

An In Depth Look at Shooting Incidents in NYC

Steven Lee

2023-10-14

Per the data.gov website, this data is a list of every shooting incident in NYC from 2006 to the previous calendar year. The Office of Management Analysis and Planning manually extract and review this data before posting. Every record is represented by each shooting incident and its related information.

Library Used

```
library(tidyverse)
library(lubridate)
```

Import Data

Import NYPD Shooting Incident data by the given URL as a .csv file.

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shootings_data_raw <- read_csv(url, show_col_types = FALSE)
head(shootings_data_raw)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##   <dbl> <chr>      <time> <chr>      <chr>              <dbl>
## 1  228798151 05/27/2021 21:30    QUEENS    <NA>              105
## 2  137471050 06/27/2014 17:40    BRONX     <NA>              40
## 3  147998800 11/21/2015 03:56    QUEENS    <NA>              108
## 4  146837977 10/09/2015 18:30    BRONX     <NA>              44
## 5    58921844 02/19/2009 22:58    BRONX     <NA>              47
## 6   219559682 10/21/2020 21:36    BROOKLYN <NA>              81
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, 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>
```

```
summary(shootings_data_raw)
```

```
##   INCIDENT_KEY      OCCUR_DATE      OCCUR_TIME      BORO
##   Min.       : 9953245 Length:27312 Length:27312 Length:27312
##   1st Qu.: 63860880  Class :character Class1:hms   Class :character
```

```

## Median : 90372218   Mode :character   Class2:difftime   Mode :character
## Mean :120860536     Mode :numeric
## 3rd Qu.:188810230
## Max. :261190187
##
## LOC_OF_OCCUR_DESC   PRECINCT   JURISDICTION_CODE LOC_CLASSFCTN_DESC
## Length:27312       Min. : 1.00   Min. :0.0000   Length:27312
## Class :character   1st Qu.: 44.00 1st Qu.:0.0000   Class :character
## Mode :character   Median : 68.00 Median :0.0000   Mode :character
##                   Mean : 65.64 Mean :0.3269
##                   3rd Qu.: 81.00 3rd Qu.:0.0000
##                   Max. :123.00 Max. :2.0000
##                   NA's :2
## LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## Length:27312       Mode :logical   Length:27312
## Class :character   FALSE:22046     Class :character
## Mode :character   TRUE :5266      Mode :character
##
##
##
## PERP_SEX           PERP_RACE           VIC_AGE_GROUP           VIC_SEX
## Length:27312       Length:27312     Length:27312     Length:27312
## Class :character   Class :character Class :character   Class :character
## Mode :character   Mode :character Mode :character   Mode :character
##
##
##
## VIC_RACE           X_COORD_CD           Y_COORD_CD           Latitude
## Length:27312       Min. : 914928   Min. :125757   Min. :40.51
## Class :character   1st Qu.:1000029 1st Qu.:182834 1st Qu.:40.67
## Mode :character   Median :1007731 Median :194487 Median :40.70
##                   Mean :1009449 Mean :208127 Mean :40.74
##                   3rd Qu.:1016838 3rd Qu.:239518 3rd Qu.:40.82
##                   Max. :1066815 Max. :271128 Max. :40.91
##                   NA's :10
## Longitude         Lon_Lat
## Min. : -74.25     Length:27312
## 1st Qu.: -73.94   Class :character
## Median : -73.92   Mode :character
## Mean : -73.91
## 3rd Qu.: -73.88
## Max. : -73.70
## NA's :10

```

Data Clean Up

I clean the raw data to include standardized occurrence dates along with the New York location where the incident occurred. Select only the necessary columns and other missing data that are not related with this report are removed. The resulting data is stored as 'shootings-data'.

