NYPD Shooting

Luis Ruiz

11/07/2022

Introduction

Our data-set contains a list of every shooting incident that occurred in NYC going back to 2006-2020. This data was manually extracted every quarter and reviewed by the Office of Management Analysis and Planning. The data-set can be found **here**.

Exploratory Data Analysis

Load the Data

```r
df <- read.csv("NYPD_Shooting_Incident_Data__Historic_.csv")
str(df)
```

```
## 'data.frame':    25596 obs. of  19 variables:
##  $ INCIDENT_KEY           : int  236168668 231008085 230717903 237712309 224465521 228252164 22695001
##  $ OCCUR_DATE             : chr  "11/11/2021" "07/16/2021" "07/11/2021" "12/11/2021" ...
##  $ OCCUR_TIME             : chr  "15:04:00" "22:05:00" "01:09:00" "13:42:00" ...
##  $ BORO                   : chr  "BROOKLYN" "BROOKLYN" "BROOKLYN" "BROOKLYN" ...
##  $ PRECINCT               : int  79 72 79 81 113 113 42 52 34 75 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LOCATION_DESC          : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "false" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "45-64" "<18" "" ...
##  $ PERP_SEX               : chr  "" "M" "M" "" ...
##  $ PERP_RACE              : chr  "" "ASIAN / PACIFIC ISLANDER" "BLACK" "" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "25-44" "25-44" "25-44" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "ASIAN / PACIFIC ISLANDER" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  996313 981845 996546 1001139 1050710 ...
##  $ Y_COORD_CD             : num  187499 171118 187436 192775 184826 ...
##  $ Latitude               : num  40.7 40.6 40.7 40.7 40.7 ...
##  $ Longitude              : num  -74 -74 -74 -73.9 -73.8 ...
##  $ Lon_Lat                : chr  "POINT (-73.95650899099996 40.68131820000008)" "POINT (-74.0086666
```

After loading our data we were able to identify a few issues right off the bat. It appears we have a number of columns with the wrong datatype, unnecessary columns and missing values. We will need to dive deeper into these issues.

Tidying & Transforming

```
df$OCCUR_DATE <- mdy(df$OCCUR_DATE)
df$OCCUR_TIME <- format(df$OCCUR_TIME, format="%H:%M:%S")
df$MONTH <- month(df$OCCUR_DATE)
df$DAY <- day(df$OCCUR_DATE)
df$YEAR <- year(df$OCCUR_DATE)
df$WEEKDAY <- weekdays(df$OCCUR_DATE)

df <-  select(df, -c("Longitude","Latitude",
        "Lon_Lat","X_COORD_CD","Y_COORD_CD",
        "JURISDICTION_CODE","LOCATION_DESC","PRECINCT" ))
```

