NYPD Shooting Analysis

JS

2022-10-07

NYPD Shooting Analysis

I have decided to focus on analyzing trends involving shooting age of victims and if it is impacted by the borough the victim is in. I'm interested in looking at where shootings are most likely to occur and who they are most likely to occur based on those factors.

Please see below steps taken to upload, clean and analyze data for the week 3 assignment on data regarding shootings in New York City.

```
## include libraries required and setup knit

knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(lubridate)
```

Initially we will import the data set. This is obtained from the NYPD via Data.gov and covers the years from 2006 to present. A link to the data page is available at:

https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

```
## import data from source

url1 <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shootdata <- read_csv(url1)
```

```
## Rows: 25596 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.
```

Inspect summary of data for future cleaning and observe column titles/types.

```
## display summary of imported data

summary(shootdata)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  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:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  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
##
```

```
str(shootdata)
```

```
## spec_tbl_df [25,596 x 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY          : num [1:25596] 2.36e+08 2.31e+08 2.31e+08 2.38e+08 2.24e+08 ...
##  $ OCCUR_DATE            : chr [1:25596] "11/11/2021" "07/16/2021" "07/11/2021" "12/11/2021" ...
##  $ OCCUR_TIME            : 'hms' num [1:25596] 15:04:00 22:05:00 01:09:00 13:42:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ BORO                  : chr [1:25596] "BROOKLYN" "BROOKLYN" "BROOKLYN" "BROOKLYN" ...
##  $ PRECINCT              : num [1:25596] 79 72 79 81 113 113 42 52 34 75 ...
##  $ JURISDICTION_CODE     : num [1:25596] 0 0 0 0 0 0 0 0 0 0 ...
##  $ LOCATION_DESC         : chr [1:25596] NA NA NA NA ...
##  $ STATISTICAL_MURDER_FLAG: logi [1:25596] FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ PERP_AGE_GROUP        : chr [1:25596] NA "45-64" "<18" NA ...
```

```
##  $ PERP_SEX               : chr [1:25596] NA "M" "M" NA ...
##  $ PERP_RACE              : chr [1:25596] NA "ASIAN / PACIFIC ISLANDER" "BLACK" NA ...
##  $ VIC_AGE_GROUP          : chr [1:25596] "18-24" "25-44" "25-44" "25-44" ...
##  $ VIC_SEX                : chr [1:25596] "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr [1:25596] "BLACK" "ASIAN / PACIFIC ISLANDER" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num [1:25596] 996313 981845 996546 1001139 1050710 ...
##  $ Y_COORD_CD             : num [1:25596] 187499 171118 187436 192775 184826 ...
##  $ Latitude               : num [1:25596] 40.7 40.6 40.7 40.7 40.7 ...
##  $ Longitude              : num [1:25596] -74 -74 -74 -73.9 -73.8 ...
##  $ Lon_Lat                : chr [1:25596] "POINT (-73.95650899099996 40.68131820000008)" "POINT (-74
##  - attr(*, "spec")=
##   .. 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_double(),
##   ..   Y_COORD_CD = col_double(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   Lon_Lat = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

After looking through the available data we will be able to remove all columns except for VIC_AGE_GROUP, BORO, and OCCUR_DATE as well as alter OCCUR_DATE to the date format. Additionally convert data as required.

```
# select only columns required
# filter out unknown age groups


cleanShootData <- shootdata %>%
    select(c(BORO, VIC_AGE_GROUP, OCCUR_DATE)) %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    mutate(BORO = factor(BORO)) %>%
    mutate(VIC_AGE_GROUP = factor(VIC_AGE_GROUP))

cleanShootData <- cleanShootData %>%
    select(c(BORO, VIC_AGE_GROUP, OCCUR_DATE))

summary(cleanShootData)
```

```
##             BORO       VIC_AGE_GROUP     OCCUR_DATE
```

```
##  BRONX        : 7402   <18    : 2681   Min.   :2006-01-01
##  BROOKLYN     :10365   18-24  : 9604   1st Qu.:2009-05-10
##  MANHATTAN    : 3265   25-44  :11386   Median :2012-08-26
##  QUEENS       : 3828   45-64  : 1698   Mean   :2013-06-13
##  STATEN ISLAND:  736   65+    :  167   3rd Qu.:2017-07-01
##                        UNKNOWN:   60   Max.   :2021-12-31
```

Noticing a group of unknown ages in the data, I decide the correct course of action would be to drop them from the analysis as they total only roughly one fifth of one percent of the data points. Following such will do one final summary check of the data.

