NYPD Shooting Incident Data Analysis

C. Cozad

2024-05-20

The NYPD shooting incident dataset lists every shooting incident that occured in New York City from 2006 to the end of the previous quarter (in this case, that is Q1 2024). Information is included about each event, such as suspect description, time, and place.

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

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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: Let's take a look at an overview of the data.

```
summary(data)
```

```
##   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
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
```

```
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Change appropriate variables to factor: The following variables are categorical, so we convert them to factor.

```
data$BORO <- factor(data$BORO)
data$LOC_OF_OCCUR_DESC <- factor(data$LOC_OF_OCCUR_DESC)
data$PRECINCT <- factor(data$PRECINCT)
data$JURISDICTION_CODE <- factor(data$JURISDICTION_CODE)
data$LOC_CLASSFCTN_DESC <- factor(data$LOC_CLASSFCTN_DESC)
data$LOCATION_DESC <- factor(data$LOCATION_DESC)
data$PERP_AGE_GROUP <- factor(data$PERP_AGE_GROUP)
data$PERP_SEX <- factor(data$PERP_SEX)
data$PERP_RACE <- factor(data$PERP_RACE)
data$VIC_AGE_GROUP <- factor(data$VIC_AGE_GROUP)
data$VIC_SEX <- factor(data$VIC_SEX)
data$VIC_RACE <- factor(data$VIC_RACE)
```

Change appropriate variables to date type: We can change `OCCUR_DATE` to a date type and `OCCUR_TIME` to a time type.

```
data$OCCUR_DATE <- as.Date(data$OCCUR_DATE, format = "%m/%d/%Y")
```

Drop unnecessary columns: We can drop `X_COORD_CD` and `Y_COORD_CD`, since these are the same as latitude and longitude, just in a different map projection and different units. We can also drop `Lon_Lat`, since it is just the latitude and longitude in a different format.

```
data <- subset(data, select = -c(X_COORD_CD, Y_COORD_CD, Lon_Lat))
```

The rest of the columns provide potentially useful information for our analysis.

Handling missing data: This dataset has quite a bit of missing data. Some variables have so few data points, it's best to drop them from the dataset entirely, since they likely won't be very helpful in an analysis.

Here is the percentage of values missing in each column:

```
missing_percentage <- colMeans(is.na(data)) * 100
missing_percentage
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##              0.000000000              0.000000000              0.000000000
##                    BORO        LOC_OF_OCCUR_DESC                 PRECINCT
##              0.000000000             89.615573139              0.000000000
##       JURISDICTION_CODE       LOC_CLASSFCTN_DESC            LOCATION_DESC
##              0.007002311             89.615573139             52.436804145
## STATISTICAL_MURDER_FLAG           PERP_AGE_GROUP                 PERP_SEX
##              0.000000000             32.714795883             32.595756600
##                PERP_RACE           VIC_AGE_GROUP                  VIC_SEX
##             32.595756600              0.000000000              0.000000000
##                 VIC_RACE                 Latitude                Longitude
##              0.000000000              0.206568167              0.206568167
```

We're going to drop any column with over half it's values missing. We'll also keep in mind that all three of the `PERP` columns have a significant number of missing values, and we might avoid them in our analysis.

```
columns_to_drop <- names(missing_percentage[missing_percentage > 50])
data <- data[, !(names(data) %in% columns_to_drop)]
```

Here's another look at the dataset before we move on.

