NYPD Shooting Incident Project

Elise Richard

2022-11-07

```
SessionInfo()
```

```
R version 4.2.2 (2022-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22000)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8  LC_CTYPE=C
[3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8
system code page: 65001

attached base packages:
[1] grDevices stats     graphics  utils     datasets  methods   base

other attached packages:
 [1] stringr_1.4.0      lubridate_1.8.0    httr_1.4.3         summarytools_1.0.1 data.table_1.14.2
 [6] knitr_1.39         forcats_0.5.1      dplyr_1.0.9        purrr_0.3.4        readr_2.1.2
[11] tidyr_1.2.0        tibble_3.1.8       ggplot2_3.3.6      tidyverse_1.3.2
```

Objectives:

1. Determine the neighborhood with the most shootings
2. Determine which neighborhood had the most incidents that were also murders
3. Determine when a shooting is most likely to occur

Data Source:

The csv file comes from the City of New York Open Data.

csv: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv

Importing the data:

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

Tidying the Data:

Used data.tables to format the data:

```
NYPD <- fread("~/R/NYPD_Shooting_Data.csv")
```

Checking columns:

```
colnames(NYPD)
```

```
 [1] "INCIDENT_KEY"           "OCCUR_DATE"             "OCCUR_TIME"             "BORO"
 [5] "PRECINCT"               "JURISDICTION_CODE"      "LOCATION_DESC"          "STATISTICAL_MURDER_FLAG"
 [9] "PERP_AGE_GROUP"         "PERP_SEX"               "PERP_RACE"              "VIC_AGE_GROUP"
[13] "VIC_SEX"                "VIC_RACE"               "X_COORD_CD"             "Y_COORD_CD"
[17] "Latitude"               "Longitude"              "Lon_Lat"
```

Removed all null values:

```
na.omit(NYPD)
```

Selected only the columns I want to use for the project:

```
NYPD2 <- NYPD %>% select(INCIDENT_KEY,
                         OCCUR_DATE,
                         OCCUR_TIME,
                         BORO,
                         STATISTICAL_MURDER_FLAG,
                         PERP_AGE_GROUP,
                         VIC_AGE_GROUP,
                         Latitude,
                         Longitude)
```

Summary:

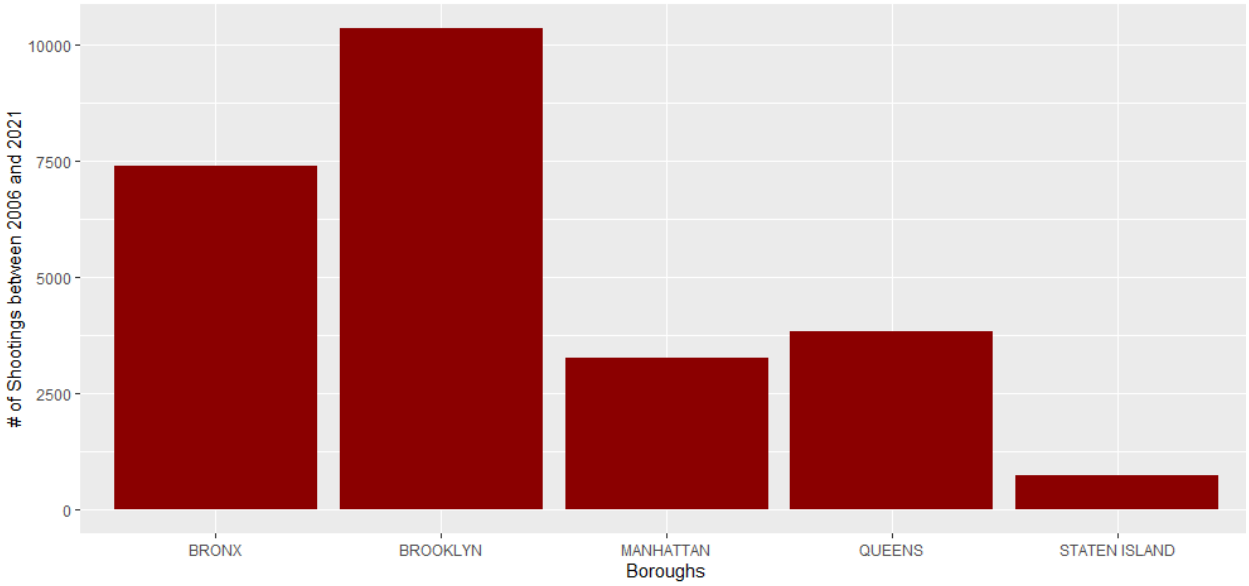
```
summary(NYPD2)
```

```
  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO            STATISTICAL_MURDER_FLAG
 Min.   :  9953245   Length:25596       Length:25596       Length:25596       Mode :logical
 1st Qu.: 61593633   Class :character   Class :character   Class :character   FALSE:20668
 Median : 86437258   Mode  :character   Mode  :character   Mode  :character   TRUE :4928
 Mean   :112382648
 3rd Qu.:166660833
 Max.   :238490103
 PERP_AGE_GROUP     VIC_AGE_GROUP         Latitude       Longitude
 Length:25596       Length:25596       Min.   :40.51   Min.   :-74.25
 Class :character   Class :character   1st Qu.:40.67   1st Qu.:-73.94
 Mode  :character   Mode  :character   Median :40.70   Median :-73.92
                                       Mean   :40.74   Mean   :-73.91
                                       3rd Qu.:40.82   3rd Qu.:-73.88
                                       Max.   :40.91   Max.   :-73.70
```

