NYPD shooting report data visualization and analysis

Kevin Juandi

2024-03-16

Installing Libraries

```
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("ggplot2")
#install.packages("repr")
```

Loading Libraries

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse, warn.conflicts = FALSE)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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 errors
```

```
library(lubridate, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
library(repr, warn.conflicts = FALSE)
options(repr.plot.width=10, repr.plot.height=8)
options(dplyr.summarise.inform = FALSE)
```

Loading Data

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

EDA

Let's start with simple EDA

```
head(NYPD, 10)
```

```
##    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
## 7      85295722 06/17/2012   22:47:00   QUEENS                        114
## 8      71662474 03/08/2010   19:41:00 BROOKLYN                         81
## 9      83002139 02/05/2012   05:45:00   QUEENS                        105
## 10     86437261 08/26/2012   01:10:00   QUEENS                        101
##    JURISDICTION_CODE LOC_CLASSFCTN_DESC           LOCATION_DESC
## 1                  0
## 2                  0
## 3                  0
## 4                  0
## 5                  0
## 6                  0
## 7                  0
## 8                  0
## 9                  0
## 10                 0                     MULTI DWELL - APT BUILD
##    STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                    false                                          18-24
## 2                    false                                          18-24
## 3                     true                                          25-44
## 4                    false                                            <18
## 5                     true          25-44        M     BLACK         45-64
## 6                     true                                          25-44
## 7                    false                                          25-44
## 8                     true                                          18-24
## 9                    false                                          25-44
## 10                   false          25-44        M     BLACK         25-44
##    VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1        M          BLACK    1058925   180924.0 40.66296 -73.73084
## 2        M          BLACK    1005028   234516.0 40.81035 -73.92494
## 3        M          WHITE    1007668   209836.5 40.74261 -73.91549
## 4        M WHITE HISPANIC    1006537   244511.1 40.83778 -73.91946
## 5        M          BLACK    1024922   262189.4 40.88624 -73.85291
## 6        M          BLACK    1004234   186461.7 40.67846 -73.92795
## 7        M          BLACK     998860   214885.0 40.75648 -73.94727
## 8        M          BLACK    1002883   192219.7 40.69426 -73.93281
## 9        M          BLACK    1054366   196628.4 40.70611 -73.74711
## 10       M          BLACK    1053937   157130.4 40.59770 -73.74906
##                                      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)
## 7   POINT (-73.94726649399996 40.75648234300007)
```

```
## 8  POINT (-73.93280863699994 40.694264056000065)
## 9  POINT (-73.74710653899996 40.706106731000034)
## 10  POINT (-73.74906464199995 40.59769719800005)
```

There seems to be a lot of blanks

