NYPD Shooting incident data analysis

5/21/2021

Introduction

purpose

In this article, I will **clean**, **visualize** and **analyze** NYPD Shooting incident data to make it easier to see the trend of cumulative shooting incidents and murder cases per boroughs.

contents

1. Data Cleaning
    i. Extract only the data needed for analysis
    ii. Date format conversion
    iii. Replace logical boolean data with integer type
    iv. Add the number of shooting occurrences and the cumulative number of it as columns.
    v. Add the cumulative number of murders as a column.
2. Visualization
    a. Cumulative number of shooting events according to the flow of the date by borough
    b. Cumulative number of murder according to the flow of the date by borough
    c. Percentage of murders in total number of shootings by borough
3. Analysis
    - Compare the total number of shooting cases and murder ones *by borough*.
    - Calculate the percentage of murders in shootings *by borough*
        - `BROOKLYN`
        - `BRONX`
        - `QUEENS`
        - `MANHATTAN`
        - `STATEN ISLAND`
4. Model
    - Compare the actual trend of shooting incidents with the predictive linear model
5. Conclusion and Bias Identification
    - Conclude the project report
    - Identify personal bias and mitigation method

1.Data Cleaning

Data used in this project is every shooting incident data list occurred in NYC from Jan. 1, 2006 to Dec. 31, 2020.

Data Importing Here you can see the original dataset below.

```
library(knitr)
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v dplyr   1.0.6
## v tibble  3.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

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

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

```
##
## -- Column specification --------------------------------------------------------
## 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_number(),
```

```
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

```
knitr::kable(head(nypd))
```

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	LOCATION_DESC	STATISTICAL_MURDER_FLAG	PERP_AGE_GROUP	PERP_SEX	PERP_RACE	VIC_AGE_GROUP	VIC_SEX	VIC_RACE	X_COORD_CD	Y_COORD_CD	Latitude	Longitude	Lon_Lat
201576	08/12/2009	23:40:00	QUEENS	0	NA	FALSE	NA	NA	NA	25-44	M	BLACK	1037453	196356	40.69781	POINT (-73.80734 40.69780814071699996)
205748	11/27/2009	19:30:00	BRONX	0	NA	FALSE	<18	M	BLACK	25-44	F	BLACK	1006738	237550	40.81870	POINT (-73.91837 40.81869973000005)
193116	03/02/2009	00:40:00	MANHATTAN	0	NA	FALSE	18-24	M	WHITE HIS-PANIC	18-24	M	BLACK HIS-PANIC	993427	227795	40.79192	POINT (-73.94547 40.791916091000076)
204192	06/24/2009	02:30:00	STATEN ISLAND	0	PVT HOUSE	TRUE	25-44	M	BLACK	25-44	F	BLACK	963814	171781	40.63806	POINT (-74.16711 40.63806398200006)
201480	08/22/2009	03:40:00	BRONX	0	NA	FALSE	25-44	M	BLACK HIS-PANIC	18-24	M	BLACK	1008225	240621	40.85455	POINT (-73.91733 40.85454734900003)
198256	06/07/2009	17:50:00	BROOKLYN	0	NA	FALSE	45-64	M	WHITE HIS-PANIC	25-44	M	BLACK	1009686	186966	40.67983	POINT (-73.90743 40.67982701600005)

Data cleaning Date format was converted for easier reading and I replaced logical boolean data, 'STATISTICAL_MURDER_FLAG' with integer type like '0' for FALSE and '1' for TRUE. And I added the number of shooting occurrences and the cumulative number of it as columns with the names of 'shooting' and 'cumshooting'. Also a column, 'cummurder', was added for the cumulative number of murders.

```
library(dplyr)
library(ggplot2)


# select only needed data
nypd_test <- drop_na(nypd) %>%
  select(-c(INCIDENT_KEY, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, JURISDIC

# change the date type
nypd_test <- nypd_test %>%
mutate(OCCUR_DATE = mdy(OCCUR_DATE))

#change logical boolean into int.
nypd_test$STATISTICAL_MURDER_FLAG [nypd_test$STATISTICAL_MURDER_FLAG == "TRUE"] <- 1
nypd_test$STATISTICAL_MURDER_FLAG [nypd_test$STATISTICAL_MURDER_FLAG == "FALSE"] <- 0
```

```
nypd_murder_boro <- nypd_test %>%
     group_by(BORO) %>%
#     summarize(STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
     select(BORO, OCCUR_DATE, STATISTICAL_MURDER_FLAG) %>%
     ungroup()

nypd_murder_boro_1 <- nypd_murder_boro %>%
     group_by(BORO, OCCUR_DATE) %>%
     summarize(STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG) %>%
     select(BORO, OCCUR_DATE, STATISTICAL_MURDER_FLAG) %>%
     ungroup()
```

```
## `summarise()` has grouped output by 'BORO', 'OCCUR_DATE'. You can override using the `.groups` argume
```

