NYPD Shooting Report

Harishawn Ramrup

2023-03-01

R Markdown

Importing Necessary Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- 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
```

```
library(dplyr)
library(ggplot2)
library(nnet)
# Not Libraries used in class
library(osmdata)
```

```
## Data (c) OpenStreetMap contributors, ODbL 1.0. https://www.openstreetmap.org/copyright
```

```
#library(ggmap)
```

Note I had errors knitting -> library(ggmap) I left the code outside of a R chunk for your references

Data Exploration

Importing police data set gathered from Data.Gov. The dataset goes back to 2006 up to 2021.

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

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
glimpse(police_df)
```

```
## Rows: 25,596
## Columns: 19
## $ INCIDENT_KEY            <dbl> 236168668, 231008085, 230717903, 237712309, 22~
## $ OCCUR_DATE              <chr> "11/11/2021", "07/16/2021", "07/11/2021", "12/~
## $ OCCUR_TIME              <time> 15:04:00, 22:05:00, 01:09:00, 13:42:00, 20:00~
## $ BORO                    <chr> "BROOKLYN", "BROOKLYN", "BROOKLYN", "BROOKLYN"~
## $ PRECINCT                <dbl> 79, 72, 79, 81, 113, 113, 42, 52, 34, 75, 32, ~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, "COMMERCIAL BLDG", NA,~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE,~
## $ PERP_AGE_GROUP          <chr> NA, "45-64", "<18", NA, NA, NA, NA, NA, NA, "2~
## $ PERP_SEX                <chr> NA, "M", "M", NA, NA, NA, NA, NA, NA, "M", "M"~
## $ PERP_RACE               <chr> NA, "ASIAN / PACIFIC ISLANDER", "BLACK", NA, N~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "25-44", "25-44", "~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "ASIAN / PACIFIC ISLANDER", "BLACK", ~
## $ X_COORD_CD              <dbl> 996313, 981845, 996546, 1001139, 1050710, 1051~
## $ Y_COORD_CD              <dbl> 187499, 171118, 187436, 192775, 184826, 196646~
## $ Latitude                <dbl> 40.68132, 40.63636, 40.68114, 40.69579, 40.673~
## $ Longitude               <dbl> -73.95651, -74.00867, -73.95567, -73.93910, -7~
## $ Lon_Lat                 <chr> "POINT (-73.95650899099996 40.68131820000008)"~
```

Summarize Data to see where I would need to clean up

```
summary(police_df)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  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
##
```

ETL

ETL Date formatting and remove unneeded columns from the dataframe

```r
#Converting string elements into Date and Time elements
police_df$OCCUR_DATE <- mdy(police_df$OCCUR_DATE)

# dropping unneeded columns
police_df<- subset(police_df, select= -c(X_COORD_CD, Y_COORD_CD))
```

Checking Dataset after ETL to ensure desired results

```r
summary(police_df)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596       Length:25596
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms         Class :character
##  Median : 86437258   Median :2012-08-26   Class2:difftime   Mode  :character
##  Mean   :112382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238490103   Max.   :2021-12-31
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  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            Latitude       Longitude
##  Length:25596       Length:25596       Min.   :40.51   Min.   :-74.25
##  Class :character   Class :character   1st Qu.:40.67   1st Qu.:-73.94
##  Mode  :character   Mode  :character   Median :40.70   Median :-73.92
##                                        Mean   :40.74   Mean   :-73.91
##                                        3rd Qu.:40.82   3rd Qu.:-73.88
##                                        Max.   :40.91   Max.   :-73.70
##
##    Lon_Lat
##  Length:25596
##  Class :character
##  Mode  :character
##
##
##
##
```

Groupings / Aggregations

Groupings

```
nypd_incident_by_date <- police_df %>%
    group_by(month=month(OCCUR_DATE), year=year(OCCUR_DATE)) %>%
    summarize(count = n())
```

