DTSA5301_wk3_shootingProj

NYPD Shooting Incident Data (Historic)

About dataset (from source):

"List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

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. Please refer to the attached data footnotes for additional information about this dataset."

---

1. Import Data

Read in the NYC shooting data from .csv:

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

#inspect raw data
head(shooting_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO PRECINCT JURISDICTION_CODE
## 1     24050482 08/27/2006   05:35:00    BRONX       52                 0
## 2     77673979 03/11/2011   12:03:00   QUEENS      106                 0
## 3    203350417 10/06/2019   01:09:00 BROOKLYN       77                 0
## 4     80584527 09/04/2011   03:35:00    BRONX       40                 0
## 5     90843766 05/27/2013   21:16:00   QUEENS      100                 0
## 6     92393427 09/01/2013   04:17:00 BROOKLYN       67                 0
##   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE
## 1                                 true
## 2                                false
## 3                                false
## 4                                false
## 5                                false
## 6                                false
##   VIC_AGE_GROUP VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1         25-44       F BLACK HISPANIC    1017542   255918.9 40.86906 -73.87963
## 2           65+       M          WHITE    1027543   186095.0 40.67737 -73.84392
## 3         18-24       F          BLACK     995325   185155.0 40.67489 -73.96008
## 4           <18       M          BLACK    1007453   233952.0 40.80880 -73.91618
```

```
## 5         18-24       M          BLACK    1041267   157133.5 40.59780 -73.79469
## 6           <18       M          BLACK    1001694   170112.9 40.63359 -73.93715
##                                       Lon_Lat
## 1  POINT (-73.87963173099996 40.86905819000003)
## 2 POINT (-73.84392019199998 40.677366895000034)
## 3 POINT (-73.96007501899999 40.674885741000026)
## 4  POINT (-73.91618413199996 40.80879780500004)
## 5 POINT (-73.79468553799995 40.597796249000055)
## 6  POINT (-73.93715330699996 40.63358818100005)
```

---

2. Tidy and Transform Data

Clean up data: * remove x/y coord and lat/long

```
shooting_data <- shooting_data %>%
  select(INCIDENT_KEY:VIC_RACE) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Get summary of dataset (shooting_data):

```
summary(shooting_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME            BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:23585       Length:23585
##  1st Qu.: 55322804   1st Qu.:2008-12-31   Class :character   Class :character
##  Median : 83435362   Median :2012-02-27   Mode  :character   Mode  :character
##  Mean   :102280741   Mean   :2012-10-05
##  3rd Qu.:150911774   3rd Qu.:2016-03-02
##  Max.   :230611229   Max.   :2020-12-31
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Length:23585
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   Class :character
##  Median : 69.00   Median :0.000     Mode  :character   Mode  :character
##  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
##  Length:23585       Length:23585
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
##
##
##
##
```