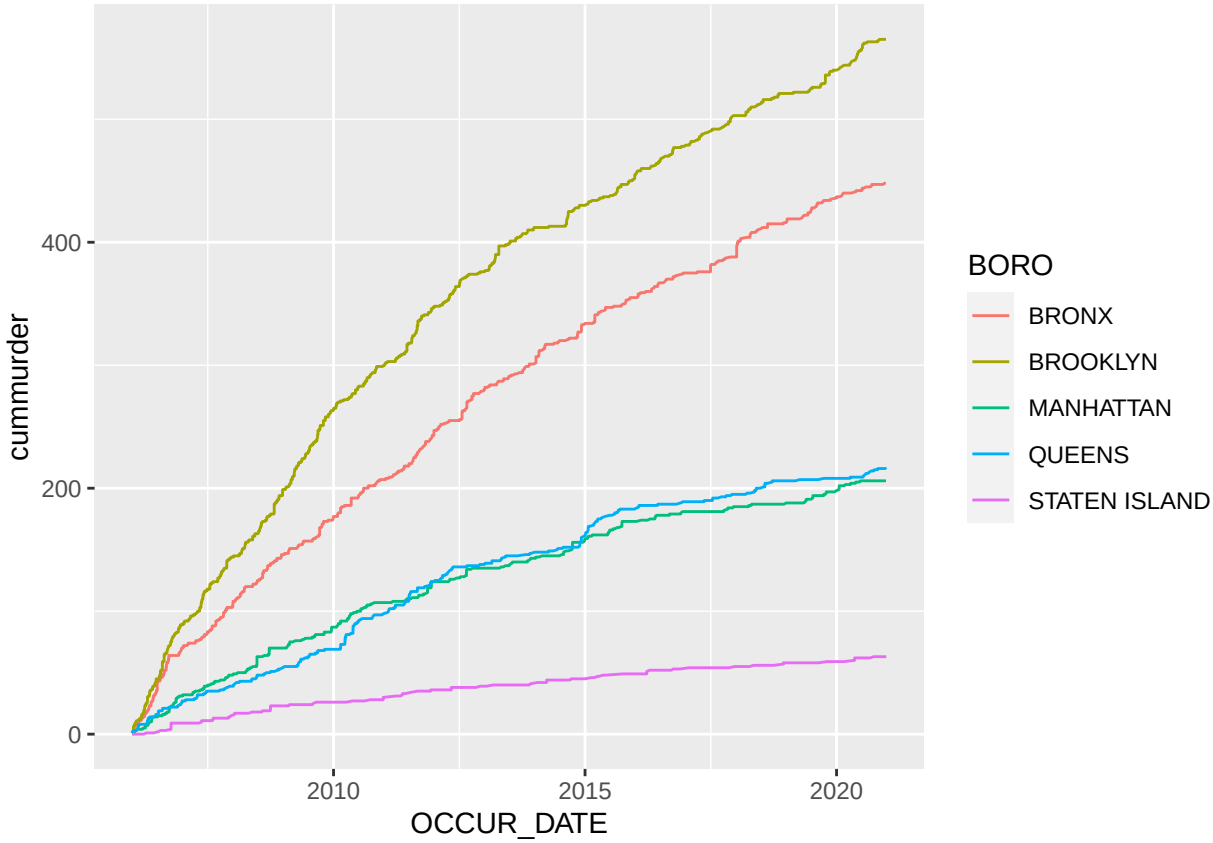
```
# add new columns
nypd_murder_boro_1$cummurder <- ave(nypd_murder_boro_1$STATISTICAL_MURDER_FLAG, nypd_murder_boro_1$BORO

nypd_murder_boro_1['shooting'] = 1

nypd_murder_boro_1$cumshooting <- ave(nypd_murder_boro_1$shooting, nypd_murder_boro_1$BORO, FUN = cumsu

nypd_murder_boro_1$murderpercent <- with(nypd_murder_boro_1, cummurder/cumshooting *100)
# show the data that will be used for analysis
knitr::kable(head(nypd_murder_boro_1))
```

