NYPD Shooting Data

Data

Data Loading and Cleanup

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.1.2  magrittr_2.0.1  fastmap_1.1.0   tools_4.1.2
##  [5] htmltools_0.5.2 yaml_2.2.1      stringi_1.7.5   rmarkdown_2.11
##  [9] knitr_1.36      stringr_1.4.0   xfun_0.27       digest_0.6.28
## [13] rlang_0.4.12    evaluate_0.14
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
library(sf)
```

```
## Linking to GEOS 3.9.1, GDAL 3.2.1, PROJ 7.2.1
```

```
theme_set(theme_bw())
```

```
#read in data
url = 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'
data = as_tibble(read.csv(url))

#replace blanks with NA in PERP_AGE_GROUP and PERP_RACE
data = data %>% mutate(
    PERP_AGE_GROUP = na_if(PERP_AGE_GROUP, ""),
    PERP_RACE = na_if(PERP_RACE, "")
)

#create OCCUR_DATE_TIME column as datetime type from OCCUR_DATE and OCCUR_TIME
data = data %>% mutate(OCCUR_DATE_TIME = mdy_hms(paste(data$OCCUR_DATE, data$OCCUR_TIME)))

#Convert STATISTICAL_MURDER_FLAG from char to logical
data = data %>% mutate(STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG))

#Convert X and Y coords from chars to doubles
data = data %>% mutate(X_COORD_CD = as.numeric(gsub(",", "",data$X_COORD_CD)))
data = data %>% mutate(Y_COORD_CD = as.numeric(gsub(",", "",data$Y_COORD_CD)))

#replace erroneous perp_age_group values with ""
data = data %>% mutate(PERP_AGE_GROUP = replace(PERP_AGE_GROUP, PERP_AGE_GROUP == "940", ""))
data = data %>% mutate(PERP_AGE_GROUP = replace(PERP_AGE_GROUP, PERP_AGE_GROUP == "1020", ""))
data = data %>% mutate(PERP_AGE_GROUP = replace(PERP_AGE_GROUP, PERP_AGE_GROUP == "224", ""))

#convert perp_age_group to factor
data = data %>% mutate(PERP_AGE_GROUP = as_factor(PERP_AGE_GROUP))

#convert perp_sex to factor
data = data %>% mutate(PERP_SEX = as_factor(PERP_SEX))

#convert perp_race to factor
data = data %>% mutate(PERP_RACE = as_factor(PERP_RACE))
```

```r
#convert VIC_AGE_GROUP, VIC_SEX, and VIC_RACE to factors
data = data %>% mutate(
    VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
    VIC_SEX = as.factor(VIC_SEX),
    VIC_RACE = as.factor(VIC_RACE)
)

#convert BORO to factor
data = data %>% mutate(BORO = as.factor(BORO))

#rearrange columns, drop LON_LAT, OCCUR_DATE, OCCUR_TIME
data = data %>% select(INCIDENT_KEY, OCCUR_DATE_TIME, everything(), -Lon_Lat, -OCCUR_DATE, -OCCUR_TIME)

