NYPD Shooting Incident - Analysis Report

Wai Lin Kyaw

2024-09-12

The dataset contains all the shooting incidents that occurred in NYC from 1st Jan of 2006 all the way to end of 2022. The dataset is published by the New York Police Department and it is still being updated regularly.

Each record contains followings:

- Date & time of the incident occurred
- Location of the incident
- Area/City where the incident occurred
- Victim information
- Perpetrator information

We will explore and analyse the all the aspects except the location.

Importing dataset

Dataset is already downloaded in the workspace. We will import the dataset and do basic inspection.

```
## Number of samples: 28562
```

```
## Number of variables: 21
```

```
## Variables:
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "LOC_OF_OCCUR_DESC"       "PRECINCT"
##  [7] "JURISDICTION_CODE"       "LOC_CLASSFCTN_DESC"
##  [9] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
## [11] "PERP_AGE_GROUP"          "PERP_SEX"
## [13] "PERP_RACE"               "VIC_AGE_GROUP"
## [15] "VIC_SEX"                 "VIC_RACE"
## [17] "X_COORD_CD"              "Y_COORD_CD"
## [19] "Latitude"                "Longitude"
## [21] "Lon_Lat"
```

Wrangling & cleaning dataset

We have about 21 variables here. We can get rid of the ones that we don't really need so that we can focus on what's important.

First, let's remove the variables related to location of the incident except borough.

```
incidents <- incidents %>%
  select(-c("PRECINCT",
           "JURISDICTION_CODE",
           "LOC_CLASSFCTN_DESC",
           "LOCATION_DESC",
           "X_COORD_CD",
           "Y_COORD_CD",
           "Latitude",
           "Longitude",
           "Lon_Lat"
           ))
```

We can inspect the dataset to observe NA and null values.

```
na_cols <- colSums(is.na(incidents) | is.null(incidents))
na_cols[na_cols > 0]
```

```
## LOC_OF_OCCUR_DESC    PERP_AGE_GROUP          PERP_SEX         PERP_RACE
##             25596              9344              9310              9310
```

As we can see, above 4 variables have a lot of missing values. It's important to impute these with appropriate values so that, when we do the analysis, it will have proper meaning and easy to interpret.

We replace those `NA` values with `Unknown`. And we will remove the columns where more than 75% of the rows have `NA`.

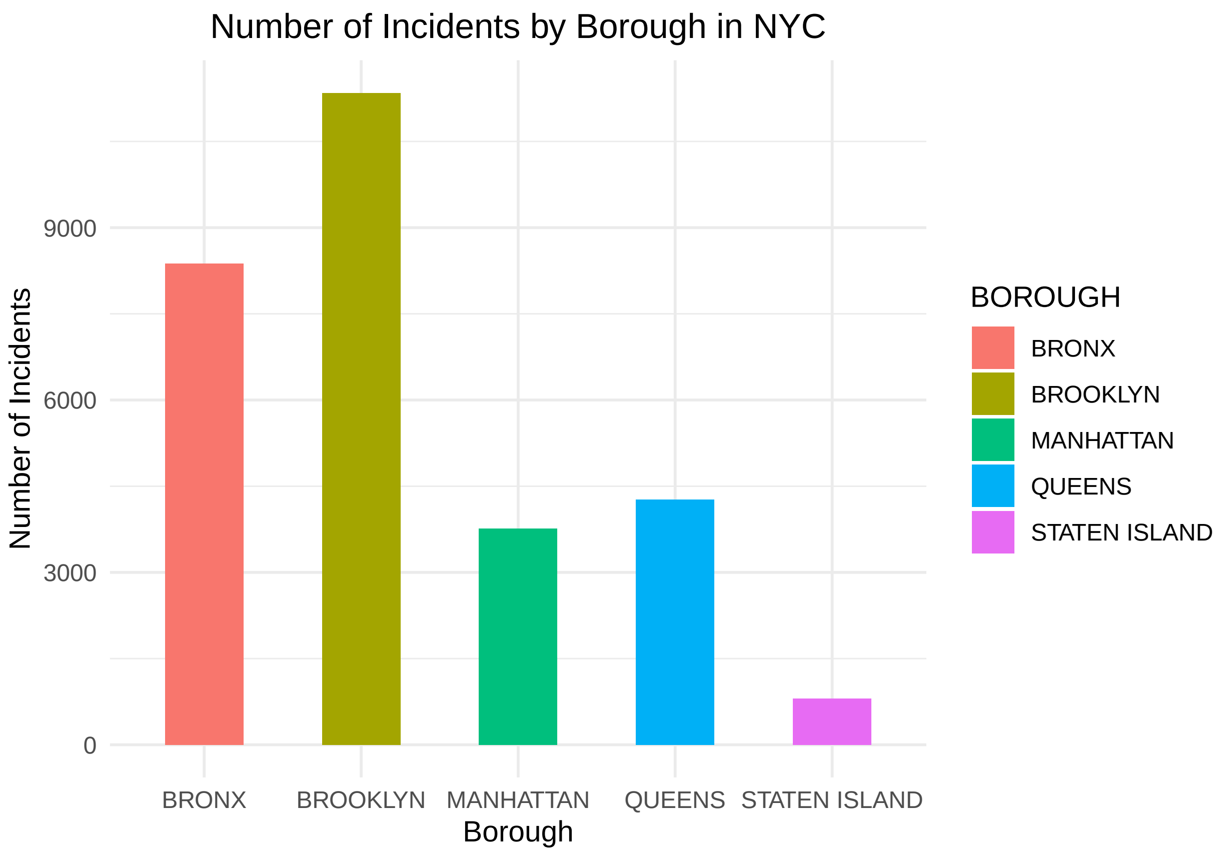
We are left with the tibble of 11 variables. It is a good time to transform the data to more relevant data types. We can typecast followings into factor (categorical value).

