NYPD Shooting Incident Data (Historic)

A. Bashar

2022-07-13

Introduction

The data we will be looking at is the NYPD Shooting Incident Data which lists every shooting incident that occurred in NYC from 2006 to 2021. We will import, clean, transform, visualize, analyze, and model the data.

Load Required Libraries

Install the packages of tidyverse and lubridate for this project.

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(lubridate)
```

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

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

Read in the Data

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

shooting_data <- read.csv(url_nypd)
```

Summary and Internal Structure of Data

```
summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class :character
##  Median : 86437258   Mode  :character   Mode  :character   Mode  :character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

```
str(shooting_data)
```

```
## 'data.frame':    25596 obs. of  19 variables:
##  $ INCIDENT_KEY           : int  24050482 77673979 226950018 237710987 224701998 225295736 231190175
##  $ OCCUR_DATE             : chr  "08/27/2006" "03/11/2011" "04/14/2021" "12/10/2021" ...
##  $ OCCUR_TIME             : chr  "05:35:00" "12:03:00" "21:08:00" "19:30:00" ...
##  $ BORO                   : chr  "BRONX" "QUEENS" "BRONX" "BRONX" ...
```

```
##  $ PRECINCT               : int  52 106 42 52 34 75 32 26 41 67 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 2 2 0 ...
##  $ LOCATION_DESC          : chr  "" "" "COMMERCIAL BLDG" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "true" "false" "true" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "65+" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "F" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  1017542 1027543 1009489 1017440 1005426 ...
##  $ Y_COORD_CD             : num  255919 186095 243050 256046 254690 ...
##  $ Latitude               : num  40.9 40.7 40.8 40.9 40.9 ...
##  $ Longitude              : num  -73.9 -73.8 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.87963173099996 40.86905819000003)" "POINT (-73.84392019
```

```
head(shooting_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME      BORO PRECINCT JURISDICTION_CODE
## 1     24050482 08/27/2006   05:35:00     BRONX       52                 0
## 2     77673979 03/11/2011   12:03:00    QUEENS      106                 0
## 3    226950018 04/14/2021   21:08:00     BRONX       42                 0
## 4    237710987 12/10/2021   19:30:00     BRONX       52                 0
## 5    224701998 02/22/2021   00:18:00 MANHATTAN       34                 0
## 6    225295736 03/07/2021   06:15:00  BROOKLYN       75                 0
##     LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1                                   true
## 2                                  false
## 3 COMMERCIAL BLDG                   true
## 4                                  false
## 5                                  false
## 6                                   true          25-44        M
##        PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD
## 1                      25-44       F BLACK HISPANIC    1017542   255918.9
## 2                        65+       M          WHITE    1027543   186095.0
## 3                      18-24       M          BLACK    1009489   243050.0
## 4                      25-44       M          BLACK    1017440   256046.0
## 5                      25-44       M BLACK HISPANIC    1005426   254690.0
## 6 BLACK HISPANIC       25-44       M WHITE HISPANIC    1020492   187865.0
##   Latitude Longitude                                      Lon_Lat
## 1 40.86906 -73.87963  POINT (-73.87963173099996 40.86905819000003)
## 2 40.67737 -73.84392 POINT (-73.84392019199998 40.677366895000034)
## 3 40.83376 -73.90880  POINT (-73.90879517699994 40.83376365400005)
## 4 40.86941 -73.88000  POINT (-73.87999831299999 40.86940749200004)
## 5 40.86572 -73.92344  POINT (-73.92344088699997 40.86572268100008)
## 6 40.68226 -73.86933  POINT (-73.86933111399996 40.68225681500007)
```

There are 19 categories in our data. By looking at the head of our data, we can already see that the perpetrator age group, perpetrator sex, and perpetrator race for example have empty data points. If these sections are used, then we will have to make a note of it.

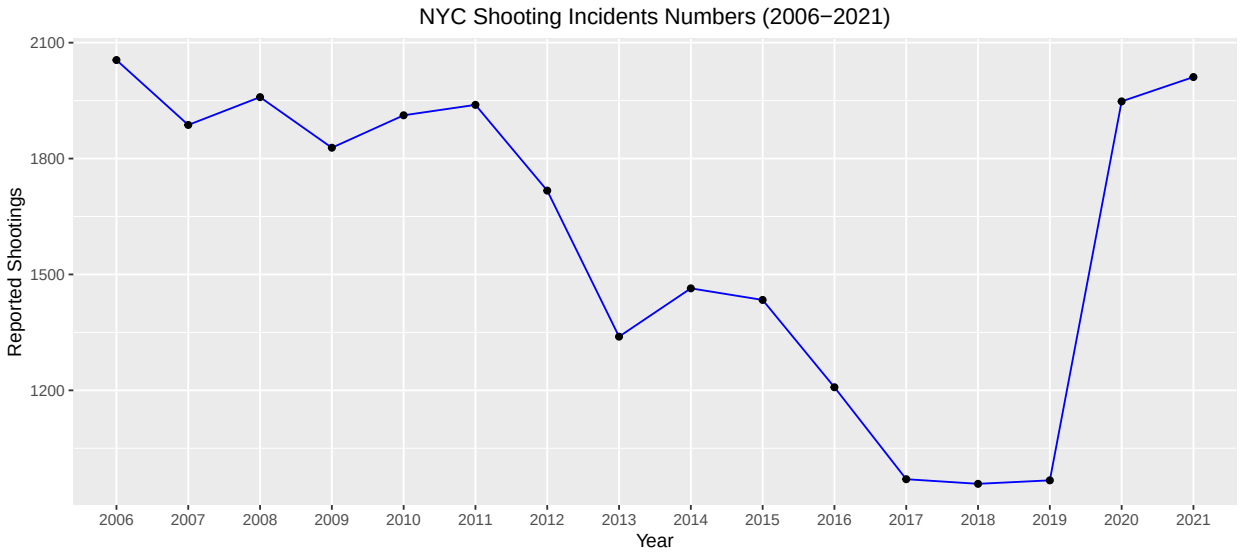
Custom Size of Future Plots

