NYPD Shooting Incident Data Report

Fei Ai

@ August 2022

OUTLINE

- ❑ Background
- ❑ Importing Data
- ❑ Tidying and Transforming
- ❑ Visualizing
- ❑ Modeling
- ❑ Analysis
- ❑ Conclusion and Bias

Background


SCHOOL
FREE DELIVERY
NYPD
POLICE DOG
KEEP BACK

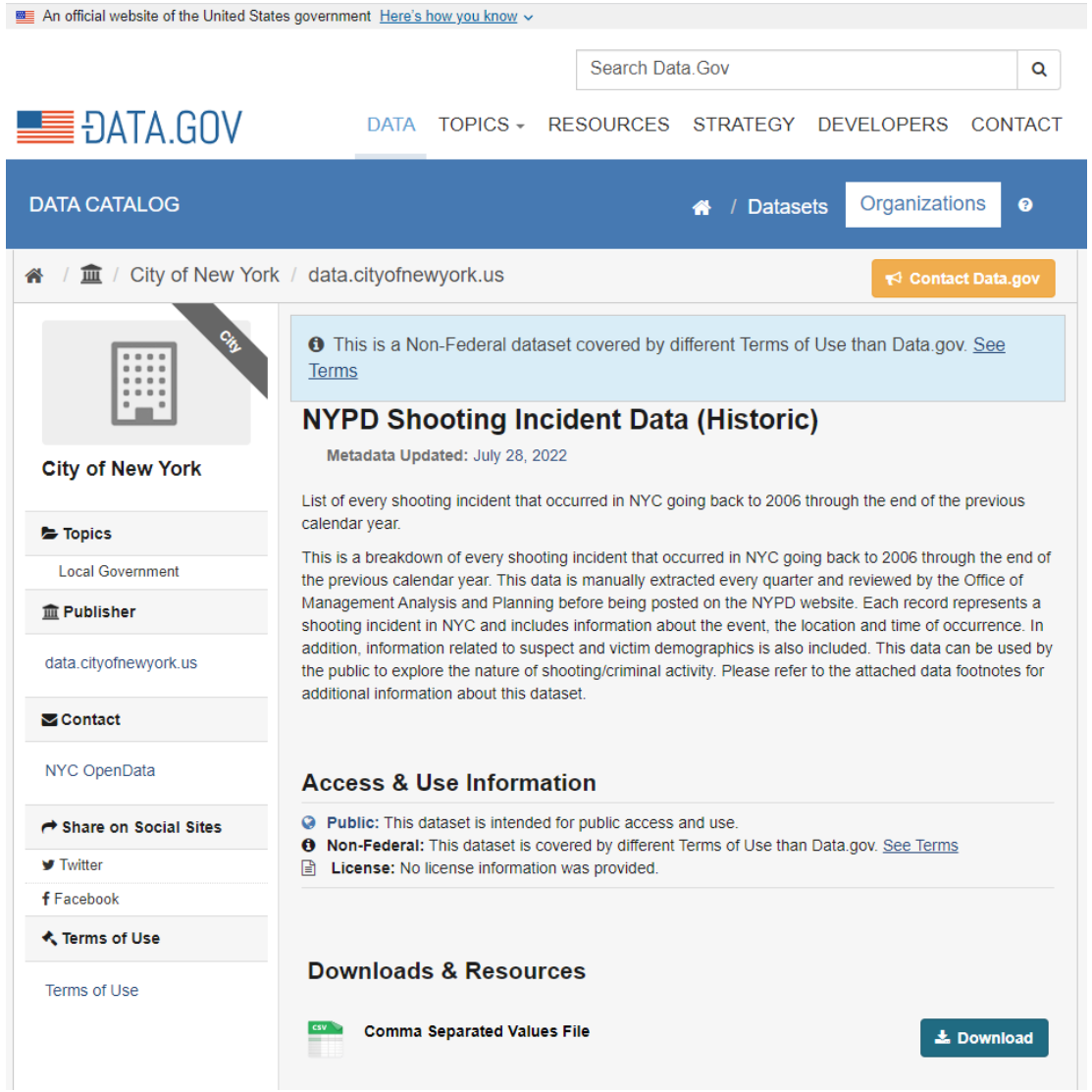

CNN on Scene
New York
3:00 PM ET
Samantha Zirkin/Earlier Today
BREAKING NEWS
NYPD ARREST SUBWAY SHOOTING SUSPECT FRANK JAMES
Apprehended "without incident" by officers in Manhattan after Crimestoppers tip
CNN
NEWSROOM


NYPD


06/17/2021 06:44:44 PM


NDTV
POLICE
BROOKLYN:
MULTIPLE PEOPLE SHOT AT
IN SUBWAY STATION

Importing Data



I start by reading **NYPD Shooting Incident Data (Historic)** from the csv file which is downloaded from this website: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic.

Code:

```
## Get historic data in the csv file
url_in <-
"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD
```

Let’s read in the data and see what we have.

```
> NYPD
# A tibble: 25,596 x 19
   INCIDENT_KEY OCCUR…¹ OCCUR…² BORO  PRECI…³ JURIS…⁴ LOCAT…⁵ STATI…⁶ PERP_…⁷ PERP_…⁸ PERP_…⁹ VIC_A…ˣ VIC_SEX VIC_R…ˣ X_COO…ˣ Y_COO…ˣ Latit…ˣ
          <dbl> <chr>    <time>   <chr>   <dbl>   <dbl> <chr>    <lgl>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>     <dbl>   <dbl>   <dbl>
 1    24050482 08/27/… 05:35    BRONX      52       0 NA       TRUE    NA      NA      NA      25-44   F       BLACK …  1.02e6 255919.    40.9
 2    77673979 03/11/… 12:03    QUEE…     106       0 NA       FALSE   NA      NA      NA      65+     M       WHITE    1.03e6 186095     40.7
