NYPD Dataset

Packages Needed

- Tidyverse
- Lubridate

```
library(tidyverse)
library(lubridate)
```

Importing the data

First, I'll import the data from https://catalog.data.gov/dataset. This data represents information about every shooting incident in New York City since 2006.

```
url_nypd <- paste0("https://data.cityofnewyork.us/api/views/833y-fsy8/",
              "rows.csv?accessType=DOWNLOAD")

nypd_shootings <- read_csv(url_nypd)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Tidying Data

Looking at the column details, I can see some columns are not the correct variable types. Therefore, I will make the following changes

- *Incident_Key* is listed as a double type
  - I want to change this to a character string type since it is a unique label for each incident.
- *Occur_Date* is listed as a string/character type
  - This needs to change to a date column using the **lubridate** package
- The following variables will need to be changed to a factor type because they are categorical
  - *BORO*
  - *JURISDICTION_CODE*
  - *PERP_AGE_GROUP*
  - *PERP_SEX*
  - *PERP_RACE*
  - *VIC_AGE_GROUP*
  - *VIC_SEX*
  - *VIC_RACE*
- I'm also removing a few variableS that I don't feel have as much impact to the analysis. INCIDENT_KEY would be important if we were joinging multiple datasets. In this case, we aren't; therefore, I am removing it along with the geographical data.

```
nypd_shootings <- nypd_shootings %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(-c(INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

factor_cols <- c("BORO", "JURISDICTION_CODE", "PERP_AGE_GROUP", "PERP_SEX",
                 "PERP_RACE", "VIC_AGE_GROUP", "VIC_SEX", "VIC_RACE")
nypd_shootings[, factor_cols] <- nypd_shootings[, factor_cols] %>%
  lapply(factor)
```

Viewing the summary, we can see that about of a third of the PERP_AGE_GROUP, PERP_SEX, AND PER_RACE are missing. Also, PERP_SEX and PERP_RACE are heavily skewed towards a specific factor. Therefore, I am dropping all three variables. If we had access to more data, I could probably fill the missing data using various methods. Also, JURISDICTION_CODE only has two observations where the data is missing, I will fill them with a random number between 0 and 2.

