NYPD Shooting Incident Data Analysis

Bianca Verlangieri

2022-08-29

Setup

This is an R Markdown document describing the NYPD Shooting Incident Data Analysis. First we load in the appropriate libraries. Note that I suppressed the output from loading in these libraries.

```
library(RCurl)
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Data Download

Next we download the data from the URL (rather than storing it locally). We also print out a summary of the data we just loaded.

```
link_to_download = getURL("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLO
nypd_data <- read.csv(text = link_to_download)
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class :character
##  Median : 86437258   Mode  :character   Mode  :character   Mode  :character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   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_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  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_SEX            VIC_RACE            X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Cleaning of the Data

Now that we have looked at a brief summary of the data, we can start cleaning it up. First, we notice there is a column called OCCUR_DATE that is currently of type character. We'll use the code below to turn these entries into doubles, so it will be easier to sort, plot, and more. We can check the type of the column before and after we change it, and access each column in the data using the $ symbol.

```
typeof(nypd_data$OCCUR_DATE)
```

```
## [1] "character"
```

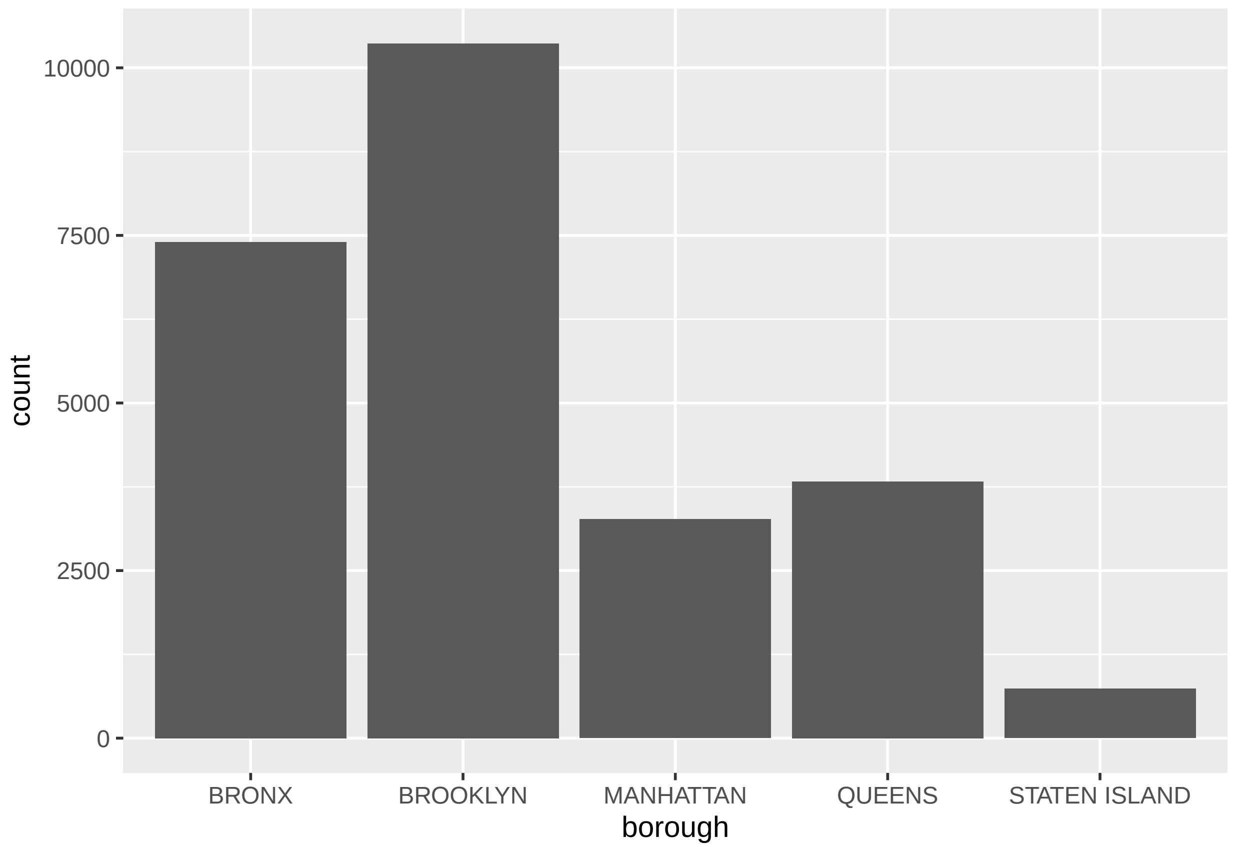
```
nypd_data$OCCUR_DATE <- mdy(nypd_data$OCCUR_DATE)
typeof(nypd_data$OCCUR_DATE)
```

```
## [1] "double"
```