```
sapply(NYPD, function(x) sum(is.na(x)))
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO        LOC_OF_OCCUR_DESC                  PRECINCT
##                       0                        0                        0
##       JURISDICTION_CODE       LOC_CLASSFCTN_DESC            LOCATION_DESC
##                       2                        0                        0
## STATISTICAL_MURDER_FLAG           PERP_AGE_GROUP                 PERP_SEX
##                       0                        0                        0
##                PERP_RACE            VIC_AGE_GROUP                  VIC_SEX
##                       0                        0                        0
##                VIC_RACE               X_COORD_CD               Y_COORD_CD
##                       0                        0                        0
##                Latitude                Longitude                  Lon_Lat
##                      10                       10                        0
```

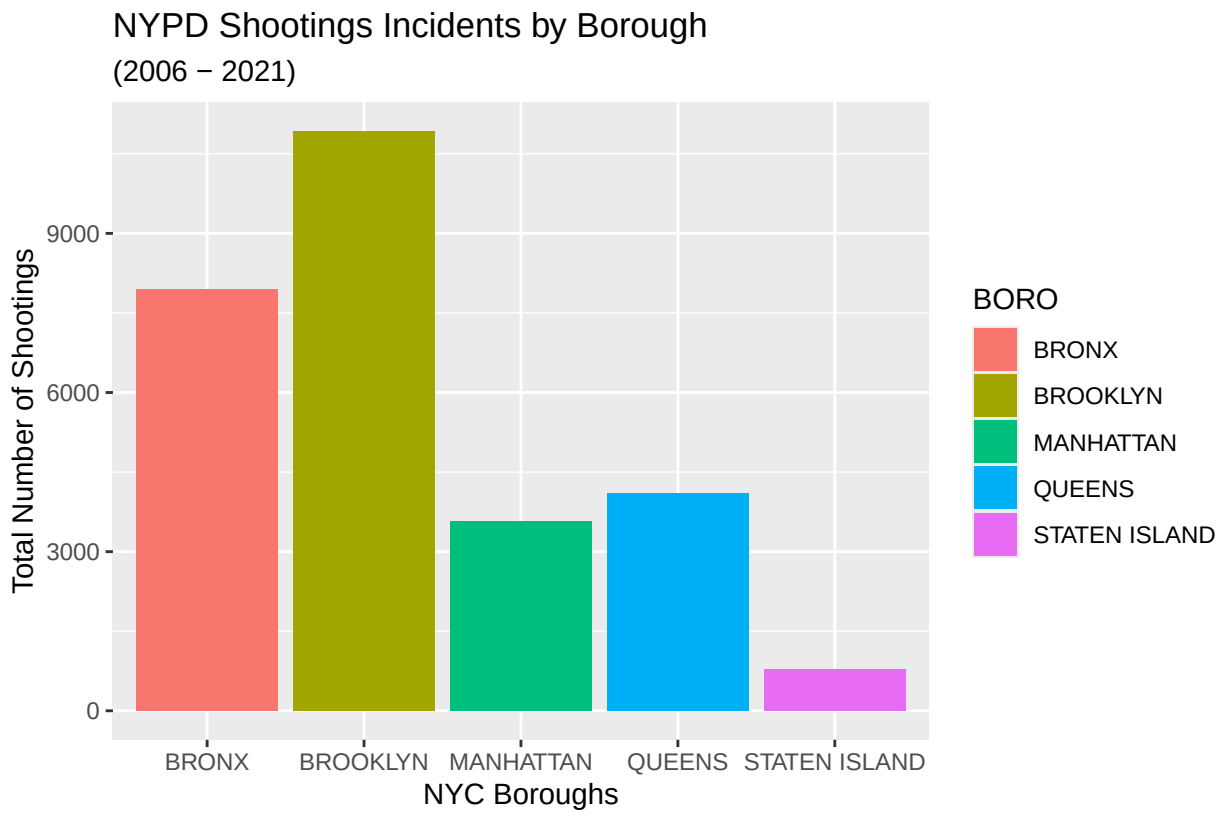
```
summary(NYPD)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  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       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        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
```

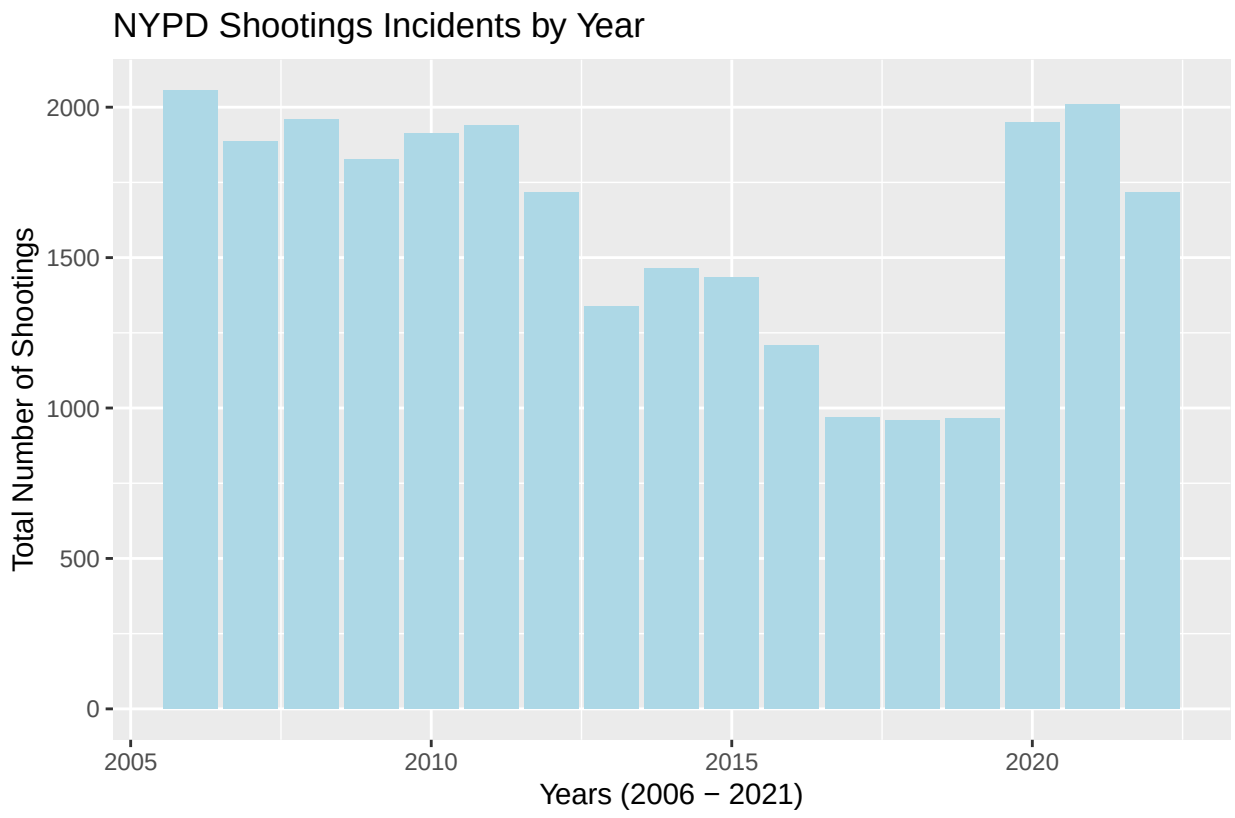
Graph Plots

```
NYPD_clean <- NYPD %>%
    select(c("OCCUR_DATE","OCCUR_TIME","BORO","PRECINCT",
             "STATISTICAL_MURDER_FLAG","PERP_RACE","VIC_AGE_GROUP","VIC_SEX","VIC_RACE")) %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE),
           OCCUR_TIME = hms(OCCUR_TIME),
           STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
           Shootings = 1,
           Year = year(OCCUR_DATE))

