at these results, we can see that most location of occurrence description fields are empty. Thus, we'll just drop it.
Same for the location classification field.

A little over half of the location description fields are empty, but of those that remain, there are a neat package of 41 unique values, potentially giving us some limited insight. Perhaps the NA fields are unknown or out in the open?

Many perpetrator details are missing, but this makes sense as not every shooting is solved. These account for roughly 1/3 of the cases and can be kept in.

Looking now at just the unique values, we have 7 age groups, which makes sense (spoiler: just by scouring manually, I spotted an age group of 940, which is…probably not accurate, so this will require further cleaning down the line).
Racial splits are similar and make sense at a brief glance.

Using
https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8
we see that jurisdiction code is also maybe not interesting at this point. Same with the precinct as we already have the borough. Might want to revisit this later, but for now, that seems sound.
Statistical murder flag, however, is interesting, denoting whether the shooting resulted in a death. So that can stay.

The values for sex are odd, at 5 and 3 for perpetrator and victim, respectively, so that may need further checking, but first let's clean up what we know we want to be rid of.

```r
drop_cols <- c("INCIDENT_KEY", "X_COORD_CD", "Y_COORD_CD",
               "Latitude", "Longitude", "Lon_Lat",
               "LOC_OF_OCCUR_DESC", "PRECINCT", "JURISDICTION_CODE",
               "LOC_CLASSFCTN_DESC")
```

```r
shooting_data <- shooting_data %>%
  mutate(`OCCUR_DATE` = mdy(`OCCUR_DATE`)) %>%
  select(-all_of(drop_cols))
```

Now that we have that out of the way, time to check the unique value in the columns that remain (not times or dates).

```r
# This boils down to: check distinct values for small-enough columns.
# There's probably a better way to do this?
shooting_data %>%
  select(-c("OCCUR_DATE", "OCCUR_TIME")) %>%
  sapply(unique)
```

```
## $BORO
## [1] "QUEENS"        "BRONX"         "BROOKLYN"      "MANHATTAN"
## [5] "STATEN ISLAND"
##
## $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"
##
## $STATISTICAL_MURDER_FLAG
## [1] FALSE  TRUE
##
## $PERP_AGE_GROUP
##  [1] NA        "25-44"   "UNKNOWN" "18-24"   "45-64"   "<18"     "65+"
##  [8] "940"     "(null)"  "224"     "1020"
##
## $PERP_SEX
## [1] NA       "M"      "U"      "F"      "(null)"
##
## $PERP_RACE
## [1] NA                               "BLACK"
## [3] "UNKNOWN"                        "BLACK HISPANIC"
## [5] "ASIAN / PACIFIC ISLANDER"       "WHITE HISPANIC"
## [7] "WHITE"                          "(null)"
## [9] "AMERICAN INDIAN/ALASKAN NATIVE"
##
## $VIC_AGE_GROUP
## [1] "18-24"   "25-44"   "<18"     "45-64"   "65+"     "UNKNOWN" "1022"
##
## $VIC_SEX
## [1] "M" "F" "U"
##
## $VIC_RACE
## [1] "BLACK"                          "WHITE"
## [3] "WHITE HISPANIC"                 "BLACK HISPANIC"
## [5] "ASIAN / PACIFIC ISLANDER"       "UNKNOWN"
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
```

Going through the list: - Boroughs look fine. - For locations, we can see "(null)" and "NONE" that can be wrapped up into NA. - Statistical murder flag looks fine. - Age groups are not great. For the PERP_AGE_GROUP, we have "UNKNOWN", "940", "(null)", "224", and "1020", which can (probably) be safely placed under NA. - Jumping to the VIC_AGE_GROUP, we see "UNKNOWN" that can be turned into NA (this may require checking how many NA values are present again…) and we see the "1022" code again. The best I can say with a Google search is maybe it's a Ten-code meaning "disregard", which doesn't fit - so we'll just say NA. - The sex fields can be boiled down to "M", "F", "NA/unknown" as well. - Racial information can also mix the "(null)" and "UNKNOWN" values into "NA".

