NYC Shooting Analysis

B McKiernan

5/7/2021

Introduction and Data Source

Gun violence is an unfortunate reality in many of America's largest urban areas. The largest US city, New York City, is no exception. Shootings effect the lives of New York City residents regardless of race, age, or boro of residence. Exploring relationships between the race, age, and boro of shooting incidents in New York City from 2006 to 2020 will help understand the degree to which specific communities of New York City residents are effected by shootings. This study will explore the distribution of age among the race classifications for both perpetrators and victims and will examine potential relationships between the boro in which the incident occurred and total number of shooting incidents. These questions will help shed light into and improve understanding of how shooting incidents effect the residents of New York City.

From the website Data.gov (catalog.data.gov), A search for "NYC Shooting incident data" was made and the data set idenitified. A complete description of the dataset is available on the source website (https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic) and cited below.

"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."

Cleaning and Processing

The data is loaded and an initial summary of the dataset was run for inspection and analysis of variables and types.

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

Examining the variables, a number that were not important to this analysis and were removed. The OCCUR_DATE and OCCUR_TIME variables were combined together and converted it into a POSIXct object for easier use in the time-based analysis later on. It also appears that a sizable number (approximately 12,000) of shooting incidents include unknown or unreported information. For the categorical analysis, rows with unknowns and empty entries were removed for cleaner analysis. This decision was made with some hesitation and caution was exercised in the subsequent analysis. The rows with unknown or empty entries remain included for the quantitative analysis to ensure proper counts of total shooting incidents. Certain variables were converted to factors to aid in potential further analysis. A summary for both versions of the initial data set was examined and the effect of removing entries noted.

```
#Remove unneeded columns
shooting_data <- shooting_data %>%
     select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, INCIDENT_KEY,
               PRECINCT, JURISDICTION_CODE))

#Change Date and Time variables to date format
shooting_data <- shooting_data %>%
     mutate(OCCUR_D_T = str_c(OCCUR_DATE, OCCUR_TIME, sep = ' ')) %>%
     mutate(OCCUR_D_T = mdy_hms(OCCUR_D_T), OCCUR_DATE = mdy(OCCUR_DATE))

#Cleaning for categorical analysis
shooting_cat <- shooting_data %>%
        filter(PERP_AGE_GROUP != "",
               PERP_SEX != " ",
               PERP_RACE != ' ',
               PERP_AGE_GROUP != ' ',
               PERP_AGE_GROUP != 'UNKNOWN',
               PERP_RACE != 'UNKNOWN',
               PERP_AGE_GROUP != 1020,
               PERP_AGE_GROUP != 940,
               PERP_AGE_GROUP != 224,
               VIC_RACE != 'UNKNOWN',
               VIC_AGE_GROUP != 'UNKNOWN')

#Change certain variables to factor variables
shooting_data <- shooting_data %>%
     mutate(BORO = as.factor(BORO), PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP),
            PERP_SEX = as.factor(PERP_SEX), PERP_RACE = as.factor(PERP_RACE),
            VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP), VIC_SEX = as.factor(VIC_SEX),
            VIC_RACE = as.factor(VIC_RACE))
```

```
#Summary of data for Categorical Analysis
summary(shooting_cat)
```

```
##    OCCUR_DATE          OCCUR_TIME            BORO           LOCATION_DESC
##  Min.   :2006-01-01   Length:11686       Length:11686       Length:11686
##  1st Qu.:2008-09-21   Class :character   Class :character   Class :character
##  Median :2011-12-12   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2012-07-22
##  3rd Qu.:2015-12-24
##  Max.   :2020-12-29
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Length:11686            Length:11686       Length:11686
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:11686       Length:11686       Length:11686       Length:11686
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##    OCCUR_D_T
##  Min.   :2006-01-01 02:00:00
##  1st Qu.:2008-09-21 22:18:00
##  Median :2011-12-13 03:35:00
##  Mean   :2012-07-23 08:29:58
##  3rd Qu.:2015-12-25 00:50:30
##  Max.   :2020-12-29 13:15:00
```

```
#Summary of data for Quantitative/Time Analysis
summary(shooting_data)
```

```
##    OCCUR_DATE          OCCUR_TIME                    BORO
##  Min.   :2006-01-01   Length:23568       BRONX        :6700
##  1st Qu.:2008-12-30   Class :character   BROOKLYN     :9722
##  Median :2012-02-26   Mode  :character   MANHATTAN    :2921
##  Mean   :2012-10-03                      QUEENS       :3527
##  3rd Qu.:2016-02-28                      STATEN ISLAND: 698
##  Max.   :2020-12-31
##
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  Length:23568       Length:23568                    :8459    : 8425
##  Class :character   Class :character        18-24    :5448   F:  334
##  Mode  :character   Mode  :character        25-44    :4613   M:13305
##                                             UNKNOWN  :3156   U: 1504
##                                             <18      :1354
##                                             45-64    : 481
##                                             (Other)  :  57
##           PERP_RACE    VIC_AGE_GROUP   VIC_SEX
##  BLACK         :9855   <18    : 2525   F: 2195
##                :8425   18-24  : 9000   M:21353
##  WHITE HISPANIC:1961   25-44  :10287   U:   20
##  UNKNOWN       :1869   45-64  : 1536
##  BLACK HISPANIC:1081   65+    :  155
##  WHITE         : 255   UNKNOWN:   65
##  (Other)       : 122
##                             VIC_RACE       OCCUR_D_T
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   :2006-01-01 02:00:00
##  ASIAN / PACIFIC ISLANDER      :  320   1st Qu.:2008-12-30 04:27:00
##  BLACK                         :16846   Median :2012-02-26 03:35:00
##  BLACK HISPANIC                : 2244   Mean   :2012-10-04 05:23:12
##  UNKNOWN                       :  102   3rd Qu.:2016-02-28 00:01:00
##  WHITE                         :  615   Max.   :2020-12-31 23:45:00
##  WHITE HISPANIC                : 3432
```