```
knitr::opts_chunk$set(fig.width=10)
```

This will make sure that our future graphs will be an appropriate size to be viewed.

Total Number of Reported Shootings in New York City by Year (From 2006 to 2021)

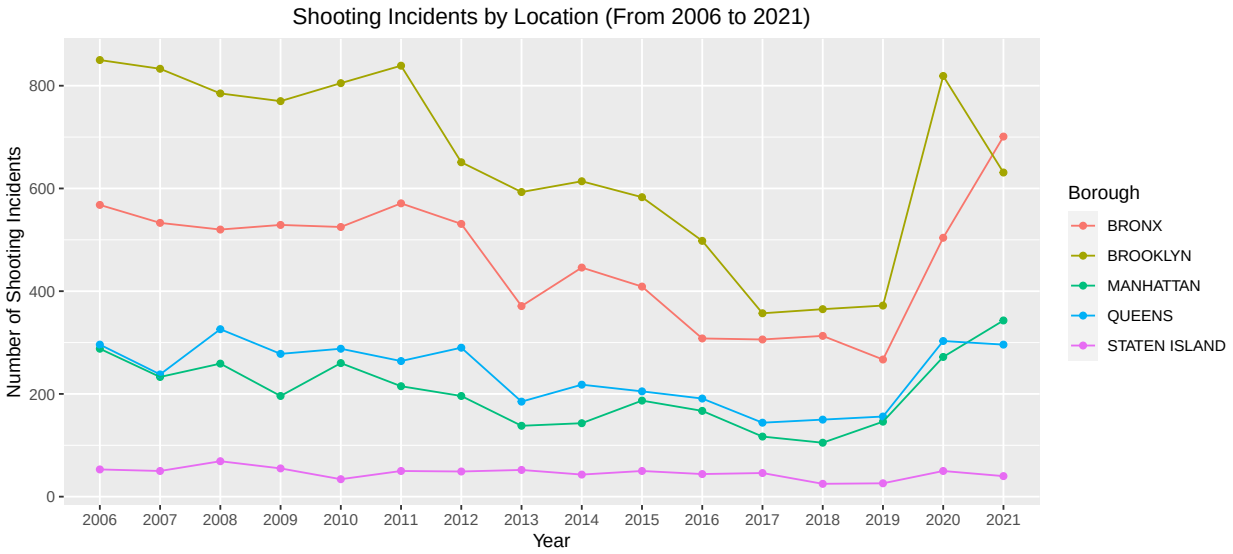
```
shooting_data %>%
    select(c(1,2)) %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    mutate(occur_year = format(as.Date(.$OCCUR_DATE),format ="%Y"))%>%
    group_by(occur_year)%>%

