NPYD Shooting Incident

2024-06-05

Data description

The data is found on this page: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year (2013 in this case). This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website.

A description of the data can be found here: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/about_data

```r
library(tidyverse)
library(plyr)
library(ggplot2)
library(lubridate)

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

```
##   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
```

Tidy and Transform

Looking at the data structure, there is no need to pivoting any columns.

1. For my analysis purpose, I will keep the following interesting information OCCUR_DATE OCCUR_TIME BORO STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE

```
shooting <- shooting %>%
  select(OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VI
```

2. Check the unique values of each column that we want to convert to factor

```
map_df(shooting %>% select(-c(OCCUR_DATE, OCCUR_TIME, BORO)), ~tibble( unique_values = toString(unique(
```

```
## # A tibble: 7 x 1
##   unique_values
##   <chr>
## 1 TRUE, FALSE
## 2 25-44, (null), NA, 18-24, 45-64, UNKNOWN, <18, 65+, 1020, 940, 224, 1028
## 3 M, (null), NA, F, U
## 4 BLACK, (null), NA, UNKNOWN, WHITE HISPANIC, BLACK HISPANIC, ASIAN / PACIFIC I~
## 5 25-44, 18-24, 45-64, 65+, <18, UNKNOWN, 1022
## 6 M, F, U
## 7 BLACK, WHITE, WHITE HISPANIC, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, UNKNO~
```

From the result we can see there are some thing needs to be cleaned up. We doing so by convert all unknown or unreasonable data to NA

```
shooting$PERP_AGE_GROUP = mapvalues(shooting$PERP_AGE_GROUP, from=c("224","940", "1020", "1028", "UNKNO
shooting$PERP_SEX = mapvalues(shooting$PERP_SEX, from=c("(null)","U"), to=rep(NA, 2))
shooting$PERP_RACE = mapvalues(shooting$PERP_RACE, from=c("(null)", "UNKNOWN"), to=rep(NA, 2))
shooting$VIC_AGE_GROUP = mapvalues(shooting$VIC_AGE_GROUP, from=c("1022", "UNKNOWN"), to=rep(NA, 2))
shooting$VIC_SEX = mapvalues(shooting$VIC_SEX, from=c("U"), to=rep(NA, 1))
shooting$VIC_RACE = mapvalues(shooting$VIC_RACE, from=c("UNKNOWN"), to=rep(NA, 1))
map_df(shooting %>% select(-c(OCCUR_DATE, OCCUR_TIME, BORO)), ~tibble( unique_values = toString(unique(
```

```
## # A tibble: 7 x 1
##   unique_values
##   <chr>
## 1 TRUE, FALSE
## 2 25-44, NA, 18-24, 45-64, <18, 65+
## 3 M, NA, F
## 4 BLACK, NA, WHITE HISPANIC, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, WHITE, A~
## 5 25-44, 18-24, 45-64, 65+, <18, NA
## 6 M, F, NA
## 7 BLACK, WHITE, WHITE HISPANIC, BLACK HISPANIC, ASIAN / PACIFIC ISLANDER, NA, A~
```

Following data type should be transformed: OCCUR_DATE: Date PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE: Factor

