NYPD Shooting Dataset Analysis

Brian Behe

April 22, 2023

Introduction

This project performs an exploratory data analysis on the NYPD Shooting Incident Dataset (Historic), generates visualizations of trends in the data, and builds a logistic regression model to predict the fatality of a shooting incident based on features in the data (see below). The data set contains information about shooting incidents in New York City and can be found here.

Import and Initial Exploration

```r
# Load required libraries
install.packages("tidyr")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("lubridate")
install.packages("caret")
library(tidyr)
library(dplyr)
library(ggplot2)
library(lubridate)
# Load the dataset directly from the URL
data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")

# Display the first few rows
head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                        105
## 2    137471050 06/27/2014   17:40:00    BRONX                         40
## 3    147998800 11/21/2015   03:56:00   QUEENS                        108
## 4    146837977 10/09/2015   18:30:00    BRONX                         44
## 5     58921844 02/19/2009   22:58:00    BRONX                         47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                         81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                    false
## 2                 0                                                    false
## 3                 0                                                     true
## 4                 0                                                    false
## 5                 0                                                     true
## 6                 0                                                     true
```

```
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK        45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                         Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

```r
#Generate Summary Statistics of the dataset
summary(data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Description of Features

- INCIDENT_KEY: A unique identifier for each shooting incident.
- OCCUR_DATE: The date of the incident. This column is stored as a character and will need to be converted to a Date type for further analysis.
- OCCUR_TIME: The time of the incident. This column is stored as a character and will need to be converted to a proper time format for further analysis.
- BORO: The borough where the incident occurred. This column is stored as a character and will need to be converted to a factor for further analysis.
- PRECINCT: The police precinct where the incident occurred.
- JURISDICTION_CODE: A code indicating the jurisdiction where the incident occurred. There are missing values in this column.
- LOCATION_DESC: A description of the location where the incident occurred. This column is stored as a character.
- STATISTICAL_MURDER_FLAG: A flag indicating whether the incident was considered a statistical murder. This column is stored as a character.
- PERP_AGE_GROUP: The age group of the perpetrator. This column is stored as a character and will need to be converted to a factor for further analysis.
- PERP_SEX: The sex of the perpetrator. This column is stored as a character and will need to be converted to a factor for further analysis.
- PERP_RACE: The race of the perpetrator. This column is stored as a character and will need to be converted to a factor for further analysis.
- VIC_AGE_GROUP: The age group of the victim. This column is stored as a character and will need to be converted to a factor for further analysis.
- VIC_SEX: The sex of the victim. This column is stored as a character and will need to be converted to a factor for further analysis.
- VIC_RACE: The race of the victim. This column is stored as a character and will need to be converted to a factor for further analysis.
- X_COORD_CD: The X-coordinate of the incident location in the New York-Long Island State Plane Coordinate System.
- Y_COORD_CD: The Y-coordinate of the incident location in the New York-Long Island State Plane Coordinate System.
- Latitude: The latitude of the incident location.