NYPD_clean %>%
    ggplot(aes(x = BORO, fill = BORO)) +
    geom_bar() +
    labs(title = "NYPD Shootings Incidents by Borough",
         subtitle = "(2006 - 2021)",
         x = "NYC Boroughs",
         y = "Total Number of Shootings",
         caption = "(Figure - 1)")
```



(Figure − 1)

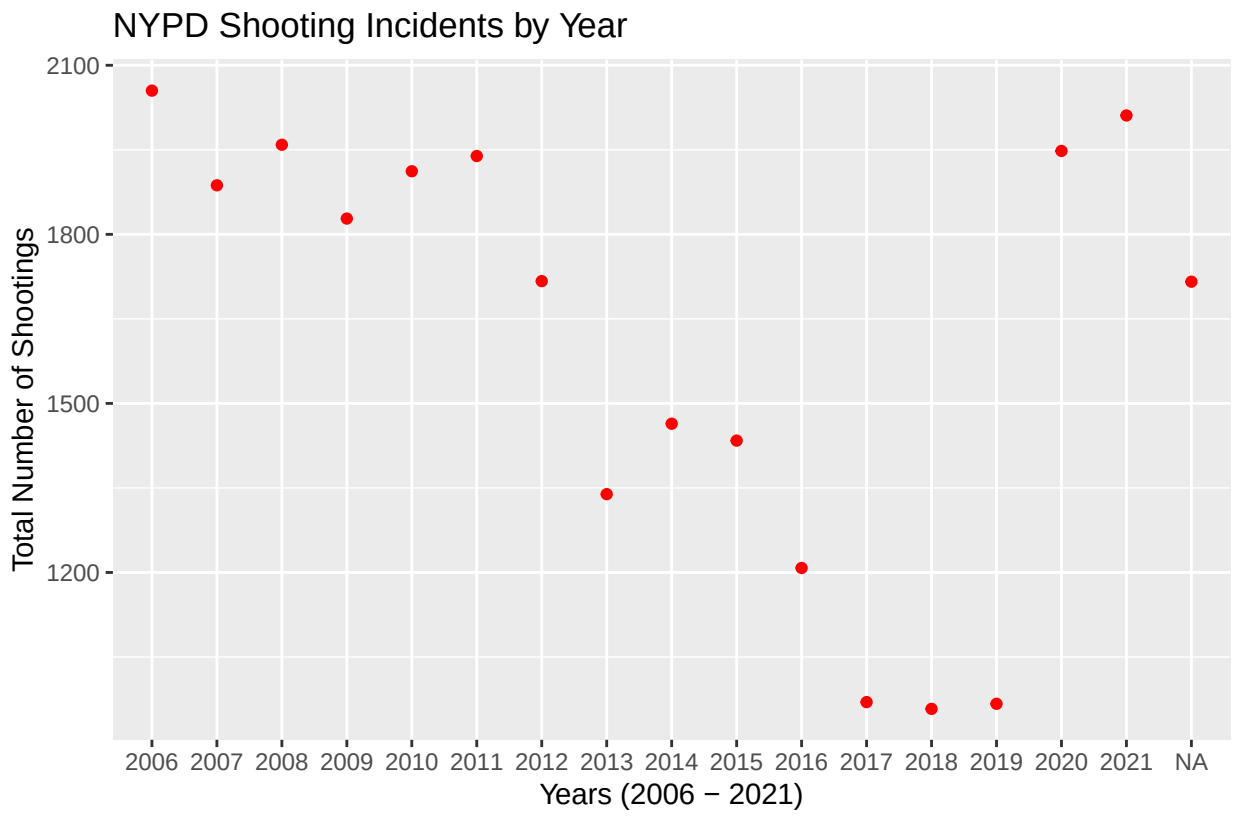
```
NYPD_clean %>%
    ggplot(aes(x = Year)) +
    geom_bar(fill = "lightblue",show.legend = FALSE) +
    labs(title = "NYPD Shootings Incidents by Year",
         x = "Years (2006 - 2021)",
         y = "Total Number of Shootings",
         caption = "(Figure - 2)")
```



(Figure − 2)

```
NYPD_year <- NYPD_clean %>%
    group_by(Year) %>%
    summarize(Shootings = sum(Shootings))

NYPD_year %>%
    ggplot(aes(x = as.factor(Year), y = Shootings)) +
    geom_line() +
    geom_point(color = "red") +
    scale_x_discrete(labels = as.character(2006:2021)) +
    labs(
        title = "NYPD Shooting Incidents by Year",
        x = "Years (2006 - 2021)",
        y = "Total Number of Shootings",
        caption = "(Figure - 3)"
    )
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



(Figure − 3)

