NYPD Historical Shooting Analysis

Quashaun Vallery

2022-07-03

Description of Data

The NYPD Shooting Incident Historical data lists every shooting incident that occurred in NYC going back to 2006 through 2021. 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.

Below is a description of all columns in the dataset:

1. INCIDENT_KEY: Randomly generated persistent ID for each incident
2. OCCUR DATE: Exact date of the shooting incident
3. OCCUR TIME: Exact time of the shooting incident
4. BORO: Borough where the shooting incident occurred
5. PRECINCT: Precinct where the shooting incident occurred
6. JURISDICTION_CODE: Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit), and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions.
7. LOCATION_DESC: Location of the shooting incident
8. STATISTICAL_MURDER_FLAG: Shooting resulted in the victim's death which would be counted as a murder
9. PERP_AGE_GROUP: Perpetrator's age within a category
10. PERP SEX: Perpetrator's age within a category
11. PERP RACE: Perpetrator's race description
12. VIC_AGE_GROUP: Victim's age within a category
13. VIC SEX: Victim's sex description
14. VIC RACE: Victim's race description
15. X_COORD_CD: Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD83, units fleet (FIPS 3104)
16. Y_COORD_CD: Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD83, units fleet (FIPS 3104)

Objective

For this analysis I answer the following questions:

1. What day of the week do shootings occur the most?
2. How many shootings are there per year? What is the trend over time?
3. What borough has the most shooting incidents? Does the borough with the most shootings change year-to-year?

Finally, I developed a linear regression model to predict murders as a function of shootings.

Import Libraries and Data

I started by importing the necessary libraries and data.

```
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   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   1.4.0     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(tidyquant)
```

```
## Loading required package: PerformanceAnalytics

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: ’zoo’

## The following objects are masked from ’package:base’:
##
##     as.Date, as.Date.numeric

##
## Attaching package: ’xts’

## The following objects are masked from ’package:dplyr’:
##
##     first, last

##
## Attaching package: ’PerformanceAnalytics’
## The following object is masked from 'package:graphics':
##
##     legend

## Loading required package: quantmod

## Loading required package: TTR

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## == Need to Learn tidyquant? ====================================================
## Business Science offers a 1-hour course - Learning Lab #9: Performance Analysis & Portfolio Optimiza
## </> Learn more at: https://university.business-science.io/p/learning-labs-pro </>
```

```
NYPD_shooting_tbl <- read_csv("NYPD_Shooting_Incident_Data__Historic_.csv")
```

```
## Warning in gzfile(file, mode): cannot open compressed file 'C:/Users/qvall/
## AppData/Local/Temp/RtmpcPqhbT\filea9e04be951c7', probable reason 'No such file
## or directory'

##
## -- 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_double(),
##   Y_COORD_CD = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Examine Data

I then examined the data for any import issues or missing values. There were no issues reading in the file. However, there were five columns with missing data.

- JURISDICTION_CODE
- LOCATION_DESC
- PERP_AGE_GROUP
- PERP_SEX
- PERP_RACE

There are the same number of observations (9,310) missing from the PERP_SEX and PERP_RACE columns. The first bias that came to mind was that these are probably shooting incidents when a suspect was unidentifiable resulting in an unsolved case. However, I don't have the evidence to prove that. Additionally, those fields are not needed for this analysis.

```
# Check for problems reading in the file
readr::problems(NYPD_shooting_tbl)
```

```
## [1] row      col      expected actual
## <0 rows> (or 0-length row.names)
```