BORO	OCCUR_DATE	STATISTICAL_MURDER_FLAG	cummurder	shooting	cumshooting	murderpercent
BRONX	2006-01-01	0	0	1	1	0
BRONX	2006-01-01	0	0	1	2	0
BRONX	2006-01-04	0	0	1	3	0
BRONX	2006-01-05	0	0	1	4	0
BRONX	2006-01-06	0	0	1	5	0
BRONX	2006-01-06	0	0	1	6	0

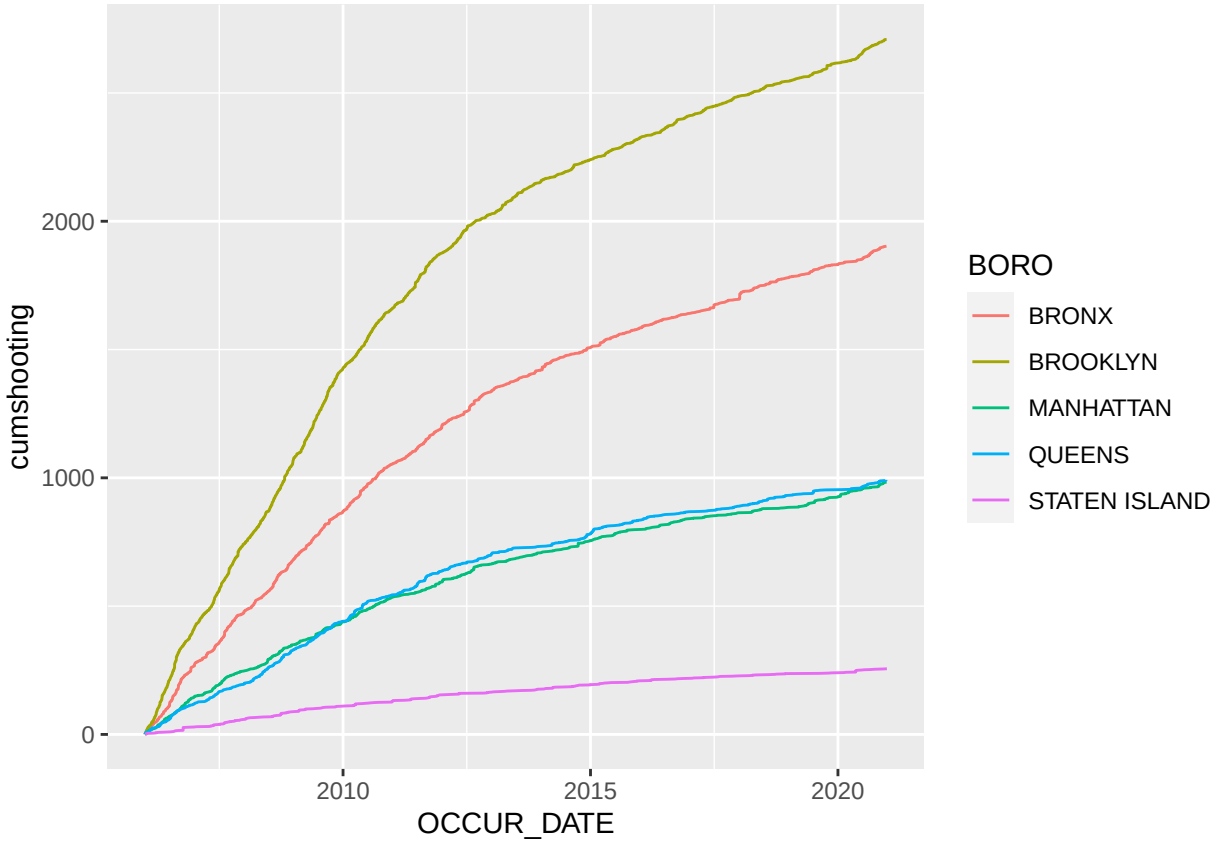
2. Visualization

Here comes three graphs. The first two shows the cumulative number of murder and of shooting events according to the flow of the date by borough. And the rest shows the percentage of murders in total number of shootings by borough.

