NYPD Shooting Incidents Analysis

D. Ivy

November 22, 2021

Data Overview

List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

Step 1: Read in data from the website

```
#Get data from website
url<-"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

#Assign data to a dataframe
data<-read.csv(url)

#Dataframe shape
dim(data)
```

```
## [1] 23585    19
```

```
#Dataframe columns
colnames(data)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "PRECINCT"                "JURISDICTION_CODE"
##  [7] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
##  [9] "PERP_AGE_GROUP"          "PERP_SEX"
## [11] "PERP_RACE"               "VIC_AGE_GROUP"
## [13] "VIC_SEX"                 "VIC_RACE"
## [15] "X_COORD_CD"              "Y_COORD_CD"
## [17] "Latitude"                "Longitude"
## [19] "Lon_Lat"
```

Step 2: Tidy and Transform Data

A detailed description of the data and what the column headers mean can be found here: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

The incident key variable is random so, by design, should not give us any information and we can remove this. We see there are several variables related to the location of the shooting. These are likely to be beyond the scope of this analysis so we can remove them. I have also chosen to remove the OCCUR_TIME variable. I dont plan on using the time in this analysis in any fashion. After this initial parsing of the data, lets take a look to see what we are left with.

```
data_cleaned<-subset(data,select=-c(INCIDENT_KEY,OCCUR_TIME,X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon

head(data_cleaned)
```

```
##   OCCUR_DATE     BORO PRECINCT JURISDICTION_CODE LOCATION_DESC
## 1 08/27/2006    BRONX       52                 0
## 2 03/11/2011   QUEENS      106                 0
## 3 10/06/2019 BROOKLYN       77                 0
## 4 09/04/2011    BRONX       40                 0
## 5 05/27/2013   QUEENS      100                 0
## 6 09/01/2013 BROOKLYN       67                 0
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                    true                                           25-44
## 2                   false                                             65+
## 3                   false                                           18-24
## 4                   false                                             <18
## 5                   false                                           18-24
## 6                   false                                             <18
##   VIC_SEX       VIC_RACE
## 1       F BLACK HISPANIC
## 2       M          WHITE
## 3       F          BLACK
## 4       M          BLACK
## 5       M          BLACK
## 6       M          BLACK
```

We see that the OCCUR_DATE variable needs to be read in as a date, the STATISTICAL_MURDER_FLAG variable is boolean, and all others are categorical.

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

data_cleaned$STATISTICAL_MURDER_FLAG<-as.integer(as.logical(data_cleaned$STATISTICAL_MURDER_FLAG))
non_factor_cols<-c("OCCUR_DATE","STATISTICAL_MURDER_FLAG")
factor_cols<-names(data_cleaned[names(data_cleaned)%in%non_factor_cols==FALSE])

data_cleaned[factor_cols]<-lapply(data_cleaned[factor_cols],as.factor)
```

We can now view the summary of the dataset to determine the next steps