Exploratory Analysis

Much information about the demographics of those involved in shooting incidents is available in the dataset. Specifically, age group, race, and boro describe various communities of New York City. Examining their relationship will give insight into how each is effected by shooting incidents. The first analysis performed explored potential associations between the reported race and reported age of the shooting victims and perpetrators.

Age and Race of Victim

```
#Tables summarizing Victim demographics
addmargins(table(shooting_cat$VIC_RACE, shooting_cat$VIC_AGE_GROUP))
```

```
##
##                                   <18 18-24 25-44 45-64   65+   Sum
##   AMERICAN INDIAN/ALASKAN NATIVE    1     1     1     0     0     3
##   ASIAN / PACIFIC ISLANDER          8    60    94    27     2   191
##   BLACK                           951  2948  3430   521    52  7902
##   BLACK HISPANIC                  159   445   497    84     8  1193
##   WHITE                            19    88   177    92    18   394
##   WHITE HISPANIC                  212   737   889   146    19  2003
##   Sum                            1350  4279  5088   870    99 11686
```

```
round(prop.table(table(shooting_cat$VIC_RACE,
                       shooting_cat$VIC_AGE_GROUP), 1), 4)
```

```
##
##                                    <18  18-24  25-44  45-64    65+
##   AMERICAN INDIAN/ALASKAN NATIVE 0.3333 0.3333 0.3333 0.0000 0.0000
##   ASIAN / PACIFIC ISLANDER       0.0419 0.3141 0.4921 0.1414 0.0105
##   BLACK                          0.1203 0.3731 0.4341 0.0659 0.0066
##   BLACK HISPANIC                 0.1333 0.3730 0.4166 0.0704 0.0067
##   WHITE                          0.0482 0.2234 0.4492 0.2335 0.0457
##   WHITE HISPANIC                 0.1058 0.3679 0.4438 0.0729 0.0095
```

The tables above show both a discrepancy between race of victims and dissimilar distributions of age group for each race. It is clear from the first table that individuals who were identified as Black, Black Hispanic, or White Hispanic make up a vast majority of the victims of shooting incidents. The second table shows that in the distribution of age groups of the victims, those reported to be Black, Black Hispanic, and White Hispanic had higher likelihoods of being victims if they were in the younger age groups (<18 and 18-24). It is notable that those victims reported as White or Asian/Pacific Islander have a slightly different distribution of age groups with a higher percentage of victims in the 24-44 and 45-64 age group and lower percentage in the <18 and 18-24 age groups. Across all race designations, the 25-44 age group consistently has the highest percentage of victims with each designation. A disproportionate number of identified victims tend to be identified from black and brown communities and in younger age groups.

