NYPD Shooting Data Project

M Rho

2022-09-24

```
library(tidyverse)
library(lubridate)
```

Load and preview The Data

Make sure you have the data file in the same directory as your RMD file.

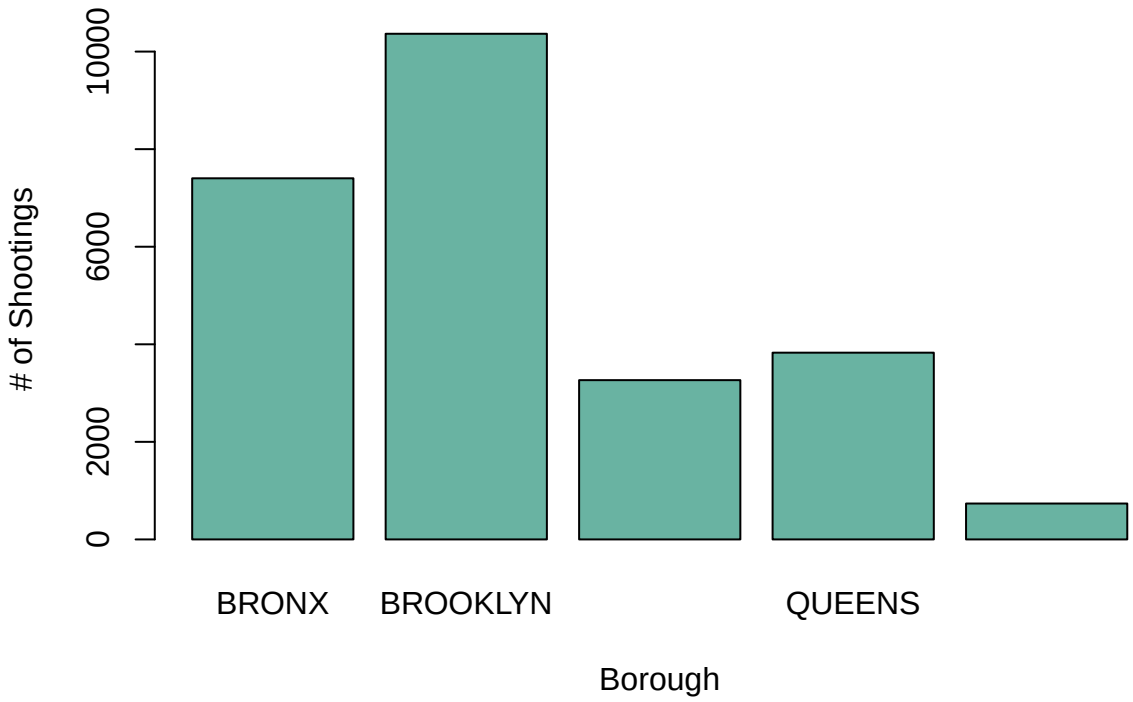
```
nyc_shootings <- read.csv("NYPD_Shooting_Incident_Data__Historic_.csv", header=TRUE, stringsAsFactors=FA
glimpse(nyc_shootings)
```

```
## Rows: 25,596
## Columns: 19
## $ INCIDENT_KEY            <int> 236168668, 231008085, 230717903, 237712309, 22~
## $ OCCUR_DATE              <chr> "11/11/2021", "07/16/2021", "07/11/2021", "12/~
## $ OCCUR_TIME              <chr> "15:04:00", "22:05:00", "01:09:00", "13:42:00"~
## $ BORO                    <chr> "BROOKLYN", "BROOKLYN", "BROOKLYN", "BROOKLYN"~
## $ PRECINCT                <int> 79, 72, 79, 81, 113, 113, 42, 52, 34, 75, 32, ~
## $ JURISDICTION_CODE       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0~
## $ LOCATION_DESC           <chr> "", "", "", "", "", "", "COMMERCIAL BLDG", "",~
## $ STATISTICAL_MURDER_FLAG <chr> "false", "false", "false", "false", "false", "~
## $ PERP_AGE_GROUP          <chr> "", "45-64", "<18", "", "", "", "", "", "", "2~
## $ PERP_SEX                <chr> "", "M", "M", "", "", "", "", "", "", "M", "M"~
## $ PERP_RACE               <chr> "", "ASIAN / PACIFIC ISLANDER", "BLACK", "", "~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "25-44", "25-44", "~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "ASIAN / PACIFIC ISLANDER", "BLACK", ~
## $ X_COORD_CD              <dbl> 996313, 981845, 996546, 1001139, 1050710, 1051~
## $ Y_COORD_CD              <dbl> 187499, 171118, 187436, 192775, 184826, 196646~
## $ Latitude                <dbl> 40.68132, 40.63636, 40.68114, 40.69579, 40.673~
## $ Longitude               <dbl> -73.95651, -74.00867, -73.95567, -73.93910, -7~
## $ Lon_Lat                 <chr> "POINT (-73.95650899099996 40.68131820000008)"~
```