We'll see how that affects our information in a bit.

```r
# Probably a better way again - but this time for sure.
# Maybe mutate?
shooting_data$LOCATION_DESC <- shooting_data$LOCATION_DESC %>%
  na_if("NONE")
shooting_data$LOCATION_DESC <- shooting_data$LOCATION_DESC %>%
  na_if("(null)")

shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  na_if("UNKNOWN")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  na_if("(null)")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  na_if("940")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  na_if("224")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  na_if("1020")

shooting_data$VIC_AGE_GROUP <- shooting_data$VIC_AGE_GROUP %>%
  na_if("UNKNOWN")
shooting_data$VIC_AGE_GROUP <- shooting_data$VIC_AGE_GROUP %>%
  na_if("1022")

shooting_data$PERP_SEX <- shooting_data$PERP_SEX %>%
  na_if("U")
shooting_data$PERP_SEX <- shooting_data$PERP_SEX %>%
  na_if("(null)")

# I guess unknown makes sense here, actually, but NA to match the perp.
# Maybe a mistake.
shooting_data$VIC_SEX <- shooting_data$VIC_SEX %>%
  na_if("U")

# Similar to sex above, since we have a victim, unknown makes sense.
# Still I march on.
shooting_data$VIC_RACE <- shooting_data$VIC_RACE %>%
  na_if("UNKNOWN")

shooting_data$PERP_RACE <- shooting_data$PERP_RACE %>%
  na_if("UNKNOWN")
shooting_data$PERP_RACE <- shooting_data$PERP_RACE %>%
  na_if("(null)")
```

Now just to re-examine and make sure I didn't break or miss anything. A good opportunity to revisit the null values and see if anything has become useless.

```r
sapply(shooting_data, function(x) n_distinct(x))
```

```
##              OCCUR_DATE              OCCUR_TIME                    BORO
##                    5761                    1421                       5
##           LOCATION_DESC STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                      39                       2                       6
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                       3                       7                       6
##                 VIC_SEX                VIC_RACE
##                       3                       7
```

```r
shooting_data %>%
  select(-c("OCCUR_DATE", "OCCUR_TIME")) %>%
  sapply(unique)
```

```
## $BORO
## [1] "QUEENS"        "BRONX"         "BROOKLYN"      "MANHATTAN"    
## [5] "STATEN ISLAND"
## 
## $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"                 "COMMERCIAL BLDG"          
## [11] "SMALL MERCHANT"            "BEAUTY/NAIL SALON"        
## [13] "FAST FOOD"                 "DRUG STORE"               
## [15] "TELECOMM. STORE"           "DRY CLEANER/LAUNDRY"      
## [17] "RESTAURANT/DINER"          "HOTEL/MOTEL"              
## [19] "SOCIAL CLUB/POLICY LOCATI" "SUPERMARKET"              
## [21] "CHAIN STORE"               "HOSPITAL"                 
## [23] "LIQUOR STORE"              "STORE UNCLASSIFIED"
## [25] "FACTORY/WAREHOUSE"         "DEPT STORE"
## [27] "SHOE STORE"                "VARIETY STORE"
## [29] "BANK"                     "ATM"
## [31] "DOCTOR/DENTIST"            "GYM/FITNESS FACILITY"
## [33] "CANDY STORE"               "VIDEO STORE"
## [35] "SCHOOL"                    "LOAN COMPANY"
## [37] "PHOTO/COPY STORE"          "CHECK CASH"
## [39] "STORAGE FACILITY"
##
## $STATISTICAL_MURDER_FLAG
## [1] FALSE  TRUE
##
## $PERP_AGE_GROUP
## [1] NA      "25-44" "18-24" "45-64" "<18"   "65+"
##
## $PERP_SEX
## [1] NA  "M" "F"
##
## $PERP_RACE
## [1] NA                               "BLACK"
## [3] "BLACK HISPANIC"                 "ASIAN / PACIFIC ISLANDER"
## [5] "WHITE HISPANIC"                 "WHITE"
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
##
## $VIC_AGE_GROUP
## [1] "18-24" "25-44" "<18"   "45-64" "65+"   NA
##
## $VIC_SEX
## [1] "M" "F" NA
##
## $VIC_RACE
## [1] "BLACK"                          "WHITE"
## [3] "WHITE HISPANIC"                 "BLACK HISPANIC"
## [5] "ASIAN / PACIFIC ISLANDER"       NA
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
```

Well, that looks tidier already.
Depending on how this turns out, I may need to either remove rows with NA values (victim age group being unknown could mean it was hard to tell or that we are not sure if there was a victim, for example) or change the values to something easier to visualize (changing the racial data to unknown might work better in a bar graph, for example). If it remains a problem, perhaps the column needs to go. Maybe the location of the shooting does not give us enough information to be useful even if logically it would (are ATMs frequent targets in a particular borough?).

