NYPD_Shooting_Incident

2024-08-19

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

NYPD Shooting Data

List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is made available for analysis via https://catalog.data.gov/dataset

Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?"

NYPD_Shooting_Data <- read_csv(url_in)

summary(NYPD_Shooting_Data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
```

```
##  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:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Cleaning Data

Now we are going to clean the data to make it more readable by useful.

- Unwanted location and jurisdiction columns can be removed.
- Longitude and latitude can be removed.
- OCCUR_DATE variable will be set to date format and OCCUR_TIME can be set as time format.
- Columns BORO, PRECINCT, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, and VIC_RACE can be set as factor

There seems to enough data available in data set to perform multiple analysis. More data may be required for other analysis where total population based on age, sex, or race may be required for each Borough or Precinct. For such cases we will have to find the data sets for NY that provide population information on same time period. Once that data is available we can join that data set with current data set to get more comprehensive data.

```
NYPD_Clean_Shooting_Data <- NYPD_Shooting_Data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE), OCCUR_TIME = hms(OCCUR_TIME)) %>%
  mutate (PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP)) %>%
  mutate (PERP_SEX = as.factor(PERP_SEX)) %>%
  mutate (PERP_RACE = as.factor(PERP_RACE)) %>%
  mutate (VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP)) %>%
  mutate (VIC_SEX = as.factor(VIC_SEX)) %>%
  mutate (VIC_RACE = as.factor(VIC_RACE)) %>%
  mutate (BORO = as.factor(BORO)) %>%
  mutate (PRECINCT = as.factor(PRECINCT)) %>%
  select (-JURISDICTION_CODE,  -LOC_OF_OCCUR_DESC, -LOC_CLASSFCTN_DESC, -LOCATION_DESC,
          -X_COORD_CD, -Y_COORD_CD, -Latitude, -Longitude, -Lon_Lat)

summary (NYPD_Clean_Shooting_Data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 65439914   1st Qu.:2009-09-04   1st Qu.:3H 30M 0S
##  Median : 92711254   Median :2013-09-20   Median :15H 15M 0S
##  Mean   :127405824   Mean   :2014-06-07   Mean   :12H 44M 16.7131153280575
##  3rd Qu.:203131993   3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S
##  Max.   :279758069   Max.   :2023-12-29   Max.   :23H 59M 0S
##
##             BORO          PRECINCT     STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  BRONX        : 8376   75     : 1628   Mode :logical           18-24  :6438
##  BROOKLYN     :11346   73     : 1500   FALSE:23036             25-44  :6041
##  MANHATTAN    : 3762   67     : 1259   TRUE :5526              UNKNOWN:3148
##  QUEENS       : 4271   44     : 1076                           <18    :1682
##  STATEN ISLAND:  807   79     : 1045                           (null) :1141
##                        47     : 1006                           (Other): 768
##                        (Other):21048                           NA's   :9344
##    PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  (null): 1141   BLACK         :11903   <18    : 2954   F: 2760
##  F     :  444   WHITE HISPANIC: 2510   1022   :    1   M:25790
##  M     :16168   UNKNOWN       : 1837   18-24  :10384   U:   12
##  U     : 1499   BLACK HISPANIC: 1392   25-44  :12973
##  NA's  : 9310   (null)        : 1141   45-64  : 1981
##                 (Other)       :  469   65+    :  205
##                 NA's          : 9310   UNKNOWN:   64
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   11
##  ASIAN / PACIFIC ISLANDER      :  440
##  BLACK                         :20235
##  BLACK HISPANIC                : 2795
##  UNKNOWN                       :   70
##  WHITE                         :  728
##  WHITE HISPANIC                : 4283
```

Focus of Analysis

While the available data can be used for different analysis spanning from trends over the years, geographical analysis based on location of incident, demographic details for victims and perpetrator, as well as characteristic of incidents, In this R markdown document I am focusing on data available for each borough. What we are trying to find how each borough compares to others based on different parameters and who is most likely to be victim of these incidents.

