Data Analysis of NYPD Shooting Incident Data (Historic)_KJ

Kexin Jiao

2023-11-18

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
## Warning: package 'wesanderson' was built under R version 4.3.2
```

This is a manually extracted data listing all the shooting incidents that occurred in NYC between 2006 and 2022. NYPD Shooting Incident Data (Historic).CSV file

```
data_url<-"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
df<-read.csv(data_url)
```

1. Data rangling

1) Quick view of the data structure

```
head(df)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
## 2    137471050 06/27/2014   17:40:00    BRONX                        40
## 3    147998800 11/21/2015   03:56:00   QUEENS                       108
## 4    146837977 10/09/2015   18:30:00    BRONX                        44
## 5     58921844 02/19/2009   22:58:00    BRONX                        47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                        81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                   false
## 2                 0                                                   false
## 3                 0                                                    true
## 4                 0                                                   false
## 5                 0                                                    true
## 6                 0                                                    true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
```

```
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                             25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                      Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

2) Remove the columns that are not related to our analysis

We plan to visualize and analyze the dependence of shooting cases on year and borough zones. There are several columns in the data frame that are not highly related to our goal, for example, the latitude and the longitude where the shooting happened. We will first of all remove those irrelated columns.

```
df_cln<-df %>% select(-c("LOC_CLASSFCTN_DESC","LOCATION_DESC","X_COORD_CD","Y_COORD_CD","Latitude","Long
str(df_cln)
```

```
## 'data.frame':    27312 obs. of  14 variables:
##  $ INCIDENT_KEY           : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE             : chr  "05/27/2021" "06/27/2014" "11/21/2015" "10/09/2015" ...
##  $ OCCUR_TIME             : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                   : chr  "QUEENS" "BRONX" "QUEENS" "BRONX" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "" "" "" "" ...
##  $ PRECINCT               : int  105 40 108 44 47 81 114 81 105 101 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
```

3) Remove empty values, null values, or nonsense values

Some columns in this data frame contain too many empty (>10% by number) values and/or null values. We will remove those columns before further data cleaning steps. After that, we will delete the rows that have empty, null, or nonsense values to completely clean the data frame. Finally, we will convert the occur date of shooting into standerized date-time format.

```
for (i in colnames(df_cln)) {
  l=length(df_cln[,i][df_cln[,i] != ""] %>% .[. != "(null)"] %>% .[. != "UNKNOWN"])
  #remove the columns contain too many (>10%) empty or null values
  if (l<=25000) {
    df_cln[,i]<-NULL
  }else if (l<nrow(df_cln)) {
  # remove the rows having empty and null values in the rest columns
```

```
    df_cln<-subset(df_cln,df_cln[,i]!="UNKNOWN")
   }
}
#remove the weird values that do not make sense
unique(df_cln$VIC_AGE_GROUP)
```

```
## [1] "18-24" "25-44" "<18"   "45-64" "65+"   "1022"
```

```
x<-which(grepl("1022",df_cln$VIC_AGE_GROUP))
df_cln<-df_cln[-x,]
#convert occur date into data-time format
df_cln<-df_cln %>% mutate(OCCUR_DATE=mdy(OCCUR_DATE))

str(df_cln)
```

```
## 'data.frame':    27200 obs. of  10 variables:
##  $ INCIDENT_KEY          : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE            : Date, format: "2021-05-27" "2014-06-27" ...
##  $ OCCUR_TIME            : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                  : chr  "QUEENS" "BRONX" "QUEENS" "BRONX" ...
##  $ PRECINCT              : int  105 40 108 44 47 81 114 81 105 101 ...
##  $ JURISDICTION_CODE     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ VIC_AGE_GROUP         : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX               : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE              : chr  "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
```

4) Extract the year number from OCCUR_DATE and save it to a new column

```
df_cln$YEAR<-as.numeric(format(df_cln$OCCUR_DATE,"%Y"))

str(df_cln)
```

```
## 'data.frame':    27200 obs. of  11 variables:
##  $ INCIDENT_KEY          : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE            : Date, format: "2021-05-27" "2014-06-27" ...
##  $ OCCUR_TIME            : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                  : chr  "QUEENS" "BRONX" "QUEENS" "BRONX" ...
##  $ PRECINCT              : int  105 40 108 44 47 81 114 81 105 101 ...
##  $ JURISDICTION_CODE     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ VIC_AGE_GROUP         : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX               : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE              : chr  "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
##  $ YEAR                  : num  2021 2014 2015 2015 2009 ...
```