- Longitude: The longitude of the incident location.
- Lon_Lat: A combination of the longitude and latitude values. This column is stored as a character.

```
# Convert date columns to appropriate format
data$OCCUR_DATE <- mdy(data$OCCUR_DATE)
data$OCCUR_YEAR <- year(data$OCCUR_DATE)
data$OCCUR_MONTH <- month(data$OCCUR_DATE)

# Change appropriate variables to factors
data$PRECINCT <- as.factor(data$PRECINCT)
data$JURISDICTION_CODE <- as.factor(data$JURISDICTION_CODE)
data$BORO <- as.factor(data$BORO)
data$VIC_SEX <- as.factor(data$VIC_SEX)
data$VIC_RACE <- as.factor(data$VIC_RACE)
data$PERP_SEX <- as.factor(data$PERP_SEX)
data$PERP_RACE <- as.factor(data$PERP_RACE)

# Remove unnecessary columns:  not using geo location data for this analysis
data$INCIDENT_KEY <- NULL
data$X_COORD_CD <- NULL
data$Y_COORD_CD <- NULL
data$Longitude <- NULL
data$Latitude <- NULL
data$Lon_Lat <- NULL
data$JURISDICTION_CODE <- NULL

head(data)
```

```
##   OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT LOC_CLASSFCTN_DESC
## 1 2021-05-27   21:30:00   QUEENS                        105
## 2 2014-06-27   17:40:00    BRONX                         40
## 3 2015-11-21   03:56:00   QUEENS                        108
## 4 2015-10-09   18:30:00    BRONX                         44
## 5 2009-02-19   22:58:00    BRONX                         47
## 6 2020-10-21   21:36:00 BROOKLYN                         81
##   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE
## 1                                 false
## 2                                 false
## 3                                  true
## 4                                 false
## 5                                  true          25-44        M     BLACK
## 6                                  true
##   VIC_AGE_GROUP VIC_SEX       VIC_RACE OCCUR_YEAR OCCUR_MONTH
## 1         18-24       M          BLACK       2021           5
## 2         18-24       M          BLACK       2014           6
## 3         25-44       M          WHITE       2015          11
## 4           <18       M WHITE HISPANIC       2015          10
## 5         45-64       M          BLACK       2009           2
## 6         25-44       M          BLACK       2020          10
```

```
# Summary of the cleaned dataset
summary(data)
```

```
##    OCCUR_DATE           OCCUR_TIME                   BORO
```

```
##  Min.   :2006-01-01   Length:27312       BRONX        : 7937
##  1st Qu.:2009-07-18   Class :character   BROOKLYN     :10933
##  Median :2013-04-29   Mode  :character   MANHATTAN    : 3572
##  Mean   :2014-01-06                      QUEENS       : 4094
##  3rd Qu.:2018-10-15                      STATEN ISLAND:  776
##  Max.   :2022-12-31
##
##  LOC_OF_OCCUR_DESC     PRECINCT     LOC_CLASSFCTN_DESC LOCATION_DESC
##  Length:27312       75     : 1557   Length:27312       Length:27312
##  Class :character   73     : 1452   Class :character   Class :character
##  Mode  :character   67     : 1216   Mode  :character   Mode  :character
##                     44     : 1020
##                     79     : 1012
##                     47     :  953
##                     (Other):20102
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Length:27312            Length:27312            : 9310
##  Class :character        Class :character   (null):  640
##  Mode  :character        Mode  :character   F     :  424
##                                             M     :15439
##                                             U     : 1499
##
##
##           PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  BLACK         :11432   Length:27312       F: 2615
##                : 9310   Class :character   M:24686
##  WHITE HISPANIC: 2341   Mode  :character   U:   11
##  UNKNOWN       : 1836
##  BLACK HISPANIC: 1314
##  (null)        :  640
##  (Other)       :  439
##                             VIC_RACE      OCCUR_YEAR    OCCUR_MONTH
##  AMERICAN INDIAN/ALASKAN NATIVE:   10   Min.   :2006   Min.   : 1.000
##  ASIAN / PACIFIC ISLANDER      :  404   1st Qu.:2009   1st Qu.: 5.000
##  BLACK                         :19439   Median :2013   Median : 7.000
##  BLACK HISPANIC                : 2646   Mean   :2013   Mean   : 6.825
##  UNKNOWN                       :   66   3rd Qu.:2018   3rd Qu.: 9.000
##  WHITE                         :  698   Max.   :2022   Max.   :12.000
##  WHITE HISPANIC                : 4049
```

Data Analysis & Visualizations

Here we look descriptive states of the data:
1) Shootings per Year
2) Shootings by Borough.

