NYPD Shootings Historic

Joe Harter

2024-04-07

Pre-requisites

```
#Needed for the map
install.packages("tmap")
install.packages("sf")
# See Installation at https://r-tmap.github.io/tmap/ if you encounter any issues.

# Needed for NYC Geographical Data
install.packages("remotes")
remotes::install_github("mfherman/nycgeo")

#If you want to use the calendar heat map
install.packages("lattice)
```

Tidying the data

Questions to Answer As I got more familiar with the data and comfortable with R there were 3 questions I wanted to answer: 1. Does time of the year have an effect on the number of shootings? 2. Are the shootings grouped in certain precincts? 3. Does population size effect the number of shootings?

```r
shooting_data <-
  read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD",
           show_col_types=FALSE)
clean_shooting_data <- shooting_data %>% select(OCCUR_DATE, PRECINCT)
```

What data am I working with? The City of New York releases a lot of data through their open data project. I am looking specifically at their shooting data which lists the location, date, time, victim, and perpetrator (if known). For the visualizations and analysis I really only need to deal with the date and the precinct, but one could think of all kinds of interesting analysis to do with this data.

Also of note is the geographical data I used from nycgeo to display the map.

Picking Columns I struggled with this part because I didn't know what data I'd be interested in so it all felt fine for me but once I decided I cared about dates and maybe something based on the police precinct I could drop lot of columns. I'm interested in reasons I need to do this as I could always select later which columns I needed but for the sake of the assignment here is me only selecting the columns I want:

Data Validation In addition to choosing my columns I should consider whether the data I have is valid and requires any adjustment.

You'll see below that I want the month name so I use lubridate's `month` function to turn the date into a month label. Also, because I care about comparing months I want to make sure I have a full dataset. The website suggested I did but to confirm I ran:

```
min(clean_shooting_data$OCCUR_DATE)
```

```
## [1] "01/01/2006"
```

... and to make sure I ended on Dec 31st I ran:

```
max(clean_shooting_data$OCCUR_DATE)
```

```
## [1] "12/31/2022"
```

So it looks like my first year starts on 1/1 and my last year ends on 12/31. I'm only focusing on numbers of shootings beyond that so I wasn't worried about any NAs in this case.

One thing I noticed later on was an outlier in the number of shootings on one date. July 5th 2020 had 47 shootings. I momentarily considered removing this as an outlier, but this didn't appear to be from a single mass shooting so I thought it best to keep in the dataset.

Visual Representation of shootings by date

Heatmap False Start My original thought was to see if there is a pattern by season or month so I looked into different calendar heat maps. I found calendarHeat.R which uses ggplot2 under the hood I believe.

So my first try was to just get the shooting data to display using this method. (I won't evaluate it for this Rmd file but if you want to evaluate it then change eval=FALSE to eval=TRUE)