```
# High-level view
NYPD_shooting_tbl %>% glimpse()
```

```
## Rows: 25,596
## Columns: 19
## $ INCIDENT_KEY            <dbl> 24050482, 77673979, 226950018, 237710987, 2247~
## $ OCCUR_DATE              <chr> "08/27/2006", "03/11/2011", "04/14/2021", "12/~
## $ OCCUR_TIME              <time> 05:35:00, 12:03:00, 21:08:00, 19:30:00, 00:18~
## $ BORO                    <chr> "BRONX", "QUEENS", "BRONX", "BRONX", "MANHATTA~
## $ PRECINCT                <dbl> 52, 106, 42, 52, 34, 75, 32, 26, 41, 67, 43, 6~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0~
## $ LOCATION_DESC           <chr> NA, NA, "COMMERCIAL BLDG", NA, NA, NA, NA, "MU~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, FALSE, TRUE, FALSE, FALSE, TRUE, FALSE, ~
## $ PERP_AGE_GROUP          <chr> NA, NA, NA, NA, NA, "25-44", "25-44", NA, "25-~
## $ PERP_SEX                <chr> NA, NA, NA, NA, NA, "M", "M", NA, "M", NA, NA,~
## $ PERP_RACE               <chr> NA, NA, NA, NA, NA, "BLACK HISPANIC", "BLACK",~
## $ VIC_AGE_GROUP           <chr> "25-44", "65+", "18-24", "25-44", "25-44", "25~
## $ VIC_SEX                 <chr> "F", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK HISPANIC", "WHITE", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1017542, 1027543, 1009489, 1017440, 1005426, 1~
## $ Y_COORD_CD              <dbl> 255918.9, 186095.0, 243050.0, 256046.0, 254690~
## $ Latitude                <dbl> 40.86906, 40.67737, 40.83376, 40.86941, 40.865~
## $ Longitude               <dbl> -73.87963, -73.84392, -73.90880, -73.88000, -7~
## $ Lon_Lat                 <chr> "POINT (-73.87963173099996 40.86905819000003)"~
```

```
# Check for empty fields
find_NAs <- colSums(is.na(NYPD_shooting_tbl))
find_NAs
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO                PRECINCT       JURISDICTION_CODE
##                       0                       0                       2
##           LOCATION_DESC STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                   14977                       0                    9344
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                    9310                    9310                       0
##                 VIC_SEX                VIC_RACE              X_COORD_CD
##                       0                       0                       0
##              Y_COORD_CD                Latitude               Longitude
##                       0                       0                       0
##                 Lon_Lat
##                       0
```

Tidy and Transform

```
# convert occur_date from character to date field
NYPD_shooting_tbl <- NYPD_shooting_tbl %>%
    mutate(OCCUR_DATE= as.Date(OCCUR_DATE, format = "%m/%d/%Y")) %>%
    mutate(year_of_shooting = year(OCCUR_DATE)) %>%

# Remove columns with missing data and select fields needed to answer my questions
    select(-JURISDICTION_CODE, -LOCATION_DESC, -PERP_SEX,
           -PERP_AGE_GROUP, -PERP_RACE) %>% glimpse()
```

```
## Rows: 25,596
## Columns: 15
## $ INCIDENT_KEY            <dbl> 24050482, 77673979, 226950018, 237710987, 2247~
## $ OCCUR_DATE              <date> 2006-08-27, 2011-03-11, 2021-04-14, 2021-12-1~
## $ OCCUR_TIME              <time> 05:35:00, 12:03:00, 21:08:00, 19:30:00, 00:18~
## $ BORO                    <chr> "BRONX", "QUEENS", "BRONX", "BRONX", "MANHATTA~
## $ PRECINCT                <dbl> 52, 106, 42, 52, 34, 75, 32, 26, 41, 67, 43, 6~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, FALSE, TRUE, FALSE, FALSE, TRUE, FALSE, ~
## $ VIC_AGE_GROUP           <chr> "25-44", "65+", "18-24", "25-44", "25-44", "25~
## $ VIC_SEX                 <chr> "F", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK HISPANIC", "WHITE", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1017542, 1027543, 1009489, 1017440, 1005426, 1~
## $ Y_COORD_CD              <dbl> 255918.9, 186095.0, 243050.0, 256046.0, 254690~
## $ Latitude                <dbl> 40.86906, 40.67737, 40.83376, 40.86941, 40.865~
## $ Longitude               <dbl> -73.87963, -73.84392, -73.90880, -73.88000, -7~
## $ Lon_Lat                 <chr> "POINT (-73.87963173099996 40.86905819000003)"~
## $ year_of_shooting        <dbl> 2006, 2011, 2021, 2021, 2021, 2021, 2021, 2021~
```