First Visualization

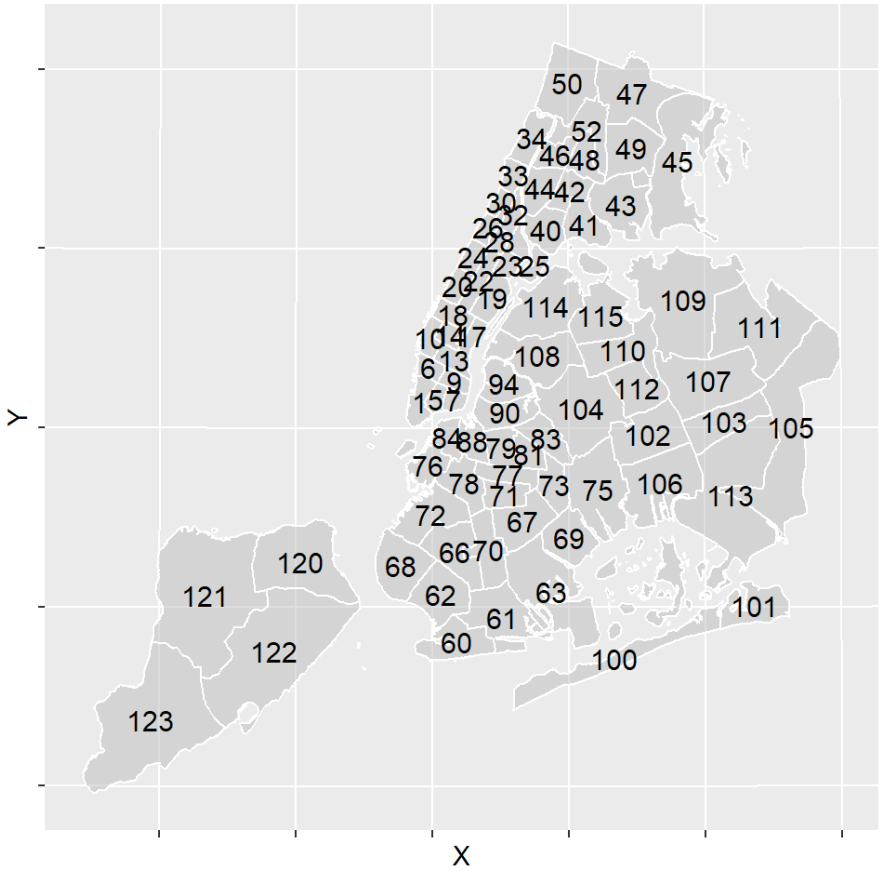
I decided to go with shootings by borough for my first visualization.

```
boros <- nyc_shootings$BORO
boros.freq <- table(boros)
barplot(boros.freq, col="#69b3a2", xlab="Borough", ylab="# of Shootings")
```



