NYPD Shooting Incident Data Report

Anurag Balaji

2025-02-27

NYPD Data-Load and Tidy

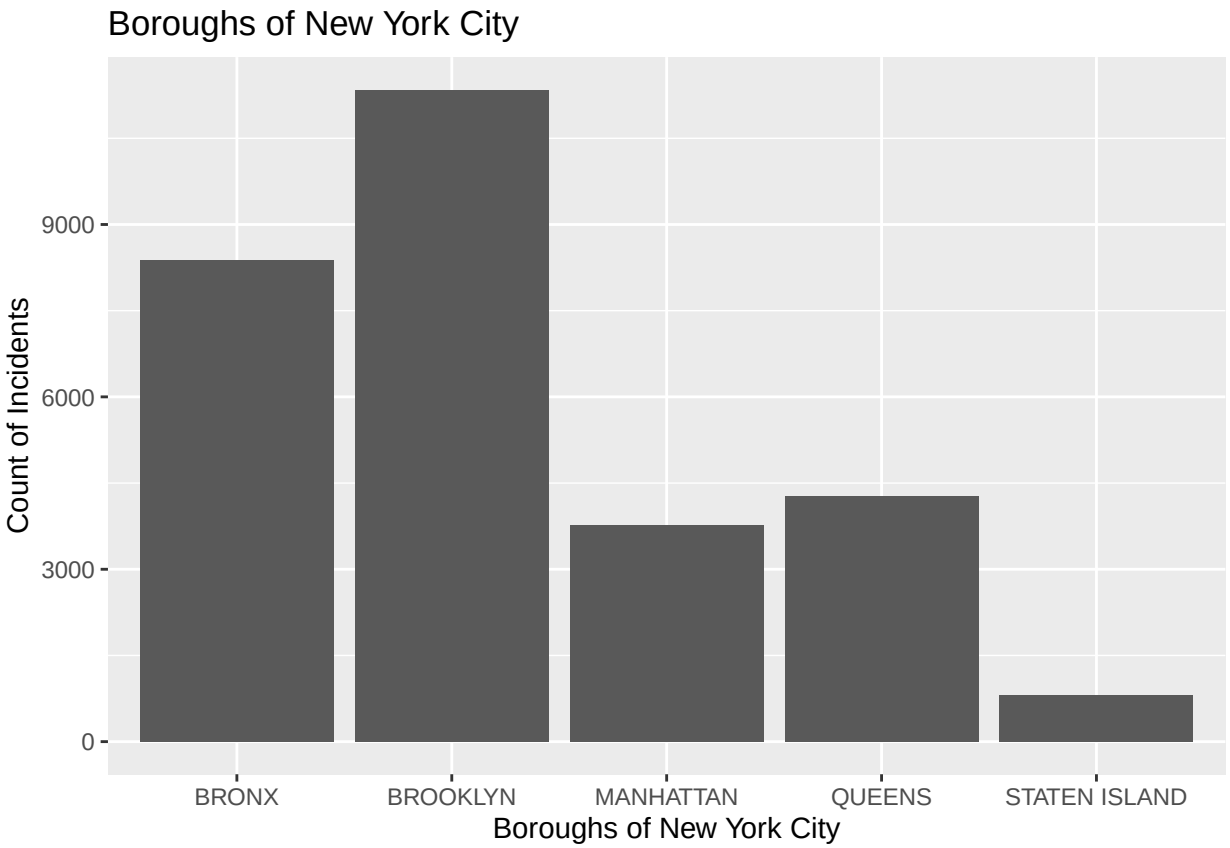
Below is a summary of the NYPD Shooting Data set. We begin by loading in the dataset and selecting only the relevant columns for our analysis. We then summarize the dataset

```
nypd=read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD",show_col
nypd = nypd %>% select(INCIDENT_KEY,
                    OCCUR_DATE,
                    OCCUR_TIME,
                    BORO,
                    STATISTICAL_MURDER_FLAG,
                    PERP_AGE_GROUP,
                    PERP_SEX,
                    PERP_RACE,
                    VIC_AGE_GROUP,
                    VIC_SEX,
                    VIC_RACE)
summary(nypd)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms         Class :character
##  Median : 92711254   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:28562       Length:28562
##  FALSE:23036             Class :character   Class :character
##  TRUE :5526              Mode  :character   Mode  :character
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

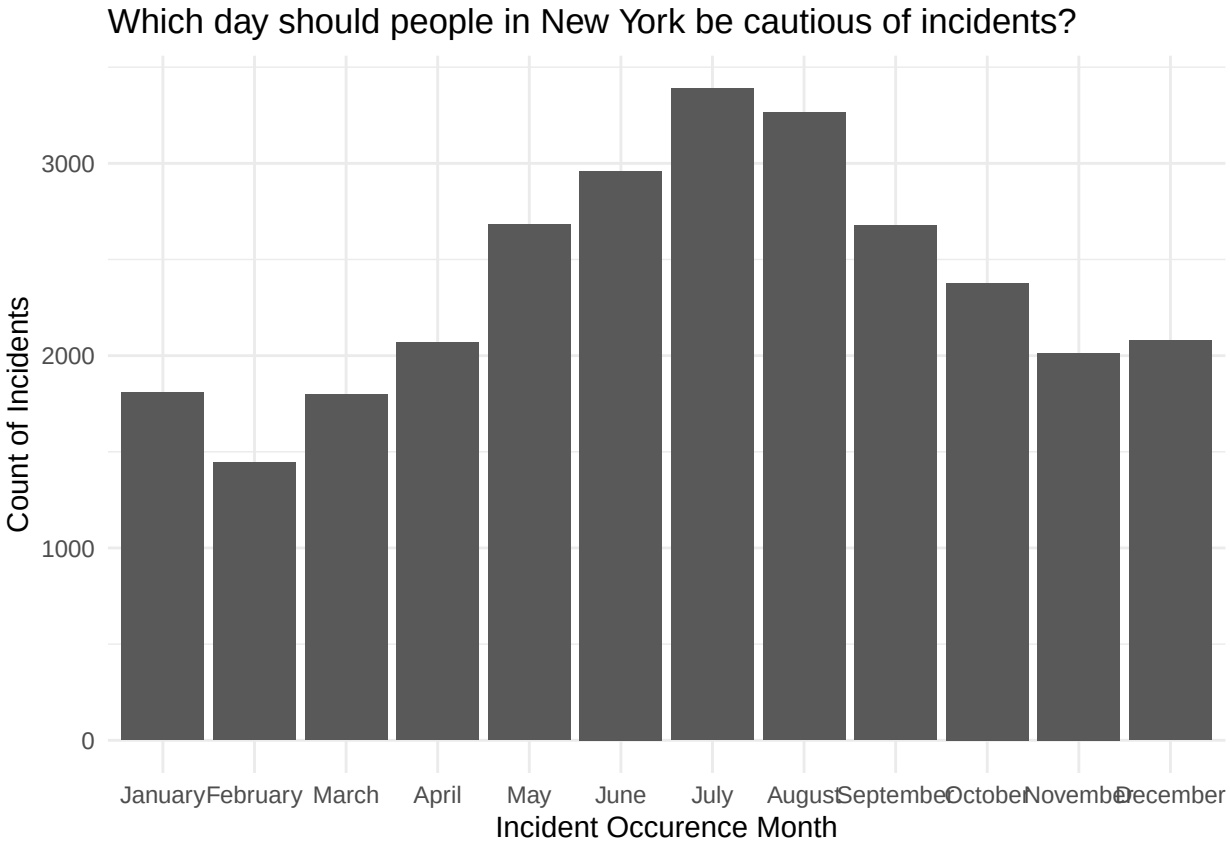
Analysis and Research Questions

1) What areas of New York City has the most number of incidents.



From the above graph, we can clearly see that Brooklyn is the neighbourhood with the most number of incidents in New York City with over 10000 incidents. Staten Island is the neighbourhood with the least number of incidents with less than 1000 incidents being reported. Individuals in Brooklyn should be more cautious and vigilant in New York City.

2) Which months saw the most incidents/ murders