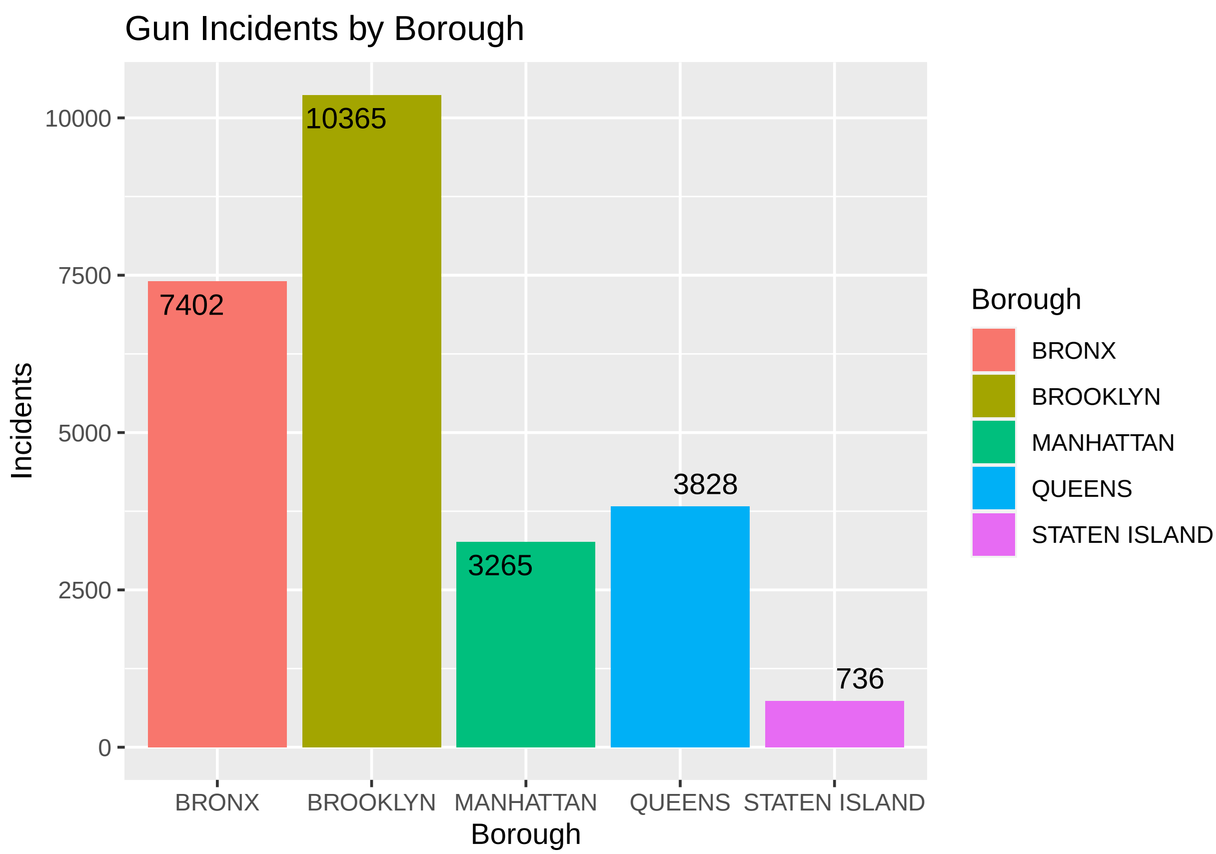
```
missingRace <- df[!(is.na(df$PERP_RACE) | df$PERP_RACE==""),]
missingAge<- df[!(is.na(df$PERP_AGE_GROUP) | df$PERP_AGE_GROUP==""),]
```

After manually going through the data-set I was able to convert OCCUR_DATE and OCCUR_TIME to the correct datatype. I also created four new columns . This will allow for a deeper analysis. There were several columns that were no use to us as it contained coordinates. I omitted observations that did not contain the age and race.