 3   226950018 04/14/… 21:08    BRONX      42       0 COMMER… TRUE    NA      NA      NA      18-24   M       BLACK    1.01e6 243050     40.8
 4   237710987 12/10/… 19:30    BRONX      52       0 NA       FALSE   NA      NA      NA      25-44   M       BLACK    1.02e6 256046     40.9
 5   224701998 02/22/… 00:18    MANH…      34       0 NA       FALSE   NA      NA      NA      25-44   M       BLACK …  1.01e6 254690     40.9
 6   225295736 03/07/… 06:15    BROO…      75       0 NA       TRUE    25-44   M       BLACK … 25-44   M       WHITE …  1.02e6 187865     40.7
 7   231190175 07/21/… 00:40    MANH…      32       0 NA       FALSE   25-44   M       BLACK   25-44   M       BLACK    9.98e5 235038     40.8
 8   233429421 09/11/… 20:20    MANH…      26       2 MULTI … FALSE   NA      NA      NA      18-24   M       BLACK    9.96e5 235674     40.8
 9   227950661 05/09/… 02:50    BRONX      41       2 MULTI … TRUE    25-44   M       BLACK   25-44   M       BLACK …  1.01e6 240068     40.8
10   227344198 04/23/… 13:25    BROO…      67       0 NA       FALSE   NA      NA      NA      18-24   M       BLACK    1.00e6 180307     40.7
# … with 25,586 more rows, 2 more variables: Longitude <dbl>, Lon_Lat <chr>, and abbreviated variable names ¹OCCUR_DATE, ²OCCUR_TIME,
#   ³PRECINCT, ⁴JURISDICTION_CODE, ⁵LOCATION_DESC, ⁶STATISTICAL_MURDER_FLAG, ⁷PERP_AGE_GROUP, ⁸PERP_SEX, ⁹PERP_RACE, ˣVIC_AGE_GROUP,
#   ˣVIC_RACE, ˣX_COORD_CD, ˣY_COORD_CD, ˣLatitude
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
>
```

