NYPD Shooting Incident Data Report

Fei Ai

2022-08-14

(Notice: I break my work into several steps for more readable, but it's not exactly the same as the project steps definition, please review and take it as a reference, thanks!)

Background

For every couple of days, we can hear the frightening shooting news at some corner of the country. And every time, there will be someone being hurt, or even being killed, that's so terrible, right? Because of endless gun violence, it seems like the city becomes more dangerous year by year. It this true? I'm interesting about this question and I will conduct a related investigation.

Importing data

Firstly, I will 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.

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

Step 2: Let's read in the data and see what we have. The table "NYPD" looks like below:

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

```
## # A tibble: 25,596 x 19
##    INCID~1 OCCUR~2 OCCUR~3 BORO  PRECI~4 JURIS~5 LOCAT~6 STATI~7 PERP_~8 PERP_~9
##      <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>   <chr>
##  1  2.41e7 08/27/~ 05:35   BRONX      52       0 <NA>    TRUE    <NA>    <NA>
##  2  7.77e7 03/11/~ 12:03   QUEE~     106       0 <NA>    FALSE   <NA>    <NA>
##  3  2.27e8 04/14/~ 21:08   BRONX      42       0 COMMER~ TRUE    <NA>    <NA>
##  4  2.38e8 12/10/~ 19:30   BRONX      52       0 <NA>    FALSE   <NA>    <NA>
##  5  2.25e8 02/22/~ 00:18   MANH~      34       0 <NA>    FALSE   <NA>    <NA>
##  6  2.25e8 03/07/~ 06:15   BROO~      75       0 <NA>    TRUE    25-44   M
##  7  2.31e8 07/21/~ 00:40   MANH~      32       0 <NA>    FALSE   25-44   M
##  8  2.33e8 09/11/~ 20:20   MANH~      26       2 MULTI ~ FALSE   <NA>    <NA>
##  9  2.28e8 05/09/~ 02:50   BRONX      41       2 MULTI ~ TRUE    25-44   M
## 10  2.27e8 04/23/~ 13:25   BROO~      67       0 <NA>    FALSE   <NA>    <NA>
## # ... with 25,586 more rows, 9 more variables: 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>, and
## #   abbreviated variable names 1: INCIDENT_KEY, 2: OCCUR_DATE, 3: OCCUR_TIME,
## #   4: PRECINCT, 5: JURISDICTION_CODE, 6: LOCATION_DESC,
## #   7: STATISTICAL_MURDER_FLAG, 8: PERP_AGE_GROUP, 9: PERP_SEX
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Tidying and Transforming

Step 3: After looking at the NYPD, I would like to tidy the dataset, I don't need INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude and Lon_Lat for the analysis I am planning, so I will get rid of these. Then the table "NYPD_summary1" looks like below:

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