Age and Race of Perpetrator

```
#Tables summarizing Perpetrator demographics
addmargins(table(shooting_cat$PERP_RACE, shooting_cat$PERP_AGE_GROUP))
```

```
##
##                                  <18 18-24 25-44 45-64   65+   Sum
##   AMERICAN INDIAN/ALASKAN NATIVE    0     1     1     0     0     2
##   ASIAN / PACIFIC ISLANDER         11    33    58     5     0   107
##   BLACK                           995  3907  3321   297    25  8545
##   BLACK HISPANIC                  110   502   325    38     5   980
##   WHITE                             7    40   127    51    16   241
##   WHITE HISPANIC                  207   861   662    74     7  1811
##   Sum                            1330  5344  4494   465    53 11686
```

```
round(prop.table(table(shooting_cat$PERP_RACE,
                       shooting_cat$PERP_AGE_GROUP), 1), 4)
```

```
##
##                                    <18  18-24  25-44  45-64    65+
##   AMERICAN INDIAN/ALASKAN NATIVE 0.0000 0.5000 0.5000 0.0000 0.0000
##   ASIAN / PACIFIC ISLANDER       0.1028 0.3084 0.5421 0.0467 0.0000
##   BLACK                          0.1164 0.4572 0.3886 0.0348 0.0029
##   BLACK HISPANIC                 0.1122 0.5122 0.3316 0.0388 0.0051
##   WHITE                          0.0290 0.1660 0.5270 0.2116 0.0664
##   WHITE HISPANIC                 0.1143 0.4754 0.3655 0.0409 0.0039
```

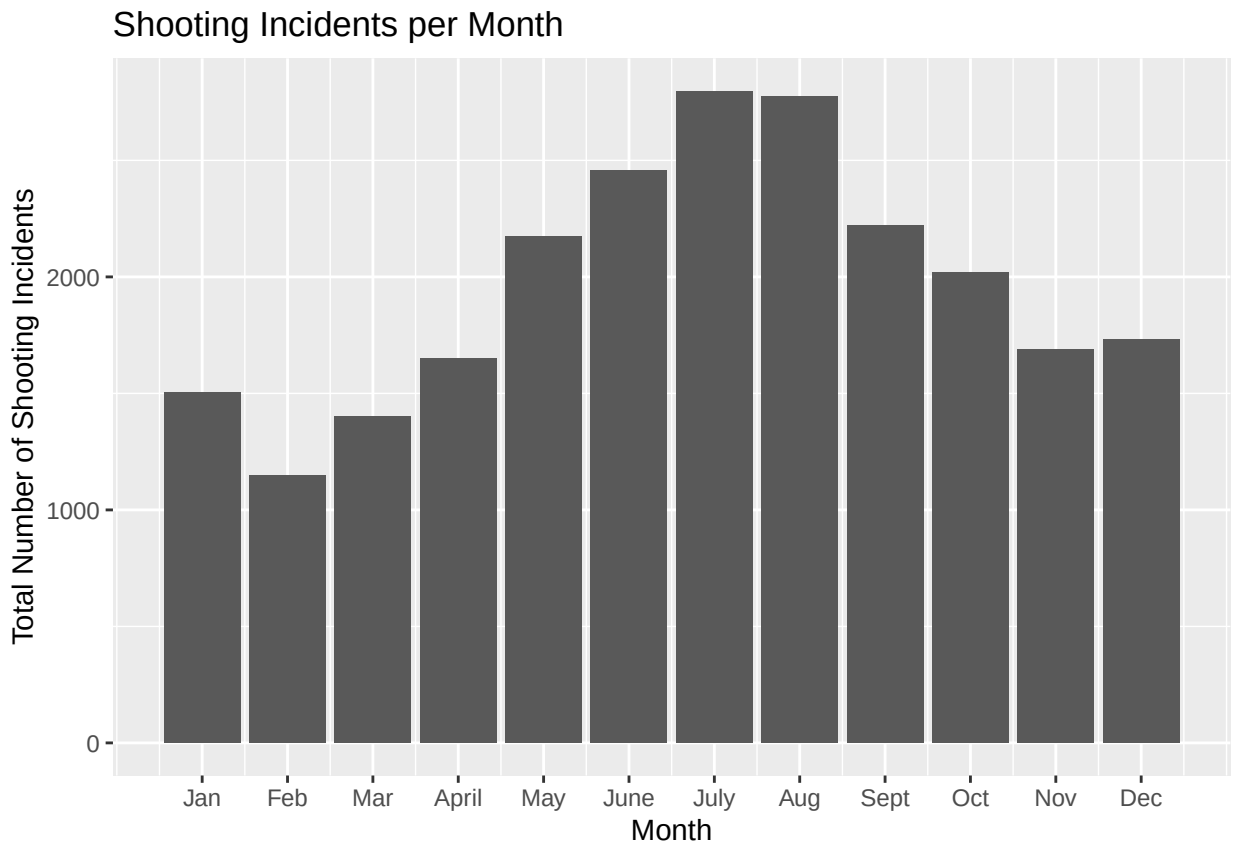
Similar age group distribution patterns are seen here as in the tables describing victims. The first table shows that perpetrators of shooting incidents are predominantly identified as Black, Black Hispanic or White Hispanic. Perpetrators identified as Black, Black Hispanic or White Hispanic are predominantly in the <18 and 18-24 age groups while perpetrators identified as White or Asian/Pacific Islander are predominantly in the 18-24 and 24-44 age groups. Similar to what was shown in the victim tables, identified perpetrators are disportionately identified as from black and brown communities and tend to be younger than other perpetrators.