From the above graph, we can see that July and August are the months with the most number of incidents in the year and February has the least number of incidents. We are also able to see overall that summer months and warmer months tend to have more incidents that the winter months. This may be mainly due to the fact that people tend to spend a lot more time outdoors during the warmer months making them potentially more vulnerable to shooting incidents in New York City.

3) Building a logistic regression model to predict whether an incident is a murder or not

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     BORO, family = binomial, data = nypd)
##
## Coefficients: (1 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -1.66962    0.08812 -18.946  < 2e-16
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -8.58355   84.06292  -0.102 0.918670
## PERP_RACEASIAN / PACIFIC ISLANDER        1.31850    0.30100   4.380 1.18e-05
## PERP_RACEBLACK                           0.82350    0.25016   3.292 0.000995
## PERP_RACEBLACK HISPANIC                  0.73988    0.25800   2.868 0.004135
## PERP_RACEUNKNOWN                        -0.82964    0.13101  -6.333 2.41e-10
## PERP_RACEWHITE                           1.62726    0.27744   5.865 4.48e-09
## PERP_RACEWHITE HISPANIC                  0.99983    0.25378   3.940 8.16e-05
## PERP_SEXF                               -0.12340    0.25836  -0.478 0.632913
## PERP_SEXM                               -0.39075    0.23501  -1.663 0.096377
## PERP_SEXU                                     NA         NA      NA       NA
```

```
## BOROBROOKLYN                              -0.13276    0.04564  -2.909 0.003624
## BOROMANHATTAN                             -0.15556    0.05861  -2.654 0.007946
## BOROQUEENS                                -0.13138    0.05767  -2.278 0.022716
## BOROSTATEN ISLAND                         -0.10203    0.10166  -1.004 0.315574
##
## (Intercept)                               ***
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER         ***