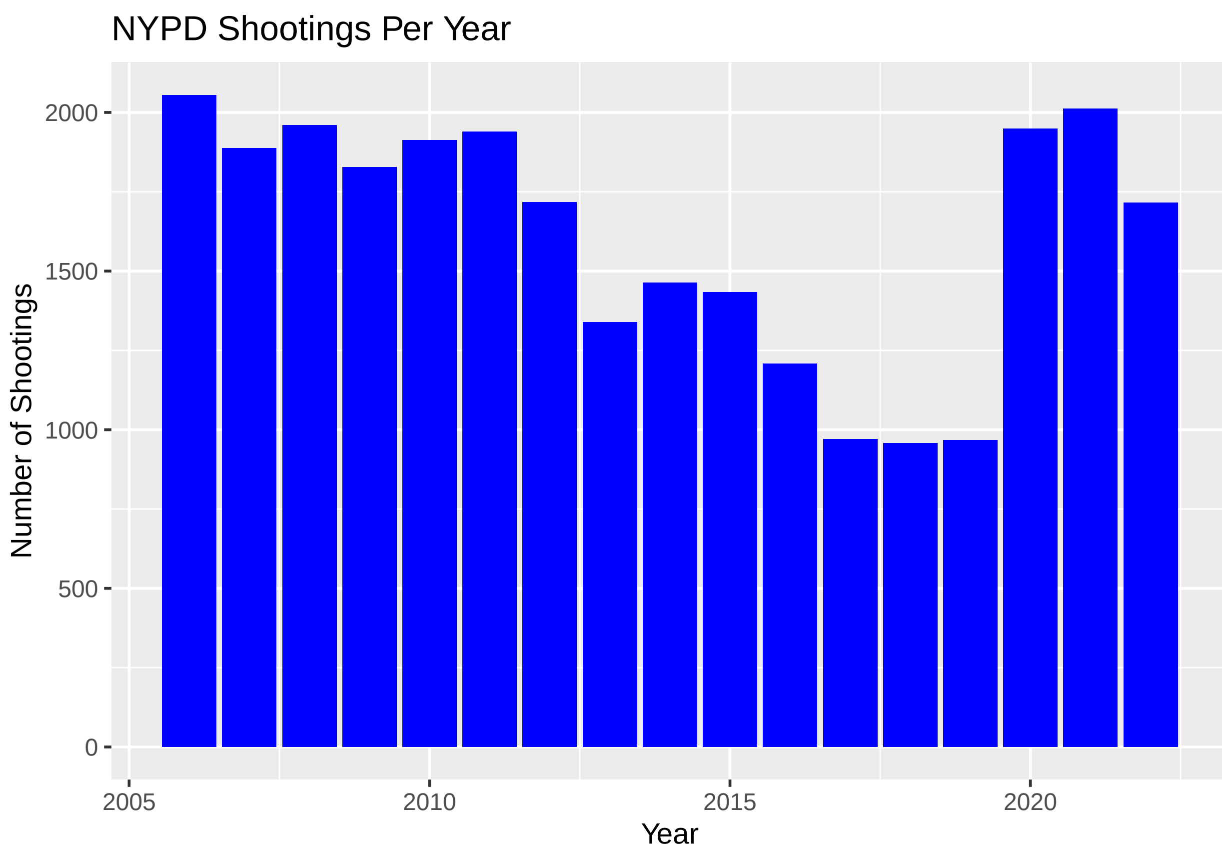
These are two very straightforward analytics to provide some trend analysis of the data.

In the first visualization, we find shootings by year to roughly then sharply trend downward over a decade then spike to previous highs over 2020 and beyond. This first visualization should provoke analysis from policymakers and law enforcement to try to understand the reasons behind the trend. Why the decrease in shootings? What contributed to that successful reduction in gun violence? Why the spike in 2020? What contributed to the spike and what mitigations might we put in place to continue the original downward trend again?

The second visualization highlights quantity of shootings by borough and here we see two boroughs Bronx and Brooklyn with the most. This might prompt reflection on where city resources (dollars, law enforcement, community programs, and policy changes etc) might best be allocated to mitigate these crimes.