Visualization

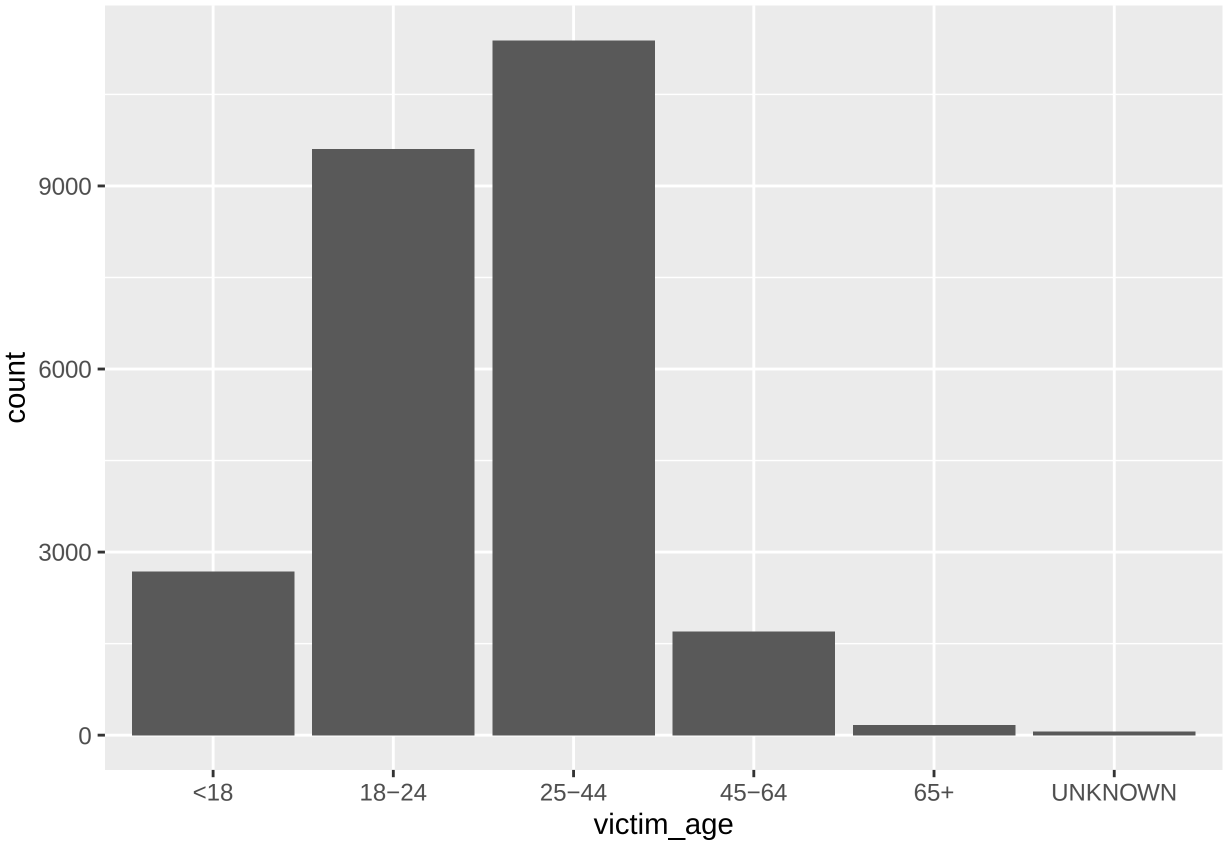
Now that we've cleaned up the data a bit let's visualize the data and see if we find anything interesting. First, let's look at a bar plot to show the number of shooting incidents per borough or town within New York City. Since this is categorical data, we need to convert it to a factor first. We'll use ggplot to help us plot the data.

```
borough = as.factor(nypd_data$BORO)
ggplot(data.frame(borough), aes(x=borough)) + geom_bar()
```