---

3. Visualize and Analyze Data

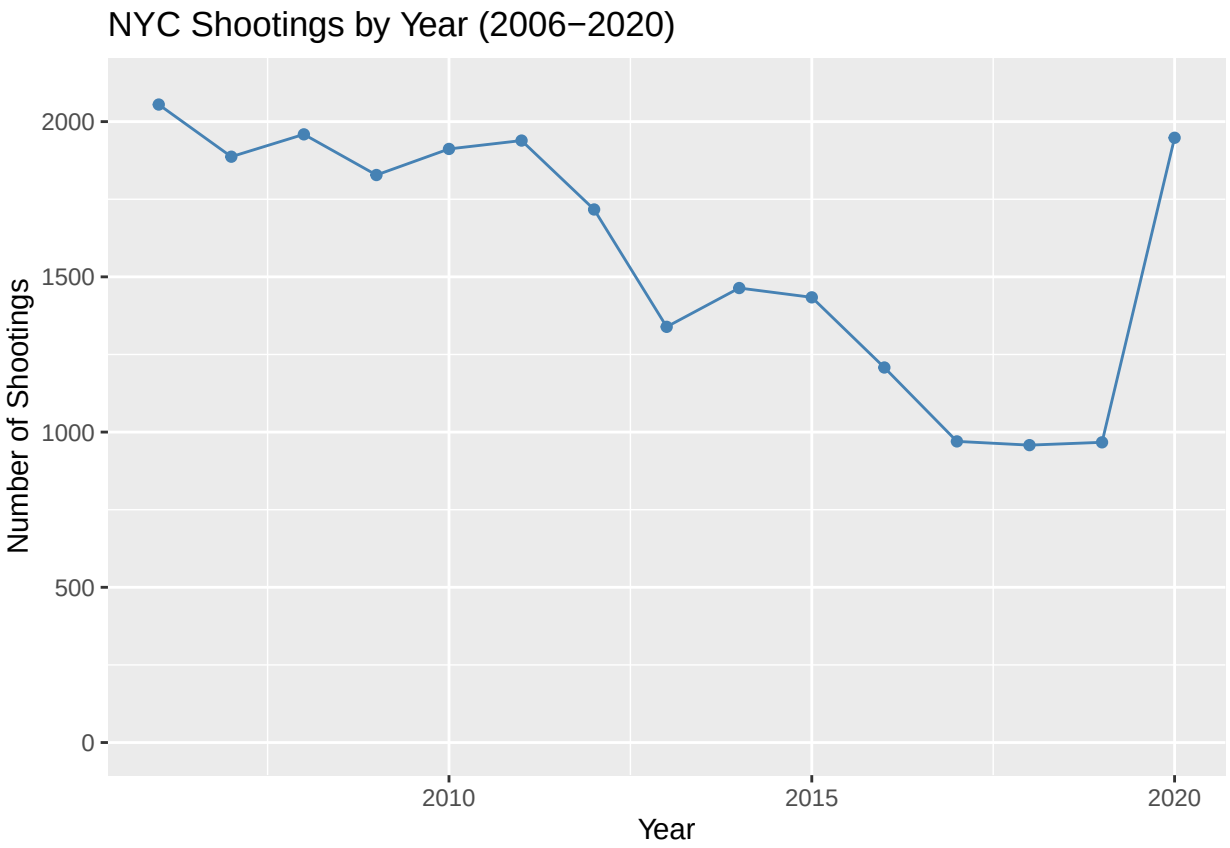
There are many questions we could ask regarding this dataset.

I'm going to start by asking if the rate of shootings has changed year over year.

```
shooting_data <- shooting_data %>%
  mutate(date_as_year = year(OCCUR_DATE))

shooting_data_by_year <- shooting_data %>%
  count(date_as_year)

ggplot(shooting_data_by_year, aes(x = date_as_year, y = n))  +
  geom_point(color = "steelblue") +
  geom_line(color = "steelblue") +
  ylim(0,2100) +
  labs(title = "NYC Shootings by Year (2006-2020)", y = "Number of Shootings", x = "Year")
```



Analysis of the year over year shooting rates show a steady decline in NYC shootings from 2006 to 2019, then a sudden spike in shootings in 2020.

Create Model

To objectively highlight the spike in shootings in 2020, and to mitigate any potential personal bias from the conclusions drawn, I will create a model so we can compare predicted vs actual.

