NYPD Historical Shooting Data

Amy Franks

2022-09-29

This public data set is a record of shooting incidents in New York City beginning in 2006. It contains available records of demographic information on the perpetrators, victims and associated information about the event. The data is posted on the NYPD website after extraction and review.

Step 1: Import Library

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

Step 2: Load Data

read_csv() reads comma delimited files

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

```
## 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.
```

```
spec(nypd)
```

```
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
```

```
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_double(),
##   Y_COORD_CD = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Step 3: Tidy and Transform Data

First remove columns that are unnecessary for the analysis, then ensure that the data types are correct. Then, explore the data set noting any missing data.

```
nypd <- nypd %>%
  select(-c(JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

nypd$OCCUR_DATE <- mdy(nypd$OCCUR_DATE)

colnames(nypd)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "PRECINCT"                "STATISTICAL_MURDER_FLAG"
##  [7] "PERP_AGE_GROUP"          "PERP_SEX"
##  [9] "PERP_RACE"               "VIC_AGE_GROUP"
## [11] "VIC_SEX"                 "VIC_RACE"
```

```
summary(nypd)
```

```
##   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      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Min.   :  1.00   Mode :logical           Length:25596       Length:25596
##  1st Qu.: 44.00   FALSE:20668             Class :character   Class :character
##  Median : 69.00   TRUE :4928              Mode  :character   Mode  :character
##  Mean   : 65.87
##  3rd Qu.: 81.00
##  Max.   :123.00
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
colSums(is.na(nypd))
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO                PRECINCT STATISTICAL_MURDER_FLAG
##                       0                        0                        0
##          PERP_AGE_GROUP                PERP_SEX                PERP_RACE
##                    9344                     9310                     9310
##           VIC_AGE_GROUP                 VIC_SEX                 VIC_RACE
##                       0                        0                        0
```

Step 4: Add Visualizations and Analysis

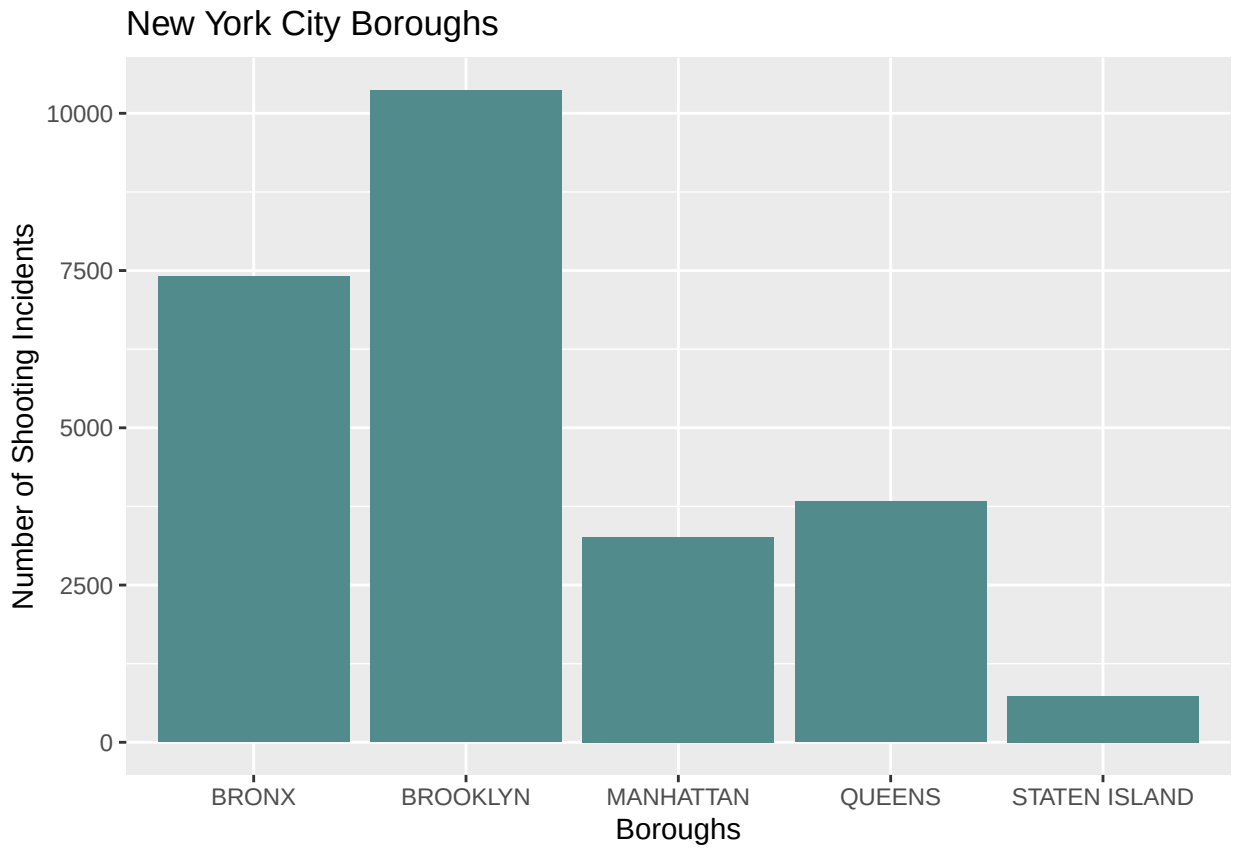
We can observe that a large proportion of the demographic data is missing. These discrepancies could have numerous points of origins including on going investigation, procedural differences between precincts or even record keeping errors. The precise reasons for this data's absence will not be speculated on here as a great deal of additional information would need to be included in the analysis. However, we can make some observations about what data is missing and from where. We can begin by noting which boroughs are most frequently represented.