```
summary(data_cleaned)
```

```
##    OCCUR_DATE                          BORO           PRECINCT      JURISDICTION_CODE
```

```
##  Min.   :2006-01-01   BRONX        :6701   75      : 1375   0   :19629
##  1st Qu.:2008-12-31   BROOKLYN     :9734   73      : 1284   1   :   54
##  Median :2012-02-27   MANHATTAN    :2922   67      : 1101   2   : 3900
##  Mean   :2012-10-05   QUEENS       :3532   79      :  921   NA's:    2
##  3rd Qu.:2016-03-02   STATEN ISLAND: 696   44      :  841
##  Max.   :2020-12-31                        47      :  818
##                                            (Other):17245
##                   LOCATION_DESC   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##                          :13581   Min.   :0.0000                 :8295
##  MULTI DWELL - PUBLIC HOUS: 4240   1st Qu.:0.0000          18-24   :5508
##  MULTI DWELL - APT BUILD  : 2553   Median :0.0000          25-44   :4714
##  PVT HOUSE                :  857   Mean   :0.1908          UNKNOWN:3148
##  GROCERY/BODEGA           :  574   3rd Qu.:0.0000          <18     :1368
##  BAR/NIGHT CLUB           :  562   Max.   :1.0000          45-64   : 495
##  (Other)                  : 1218                           (Other):   57
##  PERP_SEX           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##   : 8261   BLACK         :10025   <18    : 2525   F: 2204
##  F:  335                 : 8261   18-24  : 9003   M:21370
##  M:13490   WHITE HISPANIC: 1988   25-44  :10303   U:   11
##  U: 1499   UNKNOWN       : 1836   45-64  : 1541
##            BLACK HISPANIC: 1096   65+    :  154
##            WHITE         :  255   UNKNOWN:   59
##            (Other)       :  124
##                           VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  327
##  BLACK                         :16869
##  BLACK HISPANIC                : 2245
##  UNKNOWN                       :   65
##  WHITE                         :  620
##  WHITE HISPANIC                : 3450
```

There are two NAs in JURISDICTION_CODE. Given the large dataset, I think we can just remove these 2 observations without changing the results of any analysis we do. LOCATION_DESC has many categories and likely isnt going to tell us much so will remove that variable as well.

The 3 different PERP variables (age, sex, race) are somewhat sparse, perhaps these are unsolved cases at the time of input into the database. These variables have blanks as well as UNKNOWN or U entries. I will combine all of these entries into UNKNOWN or U.

```r
data_cleaned<-subset(data_cleaned,select=-c(LOCATION_DESC))
data_cleaned<-na.omit(data_cleaned)

data_cleaned$PERP_AGE_GROUP[data_cleaned$PERP_AGE_GROUP==""]<-"UNKNOWN"
data_cleaned$PERP_SEX[data_cleaned$PERP_SEX==""]<-"U"
data_cleaned$PERP_RACE[data_cleaned$PERP_RACE==""]<-"UNKNOWN"

