NYPD Shooting Data

4/22/2022

Background

New York City offers a vast amount of open data on their portal[1]. One such dataset, provided by the New York Police Department, provides information on all shooting incidents that have occurred within the city since 2006, up to the end of the previous calendar year. We will analyze this dataset to answer some questions, such as:

1. What areas of the city had the most shooting incidents?
2. Which age groups were most represented as perpetrators and victims?
3. Are shootings and murders trending upward or downward over time?

Load and Standardize

First, we load the data by URL and perform some simple transformations to make it easier to work with: dates become real date objects, and strings become factors.

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

```
nypd_data <- nypd_data %>%
                mutate(OCCUR_DATE = mdy(OCCUR_DATE),
                       BORO = as_factor(BORO),
                       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),
```

[1] https://data.cityofnewyork.us/

```
                    VIC_RACE = as_factor(VIC_RACE),
                    STATISTICAL_MURDER_FLAG = as_factor(STATISTICAL_MURDER_FLAG)
                ) %>%
                select(-c(X_COORD_CD, Y_COORD_CD, Lon_Lat, PRECINCT,
                          JURISDICTION_CODE, LOCATION_DESC))
```

Summary of Input Data

Let's take a quick look at the data we have so far.

```
summary(nypd_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms
##  Median : 86437258   Median :2012-08-26   Class2:difftime
##  Mean   :112382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238490103   Max.   :2021-12-31
##
##             BORO       STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  BRONX        : 7402   FALSE:20668             18-24  :5844   M   :14416
##  QUEENS       : 3828   TRUE : 4928             25-44  :5202   F   :  371
##  MANHATTAN    : 3265                           UNKNOWN:3148   U   : 1499
##  BROOKLYN     :10365                           <18    :1463   NA's: 9310
##  STATEN ISLAND:  736                           45-64  : 535
##                                                (Other):  60
##                                                NA's   :9344
##           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK         :10668   25-44  :11386   F: 2403
##  WHITE HISPANIC: 2164   65+    :  167   M:23182
##  UNKNOWN       : 1836   18-24  : 9604   U:   11
##  BLACK HISPANIC: 1203   <18    : 2681
##  WHITE         :  272   45-64  : 1698
##  (Other)       :  143   UNKNOWN:   60
##  NA's          : 9310
##                            VIC_RACE        Latitude       Longitude
##  BLACK HISPANIC                : 2485   Min.   :40.51   Min.   :-74.25
##  WHITE                         :  660   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :18281   Median :40.70   Median :-73.92
##  WHITE HISPANIC                : 3742   Mean   :40.74   Mean   :-73.91
##  ASIAN / PACIFIC ISLANDER      :  354   3rd Qu.:40.82   3rd Qu.:-73.88
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Max.   :40.91   Max.   :-73.70
##  UNKNOWN                       :   65
```

There are a number of data points in here that have "invalid" values. We will ignore those for this initial analysis, but we should investigate them more deeply before drawing firm conclusions. Not doing so could introduce bias into our results.

Shootings by Age Group

Next, we'll look at the shooting incidents aggregated by the age group of both the perpetrator and the victim. A bubble plot is a nice way to visualize this since we have three dimensions of data to represent.

```
age_groups <- c("<18", "18-24", "25-44", "45-64", "65+")

# Summarize the data into buckets by perpetrator and victim age groups
nypd_data_by_age <- nypd_data %>%
    group_by(PERP_AGE_GROUP, VIC_AGE_GROUP) %>%
    tally() %>%
    filter(PERP_AGE_GROUP %in% age_groups, VIC_AGE_GROUP %in% age_groups) %>%
    ungroup()

# Order the age groups sensibly
nypd_data_by_age$VIC_AGE_GROUP <-
  fct_relevel(nypd_data_by_age$VIC_AGE_GROUP, age_groups)
nypd_data_by_age$PERP_AGE_GROUP <-
  fct_relevel(nypd_data_by_age$PERP_AGE_GROUP, age_groups)

# Now, create a bubble plot with the data
nypd_data_by_age %>%
    ggplot(aes(x=PERP_AGE_GROUP, y=VIC_AGE_GROUP)) +
    geom_point(aes(size=n)) +
    scale_size(range = c(1, 15)) +
    labs(title = "NYPD Shootings by Age Group", x = "Perpetrator Age", y = "Victim Age") +
    theme_bw()
```