looks like some entries are missing dates

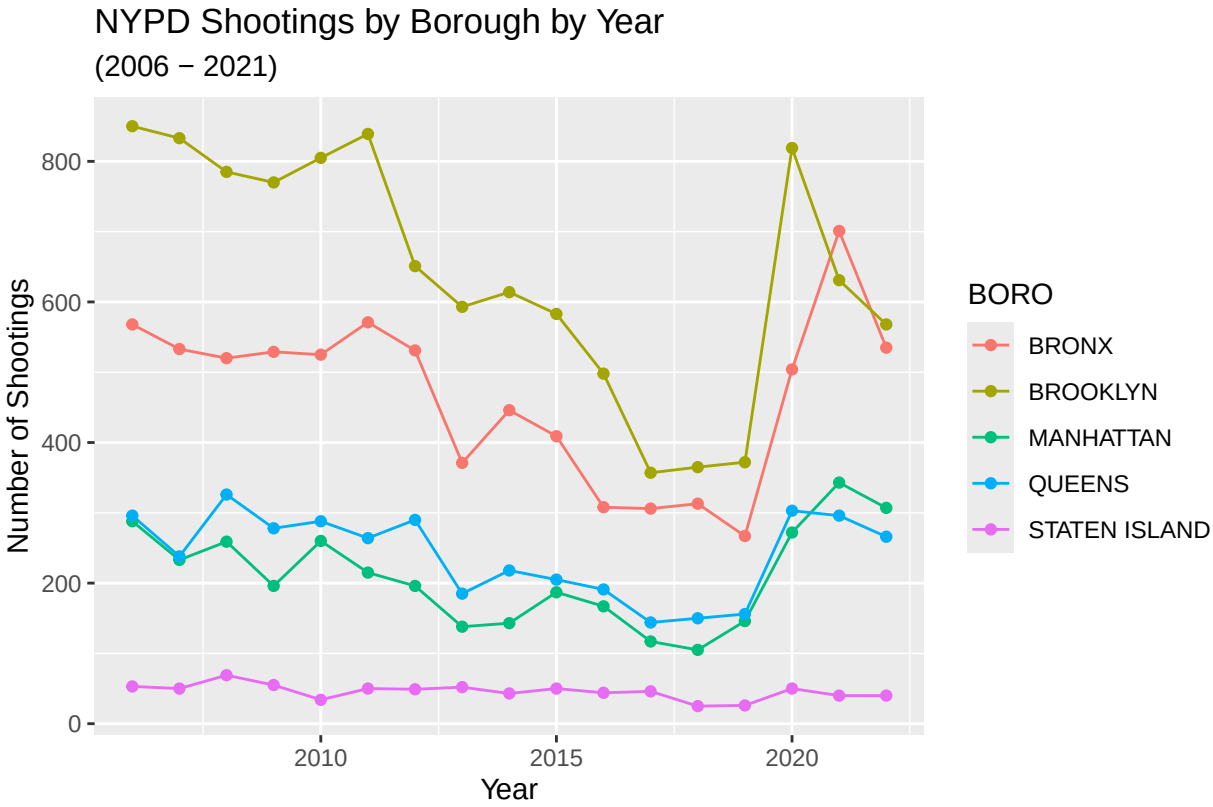
```
NYPD_boro <- NYPD_clean %>%
    group_by(BORO, OCCUR_DATE,Shootings) %>%
    summarize(Shootings = sum(Shootings),
              STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG),
              .groups = 'drop') %>%
    select(BORO,OCCUR_DATE,Shootings,STATISTICAL_MURDER_FLAG) %>%
    ungroup()

NYPD_boro_year <- NYPD_clean %>%
    mutate(Year = year(OCCUR_DATE)) %>%
    group_by(BORO, Year, Shootings) %>%
    summarize(Shootings = sum(Shootings),
              STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG),
              .groups = 'drop') %>%
    select(BORO, Year,Shootings,STATISTICAL_MURDER_FLAG) %>%
    ungroup()

NYPD_boro_total <- NYPD_boro_year %>%
    group_by(BORO) %>%
    summarize(Shootings = sum(Shootings))
(7402 + 10365)/ sum(NYPD_boro_total$Shootings)
```

```
## [1] 0.6505199
```

```
736/ sum(NYPD_boro_total$Shootings)
```

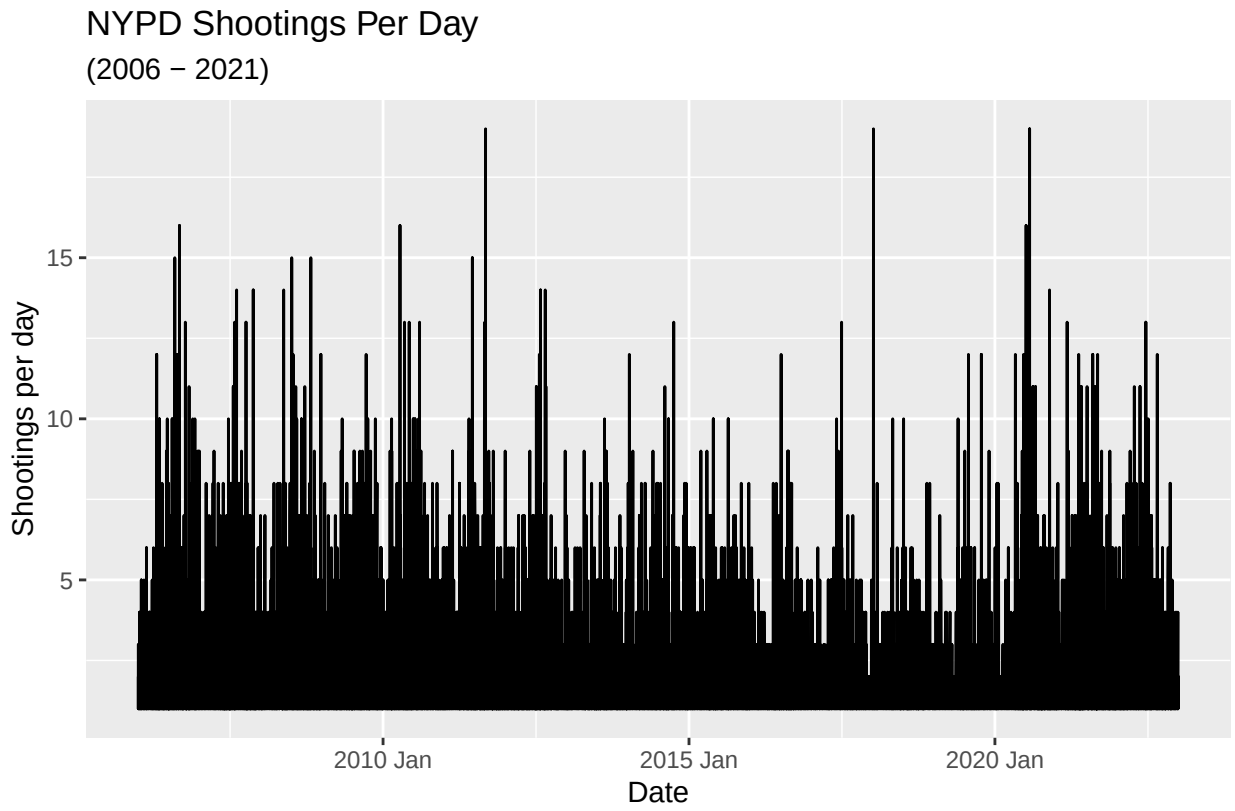
```
## [1] 0.02694786
```

```
NYPD_boro_year %>%
    ggplot(aes(x = Year, y = Shootings,color = BORO)) +
    geom_line() +
    geom_point() +
    labs(title = "NYPD Shootings by Borough by Year",
         subtitle = "(2006 - 2021)",
         x = "Year",
         y = "Number of Shootings",
         caption = "(Figure - 4)")
```



