Shooting Cases in NYC

Duc Huynh

2024-06-06

Packages used: tidyverse, utils, forecast and zoo

Shooting Cases in NYC:

```
#Input data set:
library(tidyverse)
library(utils)
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
read_dataset <- read.csv(url_in)
#Trim away unnecessary columns:
new_dataset <- read_dataset[,-c(1,5,6,7,8,9,17,18,19,20,21)]
head(new_dataset, 5)
```

```
##   OCCUR_DATE OCCUR_TIME      BORO STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1 05/05/2022   00:10:00 MANHATTAN                    true          25-44
## 2 07/04/2022   22:20:00     BRONX                    true         (null)
## 3 05/27/2012   19:35:00    QUEENS                   false
## 4 09/24/2019   21:00:00     BRONX                   false          25-44
## 5 02/25/2007   21:00:00  BROOKLYN                   false          25-44
##   PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE
## 1        M     BLACK         25-44       M    BLACK
## 2   (null)    (null)         18-24       M    BLACK
## 3                            18-24       M    BLACK
## 4        M   UNKNOWN         25-44       M    BLACK
## 5        M     BLACK         25-44       M    BLACK
```

```
tail(new_dataset,5 )
```

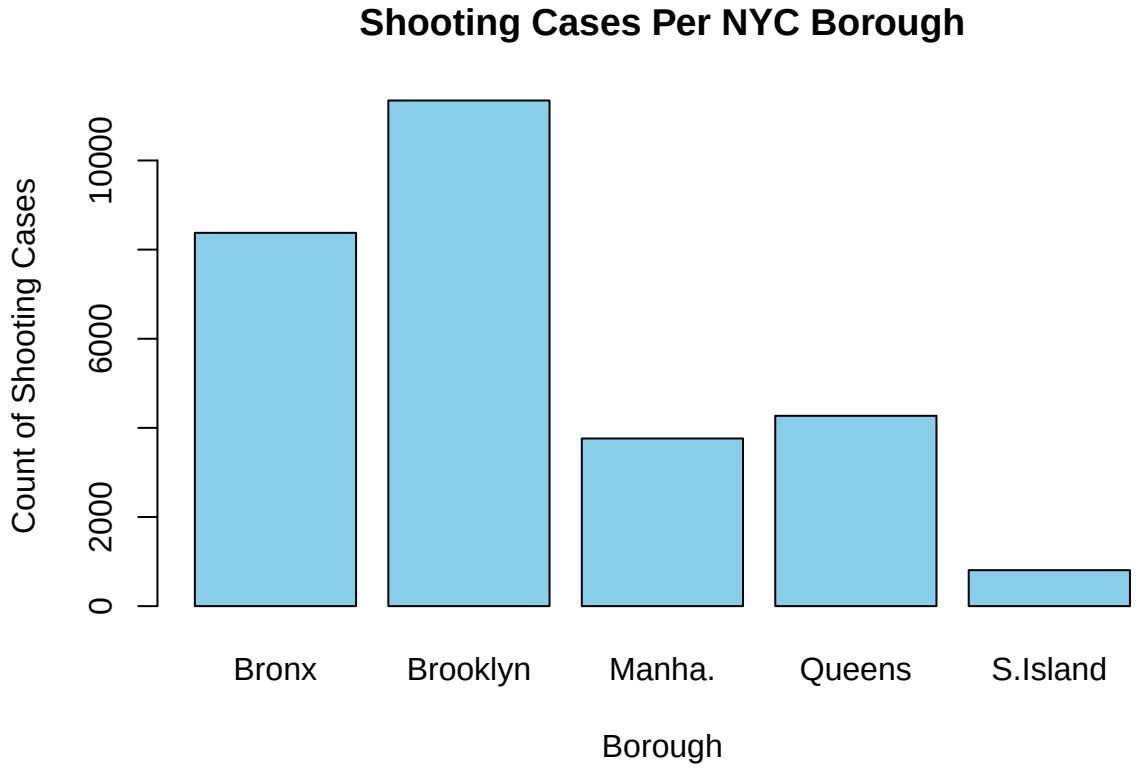
```
##       OCCUR_DATE OCCUR_TIME      BORO STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 28558 03/19/2023   23:48:00     BRONX                    true          18-24
## 28559 08/16/2023   02:46:00     BRONX                   false          25-44
## 28560 06/27/2023   12:27:00     BRONX                    true          25-44
## 28561 07/08/2023   11:27:00    QUEENS                   false          25-44
## 28562 07/24/2023   23:38:00 MANHATTAN                   false         (null)
##       PERP_SEX      PERP_RACE VIC_AGE_GROUP VIC_SEX                VIC_RACE
## 28558        M          BLACK         18-24       M                   BLACK
## 28559        F          BLACK         45-64       M                   BLACK
## 28560        M          BLACK         25-44       M                   BLACK
## 28561        M WHITE HISPANIC           65+       M ASIAN / PACIFIC ISLANDER
## 28562   (null)         (null)         25-44       M                   BLACK
```