Code:

```
library(tidyverse)
NYPD <- read_csv(url_in)
```

Tidying and Transforming

After looking at the NYPD, I would like to tidy the dataset, I don't need several columns for the analysis, like INCIDENT_KEY, etc. So I get rid of these.

```
> NYPD_summary1
# A tibble: 25,596 × 10
   OCCUR_DATE OCCUR_TIME BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE      VIC_AGE_GROUP VIC_SEX VIC_RACE
   <chr>      <time>     <chr>     <lgl>                   <chr>          <chr>    <chr>          <chr>         <chr>   <chr>
 1 08/27/2006 05:35      BRONX     TRUE                    NA             NA       NA             25-44         F       BLACK HISPANIC
 2 03/11/2011 12:03      QUEENS    FALSE                   NA             NA       NA             65+           M       WHITE
 3 04/14/2021 21:08      BRONX     TRUE                    NA             NA       NA             18-24         M       BLACK
 4 12/10/2021 19:30      BRONX     FALSE                   NA             NA       NA             25-44         M       BLACK
 5 02/22/2021 00:18      MANHATTAN FALSE                   NA             NA       NA             25-44         M       BLACK HISPANIC
 6 03/07/2021 06:15      BROOKLYN  TRUE                    25-44          M        BLACK HISPANIC 25-44         M       WHITE HISPANIC
 7 07/21/2021 00:40      MANHATTAN FALSE                   25-44          M        BLACK          25-44         M       BLACK
 8 09/11/2021 20:20      MANHATTAN FALSE                   NA             NA       NA             18-24         M       BLACK
 9 05/09/2021 02:50      BRONX     TRUE                    25-44          M        BLACK          25-44         M       BLACK HISPANIC
10 04/23/2021 13:25      BROOKLYN  FALSE                   NA             NA       NA             18-24         M       BLACK
# … with 25,586 more rows
# i Use `print(n = ...)` to see more rows
>
```

Code:

```
NYPD_summary1 = subset(NYPD, select = -
c(INCIDENT_KEY, PRECINCT,
JURISDICTION_CODE,
LOCATION_DESC, X_COORD_CD,
Y_COORD_CD,
Latitude, Longitude, Lon_Lat))
```

And then I rename OCCUR_DATE and OCCUR_TIME to be more R friendly, and change date types to display.

```
> NYPD_summary2
# A tibble: 25,596 × 10
   DATE       TIME    BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE      VIC_AGE_GROUP VIC_SEX VIC_RACE
   <date>     <time>  <chr>     <lgl>                   <chr>          <chr>    <chr>          <chr>         <chr>   <chr>
 1 2006-08-27 05:35   BRONX     TRUE                    NA             NA       NA             25-44         F       BLACK HISPANIC
 2 2011-03-11 12:03   QUEENS    FALSE                   NA             NA       NA             65+           M       WHITE
 3 2021-04-14 21:08   BRONX     TRUE                    NA             NA       NA             18-24         M       BLACK
 4 2021-12-10 19:30   BRONX     FALSE                   NA             NA       NA             25-44         M       BLACK
 5 2021-02-22 00:18   MANHATTAN FALSE                   NA             NA       NA             25-44         M       BLACK HISPANIC
 6 2021-03-07 06:15   BROOKLYN  TRUE                    25-44          M        BLACK HISPANIC 25-44         M       WHITE HISPANIC
 7 2021-07-21 00:40   MANHATTAN FALSE                   25-44          M        BLACK          25-44         M       BLACK
 8 2021-09-11 20:20   MANHATTAN FALSE                   NA             NA       NA             18-24         M       BLACK
 9 2021-05-09 02:50   BRONX     TRUE                    25-44          M        BLACK          25-44         M       BLACK HISPANIC
10 2021-04-23 13:25   BROOKLYN  FALSE                   NA             NA       NA             18-24         M       BLACK
# … with 25,586 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

Code:

```
library(lubridate)
NYPD_summary2 <- NYPD_summary1 %>%
rename(DATE = `OCCUR_DATE`,
TIME = `OCCUR_TIME`) %>%
mutate(DATE = mdy(DATE))
```

Furthermore, I find there are many missing data which shows “NA”,
I remove all the rows containing “NA”.

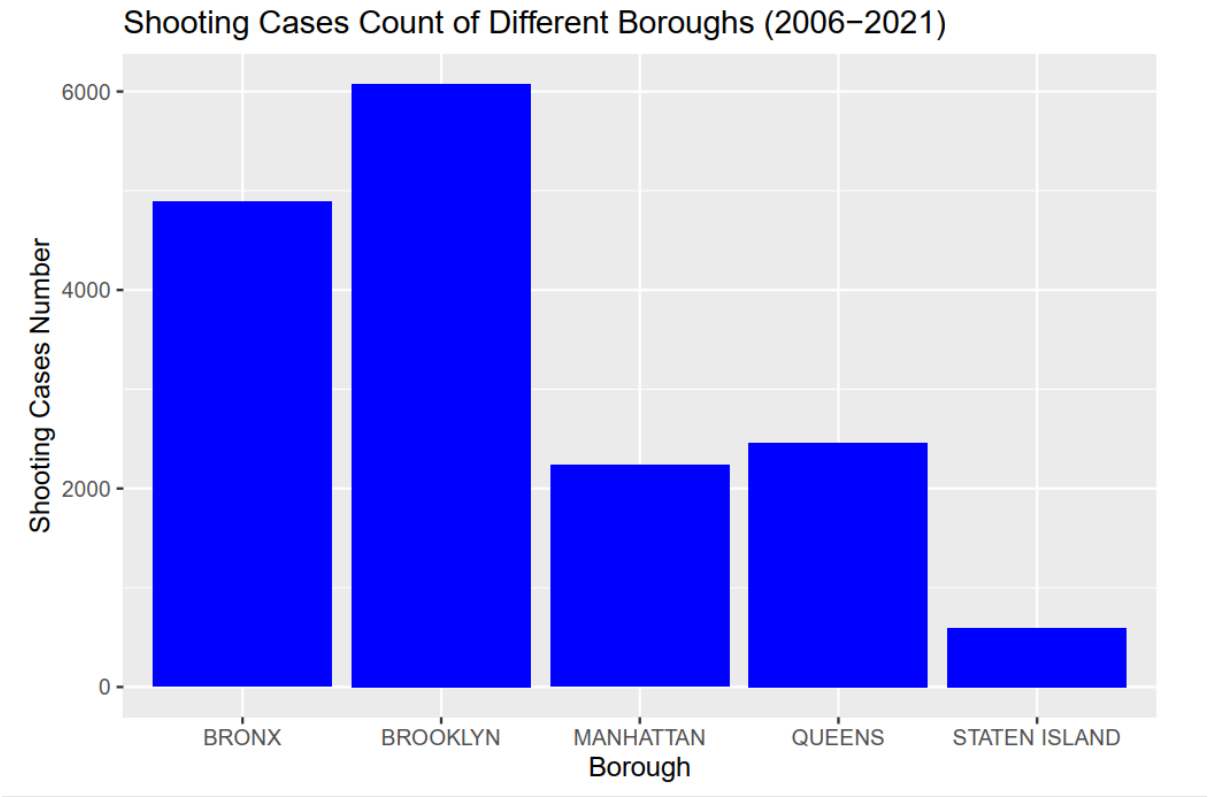
```
> NYPD_summary
# A tibble: 16,252 x 11
   DATE       TIME    BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE      VIC_AGE_GROUP VIC_SEX VIC_RACE
   <date>     <time>  <chr>     <lgl>                   <chr>          <chr>    <chr>          <chr>         <chr>   <chr>
 1 2021-03-07 06:15   BROOKLYN  TRUE                    25-44          M        BLACK HISPANIC 25-44         M       WHITE HISPANIC
 2 2021-07-21 00:40   MANHATTAN FALSE                   25-44          M        BLACK          25-44         M       BLACK
 3 2021-05-09 02:50   BRONX     TRUE                    25-44          M        BLACK          25-44         M       BLACK HISPANIC
 4 2021-06-16 23:22   BRONX     TRUE                    25-44          M        BLACK          25-44         F       BLACK
 5 2021-01-12 22:12   BROOKLYN  FALSE                   18-24          M        BLACK          18-24         M       BLACK
 6 2021-09-04 20:18   MANHATTAN FALSE                   18-24          M        WHITE HISPANIC 18-24         M       ASIAN / PACIFIC IS…
 7 2021-06-16 23:22   BRONX     FALSE                   18-24          M        WHITE HISPANIC 25-44         F       BLACK
 8 2021-09-29 12:50   BRONX     FALSE                   18-24          M        BLACK          <18           M       BLACK HISPANIC
 9 2021-03-10 07:30   MANHATTAN FALSE                   25-44          M        BLACK          18-24         M       BLACK
10 2021-08-13 01:00   QUEENS    TRUE                    18-24          M        BLACK          18-24         F       BLACK
# … with 16,242 more rows
# i Use `print(n = ...)` to see more rows
>
```

Code:

```
NYPD_summary <- NYPD_summary2 %>%
na.omit(NYPD_summary)
```

Visualizing

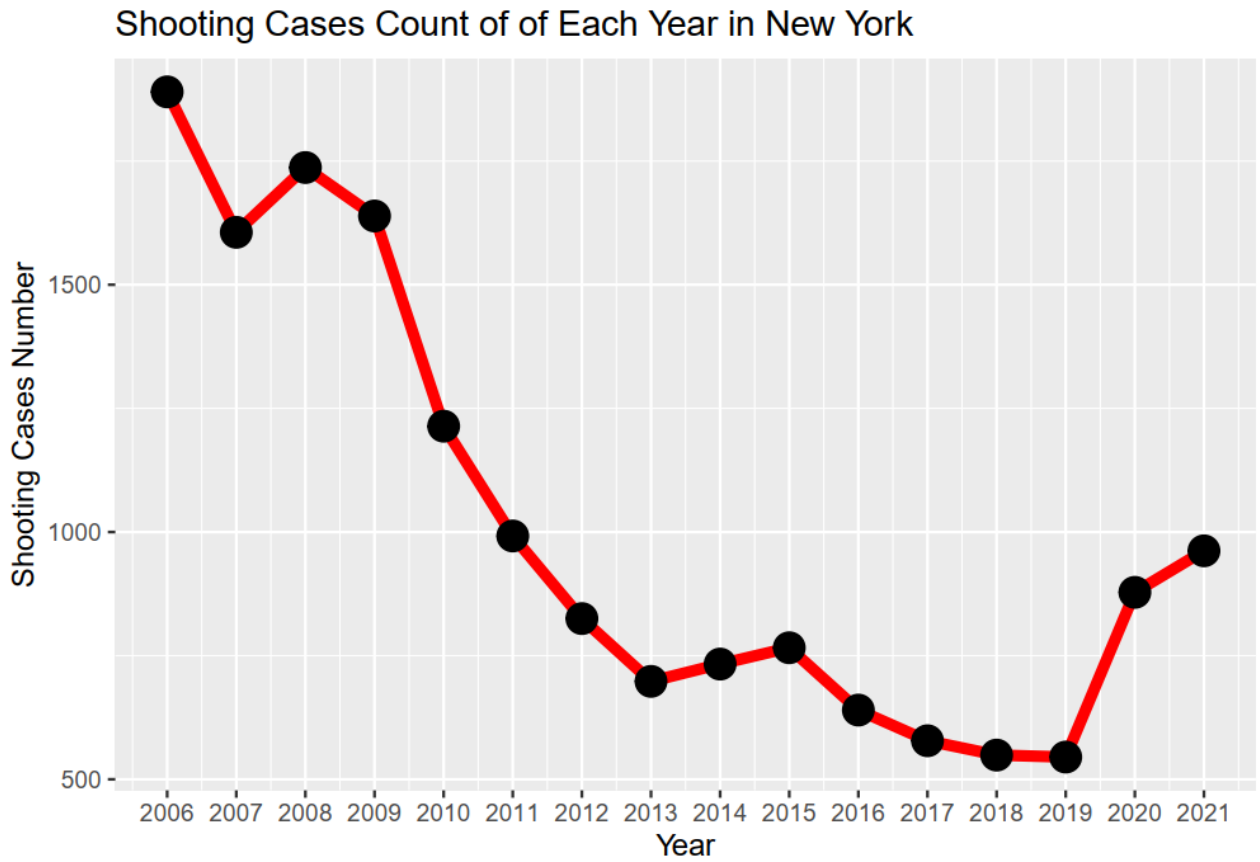
Generate a new table “Different_boroughs” by counting shooting cases from 2006 to 2021 of different boroughs in New York. Visualizing the by a bar plot.



Code:

