NYPD Shooting Incident Data Report

2024-03-12

Introduction

In this project we aim to explore the NYPD Shooting Incident Data. According to data.gov, this is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. We will explore this data for patterns such as fatality, comparison of cases in boroughs, trends over time, victims' profile.

Data

Data is available at City of New York's website: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

More information about the data can be found at: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Loading the Data

First, we need to load the necessary libraries.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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)
library(ggplot2)
```

We load the NYPD Shooting Incident Data, take a look at the first few rows and examine the summary of the data.

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read.csv(url_in)
head(nypd_data, 2)
```

```
##   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
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                    false
## 2                 0                                                    false
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE X_COORD_CD
## 1                                           18-24       M    BLACK    1058925
## 2                                           18-24       M    BLACK    1005028
##   Y_COORD_CD Latitude Longitude                                       Lon_Lat
## 1     180924 40.66296 -73.73084 POINT (-73.73083868899994 40.662964620000025)
## 2     234516 40.81035 -73.92494  POINT (-73.92494232599995 40.81035186300006)
```

```
summary(nypd_data)
```

```
##   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.:1000028   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
```

Tidying and Transforming the Data

From the summary and head of the data we can see that there are several variables that we are not going to use for our analysis. We start by removing variables INCIDENT_KEY, LOC_OF_OCCUR_DESC, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, LOC_CLASSFCTN_DESC, LOCATION_DESC, Latitude, Longitude, Lon_Lat.

```
nypd_data <- select(nypd_data, -c(INCIDENT_KEY, LOC_OF_OCCUR_DESC, JURISDICTION_CODE,
                                  X_COORD_CD, Y_COORD_CD, LOC_CLASSFCTN_DESC, LOCATION_DESC,
                                  Latitude, Longitude, Lon_Lat))
```

We can see in the summary that OCCUR_DATE and OCCUR_TIME are of character type, we convert them to date and time types respectively.

```
nypd_data <- mutate(nypd_data, OCCUR_DATE = mdy(OCCUR_DATE))
nypd_data <- mutate(nypd_data, OCCUR_TIME = hms(OCCUR_TIME))
```

We will also convert STATISTICAL_MURDER_FLAG, VIC_SEX, VIC_RACE, VIC_AGE_GROUP, BORO to factor variables.

```
nypd_data <- mutate(nypd_data, STATISTICAL_MURDER_FLAG = as.factor(STATISTICAL_MURDER_FLAG))
nypd_data <- mutate(nypd_data, VIC_SEX = as.factor(VIC_SEX))
nypd_data <- mutate(nypd_data, VIC_RACE = as.factor(VIC_RACE))
nypd_data <- mutate(nypd_data, VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP))
nypd_data <- mutate(nypd_data, BORO = as.factor(BORO))
```

Next, we summarize to see where we are at in terms of cleaning the data.

```
summary(nypd_data)
```

```
##    OCCUR_DATE           OCCUR_TIME                                   BORO
##  Min.   :2006-01-01   Min.   :0S                            BRONX        : 7937
##  1st Qu.:2009-07-18   1st Qu.:3H 27M 0S                     BROOKLYN     :10933
##  Median :2013-04-29   Median :15H 11M 0S                    MANHATTAN    : 3572
##  Mean   :2014-01-06   Mean   :12H 41M 31.7091388399567S   QUEENS       : 4094
##  3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S                    STATEN ISLAND:  776
##  Max.   :2022-12-31   Max.   :23H 59M 0S
##
##     PRECINCT      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Min.   :  1.00   false:22046             Length:27312       Length:27312
##  1st Qu.: 44.00   true : 5266             Class :character   Class :character
##  Median : 68.00                           Mode  :character   Mode  :character
##  Mean   : 65.64
##  3rd Qu.: 81.00
##  Max.   :123.00
##
##   PERP_RACE         VIC_AGE_GROUP   VIC_SEX
##  Length:27312       <18    : 2839   F: 2615
##  Class :character   1022   :    1   M:24686
##  Mode  :character   18-24  :10086   U:   11
##                     25-44  :12281
##                     45-64  : 1863
##                     65+    :  181
##                     UNKNOWN:   61
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19439
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4049
```