It seems like number of population does not reflect the number of incident. Bronx and Manhattan has the same magnitude of population, so does Brooklyn and Queens. I would assume that this is because of the presence of gangs and organized crime syndicates which Bronx and Brooklyn has the reputation for. It also worth to note that Bronx is the poorest borough.

```
NYPD_boro %>%
    ggplot(aes(x = OCCUR_DATE, y = Shootings)) +
    geom_line() + scale_x_date(date_labels = "%Y %b") +
    labs(title = "NYPD Shootings Per Day",
         subtitle = "(2006 - 2021)",
         x = "Date",
         y = "Shootings per day",
         caption = "(Figure - 5)")
```



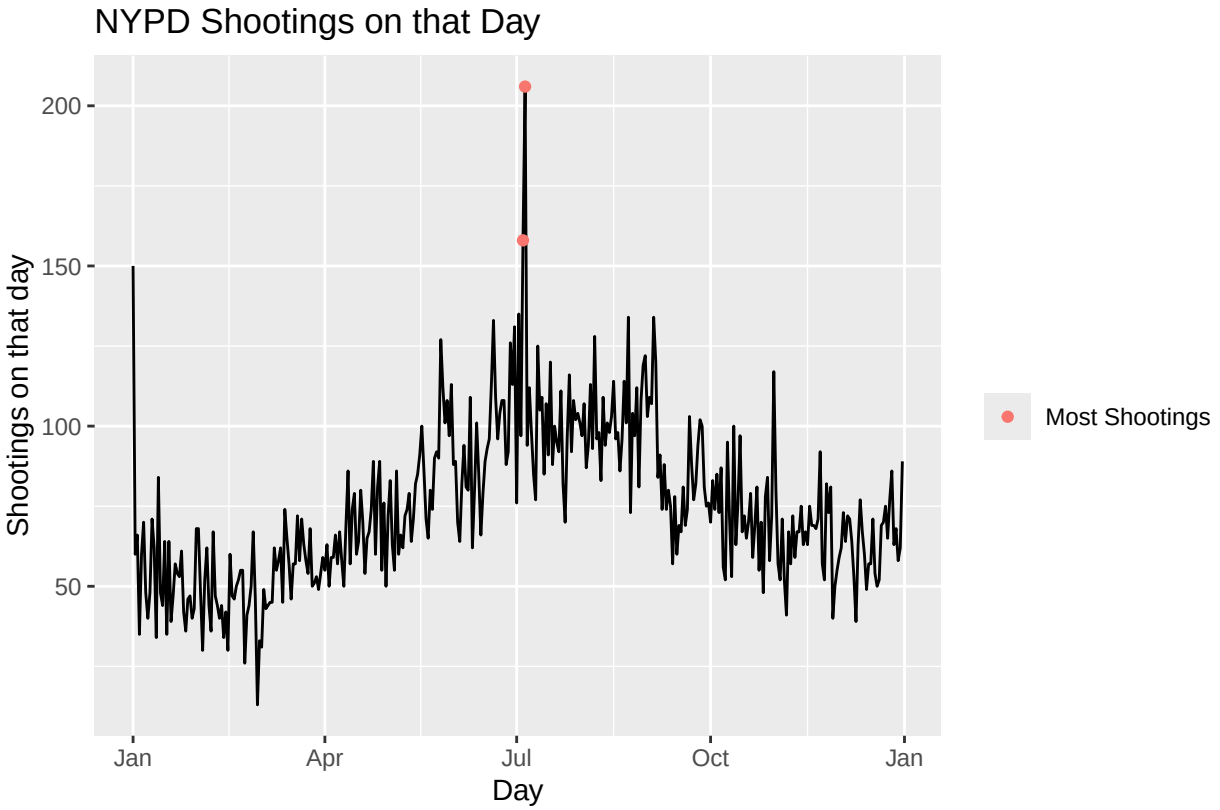
(Figure − 5)

```
NYPD_time_year <- NYPD_clean %>%
    mutate(Time_year = format(as.Date(OCCUR_DATE), "%m/%d")) %>%
    mutate(Time_year = as.Date(Time_year,"%m/%d")) %>%
    group_by(Time_year,Shootings) %>%
    summarize(Shootings = sum(Shootings),
              STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG),
              .groups = 'drop') %>%
    select(Time_year,Shootings,STATISTICAL_MURDER_FLAG) %>%
    ungroup()

NYPD_time_year %>% slice_max(Shootings, n = 2)
```

```
## # A tibble: 2 x 3
##   Time_year  Shootings STATISTICAL_MURDER_FLAG
##   <date>         <dbl>                  <int>
## 1 2024-07-05       206                     33
## 2 2024-07-04       158                     26
```

```
NYPD_July_5 <- NYPD_clean %>%
    mutate(Time_year = format(as.Date(OCCUR_DATE), "%m/%d"),
           Hour = hour(OCCUR_TIME)) %>%
    mutate(Time_year = as.Date(Time_year,"%m/%d")) %>%
    filter(Time_year == "2022-07-05") %>%
    group_by(Hour,Shootings) %>%
    summarize(Shootings = sum(Shootings),
              .groups = 'drop')
```