2. Data visualization

1) Plotting of shooting cases that happened in each borough per year

Create data frames grouped by borough and year.

```
df_by_Borough_Year<-df_cln %>%
  group_by(BORO,YEAR) %>%
  tally() %>%
  mutate(CASES=n) %>%
```

```
  select(-n)
head(df_by_Borough_Year)
```

```
## # A tibble: 6 x 3
## # Groups:   BORO [1]
##   BORO   YEAR CASES
##   <chr> <dbl> <int>
## 1 BRONX  2006   568
## 2 BRONX  2007   530
## 3 BRONX  2008   517
## 4 BRONX  2009   523
## 5 BRONX  2010   523
## 6 BRONX  2011   571
```

plot the number of shooting cases for each borough as a function of year.

```
df_BRONX<-subset(df_by_Borough_Year,df_by_Borough_Year[,"BORO"]=="BRONX")
df_BROOKLYN<-subset(df_by_Borough_Year,df_by_Borough_Year[,"BORO"]=="BROOKLYN")
df_MANHATTAN<-subset(df_by_Borough_Year,df_by_Borough_Year[,"BORO"]=="MANHATTAN")
df_QUEENS<-subset(df_by_Borough_Year,df_by_Borough_Year[,"BORO"]=="QUEENS")
df_STATEN_ISLAND<-subset(df_by_Borough_Year,df_by_Borough_Year[,"BORO"]=="STATEN ISLAND")


color_A<-c("BRONX"="maroon","BROOKLYN"="springgreen4","MANHATTAN"="steelblue","QUEENS"="red","STATEN IS
ggplot()+
  geom_line(data=df_BRONX,aes(x=YEAR,y=CASES,color="BRONX"),linewidth=1)+
  geom_point(data=df_BRONX,aes(x=YEAR,y=CASES,color="BRONX"),size=3)+
  geom_line(data=df_BROOKLYN,aes(x=YEAR,y=CASES,color="BROOKLYN"),linewidth=1)+
  geom_point(data=df_BROOKLYN,aes(x=YEAR,y=CASES,color="BROOKLYN"),size=3)+
  geom_line(data=df_MANHATTAN,aes(x=YEAR,y=CASES,color="MANHATTAN"),linewidth=1)+
  geom_point(data=df_MANHATTAN,aes(x=YEAR,y=CASES,color="MANHATTAN"),size=3)+
  geom_line(data=df_QUEENS,aes(x=YEAR,y=CASES,color="QUEENS"),linewidth=1)+
  geom_point(data=df_QUEENS,aes(x=YEAR,y=CASES,color="QUEENS"),size=3)+
  geom_line(data=df_STATEN_ISLAND,aes(x=YEAR,y=CASES,color="STATEN ISLAND"),linewidth=1)+
  geom_point(data=df_STATEN_ISLAND,aes(x=YEAR,y=CASES,color="STATEN_ISLAND"),size=3)+

  labs(title="Cases in every year in each boroughs",x="Year",y="Cases",color="Legend",tag="A")+
  scale_color_manual(name="NYPD Boroughs", values=color_A)
```