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

Create Grouping for Graphs by Month

```
nypd_incident_by_month <- police_df %>%
group_by(month=month(OCCUR_DATE)) %>%
summarize(count = n())
```

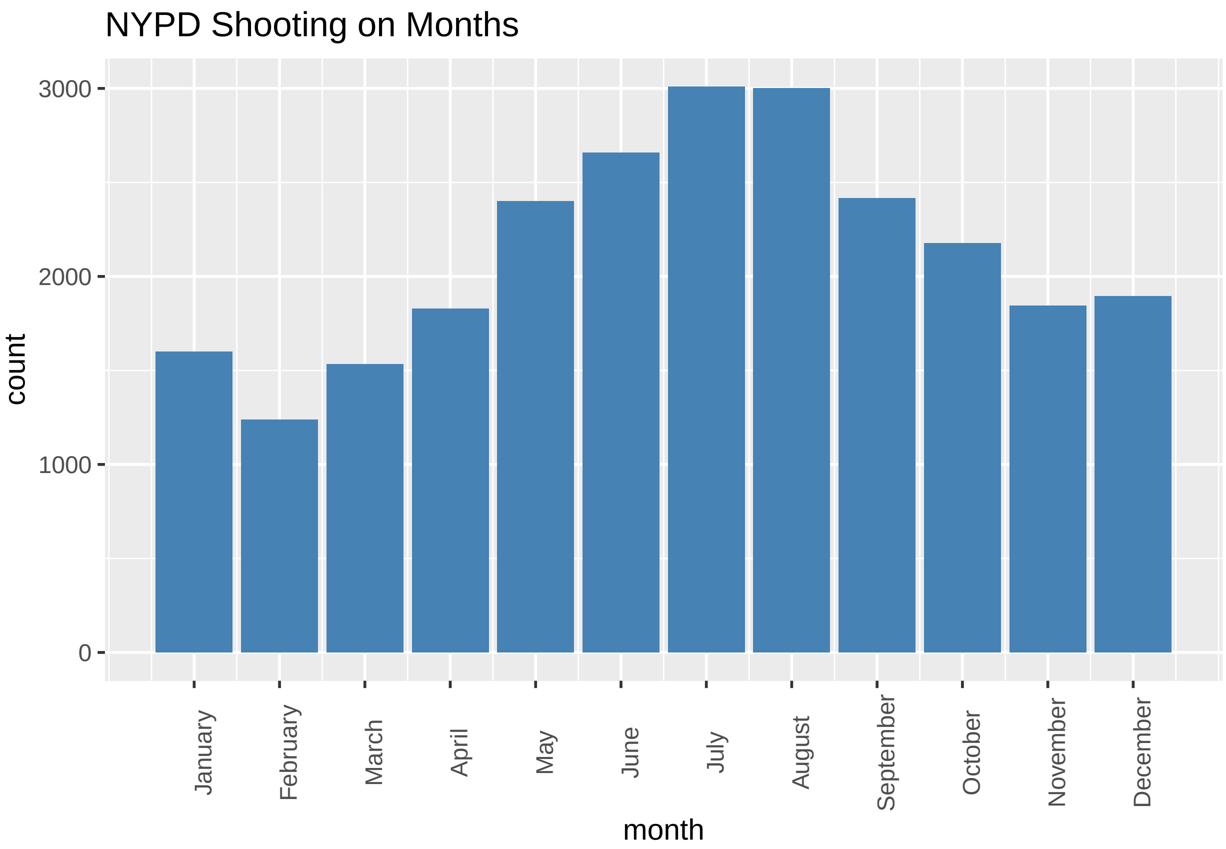
Create Grouping for Logistic Model

```
# Create data group for Model
murdered_data <- police_df %>%
  select(VIC_SEX, STATISTICAL_MURDER_FLAG, BORO, PERP_RACE) %>%
  filter(!is.na(VIC_SEX), !is.na(STATISTICAL_MURDER_FLAG))
```