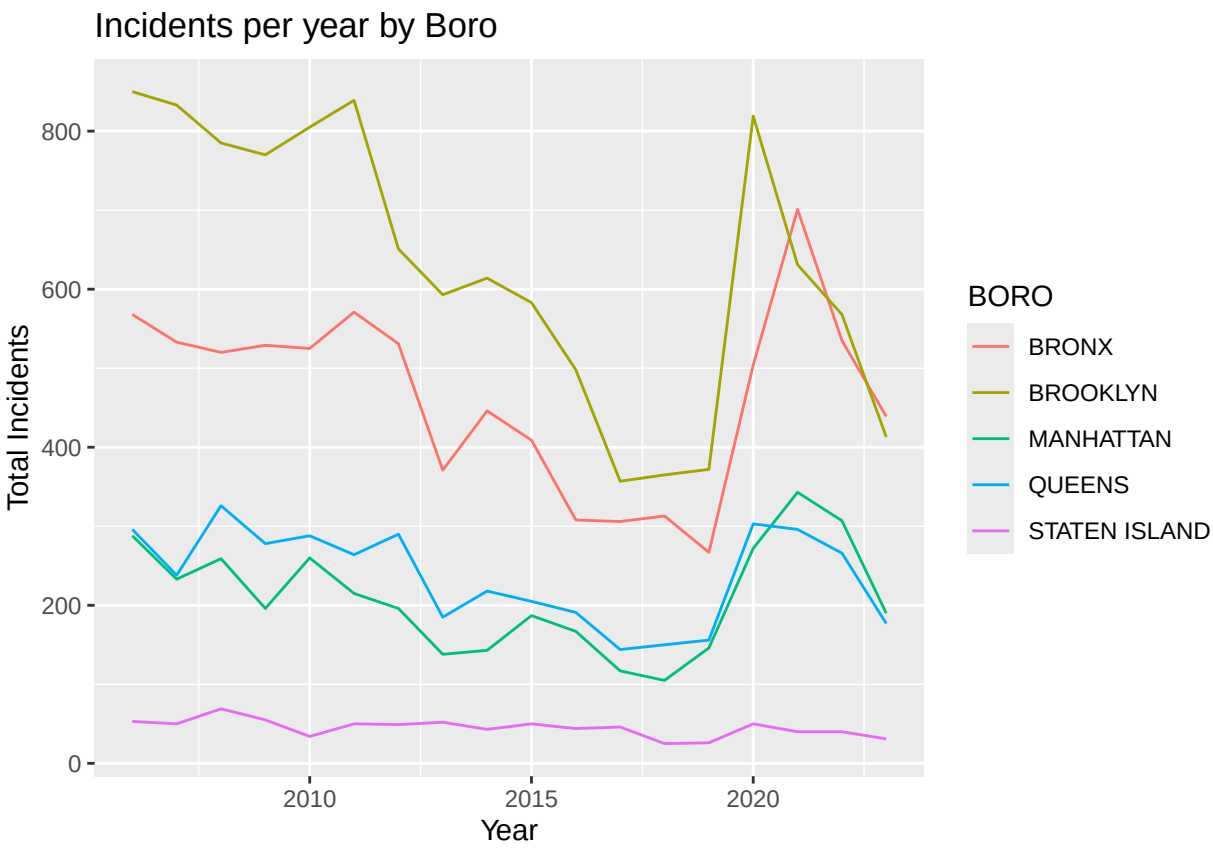
Shooting Incidents per Year

First we will look at shooting incidents for each borough on yearly basis. To achieve that we need to find out the yearly data for each borough using the cleaned data. Once the data is available for yearly counts we can plot the total incidents for each year based on borough. All areas seems to follow the same trend where reported incidents were on decline but then a sudden jump was reported in 2020-2022.

```
NYPD_Shooting_Data_Per_Year <- NYPD_Clean_Shooting_Data %>%
  mutate(Year = year(OCCUR_DATE))

NYPD_Shooting_Data_Per_Year %>%
```