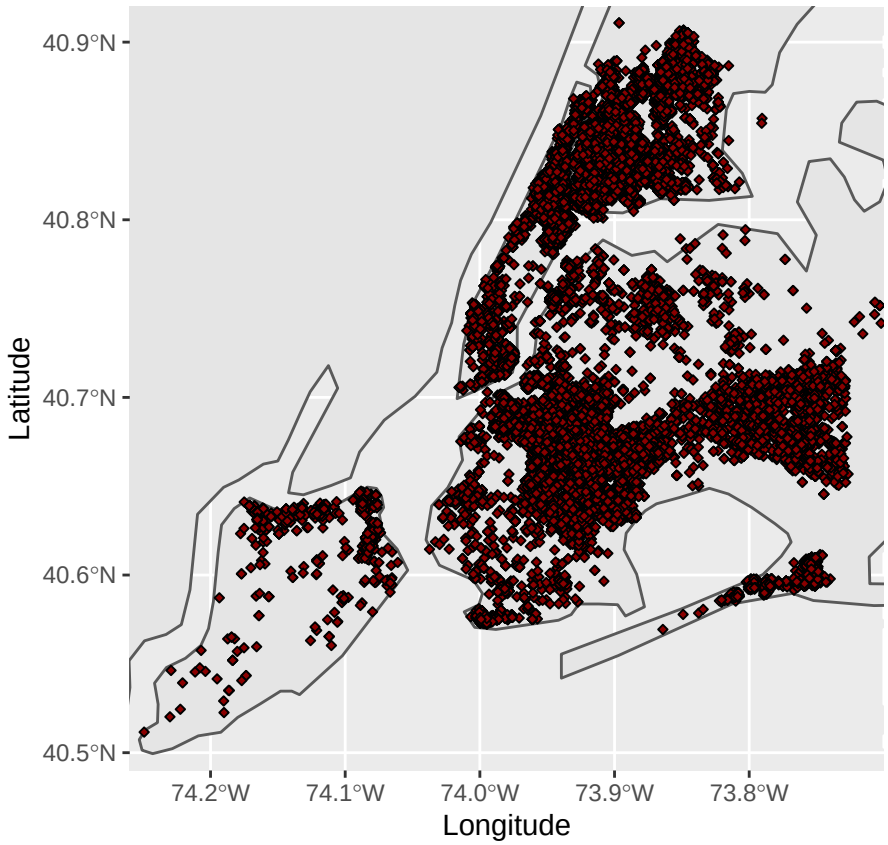
The majority of shooting incidents appear to be where both perpetrator and victim are in the 18-24 and 25-44 age groups, with a somewhat higher number of shootings where the perpetrator and the victim are in the same age group. This would be an interesting question for further investigation: how strongly does the perpetrator's age group predict the victim's age group?

Geographic Plot

Our data have latitude and longitude, which we can use to plot the shootings on a map.

```
world <- ne_countries(scale="large", returnclass="sf")

ggplot(data = world) +
     geom_sf() +
     geom_point(data=nypd_data, aes(x = Longitude, y = Latitude), size=1, shape=23, fill="darkred") +
     coord_sf(xlim=c(-74.26, -73.7), ylim=c(40.49, 40.92), expand = FALSE)
```