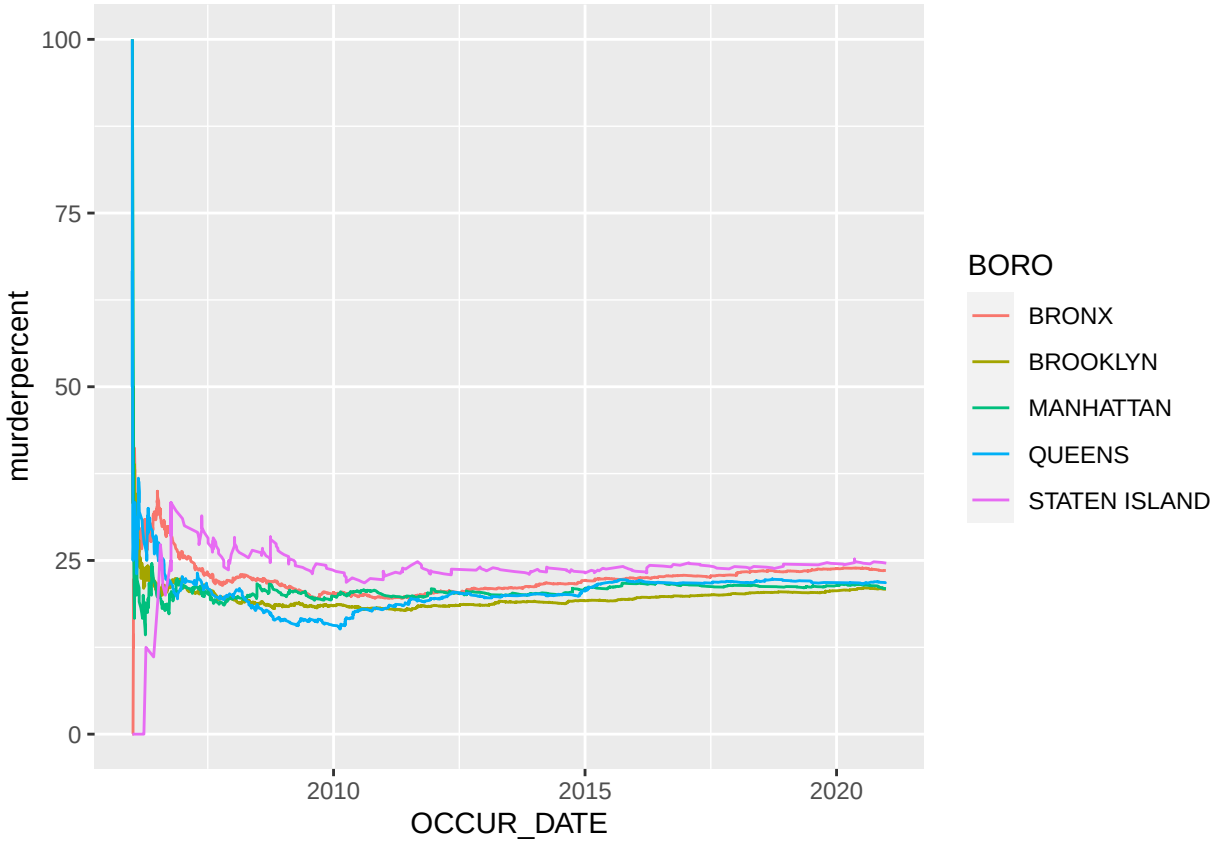
```
#Visualization

nypd_murder_boro_1 %>%
  ggplot(aes(x = OCCUR_DATE, y=cummurder, group=BORO, color=BORO))+
  geom_line()
```




```
nypd_murder_boro_1 %>%
  ggplot(aes(x = OCCUR_DATE, y=cumshooting, group=BORO, color=BORO))+
  geom_line()
```




```
nypd_murder_boro_1 %>%
  ggplot(aes(x = OCCUR_DATE, y=murderpercent, group=BORO, color=BORO))+
  geom_line()
```