```
summary(nypd_shootings)
```

```
##    OCCUR_DATE          OCCUR_TIME                     BORO          PRECINCT
##  Min.   :2006-01-01   Length:23568      BRONX        :6700   Min.   :  1.00
##  1st Qu.:2008-12-30   Class1:hms        BROOKLYN     :9722   1st Qu.: 44.00
##  Median :2012-02-26   Class2:difftime   MANHATTAN    :2921   Median : 69.00
##  Mean   :2012-10-03   Mode  :numeric    QUEENS       :3527   Mean   : 66.21
##  3rd Qu.:2016-02-28                     STATEN ISLAND: 698   3rd Qu.: 81.00
##  Max.   :2020-12-31                                          Max.   :123.00
##
##  JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  0   :19624        Length:23568       Mode :logical           18-24  :5448
##  1   :   54        Class :character   FALSE:19080             25-44  :4613
##  2   : 3888        Mode  :character   TRUE :4488              UNKNOWN:3156
##  NA's:    2                                                   <18    :1354
##                                                               45-64  : 481
##                                                               (Other):  57
##                                                               NA's   :8459
##  PERP_SEX              PERP_RACE    VIC_AGE_GROUP   VIC_SEX
##  F   :  334   BLACK         :9855   <18    : 2525   F: 2195
##  M   :13305   WHITE HISPANIC:1961   18-24  : 9000   M:21353
##  U   : 1504   UNKNOWN       :1869   25-44  :10287   U:   20
##  NA's: 8425   BLACK HISPANIC:1081   45-64  : 1536
##               WHITE         : 255   65+    :  155
##               (Other)       : 122   UNKNOWN:   65
##               NA's          :8425
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  320
##  BLACK                         :16846
##  BLACK HISPANIC                : 2244
##  UNKNOWN                       :  102
##  WHITE                         :  615
##  WHITE HISPANIC                : 3432
```

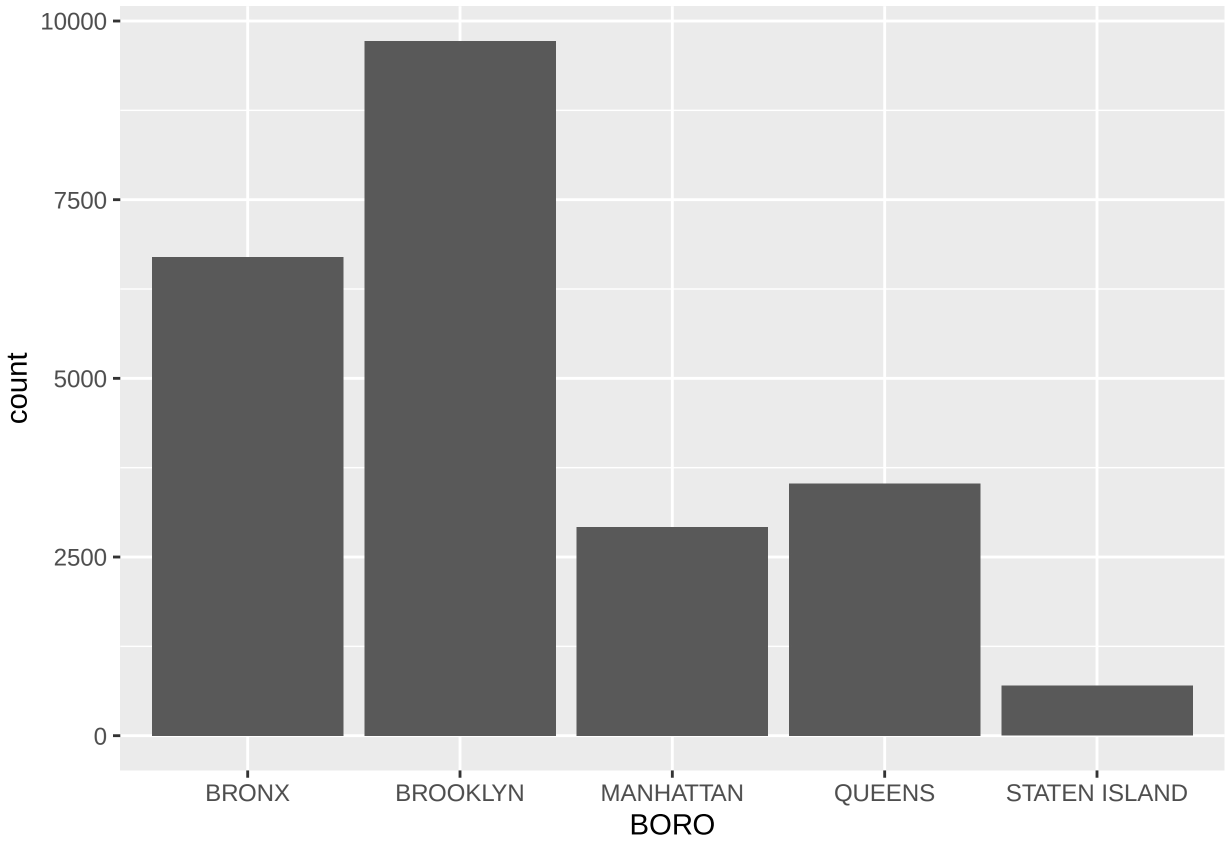
```
nypd_shootings <- nypd_shootings %>%
  mutate(JURISDICTION_CODE = replace(JURISDICTION_CODE, is.na(JURISDICTION_CODE)
                                     , sample(0:2, 1))) %>%
  select(-c(PERP_AGE_GROUP, PERP_SEX, PERP_RACE))

