# NYPD Shooting Incident Report

Gabor Schulz

5/19/2021

## Step 1: Import the project dataset

Data source: Shooting incident data recorded in NYC since 2006. https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

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

```
##
## -- Column specification --------------------------------------------------------
## 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_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

## Step 2: Tidy and transform the data

The data contains a date field which is currently stored as a string. We should convert that into a date. Also, we should convert categorical columns into factor columns. I'm replacing the UNKNOWN values in the PERP_RACE column as this is one of the factors I'm going to analyze in Step 3. Finally, we should drop columns we don't need, like the exact latitude and longitude.

```r
shooting <- shooting %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(BORO = fct_recode(BORO)) %>%
  mutate(PRECINCT = factor(PRECINCT)) %>%
  mutate(JURISDICTION_CODE = factor(JURISDICTION_CODE)) %>%
  mutate(PERP_AGE_GROUP = factor(PERP_AGE_GROUP)) %>%
  mutate(PERP_SEX = fct_recode(PERP_SEX)) %>%
  mutate(PERP_RACE = fct_recode(PERP_RACE)) %>%
  mutate(VIC_AGE_GROUP = fct_recode(VIC_AGE_GROUP)) %>%
  mutate(VIC_SEX = fct_recode(VIC_SEX)) %>%
  mutate(VIC_RACE = fct_recode(VIC_RACE)) %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

shooting$PERP_RACE[shooting$PERP_RACE == 'UNKNOWN'] <- NA
summary(shooting)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:23568
##  1st Qu.: 55317014   1st Qu.:2008-12-30   Class1:hms
##  Median : 83365370   Median :2012-02-26   Class2:difftime
##  Mean   :102218616   Mean   :2012-10-03   Mode  :numeric
##  3rd Qu.:150772442   3rd Qu.:2016-02-28
##  Max.   :222473262   Max.   :2020-12-31
##
##             BORO         PRECINCT     JURISDICTION_CODE LOCATION_DESC
##  BRONX        :6700   75      : 1367   0    :19624       Length:23568
##  BROOKLYN     :9722   73      : 1282   1    :   54       Class :character
##  MANHATTAN    :2921   67      : 1102   2    : 3888       Mode  :character
##  QUEENS       :3527   79      :  920   NA's:    2
##  STATEN ISLAND: 698   44      :  842
##                       47      :  815
##                       (Other):17240
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  Mode :logical           18-24  :5448   F   :  334
##  FALSE:19080             25-44  :4613   M   :13305
##  TRUE :4488              UNKNOWN:3156   U   : 1504
##                          <18    :1354   NA's: 8425
##                          45-64  : 481
##                          (Other):  57
##                          NA's   :8459
##                       PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK                   : 9855   <18    : 2525   F: 2195
##  WHITE HISPANIC          : 1961   18-24  : 9000   M:21353
##  BLACK HISPANIC          : 1081   25-44  :10287   U:   20
##  WHITE                   :  255   45-64  : 1536
##  ASIAN / PACIFIC ISLANDER:  120   65+    :  155
##  (Other)                 :    2   UNKNOWN:   65
##  NA's                    :10294
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  320
##  BLACK                         :16846
##  BLACK HISPANIC                : 2244
##   UNKNOWN                         :  102
##   WHITE                           :  615
##   WHITE HISPANIC                  : 3432
```

There are missing values in several columns: 1. **JURISDICTION_CODE**: 2 missing values. We could simply drop those 2 rows. 2. **PERP_AGE_GROUP**: contains two missing values: NA and UNKNOWN. These should be harmonized. If an age analysis is being done, we should drop those rows that do not contain the required data. Alternatively we could try to impute values, which could, however, distort the result.

**PERP_SEX** are **PERP_RACE** similar to **PERP_AGE_GROUP** in that they also have two different unknown values.

### Step 3: Visualizations and Analysis

```
shootings_per_boro <- shooting %>% group_by(BORO) %>% summarize(cases = n())
murders_per_boro <- merge(shooting %>% group_by(BORO, STATISTICAL_MURDER_FLAG) %>% summarize(cases = n(
```

```
## ‘summarise()‘ has grouped output by ’BORO’. You can override using the ‘.groups‘ argument.
```

