NYPD Shooting Incident Analysis

2022-11-23

Project Description

This project is an analysis of NYPD shooting incident data gathered from the Office of Management Analysis and Planning. This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

Import and Set-up Data

First, I'm going to import the relevant data into the R session using `read_csv`.

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

```
## Rows: 25596 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.
```

Tidying the Dataset

You'll notice that the `OCCUR_DATE` variable is currently stored as a character vector. I'll use lubridate to make this a proper date object.

```
data$OCCUR_DATE <- mdy(data$OCCUR_DATE)
```

There are also some variables in the dataset that we won't use for the purpose of this analysis.

Namely, these include:

- LOCATION_DESC
- X_COORD_CD
- Y_COORD_CD
- Lon_Lat

The following code block will remove these columns.

```
data = subset(data, select = -c(LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat))
```

I'll run a summary of the data now to make sure that everything looks good.

```
summary(data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME            BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596       Length:25596
```

```
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms        Class :character
##  Median : 86437258   Median :2012-08-26   Class2:difftime   Mode  :character
##  Mean   :112382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238490103   Max.   :2021-12-31
##
##     PRECINCT      JURISDICTION_CODE STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Min.   :  1.00   Min.   :0.0000    Mode :logical           Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    FALSE:20668             Class :character
##  Median : 69.00   Median :0.0000    TRUE :4928              Mode  :character
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE            Latitude       Longitude
##  Length:25596       Min.   :40.51   Min.   :-74.25
##  Class :character   1st Qu.:40.67   1st Qu.:-73.94
##  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
##
```

Looking at the summary, things appear to be good to go. Since a lot of this data is categorical, there aren't too many outliers to deal with at this point.

Analyzing the Data

Shootings per Borough

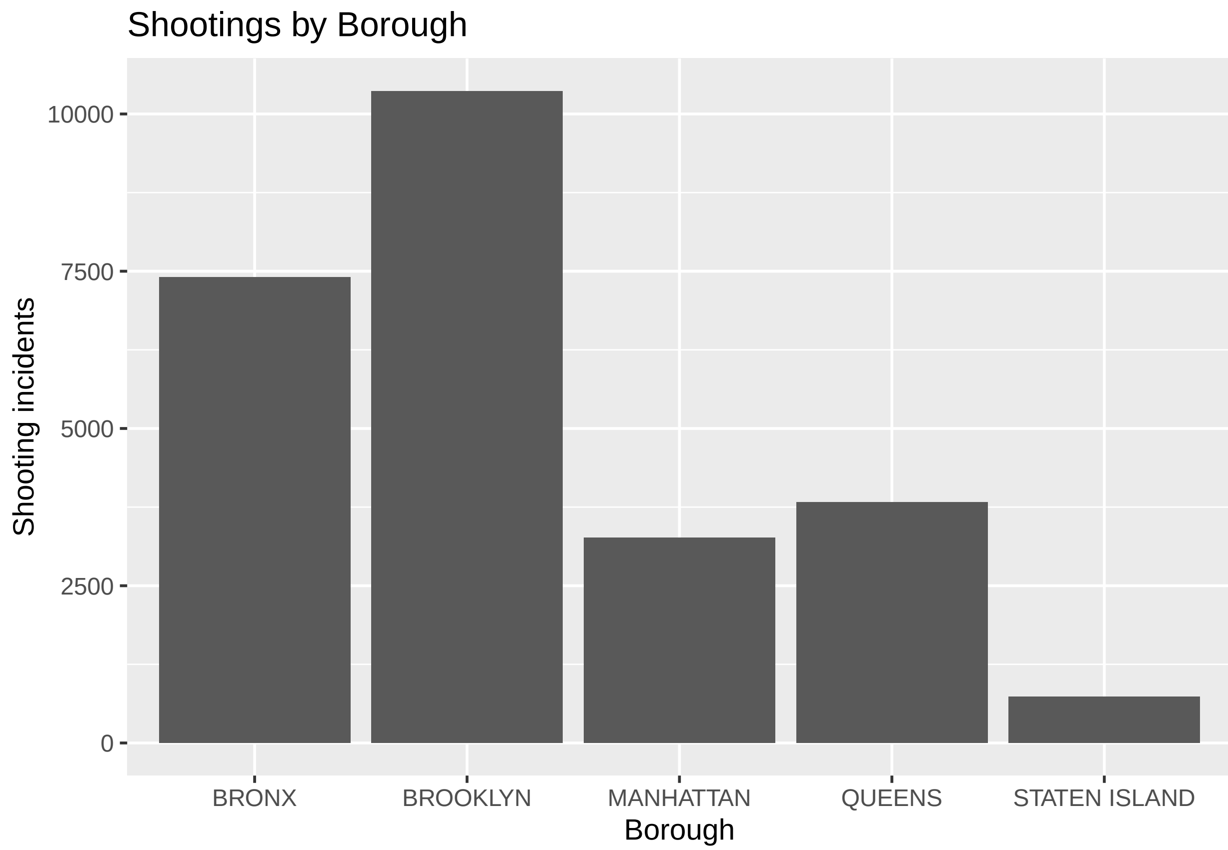
At this point, the data is present in the environment and ready to be analyzed.