```
## # A tibble: 25,596 x 10
##    OCCUR~1 OCCUR~2 BORO  STATI~3 PERP_~4 PERP_~5 PERP_~6 VIC_A~7 VIC_SEX VIC_R~8
##    <chr>   <time>  <chr> <lgl>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1 08/27/~ 05:35   BRONX TRUE    <NA>    <NA>    <NA>    25-44   F       BLACK ~
##  2 03/11/~ 12:03   QUEE~ FALSE   <NA>    <NA>    <NA>    65+     M       WHITE
##  3 04/14/~ 21:08   BRONX TRUE    <NA>    <NA>    <NA>    18-24   M       BLACK
##  4 12/10/~ 19:30   BRONX FALSE   <NA>    <NA>    <NA>    25-44   M       BLACK
##  5 02/22/~ 00:18   MANH~ FALSE   <NA>    <NA>    <NA>    25-44   M       BLACK ~
##  6 03/07/~ 06:15   BROO~ TRUE    25-44   M       BLACK ~ 25-44   M       WHITE ~
##  7 07/21/~ 00:40   MANH~ FALSE   25-44   M       BLACK   25-44   M       BLACK
##  8 09/11/~ 20:20   MANH~ FALSE   <NA>    <NA>    <NA>    18-24   M       BLACK
##  9 05/09/~ 02:50   BRONX TRUE    25-44   M       BLACK   25-44   M       BLACK ~
## 10 04/23/~ 13:25   BROO~ FALSE   <NA>    <NA>    <NA>    18-24   M       BLACK
## # ... with 25,586 more rows, and abbreviated variable names 1: OCCUR_DATE,
## #   2: OCCUR_TIME, 3: STATISTICAL_MURDER_FLAG, 4: PERP_AGE_GROUP, 5: PERP_SEX,
## #   6: PERP_RACE, 7: VIC_AGE_GROUP, 8: VIC_RACE
## # i Use `print(n = ...)` to see more rows
```

Step 4: And then I will rename OCCUR_DATE and OCCUR_TIME to be more R friendly, and change date types to display. Now the table "NYPD_summary2" looks like below:

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