summary(data_cleaned)
```

```
##    OCCUR_DATE                    BORO         PRECINCT     JURISDICTION_CODE
##  Min.   :2006-01-01   BRONX        :6701   75      : 1375   0:19629
##  1st Qu.:2008-12-31   BROOKLYN     :9734   73      : 1284   1:   54
##  Median :2012-02-27   MANHATTAN    :2921   67      : 1101   2: 3900
##  Mean   :2012-10-05   QUEENS       :3531   79      :  921
```

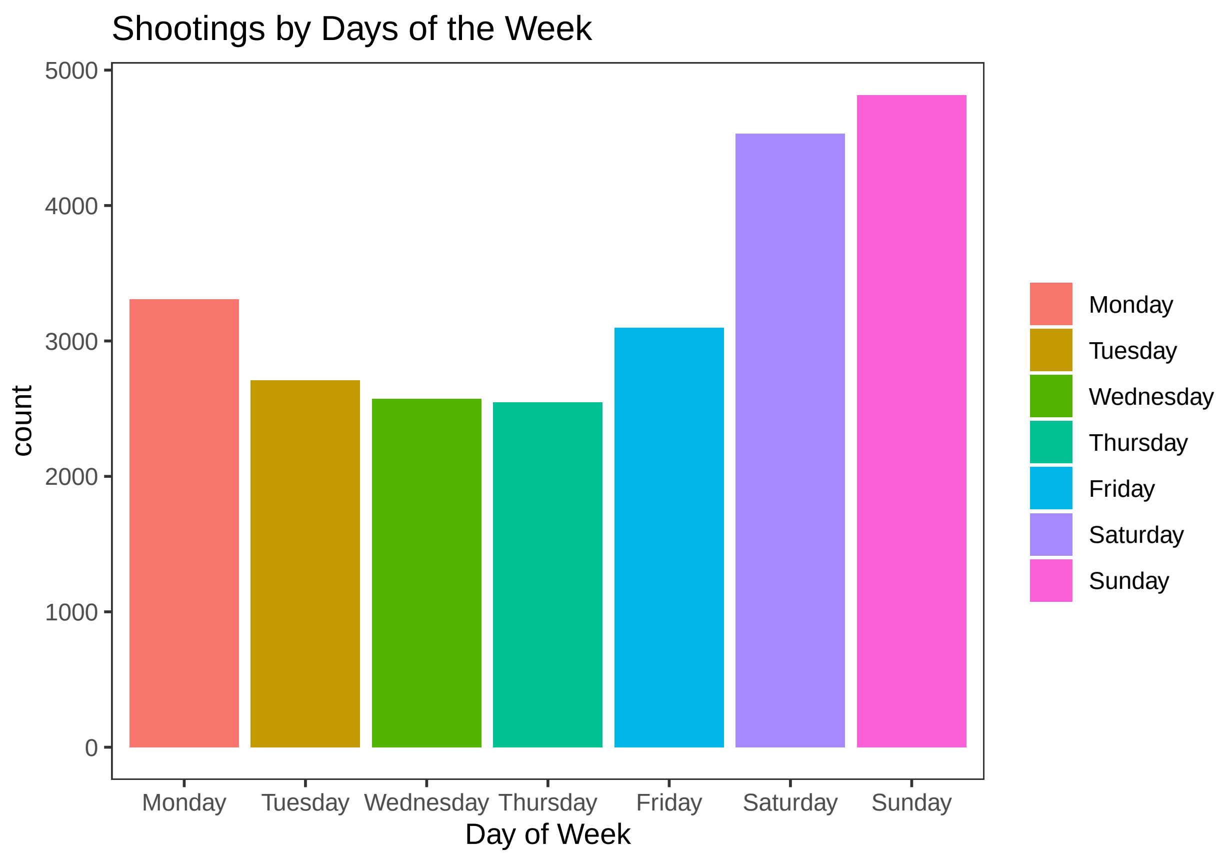
```
##  3rd Qu.:2016-03-01   STATEN ISLAND: 696   44     :  841
##  Max.   :2020-12-31                        47     :  818
##                                            (Other):17243
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP  PERP_SEX
##  Min.   :0.0000          UNKNOWN:11442    :    0
##  1st Qu.:0.0000          18-24  : 5508   F:  335
##  Median :0.0000          25-44  : 4714   M:13488
##  Mean   :0.1908          <18    : 1367   U: 9760
##  3rd Qu.:0.0000          45-64  :  495
##  Max.   :1.0000          65+    :   54
##                          (Other):    3
##                     PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  UNKNOWN                 :10097   <18    : 2525   F: 2204
##  BLACK                   :10024   18-24  : 9002   M:21368
##  WHITE HISPANIC          : 1987   25-44  :10302   U:   11
##  BLACK HISPANIC          : 1096   45-64  : 1541
##  WHITE                   :  255   65+    :  154
##  ASIAN / PACIFIC ISLANDER:  122   UNKNOWN:   59
##  (Other)                 :    2
##                              VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  327
##  BLACK                         :16868
##  BLACK HISPANIC                : 2245
##  UNKNOWN                       :   65
##  WHITE                         :  620
##  WHITE HISPANIC                : 3449
```

So, we see here we have a clean dataset and can continue onto visualizations and analysis.

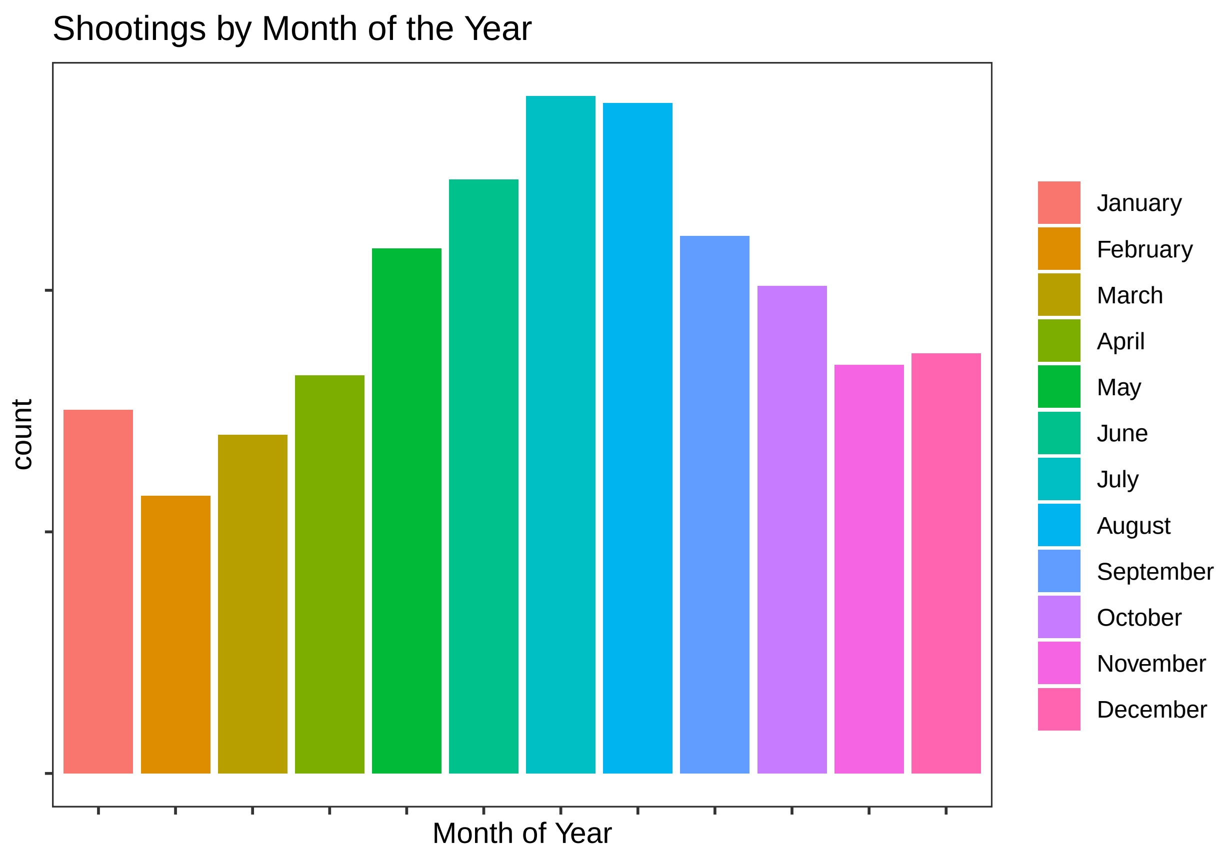
Step 3: Visualizations and Analysis

For the first couple of visualizations, I want to check for possible seasonality in the data. Do more shootings occur on the weekends? What about during the summer months?