```
summary(data)
```

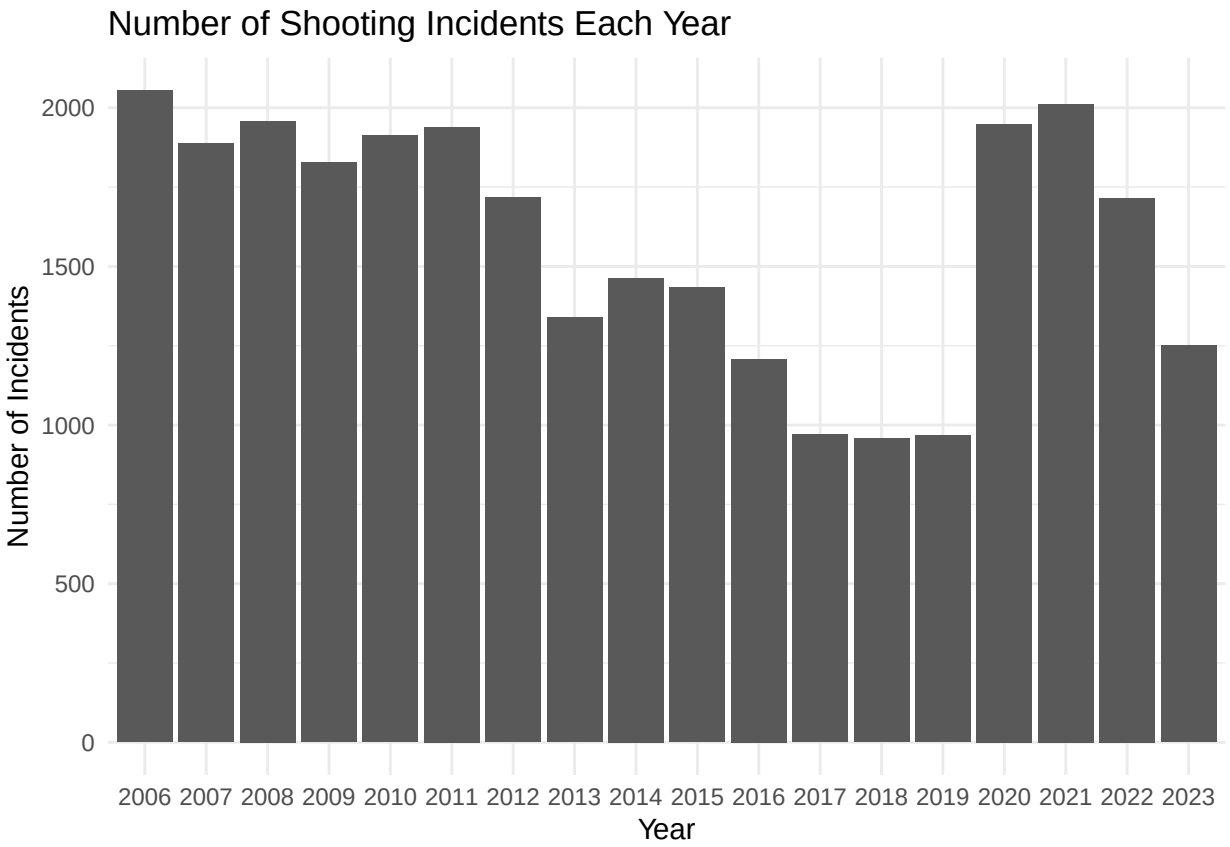
```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms
##  Median : 92711254   Median :2013-09-20   Class2:difftime
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##             BORO          PRECINCT     JURISDICTION_CODE
##  BRONX        : 8376   75     : 1628   0   :23923
##  BROOKLYN     :11346   73     : 1500   1   :   81
##  MANHATTAN    : 3762   67     : 1259   2   : 4556
##  QUEENS       : 4271   44     : 1076   NA's:    2
##  STATEN ISLAND:  807   79     : 1045
##                        47     : 1006
##                        (Other):21048
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX              PERP_RACE
```