```
#filter out unknown age groups

cleanShootData <- cleanShootData %>%
    filter(VIC_AGE_GROUP!='UNKNOWN')

cleanShootData <- cleanShootData %>%
    select(c(BORO, VIC_AGE_GROUP, OCCUR_DATE))

summary(cleanShootData)
```

```
##             BORO       VIC_AGE_GROUP     OCCUR_DATE
##  BRONX        : 7385   <18    : 2681   Min.   :2006-01-01
##  BROOKLYN     :10339   18-24  : 9604   1st Qu.:2009-05-10
##  MANHATTAN    : 3260   25-44  :11386   Median :2012-08-26
##  QUEENS       : 3817   45-64  : 1698   Mean   :2013-06-14
##  STATEN ISLAND:  735   65+    :  167   3rd Qu.:2017-07-02
##                        UNKNOWN:    0   Max.   :2021-12-31
```

The initial visualization to undertake will simply be a tracking of shootings by year in New York.

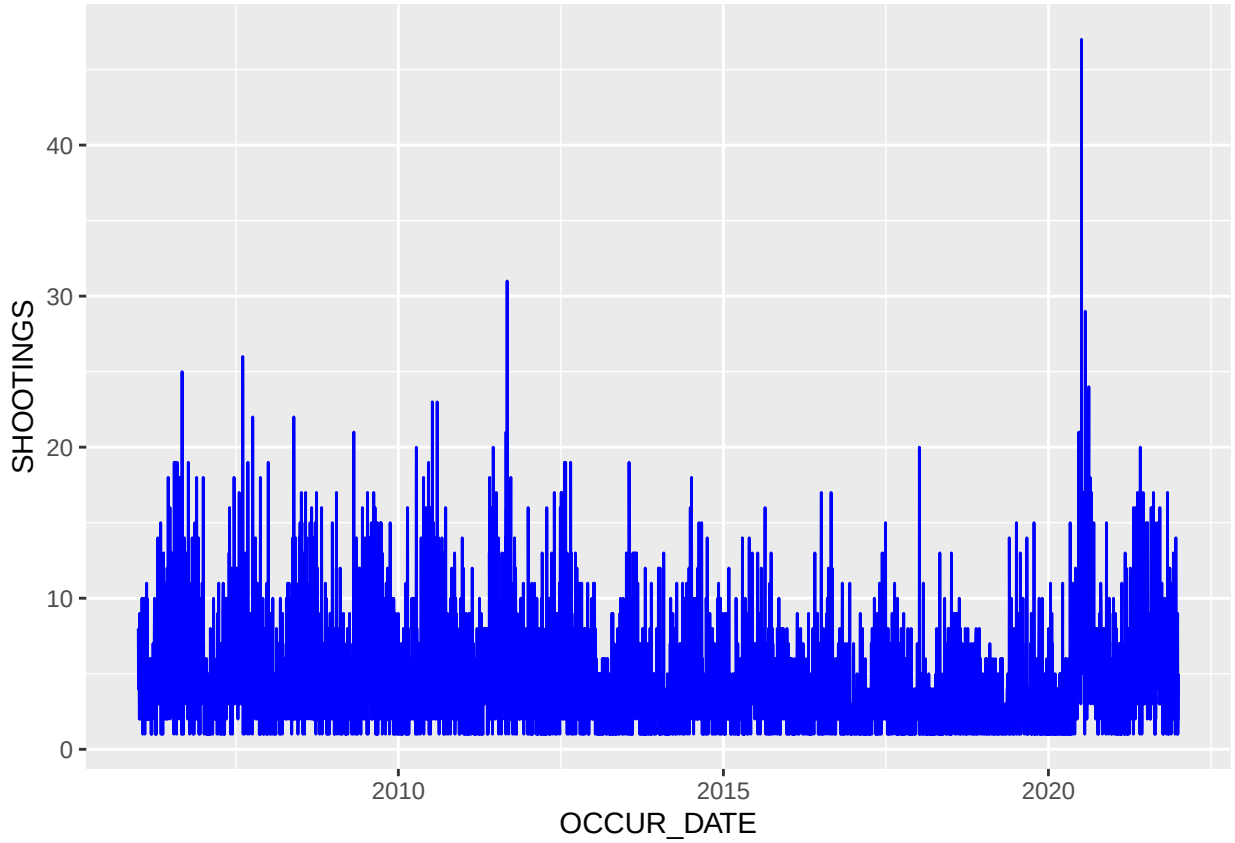
We can see clearly that while shootings were generally trending downwards, they spiked sharply around 2020.