```
nypd_w_freq <- nypd %>%
  group_by(BORO) %>%
  summarise(n = n()) %>%
  mutate(Freq = n/sum(n))

nypd_w_freq
```

```
## # A tibble: 5 x 3
##   BORO              n   Freq
##   <chr>         <int>  <dbl>
## 1 BRONX          7402 0.289
## 2 BROOKLYN      10365 0.405
## 3 MANHATTAN      3265 0.128
## 4 QUEENS         3828 0.150
## 5 STATEN ISLAND   736 0.0288
```

```
ggplot(data = nypd, aes(x = BORO)) +
  geom_bar(fill = "darkslategray4") +
  labs(title = "New York City Boroughs",
       x = "Boroughs",
       y = "Number of Shooting Incidents")
```



Now identify what proportion of the demographic data is missing.

```
nypd_count_na <- nypd %>%
     group_by(BORO) %>%
     summarise(across(PERP_AGE_GROUP:PERP_RACE,  ~sum(is.na(.))))

nypd_count_na
```

```
## # A tibble: 5 x 4
##   BORO          PERP_AGE_GROUP PERP_SEX PERP_RACE
##   <chr>                  <int>    <int>     <int>
## 1 BRONX                   2512     2506      2506
## 2 BROOKLYN                4291     4281      4281
## 3 MANHATTAN               1030     1024      1024
## 4 QUEENS                  1366     1357      1357
## 5 STATEN ISLAND            145      142       142
```

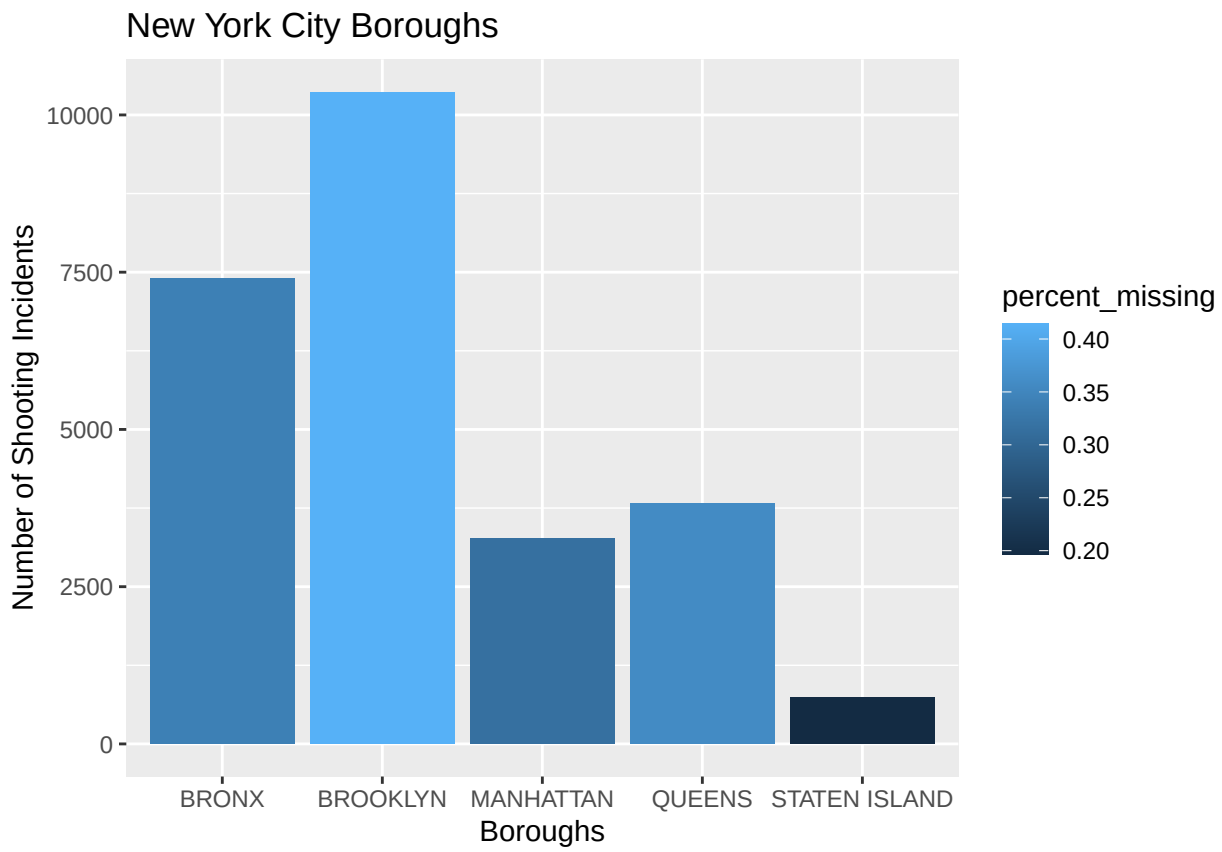
```
nypd_count_na <- nypd_count_na %>%
  full_join(nypd_w_freq) %>%
  mutate(percent_missing = PERP_AGE_GROUP /n)%>%
  mutate(percent_recorded = 1 - percent_missing)
```

