Shooting Data

Josh Garvey

2022-09-07

Step 1: Obtaining Data

First thing I will do is read in the shooting data from the web.

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

Step 2: Tidy Up the data

Now that I have data in the shooting_data variable I can work on tidying it up. I would like to look at the amount of deaths/shooting/precinct. I left in victim and perpetrator sex and race in case I want to start pulling those factors in as well

```
sub_shoot = subset(shooting_data, select = -c(INCIDENT_KEY,
                                              OCCUR_DATE,
                                              OCCUR_TIME,
                                              BORO,
                                              JURISDICTION_CODE,
                                              LOCATION_DESC,
                                              X_COORD_CD,
                                              Y_COORD_CD,
                                              VIC_AGE_GROUP,
                                              Latitude,
                                              Longitude,
                                              Lon_Lat
                                              ))
# columns removed
sub_shoot
```

```
## # A tibble: 25,596 x 7
##    PRECINCT STATISTICAL_MURDER_FLAG PERP_AGE_G~1 PERP_~2 PERP_~3 VIC_SEX VIC_R~4
##       <dbl> <lgl>                   <chr>        <chr>   <chr>   <chr>   <chr>
##  1       52 TRUE                    <NA>         <NA>    <NA>    F       BLACK ~
##  2      106 FALSE                   <NA>         <NA>    <NA>    M       WHITE
##  3       42 TRUE                    <NA>         <NA>    <NA>    M       BLACK
##  4       52 FALSE                   <NA>         <NA>    <NA>    M       BLACK
##  5       34 FALSE                   <NA>         <NA>    <NA>    M       BLACK ~
##  6       75 TRUE                    25-44        M       BLACK ~ M       WHITE ~
##  7       32 FALSE                   25-44        M       BLACK   M       BLACK
##  8       26 FALSE                   <NA>         <NA>    <NA>    M       BLACK
##  9       41 TRUE                    25-44        M       BLACK   M       BLACK ~
## 10       67 FALSE                   <NA>         <NA>    <NA>    M       BLACK
## # ... with 25,586 more rows, and abbreviated variable names 1: PERP_AGE_GROUP,
## #   2: PERP_SEX, 3: PERP_RACE, 4: VIC_RACE
```

```
# summary
summary(sub_shoot)
```

```
##     PRECINCT      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Min.   :  1.00   Mode :logical           Length:25596       Length:25596
##  1st Qu.: 44.00   FALSE:20668             Class :character   Class :character
##  Median : 69.00   TRUE :4928              Mode  :character   Mode  :character
##  Mean   : 65.87
##  3rd Qu.: 81.00
##  Max.   :123.00
##   PERP_RACE           VIC_SEX            VIC_RACE
##  Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

There are some null points in the data set. I plan on throwing those data points out if they are part of an analysis.

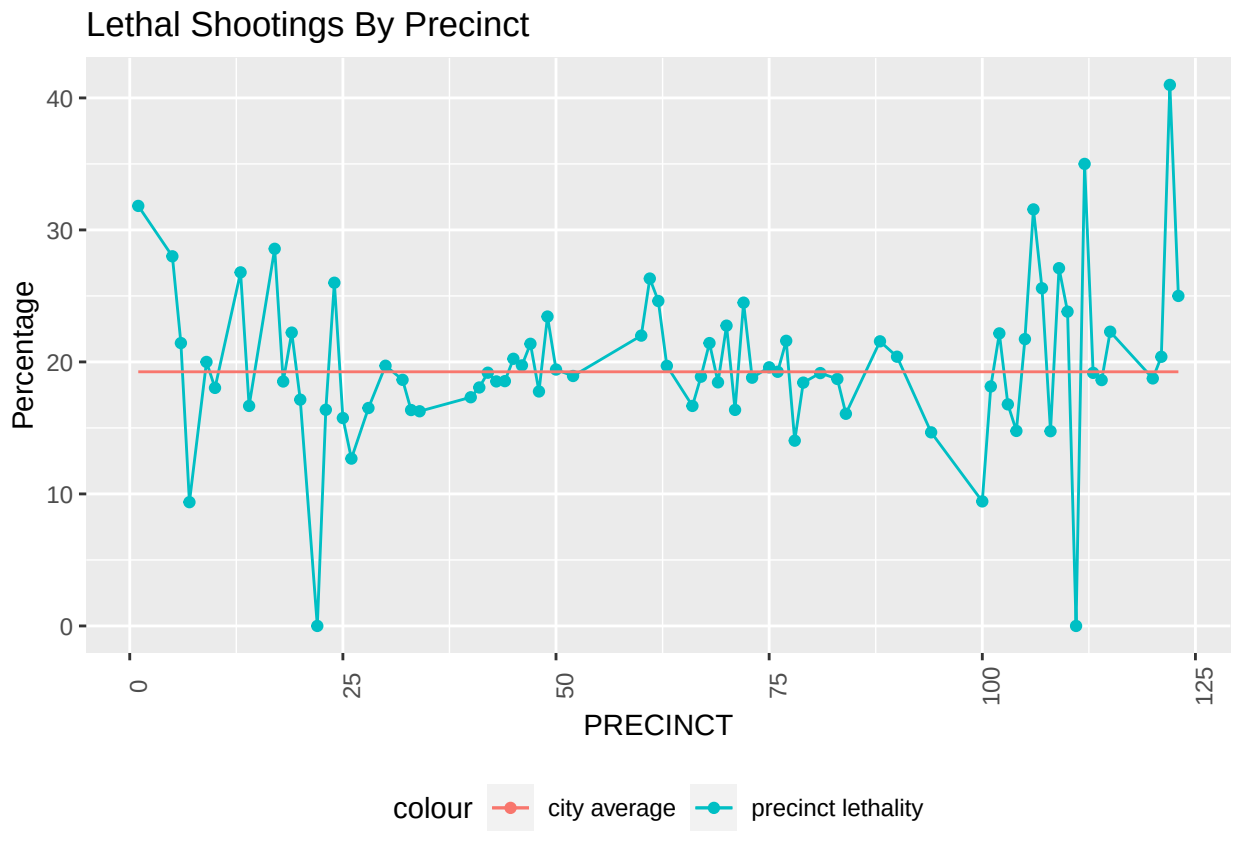
Step 3: Visualize the Data and Analyze

With the data parsed down I can start to explore different factors leading to shooting deaths to try and find correlations. To start with I will select the columns for whether or not there was a death, and what precinct it occured in.

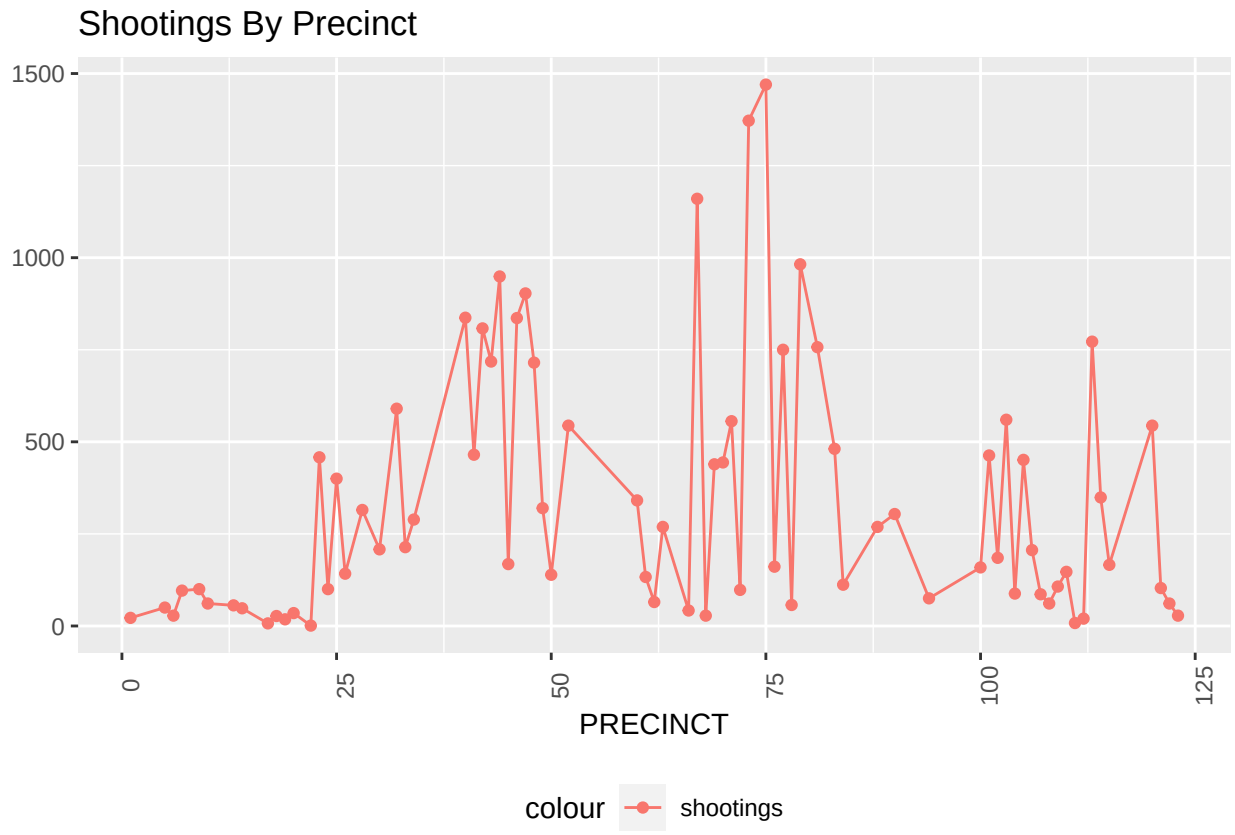
```
deaths = subset(sub_shoot, select = c(PRECINCT, STATISTICAL_MURDER_FLAG))
summary(deaths)
```

