NYPD_Shooting

DTSA

April 23, 2023

Step 1 Importing Data

The data use for this Analysis is the shooting incident data that occured in NYC going back to 2006 through the end of 2022. The data is imported from[data.gov website. The first think we are going to do before starting our Analysis is to import tidyverse package bacause we are going to use them for data wrangling. We also need to import the lubridade package since we are going to deal with date and time for our analys.

```
###call the tidyverse library
## use url to import data
library("tidyverse")
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3

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

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

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

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.3

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

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

Let's read in the data and see what we have.

```
NYPD <- read_csv(url[1])
```

```
NYPD
```

```
## # A tibble: 25,596 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>        <dbl>             <dbl>
##  1    236168668 11/11/2021 15:04      BROOKLYN        79                 0
##  2    231008085 07/16/2021 22:05      BROOKLYN        72                 0
##  3    230717903 07/11/2021 01:09      BROOKLYN        79                 0
##  4    237712309 12/11/2021 13:42      BROOKLYN        81                 0
##  5    224465521 02/16/2021 20:00      QUEENS         113                 0
##  6    228252164 05/15/2021 04:13      QUEENS         113                 0
##  7    226950018 04/14/2021 21:08      BRONX           42                 0
##  8    237710987 12/10/2021 19:30      BRONX           52                 0
##  9    224701998 02/22/2021 00:18      MANHATTAN       34                 0
## 10    225295736 03/07/2021 06:15      BROOKLYN        75                 0
## # ... with 25,586 more rows, and 13 more variables: 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>
```

Data Descritption.

Our data contatins 19 columns, the description can be found here. The following columns are the columns that we need for our analysis.

1. INCIDENT_KEY: Randomly generated persistent ID for each arrest

2.OCCUR_DATE: Exact date of the shooting incident

3.OCCUR_TIME:Exact time of the shooting incident

4.BORO: Borough where the shooting incident occurred

5.PRECINCT: Precinct where the shooting incident occurred. The list of precinct can be found here

6.JURISDICTION_CODE:Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD while codes 3 and more represent non NYPD jurisdictions.

7.LOCATION_DESC :Location of the shooting incident

8. STATISTICAL_MURDER_FLAG
   Shooting resulted in the victim's death which would be counted as a murder

9.VIC_AGE_GROUP:Victim's age within a category

10.VIC_SEX:Victim's sex description

11.VIC_RACE :Victim's race description

Step 2 Exploratory Data Analysis

We are going to conduct some exploratory data analysis in order to learn more about our dataset. ### Shape of our dataset

Here the function glimpse will provide the shape of our dataset, we are going to be able to know the number of columns which are the attributes of our dataset and the number of rows which is consider as the number of record or entry of our dataset.The function glimpse diplay also the name of each attributes and its corresponding variable type.

```
glimpse(NYPD)
```

```
## Rows: 25,596
## Columns: 19
## $ INCIDENT_KEY            <dbl> 236168668, 231008085, 230717903, 23771230~
## $ OCCUR_DATE              <chr> "11/11/2021", "07/16/2021", "07/11/2021",~
## $ OCCUR_TIME              <time> 15:04:00, 22:05:00, 01:09:00, 13:42:00, ~
## $ BORO                    <chr> "BROOKLYN", "BROOKLYN", "BROOKLYN", "BROO~
## $ PRECINCT                <dbl> 79, 72, 79, 81, 113, 113, 42, 52, 34, 75,~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0,~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, "COMMERCIAL BLDG"~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
## $ PERP_AGE_GROUP          <chr> NA, "45-64", "<18", NA, NA, NA, NA, NA, N~
## $ PERP_SEX                <chr> NA, "M", "M", NA, NA, NA, NA, NA, NA, "M"~
## $ PERP_RACE               <chr> NA, "ASIAN / PACIFIC ISLANDER", "BLACK", ~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "25-44", "25-4~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "ASIAN / PACIFIC ISLANDER", "BLA~
## $ X_COORD_CD              <dbl> 996313, 981845, 996546, 1001139, 1050710,~
## $ Y_COORD_CD              <dbl> 187499, 171118, 187436, 192775, 184826, 1~
## $ Latitude                <dbl> 40.68132, 40.63636, 40.68114, 40.69579, 4~
## $ Longitude               <dbl> -73.95651, -74.00867, -73.95567, -73.9391~
## $ Lon_Lat                 <chr> "POINT (-73.95650899099996 40.68131820000~
```

Our dataset contains 19 rows(attributes) and 25596 columns. We don't need drop some of attributes for our analysis. When we look at the variable type of each attribute. We can see that the attribut Occur_date is a charactere and we will like to convert it to date data type.

Let also check the percentage of missing value for each attributes.