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

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

Step 5: Furthermore, I find there are many missing data which shows "NA", I will remove all the rows containing "NA". After all process, I got my table "NYPD_summary" like below:

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

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

Step 6: I'm interested in which district is safer in New York. So I generate a new table "Different_boroughs" by counting shooting cases from 2006 to 2021 of different boroughs in New York. This is the table looks like below:

```
library(dplyr)
Different_boroughs <- NYPD_summary %>% group_by(BORO) %>% summarise(count_nums = n())
Different_boroughs
```

```
## # A tibble: 5 x 2
```

```
##   BORO          count_nums
##   <chr>              <int>
## 1 BRONX               4890
## 2 BROOKLYN            6074
## 3 MANHATTAN           2235
## 4 QUEENS              2462
## 5 STATEN ISLAND        591
```

Step 7: Also I want to know the shooting cases trend by yearly. So I generate another new table "Cases_Per_Year" based on the total shooting cases counting by each year from 2006 to 2021 in New York.This is the table looks like below:

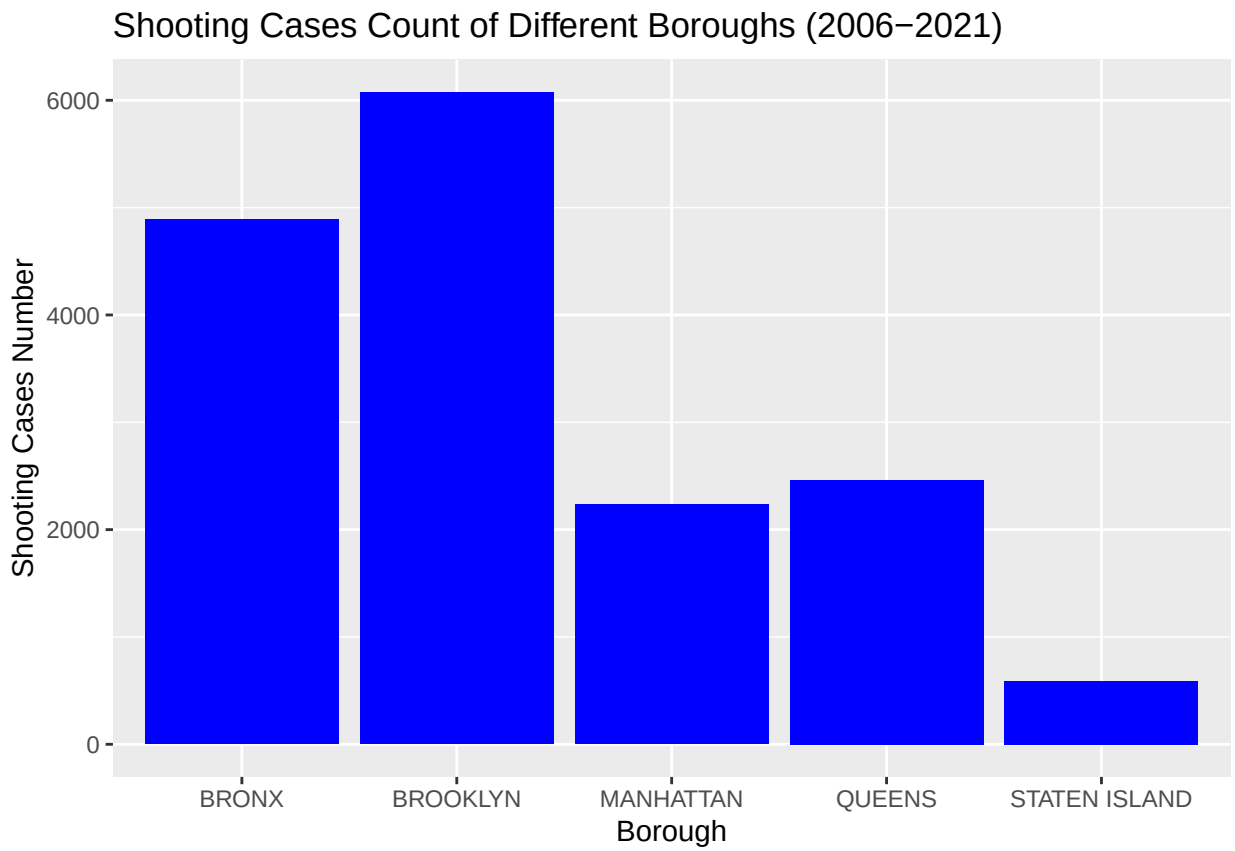
```
NYPD_summary$Year <- as.numeric(format(NYPD_summary$DATE, "%Y"))
Cases_Per_Year <- NYPD_summary %>% group_by(Year) %>% summarise(count_nums = n())
Cases_Per_Year
```

