NYPD Shooting data analysis

2022-10-31

NYPD Shooting Data source

(Data description from the hosting website, https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic)

“List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.”

```
URL = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
NYPD_data = read_csv(URL)
```

NYPD Shooting Data analysis

Summary of data set:

```
summary(NYPD_data)
```

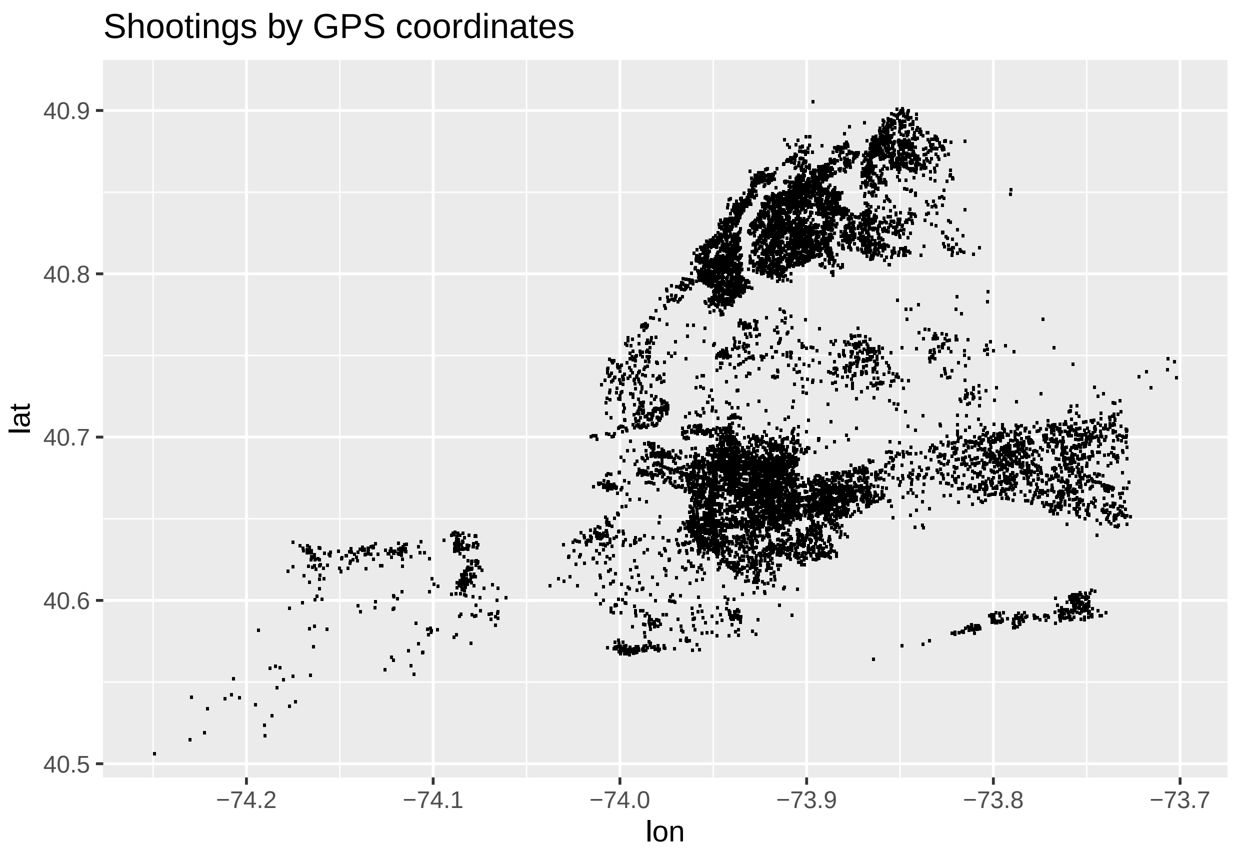
```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  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
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

There are a large number of missing values for each of the columns, so analysis will be focused on the take this into account.

Plots

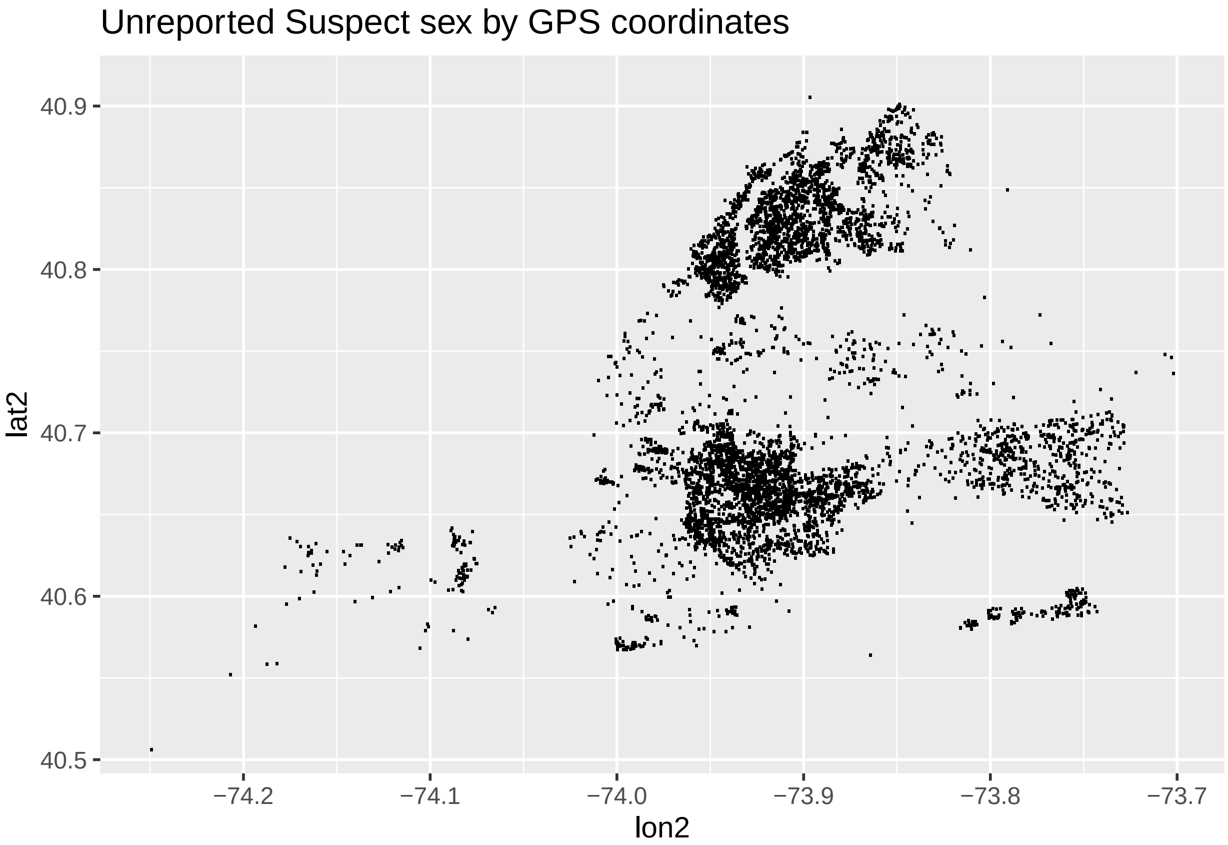
Shootings in NYC are local phenomenom. Concentrated in particular areas.

