Shooting Project

2023-02-08

Shooting Project

Dataset

The Historic NYPD Shooting Incident Data lists all shooting incidences in NYC from 2006 through the end of the most recent calendar year (2022 at time of publishing).

Every record represents a single shooting incident and includes the following:

- location event occurred
- time event occurred
- suspect information
- victim demographics

Libraries Used

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Data Collection and Inspection

```
url = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
shooting_data <- read_csv(url)
```

```
## 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.
```

```
summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms         Class :character
##  Median : 86437258   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD         Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

```
head(shooting_data)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>       <dbl>             <dbl>
## 1    236168668 11/11/2021 15:04      BROOKLYN       79                 0
## 2    231008085 07/16/2021 22:05      BROOKLYN       72                 0
## 3    230717903 07/11/2021 01:09      BROOKLYN       79                 0
## 4    237712309 12/11/2021 13:42      BROOKLYN       81                 0
## 5    224465521 02/16/2021 20:00      QUEENS        113                 0
## 6    228252164 05/15/2021 04:13      QUEENS        113                 0
## # ... with 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

Initial observations of variables:

- INCIDENT_KEY: numeric value (unique identifier)
- OCCUR_DATE: character value in form MM/DD/YYYY – should be converted to date.
- OCCUR_TIME: numeric in form hh:mm:ss
- BORO: character value
- PRECINCT: numeric value
- JURISDICTION_CODE: numeric value
- LOCATION_DESC: Character value
- STATISTICAL_MURDER_FLAG: logical
- PERP_AGE_GROUP: character. NA values numerous, presumably due to unknown perpetrator.
- PERP_SEX: character – should be a factor. NA values numerous, presumably due to unknown perpetrator.
- PERP_RACE: character. NA values numerous, presumably due to unknown perpetrator.
- VIC_AGE_GROUP: character
- VIC_SEX: character - should be a factor
- VIC_RACE: character
- X_COORD_CD: numeric
- Y_COORD_CD: numeric
- Latitude: numeric
- Longitude: numeric
- Lon_Lat: character

Tidying Data

1. Convert VIC_RACE and VIC_SEX to factors (wait to do the same for PERP_ after addressing NA values)
2. Convert OCCUR_DATE to date format

3. Remove unnecessary columns for our analysis (x & y coordinates, latitude, longitude, precinct, jurisdiction code)

```
shooting_data <- shooting_data %>%
  # Convert victim's sex and race columns to factors, convert date from number to date
  mutate(OCCUR_DATE=mdy(OCCUR_DATE),VIC_SEX = as.factor(VIC_SEX),VIC_RACE = as.factor(VIC_RACE)) %>%
  # Exclude the following columns from final data set
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, PRECINCT, JURISDICTION_CODE))
```

Missing Values

Displaying missing values for all remaining columns:

```
col_na <- colSums(is.na(shooting_data))
data.frame(na_count = col_na)
```

```
##                         na_count
## INCIDENT_KEY                   0
## OCCUR_DATE                     0
## OCCUR_TIME                     0
## BORO                           0
## LOCATION_DESC              14977
## STATISTICAL_MURDER_FLAG        0
## PERP_AGE_GROUP              9344
## PERP_SEX                    9310
## PERP_RACE                   9310
## VIC_AGE_GROUP                  0
## VIC_SEX                        0
## VIC_RACE                       0
```

Analysis It makes sense for similar (large) amounts of NA values to exist for PERP_AGE_GROUP, PERP_SEX, and PERP_RACE – presumably, these were incidents where the perpetrator was unidentified. There are slightly more observations with a missing PERP_AGE_GROUP than for PERP_SEX and PERP_RACE (which both had the same value), presumably because it is easier for a victim to guess at an attacker's sex and race than their age range if they were not apprehended. We could change these to "unknown" so that they may qualify as a separate factor.

There are also a large amount of values missing for LOCATION_DESC, which is more confusing, as this is more likely to be identified. The simplest explanation is that it is simply not considered a vital aspect of an incident report and is more frequently left out. As over 50% of the observations have NA as LOCATION_DESC, we could shift it to "unknown" but in this case I will instead drop the column to remove the risk of skewed data.