3. Analysis

Comparing the total number of shooting cases and murder ones by borough I calculated the percentage of murders in shootings and found that in STATEN ISLAND the total number of shootings is the lowest, but the proportion of deaths from shootings is the highest.

```
# analysis

aggregate(nypd_murder_boro_1$STATISTICAL_MURDER_FLAG, by=list(BORO = nypd_murder_boro_1$BORO), FUN=sum)
```

```
##            BORO   x
## 1         BRONX 448
## 2      BROOKLYN 565
## 3     MANHATTAN 206
## 4        QUEENS 217
## 5 STATEN ISLAND  63
```

```
aggregate(nypd_murder_boro_1$shooting, by=list(BORO = nypd_murder_boro_1$BORO), FUN=sum)
```

```
##            BORO    x
## 1         BRONX 1903
## 2      BROOKLYN 2709
## 3     MANHATTAN  983
## 4        QUEENS  992
## 5 STATEN ISLAND  256
```

```
city <- "BRONX"
nypd_murder_boro_BRONX <- nypd_murder_boro_1 %>%
     filter(BORO == city) %>%
     group_by(BORO, OCCUR_DATE) %>%
     #summarize(STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG) %>%
     select(BORO, OCCUR_DATE, shooting, cumshooting, STATISTICAL_MURDER_FLAG, cummurder, murderpercent)
     ungroup()
knitr::kable(tail(nypd_murder_boro_BRONX))
```

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
BRONX	2020-11-15	1	1898	0	447	23.55111
BRONX	2020-11-26	1	1899	0	447	23.53870
BRONX	2020-12-04	1	1900	0	447	23.52632
BRONX	2020-12-04	1	1901	0	447	23.51394
BRONX	2020-12-14	1	1902	1	448	23.55415
BRONX	2020-12-24	1	1903	0	448	23.54178

```
city <- "BROOKLYN"
nypd_murder_boro_BROOKLYN <- nypd_murder_boro_1 %>%
     filter(BORO == city) %>%
     group_by(BORO, OCCUR_DATE) %>%
     #summarize(STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG) %>%
     select(BORO, OCCUR_DATE, shooting, cumshooting, STATISTICAL_MURDER_FLAG, cummurder, murderpercent)
     ungroup()
knitr::kable(tail(nypd_murder_boro_BROOKLYN))
```

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
BROOKLYN	2020-12-07	1	2704	0	565	20.89497
BROOKLYN	2020-12-07	1	2705	0	565	20.88725
BROOKLYN	2020-12-07	1	2706	0	565	20.87953
BROOKLYN	2020-12-09	1	2707	0	565	20.87181
BROOKLYN	2020-12-11	1	2708	0	565	20.86411
BROOKLYN	2020-12-25	1	2709	0	565	20.85640

```
city <- "STATEN ISLAND"
nypd_murder_boro_STATENISLAND <- nypd_murder_boro_1 %>%
     filter(BORO == city) %>%
     group_by(BORO, OCCUR_DATE) %>%
     #summarize(STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG) %>%
     select(BORO, OCCUR_DATE, shooting, cumshooting, STATISTICAL_MURDER_FLAG, cummurder, murderpercent)
     ungroup()
knitr::kable(tail(nypd_murder_boro_STATENISLAND))
```

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
STATEN ISLAND	2020-06-23	1	251	0	62	24.70120
STATEN ISLAND	2020-07-13	1	252	0	62	24.60317