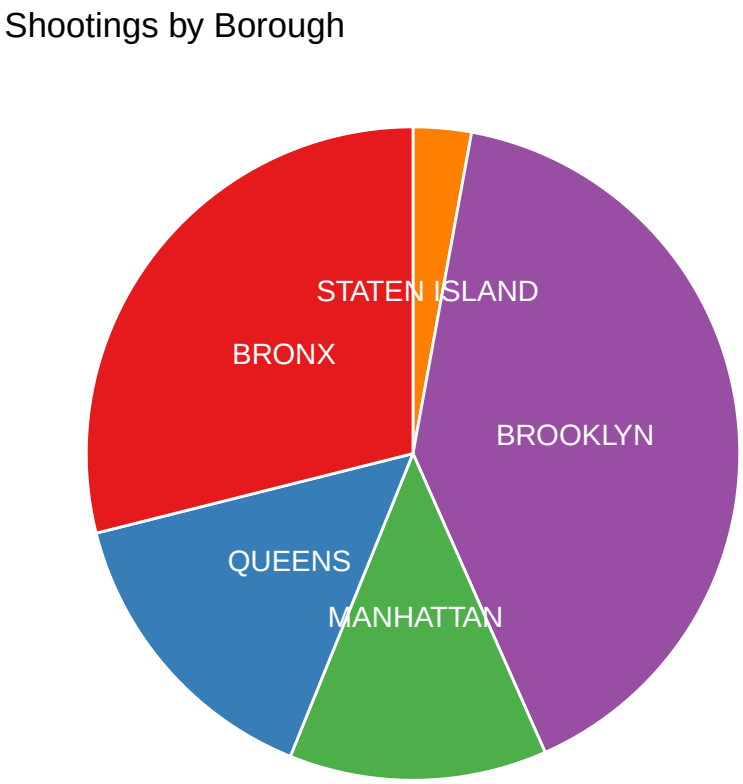


Shooting Incidents by Borough

Our map seems to show that the majority of shootings occurred in the Brooklyn and Queens boroughs. Let's count the shootings within each borough to see if that's true.

```
# Summarize the data by borough (counting the number of incidents in each)
nypd_data_by_boro <- nypd_data %>%
  group_by(BORO) %>%
  tally() %>%
  ungroup() %>%
  # these prepare our data for display in the pie chart
  arrange(desc(BORO)) %>%
  mutate(prop = n / sum(n) * 100) %>%
  mutate(ypos = cumsum(prop) - 0.5*prop)

# now generate the pie chart
nypd_data_by_boro %>%
  ggplot(aes(x="", y=prop, fill = BORO)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  theme_void() +
  labs(title = "Shootings by Borough") +
  theme(legend.position="none") +
  geom_text(aes(y = ypos, label = BORO), color = "white", size=4) +
  scale_fill_brewer(palette="Set1")
```



Indeed, those two boroughs do have da majority of the shootings in the data.

Why would that be the case? And, are there smaller geographical units (zip codes, precincts, census tracts, etc) that are over- or under-represented? These would be interesting questions to pursue.

Murders

According to the Open Data Network description of this data set[2], the STATISTICAL_MURDER_FLAG column is a boolean value that indicates that the "[s]shooting resulted in the victim's death which would be counted as a murder."

Let's plot the murders by month along with the numbers of shootings by month to see if they seem to correlate, and if they are trending upward or downward.

```
# Summarize the data by date and murder flag
nypd_months <- nypd_data %>%
  mutate(ym = floor_date(OCCUR_DATE, unit="month")) %>%
  group_by(ym, STATISTICAL_MURDER_FLAG) %>%
  add_count() %>%
  select(c(ym, STATISTICAL_MURDER_FLAG, n)) %>%
  arrange(ym, STATISTICAL_MURDER_FLAG) %>%
  summarize(shootings = n()) %>%
  select(ym, STATISTICAL_MURDER_FLAG, shootings) %>%
  rename(event = STATISTICAL_MURDER_FLAG)
```

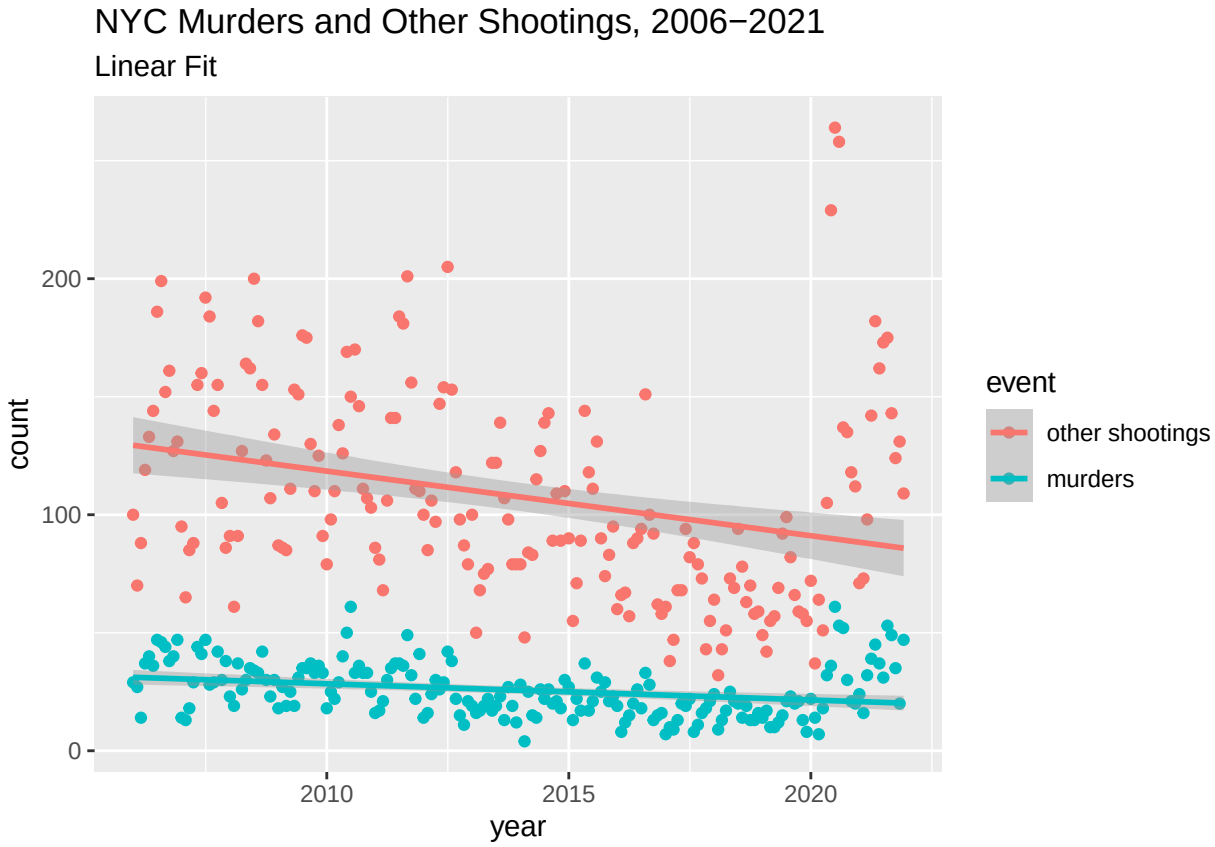
[2]https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/833y-fsy8