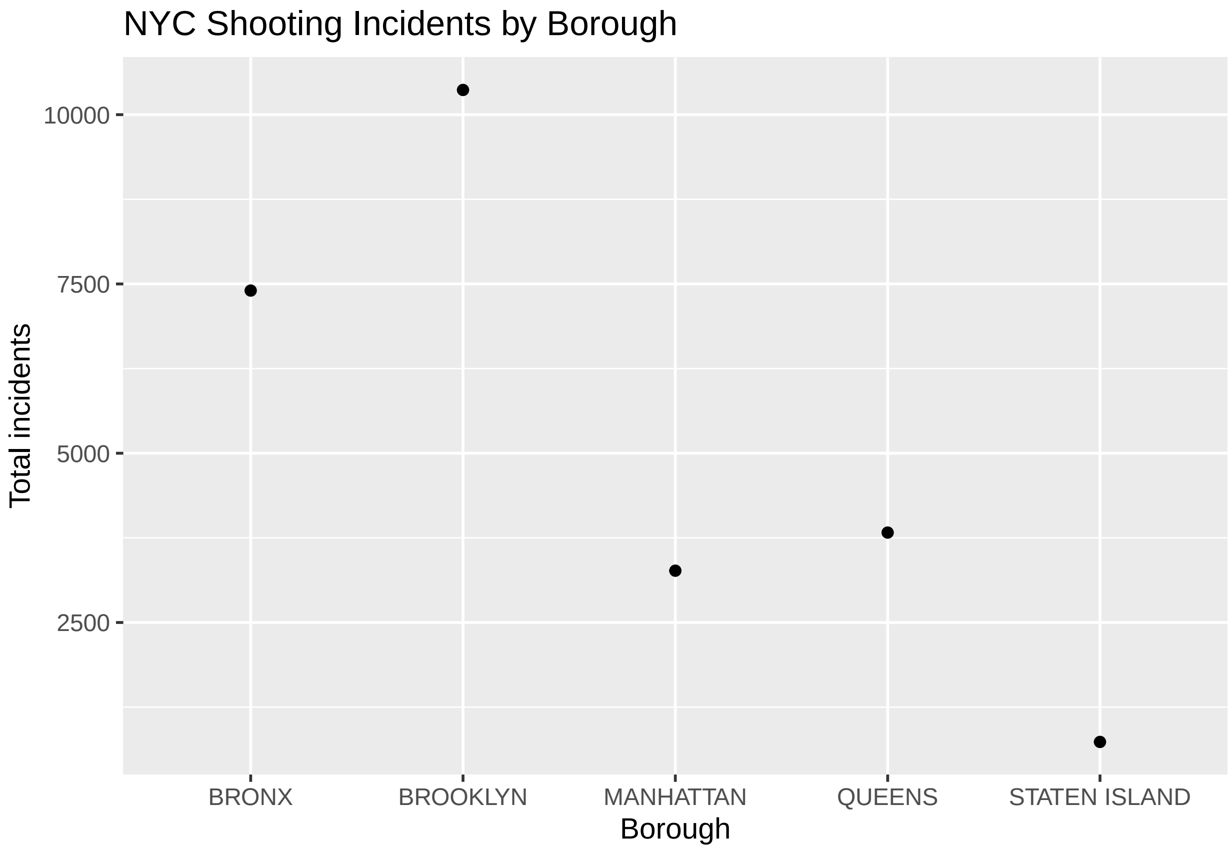
```
shooting_data <- shooting_data %>%
  # Convert any NA values to "unknown" in following three columns
  mutate(PERP_SEX = as.factor(ifelse(is.na(PERP_SEX), "unknown", PERP_SEX)),
         PERP_RACE = as.factor(ifelse(is.na(PERP_RACE), "unknown", PERP_RACE)),
         PERP_AGE_GROUP = ifelse(is.na(PERP_AGE_GROUP), "unknown", PERP_AGE_GROUP)) %>%
  # Exclude LOCATION_DESC column from final data set
  select(-LOCATION_DESC)
```

VISUALIZATION AND ANALYSIS

Total Incidents by Borough

To start, I want to see the relationship between location and the total number of incidents. I will use BORO and COUNT(INCIDENT_KEY) for this.

```r
# Create new data frame with just the borough name and total number of incidents.
numbers_by_boro <- as.data.frame(table(shooting_data$BORO))
names(numbers_by_boro)[c(1,2)] <- c("Boro", "Total")
# Plot the relationship
ggplot(data = numbers_by_boro, aes(Boro, Total)) +
       geom_point() +
       labs(title = "NYC Shooting Incidents by Borough",
            x = "Borough",
            y = "Total incidents")
```



