Covid 19 Data Report

Tsuyoshi Akiyama

2022-06-01

Abstract

This report's purpose is to analyze the time-series covid 19 data, including confirmed cases and deaths. The report follows the 6 steps, prerequisite, loading, tidy, visualizing, modeling, and considering biases. The main research is to find correlaition between them.

Introduction

Prerequisite

- clear the current environment variables
- install necessary libraries

```
# Clear all variables
rm(list=ls())
# Dependencies
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Load the data

Get every shooting incident data in NYC from 2006 to 2020.

```
source_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
# Get the data
source_data <- read_csv(source_url)
```

See the data's dimension, spec, and summary.

```
dim(source_data)
# See original data specification
spec(source_data)
summary(source_data)
```

Tidy the data

Clean

- Convert characterized date, OCCUR_DATE and OCCUR_TIME, to lubridate:datetime format.
- Select necessary columns for later analysis.
- Change gender label, from (M, F, U) to (Male, Female, Unknown)

```
shooting_incidents <- source_data %>%
  select(OCCUR_DATE, OCCUR_TIME, VIC_AGE_GROUP, VIC_SEX) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_SEX = factor(VIC_SEX,
                          levels = c("M", "F", "U"),
                          labels = c("Male", "Female", "Unknown")))
```

Convert the data

The 4 converted data will be used at visualization step.

```
shooting_incidents_per_year <- shooting_incidents %>%
  mutate(OCCUR_DATE = year(OCCUR_DATE)) %>%
  group_by(OCCUR_DATE) %>%
  summarise(n = n()) %>%
  arrange(OCCUR_DATE) %>%
  ungroup()

shooting_incidents_per_hour <- shooting_incidents %>%
  mutate(OCCUR_TIME = hour(OCCUR_TIME)) %>%
  group_by(OCCUR_TIME) %>%
  summarise(n = n()) %>%
  arrange(OCCUR_TIME) %>%
  ungroup()

shooting_incidents_per_vict <- shooting_incidents %>%
  filter(VIC_AGE_GROUP != "UNKNOWN") %>%
  group_by(VIC_AGE_GROUP, VIC_SEX) %>%
  summarise(n = n()) %>%
  arrange(VIC_AGE_GROUP) %>%
  ungroup()

shooting_incidents_per_date_gender <- shooting_incidents %>%
  group_by(OCCUR_DATE, VIC_SEX) %>%
  summarise(n = n()) %>%
  ungroup()
```

`shooting_incidents_per_gender` will be used at model step.

