NYPD_Shooting

Edison

2022-06-20

Introduction

Data is on every shooting that occurred in NYC from 2006 to the end of the previous calender year which should mean end of December 2021.

Every row in the data corresponds to a single shooting incident. Location and demographic data are included.

Load Packages

```
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(pander) # only for the single markdown table, can skip
```

Read Data

Let's load in the data from: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

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

```
## 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.
```

Preparing Data

```
summary(data_raw)
```

```
##   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(data_raw)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>        <dbl>             <dbl>
## 1     24050482 08/27/2006 05:35      BRONX           52                 0
```

```
## 2     77673979 03/11/2011 12:03      QUEENS         106                 0
## 3    226950018 04/14/2021 21:08      BRONX           42                 0
## 4    237710987 12/10/2021 19:30      BRONX           52                 0
## 5    224701998 02/22/2021 00:18      MANHATTAN       34                 0
## 6    225295736 03/07/2021 06:15      BROOKLYN        75                 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>
```

Let’s take a look at the column descriptions:

source: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

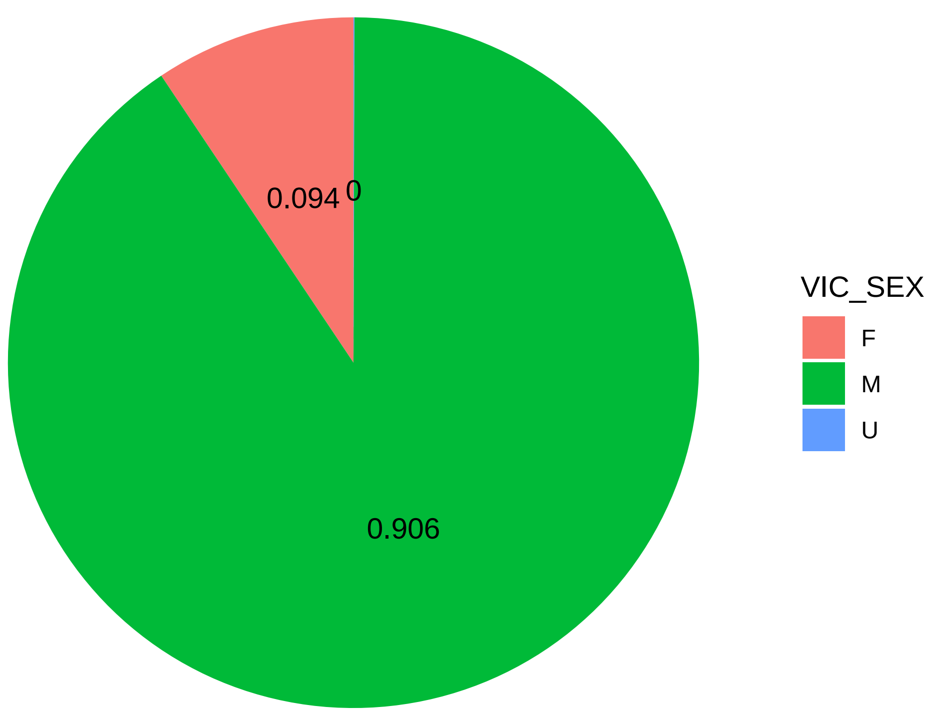
Column Name	Description
INCIDENT_KEY	Randomly generated persistent ID for each arrest
OCCUR_DATE	Exact date of the shooting incident
OCCUR_TIME	Exact time of the shooting incident
BORO	Borough where the shooting incident occurred
PRECINCT	Precinct where the shooting incident occurred
JURISDICTION_CODE	Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions
LOCATION_DESC	Location of the shooting incident
STATISTICAL_MURDER_FLAG	Shooting resulted in the victim’s death which would be counted as a murder
PERP_AGE_GROUP	Perpetrator’s age within a category
PERP_SEX	Perpetrator’s sex description
PERP_RACE	Perpetrator’s race description
VIC_AGE_GROUP	Victim’s age within a category
VIC_SEX	Victim’s sex description
VIC_RACE	Victim’s race description
X_COORD_CD	Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Lon_Lat	Longitude and Latitude Coordinates for mapping

Data Visualization

Let’s see how the genders stacked up in the dataset.