```
##  Mode :logical           18-24  :6438   (null): 1141   BLACK         :11903
##  FALSE:23036             25-44  :6041   F     :  444   WHITE HISPANIC: 2510
##  TRUE :5526              UNKNOWN:3148   M     :16168   UNKNOWN       : 1837
##                          <18    :1682   U     : 1499   BLACK HISPANIC: 1392
##                          (null) :1141   NA's  : 9310   (null)        : 1141
##                          (Other): 768                  (Other)       :  469
##                          NA's   :9344                  NA's          : 9310
##  VIC_AGE_GROUP   VIC_SEX                          VIC_RACE
##  <18    : 2954   F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  1022   :    1   M:25790   ASIAN / PACIFIC ISLANDER      :  440
##  18-24  :10384   U:   12   BLACK                         :20235
##  25-44  :12973             BLACK HISPANIC                : 2795
##  45-64  : 1981             UNKNOWN                       :   70
##  65+    :  205             WHITE                         :  728
##  UNKNOWN:   64             WHITE HISPANIC                : 4283
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :59      NA's   :59
```

Visualization #1: Let's take a look at a bar chart showing the number of shooting incidents per year. We can see that shooting incidents were generally declining each year, until spiking back up in 2020, likely related to the COVID-19 pandemic.

Additional questions that this visualization prompts include: - Does the decline in shooting incidents after 2020 correlate with the decline in new COVID-19 cases? - Why did the number of shooting incidents stop decreasing and instead plateau from 2017 to 2019?

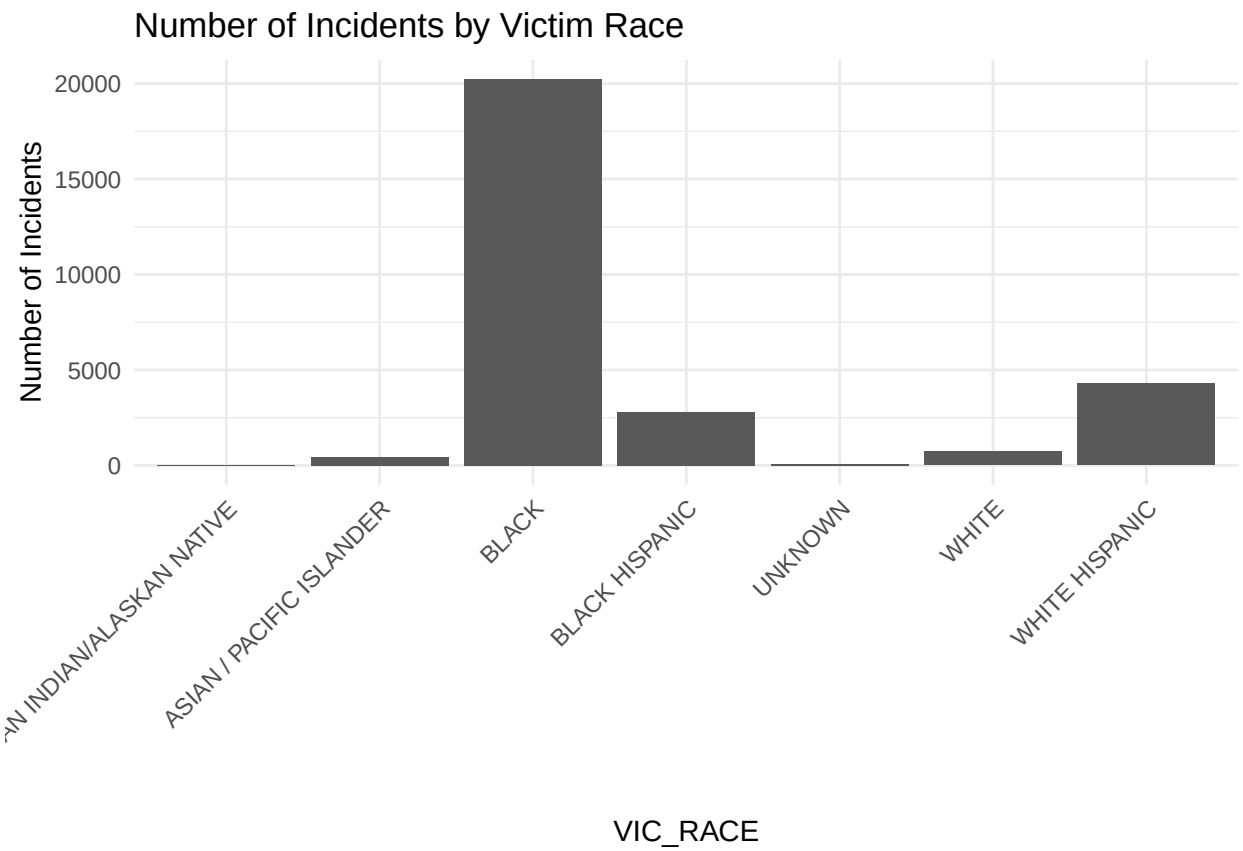
```r
data$year <- year(data$OCCUR_DATE)
incident_counts <- table(data$year)
incident_counts_df <- as.data.frame(incident_counts)
names(incident_counts_df) <- c("year", "incident_count")

