# NYPD Shooting Incident Data Report

Lin Azhi

2024-03-02

```
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 1.9.3     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
```

## What is this?

This is the maiden work of a data science student to complete all steps in the data science process in a reproducible manner, by using the NYPD Shooting Incident data.

## Step 1 - Important the data in a reproducible manner

```
url = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shooting = read.csv(url)
```

## Step 2 - Tidy the data

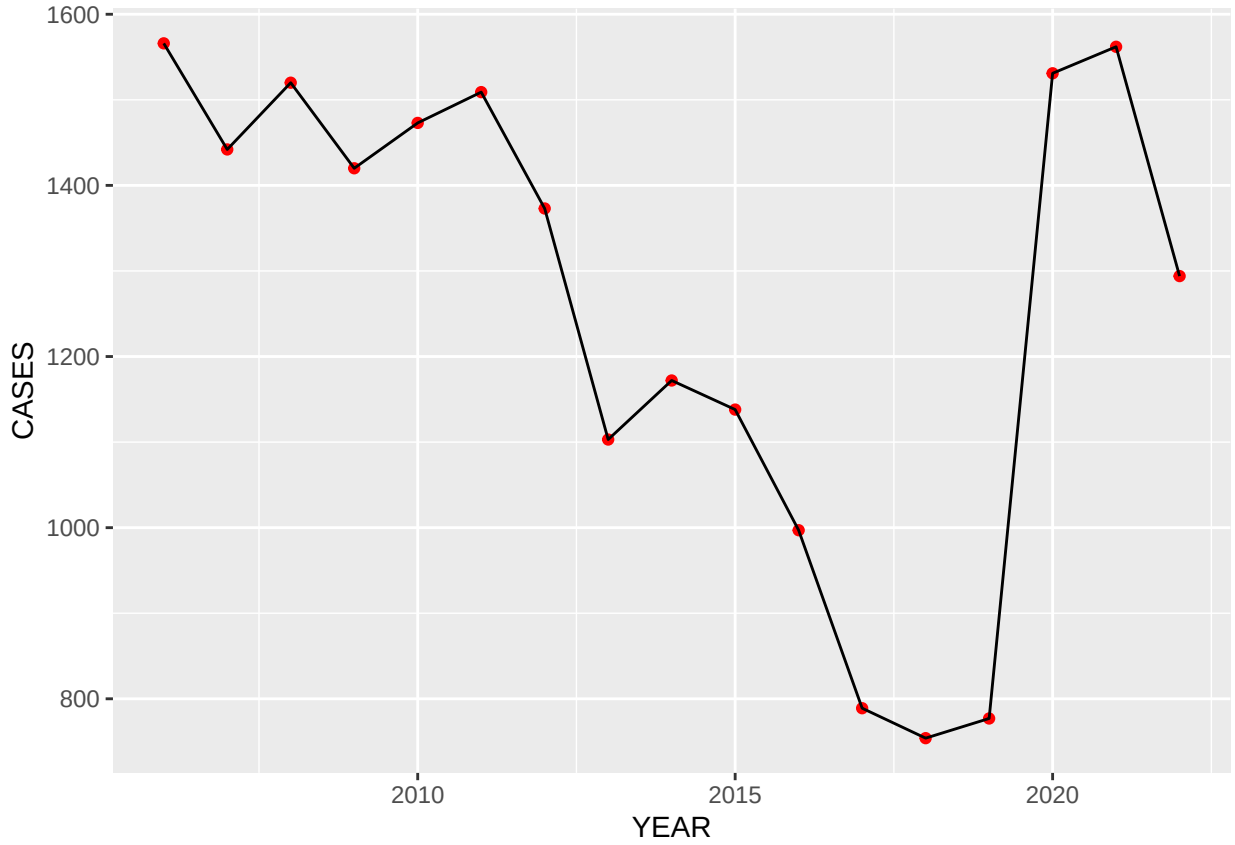
```
nypd_shooting <- nypd_shooting %>%
  select(INCIDENT_KEY:VIC_RACE) %>% #no need to include columns with coordinate information
  select(!contains("DESC")) %>% # no need to include columns with description information
  distinct(INCIDENT_KEY,.keep_all=TRUE) %>% # remove rows with duplicate incident_key
  mutate(OCCUR_DATE=mdy(OCCUR_DATE),OCCUR_TIME=hms(OCCUR_TIME)) %>% #convert to date/time format
  mutate(PERP_RACE=replace(PERP_RACE,PERP_RACE %in% c("","(null)"),"UNKNOWN")) %>% #standardize those w
