NYPD shooting

Jamil Garro

4/15/2024

Objective and dataset

Dataset

List of all the shooting that occurs in New York city from 2006 to 2022. Published by New York city. 27000 rows and 21 columns.

Objective

Yearly trends about shooting incidents in New York city. Time of day incidence occurrence. Day of week incidence occurring.

Preparation

R packages

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(ggplot2)
library(forcats)
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.3

## Warning: package 'lubridate' was built under R version 4.1.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v lubridate 1.9.2     v stringr   1.5.1
## v purrr     1.0.1     v tibble    3.2.1
## v readr     2.1.4     v tidyr     1.3.0

## -- 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 error
```

```r
library(lubridate)
```

Data import

From: https://catalog.data.gov/dataset?q=NYPD+shooting+incident&sort=views_recent+desc&ext_location=&ext_bbox=&ext_prev_extent=

```r
url <- c('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')
shooting <- read_csv(url)
```

```
## Rows: 28562 Columns: 21-- Column specification -----------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

Exploration and tidying of data

Exploration

Summary

```
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
```

Sample data

```
head(shooting, n=5)
```

```
## # A tibble: 5 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   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>
```

Tidying

```

shooting_tidy <- shooting %>% select(-c(X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat))
```

Removal of unnecessary information

removal of perpetrator info that are unuselful for our analysis

```
shooting_tidy <- shooting_tidy %>%
  select(-c(PERP_AGE_GROUP,PERP_SEX,PERP_RACE)) %>%
  select(-c(LOCATION_DESC,JURISDICTION_CODE,LOCATION_DESC))
```

converting date to date format

```
shooting_tidy$OCCUR_DATE <- as.Date(shooting_tidy$OCCUR_DATE,
                                    format = "%m/%d/%Y")
```

Data analysis

Day of shooting incidents

Will be studied the relation between day of occurrence of incidence against number of occurrence of incidence. A new variable will be created "day of week".

```
shooting_dayofweek <- shooting_tidy %>%
  mutate(day_of_week = wday(OCCUR_DATE, week_start = 1))
```

Aggregate of shooting on the different days of the week will be calculated.

```
shooting_dayofweek_agg <- shooting_dayofweek %>% group_by(day_of_week) %>%
  summarize(count = n())
```

Trends of shooting per years

Will be analyzed if there is any yearly trends in shooting incidents. Year of occurence will be derived.

```
shooting_year <- shooting_tidy %>% mutate(year = year(OCCUR_DATE))
```

total account for year and gender will be calculated.

```
shooting_year_agg <- shooting_year %>%
  group_by(victim_gender = VIC_SEX, year = year(OCCUR_DATE)) %>%
  summarize(count = n())
```

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

incidence by time of the day

Will be analyzed shooting incidence occurrence by time of the day, a new column will be created related to time of the day of the shooting.

```
shooting_hour <- shooting_tidy %>%
  mutate(hour = format(as.POSIXct(OCCUR_TIME,format="%H:%M:%S"),"%H"))
```

Now the sum of shooting by hour of the day.

```
shooting_hour_agg <- shooting_hour %>% group_by(hour) %>%
  summarize(count = n())
```