ggplot(incident_counts_df, aes(x = year, y = incident_count)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Shooting Incidents Each Year",
       x = "Year",
       y = "Number of Incidents") +
  theme_minimal()
```



Visualization #2: Let's take a look at a bar chart showing the number of shooting incidents grouped by the victim's race. The largest number of shooting victims in New York City are Black, likely since this group tends to face disadvantages that make them more likely to be shooting victims.

Additional questions that this visualization prompts include: - Has the proportion of shooting incidents by race changed over the years? - What are the number of shooting incidents *per capita* by race?

```
ggplot(data, aes(x = VIC_RACE)) +
  geom_bar() +
  labs(title = "Number of Incidents by Victim Race",
       x = "VIC_RACE",
       y = "Number of Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Model: Let's use a logistic regression model to predict which shooting incidents are fatal.

```
logit_model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + JURISDICTION_CODE + PERP_AGE_GROUP + PERP_SEX + PERP

summary(logit_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + JURISDICTION_CODE +
##     PERP_AGE_GROUP + PERP_SEX + PERP_RACE + VIC_AGE_GROUP, family = binomial,
##     data = data)
##
## Coefficients: (2 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -1.97863    0.10881 -18.184  < 2e-16
## BOROBROOKLYN                          -0.10492    0.04657  -2.253 0.024254
## BOROMANHATTAN                         -0.14845    0.05984  -2.481 0.013103
## BOROQUEENS                            -0.14389    0.05896  -2.440 0.014675
## BOROSTATEN ISLAND                     -0.16015    0.10290  -1.556 0.119641
## JURISDICTION_CODE1                    -0.02843    0.29174  -0.097 0.922377
## JURISDICTION_CODE2                    -0.16362    0.05447  -3.004 0.002667
## PERP_AGE_GROUP<18                      2.17027    0.28718   7.557 4.12e-14
## PERP_AGE_GROUP1020                    -8.96257  324.74382  -0.028 0.977982
## PERP_AGE_GROUP1028                    -8.96604  324.74383  -0.028 0.977974
## PERP_AGE_GROUP18-24                    2.30557    0.28104   8.204 2.33e-16
## PERP_AGE_GROUP224                     -9.10567  324.74382  -0.028 0.977631
```

```
## PERP_AGE_GROUP25-44                        2.55756    0.28126   9.093  < 2e-16
## PERP_AGE_GROUP45-64                        2.88285    0.29116   9.901  < 2e-16
## PERP_AGE_GROUP65+                          2.84478    0.38453   7.398 1.38e-13
## PERP_AGE_GROUP940                         -9.12835  324.74381  -0.028 0.977575
## PERP_AGE_GROUPUNKNOWN                     -0.28284    0.25945  -1.090 0.275653
## PERP_SEXF                                 -1.57757    0.28487  -5.538 3.06e-08
## PERP_SEXM                                 -1.73450    0.26257  -6.606 3.95e-11
## PERP_SEXU                                       NA         NA      NA       NA
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -11.68325  229.60045  -0.051 0.959417
## PERP_RACEASIAN / PACIFIC ISLANDER          0.26489    0.17820   1.487 0.137145
## PERP_RACEBLACK                            -0.10708    0.05355  -2.000 0.045533
## PERP_RACEBLACK HISPANIC                   -0.23107    0.08275  -2.792 0.005231
## PERP_RACEUNKNOWN                          -0.77573    0.22543  -3.441 0.000579
## PERP_RACEWHITE                             0.39370    0.13452   2.927 0.003425
## PERP_RACEWHITE HISPANIC                         NA         NA      NA       NA
## VIC_AGE_GROUP1022                        -10.90296  324.74371  -0.034 0.973217
## VIC_AGE_GROUP18-24                         0.25274    0.07237   3.492 0.000479
## VIC_AGE_GROUP25-44                         0.38034    0.07160   5.312 1.08e-07
## VIC_AGE_GROUP45-64                         0.40334    0.09313   4.331 1.48e-05
## VIC_AGE_GROUP65+                           0.80987    0.18804   4.307 1.66e-05
## VIC_AGE_GROUPUNKNOWN                       0.04233    0.33327   0.127 0.898940
##
## (Intercept)                              ***
## BOROBROOKLYN                             *
## BOROMANHATTAN                            *
## BOROQUEENS                               *
## BOROSTATEN ISLAND
## JURISDICTION_CODE1
## JURISDICTION_CODE2                       **
## PERP_AGE_GROUP<18                        ***
## PERP_AGE_GROUP1020
## PERP_AGE_GROUP1028
## PERP_AGE_GROUP18-24                      ***
## PERP_AGE_GROUP224
## PERP_AGE_GROUP25-44                      ***
## PERP_AGE_GROUP45-64                      ***
## PERP_AGE_GROUP65+                        ***
## PERP_AGE_GROUP940
## PERP_AGE_GROUPUNKNOWN
## PERP_SEXF                                ***
## PERP_SEXM                                ***
## PERP_SEXU
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER
## PERP_RACEBLACK                           *
## PERP_RACEBLACK HISPANIC                  **
## PERP_RACEUNKNOWN                         ***
## PERP_RACEWHITE                           **
## PERP_RACEWHITE HISPANIC
## VIC_AGE_GROUP1022
## VIC_AGE_GROUP18-24                       ***
## VIC_AGE_GROUP25-44                       ***
## VIC_AGE_GROUP45-64                       ***
## VIC_AGE_GROUP65+                         ***
```

```
## VIC_AGE_GROUPUNKNOWN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19167  on 19215  degrees of freedom
## Residual deviance: 17967  on 19185  degrees of freedom
##   (9346 observations deleted due to missingness)
## AIC: 18029
##
## Number of Fisher Scoring iterations: 11
```