## PERP_RACEBLACK                            ***
## PERP_RACEBLACK HISPANIC                   **
## PERP_RACEUNKNOWN                          ***
## PERP_RACEWHITE                            ***
## PERP_RACEWHITE HISPANIC                   ***
## PERP_SEXF
## PERP_SEXM                                 .
## PERP_SEXU
## BOROBROOKLYN                              **
## BOROMANHATTAN                             **
## BOROQUEENS                                *
## BOROSTATEN ISLAND
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19230  on 19251  degrees of freedom
## Residual deviance: 18827  on 19238  degrees of freedom
##   (9310 observations deleted due to missingness)
## AIC: 18855
##
## Number of Fisher Scoring iterations: 9
```

Finally, we can build a logistic regression model to help predict whether an incident is going to be a murder or not. I will use the perpetrator sex, race and boroughs to determine if an incident is a murder or not. From the above model, we can see that an incident ocurring in Brooklyn changes the log odds of murder by -0.13.

4) Potential Sources of Bias

Potential sources of bias in analyzing NYPD shooting data include reporting bias, where certain incidents may be underreported or misclassified; selection bias, if the dataset does not capture all relevant cases; and data collection bias, influenced by how and why data is recorded. Demographic bias may arise if certain groups are over- or underrepresented due to systemic factors. Additionally, analytical bias, such as choosing specific metrics or framing results in a particular way, can impact interpretations. For example, one might feel that women might be more vulnerable to incidents or certain neighbourhoods like the Bronx might have the most incidents due to personal experiences or assumptions. Acknowledging these biases and testing them first is crucial to ensuring a fair, transparent, and data-driven analysis.