  mutate(PERP_SEX=replace(PERP_SEX,PERP_SEX %in% c("","(null)"),"U")) %>% #standardize so that 'U' mean
  mutate(VIC_RACE=replace(VIC_RACE,VIC_RACE %in% c("","(null)"),"UNKNOWN")) %>%  #standardize those wit
  mutate(VIC_SEX=replace(VIC_SEX,VIC_SEX %in% c("","(null)"),"U"))  #standardize so that 'U' means sex
nypd_shooting %>% summary() #produce summary
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 64394528   1st Qu.:2009-08-02   1st Qu.:3H 28M 0S
##  Median : 91165008   Median :2013-06-14   Median :15H 5M 0S
##  Mean   :121166392   Mean   :2014-01-17   Mean   :12H 41M 9.296918767504395
##  3rd Qu.:188062788   3rd Qu.:2018-09-25   3rd Qu.:20H 43M 0S
##  Max.   :261190187   Max.   :2022-12-31   Max.   :23H 59M 0S
##
##      BORO              PRECINCT      JURISDICTION_CODE STATISTICAL_MURDER_FLAG
##  Length:21420       Min.   :  1.00   Min.   :0.0000    Length:21420
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 69.00   Median :0.0000    Mode  :character
##                     Mean   : 66.12   Mean   :0.3373
##                     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:21420       Length:21420       Length:21420       Length:21420
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE
##  Length:21420       Length:21420
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

## Step 3 - Visualizing, Analyizing and Modealing Data

### Visualizing Data - 1

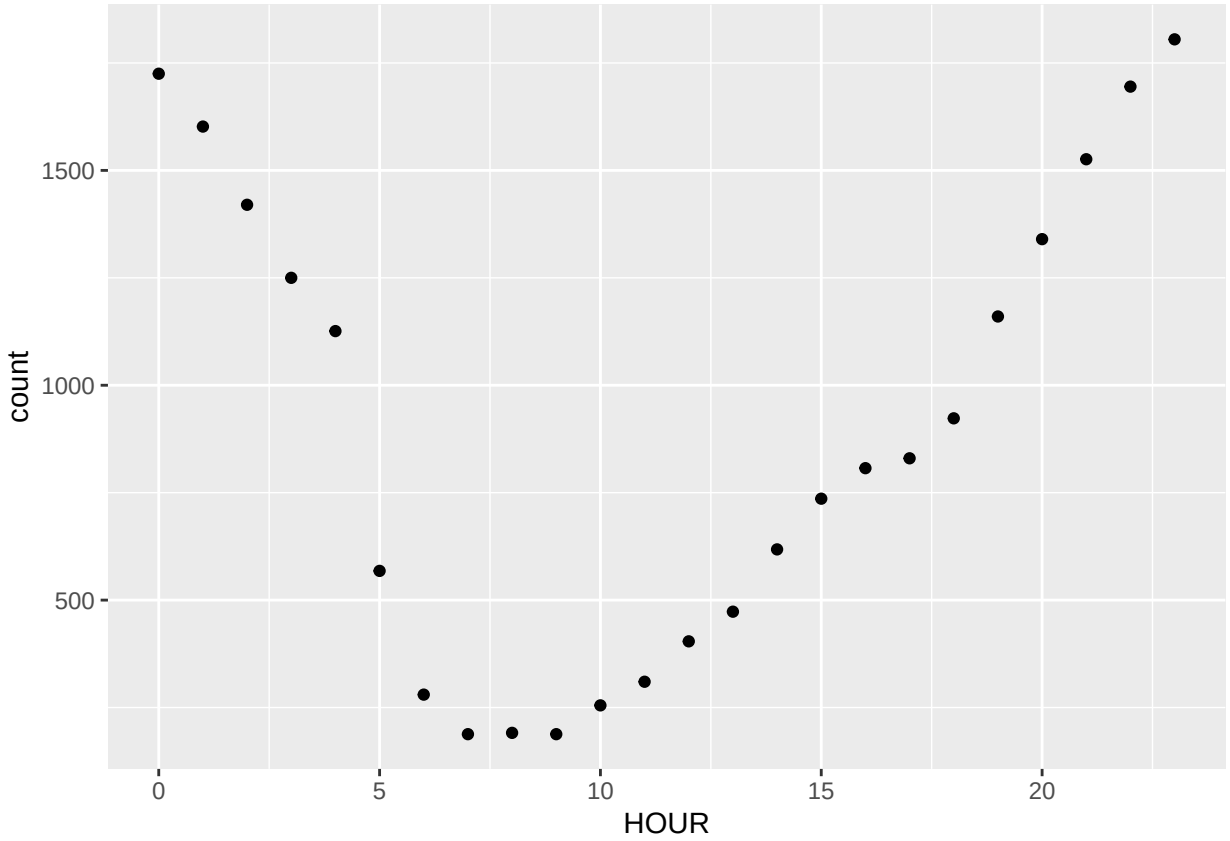
I am curious on the yearly trend of shooting cases in New York City. By visualizing the data, it seems that the year cases had gone down from 2011 til 2019. The number went up significantly from 2019 to 2020, went further up in 2021 and then dropped in 2022. The trend from 2019 can easily be linked to the COVID19 pandemic.




### Visualizing Data - 2

I'm also interested to know if number of shooting cases are higher on certain hours and lower on certain hours. The answer from the chart below suggests yes!

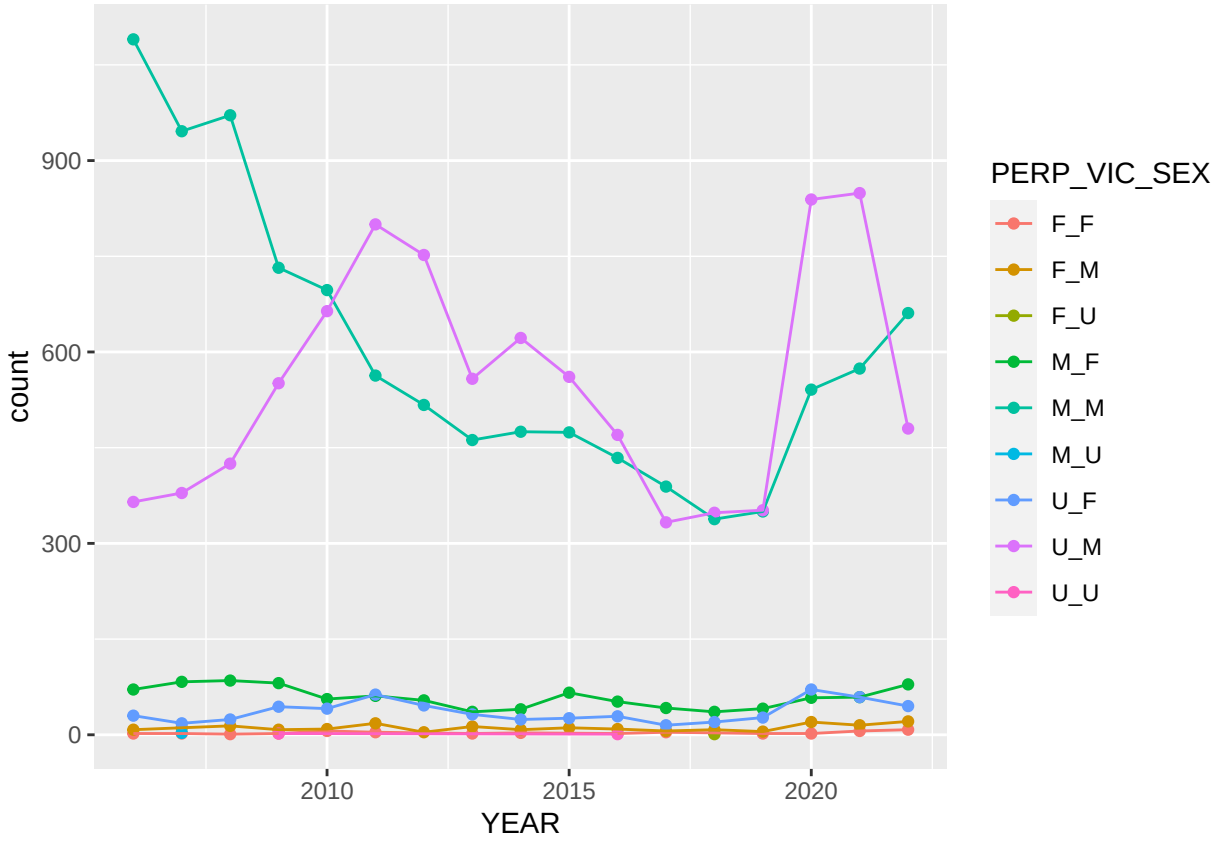
```
#get number of cases by hour
cases_by_hour <- nypd_shooting %>%
  mutate(HOUR=hour(OCCUR_TIME)) %>%
  group_by(HOUR) %>%
  summarize(count=n())

