NYPD Shooting Incident Data Project

E. Song

2022-05-12

NYPD Shooting Incident Data

Step 1. Start an Rmd Document

This is an Rmd document that imports and analyzes the shooting incident data set. Through this data analysis, I will check whether there is a difference in the number of shootings in New York by borough and the trend by year. The data is from the website https://catalog.data.gov/dataset. It is a list of shooting incident that occurred in New York City going back to 2006 through the end of 2020.

```
library(htmltools)
library(readr)

shooting <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
```

```
## Rows: 23585 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Step 2. Tidy and Transform Your Data

```
library(dplyr)
library(lubridate)
```

```
shooting <- shooting %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

```
shooting <- shooting %>%
  select(c(OCCUR_DATE,BORO,STATISTICAL_MURDER_FLAG))
shooting$YEAR <- as.Date(cut(shooting$OCCUR_DATE,
  breaks = "year"))
```

```
shooting
```

```
## # A tibble: 23,585 x 4
##    OCCUR_DATE BORO     STATISTICAL_MURDER_FLAG YEAR
##    <date>     <chr>    <lgl>                   <date>
##  1 2006-08-27 BRONX    TRUE                    2006-01-01
##  2 2011-03-11 QUEENS   FALSE                   2011-01-01
##  3 2019-10-06 BROOKLYN FALSE                   2019-01-01
##  4 2011-09-04 BRONX    FALSE                   2011-01-01
##  5 2013-05-27 QUEENS   FALSE                   2013-01-01
##  6 2013-09-01 BROOKLYN FALSE                   2013-01-01
##  7 2010-06-05 BROOKLYN FALSE                   2010-01-01
##  8 2020-03-20 BROOKLYN FALSE                   2020-01-01
##  9 2014-07-04 QUEENS   FALSE                   2014-01-01
## 10 2015-10-18 QUEENS   FALSE                   2015-01-01
## # ... with 23,575 more rows
```

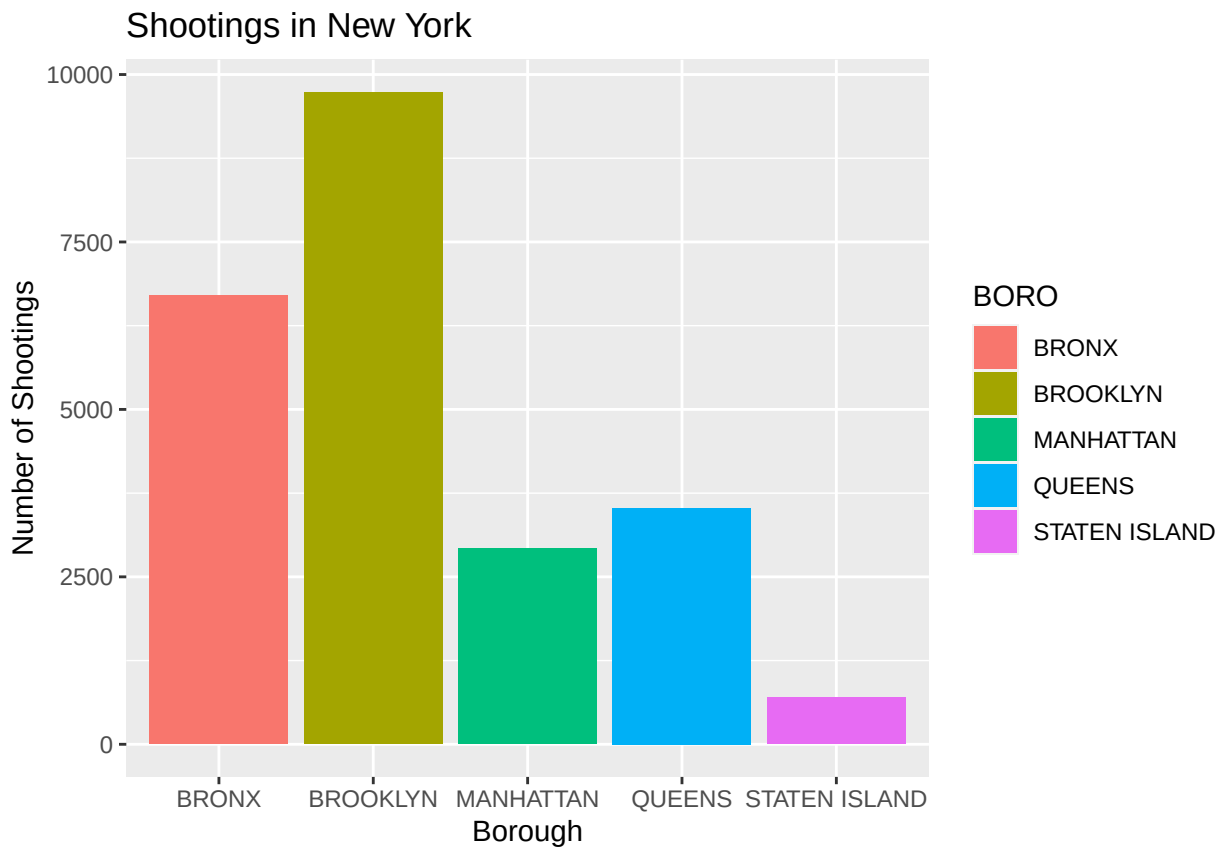
```
summary(shooting)
```

```
##    OCCUR_DATE             BORO           STATISTICAL_MURDER_FLAG
##  Min.   :2006-01-01   Length:23585       Mode :logical
##  1st Qu.:2008-12-31   Class :character   FALSE:19085
##  Median :2012-02-27   Mode  :character   TRUE :4500
##  Mean   :2012-10-05
##  3rd Qu.:2016-03-02
##  Max.   :2020-12-31
##       YEAR
##  Min.   :2006-01-01
##  1st Qu.:2008-01-01
##  Median :2012-01-01
##  Mean   :2012-03-26
##  3rd Qu.:2016-01-01
##  Max.   :2020-01-01
```

