NYPDIncident

5/18/2022

Introduction

This data-set records shooting incidents occurred in NYC from 2006 to 2020. The data is manually extracted and reviewed by the Office of Management Analysis and Planning every quarter and is published on the NYPD website.

```
#load packages
library(tidyverse)
library(lubridate)
library(ggplot2)

#get data from website
url<- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shooting_incidents <-read.csv(url, na.strings=c("","NA"))
shooting_incidents$STATISTICAL_MURDER_FLAG<- as.logical(shooting_incidents$STATISTICAL_MURDER_FLAG)
#change column data type from chr to lgl
```

There are 23585 rows and 19 columns where each row is a shooting incident and each column can be explained as the following:
- *INCIDENT_KEY* contains a randomly generated persistent ID for each arrest.
- *OCCUR_DATE* contains the exact date of the shooting incident.
- *OCCUR_TIME* contains the exact time of the shooting incident.
- *BORO* contains the borough where the shooting incident occurred.
- *PRECINCT* contains the precinct where the shooting incident occurred.
- *JURISDICTION_CODE* contains the jurisdiction where the shooting incident occurred. 0 for Patrol; 1 for Transit; 2 for Housing; 3 for non NYPD jurisdictions.
- *LOCATION_DESC* contains the location of the shooting incident.
- *STATISTICAL_MURDER_FLAG* contains the shooting resulted in the victim's death which would be counted as murder.
- *PERP_AGE_GROUP* contains the perpetrator's age within the category.
- *PERP_SEX* contains the perpetrator's sex description.
- *VIC_AGE_GROUP* contains the victim's age within a category.
- *VIC_SEX* contains the victim's sex description.
- *X_COORD_CD* contains the mid block X-coordinate for the New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet.
- *Y_COORD_CD* contains the mid block Y-coordinate for the New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet.
- *Latitude* contains the latitude coordinate for Global Coordinate System,decimal degrees.
- *Longitude* contains the longitude coordinate for Global Coordinate System,decimal degrees.
- *Lon_Lat* contains the longitude and latitude coordinates for mapping.

Tidying and Transforming Data

Tidy the data set first by removing *INCIDENT_KEY*, *LOCATION_DESC*, *X_COORD_CD*, *Y_COORD_CD*, *Latitude*, *Longitude*, and *Lon_Lat* since they are not required in the process of visualizing, analyzing, and modeling data.

```
shooting_incidents<- shooting_incidents%>%
  mutate(OCCUR_DATE=mdy(OCCUR_DATE)) %>% #coerce date into correct type
  select(-c(INCIDENT_KEY, LOCATION_DESC), -(X_COORD_CD:Lon_Lat))
```

Check for missing data.

```
sum(is.na(shooting_incidents$OCCUR_DATE))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$OCCUR_TIME))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$BORO))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$PRECINCT))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$JURISDICTION_CODE))
```

```
## [1] 2
```

```
sum(is.na(shooting_incidents$STATISTICAL_MURDER_FLAG))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$PERP_AGE_GROUP))
```

```
## [1] 8295
```

```
sum(is.na(shooting_incidents$PERP_SEX))
```

```
## [1] 8261
```

```
sum(is.na(shooting_incidents$PERP_RACE))
```

```
## [1] 8261
```

```
sum(is.na(shooting_incidents$VIC_AGE_GROUP))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$VIC_SEX))
```

```
## [1] 0
```

```
sum(is.na(shooting_incidents$VIC_RACE))
```

```
## [1] 0
```

Since there are only 2 missing values in *JURISDICTION_CODE*, we can safely remove these 2 incidents since it would not change the data significantly. However, with 8295 missing values in *PERP_AGE_GROUP*, 8261 missing values in *PERP_SEX*, and 8261 missing values in *PERP_RACE*, the missing values cannot be remove as it will change the results significantly. Therefore, we will remove *PERP_AGE_GROUP*, *PERP_SEX*, *PERP_RACE* columns and not use them in data analysis.

```
shooting_incidents <- shooting_incidents[!is.na(shooting_incidents$JURISDICTION_CODE), ]
shooting_incidents <- shooting_incidents %>%
  select(-(PERP_AGE_GROUP:PERP_RACE))
```

Data Analysis

Shooting cases and deaths by boro.

