NYPD Shooting Incident - Analysis Report

Wai Lin Kyaw

2024-09-09

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. We will keep `LOC_OF_OCCUR_DESC`.

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

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

After we typecast, 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
```

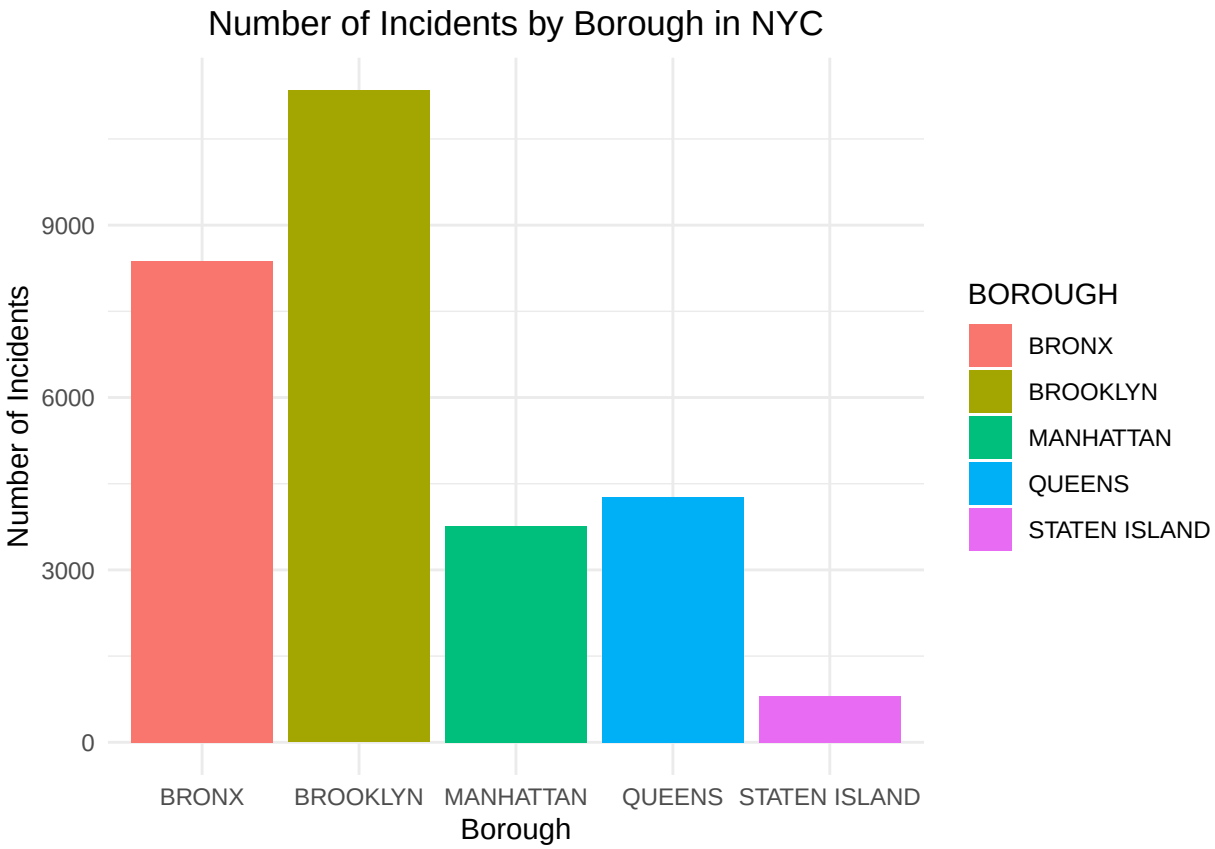
As we can see, above 4 variables have a lot of missing values. It's important to impute these with appropriate values so that, when we do the analysis, it will have proper meaning and easy to interpret.

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

Data Visualizations

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

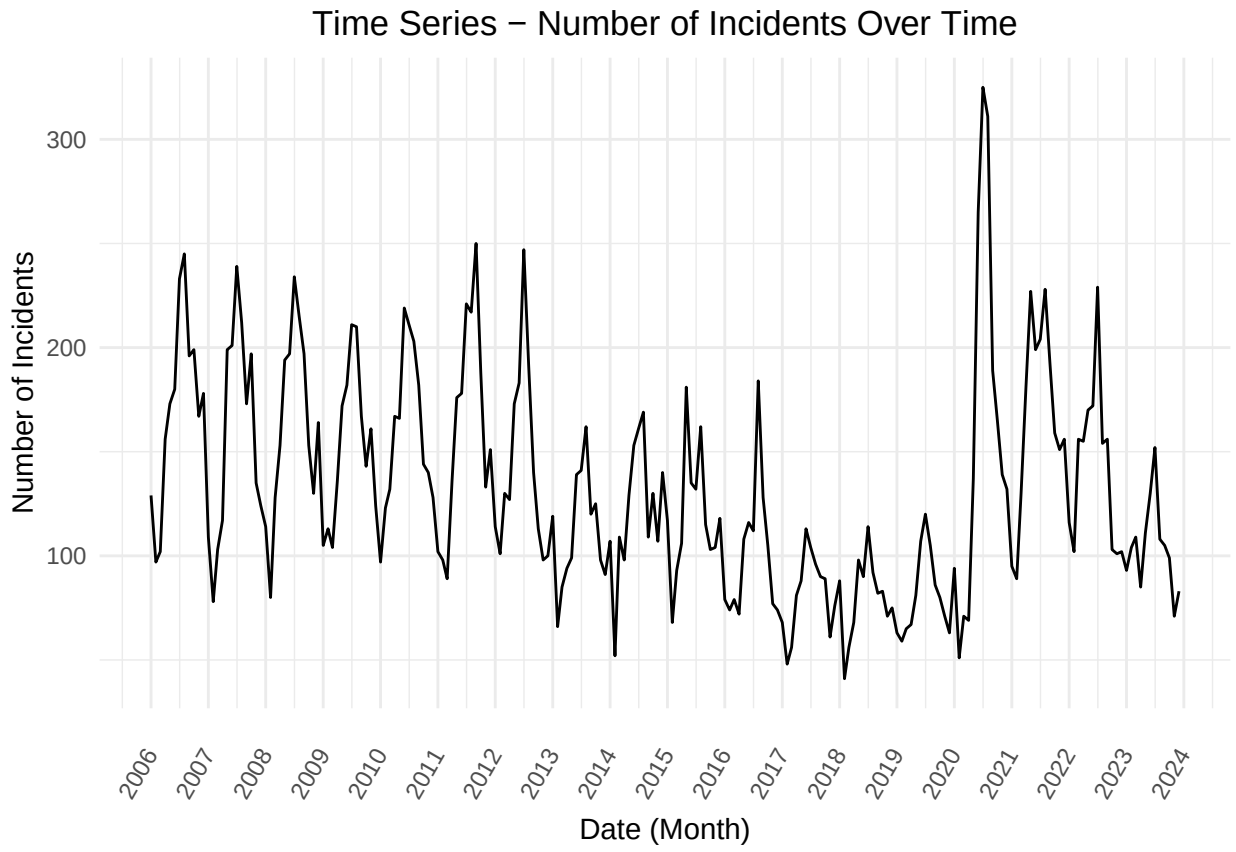
- Shootings by borough
- Time series of shooting incidents
- Victim age by incident



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.


Time Series − Number of Incidents Over Time
Number of Incidents
300
200
100
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Date (Month)