```
VIZ_DATA <- cleanShootData

VIZ_DATA <- VIZ_DATA %>%
    mutate((OCCUR_DATE = year(OCCUR_DATE))) %>%
    group_by(OCCUR_DATE) %>%
    count(name = "SHOOTINGS")

GRAPH_SHOOTING <- VIZ_DATA %>%
    ggplot(aes(x = OCCUR_DATE, y = SHOOTINGS))+geom_line (color = "blue")

GRAPH_SHOOTING
```



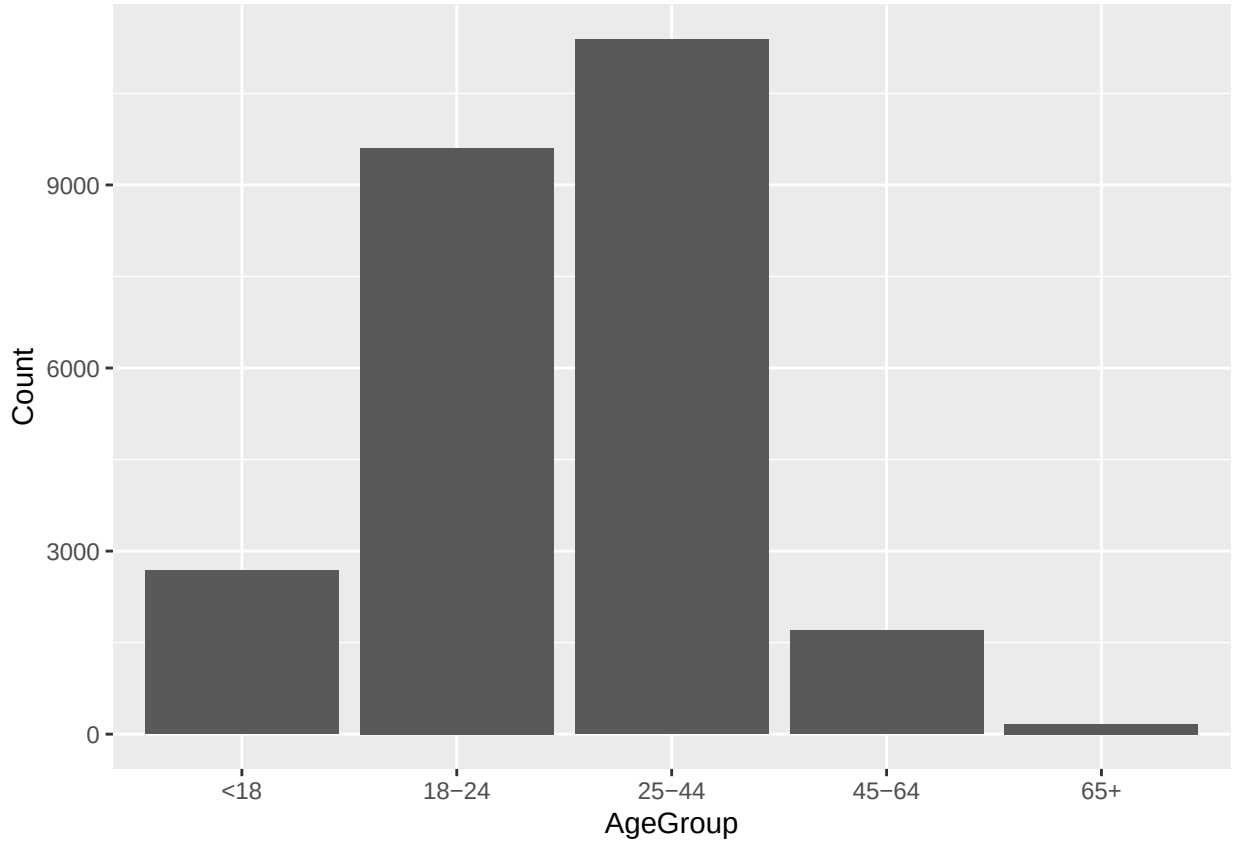
The next visual analysis to undertake will simply be tracking shootings victims by age group. As we see below, young adults (18-44) are far more likely to be the victim of a shooting than other age groups. We must be cautious to draw conclusions though as this does not account for factors such as population. If there are 10x as many of an age bracket, it would be reasonable to assume 10x as many shootings would be plausible.