Total Number of Shooting Incidents

The geographic region of the city could also give insight into the effect shootings have upon the residents of New York City. The shooting incidents were grouped by month and total number of incidents per month were calculated from the size of the sub-datasets. The subsequent chart displays a clear pattern showing a higher number of shooting incidents occurring in the warmer months of the year.

```
#Group by month and renmae
data_2 <- shooting_data %>% group_by(month(OCCUR_DATE)) %>%
        summarize(count=n())

data_2<-rename(data_2, c(Months = 'month(OCCUR_DATE)', Num_Shooting_Incidents = count))

#PLot month-by-month results
ggplot(data = data_2, aes(x=Months,y=Num_Shooting_Incidents)) +
        geom_bar(stat = 'identity') +
```

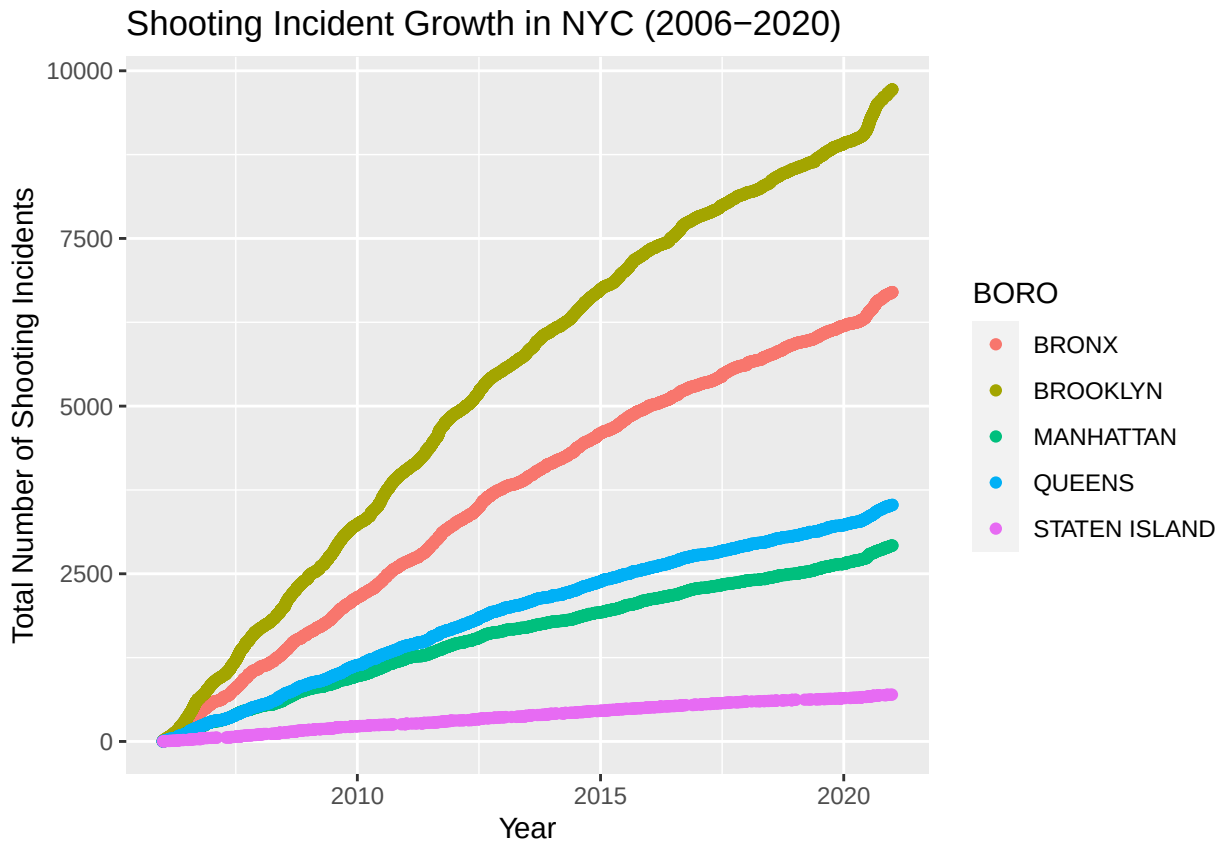
```
        scale_x_continuous(breaks=1:12,
                           labels=c('Jan', 'Feb', 'Mar', 'April', 'May', 'June', 'July',
                                    'Aug','Sept', 'Oct', 'Nov', 'Dec')) +
        labs(title="Shooting Incidents per Month",
             x = "Month",
             y="Total Number of Shooting Incidents")
```