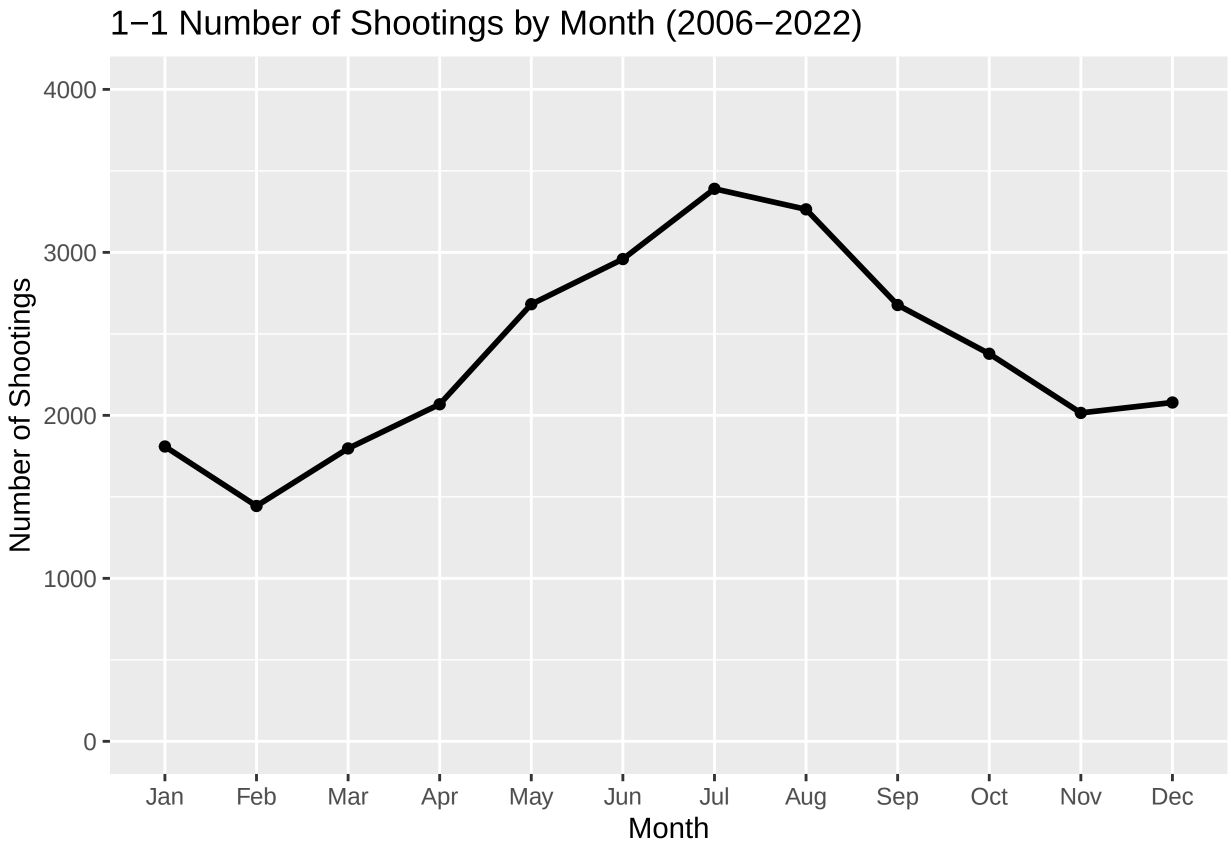
```
source("https://raw.githubusercontent.com/iascchen/VisHealth/master/R/calendarHeat.R")
# Group all shooting by a date
shooting_data_by_date <- shooting_data %>% group_by(OCCUR_DATE) %>%
  summarise(total_shootings = n())

# Set a gradient so the more shootings, the more red a cell is
g2r <- c('#00FF00', '#FF0000')

# Generates a calendar heat map
calendarHeat(mdy(shooting_data_by_date$OCCUR_DATE),
             shooting_data_by_date$total_shootings,
             ncolors = 99,
             color = 'g2r',
             varname="Shootings in New York City by Date")
```

But I encountered a couple things I didn't like:

- A different block of cells is generated for each year, but I really want to see a generic 'year' Jan 1st to Dec 31st and whether shootings group around any days. I don't want to see this year after year.
- The output does some kind of paging instead of prints all of them together.
- There were outlines for some of the years but not others.
- Apparently a calendar heat map might not be the best tool for what I want. (At least according to https://www.columbia.edu/~sg3637/blog/Time_Series_Heatmaps.html where it was suggested I use a simple line chart for what I'm after)

