NYPD Shooting Incidents

K. Al Romaithi

6/4/2021

Introduction

In this project, I am going to analyze the **NYPD Shooting Incident Data**. This dataset can be found in the following URL https://catalog.data.gov/dataset and searching for the dataset with the same name.

I am mainly interested in analyzing the frequency of those incidents as well as the ages and races of the people involved.

Data Science Process

i) Importing the data

For reproducibility purposes, I copy the link address of the dataset by right-clicking on the csv version link. The following is a code-snippet for importing the dataset into R studio.

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

ii) Exploring the data

After importing the data, we can see that it has 23,568 rows and 19 columns. It's very important to understand what the columns contain in order to identify what is necessary for our exploration. The columns and their explanations are as follows:

1. INCIDENT KEY - this is a randomly generated ID
2. OCCUR_DATE - date of the shooting incident
3. OCCUR_TIME - time of the shooting incident
4. BORO - borough where the incident occurred
5. PRECINCT - precinct where the incident occurred
6. JURISDICTION_CODE - this column includes a jurisdiction where the shooting incident occurred. Examples: 0 Patrol, 1 Transit and 2 Housing
7. LOCATION_DESC - Location of the incident
8. STATISTICAL_MURDER_FLAG - whether the shooting resulted in the victim's death
9. PERP_AGE_GROUP - age group of the perpetrator
10. PERP_SEX - perpetrator's sex
11. PERP_RACE - perpetrator's race
12. VIC_AGE_GROUP - age group of the victim
13. VIC_SEX - victim's sex

14. VIC_RACE - victim's race
15. X_COORD_CD - Midblock X-coordinate for New York State Plane Coordinate System,
16. Y_COORD_CD - Midblock Y-coordinate for New York State Plane Coordinate System,
17. Latitude - Latitude coordinate for Global Coordinate System
18. Longitude - Longitude coordinate for Global Coordinate System
19. Lon_Lat - Longitude and Latitude coordinates used for mapping

---

Reading the Data Footnotes from the website, we can understand the collection process of this data better. The footnotes note that some null values do occur in this data due to the data capturing method at that point. These are marked with "Unknown/Not Available/Not Reported". In addition, the footnotes mention that the incident_key could be duplicated in case a shooting incident resulted in multiple victims.

Running a summary of the shooting_incidents dataframe, we can notice that there are some values labeled as *NA*. This is worth keeping in mind to know what data may have missing values. In the next section, I will outline all the steps taken in order to tidy up this data.

iii) Tidying and transforming the data

As stated in the introduction of this report, I will be looking at a more high-level analysis of the data. I am interested in exploring the frequency of the shooting incidents according to the New York boroughs. In addition, I would like to look into the age, sex and race of both the perpetrator and victim (where available). Location information is not of interest at the moment.

To start, we can drop a few columns such as the precinct, jurisdiction_code, location_desc, as well as the location information (columns 15 to 19). I also notice that the column **OCCUR_DATE** is a character, so I will need to convert that into a date variable.

```
shooting_incidents <- raw_shooting_incidents[-c(5:7, 15:19)]
shooting_incidents$OCCUR_DATE <- mdy(shooting_incidents$OCCUR_DATE)
```

Printing out the **shooting_incidents** dataframe, I can now see that I've picked out the columns I'm more interested in analyzing. In addition, the data type of the date column is not correct. Running `summary(shooting_incidents)`, we can see that the data ranges from 1st of January 2006 up until the 31st of December 2020.

I can also see that there are some rows under the perpetrator columns that are either marked as **NA**, i.e. null values, or UNKNOWN, which I've previously described. For the purpose of this analysis, I am going to drop rows where these values occur.

```
shooting_incidents <- na.omit(shooting_incidents)
shooting_incidents_complete <- shooting_incidents[!(shooting_incidents$PERP_AGE_GROUP ==
    "UNKNOWN" | shooting_incidents$PERP_RACE == "UNKNOWN" | shooting_incidents$PERP_SEX ==
    "U"), ]
```

It's worth noting that after omitting the rows with null values, we reduce the number of rows from 23,568 to 15,109. By further removing rows where the perpetrator information is listed as **UNKNOWN**, we reduce the number of rows to 11,768 observations.