```
shooting <- shooting %>%
  mutate(OCCUR_DATE=mdy(OCCUR_DATE)) %>%
  mutate(PERP_AGE_GROUP=factor(PERP_AGE_GROUP)) %>%
  mutate(PERP_SEX=factor(PERP_SEX)) %>%
  mutate(PERP_RACE=factor(PERP_RACE)) %>%
  mutate(VIC_AGE_GROUP=factor(VIC_AGE_GROUP)) %>%
  mutate(VIC_SEX=factor(VIC_SEX)) %>%
  mutate(VIC_RACE=factor(VIC_RACE))
summary(shooting)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :2006-01-01   Length:28562      Length:28562
##  1st Qu.:2009-09-04   Class1:hms        Class :character
##  Median :2013-09-20   Class2:difftime   Mode  :character
##  Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:2019-09-29
##  Max.   :2023-12-29
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  Mode :logical           <18  : 1682    F   :  444
##  FALSE:23036             18-24: 6438    M   :16168
##  TRUE :5526              25-44: 6041    NA's:11950
##                          45-64:  699
##                          65+  :   65
##                          NA's :13637
##
##                            PERP_RACE     VIC_AGE_GROUP VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18  : 2954   F   : 2760
##  ASIAN / PACIFIC ISLANDER      :  169   18-24:10384   M   :25790
##  BLACK                         :11903   25-44:12973   NA's:   12
##  BLACK HISPANIC                : 1392   45-64: 1981
##  WHITE                         :  298   65+  :  205
##  WHITE HISPANIC                : 2510   NA's :   65
##  NA's                          :12288
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   11
##  ASIAN / PACIFIC ISLANDER      :  440
##  BLACK                         :20235
##  BLACK HISPANIC                : 2795
##  WHITE                         :  728
##  WHITE HISPANIC                : 4283
##  NA's                          :   70
```

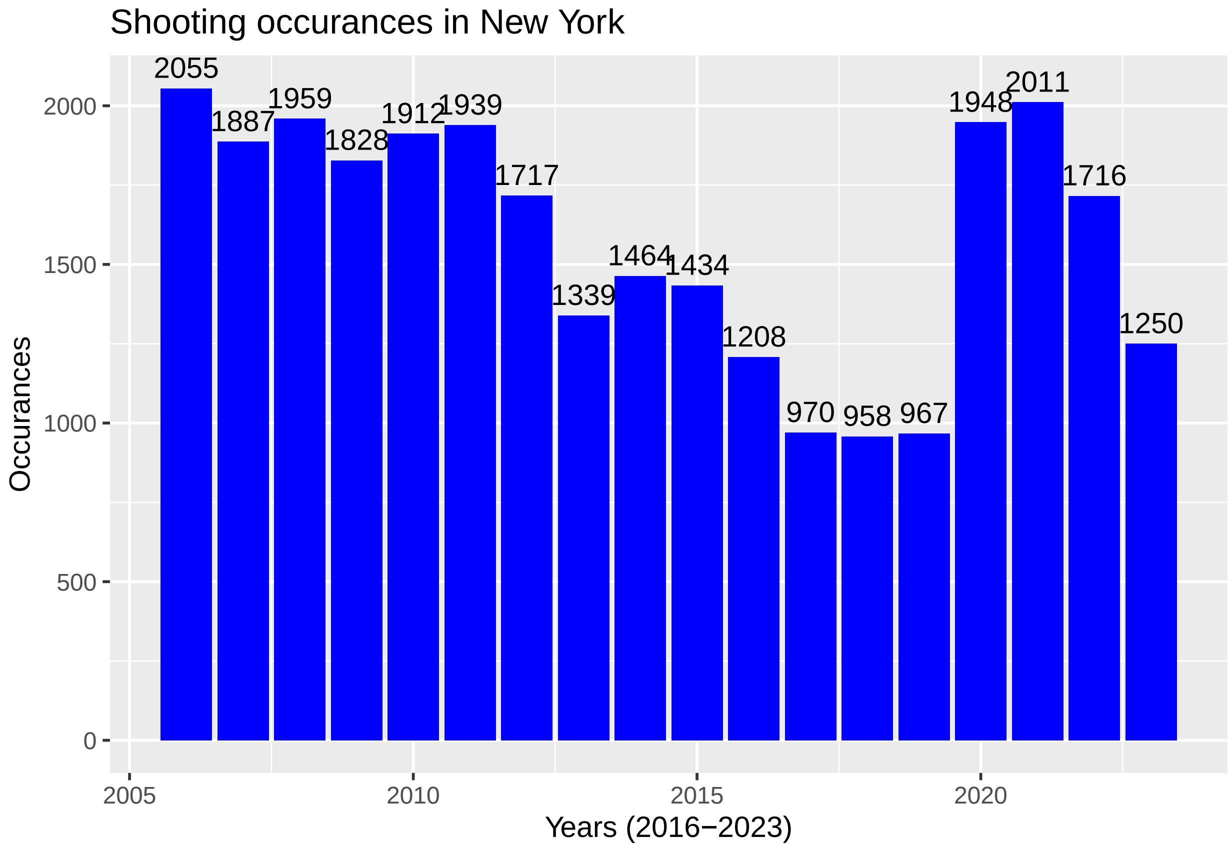
After that, we can see it contains reasonable data or NA's. For the NA's we will leave them as the are, and we will probabaly convert/filter them later when necessary.

Analysis and Visualisation

Shooting occurrances by year

At first, I want to group the shootings by year for the whole city, and plot it.