Since date and time were provided for each shooting incident, a cumulative sum of shooting incidents can be calculated. With each new shooting incident, the cumulative sum will grow. This allows for the rate of increase to be visualized and analyzed. A positive trend is anticipated and natural, but the rate of increase will describe how the total number of shooting incidents changing over time. The shooting incidents were first divided by boro in order to see the comparative growth and number of shooting incidents for each boro.

```
#Find Boro specific incidents and total numbers
shooting_data_Bronx <- shooting_data %>% filter(BORO == 'BRONX') %>%
        arrange(OCCUR_D_T)
shooting_data_Bronx$Total_Incidents <- seq(1:nrow(shooting_data_Bronx))

shooting_data_Man <- shooting_data %>% filter(BORO == 'MANHATTAN') %>%
        arrange(OCCUR_D_T)
shooting_data_Man$Total_Incidents <- seq(1:nrow(shooting_data_Man))

shooting_data_Brook <- shooting_data %>% filter(BORO == 'BROOKLYN') %>%
        arrange(OCCUR_D_T)
shooting_data_Brook$Total_Incidents <- seq(1:nrow(shooting_data_Brook))
```

```r
shooting_data_Queens <- shooting_data %>% filter(BORO == 'QUEENS') %>%
        arrange(OCCUR_D_T)
shooting_data_Queens$Total_Incidents <- seq(1:nrow(shooting_data_Queens))

shooting_data_StatIsl <- shooting_data %>% filter(BORO == 'STATEN ISLAND') %>%
        arrange(OCCUR_D_T)
shooting_data_StatIsl$Total_Incidents <- seq(1:nrow(shooting_data_StatIsl))

#combine into one dataset
shooting_total <- rbind(shooting_data_Bronx, shooting_data_Brook,
                        shooting_data_Man, shooting_data_Queens,
                        shooting_data_StatIsl)

#plot data set
ggplot(data=shooting_total) +
        geom_point(aes(x=OCCUR_D_T, y=Total_Incidents, color=BORO)) +
        labs(title="Shooting Incident Growth in NYC (2006-2020)",
             x = "Year",
             y="Total Number of Shooting Incidents")
```



While each boro shows the expected and inevitable increase of total shooting incident cases from 2006 to 2020, the growth is not consistent. The accumulation is steep initially but there is a slight curve and the plots start to level off slightly around 2012 and particularly after 2015. It is also notable that there is an uptick of recorded shooting incidents from 2019 to 2020 in all boros except Staten Island, potentially due to the effect of the COVID-19 pandemic. The plot also shows a clear distinction between the incident numbers within the five boros of New York City. The Bronx and Brooklyn could be categorized with a high number of

incidents, Manhattan and Queens a moderate number, and Staten Island a low number, relatively speaking.