 summarise(n=n()) %>%
 ggplot(aes(x = occur_year,y=n))+
 geom_line(group=1, color="blue")+
 geom_point()+
 theme(plot.title = element_text(hjust = 0.5))+
 ggtitle("NYC Shooting Incidents Numbers (2006-2021)") +
 labs(y="Reported Shootings", x="Year")
```



Overall shooting incidents had been trending downwards since 2006 and hitting a low in 2018. The shootings for 2020 and 2021 then jumped to some of the highest levels. Let us adjust the graph to see how this same trend would look like when separated by each borough in New York.

Number of Shooting Incidents by Location each Year (from 2006 to 2021)

```
shooting_data %>%
  select(c(1,2,4,10,17,18)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(occur_year = format(as.Date(.$OCCUR_DATE),format="%Y")) %>%
  group_by(occur_year,BORO) %>%
  summarise(n=n()) %>%

  ggplot(aes(x=occur_year, y=n, group=BORO)) +
  geom_point(aes(color=BORO)) +
  geom_line(aes(color=BORO)) +
  labs(x = "Year", y = "Number of Shooting Incidents", color = "Borough") +
  ggtitle("Shooting Incidents by Location (From 2006 to 2021)") +
  theme(plot.title = element_text(hjust = 0.5))
```

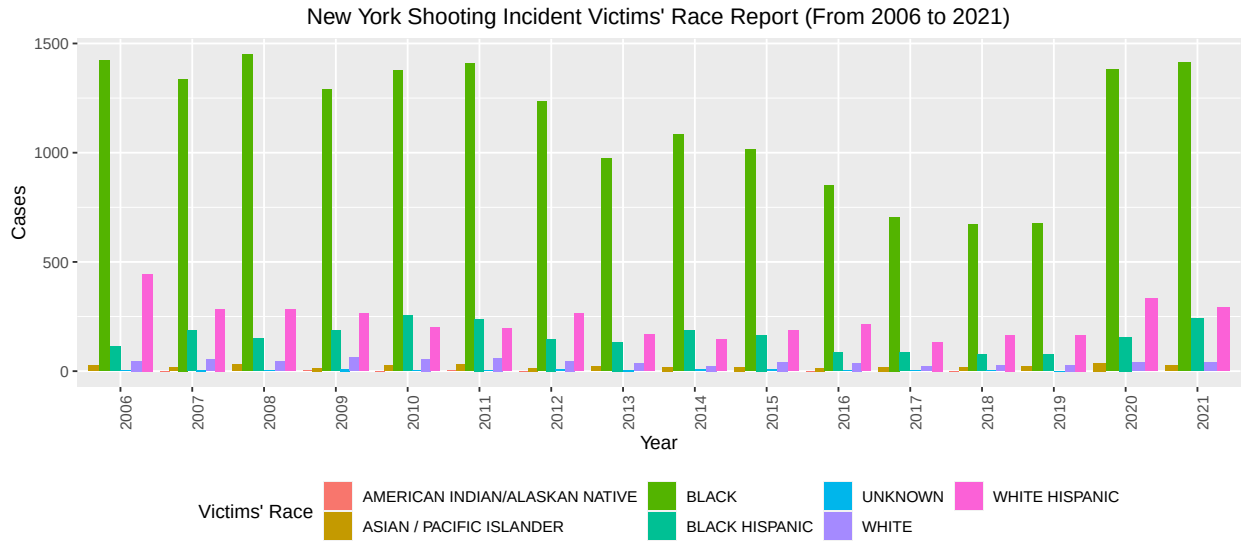


Brooklyn had generally been the borough in New York City with the highest number of shooting incidents until 2021 where Bronx took over. Staten Island had consistently the lowest number of shooting incidents and stayed in a relatively close range from 2006 to 2021. Queens had generally a slightly higher amount of shooting incidents than Manhattan except for 2021 where Manhattan had taken over. Let us check to see for the victims' race over the same timeline.

Victims' Race Over the Years (From 2006 to 2021)