```r
library(ggplot2)
data_cleaned$Weekday<-weekdays(data_cleaned$OCCUR_DATE)
data_cleaned$Weekday<-factor(data_cleaned$Weekday,c("Monday","Tuesday","Wednesday","Thursday","Friday",

g_weekdays<-ggplot(data=data_cleaned,aes(data_cleaned$Weekday,fill=data_cleaned$Weekday))+geom_bar()
g_weekdays<-g_weekdays+theme_bw()+theme(panel.grid.major=element_blank(),panel.grid.minor=element_blank
g_weekdays
```



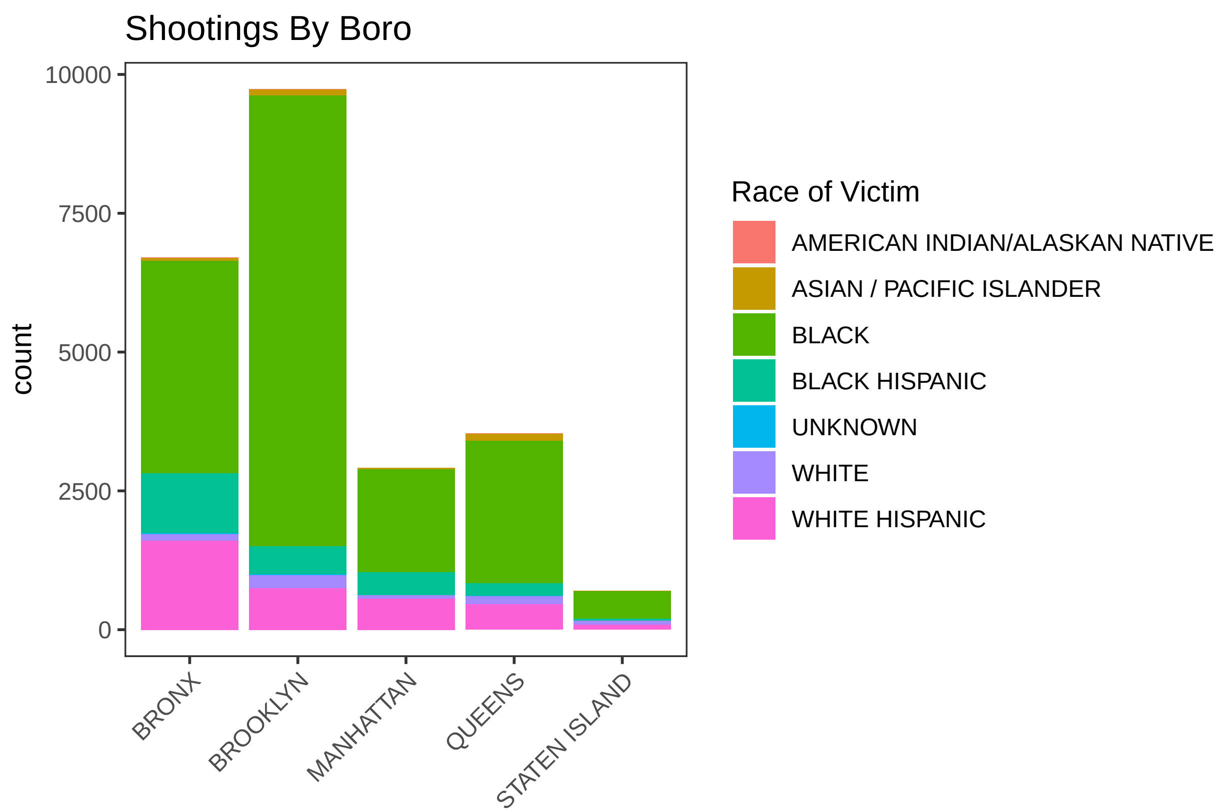

```
data_cleaned$Months<-months(data_cleaned$OCCUR_DATE)
data_cleaned$Months<-factor(data_cleaned$Months,c("January","February","March","April","May","June","Ju

g_months<-ggplot(data=data_cleaned,aes(data_cleaned$Months,fill=data_cleaned$Months))+geom_bar()
g_months<-g_months+theme_bw()+theme(panel.grid.major=element_blank(),panel.grid.minor=element_blank())+
g_months
```



We can clearly see that there appear to be more shootings on the weekends and during the warmer months of the year.

For the next visualization, I would like to look into the racial breakdown of shootings by Boro.

```
g_boro<-ggplot(data_cleaned,aes(BORO,fill=VIC_RACE))+geom_bar(position="stack")
g_boro<-g_boro+theme_bw()+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())+
g_boro
```



We can see here that most shooting victims, by far, were Black. We also see very little difference in the victims racial identification based on the Boro in which the incident took place.

Analysis

For analysis, I would like to run a logistic regression to predict how likely an incident will be a homicide, given the other variables available. An analysis like this could help the NYPD send the correct crime scene investigation unit to a reported incident.