Shooting Cases in NYC per Borough:

```
#Count number of shooting cases per borough:
count_boro <- read_dataset %>%
  count(BORO)
count_boro
```

```
##            BORO     n
## 1         BRONX  8376
## 2      BROOKLYN 11346
## 3     MANHATTAN  3762
## 4        QUEENS  4271
## 5 STATEN ISLAND   807
```

```
#Create a barplot for number of shooting cases per borough:
colnames <- c("Bronx","Brooklyn","Manha.","Queens","S.Island")
barplot(count_boro$n,
        names.arg = colnames,
        main = "Shooting Cases Per NYC Borough",
        xlab = "Borough",
        ylab = "Count of Shooting Cases",
        col = "skyblue",
        border = "black")
```

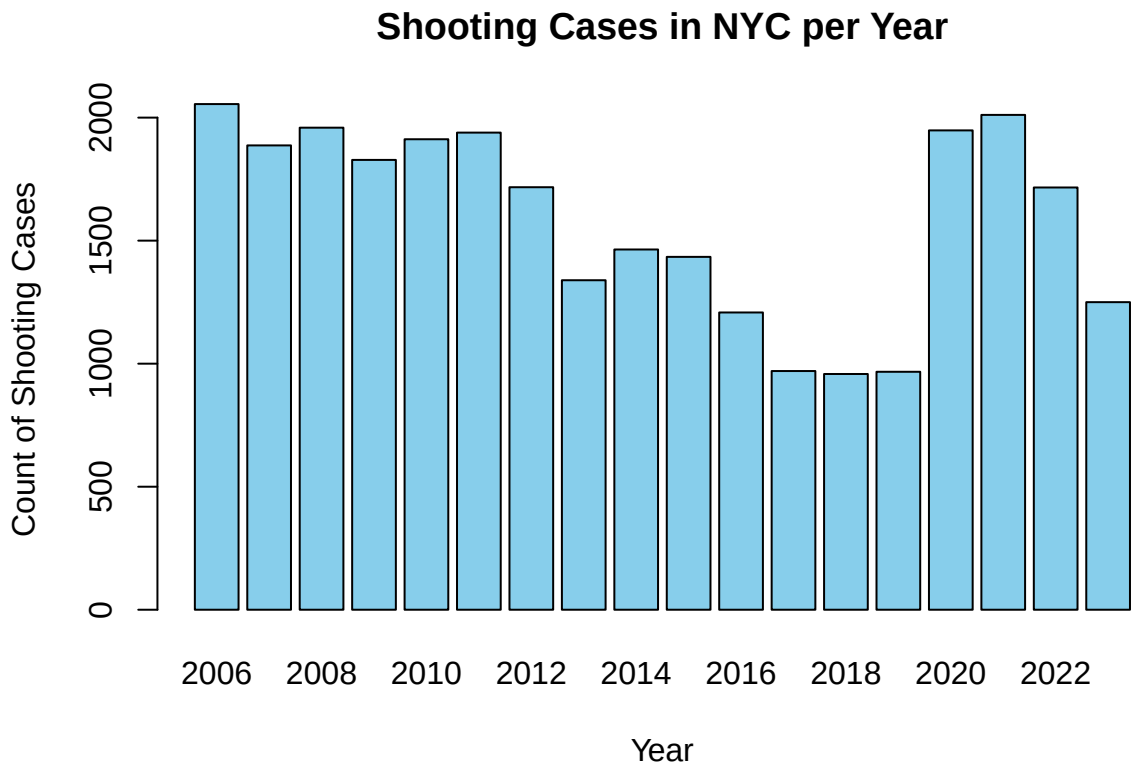


Shooting Cases in NYC per year:

```
#Transform date data into process able version:
new_date <-as.Date(read_dataset$OCCUR_DATE, format = "%m/%d/%Y")
#Count number of shooting cases per year:
a <- format(new_date, "%Y")
count_year <- table(a)
count_year
```

```
## a
## 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
## 2055 1887 1959 1828 1912 1939 1717 1339 1464 1434 1208  970  958  967 1948 2011
## 2022 2023
## 1716 1250
```

```
#Create a barplot for number of shooting cases per year:
barplot(count_year,
        main = "Shooting Cases in NYC per Year",
        xlab = "Year",
        ylab = "Count of Shooting Cases",
        col = "skyblue",
        border = "black")
```