Visualizations

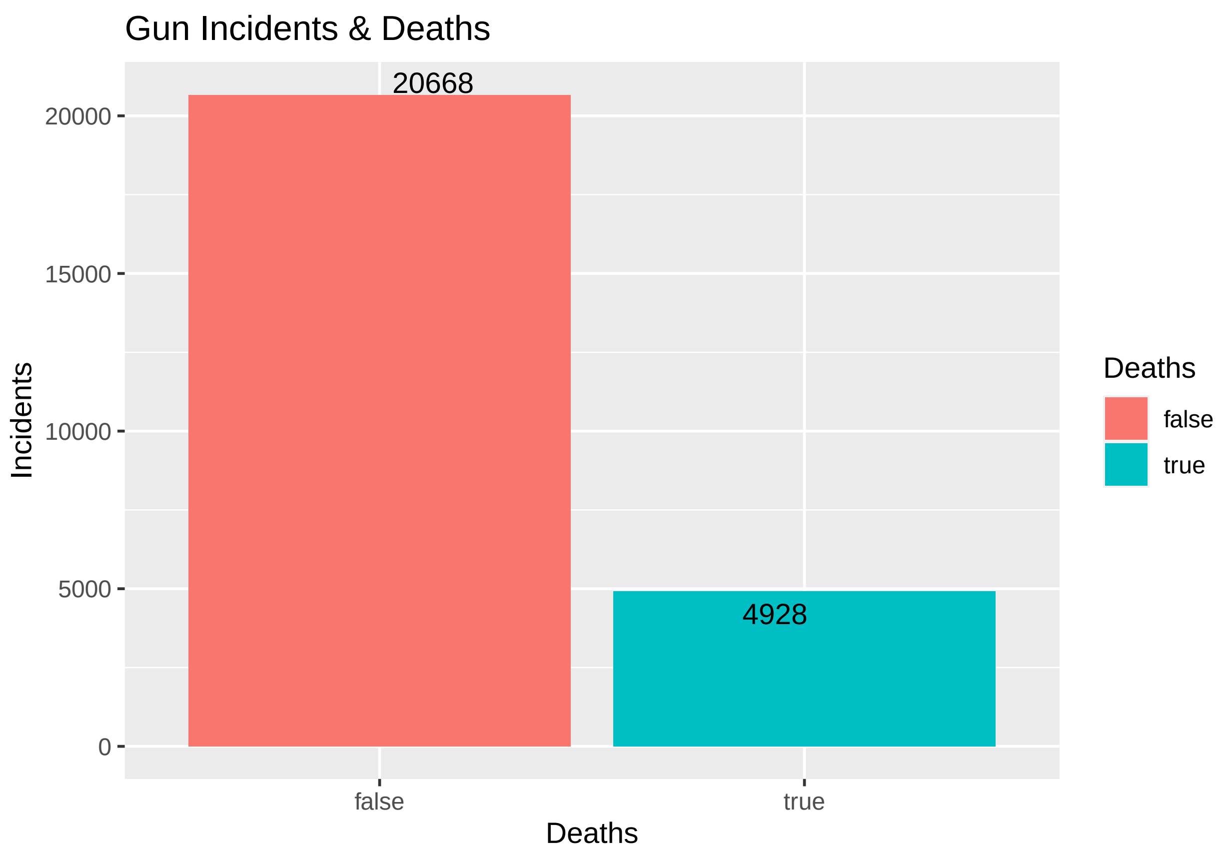
```
boroughs <- as.data.frame(table(df$BORO, dnn=list('Borough')), responseName='Incidents')

ggplot(boroughs, aes(x=Borough, y=Incidents, fill=Borough))+
  ggtitle("Gun Incidents by Borough") +
  geom_bar(stat="identity") +
  geom_text_repel(data=boroughs, aes(label=Incidents))
```



Grouping the data by borough gave us a rough overview of where the shooting are taking place in NYC. The borough with the highest number of incidents is Brooklyn at over 10,000 while Staten Island seems to be an outlier with less than 1000 incidents.

```
deaths <- as.data.frame(table(df$STATISTICAL_MURDER_FLAG,
      dnn=list('Deaths')), responseName='Incidents')
ggplot(deaths, aes(x=Deaths, y=Incidents,fill=Deaths)) +     ggtitle("Gun Incidents & Deaths") +
  geom_bar(stat="identity") +
   geom_text_repel(data=deaths,aes(label=Incidents))
```



We have a total of 25,596 gun incidents in NYC and out of those incidents about 19% or 4,928 ended up as murders. This leaves more questions to be answered. Such as, how many are classified as murders? Victims' age? Victims' race? Gender? We'll try to answer those questions deeper in our analysis.