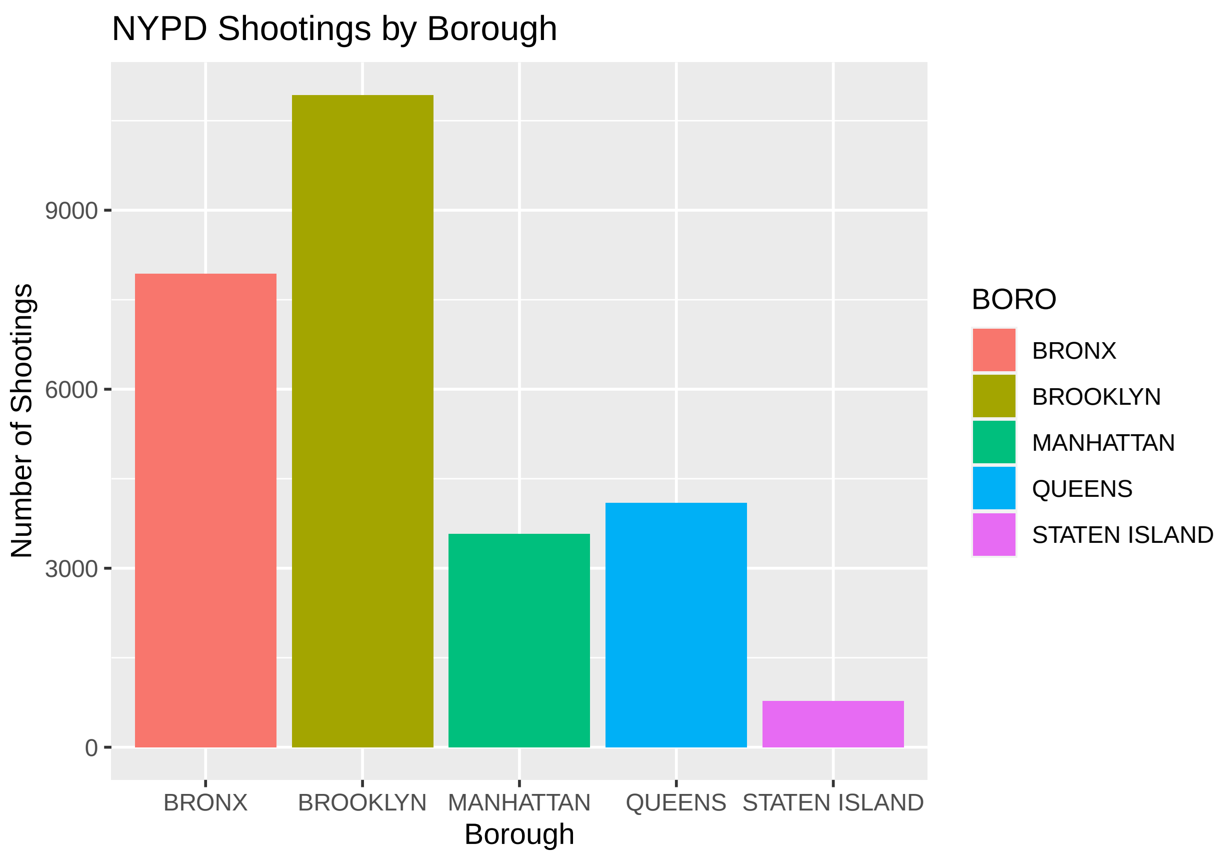
```r
# Number of shootings per year
shootings_per_year <- data %>%
  group_by(OCCUR_YEAR) %>%
  summarise(Shootings = n())

# Plot shootings per year
ggplot(shootings_per_year, aes(x = OCCUR_YEAR, y = Shootings)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "NYPD Shootings Per Year", x = "Year", y = "Number of Shootings")
```




```r
# Number of shootings by borough
shootings_by_borough <- data %>%
  group_by(BORO) %>%
  summarise(Shootings = n())

# Plot shootings by borough
ggplot(shootings_by_borough, aes(x = BORO, y = Shootings, fill = BORO)) +
  geom_bar(stat = "identity") +
  labs(title = "NYPD Shootings by Borough", x = "Borough",y = "Number of Shootings")
```