We notice that there is a value of 1022 in VIC_AGE_GROUP, that must be a typo, so we change that to "UNKNOWN" and drop that factor level.

```
nypd_data$VIC_AGE_GROUP <- replace(nypd_data$VIC_AGE_GROUP,
                                   nypd_data$VIC_AGE_GROUP==1022, "UNKNOWN")
nypd_data$VIC_AGE_GROUP <- droplevels(nypd_data$VIC_AGE_GROUP)
```

Next, we clean up PERP_SEX, PERP_AGE_GROUP and PERP_RACE variables and factorize them. (null), empty variables are changed to "UNKNOWN" or "U".

```
nypd_data$PERP_SEX <- replace(nypd_data$PERP_SEX, nypd_data$PERP_SEX == "(null)"|
                                    nypd_data$PERP_SEX == "", "U")
nypd_data$PERP_AGE_GROUP <- replace(nypd_data$PERP_AGE_GROUP,
                                    nypd_data$PERP_AGE_GROUP == "(null)" |
                                          nypd_data$PERP_AGE_GROUP == "", "UNKNOWN")
nypd_data <- mutate(nypd_data, PERP_SEX = as.factor(PERP_SEX))
nypd_data <- mutate(nypd_data, PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP))
nypd_data$PERP_RACE <- replace(nypd_data$PERP_RACE, nypd_data$PERP_RACE == "(null)" |
                                     nypd_data$PERP_RACE == "", "UNKNOWN")
nypd_data <- mutate(nypd_data, PERP_RACE = as.factor(PERP_RACE))
```

Let's check the data again:

```
summary(nypd_data)
```

```
##    OCCUR_DATE           OCCUR_TIME                                   BORO
##  Min.   :2006-01-01   Min.   :0S                          BRONX        : 7937
##  1st Qu.:2009-07-18   1st Qu.:3H 27M 0S                   BROOKLYN     :10933
##  Median :2013-04-29   Median :15H 11M 0S                  MANHATTAN    : 3572
##  Mean   :2014-01-06   Mean   :12H 41M 31.7091388399567S   QUEENS       : 4094
##  3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S                  STATEN ISLAND:  776
##  Max.   :2022-12-31   Max.   :23H 59M 0S
##
##     PRECINCT      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP  PERP_SEX
##  Min.   :  1.00   false:22046             UNKNOWN:13132   F:  424
##  1st Qu.: 44.00   true : 5266             18-24  : 6222   M:15439
##  Median : 68.00                           25-44  : 5687   U:11449
##  Mean   : 65.64                           <18    : 1591
##  3rd Qu.: 81.00                           45-64  :  617
##  Max.   :123.00                           65+    :   60
##                                           (Other):    3
##                            PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2839   F: 2615
##  ASIAN / PACIFIC ISLANDER      :  154   18-24  :10086   M:24686
##  BLACK                         :11432   25-44  :12281   U:   11
##  BLACK HISPANIC                : 1314   45-64  : 1863
##  UNKNOWN                       :11786   65+    :  181
##  WHITE                         :  283   UNKNOWN:   62
##  WHITE HISPANIC                : 2341
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19439
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4049
```

After doing all the clean up we can see that there are many unknown values in the perpetrator's race, age and sex. So we decided to concentrate our analysis on victim's profile instead.