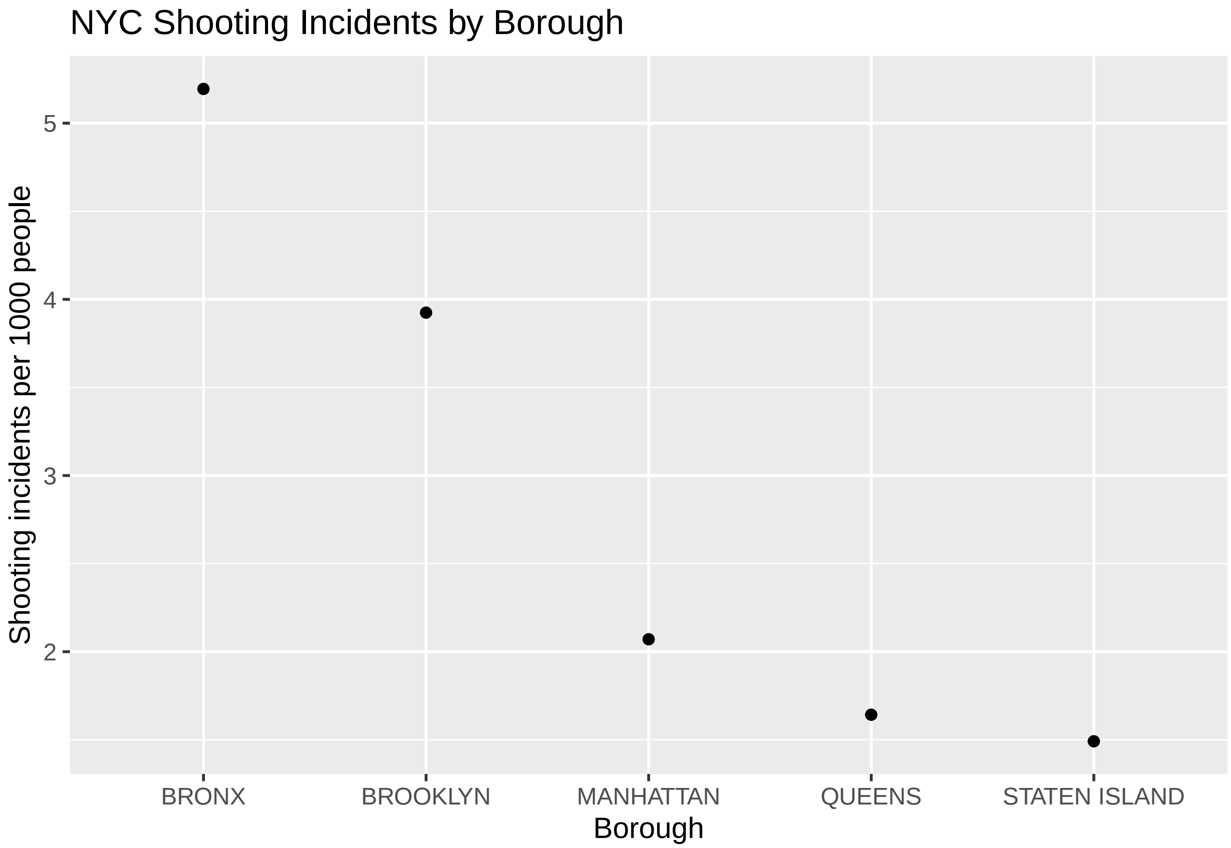
According to this graph, the most incidents occurred in Brooklyn and the Bronx, followed at a distance by Queens, Manhattan, and in the lowest set of incidents, Staten Island.

Are Brooklyn and the Bronx more dangerous? Or are they simply far more populous? The original data did not include population figures for the five boroughs, so I will need to track this down to analyze further.

Incidents by Borough Relative to Population

Per the data available from https://www.citypopulation.de/en/usa/newyorkcity/, as of the 2021 census, the boroughs have the following population:

```
# Add population data from 2021 census
Pop <- c(1424948, 2641052, 1576876, 2331143, 493494)
# Include column for deaths relative to population
numbers_by_boro <- numbers_by_boro %>%
  mutate(Population = Pop, Shootings_Per_Thousand = Total*1000/Population)
# View new graph
ggplot(data = numbers_by_boro, aes(Boro, Shootings_Per_Thousand)) +
       geom_point() +
       labs(title = "NYC Shooting Incidents by Borough",
            x = "Borough",
            y = "Shooting incidents per 1000 people")
```