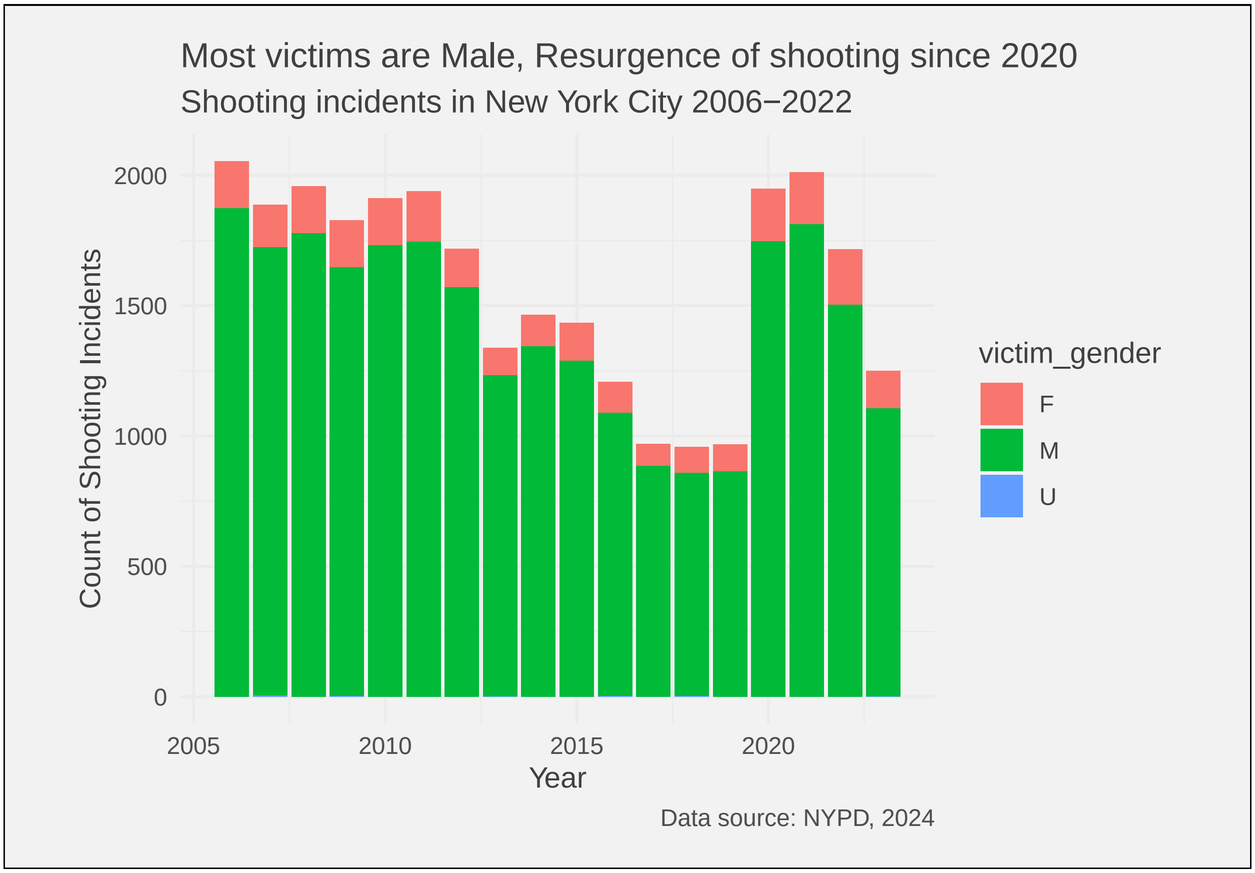
Visualisation of data

theme for visualisation

```
theme_shooting <- function() {theme_minimal() +
    theme(text = element_text(color="gray25"),
          plot.subtitle = element_text(size = 12),
          plot.caption = element_text(color = "gray30"),
          plot.background = element_rect(fill = "gray95"),
          plot.margin = unit(c(5, 10, 5, 10), units = "mm"))}
```

historical trend by years and by gender

```
ggplot(data = shooting_year_agg, aes(x = year, y = count, fill=victim_gender))+
  geom_bar(stat = "identity") +
  labs(x = "Year", y = "Count of Shooting Incidents",
       title = "Most victims are Male, Resurgence of shooting since 2020",
       subtitle = "Shooting incidents in New York City 2006-2022",
       caption = "Data source: NYPD, 2024") + theme_shooting()
```