Visualization, Analysis, and Modeling

"Add at least two different visualizations & some analysis to your Rmd. Does this raise additional questions that you should investigate?"

Keeping it short, and keeping away from Minority Report territory, let's stick to victim statistics.
Let's look at the rate of murders in New York over the data set as a function of time for our first visualization. The main question here is whether murder rates are going up or down as time goes on - or how safe are people?

For the second visualization, I want to see what kinds of people are victims of shootings in general (that is, murder or not, just involvement is enough to count), broken down by race, sex, and age group.
This will necessitate removing the NA/Unknown variables from the table.

Visualization #1: Murder rate over time

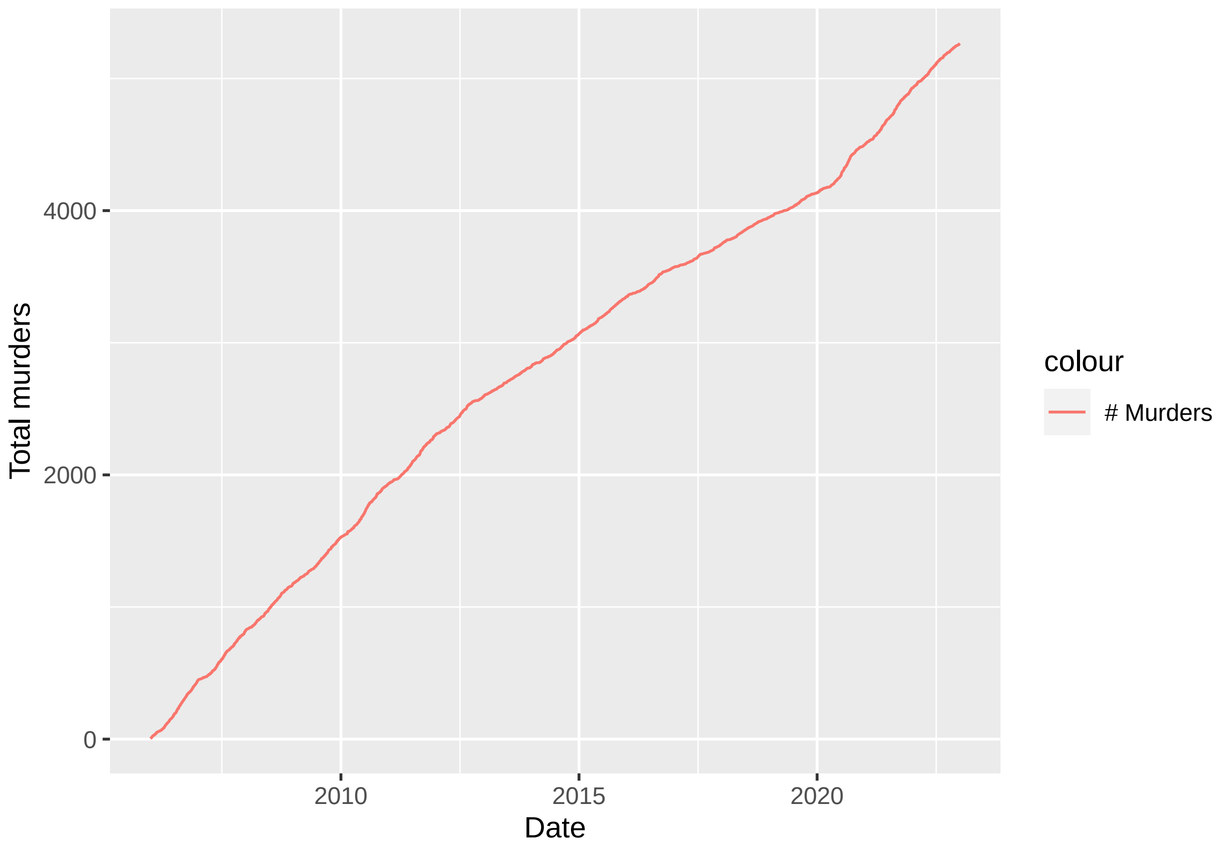
Some notes: I first visualized this over the whole time frame. Obviously there are many days in the set, and despite what some pearl-clutchers say, there are many days without a murder, so everything gets bunched together and is illegible.
As such, I opted to look at cumulative murders over the time - this has the benefit of showing a trend line that'll always be increasing, but the slope will tell us if things are getting better or worse.

```r
murders_per_day_NYC <- shooting_data %>%
  # Select only murders
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  # Only occurrence date matters - though now I feel like I oversimplified.
  group_by(OCCUR_DATE) %>%
  summarize(TOTAL_MURDERS = n())

# Just to get a vague idea before visualizing.
murders_per_day_NYC
```

```
## # A tibble: 2,886 x 2
##    OCCUR_DATE TOTAL_MURDERS
##    <date>             <int>
##  1 2006-01-01             4
##  2 2006-01-02             1
##  3 2006-01-03             1
##  4 2006-01-07             1
##  5 2006-01-08             1
##  6 2006-01-09             5
##  7 2006-01-14             4
##  8 2006-01-15             1
##  9 2006-01-16             1
## 10 2006-01-17             2
## # i 2,876 more rows
```

```r
# Actually plot the data.
murders_per_day_NYC %>%
  ggplot(aes(x = OCCUR_DATE, y = cumsum(TOTAL_MURDERS))) +
  # geom_point(aes(color = "# Murders")) +
  geom_line(aes(color = "# Murders")) +
  labs(x = "Date", y = "Total murders")
```