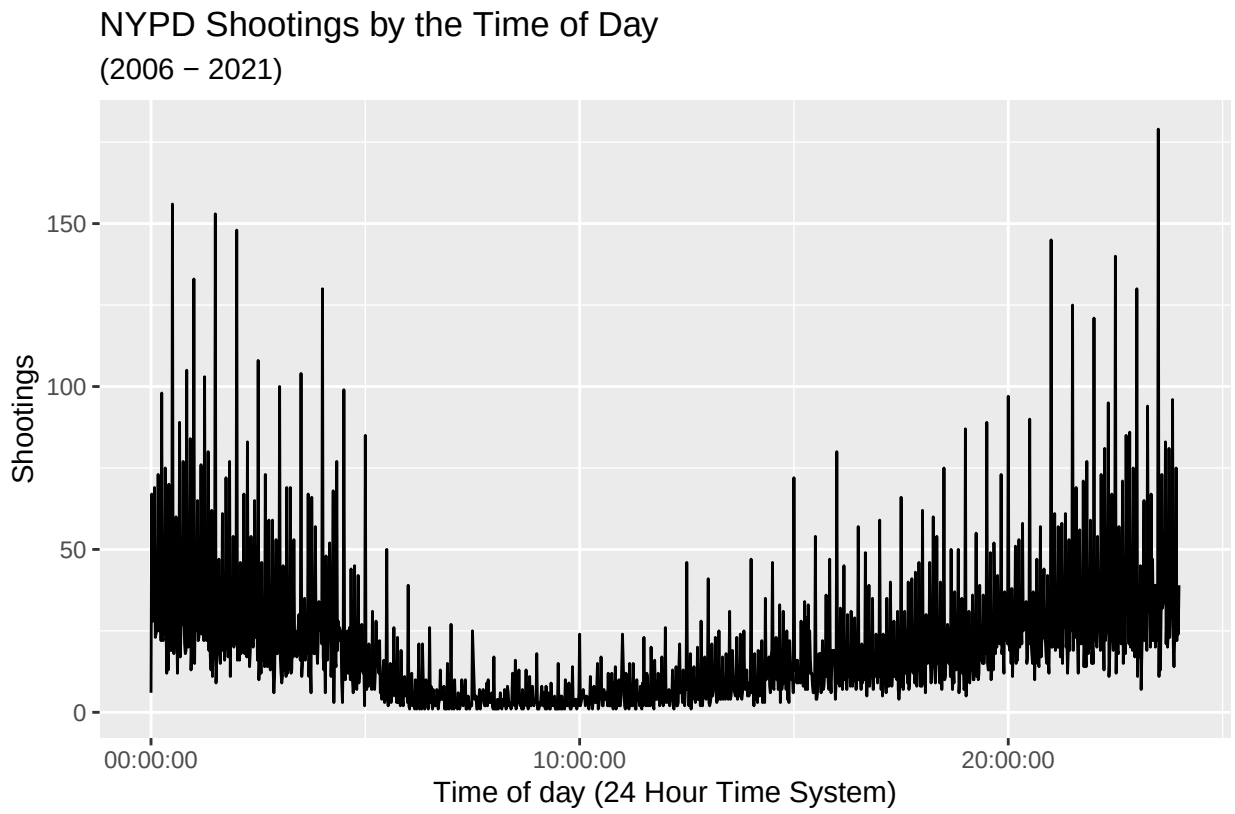
```
NYPD_time_year %>%
    ggplot(aes(x = Time_year, y = Shootings)) +
    geom_line() +
    geom_point(data = NYPD_time_year %>% slice_max(Shootings, n = 2),
               aes(color="Most Shootings")) +
    scale_x_date(date_labels = "%b") +
    labs(title = "NYPD Shootings on that Day",
         subtile = "(2006 - 2021)",
         colour = "",
         x = "Day",
         y = "Shootings on that day",
         caption = "(Figure - 6)")
```



(Figure – 6)

```
NYPD_time_day <- NYPD_clean %>%
    group_by(OCCUR_TIME,Shootings) %>%
    summarize(Shootings = sum(Shootings),
              STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG),
              .groups = 'drop') %>%
    select(OCCUR_TIME,Shootings,STATISTICAL_MURDER_FLAG)

NYPD_time_day %>%
    ggplot(aes(x = OCCUR_TIME, y = Shootings)) +
    geom_line() +
    scale_x_time() +
    labs(title = "NYPD Shootings by the Time of Day",
         subtitle = "(2006 - 2021)",
```