- boro (renaming this to borough)
- location description
- age group
- sex
- race

Exploratory Data Analysis

We are going to visualize the following three key aspects of the incidents:

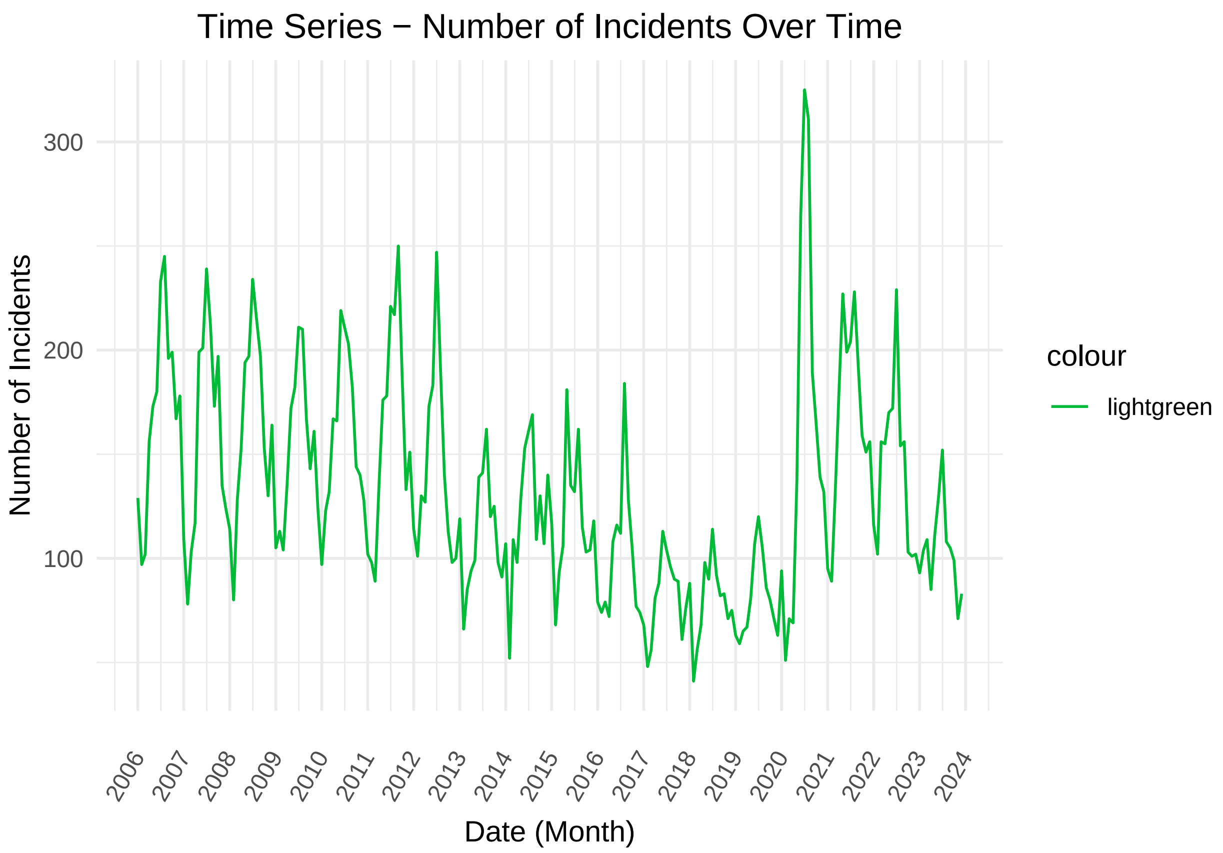
- Shootings by borough
- Time series of shooting incidents
- Victim age by incident



From this we can see that **Brooklyn** has the highest number of incidents followed by Bronx. Queens and Manhattan has the same amount of incidents.

