NYPD_analysis

Bruno Lecona

4/25/2022

NYPD Shooting Incident Data (Historic) list every shooting incident that occurred in NYC, going back to 2006 through March 2022.

This data is reviewed by the Office of Management Analysis and Planning, before being posted on the NYPD website and each record includes information of the event, location and time of occurrence as well as information related to the suspect and victim.

Being said that, my main questions are. Where have occurred the most cases in NYC? Have they been increasing along the time?

Let’s start by loading the data from the database mentioned above

```
NYPD_shooting_data <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=D
```

And reading it to know more about the variables of the data set and look what are we going to use and what we don’t need for this analysis

```
str(NYPD_shooting_data)
```

```
## spec_tbl_df [23,585 x 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY           : num [1:23585] 2.41e+07 7.77e+07 2.03e+08 8.06e+07 9.08e+07 ...
##  $ OCCUR_DATE             : chr [1:23585] "08/27/2006" "03/11/2011" "10/06/2019" "09/04/2011" ...
##  $ OCCUR_TIME             : 'hms' num [1:23585] 05:35:00 12:03:00 01:09:00 03:35:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ BORO                   : chr [1:23585] "BRONX" "QUEENS" "BROOKLYN" "BRONX" ...
##  $ PRECINCT               : num [1:23585] 52 106 77 40 100 67 77 81 101 106 ...
##  $ JURISDICTION_CODE      : num [1:23585] 0 0 0 0 0 0 0 0 0 0 ...
##  $ LOCATION_DESC          : chr [1:23585] NA NA NA NA ...
##  $ STATISTICAL_MURDER_FLAG: logi [1:23585] TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ PERP_AGE_GROUP         : chr [1:23585] NA NA NA NA ...
##  $ PERP_SEX               : chr [1:23585] NA NA NA NA ...
##  $ PERP_RACE              : chr [1:23585] NA NA NA NA ...
##  $ VIC_AGE_GROUP          : chr [1:23585] "25-44" "65+" "18-24" "<18" ...
##  $ VIC_SEX                : chr [1:23585] "F" "M" "F" "M" ...
##  $ VIC_RACE               : chr [1:23585] "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num [1:23585] 1017542 1027543 995325 1007453 1041267 ...
##  $ Y_COORD_CD             : num [1:23585] 255919 186095 185155 233952 157134 ...
##  $ Latitude               : num [1:23585] 40.9 40.7 40.7 40.8 40.6 ...
##  $ Longitude              : num [1:23585] -73.9 -73.8 -74 -73.9 -73.8 ...
##  $ Lon_Lat                : chr [1:23585] "POINT (-73.87963173099996 40.86905819000003)" "POINT (-73
##  - attr(*, "spec")=
##   .. cols(
##   ..   INCIDENT_KEY = col_double(),
##   ..   OCCUR_DATE = col_character(),
##   ..   OCCUR_TIME = col_time(format = ""),
##   ..   BORO = col_character(),
##   ..   PRECINCT = col_double(),
##   ..   JURISDICTION_CODE = col_double(),
##   ..   LOCATION_DESC = col_character(),
##   ..   STATISTICAL_MURDER_FLAG = col_logical(),
##   ..   PERP_AGE_GROUP = col_character(),
##   ..   PERP_SEX = col_character(),
##   ..   PERP_RACE = col_character(),
##   ..   VIC_AGE_GROUP = col_character(),
##   ..   VIC_SEX = col_character(),
##   ..   VIC_RACE = col_character(),
##   ..   X_COORD_CD = col_double(),
##   ..   Y_COORD_CD = col_double(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   Lon_Lat = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

*Tidy and transform data.

Location_desc has a lot of missing data but we will not be using it so we will be ignoring that variable for now. If we have had to use it we would be transforming those empty rows to a MISSING description.

As we saw, half of our data is char, we are going to transform them to factor. The Occur_date and Occur_time keys need to be transformed to data and hour formats respectively so we can use them properly, the statistical_murder_flag could work better with a logical format. Finally we got rid of the X_COORD_CD,Y_COORD_CD, Lot_Lat keys because we have the Longitude and Latitude keys on their own and are the ones we will be using.

*Visualization

Taking a look to the Precincts with the most and less cases

4 of the 5 with higher cases are in Brooklyn(75,73,67,79) and one is in the Bronx (44)

For less cases we divide them between Manhattan (19,17,22) and Queens(112,111)

Now we know why everyone wants to live in Manhattan. (Just kidding)

These are the 5 Precincts with the most historical cases in NYC