A

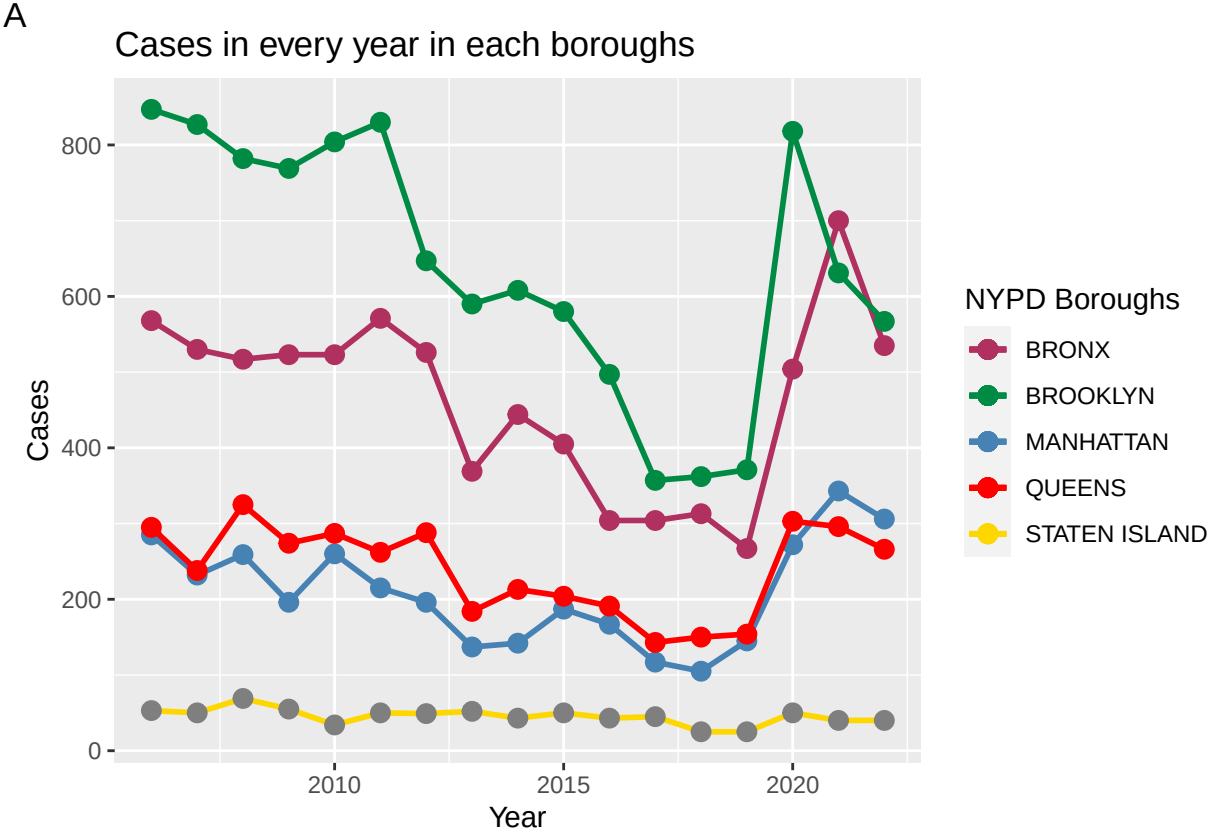

Figure A shows the number of cases of shooting happened in each borough per year.An overall decreasing trend of cases before 2019 and a sudden increase after 2020 for most of the boroughs expect STATEN ISLAND are observed.

2) plotting of cases of shooting as a function of borough

Create data frames grouped by borough.

```
df_by_Borough<-df_cln %>%
  group_by(BORO) %>%
  tally() %>%
  mutate(CASES=n) %>%
  select(-n)
head(df_by_Borough)
```

```
## # A tibble: 5 x 2
##   BORO          CASES
##   <chr>         <int>
## 1 BRONX          7903
## 2 BROOKLYN      10887
## 3 MANHATTAN      3564
## 4 QUEENS         4073
## 5 STATEN ISLAND   773
```

plot the number of shooting cases as a function of borough.

```
ggplot(data=df_by_Borough,aes(x=BORO,y=CASES))+
  geom_bar(stat="identity",fill=color_A,width=0.5)+
```

```
  labs(title="Total Cases In Each Borough",x="Boroughs",y="Cases",tag="B")
```