```
## Joining, by = "BORO"
```

```
nypd_count_na
```

```
## # A tibble: 5 x 8
##   BORO          PERP_AGE_GROUP PERP_SEX PERP_RACE     n   Freq percent~1 perce~2
##   <chr>                  <int>    <int>     <int> <int>  <dbl>     <dbl>   <dbl>
## 1 BRONX                   2512     2506      2506  7402 0.289      0.339   0.661
## 2 BROOKLYN                4291     4281      4281 10365 0.405      0.414   0.586
## 3 MANHATTAN               1030     1024      1024  3265 0.128      0.315   0.685
## 4 QUEENS                  1366     1357      1357  3828 0.150      0.357   0.643
## 5 STATEN ISLAND            145      142       142   736 0.0288     0.197   0.803
## # ... with abbreviated variable names 1: percent_missing, 2: percent_recorded
```

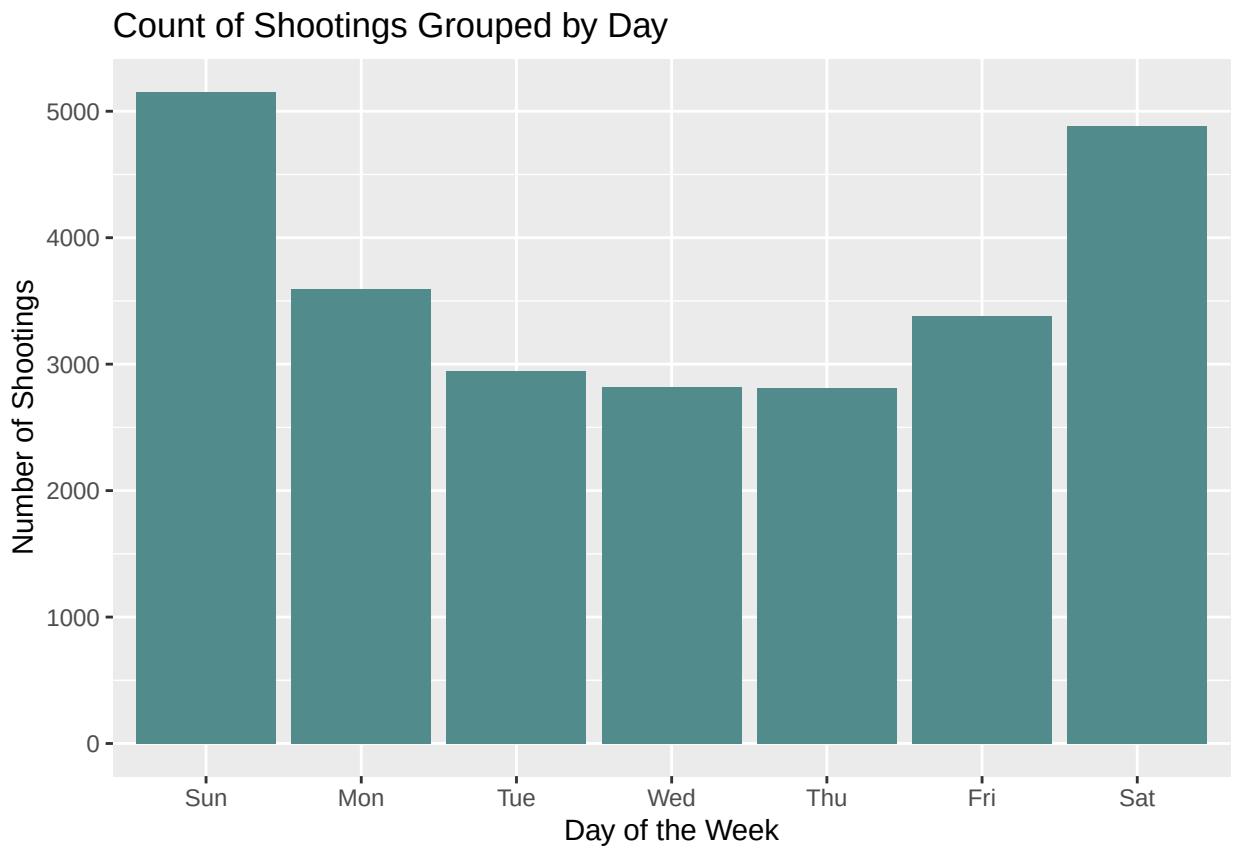
```
ggplot(data = nypd_count_na, aes(
  x = BORO,
  y = n,
  fill = percent_missing)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "New York City Boroughs",
       x = "Boroughs",
       y = "Number of Shooting Incidents")
```



We can see that Brooklyn is both missing the most data and is represented in our data set most frequently. Now we can also observe what day is most frequently represented.