```
sum(is.na(NYPD))
```

```
## [1] 42943
```

```
# calculating percentage of missing values
(colMeans(is.na(NYPD)))*100
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##             0.000000000             0.000000000             0.000000000
##                    BORO                PRECINCT       JURISDICTION_CODE
##             0.000000000             0.000000000             0.007813721
##           LOCATION_DESC STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##            58.513048914             0.000000000            36.505704016
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##            36.372870761            36.372870761             0.000000000
##                 VIC_SEX                VIC_RACE              X_COORD_CD
##             0.000000000             0.000000000             0.000000000
##              Y_COORD_CD                Latitude               Longitude
##             0.000000000             0.000000000             0.000000000
##                 Lon_Lat
##             0.000000000
```

Among our 19 Attributes, 5 have missing values. LOCATION_DESC has 58.5 percent of missing value, PERP_SEX has 36.37 percent of missing values, PERP_RACE has 36.37 percent of missing values, JURIDICTION_code has 0.008 percent of missing values and PERP_AGE_GROUP has 36.50 percent of missing values.

Let drop all the columns with more than 20% of missing values.

```
NYPDShooting = select(NYPD,-c(LOCATION_DESC, PERP_SEX, PERP_RACE,PERP_AGE_GROUP ))
```

We have dropped the attributes we more than 20% of missing values, now we are going to get ride of the attribute we don't need for our analysis. Let drop Longitude, Latitude, lon_lat, X_COORD_CD, Y_COORD_CD, INCIDENT_KEY

```
NYPDShooting = select(NYPDShooting, -c(Longitude, Latitude, Lon_Lat,X_COORD_CD, Y_COORD_CD, INCIDENT_KE
```

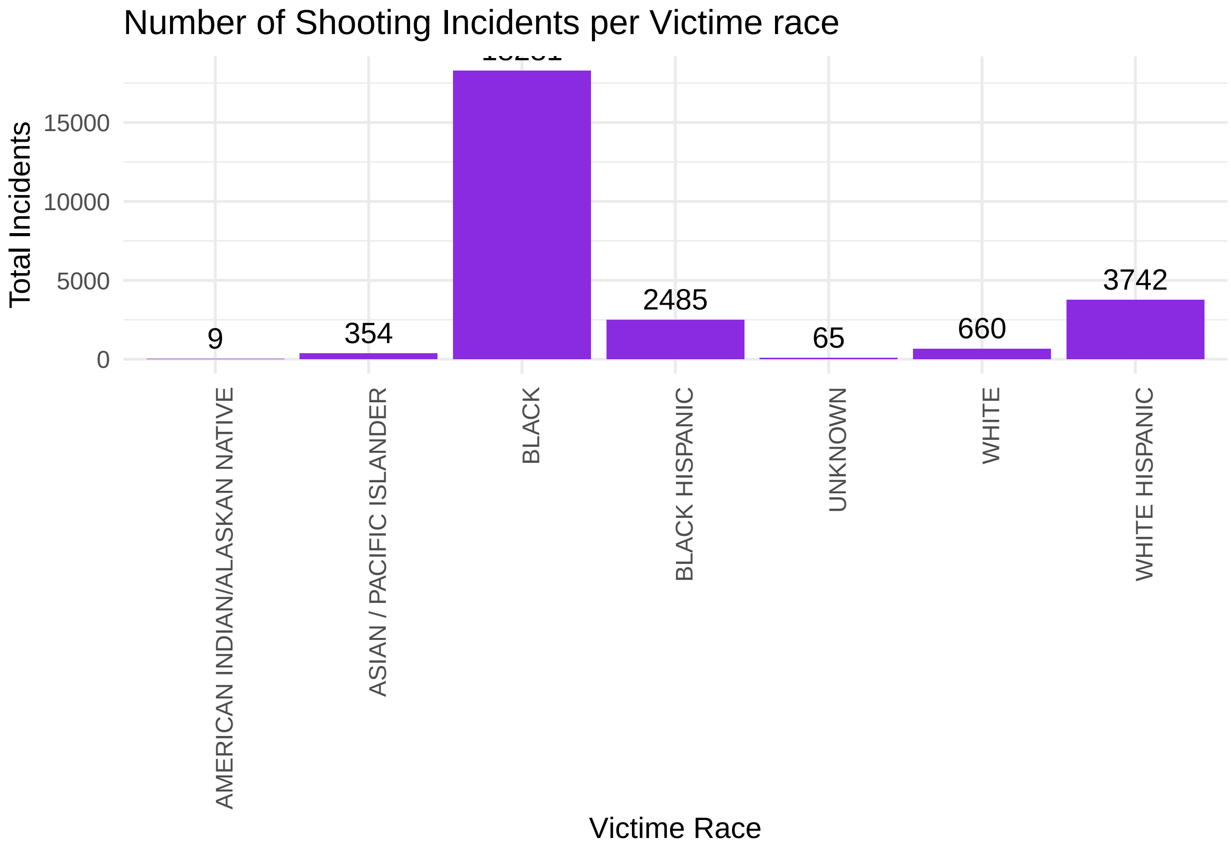
```
glimpse(NYPDShooting)
```