#plot the data
cases_by_hour %>%
 ggplot() +
 geom_point(aes(HOUR,count))
```




### Visualizing data - 3

I wonder a) if there is any trend in perpetrator-victim sex combination over the years and b) which combination is more prominent. Below is the answer! I know I'm being biased when it comes to gender but the result does suggest that gender matters when it comes to both perpetrator and victim! Violent men and poor men!

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




### Model - 1

I'm curious if there is any big difference in number of victims amongst races. Below codes quickly gives me the answer yes.

```
nypd_shooting %>% group_by(VIC_RACE) %>% summarise(count=n())
```

```
## # A tibble: 7 x 2
##   VIC_RACE                         count
##   <chr>                            <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE       7
## 2 ASIAN / PACIFIC ISLANDER           283
## 3 BLACK                            15632
## 4 BLACK HISPANIC                    1988
## 5 UNKNOWN                             50
## 6 WHITE                              540
## 7 WHITE HISPANIC                    2920
```

Based on the above, it appears that Black is having the highest number in victim.I'm being biased in terms of race. Hence, I would like to further find out if there is any linear relationship between number of Black victims vs total number of victims. Below codes give the answer and it's yes.

```
# a) get number of total victims by each month
mth_vic_count <- nypd_shooting %>%
  mutate(MONTH=format(as.Date(OCCUR_DATE),"%Y-%m")) %>%
  group_by(MONTH) %>%
```

```
  summarize(count=n())
mth_vic_count
```

```
## # A tibble: 204 x 2
##    MONTH   count
##    <chr>   <int>
##  1 2006-01   112
##  2 2006-02    81
##  3 2006-03    79
##  4 2006-04   113
##  5 2006-05   134
##  6 2006-06   146
##  7 2006-07   169
##  8 2006-08   177
##  9 2006-09   152
## 10 2006-10   150
## # i 194 more rows
```

```
# b) get number of Black victims by each month
mth_vic_black_count <- nypd_shooting %>%
  mutate(MONTH=format(as.Date(OCCUR_DATE),"%Y-%m")) %>%
  filter(VIC_RACE == "BLACK") %>%
  group_by(MONTH) %>%
  summarize(count_black=n())
mth_vic_black_count
```

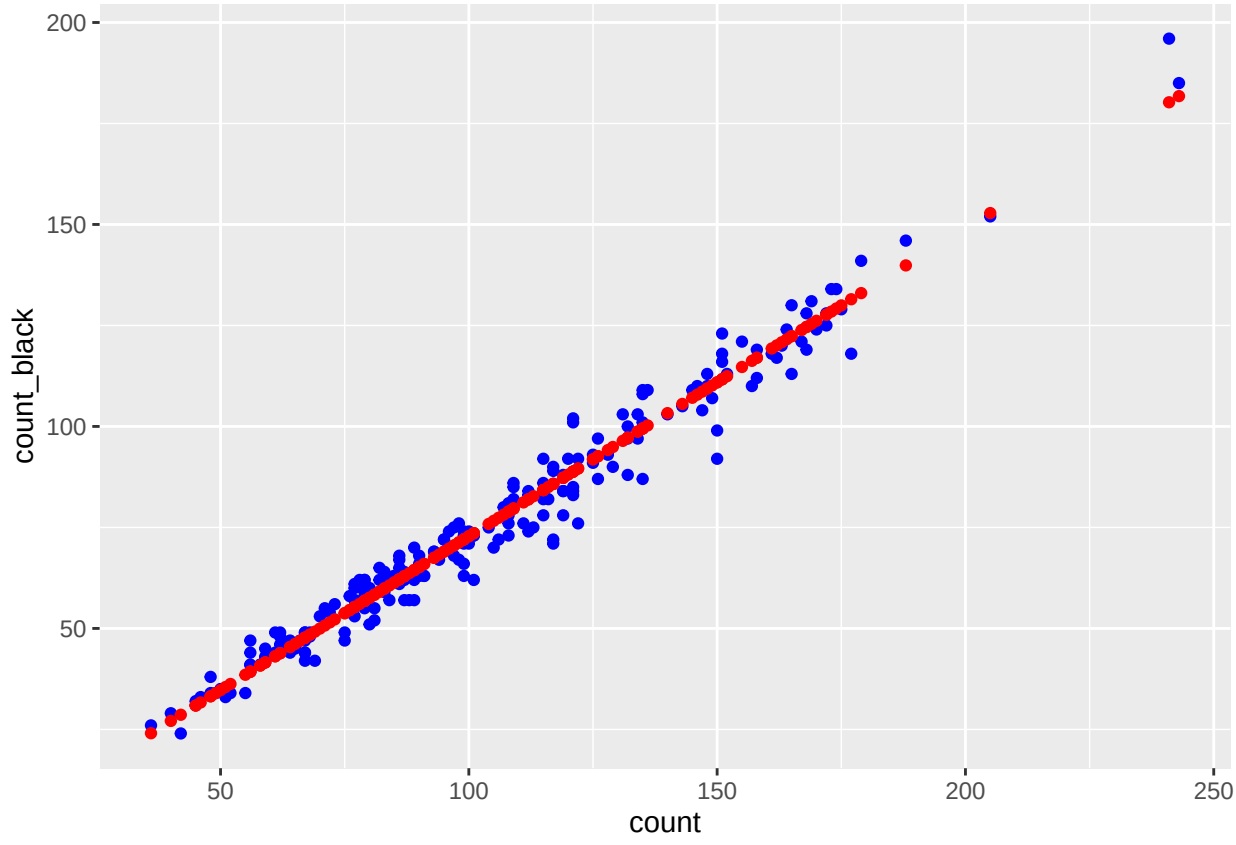
```
## # A tibble: 204 x 2
##    MONTH   count_black
##    <chr>         <int>
##  1 2006-01          74
##  2 2006-02          55
##  3 2006-03          57
##  4 2006-04          75
##  5 2006-05          97
##  6 2006-06         110
##  7 2006-07         131
##  8 2006-08         118
##  9 2006-09         113
## 10 2006-10          92
## # i 194 more rows
```

