myfirstRmd

2022-03-21

Shooting Project Data set

For this project, I am going to analyze trends in the NYPD shooting Incident. First step is to read in the data The summary of the data is shown below.

```
#imports
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("readr")
#install.packages("utils")
#install.packages("http://cran.rstudio.com/bin/windows/contrib/3.1/plyr_1.8.2.zip", repos = NULL)
#install.packages("pROC", dependencies=TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## 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(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

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

firstdata <- read_csv(url_in)
```

```
## Rows: 23585 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.
```

```
View(firstdata)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## modules/R_de.so'' had status 1
```

What does the data contain? I will be taking a look at the summary. Since I'm interested in finding out how victim race, victim age group, victim sex, and shooting location relates to murder (statistical_murder_flag), the first step is to delete variables I won't be using for sure in this analysis.

1. Delete variable "INCIDENT_KEY" (Unique to each incident)

2. Delete "JURISTICTION_CODE" and "LOCATION_DESC", and utilize other location data for this analysis.

3. Delete all long/lat data that's unique to each incident.

4. Delete PERP_AGE_GROUP, PERP_SEX, PERP_RACE. Since this is an open ended project where I can choose what to analyze, I will use the victim data and delete perp data (since it is less filled out).

```
summary(firstdata)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23585       Length:23585      Length:23585
##  1st Qu.: 55322804   Class :character   Class1:hms        Class :character
##  Median : 83435362   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102280741                      Mode  :numeric
##  3rd Qu.:150911774
##  Max.   :230611229
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   FALSE:19085
##  Median : 69.00   Median :0.000     Mode  :character   TRUE :4500
##  Mean   : 66.21   Mean   :0.333
```

```
##  3rd Qu.: 81.00   3rd Qu.:0.000
##  Max.   :123.00   Max.   :2.000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  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:23585       Length:23585       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999925   1st Qu.:182539
##  Mode  :character   Mode  :character   Median :1007654   Median :193470
##                                        Mean   :1009379   Mean   :207300
##                                        3rd Qu.:1016782   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23585
##  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
##
```

Delete the above mentioned variables.

```
df <- select (firstdata, -c ("INCIDENT_KEY", "JURISDICTION_CODE", "LOCATION_DESC", "PERP_AGE_GROUP", "PE
```

OCCUR_DATE is a character based on the summary. I will convert it to object.

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

```
##    OCCUR_DATE          OCCUR_TIME           BORO              PRECINCT
##  Min.   :2006-01-01   Length:23585       Length:23585       Min.   :  1.00
##  1st Qu.:2008-12-31   Class1:hms         Class :character   1st Qu.: 44.00
##  Median :2012-02-27   Class2:difftime    Mode  :character   Median : 69.00
##  Mean   :2012-10-05   Mode  :numeric                        Mean   : 66.21
##  3rd Qu.:2016-03-02                                         3rd Qu.: 81.00
##  Max.   :2020-12-31                                         Max.   :123.00
##  STATISTICAL_MURDER_FLAG VIC_AGE_GROUP        VIC_SEX
##  Mode :logical           Length:23585       Length:23585
##  FALSE:19085             Class :character   Class :character
##  TRUE :4500              Mode  :character   Mode  :character
##
##
##
##    VIC_RACE
##  Length:23585
```

```
##  Class :character
##  Mode  :character
##
##
##
```