Data Analysis:

The first item on the agenda is to determine which borough has the largest number of incidents:

```
ggplot(NYPD2) +
  geom_bar(aes(x = BORO), fill = "red4") +
  labs(x="Boroughs", y="# of Shootings between 2006 and 2021")
```



We can easily see that the answer is Brooklyn, and Staten Island with the least number.

Now, how many of these shootings were murder cases? We can just make a table from the two columns to see the results.

```
table(NYPD2$BORO, NYPD2$STATISTICAL_MURDER_FLAG)
```

```
               FALSE TRUE
  BRONX         5985 1417
  BROOKLYN      8345 2020
  MANHATTAN     2691  574
  QUEENS        3066  762
  STATEN ISLAND  581  155
```

The number of murder cases shows the same trend as the total number of shootings, with Brooklyn at the top of the list.

To answer the day and time that NYC is the most dangerous, we'll create new columns in the NYPD2 data frame for the day and time data:

```
NYPD2$OCCUR_DAY = mdy(NYPD2$OCCUR_DATE)
NYPD2$OCCUR_DAY = wday(NYPD2$OCCUR_DAY, label = TRUE)
NYPD2$OCCUR_HOUR = hour(hms(as.character(NYPD2$OCCUR_TIME)))
```

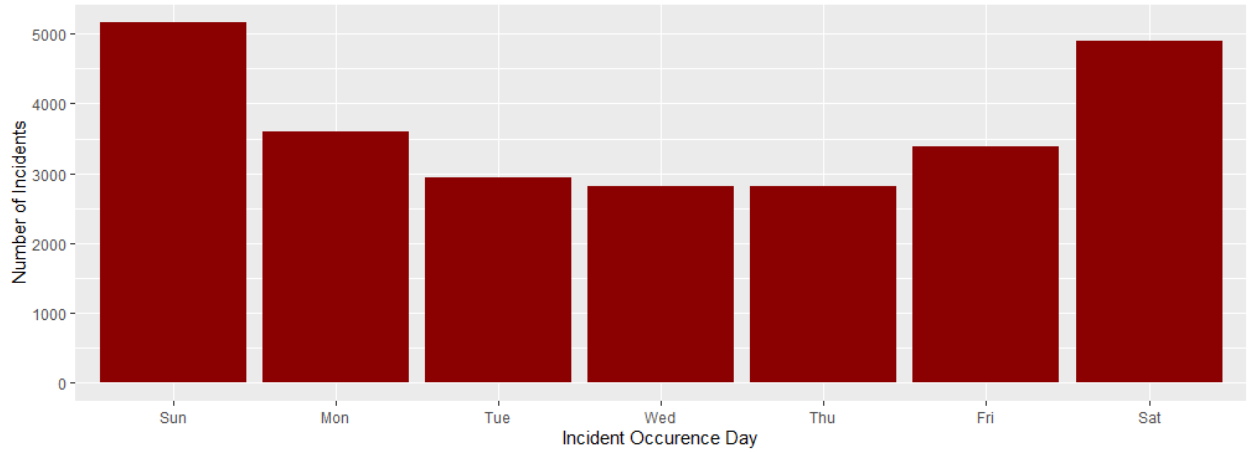
Next, we can create separate data frames to count the number of incidents per day and also per hour:

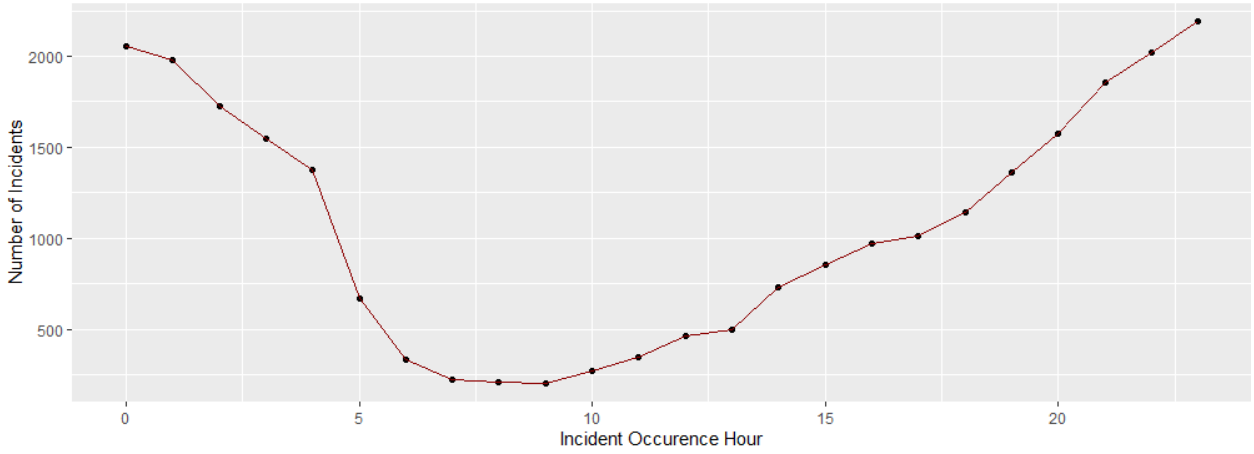
```
NYPD3 <- NYPD2 %>%
  group_by(OCCUR_DAY) %>%
  count()

NYPD4 <- NYPD2 %>%
  group_by(OCCUR_HOUR) %>%
  count()
```

And now we can visualize our data:

```
ggplot(NYPD3, aes(x = OCCUR_DAY, y = n, )) +
  geom_col(fill = "red4") +
  labs(x = "Incident Occurence Day",
       y = "Number of Incidents")
```

```
ggplot(data = NYPD4, aes(x = OCCUR_HOUR, y = n)) +
  geom_point() +
  geom_line(color = "red4") +
  labs(x = "Incident Occurence Hour",
       y = "Number of Incidents")
```





According to our analysis, Saturday and Sunday are the days with the most occurrences, and it is most dangerous to be out between the hours of approximately 8pm and 1am.

Making a model:

After the data analysis thus far, I became interested in how likely it would be for each age group to be involved in a shooting incident, so I created a model based on the victims' ages:

```
modelone.fit <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, data = NYPD)
summary(modelone.fit)
```

```
Call:
glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, data = NYPD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3234  -0.2196  -0.1651  -0.1302   0.8698

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.130175   0.007583  17.168  < 2e-16 ***
VIC_AGE_GROUP18-24   0.034964   0.008576   4.077 4.58e-05 ***
VIC_AGE_GROUP25-44   0.089393   0.008428  10.606  < 2e-16 ***
VIC_AGE_GROUP45-64   0.119530   0.012177   9.816  < 2e-16 ***
VIC_AGE_GROUP65+     0.193178   0.031314   6.169 6.97e-10 ***
VIC_AGE_GROUPUNKNOWN 0.119825   0.051251   2.338   0.0194 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1541482)

    Null deviance: 3979.2  on 25595  degrees of freedom
Residual deviance: 3944.7  on 25590  degrees of freedom
AIC: 24786

Number of Fisher Scoring iterations: 2
```

From the data we can see that it is very likely that individuals between the ages of 25-44 will be a victim of a shooting incident in NYC.

Identifying Bias:

I, personally have visited NYC one time in my life and were not the victim of any crime during the visit. However, that doesn't mean there is no crime in the areas I saw or during the time of the visit. Most of my knowledge about what it's like to live in NYC comes from television. The popular shows "Brooklyn Nine Nine," and "Two Broke Girls," take place in Brooklyn and do showcase quite a bit of crime across each series. On the other end of the spectrum is "Gossip Girl," which takes place in Manhattan, which on the show, appears to be a mostly crime-free, wealthy area. So with little other information on crime statistics in NYC, most of my bias would come from watching television shows made to emulate life in NYC.

Does my bias align with the data analysis?

While I expected Brooklyn to have the most shootings and murder incidents, I also expected Manhattan to be the lowest when Staten Island actually had the least number of incidents. Because we haven't analyzed the number of incidents based on local population (population data was not included in this data set), it's uncertain whether Manhattan or Staten Island is the more dangerous borough. In short, my personal bias did affect my initial hypothesis of the data, but did not hinder me from making an objective analysis.

Conclusion:

It was determined from a rough analysis of the NYPD Shooting Incident data that:

1. The borough with the most incidents is Brooklyn.
2. The borough with the most shootings that were also murders is Brooklyn.
3. Saturday and Sunday are the days with the most occurrences, and it is most dangerous to be out between the hours of approximately 8pm and 1am.