Demographics

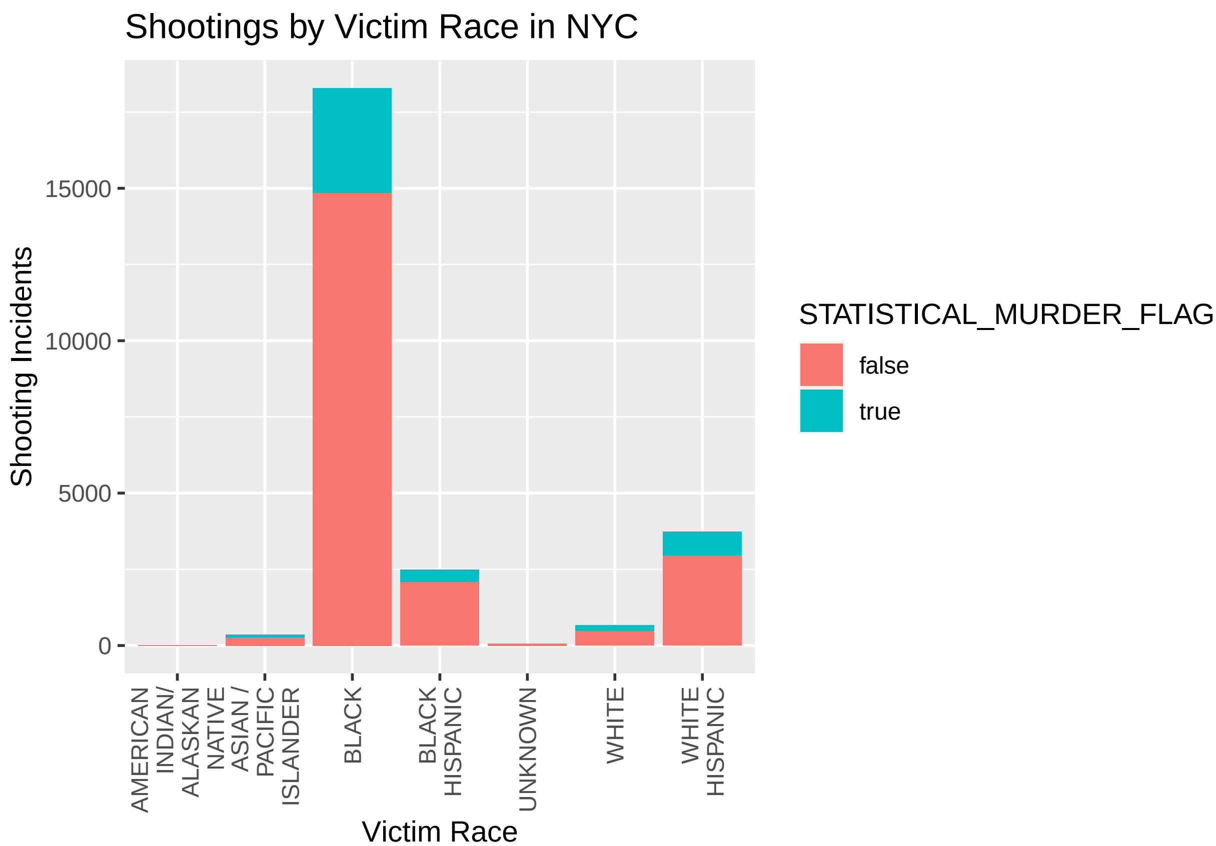
```
victim_race <- df %>%
  select(YEAR, BORO, VIC_RACE, STATISTICAL_MURDER_FLAG) %>%
  group_by(YEAR, BORO, VIC_RACE, STATISTICAL_MURDER_FLAG) %>%
  count(YEAR, BORO, VIC_RACE, STATISTICAL_MURDER_FLAG)

victim_age <- df %>%
  select(YEAR, BORO, VIC_AGE_GROUP, STATISTICAL_MURDER_FLAG) %>%
  group_by(YEAR, BORO, VIC_AGE_GROUP, STATISTICAL_MURDER_FLAG) %>%
  count(YEAR, BORO, VIC_AGE_GROUP, STATISTICAL_MURDER_FLAG)

victim_gender <- df %>%
  select(YEAR, BORO, VIC_SEX, STATISTICAL_MURDER_FLAG) %>%
  group_by(YEAR, BORO, VIC_SEX, STATISTICAL_MURDER_FLAG) %>%
  count(YEAR, BORO, VIC_SEX, STATISTICAL_MURDER_FLAG)
```