```
## Rows: 25,596
## Columns: 9
## $ OCCUR_DATE              <chr> "11/11/2021", "07/16/2021", "07/11/2021",~
## $ OCCUR_TIME              <time> 15:04:00, 22:05:00, 01:09:00, 13:42:00, ~
## $ BORO                    <chr> "BROOKLYN", "BROOKLYN", "BROOKLYN", "BROO~
## $ PRECINCT                <dbl> 79, 72, 79, 81, 113, 113, 42, 52, 34, 75,~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0,~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "25-44", "25-4~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "ASIAN / PACIFIC ISLANDER", "BLA~
```

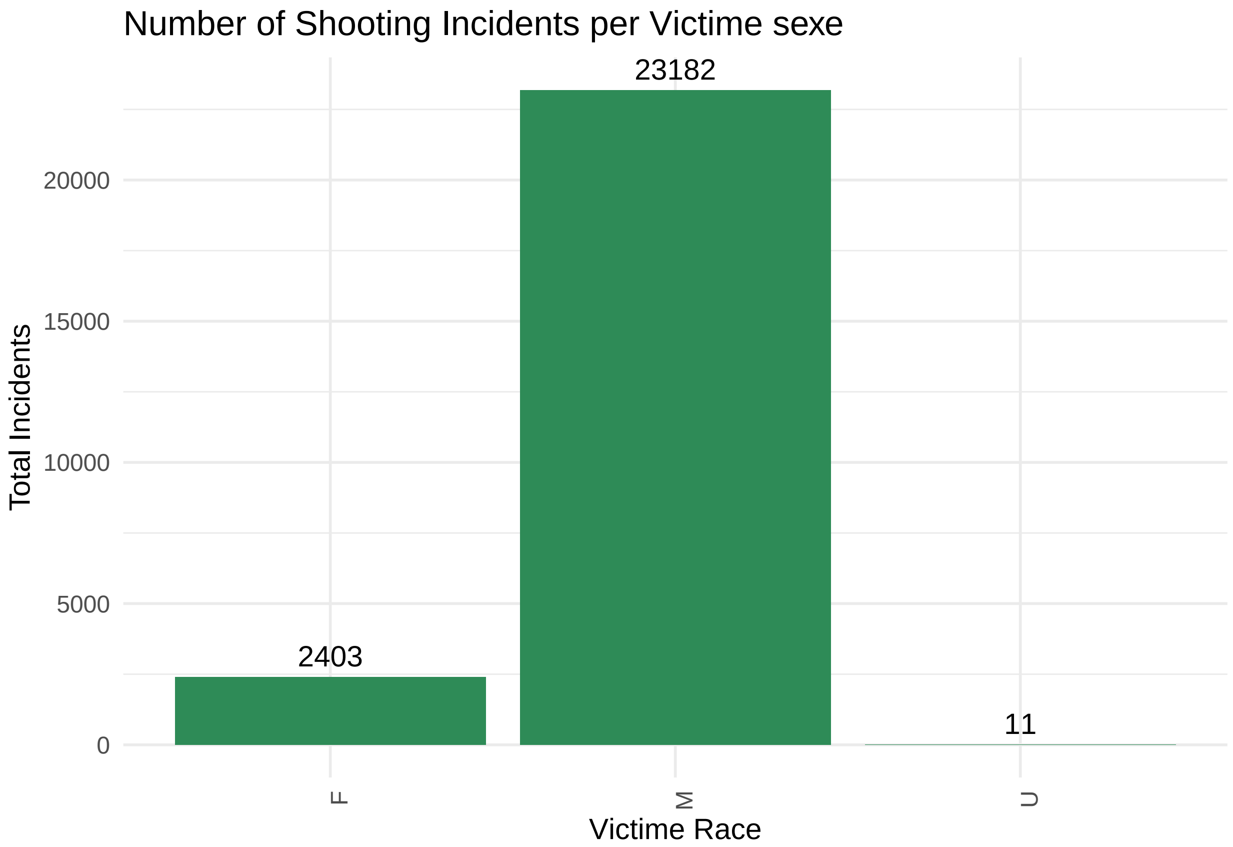
Step 3 Data Vizualization

Plot number of shooting per victime race, victime sexe or victime GE GROUP

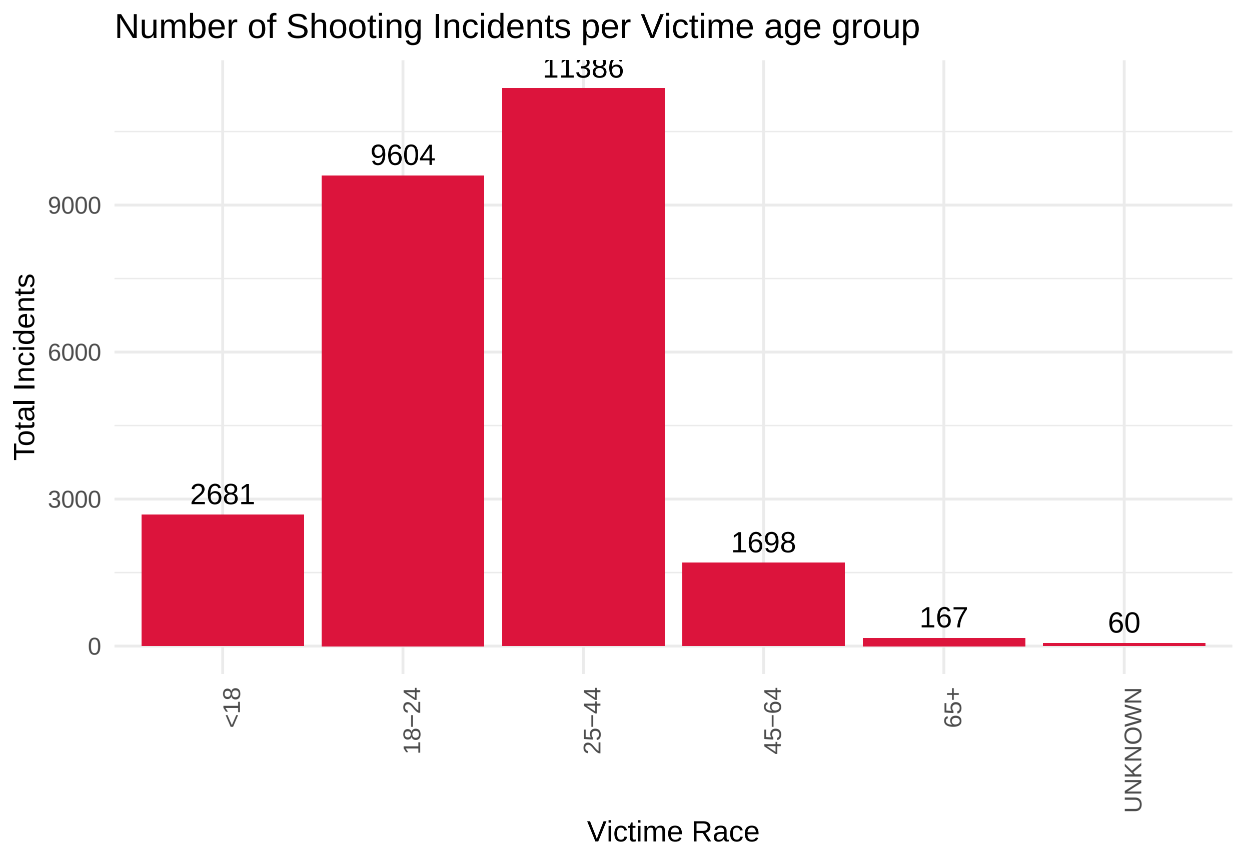
```
# Group the data by jurisdiction_code and calculate the total number of incidents in each jurisdiction
nypd_shooting_counts <- NYPDShooting %>%
                        group_by(VIC_RACE) %>%
                        summarize(total_incidents = n())

