NYPD Shooting Incident Project

S. Mauri

2023-09-01

This dataset is a list of every shooting incident that occurred in NYC from 2006 through the end of 2022. This dataset can be found at https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8. It is provided by the NYPD and available for use by the public. Each record represents a shooting in NYC and includes information about the event, location, and the date/time. Also, information about the gender, age, and race of the victim and suspect is included. I'm going to do an analysis to see how the number of shootings changes over the years. I'm also going to see if the number of shootings that occurred differed based on the month and the time of day. Then I used a model to see if the Borough, the time of day, and the time of year affected whether or not the shooting led to a murder.

I installed these packages for this analysis.

```
library(tidyverse)
library(lubridate)
```

First, I will read in the csv file for the data from the website.

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

Now I will read in the data and see what it looks like.

```
shooting_data <- read_csv(url_in)
```

After looking at shooting_data, I am going to get rid of data that I don't need including INCIDENT_KEY, LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, and Lon_Lat. I am going to rename the variables that I am keeping. The STATISTICAL_MURDER_FLAG (renamed Flag) is a True/False based on whether the shooting resulted in the victim's death which would be counted as a murder.

```
shooting_data <- shooting_data %>%
  select(STATISTICAL_MURDER_FLAG, OCCUR_TIME,
         OCCUR_DATE, BORO, PERP_AGE_GROUP,
         PERP_SEX, PERP_RACE,
         VIC_AGE_GROUP,
         VIC_SEX, VIC_RACE) %>%
  rename(Flag = STATISTICAL_MURDER_FLAG,
         Time = OCCUR_TIME,
         Date = OCCUR_DATE,
         Boro = BORO,
         Perpetrator_Age = PERP_AGE_GROUP,
         Perpetrator_Sex = PERP_SEX,
```

```
         Perpetrator_Race = PERP_RACE,
         Victim_Age = VIC_AGE_GROUP,
         Victim_Sex = VIC_SEX,
         Victim_Race = VIC_RACE)
```

I will also change the Date to date data type and the Perpetrator and Victim Age, Sex, and Race, and the Borough to the factor data type.

```
shooting_data$Date <- mdy(shooting_data$Date)
shooting_data$Perpetrator_Age <- as.factor(shooting_data$Perpetrator_Age)
shooting_data$Perpetrator_Sex <- as.factor(shooting_data$Perpetrator_Sex)
shooting_data$Perpetrator_Race <- as.factor(shooting_data$Perpetrator_Race)
shooting_data$Victim_Age <- as.factor(shooting_data$Victim_Age)
shooting_data$Victim_Sex <- as.factor(shooting_data$Victim_Sex)
shooting_data$Victim_Race <- as.factor(shooting_data$Victim_Race)
shooting_data$Boro <- as.factor(shooting_data$Boro)
```

Here is a summary of the data:

```
summary(shooting_data)
```

```
##    Flag             Time               Date                          Boro
##  Mode :logical   Length:27312      Min.   :2006-01-01   BRONX        : 7937
##  FALSE:22046     Class1:hms        1st Qu.:2009-07-18   BROOKLYN     :10933
##  TRUE :5266      Class2:difftime   Median :2013-04-29   MANHATTAN    : 3572
##                  Mode  :numeric    Mean   :2014-01-06   QUEENS       : 4094
##                                    3rd Qu.:2018-10-15   STATEN ISLAND:  776
##                                    Max.   :2022-12-31
##
##  Perpetrator_Age Perpetrator_Sex       Perpetrator_Race   Victim_Age
##  18-24  :6222    (null):  640    BLACK          :11432   <18    : 2839
##  25-44  :5687    F     :  424    WHITE HISPANIC: 2341   1022   :    1
##  UNKNOWN:3148    M     :15439    UNKNOWN        : 1836   18-24  :10086
##  <18    :1591    U     : 1499    BLACK HISPANIC: 1314   25-44  :12281
##  (null) : 640    NA's  : 9310    (null)         :  640   45-64  : 1863
##  (Other): 680                    (Other)        :  439   65+    :  181
##  NA's   :9344                    NA's           : 9310   UNKNOWN:   61
##  Victim_Sex                         Victim_Race
##  F: 2615    AMERICAN INDIAN/ALASKAN NATIVE:   10
##  M:24686    ASIAN / PACIFIC ISLANDER      :  404
##  U:   11    BLACK                         :19439
##             BLACK HISPANIC                : 2646
##             UNKNOWN                       :   66
##             WHITE                         :  698
##             WHITE HISPANIC                : 4049
```