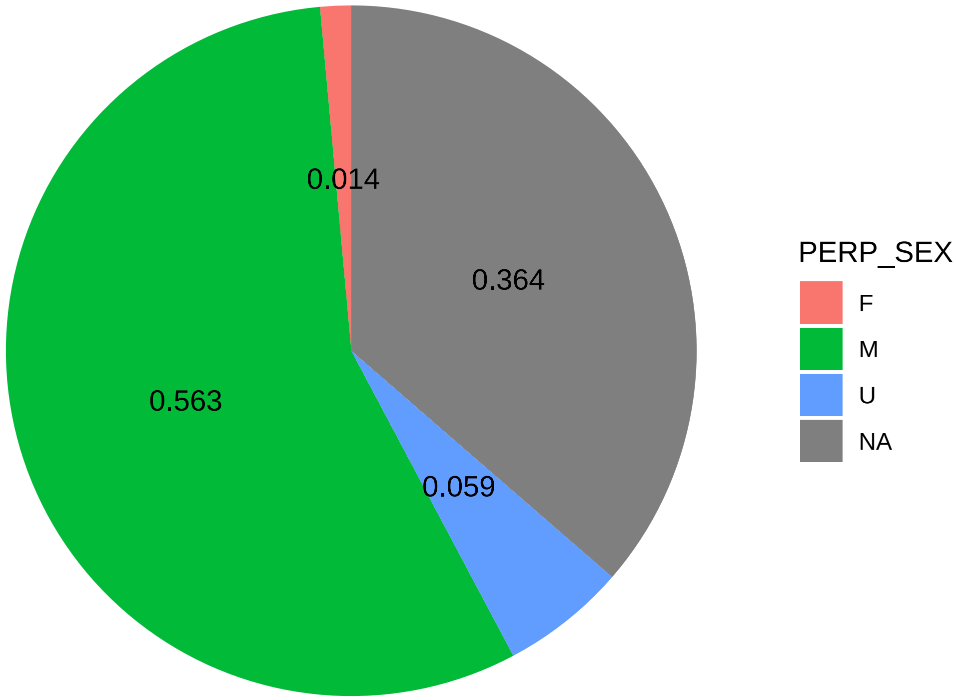
First we’ll take a look at the victims.

```r
vic_sex <- data_raw %>% count(VIC_SEX)
vic_sex$n <- vic_sex$n / nrow(data_raw)
ggplot(vic_sex, aes(x = "", y=n, fill=VIC_SEX)) +
  geom_col() +
  geom_text(aes(label = round(n, digits=3)),
            position = position_stack(vjust = 0.5)) +
  theme_void() +
  coord_polar(theta = "y")
```



Now let's take a look at the perp's sex distribution.

```r
perp_sex <- data_raw %>% count(PERP_SEX)
perp_sex$n <- perp_sex$n / nrow(data_raw)
ggplot(perp_sex, aes(x = "", y=n, fill=PERP_SEX)) +
  geom_col() +
  geom_text(aes(label = round(n, digits=3)),
            position = position_stack(vjust = 0.5)) +
  theme_void() +
  coord_polar(theta = "y")
```



Males make the majority of victims and perps. Of the known data, women are even less likely to be the perp. The perp data does have a lot of NA and UNKNOWN data.

What else could we explore with this data? Well we could see if a popular statistic is also true for this dataset.

I believe the statistic is that murders tend to be largely within race.

Let's trim this down to relevant columns.

```
data <- data_raw[c("PERP_RACE", "VIC_RACE")]
colSums(is.na(data))
```

```
## PERP_RACE  VIC_RACE
##      9310         0
```

The missing data for these these two columns are only in PERP_RACE. It must be when the perp is not caught and is unknown.

Let's remove the rows with NA.

```
print(nrow(data))
```

```
## [1] 25596
```

```r
data_no_na <- na.omit(data)
print(nrow(data_no_na))
```

```
## [1] 16286
```

Over a third of the data is dropped which is quite a lot. However, having missing PERP_RACE data would not be helpful for our purposes.