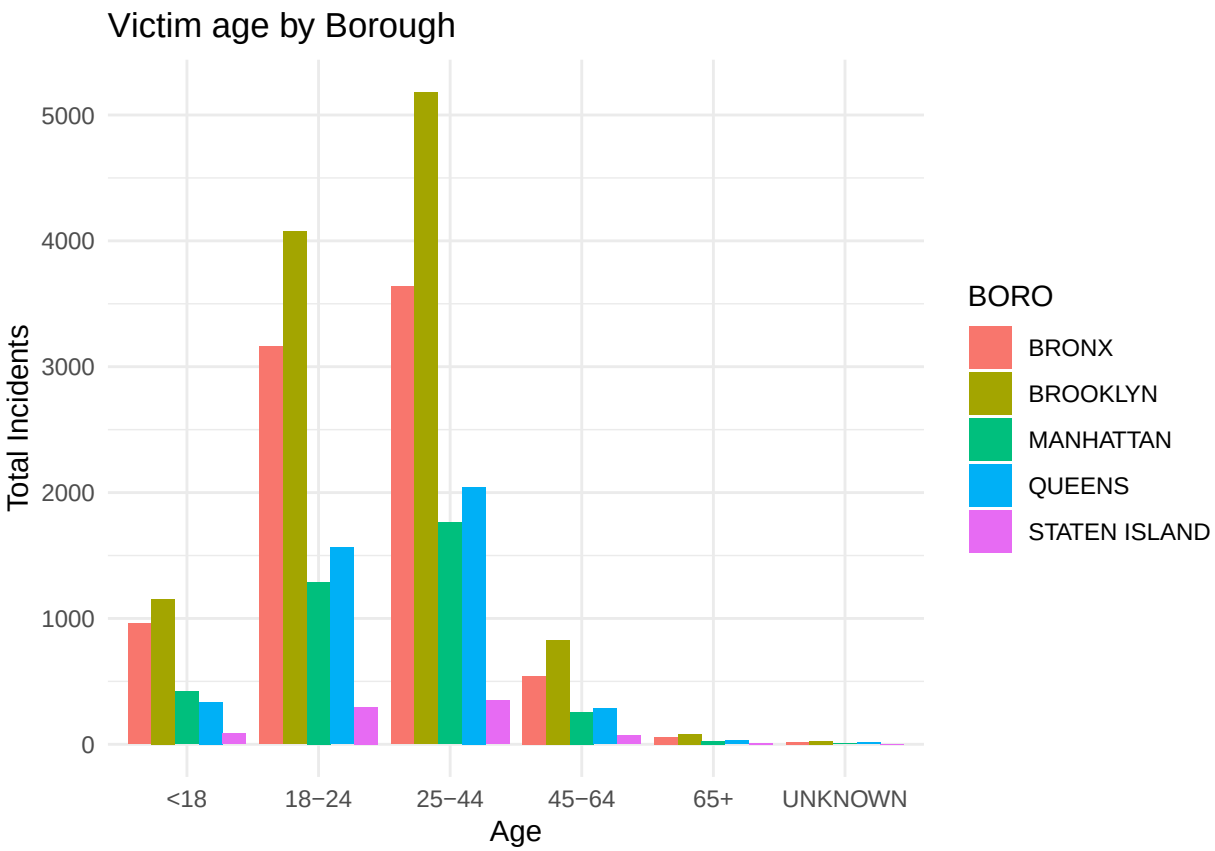
```
  group_by (Year, BORO) %>%
  summarise( Total = n()) %>%
  ggplot(aes(x=Year, y=Total, color=BORO)) +
  geom_line() +
  labs(title = "Incidents per year by Boro", x="Year", y="Total Incidents", color = "BORO")
```



Victim Age Distribution

We can also analyze the data to find how different age groups are impacted by these incidents in each borough. Looking at summary of cleaned data we found that there is an entry for which victim age is not entered properly, this entry can be ignored in analysis. People in age group 18-24 and 25-44 are the most reported victims of the incidents.