There is quite a bit of missing data in the perpetrator age, sex, and race categories, likely because in these cases it is not known who the perpetrator is. There is also some missing data in the victim age, sex, and race categories. I also noticed there was one victim age that was 1022, which I'm assuming is a typo. I'm deciding not to use these variables in any analysis since there is so much missing data.

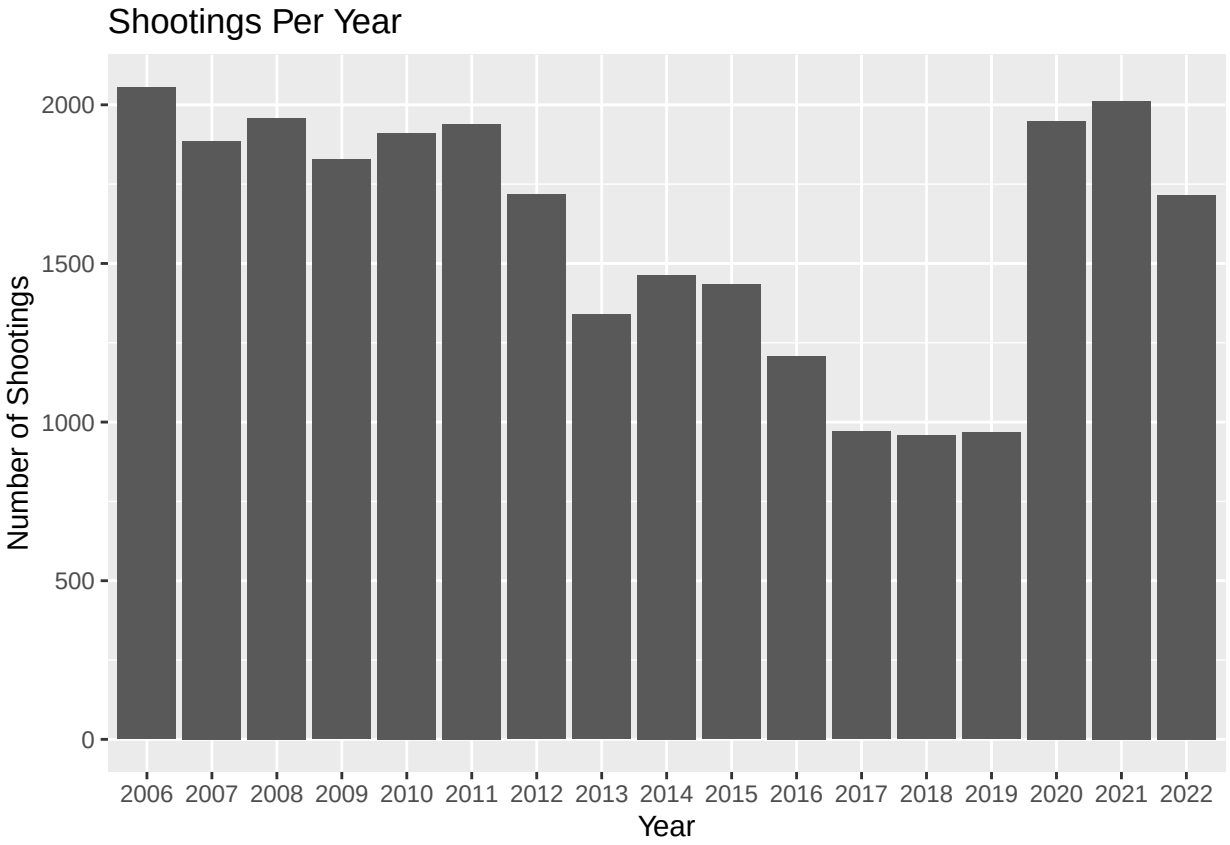
Visualizations

I was interested to see how the number of shootings changes as time went on, so I grouped the data by year and then plotted the shootings per year.

```
shooting_data <- shooting_data %>% mutate(year = year(Date))

shootings_per_year <- shooting_data %>% group_by(year) %>% summarize(incidents_per_year = n())

ggplot(data = shootings_per_year, mapping = aes(x = factor(year), y = incidents_per_year)) +
  geom_bar(stat="identity") +
  xlab("Year") + ylab("Number of Shootings") + ggtitle("Shootings Per Year")