Modeling of Shooting Incidents over Time

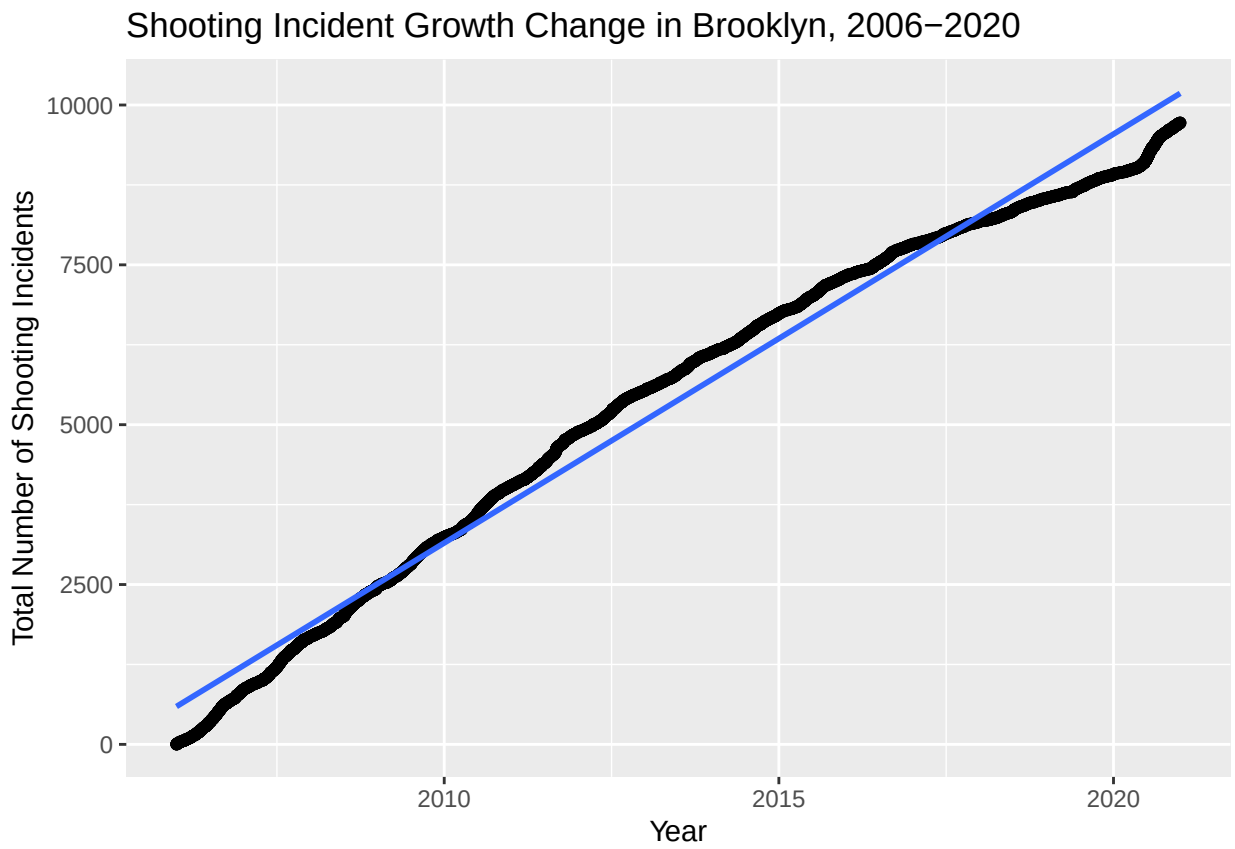
The visual above shooting the accumulation of shooting incidents over time stirred some further questions about the potential changing rate of shooting incidents over the years. The total number of shooting incidents from Brooklyn, the boro with the highest number of shooting incidents, was plotted and a linear model calculated.

```
#Only keep Brookylyn incidents
shooting_data_Brook <- shooting_data %>% filter(BORO == 'BROOKLYN') %>%
        arrange(OCCUR_DATE)

shooting_data_Brook$Total_Incidents <- seq(1:nrow(shooting_data_Brook))

#Calculate and view model
Brook_model <- lm(shooting_data_Brook$Total_Incidents ~ shooting_data_Brook$OCCUR_DATE)

summary(Brook_model)
```

```
##
## Call:
## lm(formula = shooting_data_Brook$Total_Incidents ~ shooting_data_Brook$OCCUR_DATE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -756.85 -332.57   60.83  352.68  517.50
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.243e+04  3.691e+01  -607.7   <2e-16 ***
## shooting_data_Brook$OCCUR_DATE  1.751e+00  2.356e-03   743.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369.1 on 9720 degrees of freedom
## Multiple R-squared:  0.9827, Adjusted R-squared:  0.9827
## F-statistic: 5.524e+05 on 1 and 9720 DF,  p-value: < 2.2e-16
```

```
#Plot Brooklyn incidents with model
ggplot(data=shooting_data_Brook) +
        geom_point(aes(x=OCCUR_DATE, y=Total_Incidents)) +
        geom_smooth(method='lm', aes(x=OCCUR_DATE, y=Total_Incidents)) +
        labs(title="Shooting Incident Growth Change in Brooklyn, 2006-2020",
             x = "Year",
             y="Total Number of Shooting Incidents")
```