summary(nypd_shootings)
```

```
##    OCCUR_DATE          OCCUR_TIME                  BORO          PRECINCT
##  Min.   :2006-01-01   Length:23568       BRONX        :6700   Min.   :  1.00
##  1st Qu.:2008-12-30   Class1:hms         BROOKLYN     :9722   1st Qu.: 44.00
##  Median :2012-02-26   Class2:difftime    MANHATTAN    :2921   Median : 69.00
##  Mean   :2012-10-03   Mode  :numeric     QUEENS       :3527   Mean   : 66.21
##  3rd Qu.:2016-02-28                      STATEN ISLAND: 698   3rd Qu.: 81.00
##  Max.   :2020-12-31                                           Max.   :123.00
##
##  JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG VIC_AGE_GROUP
##  0:19624           Length:23568       Mode :logical           <18    : 2525
##  1:   56           Class :character   FALSE:19080             18-24  : 9000
##  2: 3888           Mode  :character   TRUE :4488              25-44  :10287
##                                                               45-64  : 1536
##                                                               65+    :  155
##                                                               UNKNOWN:   65
##
##  VIC_SEX                             VIC_RACE
##  F: 2195   AMERICAN INDIAN/ALASKAN NATIVE:    9
##  M:21353   ASIAN / PACIFIC ISLANDER      :  320
##  U:   20   BLACK                         :16846
##            BLACK HISPANIC                : 2244
##            UNKNOWN                       :  102
##            WHITE                         :  615
##            WHITE HISPANIC                : 3432
```

Visualization and Analysis

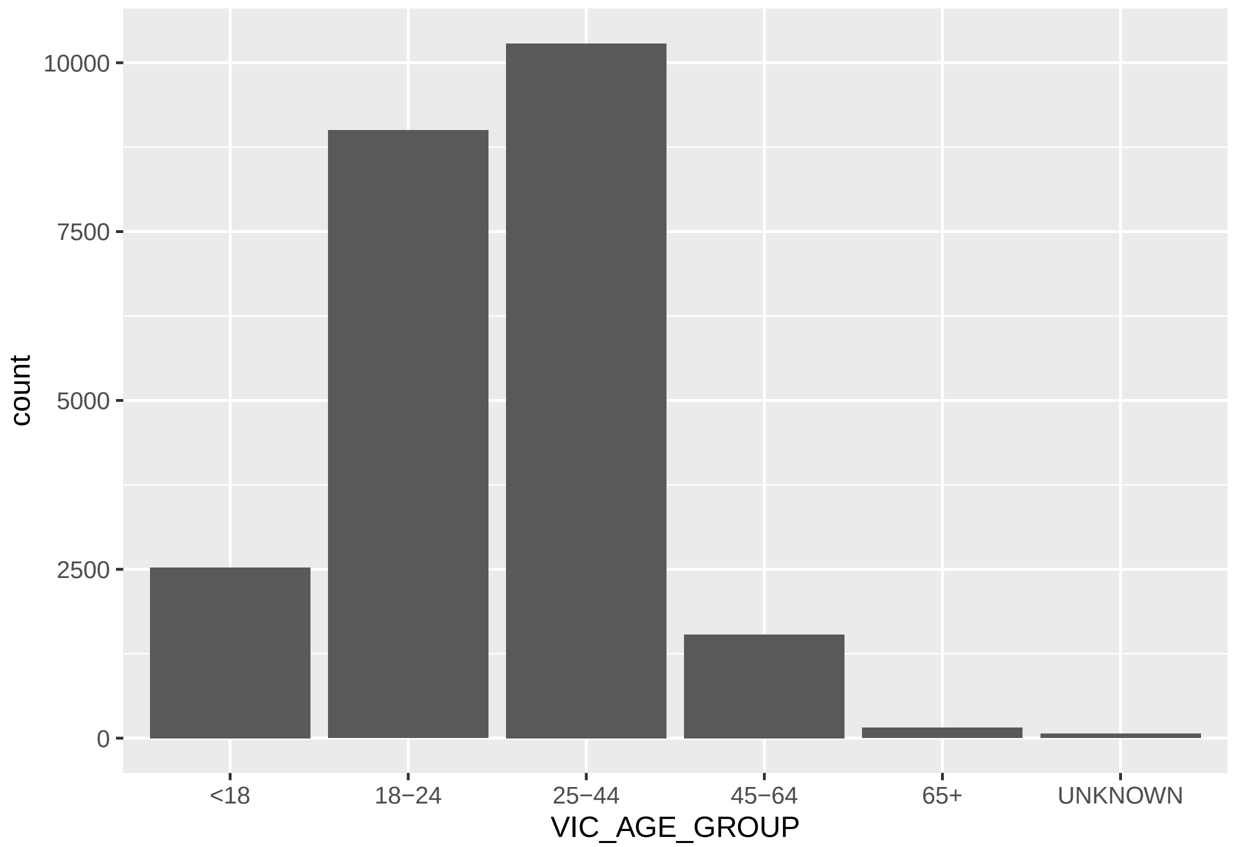
First, it would be a good idea to view the total amount of shootings between all of the boroughs. We can see that there are a higher number of shootings between Brooklyn and the Bronx compared to Manhattan, Queens and Staten Island. Bronx and Brooklyn are known for being the poorest boroughs in the area. Could wealth be related to the number of incidents?