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

```
nypd_months$event <-
  recode(nypd_months$event, "TRUE" = "murders", "FALSE" = "other shootings")

ggplot(nypd_months, aes(x = ym, y = shootings, color = event)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(title = "NYC Murders and Other Shootings, 2006-2021", subtitle = "Linear Fit", x = "year", y="co
```

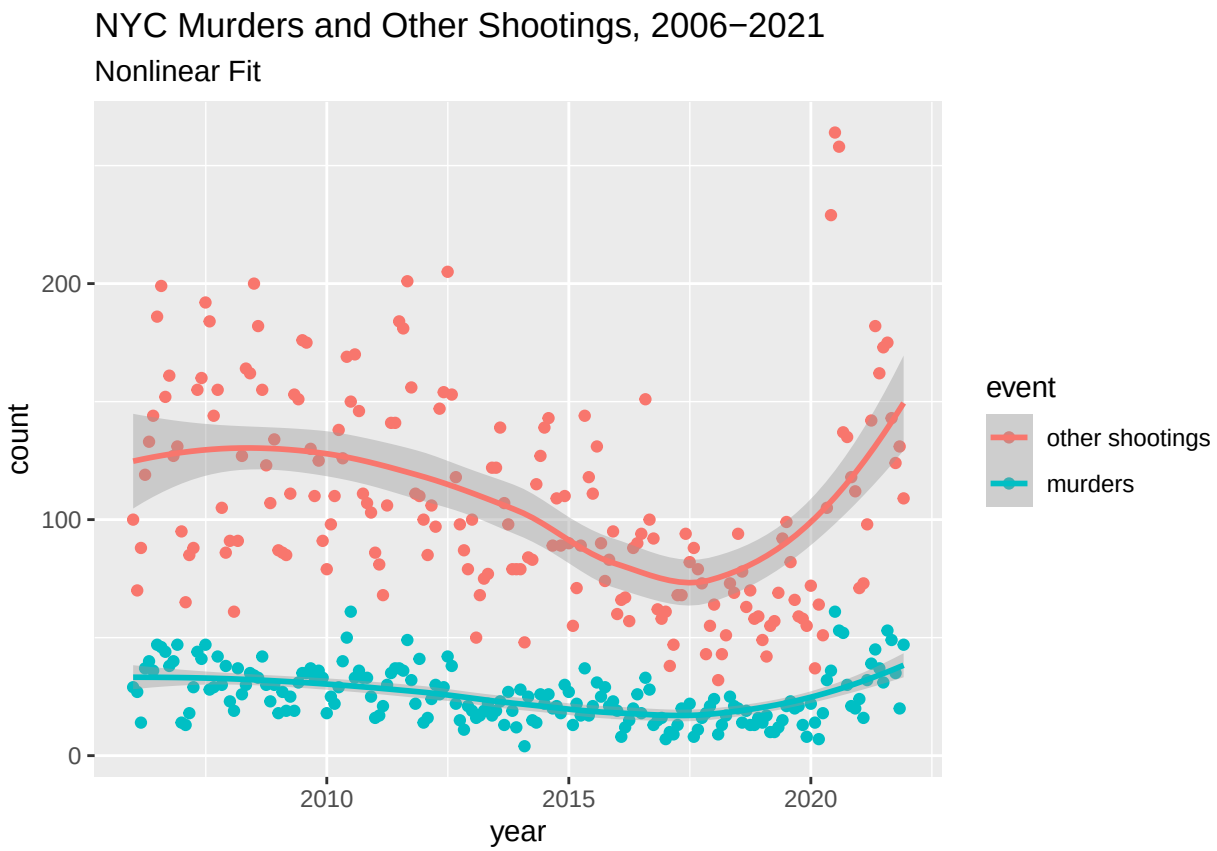
```
## `geom_smooth()` using formula 'y ~ x'
```