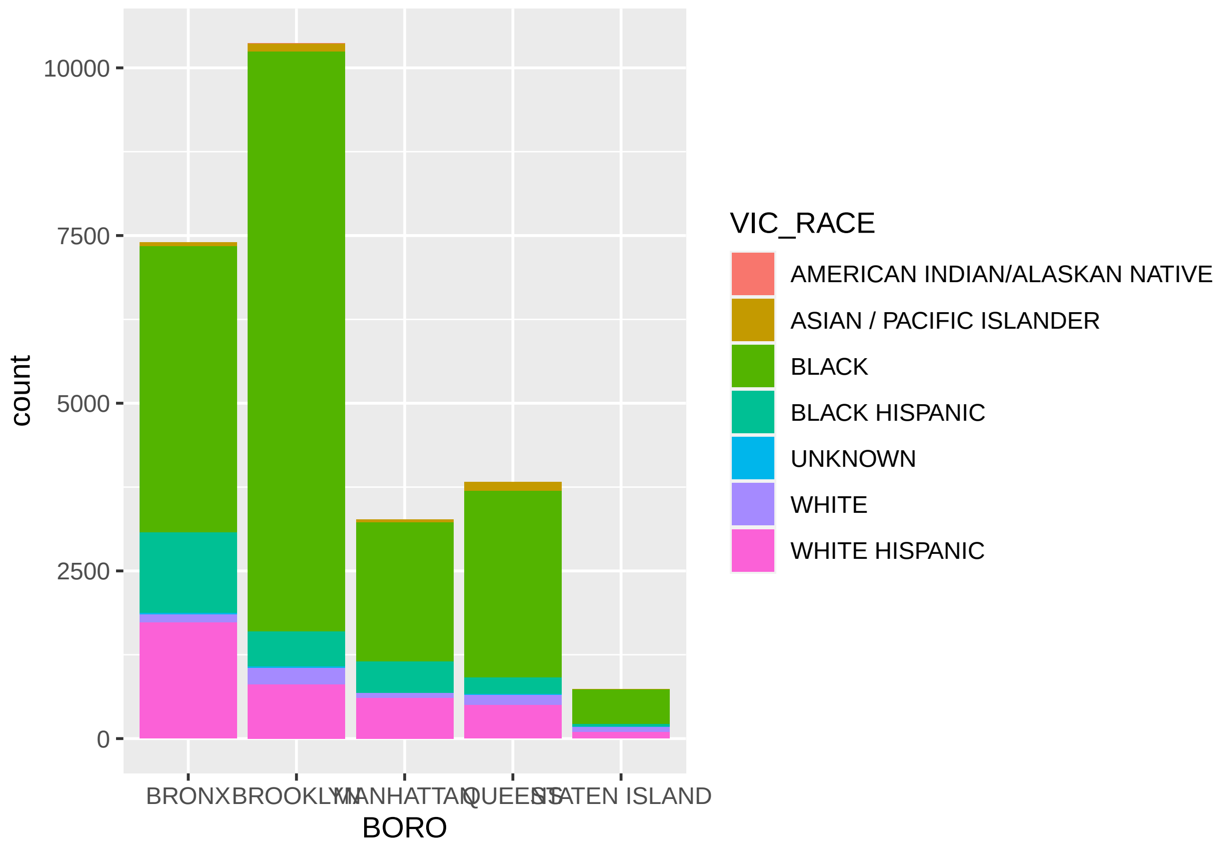


With this new information, we can see that Bronx and Brooklyn do indeed have the most incidents by population but it is the Bronx rather than Brooklyn with the highest per-capita number of incidents. And due to its low population, Staten Island is pretty similar to Queens in terms of per-capita incidents.

Total Incidents Broken Down by Age, Race and Gender

In order to break this data down further, I will use a bar plot to show how incidents break down by race of the victim in the five boroughs.