```
df_shootings <- shootings_data_raw %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, PERP_AGE_GROUP)
df_shootings$OCCUR_DATE <- mdy(df_shootings$OCCUR_DATE)

summary(df_shootings)
```

```
## INCIDENT_KEY      OCCUR_DATE      OCCUR_TIME      BORO
## Min.   : 9953245   Min.   :2006-01-01   Length:17958   Length:17958
## 1st Qu.: 49834899  1st Qu.:2008-08-05   Class1:hms     Class :character
## Median : 81778268  Median :2011-11-17   Class2:difftime Mode  :character
## Mean   :112574247  Mean   :2013-05-10   Mode :numeric
## 3rd Qu.:178508294  3rd Qu.:2018-04-22
## Max.   :261190187  Max.   :2022-12-31

## PERP_AGE_GROUP    PERP_SEX      PERP_RACE      VIC_AGE_GROUP
## Length:17958      Length:17958      Length:17958   Length:17958
## Class :character   Class :character   Class :character Class :character
## Mode  :character   Mode  :character   Mode  :character Mode  :character
##
##
##
## VIC_SEX          VIC_RACE          Latitude      Longitude
## Length:17958      Length:17958      Min.   :40.52  Min.   : -74.23
## Class :character   Class :character   1st Qu.:40.67  1st Qu.: -73.94
## Mode  :character   Mode  :character   Median :40.71  Median : -73.91
##
##                   Mean   :40.74  Mean   : -73.91
##                   3rd Qu.:40.83  3rd Qu.: -73.88
##                   Max.   :40.91  Max.   : -73.71
```

Transforming

Because much of perpetrator data remain unidentifiable, I want to designate them as “Unknown” and change their data type to “factor” along with other categorical data types. I also want to rename some variables to have a cleaner visualizations.

```
df_shootings <- df_shootings %>% mutate(VIC_RACE = case_when(VIC_RACE == "American Indian/Alaskan Native" ~ "Unknown",
  df_shootings = df_shootings %>%
    replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))

df_shootings$PERP_AGE_GROUP = recode(df_shootings$PERP_AGE_GROUP, UNKNOWN = "Unknown")
df_shootings$PERP_SEX = recode(df_shootings$PERP_SEX, U = "Unknown")
df_shootings$PERP_RACE = recode(df_shootings$PERP_RACE, UNKNOWN = "Unknown")
df_shootings$VIC_SEX = recode(df_shootings$VIC_SEX, U = "Unknown")
df_shootings$VIC_RACE = recode(df_shootings$VIC_RACE, UNKNOWN = "Unknown")
df_shootings$INCIDENT_KEY = as.character(df_shootings$INCIDENT_KEY)
df_shootings$BORO = as.factor(df_shootings$BORO)
df_shootings$PERP_AGE_GROUP = as.factor(df_shootings$PERP_AGE_GROUP)
df_shootings$PERP_SEX = as.factor(df_shootings$PERP_SEX)
df_shootings$PERP_RACE = as.factor(df_shootings$PERP_RACE)
df_shootings$VIC_AGE_GROUP = as.factor(df_shootings$VIC_AGE_GROUP)
df_shootings$VIC_SEX = as.factor(df_shootings$VIC_SEX)
df_shootings$VIC_RACE = as.factor(df_shootings$VIC_RACE)
```

```
summary(df_shootings)
```

```
## INCIDENT_KEY      OCCUR_DATE      OCCUR_TIME      BORO
## Length:17958      Min.   :2006-01-01      Length:17958      BRONX      :5425
## Class :character   1st Qu.:2008-08-05      Class1:hms        BROOKLYN    :6638
## Mode  :character   Median :2011-11-17      Class2:difftime   MANHATTAN   :2538
##                                     Mean   :2013-05-10      Mode :numeric     QUEENS      :2726
##                                     3rd Qu.:2018-04-22                                     STATEN ISLAND: 631
##                                     Max.   :2022-12-31
##
## PERP_AGE_GROUP     PERP_SEX      PERP_RACE      VIC_AGE_GROUP
## 18-24 :6219      (null) : 635    BLACK          :11429    <18      :2027
## 25-44 :5686      F      : 424    WHITE HISPANIC: 2339    1022     : 1
## Unknown:3148     M      :15434   Unknown        : 1802    18-24    :6513
## <18      :1590    Unknown: 1465   BLACK HISPANIC: 1314    25-44    :7934
## (null) : 635                                     (null)      : 635    45-64    :1290
## 45-64 : 617                                     WHITE        : 283    65+      : 137
## (Other): 63                                     (Other)      : 156    UNKNOWN: 56
## VIC_SEX      VIC_RACE      Latitude
## F      : 1920    AMERICAN INDIAN/ALASKAN NATIVE: 8    Min.   :40.52
## M      :16030    ASIAN / PACIFIC ISLANDER      : 307    1st Qu.:40.67
## Unknown: 8      BLACK          :12246    Median :40.71
##                                     BLACK HISPANIC : 1798    Mean   :40.74
##                                     Unknown        : 48     3rd Qu.:40.83
##                                     WHITE          : 552    Max.   :40.91
##                                     WHITE HISPANIC : 2999
## Longitude
## Min.   :-74.23
## 1st Qu.: -73.94
## Median : -73.91
## Mean   : -73.91
## 3rd Qu.: -73.88
## Max.   : -73.71
##
```

I think it's interesting to look at a cumulative graph of all incidents in all of NY as well as each of the Boroughs over time to see if we can observe any trends in this aspect.