```
# Make sure all missing data has been resolved
find_NAs <- colSums(is.na(NYPD_shooting_tbl))
find_NAs
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO                PRECINCT STATISTICAL_MURDER_FLAG
##                       0                       0                       0
##           VIC_AGE_GROUP                 VIC_SEX                VIC_RACE
##                       0                       0                       0
##              X_COORD_CD              Y_COORD_CD                Latitude
##                       0                       0                       0
##               Longitude                 Lon_Lat        year_of_shooting
##                       0                       0                       0
```

Analysis and Visualization

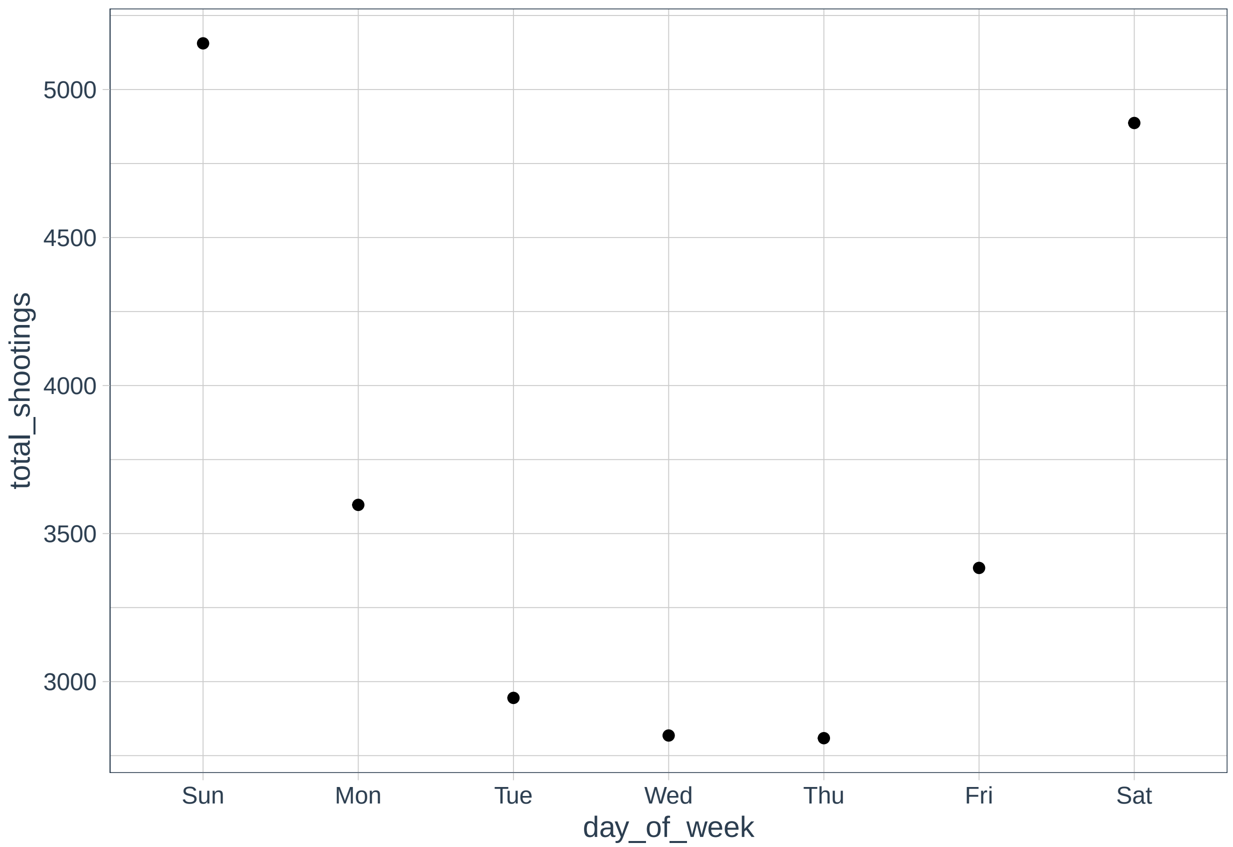
What day of the week do shootings occur the most?

Most shootings happened on the weekend with Sunday and Saturday accounting for 20 and 19 percent, respectively, of all shootings throughout the week.

