NYPD Shooting Incident Data (Historic)

2022-03-15

Read data from website

```
library(tidyverse)
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
data <- read_csv(url)
```

Initial Summary

```
head(data)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>       <dbl>              <dbl>
## 1     24050482 08/27/2006 05:35      BRONX          52                  0
## 2     77673979 03/11/2011 12:03      QUEENS        106                  0
## 3    203350417 10/06/2019 01:09      BROOKLYN       77                  0
## 4     80584527 09/04/2011 03:35      BRONX          40                  0
## 5     90843766 05/27/2013 21:16      QUEENS        100                  0
## 6     92393427 09/01/2013 04:17      BROOKLYN       67                  0
## # ... with 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

Tidy data

1. select specific columns
2. convert date to Date type

```
data <- data %>% select(OCCUR_DATE,BORO,LOCATION_DESC,PERP_AGE_GROUP,PERP_SEX,PERP_RACE,VIC_AGE_GROUP,V
head(data)
```

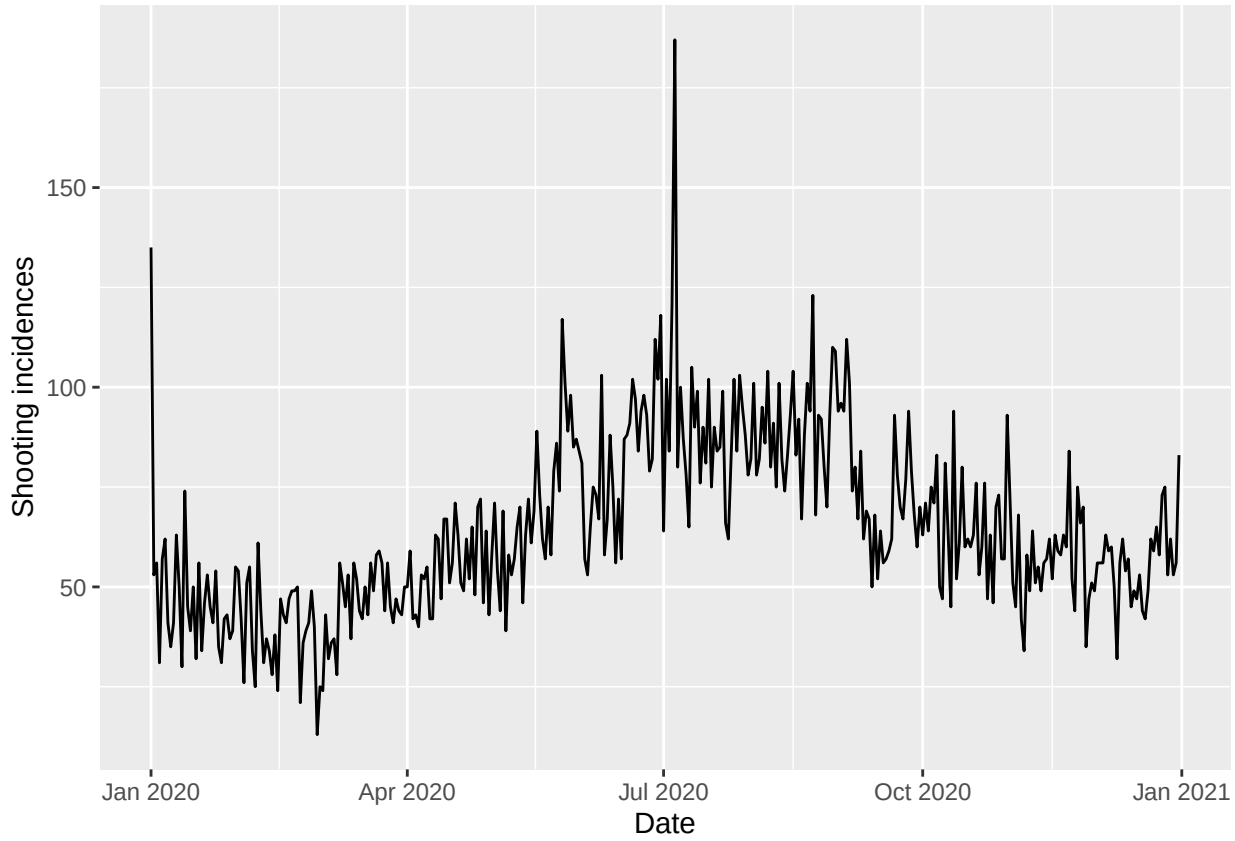
```
## # A tibble: 6 x 9
##   OCCUR_DATE BORO  LOCATION_DESC PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
##   <date>     <chr> <chr>         <chr>          <chr>    <chr>     <chr>
## 1 2020-08-27 BRONX <NA>          <NA>           <NA>     <NA>      25-44
## 2 2020-03-11 QUEE~ <NA>          <NA>           <NA>     <NA>      65+
## 3 2020-10-06 BROO~ <NA>          <NA>           <NA>     <NA>      18-24
```