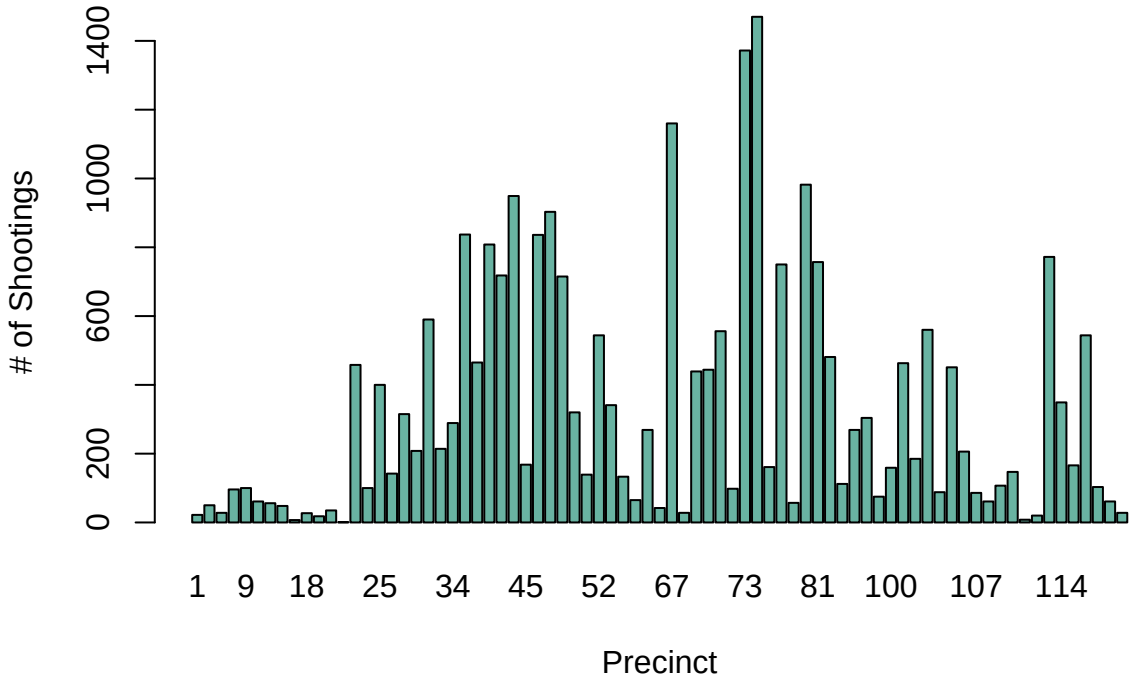
Second Visualization - Precincts

After noticing that precinct number was also provided in the data, I figured it made sense to get a little more granular. Taking a quick look at a map of precincts, the numbering of precincts is pretty well organized. Precincts with numbers close to one another also tend to be close to each other (Except at the boundaries of the



islands). I generated a shooting

```
precincts <- nyc_shootings$PRECINCT
precincts.freq <- table(precincts)
barplot(precincts.freq, col="#69b3a2", xlab="Precinct", ylab="# of Shootings")
```



Potential biases

My initial expectation that precincts near each other would have similar crime numbers was immediately crushed by this output. A lot of wild spikes and variations as we move from precinct to precinct. Thinking on this further, it was pretty naive to assume neighboring precincts would have similar numbers. A lot of data is missing. Crucial pieces of social/economic/demographic data are missing about each precinct (like population, size, police funding, age demographics, median income, etc). Without that information, there's a high potential for bias. Sourcing the necessary data to remove bias and provide a more meaningful analysis would be too time intensive for the scope of this assignment.

I decided to move on to something a little less impacted by the lack of socioeconomic data. I've always heard that crime rises as temperature goes up. I've never seen any verification of it. Since I have this data in front of me, I can use this opportunity to kick the tires on this common claim.

Initial Cleanup

Removing a few columns to focus in on the interesting stuff. Not doing any map work, so the geospatial stuff can go. Location description is rarely populated. Jurisdiction Code is not needed. The demographic stuff won't be particular relevant to analyzing weather, let's remove those as well.

Also converting the dates from characters into actual dates

```
nyc_shootings <- nyc_shootings %>% select(-one_of(c("Latitude", "Longitude", "Lon_Lat", "Y_COORD_CD", "

nyc_shootings$OCCUR_DATE <- as_date(nyc_shootings$OCCUR_DATE, format="%m/%d/%Y")
nyc_shootings <- nyc_shootings %>% rename(DATE = OCCUR_DATE)
```