```
day_shooting_tbl <- NYPD_shooting_tbl %>%
    select(OCCUR_DATE, year_of_shooting) %>%
    mutate(day_of_week = wday(OCCUR_DATE, label = TRUE, abbr = TRUE)) %>%
    mutate(count = 1)

day_shooting_summary_tbl <- day_shooting_tbl %>%
    group_by(day_of_week) %>%
    summarize(total_shootings = sum(count)) %>%
    arrange(desc(total_shootings)) %>%
    mutate(pct = total_shootings / sum(total_shootings)) %>%
    mutate(pct = scales::percent(pct, accuracy = 1.)) %>%
    ungroup()

day_shooting_summary_tbl
```

```
## # A tibble: 7 x 3
##   day_of_week total_shootings pct
##   <ord>                 <dbl> <chr>
## 1 Sun                    5156 20%
## 2 Sat                    4887 19%
## 3 Mon                    3597 14%
## 4 Fri                    3384 13%
## 5 Tue                    2945 12%
## 6 Wed                    2818 11%
## 7 Thu                    2809 11%
```

```
day_shooting_plot <- day_shooting_summary_tbl %>%
    ggplot(aes(x = day_of_week, y = total_shootings)) +
    geom_point() +
    theme_tq()

day_shooting_plot
```



How many shootings are there per year? What is the trend over time?

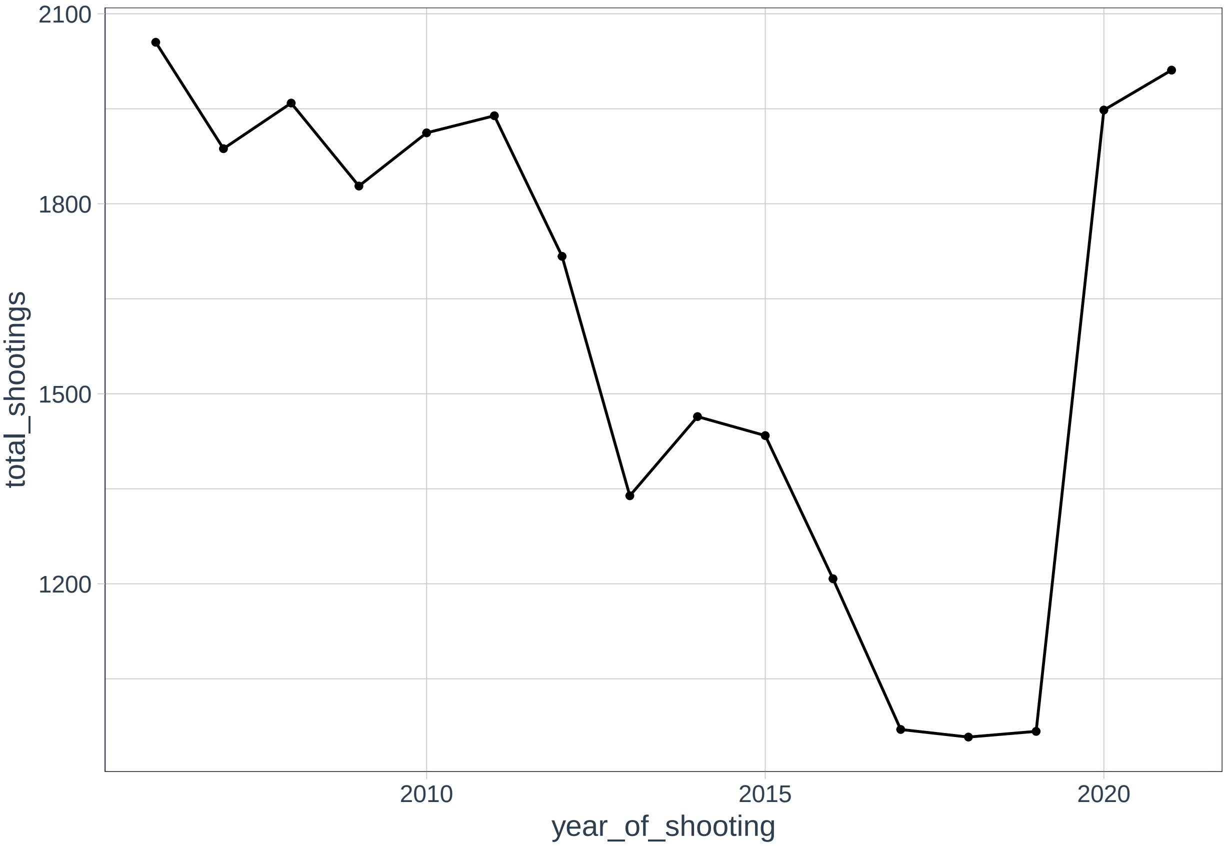
The number of shootings per year ranged from 967 to 2,055 over the 15-year period. Shootings had been steadily declining from 2006, with big drops in 2011 through 2013 and 2015 through 2017. However, shootings increased sharply after 2019, and were back to 2006 levels by 2021.

After seeing this trend, I was curious whether the COVID-19 pandemic had something to do with the increased shootings. However, I didn't go further with this line of thinking because that would have required gathering additional data sources.