Let's see how the PERP_RACE values look like:

```r
data_no_na %>% count(PERP_RACE)
```

```
## # A tibble: 7 x 2
##   PERP_RACE                          n
##   <chr>                          <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE     2
## 2 ASIAN / PACIFIC ISLANDER         141
## 3 BLACK                          10668
## 4 BLACK HISPANIC                  1203
## 5 UNKNOWN                         1836
## 6 WHITE                            272
## 7 WHITE HISPANIC                  2164
```

Now let's take a look at the corresponding VIC_RACE values:

```r
data_no_na %>% count(VIC_RACE)
```

```
## # A tibble: 7 x 2
##   VIC_RACE                           n
##   <chr>                          <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE     7
## 2 ASIAN / PACIFIC ISLANDER         257
## 3 BLACK                          11117
## 4 BLACK HISPANIC                  1641
## 5 UNKNOWN                           47
## 6 WHITE                            515
## 7 WHITE HISPANIC                  2702
```

Seems like there is a lot of similarity across the two. The UNKNOWN labels seem like they would be problematic as the number for UNKNOWN is a lot higher than the UNKNOWN for the VIC_RACE.

I will drop any row that contains UNKNOWN as well.

```r
data_complete <- data_no_na[rowSums(data_no_na == "UNKNOWN")==0, , drop = FALSE]
data_complete %>% count(VIC_RACE)
```

```
## # A tibble: 6 x 2
##   VIC_RACE                           n
##   <chr>                          <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE     4
## 2 ASIAN / PACIFIC ISLANDER         241
```

```
## 3 BLACK                          9758
## 4 BLACK HISPANIC                 1486
## 5 WHITE                           473
## 6 WHITE HISPANIC                 2447
```

```
data_complete %>% count(PERP_RACE)
```

```
## # A tibble: 6 x 2
##   PERP_RACE                          n
##   <chr>                          <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE     2
## 2 ASIAN / PACIFIC ISLANDER         141
## 3 BLACK                          10644
## 4 BLACK HISPANIC                  1198
## 5 WHITE                            271
## 6 WHITE HISPANIC                  2153
```

Data distribution in PERP and VIC race seems pretty similar.

Data went from 25596 rows to 14409 rows so we lost about half our data which is a lot.

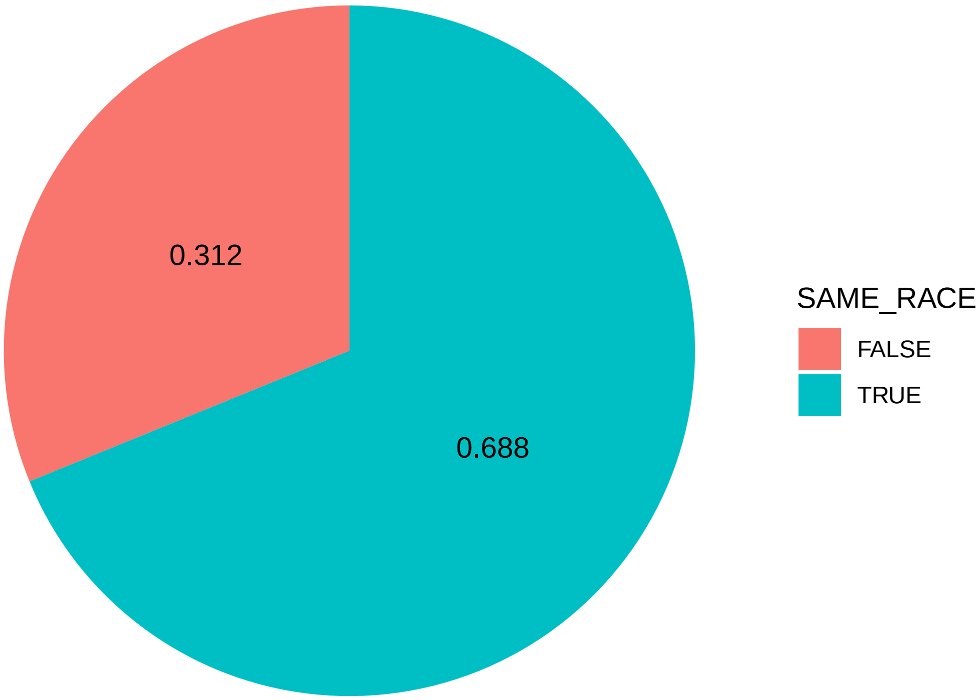
Let's make a new column that prints 1 if both VIC_RACE and PERP_RACE is exactly the same.

```
data_complete$SAME_RACE <- as.factor(data_complete$VIC_RACE == data_complete$PERP_RACE)
frequency_table <- data_complete %>% count(SAME_RACE)
frequency_table$n <- frequency_table$n / nrow(data_complete)
frequency_table
```

```
## # A tibble: 2 x 2
##   SAME_RACE     n
##   <fct>     <dbl>
## 1 FALSE     0.312
## 2 TRUE      0.688
```

Let's visualize this really quick

```
ggplot(frequency_table, aes(x = "", y=n, fill=SAME_RACE)) +
  geom_col() +
  geom_text(aes(label = round(n, digits=3)),
            position = position_stack(vjust = 0.5)) +
  theme_void() +
  coord_polar(theta = "y")
```



While not high as I expected, in about 69% of all shootings, where both the victim and perp races are known, they are of the same race label.