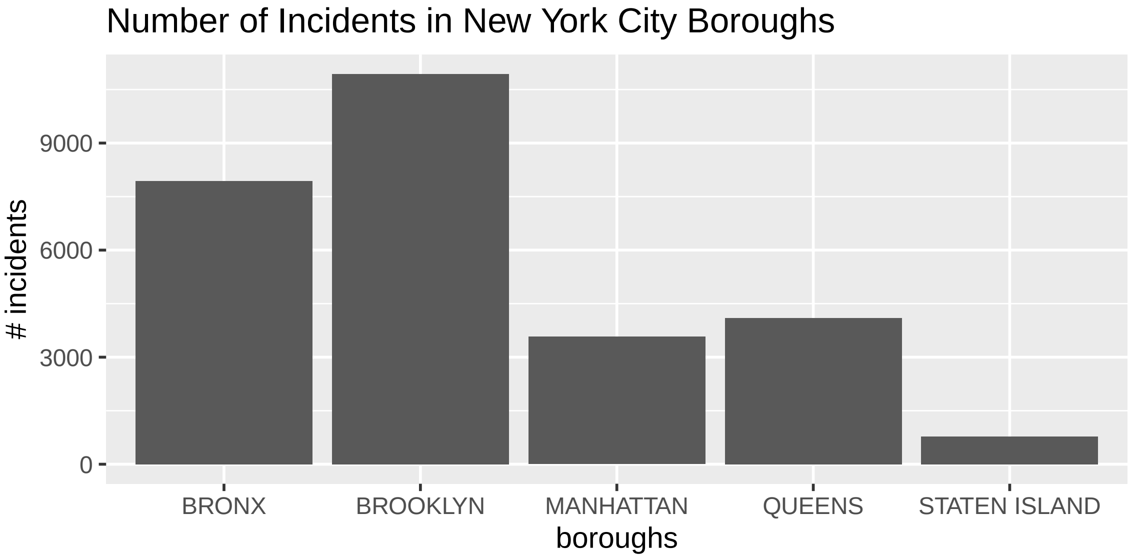
Visualizations and Analysis

Now that we cleaned and prepared our data, we are ready to do some exploratory analysis.

Number of Incidents in New York City Boroughs

First, we will take a look at the total number of incidents in each borough of New York City.

```
ggplot(nypd_data, aes(x=BORO)) +
      geom_bar() +
      labs(title = "Number of Incidents in New York City Boroughs",
           x = "boroughs",
           y = "# incidents")
```

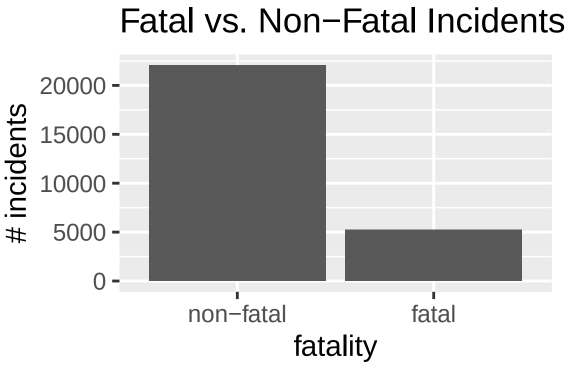


Brooklyn is the borough with the highest number of incidents and Staten Island has the lowest number of incidents.

Fatal vs. Non-Fatal Incidents

Next, we will contrast fatal versus non-fatal incidents.

```
ggplot(nypd_data, aes(x=STATISTICAL_MURDER_FLAG)) +
      geom_bar() +
      labs(title = "Fatal vs. Non-Fatal Incidents",
           x = "fatality",
          y = "# incidents") +
      scale_x_discrete(labels=c("non-fatal", "fatal"))
```

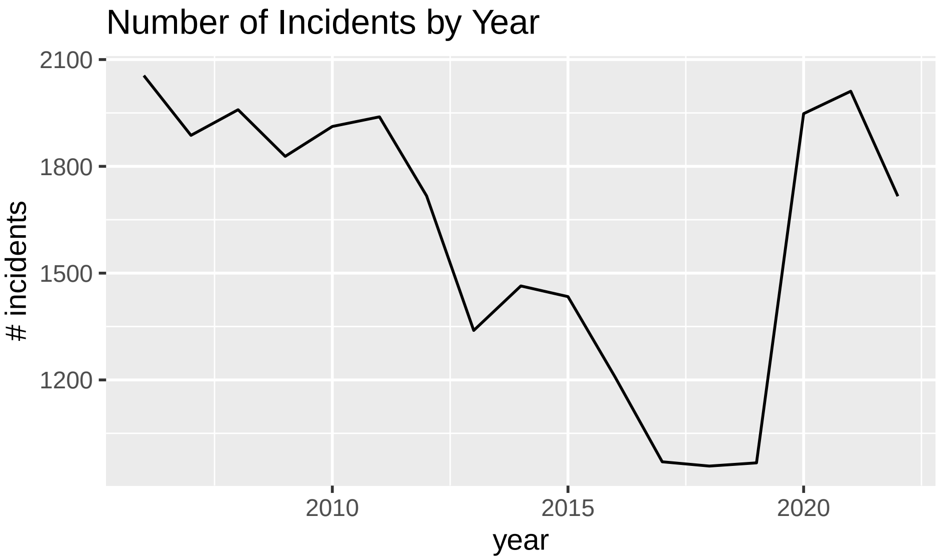


We see that vast majority of incidents are non-fatal.