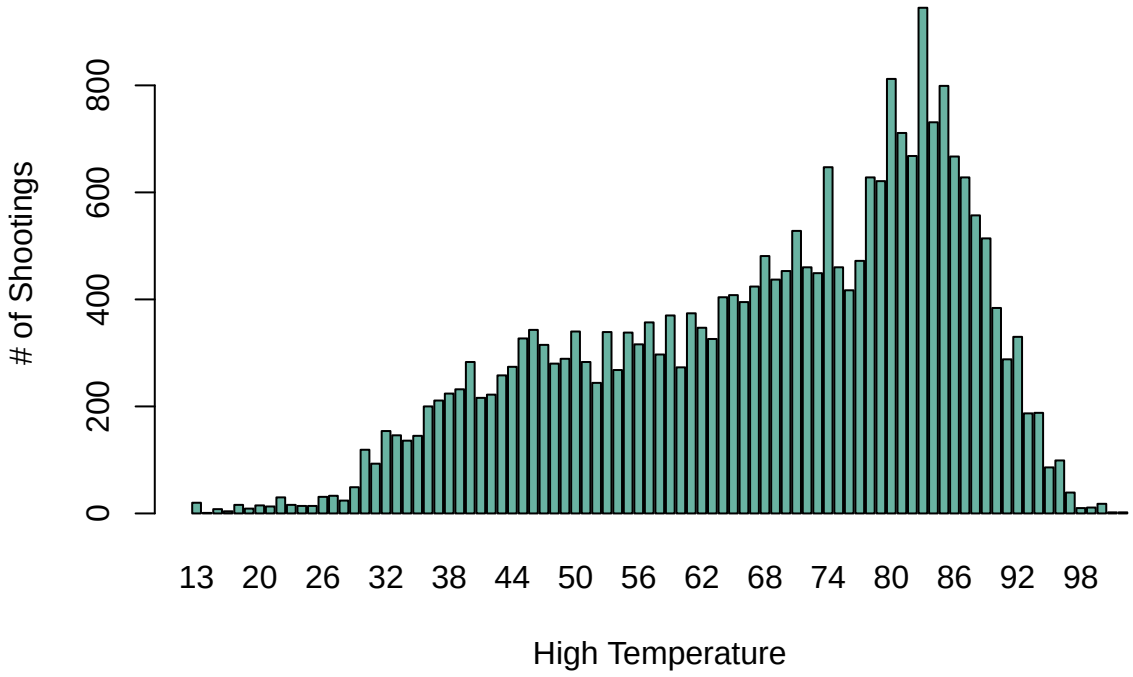
Weather

I was able to find free historical weather data for NYC pretty easily. I grabbed this a file from the climate data store on noaa.gov (https://www.ncdc.noaa.gov/cdo-web/search)

I loaded this into R and converted the date field on both tables to a date type. I then joined them on the date column. The combined table will now show me the high temperature on the day that each shooting occurred.

```
nyc_weather <- read.csv("NYC_DAILY_WEATHER.csv", header=TRUE, stringsAsFactors=FALSE )
nyc_weather$DATE <- as_date(nyc_weather$DATE, format="%Y-%m-%d")
nyc_weather <- nyc_weather %>% select(one_of(c("DATE", "TMAX")))
nyc_weather <- nyc_weather %>% rename(HIGH_TEMPERATURE = TMAX)

combined <- merge(nyc_shootings, nyc_weather,by=c("DATE"))
temperatures <- combined$HIGH_TEMPERATURE
temperatures.freq <- table(temperatures)
barplot(temperatures.freq, col="#69b3a2", xlab="High Temperature", ylab="# of Shootings")
```



Weather

This is a good start but we're looking at total number of shootings and temperatures that occur more often than others will see more shootings just because those temperatures account for more days worth of data.

To solve this, I will get the number of shootings that occurred at each temperature and the number of days that occurred at each temperature. In my new table, I can use those two columns to find the average number of shootings per day for each temperature

```
#This is the number of shootings that occurred on days with each temperature
shooting_temperature_counts = aggregate(combined[, c("HIGH_TEMPERATURE")],by=list(combined$HIGH_TEMPERA
shooting_temperature_counts <- shooting_temperature_counts %>% rename(HIGH_TEMPERATURE = Group.1)
shooting_temperature_counts <- shooting_temperature_counts %>% rename(SHOOTINGS_AT_THIS_TEMPERATURE = x
#shooting_temperature_counts

#This is the number of days that have experienced a given temperature
temperature_day_counts = aggregate(nyc_weather[, c("HIGH_TEMPERATURE")],by=list(nyc_weather$HIGH_TEMPERA
temperature_day_counts <- temperature_day_counts %>% rename(HIGH_TEMPERATURE = Group.1)
temperature_day_counts <- temperature_day_counts %>% rename(DAYS_WITH_THIS_TEMP = x)
#temperature_day_counts


temperatures_and_shootings  <- merge(temperature_day_counts, shooting_temperature_counts,by=c("HIGH_TEM
#temperatures_and_shootings

temperatures_and_shootings$SHOOTINGS_PER_DAY <- temperatures_and_shootings$SHOOTINGS_AT_THIS_TEMPERATUR
#temperatures_and_shootings
```