```
shooting_incidents_per_gender <- shooting_incidents %>%
  group_by(OCCUR_DATE, VIC_SEX) %>%
  summarise(n = n()) %>%
```

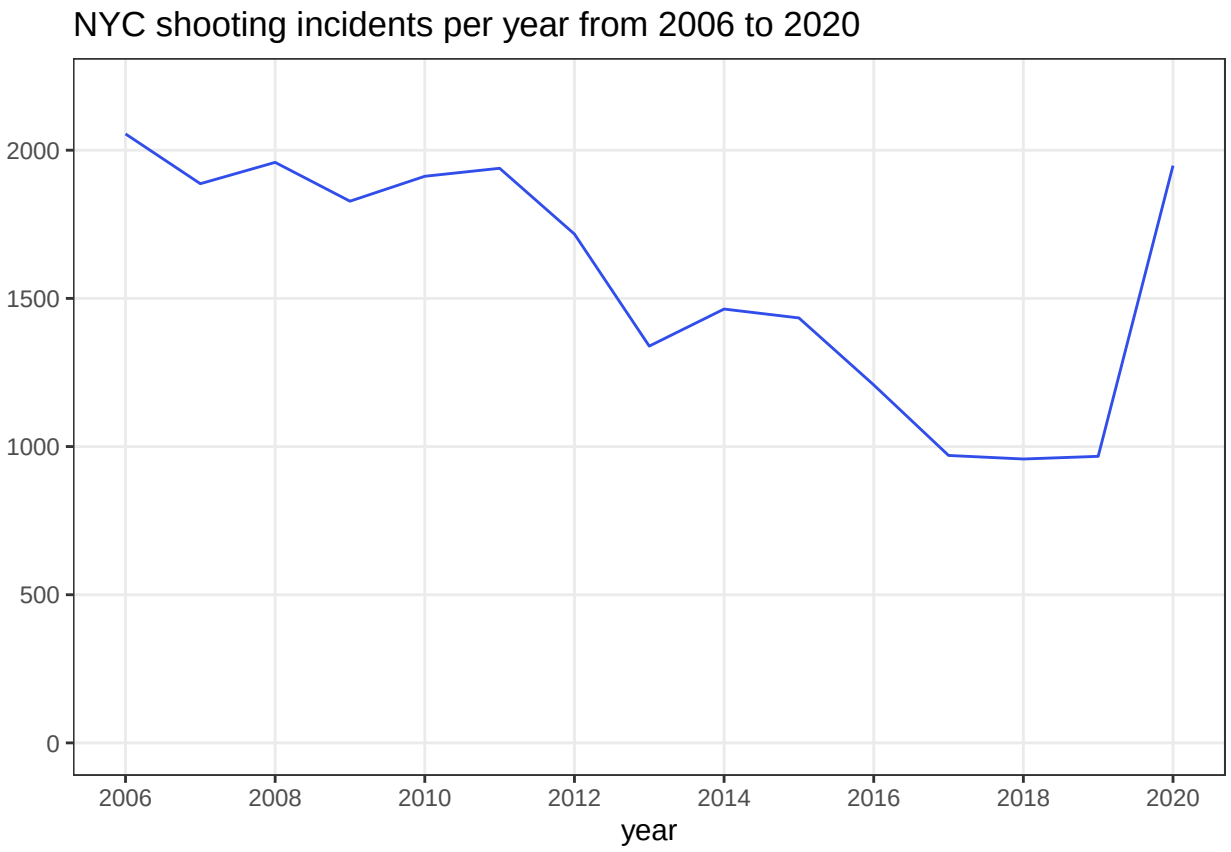
```
  spread(key = VIC_SEX, value = n) %>%
  mutate(Female = replace_na(Female, 0),
         Male = replace_na(Male, 0)) %>%
  ungroup()
```

Visualize the data

Incidents number per year

The shooting incident per year had decreased yearly from 2006 to 2019, 2000 to 1000 incidents/year. However, in 2020, the incidents number nearly doubled as 2019.

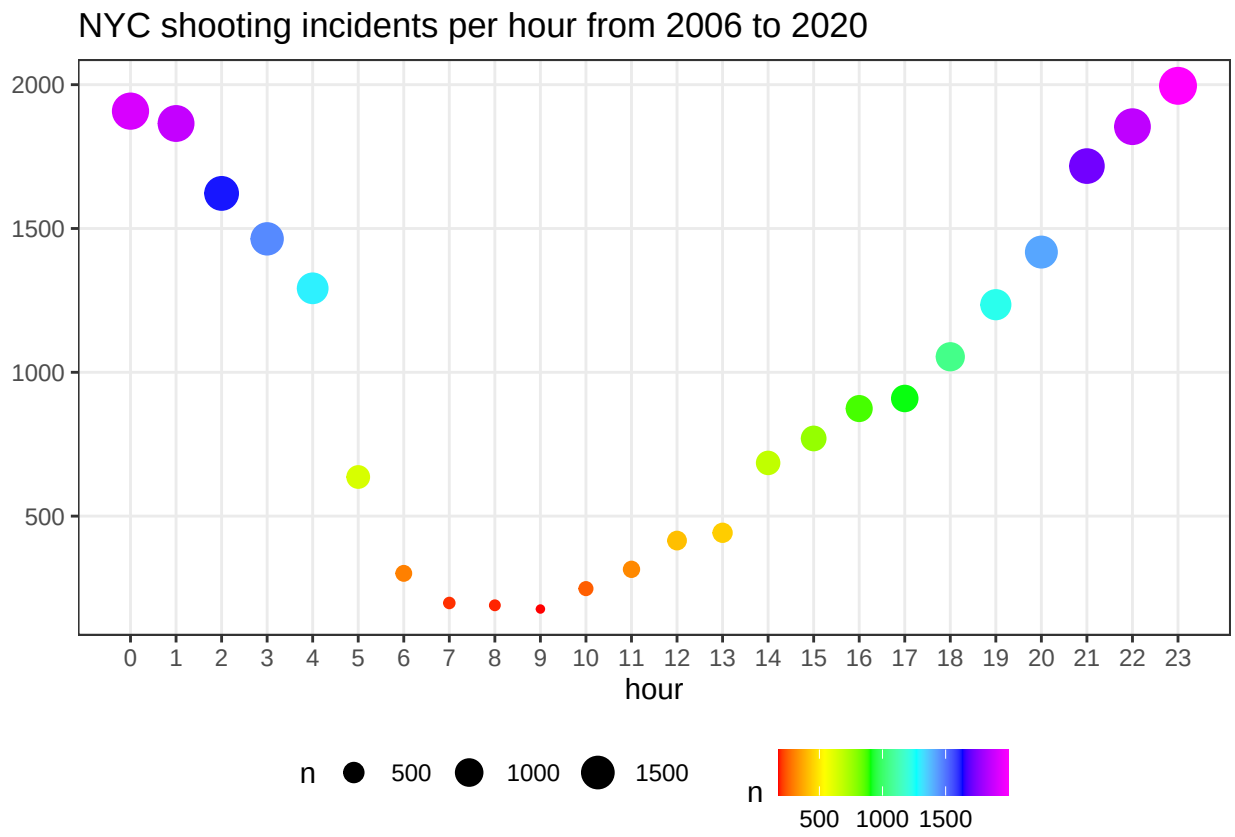