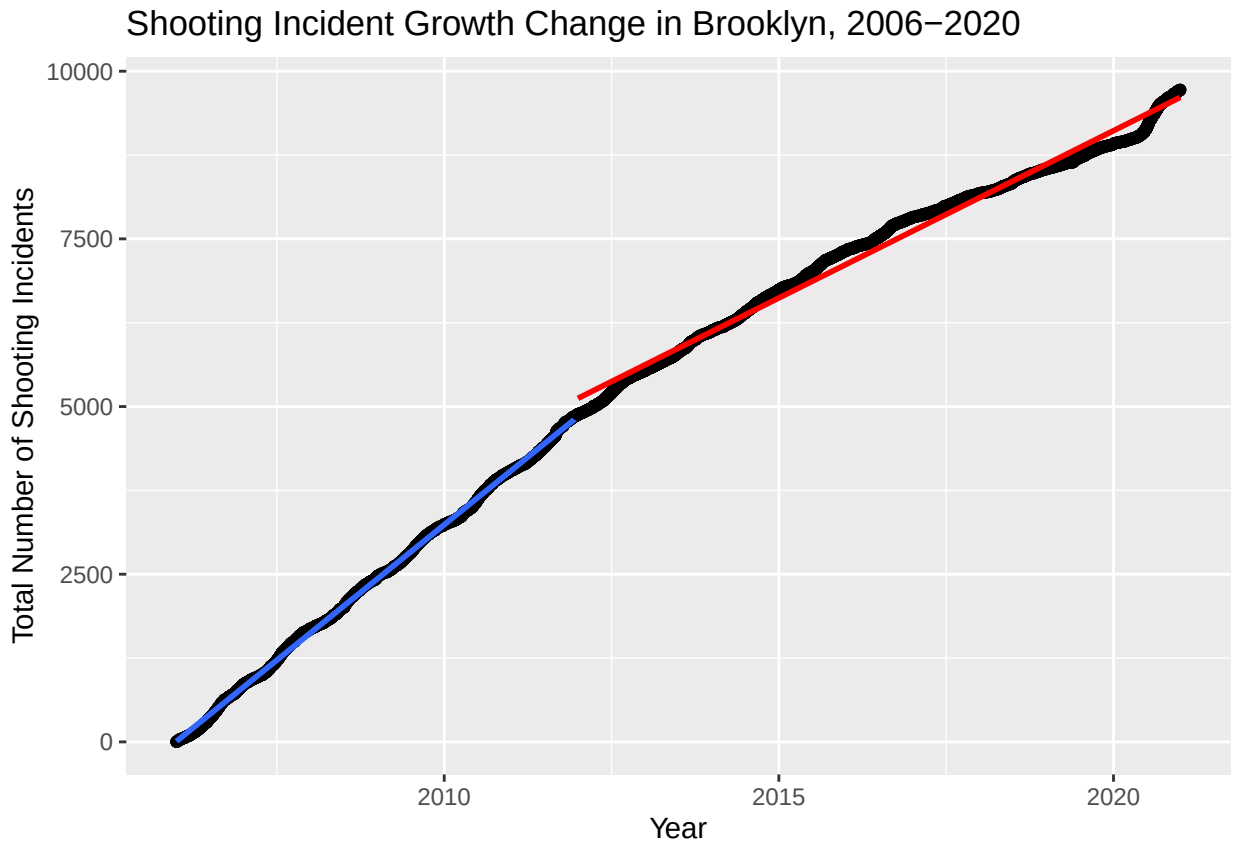
```
## `geom_smooth()` using formula 'y ~ x'
```



The linear model matched the general trend and provides some insight that each day brought, on average, 1.75 shooting incidents per day over the years 2006 to 2020. And while the results of the model are statistically significant, the model does not convey the curve in the underlying data. Specifically, it under estimates the rate from 2006 to approximately 2012 and overestimates the rate from approximately 2012 through to the end of 2020. This can clearly be seen in the plot above. The data was subsetted into two sets, one including incidents from 2002 through the end of 2011 and the other from the start of 2012 through 2020. The below plot displays the change in pattern and a clear, shallower regression line can be seen during the years 2012-2020.

```
#Subet Brooklyn Incidents int0 06-12 and 12-20
Brook_0612 <- shooting_data_Brook %>%
        filter(OCCUR_DATE >= "2006-01-01" & OCCUR_DATE < "2012-01-01")

Brook_1220 <- shooting_data_Brook %>%
        filter(OCCUR_DATE > "2011-12-31" & OCCUR_DATE <= "2020-12-31")

#Plot both sets with model for each set
ggplot() +
        geom_point(data = Brook_0612, aes(x=OCCUR_DATE, y=Total_Incidents)) +
        geom_smooth(data = Brook_0612, method='lm', aes(x=OCCUR_DATE, y=Total_Incidents)) +
        geom_point(data = Brook_1220, aes(x=OCCUR_DATE, y=Total_Incidents)) +
        geom_smooth(data = Brook_1220, method='lm', aes(x=OCCUR_DATE, y=Total_Incidents), color='red') +
        labs(title="Shooting Incident Growth Change in Brooklyn, 2006-2020",
             x = "Year",
             y="Total Number of Shooting Incidents")
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
#calculate both models for summary analysis
model_0612 <- lm(Brook_0612$Total_Incidents ~ Brook_0612$OCCUR_DATE)

model_1220 <- lm(Brook_1220$Total_Incidents ~ Brook_1220$OCCUR_DATE)

coefficients(model_0612)
```

```
##           (Intercept) Brook_0612$OCCUR_DATE
##         -29053.291630              2.210193
```

```
coefficients(model_1220)
```

```
##           (Intercept) Brook_1220$OCCUR_DATE
##         -15825.556489              1.365583
```