```
shooting_by_yr_tbl <- NYPD_shooting_tbl %>%
        select(OCCUR_DATE, year_of_shooting) %>%
        mutate(count = 1) %>%

        group_by(year_of_shooting) %>%
        summarize(total_shootings = sum(count)) %>%
        ungroup()

    shooting_by_yr_plot <-  shooting_by_yr_tbl %>%
        ggplot(aes(x = year_of_shooting, y = total_shootings)) +
        geom_line() +
        geom_point(shape = 20) +
        theme_tq()


    shooting_by_yr_tbl
```

```
## # A tibble: 16 x 2
##    year_of_shooting total_shootings
##               <dbl>           <dbl>
##  1             2006            2055
##  2             2007            1887
##  3             2008            1959
##  4             2009            1828
##  5             2010            1912
##  6             2011            1939
##  7             2012            1717
##  8             2013            1339
##  9             2014            1464
## 10             2015            1434
## 11             2016            1208
## 12             2017             970
## 13             2018             958
## 14             2019             967
## 15             2020            1948
## 16             2021            2011
```

```
shooting_by_yr_plot
```



What borough has the most shooting incidents? Does the borough with the most shootings change year-to-year?

Forty percent of all shootings over the 15-year period occurred in Brooklyn. I was curious if this was the case for all years and found that most shooting incidents occurred in Brooklyn except for the year 2021. Most shootings occurred in the Bronx in 2021. Lastly, I created box plots for each borough to see the distribution of shootings by borough.

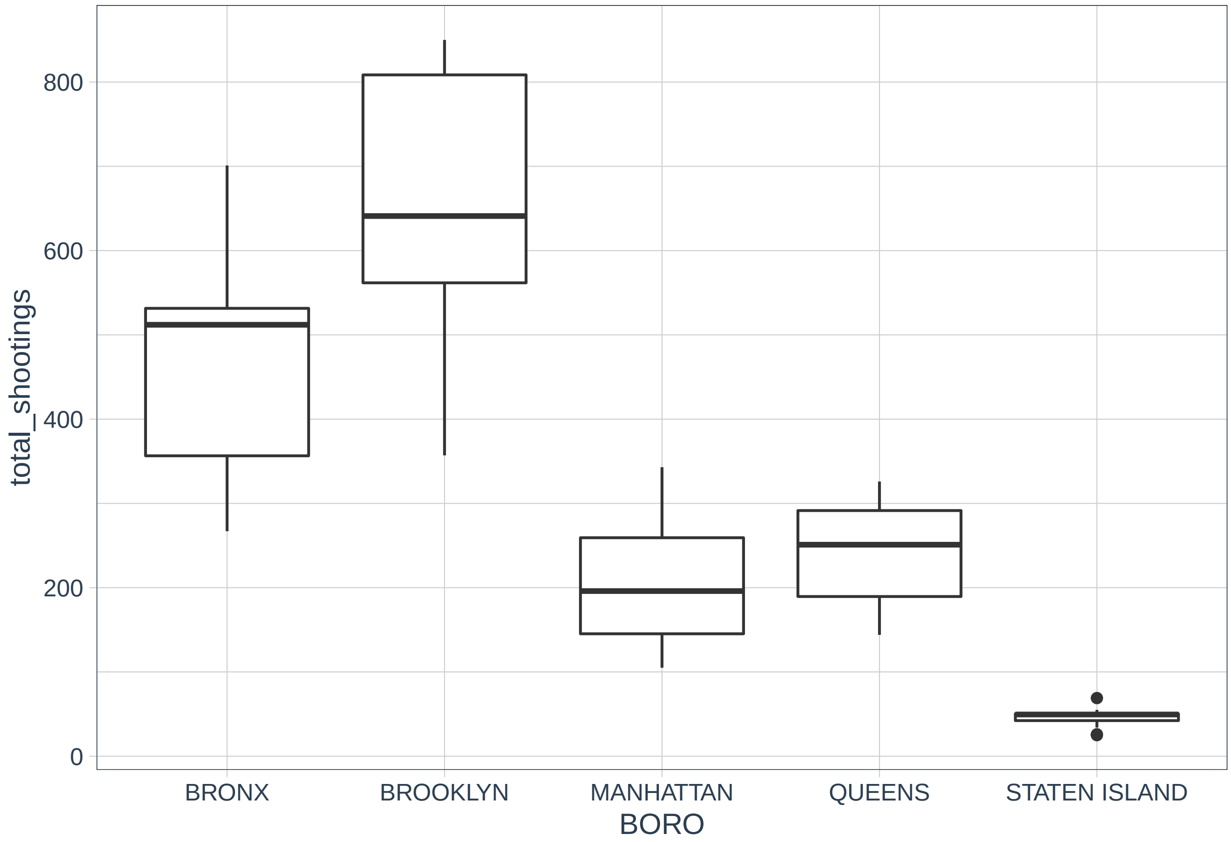
```
boro_shooting_tbl <- NYPD_shooting_tbl %>%
    select(BORO, PRECINCT, year_of_shooting) %>%
    mutate(count = 1)

boro_shooting_summary_tbl <- boro_shooting_tbl %>%
    group_by(BORO) %>%
    summarize(total_shootings = sum(count)) %>%
    arrange(desc(total_shootings)) %>%
    mutate(pct = total_shootings / sum(total_shootings)) %>%