Incidents by Year

Let's examine how the number of incidents changed over the years.

```
nypd_year <- nypd_data %>% reframe(O_YEAR = year(nypd_data$OCCUR_DATE)) %>%
      group_by(O_YEAR) %>% summarize(YEAR_N = n())
ggplot(nypd_year, aes(x=O_YEAR, y= YEAR_N)) +
      geom_line() +
      labs(title = "Number of Incidents by Year",
          x = "year",
          y = "# incidents")
```



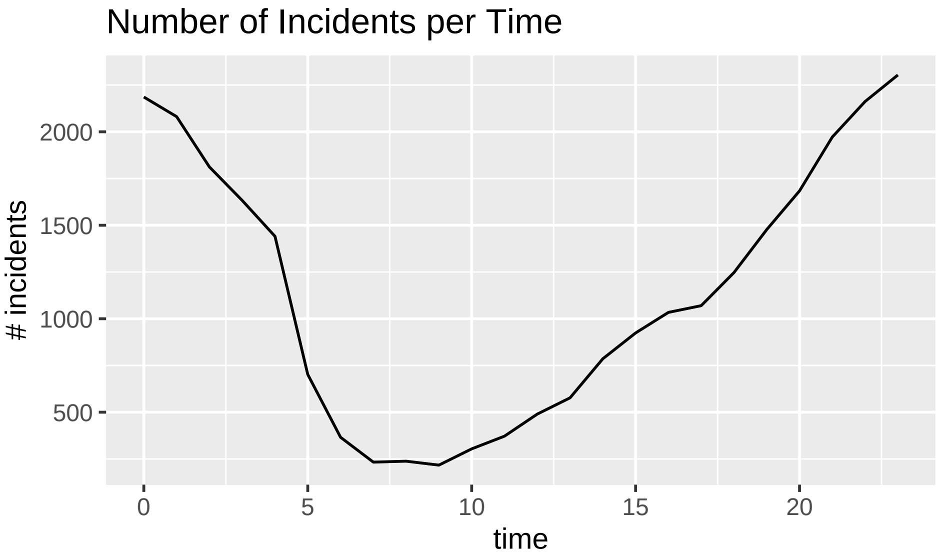
Based on our analysis number of incidents were steadily going down but peaked around 2020. This might be because of the pandemic. This could be an area for further investigation in the future.

Incidents by Time of the Day

Here we will examine the hours of the day when incidents occured.

```
nypd_time <- nypd_data %>% reframe(O_HOUR = hour(nypd_data$OCCUR_TIME)) %>%
      group_by(O_HOUR) %>% summarize(HOUR_N = n()) %>% as.data.frame()
```

```r
ggplot(nypd_time, aes(x=O_HOUR, y = HOUR_N)) +
      geom_line() +
      labs(title = "Number of Incidents per Time",
      x = "time",
      y = "# incidents")
```



Late night has the most number of incidents, morning hours are safest.