One question that I would like to explore is which boroughs of New York City have the most shooting incidents.

We can perform a visualization on this dataset to get the answer to this question.

Using `ggplot2`, we can create a bar graph which will show incidents per borough.

```
ggplot(data=data, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Shootings by Borough",
       x = "Borough",
       y = "Shooting incidents")
```

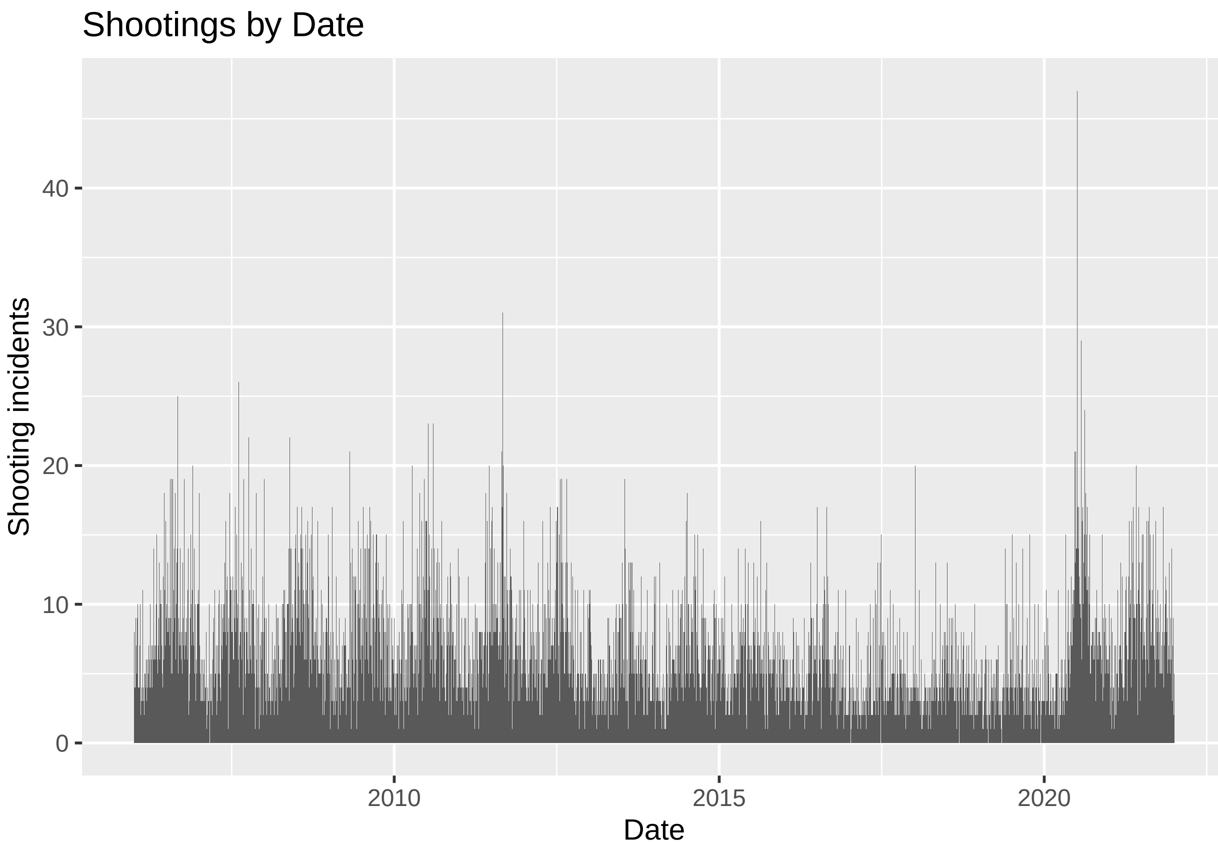


Here, we can see that Brooklyn has significantly more shooting incidents than other boroughs. Please keep in mind that this is raw total shootings, and isn't controlled for population. We can't say that Brooklyn is more dangerous than the Bronx, for example, since this analysis is not per-capita shooting incidents.

Shooting Incidents by Date

Another question that I would like to explore is whether or not certain days are more dangerous than others with regard to shooting incidents.

```
ggplot(data=data, aes(x = OCCUR_DATE)) +
  geom_bar() +
  labs(title = "Shootings by Date",
       x = "Date",
       y = "Shooting incidents")
```



As we can see from the barchart, it looks like some days are definitely more dangerous than others. There also seems to be some kind of frequency to which days are more deadly, at first glance. My suspicion is that this is different days of the week which are more deadly than others. I'll attempt to demonstrate this here.

Since we have the date of each occurrence, we can use the `wday` function from `lubridate` to find the day of the week, as follows:

```
data$DAY_OF_WEEK <- wday(data$OCCUR_DATE, label = TRUE)

table(data$DAY_OF_WEEK)
```

```
##
##  Sun  Mon  Tue  Wed  Thu  Fri  Sat
## 5156 3597 2945 2818 2809 3384 4887
```

As we can see, the number of murders is significantly higher around the weekend.