```
shooting_incidents_per_year %>%
  ggplot(aes(x = OCCUR_DATE, y = n)) +
  geom_line(color='#314DEA') +
  scale_x_continuous(breaks = seq(2006, 2020, by = 2)) +
  theme_bw() +
  theme(legend.position = "bottom", panel.grid.minor = element_blank()) +
  labs(title = "NYC shooting incidents per year from 2006 to 2020", y = NULL, x = "year") +
  ylim(0, 2200)
```



Incidents number per hour

Peaking at 23:00, the number of crimes decreases until 8:00.

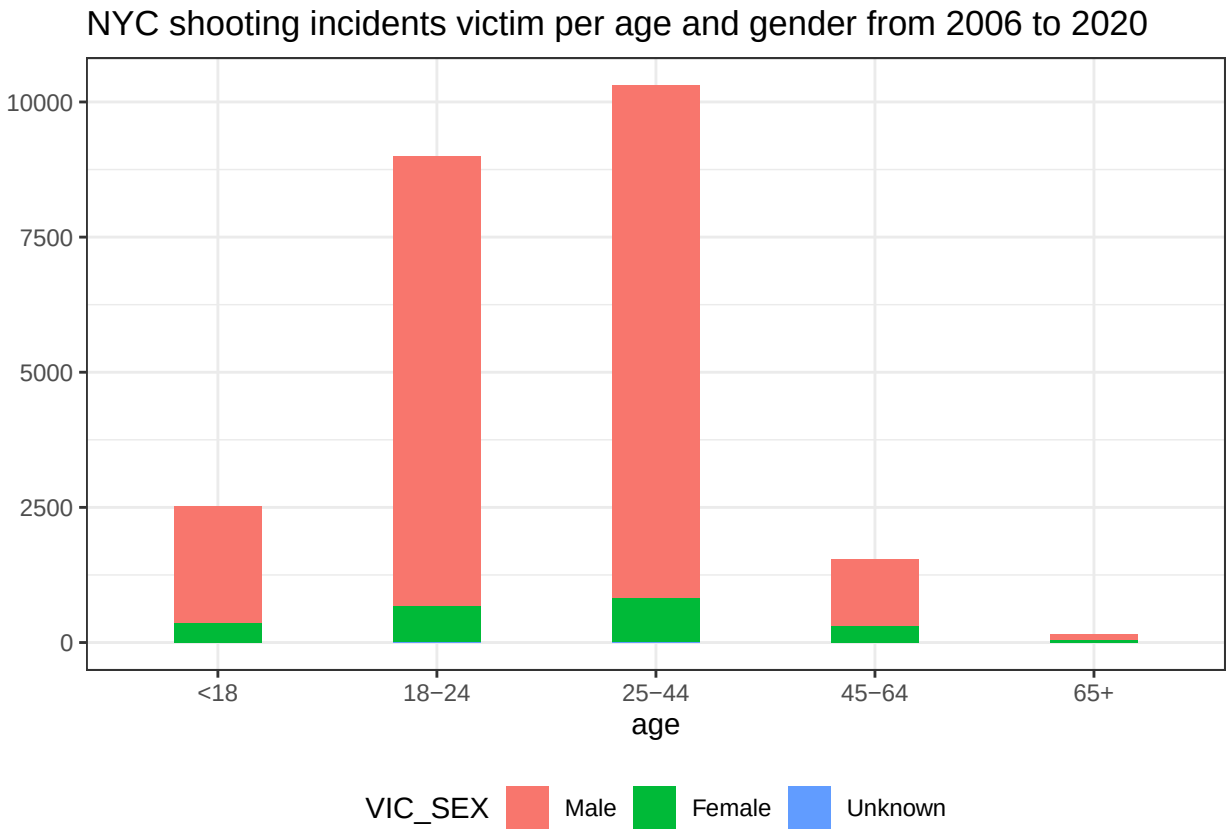
```
shooting_incidents_per_hour %>%