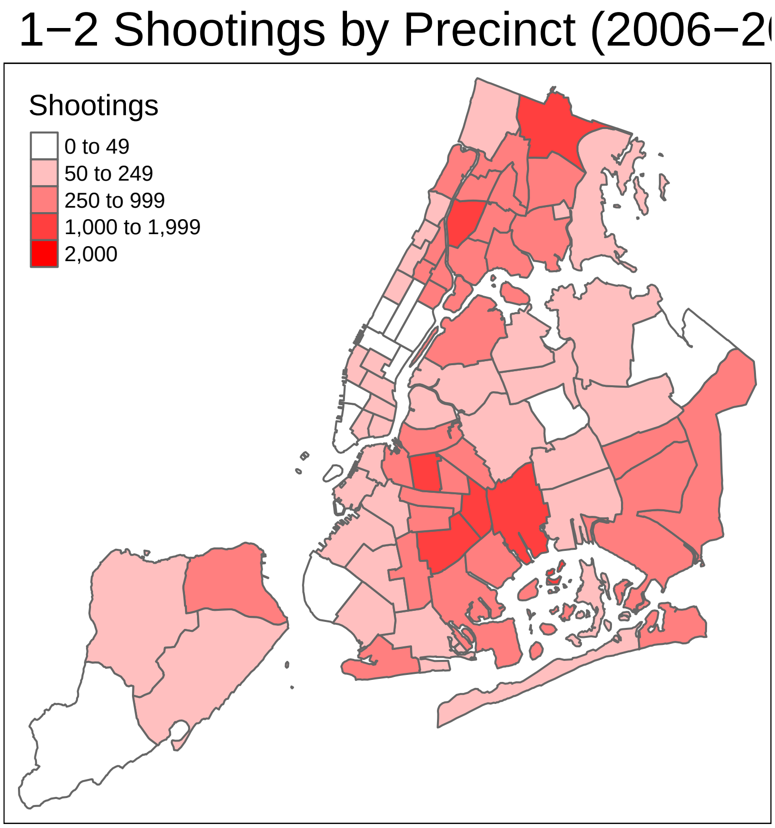
Line Chart So I decided to try a simple line chart. I'll group all of the shootings by the month they occurred and plot that:

```r
# I'm not sure if this is the best way to get labels but using lubridates month method
# I could label a column and the group by that column. The line plot then ordered it
# correctly without needing to include a month_value column
shooting_data_by_month <- shooting_data %>%
  mutate(month=month(mdy(OCCUR_DATE), label=TRUE)) %>%
  select(month, everything()) %>% group_by(month) %>%
  summarise(total_shootings = n())

# This will actually plot the shootings by month
# By default it set a floor of around 1200 shootings which made the scale and
# differences some what misleading in my opinion so I set the floor to 0
ggplot(data=shooting_data_by_month, aes(x=month, y=total_shootings, group=1)) +
  geom_line(linetype="solid", size=1) +
  geom_point() +
  ylim(0, 4000) +
  ggtitle("1-1 Number of Shootings by Month (2006-2022)") +
  ylab("Number of Shootings") +
  xlab("Month")
```



Precinct Map Seeing a visual representation of shootings by police precinct was interesting to me.

```
map_nyc_precincts <- nyc_boundaries(geography = "police")
map_nyc_precincts <- map_nyc_precincts %>% select("police_precinct_id", "geometry")
shooting_precinct_count <- clean_shooting_data %>% group_by(PRECINCT) %>% summarise(total_shootings = n

# Rename the column names to make joining easier
colnames(shooting_precinct_count) <- c("police_precinct_id", "shooting_count")

# map_nyc_precincts has precincts as characters so this should too
shooting_precinct_count$police_precinct_id <-
  as.character(shooting_precinct_count$police_precinct_id)

# Now add the number of shootings as a column where the precinct id matches
merged_shooting_data <- left_join(map_nyc_precincts,
                                  shooting_precinct_count,
                                  by="police_precinct_id")

# tmap will easily now make a map of all of the police precincts color coded by number of shootings
tm_shape(merged_shooting_data) +
    tm_polygons("shooting_count",
                title="Shootings",
                breaks = c(0, 50, 250, 1000, 2000, 2000),
                palette=c("white", "red"),
                legend.title="Shootings") +
    tm_layout(main.title = "1-2 Shootings by Precinct (2006-2022)")
```



1−2 Shootings by Precinct (2006−2(

Model the data

How many shootings will New York likely have in 10 years? One variable likely to impact that would be the population. So let's model the data.