Modeling

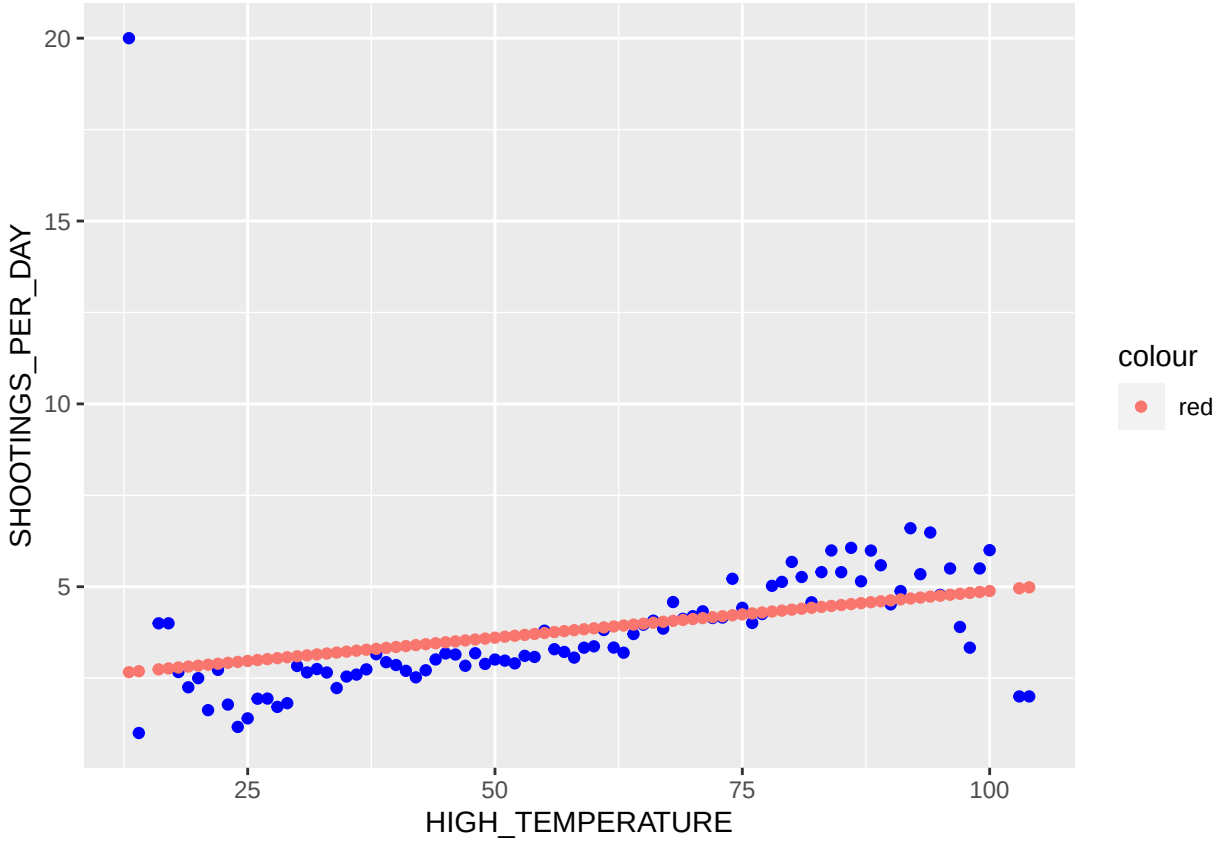
I'm going to use a linear model to see if we make a case for any kind of validity to the idea that hotter days see more violence.

```
model = lm(SHOOTINGS_PER_DAY ~ HIGH_TEMPERATURE, data = temperatures_and_shootings)
summary(model)
```

```
##
## Call:
## lm(formula = SHOOTINGS_PER_DAY ~ HIGH_TEMPERATURE, data = temperatures_and_shootings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9844 -0.6781 -0.3053  0.2284 17.3359
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.332659   0.543073   4.295 4.52e-05 ***
## HIGH_TEMPERATURE 0.025498   0.008552   2.981  0.00372 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.082 on 87 degrees of freedom
## Multiple R-squared:  0.0927, Adjusted R-squared:  0.08227
## F-statistic: 8.889 on 1 and 87 DF,  p-value: 0.00372
```

```
real_plus_pred <- temperatures_and_shootings %>% mutate(pred=predict(model))

real_plus_pred %>% ggplot(xlab="High Temperature", ylab="Avg Shootings per Day") + geom_point(aes(x=HIGH
```



Conclusion

This is a fairly shallow/simplified analysis for purposes of the assignment but in the end we showed modest support for the claim that warmer temperature days tend to have more violence.