# Customize the plot
bar_color <- "#8A2BE2" # Change the bar color to blue

ggplot(nypd_shooting_counts, aes(x = VIC_RACE, y = total_incidents, fill=VIC_RACE)) +
  geom_bar(stat = "identity", fill = bar_color) +
  labs(x = "Victime Race", y = "Total Incidents", title = "Number of Shooting Incidents per Victime rac
  theme_minimal() + # Use a minimalistic theme
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + # Rotate the x-axis labels for readability
  geom_text(aes(label = total_incidents), vjust = -0.5) # Add labels to the bars
```



```
# Group the data by jurisdiction_code and calculate the total number of incidents in each jurisdiction
nypd_shooting_counts <- NYPDShooting %>%
                        group_by(VIC_SEX) %>%
                        summarize(total_incidents = n())

# Customize the plot
bar_color <- "#2E8B57" # Change the bar color to blue

ggplot(nypd_shooting_counts, aes(x = VIC_SEX, y = total_incidents, fill=VIC_SEX)) +
  geom_bar(stat = "identity", fill = bar_color) +
  labs(x = "Victime Race", y = "Total Incidents", title = "Number of Shooting Incidents per Victime sex
  theme_minimal() + # Use a minimalistic theme
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + # Rotate the x-axis labels for readability
  geom_text(aes(label = total_incidents), vjust = -0.5) # Add labels to the bars
```



```
# Group the data by jurisdiction_code and calculate the total number of incidents in each jurisdiction
nypd_shooting_counts <- NYPDShooting %>%
                        group_by(VIC_AGE_GROUP) %>%
                        summarize(total_incidents = n())

# Customize the plot
bar_color <- "#DC143C" # Change the bar color to blue

ggplot(nypd_shooting_counts, aes(x = VIC_AGE_GROUP, y = total_incidents, fill=VIC_AGE_GROUP)) +
  geom_bar(stat = "identity", fill = bar_color) +
  labs(x = "Victime Race", y = "Total Incidents", title = "Number of Shooting Incidents per Victime age
  theme_minimal() + # Use a minimalistic theme
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + # Rotate the x-axis labels for readability
  geom_text(aes(label = total_incidents), vjust = -0.5) # Add labels to the bars
```



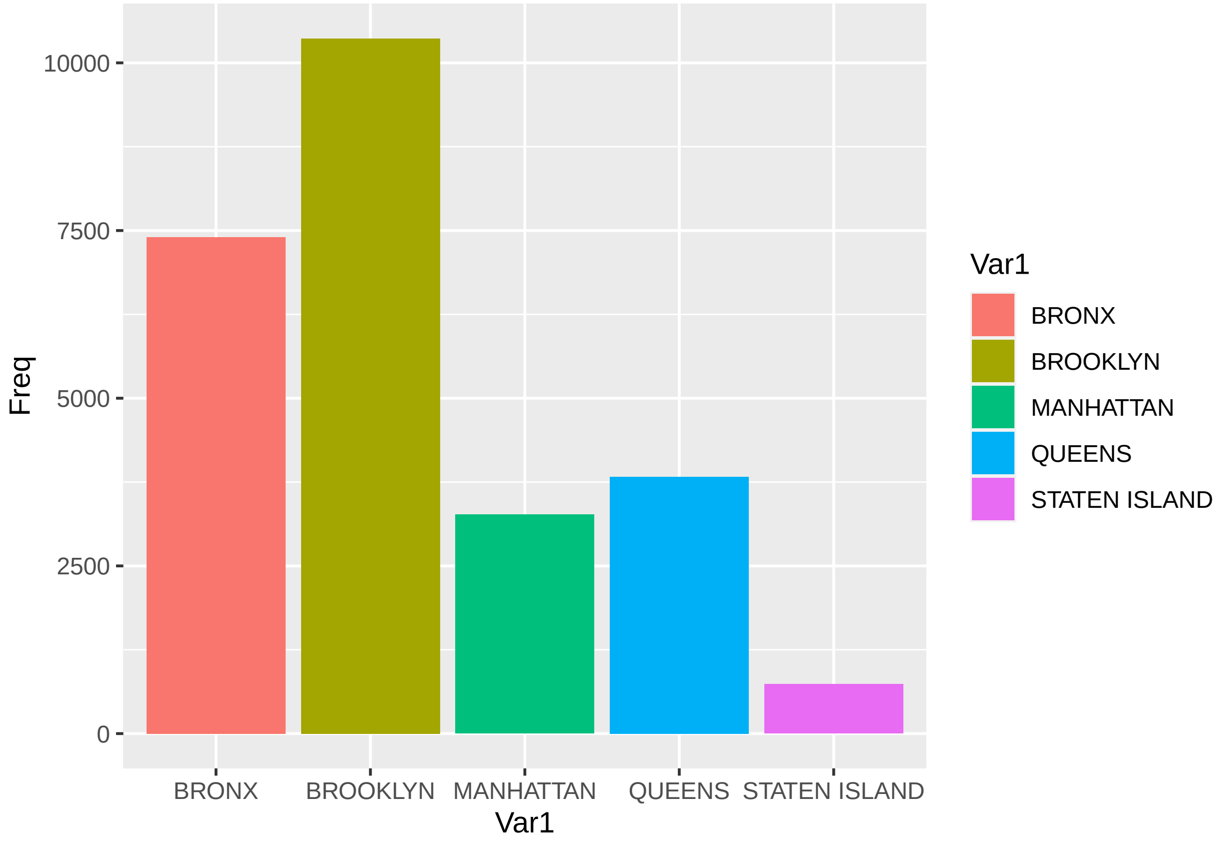
#Percentage of crime in each boro

We want to know which boro has the highest crime.