```
murders_per_boro <- murders_per_boro %>% rename(cases = cases.x, total_cases = cases.y)
murders_per_boro <- murders_per_boro %>% mutate(pct = round(cases / total_cases * 100, 2))

shootings_per_perp_race <- shooting %>% group_by(PERP_RACE) %>% summarize(cases = n())
shootings_perp_race_vic_race <- merge(shooting %>% group_by(PERP_RACE, VIC_RACE) %>% summarize(cases = r
```

```
## ‘summarise()‘ has grouped output by ’PERP_RACE’. You can override using the ‘.groups‘ argument.
```

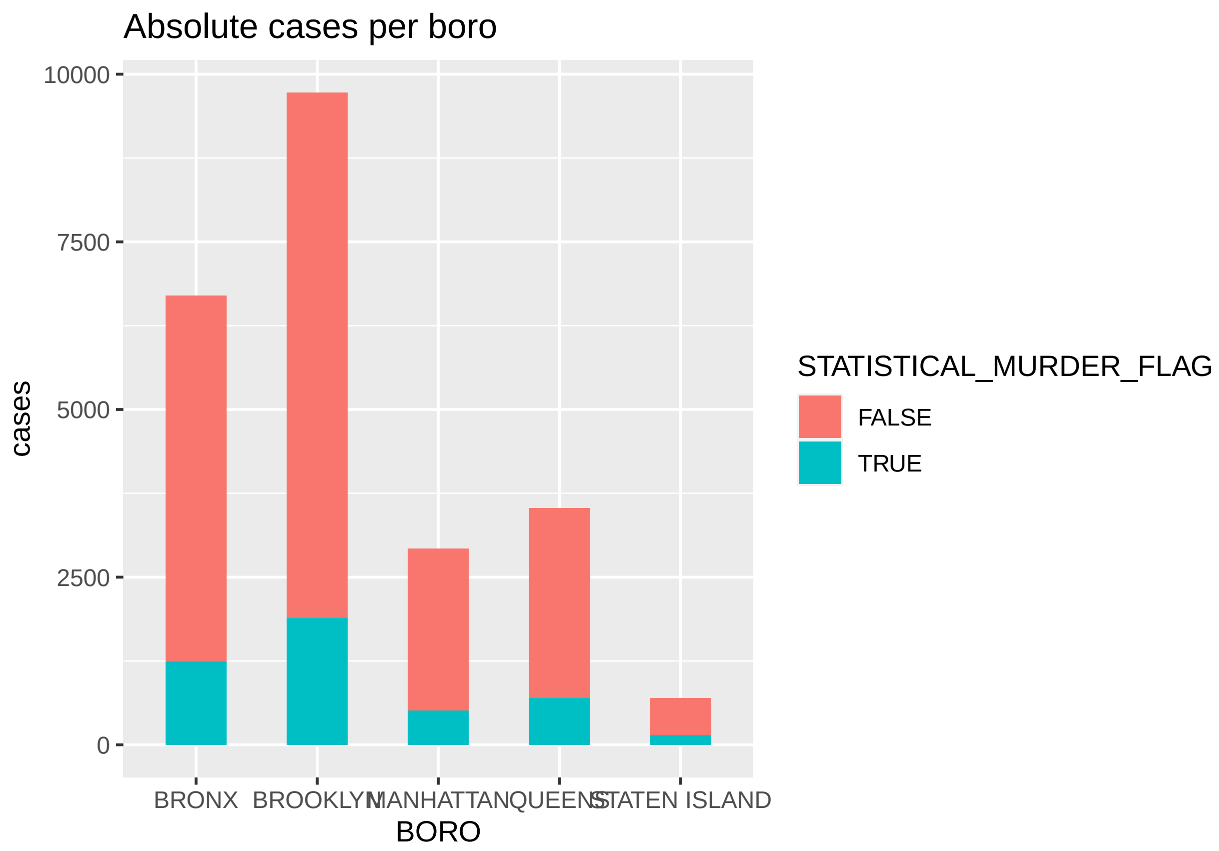
```
shootings_perp_race_vic_race <- shootings_perp_race_vic_race %>% rename(cases = cases.x, total_cases = c
shootings_perp_race_vic_race <- shootings_perp_race_vic_race %>% mutate(pct = round(cases / total_cases
```

Let’s look at the murder vs non-murder shootings per boro first:

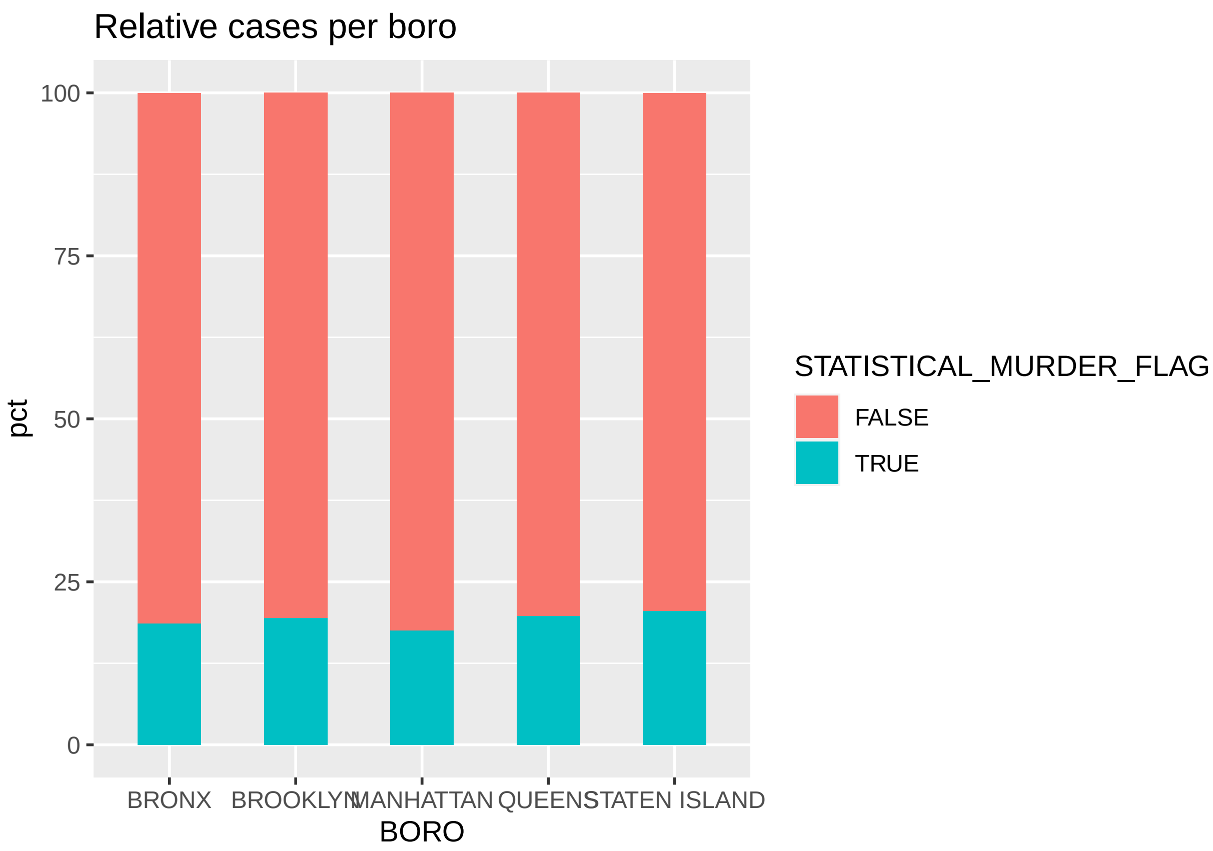
```
murders_per_boro
```