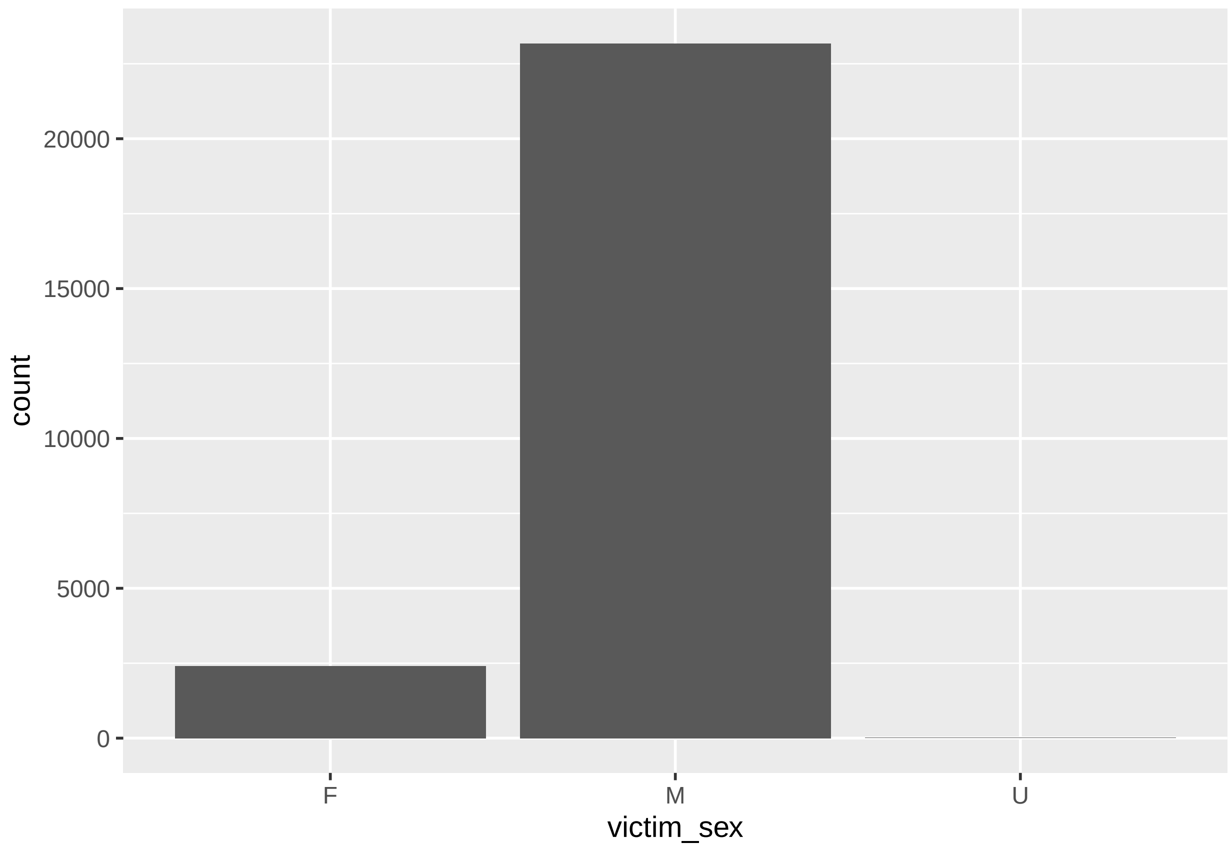


It looks like Brooklyn had the most shooting incidents in this data set, compared to the other boroughs. Next let's look at some victim categorical data, like age and sex.

```
victim_age = as.factor(nypd_data$VIC_AGE_GROUP)
ggplot(data.frame(victim_age), aes(x=victim_age)) + geom_bar()
```



```
victim_sex = as.factor(nypd_data$VIC_SEX)
ggplot(data.frame(victim_sex), aes(x=victim_sex)) + geom_bar()
```



From these counts, it looks like most victims were age 25-44 and male.

Analysis

Let's see if there are any relationships in this data by creating a binomial logistic regression model. Since male victims and shooting incidents in Brooklyn are most dominant in the shooting data, let's put those into our model.

```
nypd_data$isMale <- factor(nypd_data$VIC_SEX=="M")
nypd_data$isBrooklyn <- factor(nypd_data$BORO=="BROOKLYN")
mod <- glm(nypd_data$isMale ~ nypd_data$isBrooklyn, family = binomial())
summary(mod)
```

```
##
## Call:
## glm(formula = nypd_data$isMale ~ nypd_data$isBrooklyn, family = binomial())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1764   0.4434   0.4434   0.4476   0.4476
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.27016    0.02782  81.615   <2e-16 ***
## nypd_data$isBrooklynTRUE -0.01983    0.04350  -0.456    0.649
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15992  on 25595  degrees of freedom
## Residual deviance: 15992  on 25594  degrees of freedom
## AIC: 15996
##
## Number of Fisher Scoring iterations: 5
```

Since the p value is significant (under 0.05), these independent variables are significant.

Conclusions

We have just scraped the surface with this data. We identified in which borough the most shooting incidents occurred (Brooklyn) and that more males were victims than females. In using a binomial logistic regression model we identified that these independent variables are significant. This data has been collected since 2006, but is only reported and logged data, so there could exist bias in the data from incidents that were unreported. For example, there could be more responding officers in Brooklyn, so more shooting incidents are reported. Just an example, more investigation would need to be done to identify bias.