Some Analysis:

Many shooting cases in NYC became fatal, which we call murder cases. We will make a graph to see if the number of murder cases is proportional to shooting cases per borough.

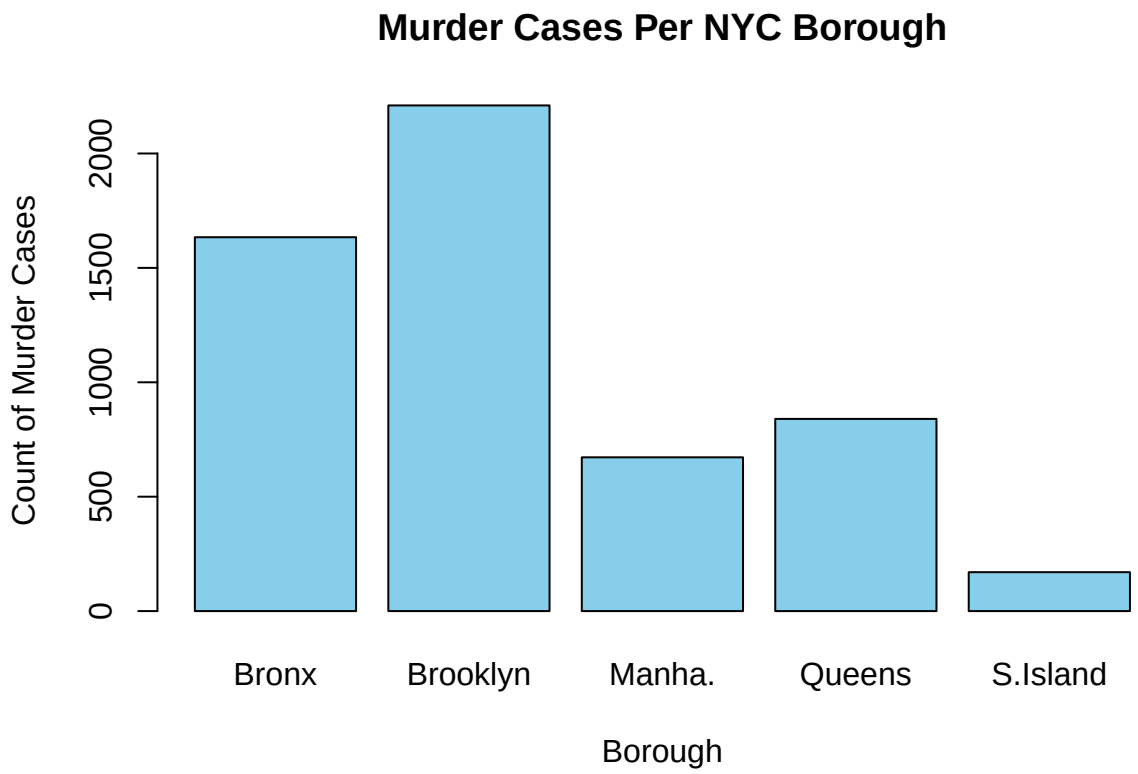
```
#Take out important columns to analyze:
new_dataset2 <-new_dataset [,c(1,3,4)]
#Choose cases that resulted in deaths:
new_dataset3 <-subset(new_dataset2, new_dataset2$STATISTICAL_MURDER_FLAG== "true")
new_dataset3$OCCUR_DATE<-as.Date(new_dataset3$OCCUR_DATE, "%m/%d/%Y")
head(new_dataset3,5)
```

```
##    OCCUR_DATE      BORO STATISTICAL_MURDER_FLAG
## 1  2022-05-05 MANHATTAN                   true
## 2  2022-07-04     BRONX                   true
## 7  2021-06-07    QUEENS                   true
## 10 2021-12-22     BRONX                   true
## 12 2021-12-23    QUEENS                   true
```

```
#Count number of murder cases per borough:
count_boro2 <- new_dataset3%>%
  count(BORO)
count_boro2
```

```
##            BORO    n
## 1         BRONX 1634
## 2      BROOKLYN 2210
## 3     MANHATTAN  672
## 4        QUEENS  840
## 5 STATEN ISLAND  170
```

```
#Create a barplot for number of murder cases per borough:
barplot(count_boro2$n,
        names.arg = colnames,
        main = "Murder Cases Per NYC Borough",
        xlab = "Borough",
        ylab = "Count of Murder Cases",
        col = "skyblue",
        border = "black")
```