We can do a quick analysis of our model's performance, showing that it can predict whether a shooting is fatal about 80% of the time, based on the variables we gave it to train on.

```
predicted <- predict(logit_model, type = "response")
predicted_classes <- ifelse(predicted > 0.5, 1, 0)
accuracy <- mean(predicted_classes == data$STATISTICAL_MURDER_FLAG)
```

```
## Warning in predicted_classes == data$STATISTICAL_MURDER_FLAG: longer object
## length is not a multiple of shorter object length
```

```
accuracy
```

```
## [1] 0.8054758
```

Conclusion, recognition of bias, and bias mitigation

There's a couple sources of bias to be aware of in this analysis:

- I used to live near New York City, and have lived in urban areas for the past several years. I certainly have opinions on which variables might be more correlated to shooting incidents, based on my personal experiences. I've attempted to mitigate that by examining each variable thoroughly, rather than cherry picking the ones my intuition thinks are important.
- The number of missing variables in the dataset is a cause for concern, especially related to the way the data was collected. Why do some variables have 80% of their values missing? Are there legal reasons the some data has to be redacted, or are there problems with the reliability of the data collection process used? I dropped variables that were missing an overwhelming number of values in an attempt to mitigate this.

To conclude, this NYC shooting incident dataset is full of valuable information about crime and safety in New York City. We learned that the number of shooting incidents can vary widely from year to year, and that there are patterns in the data with respect to race. We can also create a predictive model to make an educated guess about which incidents are fatal. This was just a brief look into the data, which could certainly be built upon in the future.