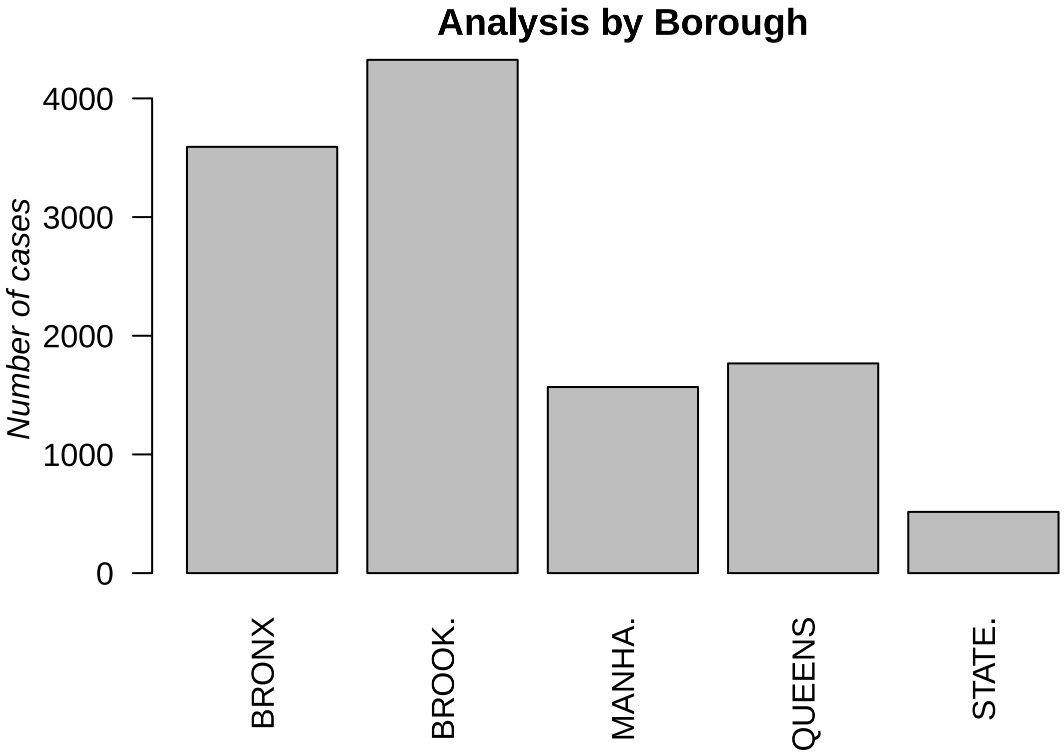
iv) Visualizations & Analysis

I will start with a simple grouping of the data based on the five boroughs of New York. I'll first look at the counts and then drill down into more information per borough.

```
by_borough <- (shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$BORO) %>%
    summarize(n()))
```

```
par(mar = c(7, 4, 2, 4))

barplot(height = by_borough$`n()`, names = c("BRONX", "BROOK.", "MANHA.", "QUEENS",
    "STATE."), ylab = "Number of cases", main = "Analysis by Borough", font.lab = 3,
    las = 2)
```



From the above data, we can see that Brooklyn has the most number of shooting incidents with a total of 4,382 and Staten Island has the least with a total of 521 incidents over the 2006-2020 identified period.

Next, we can analyze the same data, but perform the grouping on the race of the perpetrator.

```
shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$PERP_RACE) %>%
    summarize(n())
```

```
## # A tibble: 6 x 2
##   `shooting_incidents_complete$PERP_RACE` `n()`
##   <chr>                                   <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE              2
## 2 ASIAN / PACIFIC ISLANDER                  109
## 3 BLACK                                    8594
## 4 BLACK HISPANIC                            986
## 5 WHITE                                     243
## 6 WHITE HISPANIC                           1834
```

From the output above, we can see the number of incidents caused by each race. We can do a similar analysis on the victims.

```
shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$VIC_RACE) %>%
    summarize(n())
```

```
## # A tibble: 7 x 2
##   ‘shooting_incidents_complete$VIC_RACE‘ ‘n()‘
##   <chr>                                  <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE             3
## 2 ASIAN / PACIFIC ISLANDER                 193
## 3 BLACK                                   7903
## 4 BLACK HISPANIC                          1197
## 5 UNKNOWN                                   55
## 6 WHITE                                    403
## 7 WHITE HISPANIC                          2014
```