```
Borough <- table(NYPDShooting$BORO)
Borough <- as.data.frame(Borough)
Borough$Percent <- round((Borough$Freq / sum(Borough$Freq)*100),2)
Borough
```

```
##              Var1  Freq Percent
## 1           BRONX  7402   28.92
## 2        BROOKLYN 10365   40.49
## 3       MANHATTAN  3265   12.76
## 4          QUEENS  3828   14.96
## 5 STATEN ISLAND   736    2.88
```

```
ggplot(Borough, aes(x=Var1, y=Freq, fill=Var1)) + geom_bar(stat="identity")
```



From our graph we can see that Brooklyn has the highest number of shooting.

Plotting Graph Between Number of Cases and Month on Each BORO

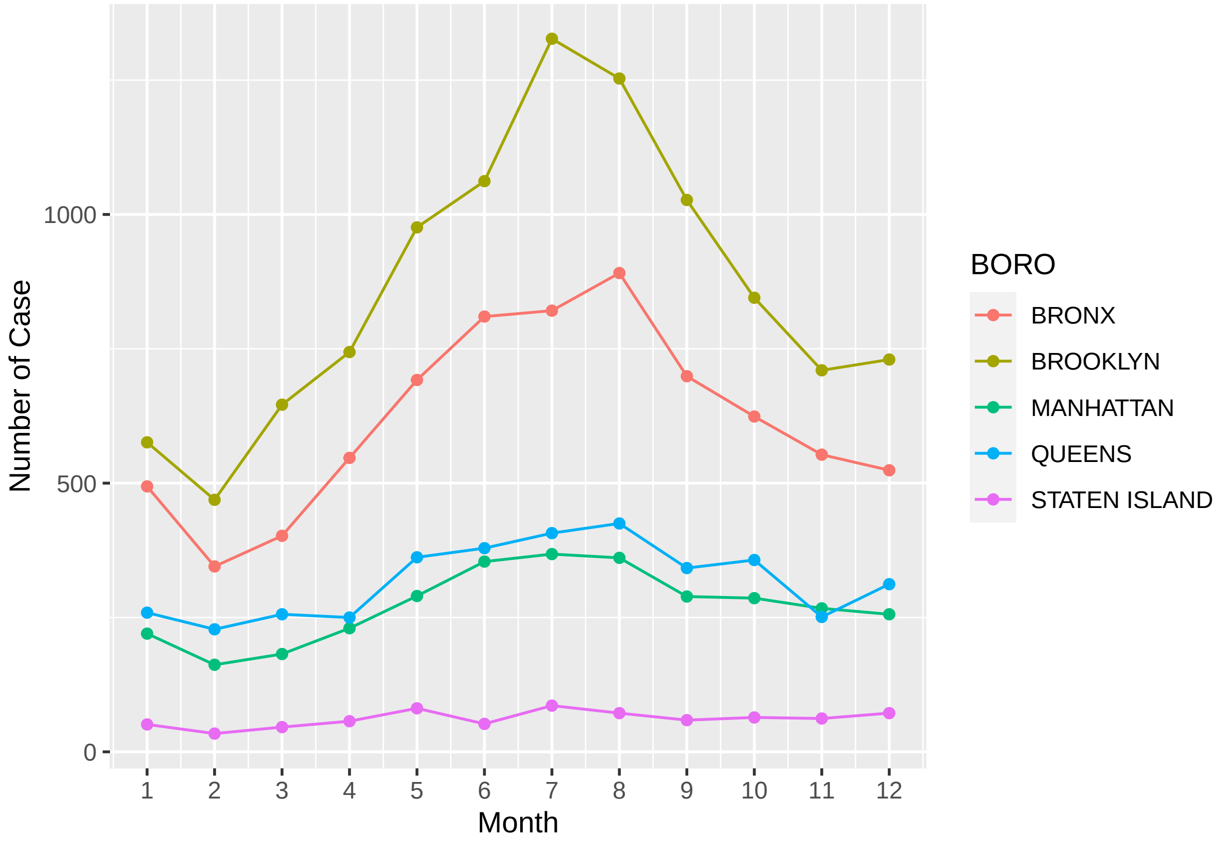
```
NYPD <- NYPDShooting %>%
  select(c(1,2,3,4)) %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, "%m/%d/%Y"),
         case = 1)

NYPD = NYPD%>%
  mutate(OCCUR_MONTH = as.numeric(format(NYPD$OCCUR_DATE, '%m')))
summary(NYPD)
```