```r
nypd$day <- wday(nypd$OCCUR_DATE, label = TRUE)

ggplot(data = nypd, aes(x = day)) +
  geom_bar(fill = "darkslategray4") +
  labs(title = "Count of Shootings Grouped by Day", x = "Day of the Week", y = "Number of Shootings")
```



It appears that the weekends in New York are the most dangerous.

Step 5: Modeling

We now apply a simple linear model comparing the incidents flagged as murders against the day of the occurrence.

```r
nypd = nypd %>%
  replace_na(list(PERP_AGE_GROUP = "N/A", PERP_SEX = "N/A", PERP_RACE = "N/A"))

summary(lm(nypd$STATISTICAL_MURDER_FLAG ~ nypd$day))
```

```
##
## Call:
## lm(formula = nypd$STATISTICAL_MURDER_FLAG ~ nypd$day)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2065 -0.1958 -0.1905 -0.1786  0.8214
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1939190  0.0025329  76.560   <2e-16 ***
## nypd$day.L  -0.0037363  0.0060423  -0.618   0.5363
## nypd$day.Q  -0.0152227  0.0063662  -2.391   0.0168 *
## nypd$day.C  -0.0091118  0.0065728  -1.386   0.1657
## nypd$day^4  -0.0028903  0.0067514  -0.428   0.6686
## nypd$day^5  -0.0006606  0.0070636  -0.094   0.9255
## nypd$day^6  -0.0111925  0.0073311  -1.527   0.1268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3943 on 25589 degrees of freedom
## Multiple R-squared:  0.0004503,  Adjusted R-squared:  0.000216
## F-statistic: 1.922 on 6 and 25589 DF,  p-value: 0.07338
```

Based on this analysis, while shootings are more likely to occur on the weekends there is not a relationship between the day of the shooting and whether or not it becomes a murder.

Step 6: Identify Bias

I tend to be immediately suspicious of missing data. My background is in regulated laboratories where a well established chain of custody of samples is critical and all data management must be highly transparent. In order to address this data set without bias, I had to set aside my natural suspicion of missing and proceed with the assumption that no malice was intended and there were other explanations for it's absence.