  ggplot(aes(x = OCCUR_TIME, y = n, size = n, color = n)) +
  geom_point() +
  scale_color_gradientn(colors = rainbow(6))+
  scale_x_continuous(breaks = seq(0, 23, by = 1)) +
  theme_bw() +
  theme(legend.position = "bottom", panel.grid.minor = element_blank()) +
  labs(title = "NYC shooting incidents per hour from 2006 to 2020", y = NULL, x = "hour")
```



Incidents victim per age and gender

The male victims account for a large portion of total victims at any ages. 25-44 years old has the most high possibility to be victim, while over 65 has the lowest of that.

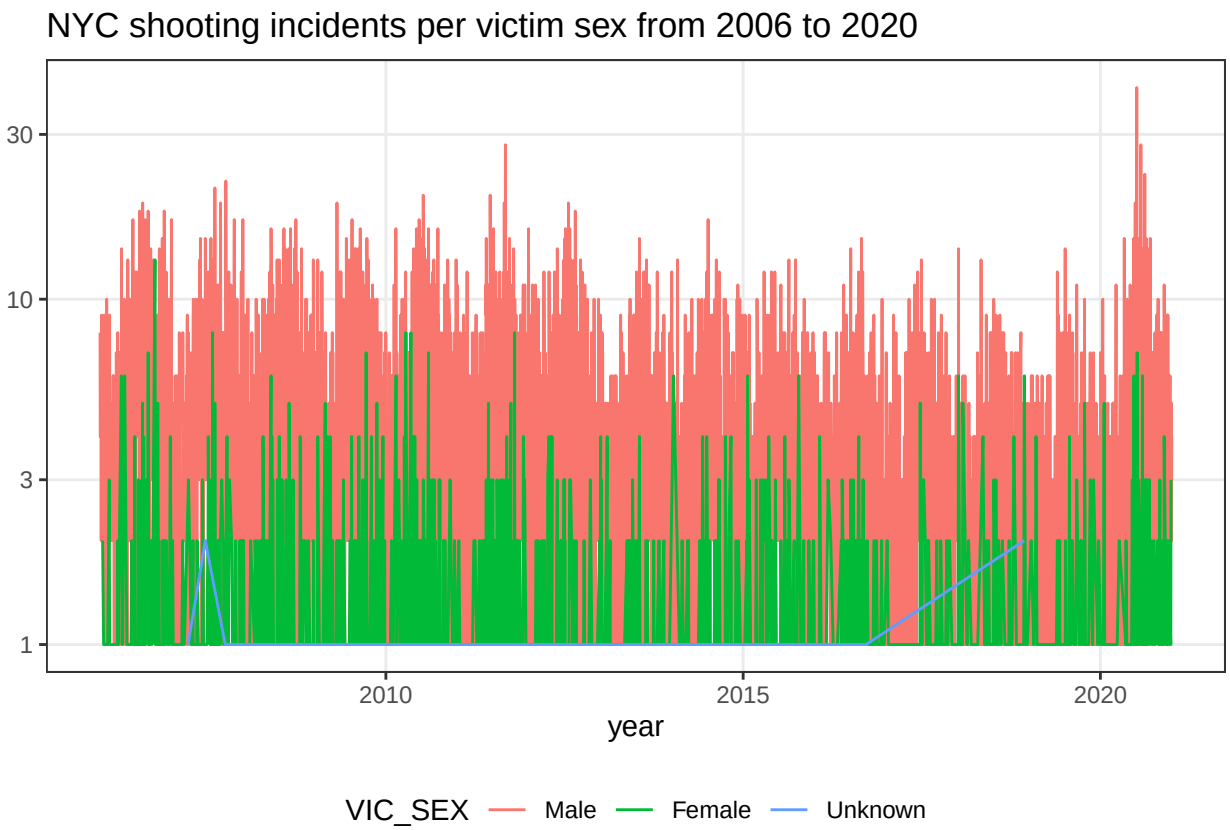
```
shooting_incidents_per_vict %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = n, fill = VIC_SEX)) +
  geom_bar(stat = "identity", width = 0.4) +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "NYC shooting incidents victim per age and gender from 2006 to 2020", y = NULL, x = "age
```



Incidents victim per gender

Incidents number by victim gender shows on time series, from 2006 to 2020. While male victim number mostly is higher than female's, the both seem to correlate each other.