    mutate(pct = scales::percent(pct, accuracy = 1.)) %>%
    ungroup()

boro_shooting_by_yr_tbl <- boro_shooting_tbl %>%
    select(BORO, year_of_shooting, count) %>%
    group_by(year_of_shooting, BORO) %>%
    summarize(total_shootings = sum(count)) %>%
    ungroup()
```

```
## ‘summarise()‘ has grouped output by ’year_of_shooting’. You can override using the ‘.groups‘ argumen
```

```
boro_shooting_by_yr_spread <- boro_shooting_by_yr_tbl %>%
    spread(key = BORO, value = total_shootings)

boro_shooting_by_yr_spread$row_max <- apply(boro_shooting_by_yr_spread[,-1], 1, max)
boro_shooting_by_yr_spread
```

```
## # A tibble: 16 x 7
##    year_of_shooting BRONX BROOKLYN MANHATTAN QUEENS ‘STATEN ISLAND‘ row_max
##               <dbl> <dbl>    <dbl>     <dbl>  <dbl>           <dbl>   <dbl>
##  1             2006   568      850       288    296              53     850
##  2             2007   533      833       233    238              50     833
##  3             2008   520      785       259    326              69     785
##  4             2009   529      770       196    278              55     770
##  5             2010   525      805       260    288              34     805
##  6             2011   571      839       215    264              50     839
##  7             2012   531      651       196    290              49     651
##  8             2013   371      593       138    185              52     593
##  9             2014   446      614       143    218              43     614
## 10             2015   409      583       187    205              50     583
## 11             2016   308      498       167    191              44     498
## 12             2017   306      357       117    144              46     357
## 13             2018   313      365       105    150              25     365
## 14             2019   267      372       146    156              26     372
## 15             2020   504      819       272    303              50     819
## 16             2021   701      631       343    296              40     701
```