```
shooting %>% mutate(year=(year(OCCUR_DATE))) %>%
  ggplot(aes(x=year))  +
  geom_bar(fill = "blue", show.legend = FALSE) +
  geom_text(stat='count', aes(label=after_stat(count)), vjust=-0.5) +
  labs(title="Shooting occurances in New York",
       x="Years (2016-2023)", y="Occurances")
```

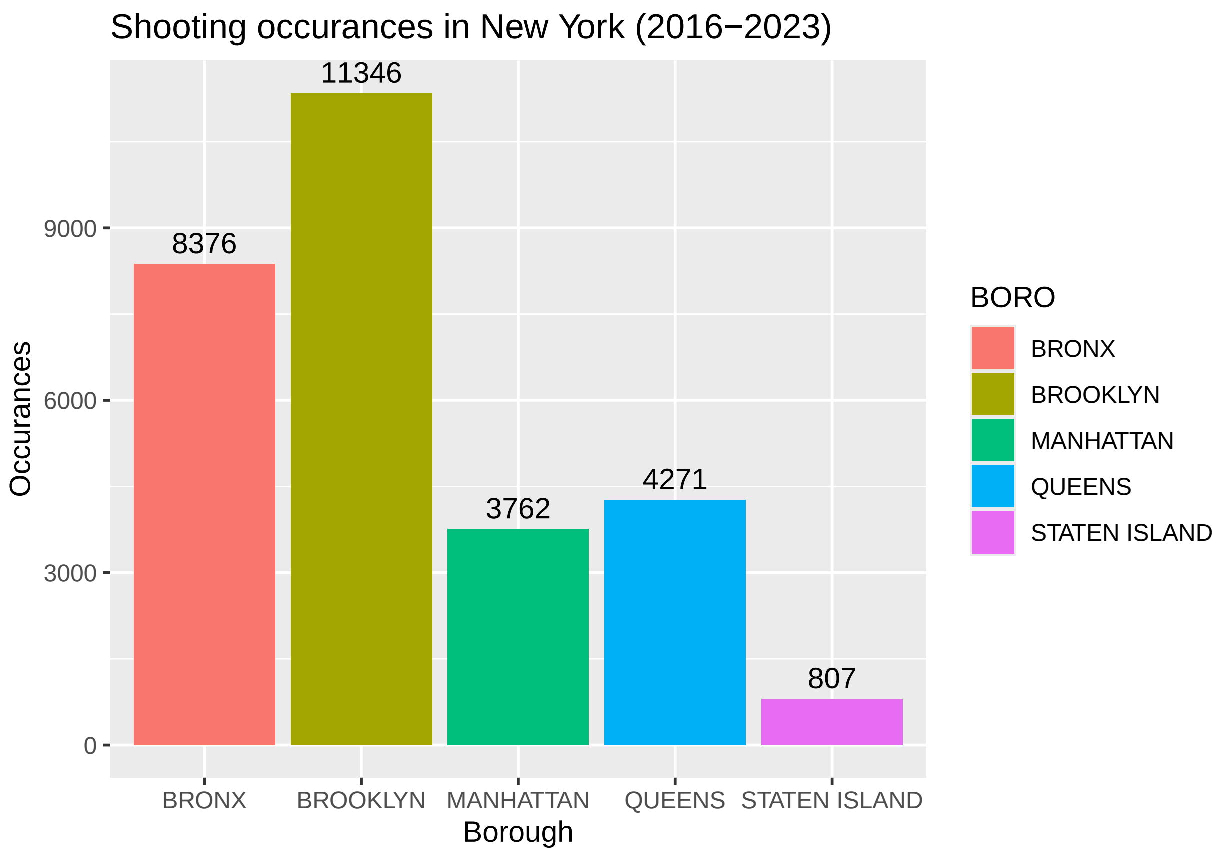


From the plot we can see the shooting occurances in NewYork decreases in general from 2005 to 2019, and there is a sudden incrase in 2020, 2021 and then going down slowly. It seems unusual that after several years of decrease it increased suddenly, it may worth for further investigation.

Shooting occurance by borough

Next I would like to visualize over all the years the occurances in each brough.