```
ggplot(nypd_shootings, aes(x=BORO)) + geom_bar()
```



Next, we can see that a majority of the victims of the shooting are between the ages of 18-44. This is questions whether these incidents may be gang related. Are these incidents happening due low-income parents looking for resources? Further investigation into the economy may help answere those questions. Looking back at the summary, why are most of these incidents involving harm to african american males?

```
ggplot(nypd_shootings, aes(x=VIC_AGE_GROUP)) +
  geom_bar()
```



Data Modeling

I would like to implement a binary logistic model to see whether a model can help predict if a shooting was a statistical murder or not. As we can see, shootings with a jurisdiction of 2 has a negative, significant correlation on if the shooting was a murder or not. However, we can see that when a victim that is 65 years or older is a victim, there is almost no affect on whether it was a murder or not. Could the reason for this be because they are not seen as a threat compared to the younger generations? Also, according to the model, Manhattan shootings are less likely to be considered a murder compared to the other boroughs.

```
nypd_model_data <- nypd_shootings %>%
select(-c("OCCUR_DATE", "OCCUR_TIME", "LOCATION_DESC"))
```

```
logit_1 <- glm(STATISTICAL_MURDER_FLAG ~., family = binomial, data=nypd_model_data)

summary(logit_1)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ ., family = binomial,
##     data = nypd_model_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0347  -0.7036  -0.6117  -0.5282   2.5362
```

```
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.290e+01  1.075e+02  -0.120  0.90445
## BOROBROOKLYN                      5.590e-02  9.403e-02   0.594  0.55222
## BOROMANHATTAN                    -6.832e-02  8.063e-02  -0.847  0.39676
## BOROQUEENS                        3.713e-02  1.853e-01   0.200  0.84123
## BOROSTATEN ISLAND                 6.807e-02  2.374e-01   0.287  0.77427
## PRECINCT                         -3.089e-04  2.837e-03  -0.109  0.91328
## JURISDICTION_CODE1                2.870e-02  3.298e-01   0.087  0.93065
## JURISDICTION_CODE2               -2.591e-01  4.854e-02  -5.337 9.46e-08 ***
## VIC_AGE_GROUP18-24                2.812e-01  6.664e-02   4.220 2.44e-05 ***
## VIC_AGE_GROUP25-44                6.375e-01  6.480e-02   9.839  < 2e-16 ***
## VIC_AGE_GROUP45-64                7.881e-01  8.470e-02   9.304  < 2e-16 ***
## VIC_AGE_GROUP65+                  1.136e+00  1.821e-01   6.238 4.44e-10 ***
## VIC_AGE_GROUPUNKNOWN              8.538e-01  3.070e-01   2.781  0.00542 **
## VIC_SEXM                         -2.678e-02  5.742e-02  -0.466  0.64092
## VIC_SEXU                         -1.514e+00  1.045e+00  -1.448  0.14771
## VIC_RACEASIAN / PACIFIC ISLANDER  1.130e+01  1.075e+02   0.105  0.91629
## VIC_RACEBLACK                     1.104e+01  1.075e+02   0.103  0.91823
## VIC_RACEBLACK HISPANIC            1.083e+01  1.075e+02   0.101  0.91976
## VIC_RACEUNKNOWN                   1.088e+01  1.075e+02   0.101  0.91936
## VIC_RACEWHITE                     1.142e+01  1.075e+02   0.106  0.91542
## VIC_RACEWHITE HISPANIC            1.117e+01  1.075e+02   0.104  0.91723
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22948  on 23567  degrees of freedom
## Residual deviance: 22628  on 23547  degrees of freedom
## AIC: 22670
##
## Number of Fisher Scoring iterations: 11
```

Conclusion

After reviewing the data, it seems as much of the shooting incidents occur in the low income boroughs of New York City. We may need to stay aware that these are shooting incidents that were investigated by New York. It’s possible that New York is more involved policing the lower income areas of the city. Some incidents in Staten Island, Queens and Manhattan may not have been recorded. Also, I may have exhibited some bias as to how I presented the data in this case. I chose to review the income inequality and how it compares to age group. As someone who was raised in a low-income household, this was a very interesting point for me to review. Also, the model I developed had a lot of bias as well. I didn’t check for correlation between variables nor tune the model at all. This can lead to significant issues within the model and possible statistical errors. As I learn more about the models, I can develop skills to reduce these biases!