```
shooting_data %>%
  select(c(1,2,14)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(occur_year = format(as.Date(.$OCCUR_DATE),format="%Y")) %>%
  group_by(occur_year,VIC_RACE) %>%
  summarise(n=n()) %>%

ggplot(aes(x = occur_year, y = n, fill = VIC_RACE), color = VIC_RACE) +
  geom_bar(stat = 'identity', position = 'dodge') +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90),
                          plot.title = element_text(hjust = 0.5)) +
```

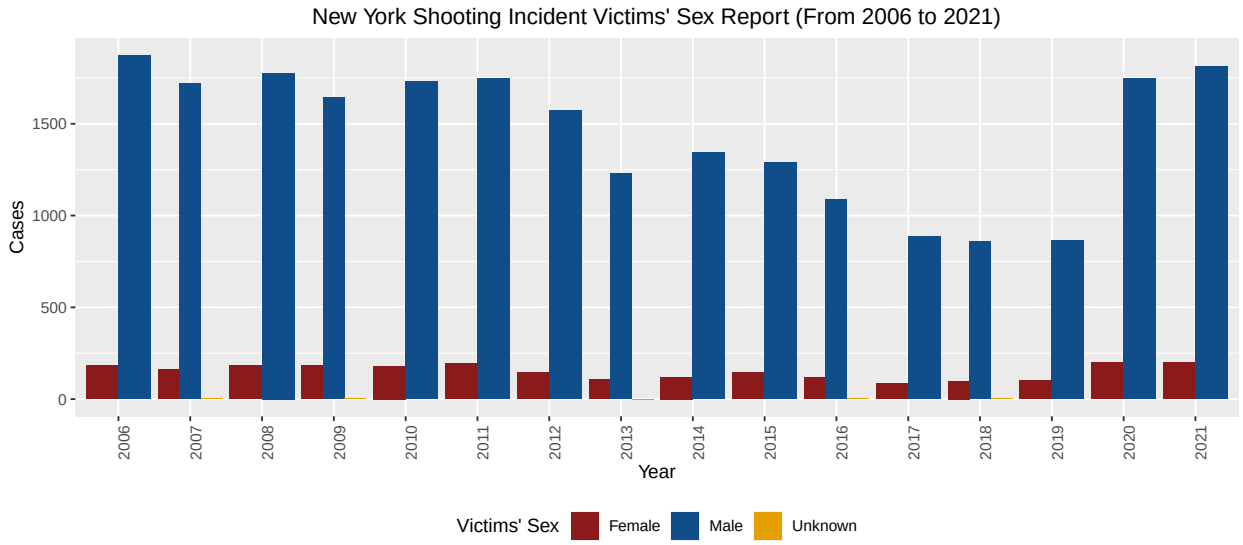
```
labs(title = "New York Shooting Incident Victims' Race Report (From 2006 to 2021)",
                        y = "Cases", x = "Year", fill = "Victims' Race")
```



From our visualization we can see that black individuals were most often the victim of shooting incidents in New York City from 2006 to 2021. There is a large drop off to the next group which has usually been White Hispanics.

Victims' Sex Over the Years (From 2006 to 2021)