To get some insight on why it might not be higher, let's take a look at the counts between PERP_VIC of each race type.

```
data_complete$PERP_VIC <- paste(data_complete$PERP_RACE, "-", data_complete$VIC_RACE)
data_complete %>% count(PERP_VIC)
```

```
## # A tibble: 27 x 2
##    PERP_VIC                                                 n
##    <chr>                                                <int>
##  1 AMERICAN INDIAN/ALASKAN NATIVE - BLACK                   2
##  2 ASIAN / PACIFIC ISLANDER - ASIAN / PACIFIC ISLANDER     43
##  3 ASIAN / PACIFIC ISLANDER - BLACK                        51
##  4 ASIAN / PACIFIC ISLANDER - BLACK HISPANIC               13
##  5 ASIAN / PACIFIC ISLANDER - WHITE                        11
##  6 ASIAN / PACIFIC ISLANDER - WHITE HISPANIC               23
##  7 BLACK - AMERICAN INDIAN/ALASKAN NATIVE                   4
##  8 BLACK - ASIAN / PACIFIC ISLANDER                       135
##  9 BLACK - BLACK                                         8471
## 10 BLACK - BLACK HISPANIC                                 749
## # ... with 17 more rows
```

As we can see BLACK HISPANIC to BLACK wouldn't count as same race using our current matching method, so if we wanted to include those, we would need a function more complex. The TRUE rate would be higher if we included these cases.

Let's leave it as is, but let's be aware of this issue.

Model

Let's see if any PERP race can be used to predict the created SAME_RACE factor better than others.

```r
model <- glm(SAME_RACE ~ PERP_RACE, data = data_complete, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = SAME_RACE ~ PERP_RACE, family = "binomial", data = data_complete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7826  -0.8530   0.6758   0.6758   1.6249
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -11.57     139.28  -0.083    0.934
## PERP_RACEASIAN / PACIFIC ISLANDER    10.74     139.28   0.077    0.939
## PERP_RACEBLACK                       12.93     139.28   0.093    0.926
## PERP_RACEBLACK HISPANIC              10.56     139.28   0.076    0.940
## PERP_RACEWHITE                       11.87     139.28   0.085    0.932
## PERP_RACEWHITE HISPANIC              11.29     139.28   0.081    0.935
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17881  on 14408  degrees of freedom
## Residual deviance: 15650  on 14403  degrees of freedom
## AIC: 15662
##
## Number of Fisher Scoring iterations: 10
```

As we can see, none of the variables are significant. It seems like individual PERP race is not a good indicator of whether the VIC_RACE will be exactly the same.

Since percentage of SAME_RACE shooting is about 69% and likely over if we include race subcategory Hispanic, it would likely be that any model will predict that SAME_RACE is true regardless of PERP_RACE category.

Conclusion

In this report, we examined our NYPD dataset to see if it fit a popular FBI statistic that murders largely tend to have victims of the same race. We saw that the NYPD does seem to match this statistic where nearly 69% of the shooting cases where both perpetrator and victim races are known, both are of the same race. There are some conflicts with this category with how Hispanic sub categories can be counted, but that will be left for future analysis. We took this SAME_RACE variable and regressed it on PERP_RACE and found that there wasn't enough evidence to conclude that any race is more likely than others to shoot someone who is of the race.

Is there any possible bias in the data source? Well it is reported by the police department and since the perpetrator can sometimes be missing or unknown, it is difficult to say whether a source of bias can be induced. I am not sure how the perpetrator race factors are determined either and could be based off eyewitness testimony. Those could be unreliable as well as people don't have the best of memory and can be biased as well. Victim race seems a lot less likely to be unbiased since the data would be more accurate due to being dead or at the incident report themselves. Further analysis can be done to see if there are any good indicators to predict missing race or age group categories based on location, borough, and victim demographics.