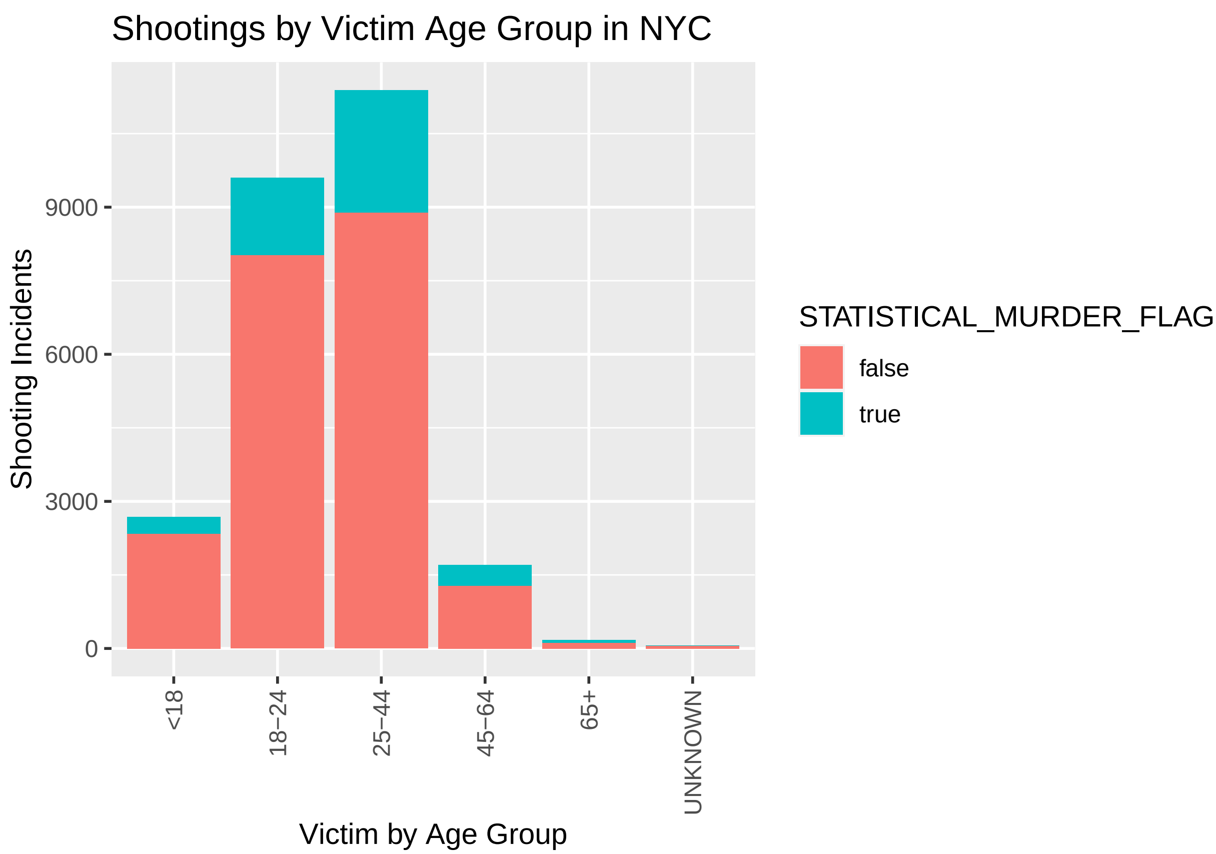
```
ggplot(data = victim_race, mapping = aes(x = VIC_RACE, fill=STATISTICAL_MURDER_FLAG, y=n)) +
geom_bar(position = position_stack(reverse = TRUE), stat="identity") +
labs(x = "Victim Race", y="Shooting Incidents", title="Shootings by Victim Race in NYC") +
```

```
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
scale_x_discrete(labels = function(x) str_wrap(x, width = 10))
```