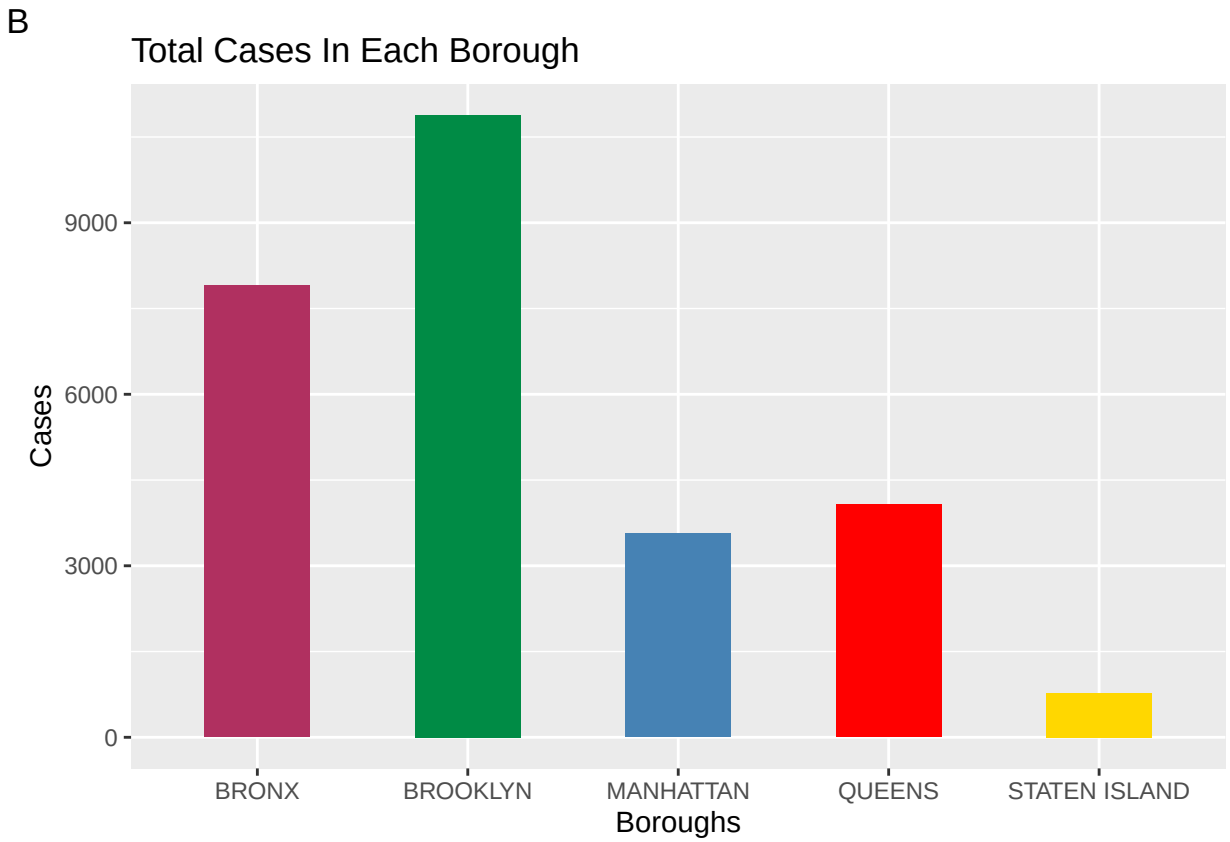

Figure B shows the total shooting cases happened in each borough from 2006 to 2022.BROOKLYN has the greatest number of shooting cases while Staten Island has the smallest number of shooting cases.

3) plotting of cases of shooting as a function of year

Create data frames grouped by year.

```
df_by_Year<-df_cln %>%
  group_by(YEAR) %>%
  tally() %>%
  mutate(CASES=n) %>%
  select(-n)
head(df_by_Year)
```

```
## # A tibble: 6 x 2
##    YEAR CASES
##   <dbl> <int>
## 1  2006  2048
## 2  2007  1877
## 3  2008  1952
## 4  2009  1817
## 5  2010  1908
## 6  2011  1928
```

plot the number of shooting cases as a function of year.

```
ggplot(data=df_by_Year,aes(x=YEAR,y=CASES))+
  geom_bar(stat="identity",fill="steelblue",width=0.3)+
  labs(title="Total Cases In Each Year",x="Years",y="Cases",tag="C")
```