BORO, PRECINCT, VIC_AGE_GROUP, VIC_SEX, VIC_RACE are categorical variables. I will convert them to be used as factors.

```
df$BORO <- as.factor(df$BORO)
df$PRECINCT <-as.factor(df$PRECINCT)
df$VIC_AGE_GROUP <-as.factor(df$VIC_AGE_GROUP)
df$VIC_SEX <-as.factor(df$VIC_SEX)
df$VIC_RACE <-as.factor(df$VIC_RACE)
summary(df)
```

```
##    OCCUR_DATE          OCCUR_TIME                  BORO         PRECINCT
##  Min.   :2006-01-01   Length:23585      BRONX        :6701   75     : 1375
##  1st Qu.:2008-12-31   Class1:hms        BROOKLYN     :9734   73     : 1284
##  Median :2012-02-27   Class2:difftime   MANHATTAN    :2922   67     : 1101
##  Mean   :2012-10-05   Mode  :numeric    QUEENS       :3532   79     :  921
##  3rd Qu.:2016-03-02                     STATEN ISLAND: 696   44     :  841
##  Max.   :2020-12-31                                          47     :  818
##                                                              (Other):17245
##  STATISTICAL_MURDER_FLAG VIC_AGE_GROUP   VIC_SEX
##  Mode :logical           <18    : 2525   F: 2204
##  FALSE:19085             18-24  : 9003   M:21370
##  TRUE :4500              25-44  :10303   U:   11
##                          45-64  : 1541
##                          65+    :  154
##                          UNKNOWN:   59
##
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  327
##  BLACK                         :16869
##  BLACK HISPANIC                : 2245
##  UNKNOWN                       :   65
##  WHITE                         :  620
##  WHITE HISPANIC                : 3450
```

STATISTIAL_MURDER_FLAG will the independent variable for this project. If shooting resulted in murder, it will have the value 1 (0 if no murder occurred). Then, this variable is converted as factor.

```
df$STATISTICAL_MURDER_FLAG[which(df$STATISTICAL_MURDER_FLAG == 'FALSE')] <- 0
df$STATISTICAL_MURDER_FLAG[which(df$STATISTICAL_MURDER_FLAG == 'TRUE')] <- 1
df$STATISTICAL_MURDER_FLAG <- as.factor(df$STATISTICAL_MURDER_FLAG)
```

The data looks good to start modeling. I've decided to just use the following variables for my analysis: BORO, VIC_AGE_GROUP, VIC_SEX, VIC_RACE. Does the victim information and location predict the outcome of the shooting?

```
summary(df)
```