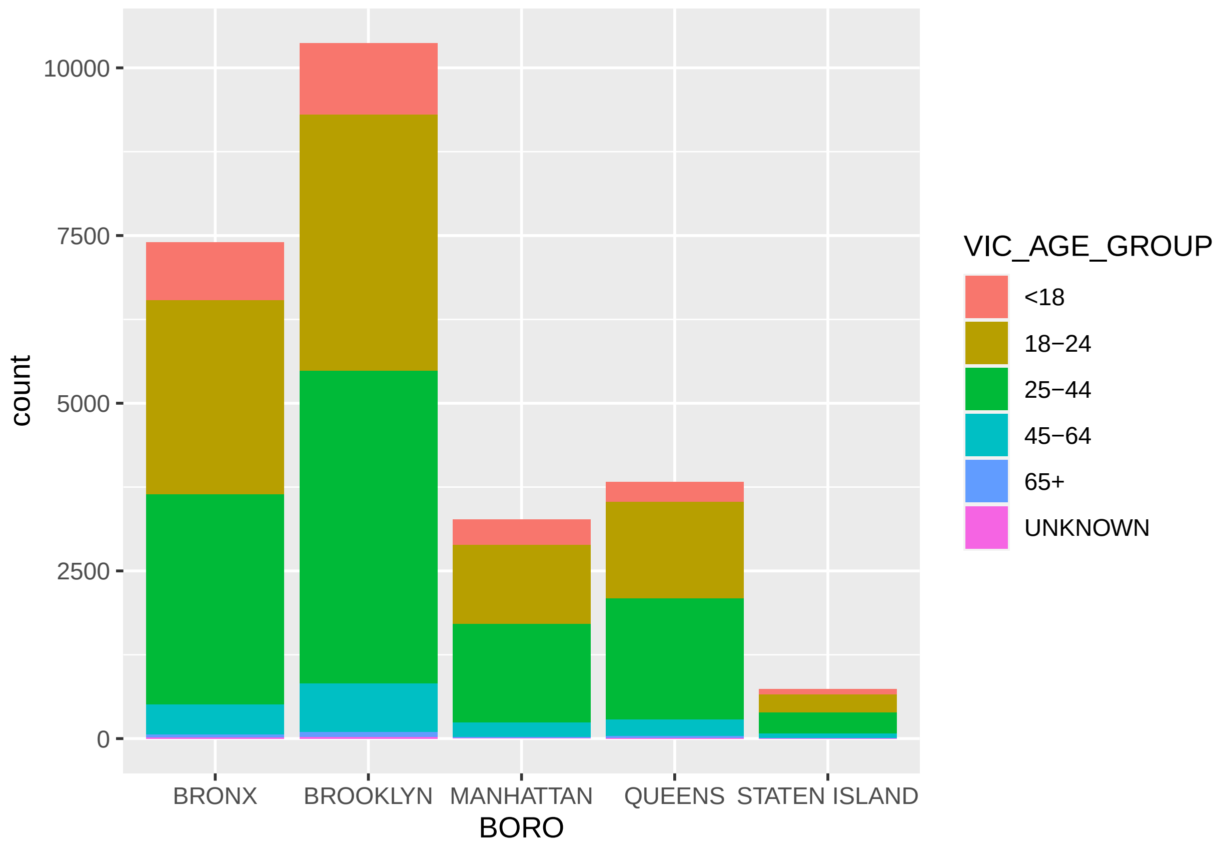
```
ggplot(data=shooting_data, aes(fill=VIC_RACE, x=BORO)) +
  geom_bar()
```



In all five boroughs, the vast majority of victims are Black (far more than the proportion of NYC's overall Black population), but the discrepancy is especially stark in Brooklyn and Queens where over 75% of victims are Black. The next largest groups are White Hispanic and Black Hispanic.