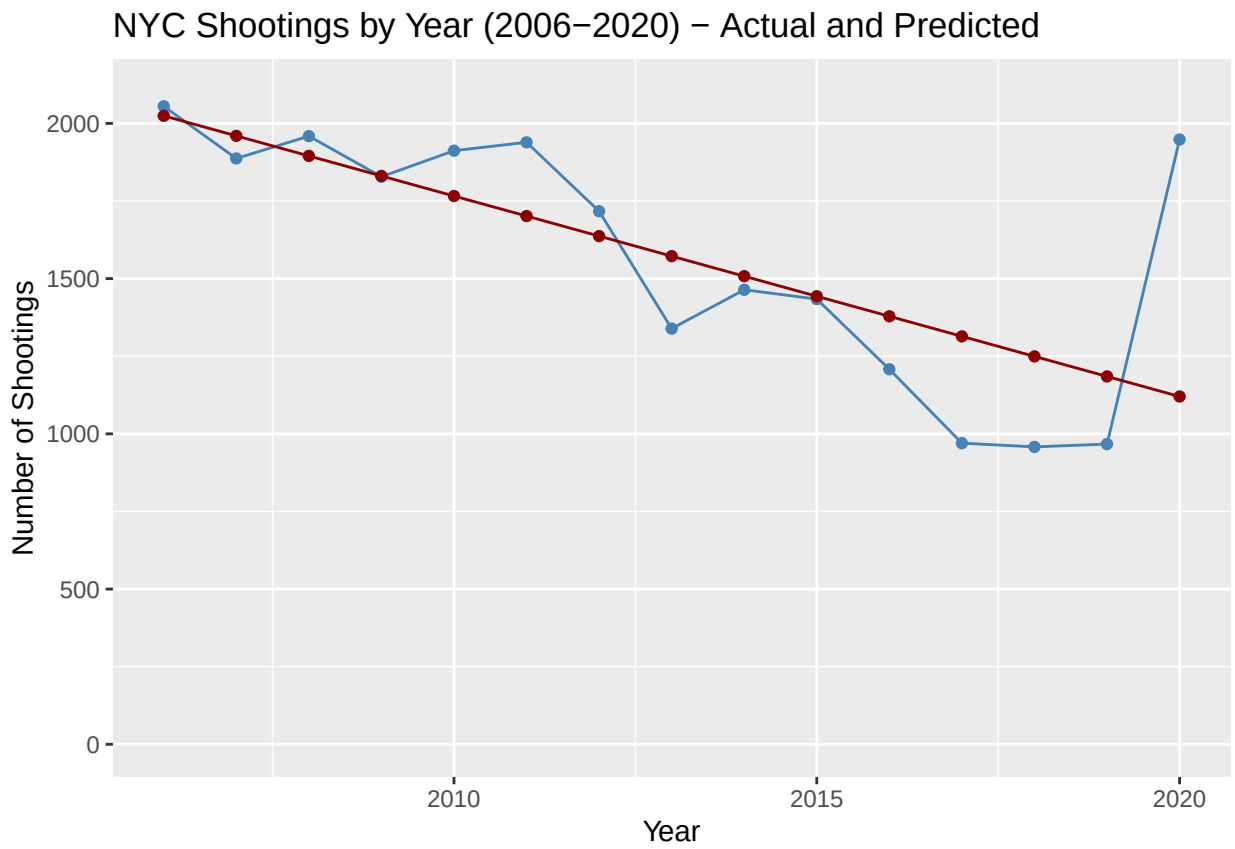
```
#create linear model
mod_by_year <- lm(n ~ date_as_year, data = shooting_data_by_year)

#create new dataset with predicted values for each year
shooting_data_by_year_pred <- shooting_data_by_year %>% mutate(pred = predict(mod_by_year))

#show prediction data
shooting_data_by_year_pred
```

```
##    date_as_year    n     pred
## 1          2006 2055 2024.358
## 2          2007 1887 1959.783
## 3          2008 1959 1895.208
## 4          2009 1828 1830.633
## 5          2010 1912 1766.058
## 6          2011 1939 1701.483
## 7          2012 1717 1636.908
## 8          2013 1339 1572.333
## 9          2014 1464 1507.758
## 10         2015 1434 1443.183
## 11         2016 1208 1378.608
## 12         2017  970 1314.033
## 13         2018  958 1249.458
## 14         2019  967 1184.883
## 15         2020 1948 1120.308
```

```
#plot actual shootings vs predicted shootings
ggplot(shooting_data_by_year_pred, aes(x = date_as_year))  +
  geom_point(aes(y = n), color = "steelblue") +
  geom_line(aes(y = n), color = "steelblue") +
  geom_point(aes(y = pred), color = "darkred") +
  geom_line(aes(y = pred), color = "darkred") +
  ylim(0,2100) +
  labs(title = "NYC Shootings by Year (2006-2020) - Actual and Predicted", y = "Number of Shootings", x
```