```
ts_df_shootings <- df_shootings %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>% ungroup()
summary(ts_df_shootings)
```

```
## OCCUR_DATE      COUNT
## Min.   :2006-01-01      Min.   : 1.000
## 1st Qu.:2009-08-14      1st Qu.: 1.000
## Median :2013-11-01      Median : 3.000
## Mean   :2014-02-26      Mean   : 3.582
## 3rd Qu.:2018-08-18      3rd Qu.: 5.000
## Max.   :2022-12-31      Max.   :25.000
```

```

bronx_n <- df_shootings %>% filter(BORO == "BRONX") %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>%
brooklyn_n <- df_shootings %>% filter(BORO == "BROOKLYN") %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>%
manhattan_n <- df_shootings %>% filter(BORO == "MANHATTAN") %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>%
queens_n <- df_shootings %>% filter(BORO == "QUEENS") %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>%
staten_n <- df_shootings %>% filter(BORO == "STATEN ISLAND") %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>%

```

Visualizations

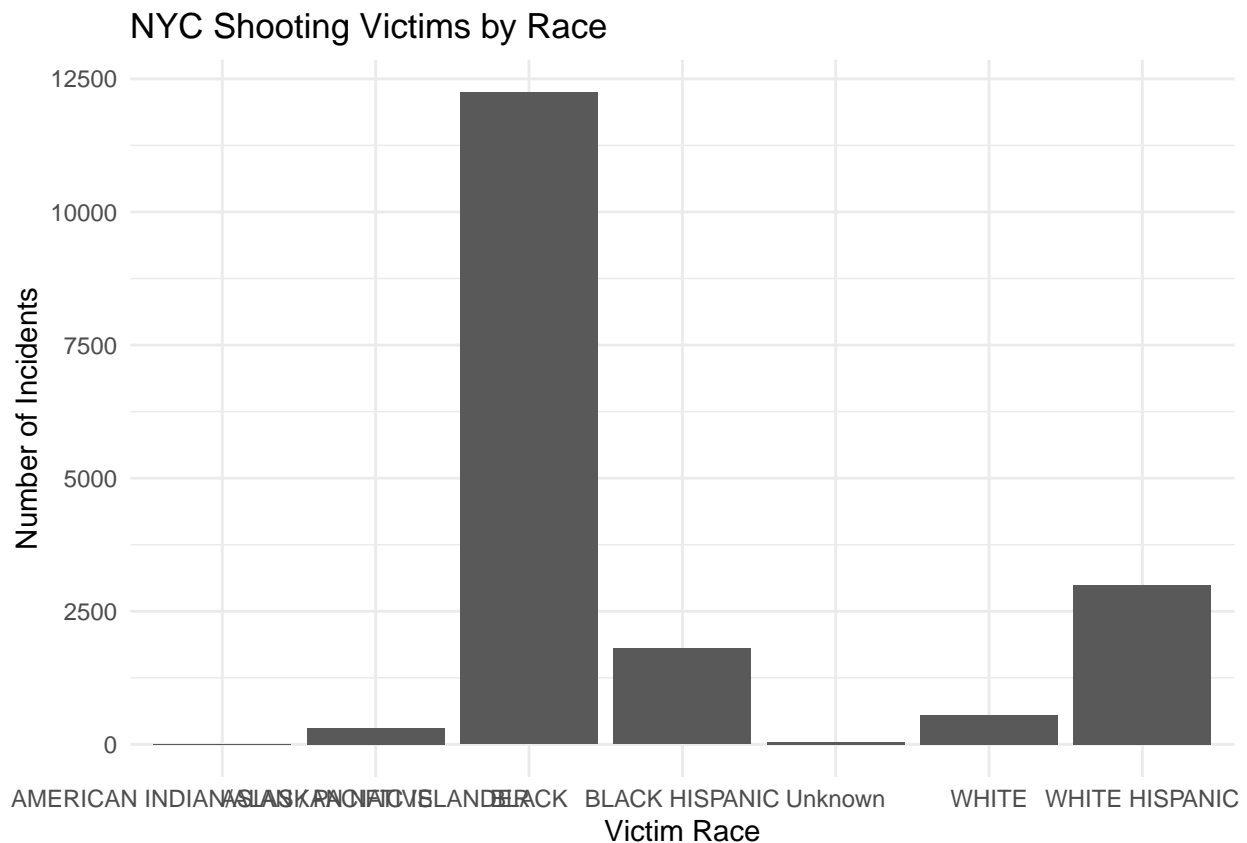
Now that all the data is organized, we can make some visualizations for questions we want answered.

1. Which victim race is involved in the most incidents?