```



This graph showing the number of shootings per year is the opposite of what I thought it would be. It shows the number of shootings per year trending down until 2020, where the number of shootings is nearly as many as in 2006. I would have expected that criminal activity would decrease when the COVID lockdowns went into effect.

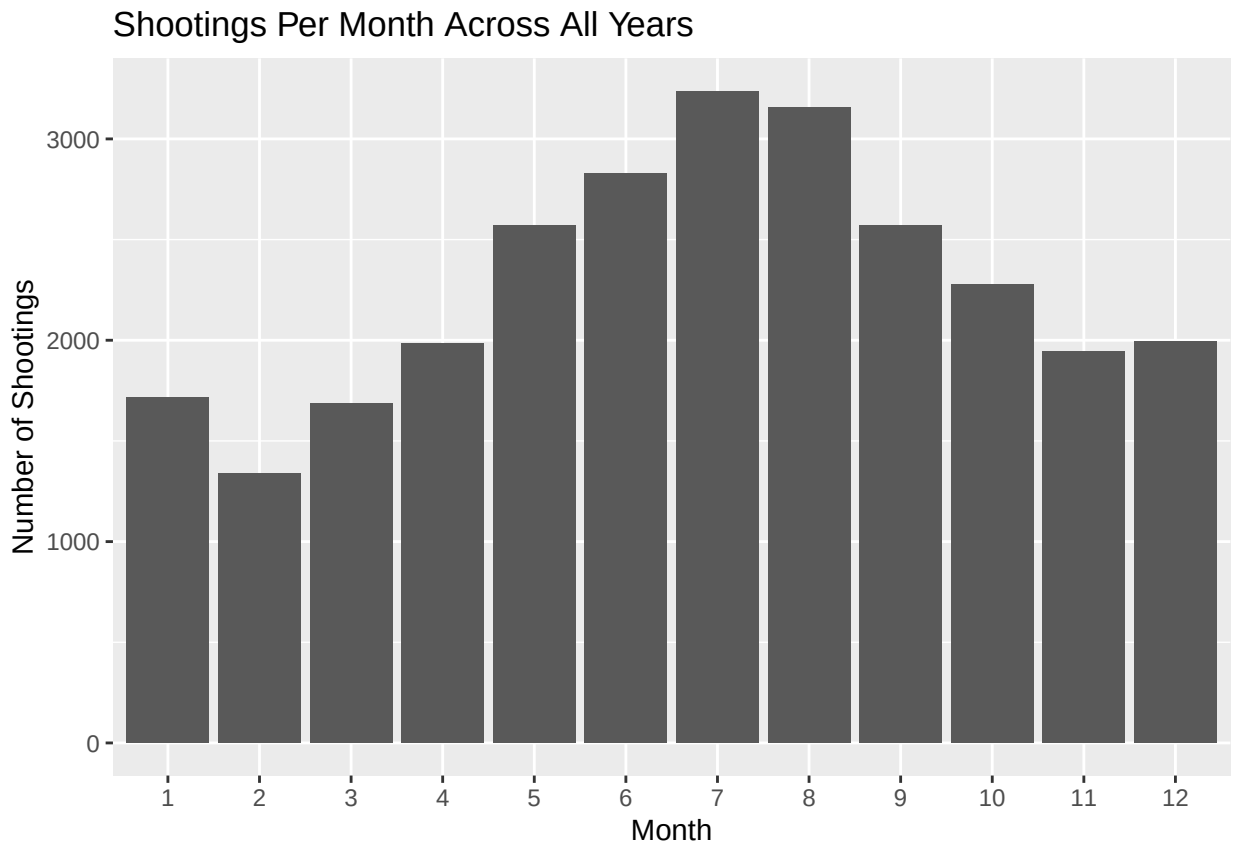
I was also interested to see how many shootings occurred each month, so I grouped the data by month and then plotted the total number of shootings per month.

```
shooting_data <- shooting_data %>% mutate(month = month(Date))

shootings_per_month <- shooting_data %>% group_by(month) %>% summarize(incidents_per_month = n())

ggplot(data = shootings_per_month, mapping = aes(x = factor(month), y = incidents_per_month)) +
```

```
  geom_bar(stat = "identity") +
  xlab("Month") + ylab("Number of Shootings") + ggtitle("Shootings Per Month Across All Years")
```