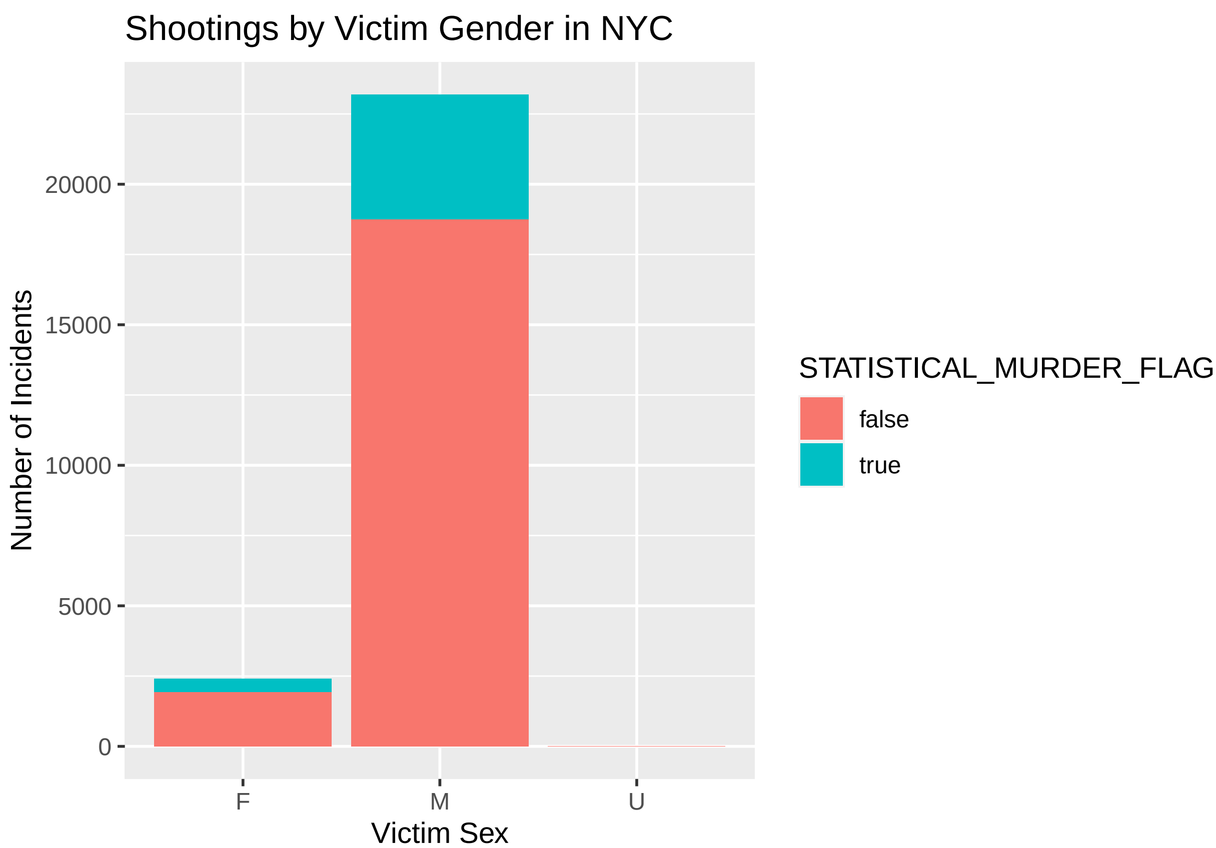
As we can see African American are disproportionately more likely to be involve in gun incidents and die because of it. Followed by White Hispanics and Black Hispanics. The disproportionate amount of Black victims leaves many questions to be answered. Why are they more likely to be involve considering they make up less than 30% of the population in NYC. Who is involve in these incidents? Police officers?

```
ggplot(data = victim_age, mapping = aes(x = VIC_AGE_GROUP, fill=STATISTICAL_MURDER_FLAG, y=n)) +
geom_bar(position = position_stack(reverse = TRUE), stat="identity") +
labs(x = "Victim by Age Group", y="Shooting Incidents", title="Shootings by Victim Age Group in NYC") +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
scale_x_discrete(labels = function(x) str_wrap(x, width = 10))
```



The most frequent age group involved in gun incident are young adults between the ages of 25-44. Closely followed by the adults between the ages of 25-44. It is unfortunate that our data-set grouped the ages instead of discrete values.

```
ggplot(data = victim_gender, mapping = aes(x = VIC_SEX, fill = STATISTICAL_MURDER_FLAG, y = n)) +
geom_bar(position = position_stack(reverse = TRUE), stat = "identity") +
labs(x = "Victim Sex", y = "Number of Incidents", title = "Shootings by Victim Gender in NYC")
```



Overwhelming majoring of the victims are Males.