We are interested to see the number of cases, deaths, proportion of cases by each boro to total cases, and whether there are some boro shootings more likely to result in death?

```
boro_deaths_rate <- shooting_incidents %>%
group_by(BORO) %>%
summarize(cases= n(), deaths=sum(STATISTICAL_MURDER_FLAG, na.rm=TRUE)) %>%
mutate(cases_prop= round(cases/nrow(shooting_incidents), 4)) %>%
mutate(deaths_rate= round(deaths/cases, 4)) %>%
arrange(desc(cases)) %>%
ungroup()
boro_deaths_rate
```

```
## # A tibble: 5 x 5
##   BORO          cases deaths cases_prop deaths_rate
##   <chr>         <int>  <int>      <dbl>       <dbl>
## 1 BROOKLYN       9734   1898     0.413        0.195
## 2 BRONX          6701   1247     0.284        0.186
## 3 QUEENS         3531    697     0.150        0.197
## 4 MANHATTAN      2921    515     0.124        0.176
## 5 STATEN ISLAND   696    143     0.0295       0.206
```

Top 5 shooting cases by precincts.

We are interested to see the number of cases, deaths, proportion of cases in each of top 5 precincts to total cases, and whether there are some precincts shootings more likely to result in death?

```
precinct_deaths_rate <- shooting_incidents %>%
  group_by(PRECINCT) %>%
  summarize(cases= n(), deaths=sum(STATISTICAL_MURDER_FLAG, na.rm=TRUE)) %>%
  mutate(cases_prop= round(cases/nrow(shooting_incidents), 4)) %>%
  mutate(deaths_rate= round(deaths/cases, 4)) %>%
  arrange(desc(cases)) %>%
  top_n(5) %>%
  ungroup()
```

```
## Selecting by deaths_rate
```

```
precinct_deaths_rate
```

```
## # A tibble: 5 x 5
##   PRECINCT cases deaths cases_prop deaths_rate
##      <int> <int>  <int>      <dbl>       <dbl>
## 1      106   185     60     0.0078       0.324
## 2      122    58     24     0.0025       0.414
## 3        1    22      7     0.0009       0.318
## 4      112    19      7     0.0008       0.368
## 5       17     6      2     0.0003       0.333
```

Shooting cases and deaths by jurisdiction.

We are interested to see the number of cases, deaths, proportion of cases by jurisdiction code, and whether there are some juridistion code shootings more likely to result in death?

```
jurisdiction_deaths_rate <- shooting_incidents %>%
  group_by(JURISDICTION_CODE) %>%
  summarize(cases= n(), deaths=sum(STATISTICAL_MURDER_FLAG, na.rm=TRUE)) %>%
  mutate(cases_prop= round(cases/nrow(shooting_incidents), 4)) %>%
  mutate(deaths_rate=round(deaths/cases, 4)) %>%
  arrange(desc(cases)) %>%
  ungroup()
jurisdiction_deaths_rate
```

```
## # A tibble: 3 x 5
##   JURISDICTION_CODE cases deaths cases_prop deaths_rate
##               <int> <int>  <int>      <dbl>       <dbl>
## 1                 0 19629   3883     0.832        0.198
## 2                 2  3900    605     0.165        0.155
## 3                 1    54     12     0.0023       0.222
```

Shooting cases and deaths by victims age group.

We are interested to see the number of cases, deaths, proportion of cases by victims age to total cases, and whether there are some age group shootings more likely to result in death?

```
victimage_deaths_rate<-shooting_incidents %>%
  group_by(VIC_AGE_GROUP) %>%
  summarize(cases= n(), deaths=sum(STATISTICAL_MURDER_FLAG, na.rm=TRUE)) %>%
  mutate(cases_prop= round(cases/nrow(shooting_incidents), 4)) %>%
  mutate(deaths_rate=round(deaths/cases, 4)) %>%
  arrange(desc(cases)) %>%
  ungroup()
victimage_deaths_rate
```

```
## # A tibble: 6 x 5
##   VIC_AGE_GROUP cases deaths cases_prop deaths_rate
##   <chr>         <int>  <int>      <dbl>       <dbl>
## 1 25-44         10302   2257     0.437        0.219
## 2 18-24          9002   1466     0.382        0.163
## 3 <18            2525    320     0.107        0.127
## 4 45-64          1541    390     0.0653       0.253
## 5 65+             154     52     0.0065       0.338
## 6 UNKNOWN          59     15     0.0025       0.254
```

Shooting cases and deaths by victims sex.