```
# this is super lazy and should pull the numbers, no copy paste

AGE_FRAME <- data.frame((AgeGroup=c("<18", "18-24", "25-44", "45-64", "65+")),
      Count = c(2681, 9604, 11386, 1698, 167))

AGE_BAR <- ggplot(data=AGE_FRAME, aes(x=AgeGroup, y=Count)) +
        geom_bar(stat="identity")

AGE_BAR
```



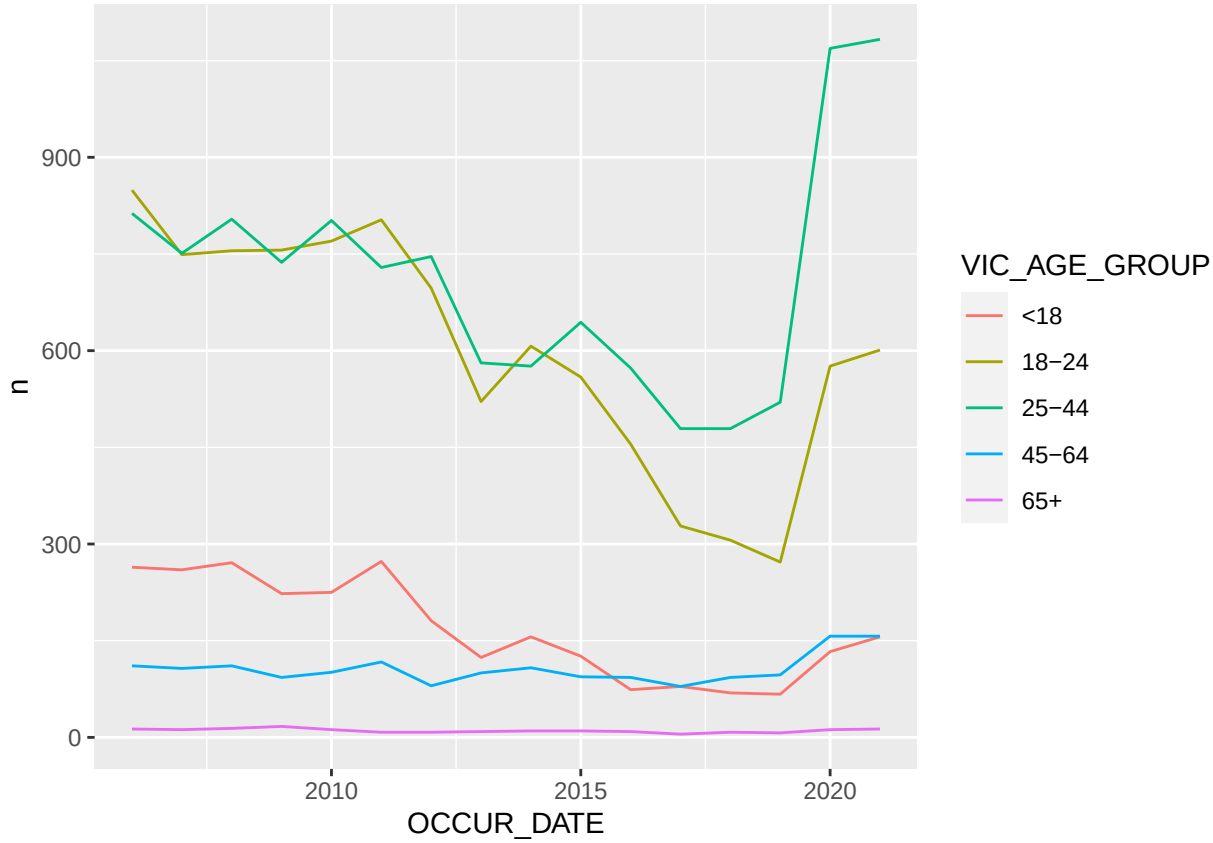
Taking that data, I'd like to find out if murders by age move in step together, so will plot out murders by age group per year.

It certainly appears that the amount of victims shot by age generally move in step with each other, indicating a larger outside factor than age as a cause. It should be noted that this could be misleading as it does not account for population of the age groups.