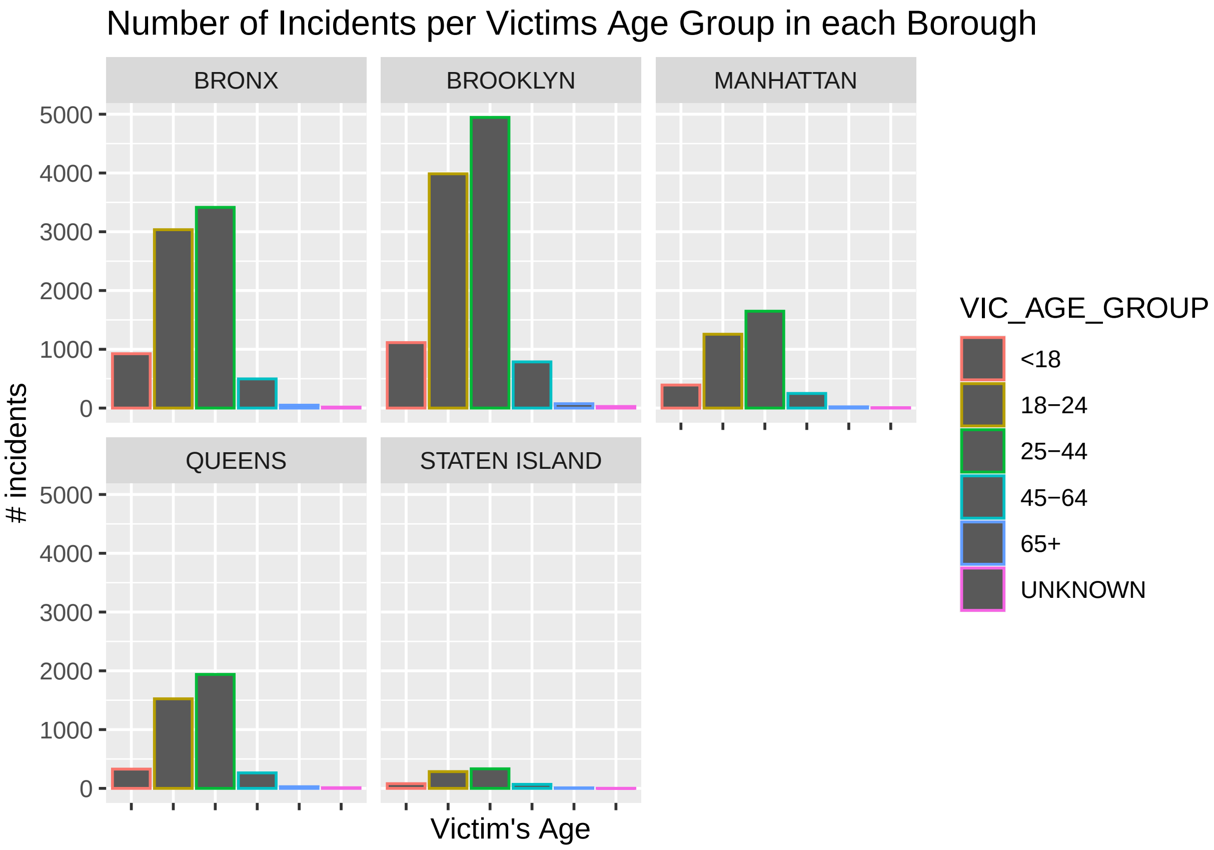
Incidents in each Borough per Victim Age Group

Here we will look at victim's age groups in each borough of New York City.

```r
nypd_ageboro <- nypd_data %>% group_by(BORO, VIC_AGE_GROUP) %>% summarize(incidents = n())
```

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

```r
ggplot(nypd_ageboro, aes(VIC_AGE_GROUP, incidents, col = VIC_AGE_GROUP)) +
      geom_bar(stat = "identity") +
      facet_wrap(vars(BORO)) +
      theme(axis.text.x = element_blank()) +
      labs(title = "Number of Incidents per Victims Age Group in each Borough",
      x = "Victim's Age",
      y = "# incidents")
```



The age group 25-44 has the highest number of victims in each borough with agaim Brooklyn taking the lead. One might be safer in New York past 45 years of age, this might be another point for further analysis in the future.