Modeling

Train Test Split

```
library(caTools)
# Convert column to 1s and 0s
df$STATISTICAL_MURDER_FLAG <- as.integer(as.logical(df$STATISTICAL_MURDER_FLAG))
# Split data into a train and test set
split <- sample.split(df$STATISTICAL_MURDER_FLAG, SplitRatio = 0.7)
train <- subset(df, split == TRUE)
test <- subset(df, split == FALSE)
# Rows and Cols for training dataset
dim(train)
```

```
## [1] 17918    15
```

```
# Rows and Cols for testing dataset
dim(test)
```

```
## [1] 7678   15
```

```r
# Break down of deaths(1) vs No-deaths(0)
prop.table(table(train$STATISTICAL_MURDER_FLAG))
```

```
##
##         0         1
## 0.8074562 0.1925438
```

```r
prop.table(table(test$STATISTICAL_MURDER_FLAG))
```

```
##
##        0        1
## 0.807502 0.192498
```

```r
# Created logistic regression model
logitModel <- glm(STATISTICAL_MURDER_FLAG ~ BORO + PERP_AGE_GROUP + PERP_SEX + PERP_RACE + VIC_AGE_GROU
```

```r
summary(logitModel)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + PERP_AGE_GROUP +
##     PERP_SEX + PERP_RACE + VIC_AGE_GROUP + VIC_SEX + VIC_RACE,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4919  -0.6967  -0.6082  -0.2024   3.1619
##
## Coefficients: (1 not defined because of singularities)
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -13.60403  201.24197  -0.068  0.94610
## BOROBROOKLYN                              0.01967    0.04905   0.401  0.68837
## BOROMANHATTAN                            -0.11729    0.06648  -1.764  0.07770
## BOROQUEENS                                0.01529    0.06230   0.245  0.80608
## BOROSTATEN ISLAND                        -0.13018    0.12044  -1.081  0.27975
## PERP_AGE_GROUP<18                        -0.43810    0.49632  -0.883  0.37740
## PERP_AGE_GROUP1020                      -12.69851  535.41139  -0.024  0.98108
## PERP_AGE_GROUP18-24                      -0.29823    0.49049  -0.608  0.54317
## PERP_AGE_GROUP224                       -12.60487  535.41140  -0.024  0.98122
## PERP_AGE_GROUP25-44                      -0.04468    0.49060  -0.091  0.92744
## PERP_AGE_GROUP45-64                       0.27174    0.50194   0.541  0.58825
## PERP_AGE_GROUP65+                         0.18780    0.59406   0.316  0.75190
## PERP_AGE_GROUP940                       -12.87432  535.41140  -0.024  0.98082
## PERP_AGE_GROUPUNKNOWN                    -3.02961    0.44500  -6.808 9.89e-12
## PERP_SEXF                                 0.81869    0.51434   1.592  0.11145
## PERP_SEXM                                 0.58756    0.49488   1.187  0.23512
## PERP_SEXU                                 2.19451    0.49857   4.402 1.07e-05
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -12.78132  378.56167  -0.034  0.97307
## PERP_RACEASIAN / PACIFIC ISLANDER         0.29448    0.23512   1.252  0.21040
## PERP_RACEBLACK                           -0.09134    0.07250  -1.260  0.20768
## PERP_RACEBLACK HISPANIC                  -0.24833    0.11017  -2.254  0.02419
```

```
## PERP_RACEUNKNOWN                         -0.70654    0.26725  -2.644  0.00820
## PERP_RACEWHITE                            0.22600    0.18219   1.240  0.21480
## PERP_RACEWHITE HISPANIC                        NA         NA      NA       NA
## VIC_AGE_GROUP18-24                        0.22991    0.07831   2.936  0.00332
## VIC_AGE_GROUP25-44                        0.47969    0.07706   6.225 4.82e-10
## VIC_AGE_GROUP45-64                        0.55900    0.10022   5.578 2.44e-08
## VIC_AGE_GROUP65+                          0.91298    0.22594   4.041 5.33e-05
## VIC_AGE_GROUPUNKNOWN                      0.65567    0.38534   1.702  0.08884
## VIC_SEXM                                  0.01641    0.06595   0.249  0.80351
## VIC_SEXU                                -11.82615  201.47140  -0.059  0.95319
## VIC_RACEASIAN / PACIFIC ISLANDER         11.87386  201.24201   0.059  0.95295
## VIC_RACEBLACK                            11.74416  201.24195   0.058  0.95346
## VIC_RACEBLACK HISPANIC                   11.53896  201.24196   0.057  0.95428
## VIC_RACEUNKNOWN                          10.96288  201.24256   0.054  0.95656
## VIC_RACEWHITE                            11.92894  201.24198   0.059  0.95273
## VIC_RACEWHITE HISPANIC                   11.80895  201.24195   0.059  0.95321
##
## (Intercept)
## BOROBROOKLYN
## BOROMANHATTAN                           .
## BOROQUEENS
## BOROSTATEN ISLAND
## PERP_AGE_GROUP<18
## PERP_AGE_GROUP1020
## 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_GROUP18-24                      **
## VIC_AGE_GROUP25-44                      ***
## VIC_AGE_GROUP45-64                      ***
## VIC_AGE_GROUP65+                        ***
## VIC_AGE_GROUPUNKNOWN                    .
## VIC_SEXM
## VIC_SEXU
## VIC_RACEASIAN / PACIFIC ISLANDER
## VIC_RACEBLACK
## VIC_RACEBLACK HISPANIC
## VIC_RACEUNKNOWN
## VIC_RACEWHITE
## VIC_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: 17556  on 17917  degrees of freedom
## Residual deviance: 16628  on 17882  degrees of freedom
## AIC: 16700
##
## Number of Fisher Scoring iterations: 12
```