```

g <- ggplot(df_shootings, aes(x = VIC_RACE)) +
  geom_bar() +
  labs(title = "NYC Shooting Victims by Race",
       x = "Victim Race",
       y = "Number of Incidents") +
  theme_minimal()
g

```

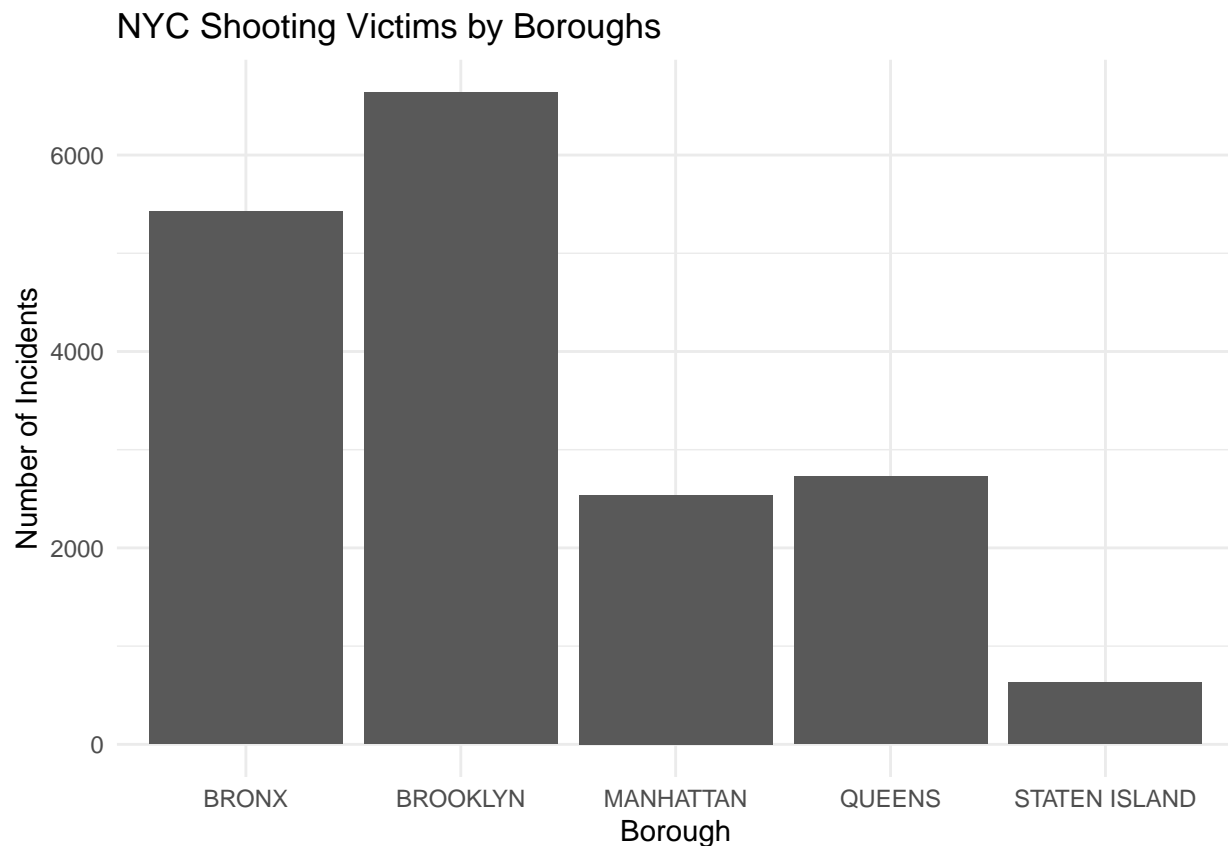


We can see that shooting incidents most often involve blacks.

2. Which borough is involved in the most incidents?