```
shooting %>%
  ggplot(aes(x=BORO, fill=BORO)) + geom_bar() +
  geom_text(stat='count', aes(label=after_stat(count)), vjust=-0.5) +
  labs(title="Shooting occurances in New York (2016-2023)",
       x="Borough", y="Occurances")
```

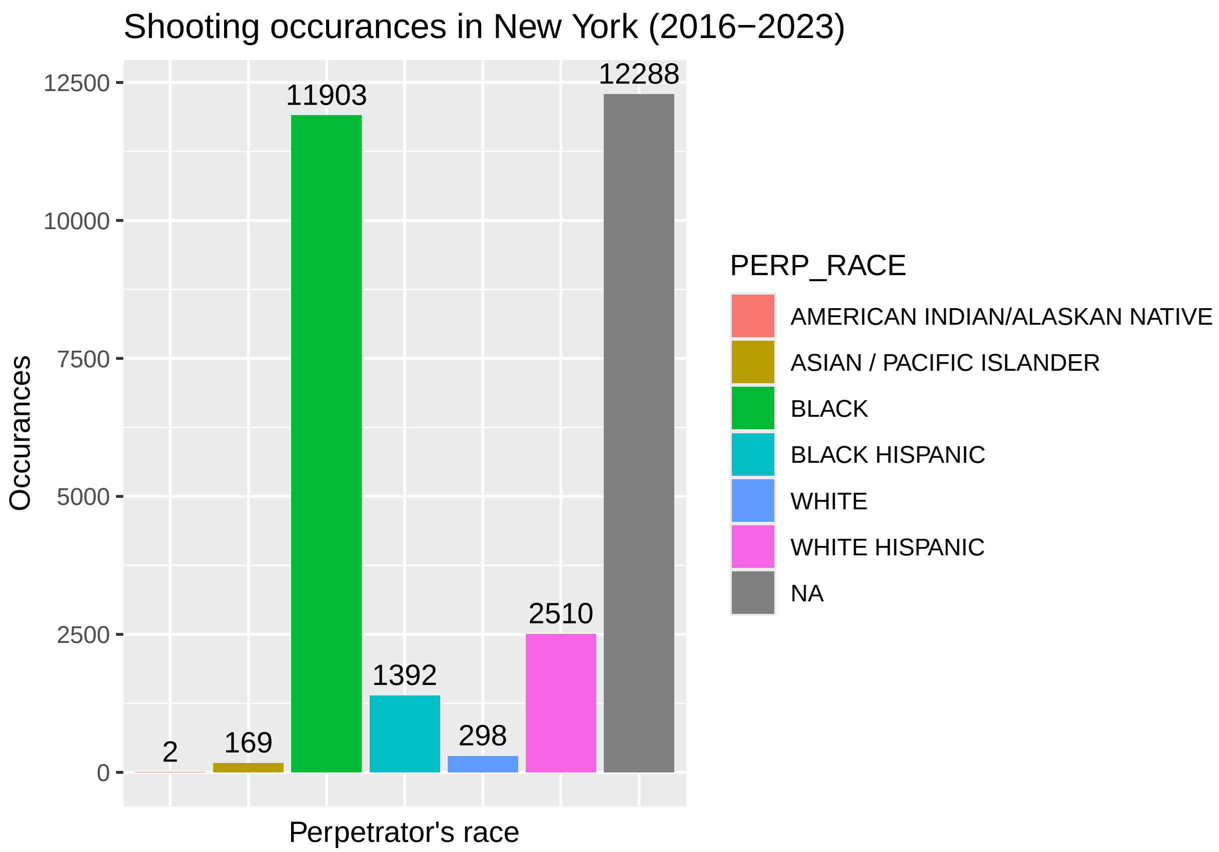


From the plot we can see the occurances vary quite much, the question is why? Is it because some borough is safer than others or it is much smaller so the occurances are also smaller? It may also worth to further investigat.

Shooting occurance by Perpetrator's race

Next I would like to visualize the occurances by perpetrator's race