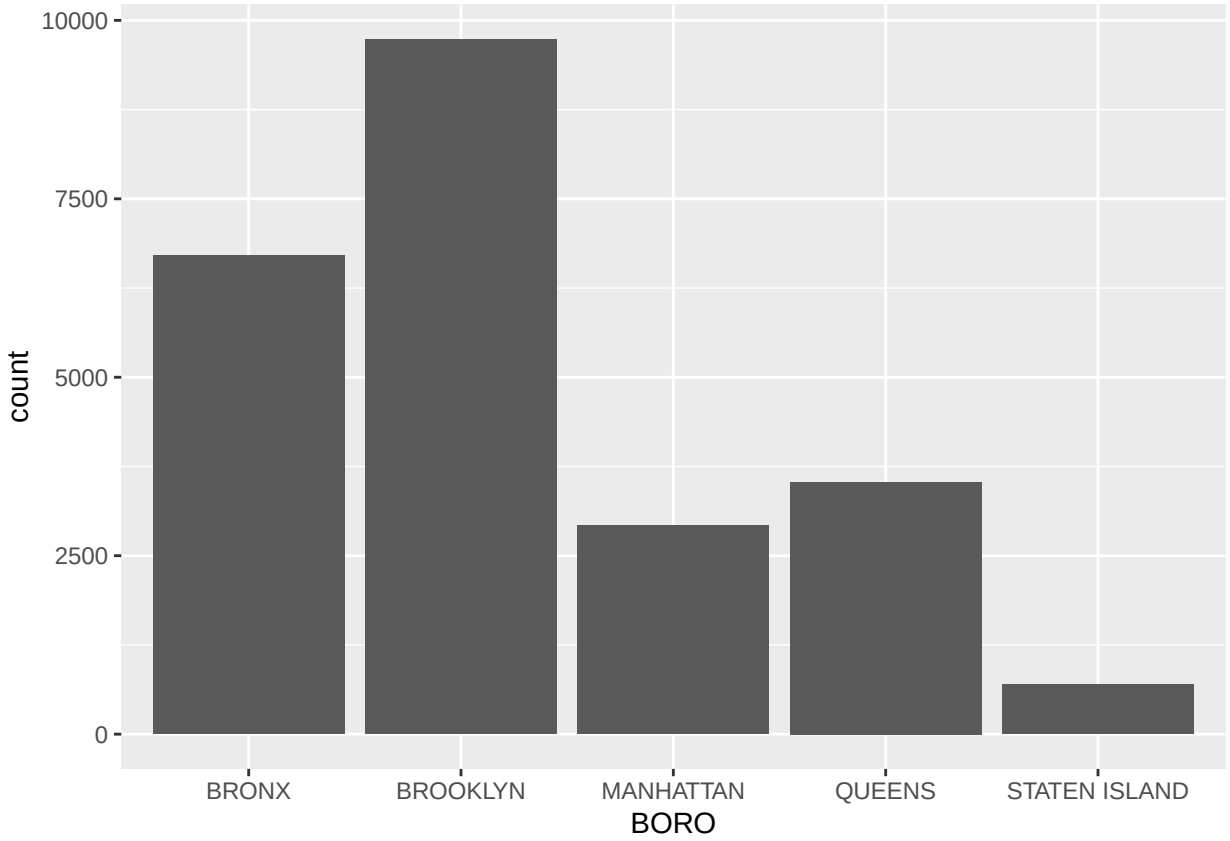
```
##    OCCUR_DATE          OCCUR_TIME                  BORO         PRECINCT
##  Min.   :2006-01-01   Length:23585      BRONX        :6701   75     : 1375
##  1st Qu.:2008-12-31   Class1:hms        BROOKLYN     :9734   73     : 1284
##  Median :2012-02-27   Class2:difftime   MANHATTAN    :2922   67     : 1101
##  Mean   :2012-10-05   Mode  :numeric    QUEENS       :3532   79     :  921
##  3rd Qu.:2016-03-02                     STATEN ISLAND: 696   44     :  841
##  Max.   :2020-12-31                                          47     :  818
##                                                              (Other):17245
##  STATISTICAL_MURDER_FLAG VIC_AGE_GROUP   VIC_SEX
##  0:19085                 <18    : 2525   F: 2204
##  1: 4500                 18-24  : 9003   M:21370
##                          25-44  :10303   U:   11
##                          45-64  : 1541
##                          65+    :  154
##                          UNKNOWN:   59
##
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  327
##  BLACK                         :16869
##  BLACK HISPANIC                : 2245
##  UNKNOWN                       :   65
##  WHITE                         :  620
##  WHITE HISPANIC                : 3450
```

Check to make sure there is no missing data. This dataset does not, but if it did, we would have to fill in or delete missing values.

```
colSums(is.na(df))
```