```
##             BORO STATISTICAL_MURDER_FLAG cases total_cases   pct
## 1          BRONX                   FALSE  5456        6700 81.43
## 2          BRONX                    TRUE  1244        6700 18.57
## 3       BROOKLYN                   FALSE  7830        9722 80.54
## 4       BROOKLYN                    TRUE  1892        9722 19.46
## 5      MANHATTAN                   FALSE  2409        2921 82.47
## 6      MANHATTAN                    TRUE   512        2921 17.53
## 7         QUEENS                   FALSE  2830        3527 80.24
## 8         QUEENS                    TRUE   697        3527 19.76
## 9  STATEN ISLAND                   FALSE   555         698 79.51
## 10 STATEN ISLAND                    TRUE   143         698 20.49
```

```
murders_per_boro %>% ggplot(aes(fill=STATISTICAL_MURDER_FLAG, x=BORO, y=cases)) +
  geom_bar(position="stack", stat="identity", width=0.5) +
  labs(title='Absolute cases per boro')
```

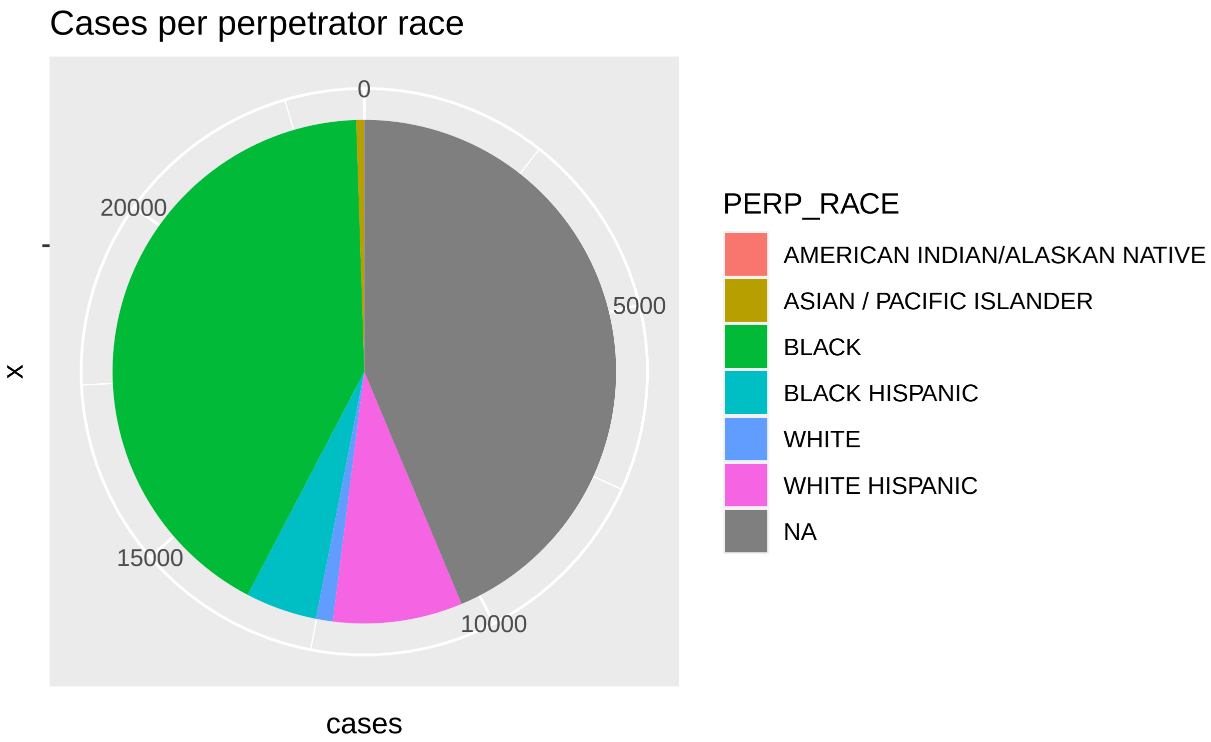


```
murders_per_boro %>% ggplot(aes(fill=STATISTICAL_MURDER_FLAG, x=BORO, y=pct)) +
  geom_bar(position="stack", stat="identity", width=0.5) +
  labs(title='Relative cases per boro')
```



It appears that Staten Island had the lowest number of shootings (698), but the highest proportion of murder cases (20.49 %). The highest number of shootings happened in Brooklyn (9722). Manhattan had the lowest proportion of murders (17.53 %).

```
shootings_per_perp_race %>% ggplot(aes(fill=PERP_RACE, x='', y=cases)) +
  geom_bar(position="stack", stat="identity", width=1) +
  coord_polar("y", start=0) +
  labs(title='Cases per perpetrator race')
```


Cases per perpetrator race

Looking at the race of perpetrators it is immediately visible that there is a huge proportion of unknown values. The 2nd largest group is black, while the smallest one is American Indian/Alaskan native.

## Conclusion and bias identification

There could be several sources of bias, both in the data and the analysis.

1. Sources of bias in the data

- The way the data is collected may be biased. E.g. there may be more points recorded in certain neighborhoods simply because of more intensive police activity in the area.
- There is a huge number of incomplete samples, which could make it more difficult to extract meaningful insights from the data

2. Sources of bias in the analysis

- The person performing the analysis could be influenced by their personal position on firearms, their race, their gender, etc.

In conclusion, this is a challenging data set because of the large number of missing values but also due to the potential political implications of the outcomes. If it's used for taking policy decisions, very thorough data cleaning is required which should involve a careful analysis of the potential effects of the decisions taken. This could best be done in co-operation with subject matter experts.