```
shooting_incidents_per_date_gender %>%
  ggplot(aes(x = OCCUR_DATE, y = n, color = VIC_SEX)) +
  geom_line() +
  scale_y_log10() +
  theme_bw() +
  theme(legend.position = "bottom", panel.grid.minor = element_blank()) +
  labs(title = "NYC shooting incidents per victim sex from 2006 to 2020", y = NULL, x = "year")
```



Model the data

Hypothesis

There are correlation between male and female victim. If so, the more male victims increase, the more female victims increase.

Make the model

Utilize linear model to `shooting_incidents_per_gender` to get the statistical relationships between male and female victim.

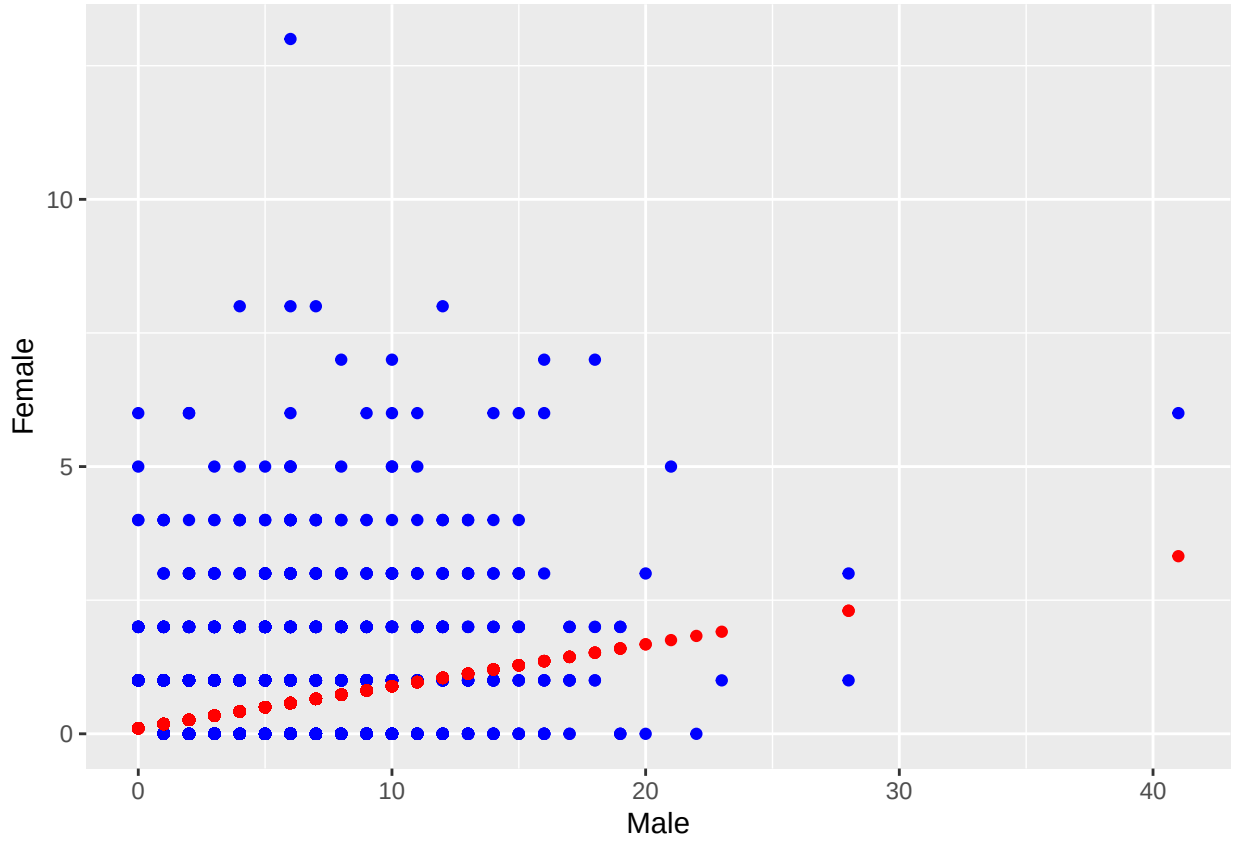
```
model <- lm(Female ~ Male, data = shooting_incidents_per_gender)
summary(model)
```

Visualize model's prediction

See the prediction to see the relationship. Blue points show the actual data, while red points are the prediction.

```
shooting_incidents_per_gender %>%
  mutate(pred = predict(model)) %>%
  ggplot() +
```

```
  geom_point(aes(x = Male, y = Female), color = 'blue') +
  geom_point(aes(x = Male, y = pred), color = 'red')
```



Assess prediciton

The slope of prediction line is small so there is a weak correlation between male and female victim.

Possible biases

One possible bias is a distribution bias by gender. The most data are the male victims under 44 years old. I need to gather more female data, or normalize the data to ease the one sided distribution.

Another bias is variable selection bias. I use `VIC_SEX` column to analyse the relationship between male and female victim. The selected variable is not good enough to see the expected results. For instance, the incidents number in 2020 nearly doubled as 2019, but any variables in the dataset answer this change. There might be another factor, outside of datasets, to affect shooting incidents in NYC.