Well, it certainly looks more bleak when put this way than just looking at individual murders per day.
We can see some variations in slope, but it's more-or-less linear, with a notable bump around the time COVID rolled around.

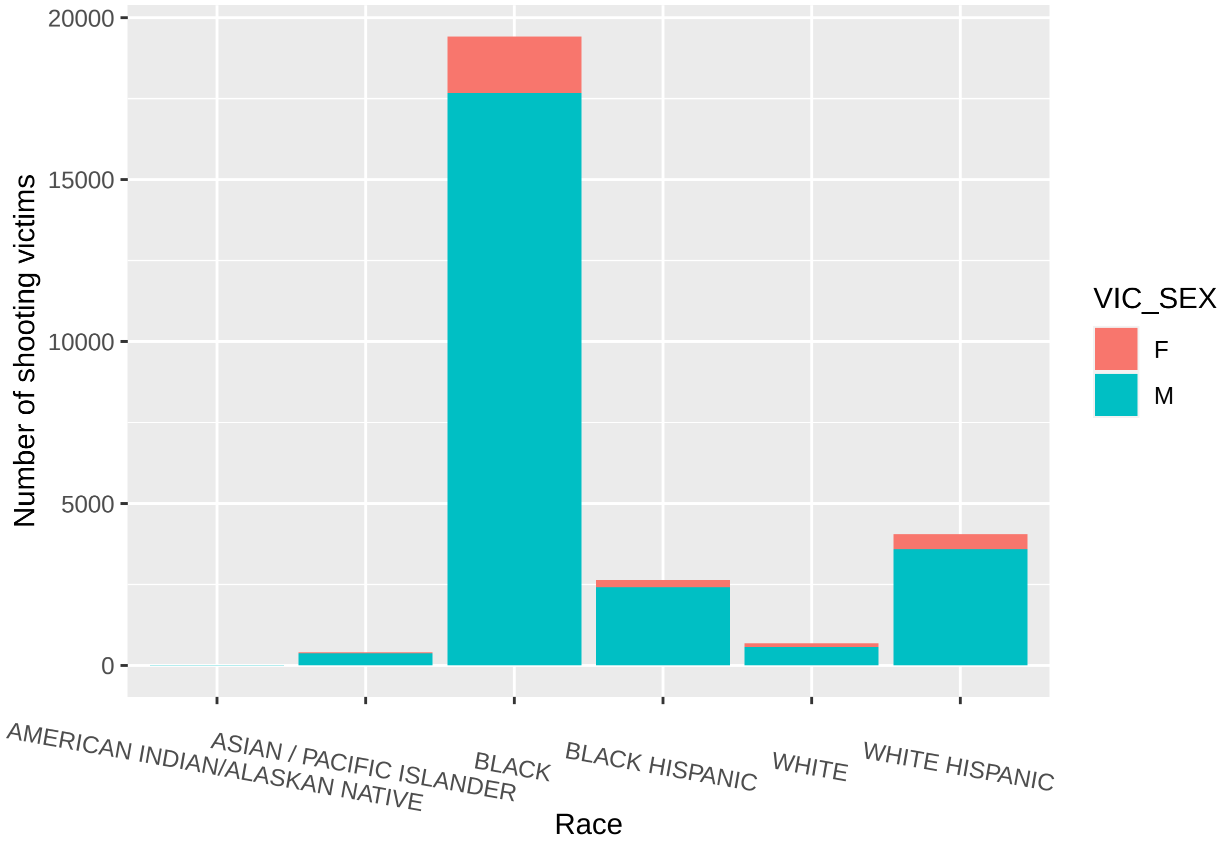
Some further questions that arise from even this simple visualization (and I'll keep it short, it just occurred to me that some of my peers are going to be skimming this, sorry for it being so long!) is what precisely contributed to the uptick? Does COVID just make people crazy? Probably not. Maybe more hurting for money, maybe more resources were dedicated to sending out police to shooting sites. How severe was the lock-down in the state? Where did these murders occur? For now, those questions will go unanswered.