```
# join a) & b). left_join because there is a possibility that certain month there is victim but no Blac
mth_vic_all_vs_black <- left_join(mth_vic_count,mth_vic_black_count,by=join_by(MONTH))
mth_vic_all_vs_black
```

```
## # A tibble: 204 x 3
##    MONTH   count count_black
##    <chr>   <int>       <int>
##  1 2006-01   112          74
##  2 2006-02    81          55
##  3 2006-03    79          57
##  4 2006-04   113          75
```

```
##  5 2006-05   134          97
##  6 2006-06   146         110
##  7 2006-07   169         131
##  8 2006-08   177         118
##  9 2006-09   152         113
## 10 2006-10   150          92
## # i 194 more rows
```

```r
# create the model by using the lm function
mod <- lm(count_black~count,mth_vic_all_vs_black)

# create a new column for the predicted number of Black victims by using the model
mth_vic_all_vs_black <- mth_vic_all_vs_black %>% mutate(pred=predict(mod))

# plot actual number of Black victims and predicted number of Black victims as compared to total number

mth_vic_all_vs_black %>%
ggplot() +
 geom_point(aes(count,count_black),color="blue") +
 geom_point(aes(count,pred),color="red")
```




```r
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
## Running under: Windows 11 x64 (build 22621)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: Asia/Singapore
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 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] gtable_0.3.4      highr_0.10        compiler_4.3.2    tidyselect_1.2.0
##  [5] scales_1.3.0      yaml_2.3.8        fastmap_1.1.1     R6_2.5.1
##  [9] labeling_0.4.3    generics_0.1.3    knitr_1.45        munsell_0.5.0
## [13] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.2       utf8_1.2.4
## [17] stringi_1.8.3     xfun_0.41         timechange_0.2.0  cli_3.6.2
## [21] withr_2.5.2       magrittr_2.0.3    digest_0.6.33     grid_4.3.2
## [25] rstudioapi_0.15.0 hms_1.1.3         lifecycle_1.0.4   vctrs_0.6.5
## [29] evaluate_0.23     glue_1.6.2        farver_2.1.1      fansi_1.0.6
## [33] colorspace_2.1-0  rmarkdown_2.25    tools_4.3.2       pkgconfig_2.0.3
## [37] htmltools_0.5.7
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.