Processing Missing Values

```
# Check for missing data
missing_data <- sapply(data, function(x) sum(x == ""))
missing_data
```

```
##            OCCUR_DATE            OCCUR_TIME                  BORO
##                    NA                     0                     0
##       LOC_OF_OCCUR_DESC              PRECINCT     LOC_CLASSFCTN_DESC
##                 25596                     0                 25596
##         LOCATION_DESC STATISTICAL_MURDER_FLAG        PERP_AGE_GROUP
##                 14977                     0                  9344
##              PERP_SEX             PERP_RACE         VIC_AGE_GROUP
##                  9310                  9310                     0
##               VIC_SEX              VIC_RACE            OCCUR_YEAR
##                     0                     0                     0
##           OCCUR_MONTH
##                     0
```

```
# Remove rows with NA values
data <- na.omit(data)
# Remove rows with missing values in specific columns with values of na or empty string
data <- data %>%
```

```
  filter(!is.na(PERP_SEX) & PERP_SEX != "",
         !is.na(PERP_RACE) & PERP_RACE != "",
         !is.na(VIC_SEX) & VIC_SEX != "",
         !is.na(VIC_RACE) & VIC_RACE != "")
```

Missing Value Analysis

We have strategies for handling missing values from imputation to removing rows with missing data entirely from the data set. In this instance, I chose to remove rows with missing values from columns used in a Logistic Regression Analysis below. I chose to remove rows with missing values from columns listed below rather than utilize an imputation strategy out of concern that bias introduced by such a strategy could have real world policy implications, particularly given the nature of this data set. Of course, removing rows also introduces distribution skew and may result in under-representation of examples in the dataset affecting model performance on real world data.

The challenge is to understand why the values are missing (missing at random, missing completely at random, missing not at random) and while there are techniques to assess the type of missingness in the data (for example running an MCAR test), this isn't necessarily definitive. For brevity, let's assume missing completely at random and drop examples with missing values, and acknowledge this bias introduced into the analyssis in the conclusion.

Logistic Regression Model

Here we look to try to predict fatality of a shooting based on various dimensions of the data

```
# Load required libraries
library(tidyr)
library(caret)

# Preprocess data for modeling
 model_data <- data %>%
   select(STATISTICAL_MURDER_FLAG, VIC_AGE_GROUP, BORO, PERP_RACE, PERP_SEX) %>%
   filter(!is.na(STATISTICAL_MURDER_FLAG) & !is.na(VIC_AGE_GROUP) & !is.na(PERP_SEX) & !is.na(PERP_RACE
   mutate(STATISTICAL_MURDER_FLAG = as.factor(STATISTICAL_MURDER_FLAG))

# Split the data into training and test sets
set.seed(123)
train_index <- createDataPartition(model_data$STATISTICAL_MURDER_FLAG, p = 0.8, list = FALSE)
train_data <- model_data[train_index, ]
test_data <- model_data[-train_index, ]

# Fit the logistic regression model
logistic_model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP + PERP_SEX + PERP_RACE,
                      data = train_data, family = binomial(link = "logit"))

# Model summary
summary(logistic_model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP +
##     PERP_SEX + PERP_RACE, family = binomial(link = "logit"),
```

```
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2084  -0.7038  -0.6403  -0.3692   2.7150
##
## Coefficients: (1 not defined because of singularities)
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -2.06456    0.14725 -14.021  < 2e-16
## BOROBROOKLYN                             -0.16703    0.05273  -3.167 0.001538
## BOROMANHATTAN                            -0.23455    0.06881  -3.409 0.000652
## BOROQUEENS                               -0.16790    0.06699  -2.506 0.012198
## BOROSTATEN ISLAND                        -0.08223    0.11405  -0.721 0.470890
## VIC_AGE_GROUP1022                        -9.66831  196.96770  -0.049 0.960851
## VIC_AGE_GROUP18-24                        0.34979    0.08150   4.292 1.77e-05
## VIC_AGE_GROUP25-44                        0.56094    0.07928   7.075 1.49e-12
## VIC_AGE_GROUP45-64                        0.68074    0.10346   6.580 4.71e-11
## VIC_AGE_GROUP65+                          0.89982    0.21330   4.218 2.46e-05
## VIC_AGE_GROUPUNKNOWN                      0.53851    0.36827   1.462 0.143667
## PERP_SEXF                                 0.72735    0.18202   3.996 6.44e-05
## PERP_SEXM                                 0.55647    0.13681   4.067 4.75e-05
## PERP_SEXU                                 1.14771    0.30520   3.761 0.000170
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -10.69352  138.38607  -0.077 0.938406
## PERP_RACEASIAN / PACIFIC ISLANDER         0.36420    0.20428   1.783 0.074613
## PERP_RACEBLACK                           -0.15511    0.06124  -2.533 0.011313
## PERP_RACEBLACK HISPANIC                  -0.19851    0.09405  -2.111 0.034797
## PERP_RACEUNKNOWN                         -1.91749    0.26394  -7.265 3.74e-13
## PERP_RACEWHITE                            0.68086    0.14454   4.710 2.47e-06
## PERP_RACEWHITE HISPANIC                        NA         NA      NA       NA
##
## (Intercept)                             ***
## BOROBROOKLYN                            **
## BOROMANHATTAN                           ***
## BOROQUEENS                              *
## BOROSTATEN ISLAND
## VIC_AGE_GROUP1022
## VIC_AGE_GROUP18-24                      ***
## VIC_AGE_GROUP25-44                      ***
## VIC_AGE_GROUP45-64                      ***
## VIC_AGE_GROUP65+                        ***
## VIC_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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14362  on 14401  degrees of freedom
## Residual deviance: 13964  on 14382  degrees of freedom
## AIC: 14004
##
## Number of Fisher Scoring iterations: 10
```

```
#Predict on test data
predictions <- predict(logistic_model, newdata = test_data, type="response")
predicted_output <- ifelse(predictions > 0.5, "1", "0")
confusion_matrix_output <- table(Predicted = predicted_output, Actual = test_data$STATISTICAL_MURDER_FL

# Calculate accuracy
accuracy <- sum(diag(confusion_matrix_output)) / sum(confusion_matrix_output)
paste("Accuracy:", round(accuracy * 100, 2), "%")
```

```
## [1] "Accuracy: 80.14 %"
```