We are interested to see the number of cases, deaths, proportion of cases by victims sex to total cases, and whether there is a shooting that is more likely to result death of a sex more than another.

```
victimsex_deaths_rate <- shooting_incidents %>%
group_by(VIC_SEX) %>%
summarize(cases= n(), deaths=sum(STATISTICAL_MURDER_FLAG, na.rm=TRUE)) %>%
mutate(cases_prop= round(cases/nrow(shooting_incidents), 4)) %>%
mutate(deaths_rate=round(deaths/cases, 4)) %>%
arrange(desc(cases)) %>%
ungroup()
victimsex_deaths_rate
```

```
## # A tibble: 3 x 5
##   VIC_SEX cases deaths cases_prop deaths_rate
##   <chr>   <int>  <int>      <dbl>       <dbl>
## 1 M       21368   4061     0.906       0.190
## 2 F        2204    438     0.0935      0.199
## 3 U          11      1     0.0005      0.0909
```

Shooting cases and deaths by victims race.

We are interested to see the number of cases, deaths, proportion of cases by races to total cases, and whether there are some races shootings more likely result in death?

```
victimrace_deaths_rate <- shooting_incidents %>%
group_by(VIC_RACE) %>%
summarize(cases= n(), deaths=sum(STATISTICAL_MURDER_FLAG, na.rm=TRUE)) %>%
  mutate(cases_prop= round(cases/nrow(shooting_incidents), 4)) %>%
mutate(deaths_rate=round(deaths/cases, 4)) %>%
```

```
arrange(desc(cases)) %>%
ungroup()
victimrace_deaths_rate
```

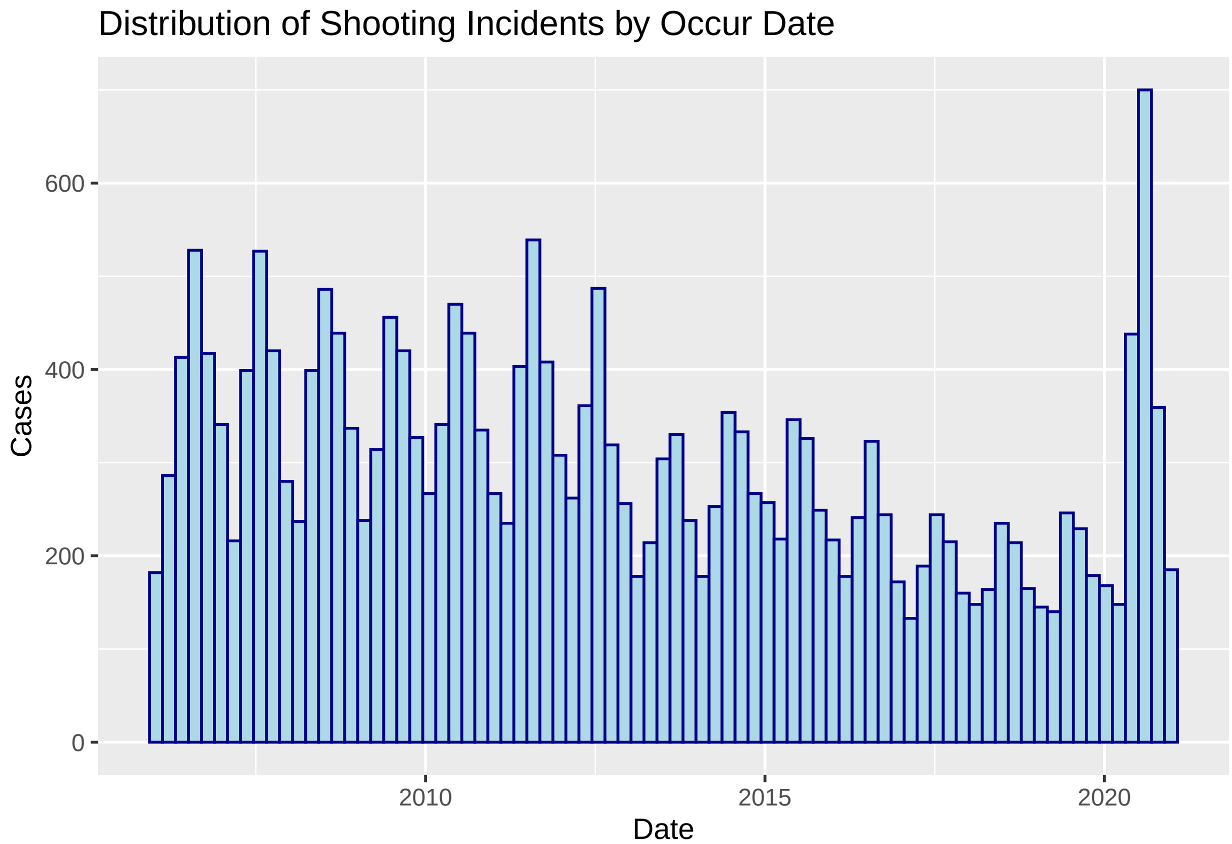
```
## # A tibble: 7 x 5
##   VIC_RACE                        cases deaths cases_prop deaths_rate
##   <chr>                           <int>  <int>      <dbl>       <dbl>
## 1 BLACK                           16868   3155     0.715        0.187
## 2 WHITE HISPANIC                   3449    725     0.146        0.210
## 3 BLACK HISPANIC                   2245    352     0.0952       0.157
## 4 WHITE                             620    178     0.0263       0.287
## 5 ASIAN / PACIFIC ISLANDER          327     83     0.0139       0.254
## 6 UNKNOWN                            65      7     0.0028       0.108
## 7 AMERICAN INDIAN/ALASKAN NATIVE      9      0     0.0004       0
```