```
SHOT_BY_TIME <- cleanShootData

SHOT_BY_TIME <- SHOT_BY_TIME %>%
    mutate(OCCUR_DATE = year(OCCUR_DATE)) %>%
    count(OCCUR_DATE,VIC_AGE_GROUP) %>%
    ggplot(mapping = aes(x = OCCUR_DATE, y=n, color = VIC_AGE_GROUP)) + geom_line()

SHOT_BY_TIME
```

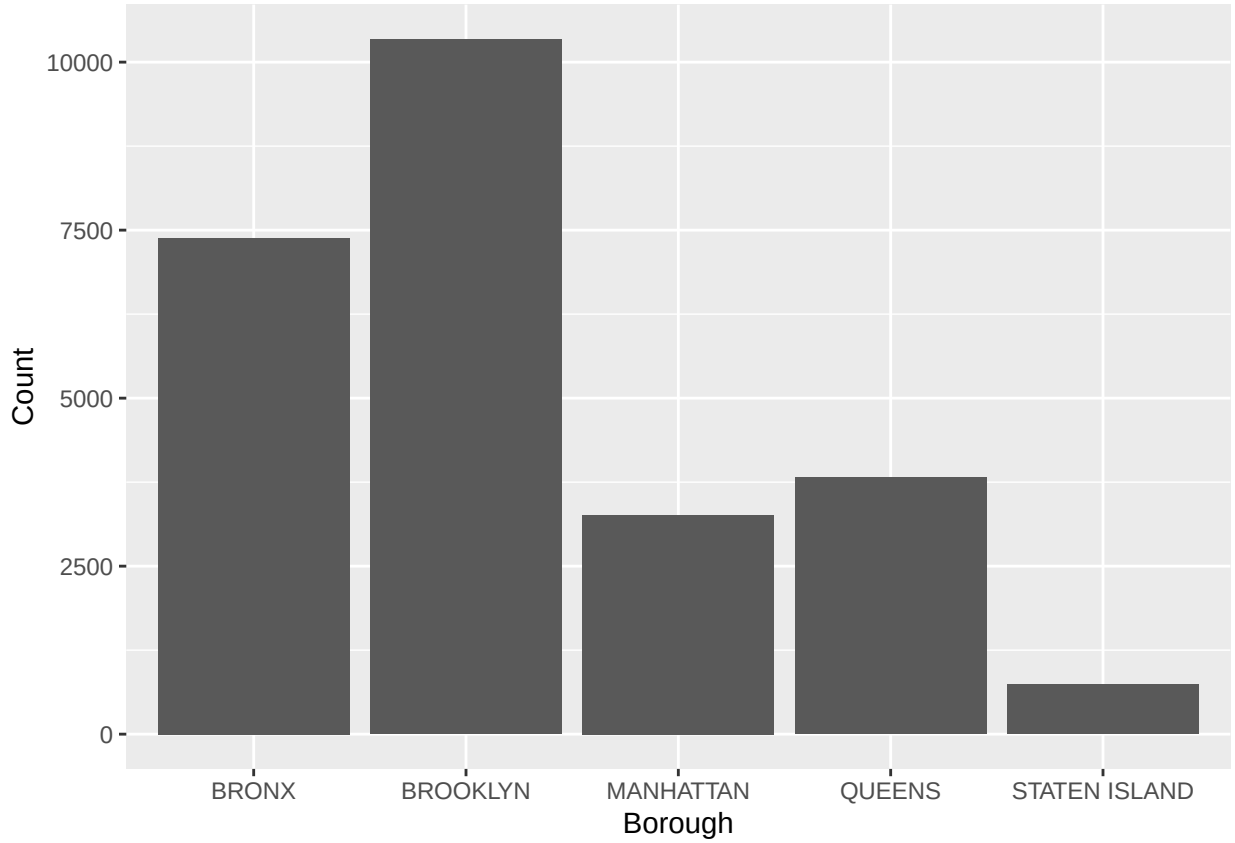


Next we will look at shooting by borough. As with the shootings by age, we can see that certain boroughs such as Brooklyn or Bronx account for a far greater percentage of the shootings as opposed Staten Island or Queens. We must again acknowledge though this could be misleading as it does not account for population of the boroughs.