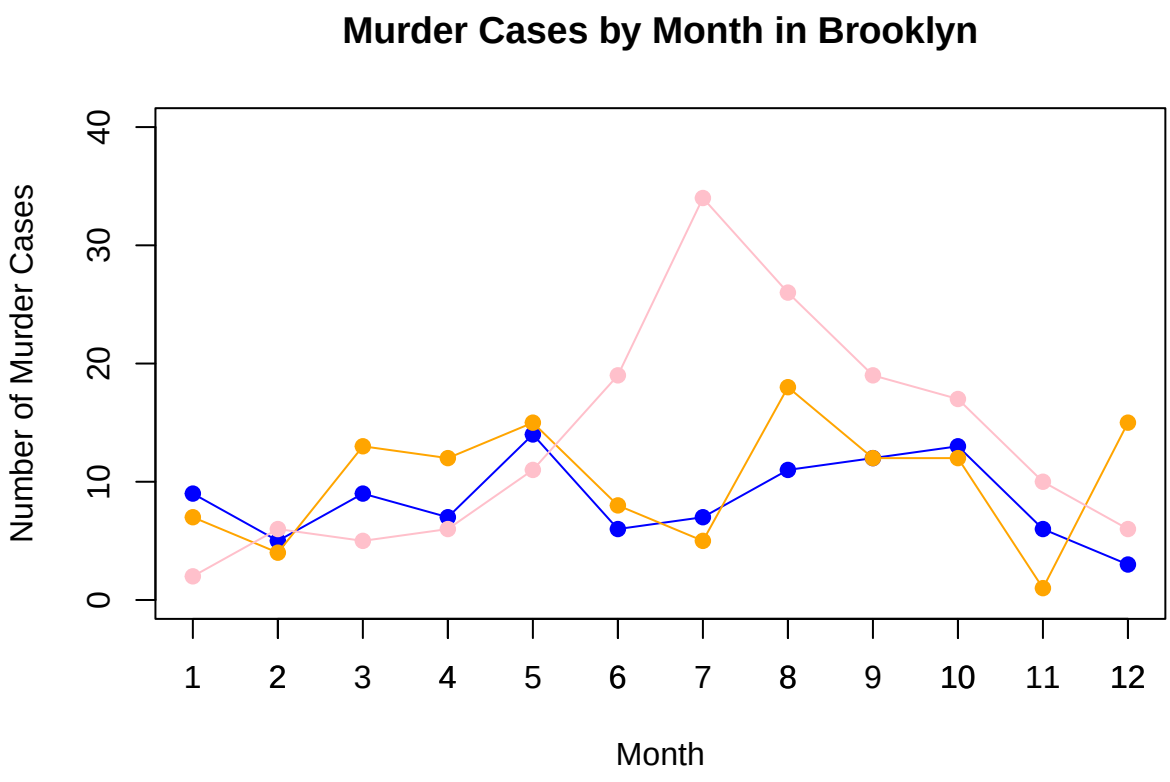


Conclusion: number of murder cases per borough seems to be proportional to number of shooting cases per borough. Now lets take a closer look and analyze number of murder cases in the most shooting borough, Brooklyn, over the year of 2020, 2021 and 2022:

```
#Count number of murder cases per month in year 2020:
date<-new_dataset3$OCCUR_DATE
murder_cases<-nrow(new_dataset3)
Year2020<- new_dataset3 %>% filter(between(date, as.Date('2020-01-01'), as.Date('2020-12-31')))
Brook20a<- Year2020 %>% filter(BORO== 'BROOKLYN')
Brook20b<- replace(Brook20a, 3, 1 )
Brook20c<-aggregate(Brook20b$STATISTICAL_MURDER_FLAG, by=list(Brook20b$OCCUR_DATE), sum)
Brook20d<-Brook20c %>%
  group_by(month(Group.1)) %>%
  summarize(Murder_Cases = sum(x))
colnames(Brook20d)[1] ="Month"
#Count number of murder cases per month in year 2021:
Year2021<- new_dataset3 %>% filter(between(date, as.Date('2021-01-01'), as.Date('2021-12-31')))
Brook21a<- Year2021 %>% filter(BORO== 'BROOKLYN')
Brook21b<- replace(Brook21a, 3, 1 )