Graphs

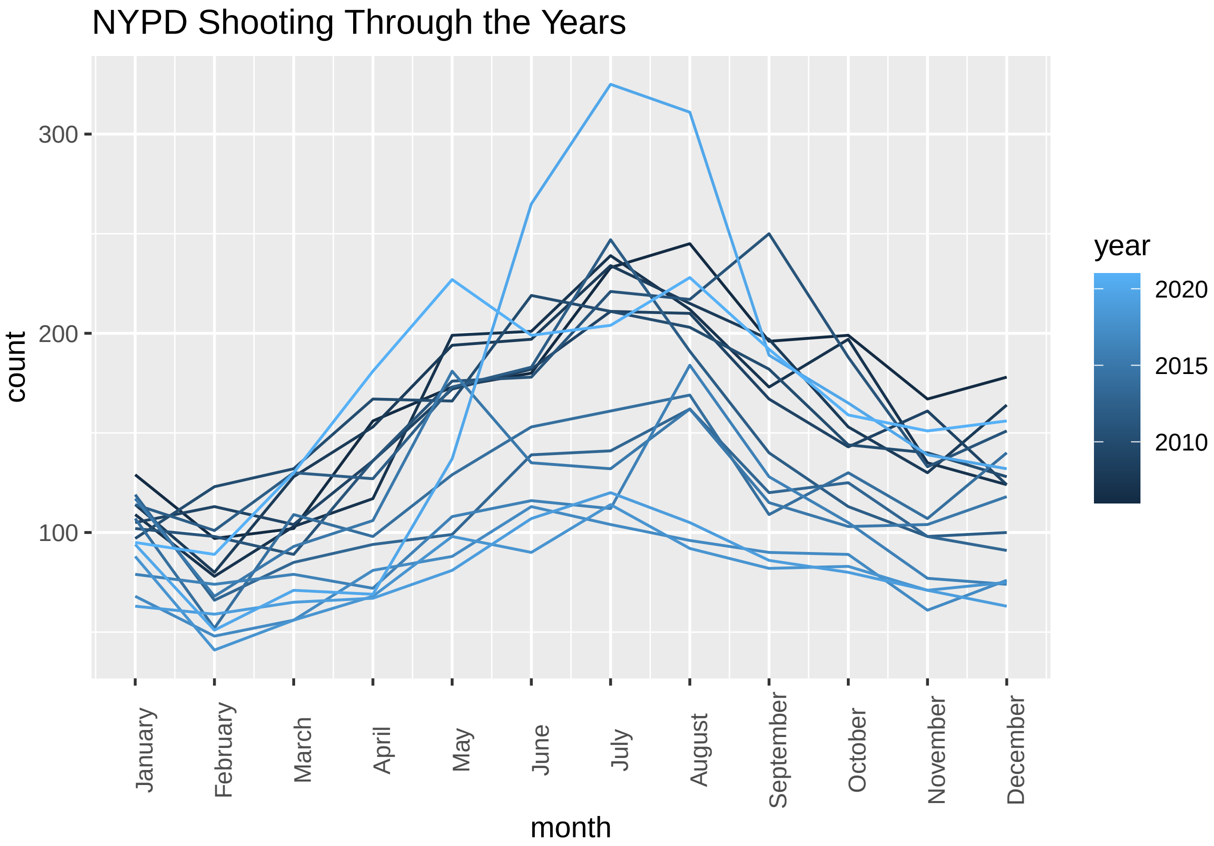
Graph shootings by month

```
ggplot(nypd_incident_by_month, aes(x=month, y = count)) +
geom_bar(stat='identity', fill = "steelblue") +
scale_x_continuous(breaks=1:12, labels = month.name) +
labs( title = "NYPD Shooting on Months") +
theme(axis.text.x = element_text(angle = 90))
```