```
g <- ggplot(df_shootings, aes(x = BORO)) +
  geom_bar() +
  labs(title = "NYC Shooting Victims by Boroughs",
        x = "Borough",
        y = "Number of Incidents") +
  theme_minimal()
```

g

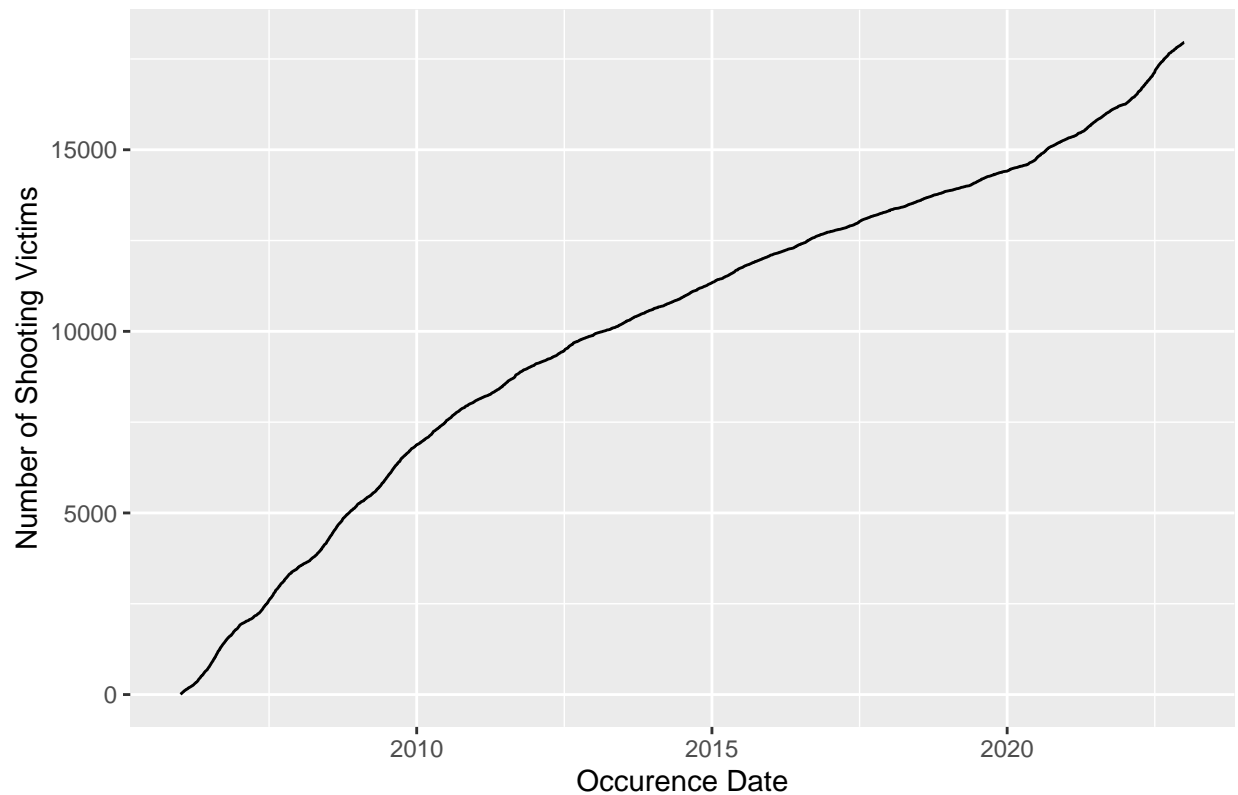


We can see that Brooklyn is the Borough with the most shooting incidents in this time period.

3. How has shooting incidents changed over time in NYC?