Boroughs and race aside, we can also group by age groups of the perpetrators and victims.

```
shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$PERP_AGE_GROUP) %>%
    summarize(n())
```

```
## # A tibble: 8 x 2
##   ‘shooting_incidents_complete$PERP_AGE_GROUP‘ ‘n()‘
##   <chr>                                        <int>
## 1 <18                                           1335
## 2 1020                                             1
## 3 18-24                                         5360
## 4 224                                              1
## 5 25-44                                         4540
## 6 45-64                                          477
## 7 65+                                             53
## 8 940                                              1
```

From the output above, you can notice odd values such as *224*, *940* and *1020*, which were not caught in the data cleaning process. Running `summary()` on this data did not pick up these values as this column is of a string type. With those three rows aside, we can see that the majority of the perpetrators fall between 18-24 years old. However, there's a very alarming number of 1,335 incidents caused by minors.

We can also run a similar grouping over the victims as shown below.

```
shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$VIC_AGE_GROUP) %>%
    summarize(n())
```

```
## # A tibble: 6 x 2
##   `shooting_incidents_complete$VIC_AGE_GROUP` `n()`
##   <chr>                                       <int>
## 1 <18                                          1349
## 2 18-24                                        4285
## 3 25-44                                        5107
## 4 45-64                                         873
## 5 65+                                            99
## 6 UNKNOWN                                        55
```

We can note that there seems to be a similarity between the counts of the age groups. A question that comes up is: is there a link between the perpetrator's age group and victim's age group?

```
metadata <- (shooting_incidents_complete %>%
    filter(shooting_incidents_complete$PERP_AGE_GROUP == shooting_incidents_complete$VIC_AGE_GROUP))

metadata %>%
    group_by(metadata$PERP_AGE_GROUP) %>%
    summarize(n())
```

```
## # A tibble: 5 x 2
##   `metadata$PERP_AGE_GROUP` `n()`
##   <chr>                     <int>
## 1 <18                         404
## 2 18-24                      2408
## 3 25-44                      2584
## 4 45-64                       131
## 5 65+                          10
```

Another question that might arise, is how many of those incidents led to murder?

```
shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$STATISTICAL_MURDER_FLAG) %>%
    summarize(n())
```

```
## # A tibble: 2 x 2
##   `shooting_incidents_complete$STATISTICAL_MURDER_FLAG` `n()`
##   <lgl>                                                 <int>
## 1 FALSE                                                  9013
## 2 TRUE                                                   2755
```

From the 11,768 filtered down cases, we have 2,755 murders.

We can look into the sex breakdown of the perpetrators.

```
shooting_incidents_complete %>%
    group_by(shooting_incidents_complete$PERP_SEX) %>%
    summarize(n())
```

```
## # A tibble: 2 x 2
##   `shooting_incidents_complete$PERP_SEX` `n()`
##   <chr>                                  <int>
## 1 F                                        314
## 2 M                                      11454
```