Step 3. Add Visualizations and Analysis

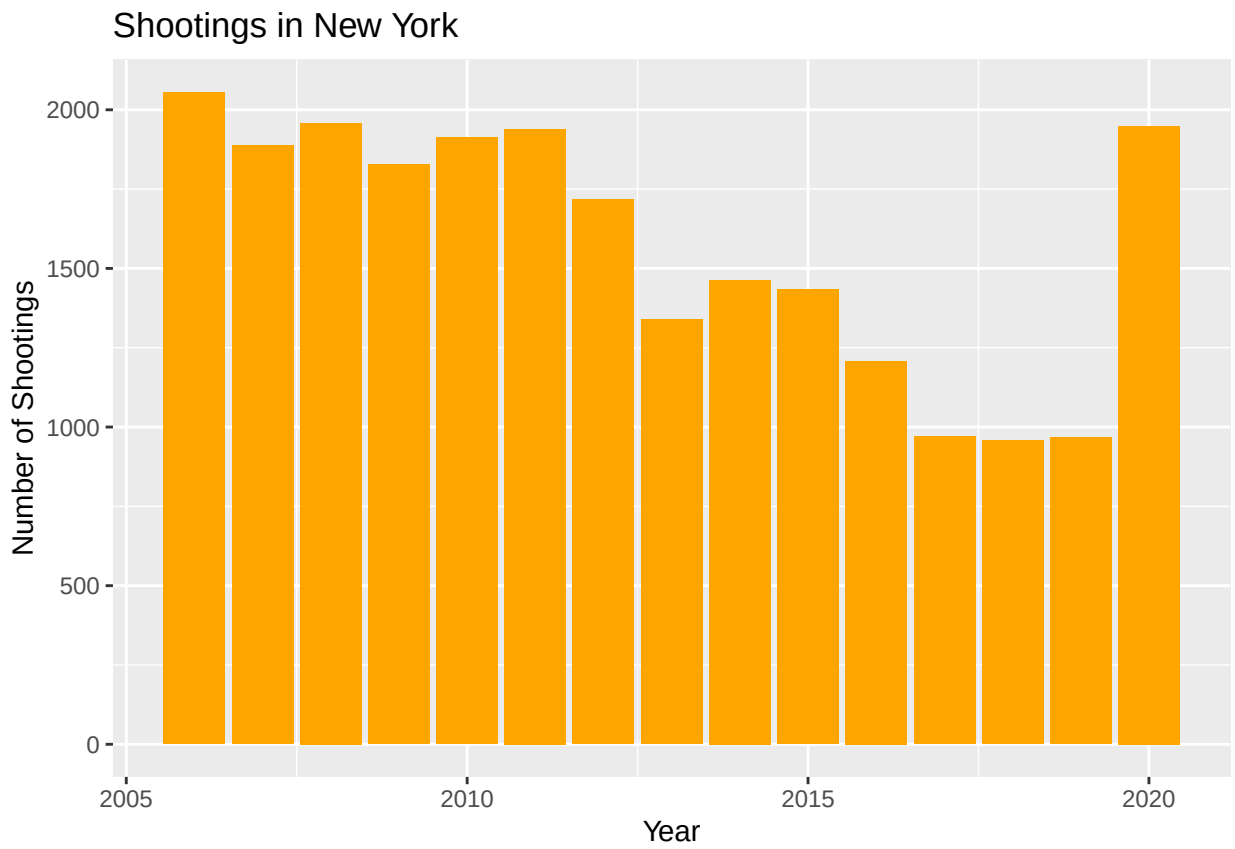
```
library(ggplot2)

s_boro <- ggplot(data=shooting, aes(x=BORO,fill=BORO))+geom_bar()
s_boro+ggtitle("Shootings in New York")+xlab("Borough")+ylab("Number of Shootings")
```