```
#Percent of calls that are murders
perc_murders<-sum(data_cleaned$STATISTICAL_MURDER_FLAG)/nrow(data_cleaned)
perc_murders
```

```
## [1] 0.1908154
```

We immediately see there may be an issue of the dataset being unbalanced (there are far more shootings not resulting in murders that there are shootings that do result in a murder). The problems that this may cause are likely beyond the scope of this course so I will just ignore this issue for now and proceed as if the dataset were balanced.

```
#split training and testing sets
set.seed(12345)
train_idx<-sample(nrow(data_cleaned),size=0.8*nrow(data_cleaned),replace=FALSE)
train_set<-data_cleaned[train_idx,]
test_set<-data_cleaned[-train_idx,]
```

```r
#fit logistic regression model for binary classification on training set
lr_model_full<-glm(STATISTICAL_MURDER_FLAG~BORO+JURISDICTION_CODE+VIC_AGE_GROUP+VIC_RACE+VIC_SEX-1,data=
lr_model<-glm(STATISTICAL_MURDER_FLAG~JURISDICTION_CODE+VIC_AGE_GROUP+VIC_SEX-1,data=train_set, family=

summary(lr_model_full)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + JURISDICTION_CODE +
##     VIC_AGE_GROUP + VIC_RACE + VIC_SEX - 1, family = "binomial",
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0346  -0.7134  -0.6110  -0.5282   2.4587
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## BOROBRONX                       -12.91803  114.19102  -0.113    0.910
## BOROBROOKLYN                    -12.92839  114.19102  -0.113    0.910
## BOROMANHATTAN                   -12.99127  114.19103  -0.114    0.909
## BOROQUEENS                      -12.90310  114.19103  -0.113    0.910
## BOROSTATEN ISLAND               -12.87014  114.19107  -0.113    0.910
## JURISDICTION_CODE1                0.16157    0.35132   0.460    0.646
## JURISDICTION_CODE2               -0.27363    0.05418  -5.051 4.40e-07 ***
## VIC_AGE_GROUP18-24                0.30280    0.07440   4.070 4.71e-05 ***
## VIC_AGE_GROUP25-44                0.65798    0.07239   9.090  < 2e-16 ***
## VIC_AGE_GROUP45-64                0.81384    0.09451   8.611  < 2e-16 ***
## VIC_AGE_GROUP65+                  1.16670    0.20984   5.560 2.70e-08 ***
## VIC_AGE_GROUPUNKNOWN              0.66922    0.39470   1.696    0.090 .
## VIC_RACEASIAN / PACIFIC ISLANDER 11.26015  114.19109   0.099    0.921
## VIC_RACEBLACK                    11.00970  114.19100   0.096    0.923
## VIC_RACEBLACK HISPANIC           10.76737  114.19101   0.094    0.925
## VIC_RACEUNKNOWN                   9.93376  114.19230   0.087    0.931
## VIC_RACEWHITE                    11.33572  114.19104   0.099    0.921
## VIC_RACEWHITE HISPANIC           11.17224  114.19101   0.098    0.922
## VIC_SEXM                          0.02199    0.06444   0.341    0.733
## VIC_SEXU                         -0.22023    1.09556  -0.201    0.841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26154  on 18866  degrees of freedom
## Residual deviance: 18152  on 18846  degrees of freedom
## AIC: 18192
##
## Number of Fisher Scoring iterations: 11
```