Brook21c<-aggregate(Brook21b$STATISTICAL_MURDER_FLAG, by=list(Brook21b$OCCUR_DATE), sum)
Brook21d<-Brook21c %>%
  group_by(month(Group.1)) %>%
  summarize(Murder_Cases = sum(x))
colnames(Brook21d)[1] ="Month"
#Count number of murder cases per month in year 2022:
Year2022<- new_dataset3 %>% filter(between(date, as.Date('2022-01-01'), as.Date('2022-12-31')))
Brook22a<- Year2022 %>% filter(BORO== 'BROOKLYN')
```

```r
Brook22b<- replace(Brook22a, 3, 1 )
Brook22c<-aggregate(Brook22b$STATISTICAL_MURDER_FLAG, by=list(Brook22b$OCCUR_DATE), sum)
Brook22d<-Brook22c %>%
  group_by(month(Group.1)) %>%
  summarize(Murder_Cases = sum(x))
colnames(Brook22d)[1] ="Month"
#Graph number of murder cases per month in year 2020, 2021 and 2022:
x  <- Brook22d$Month
y1 <- Brook22d$Murder_Cases
y2 <- Brook21d$Murder_Cases
y3 <- Brook20d$Murder_Cases
plot(y1, type="o",xlim=c(1,12),ylim=c(0,40), col='blue', pch=19,lty=1,main="Murder Cases by Month in Br
axis(side=1, at=1:12)

points(x, y2, col="orange", pch=19)
lines(x, y2, col="orange")
points(x, y3, col="pink", pch=19)
lines(x, y3, col="pink")
```