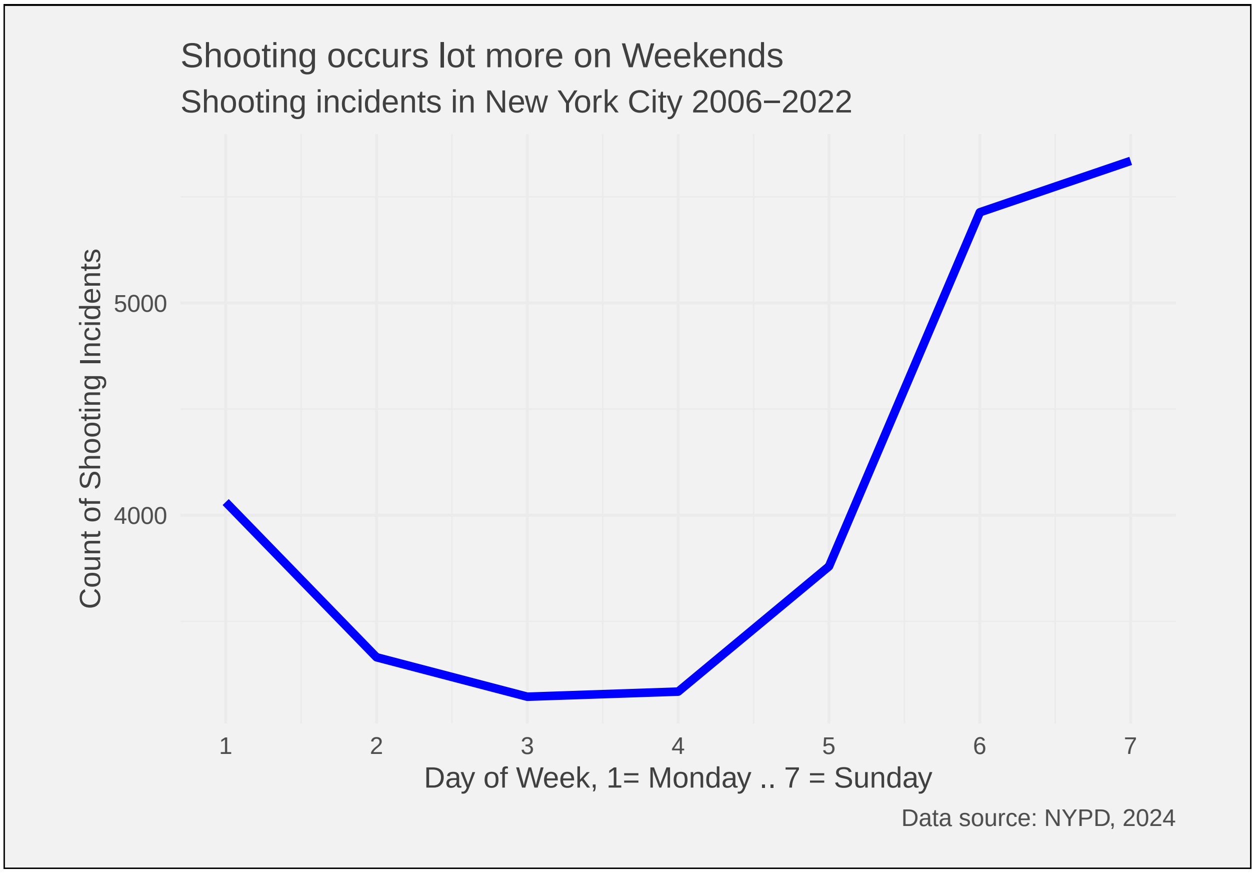

Most victims are Male, Resurgence of shooting since 2020

Shooting incidents in New York City 2006−2022

Data source: NYPD, 2024

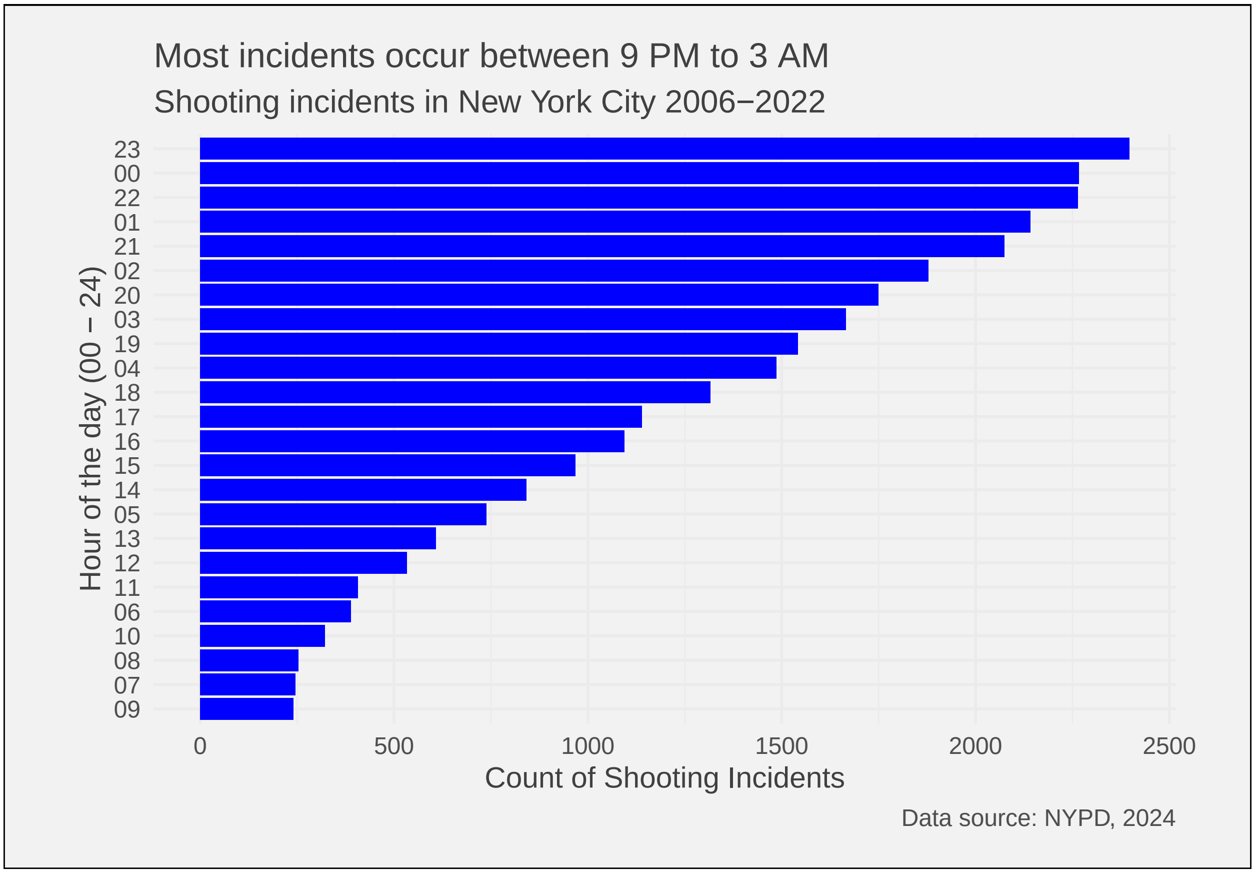
day of occurence of shooting

```
ggplot(data = shooting_dayofweek_agg,aes(x = day_of_week, y = count)) +
  geom_line(color = "blue", size =1.5) +
  scale_x_continuous(breaks = shooting_dayofweek_agg$day_of_week,
                     labels = shooting_dayofweek_agg$day_of_week) +
  labs(x = "Day of Week, 1= Monday .. 7 = Sunday",
       y = "Count of Shooting Incidents", title = "Shooting occurs lot more on Weekends",
       subtitle = "Shooting incidents in New York City 2006-2022",
       caption = "Data source: NYPD, 2024") + theme_shooting()
```