```r
summary(lr_model)
```

```
##
```

```
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ JURISDICTION_CODE + VIC_AGE_GROUP +
##     VIC_SEX - 1, family = "binomial", data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9184  -0.7192  -0.6118  -0.5296   2.2506
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## JURISDICTION_CODE0   -1.90542    0.08719 -21.854  < 2e-16 ***
## JURISDICTION_CODE1   -1.66601    0.36008  -4.627 3.71e-06 ***
## JURISDICTION_CODE2   -2.19667    0.09776 -22.470  < 2e-16 ***
## VIC_AGE_GROUP18-24    0.31258    0.07430   4.207 2.58e-05 ***
## VIC_AGE_GROUP25-44    0.67288    0.07222   9.317  < 2e-16 ***
## VIC_AGE_GROUP45-64    0.85197    0.09369   9.094  < 2e-16 ***
## VIC_AGE_GROUP65+      1.24814    0.20816   5.996 2.02e-09 ***
## VIC_AGE_GROUPUNKNOWN  0.57913    0.38532   1.503    0.133
## VIC_SEXM              0.01215    0.06431   0.189    0.850
## VIC_SEXU             -0.85702    1.06863  -0.802    0.423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26154  on 18866  degrees of freedom
## Residual deviance: 18201  on 18856  degrees of freedom
## AIC: 18221
##
## Number of Fisher Scoring iterations: 4
```

I ran several models but only included two here for simplicity and comparison. We see fromthe full model, BORO and VIC_RACE are unlikely to be significant, while VIC_SEX is very borderline. JURISDICTION_CODE and VIC_AGE both appear to be very significant. This makes some sense. Older victims are less likely to survive being shot (hence the higher coefficient indicating a higher probability this incident will result in a murder). What is interesting is shootings in the housing projects are less likely to be murder that those incidents on the transit system. I would not have expected this, but perhaps this indicates the shootings in the transit system are much more likely to be at close range (and hence more deadly), while shootings in the housing projects could also very likely include accidental, self inflicted wounds.

```r
preds<-predict(lr_model,newdata=test_set)
preds<-exp(preds)
binary_prediction<-ifelse(preds>0.5,1,0)
true_vals<-test_set$STATISTICAL_MURDER_FLAG
accuracy<-mean(true_vals==binary_prediction)
accuracy
```

```
## [1] 0.8096248
```

So, we see an accuracy of our predictions of approximately 80%. This is an okay starting point, but doesnt really improve on the naive model of just assuming no shooting incident is a murder (remember only about 19% of the incidents were murders). This is likely a problem of having the class imbalance in the training set. Perhaps oversampling the minority class of our response variable will hep improve the number.

Step 4: Identifying Biases

For this task, I initially wanted to look at any possible seasonality of these incidents. I had an inclination that weekends and summer months were going to have more incidents, but I tried to mitigate my personal assumptions and let the numbers speak for themselves. If anything, they confirmed my expectations. It was interesting to see that victime race was not really a significant predictor according to the model, however the model clearly needs improvement and the data may not be the most trustworthy either as discussed below.

There are potentially significant sources of bias in the dat. The overwhelming majority of shooting victims were listed as Black. This seemed odd. The proportion of victims was significantly higher than the actual black population of New York City and also significantly higher than the other possible racial classifications in the study. We must question how this information is observed and confirmed. Are shooting victims actually filling out this information while they suffer from this traumatic experience? Is it simply the responding officer filling in this information solely based off observation? How do they know a person flagged as Blag is not, in fact, Black Hispanic? Similarly, will a fairer skinned victim be marked White Hispanic even if they are not of Hispanic descent? The significant racial component of this data has a very high likelihood of being affected by racial biases.

Session Info:

```r
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9        pillar_1.6.4     compiler_4.1.2   tools_4.1.2
##  [5] digest_0.6.28    evaluate_0.14    lifecycle_1.0.1  tibble_3.1.5
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.12     DBI_1.1.1
## [13] yaml_2.2.1       xfun_0.27        fastmap_1.1.0    withr_2.4.2
## [17] stringr_1.4.0    dplyr_1.0.7      knitr_1.36       generics_0.1.1
## [21] vctrs_0.3.8      grid_4.1.2       tidyselect_1.1.1 glue_1.4.2
## [25] R6_2.5.1         fansi_0.5.0      rmarkdown_2.11   purrr_0.3.4
## [29] farver_2.1.0     magrittr_2.0.1   scales_1.1.1     ellipsis_0.3.2
## [33] htmltools_0.5.2  assertthat_0.2.1 colorspace_2.0-2 labeling_0.4.2
## [37] utf8_1.2.2       stringi_1.7.5    munsell_0.5.0    crayon_1.4.2
```