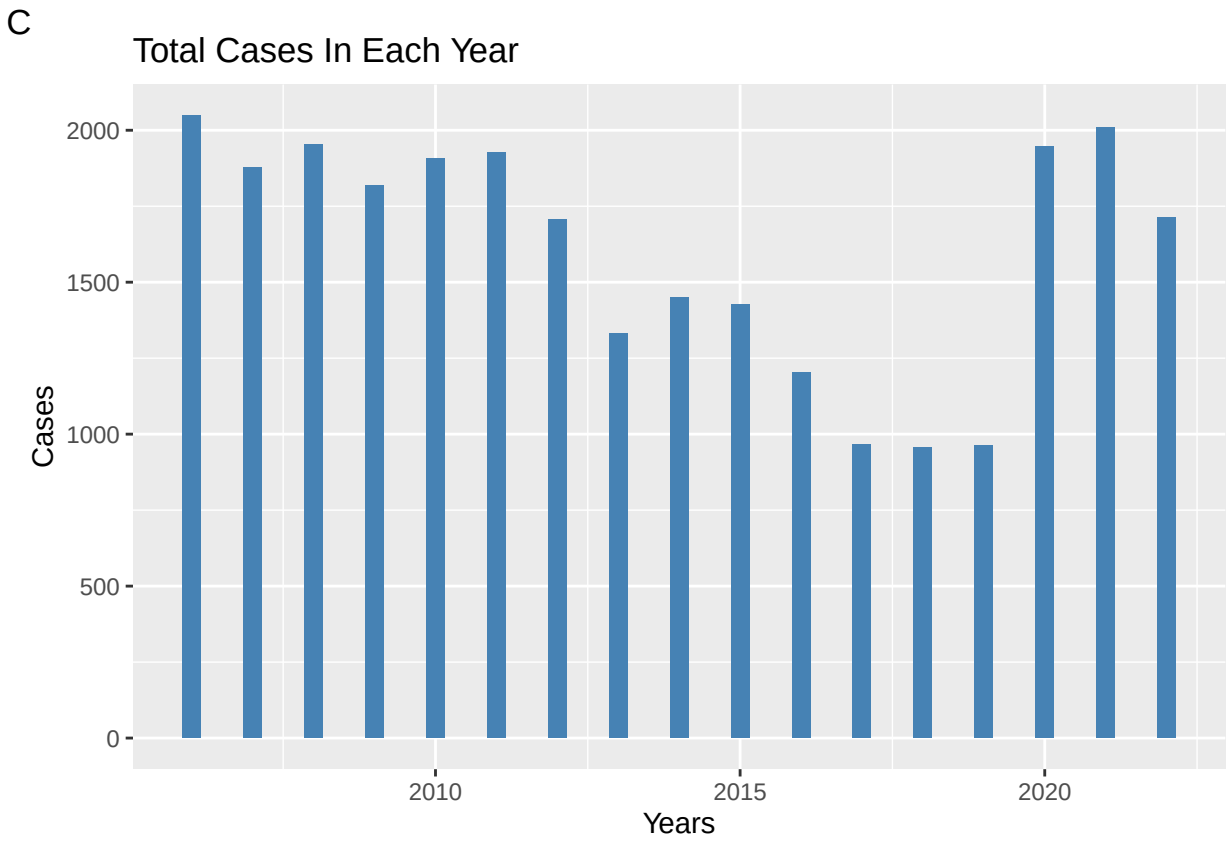

Figure C clearly reveals the decreasing of total shooting cases per year from 2006 to 2019 and a sudden increase at 2020. We would attribute this change to two factors. First, the global COVID-19 panic since early 2020 significantly affect the economy in NYC.The increased number of unemployment and the fatal health risk may negatively affect and destruct the socialeconomic structure in NYC and result in increased crime rate. Second, the NYPD's budget has been reduced by about $1 billion between fiscal years 2020 and 2021.One may expect a reduced level of patrol strength throughout the communities which may lead to the increased number shooting cases.Further investigation and data are required to support of deny out assumption.