I originally was toying around with using census data but in the end didn't need to import all of that information because I just needed the total population for 2000, 2010, and 2020 to create a predictive population model. So I googled the census information for 2000 (source) and 2010 and 2020 (source)

I now have 3 data points for total NYC population. I'll do a linear regression on that data to then be able to predict the population in New York City for the years in between to model the shootings against.

```
nyc_population_2000 = 8008278
nyc_population_2010 = 8175133
nyc_population_2020 = 8804190

# Not every year is the same length due to leap years so I used days since the
# 2000 census instead of years
days_since_2000_census <- c(0, 3652, 7305)
population <- c(nyc_population_2000, nyc_population_2010, nyc_population_2020)
data = data.frame(days_since_2000_census, population)
colnames(data) = c ("days_since_2000_census", "population")
rownames(data)=c("2000", "2010", "2020")

# Using historical data we'll make a linear regression to help predict future
# population
population_mod = lm(population ~ days_since_2000_census, data=data)
```

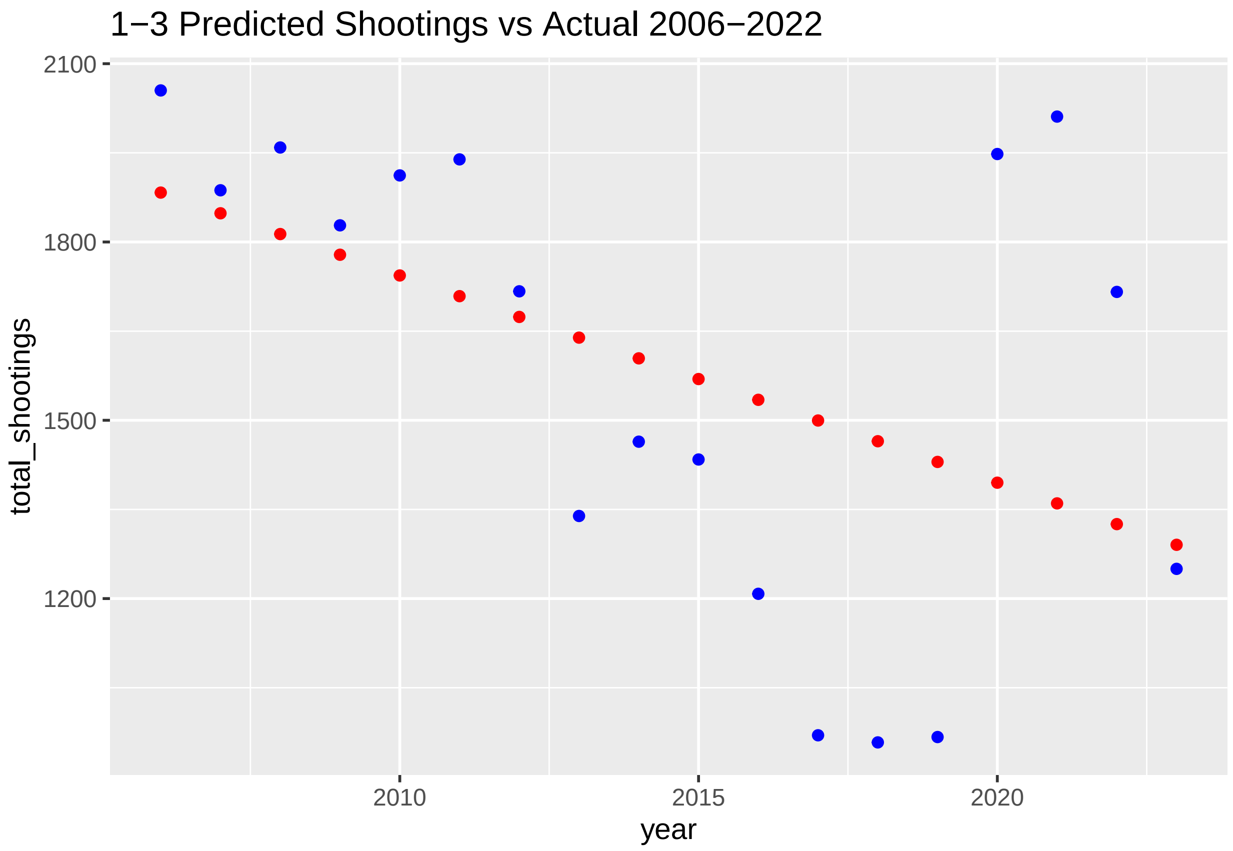
```
# Total number of shootings by year
shooting_data_by_year <- shooting_data %>%
  mutate(year=year(mdy(OCCUR_DATE))) %>%
  select(year, everything()) %>%group_by(year) %>%
  summarise(total_shootings = n())
shooting_data_by_year <- shooting_data_by_year %>%
  mutate(days_since_2000_census=as.numeric( as.Date(sprintf("%s-12-31", year))-as.Date("2000-04-01")))

# include the predicted population for that year
shooting_data_by_year <- shooting_data_by_year %>%
  mutate(pred_population = predict(population_mod,
                                    data.frame(days_since_2000_census=days_since_2000_census)))

# Now make a new linear regression for the shootings per population
shooting_mod = lm(total_shootings ~ pred_population, data=shooting_data_by_year)
summary(shooting_mod)
```