```
precinct_head<-NYPD_sh_t %>%
  group_by(PRECINCT) %>%
  tally()
 precinct_head<-precinct_head %>%
   arrange(desc(n)) %>%
   head(n=5)
 precinct_head
```

```
## # A tibble: 5 x 2
##   PRECINCT     n
##   <fct>    <int>
## 1 75        1375
## 2 73         1284
## 3 67         1101
## 4 79          921
## 5 44          841
```

These are the 5 Precincts with less historical cases in NYC

```
precinct_tail<-NYPD_sh_t %>%
 group_by(PRECINCT) %>%
 tally()
precinct_tail<-precinct_tail %>%
  arrange(desc(n)) %>%
  tail(n=5)
precinct_tail
```

```
## # A tibble: 5 x 2
##   PRECINCT     n
##   <fct>    <int>
## 1 112         19
## 2 19          11
## 3 17           6
## 4 111          6
## 5 22           1
```

Having Precincts with the most cases does not necessarily mean they are indeed where there have been the most cases in general.

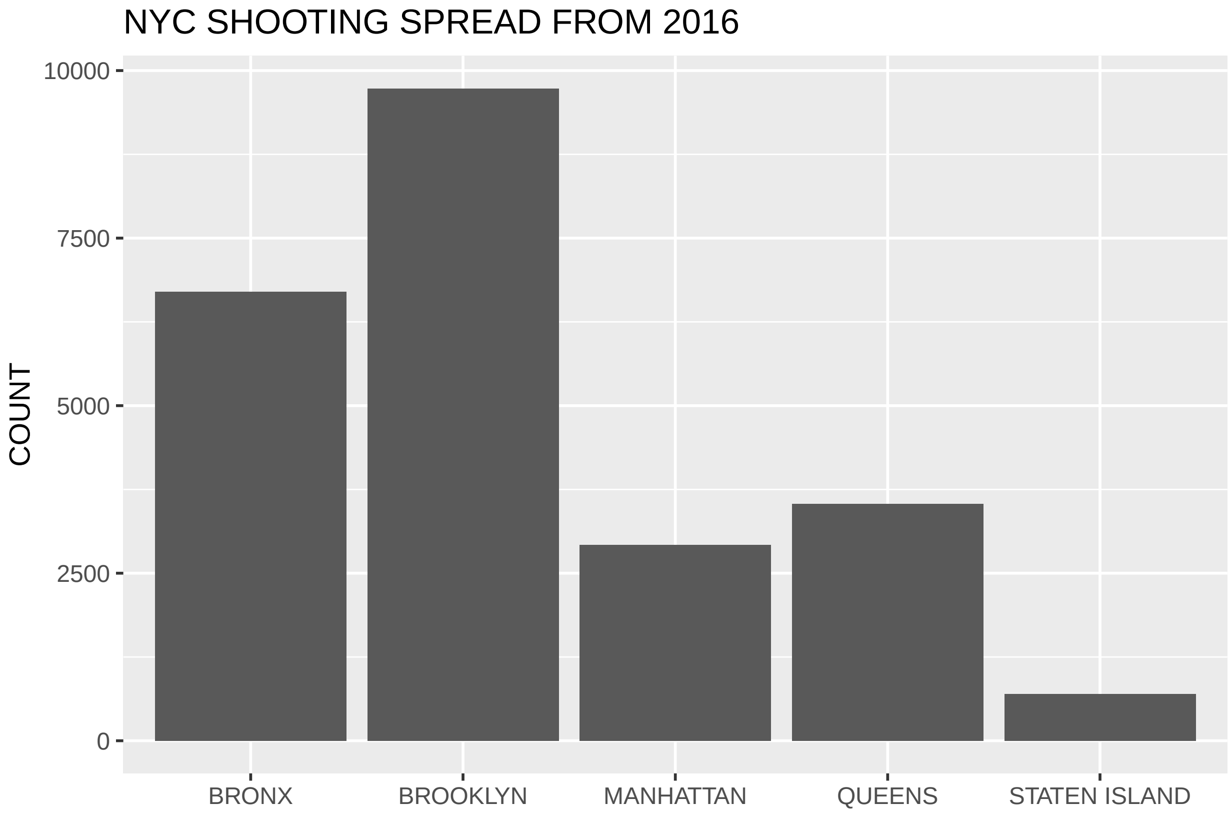