Modeling the Data

Next, we want to use this data to create a model that can predict future data points.

This dataset includes a variable called `STATISTICAL_MURDER_FLAG` which indicates if the shooting incident is likely a murder or not. Next, I will attempt to use regression to determine if a given observation will trigger this statistical murder flag using the other variables at our disposal.

This is a primary candidate for a type of regression called logistic regression - which is used to predict the likely outcome of a situation based on variables at our disposal. In this case, we are trying to predict `STATISTICAL_MURDER_FLAG` from the other variables in our data.

```r
glm.fit <- glm(STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + DAY_OF_WEEK + Latitude
summary(glm.fit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     PERP_AGE_GROUP + DAY_OF_WEEK + Latitude + Longitude, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4912  -0.7434  -0.6543  -0.1967   3.0187
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -18.67443  231.03650  -0.081 0.935578
## PERP_RACEASIAN / PACIFIC ISLANDER  12.02026  229.60727   0.052 0.958249
## PERP_RACEBLACK                   11.55532  229.60720   0.050 0.959862
## PERP_RACEBLACK HISPANIC          11.38060  229.60721   0.050 0.960469
## PERP_RACEUNKNOWN                 10.89605  229.60730   0.047 0.962150
## PERP_RACEWHITE                   12.20271  229.60724   0.053 0.957616
## PERP_RACEWHITE HISPANIC          11.66319  229.60720   0.051 0.959488
## PERP_SEXM                        -0.19233    0.12105  -1.589 0.112109
## PERP_SEXU                         1.52051    0.28849   5.270 1.36e-07 ***
## PERP_AGE_GROUP1020              -11.15860  324.74371  -0.034 0.972589
## PERP_AGE_GROUP18-24               0.17709    0.07526   2.353 0.018619 *
## PERP_AGE_GROUP224               -11.10836  324.74371  -0.034 0.972712
## PERP_AGE_GROUP25-44               0.50935    0.07494   6.797 1.07e-11 ***
## PERP_AGE_GROUP45-64               0.82987    0.11451   7.247 4.26e-13 ***
## PERP_AGE_GROUP65+                 1.03713    0.28275   3.668 0.000244 ***
## PERP_AGE_GROUP940               -11.11327  324.74371  -0.034 0.972700
## PERP_AGE_GROUPUNKNOWN            -2.39628    0.18068 -13.263  < 2e-16 ***
## DAY_OF_WEEK.L                    -0.06240    0.05013  -1.245 0.213289
## DAY_OF_WEEK.Q                    -0.01663    0.05241  -0.317 0.751013
## DAY_OF_WEEK.C                    -0.04430    0.05397  -0.821 0.411714
## DAY_OF_WEEK^4                    -0.05088    0.05471  -0.930 0.352369
## DAY_OF_WEEK^5                    -0.02077    0.05745  -0.362 0.717714
## DAY_OF_WEEK^6                    -0.04300    0.05904  -0.728 0.466417
## Latitude                          0.50248    0.23932   2.100 0.035764 *
## Longitude                         0.19868    0.29652   0.670 0.502837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16186  on 16251  degrees of freedom
## Residual deviance: 15081  on 16227  degrees of freedom
##   (9344 observations deleted due to missingness)
## AIC: 15131
##
## Number of Fisher Scoring iterations: 11
```

From this model, we can see that there are several statistically significant variables in our data. These include:

- Perpetrator Sex Unknown (PERP_SEXU)

- Perpetrator Age Group 18-24 (PERP_AGE_GROUP18-24)
- Perpetrator Age Group 25-44 (PERP_AGE_GROUP25-44)
- Perpetrator Age Group 45-64 (PERP_AGE_GROUP18-24)
- Perpetrator Age Group 65+ (PERP_AGE_GROUP65+)
- Perpetrator Age Group Unknown (PERP_AGE_GROUPUNKNOWN)
- Latitude

Interestingly, you will notice that the day of the week is NOT statistically significant in this model - meaning that even though there are more shooting incidents on the weekends, it is not more likely that a murder will occur on any given day of the week.

Identifying Model Bias

In any model, it is of course important to factor in bias that may be affecting the observations in our dataset.

Right now in the United States, policing is a relatively controversial issue. Data shows that some populations in our country are unfairly targeted by police, which can lead to observations that are tainted with the same bias affecting those undeserved populations.

We need to consider that the source of these observations is the police department itself, so it could very well be possible that the data points are affected by the perspective of the police department. If that were to be true, however, we would expect to see some kind of significant skew (or other external pressure) applied to the dataset.

In this case, I was not able to find evidence of some type of skew or unexpected relationship in the data. It makes sense to me that the age of the perpetrator is significant to the murder flag. Also, it makes sense that being unable to determine the sex of the perpetrator is significant. I would assume that usually those who commit murders are trying to conceal their identity.

All things considered, while I am concious of the bias in this case, I don't think that there is any significant evidence of bias in my model at this time.