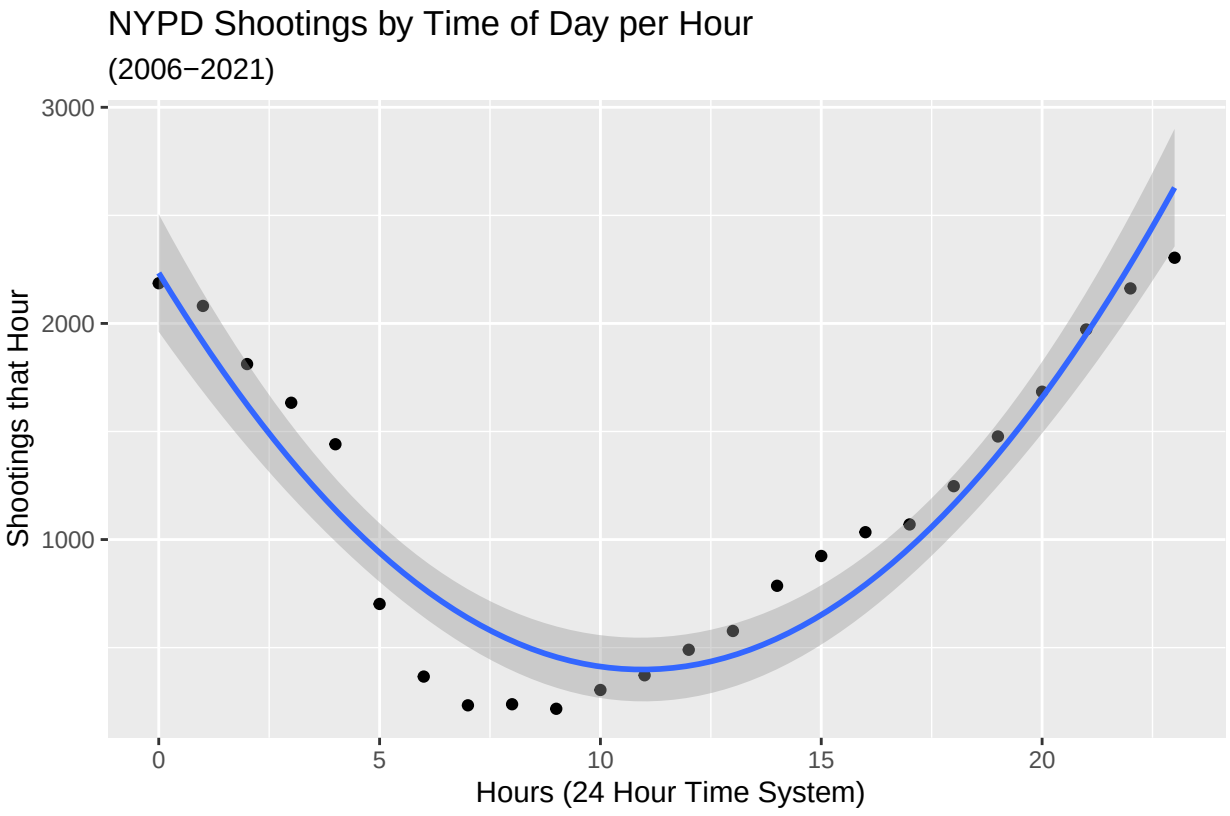
```
        x = "Time of day (24 Hour Time System)",
        y = "Shootings",
        caption = "(Figure - 7)")
```



(Figure − 7)

```
NYPD_time_hour <- NYPD_clean %>%
  mutate(Hour = hour(OCCUR_TIME)) %>%
  group_by(Hour,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG),
            .groups = 'drop') %>%
  mutate(Hour2 = Hour^2) %>%
  select(Hour,Shootings,STATISTICAL_MURDER_FLAG, Hour2)
```

```
NYPD_time_hour %>%
    ggplot(aes(x = Hour, y = Shootings)) +
    geom_point() +
    stat_smooth(method = "lm", formula = y ~ x + I(x^2), linewidth = 1) +
    labs(title = "NYPD Shootings by Time of Day per Hour",
         subtitle = "(2006-2021)",
         x = " Hours (24 Hour Time System)",
         y = "Shootings that Hour",
         caption = "(Figure - 8)")
```



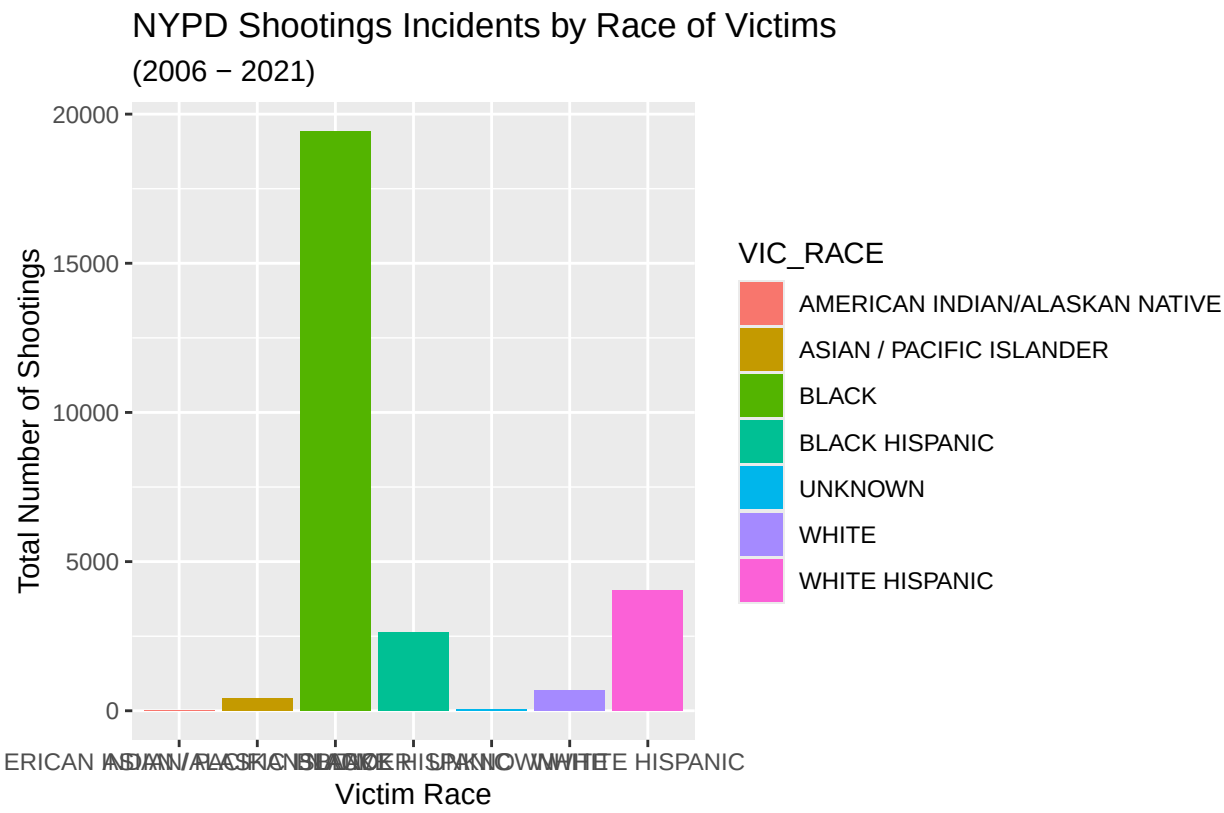
(Figure − 8)