By creating this Bar plot of NYC shootings it is clearer how the cases are distributed.

Looking at you Brooklyn, our first place.

But we should be careful, more cases does not mean it is more dangerous. The size of the Borough and population, between other data that we could add to make a deeper analysis could be part to get rid of any possible bias.

But what we do know is that Brooklyn is the Borough with the highest number of shootings since 2006.

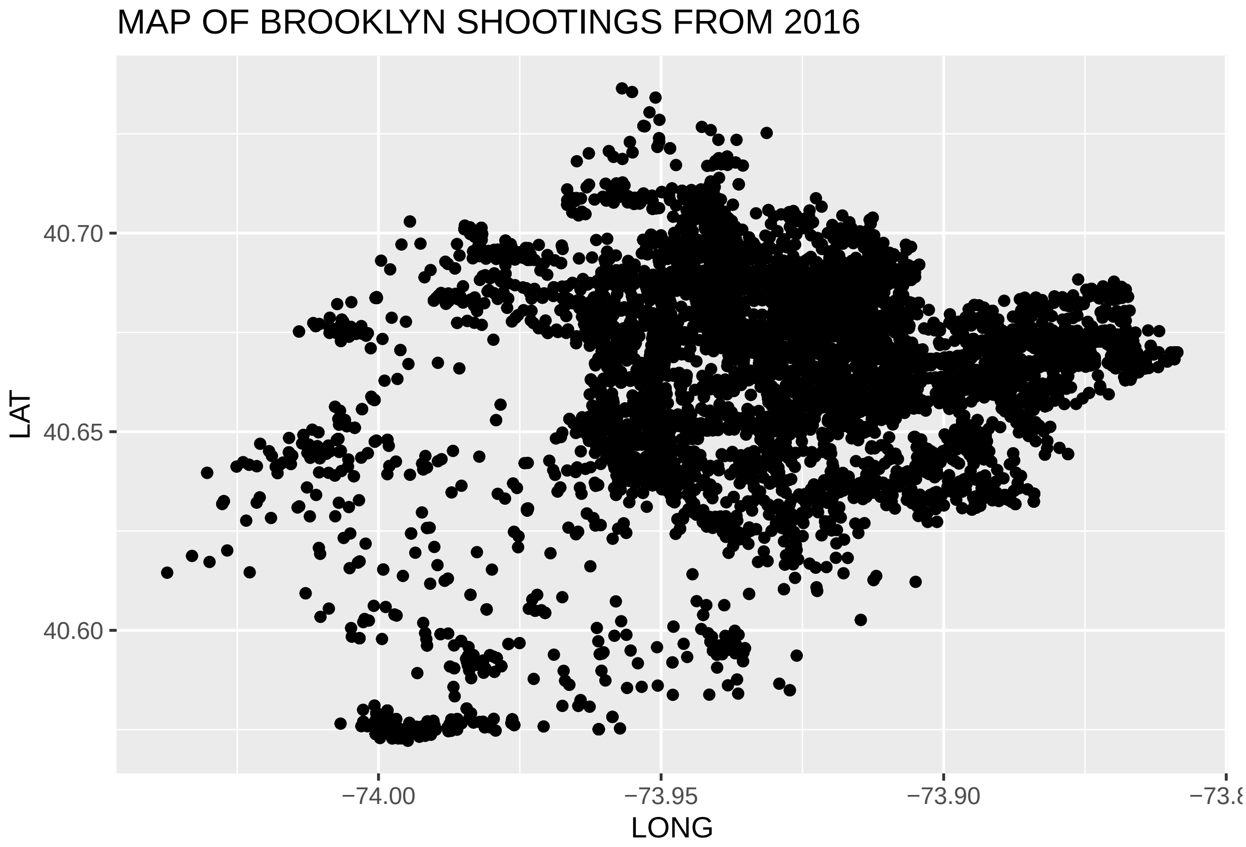
```
barplot_NYC<-NYPD_sh_t %>%
  ggplot(aes(x=BORO))+geom_bar()+xlab("")+ ylab("COUNT")+ggtitle("NYC SHOOTING SPREAD FROM 2016")
barplot_NYC
```



Additionally We will visualize a map of Brooklyn to have an idea of which part of it has the most density

Basically, in the upper right quadrant of our point graph is where it is concentrated the majority of the cases registered so far.