Model

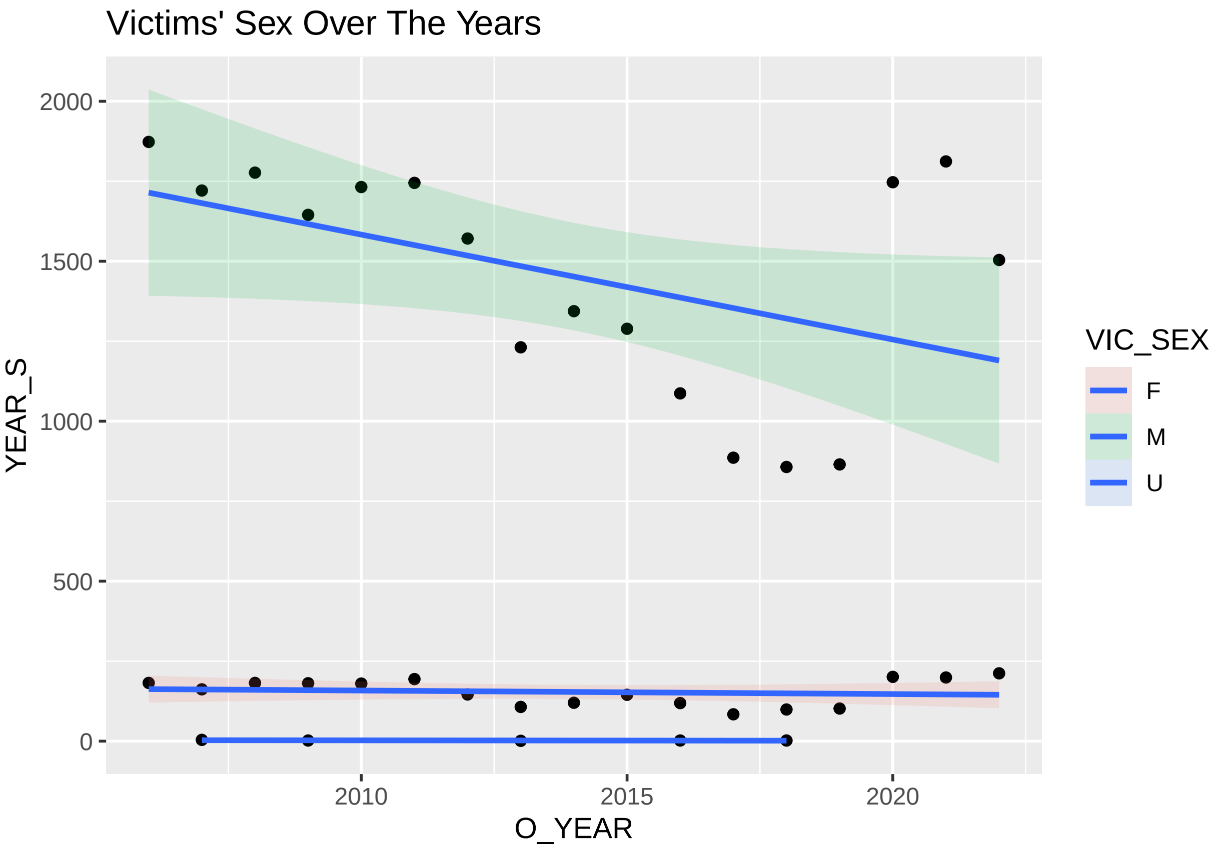
Here we will investigate the change of victim profile over the years, first the sex of the victim. We group the data by year and victim's sex.

```
nypd_years <- nypd_data
nypd_years <- mutate(nypd_data, O_YEAR =  year(nypd_data$OCCUR_DATE))
nypd_years <- nypd_years %>% group_by(O_YEAR, VIC_SEX) %>% summarize(YEAR_S = n())
```