Training Data Split

```
library(caret)
p1 <- predict(logitModel, train, type = "response")

pred1 <- ifelse(p1>0.5, 1,0)
tab1 <- table(Predicted = pred1, Actual = train$STATISTICAL_MURDER_FLAG)
```

Testing Data Split

```
p2 <- predict(logitModel, test, type = "response")

pred2 <- ifelse(p2>0.5, 1,0)
tab2 <- table(Predicted = pred2, Actual = test$STATISTICAL_MURDER_FLAG)
```

Evaluation

```
confusionMatrix(tab1)
```

```
## Confusion Matrix and Statistics
##
##          Actual
## Predicted     0     1
##         0 14455  3438
##         1    13    12
##
##                Accuracy : 0.8074
##                  95% CI : (0.8015, 0.8132)
##     No Information Rate : 0.8075
##     P-Value [Acc > NIR] : 0.5121
##
##                   Kappa : 0.0041
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.999101
##             Specificity : 0.003478
```

```
##          Pos Pred Value : 0.807858
##          Neg Pred Value : 0.480000
##              Prevalence : 0.807456
##          Detection Rate : 0.806731
##    Detection Prevalence : 0.998605
##       Balanced Accuracy : 0.501290
##
##        'Positive' Class : 0
##
```

```
confusionMatrix(tab2)
```

```
## Confusion Matrix and Statistics
##
##          Actual
## Predicted    0    1
##         0 6191 1470
##         1    9    8
##
##                Accuracy : 0.8074
##                  95% CI : (0.7984, 0.8161)
##     No Information Rate : 0.8075
##     P-Value [Acc > NIR] : 0.5185
##
##                   Kappa : 0.0064
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.998548
##             Specificity : 0.005413
##          Pos Pred Value : 0.808119
##          Neg Pred Value : 0.470588
##              Prevalence : 0.807502
##          Detection Rate : 0.806330
##    Detection Prevalence : 0.997786
##       Balanced Accuracy : 0.501981
##
##        'Positive' Class : 0
##
```

Our model on the training data set contained 17,865 observations. It was able to predict those murdered with an 80% accuracy in both training and testing data. It does a good job in generalizing our data.

Conclusion

After processing, analyzing, modeling, and evaluating our NYC shooting data-set we were able to draw several important conclusions. But, first, we must discuss the underlying bias within our data-set. It is clear that African Americans are disproportionately more involved in gun incidents compared to other races. Why? Is the police targeting African Americans. We must understand the ethics behind our data and how imbalance in the data will influence future machine learning models.

Trying to tackle gun violence in the US is difficult but reduce the number of victims. We know largest group of victims are African Americans males between the ages of 25-44. Followed by victims between the ages of 18-24.