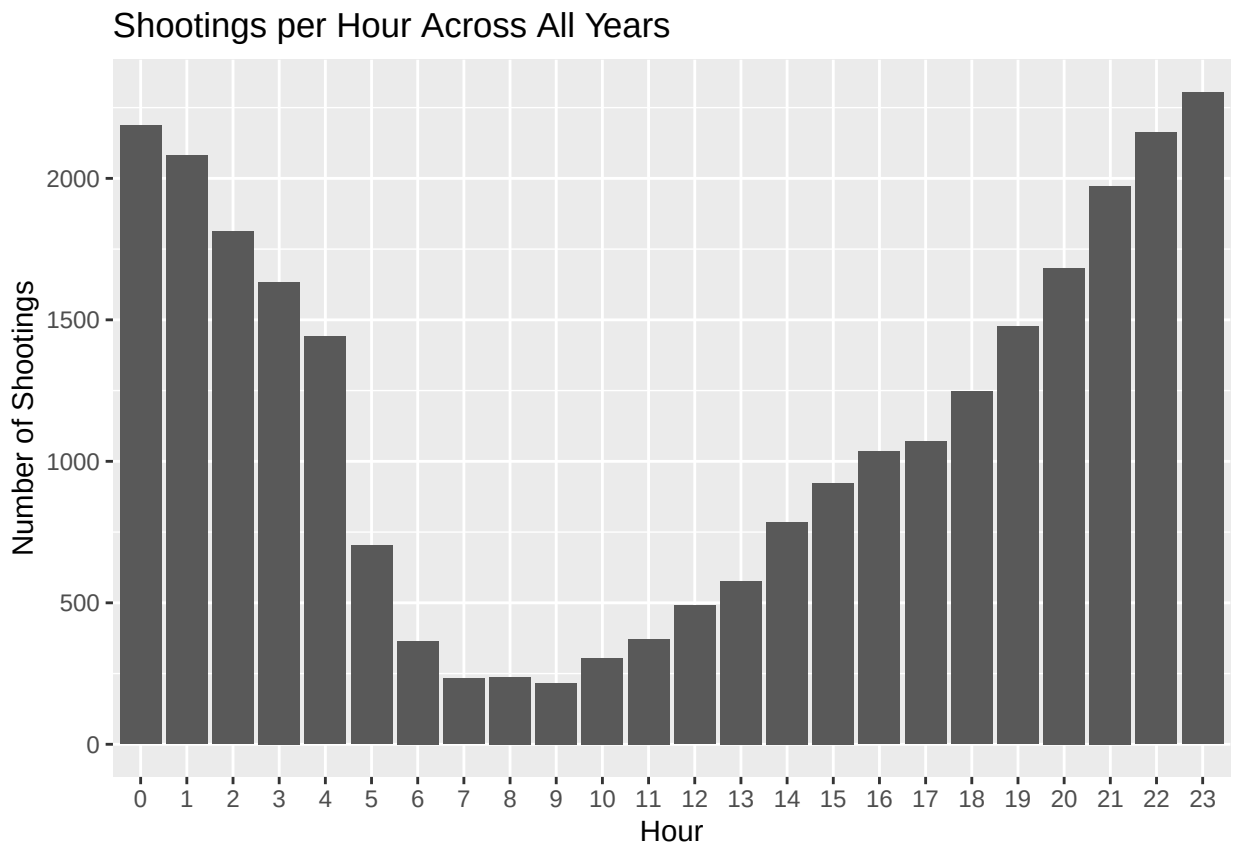
This second graph shows that across all the years of data, the most shootings occurred in the summer, and the fewest in the winter.