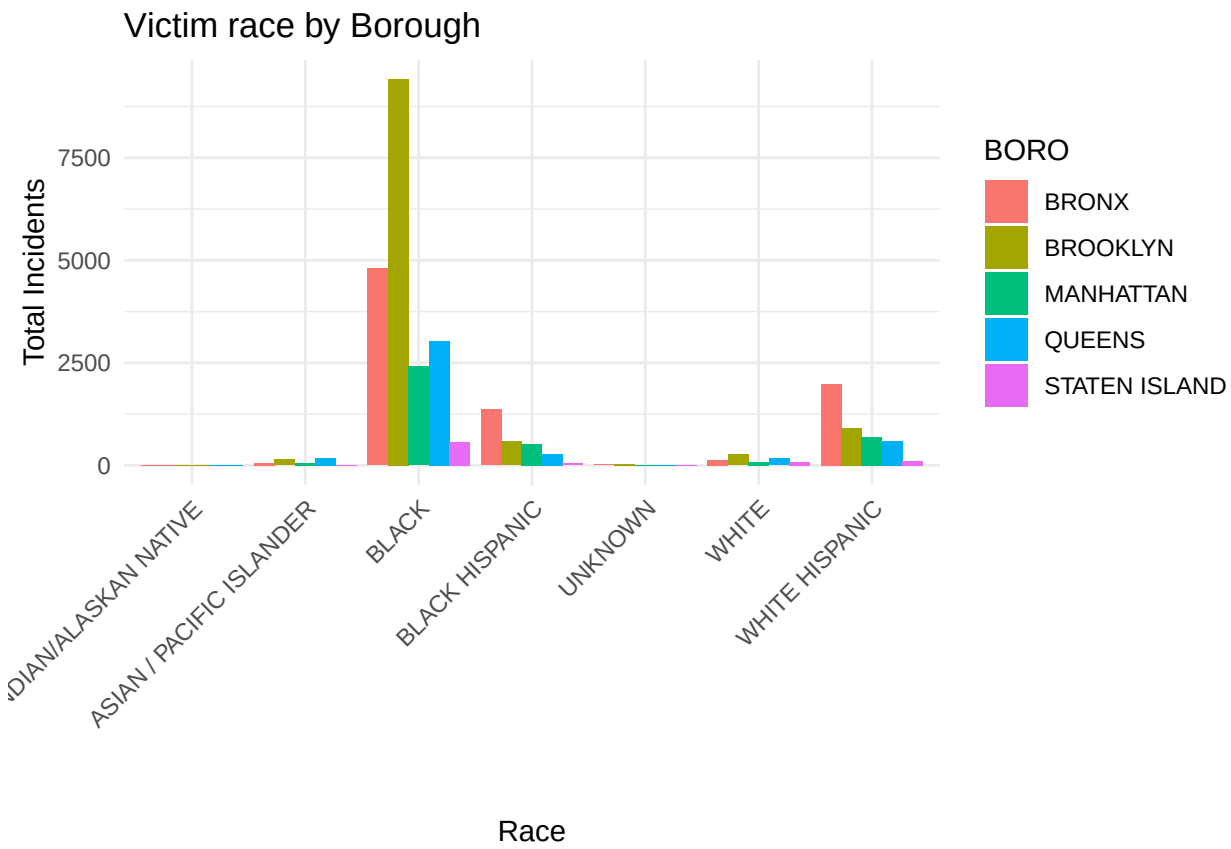
```
NYPD_Clean_Shooting_Data %>%
 filter(VIC_AGE_GROUP != 1022) %>% ##Remove a wrong entry from data set
 group_by(VIC_AGE_GROUP, BORO) %>%
 summarise(Total = n()) %>%
 ggplot(aes(x=VIC_AGE_GROUP, y=Total, fill=BORO)) +
 geom_bar(stat = "Identity", position = position_dodge()) +
 labs(title = "Victim age by Borough",
      x="Age",
      y="Total Incidents",
      color = "BORO") +
 theme_minimal()
```



Victim Race Distribution

We can also how people from different races are impacted in each borough. This graph show clearly that a black person is more likely to be victim.