```
shooting %>%
  ggplot(aes(x = PERP_RACE, fill = PERP_RACE)) +
  geom_bar() +
  geom_text(stat='count', aes(label=after_stat(count)), vjust=-0.5) +
  labs(title="Shooting occurances in New York (2016-2023)",
       x="Perpetrator's race", y="Occurances") +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



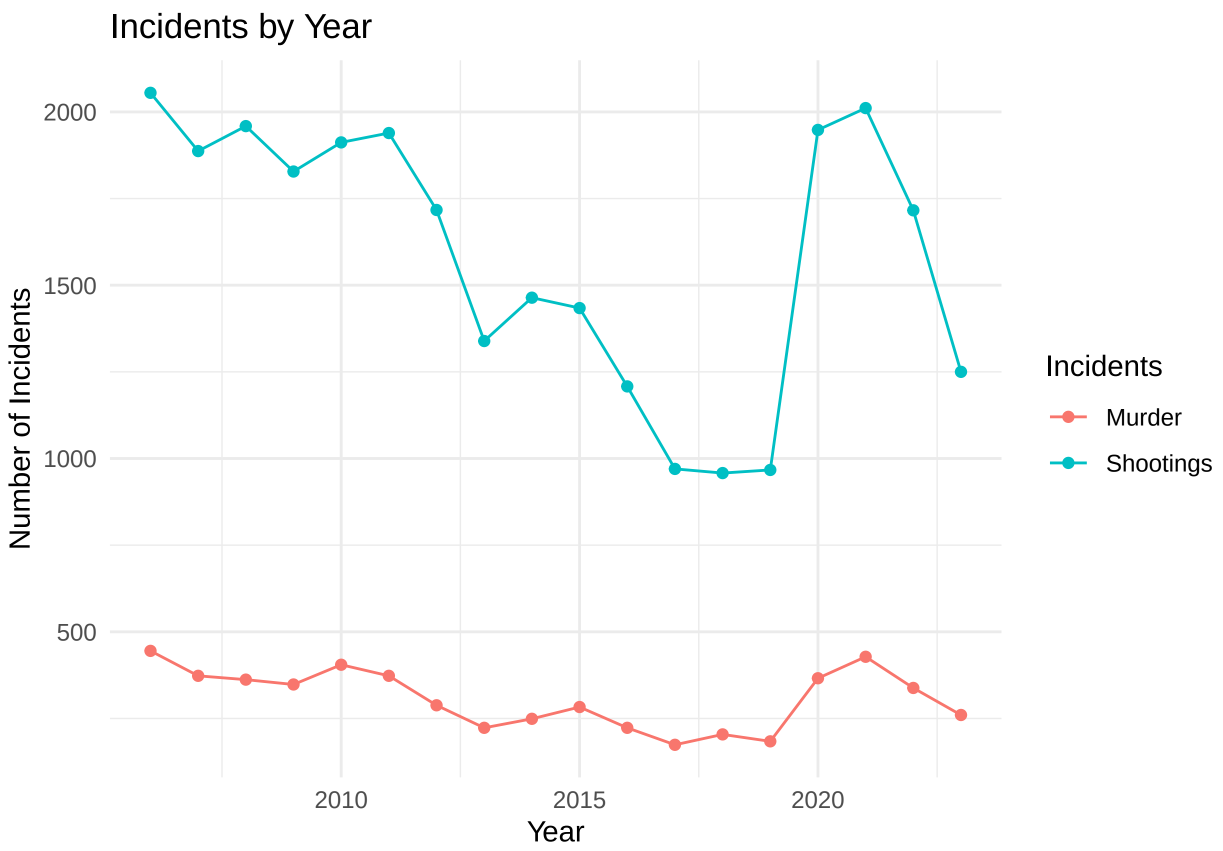
From the plot we can see there are quite some missing data, and for those not missing, the race “BLACK” is quite high. It may worth to furhter investigate why. If it is true that most of the shootings are by blacks? or is there some possible issue in data collection?