A comparison of the models shows a decrease in shooting incident per day, on average, in the 2012-2020 model compared to the 2006-2012 model. This confirms the initial observation examining the boro-by-boro trends. It is worth noting that the 2012-2020 model did not fit its underlying data well and failed to properly describe the curve present in the plot. Further analysis should be preformed to assess any additional change in the rate of shooting incident growth during this year range.

Conclusion and Analysis of Bias

Shooting incidents occur in varying numbers across the five boros of New York City while effecting populations of color in greater numbers and younger individuals within those communities at a higher rate. Further analysis should be preformed in order to discover and determine associations between racial groups, the socio-economic status of the boro in which the shooting occurred, the age group of the victim and perpetrator, as well as other demographic and socio-economic metrics to get a more complete picture of how shooting effect these various communities.

While each boro accumulated shooting incidents over time, an unfortunate reality in a a large, urban environment, they were not distributed evenly among the boros. The Bronx and Brooklyn saw noticeably higher numbers of shooting incidents while Staten Island saw the fewest during these years. Further analysis of these numbers including exploring racial and age distributions within the boros could provide more insight into potential causes the growth over time of shooting incidents in New York City.

There is also an observed change in the rate of increase of shooting incidents in the boro of Brooklyn. This reduction is welcome as it extrapolates out to roughly 3,000 fewer individuals involved in shooting in Brooklyn from 2012-2020 than if the previous trend continued. More research is needed to further define rate changes as well as explore and possible policing, city policy, or community changes which could have contributed to this decrease. Further analysis would also include similar sub-setting and exploration of other boros.

Since it is unclear how racial labels were collected, caution is needed when using the results of this analysis. Mislabeling shooting victims or perpetrators is possible and would result in miscounts and inflated numbers. Mislabeling biases the results against certain communities. Also, the summary tables at the beginning of this report show that about half of the incidents were removed because race or age was unknown. While this decision was made in order to provide a cleaner critique of the data, the results are biased as individuals labeled as “Unknown” would change counts and potentially relative percentages. Further analysis should be performed to identify the effect that removing so many entries had upon the results of the study. Finally, personal bias could have effected the creation of questions and types of analysis for this report. It is widely reported that shootings and gun violence disproportionately effects black and brown communities and, while knowing this fact, I attempted to approach the data impartially. Confirmation bias was always present as the data ultimately reinforced my beliefs. I attempted to mitigate this personal bias by describing only what I see in the data and not what I believe the data should show.

```
sessionInfo()
```

```
## R version 4.0.4 (2021-02-15)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] stringr_1.4.0    ggplot2_3.3.3    tidyr_1.1.3      lubridate_1.7.10
## [5] dplyr_1.0.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6        highr_0.8         pillar_1.6.0      compiler_4.0.4
##  [5] tools_4.0.4       digest_0.6.27     lattice_0.20-41   nlme_3.1-152
##  [9] evaluate_0.14     lifecycle_1.0.0   tibble_3.1.0      gtable_0.3.0
## [13] mgcv_1.8-33       pkgconfig_2.0.3   rlang_0.4.10      Matrix_1.3-2
## [17] DBI_1.1.1         yaml_2.2.1        xfun_0.22         withr_2.4.1
## [21] knitr_1.31        generics_0.1.0    vctrs_0.3.6       grid_4.0.4
## [25] tidyselect_1.1.0  glue_1.4.2        R6_2.5.0          fansi_0.4.2
## [29] rmarkdown_2.7     farver_2.1.0      purrr_0.3.4       magrittr_2.0.1
## [33] splines_4.0.4     scales_1.1.1      ellipsis_0.3.1    htmltools_0.5.1.1
## [37] assertthat_0.2.1  colorspace_2.0-0  labeling_0.4.2    utf8_1.2.1
## [41] stringi_1.5.3     munsell_0.5.0     crayon_1.4.1
```