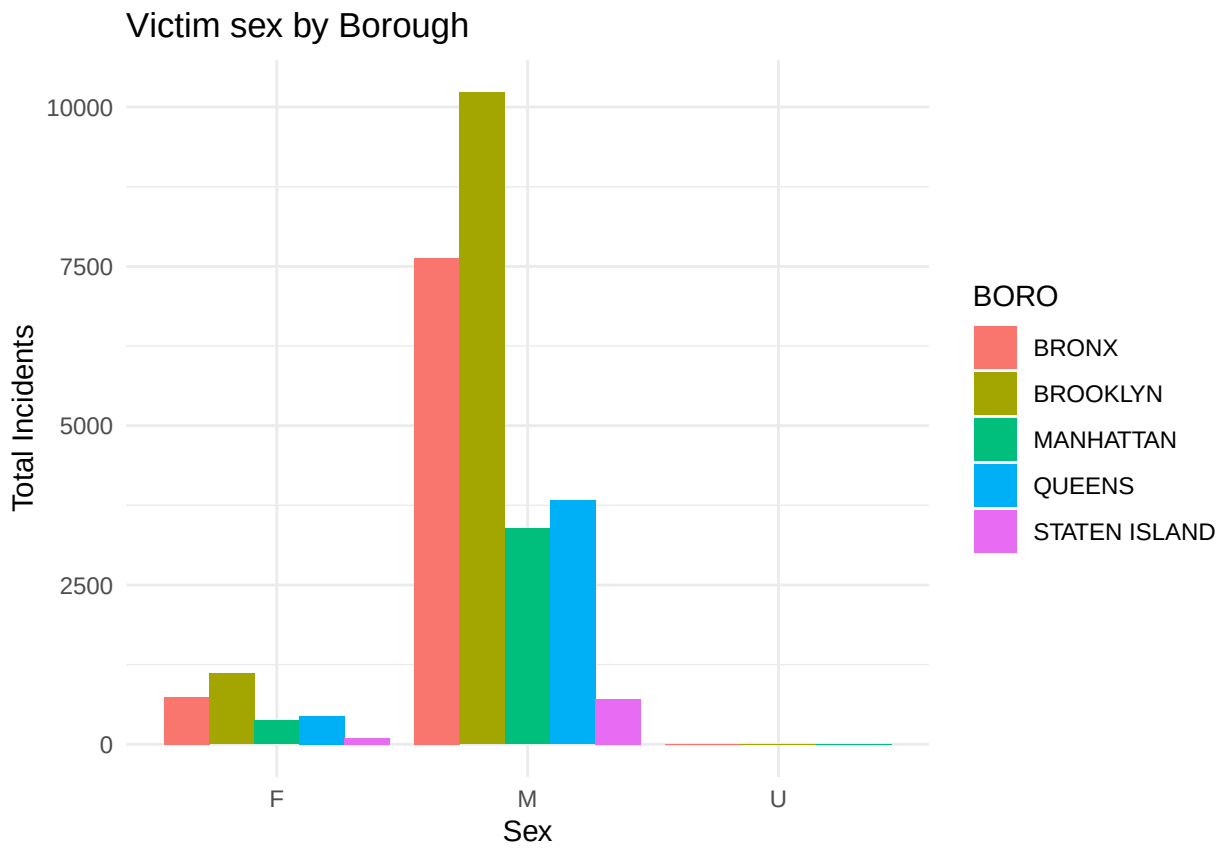
```
NYPD_Clean_Shooting_Data %>%
 group_by(VIC_RACE, BORO) %>%
 summarise(Total = n()) %>%
 ggplot(aes(x=VIC_RACE, y=Total, fill=BORO)) +
 geom_bar(stat = "Identity", position = position_dodge()) +
 labs(title = "Victim race by Borough",
      x="Race",
      y="Total Incidents",
      color = "BORO") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Victim Sex Distribution

Lets have look at how people from different sex are impacted in each borough. This visualization shows that males more prone to be victim when compare to other sex.

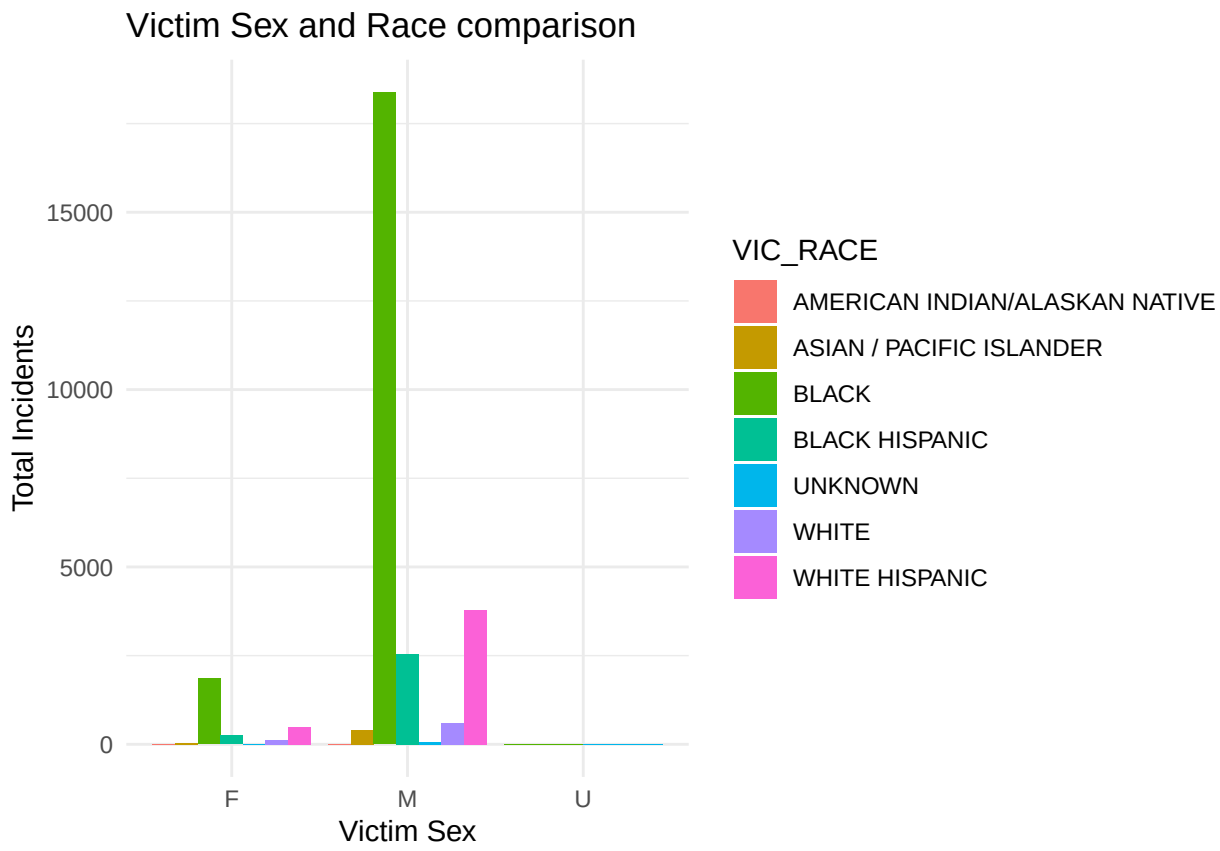
```
NYPD_Clean_Shooting_Data %>%
 group_by(VIC_SEX, BORO) %>%
 summarise(Total = n()) %>%
 ggplot(aes(x=VIC_SEX, y=Total, fill=BORO)) +
 geom_bar(stat = "Identity", position = position_dodge()) +
 labs(title = "Victim sex by Borough",
      x="Sex",
      y="Total Incidents",
      color = "BORO") +
 theme_minimal()
```



Victime Race/Sex Distribution

A different view to visualize how victims of different sex and race are impacted in these shooting incidents. This shows how black males are more at risk.