summary(data)
```

```
##   INCIDENT_KEY       OCCUR_DATE_TIME                          BORO
##  Min.   :  9953245   Min.   :2006-01-01 02:00:00   BRONX        :6700
##  1st Qu.: 55317014   1st Qu.:2008-12-30 04:27:00   BROOKLYN     :9722
##  Median : 83365370   Median :2012-02-26 03:35:00   MANHATTAN    :2921
##  Mean   :102218616   Mean   :2012-10-04 05:23:12   QUEENS       :3527
##  3rd Qu.:150772442   3rd Qu.:2016-02-28 00:01:00   STATEN ISLAND: 698
##  Max.   :222473262   Max.   :2020-12-31 23:45:00
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  Mean   : 66.21   Mean   :0.3323
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP PERP_SEX           PERP_RACE    VIC_AGE_GROUP   VIC_SEX
##  18-24  :5448    : 8425   BLACK          :9855   <18    : 2525   F: 2195
##  25-44  :4613   M:13305   WHITE HISPANIC:1961   18-24  : 9000   M:21353
##  UNKNOWN:3156   F:  334   UNKNOWN        :1869   25-44  :10287   U:   20
##  <18    :1354   U: 1504   BLACK HISPANIC:1081   45-64  : 1536
##  45-64  : 481             WHITE          : 255   65+    :  155
##  (Other):  57             (Other)        : 122   UNKNOWN:   65
##  NA's   :8459             NA's           :8425
##                            VIC_RACE       X_COORD_CD        Y_COORD_CD
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   : 914928   Min.   :125757