I would say Staten Island has smaller number of incidents compared to the rest.

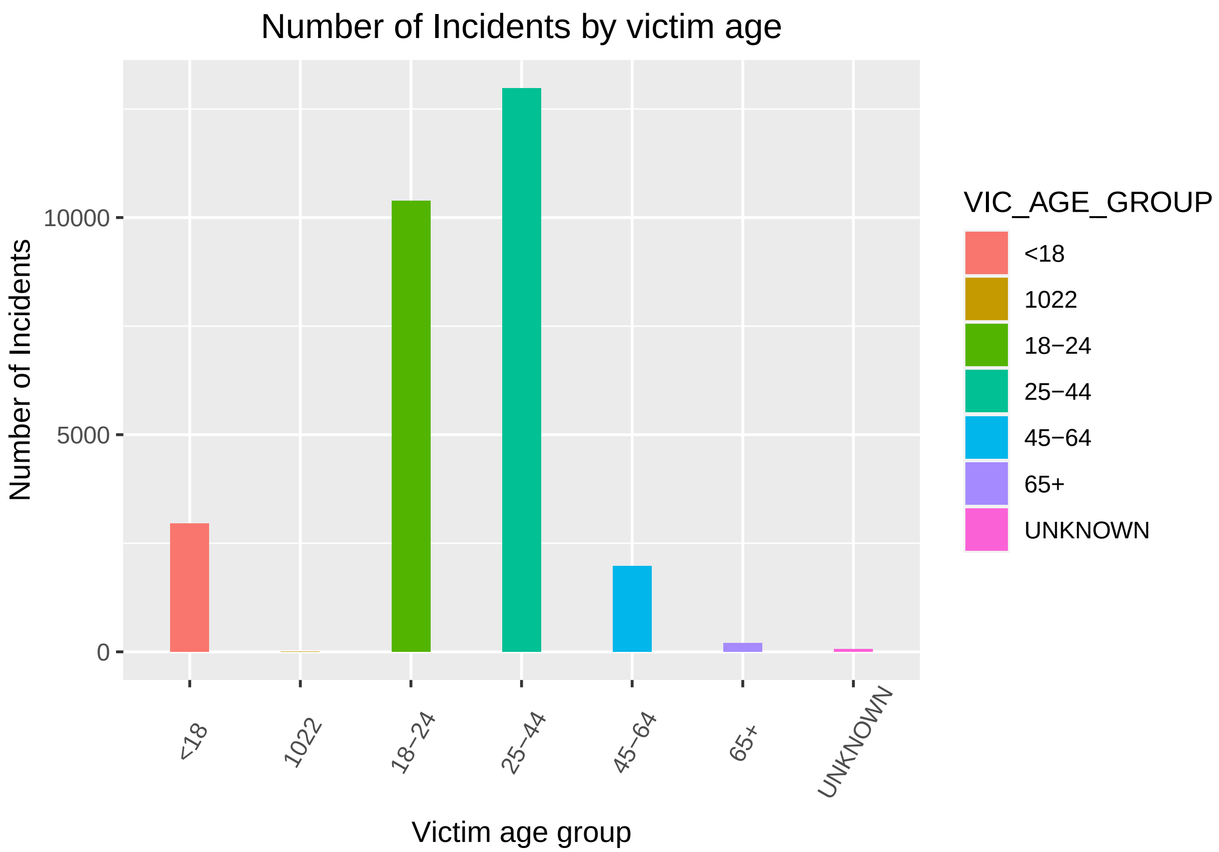
It is interesting that the rate in Brooklyn is super high while the Staten Island, only a bridge away has small number of incidents.



Here are some findings from this time series - visualization.

1. Ongoing incident rate on these 5 years term is same as the time back to 2006 all way to 2013.
2. From 2013 to 2020, it decreases bit by bit reaching the near bottom in 2020.
3. Interestingly enough, there was a spike during 2021. This is the lock down period when in Covid19 pandemic. Logically, people stay home, and the ongoing pandemic, it should've been lower than usual.

It is declining since after mid 2023. I hope this will continue.



As we can see from the last, those between 25-44 and 18-24 age groups are the most prone to the incidents.

Building a Model

Let's split the data into training and testing dataset first.

Now, we will build a machine learning model to predict whether is not a case is an instance of statistical murder or not.

To get a probability, we will use logistic regression which is design to squeeze the Infinite possibilities into the range of 0 and 1.