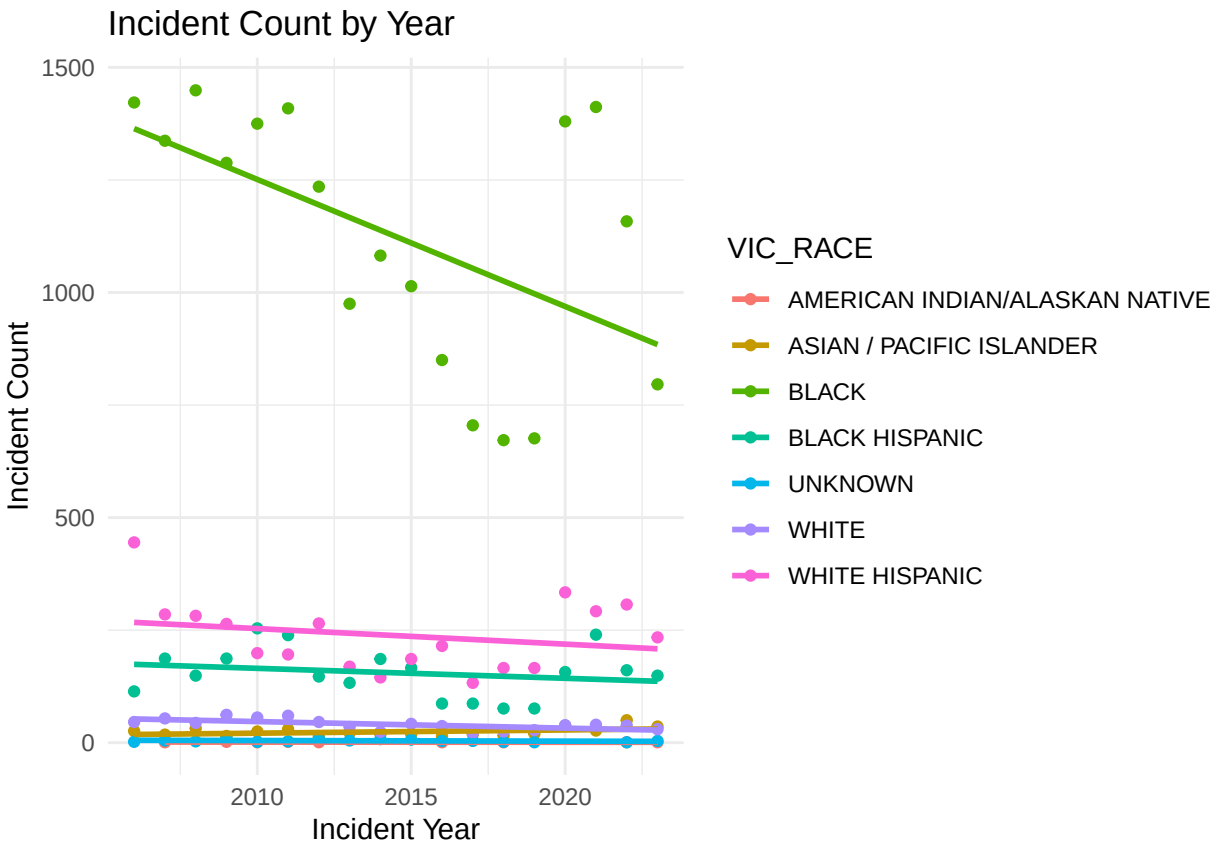
```
NYPD_Clean_Shooting_Data %>%
 group_by(VIC_SEX, VIC_RACE) %>%
 summarise(Total = n()) %>%
 ggplot(aes(x=VIC_SEX, y=Total, fill=VIC_RACE)) +
 geom_bar(stat = "Identity", position = position_dodge()) +
 labs(title = "Victim Sex and Race comparison",
      x="Victim Sex",
      y="Total Incidents",
      color = "Victim Race") +
 theme_minimal()
```



Model

One basic model that can be used for this data set is to visualize a trend on number of incidents over the years and how it co-relates to victims race.

```
NYPD_Shooting_Data_Per_Year %>%
  group_by(Year, VIC_RACE) %>%
  summarise(Total = n()) %>%
  ggplot(aes(x = Year, y = Total, color = VIC_RACE)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Incident Count by Year",
       x = "Incident Year",
       y = "Incident Count") +
  theme_minimal()
```



Conclusion

NYPD Shooting incident data provides all the shooting incidents since 2006 to 2023. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.

This data can be used to analyze and produce models to show who are the common victims by the age group or by the sex. We can also see if any specific race is more impacted by these crimes as compared to others. Locations for these crimes can also indicate if an particular area is more prone to these crimes and help improving the security measure on these areas.

Above analysis provided in this R Markdown documents is able to help us identify that a black male in age 18-44 and living in Brooklyn is most likely to victim of shooting incident followed by similar person on Bronx. Staten Islands on the other hand seems to be a much better place for all ages, gender and race.

One has to be careful when interpreting this data to produce models so that personal biases are not influencing the results and model are not leaning towards those biases. My personal bias after first look at data was related to location of the incidents. It seemed like most of the incident reported happened at the multi dwelling units. Since analysis found that not all incident reported has a location, and additional data will be required to make that as an outcome of this analysis.