```
## # A tibble: 16 x 2
##     Year count_nums
##    <dbl>      <int>
##  1  2006       1890
##  2  2007       1606
##  3  2008       1737
##  4  2009       1639
##  5  2010       1214
##  6  2011        992
##  7  2012        825
##  8  2013        698
##  9  2014        733
## 10  2015        766
## 11  2016        640
## 12  2017        578
## 13  2018        549
## 14  2019        545
## 15  2020        878
## 16  2021        962
```

Visualizations

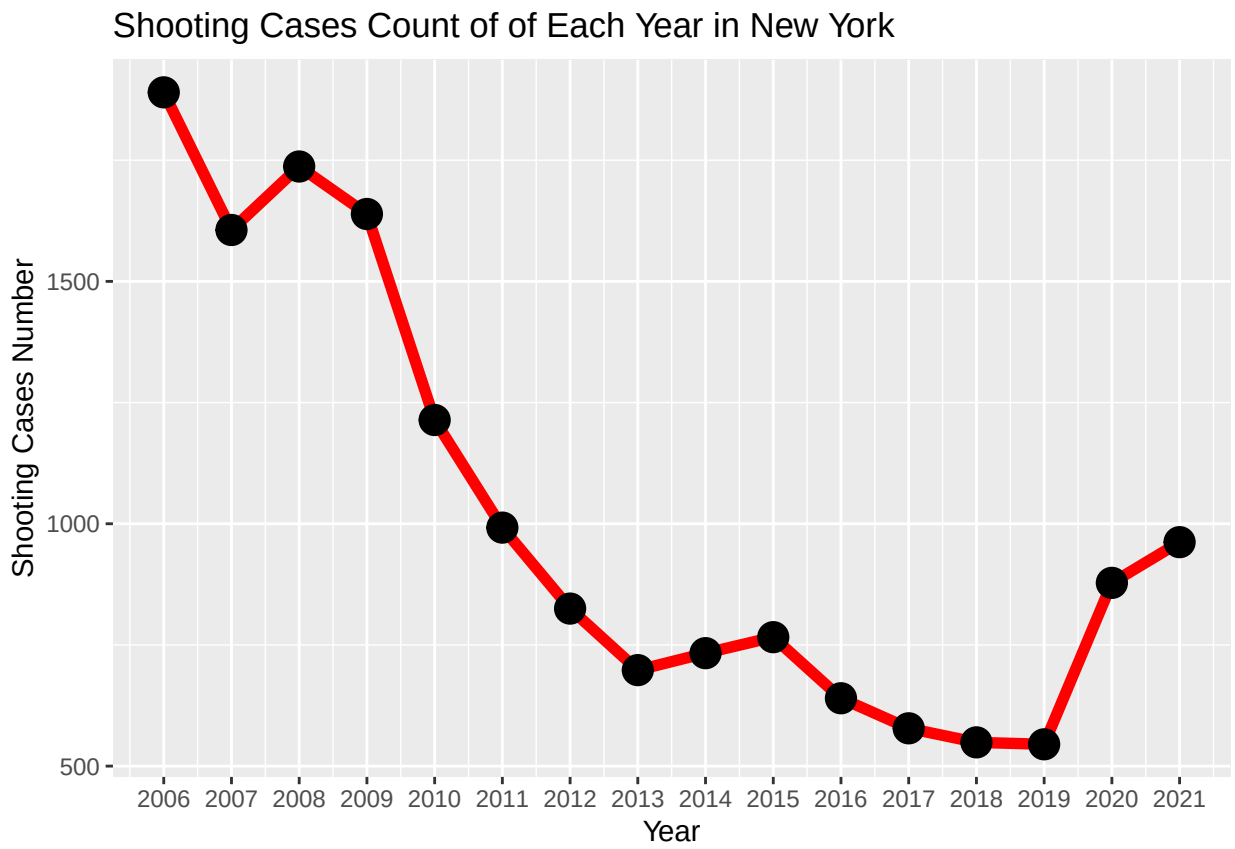
Step 8: Visualizing the data in the table1 "Different_boroughs" by a bar plot.