```
## 4 2020-09-04 BRONX <NA>          <NA>           <NA>     <NA>      <18
## 5 2020-05-27 QUEE~ <NA>          <NA>           <NA>     <NA>      18-24
## 6 2020-09-01 BROO~ <NA>          <NA>           <NA>     <NA>      <18
## # ... with 2 more variables: VIC_SEX <chr>, VIC_RACE <chr>
```

Analyze data

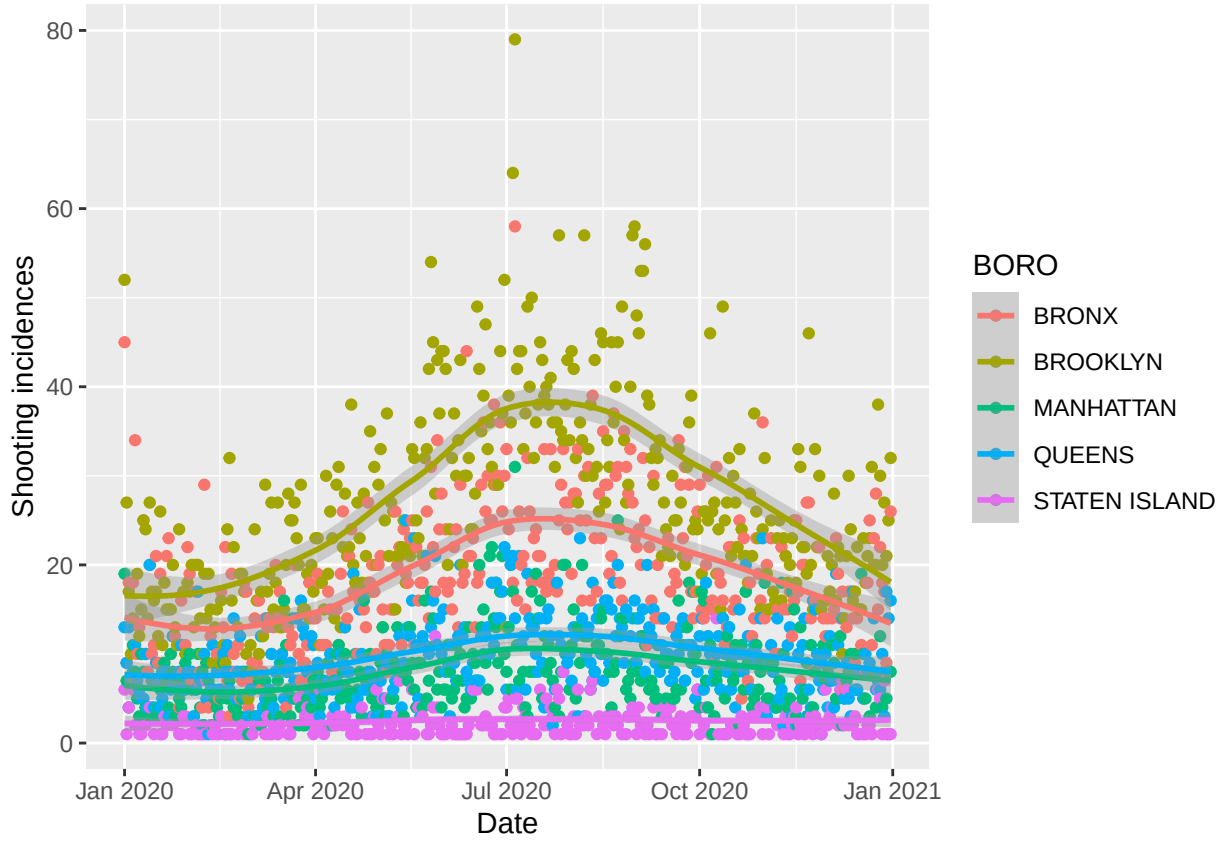
Plot count of shootings by date.

```
shootings_by_date <- data %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n())
ggplot( data = shootings_by_date, aes( OCCUR_DATE, COUNT )) +
geom_line() +
xlab("Date") + ylab("Shooting incidences")
```



Maybe separate by Borough and add moving averages to visualize.

```
shootings_by_borough <- data %>% group_by(OCCUR_DATE,BORO) %>% summarise(COUNT = n())
ggplot( data = shootings_by_borough, aes( OCCUR_DATE, COUNT, group=BORO )) +
geom_point(aes(color=BORO)) + geom_smooth(aes(color=BORO)) +
xlab("Date") + ylab("Shooting incidences")
```



So this shows that the incidences in Staten Island are low compared to the other Boroughs. And also Brooklyn and the Bronx being comparatively high.

Hmmm... what's that date in July 2020 with a high count?