Graph shootings through the years

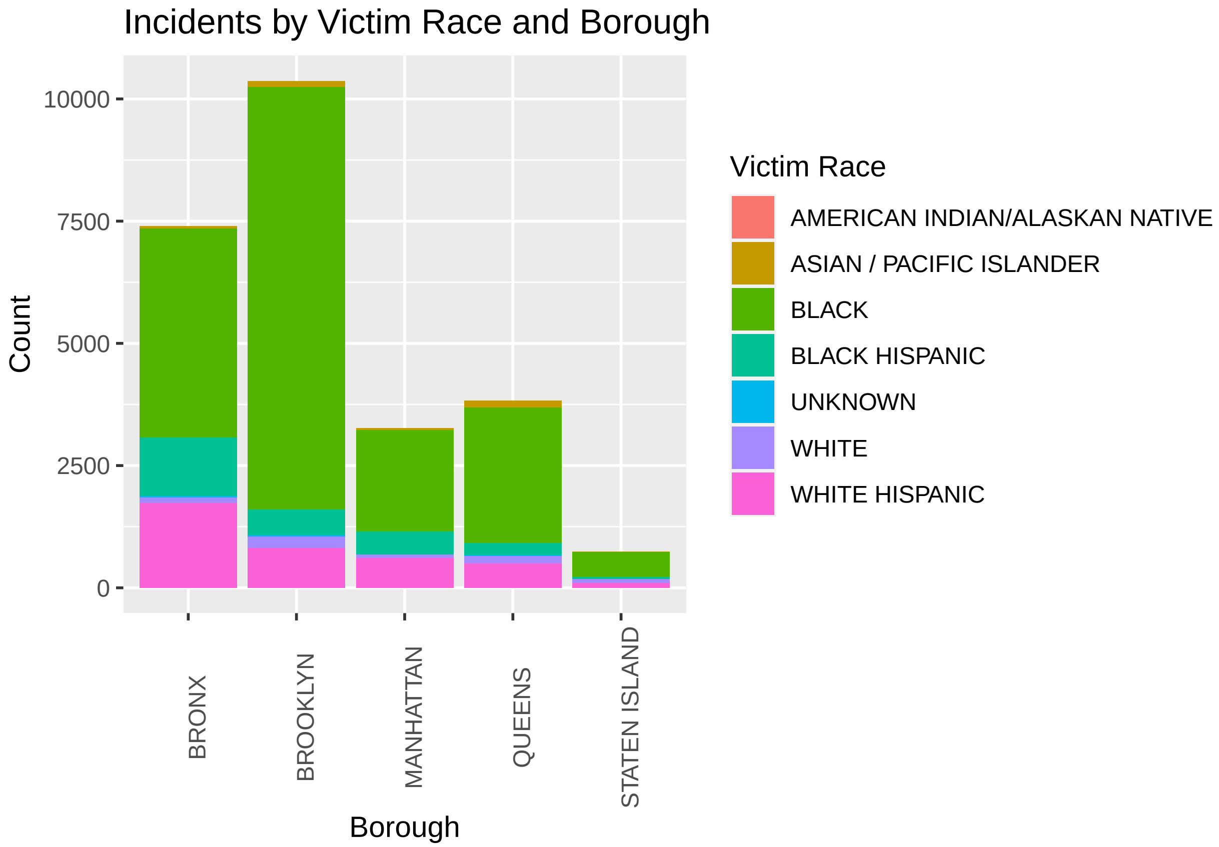
```r
ggplot(nypd_incident_by_date, aes(x=month, y = count, group = year, color=year)) +
    geom_line() +
    scale_x_continuous(breaks=1:12, labels = month.name) +
    labs( title = "NYPD Shooting Through the Years") +
    theme(axis.text.x = element_text(angle = 90))
```