Now I'm curious about what time of day the most shootings occur, so I'm going to graph that.

```
shooting_data <- shooting_data %>% mutate(hour = hour(Time))

shootings_per_hour <- shooting_data %>% group_by(hour) %>% summarize(incidents_per_hour = n())

ggplot(data = shootings_per_hour, mapping = aes(x = factor(hour), y = incidents_per_hour)) +
  geom_bar(stat = "identity") +
  xlab("Hour") + ylab("Number of Shootings") + ggtitle("Shootings per Hour Across All Years")
```

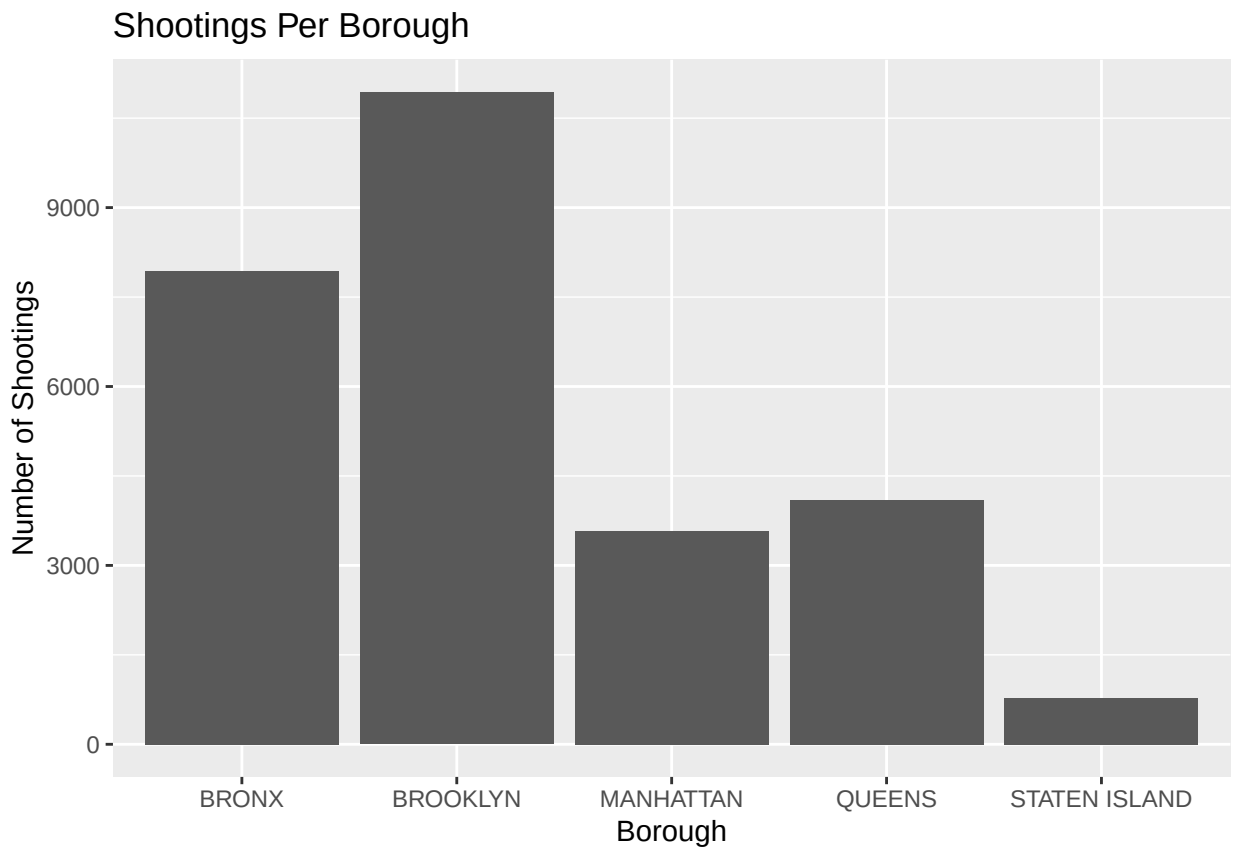


This shows that most incidents happen late at night, between the hours of 10 PM and 2 AM.

I want to plot the number of shootings that have occurred since 2006 for each Borough so I grouped the data and counted the shootings in each.

```
boro_shootings <- shooting_data %>% group_by(Boro) %>% summarize(incidents = n())

ggplot(data = boro_shootings, mapping = aes(x = Boro, y = incidents)) +
    geom_bar(stat = "identity") +
    xlab("Borough") + ylab("Number of Shootings") + ggtitle("Shootings Per Borough")
```



Brooklyn had the highest number of incidents, followed by the Bronx and Queens.