Visualize data

Distribution of Shooting Incidents by Occur Date

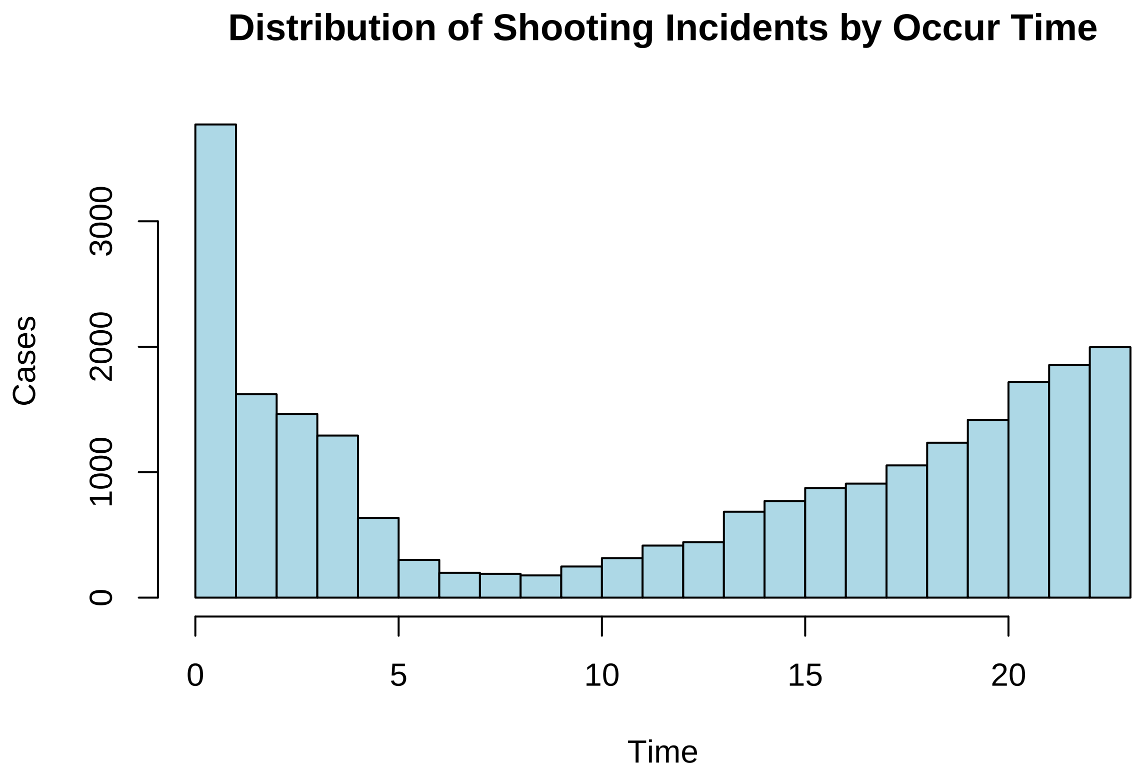
We are interested to visualize the distribution of shooting incident by occur date using histogram to see whether there are a trend in which months shootings occur the most.

```
date_histogram<-shooting_incidents %>%
ggplot(aes(x= OCCUR_DATE)) +
geom_histogram(binwidth = 70, color="darkblue", fill="lightblue") +
labs(title= "Distribution of Shooting Incidents by Occur Date", x="Date", y="Cases")
date_histogram
```



##Distribution of Shooting Incidents by Occur Time

```
hist(x=as.numeric(substr(shooting_incidents$OCCUR_TIME, 1,2)), breaks=0:23, main="Distribution of Shooti
```