```
##    OCCUR_DATE          OCCUR_TIME          BORO
##  Min.   :2006-01-01   Length:25596      Length:25596
##  1st Qu.:2009-05-10   Class1:hms        Class :character
##  Median :2012-08-26   Class2:difftime   Mode  :character
##  Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:2017-07-01
##  Max.   :2021-12-31
##     PRECINCT          case     OCCUR_MONTH
##  Min.   :  1.00   Min.   :1   Min.   : 1.000
##  1st Qu.: 44.00   1st Qu.:1   1st Qu.: 5.000
##  Median : 69.00   Median :1   Median : 7.000
##  Mean   : 65.87   Mean   :1   Mean   : 6.857
##  3rd Qu.: 81.00   3rd Qu.:1   3rd Qu.: 9.000
##  Max.   :123.00   Max.   :1   Max.   :12.000
```

```
NYPDMonth = NYPD%>%
  group_by(OCCUR_MONTH, BORO)%>%
  summarise(case = sum(case))
```

```
## ‘summarise()‘ has grouped output by ’OCCUR_MONTH’. You can override using the ‘.groups‘ argument.
```

```
NYPDMonth %>%
  ggplot(aes(x = OCCUR_MONTH, y = case)) +
  geom_point(aes(color = BORO)) +
  geom_line(aes(color = BORO)) +
  scale_x_continuous(breaks=c(1:12)) +
  labs(x = "Month", y = "Number of Case")
```

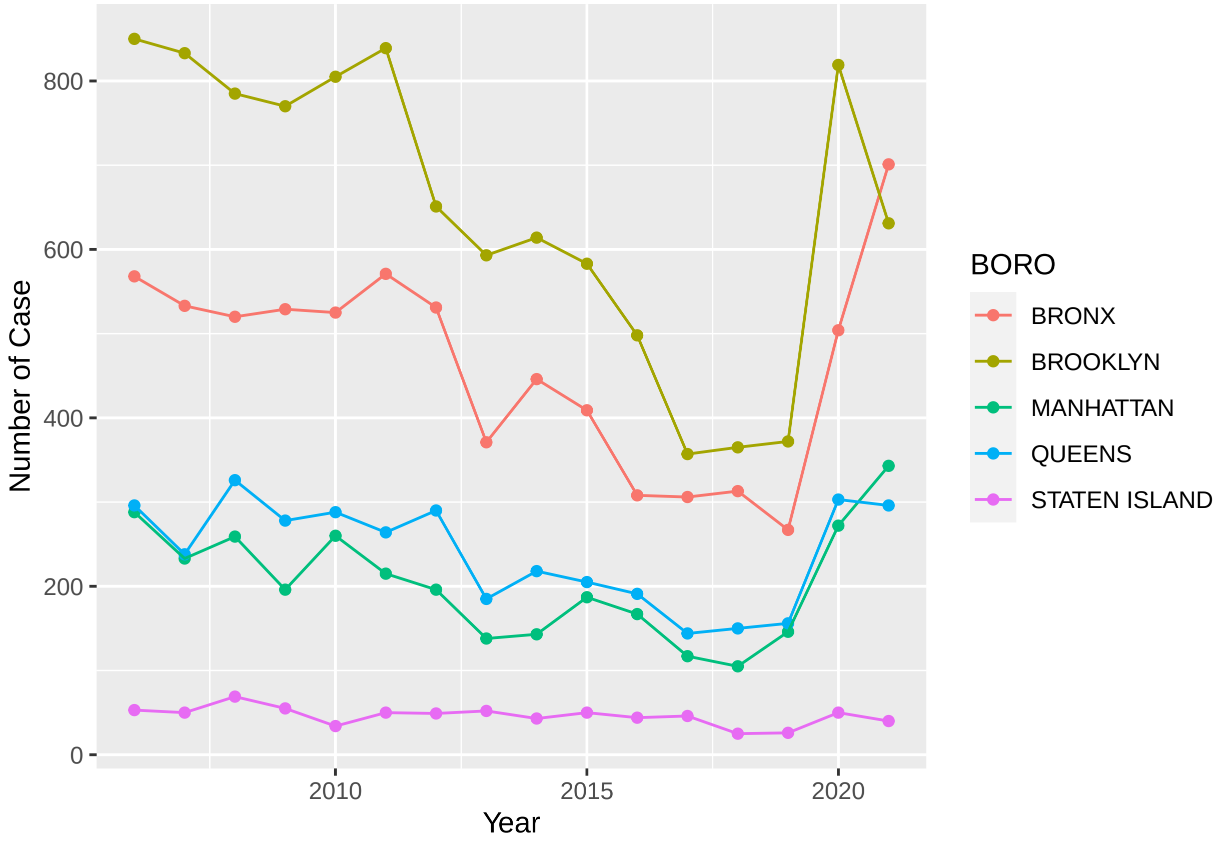


Plotting Graph Between Number of Cases and Year on Each BORO