```
filter(shootings_by_date, COUNT > 150)
```

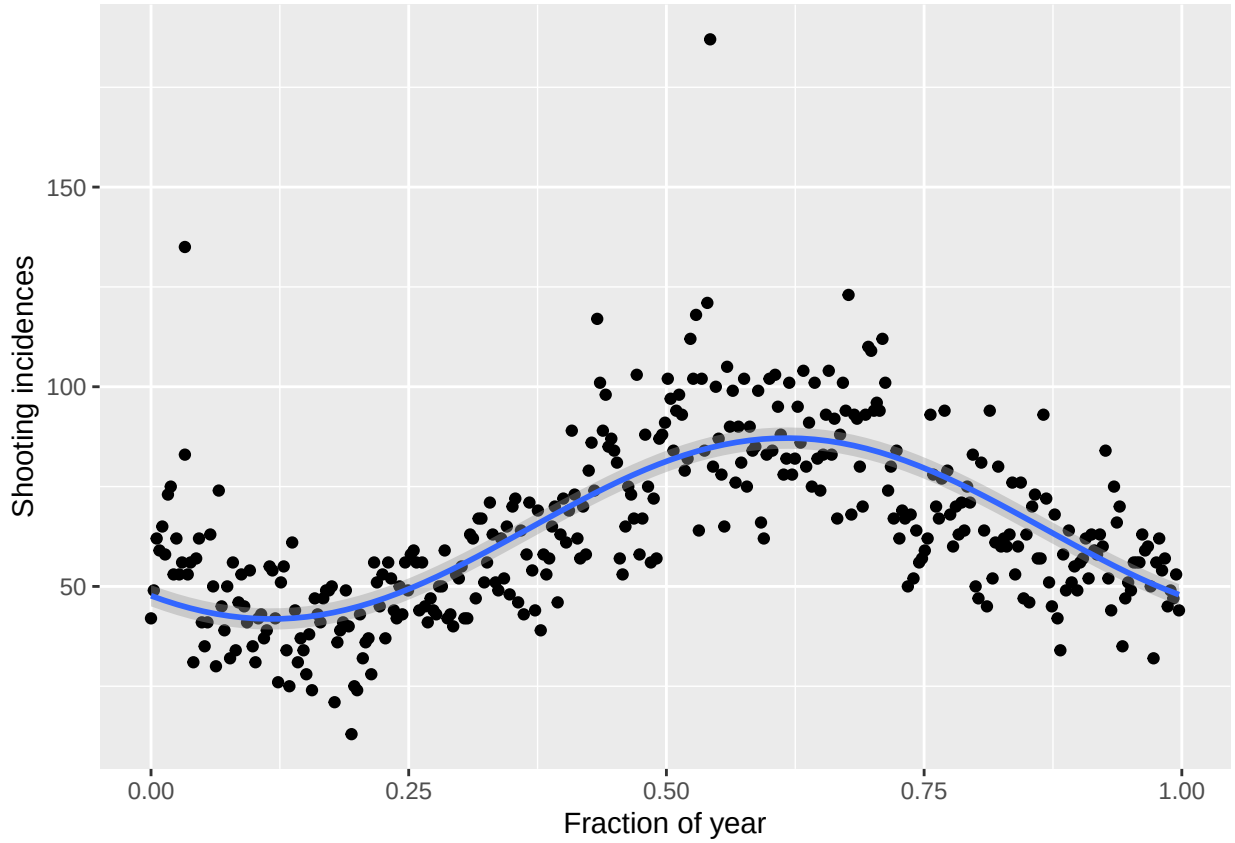
```
## # A tibble: 1 x 2
##   OCCUR_DATE COUNT
##   <date>     <int>
## 1 2020-07-05   187
```

A quick search on July 5th 2020 pulls up this article: https://nypost.com/2020/07/05/violent-july-4th-weekend-sees-at-least-10-shot-2-dead-in-nyc/ which describes an unusually high holiday weekend of shootings.

Model fitting

The data looks like it tracks with date of the year. Let's try to fit a sine of period one year.

```
ggplot( data = shootings_by_date, aes( x=(julian(OCCUR_DATE)%%365)/365, y=COUNT )) +
  geom_point() +
  geom_smooth(method="lm", formula= y ~ sin(2*pi*x)+cos(2*pi*x) ) +
  xlab("Fraction of year") + ylab("Shooting incidences")
```



Bias

The data shown here is likely only data documented by police reports. It is probable that shootings occur that go unreported. It maybe that Staten Island is under reported but it is more likely that areas with already high shooting rates are under reported.