```
ggplot() +
  geom_line(data=ts_df_shootings, aes(x=OCCUR_DATE, y=cumsum(COUNT))) +
  labs(title = "Cumulative Shooting Victims in New York") +
  labs(y="Number of Shooting Victims", x="Occurence Date")
```

Cumulative Shooting Victims in New York

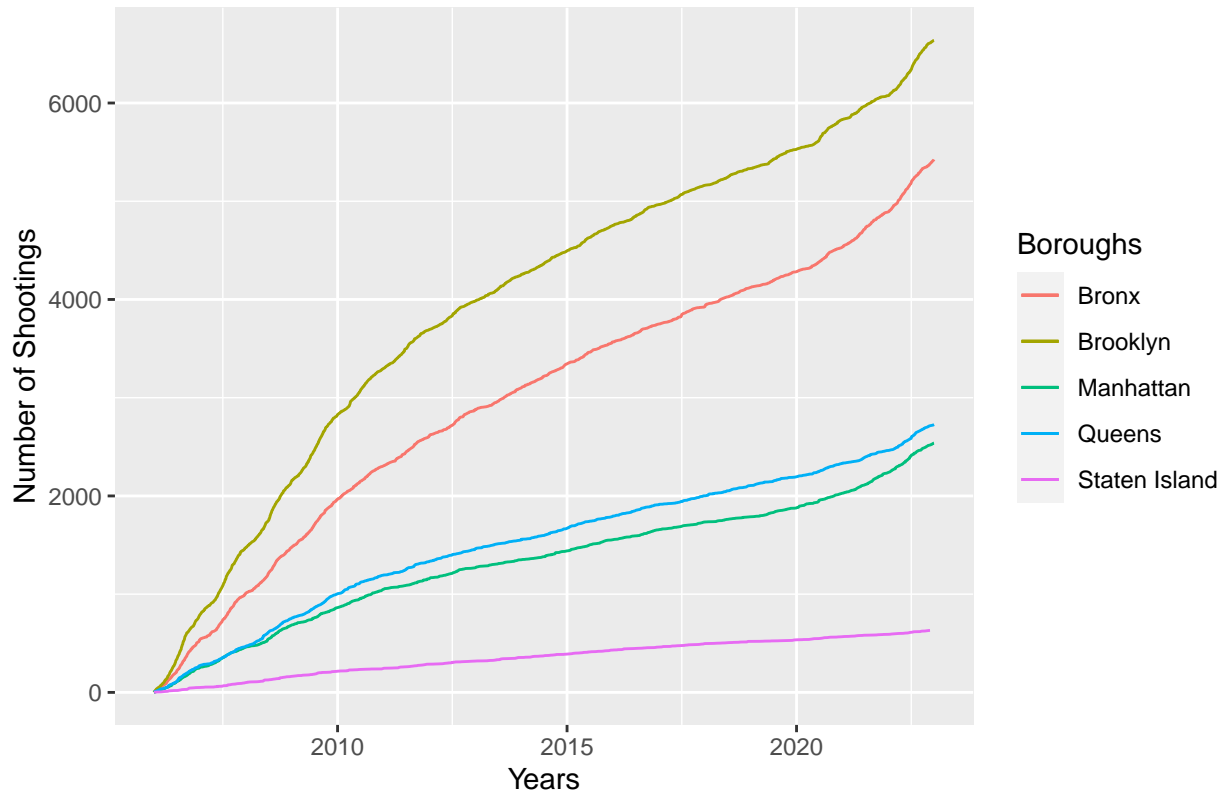


We can see that the slope of the line changes at around 2011-2012 and then again at 2020. Let's see if all the Boroughs reflex this change in the same manner.

4. How has shooting incidents changed over time in each of the Boroughs?