```
brooklyn_map<-NYPD_shooting_data %>%
  filter(BORO=="BROOKLYN") %>%
  ggplot(aes(x=Longitude, y=Latitude))+geom_point()+xlab("LONG")+ ylab("LAT")+ggtitle("MAP OF BROOKLYN S
brooklyn_map
```



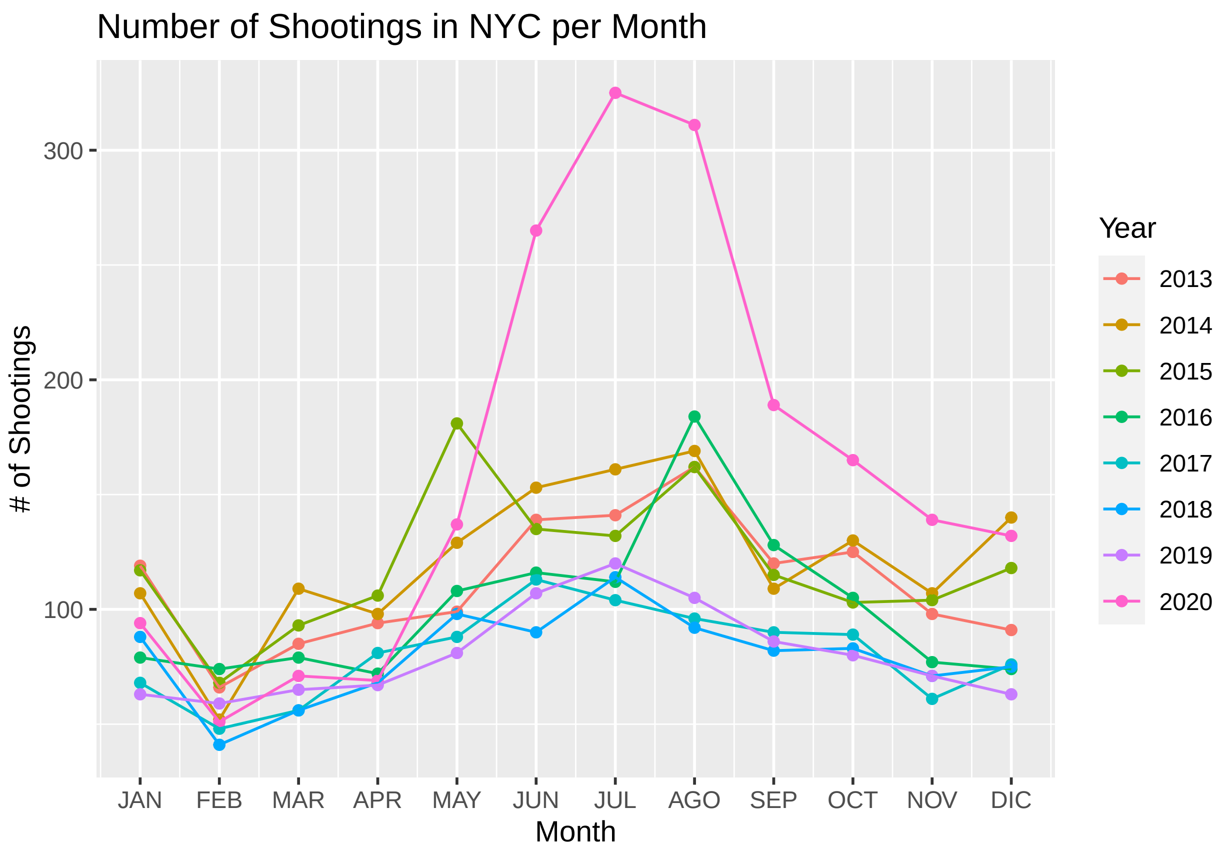
I'm left with some questions.

I would like to know, historically, in which months of the year NYC has had the most cases. We made this one!

Note that these could be influenced by tourism, weather, population, elections, etc. But we are not looking at those factors right now.

But we can clearly see a yearly trend, the spikes are near July, however something to worry about would be the this mountain that happened when the pandemic started andd went went way higher than the average years.