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

```
ggplot(nypd_years, aes(x=O_YEAR, y=YEAR_S), color=VIC_SEX) +
      ggtitle("Victims' Sex Over The Years") +
      geom_point() +
      geom_smooth(method = "lm", alpha = .15, aes(fill = VIC_SEX))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Here we can see that the number of incidents involving female victims stayed generally the same over the years and is well below male victims. The male victims are on a downward trend in general with the exception that we noted in around year 2020 in our previous analysis of Incidents per Year. This leads to our next analysis of victim profile.

Victim Profile Modeling

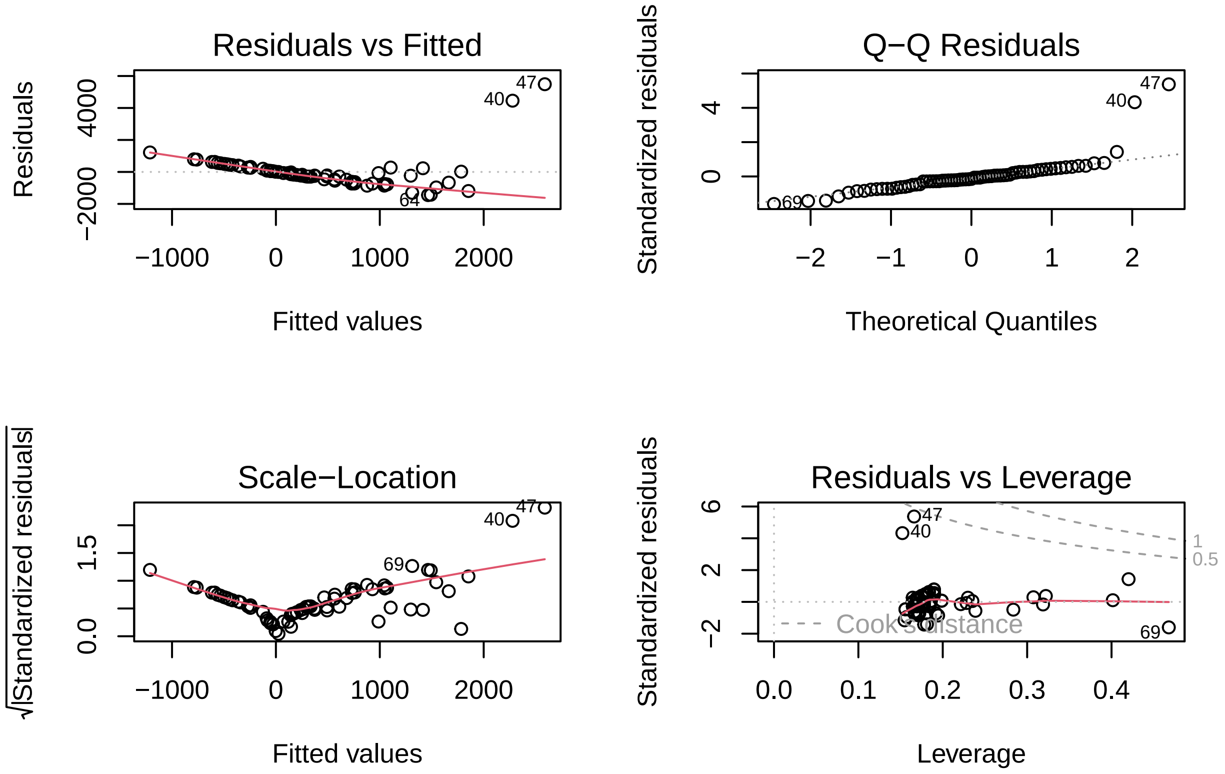
Here we will attempt to profile the victims of shooting incidents, grouping them by sex, race and age group. We will fit a linear model.

```
nypd_profile <- nypd_data
nypd_profile <- nypd_profile %>% group_by(VIC_SEX, VIC_AGE_GROUP, VIC_RACE) %>%
      summarize(N_INCIDENT = n())
```

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

```
fitpro <- lm(N_INCIDENT ~ VIC_SEX + VIC_AGE_GROUP + VIC_RACE, data = nypd_profile)

par(mfrow = c(2,2))
plot(fitpro)
```



```r
summary(fitpro)
```

```
##
## Call:
## lm(formula = N_INCIDENT ~ VIC_SEX + VIC_AGE_GROUP + VIC_RACE,
##     data = nypd_profile)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1448.3  -467.1  -141.7   270.2  5484.9
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1074.6      671.7  -1.600  0.11513
## VIC_SEXM                           734.8      281.0   2.615  0.01141 *
## VIC_SEXU                          -231.2      756.2  -0.306  0.76089
## VIC_AGE_GROUP18-24                 494.6      451.6   1.095  0.27799
## VIC_AGE_GROUP25-44                 804.8      439.8   1.830  0.07251 .
## VIC_AGE_GROUP45-64                -119.6      469.6  -0.255  0.79983
## VIC_AGE_GROUP65+                  -292.5      483.6  -0.605  0.54762
## VIC_AGE_GROUPUNKNOWN              -319.0      488.2  -0.653  0.51616
## VIC_RACEASIAN / PACIFIC ISLANDER   578.3      671.4   0.861  0.39272
## VIC_RACEBLACK                     2123.1      663.0   3.202  0.00223 **
## VIC_RACEBLACK HISPANIC             833.0      667.9   1.247  0.21742
## VIC_RACEUNKNOWN                    412.5      718.6   0.574  0.56817
## VIC_RACEWHITE                      605.0      671.4   0.901  0.37136
## VIC_RACEWHITE HISPANIC             950.0      667.9   1.422  0.16040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1119 on 57 degrees of freedom
## Multiple R-squared:  0.3642, Adjusted R-squared:  0.2193
## F-statistic: 2.512 on 13 and 57 DF,  p-value: 0.00858
```

We see in the Residuals vs Fitted plot two extreme values. We check what those are.