```
shooting_data %>%
    select(c(1,2,13)) %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    mutate(occur_year = format(as.Date(.$OCCUR_DATE),format="%Y")) %>%
    group_by(occur_year,VIC_SEX) %>%
    summarise(n=n()) %>%

    ggplot(aes(x = occur_year, y = n, fill = VIC_SEX), color = VIC_SEX) +
    scale_fill_manual(name="Victims' Sex",labels=c("Female", "Male", "Unknown"),
                      values=c("firebrick4", "dodgerblue4", "#E69F00")) +
    geom_bar(stat = 'identity', position = 'dodge') +
    theme(legend.position = "bottom", axis.text.x = element_text(angle = 90),
          plot.title = element_text(hjust = 0.5)) +
    labs(title = "New York Shooting Incident Victims' Sex Report (From 2006 to 2021)",
         y = "Cases", x = "Year")
```



Male victims are much more common over female victims through 2006 to 2021. Let us see if female victims are correlated with an increase in male victims.

Modelling of our Data for Correlation between Male Victims and Female Victims

```
knitr::opts_chunk$set(fig.width=7)

victim_sex <- shooting_data %>%
     group_by(OCCUR_DATE, VIC_SEX) %>%
     summarise(n = n()) %>%
     spread(key = VIC_SEX, value = n) %>%
     mutate(Female = replace_na(F, 0),
            Male = replace_na(M, 0)) %>%
     ungroup()
```

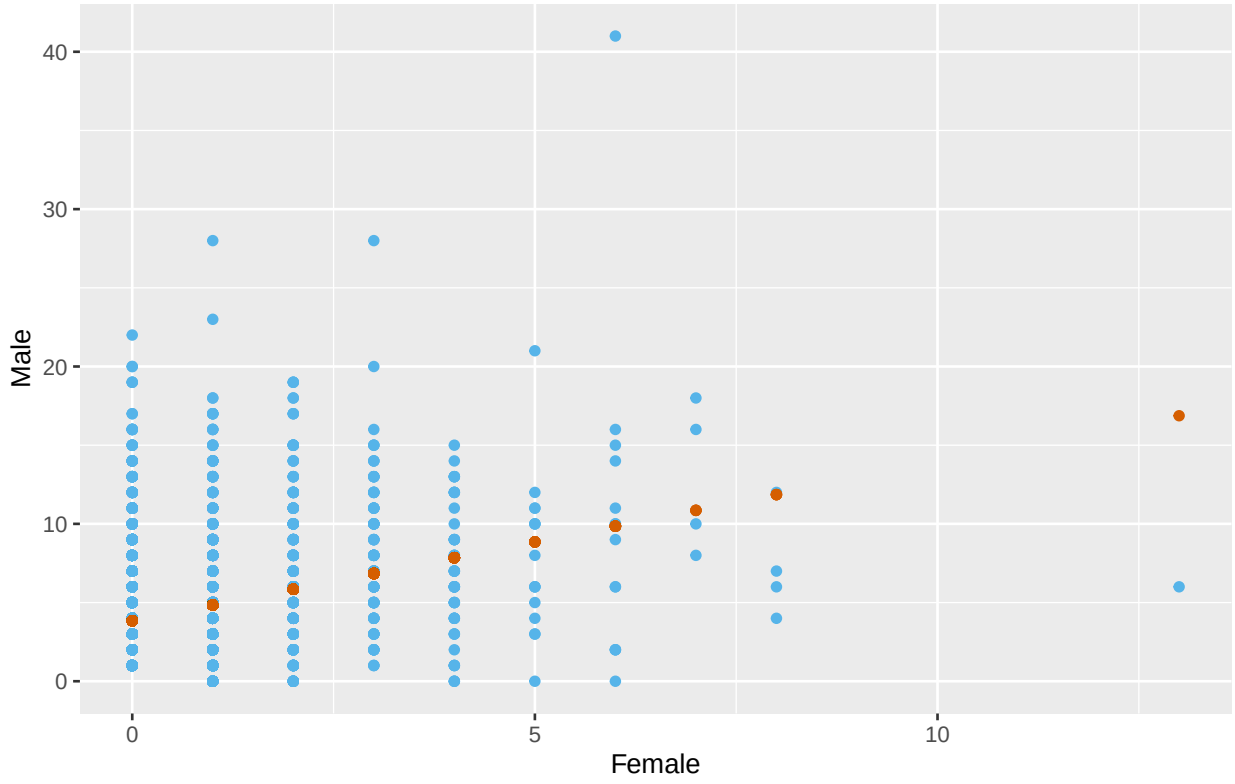
```
model <- lm(Male ~ Female, data = victim_sex)