From the graph above there is a significant increase in shootings in the Summer months 2020 and beyond. The levels pre 2010 shootings incidents where also higher than shootings that occured post 2010 with the exception of 2020. In the months of June, July, and August you can identify a massive uptick in shootings that are reported. In the winter months the averages are lower with Feburary having the least reported number of incidents in respect to other months

Here we look at the victim race per Borough

```r
ggplot(data = police_df, aes(x = BORO, fill = VIC_RACE)) +
  geom_bar() +
  labs(title = "Incidents by Victim Race and Borough", x = "Borough", y = "Count", fill = "Victim Race"
  theme(axis.text.x = element_text(angle = 90))
```



From this graph we can see Brooklyn has the highest number of Victims who were Black comparatively to the other boroughs while also having the most number of incidents overall.

Map of NY City and shootings

Graph shootings in relation to where they occurred on a map of the City

```
library(ggmap)

# Get a map of New York City using ggmap
nyc_map <- get_map(getbb("New York City"), source= ’stamen’)

# Plot the shootings on the map using geom_point
ggmap(nyc_map) +
  geom_point(data = police_df, aes(x = Longitude, y = Latitude), alpha = 0.5, color = "red")
```

Model

This model uses the Murder Flag as dependent and determines if the sex or the victim and borough can model if the shooting resulted in Murder

```
# Build the logistic regression model
model <- lm(STATISTICAL_MURDER_FLAG ~ VIC_SEX + BORO + PERP_RACE, data=murdered_data)
```

```
# Print the summary of the model
summary(model)
```

```
## 
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ VIC_SEX + BORO + PERP_RACE, 
##     data = murdered_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.43200 -0.22064 -0.19906 -0.05706  0.94537 
## 
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)   
## (Intercept)                        0.031209   0.279722   0.112  0.91116   
## VIC_SEXM                          -0.022715   0.010119  -2.245  0.02480 * 
## VIC_SEXU                          -0.081116   0.140100  -0.579  0.56261   
## BOROBROOKLYN                      -0.021579   0.007784  -2.772  0.00558 **
## BOROMANHATTAN                     -0.024014   0.010094  -2.379  0.01737 * 
## BOROQUEENS                        -0.016988   0.009849  -1.725  0.08459 . 
## BOROSTATEN ISLAND                 -0.017363   0.017340  -1.001  0.31669   
## PERP_RACEASIAN / PACIFIC ISLANDER  0.314980   0.281484   1.119  0.26316   
## PERP_RACEBLACK                     0.212147   0.279551   0.759  0.44793   
## PERP_RACEBLACK HISPANIC            0.191944   0.279754   0.686  0.49265   
## PERP_RACEUNKNOWN                   0.070148   0.279674   0.251  0.80196   
## PERP_RACEWHITE                     0.400794   0.280559   1.429  0.15315   
## PERP_RACEWHITE HISPANIC            0.241854   0.279644   0.865  0.38713   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3953 on 16273 degrees of freedom
##   (9310 observations deleted due to missingness)
## Multiple R-squared:  0.0202, Adjusted R-squared:  0.01948 
## F-statistic: 27.96 on 12 and 16273 DF,  p-value: < 2.2e-16
```

Results of Model

Overall, this model suggests that there are some significant relationships between VIC_SEX, BORO, PERP_RACE and STATISTICAL_MURDER_FLAG, but the R-squared value is very low, which suggests that the model explains only a small amount of the variation in the dependent variable (STATISTICAL_MURDER_FLAG). Including additional features can narrow down this relationship and possibly improve the R-Squared Value.

Biases

The biases that could be visible in this data set is a lack of reporting of actual. With the increase of public outrage in the excess forces that some police departments have been conducting in and with a growing social distress, departments may be more inclined to not reporting actual shootings or to try to improve public image and may lead to under reporting of shootings. In 2021, around July and August you can see an increase in shootings which is what sparked my interest in looking to if race had an effect in that. This also plays into my own bias, as I had a pre-disposition on the topic because of current events and social media.