```
ggplot() +
  geom_line(data=bronx_n, aes(x=OCCUR_DATE, y=cumsum(COUNT), color='Bronx')) +
  geom_line(data=brooklyn_n, aes(x=OCCUR_DATE, y=cumsum(COUNT), color='Brooklyn')) +
  geom_line(data=manhattan_n, aes(x=OCCUR_DATE, y=cumsum(COUNT), color='Manhattan')) +
  geom_line(data=queens_n, aes(x=OCCUR_DATE, y=cumsum(COUNT), color='Queens')) +
  geom_line(data=staten_n, aes(x=OCCUR_DATE, y=cumsum(COUNT), color='Staten Island')) +
  labs(title = "Shooting Incident Count in Each Borough") +
  labs(y="Number of Shootings", x="Years", color="Boroughs")
```

Shooting Incident Count in Each Borough



We can see that all the Boroughs follow a similar trend with the overall picture except Staten Island.

Modeling

Because Staten Island is the unique Borough, let's make a linear regression model in a scatter plot to explore if there is any correlation between the highest shooting incident Borough of Brooklyn and the lowest in Staten Island. It's easier to see if we transform the incidents into yearly totals for both Boroughs first.

Transforming

```
yearly_data <- ts_df_shootings
yearly_data$OCCUR_DATE <- yearly_data$OCCUR_DATE %>% year()
yearly_data <- yearly_data %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n())

brooklyn_yearly <- brooklyn_n
brooklyn_yearly$OCCUR_DATE <- brooklyn_yearly$OCCUR_DATE %>% year()
brooklyn_yearly <- brooklyn_yearly %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n())

staten_yearly <- staten_n
staten_yearly$OCCUR_DATE <- staten_yearly$OCCUR_DATE %>% year()
staten_yearly <- staten_yearly %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n())

summary(brooklyn_yearly)
```



```
##      OCCUR_DATE      COUNT
## Min.      :2006    Min.      :109.0
## 1st Qu.:2010    1st Qu.:124.0
## Median :2014    Median :154.0
## Mean      :2014    Mean      :181.3
## 3rd Qu.:2018    3rd Qu.:237.0
## Max.      :2022    Max.      :289.0
```

```
summary(staten_yearly)
```

```
##      OCCUR_DATE      COUNT
## Min.      :2006    Min.      :14.00
## 1st Qu.:2010    1st Qu.:22.00
## Median :2014    Median :27.00
## Mean      :2014    Mean      :26.76
## 3rd Qu.:2018    3rd Qu.:31.00
## Max.      :2022    Max.      :42.00
```

Prediction Model

```
merge_data <- merge(brooklyn_yearly[-1,],staten_yearly[-1,], by="OCCUR_DATE")
```