##  ASIAN / PACIFIC ISLANDER      :  320   1st Qu.: 999900   1st Qu.:182565
##  BLACK                         :16846   Median :1007645   Median :193482
##  BLACK HISPANIC                : 2244   Mean   :1009363   Mean   :207312
##  UNKNOWN                       :  102   3rd Qu.:1016807   3rd Qu.:239163
##  WHITE                         :  615   Max.   :1066815   Max.   :271128
##  WHITE HISPANIC                : 3432
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
```

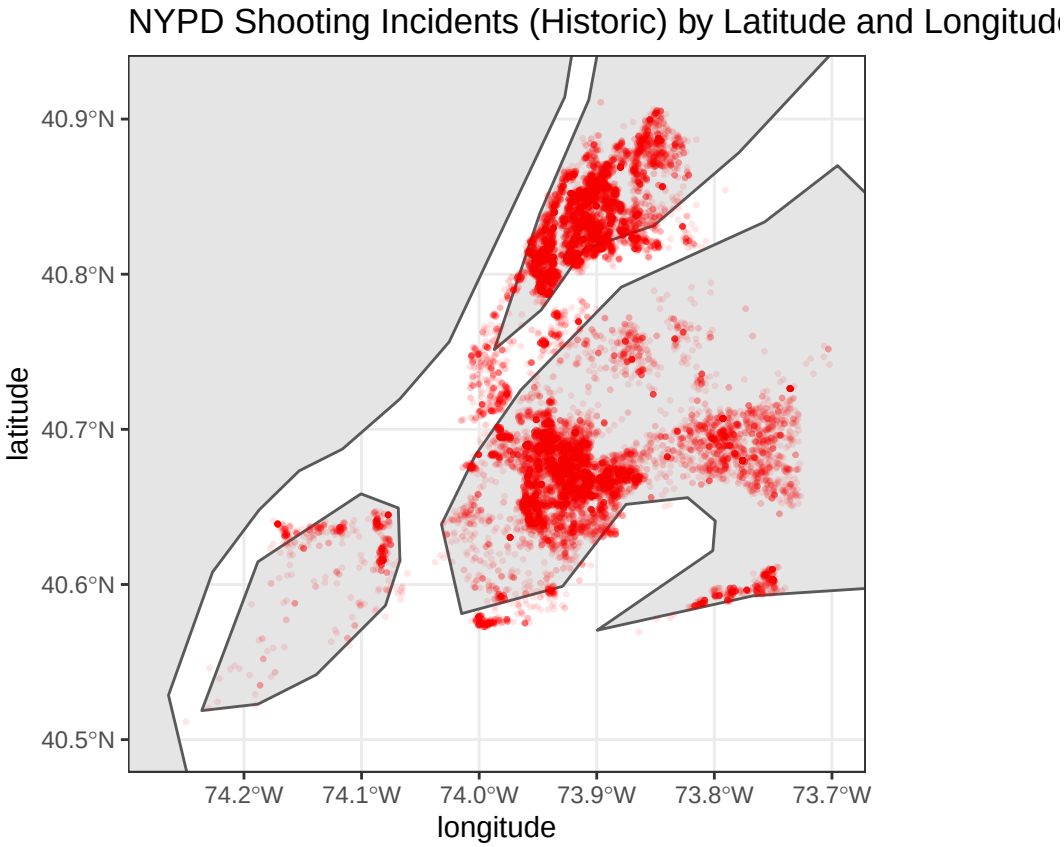
```
##  Max.   :40.91   Max.   :-73.70
##
```

A few rows had obviously erroneous age information, which was replaced with empty strings " ".

Many rows had missing information in the PERP_RACE and / or PERP_AGE_GROUP columns. I replaced the missing values with NA's. I left the value "UNKNOWN" in place because it is not necessarily the same as an empty string / missing information, so to change it would be to lose or skew the data.

Plots

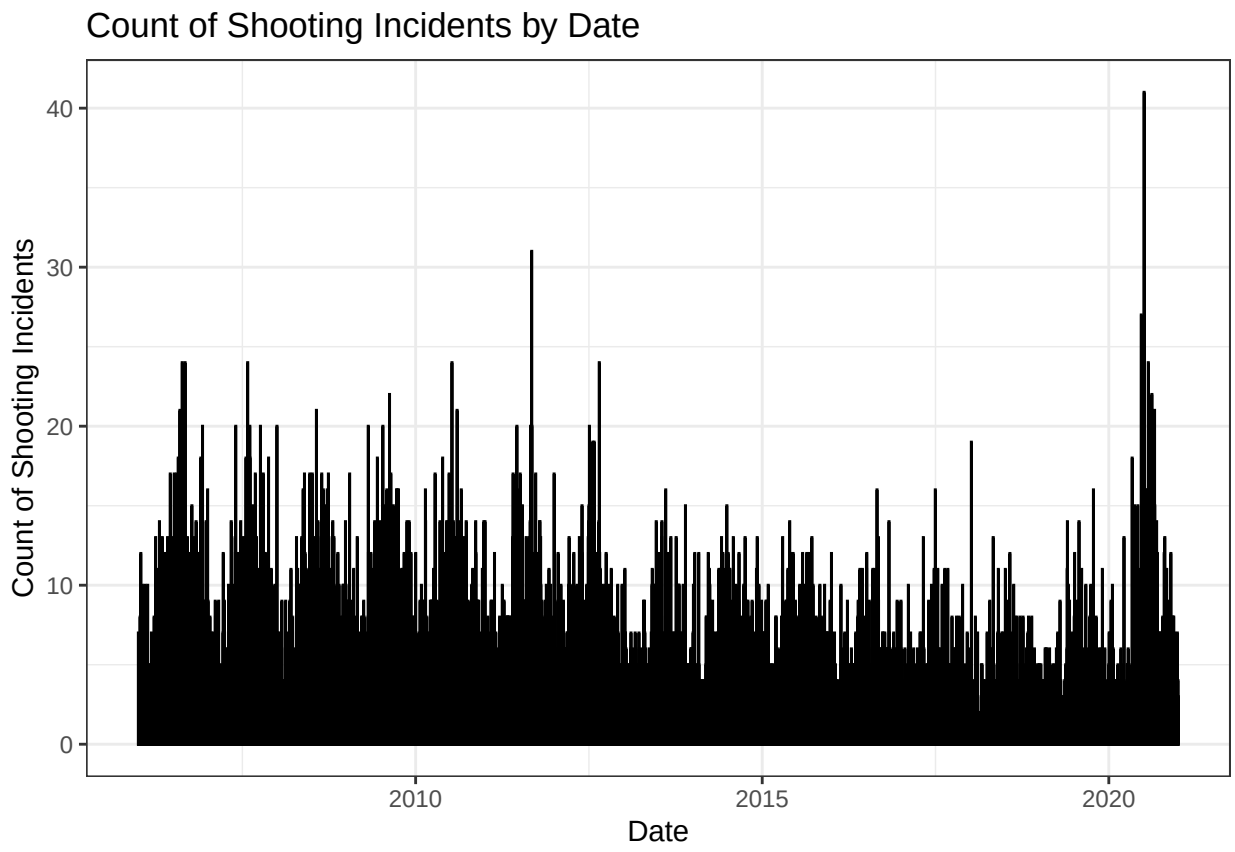
```
#ggplot(data, aes(y=Latitude, x=Longitude))+geom_point(alpha = .1, color = "red", size = .5) + labs(tit

library("rnaturalearth")
library("rnaturalearthdata")

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

sites = data.frame(latitude = data$Latitude, longitude = data$Longitude)
ggplot(data = world)+
     geom_sf() +
     geom_point(data = sites, aes(x=longitude, y = latitude),alpha = .1, color = "red", size = .5)+
     coord_sf(xlim = c(-74.27, -73.7), ylim = c(40.5, 40.92), expand = TRUE) +
  labs(title = "NYPD Shooting Incidents (Historic) by Latitude and Longitude")
```



This visualization indicates that shootings are not evenly spread across the city and instead are concentrated in at least two distinct areas. Some additional questions that this raises:

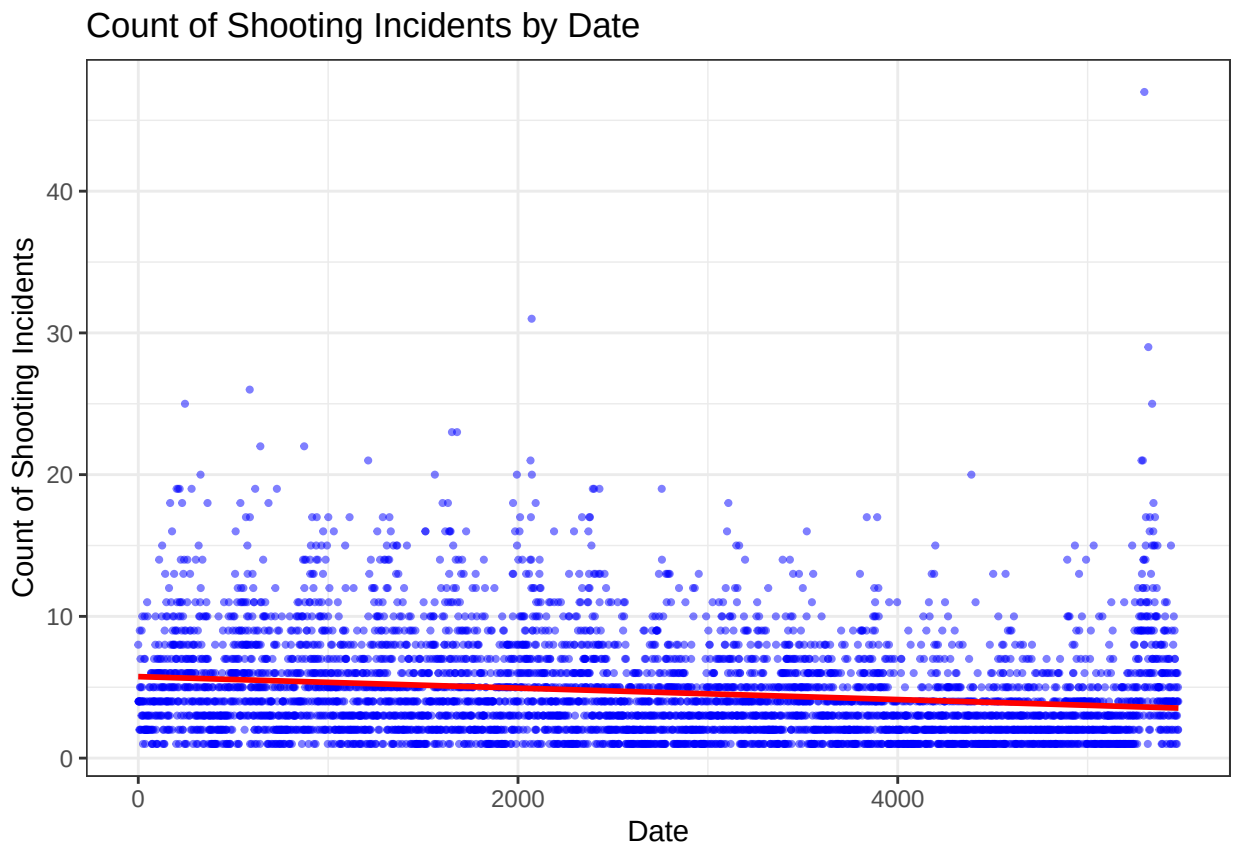
- Are there factors other than location that correlate with the number of shootings?
  - Average income?
  - Racial composition?
  - Age composistion?
- Do the disparities persist when the number of shootings is normalized by population