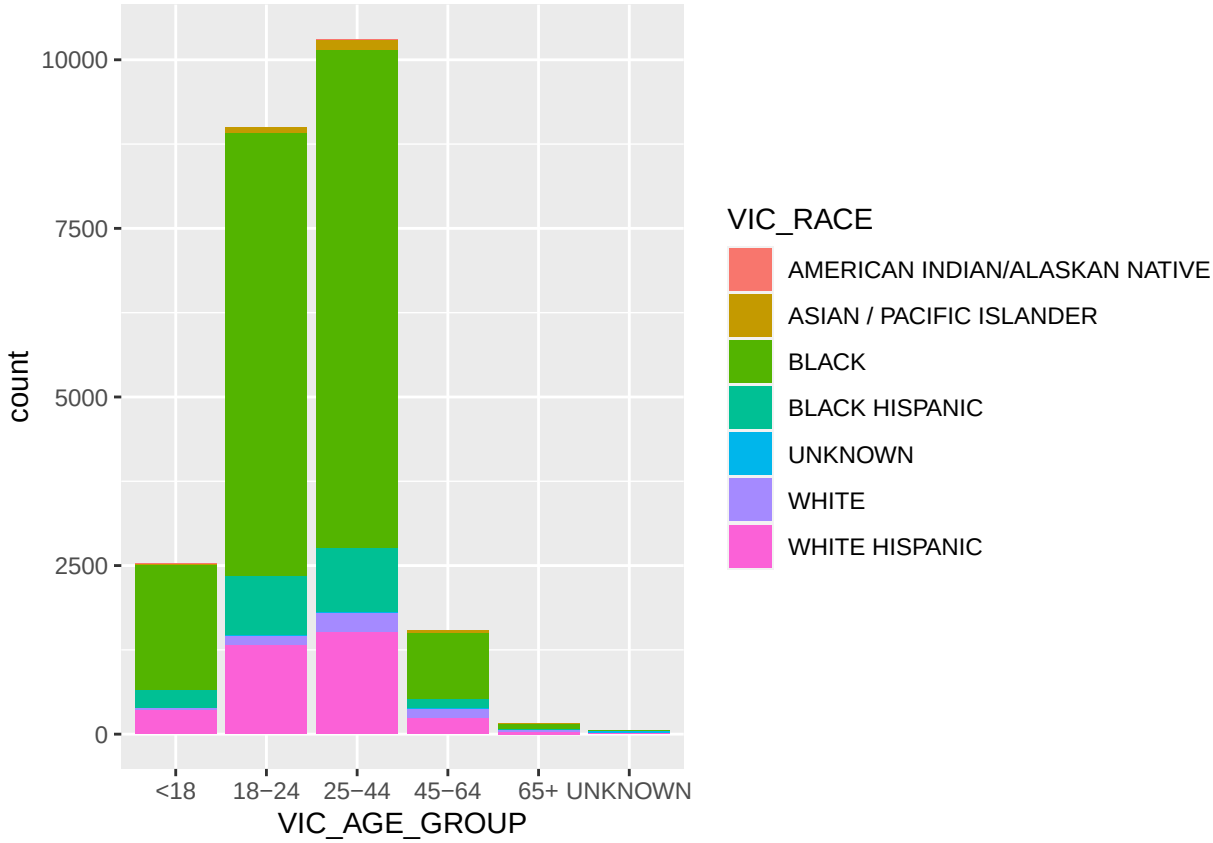
```
##              OCCUR_DATE              OCCUR_TIME                    BORO
##                       0                       0                       0
##                PRECINCT STATISTICAL_MURDER_FLAG           VIC_AGE_GROUP
##                       0                       0                       0
##                 VIC_SEX               VIC_RACE
##                       0                       0
```

There are no missing data. Let's visualize some data.



Simple visualization of BORO (location where shooting occured) and the count of shootings at each location. While the data shows Brooklyn had the most shootings and Staten Island had the least, this graph is misleading since the count isn't in relation to the population density. A better analysis would be count/per certain number of people in population (for example, count/1000 people).

```
graph2<- ggplot(df, aes(x=VIC_AGE_GROUP, fill=VIC_RACE)) +    geom_bar()
graph2
```



This graph shows that most victims were in age groups 18-24 and 25-44. The majority of the victims were black and the least is American Indian/Alaskan Native but most races are seen across most groups. Again, this graph is could be improved with information of the general population's race percentages/count. For example, the most likely reason American Indian/Alaskan Native has so few victims is likely due to the low percentage of these individuals in total population. The same logic applies to age groups. It would be also interesting to see if behavior among age groups groups varies (such as more people between ages 18-44 are out late at night).

```
#Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample(c(TRUE, FALSE), nrow(df), replace=TRUE, prob=c(0.7,0.3))
train <- df[sample, ]
test <- df[!sample, ]
```

```
set.seed(100)
model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP + VIC_SEX + VIC_RACE,  data = train, family
```