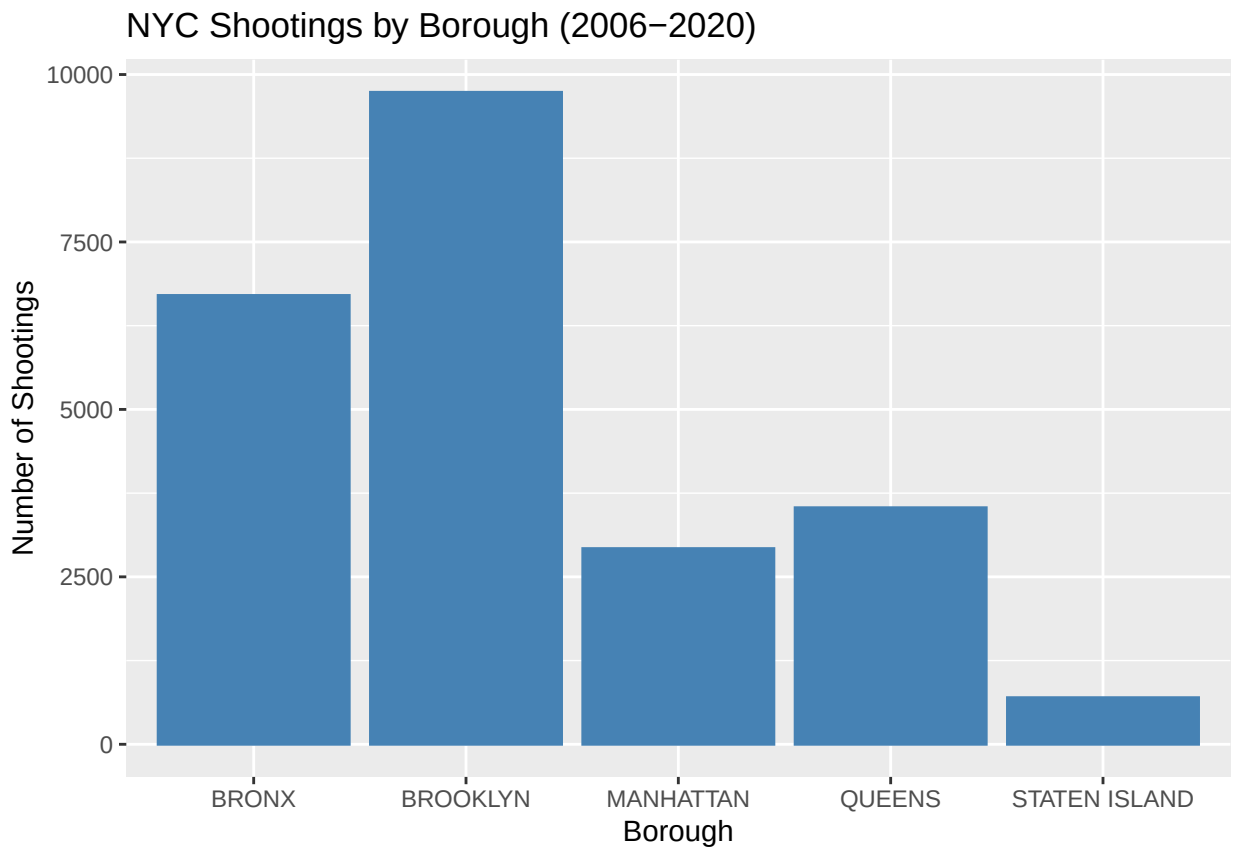


The actual (blue) vs predicted (red) values backup the previous conclusion citing that there was a downward trend in shooting from 2006 to 2019 with an unexpected increase in 2020.

The second visualization I'm going to look at is shootings by borough.

```
shooting_data_by_boro <- shooting_data %>%
  count(BORO)

ggplot(shooting_data_by_boro, aes(x = BORO, y = n)) +
  geom_bar(stat = "identity", color = "steelblue", fill = "steelblue") +
  labs(title = "NYC Shootings by Borough (2006-2020)", y = "Number of Shootings", x = "Borough")
```