```
boro_shooting_by_yr_plot <- boro_shooting_by_yr_tbl %>%
    group_by(year_of_shooting) %>%
    ggplot(aes(x = BORO, y = total_shootings)) +
    geom_boxplot() +
    theme_tq()

boro_shooting_by_yr_plot
```



Modeling

For my model, I chose the predict the number of murders as a function of shootings. The model does a pretty good job at predicting murders illustrating that the number of shooting is a pretty good indicator for murders.

```
# create total_num_yrs column
NYPD_shooting_tbl <- NYPD_shooting_tbl %>%
    mutate(total_num_yrs = max(year_of_shooting) - min(year_of_shooting))

# create shooting tbl
shooting_by_yr_tbl <- NYPD_shooting_tbl %>%
    select(OCCUR_DATE, year_of_shooting, total_num_yrs) %>%
    mutate(count = 1) %>%

    group_by(year_of_shooting, total_num_yrs) %>%
    summarize(total_shootings = sum(count)) %>%
    ungroup()
```

```
## `summarise()` has grouped output by 'year_of_shooting'. You can override using the `.groups` argumen
```

```
# create deaths tibble
deaths_by_yr_tbl <- NYPD_shooting_tbl %>%
    select(OCCUR_DATE, year_of_shooting, total_num_yrs, STATISTICAL_MURDER_FLAG) %>%
    mutate(count = 1) %>%
    mutate(murder_count = case_when(
        STATISTICAL_MURDER_FLAG == TRUE ~ 1,
        STATISTICAL_MURDER_FLAG == FALSE ~ 0)) %>%

    group_by(year_of_shooting, total_num_yrs) %>%
    summarize(total_murders = sum(murder_count)) %>%
    ungroup()
```

```
## `summarise()` has grouped output by 'year_of_shooting'. You can override using the `.groups` argumen
```

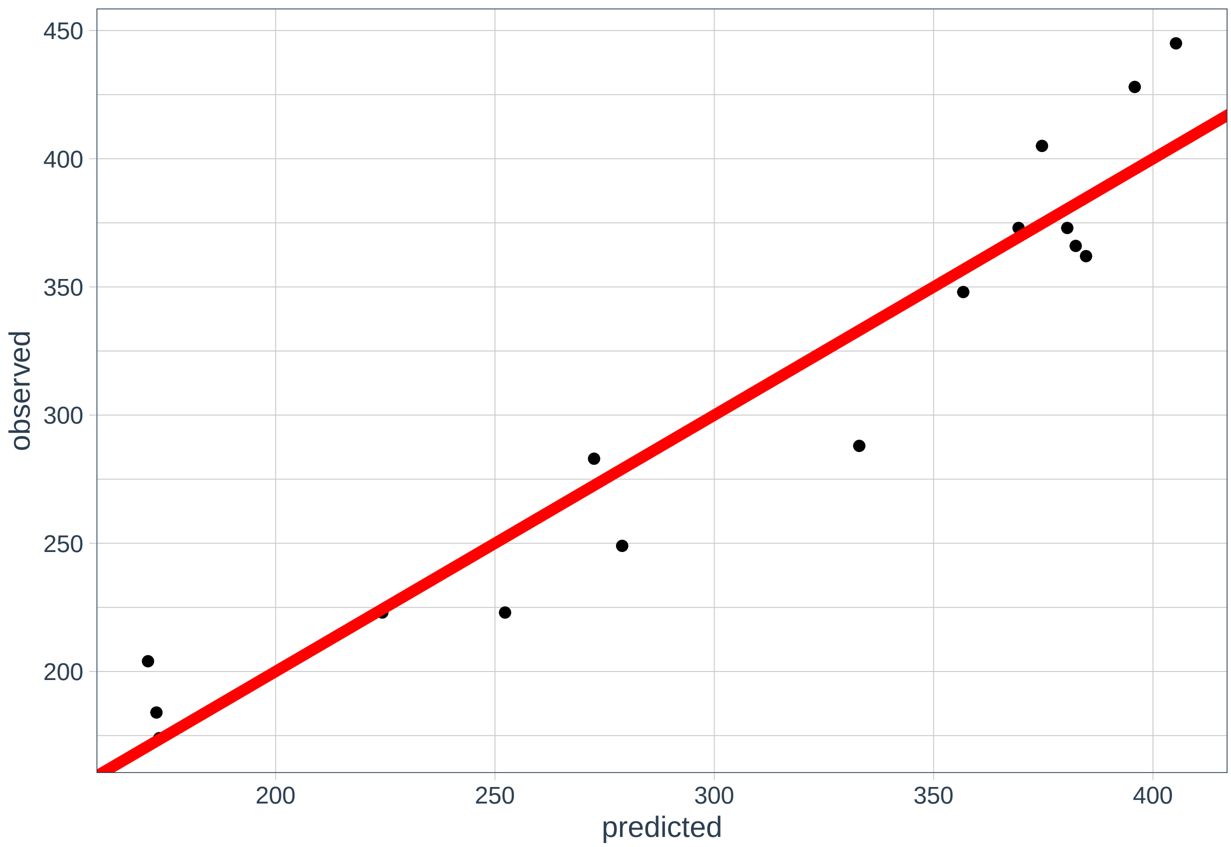
```
# join tibbles
joined_tbl <- left_join(shooting_by_yr_tbl, deaths_by_yr_tbl,
                        by = c("year_of_shooting" = "year_of_shooting"))