A basic model as STATISTICAL_MURDER_FLAG as y variable and boro, vic_age_group, vic_sex, and vic_race as x is performed.

```
summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + VIC_AGE_GROUP +
##     VIC_SEX + VIC_RACE, family = binomial, data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0695  -0.6943  -0.5899  -0.5353   2.3408
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -12.840117 121.903849  -0.105  0.91611
## BOROBROOKLYN                       0.008423   0.050402   0.167  0.86728
## BOROMANHATTAN                     -0.108963   0.069250  -1.573  0.11561
## BOROQUEENS                        -0.005429   0.064470  -0.084  0.93289
## BOROSTATEN ISLAND                 -0.052062   0.122345  -0.426  0.67044
## VIC_AGE_GROUP18-24                 0.201781   0.078043   2.585  0.00972 **
## VIC_AGE_GROUP25-44                 0.570794   0.075726   7.538 4.79e-14 ***
## VIC_AGE_GROUP45-64                 0.703972   0.100145   7.029 2.07e-12 ***
## VIC_AGE_GROUP65+                   1.076447   0.208967   5.151 2.59e-07 ***
## VIC_AGE_GROUPUNKNOWN               0.772612   0.391095   1.976  0.04821 *
## VIC_SEXM                          -0.055960   0.067249  -0.832  0.40534
## VIC_SEXU                          -0.249466   1.111832  -0.224  0.82247
## VIC_RACEASIAN / PACIFIC ISLANDER  11.375624 121.903914   0.093  0.92565
## VIC_RACEBLACK                     11.025476 121.903823   0.090  0.92793
## VIC_RACEBLACK HISPANIC            10.816912 121.903842   0.089  0.92929
## VIC_RACEUNKNOWN                   10.214726 121.904871   0.084  0.93322
## VIC_RACEWHITE                     11.509831 121.903868   0.094  0.92478
## VIC_RACEWHITE HISPANIC            11.210571 121.903832   0.092  0.92673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16127  on 16583  degrees of freedom
## Residual deviance: 15920  on 16566  degrees of freedom
## AIC: 15956
##
## Number of Fisher Scoring iterations: 11
```

The results show that the age group of the victims was statistically significant. Race, sex, and boro was not statistically significant in predicting murder.

This model has several issues/biases.

1. There might be a strong correlation between some of these variables that might strongly affect the results (I haven't done a correlation analysis).
2. I picked the variables I wanted to look into (victim information and boro) because I wanted to see which of these variables affect the outcome. However, for a real analysis, I would look at each individual potential x variable to see if it is worth looking into and perform more data transformations.
3. There's a lot of bias, starting with the choices I've made as x variables, questions I wanted answered, to which visualizations I've selected to include in this project.
4. With a basic logistic linear model, I would have ideally check each individual variable to y outcome, and combined effects of variables before I build a final model.
5. Ideally, I would run several models to compare to this one and select the best one.
6. Another possible bias is that I deleted the perp data and other locational data. It's possible that if I include the data that I've excluded, then the results could change (for example, victim age might be be statistically significant anymore).

The results show that age group is important for the outcome of murder, but not "how". Are older individuals more likely to be murdered? Are there more murders in groups with great % population? It would be interesting to investigate further.

```
predicted <- predict(model, test, type="response")
```

To evaluate how well my model predicts, I used the test set to predict the outcome probability. Then, AUC of the model was evaluated.

```
library(pROC)
auc(test$STATISTICAL_MURDER_FLAG, predicted)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.5732
```

The results of AUC is 0.57. If this value was close to .5, the porbability is close to chance (a value close to 1 would indicate an great predictor model). Based on these results, my model built using the victim and boro information, were not good predictors for the outcome of statistical murder flag. The model performed slighly above random chance. It is important to note that this model has a lot of potential concerns to think about. Here are some that comes to mind: 1. Were there enough deaths to build an accurate model? Most of the outcome was 0 (19085=0 vs. 45000=1). There might've not been enough 0 values to generate an accurate model. It would be usefult to look into this further. One possible solution could be placing more weight on the value of interest (1).
2. Would more data manipulations, such as polynomials, affected the outcome? 3. What variables could we have included into this dataset that might improve the model? Maybe victim's home location? Crime rates at each boro location? number of shots the victim had? and so on. 4. How would machine learning algorithms perform? Building multiple models and comparing would be ideal and interesting to explore.