3. Data Analysis

The maximum, minimum, and average number of shooting cases in each year

```
df_by_Year_Borough<-df_cln %>%
  group_by(YEAR,BORO) %>%
  tally() %>%
  mutate(CASES=n) %>%
  select(-n)

years<-unique(df_by_Year$YEAR)
max_cases<-tapply(df_by_Year_Borough$CASES,df_by_Year_Borough$YEAR,max)
min_cases<-tapply(df_by_Year_Borough$CASES,df_by_Year_Borough$YEAR,min)
```

```
avg_cases<-tapply(df_by_Year_Borough$CASES,df_by_Year_Borough$YEAR,mean)
max_boroughs<-rep("a",17)
min_boroughs<-rep("b",17)
df_case_by_year_analysis<-cbind(avg_cases,max_cases,min_cases)
df_case_by_year_analysis<-cbind(years,df_case_by_year_analysis)
rownames(df_case_by_year_analysis)=1:17
df_case_by_year_analysis<-as.data.frame(df_case_by_year_analysis)

for (i in 1:17){
  max_boroughs[i]<-df_by_Year_Borough$BORO[which(df_by_Year_Borough$YEAR==df_case_by_year_analysis$years
  min_boroughs[i]<-df_by_Year_Borough$BORO[which(df_by_Year_Borough$YEAR==df_case_by_year_analysis$years
}
df_case_by_year_analysis<-cbind(df_case_by_year_analysis,max_boroughs,min_boroughs)
df_case_by_year_analysis
```

```
##    years avg_cases max_cases min_cases max_boroughs  min_boroughs
## 1   2006     409.6       847        53     BROOKLYN STATEN ISLAND
## 2   2007     375.4       827        50     BROOKLYN STATEN ISLAND
## 3   2008     390.4       782        69     BROOKLYN STATEN ISLAND
## 4   2009     363.4       769        55     BROOKLYN STATEN ISLAND
## 5   2010     381.6       804        34     BROOKLYN STATEN ISLAND
## 6   2011     385.6       830        50     BROOKLYN STATEN ISLAND
## 7   2012     341.2       647        49     BROOKLYN STATEN ISLAND
## 8   2013     266.4       590        52     BROOKLYN STATEN ISLAND
## 9   2014     290.0       608        43     BROOKLYN STATEN ISLAND
## 10  2015     285.2       580        50     BROOKLYN STATEN ISLAND
## 11  2016     240.4       497        43     BROOKLYN STATEN ISLAND
## 12  2017     193.2       357        45     BROOKLYN STATEN ISLAND
## 13  2018     191.0       362        25     BROOKLYN STATEN ISLAND
## 14  2019     192.4       371        25     BROOKLYN STATEN ISLAND
## 15  2020     389.4       818        50     BROOKLYN STATEN ISLAND
## 16  2021     402.0       700        40        BRONX STATEN ISLAND
## 17  2022     342.8       567        40     BROOKLYN STATEN ISLAND
```

The data frame lists the maximum, minimum, and average number of shooting cases happened in each year. The boroughs that have the maximum and minimum cases are shown. We found that the borough Staten Island always holds the least shooting cases from 2016 to 2022 while the borough Brooklyn always has the most cases except year 2021. This is clear evidence that reveals the social security level difference between different boroughs in NYC. Many factors, for example, community economy, population composition, and education level may affect the local criminal level. Further research are required for an detailed explanation.

4. Modeling for the trend of the number of shooting cases in BRONX

We investigated the change of number of shooting cases in BROOKLYN from 2006 to 2019, which is the year right before COCID-19 panic started (2020). It is found that although BROOKLYN has the highest number of cases, it decreased by years steadily. We used a linear model to fit this trend which will be reported in the coming section.

```
df_lm_BROOKLYN<-subset(df_BROOKLYN,df_BROOKLYN$YEAR<=2019)
mod_BROOKLYN<-lm(data=df_lm_BROOKLYN,CASES~YEAR)
summary(mod_BROOKLYN)
```

```
##
## Call:
## lm(formula = CASES ~ YEAR, data = df_lm_BROOKLYN)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -89.97 -37.44  -8.61  29.57 134.13
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84119.220   7978.610   10.54 2.02e-07 ***
## YEAR          -41.484      3.965  -10.46 2.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.8 on 12 degrees of freedom
## Multiple R-squared:  0.9012, Adjusted R-squared:  0.893
## F-statistic: 109.5 on 1 and 12 DF,  p-value: 2.19e-07
```

According to the summary, a decrease rate of about -41.5 cases/year (the slope of the straight line in Figure D) was determined with a reasonable p-value.

```
df_lm_BROOKLYN<-df_lm_BROOKLYN %>% mutate(Predicted=predict(mod_BROOKLYN))
color_D<-c("BROOKLYN"="steelblue","Predicted"="maroon")
ggplot()+
  geom_point(data=df_lm_BROOKLYN,aes(x=YEAR,y=CASES,color="BROOKLYN"),size=3)+
  geom_line(data=df_lm_BROOKLYN,aes(x=YEAR,y=CASES,color="BROOKLYN"),linewidth=1)+
  geom_point(data=df_lm_BROOKLYN,aes(x=YEAR,y=Predicted,color="Predicted"),size=3)+
  geom_line(data=df_lm_BROOKLYN,aes(x=YEAR,y=Predicted,color="Predicted"),linewidth=1)+
  labs(title="Observation and Prediction of the Cases in BROOKLYN",x="Year",y="Cases",color="Legend",tag
  scale_color_manual(name="Observation vs \n Prediction",values=color_D)
```