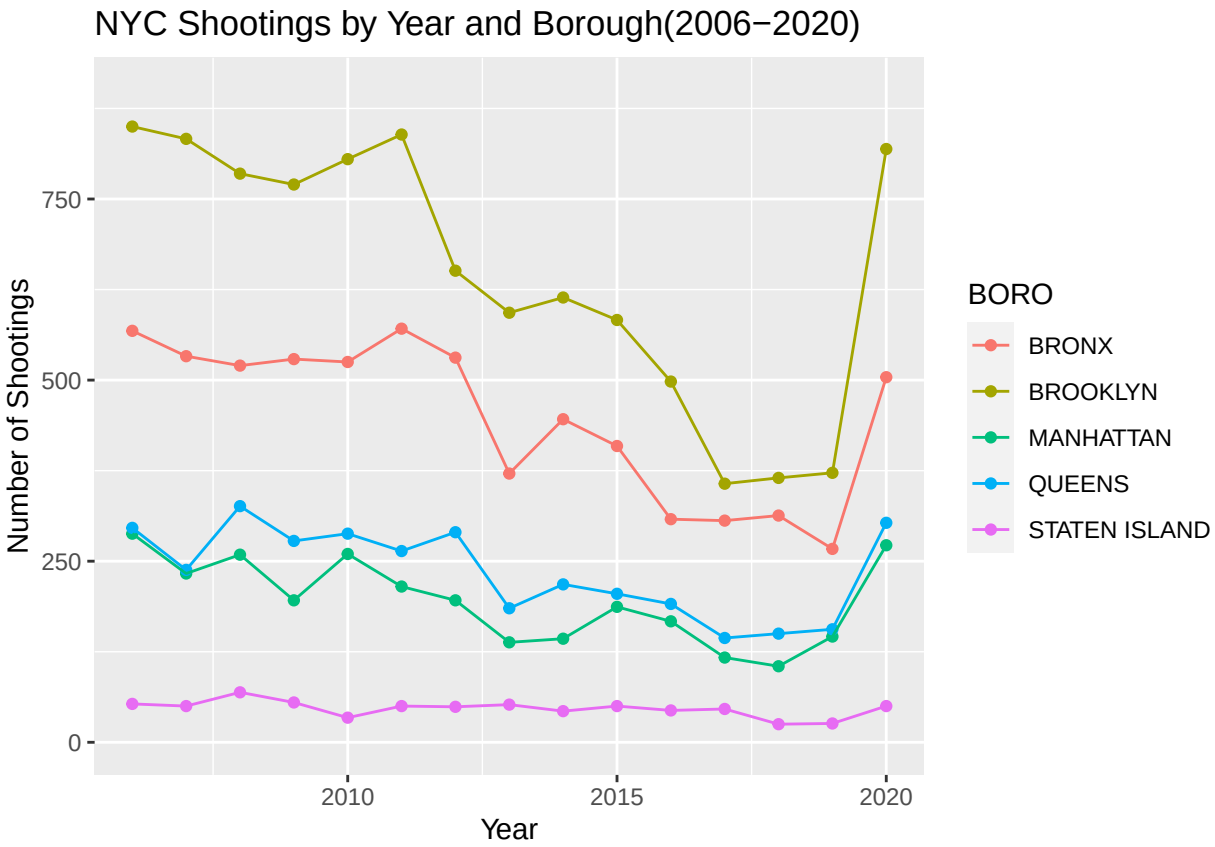


Analysis of the shootings by borough shows a large disparity in the number of shootings per borough, with Brooklyn clearly having the most followed by the Bronx. Staten Island had the least amount of shootings between 2006 and 2020.

Looking at these two visualizations leads me to question if the trend seen in the first graph is present for each borough.

```
shooting_data_by_year_boro <- shooting_data %>%
  count(date_as_year, BORO)

ggplot(shooting_data_by_year_boro, aes(x = date_as_year, y = n))  +
  geom_point(aes(x = date_as_year, y = n, colour = BORO)) +
  geom_line(aes(group = BORO, colour = BORO)) +
  ylim(0,900) +
  labs(title = "NYC Shootings by Year and Borough(2006-2020)", y = "Number of Shootings", x = "Year")
```



Separating shootings by year and borough shows that the overall trend of steady decline in number of shootings from 2006 to 2019 with a jump in 2020 was present across the 5 boroughs of NYC.

---

4. Identify Bias

The potential for personal bias affecting the analysis of a dataset like this NYC shooting data is extremely high. Themes such as **race** (PERP_RACE and VIC_RACE), **sex** (PERP_SEX, VIC_SEX), **age** (PERP_AGE_GROUP, VIC_AGE_GROUP) and **class and money** (BORO - i.e. Manhattan and Statten Island vs Brooklyn and the Bronx) all have the potential to come into play.

The analysis the I chose to perform (shootings year over year and shootings by borough) did not dig deeply into the more obvious sources of potential bias listed above. I can see, however, that even with these relatively objective analysis, a personal bias can come into play.

For instance, 2020 was a very politically volatile time in the US. With the first analysis, NYC Shooting by Year (2006-2020), one could make the argument that an analysis showing a general trend of decreasing violence over a 14 year period with a sharp increase in the final year (2020) could be driven by a political bias in one direction or the other. For this analysis, including a model to highlight the difference of actual vs expected is helpful in showing objectivity in the conclusion, thus minimizing the potential for personal bias.