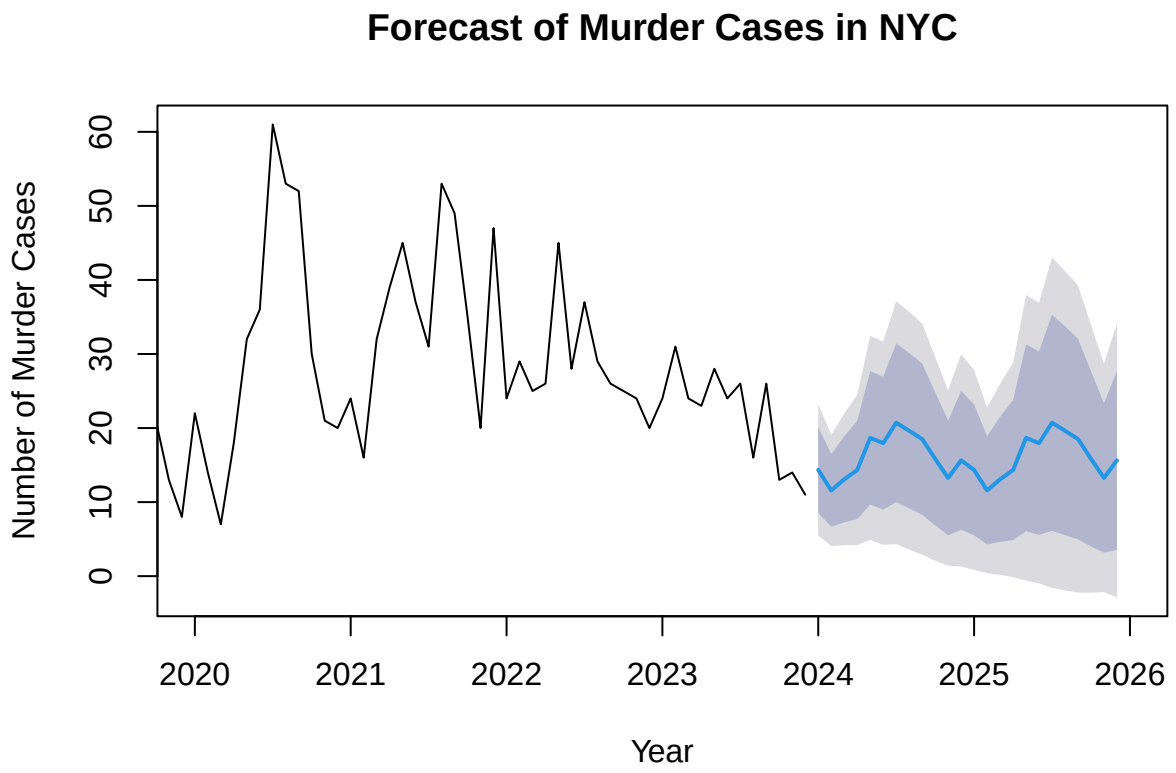
Pink: Year 2020. Orange: Year 2021. Blue: Year 2022.

From the graph we can see that highest number of murder cases usually occurred around August though the years, when people were out for summer activities. Hence, more police forces should be deployed in the summer to ensure security of the area.

Modeling Data:

I would like to predict the pattern in murder cases by gun in NYC up to year 2026:

```
library(forecast)
library(zoo)
#Add case count:
new_dataset3$case <-1
aggYears<-aggregate(new_dataset3$case,by=list(new_dataset3$OCCUR_DATE),sum)
aggMonth<-aggYears %>%
group_by(month = lubridate::floor_date(Group.1,"month")) %>%
summarize(summary_variable = sum(x))
#Convert to data frame:
frame<-as.data.frame(aggMonth)
tsdata<-as.ts(read.zoo(frame,FUN=as.yearmon))
prediction<-forecast(tsdata)
#Graph the prediction:
plot(prediction,xlim=c(2020,2026),
main="Forecast of Murder Cases in NYC",
xlab="Year",ylab="Number of Murder Cases")
```



The prediction still tells us that number of murder cases usually spikes in summertime and decreases in the end of year. More importantly, it also tells us that the number of murder cases will slowly decrease over the years.

Bias Sources:

There can be various bias sources in collecting shooting cases in NYC, such as Reporting Bias, when not all shootings are reported to the police (due to fear of retaliation, or reporting avoidance of undocumented groups, etc); or Sampling Bias, when reports only consider data from some specific agencies, hospitals or neighborhoods..., not the whole facilities. Selection Bias also counts when authorities decide to hide and unreport some incidents due to sensitive reasons. Misclassification can be another source of bias when sometimes shootings are misclassified under different crime categories such as assault, robbery.... due to unclear circumstances.

In order to mitigate these biases, it is important to collect data from various sources, and cross-reference reports from police departments, hospitals, medias or community surveys... Also, ones need to perform appropriate statistical methods to adjust for known biases, to provide a more comprehensive picture of shooting cases in NYC.