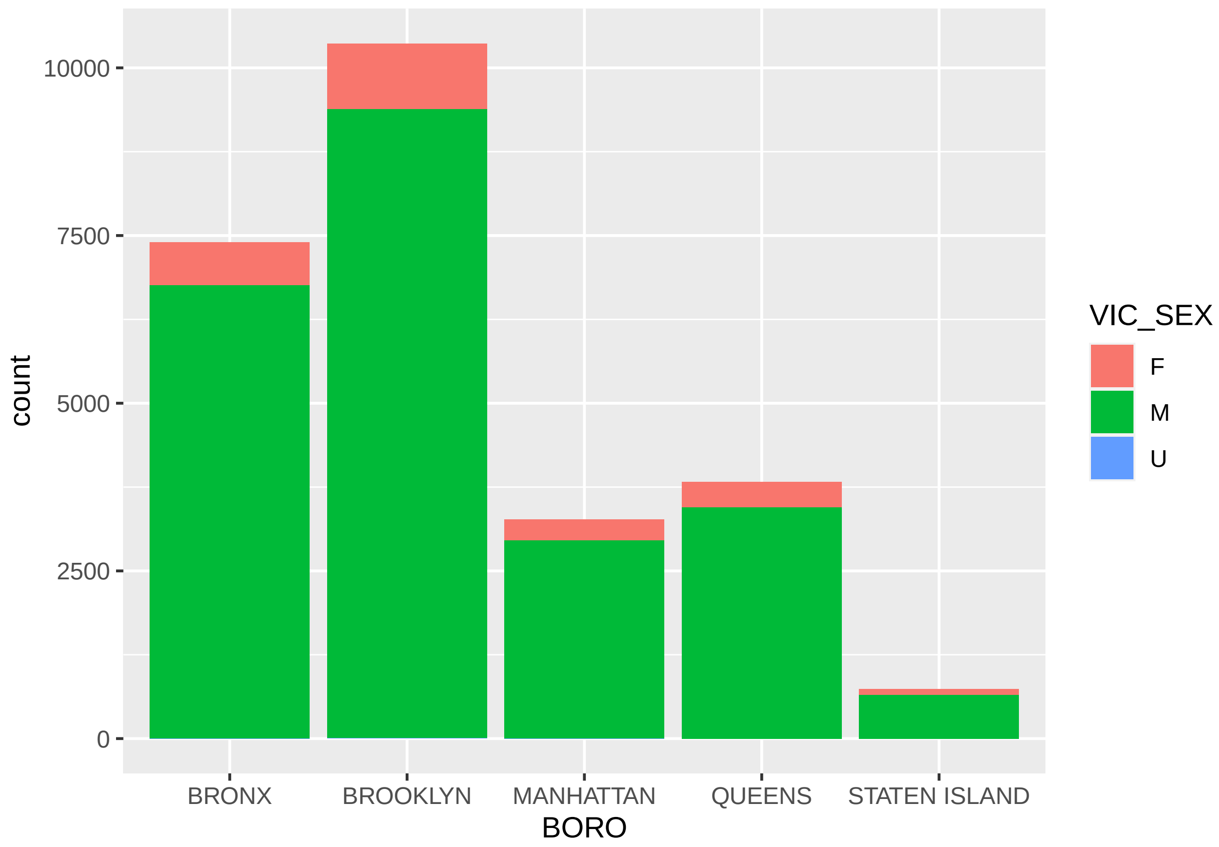
In contrast, when breaking victims down by age, we see more consistent proportions across the boroughs.

```
ggplot(data=shooting_data, aes(fill=VIC_AGE_GROUP, x=BORO)) +
  geom_bar()
```



In all five boroughs, the majority of victims are evenly split between the 18-24 group and the 25-44 group. Finally, when breaking down the data by sex, we find that victims are overwhelmingly male.

```
ggplot(data=shooting_data, aes(fill=VIC_SEX, x=BORO)) +
  geom_bar()
```



CONCLUSION

Per-capita shootings are highest in the Bronx followed by Brooklyn, but across all five boroughs the victims are disproportionately Black and/or Hispanic, young adults, and male. At first glance, this suggests that most of NYC's shootings may be gang-related, however more data should be collected before concluding this, as the data does not include whether or not victims were associated with any gangs. Aside from this data, more useful data points would be education level, income, and employment status of the victims for futher analysis.

FUTURE QUESTIONS

For future examination of this data set, here are the questions I have:

1. How do age, race, sex, education, and poverty play roles?

- Do areas with higher levels of shooting have a high predominance of shootings by (and against) young males in poorer areas who may have less education? If so, it could suggest gang involvement.
- In contrast, in areas where victims tend to be wealthier and/or better educated, it may suggest a motive related to robbery.
- Clustered incidents where victims are disproportionately young (school age) suggests it may be a school shooting, whereas disparate cases of adult female victims may point more towards domestic violence issues.

2. How was the data collected? Why was there a gap in location description?

- As hypothesized earlier, the officers recording data may have considered location description unimportant but it certainly could not have been unknown in as many incidents as the N/A entry suggests. This is because those same incidents in general did NOT lack data on latitude and longitude, so clearly the location of the incident was known upon reporting.

BIAS IDENTIFICATION

Potential sources of bias:

1. Personal - as an American who has spent most of my life in urban areas, I assume that the shooting patterns in NYC mirror those in other cities of similar size, hence the future questions pointing to poverty and gang association as likely contributors that could be assessed if the data included education and wealth, though collecting data on gang affiliation of victims would be useful to confirming or refuting this assumption.
2. Data collection - There is no available information on how the data is collected (or why, for example, location description is missing in so many incidents despite location being known). If the lack of entry is a choice on the part of the reporting officer (rather than lack of knowledge), then the N/A values in race, age, and sex may be biased.
3. Data treatment - I made the choice to eliminate the factor of location description as it was missing from so many entries, but it is possible that a trend may have been present there that this decision now obfuscates.