Conclusion

The model took into account several factors, including the presence of a murder flag, borough, victim age group, perpetrator sex, and perpetrator race. By employing an 80/20 train-test split and excluding rows with missing values, the model achieved an accuracy rate of 80.13%.

Model Features:

The logistic regression model was designed to predict fatalities based on the following features:

{STATISTICAL_MURDER_FLAG, BORO. VIC_AGE_GROUP, PERP_SEX, PERP_RACE}

Model Performance: To assess the model's performance, the data set was split into training and testing sets using a standard 80/20 ratio. The logistic regression model was trained on the 80% training data set and then used to predict fatalities on the 20% test data set.

The model predictions were transformed into binary classes with predicted fatalities being labeled as "1" and non-fatalities as "0". A confusion matrix was constructed by comparing the predicted classes against the actual data for the test dataset.

Accuracy Calculation: The accuracy of the model was calculated by dividing the sum of correctly predicted fatalities and non-fatalities by the total number of predictions. This resulted in an accuracy rate of 80.13%.

Conclusion: The logistic regression model trained here demonstrated a promising accuracy rate of 80.13% in predicting fatalities based on the selected features. This suggests that the model has the potential to be a valuable tool in analyzing and understanding crime patterns/gun violence patterns in New York City. Refinement of the model, along with the inclusion of additional factors, could lead to more accurate predictions and a deeper understanding of the factors that contribute to fatal incidents.

A note on bias:

1. Outlier Analysis: A more robust analysis would have included evaluating outliers. Logistic regression is sensitive to the presence of outliers because it estimates the probability of a certain outcome (usually coded as 0 or 1) based on the values of predictor variables. Outliers can influence the estimates of the regression coefficients, which in turn can affect the predicted probabilities of the outcomes. In this case ~80% accuracy performance is fairly good. Removing outliers may improve the analysis.

2. Missing Values: As discussed, the strategy for handling outliers was to remove rows with missing values. If the missing data was missing completely at random, this strategy is fine. If it was missing at random or missing not at random, imputation would be the preferred strategy.

3. Personal Bias: I would not suggest that any personal feelings on this subject matter influenced this specific analysis however as noted in class personal bias is real and good data science practitioners should be aware of these when beginning any type of analysis.