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BOROUGH + OCCUR_DATE +
##     OCCUR_TIME + PERP_AGE_GROUP + PERP_SEX + PERP_RACE + VIC_AGE_GROUP +
##     VIC_SEX + VIC_RACE, family = binomial, data = train_incidents)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.259e+01  3.424e+02  -0.066 0.947391
## BOROUGHBROOKLYN                  3.703e-02  4.338e-02   0.854 0.393315
## BOROUGHMANHATTAN                -1.109e-01  5.839e-02  -1.899 0.057591 .
## BOROUGHQUEENS                   -1.150e-02  5.537e-02  -0.208 0.835399
## BOROUGHSTATEN ISLAND            -1.540e-01  1.061e-01  -1.452 0.146620
## OCCUR_DATE                      -8.129e-05  9.389e-06  -8.658  < 2e-16 ***
## OCCUR_TIME                          1.095e-07  5.702e-07   0.192 0.847758
## PERP_AGE_GROUP1020                 -1.126e+01  3.247e+02  -0.035 0.972343
## PERP_AGE_GROUP1028                 -1.087e+01  3.247e+02  -0.033 0.973286
## PERP_AGE_GROUP18-24                 7.801e-02  8.026e-02   0.972 0.331103
## PERP_AGE_GROUP25-44                 3.693e-01  8.093e-02   4.563 5.04e-06 ***
## PERP_AGE_GROUP45-64                 6.733e-01  1.180e-01   5.706 1.16e-08 ***
## PERP_AGE_GROUP65+                   7.233e-01  3.074e-01   2.353 0.018638 *
## PERP_AGE_GROUP940                  -1.160e+01  3.247e+02  -0.036 0.971512
## PERP_AGE_GROUPUnknown              -2.613e+00  1.978e-01 -13.208  < 2e-16 ***
## PERP_SEXM                          -1.184e-01  1.271e-01  -0.931 0.351669
## PERP_SEXU                           1.695e+00  3.111e-01   5.447 5.13e-08 ***
## PERP_SEXUnknown                     2.934e+00  3.024e-01   9.704  < 2e-16 ***
## PERP_RACEASIAN / PACIFIC ISLANDER  1.197e+01  3.247e+02   0.037 0.970607
## PERP_RACEBLACK                      1.171e+01  3.247e+02   0.036 0.971233
## PERP_RACEBLACK HISPANIC             1.163e+01  3.247e+02   0.036 0.971441
## PERP_RACEUnknown                    1.117e+01  3.247e+02   0.034 0.972569
## PERP_RACEWHITE                      1.217e+01  3.247e+02   0.037 0.970117
## PERP_RACEWHITE HISPANIC             1.188e+01  3.247e+02   0.037 0.970825
## VIC_AGE_GROUP1022                  -1.050e+01  3.247e+02  -0.032 0.974203
## VIC_AGE_GROUP18-24                  2.638e-01  6.917e-02   3.813 0.000137 ***
## VIC_AGE_GROUP25-44                  5.149e-01  6.817e-02   7.553 4.26e-14 ***
## VIC_AGE_GROUP45-64                  6.047e-01  8.845e-02   6.836 8.12e-12 ***
## VIC_AGE_GROUP65+                    9.798e-01  1.829e-01   5.357 8.47e-08 ***
## VIC_AGE_GROUPUNKNOWN                5.521e-01  3.364e-01   1.641 0.100714
## VIC_SEXM                            4.697e-02  5.841e-02   0.804 0.421341
## VIC_SEXU                           -4.573e-01  1.087e+00  -0.421 0.674061
## VIC_RACEASIAN / PACIFIC ISLANDER   1.060e+01  1.085e+02   0.098 0.922164
## VIC_RACEBLACK                       1.047e+01  1.085e+02   0.097 0.923097
## VIC_RACEBLACK HISPANIC              1.026e+01  1.085e+02   0.095 0.924653
## VIC_RACEUNKNOWN                     9.559e+00  1.085e+02   0.088 0.929801
## VIC_RACEWHITE                       1.055e+01  1.085e+02   0.097 0.922547
## VIC_RACEWHITE HISPANIC              1.054e+01  1.085e+02   0.097 0.922616
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22449  on 22849  degrees of freedom
## Residual deviance: 21357  on 22812  degrees of freedom
## AIC: 21433
##
## Number of Fisher Scoring iterations: 11
```

Conclusion & Biases

We have built the logistic regression model here to predict the statistical murder case or not.

As next steps we could add - testing - evaluation - various metrics about the model that we just built.

Inspecting and looking through the dataset, having some exploratory analysis through visualization is a good approach to get ourselves to gain more exposure with the data that we are working with.

We should always do that before we dive into the modeling first.

Potential Biases

Here when we build the model, we are generalizing most of the features that we have. Although this seems to work, it might not be the case in practical.

For example, the factors contributing to the cases might be different during covid crisis compared to the ones before and after.

Or even there might be some other aspects that doesn't even included in the dataset.

Trying out with the real scenarios, talking with the industry experts and verify with the facts can reduce the potential biases.