Visualization #2: Victim statistics

Since the previous visualization raised a lot of questions despite its simplicity, I won't break the bank on this one.
I am interested in the kinds of people at the wrong end of shooting incidents. Are people of a certain race, sex, or age group more likely to get shot at?
Since I want to look at all of these together, I should remove any rows in the data where any of these are missing.

```r
shooting_data %>%
  # Only grab the info we care about for this question
  select(VIC_AGE_GROUP, VIC_RACE, VIC_SEX) %>%
  # ignoring the incomplete data
  drop_na() %>%
  # And create some bar graphs.
  # Man, I wish I was more familiar with this stuff than I am.
```

```r
ggplot(aes(x = VIC_RACE)) +
geom_bar(aes(fill = VIC_SEX)) +
# geom_bar(aes(x = VIC_AGE_GROUP)) +
theme(
      # label length led to overlap - this doesn't look great, but allows the information to show.
      axis.text.x = element_text(angle = -10)) +
labs(x = "Race", y = "Number of shooting victims")
```



At a glance, age grouping was not that interesting, pretty much what would be expected.
That is, most victims were between 18 and 44 with either extreme of the age group having few cases.
This cluttered the visualization significantly and felt like maybe it counted as a separate question, so it was removed.
Now, for what remains, looking at the data by race and sex, we see a huge number of victims are black, at a glance, more than all other groups combined. This raises many questions in itself.
Is this an over-representation based on racial makeup of the city? By what margin? How are groups determined (Who decides what is black versus black Hispanic, for instance)? Are these crimes limited to specific boroughs or even locations within boroughs (the location data we purged at the beginning may be of interest now)? What situations do these crimes happen in, and can they be attributed to something other than or co-morbid (?) with race? What do we even make of the lopsided sex results? Is this a function of population density? Poverty? How about how this compares to other states?

There's a lot to unpack here!

Bias Identification

“Write the conclusion to your project report and include any possible sources of bias. Be sure to identify what your personal bias might be and how you have mitigated that.”

Starting with the personal. I am quite left leaning (leaning might be too gentle a term). Upon seeing the name of the data set, I just assumed it was shootings involving officers as opposed to responses to shooting crimes, so that was a bad start. I am non-American, so while I don’t have direct exposure to the culture, my country is constantly inundated with news about the place, so ideas about crime rates and racial/gun politics for sure bleeds through. In fact, before I used the cumulative sum of murders, I was shocked with how low the daily deaths were, with many days without murders.
Some biases were confirmed, though. For example, the assumption that most crime is targeted towards minorities.
A bias I did not investigate is the assumption that most perpetrators are also minorities, specifically black males.
I mitigated these biases by looking only at victim information and total murders, I tried not to consider things like perpetrator information or splitting data up by borough (aha, these cops are lazier!), focusing only on the results.

In terms of bias from the data, much of the data was incomplete, likely based on the reports of the responding officers at the time, so it may not be entirely trustworthy. For instance, there were a few rows that had strange values for age ranges or unknown/missing information for age/race. I dealt with these just by ignoring those lines (there weren’t many, thankfully).