shooting by hour of the day

```
ggplot(data = shooting_hour_agg, aes(x = reorder(hour, count), y = count)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Hour of the day (00 - 24)",
       y = "Count of Shooting Incidents",
       title = "Most incidents occur between 9 PM to 3 AM",
       subtitle = "Shooting incidents in New York City 2006-2022",
       caption = "Data source: NYPD, 2024") + theme_shooting() + coord_flip()
```



Model: day of occurence of shooting

week-end variable

This is a variable that assess if shooting occurred on week-end vs normal day of the week.

```
shooting_dow_agg <- mutate(shooting_dayofweek_agg,
                           is_weekendd = ifelse(shooting_dayofweek_agg$day_of_week < 6,0,1))
```

Creating a linear model

to predict relationship between is_weekend and count of shooting incidents

```
shooting_model <- lm(is_weekendd ~ count, data=shooting_dow_agg)
```

the model

```
shooting_model
```

```
##
## Call:
## lm(formula = is_weekendd ~ count, data = shooting_dow_agg)
##
```

```
## Coefficients:
## (Intercept)        count
##  -1.4998657    0.0004376
```

performance of the model

```
summary(shooting_model)
```

```
##
## Call:
## lm(formula = is_weekendd ~ count, data = shooting_dow_agg)
##
## Residuals:
##        1        2        3        4        5        6        7
## -0.27771  0.04218  0.12358  0.11307 -0.14512  0.12495  0.01905
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.500e+00  2.749e-01  -5.455  0.00281 **
## count        4.376e-04  6.552e-05   6.679  0.00114 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1697 on 5 degrees of freedom
## Multiple R-squared:  0.8992, Adjusted R-squared:  0.879
## F-statistic: 44.61 on 1 and 5 DF,  p-value: 0.001137
```

Discussing the model

The p_value of the model is approximately 0.0011 well below 0.05 which gives validity to the model. There is a significant relationship between number of shooting incidents and day of week.

comparing linear model to actual shooting incidents

```
shooting_dow_agg <- shooting_dow_agg %>% mutate(pred = predict(shooting_model))
```