summary(model)
```

```
##
## Call:
## lm(formula = Male ~ Female, data = victim_sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8719  -1.8429  -0.8405   1.1595  31.1450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.84049    0.04669   82.26   <2e-16 ***
## Female       1.00242    0.04629   21.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.083 on 5407 degrees of freedom
## Multiple R-squared:  0.07982,    Adjusted R-squared:  0.07965
## F-statistic:   469 on 1 and 5407 DF,  p-value: < 2.2e-16
```

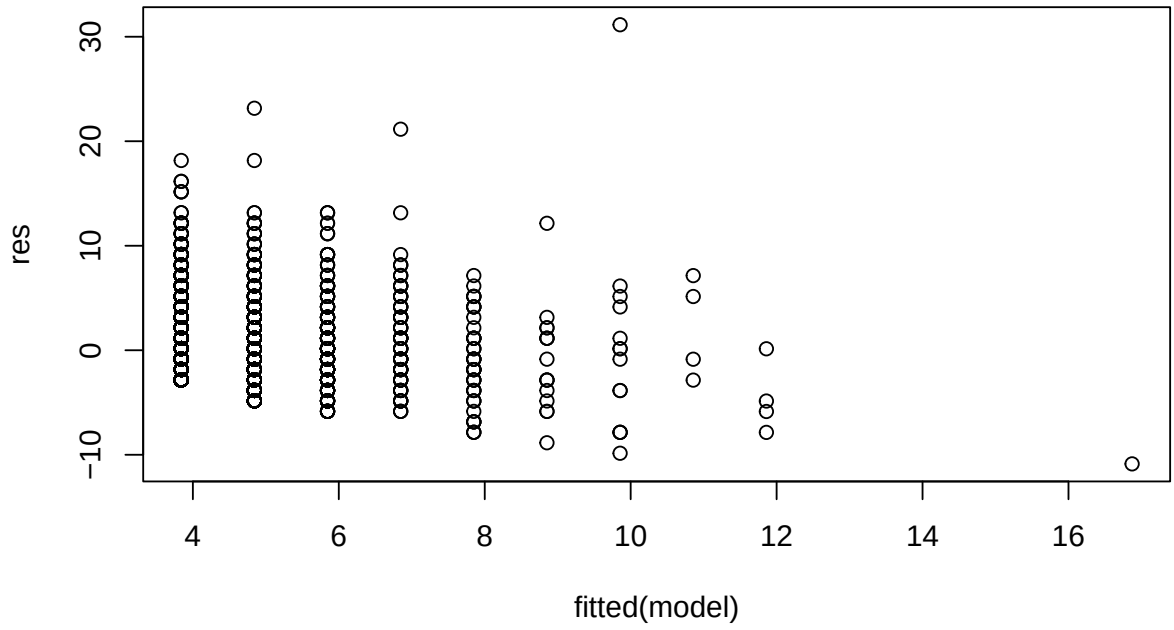
The Multiple R-squared is very small at 8% which means our model has a low chance for the explanatory variables to predict the value of the response variable. The p_value is a very small number close to 0, but that enough will not help. Let us turn our prediction into a plot.