Summary of the linear regression (shootings to predicted population)

```
##
## Call:
## lm(formula = total_shootings ~ pred_population, data = shooting_data_by_year)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -529.60 -260.10   40.91  170.99  650.85
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.067e+03  3.454e+03   2.625   0.0184 *
## pred_population -8.760e-04  4.044e-04  -2.166   0.0457 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 354.2 on 16 degrees of freedom
## Multiple R-squared:  0.2268, Adjusted R-squared:  0.1785
## F-statistic: 4.693 on 1 and 16 DF,  p-value: 0.04573
```

```
# Add a new column showing the prediction to then plot the points
shooting_data_by_year <- shooting_data_by_year %>%
  mutate(shooting_pred=predict(shooting_mod, data.frame(pred_population=pred_population)))
shooting_data_by_year %>% ggplot() +
  geom_point(aes(x=year, y=total_shootings), color="blue") +
  geom_point(aes(x=year, y=shooting_pred), color="red") +
  ggtitle("1-3 Predicted Shootings vs Actual 2006-2022")
```



Conclusion

There were a few questions: 1. Does time of the year have an effect on the number of shootings? 2. Are the shootings grouped in certain precincts? 3. Does population size effect the number of shootings?

1. Does time of the year have an effect? Given the analysis we did to generate figure 1-1 we saw that the warmer months generally had more shootings than colder months with a peak in July and a trough in February. Getting more specific temperature data would be interesting.

2. Are the shootings grouped in certain precincts? Our map visualization in figure 1-2 shows that there are indeed some precincts that are dealing with more shootings than others. Normalizing the data to use shootings / 1000 would have been interesting and useful.

3. Does population size effect the number of shootings? There does not appear to be a correlation between population and number of shootings based on this analysis given the p-value > .05. (See figure 1-3) I think looking at other variables like income or number age of the population might be interesting.

Bias in the results. At least when looking at the specific variables and data there doesn't appear to be much concern for bias, but we should consider that this data was collected by humans and perhaps there is under-reporting and over-reporting based on the biases of the people who conducted the census and tabulated the shooting information. It's worth noting that racial data is often used when analyzing shootings and I did avoid looking at that because it feels uncomfortable. That in itself is a type of bias in the level of analysis I wanted to conduct.