```
library(dplyr)
Different_boroughs <- NYPD_summary %>%
group_by(BORO) %>% summarise(count_nums = n())
library(ggplot2)
ggplot(Different_boroughs, aes(x=BORO,
y=count_nums)) +
geom_bar(stat="identity", fill ="blue") +
labs(x="Borough", y="Shooting Cases Number")+
ggtitle("Shooting Cases Count of Different Boroughs
(2006-2021)")
```

Generate another new table “Cases_Per_Year” based on the total shooting cases counting by each year from 2006 to 2021 in New York. Visualizing the data by a line plot.

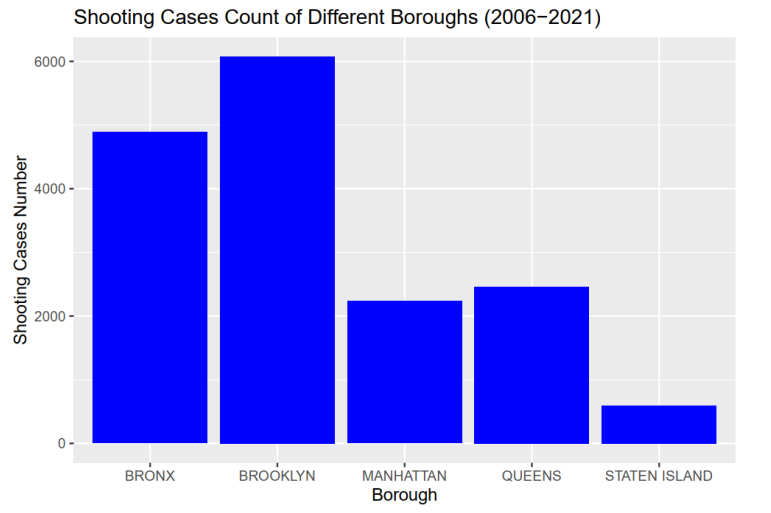


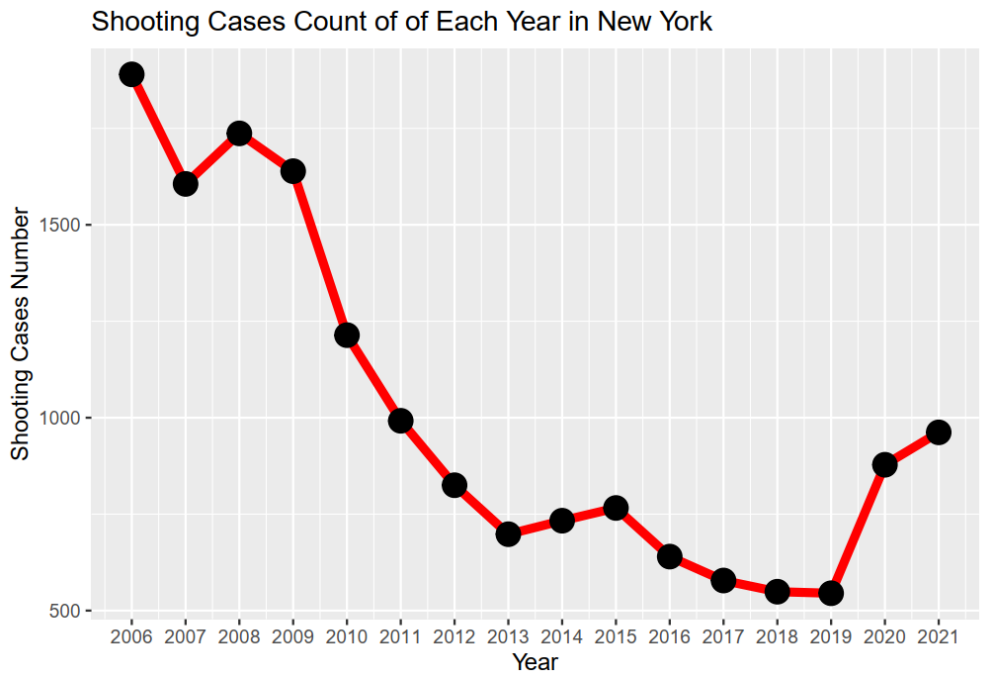
Code:

```
NYPD_summary$Year <- as.numeric(format(NYPD_summary$DATE, "%Y"))
Cases_Per_Year <- NYPD_summary %>% group_by(Year) %>% summarise(count_nums = n())
ggplot(Cases_Per_Year, aes(x=Year,y=count_nums))+
geom_line(color="red", size=2)+
geom_point(shape=21, color="black", fill="black", size=5)+
labs(x="Year", y="Shooting Cases Number")+
scale_x_continuous(breaks=seq(from = 2006, to = 2021, by = 1))+
ggtitle("Shooting Cases Count of of Each Year in New York")
```

Analysis

According to the bar plot, the total maximum cases borough is Brooklyn, the number is 6074. And the total minimum cases borough is Staten Island, the number is 591. If we look at the map of New York below, we will find the area of Brooklyn and Staten Island is about the same, but the total cases of the former is almost 10 times as the latter. That's interesting, maybe one of the reason is that Brooklyn is close to the city center and there are downtowns, and Staten Island is just at the suburb, far from the crowd area.

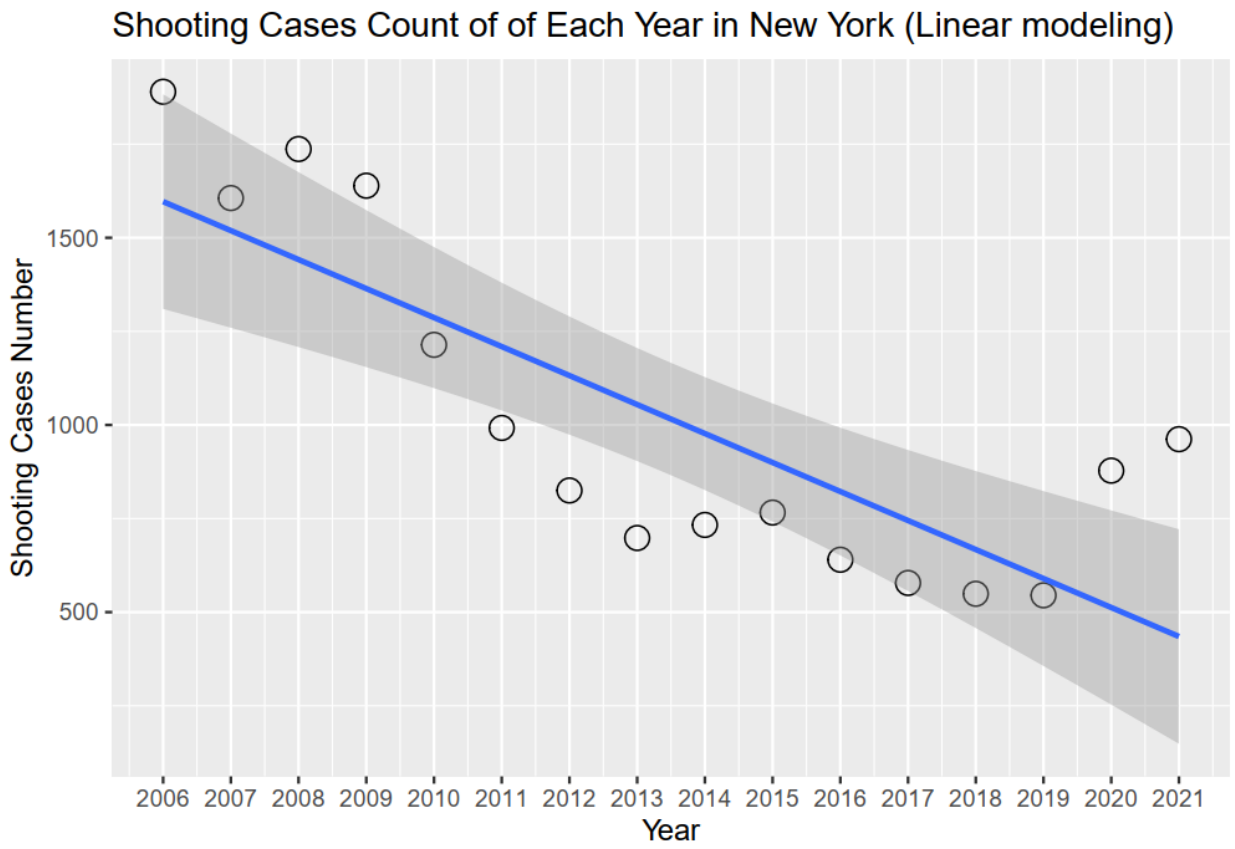




Based on the line plot, the total maximum cases year is 2006, the number is 1890. And the total minimum cases year is 2019, the number is 545. Since 2006, the cases number are mainly decreasing with the year until 2019, but then has an obvious increase in 2020 and 2021. There are several question worthy deep investigation. Why did the cases has a period peak in 2008, due to the financial crisis that year? Why did the shooting cases increase in recent two years? Does that has any relation to the COVID-19 pandemic? If so, with the ending of pandemic, can we predict that the decay trend (from 2006 to 2019) will continue by tracking the case number after 2022?

Modeling

Modeling the plot above with a linear fashion.



Code:

```
ggplot(Cases_Per_Year, aes(x=Year,y=count_nums)) +
geom_point(shape=21, color="black", size=4) +
stat_smooth(method = lm)+
labs(x="Year", y="Shooting Cases Number")+
scale_x_continuous(breaks=seq(from = 2006, to = 2021, by = 1))+
ggtitle("Shooting Cases Count of of Each Year in New York (Linear modeling)")
```

Conclusion and Bias

If we review the bar plot of, we can draw a conclusion that Brooklyn is the most dangerous borough in New York because it has the biggest number, and Staten Island is the safest borough. To my personal feeling, downtown area are always the heaven of all kinds of crime. But that maybe has some bias, we should consider the population of each borough, and then we can compare cases number based on per thousand people. If we have that data, maybe we can get another picture and tell a different story.

From a 15 years observation window, we found that Shooting Cases are decreasing year by year except for recent two year, we can draw a conclusion that it becomes safer now in New York comparing with the past. Is this kind of decrease a long term trend? I hope so and but it maybe my personal bias. I expect that the gun violence will reduce and the city around us will be a better and safer place in the future. But I need more data after 2022 to support this optimistic expectation.