```
mod <- lm(COUNT.y ~ COUNT.x, data = merge_data)
```

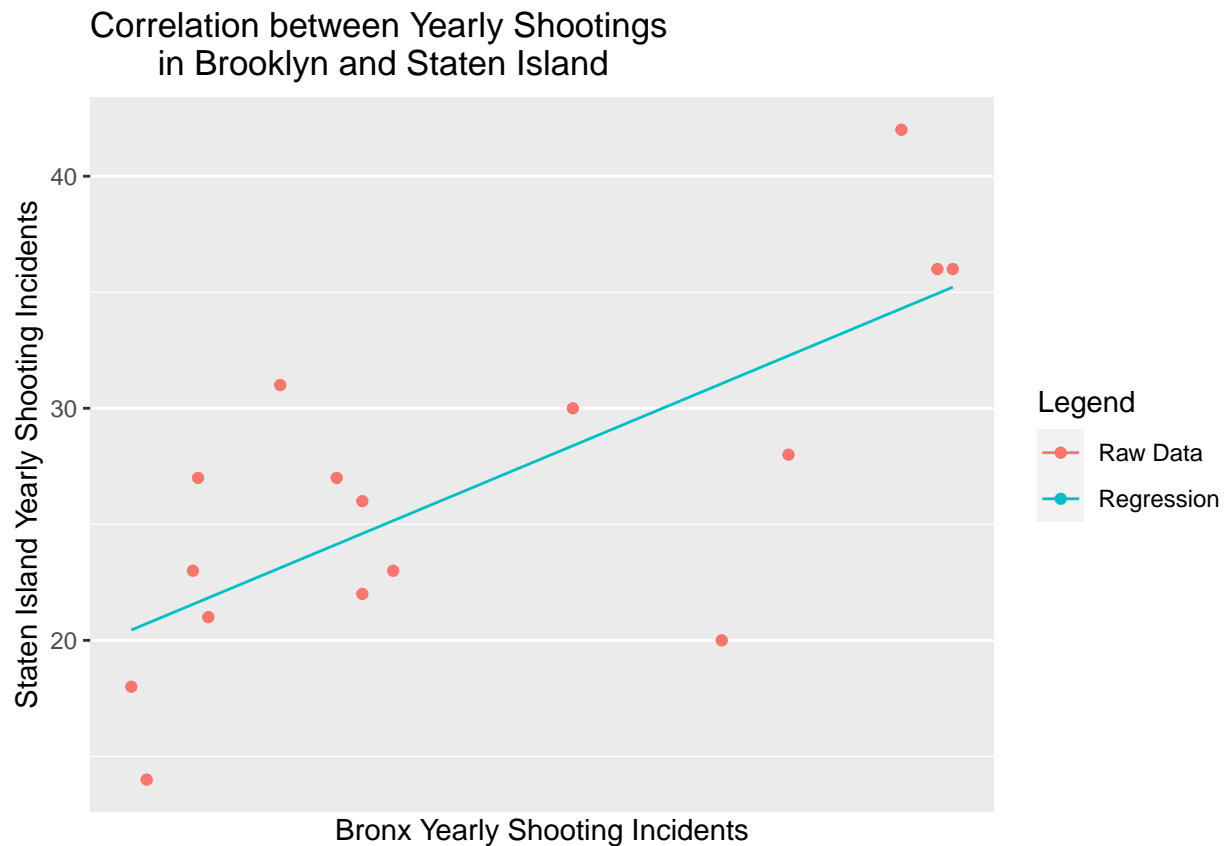
```
summary(mod)
```

```
##
## Call:
## lm(formula = COUNT.y ~ COUNT.x, data = merge_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0645  -2.4854   0.9192   1.9248   7.8758
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.38283    4.19352   2.476  0.02668 *
## COUNT.x       0.09233    0.02287   4.038  0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.144 on 14 degrees of freedom
## Multiple R-squared:  0.538, Adjusted R-squared:  0.505
## F-statistic: 16.3 on 1 and 14 DF, p-value: 0.001222
```

```
regs <- merge_data %>% mutate(reg = predict(mod))
```

```
regs %>% ggplot() + geom_point(aes(x=COUNT.x,y=COUNT.y,color = "Raw Data")) +
  geom_line(aes(x=COUNT.x,y=reg, color = "Regression")) +
  scale_x_continuous(breaks = pretty(yearly_data$OCCUR_DATE, n = 1)) +
  labs(title = "Correlation between Yearly Shootings
```

```
in Brooklyn and Staten Island") +
labs(y="Staten Island Yearly Shooting Incidents",
x="Bronx Yearly Shooting Incidents", color="Legend")
```



From the chart we can see that the correlation between Brooklyn and Staten Island is generally that both sets are increasing over time, though

Bias

Bias could stem from inconsistent reporting in certain areas, which may lead to data gaps in some places and an overabundance in others. The data also fail to take into account socioeconomic status and other outside factors like weather or the COVID 19 pandemic. More info on these variables may needed to enhance the conclusions that can be made in this analysis.