```
BORO_FRAME <- data.frame((Borough = c("BRONX", "BROOKLYN", "MANHATTAN", "QUEENS", "STATEN ISLAND")),
      Count = c(7385, 10339, 3260, 3817, 735))

BORO_BAR <- ggplot(data=BORO_FRAME, aes(x=Borough, y=Count)) +
        geom_bar(stat="identity")

BORO_BAR
```

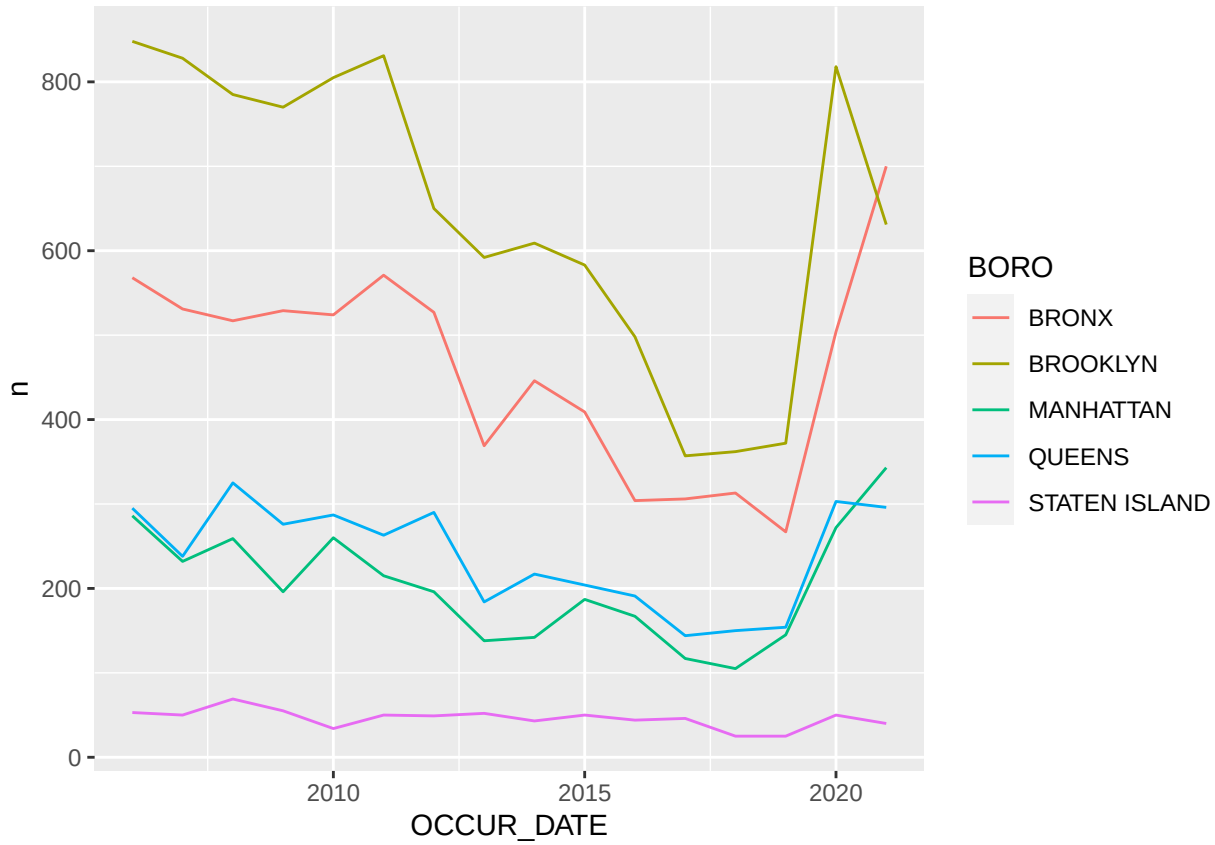


Lastly we'll look at borough shootings by year to see if there is a visual relationship. As we can see, it certainly appears to be that way for most, though Staten Island does not neccessarily move in step.