Modeling

First plot the totoal incidents and murder per year

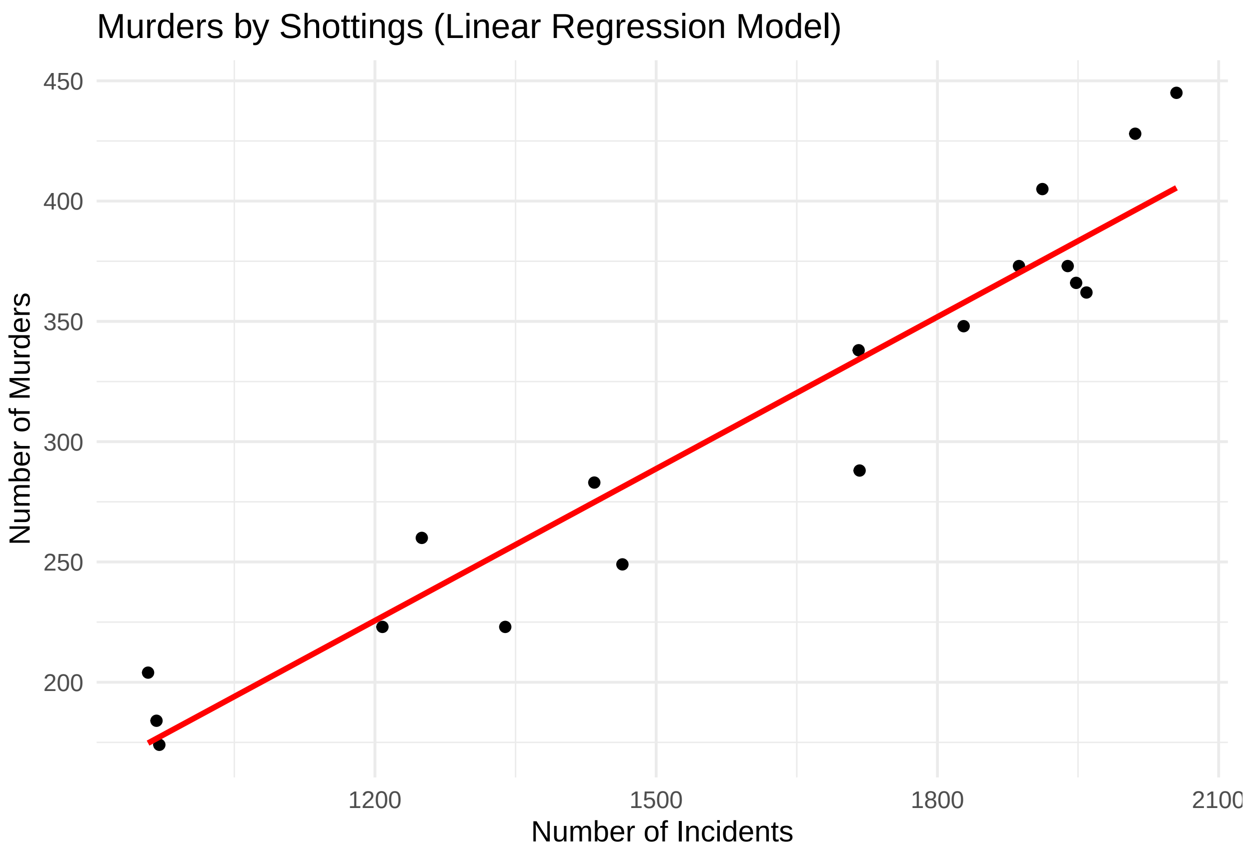
```r
shooting_by_year_with_murder <- shooting %>%
  mutate(Year = year(OCCUR_DATE)) %>%
  group_by(Year) %>%
  dplyr::summarize(Incidents = n(), Murder = sum(STATISTICAL_MURDER_FLAG), .groups = "drop")

shooting_by_year_with_murder %>%
  ggplot(aes(x = Year)) +
  geom_line(aes(y = Incidents, color = "Shootings")) +
  geom_line(aes(y = Murder, color = "Murder")) +
  geom_point(aes(y = Incidents, color = "Shootings")) +
  geom_point(aes(y = Murder, color = "Murder")) +
  labs(title = "Incidents by Year",
  x = "Year",
  y = "Number of Incidents",
  color = "Incidents") +
  theme_minimal()
```