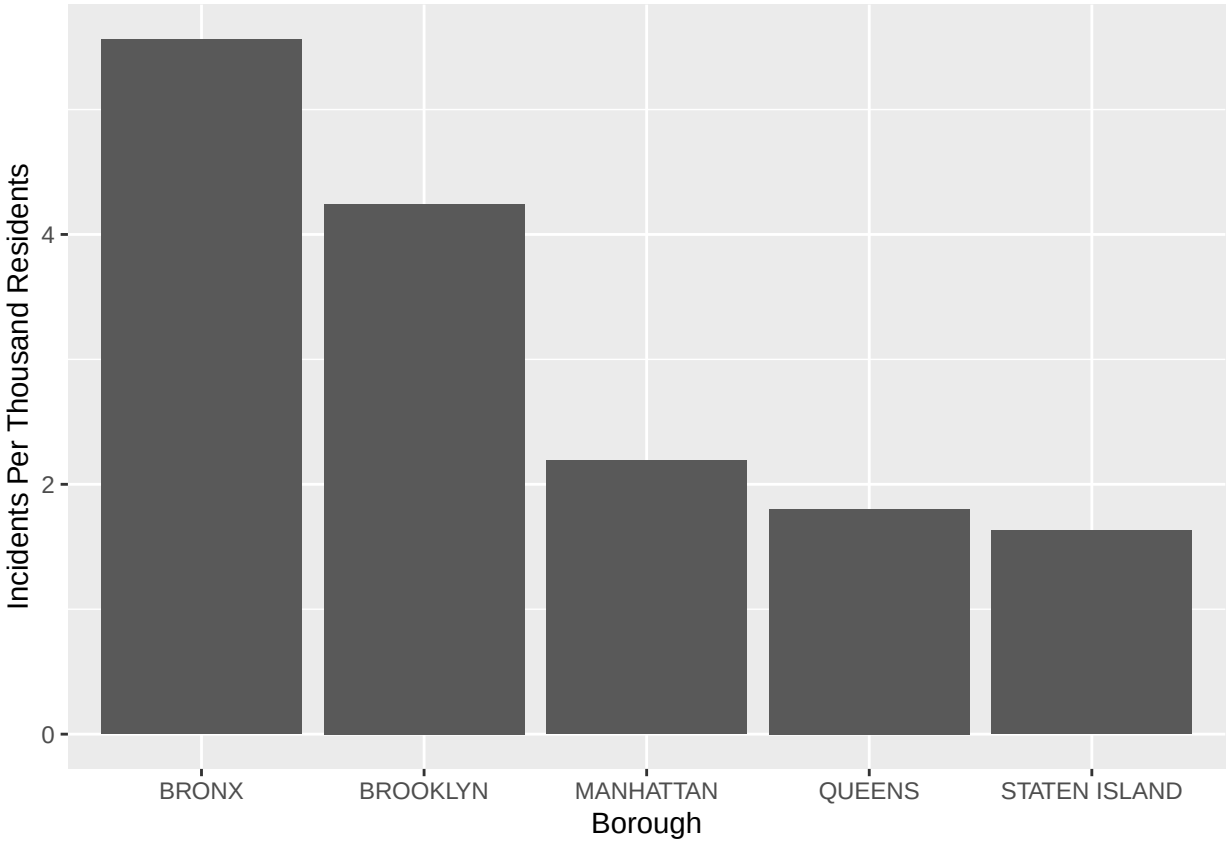
Next I added the population for each Borough.

```
shooting_data <- shooting_data %>%
  mutate(Population = case_when(
    Boro == "BROOKLYN" ~ 2576771,
    Boro == "QUEENS" ~ 2270976,
    Boro =="BRONX" ~ 1427056,
    Boro == "MANHATTAN" ~ 1629153,
    Boro =="STATEN ISLAND" ~ 475596))
```

I also wanted to add the number of incidents per 1000 residents using the population for each Borough and visualize this in order to see which Borough is the most dangerous.

```
rate <- shooting_data %>%
  group_by(Boro) %>% summarize(total = n(),
  population = unique(Population)) %>%
  mutate(deaths_per_thou = total / population * 1000)

ggplot(data = rate, mapping = aes(x = Boro, y = deaths_per_thou)) +
  geom_bar(stat = "identity") +
  xlab("Borough") + ylab("Incidents Per Thousand Residents")
```



This visualization shows that even though Brooklyn had the largest number of incidents, the Bronx has the highest shooting rate per thousand residents.

Conclusion and Bias

From the analysis, we can see that the total number of shootings was on a steady decline until 2020, where the number of shootings jumped to almost the level it was at in 2006, which had the highest number of shootings of any year in this dataset. The highest number of shootings occurred in the summer months. The most shootings also occurred in the middle of the night, between the hours of 10 pm and 2 am. Brooklyn had the most shootings overall, and Staten Island had the fewest. Brooklyn also has the highest population of all the Boroughs, and Staten Island had the smallest population. However, calculating the shooting rate per 1000 residents showed us that the Bronx actually had the highest shooting rate.

This topic can bring out a lot of bias. This can start from the moment the data was collected and happen every step of the way up to and including the analysis of the data. There was quite a bit of missing data in the perpetrator age, sex, and race categories, likely because in these cases it is not known who the perpetrator is. I decided not to do any analysis involving this data because any analysis may not have been accurate since so much was missing. I was surprised that there are significantly more incidents involving male victims than female victims. Not knowing much about NYC, I was surprised that Brooklyn had the largest number of incidents overall, I would have assumed it would be the Bronx or Queens. This was an interesting dataset to look at to see that some of my first assumptions were incorrect.