```
NYPD_sh_t$Yr <- year(NYPD_sh_t$OCCUR_DATE)

nymonths <- NYPD_sh_t %>%
     filter(Yr > 2012) %>%
     mutate("Month"= month(OCCUR_DATE), "Year" = as.factor(Yr)) %>%
     group_by(Year,Month) %>%
     summarize(shootings = n(), murders_gun = sum(STATISTICAL_MURDER_FLAG))



nymonthsgraph <- ggplot(nymonths) + geom_point(aes(x=Month, y=shootings, colour=Year)) + geom_line(aes(
nymonthsgraph
```



*Model

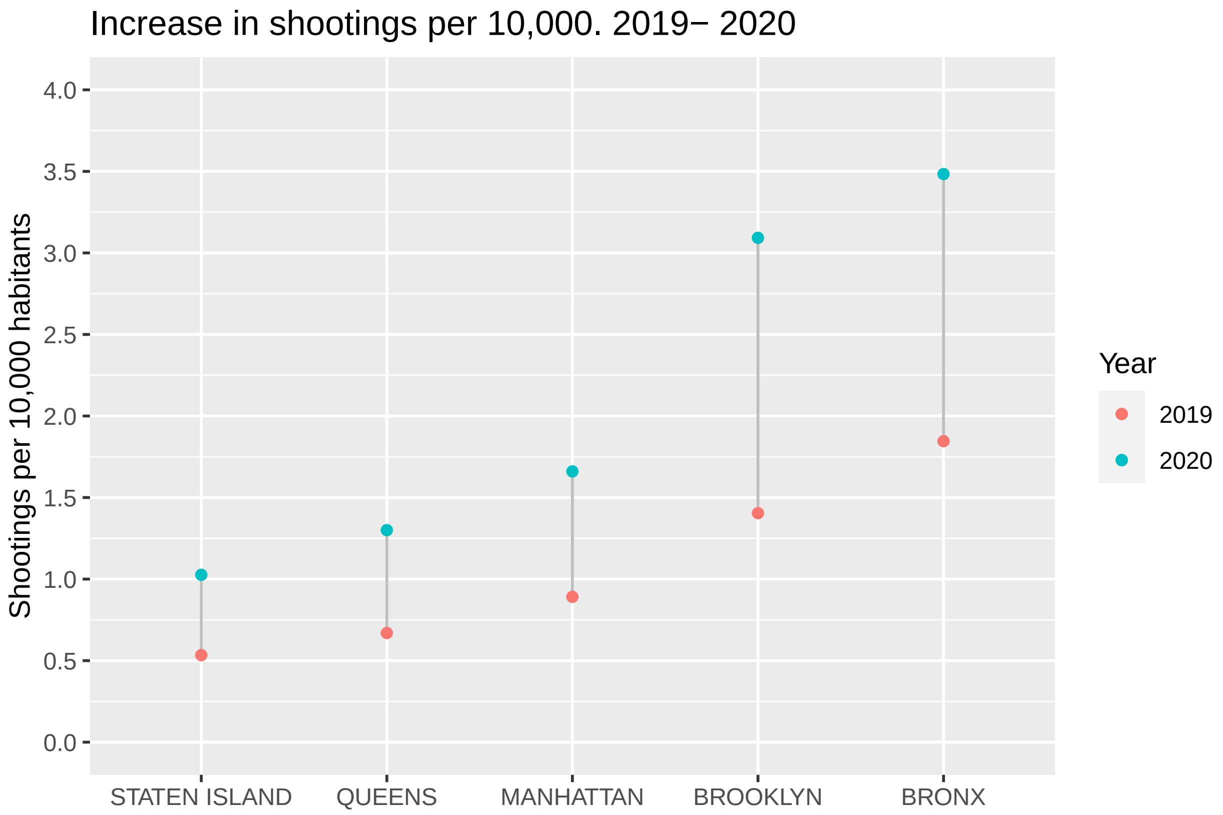
We should end this with our model, that will answer our two main questions.

Where have occurred the most cases in NYC? Have they been increasing along the time?

We took Population as a new variable from the website of newyorkcity database to be able to answer that.

And what do we see in this last plot? even though Brooklyn had the highest number of shootings in our analysis Bronx is the top at shootings per 10,000 people.

```
pop <- read_csv("https://data.cityofnewyork.us/api/views/xywu-7bv9/rows.csv")

population <- pop %>%
     select(Borough,`2020`) %>%
     rename('population'='2020', 'BORO'='Borough') %>%
     filter(BORO!= "NYC Total")%>%
     mutate(BORO = toupper(BORO))

nydbmodel <- NYPD_sh_t %>%
     filter(Yr > 2018) %>%
 mutate("Year" = as.factor(Yr)) %>%
     group_by(Year, BORO) %>%
     summarize(shootings = n()) %>%
     pivot_wider(names_from = Year, values_from = shootings,names_prefix = 'Y')

nydbmodel <-left_join(nydbmodel,population,by = 'BORO')

nyfinalmodel<- nydbmodel %>%
     mutate('S_2019_per_10000'=10000*Y2019/population,'S_2020_per_10000'=10000*Y2020/population ) %>%
     rowwise()%>%
     mutate(mymean = mean(c(S_2019_per_10000,S_2020_per_10000))) %>%
     arrange(mymean) %>%
     mutate(Borough = factor(BORO, BORO), Increase =S_2020_per_10000/S_2019_per_10000 )


ggplot(nyfinalmodel)+ geom_segment(aes(x = Borough, xend = Borough, y =S_2019_per_10000, yend=S_2020_pe
```



At the start of the project I would've thought that the most incidents could have been in Manhattan based on being one of the zones with most tourists in the world all of the year.

And I am glad I was wrong, because trying to look at it from the start or trying to explain why these kind of tendencies are happening, based on the little information I know about the US, could have been a bias to worry about.