joined_tbl <- joined_tbl %>%
    select(-total_num_yrs.y)

joined_tbl
```

```
## # A tibble: 16 x 4
##    year_of_shooting total_num_yrs.x total_shootings total_murders
##               <dbl>           <dbl>           <dbl>         <dbl>
##  1             2006              15            2055           445
##  2             2007              15            1887           373
##  3             2008              15            1959           362
##  4             2009              15            1828           348
##  5             2010              15            1912           405
##  6             2011              15            1939           373
##  7             2012              15            1717           288
##  8             2013              15            1339           223
##  9             2014              15            1464           249
## 10             2015              15            1434           283
## 11             2016              15            1208           223
## 12             2017              15             970           174
## 13             2018              15             958           204
## 14             2019              15             967           184
## 15             2020              15            1948           366
## 16             2021              15            2011           428
```

```
# my model: estimate linear regression
my_mod <- lm(total_murders ~ total_shootings, joined_tbl)
summary(my_mod)
```

```
##
## Call:
## lm(formula = total_murders ~ total_shootings, data = joined_tbl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.047 -17.980  -0.395  15.950  39.750
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -33.73553   27.45209  -1.229    0.239
## total_shootings   0.21362    0.01667  12.817    4e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 26.16 on 14 degrees of freedom
## Multiple R-squared:  0.9215, Adjusted R-squared:  0.9159
## F-statistic: 164.3 on 1 and 14 DF,  p-value: 3.996e-09
```

```r
# create data for ggplot
data_mod_tbl <- tibble(predicted = predict(my_mod),
                       observed = joined_tbl$total_murders)

# create plot
ggplot(data_mod_tbl,
       aes(x = predicted,
           y = observed)) +
    geom_point() +
    geom_abline(intercept = 0,
                slope = 1,
                color = "red",
                size = 2) +
    theme_tq()
```



Bias

A bias that I had going into the project was that borough with the highest rate of poverty in NYC would also have the highest number of shootings. I also assumed that because the cost of living in Manhattan is astronomical, that it would have the least amount of shootings. I was wrong about both!

Brooklyn had the most shootings, which has less poverty than the Bronx, and Staten Island had the least shootings, not Manhattan. However, I avoided confirmation bias just by letting the data speak for itself throughout my analysis. I did not modify any observations or cherry pick data.