```
SHOT_BY_BORO <- cleanShootData

SHOT_BY_BORO <- SHOT_BY_BORO %>%
    mutate(OCCUR_DATE = year(OCCUR_DATE)) %>%
    count(OCCUR_DATE,BORO) %>%
    ggplot(mapping = aes(x = OCCUR_DATE, y=n, color = BORO)) + geom_line()

SHOT_BY_BORO
```



Lastly we'll look to model if there is a relationship between shootings in Brooklyn and shootings in Queens and if they could be a predictive factor.

It certainly appears that a relationship is likely judging by an adjust R-squared value of approximately 0.72, however this is far too basic of an overview to draw any conclusions.

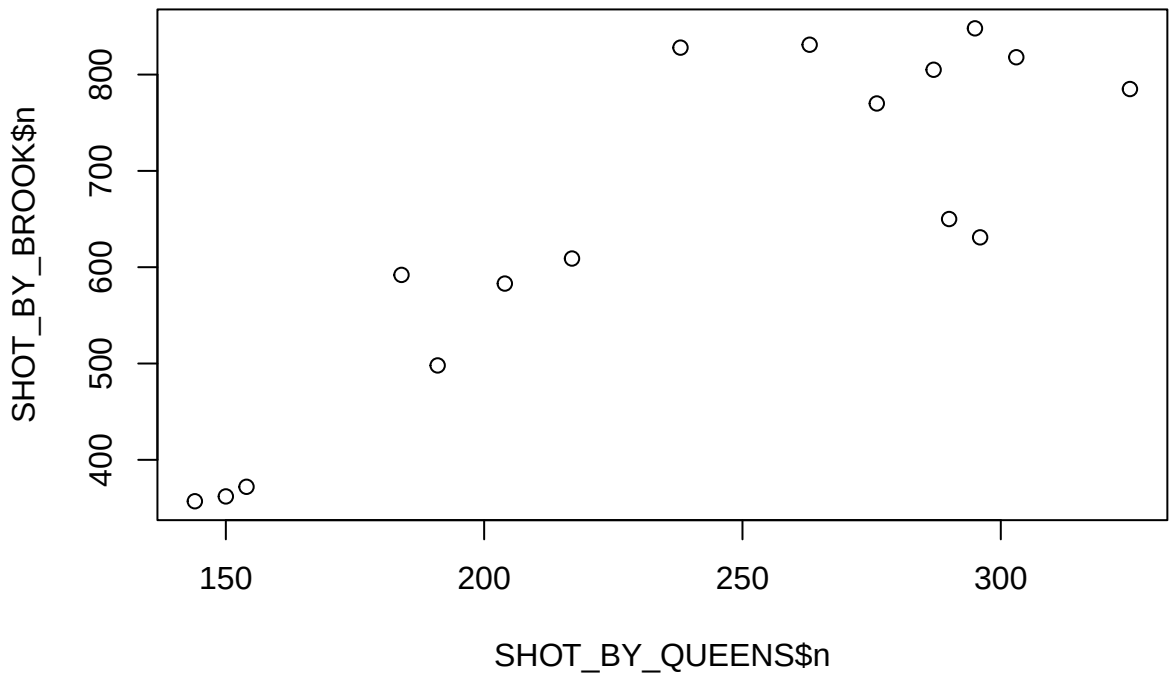
```
SHOT_BY_BROOK <- cleanShootData %>%

mutate(OCCUR_DATE = year(OCCUR_DATE)) %>%
   filter(BORO == "BROOKLYN") %>%
  group_by(OCCUR_DATE) %>%
  tally


SHOT_BY_QUEENS <- cleanShootData %>%

mutate(OCCUR_DATE = year(OCCUR_DATE)) %>%
  filter(BORO == "QUEENS") %>%
  group_by(OCCUR_DATE) %>%
  tally


  plot(SHOT_BY_QUEENS$n, SHOT_BY_BROOK$n)
```




```
mod1 <- lm(SHOT_BY_QUEENS$n ~ SHOT_BY_BROOK$n)


summary(mod1)
```

```
##
## Call:
## lm(formula = SHOT_BY_QUEENS$n ~ SHOT_BY_BROOK$n)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.720 -11.799  -3.549   4.163  61.961
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     46.08004   31.42737   1.466    0.165
## SHOT_BY_BROOK$n  0.29787    0.04702   6.335 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.14 on 14 degrees of freedom
## Multiple R-squared:  0.7414, Adjusted R-squared:  0.7229
## F-statistic: 40.14 on 1 and 14 DF,  p-value: 1.843e-05
```

Bias

A large concern for bias is that I selected what and how I would analyze the data after already inspecting it. Additionally data is not controlled for any factors such as population, a rate analysis would probably be more beneficial in painting an accurate picture. Lastly, a common bias that will occur is the tendency for us in the earlier stages of learning to select methods that we are more comfortable with instead of what may be the best analytic techniques. This can lead to weak observations and conclusions.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.