```
NYPD = NYPD%>%
  mutate(OCCUR_YEAR = as.numeric(format(NYPD$OCCUR_DATE, '%Y')))


NYPDYear = NYPD%>%
  group_by(OCCUR_YEAR, BORO)%>%
  summarise(case = sum(case))
```

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

```
NYPDYear %>%
  ggplot(aes(x = OCCUR_YEAR, y = case)) +
  geom_point(aes(color = BORO)) +
  geom_line(aes(color = BORO))+
  labs(x = "Year", y = "Number of Case")
```



Step 4 Fit the model

In this step we are going to build a linear regression model our target variable is STATISTICAL_MURDER_FLAG which record if the shooting result in murder or not. We going to fit our model with the variable OCCUR_TIME, VIC_AGE_GROUP, VIC_SEX, VIC_RACE.

```
model1=lm(STATISTICAL_MURDER_FLAG~OCCUR_TIME+VIC_AGE_GROUP+ VIC_SEX+ VIC_RACE, data = NYPDShooting)

#view model summary
summary(model1)
```

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ OCCUR_TIME + VIC_AGE_GROUP +
##     VIC_SEX + VIC_RACE, data = NYPDShooting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3819 -0.2164 -0.1643 -0.1297  0.9619
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -4.202e-02  1.312e-01  -0.320   0.7487
## OCCUR_TIME                        3.939e-08  8.010e-08   0.492   0.6229
## VIC_AGE_GROUP18-24                3.448e-02  8.603e-03   4.008 6.14e-05
## VIC_AGE_GROUP25-44                8.767e-02  8.465e-03  10.356  < 2e-16
## VIC_AGE_GROUP45-64                1.123e-01  1.225e-02   9.169  < 2e-16
## VIC_AGE_GROUP65+                  1.778e-01  3.144e-02   5.657 1.56e-08
## VIC_AGE_GROUPUNKNOWN              1.313e-01  5.337e-02   2.459   0.0139
## VIC_SEXM                         -5.301e-03  8.483e-03  -0.625   0.5320
## VIC_SEXU                         -7.292e-02  1.241e-01  -0.588   0.5567
## VIC_RACEASIAN / PACIFIC ISLANDER  2.256e-01  1.324e-01   1.704   0.0885
## VIC_RACEBLACK                     1.740e-01  1.308e-01   1.330   0.1835
## VIC_RACEBLACK HISPANIC            1.480e-01  1.310e-01   1.129   0.2587
## VIC_RACEUNKNOWN                   8.263e-02  1.405e-01   0.588   0.5566
## VIC_RACEWHITE                     2.431e-01  1.317e-01   1.846   0.0649
## VIC_RACEWHITE HISPANIC            1.951e-01  1.309e-01   1.490   0.1363
##
## (Intercept)
## OCCUR_TIME
## VIC_AGE_GROUP18-24               ***
## VIC_AGE_GROUP25-44               ***
## VIC_AGE_GROUP45-64               ***
## VIC_AGE_GROUP65+                 ***
## VIC_AGE_GROUPUNKNOWN             *
## VIC_SEXM
## VIC_SEXU
## VIC_RACEASIAN / PACIFIC ISLANDER .
## VIC_RACEBLACK
## VIC_RACEBLACK HISPANIC
## VIC_RACEUNKNOWN
## VIC_RACEWHITE                    .
## VIC_RACEWHITE HISPANIC
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3923 on 25581 degrees of freedom
## Multiple R-squared:  0.01076,    Adjusted R-squared:  0.01022
## F-statistic: 19.87 on 14 and 25581 DF,  p-value: < 2.2e-16
```

Conclusion

Per the data visualization above, it seems like the race with the highest victim is black followed by white hispanic than black hispanic. There is more male as victims of shooting than female and the age groups whith more shooting victims are 18-24 and 25-44.

We can also rank the BORO from the highest number of shooting to the lowest number of shooting as follow:Brooklyn, Bronx, Queens, Manathan than Staten Island.

The months when the crime increase to the highest are between June and September, that lead us to the conclusion that there is a lot of shooting commited during summer and the law enforcement need to take proper measure to mitigate shooting especially during summer.

We also plot the crime count per year for each boro, Brooklyn, Queens and staten Island crime decrease few month after the begining of 2020 which likely correspond to the start of covid 19 pandemic in the US but Manathan and Bronx number of shooting increase sharply during the same period.

Bias sources

From this data, I was not able to see race, sex and age group for the perpertor of the crime, since this attributes had more than 30% of missing values. Drawing conclusions form this attributes could lead us to bias since we don't know which race, sex or age group of crime perpetor are missings and why this data has high rate of missing values on such important attributes. The race with the highest victims is black, it is bias to think that the back was target on the crime when we don't have information about the population rate of black people in New York.