```
nypd_profile[47,]
```

```
## # A tibble: 1 x 4
## # Groups:   VIC_SEX, VIC_AGE_GROUP [1]
##   VIC_SEX VIC_AGE_GROUP VIC_RACE N_INCIDENT
##   <fct>   <fct>         <fct>         <int>
## 1 M       25-44         BLACK          8073
```

```
nypd_profile[40,]
```

```
## # A tibble: 1 x 4
## # Groups:   VIC_SEX, VIC_AGE_GROUP [1]
##   VIC_SEX VIC_AGE_GROUP VIC_RACE N_INCIDENT
##   <fct>   <fct>         <fct>         <int>
## 1 M       18-24         BLACK          6733
```

According to our model being black, male and in the age group 25-44 is the most common victim profile. When we check the residual plot also we can see this profile as the most extreme outlier with black, male, age group 18-24 being the other extreme value.

Conclusion

In our analysis we found that being black, male and between ages 25-44 is the most common victim profile. We saw that female gun incident victims are lower in number and relatively steady over time. While male victim numbers are significantly higher, they seem to be on a downward trend with the exception around year 2020.

Brooklyn has the most number of gun related incidents. Most victims are between ages 25-44. Late night hours have the most number of incidents and vast majority of incidents are non-fatal.

Our conclusion is being black, male, in the age group 25-44, being in Brooklyn late at night make one most likely to fall victim of a gun related incident. Thankfully it is most likely to be non-fatal and the number of incidents seem to be on a downward trend again after 2020.

Suggestions for Further Research

There is a downward trend in incidents but there is a peak around 2020 and again it is coming down. The reasons behind this pattern could be subject for further research, for example possible effect of covid-19.

After the age of 45 the victims' numbers drop dramatically, the reasons for this could be investigated taking into account overall demographic information about the population in the area.

When comparing number of incidents in boroughs population data could be taken into consideration.

Potential Sources of Bias

There are many missing values in the data set, how we impute them can be potential source of bias. In particular, the PERP_RACE, perpetrator's race variable has more unknown or empty values than the highest factor category (BLACK: 11432 vs. UNKNOWN: 11786). With so many missing values we chose to leave out the perpetrator's demographics from our analysis.

Of course our own biases can influence our analysis as well. Personally, I feel strongly that there should be stricter gun control laws and it should not be so easy for people to get guns. However, I realize that many people see having guns as their right. My bias would suggest that most incidents involving guns would be fatal, but the data shows otherwise as in the Fatal vs Non-Fatal section.

Another source of personal bias is the view that violence against women are on the rise. However according to this data that is not true, it seems pretty much level and is well below male's.

Appendix

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.4
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib 
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: Europe/Istanbul
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] lubridate_1.9.2 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4    
##  [5] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1   
##  [9] ggplot2_3.4.4   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] Matrix_1.6-3      gtable_0.3.4      highr_0.10        compiler_4.3.1   
##  [5] tidyselect_1.2.0  splines_4.3.1     scales_1.3.0      yaml_2.3.8       
##  [9] fastmap_1.1.1     lattice_0.22-5    R6_2.5.1          labeling_0.4.3   
## [13] generics_0.1.3    knitr_1.45        munsell_0.5.0     pillar_1.9.0     
## [17] tzdb_0.4.0        rlang_1.1.2       utf8_1.2.4        stringi_1.8.3    
## [21] xfun_0.41         timechange_0.2.0  cli_3.6.2         mgcv_1.8-42      
## [25] withr_2.5.2       magrittr_2.0.3    digest_0.6.33     grid_4.3.1       
## [29] rstudioapi_0.15.0 hms_1.1.3         nlme_3.1-162      lifecycle_1.0.4  
## [33] vctrs_0.6.5       evaluate_0.23     glue_1.6.2        farver_2.1.1     
## [37] fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.25    tools_4.3.1      
## [41] pkgconfig_2.0.3   htmltools_0.5.7
```