Clearly, both seem to be trending downward over the scope of the graph, using a linear fit. What if we don't require the fit to be linear? Perhaps that will confirm our perception of an upward trend at the end?

```
ggplot(nypd_months, aes(x = ym, y = shootings, color = event)) +
  geom_point() +
  geom_smooth() +
  labs(title = "NYC Murders and Other Shootings, 2006-2021", subtitle = "Nonlinear Fit", x = "year", y=
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Sure enough, these models confirm the recent upward trend in both murders and other shootings. And, yes, it seems like the numbers of murders and other shootings trend together.

Bias

There are many possible sources of bias in this analysis, including:

1. Each borough of the city may have its own standards about how/if shootings are recorded. For example, if Brooklyn has a very robust policy toward recording shootings whereas Staten Island's policy is more lax, we could erroneously conclude that Brooklyn has a disproportionate number.
2. There are many data points where the victim and/or perpetrator age groups are "Other," "Unknown," or "NA." Could it be that certain age groups are more likely than others to have one of these values?

Session Information

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
```

```
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] rnaturalearth_0.1.0 sf_1.0-7            lubridate_1.8.0
##  [4] forcats_0.5.1       stringr_1.4.0       dplyr_1.0.7
##  [7] purrr_0.3.4         readr_2.0.2         tidyr_1.1.4
## [10] tibble_3.1.5        ggplot2_3.3.5       tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2               splines_4.1.1            bit64_4.0.5
##  [4] vroom_1.5.5              jsonlite_1.7.2           modelr_0.1.8
##  [7] rnaturalearthhires_0.2.0 assertthat_0.2.1         sp_1.4-6
## [10] highr_0.9                cellranger_1.1.0         yaml_2.2.1
## [13] pillar_1.6.4             backports_1.3.0          lattice_0.20-44
## [16] glue_1.4.2               digest_0.6.28            RColorBrewer_1.1-2
## [19] rvest_1.0.2              colorspace_2.0-2         Matrix_1.3-4
## [22] htmltools_0.5.2          pkgconfig_2.0.3          broom_0.7.10
## [25] haven_2.4.3              s2_1.0.7                 scales_1.1.1
## [28] tzdb_0.2.0               proxy_0.4-26             mgcv_1.8-36
## [31] generics_0.1.1           farver_2.1.0             ellipsis_0.3.2
## [34] withr_2.4.2              cli_3.1.0                magrittr_2.0.1
## [37] crayon_1.4.2             readxl_1.4.0             evaluate_0.14
## [40] fs_1.5.0                 fansi_0.5.0              nlme_3.1-152
## [43] xml2_1.3.2               class_7.3-19             tools_4.1.1
## [46] hms_1.1.1                lifecycle_1.0.1          munsell_0.5.0
## [49] reprex_2.0.1             compiler_4.1.1           e1071_1.7-9
## [52] rlang_0.4.12             classInt_0.4-3           units_0.8-0
## [55] grid_4.1.1               rstudioapi_0.13          labeling_0.4.2
## [58] rmarkdown_2.13           wk_0.6.0                 gtable_0.3.0
## [61] DBI_1.1.2                curl_4.3.2               R6_2.5.1
## [64] knitr_1.36               fastmap_1.1.0            bit_4.0.4
## [67] utf8_1.2.2               KernSmooth_2.23-20       stringi_1.7.5
## [70] parallel_4.1.1           Rcpp_1.0.7               vctrs_0.3.8
## [73] dbplyr_2.1.1             tidyselect_1.1.1         xfun_0.30
```