```
#hist(data$OCCUR_DATE_TIME, breaks = "days", freq = TRUE, xlab = "Date", ylab = "Number of Shootings",

dates = data$OCCUR_DATE_TIME
ggplot(data = data, aes(x=dates)) +
     geom_histogram(bins = 5479, color = "black") +
     ggtitle("Count of Shooting Incidents by Date") +
     xlab("Date") +
     ylab("Count of Shooting Incidents")
```




```
     #geom_density(alpha = .2, fill="#FF6666")

data2 = data %>% mutate(date = date(OCCUR_DATE_TIME)) %>%
    select(date, PRECINCT)
date_data = data2 %>%
    group_by(date) %>%
    summarize(count = n())
date_data = date_data %>% mutate(day = as.numeric(date(date) - min(date)))
#date_data = date_data %>% mutate(pred = predict(mod))
date_data %>% ggplot(aes(day, count)) +
    geom_point(aes(x=day, y = count), color = "blue", alpha = .5, size = .75) +
    geom_smooth(method = "lm", color = "red") +
    ggtitle("Count of Shooting Incidents by Date") +
     xlab("Date") +
     ylab("Count of Shooting Incidents")
```

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




```
    #geom_text()
```

This visualization shows that there is a strong periodicity to the number of shootings, as well as a multi-year downward trend that was reversed in 2020. Some additional questions:

- What is the cause of the periodicity?
  - Weather?
  - School schedules?
  - Seasonal employment cycles?
  - Sports seasons?
- What caused the spike in 2020? COVID seems like a likely explanation, but what specific facet of the pandemic caused an increase in shootings?

Bias

Broadly speaking, there are two potential sources of bias: biased **data** or biased **analysis**.

Data Bias

The data could be biased in a number of ways:

- Bias in which incidents are recorded. It is possible that some precincts are less likely to respond to shooting incidents, or less likely to find a perpetrator if they do. Perhaps there are events which could or could not be classified as shooting incidents based on the discretion of the responding officer.
- Bias in racial categorization of perpetrators and / or victims. Age categorization is obviously objective, but racial categorization is not necessarily so. Do responding officers perform the racial categorization or do the victims and perpetrators self identify? Are the given racial categories the most relevant and appropriate? Could some categories be combined or split?
- Bias in police coverage. If some areas have a higher police presence they might appear to have more shootings than other, more lightly policed areas, even if the two areas have the same rate of shooting incidents, simply because there are more police to respond to and record events in the first area.

Analysis Bias

The analysis of the data could also be biased in a number of ways:

- Bias in data selection. In this case the data was pre-determined, but in general the selection of data can introduce bias. Which data sources are deemed trustworthy?
- Bias in data cleanup. Handling missing, incomplete, or erroneous data is left to the discretion of the analyst, which means it is a potential source of bias. Are outliers left in or removed? How is missing data handled: is the entire row ignored or are blank columns tolerated? Either decision will affect the final analysis.
- Bias in goal of analysis. Generally, data does not speak for itself. The analyst must decide what questions they are trying to answer with a given set of data, and that decision will almost certainly introduce bias. Why study shootings rather than petty theft or embezzlement? Should the purpose of the analysis be to find out the root causes of shootings so a governmental body can minimize them, or to provide a guide on how an individual can best avoid danger? There are no wrong answers to these questions, but any answer will introduce some of the analyst's own bias.

Personally, I am skeptical of race-based explanations for crime; I think material / economic conditions have much more explanatory power. In this case, I chose not to analyze the data using race at all, but if I was required to, for whatever reason, I would make sure that I gave as much attention to racial factors as I did to economic ones.