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
STATEN ISLAND	2020-08-31	1	253	0	62	24.50593
STATEN ISLAND	2020-09-27	1	254	1	63	24.80315
STATEN ISLAND	2020-11-27	1	255	0	63	24.70588
STATEN ISLAND	2020-12-27	1	256	0	63	24.60938

```
city <- "MANHATTAN"
nypd_murder_boro_MANHATTAN <- nypd_murder_boro_1 %>%
     filter(BORO == city) %>%
     group_by(BORO, OCCUR_DATE) %>%
     #summarize(STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG) %>%
     select(BORO, OCCUR_DATE, shooting, cumshooting, STATISTICAL_MURDER_FLAG, cummurder, murderpercent)
     ungroup()
knitr::kable(tail(nypd_murder_boro_MANHATTAN))
```

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
MANHATTAN	2020-11-26	1	978	0	206	21.06339
MANHATTAN	2020-12-03	1	979	0	206	21.04188
MANHATTAN	2020-12-03	1	980	0	206	21.02041
MANHATTAN	2020-12-04	1	981	0	206	20.99898
MANHATTAN	2020-12-09	1	982	0	206	20.97760
MANHATTAN	2020-12-25	1	983	0	206	20.95626

```
city <- "QUEENS"
nypd_murder_boro_QUEENS <- nypd_murder_boro_1 %>%
     filter(BORO == city) %>%
     group_by(BORO, OCCUR_DATE) %>%
     #summarize(STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG) %>%
     select(BORO, OCCUR_DATE, shooting, cumshooting, STATISTICAL_MURDER_FLAG, cummurder, murderpercent)
     ungroup()
knitr::kable(tail(nypd_murder_boro_QUEENS))
```

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
QUEENS	2020-11-05	1	987	0	216	21.88450
QUEENS	2020-11-05	1	988	0	216	21.86235
QUEENS	2020-11-30	1	989	0	216	21.84024
QUEENS	2020-12-20	1	990	0	216	21.81818
QUEENS	2020-12-21	1	991	0	216	21.79617
QUEENS	2020-12-21	1	992	1	217	21.87500

4. Model

I made a linear model to predict the shooting incident of boroughs and compared it to the actual trend of shooting incidents. In graph below, the blue line represents the actual trend of shooting incidents and the

red does the predictive model prediction.

```
# Modeling Data

mod <- lm(cumshooting ~ cummurder, data = nypd_murder_boro_1)
summary(mod)
```

```
##
## Call:
## lm(formula = cumshooting ~ cummurder, data = nypd_murder_boro_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -344.44  -41.04   -6.85   52.37  205.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.086018   1.911421   1.615    0.106
## cummurder   4.951257   0.008042 615.697   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.14 on 6841 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9823
## F-statistic: 3.791e+05 on 1 and 6841 DF,  p-value: < 2.2e-16
```

```
nypd_murder_boro_1 %>% slice_min(cumshooting)
```