```
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)")
```



Step 9: Visualizing the data in the table2 "Cases_Per_Year" by a line plot.

```
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")
```

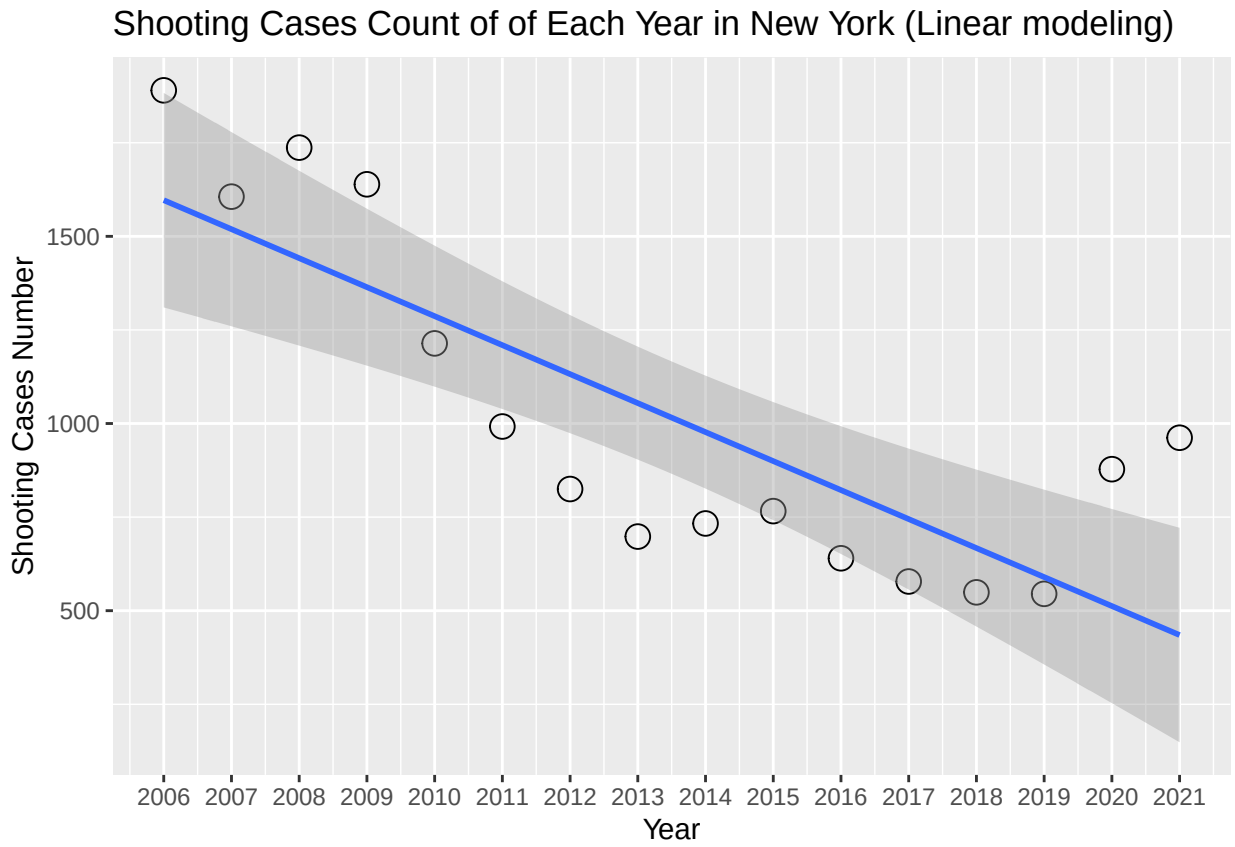


Modeling

Step 10: Modeling the plot above with a linear fashion.

```
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)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Analysis

According to the bar plot of "Shooting Cases Count of Different Boroughs (2006-2021)", there are case data collected from five districts in New York, they are: Bronx, Brooklyn, Manhattan, Queens and Staten Island. The total maximum cases borough is Brooklyn, the number is 6074, which reach 37.4% of the total cases in New York. And the total minimum cases borough is Staten Island, the number is 591, the percentage is about 3.6%. If we look at the map of New York on Google Maps, 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 of "Shooting Cases Count of of Each Year in New York", there are case data collected from 2006 to 2021 in New York. 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 (except 2008) with the year until 2019, but then has an obvious increase in 2020 and 2021. There are several question worthy deep investigation. Does 2006 has the highest case number in all the recorded history? We should explore more data before 2006, and then make a judge. 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?

Conclusion and Bias Identification

If we review the bar plot of "Shooting Cases Count of Different Boroughs (2006-2021)", we can draw a conclusion that Brooklyn is the most dangerous borough in New York because it has the biggest number,

so many shooting cases and many people were killed, and Staten Island is the safest borough in New York because it has the smallest number. That seems make sense. To my personal feeling, downtown area are always the heaven of all kinds of crime. But that maybe has some bias, to some extent, we should consider the population of each borough, and then we can compare cases number based on per million/thousand people. I don't know, if we have that data, maybe we can get another picture and tell a different story.

For the "Shooting Cases Count of of Each Year in New York", 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 I believe 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. On the other hand, is the decrease a result of the implementation of the law "Ammunition Regulation"? Maybe. Also, I notice that the case number increase suddenly in 2008 and 2020, which most likely related with economic crisis. I think this is my second possible bias, I need to import specific economic data in the history of New York (2006 to 2021 and the year before 2006), take a look and conduct analysis. If that's true, perhaps, the ultimate solution for a city to reduce gun violence is developing economic and make everyone employed.