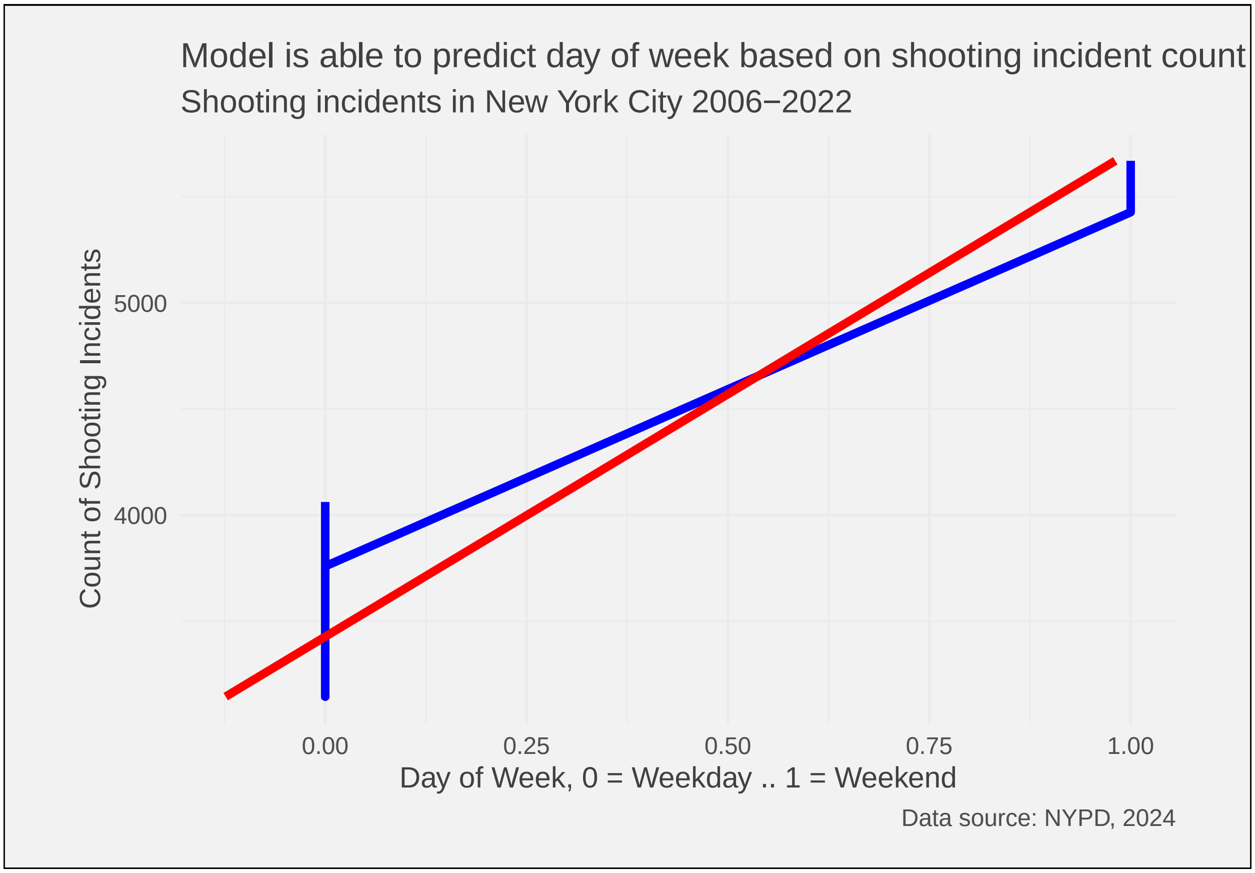
model performance

```
ggplot(data = shooting_dow_agg) + geom_line(color = "blue",
                                            size =1.5, aes(x = shooting_dow_agg$is_weekendd, y= shootin
  geom_line(color = "red", size =1.5, aes(x = shooting_dow_agg$pred,
                                          y= shooting_dow_agg$count)) +
  labs(x = "Day of Week, 0 = Weekday .. 1 = Weekend",
       y = "Count of Shooting Incidents",
       title = "Model is able to predict day of week based on shooting incident count",
       subtitle = "Shooting incidents in New York City 2006-2022",
       caption = "Data source: NYPD, 2024") + theme_shooting()
```

```
## Warning: Use of ‘shooting_dow_agg$is_weekendd‘ is discouraged. Use ‘is_weekendd‘
## instead.
## Warning: Use of ‘shooting_dow_agg$count‘ is discouraged. Use ‘count‘ instead.

## Warning: Use of ‘shooting_dow_agg$pred‘ is discouraged. Use ‘pred‘ instead.

## Warning: Use of ‘shooting_dow_agg$count‘ is discouraged. Use ‘count‘ instead.
```



data bias

Information in this data set are provided by NYPD, so it only register data that were reported to the police. If there was a lot of shooting incidents that were not reported to the police, a bias could exist.

If shooting incidents were not registered at the actual time of occurrence but only at a rounding time, a bias could occur as in this study we are considering the time of occurrence specifically.

conclusion

According to this study it appears that most of the shooting significantly occurred on week-end. It appears that activity of population increase in week-end and so the occurrence of shooting incidents.

Most victims are males. It appears in this study that more than 80% of victims are males. Males are more likely to be involved in violent activities.

Violence and shooting incidences has significantly increased from 2020 till now as clarified by this study, reasons for that are not known but it might be related to the COVID 19 pandemic.