From the beginning of 2006 to the end of 2020, the most common shootings in New York City occurred in Brooklyn, followed by the Bronx, Queens, Manhattan and Staten Island. Brooklyn's shootings were over ten times more common than on Staten Island.

```
s_year <- ggplot(data=shooting, aes(x=YEAR))+geom_bar(fill='orange')
s_year+ggtitle("Shootings in New York")+xlab("Year")+ylab("Number of Shootings")
```



From 2006 to 2020, the year with the highest number of shooting incidents in New York City was 2006, and the year with the lowest was 2018.

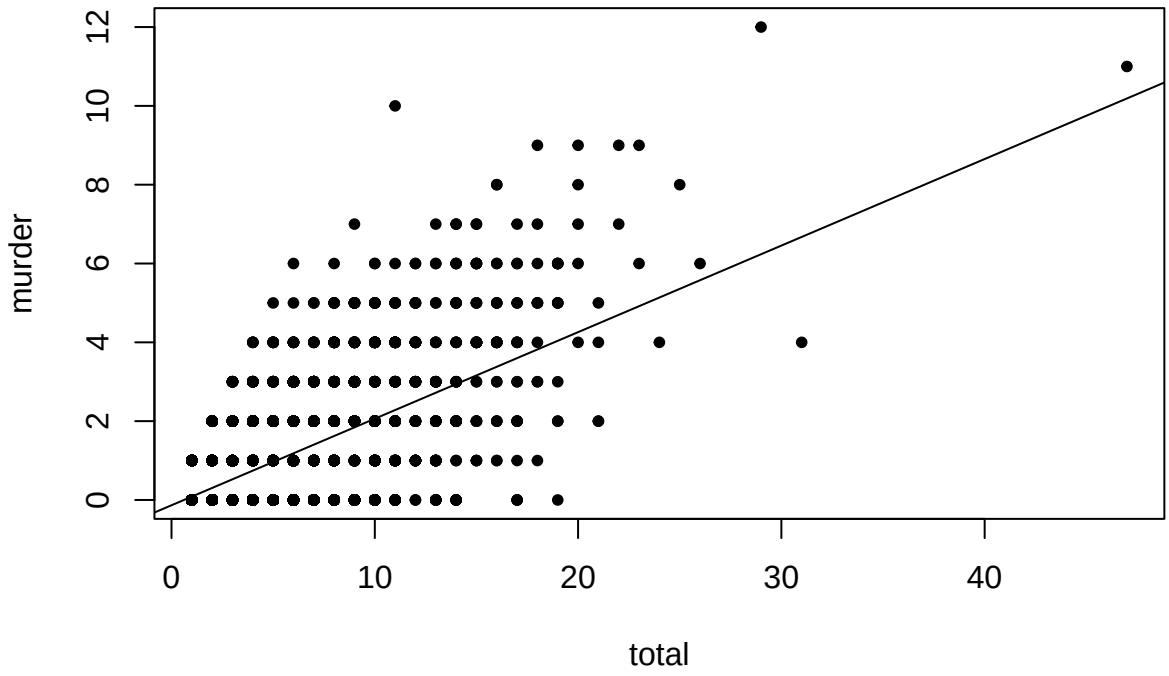
Modeling

```
shooting <- table(shooting$OCCUR_DATE,shooting$STATISTICAL_MURDER_FLAG)
murder <- shooting[,2]
total <- shooting[,1]+shooting[,2]
shooting_m <- data.frame(murder,total)
mod = lm(murder~total,data=shooting_m)
summary(mod)
```

```
##
## Call:
## lm(formula = murder ~ total, data = shooting_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0387 -0.5243 -0.0850  0.4757  7.7185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.134615   0.022655  -5.942    3e-09 ***
## total        0.219646   0.003859  56.924   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9774 on 5052 degrees of freedom
## Multiple R-squared:  0.3908, Adjusted R-squared:  0.3906
## F-statistic:  3240 on 1 and 5052 DF,  p-value: < 2.2e-16
```

```
plot(total,murder,pch=20)
abline(lm(murder~total))
```



The linear regression model shows that there is a relationship between the total number of shootings and the shootings with murder flag.

Step 4. Add Bias Identification

There is a possibility that there may be bias in this data analysis. This was analyzed based on the shooting incident data provided by the NYPD, and undefined/non-response answers were also reflected. In addition, the statistics based on the data may differ from the reality because the NYPD data does not include unreported events.

Conclusion

The conclusion of this analysis is that from 2006 to 2019, there was regional variation in the number of shootings in New York City, and the number maintained or decreased, and then increased again in 2020. Fatal shootings with a murder flag also increased with the total number of shootings.