```
##     PRECINCT      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Mode :logical
##  1st Qu.: 44.00   FALSE:20668
##  Median : 69.00   TRUE :4928
##  Mean   : 65.87
##  3rd Qu.: 81.00
##  Max.   :123.00
```

```r
# Total shootings and percentage that are lethal (a baseline for the area)
nLethal = sum(!deaths[2])
Lethal = sum(deaths[2])
total_shootings = Lethal + nLethal
percent_lethal_total = (Lethal/total_shootings)*100

# Group the data to be graphed
deaths_by_precinct <- deaths %>%
  group_by(PRECINCT) %>%
  summarize(num_deaths = sum(STATISTICAL_MURDER_FLAG),
            shootings = sum(STATISTICAL_MURDER_FLAG)+sum(!STATISTICAL_MURDER_FLAG)) %>%
  mutate(perc_lethal = (num_deaths / shootings) *100) %>%
  ungroup()

# Graph the lethal percentages by precinct
deaths_by_precinct %>%
  ggplot(aes(x = PRECINCT, y = perc_lethal)) +
  geom_line(aes(color = "precinct lethality")) +
  geom_point(aes(color = "precinct lethality")) +
  geom_line(aes(y = percent_lethal_total, color = "city average")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Lethal Shootings By Precinct", y = "Percentage")
```



```r
# Graph number of shootings versus total shooting to see if there is a trigger happy precinct
shootings_by_precinct <- deaths %>%
  group_by(PRECINCT) %>%
  summarize(num_shootings = sum(STATISTICAL_MURDER_FLAG)+sum(!STATISTICAL_MURDER_FLAG)) %>%
  ungroup()

shootings_by_precinct %>%
  ggplot(aes(x = PRECINCT, y = num_shootings)) +
  geom_line(aes(color = "shootings")) +
  geom_point(aes(color = "shootings")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Shootings By Precinct", y = NULL)
```



Step 4: Sources of Bias

I think the biggest bias source is the fact that I chose to include factors like race and sex into my analysis but left out age range, location, the date, etc. There are infinitely many factors that go into an officers decision to shoot and by handpicking these ones to analyze I may be missing some huge correlation. An example could be that the rate of deaths by shooting goes way up in one location due to it being further from medical attention than another.

I could try to mitigate by picking a few columns for data to hopefully give me an idea when something is an outlier. This is going to be hard to pick arbitrarily. Alternatively, I can accept that there is bias in the

study and append a synopsis of how I got my results to the report so others can decide if its worthy of their use.