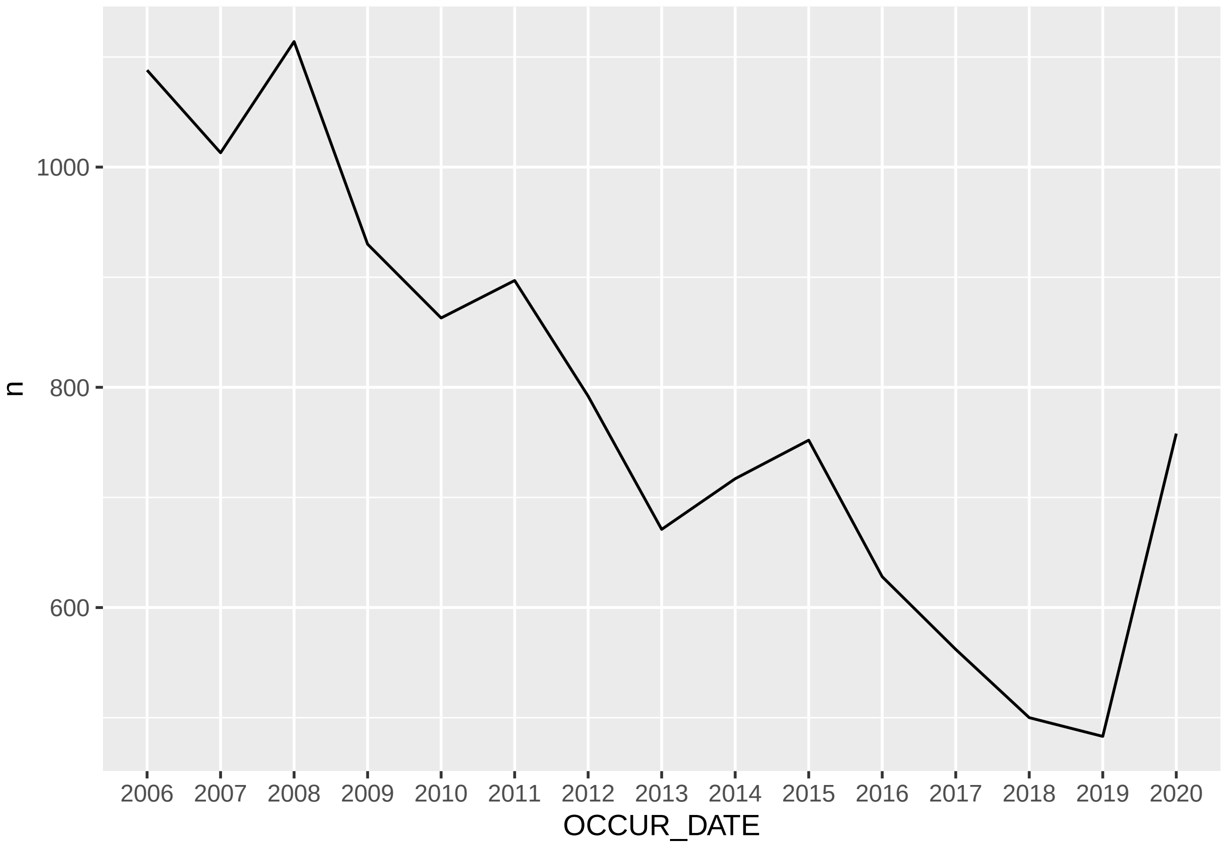
As we can see, an overwhelming majority of the shooting incidents have been caused by males.

We can also look at the trend of the shooting incidents with the years. First of all, our date column contains the full date. To be able to perform this analysis, we can format the dataframe with that specific column to just include the year.

```
shooting_incidents_with_year <- shooting_incidents_complete
shooting_incidents_with_year$OCCUR_DATE = format(shooting_incidents_with_year$OCCUR_DATE,
    format = "%Y")
```

Now we can apply a similar grouping and plot the results.

```
yearly_counts <- shooting_incidents_with_year %>%
    count(OCCUR_DATE)
ggplot(data = yearly_counts, aes(x = OCCUR_DATE, y = n, group = 1)) + geom_line()
```



Based on the grouping by year, we can actually see a trend of reduction in the number of incidents from 2006 to 2019 but then a sharp increase from 2019 to 2020. With the current dataset, we do not have enough information to make any conclusions from this trend. We can gather more data to identify reasons why there are increases in the number of incidents (such as election years or other contributing factors).

v) Bias analysis

With regards to bias, I personally do not come from the U.S. to understand the culture in New York in particular. However, hearing about issues with racism there, I can gather that the police in general do collect this data with bias. As previously mentioned, some of this data is collected based on patrols. Bias in police officers could cause them to patrol areas that are primarily populated with a certain race. This would make the probability of them encountering an incident by a specific race higher. This becomes a self-fulfilling prophecy for police officers, who as a result increase patrols in those areas.

As for the data science process I have followed, it is important to note that by dropping null values and unknowns, we have lost more than half of the data that we originally had. I do not know what the effect of losing this data does for the results shown in this report. It would have been best to note down any results before and after removing this data. This would ensure that dropping these rows did not lead to any biases.

Conclusion

In this report, I have examined the data from NYPD shooting incidents over the years. I've applied several transformations to the data to allow me to work with it and perform my analysis. Analysis was done based on the boroughs, race, sex, age group and year.