```
victim_sex %>%
  mutate(pred = predict(model)) %>%
  ggplot() +
  geom_point(aes(x = Female, y = Male), color = '#56B4E9') +
  geom_point(aes(x = Female, y = pred), color = '#D55E00')
```



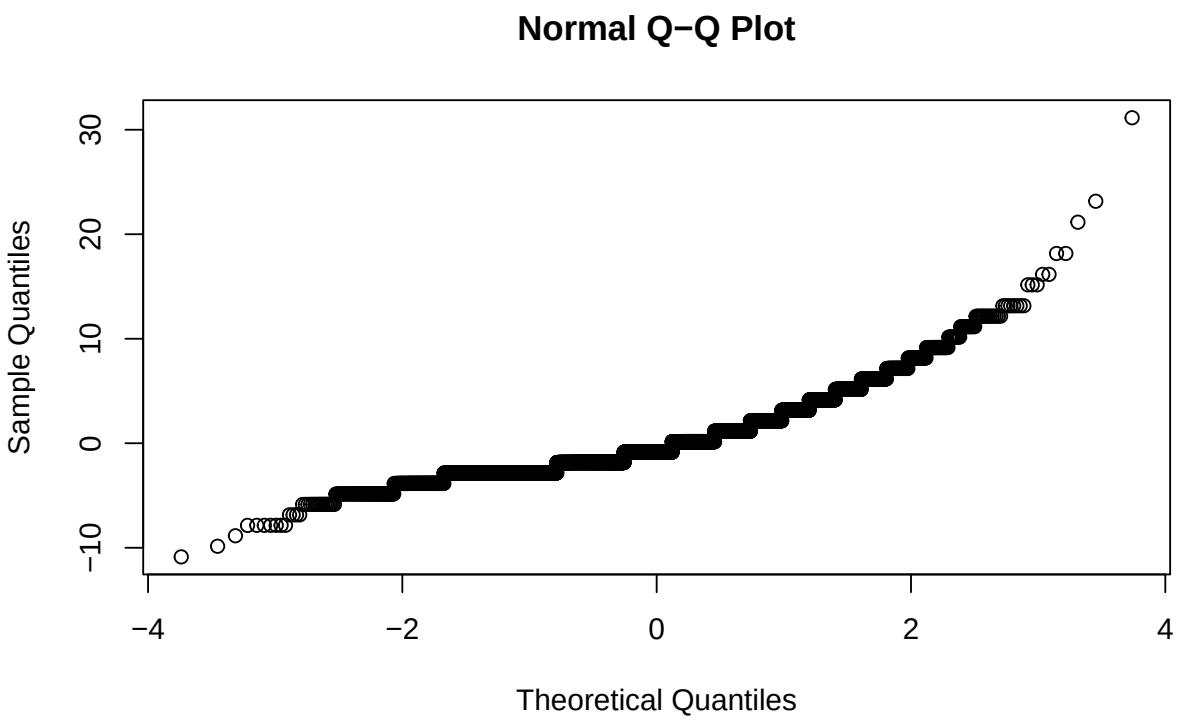
The orange points is the prediction line and the light blue points is our data. There does not look to be strong correlation between female and male victims. Let us look further into the data.

```
res <- resid(model)

plot(fitted(model), res)
```



This residual plot is useful for testing homoscedasticity. Since most of our points are not evenly distributed around the value zero, then we can assume that homoscedastricity has been violated.

```
qqnorm(res)
```



Our model displays skewness which can indicate a sample from a population and may not have a relation between the values. There are more extreme values than would be expected from a normal distribution.

Bias in the Data

- One potential bias is that this is only the reported shooting incidents and there is no estimate on non-reported incidents.
- Our graphs have shown that black men in particular are most commonly the victims of shooting incidents and that certain boroughs will have more reported shootings. Due to this, the NYPD may patrol areas with larger populations of black residents that may show bias in the data.

- Another bias in my own analysis may have been with sample sizes. There is a much larger amount of male victims than there are female victims and that would make correlations with the two to be difficult. There has to be other reasons as to why males are largely the victims in NYC shooting incidents.

Conclusion

We looked at the data from every shooting incident that occurred in NYC from 2006 to 2021. We were able to create a few visualizations that gave us a glimpse into the trends seen in the shooting incidents. Our model was unable to find a strong correlation between male and female shooting victims.

Data link: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic