NYC Shooting Project

SSH

2023-02-19

Peer graded assignment NYPD shooting incident

The objective of this assignment is to test the capability of the student in applying the data science concepts taught and provide effective outcomes.

Step 1: Import data from the server

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/"
file_names <- c("rows.csv")
urls <- str_c(url_in,file_names)
```

```
NYPD_Shooting <- read_csv(urls)
```

Step2: Tidying the data:
Identified the suitable fields for the analysis and removed the unwanted fields.
Changed the date fields in accordance with the R suited format.

```
##Removing the least relevant fields
library(tidyr)
library(tidyverse)
library(lubridate)
```

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

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

```
library(dplyr)
library(ggplot2)
NYPD_Shooting <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLO
```

```
## Rows: 25596 Columns: 19
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
knitr::kable(head(NYPD_Shooting))
```

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE	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
236168668	11/11/2021	15:04:00	BROOKLYN	79	0	NA	FALSE	NA	NA	NA	18-24	M	BLACK	996313	187499	40.68132	-73.95651	POINT (-73.95650899099996 40.68131820000008)
231008085	07/16/2021	22:05:00	BROOKLYN	72	0	NA	FALSE	45-64	M	ASIAN / PACIFIC ISLANDER	25-44	M	ASIAN / PACIFIC ISLANDER	981845	171118	40.63636	-74.00867	POINT (-74.00866668999998 40.63636384100005)
230717903	07/11/2021	01:09:00	BROOKLYN	79	0	NA	FALSE	<18	M	BLACK	25-44	M	BLACK	996546	187436	40.68115	-73.95567	POINT (-73.95566903799994 40.68114495900005)
237712309	12/11/2021	13:42:00	BROOKLYN	81	0	NA	FALSE	NA	NA	NA	25-44	M	BLACK	1001139	192775	40.69579	-73.93910	POINT (-73.939095905 40.69579171600003)
224465521	02/16/2021	20:00:00	QUEENS	113	0	NA	FALSE	NA	NA	NA	25-44	M	BLACK	1050711	184826	40.67374	-73.76041	POINT (-73.76041066999993 40.67374017600008)
228252164	05/15/2021	04:30:00	QUEENS	113	0	NA	TRUE	NA	NA	NA	25-44	M	BLACK	1051324	196640	40.70618	-73.75806	POINT (-73.75806147399999 40.70617856900003)

```
nypd_cleansed <- drop_na(NYPD_Shooting) %>% select(-c(INCIDENT_KEY, LOCATION_DESC, X_COORD_CD, Y_COORD
##Changing the date to the convenience
nypd_cleansed <- nypd_cleansed %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
##Converting the boolean values to integers
```

```
nypd_cleansed$STATISTICAL_MURDER_FLAG[nypd_cleansed$STATISTICAL_MURDER_FLAG == "TRUE"] <- 1
nypd_cleansed$STATISTICAL_MURDER_FLAG[nypd_cleansed$STATISTICAL_MURDER_FLAG == "FALSE"] <- 0
nypd_boro <- nypd_cleansed  %>% group_by(BORO, OCCUR_DATE) %>% summarize(STATISTICAL_MURDER_FLAG = STA
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.

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

```
nypd_boro$cummurder <- ave(nypd_boro$STATISTICAL_MURDER_FLAG,nypd_boro$BORO,FUN=cumsum)
nypd_boro['shooting']=1
nypd_boro$cumshooting <- ave(nypd_boro$shooting,nypd_boro$BORO, FUN = cumsum)
nypd_boro$murderpercent <- with(nypd_boro, cummurder/cumshooting *100)
#show the final data for the anlaysis
knitr::kable(head(nypd_boro))
```

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

Step 3: Data Analysis
:
Aggregated the required measures based on the suitable dimensions such as date,BORO

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

```
##            BORO   x
## 1         BRONX 502
## 2      BROOKLYN 607
## 3     MANHATTAN 234
## 4        QUEENS 235
## 5 STATEN ISLAND  70
```

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

```
##            BORO    x
## 1         BRONX 2019
## 2      BROOKLYN 2840
## 3     MANHATTAN 1062
## 4        QUEENS 1055
## 5 STATEN ISLAND  267
```

```
city <- "BRONX"
nypd_murder_boro_BRONX <- nypd_boro %>%
  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	2021-12-02	1	2014	1	499	24.77656
BRONX	2021-12-03	1	2015	0	499	24.76427
BRONX	2021-12-03	1	2016	0	499	24.75198
BRONX	2021-12-11	1	2017	1	500	24.78929
BRONX	2021-12-11	1	2018	1	501	24.82656
BRONX	2021-12-11	1	2019	1	502	24.86379

```
city <- "BROOKLYN"
nypd_murder_boro_BROOKLYN <- nypd_boro %>%
  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	2021-12-14	1	2835	1	604	21.30511
BROOKLYN	2021-12-17	1	2836	0	604	21.29760
BROOKLYN	2021-12-17	1	2837	0	604	21.29010
BROOKLYN	2021-12-17	1	2838	1	605	21.31783
BROOKLYN	2021-12-17	1	2839	1	606	21.34554
BROOKLYN	2021-12-18	1	2840	1	607	21.37324

```
city <- "STATEN ISLAND"
nypd_murder_boro_STATENISLAND <- nypd_boro %>%
  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	2021-04-18	1	262	0	66	25.19084
STATEN ISLAND	2021-04-28	1	263	1	67	25.47529

BORO	OCCUR_DATE	shooting	cumshooting	STATISTICAL_MURDER_FLAG	cummurder	murderpercent
STATEN ISLAND	2021-06-22	1	264	1	68	25.75758
STATEN ISLAND	2021-07-30	1	265	0	68	25.66038
STATEN ISLAND	2021-11-21	1	266	1	69	25.93985
STATEN ISLAND	2021-12-31	1	267	1	70	26.21723

```
city <- "MANHATTAN"
nypd_murder_boro_MANHATTAN <- nypd_boro %>%
  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	2021-11-15	1	1057	0	231	21.85430
MANHATTAN	2021-11-17	1	1058	1	232	21.92817
MANHATTAN	2021-11-20	1	1059	1	233	22.00189
MANHATTAN	2021-12-03	1	1060	0	233	21.98113
MANHATTAN	2021-12-16	1	1061	1	234	22.05467
MANHATTAN	2021-12-20	1	1062	0	234	22.03390

```
city <- "QUEENS"
nypd_murder_boro_QUEENS <- nypd_boro %>%
  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	2021-10-10	1	1050	1	232	22.09524
QUEENS	2021-11-02	1	1051	1	233	22.16936
QUEENS	2021-12-06	1	1052	1	234	22.24335
QUEENS	2021-12-06	1	1053	0	234	22.22222
QUEENS	2021-12-11	1	1054	1	235	22.29602
QUEENS	2021-12-19	1	1055	0	235	22.27488

Step 4: Applying Linear model on the data and Visualization

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

```
##
## Call:
## lm(formula = cumshooting ~ cummurder, data = nypd_boro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415.67  -51.23  -11.36   57.02  230.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.468517   2.053160   8.021 1.21e-15 ***
## cummurder    4.817127   0.008074 596.608  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 104.5 on 7241 degrees of freedom
## Multiple R-squared:  0.9801, Adjusted R-squared:  0.9801
## F-statistic: 3.559e+05 on 1 and 7241 DF,  p-value: < 2.2e-16
```

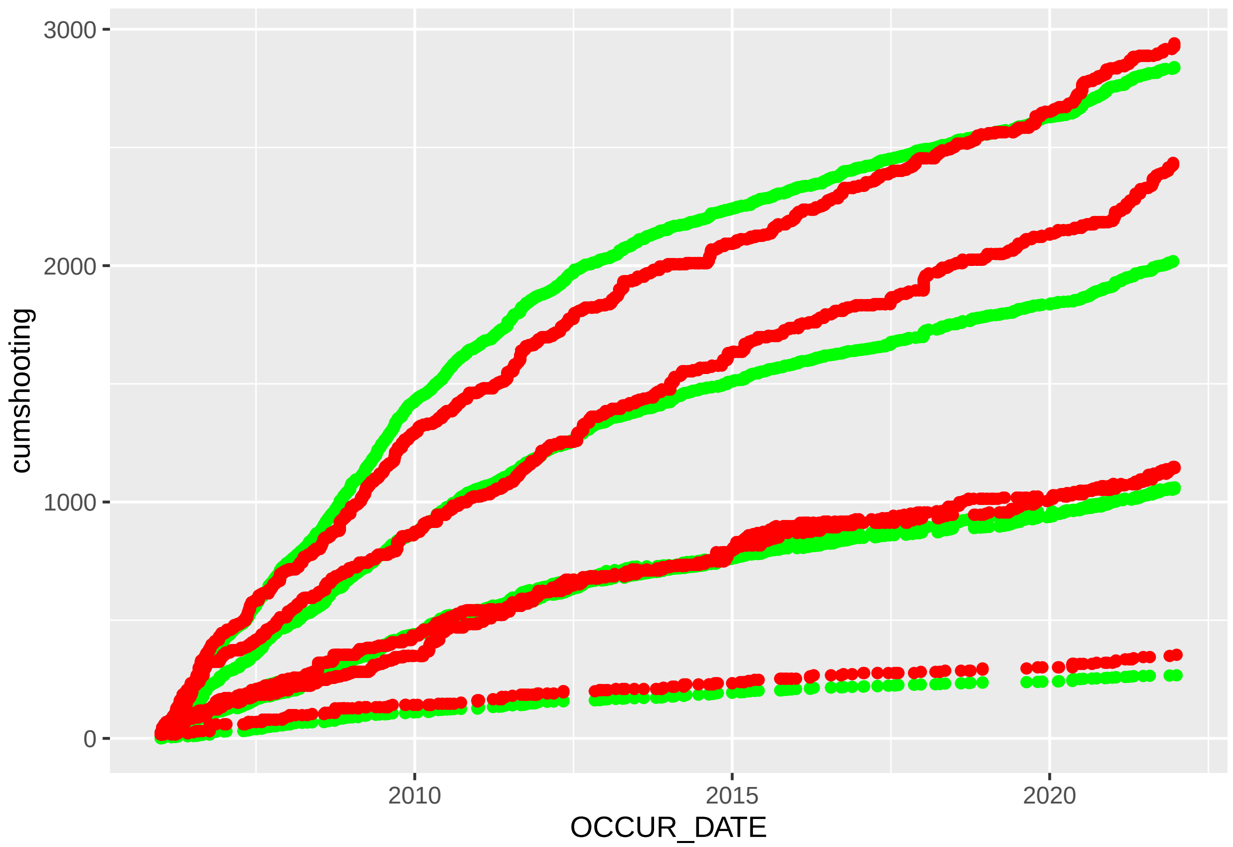
```
nypd_boro %>% slice_min(cumshooting)
```

```
## # A tibble: 5 x 7
##   BORO          OCCUR_DATE STATISTICAL_MURDER_~1 cummu~2 shoot~3 cumsh~4 murde~5
##   <chr>         <date>                     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 BRONX         2006-01-01                     0       0       1       1       0
## 2 BROOKLYN      2006-01-02                     1       1       1       1     100
## 3 MANHATTAN     2006-01-01                     1       1       1       1     100
## 4 QUEENS        2006-01-01                     0       0       1       1       0
## 5 STATEN ISLAND 2006-01-02                     0       0       1       1       0
## # ... with abbreviated variable names 1: STATISTICAL_MURDER_FLAG, 2: cummurder,
## #   3: shooting, 4: cumshooting, 5: murderpercent
```

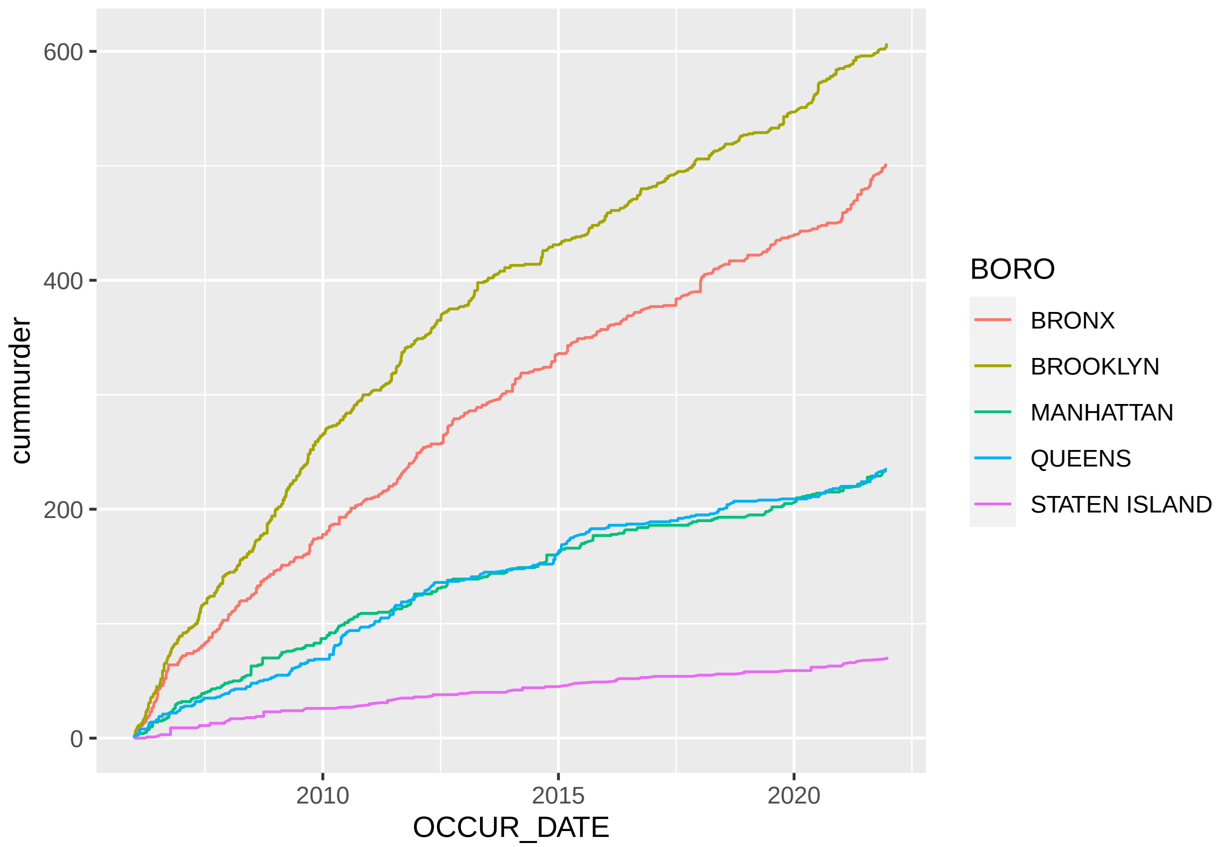
```
nypd_boro %>% slice_max(cumshooting)
```

```
## # A tibble: 1 x 7
##   BORO     OCCUR_DATE STATISTICAL_MURDER_FLAG cummurder shooting cumsh~1 murde~2
##   <chr>    <date>                       <dbl>     <dbl>    <dbl>   <dbl>   <dbl>
## 1 BROOKLYN 2021-12-18                       1       607        1    2840    21.4
## # ... with abbreviated variable names 1: cumshooting, 2: murderpercent
```

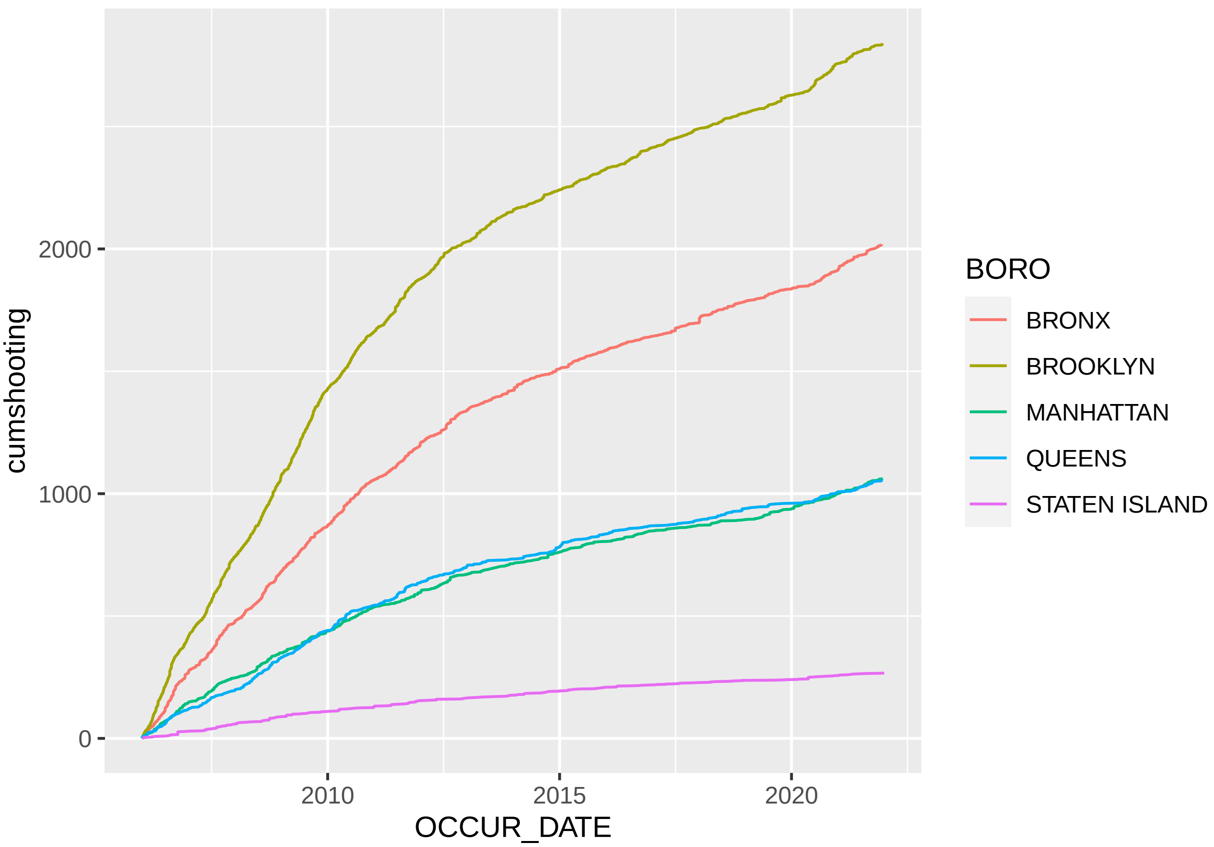
```
x_grid <- seq(0, 3000)
new_df <- tibble(cumshooting = x_grid)
nypd_pred <- nypd_boro %>% mutate(pred = predict(mod))
# nypd_pred
nypd_pred %>% ggplot() +
  geom_point(aes(x = OCCUR_DATE, y=cumshooting), color= "green")+
  geom_point(aes(x = OCCUR_DATE, y = pred), color = "red")
```



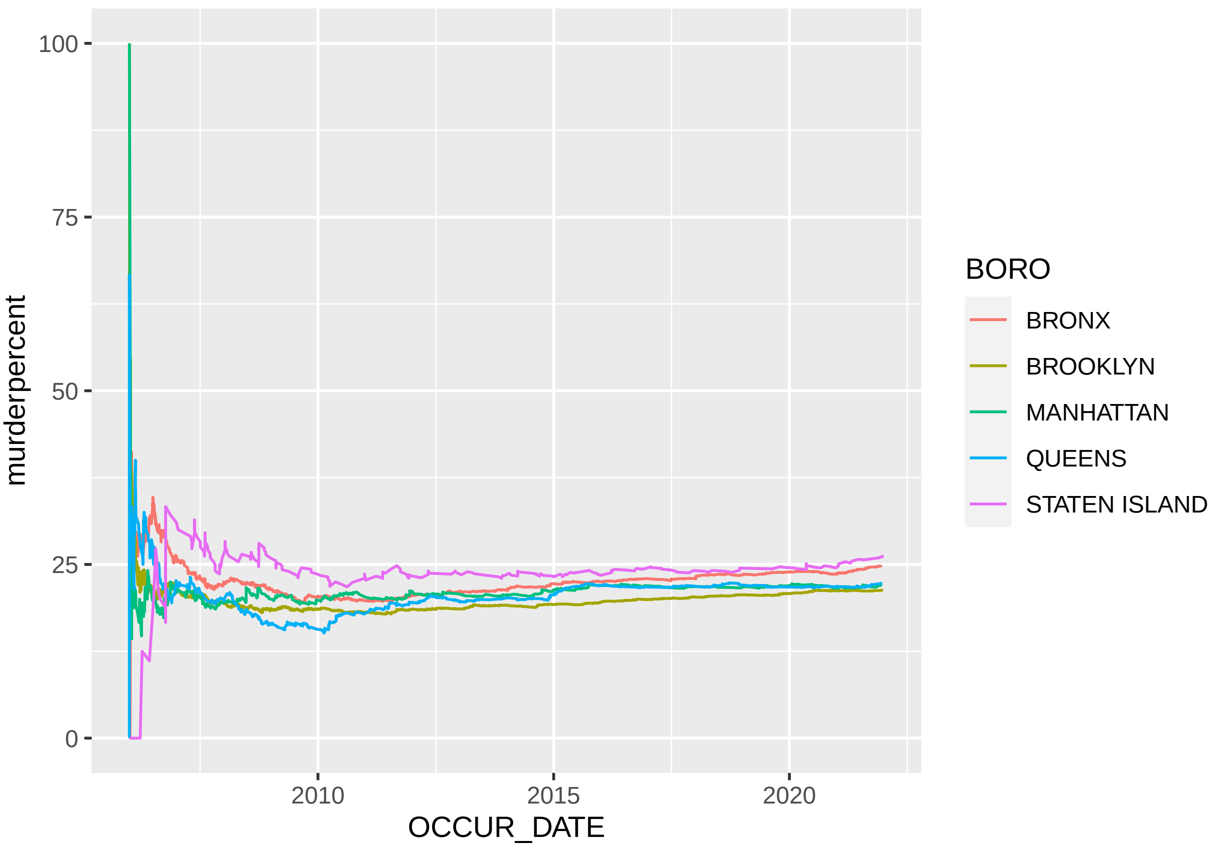
```
#Visualization of data

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



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



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



Step 5: Adding Bias Identification - Being foreigner I couldn't imagine the incident in the way it happened and that's a potential Bias.I have great fear of shooting and disbelief of the society in which it is carried out.Had to do lot of studies to understand the incident and this could lead me to the way it was portrayed.