It's rather unsurprising that we have more incident at night to the wee hours

Bias

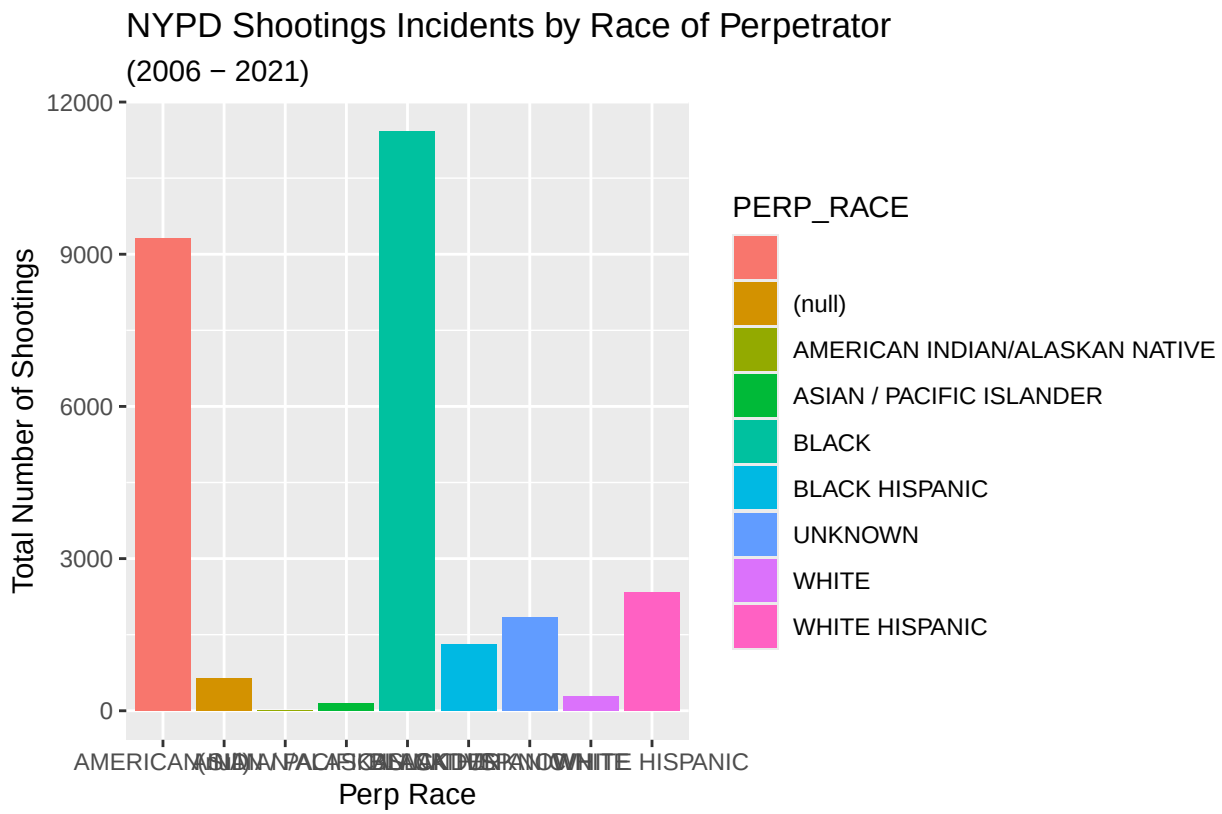
Are there racial bias?

```
NYPD_clean %>%
    ggplot(aes(x = VIC_RACE, fill = VIC_RACE)) +
    geom_bar() +
    labs(title = "NYPD Shootings Incidents by Race of Victims",
         subtitle = "(2006 - 2021)",
         x = "Victim Race",
         y = "Total Number of Shootings",
         caption = "(Figure - 9)")
```



(Figure − 9)

```
NYPD_clean %>%
    ggplot(aes(x = PERP_RACE, fill = PERP_RACE)) +
    geom_bar() +
    labs(title = "NYPD Shootings Incidents by Race of Perpetrator",
         subtitle = "(2006 - 2021)",
         x = "Perp Race",
         y = "Total Number of Shootings",
         caption = "(Figure - 10)")
```



(Figure − 10)

It seems like Afro-American made up significant amount of the data which worth more investigation but I shall stop here before stepping into more sensitive topic.