It looks the number of murders is correlated with the total incidents, let's try to model it using linear model.

```
ggplot(shooting_by_year_with_murder, aes(x = Incidents, y = Murder)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "red") +
labs(title = "Murders by Shottings (Linear Regression Model)",
x = "Number of Incidents",
y = "Number of Murders") +
theme_minimal()
```



From the plot we can see these two variables correlate with each other quite well, i.e. when there are more shootings, there are more murders.

Conclusion and possible sources of bias

In this small project I imported, tidied, transformed and visualized the shooting data in New York bwetten 2006 and 2023.

There are some thing unusual observed from the visualization, and identified some questions that may worth to further investigate.

There are might be some sources of bias in the data, e.g.

1. How the data is collection?
2. Is it complete, could there be systematic bias that course certrain data missing?

And there could personal biases during the process and analysis, e.g.

1. One may have a biased impression of which boroughs is safe/unsafe
2. One may have a biased impression of races/sexes

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.5
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib 
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
## 
## 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/Berlin
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] plyr_1.8.9      lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1  
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1    
##  [9] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    lattice_0.22-6    stringi_1.8.4    
##  [5] hms_1.1.3         digest_0.6.35     magrittr_2.0.3    evaluate_0.23    
##  [9] grid_4.4.0        timechange_0.3.0  fastmap_1.2.0     Matrix_1.7-0     
## [13] mgcv_1.9-1        fansi_1.0.6       scales_1.3.0      cli_3.6.2        
## [17] rlang_1.1.3       crayon_1.5.2      splines_4.4.0     bit64_4.0.5      
## [21] munsell_0.5.1     withr_3.0.0       yaml_2.3.8        tools_4.4.0      
## [25] parallel_4.4.0    tzdb_0.4.0        colorspace_2.1-0  curl_5.2.1       
## [29] vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4   bit_4.0.5        
## [33] vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0      gtable_0.3.5     
## [37] glue_1.7.0        Rcpp_1.0.12       highr_0.11        xfun_0.44        
## [41] tidyselect_1.2.1  rstudioapi_0.16.0 knitr_1.47        farver_2.1.2     
## [45] nlme_3.1-164      htmltools_0.5.8.1 rmarkdown_2.27    labeling_0.4.3   
## [49] compiler_4.4.0
```