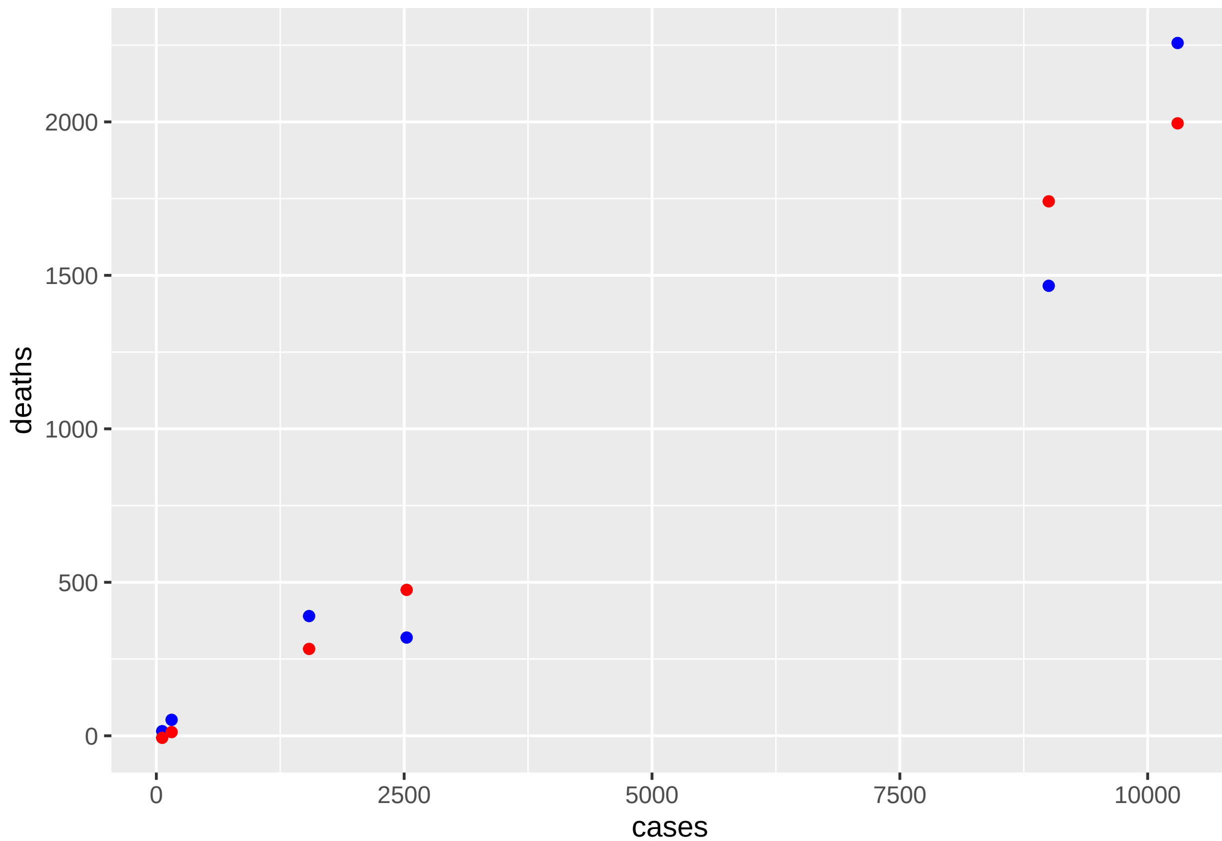
Model victim age group

```
mod <- lm(deaths~cases, data = victimage_deaths_rate)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = victimage_deaths_rate)
##
## Residuals:
##       1       2       3       4       5       6
##  261.77 -275.16 -155.31  107.00   40.07   21.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.16401  119.92418  -0.151  0.88694
## cases         0.19544    0.02099   9.313  0.00074 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.2 on 4 degrees of freedom
## Multiple R-squared:  0.9559, Adjusted R-squared:  0.9449
## F-statistic: 86.73 on 1 and 4 DF,  p-value: 0.0007398
```

```
victimage_w_pred <- victimage_deaths_rate %>% mutate(pred = predict(mod))
victimage_w_pred %>% ggplot() +
```

```
geom_point(aes(x= cases, y= deaths), color= "blue") +
geom_point(aes(x=cases, y=pred),color="red")
```



Conclusion

In conclusion, the data shows that some of the top shootings happened in Brooklyn, precinct 106, and patrol jurisdiction; victims aged 25-44, male, and black get involved in the most shootings incidents in New York. In the data visualization section, we can see that shooting incidents occur the most in 2020, and during midnight, the least shootings occur between 5 am and 10 am. There is clear indication that cases are indication for deaths where the actual cases and deaths closely follows the predicted model for victim age group.

Bias

My personal bias in regard to shooting incidents is that they would occur the most during the night when most people are asleep, to mitigate my personal bias by looking at the shooting incidents occur by time to see when shooting incidents occur and it was found that indeed, most of the shooting incidents occur at night, and peaks at midnight.