D

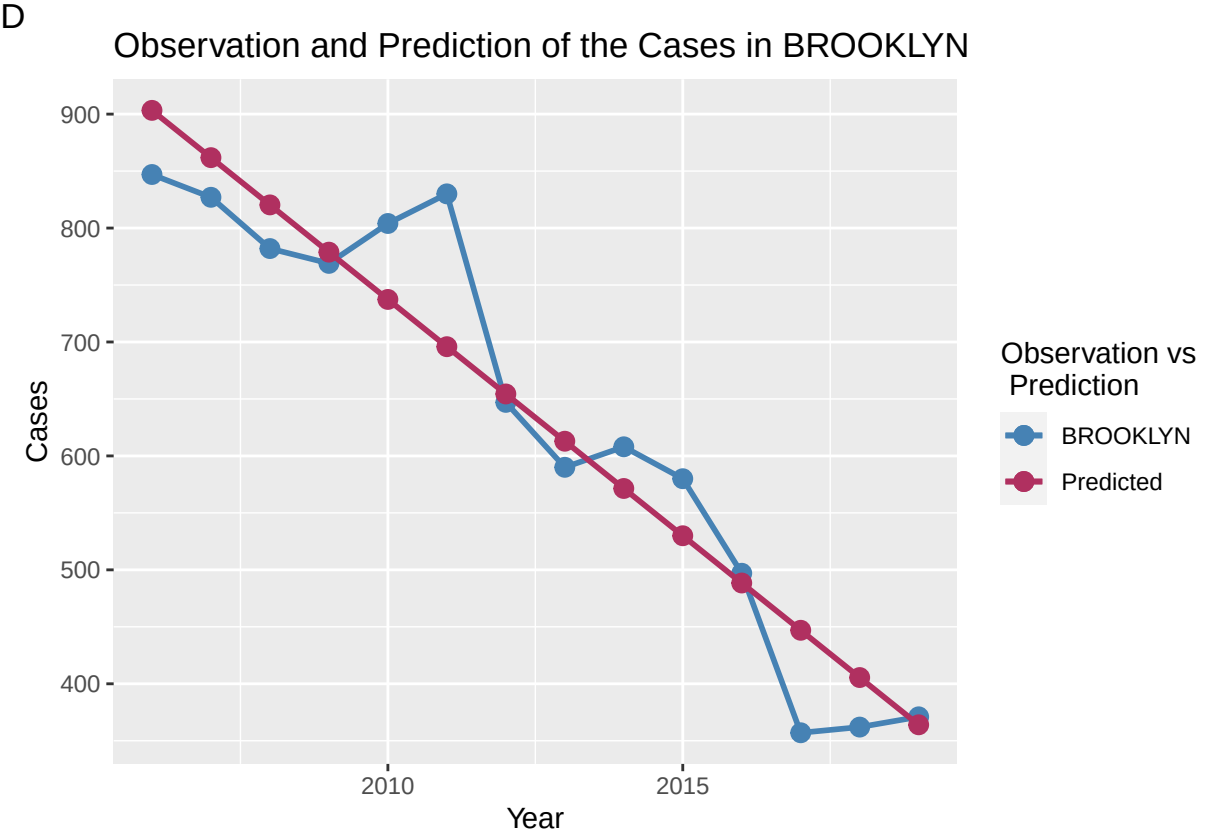

Figure D shows a good fitting of the predicted data for the original data of BROOKLYN between 2006 and 2019. It is reasonable to predict that the shooting cases will decrease again after COVID-19 panic completely ended if the NYC or BROOKLYN will follow what they have done between 2006 and 2019.

5. Discussion of possible bias in the data and analysis

1. The data was manually extracted from local reports. The number of reports may be dependent on the tendency that the local people would like to report to police and the local available police service.
2. The data only contains the shooting cases reported in NYC, it may not be able to reflect an overall crime situation in diferent boroughs.
3. During the COVID-19 panic (after 2019), the efficiency of report collecting may be varied due to the lack of common resource, which may affect the consistency of data.
4. A linear fitting of the cases number before 2019 may not reflect the real trend.