```
lat = NYPD_data$Latitude
lon = NYPD_data$Longitude
ggplot(data=NYPD_data, aes(x = lon, y = lat)) + geom_text(label = ".") + ggtitle('Shootings by GPS coor
```



I was intrigued by the missing data on perpetrator sex. Several factors might account for missing information about the suspect, but I wanted to see if there were certain areas of NY where that information was being withheld more often than others. My Hypothesis was that a particular neighborhood might be less willing to talk to police for example.

```
data = filter(NYPD_data, is.na(NYPD_data$PERP_SEX))
lat2 = data$Latitude
lon2 = data$Longitude
ggplot(data=data, aes(x = lon2, y = lat2)) + geom_text(label = ".") + ggtitle('Unreported Suspect sex by
```



However both were distributed more or less in the same pattern, challenging the idea that the missing information was localized even further than the shooting were at the outset.

Another area of analysis is whether or not the sex of a victim is correlated to the sex of the suspect.

```
complete_data = drop_na(NYPD_data, PERP_SEX)
complete_data$PERP_SEX = replace(complete_data$PERP_SEX, complete_data$PERP_SEX == 'M', 1)
complete_data$PERP_SEX = replace(complete_data$PERP_SEX, complete_data$PERP_SEX == 'F', 2)
complete_data$PERP_SEX = replace(complete_data$PERP_SEX, complete_data$PERP_SEX == 'U', 0)
complete_data$VIC_SEX = replace(complete_data$VIC_SEX, complete_data$VIC_SEX == 'M', 1)
complete_data$VIC_SEX = replace(complete_data$VIC_SEX, complete_data$VIC_SEX == 'F', 2)
complete_data$VIC_SEX = replace(complete_data$VIC_SEX, complete_data$VIC_SEX == 'U', 0)

lm(complete_data$PERP_SEX ~ complete_data$VIC_SEX, data = complete_data)
```

```
##
## Call:
## lm(formula = complete_data$PERP_SEX ~ complete_data$VIC_SEX,
##     data = complete_data)
##
## Coefficients:
##            (Intercept)  complete_data$VIC_SEX1  complete_data$VIC_SEX2
##                1.00000                -0.07372                -0.03158
```

In both cases, the correlation coefficient is very small, less that -.10 in both cases, indicating a very weak link between Suspect and Victim Sex.

Potential Bias

I am human and therefore subject to bias. I am not a New York resident and likely less passionate about New York issues as a result. I started the analysis with certain hypotheses and that made me more likely to find evidence of those conclusions.