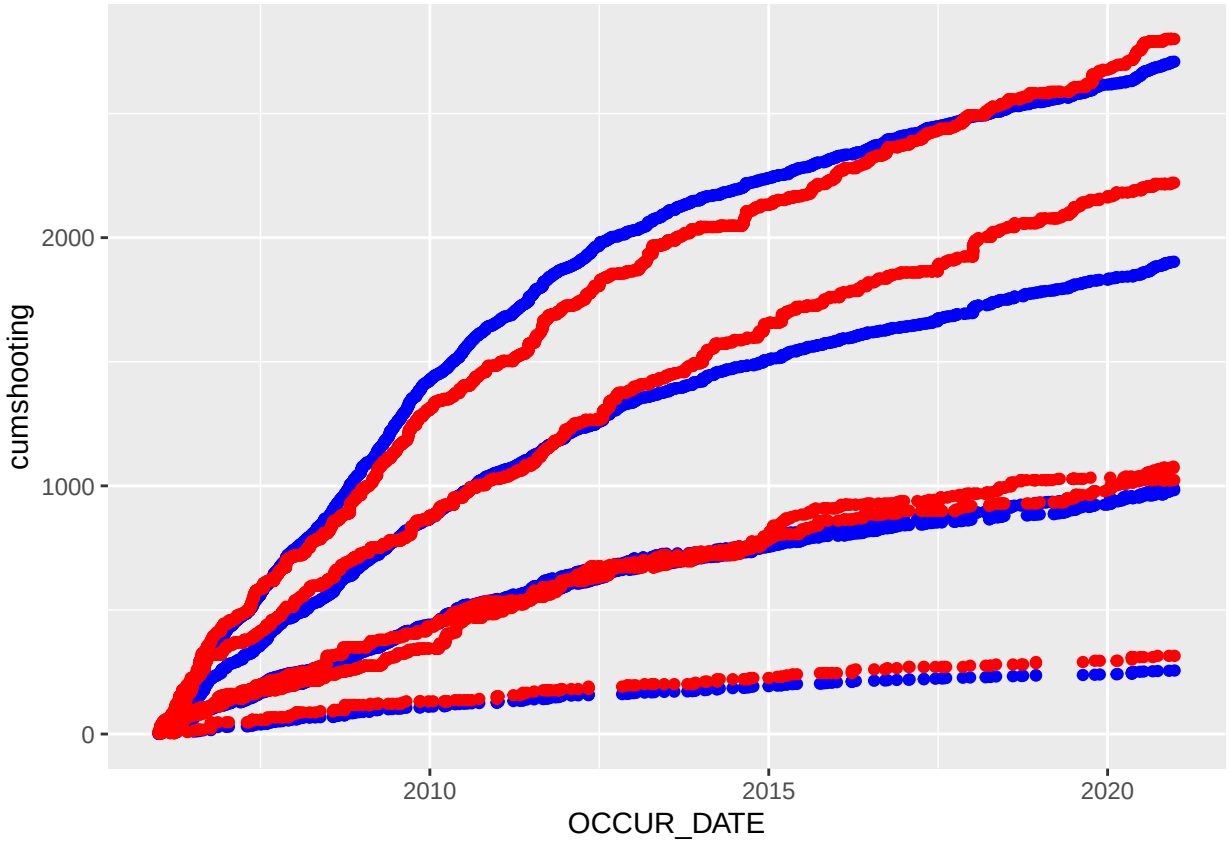
```
## # A tibble: 5 x 7
##   BORO  OCCUR_DATE STATISTICAL_MUR~ cummurder shooting cumshooting murderpercent
##   <chr> <date>                <dbl>     <dbl>    <dbl>       <dbl>         <dbl>
## 1 BRONX 2006-01-01                0         0        1           1             0
## 2 BROO~ 2006-01-02                1         1        1           1           100
## 3 MANH~ 2006-01-01                1         1        1           1           100
## 4 QUEE~ 2006-01-01                1         1        1           1           100
## 5 STAT~ 2006-01-02                0         0        1           1             0
```

```
nypd_murder_boro_1 %>% slice_max(cumshooting)
```

```
## # A tibble: 1 x 7
##   BORO  OCCUR_DATE STATISTICAL_MUR~ cummurder shooting cumshooting murderpercent
##   <chr> <date>                <dbl>     <dbl>    <dbl>       <dbl>         <dbl>
## 1 BROO~ 2020-12-25                0       565        1        2709          20.9
```

```
x_grid <- seq(0, 3000)
new_df <- tibble(cumshooting = x_grid)

nypd_pred <- nypd_murder_boro_1 %>% mutate(pred = predict(mod))

# nypd_pred

nypd_pred %>% ggplot() +
```

```
  geom_point(aes(x = OCCUR_DATE, y=cumshooting), color= "blue")+
  geom_point(aes(x = OCCUR_DATE, y = pred), color = "red")
```



5. Conclusion and Bias Identification

The analysis shows that BROOKLYN is the place where the most shootings occurred in the data. However, STATEN ISLAND has the highest rate of deaths from shooting. If social policy is established based on this data, I think that prevention education for shootings should be approached in a different way in BROOKLYN and STATEN ISLAND. Born in a country in which possession of firearms itself is illegal, I have great fear of shooting itself and distrust of the society in which it is carried out. So, rather than seeing and understanding the data of citizens who can legally own firearms as an important part of society and thinking about countermeasures, the fact that there are many gunshots is just bad. However, while doing science, it has changed that I try to accept the phenomenon itself and recognize the value of research. Not making judgments about certain facts but trying to accept and understand the phenomenon in a neutral way is a way for me to step into a bigger world, and for our society to make the rules fairer.